

Estadística Bayesiana

Regresión Lineal

Lizbeth Naranjo Albarrán

Facultad de Ciencias, UNAM

April 7, 2022

Índice

1	Estadística Clásica	1
1.1	Modelo	1
1.2	Predicción	3
2	Estadística Bayesiana	4
2.1	Modelo	4
2.2	Distribución Inicial	6
2.2.1	Familia Conjugada	6
2.2.2	Distribución Inicial No informativa	6
2.3	Distribución Final	7
2.4	Inferencia y Predicción	8
2.5	Ejemplos	11

Capítulo 1

Análisis de Regresión Lineal con Métodos de Estadística Clásica

1.1 Modelo

Las técnicas de regresión se cuentan entre los métodos mas utilizados en la Estadística Aplicada. Dada una variable de respuesta Y y un conjunto de covariables $z = (z_1, \dots, z_r)'$, es de interés estimar una supuesta relación funcional entre Y y z , así como predecir el valor de observaciones futuras para distintos valores de las covariables.

Una manera de modelar dicha relación consiste en representar el valor esperado de Y como

$$E(Y|z) = \mu(z),$$

donde, en general, $\mu(\cdot)$ es una función desconocida. En la practica es común aproximar a $\mu(\cdot)$ a través de una función paramétrica simple.

$$\mu(z) = \psi(z; \beta),$$

donde $\beta = (\beta_0, \beta_1, \dots, \beta_k)'$ denota un vector de parámetros desconocidos. Más aún, en muchos casos se supone que $\psi(\cdot; \beta)$ es una función lineal de β en alguna escala apropiada, *i.e.*

$$\psi(z; \beta) = h(\beta_0 + \beta_1 s_1(z) + \dots + \beta_k s_k(z))$$

para alguna transformación uno a uno $h(\cdot)$, conocida, y para algunas funciones suaves $\{s_j(\cdot) : j = 1, \dots, k\}$, también conocidas. Esta función es tratada entonces como si fuera la verdadera función de regresión $\mu(\cdot)$, por lo que el problema se reduce a hacer inferencias sobre el valor del parámetro β .

Este planteamiento da lugar a los llamados modelos lineales generalizados. En esta sección nos concentraremos en el modelo de regresión usual, donde la variable Y tiene una distribución Normal y $h(\cdot)$ es la función identidad.

El modelo Normal

Supongamos que se tienen n observaciones independientes $(Y_1, z_1), \dots, (Y_n, z_n)$ del modelo

$$Y_i \sim N(y_i | \mu(z_i), \sigma^2) \quad (\sigma^2 > 0, \text{desconocida}),$$

donde

$$\mu(z_i) = \beta_0 + \beta_1 s_1(z_i) + \dots + \beta_k s_k(z_i).$$

Sean $x_{i1} = 1$ ($i = 1, \dots, n$) y

$$x_{ij} = s_j(z_i), \quad i = 1, \dots, n; \quad j = 1, \dots, k.$$

Entonces podemos expresar el modelo como

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2) \quad (i = 1, \dots, n).$$

Resulta conveniente escribir el modelo de forma matricial. Sea $p = k + 1$. Entonces

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

donde $\mathbf{Y} = (Y_1, \dots, Y_n)'$, $\mathbf{X} = [x_{ij}]$ es una matriz n *times* p , $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$, e \mathbf{I}_n denota a la matriz identidad de orden n . Dicho de otra forma

$$\mathbf{Y} \sim N_n(\mathbf{y} | \mathbf{X}\beta, \sigma^2 \mathbf{I}_n). \quad (1.1)$$

En el resto de esta sección supondremos que \mathbf{X} es de rango completo, p .

La función de verosimilitud para el modelo (2.1) es de la forma

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

Recordemos que los estimadores de máxima verosimilitud para β y σ^2 están dados por $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ y $\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})$, respectivamente, y notemos que

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) + (\beta - \hat{\beta})'\mathbf{X}'\mathbf{X}(\beta - \hat{\beta}).$$

Así, la función de verosimilitud puede escribirse como

$$L(\beta, \sigma^2; \mathbf{y}) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \left[(\beta - \hat{\beta})'\mathbf{X}'\mathbf{X}(\beta - \hat{\beta}) + n\hat{\sigma}^2 \right] \right\}.$$

1.2 Predicción

Predicción

Supongamos que se desea predecir Y_* , un nuevo valor de la variable de respuesta, dado el vector de covariables $\mathbf{x}'_* = (1, x_{1*}, \dots, x_{k*})$. De acuerdo con el modelo,

$$\begin{aligned} Y_* &= \beta_0 + \beta_1 x_{1*} + \dots + \beta_k x_{k*} + \epsilon_* \\ &= \mathbf{x}'_* \beta + \epsilon_*, \end{aligned}$$

donde $\epsilon_* \sim N(0, \sigma^2)$ es independiente de ϵ . Entonces,

$$\mu_* \stackrel{def}{=} E(Y_* | \beta, \sigma^2) = \mathbf{x}'_* \beta$$

Y

$$Var(Y_* | \beta, \sigma^2) = \sigma^2.$$

El problema de predicción puede abordarse de dos maneras:

(a) *Inferencia sobre μ_** . El parámetro μ_* , que corresponde al valor esperado de la observación futura Y_* , no es mas que una combinación lineal de los coeficientes de regresión.

En particular, se tiene que para una observación

$$\begin{aligned} E(y | \mathbf{x}'_*) &= \mathbf{x}'_* \beta \\ Var(y | \mathbf{x}'_*) &= \sigma^2 \mathbf{x}'_* (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_* \end{aligned}$$

En este caso, el intervalo del $(1 - \alpha) \times 100\%$ de confianza está dado por

$$\mathbf{x}'_* \hat{\beta} \pm t_{(n-p)}^{1-\alpha/2} \tilde{\sigma} \sqrt{\mathbf{x}'_* (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_*},$$

donde $t_{(n-p)}^{1-\alpha/2}$ es el cuantil orden $(1 - \alpha/2)$ de una distribución t de Student estandarizada con $(n - p)$ grados de libertad.

(b) *Inferencia sobre Y_** . En este caso interesa calcular un intervalo para Y_* .

En particular, se tiene que para una observación

$$\begin{aligned} E(y | \mathbf{x}'_*) &= \mathbf{x}'_* \beta \\ Var(y | \mathbf{x}'_*) &= \sigma^2 \{1 + \mathbf{x}'_* (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_*\} \end{aligned}$$

En este caso, el intervalo del $(1 - \alpha) \times 100\%$ de confianza esta dado por

$$\mathbf{x}'_* \hat{\beta} \pm t_{(n-p)}^{1-\alpha/2} \tilde{\sigma} \sqrt{1 + \mathbf{x}'_* (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_*}.$$

Capítulo 2

Análisis de Regresión Lineal con Métodos de Estadística Bayesiana

2.1 Modelo

Las técnicas de regresión se cuentan entre los métodos mas utilizados en la Estadística Aplicada. Dada una variable de respuesta Y y un conjunto de covariables $z = (z_1, \dots, z_r)'$, es de interés estimar una supuesta relación funcional entre Y y z , así como predecir el valor de observaciones futuras para distintos valores de las covariables.

Una manera de modelar dicha relación consiste en representar el valor esperado de Y como

$$E(Y|z) = \mu(z),$$

donde, en general, $\mu(\cdot)$ es una función desconocida. En la practica es común aproximar a $\mu(\cdot)$ a través de una función paramétrica simple.

$$\mu(z) = \psi(z; \beta),$$

donde $\beta = (\beta_0, \beta_1, \dots, \beta_k)'$ denota un vector de parámetros desconocidos. Más aún, en muchos casos se supone que $\psi(\cdot; \beta)$ es una función lineal de β en alguna escala apropiada, *i.e.*

$$\psi(z; \beta) = h(\beta_0 + \beta_1 s_1(z) + \dots + \beta_k s_k(z))$$

para alguna transformación uno a uno $h(\cdot)$, conocida, y para algunas funciones suaves $\{s_j(\cdot) : j = 1, \dots, k\}$, también conocidas. Esta función es tratada entonces como si fuera la verdadera función de regresión $\mu(\cdot)$, por lo que el problema se reduce a hacer inferencias sobre el valor del parámetro β .

Este planteamiento da lugar a los llamados modelos lineales generalizados. En esta sección nos concentraremos en el modelo de regresión usual, donde la variable Y tiene una distribución Normal y $h(\cdot)$ es la función identidad.

El modelo Normal

Supongamos que se tienen n observaciones independientes $(Y_1, z_1), \dots, (Y_n, z_n)$ del modelo

$$Y_i \sim N(y_i | \mu(z_i), \sigma^2) \quad (\sigma^2 > 0, \text{desconocida}),$$

donde

$$\mu(z_i) = \beta_0 + \beta_1 s_1(z_i) + \dots + \beta_k s_k(z_i).$$

Sean $x_{i1} = 1$ ($i = 1, \dots, n$) y

$$x_{ij} = s_j(z_i), \quad i = 1, \dots, n; \quad j = 1, \dots, k.$$

Entonces podemos expresar el modelo como

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2) \quad (i = 1, \dots, n).$$

Resulta conveniente escribir el modelo de forma matricial. Sea $p = k + 1$. Entonces

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

donde $\mathbf{Y} = (Y_1, \dots, Y_n)'$, $\mathbf{X} = [x_{ij}]$ es una matriz n times p , $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$, e \mathbf{I}_n denota a la matriz identidad de orden n . Dicho de otra forma

$$\mathbf{Y} \sim N_n(\mathbf{y} | \mathbf{X}\beta, \sigma^2 \mathbf{I}_n). \quad (2.1)$$

En el resto de esta sección supondremos que \mathbf{X} es de rango completo, p .

La **función de verosimilitud** para el modelo (2.1) es de la forma

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

Recordemos que los estimadores de máxima verosimilitud para β y σ^2 están dados por $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ y $\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})$, respectivamente, y notemos que

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) + (\beta - \hat{\beta})'\mathbf{X}'\mathbf{X}(\beta - \hat{\beta}).$$

Así, la función de verosimilitud puede escribirse como

$$L(\beta, \sigma^2; \mathbf{y}) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \left[(\beta - \hat{\beta})'\mathbf{X}'\mathbf{X}(\beta - \hat{\beta}) + n\hat{\sigma}^2 \right] \right\}.$$

Para facilitar la notación y el desarrollo subsecuente, es conveniente trabajar en términos de la precisión $\tau = 1/\sigma^2$ en lugar de la varianza σ^2 . La verosimilitud toma entonces la forma

$$L(\beta, \tau; \mathbf{y}) \propto (\tau)^{n/2} \exp \left\{ -\frac{\tau}{2} \left[(\beta - \hat{\beta})'\mathbf{X}'\mathbf{X}(\beta - \hat{\beta}) + n\hat{\sigma}^2 \right] \right\}. \quad (2.2)$$

2.2 Distribución Inicial

2.2.1 Familia Conjugada

Dada la forma de la verosimilitud, una familia conjugada particularmente conveniente tiene densidades de la forma

$$p(\beta, \tau) \propto \tau^{n_0/2} \exp \left\{ -\frac{\tau}{2} [(\beta - \mathbf{b}_0)' \mathbf{B}_0 (\beta - \mathbf{b}_0) + s_0] \right\},$$

donde $n_0, s_0 \in \mathbb{R}$, $\mathbf{b}_0 \in \mathbb{R}^p$ y \mathbf{B}_0 es una matriz $p \times p$ simétrica y positiva semi-definida.

Notemos que, dado el valor de τ , el *kernel* de la densidad condicional $p(\beta|\tau)$ es proporcional al de la densidad $N_p(\beta|\mathbf{b}_0, \tau^{-1}\mathbf{B}_0^{-1})$, *i.e.*

$$p(\beta|\tau) \propto \tau^{p/2} \exp \left\{ -\frac{\tau}{2} (\beta - \mathbf{b}_0)' \mathbf{B}_0 (\beta - \mathbf{b}_0) \right\},$$

de manera que los factores restantes corresponden a la densidad marginal $p(\tau)$, *i.e.*

$$p(\tau) \propto \tau^{(n_0-p)/2} \exp \{ -s_0\tau/2 \},$$

Haciendo $a = n_0 - p + 2$ y $d = s_0$, se tiene entonces que

$$\begin{aligned} p(\beta, \tau) &= p(\beta|\tau)p(\tau) \\ &= N_p(\beta|\mathbf{b}_0, \tau^{-1}\mathbf{B}_0^{-1})Ga(\tau|a/2, d/2). \end{aligned} \tag{2.3}$$

Esta distribución es conocida como *Normal-Gamma* y es propia si $a > 0$, $d > 0$ y \mathbf{B}_0 es positiva definida.

2.2.2 Distribución Inicial No informativa

En situaciones en las que se desea representar un estado de información inicial vaga acerca de (β, τ) , es común utilizar algún tipo de distribución inicial “no informativa”. Uno de los métodos más populares para obtener dichas distribuciones es la *Regla de Jeffreys*. Puede demostrarse fácilmente que en el caso del modelo de regresión (2.1) la distribución inicial de Jeffreys es

$$\pi(\beta, \tau) \propto \tau^{(p-2)/2}.$$

Esta distribución es impropia y puede obtenerse a partir de la familia conjugada (2.3) haciendo $a = 0$, $d = 0$ y $\mathbf{B}_0 = \mathbf{0}$.

Otra distribución inicial no informativa comúnmente usada en modelos de localización y escala es

$$\pi(\beta, \tau) \propto \tau^{-1},$$

la cual corresponde a la *distribución inicial de referencia* obtenida a partir del método de Bernardo (1979).

Al igual que la distribución de Jeffreys, esta distribución es impropia y es un caso limite de la familia conjugada (2.3) cuando $a = -p$, $d = 0$ y $\mathbf{B}_0 = \mathbf{0}$.

Comentario. En ambos casos $\mathbf{B}_0 = \mathbf{0}$ implica que la varianza de la distribución inicial de β es infinita, lo que generalmente se interpreta como una forma de representar un estado de información inicial vaga acerca del valor de β .

2.3 Distribución Final

Proposición 2.1. La distribución final de (β, τ) para el modelo (2.1) si se utiliza una distribución inicial conjugada de la forma (2.3) es

$$p(\beta, \tau | \mathbf{y}) = N_p(\beta | \mathbf{b}_1, \tau^{-1} \mathbf{B}_1^{-1}) Ga(\tau | a_1/2, d_1/2),$$

donde

$$\begin{aligned} \mathbf{b}_1 &= (\mathbf{X}'\mathbf{X} + \mathbf{B}_0)^{-1}(\mathbf{X}\mathbf{y} + \mathbf{B}_0\mathbf{b}_0) \\ \mathbf{B}_1 &= \mathbf{X}'\mathbf{X} + \mathbf{B}_0 \\ a_1 &= n + a \\ d_1 &= (\mathbf{y} - \mathbf{X}\mathbf{b}_1)'(\mathbf{y} - \mathbf{X}\mathbf{b}_1) + (\mathbf{b}_1 - \mathbf{b}_0)'\mathbf{B}_0(\mathbf{b}_1 - \mathbf{b}_0) + d \end{aligned}$$

Demostración. Por el Teorema de Bayes,

$$p(\beta, \tau | \mathbf{y}) \propto p(\beta, \tau) L(\beta, \tau; \mathbf{y}).$$

El resultado es evidente si se nota que

$$\begin{aligned} (\beta - \hat{\beta})'\mathbf{X}'\mathbf{X}(\beta - \hat{\beta}) + (\beta - \mathbf{b}_0)'\mathbf{B}_0(\beta - \mathbf{b}_0) = \\ (\beta - \mathbf{b}_1)'\mathbf{B}_1(\beta - \mathbf{b}_1) + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} + \mathbf{b}'_0\mathbf{B}_0\mathbf{b}_0 - \mathbf{b}'_1\mathbf{B}_1\mathbf{b}_1 \end{aligned}$$

y

$$\begin{aligned} (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} + \mathbf{b}'_0\mathbf{B}_0\mathbf{b}_0 - \mathbf{b}'_1\mathbf{B}_1\mathbf{b}_1 = \\ (\mathbf{y} - \mathbf{X}\mathbf{b}_1)'(\mathbf{y} - \mathbf{X}\mathbf{b}_1) + (\mathbf{b}_1 - \mathbf{b}_0)'\mathbf{B}_0(\mathbf{b}_1 - \mathbf{b}_0). \end{aligned}$$

□

De acuerdo con este resultado, la distribución marginal final de τ es

$$p(\tau|\mathbf{y}) = Ga(\tau|a_1/2, d_1/2),$$

lo que implica que la correspondiente distribución final para σ^2 esta dada por $IGa(\sigma^2|a_1/2, d_1/2)$.

Por otro lado, si se desea hacer inferencia sobre β entonces es necesario calcular su distribución marginal final, dada por

$$\begin{aligned} p(\beta|\mathbf{y}) &= \int p(\beta, \tau|\mathbf{y}) d\tau \\ &= \frac{\Gamma((a_1 + p)/2)}{\Gamma(a_1/2)\pi^{p/2}} \det \left\{ \frac{1}{d_1} \mathbf{B}_1 \right\}^{1/2} \left\{ 1 + \frac{1}{d_1} (\beta - \mathbf{b}_1)' \mathbf{B}_1 (\beta - \mathbf{b}_1) \right\}^{-(a_1+p)/2}. \end{aligned}$$

En otras palabras,

$$p(\beta|\mathbf{y}) = St_p(\beta|\mathbf{b}_1, \mathbf{T}_1^{-1}, a_1), \quad (2.4)$$

Donde $\mathbf{T}_1 = \left(\frac{a_1}{d_1} \right) \mathbf{B}_1$, de manera que la distribución final de β es t de Student con a_1 grados de libertad, parámetro de localización \mathbf{b}_1 , y parámetro de escala \mathbf{T}_1^{-1} .

Distribución final de referencia

Recordemos que la distribución inicial de referencia, $\pi(\beta, \tau) \propto \tau^{-1}$, corresponde a un caso limite de la familia conjugada (2.3) con $\mathbf{B}_0 = \mathbf{0}$, $a = -p$ y $d = 0$. Aunque es impropia, ésta da lugar a una distribución final propia siempre y cuando $n > p$. De hecho, en este caso se tiene que

$$\begin{aligned} \mathbf{b}_1 &= \hat{\beta} \\ \mathbf{B}_1 &= \mathbf{X}'\mathbf{X} \\ a_1 &= n - p \\ d_1 &= (n - p)\tilde{\sigma}^2, \end{aligned}$$

donde $\tilde{\sigma}^2$ es el estimador insesgado usual para σ^2 . Por lo tanto,

$$\pi(\beta, \tau|\mathbf{y}) = N_p(\beta|\hat{\beta}, \tau^{-1}(\mathbf{X}'\mathbf{X})^{-1}) Ga(\tau|(n - p)/2, (n - p)\tilde{\sigma}^2/2),$$

de donde

$$\pi(\beta|\mathbf{y}) = St_p(\beta|\hat{\beta}, \tilde{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}, n - p).$$

2.4 Inferencia y Predicción

Inferencia sobre σ^2

De acuerdo con los resultados de la sección anterior, la distribución final de σ^2 es

$$p(\sigma^2|\mathbf{y}) = IGa(\sigma^2|a_1/2, d_1/2).$$

En particular, se tiene que

$$\begin{aligned} E(\sigma^2|\mathbf{y}) &= \frac{d_1}{a_1 - 2} \\ Var(\sigma^2|\mathbf{y}) &= \frac{2d_1^2}{(a_1 - 2)^2(a_1 - 4)} \\ Moda(\sigma^2|\mathbf{y}) &= \frac{d_1}{a_1 + 2}. \end{aligned}$$

Estas cantidades pueden servir de base para hacer inferencias sobre σ^2 . También es posible construir intervalos de máxima densidad o simplemente reportar algunos percentiles de la distribución final de σ^2 .

Notemos que si se utiliza la distribución de referencia estas expresiones se reducen a

$$\begin{aligned} E(\sigma^2|\mathbf{y}) &= \frac{(n-p)\tilde{\sigma}^2}{n-p-2} \\ Var(\sigma^2|\mathbf{y}) &= \frac{2(n-p)^2\tilde{\sigma}^4}{(n-p-2)^2(n-p-4)} \\ Moda(\sigma^2|\mathbf{y}) &= \frac{(n-p)\tilde{\sigma}^2}{n-p+2}. \end{aligned}$$

Inferencia sobre β

La distribución final de β esta dada por (2.3). En particular,

$$\begin{aligned} E(\beta|\mathbf{y}) &= \mathbf{b}_1 \quad \text{si } a_1 > 1 \\ Var(\beta|\mathbf{y}) &= \frac{a_1}{a_1-2} \mathbf{T}_1^{-1} \\ &= \frac{d_1}{a_1-2} \mathbf{B}_1^{-1} \quad \text{si } a_1 > 2. \end{aligned}$$

Por otro lado, si se utiliza la distribución de referencia entonces

$$\begin{aligned} E(\beta|\mathbf{y}) &= \hat{\beta} \quad \text{si } n > p+1 \\ Var(\beta|\mathbf{y}) &= \frac{(n-p)\tilde{\sigma}^2}{n-p-2} (\mathbf{X}'\mathbf{X})^{-1} \quad \text{si } n > p+2. \end{aligned}$$

Como en el caso anterior, estas cantidades pueden servir de base para hacer inferencias sobre β . Notemos, sin embargo, que en este caso el interes se centra generalmente en combinaciones lineales de las entradas del vector β .

Sea $\gamma = \mathbf{C}\beta$, donde \mathbf{C} es una matriz $r \times p$ de rango r ($r \leq p$). Entonces

$$p(\gamma|\sigma^2, \mathbf{y}) = N_r(\gamma|\mathbf{g}, \sigma^2\mathbf{G}),$$

donde $\mathbf{g} = \mathbf{C}\mathbf{b}_1$ y $\mathbf{G} = \mathbf{C}\mathbf{B}_1^{-1}\mathbf{C}'$. Por lo tanto

$$p(\gamma|\mathbf{y}) = St_r(\gamma|\mathbf{g}, (d_1/a_1)\mathbf{G}, a_1).$$

Supongamos, por ejemplo, que $r = 1$ y $\mathbf{C} = \mathbf{e}'_i = (0, \dots, 1, \dots, 0)$ para alguna $i = 1, \dots, p$. Entonces $\gamma = \mathbf{e}'_i\beta = \beta_{i-1}$. En este caso $g = \mathbf{e}'_i\mathbf{b}_1 = b_{1,i-1}$ y $G = B_1^{ii}$, donde B_1^{ii} es la entrada (i, i) de la matriz \mathbf{B}_1^{-1} . Por lo tanto,

$$p(\beta_j|\mathbf{y}) = St\left(\beta_j|b_{1j}, (d_1/a_1)B_1^{j+1,j+1}, a_1\right) \quad (j = 0, 1, \dots, k).$$

Prediccion

Supongamos que se desea predecir Y_* , un nuevo valor de la variable de respuesta, dado el vector de covariables $\mathbf{x}'_* = (1, x_{1*}, \dots, x_{k*})$. De acuerdo con el modelo,

$$\begin{aligned} Y_* &= \beta_0 + \beta_1 x_{1*} + \dots + \beta_k x_{k*} + \epsilon_* \\ &= \mathbf{x}'_*\beta + \epsilon_*, \end{aligned}$$

donde $\epsilon_* \sim N(0, \sigma^2)$ es independiente de ϵ . Entonces,

$$\mu_* \stackrel{def}{=} E(Y_*|\beta, \sigma^2) = \mathbf{x}'_*\beta$$

Y

$$Var(Y_*|\beta, \sigma^2) = \sigma^2.$$

El problema de prediccion puede abordarse de dos maneras:

(a) *Inferencia sobre μ_** . El parámetro μ_* , que corresponde al valor esperado de la observación futura Y_* , no es mas que una combinación lineal de los coeficientes de regresión.

Sea $r = 1$ y $\mathbf{C} = \mathbf{x}'_*$. Entonces $\gamma = \mathbf{x}'_*\beta = \mu_*$, $g = \mathbf{x}'_*\mathbf{b}_1$ y $G = \mathbf{x}'_*\mathbf{B}_1^{-1}\mathbf{x}_*$, por lo que

$$p(\mu_*|\mathbf{y}) = St(\mu_*|\mathbf{x}'_*\mathbf{b}_1, (d_1/a_1)\mathbf{x}'_*\mathbf{B}_1^{-1}\mathbf{x}_*, a_1).$$

En particular, si se utiliza la distribución de referencia entonces

$$\pi(\mu_*|\mathbf{y}) = St(\mu_*|\mathbf{x}'_*\hat{\beta}, \tilde{\sigma}^2\mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*, n-p).$$

En este caso, el intervalo de máxima densidad del $(1 - \alpha) \times 100\%$ está dado por

$$\mathbf{x}'_*\hat{\beta} \pm t_{(n-p)}^{1-\alpha/2} \tilde{\sigma} \sqrt{\mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*},$$

donde $t_{(n-p)}^{1-\alpha/2}$ es el cuantil orden $(1 - \alpha/2)$ de una distribución t de Student estandarizada con $(n - p)$ grados de libertad. Este intervalo tiene la misma forma que el correspondiente intervalo frecuentista.

(b) *Inferencia sobre Y_** . En este caso interesa calcular la distribución predictiva final para Y_* , *i.e.*

$$p(y_*|\mathbf{y}) = \int \int p(y_*|\beta, \sigma^2) p(\beta, \sigma^2|\mathbf{y}) d\beta d\sigma^2.$$

Recordemos primero que $Y_* = \mathbf{x}'_*\beta + \epsilon_*$. Trabajando condicionalmente en σ^2 , tenemos que β y ϵ_* son independientes y

$$\begin{aligned} p(\beta|\sigma^2, \mathbf{y}) &= N_p(\beta|\mathbf{b}_1, \sigma^2\mathbf{B}_1^{-1}) \\ p(\epsilon_*|\sigma^2, \mathbf{y}) &= N(\epsilon_*|0, \sigma^2). \end{aligned}$$

Esto implica que

$$p(y_*|\sigma^2, \mathbf{y}) = N(y_*|\mathbf{x}'_*\mathbf{b}_1, \sigma^2\{1 + \mathbf{x}'_*\mathbf{B}_1^{-1}\mathbf{x}_*\}).$$

Finalmente, integrando con respecto a la distribución final de σ^2 ,

$$p(y_*|\mathbf{y}) = St(y_*|\mathbf{x}'_*\mathbf{b}_1, (d_1/a_1)\{1 + \mathbf{x}'_*\mathbf{B}_1^{-1}\mathbf{x}_*\}, a_1).$$

Si se utiliza la distribución de referencia entonces la distribución predictiva final toma la forma

$$p(y_*|\mathbf{y}) = St(y_*|\mathbf{x}'_*\hat{\beta}, \tilde{\sigma}^2\{1 + \mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*\}, n - p).$$

En este caso, el intervalo de máxima densidad del $(a - \alpha) \times 100\%$ esta dado por

$$\mathbf{x}'_*\hat{\beta} \pm t_{(n-p)}^{1-\alpha/2} \tilde{\sigma} \sqrt{1 + \mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*}.$$

Como en el caso anterior, este intervalo tiene la misma forma que el correspondiente intervalo frecuentista.

2.5 Ejemplos

FALTA incluir:

- Cálculos de la regresión lineal usando Muestreo de Gibbs
- Prior de Jeffreys de Normal Multivariada.

Código R: Bayes8_1LinearRegression.R

Bibliografía

- Bernardo, J. M. and A. Smith (1994). *Bayesian Theory*. Wiley.
- Box, G. E. P. and G. C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. Massachusetts: Addison-Wesley.
- Congdon, P. (2006). *Bayesian Statistical Modelling* (Second ed.). Chichester: John Wiley & Sons.
- Congdon, P. D. (2010). *Applied Bayesian Hierarchical Methods*. Boca Raton, Florida: Chapman & Hall/CRC.
- Dey, D. K., S. K. Ghosh, and B. K. Mallick (2000). *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.
- Gutiérrez-Peña, E. (1998). *Análisis Bayesiano de Modelos Jerárquicos Lineales*. México: IIMAS, UNAM.
- Montgomery, D. and E. A. Peck (2001). *Introduction to Linear Regression Analysis* (3rd ed.). John Wiley and Sons.
- Weisberg, S. (2005). *Applied Linear Regression*. John Wiley & Sons.