Regresión

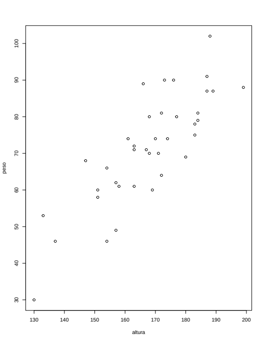
Objetivo: Predecir el peso de una persona en función de su altura.

Datos: Muestra del peso y la estatura de 38 personas adultas

altura	sexo	peso
172	h	81
133	m	53
163	m	71
...	...	...
137	m	46
166	h	89
163	h	72

¿ La altura y el peso guardan relación? ¿De qué tipo?

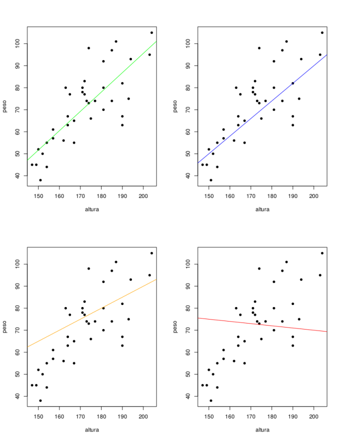
Regresión



La altura y el peso parece que tiene una relación lineal

$$Peso = \beta_0 + \beta 1 * Altura + error$$

Regresión



¿Entre todos las posibles rectas, cuál elegir y porqué?

$$Peso = \beta_0 + \beta_1 * altura$$

Regresión

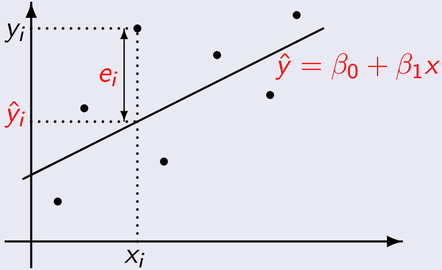
Resolución del problema:

1. Escribir en términos matemáticos el modelo de relación
2. Plantear un criterio para buscar los parámetros del modelo
3. Calcular las estimaciones óptimas de los parámetros
4. Medir la 'calidad' del modelo

Regresión

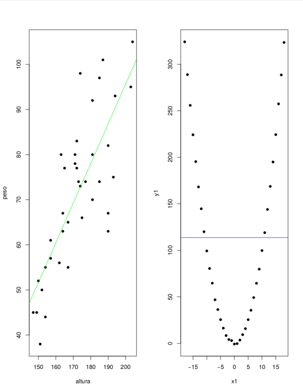
$$y_i = \beta_0 + \beta_1 \; x_i + \epsilon_i; \qquad \hat{y}_i = \beta_0 + \beta_1 \; x_i; \quad e_i = y_i - \hat{y}_i$$

$$Min_{(\beta_0,\beta_1)} \sum_{i=1}^{n} e_i^2 = Min_{(\beta_0,\beta_1)} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \; x_i)^2$$



Regresión

¿El mejor modelo lineal se ajusta bien a los datos?



Estimación

Sea x una variable aleatoria con algún parámetro θ desconocido.

Objetivo: aproximar θ a partir de una muestra $(x_1, ..., x_n)$

Estimador: transforma la muestra en un valor del parámetro

$$T : (X_1, ..., X_n) \longrightarrow \Theta$$

Se desconoce $P(cara)$ en una moneda. Se lanza cien veces y se obtienen 52 caras.

$$\hat{P}(cara) = \frac{52}{100} = 0{,}52$$

Saber la estatura media de los habitantes de una ciudad. Se elige al azar una muestra de 1000 habitantes y se mide su estatura

$$\hat{\mu} = \frac{1}{1000} \sum_{i=1}^{1000} estatura_i = 1,78$$

Estimación

Métodos de estimación:

- Máxima verosimilitud: busca el valor del parámetro que hace máxima la verosimilitud de la muestra.

$$Max_{\theta\in\Theta} L(x_1, ..., x_n; \theta)$$

- Mínimos cuadrados: busca el valor de θ que minimiza el cuadrado de la distancia entre las muestras y valores teóricos

$$Min_{\theta\in\Theta} \sum_{i=1}^{n} (x_i - teor_i(\theta))^2$$

Propiedades deseables de un estimador:

- Ser centrado y tener varianza pequeña

$$E(T) = \theta; \quad Var(T) \simeq 0$$

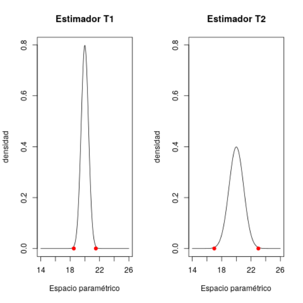
- Tener un error cuadrático medio pequeño: $E(T - \theta)^2 \simeq 0$.

Estimación

Propiedades deseables de un estimador:

- Ser centrado y tener varianza pequeña

$$E(T) = \theta; \quad Var(T) \simeq 0$$



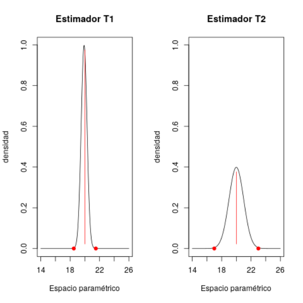
¿En general, cuál de estos dos estimadores será más preciso?

Estimación

- Tener un error cuadrático medio pequeño.

$$ECM = E(T - \theta)^2 \simeq 0$$

En este ejemplo $E(T1) \neq \theta$ y $E(T2) = \theta$



¿En general, cuál de estos dos estimadores será más preciso?

Regresión

Objetivo: buscar un modelo funcional para explicar el comportamiento de una variable Y, *variable dependiente*, a través de otras variables $X_1, \ldots, X_p$ llamadas *variables independientes*

$$Y = f(X_1, \ldots, X_p) + \epsilon.$$

Por simplicidad funciones $f(X_1, ..., X_p)$ lineales

Regresión Lineal

Modelo:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon = \vec{x}^t \vec{\beta} + \epsilon$$

- $\vec{\beta}^t = (\beta_0, \ldots, \beta_p)$, parámetros del modelo, fijos y desconocidos.
- $x_1, \ldots, x_p$, variables independientes o predictoras.
- ϵ una variable aleatoria no observable.

Regresión Lineal: Planteamiento

Modelo para una muestra con n observaciones de p variables

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x_1} + \cdots + \beta_p \vec{x_p} + \vec{\epsilon}$$

$$\vec{y} = X\vec{\beta} + \vec{\epsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} * \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{pmatrix}$$

Regresión Lineal: Planteamiento

- $\vec{y}$ el vector de n de observaciones de la variable independiente.
- X la matriz $n \times (p+1)$, de valores de las variables explicativas; se denomina matriz de regresión o de diseño, en general es de rango (p+1).
- $\vec{\beta}$ el vector de parámetros de dimensión p+1.
- $\vec{\epsilon}$ el vector de las n perturbaciones aleatorias.

Regresión Lineal: Estimaciones

Criterio de mínimos cuadrados ordinarios

$$g(\vec{\beta}) = \|\vec{y} - X\vec{\beta}\|^2 = (\vec{y} - X\vec{\beta})^t(\vec{y} - X\vec{\beta}) = \sum_{i=1}^{n}(y_i - x_i^t\vec{\beta})^2$$

$$= \vec{y}^y\vec{y} - 2\vec{y}^tX\vec{\beta} + \vec{\beta}^tX^tX\beta$$

El criterio de mínimos cuadrados busca $\widehat{\beta} = arg(\min g(\vec{\beta}))$

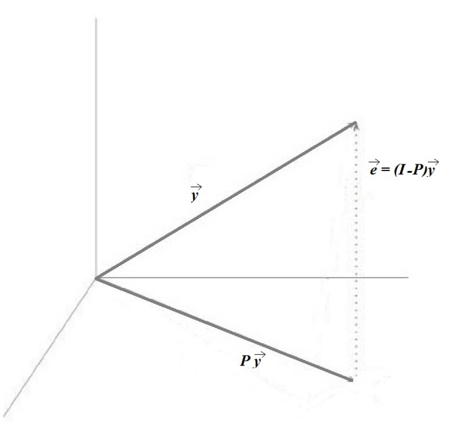
Puntos críticos $\frac{\partial g(\vec{\beta})}{\partial \vec{\beta}} = -2X^t\vec{y} + 2X^tX\widehat{\beta} = \vec{0}$ de donde

$$\widehat{\beta} = (X^tX)^{-1}X^t\vec{y} \quad \text{si existe } (X^tX)^{-1}$$

Regresión Lineal: Estimaciones

- Estimación $\widehat{\beta} = (X^t X)^{-1} X^t \vec{y}$.
- Predicciones $\widehat{y} = X\widehat{\beta} = X(X^t X)^{-1} X^t \vec{y} = P_X \vec{y}$
  P_X la matriz de proyección ortogonal sobre el subespacio generado por las columnas de la matriz X
- Residuos $\vec{e} = \vec{y} - \widehat{y} = \vec{y} - P_X \vec{y} = (I - P_X)\vec{y} = Q_X \vec{y}$
  Q_X la matriz de proyección sobre el subespacio ortogonal al generado por las columnas de la matriz X.

Regresión Lineal: Interpretación geométrica



Proyecciones ortogonales

- Si Ω es un subespacio vectorial de $\mathbb{R}^n$ entonces todo vector $y \in \mathbb{R}^n$ se puede descomponer de forma única como $y = u + v, \; u \in \Omega, v \in \Omega^{\perp}$
- Si Ω es un subespacio vectorial de $\mathbb{R}^n$ existe una única matriz P_Ω que verifique la condición: $P_\Omega y = u$. Esta matriz viene dada por $P_\Omega = X(X^t X)^{-1} X^t$, donde X es cualquier matriz cuyas columnas sean una base del subespacio.
- Si P_Ω es la matriz de proyección sobre Ω entonces $Q_\Omega = I - P_\Omega$ es la matriz de proyección sobre $\Omega^{\perp}$
- P_Ω y Q_Ω son matrices simétricas e idempotentes

Partición de la variabilidad

En el modelo de regresión se verifica $\vec{y} = \overrightarrow{e} + \widehat{y}$:

1. La media muestral de las predicciones coincide con la media muestral de los datos de la variable independiente $\overline{Y} = \overline{\widehat{Y}}$.
2. La media muestral de los residuales es cero.
3. **SCT=SCReg+SCE**
   $SCT = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \overrightarrow{y}^t H \overrightarrow{y}; \quad H = I - \overrightarrow{1}\,\overrightarrow{1}^t/n$
   $SCReg = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = \widehat{y}^t H \widehat{y}$
   $SCE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \overrightarrow{y}^t (I - P_X) \overrightarrow{y} = \overrightarrow{e}' \overrightarrow{e}$

Partición de la variabilidad: Demostración

1. $\overline{\widehat{Y}} = \overrightarrow{1}^{\,t} P_x \vec{y}/n = \overrightarrow{1}^{\,t} \overrightarrow{y}/n = \overline{Y}$
   ya que $\vec{1}' P_x = (P_x \vec{1})' = \vec{1}'$ por ser $\vec{1} \in M(X)$
2. $\overline{e} = \overrightarrow{1}^{\,t} \vec{e}/n = \overrightarrow{1}^{\,t}(I - P_x)\vec{y}/n = 0$
   puesto que $\vec{1}'(I - P_x) = \vec{1}' - \vec{1}' P_x = \vec{1}' - \vec{1}' = \vec{0}$
3. **SCT=SCReg+SCE**

$$SCT = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2$$

$$= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + 2\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y})$$

$$= \overrightarrow{e}^{\,t}\overrightarrow{e} + \widehat{y}^t H \widehat{y} + 2\overrightarrow{e}^{\,t} H \widehat{y} = \overrightarrow{e}^{\,t}\overrightarrow{e} + \widehat{y}^t H \widehat{y}$$

porque $\overrightarrow{e}^{\,t} H \widehat{y} = \overrightarrow{e}^{\,t} \widehat{y} = \overrightarrow{e}^{\,t} P_x \vec{y} = \vec{y}'(I - P_x) P_x \vec{y} = 0$

Coeficiente de determinación

Se utiliza para medir la bondad del ajuste

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

Propiedades

1. $0 \leq R^2 \leq 1$
2. $R^2 = 1 \quad \Leftrightarrow \quad e_i = 0 \quad \forall i = 1, \ldots, n$ el ajuste lineal de los datos es perfecto
3. $R^2 = 0 \quad \Leftrightarrow \quad \widehat{y}_i = \overline{y} \quad \forall i = 1, \ldots, n$. No hay relación lineal pero puede existir otro tipo de relación.

Coeficiente de determinación corregido

$$R_c^2 = 1 - \frac{SCE/(n-p-1)}{SCT7(n-1)} = 1 - \frac{S_e^2}{S_y^2}$$

Inferencia

Condiciones previas

- $E[\vec{\epsilon}] = \vec{0} \Rightarrow E[\vec{y}] = X\vec{\beta}$
- $Var[\vec{\epsilon}] = E[\vec{\epsilon}\vec{\epsilon}^t] = \sigma^2 I$ (Residuos linealmente independientes y homocedásticos). Esto implica que:
  $Var(\vec{y}) = Var(X \vec{beta} + \vec{\epsilon}) = Var(\vec{\epsilon}) = \sigma^2 I$
- $rango(X) = p + 1 \leq n$ para poder invertir la matriz $X^t X$.

Bajo las condiciones anteriores $\widehat{\beta}$ verifica:

1. $E[\widehat{\beta}] = \vec{\beta}$ (El estimador es centrado).
2. $Var(\widehat{\beta}) = \Sigma_{\widehat{\beta}} = \sigma^2 (X^t X)^{-1}$.
3. $s^2 = \frac{1}{n-(p+1)} SCE = \frac{1}{n-(p+1)} \sum_{i=1}^n e_i^2$ verifica que $E(s^2) = \sigma^2$.

Inferencia: Propiedades del estimador $\widehat{\beta}$

1. $E[\widehat{\beta}] = E((X^tX)^{-1}X^t\vec{y}) = (X^tX)^{-1}X^tE(\vec{y})$
   $= (X^tX)^{-1}X^tX\beta = \beta.$
2. $Var(\widehat{\beta}) = \Sigma_{\widehat{\beta}} = Var((X^tX)^{-1}X^t\vec{y}) =$
   $(X^tX)^{-1}X^t var(\vec{y})X(X^tX)^{-1} = \sigma^2(X^tX)^{-1}.$
3. $SCE = \sum_{i=1}^n e_i^2 = \vec{e}^t\vec{e} = \vec{y}^t(I - P_X)\vec{y} = \vec{\epsilon}^t Q_X\vec{\epsilon}$
   Por las propiedades de la traza una matriz y
   $Var(\vec{y}) = E(\vec{y}\,\vec{y}') - \vec{\mu_y}\vec{\mu_y}' = E(\vec{y}\,\vec{y}') - X\vec{\beta}\vec{\beta}^tX^t$
   $E(SCE) = E[\vec{y}^t(I - P_X)\vec{y}] = E[traza(\vec{y}^t(I - P_X)\vec{y})]$
   $= traza(E[(I - P_X)\vec{y}\vec{y}^t]) = traza[(I - P_X)(\sigma^2 I + X\vec{\beta}\vec{\beta}^tX^t)]$
   $= traza[(I - P_X)\sigma^2] = \sigma^2(n - (p + 1))$

$$\Rightarrow E(s^2) = \frac{E(SCE)}{n - (p + 1)} = \sigma^2$$

Inferencia: Propiedades del estimador $\widehat{\beta}$

Teorema de Gauss- Markov

El estimador mínimo cuadrático ordinario de $\vec{\beta}$, $\widehat{\beta}$, es el mejor entre los estimadores $\widehat{\beta}^*$ de $\vec{\beta}$ que son insesgados y funciones lineales de $\vec{y}$, $\widehat{\beta}^* = A\vec{y}$ es decir verifica que $Var(\widehat{\beta}_j^*) \geq Var(\widehat{\beta}_j)$, $j = 0, 1, ..., p$. $\widehat{\beta}$ es el mejor el estimador lineal insesgado (BLUE).

Inferencia: Teorema de Gauss- Markov. Demostracion

Por ser $\widehat{\beta}^* = A\vec{y}$ insesgado verifica:

$$E[\widehat{\beta}^*] = AE[\vec{y}] = AE[X\vec{\beta} + \vec{\epsilon}] = AX\vec{\beta} = \vec{\beta} \quad \forall\vec{\beta} \Rightarrow \ AX = I$$

$$Cov(\widehat{\beta}^*) = E[(\widehat{\beta}^* - \vec{\beta})(\widehat{\beta}^* - \vec{\beta})^t] = E[(\widehat{\beta}^* - \widehat{\beta} + \widehat{\beta} - \vec{\beta})(\widehat{\beta}^* - \widehat{\beta} + \widehat{\beta} - \vec{\beta})^t]$$

$$= Cov(\widehat{\beta}^* - \widehat{\beta}) + Cov(\widehat{\beta}) + E[(\widehat{\beta}^* - \widehat{\beta})\,(\widehat{\beta} - \vec{\beta})^t] + E[(\widehat{\beta} - \vec{\beta})\,(\widehat{\beta}^* - \widehat{\beta})^t]$$

$$= Cov(\widehat{\beta}^* - \widehat{\beta}) + Cov(\widehat{\beta})$$

ya que:

$$\widehat{\beta}^* - \widehat{\beta} = (A - (X'X)^{-1}X'\vec{\epsilon}\,; \quad \widehat{\beta} - \vec{\beta} = (X'X)^{-1}X')\vec{\epsilon}$$

$$E[(\widehat{\beta}^* - \widehat{\beta})(\widehat{\beta}^* - \vec{\beta})^t] = \sigma^2[(A - (X^tX)^{-1}X^t)\,X(X^tX)^{-1}] = \mathbf{0}$$

En consecuencia: $var(\widehat{\beta}_j^*) = var(\widehat{\beta}_j^* - \widehat{\beta}_j) + var(\widehat{\beta}_j) \geq var(\widehat{\beta}_j)$

Inferencia en el caso normal

Dado $\vec{y} = X\vec{\beta} + \vec{\epsilon}$ con $X_{n,p+1}$ una matriz fija de rango $p+1$ y $\vec{\epsilon} \equiv N(\vec{0}, \sigma^2 I)$ se cumplen las siguientes propiedades

1. $\widehat{\beta} \equiv N_{p+1}(\vec{\beta}, \sigma^2 (X^t X)^{-1})$ (El estimador es centrado y normal).
2. $\dfrac{(\widehat{\beta} - \beta)^t (X^t X)(\widehat{\beta} - \beta)}{\sigma^2} \equiv \chi^2_{p+1}$.
3. $\frac{SCE}{\sigma^2} \equiv \chi^2_{n-(p+1)}$.
4. $\widehat{\beta}$ es independiente de SCE y por tanto de $\widehat{\sigma}^2$.

Propiedades caso normal: Demostración

Se considera la distribución conjunta de $\vec{e}$ y $\widehat{\beta}$, asociada a una transformación lineal de $\vec{y}$

$$\begin{pmatrix} \widehat{\beta} \\ \vec{e} \end{pmatrix} = \begin{pmatrix} (X^tX)^{-1}X^t \\ I - P_X \end{pmatrix} \vec{y} = \begin{pmatrix} \vec{\beta} \\ \vec{0} \end{pmatrix} + \begin{pmatrix} (X^tX)^{-1}X^t \\ I - P_X \end{pmatrix} \vec{\epsilon}$$

El vector $(\vec{e}, \widehat{\beta})$ es Normal y cada componente también.

- Distribución de $\widehat{\beta}$

$$\widehat{\beta} \equiv N_{p+1}(\vec{\beta}, \sigma^2(X^tX)^{-1})$$

ya que: $E(\widehat{\beta}) = \vec{\beta}; \quad Cov(\widehat{\beta}) = \sigma^2\,(X'X)^{-1}$

Propiedades caso normal: Demostración

- $$\frac{(\widehat{\beta}-\beta)^t(X^tX)(\widehat{\beta}-\beta)}{\sigma^2} \equiv \chi^2_{p+1}$$

  Por ser $(X'X)$ simétrica e inversible $\Rightarrow (X'X) = U\Delta U'$, donde U es ortogonal y Δ diagonal con valores mayores que cero.

  $$(\widehat{\beta}-\beta)^t(X^tX)(\widehat{\beta}-\beta) = (\widehat{\beta}-\beta)^t U\Delta U'(\widehat{\beta}-\beta)$$
  $$(\widehat{\beta}-\beta)^t U\Delta^{1/2}U'\; U\Delta^{1/2}U'(\widehat{\beta}-\beta)$$

  $$Cov[U\Delta^{1/2}U'(\widehat{\beta}-\beta)] = U\Delta^{1/2}U'\; Cov(\widehat{\beta}-\beta)\; U\Delta^{1/2}U'$$
  $$= U\Delta^{1/2}U'\,\sigma^2(X'X)^{-1}U\Delta^{1/2}U' =$$
  $$= \sigma^2\; U\Delta^{1/2}U'\; U\Delta^{-1}U'\; U\Delta^{1/2}U' = \sigma^2 I$$

  Por tanto $U\Delta^{1/2}U'(\widehat{\beta}-\beta)/\sigma \equiv N(\vec{0}, I)$

  $$\frac{(\widehat{\beta}-\beta)^t(X^tX)(\widehat{\beta}-\beta)}{\sigma^2} \equiv \chi^2_{p+1}$$

Propiedades caso normal: Demostración

- Distribución de los residuales estimados

$$\vec{e} \equiv N_n(\vec{0}, \sigma^2 Q_x) \Rightarrow \frac{\vec{e}^t \, Q_x \, \vec{e}}{\sigma^2} = \frac{\vec{e}^t \, \vec{e}}{\sigma^2} \equiv \chi^2_{n-(p+1)}$$

- Independencia entre β y $\vec{e}$. Por tener una distribución normal conjunta basta demostrar que $Cov(\vec{\beta}, \vec{e}) = \mathbf{0}$

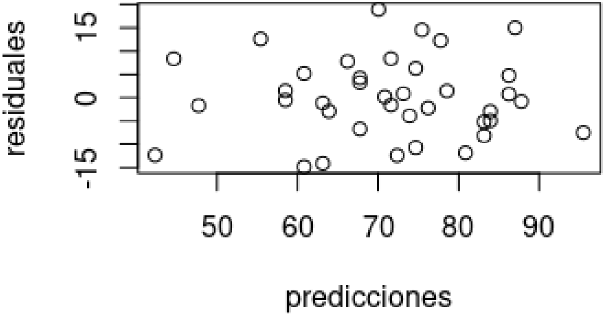
$$Cov(\vec{\beta}, \vec{e}) = cov((X^t X)^{-1} X^t \vec{y}, \; Q_X \vec{y})$$

$$= (X^t X)^{-1} X^t var(y) Q_X = \mathbf{0}$$

PROPIEDAD.
Bajo las condiciones usuales de regresión los residuales $\vec{e}$ y las predicciones $\widehat{y}$ son independientes.

Modelo de regresión: *Peso* $\sim$ *altura*.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.8735    14.9584  -3.869 0.000441
altura        0.7707     0.0889   8.669 2.44e-10
```



Estimación bajo restricciones

Estimar $\vec{\beta}$ cuando verifica $A\overrightarrow{\beta} = \overrightarrow{c}$; $rango(A) = q$

$$min_{\vec{\beta}}(\vec{y}^y\vec{y} - 2\vec{y}^t X\vec{\beta} + \vec{\beta}^t X^t X\vec{\beta})$$

$$g(\vec{\beta}, \vec{\lambda}) = \vec{y}^y\vec{y} - 2\vec{y}^t X\vec{\beta} + \vec{\beta}^t X^t X\vec{\beta} - \vec{\lambda}^t(A\vec{\beta} - \vec{c})$$

Derivando con respecto a β y λ se obtiene:

- $\partial g/\partial\vec{\beta} = \vec{0} = -2X^t\vec{y} + 2(X^t X)\widehat{\beta}_h - A'\,\vec{\lambda}_h$
  $\widehat{\beta}_h = \widehat{\beta} + (X^t X)^{-1}(1/2)A^t\vec{\lambda}_h$
- $\partial g/\partial\vec{\lambda} = A\widehat{\beta}_h - \vec{c} = \vec{0}$
  $\hat{\lambda}_h = -2[A(X'X)^{-1}A']^{-1}(A\widehat{\beta} - \vec{c})$

$$\widehat{\beta}_h = \widehat{\beta} - (X^t X)^{-1}A^t[A(X^t X)^{-1}A^t]^{-1}(A\widehat{\beta} - \vec{c})$$

Estimación bajo restricciones

1. El estimador de mínimos cuadrados restingido de $\overrightarrow{\beta}$ es:

$$\widehat{\beta}_h = \widehat{\beta} - (X^tX)^{-1}A^t[A(X^tX)^{-1}A^t]^{-1}(A\widehat{\beta} - \vec{c})$$

2. $SCE_h = (\vec{y} - X\widehat{\beta}_h)^t(\vec{y} - X\widehat{\beta}_h)$
3. $SCE_h - SCE = (A\widehat{\beta} - \vec{c})^t(A(X^tX)^{-1}A^t)^{-1}(A\widehat{\beta} - \vec{c})$

Ejemplo. Los coeficientes $\beta_1 = ... = \beta_q = 0$.

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & & \dots\dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{pmatrix} * \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_q \\ \dots \\ \beta_p \end{pmatrix} = \vec{0}$$

Contraste $H_0 : A\vec{\beta} = \vec{c}$

Dado $\vec{y} = X\vec{\beta} + \vec{\epsilon}$ con $X_{n,p+1}$ una matriz fija de rango $p+1$ y $\vec{\epsilon} \equiv N(\vec{0}, \sigma^2 I)$ y una matriz $A_{q,p+1}$ de rango q. Si $\widehat{\beta}_h$ es el EMC de $\vec{\beta}$ bajo la restricción $A\vec{\beta} = \vec{c}$ se verifica que:

1. $E(SCE_h - SCE) = q\sigma^2 + (A\vec{\beta} - \vec{c})^t (A(X^t X)^{-1} A^t)^{-1} (A\vec{\beta} - \vec{c})$
2. Hipótesis nula $H_0 : A\vec{\beta} = \vec{c}$

$$F = \frac{\frac{SCE_h - SCE}{q}}{\frac{SCE}{n-p-1}} \equiv_{H_0} F_{q,n-p-1}$$

Contraste $H_0: A\vec{\beta} = \vec{c}$. Demostración

Bajo $H_0: \widehat{\beta}_h = \widehat{\beta} - (X^tX)^{-1}A^t[A(X^tX)^{-1}A^t]^{-1}(A\widehat{\beta} - \vec{c})$

$$E(SCE_h - SCE) = E[(A\widehat{\beta} - \vec{c})^t(A(X^tX)^{-1}A^t)^{-1}(A\widehat{\beta} - \vec{c})]$$

$$= q\sigma^2 + (A\vec{\beta} - \vec{c})^t(A(X^tX)^{-1}A^t)^{-1}(A\vec{\beta} - \vec{c})$$

$\vec{e}$ y $X(X^tX)^{-1}A^t[A(X^tX)^{-1}A^t]^{-1}(A\widehat{\beta} - \vec{c})$ son independientes
$\Rightarrow SCE_h - SCE$ y SCE son independientes,
bajo H_0 son estimadores insesgados de $q\,\sigma^2$ y $(n-p-1)\,\sigma^2$, con distribuciones χ^2 de donde

$$F = \frac{\frac{SCE_h - SCE}{q}}{\frac{SCE}{n-p-1}} \equiv_{H_0} F_{q,n-p-1}$$

Comparación entre dos modelos

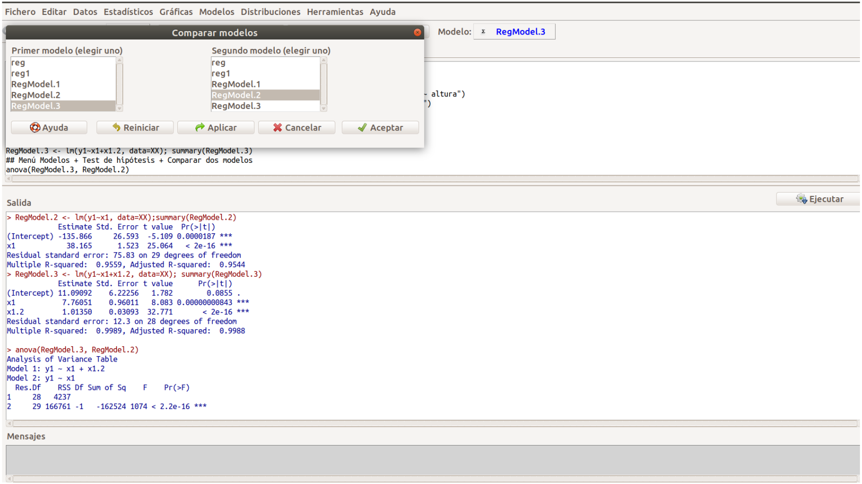
Modelo I. $y = \beta_0 + \beta_{q+1}x_{q+1} + \beta_p x_p + \epsilon$

Modelo II. $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_q + x_q + \beta_{q+1}x_{q+1} + \beta_p x_p + \epsilon$

¿Cuál es más adecuado?
Una solución razonable es comprobar si el modelo más sencillo tiene la misma capacidad de predicción que el otro modelo.

$$F = \frac{\frac{SCE_h - SCE}{q}}{\frac{SCE}{n-p-1}} \equiv_{H_0} F_{q,n-p-1}$$

Comparación entre dos modelos



Contraste de la regresión

H_0 : La variable y es linealmente independiente de todas las x

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad \Leftrightarrow \quad H_0 : A\vec{\beta} = \vec{0},\ A = (\vec{0}, I_p)$$

$SCT = SCE + SCReg$
Bajo H_0 : $\quad SCE_h = SCT$; q= rango(A)= p

$$F = \frac{\frac{SCE_h - SCE}{q}}{\frac{SCE}{n-p-1}} = \frac{\frac{SCReg}{p}}{\frac{SCE}{n-p-1}} = \frac{n-p-1}{p} \frac{R^2}{1-R^2} \equiv_{H_0} F_{p,n-p-1}$$

Regresión Lineal: Inferencia

Inferencias sobre combinaciones lineales $\vec{h}^t\vec{\beta}$

Función pivote

$$\frac{\vec{h}^t\widehat{\beta} - \vec{h}^t\vec{\beta}}{\widehat{\sigma}\sqrt{\vec{h}^t(X^tX)^{-1}\vec{h}}} \equiv t_{n-p-1}$$

Caso particular: función pivote para cada componente $\widehat{\beta}_j$:

$$\frac{(\widehat{\beta}_j - \vec{\beta}_j)\sqrt{(X^tX)_{jj}}}{(SCE/(n-p-1))} \equiv t_{1,n-p-1}$$

Regresión Lineal: Estimacion

Intervalos de confianza para $E(Y/\vec{x}_0) = \vec{x}_0^t \vec{\beta}$

Estimador insesgado de $E(Y/\vec{x}_0)$: $\quad \vec{x}_0^t \widehat{\beta}$

$$\vec{x}_0^t \widehat{\beta} \equiv N_1(\vec{x}_0^t \vec{\beta}, \sigma^2 \vec{x}_0^t (X^t X)^{-1} \vec{x}_0)$$

Función pivote

$$\frac{\vec{x}_0^t \widehat{\beta} - \vec{x}_0^t \beta}{\widehat{\sigma} \sqrt{\vec{x}_0^t (X^t X)^{-1} \vec{x}_0}} \equiv t_{n-p-1}$$

Intervalo de confianza con coeficiente $(1-\alpha)$

$$\left(\vec{x}_0^t \widehat{\beta} \pm F^{-1}_{t_{n-p-1}}(1-\alpha/2) \widehat{\sigma} \sqrt{\vec{x}_0^t (X^t X)^{-1} \vec{x}_0} \right)$$

Regresión Lineal: Estimación

Intervalo de confianza para la predicción de $Y/\vec{x}_0$

Individuo genérico con valores $\vec{x}_0$, que no pertenece a la muestra:

$$Y/\vec{x}_0 - \vec{x}_0^t\widehat{\beta} \equiv N_1\left(0, \sigma^2(\vec{x}_0^t(X^tX)^{-1}\vec{x}_0 + 1)\right)$$

Intervalo de confianza con coeficiente $(1-\alpha)$

$$\left(\vec{x}_0^t\widehat{\beta} \pm F^{-1}_{t_{n-p-1}}(1-\alpha/2)\widehat{\sigma}\sqrt{\vec{x}_0^t(X^tX)^{-1}\vec{x}_0 + 1}\right)$$

Análisis de residuales

Sirven para estudiar si un modelo de regresión está bien construido.

Bajo la hipótesis $\vec{\epsilon} = N_n(\vec{0}, \sigma^2 I_n)$ se tiene

$$\vec{e} = (I - P_X)\vec{y} \equiv N_n[\vec{0}, \sigma^2(I_n - P_X)]$$

$$var(e_i) = \sigma^2(1 - h_{ii}) \quad \text{con } h_{ii} = \vec{x}_i^t (X^t X)^{-1} \vec{x}_i = diag(P_X)$$

$$h_{ii} = \frac{1}{n}[1 + D^2(\vec{x}_i, \overline{x})]$$

Los residuos estimados no son independientes, ni homocedásticos.

Análisis de residuales

Teorema Sherman-Morrison-Woodbury:

$$(A + \vec{u}\,\vec{v}^t)^{-1} = A^{-1} - \frac{A^{-1}\,\vec{u}\vec{v}^t\,A^{-1}}{1 + \vec{v}^t\,A^{-1}\,\vec{u}}$$

Aplicación:

$$X'X = \sum i = 1^n \vec{x}_i\,\vec{x}'_i = \vec{x}_i\,\vec{x'}_i + \sum_{j \neq i}^{n} \vec{x}_i\,\vec{x'}_i = \vec{x}_i\,\vec{x'}_i + X'_{-i}X_{-i}$$

donde X_{-i} representa la matriz de datos sin la observación i.

$$X'_{-i}X_{-i} = X'X - \vec{x}_i\,\vec{x}'_i$$

Aplicando el teorme de Sherman-Morrison-Woodbury para invertir la matriz $X'_{-i}X_{-i}$ se obtienen los estimadores de la regresión cuando se elimina de la muestra el caso i.

Análisis de residuales

Estimación de $\vec{\beta}$ al eliminar la observación i-ésima de la muestr:

$$\widehat{\beta}_{(-i)} = \widehat{\beta} - (X'X)^{-1}\vec{x}_i \frac{e_i}{1-h_{ii}}$$

Residuo i-ésimo imparcial: la observación no se usa para estimar

$$\widehat{y}_{i,-i} = \widehat{y}_i - \frac{h_{ii}}{1-h_{ii}} e_i$$

$$e_{i,-i} = y_i - \widehat{y}_{i,-i} = \frac{e_i}{1-h_{ii}}$$

$$\widehat{\sigma}^2_{-i} = \sum_{j\neq i}(y_j - \vec{x}_j'\,\widehat{\beta}_{(-i)})^2 = \frac{(n-p)\widehat{\sigma}^2 - (e_i^2/(1-h_{ii})}{n-(p+1)}$$

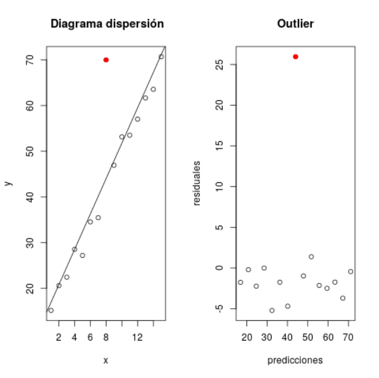
Análisis de residuales

Tipos de residuales:

1. Originales: $e_i = y_i - \widehat{y}_i$
2. Tipificados $z_i = \frac{e_i}{\widehat{\sigma}}$
3. Estudentizados internos $r_i = \frac{e_i}{\sqrt{\widehat{\sigma}^2(1-h_{ii})}}$
4. Estudentizados externos $t_i \frac{e_i}{\sqrt{\widehat{\sigma}^2_{-i}(1-h_{ii})}} \equiv t_{n-1-p-1}$.

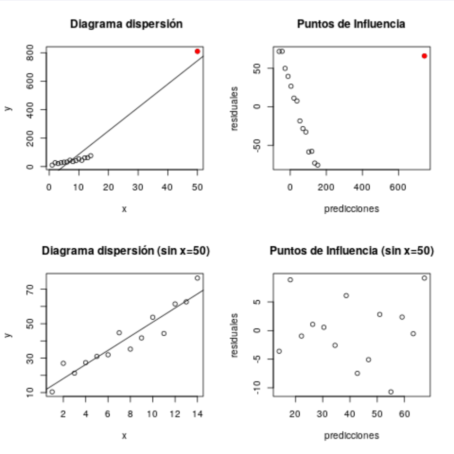
Outliers

Outliers: Observaciones que se comportan de manera muy distinta a la que indica el modelo de regresión.
Efecto: Sobreestiman la varianza de los residuales
Detección: los residuales estudentizados externos



Puntos de Influencia

Puntos de Influencia: Observaciones que modifican mucho las estimaciones de los parámetros al incluirlas en la muestra.



Puntos de Influencia

Detección:

1. **Leverage:** h_{ii}; $\frac{h_{ii}}{1-h_{ii}}$. Con h_{ii}= diagonal(P_x), i=1, ..., p+1.
2. **Distancia de Cook:**

$$D_i = \frac{(\widehat{\beta} - \widehat{\beta}_{-i})^t (X^t X)(\widehat{\beta} - \widehat{\beta}_{-i})}{(p+1)\widehat{\sigma}^2} = \frac{(\widehat{y} - \widehat{y}_{-i})^t (\widehat{y} - \widehat{y}_{-i})}{(p+1)\widehat{\sigma}^2}$$

3. **Dfbetas**

$$Dfbetas_{j,i} = \frac{\beta_j - \beta_{j,-i}}{\widehat{\sigma}_{-i}\sqrt{c_{jj}}} \quad \text{con } c_{jj} = diag_j[(XX)^{-1}]$$

4. **Dffits**

$$Dffits_i = \frac{e_i}{\widehat{\sigma}_{-i}\sqrt{1-h_{ii}}} \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

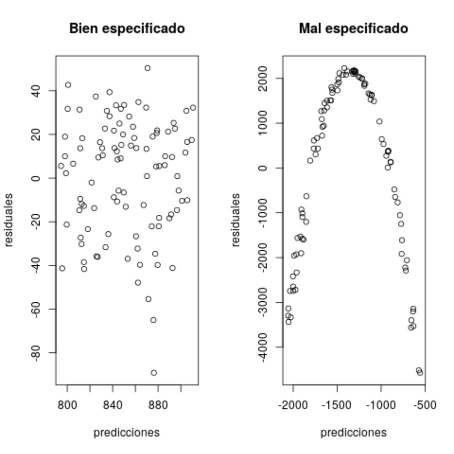
Error de especificación

Error de especificación:
El modelo no recoge de manera adecuada la relación entre la variable dependiente y las independientes.

Posibles causas:

1. Omisión de alguna variable importante.
2. Inclusión de variables innecearias.
3. Expresión funcional inadecuada de alguna variable.

Error de especificación



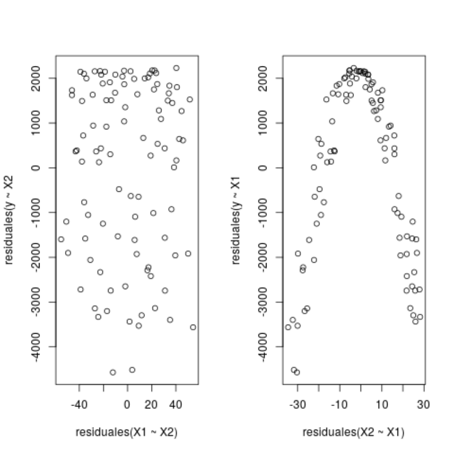
Error de especificación

Gráficos de regresiones parciales (Added variable Plots):
Miden la aportación de cada variable $\vec{x}_j$ al modelo cuando se elimina la influencia de las demás.
Se calculan de la siguiente manera:

- $\widehat{\beta}_{-j} = (X'_{-j}X_{-j})^{-1}X'_{-j}\vec{y}$
- $\vec{e}(\vec{y}|X_{-j}) = [I - X_{-j}(X'_{-j}X_{-j})^{-1}X'_{-j}\vec{y} = (I - P_{-j})\vec{y}$
- $\vec{e}(\vec{x}_j|X_{-j}) = (I - P_{-j})\vec{x}_j$
- Se representa $\vec{e}(\vec{x}_j|X_{-j})$ frente a $\vec{e}(\vec{y}|X_{-j})$

Error de especificación

Gráficos de regresiones parciales (Added variable Plots):



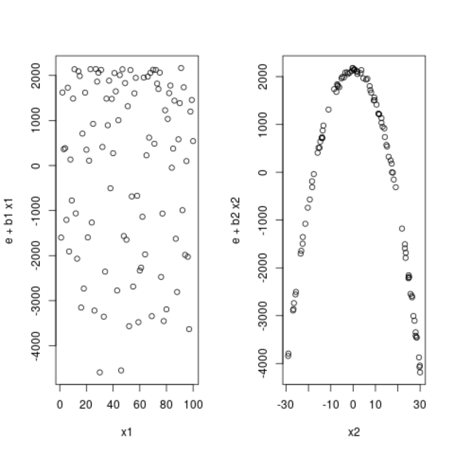
Error de especificación

Gráficos de residuales parciales (Component + residuals Plots): Tratan de determinar si la variable $\vec{x}_j$ tien una aportación no lineal al modelo

- Se obtienen los residuales de la regresión $\vec{e} = (I - P_X)\vec{y}$
- Se calcula la suma $\vec{e} + \widehat{\beta}_j \vec{x}_j$
- Se representa $\vec{e} + \widehat{\beta}_j \vec{x}_j$ frente a $\vec{x}_j$

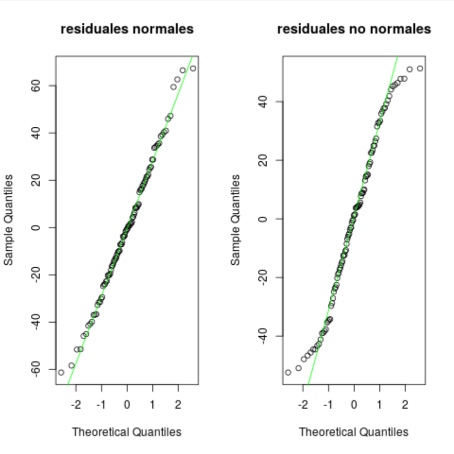
Error de especificación

Gráficos de residuales parciales (Components + residuals Plots):



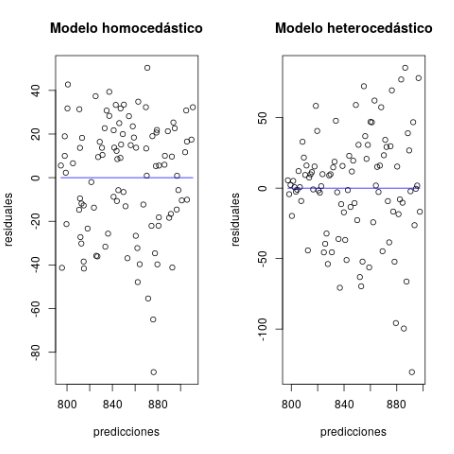
Normalidad

La normalidad es básica para la inferencia en regresión. Para comprobarla se suelen usar métodos gráficos como los QQ-plots.



Homocedasticidad

La varianza de los residuales no es constante.



Homocedasticidad

1. $\vec{\epsilon} \equiv N_n(\vec{0}, \Sigma)$ con $\Sigma = diag(\sigma_{11}, \ldots, \sigma_{nn})$ conocida o se puede estimar. Se trabaja con el modelo transformado:

$$\vec{y}^* = \Sigma^{-1/2}\vec{y} = \Sigma^{-1/2}X\vec{\beta} + \Sigma^{-1/2}\epsilon = X^*\vec{\beta} + \epsilon^*$$

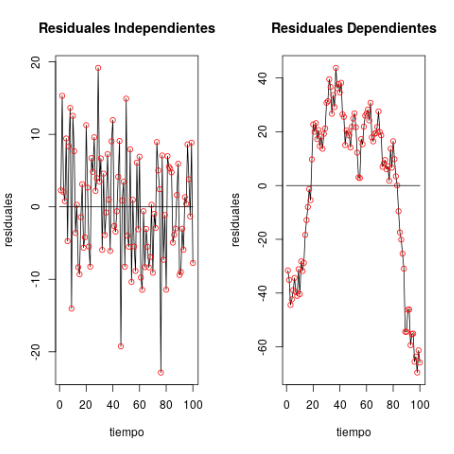
$$\widehat{\beta}^* = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-}1\vec{y}$$

2. Existe relación entre $E(y)$ y $Var(y)$.
Probar las transformaciones: $1/y$; $log(y)$; $\sqrt(y)$; y^2
3. La variable x_k es responsable de la heterocedasticidad. Utilizar por ejemplo:

$$\frac{y}{x_k} = \frac{\beta_0}{x_k} + \beta_1\frac{x_1}{x_k} + \ldots + \beta_p\frac{x_p}{x_k} + \beta_k + \epsilon$$

Independencia

La matriz $\Sigma = Var(\vec{\epsilon})$ no es diagonal.



Independencia

Test de Durbin-Watson: Sirve para contrastar si existe relación entre residuales sucesivos $H_0 : \rho_1 = 0$

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_i^2} \simeq 2(1 - r_1)$$

donde r_1 es el coeficiente de autocorrelación de orden 1.

- Si $r_1 \simeq 0 \Leftrightarrow d \simeq 2$ y no se rechaza H_0
- Si $r_1 >> 0 \Leftrightarrow d \simeq 0$. Autocorrelación positiva
- Si $r_1 << 0 \Leftrightarrow d >> 2$. Autocorrelación negativa

Test de Ljung-Box:

Independencia

1. **Mínimos cuadrados generalizados** $\vec{\epsilon} \equiv N_n(\vec{0}, \Sigma)$, donde $\Sigma \neq$ diagonal pero es conocida o se puede estimar
   Se trabaja con el modelo transformado:

$$\vec{y}^* = \Sigma^{-1/2}\vec{y} = \Sigma^{-1/2}X\vec{\beta} + \Sigma^{-1/2}\epsilon = X^*\vec{\beta} + \epsilon^*$$

$$\widehat{\beta}^* = (X'\Sigma^{-1}X)^{-1}X'\Sigma^-1\vec{y}$$

2. **Operador diferencia**

$$\bigtriangledown y_t = y_t - y_{t-1}$$

Criterios para la selección de modelos

Métodos paso a paso

1. Hacia adelante
2. Hacia atrás
3. Paso a paso

Criterios de selección de variables

- Incremento de R^2
- C_p de Mallows
- Akaike
- BIC

Criterios de parada

- No se incrementa el coeficiente de determinación ajustado.
- Minimizar AIC, (Akaike Information Criterion)

$$AIC = 2k - 2\ln(L) \simeq n\, log(2\pi) + n * log(\widehat{\sigma}_k^2) + 2k$$

  (donde k número de parámetros, L es el máximo valor de la función de verosimilitud para el modelo estimado)
- minimizar BIC (Bayesian Information Criterion)

$$BIC = n * log(\widehat{\sigma}_k^2) + k\, log(n)$$

- minimizar el coeficiente de Mallows, C_p

$$C_p = p + (n-p)\frac{\widehat{s}_R^2(p) - \widehat{s}_R^2(k)}{\widehat{s}_R^2(k)} = \frac{SCE_p}{SCE_k} - n + 2p$$

Regresión logística

Objetivo: predecir el comportamiento de la variable respuesta cuando es cualitativa, frecuentemente dicotómica.

$$E[Y/\overrightarrow{x}] = P(Y = 1/\overrightarrow{x}) = g(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$$

Función logística

$$g(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$

Odds y Odds ratio

Odds (ventaja, disparidad) Cocientre entre la probabilidad de que ocurra un suceso ($Y = 1$) y que no ocurra ($Y = 0$).

$$odds = \frac{P(Y = 1)}{1 - P(Y = 1)}$$

$$Odds \in [0, \infty)$$

$$Odds > 1 \Leftrightarrow P(Y = 1) > 0{,}50$$

Odds-Ratio: cociente de los dos odds asociados a un factor dicotómico F:

$$odds - ratio = \frac{\frac{P(Y=1/F)}{1-P(Y=1/F)}}{\frac{P(Y=1/F^c)}{1-P(Y=1/F^c)}}$$

Mide la influencia de un factor en la presencia de una enfermedad.

```
Riesgo Relativo y Odds Ratio

X= factor de riesgo para contraer una enfermedad .
Si una persona tiene ese factor: X=0 ; en otro caso X=1.

Y= presencia de cierta enfermedad
Sano: Y= 0. Enfero: Y=1.

Supongamos las siguientes probabilidades teóricas:
 P(Enfermo) = P(Y=1) = 0,10;  P(Sano)=P(Y=0)=0,90
 P(Enfermo|No)= P(Y=1|X=0)= 0,1; P(Sano|No)=P(Y=0|X=0)=0,9
 P(Enfermo|Si)= P(Y=1|X=1)= 0,5; P(Sano|Si)=P(Y=0|X=1)=0,5
```

Se elige una muestra aleatoria de 1000 personas:

Frecuencias esperadas

		Y		
		Enfermo	Sano	Total
X	Si	50	50	100
	No	90	810	900
	Total	140	860	1000

Definición de Riesgo Relativo:

RR = P(Enfermo | Si) / P(Enfermo | No)

Riesgo Relativo estimado =(50/100)/(90/900)= 0,5/0,1=5.

Resulta cinco veces más frecuente padecer esta enfermedad entre quienes tienen ese factor que en el resto.

En algunas situaciones no es posible elegir una muestra al azar de la población y es necesario recurrir a muestrear en condiciones mucho más restrictivas.

Por ejemplo: Se elijen aleatoriamente 140 paciente con esa enfermedad y 86 personas sanas

Los resultados del muestreo han sido los siguientes:

		Y		
		Enfermo	Sano	Total
X	Si	50	5	55
	No	90	81	171
	Total	140	86	226

Se mantienen las mismas proporciones que antes

El estimador del Riesgo Relativo cambia drásticamente.

RR= P(Enfemedad|Si)/ P(Enfemedad|No) =(50/55)/(90/171) = 0.909 / 0.526 = 1,727

En este caso el Riesgo Relativo no funciona bien porque:
a) La muestra no mantiene la proporción entre enfermos y sanos de la población.

b) No se estiman bien las probabilidades que intervienen en este indicador.

```
El Odds Ratio afronta este problema.
Se define como cociente de Odss:

Odd (Si)= P(Enfemedad | Si) / P(Sano | Si)

Odd (No)= P(Enfemedad | No) / P(Sano | No)

OR = Odd (Si) / Odd (No) =
[P(Enfemedad|Si)*P(Sano|No)]/[P(Sano|Si)*P(Enfemedad|No)]

Si el OR=1 la enfermedad es independiente del factor.

Cuando el OR>1 el factor favorece la enfermedad.
```

```
En las dos tablas anteriores el valor de los OR son los sig

OR(Tabla 1) = (50 * 810) / (90 * 50) = 9
OR(Tabla 2) = (50 * 81) / (90 * 5) = 9

es decir, el OR no se ve afectado por la falta de proporci

El OR es más difícil de interpretar que el RR pero
hay muchas situaciones en los que debe emplearse el Odds
(como los estudios clínicos de casos y controles, en los es
Ratio frente al Riesgo Relativo.
```

El modelo logístico

$$P(Y = 1/\overrightarrow{x}) = \frac{e^{\vec{x}^t \vec{\beta}}}{1 + e^{\vec{x}^t \vec{\beta}}} = \frac{1}{1 + e^{-\vec{x}^t \vec{\beta}}}$$

$$P(Y = 0/\overrightarrow{x}) = \frac{1}{1 + e^{\vec{x}^t \vec{\beta}}}$$

$$\frac{P(Y = 1/\overrightarrow{x})}{P(Y = 0/\overrightarrow{x})} = e^{\vec{x}^t \vec{\beta}}$$

El modelo logístico

$$Logit(\overrightarrow{x}) = log\left(\frac{P(Y=1/\overrightarrow{x})}{1-P(Y=1/\overrightarrow{x})}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$log\left(\frac{\widehat{P}(Y=1/\overrightarrow{x})}{1-\widehat{P}(Y=1/\overrightarrow{x})}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

Sean $\vec{x}' = (1, x_1, \ldots, x_i, \ldots, x_p)$; $\vec{x*}' = (1, x_1 + 1, \ldots, x_i, \ldots, x_p)$, los coeficientes β_i están relacionados con los ods-ratio

$$P(Y=1|\vec{x*})/P(Y=0|\vec{x*}) = exp(\beta_0 + \beta_1(x_1+1) + \beta_2 x_2 + \beta_p x_p)$$

$$P(Y=1|\vec{x})/P(Y=0|\vec{x}) = exp(\beta_0 + \beta_1(x_1 + \beta_2 x_2 + \beta_p x_p)$$

$$odds - ratio_1 = \frac{P(Y=1|\vec{x*})/P(Y=0|\vec{x*})}{P(Y=1|\vec{x})/P(Y=0|\vec{x})} = exp(\beta_1)$$

Estimación Máximo Verosímil

Verosimilitud de la muestra es

$$L(\vec{\beta}, \vec{y}, X)) = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{exp(\vec{x}_i^t \vec{\beta})}{1 + exp(\vec{x}_i^t \vec{\beta})} \right)^{y_i} \left(\frac{1}{1 + exp(\vec{x}_i^t \vec{\beta})} \right)^{1-y_i}$$

$$log(L(\vec{\beta}, \vec{y}, X)) = \sum_{i=1}^{n} y_i \vec{x}_i^t \vec{\beta} - \sum_{i=1}^{n} log(1 + e^{\vec{x}_i^t \vec{\beta}})$$

Estimación Máximo Verosímil

Ecuación de verosimilitud:

$$\frac{\partial L(\vec{\beta}, \vec{y}, X)}{\partial \vec{\beta}} = \vec{0}$$

Solución de la ecuación de verosimilitud (Numérica):

$$\sum_{i=1}^{n} y_i \vec{x}_i = \sum_{i=1}^{n} \vec{x}_i \frac{1}{1 + e^{-\vec{x}_i^t \widehat{\beta}}}$$

Distribución asintótica de $\widehat{\beta}$

$$\sqrt{n}(\widehat{\beta} - \beta) \longrightarrow^{\mathcal{L}} N(0, I(\beta)^{-1})$$

Estimación Máximo Verosímil

Contrastes individuales $H_0 : \beta_i = 0$

$$-2log\frac{L(X,\widehat{\beta}_0)}{L(X,\widehat{\beta})} \longrightarrow \chi^2_1 \quad \text{bajo } H_0$$

Contraste de la regresión $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$

$$-2log\frac{L(X,\widehat{\vec{\beta}}_0)}{L(X,\widehat{\vec{\beta}})} \longrightarrow \chi^2_p \quad \text{bajo } H_0$$

Test de bondad de ajuste de Hosmer-Lemeshow

$$T = \sum_{j=1}^{g}\left(\frac{(O_{0j}-E_{0j})^2}{E_{0j}} + \frac{(O_{1j}-E_{1j})^2}{E_{1j}}\right) \longrightarrow \chi^2_{g-2} \quad \text{bajo } H_0$$

Residuales

- **Residuos de Pearson**

$$e_i = \frac{y_i - \widehat{p}_i}{\sqrt{\widehat{p}_i(1 - \widehat{p}_i)}}$$

$E(e_i) = 0; Var(e_i) = 1$

$$T = \sum e_i^2 \longrightarrow \chi^2_{n-(p+1)} \quad \text{bajo } H_0$$

- **Deviances**

$$d_i^2 = -2(y_i \log(\widehat{p}_i) + (1 - y_i) \log(1 - \widehat{p}_i))$$

$$D^2 = -2l(\widehat{\beta}) = \sum d_i^2$$