Análisis de Datos e Inferencia

Basado en las clases y apuntes de Carmen del Castillo Vázquez

Autor:
Jorge Rodríguez Domínguez

Índice general

Parte I

Modelos de regresión

Capítulo 1

Modelo de regresión lineal simple

1.1. Introducción

La regresión lineal es un modelo matemático que nos permite establecer la relación de dependencia entre una variable dependiente Y y una variable independiente X. Nos interesan las relaciones de la forma $y = f(x) + u$, donde u es una variable aleatoria a la que llamamos perturbación. En el caso de la regresión lineal simple, el modelo será de la forma

$$y = \beta_0 + \beta_1 x + u,$$

con β_0 y β_1 parámetros. Llamamos intercepto a β_0 y pendiente a β_1.

1.2. Modelo. Hipótesis del modelo

Sea X una variable aleatoria cuantitativa e Y una variable aleatoria continua. Sean $(x_1, y_1), ..., (x_n, y_n)$ datos. El modelo de regresión simple que vamos a estudiar es el siguiente

$$y_i = \beta_0 + \beta_1 \cdot x_i + u_i, \quad i = 1, ..., n.$$

Hipótesis del modelo

H1. $E(u_i) = 0$, $i = 1, ..., n$

H2. $Var(u_i) = \sigma^2$, $i = 1, ..., n$ (Homocedasticidad).

H3. $u_i \sim N(0, \sigma^2)$, $i = 1, ..., n$ (Normalidad).

H4. $E(u_i - u_j) = 0$, $i, j = 1, ..., n$, $i \neq j$ (Independencia).

Traduzcamos estas hipótesis en términos de y_i.

H1. $E(y_i|x_i) = \beta_0 + \beta_1 x_i$, $i = 1, ..., n$ (Linealidad).

H2. $Var(y_i|x_i) = \sigma^2$, $i = 1, ..., n$ (Homocedasticidad).

H3. $y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, ..., n$ (Normalidad).

H4. $Cov(y_i, y_j) = 0$, $i, j = 1, ..., n$, $i \neq j$ (Independencia).

Observación 1.1.

- $E(y_i|x_i = 0) = \beta_0$, $i = 1, ..., n$.
- $E(y_i|x_i + 1) - E(y_i|x_i) = \beta_0 + \beta_1(x + 1) - \beta_0 - \beta_1 x_i = \beta_1$, $i = 1, ..., n$. Podemos decir que β_1 es la variación media que experimenta la variable respuesta (Y), cuando x_i aumenta en una unidad.

1.3. Estimación de los parámetros

1.3.1. Método de máxima verosimilitud

Por H3, tenemos que $y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, ..., n$. Así

$$f(y_i|x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right].$$

De esta forma, la función de máxima verosimilitud es

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f(y_i|x_i), \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]. \end{aligned}$$

Tomando logaritmos

$$\log L(\beta_0, \beta_1\sigma^2) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Para calcular los puntos críticos de esta función, vamos a calcular las derivadas parciales y veremos donde se anulan

$$\begin{aligned} \frac{\partial \log L}{\partial \beta_0}(\beta_0, \beta_1, \sigma^2) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i), \\ \frac{\partial \log L}{\partial \beta_1}(\beta_0, \beta_1, \sigma^2) &= \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i). \end{aligned}$$

Se ha de cumplir entonces que

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta_1} x_i) = 0, \\ \sum_{i=1}^{n} x_i(y_i - \widehat{\beta}_0 - \widehat{\beta_1} x_i) = 0, \end{aligned}$$

que se conocen como **ecuaciones de la regresión**. Llamando $\widehat{y}_i = \widehat{E(y_i|x_i)} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ y $e_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$, tenemos que

$$\begin{aligned} \sum_{i=1}^{n} e_i = 0, \\ \sum_{i=1}^{n} x_i e_i = 0. \end{aligned}$$

Calculando la otra derivada parcial que teníamos pendiente

$$\frac{\partial \log L}{\partial \sigma^2}(\beta_0, \beta_1, \sigma 2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

El estimador de máxima verosimilitud de σ^2, que denotaremos por $\widehat{\sigma^2}$, será (trás desarrollar un poco la expresión anterior)

$$\boxed{\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta_1} x_i)^2.}$$

Intentemos despejar $\widehat{\beta}_0$ y $\widehat{\beta}_1$ de las ecuaciones de la regresión. De dichas ecuaciones tenemos que

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\beta}_0 + \widehat{\beta}_1 x_i) \Longrightarrow \overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x} \Longrightarrow \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x},$$
$$\sum_{i=1}^{n} x_i y_i - \widehat{\beta}_0 \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0,$$

De la segunda ecuación

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} x_i y_i &= \frac{1}{n}\widehat{\beta}_0 \sum_{i=1}^{n} x_i + \frac{1}{n}\widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \frac{1}{n}(\overline{y} - \widehat{\beta}_1 \overline{x}) \sum_{i=1}^{n} x_i + \frac{1}{n}\widehat{\beta}_1 \sum_{i=1}^{n} x_i^2, \\ &= \overline{y} \cdot \overline{x} - \widehat{\beta}_1 \overline{x}^2 + \frac{1}{n}\widehat{\beta}_1 \sum_{i=1}^{n} x_i^2,\end{aligned}$$

es decir,

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{y} \cdot \overline{x} = \widehat{\beta}_1 \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \right) \Longleftrightarrow s_{XY} = \widehat{\beta}_1 s_X^2.$$

Luego,

$$\boxed{\widehat{\beta}_1 = \frac{s_{XY}}{s_X^2}, \quad \widehat{\beta}_0 = \overline{y} - \frac{s_{XY}}{s_X^2}\overline{x}.}$$

1.3.2. Estimación por mínimos cuadrados

Queremos minimizar la suma de los cuadrados de los errores, es decir, minimizar $\sum_{i=1}^{n} e_i^2$ donde $e_i = y_i - \hat{y}_i$. Para ello minimizamos la función $M(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$. Si calculamos las derivadas parciales de M, tenemos que

$$\frac{\partial M}{\partial \beta_0}(\beta_0, \beta_1) = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2,$$
$$\frac{\partial M}{\partial \beta_1}(\beta_0, \beta_1) = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i)^2,$$

de donde obtenemos las mismas condiciones que para los estimadores de máxima verosimilitud, así que los estimadores de β_0 y β_1 por máxima verosimilitud coinciden con los estimadores por mínimos cuadrados.

1.3.3. Estimación de la varianza

Trabajemos un poco con el estimador de máxima verosimilitud de σ^2 para llegar a una expresión equivalente

$$\begin{aligned}\widehat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y} + \widehat{\beta}_1 \overline{x} - \widehat{\beta}_1 x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}((y_i - \overline{y}) - \widehat{\beta}_1(x_i - \overline{x}))^2, \\ &= \frac{1}{n}\sum_{i=1}^{n}[(y_i - \overline{y})^2 + \widehat{\beta}_1^2(x_i - \overline{x})^2 - 2(y_i - \overline{y})\widehat{\beta}_1^2(x_i - \overline{x})], \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 + \frac{\widehat{\beta}_1^2}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 - 2\frac{\widehat{\beta}_1}{n}\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}), \\ &= s_Y^2 + \frac{s_{XY}^2}{(s_X^2)^2}s_X^2 - 2\frac{s_{XY}}{s_X^2}s_{XY} = s_Y^2 - \frac{s_{XY}^2}{s_X^2} \Longrightarrow \boxed{\widehat{\sigma^2} = s_Y^2 - \frac{s_{XY}^2}{s_X^2}.}\end{aligned}$$

Observación 1.2. En el próximo capítulo veremos que

$$\frac{\sum_{i=1}^n e_i^2}{\sigma^2} \sim \chi^2_{n-2},$$

de donde deducimos que

$$E\left(\frac{\sum_{i=1}^n e_i^2}{\sigma^2}\right) = n-2, \quad Var\left(\frac{\sum_{i=1}^n e_i^2}{\sigma^2}\right) = 2(n-2).$$

Entonces, tenemos que

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^n e_i^2}{n} \Longrightarrow E\left(\widehat{\sigma^2}\right) = \frac{1}{n}E\left(\sum_{i=1}^n e_i^2\right) = \frac{n-2}{n}\sigma^2.$$

Es decir, $\widehat{\sigma^2}$ no es un estimador insesgado para σ^2, pero esto se arregla definiendo

$$\boxed{s_R^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}.}$$

que se conoce como la **varianza residual** y es un estimador insesgado para σ^2. Además su varianza es

$$Var(s_R) = Var\left(\frac{\sum_{i=1}^n e_i^2}{n-2}\right) = \frac{n}{n-2}\left(\sigma^2\right)^2.$$

1.4. Propiedades de los estimadores

Podemos escribir $\widehat{\beta}_1$ de la forma:

$$\hat{\beta}_1 = \sum_{i=1}^n w_i y_i, \quad w_i = \frac{x_i - \overline{x}}{n s_X^2}.$$

Por las hipótesis del modelo, y_i son normales e independientes, luego $\hat{\beta}_1 \sim N$. Podemos calcular:

- $E(\widehat{\beta}_1) = \beta_1$ (estimador insesgado).
- $Var(\widehat{\beta}_1) = \dfrac{\sigma^2}{n s_X^2}$.

Por tanto, $\widehat{\beta}_1 \sim N\left(\beta_1, \dfrac{\sigma^2}{n s_X^2}\right)$.

De forma análoga, podemos escribir:

$$\widehat{\beta}_0 = \sum_{i=1}^n \left(\frac{1}{n} - \overline{x} w_i\right).$$

Como las y_i son normales e independientes, $\widehat{\beta}_0 \sim N$. Calculamos:

- $E(\widehat{\beta}_0) = \beta_0$ (estimador insesgado).
- $Var(\widehat{\beta}_0) = \dfrac{\sigma^2}{n}\left(1 + \dfrac{\overline{x}^2}{s_X^2}\right)$.

Por tanto, $\widehat{\beta}_0 \sim N\left(\beta_0, \dfrac{\sigma^2}{n}\left(1 + \dfrac{\overline{x}^2}{s_X^2}\right)\right)$.

En cuanto a s_R^2, sabemos que $\frac{1}{\sigma^2}\sum_{i_1}^n e_i^2 \sim \chi^2_{n-2}$. Obtenemos que:

- $E(s_R^2) = \sigma^2$.
- $V(s_R^2) = \dfrac{2}{n-2}(\sigma^2)^2$.

1.5. Intervalos de confianza para los parámetros

Intervalos de confianza para β_1

Caso 1: σ^2 conocida. Sabemos que

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{ns_X^2}\right) \Longrightarrow \frac{\widehat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{ns_X^2}}} \sim N(0,1)$$

Fijando $\alpha \in (0,1)$ nivel de significación, tenemos que

$$IC_{1-\alpha}(\beta_1) = \left(\widehat{\beta}_1 - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{ns_X^2}}, \widehat{\beta}_1 + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{ns_X^2}}\right),$$

es un intervalo al $(1-\alpha)100\,\%$ para β_1, siendo $z_{1-\alpha/2}$ es el cuantil de orden $1-\alpha$ de una variable aleatoria $Z \sim N(0,1)$,

Caso 2: σ^2 desconocida. Sabemos que

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{ns_X^2}\right) \Longrightarrow \frac{\widehat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{ns_X^2}}} \sim N(0,1)$$

$$\frac{\sum_{i=1}^n e_i^2}{\sigma^2} = \frac{(n-2)s_R^2}{\sigma^2} \sim \chi^2_{n-2},$$

de donde deducimos que

$$\frac{\frac{\widehat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{ns_X^2}}}}{\sqrt{\frac{(n-2)s_R^2}{\sigma^2} \cdot \frac{1}{n-2}}} = \frac{\widehat{\beta}_1 - \beta_1}{\frac{s_R}{\sqrt{ns_X^2}}} \sim t_{n-2}.$$

Fijando $\alpha \in (0,1)$ nivel de significación, tenemos que

$$IC_{1-\alpha}(\beta_1) = \left(\widehat{\beta}_1 - t_{n-2,1-\alpha/2} \cdot \frac{s_R}{\sqrt{ns_X^2}}, \widehat{\beta}_1 + t_{n-2,1-\alpha/2} \cdot \frac{s_R}{\sqrt{ns_X^2}}\right),$$

es un intervalo al $(1-\alpha)100\,\%$ para β_1, siendo $t_{n-2,1-\alpha/2}$ es el cuantil de orden $1-\alpha$ de una variable aleatoria $t \sim t_{n-2}$.

Intervalos de confianza par β_0

Caso 1: σ^2 conocida. Sabemos que

$$\widehat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{s_X^2}\right)\right).$$

Trabajando de forma análoga a como hicimos con β_1, tenemos que fijando $\alpha \in (0,1)$ nivel de significación, entonces

$$IC_{1-\alpha}(\beta_0) = \left(\widehat{\beta}_0 - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\sqrt{1 + \frac{\overline{x}^2}{s_X^2}}, \widehat{\beta}_0 + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\sqrt{1 + \frac{\overline{x}^2}{s_X^2}}\right),$$

es un intervalo al $(1-\alpha)100\,\%$ para β_0, siendo $z_{1-\alpha/2}$ es el cuantil de orden $1-\alpha$ de una variable aleatoria $Z \sim N(0,1)$,

Caso 2: σ^2 desconocida. Trabajando de forma análoga a como hicimos con β_1, tenemos que

$$\frac{\widehat{\beta}_0 - \beta_0}{\frac{s_R}{\sqrt{n}}\sqrt{1+\frac{\overline{x}^2}{s_X^2}}} \sim t_{n-2}.$$

Fijando $\alpha \in (0,1)$ nivel de significación, tenemos que

$$IC_{1-\alpha}(\beta_0) = \left(\widehat{\beta}_0 - t_{n-2,1-\alpha/2}\cdot\frac{s_R}{\sqrt{n}}\sqrt{1+\frac{\overline{x}^2}{s_X^2}}, \widehat{\beta}_0 + t_{n-2,1-\alpha/2}\cdot\frac{s_R}{\sqrt{n}}\sqrt{1+\frac{\overline{x}^2}{s_X^2}}\right),$$

es un intervalo al $(1-\alpha)100\,\%$ para β_1, siendo $t_{n-2,1-\alpha/2}$ es el cuantil de orden $1-\alpha$ de una variable aleatoria $t \sim t_{n-2}$.

Intervalo de confianza para σ^2

Sabemos que

$$\frac{\sum_{i=1}^n e_i^2}{\sigma^2} = \frac{(n-2)s_R^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Fijado $\alpha \in (0,1)$ nivel de significación, queremos encontrar $a, b \in \mathbb{R}$ tales que $P(a < \sigma^2 < b) = 1-\alpha$.

$$\begin{aligned} P(a<\sigma^2<b) &= P\left(\frac{1}{b} < \frac{1}{\sigma^2} < \frac{1}{a}\right) = P\left(\frac{(n-2)s_R^2}{b} < \frac{(n-2)s_R^2}{\sigma^2} < \frac{(n-2)s_R^2}{a}\right) \\ &= P\left(\frac{(n-2)s_R^2}{b} < \chi^2_{n-2} < \frac{(n-2)s_R^2}{a}\right). \end{aligned}$$

De esta forma

$$\begin{aligned} \frac{(n-2)s_R^2}{b} = \chi^2_{n-2,\alpha/2} &\Longrightarrow b = \frac{(n-2)s_R^2}{\chi^2_{n-2,\alpha/2}} \\ \frac{(n-2)s_R^2}{a} = \chi^2_{n-2,1-\alpha/2} &\Longrightarrow a = \frac{(n-2)s_R^2}{\chi^2_{n-2,1-\alpha/2}}. \end{aligned}$$

Tenemos entonces que

$$IC_{1-\alpha}(\sigma^2) = \left(\frac{(n-2)s_R^2}{\chi^2_{n-2,\alpha/2}}, \frac{(n-2)s_R^2}{\chi^2_{n-2,1-\alpha/2}}\right),$$

es un intervalo al $(1-\alpha)100\,\%$ de confianza para σ^2.

1.6. Contraste de la regresión

Consideramos el siguiente contraste

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Fijamos nivel de significación α.

I) Mediante intervalos de confianza.

- Aceptamos H_0 a nivel de significación α si $0 \in IC_{1-\alpha}(\beta_1)$.
- Rechazamos H_0 a nivel de significación α si $0 \notin IC_{1-\alpha}(\beta_1)$.

II) Mediante el estadístico T. Sabemos que

$$T = \frac{\widehat{\beta}_1}{\frac{s_R}{\sqrt{ns_R^2}}} \sim t_{n-2}, \quad \text{si } H_0 \text{ es cierta.}$$

Calculamos t_{exp}.

- Aceptamos H_0 a nivel de significación α si $|t_{exp}| \leq t_{n-2,1-\alpha/2}$.
- Rechazamos H_0 a nivel de significación α si $|t_{exp}| > t_{n-2,1-\alpha/2}$.

III) Mediante el p-valor.

- Si p-valor $\geq \alpha$, aceptamos H_0 a nivel de significación α.
- Si p-valor $< \alpha$, rechazamos H_0 a nivel de significación α.

IV) Mediante el estadístico F. Observamos que

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\overline{y})^2 &= \sum_{i=1}^{n}(y_i-\widehat{y}_i+\widehat{y}_i-\overline{y})^2,\\ &= \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2+\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2+2\sum_{i=1}^{n}(y_i-\widehat{y}_i)(\widehat{y}_i-\overline{y}).\end{aligned}$$

Veamos que el último sumando es igual a cero.

Demostración.

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\widehat{y}_i)(\widehat{y}_i-\overline{y}) &= \sum_{i=1}^{n} e_i\left(\widehat{\beta}_0+\widehat{\beta}_1 x_i-\widehat{\beta}_0-\widehat{\beta}_1\overline{x}\right),\\ &= \widehat{\beta}_1\sum_{i=1}^{n} e_i x_i-\widehat{\beta}_1\overline{x}\sum_{i=1}^{n} e_i,\\ &= \widehat{\beta}_1\cdot 0-\widehat{\beta}_1\overline{x}\cdot 0 = 0.\end{aligned}$$

□

Por tanto, tenemos que

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

Definición 1.3.

- **Variabilidad total**.

$$VT = \sum_{i=1}^{n}(y_i-\overline{y})^2.$$

- **Variabilidad no explicada**.

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

- **Variabilidad explicada**.

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

Observamos que

$$\frac{VNE}{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2_{n-2},$$

y como

$$\frac{\widehat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{ns_X^2}}} \sim N(0,1) \Longrightarrow \left(\frac{\widehat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{ns_X^2}}} \right)^2 \sim \chi^2_1 \Longrightarrow \frac{(\widehat{\beta}_1 - \beta_1)^2}{\frac{\sigma^2}{ns_X^2}} \sim \chi^2_1.$$

Luego

$$\frac{(\widehat{\beta}_1)^2}{\frac{\sigma^2}{ns_X^2}} = \frac{n(\widehat{\beta}_1)^2 s_X^2}{\sigma^2} = \frac{VE}{\sigma^2} \sim \chi^2_1, \quad \text{si } H_0 \text{ es cierta.}$$

Además, como $VT = VNE + VE$

$$\frac{VT}{\sigma^2} \sim \chi^2_{n-1}.$$

Consideremos ahora

$$F = \frac{\frac{VE}{\sigma^2}/1}{\frac{VNE}{\sigma^2}/(n-2)} = \frac{(n-2)VE}{VNE} = \frac{VE}{s_R^2} \sim F_{1,n-2}, \quad \text{si } H_0 \text{ es cierta.}$$

Calculamos F_{exp}.

- Aceptamos H_0 a nivel de significación α si $|F_{exp}| \leq F_{1,n-2,1-\alpha}$.
- Rechazamos H_0 a nivel de significación α si $|F_{exp}| > F_{1,n-2,1-\alpha}$.

Tabla ANOVA

Variabilidad	Suma de cuadrados	Grados de libertad	Cociente	F_{exp}	p-valor
VE	$\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$	1	VE/1	$\frac{VE/1}{VNE/(n-2)}$	p-valor
VNE	$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	$n-2$	VNE/$(n-2)$		
VT	$\sum_{i=1}^{n}(y_i - \overline{y})^2$	$n-1$			

Relación entre t_{exp} y F_{exp}

$$t_{exp} = \frac{\widehat{\beta}_1}{\frac{s_R}{\sqrt{ns_X^2}}} \Longrightarrow t^2_{exp} = \frac{(\widehat{\beta}_1)^2}{\frac{s_R^2}{ns_X^2}} = \frac{ns_X^2(\widehat{\beta}_1)^2}{s_R^2} = \frac{VE}{s_R^2} = F_{exp}.$$

Luego $\boxed{t^2_{exp} = F_{exp}.}$

Significado del contraste

Rechazar H_0 a nivel de significación α si

- β_1 no es significativo a nivel de significación α.
- La relación lineal calculada no es significativa a nivel de significación α.
- No se rechaza la hipótesis de linealidad a nivel de significación α.
- La muestra me ofrece evidencias suficientes para rechazar H_0 a nivel de significación α.

1.7. Evaluación del ajuste

Existen dos coeficientes para evaluar el ajuste del modelo: el coeficiente de correlación lineal y el coeficiente de determinación.

Coeficiente de correlación lineal

El coeficiente de correlación lineal se define como:

$$r = \frac{s_{XY}}{s_X s_Y}, \quad -1 \leq r \leq 1$$

- Si $r = 1$, se tiene dependencia lineal exacta positiva.
- Si $r = -1$, se tiene dependencia lineal exacta negativa.
- Si $r = 0$, las variables están incorreladas linealmente.

Se dice que el ajuste es bueno si $|r|$ es cercano a 1. Si por el contrario r se aproxima a 0, entonces las variables no tienen relación lineal.

Coeficiente de determinación

El coeficiente de determinación se define como:

$$R^2 = \frac{VE}{VT}, \quad 0 \leq R^2 \leq 1$$

- Si $R^2 = 1$ entonces $VE = VT$ luego $VNE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = 0$. Por tanto, $e_i = 0$ para todo $i = 1, \ldots, n$, así que el ajuste lineal es exacto.
- Si $R^2 = 0$ entonces $VE = 0$, luego $VT = VNE$. Así que el ajuste lineal es pésimo.

Teorema 1.4. *El coeficiente de determinación coincide con el coeficiente de correlación lineal al cuadrado. Es decir, $r^2 = R^2$.*

Demostración.

$$R^2 = \frac{VE}{VT} = \frac{n(\widehat{\beta}_1)^2 s_X^2}{n s_Y^2} = \frac{\left(\frac{s_{XY}}{s_X^2}\right)^2 s_X^2}{s_Y^2} = \frac{s_{XY}^2}{s_X^2 s_Y^2} = r^2.$$

□

1.8. Predicción

1.8.1. Estimación de las medias condicionadas

Llamamos $m_0 = E(y|x = x_0) = \beta_0 + \beta_1 x_0$. Observamos que m_0 es un parámetro que podemos estimar de la forma:

$$\widehat{m}_0 = E(\widehat{y|x = x_0}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

Teorema 1.5.

$$\widehat{m}_0 \sim N\left(m_0, \frac{\sigma^2}{n}\left(1 + \frac{(x_0 - \overline{x})^2}{s_X^2}\right)\right).$$

Intervalos de confianza para m_0

Caso 1: σ^2 conocida.

$$IC_{1-\alpha}(m_0) = \left(\widehat{m}_0 - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\sqrt{1 + \frac{(x_0 - \overline{x})^2}{s_X^2}}, \widehat{m}_0 + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\sqrt{1 + \frac{(x_0 - \overline{x})^2}{s_X^2}}\right).$$

Caso 2: σ^2 desconocida.

$$IC_{1-\alpha}(m_0) = \left(\widehat{m}_0 - t_{n-2,1-\alpha/2} \cdot \frac{s_R}{\sqrt{n}}\sqrt{1 + \frac{(x_0 - \overline{x})^2}{s_X^2}}, \widehat{m}_0 + t_{n-2,1-\alpha/2} \cdot \frac{s_R}{\sqrt{n}}\sqrt{1 + \frac{(x_0 - \overline{x})^2}{s_X^2}}\right).$$

1.8.2. Predicción de una observación futura

Dado un conjunto de datos $(x_1, y_1), \ldots, (x_n, y_n)$ y dado x_0 queremos predecir

$$y_0 = \beta_0 + \beta_1 x_0 + u_0,$$

donde u_0 es independiente a $u_1, \ldots, u_n$ con $u_0 \sim N(0, \sigma^2)$. Observamos que y_0 es una variable aleatoria, con estimador $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$. La estimación puntual es:

$$\widehat{y}_0 = \widehat{\beta}_0 + \hat{\beta}_1 x_0 = \widehat{m}_0.$$

Consideramos el error

$$e_0 = y_0 - \widehat{y}_0 = \beta_0 + \beta_1 x_0 + u_0 - (\widehat{\beta}_0 + \widehat{\beta}_1 x_0).$$

que también es una variable aleatoria.

Teorema 1.6.

$$e_0 \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n s_X^2}\right)\right).$$

Demostración. Tenemos que $e_0 = y_0 - \widehat{y}_0$, $y_0 \sim N$, $\widehat{y}_0 = \widehat{m}_0 \sim N$ y que y_0 e $\widehat{y}_0$ son independientes, de donde deducimos que $e_0 \sim N$. Calculemos los parámetros de la distribución de e_0.

- Media.

$$E(e_0) = E(y_0 - \widehat{y}_0) = E(y_0) - E(\widehat{y}_0) = \beta_0 + \beta_1 x_0 - E(\widehat{m}_0) = \beta_0 + \beta_1 x_0 - m_0 = 0.$$

- Varianza.

$$Var(e_0) = Var(y_0 - \widehat{y}_0) = Var(y_0) + Var(\widehat{y}_0) = \sigma^2 + Var(\widehat{m}_0) = \sigma^2 + \frac{\sigma^2}{n}\left(1 + \frac{(x_0 - \overline{x})^2}{s_X^2}\right),$$
$$= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{ns_X^2}\right).$$

□

Intervalos de pronóstico de y_0

Caso 1: σ^2 conocida.

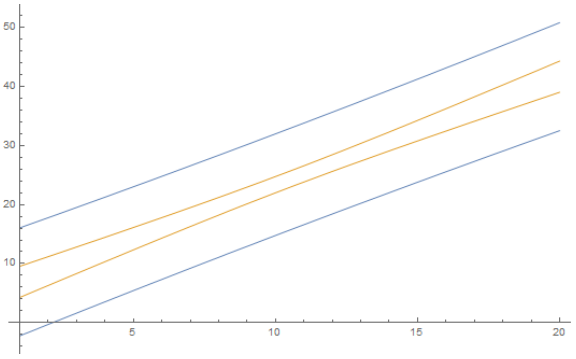
$$IP_{1-\alpha}(y_0) = \left(\widehat{y}_0 - z_{1-\alpha/2}\cdot\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{ns_X^2}}, \widehat{y}_0 + z_{1-\alpha/2}\cdot\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{ns_X^2}}\right).$$

Caso 1: σ^2 desconocida.

$$IP_{1-\alpha}(y_0) = \left(\widehat{y}_0 - t_{n-2,1-\alpha/2}\cdot s_R\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{ns_X^2}}, \widehat{y}_0 + t_{n-2,1-\alpha/2}\cdot s_R\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{ns_X^2}}\right).$$

1.8.3. Bandas de confianza y predicción

Los intervalos de pronóstico de predicción de una observaciones futuras siempre son más grandes que los intervalos de confianza de las estimaciones de las medias condicionadas. Podemos ver esta diferencia en la siguiente gráfica.



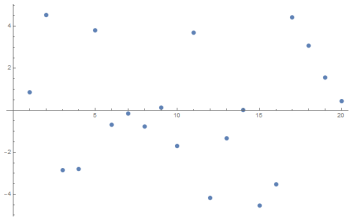
1.9. Diagnosis de las hipótesis del modelo. Análisis de residuos. Observaciones atípicas e influyentes

Lienalidad

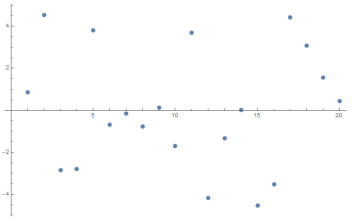
Recordemos que la hipótesis de linealidad era la siguiente

$$E(u_i) = 0 \Longleftrightarrow E(y_i|x_i) = \beta_0 + \beta_1 x_i, \quad \forall i = 1, \dots n.$$

Para ver gráficamente si se cumple esta hipótesis, podemos representar la gráfica de e_i frente a x_i, o bien la gráfica de e_i frente de $\widehat{y}_i$. Si dichas gráficas no presentan una estructura funcional, entonces podemos decir que se cumple la hipótesis de linealidad. Podemos ver que esta hipótesis se cumple en las siguientes gráficas



(a) Gráfica de e_i frente a x_i.



(b) Gráfica de e_i frente a $\widehat{y}_i$.

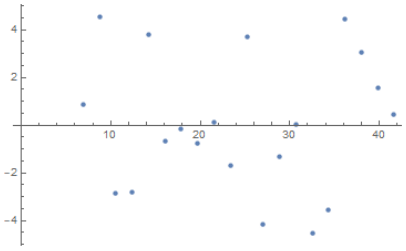
De no cumplirse la hipótesis de linealidad, podríamos tratar de aplicar alguna transformación a X.

Homocedasticidad

Recordemos que la hipótesis de homocedasticidad es la siguiente

$$Var(u_i) = \sigma^2 \iff Var(y_i|x_i) = \sigma^2, \quad \forall i = 1, \dots n.$$

Para ver gráficamente si se cumple esta hipótesis, podemos representar la gráfica de e_i frente a $\widehat{y}_i$. Si dicha gráfica está acotada por "rectas" paralelas, entonces podemos decir que si se cumple la hipótesis de homocedasticidad. Podemos ver que esta hipótesis se cumple en la siguiente gráfica

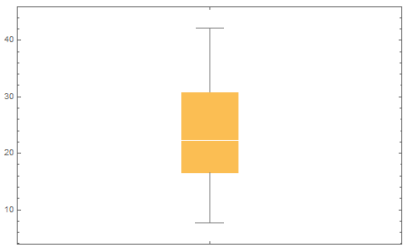


De no cumplirse la hipótesis de linealidad, podriamos tratar de aplicar alguna transformación a Y.

Normalidad

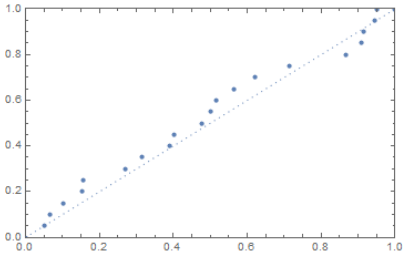
Podemos hacer lo siguiente

a) Gráficos descriptivos: Histograma de los residuos o diagrama de caja y bigotes.



b) Gráficos de probabilidad normal.

- $P - P$: Probabilidad de datos frente a lo normal.
- $Q - Q$: Cuantiles de los datos frente a la normal.



c) Graficar los residuos frente a x_i o frente a $\widehat{y}_i$.

Observaciones atípicas

Una observación atípica es un valor que es numéricamente distinto al resto de los datos. Visualmente, es un dato que se sale del patrón. Las observaciones atípicas pueden ser indicativas de errores de observación o errores en el modelo. Un error de observación se debe a datos que pertenecen a una población diferente del resto de muestras, mientras que un error en el modelo puede ser debido a que la muestra depende una variable desconocida que no se han tenido en cuenta.

Observaciones influyentes

Una observación influyente (x_A, y_A) es una observación atípica cuya exclusión produce un cambio drástico en la recta de regresión. Puede ser causada por un error de observación o por un modelo incorrecto. Algunas posibles causas de que el modelo sea incorrecto son:

- La relación entre x e y no es lineal cerca de x_A.
- La varianza aumenta mucho con x.
- Una variable desconocida ha tomado un valor distinto en x_A.

Puntos palanca

Los puntos palanca son observaciones con un valor alto de p_i. Estos tienen la capacidad de alterar en gran medida la recta de regresión.

1.10. Transformaciones

Cuando el diagrama de dispersión entre las dos variables o el de los residuos presenta indicios de incumplimiento de alguna hipótesis básica, entonces hay que abandonar el modelo inicial por uno menos simple o bien aplicar alguna transformación a los datos.

Capítulo 2

Modelo de regresión lineal múltiple

2.1. Modelo. Hipétesis del modelo

Buscamos un modelo que nos permita establecer una relación entre las variables aleatorias $X_1, \ldots, X_k$ e Y. El modelo que vamos a estudiar es el siguiente

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + u_i,\ i = 1, \ldots, n,$$

donde x_{ji} es el valor de la variable X_j en el i-ésimo individuo, $j = 1, \ldots, k$, $i = 1, \ldots, n$.

Hipótesis del modelo:

H1. $E(u_i) = 0$, $i = 1, \ldots, n$.

H2. $Var(u_i) = \sigma^2$, $i = 1, \ldots, n$ (Homocedasticidad).

H3. $u_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$ (Normalidad).

H4. $E(u_i - u_j) = 0$, $i, j = 1, .., n$, $i \neq j$ (Independencia).

H5. $n > k + 1$.

H6. No existen relaciones lineales entre las variables X_i (Ausencia de multicolinealidad).

Podemos expresar este modelo de forma matricial, si definimos

$$\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n,\ \vec{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} \in \mathbb{R}^{k+1},\ \vec{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \in \mathbb{R}^n,$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_n \end{bmatrix} \in \mathcal{M}_{n\times(k+1)}(\mathbb{R}).$$

Entonces, podemos expresar el modelo anterior como

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

Las hipótesis H1, H2, H3 y H4, se traducen en $\vec{u} \sim N_n(\vec{0}, \sigma^2 I_n)$, de donde deducimos que $\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I_n)$. La hipótesis H5 sigue siendo la misma ($n > k + 1$). La hipótesis H6 se traduce en que $rg(X) = k + 1$.

Hipótesis del modelo:

H1. $E(y_i|x_{1i},\ldots,x_{ki}) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$, $i = 1,,.,n$ (Linealidad).

H2. $Var(y_i|x_{1i},\ldots,x_{ki}) = \sigma^2$, $i = 1,\ldots,n$ (Homocedasticidad).

H3. $y_i|x_{1i},\ldots,x_{ki} \sim N(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}, \sigma^2)$, $i = 1,\ldots,n$.

H4. $Cov(y_i, y_j) = 0$, $i, j = 1,\ldots,n$, $i \neq j$.

H5. $n > k + 1$.

H6. No existen relaciones lineales entres las variables X_j (Ausencia de multicolinealidad).

La interpretacón de los β_j es la siguiente:

- $E(y_1|x_{1i} = \ldots = x_{ki} = 0) = \beta_0$.
- $E(y_i|x_{1i},\ldots,x_{ji}+1,\ldots,x_{ki}) - E(y_i|x_{1i},\ldots,x_{ji},\ldots,x_{ki}) = \beta_j$. Para $j = 1,\ldots,k$, β_j es la variación media de la variable respuesta cuando x_{ji} aumenta en una unidad y el resto de variables permanece constante.

2.2. Estimación de los parámetros

Podemos estimar $E(y_i|x_{i1},\ldots,x_{ki})$ con

$$\widehat{y_i} = \widehat{E(y_i|x_{i1},\ldots,x_{ki})} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \ldots + \widehat{\beta}_k x_{ki}$$

Apliquemos mínimos cuadrados para ver quienes serían $\widehat{\beta}' = (\widehat{\beta}_0,\ldots,\widehat{\beta}_k)$. Definimos

$$M(\beta_0,\ldots,\beta_k) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_k x_{ki})^2.$$

Sus derivadas parciales son

$$\frac{\partial M}{\partial \beta_0}(\beta_0,\ldots,\beta_k) = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_k x_{ki})$$
$$\frac{\partial M}{\partial \beta_j}(\beta_0,\ldots,\beta_k) = -2\sum_{i=1}^{n} x_{ji}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_k x_{ki}, \quad j = 1,\ldots,k$$

Haciendo algunos cálculos se llega a que se tiene que cumplir:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{x1i} - \ldots - \beta_k x_{ki}) = 0$$
$$\sum_{i=1}^{n} x_{ji}(y_i - \beta_0 - \beta_1 x_{x1i} - \ldots - \beta_k x_{ki}) = 0, \quad j = 1,\ldots,k$$

Si denotamos $e_i = y_i - \widehat{y}_i$, $i = 1,\ldots,n$, dichas condiciones se traducen como

$$\sum_{i=1}^{n} e_i = 0$$
$$\sum_{i=1}^{n} x_{ji} e_i = 0, \quad j = 1,\ldots,k,$$

que se conocen como las ecuaciones normales de la regresión. Desarrollando cálculos

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_{1i} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{2i} + \ldots + \widehat{\beta}_k \sum_{i=1}^{n} x_{ki} \\
\sum_{i=1}^{n} x_{1i} y_i &= \widehat{\beta}_0 \sum_{i=1}^{n} x_{1i} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 + \widehat{\beta}_2 \sum_{i=1}^{n} x_{1i}x_{2i} + \ldots + \widehat{\beta}_k \sum_{i=1}^{n} x_{1i}x_{ki} \\
\sum_{i=1}^{n} x_{2i} y_i &= \widehat{\beta}_0 \sum_{i=1}^{n} x_{2i} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{1i}x_{2i} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{2i}^2 + \ldots + \widehat{\beta}_k \sum_{i=1}^{n} x_{2i}x_{ki} \\
&\vdots \\
\sum_{i=1}^{n} x_{ki} y_i &= \widehat{\beta}_0 \sum_{i=1}^{n} x_{ki} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{1i}x_{ki} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{2i}x_{ki} + \ldots + \widehat{\beta}_k \sum_{i=1}^{n} x_{ki}^2
\end{aligned}$$

De manera matricial nos quedaría

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

Nótese que $(X'X)^{-1}$ existe por H6.

Observación 2.1.

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2_{n-(k+1)} \equiv \chi^2_{n-k-1}$$

Entonces

$$E\left(\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2}\right) = n-k-1 \quad \text{y} \quad Var\left(\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2}\right) = 2(n-k-1).$$

Si definimos

$$\boxed{s_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}}$$

Tenemos que

$$\begin{aligned}
E(s_R^2) &= \frac{1}{n-k-1} E\left(\sum_{i=1}^{n} e_i^2\right) = \frac{1}{n-k-1}\sigma^2(n-k-1) = \sigma^2 \\
Var(s_R^2) &= \frac{1}{(n-k-1)^2} Var\left(\sum_{i=1}^{n} e_i^2\right) = \frac{1}{(n-k-1)^2} 2(n-k-1)(\sigma^2)^2 = \frac{2(\sigma^2)^2}{n-k-1}
\end{aligned}$$

Todo esto nos dice que s_R^2 es un estimador insesgado para σ^2.

2.3. Propiedades de los estimadores

Ya hemos probado que $\widehat{\vec{\beta}} = (X'X)^{-1}X'\vec{y}$ y que $\vec{y} = N_n(X\vec{\beta}, \sigma^2 I_n)$, de donde deducimos que $\widehat{\vec{\beta}} \sim N_{k+1}(\vec{\beta}, \sigma^2(X'X)^{-1})$. Veámoslo.

Demostración.

$$\begin{aligned}
E(\widehat{\vec{\beta}}) &= (X'X)^{-1}XE(\vec{y}) = (X'X)^{-1}X\vec{\beta} = \vec{\beta} \\
Cov(\widehat{\vec{\beta}}) &= (X'X)^{-1}XCov(\vec{y})X(X'X)^{-1} = (X'X)^{-1}X'\sigma I_n X(X'X)^{-1} = \sigma^2(X'X)^{-1} \\
&= \begin{bmatrix}
Var(\widehat{\beta}_0) & Cov(\widehat{\beta}_0, \widehat{\beta}_1) & \cdots & Cov(\widehat{\beta}_0, \widehat{\beta}_k) \\
Cov(\widehat{\beta}_0, \widehat{\beta}_1) & Var(\widehat{\beta}_2) & \cdots & Cov(\widehat{\beta}_2, \widehat{\beta}_k) \\
\vdots & \vdots & \ddots & \vdots \\
Cov(\widehat{\beta}_0, \widehat{\beta}_k) & Cov(\widehat{\beta}_0, \widehat{\beta}_{k-1}) & \cdots & Var(\widehat{\beta}_k)
\end{bmatrix}
\end{aligned}$$

Así, $\widehat{\beta}_j \sim N(\beta_j, \sigma^2 q_{j+1,j+1})$, siendo $q_{j+1,j+1}$ el elemento $(j+1)$-ésimo de la diagonal de $(X'X)^{-1}$. □

2.4. Intervalos de confianza para los parámetros

2.4.1. Intervalos de confianza para β_j, $j = 0, 1, \ldots, k$

Supondremos que σ^2 es desconocido (que suele ser lo común en la realidad). Sea $j \in \{0, 1, \ldots, k\}$, entonces $\widehat{\beta}_j \sim N(\beta_j, \sigma^2 q_{j+1,j+1})$, tipificando

$$\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{q_{j+1,j+1}}} \sim N(0,1).$$

Sabemos además que

$$\frac{(n-k-1)s_R^2}{\sigma^2} = \frac{\sum_{i=1}^n e_i^2}{\sigma^2} \sim \chi^2_{n-k-1}.$$

Entonces

$$\frac{\dfrac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{q_{j+1,j+1}}}}{\sqrt{\dfrac{(n-k-1)s_R^2}{\sigma^2} \cdot \dfrac{1}{n-k-1}}} = \frac{\widehat{\beta}_j - \beta_j}{s_R\sqrt{q_{j+1,j+1}}} \sim t_{n-k-1}.$$

Entonces

$$IC_{1-\alpha}(\beta_j) = \left(\widehat{\beta}_j - t_{n-k-1,1-\alpha/2} \cdot s_R\sqrt{q_{j+1,j+1}}, \widehat{\beta}_j + t_{n-k-1,1-\alpha/2} \cdot s_R\sqrt{q_{j+1,j+1}}\right),$$

siendo $t_{n-k-1,1-\alpha/2}$ el cuantil $1-\alpha/2$ de la distribución t_{n-k-1}.

2.4.2. Intervalo de confianza para σ^2

Fijamos $\alpha \in (0,1)$. Sean $0 < a < b$, entonces

$$\begin{aligned} 1-\alpha = P(a < \sigma^2 < b) &= P\left(\frac{1}{b} < \frac{1}{\sigma^2} < \frac{1}{a}\right) \\ &= P\left(\frac{(n-k-1)s_R^2}{b} < \frac{(n-k-1)s_R^2}{\sigma^2} < \frac{(n-k-1)s_R^2}{a}\right) \end{aligned}$$

Sabemos que

$$\frac{(n-k-1)s_R^2}{\sigma^2} = \frac{\sum_{i=1}^n e_i^2}{\sigma^2} \sim \chi^2_{n-k-1}.$$

Entonces

$$IC_{1-\alpha}(\sigma^2) = \left(\frac{(n-k-1)s_R^2}{\chi^2_{n-k-1,1-\alpha/2}}, \frac{(n-k-1)s_R^2}{\chi^2_{n-k-1,\alpha/2}}\right),$$

siendo $\chi^2_{n-k-1,1-\alpha/2}$ y $\chi^2_{n-k-1,\alpha/2}$ los cuantiles $1-\alpha/2$ y $\alpha/2$ de la distribución χ^2_{n-k-1} respectivamente.

2.5. Contrastes de hipótesis para los coeficientes de regresión

2.5.1. Contrastes de significación individuales

Queremos saber si la variable X_j es influyente en el modelo. Para cada $j = 1, \ldots, k$ planteamos el contraste

$$\begin{cases} H_0 \ : \ \beta_j = 0 \\ H_1 \ : \ \beta_j \neq 0 \end{cases}$$

Este contraste quiere decir si hay suficiente evidencia en la muestra para afirmar que la variable X_j tiene una influencia lineal significativa en el modelo.

Fijemos nivel de signifación α.

I) Mediante intervalos de confianza. Sabemos que

$$IC_{1-\alpha}(\beta_j) = \left(\widehat{\beta}_j - t_{n-k-1,1-\alpha/2} \cdot s_R\sqrt{q_{j+1,j+1}}, \widehat{\beta}_j + t_{n-k-1,1-\alpha/2} \cdot s_R\sqrt{q_{j+1,j+1}}\right).$$

- Aceptamos H_0 si $0 \in IC_{1-\alpha}(\beta_j)$ a nivel de significación α.
- Rechazamos H_0 si $0 \notin IC_{1-\alpha}(\beta_j)$ a nivel de significación α.

II) Mediante el estadístico T. Sabemos que

$$T = \frac{\widehat{\beta}_j}{s_R\sqrt{q_{j+1,j+1}}} \sim t_{n-k-1}, \quad \text{si } H_0 \text{ es cierta.}$$

Calculamos t_{exp}.

- Aceptamos H_0 a nivel de significación α si $|t_{exp}| \leq t_{n-k-1,1-\alpha/2}$.
- Rechazamos H_0 a nivel de significación α si $|t_{exp}| > t_{n-k-1,1-\alpha/2}$.

III) Mediante el p-valor.

- Si p-valor $\geq \alpha$, aceptamos H_0 a nivel de significación α.
- Si p-valor $< \alpha$, rechazamos H_0 a nivel de significación α.

2.5.2. Contraste de regresión

Plantemos el siguiente contraste

$$\begin{cases} H_0 \ : \ \beta_1 = \ldots = \beta_k = 0 \\ H_1 \ : \ \exists \ i \in \{1, \ldots, k\} \ : \ \beta_i \neq 0 \end{cases}$$

Este contraste quiere decir si hay suficiente evidencia en la muestra como para afirmar que el modelo global es válido.

Fijamos nivel de significación α. Recordemos que

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

y denotemos por

$$VT = \sum_{i=1}^{n}(y_i - \overline{y})^2, \quad VNE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2, \quad VE = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$$

Entonces

$$\frac{\sum_{i=1}^n e_i^2}{\sigma^2} \sim \chi^2_{n-k-1} \Longrightarrow \frac{VNE}{\sigma^2} \sim \chi^2_{n-k-1}$$
$$\frac{VE}{\sigma^2} \sim \chi^2_k, \quad \text{si } H_0 \text{ es cierta}$$
$$\frac{VT}{\sigma^2} \sim \chi^2_{n-1}, \quad \text{si } H_0 \text{ es cierta}$$

donde usamos que $s_R^2 = \frac{VNE}{n-k-1}$. El estadístico de contraste es

$$F = \frac{\left(\frac{VE}{\sigma^2}\right)/k}{\left(\frac{VNE}{\sigma^2}\right)/(n-k-1)} \sim F_{k,n-k-1}, \quad \text{si } H_0 \text{ es cierta.}$$

Desarrollando cálculos

$$F = \frac{(n-k-1)\cdot VE}{k \cdot VNE} = \frac{VE}{k \cdot s_R^2} \sim F_{k,n-k-1}, \quad \text{si } H_0 \text{ es cierta.}$$

Cuando calculemos F_{exp}

- Si $F_{exp} \leq F_{k,n-k-1,1-\alpha}$, aceptamos H_0 a nivel de significación α.
- Si $F_{exp} > F_{k,n-k-1,1-\alpha}$, rechazamos H_0 a nivel de significación α.

2.5.3. Tabla ANOVA

Variabilidad	Suma de cuadrados	Grados de libertad	Cociente	F_{exp}	p-valor
VE	$\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2$	k	VE/k	$\frac{VE/k}{VNR/(n-k-1)}$	p-valor
VNE	$\sum_{i=1}^n (y_i - \widehat{y}_i)^2$	$n-k-1$	VNE/$(n-k-1)$		
VT	$\sum_{i=1}^n (y_i - \overline{y})^2$	$n-1$			

Demos otras fórmulas para la variabilidad.

$$\begin{aligned} ns_Y^2 = VT &= \sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n y_i^2 + n\overline{y}^2 - 2\overline{y}\sum_{i=1}^n y_i = \sum_{i=1}^n y_i^2 + n\overline{y}^2 - 2\overline{y}n\overline{y} \\ &= \sum_{i=1}^n y_i^2 - n\overline{y}^2 = \vec{y}'\vec{y} - n\overline{y}^2 \\ VNE &= \sum_{i=1}^n (y_i - \widehat{y}_i)^2 = \left(\vec{y} - \widehat{\vec{y}}\right)' \left(\vec{y} - \widehat{\vec{y}}\right) = \left(\vec{y} - X\widehat{\vec{\beta}}\right)' \left(\vec{y} - X\widehat{\vec{\beta}}\right) \\ &= \vec{y}'\vec{y} - \vec{y}'X\widehat{\vec{\beta}} - \widehat{\vec{\beta}}'X'\vec{y} + \widehat{\vec{\beta}}X'X\widehat{\vec{\beta}} \\ &= \vec{y}'\vec{y} - \widehat{\vec{\beta}}'X'\vec{y} + \left(\widehat{\vec{\beta}}'X'X - \vec{y}'X\right)\widehat{\vec{\beta}} = \vec{y}'\vec{y} - \widehat{\vec{\beta}}'X'\vec{y}, \end{aligned}$$

donde en la útlima igualdad usamos que

$$\widehat{\vec{\beta}}'X'X - \vec{y}'X = \ldots = 0$$

Finalmente,

$$VE = VT - VNE = \ldots = \widehat{\vec{\beta}}'X'\vec{y} - n\overline{y}^2$$

2.5.4. Interpretación de los contrastes sobre los coeficientes de regresión

Los casos que se pueden representar al realizar contrastes de hipótesis en un modelo de regresión son los siguientes

Casos	Contraste conjunto	Contraste individual
1	Significativo	Todos significativos
2	Significativo	Algunos significativos
3	Significativo	Ninguno significativo
4	No significativo	Todos significativos
5	No significativo	Algunos significativos
6	No significativo	Ninguno significativo

donde "significativo" indica que se rechaza la hipótesis H_0 de que el parámetro o los parámetros al o a los que se refiere la hipótesis sea(n) 0.

1. Indica que todas las variables explicativas influyen.
2. Indica que solamente influyen algunas variables explicativas, por lo que en principio, se deberían eliminar las no significativas del modelo, pero no debe hacerse mecánicamente sino estudiando más en profundidad cuál sería el modelo que se seleccionaría.
3. Indica que las variables X_j son muy dependientes entre sí y aunque conjuntamente influyen, individualmente son no signitifivas (Multicolinealidad).
4. Es poco frecuente y es un tipo de multicolinealidad especial: si dos variables influyen sobre Y pero en sentido contrario, su efecto conjunto puede ser no significativo, aunque sus efectos individuales si lo sean.
5. Es análogo a 4.
6. Ninguna de las variables parece tener efecto sobre Y, pero sólo podremos decir que sus efectos no se detectan en la muestra considerada.

2.5.5. Contrastes de grupos de coeficientes

Planteamos el siguiente contraste

$$\begin{cases} H_0 \ : \ \beta_{j_1} = \ldots = \beta_{j_i} = 0 \\ H_1 \ : \ \exists \ j \in \{j_1, \ldots, j_i\} \ : \ \beta_j \neq 0 \end{cases}$$

Denotamos por VE(k) a la variabilidad explicada por el modelo con las variables $X_1, \ldots, X_k$ como variables explicativas. VE($k-i$) a la variabilidad explicada por el modelo con todas las variables explicativas excepto $X_{j_1}, \ldots, X_{j_i}$. ΔVE(k) = VE(k) - VE($k-i$), que es aproximadamente la variablidad explicada por las variables $X_{j_1}, \ldots, X_{j_i}$. VNE(k) a la variabilidad no explicada por el modelo con las variables $X_1, \ldots, X_k$ como variables explicativas.

Fijamos nivel de significación α. El estadístico de contraste es

$$F = \frac{\Delta VE/i}{VNE(k)/(n-k-1)} \sim F_{i,n-k-1}, \quad \text{si } H_0 \text{ es cierta.}$$

Desarrollando cálculos

$$F = \frac{(n-k-1)\Delta VE}{i \cdot VNE(k)} = \frac{(n-k-1)\Delta VE}{i \cdot s_R^2(n-k-1)} = \frac{\Delta VE}{i \cdot s_R^2} \sim F_{i,n-k-1}, \quad \text{si } H_0 \text{ es cierta.}$$

Cuando calculemos F_{exp}

- Si $F_{exp} \leq F_{i,n-k-1,1-\alpha}$, aceptamos H_0 a nivel de significación α.

- Si $F_{exp} > F_{i,n-k-1,1-\alpha}$, rechazamos H_0 a nivel de significación α.

Observación 2.2. Podemos usar este contraste para contrastes individuales:

$$\begin{cases} H_0 \ : \ \beta_i = 0 \\ H_1 \ : \ \beta_i = 0 \end{cases}$$

Solo en este caso se tiene $F_{exp} = t_{exp}^2$.

2.6. Correlación en regresión múltiple

2.6.1. Coeficiente de determinación

$$R^2 = \frac{VE}{VT} = \frac{VT - VNE}{VT} = 1 - \frac{VNE}{VT} = 1 - \frac{(n-k-1)s_R^2}{VT}$$

Observación 2.3.

- $0 \leq R^2 \leq 1$.
- Si k aumenta, entonces R^2 aumenta (aún en el caso de que dichas variables no sean significativas).

2.6.2. Coeficiente de determinación ajustado o corregido

$$\begin{aligned} \overline{R}^2 &= 1 - \frac{VNE/(n-k-1)}{VT/(n-1)} = 1 - \frac{n-1}{n-k-1} \cdot \frac{VNE}{VT} = 1 - \frac{n-1}{n-k-1} \cdot \frac{(n-k-1)s_R^2}{VT} \\ &= 1 - (n-1)\frac{s_R^2}{VT} \end{aligned}$$

Observación 2.4. $\overline{R}^2$ aumenta si y solo si s_R^2 disminuye.

Observación 2.5. Relación entre R^2 y $\overline{R}^2$.

$$\begin{aligned} &R^2 = 1 - \frac{(n-k-1)s_R^2}{VT} \Longrightarrow 1 - R^2 = \frac{(n-k-1)s_R^2}{VT} \Longrightarrow \frac{s_R^2}{VT} = \frac{1-R^2}{n-k-1} \\ &\overline{R}^2 = 1 - (n-1)\frac{s_R^2}{VT} \Longrightarrow \overline{R}^2 = 1 - \frac{(n-1)(1-R^2)}{n-k-1} \end{aligned}$$

de donde concluimos $\boxed{\overline{R}^2 \leq R^2 \leq 1}$. Veámoslo.

Demostración.

$$\begin{aligned} n-k-1 \leq n-1 \Longrightarrow \frac{n-1}{n-k-1} \geq 1 \Longrightarrow \frac{n-1}{n-k-1}(1-R^2) \geq (1-R^2) \\ \Longrightarrow 1 - \overline{R}^2 \geq 1 - R^2 \Longrightarrow \overline{R}^2 \leq R^2 \end{aligned}$$

□

2.7. Predicción

2.7.1. Estimación de las medias condicionadas

Supongamos que queremos predecir

$$m_0 \equiv E(y|x_{10},\ldots,x_{k0}) = \beta_0 + \beta_1 x_{10} + \ldots + \beta_k x_{k0} = \vec{x}_0'\vec{\beta}$$

Podemos estimarla como

$$\widehat{m}_0 \equiv E(y|\widehat{x_{10},\ldots},x_{k0}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_{10} + \ldots + \widehat{\beta}_k x_{k0} = \vec{x}_0'\widehat{\vec{\beta}}$$

Recordemos que

$$\begin{aligned}\vec{\beta}' &= (\beta_0,\beta_1,\ldots,\beta_k) \in \mathbb{R}^{k+1}\\ \vec{x}_0' &= (1,x_{10},\ldots,x_{k0}) \in \mathbb{R}^{k+1}\\ \widehat{\vec{\beta}}' &= (\widehat{\beta}_0,\widehat{\beta}_1,\ldots,\widehat{\beta}_k) \in \mathbb{R}^{k+1}.\end{aligned}$$

Ya vimos que $\widehat{\vec{\beta}} \sim N_{k+1}\left(\vec{\beta},\sigma^2(X'X)^{-1}\right)$, de donde deducimos que

$$\widehat{m}_0 \sim N\left(\vec{x}_0'\vec{\beta},\vec{x}_0' Var\left(\widehat{\vec{\beta}}\right)\vec{x}_0\right) \equiv N\left(\vec{x}_0'\vec{\beta},\sigma^2\vec{x}_0'(X'X)^{-1}\vec{x}_0\right),$$

de aquí deducimos

$$\frac{\widehat{m}_0 - m_0}{\sigma\sqrt{\vec{x}_0'(X'X)^{-1}\vec{x}_0}} \sim N(0,1).$$

Si suponemos que σ es desconocido (que es lo normal en la realidad), entonces

$$\frac{\widehat{m}_0 - m_0}{s_R\sqrt{\vec{x}_0'(X'X)^{-1}\vec{x}_0}} \sim t_{n-k-1}.$$

Entonces

$$IC_{1-\alpha}(m_0) = \left(\widehat{m}_0 - t_{n-k-1,1-\alpha/2}\cdot s_R\sqrt{\vec{x}_0'(X'X)^{-1}\vec{x}_0}, \widehat{m}_0 + t_{n-k-1,1-\alpha/2}\cdot s_R\sqrt{\vec{x}_0'(X'X)^{-1}\vec{x}_0}\right)$$

2.7.2. Predicción de una nueva observación

Recodermos que

$$\begin{aligned}y_0 &= \beta_0 + \beta_1 x_{10} + \ldots + \beta_k x_{k0} + u_0 = \vec{x}_0'\vec{\beta} + u_0\\ \widehat{y}_0 &= \widehat{\beta}_0 + \widehat{\beta}_1 x_{10} + \ldots + \widehat{\beta}_k x_{k0} = \vec{x}_0'\vec{\beta} = \widehat{m}_0\end{aligned}$$

Veamos que distribución sigue $e_0 = y_0 - \widehat{y}_0$. Sabemos que

$$\begin{aligned}y_0 &\sim N\left(\vec{x}_0'\vec{\beta},\sigma^2\right)\\ \widehat{y}_0 = \widehat{m}_0 &\sim N\left(m_0,\sigma^2\vec{x}_0'(X'X)^{-1}\vec{x}_0\right),\end{aligned}$$

de donde deducimos que e_0 sigue una distribución Normal (pues es diferencia de distribuciones normales) y porque y_0 e $\widehat{y}_0$ son independientes (pues $u_0,\ldots,u_n$ son independientes). Veamos que media y varianza tiene e_0.

- $E(e_0) = E(y_0) - E\left(\widehat{y}_0\right) = \vec{x}_0'\vec{\beta} - m_0 = m_0 - m_0 = 0.$

- $Var(e_0) = Var(y_0) + Var\left(\widehat{y}_0\right) = \sigma^2 + \sigma^2 \vec{x}_0'(X'X)^{-1}\vec{x}_0 = \sigma^2\left(1 + \vec{x}_0'(X'X)^{-1}\vec{x}_0\right)$

Esto es,

$$e_0 \sim N\left(0, \sigma^2\left(1 + \vec{x}_0'(X'X)^{-1}\vec{x}_0\right)\right)$$

Calculemos un intervalo probabilistico para y_0. Tenemos que

$$\frac{e_0}{\sigma\sqrt{(1 + \vec{x}_0'(X'X)^{-1}\vec{x}_0)}} \sim N(0,1) \Longrightarrow \frac{e_0}{s_R\sqrt{(1 + \vec{x}_0'(X'X)^{-1}\vec{x}_0)}} \sim t_{n-k-1}.$$

Finalmente

$$IP_{1-\alpha}(y_0) = \left(\widehat{y}_0 - t_{n-k-1,1-\alpha/2} \cdot s_R\sqrt{(1 + \vec{x}_0'(X'X)^{-1}\vec{x}_0)}, \widehat{y}_0 + t_{n-k-1,1-\alpha/2} \cdot s_R\sqrt{(1 + \vec{x}_0'(X'X)^{-1}\vec{x}_0)}\right)$$

2.8. Diagnosis y validación del modelo

2.8.1. Multicolinealidad

El primer problema que surge es la dependencia de las variables explicativas (o regresores) entre sí, es decir, la existencia de una o más combinaciones lineales entre las columnas de la matriz X

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$$

Problema: cuando $rg(X) < k+1$. Cuando esto ocurre es dificil separar los efectos de cada variable explicativa y medir la contribución individual, con lo que los estimadores individuales serán inestables y con gran varianza. A este problema se le denomina multicolinealidad y consiste en querer extraer de la muestra más información de la que contiene.

Existen dos tipos de multicolinealidad:

1. **Multicolinealidad perfecta**. Una de las variables explicativas es combinación lineal exacta de las demás: $rg(X) < k+1 \Longrightarrow det(X'X) = 0 \Longrightarrow$ no se puede calcular $(X'X)^{-1}$. El sistema de ecuaciones que determina el vector $\widehat{\beta}$ no tiene solución única.
2. **Alta multicolincalidad**. Cuando alguna o todas las variables explicativas están altamente correlacionadas entre sí (pero el coeficiente de correlación no llega a ser 1 ni -1). En este caso las columnas de la matriz X tienen un alto grado de dependencia entre sí, pero sí puede calcularse el vector $\widehat{\beta}$, aunque:
   a) Los estimadores $\widehat{\beta}$ tendrán varianzas muy altas, lo que provocará mucha imprecisión en la estimación de los $\widehat{\beta}_j$, y, por tanto, los intervalos de confianza serán muy anchos.
   b) Los estimadores $\widehat{\beta}_j$, serán muy dependientes entre sí, puesto que tendrán altas covarianzas y habrá poca inforrmación sobre lo que ocurre al variar una variable si las demás permanecen constantes.

Consecuencias de la multicolinealidad

- Los estimadores $\widehat{\beta}_j$, serán muy sensibles a pequeñas variaciones en el tamaño muestral o a la supresión de una variable aparentemente no significativa. A pesar de esto, la predicción no tiene por qué verse afectada ante la multicolinealidad, ni ésta afecta al vector de residuos, que está siempre bien definido.
- Los coeficientes de regresión pueden ser no significativos individualmente (puesto que las varianzas de los $\widehat{\beta}_j$, van a ser grandes), aunque el contraste global del modelo sea significativo.
- La multicolinealidad puede afectar mucho a algunos parámetros y nada a otros. Los parámetros que estén asociados a variables explicativas poco correlacionadas con el resto no se verán afectados y podrán estimarse con precisión.

Identificación de la multicolinealidad

La indentificación de la multicolinealidad se realiza examinando

1) La matriz de correlaciones entre las variables explicativas, $\mathbf{R}$, y su inversa $\mathbf{R}^{-1}$.

2) Los autovalores de $X'X$ o de $\mathbf{R}$.

En el caso 1) la presencia de correlaciones altas entre variables explicativas es un indicio de multicolinealidad. Pero, es posible que exista una relación perfecta entre una variable y el resto y, sin embargo, sus coeficientes de correlación sean bajos (por ejemplo, cuando sea el caso de una relación no lineal).

Definimos la matriz de correlaciones como

$$\mathbf{R} = \begin{bmatrix} 1 & r_{11} & \cdots & r_{1k} \\ r_{12} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1k} & r_{2k} & \cdots & 1 \end{bmatrix}, \quad \text{donde} \quad r_{ij} = \frac{s_{X_i,X_j}}{s_{X_i} s_{X_j}}, \ -1 \leq r_{ij} \leq 1,$$

que es una matriz de orden k, simétrica, con unos en la diagonal.

La inversa de la matriz de correlaciones

$$\mathbf{R}^{-1} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1k} \\ \gamma_{12} & \gamma_{22} & \cdots & \gamma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{1k} & \gamma_{2k} & \cdots & \gamma_{kk} \end{bmatrix}$$

tiene en cuenta la dependencia conjunta. Los elementos de su diagonal se denominan **factores de incremento** o **de inflación de la varianza** y verifican

$$\gamma_{ii} = FIV(i) = \frac{1}{1 - R^2_{i,resto}}, \quad i = 1, \ldots, k,$$

donde $R^2_{i,resto}$ es el coeficiente de determinaciónn de la regresión de la variable X_i en función del resto de variables explicativas, es decir, $X_1, .., X_{i-1}, X_{i+1}, \ldots, X_k$. Por tanto, si para algún i se tiene que

$$\gamma_{ii} > 10 \Longleftrightarrow \frac{1}{1 - R^2_{i,resto}} > 10 \Longleftrightarrow 1 - R^2_{i,resto} < 0,1 \Longleftrightarrow R^2_{i,resto} > 0,9$$

es decir, la variable X_j se explica a un 90 % por el resto de variables explicativas, por tanto, estamos en una situación de alta multicolinealidad.

Inconveniente: $\mathbf{R}^{-1}$ se calculará con poca precisión cuando $\mathbf{R}$ sea casi singular ($det(\mathbf{R}) \approx 0$).

En el caso 2) las mejores medidas de singularidad de $X'X$ o de $\mathbf{R}$ ulilizan los autovalores de estas matrices. Un índice de singularidad, que se utiliza en cálculo numérico, es el **índice de condicionamiento** (*condition number*).

Si $\mathbf{M}$ es una matriz de orden k, simétrica y definida positiva, y $\lambda_1 < \lambda_2 < \ldots < \lambda_k$, son sus autovalores, se define el indice de condicionamiento de $\mathbf{M}$ como:

$$cond(\mathbf{M}) = \sqrt{\frac{\lambda_k}{\lambda_1}}.$$

Es claro que $cond(\mathbf{M}) \geq 1$. Es más conveniente calcular este indice para $\mathbf{R}$ que para $X'X$, con el fin de evitar la influencia de las escalas de medida de los regresores.

Para saber si existe o no multicolinealidad, calcularemos $cond(\mathbf{R})$ y si

- $cond(\mathbf{R}) > 30$, se tiene alta multicolinealidad.
- $10 < cond(\mathbf{R}) < 30$, se tiene multicolinealidad moderada.
- $cond(\mathbf{R}) < 10$, se tiene ausencia de multicolinealidad (la matriz $\mathbf{R}$ está bien definida).

Tratamiento de la multicolinealidad

Cuando la recogida de datos se diseñe a priori, la multicolinealidad puede evitarse tomando las observaciones de manera que la matriz $X'X$ sea diagonal, lo que aumentará la precisión en la estimación (los estimadores tendrán varianza pequeña).

La multicolinealidad es un problema de la muestra y, por tanto, no tiene solución simple, ya que estamos pidiendo a los datos más información de la que contienen.

Las dos únicas soluciones son

1) Eliminar regresores, reduciendo el número de parámetros a estimar.
2) Incluir información externa a los datos.

2.8.2. Análisis de los residuos

Los residuos se definen como:

$$e_i = y_i - \widehat{y_i}, \quad i = 1, \dots, n$$

Consideramos el vector de residuos:

$$\begin{aligned}\vec{e} = \vec{y} - \widehat{\vec{y}} = \vec{y} - X\widehat{\vec{\beta}} = \vec{y} - X(X'X)^{-1}X'\vec{y} = (I - X(X'X)^{-1}X')\vec{y} = \\ = (I-H)\vec{y} = (I-H)(X\vec{\beta} + \vec{u}) = X\vec{\beta} + \vec{u} - HX\vec{\beta} - H\vec{u} = \\ = X\vec{\beta} + \vec{u} - X\vec{\beta} - H\vec{u} = \vec{u} - H\vec{u} = (I-H)\vec{u}\end{aligned}$$

donde $H = X(X'X)^{-1}X'$ es una matriz simétrica e idempotente. Veamos esto último:

Demostración. $H^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = H$ □

Observación 2.6. $HX\vec{\beta} = X(X'X)^{-1}X'X\vec{\beta} = X\vec{\beta}$

Como $\vec{u} \sim N_n(\vec{0}, \sigma^2 I_n)$, entonces $\vec{e} \sim N_n(\vec{0}, (I-H)'\sigma^2 I(I-H))$. Obtengamos una expresión más simplificada usando las propiedades de H:

$$(I-H)'\sigma^2 I(I-H) = \sigma^2(I-H)^2 = \sigma^2(I-H)$$

Por tanto, $\vec{e} \sim N(\vec{0}, \sigma^2(I-H))$. Además, podemos ver que $e_i \sim N(0, \sigma^2(1-h_{ii}))$, donde h_{ii} es el elemento (i,i) de la matriz H. Este resultado es válido para la regresión lineal simple.

Se definen los residuos estandarizados como:

$$\boxed{r_i = \frac{e_i}{s_R\sqrt{1-h_{ii}}} \sim t_{n-k-1}}$$

Observación 2.7.

$$\begin{cases} \dfrac{e_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0,1) \\ \dfrac{(n-k-1)s_R^2}{\sigma^2} = \dfrac{\sum e_i^2}{\sigma^2} \sim \chi^2_{n-k-1} \end{cases} \Rightarrow \frac{\dfrac{e_i}{\sigma\sqrt{1-h_{ii}}}}{\sqrt{\dfrac{(n-k-1)s_R^2}{\sigma^2(n-k-1)}}} = \frac{e_i}{s_R\sqrt{1-h_{ii}}} \sim t_{n-k-1}$$

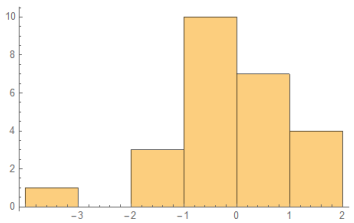
Se definen los residuos estudentizados como:

$$\boxed{t_i = \frac{e_i}{s_R(i)\sqrt{1-h_{ii}}} \sim t_{n-k-2}}$$

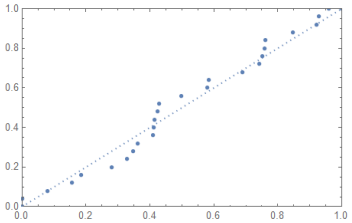
donde $s_R^2(i)$ es la varianza muestral de todos los datos excepto el i-ésimo.

Análisis gráfico de los residuos

1. **Histograma y gráfico probabilístico normal.** Sirve para detectar si hay normalidad y datos atípicos.

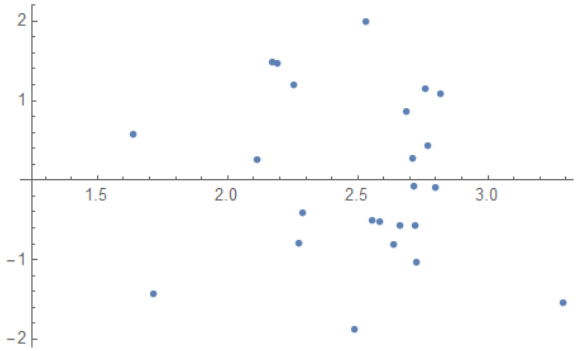


(a) Histograma.



(b) Gráfico probabilístico normal.

2. **Gráfico de residuos frente a los valores predichos.** Sirve para comprobar si hay linealidad, homocedasticidad y datos atípicos. Se representan los residuos t_i frente a los $\widehat{y_i}$.



3. **Gráficos de residuos frente a variables explicativas.** Detectan si hay linealidad, homocedasticidad y datos atípicos en cada variable. Se hacen k gráficos, cada uno representando los residuos t_i frente a cada variable X_{ji}, para $j = 1, \ldots, k$.

4. **Gráficos parciales de residuos.** Miden la influencia de cada X_i quitando todas las demás variables. Se hacen k gráficos con el siguiente procedimiento para cada X_i con $i = 1, \ldots, k$:

   a) Ajustamos el modelo con todas las variables explicativas salvo X_i.

   b) Calculamos los errores del ajuste anterior $t_j^{(i)}$ y los representamos frente a X_i.

5. **Gráfico de residuos frente a variables omitidas.** Sirve para comprobar si una variable omitida X_{k+1} debería ser tenida en cuenta en el modelo. Se representan los residuos frente a X_{k+1}. Una estructura lineal en esta gráfica indica que hay que tener en cuenta esta variable.

2.8.3. Observaciones atípicas e influyentes

La observación i-ésima es atípica a nivel de significación α si $|t_i| > t_{n-k-2,1-\frac{\alpha}{2}}$.

Una observación es influyente si se da alguno de estos casos:

- Modifica el vector $\widehat{\vec{\beta}}$ de parámetros estimado.
- Modifica el vector $\widehat{\vec{y}}$ de predicciones.
- Hace que la observación del punto sea muy buena cuando este se incluye en el modelo y mala cuando se excluye.

En general son puntos palanca.

Definimos la distancia de Cook de la observación i-ésima como:

$$\boxed{D(i) = \frac{(\widehat{\vec{\beta}} - \widehat{\vec{\beta}}_{(i)})' X' X (\widehat{\vec{\beta}} - \widehat{\vec{\beta}}_{(i)})}{(k+1)s_R^2}}$$

donde $\widehat{\vec{\beta}}_{(i)}$ es el vector de parámetros estimado sin la observación i-ésima.

Observación 2.8. Recordamos que:

$$\begin{cases} \dfrac{(\widehat{\vec{\beta}} - \vec{\beta})' X' X (\widehat{\vec{\beta}} - \vec{\beta})}{\sigma^2} \sim \chi^2_{k+1} \\ \dfrac{(n-k-1)s_R^2}{\sigma^2} = \dfrac{\sum e_i^2}{\sigma^2} \sim \chi^2_{n-k-1} \end{cases}$$

Entonces:

$$\frac{\frac{(\widehat{\vec{\beta}} - \vec{\beta})' X' X (\widehat{\vec{\beta}} - \vec{\beta})}{\sigma^2}/(k+1)}{\frac{(n-k-1)s_R^2}{\sigma^2}/(n-k-1)} = \frac{(\widehat{\vec{\beta}} - \vec{\beta})' X' X (\widehat{\vec{\beta}} - \vec{\beta})}{(k+1)s_R^2} \sim F_{k+1,n-k-1}$$

Usando esta distancia, podemos determinar que la observación i-ésima es influyente a nivel de significación α si:

Observación 2.9. Una distancia $D(i) > 1$ suele indicar que la observación es influyente.

2.9. Selección de modelos

Distinguimos dos tipos de medidas para la bondad del modelo:

1. Criterios basados en la bondas de ajuste:
   - **Coeficiente de determinación.** No sirve para comparar modelos en general, porque aquel que tenga más variables explicativas tiene un mayor R^2, incluso si no son significativas.
   - **Coeficiente de determinación ajustado.** Es mejor modelo el que tenga mayor $\overline{R}^2$.
   - **Varianza residual.** Es mejor modelo el que tenga menor s_R^2. Es equivalente al anterior criterio por la relación que hay entre $\overline{R}^2$ y s_R^2.
2. Criterios basados en buscar buenas predicciones:
   - **AIC (Akaike Information Criterion).** Es mejor modelo el que tenga menor AIC.
   - **BIC (Bayesian Information Criterion).** Es mejor modelo el que tenga menor BIC.

Si dos modelos tienen una bondad similar, siempre es preferible el más simple.

2.10. Regresión con variables cualitativas

Supongamos que tenemos dos poblaciones demográficas A y B y tenemos unas variables aleatorias $X_1, \ldots, X_k$. Sean $x_{1i}, \ldots, x_{ki}$ datos sobre dichas variables aleatorias que sabemos a que población corresponden. Podemos tener los siguientes modelos:

1. Mezclar todos los datos y hacer una regresión lineal sobre ellos, es decir, consideramos el siguiente modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + u_i,$$

y lo estimamos de la siguiente forma,

$$\widehat{y_i} = E(y_1|\widehat{x_{1i}, \ldots}, x_{ki}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \ldots + \widehat{\beta}_k x_{ki}.$$

Lo bueno de este modelo es que usamos todos los datos, pero estamos considerando que ambas poblaciones son homogéneas.

2. Ajustar cada población por serparado, es decir, hacer una regresión lineal para la población A y otra regresión lineal para la población B. Lo bueno de este modelo es que podemos hacer buenas estimaciones de ambas poblaciones de manera independiente, pero en cada modelo usamos menos datos.

3. Introducir una nueva variable, X_{k+1} (variable Dummy) tal que

$$x_{k+1,i} = \begin{cases} 1, \text{ si el dato } i\text{-ésimo pertenece a la población } A \\ 0, \text{ si el dato } i\text{-ésimo pertenece a la población } B \end{cases}.$$

Ahora, nuestros datos serían de la forma $(x_{1i}, x_{2i}, \ldots, x_{ki}, 1 \text{ ó } 0, y_i)$. Así, consideramos el modelo

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \ldots + \widehat{\beta}_k x_{ki} + \widehat{\beta}_{k+1} x_{k+1,i}$$

- Para obtener la regresión para la población A basta tomar $x_{k+1,i} = 1$,

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \ldots + \widehat{\beta}_k x_{ki} + \widehat{\beta}_{k+1}$$

- Para obtener la regresión para la población B basta tomar $x_{k+1,i} = 0$,

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \ldots + \widehat{\beta}_k x_{ki}$$

Para saber si ambas poblaciones son homogéneas, o equivalentemente, si la variable X_{k+1} es significativa, podemos plantear el siguiente contraste:

$$\begin{cases} H_0 : \beta_{k+1} = 0 \\ H_1 : \beta_{k+1} \neq 0 \end{cases}$$

Los modelos que hemos visto se llaman modelos anidados, debido a que cada uno contiene todos los términos del modelo anterior. Este último es mejor que los anteriores pero supone que el incremento de $\widehat{y}$ es igual para cada población, lo que no es cierto en general. Veremos en ejemplos que podemos mejorarlo añadiendo interacciones.

Ejemplo 2.10. Consideramos las variables Y (peso en kg) y X (altura en cm). Los datos $\{(x_i, y_i)\}$ provienen de dos poblaciones según el sexo: hombres y mujeres.

El modelo conjunto sería:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$$

Como el sexo influye en el peso de una persona, añadimos una variable ficticia, X_2, para mejorar el modelo tal que

$$x_{2i} = \begin{cases} 1, \text{ si el dato } i\text{-ésimo es hombre} \\ 0, \text{ si el dato } i\text{-ésimo es mujer.} \end{cases}$$

Codificamos los datos a la forma (x_{1i}, x_{2i}, y_i) para tener en cuenta estas nuevas variables. De esta forma, obtenemos el modelo general:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$$

Al término $\widehat{\beta}_2 x_2$ se le llama **efecto principal** del tipo de combustible.

- Para los hombres el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 = (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 x_1$$

- Para las mujeres el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$$

Podemos observar que ambas rectas tienen la misma pendiente, es decir, se supone que el incremento de los pesos es igual en cada población, lo que no es cierto en general. Para mejorar el modelo, introducimos un nuevo término $\widehat{\beta}_3 x_1 x_2$ llamado **interacción** de altura y sexo. Así que este nuevo modelo queda de la forma:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_1 x_2$$

- Para los hombres el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 + \widehat{\beta}_3 x_1 = (\widehat{\beta}_0 + \widehat{\beta}_2) + (\widehat{\beta}_1 + \widehat{\beta}_3) x_1$$

- Para las mujeres el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$$

Observamos que ahora las rectas tienen distinta pendiente.

Podemos realizar algunos contrastes de hipótesis para comprobar si este modelo es el correcto. Para determinar si el sexo tiene una influencia significativa en el peso, contrastamos:

$$\begin{cases} H_0 : \beta_2 = \beta_3 = 0 \\ H_1 : \beta_2 \neq 0 \text{ o } \beta_3 \neq 0 \end{cases}$$

Para comprobar si el incremente en el peso medio es igual para hombres y mujeres, podemos realizar el contraste:

$$\begin{cases} H_0 : \beta_3 = 0 \\ H_1 : \beta_3 \neq 0 \end{cases}$$

Ejemplo 2.11. Consideramos las variables Y (rendimiento de un motor diésel) y X (velocidad del motor). Existen tres tipos de combustible: petróleo, carbón y mezcla. Tenemos un conjunto de datos $\{(x_i, y_i)\}$ con los distintos tipos de combustible y queremos ajustar un modelo de regresión.

El modelo conjunto sería:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$$

Como es de esperar que el tipo de combustible influya en el rendimiento del motor, añadimos dos variables ficticias, X_2 y X_3 tales que:

$$x_{2i} = \begin{cases} 1, \text{ si el dato } i\text{-ésimo usa petróleo} \\ 0, \text{ si el dato } i\text{-ésimo no usa patróleo.} \end{cases}$$

$$x_{3i} = \begin{cases} 1, \text{ si el dato } i\text{-ésimo usa carbón} \\ 0, \text{ si el dato } i\text{-ésimo no usa carbón.} \end{cases}$$

Codificamos los datos a la forma $(x_{1i}, x_{2i}, x_{3i}, y_i)$ para tener en cuenta estas nuevas variables. De esta forma, obtenemos el modelo general:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3$$

Al término $\widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3$ se le llama **efecto principal** del tipo de combustible.

- Para el petróleo el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 = (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 x_1$$

- Para el carbón el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_3 = (\widehat{\beta}_0 + \widehat{\beta}_3) + \widehat{\beta}_1 x_1$$

- Para la mezcla el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$$

De nuevo, podemos observar que las tres rectas tienen la misma pendiente, es decir, se supone que el incremento del rendimiento del motor es igual para cada tipo de combustible, lo que no es cierto en general. Para corregirlo introducimos la **interacción** entre velocidad y tipo de combustible $\widehat{\beta}_4 x_1 x_2 + \widehat{\beta}_5 x_1 x_3$. Así, el nuevo modelo queda de la forma:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \widehat{\beta}_4 x_1 x_2 + \widehat{\beta}_5 x_1 x_3$$

- Para el petróleo el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 + \widehat{\beta}_4 x_1 = (\widehat{\beta}_0 + \widehat{\beta}_2) + (\widehat{\beta}_1 + \widehat{\beta}_4) x_1$$

- Para el carbón el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_3 + \widehat{\beta}_5 x_1 = (\widehat{\beta}_0 + \widehat{\beta}_3) + (\widehat{\beta}_1 + \widehat{\beta}_5) x_1$$

- Para la mezcla el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$$

Ahora las rectas tienen distinta pendiente, como queríamos.

También podemos realizar algunos contrastes de hipótesis para comprobar si este modelo es el correcto. Para determinar si el rendimiento medio del motor depende del tipo de combustible, contrastamos:

$$\begin{cases} H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : \beta_2 \neq 0 \text{ o } \beta_3 \neq 0 \text{ o } \beta_4 \neq 0 \text{ o } \beta_5 \neq 0 \end{cases}$$

Para comprobar si hay dependencia entre velocidad y tipo de combustible, podemos realizar el contraste:

$$\begin{cases} H_0 : \beta_4 = \beta_5 = 0 \\ H_1 : \beta_4 \neq 0 \text{ o } \beta_5 \neq 0 \end{cases}$$

Supongamos ahora que creemos que la relación entre el rendimiento medio de un motor diésel y la veloicidad es cuadrática. Entonces el modelo conjunto sería:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_1^2$$

Añadimos el efecto principal del tipo de combustible con las variables ficticias X_2 y X_3:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_1^2 + \widehat{\beta}_3 x_2 + \widehat{\beta}_4 x_3$$

- Para el petróleo el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_1^2 + \widehat{\beta}_3 = (\widehat{\beta}_0 + \widehat{\beta}_3) + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_1^2$$

- Para el carbón el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_1^2 + \widehat{\beta}_4 = (\widehat{\beta}_0 + \widehat{\beta}_4) + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_1^2$$

- Para la mezcla el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_1^2$$

Añadimos ahora la interacción entre la velocidad del motor y el tipo de combustible:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_1^2 + \widehat{\beta}_3 x_2 + \widehat{\beta}_4 x_3 + \widehat{\beta}_5 x_1 x_2 + \widehat{\beta}_6 x_1 x_3 + \widehat{\beta}_7 x_1^2 x_2 + \widehat{\beta}_8 x_1^2 x_3$$

- Para el petróleo el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_1^2 + \widehat{\beta}_3 + \widehat{\beta}_5 x_1 + \widehat{\beta}_7 x_1^2 = (\widehat{\beta}_0 + \widehat{\beta}_3) + (\widehat{\beta}_1 + \widehat{\beta}_5) x_1 + (\widehat{\beta}_2 + \widehat{\beta}_7) x_1^2$$

- Para el carbón el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_1^2 + \widehat{\beta}_4 + \widehat{\beta}_6 x_1 + \widehat{\beta}_8 x_1^2 = (\widehat{\beta}_0 + \widehat{\beta}_4) + (\widehat{\beta}_1 + \widehat{\beta}_6) x_1 + (\widehat{\beta}_2 + \widehat{\beta}_8) x_1^2$$

- Para la mezcla el modelo es:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_1^2$$

Para determinar si el modelo medio de un motor varía según el tipo de combustible, contrastamos:

$$\begin{cases} H_0 : \beta_i = 0, \quad \forall i \in \{3, \dots, 8\} \\ H_1 : \exists i \in \{3, \dots, 8\} : \beta_i \neq 0 \end{cases}$$

Para comprobar si un modelo de segundo orden es mejor a uno de primer orden, realizamos el contraste:

$$\begin{cases} H_0 : \beta_2 = \beta_7 = \beta_8 = 0 \\ H_1 : \exists i \in \{2, 7, 8\} : \beta_i \neq 0 \end{cases}$$

Capítulo 3

Modelos lineales generalizados

3.1. Introducción

El modelo lineal generalizado es una generalización de la regresión lineal que permite que lo y_i no sigan una distribución normal. Este modelo tiene tres componentes:

- **Componente aleatoria.** Viene dada por la variable Y. Las y_i pueden seguir varias distribuciones comunes, como la Bernoulli, Binomial, Binomial Negativa, Poisson y Gamma. Únicamente veremos el caso en el que $y_i \sim Ber(p)$.
- **Componente sistemática.** Viene dada por las variables $X_1, \ldots, X_k$, que están relacionadas mediante el predictor lineal $\vec{x}_i'\vec{\beta}$, siendo $\vec{x}_i' = (1, x_{1i}, \ldots, x_{ki})$.
- **Función enlace.** La función enlace proporciona la relación entre el predictor lineal y la media de la función de distribución.

  $$E(y_i|x_{1i}, \ldots, x_{ki}) = g_i(\vec{x}_i'\vec{\beta}) \Rightarrow \vec{x}_i'\vec{\beta} = g_i^{-1}(E(y_i|x_{1i}, \ldots, x_{ki}))$$

  La función g_i^{-1} es la **función enlace**.

  Observación 3.1. Si $g_i = g = Id$ para todo i, se corresponde con el modelo de regresión lineal múltiple.

3.2. Modelo de regresión con respuesta binaria

Los modelos de regresión con respuesta binaria son aquellos en los que $y_i \sim Ber(p)$. En estos casos tenemos datos que queremos clasificar en dos poblaciones A y B. El conocimiento de una serie de variables nos ayudará a determinar de qué población son.

Tenemos entonces un conjunto de datos $\{x_{1i}, \ldots, x_{ki}, y_i\}$, donde y_i es una variable dicotómica.

$$y_i = \begin{cases} 1 & \text{si el dato } i\text{-ésimo procede de } A \\ 0 & \text{si el dato } i\text{-ésimo procede de } B \end{cases}$$

Así que $y_i \sim Ber(p_i)$ con $p_i = P(Y_i = 1)$. Queremos estimar $\widehat{p_i} = \widehat{P(Y_i = 1)}$, es decir, la probabilidad de que el individuo i sea de la población A.

$$E(y_i|x_{1i}, \ldots, x_{ki}) = p_i = g_i(\vec{x}_i'\vec{\beta})$$
$$\widehat{E}(y_i|x_{1i}, \ldots, x_{ki}) = \widehat{p}_i$$

Sin embargo, este modelo tiene un problema y es que, queremos estimar p_i, que es una probabilidad, por tanto, $p_i \in [0, 1]$, y en principio, según este modelo, p_i podría tomar cualquier valor real. Pero este problema, se puede arreglar de la siguiente forma, tomar $p_i = F(\vec{x}_i'\vec{\beta})$, con F función de distribución. Usaremos dos funciones de distribución:

- **Función de distribución logística**.

$$F(x) = \frac{1}{1+e^{-x}}, \quad x \in \mathbb{R}$$

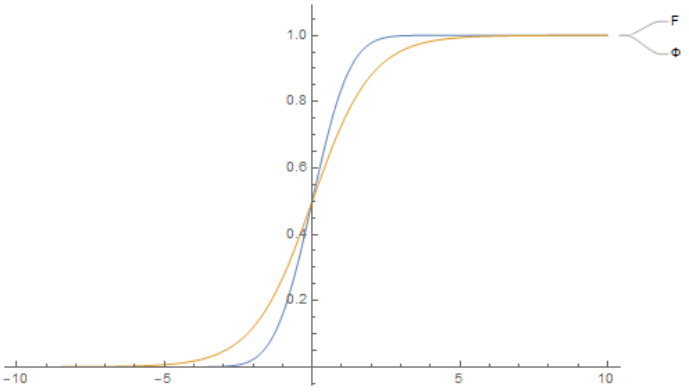
La función F^{-1} es la función enlace y se conoce como **función logit**. Podemos calcularla:

$$p_i = F(\vec{x}_i'\vec{\beta}) = \frac{1}{1+e^{-\vec{x}_i'\vec{\beta}}} \iff p_i + p_i e^{-\vec{x}_i'\vec{\beta}} = 1 \iff e^{\vec{x}_i'\vec{\beta}} = \frac{p_i}{1-p_i} \Leftrightarrow$$
$$\iff \vec{x}_i'\vec{\beta} = \log\left(\frac{p_i}{1-p_i}\right)$$

- **Función de distribución normal (estándar)**.

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$$

La función Φ^{-1} es la función enlace y se conoce como **función probit**.



3.3. Riesgo, oportunidad, riesgo relativo y odds ratio

Iluestremos estos conceptos con un ejemplo.

Ejemplo 3.2. En 1 de cada 200 nacimientos ocurre un parto gemelar. Entonces la probabilidad o *riesgo* de que un embarazo elegido al azar de un parto gemelar es $R_1 = \frac{1}{200}$. Hay un forma de expresar lo mismo en términos de apuestas, de 200 partos, 1 es gemelar y 199 no lo son, es decir $O_1 = \frac{1}{199}$. Nótese que

$$O_1 = \frac{R_1}{1-R_1}.$$

O_1 es la *oportunidad*. Se observó que entre 100 muejeres que habían tomado ácido fólico, 3 de cada 200 partos eran gemelar. Ahora,

$$R_2 = \frac{3}{200}, \quad O_2 = \frac{3}{197} = \frac{R_2}{1-R_2}.$$

El aumento del riesgo del embarazo gemeral se puede expresar numéricamente como

$$RR := \frac{R_2}{R_1} = 3, \quad OR := \frac{O_2}{O_1}\frac{199 \cdot 3}{197} \approx 3{,}03.$$

Definición 3.3.

- El riesgo es la probabilidad de que ocurra un resultado.
- La oportunidad es el cociente del número de eventos que producen un resultado entre el número de eventos que no lo producen.
- El riesgo relativo es el cociente de los riesgos de dos grupos de población.

$$RR = \frac{R_1}{R_2}.$$

- La razón de oportunidades es el cociente de las oportunidades de dos grupos de población.

$$OR = \frac{O_1}{O_2}.$$

Observación 3.4. Existe una relación entre el riesgo y la oportunidad:

$$O = \frac{R}{1-R}.$$

3.4. Regresión logística

El modelo de regresión logística es:

$$E(y_i|x_{1i},\dots,x_{ki}) = p_i = F(\vec{x}_i'\vec{\beta}) = \frac{1}{1+e^{-\vec{x}_i'\vec{\beta}}}$$

Además, hemos visto que:

$$\vec{x}_i'\vec{\beta} = \log\left(\frac{p_i}{1-p_i}\right)$$

Luego queremos estimar:

$$\widehat{p_i} = \frac{1}{1+e^{-\vec{x}_i'\widehat{\vec{\beta}}}}$$

Para encontrar los estimadores $\widehat{\beta_i}$ usamos el método de máxima verosimilitud. Calculamos la función de verosimilitud:

$$L(\beta_0,\dots,\beta_k) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

Tomamos logaritmos en ambos miembros de la igualdad:

$$\begin{aligned}\log L(\beta_0,\dots,\beta_k) &= \sum_{i=1}^{n} y_i\log(p_i) + \sum_{i=1}^{n}(1-y_i)\log(1-p_i)\\ &= \sum_{i=1}^{n} y_i\log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n}\log(1-p_i) = \sum_{i=1}^{n} y_i\vec{x}_i'\vec{\beta} + \sum_{i=1}^{n}\log\left(\frac{e^{-\vec{x}_i'\vec{\beta}}}{1+e^{-\vec{x}_i'\vec{\beta}}}\right)\\ &= \sum_{i=1}^{n} y_i\vec{x}_i'\vec{\beta} + \sum_{i=1}^{n}\log\left(\frac{1}{e^{\vec{x}_i'\vec{\beta}}+1}\right) = \sum_{i=1}^{n} y_i\vec{x}_i'\vec{\beta} - \sum_{i=1}^{n}\log(1+e^{\vec{x}_i'\vec{\beta}})\end{aligned}$$

Así que, derivando tenemos que:

$$\frac{\partial \log L}{\partial \vec{\beta}} = \sum_{i=1}^{n} y_i\vec{x}_i - \sum_{i=1}^{n}\frac{\vec{x}_i e^{\vec{x}_i'\vec{\beta}}}{1+e^{\vec{x}_i'\vec{\beta}}}$$

Para hallar los $\widehat{\beta}_i$ hay que resolver el sistema $\frac{\partial \log L}{\partial \beta_i} = 0$ para todo $i = 0, \ldots, k$ numéricamente.

Podemos darles significado a los β_j. Para ello calculamos la oportunidad:

$$O(x_{1i}, \ldots, x_{ki}) = \frac{p_i}{1-p_i} = \frac{P(Y_i = 1)}{P(Y_i = 0)} = \frac{\frac{1}{1+e^{-\vec{x}_i'\vec{\beta}}}}{\frac{e^{-\vec{x}_i'\vec{\beta}}}{1+e^{-\vec{x}_i'\vec{\beta}}}} = e^{\vec{x}_i'\vec{\beta}} = e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_k x_{ki}}$$

Calculamos ahora la razón de oportunidades cuando aumenta x_{ji} en una unidad:

$$OR_j = \frac{O(x_{1i}, \ldots, x_{ji}+1, \ldots, x_{ki})}{O(x_{1i}, \ldots, x_{ji}, \ldots, x_{ki})} = e^{\beta_j}$$

Así que e^{β_j} es lo que varía la oportunidad cuando aumenta la componente j-ésima en una unidad. Luego OR_j determina si las variable j-ésima es significativa.

Parte II

Inferencia Bayesiana

Capítulo 4

Inferencia Bayesiana

4.1. Introducción

La estadistica Bayesiana le debe su nombre al trabajo pionero del reverendo Thomas Bayes titulado "*An Essay towards solving a Problem in the Doctrine of Chances*" publicado póstumamente en 1764 en la "*Philosophical Transactions of the Royal Society of London*". El artículo fue enviado a la Real Sociedad de Londres por Richard Price, amigo de Bayes, en 1763, quién escribió:

> "Yo ahora le mando un ensayo que he encontrado entre los papeles de nuestro fallecido amigo Thomas Bayes, y el cual, en mi opinión, tiene un gran mérito, y bien merece ser preservado ... En una introducción que él ha escrito para este ensayo, él dice, que su objetivo en un principio fue, descubrir un método por el cual se pueda juzgar la probabilidad de que un evento tenga que ocurrir bajo circunstancias dadas, y bajo la suposición de que nada es conocido sobre dicho evento, salvo que, bajo las mismas circunstancias, éste ha ocurrido un cierto número de veces y fallado otro tanto ... Cualquier persona juiciosa verá que el problema aquí mencionado no es de ninguna manera una simple especulación producto de la curiosidad, sino un problema que se necesita resolver para contar con un fundamento seguro para todos nuestros razonamientos concernientes a hechos pasados y a lo que probablemente ocurra de ahí en adelante ... El propósito a mí me parece es, mostrar qué razones nosotros tenemos para creer que en la constitución de las cosas existen leyes fijas de acuerdo con las cuales las cosas pasan, y que, por lo tanto, el funcionamiento del mundo debe ser el efecto de la sabiduria y el poder de una causa inteligente, y así, confirmar el argumento tomado desde las causas finales para la existencia de la deidad."

Aunque la obra de Thomas Bayes data ya de hace más de dos siglos, la estadistica Bayesiana es relativamente nueva, y actualmente ostenta un gran desarrollo aunque no ajeno a también grandes controversias. El marco teórico en el cual se desarrolla la inferencia Bayesiana es idéntico al de la teoría clásica. Se tiene un parámetro poblacional θ sobre el cual se desen hacer inferencias y se tiene un modelo de probabilidad $f(x/\theta)$ el cual determina la probabilidad de los datos observados r bajo diferentes valores de θ. La diferencia fundamental entre la teoría clásica y la bayesiana está en que θ es tratado como una cantidad aleatoria. Asi, la inferencia Bayesiana se basa en $f(\theta/x)$ en vez de $f(x/\theta)$, esto es, en la distribución de probabilidades del parámetro dados los datos.

La inferencia Bayesiana, se puede resumir como el proceso de ajustar un modelo de probabilidad a un conjunto de datos y resumir los resultados me diante una distribución de probabilidades para los parámetros del modelo y para cantidades desconocidas pero observables tales como predicciones para nuevas observaciones. La caracteristica esencial de los métodos Bayesianos está en su uso explícito de probabilidades para cuantificar la incertidumbre en inferencias basadas en el análisis estadístico de los datos. Esto permite un manejo mucho más natural e intuitivo de la inferencia, salvando por ejemplo el problema de la interpretación frecuencial de los resultados. Sin embargo, para hacer uso de un enfoque Bayesiano, es necesario especificar una distribución de probabilidades a priori $f(\theta)$, la cual representa el conocimiento que se tiene sobre la distribución de θ previo a

la obtención de los datos. Esta noción de una distribución a priori para el parámetro constituye el centro del pensamiento Bayesiano y, dependiendo de si se es un defensor o un opositor a esta metodología, su principal ventaja sobre la teoría clásica o su mayor vulnerabilidad.

4.2. Teorema de Bayes

Teorema 4.1 (de Bayes)**.** *Sea $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad y sea $A_1, \ldots, A_n \in \mathcal{A}$ una partición de Ω, es decir, $\Omega = \cup_{i=1}^n A_i$ y $A_i \cap A_j = \emptyset$ para todo $i \neq j$ y tales que $P(A_i) > 0$ para todo i. Sea $B \in \mathcal{A}$ tal que $P(B) > 0$ y tal que $P(B|A_i)$ son conocidas para todo i. Entonces*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)}, \quad i = 1, \ldots, n.$$

Esta fórmula de conoce como la fórmula de Bayes y

- $P(A_j)$ son las **probabilidades a priori**,
- $P(B|A_j)$ son las **verosimilitudes**,
- $P(A_j|B)$ son las **probabilidades a posteriori**,

para todo $j = 1 \ldots, n$.

Ejemplo 4.2. Supogamos que tenemos una caja con una moneda legal, M_1, y otra moneda con una cara en cada lado, M_2.

a) Se selecciona una moneda al azar, se lanza y se obtiene cara, ¿qué probabilidad hay de que la moneda elegida sea M_1? Sea $C =$"obtener cara". Tenemos

Probs a priori	Verosimilitudes
$P(M_1) = \frac{1}{2}$	$P(C\|M_1) = \frac{1}{2}$
$P(M_2) = \frac{1}{2}$	$P(C\|M_2) = 1$

Por el Teorema de Bayes

$$P(M_1|C) = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2} \cdot \frac{1}{2} + 1 + \frac{1}{2}} = \frac{1}{3},$$
$$P(M_2|C) = 1 - P(M_1|C) = \frac{2}{3}.$$

b) Lanzamos de nuevo la moneda elegida y se obtiene otra cara, ¿qué probabilidad hay de que la moneda elegida sea M_1? Utilizado el carácter secuencial, tenemos que las probabilidades a posterior de antes pasan a ser nuestras nuevas probabilidades a priori, es decir,

Probs a priori	Verosimilitudes
$P(M_1\|C_1) = \frac{1}{3}$	$P(C_2\|M_1) = \frac{1}{2}$
$P(M_2\|C_1) = \frac{2}{3}$	$P(C_2\|M_2) = 1$

siendo $C_i =$ "obtener cara en el i-ésimo lanzamiento", $i = 1, 2$. Por el Teorema de Bayes

$$P(M_1|C_1 \cap C_2) = \frac{\frac{1}{3} \cdot \frac{1}{2}}{\frac{1}{3} \cdot \frac{1}{2} + 1 + \frac{2}{3}} = \frac{1}{5},$$
$$P(M_2|C_1 \cap C_2) = 1 - P(M_1|C_1 \cap C_2) = \frac{4}{5}.$$

4.3. Teorema de Bayes generalizado

Queremos hacer inferencia sobre un parámetro θ y tenemos una muestra aleatoria simple $x_1, \ldots, x_n$ que sabemos de la población que procede. En la inferencia Bayesiana θ es una variable aleatoria. Denotamos por f_θ a la **distribución a priori** de θ. Al igual que en la inferencia clásica, conocemos la **función de verosimilitud** $L(\vec{x};\theta)$. Mediante el Teorema de Bayes, calculamos $f(\theta|\vec{x})$, que se conoce como **distribución a posteriori** de θ. Así

$$f(\theta|\vec{x}) = \frac{f(\vec{x}|\theta)f_\theta(\theta)}{f(\vec{x})}. \quad \text{siendo} \quad f(\vec{x}) = \begin{cases} \sum_i f(x_i|\theta_i)f_{\theta_i}(\theta_i), & \theta \text{ es discreta,} \\ \int_\Theta f(\vec{x}|\theta)f_\theta(\theta)\ d\theta, & \theta \text{ es continua.} \end{cases}$$

Ejemplo 4.3. Supongamos que tenemos una moneda y queremos estimar la probabilidad de obtener una cara, que denotaremos como p. Supongamos que nuestras creencias a priori sobre p se pueden describir con una distribución $U(0,1)$, es decir, todos los valores entre 0 y 1 que puede tomar p son igualmente probables. Realizamos el experimento de tirar la moneda 12 veces y obtenemos 9 caras y 3 cruces. Determinar la distribución a posteriori de p.

Tenemos que

$$x|p \sim Ber(p), \quad x = \begin{cases} 1 & \text{si "sale cara",} \\ 0 & \text{si "sale cruz".} \end{cases}$$

Es claro que $P(\text{Obtener Cara}) = P(x = 1) = p$. Nos dicen que la distribución a priori de p sigue una $U(0,1)$, así que $f_p(p) = 1$, si $p \in (0,1)$ (y 0 en cualquier otro caso). Sea $x_1, ..., x_n$ una muestra aleatoria simple de una x. Entonces la función de verosimilitud es

$$L(\vec{x};p) = \prod_{i=1}^{12} f(x_i|p) = \prod_{i=1}^{12} p^{x_i}(1-p)^{1-x_i} p^{\sum_{i=1}^{12} x_i}(1-p)^{12-\sum_{i=1}^{12} x_i} = p^9(1-p)^3.$$

Nótese que la útlima igualdad se da porque de los 12 lanzamientos, 9 son caras, por tanto $\sum_{i=1}^{12} x_i = 9$. Finalmente, la distribución a posteriori es

$$f(p|\vec{x}) \propto f_p(p)L(\vec{x};p) = p^9(1-p)^3, \quad 0 < p < 1,$$

de donde concluimos que $p|\vec{x} \sim Be(10,4)$.

4.4. Familias de distribuciones conjugadas

Son aquellas en las que las distribuciones a priori y a posteriores son de la misma familia.

Familia conjugada de la Bernoulli

Consideremos $x|\theta \sim Ber(\theta)$ y supongamos que $\theta \sim Be(p,q)$. Sea $\vec{x} = (x_1, \ldots, x_n)$ nuestro vector de datos. La función de distribución a priori de θ es $f_\theta(\theta) \propto \theta^{p-1}(1-\theta)^{q-1}$, $0 < \theta < 1$. La función de verosimilitud es

$$L(\vec{x};\theta) = \prod_{i=1}^{n} f(x_i|\theta) \propto \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Finalmente, la función de dsitribución a posteriori es

$$f(\theta|\vec{x}) \propto f_\theta(\theta)L(\vec{x};\theta) \propto \theta^{p+\sum_{i=1}^n x_i - 1}(1-\theta)^{n+q-\sum_{i=1}^n x_i - 1}.$$

Así, hemos llegado a que $\theta|\vec{x} \sim Be\left(p + \sum_{i=1}^n x_i, n + q - \sum_{i=1}^n x_i\right)$ y por tanto la distribución Beta es una familia conjugada respecto de muestras de la Bernoulli.

Familia conjugada de la Poisson

Consideremos $x|\lambda \sim Po(\lambda)$ y supongamos que $\lambda \sim Ga(a,p)$. Sea $\vec{x} = (x_1, \ldots, x_n)$ nuestro vector de datos. La función de distribución a priori de λ es

$$f_\lambda(\lambda) = \frac{a^p}{\Gamma(p)} e^{-a\lambda} \lambda^{p-1}, \quad \lambda > 0, \quad a, p > 0.$$

La función de verosimilitud es

$$L(\vec{x}; \lambda) = \prod_{i=1}^{n} f(x_i|\lambda) \propto \prod_{i=1}^{n} e^{-\lambda} \lambda^{x_i} = e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}.$$

Finalmennte, la función de distribución a posteriori es

$$f(\lambda|\vec{x}) \propto f_\lambda(\lambda) L(\vec{x}; \lambda) \propto e^{-(a+n)\lambda} \lambda^{\sum_{i=1}^n x_i + p - 1}.$$

Así, hemos llegado a que $\lambda|\vec{x} \sim Ga\left(a+n, \sum_{i=1}^n x_i + p\right)$ y por tanto la Gamma es una familia conjugada respecto de muestras de la Poisson.

Familia conjugada de la Normal (media desconocida)

Consideremos $x|\mu \sim N(\mu, p)$, donde $p = 1/\sigma^2$ es la precisión, y supongamos que $\mu \sim N(m_0, p_0)$. Sea $\vec{x} = (x_1, \ldots, x_n)$ nuestro vector de datos. La función de densidad de $x|\mu$ es

$$f(x|\mu) = \frac{\sqrt{p}}{\sqrt{2\pi}} e^{-\frac{p}{2}(x-\mu)^2}, \quad x \in \mathbb{R}, \quad \mu \in \mathbb{R}.$$

La función de verosimilitud es entonces

$$L(\vec{x}; \mu) = \prod_{i=1}^{n} f(x_i|\mu) \propto \prod_{i=1}^{n} e^{-\frac{p}{2}(x-\mu)^2} \propto e^{-\frac{p}{2}\sum_{i=1}^n (x-\mu)^2}$$

Observación 4.4.

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \overline{x} + \overline{x} - \mu)^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 n(\overline{x} - \mu)^2 + 2(\overline{x} - \mu) \sum_{i=1}^{n} (x_i - \overline{x}),$$

pero este último término sabemos que es 0, por tanto

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 n(\overline{x} - \mu)^2 = (n-1)s^2 + n(\overline{x} - \mu)^2,$$

siendo

$$s^2 = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1}.$$

Usando esto, llegamos que

$$L(\vec{x}; \mu) \propto e^{-\frac{p}{2}[(n-1)s^2 + (\overline{x}-\mu)^2]} \propto e^{-\frac{p}{2} n(\overline{x}-\mu)^2}.$$

La función de distribución a priori es

$$f_\mu(\mu) = \frac{\sqrt{p_0}}{\sqrt{2\pi}} e^{-\frac{p_0}{2}(\mu - m_0)^2} \propto e^{-\frac{p_0}{2}(\mu - m_0)^2}.$$

Para calcular la función de distribución a posterior vamos a usar el siguiente Lema.

Lema 4.5.

$$A(z-a)^2 + B(z-b)^2 = (A+B)\left(z - \frac{Aa+Bb}{A+B}\right)^2 + \frac{AB}{A+B}(a-b)^2$$

La función de distribución a posterior es

$$\begin{aligned} f(\mu|\vec{x}) &\propto f_\lambda(\lambda) L(\vec{x}|\mu) \propto e^{-\frac{p_0}{2}(\mu-m_0)} e^{-\frac{p}{2}n(\overline{x}-\mu)^2} = e^{-\frac{1}{2}[p_0(\mu-m_0)^2 + np(\overline{x}-\mu)^2]} \\ &= e^{-\frac{1}{2}[a(\mu-b)^2]+c} \propto e^{\frac{-a}{2}(\mu-b)^2}, \end{aligned}$$

que por el Lema anterior, sabemos que

$$a = p_0 + np, \quad b = \frac{p_0 m_0 + np\overline{x}}{p_0 + np}, \quad c = \frac{p_0 np}{p_0 + np}(m_0 - x)^2.$$

Por tanto,

$$\mu|\vec{x} \sim N\left(\frac{p_0 m_0 + np\overline{x}}{p_0 + np}, p_0 + np\right), \quad p_0 + np \text{ es la precisón,}$$

y por tanto, la Normal es una familia conjugada respecto de muestras de la Normal con media desconocida y precisión conocida.

Familia conjugada de la Normal (precisión desconocida)

Consideremos $x|\tau \sim N(\mu, \tau)$, donde $\tau = 1/\sigma^2$ es la precisión, y supongamos que $\tau \sim Ga(a_0, p_0)$. Sea $\vec{x} = (x_1, \ldots, x_n)$ nuestro vector de datos. Tenemos que

$$\begin{aligned} f(x|\tau) &= \frac{\sqrt{\tau}}{\sqrt{2\pi}} e^{-\frac{\tau}{2}(x-\mu)^2}, \quad x \in \mathbb{R}, \tau > 0, \\ f_\tau(\tau) &= \frac{a_0^{p_0}}{\Gamma(p_0)} e^{-a_0\tau} \tau^{p_0-1}. \end{aligned}$$

La función de verosimilitud es entonces

$$L(\vec{x};\tau) = \prod_{i=1}^n f(x_i|\tau) \propto \prod_{i=1}^n \sqrt{\tau} \exp\left[-\frac{\tau}{2}(x_i-\mu)\right] = \tau^{\frac{n}{2}} \exp\left[-\frac{\tau}{2}\sum_{i=1}^n (x_i-\mu)^2\right].$$

La función de distribución a posteriori es

$$f(\tau|\vec{x}) \propto f_\tau(\tau) L(\vec{x}|\tau) = \tau^{\frac{n}{2}+p_0-1} \exp\left[-\tau\left(a_0 + \frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2\right)\right].$$

Así, hemos llegado a que $\tau|\vec{x} \sim Ga(a_n, p_n)$, siendo

$$a_n = a_0 + \frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2, \quad p_n = \frac{n}{2} + p_0.$$

Por tanto, la Gamma es una familia conjugada respecto de muestras de la Normal con media conocida y precisión desconocida.

Definición 4.6. Si $X \sim Ga(a,p)$ y definimos $Y = 1/X$, entonces $Y \sim GaI(a,p)$.

Calculemos la función de densidad de $Y \sim GaI(a,p)$. Sabemos que la densidad de $X \sim Ga(a,p)$ es

$$f_X(x) = \frac{a^p}{\Gamma(p)} e^{-ax} x^{p-1}, \quad x > 0, a, p > 0.$$

Tenemos que $Y = 1/X$, es decir, $X = 1/Y$. Aplicando el Teorema de cambio de variable, tenemos que la densidad de Y es

$$f_Y(y) = f_X\left(\frac{1}{y}\right)\left|\frac{d}{dy}\left(\frac{1}{y}\right)\right| = \frac{a^p}{\Gamma(p)} e^{-a/y}\left(\frac{1}{y}\right)^{p-1}\frac{1}{y^2} = \frac{a^p}{\Gamma(p)} e^{-a/y} y^{-(p+1)}, \quad y > 0.$$

Familia conjugada de la Normal (varianza desconocida)

Consideremos $x|\sigma^2 \sim N(\mu, \sigma^2)$ y supongamos que $\sigma^2 \sim GaI(a_0, p_0)$. Sea $\vec{x} = (x_1, \ldots, x_n)$ nuestro vector de datos. Tenemos que

$$f(x|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad x \in \mathbb{R}, \sigma^2 > 0,$$
$$f_{\sigma^2}(\sigma^2) = \frac{a_0^{p_0}}{\Gamma(p_0)} \left(\sigma^2\right)^{-(p_0+1)}, \quad a_0, p_0 > 0.$$

La función de verosimilitud es entonces

$$L(\vec{x}; \sigma^2) = \prod_{i=1}^{n} f(x_i|\sigma^2) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} = \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right].$$

La función de distribución a posteriori es

$$f(\sigma^2|\vec{x}) \propto f_{\sigma^2}(\sigma^2) L(\vec{x}|\sigma^2) \propto \left(\sigma^2\right) - (p_0 + n/2 + 1) \exp\left[-\frac{1}{\sigma^2}\left(a_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\right].$$

Así, hemos llegado a que $\sigma^2|\vec{x} \sim GaI(a_n, p_n)$, siendo

$$a_n = a_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2, \quad p_n = p_0 + \frac{n}{2}.$$

Por tanto, la Gamma Inversa es una familia conjugada respecto de muestras de la Normal con media conocida y varianza desconocida.

Definición 4.7. Decimos que $(\mu, \tau) \sim NGa(m_0, \tau_0, a_0, p_0)$ (Normal-Gamma) siendo $m_0 \in \mathbb{R}$, $\tau_0, a_0, p_0 > 0$ si

$$\mu|\tau \sim N(m_0, \tau\tau_0), \quad \mu \in \mathbb{R},$$
$$\tau \sim Ga(a_0, p_0), \quad \tau > 0.$$

Calculemos la función de densidad de $(\mu, \tau) \sim NGa(m_0, \tau_0, a_0, p_0)$. Tenemos que

$$f(\mu|\tau) = \frac{\sqrt{\tau\tau_0}}{\sqrt{2\pi}} e^{-\frac{\tau\tau_0}{2}(\mu-m_0)^2}, \quad \mu \in \mathbb{R},$$
$$f_\tau(\tau) = \frac{a_0^{p_0}}{\Gamma(p_0)} e^{-a_0\tau} \tau^{p_0-1}, \quad \tau > 0.$$

Así,

$$f(\mu, \tau) = f(\mu|\tau) f_\tau(\tau) = \frac{\sqrt{\tau_0}}{\sqrt{2\pi}} \frac{a_0^{p_0}}{\Gamma(p_0)} \tau^{p_0-1/2} e^{-\tau\left(a_0 + \frac{\tau_0}{2}(\mu-m_0)^2\right)}.$$

Podemos darle una vuelta a la definición anterior de la siguiente manera

Definición 4.8. Decimos que $(\mu, \tau) \sim NGa(m_0, \tau_0, a_0, p_0)$ (Normal-Gamma) si

$$\frac{\sqrt{\tau\tau_0}}{\sqrt{2\pi}} \frac{a_0^{p_0}}{\Gamma(p_0)} \tau^{p_0-1/2} e^{-\tau\left(a_0 + \frac{\tau_0}{2}(\mu-m_0)^2\right)},$$

siendo $\mu \in \mathbb{R}$, $\tau > 0$, $m_0 \in \mathbb{R}$, $\tau_0, a_0, p_0 > 0$.

Recordemos que la función de densidad de la t-Student con n grados de libertad es

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)\sqrt{n}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad t \in \mathbb{R}.$$

Además,

$$E[T] = 0, \quad \text{si } n > 1, \quad Var[T] = \frac{n}{n-2}, \quad \text{si } n > 2.$$

Ahora vamos a generalizar la distribución t-Student, permitiendo desplazarla (cambiar la media) y aplastarla (cambiar la "varianza")

Definición 4.9. Si $T \sim t_n$, entonces $X = \mu + \frac{1}{\sqrt{p}}T \sim t(\mu, p, n)$.

Tenemos que

$$E[X] = E\left[\mu + \frac{1}{\sqrt{p}}T\right] = \mu + \frac{1}{\sqrt{p}}E[T] = \mu,$$
$$Var[X] = Var\left[\mu + \frac{1}{\sqrt{p}}T\right] = \frac{1}{p}Var[T] = \frac{1}{p}\cdot\frac{n}{n-2}.$$

Llamaremos *precisión* de X a

$$\text{precisión}(X) = \frac{1}{p}\cdot\frac{n}{n-2}.$$

Calculemos ahora la función de densidad de X. Tenemos que

$$X = \mu + \frac{1}{\sqrt{p}}T \iff T = \sqrt{p}(X-\mu).$$

Aplicando el Teorema de cambio de variable

$$f_X(x) = f_T\left(\sqrt{p}(x-\mu)\right)\left|\frac{d}{dx}\left(\sqrt{p}(x-\mu)\right)\right| = \frac{\Gamma\left(\frac{n+1}{2}\right)\sqrt{p}}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)\sqrt{n}}\left(1+\frac{p(x-\mu)^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in \mathbb{R}.$$

Observación 4.10.

- t_1 es la conocida distribución de Cauchy.
- $t_n \xrightarrow{n\to\infty} N(0,1)$.

Teorema 4.11. *Si* $(\mu, \tau) \sim NGa(m_0, \tau_0, a_0, p_0)$, *entonces*

$$\mu \sim t\left(m_0, \frac{\tau_0 p_0}{2}, 2p_0\right).$$

Observación 4.12. Si tenemos que $\mu|\tau \sim N(0,\tau)$ y $\tau \sim Ga\left(\frac{n}{2}, \frac{n}{2}\right)$, entonces $\mu \sim t(0,1,n) \equiv t_n$ (Génesis Bayesiana de la t-Student).

Además, si $\mu|\sigma^2 \sim N(\mu, \sigma^2)$ y $\sigma^2 \sim GaI\left(\frac{n}{2}, \frac{n}{2}\right)$, entonces $\mu \sim t_n$.

Familia conjugada de la Normal (media y precisión desconocidas)

Consideremos $x|\mu,\tau \sim N(\mu,\sigma^2)$, siendo $\tau = 1/\sigma^2$ es la precisión y supongamos que $(\mu,\tau) \sim NGa(m_0,\tau_0,a_0,p_0)$. Sea $\vec{x} = (x_1,\ldots,x_n)$ nuestro vector de datos. Tenemos que

$$f(\mu,\tau) = \frac{\sqrt{\tau_0}}{\sqrt{2\pi}}\frac{a_0^{p_0}}{\Gamma(p_0)}\tau^{p_0-1/2}e^{-\tau\left(a_0+\frac{\tau_0}{2}(\mu-m_0)^2\right)},$$
$$f(x|\mu,\tau) = \frac{\tau}{\sqrt{2\pi}}e^{-\frac{\tau}{2}(x-\mu)^2}, \quad x\in\mathbb{R}, \mu\in\mathbb{R}, \tau\geq 0.$$

La función de verosimilitud es entonces

$$L(\vec{x};\mu,\tau) = \prod_{i=1}^n f(x_i|\mu,\tau) \propto \tau^{\frac{n}{2}}\exp\left[-\frac{\tau}{2}\sum_{i=1}^n(x_i-\mu)^2\right] = \tau^{\frac{n}{2}}e^{(n-1)s^2+n(\overline{x}-\mu)^2},$$

donde

$$s^2 = \frac{\sum_{i=1}^n(x_i-\overline{x})^2}{n-1}.$$

La fución de distribución a posteriori es

$$f(\mu,\tau|\vec{x}) \propto f(\mu,\tau)L(\vec{x};\mu,\tau) \propto \ldots \propto \tau^{n/2-p_0-1/2}e^{-\frac{\tau}{2}(2a_0+(n-1)s^2+a(\mu-b)^2)+c)},$$

siendo

$$a = \tau_0+n, \quad b = \frac{\tau_0 m_0+n\overline{x}}{\tau_0+n}, \quad c = \frac{\tau_0 n}{\tau_0+n}(\overline{x}-m_0)^2.$$

Así, hemos llegado a que $(\mu,\tau)|\vec{x} \sim NGa\left(b,a,\frac{2a_0+(n-1)s^2+c}{2},\frac{n}{2}+p_0\right)$, es decir, $(\mu,\tau)|\vec{x} \sim NGa(m_n,\tau_n,a_n,p_n)$, siendo

$$m_n = \frac{\tau_0 m_0+n\overline{x}}{\tau_0+n}, \quad \tau_n = \tau_0+n, \quad a_n = \frac{2a_0+(n-1)s^2+\frac{\tau_0 n}{\tau_0+n}(\overline{x}-m_0)^2}{2}, \quad p_n = \frac{n}{2}+p_0.$$

Por tanto, la Normal-Gamma es una familia conjugada respecto de muestras de la Normal con media y precisión desconocidas.

4.5. Distribuciones a priori no informativas

Se deducen de la regla de Jeffreys:

$$f_\theta(\theta) \propto \sqrt{J(\theta)}, \quad J(\theta) = -E\left[\frac{\partial^2 \log f(x|\theta)}{\partial\theta^2}\right].$$

$J(\theta)$ se conoce como *información de Fisher*. Es importante saber que la regla de Jeffreys no siempre proporiciona funciones de densidad, en realidad, proporciona densidades impropias.

Observación 4.13. En la asignatura de Inferencia Estadística vimos que si $\widehat{\theta}$ es un estimador insesgado para θ, entonces

$$Var[\widehat{\theta}] = \frac{1}{J(\widehat{\theta})},$$

que se conoce como la Cota de Frechet-Cramer-Rao.

Distribución no informativa de la Bernoulli

Supongamos que $x|\theta \sim Ber(\theta)$. Entonces

$$\begin{aligned} f(x|\theta) &= \theta^x(1-\theta)^{1-x}, \quad x = 0, 1, \quad 0 < \theta < 1, \\ \log f(x|\theta) &= x \log \theta + (1-x)\log(1-\theta), \\ \frac{\partial \log f(x|\theta)}{\partial \theta} &= \frac{x}{\theta} - \frac{1-x}{1-\theta}, \\ \frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} &= -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}. \end{aligned}$$

Entonces, la información de Fisher es

$$\begin{aligned} J(\theta) &= E\left[\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}\right] = \frac{1}{\theta}E[x] + \frac{1}{(1-\theta)^2}E[1-x] = \frac{1}{\theta^2}\theta + \frac{1}{(1-\theta)^2}(1-\theta) \\ &= \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)} \end{aligned}$$

Así, la regla de Jeffreys nos dice que

$$f_\theta(\theta) \propto \sqrt{J(\theta)} = \sqrt{\frac{1}{\theta(1-\theta)}} = \theta^{-1/2}(1-\theta)^{-1/2}.$$

De donde deducimos que $\theta \sim Be\left(\frac{1}{2}, \frac{1}{2}\right)$.

Distribución no informativa de la Poisson

Supongamos que $x|\lambda \sim Po(\lambda)$. Entonces

$$\begin{aligned} f(x|\lambda) &= \frac{e^{-\lambda}\lambda^x}{x!} \propto e^{-\lambda}\lambda^x, \quad x \in \mathbb{N} \cup \{0\}, \quad \lambda > 0, \\ \log f(x|\lambda) &\propto -\lambda + x \log \lambda, \\ \frac{\partial \log f(x|\lambda)}{\partial \lambda} &\propto -1 + \frac{x}{\lambda}, \\ \frac{\partial^2 \log f(x|\lambda)}{\partial \lambda^2} &\propto -\frac{x}{\lambda^2}. \end{aligned}$$

Entonces, la información de Fisher es

$$J(\lambda) \propto E\left[\frac{x}{\lambda^2}\right] = \frac{1}{\lambda^2}E[x] = \frac{1}{\lambda^2}\lambda = \frac{1}{\lambda}.$$

Así, la regla de Jeffreys nos dice que

$$f_\lambda(\lambda) \propto \sqrt{J(\lambda)} \propto \sqrt{\frac{1}{\lambda}} = \lambda^{-1/2}.$$

Distribución no informativa de la Normal (media desconocida)

Supongamos que $x|\mu \sim N(\mu, p)$, siendo $p = 1/\sigma^2$ la precisión. Entonces

$$\begin{aligned} f(x|\mu) &= \frac{\sqrt{p}}{\sqrt{2\pi}} e^{-\frac{p}{2}(x-\mu)^2} \propto e^{-\frac{p}{2}(x-\mu)^2} \\ \log f(x|\mu) &\propto -\frac{p}{2}(x-\mu)^2, \\ \frac{\partial \log f(x|\mu)}{\partial \mu} &\propto p(x-\mu), \\ \frac{\partial^2 \log f(x|\mu)}{\partial \mu^2} &\propto -p \propto -1. \end{aligned}$$

Entonces, la información de Fisher es

$$J(\mu) \propto E\left[1\right] = 1.$$

Así, la regla de Jeffreys nos dice que

$$f_\mu(\mu) \propto \sqrt{J(\mu)} \propto 1.$$

Distribución no informativa de la Normal (desviación típica desconocida)

Supongamos que $x|\sigma \sim N(\mu, \sigma)$. Entonces

$$\begin{aligned} f(x|\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \propto \frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\ \log f(x|\sigma) &\propto -\log\sigma - \frac{1}{2\sigma^2}(x-\mu)^2, \\ \frac{\partial \log f(x|\sigma)}{\partial\sigma} &\propto -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3} \\ \frac{\partial^2 \log f(x|\sigma)}{\partial\sigma^2} &\propto \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}. \end{aligned}$$

Entonces, la información de Fisher es

$$J(\sigma) \propto E\left[-\frac{1}{\sigma^2} + \frac{3(x-\mu)^2}{\sigma^4}\right] \propto -\frac{1}{\sigma^2} + \frac{3}{\sigma^4} E\left[(x-\mu)^2\right] = -\frac{1}{\sigma^2} + \frac{3}{\sigma^4}\sigma^2 \propto \frac{1}{\sigma^2}.$$

Así, la regla de Jeffreys nos dice que

$$f_\sigma(\sigma) \propto \sqrt{J(\sigma)} \propto \sqrt{\frac{1}{\sigma^2}} = \frac{1}{\sigma}.$$

Distribución no informativa de la Normal (desviación típica desconocida)

Supongamos que $x|\sigma^2 \sim N(\mu, \sigma^2)$. En lugar de repetir todo el proceso anterior, podemos hacer lo siguiente. Definimos $x = \sigma$ e $y = \sigma^2$. Es claro que $y = x^2$, es decir, $x = \sqrt{y}$. Acabamos de probar que $f_x(x) \propto 1/x$. Usando el Teoremas de cambio de variables

$$f_y(y) = f_x\left(\sqrt{y}\right)\left|\frac{d}{dy}\left(\sqrt{y}\right)\right| \propto \frac{1}{\sqrt{y}} \cdot \frac{1}{2\sqrt{y}} \propto \frac{1}{y}.$$

Así,

$$f_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2}.$$

Distribución no informativa de la Normal (precisión desconocida)

Supongamos que $x|\tau \sim N(\mu, \tau)$, siendo $\tau = 1/\sigma^2$ la precisión. En lugar de repetir todo el proceso anterior, podemos hacer lo siguiente. Definimos $x = \sigma$ e $y = 1/\sigma^2$. Es claro que $y = 1/x^2$, es decir, $x = 1/\sqrt{y}$. Ya hemos probado que $f_x(x) \propto 1/x$. Usando el Teoremas de cambio de variables

$$f_y(y) = f_x\left(1/\sqrt{y}\right)\left|\frac{d}{dy}\left(1/\sqrt{y}\right)\right| \propto \sqrt{y} \cdot \frac{2}{y\sqrt{y}} \propto \frac{1}{y}.$$

Así,

$$f_\tau(\tau) \propto \frac{1}{\tau}.$$

Distribución no informativa de la Normal (media y desviación típicas desconocidas)

Supongamos que $x|\mu,\sigma \sim N(\mu,\sigma)$. Entonces

$$f(\mu,\sigma) = f_\mu(\mu)\cdot f_\sigma(\sigma) \propto \frac{1}{\sigma}.$$

Distribución no informativa de la Normal (media y varianza desconocidas)

Supongamos que $x|\mu,\sigma^2 \sim N(\mu,\sigma^2)$. Entonces

$$f(\mu,\sigma^2) = f_\mu(\mu)\cdot f_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2}.$$

Distribución no informativa de la Normal (media y precisión desconocidas)

Supongamos que $x|\mu,\tau \sim N(\mu,\tau)$, siendo $\tau = 1/\sigma^2$ la precisión. Entonces

$$f(\mu,\tau) = f_\mu(\mu)\cdot f_\tau(\tau) \propto \frac{1}{\tau}.$$

Regla de Oro

Todo sale de la distribución a priori.

4.6. Estimación puntual

La estimación puntual Bayesiana es un problema de decisión. Debemos tener una función de pérdida, $\mathrm{L} : \Theta\times\Theta \longrightarrow \mathbb{R}$, siendo $L(\theta,t)$ la pérdida si estimamos el parámetro por t siendo θ su verdadero valor. Las funciones de pérdida más usuales son:

- Función de pérdida cuadrática: $L(\theta,t) = (\theta-t)^2$.
- Función pérdida de valor absoluto: $L(\theta,t) = |\theta-t|$.
- Función de pérdida 0-1:

$$L(\theta,t) = \begin{cases} 0, & |\theta-t| \le \varepsilon, \\ 1, & |\theta-t| > \varepsilon. \end{cases}$$

Lo ideal sería minimizar esta función, pero θ es desconocido. Así que, minimizamos $E[L(\theta,t)|\vec{x}]$, siendo $\vec{x}$ nuestros datos. Elegimos $\widehat{\theta}$ tal que

$$E[L(\theta,\widehat{\theta})] = \min_t E[L(\theta,t)|\vec{x}] = \min_t \int_\Theta L(\theta,t) f(\theta|\vec{x})\ d\theta.$$

Supongamos que tenemos como función de pérdida $L(\theta,t) = (\theta-t)^2$. Entonces

$$\begin{aligned}\psi(t) = E[L(\theta,t)|\vec{x}] &= \int_\Theta (\theta-t)^2 f(\theta|\vec{x})\ d\theta \\ &= \int_\Theta \theta^2 f(\theta|\vec{x})\ d\theta + t^2\int_\Theta f(\theta|\vec{x})\ d\theta - 2t\int_\Theta \theta f(\theta|\vec{x})\ d\theta \\ &= \int_\Theta \theta^2 f(\theta|\vec{x})\ d\theta + t^2 - 2tE[\theta|\vec{x}]\end{aligned}$$

Calculemos el mínimo de ψ (respecto de t).

$$\psi'(t) = 2t - 2E[\theta|\vec{x}] \Longrightarrow \psi'(t) = 0 \Longleftrightarrow t = E[\theta|\vec{x}]$$

Se comprueba facilmente que, efectivamente, es un mínimo. Por tanto, tomamos $\widehat{\theta} = E[\theta|\vec{x}]$, es decir, el estimador Bayesiano de θ bajo la función de pérdida cuadrática es la media a posteriori de θ, es decir, $E[\theta|\vec{x}]$.

Si repitiésemos esto para las otras funciones de pérdida, llegaríamos a que

- El estimador Bayesiano de θ bajo la función de pérdida de valor absoluto es la mediana a posteriori de θ.
- El estimador Bayesiano de θ bajo la función de pérdida 0-1 es la moda a posteriori de θ.

Observación 4.14. Se puede probar que la media a posteriori es una media ponderada de la media a priori y del estimador de máxima verosimilitud, es decir, existe $\omega \in (0,1)$ tal que

$$E[\theta|\vec{x}] = \omega E[\theta] + (1-\omega)\widehat{\theta}_{EMV}$$

4.7. Intervalos de credibilidad

Definición 4.15. Un intervalo de credibilidad para θ de contenido probabilístico $1-\alpha$, es un intervalo (a,b) tal que

$$P(a < \theta|\vec{x} < b) = 1-\alpha.$$

Definición 4.16. Una región R es de máxima densidad a posteriori (MDP) de θ de contenido probabilístico $1-\alpha$, si

1. $P(\theta|\vec{x} \in R) = 1-\alpha$.
2. Si $\theta_1 \in R$ y $\theta_2 \notin R$, entonces $f(\theta_1|\vec{x}) > f(\theta_2|\vec{x})$.

Observación 4.17. Si la distribución de $\theta|\vec{x}$ es unimodal, la región MDP de contenido probabilistico $1-\alpha$ para θ es un intervalo (a,b) tal que

1. $P(a < \theta|\vec{x} < b) = 1-\alpha$.
2. $f(a|\vec{x}) = f(b|\vec{x})$.

4.8. Contrastes de hipótesis

Los posibles contrastes de hipótesis que plantearemos serán los siguientes:

$$1.\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases} \qquad 2.\begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases} \qquad 3.\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

Resolveremos estos contrastes de la siguiente manera

1. Calculamos $P(H_0|\vec{x}) = P(\theta \leq \theta_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\theta > \theta_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
2. Calculamos $P(H_0|\vec{x}) = P(\theta \geq \theta_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\theta < \theta_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
3. Fijamos nivel de significación α.
   - Si $\theta_0 \in MDP_{1-\alpha}(\theta|\vec{x})$, aceptamos H_0 a nivel de significación α.
   - Si $\theta_0 \notin MDP_{1-\alpha}(\theta|\vec{x})$, rechazamos H_0 a nivel de significación α.

4.9. Distribución predictiva

Supongamos que $\vec{x} = (x_1, \ldots, x_n)$ es nuestro vector de datos y supongamos que queremos predecir una nueva observación x_{n+1}.

$$f(x_{n+1}|\vec{x}) = \int_\Theta f(x_{n+1}|\theta) f(\theta|\vec{x}) \, d\theta.$$

Ejemplo 4.18.

4.10. Análisis Bayesiano para datos de Bernoulli

Supongamos que $x|\theta \sim Ber(\theta)$ y sea $\vec{x} = (x_1, \ldots, x_n)$ nuestro vector de datos.

Caso informativo

Ya habiamos calculado la familia conjugada para muestras de la Bernoulli y vimos que

$$\begin{aligned} \theta &\sim Be(p, q), \\ \theta|\vec{x} &\sim Be\left(\sum_{i=1}^n x_i + p, n - \sum_{i=1}^n x_i + q\right). \end{aligned}$$

Estimación puntual

- Si $L(\theta, t)$ es la función de pérdida cuadrática,

$$\widehat{\theta} = E[\theta|\vec{x}] = \frac{\sum_{i=1}^n x_i + p}{n + p + q}.$$

- Si $L(\theta, t)$ es la función de pérdida de valor absoluto, $\widehat{\theta} = Me[\theta|\vec{x}]$.
- Si $L(\theta, t)$ es la función de pérdida 0-1,

$$\widehat{\theta} = Mo[\theta|\vec{x}] = \frac{\sum_{i=1}^n x_i + p}{n + p + q - 2}.$$

Calculemos $\omega \in (0, 1)$ tal que

$$\begin{aligned} E[\theta|\vec{x}] &= \omega E[\theta] + (1 - \omega)\widehat{\theta}_{EMV} \\ \frac{\sum_{i=1}^n x_i + p}{n + p + q} &= \omega \frac{p}{p + q} + (1 - \omega)\frac{\sum_{i=1}^n x_i}{n}. \end{aligned}$$

Es fácil comprobar que $\omega = \frac{p+q}{n+p+q}$. Observamos que $\omega \xrightarrow[n\to\infty]{} 0$, lo que nos dice que cuantos más datos tenemos, el estimador de máxima verosimilitud de θ tiene más peso.

Intervalos de credibilidad: Un intervalo de credibilidad de contenido probabilístico $1 - \alpha$ es un intervalo (a, b) tal que

$$1 - \alpha = P(a < \theta|\vec{x} < b) = \int_a^b f(\theta|\vec{x}) \, d\theta.$$

El intervalo MDP de contenido probabilístico $1 - \alpha$ es un intervalo (a, b) tal que

1. $P(a < \theta|\vec{x} < b) = 1 - \alpha$.
2. $f(a|\vec{x}) = f(b|\vec{x})$.

Contrastes de hipótesis

$$1.\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases} \qquad 2.\begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases} \qquad 3.\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

1. Calculamos $P(H_0|\vec{x}) = P(\theta \leq \theta_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\theta > \theta_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
2. Calculamos $P(H_0|\vec{x}) = P(\theta \geq \theta_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\theta < \theta_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
3. Fijamos nivel de significación α.
   - Si $\theta_0 \in MDP_{1-\alpha}(\theta|\vec{x})$, aceptamos H_0 a nivel de significación α.
   - Si $\theta_0 \notin MDP_{1-\alpha}(\theta|\vec{x})$, rechazamos H_0 a nivel de significación α.

Distribución predictiva: Supongamos que queremos hacer una predicción x_{n+1}.

$$\begin{aligned} P(x_{n+1} = 1|\vec{x}) &= \int_0^1 P(x_{n+1} = 1|\theta) f(\theta|\vec{x})\ d\theta \\ &= \int_0^1 \theta \frac{\Gamma(n+p+q)}{\Gamma(\sum_{i=1}^n x_i + p)\Gamma(n - \sum_{i=1}^n x_i + q)} \theta^{\sum_{i=1}^n x_i + p - 1}(1-\theta)^{n - \sum_{i=1}^n x_i - 1 - q}\ d\theta \\ &= \frac{\Gamma(n+p+q)}{\Gamma(\sum_{i=1}^n x_i + p)\Gamma(n - \sum_{i=1}^n x_i + q)} \int_0^1 \theta^{\sum_{i=1}^n x_i + p}(1-\theta)^{n - \sum_{i=1}^n x_i - 1 - q}\ d\theta \end{aligned}$$

Lo que tenemos dentro de la integral es la función de densidad (salvo constantes) de una $Be\left(\sum_{i=1}^n x_i + p + 1, n - \sum_{i=1}^n x_i + q\right)$. Por tanto,

$$\begin{aligned} P(x_{n+1} = 1|\vec{x}) &= \frac{\Gamma(n+p+q)}{\Gamma(\sum_{i=1}^n x_i + p)\Gamma(n - \sum_{i=1}^n x_i + q)} \cdot \frac{\Gamma(\sum_{i=1}^n x_i + p + 1)\Gamma(n - \sum_{i=1}^n x_i - q)}{\Gamma(n+p+q+1)} \\ &= \frac{\sum_{i=1}^n x_i + p}{n+p+q}. \end{aligned}$$

Podemos actuar de forma análoga y calcular la probabilidad de que en las siguientes m pruebas haya r éxistos (y por tanto $m - r$ fracasos) de la siguiente manera:

$$\binom{m}{r} \int_0^1 \theta^r (1-\theta)^{m-r} f(\theta|\vec{x})\ d\theta.$$

Caso no informativo

Recordemos que la regla de Jeffreys nos decía que $\theta \sim Be\left(\frac{1}{2}, \frac{1}{2}\right)$. Si repetimos todo el proceso imponiendo $p = q = \frac{1}{2}$, nos ahorraremos muchos cálculos.
Estimación puntual

- Si $L(\theta, t)$ es la función de pérdida cuadrática,

$$\widehat{\theta} = E[\theta|\vec{x}] = \frac{\sum_{i=1}^n x_i + \frac{1}{2}}{n + \frac{1}{2} + \frac{1}{2}} = \frac{\sum_{i=1}^n x_i + \frac{1}{2}}{n+1}.$$

- Si $L(\theta, t)$ es la función de pérdida de valor absoluto, $\widehat{\theta} = Me[\theta|\vec{x}]$.

- Si $L(\theta, t)$ es la función de pérdida 0-1,

$$\widehat{\theta} = Mo[\theta|\vec{x}] = \frac{\sum_{i=1}^{n} x_i + \frac{1}{2}}{n + \frac{1}{2} + \frac{1}{2} - 2} = \frac{\sum_{i=1}^{n} x_i + \frac{1}{2}}{n - 1}.$$

Calculemos $\omega \in (0, 1)$ tal que

$$E[\theta|\vec{x}] = \omega E[\theta] + (1 - \omega)\widehat{\theta}_{EMV}$$
$$\frac{\sum_{i=1}^{n} x_i + \frac{1}{2}}{n + 1} = \omega \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{2}} + (1 - \omega)\frac{\sum_{i=1}^{n} x_i}{n}.$$

Es fácil comprobar que $\omega = \frac{1}{n+1}$. Observamos que $\omega \xrightarrow[n\to\infty]{} 0$, lo que nos dice que cuantos más datos tenemos, el estimador de máxima verosimilitud de θ tiene más peso.

Intervalos de credibilidad: Un intervalo de credibilidad de contenido probabilístico $1 - \alpha$ es un intervalo (a, b) tal que

$$1 - \alpha = P(a < \theta|\vec{x} < b) = \int_a^b f(\theta|\vec{x})\ d\theta.$$

El intervalo MDP de contenido probabilístico $1 - \alpha$ es un intervalo (a, b) tal que

1. $P(a < \theta|\vec{x} < b) = 1 - \alpha$.
2. $f(a|\vec{x}) = f(b|\vec{x})$.

Contrastes de hipótesis

$$1.\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases} \qquad 2.\begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases} \qquad 3.\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

1. Calculamos $P(H_0|\vec{x}) = P(\theta \leq \theta_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\theta > \theta_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
2. Calculamos $P(H_0|\vec{x}) = P(\theta \geq \theta_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\theta < \theta_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
3. Fijamos nivel de significación α.
   - Si $\theta_0 \in MDP_{1-\alpha}(\theta|\vec{x})$, aceptamos H_0 a nivel de significación α.
   - Si $\theta_0 \notin MDP_{1-\alpha}(\theta|\vec{x})$, rechazamos H_0 a nivel de significación α.

Distribución predictiva: Supongamos que queremos hacer una predicción x_{n+1}.

$$P(x_{n+1} = 1|\vec{x}) \int_0^1 P(x_{n+1} = 1|\theta) f(\theta|\vec{x})\ d\theta = \frac{\sum_{i=1}^{n} x_i + \frac{1}{2}}{n + \frac{1}{2} + \frac{1}{2}} = \frac{\sum_{i=1}^{n} x_i + \frac{1}{2}}{n + 1}.$$

Podemos actuar de forma análoga y calcular la probabilidad de que en las siguientes m pruebas haya r éxistos (y por tanto $m - r$ fracasos) de la siguiente manera:

$$\binom{m}{r} \int_0^1 \theta^r (1 - \theta)^{m-r} f(\theta|\vec{x})\ d\theta.$$

4.11. Análisis Bayesiano para datos de Poisson

Supongamos que $x|\lambda \sim Po(\lambda)$ y sea $\vec{x} = (x_1, \ldots, x_n)$ nuestro vector de datos.

Caso informativo

Ya habiamos calculado la familia conjugada para muestras de la Poisson y vimos que

$$\lambda \sim Ga(p, q),$$
$$\lambda|\vec{x} \sim Ga\left(a+n, \sum_{i=1}^{n} x_i + p\right).$$

Estimación puntual

- Si $L(\lambda, t)$ es la función de pérdida cuadrática,

$$\widehat{\lambda} = E[\lambda|\vec{x}] = \frac{\sum_{i=1}^{n} x_i + p}{a+n}.$$

- Si $L(\lambda, t)$ es la función de pérdida de valor absoluto, $\widehat{\lambda} = Me[\lambda|\vec{x}]$.
- Si $L(\lambda, t)$ es la función de pérdida 0-1,

$$\widehat{\lambda} = Mo[\lambda|\vec{x}] = \frac{\sum_{i=1}^{n} x_i + p - 1}{a+n}$$

Calculemos $\omega \in (0,1)$ tal que

$$E[\lambda|\vec{x}] = \omega E[\lambda] + (1-\omega)\widehat{\lambda}_{EMV}$$
$$\frac{\sum_{i=1}^{n} x_i + p}{a+n} = \omega\frac{a}{p} + (1-\omega)\frac{\sum_{i=1}^{n} x_i}{n}.$$

Se puede comprobar que $\omega \xrightarrow[n\to\infty]{} 0$, lo que nos dice que cuantos más datos tenemos, el estimador de máxima verosimilitud de λ tiene más peso.

Intervalos de credibilidad: Un intervalo de credibilidad de contenido probabilístico $1-\alpha$ es un intervalo (λ_1, λ_2) tal que

$$1 - \alpha = P(\lambda_1 < \lambda|\vec{x} < \lambda_2) = \int_{\lambda_1}^{\lambda_2} f(\lambda|\vec{x})\, d\lambda.$$

Observación 4.19. Si $X \sim Ga(a, p)$, entonces $2aX \sim \chi^2_{2p}$.

Usando esta observación, $2(a+n)\cdot\lambda|\vec{x} \sim \chi^2_{\sum_{i=1}^{n} x_i + p}$.

$$1 - \alpha = P(\lambda_1 < \lambda|\vec{x} < \lambda_2) = P(2(a+n)\lambda_1 < 2(a+n)\cdot\lambda|\vec{x} < 2(a+n)\lambda_2)$$

Tomamos

$$2(a+n)\lambda_1 = \chi^2_{\sum_{i=1}^{n} x_i + p, \alpha/2} \Longrightarrow \lambda_1 = \frac{\chi^2_{\sum_{i=1}^{n} x_i + p, \alpha/2}}{2(a+n)}$$
$$2(a+n)\lambda_2 = \chi^2_{\sum_{i=1}^{n} x_i + p, 1-\alpha/2} \Longrightarrow \lambda_2 = \frac{\chi^2_{\sum_{i=1}^{n} x_i + p, 1-\alpha/2}}{2(a+n)}$$

El intervalo MDP de contenido probabilístico $1-\alpha$ es un intervalo $(\lambda_1\lambda_2)$ tal que

1. $P(\lambda_1 < \lambda|\vec{x} < \lambda_2) = 1 - \alpha$.

2. $f(\lambda_1|\vec{x}) = f(\lambda_2|\vec{x})$.

Contrastes de hipótesis

$$1.\begin{cases} H_0 : \lambda \leq \lambda_0 \\ H_1 : \lambda > \lambda_0 \end{cases} \qquad 2.\begin{cases} H_0 : \lambda \geq \lambda_0 \\ H_1 : \lambda < \lambda_0 \end{cases} \qquad 3.\begin{cases} H_0 : \lambda = \lambda_0 \\ H_1 : \lambda \neq \lambda_0 \end{cases}$$

1. Calculamos $P(H_0|\vec{x}) = P(\lambda \leq \lambda_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\lambda > \lambda_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
2. Calculamos $P(H_0|\vec{x}) = P(\lambda \geq \lambda_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\lambda < \lambda_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
3. Fijamos nivel de significación α.
   - Si $\lambda_0 \in MDP_{1-\alpha}(\lambda|\vec{x})$, aceptamos H_0 a nivel de significación α.
   - Si $\lambda_0 \notin MDP_{1-\alpha}(\lambda|\vec{x})$, rechazamos H_0 a nivel de significación α.

Distribución predictiva: Supongamos que queremos hacer una predicción x_{n+1}.

$$\begin{aligned} P(X_{n+1} = x_{n+1}|\vec{x}) &= \int_0^\infty P(X_{n+1} = x_{n+1}|\lambda) f(\lambda|\vec{x})\ d\lambda \\ &= \int_0^\infty \frac{e^{-\lambda}\lambda^{x_{n+1}}}{x_{n+1}!} \cdot \frac{(a+n)^{\sum_{i=1}^n x_i + p}}{\Gamma(\sum_{i=1}^n x_i + p)} e^{-(a+n)\lambda} \lambda^{\sum_{i=1}^n x_i + p - 1}\ d\lambda \\ &= \frac{(a+n)^{\sum_{i=1}^n x_i + p}}{x_{n+1}!\Gamma(\sum_{i=1}^n x_i + p)} \int_0^\infty e^{-\lambda(a+n+1)} \lambda^{x_{n+1} + \sum_{i=1}^n x_i + p - 1}\ d\lambda \end{aligned}$$

Lo que tenemos dentro de la integral es la función de densidad (salvo constantes) de una $Ga\left(a+n+1, x_{n+1} + \sum_{i=1}^n x_i + p\right)$. Por tanto,

$$P(X_{n+1} = x_{n+1}|\vec{x}) = \frac{(a+n)^{\sum_{i=1}^n x_i + p}}{x_{n+1}!\Gamma(\sum_{i=1}^n x_i + p)} \cdot \frac{\Gamma\left(x_{n+1} + \sum_{i=1}^n x_i + p\right)}{(a+n+1)^{x_{n+1} + \sum_{i=1}^n x_i + p}}$$

Si consideramos que $p \in \mathbb{Z}$ podemos ir un poco más allá en esta integral.

$$\begin{aligned} P(X_{n+1} = x_{n+1}|\vec{x}) &= \frac{(x_{n+1} + \sum_{i=1}^n x_i + p)!}{x_{n+1}!\,(x_{n+1} + \sum_{i=1}^n x_i + p)!} \cdot \left(\frac{a+n}{a+n+1}\right)^{\sum_{i=1}^n x_i + p} \cdot \left(\frac{1}{a+n+1}\right)^{x_{n+1}} \\ &= \binom{x_{n+1} + \sum_{i=1}^n x_i + p}{x_{n+1}} \left(\frac{a+n}{a+n+1}\right)^{\sum_{i=1}^n x_i + p} \cdot \left(\frac{1}{a+n+1}\right)^{x_{n+1}}, \quad p \in \mathbb{Z} \end{aligned}$$

Así, hemos llegado a que $x_{n+1}|\vec{x} \sim BN\left(\sum_{i=1}^n x_i + p, \frac{a+n}{a+n+1}\right)$.

Caso no informativo

Recordemos que la Regla de Jeffreys nos decía que $f_\lambda(\lambda) \propto \lambda^{-1/2}$. Si repetimos todo el proceso imponiendo $a = 0$ y $p = \frac{1}{2}$, nos ahorraremos muchos cálculos.
Estimación puntual

- Si $L(\lambda, t)$ es la función de pérdida cuadrática,

$$\widehat{\lambda} = E[\lambda|\vec{x}] = \frac{\sum_{i=1}^n x_i + \frac{1}{2}}{n}.$$

- Si $L(\lambda, t)$ es la función de pérdida de valor absoluto, $\widehat{\lambda} = Me[\lambda|\vec{x}]$.
- Si $L(\lambda, t)$ es la función de pérdida 0-1,

$$\widehat{\theta} = Mo[\lambda|\vec{x}] = \frac{\sum_{i=1}^{n} x_i + \frac{1}{2} - 1}{n}.$$

Calculemos $\omega \in (0, 1)$ tal que

$$E[\lambda|\vec{x}] = \omega E[\lambda] + (1-\omega)\widehat{\lambda}_{EMV}$$
$$\frac{\sum_{i=1}^{n} x_i + \frac{1}{2}}{n} = \omega \frac{p}{a} + (1-\omega)\frac{\sum_{i=1}^{n} x_i}{n}.$$

Se puede comprobar que $\omega \xrightarrow[n\to\infty]{} 0$, lo que nos dice que cuantos más datos tenemos, el estimador de máxima verosimilitud de λ tiene más peso.

Intervalos de credibilidad: Un intervalo de credibilidad de contenido probabilístico $1-\alpha$ es un intervalo $(\lambda_1\lambda_2)$ tal que

$$1 - \alpha = P(\lambda_1 < \lambda|\vec{x} < \lambda_2) = \int_{\lambda_1}^{\lambda_2} f(\lambda|\vec{x})\ d\lambda.$$

Observación 4.20. Si $X \sim Ga(a, p)$, entonces $2aX \sim \chi^2_{2p}$.

Usando esta observación, $2n \cdot \lambda|\vec{x} \sim \chi^2_{\sum_{i=1}^{n} x_i + p}$.

$$1 - \alpha = P(\lambda_1 < \lambda|\vec{x} < \lambda_2) = P(2n\lambda_1 < 2n \cdot \lambda|\vec{x} < 2n\lambda_2)$$

Tomamos

$$2n\lambda_1 = \chi^2_{\sum_{i=1}^{n} x_i + \frac{1}{2}, \alpha/2} \Longrightarrow \lambda_1 = \frac{\chi^2_{\sum_{i=1}^{n} x_i + \frac{1}{2}, \alpha/2}}{2n}$$
$$2n\lambda_2 = \chi^2_{\sum_{i=1}^{n} x_i + \frac{1}{2}, 1-\alpha/2} \Longrightarrow \lambda_2 = \frac{\chi^2_{\sum_{i=1}^{n} x_i + \frac{1}{2}, 1-\alpha/2}}{2n}$$

El intervalo MDP de contenido probabilístico $1-\alpha$ es un intervalo $(\lambda_1\lambda_2)$ tal que

1. $P(\lambda_1 < \lambda|\vec{x} < \lambda_2) = 1 - \alpha$.
2. $f(\lambda_1|\vec{x}) = f\lambda_2|\vec{x})$.

Contrastes de hipótesis

$$1. \begin{cases} H_0 : \lambda \leq \lambda_0 \\ H_1 : \lambda > \lambda_0 \end{cases} \qquad 2. \begin{cases} H_0 : \lambda \geq \lambda_0 \\ H_1 : \lambda < \lambda_0 \end{cases} \qquad 3. \begin{cases} H_0 : \lambda = \lambda_0 \\ H_1 : \lambda \neq \lambda_0 \end{cases}$$

1. Calculamos $P(H_0|\vec{x}) = P(\lambda \leq \lambda_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\lambda > \lambda_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
2. Calculamos $P(H_0|\vec{x}) = P(\lambda \geq \lambda_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\lambda < \lambda_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
3. Fijamos nivel de significación α.
   - Si $\lambda_0 \in MDP_{1-\alpha}(\lambda|\vec{x})$, aceptamos H_0 a nivel de significación α.

- Si $\lambda_0 \notin MDP_{1-\alpha}(\lambda|\vec{x})$, rechazamos H_0 a nivel de significación α.

Distribución predictiva: Supongamos que queremos hacer una predicción x_{n+1}.

$$\begin{aligned} P(X_{n+1} = x_{n+1}|\vec{x}) &= \int_0^\infty P(X_{n+1} = x_{n+1}|\lambda) f(\lambda|\vec{x})\ d\lambda \\ &= \frac{n^{\sum_{i=1}^n x_i + \frac{1}{2}}}{x_{n+1}! \Gamma(\sum_{i=1}^n x_i + \frac{1}{2})} \cdot \frac{\Gamma\left(x_{n+1} + \sum_{i=1}^n x_i + \frac{1}{2}\right)}{(n+1)^{x_{n+1} + \sum_{i=1}^n x_i + \frac{1}{2}}} \end{aligned}$$

Si consideramos que $p \in \mathbb{Z}$ podemos ir un poco más allá en esta integral.

$$P(X_{n+1} = x_{n+1}|\vec{x}) = \quad = \binom{x_{n+1} + \sum_{i=1}^n x_i + p}{x_{n+1}} \left(\frac{n}{n+1}\right)^{\sum_{i=1}^n x_i + \frac{1}{2}} \cdot \left(\frac{1}{n+1}\right)^{x_{n+1}}, \quad p \in \mathbb{Z}$$

Así, hemos llegado a que $x_{n+1}|\vec{x} \sim BN\left(\sum_{i=1}^n x_i + \frac{1}{2}, \frac{n}{n+1}\right)$.

4.12. Análisis Bayesiano para datos Normales

4.12.1. Normales con media desconocida y precisión conocida

Supongamos que $x|\mu \sim N(\mu, p)$, siendo $p = 1/\sigma^2$ la precisión, y sea $\vec{x} = (x_1, \ldots, x_n)$ nuestro vector de datos.

Caso informativo

Ya habiamos calculado la familia conjugada para muestras de la Normal con media desconocida y precisión conocida, y vimos que

$$\begin{aligned} \mu &\sim N(m_0, a_0), \\ \mu|\vec{x} &\sim N(m_n, p_n) \equiv N\left(\frac{np\overline{x} + p_0 m_0}{np + p_0}, np + p_0\right). \end{aligned}$$

Estimación puntual

- Si $L(\mu, t)$ es la función de pérdida cuadrática,

$$\widehat{\mu} = E[\mu|\vec{x}] = \frac{np\overline{x} + p_0 m_0}{np + p_0}.$$

- Si $L(\mu, t)$ es la función de pérdida de valor absoluto,

$$\widehat{\mu} = Me[\mu|\vec{x}] = \frac{np\overline{x} + p_0 m_0}{np + p_0}.$$

- Si $L(\mu, t)$ es la función de pérdida 0-1,

$$\widehat{\mu} = Mo[\mu|\vec{x}] = \frac{np\overline{x} + p_0 m_0}{np + p_0}.$$

Calculemos $\omega \in (0, 1)$ tal que

$$\begin{aligned} E[\mu|\vec{x}] &= \omega E[\mu] + (1 - \omega)\widehat{\mu}_{EMV} \\ \frac{np\overline{x} + p_0 m_0}{np + p_0} &= \omega m_0 + (1 - \omega)\frac{\sum_{i=1}^n x_i}{n}. \end{aligned}$$

Se puede comprobar que $\omega = \frac{p_0}{np + p_0} \xrightarrow[n\to\infty]{} 0$, lo que nos dice que cuantos más datos tenemos, el estimador de máxima verosimilitud de μ tiene más peso.

Intervalos de credibilidad: El intervalo MDP de contenido probabilístico $1 - \alpha$ para μ es un intervalo (μ_1, μ_2) tal que

- $P(\mu_1 < \mu|\vec{x} < \mu_2) = 1 - \alpha$.
- $f(\mu_1|\vec{x}) = f(\mu_2|\vec{x})$.

$$1 - \alpha = P(\mu_1 < \mu|\vec{x} < \mu_2) = P(\sqrt{p_n}(\mu_1 - m_n) < \sqrt{p_n}(\mu - m_n)|\vec{x} < \sqrt{p_n}(\mu_2 - m_n))$$

Tenemos que $\sqrt{p_n}(\mu - m_n)|\vec{x} \sim N(0,1)$, por lo que tomamos

$$\sqrt{p_n}(\mu_1 - m_n) = -z_{1-\alpha/2} \Longrightarrow \mu_1 = m_n - \frac{1}{\sqrt{p_n}} z_{1-\alpha/2}$$
$$\sqrt{p_n}(\mu_2 - m_n) = z_{1-\alpha/2} \Longrightarrow \mu_2 = m_n + \frac{1}{\sqrt{p_n}} z_{1-\alpha/2}$$

Contrastes de hipótesis

$$1. \begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases} \qquad 2. \begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \qquad 3. \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

1. Calculamos $P(H_0|\vec{x}) = P(\mu \leq \mu_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\mu > \mu_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
2. Calculamos $P(H_0|\vec{x}) = P(\mu \geq \mu_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\mu < \mu_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
3. Fijamos nivel de significación α.
   - Si $\mu_0 \in MDP_{1-\alpha}(\mu|\vec{x})$, aceptamos H_0 a nivel de significación α.
   - Si $\mu_0 \notin MDP_{1-\alpha}(\mu|\vec{x})$, rechazamos H_0 a nivel de significación α.

Distribución predictiva

$$\begin{aligned} f(x_{n+1}|\vec{x}) &= \int_{\mathbb{R}} f(x_{n+1}|\mu) f(\mu|\vec{x})\, d\mu = \int_{\mathbb{R}} \frac{\sqrt{p}}{\sqrt{2\pi}} e^{-\frac{p}{2}(x_{n+1}-\mu)^2} \frac{\sqrt{p_n}}{\sqrt{2\pi}} e^{-\frac{p_n}{2}(\mu - m_n)^2}\, d\mu \\ &= \frac{\sqrt{p}\sqrt{p_n}}{2\pi} \int_{\mathbb{R}} e^{-\frac{1}{2}(p(x_{n+1}-\mu)^2 + p_n(\mu - m_n)^2)}\, d\mu \end{aligned}$$

Usando el Lema, podemos hacer

$$p(x_{n+1} - \mu)^2 + p_n(\mu - m_n)^2) = \alpha(\mu - \beta)^2 + \gamma,$$

donde,

$$\alpha = p + p_n, \quad \beta = \frac{p x_{n+1} + p_n m_n}{p + p_n}, \quad \gamma = \frac{\sqrt{p p_n}}{\sqrt{2\pi}} (x_{n+1} - m_n)^2.$$

Así,

$$f(x_{n+1}|\vec{x}) = \frac{\sqrt{p}\sqrt{p_n}}{2\pi} \int_{\mathbb{R}} e^{-\frac{1}{2}(\alpha(\mu-\beta)^2 + \gamma)}\, d\mu = \frac{\sqrt{p}\sqrt{p_n}}{2\pi} e^{-\frac{1}{2}\gamma} \int_{\mathbb{R}} e^{-\frac{\alpha}{2}(\mu-\beta)^2}\, d\mu$$

Observamos que lo de dentro de la última integral es la función de densidad (salvo consntantes) de una $N(\beta, \alpha)$, (α precisión), por tanto

$$f(x_{n+1}|\vec{x}) = \frac{\sqrt{p}\sqrt{p_n}}{2\pi} e^{-\frac{1}{2}\gamma} \cdot \frac{\sqrt{2\pi}}{\sqrt{\alpha}} = \frac{\sqrt{p}\sqrt{p_n}}{2\pi} \cdot \frac{1}{\sqrt{p + p_n}} e^{\frac{p p_n}{p + p_n}(x_{n+1} - m_n)^2},$$

donde concluimos que $x_{n+1}|\vec{x} \sim N\left(m_n, \frac{p p_n}{p + p_n}\right)$ (siendo el segundo parámetro la precisión).

Caso no informativo

Recordemos que la regla de Jeffreys nos decía que en este caso, $f_\mu(\mu) \propto 1$. Si repetimos todo el proceso anterior imponiendo $p_0 = 0$ nos ahorraremos muchos cálculos.

Estimación puntual

- Si $L(\mu, t)$ es la función de pérdida cuadrática,

$$\widehat{\mu} = E[\mu|\vec{x}] = \frac{np\overline{x}}{np} = \overline{x}.$$

- Si $L(\mu, t)$ es la función de pérdida de valor absoluto,

$$\widehat{\mu} = Me[\mu|\vec{x}] = \frac{np\overline{x}}{np} = \overline{x}.$$

- Si $L(\mu, t)$ es la función de pérdida 0-1,

$$\widehat{\mu} = Mo[\mu|\vec{x}] = \frac{np\overline{x}}{np} = \overline{x}.$$

Calculemos $\omega \in (0, 1)$ tal que

$$E[\mu|\vec{x}] = \omega E[\mu] + (1-\omega)\widehat{\mu}_{EMV}$$
$$\overline{x} = \omega m_0 + (1-\omega)\overline{x}.$$

Se tiene que $\omega = 0$, es decir, la media a posteriori de μ conincide con el estimador de máxima verosimilitud.

Intervalos de credibilidad: El intervalo MDP de contenido probabilístico $1-\alpha$ para μ es un intervalo (μ_1, μ_2) tal que

- $P(\mu_1 < \mu|\vec{x} < \mu_2) = 1 - \alpha$.
- $f(\mu_1|\vec{x}) = f(\mu_2|\vec{x})$.

$$1 - \alpha = P(\mu_1 < \mu|\vec{x} < \mu_2) = P(\sqrt{pn}(\mu_1 - \overline{x}) < \sqrt{pn}(\mu - \overline{x})|\vec{x} < \sqrt{pn}(\mu_2 - \overline{x}))$$

Tenemos que $\sqrt{pn}(\mu - \overline{x})|\vec{x} \sim N(0, 1)$, por lo que tomamos

$$\sqrt{pn}(\mu_1 - \overline{x}) = -z_{1-\alpha/2} \Longrightarrow \mu_1 = \overline{x} - \frac{1}{\sqrt{pn}} z_{1-\alpha/2}$$
$$\sqrt{pn}(\mu_2 - \overline{x}) = z_{1-\alpha/2} \Longrightarrow \mu_2 = \overline{x} + \frac{1}{\sqrt{pn}} z_{1-\alpha/2}$$

Tenemos entonces que

$$MDP_{1-\alpha} = \left(\overline{x} - \frac{1}{\sqrt{pn}} z_{1-\alpha/2}, \overline{x} + \frac{1}{\sqrt{pn}} z_{1-\alpha/2}\right) = \left(\overline{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \overline{x} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right),$$

que coincide con el intervalo al $(1-\alpha)\cdot 100\,\%$ de confianza (clásico) para μ.

Contrastes de hipótesis:

$$1. \begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases} \qquad 2. \begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \qquad 3. \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

1. Calculamos $P(H_0|\vec{x}) = P(\mu \leq \mu_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\mu > \mu_0|\vec{x})$.

- Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
- Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.

2. Calculamos $P(H_0|\vec{x}) = P(\mu \geq \mu_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\mu < \mu_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
3. Fijamos nivel de significación α.
   - Si $\mu_0 \in MDP_{1-\alpha}(\mu|\vec{x})$, aceptamos H_0 a nivel de significación α.
   - Si $\mu_0 \notin MDP_{1-\alpha}(\mu|\vec{x})$, rechazamos H_0 a nivel de significación α.

Distribución predictiva: Repitiendo todos los cálculos con $p_0 = 0$, concluimos que $x_{n+1}|\vec{x} \sim N\left(\overline{x}, \frac{np}{n+1}\right)$ (siendo el segundo parámetro la precisión).

4.12.2. Normales con varianza desconocida y media conocida

Supongamos que $x|\sigma^2 \sim N(\mu, \sigma^2)$ y sea $\vec{x} = (x_1, \ldots, x_n)$ nuestro vector de datos.

Caso informativo

Ya habiamos calculado la familia conjugada para muestras de la Normal con media desconocida y precisión conicda, y vimos que

$$\begin{aligned} \sigma^2 &\sim GaI(a_0, p_0), \\ \sigma^2|\vec{x} &\sim GaI(a_n, p_n) \equiv GaI\left(\frac{2a_0 + \sum_{i=1}^n (x_i - \mu)^2}{2}, \frac{n}{2} + p_0\right). \end{aligned}$$

Estimación puntual

- Si $L(\sigma^2, t)$ es la función de pérdida cuadrática,

$$\widehat{\sigma^2} = E[\sigma^2|\vec{x}] = \frac{\frac{2a_0 + \sum_{i=1}^n (x_i - \mu)^2}{2}}{\frac{n}{2} + p_0 - 1}.$$

- Si $L(\sigma^2, t)$ es la función de pérdida de valor absoluto, $\widehat{\sigma^2} = Me[\sigma^2|\vec{x}]$.
- Si $L(\sigma^2, t)$ es la función de pérdida 0-1,

$$\widehat{\sigma^2} = Mo[\sigma^2|\vec{x}] = \frac{\frac{2a_0 + \sum_{i=1}^n (x_i - \mu)^2}{2}}{\frac{n}{2} + p_0 + 1}$$

Calculemos $\omega \in (0, 1)$ tal que

$$\begin{aligned} E[\sigma^2|\vec{x}] &= \omega E[\sigma^2] + (1 - \omega)\widehat{\sigma^2}_{EMV} \\ \frac{\frac{2a_0 + \sum_{i=1}^n (x_i - \mu)^2}{2}}{\frac{n}{2} + p_0 - 1} &= \omega m_0 + (1 - \omega)\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}. \end{aligned}$$

Se puede comprobar que $\omega \xrightarrow[n \to \infty]{} 0$, lo que nos dice que cuantos más datos tenemos, el estimador de máxima verosimilitud de σ^2 tiene más peso.

Intervalos de credibilidad: Un intervalo de credibilidad de contenido probabilístico $1 - \alpha$ para σ^2 es un intervalo (σ_1^2, σ_2^2) tal que

$$1 - \alpha = P(\sigma_1^2 < \sigma^2|\vec{x} < \sigma_2^2) = P\left(\frac{1}{\sigma_2^2} < \frac{1}{\sigma^2}|\vec{x} < \frac{1}{\sigma_1^2}\right)$$

Nótese que $\frac{1}{\sigma^2}|\vec{x}$ sigue una distribución Gamma, puesto que $\sigma^2|\vec{x}$ sigue una distribución Gamma Inversa. Así,

$$1-\alpha = P\left(\frac{1}{\sigma_2^2} < \frac{1}{\sigma^2}|\vec{x} < \frac{1}{\sigma_1^2}\right) = P\left(\frac{2a_n}{\sigma_2^2} < \frac{2a_n}{\sigma^2}|\vec{x} < \frac{2a_n}{\sigma_1^2}\right),$$

donde $\frac{2a_n}{\sigma^2}|\vec{x}$ sigue una $\chi^2_{2p_n}$. El intervalo MDP de contenido probabilístico $1-\alpha$ para σ^2 debe verificar además que $f(\sigma_1^2|\vec{x}) = f(\sigma_2^2|\vec{x})$.

Contrastes de hipótesis:

$$1.\begin{cases} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases} \qquad 2.\begin{cases} H_0 : \sigma^2 \geq \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{cases} \qquad 3.\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$$

1. Calculamos $P(H_0|\vec{x}) = P(\sigma^2 \leq \sigma_0^2|\vec{x})$ y $P(H_1|\vec{x}) = P(\sigma^2 > \sigma_0^2|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
2. Calculamos $P(H_0|\vec{x}) = P(\sigma^2 \geq \sigma_0^2|\vec{x})$ y $P(H_1|\vec{x}) = P(\sigma^2 < \sigma_0^2|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
3. Fijamos nivel de significación α.
   - Si $\sigma_0^2 \in MDP_{1-\alpha}(\sigma^2|\vec{x})$, aceptamos H_0 a nivel de significación α.
   - Si $\sigma_0^2 \notin MDP_{1-\alpha}(\sigma^2|\vec{x})$, rechazamos H_0 a nivel de significación α.

Distribución predictiva:

$$\begin{aligned} f(x_{n+1}|\vec{x}) &= \int_0^\infty f(x_{n+1}|\sigma^2) f(\sigma^2|\vec{x})\ d\sigma^2 \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_{n+1}-\mu)^2} \cdot \frac{a_n^{p_n}}{\Gamma(p_n)} e^{-\frac{a_n}{\sigma^2}} \left(\sigma^2\right)^{-(p_n+1)}\ d\sigma^2 \\ &= \frac{1}{\sqrt{2\pi}} \frac{a_n^{p_n}}{\Gamma(p_n)} \int_0^\infty \left(\sigma^2\right)^{-\left(p_n+\frac{1}{2}+1\right)} e^{-\frac{1}{2\sigma^2}((x_{n+1}-\mu)^2+2a_n)}\ d\sigma^2. \end{aligned}$$

Observamos que lo que hay dentro de la integral es la fución de densidad (salvo constantes) de una $GaI\left(\frac{(x_{n+1}-\mu)^2}{2} + a_n, p_+\frac{1}{2}\right)$, por tanto

$$\begin{aligned} f(x_{n+1}|\vec{x}) &= \frac{1}{\sqrt{2\pi}} \frac{a_n^{p_n}}{\Gamma(p_n)} \frac{\Gamma\left(p_n+\frac{1}{2}\right)}{\left(\frac{(x_{n+1}-\mu)^2}{2}+a_n\right)^{p_n+\frac{1}{2}}} \\ &\vdots \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(p_n)} \frac{\Gamma\left(p_n+\frac{1}{2}\right)}{\sqrt{a_n}} \left(1+\frac{1}{2a_n}(x_{n+1}-\mu)^2\right)^{-\left(p_n+\frac{1}{2}\right)}, \end{aligned}$$

de donde concluímos que $x_{n+1}|\vec{x} \sim t\left(\mu, \frac{p_n}{a_n}, 2p_n\right) \equiv t\left(\mu, \frac{n+2p_0}{2a_0+\sum_{i=1}^n (x_i-\mu)^2}, n+2p_0\right)$.

Caso no informativo

Recordemos que la regla de Jeffreys nos decía que en este caso, $f_{\sigma^2}(\sigma^2) \propto \left(\sigma^2\right)^{-1}$. Si repetimos todo el proceso anterior imponiendo $a_0 = p_0 = 0$ nos ahorraremos muchos cálculos. Ahora

$$\begin{aligned}\sigma^2 &\sim GaI(a_0, p_0),\\ \sigma^2|\vec{x} &\sim GaI(a_n, p_n) \equiv GaI\left(\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}, \frac{n}{2}\right).\end{aligned}$$

Estimación puntual

- Si $L(\sigma^2, t)$ es la función de pérdida cuadrática,

$$\widehat{\sigma^2} = E[\sigma^2|\vec{x}] = \frac{\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}}{\frac{n}{2}-1} = \frac{\sum_{i=1}^n (x_i-\mu)^2}{n-2}.$$

- Si $L(\sigma^2, t)$ es la función de pérdida de valor absoluto, $\widehat{\sigma^2} = Me[\sigma^2|\vec{x}]$.
- Si $L(\sigma^2, t)$ es la función de pérdida 0-1, $\widehat{\sigma^2} = Mo[\sigma^2|\vec{x}]$.

Calculemos $\omega \in (0,1)$ tal que

$$\begin{aligned}E[\sigma^2|\vec{x}] &= \omega E[\sigma^2] + (1-\omega)\widehat{\sigma^2}_{EMV}\\ \frac{\sum_{i=1}^n (x_i-\mu)^2}{n-2} &= \omega m_0 + (1-\omega)\frac{\sum_{i=1}^n (x_i-\mu)^2}{n-2}.\end{aligned}$$

Es fácil ver que $\omega = 0$, lo que nos dice que la media a posteriori de σ^2 coincide con el estimador de máxima verosimilitud.

Intervalos de credibilidad: Un intervalo de credibilidad de contenido probabilístico $1-\alpha$ para σ^2 es un intervalo (σ_1^2, σ_2^2) tal que

$$1-\alpha = P(\sigma_1^2 < \sigma^2|\vec{x} < \sigma_2^2) = P\left(\frac{1}{\sigma_2^2} < \frac{1}{\sigma^2}|\vec{x} < \frac{1}{\sigma_1^2}\right)$$

Nótese que $\frac{1}{\sigma^2}|\vec{x}$ sigue una distribución Gamma, puesto que $\sigma^2|\vec{x}$ sigue una distribución Gamma Inversa. Así,

$$1-\alpha = P\left(\frac{1}{\sigma_2^2} < \frac{1}{\sigma^2}|\vec{x} < \frac{1}{\sigma_1^2}\right) = P\left(\frac{2a_n}{\sigma_2^2} < \frac{2a_n}{\sigma^2}|\vec{x} < \frac{2a_n}{\sigma_1^2}\right),$$

donde $\frac{2a_n}{\sigma^2}|\vec{x}$ sigue una $\chi^2_{2p_n}$. El intervalo MDP de contenido probabilístico $1-\alpha$ para σ^2 debe verificar además que $f(\sigma_1^2|\vec{x}) = f(\sigma_2^2|\vec{x})$.

Contrastes de hipótesis:

$$1.\begin{cases}H_0 : \sigma^2 \le \sigma_0^2\\ H_1 : \sigma^2 > \sigma_0^2\end{cases} \qquad 2.\begin{cases}H_0 : \sigma^2 \ge \sigma_0^2\\ H_1 : \sigma^2 < \sigma_0^2\end{cases} \qquad 3.\begin{cases}H_0 : \sigma^2 = \sigma_0^2\\ H_1 : \sigma^2 \ne \sigma_0^2\end{cases}$$

1. Calculamos $P(H_0|\vec{x}) = P(\sigma^2 \le \sigma_0^2|\vec{x})$ y $P(H_1|\vec{x}) = P(\sigma^2 > \sigma_0^2|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \le P(H_1|\vec{x})$, rechazamos H_0.
2. Calculamos $P(H_0|\vec{x}) = P(\sigma^2 \ge \sigma_0^2|\vec{x})$ y $P(H_1|\vec{x}) = P(\sigma^2 < \sigma_0^2|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \le P(H_1|\vec{x})$, rechazamos H_0.

3. Fijamos nivel de significación α.

   - Si $\sigma_0^2 \in MDP_{1-\alpha}(\sigma^2|\vec{x})$, aceptamos H_0 a nivel de significación α.
   - Si $\sigma_0^2 \notin MDP_{1-\alpha}(\sigma^2|\vec{x})$, rechazamos H_0 a nivel de significación α.

Distribución predictiva: Concluímos que $x_{n+1}|\vec{x} \sim t\left(\mu, \frac{n}{\sum_{i=1}^n (x_i-\mu)^2}, n\right)$.

4.12.3. Normales con media y precisión desconocidas

Supongamos que $x|\mu,\tau \sim N(\mu,\tau)$, siendo $\tau = 1/\sigma^2$ la precisión, y sea $\vec{x} = (x_1,\ldots,x_n)$ nuestro vector de datos.

Caso informativo

Ya habiamos calculado la familia conjugada para muestras de la Normal con media y precisión desconocidas, y vimos que

$$(\mu,\tau) \sim NGa(m_0,\tau_0,a_0,p_0)$$
$$(\mu,\tau)|\vec{x} \sim NG(m_n,\tau,a_n,p_n) \equiv NGa\left(\frac{\tau_0 m_0 + n\overline{x}}{\tau_0+n}, \tau_0+n, a_0 + \frac{(n-1)s^2}{2} + \frac{\tau_0 n(\overline{x}-m_0)^2}{2(\tau_0+n)^2}, \frac{n}{2}+p_0\right).$$

Recordemos que por la definición de la Normal-Gamma, tenemos que

$$\mu|\vec{x} \sim t\left(m_n, \frac{p_n\tau_n}{a_n}, 2p_n\right)$$
$$\tau|\vec{x} \sim Ga(a_n,p_n).$$

Estimación puntual para μ

- Si $L(\mu,t)$ es la función de pérdida cuadrática, $\widehat{\mu} = E[\mu|\vec{x}] = m_n$.
- Si $L(\mu,t)$ es la función de pérdida de valor absoluto, $\widehat{\mu} = Me[\mu|\vec{x}]$.
- Si $L(\mu,t)$ es la función de pérdida 0-1, $\widehat{\mu} = Mo[\mu|\vec{x}] = m_n$.

Calculemos $\omega \in (0,1)$ tal que

$$E[\mu|\vec{x}] = \omega E[\mu] + (1-\omega)\widehat{\mu}_{EMV}$$

Se puede comprobar que $\omega \xrightarrow[n\to\infty]{} 0$, lo que nos dice que cuantos más datos tenemos, el estimador de máxima verosimilitud de μ tiene más peso.

Estimación puntual para τ

- Si $L(\tau,t)$ es la función de pérdida cuadrática, $\widehat{\tau} = E[\tau|\vec{x}] = \frac{a_n}{p_n}$.
- Si $L(\tau,t)$ es la función de pérdida de valor absoluto, $\widehat{\tau} = Me[\tau|\vec{x}]$.
- Si $L(\tau,t)$ es la función de pérdida 0-1, $\widehat{\tau} = Mo[\tau|\vec{x}]$.

Calculemos $\omega \in (0,1)$ tal que

$$E[\tau|\vec{x}] = \omega E[\tau] + (1-\omega)\widehat{\tau}_{EMV}$$

Se puede comprobar que $\omega \xrightarrow[n\to\infty]{} 0$, lo que nos dice que cuantos más datos tenemos, el estimador de máxima verosimilitud de μ tiene más peso.

Intervalos de credibilidad para μ: Un intervalo de credibilidad de contenido probabilístico para μ es un intervalo (μ_1, μ_2) tal que $P(\mu_1 < \mu|\vec{x} < \mu_2) = 1 - \alpha$. Nótese que

$$\mu|\vec{x} \sim t\left(m_n, \frac{p_n\tau_n}{a_n}, 2p_n\right) \Longrightarrow \sqrt{\frac{p_n\tau}{a_n}}(\mu - m_n)|\vec{x} \sim t_{2p_n}.$$

Así

$$1 - \alpha = P(\mu_1 < \mu|\vec{x} < \mu_2) = P\left(\sqrt{\frac{p_n\tau}{a_n}}(\mu_1 - m_n) < \sqrt{\frac{p_n\tau}{a_n}}(\mu - m_n)|\vec{x} < \sqrt{\frac{p_n\tau}{a_n}}(\mu_2 - m_n)\right).$$

El intervalo MDP de contenido probabilístico $1 - \alpha$ para μ debe verificar además que $f(\mu_1|\vec{x}) = f(\mu_2|\vec{x})$, lo que se traduce en que

$$\sqrt{\frac{p_n\tau}{a_n}}(\mu_1 - m_n) = -t_{2p_n,1-\alpha/2} \Longrightarrow \mu_1 = m_n - \sqrt{\frac{a_n}{p_n\tau_n}}t_{2p_-1-\alpha/2}$$
$$\sqrt{\frac{p_n\tau}{a_n}}(\mu_2 - m_n) = t_{2p_n,1-\alpha/2} \Longrightarrow \mu_2 = m_n + \sqrt{\frac{a_n}{p_n\tau_n}}t_{2p_-1-\alpha/2}$$

Intervalos de credibilidad para τ: Nótese que

$$\tau|\vec{x} \sim Ga(a_n, p_n) \Longrightarrow 2a_n \cdot \tau|\vec{x} \sim \chi^2_{2p_n}.$$

Un intervalo de credibilidad de contenido probabilístico para τ es un intervalo (τ_1, τ_2) tal que $P(\tau_1 < \tau|\vec{x} < \tau_2) = 1 - \alpha$. El intervalo MDP de contenido probabilístico $1 - \alpha$ para μ debe verificar además que $f(\tau_1|\vec{x}) = f(\tau_2|\vec{x})$.

Contrastes de hipótesis para μ:

$$1.\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases} \qquad 2.\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \qquad 3.\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

1. Calculamos $P(H_0|\vec{x}) = P(\mu \leq \mu_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\mu > \mu_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
2. Calculamos $P(H_0|\vec{x}) = P(\mu \geq \mu_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\mu < \mu_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
3. Fijamos nivel de significación α.
   - Si $\mu_0 \in MDP_{1-\alpha}(\mu|\vec{x})$, aceptamos H_0 a nivel de significación α.
   - Si $\mu_0 \notin MDP_{1-\alpha}(\mu|\vec{x})$, rechazamos H_0 a nivel de significación α.

Contrastes de hipótesis para τ:

$$1.\begin{cases} H_0 : \tau \leq \tau_0 \\ H_1 : \tau > \tau_0 \end{cases} \qquad 2.\begin{cases} H_0 : \tau \geq \tau_0 \\ H_1 : \tau < \tau_0 \end{cases} \qquad 3.\begin{cases} H_0 : \tau = \tau_0 \\ H_1 : \tau \neq \tau_0 \end{cases}$$

1. Calculamos $P(H_0|\vec{x}) = P(\tau \leq \tau_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\tau > \tau_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
2. Calculamos $P(H_0|\vec{x}) = P(\tau \geq \tau_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\tau < \tau_0|\vec{x})$.

- Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
- Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.

3. Fijamos nivel de significación α.
   - Si $\tau_0 \in MDP_{1-\alpha}(\tau|\vec{x})$, aceptamos H_0 a nivel de significación α.
   - Si $\tau_0 \notin MDP_{1-\alpha}(\tau|\vec{x})$, rechazamos H_0 a nivel de significación α.

Distribución predictiva:

$$\begin{aligned} f(x_{n+1}|\vec{x}) &= \int_0^\infty \int_{\mathbb{R}} f(x_{n+1}|\mu,\tau) f(\mu,\tau|\vec{x})\ d\mu d\tau \\ &= \int_0^\infty \int_{\mathbb{R}} \frac{\sqrt{\tau}}{\sqrt{2\pi}} e^{-\frac{\tau}{2}(x_{n+1}-\mu)^2} \cdot \frac{\sqrt{\tau_n} a_n^{p_n}}{\sqrt{2\pi}\Gamma(p_n)} e^{-\frac{\tau}{2}(2a_n+\tau_n(\mu-m_n)^2)} \tau^{p_n-\frac{1}{2}}\ d\mu d\tau \\ &\ \ \vdots \\ &= (2a_n)^{-\left(p_n+\frac{1}{2}\right)} \left(1+\frac{\tau_n}{2a_n(\tau_n+1)(x_{n+1}-n)^2}\right)^{-\left(p_n+\frac{1}{2}\right)}, \end{aligned}$$

de donde concluímos que $x_{n+1}|\vec{x} \sim t\left(m_n, \frac{p_n\tau_n}{a_n(\tau_n+1)}, 2p_n\right)$.

Caso no informativo

Recordemos que la regla de Jeffreys nos decía que en este caso, $f_{(\mu,\tau)}(\mu,\tau) \propto \tau^{-1}$. Si repetimos el proceso anterior para $a_0 = \tau_0 = 0$ y $p_0 = \frac{1}{2}$, nos ahorraremos muchos cálculos. Así,

$$\begin{aligned} (\mu,\tau)|\vec{x} &\sim NGa\left(\overline{x}, n, \frac{n-1}{2}s^2, \frac{n-1}{2}\right), \\ \mu|\vec{x} &\sim t\left(\overline{x}, \frac{n}{s^2}, n-1\right), \\ \tau|\vec{x} &\sim Ga\left(\frac{n-1}{2}s^2, \frac{n-1}{2}\right). \end{aligned}$$

Estimación puntual para μ

- Si $L(\mu,t)$ es la función de pérdida cuadrática, $\widehat{\mu} = E[\mu|\vec{x}] = \overline{x}$.
- Si $L(\mu,t)$ es la función de pérdida de valor absoluto, $\widehat{\mu} = Me[\mu|\vec{x}]$.
- Si $L(\mu,t)$ es la función de pérdida 0-1, $\widehat{\mu} = Mo[\mu|\vec{x}] = \overline{x}$.

Calculemos $\omega \in (0,1)$ tal que

$$\begin{aligned} E[\mu|\vec{x}] &= \omega E[\mu] + (1-\omega)\widehat{\mu}_{EMV} \\ \overline{x} &= \omega E[\mu] + (1-\omega)\overline{x} \end{aligned}$$

Es fácil ver que $\omega = 0$, lo que nos dice que la media a posteriori de μ coincide con el estimador de máxima verosimilitud.

Estimación puntual para τ

- Si $L(\tau,t)$ es la función de pérdida cuadrática,

$$\widehat{\tau} = E[\tau|\vec{x}] = \frac{\frac{n-1}{2}}{\frac{n-1}{2}s^2} = \frac{1}{s^2}.$$

- Si $L(\tau, t)$ es la función de pérdida de valor absoluto, $\widehat{\tau} = Me[\tau|\vec{x}]$.
- Si $L(\tau, t)$ es la función de pérdida 0-1, $\widehat{\tau} = Mo[\tau|\vec{x}]$.

Calculemos $\omega \in (0, 1)$ tal que

$$E[\tau|\vec{x}] = \omega E[\tau] + (1-\omega)\widehat{\tau}_{EMV}$$
$$\frac{n-1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \omega E[\tau] + (1-\omega)\frac{n-1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Es fácil ver que $\omega = 0$, lo que nos dice que la media a posteriori de τ coincide con el estimador de máxima verosimilitud.

Intervalos de credibilidad para μ: Un intervalo de credibilidad de contenido probabilístico para μ es un intervalo (μ_1, μ_2) tal que $P(\mu_1 < \mu|\vec{x} < \mu_2) = 1 - \alpha$. Nótese que

$$\mu|\vec{x} \sim t\left(\overline{x}, \frac{n}{s^2}, n-1\right) \Longrightarrow \sqrt{\frac{n}{s^2}}(\mu - \overline{x})|\vec{x} \sim t_{n-1}.$$

El intervalo MDP de contenido probabilístico $1 - \alpha$ para μ es

$$MDP_{1-\alpha}(\mu) = \left(\overline{x} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right),$$

que coincide con el intervalo de confianza clásico.

Intervalos de credibilidad para τ: Ahora tenemos que

$$\tau|\vec{x} \sim Ga\left(\frac{n-1}{2}s^2, \frac{n-1}{2}\right) \Longrightarrow (n-1)s^2 \cdot \tau|\vec{x} \sim \chi^2_{n-1}.$$

Un intervalo de credibilidad de contenido probabilístico para τ es un intervalo (τ_1, τ_2) tal que $P(\tau_1 < \tau|\vec{x} < \tau_2) = 1 - \alpha$. Por ejemplo:

$$IC_{1-\alpha}(\tau) = \left(\frac{\chi^2_{n-1,\alpha/2}}{(n-1)s^2}, \frac{\chi^2_{n-1,1-\alpha/2}}{(n-1)s^2}\right)$$

El intervalo MDP de contenido probabilístico $1 - \alpha$ para μ debe verificar además que $f(\tau_1|\vec{x}) = f(\tau_2|\vec{x})$.

Contrastes de hipótesis para μ:

$$1.\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases} \qquad 2.\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \qquad 3.\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

1. Calculamos $P(H_0|\vec{x}) = P(\mu \leq \mu_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\mu > \mu_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
2. Calculamos $P(H_0|\vec{x}) = P(\mu \geq \mu_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\mu < \mu_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
3. Fijamos nivel de significación α.
   - Si $\mu_0 \in MDP_{1-\alpha}(\mu|\vec{x})$, aceptamos H_0 a nivel de significación α.
   - Si $\mu_0 \notin MDP_{1-\alpha}(\mu|\vec{x})$, rechazamos H_0 a nivel de significación α.

Contrastes de hipótesis para τ:

$$1.\begin{cases} H_0 : \tau \leq \tau_0 \\ H_1 : \tau > \tau_0 \end{cases} \qquad 2.\begin{cases} H_0 : \tau \geq \tau_0 \\ H_1 : \tau < \tau_0 \end{cases} \qquad 3.\begin{cases} H_0 : \tau = \tau_0 \\ H_1 : \tau \neq \tau_0 \end{cases}$$

1. Calculamos $P(H_0|\vec{x}) = P(\tau \leq \tau_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\tau > \tau_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
2. Calculamos $P(H_0|\vec{x}) = P(\tau \geq \tau_0|\vec{x})$ y $P(H_1|\vec{x}) = P(\tau < \tau_0|\vec{x})$.
   - Si $P(H_0|\vec{x}) > P(H_1|\vec{x})$, aceptamos H_0.
   - Si $P(H_0|\vec{x}) \leq P(H_1|\vec{x})$, rechazamos H_0.
3. Fijamos nivel de significación α.
   - Si $\tau_0 \in MDP_{1-\alpha}(\tau|\vec{x})$, aceptamos H_0 a nivel de significación α.
   - Si $\tau_0 \notin MDP_{1-\alpha}(\tau|\vec{x})$, rechazamos H_0 a nivel de significación α.

Distribución predictiva: Tenemos que $x_{n+1}|\vec{x} \sim t\left(\overline{x}, \frac{n}{(n-1)s^2}, n-1\right)$.

4.13. Influencia de la distribución a priori según el tamaño muestral

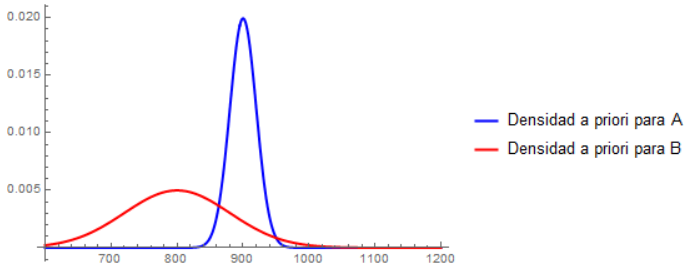
Veamos essta influencia con un ejemplo.

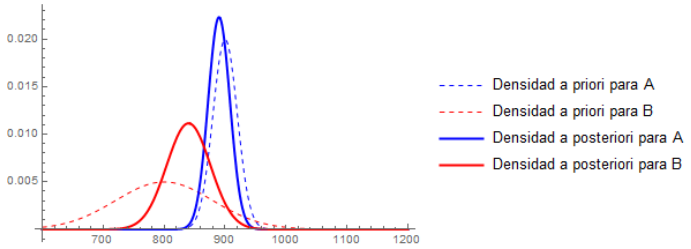
Ejemplo 4.21. Supongamos que dos físicos A y B están interesados en obtener estimaciones lo más precisas posibles de una constante física θ conocida previamente solo de forma aproximada. Supongamos que la opinión del físico A, muy familiarizado con este estudio, es que $\theta_A \sim (900, \sigma_A = 20)$. Supongamos que la opinión del físico B, con poca experiencia en este área, es que $\theta_B \sim N(800, \sigma_B = 80)$.

Supogamos que un método experimetal de medida está disponible, y que tenemos una observación $y|\theta \sim (\theta, \sigma = 40)$, tomando y el valor 850. Calulemos la distribución a posteriori en ambos casos. Estos cálculos ya los habíamos hecho anteriormente, y tenemos que

$$\theta_A|y \sim N(890, 17'9), \quad \theta_B|y \sim N(840, 35'7),$$

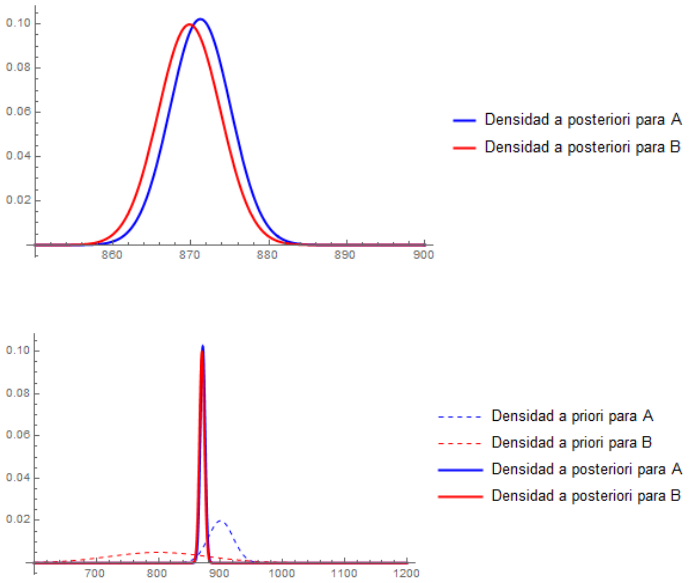
donde el segundo parámetro es la desviación típica.





Supogamos ahora que tenemos una muestra $\vec{y} = (y_1, \ldots, y_{100})$ y que $\overline{y} = 870$. Ahora tendríamos que

$$\theta_A|\vec{y} \sim N(871'2, 3'9), \quad \theta_B|\vec{y} \sim N(869'8, 3'995),$$



4.14. Inferencia Clásica vs. Inferencia Bayesiana

Clásicos	**Bayesianos**
Parámetros fijos	Parámetros variables
Probabilidad como frecuencia límite	Probabilidad como incertidumbre
No incluye información previa	Inclusión de información previa
Estimadores de máxima verosimilitud	La estiamción es un problema de decisión
Intervalos de confianza	Intervalos de credibilidad
Método de muestreo muy importante	El método de muestreo no importa

Ventajas métodos bayesianos

- Permiten una inferencia más natural y útil que los métodos clásicos.
- Tienen una interpretación más directa que los intervalos de confianza, contrastes de hipótesis. clásicos y p-valor.
- Hacen uso de mayor cantidad de información disponible, lo que suele implicar resultados más consistentes que los obtenidos por el método clásico.
- Permiten ir actualizando resultados a medida que se incorpora nueva información.
- Permiten resolver problemas más complejos que los métodos clásicos.
- Son fundamentales para resolver problemas de decisión, mientras que los métodos clásicos están limitados a análisis estadíssticos que informan de la decisión sólo indirectamente.
- Son muy útiles en el caso de tamaños muestrales pequeños. No asumen mustras infinitas ni normalidad.

Crísticas métodos bayesianos

- Los métodos bayesianos incorporan un elemento de subjetividad que no aparece de forma directa en los métodos clásicos.
- Las conclusiones dependenn de la selección de la distribución a priori.
- Se puede obtener resultados complejos que requiere del uso de métodos computacionales.
- En la práctica, la información extra que los métodos bayesianos utilizan es difícil de especificar de forma exacta.

Parte III

Inferencia no paramétrica

Capítulo 5

Inferencia no paramétrica

5.1. Introducción

Hasta ahora los tests de hipótesis han sido utilizados para contrastar la veracidad de hipótesis acerca de los parámetros de la distribución de una población. Sin embargo, en muchas ocasiones, es necesario emitir un juicio estadísstico sobre la distribución poblacional en su conjunto.

Los problemas de este tipo que se plantean de manera habitual son los siguientes:

- Decidir, a la vista de una muestra aleatoria de una población, si puede admitirse que la distribución poblacional coincide con una cierta distribución dada o pertenece a un determinado tipo de distribuciones ⇒ **Contrastes de bondad de ajuste**.

- Analizar si varias muestras aleatorias provienen de poblaciones con la misma distribución teórica, de forma que puedan ser utilizadas conjuntamente para inferencia posteriores acerca de ésta; o, por el contrario, son muestras de poblaciones con distinta distribución, que no pueden agruparse como información homogénea acerca de una única distribución ⇒ **Contrastes de homogeneidad**.

- Estudiar, en el caso en que se observen dos o más características de los elementos de la población si las características observadas pueden considerarse independientes, y se puede proceder a su análisis por separado, o, por el contrario, existe relación estadística entre ellas ⇒ **Contrastes de independencia**.

5.2. Contrastes de bondad de ajuste

Vamos a centrarnos en los siguientes contrastes:

- Contraste χ^2 de bondad de ajuste (primer caso).

- Contraste χ^2 de bondad de ajuste (segundo caso).

Contraste χ^2 de bondad de ajuste (primer caso)

Consideramos un muestra aleatoria $(X_1, \ldots, X_n)$ de una variable aleatoria X con distribución desconocida. Queremos dar respuesta a:

- ¿A la vista de la muestra, es razonable admitir que la distribución de X viene dada por un determinado modelo de probabilidad P?

- ¿Se ajustan bien los datos a P?

Resolveremos el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \text{El modelo de probabilidad de } X \text{ es } P \\ H_1 : \text{El modelo de probabilidad de } X \text{ no es } P \end{cases}$$

El modelo P debe estar completamente especificado.

Para contrastar H_0 frente a H_1 hacemos una partición (arbitraria) del espacio muestral de la población (posibles valores de X) en k clases $A_1, \ldots, A_k$. Después, para cada A_i, $i = 1, \ldots, k$, consideramos las siguientes frecuencias (absolutas)

$$\begin{aligned} O_i &= \text{frecuencia observada en } A_i \\ &= \text{número de elementos de la muestra que están en la clase } A_i \end{aligned}$$

$$\begin{aligned} e_i &= \text{frecuencia esperada de la clase } A_i \text{ si la hipótesis } H_0 \text{ es cierta} \\ &= n \cdot P(A_i) \end{aligned}$$

El estadístico que utilizaremos es:

$$\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i},$$

que tiene aproximadamente (cuando n es grande) una distribución χ^2_{k-1} si H_0 es cierta.

Para que la aproximación sea razonablemente "buena", además de tener una muestra suficientemente grande, es necesario que el valor esperado de cada clase sea "suficientemente grande". Si la muestra procede de P, es de esperar que haya valores parecidos para O_i y e_i, y por tanto, este estadístico debería tomar valores próximos a cero. En consecuencia, rechazaremos la hipótesis nula cuando los valores de este estadístico sean "grandes" y la aceptaremos cuando sean "pequeños".

Rechazaremos H_0 a nivel de significación α si

$$\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} > \chi^2_{k-1,1-\alpha}.$$

En caso contrario, aceptaremos H_0 a nivel de significación α.

Contraste χ^2 de bondad de ajuste (segundo caso)

El contraste de bondad de ajuste se puede plantear también en una situación más general. Consideramos un muestra aleatoria $(X_1, \ldots, X_n)$ de una variable aleatoria X con distribución desconocida. Queremos dar respuesta a:

- ¿A la vista de la muestra, es razonable admitir que la distribución de X viene dada por algún modelo de probabilidad de la familia P_θ, donde $\theta = (\theta_1, \ldots, \theta_r)$?
- ¿Se ajustan bien los datos a un modelo de probabilidad de la familia $\{P_\theta : \theta \in \Theta\}$?

Resolveremos el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \text{El modelo de probabilidad de } X \text{ es de la familia } \{P_\theta : \theta \in \Theta\} \\ H_1 : \text{El modelo de probabilidad de } X \text{ no es de la familia } \{P_\theta : \theta \in \Theta\} \end{cases}$$

Para contrastar H_0 frente a H_1 hacemos una partición (arbitraria) del espacio muestral de la población (posibles valores de X) en k clases $A_1, \ldots, A_k$. Después, para cada A_i, $i = 1, \ldots, k$,

consideramos las siguientes frecuencias (absolutas)

$$\begin{aligned} O_i &= \text{frecuencia observada en } A_i \\ &= \text{número de elementos de la muestra que están en la clase } A_i \end{aligned}$$

$$\begin{aligned} e_i &= \text{frecuencia esperada de la clase } A_i \text{ si la hipótesis } H_0 \text{ es cierta} \\ &= n \cdot P_\theta(A_i) \approx n \cdot P_{\widehat{\theta}}(A_i) \end{aligned}$$

donde $\widehat{\theta}$ es el estimador de máxima verosimilitud de θ.

El estadístico que utilizaremos es:

$$\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i},$$

que tiene aproximadamente (cuando n es grande) una distribución χ^2_{k-r-1} si H_0 es cierta.

Rechazaremos H_0 a nivel de significación α si

$$\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} > \chi^2_{k-r-1,1-\alpha}.$$

En caso contrario, aceptaremos H_0 a nivel de significación α.

5.3. Contrastes de homogeneidad

Contraste χ^2 de homogeneidad

Supongamos que disponemos de p muestras aleatorias independientes tomadas de p poblaciones

$$\begin{gathered} (X_{11}, X_{12} \ldots, X_{1n_1}) \\ (X_{21}, X_{22} \ldots, X_{2n_2}) \\ \vdots \\ (X_{p1}, X_{p2} \ldots, X_{pn_p}) \end{gathered}$$

y sea $n = n_1 + n_2 + \ldots + n_p$. Queremos ver si, a la vista de las muestras obtenidas, es razonable admitir que todas las poblaciones tienen una distribución común, es decir, queremos ver si son poblaciones homogéneas. Por tanto, tenemos:

$$\begin{cases} H_0 : \text{las } p \text{ poblaciones tienen una distribución común} \\ H_1 : \text{las } p \text{ poblaciones no tienen una distribución común} \end{cases}$$

Para contrastar H_0 frente a H_1 hacemos nuevamente una partición (arbitraria) del espacio muestral común a las p poblaciones en k clases $A_1, \ldots, A_k$. Después, definimos para la clase A_i, $i = 1, \ldots, k$ y para la muestra de la población j-ésima ($j = 1, \ldots, p$)

$$O_{ij} = \text{frecuencia observada en la clase } A_i \text{ con la } j\text{-ésima muestra}$$

$$\begin{aligned} e_{ij} &= \text{frecuencia esperada en la clase } A_i \text{ con la } j\text{-ésima muestra si la hipótesis } H_0 \text{ es cierta} \\ &= n_j \cdot P_\theta(A_i) \end{aligned}$$

Entonces, tenemos para la muestra j-ésima

$$\sum_{i=1}^{k} \frac{(O_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{k-1}, \quad \text{(aproximadamente) si } H_0 \text{ es cierta}$$

Si sumamos los p estadísticos obtenidos, uno para cada muestra, tenemos:

$$\sum_{j=1}^{p}\sum_{i=1}^{k}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{p(k-1)}, \quad \text{(aproximadamente) si } H_0 \text{ es cierta}$$

Sin embargo, rodavía tenemos un problema por resolver: el valor de este se podría calcular si supiéramos cuál es la distribución P común a las p poblaciones. Normalmente, lo único que queremos contrastar es si tienen una distribución común, pero sin que sepamos, ni nos importe, cuál es esa distribución común puede ser cualquiera). Por tanto, tenemos que estimar $P(A_i)$, $i = 1, \ldots, k$, a partir de las observaciones. Esta estimación se hace mediante:

$$\widehat{P(A_i)} = \frac{\sum_{j=1}^{p} O_{ij}}{n}, \quad j = 1, \ldots, k.$$

Las frecuencias esperadas serán, entonces

$$e_{ij} = n_j\widehat{P(A_i)} = n_j\frac{\sum_{j=1}^{p} O_{ij}}{n} = \frac{\left(\sum_{i=1}^{k} O_{ij}\right)\left(\sum_{j=1}^{p} O_{ij}\right)}{n}$$

En definitiva, el estadístico utilizdo es, para los valores de las frecuencias esperadas dadas anteriormente

$$\sum_{j=1}^{p}\sum_{i=1}^{k}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{(p-1)(k-1)}, \quad \text{(aproximadamente) si } H_0 \text{ es cierta}$$

Rechazaremos H_0 a nivel de significación α si

$$\sum_{j=1}^{p}\sum_{i=1}^{k}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} > \chi^2_{(p-1)(k-1),1-\alpha}.$$

En caso contrario, aceptaremos H_0 a nivel de significación α.

5.4. Contrastes de independencia

Contraste χ^2 de independencia

Supongamos que queremos estudiar si dos características X e Y de una población están relacionadas o no. Para hacer este estudio, obtenemos una muestra aleatoria de n pares de valores de estas características

$$((X_1, Y_1), \ldots, (X_n, Y_n)).$$

Queremos ver si, a la vista de la muestra tiene sentido admitir que X e Y son independientes Por tanto, tenemos el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : X \text{ e } Y \text{ son independientes} \\ H_1 : X \text{ e } Y \text{ no son independientes} \end{cases}$$

Tomamos una partición (arbitraria) del espacio muestral (correspondientes a los posibles valores de X e Y) en $k \cdot p$ clases $A_1 \times B_1, A_1 \times B_2, \ldots, A_k \times B_p$. Estas $k \cdot p$ clases corresponden a tomar clases $A_1, \ldots, A_k$ para la característica X, y las clases $B_1, \ldots, B_p$ para la característica Y.

Llamamos

$$O_{ij} = \text{frecuencia osbervada en la clase } A_i \times B_j$$

$$\begin{aligned} e_{ij} &= \text{frecuencia esperada en la clase } A_i \times B_j \text{ si la hipótesis } H_0 \text{ es cierta} \\ &= n \cdot P(A_i) \cdot P(B_j) \end{aligned}$$

Entonces tenemos que

$$\sum_{j=1}^{p}\sum_{i=1}^{k}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{(p-1)(k-1)}, \quad \text{(aproximadamente) si } H_0 \text{ es cierta}$$

Pero otra vez tenemos el mismo problema de antes: los valores de $P(A_i)$ y $P(B_j)$ tienen que ser estimados a partir de la muestra; esto se hace de la forma:

$$\widehat{P(A_i)} = \frac{\sum_{j=1}^{p} O_{ij}}{n}, \quad \widehat{P(B_j)} = \frac{\sum_{i=1}^{k} O_{ij}}{n}.$$

Las frecuencias esperadas serán, entonces

$$e_{ij} = n \cdot \widehat{P(A_i)} \cdot \widehat{P(B_j)} = n \cdot \frac{\sum_{j=1}^{p} O_{ij}}{n} \cdot \frac{\sum_{i=1}^{k} O_{ij}}{n} = \frac{\left(\sum_{i=1}^{k} O_{ij}\right)\left(\sum_{j=1}^{p} O_{ij}\right)}{n}$$

En definitiva, el estadístico utilizdo es, para los valores de las frecuencias esperadas dadas anteriormente

$$\sum_{j=1}^{p}\sum_{i=1}^{k}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{(p-1)(k-1)}, \quad \text{(aproximadamente) si } H_0 \text{ es cierta}$$

Rechazaremos H_0 a nivel de significación α si

$$\sum_{j=1}^{p}\sum_{i=1}^{k}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} > \chi^2_{(p-1)(k-1),1-\alpha}.$$

En caso contrario, aceptaremos H_0 a nivel de significación α.

Observación 5.1. Los contrastes χ^2 tienen los siguientes inconvenientes:

- Son poco precisos para muestras pequeñas por ser tests asintóticos.
- Para variables continuas, se desprecia información al agrupar datos en clases..

Capítulo 6

Estimación no paramétrica de densidades

En la estimación de la densidad, como en la Inferencia en general, existen dos posibles vías de estudio.

- **Estimación paramétrica**, en la que se asume determinada distribución de la variable y se emplean datos para la estimación de los correspondientes parámetros.
- **Estimación no paramétrica**, que no asume ninguna distribución, únicamente utiliza la información proporcionada por la muestra.

Tanto en la inferencia paramétrica como la no paramétrica poseen numerosos simpatizantes y detractores, pues ambas metodologías de trabajo tienen ventajas e inconvenientes que han sido ampliamente estudiados a lo largo de los años.

La suposicion inicial de que la población de la que proceden los datos sigue un modelo paramétrico puede limitar mucho el ajuste del modelo. En caso deser correcta dicha suposición, el ajuste será muy bueno, pero si el modelo paramétrico es incorrecto, las conclusiones podrían ser totalmente erróneas. Por ello, es deseable considerar técnicas no paramétricas que olviden cualquier hipótesis previa y trabajen únicamente con la información que proporcionan los datos, teniendo siempre presente la aleatoriedad intrínseca a los mismos.

Las principios de la estimación no paramétrica de la densidad datan de finales del siglo XIX, cuando Karl Pearson introdujo el **histograma**, que no es más que la representación de las frecuencias por clases. El histograma es un estimador discontinuo, que además depende de la elección de un punto inicial y de un parámetro ventana, con gran influencia por parte de ambos en el resultado final.

Para solventar el problema de la dependencia del punto inicial, hay que esperar hasta mediados del siglo XX, cuando se desarrolló el denominado **histograma móvil** o **estimador naive**, que sigue siendo discontinuo y dependiente de la ventana. Posteriormente, Parzen, en 1962, y Rosenblatt (1956), propusieron el estimador tipo núcleo, que sí es continuo y que, por lo tanto, en la mayor parte de las veces, se ajusta mejor a la realidad de los modelos estudiados, aunque también depende en gran medida de la elección de un parámetro ventana.

En la literatura estadística ha sido ampliamente estudiado el papel fundamental del parámetro ventana en el estimador tipo núcleo. Dicho parámetro es el que controla el grado de suavización del estimador, y una mala elección del mismo puede derivar en un estimador tanto infra como sobresuavizado. Debido a esto, la segunda mitad del siglo XX fue muy prolífica en cuanto a métodos de seleccioón de ventana, entre los que destacan el propuesto por Silverman (1986), el meétodo de Shealter y Jones (1991) y el de Bowman (1984).

Definición 6.1. Sea $X_1, \ldots, X_n$ una muestra de datos que proviene de una variable aleatoria X continua con función de densidad desconocida, f. El **estimador de núcleos** de f se define por

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h},\right)$$

donde K es una función, denominada **función kernel, función núcleo o función peso**, que satisface ciertas condiciones de regularidad, generalmente es una función de densidad simétrica, con media 0 y varianza 1, $h > 0$ es el **parámetro de suavizado o ancho de banda**.

El estimador de núcleos se puede ver como una suma de pequeñas montañas o protuberancias situadas en las observaciones. La funcion núcleo determina la forma de estas protuberancias, mientras que el paámetro de suavizado determina su anchura.

Cada pequeña montaña o protuberancia está centrada en una observación X_i y tiene una superficie de $1/n$.

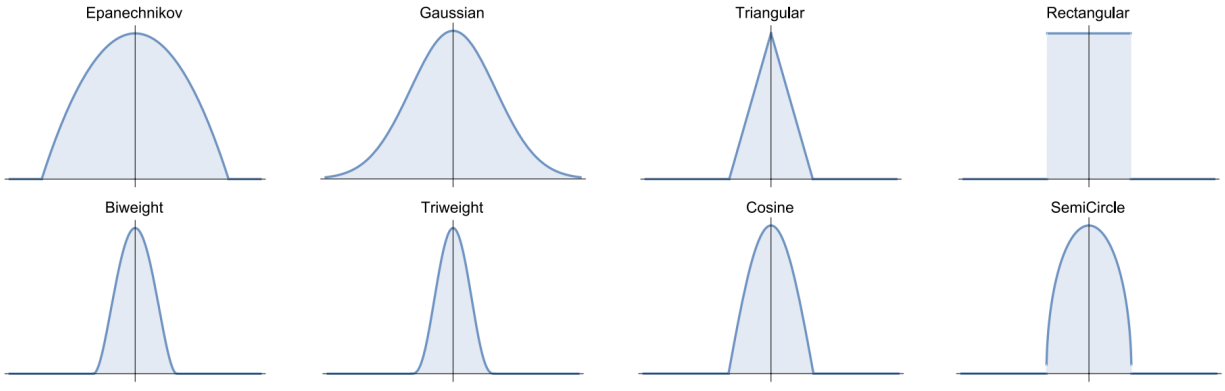
Si h es demasiado pequeño, las montañas estarán muy separadas y observaremos muchos picos.

Si h es demasiado grande, observaremos una sola montaña plana.

Un valor intermedio de h debería ser la mejor elección.

Algunas de las funciones núcleo más comunes son

- **Epanechnikov**, $K(t) = \frac{3}{4}(1 - t^2)$, $|t| < 1$.
- **Gauss**, $K(t) = \frac{1}{2\pi} e^{-\frac{t^2}{2}}$, $t \in \mathbb{R}$.
- **Triangular**, $K(t) = 1 - |t|$, $|t| < 1$.
- **Rectangular**, $K(t) = \frac{1}{2}$, $|t| < 1$.
- **Biweight**, $K(t) = \frac{15}{16}(1 - t^2)^2$, $|t| < 1$.
- **Triweight**, $K(t) = \frac{35}{32}(1 - t^2)^3$, $|t| < 1$.
- **Coseno**, $K(t) = \frac{\pi}{4} \cos\left(\frac{\pi t}{2}\right)$, $|t| < 1$.
- **Semicírculo**, $K(t) = \frac{2}{\pi}\sqrt{1 - t^2}$, $|t| < 1$.



Apliquemos todo esto a un ejemplo real. Los datos que vamos a analizar fueron analizados en Azzalini y Bowman (1990), ("Applied Smoothing Techniques for Data Analysis") quienes registraron el tiempo (en minutos) que dura una erupcion´ del geyser Old Faithful que se encuentra en el parque nacional de Yellowstone (Wyoming, USA). Las medidas (27 erupciones en total) fueron tomadas entre el 1 y el 15 de Agosto de 1985.

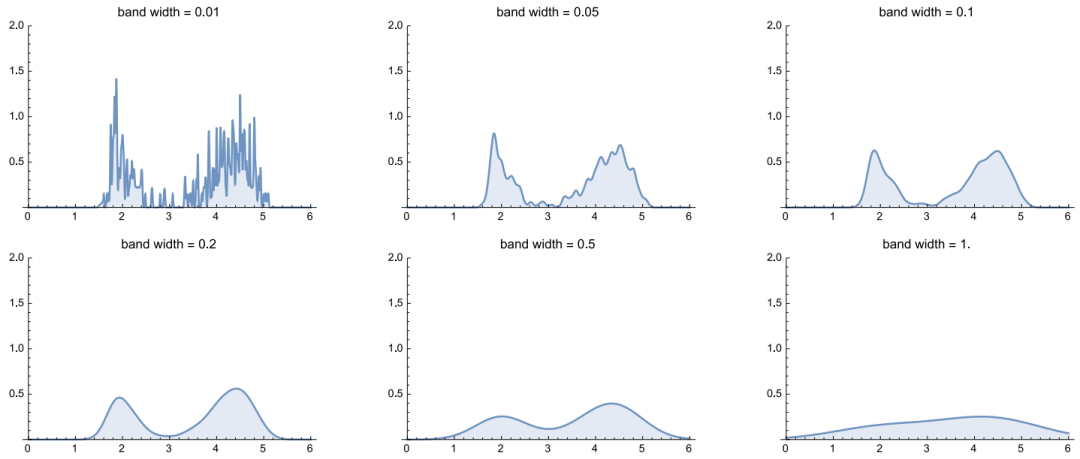


Figura 6.1: Ajustando diferentes valores de h con la función núcleo que usa Mathmatica por defecto.

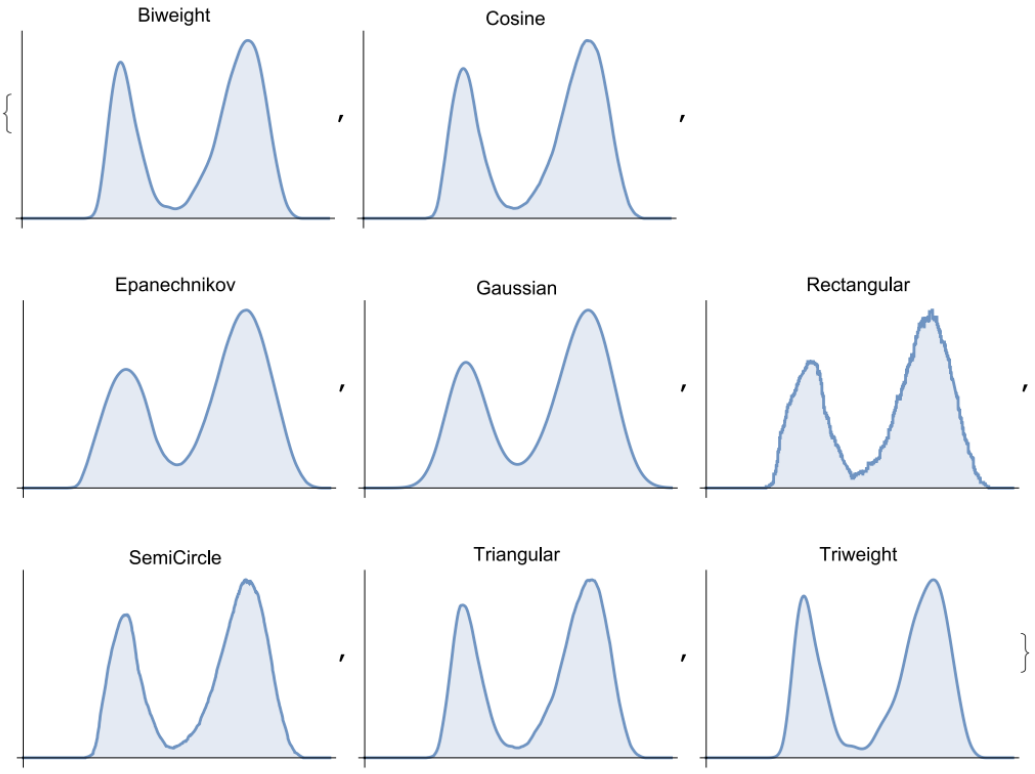


Figura 6.2: Funciónes núcleo vistas anteriormente.