Modelos de Regresión y Series de Tiempo (MRST)
2025 - 02

Clase 4 – MRLS
Inferencia sobre el modelo de regresión (PH), prueba de utilidad y ANOVA

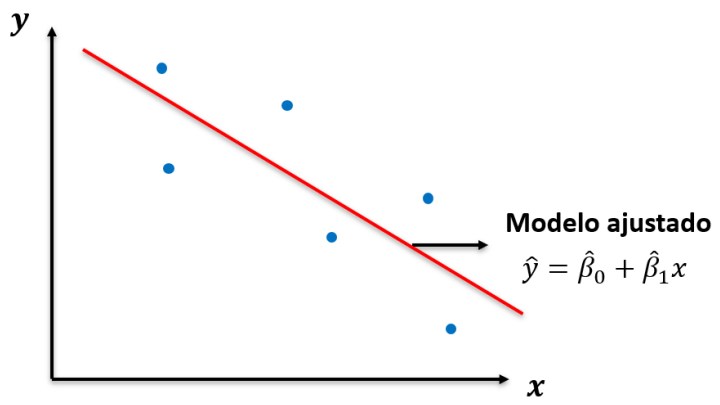
Docente: Natalia Jaramillo Quiceno

Escuela de Ingenierías

natalia.jaramilloq@upb.edu.co

Regresión lineal simple

Paso a paso muy simplificado



- Análisis **descriptivo** de lo datos
  - Gráfico de dispersión, coeficiente de correlación de Pearson (**R**)
- A la **ecuación** ajustada se le llama **modelo de regresión**
  - El ajuste se realiza mediante el método de los **mínimos cuadrados**
  - Se debe evaluar si el modelo es útil para describir la variable Y en función de X
- Es necesario evaluar la calidad del ajuste que presenta el modelo:
  - **Coeficiente de determinación r^2 o R^2**: cuánta variabilidad de los datos explica el modelo lineal
- Se debe comprobar que el modelo cumple unos supuestos:
  - Validación de supuestos sobre los residuos (errores): **normalidad, varianza constante**

Regresión lineal simple
Inferencia sobre β_0 y β_1

Propiedades de $\widehat{\boldsymbol{\beta}}_1$

$\hat{\beta}_1$ es una función lineal de variables aleatorias independientes $Y_1, Y_2, \ldots, Y_n$, cada una de las cuales está normalmente distribuida

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})Y_i}{S_{xx}} = \sum c_i Y_i \qquad \text{donde } c_i = (x_i - \bar{x})/S_{xx}$$

Así, se tiene que:

- $\boldsymbol{E(\widehat{\beta}_1) = \beta_1}$: estimador insesgado. La distribución de $\hat{\beta}_1$ siempre está centralizada en el valor β_1.

- $\mathbf{V}(\widehat{\boldsymbol{\beta}}_1) = \frac{\sigma^2}{S_{xx}}$ y $\mathbf{se}(\widehat{\boldsymbol{\beta}}_1) = \frac{\sigma}{\sqrt{S_{xx}}}$ (**se:** error estándar estimado)

- El estimador $\hat{\beta}_1$ tiene una distribución normal.

Regresión lineal simple
Inferencia sobre β_0 y β_1

Propiedades de $\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}$

Para el intercepto, se puede demostrar de la misma manera que:

- $\boldsymbol{E}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}) = \beta_0$: estimador insesgado. La distribución de $\hat{\beta}_0$ siempre está centralizada en el valor β_0.

- $\mathbf{V}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$ y

- $\mathbf{se}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}) = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$ (**se:** error estándar estimado)

- El estimador $\hat{\beta}_0$ tiene una distribución normal.

Regresión lineal simple
Inferencia sobre β_0 y β_1

Resumiendo, tenemos que:

- β_0 y β_1 son combinaciones lineales de los Y_i's que se distribuyen normal. Por tanto:

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx}) \quad \text{y} \quad \hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

- Los procedimientos inferenciales se basan en estandarizar un estimador, restando su valor medio y dividiéndolo entre su desviación estándar.

Así, para realizar inferencias (PH e IC) sobre β_0 y β_1 tenemos las siguientes variables estándar:

Para $\boldsymbol{\beta_0}$

$$T_0 = \frac{\hat{\beta}_o - \beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

Para $\boldsymbol{\beta_1}$

$$T_0 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2/S_{xx}}}$$

Ambos tienen una **distribución t** Con **n-2** grados de libertad

Nos concentraremos en $\boldsymbol{\beta_1}$

Regresión lineal simple

Pruebas de hipótesis para β_1

Ahora supongamos que se desea probar la hipótesis de que la pendiente es igual a una constante, por ejemplo, $\beta_{1,0}$. Las hipótesis correspondientes son:

$$H_0: \beta_1 = \beta_{1,0}$$

$$H_a: \beta_1 \neq \beta_{1,0}$$

Así, el estadístico de prueba es: $$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se\left(\hat{\beta}_1\right)}$$

Y la **región de rechazo** está dada por: $\left|t_0\right| > t_{\alpha/2,n-2}$

Luego, se **rechaza $\boldsymbol{H_0}$** con un nivel de significancia α si el valor del estadístico de prueba t_0 cae en la región de rechazo o, dicho de otra manera, si $\boldsymbol{Valor\ p < \alpha}$.

Regresión lineal simple

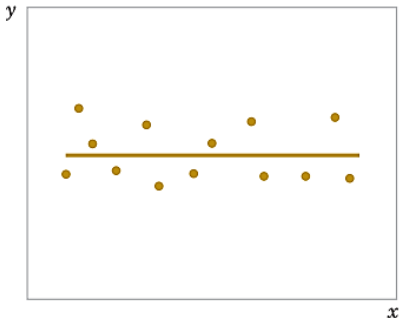
Utilidad (significancia) del modelo - Pruebas de hipótesis **especial** para $\boldsymbol{\beta_1}$

Caso especial muy importante:

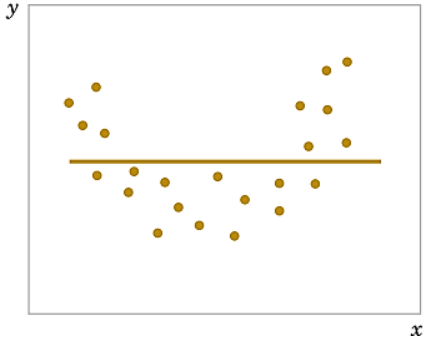
$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

Estas hipótesis se relacionan con la **utilidad de la regresión**

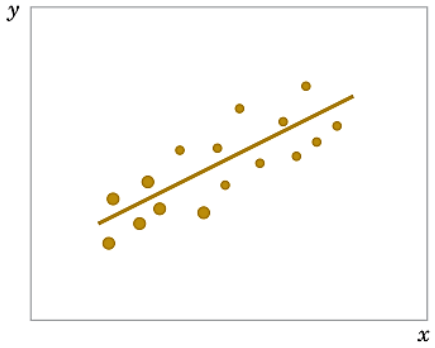
No rechazar H_0: no hay relación lineal entre x y y.



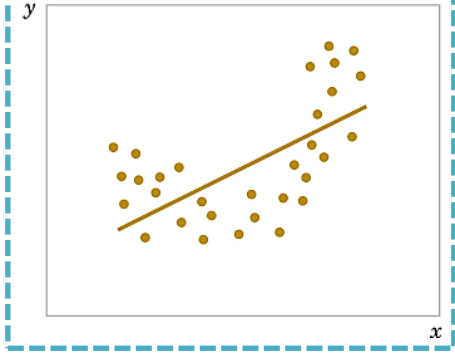
El mejor estimador para cualquier x es $\hat{y} = \bar{y}$



Rechazar H_0: x sí tiene valor para explicar la variabilidad de y.



Se recomienda agregar términos polinomiales



Regresión lineal simple

Utilidad (significancia) del modelo - Pruebas de hipótesis **especial** para $\boldsymbol{\beta_1}$

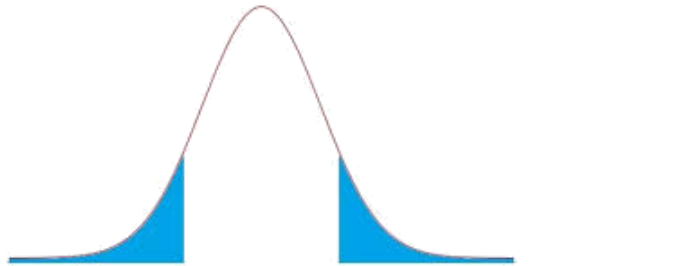
Prueba de utilidad para el ejemplo de inversión en I+D y ganancias de la empresa

Valor dado por H_0

- Estadístico de prueba:

  Ayudida → $S_{xx} = 50$

- Región de rechazo:



$-t_{0.025,4} = -2.776$ $\quad$ $t_{0.025,4} = 2.776$

Regresión lineal simple

Prueba de utilidad con – Ejemplo ganancias en R

```
> summary(modelo)
```

Comando en R para generar resumen del modelo

```
Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)  20.0000 ←  2.6458     7.559   0.00164 **
inv           2.0000 ←  0.4583     4.364   0.01202 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.24 on 4 degrees of freedom
Multiple R-squared:  0.8264,    Adjusted R-squared:  0.7831
F-statistic: 19.05 on 1 and 4 DF,  p-value: 0.01202
```

$\widehat{\beta}_0$

$\widehat{\beta}_1$

Valor P de la prueba de utilidad, usando **estadístico t para β_1**

¡Valores P equivalentes!

Regresión lineal simple

Sumas cuadráticas y análisis de varianza - ANOVA

Variabilidad total de la variable *y* (SST)

Explicada por la línea de regresión (SSR)	Explicada por el error (SSE)

La identidad fundamental del **análisis de varianza** para un modelo de regresión:

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

SST = SSR + SSE

Para efectos de cálculo de las sumas de cuadrados se utilizan las siguientes expresiones:

$$\boldsymbol{SST}=\sum_{i=1}^{n}y_i^2-n\bar{y}^2=\frac{S_{yy}}{n-1}\quad \boldsymbol{SSR}=\hat{\beta}_1 S_{xy}\quad \text{y}\quad \boldsymbol{SSE}=SST-SSR$$

Regresión lineal simple

Sumas cuadráticas y análisis de varianza - ANOVA

Cada una de las sumas de cuadrados tiene asociado un parámetro llamado **grados de libertad (gl)**, que define el número de observaciones independientes disponibles en la suma, y que están dados por:

$$\boldsymbol{SST} = \boldsymbol{SSR} + \boldsymbol{SSE}$$

$$(n-1) = 1 + (n-2)$$

Con base a lo anterior se construyen estimaciones independientes para la **varianza explicada por cada componente (modelo y error)**, usando la respectiva suma de cuadrados dividida por sus grados de libertad, así:

$$\boldsymbol{MSR} = SSR/1 \quad \text{y} \quad \boldsymbol{MSE} = SSE/n-2$$

Al analizar la razón entre el $\boldsymbol{MSR}$ y el $\boldsymbol{MSE}$ estamos comparando varianzas…a este ejercicio le llamaremos:

Análisis de varianza → Método **alternativo** para realizar la prueba de utilidad (significancia) del MRLS.

Regresión lineal simple

Prueba de utilidad (significancia) utilizando ANOVA

¿Cómo evaluamos la relación entre el MSR y el MSE?

Aquí aparece el **estadístico F**:

$$\boldsymbol{F_0} = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/n-2} \sim F_{1,n-2}$$

Este también puede utilizarse para evaluar la hipótesis de la **prueba de utilidad:**

$$\begin{cases} \boldsymbol{H_0}: \beta_1 = 0 \\ \boldsymbol{H_a}: \beta_1 \neq 0 \end{cases}$$

¿Cómo interpretamos el estadístico F y sacamos una conclusión?

1. Definir nivel de significancia de $\boldsymbol{\alpha}$ (0,05 por ejemplo…el más común)

2a. Hallar $\boldsymbol{F_{\alpha/2,1,n-2}}$ (tablas) y se **rechaza $\boldsymbol{H_0}$** si $\boldsymbol{F_0 > F_{\alpha/2,1,n-2}}$ o…

2b. Obtener $\boldsymbol{Valor\ p}$ (herramienta computacional, R o Excel) y se **rechaza $\boldsymbol{H_0}$** si $\boldsymbol{Valor\ p < \alpha}$

¡Siempre nos salva!

Regresión lineal simple

Prueba de utilidad utilizando ANOVA

Toda la información de la prueba de utilidad (significancia) de la regresión se puede resumir en una tabla conocida como tabla ANOVA:

Fuente de variación	SS	gl	MS	F	Valor P
Regresión	$SSR = \hat{\beta}_1 S_{xy}$	1	$MSR = SSR/1$		
Error	$SSE = SST - \hat{\beta}_1 S_{xy}$	$n-2$	$MSE = SSE/n-2$	$F_0 = \frac{MSR}{MSE}$	Clave
Total	SST	$n-1$			

Nota: Ambas formas para probar la utilidad (significancia) de la regresión son equivalentes y el valor P de las dos pruebas es el mismo.

Regresión lineal simple

Prueba de utilidad con – Ejemplo ganancias en R

```
> anova(modelo)
Analysis of Variance Table

Response: gan
          Df Sum Sq Mean Sq F value  Pr(>F)
inv        1    200   200.0  19.048 0.01202 *
Residuals  4     42    10.5
---
Signif. codes:  0 ‘***’ ‘**’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Comando en R para generar ANOVA del modelo

MS

SSR (200)

SSE (42)

F_0 (19.048)

```
> summary(modelo)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.0000     2.6458   7.559  0.00164 **
inv           2.0000     0.4583   4.364  0.01202 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.24 on 4 degrees of freedom
Multiple R-squared:  0.8264,    Adjusted R-squared:  0.7831
F-statistic: 19.05 on 1 and 4 DF,  p-value: 0.01202
```

$\hat{\beta}_1$ (2.0000)

¡Valores P equivalentes!

Regresión lineal simple

Coeficiente de determinación

Partiendo de las sumas de cuadrados, podríamos obtener un coeficiente que nos indique:

Qué fracción o proporción de la variabilidad de $\boldsymbol{y}$ es explicada por el modelo ajustado

¿Cómo?

Evaluando la razón → $\frac{SSR}{SST} = \boldsymbol{R^2}$

A esta medida se le conoce como **COEFICIENTE DE DETERMINACIÓN** y me indica qué tan bueno es el ajuste realizado.

El coeficiente de determinación se expresa siempre como $\boldsymbol{r^2}$ o $\boldsymbol{R^2}$ y toma valores entre 0 y 1.

Truquito: en el caso de RLS el coeficiente de determinación se puede obtener al elevar al cuadrado el coeficiente de correlación.

Regresión lineal simple

Coeficiente de determinación - Ejemplo ganancias en R

```
> summary(modelo)
```

Comando en R para generar resumen del modelo

```
Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)  20.0000        2.6458   7.559  0.00164 **
inv           2.0000        0.4583   4.364  0.01202 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.24 on 4 degrees of freedom
Multiple R-squared:  0.8264,    Adjusted R-squared:  0.7831
F-statistic: 19.05 on 1 and 4 DF,  p-value: 0.01202
```

Regresión lineal simple
Preguntas de interpretación

1 ¿Cuál de las siguientes es una interpretación correcta del R^2 obtenido para el modelo ajustado a las ganancias de la empresa como función de la inversión en I+D?

a) El 82,64% de las veces la inversión en I+D predice correctamente las ganancias esperadas.

b) El 17,36% de la variabilidad de las ganancias es explicada por el modelo.

c) El 82,64% de la variabilidad de la inversión en I+D es explicada por el modelo.

d) El 82,64% de la variabilidad de las ganacias es explicada por el modelo.

Regresión lineal simple
Preguntas de interpretación

2 Si para otro caso, se tiene un coeficiente de correlación de -0.75 ¿Cuál será el valor de R^2?

3 Se tiene el siguiente gráfico de dispersión, el valor de R^2 para la relación descrita es de 92,16% ¿Cuál será entonces el coeficiente de correlación para este caso?

Fuente: https://www.coursera.org/learn/linear-regression-model

MUCHAS GRACIAS

Natalia Jaramillo Quiceno

e-mail: natalia.jaramilloq@upb.edu.co