Regresión Lineal y OLS

Camila J. Pulgar F.- Paz Montaño Kerdy,

Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Industrial

28 de mayo de 2021

Contenidos

Repaso Regresión Lineal - OLS

- OLS es una forma de estimar el modelo poblacional de regresión lineal:
  - Modelo Poblacional:

$$y = aX + b$$

  - Estimación:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta} X + \epsilon$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_N x_N + \epsilon$$

Repaso Regresión Lineal - OLS

- Modelo Poblacional:

$$y = aX + b$$

Componentes del modelo:

  - y : Variable Endógena, Variable Dependiente, Variable Explicativa (explicada), Variable a Predecir.

- Estimación:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta} X + \epsilon$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_N x_N + \epsilon$$

Componentes del modelo:

  - $\hat{y}$: Estimación de la Variable dependiente.
  - $\hat{\beta}_j$: Coeficientes de la regresión (Slope).
  - $\hat{\beta}_0$: Intercepto (intercept)
  - $x_1, \ldots, x_N$: Variables independientes, variables exógenas, variables explicativas, predictores regresores, variables de control, etc.

Residual vs Errror

- ¿Residual vs Error?
- Residual: $\hat{\epsilon_i} = y_i - \hat{y_i}$
- Error: Representa factores distintos a x que afectan a y. Es no observable, por lo que no se puede calcular. Además, es considerado como una variable aleatoria.

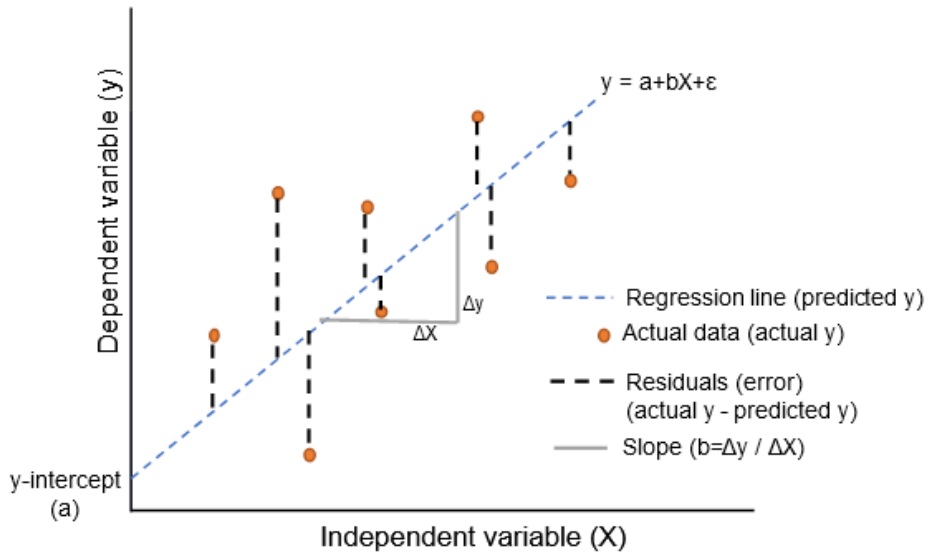

Figura 1: Componentes de la regresión

Mínimos cuadrados ordinarios (OLS)

OLS: Ordinary Least Square

- Las estimaciones las obtenemos utilizando el error cuadrático medio (MSE), la cual pretende minimizar la variable aleatoria del error, aquel que desconocemos.

$$MSE = \mathbb{E}[(Y - m(X))^2]$$

- Que en este caso:

$$MSE = \mathbb{E}[(Y - \hat{\beta}_0 - \hat{\beta}_1 * X)^2]$$

- Como estamos trabajando con una muestra aleatoria, la esperanza la estimamos como el promedio de los residuos al cuadrado:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \hat{\epsilon}_i^2$$

Mínimos cuadrados ordinarios (OLS)

OLS: Ordinary Least Square

- Residuos

$$\hat{\epsilon_i} = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

- Optimización: Minimizar el error, es decir la distancia a la recta

$$min_{\beta_0,\beta_1} \sum_i^N (\hat{\epsilon_i})^2 = min_{\beta_0,\beta_1} \sum_i^N (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

Modelo de Regresión Poblacional

$$m(x) = a^* + b^* x$$

- $a^* = \mathbb{E}(Y) - b\,\mathbb{E}(X)$
- $b^* = \frac{Cov(Y,X)}{Var(X)}$

Modelo de Regresión Lineal OLS

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon_i}$$

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- $\hat{\beta}_1 = \frac{\sum_i x_i (y_i - \bar{y})}{\sum_i x_i (x_i - \bar{x})}$

Sesgo

OLS: Ordinary Least Square

$$b^* = \frac{Cov(Y,X)}{Var(X)} = \hat{\beta}_1 + \underbrace{\frac{Cov(X,\epsilon)}{Var(X)}}_{\textbf{Sesgo}}$$

- Cuando

$$Cov(X,\epsilon) = 0 \Rightarrow b^* = \beta_1$$

Si $Cov(X,\epsilon) = 0$ entonces el β_1 es igual al beta óptimo del modelo de regresión poblacional. Si es distinto, mi estimador de regresión lineal de OLS esta sesgado respecto a mi modelo de regresión poblacional.

- X: Información que poseemos, exógena.
- ϵ: Todo lo que no podemos observar y que por ende no estamos incorporando en el modelo, es todo lo que no explica X.

Demostración Clase 13 (parte 2) - Modelo de regresión linea.

R^2: Coeficiente de Determinación Múltiple

El coeficiente de Determinación Múltiple R^2

El R^2 es una métrica que nos indica la cantidad de información (variabilidad) que es capturada por el modelo versus el total de variabilidad que hay en la variable de interés (Y).

Y se construye a partir de:

- Suma de Cuadrados Totales (*Sum of Squares Total*):
  $SST = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$
- Suma de Cuadrados de la Regresión (*Sum of Squares Regression*):
  $SSR = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$
- Suma de Cuadrados de Error (*Sum of Squares Error*):
  $SSE = \sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2$

Y, se cumple la siguiente identidad con estas sumas de cuadrados:

$$SST = SSR + SSE$$

R^2: El coeficiente de Determinación Múltiple

En base a esto es que R^2 se define como:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Es decir, es el ratio de la variabilidad que es explicada por el modelo del total de la variabilidad en Y.
- Por ende, sus valores van desde 0 hasta 1, donde a mayor valor, mejor es el modelo en explicar la variable dependiente Y.

Supuestos OLS

- Linealidad en los parámetros
- Muestreo Aleatorio
- No Multicolinealidad
- Exogeneidad
- Homocedasticidad

Supuestos OLS

Linealidad en los parámetros β y NO EN LOS x_i

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_N x_N + \epsilon$$

- Error se integra de forma aditiva
- $E[X] = 0$, ϵ sea iid.
- Si la esperanza del error no es cero no es tan terrible, modelamos aditivamente restando su esperanza y creando un nuevo error para que $E[X] = 0$ manteniendo las características.
- No necesitamos que los errores tengan una distribución especifica tales como; Normal, Logarítmico,etc. Sino que simplemente provengan de la misma distribución, de esta forma cumplan con ser independientes e identicamente distribuidos.

Supuestos OLS

Muestra aleatoria:

$$\{x_i, y_i\}_{i=1}^{N}$$

Se requiere que la muestra represente bien la variabilidad de los datos

- Si $Var(x_k)$ es muy chica, entonces $Var(\hat{\beta}_k)$ será muy grande y de esta forma el estimador va a ser menos preciso en la estimación.
- Mientras mayor información aporta x - **A mayor dispersión de x**, más chico va a ser la varianza o el intervalo de confianza para el estimador por OLS.

Demostración Próxima clase

No Multicolinealidad:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$
$$\hat{\beta} = \beta + \left(X^T X\right)^{-1} X^T \epsilon$$

La matriz $X^T X$ debe ser de rango completo. Es decir, sus filas deben ser linealmente independiente y sus columnas deben ser linealmente independiente.

- Necesitamos que ninguna variable X_k que agregamos al modelo pueda ser explicada al 100 % por las otras X_{1-k} variables en el modelo
- Si queremos verificar si una variable X_k puede ser explicada por otras variables podemos proponer la siguiente regresión :

$$X_k = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 ... \alpha_{k-1} X_{k-1} + \alpha_k X_k + \check{\delta}$$

- Si

$$R^2 = 1 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$\Rightarrow \underbrace{\frac{SSR}{SST}}_{Var\ explican\ toda\ la\ variabilidad\ en\ X_k} = 1 - \underbrace{\frac{SSE}{SST}}_{No\ explican\ los\ residuos}$$

Exogeneidad o No Endogeneidad:

Supuesto sobre el error:

$$E\left[\epsilon_i / x_i\right] = 0$$
$$E\left[\epsilon_i\right] = 0$$
$$Cov(X, \epsilon)) = 0$$

- No existe correlación entre los regresores del modelo y el error. ¿Esto se cumple en la realidad? ¿Cómo validar este supuesto?
- Este supuesto asegura que el estimador OLS sea insegado, el efecto esta correctamente identificado y de esta forma podamos atribuir causalidad al efecto del tratamiento que estamos estudiando.
- No hay Endogeneidad $E[\epsilon_i|X] = 0$. Valor esperado del error aleatorio no varía condicional a las observaciones, es decir, variables no observadas no están relacionadas con las incluidas en el modelo. La media del termino del error no observable es simpre la misma e igual a cero.

Supuestos OLS

Homocedasticidad:

El error posee varianza constante sobre las observaciones y variables explicativas del modelo.

$$\text{Var}\left(\epsilon_i\right) = \sigma^2$$

- Necesitamos este supuesto para la **eficiencia** del estimador OLS.
- Eficiencia: Corresponde a una propiedad de los estimadores. Un estimador es más eficiente que otro cuando tiene menor varianza que el otro estimador.

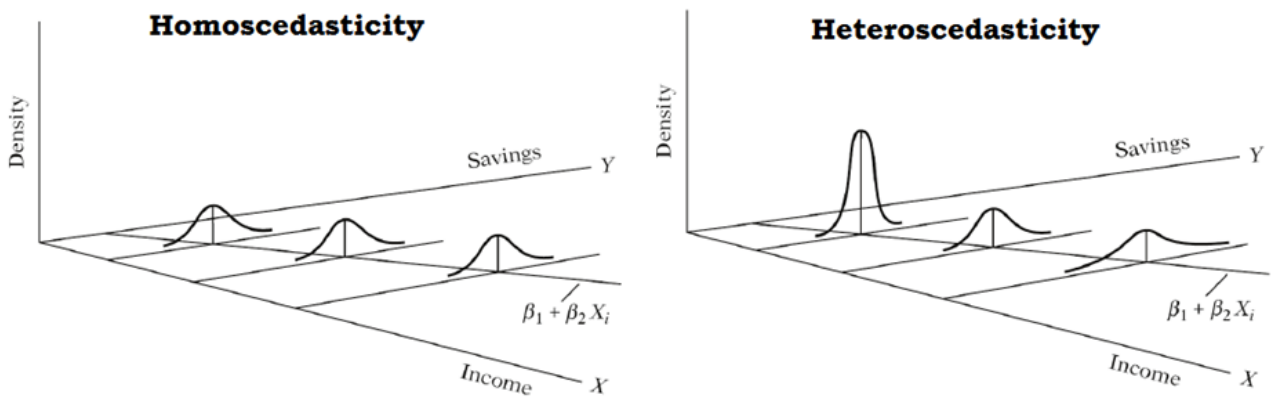

Figura 2: Homocedasticidad "Izquierda" - Heterocedasticidad "Derecha"

Teorema Gauss Markov

Estimador BLUE.

Bajo los 5 supuestos anteriores, el estimador de OLS β_{OLS} es el Mejor Estimador Lineal Insesgado de Menor Varianza (Best Linear Unbiased Estimator, BLUE):

- Mejor (Best): Mínima varianza. Gracias a los supuestos de varianza.
- Lineal (Linear): Primer supuesto, linealidad sobre los parámetros.
- Insesgado (Unbiased): $E[\hat{\beta}] = \beta$. Gracias a los supuestos de muestreo aleatorio y exogeneneiad.
- Estimador (Estimator).

Figura 3:

Interpretación de Coeficientes

Tabla 1: Interpretación de Coeficientes.

Modelo	**Regresión**	**Variable Dep (Y)**	**Variable Ind (X)**	**Interpretación del Regresor (β_1)**
Nivel - Nivel	$Y = \beta_0 + \beta_1 X + \epsilon$	Y	X	$\Delta Y = \beta_1 \Delta X$
Log- Nivel	$\log(Y) = \beta_0 + \beta_1 X + \epsilon$	log(Y)	X	$\%\Delta Y = 100\beta_1 \Delta X$
Nivel - Log	$Y = \beta_0 + \beta_1 \log(X) + \epsilon$	Y	log(X)	$\Delta Y = \frac{\beta_1}{100} \Delta X$
Log - Log	$\log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon$	log(Y)	log(X)	$\%\Delta Y = \beta_1 \%\Delta X$

Interpretación de Coeficientes

- Level-Level: Si hay un cambio de 1 unidad en x, se espera un cambio de β_1 en y, es decir, $\Delta y = \beta_1 \Delta x$

$$y = \beta_0 + \beta_1 x + \epsilon.$$

- Log-Level: Si hay un cambio de 1 unidad en x, se espera un cambio porcentual de $100\beta_1$ en y, es decir, $\%\Delta y = 100\beta_1 \Delta x$

$$\log(y) = \beta_0 + \beta_1 x + \epsilon.$$

- Level-Log: Si hay un cambio de $1\,\%$ en x, se espera un cambio de $\beta_1/100$ unidades en y, es decir, $\Delta y = \frac{\beta_1}{100}\Delta x$

$$y = \beta_0 + \beta_1 \log(x) + \epsilon.$$

- Log-Log: Si hay un cambio de $1\,\%$ en x, se espera un cambio de $\beta_1\,\%$ en y, es decir, $\%\Delta y = \beta_1\,\%\Delta x$

$$\log(y) = \beta_0 + \beta_1 \log(x) + \epsilon$$

Test de Significancia

- Test de significancia individual sobre parámetros. Asumiendo $E\left[\epsilon_i / x_i\right] = 0$ $\epsilon_i \sim N\left(0, \sigma^2\right)$

$$H_0 : \beta_i = 0$$
$$H_A : \beta_i \neq 0$$
$$t = \frac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)} \sim t_{n-k-1}$$

- No se conoce σ^2, por tanto se utiliza $SE\left(\hat{\beta}_j\right)$ y una distribución t-Student.

Dudas o consultas al:

$camilapulgarf@gmail.com$

$sofia.montano@ug.uchile.cl$

Gracias por su atención