Análisis de datos e inferencia

20 de diciembre de 2022

Índice general

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 e hipótesis

Sea X una variable aleatoria cuantitativa, Y una variable aleatoria continua y $(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$ un conjunto de datos. Entonces el modelo de regresión lineal simple es

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

Hipótesis del modelo

1. $E(u_i) = 0, \quad \forall i = 1, \dots n.$
2. $V(u_i) = \sigma^2, \quad \forall i = 1, \dots n$ (homocedasticidad)

3. $u_i \sim N(0, \sigma^2), \quad \forall i = 1, \dots n$ (normalidad)
4. $E(u_i u_j) = 0, \quad \forall i \neq j$ (independencia)

Nota. En realidad, la cuarta hipótesis es de incorrelación ($Cov(u_i, u_j) = 0$).

$$Cov(u_i, u_j) = E(u_i u_j) - E(u_i)E(u_j) = E(u_i u_j)$$

Sin embargo, bajo normalidad la incorrelación y la independencia son equivalentes.

Podemos escribir las mismas hipótesis en términos de $y_i, \quad \forall i = 1, \dots n$.

1. $E(y_i|x_i) = E(\beta_0 + \beta_1 x_i + u_i) = \beta_0 + \beta_1 x_i, \quad \forall i = 1, \dots n$ (linealidad)
2. $V(y_i|x_i) = V(\beta_0 + \beta_1 x_i + u_i) = \sigma^2, \quad \forall i = 1, \dots n$ (homocedasticidad)
3. $y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \quad \forall i = 1, \dots n$ (normalidad)
4. $Cov(y_i, y_j) = 0, \quad \forall i \neq j$ (independencia)

Podemos dar un significado real a β_0 y β_1:

- β_0 es el valor medio de la variable Y cuando x_i toma el valor 0.

$$E(y_i|x_i = 0) = \beta_0, \quad i = 1, \dots n$$

- β_1 es la variación media que experimenta la variable Y cuando X_i aumenta en una unidad.

$$E(y_i|x_i + 1) - E(y_i|x_i) = \beta_1, \quad i = 1, \dots n$$

1.3. Estimación de los parámetros

Queremos estimar β_0, β_1 y σ^2. Con los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ podemos estimar

$$\hat{E}(y_i|x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \dots n$$

Método de máxima verosimilitud

$y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, \dots, n$, así que podemos encontrar estimadores de máxima verosimilitud para los parámetros y para σ^2.

Usando el método de máxima verosimilitud llegamos las ecuaciones normales de la regresión:

$$\begin{cases} \dfrac{\partial \log L}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \dfrac{\partial \log(L)}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

Notación. $\hat{y}_i = \hat{E}(y_i|x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Si definimos el error o residuo como $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, podemos escribir las ecuaciones normales de regresión de la siguiente forma:

$$\begin{cases} \sum_{i=1}^n e_i = 0 \\ \sum_{i=1}^n x_i e_i = 0 \end{cases}$$

Resolviendo este sistema, obtenemos los estimadores:

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$$
$$\hat{\beta}_0 = \bar{y} - \frac{s_{XY}}{s_X^2}\bar{x}$$
$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n}\sum_{i=1}^n e_i^2$$

La ecuación de la recta resultante es:

$$\hat{y_i} = \bar{y} + \frac{s_{XY}}{s_X^2}(x_i - \bar{x})$$

Estimación por mínimos cuadrados

Queremos minimizar la suma de los cuadrados de los errores $\sum_{i=1}^n e_i^2$, donde $e_i = y_i - \hat{y_i}$. Para ello minimizamos la función $M(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$.

$$\begin{cases} \frac{\partial M}{\partial \beta_0}(\beta_0, \beta_1) = -2\sum(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial M}{\partial \beta_1}(\beta_0, \beta_1) = -2\sum x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

Simplificando obtenemos las ecuaciones normales de la regresión, como antes. Así que los estimadores de β_0 y β_1 por máxima verosimilitud coinciden con los estimadores por mínimos cuadrados.

Estimación de la varianza

Partiendo del estimador $\bar{\sigma^2} = \frac{1}{n}\sum_{i=1}^n e_i^2$ obtenido previamente, podemos llegar a una expresión equivalente:

$$\hat{\sigma^2} = s_Y^2 - \frac{s_{XY}^2}{s_X^2}$$

Veamos si este estimador es insesgado calculando su esperanza.

$$E(\hat{\sigma^2}) = E(\frac{\sum_{i=1}^n e_i^2}{n}) = \frac{1}{n}E(\sum_{i=1}^n e_i^2) = \frac{1}{n}\sigma^2(n-2)$$

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

Observamos que este estimador no es insesgado. Consideramos entonces:

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

Este sí es un estimador insesgado de σ^2 y le llamamos varianza residual. Tenemos la relación $s_R^2 = \frac{n}{n-2}\hat{\sigma^2}$.

1.4. Propiedades de los estimadores

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

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

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

- $E(\hat{\beta}_1) = \beta_1$ (estimador insesgado)
- $V(\hat{\beta}_1) = \frac{\sigma^2}{n s_X^2}$

Por tanto, $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{n s_X^2})$.

De forma análoga, podemos escribir:

$$\hat{\beta}_0 = \sum_{i=1}^n (\frac{1}{n} - \bar{x} w_i)$$

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

- $E(\hat{\beta}_0) = \beta_0$ (estimador insesgado)
- $V(\hat{\beta}_0) = \frac{\sigma^2}{n}(1 + \frac{\bar{x}^2}{s_X^2})$

Por tanto, $\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{n}(1 + \frac{\bar{x}^2}{s_X^2}))$.

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:

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

1.5. Intervalos de confianza para los parámetros

Intervalos de confianza para β_1

Caso 1: σ^2 conocida

Sabemos que $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{ns_X^2})$. Entonces:

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

Por tanto, el intervalo de confianza para β_1 a nivel de significación α es:

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

donde $z_{1-\frac{\alpha}{2}}$ es el percentil de orden $(1-\frac{\alpha}{2})100\,\%$ de una variable aleatoria $Z \sim N(0,1)$.

Caso 2: σ^2 desconocida

$$\begin{cases} \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{ns_X^2}}} \sim N(0,1) \\ \frac{\sum_{i=1}^n e_i^2}{\sigma^2} = \frac{(n-2)s_R^2}{\sigma^2} \sim \chi^2_{n-2} \end{cases} \Rightarrow \frac{\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{ns_X^2}}}}{\sqrt{\frac{(n-2)s_R^2}{\sigma^2}\frac{1}{n-2}}} = \frac{\hat{\beta}_1 - \beta_1}{\frac{s_R}{\sqrt{ns_X^2}}} \sim t_{n-2}$$

Luego el intervalo de confianza para β_1 a nivel de significación α es:

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

donde $s_R = +\sqrt{s_R^2}$ y $t_{n-2,1-\frac{\alpha}{2}}$ es el percentil de orden $(1-\frac{\alpha}{2})100\,\%$ de una variable aleatoria $T \sim t_{n-2}$.

Intervalos de confianza para β_0

Caso 1: σ^2 conocida

Sabemos que $\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}}{s_X^2}\right)\right)$. Entonces, el intervalo de confianza para β_0 a nivel de significación α es:

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

donde $z_{1-\frac{\alpha}{2}}$ es el percentil de orden $(1-\frac{\alpha}{2})100\,\%$ de una variable aleatoria $Z \sim N(0,1)$.

Caso 2: σ^2 desconocida

Razonando de forma análoga al caso de β_1, tenemos que:

$$\frac{\hat{\beta}_0 - \beta_0}{\frac{s_R}{n}\sqrt{1 + \frac{\bar{x}^2}{s_X^2}}}$$

Por tanto, el intervalo de confianza para β_0 a nivel de significación α es:

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

donde $t_{n-2,1-\frac{\alpha}{2}}$ es el percentil de orden $(1-\frac{\alpha}{2})100\,\%$ de una variable aleatoria $T \sim t_{n-2}$.

Intervalos de confianza para σ^2

Sabemos que $\frac{\sum_{i=1}^n e_i^2}{\sigma^2} = \frac{(n-2)s_R^2}{\sigma^2} \sim \chi^2_{n-2}$. Queremos que $P(a < \sigma^2 < b) = 1 - \alpha$.

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

Luego:

$$\begin{cases} \frac{(n-2)s_R^2}{b} = \chi^2_{n-2,\frac{\alpha}{2}} \Rightarrow b = \frac{(n-2)s_R^2}{\chi^2_{n-2,\frac{\alpha}{2}}} \\ \frac{(n-2)s_R^2}{a} = \chi^2_{n-2,1-\frac{\alpha}{2}} \Rightarrow a = \frac{(n-2)s_R^2}{\chi^2_{n-2,1-\frac{\alpha}{2}}} \end{cases}$$

Por tanto, el intervalo de confianza para σ^2 a nivel de significación α tiene por extremos a y b, es decir:

$$IC_{1-\alpha}(\sigma^2) = (a, b)$$

1.6. Contraste de la regresión

Consideramos el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \beta_1 = 0 \Leftrightarrow E(y|x) = \beta_0 \\ H_1 : \beta_1 \neq 0 \Leftrightarrow E(y|x) = \beta_0 + \beta_1 x \end{cases}$$

Fijamos el nivel de significación α. Podemos resolverlo de cuatro formas distintas.

Intervalos de confianza

Sea $IC_{1-\alpha}(\beta_1)$ el intervalo de confianza para β_1 a nivel de significación α. Entonces:

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

Estadístico T

Sabemos que $\frac{\hat{\beta_1}-\beta_1}{\frac{s_R}{\sqrt{ns_X^2}}} \sim t_{n-2}$. Entonces $T = \frac{\hat{\beta_1}}{\frac{s_R}{\sqrt{ns_X^2}}} \sim t_{n-2}$ si H_0 es cierto.

Tomamos un t_{exp}.

- Si $t_{exp} \in (-t_{n-2,1-\frac{\alpha}{2}}, t_{n-2,1-\frac{\alpha}{2}})$, o equivalentemente $|t_{exp}| \leq t_{n-2,1-\frac{\alpha}{2}}$, aceptamos H_0 a nivel de significación α.
- En caso contrario, rechazamos H_0 a nivel de significación α.

Valor p

Sea p el valor p o p-valor de la distribución. Entonces:

- Si $p \geq \alpha$, aceptamos H_0 a nivel de significación α.
- En caso contrario, rechazamos H_0 a nivel de significación α.

Tabla ANOVA

Partimos de que podemos escribir:

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

Definimos:

- Variabilidad total:
$$VT = \sum_{i=1}^{n}(y_i - \bar{y})^2 = ns_Y^2$$
- Variabilidad no explicada:
$$VNE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = (n-2)s_R^2 = n\hat{\sigma^2} = n(s_Y^2 - \hat{\beta_1}^2 s_X^2)$$
- Variabilidad explicada:
$$VE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = VT - VNE = n\hat{\beta_1}^2 s_X^2$$

Observamos que:

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

Además, como $\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{ns_X^2}}} \sim N(0,1)$, entonces:

$$\frac{(\hat{\beta}_1 - \beta_1)^2}{\frac{\sigma^2}{ns_X^2}} \sim \chi^2_1$$

Luego $\frac{\hat{\beta_1}^2}{\frac{\sigma^2}{ns_X^2}} = \frac{n\hat{\beta_1}^2 s_X^2}{\sigma^2} = \frac{VE}{\sigma^2} \sim \chi^2_1$ si H_0 es cierta.

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

Tomamos un F_{exp}.

- Aceptamos H_0 a nivel de significación α si $F_{exp} \leq F_{1,n-2,1-\alpha}$.
- En caso contrario, rechazamos H_0 a nivel de significación α.

La tabla ANOVA es de la forma:

Fuentes	Suma de cuadrados	Grados de libertad	Cocientes
VE	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	$VE/1$
VNE	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$VNE/(n-2)$
VT	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$	

También se incluyen columnas para F_{exp} y p-valor.

Observación. Existe la siguiente relación entre t_{exp} y F_{exp}:

$$t_{exp}^2 = F_{exp}$$

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 $VNE = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 = \sum_{i=1}^{n} e_i^2 = 0$. Por tanto, $e_i = 0$ para todo $i = 1, \dots n$, 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 1.1. *El coeficiente de determinación coincide con el coeficiente de correlación lineal al cuadrado. Es decir,*

$$r^2 = R^2$$

Demostración.

$$R^2 = \frac{VE}{VT} = \frac{n\hat{\beta_1}^2 s_X^2}{n s_Y^2} = \frac{\left(\frac{s_{XY}}{s_X^2}\right)^2 s_X^2}{s_Y^2} = \frac{s_{XY}^2}{s_X^2 s_Y^2} = r^2$$

□

1.8. Predicción

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:

$$\hat{m_0} = \hat{E}(y|x = x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Teorema 1.2.

$$\hat{m_0} \sim N\left(m_0, \frac{\sigma^2}{n}\left(1 + \frac{(x_0 - \bar{x})^2}{s_X^2}\right)\right)$$

Intervalos de confianza para m_0

Podemos calcular los intervalos de confianza para m_0 con nivel de confianza de $100(1-\alpha)\,\%$.

Si σ^2 es conocida,

$$IC_{1-\alpha}(m_0) = \left(\hat{m_0} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{1 + \frac{(x_0 - \bar{x})^2}{s_X^2}}, \right.$$
$$\left. \hat{m_0} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{1 + \frac{(x_0 - \bar{x})^2}{s_X^2}} \right)$$

Si σ^2 es desconocida,

$$IC_{1-\alpha}(m_0) = \left(\hat{m_0} - t_{n-2,1-\frac{\alpha}{2}} \frac{s_R}{\sqrt{n}} \sqrt{1 + \frac{(x_0 - \bar{x})^2}{s_X^2}}, \right.$$
$$\left. \hat{m_0} + t_{n-2,1-\frac{\alpha}{2}} \frac{s_R}{\sqrt{n}} \sqrt{1 + \frac{(x_0 - \bar{x})^2}{s_X^2}} \right)$$

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 $\hat{y_0} = \hat{\beta_0} + \hat{\beta_1} x_0$. La estimación puntual es:

$$\hat{y_0} = \hat{\beta_0} + \hat{\beta_1} x_0 = \hat{m_0}$$

Consideramos el error:

$$e_0 = y_0 - \hat{y_0} = \beta_0 + \beta_1 x_0 + u_0 - (\hat{\beta_0} + \hat{\beta_1} x_0)$$

que también es una variable aleatoria.

Teorema 1.3.

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

Intervalos de pronóstico para y_0

Podemos calcular los intervalos de pronóstico para y_0 con contenido probabilístico $1 - \alpha$.

Si σ^2 es conocida,

$$IP_{1-\alpha}(y_0) = \left(\hat{y_0} - z_{1-\frac{\alpha}{2}} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_X^2}}, \right.$$
$$\left. \hat{y_0} + z_{1-\frac{\alpha}{2}} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_X^2}} \right)$$

Si σ^2 es desconocida,

$$IP_{1-\alpha}(y_0) = \left(\hat{y_0} - t_{n-2,1-\frac{\alpha}{2}} s_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_X^2}}, \right.$$
$$\left. \hat{y_0} + t_{n-2,1-\frac{\alpha}{2}} s_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_X^2}} \right)$$

1.9. Análisis de residuos y observaciones atípicas e influyentes

Residuos

El residuo de un dato es la diferencia entre su valor y la predicción mediante el modelo.

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

El análisis de los residuos puede darnos información sobre el ajuste del modelo.

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 e hipótesis

Sean $X_1, \ldots, X_n$ variables explicativas, Y una variable aleatoria continua y $(x_{11}, x_{21}, \ldots, x_{k1}, y_1), \ldots, (x_{1n}, x_{2n}, \ldots, x_{kn}, y_n)$ un conjunto de datos. Entonces el modelo de regresión lineal múltiple es:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i, \quad i = 1, \ldots, n$$

Hipótesis del modelo

- $E(u_i) = 0, \quad \forall i = 1, \ldots n.$
- $V(u_i) = \sigma^2, \quad \forall i = 1, \ldots n$ (homocedasticidad)
- $u_i \sim N(0, \sigma^2), \quad \forall i = 1, \ldots n$ (normalidad)
- $E(u_i u_j) = 0, \quad \forall i \neq j$ (independencia)
- $n > k + 1$
- No existen relaciones lineales entre los X_i (ausencia de multicolinealidad)

El modelo se puede escribir de forma matricial. Definimos:

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

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{pmatrix} \in \mathcal{M}_{n\times(k+1)}$$

Entonces el modelo es equivalente a:

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

Las hipótesis del modelo se pueden reescribir como:

- $\vec{u} \sim N_n(\vec{0}, \sigma^2 I_n)$
- $n > k + 1$
- Ausencia de multicolinealidad.

Podemos escribir las mismas hipótesis iniciales en términos de $y_i, \quad \forall 1, \dots, n$.

1. $E(y_i|x_{1i}, \dots, x_{ki}) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}, \quad \forall i = 1, \dots n$ (linealidad)
2. $V(y_i|x_{1i}, \dots, x_{ki}) = \sigma^2, \quad \forall i = 1, \dots n$ (homocedasticidad)
3. $y_i|x_{1i}, \dots, x_{ki} \sim N(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k xki, \sigma^2), \quad \forall i = 1, \dots n$ (normalidad)
4. $Cov(y_i, y_j) = 0, \quad \forall i \neq j$ (independencia)
5. $n > k + 1$
6. No existen relaciones lineales entre los X_i (ausencia de multicolinealidad)

Escritas para el modelo en forma matricial quedan:

- $\vec{y} \sim N_n(\vec{x}\vec{\beta}, \sigma^2 I_n)$
- $n > k + 1$
- $rang(X) = k + 1$

Podemos darle un significado real a β_0 y β_j, para $j = 1, \dots k$.

- β_0 es el valor medio de la variable Y cuando todas las variables explicativas toman el valor 0.

$$E(y_i|x_{1i} = \dots = x_{ki} = 0) = \beta_0$$

- β_j es la variación media que experimenta la variable Y cuando X_j aumenta en una unidad y las demás variables explicativas permanecen constantes.

$$E(y_i|x_{1i}, \dots, x_{ji} + 1, \dots, x_{ki}) - E(y_i|x_{1i}, \dots, x_{ji}, \dots, x_{ki}) = \beta_j$$

2.2. Estimación de los parámetros

Queremos estimar $\beta_0, \beta_1, \ldots, \beta_k$, o análogamente $\hat{\vec{\beta}}$, y σ^2. Con los estimadores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ podemos estimar:

$$\hat{E}(y_i|x_{1i}, \ldots, x_{ki}) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}, \quad i = 1, \ldots n$$

Procedemos mediante el método de mínimos cuadrados. La función a minimizar es:

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

Planteamos las ecuaciones:

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

Estas son las ecuaciones normales de la regresión.

Resolviendo este sistema, llegamos a que M alcanza el mínimo si:

$$X'\vec{y} = X'X\hat{\vec{\beta}}$$

Por la hipótesis de ausencia de multicolinealidad $X'X$ tiene inversa, así que podemos escribir:

$$\hat{\vec{\beta}} = (X'X)^{-1}X'\vec{y}$$

Para estimar la varianza σ^2 usaremos la varianza residual:

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

Nota.

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

2.3. Propiedades de los estimadores

Sobre el estimador $\hat{\vec{\beta}}$, sabemos que:

$$\begin{cases} \hat{\vec{\beta}} = (X'X)^{-1}X'\vec{y} \\ \hat{y} \sim N_n(X\vec{\beta}, \sigma^2 I_n) \end{cases} \Rightarrow \hat{\vec{\beta}} \sim N_{k+1}(\vec{\beta}, \sigma^2(X'X)^{-1})$$

Nota.

$$\begin{cases} \vec{x} \sim N_n(\vec{\mu}, \Sigma) \\ \vec{y} = A\vec{x} \end{cases} \Rightarrow \vec{y} \sim N_k(A\vec{\mu}, A\Sigma A')$$

Tenemos además que:

$$Cov(\hat{\vec{\beta}}) = \sigma^2(X'X)^{-1} = \begin{pmatrix} V(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) & \dots & Cov(\hat{\beta}_0, \hat{\beta}_k) \\ Cov(\hat{\beta}_1, \hat{\beta}_0) & V(\hat{\beta}_1) & \dots & Cov(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_k, \hat{\beta}_0) & Cov(\hat{\beta}_k, \hat{\beta}_1) & \dots & V(\hat{\beta}_k) \end{pmatrix}$$

Así que $\hat{\beta}_j \sim N(\beta_j, \sigma^2 q_{j+1,j+1})$ para $j = 0, \dots, k$, donde $q_{j+1,j+1}$ es el elemento $(j+1, j+1)$ de $(X'X)^{-1}$. Equivalentemente, es el elemento $(j+1)$-ésimo de la diagonal principal de $(X'X)^{-1}$.

En cuanto a s_R^2,

$$\begin{cases} E(\frac{1}{\sigma^2}\sum_{i=1}^n e_i^2) = n-k-1 \\ V(\frac{1}{\sigma^2}\sum_{i=1}^n e_i^2) = 2(n-k-1) \end{cases} \Rightarrow \begin{cases} E(s_R^2) = \sigma^2 \\ V(s_R^2) = \frac{2(\sigma^2)^2}{n-k-1} \end{cases}$$

2.4. Intervalos de confianza para los parámetros

Intervalos de confianza para β_j, $j = 0, \dots, k$

Supondremos σ^2 desconocida.

Sea $j \in \{0, \dots, k\}$, sabemos que $\hat{\beta}_j \sim N(\beta_j, \sigma^2 q_{j+1,j+1})$. Así que:

$$\begin{cases} \dfrac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{q_{j+1,j+1}}} \sim N(0,1) \\ \dfrac{(n-k-1)s_R^2}{\sigma^2} \sim \chi^2_{n-k-1} \end{cases} \Rightarrow \frac{\hat{\beta}_j - \beta_j}{s_R\sqrt{q_{j+1,j+1}}} \sim t_{n-k-1}$$

Luego el intervalo de confianza para β_j a nivel de significación α es:

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

Intervalos de confianza para σ^2

Sabemos que $\frac{(n-k-1)s_R^2}{\sigma^2} \sim \chi^2_{n-k-1}$. Usando un desarrollo análogo al que hicimos para el modelo de regresión lineal simple, llegamos a que el intervalo de confianza para σ^2 a nivel de significación α es:

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

2.5. Contrastes de hipótesis para los coeficientes de regresión

Contrastes de significación individuales

Consideramos el contraste de hipótesis:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases} \qquad j = 1, \ldots, k$$

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

Fijamos el nivel de significación α. Hay tres formas de resolver el contraste.

Intervalos de confianza

Sea $IC_{1-\alpha}(\beta_j)$ el intervalo de confianza para β_j a nivel de significación α. Entonces:

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

Estadístico T

Sabemos que $\frac{\hat{\beta}_j - \beta_j}{s_R\sqrt{q_{j+1,j+1}}} \sim t_{n-k-1}$. Entonces $T = \frac{\hat{\beta}_j}{s_R\sqrt{q_{j+1,j+1}}} \sim t_{n-k-1}$ si H_0 es cierto. Tomamos un t_{exp}.

- Si $|t_{exp}| \leq t_{n-k-1,1-\frac{\alpha}{2}}$, aceptamos H_0 a nivel de significación α.
- En caso contrario, rechazamos H_0 a nivel de significación α.

Valor p

Sea p el p-valor de la distribución. Entonces:

- Si $p \geq \alpha$, aceptamos H_0 a nivel de significación α.
- En caso contrario, rechazamos H_0 a nivel de significación α.

Contraste de regresión

Consideramos ahora el contraste de hipótesis:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 : \exists i \in \{1, \ldots, k\} : \beta_i \neq 0 \end{cases}$$

Este contraste indica si hay suficiente evidencia en la muestra para afirmar que el modelo es globalmente o conjuntamente válido.

Recordamos que:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$
$$VT=VNE+VE$$

Observamos que:

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

Se verifica que $\frac{VE}{\sigma^2}\sim\chi^2_k$ si H_0 es cierta. Así que $\frac{VT}{\sigma^2}\sim\chi^2_{n-1}$ si H_0 es cierta.

Consideramos entonces el estadístico de contraste:

$$F=\frac{\frac{VE}{\sigma^2}/k}{\frac{VNE}{\sigma^2}/(n-k-1)}=\frac{(n-k-1)VE}{ks_R^2}$$

Observamos que $F\sim F_{k,n-k-1}$ si H_0 es cierta.

Tomamos un F_{exp}.

- Aceptamos H_0 a nivel de significación α si $F_{exp}\leq F_{k,n-k-1,1-\alpha}$.
- En caso contrario, rechazamos H_0 a nivel de significación α.

La tabla ANOVA es de la forma:

Fuentes	Suma de cuadrados	Grados de libertad	Cocientes
VE	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	k	VE/k
VNE	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$n-k-1$	$VNE/(n-k-1)$
VT	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$	

También se incluyen columnas para F_{exp} y p-valor.

Veamos algunas expresiones para VT, VNE y VE.

$$VT=\sum(y_i-\bar{y})^2=\sum y_i^2+n\bar{y}^2-2\bar{y}\sum y_i=\sum y_i^2+n\bar{y}^2=\vec{y}'\vec{y}-n\bar{y}^2$$
$$VNE=\sum(y_i-\hat{y}_i)^2=(\vec{y}-\hat{\vec{y}})'(\vec{y}-\hat{\vec{y}})=(\vec{y}-X\hat{\vec{\beta}})'(\vec{y}-X\hat{\vec{\beta}})=$$
$$=\vec{y}'\vec{y}-\hat{\vec{\beta}}X'\vec{y}+(\hat{\vec{\beta}}'X'X-\vec{y}'X)\vec{\beta}=\vec{y}'\vec{y}-\hat{\vec{\beta}}X'\vec{y}$$
$$VE=VT-VNE=\vec{y}'\vec{y}-n\bar{y}^2-\vec{y}'\vec{y}+\hat{\vec{\beta}}X'\vec{y}=\hat{\vec{\beta}}X'\vec{y}-n\bar{y}^2$$

Nota.

$$\hat{\beta}'X'X-\vec{y}'X=y'X(X'X)^{-1}X'X-y'X=y'X-y'X=0$$

Interpretación de los contrastes sobre los coeficientes de regresión

Los casos que se pueden presentar 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

Significativo indica que se rechaza la hipótesis H_0 de que el parámetro o parámetros a los que se refiere la hipótesis sea 0.

Analicemos cada uno de los casos:

- El caso 1 indica que todas las variables explicativas influyen.
- El caso 2 indica que solo influyen algunas variables explicativas, por lo que en principio se deberían eliminar las no significativas del modelo. Esto no debe hacerse mecánicamente, sino estudiando en profundidad cuál sería el modelo que se seleccionaría.
- El caso 3 corresponde al caso en que las x son muy dependientes entre sí y, aunque conjuntamente infuyen, individualmente no son significativas. Es decir, se tiene multicolinealidad.
- El caso 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 sí lo sean.
- El caso 5 es análogo al 4.
- En el caso 6 ninguna de las variables parece tener efecto sobre y pero solo podremos decir que sus efectos no se detectan en la muestra considerada.

Contrastes de grupos de cocientes

Consideramos ahora el contraste de hipótesis:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_i = 0 \\ H_1 : \exists j \in \{1, \ldots, i\} : \beta_j \neq 0 \end{cases}$$

Definimos:

- $VE(k)$: variabilidad explicada por el modelo con $x_1, \ldots, x_k$ como variables explicativas.

- $VE(k-i)$: variabilidad explicada por el modelo con todas las variables explicativas excepto $x_1, \ldots, x_k$.
- $\Delta VE = VE(k) - VE(k-i)$: variabilidad explicada por $x_1, \ldots, x_i$.
- $VNE(k)$: variabilidad no explicada por el modelo con $x_1, \ldots, x_k$ como variables explicativas.

Consideramos el estadístico de contraste:

$$F = \frac{\Delta VE/i}{VNE(k)/(n-k-1)} = \frac{(n-k-1)\Delta VE}{iVNE(k)} = \frac{(n-k-1)\Delta VE}{is_R^2(n-k-1)} = \frac{\Delta VE}{is_R^2}$$

Observamos que $F \sim F_{i,n-k-1}$ si H_0 es cierta.

Tomamos un F_{exp}.

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

Este constraste se puede utilizar para contrastes individuales:

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

En este caso, $F_{exp} = t_{exp}^2$.

2.6. Correlación en regresión múltiple

Coeficiente de determinación

Definimos el coeficiente de determinación como:

$$R^2 = \frac{VE}{VT} = \frac{VT - VNE}{VT} = 1 - \frac{VNE}{VT} = 1 - \frac{(n-k-1)s_R^2}{VT}, \quad 0 \leq R^2 \leq 1$$

Observamos que R^2 aumenta si el número de variables explicativas aumenta k, aunque las variables no sean significativas.

Coeficiente de determinación ajustado

Definimos el coeficiente de determinación ajustado o corregido como:

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

Observamos que $\bar{R}^2$ aumenta si y solo si s_R^2 disminuye.

Nota. $\bar{R}^2$ puede ser negativo.

Veamos qué relación hay entre R^2 y $\bar{R}^2$.

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

Por tanto,

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

Además,

$$n-k-1 \leq n-1 \Rightarrow \frac{n-1}{n-k-1}(1-R^2) \geq 1-R^2 \Rightarrow 1-\bar{R}^2 \geq 1-R^2 \Rightarrow \bar{R}^2 \leq R^2$$

Luego $\bar{R}^2 \leq R^2 \leq 1$.

2.7. Predicción

Estimación de las medias condicionadas

Queremos estimar:

$$m_0 = E(y|x_{10}, \ldots, x_{k0}) = \beta_0 + \beta_1 x_{10} + \cdots + \beta_k x_{k0} = \vec{x_0}'\vec{\beta}$$

con $\vec{x_0} = (1, x_{10}, \ldots, x_{k0})$. Para ello usamos el estimador:

$$\hat{m_0} = \hat{E}(y|x_{10}, \ldots, x_{k0}) = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \cdots + \hat{\beta}_k x_{k0}$$

Como $\hat{\vec{\beta}} \sim N_{k+1}(\vec{\beta}, \sigma^2(X'X)^{-1})$,

$$\hat{m_0} \sim N(\vec{x_0}'\vec{\beta}, \vec{x_0}'V(\hat{\vec{\beta}})\vec{x_0}) \equiv N(m_0, \sigma^2\vec{x_0}'(X'X)^{-1}\vec{x_0})$$

Observamos que $\hat{m_0}$ es un estimador insesgado.

Para obtener intervalos de confianza vemos que:

$$\frac{\hat{m_0} - m_0}{\sigma\sqrt{\vec{x_0}'(X'X)^{-1}\vec{x_0}}} \sim N(0,1) \Rightarrow \frac{\hat{m_0} - m_0}{s_R\sqrt{\vec{x_0}'(X'X)^{-1}\vec{x_0}}} \sim t_{n-k-1}$$

Por tanto, el intervalo de confianza para m_0 a nivel de significación α es:

$$IC_{1-\alpha}(m_0) = \Bigg(\hat{m_0} - t_{n-k-1,1-\frac{\alpha}{2}} s_R\sqrt{\vec{x_0}'(X'X)^{-1}\vec{x_0}},$$
$$\hat{m_0} + t_{n-k-1,1-\frac{\alpha}{2}} s_R\sqrt{\vec{x_0}'(X'X)^{-1}\vec{x_0}}\Bigg)$$

Predicción de una nueva observación

Queremos predecir la variable aleatoria

$$y_0 = \beta_0 + \beta_1 x_{10} + \cdots + \beta_k x_{k0} + u_0 = \vec{x_0}'\vec{\beta} + u_0$$

Su estimación puntual es:

$$\hat{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \cdots + \hat{\beta}_k x_{k0} = \vec{x_0}'\hat{\vec{\beta}} = \hat{m}_0$$

Consideramos la variable aleatoria error:

$$e_0 = y_0 - \hat{y_0}$$

Sabemos que:

$$\begin{cases} y_0 \sim N(\vec{x_0}'\hat{\beta}, \sigma^2) \\ \vec{y_0} \sim N(m_0, \sigma^2 \vec{x_0}'(X'X)^{-1}\vec{x_0}) \end{cases}$$

Como además y_0 e $\hat{y_0}$ son independientes, $e_0 \sim N$.

$$E(e_0) = E(y_0) - E(\hat{y_0}) = \vec{x_0}'\vec{\beta} - m_0 = 0$$
$$V(e_0) = V(y_0) + V(\hat{y_0}) = \sigma^2 + \sigma^2 \vec{x_0}'(X'X)^{-1}\vec{x_0} = \sigma^2(1 + \vec{x_0}'(X'X)^{-1}\vec{x_0})$$

Por tanto, $e_0 \sim N(0, \sigma^2(1 + \vec{x_0}'(X'X)^{-1}\vec{x_0}))$.

Para obtener intervalos de pronóstico observamos que:

$$\frac{e_0}{\sigma\sqrt{1 + \vec{x_0}'(X'X)^{-1}\vec{x_0}}} \sim N(0,1) \Rightarrow \frac{e_0}{s_R\sqrt{1 + \vec{x_0}'(X'X)^{-1}\vec{x_0}}} \sim t_{n-k-1}$$

Así que el intervalo de pronóstico para y_0 con contenido probabilístico $1-\alpha$ es:

$$IP_{1-\alpha}(y_0) = \Bigg(\hat{y_0} - t_{n-k-1,1-\frac{\alpha}{2}} s_R \sqrt{1 + \vec{x_0}'(X'X)^{-1}\vec{x_0}},$$
$$\hat{y_0} + t_{n-k-1,1-\frac{\alpha}{2}} s_R \sqrt{1 + \vec{x_0}'(X'X)^{-1}\vec{x_0}}\Bigg)$$

2.8. Diagnosis y validación del modelo

Multicolinealidad

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

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{pmatrix}, \quad rang(X) \leq k+1$$

Esto es equivalente a que $rang(X) < k + q$.

Cuando esto ocurre es difícil 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.** Se da cuando una de las variables explicativas es combinación lineal exacta de las demás. En este caso $rang(X) < k + 1$ así que $\det(X'X) = 0$ y no se puede calcular $(X'X)^{-1}$. El sistema de ecuaciones que determina el vector $\hat{\beta}$ no tiene solución única.
2. **Alta multicolinealidad.** Se da 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 $\hat{\beta}$. Sin embargo, presenta algunos problemas.
   - Los estimadores $\hat{\beta}_j$ tendrán varianzas muy altas, lo que provocará mucha imprecisión en la estimación de los $\hat{\beta}_j$. En consecuencia, los intervalos de confianza serán muy anchos.
   - Los estimadores $\hat{\beta}_j$ serán muy dependientes entre sí, puesto que tendrán altas covarianzas y habrá poca información sobre lo que ocurre al variar una variable si las demás permanecen constantes.

Consecuencias de la multicolinealidad

- Los estimadores $\hat{\beta}_j$ serán muy sensibles a pequeñas variaciones en el tamaño muestral o la supresión de una variable aparentemente no significativa. A pesar de esto, la predicción no tiene por qué verse afectada ante la multicolinealidad, ni esta 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 $\hat{\beta}_j$ 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 identificación de variables correlacionadas se realiza de una de las siguientes formas:

1. Examinando la matriz de correlaciones entre las variables explicativas R y su inversa. La presencia de correlaciones altas entre variables explicativas es un indicio de multicolinealidad. Aun así, es posible que exista una relación perfecta entre una variable y el resto y, sin embargo, sus coeficientes de correlación sean bajos.

   Definimos la matriz de correlaciones como:

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1k} \\ r_{12} & 1 & r_{23} & \dots & r_{2k} \\ r_{13} & r_{23} & 1 & \dots & r_{3k} \\ \vdots & \vdots & & \ddots & \vdots \\ r_{1k} & r_{2k} & r_{3k} & \dots & 1 \end{pmatrix}, \quad r_{ij} = \frac{s_{X_i X_j}}{s_{X_i} s_{X_j}}, \quad -1 \leq r_{ij} \leq 1$$

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

   La inversa de la matriz de correlaciones:

$$R^{-1} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \dots & \gamma_{1k} \\ \gamma_{21} & \gamma_{22} & \dots & \gamma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{k1} & \gamma_{k2} & \dots & \gamma_{kk} \end{pmatrix}$$

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

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

   donde $R^2_{i,r}$ es el coeficiente de determinación de la regresión de la variable X_i en función del resto de variables explicativas. Por tanto, si para algún i se tiene que:

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

   es decir, la variable X_i se explica como mínimo en un 90 % por el resto de variables explicativas. Luego estamos en una situación de alta multicolinealidad.

   R^{-1} se calculará con poca precisión cuando R sea casi singular.

2. Examinando los autovalores de $X'X$ o de R. Las mejores medidas de singularidad de $X'X$ o de R utilizan los autovalores de estas matrices. Un índice de singularidad que se utiliza en cálculo numérico es el índice de condicionamiento.

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

$$cond(M) = \sqrt{\frac{\lambda_k}{\lambda_1}} \geq 1$$

Es más conveniente calcular este índice para 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 $cond(R)$ y:

- Si $cond(R) > 30$, se tiene alta multicolinealidad.
- Si $10 < cond(R) < 30$, se tiene multicolinealidad moderada.
- Si $cond(R) < 10$, se tiene ausencia de multicolinealidad.

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.

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:

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

Análisis de los residuos

Los residuos se definen como:

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

Consideramos el vector de residuos:

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

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

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

Nota.

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

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:

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

Nota.

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

Por tanto:

$$\frac{\dfrac{e_i}{\sigma\sqrt{1-h_{ii}}}}{\sqrt{\dfrac{(n-k-1)s_R^2}{\sigma^2(n-k-1)}}} = \frac{e_i}{s_R\sqrt{1-h_{ii}}} \sim t_{n-k-1}$$

Se definen los residuos estudentizados como:

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

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.
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 $\hat{y}_i$.
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.

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 $\hat{\vec{\beta}}$ de parámetros estimado.
- Modifica el vector $\hat{\vec{y}}$ 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:

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

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

Nota. Recordamos que:

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

Entonces:

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

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

$$D(i) > F_{k+1,n-k-1,1-\alpha}$$

Nota. 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 $\bar{R}^2$.
   - **Varianza residual.** Es mejor modelo el que tenga menor s_R^2. Es equivalente al anterior criterio por la relación que hay entre $\bar{R}^2$ 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

Consideramos un conjunto de datos $\{(x_{1i}, \ldots, x_{ki})\}$ proveniente de dos poblaciones A y B. Hay dos modelos de regresión para estos datos que no son muy recomendables:

1. **Modelo conjunto.** Se ajusta un único modelo para todos los datos, sin importar la población de la que provienen. El modelo es por tanto sencillo y se consideran todos los datos. Sin embargo, se suponen homogéneas las poblaciones y esto no es cierto en general.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$$

2. **Modelos individuales.** Se ajustan dos modelos, uno para cada población por separado. Las predicciones tienen sentido pero se tienen menos datos para cada modelo.

Para obtener un mejor modelo añadimos una variable ficticia o *dummy*:

$$X_{k+1} = \begin{cases} 1 & \text{si el dato procede de } A \\ 0 & \text{si el dato procede de } B \end{cases}$$

Consideramos entonces el nuevo modelo general:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki} + \hat{\beta}_{k+1} x_{(k+1)i}$$

Al término $\hat{\beta}_{k+1}x_{(k+1)i}$ se le llama efecto principal.

- Para la población A, el modelo es:

$$\hat{y_i} = (\hat{\beta}_0 + \hat{\beta}_{k+1}) + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$$

- Para la población B, el modelo es:

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$$

Para comprobar si la variable X_{k+1} es significativa a nivel de significación α, consideramos el contraste de hipótesis:

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

Aceptar H_0 significa que los datos son homogéneos a nivel de significación α.

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 $\hat{y}$ es igual para cada población, lo que no es cierto en general. Veremos en ejemplos que podemos mejorarlo añadiendo interacciones.

Ejemplo. 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:

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

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

$$X_2 = \begin{cases} 1 & \text{si es hombre} \\ 0 & \text{si es mujer} \end{cases}$$

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:

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

- Para los hombres el modelo es:

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

- Para las mujeres el modelo es:

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

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 $\hat{\beta}_3 x_1 x_2$ llamado interacción de altura y sexo. Así que este nuevo modelo queda de la forma:

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

- Para los hombres el modelo es:

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

- Para las mujeres el modelo es:

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

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. 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:

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

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

$$X_2 = \begin{cases} 1 & \text{si usa petróleo} \\ 0 & \text{si no usa petróleo} \end{cases} \qquad X_3 = \begin{cases} 1 & \text{si usa carbón} \\ 0 & \text{si no usa carbón} \end{cases}$$

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:

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

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

- Para el petróleo el modelo es:

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

- Para el carbón el modelo es:

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

- Para la mezcla el modelo es:

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

De nuevo, podemos observar que las tres rectas tienen la misma pendiente, es decir, se supone que el incremento del rednimiento 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 $\hat{\beta}_4 x_1 x_2 + \hat{\beta}_5 x_1 x_3$. Así, el nuevo modelo queda de la forma:

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

- Para el petróleo el modelo es:

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

- Para el carbón el modelo es:

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

- Para la mezcla el modelo es:

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

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:

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

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

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

- Para el petróleo el modelo es:

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

- Para el carbón el modelo es:

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

- Para la mezcla el modelo es:

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

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

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

- Para el petróleo el modelo es:

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

- Para el carbón el modelo es:

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

- Para la mezcla el modelo es:

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

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

$$\begin{cases} H_0 : \beta_3 = \cdots = \beta_8 = 0 \\ H_1 \end{cases}$$

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

$$\begin{cases} H_0 : \beta_2 = \beta_7 = \beta_8 = 0 \\ H_1 \end{cases}$$

Capítulo 3

Modelo lineal generalizado

3.1. Introducción

El modelo lineal generalizado es una generalización de la regresión lineal que permite que los 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 $\vec{x}_i'\vec{\beta}$.

  Nota. $\vec{x}_i' = (1, x_{1i}, \ldots, x_{ki})$.
- **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. 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.

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

Así que $y_i \sim Ber(p_i)$ con $p_i = P(Y_i = 1)$.

Queremos estimar $\hat{p_i} = \hat{P}(Y_i = 1) = \hat{P}(A)$, es decir, la probabilidad de que el individuo i sea de la población A.

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

No podemos usar el modelo de regresión múltiple porque necesitamos que $p_i \in (0, 1)$. En su lugar podemos 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:

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

- **Función de distribución normal.**

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

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

3.3. Riesgo, oportunidad, riesgo relativo y razón de oportunidades

Definición 3.1 (Riesgo)**.** El riesgo es la probabilidad de que ocurra un resultado.

Definición 3.2 (Oportunidad)**.** La oportunidad es el cociente del número de eventos que producen un resultado entre el número de eventos que no lo producen.

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

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

Definición 3.3 (Riesgo relativo)**.** El riesgo relativo es el cociente de los riesgos de dos grupos de población.

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

Definición 3.4 (Razón de oportunidades)**.** La razón de oportunidades es el cociente de las oportunidades de dos grupos de población.

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

Ejemplo. En uno de cada 200 nacimientos ocurre un parto gemelar. La probabilidad o riesgo de que un embarazo elegido al azar dé lugar a gemelos es:

$$R_1 = \frac{1}{200}$$

Desde el punto de vista de la oportunidad, de 200 partos 1 es gemelar y 199 no lo son. Luego la oportunidad es:

$$O_1 = \frac{1}{199} = \frac{R_1}{1-R_1}$$

Introducimos un factor de riesgo. Se observó que, entre 100 mujeres que tomaron ácido fólico, 3 de cada 200 partos fueron gemelares. En este caso:

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

El aumento del riesgo del embarazo gemelar se puede expresar numéricamente como:

$$RR = \frac{R_2}{R_1} = 3, \quad OR = \frac{O_2}{O_1} = \frac{3/197}{1/199} \approx 3{,}03$$

Estudiaremos los modelos de regresión logística en términos de la razón de oportunidades.

3.4. Modelo de regresión logística

El modelo de regresión logística es:

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

Además, hemos visto que:

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

Luego queremos estimar:

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

Para encontrar los estimadores $\hat{\beta}_i$ usamos el método de máxima verosimilitud. Calculamos la función de verosimilitud:

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

Tomamos logaritmos en ambos miembros de la igualdad:

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

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 $\hat{\beta}_i$ 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 variables son significativas.

Capítulo 4

Inferencia bayesiana

4.1. Teorema de Bayes

Teorema 4.1 (Teorema de Bayes)**.** *Sea $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad. Sea $\{A_1, \ldots, A_n\} \subset \mathcal{A}$ una partición de Ω y sea $B \in \mathcal{A}$ tal que $P(B) > 0$ y del que se conocen $P(B|A_i)$, para $i = 1, \ldots, n$. Entonces*

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

donde:

- *$P(A_j)$, $j = 1, \ldots, n$, se llaman probabilidades a priori.*
- *$P(B|A_j)$, $j = 1, \ldots, n$, se llaman verosimilitudes.*
- *$P(A_j|B)$, $j = 1, \ldots, n$, se llaman probabilidades a posteriori.*

Esta se conoce como la fórmula de Bayes.

Observación. Las probabilidades a posteriori son proporcionales al producto de verosimilitudes y probabilidades a priori.

$$P(A_i|B) \propto P(B|A_i)P(A_i)$$

Ejemplo. Una caja contiene dos monedas: una moneda legal M_1 y otra con una cara en cada lado M_2.

En primer lugar, se selecciona una de las dos monedas al azar, se lanza y sale cara. Veamos cuál es la probabilidad de que la moneda lanzada sea la legal.

Para ello definimos los sucesos:

- C_i: en el lanzamiento i sale cara.

- F_i: en el lanzamiento i sale cruz.

Usamos el teorema de Bayes:

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

Lanzamos de nuevo la moneda elegida y se obtiene otra cara. Veamos cuál es la probabilidad de que la moneda lanzada sea la legal.

Podemos usar el carácter secuencial del teorema de Bayes y usar los resultados anteriores.

Probabilidad a priori	Verosimilitudes	Probabilidad a posteriori
$P(M_1\|C_1) = \frac{1}{3}$	$P(C_2\|M_1) = \frac{1}{2}$	$P(M_1\|C_1 \cap C_2) = \frac{1}{5}$
$P(M_2\|C_1) = \frac{2}{3}$	$P(C_2\|M_2) = 1$	$P(M_2\|C_1 \cap C_2) = \frac{4}{5}$

4.2. Teorema de Bayes generalizado

Teorema 4.2 (Teorema de Bayes generalizado)**.** *Sean $\vec{x} = (x_1, \ldots, x_n)$ una muestra y θ una variable aleatoria. Sea f_θ la distribución a priori y $f(\vec{x}|\theta)$ la función de verosimilitud. Entonces:*

$$f(\theta|\vec{x}) = \frac{f(\vec{x}|\theta) f_\theta(\theta)}{f(\vec{x})}$$

donde:

$$f(\vec{x}) = \begin{cases} \sum_{i=1}^{n} f(\vec{x}|\theta_i) f_\theta(\theta_i) & \textit{si es discreta} \\ \int_\Theta f(\vec{x}, \theta) f_\theta(\theta) d\theta & \textit{si es continua} \end{cases}$$

Observación. Para los clásicos, θ es un parámetro fijo y desconocido. En cambio, para los bayesianos θ es una variable aleatoria.

Ejemplo. Supongamos que tenemos una moneda y queremos estimar la probabilidad p de obtener cara. Supongamos que nuestras creencias a priori sobre p se pueden describir por una distribución uniforme en $(0, 1)$. Realizamos el experimento de tirar la moneda 12 veces y obtenemos 9 caras y 3 cruces.

Definimos la variable aleatoria:

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

Queremos estimar $P(C) = P(X = 1) = p$ a partir de nuestra muestra $\vec{x} = (x_1, \ldots, x_{12})$, com $\sum_{i=1}^{12} x_i = 9$.

Como $p \sim U(0, 1)$, su distribución a priori es $f(p) = 1$ si $p \in (0, 1)$. Calculamos la función de verosimilitud:

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

Podemos hallar la distribución a posteriori:

$$f(p|\vec{x}) \propto p^9(1-p)^3 \Rightarrow p|\vec{x} \sim Be(10,4)$$

Nota. Si $X \sim Be(p,q)$ beta, entonces:

$$f_X(x) \propto x^{p-1}(1-x)^{q-1}$$

4.3. Familias de distribución conjugadas

Las familias de distribución conjugadas son aquellas en las que las distribuciones a priori y a posteriori son de la misma familia.

Muestras de la distribución Bernoulli

Sean $x_i|\theta \sim Ber(\theta)$ y $\theta \sim Be(p,q)$. Su distribución a priori es:

$$f_\theta(\theta) \propto \theta^{p-1}(1-\theta)^{q-1}, \quad \theta \in (0,1)$$

Dada una muestra $\vec{x}$, calculamos la función de verosimilitud:

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

Luego la distribución a posteriori es:

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

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

Por tanto, la beta es una familia conjugada respecto de muestras de la Bernoulli.

Muestras de la distribución de Poisson

Sean $x_i|\lambda \sim Po(\lambda)$ y $\lambda \sim Ga(a,p)$. Su distribución a priori es:

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

Dada una muestra $\vec{x}$, calculamos la función de verosimilitud:

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

Luego la distribución a posteriori es:

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

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

Por tanto, la gamma es una familia conjugada respecto de muestras de la Poisson.

Muestras de la distribución normal

Lema 4.3.

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

Media desconocida y precisión conocida

Sea $x_i|\mu \sim N(\mu, p)$ con media μ desconocida y precisión p conocida y sea $\mu \sim N(m_0, p_0)$. Su distribución a priori es:

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

Dada una muestra $\vec{x}$, calculamos la función de verosimilitud:

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

Nota.

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2 = \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 + 2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x}) = \\ &= (n-1)s^2 + n(\bar{x}-\mu)^2\end{aligned}$$

Así que:

$$L(\vec{x}, \mu) = e^{-\frac{p}{2}((n-1)s^2 + n(\bar{x}-\mu)^2)} \propto e^{-\frac{np}{2}(\bar{x}-\mu)^2}$$

Luego la distribución a posteriori es:

$$f(\mu|\vec{x}) \propto f_\mu(\mu) L(\vec{x}, \mu) \propto e^{-\frac{p}{2}(\mu-m_0)^2} e^{-\frac{np}{2}(\bar{x}-\mu)^2} = e^{-\frac{1}{2}(p_0(\mu-m_0)^2 + np(\bar{x}-\mu)^2)}$$

Usando el lema previo, queda:

$$\begin{gathered}f(\mu, \vec{x}) \propto e^{-\frac{1}{2}(a(\mu-b)^2+c)} \propto e^{-\frac{a}{2}(\mu-b)^2} \\ \Rightarrow \mu|\vec{x} \sim N(b, pr = a)\end{gathered}$$

donde

$$a = p_0 + np, \quad b = \frac{p_0 m_0 + np\bar{x}}{p_0 + np}$$

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

Media conocida y precisión desconocida

Sea $x_i|\tau \sim N(\mu, \tau)$ y sea $\tau \sim Ga(a_0, p_0)$. Su distribución a priori es:

$$f_\tau(\tau) = \frac{a_0^p}{\Gamma(p_0)} e^{-a_0\tau} \tau^{p_0-1}, \quad \tau, a_0, p_0 > 0$$

Dada una muestra $\vec{x}$, calculamos la función de verosimilitud:

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

Luego la distribución a posteriori es:

$$\begin{aligned} f(\tau|\vec{x}) &\propto f_\tau(\tau) L(\vec{x}, \tau) \propto \tau^{\frac{n}{2}+p_0-1} e^{-\tau\left(a_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)} \\ &\Rightarrow \tau|\vec{x} \sim Ga(a_n, p_n) \end{aligned}$$

donde

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

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

Media conocida y varianza desconocida

Definición 4.1. Sea $X \sim Ga(a, p)$, consideramos $Y = \frac{1}{X}$. Entonces $Y \sim GaI(a, p)$ gamma invertida. Su función de densidad es:

$$f_Y(y) = \frac{a^p}{\Gamma(p)} e^{-\frac{a}{y}} y^{-(p+1)}, \quad y > 0$$

Sea $x_i|\sigma^2 \sim N(\mu, \sigma^2)$ con varianza σ^2 desconocida y sea $\sigma^2 \sim GaI(a_0, p_0)$. Su distribución a priori es:

$$f_{\sigma^2}(\sigma^2) = \frac{a_0^{p_0}}{\Gamma(p_0)} e^{-\frac{a_0}{\sigma^2}} (\sigma^2)^{-(p_0+1)}, \quad a_0, p_0 > 0$$

Dada una muestra $\vec{x}$, calculamos la función de verosimilitud:

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

Luego la distribución a posteriori es:

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

donde

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

Por tanto, la gamma invertida es una familia conjugada respecto de muestras de la normal con media conocida y varianza desconocida.

Media y precisión desconocidas

Definición 4.2. Decimos que (μ, τ) sin $NGa(m_0, \tau_0, a_0, p_0)$ normal gamma, con $m_0 \in \mathbb{R}, \tau_0, a_0, p_0 > 0$, si:

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

Su función de densidad es:

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

Definición 4.3. Si $T \sim t_n$ y $X = \mu + \frac{1}{\sqrt{p}}T$, entonces $X \sim t(\mu, p, n)$, donde μ es la media y p es el parámetro de escala. Su función de densidad es:

$$f_X(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)\sqrt{p}}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)\sqrt{n}} \left(1 + \frac{p}{n}(x - \mu)^2\right)^{-\frac{n+1}{2}}$$

Verifica que:

$$E(X) = \mu, \quad V(X) = \frac{1}{p}\frac{n}{n-2}$$

Observación. La distribución t_1 se llama distribución de Cauchy. Además:

$$t_n \xrightarrow[n\to\infty]{} N(0, 1)$$

Teorema 4.4. *Si $(\mu, \tau) \sim NGa(m_0, \tau_0, a_0, p_0)$, entonces:*

$$\mu \sim t\left(m_0, \frac{p_0\tau_0}{a_0}, 2p_0\right)$$

Corolario 4.5 (Génesis bayesiana de la t de Student)**.**

$$\begin{cases} \mu|\tau \sim N(0, \tau) \\ \tau \sim Ga\left(\frac{n}{2}, \frac{n}{2}\right) \end{cases} \Rightarrow \mu \sim t(0, 1, n) \equiv t_n$$

Sean $x_i|\mu, \tau \sim N(\mu, \tau)$ y $(\mu, \tau) \sim NGa(m_0, \tau_0, a_0, p_0)$. Se puede comprobar que la normal gamma es una familia conjugada respecto de muestras de la normal con media y precisión desconocidas.

4.4. Distribuciones a priori no informativas

Definición 4.4. La información de Fisher para θ se define como:

$$J(\theta) = -E\left(\frac{\partial^2 \log(f(x|\theta))}{\partial \theta^2}\right)$$

Proposición 4.6 (Regla de Jeffreys)**.**

$$f_\theta(\theta) \propto \sqrt{J(\theta)}$$

Observación. La regla de Jeffreys no da densidades en general. A aquellas que no son densidades se les llama densidades impropias.

Muestras de la distribución Bernoulli

Sea $x|\theta \sim Ber(\theta)$. Entonces:

$$f(x|\theta) = \theta^x(1-\theta)^{1-x}, \quad x = 0, 1, 0 < \theta < 1$$

Calculamos:

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

Luego la información de Fisher para θ es:

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

Por tanto:

$$f_\theta(\theta) \propto \sqrt{J(\theta)} \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \Rightarrow \theta \sim Be\left(\frac{1}{2}, \frac{1}{2}\right)$$

Muestras de la distribución de Poisson

Sea $x|\lambda \sim Po(\lambda)$. Entonces:

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \propto e^{-\lambda}\lambda^x, \quad x = 0, 1, \lambda > 0$$

Calculamos:

$$\begin{aligned}\log(f(x|\lambda)) &\propto -\lambda + x\log(\lambda)\\ \frac{\partial \log(f(x|\lambda))}{\partial \lambda} &\propto -1 + \frac{x}{\lambda}\\ \frac{\partial^2 \log(f(x|\lambda))}{\partial \lambda^2} &\propto -\frac{x}{\lambda^2}\end{aligned}$$

Luego la información de Fisher para λ es:

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

Por tanto:

$$f_\lambda(\lambda) \propto \frac{1}{\sqrt{\lambda}}$$

Observamos que f_λ no es una densidad.

Muestras de la distribución normal

Media desconocida y precisión conocida

Sea $x|\mu \sim N(\mu, p)$. Entonces:

$$f(x|\mu) = \frac{\sqrt{p}}{\sqrt{2\pi}} e^{-\frac{p}{2}(x-\mu)^2} \propto e^{-\frac{p}{2}(x-\mu)^2}$$

Calculamos:

$$\begin{aligned}\log(f(x|\mu)) &\propto -\frac{p}{2}(x-\mu)^2\\ \frac{\partial \log(f(x|\mu))}{\partial \mu} &\propto p(x-\mu)\\ \frac{\partial^2 \log(f(x|\mu))}{\partial \mu^2} &\propto -p \propto -1\end{aligned}$$

Luego la información de Fisher para μ es:

$$J(\mu) \propto E(1) = 1$$

Por tanto:

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

Observamos que f_μ no es una densidad.

Media conocida y desviación típica desconocida

Sea $x|\sigma \sim N(\mu, \sigma)$. Entonces:

$$f(x|\sigma) = \frac{1}{\sqrt{1\pi\sigma^2}} e^{-\frac{p}{2}(x-\mu)^2} \propto e^{-\frac{p}{2}(x-\mu)^2}$$

Calculamos:

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

Luego la información de Fisher para σ es:

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

Por tanto:

$$f_\sigma(\sigma) \propto \frac{1}{\sigma}$$

Observamos que f_σ no es una densidad.

Media conocida y varianza desconocida

Sea $x|\sigma^2 \sim N(\mu, \sigma^2)$. Entonces:

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

Media conocida y precisión desconocida

Sea $x|\tau \sim N(\mu, \tau)$. Entonces:

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

Media y desviación típica desconocidas

Sea $x|\mu, \sigma \sim N(\mu, \sigma)$. Entonces:

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

Media y varianza desconocidas

Sea $x|\mu, \sigma^2 \sim N(\mu, \sigma^2)$. Entonces:

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

Media y precisión desconocidas

Sea $x|\mu,\tau \sim N(\mu,\tau)$. Entonces:

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

Observación (Regla de oro). Todo sale de la distribución a priori.

4.5. Estimación puntual

La estimación puntual bayesiana es un problema de decisión.

Una función de pérdida es una función $L : \Theta \times \Theta \to \mathbb{R}$ donde $L(\theta,t)$ es 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 de pérdida valor absoluto:

$$L(\theta,t) = |\theta - t|$$

- Función de pérdida 0-1:

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

Queremos minimizar respecto de t la función pero el valor de θ es desconocido y no podemos calcular la pérdida. Así que minimizamos la esperanza de la función de pérdida a posteriori $E(L(\theta,t)|\vec{x})$. Elegimos $\hat{\theta}$ tal que:

$$\min_{t\in\Theta} E(L(\theta,t)|\vec{x}) = \min_{t\in\Theta} \int_\Theta L(\theta,t)f(\theta|\vec{x})d\theta = E(L(\theta,\vec{\theta}))$$

Ejemplo. Tomamos $L(\theta,t) = (\theta - t)^2$.

$$\begin{aligned} \psi(t) = E(L(\theta,t)|\vec{x}) &= \int_\Theta (\theta - t)^2 f(\theta|\vec{x})d\theta = \\ &= \int_\Theta \theta^2 f(\theta|\vec{x})d\theta + t^2 \int_\Theta f(\theta|\vec{x})d\theta - 2t\int_\Theta \theta f(\theta|\vec{x})d\theta = \\ &= \int_\Theta \theta^2 f(\theta|\vec{x})d\theta + t^2 - 2tE(\theta|\vec{x}) \\ \psi'(t) &= 2t - 2E(\theta|\vec{x}) \end{aligned}$$

Hallamos el mímimo de ψ:

$$\psi'(t) = 0 \Leftrightarrow t = E(\theta|\vec{x}) \Rightarrow \hat{\theta} = E(\theta|\vec{x})$$

De forma análoga podemos concluir:

- El estimador bayesiano de θ bajo la función de pérdida cuadrática es la media a posteriori de θ.
- El estimador bayesiano de θ bajo la función de pérdida 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. La media a posteriori es una media ponderada de la media a priori y del estimador de máxima verosimilitud.

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

4.6. Intervalos de credibilidad

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 4.5. Un región R es de máxima densidad a posteriori de θ con 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 $f(\theta_1|\vec{x}) > f(\theta_2|\vec{x})$.

Se escribe $MDP_{1-\alpha}$.

Nota. Si la distribución de $\theta|\vec{x}$ es unimodal, la región de máxima densidad a posteriori de contenido probabilístico $1-\alpha$ para θ es un intervalo (a,b) tal que:

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

4.7. Contrastes de hipótesis

Contraste unilateral a la derecha

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

Calculamos:

$$P(H_0|\vec{x}) = P(\theta \leq \theta_0|\vec{x})$$
$$P(H_1|\vec{x}) = P(\theta > \theta_0|\vec{x})$$

Entonces:

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

Contraste unilateral a la izquierda

$$\begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

Calculamos:

$$P(H_0|\vec{x}) = P(\theta \geq \theta_0|\vec{x})$$
$$P(H_1|\vec{x}) = P(\theta < \theta_0|\vec{x})$$

Entonces:

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

Contraste bilateral

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

- Si $\theta_0 \in MDP_{1-\alpha}(\theta|\vec{x})$, aceptamos H_0.
- En caso contrario, rechazamos H_0.

4.8. Distribuciones predictivas

Sea $\vec{x} = (x_1, \ldots, x_n)$, definimos la distribución predictiva como:

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

Ejemplo. Consideramos los tres siguientes casos:

1. A una anciana se le somete a diez pruebas para ver si acierta o no si en una taza de té con leche se ha echado antes el té o la leche. Acierta las diez pruebas.
2. A un experto en música se le somete a diez pruebas para ver si acierta si una pieza musical es de Mozart o de Haydn. Acierta en las diez ocasiones.
3. A un borracho se le somete a diez pruebas para ver si acierta si en el lanzamiento de una moneda sale cara o cruz. Acierta en las diez ocasiones.

Estudiemos qué ocurrirá si realizamos en cada uno de estos casos una prueba más.

Sea $X \sim Ber(\theta)$, con:

$$X = \begin{cases} 1 & \text{si acierta} \\ 0 & \text{si falla} \end{cases}, \quad \theta = P(X = 1)$$

Entonces:

$$f(x_i|\theta) \propto \theta^{x_i}(1-\theta)^{10-x_i}$$

Sea $\vec{x} = (x_1, \ldots, x_{10}) = (1, \ldots, 1)$, calculamos la función de verosimilitud:

$$L(\vec{x}, \theta) = \prod_{i=1}^{10} f(x_i|\theta) \propto \theta^{\sum_{i=1}^{10} x_i}(1-\theta)^{10-\sum_{i=1}^{10} x_i} = \theta^{10}$$

Sea $\theta \sim Be(p, q)$. Su distribución a priori es:

$$f_\theta(\theta) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\theta^{p-1}(1-\theta)^{q-1}, \quad 0 < \theta < 1$$

Luego la distribución a posteriori es:

$$f(\theta|\vec{x}) \propto f_\theta(\theta)L(\vec{x}, \theta) \propto \theta^{p-1+10}(1-\theta)^{q-1} \Rightarrow \theta|\vec{x} \sim Be(p+10, q)$$

Una estimación puntual de θ con la función de pérdida cuadrática es:

$$E(\theta|\vec{x}) = \frac{p+10}{p+q+10}$$

Hallamos la distribución predictiva:

$$\begin{aligned} P(x_{11} = 1|\vec{x}) &= \int_0^1 P(x_{11} = 1|\theta) f(\theta|\vec{x}) d\theta = \\ &= \int_0^1 \theta \frac{\Gamma(p+10+q)}{\Gamma(p+10)\Gamma(q)}\theta^{p+10}(1-\theta)^{q-1}d\theta = \\ &= \frac{\Gamma(p+10+q)}{\Gamma(p+10)\Gamma(q)} \int_0^1 \theta^{p+10}(1-\theta)^{q-1}d\theta \end{aligned}$$

Observamos que el interior de la integral es la función de densidad de una distribución $Be(p+11, q)$ salvo constantes. Por tanto:

$$P(x_{11} = 1|\vec{x}) = \frac{\Gamma(p+10+q)}{\Gamma(p+10)\Gamma(q)} \frac{\Gamma(p+11)\Gamma(q)}{\Gamma(p+q+11)} = \frac{p+10}{p+q+10}$$

Luego $P(x_{11} = 0|\vec{x}) = \frac{q}{p+q+10}$.

Analizamos cada caso:

1. Podemos usar una distribución no informativa según la regla de Jeffreys, como $p = q = \frac{1}{2}$. En este caso:

$$P(x_{11}|\vec{x}) = \frac{1/2 + 10}{1/2 + 1/2 + 10} = \frac{21}{22} \approx 0,9545$$

2. Podemos tomar $p, q \to \infty$, de forma que $\frac{p}{p+q} \xrightarrow[p,q\to\infty]{} 1$. En este caso:

$$P(x_{11}|\vec{x}) = \frac{p + 10}{p + q + 10} \xrightarrow[p,q\to\infty]{} 1$$

3. Podemos tomar $p = q \to \infty$. En este caso:

$$P(x_{11}|\vec{x}) = \frac{p + 10}{2p + 10} \xrightarrow[p\to\infty]{} \frac{1}{2}$$

La probabilidad de que en tres pruebas siguientes los resultados sean acierto, fallo y fallo, en ese orden, podemos calcularla como:

$$P(x_{11} = 1, x_{12} = 1, x_{13} = 0|\vec{x}) = \int_0^1 P(x_{11} = 1, x_{12} = 1, x_{13} = 0|\theta) f(\theta|\vec{x}) d\theta = \\ = \int_0^1 \theta^2(1 - \theta) f(\theta|\vec{x}) d\theta$$

La probabilidad de que en tres pruebas siguientes haya dos aciertos y un fallo, podemos calcularla como $3P(x_{11} = 1, x_{12} = 1, x_{13} = 0)$.

En general, la probabilidad de que en las siguientes m pruebas haya r aciertos se calcula como $\binom{m}{r} P(x_{n+1} = \cdots = x_{n+r} = 1, x_{n+r+1} = \cdots = x_{n+m} = 0)$.

4.9. Análisis bayesiano para datos de Bernoulli

Sea $x|\theta \sim Ber(\theta)$. Estudiaremos el caso informativo con $\theta \sim Be(p, q)$. Sabíamos que:

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

Como por la regla de Jeffreys $\theta \sim Be\left(\frac{1}{2}, \frac{1}{2}\right)$, el caso no informativo es análogo con $p = q = \frac{1}{2}$.

Estimación puntual

Analizamos la estimación bajo cada una de las funciones de pérdida usuales.

- Si $L(\theta, t) = (\theta - t)^2$, entonces:

$$\hat{\theta} = E(\theta|\vec{x}) = \frac{\sum_{i=1}^{n} x_i + p}{n + p + q}$$

- Si $L(\theta, t) = |\theta - t|$, entonces:

$$\hat{\theta} = Me(\theta|\vec{x})$$

- Si $L(\theta, t) = \begin{cases} 0 & \text{si } |\theta - t| \leq \varepsilon \\ 1 & \text{si } |\theta - t| > \varepsilon \end{cases}$, entonces:

$$\hat{\theta} = Mo(\theta|\vec{x}) = \frac{\sum_{i=1}^{n} x_i + p - 1}{n + p + q - 2}$$

Vimos que la media a posteriori era una media ponderada de la media a priori y del estimador de máxima verosimilitud. Es decir:

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

Calculamos los pesos:

$$\frac{\sum_{i=1}^{n} x_i + p}{n + p + q} = \omega \frac{p}{p + q} + (1 - \omega)\frac{1}{n}\sum_{i=1}^{n} x_i \Rightarrow \omega = \frac{p + q}{n + p + q}$$

Observamos que:

$$\omega = \frac{p + q}{n + p + q} \xrightarrow[n\to\infty]{} 0, \quad 1 - \omega = \frac{n}{n + p + q} \xrightarrow[n\to\infty]{} 1$$

Es decir, en la media a posteriori tiene mayor influencia el estimador de máxima verosimilitud si tenemos muchos datos.

Intervalos de credibilidad

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

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

El intervalo de máxima densidad a posteriori de θ con contenido probabilístico $1 - \alpha$ es un intervalo (a, b) tal que:

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

Nota. El intervalo de máxima densidad a posteriori no se puede calcular a mano.

Distribución predictiva

Calculamos la distribución predictiva.

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

Observamos que el interior de la integral es la función de densidad de una distribución $Be(\sum_{i=1}^n x_i + p + 1, n - \sum_{i=1}^n x_i + q)$ salvo constantes. Por tanto:

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

La probabilidad de que en las siguientes m pruebas haya r éxitos y $m-r$ fracasos se calcula como:

$$\binom{m}{r} P(x_{n+1} = \cdots = x_{n+r} = 1, x_{n+r+1} = \cdots = x_{n+m} = 0|\vec{x}) =$$
$$= \binom{m}{r} \int_0^1 \theta^r (1-\theta)^{m-r} f(\theta|\vec{x}) d\theta$$

4.10. Análisis bayesiano para datos de Poisson

Sea $x|\lambda \sim Po(\lambda)$. Estudiaremos el caso informativo con $\lambda \sim Ga(a,p)$. Sabíamos que:

$$f_\lambda(\lambda) = \frac{a^p}{\Gamma(p)} e^{-a\lambda} \lambda^{p-1}, \quad \lambda > 0$$
$$\lambda|\vec{x} \sim Ga(a+n, p + \sum_{i=1}^n x_i)$$

Como por la regla de Jeffreys $f_\lambda(\lambda) \propto \lambda^{-\frac{1}{2}}$, el caso no informativo es análogo con $a=0$ y $p=\frac{1}{2}$.

Estimación puntual

Analizamos la estimación bajo la función de pérdida cuadrática:

$$\hat{\lambda} = E(\lambda|\vec{x}) = \frac{a+n}{\sum_{i=1}^n x_i + p}$$

Intervalos de credibilidad

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

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

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

Así que:

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

Por tanto:

$$1-\alpha = P(\lambda_1 < \lambda|\vec{x} < \lambda_2) = P(2(a+n)\lambda_1 < \chi^2_{2(\sum_{i=1}^n x_i - p)} < 2(a+n)\lambda_2)$$

El intervalo de máxima densidad a posteriori de λ con contenido probabilístico $1-\alpha$ es un intervalo (λ_1, λ_2) tal que:

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

Nota. El intervalo de máxima densidad a posteriori no se puede calcular a mano.

Distribución predictiva

Calculamos la distribución predictiva.

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

Observamos que el interior de la integral es la función de densidad de una distribución $Ga(a+n+1, x_{n+1} + \sum_{i=1}^n x_i + p)$ salvo constantes. Por tanto:

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

4.11. Análisis bayesiano para datos normales

Media desconocida y precisión conocida

Sea $x|\mu \sim N(\mu, p)$. Estudiaremos el caso informativo con $\mu \sim N(m_0, p_0)$. Sabíamos que:

$$f_\mu(\mu) = \frac{\sqrt{p_0}}{\sqrt{2\pi}} e^{\frac{p_0}{2}(\mu - m_0)^2} \propto e^{-\frac{p_0}{2}(\mu - m_0)^2}$$
$$\mu|\vec{x} \sim N(m_n, p_n)$$

donde

$$m_n = \frac{p_0 m_0 + np\bar{x}}{p_0 + np}, \quad p_n = p_0 + np$$

Como por la regla de Jeffreys $f_\mu(\mu) \propto 1$, el caso no informativo es análogo con $p_0 = 0$.

Estimación puntual

Analizamos la estimación bajo la función de pérdida cuadrática:

$$\hat{\mu} = E(\mu|\vec{x}) = m_n = \frac{p_0 m_0 + np\bar{x}}{p_0 + np}$$

Intervalos de credibilidad

El intervalo de máxima densidad a posteriori de μ con contenido probabilístico $1-\alpha$ es un intervalo (μ_1, μ_2) tal que:

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

Es decir, tiene que verificar:

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

Además:

$$\sqrt{p_n}(\mu_1 - m_n) = -z_{1-\frac{\alpha}{2}} \Rightarrow \mu_1 = m_n - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{p_n}}$$

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

Por tanto:

$$MDP_{1-\alpha}(\mu) = \left(m_n - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{p_n}}, m_n + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{p_n}} \right)$$

Contrastes de hipótesis

Consideramos el contraste:

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

Entonces:

$$\begin{aligned} P(H_0|\vec{x}) &= P(\mu \leq \mu_0|\vec{x}) = P(\sqrt{p_n}(\mu - m_n) \leq \sqrt{p_n}(\mu_0 - m_n)) = \\ &= P(Z \leq \sqrt{p_n}(\mu_0 - m_n)) = \Phi(\sqrt{p_n}(\mu_0 - m_n)) \\ P(H_1|\vec{x}) &= 1 - \Phi(\sqrt{p_n}(\mu_0 - m_n)) \end{aligned}$$

Aceptamos H_0 si $P(H_0|\vec{x}) \geq P(H_1|\vec{x})$.

Consideramos ahora el contraste:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Aceptamos H_0 si $\mu_0 \in MDP_{1-\alpha}(\mu)$, para α fijo.

Distribución predictiva

Calculamos la distribución predictiva.

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

Usando el lema, queda:

$$\begin{aligned} f(x_{n+1}|\vec{x}) &= \frac{\sqrt{pp_n}}{2\pi} \int_{\mathbb{R}} e^{-\frac{1}{2}(\alpha(\mu-\beta)^2+\gamma)} d\mu = \\ &= \frac{\sqrt{pp_n}}{2\pi} e^{-\frac{\gamma}{2}} \int_{\mathbb{R}} e^{-\frac{\alpha}{2}(\mu-\beta)^2} \end{aligned}$$

Observamos que el interior de la integral es la función de densidad de una distribución $N(\beta, \alpha)$ salvo constantes. Por tanto:

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

Media conocida y varianza desconocida

Sea $x|\sigma^2 \sim N(\mu, \sigma^2)$. Estudiaremos el caso informativo con $\sigma^2 \sim GaI(a_0, p_0)$. Sabíamos que:

$$\begin{aligned} f_{\sigma^2}(\sigma^2) &= \frac{a_0^{p_0}}{\Gamma(p_0)} e^{-\frac{a_0}{\sigma^2}} (\sigma^2)^{-(p_0+1)} \\ \sigma^2|\vec{x} &\sim GaI(a_n, p_n) \end{aligned}$$

donde

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

Como por la regla de Jeffreys $f_{\sigma^2}(\sigma^2) \propto (\sigma^2)^{-1}$, el caso no informativo es análogo con $a_0 = p_0 = 0$.

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:

$$1-\alpha = P(\sigma_1^2 < \sigma^2 | \vec{x} < \sigma_2^2) = P(\frac{1}{\sigma_2^2} < \frac{1}{\sigma^2} | \vec{x} < \frac{1}{\sigma_1^2}) =$$
$$= P\left(\frac{2a_0 + \sum_{i=1}^n (x_i-\mu)^2}{\sigma_2^2} < \chi^2_{n+2p_0} < \frac{2a_0 + \sum_{i=1}^n (x_i-\mu)^2}{\sigma_1^2}\right)$$

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

Distribución predictiva

Calculamos la distribución predictiva.

$$f(x_{n+1}|\vec{x}) = \int_0^\infty f(x_{n+1}|\sigma^2) f(\sigma^2|\vec{x}) d\sigma^2 =$$
$$= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \frac{a_n^{p_n}}{\Gamma(p_n)} \int_0^\infty (\sigma^2)^{-(p_n+\frac{1}{2}+1)} e^{-\frac{1}{2\sigma^2}((x_{n+1}-\mu)^2+2a_n)} d\sigma^2$$

Observamos que el interior de la integral es la función de densidad de una distribución $GaI\left(\frac{(x_{n+1}-\mu)^2}{2} + a_n, p_n + \frac{1}{2}\right)$ salvo constantes. Por tanto:

$$f(x_{n+1}|\vec{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{a_n^{p_n}}{\Gamma(p_n)} \frac{\Gamma\left(p_n+\frac{1}{2}\right)}{\left(\frac{(x_n-\mu)^2}{2} + a_n\right)^{p_n+\frac{1}{2}}}$$

4.12. Influencia de la distribución a priori según el tamaño muestral

Ejemplo. 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 a priori $\theta_A \sim N(900, \sigma_A = 20)$. Supongamos que B tiene poca experiencia en este área y su opinión más bien vaga es que a priori $\theta_B \sim N(200, \sigma_B = 80)$. Supongamos que un método experimental de medidas está disponible y que una observación y medida por este método sigue $y|\theta \sim N(\theta, \sigma = 40)$. Supongamos que $y = 850$.

Calculamos la función de verosimilitud:

$$L(\theta, y) = \frac{1}{\sqrt{2\pi\frac{1600}{40^2}}} e^{-\frac{1}{20\cdot 40^2}(850-\theta)^2}$$

Hallamos sus distribuciones a posteriori:

$$\theta_A|y \sim N\left(890, \frac{1}{390}\right) \equiv N(890, 17{,}9)$$
$$\theta_B|y \sim N\left(840, \frac{1}{1280}\right) \equiv N(840, 35{,}7)$$

Supongamos ahora que tenemos 100 observaciones $y_1, \ldots, y_{100}$ tales que $\bar{y} = 870$. Calculamos de nuevo la función de verosimilitud:

$$L(\theta, y_1, \ldots, y_{100}) = \prod_{i=1}^{100} e^{-\frac{1}{2\cdot 40^2}(y_i-\theta)^2} \propto e^{-\frac{1}{2\cdot 40^2}\sum_{i=1}^{100}(y_i-\theta)^2}$$

Observamos que $\sum_{i=1}^{100}(y_i - \theta)^2 = \sum_{i=1}^{100}(y_i - \bar{y} + \bar{y} - \theta)^2 = \sum_{i=1}^{100}(y_i - \bar{y})^2 + 100(\bar{y} - \theta)^2$. Por tanto:

$$L(\theta, y_1, \ldots, y_{100}) \propto e^{-\frac{1}{2\cdot 40^2}100(\bar{y}-\theta)^2}$$

Hallamos sus distribuciones a posteriori:

$$\theta_A|y_1, \ldots, y_{100} \sim N(871{,}2, 3{,}9)$$
$$\theta_B|y_1, \ldots, y_{100} \sim N(869{,}8, 3{,}995)$$

Observamos que son muy similares.