MÈTODES DE CAPTACIÓ, ANÀLISI I INTERPRETACIÓ DE DADES

MASTER DE LOGÍSTICA, TRANSPORT I MOBILITAT

MASTER D'ESTADÍSTICA I INVESTIGACIÓ OPERATIVA

APUNTS DE CLASSE PROF. LÍDIA MONTERO:

TEMA 3: MODELS DE RESPOSTA NORMAL. ESTIMACIÓ I INFERÈNCIA

AUTORA:
Lídia Montero Mercadé
Departament d'Estadística i Investigació Operativa

Versió 1.4
Setembre del 2.012

TABLA DE CONTENIDOS

TABLA DE CONTENIDOS

3-1-1. BIBLIOGRAFÍA

Referències bàsiques:

- Fox, J. *Applied Regression Analysis and Generalized Linear Models.* Sage Publications, 2nd Edition 2008.
- Fox, J. *An R Companion to Applied Regression.* Sage Publications, 2nd Edition 2010.
- Dalgaard, P. *Introductory Statistics with R.* Springer 2002
- Dobson, A. J.: *An Introduction to Generalized Linear Models.* Chapman and Hall, 1990.
- Faraway, J.J: *Extending the Linear Model with R.* Chapman and Hall/CRC, 2006.
- Peña, D.: *Estadística. Modelos y métodos. Vol. 2, Modelos lineales y series temporales.* Alianza Universidad Textos, 1989.

3-1-2. TEMA 3: INTRODUCCIÓN A LOS MODELOS DE RESPUESTA NORMAL

Sea un vector de observaciones de *n* componentes, $\mathbf{y}^T = (y_1, \ldots, y_n)$, realización de un vector aleatorio $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$, cuyas componentes son estadísticamente independientes y distribuidas con medias $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_n)$:

- En los modelos lineales ordinarios, las componentes de la respuesta $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$ tienen distribuciones normales, independientes con varianza constante, con $\mathrm{E}[\mathbf{Y}] = \boldsymbol{\mu}$ y varianza σ^2.

- **La componente sistemática** del modelo consiste en la especificación de un vector $\boldsymbol{\eta}$, el predictor lineal a partir de un número reducido de parámetros a estimar y regresores; parámetros $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ y regresores $\mathbf{X}^T = (X_1, \ldots, X_p)$. Esta especificación responde, en notación matricial a $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ donde $\boldsymbol{\eta}$ es *nx1,* $\mathbf{X}$ es *nxp* y $\boldsymbol{\beta}$ es *px1.*

- El vector $\boldsymbol{\mu}$ está funcionalmente relacionado con el predictor lineal $\boldsymbol{\eta}$, a través de la ***función de link identidad*, en** los modelos lineales ordinarios $\boldsymbol{\eta} = \boldsymbol{\mu}$.

TEMA 3: INTRODUCCIÓN A LOS MODELOS DE RESPUESTA NORMAL

Clasificación de los métodos estadísticos de análisis:

Variables Explicativas	Variable de respuesta				
	Binaria	***Politómica***	***Cuantitativa Discreta***	***Cuantitativa Continua***	
				Normal	***Tiempo entre eventos***
Binaria	Tablas de contingencia Regresión logística Modelos log-lineales	Tablas de contingencia * Modelos log-lineales	Modelos log-lineales	**Tests en medias de 2 grupos: t.test**	Análisis de la Supervivencia
Politómicas	Tablas de contingencia Regresión logística Modelos log-lineales	Tablas de contingencia Modelos log-lineales	Modelos log-lineales	**ONEWAY, ANOVA**	Análisis de la Supervivencia
Continuas	Regresión logística	*	Modelos log-lineales	**Regresión Múltiple**	Análisis de la Supervivencia
Factores y covariables	Regresión logística	*	Modelos log-lineales	**ANCOVA**	Análisis de la Supervivencia
Efectos Aleatorios	Modelos mixtos	Modelos mixtos	Modelos mixtos	Modelos mixtos	Modelos mixtos

TEMA 3: INTRODUCCIÓN A LOS MODELOS DE RESPUESTA NORMAL

Una distribución pertenece a la familia exponencial si puede escribirse de la siguiente manera:

$$f_Y(y,\theta,\phi)=\exp\left(\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right)$$

Donde *a(.), b(.) y c(.)* son funciones específicas con ϕ conocido y donde se denomina al único parámetro θ : *parámetro canónico*.

➡ En la distribución normal:

$$f_Y(y,\theta,\phi)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)=\exp\left(\frac{y\mu-\mu^2/2}{\sigma^2}-\frac{1}{2}\left(\frac{y^2}{\sigma^2}+\log\left(2\pi\sigma^2\right)\right)\right)$$

donde $a(\phi)=\phi=\sigma^2$, $b(\theta)=\frac{\theta^2}{2}=\frac{\mu^2}{2}$ (es decir, $\theta=\mu$) y $c(y,\phi)=-\frac{1}{2}\left(\frac{y^2}{\phi}+\log(2\pi\phi)\right)$.

TEMA 3: INTRODUCCIÓN A LOS MODELOS DE RESPUESTA NORMAL

➡ …En la distribución normal:

$$\ell(\theta,\phi,y)=\log f_Y(y,\theta,\phi)=\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)=\frac{y\theta-\theta^2/2}{\phi}-\frac{1}{2}\left(\frac{y^2}{\phi}+\log(2\pi\phi)\right)$$

➡ Para la ley normal, la devianza escalada toma por expresión dada una colección de n observaciones:

$$D'(\mathbf{y},\hat{\boldsymbol{\mu}})=2\,\ell(\mathbf{y},\phi,\mathbf{y})-2\,\ell(\hat{\boldsymbol{\mu}},\phi,\mathbf{y})=\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\phi}$$

Para la ley normal, la devianza toma por expresión dada una colección de n observaciones:

$$D(\mathbf{y},\hat{\boldsymbol{\mu}})=D'(\mathbf{y},\hat{\boldsymbol{\mu}})\phi=\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$$

3-1-3. TEMA 3: ESTIMACIÓN POR MÍNIMOS CUADRADOS

Sea el modelo con variable de respuesta contínua,

$$\mathbf{Y} = \mu + \varepsilon = \mathbf{X}\beta + \varepsilon$$, donde $\mathbf{Y}$, $\eta = \mu$ son *nx1,* $\mathbf{X}$ es *nxp* y β es *px1*

en primera instancia sin ninguna hipótesis sobre la distribución de la variable **Y**, ni los errores.

La estimación de los parámetros β puede caracterizarse de manera genérica como,

$$Min_{\beta} \sum_k M(\varepsilon_k) = \sum_k M\left(Y_k - \mathbf{x}_k^T \beta\right)$$

donde la función real de los errores puede ser:

- $M(x)=|x|$, da lugar a procedimientos de estimación robusta basados en la norma 1.
- $M(x)=x^2$, que da lugar al método de los mínimos cuadrados.

El método de los mínimos cuadrados de entrada no requiere de ninguna hipótesis sobre la distribución de las observaciones.

La función objetivo a minimizar se suele denominar $S(\beta)$ y se define $S(\beta) = \|\varepsilon\|^2 = \varepsilon^T \cdot \varepsilon = \sum_k \varepsilon_k^2$

o bien,

TEMA 3: ESTIMACIÓN POR MÍNIMOS CUADRADOS

… La función objetivo a minimizar es …

$$S(\boldsymbol{\beta}) = \|\boldsymbol{\varepsilon}\|^2 = \boldsymbol{\varepsilon}^T \cdot \boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \cdot (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \sum_k \left(Y_k - \mathbf{x}_k^T \boldsymbol{\beta}\right)^2 = \mathbf{Y}^T\mathbf{Y} + \boldsymbol{\beta}^T \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T \mathbf{X}^T\mathbf{Y}$$

➡ Las condiciones de primer orden de mínimo de $S(\boldsymbol{\beta})$ son:

$$\nabla_\beta S(\boldsymbol{\beta}) = \mathbf{0} \leftrightarrow \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_i} = 0 \quad i = 1, \ldots, p$$

Derivando vectorialmente la expresión se obtienen las bien conocidas ecuaciones normales,

$$\nabla_\beta S(\boldsymbol{\beta}) = \mathbf{0} \leftrightarrow \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^T\mathbf{Y} = \mathbf{0} \rightarrow \mathbf{b} = \hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Si la matriz de diseño es no singular, es decir de rango *p*, entonces la solución es única. Si $\mathbf{X}$ no es de pleno rango, existen infinitas soluciones a las ***ecuaciones normales***,

$$2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 2\mathbf{X}^T\mathbf{Y} \rightarrow \mathbf{b} = \hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-}\mathbf{X}^T\mathbf{Y} + \left(\mathbf{I} - \left(\mathbf{X}^T\mathbf{X}\right)^{-}\mathbf{X}^T\mathbf{X}\right)w \quad w \in \Re$$

pero todas ellas facilitan un predictor lineal $\hat{\mathbf{y}} = \hat{\boldsymbol{\mu}} = \mathbf{X}\mathbf{b}$ idéntico y un mínimo de la función objetivo por tanto idéntico.

TEMA 3: ESTIMACIÓN POR MÍNIMOS CUADRADOS

- En modelos específicos ANOVA y ANCOVA, las dependencias lineales entre las columnas de la matriz de diseño pueden eliminarse mediante una reparametrización, que aconsejablemente debe aplicar el estadístico, ya que la interpretabilidad de los parámetros resultantes depende de ella.

$\mathbf{A}^- = \left(\mathbf{X}^T\mathbf{X}\right)^-$ es una g-inversa o inversa generalizada y satisface $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$.

$\mathbf{A}^-$ siempre existe, pero no es única. Si además satisface:

1. $\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-$ 2. $\left(\mathbf{A}\mathbf{A}^-\right)^T = \mathbf{A}\mathbf{A}^-$ 3. $\left(\mathbf{A}^-\mathbf{A}\right)^T = \mathbf{A}^-\mathbf{A}$

Entonces $\mathbf{A}^-$ es única y se denomina inversa de Moore-Penrose o p-inversa y nota como $\mathbf{A}^+$.

- Las condiciones de segundo orden de suficiencia de mínimo requieren que la hessiana de $S(\boldsymbol{\beta})$ sea definida positiva, es decir ,

$$\nabla_{\beta}{}^2 S(\boldsymbol{\beta}) > \mathbf{0} \leftrightarrow \left[\frac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\right] > 0 \quad i, j = 1, \ldots, p$$

TEMA 3: ESTIMACIÓN POR MÍNIMOS CUADRADOS

$$\nabla_{\beta}{}^{2} S(\boldsymbol{\beta}) > \mathbf{0} \leftrightarrow \left[\frac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\right] = 2\mathbf{X}^{\mathrm{T}}\mathbf{X} > \mathbf{0} \quad i, j = 1,\ldots,p$$ si la matriz de diseño **X** es no singular entonces la hessiana es definida positiva y el punto que satisfaga las condiciones de primer orden es un mínimo global.

3-1-3.1 Propiedades geométricas

Sea $\Re(\mathbf{X})$ el espacio generado por las columnas de **X**, $\Re(\mathrm{X}) = \left\{\boldsymbol{\mu} \middle| \boldsymbol{\mu} = \mathrm{X}\boldsymbol{\beta} \quad \boldsymbol{\beta} \in \Re^p\right\} \subset \Re^n$.

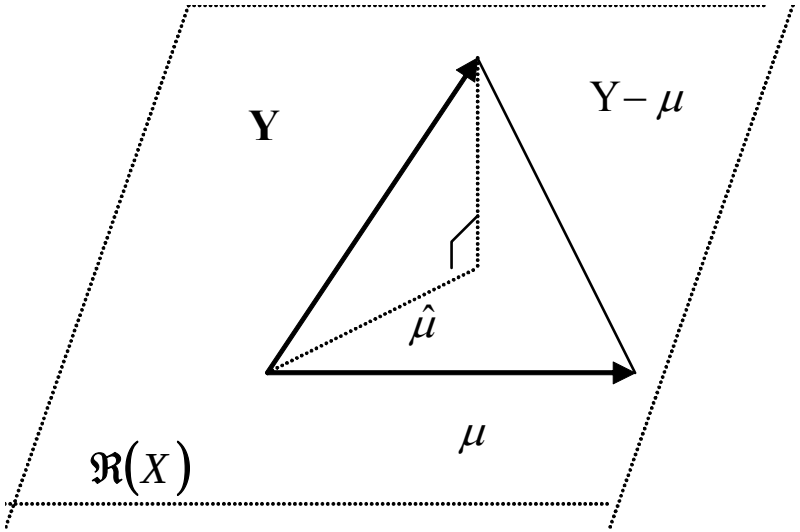
Sea $\hat{\boldsymbol{\mu}}$ la solución del problema de minimización $Min_{\eta} \|\mathrm{Y} - \boldsymbol{\mu}\|^2 = \hat{\boldsymbol{\mu}}$,

➡ Entonces se puede demostrar que $\hat{\boldsymbol{\mu}}$ es la proyección ortogonal de **Y** y es única, siendo el operador de proyección, la denominada *matriz sombrero*, $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathrm{T}}$, puesto que su aplicación a **Y** facilita los valores ajustados o predichos de **Y**, notados $\hat{\mathrm{Y}}$,

$\mathbf{P} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathrm{T}}$ pone el sombrero ^ a la **Y**: $\hat{\mathrm{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{P}\mathbf{Y}$.

TEMA 3: ESTIMACIÓN POR MÍNIMOS CUADRADOS

Gráficamente,



$Y-\hat{\mu}$ es perpendicular al espacio engendrado por las columnas de la matriz de diseño $\Re(X)$.

- Los valores ajustados se notan $\hat{Y}=\hat{\mu}=X\hat{\beta}=PY$ (son únicos).
- Los residuos se definen como las diferencias entre los valores observados y los valores ajustados: $e=Y-\hat{Y}=Y-\hat{\mu}=Y-X\hat{\beta}=(I-P)Y$ (son únicos).
- No ***confundir*** los residuos, con los errores: $\varepsilon=Y-\hat{\mu}=Y-X\beta$.

TEMA 3: ESTIMACIÓN POR MÍNIMOS CUADRADOS

3-1-3.2 Propiedades básicas de inferencia

➡ Sea el modelo con variable de respuesta continua,

$\mathrm{Y} = \mu + \varepsilon = \mathrm{X}\beta + \varepsilon$, donde $\mathbf{Y}$, μ son *nx1,* $\mathbf{X}$ es *nxp* de rango *p* y β es *px1*

y los errores son no sesgados de varianza constante, independientes y distribuidos normalmente: $\boldsymbol{\varepsilon} \approx \mathbf{N}_n\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right)$ o equivalentemente $\mathbf{Y} \approx \mathbf{N}_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right)$.

➡ …Entonces, el **estimador no sesgado de mínima varianza** de β es $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$, estimador por mínimos cuadrados y coincide con el estimador de β por maximización de verosimilitud $\hat{\beta}_{MV}$.

➡ En ausencia de normalidad, los estimadores por mínimos cuadrados no son eficientes, es decir tienen una varianza superior a la varianza de los estimadores MV.

➡ En el modelo, el estimador no sesgado de σ^2 , notado s^2 es eficiente (de mínima varianza),

$$s^2 = \frac{\mathbf{e}^T \cdot \mathbf{e}}{n-p} = \frac{\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^T \cdot \left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n-p} = \frac{SCR}{n-p}$$

TEMA 3: ESTIMACIÓN M. C.

➡ **Teorema de distribución de los estimadores de los parámetros (Th 3.5 Seber (77) pag. 54) :**

1. $\hat{\boldsymbol{\beta}} \approx \mathbf{N}_p\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\right)$.
2. $\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T \mathbf{V}\left[\hat{\boldsymbol{\beta}}\right]^{-1}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)=\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T \mathbf{X}^{\mathbf{T}}\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)/\sigma^2 \approx \chi^2_p$.
3. $\hat{\boldsymbol{\beta}}=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ es independiente de s^2.
4. $SCR/\sigma^2=(n-p)s^2/\sigma^2 \approx \chi^2_{n-p}$.

TEMA 3: ESTIMACIÓN M. C. REGRESIÓN LINEAL SIMPLE

3-1-3.3 Caso particular: la regresión lineal simple

Sea el modelo con variable de respuesta contínua,

$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, donde $\mathbf{Y}$, $\boldsymbol{\eta} = \boldsymbol{\mu}$ son *nx1*, $\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x} \end{bmatrix}$ es *nx2* de rango *2* y $\boldsymbol{\beta}$ es *2x1*

y los errores son no sesgados de varianza constante, independientes y distribuidos normalmente: $\boldsymbol{\varepsilon} \approx \mathbf{N}_n\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right)$ o equivalentemente $\mathbf{Y} \approx \mathbf{N}_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right)$.

- El modelo ordinario de regresión lineal simple presupone una recta de relación entre la variable explicativa **X** asociada al parámetro β_2 y la variable de respuesta **Y**.
- La recta no tiene porqué pasar por el origen y por tanto la matriz de diseño contiene una columna constante con valores 1 a la que se asocia el denominado ***término independiente***, parámetro β_1.
- En resumen, y después de haber descrito la particularización de la notación general al modelo de RLS clásico, lo que se persigue es determinar "de la mejor manera posible" los coeficientes de la recta de regresión que relaciona la variable de respuesta **Y** con la variable explicativa **X**: $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$.

TEMA 3: ESTIMACIÓN M. C. REGRESIÓN LINEAL SIMPLE

Los estimadores por mínimos cuadrados ordinarios se obtienen planteando las ecuaciones normales:

$$S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \cdot (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \sum_k (Y_k - \beta_1 - \beta_2 x_k)^2$$

➡ Las condiciones de primer orden de mínimo de $S(\boldsymbol{\beta})$ son:

$$\nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = \mathbf{0} \leftrightarrow \begin{array}{l} \dfrac{\partial S(\boldsymbol{\beta})}{\partial \beta_2} = 0 = -2\sum_k (Y_k - \beta_1 - \beta_2 x_k) x_k \rightarrow \sum_k Y_k x_k = \hat{\beta}_1 \sum_k x_k + \hat{\beta}_2 \sum_k x_k^2 \\ \dfrac{\partial S(\boldsymbol{\beta})}{\partial \beta_1} = 0 = -2\sum_k (Y_k - \beta_1 - \beta_2 x_k) \rightarrow \sum_k Y_k = n\hat{\beta}_1 + \hat{\beta}_2 \sum_k x_k \rightarrow \overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{x} \end{array}$$

La segunda ecuación indica que la recta de regresión siempre pasa por el punto $(\overline{x}, \ \overline{y})$ y dividiendo la primera por el número de observaciones *n* y restando la segunda se obtiene:

$$\rightarrow \frac{\sum_k Y_k x_k}{n} - \overline{Y}\,\overline{x} = \hat{\beta}_1 \left(\frac{\sum_k x_k}{n} - \overline{x} \right) + \hat{\beta}_2 \left(\frac{\sum_k x_k^2}{n} - \overline{x}^2 \right) \rightarrow Cov(Y, x) = \hat{\beta}_2 s_x^2 \rightarrow \hat{\beta}_2 = \frac{Cov(Y, x)}{s_x^2}$$

➡ La estimación de la varianza del modelo resulta: $s^2 = \dfrac{SCR}{n-2} = \dfrac{\sum_k e_k^2}{n-2}$.

TEMA 3: ESTIMACIÓN M. C. REGRESIÓN LINEAL SIMPLE

➡ La varianza de los estimadores $\hat{\boldsymbol{\beta}} \approx \mathbf{N}_2\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\right)$ puede expresarse en este caso particular como:

$$V\left[\hat{\beta}_2\right] = \frac{\sigma^2}{ns_x^2}$$

$$V\left[\hat{\beta}_1\right] = \frac{\sigma^2}{ns_x^2}\left(1 + \frac{\bar{x}^2}{s_x^2}\right)$$

3-1-3.4 Mínimos cuadrados generalizados

Sea el modelo con variable de respuesta contínua,

$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, donde $\mathbf{Y}$, $\boldsymbol{\eta} = \boldsymbol{\mu}$ son *nx1,* $\mathbf{X}$ es *nxp* de rango *p* y $\boldsymbol{\beta}$ es *px1*

y los errores son no sesgados y correlacionados con distribución normal: $\boldsymbol{\varepsilon} \approx \mathbf{N}_n\left(\mathbf{0}, \sigma^2\mathbf{W}\right)$ o equivalentemente $\mathbf{Y} \approx \mathbf{N}_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}\right)$, donde $\mathbf{W}$ es una matriz simétrica y definida positiva de dimensión *nxn* .

TEMA 3: ESTIMACIÓN M. C. MÍNIMOS CUADRADOS GENERALIZADOS

Si $\mathbf{W}$ en el modelo $\mathbf{Y} \approx \mathbf{N}_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}\right)$ es simétrica y definida positiva entonces puede calcularse una matriz no singular **K** triangular inferior de dimensión *nxn* tal que $\mathbf{W} = \mathbf{K}\mathbf{K}^{\mathbf{T}}$, esta matriz es única, es la factorización de Cholesky de $\mathbf{W}$. Los elementos de **K** se pueden determinar a partir de la factorización de Cholesky o a partir de la descomposición triangular (vista en Optimización Continua) ,

$$\mathbf{W} = \mathbf{L}\mathbf{U} = \mathbf{L}\left(\mathbf{D}\mathbf{L}^{\mathbf{T}}\right) = \mathbf{L}\mathbf{D}^{\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{D}^{\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{L}^{\mathbf{T}} = \left(\mathbf{L}\mathbf{D}^{\frac{1}{2}}\right)\left(\mathbf{D}^{\frac{1}{2}}\mathbf{L}^{\mathbf{T}}\right) = \mathbf{K}\mathbf{K}^{\mathbf{T}}.$$

La matriz **K** que aparece en la factorización de **W** permite calcular las transformaciones lineales de las observaciones (notado $\tilde{\mathbf{Y}}$), la matriz de diseño $\tilde{\mathbf{X}}$ y los errores $\tilde{\boldsymbol{\varepsilon}}$:

1. $\tilde{\mathbf{Y}} = \mathbf{K}^{-1}\mathbf{Y}$ 2. $\tilde{\mathbf{X}} = \mathbf{K}^{-1}\mathbf{X}$ 3. $\tilde{\boldsymbol{\varepsilon}} = \mathbf{K}^{-1}\boldsymbol{\varepsilon}$
4. $\mathbf{K}^{-1}\mathbf{Y} = \mathbf{K}^{-1}\boldsymbol{\eta} + \mathbf{K}^{-1}\boldsymbol{\varepsilon} = \mathbf{K}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{K}^{-1}\boldsymbol{\varepsilon} \leftrightarrow \tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$.

➡ Ahora el modelo transformado $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$ tiene errores no correlacionados de varianza σ^2 ,

$$\mathbf{V}\left[\tilde{\mathbf{Y}}\right] = \mathbf{V}\left[\tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}\right] = \mathbf{V}\left[\tilde{\boldsymbol{\varepsilon}}\right] = \mathbf{V}\left[\mathbf{K}^{-1}\boldsymbol{\varepsilon}\right] = \mathbf{K}^{-1}\mathbf{V}\left[\boldsymbol{\varepsilon}\right]\left(\mathbf{K}^{-1}\right)^{T} = \sigma^2\mathbf{K}^{-1}\mathbf{W}\mathbf{K}^{-\mathbf{T}} = \sigma^2\mathbf{K}^{-1}\mathbf{K}\mathbf{K}^{T}\mathbf{K}^{-\mathbf{T}} = \sigma^2\mathbf{I}_n$$

TEMA 3: ESTIMACIÓN M. C. MÍNIMOS CUADRADOS GENERALIZADOS

Síntesis de la estimación e inferencia por mínimos cuadrados con respuesta normal:

MODELO	$\mathbf{Y}=\mathbf{X}\beta+\varepsilon,\ \mathbf{V}[\varepsilon]=\sigma^2\mathbf{I}$	$\mathbf{Y}=\mathbf{X}\beta+\varepsilon,\ \mathbf{V}[\varepsilon]=\sigma^2\mathbf{D}$	$\mathbf{Y}=\mathbf{X}\beta+\varepsilon,\ \mathbf{V}[\varepsilon]=\sigma^2\mathbf{W}=\sigma^2\mathbf{K}\mathbf{K}^T$
Transformación	$\mathbf{Y}\to\mathbf{Y}\ ,\ \mathbf{X}\to\mathbf{X}$	$\mathbf{Y}\to\mathbf{D}^{-\frac{1}{2}}\mathbf{Y}\ ,\ \mathbf{X}\to\mathbf{D}^{-\frac{1}{2}}\mathbf{X}$	$\mathbf{Y}\to\mathbf{K}^{-1}\mathbf{Y}\ ,\ \mathbf{X}\to\mathbf{K}^{-1}\mathbf{X}$
$S(\beta)$ *Forma cuadrádica a minimizar*	$(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)$	$(\mathbf{Y}-\mathbf{X}\beta)^T\mathbf{D}^{-1}(\mathbf{Y}-\mathbf{X}\beta)$	$(\mathbf{Y}-\mathbf{X}\beta)^T\mathbf{W}^{-1}(\mathbf{Y}-\mathbf{X}\beta)$
Ecuaciones normales $\boxed{\mathbf{S}\beta=\mathbf{Q}}$	$\mathbf{S}=\mathbf{X}^T\mathbf{X},\ \mathbf{Q}=\mathbf{X}^T\mathbf{Y}$	$\mathbf{S}=\mathbf{X}^T\mathbf{D}^{-1}\mathbf{X},\ \mathbf{Q}=\mathbf{X}^T\mathbf{D}^{-1}\mathbf{Y}$	$\mathbf{S}=\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X},\ \mathbf{Q}=\mathbf{X}^T\mathbf{W}^{-1}\mathbf{Y}$
Estimadores $\hat{\beta}$	$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$	$(\mathbf{X}^T\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}^{-1}\mathbf{Y}$	$(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-1}\mathbf{Y}$
Varianza de $\hat{\beta}$	$\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$	$\sigma^2(\mathbf{X}^T\mathbf{D}^{-1}\mathbf{X})^{-1}$	$\sigma^2(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}$
SCR	$\mathbf{Y}^T\mathbf{Y}-(\mathbf{X}^T\mathbf{Y})^T\hat{\beta}$	$\mathbf{Y}^T\mathbf{D}^{-1}\mathbf{Y}-(\mathbf{X}^T\mathbf{D}^{-1}\mathbf{Y})^T\hat{\beta}$	$\mathbf{Y}^T\mathbf{W}^{-1}\mathbf{Y}-(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{Y})^T\hat{\beta}$

3-1-4. TEMA 3: CONTRASTES DE HIPÓTESIS EN MODELOS NORMALES

➡ Sea el modelo con variable de respuesta continua y normal,

$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, donde $\mathbf{Y}$, $\boldsymbol{\eta} = \boldsymbol{\mu}$ son *nx1,* $\mathbf{X}$ es *nxp* de rango *p* y $\boldsymbol{\beta}$ es *px1*

y los errores son no sesgados y no correlacionados de varianza constante: $\mathbf{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ y $\mathbf{V}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n$. Los estimadores ordinarios por mínimos cuadrados se vienen notando $\hat{\boldsymbol{\beta}}$.

➡ Sea el modelo con variable de respuesta continua y normal,

$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, donde $\mathbf{Y}$, $\boldsymbol{\eta} = \boldsymbol{\mu}$ son *nx1,* $\mathbf{X}$ es *nxp* de rango *p* y $\boldsymbol{\beta}$ es *px1*

y los errores son no sesgados y no correlacionados de varianza constante: $\mathbf{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ y $\mathbf{V}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n$ sujeto a un conjunto de restricciones lineales $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ que ***definen una hipótesis a contrastar denominada H,*** donde **A** es una matriz qxp de rango *q < p.* Los estimadores ordinarios por mínimos cuadrados se notarán $\hat{\boldsymbol{\beta}}_H$.

CONTRASTES EN MODELOS NORMALES POR VARIANZA INCREMENTAL

3. Si la hipótesis H es cierta, entonces se puede demostrar que

$$F = \frac{(SCR_H - SCR)/q}{SCR/(n-p)} = \frac{\left(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)^T \left(\mathbf{A}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{A}^T\right)^{-1}\left(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)}{q\,s^2} \rightarrow F_{q,n-p}$$

➡La justificación del test puede realizarse a partir de la estimación de la varianza del modelo por la suma de cuadrados residual del modelo restringido s_H^2, tiene por esperanza matemática:

$$\mathrm{E}\left[s_H^2\right] = \mathrm{E}\left[\frac{SCR_H - SCR}{q}\right] = \sigma^2 + \frac{\left(\mathbf{A}\beta - \mathbf{c}\right)^T \left(\mathbf{A}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathbf{A}^T\right)^{-1}\left(\mathbf{A}\beta - \mathbf{c}\right)}{q}$$

➡… pero $\mathbf{E}\left[s_H^2\right] = \sigma^2 + \delta$ con $\delta \geq 0$ facilita una estimación sesgada de la varianza del modelo ya que, $\mathbf{E}\left[s^2\right] = \mathbf{E}\left[\frac{SCR}{n-p}\right] = \sigma^2$. Sin embargo, si H es cierta entonces $\delta = 0$ tanto s_H^2 como s^2 son estimadores no sesgados de σ^2 y por tanto el estadístico $\boxed{F = \frac{s_H^2}{s^2}}$ toma un valor alrededor de 1. H es rechazada si el estadístico F toma valores significativamente grandes.

CONTRASTES EN MODELOS NORMALES. REGIONES E INTERVALOS

3-1-4.1 Cálculo de intervalos de confianza

Los intervalos de confianza para los coeficientes individuales β_i se obtienen mediante la fórmula habitual:

$$t = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} \approx t_{n-p} \quad \rightarrow \quad \hat{\beta}_i \pm t_{n-p}^{\alpha/2} \hat{\sigma}_{\hat{\beta}_i} \quad \text{donde} \quad \hat{\sigma}_{\hat{\beta}_i} = s\sqrt{\left(X^T X\right)^{-1}_{ii}} \quad \text{y} \quad s = \hat{\sigma} = \sqrt{\frac{SCR}{n-p}}$$

$t_{n-p}^{\alpha/2}$ es el valor correspondiente al estadístico *t de Student* para el cálculo de un intervalo de confianza bilateral a un nivel α con los grados de libertad correspondientes a la estimación de varianza del modelo (*n-p*).

➡ Los contrastes de significación se realizarán a partir del cálculo del estadístico $t = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} \approx t_{n-p}$ y su comparación con el valor de la ley *t de Student* de *n-p* grados de libertad al nivel de confianza, unilateral o bilateral, deseado.

Los coeficientes β_i son estadísticamente dependientes y por tanto los intervalos de confianza individuales pueden dar una imagen errónea de sus valores conjuntos.

3-1-5. TEMA 3: EL COEFICIENTE DE CORRELACIÓN MÚLTIPLE

➡ Una medida de la bondad del modelo ajustado a los datos, en los modelos lineales normales, es el ***coeficiente de correlación múltiple R,*** definido como el coeficiente de correlación muestral entre los datos y_k y los valores ajustados $\hat{y}_k$:

$$R = \frac{\sum_k \left(y_k - \bar{y}\right)\left(\hat{y}_k - \bar{\hat{y}}\right)}{\left\{\sum_k \left(y_k - \bar{y}\right)^2 \sum_k \left(\hat{y}_k - \bar{\hat{y}}\right)^2\right\}^{1/2}}$$

➡ El estadístico $\mathbf{R^2}$ se denomina ***coeficiente de determinación***.

➡ La descomposición de la suma de cuadrados total (SCT) como suma de cuadrados explicada (SCE) por el modelo más suma de cuadrados residual (SCR) es un resultado conocido y muy útil en modelos lineales que incluyen término independiente (por simplicidad, supóngase que es el primer parámetro β_1)

1. SCT= $\sum_k \left(y_k - \bar{y}\right)^2$ donde $\bar{y} = \frac{1}{n}\sum_k y_k$ es la media muestral de las observaciones.
2. SCE= $\sum_k \left(\hat{y}_k - \bar{y}\right)^2$ y SCR= $\sum_k \left(y_k - \hat{y}_k\right)^2$.

TEMA 3: EL COEFICIENTE DE CORRELACIÓN MÚLTIPLE …

3. **SCT=SCE+SCR**, es decir, $\sum_k (y_k - \bar{y})^2 = \sum_k (\hat{y}_k - \bar{y})^2 + \sum_k (y_k - \hat{y}_k)^2$, ya que

$$\sum_k (y_k - \bar{y})^2 = \sum_k ((y_k - \hat{y}_k) + (\hat{y}_k - \bar{y}))^2 = \sum_k (\hat{y}_k - \bar{y})^2 + \sum_k (y_k - \hat{y}_k)^2 + 2\sum_k (y_k - \hat{y}_k)(\hat{y}_k - \bar{y}) =$$
$$\sum_k (\hat{y}_k - \bar{y})^2 + \sum_k (y_k - \hat{y}_k)^2$$

donde,

$$\sum_k (y_k - \hat{y}_k)(\hat{y}_k - \bar{y}) = \sum_k (y_k - \hat{y}_k)\hat{y}_k - \bar{y}\sum_k (y_k - \hat{y}_k) = \sum_k (y_k - \hat{y}_k)\hat{y}_k = (\mathbf{Y} - \hat{\mathbf{Y}})^{\mathrm{T}} \hat{\mathbf{Y}} =$$
$$= (\mathbf{Y} - \mathbf{PY})^{\mathrm{T}} \mathbf{PY} = \mathbf{Y}^{\mathrm{T}} (\mathbf{I} - \mathbf{P}) \mathbf{PY} = \mathbf{0}$$

- El ***coeficiente de determinación*** puede reescribirse ahora:
- O equivalentemente

$$SCR = (1 - R^2) SCT .$$

$$R^2 = \frac{\sum_k (\hat{y}_k - \bar{\hat{y}})^2}{\sum_k (y_k - \bar{y})^2} = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

TEMA 3: EL COEFICIENTE DE CORRELACIÓN MÚLTIPLE …

3-1-5.1 Propiedades del coeficiente de correlación múltiple

1. $|R| \leq 1$ y si $|R| = 1$ existe una relación funcional exacta entre la respuesta y los regresores.
2. R es el coeficiente de correlación lineal simple entre los valores observados y los valores ajustados.
3. $100(1 - R^2)$ representa el % de variabilidad no explicada por el modelo.

3-1-5.2 R^2-adjusted

➡ Algunos estadísticos prefieren emplear al valorar los modelos de regresión ordinarios, el denominado ***coeficiente de determinación ajustado***, ajuste que se refiere a la introducción de los correspondientes grados de libertad de *SCT* y *SCR*, muy relacionado con el estadístico **C_p de Mallows**:

$$R_a^2 = 1 - \frac{SCR/n-p}{SCT/n-1} = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-p}\right)$$

➡ El coeficiente de determinación ajustado siempre es inferior al coeficiente de determinación y puede tomar valores negativos. Si R^2 siempre crece al incrementar el número de regresores, ya que la suma de cuadrados residual siempre se reduce, R_a^2 únicamente sufre un incremento al añadir uno o más nuevos regresores, si el estadístico de Fisher F correspondiente al test de significación de los nuevos regresores toma un valor superior a 1.

3-1-6. TEMA 3: TEST GLOBAL DE REGRESIÓN. TABLA ANOVA

- El test global de regresión es un caso particular del contraste de hipótesis múltiples en modelos con término independiente, donde la hipótesis **H** a contrastar es que todos los parámetros son cero, excepto el correspondiente al término independiente: $\beta_2 = 0, \ldots, \beta_p = 0$.

$$F = \frac{(SCR_H - SCR)/q}{SCR/(n-p)} = \frac{(SCT - SCR)/(p-1)}{SCR/(n-p)} = \frac{SCE/(p-1)}{SCR/(n-p)} = \frac{SCE}{(p-1)s^2} \approx F_{p-1,n-p,}$$

- El contraste global de regresión se ve clarificado si la descomposición de la varianza se escribe en forma de tabla ANOVA, algo habitual en los paquetes estadísticos:

TABLA ANOVA	*Descomposición*	*Grados libertad*	*Varianza*	*Contraste*
SCE	$\sum_k (\hat{y}_k - \bar{y})^2$	*p-1*	$s^2_{\exp}$ *=SCE/(p-1)*	
SCR	$\sum_k (y_k - \hat{y}_k)^2$	*n-p*	s^2 *=SCR/(n-p)*	*F=* $S^2_{\exp} / s^2$
SCT	$\sum_k (y_k - \bar{y})^2$	*n-1*	s^2_Y *=SCT/(n-1)*	

3-1-7. TEMA 3: DISTRIBUCIÓN DE LOS VALORES AJUSTADOS

➡ Sea $\hat{Y}_k$ el valor ajustado para la observación k-ésima que tiene por valores de los regresores $\mathbf{x}_\mathbf{k}^\mathbf{T} = \begin{pmatrix} 1 & x_2 & \dots & x_p \end{pmatrix}$: $\hat{Y}_k = \mathbf{x}_\mathbf{k}^\mathbf{T}\hat{\boldsymbol{\beta}}$.

1. $\mathrm{E}\left[\hat{Y}_k\right] = \mathrm{E}\left[\mathbf{x}_\mathbf{k}^\mathbf{T}\hat{\boldsymbol{\beta}}\right] = \mathbf{x}_\mathbf{k}^\mathbf{T}\boldsymbol{\beta}$.
2. $V\left[\hat{Y}_k\right] = V\left[\mathbf{x}_\mathbf{k}^\mathbf{T}\hat{\boldsymbol{\beta}}\right] = \mathbf{x}_\mathbf{k}^\mathbf{T} V\left[\hat{\boldsymbol{\beta}}\right]\mathbf{x}_\mathbf{k} = \sigma^2 \mathbf{x}_\mathbf{k}^\mathbf{T}\left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-1}\mathbf{x}_\mathbf{k} = \sigma^2 p_{kk}$, donde p_{kk} es el término diagonal k-ésimo de la matriz de proyección, que por ser idempotente tiene valores entre *1/n* y *1*. La varianza del valor ajustado es mínima si se encuentra en el centro de gravedad de los regresores.
3. Se distribuye normalmente.

➡ En la literatura sobre modelos de regresión, la distribución de los valores ajustados se suele denominar distribución de los valores medios (no se considera el término de residuo) y los intervalos de confianza se calculan en base a la ley *t de Student de n-p grados de libertad*:

$$\frac{\hat{Y} - \mathbf{x}^\mathbf{T}\boldsymbol{\beta}}{\sigma_{\hat{Y}}} \approx N(0,1) \rightarrow \; t = \frac{\hat{Y} - \mathbf{x}^\mathbf{T}\boldsymbol{\beta}}{\hat{\sigma}_{\hat{Y}}} \approx t_{n-p} \quad \text{donde } \hat{\sigma}_{\hat{Y}} = \hat{\sigma}\sqrt{\mathbf{x}^\mathbf{T}\left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-1}\mathbf{x}} = s\sqrt{\mathbf{x}^\mathbf{T}\left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-1}\mathbf{x}}$$

Con lo que un intervalo de confianza bilateral a un nivel α para el verdadero valor medio viene determinado por: $\hat{Y} \pm t_{n-p}^{\alpha/2}\hat{\sigma}_{\hat{Y}}$.

3-1-8. TEMA 3: DIAGNOSIS Y VALIDACIÓN DEL MODELO

➡ El análisis de los residuos constituye una herramienta práctica y que entra por los ojos para la validación de las hipótesis aparentemente muy teóricas de la regresión lineal y por tanto, para garantizar las propiedades estadísticas de los estimadores del *modelo asumido*:

Sea el modelo con variable de respuesta contínua,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$, donde $\mathbf{Y}$, $\boldsymbol{\eta} = \boldsymbol{\mu}$ son *nx1,* $\mathbf{X}$ es *nxp* de rango *p* y $\boldsymbol{\beta}$ es *px1*

y los errores son no sesgados de varianza constante, no correlacionados y distribuidos normalmente: $\boldsymbol{\varepsilon} \approx \mathbf{N}_n\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right)$ o equivalentemente $\mathbf{Y} \approx \mathbf{N}_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right)$.

➡ Los residuos son las diferencias entre los valores observados y los valores ajustados por el modelo: $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P})\mathbf{Y} = (\mathbf{I} - \mathbf{P})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} - \mathbf{P}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{P})\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{P})\boldsymbol{\varepsilon}$ y de ahí que a veces se les denomine errores observados, término que únicamente debiera emplearse si el modelo es correcto.

TEMA 3: DIAGNOSIS Y VALIDACIÓN DEL MODELO

➡ Los residuos tienen una distribución:

1. $\mathrm{E}[\mathbf{e}] = \mathrm{E}[\mathbf{Y} - \hat{\mathbf{Y}}] = \mathrm{E}[\mathbf{Y}] - \mathrm{E}[\mathbf{X}\hat{\boldsymbol{\beta}}] = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$.
2. $V[\mathbf{e}] = V[(\mathbf{I}-\mathbf{P})\mathbf{Y}] = (\mathbf{I}-\mathbf{P})^T V[\mathbf{Y}](\mathbf{I}-\mathbf{P}) = \sigma^2(\mathbf{I}-\mathbf{P})^2 = \sigma^2(\mathbf{I}-\mathbf{P})$.

➡ Aunque los errores $\boldsymbol{\varepsilon}$ sean independientes y de varianza constante, los residuos $\mathbf{e}$ no son independientes, ni tienen la misma varianza: $\mathbf{V}[\mathbf{e}] = \sigma^2(\mathbf{I}-\mathbf{P}) = \sigma^2\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right)$ o bien individualmente, la varianza del residuo i-ésimo es $V[e_i] = \sigma^2(1-p_{ii})$, donde p_{ii} es el elemento diagonal i-ésimo de la matriz de proyección $\mathbf{P}$.

➡ **Para comparar los residuos entre sí, suele ser más ilustrativo transformarlos, encontrándose en la literatura y los paquetes estadísticos diversas posibilidades:**

1. El residuo escalado c_i se define como el residuo dividido por el estimador de la desviación estandard del modelo s, $c_i = \dfrac{e_i}{s}$, lo cual no es demasiado incorrecto si no existen grandes variaciones en $V[e_i] = \sigma^2(1-p_{ii})$.

TEMA 3: DIAGNOSIS Y VALIDACIÓN DEL MODELO

➡ Sin embargo, en los residuo estandarizados numerador y denominador son dependientes, ya que e_i se ha empleado para estimar la varianza del modelo $\hat{\sigma}^2 = s^2$, lo que puede solucionarse eliminando la observación i-ésima de los cálculos y estimando una regresión con los restantes *n-1* datos, lo que daría lugar a los estimadores $\hat{\beta}_{(i)}$ y $s^2_{(i)}$ que podrían relacionarse formalmente con los estimadores $\hat{\beta}$ y permiten definir los denominados residuos estudentizados.

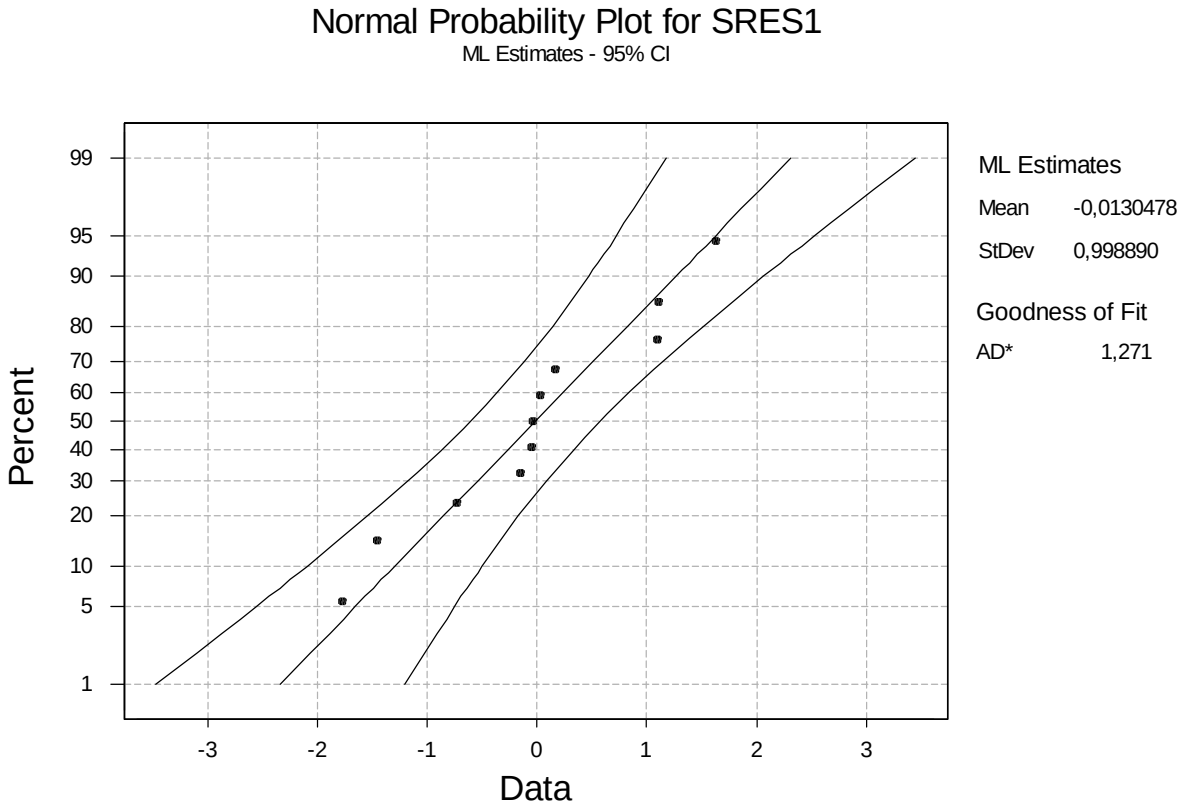
2. El residuo estandarizado d_i se define como el residuo dividido por su desviación tipo: $$d_i = \frac{e_i}{s\sqrt{1-p_{ii}}}.$$

3. El residuo estudentizado r_i se define como $$r_i = \frac{e_i}{s_{(i)}\sqrt{1-p_{ii}}}$$ donde $$s^2_{(i)} = \frac{(n-p)s^2 - e_i^2/(1-p_{ii})}{n-p-1}.$$

Los residuos estudentizados siguen una distribución *t de Student* con *n-p-1* grados de libertad bajo modelos de respuesta continua sujetos a las hipótesis ordinarias.

TEMA 3: DIAGNOSIS Y VALIDACIÓN DEL MODELO

El análisis de los residuos permite concluir si las hipótesis asumidas son correctas o no son correctas (y por qué no lo son) y se realiza en base a herramientas gráficas de la estadística descriptiva:



1. Histograma de los residuos: se persigue ver que esten centrados en el cero y que su distribución sea aproximadamente normal. Un boxplot puede ayudar a identificar los outliers de los residuos.

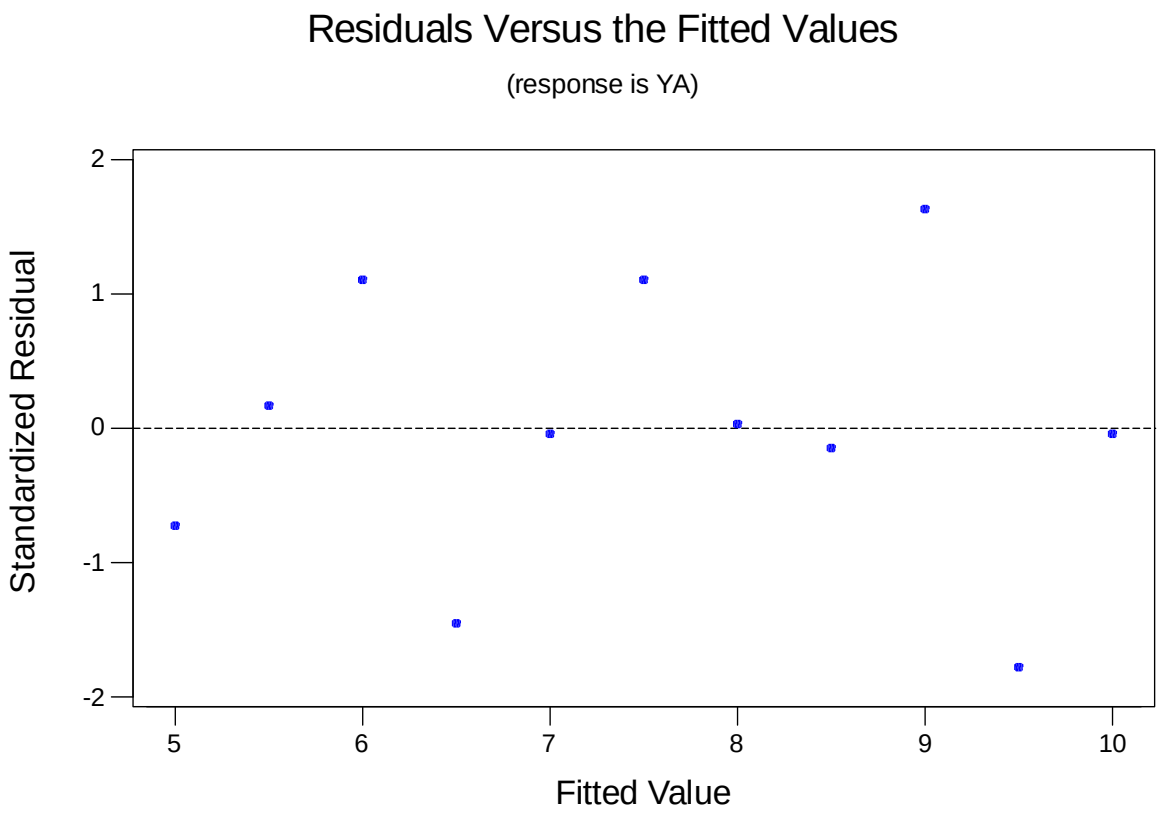
2. Tambien se puede emplear un *Normal Probability Plot* de los residuos estandarizados o estudentizados (recta de Henry).

TEMA 3: DIAGNOSIS Y VALIDACIÓN DEL MODELO

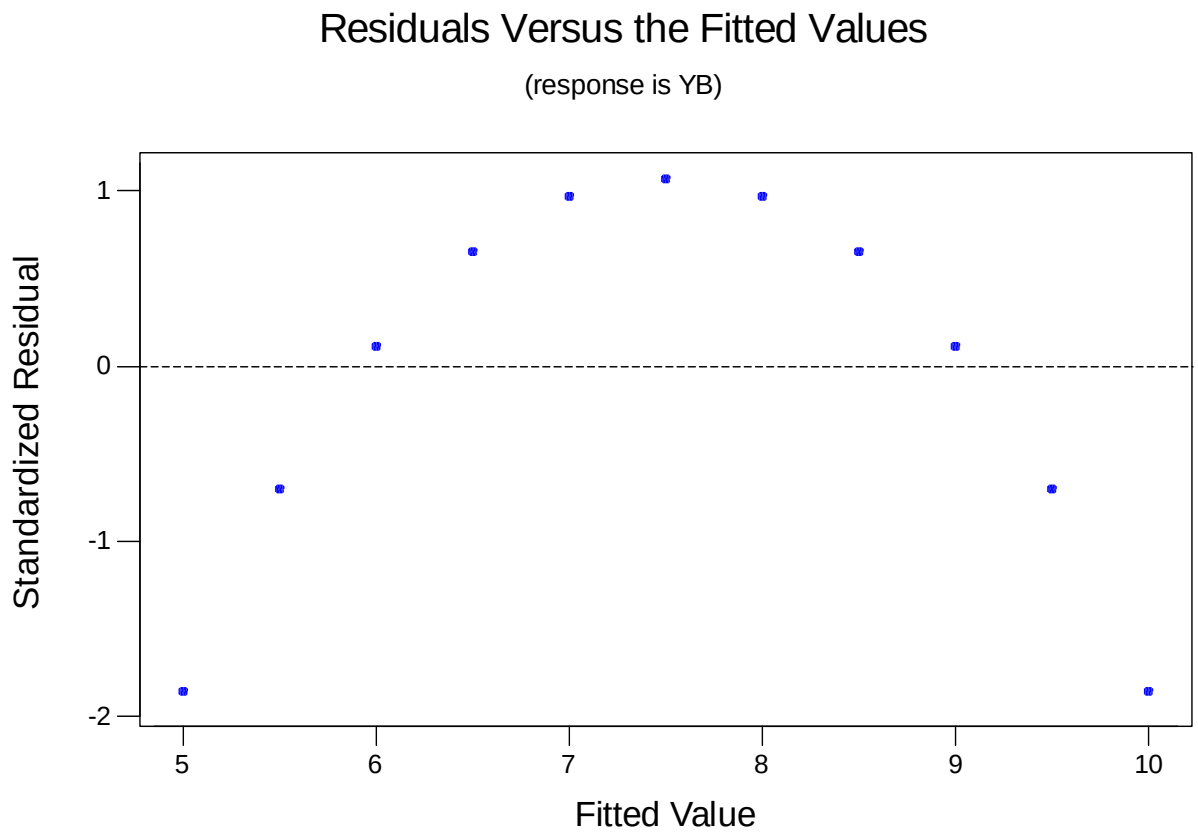
- Ante desviaciones de la normalidad los contrastes basados en las leyes t de Student y F de Fisher, se convierten en aproximados, además de perder la eficiencia de los estimadores, se debe proceder mediante la aplicación de una transformación adecuada a la variable de respuesta **Y**.
- La potencia de los tests de normalidad es baja, ya que aunque los errores no sean normales, los residuos son combinación lineal de los errores y en virtud del Teorema Central del Límite tienden a la normalidad, a pesar de ser dependientes (para *n>30*, no suele ser crucial este último aspecto).
- Los residuos suelen estar correlacionados con las observaciones Y_i, pero no con los valores ajustados $\hat{Y}_i$, de ahí la selección de estos últimos para la realización de los gráficos indicados. (Ejemplo de Ascombe, Peña pp. 263-264 vol.2).

3. Diagrama bivariante de los residuos (en ordenadas) frente a los valores ajustados: $\boxed{e_i \ vs \ \hat{Y}_i}$, o mejor $\boxed{d_i \ vs \ \hat{Y}_i}$ o $\boxed{r_i \ vs \ \hat{Y}_i}$. Estos gráficos suelen indicar la falta de linealidad (requiere transformación de las observaciones Y_i y/o introducción de nuevos regresores), heterocedasticidad (requiere transformación de las observaciones Y_i) e identifican valores atípicos.

TEMA 3: DIAGNOSIS Y VALIDACIÓN DEL MODELO

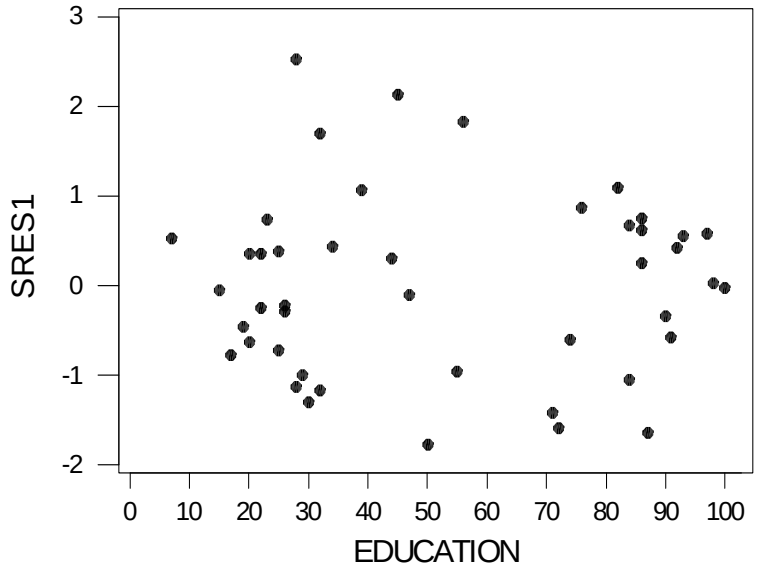


TEMA 3: DIAGNOSIS Y VALIDACIÓN DEL MODELO



TEMA 3: DIAGNOSIS Y VALIDACIÓN DEL MODELO

➡ La visualización de una banda horizontal de residuos centrada en $Y=0$ indica satisfacción de las hipótesis. Los residuos estudentizados tienen una distribución de referencia y por tanto, la presencia de puntos con valores negativos o positivos más allá de un cierto nivel de confianza seleccionado para la distribución de referencia (t de Student de *n-p-1* grados de libertad) indica un valor atípico del residuo o *outlier* en los residuos.



4. ***Diagramas bivariantes de los residuos (en ordenadas) frente a cada uno de los regresores*** (excepto el término independiente).

➡ Ayudan a identificar si la falta de linealidad (paliable mediante la transformación de las observaciones **Y** o mediante la introducción de un término cuadrático del regresor correspondiente) o la heterocedasticidad (paliable mediante transformación de **Y**) son debidas a algún regresor en particular. La visualización de una banda horizontal de residuos indica satisfacción de las hipótesis.

TEMA 3: DIAGNOSIS Y VALIDACIÓN DEL MODELO

5. Representación de los residuos (en ordenadas) frente a variables explicativas omitidas en el modelo.

- Se emplean para detectar la posible influencia en los residuos de una variable no incluída en el modelo.
- Un caso particular consiste en la representación de los residuos en función del tiempo u orden de los datos; en este caso se pueden calcular los coeficientes de autocorrelación observados de los residuos de orden *k*, $r(k)=\dfrac{\sum_i e_i e_{i+k}}{\sum_i e_k^2}$ y aplicar el siguiente test para muestras grandes: bajo la hipótesis de independencia (todos los coeficientes de autocorrelación teórica de orden *k* igual a cero), entonces el estadístico $Q=n(n+2)\sum_{k=1}^{m}\dfrac{r^2(k)}{n-k+1}\approx \chi^2_{m-2}$.
- Un test alternativo es el test de Durbin-Watson, basado en el estadístico del mismo nombre para el contraste de la hipótesis nula que el coeficiente de autocorrelación de primer orden (r(1)) es cero, cuya justificación y examen de las tablas correspondientes no es nada trivial.

TEMA 3: DIAGNOSIS Y VALIDACIÓN DEL MODELO

3-1-8.1 Transformación de Box-Cox

- La familia Box-Cox es una familia de transformaciones de variables aleatorias que se emplean para conseguir normalidad o homocedasticidad:

$$h(Y) = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \log Y & \lambda = 0 \end{cases}$$

- La justificación de la definición viene de una propiedad básica que relaciona las varianzas de transformaciones de la variables aleatorias y que trasladada a la notación empleada en la presente sección es:

$$V\left[h(Y)\right] = V\left[Y\right]\left|h'(Y)\right|^2$$

Por tanto, si la varianza de la variable transformada se desea que sea constante, entonces la derivada de la transformación debe ser inversamente proporcional a la desviación típica de la variable original.

- La constante λ de la transformación puede estimarse gráficamente o por máxima verosimilitud.

TEMA 3: DIAGNOSIS Y VALIDACIÓN DEL MODELO. FAMILIA BOX-COX

Algunos casos particulares son:

- Si la desviación típica de Y es proporcional al cuadrado de su media, $S \propto \overline{Y}^2$, la constante $\boxed{\lambda = -1}$ facilita la transformación recíproca adecuada para esta situación.
- Si la desviación típica de Y es $S \propto \overline{Y}^{3/2}$, la constante $\boxed{\lambda = -1/2}$ facilita la transformación inversa de la raiz adecuada para esta situación.
- Si la desviación típica de Y es proporcional a su media, $S \propto \overline{Y}$, la constante $\boxed{\lambda = 0}$ facilita la transformación logarítmica adecuada para esta situación.
- Si la desviación típica de Y es proporcional a la raiz cuadrada de su media, $S \propto \overline{Y}^{1/2}$, la constante $\boxed{\lambda = 1/2}$ facilita la transformación raiz cuadrada adecuada para esta situación.

3-1-9. TEMA 3: OBSERVACIONES INFLUYENTES A PRIORI Y A POSTERIORI

Resulta fácil desarrollar ejemplos que ponen de manifiesto que existen observaciones que tienen mucha mayor influencia en las propiedades del modelo que otras, hasta el extremo que en presencia de 100 valores observados, las propiedades de los estimadores dependan únicamente de unos pocos de esos valores.

- Este aspecto está relacionado con la fiabilidad del modelo en la realización de predicciones, y parece más conveniente un modelo que venga avalado por la totalidad de la muestra empleada para su estimación, que no aquel otro que sólo dependa de unas pocas observaciones.
- El estudio de los valores influyentes a priori determinará la robustez del diseño de recogida de los datos y el estudio de los valores influyentes a posteriori determinará la robustez de los parámetros estimados.
- No hay que confundir observaciones influyentes con residuos atípicos: una observación influyente puede tener o no un residuo estadísticamente grande, y viceversa, un residuo atípico no fuerza no implica que la observación correspondiente sea influyente.

  Por ejemplo, en regresión lineal simple se puede introducir una observación muy atípica (residuo muy alto) en el valor medio de la variable explicativa, la observación no resultará influyente, sin embargo el coeficiente de determinación se resentirá y disminuirá debido al incremento de la suma de cuadrados residuales.

TEMA 3: OBSERVACIONES INFLUYENTES A PRIORI Y A POSTERIORI

3-1-9.1 Observaciones influyentes a priori

Los puntos **x** ($\mathbf{x} \in \Re^p$) heterogéneos respecto el centro de gravedad de los valores de los regresores identifican las observaciones influyentes *a priori* y corresponden a valores grandes en la diagonal de la matriz de proyección $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$ notados $p_{ii} = \mathbf{x_i}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x_i}$ (al ser simétrica e idempotente tiene *p* vaps 1 y *n-p* vaps 0) y se puede demostrar que:

1. $\frac{1}{n} \le p_{ii} \le 1$

2. Al tratarse de una matriz idempotente y simétrica: $\mathbf{dim}(\mathbf{P}) = traza(\mathbf{P}) = \sum_i p_{ii} = p$.

➡ Lo que permite determinar su media $\overline{p} = \frac{\sum_i p_{ii}}{n} = \frac{p}{n}$ y a partir de otros estadísticos descriptivos calculables, los resultados de Belsley *et al.*, indican que si las variables explicativas proceden de una distribución normal indican que pueden considerarse valores influyentes *a priori* aquellos puntos con $p_{ii} > 2\overline{p}$.

TEMA 3: OBSERVACIONES INFLUYENTES A PRIORI Y A POSTERIORI

➡ Los valores p_{ii} suelen denominarse en los paquetes estadísticos factores de anclaje o *leverage* y miden la distancia entre una observación $\mathbf{x_i}$ y el centro de gravedad de las observaciones,

1. Si la observación está muy alejada $p_{ii} \to 1$ y $V[e_i] = \sigma^2(1 - p_{ii}) = 0$, indicando que sea cual sea el valor observado Y_i, su residuo es siempre igual a su valor esperado, cero, por lo que la ecuación de regresión estimada por mínimos cuadrados ordinarios pasará siempre por dicho punto.
2. Si la observación está en el centro de gravedad entonces $p_{ii} \to 1/n$ y $V[e_i]$ es máxima, indicando que sea cual sea el valor observado Y_i, puede tener por efecto una reducción drástica del coeficiente de determinación; sin embargo, nunca será un valor detectado como influyente a priori según los criterios descritos.

TEMA 3: OBSERVACIONES INFLUYENTES A PRIORI Y A POSTERIORI

3-1-9.2 Observaciones influyentes a posteriori

Una observación influyente a posteriori implica que su inclusión:

1. Modifica el vector de parámetros estimados $\hat{\boldsymbol{\beta}}$.
2. Modifica los valores ajustados $\hat{\mathbf{Y}}$.
3. Su valor ajustado es muy bueno cuando se incluye la observación en el proceso de estimación por mínimos cuadrados ordinarios, pero su valor ajustado es muy malo si se ha omitido la observación en el proceso de estimación.

➡ La influencia de una observación en la determinación de los estimadores por mínimos cuadrados ordinarios se determina mediante la distancia de Cook. Una observación muy influyente a priori, puede que no sea influyente a posteriori (se pueden construir ejemplos fácilmente). Cook propuso una medida de la influencia a posteriori de una observación muy efectiva, a través de la distancia de Cook D_i :

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)^{\mathrm{T}} \mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)}{p\,s^2} = \left(\frac{e_i}{s\sqrt{1-p_{ii}}}\right)^2 \left(\frac{p_{ii}}{1-p_{ii}}\right)\frac{1}{p} \approx F_{p,n-p}$$

donde $\hat{\boldsymbol{\beta}}_{(i)}$ son los estimadores obtenidos después de la supresión de la observación i-ésima.

TEMA 3: OBSERVACIONES INFLUYENTES A PRIORI Y A POSTERIORI

- De manera que un criterio para la determinación de las observaciones influyentes a posteriori consiste en comparar su distancia de Cook D_i con el valor de la ley de Fisher correspondiente al nivel de confianza seleccionado $F^{\alpha}_{p,n-p}$: un valor con distancia de Cook elevada, $D_i > F^{\alpha}_{p,n-p}$, denota una observación influyente a posteriori.

- Recordar que la esperanza matemática de una ley de Fisher de p y q grados de libertat es $q/(q-2)$ $q>2$.

- Un criterio práctico propuesto por Chatterjee y Hadi (88) justifica un umbral máximo para la distancia de Cook de 4/(n-p).

3-1-10. TEMA 3: SELECCIÓN DEL MEJOR MODELO

El establecer una ecuación de regresión para una respuesta Y en términos de unos predictores o regresores $(X_1,\ldots,X_p)$ que pueden ser transformaciones de las variables explicativas originales $(Z_1,\ldots,Z_q)$ sintetiza dos criterios opuestos, lo que se denomina *criterio de parsimonia*:

1. La ecuación tiene que ser útil para finalidades predictivas, de manera que se incluíran tantos regresores como sea necesario para que los valores ajustados sean fiables.
2. Los modelos con muchos regresores tienen un alto coste de obtención y mantenimiento de la información, de manera que el modelo debe incluir el mínimo de regresores necesario.

➡ En la práctica, ***es inviable*** la generación y análisis de todos las posibles ecuaciones de regresión, para la selección de la más conveniente. Falta indicar que un buen modelo, debe mostrar un análisis de los residuos satisfactorio y un estudio de los valores influyentes, sería deseable la consecución de modelos sin residuos atípicos, ni valores influyentes *a posteriori*.

TEMA 3: SELECCIÓN DEL MEJOR MODELO

Los elementos que se han expuesto hasta el momento y que permiten valorar la calidad de una ecuación de regresión son:

1. El coeficiente de determinación, R^2. Se estabiliza cuando el número de regresores incluídos es satisfactorio, aunque puede haber más de la cuenta, ya que se incrementa (no linealmente) al incrementarse el número de regresores.

   Considerar para facilitar la tarea el coeficiente de determinación ajustado, R_a^2.

2. La estabilización del estimador clásico de la varianza del modelo, que ante modelos insatisfactorios recuérdese que se ha visto que es sesgado y por tanto, debe denominarse residuo cuadrático medio.
3. El análisis de los residuos.
4. El estudio de los valores influyentes a priori y a posteriori.
5. Se va a añadir un último elemento, el denominado C_p de Mallows.

La combinación de los 5 puntos anteriores permitirá seleccionar dentro de un conjunto de ecuaciones de regresión (quizás incluso en el caso hipotético de todas) la *mejor*.

TEMA 3: SELECCIÓN DEL MEJOR MODELO. C_P MALLOWS

El C_p de Mallows se define como $C_p = \frac{SCR_p}{s^2} - (n-2p) = \frac{AIC_p}{s^2} - n$, donde SCR_p es la suma de cuadrados residual de un modelo con p regresores y el estimador de la varianza del modelo procede del modelo maximal (se intenta garantizar así la ausencia de sesgo).

- La esperanza matemática del estadístico C_p es el número de parámetros del modelo: $\mathrm{E}[C_p] = p$.
- El procedimiento a seguir consiste en representar en un diagrama bivariante C_p frente *p*: los modelos satisfactorios quedarán cerca de la bisectriz, el modelo con *p* más bajo, pero sobre la bisectriz resulta el más satisfactorio bajo el criterio de Mallows. La justificación del procedimiento procede de las siguientes consideraciones:

1. Un modelo no adecuado facilitará una SCR_p elevada, con $C_p > p$. De alguna manera, el estadístico de Mallows se desvía de la bisectriz indicando que existe sesgo en la estimación de la varianza del modelo: un error cuadrático medio (varianza + sesgo^2) distinto de la varianza real del modelo.
2. Un modelo con exceso de regresores ajusta bien los datos y $C_p \approx p$, pero *p* es mayor que en otro modelo satisfactorio con valor del estadístico de Mallows sobre la bisectriz de la gráfica y menor número de parámetros.

TEMA 3: SELECCIÓN DEL MEJOR MODELO

Exemplo: Datos DUNCAN1 resultados del análisis mediante el Cp de Mallows de todos los modelos posibles para explicar el 'PRESTIGE' a partir de los ingresos (INCOME) y la EDUCATION.

```
MTB > BReg 'PRESTIGE' 'INCOME' 'EDUCATION' ;
SUBC>   NVars 1 2;
SUBC>   Best 2;
SUBC>   Constant.
```

Best Subsets Regression: PRESTIGE versus INCOME; EDUCATION

```
Response is PRESTIGE

                                                    E
                                                    D
                                                 I  U
                                                 N  C
                                                 C  A
                                                 O  T
                                                 M  I
Vars   R-Sq    R-Sq(adj)        C-p          S   E  O

   1   72,6         71,9       26,0     16,692      X
   1   70,2         69,5       31,9     17,403   X
   2   82,8         82,0        3,0     13,369   X  X


MTB >
```

TEMA 3: SELECCIÓN DEL MEJOR MODELO

3-1-10.1 Procedimiento de "backward elimination"

Procedimiento económico que no requiere del cálculo de un número elevado de ecuaciones de regresión. Los pasos básicos son:

1. Calcular la ecuación de regresión maximal, es decir, que contenga todos los regresores disponibles.
2. Para cada regresor se efectúa un test de Fisher de la hipótesis $H_i : \beta_i = 0$, sea el valor del estadístico de Fisher correspondiente al test de la hipótesis nula del *i*-ésimo regresor F_i.
3. Se selecciona el regresor tal que el estadístico de Fisher correspondiente es mínimo, sea el regresor *l*-ésimo: $F_l = min\{F_1, F_2, \ldots\}$ y se compara con el valor del correspondiente a un cierto nivel de significación de la ley de Fisher correspondiente denominado en muchos paquetes estadísticos "*F to remove*": $\underline{F}^{\alpha}$.

   Si $F_l < \underline{F}^{\alpha}$ entonces se elimina el regresor *l*-ésimo del modelo. Se repite a partir del punto 2.

   Sinó el modelo ya es satisfactorio.

TEMA 3: SELECCIÓN DEL MEJOR MODELO. BACKWARD & FORWARD PROCEDURES

- Backward Elimination es adecuado para la regresión polinómica y robusto, una vez eliminada una variable nunca vuelve a aparecer en la ecuación, ni tampoco ningún modelo alternativo que la contenga, puede dar como resultado modelos que no son significativamente los mejores.
- La construcción de la regresión maximal podía ser un inconveniente hace unos años, por el mal condicionamiento posible de la matriz de diseño maximal, actualmente los procedimientos de optimización empleados son muy robustos.
- De manera análoga, algunos paquetes estadísticos disponen del procedimiento *forward inclusion*, que parte del modelo minimal (únicamente con el término independiente) y va añadiendo regresores siempre que el test de inclusión basado en el estadístico de Fisher resulte significativo para alguno de los regresores no incluidos hasta el momento.

- *Forward inclusion* es un procedimiento menos robusto que *la backward elimination*, ya que a veces un regresor incluído en una etapa anterior, podría ser eliminado por falta de significación. Este inconveniente lleva directamente al diseño de un procedimiento híbrido denominada regresión paso o paso o *stepwise regression*, que aparece en todos los paquetes (al menos, los conocidos por la autora) y cuyo empleo se incentiva en las clases de laboratorio.

TEMA 3: SELECCIÓN DEL MEJOR MODELO. *STEPWISE REGRESSION*

3-1-10.2 Regresión paso o paso (stepwise regression)

Procedimiento de selección de la mejor ecuación de regresión (mejor modelo), parte de un conjunto reducido de regresores y lo va engrandeciendo hasta hallar el modelo satisfactorio. Las etapas pueden resumirse en los siguientes puntos:

1. Seleccionar el regresor más correlacionado con la variable de respuesta **Y**, sea $\mathbf{x}_m$. Calcular la ecuación de regresión.
2. Para cada regresor *i* no incluido hasta el momento se calcula el coeficiente de correlación parcial con la variable de respuesta **Y** (técnicamente supone calcular la correlación entre los residuos del modelo actual y los residuos de una ecuación de regresión auxiliar que especifica como variable de repuesta el regresor *i* no presente en el modelo y variables explicativas todos los regresores presentes en el modelo actual).
3. Se selecciona el regresor con coeficiente de correlación parcial más elevado, sea $\mathbf{x}_m$ y se recalcula la ecuación de regresión con el modelo incrementado, aceptando el nuevo regresor $\mathbf{x}_m$ si el estadístico de Fisher para el contraste de la hipótesis nula $H_m : \beta_m = 0$, sea F_m es superior a un cierto valor de referencia mínima de la ley de Fisher correspondiente denominado "**F to enter**": $\overline{F}^{\beta}$.

   Si $F_m > \overline{F}^{\beta}$ entonces se incluye el regresor *m*-ésimo en el modelo.

TEMA 3: SELECCIÓN DEL MEJOR MODELO. *STEPWISE REGRESSION (Cont)*

4. Para cada regresor incluido hasta el momento se efectua un test de Fisher de la hipótesis $H_i : \beta_i = 0$, sea el valor del estadístico de Fisher correspondiente al test de la hipótesis nula del *i*-ésimo regresor F_i.

5. Se selecciona el regresor tal que el estadístico de Fisher correspondiente es mínimo, sea el regresor *l*-ésimo: $F_l = min\{F_1, F_2, \ldots\}$ y se compara con el valor del correspondiente a un cierto nivel de significación de la ley de Fisher correspondiente denominado en muchos paquetes estadísticos "*F to remove*": $\underline{F}^{\alpha}$.

   Si $F_l < \underline{F}^{\alpha}$ entonces se elimina el regresor *l*-ésimo del modelo. **Volver** al punto 2.

6. El procedimiento finaliza cuando ningún regresor satisfaga el criterio de entrada y ningún regresor satisfaga el criterio de salida. Si los niveles de significación para la entrada y la salida de regresores estan bien seleccionados da buenos resultados: $\alpha = \beta = 0.05$ es una selección habitual.