MODELOS ESTADISTICOS I

Objetivos:

1. Presentar los resultados básicos acerca del modelo de regresión lineal.
2. Discutir técnicas de diagnóstico y enfoques de análisis en caso de violaciones a supuestos.
3. Alternativas de modelación tales como: Regresión logística, Poisson, nolineal, noparamétrica.

Temario

1. **El modelo de regresión lineal**
   a) Estructura del modelo
   b) Estimadores suficientes
   c) Estimación vía mínimos cuadrados
   d) Estimación máximo verosímil
   e) Intervalos y regiones de confianza
   f) Teorema de Gauss-Markov (BLUE's)

2. **Diagnóstico en modelos de regresión lineal**
   a) Análisis de residuales
      i. Residuales estandarizados
      ii. Análisis gráfico
   b) Observaciones influyentes
      i. Puntos palanca (diagonal de matriz de proyección)
      ii. DFBETAS
      iii. D de Cook
   c) Factores de inflación de varianza
      i. Detección de colinealidades

3. **Alternativas ante violaciones de supuestos**
   a) Colinealidad
      i. Regresión ridge
      ii. Regresión en componentes principales
   b) Transformación de variables
      i. Transformaciones estabilizadoras de varianza
      ii. Transformaciones Box-Cox
   c) Heterogeneidad de varianza y correlación
      i. Mínimos cuadrados generalizados
      ii. Mínimos cuadrados ponderados
   d) Selección de variables
      i. Criterios para la selección de subconjuntos

ii. Métodos de selección por pasos (Stepwise)

4. **Modelos lineales generalizados**
   a) Estructura de los modelos lineales generalizados
      i. La familia exponencial
      ii. Funciones liga
      iii. Devianza
      iv. Ajuste vía mínimos cuadrados ponderados iterativamente
   b) Casos específicos
      i. Regresión logística
      ii. Regresión Poisson

5. **Temas especiales**
   a) Regresión noparamétrica
   b) Regresión nolineal
   c) Regresión robusta
   d) Regresión para cuantiles

Bibliografía

1. **Rawlings, J.O., Pantula, S.G. & Dickey, D.A.** (1998). Applied regression analysis: a research tool (2^{nd} ed.) Springer.
2. **McCullagh, P. & Nelder, J.A.** (1989). Generalized linear models (2^{nd} ed.) Chapman & Hall.
3. **Belsley, D.A., Kuh, E. & Welsch, R.E.** (1980). Regression diagnostics. Wiley.
4. **Carroll, R.J. & Ruppert, D.** (1988). Transformations and weighting in regression. Chapman & Hall.
5. **Bates, D.M. & Watts, D.G.** (1988). Nonlinear regression analysis and its applications. Wiley.
6. **Hardle, W.** (1990). Applied nonparametric regression. Cambridge.
7. **Koenker, R.** (2005). Quantile regression. Cambbridge.
8. **Christensen, R.** (1996). Plane answers to complex questions: The theory of linear models. (2^{nd} ed.) Springer.

Evaluación

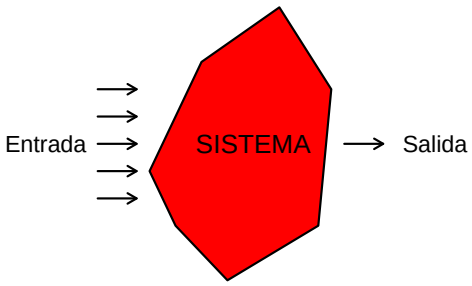
La nota final del curso estará basada en dos exámenes parciales (50%), un exámen final (30%) y resultados de tareas (20%).

Instructores del Curso

- Rogelio Ramos Quiroga, rramosq@cimat.mx, Oficina G1, ext. 49555
- Angélica Hernández Quintero, ahernandez@cimat.mx, Oficina B3, ext. 49616
- Guillermo Basulto Elías, guillermo@cimat.mx, Oficina M2, ext. 49677

Resumen de Clase 1: Miércoles 26 de enero

- Problemas en regresión:
  - modelar (de alguna forma) la relación entre la entrada y la salida de un sistema.
  - predecir y para un conjunto dado de covariables.
  - evaluar y comparar el impacto de diferentes covariables sobre la respuesta.



- En algunos casos es posible tener el grado de conocimiento del fenómeno, de forma que se puede escribir explícitamente un modelo matemático que describe el comportamiento del mismo

$$y = f(x_1, x_2, \cdots, x_k)$$

  Sin embargo, las técnicas de análisis que veremos se refieren precisamente al caso en el que no tenemos tal detalle de conocimiento (salvo cuando veamos regresión nolineal).

- Datos: $(x_1^T, y_1), \cdots, (x_n^T, y_n)$. El modelo de regresión implica

$$\mathsf{E}(y_i) = x_i^T \beta \qquad \text{y} \qquad \mathsf{Var}(y_i) = \sigma^2$$

  además de y_i's independientes y normales; i.e.

$$y \sim N_n(X\beta, \sigma^2 I)$$

- La validación del modelo implica revisar:
  - dependencia lineal de $\mathsf{E}(y_i)$ sobre x_i
  - heterogeneidad de varianza
  - falta de independencia
  - no normalidad
  - valores atípicos
  - variables regresoras estocásticas
  - colinealidad

- Estimación vía mínimos cuadrados: $\widehat{\beta}$ minimiza $\mathsf{SCE}(\beta) = (y - X\beta)^T(y - X\beta) = ||y - X\beta||^2$. Cuando β varía, $X\beta$ determina el subespacio de R^n generado por las columnas de X, de modo que $||y - X\beta||$ es la distancia de y al plano determinado por las columnas de X, sabemos que el vector que minimiza esta distancia es la proyección, p, de y sobre ese plano, así que p debe ser de la forma $p = X\widehat{\beta}$. Ahora, $y - p$ debe ser ortogonal a todos los vectores del plano, esto es $(y - p) \perp X\gamma$ para toda γ; entonces $(y - X\widehat{\beta})^T X\gamma = 0$ para toda γ, de modo que se debe cumplir que (ecuaciones normales) $X^T X\widehat{\beta} = X^T y$ y, de aquí que $\widehat{\beta} = (X^T X)^{-1} X^T y$. Notación: $\mathsf{SCE} = (y - X\widehat{\beta})^T(y - X\widehat{\beta})$.

- Estimación máximo verosímil: En general, la densidad normal multivariada, $y \sim N_n(\mu, \Sigma)$, es

$$f(y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \mathsf{exp}\left\{-\frac{1}{2}(y - \mu)\Sigma^{-1}(y - \mu)\right\}$$

en el caso $\mu = X\beta$ y $\Sigma = \sigma^2 I$, tenemos

$$f(y) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \mathsf{exp}\left\{-\frac{1}{2\sigma^2}(y - X\beta)(y - X\beta)\right\}$$

reescribimos el exponente como

$$\begin{aligned}(y - X\beta)(y - X\beta) &= (y - X\widehat{\beta} + X\widehat{\beta} - X\beta)^T(y - X\widehat{\beta} + X\widehat{\beta} - X\beta) \\ &= \left[(y - X\widehat{\beta}) + X(\widehat{\beta} - \beta)\right]^T \left[(y - X\widehat{\beta}) + X(\widehat{\beta} - \beta)\right] \\ &= \mathsf{SCE} + (\widehat{\beta} - \beta)^T X^T X(\widehat{\beta} - \beta)\end{aligned}$$

así que la verosimilitud se escribe como

$$\mathsf{L}(\beta, \sigma^2; y) = f(y) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \mathsf{exp}\left\{-\frac{1}{2\sigma^2}\mathsf{SCE} - \frac{1}{2\sigma^2}(\widehat{\beta} - \beta)^T X^T X(\widehat{\beta} - \beta)\right\}$$

de aquí, es fácil ver que los estimadores de máxima verosimilitud son

$$\begin{aligned}\widehat{\beta}_{\mathsf{MV}} &= \widehat{\beta} \\ \widehat{\sigma}^2 &= \mathsf{SCE}/n\end{aligned}$$

- Nota: En clase comentamos que el modelo lineal es una aproximación a la media condicional de y dado el vector de covariables x (p-dimensional) cuando (y, x^T) tiene una distribución normal multivariada. Para ver esto supongamos

$$\begin{bmatrix} y \\ x \end{bmatrix} \sim N_{1+p}\left(\begin{bmatrix} \mu_y \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}\right)$$

entonces $\mathsf{E}(y|x) = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu) = (\mu_y - \Sigma_{yx}\Sigma_{xx}^{-1}\mu) + \Sigma_{yx}\Sigma_{xx}^{-1}x \equiv \beta_0 + \beta_1^T x$

- Nota: En el caso de nonormalidad, tenemos que $\mathsf{E}(y|x) = h(x)$ y entonces, una aproximación de primer orden es:

$$\mathsf{E}(y|x) = h(x) \approx h(\mu) + \nabla^T h(\mu)(x - \mu) = [h(\mu) - \nabla^T h(\mu)\mu] + \nabla^T h(\mu)x \equiv \beta_0 + \beta_1^T x$$

Esto es, en ambos casos tenemos que el modelo lineal aproxima a la media condicional de y dado x.

Resumen de Clase 2: Lunes 31 de enero

El modelo estándar de regresión lineal es $y \sim N_n(X\beta, \sigma^2 I)$; la correspondiente verosimilitud, vimos en la clase pasada que se puede escribir como

$$\mathsf{L}(\beta, \sigma^2; y) = f(y) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathsf{SCE} - \frac{1}{2\sigma^2}(\widehat{\beta} - \beta)^T X^T X(\widehat{\beta} - \beta)\right\}$$

de aquí que, por el teorema de factorización, tenemos que SCE y $\widehat{\beta}$ son suficientes para β y σ^2.

Algo de terminología:

- Matriz de proyección: $P = X(X^TX)^{-1}$
- Valores predichos: $\widehat{y} = X\widehat{\beta} = X(X^TX)^{-1}X^Ty = Py$
- Residuales: $r = y - X\widehat{\beta} = (I - P)y$

Un paréntesis sobre matrices de proyección:

- Supongamos que X es $n \times p$ de rango p. Sea $\mathcal{L} = \mathcal{C}(X)$, el espacio de columnas de X. Decimos que P, $n \times n$, es la matriz de proyección sobre $\mathcal{L}$ si se cumplen:
  1. Si $u \in \mathcal{L} \quad \Rightarrow \quad Pu = u$
  2. Si $v \in \mathcal{C}^{\perp}(X) \quad \Rightarrow \quad Pv = 0$
- Considere $P = X(X^TX)^{-1}X^T$; veamos que P satisface las condiciones anteriores.
  1. Si $u \in \mathcal{L} \Rightarrow u = Xa$ para algún a. Entonces $Pu = X(X^TX)^{-1}X^TXa = Xa = u$. ✓
  2. Si $v \in \mathcal{C}^{\perp}(X) \Rightarrow X^Tv = 0$. Entonces $Pv = X(X^TX)^{-1}X^Tv = 0$. ✓
- Cuando decimos la matriz de proyección estamos implicando que P es única, efectivamente, supongamos que H, $n \times n$, también satisface las condiciones de matriz de proyección. Sea $y \in R^n$ cualquier vector, entonces, en forma única, podemos escribir $y = u + v$ con $u \in \mathcal{L}$ y $v \in \mathcal{L}^{\perp}$

  $$Hy = H(u + v) = Hu + Hv = u$$

  similarmente se ve que $Py = u$, por lo tanto $Hy = Py$ para toda y, de aquí que $P = H$.
- Nota: Si P es la matriz de proyección sobre $\mathcal{C}(X)$ entonces $\mathcal{C}(P) = \mathcal{C}(X)$. Veamos esto:
  - Sea $x \in \mathcal{C}(X)$ entonces $Px = x$ y por lo tanto $x \in \mathcal{C}(P)$, esto es $\mathcal{C}(X) \subset \mathcal{C}(P)$
  - Sea $p \in \mathcal{C}(P)$ entonces $p = Pa = P(u+v) = Pu+Pv = Pu+0 = u \in \mathcal{C}(X)$, esto es $\mathcal{C}(X) \supset \mathcal{C}(P)$
- Supongamos A simétrica $n \times n$, con $\mathcal{C}(A) = \mathcal{L} = \mathcal{C}(X)$, entonces A es la matriz de proyección. Para ver esto, tenemos que $\mathcal{C}(A) = \mathcal{C}(X)$ implica que existen matrices B y R tales que $X = AB$ y $A = XR$; ahora vemos que A satisface las condiciones de matriz de proyección:
  1. Si $u \in \mathcal{L} \Rightarrow u = Xa$, entonces $Au = AXa = AABa = ABa = Xa = u$ (aquí se uso que A es idempotente)
  2. Si $v \in \mathcal{L}^{\perp}(L) \Rightarrow X^Tv = 0$, entonces $Av = A^Tv = R^TX^Tv = 0$ (aquí se uso que A es simétrica)
- Ahora supongamos que P es la matriz de proyección sobre $\mathcal{L}$, queremos ver que P necesariamente es simétrica e idempotente.
  - Idempotencia: $PP = P[p_1, \cdots, p_n] = [Pp_1, \cdots, Pp_n] = [p_1, \cdots, p_n] = P$

– Simetría: Si $P^T(I-P)=0$ fuera cierto, entonces $P^T = P^TP$ y, como el lado derecho es simétrico entonces P^T sería simétrica, i.e. P sería simétrica.

Sean a y b vectores arbitrarios en R^n y escribamos $a = u_1 + v_1$ y $b = u_2 + v_2$, con $u_1, u_2 \in \mathcal{C}(P)$ $(= \mathcal{C}(X) = \mathcal{L})$ y $v_1, v_2 \in \mathcal{C}^{\perp}(P)$. Ahora

$$\begin{aligned} a^TP^T(I-P)b &= a^TP^Tb - a^TP^TPb = (Pa)^Tb - (Pa)^T(Pb) \\ &= u_1^T(u_2+v_2) - u_1^Tu_2 = u_1^Tv_2 = 0 \end{aligned}$$

como $a^TP^T(I-P)b = 0$ para todo a y b de R^n entonces $P^T(I-P) = 0$.

- Resumiendo: P es la proyección sobre $\mathcal{C}(P)$ si y solo si P es simétrica e idempotente.
- (ver el apéndice B.3 del libro de Christensen para ver más resultados sobre proyecciones)

Algunos resultados sobre esperanza y varianza. Supongamos que y es un vector aleatorio, entonces

1. $\mathsf{E}(y) = (\mathsf{E}(y_1), \cdots, \mathsf{E}(y_n))^T = (\mu_1, \cdots, \mu_n)^T = \mu$
2. $\mathsf{Var}(y) = \mathsf{E}[(y-\mu)(y-\mu)^T] = \mathsf{E}(yy^T) - \mu\mu^T$

$$\mathsf{Var}(y) = \Sigma = \begin{bmatrix} \mathsf{Var}(y_1) & \cdots & \mathsf{Cov}(y_1, y_n) \\ \vdots & \ddots & \vdots \\ \mathsf{Cov}(y_n, y_1) & \cdots & \mathsf{Var}(y_n) \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix}$$

3. $\mathsf{E}(Ay) = A\mathsf{E}(y)$
4. $\mathsf{Var}(Ay) = A\mathsf{Var}(y)A^T$
5. Comentamos que este resultado era consecuencia de que el ij-ésimo elemento de ABC es $a_i^TBc_j$ donde a_i^T es el i-ésimo renglón de A y c_j es la j-ésima columna de C

Resumen de Clase 3: Viernes 4 de febrero

- Teorema de Gauss-Markov: Supongamos que y es tal que $\mathsf{E}(y) = X\beta$ (rango completo para X) y $\mathsf{Var}(y) = \sigma^2 I$. Sea $\phi = c^T\beta$ cualquier combinación lineal de parámetros de regresión. Sea t cualquier estimador lineal e insesgado de ϕ. Entonces $\widehat{\phi} = c^T\widehat{\beta}$ es el "mejor" estimador para ϕ ("mejor" significa "de varianza mínima").

  - t lineal implica que $t = a^T y$ para algún a; insesgadez implica que $c^T\beta = \mathsf{E}(t) = \mathsf{E}(a^T y) = a^T\mathsf{E}(y) = a^T X\beta$; esto es $c^T\beta = a^T X\beta$ para toda β; de aquí que $c^T = a^T X$.
  - La idea básica de la prueba del teorema de Gauss-Markov es ver que $\mathsf{Var}(t) = \mathsf{Var}(c^T\widehat{\beta}) + \Delta$ donde $\Delta \geq 0$:

$$\begin{aligned}\mathsf{Var}(t) = \mathsf{Var}(a^T y) &= \mathsf{Var}(a^T y - c^T\widehat{\beta} + c^T\widehat{\beta}) = \mathsf{Var}(a^T y - a^T X\widehat{\beta} + c^T\widehat{\beta}) \\ &= \mathsf{Var}[a^T(y - X\widehat{\beta}) + c^T\widehat{\beta}] = \mathsf{Var}(c^T\widehat{\beta}) + \mathsf{Var}[a^T(y - X\widehat{\beta})] + 2\mathsf{Cov}[a^T(y - X\widehat{\beta}), c^T\widehat{\beta}]\end{aligned}$$

    ahora, $\mathsf{Cov}[a^T(y - X\widehat{\beta}), c^T\widehat{\beta}] = \mathsf{Cov}[a^T(I - P)y, a^T Py] = \sigma^2 a^T(I - P)Pa = 0$. De aquí que $\mathsf{Var}(t) = \mathsf{Var}(c^T\widehat{\beta}) + \Delta$, donde $\Delta = \mathsf{Var}[a^T(y - X\widehat{\beta})]$ y esto muestra el teorema.
  - Note que $\Delta = 0$ si y sólo si $a^T(y - X\widehat{\beta}) =$ cte.; esto es $t - c^T\widehat{\beta} =$ cte., y como tanto t como $c^T\widehat{\beta}$ tienen el mismo valor esperado ($= c^T\beta$) entonces necesariamente cte. $= 0$. Esto es, $\mathsf{Var}(c^T\widehat{\beta}) < \mathsf{Var}(t)$ para todo t lineal e insesgado ($t \neq c^T\widehat{\beta}$).

- Algunas propiedades distribucionales del estimador de máxima verosimilitud de β bajo el modelo lineal estándar $y \sim N(X\beta, \sigma^2 I)$, que nos permitirán construír intervalos y regiones de confianza:

  - Como $\widehat{\beta} = (X^T X)^{-1} X^T y$, entonces $\widehat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$. Esta propiedad es consecuencia de que transformaciones lineales de vectores normales son normales.
  - Como $\widehat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$ entonces $\widehat{\beta}_i \sim N(\beta_i, \sigma^2(X^T X)^{-1}_{ii})$. Esta propiedad es consecuencia de que marginales de vectores normales son normales.
  - Estas y otras propiedades de la normal serán tratadas en la siguiente clase.

- Pruebas de hipótesis sobre los parámetros de regresión. Consideremos la hipótesis

$$H_0 : \beta \in \Omega_0$$

  - La prueba estándar en Estadística se basa en el estadístico Cociente de Verosimilitudes

$$\Lambda = \frac{\max\limits_{\beta,\sigma^2 \in \Omega_0} L(\beta, \sigma^2; y)}{\max\limits_{\beta,\sigma^2} L(\beta, \sigma^2; y)}$$

  - Es claro que el máximo en el denominador se alcanza en $\widehat{\beta} = (X^T X)^{-1} X^T y$ y $\widehat{\sigma}^2 = (y - X\widehat{\beta})^T(y - X\widehat{\beta})/n$; así que

$$\max_{\beta,\sigma^2} L(\beta, \sigma^2; y) = \frac{1}{(2\pi)^{n/2}(\widehat{\sigma}^2)^{n/2}} \exp\left\{-\frac{1}{2\widehat{\sigma}^2}(y - X\widehat{\beta})^T(y - X\widehat{\beta})\right\} = \frac{e^{-n/2}}{(2\pi)^{n/2}(\widehat{\sigma}^2)^{n/2}} = \frac{c}{(\widehat{\sigma}^2)^{n/2}}$$

  - Con respecto al numerador, note que la verosimilitud es

$$L(\beta, \sigma^2; y) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\}$$

y la logverosimilitud

$$l(\beta, \sigma^2; y) = -\frac{1}{2\sigma^2}||y - X\beta||^2 - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2)$$

de aquí, el máximo para β corresponde al mínimo de $||y - X\beta||$ restringiendo los posibles valores de β al espacio Ω_0; más adelante veremos formas explícitas de este mínimo (optimización con restricciones), por lo pronto no lo necesitamos explícitamente; sea $\widehat{\beta}_R$ ese valor mínimo, es fácil ver que la solución para σ^2 es $\widehat{\sigma}^2 = ||y - \widehat{\beta}_R||^2/n$ y así, el numerador del cociente de verosimilitudes es:

$$\max_{\beta, \sigma^2 \in \Omega_0} L(\beta, \sigma^2; y) = L(\widehat{\beta}_R, \widehat{\sigma}^2_R; y) = \frac{c}{(\widehat{\sigma}^2_R)^{n/2}}$$

- Por lo tanto, el estadístico de prueba de $H_0 : \beta \in \Omega_0$ es

$$\Lambda = \left(\frac{\widehat{\sigma}^2}{\widehat{\sigma}^2_R}\right)^{n/2}$$

- La regla de decisión consiste en rechazar H_0 si Λ es pequeño $\cdots$ el que tan pequeño es pequeño está en función de las propiedades distribucionales de Λ. El valor crítico, λ, se escoge de tal forma que $P(\Lambda < \lambda \mid H_0) = \alpha$, donde α es la probabilidad de tomar una mala decisión cuando H_0 es cierta. En general, la distribución de Λ puede ser difícil de determinar y es por ello que con frecuencia se usa una aproximación válida asintóticamente: $-2\log(\Lambda) \sim \chi^2_g$. Ahora bien, para el caso del modelo lineal estándar vamos a poder calcular la distribución exacta de (una transformación de) Λ.
- Note que $\Lambda < \lambda$ si y sólo si $(\widehat{\sigma}^2_R/\widehat{\sigma}^2) > c$, donde c es una constante. Equivalentemente

$$\Lambda < \lambda \quad \Leftrightarrow \quad \frac{SCE_0 - SCE}{SCE} > d, \qquad d \text{ una constante}$$

donde $SCE = ||y - X\widehat{\beta}||^2$ y $SCE_R = ||y - X\widehat{\beta}_R||^2$

- Veremos que, bajo H_0, $SCE_0 - SCE \sim \sigma^2\chi^2_{\nu_0}$ y que se distribuye en forma independiente de $SCE \sim \sigma^2\chi^2_{\nu}$. Por lo tanto, bajo H_0:

$$F = \frac{(SCE_0 - SCE)/\nu_0}{SCE/\nu} \sim F^{\nu_0}_{\nu}$$

esto es

$$\Lambda < \lambda \quad \Leftrightarrow \quad F > f$$

de modo que el valor crítico, f, para la prueba puede obtenerse como un cuantil de una distribución conocida.

- Note que $SCE = ||y - X\widehat{\beta}||^2 = (y - X\widehat{\beta})^T(y - X\widehat{\beta}) = y^T(I - P)y$ y algo similar se obtendrá para $SCE_R - SCE$, de modo que para obtener los resultados distribucionales antes mencionados necesitaremos estudiar el comportamiento distribucional de expresiones de la forma y^TAy.

Tarea 1. Modelos Estadísticos I

1. Sea P la matriz de proyección sobre $\mathcal{C}(P)$. Muestre que
   (a) $\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}^2 = \text{rango}(P)$
   (b) $(I-P)$ es la matriz de proyección de $\mathcal{C}^{\perp}(P)$

2. (a) Si A es idempotente, muestre que los valores propios de A son 1's y 0's
   (b) Si X es $n \times p$, muestre que $\mathcal{C}(XX^T) = \mathcal{C}(X)$

3. Considere el modelo de regresión lineal simple $y_i = \beta_0 + x_i\beta_1 + e_i,\ i = 1, \cdots, n$ con las e_i's independientes con media 0 y varianza constante. Considere las matrices

$$X_1 = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{y} \quad X_2 = \begin{bmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{bmatrix}$$

   (a) Muestre que $\mathcal{C}(X_1) = \mathcal{C}(X_2)$. Muestre que $\hat{y}_i$ tiene el mismo valor independientemente de si usamos el modelo $\text{E}(y_i) = \beta_0 + x_i\beta_1$ o $\text{E}(y_i) = \beta_0 + (x_i - \bar{x})\beta_1$.
   (b) Muestre que los estimadores de los coeficientes de regresión están no correlacionados bajo el segundo modelo mientras que si lo pueden estar bajo el primero.

4. (a) Si $x \sim N_d(\mu, \Sigma)$, veremos que $a^T x$ tiene una distribución normal univariada para toda a no nula. Muestre que el converso también es cierto, esto es, muestre que si para toda a no nula, $a^T x$ tiene una distribución normal univariada, entonces y tiene una distribución normal multivariada (use la función generatriz de momentos) (este resultado es llamado Teorema de Cramér-Wold).
   (b) Suponga que $x \sim N_d(\mu, \Sigma)$. En clase veremos, usando la función generadora de momentos, que Ax tiene una distribución normal. Suponga que A es nosingular, muestre directamente que Ax es normal, esto es, encuentre la densidad de $y = Ax$ usando el resultado para transformaciones de variables: Si $x \sim f(x)$ entonces $y = H(x) \sim g(y)$ donde $g(y) = f(H^{-1}(y))|J_{H^{-1}}|$.

5. Suponga que $x \sim N_3(0, \Sigma)$, donde

$$\Sigma = \begin{bmatrix} 7 & 3 & 2 \\ 3 & 4 & 1 \\ 2 & 1 & 2 \end{bmatrix}$$

   (a) Encuentre la distribución marginal de $(x_1, x_3)^T$.
   (b) Encuentre la distribución condicional de $(x_1, x_3)^T$ dado que $x_2 = 1$.

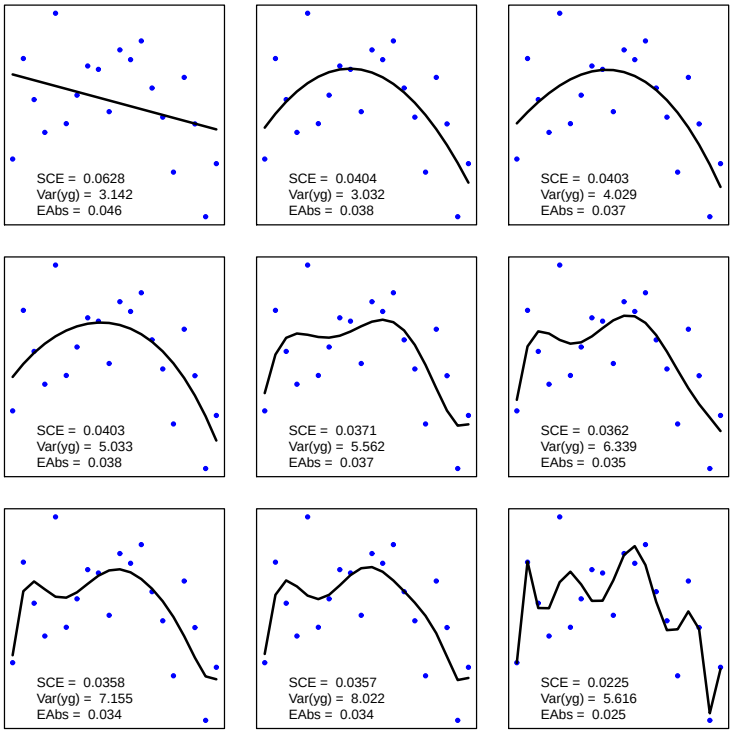
6. Ejecute las siguientes líneas en R:

```
n  <- 20
x  <- seq(0,1,length=n)
set.seed(76767)
b0 <- 5; b1 <- .2; b2 <- -.3
y  <- b0+b1*x+b2*x^2 + rnorm(n,0,.05)
X  <- rep(1,n)
```

```
par(mfrow=c(3,3),mar=c(1,1,1,1))
for(i in 1:9){
 plot(x,y,xaxt="n",yaxt="n",pch=20,col="blue")
 X   <- cbind(X,x^i)
 bg  <- solve(t(X)%*%X,t(X)%*%y)
 P   <- X%*%solve(t(X)%*%X,t(X))
 yg  <- X%*%bg
 lines(x,yg,lwd=2)
 sce <- sum((y-yg)^2)
 vyg <- (sce/n)*mean(diag(P))
 eab <- mean(abs(y-yg))
 legend(0,4.93,legend=c(paste("SCE = ",round(sce,4)),
  paste("Var(yg) = ",round((10^4)*vyg,3)),
  paste("EAbs = ",round(eab,3))),bty="n") }
```

(a) ¿Qué son esas gráficas?, ¿Qué cálculos se están haciendo?.

(b) Dé algunas ideas sobre cómo elegir un modelo.



Fecha de entrega: Jueves 17 de febrero.

Problemas Extra 1. Modelos Estadísticos I

1. Sean $y_1, n\times 1$, y $y_2, m\times 1$, vectores aleatorios con $\mathsf{E}(y_i)=\mu_i,\ \mathsf{Var}(y_i)=\Sigma_i,\ i=1,2,\ \mathsf{Cov}(y_1,y_2)=\Sigma_{12}$ y $\mathsf{Cov}(y_2,y_1)=\Sigma_{21}$. Sean A y B matrices $r\times n$ y $r\times m$ respectivamente. Demuestre que

$$\mathsf{Var}(Ay_1+By_2)=A\Sigma_1A^T+A\Sigma_{12}B^T+B\Sigma_{21}A^T+B\Sigma_2B^T$$

2. Sean C y D matrices de tamaños adecuados. Con la notación del problema anterior, calcule

$$\mathsf{Cov}(\,Ay_1+By_2\,,\,Cy_1+Dy_2\,)$$

3. Sea A una matriz $n\times n$ y b un vector $n\times 1$. Encuentre el valor, λ^*, que minimiza $\|Ab-\lambda b\|^2$.

4. Sea X una matriz $n\times p$ de rango r, con $r<p$, $(p\leq n)$. Considere el espacio de columnas de X, $\mathcal{C}(X)$. La dimensión de $\mathcal{C}(X)$ es r, así que cualquiera de sus bases consta de r elementos.

   (a) Sean B_1 y B_2 matrices $n\times r$ cuyas columnas son bases de $\mathcal{C}(X)$, suponga además que las columnas de B_2 son ortonormales. Pruebe que

$$B_1(B_1^TB_1)^{-1}B_1^T=B_2B_2^T$$

   (b) Sea P una matriz simétrica e idempotente tal que $\mathcal{C}(P)=\mathcal{C}(X)$. Pruebe que

$$P=B_2B_2^T$$

5. Considere el sistema de ecuaciones lineales

$$\begin{aligned}2x_1+x_2-x_3&=1\\4x_1+2x_2-2x_3&=3\end{aligned}$$

   (a) Muestre que el sistema no es consistente.

   (b) Escriba el sistema como $Ax=b$. Si x_0 es un vector 3×1 cualquiera, entonces Ax_0 no es exactamente igual a b, sin embargo deseamos de todos modos obtener una solución aproximada; defina el "residual de x_0" como $r_0=Ax_0-b$. Encuentre el vector x que tenga el residual más pequeño en magnitud.

   (c) Interprete geométricamente el resultado.

6. Denotemos por 1_{n_j} al vector $n_j\times 1$ formado por puros unos. Considere la matriz

$$X=\begin{bmatrix}1_{n_1}&1_{n_1}&0&0&\cdots&0\\1_{n_2}&0&1_{n_2}&0&\cdots&0\\1_{n_3}&0&0&1_{n_3}&\cdots&0\\\vdots&\vdots&\vdots&\vdots&\ddots&\vdots\\1_{n_k}&0&0&0&\cdots&1_{n_k}\end{bmatrix}$$

   donde los vectores de ceros son de tamaños apropiados. Esta es la matriz de diseño para un modelo con un criterio de clasificación. ("Diseño completamente al azar con diferente número de repeticiones por tratamiento").

(a) Encuentre una base ortogonal para $\mathcal{C}(X)$.

(b) Encuentre la matriz de proyección sobre $\mathcal{C}(X)$.

7. Suponga que $y \sim (\mu, \Sigma)$ (no necesariamente normal). Pruebe que $\mathsf{E}(y^T Ay) = \mu^T A\mu + \text{traza}(A\Sigma)$
Sugerencia: $\mathsf{E}(y^T Ay) = \mathsf{E}(\text{traza}(y^T Ay)) = \mathsf{E}(\text{traza}(Ayy^T)) = \text{traza}(A\mathsf{E}(yy^T)) = \text{etc.}$

8. Suponga que $y \sim N_2(\mu, \Sigma)$, donde $y = (y_1, y_2)^T$, $\mu = (\mu_1, \mu_2)^T$ y $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$ con Σ positiva definida. Escriba explícitamente la densidad conjunta de y_1 y y_2 (explícitamente quiere decir que hay que calcular explícitamente $|\Sigma|$ y Σ^{-1}).

9. Sea y un vector aleatorio con $\mathsf{E}(y) = 0$ y $\text{Var}(y) = \Sigma$ con Σ $n \times n$ de rango r.

(a) Pruebe que es posible escribir

$$y_j = \sum_{i \neq j} c_i y_i$$

para alguna j y constantes c_i's. En otras palabras, hay que probar que si y tiene varianza semidefinida positiva entonces podemos encontrar relaciones determinísticas entre sus elementos. *Sugerencia*: Si $r(\Sigma) < n$ entonces $\Sigma a = 0$ para algún $a \neq 0$, considere ahora a la variable aleatoria $a^T y$.

(b) Pruebe que $y \in \mathcal{C}(\Sigma)$ con probabilidad 1. *Sugerencia*: Ver que lo obtenido en el inciso anterior es válido para toda $a \in \mathcal{C}^{\perp}(\Sigma)$.

10. Supongamos que $u_i \sim N_1(\mu_i, \sigma^2)$, $i = 1, \cdots, n$ y son independientes entre sí. Definamos

$$V = \sum_{i=1}^{n} b_i u_i \quad \text{donde} \quad b_i = \frac{\mu_i}{\sqrt{\sum_{j=1}^{n} \mu_j^2}} \quad \text{y} \quad W = \sum_{i=1}^{n} u_i^2 - V^2$$

Pruebe que V y W son independientes.

Resumen de Clase 4: Miércoles 9 de febrero

Resultados relacionados con la Normal y formas cuadráticas.

1. Sea y un vector aleatorio $n \times 1$ con densidad

$$f(y) \;=\; \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2} y^T y\right\} \;=\; \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} \exp\left\{-\frac{1}{2} y_i^2\right\}$$

   en este caso diremos que y tiene una distribución normal estándar multivariada y lo denotaremos por $y \sim N(0, I_n)$.

2. Sea y un vector aleatorio $n \times 1$ con $\mathsf{E}(y) = \mu$ y $\mathsf{Var}(y) = \Sigma$ con $\text{rango}(\Sigma) = k$ (k puede ser $< n$). Por ser Σ nonegativa tenemos que $\Sigma = \Gamma\,\Gamma^T$ donde Γ es $n \times k$ con columnas ortogonales.

   Diremos que y tiene una distribución normal multivariada de rango k si y tiene la misma distribución que el vector $\mu + \Gamma z$, donde $z \sim N(0, I_k)$.

   Notación: $y \sim N(\mu, \Sigma)$. Note que $\mathsf{E}(y) = \mu$ y $\mathsf{Var}(y) = \Gamma\,\Gamma^T = \Sigma$.

3. Sea y u vector aleatorio $n \times 1$. Decimos que $M_y(t) = \mathsf{E}(e^{t^T y})$ es la función generatriz de momentos de y si esa esperanza existe en el cubo $-h < t_i < h,\ i = 1, \cdots, n$, para algún $h > 0$.

4. Si $y \sim N(0, I_n)$ entonces $M_y(t) = e^{t^T t/2}$. Puede verse que

$$\begin{aligned} M_y(t) = \mathsf{E}(e^{t^T y}) &= \int_{y \in R^n} \exp\left\{t^T y\right\}\ \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2} y^T y\right\} dy \\ &= \prod_{i=1}^{n} \int_{y_i} \frac{1}{(2\pi)^{1/2}} \exp\left\{-\frac{1}{2}(y_i^2 - 2t_i y_i)\right\} dy_i \;=\; \cdots \;= e^{t^T t/2} \end{aligned}$$

5. Si $y \sim N(\mu, \Sigma)$ entonces $M_y(t) = \exp\left\{t^T \mu + \frac{1}{2} t^T \Sigma t\right\}$. Para ver esto, notamos que:

$$M_y(t) = \mathsf{E}(e^{t^T y}) = \mathsf{E}(e^{t^T(\mu + \Gamma z)}) = \mathsf{E}(e^{t^T \mu + t^T \Gamma z}) = e^{t^T \mu}\, \mathsf{E}(e^{(\Gamma^T t)^T z}) = e^{t^T \mu}\, M_z(\Gamma^T t) = e^{t^T \mu + \frac{1}{2} t^T \Sigma t}$$

6. Si x es un vector aleatorio $p \times 1$ con $x \sim N(\mu, \Sigma)$, entonces, para cualquier matriz B, $q \times p$, de constantes y cualquier a vector, se cumple que $y = a + Bx \sim N(a + B\mu, B\Sigma B^T)$. Esto es, transformaciones lineales de normales son normales. Para ver esto usar la función generatriz d emomentos.

7. Sea $y \sim N(\mu, \Sigma)$, particionemos los diferentes vectores y matrices como

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \begin{matrix} n_1 \times 1 \\ n_2 \times 1 \end{matrix}, \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

   Entonces $y_1 \sim N(\mu_1, \Sigma_{11})$; i.e. Marginales de normales son normales. Para ver esta propiedad note que $y_1 = [I, O] \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = By$; y luego aplicamos la propiedad anterior sobre transformaciones lineales de normales.

8. Sea $y \sim N(\mu, \Sigma)$ con las particiones como en el inciso anterior. Entonces y_1 es independiente de y_2 si y sólo si, $\Sigma_{12} = 0$.

   ($\Rightarrow$) Si y_1 y y_2 son independientes entonces, sabemos que $0 = \text{Cov}(y_1, y_2) = \Sigma_{12}$.

   ($\Leftarrow$) Ahora, supongamos que $\Sigma_{12} = 0$ y supongamos además que rango$(\Sigma) = k$, entonces

   $$\Sigma = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$$

   donde rango$(\Sigma_{11}) = k_1$ y rango$(\Sigma_{22}) = k_2$, con $k_1 + k_2 = k$. Ahora, escribamos $\Sigma_{ii} = \Gamma_i \Gamma_i^T$, donde Γ_i es $n_i \times k_i$ de rango k_i $(i = 1, 2)$, entonces

   $$\Sigma = \begin{bmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{bmatrix} \begin{bmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{bmatrix}^T \equiv \Gamma \Gamma^T$$

   Por otro lado, sabemos que $y \stackrel{d}{=} \mu + \Gamma z$ donde $z \sim N(0, I_k)$; entonces

   $$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \stackrel{d}{=} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \mu_1 + \Gamma_1 z_1 \\ \mu_2 + \Gamma_2 z_2 \end{bmatrix}$$

   dada la independencia de z_1 y z_2 se sigue (como se comentó en clase) que y_1 y y_2 son independientes.

9. Distribuciones Condicionales (para el caso nosingular) (el caso singular consultarlo en el Rao). Sea $y \sim N(\mu, \Sigma)$ con las particiones como antes. Queremos obtener la distribución de $y_1 \,|\, y_2$. Consideremos la matriz

   $$A = \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix}$$

   Note que $Ay \sim N(A\mu, A\Sigma A^T)$. Ahora, haciendo los cálculos se ve que

   $$Ay = \begin{bmatrix} y_1 - \Sigma_{12}\Sigma_{22}^{-1} y_2 \\ y_2 \end{bmatrix} \equiv \begin{bmatrix} u \\ y_2 \end{bmatrix}, \quad A\mu = \begin{bmatrix} \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\ y_2 \end{bmatrix}, \quad A\Sigma A^T = \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$$

   Como $(u^T, y_2$ es normal y con Cov$(u, y_2) = 0$ entonces u es independiente de y_2 y, por lo tanto, la condicional de u dado y_2 es la misma que la marginal de u; esto es

   $$u \,|\, y_2 \sim N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

   pero $u = y_1 - \Sigma_{12}\Sigma_{22}^{-1} y_2$, entonces u condicionada a y_2 no es mas que $y_1 -$ cte, por lo tanto $y_1 \,|\, y_2 \sim N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \text{cte}, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$. Esto es

   $$y_1 \,|\, y_2 \sim N(\, \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2)\,,\, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\,)$$

10. Sea $y \sim N(\mu, \Sigma)$ con Σ positiva definida, entonces, la densidad de y es

    $$f(y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right\}$$

    Para ver esto, calcular la función generatriz de momentos asociada con f y ver que es la misma que la mostrada en el inciso 5.

11. Si $z_1 \sim N(0, 1)$ entonces $z_1^2 \sim \chi_1^2$.

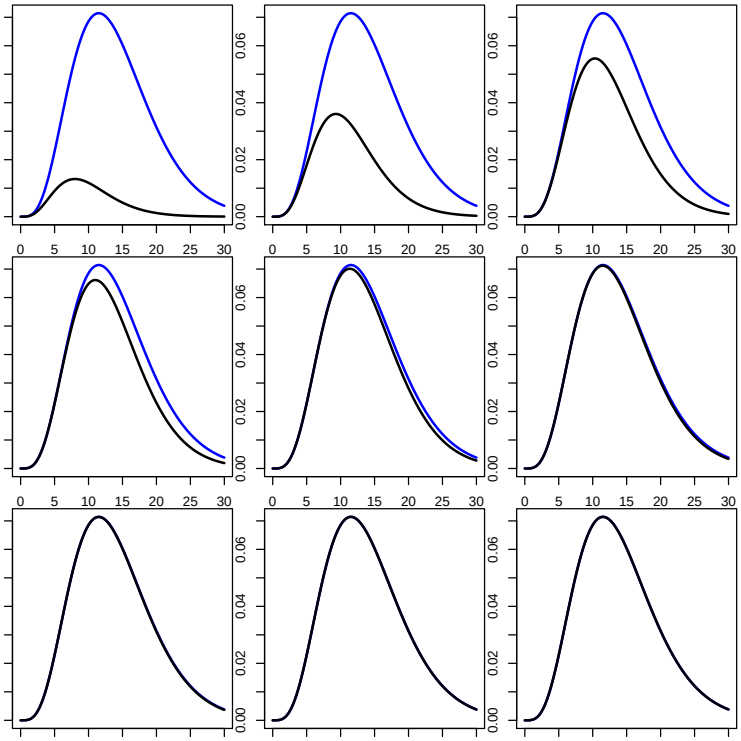
12. Si $z = (z_1, \cdots, z_n)^T \sim N(0, I)$ entonces $z^T z \sim \chi^2_n$.

13. Si $z = (z_1, \cdots, z_n)^T \sim N(\mu, I)$ entonces $z^T z \sim \chi^2_{n,\lambda}$; esto es, $z^T z$ tiene una distribución ji-cuadrada nocentral con n grados de libertad y parámetro de nocentralidad $\lambda = \frac{1}{2}\mu^T\mu$.

    Comentaremos en la siguiente clase que la densidad de $u = z^T z$ es una ji-cuadrada nocentral, $\chi^2_{n,\lambda}$ cuya forma es:

$$f(u) = \sum_{r=0}^{\infty} \frac{e^{-\lambda}\lambda^r}{r!} \chi^2_{2r+n}(u)$$

    esto es, esta densidad es una mezcla de ji-cuadradas centrales con pesos Poisson. La siguiente gráfica muestra la suma de los primeros términos de la mezcla en forma incremental.




```
# Primeras aproximaciones a una ji-cuadrada no-central
# usando una mezcla Poisson de ji-cuadradas centrales.

n   <- 10
lam <- 2
x   <- seq(0,30,length=200)
y   <- dchisq(x,n,2*lam)
yy  <- rep(0,200)

par(mfrow=c(3,3),mar=c(1,1,1,1))

for(r in 0:8){
 plot(x,y,type="l",lwd=2,col="blue",xlab="",ylab="")
 yy <- yy + dpois(r,lam)*dchisq(x,n+2*r)
 lines(x,yy,lwd=2)}
```

Resumen de Clase 5: Lunes 14 de febrero

- Comentamos la clase pasada que si $y = (y_1, \cdots, y_n)^T \sim N(\mu, I)$ entonces a la distribución de $a = y^T y$ se le llama ji-cuadrada nocentral y está parametrizada por n (grados de libertad) y por λ (parámetro de nocentralidad); denotamos esto por $a = y^T y \sim \chi^2_{n,\lambda}$. Como veremos, λ toma el valor $\lambda = \frac{1}{2}\mu^T\mu$.

  La densidad de $a = y^T y$ tiene la forma:

$$h(a) = \sum_{r=0}^{\infty} \frac{e^{-\lambda}\lambda^r}{r!} \chi^2_{2r+n}(a)$$

  esto es, esta densidad es una mezcla de ji-cuadradas centrales con pesos Poisson. Damos enseguida un esbozo sobre como se deduce esta expresión.

  i. Tenemos que $y \sim N_n(\mu, I)$, así que, para cualquier B tenemos $w = By \sim N(B\mu, BB^T)$.

  ii. Tomemos, en particular, B una matriz ortogonal cuyo primer renglón es $\mu^T/||\mu||$, de modo que $w = By \sim N(\nu, I)$ donde $\nu = (||\mu||, 0, \cdots, 0)^T$.

  iii. Note que $w^T w = y^T B^T B y = y^T y$, así que $w^T w \stackrel{d}{=} y^T y$.

  iv. Note que $w^T w = w_1^2 + w_2^2 + \cdots + w_n^2 = w_1^2 + (w_2^2 + \cdots + w_n^2) \equiv U + V$. Note que U y V son independientes con $V \sim \chi^2_{n-1}$ y $U = w_1^2$ donde $w_1 \sim N(||\mu||, 1)$. Así que la densidad de $y^T y$ es la convolución de las densidades de U y V.

  v. Puede verse que la densidad de U es

$$f(u) = \frac{1}{2\sqrt{2\pi}} u^{-\frac{1}{2}} e^{-\frac{1}{2}(u+||\mu||^2)} \left[e^{||\mu||\sqrt{u}} + e^{-||\mu||\sqrt{u}} \right]$$

  vi. Alternativamente, puede verse, expandiendo las exponenciales entre corchetes, que:

$$f(u) = \sum_{r=0}^{\infty} \frac{e^{-\lambda}\lambda^r}{r!} \chi^2_{2r+1}(u)$$

  donde $\lambda = \frac{1}{2}||\mu||^2$ y $\chi^2_{2r+1}(u)$ es la densidad de una ji-cuadrada central con $2r+1$ grados de libertad.

  vii. Finalmente, la densidad de a es la convolución de las densidades de U y V:

$$h(a) = \int_u f(u)g(a-u)du = \sum_{r=0}^{\infty} \frac{e^{-\lambda}\lambda^r}{r!} \int_u \chi^2_{2r+1}(u)\chi^2_{n-1}(a-u)du = \sum_{r=0}^{\infty} \frac{e^{-\lambda}\lambda^r}{r!} \chi^2_{2r+n}(a)$$

- Resumiendo, tenemos el Resultado 1: Si $y \sim N(\mu, I_n)$ entonces $y^T y \sim \chi^2_{n,\lambda}$, donde $\lambda = \frac{1}{2}\mu^T\mu$.

- Resultado 2: Si $y \sim N(0, I_n)$ entonces $y^T A y \sim \chi^2_k$ si y sólo si A es simétrica e idempotente de rango k.

  i. Por ser A simétrica e idempotente de rango k tenemos que existe B tal que $A = BB^T$ donde B es $n \times k$ y $B^T B = I_k$.

  ii. Ahora, $y^T A y = y^T BB^T y \equiv w^T w$ donde $w = B^T y \sim N(0, I_k)$. Por lo tanto $y^T A y \sim \chi^2_k$.

  iii. (El converso es cierto pero no lo veremos).

- Resultado 3: Si $y \sim N(\mu, I_n)$ entonces $y^T A y \sim \chi^2_{k,\lambda}$ si y sólo si A es simétrica e idempotente de rango k (en este caso $\lambda = \frac{1}{2}\mu^T A\mu$).

i. Como en el resultado anterior, tenemos que $A = BB^T$ y $y^TAy \equiv w^Tw$ donde $w = B^Ty \sim N(B^T\mu, I_k)$. Entonces, por el Resultado 1, $y^TAy \sim \chi^2_{k,\lambda}$ donde k es el rango de A y $\lambda = \frac{1}{2}\mu^TBB^T\mu = \frac{1}{2}\mu^TA\mu$.

ii. (El converso es cierto pero no lo veremos).

- Resultado 4: Si $y \sim N(\mu, \Sigma)$ con Σ positiva definida, entonces $y^TAy \sim \chi^2_{k,\lambda}$ si y sólo si $A\Sigma$ es idempotente (en este caso $\lambda = \frac{1}{2}\mu^TA\mu$ y k es el rango de A).

i. Como Σ es positiva definida tenemos que $\Sigma = VDV^T = VD^{\frac{1}{2}}D^{\frac{1}{2}}V^T = (VD^{\frac{1}{2}}V^T)(VD^{\frac{1}{2}}V^T) \equiv \Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}$, donde $\Sigma^{\frac{1}{2}}$ es simétrica e invertible.

ii. Consideremos $w = \Sigma^{-\frac{1}{2}}(y - \mu) \sim N(0, I_n)$, entonces $y = \mu + \Sigma^{\frac{1}{2}}w$ y

$$y^TAy = (\mu + \Sigma^{\frac{1}{2}}w)^TA(\mu + \Sigma^{\frac{1}{2}}w) = (\Sigma^{-\frac{1}{2}}\mu + w)^T\ \Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\ (\Sigma^{-\frac{1}{2}}\mu + w) \equiv v^TBv$$

donde $v = \Sigma^{-\frac{1}{2}}\mu + w \sim N(\Sigma^{-\frac{1}{2}}\mu, I_n)$ y $B = \Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}$

iii. Note que $A\Sigma$ idempotente implica que $A\Sigma A\Sigma = A\Sigma$ y esto implica que $A\Sigma A = A$ (esto es, Σ es una inversa generalizada de A), a su vez, esto implica que $\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}} = \Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}$, i.e. B es idempotente.

iv. Entonces, por el Resultado 3, tenemos que $y^TAy = v^TBv \sim \chi^2_{k,\lambda}$ donde k es el rango de B (esto es, k es igual al rango de A) y $\lambda = \frac{1}{2}(\Sigma^{-\frac{1}{2}}\mu)^TB(\Sigma^{-\frac{1}{2}}\mu) = \frac{1}{2}\mu^T\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\mu = \frac{1}{2}\mu^TA\mu$.

v. (El converso es cierto pero no lo veremos) (Hay también una versión de este resultado para el caso de Σ singular).

- Algunos resultados sobre independencia: Supongamos que $y \sim N(\mu, \sigma^2I_n)$, A y B son matrices $p \times n$ y $q \times n$ respectivamente, y C y D son nonegativas $n \times n$. Entonces:

a) Ay y By son independientes $\Leftrightarrow\ AB^T = 0$

b) Ay y y^TCy son independientes $\Leftrightarrow\ AC = 0$

c) y^TCy y y^TDy son independientes $\Leftrightarrow\ CD = 0$

– Para ver (a), note que

$$\begin{bmatrix} Ay \\ By \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix} y \sim N_{p+q}\left(\begin{bmatrix} A \\ B \end{bmatrix} \mu\ ,\ \sigma^2 \begin{bmatrix} AA^T & AB^T \\ BA^T & BB^T \end{bmatrix} \right)$$

de aquí que Ay y By son independientes $\Leftrightarrow\ AB^T = 0$.

Resumen de Clase 6: Miércoles 16 de febrero

- Algunos resultados sobre independencia: Supongamos que $y \sim N(\mu, \sigma^2 I_n)$, A y B son matrices $p \times n$ y $q \times n$ respectivamente, y C y D son nonegativas $n \times n$. Entonces:

  a) Ay y By son independientes $\Leftrightarrow$ $AB^T = 0$

  b) Ay y $y^T Cy$ son independientes $\Leftrightarrow$ $AC = 0$

  c) $y^T Cy$ y $y^T Dy$ son independientes $\Leftrightarrow$ $CD = 0$

  - El resultado (a) lo vimos la clase pasada.
  - Para ver (b) notamos que, por ser C nonegativa podemos escribir $C = E^T E$, entonces

    $$AC = 0 \quad \Rightarrow \quad AE^T E = 0 \quad \Rightarrow \quad AE^T EA^T = (AE^T)(AE^T)^T = 0 \quad \Rightarrow \quad AE^T = 0$$

    y, por la propiedad (a) tenemos que Ay y Ey son independientes, por lo tanto, Ay y $(Ey)^T Ey = y^T Cy$ son independientes. (La otra implicación también es cierta, pero no la veremos).
  - Similarmente, para ver (c) notamos que podemos escribir $C = E^T E$ y $D = F^T F$. Ahora

    $$CD = 0 \quad \Rightarrow \quad E^T ED^T D = 0 \quad \Rightarrow \quad \text{(después de algunos pasos)} \quad ED^T = 0$$

    por lo tanto, Ey y Dy son independientes y, entonces $y^T Cy = (Ey)^T (Ey)$ es independiente de $y^T Dy = (Fy)^T (Fy)$. (La otra implicación también es cierta, pero no la veremos).

- Consideremos el modelo usual de regresión lineal: $y \sim N(X\beta, \sigma^2 I)$. Sabemos que $\widehat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$, así que $a^T \widehat{\beta} \sim N(a^T \beta, \sigma^2 a^T (X^T X)^{-1} a)$, esta propiedad distribucional es el primer paso para tratar de probar hipótesis acerca de $a^T \beta$. Note que

  $$A \equiv \frac{a^T \widehat{\beta} - a^T \beta}{\sqrt{\sigma^2 a^T (X^T X)^{-1} a}} \sim N(0, 1)$$

- Por otro lado, $\text{SCE} = (y - \widehat{y})^T (y - \widehat{y}) = (y - X\widehat{\beta})^T (y - X\widehat{\beta}) = (y - Py)^T (y - Py) = y^T (I - P) y$. Consideremos la distribución de

  $$\frac{\text{SCE}}{\sigma^2} = y^T \frac{(I - P)}{\sigma^2} y$$

  En base a resultados sobre distribuciones de formas cuadráticas, es fácil ver que

  $$B \equiv \frac{\text{SCE}}{\sigma^2} \sim \chi^2_{n-p}$$

  los grados de libertad, $n - p$, son debido a que $\text{rango}(I - P) = n - p$.

- Independencia entre A y B. Note que en A lo único que es aleatorio es $a^T \widehat{\beta} = a^T (X^T X)^{-1} X^T y \equiv HX^T y$, esto es, A es función de $HX^T y$ y B es función de $y^T (I - P) y$. Note que $HX^T (I - P) = H[(I - P)X]^T = H(X - PX)^T = H(X - X)^T = 0$, por lo tanto la forma lineal $HX^T y$ es independiente de la forma cuadrática $y^T (I - P) y$ y por lo tanto A y B son independientes.

- Recordando la estructura de una variable aleatoria t_ν:

$$t_\nu = \frac{N(0,1)}{\sqrt{\frac{\chi^2_\nu}{\nu}}}$$

donde la $N(0,1)$ y la χ^2_ν son independientes.

- Aprovechando los resultados distribucionales que acabamos de obtener, tenemos

$$t = \frac{\frac{a^T\widehat{\beta} - a^T\beta}{\sqrt{\sigma^2 a^T(X^TX)^{-1}a}}}{\sqrt{\frac{\mathsf{SCE}}{\sigma^2(n-p)}}} = \frac{a^T\widehat{\beta} - a^T\beta}{\sqrt{\frac{\mathsf{SCE}}{n-p}a^T(X^TX)^{-1}a}} = \frac{a^T\widehat{\beta} - a^T\beta}{\sqrt{\mathsf{CME}\ a^T(X^TX)^{-1}a}} \sim t_{n-p}$$

donde CME es el Cuadrado Medio del Error $= \mathsf{SCE}/(n-p)$.

- El Cuadrado Medio del Error es un estimador insesgado de σ^2 pues note que (aprovechando el resultado del problema 7 de los problemas extra: Si $y \sim N(\mu, \Sigma)$ entonces $\mathsf{E}(y^TAy) = \mathsf{tr}(A\Sigma) + \mu^TA\mu$):

$$\mathsf{E}(\mathsf{CME}) = \frac{1}{n-p}\mathsf{E}[y^T(I-P)y] = \frac{1}{n-p}\left[\beta^TX^T(I-P)X\beta + \sigma^2\mathsf{tr}(I-P)\right] = \frac{0 + (n-p)\sigma^2}{n-p} = \sigma^2$$

- El resultado obtenido más arriba nos permite, por ejemplo, probar hipótesis de la forma $H_0 : a^T\beta = m$; en este caso, el estadístico de prueba es

$$t = \frac{a^T\widehat{\beta} - m}{\sqrt{\mathsf{CME}\ a^T(X^TX)^{-1}a}} \overset{H_0}{\sim} t_{n-p}$$

El criterio de decisión es rechazar H_0 si $|t| > t_{n-p,\alpha/2}$.

- Un caso particular son las pruebas sobre coeficientes individuales, $H_0 : \beta_j = 0$, en este caso, el estadístico de prueba se reduce a:

$$t = \frac{\widehat{\beta}_j}{\sqrt{\mathsf{CME}\ (X^TX)^{-1}_{jj}}} \overset{H_0}{\sim} t_{n-p}$$

con la misma regla de decisión. Esta prueba es reportada en todos los softwares que hacen regresiones.

- También podemos construír intervalos de confianza aprovechando los resultados anteriores:

$$\widehat{\beta}_j \pm t_{n-p,\alpha/2}\sqrt{\mathsf{CME}\ (X^TX)^{-1}_{jj}}$$

- Regiones de Confianza.
  - Considere la forma cuadrática

$$q_1 = \frac{1}{\sigma^2}(\widehat{\beta} - \beta)^T(X^TX)(\widehat{\beta} - \beta)$$

notando que $\widehat{\beta} - \beta \sim N_p(0, \sigma^2(X^TX)^{-1})$ es fácil ver que $q_1 \sim \chi^2_p$.

- Ahora consideremos $q_2 = \text{SCE}/\sigma^2$, la cual, vimos, tiene la distribución $q_2 \sim \chi^2_{n-p}$
- Note que q_1 puede escribirse como (con $\mu = X\beta$)

$$\begin{aligned}\sigma^2 q_1 &= (\widehat{\beta}-\beta)^T(X^TX)(\widehat{\beta}-\beta) = [X(\widehat{\beta}-\beta)]^T[X(\widehat{\beta}-\beta)] \\ &= (\widehat{y}-\mu)^T(\widehat{y}-\mu) = (Py-\mu)^T(Py-\mu) = (Py-P\mu+P\mu-\mu)^T(Py-P\mu+P\mu-\mu) \\ &= [P(y-\mu)-(I-P)\mu]^T[P(y-\mu)-(I-P)\mu] = (y-\mu)^TP(y-\mu)+\mu^T(I-P)\mu\end{aligned}$$

  donde el producto cruzado se cancela. Esto es, q_1 es una función de la forma cuadrática $(y-\mu)^TP(y-\mu)$ (el término $\mu^T(I-P)\mu$ es una constante).
- Por otro lado $\sigma^2 q_2 = y^T(I-P)y = (y-\mu)^T(I-P)(y-\mu)$. Es directo ver que $(y-\mu)^TP(y-\mu)$ es independiente de $(y-\mu)^T(I-P)(y-\mu)$. En otras palabras, q_1 y q_2 son independientes.
- Por lo tanto

$$\frac{q_1/p}{q_2/(n-p)} = \frac{(\widehat{\beta}-\beta)^T(X^TX)(\widehat{\beta}-\beta)}{p\ \text{CME}} \sim F^p_{n-p}$$

  de aquí que

$$\text{P}\left(\frac{1}{p\ \text{CME}}(\widehat{\beta}-\beta)^T(X^TX)(\widehat{\beta}-\beta)\ \leq\ F^p_{n-p,\alpha}\right) = 1-\alpha$$

  equivalentemente

$$\text{P}\left((\beta-\widehat{\beta})^T(X^TX)(\beta-\widehat{\beta})\ \leq\ p\ \text{CME}\, F^p_{n-p,\alpha}\right) = 1-\alpha$$

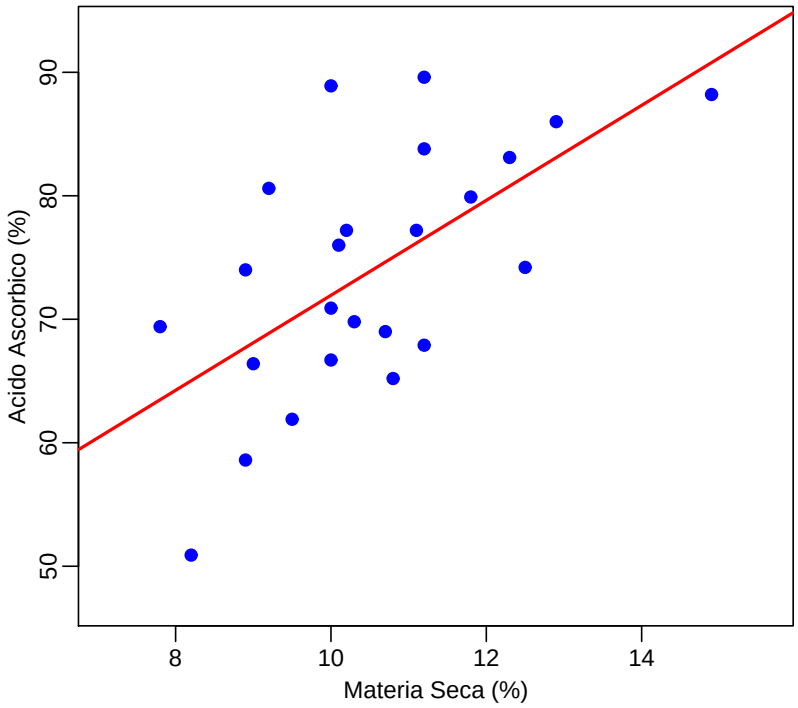
- La región de confianza del $(1-\alpha)\times 100\%$ es

$$\left\{\beta\ :\ (\beta-\widehat{\beta})^T(X^TX)(\beta-\widehat{\beta})\ \leq\ p\ \text{CME}\, F^p_{n-p,\alpha}\right\}$$

- La expresión $(\beta-\widehat{\beta})^T(X^TX)(\beta-\widehat{\beta})$, vista como función de β, define un elipsoide centrado en $\widehat{\beta}$, así que la región de confianza consiste del interior del elipsoide correspondiente al nivel $p\ \text{CME}\, F^p_{n-p,\alpha}$.

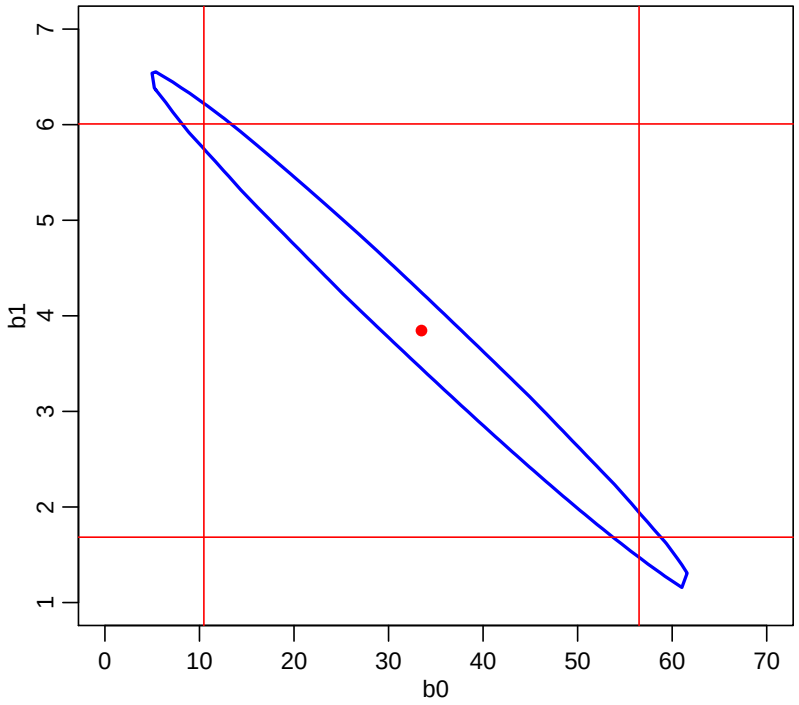
- El siguiente programa en R ilustra los cálculos asociados a un ajuste de un modelo de regresión lineal simple, así como la construcción de una región de confianza para los coeficientes del modelo.

```
# Datos del Jorgensen p. 3
# Porcentaje de Materia Seca, x, en espinacas frescas y
# porcentaje de acido ascorbico conservado, y, despues del secado.
datos <- matrix( c(
100, 89, 89, 92, 78, 101, 90, 82, 95, 108, 111, 112,
125, 123, 100, 102, 112, 112, 100, 107, 103, 129, 118, 149,
709, 740, 586, 806, 694, 760, 664, 509, 619, 652, 772, 896,
742, 831, 667, 772, 838, 679, 889, 690, 698, 860, 799, 882),
ncol=2, byrow=F)/10
x <- datos[,1]
y <- datos[,2]
xr <- range(x); d <- xr[2]-xr[1]; xr <- xr+.10*d*c(-1,1)
yr <- range(y); d <- yr[2]-yr[1]; yr <- yr+.10*d*c(-1,1)
plot(x, y, mgp=c(1.5,.5,0), col="blue", cex.axis=.9, cex.lab=.9,
 type="p", xlim=xr, ylim=yr, xlab="Materia Seca (%)",
 ylab="Acido Ascorbico (%)", pch=19 )
out <- lm(y~x)
abline(out,col="red",lwd=2)
```




```
# Un resumen basico sobre el ajuste
summary(out)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   33.482     11.098   3.017  0.00634 **
x              3.846      1.042   3.689  0.00128 **
Residual standard error: 8.052 on 22 degrees of freedom
Multiple R-squared: 0.3822,     Adjusted R-squared: 0.3541
F-statistic: 13.61 on 1 and 22 DF,  p-value: 0.001283
n   <- length(y)                      # 24
p   <- 2
gl  <- n-p                            # 22
X   <- cbind(rep(1,n),x)
XXi <- solve(t(X)%*%X)
be  <- solve(t(X)%*%X,t(X)%*%y)       # 33.481893   3.845804
sce <- sum( (y-X%*%be)^2 )            # 1426.500
cme <- sce/gl                         # 64.84089
sig <- sqrt(cme)                      #  8.05238
alf <- 0.05
# Intervalos de confianza para la ordenada y pendiente
tal <- qt(1-alf/2,gl)
li0 <- be[1]-tal*sqrt(cme*XXi[1,1])   #  10.46535
ls0 <- be[1]+tal*sqrt(cme*XXi[1,1])   #  56.49844
li1 <- be[2]-tal*sqrt(cme*XXi[2,2])   #   1.683931
ls1 <- be[2]+tal*sqrt(cme*XXi[2,2])   #   6.007677
# estos pueden obtenerse usando:  confint(out)
```

```
# Region de confianza
m   <- 40
rco <- function(beta){return(sum((X%*%(beta-be))^2))}
b0  <- seq(0,70,length=m)
b1  <- seq(1,7,length=m)
z   <- matrix(0,m,m)
for(i in 1:m){
 for(j in 1:m){
  z[i,j] <- rco(c(b0[i],b1[j]))}}
val <- p*cme*qf(1-alf,p,n-p)
contour(b0,b1,z, levels=val, mgp=c(1.5,.6,0), xlab="b0", ylab="b1",
 cex.lab=.9, cex.axis=.9, lwd=2, col="blue", drawlabels=F )
points(be[1],be[2],pch=16,col="red")
abline(v=c(li0,ls0),h=c(li1,ls1),col="red")
```



Resumen de Clase 7: Lunes 21 de febrero

- En la clase 3 consideramos la hipótesis $H_0 : \beta \in \Omega_0$, y vimos que la prueba de cociente de verosimilitudes con $\Lambda = \max_{\Omega_0} L \,/\, \max L$, donde se rechaza H_0 si $\Lambda < c$, es equivalente a rechazar H_0 si $(\mathsf{SCE}_0 - \mathsf{SCE}) \,/\, \mathsf{SCE} > d$. Veremos más de cerca esta prueba.

- Específicamente, Ω_0 lo tomaremos como el espacio formado por todas las soluciones del sistema $K^T\beta = m$, donde K^T es $q \times p$ de rango q. Esto es, consideraremos hipótesis de la forma $H_0 : K^T\beta = m$.

- Para calcular SCE_0 necesitamos el estimador de máxima verosimilitud de β cuando restringimos a que $K^T\beta = m$. Necesitamos

$$\max l(\beta, \sigma^2; y) = -\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2), \quad \text{sujeta a } K^T\beta = m$$

- El Lagrangiano es

$$l(\beta, \sigma^2; y) - \frac{1}{\sigma^2}\lambda^T(K^T\beta - m)$$

derivando con respecto a β, σ^2 y λ tenemos que $\widehat{\sigma}^2_R = \mathsf{SCE}(\widehat{\beta}_R)/n$ y que $\widehat{\beta}_R$ y $\widehat{\lambda}$ satisfacen el sistema

$$\begin{array}{rcl} X^TX\beta + K\lambda & = & X^Ty \\ K^T\beta & = & m \end{array}$$

- En clase vimos que, aplicando operaciones elementales por bloques, que la solución es

$$\widehat{\beta}_R = \widehat{\beta} - (X^TX)^{-1}K\left[K^T(X^TX)^{-1}K\right]^{-1}(K^T\widehat{\beta} - m)$$

- Por otro lado, necesitamos calcular $\mathsf{SCE}_0 = (y - X\widehat{\beta}_R)^T(y - X\widehat{\beta}_R)$.

- Después de algo de álgebra, se vé que

$$\mathsf{SCE}_0 = \mathsf{SCE} + (K^T\widehat{\beta} - m)^T\left[K^T(X^TX)^{-1}K^T\right]^{-1}(K^T\widehat{\beta} - m)$$

- De modo que, aplicando propiedades de distribuciones de formas cuadráticas, tenemos que

$$\frac{1}{\sigma^2}(\mathsf{SCE}_0 - \mathsf{SCE}) = \frac{1}{\sigma^2}(K^T\widehat{\beta} - m)^T\left[K^T(X^TX)^{-1}K^T\right]^{-1}(K^T\widehat{\beta} - m) \sim \chi^2_{k,\lambda}$$

donde $k = \text{rango}(K^T(X^TX)^{-1}K^T) = q$ y $\lambda = \frac{1}{2\sigma^2}(K^T\beta - m)^T\left[K^T(X^TX)^{-1}K^T\right]^{-1}(K^T\beta - m)$.

- Por otro lado, notamos que $\mathsf{SCE}_0 - \mathsf{SCE}$ es función de

$$K^T\widehat{\beta} = K^T(X^TX)^{-1}X^Ty = K^T(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^Ty = K^T(X^TX)^{-1}X^TPy \equiv APy$$

también vemos que $\mathsf{SCE} = y^T(I - P)y$, así que APy es independiente de $y^T(I - P)y$ y, por lo tanto, $\mathsf{SCE}_0 - \mathsf{SCE}$ es independiente de SCE

- Entonces
$$F = \frac{(K^T\widehat{\beta} - m)^T \left[K^T(X^TX)^{-1}K^T\right]^{-1}(K^T\widehat{\beta} - m)/q}{y^T(I-P)y/(n-p)} \overset{H_0}{\sim} F^q_{n-p}$$

- En general, $K^T\beta$ puede ser diferente de m y por lo tanto el numerador es una χ^2 nocentral. Esto dá lugar a que
$$F = \frac{(K^T\widehat{\beta} - m)^T \left[K^T(X^TX)^{-1}K^T\right]^{-1}(K^T\widehat{\beta} - m)/q}{y^T(I-P)y/(n-p)} \sim F^q_{n-p,\lambda}$$
esto es, F tiene una distribución F nocentral (consultar esto en la literatura).

- Algunos casos particulares
  - $H_0 : \beta_j = 0$; para esta hipótesis tenemos $K^T = (0, \cdots, 1, \cdots, 0)$ y $m = 0$, así:
$$F = \frac{\widehat{\beta}_j^2}{\mathsf{CME} \times (X^TX)^{-1}_{jj}} \sim F^1_{n-p,\lambda}$$
  - $H_0 : \beta = 0$; aquí $K^T = I_p$ y el estadístico es
$$F = \frac{\widehat{\beta}^T(X^TX)\widehat{\beta}}{\mathsf{CME}} \sim F^p_{n-p,\lambda}$$
en general, esta prueba no es de interés, ¿Porqué?
  - $H_0 : \beta_r = \beta_{r+1} = \cdots = \beta_p = 0$; esto es $K^T = [0, I]$ y $m = 0$, aquí el estadístico F toma una cierta forma al sustituír estos valores. En la práctica, sin embargo, es preferible usar
$$F = \frac{(\mathsf{SCE}_0 - \mathsf{SCE})/(p - r + 1)}{\mathsf{CME}} \sim F^{p-r+1}_{n-p,\lambda}$$
esto es fácil de calcular pues casi cualquier software que haga regresiones te dará SCE_0 y SCE.

Tarea 2. Modelos Estadísticos I

1. Considere el modelo

$$y = X_1\beta_1 + X_2\beta_2 + e = [X_1, X_2]\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} + e \equiv X\beta + e, \quad \text{donde } e \sim N(0, \sigma^2 I_n)$$

   (a) Sean P_1 y P las matrices de proyección sobre las columnas de X_1 y X respectivamente. Muestre que $P - P_1$ es una matriz de proyección (para ver esto no se necesita lo siguiente, pero es bueno saber que: $P - P_1$ proyecta sobre el subespacio de $\mathcal{C}(X)$ que es ortogonal a $\mathcal{C}(X_1)$).

   (b) Note que, trivialmente

$$y^T\left(I - \frac{1}{n}J\right)y = y^T\left(P - \frac{1}{n}J\right)y + y^T(I - P)\,y$$

   esta es una igualdad independientemente de quienes son P y J, en el caso cuando P es la proyección sobre $\mathcal{C}(X)$ y J es una matriz $n \times n$ formada por unos (estamos suponiendo que el modelo tiene una ordenada al origen), a esta expresión se le llama descomposición para el Análisis de la Varianza y usamos la notación SCT = SCR + SCE (i.e. la suma de cuadrados total es igual a la suma de cuadrados de regresión más la suma de cuadrados del error).
   El coeficiente de determinación de un modelo de regresión se define como

$$R^2 = 1 - \frac{\text{SCE}}{\text{SCT}}$$

   y a $100 \times R^2$ se le interpreta como el porcentaje de la variabilidad total que es explicada por el modelo. Muestre que la R^2 del modelo $y = X\beta + e$ es mayor o igual que la R^2 del modelo $y = X_1\beta_1 + e$ (i.e. siempre podemos incrementar el valor de R^2 de un modelo simplemente agregando variables, sean o no importantes).

2. Estimación Bayesiana. Considere el modelo $y = X\beta + e$, con $e|\sigma^2 \sim N(0, \sigma^2 I_n)$. Definamos un estimador Bayesiano de β como aquel estadístico $t(y)$ que minimice el Riesgo de Bayes

$$R = \mathsf{E}_{\beta,\sigma^2,y}\left[\{t(y) - \beta\}^T\{t(y) - \beta\}\right]$$

   donde la esperanza se toma sobre la distribución conjunta de β, σ^2 y y.

   (a) Muestre que

$$R = \mathsf{E}_y\mathsf{E}_{\beta,\sigma^2|y}\left[\left\{t(y) - \tilde{\beta}\right\}^T\left\{t(y) - \tilde{\beta}\right\}\right] + \mathsf{E}_y\mathsf{E}_{\beta,\sigma^2|y}\left[\left\{\beta - \tilde{\beta}\right\}^T\left\{\beta - \tilde{\beta}\right\}\right]$$

   donde $\tilde{\beta} = \mathsf{E}_{\beta,\sigma^2|y}(\beta)$, i.e. la media posterior de β.

   (b) Del inciso anterior, deduzca que el estimador de Bayes para β es $\tilde{\beta}$.

   (c) Suponga que la distribución apriori para β y σ^2 es de la forma $N(0, \frac{\sigma^2}{c}I_p) \times g(\sigma^2)$, donde $c > 0$. Muestre que

$$\tilde{\beta} = \mathsf{E}_{\beta,\sigma^2|y}(\beta) = \mathsf{E}_{\sigma^2|y}\mathsf{E}_{\beta|\sigma^2,y}(\beta) = \frac{1}{c}X^T\left(I + \frac{1}{c}XX^T\right)^{-1}y$$

   (d) Muestre que

$$\frac{1}{c}X^T\left(I + \frac{1}{c}XX^T\right)^{-1} = \left(cI + X^TX\right)^{-1}X^T$$

   y que, por lo tanto, el estimador Bayesiano de β es $\tilde{\beta} = (X^TX + cI)^{-1}X^Ty$ (i.e. el llamado Estimador Ridge).

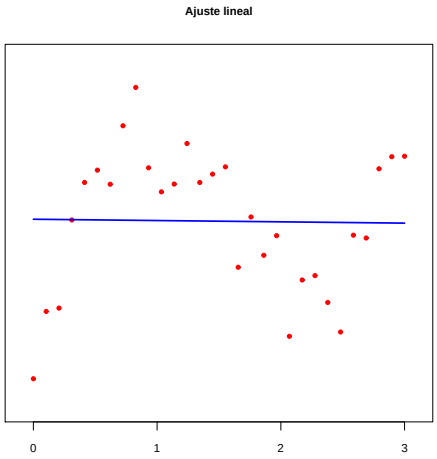
3. El siguiente código genera un conjunto de datos y produce una gráfica de los mismos

```
set.seed(79015)
n    <- 30
x    <- seq(0,3,length=n)
y    <- 3*x-3*x^2+x^3-.1*x^4 + rnorm(n,0,.15)
yr   <- range(y)
yl   <- yr + .10*c(-1,1)*(yr[2]-yr[1])
plot(x, y, xlab="", ylab="", mgp=c(1.5,.5,0), col="red", yaxt="n", xaxt="n",
 cex.axis=.7, main="Datos", cex.main=.8, pch=20, ylim=yl, xlim=c(-.1,3.1))
axis(1,at=0:3, cex.axis=.8)
```

Tenemos entonces, n parejas de puntos $(x_1, y_1), \cdots, (x_n, y_n)$. Suponga que deseamos ajustar un modelo lineal de la forma

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

con $e_i \sim$ i.i.d. $N(0, \sigma^2)$, $i = 1, \cdots, n$. Escribamos este modelo en forma matricial $y = X\beta + e$, donde $e \sim N_n(0, \sigma^2 I)$. Con los datos generados para este problema, escriba un programa para calcular $\widehat{\beta}$ y $\widehat{\text{Var}}(\widehat{\beta})$, y grafique la recta estimada de regresión, $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$, como se muestra en la siguiente gráfica.

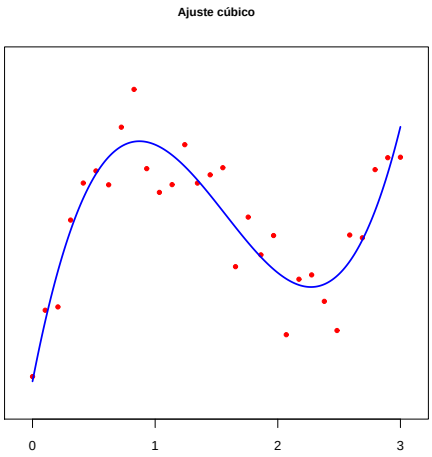


4. En el problema anterior vemos que el ajuste no es satisfactorio.

Ajuste un modelo cúbico

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i$$

con $e_i \sim$ i.i.d. $N(0, \sigma^2)$, $i = 1, \cdots, n$. Calcule $\widehat{\beta}$ y sus errores estándar y grafique la curva estimada de regresión, como se muestra en la siguiente gráfica. ¿Hay alguna diferencia importante si ajustara un modelo de orden 4?.



5. Los modelos polinomiales son importantes pero, al mismo tiempo, son algo restrictivos pues es difícil imaginar que el comportamiento de un fenómeno real sea modelable de esa forma en todo el rango de

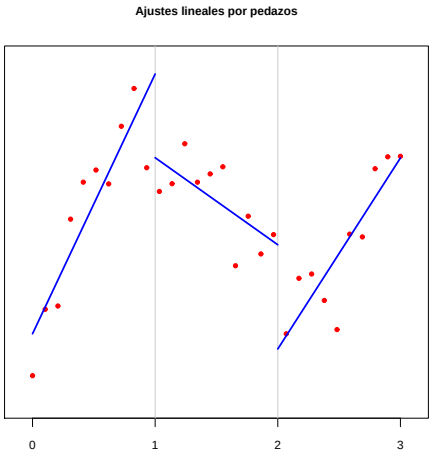
su dominio. Hay ocasiones en las que será razonable pensar que en cierto rango de valores de la variable independiente el comportamiento sigue una relación polinomial y en otro rango es modelable por otro polinomio. En este problema deseamos explorar como hacer el ajuste de un modelo lineal por pedazos.

Suponga que queremos ajustar, a los datos del problema 3, el modelo

$$E(y_i|x_i) = \begin{cases} a + bx_i & \text{si} \quad 0 \le x_i \le 1 \\ c + dx_i & \text{si} \quad 1 < x_i \le 2 \\ e + fx_i & \text{si} \quad 2 < x_i \le 3 \end{cases}$$

Este modelo también puede ponerse en la forma $y = X\beta + e$, pero con

$$X = \begin{bmatrix} 1 & x_1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{10} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{20} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x_{21} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & x_{30} \end{bmatrix} \quad \text{y} \quad \beta = \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix}$$



Calcule $\widehat{\beta}$ y sus errores estándar y grafique las rectas estimadas de regresión, como se muestra en la gráfica de la derecha.

6. Desde el punto de vista de modelación, el modelo anterior no es aceptable pues si nos acercamos a $x = 1$ por la izquierda el modelo nos predice un cierto valor para y, pero si nos acercamos a $x = 1$ por la derecha, el modelo nos predice un valor diferente para y, mucho más bajo. Lo que le falta a este modelo es "continuidad". Para hacer que las rectas se "peguen" en los puntos 1 y 2 (a estos puntos se les llama **nodos**), necesitamos que se cumpla que

$$\begin{aligned} a + k_1\, b &= c + k_1\, d \\ c + k_2\, d &= e + k_2\, f \end{aligned}$$

donde $k_1 = 1$ y $k_2 = 2$ son los nodos. En otras palabras, necesitamos que los parámetros del modelo satisfagan el sistema de ecuaciones lineales (2 ecuaciones y 6 incógnitas):

$$\begin{bmatrix} 1 & k_1 & -1 & -k_1 & 0 & 0 \\ 0 & 0 & 1 & k_2 & -1 & -k_2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \equiv \qquad K^T\beta = 0$$

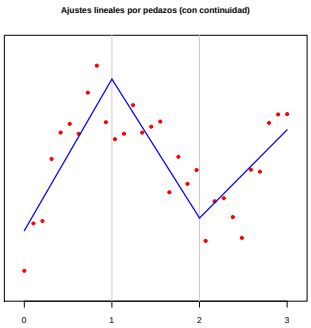
Entonces, para estimar los parámetros del modelo tenemos que resolver el siguiente problema de minimización sujeta a restricciones:

$$\min_{\beta} \ (y - X\beta)^T(y - X\beta) \qquad \text{sujeta a} \quad K^T\beta = 0$$

Este problema puede resolverse en forma explícita usando Multiplicadores de Lagrange y la solución es

$$\tilde{\beta} \ = \ \widehat{\beta} - (X^TX)^{-1}K[\ K^T(X^TX)^{-1}K\ ]^{-1}K^T\widehat{\beta}$$

donde $\widehat{\beta} = (X^T X)^{-1} X^T y$, es el estimador de β sin restricciones. Encuentre el estimador restringido $\tilde{\beta}$ y grafique el modelo resultante, como se muestra en la gráfica:



7. Una vez que ya sabemos como ajustar modelos lineales por pedazos, podemos ahora considerar modelos cúbicos por pedazos y que satisfagan condiciones de continuidad. Suponga que queremos ajustar, a los datos del problema 3, el modelo

$$E(y_i|x_i) = \begin{cases} a + bx_i + cx_i^2 + dx_i^3 & \text{si} \quad 0 \leq x_i \leq 1 \\ e + fx_i + gx_i^2 + hx_i^3 & \text{si} \quad 1 < x_i \leq 2 \\ i + jx_i + kx_i^2 + lx_i^3 & \text{si} \quad 2 < x_i \leq 3 \end{cases}$$

Por lo que aprendimos en el problema 7, esto lo podemos poner en un modelo lineal $y = X\beta + e$ con

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{10} & x_{10}^2 & x_{10}^3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x_{11} & x_{11}^2 & x_{11}^3 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & x_{20} & x_{20}^2 & x_{20}^3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & x_{21} & x_{21}^2 & x_{21}^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & x_{30} & x_{30}^2 & x_{30}^3 \end{bmatrix} \quad \text{y} \quad \beta_{12\times 1} = \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \\ i \\ j \\ k \\ l \end{bmatrix}$$

y como queremos condiciones de continuidad, entonces las ecuaciones son

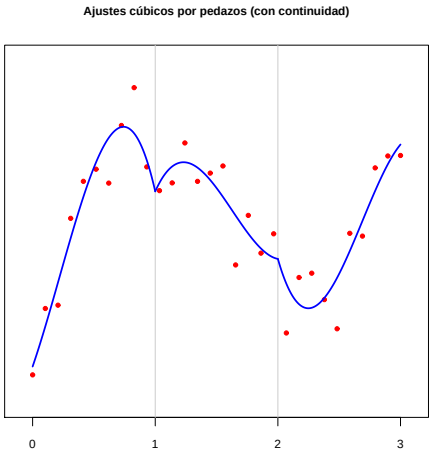
$$\begin{aligned} a + bk_1 + ck_1^2 + dk_1^3 &= e + fk_1 + gk_1^2 + hk_1^3 \\ e + fk_2 + gk_2^2 + hk_2^3 &= i + jk_2 + kk_2^2 + lk_2^3 \end{aligned}$$

En forma matricial, estas condiciones son

$$\begin{bmatrix} 1 & k_1 & k_1^2 & k_1^3 & -1 & -k_1 & -k_1^2 & -k_1^3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & k_2 & k_2^2 & k_2^3 & -1 & -k_2 & -k_2^2 & -k_2^3 \end{bmatrix} \beta_{12\times 1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \equiv \quad K^T\beta = 0$$

donde k_1 y k_2 son los nodos.

Nuevamente, como en el problema 6, lo que queremos es estimar β mediante aquel valor que minimice la suma de cuadrados residual, $SCE(\beta) = (y - X\beta)^T(y - X\beta)$, sujeta a las restricciones $K^T\beta = 0$. Ajuste el modelo cúbico por pedazos con restricciones de continuidad y muéstrelo gráficamente como enseguida:



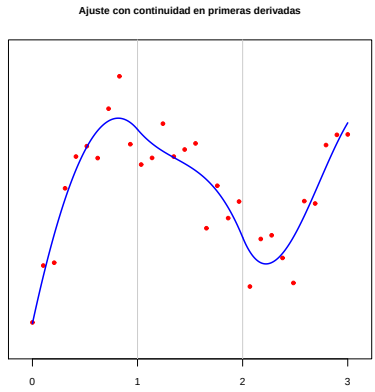
8. Inmediatamente observamos que algo no esta bien en la solución obtenida en el problema 7. Si bien, tenemos continuidad, a la gráfica le falta "suavidad" en los nodos; así que podemos pedir que las derivadas sean continuas ahí. La derivada de la primer cúbica es $b+2cx+3dx^2$, queremos que sea igual a la derivada de la segunda cúbica, $f+2gx+3hx^2$ cuando las evaluemos en $x=k_1$. Similarmente, en el segundo nodo, también queremos que las derivadas sean iguales. Estas observaciones se traducen en las ecuaciones

$$\begin{aligned} b+2ck_1+3dk_1^2 &= f+2gk_1+3hk_1^2 \\ f+2gk_2+3hk_2^2 &= j+2kk_2+3lk_2^2 \end{aligned}$$

De modo que el conjunto total de restricciones es ahora

$$\begin{bmatrix} 1 & k_1 & k_1^2 & k_1^3 & -1 & -k_1 & -k_1^2 & -k_1^3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & k_2 & k_2^2 & k_2^3 & -1 & -k_2 & -k_2^2 & -k_2^3 \\ 0 & 1 & 2k_1 & 3k_1^2 & 0 & -1 & -2k_1 & -3k_1^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2k_2 & 3k_2^2 & 0 & -1 & -2k_2 & -3k_2^2 \end{bmatrix} \beta = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \equiv K^T\beta = 0$$

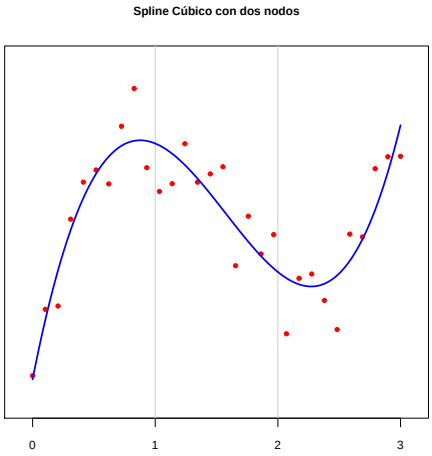
donde k_1 y k_2 son los nodos. Efectuar el ajuste y mostrarlo gráficamnete.



9. Una petición más: Queremos agregar continuidad en la segunda derivada. En otras palabras:

$$\begin{aligned} 2c + 6dk_1 &= 2g + 6hk_1 \\ 2g + 6hk_2 &= 2k + 6lk_2 \end{aligned}$$

Agregue estas dos ecuaciones a las anteriores, efectue el ajuste y muestrelo gráficamente. Al resultado de este ajuste se le llama **Spline Cúbico**.



10. Cuando ajustamos un modelo de regresión de la forma $y = X\beta + e$, lo que hacemos es obtener $\widehat{y}$ el cual es la proyección de y sobre el espacio de columnas generado por la matriz X. Ahora, cuando uno impone condiciones, en realidad lo que estamos haciendo es definiendo un subespacio del espacio de columnas de X. Para clarificar estas ideas, supongamos un modelo de regresión lineal múltiple con dos variables predictoras

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \qquad i = 1, \cdots, n$$

La matriz X en este caso es

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix}_{n \times 3}$$

Ahora, si quisieramos ajustar el modelo bajo la condición $\beta_1 = \beta_2$, esto sería equivalente a ajustar el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_1 x_{2i} + e_i \;=\; \beta_0 + \beta_1 (x_{1i} + x_{2i}) + e_i, \qquad i = 1, \cdots, n$$

cuya nueva matriz X tiene un espacio de columnas que es un subespacio del original

$$X = \begin{bmatrix} 1 & x_{11} + x_{21} \\ 1 & x_{12} + x_{22} \\ \vdots & \vdots \\ 1 & x_{1n} + x_{2n} \end{bmatrix}_{n \times 2}$$

Note que la dimensión del nuevo subespacio es 2 = 3 - 1 = dimensión original - número de restricciones.

Cuando ajustamos un spline cúbico con dos nodos, tenemos que la dimensión original es 12 y el número de restricciones es 6; así que la dimensión del subespacio del spline es 12 - 6 = 6. Ahora, como $\widehat{y}$ es una proyección sobre ese subespacio resultante, no importa cual base sea, la proyección es independiente de la

base; así que es suficiente con conocer esa base y ya esta. Puede verse que una base para el spline cúbico con dos nodos está dada por las columnas de

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{10} & x_{10}^2 & x_{10}^3 & 0 & 0 \\ 1 & x_{11} & x_{11}^2 & x_{11}^3 & (x_{11}-1)^3 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{20} & x_{20}^2 & x_{20}^3 & (x_{20}-1)^3 & 0 \\ 1 & x_{21} & x_{21}^2 & x_{21}^3 & (x_{21}-1)^3 & (x_{21}-2)^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{30} & x_{30}^2 & x_{30}^3 & (x_{30}-1)^3 & (x_{30}-2)^3 \end{bmatrix}$$

Esto es, a los elementos de la base los podemos denotar por $1,\ x,\ x^2,\ x^3,\ (x-k_1)^3_+,\ (x-k_2)^3_+$, donde la notación $(x-k)^3_+$ significa que vale $(x-k)^3$ si $x > k$ y vale 0 de otra forma.

El ajuste del modelo se hace en la forma usual de regresión $\widehat{y} = X\tilde{\beta} = X(X^TX)^{-1}X^Ty$, donde esta X es la que acabamos de definir, correspondiente al subespacio del spline. Entonces, para cualquier x, el valor estimado de y está dado por

$$\widehat{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\beta}_2 x^2 + \tilde{\beta}_3 x^3 + \tilde{\beta}_4 (x-k_1)^3_+ + \tilde{\beta}_5 (x-k_2)^3_+$$

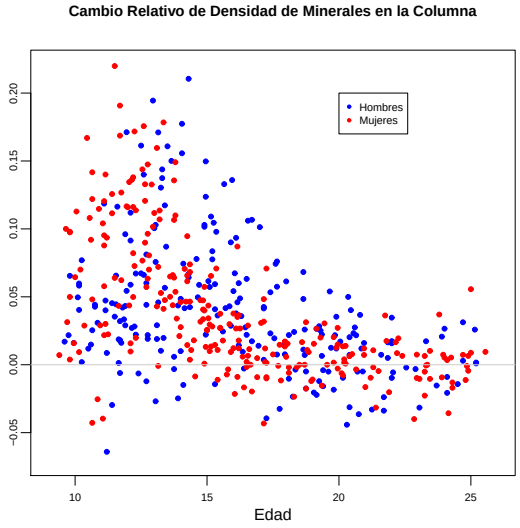
Ajuste el spline cúbico con dos nodos en $k_1 = 1$ y $k_2 = 2$, usando esta nueva matriz X (note que aquí ya no hay necesidad de resolver el problema de mínimos cuadrados con restricciones). Muestre el ajuste gráficamente (la gráfica debe ser idéntica a la obtenida en el problema 9).

Un punto importante que se desprende de este ejercicio es que es muy fácil ajustar splines, por ejemplo, si tuvieramos 4 nodos $k_1 < k_2 < k_3 < k_4$, entonces la base sería

$$1,\ x,\ x^2,\ x^3,\ (x-k_1)^3_+,\ (x-k_2)^3_+,\ (x-k_3)^3_+,\ (x-k_4)^3_+$$

11. El siguiente código produce la gráfica que se muestra enseguida

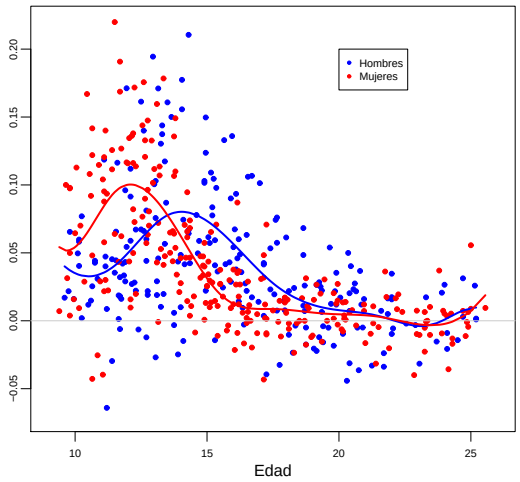
```
datos  <- read.csv("c:\\Documents and Settings\\ ... \\Bone.csv", header=TRUE)
names(datos) <- c("id","edad","genero","spnbmd")
attach(datos)
datH <- datos[genero=="male",]     # 226 x 4
datM <- datos[genero=="female",]   # 259 x 4
detach(datos)

plot(datH[,2],datH[,4], xlab="Edad", ylab="", mgp=c(1.5,.5,0), col="blue",
 cex.axis=.7,ylim=c(-0.07,0.22), xlim=c(9,26),pch=20,
 main="Cambio Relativo de Densidad de Minerales en la Columna", cex.main=.9)
points(datM[,2],datM[,4],col="red",pch=20)
abline(h=0,col=gray(.8))
legend(20,.2,legend=c("Hombres","Mujeres"),pch=20,col=c("blue","red"),cex=.7)
```



Estos datos son los cambios relativos en la densidad de minerales en la columna vertebral de cerca de 500 niños y jóvenes. Estos cambios relativos fueron registrados en visitas consecutivas (aproximadamente separadas un año). Los cambios para los hombres se indican en azul y los de las mujeres en rojo.

Ajuste splines cúbicos para los datos de hombres y mujeres y muestre gráficamente sus resultados. Las gráficas refuerzan el hecho de que el crecimiento de las mujeres antecede al de los hombres (los máximos de las gráficas están separadas por unos dos años).

La siguiente gráfica fue construída usando 7 nodos (`nodos <- seq(11.5,23.5,length=7)`). Los splines cúbicos tienen no muy buen comportamiento antes del primer nodo y después del último; hay técnicas para corregir esto (**splines naturales**). Finalmente, hay splines que le evitan a uno decidir donde poner los nodos pues usan un número maximal de nodos pero tienen que usar técnicas de regularización para su implementación (**splines suavizadores**).



Fecha de entrega: Jueves 3 de marzo.

Problemas Extra 2. Modelos Estadísticos I

1. Supongamos que tenemos el modelo $y = X\beta + \epsilon$ y que deseamos estimar β mediante aquel valor, $\widehat{\beta}$, que minimiza $(y - X\beta)^T(y - X\beta)$ sujeta a la restricción de que $\widehat{\beta}$ satisfaga $A\widehat{\beta} = 0$. Suponga además que X es de tamaño $n \times p$ y es rango completo por columnas y A es de rango completo por renglones. Esta situación aparece en el contexto de pruebas de hipótesis ($H_0 : A\beta = 0$)

   (a) Muestre que $S = \{y \mid y = X\beta$ con β tal que $A\beta = 0\}$ es un subespacio de $\mathcal{R}^n$.

   (b) Muestre que $S = \mathcal{C}(XC)$ donde C es una matriz cuyas columnas forman una base del espacio nulo de A (el espacio nulo de A es $\mathcal{C}^{\perp}(A^T)$ o, en otras palabras, todas las soluciones de $Ax = 0$).

   (c) Usando resultados de proyecciones, muestre que el estimador de mínimos cuadrados restringidos para β está dado por $\widehat{\beta} = C(C^T X^T XC)^{-1} C^T X^T y$ (esta expresión se ve diferente a la vista en clase pero puede verse que son equivalentes).

2. Sean $y_{11}, \cdots, y_{1r}$ distribuídas $N(\mu_1, \sigma^2)$ y $y_{21}, \cdots, y_{2s}$ distribuídas $N(\mu_2, \sigma^2)$, todas las y_{ij}'s independientes. Escriba esto como un modelo lineal. Encuentre estimadores de máxima verosimilitud para μ_1, μ_2, $\mu_1 - \mu_2$ y σ^2. Encuentre un intervalo del 95% de confianza para $\mu_1 - \mu_2$.

3. Suponga que $y_1, \cdots, y_n$ son observaciones independientes que se han tomado de una distribución normal con varianza desconocida σ^2, suponga además que se cuenta con una nueva observación y_{n+1}, sin embargo se sospecha que hubo un cambio en la media al observar dicho dato. Encuentre una prueba F para la hipótesis de que la media de y_{n+1} es la misma que la de las primeras n observaciones.

4. Suponga que $y = \theta + e$, donde $e \sim N_4(0, \sigma^2 I)$, $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)^T$ con $\theta_1 + \theta_2 + \theta_3 + \theta_4 = 0$, muestre que el estadístico F para probar la hipótesis $H_0 : \theta_1 = \theta_3$ es

$$F = \frac{2(y_1 - y_3)^2}{(y_1 + y_2 + y_3 + y_4)^2}$$

5. Considere el modelo de regresión lineal simple $y_i = \beta_0 + \beta_1 x_i + e_i$, $i = 1, \cdots, n$, con los supuestos usuales. Deseamos obtener el estadístico F para probar la hipótesis $H_0 : \beta_1 = c$. Pruebe que

$$F = \frac{(\widehat{\beta}_1 - c)^2}{CME / \sum (x_i - \bar{x})^2}$$

6. Considere el modelo lineal usual y la hipótesis $H_0 : K^T\beta = 0$, donde K^T es $q \times p$ de rango q. Suponga que las últimas q columnas de K^T son independientes, de tal forma que $K^T = [K_1^T, K_2^T]$ con K_2^T no-singular. Exprese a β_2 en términos de β_1, ($\beta^T = (\beta_1^T, \beta_2^T)$). Encuentre una matriz X_H tal que, bajo H_0, E$(y) = X_H\beta_1$.

7. Considere el modelo $y_i = \beta_0 + z_{i1}\beta_1 + z_{i2}\beta_2 + \epsilon_i$, $i = 1, \cdots, n$, donde las z_{i1}'s y z_{i2}'s están centradas y reescaladas (*i.e.* $\sum z_{i1} = \sum z_{i2} = 0$ y $\sum z_{i1}^2 = \sum z_{i2}^2 = 1$) y los errores ϵ_i's $\sim NID(0, \sigma^2)$. Defina ρ como $\rho = \sum z_{i1} z_{i2}$.

(a) Sea $\widehat{\beta}$ el estimador de mínimos cuadrados, verifique que

$$\mathsf{Var}(\widehat{\beta}) = \sigma^2 \begin{bmatrix} \frac{1}{n} & 0 & 0 \\ 0 & \frac{1}{k} & -\frac{\rho}{k} \\ 0 & -\frac{\rho}{k} & \frac{1}{k} \end{bmatrix}$$

donde $k = 1 - \rho^2$.

(b) Suponga que $\widehat{\theta}$ es un estimador de θ, donde este es un determinado parámetro. El error cuadrático medio de $\widehat{\theta}$ se define como $\mathsf{ECM}(\widehat{\theta}) = \mathsf{E}(\widehat{\theta} - \theta)^2$. Muestre que

$$\mathsf{ECM}(\widehat{\theta}) = \mathsf{Var}(\widehat{\theta}) + \Big(\mathsf{E}(\widehat{\theta}) - \theta\Big)^2$$

(i.e. Error cuadrático medio = Varianza + Sesgo cuadrático)

(c) Considere

$$\tilde{\beta}_1 = \frac{\sum z_{i1} y_i}{\sum z_{i1}^2}$$

Muestre que $\mathsf{ECM}(\tilde{\beta}_1) = \sigma^2 + \rho^2 \beta_2^2$.

(d) ¿Bajo que condiciones $\mathsf{ECM}(\tilde{\beta}_1) < \mathsf{ECM}(\widehat{\beta}_1)$?

8. Considere el modelo $y = \gamma_0 1 + W\gamma + \epsilon$, donde $\epsilon \sim N(0, \sigma^2 I)$ y las columnas de W están estandarizadas como en el problema anterior. El estimador Ridge para γ está dado por $\widehat{\gamma}_R = (W^T W + kI)^{-1} W^T y$. Suponga que se tiene un software que sólo calcula regresiones (i.e. calcula $(X^T X)^{-1} X^T y$), deseamos usar ese software para Ridge también. Muestre que el siguiente procedimiento realmente produce estimadores Ridge con un software que hace sólo regresiones:

- Añadir a W los renglones $\sqrt{k} I$, definiendo

$$W_R = \begin{bmatrix} W \\ \sqrt{k} I \end{bmatrix}$$

- Añadir 0's a y, definiendo

$$y_R = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

- Usar el susodicho software para calcular $(W_R^T W_R)^{-1} W_R^T y_R$. Pruebe que esta expresión es precisamente $\widehat{\gamma}_R$.

9. Considere los modelos

$$\begin{aligned} y_{1i} &= \alpha_1 + \beta_1 x_{1i} + \epsilon_{1i} \qquad i = 1, 2, \cdots, n_1 \\ y_{2i} &= \alpha_2 + \beta_2 x_{2i} + \epsilon_{2i} \qquad i = n_1 + 1, n_1 + 2, \cdots, (n_1 + n_2) \end{aligned}$$

donde los errores son independientes, normalmente distribuídos, con media cero y misma varianza. Obtenga el estadístico F basado en el criterio de cociente de verosimilitudes para probar la hipótesis $H_0 : \beta_1 = \beta_2$.

10. Considere el modelo $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \quad i = 1, 2, \cdots, n,$ con los supuestos usuales sobre los errores. Deseamos construír una prueba para la hipótesis $H_0 : \beta_1 = \beta_2 = \beta_3$, (i.e. los coeficientos son iguales, no necesariamente iguales a cero); para ello recurrimos a la prueba del cociente de verosimilitudes para hipótesis de la forma $H_0 : K^T \beta = m$ pues notamos que la hipótesis original es equivalente a probar $H_0^{(1)} : \beta_1 - \beta_2 = 0,\ \beta_1 - \beta_3 = 0,$ con la correspondiente matriz

$$K_1^T = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}$$

Notamos inmediatamente que, en vez de lo anterior, pudimos haber considerado $H_0^{(2)} : \beta_1 - \beta_3 = 0,\ \beta_2 - \beta_3 = 0$ cuya matriz es

$$K_2^T = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

El propósito de este problema es ver que el estadístico F tiene el mismo valor, independientemente de si se usó K_1 o si se usó K_2.

(a) Pruebe que los espacios de columnas de K_1 y K_2 son iguales; esto es, pruebe que

$$\mathcal{C}(K_1) = \mathcal{C}(K_2)$$

(b) Si X es la matriz del modelo (X es $n \times 4$ de rango 4), pruebe que

$$\mathcal{C}(X(X^TX)^{-1}K_1) = \mathcal{C}(X(X^TX)^{-1}K_2)$$

(c) Pruebe que

$$K_1 \left[K_1^T(X^TX)^{-1}K_1\right]^{-1} K_1^T = K_2 \left[K_2^T(X^TX)^{-1}K_2\right]^{-1} K_2^T$$

Sugerencia: Considerar las matrices de proyección sobre los espacios del inciso anterior.

(d) Pruebe que los estadísticos F_1 y F_2 para las hipótesis $H_0^{(1)}$ y $H_0^{(2)}$, son iguales.

11. Suponga que $y \sim N_k(\mu, \sigma^2 I)$, Sean A y B matrices $r \times k$ y $(k-r) \times k$ respectivamente, tales que los renglones de A son ortogonales a los renglones de B. Muestre que u y v son independientes, donde $u = Ay$ y $v = By$.

12. Sea $y = (y_1, y_2, y_3)^T \sim N(\mu, \Sigma)$, donde $\mu = (1, 2, -2)^T$ y

$$\Sigma = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 3 & 1 \\ -1 & 1 & 4 \end{bmatrix}$$

Deseamos generar números aleatorias normales trivariados de esta distribución.

(a) Si $z \sim N_3(0, I)$, pruebe que $y = \Gamma z + \mu$ tiene una distribución $N_3(\mu, \Sigma)$, donde Γ es tal que $\Sigma = \Gamma\Gamma^T$.

(b) Usando el resultado del inciso anterior, escriba un programa en R que genere números aleatorios $N_3(\mu, \Sigma)$

(c) Use la función `rmvnorm` de R para generar números aleatorios $N_3(\mu, \Sigma)$. Esta función no forma parte del conjunto básico de funciones que R trae por default, así que si no está activa entonces bájela de la red; es el paquete `mvtnorm` que se encuentra en el sitio:

`http://cran.us.r-project.org/`

13. Supongamos que $x_i \sim N_1(\mu, \sigma^2),\ i = 1, \cdots, n$ y son independientes entre sí. Definamos

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{y} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Pruebe que $\bar{x}$ y s^2 son independientes.

14. Si $a, b \in \mathcal{R}^n$, la desigualdad de Cauchy-Scwartz nos dice que $|a^Tb| \leq \|a\|\ \|b\|$

(a) Suponga que Σ es una matrix no negativa. Usando esta desigualdad, muestre que

$$|a^T \Sigma b| \leq \sqrt{a^T \Sigma a}\sqrt{b^T \Sigma b}$$

(b) Si $y = (y_1, \cdots, y_n)^T$ es un vector aleatorio, entonces el coeficiente de correlación, ρ_{ij}, entre y_i y y_j está dado por

$$\rho_{ij} = \frac{\mathsf{Cov}(y_i, y_j)}{\sqrt{\mathsf{Var}(y_i)}\sqrt{\mathsf{Var}(y_j)}}$$

Pruebe que $|\rho_{ij}| \leq 1$. (Suponga, sin pérdida de generalidad, que $\mathsf{E}(y) = \mu = 0$).

(c) ¿Porqué dice "sin pérdidad de generalidad"?. Justificar la respuesta.

15. Sea $y = (y_1^T, y_2^T)^T$ un vector aleatorio particionado de tal forma que y_1 y y_2 son de dimensiones r y $n-r$, respectivamente. Suponga que conocemos el valor de y_2; en base a este valor queremos predecir el valor de y_1 (el cual todavía no observamos). Una forma de hacer esto es buscar una función, $\lambda(y_2)$ que minimice el error cuadrático medio

$$\mathsf{ECM}(\ \lambda(y_2)\ ) \equiv \mathsf{E}\left[(\ y_1 - \lambda(y_2)\ )^T (\ y_1 - \lambda(y_2)\ ) \mid y_2\right]$$

(a) Pruebe que $\lambda^*(y_2) = \mathsf{E}(y_1 \mid y_2)$ es el mejor predictor.
En principio, este es un problema muy difícil, pues la minimización del error cuadrático se hace sobre todo el espacio de funciones $\lambda : \mathcal{R}^{n-r} \to \mathcal{R}^r$, sin embargo no es difícil si hacemos

$$\mathsf{ECM}(\ \lambda(y_2)\ ) = \mathsf{E}\left[\ \{\ (\ y_1 - \lambda^*(y_2)\ ) + (\ \lambda^*(y_2) - \lambda(y_2)\ )\ \}^T \{\ (\ y_1 - \lambda^*(y_2)\ ) + (\ \lambda^*(y_2) - \lambda(y_2)\ )\ \} \mid y_2\right]$$

y observamos cuánto vale el lado derecho después de condicionar en y_2

(b) Suponga que $y = (y_1^T, y_2^T)^T$ tiene una distribución $N_n(\mu, \Sigma)$. Encuentre el mejor predictor de y_1 dado que ya observamos y_2.

16. Sea T un estadístico insesgado para θ y con varianza finita $\mathsf{Var}(T) < \infty$ (por simplicidad, consideremos el caso de θ un parámetro escalar).

(a) Suponga que T tiene la mínima varianza posible (en la clase de estimadores insesgados para θ, con varianza finita). Pruebe que si w es cualquier estadístico tal que $\mathsf{E}(w) = 0$ (y $\mathsf{Var}(w) < \infty$), entonces

$$\mathsf{Cov}(T, w) = 0$$

Sugerencia: Considere $T + \lambda w$ donde λ es cualquier número real. $T + \lambda w$ es insesgado para θ y, por lo tanto, $\mathsf{V}(T + \lambda w) \geq \mathsf{V}(T)$ para todo λ, de aquí concluya que $\mathsf{Cov}(T, w) = 0$.

(b) Suponga ahora que $\mathsf{Cov}(T, w) = 0$ para toda w tal que $\mathsf{E}(w) = 0$ (y con $\mathsf{Var}(w) < \infty$). Pruebe que T es el estimador de mínima varianza.
Sugerencia: Sea Z cualquier otro estimador insesgado de θ, entonces $Z - T$ tiene esperanza cero y de aquí deduzca que $\mathsf{Var}(Z) \geq \mathsf{Var}(T)$.

17. Los datos de la siguiente tabla muestran el peso de semillas de soya, observadas por seis semanas consecutivas a partir del inicio del ciclo reproductivo de la planta, así como la cantidad acumulada de radiación solar. Se observaron estas variables en dos regiones diferentes expuestas a diferentes niveles de contaminación por ozono. La variable $Peso$ es el peso medio de las semillas (gramos por planta) de muestras independientes de cuatro plantas.

Ozono bajo		Ozono alto	
Radiación	*Peso*	*Radiación*	*Peso*
118.4	0.7	109.1	1.3
215.2	2.9	199.6	4.8
283.9	5.6	264.2	6.5
387.9	8.7	358.2	9.4
451.5	12.4	413.2	12.9
515.6	17.4	452.5	12.3

(a) Ajuste un modelo de regresión lineal para $Peso$ contra $Radiación$ para cada nivel de Ozono. Determine la similaridad de las dos regresiones comparando los intervalos de confianza para los dos interceptos así como para las dos pendientes. Efectúe también una comparación visual de los dos conjuntos de datos.

(b) Defina una nueva variable $Ozono$ como 1 si el nivel de ozono es alto y 0 si el nivel es bajo. Ajuste un modelo para $Peso$ en función de $Radiación$ y $Ozono$. Pruebe la hipótesis $H_0 : \beta_2 = 0$. Dé su conclusión.

(c) Suponga ahora que "Ozono bajo" es 0.025 ppm y "Ozono alto" es 0.07 ppm. Analize un modelo de regresión lineal para $Peso$ sobre las variables $Radiación$ y $Ozono$ (Utilize variables independientes centradas)(¿Por qué?). Interprete los resultados.

(d) Extienda este modelo a un modelo que incluya otra variable independiente definida como el producto de las variables centradas $Radiación$ y $Ozono$. Estime este modelo e interprete los resultados. ¿Cómo se interpreta esa variable nueva?.

18. Considere el modelo $y = X\beta + e$ con X de tamaño $n \times p$ de rango p y $e \sim N(0, \sigma^2 I)$. Dado que $y^T y = y^T P y + y^T (I - P) y$, a $y^T P y$ se le conoce como la Suma de Cuadrados del Modelo (SCM). Suponga que X puede escribirse como $X = QR$ donde Q es $n \times p$ con p columnas ortonormales y R es una matriz $p \times p$ triangular superior no-singular. Hagamos $Q = [q_1, q_2, \cdots, q_p]$. Pruebe que

$$SCM = SCM_1 + SCM_2 + \cdots + SCM_p$$

(descomposición de una suma de cuadrados del modelo en sumas de cuadrados con 1 grado de libertad) donde SCM_j es la suma de cuadrados del modelo $y = q_j \gamma + e, \quad j = 1, 2, \cdots, p$.

19. Considera el modelo $(*)$ $y = X\beta + e$ con $E(e) = 0$ y $Var(e) = \sigma^2 I$

(a) Supongamos que, en vez de observar X, nosotros registramos $Z = X + E$ (aunque $(*)$ sigue siendo el modelo correcto). Muestre que $\lambda^T \hat{\beta}_Z$, donde $\hat{\beta}_Z = (Z^T Z)^- Z^T y$, es insesgado para $\lambda^T \beta$, para toda $\lambda \in \mathcal{R}(Z)$, si y sólo si, $E = (I - P_Z)U$ para algún U. (i.e. $\mathcal{C}(E) \subset \mathcal{C}(I - P_Z)$. ($P_Z$ es la matriz de proyección sobre $\mathcal{C}(Z)$).

(b) Sea W una matriz $n \times p$, tal que $W^T X$ es no-singular (i.e. aquí suponemos que el rango de X es p). Mostrar que $\hat{\beta}_W = (W^T X)^{-1} W^T y$ es insesgado para β. ($\hat{\beta}_W$ es conocido en la literatura econométrica como "estimador de variables instrumentales").

20. Considere el modelo lineal $y = X\beta + e$ con X de tamaño $n \times p$ de rango p y $e \sim N(0, \sigma^2 I)$. La desigualdad generalizada de Cauchy–Schwartz es

$$(u^T w)^2 \leq (u^T A u)(w^T A^{-1} w)$$

la cual es válida para cualesquier vectores u y w y cualquier matriz A positiva definida.

(a) Usando esta desigualdad, encuentre intervalos de confianza simultáneos para $u^T\beta$, mostrando que

$$P\left[\, u^T\widehat{\beta} - L(u) \leq u^T\beta \leq u^T\widehat{\beta} + L(u), \quad \underline{\text{para todo}} \quad u \,\right] \geq 1-\alpha$$

donde $\widehat{\beta}$ es el estimador de mínimos cuadrados de β y $L(u)$ es una cantidad que, además de depender de u, depende también de cantidades tales como p, CME, $(X^TX)^{-1}$ y $F^p_{n-p,\alpha}$. (A estos intervalos de confianza simultáneos se les conoce como *Intervalos de Scheffé*).

Sugerencia: Tomar $w = \widehat{\beta} - \beta$ en la desigualdad de Cauchy–Schwartz.

(b) Suponga ahora que u es un vector fijo y que solo queremos un intervalo de confianza para esta $u^T\beta$ particular. ¿Cómo lo construiría?. ¿Sería más angosto o más ancho que el obtenido en el inciso anterior?.

21. Suponga que $y = X_1\beta_1 + e$, con $E(e) = 0$ y $Var(e) = \sigma^2 I$ es el modelo verdadero en cierta situación, sin embargo, por alguna razón, ajustamos el modelo sobreparametrizado $y = X_1\beta_1 + X_2\beta_2 + e$. Sea $\widehat{\beta}_{1v}$ el estimador de mínimos cuadrados de β_1 bajo el modelo verdadero. Los correspondientes estimadores de mínimos cuadrados para β_1 y β_2, bajo el modelo sobreparametrizado, están denotados por $\widehat{\beta}_{1s}$ y $\widehat{\beta}_{2s}$.

(a) Muestre que

$$\widehat{\beta}_{1s} = \widehat{\beta}_{1v} - A\widehat{\beta}_{2s}$$

y

$$\widehat{\beta}_{2s} = [X_2^T(I-P_1)X_2]^{-1}X_2^T(I-P_1)y$$

donde $A = (X_1^TX_1)^{-1}X_1^TX_2$ y P_1 es la matriz de proyección sobre $\mathcal{C}(X_1)$.

Sugerencia: Se puede usar la expresión de la inversa de una matriz particionada, sin embargo, es más fácil resolver las ecuaciones normales directamente.

(b) Los residuales de la regresión de y sobre X_1 son $r = (I-P_1)y$, estos residuales pueden interpretarse como "todo aquello de y que no puede ser explicado por X_1". Del mismo modo, la cantidad $R = (I-P_1)X_2$ puede interpretarse como una matriz de residuales, esto es, todo lo de X_2 que no puede ser explicado por X_1. Efectúe la regresion de los residuales r sobre la matriz de residuales R, mostrando que el vector estimado de parámetros de esta regresión es precisamente $\widehat{\beta}_{2s}$ del inciso anterior.

Nota: Si la matriz X_2 tuviera una sola columna, esto es, si solo hay una variable extra en el modelo sobreparametrizado, entonces la regresión de r sobre R es una regresión lineal simple que pasa por el origen; a esta gráfica se le llama "gráfica de regresión parcial" y es usada para identificar visualmente cuales puntos son los que más impactan en la magnitud del estimador de β_2.

22. Considere los datos de un estudio para relacionar el peso (en gramos) de materia seca de plantas, con el porcentaje de materia orgánica presente en el suelo (x_1) y la cantidad de nitrógeno suplementario (x_2) en kilogramos por cada 1000 m^2. Se obtuvieron datos de $n = 7$ parcelas y se ajustó un modelo de regresión lineal (con intercepto), dando la siguiente información:

$$(X^TX)^{-1} = \begin{bmatrix} 1.80 & -.07 & -.25 \\ -.07 & .01 & 0 \\ -.25 & 0 & .06 \end{bmatrix}, \quad \widehat{\beta} = \begin{bmatrix} 51.6 \\ 1.5 \\ 6.7 \end{bmatrix}, \quad \text{SCE} = 27.6$$

(a) Dé la ecuación de regresión e interprete cada coeficiente.

(b) Determine intervalos de confianza del 95% para β_1 y β_2.

(c) Suponga que por experiencia previa se cree que un punto porcentual de incremento en el porcentaje de materia orgánica es equivalente (en términos de su efecto en producción de materia seca) a 0.5 kilos de nitrógeno suplementario por cada 1000 m^2. Traduzca este estatuto en una hipótesis nula acerca de los coeficientes de regresión. Use un estadístico t para probar esta hipótesis contra una hipótesis alternativa de que el nitrógeno suplementario es más efectivo de lo que el estatuto implica.

Resumen de Clase 8: Miércoles 23 de febrero

- Comentarios acerca de intervalos de confianza simultáneos:
  - Consideremos m funciones paramétricas de la forma

    $$a_1^T\beta, a_2^T\beta, \cdots, a_m^T\beta$$

    si queremos intervalos de confianza, es natural construirlos como

    $$a_i^T\widehat{\beta} \pm t_{n-p,\alpha/2}\sqrt{\mathsf{CME}\; a_i^T(X^TX)^{-1}a_i}, \quad i = 1, \cdots, m$$

  - Sea E_i el evento "el intervalo i cubre a la correspondiente función paramétrica", entonces $\mathsf{P}(E_i) = 1-\alpha, \;\; i = 1, \cdots, m.$
  - Por otro lado, la probabilidad de que <u>todos</u> los intervalos cubran a sus respectivas funciones paramétricas es

    $$\mathsf{P}\left[\bigcap_{i=1}^m E_i\right] = 1 - \mathsf{P}\left[\bigcup_{i=1}^m E_i^c\right] \geq 1 - \sum_{i=1}^m \mathsf{P}(E_i^c) = 1 - m\alpha$$

    esto es, la probabilidad de que todos los intervalos esten haciendo su trabajo puede ser tan pequeña como $1 - m\alpha$; por ejemplo, con $\alpha = .05$ y $m = 10$ tenemos que la probabilidad puede ser tan baja como 0.5 (lo cual puede ser algo no informativo).
  - Nos gustaría que $\mathsf{P}[\cap E_i] \geq 1-\alpha$. Una forma de lograr esto es usar la llamada "corrección de Bonferroni", ajustando el nivel de confianza de los intervalos individuales: En vez de usar α, usar α/m, esto asegura un nivel de confianza simultáneo de al menos $1-\alpha$.
  - Si la dependencia entre los eventos E_i's es pequeña, entonces

    $$\mathsf{P}\left[\bigcap_{i=1}^m E_i\right] \approx \prod_{i=1}^m \mathsf{P}(E_i) = (1-\alpha)^m$$

    en este caso, también tenemos que esta probabilidad puede ser bastante más pequeña que $1-\alpha$.
  - Si bien es cierto que la corrección de Bonferroni da el nivel de confianza global que queremos, los intervalos resultantes pueden llegar a ser bastante amplios y, por lo tanto, menos informativos.
  - Recomendamos la lectura del artículo "Multiple hypothesis testing in microarray experiments" de Dudoit y Shaffer del 2003 en Statistical Science (Vol.18).

- El problema de colinealidad en Regresión. Es claro que cuando existen dependencia lineales exactas entre las variables en un modelo de regresión el ajuste del modelo no puede hacerse pues la matriz X^TX no es invertible. Ahora bien, cuando las dependencias no son exactas, esta matriz si es invertible pero esta mal condicionada y el modelo ajustado puede presentar problemas graves de interpretación.

- En la discusión que sigue haremos uso de la inversa de una matriz particionada. Puede verse que, efectuando operaciones elementales por bloques, obtenemos

  $$\left[\begin{array}{cc|cc} A & B & I & O \\ C & D & O & I \end{array}\right] \sim \cdots \sim \left[\begin{array}{cc|cc} A & O & I + BH^{-1}CA^{-1} & -BH^{-1} \\ O & H & -CA^{-1} & I \end{array}\right], \quad \text{donde } H = D - CA^{-1}B$$

  esto es

  $$\left[\begin{array}{cc} A & B \\ C & D \end{array}\right]^{-1} = \left[\begin{array}{cc} A^{-1} + A^{-1}BH^{-1}CA^{-1} & -A^{-1}BH^{-1} \\ -H^{-1}CA^{-1} & H^{-1} \end{array}\right], \quad \text{donde } H = D - CA^{-1}B$$

- Factores de inflación de varianza. Consideremos el modelo de regresión usual $y = X\beta + e$, con $e \sim N(0, \sigma^2 I)$. Suponga que estandarizamos las variables (restandoles la media y dividiendo entre su norma), de modo que podemos escribir $y = \alpha 1 + W\gamma + e$. Es fácil ver que los estimadores de α y γ, sus medias y sus varianzas son:

$$\begin{aligned} \widehat{\alpha} &= \bar{y} & \widehat{\gamma} &= (W^T W)^{-1} W^T y \\ \mathsf{E}(\widehat{\alpha}) &= \alpha & \mathsf{E}(\widehat{\gamma}) &= \gamma \\ \mathsf{Var}(\widehat{\alpha}) &= \frac{\sigma^2}{n} & \mathsf{Var}(\widehat{\gamma}) &= \sigma^2 (W^T W)^{-1} \equiv \sigma^2 Q \end{aligned}$$

- Queremos ver el impacto de colinealidades sobre las varianzas de $\widehat{\gamma}$; esto es, sobre los elementos diagonales de Q. Pongamos atención sobre q_{kk}, el último elemento diagonal de Q. Note que

$$W^T W = \begin{bmatrix} W_1^T \\ w^T \end{bmatrix} [W_1, w] = \begin{bmatrix} W_1^T W_1 & W_1^T w \\ w^T W_1^T & w^T w \end{bmatrix} \equiv \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

de las expresiones anteriores para la inversa de una matriz particionada tenemos

$$q_{kk} = H^{-1} = (D - CA^{-1}B)^{-1} = \left(w^T w - w^T W_1 (W_1^T W_1)^{-1} W_1^T w \right)^{-1} = \left(w^T w - w^T P_1 w \right)^{-1}$$

donde P_1 es la matriz de proyección sobre las columnas de W_1. Si hicieramos la regresión de w sobre las columnas de W_1, entonces la descomposición del Análisis de Varianzan sería $w^T w = w^T P_1 w + w^T (I - P_1) w$; de modo que el coeficiente de determinación de la regresión de la variable k-ésima sobre el resto es

$$R_k^2 = \frac{w^T P_1 w}{w^T w} = w^T P_1 w$$

pues las columnas de W están normalizadas y $w^T w = 1$. Entonces

$$q_{kk} = \frac{1}{1 - R_k^2}$$

de aquí que si w se puede aproximar bien mediante una combinación lineal de las demás variables, tendremos que R_k^2 estará cercana a 1 y por lo tanto q_{kk} puede ser muy grande.

- Puede verse, en general, que también los demás elementos diagonales pueden expresarse en forma similar, $q_{jj} = 1/(1 - R_j^2)$. Resumiendo, si alguna variable puede expresarse linealmente en términos de las demás entonces el corresondiente elemento diagonal de Q puede ser muy grande. En este caso, lo que sucede es que la varianza del estimador, $\mathsf{Var}(\widehat{\gamma}_j) = \sigma^2 q_{jj}$, puede hacerse muy grande, impactando la calidad de las inferencias. A los elementos q_{jj} se les llama "factores de inflación de varianza".

- Veamos ahora como detectar las variables involucradas en una colinealidad. Consideremos el modelo de regresión en términos de las variables originales $y = X\beta + e$. Note que, de la descomposición espectral, se tiene

$$X^T X = \sum_{j=1}^{p} \lambda_j v_j v_j^T \qquad \left(\text{también se tiene que:} \quad (X^T X)^{-1} = \sum_{j=1}^{p} \frac{1}{\lambda_j} v_j v_j^T \right)$$

Supongamos que $\lambda_p \approx 0$, entonces $X^T X v_p = \lambda_p v_p \approx 0$, entonces también $v_p^T X^T X v_p \approx 0$ y de aquí que $X v_p \approx 0$, por lo tanto

$$v_{1p} x_1 + v_{2p} x_2 + \cdots + v_{pp} x_p \approx 0$$

note que los elementos dominantes del vector propio identificarán a las variables involucradas en la colinealidad.

- Notemos ahora el efecto de las colinealidades en los estimadores de los coeficientes de regresión

$$\widehat{\beta} = (X^T X)^{-1} X^T y = \sum_{j=1}^{p} \frac{1}{\lambda_j} v_j v_j^T X^T y \equiv \sum_{j=1}^{p} \frac{c_j}{\lambda_j} v_j = \sum_{j=1}^{p-1} \frac{c_j}{\lambda_j} v_j + \frac{c_p}{\lambda_p} v_p$$

si $\lambda_p \approx 0$ entonces los elementos de $\widehat{\beta}$ correspondientes a los elementos dominantes del vector propio v_p se verán bastante afectados debido a $\lambda_p \approx 0$. En otras palabras, los coeficientes estimados de variables involucradas en colinealidades no son confiables.

Resumen de Clase 9: Lunes 28 de febrero

- En la sesión pasada comentamos que la presencia de colinealidades (i.e. casi dependencias lineales entre las columnas de la matriz X en un modelo de regresión) puede afectar tanto a la estimación misma de los coeficientes del modelo como a la varianza de esas estimaciones.

- Una forma de enfrentar la presencia de colinealidades es detectar conjuntos de variables que sean colineales y efectuar un proceso de eliminación.

- En caso de que deseemos quedarnos con todas las variables predictoras, una opción es usar Regresión Ridge. Supongamos el modelo usual

$$y = X\beta + e, \qquad \text{con} \quad e \sim N(0, \sigma^2 I_n)$$

  El estimador Ridge para β se define como

$$\widehat{\beta}_R = (X^T X + kI)^{-1} X^T y$$

  donde $k \geq 0$ es el parámetro Ridge.

- Note que si $X^T X = \sum_{j=1}^{p} \lambda_j v_j v_j^T$, entonces $(X^T X)^{-1} = \sum_{j=1}^{p} \frac{1}{\lambda_j} v_j v_j^T$ y también $(X^T X + kI)^{-1} = \sum_{j=1}^{p} \frac{1}{k+\lambda_j} v_j v_j^T$.

- El estimador Ridge es sesgado. Para ver esto, notemos primero que

$$(X^T X + kI)^{-1}(X^T X + kI) = I \quad \Rightarrow \quad (X^T X + kI)^{-1} X^T X = I - k(X^T X + kI)^{-1}$$

  entonces

$$\mathsf{E}(\widehat{\beta}_R) = (X^T X + kI)^{-1} X^T X \beta = \beta - k(X^T X + kI)^{-1}\beta \neq \beta$$

  de aquí que el sesgo está dado por:

$$\text{sesgo} = \beta - \mathsf{E}(\widehat{\beta}_R) = k(X^T X + kI)^{-1}\beta$$

- Sin embargo, el estimador Ridge puede tener menor Error Cuadrático Medio.

$$\mathsf{ECM}(\widehat{\beta}_R) = \mathsf{E}\left((\widehat{\beta}_R - \beta)^T(\widehat{\beta}_R - \beta)\right) = \cdots = \text{tr Var}(\widehat{\beta}_R) + k^2\beta^T(X^T X + kI)^{-2}\beta$$

$$\mathsf{ECM}(\widehat{\beta}_R) = \text{Varianza} + \text{Sesgo}^2 = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(k+\lambda_j)^2} + k^2 \sum_{j=1}^{p} \frac{(\beta^T v_j)^2}{(k+\lambda_j)^2}$$

  por otro lado, el error cuadrático medio del estimador de mínimos cuadrados es

$$\mathsf{ECM}(\widehat{\beta}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$$

  Hoerl, Kennard y Marquardt, en una serie de artículos (Technometrics, 1970) sobre el tema, mostraron que

$$\mathsf{ECM}(\widehat{\beta}_R) < \mathsf{ECM}(\widehat{\beta})$$

  si k es tomado en la región $0 < k < \sigma^2/||\beta||^2$.

Resumen de Clase 10: Miércoles 2 de marzo

- Una forma alternativa de ver a la Regresión Ridge. Consideremos el modelo estándar de regresión, $y = X\beta + e$, con $e \sim N_(0, \sigma^2 I)$ y consideremos el siguiente problema de estimación restringida:

$$\min_{\beta} \quad (y - X\beta)^T (y - X\beta) \qquad \text{s.a.} \quad ||\beta||^2 \leq c$$

- El Lagrangiano es

$$(y - X\beta)^T (y - X\beta) + \lambda(||\beta||^2 - c)$$

derivando con respecto a β e igualando a cero, puede verse que la solución del problema restringido es de la forma:

$$\widehat{\beta}_R = (X^T X + \lambda I)^{-1} X^T y$$

esto es, el estimador Ridge es la solución del problema restringido.

- Vimos que la colinealidad ocasiona inflación en la varianza y en la magnitud de los estimadores de mínimos cuadrados, de aquí que si no dejamos que las magnitudes de los coeficientes estimados sea grande es razonable esperar que la solución sea precisamente el estimador Ridge.

- Una vez que planteamos el problema de minimización restringida, sujeta a $||\beta||^2 \leq c$, es natural considerar, por ejemplo

$$\min_{\beta} \quad (y - X\beta)^T (y - X\beta) \qquad \text{s.a.} \quad ||\beta||_1^2 = \sum_{j=1}^{p} |\beta_j| \leq c$$

la solución a este problema da lugar a la Regresión Lasso.

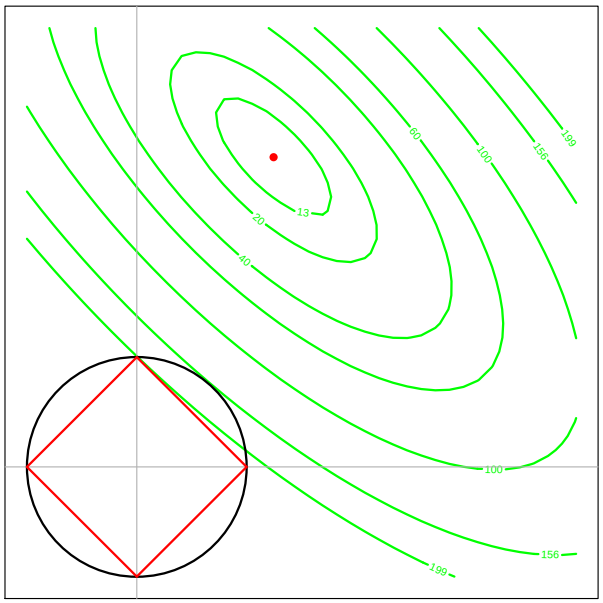
- Ridge y Lasso son casos especiales de "Regresión Penalizada" en donde minimizamos expresiones de la forma:

$$\text{Ajuste} + \text{Penalización}$$

(el ajuste de splines puede verse que también cae en esta clase de problemas).

- Una propiedad importante de Lasso es que, además de no permitir valores demasiado grandes (minimizando con ello problemas de colinealidad), también actúa como un procedimiento de selección de variables. La siguiente gráfica ilustra como Lasso elimina una variable y Ridge no.

Minimización de SCE, sujeta a restricciones Ridge y Lasso



- Incluimos el código R por si alguien quiere saber como se hizo la gráfica.

```
# Grafica Lasso y Ridge
set.seed(75757)
n   <- 20
p   <- 2
x1  <- runif(n,0,2)
x2  <- runif(n,0,2)
y   <- x1 + 3*x2 + rnorm(n)
X   <- cbind(x1,x2)
be  <- solve(t(X)%*%X,t(X)%*%y)
sce <- function(b){ return(sum((y-X%*%b)^2)) }
m   <- 40
b1  <- seq(-1,4,length=m)
b2  <- seq(-1,4,length=m)
z   <- matrix(0,m,m)
for(i in 1:m){
 for(j in 1:m){
  z[i,j] <- sce(c(b1[i],b2[j]))}}

par(mar=c(2,2,2,2))
contour(b1,b2,z, mgp=c(1.5,.6,0),xlab="",ylab="",,xaxt="n",yaxt="n",
 xlim=c(-1,4),ylim=c(-1,4),levels=c(13,20,40,60,100,156,199),
 cex.lab=.9, cex.axis=.9, lwd=2, col="green", drawlabels=T,cex.main=1,
 main="Minimizacin de SCE, sujeta a restricciones Ridge y Lasso" )
points(be[1],be[2],pch=16,col="red")
abline(h=0,v=0,col=gray(.7))
symbols(0,0,circles=1,add=T,inches=F,col="blue",lwd=2)
segments(1,0,0,1,lwd=2,col="red")
segments(0,1,-1,0,lwd=2,col="red")
segments(-1,0,0,-1,lwd=2,col="red")
segments(0,-1,1,0,lwd=2,col="red")
```

- A continuación tenemos el ejemplo visto en clase. Primero, lectura de datos, cálculo de correlaciones y una gráfica con fines exploratorios.

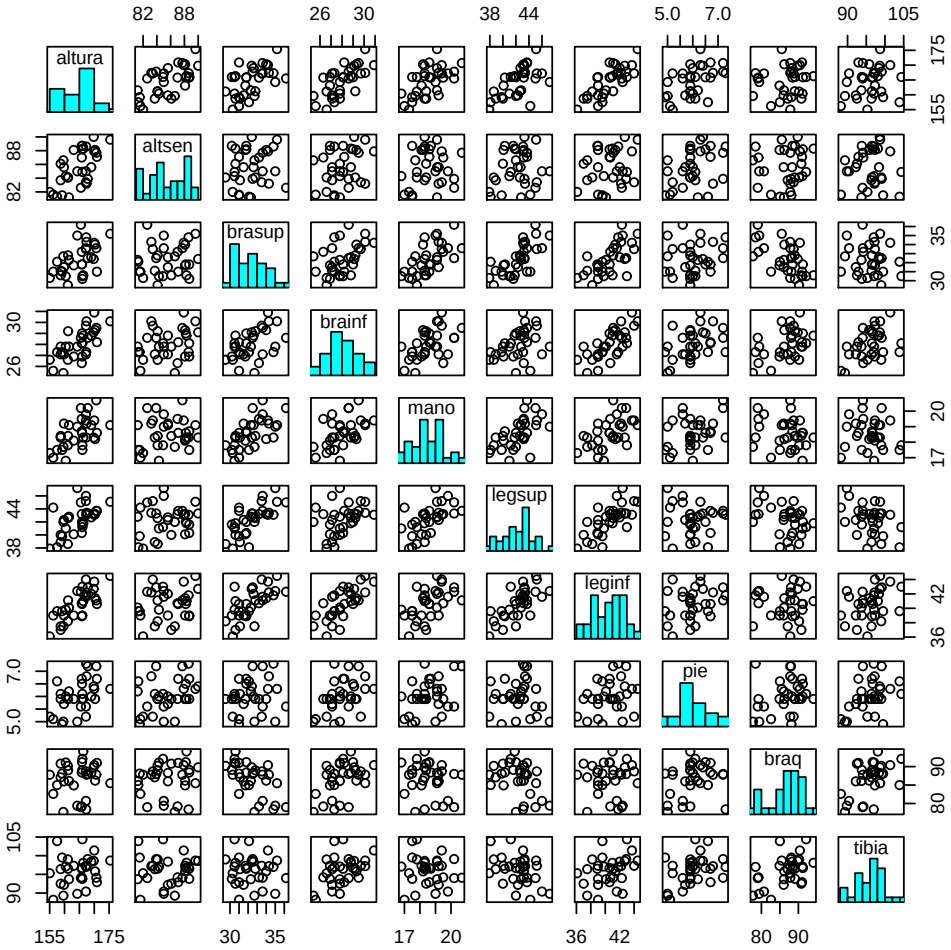
```
# Datos fisicos de solicitantes femeninas a un departamento
# de policia de Estados Unidos. Datos obtenidos de:
# Gunst,R.F. & Mason,R.L. (1980)
# Regression analysis and its application
# Marcel Dekker.
#
# altura = altura
# altsen = altura estando sentadas
# brasup = distancia del hombro al codo
# brainf = distancia del codo a la muneca
# mano   = longitud de la mano
# legsup = distancia de la cadera a la rodilla
# leginf = distancia de la rodilla al tobillo
# pie    = longitud del pie
# braq   = indice brainf/brasup x 100
# tibia  = indice tibia/femur x 100
# (Todas las longitudes en centimetros, excepto ultimas tres)
```

```
datos  <- read.csv(
"c:\\Documents and Settings\\...\\mediciones.csv",
 header=FALSE)
names(datos) <- c("altura","altsen","brasup","brainf","mano",
                  "legsup","leginf","pie","braq","tibia")
attach(datos)
```

```
      altura altsen brasup brainf mano legsup leginf pie braq tibia
 [1,]  165.8   88.7   31.8   28.1 18.7   40.3   38.9 6.7 88.4  96.5
 [2,]  169.8   90.0   32.4   29.1 18.3   43.3   42.7 6.4 89.8  98.6
 [3,]  170.7   87.7   33.6   29.5 20.7   43.7   41.1 7.2 87.8  94.1
 [4,]  170.9   87.1   31.0   28.2 18.6   43.7   40.6 6.7 91.0  92.9
...
[30,]  164.3   85.0   35.0   27.8 19.0   47.2   42.4 5.0 79.4  89.8
[31,]  165.5   82.6   36.2   28.6 20.2   45.0   42.3 5.6 79.0  94.0
[32,]  167.2   85.0   33.6   27.1 19.8   46.0   41.6 5.6 80.7  90.4
[33,]  167.2   83.4   33.5   29.7 19.4   45.2   44.0 5.2 88.7  97.3
```

```
corr <- cor(datos)
     altura altsen brasup brainf  mano legsup leginf   pie  braq tibia
altura 1.00   0.65   0.54   0.70  0.58   0.59   0.78  0.49  0.10  0.21
altsen 0.65   1.00   0.14   0.28  0.14   0.19   0.23  0.37  0.11  0.02
brasup 0.54   0.14   1.00   0.47  0.64   0.72   0.66  0.15 -0.58 -0.10
brainf 0.70   0.28   0.47   1.00  0.50   0.37   0.73  0.43  0.44  0.44
  mano 0.58   0.14   0.64   0.50  1.00   0.59   0.54  0.35 -0.19 -0.10
legsup 0.59   0.19   0.72   0.37  0.59   1.00   0.71 -0.03 -0.39 -0.41
leginf 0.78   0.23   0.66   0.73  0.54   0.71   1.00  0.28  0.00  0.34
   pie 0.49   0.37   0.15   0.43  0.35  -0.03   0.28  1.00  0.24  0.40
  braq 0.10   0.11  -0.58   0.44 -0.19  -0.39   0.00  0.24  1.00  0.51
 tibia 0.21   0.02  -0.10   0.44 -0.10  -0.41   0.34  0.40  0.51  1.00
```

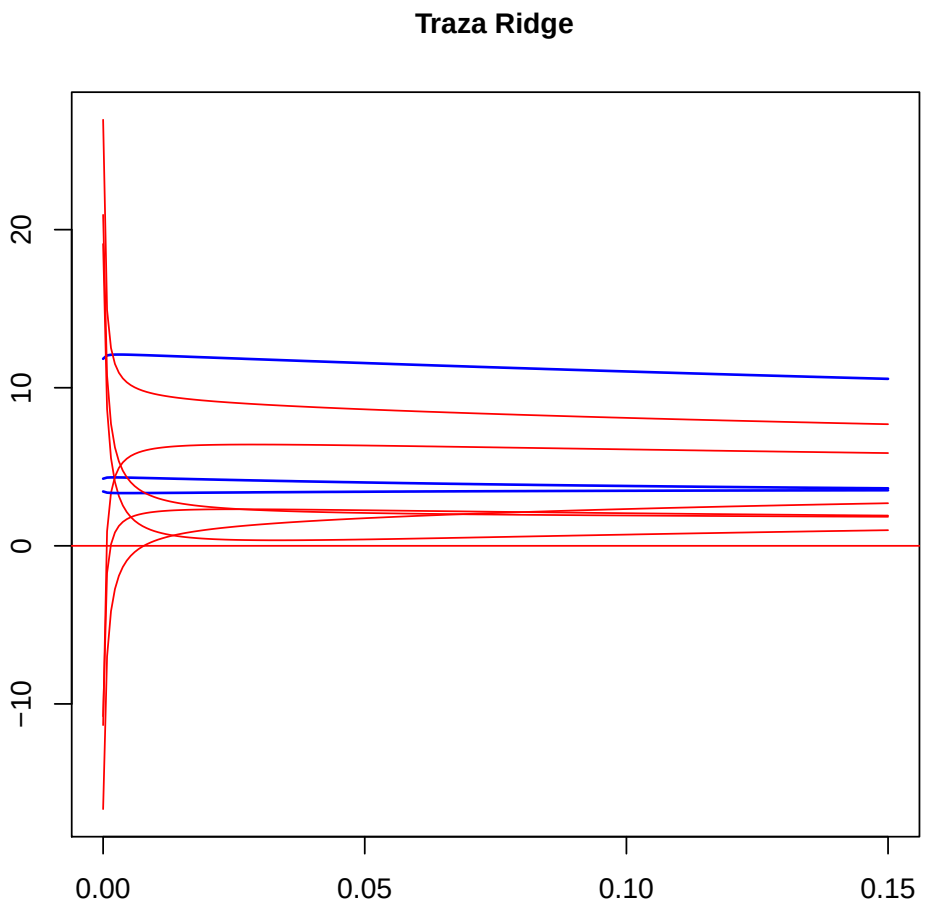
```
n   <- dim(datos)[1]
nom <- c("altura","altsen","brasup","brainf","mano",
         "legsup","leginf","pie","braq","tibia")

panel.hist <- function(x){
 usr    <- par("usr")
 on.exit(par(usr))
 par(usr = c(usr[1:2], 0, 1.5) )
 h      <- hist(x, plot = FALSE)
 breaks <- h$breaks
 nB     <- length(breaks)
 y      <- h$counts
 y      <- y/max(y)
 rect(breaks[-nB], 0, breaks[-1], y, col="cyan") }

# Grafica de todas las variables vs todas las variables
# en un panel con histogramas en la diagonal.
pairs(datos, diag.panel=panel.hist, cex.labels=1)
```



- Ahora, un ajuste estándar de regresión, notar que las variables aparentan no ser significativas (su efecto está oculto por las colinealidades), cálculo de factores de inflación de varianza y detección de colinealidades. Uso de la traza Ridge para la elección del parámetro k.

```
# Ajuste usando lm()
out <- lm(altura ~  altsen+brasup+brainf+mano+legsup+leginf+pie+braq+tibia)
summary(out)

Coefficients:
               Value Std. Error   t value  Pr(>|t|)
(Intercept)   -6.4327  220.2339    -0.0292    0.9770
     altsen    0.7933    0.1544     5.1375    0.0000
     brasup    2.0098    4.3512     0.4619    0.6485
     brainf   -2.2344    5.0047    -0.4465    0.6594
       mano    0.7921    0.5451     1.4531    0.1597
     legsup   -0.8916    4.4210    -0.2017    0.8419
     leginf    2.2742    4.6920     0.4847    0.6325
        pie    0.9548    0.6972     1.3695    0.1841
       braq    0.8529    1.6458     0.5183    0.6092
      tibia   -0.5035    1.9434    -0.2591    0.7979
```

```
Residual standard error: 1.881 on 23 degrees of freedom
Multiple R-Squared: 0.8938
F-statistic: 21.5 on 9 and 23 degrees of freedom, the p-value is 3.629e-009

# Usando variables estandarizadas
X   <- datos[,-1]
W   <- scale(X,scale=F)
W   <- t(t(W)/sqrt(diag(t(W)%*%W)))
bmc <- solve( t(W)%*%W, t(W)%*%altura )

# Factores de inflacion de varianza
WtW <- t(W)%*%W
vif <- diag(solve(WtW))

altsen brasup brainf   mano legsup leginf    pie   braq  tibia
  1.50 483.03 393.39   2.41 894.12 874.06   1.78 461.33 489.72

# Deteccion de variables colineales
eig <- eigen(WtW)

round(eig$values,4)
[1] 3.6211 2.4430 1.0129 0.7695 0.6157 0.3049 0.2318 0.0008 0.0004

round(eig$vectors,2)
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
 [1,] -0.19 -0.15  0.80 -0.29  0.36  0.24 -0.17  0.00  0.01
 [2,] -0.44  0.24 -0.10  0.22  0.26  0.31  0.40  0.58 -0.17
 [3,] -0.39 -0.33 -0.17 -0.23 -0.13  0.31  0.50 -0.51  0.17
 [4,] -0.42  0.08  0.03  0.23 -0.58  0.38 -0.54  0.00  0.00
 [5,] -0.41  0.30 -0.01 -0.35 -0.06 -0.46 -0.04  0.18  0.60
 [6,] -0.46 -0.10 -0.25 -0.17  0.27 -0.37 -0.29 -0.18 -0.60
 [7,] -0.22 -0.36  0.38  0.58 -0.19 -0.49  0.24  0.00  0.00
 [8,]  0.08 -0.55 -0.05 -0.44 -0.38 -0.04  0.04  0.57 -0.16
 [9,] -0.05 -0.53 -0.33  0.26  0.44  0.11 -0.34  0.13  0.45

 # Traza Ridge
p  <- dim(W)[2]
m  <- 200
k  <- seq(0,0.15,length=m)
bR <- matrix(0,p,m)
for(j in 1:m){ bR[,j] <- solve(t(W)%*%W+k[j]*diag(p),t(W)%*%altura) }

mbR <- min(bR)
MbR <- max(bR)
co  <- c("blue","red","red","blue","red","red","blue","red","red")
lw  <- c(1.5,1,1,1.5,1,1,1.5,1,1)
plot(1,1,xlim=c(0,.15),ylim=c(mbR,MbR),type="n",
 xlab="",ylab="",main="Traza Ridge",cex.main=1)
for(i in 1:9){ lines(k,bR[i,],col=co[i], lwd=lw[i]) }
abline(h=0,col="red")
```

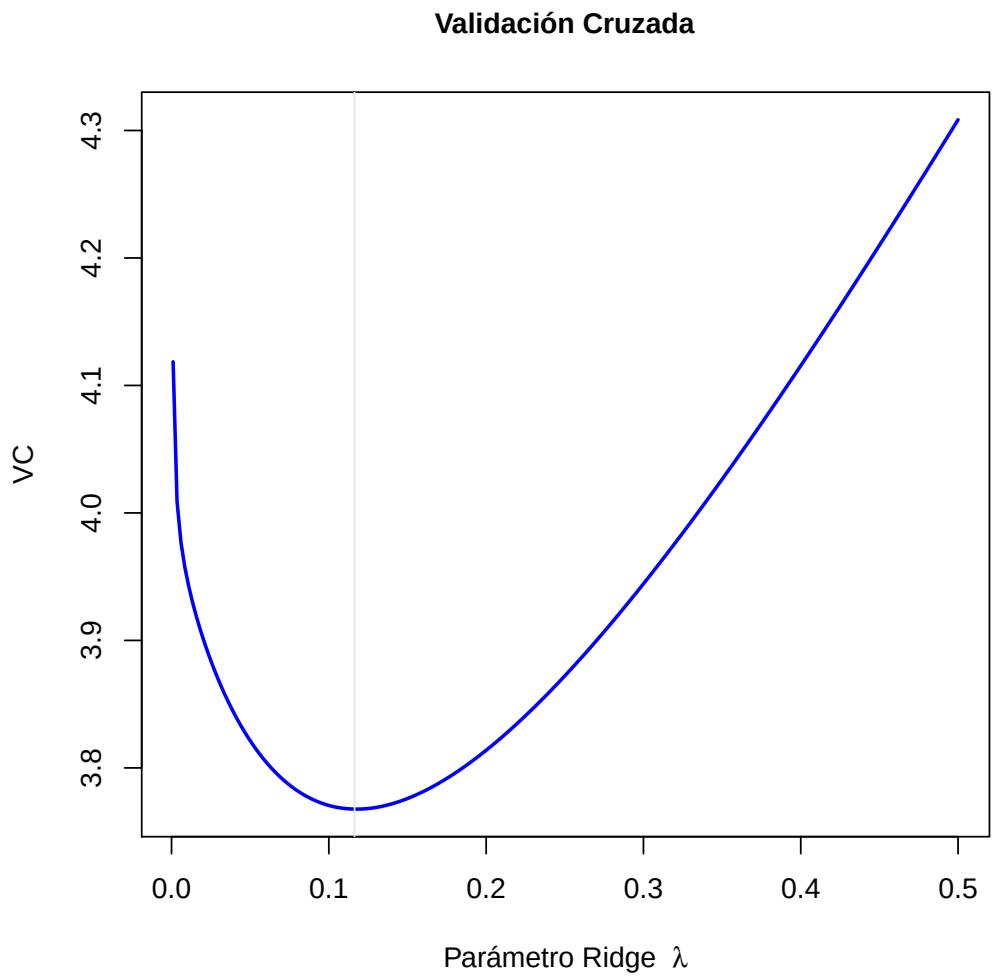


- Elegimos k, ni muy grande que incremente el sesgo, ni muy pequeño que aumente la varianza, $\cdots$ decidimos que $k = 0.06$ es ok. La siguiente tabla muestra que el uso de este valor de k, tiene impactos positivos (ver que ahora los vif's son bastante pequeños comparados con los de la regresión usual).

```
# Usando parametro Ridge k = 0.06
k    <- 0.06
bR   <- solve(t(W)%*%W+k*diag(p),t(W)%*%altura)
QR   <- ( solve(t(W)%*%W+k*diag(p),t(W)%*%W) ) %*% solve(t(W)%*%W+k*diag(p))
vifR <- diag(QR)
round( cbind(bmc,bR,vif,vifR), 2 )
         bmc    bR       vif    vifR
altsen  11.83  11.45     1.50   1.06
brasup  19.09   0.46   483.03   0.97
brainf -16.65   1.89   393.39   1.15
mano     4.25   3.96     2.41   1.68
legsup -11.34   6.30   894.12   0.77
leginf  26.95   8.51   874.06   0.84
pie      3.44   3.44     1.78   1.34
braq    20.94   2.03   461.33   0.62
tibia  -10.78   2.21   489.72   0.92
```

- Un método formal de elección del parámetro k es mediante Validación Cruzada. Enseguida tenemos la implementación computacional de este procedimiento que explicamos en clase.

```
# Seleccion del parametro Ridge usando validacion cruzada
datos  <- read.csv(
"c:\\Documents and Settings\\...\\mediciones.csv",
 header=FALSE)

X   <- as.matrix(datos[,-1])
y   <- as.vector(datos[, 1])
n   <- dim(X)[1]
p   <- dim(X)[2]

# 33 registros: 11 grupos de 3
set.seed(545)
sel <- sample(1:n)
ngV <- 11
mV  <- 3
m   <- 200
kk  <- seq(0.001,.5,length=m)
VC  <- rep(0,m)

for(i in 1:m){
 k   <- kk[i]
 eVC <- rep(0,ngV)
 for(j in 1:ngV){
  selV <- sel[-( (mV*(j-1)+1):(mV*j) )]
  Xk   <- X[selV,]
  yk   <- y[selV]
  alfa <- mean(yk)
  cent <- colMeans(Xk)
  Wk   <- t(t(Xk)-cent)
  std  <- sqrt(diag(t(Wk)%*%Wk))
  Wk   <- t(t(Wk)/std)
  beta <- solve(t(Wk)%*%Wk+k*diag(p),t(Wk)%*%yk)
  XV   <- X[-selV,]
  WXV  <- t(t(XV)-cent)
  WXV  <- t(t(WXV)/std)
  yV   <- y[-selV]
  eVC[j] <- mean( (yV-alfa-WXV%*%beta)^2 ) }
 VC[i] <- mean(eVC) }

plot(kk,VC,type="l",lwd=2,col="blue",main="Validacion Cruzada",cex.main=1,
 xlab=expression(paste("Parametro Ridge  ",lambda)))
abline(v=0.116346734,col=gray(.9))

# el minimo en 0.116346734
```


Validación Cruzada
VC
4.3
4.2
4.1
4.0
3.9
3.8
0.0
0.1
0.2
0.3
0.4
0.5
Parámetro Ridge λ

Resumen de Clase 11: Lunes 7 de marzo

- Regresión en Componentes Principales. En clases pasadas vimos a Regresión Ridge como una forma de estimar los coeficientes de regresión en el caso de variables colineales; ahora comentaremos de otra alternativa.

- Nuevamente, supongamos el modelo estandarizado $y = \alpha 1 + W\gamma + e$ con $e \sim N(0, \sigma^2 I_n)$. Supongamos que $\lambda_1 \geq \cdots \geq \lambda_r \geq \lambda_{r+1} \geq \cdots \geq \lambda_p > 0$ son los valores propios de $W^T W$ y supongamos que $\lambda_{r+1}, \cdots, \lambda_p$ son muy pequeños. Note que

$$(W^T W)^{-1} = \sum_{j=1}^{p} \frac{1}{\lambda_j} v_j v_j^T$$

  donde los v_j's son los vectores propios de $W^T W$ correspondientes a los λ_j's. Definamos

$$(W^T W)^{+} = \sum_{j=1}^{r} \frac{1}{\lambda_j} v_j v_j^T$$

- El estimador en componentes principales de γ se define como

$$\tilde{\gamma} = (W^T W)^{+} W^T y$$

- Note que $\tilde{\gamma}$ es sesgado. Para ver esto, note que

$$(W^T W)^{-1} = \sum_{j=1}^{p} \frac{1}{\lambda_j} v_j v_j^T = \sum_{j=1}^{r} \frac{1}{\lambda_j} v_j v_j^T + \sum_{j=r+1}^{p} \frac{1}{\lambda_j} v_j v_j^T = (W^T W)^{+} + \sum_{j=r+1}^{p} \frac{1}{\lambda_j} v_j v_j^T$$

$$\begin{aligned}
\mathsf{E}(\tilde{\gamma}) = (W^T W)^{+} W^T \mathsf{E}(y) &= \left((W^T W)^{-1} - \sum_{j=r+1}^{p} \frac{1}{\lambda_j} v_j v_j^T \right) W^T (\alpha 1 + W\gamma) \\
&= \left((W^T W)^{-1} - \sum_{j=r+1}^{p} \frac{1}{\lambda_j} v_j v_j^T \right) W^T W \gamma = \gamma - \left(\sum_{j=r+1}^{p} \frac{1}{\lambda_j} v_j v_j^T \right) \left(\sum_{i=1}^{p} \lambda_i v_i v_i^T \right) \gamma \\
&= \gamma - \sum_{j=r+1}^{p} v_j v_j^T \gamma \neq \gamma
\end{aligned}$$

- $\tilde{\gamma}$ es sesgado, pero, por otro lado, tiene menor varianza que el estimador de mínimos cuadrados. Para ver esto, notamos primero que

$$(W^T W)^{+} W^T W (W^T W)^{+} = (W^T W)^{+}$$

  así que

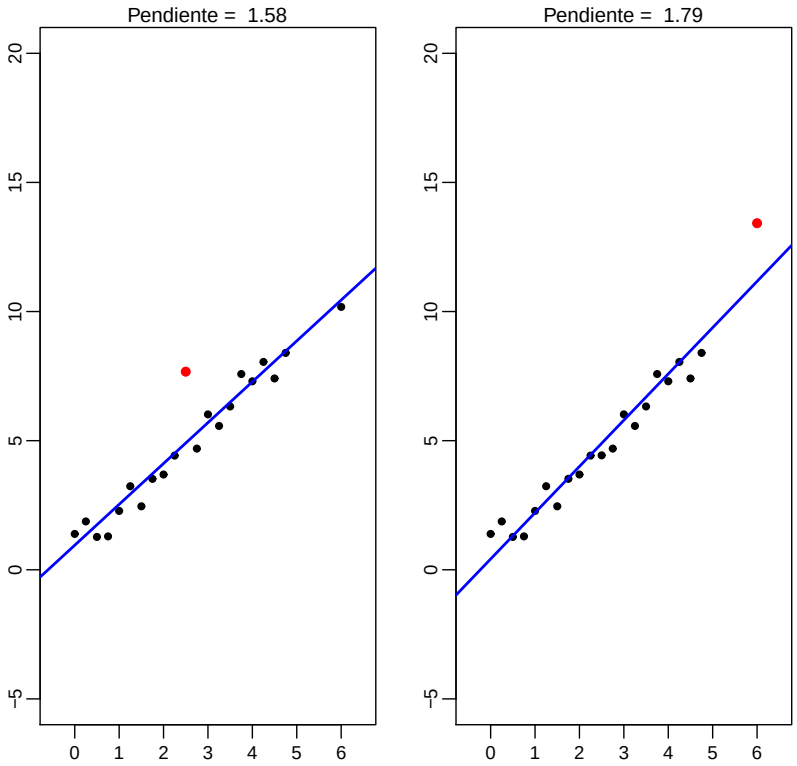
$$\mathsf{Var}(\tilde{\gamma}) = (W^T W)^{+} W^T \, \sigma^2 I \, W (W^T W)^{+} = \sigma^2 (W^T W)^{+}$$

  entonces

$$\mathsf{Var}(\tilde{\gamma}_k) = \sigma^2 \sum_{j=1}^{r} \frac{1}{\lambda_j} v_{jk}^2 \leq \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} v_{jk}^2 = \mathsf{Var}(\widehat{\gamma}_k)$$

  donde $\widehat{\gamma}_k$ es el estimador de mínimos cuadrados de γ_k.

- **Herramientas de Diagnóstico**. Comentaremos ahora sobre algunas herramientas que ayudaran en la construcción de modelos de regresión. Primero veremos como detectar observaciones con puntos influyentes (un punto es influyente si cambios o perturbaciones en esa observación pueden tener impactos fuertes en los estimadores) (el siguiente código en R ilustra dinámicamente estos conceptos)




```
library(tcltk)

grafica <- function(...){
 par(mfcol=c(1,2),mar=c(3,2,2,1))
 n    <- 21
 x    <- seq(0,5,length=n)
 x[n] <- 6
 set.seed(7587)
 y    <- 1+1.5*x+rnorm(n,mean=0,sd=.5)
 y[11]<- y[11] + as.numeric(tclvalue(sliderval))
 plot(x,y,pch=20,xlim=c(-.5,6.5),ylim=c(-5,20),mgp=c(1.5,.5,0),
  cex.lab=.8,cex.axis=.8,xlab="",ylab="")
 out <- lm(y~x)
 abline(out,lwd=2,col="blue")
 points(x[11],y[11],pch=16,col="red")
 mtext(paste("Pendiente = ",round(out$coef[2],2)),cex=.9)
 set.seed(7587)
 y    <- 1+1.5*x+rnorm(n,mean=0,sd=.5)
 y[n] <- y[n] + as.numeric(tclvalue(sliderval))
 plot(x,y,pch=20,xlim=c(-.5,6.5),ylim=c(-5,20),mgp=c(1.5,.5,0),
  cex.lab=.8,cex.axis=.8,xlab="",ylab="")
 out <- lm(y~x)
```

```
 abline(out,lwd=2,col="blue")
 points(x[n],y[n],pch=16,col="red")
 mtext(paste("Pendiente = ",round(out$coef[2],2)),cex=.9)}

sliderinicio <- 0
slidermin    <- 10
slidermax    <- -10
sliderstep   <- 0.01
tt           <- tktoplevel()
sliderval    <- tclVar(sliderinicio)
slidervallab <- tklabel(tt,text=as.character(tclvalue(sliderval)))
tkgrid(tklabel(tt,text="Algo = "),
       slidervallab,tklabel(tt,text=""))
tkconfigure(slidervallab,textvariable=sliderval)
slider <- tkscale(tt, from=slidermin, to=slidermax, showvalue=F,
                  variable=sliderval,resolution=sliderstep,command=grafica)
tkgrid(slider)
tkgrid(tklabel(tt,text="Nivel de Perturbacion"))
tkfocus(tt)
```

- Necesitamos unos resultados preliminares sobre inversas de matrices particionadas. Queremos calcular la inversa de

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

$$\begin{array}{c} -A_{21}A_{11}^{-1} \\ \hookrightarrow \end{array} \left[\begin{array}{cc|cc} A_{11} & A_{12} & I & O \\ A_{21} & A_{22} & O & I \end{array}\right] \sim \left[\begin{array}{cc|cc} A_{11} & A_{12} & I & O \\ O & A_{22}-A_{21}A_{11}^{-1}A_{12} & -A_{21}A_{11}^{-1} & I \end{array}\right] \sim (*)$$

definamos $S_{2|1} = A_{22} - A_{21}A_{11}^{-1}A_{12}$, entonces

$$(*) \sim \begin{array}{c} \uparrow \\ -A_{12}S_{2|1}^{-1} \end{array} \left[\begin{array}{cc|cc} A_{11} & A_{12} & I & O \\ O & S_{2|1} & -A_{21}A_{11}^{-1} & I \end{array}\right] \sim \begin{array}{c} A_{11}^{-1} \\ S_{2|1}^{-1} \end{array} \left[\begin{array}{cc|cc} A_{11} & O & I+A_{12}S_{2|1}^{-1}A_{21}A_{11}^{-1} & -A_{12}S_{2|1}^{-1} \\ O & S_{2|1} & -A_{21}A_{11}^{-1} & I \end{array}\right] \sim$$

$$\sim \left[\begin{array}{cc|cc} I & O & A_{11}^{-1}+A_{11}^{-1}A_{12}S_{2|1}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S_{2|1}^{-1} \\ O & I & -S_{2|1}^{-1}A_{21}A_{11}^{-1} & S_{2|1}^{-1} \end{array}\right]$$

de aquí que

$$A^{-1} = \begin{bmatrix} A_{11}^{-1}+A_{11}^{-1}A_{12}S_{2|1}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S_{2|1}^{-1} \\ -S_{2|1}^{-1}A_{21}A_{11}^{-1} & S_{2|1}^{-1} \end{bmatrix}$$

De manera similar (haciendo las operaciones elementales en otro orden) se puede ver que también:

$$A^{-1} = \begin{bmatrix} S_{1|2}^{-1} & -S_{1|2}^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}S_{1|2}^{-1} & A_{22}^{-1}+A_{22}^{-1}A_{21}S_{1|2}^{-1}A_{12}A_{22}^{-1} \end{bmatrix} \quad \text{donde} \quad S_{1|2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$$

- Ahora, sea h_{ii} el i-ésimo elemento diagonal de la matriz de proyección sobre $\mathcal{C}(X)$, $P = X(X^TX)^{-1}X^T$. Hagamos $X = [1, X_1]$, donde

$$X_1 = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$$

entonces

$$h_{ii} = [1, x_i^T] \left(\begin{bmatrix} 1^T \\ X_1^T \end{bmatrix} [1, X_1] \right)^{-1} \begin{bmatrix} 1 \\ x_i \end{bmatrix} = [1, x_i^T] \begin{bmatrix} n & 1^T X_1 \\ X_1^T 1 & X_1^T X_1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$

haciendo $A_{11} = n$, $A_{12} = 1^T X_1$, $A_{21} = X_1^T 1$ y $A_{22} = X_1^T X_1$, tenemos que

$$S_{2|1} = A_{22} - A_{21} A_{11}^{-1} A_{12} = X_1^T X_1 - X_1^T 1 \left(\frac{1}{n} \right) 1^T X_1 = X_1^T \left(I - \frac{1}{n} J \right) X_1 = S_*$$

del mismo modo se obtienen todos los términos de la inversa:

$$\begin{bmatrix} n & 1^T X_1 \\ X_1^T 1 & X_1^T X_1 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{n} + \bar{x}^T S_*^{-1} \bar{x} & -\bar{x}^T S_*^{-1} \\ -S_*^{-1} \bar{x} & S_*^{-1} \end{bmatrix}$$

de modo que

$$h_{ii} = [1, x_i^T] \begin{bmatrix} \frac{1}{n} + \bar{x}^T S_*^{-1} \bar{x} & -\bar{x}^T S_*^{-1} \\ -S_*^{-1} \bar{x} & S_*^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ x_i \end{bmatrix} = \frac{1}{n} + (x_i - \bar{x})^T S_*^{-1} (x_i - \bar{x}) = \frac{1}{n} + \frac{(DM_i)^2}{n-1}$$

donde DM_i es la distancia de Mahalanobis entre x_i y el centroide $\bar{x}$.

- De la relación anterior, tenemos que $h_{ii} \geq \frac{1}{n}$

- Por otro lado, de $P = PP$ se obtiene que

$$h_{ii} = h_{i1}^2 + \cdots + h_{ii}^2 + \cdots + h_{in}^2 \geq h_{ii}^2$$

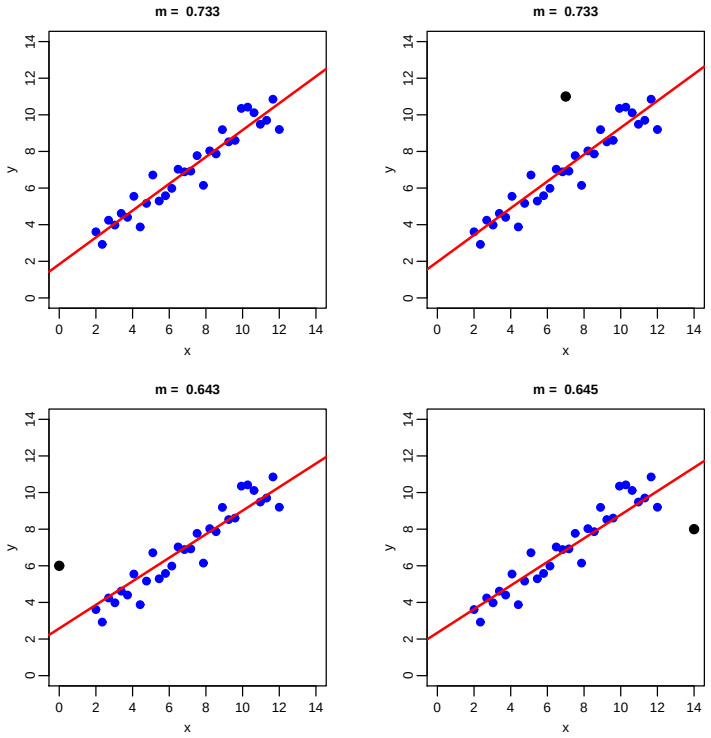
y de aquí se tiene que $h_{ii} \leq 1$. Esto es $\frac{1}{n} \leq h_{ii} \leq 1$ y los valores grandes (esto es, cercanos a 1) indicarán que las correspondientes observaciones son puntos palanca (i.e. influyentes)

- El valor promedio de las h_{ii}'s es $\text{tr}(P)/n$ y esto es lo mismo que p/n.

- Una regla empírica es indicar que los puntos que son candidatos a ser influyentes son aquellos para los cuales $h_{ii} > 2p/n$.

Resumen de Clase 12: Miércoles 9 de marzo

- En la siguiente gráfica hemos agregado tres puntos que obviamente no pertenecen al conjunto de datos que se esta modelando, a estos puntos atípicos se les llama **aberrantes** o **outliers**. Notamos que no todos los datos aberrantes alteran de igual forma los resultados del ajuste: los mostrados en las dos gráficas inferiores si tienen un impacto en el ajuste; a esta clase de datos atípicos se les denomina también datos **influyentes** o **palanca**. Notamos también que los puntos influyentes se encuentran alejados del centro del rango de variación de la variable x, de aquí que una primera herramienta de diagnóstico está basada en la distancia de las variables predictoras al centro de la nube formada por las predictoras.



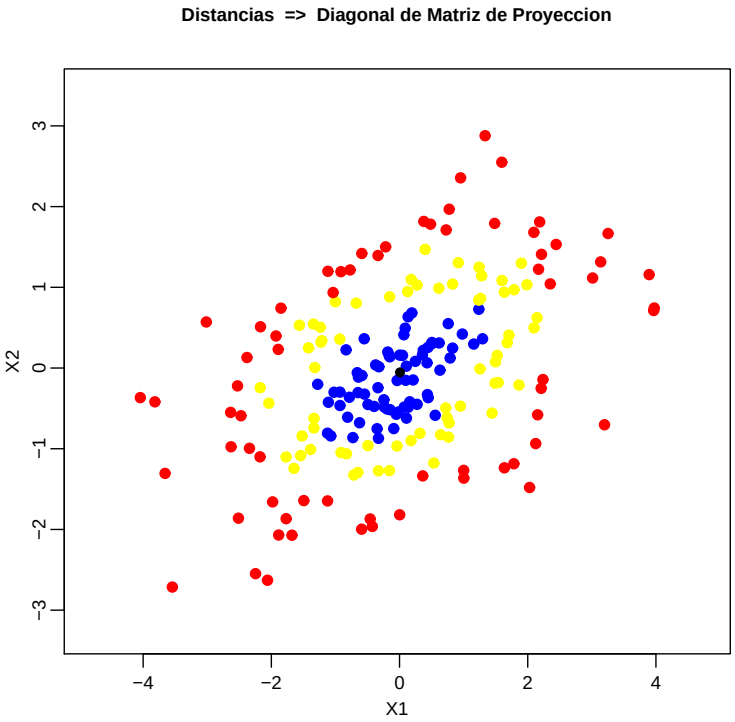

```
par(mfrow=c(2,2),mar=c(3,3,2,2))
set.seed(78321)
n   <- 30
x   <- seq(2,12,length=n)
y   <- 2 + 0.7 * x + rnorm(n,mean=0,sd=.6)
xr  <- c(0,14)
out <- lm(y~x)
plot(x,y, ylim=xr, xlim=xr, pch=19, col="blue",
 ylab="y", mgp=c(1.5,.5,0), xlab="x", cex.axis=.8, cex.lab=.8,
 main=paste("m = ",round((out$coeff)[2],3)), cex.main=.8)
abline(out,col="red",lwd=2)

xx <- c(7,x); yy <- c(11,y); out <- lm(yy~xx)
plot(xx,yy, ylim=xr, xlim=xr, pch=19, col="blue",
 ylab="y", mgp=c(1.5,.5,0), xlab="x", cex.axis=.8, cex.lab=.8,
 main=paste("m = ",round((out$coeff)[2],3)), cex.main=.8)
abline(lm(yy~xx),col="red",lwd=2)
points(xx[1],yy[1],cex=1.2,col="black",pch=19)
```

```
xx <- c(0,x); yy <- c(6,y); out <- lm(yy~xx)
plot(xx,yy, ylim=xr, xlim=xr, pch=19, col="blue",
 ylab="y", mgp=c(1.5,.5,0), xlab="x", cex.axis=.8, cex.lab=.8,
 main=paste("m = ",round((out$coeff)[2],3)), cex.main=.8)
abline(lm(yy~xx),col="red",lwd=2)
points(xx[1],yy[1],cex=1.2,col="black",pch=19)

xx <- c(14,x); yy <- c(8,y); out <- lm(yy~xx)
plot(xx,yy, ylim=xr, xlim=xr, pch=19, col="blue",
 ylab="y", mgp=c(1.5,.5,0), xlab="x", cex.axis=.8, cex.lab=.8,
 main=paste("m = ",round((out$coeff)[2],3)), cex.main=.8)
abline(lm(yy~xx),col="red",lwd=2)
points(xx[1],yy[1],cex=1.2,col="black",pch=19)
```

- En el caso de regresión lineal simple, es fácil detectar cuales puntos son influyentes pues gráficas sencillas como las anteriores los mostrarán. En el caso multivariado no es obvio ver cuales puntos son influyentes pues no es suficiente con ver las distribuciones marginales de las variables (p. ej. haciendo histogramas para cada variable).

- Una forma de detectar puntos influyentes es calcular su distancia al centroide de la nube de predictoras. La forma adecuada de calcular distancias es mediante la **distancia de Mahalanobis** y no la distancia Euclideana. La distancia de Mahalanobis es una "distancia estadística" que incorpora la correlación entre las predictoras. Esta distancia esta íntimamente ligada a la magnitud de los elementos diagonales de la matriz de proyección

$$h_{ii} = \frac{1}{n} + \frac{\mathrm{DM}_i^2}{n-1}.$$



- La gráfica anterior muestra 200 puntos bidimensionales. Se calculó

$$P = X(X^T X)^{-1} X^T$$

y se extrajeron los elementos diagonales $h_{11}, \cdots, h_{nn}$. En rojo se muestran los puntos con los valores más grandes de h_{ii}, en azul los puntos con menor valor y en amarillo los intermedios. En resumen, los elementos diagonales de P dan información acerca de la distancia de los puntos al centro. Así que los h_{ii}'s nos ayudarán a descubrir puntos potencialmente influyentes.

Como vimos la clase pasada, los valores límite para las h_{ii}'s son: $1/n \leq h_{ii} \leq 1$, su valor promedio es p/n (pues hay n valores diagonales y su suma es $\text{tr}(P) = p$). Estudios empíricos sugieren usar

$$h_{ii} > 2\frac{p}{n}$$

como punto de corte para identificar puntos palanca potenciales. (Ver Belsley, Kuh y Welsch de 1980).

```
library(MSBVAR)  # para generar normales multivariadas
set.seed(658743)
var <- matrix( c(2,.7,.7,1), ncol=2 )
mu  <- c(0,0)
n2  <- rmultnorm( 200, mu, var )
hii <- diag(n2 %*% solve( t(n2)%*%n2, t(n2) ) )
col <- ifelse(hii<0.0042,"blue","yellow")
col <- ifelse(hii>0.0117,"red",col)
yr  <- range(n2[,2]); d  <- yr[2]-yr[1]; yr <- yr+.1*d*c(-1,1)
xr  <- range(n2[,1]); d  <- xr[2]-xr[1]; xr <- xr+.1*d*c(-1,1)
plot(n2[,1],n2[,2], ylim=yr, xlim=xr, pch=19, col=col,
 ylab="X2", mgp=c(1.5,.5,0), xlab="X1", cex.axis=.8, cex.lab=.8,
 main="Distancias  =>  Diagonal de Matriz de Proyeccion", cex.main=.8)
points(mean(n2[,1]),mean(n2[,2]),pch=16)
```

- De la clase pasada, cuando vimos inversas de matrices particionadas, tenemos que:

$$S_{2|1}^{-1} = A_{22}^{-1} + A_{22}^{-1} A_{21} S_{1|2}^{-1} A_{12} A_{22}^{-1}$$

esto es,

$$(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} = A_{22}^{-1} + A_{22}^{-1}A_{21}(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1}$$

esta expresión aparentemente no nos ayuda, hasta se ve más complicada; podemos reescribirla como (haciendo $A_{22} = A$, $A_{11} = -C^{-1}$, $A_{21} = U$ y $A_{12} = V$):

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

(ver "Woodbury formula" en http://en.wikipedia.org). Una forma particular es cuando $C = -1$, $U = u$ un vector y $V = u^T$:

$$(A - uu^T)^{-1} = A^{-1} + A^{-1}u(1 - u^T A^{-1}u)^{-1}u^T A^{-1}$$

de aquí, podemos obtener la siguiente expresión, de la cual haremos uso más adelante

$$(X^T X - x_1 x_1^T)^{-1} = (X^T X)^{-1} + (X^T X)^{-1}x_1(1 - x_1^T(X^T X)^{-1}x_1)^{-1}x_1^T(X^T X)^{-1}.$$

- Los elementos diagonales de la matriz de proyección nos identifican puntos del espacio de predictoras que tienen el potencial de ser influyentes. Definimos un punto influyente como aquel que afecta el ajuste. Veamos cómo se afecta el ajuste cuando eliminamos un punto, digamos el primero.

  Queremos comparar $\widehat{\beta}_{(1)}$ (los coeficientes estimados de regresión cuando no consideramos la observación 1) contra $\widehat{\beta}$. Particionamos la información a eliminar:

$$y = \left[\begin{array}{c} y_1 \\ y_{(1)} \end{array} \right] = X\beta + e = \left[\begin{array}{c} x_1^T \\ X_1 \end{array} \right] \beta + e$$

  Queremos $\widehat{\beta}_{(1)}$, obtenido del ajuste de

$$y_{(1)} = X_1\beta + \epsilon$$

  entonces $\widehat{\beta}_{(1)} = (X_1^T X_1)^{-1} X_1^T y_{(1)}$.

- La ecuación anterior implica que si queremos calcular el impacto de cada dato, se requerirían ajustar $n+1$ regresiones. En realidad es suficiente con hacer una sola. Note que

$$X^T X = [x_1, X_1^T] \left[\begin{array}{c} x_1^T \\ X_1 \end{array} \right] = x_1 x_1^T + X_1^T X_1$$

  entonces $X_1^T X_1 = X^T X - x_1 x_1^T$ y, por la fórmula de Woodbury, tenemos

$$(X_1^T X_1)^{-1} = (X^T X)^{-1} + (X^T X)^{-1} x_1 (1 - x_1^T (X^T X)^{-1} x_1)^{-1} x_1^T (X^T X)^{-1}$$

  Similarmente

$$X_1^T y_{(1)} = X^T y - x_1 y_1$$

  entonces

$$\begin{aligned} \widehat{\beta}_{(1)} = & \ (X^T X)^{-1} X^T y + (X^T X)^{-1} x_1 (1 - x_1^T (X^T X)^{-1} x_1)^{-1} x_1^T (X^T X)^{-1} X^T y \\ & - (X^T X)^{-1} x_1 y_1 - (X^T X)^{-1} x_1 (1 - x_1^T (X^T X)^{-1} x_1)^{-1} x_1^T (X^T X)^{-1} x_1 y_1 \\ = & \ \widehat{\beta} + (X^T X)^{-1} x_1 (1 - h_{11})^{-1} x_1^T \widehat{\beta} - (X^T X)^{-1} x_1 y_1 - (X^T X)^{-1} x_1 (1 - h_{11})^{-1} h_{11} y_1 \end{aligned}$$

  de aquí que:

$$\begin{aligned} \widehat{\beta}_{(1)} = & \ \widehat{\beta} + \frac{(X^T X)^{-1}}{1 - h_{11}} \left[x_1 \widehat{y}_1 - (1 - h_{11}) x_1 y_1 - x_1 h_{11} y_1 \right] \\ \widehat{\beta}_{(1)} = & \ \widehat{\beta} - \frac{(X^T X)^{-1}}{1 - h_{11}} x_1 r_1 \\ \widehat{\beta}_{(1)} = & \ \widehat{\beta} - (X^T X)^{-1} x_1 \frac{r_1}{1 - h_{11}} \end{aligned}$$

  En general, si eliminamos la observación i obtenemos

$$\widehat{\beta}_{(i)} = \widehat{\beta} - (X^T X)^{-1} x_i \frac{r_i}{1 - h_{ii}}$$

  esto es, los impactos individuales de cada observación pueden obtenerse del ajuste de regresión original.

- **Residuales**. Definimos los residuales como la diferencia entre lo observado y lo predicho:

$$r = y - \widehat{y} = y - X\widehat{\beta} = y - X(X^T X)^{-1} X^T y = (I - P)y$$

  donde P es la matriz de proyección. Es conveniente pensar a los residuales como una realización de los errores del modelo (aunque **no** son eso), de forma que los podemos usar para detectar algunas deficiencias del modelo. Note que

$$\mathsf{E}(r) = (I - P)X\beta = (X - PX)\beta = 0 \quad \text{y} \quad \mathsf{Var}(r) = \sigma^2 (I - P)$$

  de aquí vemos que los residuales no tienen varianza constante, pero, en promedio, la varianza es $\sigma^2 \mathrm{tr}(I - P)/n = \sigma^2 (n - p)/n \approx \sigma^2$

- Los **residuales estandarizados** se definen como

$$u_i = r_i/\sqrt{\mathsf{CME}}, \quad i = 1, \cdots, n$$

y los **residuales estudentizados** se dividen por la desviación estándar exacta

$$e_i = r_i/\sqrt{\mathsf{CME}(1-h_{ii})}, \quad i = 1, \cdots, n$$

donde h_{ii} es el $i-$ésimo elemento diagonal de la matriz de proyección.

Una herramienta visual para validar el supuesto de normalidad es la gráfica QQ de residuales vs cuantiles teóricos de una normal; sin embargo, siempre tendremos un componente subjetivo en la decisión de si hay o no normalidad. Las gráficas de residuales contra valores ajustados o predictoras, no deben mostrar ningún patrón particular.

- D **de Cook**. El indicador de influencia D de Cook cuantifica el grado de disparidad entre $\widehat{\beta}$ y $\widehat{\beta}_{(i)}$, estimando una distancia estadística entre ellos:

$$D_i = \frac{(\widehat{\beta}_{(i)} - \widehat{\beta})^T (X^T X)\, (\widehat{\beta}_{(i)} - \widehat{\beta})}{p\, \mathsf{CME}}$$

Usando resultados obtenidos arriba, puede verse que dos formas equivalentes son:

$$D_i = \left(\frac{r_i}{\sqrt{\mathsf{CME}(1-h_{ii})}}\right)^2 \left(\frac{h_{ii}}{p(1-h_{ii})}\right) = \frac{e_i^2}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

donde e_i^2 es el residual estudentizado. Una observación tendrá una D de Cook grande si su residual es grande (esto es, si y_i está lejos de $\widehat{y}_i$) y el correspondiente h_{ii} es cercano a 1 (esto es si la x_i está alejada del centroide de las x's).

- En la clase 6 vimos que la región de confianza del $(1-\alpha)\times 100\%$ es

$$\left\{\beta \ : \ \frac{(\beta - \widehat{\beta})^T (X^T X)(\beta - \widehat{\beta})}{p\, \mathsf{CME}} \ \leq \ F^p_{n-p,\alpha}\right\}$$

Note la semejanza de esta expresión con la D de Cook. Ahora, si D_i es del orden de digamos $.8 \approx F^p_{n-p,.5} \approx 1$ esto representaría que los coeficientes estimados, al eliminar la observación i, son afectados con una traslación hasta la frontera de una región del 50% de confianza, y esto representa un impacto drástico (un cambio de esa magnitud es considerado, por consenso, drástico) . **Entonces D de Cook mayores a 1 son outliers**. Otra medida empírica que también es usada es: Si $D_i > 4/n$ entonces la observación i es potencialmente un outlier.

Resumen de Clase 13: Lunes 14 de marzo

- **Normalidad: Prueba de Anderson-Darling**. La prueba Anderson-Darling de bondad de ajuste es una de las pruebas basadas en la función de distribución empírica; cuantifica la discrepancia entre las distribuciones teórica y empírica. Una forma general ded medir esa discrpancia es mediante:

$$A^2 = n \int_{-\infty}^{\infty} \left[F_n(x) - F(x)\right]^2 \psi(x) f(x) dx$$

donde $F_n(x)$ es la función de distribución empírica, $F(x)$ es la función de distribución teórica o hipotetizada, $\psi(x)$ es una función ponderadora y $f(x)$ es la densidad teórica. El caso especial de la Anderson-Darling ocurre cuando se toma la función ponderadora como:

$$\psi(x) = \frac{1}{F(x)\left[1 - F(x)\right]}$$

- $\psi(x)$ tiene como objetivo darle una mayor importancia a discrepancias en las colas de la distribución, permitiéndole al estadístico A^2 una mayor capacidad para detectar otras distribuciones. La razón por la cual se le dá importancia a las colas es que la diferencia $F_n(x) - F(x)$ tiende a cero en las colas, aún cuando la función $F(x)$ no sea la distribución verdadera.

- Sea $y_1, \cdots, y_n$ una muestra aleatoria proveniente de cierta distribución. Si queremos probar la hipótesis $H_0 : y \sim F$, el estadístico A^2 toma la forma

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1) \left[\log(z_i) - \log(1 - z_{n-i+1}) \right]$$

donde $z_i = F(y_i; \theta)$. La regla de decisión consiste en rechazar H_0 para valores grandes de A^2 (para lo cual, necesitaremos la distribución de A^2).

- Primero veamos como se deduce la expresión para el cálculo de A^2. Supongamos $x_1 \leq x_2 \leq \cdots \leq x_n$:

$$\begin{aligned}
\frac{1}{n} A^2 &= \int_{-\infty}^{\infty} \left[F_n(x) - F(x)\right]^2 \psi(x) f(x) dx \\
&= \int_{-\infty}^{x_1} \frac{F(x)}{1 - F(x)} f(x) dx + \sum_{j=1}^{n-1} \int_{x_j}^{x_{j+1}} \frac{\left[\frac{j}{n} - F(x)\right]^2}{F(x)\left[1 - F(x)\right]} f(x) dx + \int_{x_n}^{\infty} \frac{1 - F(x)}{F(x)} f(x) dx \\
&= \int_0^{z_1} \frac{u}{1-u} du + \sum_{j=1}^{n-1} \int_{z_j}^{z_{j+1}} \frac{\left(\frac{j}{n} - u\right)^2}{u(1-u)} du + \int_{z_n}^{1} \frac{1-u}{u} du \equiv a + \sum_j b_j + c
\end{aligned}$$

aquí hicimos el cambio de variable $u = F(x)$ con $du = f(x) fx$.

- Tenemos que

$$\begin{aligned}
a &= \int_0^{u_1} \left(-1 - \frac{1}{1-u}\right) du = -u - \log(1-u) \,|_0^{u_1} = -u_1 - \log(1-u_1) \\
c &= \int_{u_n}^{1} \left(\frac{1}{u} - 1\right) du = \log(u) - u \,|_{u_n}^{1} = -1 - \log u_n + u_n
\end{aligned}$$

- También, con $a_j = j/n$,

$$\begin{aligned}
b_j &= \int_{u_j}^{u_{j+1}} \frac{(a_j - u)^2}{u(1-u)} du = \int_{u_j}^{u_{j+1}} \frac{\left(\left[a_j - 1\right] + \left[1 - u\right] \right)^2}{u(1-u)} du \\
&= (1 - a_j)^2 \int_{u_j}^{u_{j+1}} \frac{1}{u(1-u)} du - 2(1 - a_j) \int_{u_j}^{u_{j+1}} \frac{1}{u} du + \int_{u_j}^{u_{j+1}} \frac{1-u}{u} du \\
&= (1 - a_j)^2 \left[\log u_{j+1} - \log(1 - u_{j+1}) - \log u_j + \log(1 - u_j)\right] - 2(1 - a_j) \left[\log u_{j+1} - \log u_j\right] \\
&\quad + \log u_{j+1} - u_{j+1} - \log u_j + u_j
\end{aligned}$$

simplificando

$$b_j = a_j^2 \log u_{j+1} - (1-a_j)^2 \log(1-u_{j+1}) - u_{j+1} - a_j^2 \log u_j + (1-a_j)^2 \log(1-u_j) + u_j$$

- Entonces

$$\begin{aligned}\frac{1}{n}A^2 = & -u_1 - \log(1-u_1) - 1 - \log u_n + u_n + \sum_{j=1}^{n-1} a_j^2 \log u_{j+1} - \sum_{j=1}^{n-1}(1-a_j)^2 \log(1-u_{j+1}) - \sum_{j=1}^{n-1} u_{j+1} \\ & - \sum_{j=1}^{n-1} a_j^2 \log u_j + \sum_{j=1}^{n-1}(1-a_j)^2 \log(1-u_j) + \sum_{j=1}^{n-1} u_j\end{aligned}$$

es fácil ver que $-u_1 - \sum_{j=1}^{n-1} u_{j+1} + \sum_{j=1}^{n-1} u_j + u_n = 0$; de aquí que

$$\begin{aligned}\frac{1}{n}A^2 = & -\log(1-u_1) - 1 - \log u_n + \sum_{j=1}^{n-1} a_j^2 \log u_{j+1} - \sum_{j=1}^{n-1}(1-a_j)^2 \log(1-u_{j+1}) \\ & - \sum_{j=1}^{n-1} a_j^2 \log u_j + \sum_{j=1}^{n-1}(1-a_j)^2 \log(1-u_j)\end{aligned}$$

definiendo $a_0 \equiv 0$, tenemos

$$\begin{aligned}\frac{1}{n}A^2 &= -1 - \sum_{j=1}^{n}(1-a_{j-1})^2 \log(1-u_j) + \sum_{j=1}^{n} a_{j-1}^2 \log u_j + \sum_{j=1}^{n}(1-a_j)^2 \log(1-u_j) - \sum_{j=1}^{n} a_j^2 \log u_j \\ &= -1 - \sum_{j=1}^{n} \left[(1-a_{j-1})^2 - (1-a_j)^2 \right] \log(1-u_j) - \sum_{j=1}^{n} \left[a_j^2 - a_{j-1}^2 \right] \log u_j\end{aligned}$$

- Por otro lado, $a_j^2 - a_{j-1}^2 = (2j-1)/n^2$ y $(1-a_{j-1})^2 - (1-a_j)^2 = [\,2(n-j)+1\,]/n^2$, entonces

$$\frac{1}{n}A^2 = -1 - \frac{1}{n^2}\sum_{j=1}^{n} [\,2(n-j)+1\,] \log(1-u_j) - \frac{1}{n^2}\sum_{j=1}^{n} [\,2j-1\,] \log u_j$$

cambiando el orden de los subíndices de la primera suma, tenemos que

$$\frac{1}{n}A^2 = -1 - \frac{1}{n^2}\sum_{j=1}^{n} [\,2j-1\,] \log(1-u_{n-j+1}) - \frac{1}{n^2}\sum_{j=1}^{n} [\,2j-1\,] \log u_j$$

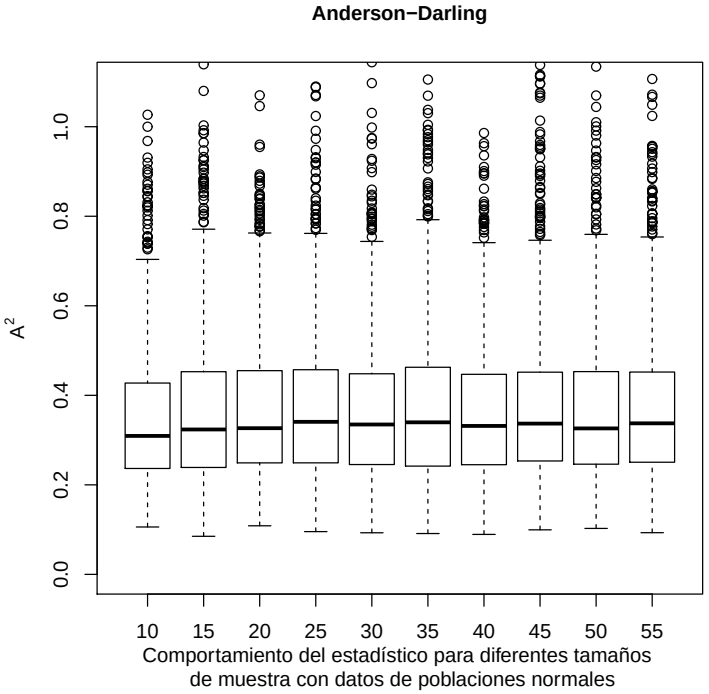
y de aquí se obtiene la forma computacional del estadístico Anderson-Darling

$$A^2 = -n - \frac{1}{n}\sum_{j=1}^{n}(2j-1)\ [\,\log u_j + \log(1-u_{n-j+1})\,]$$

- Mencionamos antes que si $y_1, \cdots, y_n$ es una muestra aleatoria proveniente de cierta distribución y si queremos probar la hipótesis $H_0 : y \sim F$, entonces rechazamos H_0 para valores grandes de A^2. Ahora, para el cálculo de A^2 tenemos que evaluar $u_j = F(x_j; \mu, \sigma^2)$, pero, ¿cúales parámetros debemos usar en F?. Para el caso de la distribución normal, es natural usar $\bar{x}$ y s^2, sin embargo, el hacer esto hace que el problema de determinar la distribución de A^2 sea extremadamente difícil. Una opción es usar simulación. Esta opción es factible por una propiedad interesante: Para familias de localización y escala (como la normal), la distribución de A^2 depende de F pero no de los parámetros de F; así que podemos simular de una normal específica, digamos $N(0,1)$, luego calculamos $u_i = F(x_i; \bar{x}, s^2)$ y la distribución resultante de A^2 será válida cuando la distribución verdadera sea cualquier $N(\mu, \sigma^2)$.

- A continuación presentamos el ejercicio de simulación para determinar la distribución de A^2:

```
# Comportamiento del Anderson-Darling estimando parametros
AD <- function(x){
 x  <- sort(x)
 n  <- length(x)
 m  <- mean(x)
 s  <- sqrt( var(x) )
 z  <- pnorm(x,m,s)
 zi <- rev(z)
 i  <- 1:n
 a2 <- -mean((2*i-1)*(log(z)+log(1-zi)))-n
 return(a2) }
N    <- 100000
mues <- c(10,15,20,25,30,35,40,45,50,55)
and  <- matrix(0,N,length(mues))
for( j in 1:length(mues) ){
 for( i in 1:N ){ z  <- rnorm( mues[j] )
  and[i,j] <- AD(z) }}
# boxplots con menos datos para que grafico en eps no este tan pesado
pocos <- sort(sample(1:N,size=1000))
boxplot(and[pocos,1], and[pocos,2], and[pocos,3], and[pocos,4],
        and[pocos,5], and[pocos,6], and[pocos,7], and[pocos,8],
        and[pocos,9], and[pocos,10], ylim=c(0,1.1),ylab=expression(A^2),
 xlab="Comportamiento del estadistico para diferentes tamanos \n
       de muestra con datos de poblaciones normales",
 names=c("10","15","20","25","30","35","40","45","50","55"),
 main="Anderson-Darling",cex.main=1 )
```



- Los cuantiles estimados son:

cuantiles	10	15	20	25	30	35	40	45	50	55
50	0.32	0.32	0.33	0.33	0.33	0.33	0.34	0.34	0.34	0.34
60	0.35	0.36	0.37	0.37	0.38	0.38	0.38	0.38	0.38	0.38
70	0.40	0.42	0.42	0.42	0.43	0.43	0.43	0.43	0.43	0.43
80	0.47	0.48	0.49	0.49	0.50	0.50	0.50	0.50	0.50	0.50
90	0.58	0.60	0.61	0.61	0.62	0.62	0.62	0.62	0.62	0.62
91	0.59	0.62	0.63	0.63	0.63	0.63	0.64	0.64	0.64	0.64
92	0.61	0.63	0.65	0.65	0.65	0.65	0.66	0.66	0.66	0.66
93	0.63	0.66	0.67	0.67	0.68	0.68	0.68	0.68	0.68	0.68
94	0.66	0.68	0.69	0.70	0.70	0.70	0.70	0.71	0.71	0.71
95	0.69	0.71	0.72	0.73	0.73	0.73	0.74	0.74	0.74	0.74
96	0.72	0.75	0.76	0.77	0.77	0.77	0.77	0.78	0.77	0.78
97	0.76	0.80	0.81	0.82	0.82	0.82	0.82	0.83	0.82	0.83
97.5	0.79	0.83	0.84	0.85	0.86	0.85	0.85	0.86	0.86	0.86
98.0	0.83	0.87	0.88	0.89	0.89	0.89	0.89	0.90	0.90	0.90
99.0	0.93	0.98	0.99	1.01	1.00	1.01	1.01	1.01	1.01	1.01
99.5	1.04	1.09	1.11	1.12	1.12	1.13	1.13	1.13	1.13	1.13
99.9	1.28	1.33	1.38	1.40	1.41	1.41	1.45	1.39	1.42	1.43

  Estos fueron calculados mediante:

```
probas  <-
 c(.5,.6,.7,.8,.90,.91,.92,.93,.94,.95,.96,.97,.975,.98,.99,.995,.999)
pes     <- length(probas)
cuantad <- matrix(0,pes,length(mues))
for(j in 1:length(mues)){cuantad[,j] <- quantile(and[,j], probs = probas)}
rbind( c(0,mues),cbind(probas,round(cuantad,2)) )
```

- La distribución de A^2 depende de n. Una forma de tener un solo juego de cuantiles críticos es usar la siguiente modificación:

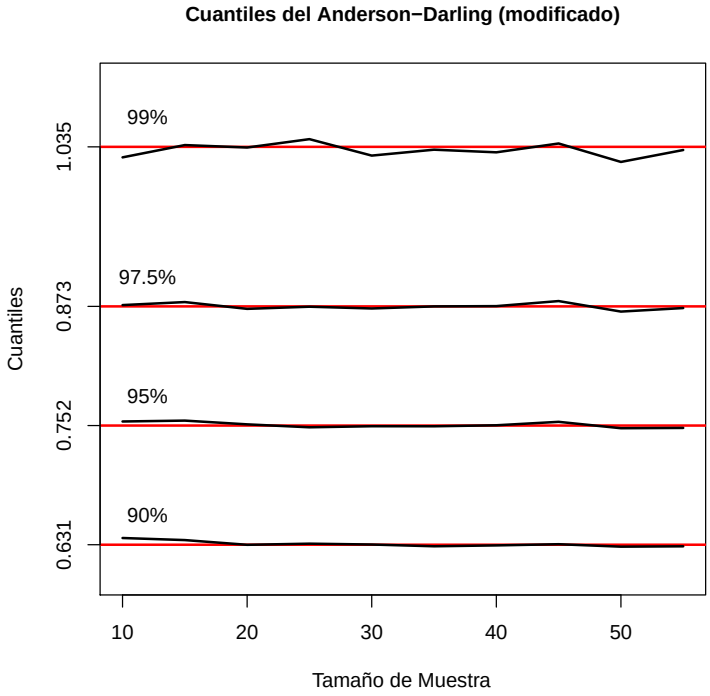
$$A^2_* = A^2\left(1 + \frac{.75}{n} + \frac{2.25}{n^2}\right)$$

  Algunos cuantiles convenientes para pruebas de bondad de ajuste son

90%	95%	97.5%	99%
0.631	0.752	0.873	1.035

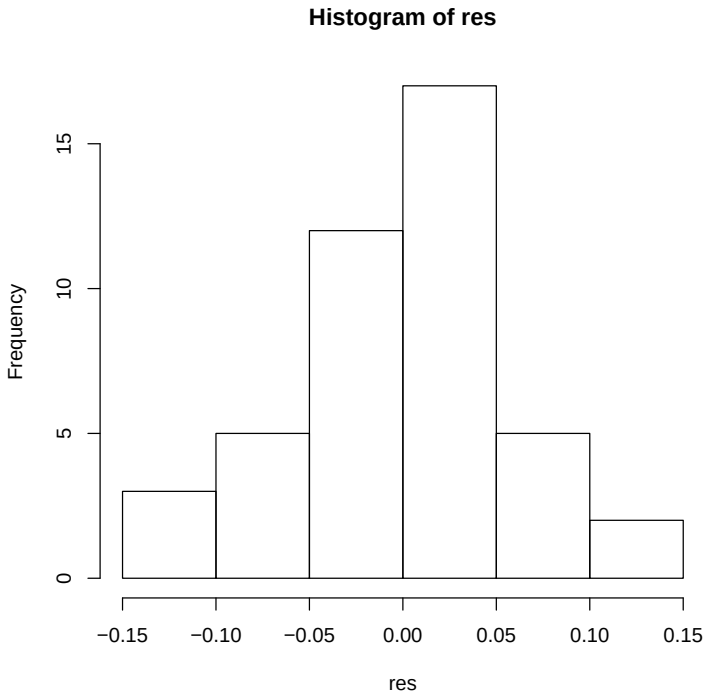
- La siguiente gráfica muestra que, efectivamente, al modificar los cuantiles de A^2 mediante la transformación anterior, tenemos que la distribución de A^2_* no depende de n, de modo que sólo necesitamos conocer los cuantiles antes mencionados.

```
sel <- c(5,10,13,15)
aa  <- cuantad[sel,]
fac <- 1 + 0.75/mues + 2.25/(mues^2)
aa  <- t(fac*t(aa))

adm <- c(.631,.752,.873,1.035)
plot(mues,aa[1,],ylim=c(0.6,1.1),type="n",lwd=2,xlab="Tamano de Muestra",
 ylab="Cuantiles",main="Cuantiles del Anderson-Darling (modificado)",
 cex.main=1,yaxt="n")
abline(h=adm,col="red",lwd=2)
for(i in 1:4){ lines(mues,aa[i,],lwd=2) }
axis(2,at=adm)
text( rep(12,4), adm+.03, labels=c("90%","95%","97.5%","99%") )
```



- Enseguida ilustramos la prueba de normalidad sobre los residuales de la regresión con los datos de Óxidos Nitrosos. La prueba no rechaza normalidad, lo cual es bastante razonable, a juzgar por el histograma que se anexa.

```
# Prueba de normalidad sobre los residuales (Anderson-Darling)
datos  <- read.csv("c:\\Documents and Settings\\...\\EmisionesNOX.csv",header=TRUE)
attach(datos)
out <- lm(NITROX ~ HUMEDAD+TEMP+PRESION)
aa  <- summary(out)
res <- aa$residuals
a2  <- AD(res)              # .2251 no se rechaza normalidad
hist(res)
AD <- function(x){
 x  <- sort(x)
 n  <- length(x)
 m  <- mean(x)
 s  <- sqrt( var(x) )
 z  <- pnorm(x,m,s)
 zi <- rev(z)
 i  <- 1:n
 a2 <- -mean((2*i-1)*(log(z)+log(1-zi)))-n
 return(a2) }
```


Histogram of res
Frequency
0
5
10
15
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
res

Primer Examen de Modelos Estadísticos I

Nombre: _____

1. **(10)** Cuando decimos "modelo lineal usual" queremos decir $y \sim N(X\beta, \sigma^2 I_n)$. Enliste y explique los supuestos que esta expresión implica.

2. **(10)** Con el fin de estimar ciertos parámetros, β_0 y β_1, se pueden hacer tres tipos de observaciones: Las del primer tipo tienen media β_0, las del segundo tipo con media $\beta_0 + \beta_1$, y las del tercer tipo con media $\beta_0 - 2\beta_1$. Suponga además que las observaciones están sujetas a errores independientes con media cero y varianza constante. Suponga que se efectúan m observaciones del primer tipo, m observaciones del segundo tipo y n observaciones del tercer tipo. Encuentre los estimadores de mínimos cuadrados de β_0 y β_1 y pruebe que esos estimadores no están correlacionados si $m = 2n$.

3. **(10)** Sean $y_1, y_2, \cdots, y_n$, variables aleatorias independientes e igualmente distribuídas $N(\mu, \sigma^2)$. Considere la cantidad

$$Q = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2$$

   Escriba a Q como una forma cuadrática en $y = (y_1, y_2, \cdots, y_n)^T$. Calcule el valor esperado de esa forma cuadrática y pruebe con ello que Q es un estimador insesgado de σ^2.

4. **(10)** Considere el modelo lineal $y = X\beta + e$ con X de tamaño $n \times p$ de rango p y $e \sim N(0, \sigma^2 I)$. La desigualdad generalizada de Cauchy–Schwartz es

$$(u^T w)^2 \leq (u^T A u)(w^T A^{-1} w)$$

   la cual es válida para cualesquier vectores u y w y cualquier matriz A positiva definida. Usando esta desigualdad, encuentre intervalos de confianza simultáneos para $u^T\beta$, mostrando que

$$P\left[u^T\widehat{\beta} - L(u) \leq u^T\beta \leq u^T\widehat{\beta} + L(u), \quad \underline{\text{para todo}} \quad u \right] \geq 1 - \alpha$$

   donde $\widehat{\beta}$ es el estimador de mínimos cuadrados de β y $L(u)$ es una cantidad que, además de depender de u, depende también de cantidades tales como p, CME, $(X^TX)^{-1}$ y $F^p_{n-p,\alpha}$. (A estos intervalos de confianza simultáneos se les conoce como *Intervalos de Scheffé*).

   Sugerencia: Tomar $w = \widehat{\beta} - \beta$ en la desigualdad de Cauchy–Schwartz.

5. **(10)** Sea A una matriz $p \times p$, positiva definda. Supongamos que $\lambda_1 \geq \cdots \geq \lambda_r \geq \lambda_{r+1} \geq \cdots \geq \lambda_p > 0$, son los valores propios de A y $v_1, \cdots, v_p$, son los vectores propios de A correspondientes a los λ_j's. Definamos

$$A^+ = \sum_{j=1}^{r} \frac{1}{\lambda_j} v_j v_j^T$$

   Muestre que $A^+AA^+ = A^+$ (esta igualdad la usamos cuando vimos regresión en componentes principales).

6. Considere el modelo $y_i = \beta_0 + z_{i1}\beta_1 + z_{i2}\beta_2 + \epsilon_i,\ i = 1, \cdots, n$, donde las z_{i1}'s y z_{i2}'s están centradas y reescaladas (i.e. $\sum z_{i1} = \sum z_{i2} = 0$ y $\sum z_{i1}^2 = \sum z_{i2}^2 = 1$) y los errores ϵ_i's $\sim NID(0, \sigma^2)$. Defina ρ como $\rho = \sum z_{i1} z_{i2}$.

   (a) **(10)** Sea $\widehat{\beta}$ el estimador de mínimos cuadrados, verifique que

$$\mathsf{Var}(\widehat{\beta}) = \sigma^2 \begin{bmatrix} \frac{1}{n} & 0 & 0 \\ 0 & \frac{1}{k} & -\frac{\rho}{k} \\ 0 & -\frac{\rho}{k} & \frac{1}{k} \end{bmatrix}$$

   donde $k = 1 - \rho^2$.

   (b) **(10)** Suponga que $\widehat{\theta}$ es un estimador de θ, donde este es un determinado parámetro. El error cuadrático medio de $\widehat{\theta}$ se define como $\mathsf{ECM}(\widehat{\theta}) = \mathsf{E}(\widehat{\theta} - \theta)^2$. Sabemos que

$$\mathsf{ECM}(\widehat{\theta}) = \mathsf{Var}(\widehat{\theta}) + \Big(\mathsf{E}(\widehat{\theta}) - \theta\Big)^2$$

   (i.e. Error cuadrático medio = Varianza + Sesgo cuadrático). Considere

$$\widetilde{\beta}_1 = \frac{\sum z_{i1} y_i}{\sum z_{i1}^2}$$

   Muestre que $\mathsf{ECM}(\widetilde{\beta}_1) = \sigma^2 + \rho^2 \beta_2^2$.

   (c) **(10)** ¿Bajo que condiciones $\mathsf{ECM}(\widetilde{\beta}_1) < \mathsf{ECM}(\widehat{\beta}_1)$?

7. Suponga el modelo lineal usual y considere los residuales estudentizados

$$e_i = \frac{r_i}{\sqrt{\mathsf{CME}\,(1 - h_{ii})}}, \quad i = 1, \cdots, n$$

   donde los residuales r_i's son tales que $r = (r_1, \cdots, r_n)^T = (I - P)y$

   (a) **(10)** Muestre que

$$u_i = \frac{r_i}{\sqrt{\sigma^2\,(1 - h_{ii})}} \sim N(0, 1)$$

   (b) **(10)** Sabemos que $\mathsf{CME} = \mathsf{SCE}/(n - p)$ y que $\chi = \mathsf{SCE}/\sigma^2 \sim \chi^2_{n-p}$, de aquí que si consideramos los residuales estudentizados

$$e_i = \frac{r_i}{\sqrt{\mathsf{CME}\,(1 - h_{ii})}} = \frac{u_i}{\sqrt{\frac{\chi}{n-p}}}$$

   de aquí que es natural esperar que los residuales estudentizados tengan una distribución t de Student; sin embargo, esto no es cierto, ¿Porqué? (justificar la respuesta).

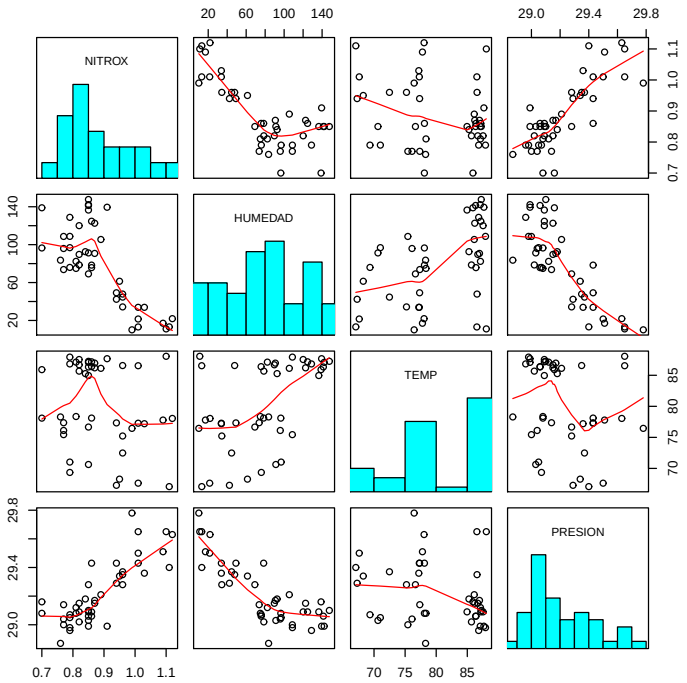
Duración del examen: Dos horas y media.

Resumen de Clase 14: Miércoles 16 de marzo

- **Ejemplo: Análisis de Emisiones**. Las emisiones de vehículos, por ley, están sujetas a ciertos estándares. Los datos del archivo EmisionesNOX.csv corresponden a un estudio sobre emisiones con camionetas diesel. Los datos son 44 registros (en una misma camioneta) de óxidos nitrosos tomados bajo diferentes condiciones ambientales de humedad, temperatura y presión barométrica. Uno de los objetivos de ese estudio era el de construír ecuaciones predictivas para ser usadas para transformar lecturas de emisiones a una escala estándar con condicones ambientales fijas. Las variables son:

NOX	—	Óxidos nitrosos (NO y NO_2)
HUMEDAD	—	Humedad (H_2O/lb de aire)
TEMP	—	Temperatura ($^\circ$F)
PRESION	—	Presión barométrica (pulg. de Hg)

La siguiente lámina tiene una representación gráfica de los datos.




```
datos  <- read.csv("c:\\Documents and Settings\\...\\EmisionesNOX.csv",header=TRUE)
histo <- function(x, ...){
 usr <- par("usr")
 on.exit(par(usr))
 par( usr=c(usr[1:2],0,1.5) )
 h      <- hist(x,plot=FALSE)
 cortes <- h$breaks
 nc     <- length(cortes)
 y      <- h$counts
 y      <- y/max(y)
 rect(cortes[-nc], 0, cortes[-1], y, col="cyan")}
pairs(datos, diag.panel=histo, cex.labels=.9, panel=panel.smooth)
```

- **Datos de Emisiones**[*].

NITROX	HUMEDAD	TEMP	PRESION	NITROX	HUMEDAD	TEMP	PRESION
0.81	74.92	78.36	29.08	0.76	83.61	78.29	28.87
0.96	44.64	72.48	29.37	0.79	75.97	69.35	29.07
0.96	34.30	75.22	29.28	0.77	108.66	75.44	29.00
0.94	42.36	67.28	29.29	0.82	78.59	85.67	29.02
0.99	10.12	76.45	29.78	1.01	33.85	77.28	29.43
1.11	13.22	67.07	29.40	0.94	49.20	77.33	29.43
1.09	17.07	77.79	29.51	0.86	75.75	86.39	29.06
0.77	73.70	77.36	29.14	0.79	128.81	86.83	28.96
1.01	21.54	67.62	29.50	0.81	82.36	87.12	29.12
1.03	33.87	77.20	29.36	0.87	122.60	86.20	29.15
0.96	47.85	86.57	29.35	0.86	24.69	87.17	29.09
1.12	21.89	78.05	29.63	0.82	120.04	87.54	29.09
1.01	13.14	86.54	29.65	0.91	139.47	87.67	28.99
1.10	11.09	88.06	29.65	0.89	105.44	86.12	29.21
0.86	78.41	78.11	29.43	0.87	90.74	86.96	29.17
0.85	69.15	76.66	29.28	0.85	142.20	87.08	28.99
0.70	96.50	78.10	29.08	0.85	136.52	84.96	29.09
0.79	108.72	87.93	28.98	0.70	138.90	85.91	29.16
0.95	61.37	68.27	29.34	0.82	89.69	86.69	29.15
0.85	91.26	70.63	29.03	0.84	92.59	85.27	29.18
0.79	96.83	71.02	29.05	0.85	147.63	87.25	29.10
0.77	95.94	76.11	29.04	0.85	141.35	86.34	29.06

(*) Gunst & Mason (1980) Regression Analysis and Its Applications. Dekker.

- **Análisis del Ejemplo**. Un análisis preliminar de los datos puede hacerse como sigue:

```
datos  <- read.csv("c:\\ ... \\EmisionesNOX.csv", header=TRUE)
attach(datos)
out <- lm(NITROX ~ HUMEDAD+TEMP+PRESION); summary(out)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.3362273  2.3048269  -3.183  0.00282 **
HUMEDAD     -0.0008559  0.0004614  -1.855  0.07096 .
TEMP         0.0012073  0.0016698   0.723  0.47388
PRESION      0.2803706  0.0792100   3.540  0.00103 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.0617 on 40 degrees of freedom
Multiple R-Squared: 0.6931,     Adjusted R-squared: 0.6701
F-statistic: 30.12 on 3 and 40 DF,  p-value: 2.371e-10
```

Parece que el nivel de emisiones no se afecta por la temperatura; el efecto de la presión barométrica es muy fuerte y la humedad del ambiente tiene un efecto marginal (¿habrá que incluír un término cuadrático?).

Ahora vemos el detalle de los cálculos mostrados arriba.

- **Ajuste del Modelo de Regresión**.
  - Coeficientes estimados, $\widehat{\beta} = (X^T X)^{-1} X^T y$

```
n  <- length(NITROX); p  <- 4
y  <- NITROX
X  <- cbind( rep(1,n), HUMEDAD, TEMP, PRESION )
b  <- solve( t(X) %*% X, t(X) %*% y )
round(as.vector(b),5)   #   -7.33623 -0.00086  0.00121  0.28037
```

- Error estándar, $\widehat{\sigma} = \sqrt{(y - X\widehat{\beta})^T (y - X\widehat{\beta})/(n-p)}$

  ```
  sig <- sqrt( sum((y - X%*%b)^2)/(n-p) )   #  sig = 0.06170448
  ```

- Errores estándar de los coeficientes, $\text{Var}(\widehat{\beta}) = \widehat{\sigma}^2 (X^T X)^{-1}$

  ```
  errstd <- sig*sqrt( diag(solve(t(X)%*%X)) ) # 2.30483 0.00046 0.00167 0.07921
  ```

- Valores de t: $t = (\widehat{\beta}_i - 0)/\sqrt{\widehat{\sigma}^2 (X^T X)^{-1}_{ii}}$

  ```
  tt <- b/errstd       # -3.183 -1.855  0.723  3.540
  ```

- $p-$valores: $2(\text{P}(t_{n-p} > |t|))$

  ```
  pval <- 2*(1-pt( abs(tt), n-p ))  # 0.00282 0.07096 0.47388 0.00103
  ```

- Coeficiente de determinación, R^2

  ```
  sct <- (n-1)*var(y); sce <- sum((y - X%*%b)^2)
  scr <- sct-sce; R2  <- scr/sct      #   R2 = 0.69313
  ```

- Coeficiente de determinación ajustado, R^2_a

  ```
  R2a <- 1-(1-R2)*(n-1)/(n-p)          #  R2a = 0.67012
  ```

- Valor del estadístico F: $F = (\text{SCR}/(p-1))/(\text{SCE}/(n-p))$

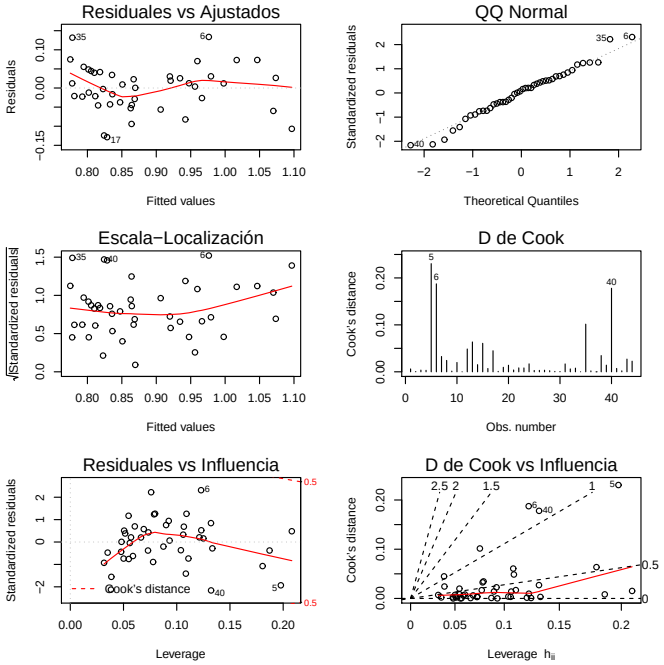
  ```
  FF <- (scr/(p-1))/(sce/(n-p))        #  FF = 30.117
  ```

- $p-$valor del estadístico F: $\text{P}(F^{p-1}_{n-p} > F)$

  ```
  pvalF <- 1-pf(FF,p-1,n-p)    #  2.371e-10
  ```

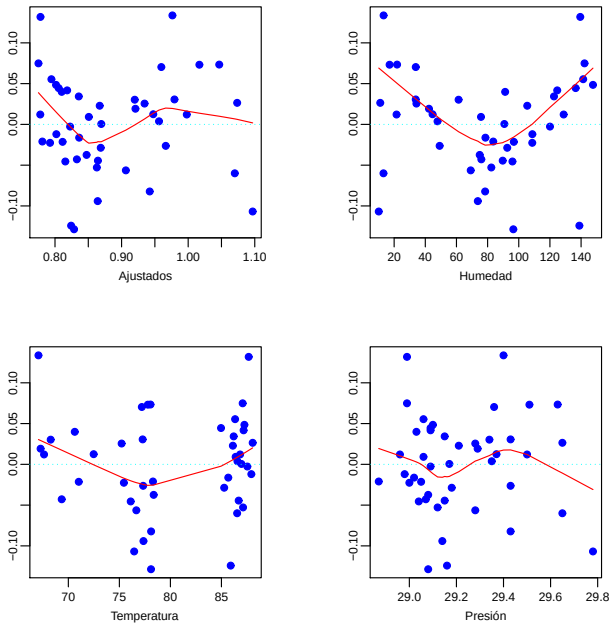
- Las gráficas de diagnóstico siguientes no muestran desviaciones fuertes de los supuestos.

```
attach(datos)
out <- lm(NITROX ~ HUMEDAD+TEMP+PRESION)
summary(out)
par(mfrow=c(3,2),mar=c(5, 5, 2, 2))
plot(out, which=1:6, caption=c(
 "Residuales vs Ajustados", "QQ Normal", "Escala-Localizacion",
 "D de Cook", "Residuales vs Influencia", "D de Cook vs Influencia"))
```



- Gráficas de residuales (vs valores ajustados y variables individuales).

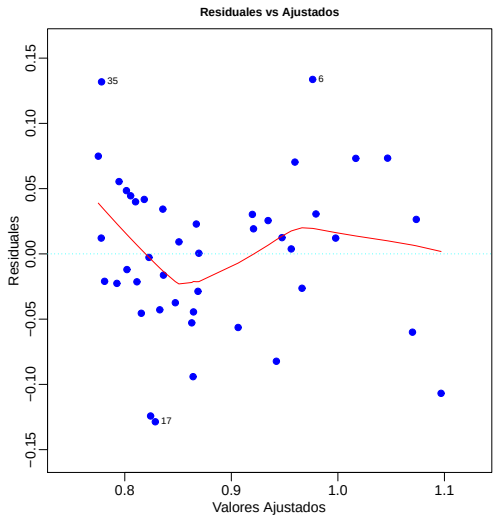
```
out <- lm(NITROX ~ HUMEDAD+TEMP+PRESION)

par(mfrow=c(2,2),mar=c(4, 4, 2, 2))
plot(out$fitted, out$res, mgp=c(1.5,.5,0), col="blue", cex.lab=.8,
 cex.axis=.8, xlab="Ajustados", ylab="", pch=19)
abline( h=0, col="cyan", lty=3 )
lines(lowess(out$fitted, out$res), col="red")

plot(HUMEDAD,out$res, mgp=c(1.5,.5,0), col="blue", cex.lab=.8,
 cex.axis=.8, xlab="Humedad", ylab="", pch=19)     # Termino cuadratico?
abline( h=0, col="cyan", lty=3 )
lines(lowess(HUMEDAD,out$res), col="red")

plot(TEMP,out$res, mgp=c(1.5,.5,0), col="blue", cex.lab=.8,
 cex.axis=.8, xlab="Temperatura", ylab="", pch=19)
abline( h=0, col="cyan", lty=3 )
lines(lowess(TEMP, out$res), col="red")

plot(PRESION,out$res, mgp=c(1.5,.5,0), col="blue", cex.lab=.8,
 cex.axis=.8, xlab="Presin", ylab="", pch=19)
abline( h=0, col="cyan", lty=3 )
lines(lowess(PRESION, out$res), col="red")
```

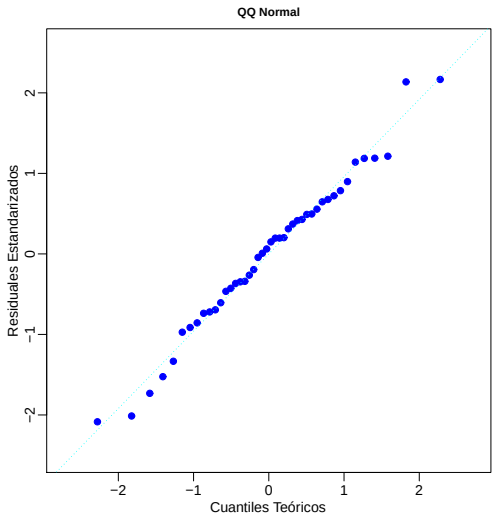


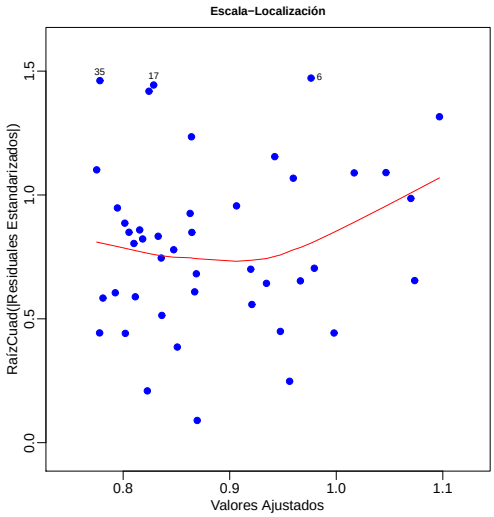
- Reproducción de las gráficas de diagnóstico producidas por `plot(out)`.

```
# Graficas de diagnostico: Datos de Emisiones
n   <- length(NITROX)
p   <- 4
y   <- NITROX
X   <- cbind( rep(1,n), HUMEDAD, TEMP, PRESION )
```

```
b   <- solve( t(X) %*% X, t(X) %*% y )
yg  <- X%*%b
res <- y - yg

par(mar=c(3, 3, 2, 2))
yr <- range(res); d  <- yr[2]-yr[1]; yr <- yr+.1*d*c(-1,1)
xr <- range(yg); d  <- xr[2]-xr[1]; xr <- xr+.1*d*c(-1,1)
plot(yg, res, ylim=yr, xlim=xr, pch=19, col="blue",
 ylab="Residuales", mgp=c(1.5,.5,0), xlab="Valores Ajustados",
 main="Residuales vs Ajustados", cex.main=.8)
abline( h=0, col="cyan", lty=3 )
lines(lowess(yg,res), col="red")
identify(yg, res, n=3, cex=.7)

# QQ-Plot de residuales estandarizados

rese <- as.vector(res/sig)
or   <- order(rese)
qteo <- qnorm(((1:n)-.5)/n)
yr <- range(rese); d  <- yr[2]-yr[1]; yr <- yr+.1*d*c(-1,1)
xr <- range(qteo); d  <- xr[2]-xr[1]; xr <- xr+.1*d*c(-1,1)
plot(qteo, rese[or], ylim=yr, xlim=xr, pch=19, col="blue",
 ylab="Residuales Estandarizados", mgp=c(1.5,.5,0), xlab="Cuantiles Teoricos",
 main="QQ Normal", cex.main=.8)
abline( lm(rese[or]~qteo), col="cyan", lty=3 )

# residuales estandarizados vs ajustados

yr <- range(sqrt(abs(rese))); d  <- yr[2]-yr[1]; yr <- yr+.1*d*c(-1,1)
xr <- range(yg); d  <- xr[2]-xr[1]; xr <- xr+.1*d*c(-1,1)
plot(yg, sqrt(abs(rese)), ylim=yr, xlim=xr, pch=19, col="blue",
 ylab="RazCuad(|Residuales Estandarizados|)", mgp=c(1.5,.5,0),
 xlab="Valores Ajustados",
 main="Escala-Localizacion", cex.main=.8)
lines(lowess(yg,sqrt(abs(rese))), col="red")
identify(yg,sqrt(abs(rese)), n=3, cex=.7)
```

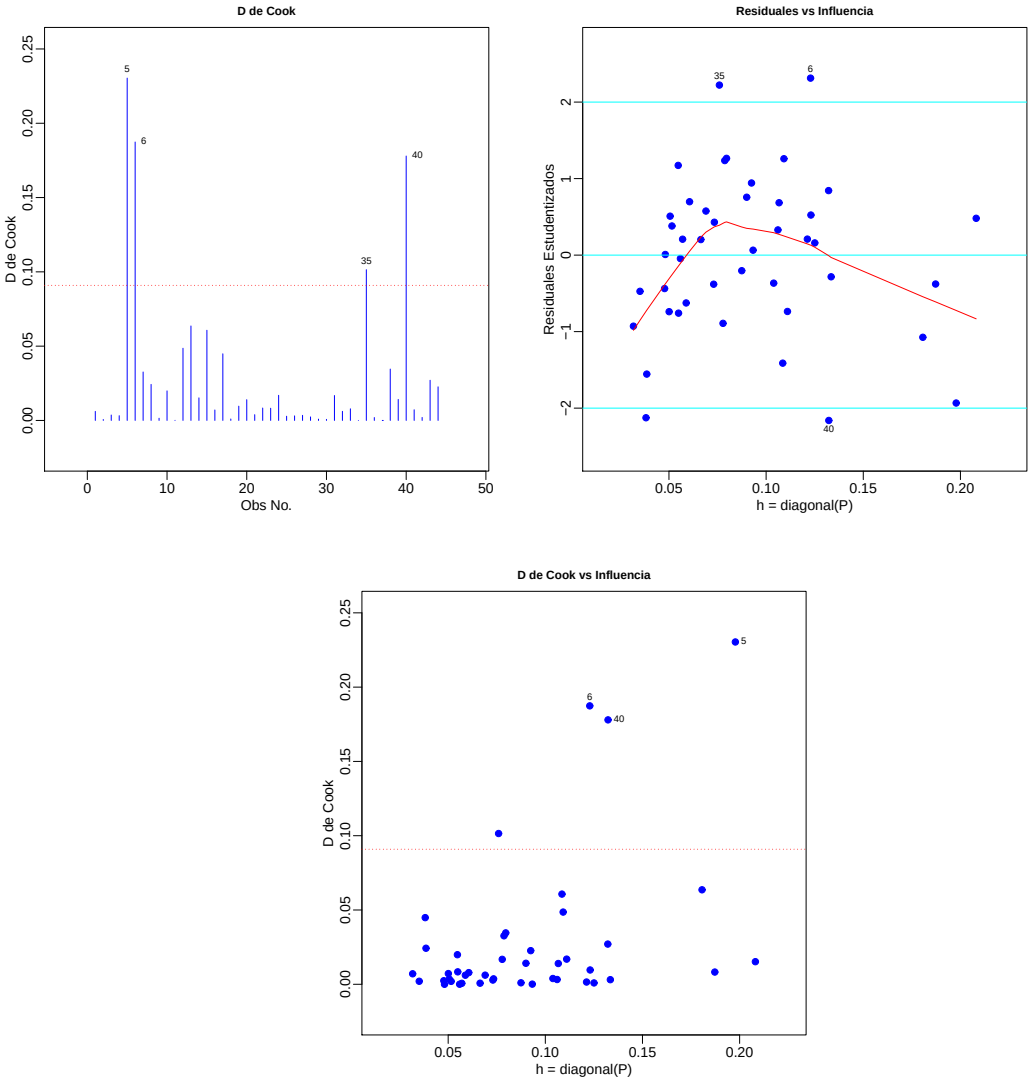






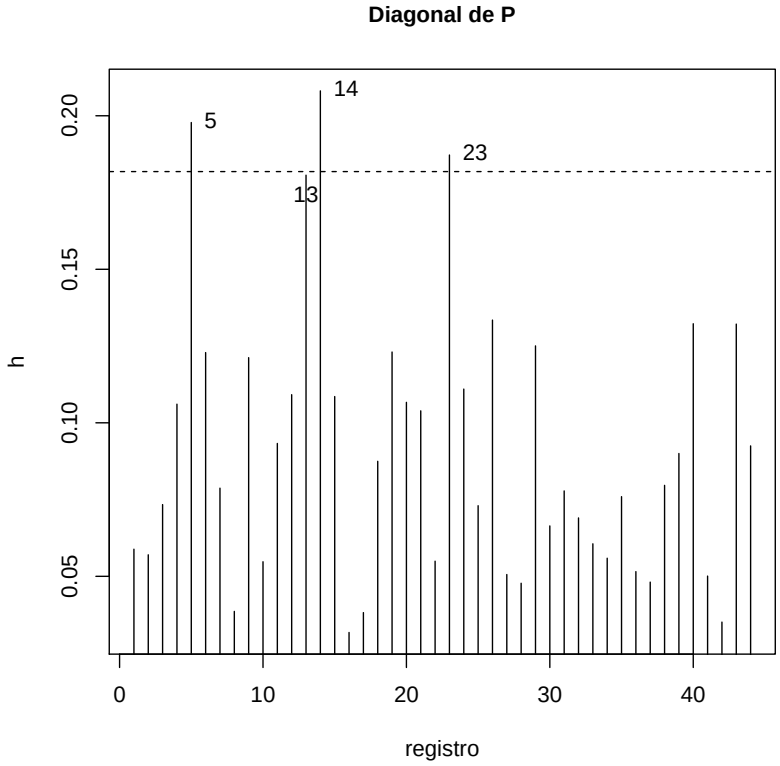
- Reproducción de las últimas tres gráficas de diagnóstico producidas por `plot(out)`.

```
X   <- cbind( rep(1,n), HUMEDAD, TEMP, PRESION )
H   <- solve( t(X)%*%X ); b   <- H %*% t(X) %*% y; ri  <- y - X%*%b
CME <- sum(ri^2)/(n-p)
P   <- X %*% H %*% t(X)
h   <- diag(P)
ris <- ri/sqrt(CME*(1-h))
Di  <- (ris)^2 * (h/(1-h)) / p

# Grafica D de Cook
yr <- range(Di); d  <- yr[2]-yr[1]; yr <- yr+.1*d*c(-1,1)
xr <- range(1:n); d  <- xr[2]-xr[1]; xr <- xr+.1*d*c(-1,1)
plot(1:n, Di, ylim=yr, xlim=xr, pch=19, col="blue",
 xlab="Obs No.", mgp=c(1.5,.5,0), ylab="D de Cook",
 main="D de Cook", cex.main=.8, type="h")
abline( h=4/n, col="red", lty=3 )
identify(1:n, Di, n=4, cex=.7, labels=1:n)

# Residuales vs "Leverage"
yr <- range(ris); d  <- yr[2]-yr[1]; yr <- yr+.1*d*c(-1,1)
xr <- range(h); d  <- xr[2]-xr[1]; xr <- xr+.1*d*c(-1,1)
plot(h, ris, ylim=yr, xlim=xr, pch=19, col="blue",
 xlab="h = diagonal(P)", mgp=c(1.5,.5,0), ylab="Residuales Estudentizados",
 main="Residuales vs Influencia", cex.main=.8, type="p")
abline(h=c(-2,0,2),col="cyan")
lines(lowess(h,ris), col="red")
identify(h, ris, n=3, cex=.7, labels=1:n)

# D de Cook vs "Leverage"
yr <- range(Di); d  <- yr[2]-yr[1]; yr <- yr+.1*d*c(-1,1)
xr <- range(h); d  <- xr[2]-xr[1]; xr <- xr+.1*d*c(-1,1)
plot(h, Di, ylim=yr, xlim=xr, pch=19, col="blue",
 xlab="h = diagonal(P)", mgp=c(1.5,.5,0), ylab="D de Cook",
 main="D de Cook vs Influencia", cex.main=.8, type="p")
abline(h=4/n, col="red", lty=3)
identify(h, Di, n=3, cex=.7, labels=1:n)
```



- Gráfica de elementos diagonales de matriz de proyección

```
# Calculo de indicadores de influencia
datos  <- read.csv(
"c:\\Documents and Settings\\...\\EmisionesNOX.csv", header=TRUE)
attach(datos)
n   <- length(NITROX)
p   <- 4
y   <- NITROX
X   <- cbind( rep(1,n), HUMEDAD, TEMP, PRESION )
A   <- solve( t(X) %*% X , t(X) )
P   <- X %*% A
h   <- diag(P)

plot(1:n,h,type="h",xlab="registro",ylab="h",main="Diagonal de P",cex.main=1)
abline(h=2*p/n,lty=2)
identify(1:n,h,n=4)
```



- Vimos que, si eliminamos la observación i, el efecto en los estimadores de los coeficientes de regresión se puede evaluar calculando

$$\widehat{\beta} - \widehat{\beta}_{(i)} = (X^T X)^{-1} x_i \frac{r_i}{1 - h_{ii}}$$

estas cantidades son las que se están calculando en `DFBETA`:

```
A      <- solve( t(X) %*% X , t(X) )
P      <- X %*% A
h      <- diag(P)
r      <- y - P%*%y
m      <- r/(1-h)
D      <- diag( as.vector(m) )
DFBETA <- A %*% D
```

- La medida anterior mide el impacto de la i-ésima observación, en todo el vector de coeficientes estimados. El siguiente indicador, `DFBETAS`, mide el impacto de la i-ésima observación sobre la j-ésima variable:

$$\mathsf{DFBETAS}_{ij} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{\sqrt{\mathsf{CME}_{(i)} (X^T X)^{-1}_{jj}}}$$

donde $\mathsf{CME}_{(i)}$ es el estimador de σ^2 que se obtiene del modelo sin la i-ésima observación. Puede verse que

$$(n - p - 1)\mathsf{CME}_{(i)} = (n - p)\mathsf{CME} - \frac{r_i^2}{1 - h_{ii}}$$

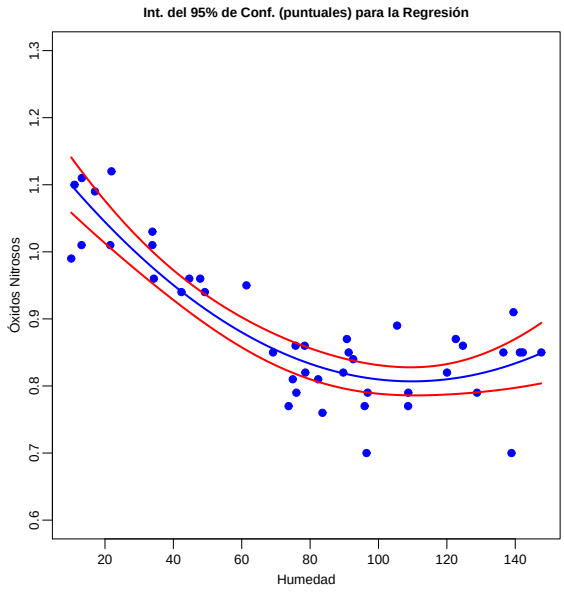
Un nivel crítico (empírico) para estos indicadores es $2/\sqrt{n}$. Enseguida presentamos estos cálculos.

```
# Estandarizacion de DFBETA
# Calculamos primero estimaciones de la varianza eliminando
# una observacion a la vez

cme     <- sum(r*r)/(n-p)
cmi     <- ( (n-p)*cme - (r*r)/(1-h) )/(n-1-p)
Sigi    <- diag( as.vector(sqrt( 1/cmi )) )
XXjj    <- sqrt(diag(1/diag( solve(t(X)%*%X) )))
DFBETAS <- XXjj %*% DFBETA %*% Sigi

# Observacion i-esima puede ser de influencia si |DFBETAS|>2/sqrt(n)

influ <- ifelse((abs(DFBETAS)>2/sqrt(n)),1,0)

# DFBETAS
      [,1] [,2] [,3] [,4]          [,1] [,2] [,3] [,4]
 [1,]    0    0    0    0    [23,]    0    0    0    0
 [2,]    0    0    0    0    [24,]    0    0    0    0
 [3,]    0    0    0    0    [25,]    0    0    0    0
 [4,]    0    0    0    0    [26,]    0    0    0    0
 [5,]    1    1    0    1    [27,]    0    0    0    0
 [6,]    0    0    1    0    [28,]    0    0    0    0
 [7,]    0    0    0    0    [29,]    0    0    0    0
 [8,]    0    0    0    0    [30,]    0    0    0    0
 [9,]    0    0    0    0    [31,]    0    0    0    0
[10,]    0    0    0    0    [32,]    0    0    0    0
[11,]    0    0    0    0    [33,]    0    0    0    0
[12,]    0    0    0    0    [34,]    0    0    0    0
[13,]    0    0    1    0    [35,]    0    0    0    0
[14,]    0    0    0    0    [36,]    0    0    0    0
[15,]    1    1    0    1    [37,]    0    0    0    0
[16,]    0    0    0    0    [38,]    0    0    0    0
[17,]    0    0    0    0    [39,]    0    0    0    0
[18,]    0    0    0    0    [40,]    1    1    0    1
[19,]    0    0    0    0    [41,]    0    0    0    0
[20,]    0    0    0    0    [42,]    0    0    0    0
[21,]    0    0    0    0    [43,]    0    0    0    0
[22,]    0    0    0    0    [44,]    0    0    0    0
```

- Finalizamos mostrando un modelo sencillo para niveles de óxido como una función cuadrática de la variable humedad.

```
# Intervalos de Confianza

y   <- NITROX
x   <- HUMEDAD
x2  <- HUMEDAD^2
n   <- length(y)
X   <- cbind(rep(1,n),x,x2)
b   <- solve(t(X)%*%X,t(X)%*%y)
cme <- sum((y-X%*%b)^2)/(n-3)
aa  <- cme*solve(t(X)%*%X)
xx  <- seq(min(x),max(x),length=200)
mm  <- b[1]+b[2]*xx+b[3]*xx^2
```

```
dd  <- rep(0,200)
for(i in 1:200){dd[i] <- sum(c(1,xx[i],xx[i]^2)*(aa%*%c(1,xx[i],xx[i]^2)))}
plot(x,y, col="blue", pch=19, mgp=c(1.5,.5,0), xlab="Humedad",
 ylab="Oxidos Nitrosos", cex.lab=.8, cex.axis=.8, cex.main=.8,
 main="Int. del 95% de Conf. (puntuales) para la Regresion",
 ylim=c(.6,1.3))
lines(xx,mm, lwd=2, col="blue")
lines(xx,mm+qt(.975,n-3)*sqrt(dd), lwd=2, col="red")
lines(xx,mm-qt(.975,n-3)*sqrt(dd), lwd=2, col="red")
```

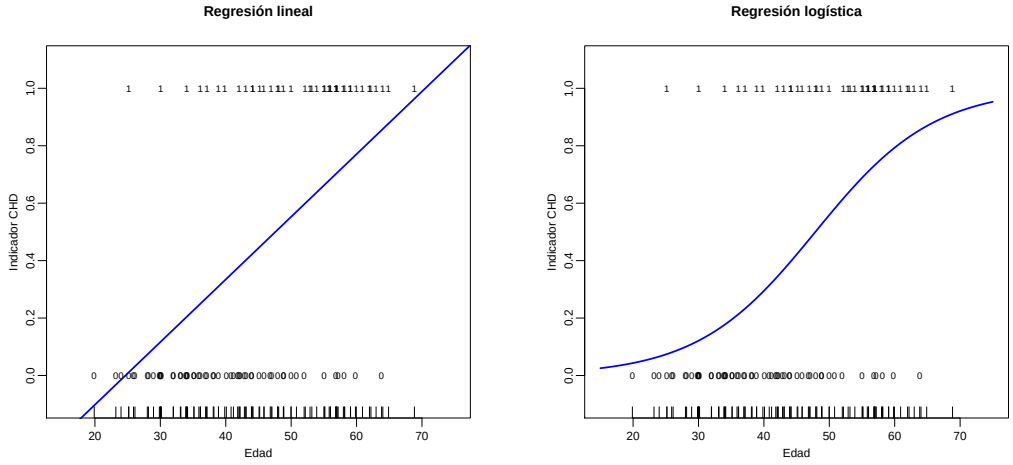


Resumen de Clase 15: Miércoles 23 de marzo

- Modelos de regresión con respuesta binaria. Los datos siguientes son edades (en años) e indicadores de presencia o ausencia de daño significativo en la coronaria de 100 individuos seleccionados para participar en el estudio.

edad	CHD	edad	CHD	edad	CHD	edad	CHD	edad	CHD	edad	CHD	edad	CHD	edad	CHD	edad	CHD	edad	CHD
20	0	30	0	34	0	37	0	41	0	44	1	48	1	53	1	57	0	60	0
23	0	30	0	34	0	37	1	42	0	44	1	48	1	53	1	57	1	60	1
24	0	30	0	34	1	37	0	42	0	45	0	49	0	54	1	57	1	61	1
25	0	30	0	34	0	38	0	42	0	45	1	49	0	55	0	57	1	62	1
25	1	30	0	34	0	38	0	42	1	46	0	49	1	55	1	57	1	62	1
26	0	30	1	35	0	39	0	43	0	46	1	50	0	55	1	58	0	63	1
26	0	32	0	35	0	39	1	43	0	47	0	50	1	56	1	58	1	64	0
28	0	32	0	36	0	40	0	43	1	47	0	51	0	56	1	58	1	64	1
28	0	33	0	36	1	40	1	44	0	47	1	52	0	56	1	59	1	65	1
29	0	33	0	36	0	41	0	44	0	48	0	52	1	57	0	59	1	69	1

Deseamos establecer una relación entre la edad de una persona y su propensión a padecer un problema en la coronaria. Las siguientes gráficas muestran dos posibles soluciones:



La cuestión aquí no es ¿cuál ajusta mejor? sino, ¿cuál es la más adecuada?.

- Regresión Logística. En regresión lineal modelamos el comportamiento medio de una variable de interés (variable de respuesta) como función de covariables

$$\mathsf{E}(y) = \beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k$$

en regresión logística (esto es, cuando la respuesta es binaria) también se modela la media como función de covariables

$$\mathsf{E}(y) = h(\beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k) = h(x^T\beta) \qquad \text{o, equivalentemente} \qquad g(\,\mathsf{E}(y)\,) = x^T\beta$$

Por ser y una variable binaria, entonces tenemos que sus dos posibles valores los toma con probabilidades

$$P(y=1) = p \qquad \text{y} \qquad P(y=0) = 1-p$$

sabemos que la media de una Bernoulli es E(y) $= p$. Una forma muy usada (aparte de que es sensata) de modelar la dependencia de p sobre covariables es mediante la función logit:

$$\text{logit}(p) \;=\; \log\left(\frac{p}{1-p}\right) \;=\; x^T\beta$$

o, equivalentemente (despejando para p), mediante la función logística (ésta es la que se muestra arriba a la derecha):

$$p = E(y|x) = \frac{1}{1+\exp(-x^T\beta)}$$

Este es el llamado Modelo de Regresión Logística.

- Problemas básicos en Estadística. En forma un tanto cuanto simplificada, muchos problemas en Estadística giran alrededor de responder las siguientes preguntas:
  - ¿Cómo modelo un fenómeno? (i.e. ¿Cómo parametrizo su comportamiento?)
  - ¿Cómo estimo, a partir de datos, los parámetros del modelo?
  - ¿Cómo valido ese modelo?
  - ¿Cómo uso ese modelo? (i.e. cómo me permite explicar un fenómeno, cómo descubre uno relaciones, cómo hago inferencias, cómo hago predicciones, etc.)

  Supongamos que tenemos observaciones independientes $(x_1^T, y_1), (x_2^T, y_2), \cdots, (x_n^T, y_n)$, sobre n individuos, donde y_i es una variable binaria (0/1) con 1 indicando la presencia de cierta característica de interés en el individuo i, y x_i es el correspondiente vector de covariables (o atributos). El modelo de regresión logística postula la relación entre la probabilidad de ocurrencia de un evento y los niveles de covariables (o variables predictoras):

$$p_i = P(y_i = 1 \mid x_i) = \frac{1}{1+\exp(-x_i^T\beta)}$$

  El método estándar para estimar los parámetros del modelo es **Máxima Verosimilitud**.

- Estimación de Parámetros. La verosimilitud de los datos independientes $(x_1^T, y_1), (x_2^T, y_2), \cdots, (x_n^T, y_n)$ es

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \qquad \text{con} \quad p_i = \frac{1}{1+\exp(-x_i^T\beta)}$$

  Estimamos β como aquel valor que maximiza $L(\beta)$, o equivalentemente, que maximiza a la logverosimilitud $l(\beta) = \log L(\beta)$

$$l(\beta) = \sum_{i=1}^{n} [\; y_i \log(p_i) + (1-y_i)\log(1-p_i) \;]$$

  Derivando la logverosimilitud:

$$\frac{\partial l(\beta)}{\partial \beta} \;=\; \sum_{i=1}^{n}\left[\; y_i\frac{1}{p_i}\frac{\partial p_i}{\partial \beta} - (1-y_i)\frac{1}{1-p_i}\frac{\partial p_i}{\partial \beta}\;\right] \;=\; \sum_{i=1}^{n}\left[\;\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i}\;\right]\frac{\partial p_i}{\partial \beta} \;=\; \sum_{i=1}^{n}\left[\;\frac{y_i - p_i}{p_i(1-p_i)}\;\right]\frac{\partial p_i}{\partial \beta}$$

  por otro lado,

$$\frac{\partial p_i}{\partial \beta} = -\left(\,1+e^{-x_i^T\beta}\,\right)^{-2} e^{-x_i^T\beta}(-x_i) \;=\; \frac{1}{1+e^{-x_i^T\beta}}\,\frac{e^{-x_i^T\beta}}{1+e^{-x_i^T\beta}}\,x_i \;=\; p_i(1-p_i)x_i$$

  entonces

$$\frac{\partial l(\beta)}{\partial \beta} \;=\; \sum_{i=1}^{n}\left[\;\frac{y_i - p_i}{p_i(1-p_i)}\;\right] p_i(1-p_i)x_i \;=\; \sum_{i=1}^{n}(y_i - p_i)\;x_i$$

Así, las ecuaciones de verosimilitud son:

$$\frac{\partial l(\beta)}{\partial \beta} \;=\; \sum_{i=1}^{n}(y_i - p_i)\, x_i \;=\; 0, \qquad \text{donde las } x_i\text{'s son } \; p \times 1$$

en general, este es un sistema de p ecuaciones nolineales en n incógnitas. El método de Newton es un procedimiento iterativo que puede ser útil para resolver este tipo de sistemas.

- Método de Newton. Las ecuaciones de verosimilitud son:

$$\frac{\partial l(\beta)}{\partial \beta} \;=\; \sum_{i=1}^{n}(y_i - p_i)\, x_i \;=\; 0$$

y, para el caso de una sola covariable, se ven como:

$$\frac{\partial l(\beta)}{\partial \beta} \;=\; \sum_{i=1}^{n}(y_i - p_i) \begin{bmatrix} 1 \\ z_i \end{bmatrix} \;=\; \begin{bmatrix} \sum(y_i - p_i) \\ \sum(y_i - p_i) z_i \end{bmatrix} \;\equiv\; \begin{bmatrix} h_1(\beta) \\ h_2(\beta) \end{bmatrix} \;\equiv\; H(\beta) \;=\; 0, \quad (\text{note que } H : \mathbb{R}^2 \to \mathbb{R}^2)$$

Para resolver $H(\beta) = 0$, hacemos una aproximación de primer orden para H alrededor de algún valor inicial razonable β_0:

$$H(\beta) \approx H(\beta_0) + \frac{\partial H(\beta_0)}{\partial \beta}(\beta - \beta_0), \qquad \text{aquí} \quad \frac{\partial H(\beta_0)}{\partial \beta} = \begin{bmatrix} \frac{\partial h_1(\beta_0)}{\partial \beta^T} \\ \frac{\partial h_2(\beta_0)}{\partial \beta^T} \end{bmatrix} \quad \text{es una matriz } 2 \times 2$$

entonces, en vez de resolver $H(\beta) = 0$, resolvemos el problema más fácil $H(\beta_0) + \frac{\partial H(\beta_0)}{\partial \beta}(\beta - \beta_0) = 0$. Así, despejando para β, obtenemos

$$\beta_1 = \beta_0 - \left[\frac{\partial H(\beta_0)}{\partial \beta} \right]^{-1} H(\beta_0)$$

ésta expresión la iteramos hasta convergencia

$$\beta_{k+1} = \beta_k - \left[\frac{\partial H(\beta_k)}{\partial \beta} \right]^{-1} H(\beta_k)$$

Ahora, las diferentes partes que intervienen en la expresión anterior son:

$$\begin{aligned} H(\beta) \;&=\; \begin{bmatrix} h_1(\beta) \\ h_2(\beta) \end{bmatrix} = \begin{bmatrix} \sum(y_i - p_i) \\ \sum(y_i - p_i) z_i \end{bmatrix} \\ \frac{\partial H(\beta)}{\partial \beta} \;&=\; \begin{bmatrix} \frac{\partial h_1(\beta)}{\partial \beta^T} \\ \frac{\partial h_2(\beta)}{\partial \beta^T} \end{bmatrix} \;=\; \begin{bmatrix} -\sum \frac{\partial p_i}{\partial \beta} \\ -\sum \frac{\partial p_i}{\partial \beta} z_i \end{bmatrix} \;=\; \begin{bmatrix} -\sum p_i(1-p_i) x_i^T \\ -\sum p_i(1-p_i) z_i x_i^T \end{bmatrix} \\ \frac{\partial H(\beta)}{\partial \beta} \;&=\; -\sum_{i=1}^{n} p_i(1-p_i) \begin{bmatrix} 1 \\ z_i \end{bmatrix} x_i^T \;=\; -\sum_{i=1}^{n} p_i(1-p_i) x_i x_i^T \end{aligned}$$

Reescribimos esto en notación matricial (más manejable para R). Definamos las matrices X y W de tamaños $n \times p$ y $n \times n$, respectivamente y los vectores y y p, $n \times 1$:

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \quad W = \begin{bmatrix} p_1(1-p_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_n(1-p_n) \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \text{y} \quad p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix}$$

con esto, es fácil ver que

$$\beta_{k+1} \;=\; \beta_k - \left[\frac{\partial H(\beta_k)}{\partial \beta} \right]^{-1} H(\beta_k) \;=\; \beta_k + (X^T W X)^{-1} X^T (y - p)$$

En la literatura de Modelos Lineales Generalizados encontrarán que la estimación se hace comunmente mediante "Mínimos Cuadrados Ponderados Iterativamente" (Iteratively Weighted Least Squares); éste método es precisamente el anterior que acabamos de ver. Si definimos el vector de "observaciones de trabajo"

$$\tilde{y} = X\beta_k + W^{-1}(y - p)$$

entonces, el Método de Newton es

$$\begin{aligned}\beta_{k+1} &= \beta_k + (X^T W X)^{-1} X^T (y - p) \;=\; (X^T W X)^{-1} X^T W X \beta_k + (X^T W X)^{-1} X^T W W^{-1} (y - p) \\ &= (X^T W X)^{-1} X^T W (\, X\beta_k + W^{-1}(y - p)\,) \;=\; (X^T W X)^{-1} X^T W \tilde{y}\end{aligned}$$

esto es $\beta_{k+1} = (X^T W X)^{-1} X^T W \tilde{y}$, y de aquí es de donde le viene el nombre de "mínimos cuadrados ponderados".

- Método de Newton en R. A continuación está el código en R para ajustar un modelo de regresión logística para los datos de CHD y edad.

```
# Hosmer, D.W. & Lemeshow, S.(1989) Applied logistic regression. Wiley
# Edad y Coronaria (danio significativo en coronaria)

edad <- c(
 20, 23, 24, 25, 25, 26, 26, 28, 28, 29, 30, 30, 30, 30, 30, 30, 32, 32, 33, 33,
 34, 34, 34, 34, 34, 35, 35, 36, 36, 36, 37, 37, 37, 38, 38, 39, 39, 40, 40, 41,
 41, 42, 42, 42, 42, 43, 43, 43, 44, 44, 44, 44, 45, 45, 46, 46, 47, 47, 47, 48,
 48, 48, 49, 49, 49, 50, 50, 51, 52, 52, 53, 53, 54, 55, 55, 55, 56, 56, 56, 57,
 57, 57, 57, 57, 57, 58, 58, 58, 59, 59, 60, 60, 61, 62, 62, 63, 64, 64, 65, 69)

coro <- c(
 0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,
 0,0,1,0,0,0,0,0,1,0,0,1,0,0,0,0,1,0,1,0,
 0,0,0,0,1,0,0,1,0,0,1,1,0,1,0,1,0,0,1,0,
 1,1,0,0,1,0,1,0,0,1,1,1,1,0,1,1,1,1,1,0,
 0,1,1,1,1,0,1,1,1,1,0,1,1,1,1,1,0,1,1,1)

# Grafica de los datos
edaj <- jitter(edad)              # solo con fines de graficacion

plot(edaj,coro,xlab="Edad",ylab="Indicador CHD",ylim=c(-.1,1.1),
 mgp=c(1.5,.5,0),cex.axis=.8,cex.lab=.8,cex.main=1,xlim=c(15,75),cex=.7,
 main="Regresion lineal",pch=ifelse(coro==1,"1","0"))
rug(edaj)
out <- lm(coro ~ edad)
abline(out,lwd=2,col="blue")      # esta es la recta de regresion

plot(edaj,coro,xlab="Edad",ylab="Indicador CHD",ylim=c(-.1,1.1),
 mgp=c(1.5,.5,0),cex.axis=.8,cex.lab=.8,cex.main=1,xlim=c(15,75),cex=.7,
 main="Regresion logistica",pch=ifelse(coro==1,"1","0"))
rug(edaj)

# Resolviendo ecuaciones de verosimilitud. Metodo de Newton.
y      <- coro
n      <- length(y)
X      <- cbind(rep(1,n),edad)
b      <- c(-10,.2)               # valores iniciales
```

```
# Las 4 lineas anteriores son especificas para los datos de enfermedades
# del corazon. Las siguientes son generales y se aplican para
# cualquier otra y y X's (incluso X con mas variables predictoras)

tolm   <- 1e-6      # tolerancia (norma minima de delta)
iterm  <- 100       # numero maximo de iteraciones
tolera <- 1         # inicializar tolera
itera  <- 0         # inicializar itera
histo  <- b         # inicializar historial de iteraciones

while( (tolera>tolm)&(itera<iterm) ){
 p      <- 1/( 1+exp( -as.vector(X%*%b) ) )
 W      <- p*(1-p)
 delta  <- as.vector( solve(t(X*W)%*%X, t(X)%*%(y-p)) )
 b      <- b + delta
 tolera <- sqrt( sum(delta*delta) )
 histo  <- rbind(histo,b)
 itera  <- itera + 1 }

#             [,1]       [,2]
# histo -10.000000 0.20000000
# b      -1.497206 0.03767488
# b      -4.380358 0.09253679
# b      -5.221685 0.10918224
# b      -5.308597 0.11090419
# b      -5.309453 0.11092114
# b      -5.309453 0.11092114

# Agregamos curva logistica a la grafica original
xx <- seq(15,75,length=200)
X  <- cbind(rep(1,n),xx)
p  <- 1/( 1+exp( -as.vector(X%*%b) ) )

lines(xx,p,lwd=2,col="blue")
```

- Interpretación de los Coeficientes del Modelo. En regresión lineal, por ejemplo

$$E(y \mid z_1) = \beta_0 + \beta_1 z_1$$

los coeficientes del modelo tienen una interpretación directa. Si los atributos, z_1, de un individuo se incrementan en una unidad, esto es, de z_1 pasa a $z_1 + 1$, entonces el impacto en la media esta dado por β_1, pues

$$E(y \mid z_1 + 1) - E(y \mid z_1) \;=\; (\beta_0 + \beta_1[z_1 + 1]) - (\beta_0 + \beta_1 z_1) \;=\; \beta_1$$

En regresión logística sucede algo semejante, pero el impacto no es en la media sino en la diferencia de logits.

$$\text{logit}(p) = \text{logit}[\, P(y = 1 \mid z_1)\,] = \beta_0 + \beta_1 z_1$$

entonces

$$\text{logit}[\, P(y = 1 \mid z_1 + 1)\,] - \text{logit}[\, P(y = 1 \mid z_1)\,] \;=\; \beta_1$$

y resulta que esta diferencia de logits tiene una interpretación importante

$$\text{logit}[\, P(y = 1 \mid z_1 + 1)\,] - \text{logit}[\, P(y = 1 \mid z_1)\,] \;=\; \log \frac{\text{Momios de } y = 1 \text{ con } z_1 + 1}{\text{Momios de } y = 1 \text{ con } z_1} \;=\; \beta_1$$

Los momios de que ocurra un evento A se definen como $P(A)/[\ 1 - P(A)\ ]$. La tasa de momios ("odds ratio") compara los momios de un evento bajo dos escenarios (el base z_1 versus cuando la variable aumenta $z_1 + 1$). La tasa de momios, para propósitos prácticos, se interpreta como la comparación de dos probabilidades; por ejemplo si la tasa de momios es igual a 2 entonces decimos que aumentar z_1 en una unidad incrementa al doble la probabilidad de ocurrencia del evento con respecto al escenario base. Entonces, la relación entre la tasa de momios y los coeficientes del modelo es, en este caso:

$$\frac{\text{Momios de } y = 1 \text{ con } z_1 + 1}{\text{Momios de } y = 1 \text{ con } z_1} \ = \ e^{\beta_1}$$

Por ejemplo, en el estudio sobre enfermedades coronarias tenemos $\widehat{\beta} = 0.1109$, entonces si comparamos dos individuos con 10 años de diferencia tenemos

$$\frac{\text{Momios de } y = 1 \text{ con } z_1 + 10}{\text{Momios de } y = 1 \text{ con } z_1} \ = \ e^{10\beta_1} \ = \ e^{1.109} \ = \ 3.03$$

de aquí que si dos personas tienen 10 años de diferencia, entonces el riesgo de tener problemas en la coronaria para la persona mayor es 3 veces más alto que el riesgo que tiene la persona menor. Esta interpretación supone que los logits se expresan linealmente en términos de las covariables **en toda la escala de las covariables**; esto pudiera no ser realista pues la comparación de los riesgos entre dos personas de 20 y 30 años podría no ser tan drástica como la comparación de los riesgos entre personas de 50 y 60 años. Esto lleva a un problema práctico el cual se podría enfrentar, por ejemplo, con transformaciones de la covariable, digamos log(edad).

Las tasas de momios tienen una propiedad importante en aplicaciones: Son invariantes con respecto al tipo de estudio (prospectivo o retrospectivo, i.e. estudios de cohortes prospectivos o estudios casos-control) (o sobresobresimplificando: "caros y lentos" versus "baratos y rápidos"). Esta propiedad de invarianza se traduce en un mayor atractivo de los modelos de regresión logística.

Resumen de Clase 16: Viernes 25 de marzo

- Errores Estándar. Una forma aproximada de calcular la matriz de varianzas y covarianzas del estimador de máxima verosimilitud es mediante

$$V(\widehat{\beta}) \doteq \left[-\frac{\partial^2 l(\widehat{\beta})}{\partial\beta\partial\beta^T} \right]^{-1}$$

esto es, la varianza (asintótica) de $\widehat{\beta}$ se aproxima por el inverso de la Matriz observada de Información. Una forma intuitiva de entender esto es: La logverosimilitud es globalmente cóncava (para los modelos lineales generalizados), por lo tanto su segunda derivada es negativa, además, la segunda derivada mide el grado de curvatura de la logverosimilitud. Mientras más grande sea la segunda derivada, más picuda es la logverosimilitud y, por lo tanto, mejor definido está su máximo. Entonces, si tomamos el recíproco del negativo de la curvatura tenemos una medida de que tan mal está nuestro estimador, (i.e. que tanta varianza tiene), a mayor curvatura menor varianza.

Para el caso de regresión logística vimos que

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta} &= H(\beta) = \sum_{i=1}^{n} (y_i - p_i) x_i \\ \frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} &= \frac{\partial H(\beta)}{\partial \beta^T} = -\sum_{i=1}^{n} p_i(1-p_i) x_i x_i^T = -X^T W X \end{aligned}$$

así que, la varianza se puede calcular como:

$$V(\widehat{\beta}) = \left[X^T W X \right]^{-1}$$

Los errores estándar de los estimadores se obtienen sacando raíz cuadrada a los elementos diagonales de V. En R, esto se calcula como:

```
# (esta es una continuacion del ejemplo que empezamos a ver la clase pasada)
y  <- coro
n  <- length(y)
X  <- cbind(rep(1,n),edad)

# nuevamente, las lineas anteriores son especificas para los datos CHD
# pero el calculo siguiente es general:

p  <- 1/( 1+exp( -as.vector(X%*%b) ) )      # (b se calculo en clase pasada)
W  <- p*(1-p)
V  <- solve( t(X*W)%*%X )
es <- sqrt( diag(V) )

# estimadores y sus desviaciones estandar
cbind(b,es)
               b         es
     -5.3094534 1.13365464
edad  0.1109211 0.02405984
```

- Ajuste del Modelo de Regresión Logística usando `glm()`. Nos pudimos haber ahorrado los cálculos anteriores pues R tiene funciones que hacen el trabajo por nosotros, las primeras dos líneas siguientes ajustan el modelo de regresión logística.

```
out <- glm(y ~ edad, family=binomial)
summary(out)

# Call:
# glm(formula = y ~ edad, family = binomial)

# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -1.9718  -0.8456  -0.4576   0.8253   2.2859

# Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
# edad         0.11092    0.02406   4.610 4.02e-06 ***
# ---
# Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

# (Dispersion parameter for binomial family taken to be 1)

#     Null deviance: 136.66  on 99  degrees of freedom
# Residual deviance: 107.35  on 98  degrees of freedom
# AIC: 111.35

# Number of Fisher Scoring iterations: 4
```

- Predicción y Clasificación. Los modelos de regresión logística contribuyen a:
  - Identificar factores de riesgo.
  - Evaluar el riego (probabilidad de ocurrencia del evento respuesta) para individuos específicos.
  - Clasificar / discriminar a grupos de individuos como alto riesgo / bajo riesgo.

  El proceso de identificación de factores de riesgo se da en forma natural (o es equivalente a) el proceso de selección de variables en el modelo de regresión logística. La formulación y estimación del modelo

  $$P(y = 1 \mid x) = \frac{1}{1 + e^{-x^T \beta}}$$

  nos dá una fórmula que nos permite evaluar el riesgo de un individuo específico (i.e. para un vector, x, de atributos específicos). Finalmente, esta misma regla la podemos aplicar a toda una población para formar grupos de riesgo; claro, bajo el supuesto de que los datos usados para formular el modelo son representativos de la población de interés (esto se dice fácil pero típicamente aquí se va mucho del esfuerzo correspondiente a un estudio y requiere de mucho input del médico responsable; esto es, la definición de la población, el diseño de muestreo, que atributos registrar, longitudinal o de sección cruzada, etc.).

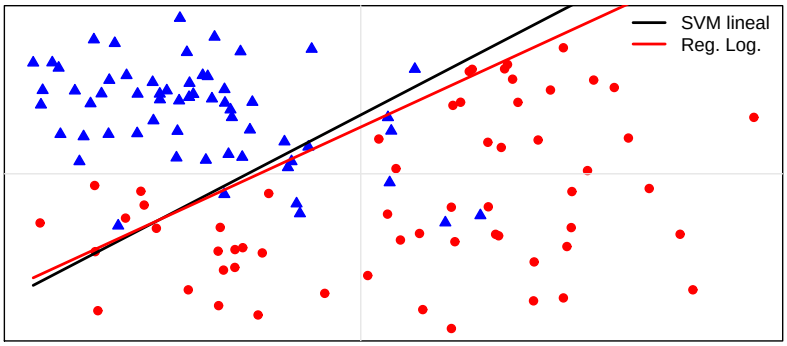
  Si deseamos producir una regla discriminante, podemos decir: Clasificamos a un individuo como de alto riesgo si la probabilidad de ocurrencia es mayor o igual a cierto umbral (definido por el especialista responsable del estudio); entonces, diremos que un individuo con atributos x es de alto riesgo si

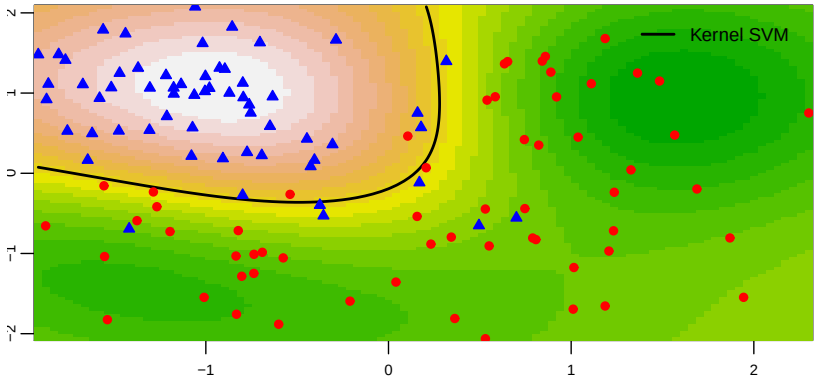
  $$\frac{1}{1 + e^{-x^T \beta}} > p$$

  esto es, si

  $$x^T \beta > \log \frac{p}{1 - p}$$

En el caso particular de $p = 0.5$ tenemos que los individuos de alto riesgo satisfacen $x^T\beta > 0$. La gráfica siguiente muestra el comportamiento de regresión logística, como función discriminante, comparada con el clasificador SVM. Regresión logística no sale bien librada, sin embargo no recomendamos la aplicación de este tipo de clasificadores en aplicaciones biomédicas pues se pierde la forma de como identificar los factores ded riesgo (aunque esto se puede salvar, pero no lo discutiremos aquí).

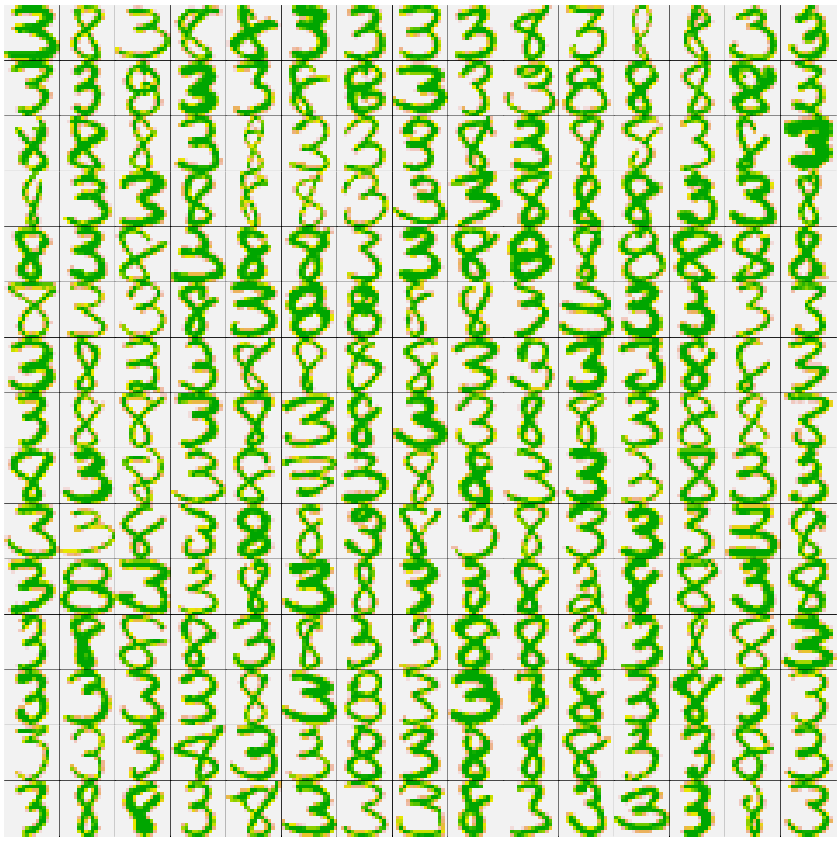




- Ejemplo de Problema de Clasificación. Los datos del siguiente ejemplo consisten de digitalizaciones de dígitos manuscritos. Cada dígito es escaneado y estandarizado a un cuadro de 16×16 pixeles; de modo que la x correspondiente es un vector con 256 atributos (niveles de gris de cada pixel). En cada caso se cuenta con y, el valor verdadero del dígito escaneado. Queremos tener una regla discriminante entre, digamos, los dígitos 8 y 3, usando las ideas expuestas antes.

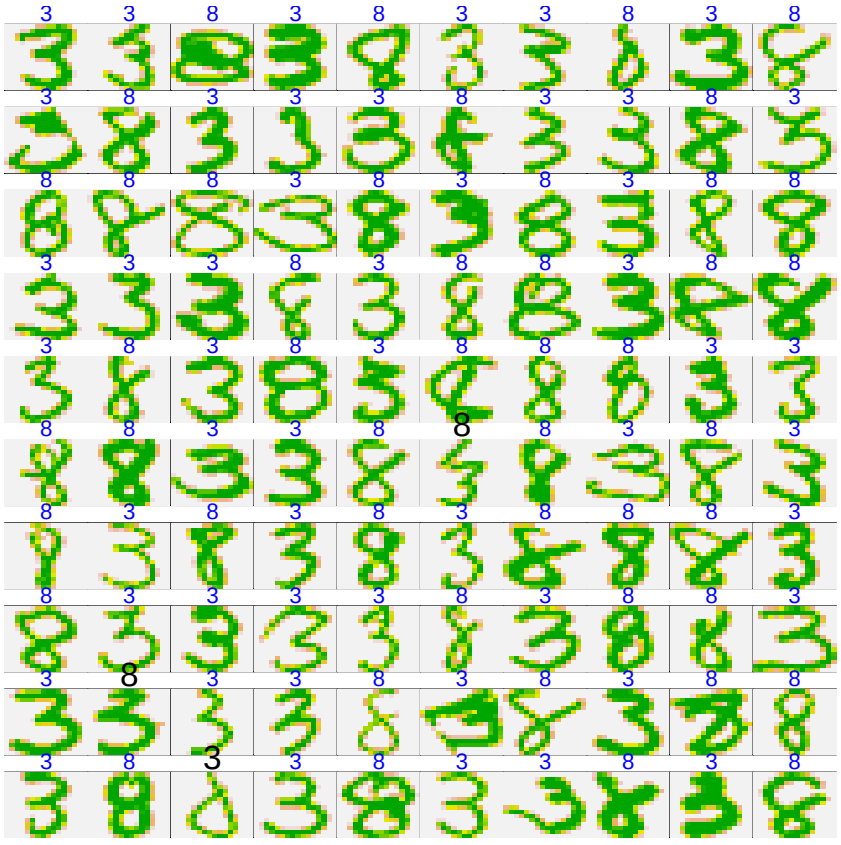
```
# 0123456789:  Clasificacion de digitos usando regresion logistica
# Leemos los datos con los que se hara el ajuste
datos <- scan(
 "C:\\Documents and Settings\\...\\digitrain.txt")
dato  <- matrix(datos,ncol=257,byrow=T)            # 7291 x 257
base  <- 8
vs    <- 3
dat   <- dato[ (dato[,1]==base)|(dato[,1]==vs) ,]  # 1200 x 257
y     <- ifelse(dat[,1]==base,0,1)
dd    <- as.data.frame(dat[,-1])
colnames(dd) <- paste("X",1:(dim(dd)[2]),sep="")

# Graficamos una muestra de los digitos
set.seed(64646)
```

```
n     <- dim(dd)[1]
M     <- 15
aa    <- as.matrix(dd[sample(1:n,size=(M^2)),])
par(mfrow=c(M,M),mar=c(0, 0, 0, 0))           # grafica de M^2 digitos
for(i in 1:(M^2)){
 bb  <- matrix(aa[i,],ncol=16,byrow=T)
 bb  <- bb[16:1,]
 image(-t(bb), cex.axis=.7, col=terrain.colors(20),mgp=c(1.5,.5,0),
       xlab="",ylab="",xaxt="n",yaxt="n")}
```



```
# Ajustamos un modelo de regresion logistica
# (no se alcanza convergencia con los defaults...)
# (pero no importa, seguir adelante))
out   <- glm(y ~ ., family=binomial, data=dd)

# Leemos los datos que usaremos para probar el modelo de clasificacion
datp  <- scan(
 "C:\\Documents and Settings\\...\\digitest.txt")
dap   <- matrix(datp,ncol=257,byrow=T)          # 2007 x 257
dp    <- dap[ (dap[,1]==base)|(dap[,1]==vs) ,]  # 332 x 257
yp    <- ifelse(dp[,1]==base,0,1)
d     <- as.data.frame(dp[,-1])
colnames(d) <- paste("X",1:(dim(d)[2]),sep="")

# Evaluamos la bondad de prediccion con los datos de prueba
pre   <- predict(out, newdata=d, type="response")
ypre  <- ifelse(round(pre)==0,base,vs)
```

```
yobs  <- ifelse(yp==0,base,vs)
pp    <- table(ypre,yobs)
    yobs
ypre   3   8
   3 152   9
   8  14 157
100*sum(diag(pp))/sum(pp)
93.07229   # porcentaje de aciertos con los datos de prueba

# Graficamos una muestra
nt   <- 100
selg <- sample(1:(dim(d)[1]),size=nt)
par(mfrow=c(sqrt(nt),sqrt(nt)),mar=c(0, 0, 1, 0))
for(i in 1:nt){
 bb  <- matrix(as.numeric(d[selg[i],]),ncol=16,byrow=T)# grafica de digitos de prueba
 bb  <- bb[16:1,]                # se resaltan los digitos mal clasif.
 image(-t(bb), cex.axis=.7, col=terrain.colors(20),mgp=c(1.5,.5,0),
       xlab="",ylab="",xaxt="n",yaxt="n")
 mtext(ypre[selg[i]], side = 3,
       col = ifelse(ypre[selg[i]]==yobs[selg[i]],"blue","black"),
       cex= ifelse(ypre[selg[i]]==yobs[selg[i]],1,1.5)) }
```



Resumen de Clase 17: Lunes 28 de marzo

- Codificación de Variables. Casi por regla general tendremos en un estudio variables en escala nominal (tratamiento, clasificación clínica, región de procedencia, religión, etc.). Para incluírlas en el modelo podemos construír variables auxiliares que sean manejables. Para una variable con r niveles necesitamos incluír $r-1$ variables auxiliares. Por ejemplo, si tenemos una variable que indica el tipo de zona en donde se encuentra la vivienda de los individuos del estudio y tenemos que los niveles son: zona rural, zona residencial urbana, zona urbana industrial; entonces podemos incluír dos variables de la siguiente forma

Zona	V_1	V_2
rural	0	0
urbana res.	1	0
urbana ind.	0	1

  Así, si la variable V_1 tiene un efecto significativo y si suponemos que su coeficiente estimado es igual a 2.7, entonces decimos que los momios de ocurrencia del evento ($y=1$) para los habitantes de la zona urbana residencial son $e^{2.7}$ veces los momios de ocurrencia del evento para los habitantes de la zona rural. De aquí que la elección del nivel base tendrá que ver con los intereses del estudio.

- Validación y comparación de modelos. Ilustraremos los conceptos con un ejemplo. Consideremos un estudio sobre pesos de recién nacidos. Se tienen 188 registros de nacimientos de los cuales 58 fueron bebés de peso bajo ($<$ 2.5 kgm).

```
     id low age lwt race smoke ptl ht ui ftv  bwt
1    10   1  29 130    1     0   0  0  1   2 1021
2    11   1  34 187    2     1   0  1  0   0 1135
3    13   1  25 105    3     0   1  1  0   0 1330
4    15   1  25  85    3     0   0  0  1   0 1474
5    16   1  27 150    3     0   0  0  0   0 1588
...
185 223   0  35 170    1     0   1  0  0   1 4174
186 224   0  19 120    1     1   0  0  0   0 4238
187 225   0  24 116    1     0   0  0  0   1 4593
188 226   0  45 123    1     0   0  0  0   1 4990
```

  Dicotomizamos la variable `bwt` haciendo `lbw=1` si `bwt < 2500` (la columna `bwt` es el peso, en gramos, al nacer). Consideraremos solo las columnas `age, lwt, race, ftv` como variables predictoras de `lbw`. A continuación está el ajuste de un primer modelo

```
lowbwt <- read.table("c:\\Documents and Settings\\My Documents\\lowbwt.dat",header=T)
names(lowbwt) <- c("id","low","age","lwt","race","smoke","ptl","ht","ui","ftv","bwt")
attach(lowbwt)
lbw    <- ifelse(bwt<2500,1,0)
r1     <- ifelse(race==2,1,0)
r2     <- ifelse(race==3,1,0)
out    <- glm(lbw ~ age+lwt+r1+r2+ftv, family=binomial)
summary(out)
# Resolviendo ecuaciones de verosimilitud
y      <- lbw
n      <- length(y)
X      <- cbind(rep(1,n),age,lwt,r1,r2,ftv)
b      <- c(1,0,0,1,.3,0)          # valores iniciales
```

```
tolm   <- 1e-6      # tolerancia (norma minima de delta)
iterm  <- 100       # numero maximo de iteraciones
tolera <- 1         # inicializar tolera
itera  <- 0         # inicializar itera
histo  <- b         # inicializar historial de iteraciones
while( (tolera>tolm)&(itera<iterm) ){
 p      <- 1/( 1+exp( -as.vector(X%*%b) ) )
 W      <- p*(1-p)
 delta  <- as.vector( solve(t(X*W)%*%X, t(X)%*%(y-p)) )
 b      <- b + delta                  # al final, b = estimadores Max.V.
 tolera <- sqrt( sum(delta*delta) )
 histo  <- rbind(histo,b)
 itera  <- itera + 1 }
# Calculo de errores estandar
y      <- lbw
n      <- length(y)
X      <- cbind(rep(1,n),age,lwt,r1,r2,ftv)
p      <- 1/( 1+exp( -as.vector(X%*%b) ) )
W      <- p*(1-p)
V      <- solve( t(X*W)%*%X )
es     <- sqrt( diag(V) )
cbind(b,es)
               b          es
     1.39579193 1.079401170
age -0.02870331 0.034131750
lwt -0.01422444 0.006558137
r1   0.99296316 0.498178969
r2   0.38498843 0.365125203
ftv -0.03457411 0.167568833
# (Igual que con glm)
```

- Una prueba global de ajuste. La herramienta básica para comparación de modelos es el **Cociente de Verosimilitudes**. Empezamos con una prueba global del ajuste de un modelo (semejante a la prueba F del análisis de varianza). El juego de hipótesis que queremos es

$$H_0 : \text{Los factores no afectan la respuesta} \qquad \text{versus} \qquad H_1 : \text{Si hay un efecto}$$

la hipótesis nula es equivalente a que $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.

El estadístico Cociente de Verosimilitudes es

$$\Lambda = \frac{\text{Verosimilitud optimizada bajo } H_0}{\text{Verosimilitud optimizada bajo el modelo completo}}$$

Una propiedad asintótica de este estadístico es que, bajo H_0: $G = -2 \log \Lambda \sim \chi^2_{gl}$, donde los grados de libertad, gl, son la diferencia entre el número de parámetros del modelo completo y el número de parámetros del modelo reducido, (en este caso $gl = 6 - 1 = 5$). Rechazamos la hipótesis nula si $G = -2 \log \Lambda > \chi^2_{5,\alpha}$. Observe que

$$G = -2 \log \Lambda \; = \; -2 \text{ logvero. bajo nula} - 2 \text{ logvero.}$$

Calculando esto en R:

```
# -2 loglik del modelo (residual deviance)
aa <- -2*(sum(y*log(p) + (1-y)*log(1-p)))                 # 220.3819
```

```
# -2 loglik del modelo nulo (null deviance)
bb <- -2*(sum(y)*log(mean(y))+(n-sum(y))*log(1-mean(y)))  # 232.3318

# -2 log(Cociente de verosimilitudes)
G  <- bb-aa                # 11.94995

pval <- 1-pchisq(G,df=5)   # 0.035 => rech Ho
```

De aquí que, con un nivel de significancia del 5%, podemos decir que las variables ayudan a explicar el peso bajo de los bebés.

- Pruebas sobre las variables individuales. Si queremos probar la significancia de, por ejemplo, la variable `ftv`, podemos hacerlo de dos formas; la primera, y más directa, es la prueba de Wald, en la que simplemente tomamos la estimación, la dividimos entre su error estándar y comparamos contra un cuantil de la normal estándar (o calculamos su p-valor).

```
z  <- b[6]/es[6]        # -0.2063278
2*(1-pnorm(abs(z)))     #  0.8365349
```

el p-valor es grande y, por lo tanto, el factor `ftv` no es significativo (esta prueba también la da `glm()` directamente).

La segunda forma consiste en aplicar la prueba de cociente de verosimilitudes, ajustando dos modelos, uno con todas las variables y el otro con todas las variables menos la que se quiere evaluar. Calculamos el estadśitico G y lo comparamos contra una χ^2_1. Explicamos este enfoque enseguida pues es un caso particular de comparación de modelos.

- Comparación de modelos. Viendo la salida de `glm()` tenemos que, aparentemente solo las variables `lwt` y raza son importantes (al menos en forma individual).

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.395792   1.079401   1.293   0.1960
age         -0.028703   0.034132  -0.841   0.4004
lwt         -0.014224   0.006558  -2.169   0.0301 *
r1           0.992963   0.498179   1.993   0.0462 *
r2           0.384988   0.365125   1.054   0.2917
ftv         -0.034574   0.167569  -0.206   0.8365
```

Así que nos gustaría considerar un modelo reducido con solo estas variables y contrastarlo con el modelo completo. Usando `glm()`, el valor de -2 veces la logverosimilitud del modelo bajo consideración se encuentra bajo el nombre `deviance`, así que basta con tomar la diferencia entre ellas (la del completo y el reducido) y compararla contra una ji-cuadrada de 2 grados de libertad

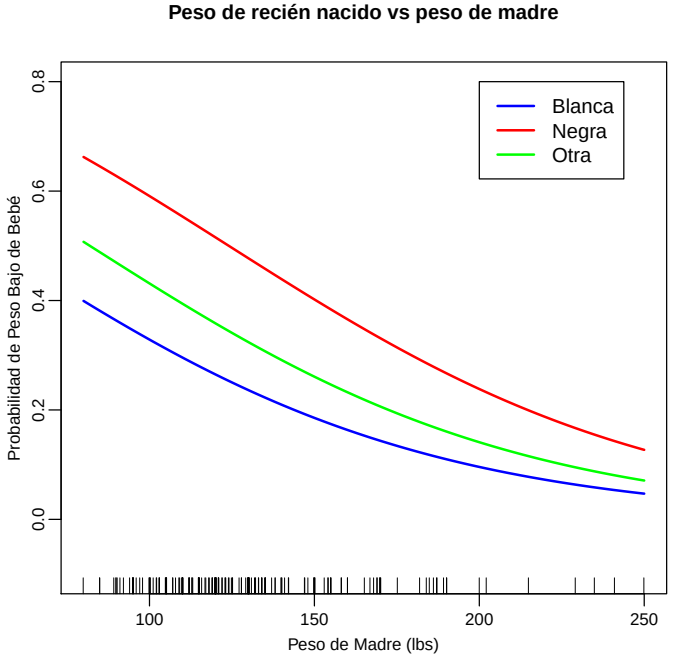
```
mcomp <- glm(lbw ~ age+lwt+r1+r2+ftv, family=binomial)
mred  <- glm(lbw ~ lwt+r1+r2, family=binomial)
G     <- mred$deviance - mcomp$deviance   # 0.8516
pval  <- 1-pchisq(G,df=2)                 # 0.6532467
```

El p-valor es grande, luego no rechazamos la hipótesis de que el modelo reducido es tan bueno como el modelo completo. Esto es, para explicar el peso bajo de bebés, un modelo basado en `lwt` y raza es tan bueno como uno que incluye las demás variables; por supuesto, un modelo parsimonioso es preferible a uno más complejo. El ajuste del modelo reducido aparece a continuación:

```
mred  <- glm(lbw ~ lwt+r1+r2, family=binomial)
summary(mred)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.815520   0.847607   0.962   0.3360
lwt         -0.015301   0.006461  -2.368   0.0179 *
r1           1.082278   0.488243   2.217   0.0266 *
r2           0.437738   0.359270   1.218   0.2231
```

En forma gráfica podemos visualizar este modelo como sigue:

```
br <- mred$coeff
xx <- seq(80,250,length=200)
X  <- cbind(rep(1,200),xx)
p  <- 1/( 1+exp( -as.vector(X%*%br[1:2]) ) )
p1 <- 1/( 1+exp( -as.vector(X%*%(br[1:2]+c(br[3],0))) ) )
p2 <- 1/( 1+exp( -as.vector(X%*%(br[1:2]+c(br[4],0))) ) )

plot(xx,p,xlab="Peso de Madre (lbs)",ylab="Probabilidad de Peso Bajo de Bebe",
 ylim=c(-.1,.8), mgp=c(1.5,.5,0),cex.axis=.8,cex.lab=.8,
 cex.main=1,xlim=c(80,250),cex=.7,lwd=2,col="blue",type="l",
 main="Peso de recien nacido vs peso de madre")
lines(xx,p1,lwd=2,col="red")
lines(xx,p2,lwd=2,col="green")
rug(jitter(lwt))
legend(200,.8,legend=c("Blanca","Negra","Otra"),lwd=2,
       col=c("blue","red","green"))
```



- Apéndice 1: Acerca de la tasa de momios. Vimos que los coeficientes de un modelo de regresión logística tienen interpretaciones en términos de tasas de momios. Haremos ahora una pequeña revisión de este tema. Los datos de la siguiente tabla provienen de uno de los primeros estudios sobre la asociación entre cáncer de pulmón y fumar.

	Cáncer	Controles
Fuma	688	650
No Fuma	21	59
	709	709

El estudio fue efectuado en 20 hospitales en Inglaterra; los controles fueron pacientes (sin cáncer) seleccionados del mismo sexo, mismos hospitales y aproximadamente de la misma edad que los pacientes con cáncer. La cantidad que es de interés es el **Riesgo Relativo**

$$RR = \frac{P(Can \mid Fum)}{P(Can \mid NoFum)}$$

sin embargo, para este estudio, estas cantidades no son estimables ¿por qué?.

Los **momios** de la ocurrencia de un evento A se definen como

$$\omega = \frac{P(A)}{1 - P(A)}$$

Así, si $A \equiv Can \mid Fum$, los momios de cáncer dado que la persona fuma, se definen como

$$\omega_1 = \frac{P(Can \mid Fum)}{1 - P(Can \mid Fum)} \equiv \frac{P(C \mid F)}{1 - P(C \mid F)}$$

y queremos comparar estos momios contra los momios de cáncer dado que la persona no fuma

$$\omega_2 = \frac{P(Can \mid NoFum)}{1 - P(Can \mid NoFum)} \equiv \frac{P(C \mid NF)}{1 - P(C \mid NF)}$$

para ello usamos la **tasa de momios**

$$\theta = \frac{\omega_1}{\omega_2} = \frac{P(C \mid F)[1 - P(C \mid NF)]}{P(C \mid NF)[1 - P(C \mid F)]}$$

Esta expresión para la tasa de momios aparentemente tiene el mismo problema de no estimabilidad de las probabilidades que la conforman; sin embargo, tenemos la siguiente relación:

$$\theta = \frac{\omega_1}{\omega_2} = \frac{P(C \mid F)[1 - P(C \mid NF)]}{P(C \mid NF)[1 - P(C \mid F)]} = \frac{P(F \mid C)[1 - P(F \mid NC)]}{P(F \mid NC)[1 - P(F \mid C)]}$$

las cuales **si** pueden ser estimadas del estudio retrospectivo.

$$\widehat{\theta} = \frac{[688/709]\ [59/709]}{[650/709]\ [21/709]} = \frac{688 \times 59}{650 \times 21} = 2.97$$

De aquí que **los momios de cáncer en fumadores son 3 veces más altos que los momios de cáncer en no fumadores**. En general, en una tabla 2×2, los momios se calculan como:

$$\widehat{\theta} = \frac{n_{11}\ n_{22}}{n_{12}\ n_{21}}$$

Nota técnica:

$$\theta = \frac{P(C \mid F)[1 - P(C \mid NF)]}{P(C \mid NF)[1 - P(C \mid F)]} = \frac{\frac{P(C,F)}{P(F)} \times \frac{P(NF) - P(C,NF)}{P(NF)}}{\frac{P(C,NF)}{P(NF)} \times \frac{P(F) - P(C,F)}{P(F)}}$$
$$= \frac{P(F,C)P(NF,NC)}{P(NF,C)P(F,NC)} = \frac{\frac{P(F,C)}{P(C)} \times \frac{P(NF,NC)}{P(NC)}}{\frac{P(NF,C)}{P(C)} \times \frac{P(F,NC)}{P(NC)}}$$
$$= \frac{P(F \mid C)}{P(NF \mid C)} \times \frac{P(NF \mid NC)}{P(F \mid NC)} = \frac{P(F \mid C)[1 - P(F \mid NC)]}{P(F \mid NC)[1 - P(F \mid C)]}$$

la cual es la relación que queríamos demostrar. Para obtener lo anterior, se usó la relación:

$$P(A) = P(A, B) + P(A, NB)$$

- Apéndice 2: Acerca de los errores estándar de las tasas de momios para datos multinomiales. La distribución multinomial es la base de muchos procedimientos para el análisis de datos categóricos. Daremos primero un resumen de algunas de sus propiedades.

  Consideremos un experimento en el que puede ocurrir alguna de c posibles categorías con probabilidades $\pi_1, \pi_2, \cdots, \pi_c$. Supongamos que efectuamos un total de n repeticiones independientes del experimento, si denotamos por y_i el total de ocurrencias de la i-ésima categoría, entonces

$$P(y_1 = n1, y_2 = n_2, \cdots, y_c = n_c) = \frac{n!}{n_1! n_2! \cdots n_c!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}$$

  donde $n_1 + n_2 + \cdots + n_c = n$ y $\pi_1 + \pi_2 + \cdots + \pi_c = 1$. El vector aleatorio $y = (y_1, y_2, \cdots, y_c)^T$ es una variable aleatoria **Multinomial**.

  En forma similar al caso de la Binomial, una variable multinomial se puede escribir como la suma de n variables "Bernoulli" independientes

$$y = w_1 + w_2 + \cdots + w_n$$

  donde $w_i = (y_{i1}, y_{i2}, \cdots, y_{ic})^T$, con $y_{ij} = 1$ si en el experimento i ocurrió la categoría j y $y_{ij} = 0$ si fue de otra forma. Note que

$$\mathsf{E}(w_i) = (\pi_1, \pi_2, \cdots, \pi_c)^T \equiv \pi$$

  además

$$\mathsf{Cov}(y_{ij}, y_{ik}) = \mathsf{E}(y_{ij} y_{ik}) - \pi_j \pi_k = \begin{cases} \pi_j(1 - \pi_j) & \text{si} \quad j = k \\ -\pi_j \pi_k & \text{si} \quad j \neq k \end{cases}$$

  De la expresión anterior tenemos

$$\mathsf{Var}(w_i) = \begin{bmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_c \\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & \cdots & -\pi_2\pi_c \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_c\pi_1 & -\pi_c\pi_2 & \cdots & \pi_c(1-\pi_c) \end{bmatrix} \equiv \Sigma$$

  Con estos resultados tenemos que la media y varianza de una variable multinomial están dadas por

$$\mathsf{E}(y) = n\pi \quad \text{y} \quad \mathsf{Var}(y) = n\Sigma$$

  Si los datos de un experimento multinomial son

$$(n_1, n_2, \cdots, n_c)^T$$

entonces, para estimar los parámetros, π_i's, del modelo multinomial, maximizamos la logverosimilitud, la cual es de la forma

$$l(\pi) = n_1 \log(\pi_1) + n_2 \log(\pi_2) + \cdots + n_c \log(\pi_c)$$

o, equivalentemente

$$l(\pi) = n_1 \log(\pi_1) + \cdots + n_{c-1} \log(\pi_{c-1}) + n_c \log\left(1 - \sum_{i=1}^{c-1} \pi_i\right)$$

derivando e igualando a 0, obtenemos

$$\frac{\partial l}{\partial \pi_k} = \frac{n_k}{\pi_k} - \frac{n_c}{\pi_c} = 0 \quad \Rightarrow \quad \pi_k = \pi_c \frac{n_k}{n_c}, \quad k = 1, \cdots, c-1$$

sumando estas expresiones:

$$\sum_{k=1}^{c-1} \pi_k = \pi_c \frac{1}{n_c} \sum_{k=1}^{c-1} n_k \quad \Rightarrow \quad 1 - \pi_c = \pi_c \frac{n - n_c}{n_c}$$

y de aquí obtenemos que $\widehat{\pi}_c = n_c/n$ y, también: $\widehat{\pi}_k = n_k/n$. Esto es, los estimadores de máxima verosimilitud para las probabilidades de ocurrencia, π_i, de las cateogorías, son las proporciones observadas n_i/n, lo cual era lo lógico de esperar.

Para obtener errores estándar, podemos usar las propiedades asintóticas del estimador de máxima verosimilitud, que nos dicen que la varianza es el inverso de la **Matriz de Información**:

$$\text{Var}(\widehat{\pi}) = \left[-\text{E}\left(\frac{\partial^2 l(\pi)}{\partial \pi \partial \pi^T}\right)\right]^{-1}$$

En realidad, la matriz de información, tal como esta escrita, es no invertible pues considera todas las entradas del vector $\widehat{\pi}$, el cual tiene entradas redundantes (suman 1). Sin embargo, si tomamos sólo la submatriz principal $(c-1) \times (c-1)$, puede verse que la varianza estimada del estimador de máxima verosimiltud es:

$$\text{Var}(\widehat{\pi}) = \frac{1}{n}\Sigma$$

(alternativamente, y más fácil, $\widehat{\pi} = y/n$ y de aquí también se sigue este resultado).

Supongamos ahora, muestreo multinomial en una tabla de contingencia 2×2:

n_{11}	n_{12}	
n_{11}	n_{12}	
		n

Tenemos que si $y = (n_{11}, n_{12}, n_{21}, n_{22})^T$ es un vector aleatorio multinomial, entonces $\text{E}(y) = n\pi$, donde $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})^T$ y además $\text{Var}(y) = n\Sigma$. En general, si quiero encontrar $\text{Var}[g(y)]$, entonces podemos usar el "método delta" que consiste en usar una aproximación de primer orden para g:

$$g(y) \approx g(\mu) + g'(\mu)^T (y - \mu)$$

entonces $\text{Var}[g(y)] \approx n g'(\mu)^T \Sigma g'(\mu)$. Consideremos el log de la tasa de momios

$$g(y) = \log(\widehat{\theta}) = \log n_{11} - \log n_{12} - \log n_{21} + \log n_{22}$$

$$g'(y) = (1/n_{11}, -1/n_{12}, -1/n_{21}, 1/n_{22})^T$$

evaluando en $\mu = n\pi = (n\pi_{11}, n\pi_{12}, n\pi_{21}, n\pi_{22})^T$:

$$g'(\mu) = (1/\pi_{11}, -1/\pi_{12}, -1/\pi_{21}, 1/\pi_{22})^T/n$$

de aquí que $\mathrm{Var}(\log\widehat{\theta})$ es aproximadamente

$$\frac{1}{n}\left(\frac{1}{\pi_{11}}, -\frac{1}{\pi_{12}}, -\frac{1}{\pi_{21}}, \frac{1}{\pi_{22}}\right)\begin{bmatrix} \pi_{11}(1-\pi_{11}) & -\pi_{11}\pi_{12} & -\pi_{11}\pi_{21} & -\pi_{11}\pi_{22} \\ -\pi_{12}\pi_{11} & \pi_{12}(1-\pi_{12}) & -\pi_{12}\pi_{21} & -\pi_{12}\pi_{22} \\ -\pi_{21}\pi_{11} & -\pi_{21}\pi_{12} & \pi_{21}(1-\pi_{21}) & -\pi_{21}\pi_{22} \\ -\pi_{22}\pi_{11} & -\pi_{22}\pi_{12} & -\pi_{22}\pi_{21} & \pi_{22}(1-\pi_{22}) \end{bmatrix}\begin{bmatrix} 1/\pi_{11} \\ -1/\pi_{12} \\ -1/\pi_{21} \\ 1/\pi_{22} \end{bmatrix}$$

$$\mathrm{Var}(\log\widehat{\theta}) \approx \frac{1}{n}(1,-1,-1,1)\begin{bmatrix} 1/\pi_{11} \\ -1/\pi_{12} \\ -1/\pi_{21} \\ 1/\pi_{22} \end{bmatrix} = \frac{1}{n}\left(\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}}\right)$$

Entonces estimamos la varianza del log(tasa de momios) mediante

$$\mathrm{Var}(\log\widehat{\theta}) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

Asintóticamente, la distribución de $\log\widehat{\theta}$ es normal, entonces un intervalo de confianza se obtiene mediante

$$\log\widehat{\theta} \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Tomando exponencial a los extremos de este intervalo obtenemos un correspondiente intervalo para θ. Así, por ejemplo, para los datos de cáncer, obtenemos intervalos de confianza del 95%:

$$0.58 < \log\theta < 1.56$$

y para la tasa de momios

$$1.79 < \theta < 4.95$$

así, podemos asegurar (con un 95% de confianza) que los momios de contraer cáncer entre fumadores son, al menos, 1.8 veces más grandes que los momios de cáncer entre no fumadores.

Resumen de Clase 18: Miércoles 30 de marzo

– Modelos Lineales Generalizados –

- Características del modelo Normal. Supongamos datos independientes $(x_1^T, y_1), \cdots, (x_n^T, y_n)$. El modelo de regresión Normal tiene tres partes:
  - La componente aleatoria: Las y_i's son variables aleatorias independientes, con distribución normal $N(\mu_i, \sigma^2)$.
  - La componente sistemática: El predictor lineal $\eta_i = x_i^T \beta$.
  - La liga entre ambas componentes: $\mu_i = x_i^T \beta$, o $g(\mu_i) = \eta_i$, donde g es la "liga identidad".

- Características del modelo Bernoulli. Supongamos datos independientes $(x_1^T, y_1), \cdots, (x_n^T, y_n)$. El modelo de regresión Logística tiene tres partes:
  - La componente aleatoria: Las y_i's son variables aleatorias independientes, con distribución Bernoulli $\mathsf{B}(p_i)$.
  - La componente sistemática: El predictor lineal $\eta_i = x_i^T \beta$.
  - La liga entre ambas componentes: $g(p_i) = x_i^T \beta$, donde $g(p_i) = \mathsf{logit}(p_i) = \log(\, p_i/(1-p_i)\,)$, a g se le llama "liga logit".

- Modelos Lineales Generalizados. Una generalización de los modelos anteriores sería tener datos independientes $(x_1^T, y_1), \cdots, (x_n^T, y_n)$, bajo un modelo con tres partes:
  - La componente aleatoria: Las y_i's son variables aleatorias independientes, con distribución perteneciente a la familia Exponencial:

$$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

  - La componente sistemática: El predictor lineal $\eta_i = x_i^T \beta$.
  - La liga entre ambas componentes: $g(\mu_i) = x_i^T \beta$, donde a g se le llama "función liga".

- La Normal pertenece a la familia Exponencial. Note que

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y-\mu)^2 \right\} = \exp\left\{ \frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \right\}$$

  de aquí que $\theta = \mu$, $b(\theta) = \frac{1}{2}\mu^2 = \frac{1}{2}\theta^2$, $a(\phi) = \phi = \sigma^2$ y $c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right)$.

- La Bernoulli pertenece a la familia Exponencial. Note que

$$f(y; p) = p^y (1-p)^{1-y} = \exp\{\, y \log p + (1-y)\log(1-p)\,\} = \exp\{\, y\,\mathsf{logit}\,(p) + \log(1-p)\,\}$$

  de aquí que $\theta = \mathsf{logit}\,(p)$, $b(\theta) = -\log(1-p) = \log(1+e^\theta)$, $a(\phi) = \phi = 1$ y $c(y, \phi) = 0$

- Algunos resultados sobre logverosimilitudes. Sea $f(y; \theta)$ cierta densidad o función de probabilidad. Sea $l(\theta) = \log f(y; \theta)$. Entonces

1. $\mathsf{E}\left(\,\partial l(\theta)\,/\,\partial\theta\,\right) = 0$. Prueba:

$$\begin{aligned}\mathsf{E}\left(\frac{\partial l(\theta)}{\partial\theta}\right) &= \int \frac{\partial l(\theta)}{\partial\theta} f(y;\theta)dy = \int \frac{\frac{\partial f(y;\theta)}{\partial\theta}}{f(y;\theta)} f(y;\theta)dy \\ &= \int \frac{\partial f(y;\theta)}{\partial\theta}dy = \frac{\partial}{\partial\theta}\int f(y;\theta)dy = \frac{\partial}{\partial\theta}(1) = 0\end{aligned}$$

2. $\mathsf{E}\left(\,\partial^2 l(\theta)\,/\,\partial\theta^2\,\right) + \mathsf{E}\left(\,\partial l(\theta)\,/\,\partial\theta\,\right)^2 = 0$. Prueba:

$$\frac{\partial^2 l}{\partial\theta^2} = \frac{\partial}{\partial\theta}\left(\frac{\partial l}{\partial\theta}\right) = \frac{\partial}{\partial\theta}\left(\frac{1}{f}\frac{\partial f}{\partial\theta}\right) = \frac{1}{f}\frac{\partial^2 f}{\partial\theta^2} + \frac{\partial f}{\partial\theta}\left(-\frac{1}{f^2}\frac{\partial f}{\partial\theta}\right) = \frac{1}{f^2}\left[f\frac{\partial^2 f}{\partial\theta^2} - \left(\frac{\partial f}{\partial\theta}\right)^2\right]$$

Entonces

$$\begin{aligned}\mathsf{E}\left(\frac{\partial^2 l(\theta)}{\partial\theta^2}\right) &= \int \frac{1}{f^2}\left[f\frac{\partial^2 f}{\partial\theta^2} - \left(\frac{\partial f}{\partial\theta}\right)^2\right] f\,dy = \int \frac{\partial^2 f}{\partial\theta^2}\,dy - \int \left(\frac{\partial f}{\partial\theta}\right)^2 \frac{1}{f}\,dy \\ &= \frac{\partial^2}{\partial\theta^2}\int f\,dy - \int \left(\frac{1}{f}\frac{\partial f}{\partial\theta}\right)^2 f\,dy = 0 - \int \left(\frac{\partial}{\partial\theta}\log f\right)^2 f\,dy = -\mathsf{E}\left(\frac{\partial l}{\partial\theta}\right)^2\end{aligned}$$

esto es

$$\mathsf{E}\left(\frac{\partial^2 l(\theta)}{\partial\theta^2}\right) + \mathsf{E}\left(\frac{\partial l(\theta)}{\partial\theta}\right)^2 = 0$$

- Media y varianza de la Exponencial. Para distribuciones en la familia Exponencial tenemos

$$f(y;\theta,\phi) = \exp\left\{\frac{y\,\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \qquad \text{y} \qquad l(\theta) = \frac{y\,\theta - b(\theta)}{a(\phi)} + c(y,\phi)$$

Usando la primera propiedad de arriba:

$$\frac{\partial l(\theta)}{\partial\theta} = \frac{y - b'(\theta)}{a(\phi)} \qquad \Rightarrow \qquad 0 = \mathsf{E}\left(\frac{\partial l(\theta)}{\partial\theta}\right) = \frac{\mu - b'(\theta)}{a(\phi)} \qquad \Rightarrow \qquad \boxed{\mathsf{E}(y) = \mu = b'(\theta)}$$

Usando la segunda propiedad:

$$\frac{\partial^2 l(\theta)}{\partial\theta^2} = -\frac{b''(\theta)}{a(\phi)} \quad \Rightarrow \quad \mathsf{E}\left(\frac{\partial l(\theta)}{\partial\theta}\right)^2 = -\mathsf{E}\left(\frac{\partial^2 l(\theta)}{\partial\theta^2}\right) = \frac{b''(\theta)}{a(\phi)}$$

pero, por otro lado:

$$\mathsf{E}\left(\frac{\partial l(\theta)}{\partial\theta}\right)^2 = \mathsf{E}\left(\frac{y-\mu}{a(\phi)}\right)^2 = \frac{1}{a^2(\phi)}\mathsf{Var}(y)$$

de aquí que

$$\boxed{\mathsf{Var}(y) = a(\phi)\,b''(\theta)}$$

- Funciones Liga. Las distribuciones de la familia Exponencial tienen:

$$f(y;\theta,\phi) = \exp\left\{\frac{y\,\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

donde θ es llamado "parámetro canónico" y ϕ es es el llamado "parámetro de dispersión", el cual, por el momento, será un parámetro de estorbo. Los modelos lineales generalizados asocian el predictor lineal $\eta = x^T\beta$ con la media $\mathsf{E}(y) = \mu$ mediante una función liga $g(\mu) = \eta$, donde g es monótona y diferenciable.

Note que $\mu = b'(\theta)$ así que μ es una función de θ, digamos $\mu = h(\theta)$, la "liga canónica", $g(\mu)$, se define como aquella función g tal que $\theta = \eta$ (esto ocurre cuando $g = h^{-1}$, pues $\eta = g(\mu) = g(h(\theta)) = h^{-1}(h(\theta)) = \theta$).

Una propiedad de la liga canónica es que asegura la existencia de un vector de estadísticos suficientes para β, tal vector de estadísticos suficientes con la misma dimensión que β. Es fácil ver que, al formar la verosimilitud, sustituyendo θ_i por $\eta_i = x_i^T\beta$, nos queda el término $y^T X\beta$, de donde $X^T y$ es el vector de estadísticos suficientes para β. En la práctica se ha encontrado que la liga canónica es razonable y que puede justificarse en el contexto de un problema (en otras palabras, la liga canónica tiene propiedades estadísticas buenas, como suficiencia, y es, en los casos usuales como Normal, Poisson y Bernoulli, bastante razonable desde el punto de vista de aplicaciones) (por supuesto, podemos usar otra función liga que creeamos sea más razonable).

- Estimación Máximo Verosímil. Supongamos datos independientes $(x_1^T, y_1), \cdots, (x_n^T, y_n)$, bajo un modelo de la familia Exponencial. La logverosimilitud es de la forma

$$l(\beta) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \equiv \sum_{i=1}^{n} l_i(\beta)$$

La estructura del procedimiento Newton-Raphson para la maximización de l es

$$\beta^{k+1} = \beta^k - \left[\frac{\partial^2 l(\beta^k)}{\partial\beta\partial^T\beta}\right]^{-1} \frac{\partial l(\beta^k)}{\partial\beta}$$

En general, el procedimiento usado en el área de GLM's es el Método de Scoring de Fisher:

$$\beta^{k+1} = \beta^k - \left[\ \mathsf{E}\left(\frac{\partial^2 l(\beta^k)}{\partial\beta\partial^T\beta}\right)\ \right]^{-1} \frac{\partial l(\beta^k)}{\partial\beta}$$

- Estructura del Gradiente.

$$\frac{\partial l(\beta)}{\partial\beta} = \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial\beta} = \sum_{i=1}^{n} \begin{bmatrix} \partial l_i(\beta)/\partial\beta_1 \\ \vdots \\ \partial l_i(\beta)/\partial\beta_p \end{bmatrix}$$

Ahora,

$$\frac{\partial l_i(\beta)}{\partial\beta_r} = \frac{\partial l_i(\beta)}{\partial\theta_i}\ \frac{\partial\theta_i}{\partial\mu_i}\ \frac{\partial\mu_i}{\partial\eta_i}\ \frac{\partial\eta_i}{\partial\beta_r} = \frac{\partial l_i(\beta)}{\partial\theta_i}\ \frac{1}{\partial\mu_i/\partial\theta_i}\ \frac{1}{\partial\eta_i/\partial\mu_i}\ \frac{\partial\eta_i}{\partial\beta_r} = \frac{y_i - \mu_i}{a(\phi)}\ \frac{1}{b''(\theta_i)}\ \frac{1}{g'(\mu_i)}\ x_{ir}$$

de aquí que

$$\frac{\partial l_i(\beta)}{\partial\beta_r} = \frac{y_i - \mu_i}{\mathsf{Var}(y_i)}\ \frac{1}{g'(\mu_i)}\ x_{ir}$$

entonces

$$\frac{\partial l(\beta)}{\partial\beta} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathsf{Var}(y_i)}\ \frac{1}{g'(\mu_i)} \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} = X^T DW(y - \mu)$$

donde $X^T = [x_1, \cdots, x_n]_{p\times n}$, $D = \mathsf{diag}(\ 1/g'(\mu_i)\ )$ y $W = \mathsf{diag}(\ 1/\mathsf{Var}(y_i)\ )$

Resumen de Clase 19: Lunes 4 de abril

- Método Scoring de Fisher en Modelos Lineales Generalizados. Supongamos datos independientes $(x_1^T, y_1), \cdots, (x_n^T, y_n)$, bajo un modelo de la familia Exponencial. La logverosimilitud es de la forma

$$l(\beta) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \equiv \sum_{i=1}^{n} l_i(\beta)$$

  El procedimiento de maximización Scoring de Fisher es un método iterativo definido por:

$$\beta^{k+1} = \beta^k - \left[\mathsf{E}\left(\frac{\partial^2 l(\beta^k)}{\partial\beta\partial\beta^T} \right) \right]^{-1} \frac{\partial l(\beta^k)}{\partial\beta}$$

- Estructura del Gradiente. Vimos en la clase pasada que

$$\frac{\partial l(\beta)}{\partial\beta} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathsf{Var}(y_i)} \frac{1}{g'(\mu_i)} x_i = X^T DW(y - \mu)$$

  donde $X^T = [x_1, \cdots, x_n]_{p\times n}$, $D = \mathsf{diag}(\ 1/g'(\mu_i)\ )$ y $W = \mathsf{diag}(\ 1/\mathsf{Var}(y_i)\ )$

- Estructura del Hessiano. La matriz de segundas derivadas de la logverosimilitud es:

$$\begin{aligned}
\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} &= \frac{\partial}{\partial\beta^T}\left(\frac{\partial l(\beta)}{\partial\beta}\right) = \frac{\partial}{\partial\beta^T}\left(\sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathsf{Var}(y_i)} \frac{1}{g'(\mu_i)} x_i\right) \\
&= \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial}{\partial\beta^T}\left(\frac{1}{\mathsf{Var}(y_i)} \frac{1}{g'(\mu_i)} x_i\right) + \sum_{i=1}^{n} \left(\frac{1}{\mathsf{Var}(y_i)} \frac{1}{g'(\mu_i)} x_i\right) \frac{\partial}{\partial\beta^T}(y_i - \mu_i)
\end{aligned}$$

  En el método Scoring de Fisher necesitamos el valor esperado de este hessiano, así que

$$\begin{aligned}
\mathsf{E}\left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}\right) &= -\sum_{i=1}^{n} \left(\frac{1}{\mathsf{Var}(y_i)} \frac{1}{g'(\mu_i)} x_i\right) \left(\frac{\partial\mu_i}{\partial\beta^T}\right) = -\sum_{i=1}^{n} \left(\frac{1}{\mathsf{Var}(y_i)} \frac{1}{g'(\mu_i)} x_i\right) \left(\frac{\partial\mu_i}{\partial\eta_i} \frac{\partial\eta_i}{\partial\beta^T}\right) \\
&= -\sum_{i=1}^{n} \left(\frac{1}{\mathsf{Var}(y_i)} \frac{1}{g'(\mu_i)} x_i\right) \left(\frac{1}{g'(\mu_i)} x_i^T\right) = -\sum_{i=1}^{n} x_i \left(\frac{1}{g'(\mu_i)} \frac{1}{\mathsf{Var}(y_i)} \frac{1}{g'(\mu_i)}\right) x_i^T \\
&= -X^T DWDX
\end{aligned}$$

- Método Scoring de Fisher = Mínimos Cuadrados Ponderados Iterativamente. Juntando los resultados anteriores y haciendo $U = DWD$, tenemos

$$\begin{aligned}
\beta^{k+1} &= \beta^k - (-X^T DWDX)^{-1} X^T DW(y - \mu) = \beta^k + (X^T UX)^{-1} X^T UD^{-1}(y - \mu) \\
&= (X^T UX)^{-1} X^T UX\beta^k + (X^T UX)^{-1} X^T UD^{-1}(y - \mu) \\
&= (X^T UX)^{-1} X^T U\left[X\beta^k + D^{-1}(y - \mu)\right] = (X^T UX)^{-1} X^T U\left[\eta + D^{-1}(y - \mu)\right]
\end{aligned}$$

  esto es,

$$\beta^{k+1} = (X^T UX)^{-1} X^T Uy^*$$

  donde $y^* = \eta + D^{-1}(y - \mu)$ son las "observaciones de trabajo". Note que en cada iteración se resuelve un problema de mínimos cuadrados ponderados.

- Estimación de Errores Estándar. La varianza asintótica del estimador de máxima verosimilitud es

$$\text{Var}(\widehat{\beta}) = \left[\text{E}\left(-\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}\right)\right]^{-1} \approx (X^T U X)^{-1}$$

donde U se evalúa en $\widehat{\beta}$.

- Ejemplo (Venables & Ripley). Los datos siguientes corresponden a un estudio de toxicidad de la cipermetrina en orugas del tabaco. Se contó con 240 orugas, 120 machos y 120 hembras. Se dividieron los 120 machos en 6 grupos de 20 orugas cada uno, similarmente las orugas hembra fueron separadas en 6 grupos. A los diferentes grupos se les aplicaron 6 dosis crecientes de cipermetrina y se registraron las diferentes tasas de mortandad en cada grupo despúes de 3 días de exposición a la cipermetrina. Se observaron los siguientes datos:

dosis	1	2	4	8	16	32
ldosis	0	1	2	3	4	5
machos	1	4	9	13	18	20
hembras	0	2	6	10	12	16

Suponemos que $y_{ij} \sim \text{Binomial}(n, p_i)$, de modo que

$$f(y_{ij}) = \begin{pmatrix} n \\ y_{ij} \end{pmatrix} p_{ij}^{y_{ij}} (1-p_{ij})^{n-y_{ij}} = \exp\left\{\frac{1}{1/n}\left(\frac{y_{ij}}{n}\log\frac{p_{ij}}{1-p_{ij}} + \log(1-p_{ij})\right) + \log\begin{pmatrix} n \\ y_{ij} \end{pmatrix}\right\}$$

así, para la variable aleatoria y_{ij}/n, i.e. proporción observada de muertes, tenemos una distribución de la familia exponencial con $a(\phi) = 1/n$, $\theta_{ij} = \text{logit}(p_{ij})$, $b(\theta_{ij}) = -\log(1-p_{ij})$, $\mu_{ij} = p_{ij}$ y con un predictor lineal, asociado con μ_{ij} vía la liga canónica:

$$\text{logit}(p_{ij}) = \eta_{ij} = x_{ij}^T\beta = \beta_0 + \beta_1 \text{género}_i + \beta_2 \text{ldosis}_j + \beta_3 \text{género}_i * \text{ldosis}_j$$

Si codificamos a los machos como 0 y a las hembras como 1, entronces, bajo la parametrización anterior, la matriz X es de la forma

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 0 & 4 & 0 \\ 1 & 0 & 5 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ 1 & 1 & 3 & 3 \\ 1 & 1 & 4 & 4 \\ 1 & 1 & 5 & 5 \end{bmatrix}$$

- Código en R. Un análisis se presenta a continuación.

```
# Datos del Venables & Ripley p.189
n   <- 20
ld  <- 0:5
X   <- rbind(cbind(1,0,ld,0), cbind(1,1,ld,ld))
y   <- c(1,4,9,13,18,20,0,2,6,10,12,16)
z   <- y/n                                  # datos son proporciones
b   <- c(-1,0,0,0)                          # valores iniciales
```

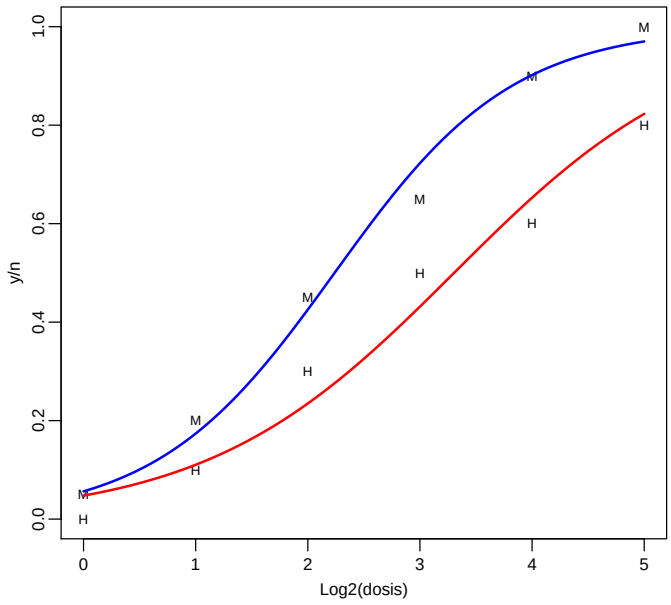
```
tolm   <- 1e-4        # tolerancia (norma minima de delta)
iterm  <- 100         # numero maximo de iteraciones
tolera <- 1           # inicializar tolera
itera  <- 0           # inicializar itera
histo  <- b           # inicializar historial de iteraciones

while( (tolera>tolm)&(itera<iterm) ){
 eta    <- as.vector(X%*%b)
 pi     <- 1/(1+exp(-eta))
 U      <- n*pi*(1-pi)
 wy     <- eta + (z-pi)/(pi*(1-pi))
 aa     <- (lm( wy ~ -1+X, weights=U ))$coeff    # minimos cuadrados ponderados
 delta  <- aa-b
 b      <- aa
 tolera <- sqrt( sum(delta*delta) )
 histo  <- rbind(histo,b)
 itera  <- itera + 1 }
# aa     <- solve(t(X)%*%(U*X), t(X)%*%(U*wy))   # minimos cuadrados ponderados

histo -1.000000  0.0000000 0.000000  0.0000000
b     -2.174121 -0.3148576 1.024498 -0.1961805
b     -2.719045 -0.2022833 1.220882 -0.3341889
b     -2.815790 -0.1764240 1.257850 -0.3521712
b     -2.818553 -0.1749886 1.258948 -0.3529122
b     -2.818555 -0.1749868 1.258949 -0.3529130
```

- De aquí, tenemos que los estimadores de máxima verosimilitud son

$$\widehat{\beta}_0 = -2.819 \quad \widehat{\beta}_1 = -0.175 \quad \widehat{\beta}_2 = 1.259 \quad \widehat{\beta}_3 = -0.353$$

```
## usando glm de R
yy  <- cbind(n*z,n-n*z)
#out <- glm( yy ~ -1+X, family=binomial(link="logit"))
out <- glm( yy ~ X[,-1], family=binomial(link="logit"))
# (nota: las formulas yy~-1+X,  yy~X[,-1], dan Devianza Nula diferentes)
summary(out)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.39849  -0.32094  -0.07592   0.38220   1.10375

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.8186     0.5480  -5.143 2.70e-07 ***
X[, -1]      -0.1750     0.7783  -0.225    0.822
X[, -1]ld     1.2589     0.2121   5.937 2.91e-09 ***
X[, -1]      -0.3529     0.2700  -1.307    0.191

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   4.9937  on  8  degrees of freedom
AIC: 43.104
```

```
Number of Fisher Scoring iterations: 4

# Replicamos los calculos anteriores de glm(), explicitamente:
pi     <- z
sel    <- (pi != 0 & pi != 1)
aa     <- z[sel]*log(pi[sel]/(1-pi[sel]))+log(1-pi[sel])
lsat   <- n*sum( aa ) + sum( lchoose(n,n*z) )

etai   <-  as.vector(X%*%b)
pf     <- 1/(1+exp(-eta))
Uf     <- n*pf*(1-pf)
errstd <- sqrt(diag(solve(t(X)%*%(U*X))))  # 0.54799 0.77831 0.21207 0.26999
pbar   <- mean(z)
lmax   <- n*sum( z*etai-log(1+exp(etai)) ) + sum( lchoose(n,n*z) )
lnull  <- n*sum( z*log(pbar/(1-pbar))+log(1-pbar) ) + sum( lchoose(n,n*z) )
NullD  <- -2*( lnull - lsat )   # 124.8756   con 13-1=12 gl
ResD   <- -2*( lmax - lsat )    # 4.993727   con 12-4= 8 gl
AIC    <- -2*lmax + 2*4         # 43.10413

# Grafica mortalidad orugas machos vs hembras
xx <- c(ld,ld)
plot(xx,z, xlab="Log2(dosis)", type="n",  mgp=c(1.5,.5,0),
 cex.lab=.8, cex.axis=.8)
points(ld,z[1:6], pch="M", cex=.7)
points(ld,z[7:12], pch="H", cex=.7)
dd <- seq(0,5,length=100)
lines(dd,1/(1+exp(-b[1]-b[3]*dd)), col="blue", lwd=2, ylab="Prop. de Muertes")
lines(dd,1/(1+exp(-b[1]-b[2]-(b[3]+b[4])*dd)), col="red", lwd=2)
```

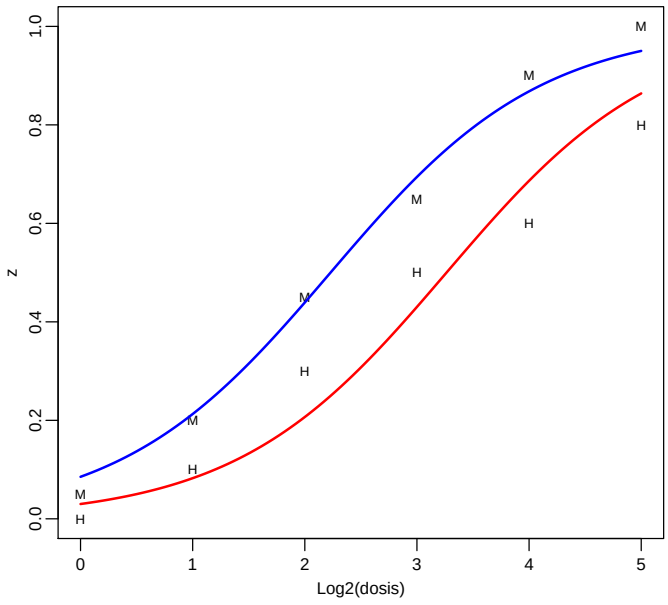



```
####################################################
```

```
# Ajuste del modelo sin interaccion
n   <- 20
ld  <- 0:5
X   <- rbind(cbind(1,0,ld), cbind(1,1,ld))
y   <- c(1,4,9,13,18,20,0,2,6,10,12,16)
z   <- y/n                                    # datos son proporciones
b   <- c(-1,0,0)                              # valores iniciales

tolm   <- 1e-4      # tolerancia (norma minima de delta)
iterm  <- 100       # numero maximo de iteraciones
tolera <- 1         # inicializar tolera
itera  <- 0         # inicializar itera
histo  <- b         # inicializar historial de iteraciones

while( (tolera>tolm)&(itera<iterm) ){
 eta    <- as.vector(X%*%b)
 pi     <- 1/(1+exp(-eta))
 U      <- n*pi*(1-pi)
 wy     <- eta + (z-pi)/(pi*(1-pi))
 aa     <- (lm( wy ~ -1+X, weights=U ))$coeff
 delta  <- aa-b
 b      <- aa
 tolera <- sqrt( sum(delta*delta) )
 histo  <- rbind(histo,b)
 itera  <- itera + 1 }
# aa     <- solve(t(X)%*%(U*X), t(X)%*%(U*wy))

histo -1.000000  0.0000000 0.000000
b     -1.928895 -0.8053089 0.926408
b     -2.303838 -1.0604790 1.038235
b     -2.370798 -1.0997547 1.063564
b     -2.372411 -1.1007428 1.064214
b     -2.372412 -1.1007434 1.064214

# o tambien
yy    <- cbind(y,n-y)
out0  <- glm( yy ~ X[,-1], family=binomial(link="logit"))
summary(out0)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3724     0.3855  -6.154 7.56e-10 ***
X[, -1]      -1.1007     0.3558  -3.093  0.00198 **
X[, -1]ld     1.0642     0.1311   8.119 4.70e-16 ***

    Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:   6.757  on  9  degrees of freedom
AIC: 42.867

pi     <- z
sel    <- (pi != 0 & pi != 1)
aa     <- z[sel]*log(pi[sel]/(1-pi[sel]))+log(1-pi[sel])
lsat   <- n*sum( aa ) + sum( lchoose(n,n*z) )
```

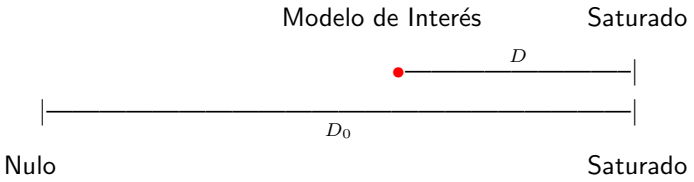
```
etai   <-  as.vector(X%*%b)
pf     <- 1/(1+exp(-eta))
Uf     <- n*pf*(1-pf)
errstd <- sqrt(diag(solve(t(X)%*%(U*X))))  # 0.3855108 0.3558271 0.1310774
pbar   <- mean(z)
lmax   <- n*sum( z*etai-log(1+exp(etai)) ) + sum( lchoose(n,n*z) )
lnull  <- n*sum( z*log(pbar/(1-pbar))+log(1-pbar) ) + sum( lchoose(n,n*z) )
NullD  <- -2*( lnull - lsat )   # 124.8756   con 13-1=12 gl
ResD0  <- -2*( lmax - lsat )    # 6.757064   con 12-3= 9 gl
AIC0   <- -2*lmax + 2*3         # 42.86747

# Comparacion del modelo con interaccion (mod completo)
# contra un modelo sin interaccion (mod reducido)

Delta <- ResD0 - ResD
pval <- 1-pchisq(Delta,4-3)  # 0.1842 no hay suf. evidencia para rechazar
# el modelo sin interaccion.

# Grafica mortalidad orugas machos vs hembras
xx <- c(ld,ld)
plot(xx,z, xlab="Log2(dosis)", type="n",  mgp=c(1.5,.5,0),
 cex.lab=.8, cex.axis=.8, main="Modelo sin interaccin")
points(ld,z[1:6], pch="M", cex=.7)
points(ld,z[7:12], pch="H", cex=.7)
dd <- seq(0,5,length=100)
lines(dd,1/(1+exp(-b[1]-b[3]*dd)), col="blue", lwd=2, ylab="Prop. de Muertes")
lines(dd,1/(1+exp(-b[1]-b[2]-b[3]*dd)), col="red", lwd=2)
```

Modelo sin interacción



Resumen de Clase 20: Miércoles 6 de abril

- Modelo saturado, modelo nulo. Los modelos nulo y saturado son dos modelos extremos. En el nulo, el predictor lineal, $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$, es de la forma $\eta_i = \beta_0$ y, básicamente, lo que suponemos es que $y_1, \cdots, y_n$ son i.i.d. $f(y;\theta)$, donde θ es un sólo parámetro. Por otro lado, en el modelo saturado tenemos un número maximal de parámetros (puede haber hasta el máximo número: n), que, por supuesto, será el modelo que mejor ajuste a los datos. El modelo de interés es un modelo intermedio:



- Devianza. La devianza nos ayuda a medir "distancias" entre modelos. En lo que sigue, consideraremos que el parámetro de dispersión, ϕ, satisface $a(\phi) = \phi/\omega$, donde ω es conocido. La devianza de un modelo particular se define como

$$D = -2\left[l(\widehat{\beta}) - l(\widehat{\beta}_{\text{sat}})\right]\phi$$

donde $l(\widehat{\beta})$ es la logverosimilitud del modelo bajo consideración, evaluada en el máximo verosímil y $l(\widehat{\beta}_{\text{sat}})$ es la logverosimilitud del modelo saturado, evaluada en el estimador máximo verosímil del parámetro de ese modelo. Es claro que esta "distancia" está basada en el valor del estadístico cociente de verosimilitudes para la comparación del modelo de interés contra el saturado.

Una medida relacionada con la devianza es la Devianza Estandarizada: $D^* = D/\phi$. Debido a su estructura,

$$D^* = -2\log\frac{L(\widehat{\beta})}{L(\widehat{\beta}_{\text{sat}})}$$

sería de esperarse que, si el modelo bajo consideración es razonable, entonces $D^* \sim \chi^2_{n-p}$; sin embargo, no es cierto en general para todos los elementos de la familia exponencial (para la Normal si es válido, de hecho, es un resultado exacto).

- Pruebas de hipótesis. Consideremos el contraste de hipótesis $H_0 : M_0$ contra $H_1 : M_1$, donde el modelo M_0 está anidado en el modelo M_1. El log del cociente de verosimilitudes es

$$\log\Lambda = l(\widehat{\beta}_0) - l(\widehat{\beta}_1)$$

y se cumple que

$$D_0^* - D_1^* = -2\left[l(\widehat{\beta}_0) - l(\widehat{\beta}_{\text{sat}})\right] + 2\left[l(\widehat{\beta}_1) - l(\widehat{\beta}_{\text{sat}})\right] = -2\left[l(\widehat{\beta}_0) - l(\widehat{\beta}_1)\right] = -2\log\Lambda$$

de modo que esperaríamos que $D_0^* - D_1^* \sim \chi^2_{p_1-p_0}$. Este resultado es asintóticamente válido para todos los elementos de la familia exponencial (suponiendo el caso de ϕ conocida). Ahora, si pudieramos considerar a $D_0^* - D_1^*$, asintóticamente independiente de D_1^*, entonces tendríamos que, asintóticamente

$$F = \frac{(D_0^* - D_1^*)/(p_1 - p_0)}{D_1^*/(n - p_1)} \overset{H_0}{\sim} F_{n-p_1}^{p_1-p_0}$$

- Prueba F. Una ventaja de la expresión F anterior es que permite cancelar ϕ del numerador y denominador, de modo que puede usarse en el caso usual en el que el parámetro de dispersión es desonocido:

$$F = \frac{(D_0 - D_1)/(p_1 - p_0)}{D_1/(n - p_1)} \overset{H_0}{\sim} F_{n-p_1}^{p_1-p_0}$$

Ahora bien, mencionamos antes que la distribución del denominador no es posible, en general, justificarla como χ^2_{n-p}, (ni el resultado necesario de independendencia), sin embargo se ha encontrado (Wood (2006)) que conduce a una prueba conservadora y puede usarse en la práctica.

Resumiendo, para probar $H_0 : M_0$ contra $H_1 : M_1$, con M_0 anidado en M_1:

  - Si ϕ no es conocida, entonces

$$F = \frac{(D_0 - D_1)/(p_1 - p_0)}{D_1/(n - p_1)} \overset{H_0}{\sim} F_{n-p_1}^{p_1-p_0}$$

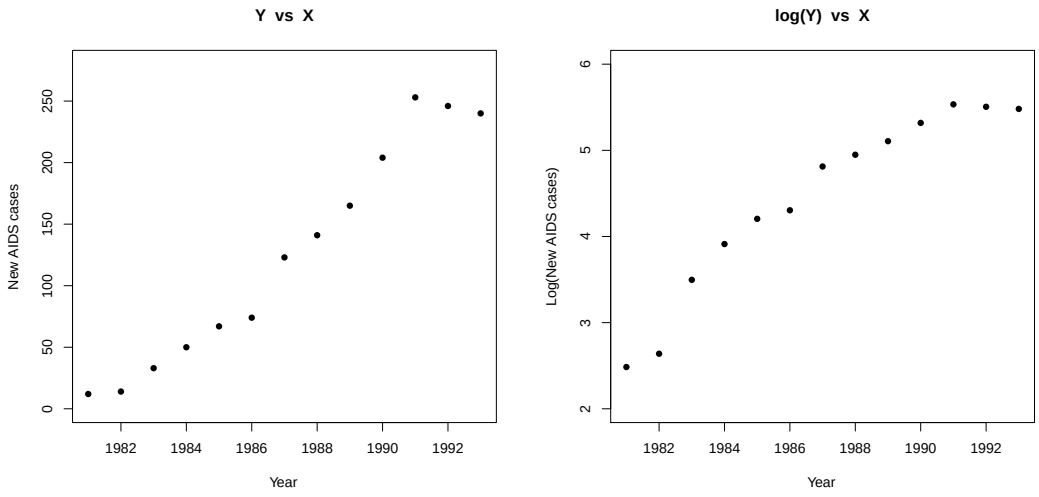
  - Si ϕ es conocida, entonces

$$D_0^* - D_1^* \overset{H_0}{\sim} \chi^2_{p_1-p_0}$$

- Ejemplo. El siguiente ejemplo considera el número de casos nuevos de SIDA en Bélgica (datos de los 80's). Supondremos que es razonable considerar el modelo Poisson.

$$y_i \sim f(y_i; \lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} = \exp\{y_i \log\lambda_i - \lambda_i - \log(y_i!)\}$$

```
y  <- c(12,14,33,50,67,74,123,141,165,204,253,246,240)
n  <- length(y)
x  <- 1:n

plot(x+1980,y,xlab="Year",ylab="New AIDS cases",ylim=c(0,280),pch=16,
 main="Y  vs  X")
plot(x+1980,log(y),xlab="Year",ylab="Log(New AIDS cases)",ylim=c(2,6),
 main="log(Y)  vs  X",pch=16)
```



La gráfica de la izquierda sugiere liga identidad y predictor con término cuadrático y lineal; la de la derecha: Liga canónica y predictor con término cuadrático y lineal. Optamos por la segunda.

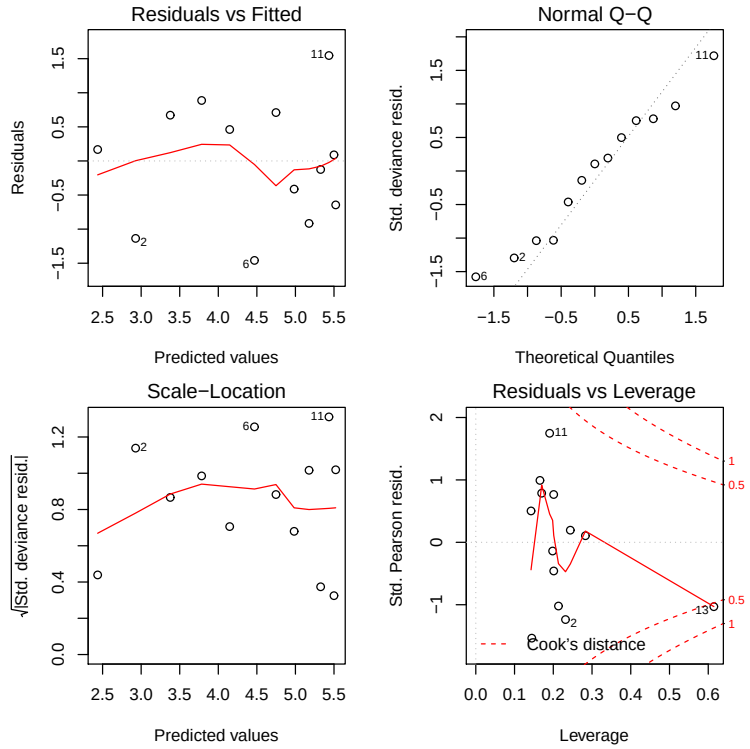
- Para el modelo Poisson, las expresiones del Scoring de Fisher son

$$\beta^{k+1} = (X^T U X)^{-1} X^T U y^*, \quad \text{donde } U = DWD \quad \text{y} \quad y^* = \eta + D^{-1}(y - \mu)$$

y es fácil ver que $D = \text{diag}(\lambda_i)$, $W = \text{diag}(\lambda_i^{-1})$, $U = \text{diag}(\lambda_i)$ y $\mu = (\lambda_1, \cdots, \lambda_n)^T$.

```
X  <- cbind( rep(1,n), x, x^2 )  # modelo cuadratico
b  <- c(2,1,0)      # valores iniciales de parametros

tolm   <- 1e-6      # tolerancia (norma minima de delta)
iterm  <- 1000      # numero maximo de iteraciones
tolera <- 1         # inicializar tolera
itera  <- 0         # inicializar itera
histo  <- b         # inicializar historial de iteraciones

while( (tolera>tolm)&(itera<iterm) ){
 eta    <- as.vector(X%*%b)
 lamb   <- exp(eta)              # liga canonica log(lamb) = eta
 z      <- eta + (y-lamb)/lamb
 aa     <- as.vector(solve(t(X*lamb)%*%X, t(X*lamb)%*%z))
 delta  <- aa-b
 b      <- aa
 tolera <- sqrt( sum(delta*delta) )
 histo  <- rbind(histo,b)
 itera  <- itera + 1 }

histo 2.0000000  1.00000000  0.0000000000
b     1.1211293  0.97953719  0.0008592407
... (14 iteraciones)
b     1.9014586  0.55600327 -0.0213462716

lsat   <- sum( y*log(y) - y - lfactorial(y) )  # logv(saturado)
eta    <- as.vector(X%*%b)
lamb   <- exp(eta)
errstd <- sqrt(diag(solve(t(X*lamb)%*%X))) # 0.186877 0.045780 0.002659
lmax   <- sum( y*log(lamb) - y - lfactorial(y) )   # logv(mod interes)
lambN  <- mean(y)
lnull  <- sum( y*log(lambN) - y - lfactorial(y) )  # logv(nulo)
NullD  <- -2*( lnull - lsat )   # 872.2058   con 13-1=12 gl
ResD   <- -2*( lmax - lsat )    # 9.240248   con 13-3=10 gl
AIC    <- -2*lmax + 2*3         # 96.92358

# Usando glm
m1 <- glm(y~x+I(x^2),poisson)
par(mfrow=c(2,2),mar=c(4,4,2,2))
plot(m1)
summary(m1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.901459   0.186877  10.175  < 2e-16 ***
t            0.556003   0.045780  12.145  < 2e-16 ***
I(t^2)      -0.021346   0.002659  -8.029 9.82e-16 ***

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 872.2058  on 12  degrees of freedom
Residual deviance:   9.2402  on 10  degrees of freedom
AIC: 96.924

Number of Fisher Scoring iterations: 4
```



- Comparamos el modelo cuadrático vs el lineal

```
m0 <- summary( glm(y~x,poisson) )
m1 <- summary( glm(y~x+I(x^2),poisson) )

# Estadistico de prueba

ji <- m0$deviance - m1$deviance
gl <- m0$df.residual - m1$df.residual

1-pchisq(ji,gl)   # = 0 => se rechaza fuertemente el modelo con solo la parte lineal
```

Segundo Examen de Modelos Estadísticos I

Nombre: _____

1. El archivo "Consumo.xls" contiene 6 columnas y 48 renglones, correspondientes a información de consumo de gasolina en 48 estados de la Unión Americana (datos de principios de los 70's). Las variables consideradas son:

   - POB = Población en 1971 (en miles).
   - IMP = Impuesto a la gasolina (centavos por galón).
   - NCO = Número de conductores.
   - ING = Ingreso per cápita (miles de dólares).
   - CAM = Miles de millas de carreteras en el estado (en 1971).
   - GAS = Consumo de gasolina en el estado (millones de galones).

   Es de particular interés estudiar GAS en función de las demás variables. Antes de iniciar el análisis observamos que unas variables son función del tamaño de la población, así que será mejor reescalar NCO y GAS dividiendolas por POB. Defina GASP = GAS/POB y NCOP = NCO/POB.

   Efectúe un análisis completo de GASP en función del resto de variables (IMP, NCOP, ING, CAM). El análisis deberá incluír desde gráficas de la variable dependiente versus variables individuales, análisis de residuales, pruebas de significancia de parámetros de los modelos y uso de herramientas de diagnósitico. Indique claramente cuál es su modelo final y para que estados es aplicable. (Nota: No porque vimos identificación de outliers, a fuerza deberán eliminar observaciones. Sin embargo, es importante justificar cada decisión).

2. Consideremos el problema de seleccionar un modelo entre una colección de modelos anidados. Una forma natural de tomar esta decisión es seleccionar el modelo con la mínima suma de cuadrados del error o con la máxima verosimilitud, sin embargo, estos criterios llevan inevitablemente a elegir el modelo con más términos. Una forma de abordar este problema es el de considerar, por ejemplo, el estadístico F, el cual nos da un medio de elegir el modelo más simple que sea consistente con los datos observados, donde consistente es en el sentido de cierto nivel de significancia.

   En este ejercicio consideraremos una forma alternativa basada en la idea de elegir el modelo con más potencial de predecir $\mu = \mathsf{E}(y)$, en vez del modelo que esté más cercano a los datos y. Como hemos visto antes, estimamos un modelo lineal encontrando β que maximiza:

   $$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}||y - X\beta||^2\right\}$$

   en general, mientras más parámetros (equivalentemente, variables) consideremos en el modelo, mejor ajustaremos a y, aunque esto signifique que estamos simplemente ajustando mejor la componente de ruido de y. Ahora, este sobreajuste podría, tal vez, resolverse si pudieramos mejor maximizar:

   $$K(\widehat{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}||\mu - X\widehat{\beta}||^2\right\}$$

   Consideremos la lógica de este enfoque. Suponga que añadimos términos al modelo y que en algún punto en la secuencia de modelos nos encontramos con el "verdadero". Si realmente conociéramos μ entonces alcanzaríamos el máximo de esta función objetivo una vez que tuvieramos el modelo suficientemente grande

para representar a μ y permaneceríamos en el máximo aún si más términos (redundantes) fueran agregados al modelo. Por otro lado, si elegimos a L como función objetivo entonces en el proceso secuencial de consideración de modelos, eventualmente maximizaríamos a K, pero al incrementar más términos (bajo el criterio L) nos alejaríamos del óptimo de K al tener que las predicciones se alejan de μ por acercarse a y. Ahora, esto suena bien pero no podemos maximizar K directamente pues no conocemos μ, sin embargo, se puede estimar.

(a) Sea P la matriz de proyección sobre $\mathcal{C}(X)$ y suponga que $y = X\beta + e$ donde e es un vector apropiado de discrepancias. Sabemos que $X\widehat{\beta} = Py$. Muestre que

$$||\mu - X\widehat{\beta}||^2 = ||\mu - Py||^2 = ||y - Py||^2 - e^T e - 2e^T\mu + 2e^T P\mu + 2e^T Pe$$

(b) Muestre que

$$\mathsf{E}\left(||\mu - X\widehat{\beta}||^2\right) = \mathsf{E}\left(||y - Py||^2\right) - n\sigma^2 + 2\,\mathrm{tr}(P)\sigma^2$$

de aquí que podemos estimar a $\mathsf{E}\left(||\mu - X\widehat{\beta}||^2\right)$ mediante $||y - Py||^2 - n\sigma^2 + 2\,\mathrm{tr}(P)\sigma^2$ y, entonces, un estimador apropiado para $\kappa = \log K$ es

$$\widehat{\kappa} = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}||y - X\widehat{\beta}||^2 + \frac{n}{2} - \mathrm{tr}(P)$$

(c) Muestre que $\widehat{\kappa}$ es maximizado por el modelo que minimize el Criterio de Información de Akaike:

$$\mathsf{AIC} = -2\,l(\widehat{\beta}, \sigma^2) + 2p$$

3. El artículo de Nelder y Wedderburn (1972) marca muy claramente el inicio de la teoría de los modelos lineales generalizados:

> Nelder, J.A. & Wedderburn, W.M. (1972). Generalized Linear Models
> *J. R. Statist. Soc.* **A**, **135**, *Part* 3, pp 370-384.

- En la página 371 dice "This procedure is a generalization of the well-known one described by Finney (1952) for maximum likelihood estimation in probit analysis". ¿Qué es y para que se usa el análisis Probit?, Muestre un ejemplo de un análisis Probit (hay que incluir los datos y el correspondiente análisis ejecutado en R).
- En la página 375 se muestran 4 devianzas. Deducir esas expresiones.
- En la página 378, en la sección sobre el modelo Poisson, se presenta un ejemplo de una tabla de contingencia 5×4. Mostrar los cálculos que llevan a las conclusiones "... that the data are adequately described by a negative linear $\times$ linear interaction" (no hay necesidad de incluir ni discutir la segunda tabla donde usan el método de Yates).

4. Los datos de la siguiente tabla son números, n, de pólizas de seguros y los correspondientes números, y, de reclamos (esto es, número de accidentes en los que se pidió el amparo de la póliza). La variable CAR es una codificación de varias clases de carros, EDAD es la edad del titular de la póliza y DIST es el distrito donde vive el titular.

CAR	EDAD	DIST$=0$ y	DIST$=0$ n	DIST$=1$ y	DIST$=1$ n
1	1	65	317	2	20
1	2	65	476	5	33
1	3	52	486	4	40
1	4	310	3259	36	316
2	1	98	486	7	31
2	2	159	1004	10	81
2	3	175	1355	22	122
2	4	877	7660	102	724
3	1	41	223	5	18
3	2	117	539	7	39
3	3	137	697	16	68
3	4	477	3442	63	344
4	1	11	40	0	3
4	2	35	148	6	16
4	3	39	214	8	25
4	4	167	1019	33	114

(a) Calcule la tasa de reclamos, y/n, para cada categoría y grafique estas tasas contra las diferentes variables para tener una idea de los efectos principales.

(b) Use regresión Poisson para estimar los efectos principales (cada variable tratada como categórica y modelada usando variables indicadoras) así como sus interacciones.

(c) Basados en los resultados del inciso anterior, los autores del artículo donde aparecieron estos datos, decidieron que ninguna interacción era importante y que podían considerar que CAR y EDAD fuesen tratadas como variables continuas. Ajuste un modelo incorporando estas observaciones y compárelo con el obtenido en (b). ¿Cuáles son las conclusiones?.

5. (a) Considere un modelo lineal $y_i = x_i^T\beta + e_i$, bajo los supuestos usuales. Muestre que

$$\frac{\partial \widehat{y}_i}{\partial y_i} = h_i$$

donde h_i es el $i-$ésimo elemento diagonal de la matriz de proyección y $\widehat{y}_i = x_i^T\widehat{\beta}$. ¿Qué interpretación le damos a esta derivada?.

(b) Suponga que $y_i \sim$ Bernoulli $(p_i),\ i = 1, \cdots, n$, las y_i's son observadas bajo condiciones de independencia y

$$p_i = F(x_i^T\beta) = \frac{\exp(x_i^T\beta)}{1 + \exp(x_i^T\beta)}$$

Sabemos que las ecuaciones correspondientes a máxima verosimilitud son de la forma

$$\sum_{i=1}^{n}(y_i - p_i)x_i = 0$$

Definamos $\widehat{y}_i = F(x_i^T\widehat{\beta})$, donde $\widehat{\beta}$ es el estimador máximo verosimil de β. Queremos encontrar $\partial\widehat{y}_i/\partial y_i$. Muestre que

$$\frac{\partial \widehat{y}_i}{\partial y_i} = F'(x_i^T\widehat{\beta})x_i^T\frac{\partial \widehat{\beta}}{\partial y_i}$$

Note que $\widehat{\beta}$ satisface:

$$\sum_{i=1}^{n}(y_i - F(x_i^T\widehat{\beta}))x_i = 0$$

Derivando ambos lados con respecto a y_i se puede encontrar la derivada de $\widehat{\beta}$ con respecto a y_i, lo cual, nos lleva a que (mostrar):

$$\frac{\partial \widehat{y}_i}{\partial y_i} = h_i$$

donde h_i es el $i-$ésimo elemento diagonal de

$$H = V^{1/2} X (X^T V X)^{-1} X^T V^{1/2}$$

y $V = \text{diag}(F'(x_1^T \widehat{\beta}), \cdots, F'(x_n^T \widehat{\beta}))$.

6. Considere el caso de observaciones independientes $y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \cdots, n$, con vectores $p \times 1$ de covariables $x_1, \cdots, x_n$.

   (a) Para este caso, escriba todos los elementos de un modelo lineal generalizado (suponga la liga canónica).

   (b) Considere el método iterativo scoring de Fisher

   $$\beta^{(k+1)} = (X^T U X)^{-1} X^T U (\eta + D^{-1}(y - \mu))$$

   Muestre que, no importa cuales sean los valores iniciales, el método converge en una iteración.

7. Sea (Y, Z) una variable aleatoria bi-dimensional. Suponga que, tanto Y como Z son variables categóricas con I niveles para Y y J niveles para Z. A la distribución conjunta de (Y, Z) se le llama $Tabla\ de\ Contingencia\ I \times J$

		Z				
		1	2	$\cdots$	J	
	1	π_{11}	π_{12}	$\cdots$	π_{1J}	$\pi_{1.}$
	2	π_{21}	π_{22}	$\cdots$	π_{2J}	$\pi_{2.}$
Y	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	π_{I1}	π_{I2}	$\cdots$	π_{IJ}	$\pi_{I.}$
		$\pi_{.1}$	$\pi_{.2}$	$\cdots$	$\pi_{.J}$	1

Por ejemplo, Y puede ser la variable nivel de educación y Z el ingreso (categorizado en intervalos de ingreso). Sea y_{ij} el número de veces que se observa, en un estudio, la ocurrencia de un individuo con los atributos (i, j) (i.e. y_{ij} es la frecuencia observada en la celda (i, j)). Suponga que es razonable modelar y_{ij} mediante una Poisson con parámetro μ_{ij}, donde $\mu_{ij} = n\pi_{ij}$ y n es el número total de observaciones (las cuales suponemos independientes).

   (a) Si usamos la liga canónica en el caso Poisson, tenemos

   $$\log(\mu_{ij}) = \alpha + \text{efecto de } Y + \text{efecto de } Z$$

   Escriba la matriz X correspondiente. Recuerde que Y y Z son categóricas, así que podría usar dummys; por ejemplo, puede escribir

   $$\log(\mu_{ij}) = \alpha + \beta_i + \gamma_j$$

   donde $\beta_I = 0$ y $\gamma_J = 0$.

   (b) Muestre que si el modelo anterior es correcto entonces Y y Z son independientes.

8. Los datos mostrados en la tabla provienen de un estudio sobre ocurrencia de infecciones en mujeres con partos por cesárea. Los factores considerados son si la cesárea fue planeada o no, si había factores de riesgo extra (por ejemplo, diabetes y/o sobrepeso) y si se usó antibióticos como preventivo de infecciones o no. Deseamos evaluar el impacto de estos factores sobre la ocurrencia de infecciones postparto en casos de nacimientos por cesárea.

		Cesárea planeada		Cesárea no planeada	
		Infección		Infección	
		si	no	si	no
Con antibióticos	Con factores de riesgo	1	17	11	87
	Sin factores de riesgo	0	2	0	0
Sin antibióticos	Con factores de riesgo	28	30	23	3
	Sin factores de riesgo	8	32	0	9

(a) Escriba un programa en R para ajustar un modelo apropiado. El programa debe incluir el cálculo de errores estándar.

(b) Compare los resultados obtenidos con la salida de la función `glm`. De sus conclusiones en el contexto del problema.

Entregar el examen el viernes 15 de abril a las 12:30.

Resumen de Clase 21: Lunes 11 de abril

- Estimación del Parámetro de Dispersión. En la clase anterior definimos la Devianza como

$$D = -2\left[l(\widehat{\beta}) - l(\widehat{\beta}_{\text{sat}})\right]\phi$$

y la Devianza Estandarizada como $D^* = D/\phi$. Comentamos también que, en algunos casos, es razonable que $D^* \sim \chi^2_{n-p}$. De aquí se deduce que, como

$$\mathsf{E}(D^*) = n - p, \quad \text{entonces} \quad \mathsf{E}\left(\frac{D}{\phi}\right) = n - p$$

y entonces, un estimador para ϕ es

$$\widehat{\phi} = \frac{1}{n-p}D$$

por supuesto, la lógica que se usa en su deducción no es completamente sólida, sin embargo, es ampliamente usado en la práctica.

- Estimación del Parámetro de Dispersión: Estimador tipo Pearson. Recuerde que, si y_i pertenece a la familia exponencial, entonces

$$\mathsf{Var}(y_i) = \frac{\phi}{w_i}\frac{\partial^2 b(\theta_i)}{\partial\theta_i^2} \equiv \phi V(\mu_i)$$

El estadístico generalizado de Pearson, X^2, se define como:

$$X^2 = \sum_{i=1}^{n}\frac{(y_i - \widehat{\mu}_i)^2}{V(\mu_i)}$$

dividiendo entre el parámetro de dispersión, tenemos

$$\frac{X^2}{\phi} = \sum_{i=1}^{n}\frac{(y_i - \widehat{\mu}_i)^2}{\phi V(\mu_i)} = \sum_{i=1}^{n}\left(\frac{y_i - \widehat{\mu}_i}{\sqrt{\mathsf{Var}(y_i)}}\right)^2$$

esta última expresión tiene una distribución aproximada χ^2_{n-p} (ver libro de Dobson (2002), pág. 126), entonces, otro estimador para el parámetro de dispersión es

$$\widehat{\phi} = \frac{1}{n-p}X^2$$

- Residuales para GLM's. Los residuales usuales, $r_i = y_i - \widehat{\mu}_i$, no tienen varianza constante, así que es difícil usarlos directamente como herramienta de diagnóstico. Dos opciones para residuales son:
  - Residuales de Pearson.

$$r_i^p = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i}}$$

si el modelo es correcto, estos residuales tienen aproximadamente media cero y varianza ϕ (sin embargo, pueden tener un comportamiento asimétrico alrededor de 0).

– Residuales basados en la Devianza. Recuerde que

$$D = -2\left[l(\widehat{\beta}) - l(\widehat{\beta}_{\text{sat}})\right]\phi = \sum_{i=1}^{n} 2w_i\left[y_i(\tilde{\theta}_i - \widehat{\theta}_i) - b(\tilde{\theta}_i) + b(\widehat{\theta}_i)\right] \equiv \sum_{i=1}^{n} d_i$$

Los residuales basados en devianzas se definen como

$$r_i^d = \text{signo}(y_i - \widehat{\mu}_i)\sqrt{d_i}$$

(como en el caso Normal, la suma de cuadrados de estos residuales dan la devianza o suma de cuadrados del error).

Resumen de Clase 22: Miércoles 27 de abril

- La función Score. Recordemos la forma de la familia Exponencial:

$$y_i \sim \exp\left\{ \frac{w_i}{\phi}\left(y_i\theta_i - b(\theta_i)\right) + c(y_i, \phi, w_i) \right\}$$

donde $\mu_i = \mathsf{E}(y_i) = b'(\theta_i)$, $\mathsf{Var}(y_i) = \phi\, b''(\theta_i)/w_i \equiv \phi\, V(\mu_i)$ y $g(\mu_i) = \eta_i = x_i^T\beta$. La logverosimilitud es de la forma

$$l(\beta) = \sum_{i=1}^{n}\left\{ \frac{w_i}{\phi}\left(y_i\theta_i - b(\theta_i)\right) + c(y_i, \phi, w_i) \right\} \equiv \sum_{i=1}^{n} l_i(\beta)$$

y la función score es:

$$\mathcal{U}(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n}\frac{\partial l_i(\beta)}{\partial \beta} = \sum_{i=1}^{n}\frac{y_i - \mu_i}{\mathsf{Var}(y_i)}\frac{1}{g'(\mu_i)}x_i \equiv \sum_{i=1}^{n}\mathcal{U}_i(\beta)$$

Note que la función score es una transformación $\mathcal{U} : \mathbb{R}^p \to \mathbb{R}^p$, con propiedades

  - Si β es el valor verdadero del parámetro, entonces $\mathsf{E}\left(\, \mathcal{U}_i(\beta)\, \right) = 0,\ \ i = 1, \cdots, n.$
  - Si β es el valor verdadero del parámetro, entonces $\mathsf{E}\left(\, \mathcal{U}(\beta)\, \right) = 0.$
  - $\mathcal{U}_1(\beta), \cdots, \mathcal{U}_n(\beta)$ son independientes.
  - $V_n \equiv \mathsf{Var}\left(\, \mathcal{U}(\beta)\, \right) = \mathsf{E}\left(\, \mathcal{U}(\beta)\mathcal{U}^T(\beta)\, \right) = \mathsf{E}\left(\frac{\partial l(\beta)}{\partial \beta}\frac{\partial l(\beta)}{\partial \beta^T} \right) = \mathsf{E}\left(-\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} \right) =$ Mat. de Información
  - Además $V_n = \mathsf{Var}\left(\, \mathcal{U}(\beta)\, \right) = \sum_{i=1}^{n}\mathsf{Var}\left(\, \mathcal{U}_i(\beta)\, \right).$

- Normalidad asintótica de la función Score. Consideremos la siguiente versión del Teorema Central del Límite: Sea $x_1, x_2, \cdots$ una sucesión de variables aleatorias k-dimensionales independientes con $\mathsf{E}(x_i) = 0$ y $\mathsf{Var}(x_i) = \Sigma_i$. Suponga que

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Sigma_i = \Sigma > 0$$

suponga además que, para cada $\epsilon > 0$,

$$\frac{1}{n}\sum_{i=1}^{n}\int_{||x_i|| > \epsilon\sqrt{n}} ||x_i||^2\, dF_i \to 0$$

donde F_i es la función de distribución de x_i. Entonces

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \xrightarrow{d} N_k(0, \Sigma)$$

(Ver libro de Rao (1973), pág. 147).

Usando este resultado, tenemos que

$$\frac{1}{\sqrt{n}}\mathcal{U}(\beta) \xrightarrow{d} N_p(0, V)$$

donde estamos suponiendo, en particular, que

$$\frac{1}{n}V_n = \frac{1}{n}\mathsf{Var}\left(\, \mathcal{U}(\beta)\, \right) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{Var}\left(\, \mathcal{U}_i(\beta)\, \right) = \frac{1}{n}\mathsf{E}\left(-\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} \right) \to V > 0$$

Una reformulación de este resultado es

$$\frac{1}{\sqrt{n}}\, V^{-1/2}\, \mathcal{U}(\beta) \xrightarrow{d} N_p(0, I)$$

y, puede verse que esto es equivalente a

$$V_n^{-1/2}\, \mathcal{U}(\beta) \xrightarrow{d} N_p(0, I)$$

donde estamos suponiendo que β es el verdadero valor del parámetro.

- Propiedades asintóticas de los estimadores de Máxima Verosimilitud en GLM's. Usando una aproximación de primer orden de la función score alrededor del valor verdadero de β tenemos que

$$\mathcal{U}(\widehat{\beta}) \approx \mathcal{U}(\beta) + \left[\frac{\partial \mathcal{U}(\beta)}{\partial \beta}\right](\widehat{\beta} - \beta)$$

pero $\widehat{\beta}$ es el estimador de máxima verosimilitud, entonces $\mathcal{U}(\widehat{\beta}) = 0$ y de aquí que

$$\widehat{\beta} - \beta \approx \left[-\frac{\partial \mathcal{U}(\beta)}{\partial \beta}\right]^{-1} \mathcal{U}(\beta)$$

entonces

$$V_n^{1/2}\,(\widehat{\beta} - \beta) \approx V_n^{1/2} \left[-\frac{\partial \mathcal{U}(\beta)}{\partial \beta}\right]^{-1} V_n^{1/2}\, V_n^{-1/2}\, \mathcal{U}(\beta)$$

Note que

$$\begin{aligned} V_n^{1/2} \left[-\frac{\partial \mathcal{U}(\beta)}{\partial \beta}\right]^{-1} V_n^{1/2} &= \left(\frac{1}{n} V_n\right)^{1/2} \left[-\frac{1}{n}\frac{\partial \mathcal{U}(\beta)}{\partial \beta}\right]^{-1} \left(\frac{1}{n} V_n\right)^{1/2} \\ &= \left(\frac{1}{n} V_n\right)^{1/2} \left[-\frac{1}{n}\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}\right]^{-1} \left(\frac{1}{n} V_n\right)^{1/2} \end{aligned}$$

y, puede verse que

$$-\frac{1}{n}\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} \xrightarrow{p} V$$

de modo que

$$V_n^{1/2} \left[-\frac{\partial \mathcal{U}(\beta)}{\partial \beta}\right]^{-1} V_n^{1/2} = \left(\frac{1}{n} V_n\right)^{1/2} \left[-\frac{1}{n}\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}\right]^{-1} \left(\frac{1}{n} V_n\right)^{1/2} \xrightarrow{p} I$$

y, por lo tanto

$$V_n^{1/2}\,(\widehat{\beta} - \beta) \xrightarrow{d} N_p(0, I)$$

Para propósitos prácticos, en aplicaciones usamos que

$$\widehat{\beta} \sim N_p(\beta, V_n^{-1})$$

donde V_n es la Matriz de Información de Fisher. Recuerde que en la clase 19 vimos que

$$V_n = \mathsf{E}\left(-\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}\right) = X^T D W D X$$

donde $X^T = [x_1, \cdots, x_n]_{p \times n}$, $D = \text{diag}(\ 1/g'(\mu_i)\ )$ y $W = \text{diag}(\ 1/\text{Var}(y_i)\ )$. Al evaluar estas expresiones en los valores verdaderos de los parámetros, β y ϕ, obtenemos que V_n^{-1} es la varianza asintótica correcta. En la práctica, usamos estas expresiones evaluadas en los estimadores de máxima verosimilitud, de modo que V_n^{-1} es la varianza asintótica aproximada.

- Referencia. Para un tratamiento formal sobre estas propiedades asintóticas ver:

  Fahrmer & Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, **Vol** 13, No. 1, 342-368.

- Cociente de Verosimilitudes. Usando una aproximación de segundo orden para la logverosimilitud, alrededor de $\widehat{\beta}$, tenemos

$$l(\beta) \approx l(\widehat{\beta}) + \frac{\partial l(\widehat{\beta})}{\partial \beta^T}(\beta - \widehat{\beta}) + \frac{1}{2}(\beta - \widehat{\beta})^T \left(\frac{\partial^2 l(\widehat{\beta})}{\partial \beta \partial \beta^T} \right) (\beta - \widehat{\beta})$$

  el segundo término del lado derecho es 0. Entonces

$$-2[l(\beta) - l(\widehat{\beta})] \approx (\widehat{\beta} - \beta)^T \left(-\frac{\partial^2 l(\widehat{\beta})}{\partial \beta \partial \beta^T} \right) (\widehat{\beta} - \beta) \approx (\widehat{\beta} - \beta)^T V_n (\widehat{\beta} - \beta)$$

  la expresión del lado derecho es una forma cuadrática en variables normales (mejor dicho: asintóticamente normales), de modo que $-2\log\Lambda$ tiene una distribución ji-cuadrada con ciertos grados de libertad, donde Λ es un cociente de verosimilitudes. Detalles formales verlos en Cox & Hinkley (1974).

Resumen de Clase 23: Viernes 28 de abril

- Ecuaciones Normales para GLM's. En la clase pasada vimos que para hacer máxima verosimilitud en el modelo GLM tenemos que resolver

$$\mathcal{U}(\beta) = \sum_{i=1}^{n} \mathcal{U}_i(\beta) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathsf{Var}(y_i)} \frac{1}{g'(\mu_i)} x_i = X^T D W (y - \mu) = 0$$

  Vamos a cambiar un poco de notación. Recuerde que

$$W = \mathsf{diag}(\ 1/\mathsf{Var}(y_i)\ ) = \frac{1}{\phi}\mathsf{diag}(\ 1/V(\mu_i)\ ) \equiv \frac{1}{\phi} B^{-1}$$

  donde $V(\mu_i) = b''(\theta_i)/w_i$ y $B = \mathsf{diag}(V(\mu_i))$. Por otro lado, sea $H^T = [\partial\mu_1/\partial\beta, \cdots, \partial\mu_n/\partial\beta]$. Note que, por la regla de la cadena, tenemos que $H^T = X^T D$. Entonces, introduciendo estos cambios de notación tenemos que las ecuaciones normales son

$$\mathcal{U}(\beta) = \sum_{i=1}^{n} \mathcal{U}_i(\beta) = \frac{1}{\phi} H^T B^{-1} (y - \mu) = 0$$

- El estimador de máxima Cuasi-Verosimilitud. Note que para especificar el sistema de ecuaciones normales necesitamos
  - Un modelo para la media en función de covariables: $\mu_i = \mu_i(\beta)$. Con esto se definen la matriz H y el vector μ.
  - Un modelo para la varianza: $\mathsf{Var}(y_i) = \phi V(\mu_i)$, donde $V(\mu_i)$ es la función de varianza y ayuda a modelar la varianza en función de la media. Con esto se define la matriz B.

  Si sólo podemos contar con estas dos componentes, $\mu_i(\beta)$ y $V(\mu_i)$, entonces, de todos modos podemos construír las ecuaciones $\mathcal{U}(\beta) = 0$. A la solución de este sistema de ecuaciones (típicamente no-lineales) se le llama "Estimador de Máxima Cuasi-Verosimilitud". Un tratamiento del tema se encuentra en:

  McCullagh, P. (1985). Quasi-likelihood functions. *The Annals of Statistics*, **Vol** 11, No. 1, 59-67.

- Cuasi-Verosimilitud. Supongamos datos independientes $y_1, \cdots, y_n$ y covariables $x_1, \cdots, x_n$; además supongamos que hemos definido los modelos $\mathsf{E}(y_i) = \mu_i(\beta)$ y $\mathsf{Var}(y_i) = \phi V(\mu_i)$. Queremos definir algo parecido a una logverosimilitud. Una forma de proceder es definir una función de Cuasi-Logverosimilitud como aquella función cuya derivada es $\frac{1}{\phi} H^T B^{-1}(y - \mu)$, pues, por analogía con los GLM's, $\frac{1}{\phi} H^T B^{-1}(y - \mu)$ es la función Score, la cual es la derivada de la logverosimilitud.

  Definimos la contribución de la observación y_i a la cuasi-logverosimilitud como:

$$q_i(\beta) = \int_{y_i}^{\mu_i} \frac{y_i - z}{\phi V(z)} dz$$

  La cuasi-logverosimilitud completa es

$$q(\beta) = \sum_{i=1}^{n} q_i(\beta)$$

  Note que si derivamos esta cuasi-logverosimilitud se obtiene la función "cuasi-score":

$$\frac{\partial q(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial q_i(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial q_i(\beta)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta} = \frac{1}{\phi} H^T B^{-1} (y - \mu)$$

  También note que si queremos maximizar la cuasi-logverosimilitud entonces tenemos que rersolver $H^T B^{-1}(y - \mu) = 0$, las cuales son las mismas ecuaciones que para los GLM's; esto implica que podemos usar el mismo algoritmo, IRWLS, para encontrar el máximo.

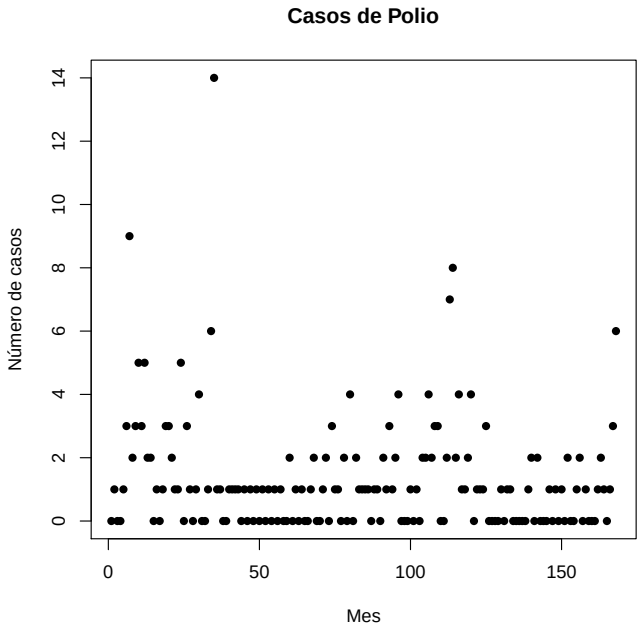
- Propiedades de la función Cuasi-Score.
  - Es una suma de variables independientes. La función cuasi-score es una suma de variables aleatorias independientes:
$$\frac{\partial q(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial q_i(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta}$$
  - Media. Si β es el valor verdadero del parámetro entonces el valor esperado de la cuasi-score es 0:
$$\mathsf{E}\left(\frac{\partial q(\beta)}{\partial \beta}\right) = \frac{1}{\phi} H^T B^{-1} [\, \mathsf{E}(y) - \mu \,] = 0$$
  - Varianza. La varianza de la cuasi-score es:
$$\mathsf{Var}\left(\frac{\partial q(\beta)}{\partial \beta}\right) = \frac{1}{\phi^2} H^T B^{-1} \mathsf{Var}(y) B^{-1} H = \frac{1}{\phi^2} H^T B^{-1} \phi B B^{-1} H = \frac{1}{\phi} H^T B^{-1} H$$
  - Matriz de Información. Si β es el valor verdadero del parámetro, entonces, es fácil ver que una expresión alternativa para la varianza de la cuasi-score es
$$\mathsf{Var}\left(\frac{\partial q(\beta)}{\partial \beta}\right) = \frac{1}{\phi} H^T B^{-1} H = \mathsf{E}\left(-\frac{\partial^2 q(\beta)}{\partial \beta \partial \beta^T}\right)$$

- Distribución asintótica del estimador de máxima cuasi-verosimilitud. Si comparamos las propiedades anteriores con las vistas en la clase pasada sobre la función score vemos que son idénticas. La deducción esbozada para ver la normalidad asintótica del estimador de máxima verosimilitud es completamente adaptable al caso cuasi-verosímil, así que, para propósitos práctico tenemos que si $\tilde{\beta}$ es el máximo cuasi-verosímil, entonces
$$\tilde{\beta} \sim N_p(\beta, V_n^{-1})$$
donde $V_n = H^T B^{-1} H/\phi$; en particular, una de las condiciones que hay que pedir es que V_n/n converja a algo positivo definido.

- Estimación del parámetro de dispersión. El estimador de Pearson para ϕ es
$$\widehat{\phi} = \frac{1}{n-p} (y - \widehat{\mu})^T B^{-1} (y - \widehat{\mu})$$

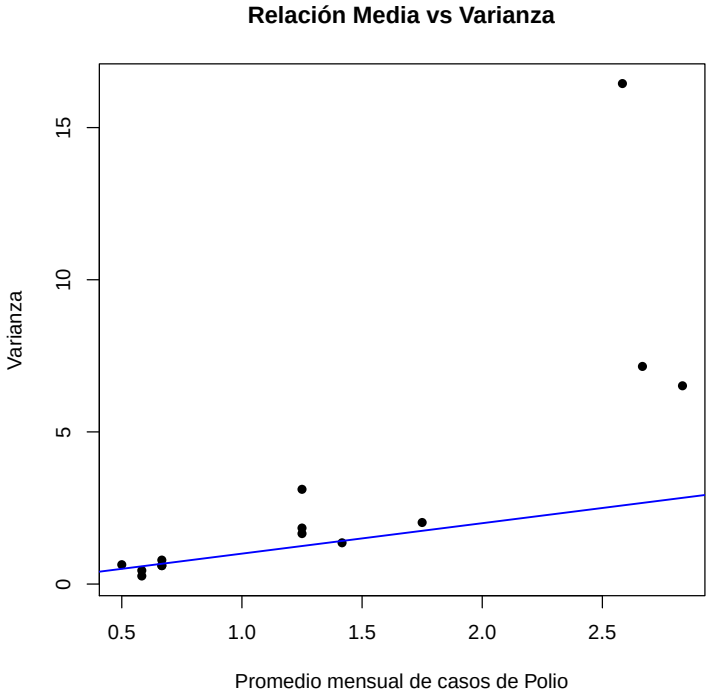
- Ejemplo. El siguiente ejemplo fue discutido en clase. Es tomado de notas de clase de Rachel MacKay Altman de la Simon Fraser University.
  - Lectura de datos y gráfica exploratoria.

```
# Polio Data: Month Count
casos  <- read.csv("c:\\...\\ModelosEstadisticosI2011\\polio.csv",header=TRUE)
attach(casos)
plot(Month,Count,pch=16,xlab="Mes",ylab="Numero de casos",main="Casos de Polio")
```



– Una gráfica que sugiere que el modelo Poisson tal vez no es adecuado por existir sobredispersión.

```
aa   <- matrix(casos[,2],12,14)
med  <- apply(aa,2,mean)
vari <- apply(aa,2,var)
plot(med,vari,pch=16,xlab="Promedio mensual de casos de Polio",
 ylab="Varianza",main="Relacion Media vs Varianza")
abline(a=0,b=1,lwd=1.5,col="blue")
```



– Ajuste del GLM Poisson y del modelo Cuasi-Poisson.

```
# GLM Poisson
out1 <- glm(Count ~ Month, family=poisson)
summary(out1)


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.626639   0.123641   5.068 4.02e-07 ***
Month       -0.004263   0.001395  -3.055  0.00225 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 343.00  on 167  degrees of freedom
Residual deviance: 333.55  on 166  degrees of freedom
AIC: 594.59

Number of Fisher Scoring iterations: 5

# Cuasi-Poisson
out2 <- glm(Count ~ Month, family=quasipoisson)
summary(out2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.626639   0.194788   3.217  0.00156 **
Month       -0.004263   0.002198  -1.939  0.05415 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for quasipoisson family taken to be 2.481976)

    Null deviance: 343.00  on 167  degrees of freedom
Residual deviance: 333.55  on 166  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Resumen de Clase 24: Lunes 2 de mayo

- Ejemplo Cuasi-Verosimilitud. El siguiente ejemplo fue discutido en clase. Es tomado del libro
  Cameron, A.C. & Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge.
  Mostramos tanto la forma fácil (usando `glm()`) como la difícil (programando el método Scoring de Fisher) (i.e. IRWLS).

```
# Ejemplo tomado del libro:
# Cameron, A.C. & Trivedi, P.K. (1998) Regression Analysis
# of Count Data. Cambridge.

casos  <- scan(
"c:\\Documents and Settings\\Rogelio\\My Documents\\ModelosEstadisticosI2011\\racd3.txt")

datos <- matrix(casos,5190,20,byrow=T)
datos <- datos[,-20]   # 5190 x 19
colnames(datos) <- c("sex","age","agesq","income",
 "levyplus","freepoor","freerepa","illness","actdays",
 "hscore","chcond1","chcond2","dvisits","nondocco","hospadmi",
 "hospdays","medicine","prescrib","nonpresc")

y   <- datos[,13]
X   <- datos[,1:12]

# Primero, un vistazo a los resumenes numericos de las variables del estudio:
table(y)
   0    1    2    3    4    5    6    7    8    9
4141  782  174   30   24    9   12   12    5    1

table(X[,1]) # sex
   0    1
2488 2702

summary(X[,2]) # age
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1900  0.2200  0.3200  0.4064  0.6200  0.7200

summary(X[,4]) # income
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2500  0.5500  0.5832  0.9000  1.5000

table(X[,5]) # levyplus
   0    1
2892 2298

table(X[,6]) # freepoor
   0    1
4968  222

table(X[,7]) # freerepa
   0    1
4099 1091
```

```
table(X[,8]) # illness
   0    1    2    3    4    5
1554 1638  946  542  274  236

table(X[,9]) # actdays
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14
4454  177  108   74   45   40   17   38   17    7   12    2    6    5  188

table(X[,10]) # hscore
   0    1    2    3    4    5    6    7    8    9   10   11   12
3026  823  446  273  187  132  104   61   42   32   21   24   19

table(X[,11]) # chcond1
   0    1
3098 2092

table(X[,12]) # chcond2
   0    1
4585  605

################################################
out1 <- glm(y ~ X, family=poisson)
summary(out1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.223848   0.189816 -11.716  < 2e-16 ***
Xsex         0.156882   0.056137   2.795  0.00520 **
Xage         1.056299   1.000780   1.055  0.29121
Xagesq      -0.848704   1.077784  -0.787  0.43102
Xincome     -0.205321   0.088379  -2.323  0.02017 *
Xlevyplus    0.123185   0.071640   1.720  0.08552 .
Xfreepoor   -0.440061   0.179811  -2.447  0.01439 *
Xfreerepa    0.079798   0.092060   0.867  0.38605
Xillness     0.186948   0.018281  10.227  < 2e-16 ***
Xactdays     0.126846   0.005034  25.198  < 2e-16 ***
Xhscore      0.030081   0.010099   2.979  0.00290 **
Xchcond1     0.114085   0.066640   1.712  0.08690 .
Xchcond2     0.141158   0.083145   1.698  0.08956 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5634.8  on 5189  degrees of freedom
Residual deviance: 4379.5  on 5177  degrees of freedom
AIC: 6737.1

Number of Fisher Scoring iterations: 6

################################################
out2 <- glm(y ~ X, family=quasipoisson)
summary(out2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.223848   0.218725 -10.167  < 2e-16 ***
Xsex         0.156882   0.064686   2.425  0.01533 *
Xage         1.056299   1.153198   0.916  0.35972
Xagesq      -0.848704   1.241930  -0.683  0.49440
Xincome     -0.205321   0.101839  -2.016  0.04384 *
Xlevyplus    0.123185   0.082551   1.492  0.13570
Xfreepoor   -0.440061   0.207197  -2.124  0.03373 *
Xfreerepa    0.079798   0.106081   0.752  0.45194
Xillness     0.186948   0.021065   8.875  < 2e-16 ***
Xactdays     0.126846   0.005801  21.868  < 2e-16 ***
Xhscore      0.030081   0.011637   2.585  0.00977 **
Xchcond1     0.114085   0.076789   1.486  0.13742
Xchcond2     0.141158   0.095808   1.473  0.14072
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for quasipoisson family taken to be 1.327793)

    Null deviance: 5634.8  on 5189  degrees of freedom
Residual deviance: 4379.5  on 5177  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
##################################################

n  <- dim(X)[1]
XX <- cbind( rep(1,n), X )
b  <- c(-2,rep(0,12)) # valores iniciales de parmetros

tolm   <- 1e-6      # tolerancia (norma minima de delta)
iterm  <- 100       # numero maximo de iteraciones
tolera <- 1         # inicializar tolera
itera  <- 0         # inicializar itera
histo  <- b         # inicializar historial de iteraciones

while( (tolera>tolm)&(itera<iterm) ){
 eta    <- XX%*%b
 U      <- as.vector(exp(eta))
 z      <- eta + (y-U)/U
 aa     <- as.vector(solve(t(XX*U)%*%XX, t(XX*U)%*%z))
 delta  <- aa-b
 b      <- aa
 tolera <- sqrt( sum(delta*delta) )
 histo  <- rbind(histo,b)
 itera  <- itera + 1 }

lsat   <- sum( y*log(y) - y - lfactorial(y) )               # logv(saturado)
lsat   <- sum( y*ifelse(y>0,log(y),0) - y - lfactorial(y) ) # logv(saturado)
eta    <- as.vector(XX%*%b)
lamb   <- exp(eta)
errstd <- sqrt(diag(solve(t(XX*lamb)%*%XX))) # errores estandar ok
```

```
lmax   <- sum( y*log(lamb) - y - lfactorial(y) )   # logv(mod interes)
lambN  <- mean(y)
lnull  <- sum( y*log(lambN) - y - lfactorial(y) )  # logv(nulo)
NullD  <- -2*( lnull - lsat )   # 5634.821   con 5190-1=5189 gl
ResD   <- -2*( lmax - lsat )    # 4379.5     con 5190-13=5177 gl
AIC    <- -2*lmax + 2*13        # 6737.083
# Estimaciones ok. Ver Cameron & Triveldi p.69

# Estimacion del parametro de dispersion
fi  <- sum( ((y-lamb)^2)/lamb )/(n-13)    # 1.327793

# Errores estandar para quasiverosimilitud (estos corresponden
# a la columna NB1 de Cameron & Triveldi p.69)

sqrt(fi)*errstd

                   sex         age       agesq      income    levyplus
0.218724882 0.064686396 1.153198186 1.241929814 0.101839386 0.082550502
   freepoor    freerepa     illness     actdays      hscore     chcond1
0.207196567 0.106080949 0.021064652 0.005800639 0.011637496 0.076788677
    chcond2
0.095808021


################### INFO ##########################
Socioeconomic:
 SEX       1 if female, 0 if male
 AGE       Age in years divided by 100
           (measured as mid-point of 10 age groups from 15-19 years to
            65-69 with 70 or more coded treated as 72)
 AGESQ     AGE squared
 INCOME    Annual income in Australian dollars divided by 1000
           (measured as mid-point of coded ranges Nil, <200, 200-1000,
            1001-, 2001-, 3001-, 4001-, 5001-, 6001-, 7001-, 8001-10000,
            10001-12000, 12001-14000, with 14001- treated as 15000

Health insurance:
 LEVYPLUS  1 if covered by private health insurance fund for private
           patient in public hospital (with doctor of choice), 0 otherwise
 FREEPOOR  1 if covered by government because low income, recent immigrant,
           unemployed, 0 otherwise
 FREEREPA  1 if covered free by government because of old-age or disability
           pension, or because invalid veteran or family of deceased
           veteran, 0 otherwise

Health status:
 ILLNESS   Number of illnesses in past 2 weeks with 5 or more coded as 5
 ACTDAYS   Number of days of reduced activity in past two weeks due to
           illness or injury
 HSCORE    General health questionnaire score using Goldberg's method.
           High score indicates bad health.
 CHCOND1   1 if chronic condition(s) but not limited in activity, 0 otherwise
 CHCOND2   1 if chronic condition(s) and limited in activity, 0 otherwise
```

```
The count variables:
 DVISITS  Number of consultations with a doctor or specialist
          in the past 2 weeks
 NONDOCCO Number of consultations with non-doctor health professionals
          (chemist, optician, physiotherapist, social worker,
           district community nurse, chiropodist or chiropractor)
           in the past 2 weeks
 HOSPADMI Number of admissions to a hospital, psychiatric hospital,
          nursing or convalescent home in the past 12 months
          (up to 5 or more admissions which is coded as 5)
 HOSPDAYS Number of nights in a hospital, etc. during most recent
          admission: taken, where appropriate, as the mid-point of the
          intervals 1, 2, 3, 4, 5, 6, 7, 8-14, 15-30, 31-60, 61-79
          with 80 or more admissions coded as 80. If no admission
          in past 12 months then equals zero.
 MEDICINE Total number of prescribed and nonprescribed medications used
          in past 2 days
 PRESCRIB Total number of prescribed medications used in past 2 days
 NONPRESC Total number of nonprescribed medications used in past 2 days
```

- Distribución Binomial Negativa. Supongamos ensayos Bernoulli, con probabilidad de éxito p. Si Y es el número de éxitos observados hasta obtener r fracasos, entonces X tiene la siguiente distribución:

$$P(Y=y) = \begin{pmatrix} y+r-1 \\ r-1 \end{pmatrix} p^y (1-p)^r, \qquad y = 0, 1, 2, \cdots$$

con $\mathsf{E}(Y) = pr/(1-p)$ y $\mathsf{Var}(Y) = pr/(1-p)^2$. Para nuestros fines, es más conveniente una reparametrización tal que

$$\mathsf{E}(Y) = \mu \qquad \text{y} \qquad \mathsf{Var}(Y) = \mu(1+\alpha\mu)$$

esto se logra haciendo $r = \alpha^{-1}$ y $p = \mu/(\mu+\alpha^{-1})$. La forma correspondiente de la Binomial Negativa es

$$P(Y=y) = \begin{pmatrix} y+\alpha^{-1}-1 \\ \alpha^{-1}-1 \end{pmatrix} \left(\frac{\mu}{\mu+\alpha^{-1}}\right)^y \left(\frac{\alpha^{-1}}{\mu+\alpha^{-1}}\right)^{\alpha^{-1}}, \qquad y = 0, 1, 2, \cdots$$

La distribución Binomial Negativa es una distribución con soporte $0, 1, 2, \cdots$ (igual que la Poisson), con media μ y varianza $\mu(1+\alpha\mu)$ (ligeramente distinta a la Poisson), por lo tanto es un modelo que es usado como un modelo para conteos, alternativo a la Poisson, pero que incorpora explícitamente el fenomeno de sobredispersión.

En el caso α conocido, la Binomial Negativa es un miembro de la familia exponencial, sin embargo raramente se usa este hecho, por ejemplo, la liga canónica no es muy interpretable, así que en aplicaciones se usa la liga log, esto es $\log(\mu) = x^T\beta$, además de que, por supuesto, difícilmente α es conocido.

- Estimación en la Binomial Negativa. Observando que

$$\begin{pmatrix} y+\alpha^{-1}-1 \\ \alpha^{-1}-1 \end{pmatrix} = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(\alpha^{-1})y!} = \frac{1}{y!}\prod_{j=0}^{y-1}(j+\alpha^{-1})$$

es directo ver que la logverosimilitud esta dada por

$$l(\alpha,\beta) = \sum_{i=1}^{n}\left\{\sum_{j=0}^{y_i-1}\log(j+\alpha^{-1}) - \log(y_i!) + y_i\log\alpha + y_i x_i^T\beta - (y_i+\alpha^{-1})\log(1+\alpha\exp(x_i^T\beta))\right\}$$

y derivando con respecto a β y α e igualando a 0 se obtienen las ecuaciones

$$A \equiv \frac{\partial l(\alpha, \beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{1 + \alpha \mu_i} x_i = 0$$

$$B \equiv \frac{\partial l(\alpha, \beta)}{\partial \alpha} = \sum_{i=1}^{n} \left\{ \frac{1}{\alpha^2} \left(\log(1 + \alpha \mu_i) - \sum_{j=0}^{y_i - 1} \frac{1}{(j + \alpha^{-1})} \right) + \frac{y_i - \mu_i}{\alpha(1 + \alpha \mu_i)} \right\} = 0$$

Un posible algoritmo para encontrar una solución es

i. Inicializar α^0.

ii. Resolver A usando IRWLS pues con α^0 conocido, tenemos un GLM usual; con esto obtenemos β^0.

iii. Dado β^0, resolver la ecuación univariada B y obtener α^1.

iv. Iterar ii y iii, con el valor actualizado de α.

Nota: La varianza asintótica es diagonal en bloques y, usando el hessiano correspondiente, se obtienen las expresiones para las varianzas de los estimadores.

Resumen de Clase 25: Miércoles 4 de mayo

- Transformaciones Estabilizadoras de Varianza. En las clases anteriores, en el tema de cuasi-Verosimilitud, hemos enfatizado el incorporar un modelo para la varianza en el proceso de inferencia. Ahora vamos a regresarnos un poco y comentar sobre el como evitar el incorporar la heterogeneidad de varianza. Esto es, retomaremos el tema de regresión lineal, pero ahora veremos algunas ideas sobre que hacer cuando algunos de los supuestos usuales no se cumplen.

  Supongamos una variable aleatoria $y \sim N(\mu, \sigma^2)$, donde $\sigma^2 = h(\mu)$. Queremos una transformación de y, $y^* = \phi(y)$, que estabilize su varianza (i.e. que la varianza no dependa de la media). Note que

  $$\text{Var}(y^*) = \text{Var}(\ \phi(y)\ ) \doteq \text{Var}\ (\ \phi(\mu) + \phi'(\mu)(y-\mu)\ ) = [\phi'(\mu)]^2\, \text{Var}(y) = [\phi'(\mu)]^2\, h(\mu)$$

  queremos que $\text{Var}(y^*) = C$, entonces ϕ debe satisfacer

  $$\phi'(\mu) = \frac{C}{\sqrt{h(\mu)}}, \qquad \text{esto es} \qquad \phi(\mu) = C \int \frac{d\mu}{\sqrt{h(\mu)}}$$

  Note que $y^* = \phi(y) \doteq \phi(\mu) + \phi'(\mu)(y-\mu)$, entonces

  $$y^* \sim N(\phi(\mu), C) \equiv N(\mu^*, C)$$

  donde μ^* es modelada en términos de las covariables.

- Transformación Raíz Cuadrada. Si sospechamos que la varianza cambia proporcionalmente con la media, i.e. si $\sigma^2 \propto \mu$, o $\sigma^2 = k\mu$, entonces

  $$\phi(\mu) = C \int \frac{d\mu}{\sqrt{k\mu}} = k_1 \sqrt{\mu} + k_2$$

  esto implica que la transformación $\phi(y) = \sqrt{y}$ estabiliza la varianza en el caso cuando la varianza es proporcional a la media.

- Transformación Arcoseno. En ocasiones, al analizar datos que estan en forma de proporciones, es común utilizar la transformación Arcoseno; la razón de ello es que la varianza de una proporción es de la forma $p(1-p)/k$ y entonces

  $$\phi(p) = C \int \frac{dp}{\sqrt{p(1-p)/k}} = K\ \text{Arcoseno}(\sqrt{p})$$

- Transformaciones a normalidad. La familia de transformaciones Box-Cox es de la forma

  $$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \text{si} \quad \lambda \neq 0 \\ \log y & \text{si} \quad \lambda = 0 \end{cases}$$

  Supongamos que el modelo lineal usual se cumple para alguna λ, esto es

  $$y^{(\lambda)} \sim N_p(X\beta, \sigma^2 I)$$

  (aquí estamos abusando un poco de la notación, $y^{(\lambda)}$ es un vector, pero en la definición de la transformación, $y^{(\lambda)}$ es un escalar). Ahora bien, ¿Cuál es la distribución conjunta de mis datos?. Tenemos la

distribución de $y^{(\lambda)}$, pero no la de y, pero y es una transformación de $y^{(\lambda)}$, digamos $y = T(y^{(\lambda)})$, de aquí que, la conjunta de y está dada por el teorema de cambio de variable:

$$f(y) = f_\lambda(T^{-1}(y))|J_{T^{-1}}| = f_\lambda(y^{(\lambda)}) \left| \prod_{i=1}^{n} y_i^{\lambda-1} \right|$$

la última expresión es porque la matriz jacobiana es diagonal y el determinante es simplemente el producto de los elementos diagonales, los cuales son de la forma $\partial[(y^\lambda - 1)/\lambda]/\partial y = y^{\lambda-1}$, (las paralelas indican valor absoluto). Por lo tanto:

$$f(y) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}(y^{(\lambda)} - X\beta)^T(y^{(\lambda)} - X\beta) \right\} \left| \prod_{i=1}^{n} y_i^{\lambda-1} \right|$$

Esta densidad conjunta es precisamente la verosimilitud de nuestros datos, por lo tanto, para hacer inferencia maximizamos esta función con respecto a β y λ.

- Elección del parámetro λ. Maximizamos la verosimilitud en dos etapas, fijamos λ y estimamos β y σ^2, luego obtenemos la verosimilitud perfil para λ.

  Para una λ dada tenemos

  $$\widehat{\beta}(\lambda) = (X^T X)^{-1} X^T y^{(\lambda)} \qquad \text{y} \qquad \widehat{\sigma}^2(\lambda) = \frac{1}{n}\text{SCE}(\lambda)$$

  La verosimilitud perfil es

  $$L_P(\lambda) = \frac{e^{-n/2}}{(2\pi)^{n/2}(\widehat{\sigma}^2(\lambda))^{n/2}} \left| \prod_{i=1}^{n} y_i^{\lambda-1} \right|$$

  Sacando logaritmo:

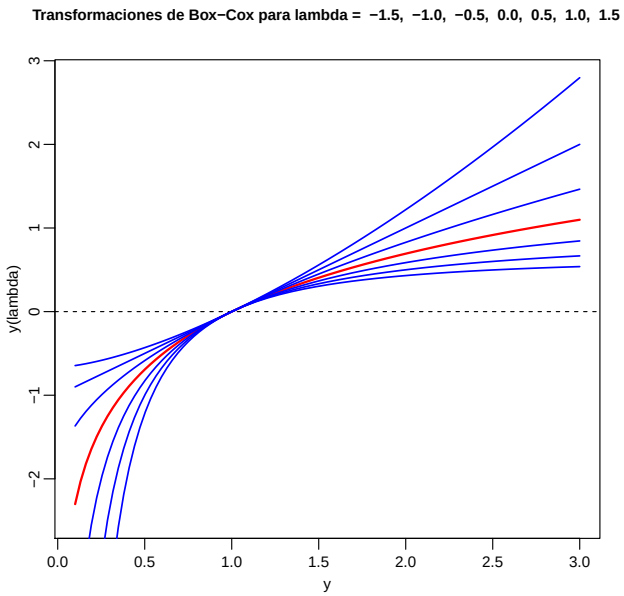
  $$l_P(\lambda) = C + (\lambda - 1)\sum_{i=1}^{n} \log(y_i) - \frac{n}{2}\log(\widehat{\sigma}^2(\lambda))$$

- Referencias. La referencia original y algunas discusiones:
  - Box & Cox (1964). An analysis of transformations. *J. R. Stat. Soc. B*, **Vol** 26, 211-246.
  - Bickel & Doksum (1981). An analysis of transformations revisited. *J. Am. Stat. Assoc.*, **Vol** 76, 296-311.
  - Box & Cox (1982). An analysis of transformations revisited, rebutted. *J. Am. Stat. Assoc.*, **Vol** 77, 209-210.

- Ejemplo. El siguiente ejemplo es el caso más sencillo de regresión, estamos suponiendo que no hay covariables asi que sólo tenemos el intercepto. Se muestran cuatro gráficas, la primera ilustra diferentes miembros de la familia de transformaciones Box-Cox, la segunda es la verosimilitud perfil para λ y las últimas dos muestra histogramas de los datos sin transformar y despues de la transformación.

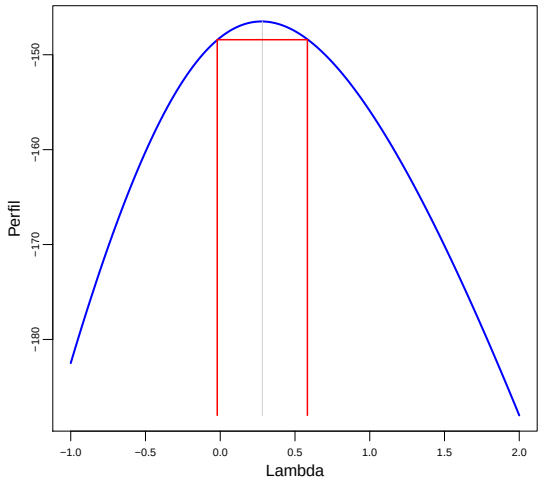
```
yt <- function( y, lam ){
    if( lam != 0 ) {return( (-1+y^lam)/lam )}
    else {return(log(y))}}
y  <- c(seq( .1,1,length=30 ),seq( 1,3,length=30) )
plot( y, yt(y,0), type = "l", lwd=2, col="red",ylab="y(lambda)", xlab="y",
    ylim=c(-2.5,2.8), mgp=c(1.5,.5,0), cex.axis=.8, cex.lab=.8)
abline(h=0, lty=2)
title( cex.main=.8, main=
```

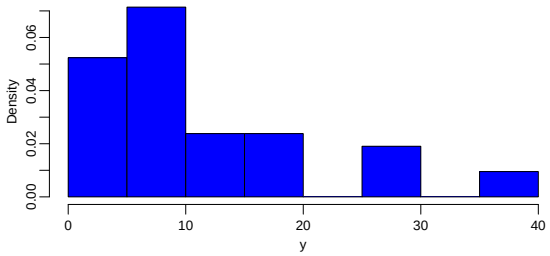
```
"Transformaciones de Box-Cox para lambda =  -1.5,  -1.0,  -0.5,  0.0,  0.5,  1.0,  1.5")

lines( y, yt(y,-1.5), lwd=1.5, col="blue" )
lines( y, yt(y,-1.0), lwd=1.5, col="blue" )
lines( y, yt(y,-0.5), lwd=1.5, col="blue" )
lines( y, yt(y, 0.5), lwd=1.5, col="blue" )
lines( y, yt(y, 1.0), lwd=1.5, col="blue" )
lines( y, yt(y, 1.5), lwd=1.5, col="blue" )
```

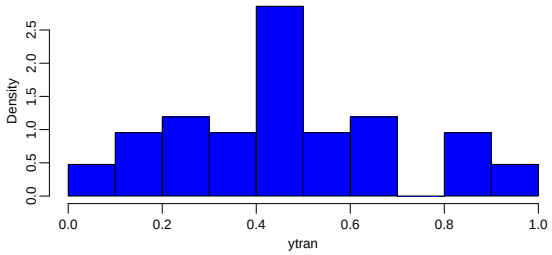



```
# Datos
y  <- c(15,  9, 18, 10,  5, 12,  8,  5,  8, 10,  7,  2,  1, 10,
        10, 10,  2, 10,  1, 40, 10,  5,  3,  5, 15, 10, 15,  9,
         8, 18, 10, 20, 11, 30,  2, 20, 20, 30, 30, 40, 30,  5)
tst <- shapiro.test(y)
# Shapiro-Wilk normality test
# W = 0.8579, p-value = 9.902e-05  (normalidad rechazada)
n    <- length(y)
m    <- 200
lam  <- seq(-1,2,length=m)
perf <- rep(0,m)
cte  <- -n*(1+log(2*pi))/2
for(j in 1:m){
 ylam    <- (y^(lam[j])-1)/lam[j]
 s2      <- (n-1)*var(ylam)/n
 perf[j] <- cte-n*log(s2)/2+(lam[j]-1)*sum(log(y))}
plot(lam,perf,xlab="Lambda",ylab="Perfil",lwd=2,
 mgp=c(1.5,.5,0), col="blue", cex.axis=.7, type="l")
lambda <- lam[ perf==max(perf) ]
segments(lambda,min(perf),lambda,max(perf),col=gray(.8))
linf  <- max(perf) - qchisq(.95,1)/2
rango <- lam[(abs(perf-linf)<.1)]
```

```
segments(rango[1],linf,rango[2],linf,col="red",lwd=1.5)
segments(rango[1],min(perf),rango[1],linf,col="red",lwd=1.5)
segments(rango[2],min(perf),rango[2],linf,col="red",lwd=1.5)
inter <- c(rango[1],lambda,rango[2]) # -0.0201  0.2814  0.5829
ytran <- (y^(lambda)-1)/lam[j]
tst   <- shapiro.test(ytran)
# Shapiro-Wilk normality test
# W = 0.9652, p-value = 0.2262
par(mfrow=c(2,1), mar=c(3, 3, 2, 2) )
hist(y, cex.axis=.8, cex.lab=.8, mgp=c(1.5,.5,0), main="", prob=T,
 col="blue", nclass=9)
hist(ytran, cex.axis=.8, cex.lab=.8, mgp=c(1.5,.5,0), main="",
 prob=T, col="blue", nclass=9 )
```







Tarea 3. Modelos Estadísticos I

1. La distribución Pareto tiene sus orígenes en la modelación de la distribución de la riqueza en una sociedad capitalista. La densidad Pareto está dada por

$$f(y;\theta) = \frac{\theta}{y^{\theta+1}}, \qquad y > 1$$

esta densidad refleja la idea de que una fracción pequeña de la sociedad posee una gran parte de la riqueza (La "ley de Pareto" o "regla 80-20" está asociada a esta distribución).

   (a) Muestre que la Pareto pertenece a la familia exponencial.

   (b) Escriba todos los elementos de un modelo lineal generalizado, así como todos los elementos necesarios para efectuar estimación vía mínimos cuadrados ponderados iterativamente.

2. Considere nuevamente la distribución Pareto del problema anterior.

   (a) Encuentre el estadístico $\mathcal{U}$ (score) y la información Var($\mathcal{U}$). Verifique explícitamente que E($\mathcal{U}$) $= 0$.

   (b) Use la distribución asintótica del estimador de máxima verosimilitud para construír un intervalo de aproximadamente un 95% de confianza.

   (c) Si U es una variable uniforme entre 0 y 1. Muestre que $U^{-1/\theta} \sim$ Pareto(θ).

   (d) Efectúe un estudio de simulación para evaluar el desempeño del intervalo de confianza asintótico. Use $\theta = 2$.

3. Los siguientes datos provienen de cierto estudio

x:	1.0	1.2	1.4	1.6	1.8	2.0
y:	3.15	4.85	6.50	7.20	8.25	16.50

   Considere el modelo

$$\mathsf{E}(y) = \log(\beta_0 + \beta_1 x + \beta_2 x^2)$$

   y suponga que hay evidencia de que es un modelo razonable para el fenómeno subyacente. Especificando los supuestos necesarios, ofrezca un análisis de este conjunto de datos. Este problema nos da un indicio de que una parte del área de Regresión Nolineal es un caso especial de los Modelos Lineales Generalizados.

4. Las propiedades asintóticas del estimador de máxima verosimilitud nos dicen que $V_n^{1/2}\,(\widehat{\beta} - \beta) \xrightarrow{d} N_p(0, I)$, donde V_n es la matriz de Información y V_n^{-1} es la llamada varianza asintótica. En general, la matriz de información depende del valor verdadero de β; cuando esto no es así, (esto es, cuando V_n no depende de β) los procedimientos inferenciales son más sencillos.

   (a) Considere el modelo lineal usual. Calcule la varianza asintótica de $\widehat{\beta}$. Observe que no depende de β.

   (b) En los GLM's es posible lograr que la matriz de información no dependa de β si usamos una función liga, $g(\mu)$, apropiada. Suponga que pedimos que la función liga satisfaga $[\,g'(\mu)\,]^2 b''(\theta) = 1$. Para este caso, muestre que la varianza asintótica no depende de β.

   (c) Considere el modelo Poisson. La liga canónica es $g(\mu) = \log(\mu)$ pero esta no tiene la propiedad mencionada. Encuentre una función liga, $g(\mu)$, tal que la varianza asintótica no dependa de β. Note la similaridad con las transformaciones estabilizadoras de varianza.

5. Considere el conjunto de datos `ldeaths`, disponibles en la distribución base de R. Son datos de el número de muertes por mes debidas a deficiencias pulmonares en Gran Bretaña durante un período de varios años. Un posible modelo considera tratar estos datos como realizaciones de una variable Poisson con una componente estacional y otra componente con la tendencia a largo plazo:

$$\mathsf{E}(\texttt{deaths}_i) = \beta_0 + \beta_1 t_i + \alpha \mathsf{Sen}(2\pi\, s_i/12 + \delta)$$

donde β_0, β_1, α y δ son parámetros, t_i es el tiempo desde el inicio del estudio y s_i es el mes del año.

   (a) Use las propiedades básicas de las funciones trigonométricas para escribir este modelo como un GLM. Ajústelo usando `glm()`. Use `as.numeric(ldeaths)` para que los datos esten en un vector numérico estándar en vez de un objeto de R del tipo de series de tiempo.

   (b) Grafique los datos como una serie temporal y grafique encima el ajuste de los datos.

   (c) Comente acerca del ajuste y dé sus conclusiones.

6. Sean $y_i, \cdots, y_n$, variables aleatorias independientes, con $\mathsf{E}(y_i) = \mu_i$ y varianza $\mathsf{Var}(y_i) = \phi V(\mu_i)$, donde V es una función conocida. Suponga que g es una función monótona y diferenciable, tal que $g(\mu_i) = \eta_i = x_i^T \beta$. Definamos

$$z_i = g'(\mu_i)(y_i - \mu_i) + \eta_i \qquad \text{y} \qquad w_i = \left\{V(\mu_i)[g'(\mu_i)]^2\right\}^{-1}$$

   (a) Muestre que $\mathsf{E}(z_i) = x_i^T \beta$.

   (b) Muestre que la matriz de covarianza de $z = (z_1, \cdots, z_n)^T$ es $\phi\, W^{-1}$, donde W es una matriz diagonal con $W_{ii} = w_i$.

   (c) Si β es estimado mediante la minimización de $\sum_{i=1}^n w_i (z_i - x_i^T \beta)^2$, muestre que la matriz de covarianza del estimador resultante, $\widehat{\beta}$, es $\phi\,(X^T W X)^{-1}$ y encuentre $\mathsf{E}(\widehat{\beta})$.

   (d) La versión multivariada del teorema central del límite nos dice que, conforme la dimensión de z tiende a infinito, $X^T W z$ va a tender a una normal multivariada. ¿Qué implicación tiene esto sobre la distribución de $\widehat{\beta}$ para muestras grandes?. ¿Es $\widehat{\beta}$ el estimador de máxima verosimilitud?.

7. El conjunto de datos `harrier` se ecuentra disponible en el paquete `gamair`. Son 37 registros de densidad de "Grouse" (un cierto tipo de aves) presentes en una región (la densidad medida en aves por km^2) y la tasa de consumo de este tipo de aves por cierto tipo de halcón "Hen Harrier". La teoría ecológica sugiere que el consumo, c, esperado, debería estar relacionado con la densidad, d, de Grouse mediante el modelo

$$\mathsf{E}(c_i) = \frac{a d_i^m}{1 + a t d_i^m}$$

donde a, t y m son parámetros desconocidos. Se espera que la varianza de la tasa de consumo sea proporcional a la tasa media de consumo.

   (a) Muestre que, para m fija, se puede obtener un GLM que relacione c_i con d_i mediante una liga recíproca.

   (b) Para $m = 1$ estime el modelo usando `glm()` con la familia `quasi`.

   (c) Grafique los residuales del modelo contra la densidad de Grouse e interprete esta gráfica.

   (d) Determine un valor adecuado para m. Una posibilidad es calcular la devianza para un rango de valores de m y tomar aquel valor de m que haga mínima la devianza.

   (e) Una vez decidido cual es el mejor modelo, construya una gráfica mostrando la curva de consumo predicho contra densidad. Añada los datos observados a la gráfica. Usando la función `predict()` (con el argumento `se` igual a `TRUE`), añada intervalos de aproximadamente un 95% de confianza.

   (f) Forma alternativa para estimar los parámetros: Obtenga la cuasi-logverosimilitud y maximízela usando algún optimizador de R.

Fecha de entrega: Viernes 20 de mayo. De los incisos 2d, 5b, 5c, 7e y 7f, entregar solo 2 (el resto de la tarea es, por supuesto, obligatorio).

Resumen de Clase 26: Lunes 9 de mayo

- Modelos No–Lineales. Una tarea importante en estadística, es el tratar de establecer relaciones entre variables. En general, el proceso de observación de un fenómeno está sujeto a ruidos de diversas naturalezas tales como, errores de observación, variabilidad biológica, o el simple desconocimiento de las relaciones funcionales subyacentes. Los modelos de regresión lineal

$$y = x^T\beta + e$$

son una herramienta poderosa para cuantificar el impacto de covariables, x, sobre una respuesta, y. Los modelos de regresión no–lineal extienden estos modelos a casos en los que se cuenta con información acerca de la naturaleza de la relación funcional subyacente (típicamente una función no–lineal):

$$y = f(x;\theta) + e$$

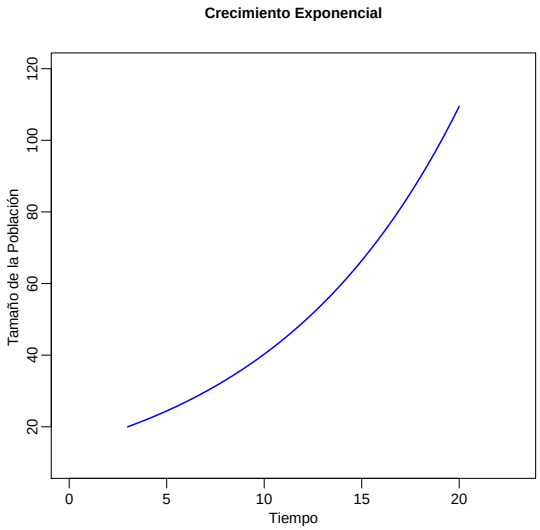
- Ejemplo: Modelos de Crecimiento. Los organismos más simples se reproducen mediante la partición binaria de sus células. Si t denota tiempo y y el tamaño de la población, esto lleva a un modelo exponencial de crecimiento

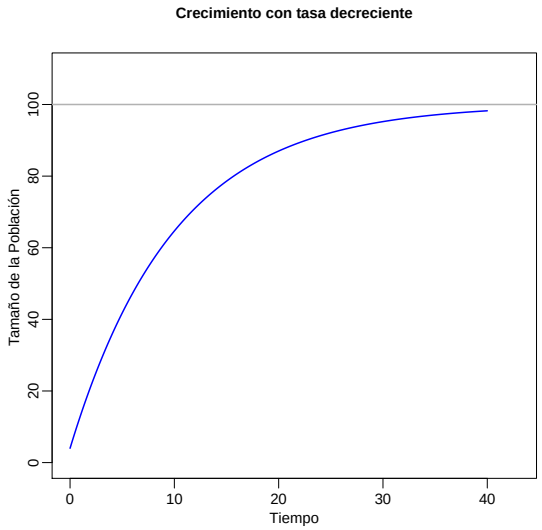
$$\frac{dy}{dt} = \kappa y, \quad \text{o} \quad y = \alpha e^{\kappa(t-\gamma)}$$

Si suponemos que la tasa de crecimiento es proporcional al tamaño remanente, entonces

$$\frac{dy}{dt} = \kappa(M - y), \quad \text{o} \quad y = M - (M-\alpha)e^{-\kappa t}$$

Comportamientos típicos se muestran en las siguientes dos gráficas.

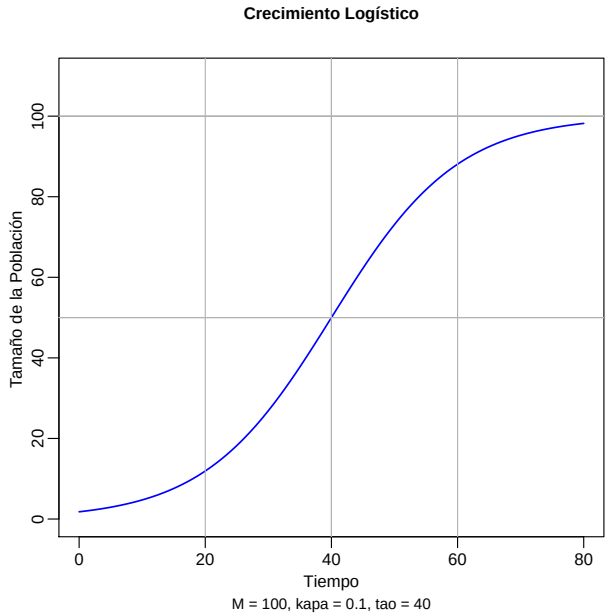




- Modelo Logístico de Crecimiento. Para muchos tipos de datos de crecimiento se tiene una combinación de crecimiento acelerado hasta cierto punto, seguido de un crecimiento desacelerado. Esto es,

$$\frac{dy}{dt} = \frac{\kappa}{M}y(M-y), \quad \text{o} \quad y = \frac{Me^{\kappa(t-\tau)}}{1+e^{\kappa(t-\tau)}}$$

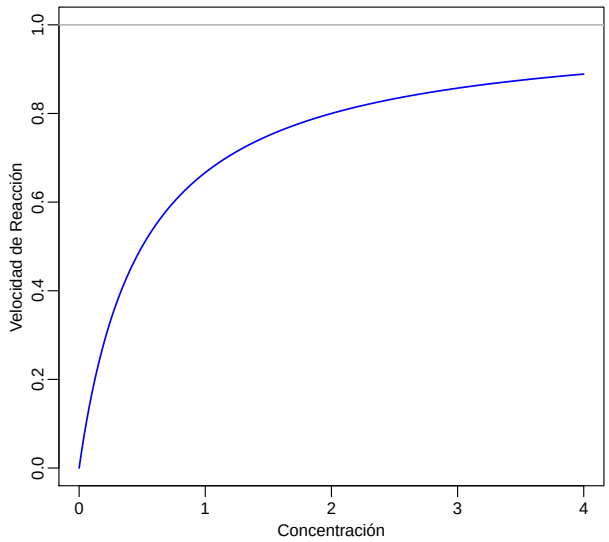
donde $y(t) \to M$ cuando $t \to \infty$, además, cuando $t = \tau$ se tiene que se alcanza el 50% del tamaño máximo, esto es, $M/2$. El modelo logístico, con su característico comportamiento sigmoidal es muy utilizado como modelo de crecimiento. El comportamiento del modelo logístico es simétrico con respecto al tiempo τ; si la población bajo estudio no tiene esta característica pudiera usarse el modelo Gompertz.



- Cinética de Enzimas. El modelo de Michaelis–Menten es uno de los modelos más usados en ciencias biológicas, particularmente, en el estudio de reacciones enzimáticas, las cuales son de importancia en investigación en medicina, biología y farmacología. El modelo es usado para describir funciones de saturación en muchos fenómenos biológicos y físicos. La forma más simple del modelo es

$$E(y \mid x) = f(x;\theta) = \frac{\theta_1 x}{\theta_2 + x}$$

donde y es la velocidad de reacción, x es la concentración de un substrato, θ_1 es la velocidad máxima obtenible en esta reacción y θ_2 es un parámetro relacionado con la concentración a la cual se alcanza un 50% de la velocidad máxima.



- Ajuste de Modelos. Consideremos el modelo

$$y_i = f(x_i;\theta) + e_i, \quad i = 1, \cdots, n$$

con $e_i \sim$ i.i.d.$N(0, \sigma^2)$. En este caso, máxima verosimilitud y mínimos cuadrados conducen a estimar θ mediante el minimizador de

$$SCR(\theta) = \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2$$

Esta función puede ser minimizada directamente por cualquier minimizador (por ejemplo en R, `optim` o `nlminb`); aquí consideraremos (en parte por tradición) el método **Gauss–Newton**, basado en una linealización del modelo $f(x_i;\theta)$.

- Gauss–Newton. Reescribimos el modelo para todos los datos en forma vectorial

$$y = f(x;\theta) + e, \quad \text{donde} \quad y,\ f \text{ y } e \quad \text{son vectores } n \times 1$$

Sea θ^0 un vector de valores iniciales para los parámetros. Una aproximación de primer orden alrededor de θ^0 para f es

$$f(x;\theta) = f(x;\theta^0) + F(\theta^0)(\theta - \theta^0)$$

donde $F(\theta^0)$ es una matriz $n \times p$ cuyo i–ésimo renglón es $\partial f(x_i;\theta^0)/\partial\theta^T$. El modelo es entonces

$$y - f(x;\theta^0) = F(\theta^0)\delta + e, \quad \text{donde} \quad \delta = \theta - \theta^0.$$

obtenemos una estimación, δ^1, de δ mediante mínimos cuadrados ordinarios y hacemos

$$\theta^1 = \theta^0 + (F^T F)^{-1} F^T (y - f(x;\theta^0))$$

iteramos este procedimiento hasta que, por ejemplo, $||\delta^k|| < \tau$ para cierta tolerancia τ.

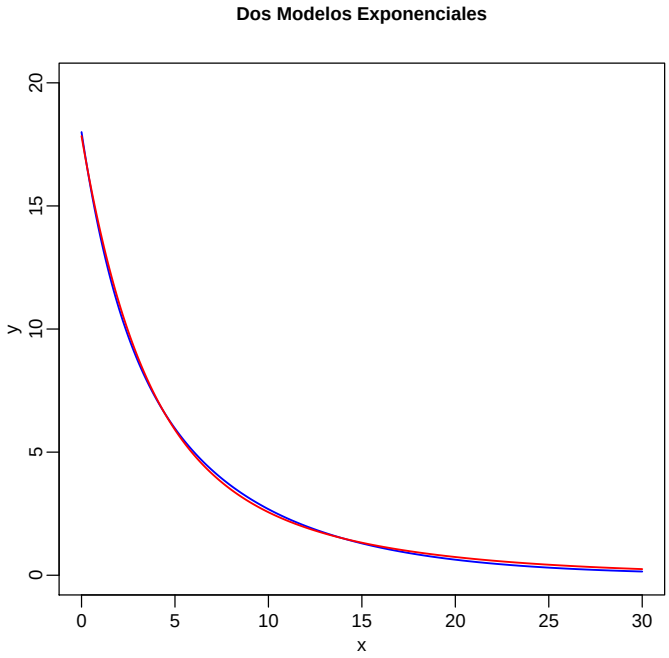
- No Identificabilidad Aproximada. Modelos de la forma

$$y = \sum_{j=1}^{k} \alpha_j e^{-\beta_j x} + e$$

aparecen en muchas areas, por ejemplo, en modelos de compartimentos en estudios de farmacocinética. La gráfica en la siguiente hoja muestra gráficas superpuestas de dos modelos con sumas de exponenciales. Son virtualmente indistinguibles, sin embargo, los parámetros correspondientes son muy diferentes. Para el ajuste de estos modelos, la decisión sobre que puntos observar (diseño experimental) se vuelve crucial.

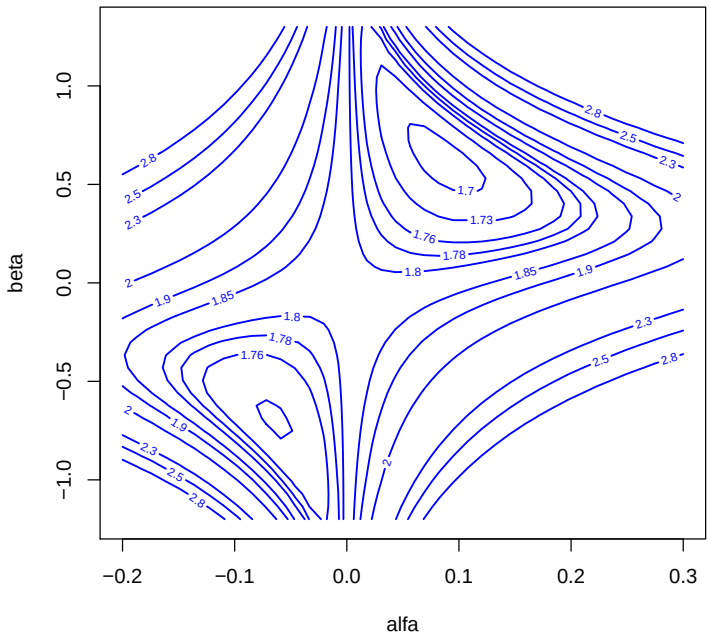
Las funciones graficadas son

$$f_1(x) = 7\, e^{-x/2} + 11\, e^{-x/7}$$
$$f_2(x) = 11.8\, e^{-x/3.1} + 6.06\, e^{-x/9.4}$$



- SCR: Modelo Exponencial mal–condicionado. Considere el modelo $y = \alpha \exp(-\beta x) + e$ el cual es observado en cuatro puntos (datos ficticios). El ajuste de este modelo (para este conjunto de datos) presenta óptimos locales múltiples. En este caso particular, la razón de este comportamiento es que simplemente el modelo exponencial no es adecuado para estos datos. El punto del ejemplo es que en modelos no-lineales, a diferencia del caso lineal, la superficie definida por la suma de cuadrados residual puede ser difícil de minimizar.

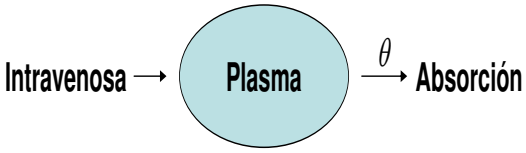
```
# Seber & Wild p.92
# Superficie de Suma de Cuadrados Residual con dos minimos relativos.

x   <- c(-2, -1,   1, 2 )
y   <- c( 0,  1, -.9, 0 )
# Estos datos no corresponden a un experimento real, solo son para ilustrar
# el caso potencial de soluciones multiples en un problema de regresion nolineal.
#
# Graficas de Contornos para SCR(teta).
N   <- 40
a   <- seq(-.2, .3, length=N)
b   <- seq(-1.2, 1.3, length=N)
SCR <- matrix(0,N,N)
for( i in 1:N ){ for( j in 1:N ){
  SCR[i,j] <- sum((y - a[i]*exp(-b[j]*x))^2) }}

contour( a, b, SCR, xlab= "alfa", ylab="beta", lwd=1.5, col="blue",
         levels = c(1.70,1.73,1.76,1.78,1.8,1.85, 1.9,2,2.3,2.5,2.8) )
```



- Modelos de Compartimientos en Farmacocinética. La farmacocinética estudia la evolución en el tiempo de un medicamento en el organismo (absorción, distribución y eliminación):
  - Administración oral, intravenosa, tópica, etc.
  - La sangre transporta el medicamento a los tejidos y órganos del cuerpo.
  - El medicamento es metabolizado o eliminado por excreción renal o biliar.

  Los modelos compartimentales consideran al cuerpo (o sistema) dividido en compartimientos y, típicamente, se asume que las tasas de transferencia de un medicamento siguen una cinética de primer orden; esto es, que el cambio en la concentración es proporcional a la concentración presente.

- Un Compartimiento.

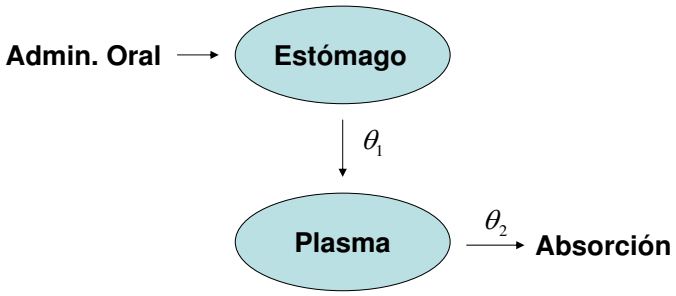


Cinética de primer orden:

$$\frac{dC(t)}{dt} = -\theta C(t) \quad \Rightarrow \quad C(t) = C_0 e^{-\theta t}$$

- Dos Compartimientos.

$$\begin{array}{rcl} \frac{dA(t)}{dt} & = & -\theta_1 A(t) \\ \frac{dC(t)}{dt} & = & \theta_1 A(t) - \theta_2 C(t) \end{array} \quad \Rightarrow \quad C(t) = \frac{A_0\theta_1}{\theta_1 - \theta_2}\left(e^{-\theta_2 t} - e^{-\theta_1 t}\right)$$



- Inferencia en Regresión No–Lineal. Consideremos el modelo

$$y_i = f(x_i;\theta) + e_i, \quad i = 1, \cdots, n$$

con $e_i \sim$ i.i.d.$N(0,\sigma^2)$. Entonces, (Ver Gallant, (1987))

$$\widehat{\theta} - \theta \approx N_p(0, \sigma^2 C^{-1}), \quad \text{donde} \quad C = F^T(\theta)F(\theta)$$
$$(n-p)CME/\sigma^2 \approx \chi^2_{n-p}$$
$$\widehat{\theta} \text{ y } CME \text{ son independientes}$$

de aquí se tiene que, intervalos de confianza para combinaciones lineales de los parámetros pueden construírse como:

$$a^T\widehat{\theta} \pm t_{n-p,\alpha/2}\sqrt{CMEa^TC^{-1}a}$$

Algunas Referencias

1. Bates, D. M. & Watts, D. G. (1988). *Nonlinear Regression Analysis and its Applications*. Wiley.
2. Gallant, A. R. (1987). *Nonlinear Statistical Models*. Wiley.
3. Huet, S., Bouvier, A., Gruet, M. A. & Jolivet, E. (1996). *Statistical Tools for Nonlinear Regression*. Springer–Verlag.
4. Seber, G. A. F. & Wild, C. J. (1989). *Nonlinear Regression*. Wiley.

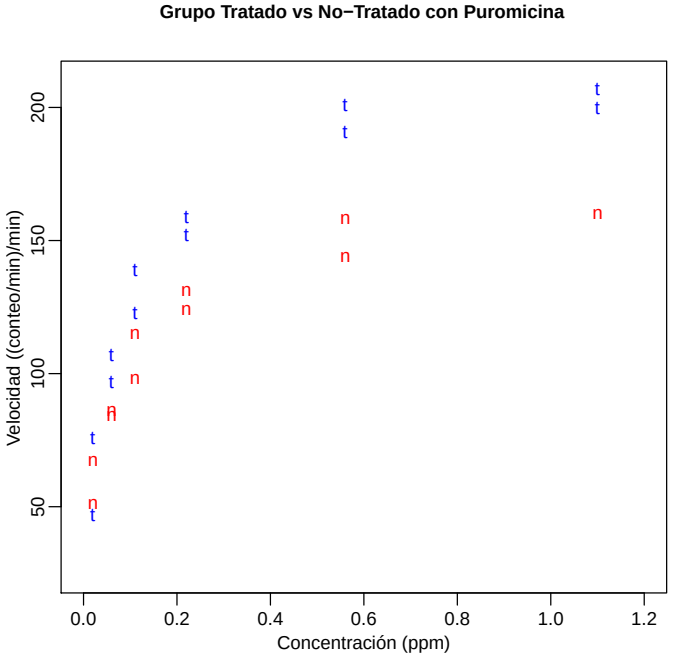
Resumen de Clase 27: Miércoles 11 de mayo

- Ejemplo: Regresión No-Lineal. Análisis de datos sobre reacciones enzimáticas tomado del libro de Bates & Watts. Datos de "velocidad" de una reacción enzimática; obtenidos por M.A. Treloar (1974) "Effects of puromycin on galactosyltransferase of Golgi membranes", Tesis de Maestría, Universidad de Toronto.

  Se reportan conteos por minuto de un material radioactivo producto de cierta reacción; a partir de estos conteos se deduce una "velocidad" inicial de reacción. El objetivo es relacionar esta velocidad inicial con la concentración del sustrato (ppm). Se supone que la velocidad depende de la concentración de acuerdo a un modelo Michaelis-Menten. Se hipotetiza que la velocidad final debe de depender de la presencia de Puromicina, pero no depende del parámetro de (velocidad final)/2. El experimento fue conducido primero con las enzimas tratadas con puromicina y luego con enzimas sin tratamiento.

  Las variables registradas son:

`vel`	$=$	(conteo/min^2)
`conc`	$=$	concentración del sustrato (ppm)
`trat`	$=$	1 si tratado con puromicina, $=$ 0 si no tratado.




```
vel  <- c(76,47,97,107,123,139,159,152,191,201,207,
          200,67,51,84,86,98,115,131,124,144,158,160)
conc <- c(2,2,6,6,11,11,22,22,56,56,110,110,2,2,6,6,11,11,22,22,56,56,110)
conc <- conc/100
trat <- c(rep(1,12),rep(0,11))
puro <- data.frame(trat,conc,vel)
ranx <- range(conc)
rany <- range(vel)
```

```
# Grafica de los datos
plot( conc[trat==1], vel[trat==1], xlim=c(0,1.2), ylim=c(25,210),
    xlab = "Concentracion (ppm)", mgp=c(1.5,.5,0),
    ylab = "Velocidad ((conteo/min)/min)", col="blue",
    cex = .9, cex.lab=.9, cex.axis=.9, pch = "t", cex.main=.9
    main = "Grupo Tratado vs No-Tratado con Puromicina")
points( conc[trat==0], vel[trat==0], cex=.9, pch="n", col="red" )
```

- Modelo Michaelis-Menten. Este modelo aparece en la solución de ciertas ecuaciones diferenciales que modelan la dinámica de reacciones enzimáticas. El modelo que relaciona la velocidad inicial de reacción con la concentracción es

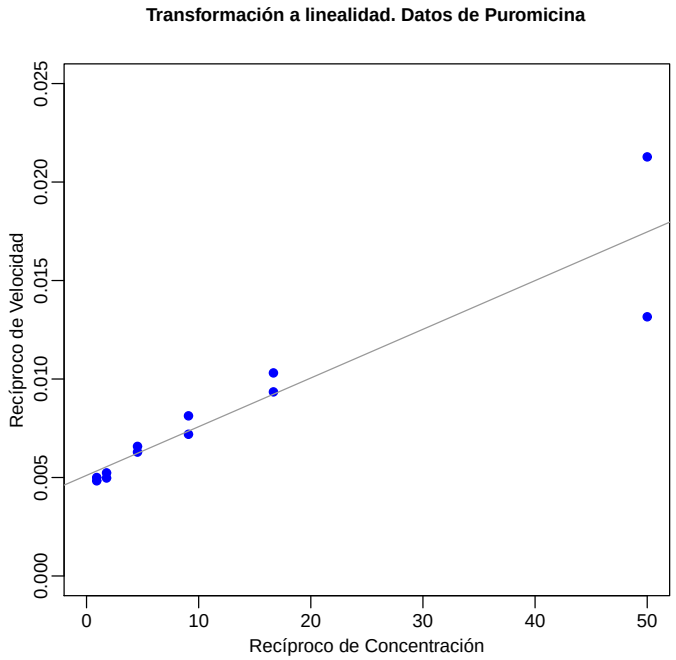
$$\text{Vel} = \frac{\theta_1 \text{Conc}}{\theta_2 + \text{Conc}}$$

  Es fácil ver que θ_1 es la asíntota de la Velocidad cuando la Concentración se hace grande y que θ_2 es la concentración que corresponde a un 50% de la velocidad inicial máxima.

- Transformación con recíprocos. Denotando $y = \text{Vel}$ y $x = \text{Conc}$, note que

$$y = \frac{\theta_1 x}{\theta_2 + x} \qquad \Leftrightarrow \qquad \frac{1}{y} = \frac{1}{\theta} + \frac{\theta_2}{\theta_1}\frac{1}{x} \qquad \equiv \qquad y^* = \alpha + \beta x^*$$

  así que el tomar recíprocos linealiza el modelo de Michaelis-Menten. La siguiente gráfica muestra los datos transformados. En este caso la varianza se afecta con la transformación; en estos casos es preferible trabajar directamente con el Michaelis-Menten en vez de linealizarlos.

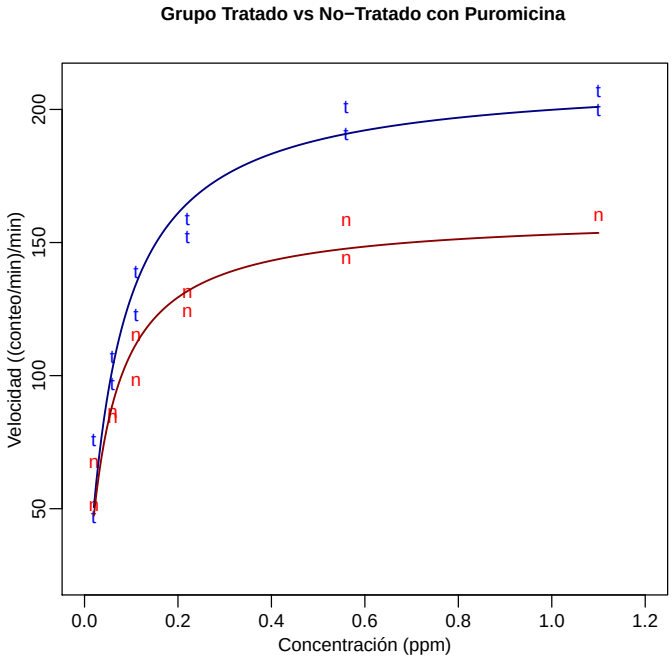



```
####  Transformacion con reciprocos
gtra <- (trat==1)    # Grupo tratado
y    <- vel[gtra]
```

```
x    <- conc[gtra]
yr   <- 1/y
xr   <- 1/x

plot( xr, yr, xlim=c(0,50), ylim=c(0,0.025),
    xlab = "Reciproco de Concentracion", mgp=c(1.5,.5,0),
    ylab = "Reciproco de Velocidad", col="blue",
    cex=.9, cex.lab=.9, cex.axis=.9, pch=19, cex.main=.9,
    main = "Transformacion a linealidad. Datos de Puromicina")
abline( lm( yr ~xr ), col=gray(.6) )
```

- Ajuste del modelo: Uso de la función `nls()`. Enseguida mostramos ajustes separados para los dos modelos.

```
# Ajuste del modelo M-M para el grupo tratado
tet0 <- c(200,.1)         # valores iniciales para los parmetros
gtra <- (trat==1)
outt <- nls(vel ~ t1*conc/(t2+conc), subset=gtra,
            start = list(t1 = tet0[1], t2 = tet0[2]))
sumt <- summary(outt)
xt   <- seq(min(conc),max(conc),length=200)
t1et <- sumt$parameters[1,1]
t2et <- sumt$parameters[2,1]
yt   <- t1et * xt/(t2et + xt)
lines(xt,yt, col="navyblue",lwd=1.5)
# t1et = 212.6836;  t2et = 0.0641211
# Salida de summary:
# Formula: vel ~ t1 * conc/(t2 + conc)
#Parameters:
#    Estimate Std. Error t value Pr(>|t|)
#t1 2.127e+02  6.947e+00  30.615 3.24e-11 ***
#t2 6.412e-02  8.281e-03   7.743 1.57e-05 ***
#Residual standard error: 10.93 on 10 degrees of freedom
# Ajuste del modelo M-M para el grupo no tratado
outn <- nls(vel ~ t1*conc/(t2+conc), subset=!gtra,
            start = list(t1 = tet0[1], t2 = tet0[2]))
sumn <- summary(outn)
xn   <- seq(min(conc),max(conc),length=200)
t1en <- sumn$parameters[1,1]
t2en <- sumn$parameters[2,1]
yn   <- t1en * xn/(t2en + xn)
lines(xn,yn, col="darkred",lwd=1.5)
# t1en = 160.28;   t2en = 0.04770812
```



- Ilustración de mínimos cuadrados (Método de Gauss-Newton).

```
gtra <- (trat==1)
y    <- vel[gtra]             # definicion de vector de respuesta
x    <- conc[gtra]            # definicion de vector de covariables
n    <- length(y)
p    <- 2                     # numero de parametros en modelo M-M
ff   <- function(tet){        # ff regresa vector nx1 con modelo M-M
 t1  <- tet[1]
 t2  <- tet[2]
 return( t1*x/(t2+x) )}
FF   <- function(tet){        # FF regresa matriz nx2 con derivadas
 t1  <- tet[1]
 t2  <- tet[2]
 return( cbind(x/(t2+x),-t1*x/(t2+x)^2) )}

tet0   <- c(200,.1)   # valores iniciales
tolm   <- 1e-4        # tolerancia (norma minima de delta)
iterm  <- 100         # numero maximo de iteraciones
tolera <- 1           # inicializar tolera
itera  <- 0           # inicializar itera
delta  <- rep(0,p)    # inicializar delta
histo  <- tet0        # inicializar historial de iteraciones

while( (tolera>tolm)&(itera<iterm) ){
 X      <- FF(tet0)
 res    <- y - ff(tet0)
 delta  <- as.vector( solve( t(X)%*%X , t(X)%*%res ) )
```

```
 tet0   <- tet0 + delta
 tolera <- sqrt( sum(delta*delta) )
 histo  <- rbind(histo,tet0)
 itera  <- itera + 1 }
teta <- tet0
histo 200.0000 0.10000000
tet0  212.0238 0.05428736
tet0  211.7728 0.06232446
tet0  212.5633 0.06392648
tet0  212.6716 0.06410228
tet0  212.6826 0.06411945
tet0  212.6836 0.06412111
tet0  212.6837 0.06412126
tet0  212.6837 0.06412128
```

- Cálculo de errores estándar e intervalos de confianza. Comentamos en la clase pasada que la distribución del estimador de máxima verosimilitud, $\widehat{\theta}$, puede aproximarse mediante

$$\widehat{\theta} \sim N_p(\theta, \sigma^2 C^{-1}), \quad \text{donde} \quad C = F^T(\theta)F(\theta)$$

y donde $F(\theta)$ es la matriz $n \times p$ cuyo i–ésimo renglón es $\partial f(x_i;\theta)/\partial\theta^T$. El estimador usual para la varianza, basado en la suma de cuadrados residual es $\widehat{\sigma}^2 = \text{SCR}(\widehat{\theta})/(n-p)$ y, como se comentó al final de la clase pasada, podemos construír intervalos de confianza mediante:

$$a^T\widehat{\theta} \pm t_{n-p,\alpha/2}\sqrt{CMEa^TC^{-1}a}$$

```
# Calculo de errores estandar e intervalos de confianza
sig2  <- sum( (y-ff(teta))^2 )/(n-p)    # sqrt(sig2) = 10.93366
X     <- FF( teta )
errst <- sqrt( diag( sig2*solve( t(X) %*% X ) ) )
lims  <- teta + abs(qt(.05/2,n-p))*errst
limi  <- teta - abs(qt(.05/2,n-p))*errst

cbind( limi,teta,lims )  # Intervalos del 95% de confianza
             limi         teta         lims
[1,] 197.20451590 212.68374209 228.16296828
[2,]   0.04567018   0.06412128   0.08257238
# errst =   6.94715510 0.00828095
```

- Incorporación de covariables al modelo. Podemos incorporar, por ejemplo, una variable indicadora de tratamiento y ajustar un modelo combinado para todos los datos.

$$\text{Vel} = \frac{(\theta_1 + \phi_1\,\text{Trat})\,\text{Conc}}{\theta_2 + \phi_2\,\text{Trat} + \text{Conc}}$$

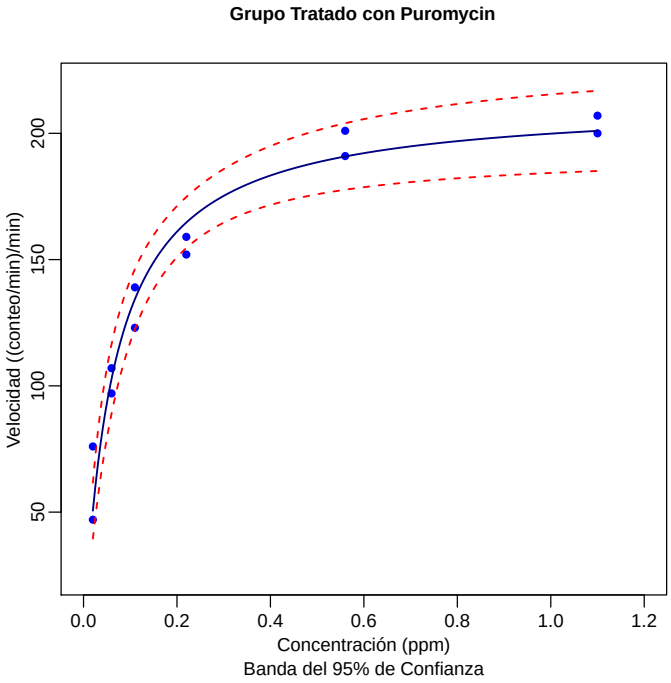
```
# Ajuste del modelo completo

out  <- nls(vel ~ (t1+fi1*trat)*conc/((t2+fi2*trat)+conc),
           start = list(t1 = 160, t2 = .04, fi1=42, fi2=.02))
sumy <- summary(out)

Formula: vel ~ (t1 + fi1 * trat) * conc/((t2 + fi2 * trat) + conc)
```

```
Parameters:
     Estimate Std. Error t value Pr(>|t|)
t1  1.603e+02  6.896e+00  23.242 2.04e-15 ***
t2  4.771e-02  8.281e-03   5.761 1.50e-05 ***
fi1 5.240e+01  9.551e+00   5.487 2.71e-05 ***
fi2 1.641e-02  1.143e-02   1.436    0.167

Residual standard error: 10.4 on 19 degrees of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 3.145e-06

# Hay indicio de que fi1 es distinto de cero, esto es,
# diferente asintota (diferente velocidad inicial maxima)
# Hay indicio de que fi2 es cero, esto es, mismo parametro
# t2 para ambos.
```

- Comparación de Modelos.

```
# Modelo Completo (4 parametros) vs Modelo Reducido (3 parametros)

outr  <- nls(vel ~ (t1+fi1*trat)*conc/(t2+conc),
           start = list(t1 = 160, t2 = .04, fi1=42))
sumr <- summary(outr)

n <- length(vel)
p <- 4
SCModComp <- (n-p)*(sumy$sigma)^2     # 19 gl
SCModRedu <- (n-p+1)*(sumr$sigma)^2   # 20 gl
Fobs <- (SCModRedu-SCModComp)/(SCModComp/(n-p))
pval <- 1-pf(Fobs,1,20)   # 0.205
# Modelo Reducido ok

Parameters:
     Estimate Std. Error t value Pr(>|t|)
t1  166.60402    5.80743  28.688  < 2e-16 ***
t2    0.05797    0.00591   9.809 4.37e-09 ***
fi1  42.02594    6.27214   6.700 1.61e-06 ***

Residual standard error: 10.59 on 20 degrees of freedom
```

- Bandas de Confianza.

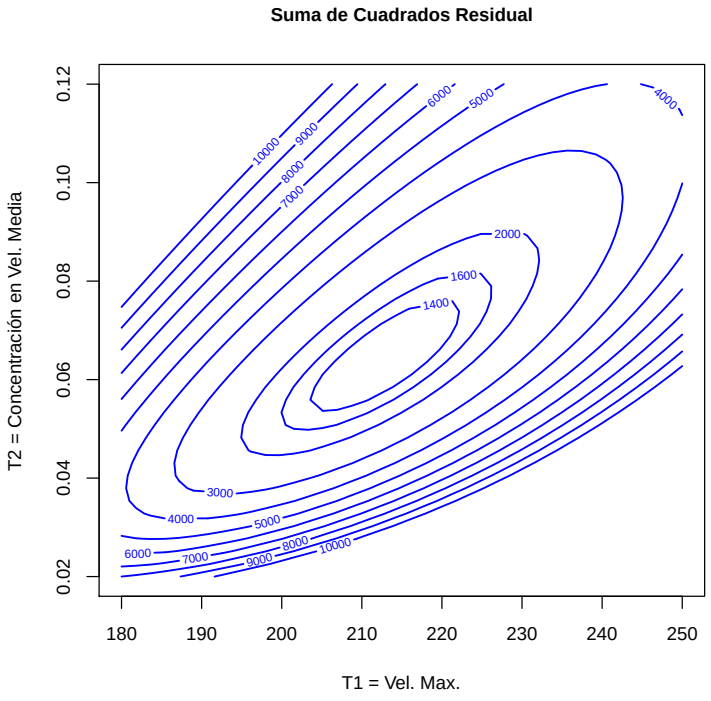



```
# Banda de confianza para respuesta esperada: Datos de Puromycin
plot( conc[trat==1], vel[trat==1], xlim=c(0,1.2), ylim=c(25,220),
    xlab = "Concentracion (ppm)", mgp=c(1.5,.5,0),
    ylab = "Velocidad ((conteo/min)/min)", col="blue",
    cex = .9, cex.lab=.9, cex.axis=.9, pch = 16,
    main = "Grupo Tratado con Puromicina", cex.main=.9,
    sub = "Banda del 95% de Confianza", cex.sub=.9)
t1et <- teta[1]
t2et <- teta[2]
xt   <- seq(min(conc),max(conc),length=200)
yt   <- t1et * xt/(t2et + xt)
lines(xt,yt, col="navyblue",lwd=1.5)
ff0   <- function(tet){
 t1  <- tet[1];  t2  <- tet[2]
 return( t1*xt/(t2+xt) )}
FF0   <- function(tet){
 t1  <- tet[1];  t2  <- tet[2]
 aa  <- xt/(t2+xt)
 return( cbind(aa,-t1*aa/(t2+xt)) )}  # FF regresa matriz Mx2 con derivadas
ff   <- function(tet){
 t1  <- tet[1];  t2  <- tet[2]
 return( t1*x/(t2+x) )}
FF   <- function(tet){
 t1  <- tet[1];  t2  <- tet[2]
 aa  <- x/(t2+x)
 return( cbind(aa,-t1*aa/(t2+x)) )}  # FF regresa matriz nx2 con derivadas
SCR <- function(teta){return( sum((y - teta[1]*x/(teta[2]+x))^2) )}
gtra  <- (trat==1)
y     <- vel[gtra]
```

```
x     <- conc[gtra]
pare  <- c(t1et, t2et)
sig2  <- (sumt$sigma)^2
f0    <- ff0( pare )
X0    <- FF0( pare )
X     <- FF( pare )
errst <- sqrt( diag( sig2*X0%*%solve( t(X)%*%X, t(X0) ) ) )   # p.59 B&W
lims  <- f0 + sqrt(qf(.95,p,n-p)*p)*errst
limi  <- f0 - sqrt(qf(.95,p,n-p)*p)*errst
lines( xt, lims, lty=2, col="red", lwd=1.5 )
lines( xt, limi, lty=2, col="red", lwd=1.5 )
```

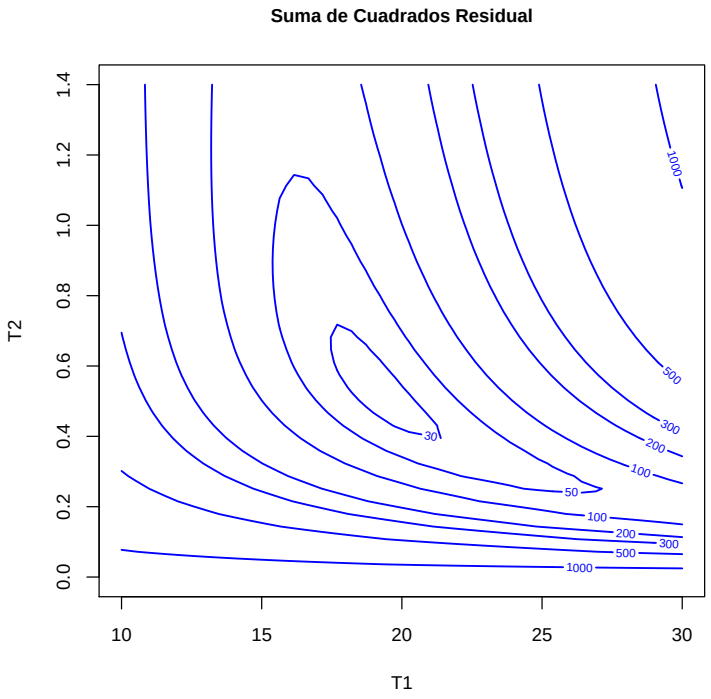
- Gráfica de contornos 1. Datos de puromicina: Contornos aproximadamente elípticos, de aquí que las distribuciones asintóticas son buenas aproximadas.

```
## Grafica de Contornos de la Suma de Cuadrados Residual
gtra <- (trat==1)    # Grupo tratado
y    <- vel[gtra]
x    <- conc[gtra]
N    <- 40
t1   <- seq(180, 250, length=N)
t2   <- seq(.02,.12, length=N)
SCR  <- matrix(0,N,N)
for( i in 1:N ){
 for( j in 1:N ){ SCR[i,j] <- sum((y - t1[i]*x/(t2[j]+x))^2)} }
contour(t1,t2,SCR, xlab = "T1 = Vel. Max.", cex.axis=.9,lwd=1.5,
        levels = c(1400,1600,seq(2000,10000,by=1000)),col="blue",
        ylab = "T2 = Concentracion en Vel. Media", cex.lab=.9,
        cex.main=.9, main="Suma de Cuadrados Residual")
```



- Gráfica de contornos 2. Para otro conjunto de datos y usando un modelo exponencial, tenemos que los contornos no son "tan elípticos" como en el caso anterior. Esto dará una indicación de que las aproximaciones asintóticas no serán tan buenas.

```
### BOD Data
x <- c(1,2,3,4,5,7,9,11)
y <- c(.47,.74,1.17,1.42,1.6,1.84,2.19,2.17)

x <- c(1,2,3,4,5,7)
y <- c(8.3,10.3,19,16,15.6,19.8)
N   <- 40
t1  <- seq(10, 30, length=N)
t2  <- seq(0,1.4, length=N)
SCR <- matrix(0,N,N)
for( i in 1:N ){
 for( j in 1:N ){
  SCR[i,j] <- sum((y - t1[i]*(1-exp(-t2[j]*x)) )^2)}}

contour(t1,t2,SCR, xlab = "T1", cex.axis=.9,lwd=1.5,
        levels = c(20,30,50,100,200,300,seq(500,5000,by=500)),col="blue",
        ylab = "T2", cex.lab=.9,
        cex.main=.9, main="Suma de Cuadrados Residual")
```



Resumen de Clase 28: Miércoles 18 de mayo

- Referencias. En esta clase veremos como cuantificar el grado de nolinealidad de un modelo; dos buenas referencias sobre el tema son:

  1. **Bates, D.M. & Watts, D.G.** (1988). Nonlinear regression analysis and its applications. Wiley.
  2. **Seber, G.A.F. & Wild, C.J.** (1989). Nonlinear regression. Wiley.

- Mínimos cuadrados en regresión nolineal. En regresión lineal estimamos θ minimizando $S(\theta) = ||y - X\theta||^2$. Para modelos nolineales, tenemos que $S(\theta) = ||y - \mu(\theta)||^2$, la cual es minimizada en $\widehat{\theta}$ cuando $y - \mu(\widehat{\theta})$ sea ortogonal al plano tangente a $\mu(\widehat{\theta})$

- Plano tangente. En general, $\mu(\theta)$ es una función de $\mathbb{R}^p$ a $\mathbb{R}^n$. El plano tangente está dado por la aproximación lineal

$$\mu(\theta) \doteq \mu(\widehat{\theta}) + F_{\cdot}(\theta - \widehat{\theta}), \qquad \text{donde} \quad \underset{n\times p}{F_{\cdot}} = \left[\frac{\partial f(x_i,\theta)}{\partial \theta^T}\right]$$

- Aproximación de segundo orden. El grado de nolinealidad de un modelo se cuantifica en base a la discrepancia que hay entre la aproximación de primer orden y una de segundo orden. Note que $\mu(\theta) = (\, f(x_1;\theta), \cdots, f(x_n;\theta)\,)^T$, si tomamos la $i-$ésima componente, tenemos que una aproximación de segundo orden es

$$f(x_i;\theta) = f(x_i;\widehat{\theta}) + \frac{\partial f(x_i;\widehat{\theta})}{\partial \theta^T}(\theta - \widehat{\theta}) + \frac{1}{2}(\theta - \widehat{\theta})^T \left[\frac{\partial f(x_i;\widehat{\theta})}{\partial \theta \partial \theta^T}\right](\theta - \widehat{\theta})$$

  Así que

$$\mu(\theta) - \mu(\widehat{\theta}) \doteq F_{\cdot}\delta + \frac{1}{2}\begin{bmatrix} \delta^T H_1 \delta \\ \vdots \\ \delta^T H_n \delta \end{bmatrix}, \qquad \text{donde} \quad \delta = \theta - \widehat{\theta} \quad \text{y} \quad H_i = \left[\frac{\partial f(x_i;\widehat{\theta})}{\partial \theta \partial \theta^T}\right]_{p\times p}$$

  Las medidas de nolinealidad están basadas en el arreglo $p \times p \times n$: $[H_1, \cdots, H_n]$.

- El término cuadrático. Sea g_{rs} el vector $n \times 1$ definido por

$$g_{rs} = \begin{bmatrix} \frac{\partial^2 f(x_1,\widehat{\theta})}{\partial \theta_r \partial \theta_s} \\ \vdots \\ \frac{\partial^2 f(x_n,\widehat{\theta})}{\partial \theta_r \partial \theta_s} \end{bmatrix}$$

  Ahora, cada forma cuadrática $\delta^T H_i \delta$ es de la forma $\sum_r \sum_s \delta_r \delta_s \partial^2 f(x_i,\widehat{\theta})/\partial\theta_r\partial\theta_s$, por lo tanto

$$\delta^T F_{\cdot\cdot}\delta \equiv \begin{bmatrix} \delta^T H_1 \delta \\ \vdots \\ \delta^T H_n \delta \end{bmatrix} = \sum_r \sum_s \delta_r \delta_s g_{rs}$$

  Cada vector g_{rs} puede ser descompuesto como

$$g_{rs} = g_{rs}^{\mathsf{T}} + g_{rs}^{\perp}, \qquad \text{donde} \quad g_{rs}^{\mathsf{T}} = P g_{rs} \quad \text{y} \quad g_{rs}^{\perp} = (I - P) g_{rs}$$

con P la matriz de proyección sobre el espacio de columnas de $F_{\cdot}$ (i.e. una base del espacio tangente afin). Entonces el término cuadrático de la expansión de $\mu(\theta)$ es

$$\delta^T F_{\cdot\cdot}\delta = \begin{bmatrix} \delta^T H_1 \delta \\ \vdots \\ \delta^T H_n \delta \end{bmatrix} = \sum_r \sum_s \delta_r \delta_s (g^{\mathsf{T}}_{rs} + g^{\perp}_{rs}) \equiv P^{\mathsf{T}} + P^{\perp}$$

Por propiedades de ortogonalidad tenemos

$$||\delta^T F_{\cdot\cdot}\delta||^2 = ||P^{\mathsf{T}}||^2 + ||P^{\perp}||^2$$

En las referencias mencionadas arriba se muestra que $||\delta^T F_{\cdot\cdot}\delta||$ es una medida que depende de la forma en que se parametriza el modelo, pero en la descomposición resulta que sólo el primer término, $||P^{\mathsf{T}}||$, depende de la parametrización y el segundo no.

- Medidas de nolinealidad 1.

  a) Curvatura debida a la parametrización.

  $$K^{\mathsf{T}}_{\delta} = \frac{||P^{\mathsf{T}}||}{||F_{\cdot}\delta||^2}$$

  b) Curvatura intrínseca.

  $$K^{\perp}_{\delta} = \frac{||P^{\perp}||}{||F_{\cdot}\delta||^2}$$

- Medidas de nolinealidad 2. En un modelo con p parámetros y con un estimador de la varianza denotado por CME, definamos $\rho = p\,\text{CME}$. Para eliminar efectos de escala en las medidas anteriores definimos

  a) Curvatura debida a la parametrización.

  $$\gamma^{\mathsf{T}}_{\delta} = \rho K^{\mathsf{T}}_{\delta}$$

  b) Curvatura intrínseca.

  $$\gamma^{\perp}_{\delta} = \rho K^{\perp}_{\delta}$$

- Medidas de nolinealidad 3. Finalmente, para eliminar el efecto de dirección presente en δ, se toma el máximo sobre δ de las medidas anteriores

  a) Curvatura debida a la parametrización.

  $$\gamma^{\mathsf{T}}_{\mathsf{max}} = \max_{\delta} \gamma^{\mathsf{T}}_{\delta}$$

  b) Curvatura intrínseca.

  $$\gamma^{\perp}_{\mathsf{max}} = \max_{\delta} \gamma^{\perp}_{\delta}$$

  Se ha propuesto a

  $$\frac{1}{2\sqrt{F^p_{n-p,\alpha}}}$$

  como un valor razonable tal que si $\gamma^{\perp}_{\mathsf{max}}$ es menor que este valor entonces son adecuadas las inferencias usuales basadas en propiedades asintóticas del estimador de máxima verosimilitud (las cuales usan la aproximación de primer orden). También se ha sugerido este mismo valor límite para $\gamma^{\mathsf{T}}_{\mathsf{max}}$ aunque no es tan contundente la afirmación como en el otro caso.

Examen Final de Modelos Estadísticos I

Nombre: _____

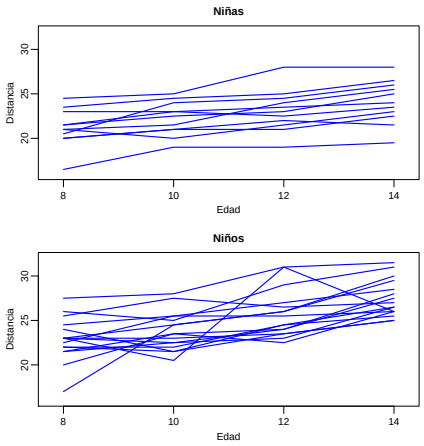
1. En un estudio, llevado a cabo por ortodoncistas, se midió la distancia entre la glándula pituitaria y la fisura maxilar de 27 niños (11 niñas y 16 niños) (ambos puntos fácilmente identificables en placas de rayos-X). Estas mediciones fueron efectuadas cada dos años, iniciando el estudio cuando los niños tenían 8 años y terminando a los 14.

   De la gráfica observamos:

   - Tendencia creciente de los perfiles individuales.
   - Diferentes ordenadas al origen por individuo.
   - Pendientes similares dentro de cada grupo.
   - Aparentemente los perfiles de los niños tienden a estar a un mayor nivel que los de las niñas.

   Los intereses del estudio son:

   - Caracterizar el crecimiento de la distancia pituitaria-fisura.
   - Efectuar la caracterización por grupos (niñas, niños) si es que es necesario.



   Las gráficas sugieren un modelo de **interceptos aleatorios**

$$y_{ij} = \begin{cases} \beta_{0i} + \beta_{1m} t_j + e_{ij} & \text{para } i = 1, \cdots, 11 \\ \beta_{0i} + \beta_{1h} t_j + e_{ij} & \text{para } i = 12, \cdots, 27 \end{cases} \qquad j = 1, \cdots, 4.$$

   aquí suponemos que los interceptos, β_{0i}'s, son aleatorios e independientes con distribución $N(\beta_0, \sigma_0^2)$; también asumimos que son independientes de las desviaciones e_{ij}'s las cuales son i.i.d $N(0, \sigma^2)$. En otras palabras, condicionado a la realización de una ordenada al origen, el perfil de crecimiento de un individuo consiste de una tendencia lineal creciente afectada por ruido aleatorio.

   a) Si $y_i = (y_{i1}, \cdots, y_{i4})^T$, muestre que

$$\text{Var}(y_i) = \tau^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix} \equiv \tau^2 V_0, \quad \text{donde } \tau^2 = \sigma_0^2 + \sigma^2$$

   y $\rho = \sigma_0^2/(\sigma_0^2 + \sigma^2)$, (a la estructura de la matriz V_0 se le denomina **matriz de correlación uniforme**).

   b) Si escribimos $\beta_{0i} = \beta_0 + \delta_i$, con $\delta_i \sim N(0, \sigma_0^2)$, entonces

$$y_{ij} = \beta_0 + \beta_{1m} x_i t_j + \beta_{1h}(1 - x_i) t_j + \delta_i + e_{ij}, \text{ con } x_i = \begin{cases} 1 & \text{si } i = 1, \cdots, 11 \\ 0 & \text{si } i = 12, \cdots, 27 \end{cases}$$

   Tenemos que los vectores $y_1, \cdots, y_n$ son independientes, donde

$$y_1 \sim N(X_1\beta, \tau^2 V_0), \quad \cdots \quad y_n \sim N(X_n\beta, \tau^2 V_0)$$

con $\beta = (\beta_0, \beta_{1m}, \beta_{1h})^T$,

$$X_i = \begin{bmatrix} 1 & t_1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & t_4 & 0 \end{bmatrix}, \quad \text{si } i = 1, \cdots, 11 \qquad \text{y} \qquad X_i = \begin{bmatrix} 1 & 0 & t_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & t_4 \end{bmatrix}, \quad \text{si } i = 12, \cdots, 27$$

Muestre que la logverosimilitud es

$$l(\beta, \tau^2, V_0) = C - \frac{1}{2\tau^2} \sum_{i=1}^{n} (y_i - X_i\beta)^T V_0^{-1} (y_i - X_i\beta) - \frac{mn}{2} \log(\tau^2) - \frac{n}{2} \log|V_0|, \quad m = 4$$

c) Si fijamos V_0, muestre que los estimadores de máxima verosimilitud son:

$$\widehat{\beta}(V_0) = \left(\sum_{i=1}^{n} X_i^T V_0^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{n} X_i^T V_0^{-1} y_i \right)$$

$$\widehat{\tau}^2(V_0) = \frac{1}{mn} \text{SCE}(V_0) = \frac{1}{nm} \sum_{i=1}^{n} (y_i - X_i\widehat{\beta})^T V_0^{-1} (y_i - X_i\widehat{\beta})$$

y que la logverosimilitud perfil es

$$l(V_0) = C - \frac{mn}{2} \log[\text{SCE}(V_0)] - \frac{n}{2} \log|V_0|$$

Note que aquí V_0 corresponde a una matriz de correlación uniforme, sin embargo, se podrían usar otras estructuras de correlación que se consideraran convenientes.

2. Considere el modelo de regresión nolineal

$$y_i = f(x_i, \theta) + e_i, \qquad i = 1, 2, \cdots, n$$

donde las e_i's con i.i.d. $N(0, \sigma^2)$ y θ es un parámetro $p-$dimensional. Sea $\widehat{\theta}$ el estimador de Máxima Verosimilitud de θ.

a) Considere la función Score, $\mathcal{U}(\theta) = \partial l(\theta)/\partial\theta$. Muestre que $\mathcal{U}(\theta)/\sqrt{n}$ es asintóticamente normal. Especifique quién es la varianza asintótica.

b) Usando una aproximación de primer orden para $\mathcal{U}(\widehat{\theta})$ deduzca que $\widehat{\theta}$ es asintóticamente normal. Especifique quién es la varianza asintótica.

Puede usar el siguiente resultado: Sea $x_1, x_2, \cdots$ una sucesión de variables aleatorias k-dimensionales independientes con $\text{E}(x_i) = 0$ y $\text{Var}(x_i) = \Sigma_i$. Suponga que

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Sigma_i = \Sigma > 0$$

suponga además que, para cada $\epsilon > 0$,

$$\frac{1}{n} \sum_{i=1}^{n} \int_{||x_i|| > \epsilon\sqrt{n}} ||x_i||^2 \, dF_i \to 0$$

donde F_i es la función de distribución de x_i. Entonces

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \xrightarrow{d} N_k(0, \Sigma).$$

3. Suponga $y_i \sim \text{Bin}(m_i, \pi_i)$, con $y_1, \cdots, y_n$ independientes. Usaremos la notación $\pi = (\pi_1, \cdots, \pi_n)^T$ y $y = (y_1, \cdots, y_n)^T$. La devianza residual se define como dos veces la diferencia entre el máximo valor alcanzable de la logverosimilitud y el máximo valor alcanzable de la logverosimilitud bajo el modelo bajo estudio. Esto es

$$D(y\,;\,\widehat{\pi}) = 2l(\tilde{\pi}\,;\,y) - 2l(\widehat{\pi}\,;\,y)$$

Es fácil ver que el máximo valor alcanzable ocurre con $\tilde{\pi}_i = y_i/m_i, \;\; i = 1, \cdots, n.$

a) Muestre que

$$D(y\,;\,\widehat{\pi}) = 2\sum_i \left\{ y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \widehat{\mu}_i}\right)\right\}$$

donde $\widehat{\mu}_i = m_i\widehat{\pi}_i$

b) Considere dos modelos anidados $H_0 \subset H_1$, con respectivas devianzas $D(y\,;\,\widehat{\pi}^0)$ y $D(y\,;\,\widehat{\pi}^1)$. Muestre que el modelo de regresión logística satisface

$$D(y\,;\,\widehat{\pi}^0) = D(y\,;\,\widehat{\pi}^1) + D(\widehat{\mu}^1\,;\,\widehat{\pi}^0)$$

donde $\widehat{\mu}^1 = (\widehat{\mu}^1_1, \cdots, \widehat{\mu}^1_n)^T = (m_1\widehat{\pi}^1_1, \cdots, m_n\widehat{\pi}^1_n)^T$

c) Muestre que $D(\widehat{\mu}^1\,;\,\widehat{\pi}^0)$ es el estadístico de prueba basado en el cociente de verosimilitudes para probar H_0 versus H_1

4. Sean $y_1, \cdots, y_n$, variables aleatorias independientes, con $\text{E}(y_i) = \mu_i$ y varianza $\text{Var}(y_i) = \phi V(\mu_i)$, donde V es una función conocida. Suponga que g es una función monótona y diferenciable, tal que $g(\mu_i) = \eta_i = x_i^T\beta$. Definamos

$$z_i = g'(\mu_i)(y_i - \mu_i) + \eta_i \qquad \text{y} \qquad w_i = \left\{V(\mu_i)[g'(\mu_i)]^2\right\}^{-1}$$

a) Note que $\text{E}(z_i) = x_i^T\beta$ y que la matriz de covarianza de $z = (z_1, \cdots, z_n)^T$ es ϕW^{-1}, donde W es una matriz diagonal con $W_{ii} = w_i$. Si β es estimado mediante la minimización de

$$S(\beta) = \sum_{i=1}^{n} w_i(z_i - x_i^T\beta)^2,$$

muestre que la matriz de covarianza del estimador resultante, $\widehat{\beta}$, es $\phi\,(X^TWX)^{-1}$ y encuentre $\text{E}(\widehat{\beta})$.

b) La versión multivariada del teorema central del límite nos dice que, conforme la dimensión de z tiende a infinito, X^TWz va a tender a una normal multivariada. ¿Qué implicación tiene esto sobre la distribución de $\widehat{\beta}$ para muestras grandes?. ¿Es $\widehat{\beta}$ el estimador de máxima verosimilitud?.

El examen es individual y tiene una duración de 3 horas. Hay un total de 10 incisos, entregar solo 9, indicando claramente cuáles 9 son los que serán calificados.

FIN DEL CURSO DE MODELOS ESTADÍSTICOS I