Bondad de Ajuste y Propiedades Estadísticas del Modelo de Regresión

Métodos cuantitativos aplicados a estudios urbanos II - MEU UTDT

Ricardo Pasquini - rpasquini@utdt.edu

June 20, 2025

Notación para modelización usando regresión

Ejemplo Modelo Hedonico de Alquileres

$$\underbrace{Y_i}_{\substack{\text{Valor} \\ \text{Alquiler}}} = \underbrace{\beta_0}_{\text{Intercepto}} + \beta_1 \cdot \underbrace{X_{1,i}}_{\substack{\text{Número de} \\ \text{Habitaciones}}} + \beta_2 \cdot \underbrace{X_{2,i}}_{\substack{\text{Número de} \\ \text{Baños}}} + \cdots + \underbrace{\varepsilon_i}_{\text{Error}}$$

- $\beta_0, \beta_2, \ldots, \beta_k$ son los coeficientes a ser estimados.
- Una vez estimados los denotamos con $\hat{\beta}_0, \hat{\beta}_2, \ldots, \hat{\beta}_k$.
- Los coeficientes medirán la contribución a la variable explicada por cada unidad de la variable explicativa.

Explicación y Predicción

- **Explicación**
- Los coeficientes estimados $\hat{\beta}_0, \hat{\beta}_2, \ldots, \hat{\beta}_k$ son la base para la explicación.
- Si bien estrictamente hablando, los coeficientes solo capturan variaciones, y por lo tanto no representan necesariamente causalidad, son la base bajo las cuales buscaremos identificar efectos causales.
- **Predicción**
- La predicción se obtiene reemplazando los coeficientes estimados en la ecuación, y utilizando estos componentes para calcular el valor de que predice el modelo.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \cdots + \hat{\beta}_k X_{k,i}$$

Medidas de Bondad de Ajuste

R^2

- El R^2 es una medida de la bondad de ajuste del modelo en su conjunto.
- Se define como el cociente entre la varianza explicada por el modelo y la varianza total de la variable dependiente.
- Se interpreta como la proporción de la varianza de la variable dependiente que es explicada por el modelo.

Bondad de Ajuste: R^2

- Puesto que vale que

$$Y_i = \hat{Y}_i + \hat{\varepsilon}_i$$

- También vale que la varianza de Y_i es la suma de la varianza de la predicción y la varianza de los residuos.
- Esta igualdad nos permite definir una medida de la bondad de ajuste del modelo: el R^2 es la proporción de la varianza de la variable dependiente que es explicada por el modelo.

$$1 = \frac{Var(\hat{Y}_i)}{Var(Y_i)} + \frac{Var(\hat{\varepsilon}_i)}{Var(Y_i)}$$

$$R^2 = \frac{Var(\hat{Y}_i)}{Var(Y_i)} = 1 - \frac{Var(\hat{\varepsilon}_i)}{Var(Y_i)}$$

- Notar que es un valor entre 0 y 1.

Bondad de Ajuste: R^2

Ejemplo Modelo Hedonico de Alquileres

OLS Regression Results

Dep. Variable:	price	R-squared:	0.049
Model:	OLS	Adj. R-squared:	0.049
Method:	Least Squares	F-statistic:	1443.
Date:	Fri, 13 Jun 2025	Prob (F-statistic):	7.68e-308
Time:	11:33:08	Log-Likelihood:	-3.0876e+05
No. Observations:	27879	AIC:	6.175e+05
Df Residuals:	27877	BIC:	6.175e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.151e+04	205.145	56.088	0.000	1.11e+04	1.19e+04
bedrooms	4253.5067	111.984	37.983	0.000	4034.013	4473.000

Omnibus:	39834.368	Durbin-Watson:	1.851
Prob(Omnibus):	0.000	Jarque-Bera (JB):	59814269.062
Skew:	7.918	Prob(JB):	0.00
Kurtosis:	229.365	Cond. No.	5.02

Medidas de Bondad de Ajuste

Error Cuadrático Medio

- El error cuadrático medio (MSE) es otra medida de la bondad de ajuste del modelo en su conjunto.
- Es una medida que busca cuantificar la magnitud de los errores.

$$MSE = \frac{\sum \hat{\varepsilon}_i^2}{n - (k+1)}$$

- El $-(k+1)$ en el denominador es una corrección estadística al promedio, dada por el número de coeficientes estimados.
- La raiz cuadrada del MSE (o RMSE) es una medida que puede compararse en magnitud con la variable dependiente.

Medidas de Bondad de Ajuste: Error Cuadrático Medio

Ejemplo Modelo Hedonico de Alquileres

	price	bedrooms
count	27879.000000	27879.000000
mean	18441.120341	1.630403
std	16017.027024	0.835310
min	50.000000	-2.000000
25%	10500.000000	1.000000
50%	15000.000000	1.000000
75%	22900.000000	2.000000
max	666666.000000	13.000000

```
resultados.mse_resid
```

```
np.float64(243930120.92914465)
```

```
[54] # Le tomo raiz
     resultados.mse_resid**0.5
```

```
np.float64(15618.262417091879)
```

Propiedades Estadísticas de los Coeficientes

Insesgadez

- **Ausencia de Sesgo (Insesgadez):** Esta propiedad establece que, *en valor esperado*, los coeficientes estimados serán iguales a los verdaderos coeficientes poblacionales.
- Formalmente, se define como:

$$E(\hat{\beta}_j) = \beta_j$$

- Intuición: Aunque las estimaciones individuales pueden variar debido a la aleatoriedad inherente del muestreo, si pudiéramos repetir esta estimación en múltiples muestras, el promedio de estas estimaciones coincidiría con el valor verdadero.

Propiedades Estadísticas de los Coeficientes

Insesgadez

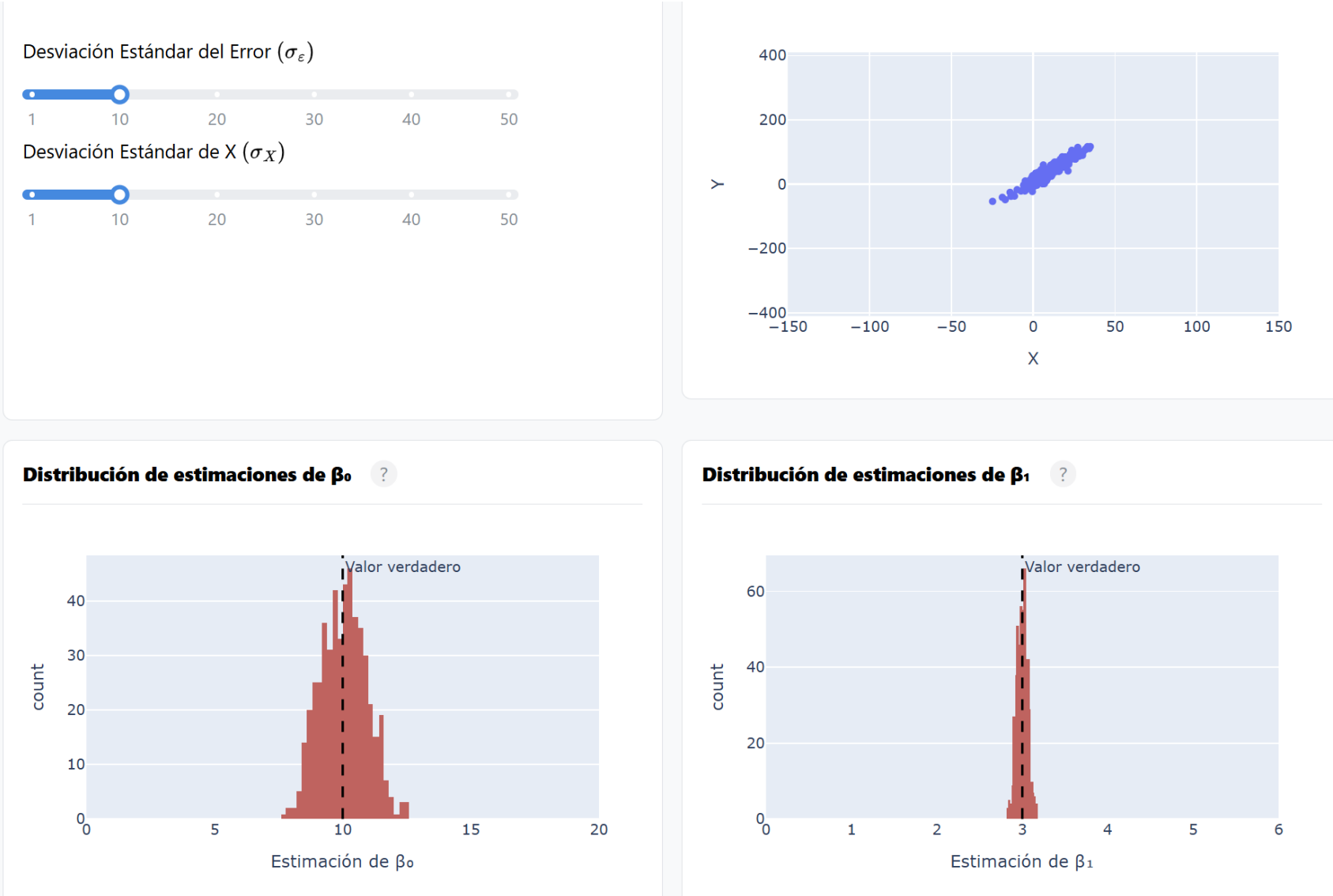


Figure: Simulación. Ver en SimuEcon.com

Propiedades Estadísticas de los Coeficientes

Varianza

- Aunque insesgadas, nuestras estimaciones siempre exhibirán cierto grado de **varianza**, que cuantifica la incertidumbre alrededor del coeficiente estimado.
- Formalmente, se define como:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}$$

- σ^2 es la varianza de los residuos.
- $\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$ es la suma de los cuadrados de las desviaciones de la variable independiente X_j con respecto a su media.
- Con múltiples variables se incorporará un factor de corrección adicional que dependerá de la correlación entre las variables independientes. (próxima clase)

Propiedades Estadísticas de los Coeficientes

Varianza

- La teoria indica que la varianza de nuestras estimaciones está influenciada por dos factores clave:
  - **Error en el modelo:** La presencia de variación no explicada en la variable dependiente (Y) contribuye a la varianza de nuestras estimaciones. Este error puede atribuirse a factores no incluidos en el modelo o a la aleatoriedad inherente en los datos.
  - **Variabilidad de la variable independiente:** Una mayor dispersión en los valores de nuestra variable independiente (X) conduce a una menor varianza en nuestras estimaciones de coeficientes. Esto se debe a que un rango más amplio de valores de X proporciona más información para estimar la relación con Y.

Propiedades Estadísticas de los Coeficientes

Distribución de probabilidad de los Coeficientes

- Bajo ciertos supuestos, los coeficientes siguen una distribución normal:

$$\hat{\beta} \sim N(\beta, Var(\beta))$$

- El valor estandarizado sigue una Normal Estándar:

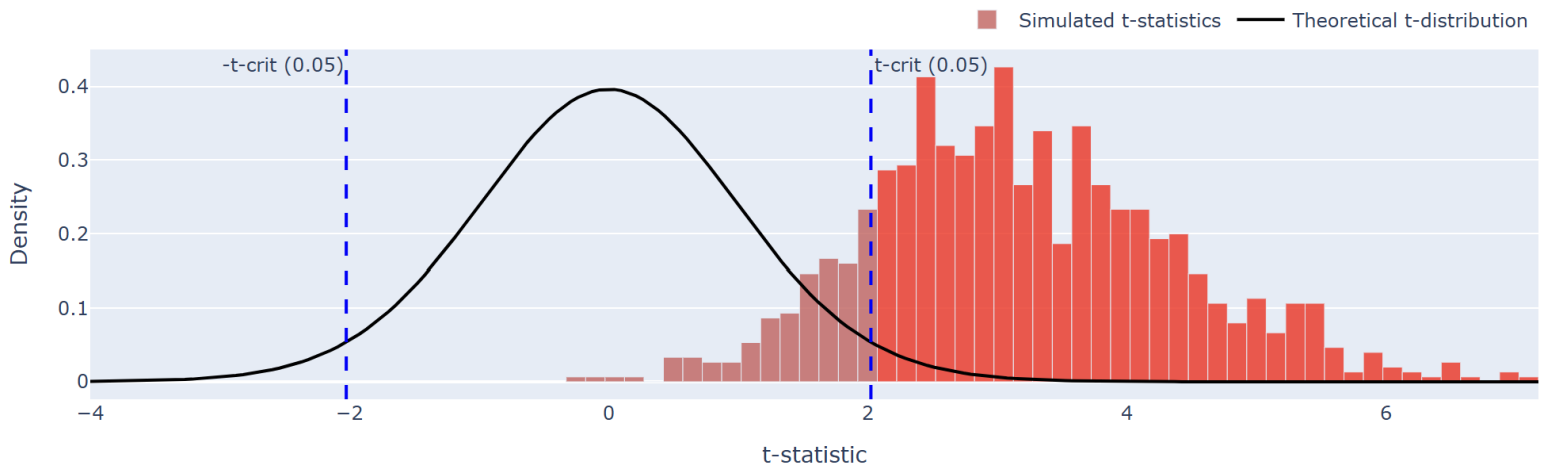
$$Z = \frac{\hat{\beta} - \beta}{\sqrt{Var(\beta)}} \sim N(0, 1)$$

- En la práctica, usamos:

$$T = \frac{\hat{\beta} - \beta}{\sqrt{\hat{Var}(\beta)}} \sim T_{n-1}$$

Test de Hipótesis

Distribución T y Potencia



Nivel de Significancia (α)	Potencia (%)
α = 0.01	62.50%
α = 0.05	85.10%
α = 0.10	91.40%

Figure: Distribución del estadístico T bajo H_0

Test de Hipótesis

- Test típico:

$$\begin{cases} H_0 : \beta = 0 \\ H_a : \beta \neq 0 \end{cases}$$

- Interpretación: Si H_0 es válida, X no tiene efecto sobre Y
- **Lógica del Test:**
  1. Asumimos $\beta = 0$ (Hipótesis nula)
  2. Derivamos la distribución de probabilidad según esa hipótesis
  3. Observamos nuestra estimación $\hat{\beta}$ y construímos el estadístico de prueba $\hat{T}$
  4. Si $\hat{T}$ es muy atípico, rechazamos H_0
- Cuán atípico es $\hat{T}$? **P-valor:** Probabilidad de observar un valor tan extremo como $\hat{T}$ bajo H_0

Test de Hipótesis

OLS Regression Results

Dep. Variable:	price	R-squared:	0.049
Model:	OLS	Adj. R-squared:	0.049
Method:	Least Squares	F-statistic:	1443.
Date:	Fri, 13 Jun 2025	Prob (F-statistic):	7.68e-308
Time:	11:33:08	Log-Likelihood:	-3.0876e+05
No. Observations:	27879	AIC:	6.175e+05
Df Residuals:	27877	BIC:	6.175e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.151e+04	205.145	56.088	0.000	1.11e+04	1.19e+04
bedrooms	4253.5067	111.984	37.983	0.000	4034.013	4473.000

Omnibus:	39834.368	Durbin-Watson:	1.851
Prob(Omnibus):	0.000	Jarque-Bera (JB):	59814269.062
Skew:	7.918	Prob(JB):	0.00
Kurtosis:	229.365	Cond. No.	5.02

Figure: Interpretación del P-valor. Ver en SimuEcon.com

Test de Hipótesis

Ejemplo estimando velocidad de buses con BRT - Gonzalez y Silva 2025 JUE

Table 2
Priority infrastructure, bus speed, and travelers.

Dependent variable:	Log bus speed (km/hr)			Log million travelers		
	Work days			Work days		
	Peakhours	Off-peakhours	Weekend	Peakhours	Off-peakhours	Weekend
Panel A	(1)	(2)	(3)	(4)	(5)	(6)
Percentage route with bus corridors	0.197***	0.155***	−0.050	−0.177	0.246*	−0.001
	(0.042)	(0.043)	(0.039)	(0.207)	(0.147)	(0.101)
Percentage route with bus lanes	0.055	0.045	0.066	0.098	0.462*	0.603**
	(0.081)	(0.084)	(0.074)	(0.207)	(0.253)	(0.268)
Panel B						
Indicator route with bus corridors	0.044***	0.033***	−0.002	−0.030	0.036	0.017
	(0.011)	(0.010)	(0.008)	(0.065)	(0.030)	(0.016)
Indicator route with bus lanes	0.008	0.005	0.006	−0.008	−0.029	−0.006
	(0.008)	(0.008)	(0.008)	(0.033)	(0.030)	(0.028)
Observations	2,028	1,768	1,680	2,028	1,768	1,680
Bus routes	507	442	420	507	442	420
Trips (in millions)	16.1	20.3	17.5	16.1	20.3	20.4
Avg. dependent variable (levels)	19.22	20.87	23.88	0.31	0.33	0.23
Route fixed effects	Y	Y	Y	Y	Y	Y
Year fixed effects	Y	Y	Y	Y	Y	Y

* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

Notes: Panel A shows two-way fixed effects estimates between priority infrastructure (bus corridors, bus lanes) and (i) bus speed in columns 1-3, and (ii) travelers in columns 4-6. The unit of observation is a route in a given year between 2016 and 2019. Panel B presents estimates of the same relationship but using the method proposed by Borusyak et al. (2024). All regression specifications include route and year fixed effects. Panel A uses the percentage of the route with priority infrastructure as right-hand side variable while Panel B uses an indicator for routes with more than 10% of priority infrastructure. Each coefficient and standard error comes from a separate regression. Peak hours are from 6.30 to 8.29 h in the morning and from 17.30 to 20.29 h in the afternoon. Off-peak hours are from 9.30 to 12.29 h in the morning, from 14.00 to 17.29 h in the afternoon, and from 21.30 to 22.59 h at night. The remaining hours of the day correspond to "transition" or "night" hours. Work days include days from Monday to Friday that are not a holiday. Weekend hours include all hours on Saturdays, Sundays, and holidays. Robust standard errors are clustered at the route level.

Figure: Gonzalez y Silva 2025

Takeaways

- Mejorar la bondad de ajuste del modelo y evaluar las significatividad de variables individuales son objetivos diferentes y donde podemos alcanzar conclusiones diferentes (bajo ajuste general, alta significatividad individual, o viceversa)
- Reducir el error del modelo ayuda a mejorar la precisión de los coeficientes individuales
- Buscar ampliar la varianza de las variables explicativas incrementa la precisión.
- El test de hipótesis se realiza bajo una distribución de probabilidades centrada en la hipótesis nula (típicamente de 0 efecto). La rechazamos si encontramos un valor atípicamente alto (bajo). El p-valor, es la medida de cuán atípico es el valor encontrado.

Takeaways

- Puesto que solo podemos testear usando el estadístico T, y este estadístico considera el efecto en relación a su standard error, la *significatividad estadística* solo habla de un concepto relativo: cuan grande o chico es un efecto en relación a la precisión con la que fue estimado.