Econometría I

regresión simple

Leonardo Manríquez M.
(lmanriquez@ucsc.cl)

August 20, 2023

*Las gráficas y cálculos[1] de esta presentación están elaborados en base a dataset_class1.csv

[1] Atención: en esta presentación no se consideran todas las cifras decimales. Por lo tanto, pueden existir diferencias entre el cálculo expuesto en esta presentación y el cálculo considerando todas las cifras decimales.

Recitación

- Introducción
- Método de Mínimos Cuadrados Ordinarios, MCO/OLS
- Descomposición de varianza
- Propiedades de los estimadores de MCO/OLS
- Inferencia estadística en el modelo de RLS
- Predicción en el modelo de regresión lineal simple
- Formas funcionales alternativas

Introducción

- El modelo de **regresión lineal simple** trata de modelar la relación lineal únicamente entre dos variables.

$$y = f(x), \tag{1}$$

donde $f(...)$ es una función de x.

Introducción

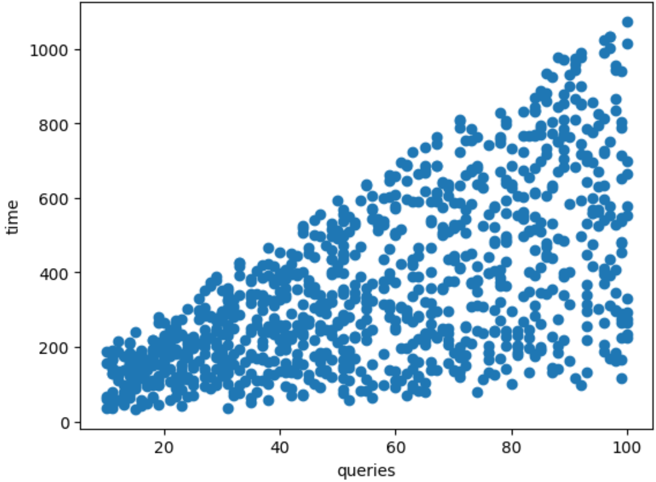

Figure: Relación tiempo de respuesta y y *queries* x

Introducción

- En este punto podemos distiguir dos tipos de relaciones entre y e x:
    1. **Determinista**
    2. **Estadística**

Introducción

- Supongamos que la relación entre tiempo de respuesta de un sistema y y el número de consultas x es:

$$y = 100 + 2 \cdot x - 3 \cdot x^2. \tag{2}$$

- Esta es una relación **determinista** entre y e x porque para cada valor de número de *queries* x conocemos con certeza el valor de tiempo de respuesta y.
- Por otra parte, supongamos que ahora la relación entre tiempo de respuesta de un sistema y y el número de *queries* x es:

$$y = 100 + 2 \cdot x - 3 \cdot x^2 + \mu, \tag{3}$$

  donde μ puede tomar los valores $\mu = 10$ con probabilidad de $\frac{1}{2}$ y $\mu = -10$ con probabilidad de $\frac{1}{2}$.
- Ahora los diferentes valores de y para diferentes valores de x no pueden determinarse exactamente, pero son descritos en términos **probabilísticos**

Introducción

- Consideraremos que $f(...)$ es lineal en x, esto es:

$$f(x) = \beta_0 + \beta_1 \cdot x. \tag{4}$$

- Y asumiendo una relación **estocástica** tenemos que:

$$y = f(x) + \mu = \underbrace{\beta_0 + \beta_1 \cdot x}_{\text{Determinista}} + \underbrace{\mu}_{\text{Estocástica}}. \tag{5}$$

- El término μ es llamada perturbación o error (con alguna distribución de probabilidad conocida)[2]
- β_0, β_1 son llamados coeficientes de la regresión o parámetros de la regresión, y son estimados a través de la muestra de y e x.

[2] Es decir, es una variable aleatoria.

Introducción

- ¿Por qué permitir el término del error μ?
  1. Elementos impredecibles;
  2. Variables omitidas; y
  3. Errores de medida en y.
- Luego, si tenemos n observaciones para y e x:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \mu_i, \ \forall i = 1, .., n. \tag{6}$$

- Por lo tanto, el objetivo es estimar los parámetros desconocidos (β_0, β_1) dados los n valores de y e x (la muestra).
- Para lograr esto debemos considerar algunos supuestos respecto al término del error.

Introducción

- A través de los supuestos hacemos que:
  1. Valor medio cero, $\mathbb{E}[\mu] = 0$;
  2. Varianza común, $\mathbb{V}[\mu] = \sigma_\mu^2$;
  3. Independencia entre los términos del error μ_i y μ_j, $\forall\ i \neq j$;
  4. Independencia entre μ_i y x_j, $\forall\ i, j$; y
  5. Normalidad, μ_i se distribuye normal $\forall i$.

  Finalmente, si consideramos los supuestos 1, 2, 3 y 5 podemos establcer lo siguiente para el término del error:

$$\mu_i \sim IN(0, \sigma_\mu^2). \tag{7}$$

Introducción

- Utilizando el supuesto 1, tenemos que:

$$\mathbb{E}[y_i] = \beta_0 + \beta_1 \cdot x_i. \tag{8}$$

- Esta expresión se conoce como la **función de regresión poblacional** , donde si utilizamos los parámetros estimados tendremos la **función de regresión muestral**

Mínimos Cuadrados Ordinarios: MCO/OLS

- Si se define la función $RSS(...)$ como la **Suma de Cuadrados Residuales** tenemos que:

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i}^{n} \hat{\mu}_i^2 = (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i)^2. \tag{9}$$

- A través del método de MCO/OLS se encuentran los parámetros $\hat{\beta}_0, \hat{\beta}_1$ tal que:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} RSS(\hat{\beta}_0, \hat{\beta}_1). \tag{10}$$

Mínimos Cuadrados Ordinarios: MCO/OLS

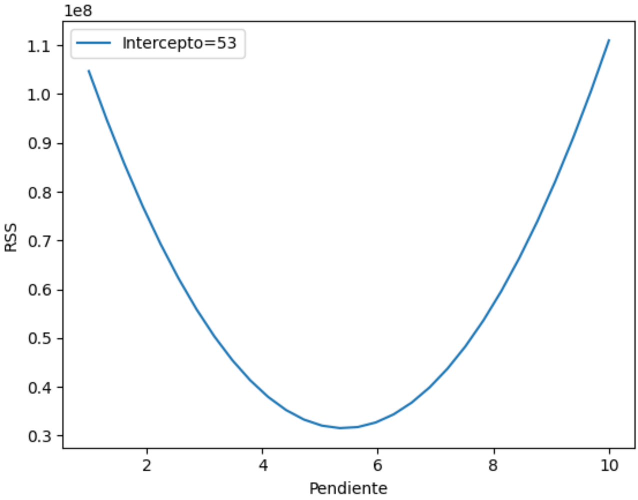

Figure: RSS con intercepto fijo en 53.

Mínimos Cuadrados Ordinarios: MCO/OLS

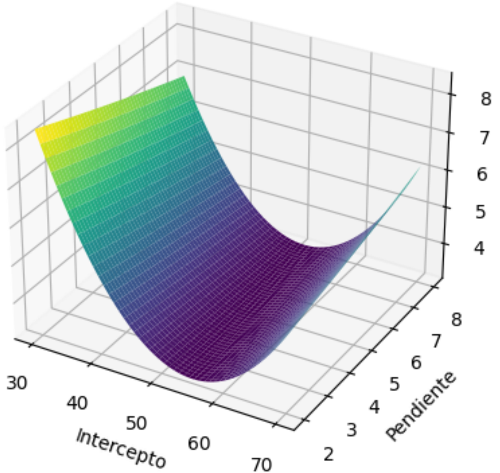

Figure: RSS para diferentes valores de intercepto y pendiente

Mínimos Cuadrados Ordinarios: MCO/OLS

- Entonces, para encontrar los valores $\hat{\beta}_0, \hat{\beta}_1$ que minimizan la función $RSS(...)$ debemos resolver el siguiente problema:

$$\begin{aligned}
\frac{\partial RSS}{\partial \hat{\beta}_0} &= 0 \rightarrow \sum_i^n 2 \cdot (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i)(-1) = 0 \\
\frac{\partial RSS}{\partial \hat{\beta}_1} &= 0 \rightarrow \sum_i^n 2 \cdot (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i)(-x_i) = 0
\end{aligned} \tag{11}$$

- Desde $\frac{\partial RSS}{\partial \hat{\beta}_0} = 0$ podemos expresar $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$.
- Desde $\frac{\partial RSS}{\partial \hat{\beta}_1} = 0$ podemos expresar $\sum_i^n y_i x_i = \hat{\beta}_0 \cdot \sum_i^n x_i + \hat{\beta}_1 \cdot \sum_i^n x_i^2$.

Mínimos Cuadrados Ordinarios: MCO/OLS

- Al reemplazar $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$ en $\sum_i^n y_i x_i = \hat{\beta}_0 \cdot \sum_i^n x_i + \hat{\beta}_1 \cdot \sum_i^n x_i^2$ tenemos que:

$$
\begin{aligned}
\sum_i^n y_i x_i &= (\bar{y} - \hat{\beta}_1 \cdot \bar{x}) \cdot \sum_i^n x_i + \hat{\beta}_1 \cdot \sum_i^n x_i^2 \\
\sum_i^n y_i x_i &= (\bar{y} - \hat{\beta}_1 \cdot \bar{x}) \cdot n \cdot \bar{x} + \hat{\beta}_1 \cdot \sum_i^n x_i^2
\end{aligned}
\tag{12}
$$

- Desde donde podemos encontrar que:

$$
\hat{\beta}_1 = \frac{\sum_i^n y_i \cdot x_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_i^n x_i - n \cdot \bar{x}^2} = \frac{\mathsf{cov}(x, y)}{\mathbb{V}(x)}. \tag{13}
$$

Mínimos Cuadrados Ordinarios: MCO/OLS

- Desde el resultado anterior tenemos que los parámetros $\hat{\beta}_0, \hat{\beta}_1$ que minimizan la función $RSS(...)$ corresponde a:

$$\hat{\beta}_1 = \frac{\texttt{cov}(y, x)}{\mathbb{V}(x)} \quad \text{y} \quad \hat{\beta}_0 = \bar{y} - \bar{x} \cdot \hat{\beta}_1. \tag{14}$$

- Recordando la **función de regresión poblacional** tenemos que $y_i = \beta_0 + \beta_1 \cdot x_i + \mu_i$, tenemos que:
- $\hat{\beta}_0$ es el intercepto estimado.
- $\hat{\beta}_1$ es la pendiente estimada de la recta.

Mínimos Cuadrados Ordinarios: MCO/OLS

- Recordemos que en nuestro ejemplo: y es el tiempo de respuesta de un sistema y x el número de *queries*.
- Para nuestro caso tenemos que $\texttt{cov}(y, x) = 3685.93$, $\mathbb{V}(x) = 682.93$, $\bar{x} = 55.57$ y $\bar{y} = 353.69$.
- Por lo tanto, $\hat{\beta}_1 = \frac{3685.93}{682.93} = 5.39$ y $\hat{\beta}_0 = 353.69 - 5.39 \cdot 55.57 = 54.16$.
- Ahora la **función de regresión muestral** estimada viene dada por:
$$\hat{y}_i = 54.16 + 5.39 \cdot x_i. \tag{15}$$
- **Interpretación intercepto**: cuando el número de *queries* es igual a 0 entonces el tiempo de espera será de 54.16 segundos.
- **Interpretación pendiente**: cuando las *queries* aumenta en 1 unidad entonces el tiempo de espera aumenta en 5.59 segundos.

Mínimos Cuadrados Ordinarios: MCO/OLS

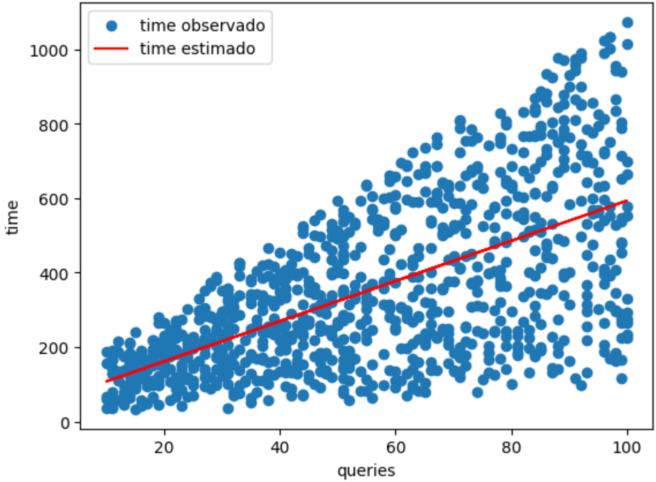

Figure: Observados v.s. recta de MCO/OLS

Descomposición de varianza

- La descomposición de varianza esta basada en el análisis de varianzas (ANOVA).
- Sabemos que para cada observación $y_i = \hat{y}_i + \hat{\mu}_i$.
- Y si tomamos varianza tenemos que:

$$\begin{aligned} \mathbb{V}(y_i) &= \mathbb{V}(\hat{y} + \hat{\mu}) \\ \mathbb{V}(y_i) &= \mathbb{V}(\hat{y}) + \mathbb{V}(\hat{\mu}) + 2 \cdot \texttt{cov}(\hat{y}, \hat{\mu}) \end{aligned} \tag{16}$$

- Sin embargo, por los supuestos tenemos que $\texttt{cov}(\hat{y}, \hat{\mu}) = 0$ y si reexpresamos de manera explícita:

$$\frac{1}{n} \cdot \sum_{i}^{n} (y_i - \bar{y})^2 = \frac{1}{n} \cdot \sum_{i}^{n} (\hat{y}_i - \bar{\hat{y}})^2 + \frac{1}{n} \cdot \sum_{i}^{n} (\hat{\mu}_i - \bar{\hat{\mu}})^2 \tag{17}$$

Descomposición de varianza

- Y por los resultados de MCO/OLS podemos expresar que $\bar{\hat{y}} = \bar{y}$ y $\bar{\hat{\mu}} = 0$. De modo que:

$$\underbrace{\frac{1}{n} \cdot \sum_{i}^{n} (y_i - \bar{y})^2}_{TSS} = \underbrace{\frac{1}{n} \cdot \sum_{i}^{n} (\hat{y}_i - \bar{y})^2}_{ESS} + \underbrace{\frac{1}{n} \cdot \sum_{i}^{n} (\hat{\mu}_i)^2}_{RSS} \qquad (18)$$

- Donde TSS es la **Suma Cuadrados Totales**, ESS es la **Suma de Cuadrados Estimados** y RSS es la **Suma de Cuadrados Residuales**

Descomposición de varianza

- Siguiendo la relación anterior podemos definir que:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}. \tag{19}$$

- Entonces, R^2 corresponde a la proporción de la varianza muestral de y que esta siendo explicada por la regresión MCO/OLS.

Descomposición de varianza

- Para nuestro caso tenemos que $RSS = 31538.61$ y $TSS = 51392.54$. Por lo tanto:

$$R^2 = 1 - \frac{31538.61}{51392.54} = 0.3863. \tag{20}$$

- Esto implica que un 38.63% de la variabilidad del tiempo de respuesta (y) esta siendo explicado a través de la variación de las *queries* (x).

Propiedades de los estimadores de MCO/OLS

- El **Teorema de Gauss-Márkov** indica que si se cumplen los supuestos que hemos establecido anteriormente para el modelo de regresión lineal, entonces los estimadores de MCO/OLS serán los **mejores** estimadores **lineales** e **insesgados** (MELI/BLUE)

Propiedades de los estimadores de MCO/OLS

- Consideremos el modelo de regresión:

$$y_i = \beta \cdot x_i + \mu_i, \ \forall 1, 2, ..., n. \tag{21}$$

- Por simplicidad omitiremos el término del intercepto. Asumimos que μ_i es distribuido independientemente con media 0 y varianzas σ_μ^2.
- Como x_i son términos constantes, $\mathbb{E}(y_i) = \beta \cdot x_i$ y $\mathbb{V}(y) = \sigma^2$. El estimador de MCO/OLS:
-
$$\hat{\beta} = \frac{\sum_i^n x_i \cdot y_i}{\sum_i^n x_i^2} = \sum_i^n c_i \cdot y_i, \tag{22}$$

- Donde $c_i = \frac{x_i}{\sum_i^n x_i^2}$. De este modo, $\hat{\beta}$ es una función **lineal** de las observaciones muestrales de y_i y por lo tanto es llamado **estimador lineal**

Propiedades de los estimadores de MCO/OLS

- Ahora, si aplicamos esperanza al estimador de MCO/OLS, tenemos que:

$$
\begin{aligned}
\mathbb{E}(\hat{\beta}) =& \mathbb{E}(\sum_{i}^{n} c_i \cdot y_i), \\
\mathbb{E}(\hat{\beta}) =& \sum_{i}^{n} c_i \cdot \mathbb{E}(y_i), \\
\mathbb{E}(\hat{\beta}) =& \sum_{i}^{n} \left(\frac{x_i}{\sum_i^n x_i^2} \right) \cdot (\beta \cdot x_i), \\
\mathbb{E}(\hat{\beta}) =& \beta \frac{\sum_i^n x_i^2}{\sum_i^n x_i^2}, \\
\mathbb{E}(\hat{\beta}) =& \beta.
\end{aligned}
\tag{23}
$$

- Por lo tanto, $\hat{\beta}$ es un **estimador lineal insesgado**.

Propiedades de los estimadores de MCO/OLS

- Finalmente, para demostrar que este estimador es el mejor debemos demostrar que tiene **mínima varianza** entre la clase de estimadores lineales insesgados.
- Consideremos el siguiente estimador:

$$\tilde{\beta} = \sum_{i}^{n} d_i y_i. \tag{24}$$

- Luego, si este estimador es insesgado tenemos que:

$$\mathbb{E}(\tilde{\beta}) = \beta. \tag{25}$$

- Tal que necesitamos tener que $\sum_i^n d_i \cdot x_i = 1$. Luego, desde que y_i es independiente con una varianza común σ^2, tenemos que:

$$\mathbb{V}(\tilde{\beta}) = \sum_{i}^{n} d_i^2 \cdot \sigma^2. \tag{26}$$

Propiedades de los estimadores de MCO/OLS

- Es necesario encontrar d_i de modo que esta varianza, sujeta a la condición $\sum_i^n d_i \cdot x_i = 1$, sea **mínima**.
- Por lo tanto minimizamos:

$$\sum_i^n d_i^2 - \lambda \cdot (\sum_i^n d_i \cdot x_i - 1), \tag{27}$$

- Donde λ es el multiplicador Lagrangiano. Ahora, si diferenciamos esta última expresión respecto a d e igualamos a 0, tenemos que:

$$\frac{\partial(...)}{\partial d} = 2 \cdot d_i - \lambda \cdot x_i = 0 \rightarrow d_i = \frac{\lambda}{2} \cdot x_i. \tag{28}$$

- Ahora, si multiplicamos ambos lados de la igualdad por x_i y consideramos $\sum_i^n$ tenemos que:

$$\sum_i^n d_i \cdot x_i = \frac{\lambda}{2} \cdot \sum_i^n x_i^2. \tag{29}$$

Propiedades de los estimadores de MCO/OLS

- Pero $\sum_i^n d_i \cdot x_i = 1$. Por lo tanto,

$$\lambda = \frac{2}{\sum_i^n x_i^2}. \tag{30}$$

- Así obtenemos que:

$$d_i = \frac{\lambda}{2} \cdot x_i = \frac{x_i}{\sum_i^n x_i^2}. \tag{31}$$

- Notar que la última expresión de d_i corresponde al coeficiente c_i de MCO/OLS.
- De este modo el estimador de MCO/OLS tiene **mínima varianza** en la clase de estimador lineal insesgado.

Propiedades de los estimadores de MCO/OLS

- La **mínima varianza** es:

$$\begin{aligned}
\mathbb{V}(\hat{\beta}) =& \sum_{i}^{n} c_i^2 \cdot \sigma^2, \\
\mathbb{V}(\hat{\beta}) =& \sum_{i}^{n} \left(\frac{x_i}{\sum_i^n x_i^2} \right)^2 \cdot \sigma^2, \\
\mathbb{V}(\hat{\beta}) =& \frac{\sigma^2}{\sum_i^n x_i^2}.
\end{aligned} \tag{32}$$

Inferencia estadística en el modelo de RLS

- Notar que hasta el momento, para encontrar los estimadores pode MCO/OLS no hemos necesitado asumir alguna distribución de probabilidad particular para el término del error μ_i.
- Sin embargo, para estimar intervalos de confianza para los parámetros y aplicar contrastes de hipótesis necesitamos asumir que el término del error μ_i sigue una distribución de probabilidad normal.
- De los resultados anteriores obtuvimos que los estimadores de MCO/OLS son **insesgados** y tienen **mínima varianza** entre la clase de estimadores lineales insesgados.

Inferencia estadística en el modelo de RLS

- Estas dos propiedades pueden permanecer independiente que el término del error μ siga una distribución normal siempre que las otras suposiciones que hemos hecho se cumplan. Estas suposiones son.
  1. $\mathbb{E}(\mu_i) = 0$,
  2. $\mathbb{V}(\mu_i) = \sigma^2$,
  3. u_i y u_j son independientes $\forall\ i \neq j$; y
  4. x_i no es estocástica.
- Ahora, asumiremos adicionalmente que los términos del error están distribuidos normalmente y encontraremos los intervalos de confianza para β_0 y β_1 y el test de hipótesis para β_0 y β_1.

Inferencia estadística en el modelo de RLS

- Primero, necesitamos la distribución muestral de $\hat{\beta}_0$ y $\hat{\beta}_1$. Se puede probar que tienen una distribución normal (material anexo), para lo que tenemos los siguientes resultados:
- $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen distribución normal conjunta con:

$$\begin{aligned} \mathbb{E}(\hat{\beta}_0) = \beta_0, \quad \mathbb{V}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathbb{V}(x)}\right), \\ \mathbb{E}(\hat{\beta}_1) = \beta_1, \quad \mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{\mathbb{V}(x)} \quad \text{y} \\ \mathtt{cov}(\hat{\beta}_0, \hat{\beta}_1) = \sigma^2\left(\frac{-\bar{x}}{\mathbb{V}(x)}\right). \end{aligned} \tag{33}$$

- Estos resultados serían de gran utilidad si se conociera la varianza del error σ^2.

Inferencia estadística en el modelo de RLS

- Lamentablemente, σ^2 es desconocido y se requiere estimarlo.
- Si RSS es la Suma de Cuadrados Residual, entonces:

$$\hat{\sigma^2} = \frac{RSS}{n-k}, \text{ es un estimador insesgado de } \sigma^2. \qquad (34)$$

- k es el número de coeficientes. En el modelo de regresión lineal simple el número de coeficientes es $k = 2$.
- También sabemos que $\hat{\sigma^2} = \frac{RSS}{n-k}$ tiene una distribución $\tilde{\chi}^2$ con $(n-k)$ grados de libertad (g.l).
- La distribución de RSS es independiente de la distribución de $\hat{\beta}_0$ y $\hat{\beta}_1$ (material anexo)
- Este resultado puede ser utilizado para obtener los intervalos de confianza de σ^2 o para una prueba de hipótesis respecto de σ^2.

Inferencia estadística en el modelo de RLS

- Sin embargo, por el momento, mantendremos el problema a realizar inferencia sobre $\hat{\beta}_0$ y $\hat{\beta}_1$. Para este propósito utilizaremos la distribución t.
- Recordemos que para dos variables aleatorias $z \sim N(0,1)$ y $q \sim \tilde{\chi}^2$ con j grados de libertad y si z y q son independientes, tenemos que:

$$x = \frac{z}{\frac{q}{j}} \sim \; t \text{ con } k \text{ g.l.} \tag{35}$$

Inferencia estadística en el modelo de RLS

- En nuestro caso tenemos que:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\mathbb{V}(x)}}} \sim N(0,1). \tag{36}$$

- Notar que a $\hat{\beta}_1$ se le resta la media y divide por la desviación estándar.
- También, sabemos que $\frac{RSS}{\sigma^2} \sim \tilde{\chi}^2_{n-k}$ y que ambas distribuciones son independientes. Por lo tanto:

$$\frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\mathbb{V}(x)}}}}{\sqrt{\frac{\frac{RSS}{\sigma^2}}{n-k}}} \sim t \text{ con } n-k \text{ g.l.} \tag{37}$$

Inferencia estadística en el modelo de RLS

- Lo cual se puede simplificar a:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma^2}}{\mathbb{V}(x)}}} \sim t \text{ con } n-k \text{ g.l.} \tag{38}$$

- Notar que el estimador de la varianza de $\hat{\beta}_1$ es $\frac{\sigma^2}{\mathbb{V}(x)}$ y su raíz cuadrada es llamada como **error estándar**, $se(\hat{\beta}_1)$.
- Sin pérdida de generalidad, podemos seguir un procedimiento similar para $\hat{\beta}_0$.
- $\hat{\sigma}$ es llamado usualmente como **error estándar de la regresión**, SER.

Inferencia estadística en el modelo de RLS

- Para nuestro ejemplo: y tiempo de respuesta y x número de *queries*. Tenemos la siguiente información:
- $n = 999$, $\bar{x} = 55.57$, $\mathbb{V}(y) = 51392.54$, $\mathbb{V}(x) = 682.93$ y `cov(x,y)` $= 3685.93$. Por lo tanto:

$$\begin{aligned} \mathbb{V}(\hat{\beta}_0) &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathbb{V}(x)}\right) = \sigma^2 \cdot 4.52, \\ \mathbb{V}(\hat{\beta}_1) &= \frac{\sigma^2}{\mathbb{V}(x)} = \frac{\sigma^2}{682.93}, \\ \hat{\sigma^2} &= \frac{1}{n-k} \cdot \left(\mathbb{V}(y) - \frac{\text{cov}(x,y)^2}{\mathbb{V}(x)}\right) = 31.59. \end{aligned} \tag{39}$$

Por lo tanto: $se(\hat{\beta}_0) = 31.59 \cdot 4.52 = 11.91$ y $se(\hat{\beta}_1) = \sqrt{\frac{31.59}{682.93}} = 0.21$

Inferencia estadística en el modelo de RLS: intervalos de confianza

- Desde que $\frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)}$ y $\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}$ tienen distribución t con $n-k$ grados de libertar, obtenemos que:

$$\begin{aligned} \text{Prob}\left[-t_{n-k} < \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} < t_{n-k} \right] = 0.95, \\ \text{Prob}\left[-t_{n-k} < \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} < t_{n-k} \right] = 0.95. \end{aligned} \tag{40}$$

Inferencia estadística en el modelo de RLS: intervalos de confianza

- De los resultados anteriores podemos definir los intervalos de confianza para $\hat{\beta}_0$ y $\hat{\beta}_1$:

$$
\begin{aligned}
IC_{\hat{\beta}_0} &= \hat{\beta}_0 \pm t_{n-k} \cdot se(\hat{\beta}_0), \\
IC_{\hat{\beta}_1} &= \hat{\beta}_1 \pm t_{n-k} \cdot se(\hat{\beta}_1).
\end{aligned}
\tag{41}
$$

- Donde t_{n-k} cambiará dependiendo el nivel de confianza a utilizar y los grados de libertad. Debemos recurrir a la tabla de valores t.
- La expresión que resta (-) se conoce como el límite inferior y la que suma $(+)$ como el límite superior.
- Notar que se han definido los intervalos para caras. Si estamos interesados en calcular para una cara, tenemos que $\texttt{Prob}[t < t_{n-k}]$ y $\texttt{Prob}[t > -t_{n-k}]$, donde para nosotros $t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}$ (si hablamos de $\hat{\beta}_1$).

Inferencia estadística en el modelo de RLS: intervalos de confianza

- Recordemos que para nuestro ejemplo encontramos que $se(\hat{\beta}_0) = 11.91$ y $se(\hat{\beta}_1) = 0.21$. Ahora, si consideramos un nivel de confianza de 95% tenemos que:

$$t_{997,\frac{0.05}{2}} = 1.96. \tag{42}$$

- Por lo tanto, los intervalos de confianza para los parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$ corresponden a:

$$\begin{aligned} IC_{\hat{\beta}_0} &= 54.16 \pm 1.96 \cdot 11.91 = [30.81; 77.50], \\ IC_{\hat{\beta}_1} &= 5.39 \pm 1.96 \cdot 0.21 = [4.97; 5.80]. \end{aligned} \tag{43}$$

Inferencia estadística en el modelo de RLS: pruebas de hipótesis

- Supongamos que se desea probar la hipótesis de que el verdadero valor de $\hat{\beta}_1$ es 2.0. Se sabe que

$$t_c = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{n-k}. \tag{44}$$

- Sea el valor t_c el valor t observable. Si la hipótesis esta planteada de la siguiente manera:

$$\begin{aligned} H_0 &= \beta_1 = 2 \\ H_1 &= \beta_1 \neq 2 \end{aligned} \tag{45}$$

- Entonces es preciso considerar $|t_c|$ como estadístico de prueba.

Inferencia estadística en el modelo de RLS: pruebas de hipótesis

- Por lo tanto, si el valor verdadero de β_1 es 2, tenemos que $|t_c|$ viene dado por:

$$|t_c| = |\frac{5.39 - 2}{0.21}| = 16.14. \tag{46}$$

- Al observar la tabla t para $n - k = 997$ grados de libertad y un nivel de confianza de 95% (es decir, $\alpha = 0.05$ y $\frac{\alpha}{2} = 0.025$) tenemos que:

$$\texttt{Prob}(t > 1.96) = 0.025. \tag{47}$$

- Y así,

$$\texttt{Prob}(|t_c| > 16.14) = 1.48e - 52. \tag{48}$$

Inferencia estadística en el modelo de RLS: pruebas de hipótesis

- De manera general esta prueba de hipótesis se conoce como una **prueba de hipótesis de dos colas**. Donde se rechaza H_0 si:
  1. $t_c \in (t, +\infty)$ (es decir, $t_c > t$) o $t_c \in (-\infty, -t)$ (es decir, $t_c < -t$). El valor t se busca para $\frac{\alpha}{2}$[3].
  2. $2 \cdot \texttt{Prob}(|t_c| > |\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}|) < \alpha$.
- Para nuestro ejemplo anterior tenemos que $|t_c| = 16.14$, $t = 1.96$, $\alpha = 0.05$ y valor-p$=1.148e-52$. De modo que tenemos evidencia para rechazar la hipótesis nula H_0.
- Como en este caso, para $\alpha = 0.05$ existe evidencia estadística para rechazar H_0 diremos que: *con un 95% de probabilidad se rechaza H_0, lo que implica que el verdadero valor de β_0 es estadísticamente diferente de 2.*

[3] Esto por α se distribuye simétricamente en las dos colas de la distribución.

Inferencia estadística en el modelo de RLS: pruebas de hipótesis

- Recordemos nuestro ejemplo: y es el tiempo de respuesta de un sistema y x el número de *queries*. Para el cual tenemos la siguiente función de regresión poblacional:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \mu_i. \tag{49}$$

- Por un momento nos podemos preguntar, ¿cuál es el efecto de las *queries* (x) sobre el tiempo de respuesta (y)?
- Para esto podemos responder que el efecto viene dado a través de $\frac{\partial y}{\partial x} = \beta_1$. Y para el parámetro β_1 (por MCO/OLS) tenemos que $\hat{\beta_1} = 5.39$. Por lo tanto, si es que el número de *queries* aumenta en 1, entonces el tiempo de respuesta aumentará en 5.39 segundos.

Inferencia estadística en el modelo de RLS: pruebas de hipótesis

- Hasta el momento, bien. Sin embargo, respetando la naturaleza estadística de los resultados nos debemos preguntar lo siguiente: **¿es estadísticamente diferente de cero el efecto de x sobre y?**
- Esta pregunta es relevante por lo siguiente: desde $\frac{\partial y}{\partial x} = \beta_1$ sabemos que $\partial y = \beta_1 \cdot \partial x$ (el cambio en y es β_1 veces el cambio en x). De modo que si con alguna probabilidad $\beta_1 = 0$ entonces tendremos que $\partial y = 0 \cdot \partial x$ lo que indica que el cambio en x no afecta significativamente a y.

Inferencia estadística en el modelo de RLS: pruebas de hipótesis

- Para responder la pregunta anterior, podemos utilizar la noción de pruebas de hipótesis que se presento anteriormente.
- Ahora, si estamos interesados en evaluar que el verdadero valor de β_1 sea diferente de cero, podemos plantear que:

$$\begin{aligned} H_0 &: \beta_1 = 0, \\ H_1 &: \beta_1 \neq 0. \end{aligned} \tag{50}$$

- Luego tendremos que nuestro estadístico de prueba toma la siguiente forma:

$$t_c = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}. \tag{51}$$

Inferencia estadística en el modelo de RLS: pruebas de hipótesis

- Como estamos en una prueba de 2 colas sabemos que H_0 se rechaza si $|t_c| > t$ o `valor-p`$< \alpha$.
- Ahora:
  1. Si **se rechaza** H_0 para un nivel de α dado tenemos que el parámetro β_1 **es** estadísticamente diferente de 0. Es decir, la variable x **tiene efecto significativo** sobre la variable y.
  2. Si **no se rechaza** H_0 para un nivel de α dado tenemos que el parámetro β_1 **no es** estadísticamente diferente de 0. Es decir, la variable x **no tiene efecto significativo** sobre la variable y.

Inferencia estadística en el modelo de RLS: pruebas de hipótesis

- En nuestro ejemplo: $t_c = \frac{5.39}{0.21} = 25.66$ y $t_{997,\frac{\alpha}{2}}$ para $\alpha = 0.05$ es $t = 1.96$. Por lo tanto, se rechaza la hipótesis nula H_0 y esto implica que las *queries* tiene un impacto significativo sobre el tiempo de respuesta del sistema.

Inferencia estadística en el modelo de RLS: pruebas de hipótesis

- Cuando hablamos de pruebas de hipótesis de **una cola** los criterios de decisión cambian.
  1. Prueba de cola derecha:
     - El planteamiento de hipótesis corresponde a:

$$\begin{aligned} H_0 :&\beta_1 \leq 0, \\ H_1 :&\beta_1 > 0. \end{aligned} \tag{52}$$

     - El valor t se encuentra en la tabla para $t_{n-k,\alpha}$ y se rechaza H_0 cuando $t_c \in (t, +\infty)$
  2. Prueba de cola izquierda:
     - El planteamiento de hipótesis corresponde a:

$$\begin{aligned} H_0 :&\beta_1 \geq 0, \\ H_1 :&\beta_1 < 0. \end{aligned} \tag{53}$$

     - El valor t se encuentra en la tabla para $t_{n-k,\alpha}$ y se rechaza H_0 cuando $t_c \in (-\infty, -t)$

Inferencia estadística en el modelo de RLS: relación intervalos de confianza y pruebas de hipótesis

- Para nuestro ejemplo. Si estamos interesados en probar:

$$\begin{aligned} H_0 &: \beta_1 = 0, \\ H_1 &: \beta_1 \neq 0. \end{aligned} \tag{54}$$

- Podemos recurrir a los intervalos de confianzas. Recordemos que anteriormente encontramos que el intervalo de confianza para $\hat{\beta}_1$ con 95% es $IC_{\hat{\beta}_1} = [4.97; 5.80]$.
- Notar a través de nuestra hipótesis queremos probar que $\beta_1 = 0$. Valor que no se encuentra contenido en el intervalo de confianza, por lo tanto con 95% de confianza rechazamos H_0 y el verdadero valor de β_1 no será 0.
- En caso que el "valor" que estamos interesados en probar se encuentre contenido en el intervalo de confianza, entonces no podremos rechazar la idea de H_0.

Predicción en el modelo de regresión lineal simple

- La ecuación estimada de regresión $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$ se utiliza para predecir el valor de y para determinados valores de x.
- Sea x_0 el valor determinado de x. Entonces, podemos predecir y_0 como:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_0. \tag{55}$$

- El verdadero valor de y_0 está dado por:

$$y_0 = \beta_0 + \beta_1 \cdot x_0 + \mu_0, \tag{56}$$

- Donde μ_0 es el término del error.

Predicción en el modelo de regresión lineal simple

- Por lo tanto, el error de la predicción corresponde a:

$$\hat{y}_0 - y_0 = (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) \cdot x_0 - \mu_0. \tag{57}$$

- Y como $(E)(\hat{\beta}_0 - \beta_0) = 0$ y $(E)(\hat{\beta}_1 - \beta_1) = 0$, tenemos que

$$\mathbb{E}(\hat{y_0} - y_0) = 0. \tag{58}$$

- Esta ecuación muestra que el predictor que se utiliza es insesgado.

Predicción en el modelo de regresión lineal simple

- Ahora, la varianza del error de predicción es:

$$
\begin{aligned}
\mathbb{V}(\hat{y_0} - y_0) &= \mathbb{V}(\hat{\beta_0} - \beta_0) + x_0^2 \cdot \mathbb{V}(\hat{\beta_1} - \beta_1) + 2 \cdot x_0 \cdot \texttt{cov}(\hat{\beta_0} - \beta_0, \hat{\beta_0} - \beta_0) + \mathbb{V}(\mu_0), \\
\mathbb{V}(\hat{y_0} - y_0) &= \sigma^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{\mathbb{V}(x)} \right) + \sigma^2 \cdot \frac{x_0^2}{\mathbb{V}(x)} - 2 \cdot x_0 \cdot \sigma^2 \cdot \frac{\bar{x}}{\mathbb{V}(x)} + \sigma^2, \\
\mathbb{V}(\hat{y_0} - y_0) &= \sigma^2 \cdot \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathbb{V}(x)} \right).
\end{aligned}
\tag{59}
$$

- Notar que la varianza se incrementa conforme el valor de x_0 se aleja de $\bar{x}$ sobre la que se calcularon los parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$.

Predicción en el modelo de regresión lineal simple

- Por lo tanto, si x_0 se encuentra dentro del rango de observaciones muestrales de x es posible llamar a la predicción como **predicción dentro de la muestra** (*in-sanple*). Por otra parte, cuando x_0 se encuentra fuera de dicho rango, la predicción como **predicción fuera de la muestra** (*out-sanple*).

Predicción en el modelo de regresión lineal simple

- Recordemos que nuestro ejemplo tenemos que: $\bar{x} =$55.57, $\hat{\sigma^2} = 31.59$, $\mathbb{V}(x) = 682.93$ y $n = 999$.
  1. Predicción *in-sample*: consideremos que $x_0 = 50$.

$$\begin{aligned} \hat{y_0} &= 54.16 + 5.39 \cdot 50 = 323.66, \\ se(\hat{y_0}) &= \sqrt{31.39 \cdot (1 + \frac{1}{999} + \frac{(50 - 55.57)^2}{682.93})} = 5.73. \end{aligned} \tag{60}$$

- Por lo tanto, el valor predicho del tiempo de respuesta (y) cuando el número de *queries* (x) es 50 unidades es de 323.66 segundos.
- Notar que como conocemos el error estándar, podemos calcular el intervalo de confianza de la predicción como:

$$IC_{\hat{y_0}} = \hat{y_0} \pm t_{n-k,\frac{\alpha}{2}} \cdot se(\hat{y_0}). \tag{61}$$

Predicción en el modelo de regresión lineal simple

- Recordemos que nuestro ejemplo tenemos que: $\bar{x} =$55.57, $\hat{\sigma^2} = 31.59$, $\mathbb{V}(x) = 682.93$ y $n = 999$.
  2. Predicción *out-sample*: consideremos que $x_0 = 150$.

$$\hat{y_0} = 54.16 + 5.39 \cdot 150 = 862.66,$$
$$se(\hat{y_0}) = \sqrt{31.39 \cdot (1 + \frac{1}{999} + \frac{(150 - 55.57)^2}{682.93})} = 21.00. \tag{62}$$

- Por lo tanto, el valor predicho del tiempo de respuesta (y) cuando el número de *queries* (x) es 150 unidades es de 862.66 segundos.
- Notar que como conocemos el error estándar, podemos calcular el intervalo de confianza de la predicción como:

$$IC_{\hat{y_0}} = \hat{y_0} \pm t_{n-k,\frac{\alpha}{2}} \cdot se(\hat{y_0}). \tag{63}$$

Formas funcionales alternativas

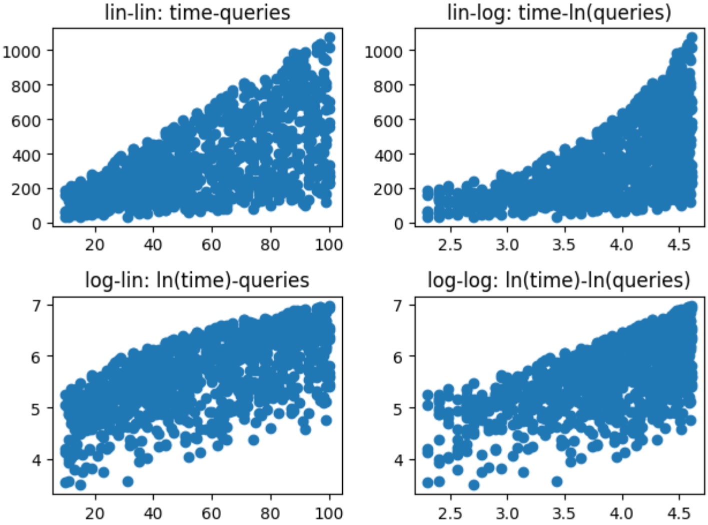

Figure: Relación tiempo y *queries*

Formas funcionales alternativas: lin-log

- Para la siguiente especificación

$$y_i = \beta_0 + \beta_1 \ln(x_i) + \mu_i \tag{64}$$

- β_1 corresponde a

$$\beta_1 = \frac{\partial y_i}{\partial \ln(x_i)} = \frac{\partial Y_i}{\frac{1}{x_i}\partial x_i} = \frac{\Delta y}{\Delta \% x} \tag{65}$$

- Ahora, β_1 corresponde a una **semielasticidad** de y con respecto a x. Muestra como cambia el nivel de y ante un cambio porcentual de x.

Formas funcionales alternativas: log-lin

- Para la siguiente especificación

$$\ln(y_i) = \beta_0 + \beta_1 x_i + \mu_i \tag{66}$$

- β_1 corresponde a

$$\beta_1 = \frac{\partial \ln(y_i)}{\partial x_i} = \frac{\frac{1}{y_i}\partial y_i}{\partial x_i} = \frac{\Delta \%_0 y}{\Delta x} \tag{67}$$

- Ahora, β_1 corresponde a una **semielasticidad** de Y con respecto a X. Muestra como cambia porcentualmente Y ante un cambio en una unidad en X.

Formas funcionales alternativas: log-log

- Para la siguiente especificación

$$\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + \mu_i \tag{68}$$

- β_1 corresponde a

$$\beta_1 = \frac{\partial \ln(y_i)}{\partial \ln(x_i)} = \frac{\frac{1}{y_i}\partial y_i}{\frac{1}{x_i}\partial x_i} = \frac{\Delta\% y}{\Delta\% x} \tag{69}$$

- Ahora, β_1 corresponde es una **elasticidad**. Muestra el cambio porcentual que experimenta y ante un cambio porcentual en x

Formas funcionales alternativas

- Para el caso de regresión lineal simple

Modelo	V.a.E	V.E.	$\hat{\beta}_1$
lin-lin	tiempo	*queries*	5.39
lin-log	tiempo	$\ln$(*queries*)	230.72
log-lin	$\ln$(tiempo)	*queries*	0.01
log-log	$\ln$(tiempo)	$\ln$(*queries*)	0.74

Nota: V.a.E. se refiere a Variable a Explicar (y) y V.E. se refiere a Variable Explicativa (x).

Formas funcionales alternativas

- Para modelo lin-lin

Si las *queries* aumentan en una unidad, entonces el tiempo de respuesta del sistema aumentará en 5.39 segundos.

- Para modelo lin-log

Si las *queries* aumentan en 1%, entonces el tiempo de respuesta del sistema aumentará en 2.30 segundos.

- Para modelo log-lin

Si las *queries* aumentan en una unidad, entonces el tiempo de respuesta del sistema aumentará en 1%.

- Para modelo log-log

Si las *queries* aumentan en 1% entonces el tiempo de respuesta del sistema aumentará en 0.74%.

Econometría I

regresión simple

Leonardo Manríquez M.
(lmanriquez@ucsc.cl)

August 20, 2023