Capítulo 3

MODELOS DE REGRESIÓN

ESTIMACIÓN

3.1 Estimación puntual de los parámetros del modelo

Existen diversos métodos para hallar estimadores, los más usados son el de máxima de verosimilitud y el de mínimos cuadrados ordinarios. Para poder usar el método de máxima verosimilitud es necesario conocer la distribución de probabilidad de los y_i, los cuales en este caso se suponen normales. En el caso del método de mínimos cuadrados no es necesario conocer la distribución de probabilidad de los y_i. En primer lugar se determinan los estimadores por el método de máxima verosimilitud y luego se hará por el método de mínimos cuadrados.

3.1.1 Estimadores por el método de mínimos cuadrados.

3.1.1.1 Estimador de β En esta sección discutimos el método de los mínimos cuadrados para estimar $\boldsymbol{\beta}$. Ningún supuesto sobre la distribución de $\boldsymbol{y}$ es requerido para obtener los estimadores.

El método de mínimos cuadrados consiste en determinar los estimadores de los parámetros que minimizan la suma de los errores elevados al cuadrado. En este caso, dado que el modelo es $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, el método de mínimos cuadrados consiste en determinar los valores de $\boldsymbol{\beta}$, digamos $\hat{\boldsymbol{\beta}}$, para los cuales $\sum_{i=1}^{n} \varepsilon_i^2$ es mínima. Dado que $\sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ y $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$, entonces se quiere

$$\min \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \min(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{3.1}$$

El resultado está dado en el siguiente teorema

Teorema 3.1 *Sea $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ donde $\boldsymbol{X}$ es una matriz $n \times p$ de rango $p < n$, entonces el valor de $\boldsymbol{\beta}$ que minimiza $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ es*

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}^{-}\boldsymbol{y} \tag{3.2}$$

Prueba. Queremos hallar el valor de $\boldsymbol{\beta}$

$$\min \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \min(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

Desarrollando el producto $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ se tiene que

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

Ahora para obtener $\boldsymbol{\beta}$ que minimiza $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ se deriva $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ con respecto a $\boldsymbol{\beta}$ y se iguala a cero obteniéndose de esta manera la siguiente ecuación

$$\frac{\partial \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0$$

despejando se obtienen las siguientes ecuaciones

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

la cual es conocida como las **ecuaciones normales**. Ahora, suponiendo que la matriz $\boldsymbol{X}$ es de rango completo, $\boldsymbol{X}$?‘$\boldsymbol{X}$ es nosingular, se obtiene el estimador de $\boldsymbol{\beta}$, el cual está dado por

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

lo cual demuestra el teorema.

Ya que $\hat{\boldsymbol{\beta}}$ en 3.2 minimiza $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$, $\hat{\boldsymbol{\beta}}$ es llamado el *estimador de mínimos cuadrados*.. Note que cada $\hat{\beta}_j$ en $\hat{\boldsymbol{\beta}}$ es una función lineal de $\boldsymbol{y}$; esto es, $\hat{\beta}_j = \mathbf{a}_j'\boldsymbol{y}$ donde $\mathbf{a}_j'$ es la $j - esima$ fila de $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

Ahora demostraremos que $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ minimiza $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$.

$$\begin{aligned}
\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&\quad +2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}})
\end{aligned} \tag{3.3}$$

El tercer termino de lado derecho de 3.3 desaparece ya que de las ecuaciones normales $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$. El segundo termino es una forma cuadrática definida positiva (asumiendo que $\boldsymbol{X}$ es de rango completo), por lo tanto $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ es minimizada cuando $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

Actividad: Determinar los estimadores de mínimos cuadrados en el caso del modelo lineal simple.

La estructura de la matriz $\boldsymbol{X'X}$ y del vector $\boldsymbol{X'y}$ es la siguiente

$$\begin{aligned}\boldsymbol{X'X} &= \begin{pmatrix}1 & x_{11} & x_{12} & \dots & x_{1k}\\ 1 & x_{21} & x_{22} & \dots & x_{2k}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & x_{n2} & \dots & x_{nk}\end{pmatrix}' \begin{pmatrix}1 & x_{11} & x_{12} & \dots & x_{1k}\\ 1 & x_{21} & x_{22} & \dots & x_{2k}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & x_{n2} & \dots & x_{nk}\end{pmatrix}\\ &= \begin{pmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} & \dots & \sum_i x_{ik}\\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \dots & \sum_i x_{i1}x_{ik}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \sum_i x_{ik} & \sum_i x_{i1}x_{ik} & \sum_i x_{i2}x_{ik} & \dots & \sum_i x_{ik}^2\end{pmatrix}\end{aligned}$$

$$\begin{aligned}\boldsymbol{X'y} &= \begin{pmatrix}1 & x_{11} & x_{12} & \dots & x_{1k}\\ 1 & x_{21} & x_{22} & \dots & x_{2k}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & x_{n2} & \dots & x_{nk}\end{pmatrix}\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_k\end{pmatrix}\\ &= \begin{pmatrix}\sum_i y_i\\ \sum_i x_{i1}y_i\\ \vdots\\ \sum_i x_{ik}y_i\end{pmatrix}\end{aligned}$$

Si $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$, entonces

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{3.4}$$

es el vector de residuales, $\hat{\varepsilon}_1 = y_1 - \hat{y}_1$, $\hat{\varepsilon}_2 = y_2 - \hat{y}_2$,..., $\hat{\varepsilon}_n = y_n - \hat{y}_n$. El vector de residuales $\hat{\boldsymbol{\varepsilon}}$ estima a ε en el modelo $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ y puede ser usado para chequear la validez del modelo y evaluar los supuestos. Esto se verá mas adelante.

EJEMPLO 3.1 Estimación de los parámetros en el Modelo de Regresión Lineal Simple

Considere el modelo lineal simple

$$\begin{aligned}&Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i;\ i = 1, 2, ..., n;\ n \geq 2\\ &\varepsilon_i \sim NID(0, \sigma^2)\\ &\Omega = \{(\beta_0, \beta_1, \sigma^2) : \beta_0 \in E_1, \beta_1 \in E_1, \sigma^2 > 0\}\end{aligned} \tag{3.5}$$

Cuya forma matricial es

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

$$\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix} \\ &= \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + n\sum_{i=1}^{n} x_i y_i \end{pmatrix} \\ &= \begin{pmatrix} \bar{y} - \hat{\beta}_1 \bar{x} \\ \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{pmatrix} \end{aligned}$$

Así,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.6}$$

EJEMPLO 3.2

Usando los datos de la tabla 2.2 ilustramos los cálculos de $\hat{\boldsymbol{\beta}}$ usando (3.2). Los datos matricialmente son lo siguientes

$$\boldsymbol{y} = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 7 \\ 6 \\ 8 \\ 10 \\ 7 \\ 8 \\ 12 \\ 11 \\ 14 \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 6 \\ 1 & 2 & 7 \\ 1 & 2 & 5 \\ 1 & 4 & 9 \\ 1 & 4 & 8 \\ 1 & 4 & 7 \\ 1 & 6 & 10 \\ 1 & 6 & 11 \\ 1 & 6 & 9 \\ 1 & 8 & 15 \\ 1 & 8 & 13 \end{pmatrix}$$

entonces,

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 12 & 52 & 102 \\ 52 & 395 & 536 \\ 102 & 536 & 1004 \end{pmatrix}, \qquad \begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix}$$

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 0.97476 & 0.24290 & -0.22871 \\ 0.24290 & 0.16207 & -0.11120 \\ -0.22871 & -0.11120 & 0.08360 \end{pmatrix}$$

por lo tanto,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} 5.3754 \\ 3.0118 \\ -1.2855 \end{pmatrix}$$

3.1.1.2 Propiedades del Estimador de Mínimos Cuadrados $\hat{\boldsymbol{\beta}}$ El estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}$ en el teorema 3.1 fue obtenido sin usar los supuestos de que $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ y $cov(\boldsymbol{y}) = \sigma^2\mathbf{I}$ dados en la sección 2.2.2. Solamente se postuló un modelo $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ y se ajustó. Si $E(\boldsymbol{y}) \neq \boldsymbol{X}\boldsymbol{\beta}$, el modelo $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ podría no ajustarse a los datos, en cuyo caso, $\hat{\boldsymbol{\beta}}$ podría tener propiedades pobres. Si $cov(\boldsymbol{y}) \neq \sigma^2\mathbf{I}$, esto podría tener efectos

adversos adicionales sobre $\hat{\boldsymbol{\beta}}$. Sin embargo, si $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ y $cov(\boldsymbol{y}) = \sigma^2\mathbf{I}$, $\hat{\boldsymbol{\beta}}$ tiene algunas propiedades buenas, como se notará en los cuatro teoremas de esta sección. Note que $\hat{\boldsymbol{\beta}}$ es un vector aleatorio. Discutimos su vector de media y matriz de covarianza (sin el supuesto de la distribución de $\boldsymbol{y}$) y su distribución (asumiendo que las variables y son normales) se verá más adelante. Nuevamente en los siguientes teoremas, asumimos que $\boldsymbol{X}$ es fija y de rango completo.

Teorema 3.2 *Si $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$, entonces $\hat{\boldsymbol{\beta}}$ es un estimador insesgado para $\boldsymbol{\beta}$*

Prueba.

$$E(\hat{\boldsymbol{\beta}}) = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

Teorema 3.3 *Si $cov(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$, entonces la matriz de covarianza para $\hat{\boldsymbol{\beta}}$ está dado por $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$*

Prueba.

$$\begin{aligned}
cov(\hat{\boldsymbol{\beta}}) &= cov[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] \\
&= [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']cov(y)[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' && (3.7)\\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\mathbf{I}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} && (3.8)\\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} && (3.9)\\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} && (3.10)
\end{aligned}$$

EJEMPLO 3.3

Para los datos de la tabla 2.2, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ esta dada en el ejemplo 3.2. Por lo tanto, $cov(\hat{\boldsymbol{\beta}})$ está dada por

$$cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2\begin{pmatrix} 0.97476 & 0.24290 & -0.22871 \\ 0.24290 & 0.16207 & -0.11120 \\ -0.22871 & -0.11120 & 0.08360 \end{pmatrix}$$

El valor negativo de $cov(\hat{\beta}_1, \hat{\beta}_2) = -0.11120$ indica que en muestreos repetidos (usando los mismos 12 valores de x_1 y x_2), $\hat{\beta}_1$ y $\hat{\beta}_2$ tenderán a moverse en direcciones opuestas; es decir, un incremento en una debería estar acompañada por un decrecimiento en la otra.

Además de que $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ y $cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, una tercera propiedad importante de $\hat{\boldsymbol{\beta}}$ es que bajo los supuestos estandar, la varianza de cada $\hat{\beta}_j$ es mínima. Veamoslo en el siguiente teorema.

Teorema 3.4 (Teorema de Gauss-Markov) *Si $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ y $cov(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$, el estimador de mínimos cuadrados $\hat{\beta}_j$, $j = 0, 1, ..., k$ tiene varianza mínima entre entre todos los estimadores lineales insesgados.*

Prueba. La prueba se consigue en la página 147 del Rencher (2008).

El teorema de Gauss-Markov es algunas veces establecido de la siguiente manera. Si $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ y $cov(\boldsymbol{y}) = \sigma^2\mathbf{I}$, los estimadores de mínimo cuadrado $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ son los *mejores estimadores lineales insesgados (BLUE)*. En esta expresión, *mejor* significa varianza mínima y *lineal* indica que los estimadores son funciones lineales de $\boldsymbol{y}$.

La característica resaltante del teorema de Gauss-Markov es su generalidad distribucional. El resultado se obtiene para cualquier distribución de $\boldsymbol{y}$; la normalidad no es requerida. Los únicos supuestos usados en la prueba son $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ y $cov(\boldsymbol{y}) = \sigma^2\mathbf{I}$. Si estos supuestos no se cumplen, $\hat{\boldsymbol{\beta}}$ podría ser sesgado o cada $\hat{\beta}_j$ podría tener varianza más grande que algún otro estimador.

El teorema de Gauss-Markov se puede extender fácilmente a una combinación lineal de los $\hat{\boldsymbol{\beta}}$ como sigue

Corolario 3.1 *Si $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ y $cov(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$, el mejor estimador lineal insesgado de $\boldsymbol{a}'\boldsymbol{\beta}$ es $\boldsymbol{a}'\hat{\boldsymbol{\beta}}$, donde $\hat{\boldsymbol{\beta}}$ es el estimador de mínimo cuadrados.*

3.1.1.3 Estimador para σ^2 La minimización de la suma de cuadrados del error $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ no provee un estimador de σ^2; sin embargo, podemos obtener un estimador insesgado de σ^2 basado en el estimador de mínimos cuadrados de $\boldsymbol{\beta}$. Por el supuesto de que $cov(\boldsymbol{y}) = \sigma^2\mathbf{I}$, se tiene que σ^2 es el mismo para cada y_i, $i = 1, 2, ..., n$. Sabemos que σ^2 esta definido por $\sigma^2 = E[y_i - E(y_i)]^2$, y como

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \boldsymbol{x}_i'\boldsymbol{\beta}$$

donde $\boldsymbol{x}_i'$ es la $i - esima$ fila de $\boldsymbol{X}$. Por lo tanto

$$\sigma^2 = E[y_i - \boldsymbol{x}_i'\boldsymbol{\beta}]^2$$

Podemos estimar σ^2 por

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2 \tag{3.11}$$

donde n es el tamaño de la muestra y p es el numero de parámetros. En terminos matriciales se tiene

$$s^2 = \frac{1}{n-p}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{3.12}$$

lo cual es igual a

$$s^2 = \frac{\boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}}{n-p} = \frac{SCE}{n-p} \tag{3.13}$$

donde $SCE = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$. Con el denominador $n - p$, s^2 es un estimador insesgado de σ^2, lo cual se demuestra a continuación.

Teorema 3.5 *Si s^2 es definido por (3.11), (3.12) o (3.13) y si $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$ y $cov(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$, entonces*

$$E(s^2) = \sigma^2 \tag{3.14}$$

Prueba. SCE se puede escribir como la siguiente forma cuadratica

$$SCE = \boldsymbol{y}'\boldsymbol{y} - [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}]'\boldsymbol{X}'\boldsymbol{y} \tag{3.15}$$

$$= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{y}'[\mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{y} \tag{3.16}$$

Aplicando la Esperanza de una forma cuadratica, se tiene que

$$\begin{aligned} E(SCE) &= tr\{[\mathbf{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\sigma^2\mathbf{I}\} + E(\boldsymbol{y}')[\mathbf{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']E(\boldsymbol{y}) \\ &= \sigma^2 tr[\mathbf{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'] + \boldsymbol{\beta}'\boldsymbol{X}'[\mathbf{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{X}\boldsymbol{\beta} \\ &= \sigma^2\{n - tr[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \sigma^2\{n - tr[\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}]\} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \sigma^2[n - tr(\mathbf{I}_p)] = \sigma^2(n-p) \end{aligned}$$

despejando se obtiene el resultado esperado.

Corolario 3.2 *Un estimador insesgado de* $cov(\hat{\boldsymbol{\beta}})$ *está dado por*

$$\hat{cov}(\hat{\boldsymbol{\beta}}) = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{3.17}$$

Note la correspondencia entre $n-p$ y $\boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$; hay n terminos en $\boldsymbol{y}'\boldsymbol{y}$ y p términos en $\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$. Una propiedad de la muestra es que cada x (y $\hat{\boldsymbol{\beta}}$) adicional en el modelo reduce la SCE.

Ya que SCE es una función cuadrática de $\boldsymbol{y}$ este no es un mejor estimador lineal insesgado. La propiedad óptima de s^2 está dada en el siguiente teorema.

Teorema 3.6 *Si* $E(\boldsymbol{\varepsilon}) = 0$, $cov(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$ *y* $E(\varepsilon_i^4) = 3\sigma^4$ *para el modelo lineal* $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, *entonces* s^2 *en (3.12) o (3.13) es el mejor (mínima varianza) estimador cuadrático insesgado de* σ^2.

Prueba. Ver Graybill (1954).

EJEMPLO 3.4

Para los datos en 2.2, tenemos

$$\begin{aligned} SCE &= \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} \\ &= 840 - \begin{pmatrix} 5.3754 & 3.0118 & -1.2855 \end{pmatrix} \begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix} \\ &= 840 - 814.541 = 25.459 \end{aligned}$$

entonces

$$s^2 = \frac{SCE}{n-p} = \frac{25.459}{12-3} = 2.829$$

Nota 3.1 *Las ecuaciones para estimar* $\boldsymbol{\beta}$ *y* σ^2 *son a menudo llamadas ecuaciones normales o ecuaciones de mínimos cuadrados. el primer termino, comúnmente usado, se refiere a una interpretación geométrica de la maximización. El segundo termino surge al notar que estas son las ecuaciones estacionarias para la minimización de* $Q(\boldsymbol{\beta}) = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$ *con respecto a* $\boldsymbol{\beta}$ *y reconociendo que* $Q(\boldsymbol{\beta})$ *es la suma de las diferencias cuadráticas entre las observaciones y sus valores esperados. Así,* $\tilde{\boldsymbol{\beta}}$ *es también llamado el estimador de mínimos cuadrados de* $\boldsymbol{\beta}$. *Una interpretación geométrica de la estimación de mínimos cuadrados está dada en el capítulo 2 en la discusión de modelos de regresión de Hocking.*

3.1.2 Geometría de los Mínimos Cuadrados

En las secciones anteriores presentamos el modelo de regresión lineal múltiple como la ecuación matricial $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Definimos el principio de la estimación de mínimos cuadrados en términos de desviaciones desde el modelo, y luego usando calculo matricial y álgebra de matrices derivamos los estimadores de $\boldsymbol{\beta}$ y σ^2. Ahora presentamos una derivación alterna pero equivalente de estos estimadores basado completamente en ideas geométricas.

Es importante aclarar en primer lugar lo que no es el enfoque geométrico de los mínimos cuadrados. En dos dimensiones, ilustramos el principio de mínimos cuadrados al crear un gráfico de dispersión de dos dimensiones de los n puntos $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Luego visualizamos la recta de regresión de mínimos cuadrados como la línea recta que mejor se ajusta a los datos. Este enfoque puede generalizarse para presentar la estimación de mínimos cuadrados en regresión lineal múltiple sobre la base del hiperplano en un espacio p-dimensional que mejor se ajusta a los n puntos $(x_{11}, x_{12}, ..., x_{1k}, y_1)$, $(x_{21}, x_{22}, ..., x_{2k}, y_2), ..., (x_{n1}, x_{n2}, ..., x_{nk}, y_n)$. Aunque este enfoque es algo usado en visualizar regresión lineal múltiple, el enfoque geométrico para la estimación de mínimos cuadrados en regresión lineal múltiple no envuelve esta generalización de alta dimensión.

El enfoque geométrico discutido más abajo es atractivo debido a su elegancia matemática. Por ejemplo, el estimador es derivado sin el uso de calculo matricial, Además, el enfoque geométrico proporciona una visión más profunda de la inferencia estadística. Varios métodos de estadística avanzada incluidos suavización kernel (Eubank y Eubank 1999), análisis de Fourier (Bloomfield 2000), y análisis Wavelet (Ogden 1997) pueden entenderse como generalizaciones de este enfoque geométrico. El enfoque geométrico de modelos lineales fue primero propuesto por Fisher (Mahalanobis 1964). Christensen (1996) y Jammalamadaka y Sengupta (2003) discuten el modelo estadístico lineal casi completamente desde la perspectiva geométrica.

3.1.2.1 Espacio del Parámetro, Espacio de Datos y Espacio de Predicción El enfoque geométrico de mínimos cuadrados comienza con dos espacios de altas dimensiones, un espacio $p-$dimensional y un espacio $n-$dimensional. El vector de parámetros desconocido $\boldsymbol{\beta}$ puede verse como un punto en el espacio $p-$dimensional, con ejes correspondientes a los p coeficientes de regresión $\beta_0, \beta_1, ..., \beta_k$. Por lo tanto llamamos a este espacio el *espacio del parámetro* (ver figura ***). De manera similar, el vector de datos $\boldsymbol{y}$ puede verse como un punto en el espacio $n-$dimensional con ejes correspondientes a las n observaciones. Llamamos este espacio el *espacio de datos*.

La matriz $\boldsymbol{X}$ del modelo de regresión múltiple puede escribirse como una matriz particionada en términos de sus p columnas como

$$\boldsymbol{X} = \begin{pmatrix} \mathbf{j} & \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & ... & \boldsymbol{x}_k \end{pmatrix}$$

Las columnas de $\boldsymbol{X}$, incluyendo a $\mathbf{j}$, son todos vectores n-dimensionales y son, por lo tanto, puntos en el espacio de datos. Note que debido a que asumimos que $\boldsymbol{X}$ es de rango p, estos vectores son linealmente independientes. El conjunto de todas las posibles combinaciones lineales de las columnas de $\boldsymbol{X}$ constituyen un subconjunto del espacio de datos (ver sección 2.3 del Rencher). Elementos de este subconjunto pueden escribirse como

$$\boldsymbol{X}\mathbf{b} = b_0\mathbf{j} + b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + \cdots + b_k\boldsymbol{x}_k$$

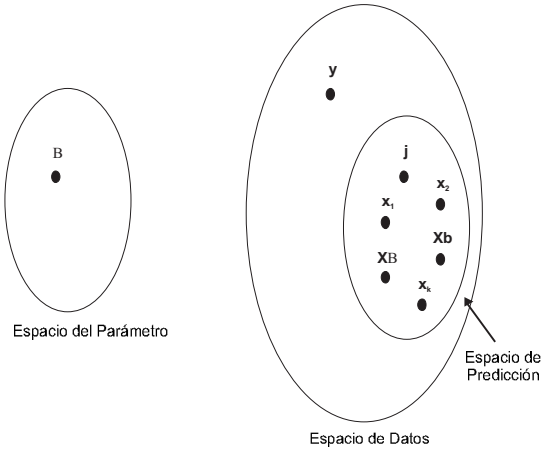

Figure 3.1 Espacio del parámetro, espacio de datos y espacio de predicción con elementos representativos

donde **b** es cualquier vector $p \times 1$, es decir, es cualquier vector en el espacio del parámetro. Este subconjunto tiene el estatus de subespacio ya que este es cerrado bajo la adición y la multiplicación escalar (Harville 1997, pp. 28 - 29). Este subconjunto se dice ser el subespacio generado por las columnas de $\boldsymbol{X}$, y nosotros llamaremos este subespacio como el *espacio de predicción*. Las columnas de $\boldsymbol{X}$ constituyen un *conjunto base* para el espacio de predicción.

3.1.2.2 Interpretación Geométrica del Modelo de Regresión Lineal Múltiple El modelo de regresión lineal múltiple establece que $\boldsymbol{y}$ es igual a un vector en el espacio de predicción, $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$, mas un vector de errores aleatorios, $\boldsymbol{\varepsilon}$ (figura 3.2). El problema es que ni $\boldsymbol{\beta}$ ni $\boldsymbol{\varepsilon}$ son conocidos. Sin embargo, el vector de datos $\boldsymbol{y}$, el cual no está en el espacio de predicción, es conocido. Y se sabe que $E(\boldsymbol{y})$ está en el espacio de predicción.

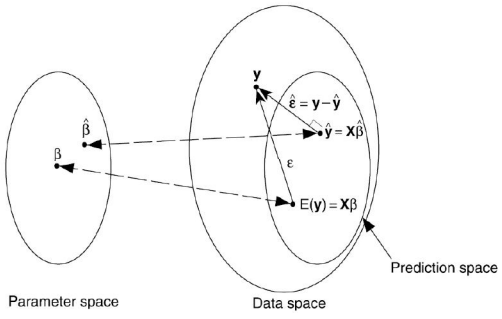

Figure 3.2 Relación geométrica de los vectores asociados con el modelo de regresión lineal múltiple

La regresión lineal múltiple puede entenderse geométricamente como el proceso de hallar una estimación sensible de $E(\boldsymbol{y})$ en el espacio de predicción y luego determinar el vector

en el espacio del parámetro que esta asociado con esta estimación (figura 3.2). La estimación de $E(\boldsymbol{y})$ es denotada como $\hat{\boldsymbol{y}}$, y el vector asociado en el espacio del parámetro es denotado por $\hat{\boldsymbol{\beta}}$.

Una idea geométrica razonable es estimar $E(\boldsymbol{y})$ usando el punto en el espacio de predicción que esta mas cercano a $\boldsymbol{y}$. Resulta en que $\hat{\boldsymbol{y}}$, el punto más cercano en el espacio de predicción a $\boldsymbol{y}$, puede hallarse al notar que el vector diferencia $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ debe ser ortogonal (perpendicular) al espacio de predicción (Harville 1997, p. 170). Además, ya que el espacio de predicción es generado por las columnas de $\boldsymbol{X}$, el punto $\hat{\boldsymbol{y}}$ debe ser tal que $\hat{\boldsymbol{\varepsilon}}$ es ortogonal a las columnas de $\boldsymbol{X}$. Es decir, que $\hat{\boldsymbol{y}}$ es tal que

$$\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = 0$$

o

$$\boldsymbol{X}'(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0$$

lo cual implica que

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

Por lo tanto, usando ideas solamente geométricas, obtenemos las ecuaciones normales y en consecuencia el usual estimado1r de mínimos cuadrados $\hat{\boldsymbol{\beta}}$. Podemos entonces calcular $\hat{\boldsymbol{y}}$ como $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$. Además, $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \boldsymbol{H})\boldsymbol{y}$ puede ser tomado como un estimador de $\boldsymbol{\varepsilon}$. Ya que $\hat{\boldsymbol{\varepsilon}}$ es un vector en el espacio $(n-p))$ dimensional, es razonable estimar a σ^2 como el cuadrado de la longitud de $\hat{\boldsymbol{\varepsilon}}$ dividido por $n - p$. En otras palabras, un estimador sensible de σ^2 es $s^2 = \boldsymbol{y}'(\mathbf{I} - \boldsymbol{H})\boldsymbol{y}/(n - p)$.

3.1.3 Estimadores por el método de máxima verosimilitud

.
Hasta el momento no se ha hecho ningún supuesto sobre la distribución de las variables aleatorias $y_1, y_2, ..., y_n$.

Supongamos que $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$ lo cual es equivalente a $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Bajo este supuesto podemos entonces hallar los estimadores máximo verosimiles de $\boldsymbol{\beta}$ y σ^2.
Como $\boldsymbol{Y}$ se distribuye $N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, la función de verosimilitud es:

$$\begin{aligned} L &= L(\boldsymbol{\beta}, \sigma^2 : y_1, y_2, ..., y_n; x_1, x_2, ..., x_n) = L(\boldsymbol{\beta}, \sigma^2 : \boldsymbol{y}; \boldsymbol{X}) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} exp\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\} \end{aligned} \tag{3.18}$$

Ya que L es una función cuyo rango es positivo y la función logaritmo es una función monótona que mantiene el orden, es decir si $f(x)$ alcanza un máximo en x_0, entonces, $\log f(x)$ alcanza también el máximo en x_0, por lo tanto, en vez de hallar el máximo en L, se hallará el máximo del $\log L$, pues los cálculos son más fáciles. Así,

$$\begin{aligned}
\log L(\boldsymbol{\beta}, \sigma^2 : \boldsymbol{y}; \boldsymbol{X}) &= \log\left\{\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right]\right\} \\
&= \log\left\{\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\right\} + \log\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\right] \\
&= \frac{n}{2}\log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \quad (3.19)
\end{aligned}$$

El espacio del parámetro es

$$\Omega = \{(\boldsymbol{\beta}, \sigma^2) : \sigma^2 > 0, -\infty < \beta_i < \infty; i_1, 2, ..., p\} \quad (3.20)$$

Para hallar los valores de $\boldsymbol{\beta}$ y σ^2 en Ω que maximizan la función de verosimilitud se iguala a cero las derivadas parciales con respecto a $\boldsymbol{\beta}$ y σ^2.

Antes de derivar se reescribe la ecuación

$$\log L = \frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta})$$

Haciendo $\boldsymbol{V} = \boldsymbol{y}'\boldsymbol{X}$ y $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{X}$

$$\log L = \frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{V}'\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{V} + \boldsymbol{\beta}'\boldsymbol{W}\boldsymbol{\beta})$$

Derivando con respecto a $\boldsymbol{\beta}$ y σ^2 e igualando a cero (ver **??**), se obtiene

$$\begin{aligned}
\frac{\partial \log L}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma^2}(-\boldsymbol{V} - \boldsymbol{V} + 2\boldsymbol{W}\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}(-2\boldsymbol{V} + 2\boldsymbol{W}\boldsymbol{\beta}) \\
&= \frac{1}{\sigma^2}(\boldsymbol{V} - \boldsymbol{W}\boldsymbol{\beta}) = \frac{1}{\sigma^2}(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}) = 0 \quad (3.21) \\
\frac{\partial \log L}{\partial \sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0 \quad (3.22)
\end{aligned}$$

Sean $\tilde{\boldsymbol{\beta}}$ y $\tilde{\sigma}^2$ las soluciones de las ecuaciones para $\boldsymbol{\beta}$ y σ^2 cuando las derivadas son en conjunto iguales a cero. Entonces

$$\frac{1}{\tilde{\sigma}^2}(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}}) = 0 \Rightarrow \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} = 0 \Rightarrow \boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \quad (3.23)$$

$$-\frac{n}{2}\frac{1}{\tilde{\sigma}^2} = -\frac{1}{2\hat{\sigma^4}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \Rightarrow \tilde{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) \quad (3.24)$$

Ahora como $\boldsymbol{X}$ es de rango p, entonces $\boldsymbol{X}'\boldsymbol{X}$ es también de rango p la cual es su dimensión. Así, $\boldsymbol{X}'\boldsymbol{X}$ es invertible. (ver **??** y **??**)
Por lo tanto, los estimadores máximos verosímiles de $\boldsymbol{\beta}$ y σ^2 son

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{3.25}$$

$$\begin{aligned}
\tilde{\sigma}^2 &= \frac{1}{n}\left[\boldsymbol{y}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right]'\left[\boldsymbol{y}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right] \\
&= \frac{1}{n}\left[\boldsymbol{y}'-\boldsymbol{y}'\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^{-1}]'\boldsymbol{X}'\right]\left[\boldsymbol{y}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right] \\
&= \frac{1}{n}\left[\boldsymbol{y}'-\boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\left[\boldsymbol{y}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right] \\
&= \frac{1}{n}\boldsymbol{y}'\left[\mathbf{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\left[\mathbf{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\boldsymbol{y}
\end{aligned} \tag{3.26}$$

Sea $\boldsymbol{K} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Se puede notar que $\boldsymbol{K}$ es idempotente (ver **??**), pues

$$\boldsymbol{K}\boldsymbol{K} = \boldsymbol{X}\overbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}^{\mathbf{I}}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{K}$$

Entonces,

$$\tilde{\sigma}^2 = \frac{1}{n}\boldsymbol{y}'(\mathbf{I}-\boldsymbol{K})'(\mathbf{I}-\boldsymbol{K})\boldsymbol{y}$$

Además,

$$(\mathbf{I}-\boldsymbol{K})'(\mathbf{I}-\boldsymbol{K}) = \mathbf{I}-\boldsymbol{K}-\boldsymbol{K}+\boldsymbol{K}\boldsymbol{K} = \mathbf{I}-\boldsymbol{K}-\boldsymbol{K}+\boldsymbol{K} = \mathbf{I}-\boldsymbol{K}$$

Así,

$$\tilde{\sigma}^2 = \frac{1}{n}\boldsymbol{y}'(\mathbf{I}-\boldsymbol{K})\boldsymbol{y}$$

Haciendo $\boldsymbol{M} = \mathbf{I}-\boldsymbol{K}$ se tiene que

$$\tilde{\sigma}^2 = \frac{1}{n}\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}$$

Para demostrar que $\tilde{\boldsymbol{\beta}}$ y $\tilde{\sigma}^2$ son los estimadores máximo verosímiles de $\boldsymbol{\beta}$ y σ^2 se debe demostrar que ellos maximizan la función de verosimilitud (La prueba esta dada en la sección 9.3 del Graybill).

3.1.3.1 Propiedades de los estimadores máximo verosimiles Ahora se examinan las propiedades de los estimadores $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ y se determinan sus distribuciones.

3.1.3.2 Suficiencia Primero se demostrará que los $p+1$ estimadores $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p, \tilde{\sigma}^2$ son estadísticos suficientes. Para ello se usa el siguiente teorema

Teorema 3.7 *Sea $f_y(y;\theta)$ la fdp de una muestra observable $y_1, y_2, ..., y_n$. Los estadísticos $S_1, S_2, ..., S_r$ son estadísticos suficientes si y sólo si la fdp puede factorizarse como*

$$f_Y(y;\theta) = g(S_1, S_2, ..., S_r;\theta)h(y)$$

donde $g(S_1, S_2, ..., S_r; \theta)$ es una función no negativa de solamente los estadísticos $S_1, S_2, ..., S_r$ y el parámetro θ, y $h(y)$ no es una función de θ.

Entonces se quiere demostrar que la fdp de $\boldsymbol{y}$, llámese $f(\boldsymbol{y} : \boldsymbol{X}; \boldsymbol{\beta}; \sigma^2)$, se puede factorizar como $g(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2 : \boldsymbol{\beta}; \sigma^2)h(\boldsymbol{y})$ donde $g(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2 : \boldsymbol{\beta}; \sigma^2)$ contiene las observaciones $\boldsymbol{y}$ en la forma de $\tilde{\boldsymbol{\beta}}$ y $\tilde{\sigma}^2$ solamente, y $h(\boldsymbol{y})$ no contiene los parámetros $\boldsymbol{\beta}$ y σ^2.

La fdp de $\boldsymbol{y}$ está dada por

$$f(\boldsymbol{y} : \boldsymbol{X}; \boldsymbol{\beta}; \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right] \tag{3.27}$$

Trabajando con una porción del exponente se obtiene que

$$\begin{aligned}
(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \left[(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\right]' \left[(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\right] \\
&= \left[(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]' \left[(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right] \\
&= \left[(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})' - (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\right] \left[(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right] \\
&= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\
&\quad -(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\
&= n\tilde{\sigma}^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})
\end{aligned} \tag{3.28}$$

ya que,

$$\begin{aligned}
(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{X} &= \boldsymbol{y}'\boldsymbol{X} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{y}'\boldsymbol{X} - [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}]'\boldsymbol{X}'\boldsymbol{X} \\
&= \boldsymbol{y}'\boldsymbol{X} - \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{y}'\boldsymbol{X} - \boldsymbol{y}'\boldsymbol{X} = 0
\end{aligned}$$

y de la misma manera, $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0$.

Sustituyendo 3.28 en 3.27 se obtiene

$$\begin{aligned}
f(\boldsymbol{y} : \boldsymbol{X}; \boldsymbol{\beta}; \sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} exp\left[-\frac{1}{2\sigma^2}[n\tilde{\sigma}^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})]\right] \\
&= g(\hat{\boldsymbol{\beta}}, \tilde{\sigma}^2 : \boldsymbol{\beta}; \sigma^2)h(\boldsymbol{y})
\end{aligned}$$

donde $h(\boldsymbol{Y}) = 1$. Por lo tanto, $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ son estadísticos suficientes.

3.1.3.3 Completitud. Ahora se probará que $\tilde{\boldsymbol{\beta}}$ y $\tilde{\sigma}^2$ son completos. Para ello es necesario recordar la siguiente definición y el subsiguiente teorema

Definición 3.1 (Familia Exponencial de Densidades) *Sea $f_y(y; \theta), \theta \in \Omega$ la fdp conjunta de una muestra observable $y_1, y_2, ..., y_n$; donde $\theta' = [\theta_1, \theta_2, ..., \theta_k]$. La fdp $f_y(y; \theta)$*

se define a pertenecer a una familia exponencial de densidades si esta puede escribirse como

$$f_y(y;\theta) = h(\theta)g(y)\exp\left[\sum_{j=1}^{k} S_j(y)P_j(\theta)\right] \quad a_i < y_i < b_i; i = 1,2,...,n \qquad (3.29)$$

donde

1. a_i *y* b_i *no dependen de los* θ_j, a_i *puede ser* $-\infty$ *y* b_i *puede ser* $+\infty$.
2. Ω, *el espacio del parámetro, contiene un espacio* K*-dimensional no degenerado (rectángulo).*
3. $h(\theta)$ *y los* $P_j(\theta)$ *no dependen de los* y_j.
4. $g(y)$ *y los* $S_j(y)$ *no dependen de los* θ_i; $g(y)$ *es no negativa.*

Teorema 3.8 *Sea* $Y_1, Y_2, ..., Y_n$ *una muestra observable con fdp* $f_Y(\boldsymbol{y};\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Omega$. *Supóngase que* $f_Y(y;\theta)$ *pertenece a una familia exponencial, esto es,* $f_Y(\boldsymbol{y};\theta)$ *puede escribirse en la forma dada en la ecuación 3.29 y satisface a,b,c y d. Supóngase que las siguientes condiciones también se satisfacen*

1. *Los* $S_j(y)$ *son funcionalmente independientes para* $j = 1,2,...,k$
2. $\partial[S_j(y)]/\partial y_i$ *existe y es continua para todos los* $j = 1,2,...,k$ *y* $i = 1,2,...,n$
3. $P_j(\theta)$ *es una función continua de* θ *para* $j = 1,2,...,k$
4. *Si se define* $\gamma' = [P_1(\theta), P_2(\theta), ..., P_k(\theta)]$, *entonces el conjunto de valores de* γ *(Como* θ *varía sobre* Ω*) contiene un rectángulo K dimensional no degenerado.*

Entonces el conjunto de K estadísticos $S_1(y), S_2(y), ..., S_k(y)$, *es un conjunto de estadísticos mínimo suficientes.*

Ahora, para establecer la completitud se enuncia el siguiente teorema

Teorema 3.9 *Sean* $y_1, y_2, ..., y_n$ *una muestra observable con fdp* $f_Y(y;\theta)$, $\theta \in \Omega$. *Si* $f_Y(y;\theta)$ *pertenece a una familia exponencial como se define en 3.1 y si las condiciones a,b,c,d del teorema 3.8 se satisfacen, entonces los K estadísticos* $S_1(y), S_2(y), ..., S_k(y)$, *son estadísticos suficientes completos.*

Por lo tanto, para probar la completitud primero se debe probar que $f(\boldsymbol{y} : \boldsymbol{X};\boldsymbol{\beta};\sigma^2)$ pertenece a una familia exponencial, es decir, $f(\boldsymbol{y} : \boldsymbol{X};\boldsymbol{\beta};\sigma^2)$ debe tener la forma

$$f(\boldsymbol{y} : \boldsymbol{X};\boldsymbol{\beta};\sigma^2) = h(\boldsymbol{\beta},\sigma^2)g(\boldsymbol{y},\boldsymbol{X})\exp\left[S_1(\boldsymbol{y},\boldsymbol{X})P_1(\boldsymbol{\beta},\sigma^2) + S_2(\boldsymbol{y},\boldsymbol{X})P_2(\boldsymbol{\beta},\sigma^2)\right] \qquad (3.30)$$

La fdp de $\boldsymbol{y}$ es

$$
\begin{aligned}
f(\boldsymbol{y}:\boldsymbol{X};\boldsymbol{\beta};\sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right] \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} exp\left[-\frac{\boldsymbol{y'y}}{2\sigma^2}+\frac{2\boldsymbol{y'X\beta}}{2\sigma^2}-\frac{\boldsymbol{\beta'X'X\beta}}{2\sigma^2}\right] \\
&= \left(\frac{1}{2\pi}\right)^{n/2}\left(\frac{1}{\sigma^2}\right)^{n/2} exp\left[-\frac{\boldsymbol{\beta'X'X\beta}}{2\sigma^2}\right] exp\left[-\frac{\boldsymbol{y'y}}{2\sigma^2}+\frac{2\boldsymbol{y'X\beta}}{2\sigma^2}\right] \\
&= \left(\frac{1}{2\pi}\right)^{n/2}\left(\frac{1}{\sigma^2}\right)^{n/2} exp\left[-\frac{\boldsymbol{\beta'X'X\beta}}{2\sigma^2}\right] \\
&\quad exp\left[\boldsymbol{-y'y}\frac{1}{2\sigma^2}+\boldsymbol{y'X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'X}\frac{\boldsymbol{\beta}}{\sigma^2}\right] \\
&= \left(\frac{1}{2\pi}\right)^{n/2}\left(\frac{1}{\sigma^2}\right)^{n/2} exp\left[-\frac{\boldsymbol{\beta'X'X\beta}}{2\sigma^2}\right] exp\left[\boldsymbol{-y'y}\frac{1}{2\sigma^2}+\hat{\boldsymbol{\beta}}'\boldsymbol{X'X}\frac{\boldsymbol{\beta}}{\sigma^2}\right]
\end{aligned}
$$

Así,

$h(\boldsymbol{\beta},\sigma^2)=\left(\frac{1}{\sigma^2}\right)^{n/2} exp\left[-\frac{\boldsymbol{\beta'X'X\beta}}{2\sigma^2}\right]$, $g(\boldsymbol{y},\boldsymbol{X})=\left(\frac{1}{2\pi}\right)^{n/2}$, $S_1(\boldsymbol{y})=\boldsymbol{y'y}$, $P_1(\boldsymbol{\beta},\sigma^2)=-\frac{1}{2\sigma^2}$, $S_2(\boldsymbol{Y})=\hat{\boldsymbol{\beta}}'$ y $P_2(\boldsymbol{\beta},\sigma^2)=\boldsymbol{X'X}\frac{\boldsymbol{\beta}}{\sigma^2}$.

Por lo tanto, $\boldsymbol{y'y}$ y $\hat{\boldsymbol{\beta}}$ son completos.

3.1.3.4 Distribución de $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$

Distribución de $\hat{\boldsymbol{\beta}}$. Para hallar la distribución de $\hat{\boldsymbol{\beta}}$ es necesario primero enunciar el siguiente teorema

Teorema 3.10 *Sea $\boldsymbol{X}$ un vector aleatorio $p\times 1$ que se distribuye $N(\mu,\Sigma)$, donde Σ tiene rango k, Sea $\boldsymbol{B}$ cualquier matriz $q\times p$ de constantes, y sea $\boldsymbol{b}$ cualquier vector $q\times 1$ de constantes. Entonces el vector $\boldsymbol{y}$ $q\times 1$ definido por $\boldsymbol{y}=\boldsymbol{BX}+\boldsymbol{b}$ ($\boldsymbol{Y}$ es una función lineal de $\boldsymbol{X}$) se distribuye $N(\boldsymbol{B}\mu+\boldsymbol{b},\boldsymbol{B}\Sigma\boldsymbol{B}')$*

Como $\hat{\boldsymbol{\beta}}=(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$, donde $\boldsymbol{y}\sim N(\boldsymbol{X\beta};\sigma^2\mathbf{I})$ y haciendo $\boldsymbol{B}=(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$ ($\boldsymbol{B}$ es una matriz $p\times n$) y $b=0$, entonces $\hat{\boldsymbol{\beta}}=\boldsymbol{BY}$ es una función lineal de $\boldsymbol{Y}$, por lo tanto, por el teorema 3.10 se tiene que $\hat{\boldsymbol{\beta}}$ se distribuye normal, cuyos parámetros son

$$
E(\hat{\boldsymbol{\beta}})=E[(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}]=(\boldsymbol{X'X})^{-1}\boldsymbol{X'}E(\boldsymbol{Y})=(\boldsymbol{X'X})^{-1}\boldsymbol{X'X\beta}=\boldsymbol{\beta} \tag{3.31}
$$

lo cual indica que $\hat{\boldsymbol{\beta}}$ es un estimador insesgado de $\boldsymbol{\beta}$,

$$
\begin{aligned}
Cov(\tilde{\boldsymbol{\beta}}) &= E[(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}))(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}))'] = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\
&= E(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' - \hat{\boldsymbol{\beta}}\boldsymbol{\beta}' - \boldsymbol{\beta}\hat{\boldsymbol{\beta}}' + \boldsymbol{\beta}\boldsymbol{\beta}') = E(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}') - \boldsymbol{\beta}\boldsymbol{\beta}' - \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\beta}\boldsymbol{\beta}') \\
&= E(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}') - \boldsymbol{\beta}\boldsymbol{\beta}' = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}] - \boldsymbol{\beta}\boldsymbol{\beta}' \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{Y}\boldsymbol{Y}')\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}' \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[Cov(\boldsymbol{Y}) + E(\boldsymbol{y})E(\boldsymbol{y})']\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}' \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[\sigma^2\mathbf{I} + (\boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{X}\boldsymbol{\beta})']\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}' \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}' \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2 + \boldsymbol{\beta}\boldsymbol{\beta}' - \boldsymbol{\beta}\boldsymbol{\beta}' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2 \qquad (3.32)
\end{aligned}
$$

Por lo tanto,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2)$$

Distribución de $\hat{\sigma}^2$. Para hallar la distribución de $\hat{\sigma}^2$ se usa el siguiente teorema

Teorema 3.11 *Sea $\boldsymbol{y}$ un vector aleatorio $n \times 1$ que se distribuye $N(\mu, \Sigma)$, donde Σ tiene rango n. Entonces la forma cuadrática $U = \boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} \sim \chi^2(p, \lambda)$, donde $\lambda = \frac{1}{2}\mu'\boldsymbol{A}\mu$, si y sólo si cualquiera de las siguientes 3 condiciones se satisfacen:*

1. *$\boldsymbol{A}\Sigma$ es una matriz idempotente de rango p.*
2. *$\Sigma\boldsymbol{A}$ es una matriz idempotente de rango p.*
3. *Σ es una c-inversa de $\boldsymbol{A}$ y $\boldsymbol{A}$ tiene rango p.*

Como $\tilde{\sigma}^2$ está dado por

$$\tilde{\sigma}^2 = \frac{1}{n}\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}$$

donde $\boldsymbol{y} \sim N(N(\boldsymbol{X}\boldsymbol{\beta}; \sigma^2\mathbf{I}))$ y $\boldsymbol{M}$ es una matriz simétrica e idempotente $n \times n$ cuyo rango es

$$r(\boldsymbol{M}) = r(\mathbf{I} - \boldsymbol{K}) = tr(\mathbf{I} - \boldsymbol{K}) = tr(\mathbf{I}) - tr(\boldsymbol{K}) = r(\mathbf{I}) - r(\boldsymbol{K}) = n - p$$

entonces por el teorema anterior

$$U = \frac{\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}}{\sigma^2} \sim \chi^2(n - p, \lambda)$$

donde,

$$\begin{aligned}\lambda &= \frac{1}{2}[E(\boldsymbol{y})]'\boldsymbol{M}[E(\boldsymbol{y})] = \frac{1}{2}(\boldsymbol{X}\boldsymbol{\beta})'(\mathbf{I}-\boldsymbol{K})(\boldsymbol{X}\boldsymbol{\beta}) \\ &= \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{X}'[\mathbf{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{X}\boldsymbol{\beta} \\ &= \frac{1}{2}\boldsymbol{\beta}'[\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}] \\ &= \frac{1}{2}\boldsymbol{\beta}'[\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{X}'\boldsymbol{X}]\boldsymbol{\beta} = 0\end{aligned}$$

Por lo tanto,

$$U = \frac{\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}}{\sigma^2} \sim \chi^2(n-p)$$

Dado que el primer momento de la distribución chi-cuadrado central son sus grados de libertad, entonces $E(U) = n - p$, así

$$E(U) = E\left(\frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}\right) = E\left(\frac{1}{\sigma^2}n\tilde{\sigma}^2\right) = n-p$$

entonces,

$$E(\tilde{\sigma}^2) = \frac{n-p}{n}\sigma^2$$

lo cual indica que $\tilde{\sigma}^2$ es un estimador sesgado de σ^2. Por lo tanto,

$$\hat{\sigma}^2 = \frac{n}{n-p}\tilde{\sigma}^2 = \frac{1}{n-p}\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}$$

es un estimador insesgado de σ^2.

Nota 3.2 *$\hat{\sigma}^2$ tiene distintas formas de calcularse*

1. $\hat{\sigma}^2 = \frac{1}{n-p}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$ *donde* $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}^{-}\boldsymbol{y}$
2. $\hat{\sigma}^2 = \frac{1}{n-p}\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}$ *donde* $\boldsymbol{M} = \boldsymbol{I}-\boldsymbol{K} = \boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$
3. $\hat{\sigma}^2 = \frac{1}{n-p}\boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$

3.1.3.5 $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ son independientes Para demostrar la independencia se hace uso del siguiente teorema

Teorema 3.12 *Sea $\boldsymbol{y}$ un vector aleatorio $n \times 1$ que se distribuye $N(\mu, \Sigma)$, donde Σ tiene rango n. Si $\boldsymbol{B}\Sigma\boldsymbol{A} = 0$. La forma cuadrática $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}$ es independiente de la forma lineal $\boldsymbol{B}\boldsymbol{y}$, donde $\boldsymbol{B}$ es una matriz $q \times n$.*

Como

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}^{-}\boldsymbol{y} \\ \hat{\sigma}^2 &= \frac{1}{n-p}\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}\end{aligned}$$

Entonces haciendo $\boldsymbol{B} = \boldsymbol{X}^{-}$, $\Sigma = \sigma^2\mathbf{I}$ y $\boldsymbol{A} = \boldsymbol{M}$ entonces,

$$\begin{aligned}\boldsymbol{B}\Sigma\boldsymbol{A} &= \boldsymbol{X}^{-}\sigma^2\mathbf{I}\boldsymbol{M} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\mathbf{I}-\boldsymbol{K}) \\ &= \sigma^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right] \\ &= \sigma^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right] \\ &= \sigma^2(0) = 0\end{aligned}$$

Por lo tanto, $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ son independientes.

3.1.3.6 $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ son estimadores insesgados de mínima varianza Se puede probar que $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ son estimadores insesgados de mínima varianza pero se escapa del objetivo de este curso la demostración de este resultado.

Las propiedades de $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ se resumen en el siguiente teorema

Teorema 3.13 *Sea $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ donde $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Se cumplen las siguientes propiedades*

1. *$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}^{-}\boldsymbol{y}$ es el estimador de máxima verosimilitud de $\boldsymbol{\beta}$.*
2. *$\hat{\sigma}^2 = \frac{1}{n-p}\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y}$ es el estimador de máxima verosimilitud de σ^2.*
3. *$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2)$*
4. *$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$*
5. *$\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ son independientes.*
6. *$\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ son estadísticos suficientes.*
7. *$\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ son estadísticos completos.*

La prueba de este teorema se desarrollo durante toda la sección.

A continuación se muestra un resultado muy importante sobre las propiedades de los estimadores, previo es necesario enunciar el siguiente teorema

Teorema 3.14 *Sea $Y_1, Y_2, ..., Y_n$ una muestra observable con fdp $f_Y(\boldsymbol{y} : \boldsymbol{\theta})$ perteneciente a una familia exponencial, si las condiciones a, b, c y d del teorema 3.8 se satisfacen, entonces los K estadísticos $s_1\boldsymbol{y}, s_2\boldsymbol{y}, ..., s_K\boldsymbol{y}$ son estadísticos suficientes completos.*

Teorema 3.15 *Sea $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, donde $\boldsymbol{\varepsilon}$ se distribuye $N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, dado en la definición* **??**. *Sea $t(\boldsymbol{\beta}, \sigma^2)$ cualquier función de los parámetros $\boldsymbol{\beta}$ y σ^2 para los cuales un estimador*

insesgado existe. Entonces existe una función de los estadísticos suficientes $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$, digamos $q(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$, que es también un estimador insesgado de $t(\boldsymbol{\beta}, \sigma^2)$. En adición, $q(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ es el estimador insesgado uniformemente de varianza mínima para $t(\boldsymbol{\beta}, \sigma^2)$.

Nota: Por la condición (6) del teorema 3.13 se sabe que no se necesita el vector $\boldsymbol{y}$ ni la matriz $\boldsymbol{X}$ excepto para obtener los estadísticos suficientes $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$. Ya que los $p+1$ variables aleatorias $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p, \hat{\sigma}^2$ son estadísticos suficientes, ellos contienen toda la "información" acerca del modelo (para el caso 1) que esta contenida en los $n+np$ elementos de $\boldsymbol{y}$ y $\boldsymbol{X}$. Ya que se conoce la distribución conjunta de las $p+1$ variables aleatorias $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p, \hat{\sigma}^2$, se puede usar esa distribución para la estimación puntual, estimación por intervalo y prueba de hipótesis, sin sacrificar "información" por no usar las observaciones originales $\boldsymbol{y}$ y $\boldsymbol{X}$.

3.1.4 Propiedades de los estimadores para muestras grandes

Teorema 3.16 *Considere una sucesión de modelos lineal general*

$$\boldsymbol{y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n \qquad \boldsymbol{\varepsilon}_n \sim N(\boldsymbol{0}_n, \sigma^2\boldsymbol{I}_n) \qquad n = p+1, p+2, ...$$

donde $\boldsymbol{y}_n$ es un vector $n \times 1$ de variables aleatorias observables, $\boldsymbol{X}_n$ es una matriz de rango p (para cada n) de variables no aleatorias observables, $\boldsymbol{\beta}$ es un vector $p/times1$ de constantes no observables, y $\boldsymbol{\varepsilon}_n$ es un vector $n \times 1$ de variables aleatorias no observables. Sean $\hat{\boldsymbol{\beta}}_n$ y $\hat{\sigma}_n^2$, respectivamente, los estimadores de máxima verisimilitud ($\hat{\sigma}_n^2$ ajustado por sesgo) de bee y σ^2 en el n-esimo modelo. Por lo tanto,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_n &= (\boldsymbol{X}'_n\boldsymbol{X}_n)^{-1}\boldsymbol{X}'_n\boldsymbol{y}_n \\ \hat{\sigma}^2 &= (n-p)^{-1}\boldsymbol{Y}'_n[\boldsymbol{I}_n - \boldsymbol{X}_n(\boldsymbol{X}'_n\boldsymbol{X}_n)^{-1}\boldsymbol{X}'_n]\boldsymbol{y}_n \qquad n = p+1, p+2, ... \end{aligned}$$

1. *Si $\lim_{n\to\infty}(\boldsymbol{X}'_n\boldsymbol{X}_n)^{-1} = 0$, entonces la sucesión de estimadores $\{\boldsymbol{\ell}'\hat{\boldsymbol{\beta}}_n\}$ es un estimador consistente en media cuadrática (y simple) de $\{\boldsymbol{\ell}'\boldsymbol{\beta}_n\}$ (donde $\boldsymbol{\ell}$ es un vector $p \times 1$ de constantes).*
2. *La sucesión de estimadores $\{\hat{\sigma}_n^2\}$ es un estimador consistente en media cuadrática (y simple) de σ^2.*

Prueba 3.1 *Hay que demostrar que $\hat{\sigma}_n^2$ y $\boldsymbol{\ell}'\hat{\boldsymbol{\beta}}_n$ son insesgados en el límite y que las varianzas de $\hat{\sigma}_n^2$ y $'\hat{\boldsymbol{\beta}}_n$ son cero en el límite. Se dejan los detalles como ejercicio.*

Corolario 3.3 *Bajo las condiciones del teorema 3.16, cada elemento de $\hat{\boldsymbol{\beta}}_n$, es un estimador consistente en media cuadrática (y simple) del correspondiente elemento en $\boldsymbol{\beta}$.*

3.1.5 Estimación Puntual de funciones lineales de los parámetros

Para un vector $\boldsymbol{x}$ en el dominio $\mathbb{D}$, se sigue que

$$\mu(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x} = \sum_{i=1}^{p}\beta_i x_i$$

y esta es una combinación lineal dad de los β_i. Por lo tanto por las partes 6 y 7 del teorema 3.13, y más específicamente por el teorema 3.15, se puede hacer $t(\boldsymbol{\beta}, \sigma^2)$ igual a

$\boldsymbol{\beta}'\boldsymbol{x}$, y se sigue que el estimador insesgado de varianza uniformemente mínima de $\mu(\boldsymbol{x})$ es $\hat{\mu}(\boldsymbol{x}) = \hat{\boldsymbol{\beta}}'\boldsymbol{x}$ para cualquier vector $\boldsymbol{x}$ en el dominio $\mathbb{D}$, donde $\hat{\boldsymbol{\beta}}$ está dado como en la parte 1 del teorema3.13.

Para cualquier vector de constantes $\boldsymbol{\ell}$ $p \times 1$, el EIVUM de $\boldsymbol{\ell}'\boldsymbol{\beta}$, cualquier combinación lineal de $\boldsymbol{\beta}$, está dado por $\boldsymbol{\ell}'\hat{\boldsymbol{\beta}}$. Esto es,

$$\widehat{\boldsymbol{\ell}'\boldsymbol{\beta}} = \boldsymbol{\ell}'\hat{\boldsymbol{\beta}}$$

Por supuesto, $\mu(\boldsymbol{x})$ es un caso especial de $\boldsymbol{\ell}'\boldsymbol{\beta}$, como se puntualizará en el siguiente ejemplo

EJEMPLO 3.5

Considere el ejemplo **??**, donde $\mu(x) = \beta_0 + \beta_1 x$ para x en el intervalo $50 \leq x \leq 300$ millas. Si se recolectan los datos y se usan las formulas del ejemplo 3.1, se puede estimar β_0 y β_1. Suponga que se desea estimar la media de la variable aleatoria $Y_{(60)}$, es decir, $\mu(60)$. Ya que $\mu(60) = \beta_0 + \beta_1$, el EIUVM de $\mu(60)$ es

$$\hat{\mu}(60) = \hat{\beta}_0 + 60\hat{\beta}_1$$

Por otro lado, suponga que el investigador quiere estimar $30\beta_0 - 80\beta_1$, lo cual se denota por $\boldsymbol{\ell}\boldsymbol{\beta}$ donde $\boldsymbol{\ell}' = [30, -80]$. De acuerdo con el modelo de la situación del mundo real, $30\beta_0 - 80\beta_1$ no es la media $\mu(x)$ de cualquiera de la distribuciones guajiras. Sin embargo, se puede estimar $\boldsymbol{\ell}'\boldsymbol{\beta}$, y por el teorema 6.2.2 el EIVUM es $\boldsymbol{\ell}'\hat{\boldsymbol{\beta}} = 30\hat{\beta}_0 - 80\hat{\beta}_1$.

3.2 Valores Ajustados y Residuales

Una vez obtenido el estimador $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$. Si sustituimos esos valores en la ecuación del modelo

$$E[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}$$

obtenemos la ecuación de regresión estimada

$$\widehat{E[\boldsymbol{y}]} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{3.33}$$

que por costumbre usamos $\widehat{\boldsymbol{y}}$ en vez de $\widehat{E[\boldsymbol{y}]}$.

Cuando sustituimos los valores observados en la ecuación 3.33, obtenemos el **vector de valores ajustados**, representado por $\widehat{\boldsymbol{y}}$. Otra manera de representar el vector de valores ajustados es

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \tag{3.34}$$

donde

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \tag{3.35}$$

La diferencia entre el vector de valores observados $\boldsymbol{y}$ y el vector de valores ajustados $\widehat{\boldsymbol{y}}$ es conocido como el **vector de residuales** o simplemente **residuales**, y se representa con la letra $\boldsymbol{e}$, es decir

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{3.36}$$

el cual se puede escribir como

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\mathbf{I} - \boldsymbol{H})\boldsymbol{y} \tag{3.37}$$

3.3 Evaluación del Modelo

La regresión lineal calcula una ecuación que minimiza la distancia entre la recta ajustada y todos los puntos de los datos. Técnicamente, el método de mínimos cuadrados minimiza la suma de cuadrados de los residuos. En general, un modelo se ajusta bien a los datos si las diferencias entre los valores observados y los valores predichos (ajustados) son pequeños e insesgados.

Antes de mirar las medidas estadísticas para evaluar la bondad del ajuste, se deben chequear los gráficos de los residuos. Los mismo pueden revelar patrones que indican resultados sesgados de manera más eficaz que los números. Cuando los gráfico de residuos pasan la prueba, se deben usar números y comprobar las estadísticas de bondad de ajuste.

3.3.1 Partición de las Sumas de Cuadrados

La variación de un conjunto de datos es convencionalmente medida en términos de las desviaciones de los y_i alrededor de la media $\overline{y}$. Estas desviaciones se muestran por la linea vertical en la figura 3.3.

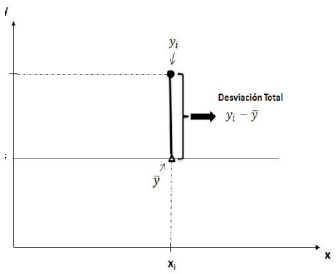

Figure 3.3 Desviación Total

La medida de la variación total, denotada por SCT, es la suma de las desviaciones al cuadrado, es decir

$$SCT = \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{3.38}$$

Aquí SCT denota la *suma de cuadrados total*. Si todas las observaciones y_i son las mismas, $SCT = 0$. Cuanto mayor es la variación entre las observaciones y_i, mayor es SCT. Por lo tanto, SCT es la variación de las observaciones y_i cuando no se toman en cuenta las variables explicatorias.

Cuando utilizamos la variable predictora x, la variación que refleja la incertidumbre con respecto a la variable y es la de las observaciones y_i alrededor de la regresión ajustada:

$$y_i - \hat{y}_i$$

Estas desviaciones se muestran en la línea vertical en la figura 3.4.

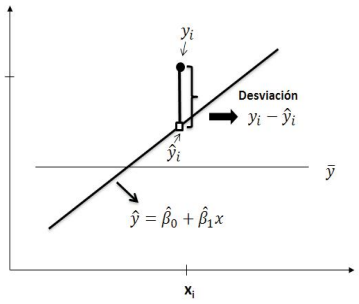

Figure 3.4 Desviación de las observaciones con respecto a la regresión

La medida de la variación en las observaciones y_i que está presente cuando la variable predictora x se tiene en cuenta es la suma de las desviaciones $y_i - \hat{y}_i$ al cuadrado, es decir

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.39}$$

Aquí SCE denota la *suma de cuadrado del error*. Si todas las observaciones y_i caen sobre la recta de regresión ajustada, $SCE = 0$. Cuanto mayor sea la variación de las observaciones y_i alrededor de la recta de regresión ajustada, mayor será la SCE.

La diferencia entre estas dos medida también es una suma de cuadrados, la cual está dada por

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3.40}$$

Note que la SCR es una suma de desviaciones al cuadrado, las desviaciones son

$$\hat{y}_i - \bar{y}$$

Estas desviaciones se muestran por la línea vertical em la figura 3.5. Cada desviación es simplemente la diferencia entre el valor ajustado de la recta de regresión y la media de las observaciones (la cual es la misma media de los valores ajustados). Si la recta de regresión es horizontal de manera que $\hat{y}_i - \bar{y} = 0$, entonces $SCR = 0$, de lo contrario, SCR será positiva.
La SCR puede considerarse como una medida de la parte de la variabilidad de los y_i la cual está asociada con la recta de regresión. Cuanto mayor sea SCR en relación a la SCT, mayor será el efecto de la relación de regresión para explicar la variación total de las observaciones y_i.

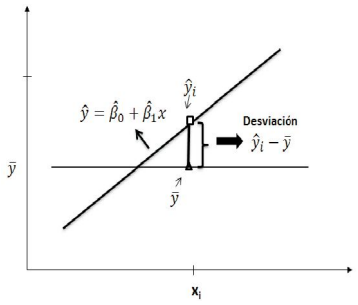

Figure 3.5 Desviación de la regresión con respecto a la media

3.3.1.1 Desarrollo Formal de la Partición La desviación total, usada en la medida de la variación total de las observaciones $y_i - \bar{y}$, usada en la medida de la variación total de las observaciones y_i sin tomar en cuenta la variable predictora, se puede descomponer en dos componentes:

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y}_i \tag{3.41}$$

Los dos componentes son:

1. La desviación del valor ajustado $\hat{y}_i$ alrededor de la media $\bar{y}$, $(\hat{y}_i - \bar{y})$
2. La desviación de la observación y_i alrededor de la recta de regresión ajustada, $(y_i - \hat{y}_i)$

Es una propiedad remarcable que la suma de cuadrados de estas desviaciones cuadrados tienen la misma relación:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.42}$$

o, usando la notación en (3.44), (3.39) y (3.40):

$$SCT = SCR + SCE \tag{3.43}$$

Para probar este resultado procedemos de la siguiente manera:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}[(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2 \\
&= \sum_{i=1}^{n}[(\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 + 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)] \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)
\end{aligned}$$

Se puede probar fácilmente que el tercer termino del lado derecho (la suma cruzadas es igual a cero).

3.3.1.2 Sumas de cuadrados en términos matriciales En la ecuación (3.44) tenemos una expresión de la SCT, pero la misma se puede escribir como

$$SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \tag{3.44}$$

Pero $\sum_{i=1}^{n} y_i^2$ se puede escribir en términos matricial como $\boldsymbol{y'y}$. Además

$$\frac{(\sum_{i=1}^{n} y_i)^2}{n} = \left(\frac{1}{n}\right) \boldsymbol{y'Jy}$$

donde $\boldsymbol{J}$ es una matriz de unos. Por lo tanto

$$SCT = \boldsymbol{y'y} - \left(\frac{1}{n}\right) \boldsymbol{y'Jy} = \boldsymbol{y'}\left[\mathbf{I} - \left(\frac{1}{n}\right) \boldsymbol{J}\right] \boldsymbol{y} \tag{3.45}$$

Así mismo, la SCE esta dada por la ecuación (3.39). Si denotamos $e_i = y_i - \hat{y}_i$ entonces $SCE = \sum_{i=1}^{n} e_i^2$, lo cual en terminos matriciales sería

$$SCE = \boldsymbol{e'e} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{3.46}$$

que también se puede escribir como

$$SCE = \boldsymbol{e'e} = \boldsymbol{y'y} - \hat{\boldsymbol{\beta}}' \boldsymbol{X'y} \tag{3.47}$$

Finalmente, la suma de cuadrados de regresión se puede escribir como

$$SCR = \hat{\boldsymbol{\beta}}' \boldsymbol{X'y} - \left(\frac{1}{n}\right) \boldsymbol{y'Jy} \tag{3.48}$$

3.4 Coeficiente de Determinación

El coeficiente de determinación, denotado por R^2 es una medida estadística de cuan cerca de los datos está la ecuación ajustada. La definición de R^2 es muy simple, este es el porcentaje de la variación de la variable respuesta que es explicada por el modelo lineal, es decir,

$$R^2 = \frac{\text{Variación Explicada}}{\text{Variación Total}}$$

De acuerdo con lo estudiado anteriormente la variación explicada está dada por la *suma de cuadrados de regresión*, $SCR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, y la variación total está dada por *la suma de cuadrados total*, $SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2$. Por lo tanto,

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.49}$$

Forma matricial: En la sección anterior se llegó a las siguientes expresiones de las sumas de cuadrados en forma matricial

- $SCT = \boldsymbol{y'}\left[\mathbf{I} - \left(\frac{1}{n}\right) \boldsymbol{J}\right] \boldsymbol{y}$
- $SCE = \boldsymbol{e'e} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$

- $SCR = \hat{\boldsymbol{\beta}}' \boldsymbol{X}'\boldsymbol{y} - \left(\frac{1}{n}\right) \boldsymbol{y}'\boldsymbol{J}\boldsymbol{y}$

Por lo tanto el coeficiente de terminación en forma matricial está dado por

$$R^2 = \frac{SCR}{SCT} = \frac{\hat{\boldsymbol{\beta}}' \boldsymbol{X}'\boldsymbol{y} - \left(\frac{1}{n}\right) \boldsymbol{y}'\boldsymbol{J}\boldsymbol{y}}{\boldsymbol{y}' \left[\mathbf{I} - \left(\frac{1}{n}\right) \boldsymbol{J}\right] \boldsymbol{y}} = \frac{\hat{\boldsymbol{\beta}}' \boldsymbol{X}'\boldsymbol{y} - n\bar{y}^2}{\boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2} \tag{3.50}$$

3.4.1 Propiedades del coeficiente de determinación

Algunas propiedades de R^2 se listan a continuación.

1. El rango de R^2 es $0 \leq R^2 \leq 1$. Si todos los $\hat{\beta}_j$ fuesen cero, excepto $\hat{\beta}_0$, R^2 debería ser cero. Si todos los valores de y caen sobre la superficie ajustada, es decir $y_i = \hat{y}_i, i = 1, 2, ..., n$, entonces r^2 debería ser 1.
2. Adicionar una variable x al modelo incrementa (nunca decrementa) el valor de R^2.
3. Si $\hat{\beta}_1 = \hat{\beta}_2 = \cdots = \hat{\beta}_k = 0$, entonces

$$E(R^2) = \frac{k}{n-1} \tag{3.51}$$

   Note que los $\hat{\beta}_j$ no serían 0 cuando los β_j son 0.
4. R^2 es invariante a transformaciones lineales de rango completo sobre los x y a un cambio de escala sobre y (pero no invariante a transformaciones lineales conjuntas de y y las x).

En las propiedades 2 y 3 vemos que si k es una fracción relativamente grande de n, es posible tener un valor grande de R^2 que no es significativo. En este caso, los x que no contribuyen a predecir y pueden aparecer para hacerlo en un ejemplo particular, y la ecuación de regresión estimada puede no ser un estimador útil del modelo poblacional. Para corregir esta tendencia, un R^2 ajustado, denotado por R_a^2 fue propuesto por Ezekiel (1930), el cual veremos más adelante.

3.4.2 Observaciones sobre el coeficiente de determinación

1. No permite determinar si los coeficientes estimados y las predicciones sos sesgadas.
2. No indica si un modelo de regresión es adecuado. Se puede tener un valor bajo de R^2 para un modelo bueno, o una valor alto de R^2 para un modelo que no se ajusta a los datos.

   (a) R^2 **bajos no necesariamente son malos.** En algunos campos, es completamente esperado que el valor de R^2 sea bajo. Por ejemplo, al intentar predecir el comportamiento humano, tal como en psicología, típicamente se tienen valores de R^2 por debajo del 50%. Los humanos simplemente son más difíciles de predecir que, digamos, los procesos físicos.

   Si el valor de R^2 es bajo pero se tienen predictores estadísticamente significativos, se pueden sacar conclusiones importantes sobre como cambios en los valores de los predictores están asociados con cambios en la variable respuesta. Independientemente del R^2, los coeficientes significativos aún representan el cambio medio

en la respuesta por el cambio en una unidad en el predictor mientras los otros predictores se mantienen constantes. Obviamente, este tipo de información puede ser extremadamente valiosa.

Por lo tanto, no debe sorprendernos tener resultados con R^2 bajos y $p - valor$ bajos.

(b) R^2 **altos no necesariamente son buenos.** Un R^2 alto no necesariamente indica que el modelo tiene un buen ajuste, esto puede ser una sorpresa. Pero existen situaciones en las que se puede observar gráficamente que la linea ajustada (modelo de regresión lineal simple) tiene un comportamiento alejado de los datos y aún así el R^2 puede ser alto. Esto es posible por muchas razones, una de ellas es que la mejor manera de modelar la variable respuesta es a través de regresión no lineal; otra es que en la construcción del modelo no se están usando los predictores importantes o que se han incluido muchos predictores. Existen otras razones que se dejan como tema de investigación a los estudiantes.

3. Una pregunta común al realizar un análisis de regresión es ¿Qué tan alto debe ser el R^2?. La respuesta a esta pregunta es... **depende.**. Depende del objetivo principal que se tenga al realizar el análisis. Dos objetivos frecuentes en el análisis de regresión son

   (a) Describir la relación entre los predictores y la variable respuesta, o

   (b) Predecir la variable respuesta

   Si el principal objetivo es determinar cuales predictores son estadísticamente significativos y como cambios en los predictores se relacionan a cambios en la variable respuesta (objetivo (a)), entonces el valor de R^2 es irrelevante. Por ejemplo, suponga que se modela la relación entre X y Y, se encuentra que el $p-$valor para X es significativo, además que su coeficiente es 2 y que los supuestos se cumplen, estos resultados indican que un incremento de una unidad en X esta asociado a un incremento promedio de 2 unidades en Y. Esta interpretación es correcta aún si el R^2 es 0.95 o 0.25. Por lo tanto preguntarse ¿cuán grande debe ser el R^2? en este contexto no tiene sentido porque este no es relevante. Un R^2 bajo ni niega un predictor significativo, ni cambia el sentido de su coeficiente. El R^2 es simplemente cualquier valor, y no necesita ser un valor particular para permitir una interpretación válida.

   Por el contrario, si el objetivo principal es producir predicciones precisas, el valor de R^2 es importante. Lasa predicciones no son tan simples como solo predecir un valor, ya que ellas incluyen un margen de error, predicciones más precisas tienen menos error. El R^2 entra en escena porque un R^2 bajo indica que el modelo tiene mucho error. Por lo tanto, un R^2 puede advertir de predicciones imprecisas. Sin embargo, no se puede usar el R^2 para determinar si las predicciones son lo suficientemente precisas para las necesidades.

3.4.3 Coeficiente de determinación ajustado

Hemos visto que una de los propiedades que tiene el R^2 es que este aumenta (o por lo menos no disminuye) a medida que se van añadiendo variables al modelo. En consecuencia un modelo con más coeficientes puede parecer tener un mejor ajuste simplemente porque tiene más terminos, situación que no necesariamente es cierta.

El coeficiente de determinación ajustado es una versión modificada del R^2 que se ajusta por el número de predictores en el modelo. El R^2 ajustado se incremente solo si el nuevo termino mejora al modelo más lo que se esperaba por azar . Este disminuye cuando cuando un predictor mejora al modelo por menos de los esperado al azar. Siempre es mas bajo que el R^2.

Para obtener R_a^2, primero restamos $k/(n-1)$ en 3.51 desde R^2 para corregir por el sesgo cuando $\beta_1 = \beta_2 = ... = \beta_k = 0$. Esta corrección, sin embargo, debería hacer R_a^2 más pequeña cuando los $\hat{\beta}$ son grandes, por lo que una modificación adicional se hace para que $R_a^2 = 1$ cuando $R^2 = 1$. Por lo tanto R_a^2 es definido como

$$R_a^2 = \frac{(R^2 - \frac{k}{n-1})(n-1)}{n-k-1} = \frac{(n-1)R^2 - k}{n-k-1} \tag{3.52}$$