Estadística

Clase 7 - Regresión Lineal Simple

Rodrigo Del Rosso
RDelRosso-ext@austral.edu.ar

28 de Mayo de 2022

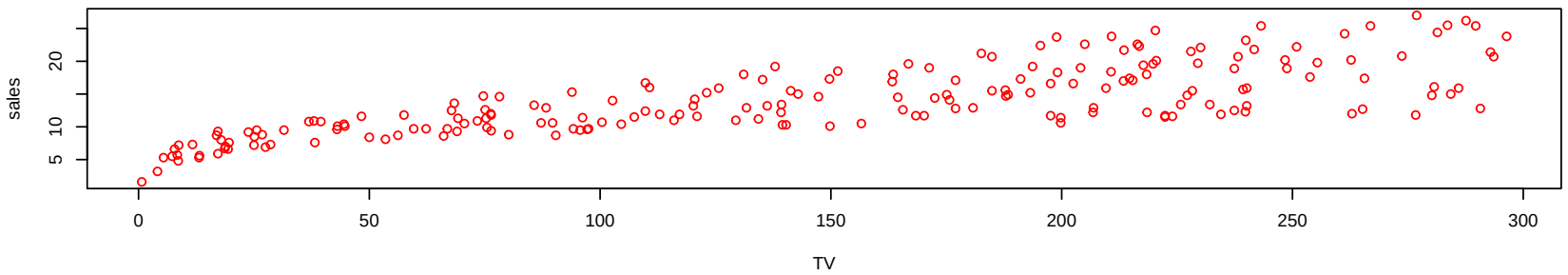
Introducción

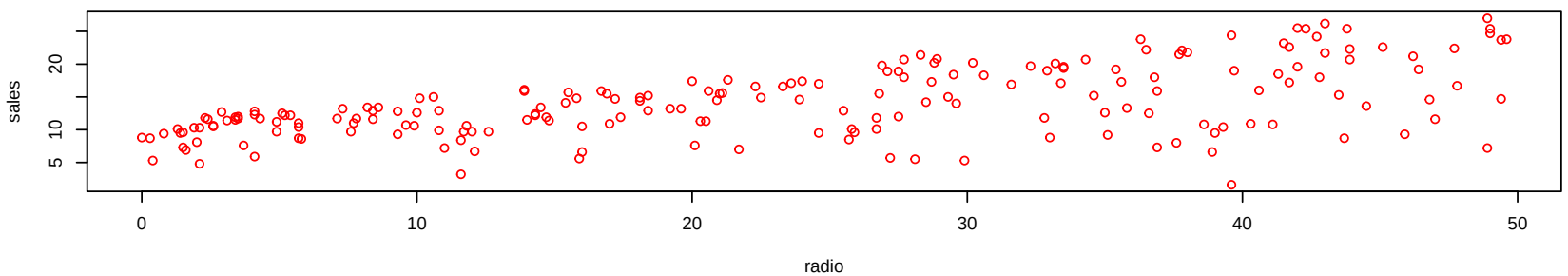
Los métodos estadísticos estudiados anteriormente son utilizados para tener información acerca de una única variable observada a cada unidad experimental.

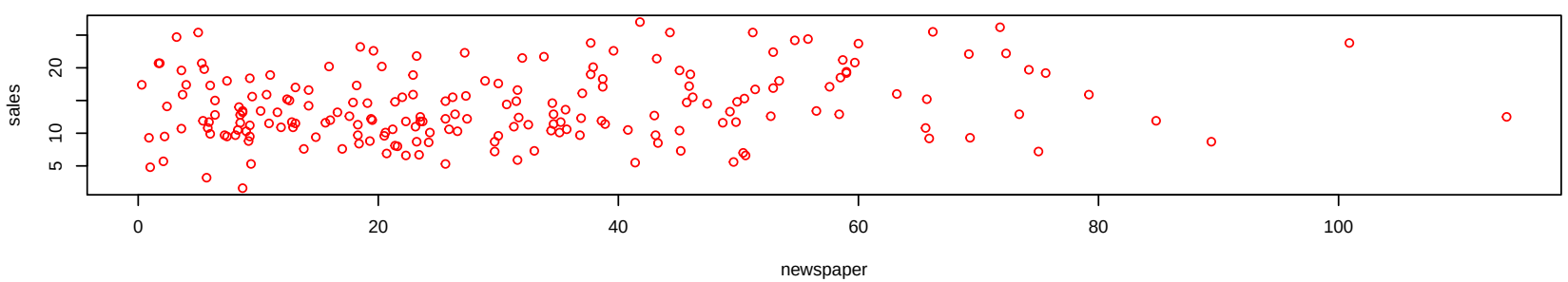
A esta única variable, se le examinaron varias medidas que describen su comportamiento y se aplicaron diversas técnicas de inferencia estadística, tales como **intervalos de confianza** y **pruebas de hipótesis**, para hacer estimaciones y sacar conclusiones acerca de ella.

En esta parte de la materia se estudiarán métodos que permiten establecer las posibles relaciones existentes entre dos o más variables cuantitativas observadas a cada unidad experimental.

Motivación


sales
5
10
20
0
50
100
150
200
250
300
TV


sales
5
10
20
0
10
20
30
40
50
radio


sales
5
10
20
0
20
40
60
80
100
newspaper

Motivación

Los **presupuestos publicitarios** son variables de **entrada**, mientras que las **Ventas** (Sales) son una variable de **salida**.

Las variables de entrada generalmente se denotan usando las últimas letras del abecedario X, con un subíndice para distinguirlas.

Entonces X_1 podría ser el presupuesto de TV, X_2 el presupuesto de radio y X_3 el presupuesto del periódico.

Las entradas tienen diferentes nombres,

- Predictores
- Variables independientes
- Features (características)
- Variable Explicativa del predictor
- Regresores

La variable de salida (en este caso, ventas) es una variable que a menudo se llama respuesta, dependiente o explicada, y generalmente se denota como variable dependiente de respuesta utilizando la letra Y.

A lo largo de las clases se utilizarán estos términos de manera indistinta.

Definiciones

Se denomina **Análisis de Regresión** a un método estadístico que permite explicar el comportamiento de una variable cuantitativa, a partir del comportamiento de otra u otras variables que puedan estar relacionadas, estableciendo la expresión funcional del modelo matemático que describa dicho comportamiento.

Se denomina **Variable Explicada** a aquella variable cuantitativa cuyo comportamiento se desea describir a partir del comportamiento de otra u otras variables.

Se denomina **Variables Explicativas** a aquellas variables que explican el comportamiento de la variable explicada.

El Análisis de Regresión consiste en construir un modelo que permita predecir el valor de la **variable explicada** mediante los valores de k **variables explicativas**.

Análisis de Regresión

Este modelo consta de dos partes bien diferenciadas,

- Una función real o modelo matemático
- Una variable aleatoria, que representa la variable no controlada por las k **variables explicativas**. En términos formales,

$$Y_i = f(X_{1i}, X_{2i}, ..., X_{ki}) + \epsilon_i$$

Esta expresión representa la explicación del fenómeno Y en función de una cantidad de regresores y un término aleatorio. El subindice i representa la i-ésima observación de la variable. Al primer sumando se lo denomina **Función de Regresión** y representa el modelo matemático que interviene en el modelo estadístico de regresión (Modelo de Regresión Poblacional).

El segundo sumando se denomina **Perturbación Aleatoria**, Error Aleatoria, Cantidad Aleatoria, Shock Aleatoria, **Ruido Blanco** entre otras formas. Este término aleatorio representa los errores de observaciones producto de todos aquellos factores que influyen en la variable respuesta Y que no fueron considerados en la construcción del modelo.

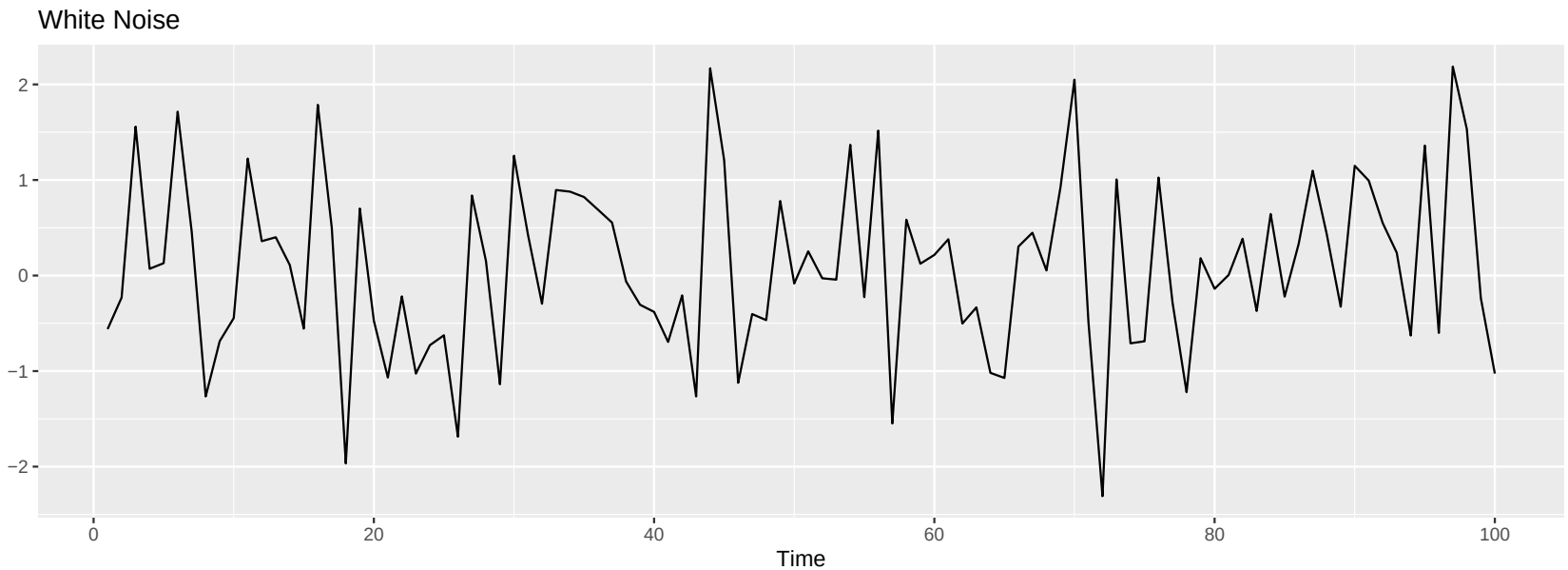
Supuestos de Gauss-Markov

Para realizar el análisis inferencial al Modelo Estadístico de Regresión deben cumplirse con los siguientes supuestos, los cuales deben ser testeados (en esta materia se considerarán que se cumplen siempre... esto lo verán en Econometría).

1. $E(\epsilon_i) = 0$
2. $Var(\epsilon_i) = \sigma_\epsilon^2$ (Homocedasticidad)
3. $Cov(\epsilon_i; \epsilon_j) = 0 \quad \forall i \neq j$ (Incorrelación)
4. $Cov(\epsilon; X_j) = 0 \quad \forall\, j = 1, 2, ..., k$
5. $\epsilon \sim N$ (Normalidad)

Si ϵ cumple con los primeros tres supuestos se dice que es un **Ruido Blanco** (White Noise) y si se le adiciona el último supuesto se denomina **Ruido Blanco Normal**. En la siguiente sección se exhibe un gráfico del mismo.

White Noise



Regresión Lineal Simple

La función de Regresión que capta la relación entre los regresores y la variable explicada se supone que es lineal y que únicamente un solo regresor X explica el comportamiento de Y. En términos formales,

$$f(X_{1i}, X_{2i}, ..., X_{ki}) = f(X_{1i}) = \beta_0 + \beta_1 X_{1i}$$

A esta recta se denomina **Recta de Regresión**. Al coeficiente β_1 se denomina **Coeficiente de Regresión** y al coeficiente β_0 es el intercepto de la recta.

Por lo tanto Y se puede reexpresar de la forma siguiente,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

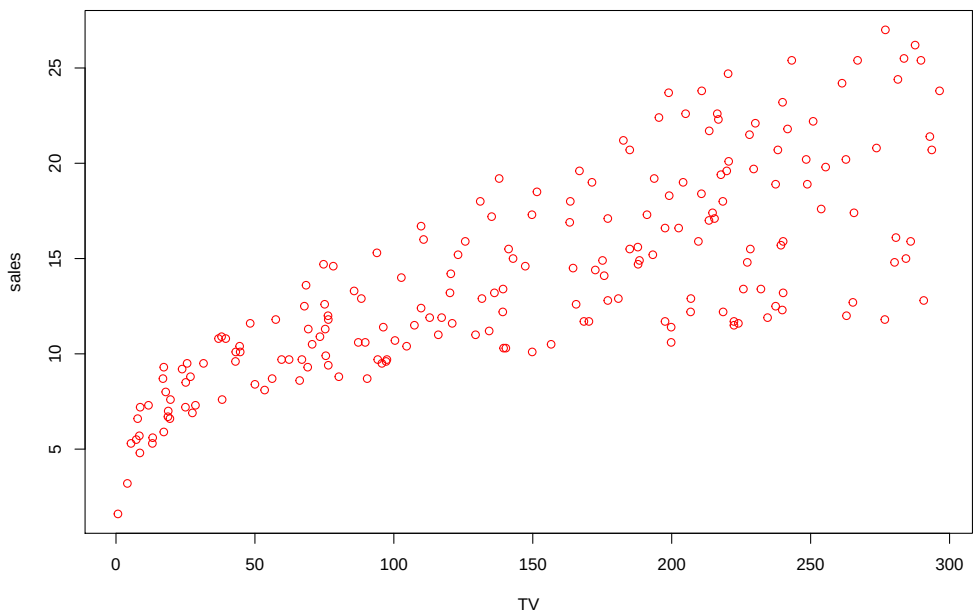
Dado que en el curso se verá la utilización de un solo regresor, sin pérdida de generalidad se reexpresa de la forma siguiente,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Regresión Lineal Simple

Los valores individuales de cada una de las variables que se miden se presentan como pares ordenados (x_i, y_i).

Cada par ordenado representa un punto en un plano. Todos los posibles pares se pueden presentar en un gráfico mediante las coordenadas cartesianas ortogonales, al cual se denomina **Diagrama de Puntos (Dispersión)** a la representación gráfica de los pares ordenados de los valores de las variables que intervienen en el análisis de regresión lineal simple.



Regresión Lineal Simple

La siguiente ecuación representa el **Modelo de Regresión Lineal Simple** polinómico de primer grado, que se forma cuando la función de regresión es una recta. En términos formales,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Donde

- Y es la Variable Explicada
- X es la Variable Explicativa
- β_0 es la Ordenada al Origen, que representa el valor de Y cuando X es igual a cero.
- β_1 es el Coeficiente de Regresión, que representa la variación promedio de Y cuando X varía en una unidad.
- ϵ es el término aleatorio denominado "Ruido Blanco"

Esta última medida se suele asociar con el concepto de **Elasticidad** muy utilizado en Economía o el concepto de **Pendiente** en Análisis Matemático.

Regresión Lineal Simple

De cumplirse los supuestos referidos a la Esperanza Matemática y a la Varianza del Ruido Blanco, entonces

- El valor esperado de Y para cada valor del Regresor X (Esperanza Matemática Condicionada) es la **Recta de Regresión**. En términos formales,

$$E\left(Y_i/X_i\right) = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad \forall\ i$$

- La Varianza de Y para cada valor del Regresor X (Varianza Condicionada) es igual a la Varianza del Ruido Blanco. En términos formales,

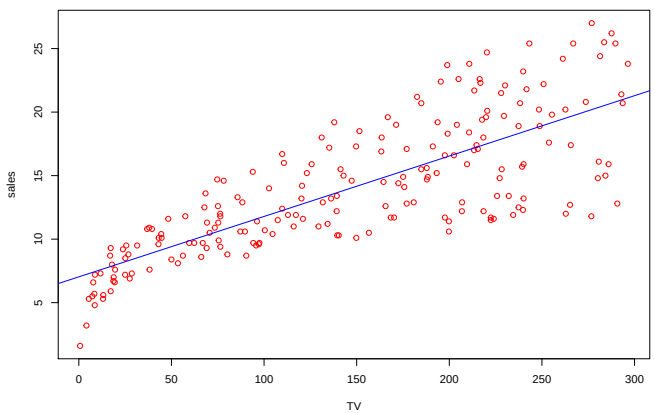
$$Var\left(Y_i/X_i\right) = \sigma_\epsilon^2 \qquad \forall\ i$$

- La diferencia entre el valor real observador de Y_i y el estimado por el modelo de regresión $\hat{Y}_i$ se denomina **Residuo**. En términos formales,

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

Regresión Lineal Simple

Si cada valor de la Recta de Regresión es el valor esperado de Y condicionado al regresor X, entonces dicha recta pasa entre los puntos del **Diagrama de Puntos** compensando los desvíos que se producen. El siguiente gráfico exhibe lo mencionado,



$$Sales_i = \beta_0 + \beta_1 TV_i + \epsilon_i$$

$$\hat{Sales}_i = \hat{\beta}_0 + \hat{\beta}_1 TV_i$$

$$\hat{\epsilon}_i = Sales_i - \hat{Sales}_i$$

Aplicaciones Económicas

- Función de Consumo Keynesiana (Consumo - Ingreso)
  - Y = Consumo
  - X = Ingreso
  - β_0 = Consumo Autónomo
  - β_1 = Propensión Marginal a Consumir (PMgC)

Donde la PMgC es la cantidad que se destina al Consumo por cada peso adicional del Ingreso.

- Función de Costo Total
  - Y = Costo Total
  - X = Unidades Producidas
  - β_0 = Costo Fijo
  - β_1 = Costo Medio Variable

Donde el Costo Medio Variable representa el incremento del Costo Total por cada unidad que se produce.

Aplicaciones Económicas

¿Qué otras aplicaciones se les ocurre?

Parámetros

Tal como hemos visto la recta de regresión lineal poblacional es,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- β_0 es la Ordenada al Origen, que representa el valor de Y cuando X es igual a cero.
- β_1 es el Coeficiente de Regresión, que representa la variación promedio de Y cuando X varía en una unidad.

β_1 se calcula como el cociente entre la Covarianza entre las dos variables y la Varianza del Regresor. En términos formales,

$$\beta_1 = \frac{Cov(X, Y)}{Var(X)} = \frac{\sigma_{XY}}{\sigma_X^2}$$

La ordenada al origen de la Recta de Regresión, o sea el parámetro β_0 es,

$$\beta_0 = \mu_Y - \beta_1 \mu_X$$

Uno de los supuestos implícitos es que los parámetros son desconocidos y deben ser inferidos mediante los datos provistos por una muestra.

Estimadores de los Parámetros del Modelo

Para estimar los parámetros se toma una muestra al azar de n unidades experimentales. Esta muestra proporciona n pares de valores (x_i, y_i). Estos valores, para una mejor interpretación se disponen en una tabla de las siguientes características,

X	Y
x_1	y_1
x_2	y_2
...	...
x_n	y_n

A la representación de estos datos en un momento del tiempo se denomina **Datos de Corte Transversal** y se trata de explicar el comportamiento de la variable Y a partir de una muestra de valores de X e Y en un momento determinado del tiempo. En materias posteriores verán que los datos pueden provenir de,

- Datos de Corte Transversal, Series Temporales o de Panel

Estimadores de los Parámetros del Modelo

Para obtener los Estimadores de los Parámetros del Modelo de Regresión Lineal Simple se procede a emplear el Método de Estimación por Mínimos Cuadrados Ordinarios (MCO) (Ordinary Least Squares - OLS). Este método es uno de los tantos métodos de estimación de parámetros que existen. MCO es un método numérico que se utiliza para ajustar unos puntos a una función de pérdida determinada.

Se guiará la búsqueda de los parámetros buscando minimizar la suma de los residuos al cuadrado (Suma de Cuadrados Residual) que será la función de pérdida en este modelo.

Se define la función de pérdida del problema en cuestión de la forma siguiente,

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

El método de MCO trata de minimizar la expresión anterior, de forma tal de encontrar los valores de β_0 y β_1 que hagan mínima esa expresión.

Estimadores de los Parámetros del Modelo

Se demuestra que la expresión anterior es una función de dos variables, en este caso de los parámetros del modelo de regresión lineal simple. En términos formales,

$$f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

Para obtener los estimadores se procede a minimizar dicha expresión, y al ser una función de dos variables se debe proceder de la forma siguiente,

$$\frac{\partial f}{\partial \hat{\beta}_0} = 0 \qquad \frac{\partial f}{\partial \hat{\beta}_1} = 0$$

La expresión anterior representa la CPO (Condición de Primer Orden), denominada Condición Necesaria para que exista un óptimo. De esta expresión salen puntos críticos que deben ser evaluados en la derivada segunda de la función (o que el Hessiano sea Positivo . . . lo verán más adelante - no se preocupen -).

Estimadores de los Parámetros del Modelo

De la derivación de la función respecto a cada uno de los parámetros, luego de haber verificado que son óptimos y mínimos de la expresión se llegan a las expresiones funcionales de los estimadores:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

A partir de los datos muestrales se calcula la estimación puntual de cada uno de estos parámetros poblacionales, mediante el reemplazo en las fórmulas anteriores. Si se distribuyen los términos en la fórmula del Coeficiente de Regresión se llega a la siguiente expresión,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

Distribución de los Estimadores

Es necesario conocer el modelo probabilístico que explique el comportamiento de estas variables aleatorias, es decir hay que determinar la **Distribución de Probabilidad** de los **Estimadores**.

Se demuestra bajo ciertas condiciones que si el Ruido Blanco es Normal, entonces los estimadores se distribuyen con igual distribución (esto habría que testearlo mediante una Prueba de Hipótesis que verán en otra materia) con los siguientes parámetros,

$$\hat{\beta}_0 \sim N\left(\beta_0; \sqrt{\sigma_\epsilon^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right)}\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1; \sqrt{\frac{\sigma_\epsilon^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}}\right)$$

Se demuestra que ambos estimadores son **insesgados**, además de cumplir con otras propiedades deseables de los estimadores, se los conoce como estimadores MELI (Mejores Estimadores Linealmente Insesgados).

Distribución de los Estimadores

El problema de las expresiones anteriores es que **se desconoce la varianza del error poblacional** σ_ϵ^2, por lo tanto se infieren a partir de los residuos que surgen de la estimación del modelo lineal.

El estimador insesgado de la varianza del error poblacional se determina de la forma siguiente,

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-2} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-2}$$

Es decir la varianza del error poblacional se estima mediante un cociente entre la suma de cuadrados residual y los grados de libertad (**n-2**), que representa la cantidad de observaciones menos la cantidad de parámetros estimados en el modelo.

A partir de la distribución de probabilidad de los estimadores es posible realizar Pruebas de Hipótesis acerca del valor poblacional de los parámetros o construir Intervalos de Confianza para mejorar la precisión de las estimaciones puntuales de los mismos.

Test de Significancia Individual

En la construcción de modelos econométricos es de interés fundamental evaluar si los coeficientes estimados del modelo son significativos o no. Para tal fin es necesario poner a prueba ciertas conjeturas acerca de los parámetros del mismo.

Una de las conjeturas que se posee es que los coeficientes del modelo a nivel poblacional son significativos, es decir distinto a cero. En la clase siguiente se analizará que este planteo tiene fuerte vinculación con la **correlación lineal** entre las variables incluidas en el modelo. Las hipótesis que ponen a prueba son,

$$H_0 : \beta_0 = 0 \; H_1 : \beta_0 \neq 0$$

$$H_0 : \beta_1 = 0 \; H_1 : \beta_1 \neq 0$$

Test de Significancia Individual

El planteo de las hipótesis sobre el Coeficiente de Regresión son,

$$H_0 : \beta_1 = 0 \; H_1 : \beta_1 \neq 0$$

El estadístico que se **debería** emplear es,

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim N(0;1)$$

El problema con la fórmula anterior es que se desconoce la varianza del error poblacional, por lo tanto se empleará **siempre** el siguiente estadístico,

$$t^e = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t_{n-2}$$

Los siguientes pasos del Test de Hipótesis son los mismos pasos que se han visto en clases anteriores.

Test de Significancia Individual

El planteo de las hipótesis sobre el Intercepto del modelo son,

$$H_0 : \beta_0 = 0 \; H_1 : \beta_0 \neq 0$$

El estadístico que se **debería** emplear es,

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}} \sim N(0; 1)$$

El problema con la fórmula anterior es que se desconoce la varianza del error poblacional, por lo tanto se empleará **siempre** el siguiente estadístico,

$$t^e = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}_\varepsilon^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}} \sim t_{n-2}$$

Los siguientes pasos del Test de Hipótesis son los mismos pasos que se han visto en clases anteriores.

Intervalos de Confianza

Los Intervalos de Confianza (IC) se construyen para estimar, con un nivel de confianza igual a $1-\alpha$, tanto los parámetros de la Recta de Regresión como los valores de la Recta Poblacional para un valor dado de la variable explicativa X.

Todos estos intervalos son de tipo aditivo, es decir a la estimación puntual se le suma y resta el error de estimación.

El IC para el intercepto del modelo lineal es,

$$\hat{\beta}_0 \pm \ t_{1-\frac{\alpha}{2};n-2}\sqrt{\hat{\sigma}^2_\varepsilon\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\right)}$$

El IC para el Coeficiente de Regresión del modelo lineal es,

$$\hat{\beta}_1 \pm \ t_{1-\frac{\alpha}{2};n-2}\sqrt{\frac{\hat{\sigma}^2_\varepsilon}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}}$$

Intervalos de Confianza

Para estimar la Recta de Regresión Poblacional se procede a obtener la distribución de probabilidad y parámetros resultantes de la misma. Dado que ambos estimadores se distribuyen en forma normal, entonces la suma de variables normales por definición es una normal. Por lo tanto se puede demostrar lo siguiente,

La recta de regresión estimada para un valor X_0 dado de X es,

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

La distribución resultante es,

$$\hat{Y}_0 \sim N\left(Y_0; \sqrt{\sigma_\varepsilon^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}\right)$$

El IC para dicha estimación puntual es,

$$\hat{Y}_0 \pm \ t_{1-\frac{\alpha}{2};n-2}\sqrt{\hat{\sigma}_\varepsilon^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

Bondad del Modelo

La variabilidad total de Y se puede descomponer como la suma de dos variabilidades: una que se encuentra captada por el modelo (Variabilidad Explicada) y otra que no es captada por el mismo (Variabilidad Residual).

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

MCO encontró aquellos parámetros del modelo que mejor se ajusten a partir de minimizar la segundad variabilidad, "pedir" que la variabilidad residual sea la más pequeña es lo mismo a maximizar la variabilidad explicada por el modelo.

$$SCT = SCE + SCR$$

Donde,

- $SCT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- $SCE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$
- $SCR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

Bondad del Modelo

De estas medidas surge una métrica que será importante a esta altura de la materia, que se denomina **Coeficiente de Determinación** o simplemente R^2.En términos formales,

$$R^2 = \frac{SCE}{SCT}$$

Es decir, de la variabilidad total de la variable que se intenta explicar, que proporción esta captada por el modelo econométrico.

Un $R^2 = 1$ indica que la variabilidad total es captada completamente por el modelo.

Un $R^2 = 0$ indica que la variabilidad residual coincide con la variabilidad total y por lo tanto el modelo es malo para explicar el comportamiento de la variable Y.

En general esta medida se encuentra entre ambos extremos, es decir,

$$0 < R^2 < 1$$

Bondad del Modelo

La variabilidad explicada (SCE) se calcula de la forma siguiente,

$$SCE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 = \hat{\beta}_1^2 \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right)$$

A partir de esta expresión se reemplaza en la fórmula de R^2 y se obtiene lo siguiente,

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = \hat{\beta}_1^2 \left(\frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2} \right)$$

Aplicaciones

Table 1:

	Variable Dependiente:		
	sales		
	(1)	(2)	(3)
TV	0.048***		
	(0.003)		
radio		0.202***	
		(0.020)	
newspaper			0.055***
			(0.017)
Constant	7.033***	9.312***	12.351***
	(0.458)	(0.563)	(0.621)
Observations	200	200	200
R^2	0.612	0.332	0.052
Adjusted R^2	0.610	0.329	0.047
Residual Std. Error (df = 198)	3.259	4.275	5.092
F Statistic (df = 1; 198)	312.145***	98.422***	10.887***
Note:			*p<0.1; **p<0.05; ***p<0.01

Aplicaciones

Table 2:

	Variable Dependiente:	
	sales	
	(4)	(5)
TV	0.046^{***}	0.046^{***}
	(0.001)	(0.001)
radio	0.189^{***}	0.188^{***}
	(0.009)	(0.008)
newspaper	-0.001	
	(0.006)	
Constant	2.939^{***}	2.921^{***}
	(0.312)	(0.294)
Observations	200	200
R^2	0.897	0.897
Adjusted R^2	0.896	0.896
Residual Std. Error	1.686 (df = 196)	1.681 (df = 197)
F Statistic	570.271^{***} (df = 3; 196)	859.618^{***} (df = 2; 197)
Note:		*p<0.1; **p<0.05; ***p<0.01

Bondad del Modelo

Bondad del Modelo

Si los dos modelos son lineales, y ambos utilizan una variable explicativa.

A la hora de comparar, selecciono aquel que tenga un R^2 mayor, o sea que es buena medida comparativa.

Si en cambio el modelo lineal se construye agregando modularmente los distintos regresores, que es el caso más habitual en la práctica.

Se puede demostrar que siempre que se agregue un regresor al modelo (Ver Green), se verificará, como una consecuencia de los cuadrados mínimos,

$$R^2_{y/xz} \geq R_{y/x}$$

incluso en el caso en que dicho regresor no contribuya a explicar el comportamiento de y.

Bondad del Modelo

El inconveniente de agregar al modelo un regresor que no contribuya a la explicación, radica en que se pierde la condición de parsimonia, según la cual, ante dos modelos de igual calidad, siempre debe elegirse el más simple, o sea el que tenga menor cantidad de regresores.

A medida que se agregan regresores, el coeficiente de determinación global siempre crece, como consecuencia de los cuadrados mínimos; lo que interesa en un caso particular, es ver si R^2 crece por sobre lo que aportan los cuadrados mínimos, o por debajo, de modo que pueda determinarse si el regresor adicional contribuye a explicar a y.

Las medidas que permiten detectar este tipo de calidad, se denominan **medidas con término de penalización**.

Es decir: se agrega el regresor, y se calcula R^2. Si el incremento en R^2 es mayor que el término de penalización, se justifica el agregado del regresor adicional.

Bondad del Modelo

". . . es una medida de la contribución relativa del modelo lineal estimado por el método de mínimos cuadrados. Relativa a un modelo Naive consistente en la media muestral de la variable de interés (...)" (Sosa Escudero, 2015, p.30)

"El R^2 es una medida de calidad en relación con la pregunta que uno se hizo inicialmente, es decir, el R^2 no juzga la respuesta ni la pregunta, sino la adecuación de la respuesta a la pregunta (...)" (Sosa Escudero, 2015, p.31)

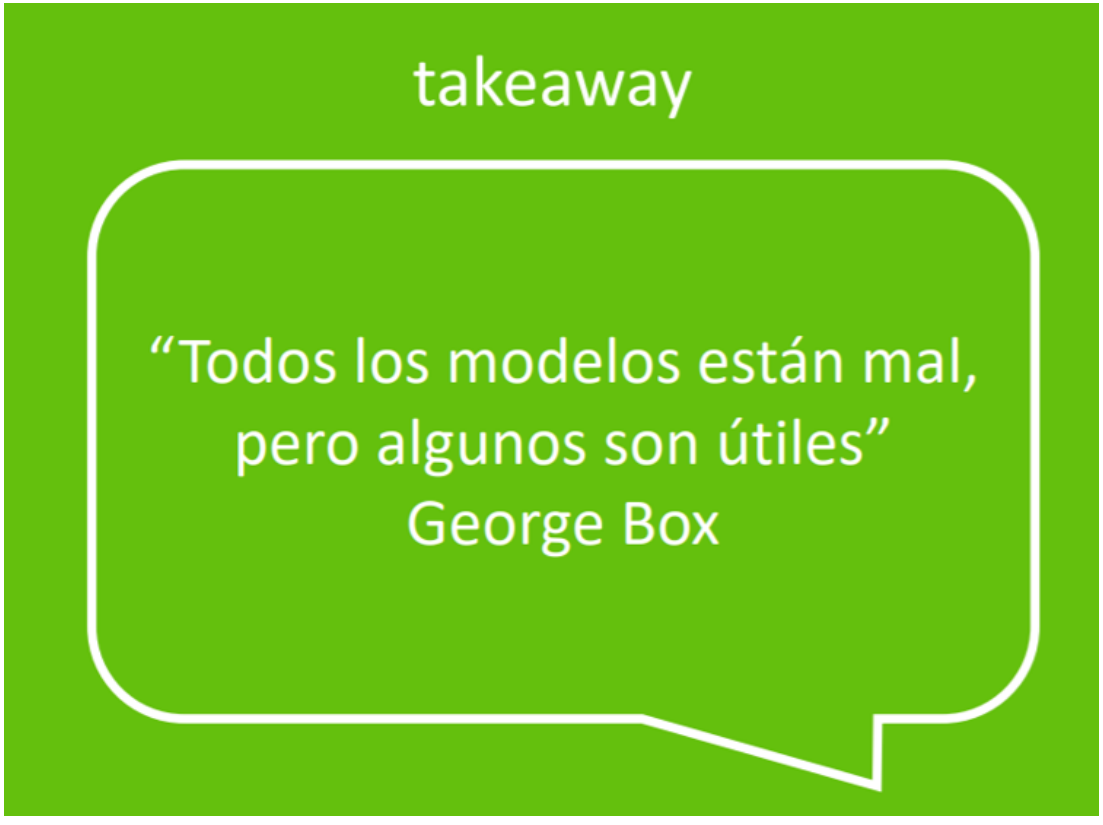
Bondad del Modelo

"Comparar modelos nada más que sobre la base del R^2 es como comparar coches sobre la base de su tamaño. Sin otra mención en particular, creer que un modelo es mejor que otro porque tiene R^2 más alto es como creer que un desvencijado ómnibus es mejor que un Porsche solo porque es más grande (...)" (Sosa Escudero, 2015, p.30)

"La enorme popularidad del R^2 tiene que ver con hacerles creer a los principiantes que se trata de "la" medida de calidad (...)" (Sosa Escudero, 2015, p.31)

Sosa Escudero, W. (2015). El Lado Oscuro de la Econometría. Editorial Temas. Primera Edición. Buenos Aires, Argentina.

Bondad del Modelo



Bondad del Modelo

El coeficiente de determinación ajustado $\bar{R}^2$ se utiliza en la **Regresión Múltiple** para ver el grado de intensidad o efectividad que tienen las variables independientes en explicar la variable dependiente.

El uso de este coeficiente se justifica en que a medida que añadimos variables a una regresión, el coeficiente de determinación sin ajustar tiende a aumentar. Incluso cuando la contribución marginal de cada una de las nuevas variables añadidas no tiene relevancia estadística.

Por lo tanto, al añadir variables al modelo, el coeficiente de determinación podría aumentar y podríamos pensar, de manera errónea, que el conjunto de variables elegido es capaz de explicar una mayor parte de la variación de la variable independiente. A este problema se le conoce comúnmente como **"sobreestimación del modelo"**.

Bondad del Modelo

Pertenece a la familia de medidas con términos de penalización ya que penaliza a medida que incorpora un regresor. Esta medida sirve para comparar distintos modelos.

$$\bar{R}^2 = R^2 - f(k)$$

Donde k representa la cantidad de regresores. Para $k = 1$ se obtiene la siguiente expresión,

$$\bar{R}^2 = 1 - \frac{\frac{SCR}{n-2}}{\frac{SCT}{n-1}} = 1 - \left(\frac{n-1}{n-2}\right)\frac{SCR}{SCT} = 1 - \left(\frac{n-1}{n-2}\right)\left(1 - R^2\right)$$

Para un valor genérico de k,

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(1 - R^2\right)$$

Bondad del Modelo

Teniendo en cuenta que $1 - R^2$ es un número constante y que n es mayor que k, a medida que añadimos variables al modelo, el cociente entre paréntesis se hace más grande. Consecuentemente. también el resultado de multiplicar este por $1 - R^2$.

Con lo cual se observa que la fórmula está construida para ajustar y penalizar la inclusión de coeficientes en el modelo.

Además de la ventaja anterior, el ajuste empleado en la fórmula anterior, permite también comparar modelos con distinto número de variables independientes. De nuevo, la fórmula ajusta el número de variables entre un modelo y otro y **permite realizar una comparación homogénea**.

Se puede deducir que el $\bar{R}^2$ será siempre igual o menor que el R^2. Al contrario que el R^2 que varía entre 0 y 1, el $\bar{R}^2$ podría ser negativo por 2 motivos:

1. Cuanto más se aproxime k a n.
2. Cuanto menor sea el R^2.

Análisis de Correlación

Análisis de Correlación

Es un método estádistico que permite **medir el grado de asociación entre las variables**.

El análisis de **correlación lineal simple** se lleva a cabo cuando la Función de Regresión que explica el comportamiento conjunto de las variables es una recta.

La intensidad de la relación lineal entre las variables se mide con el **Coeficiente de Correlación Lineal** ρ. Este coeficiente surge del cociente entre la Covarianza entre las variables y el producto de los Desvíos Estándares de cada una de ellas. En términos formales,

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

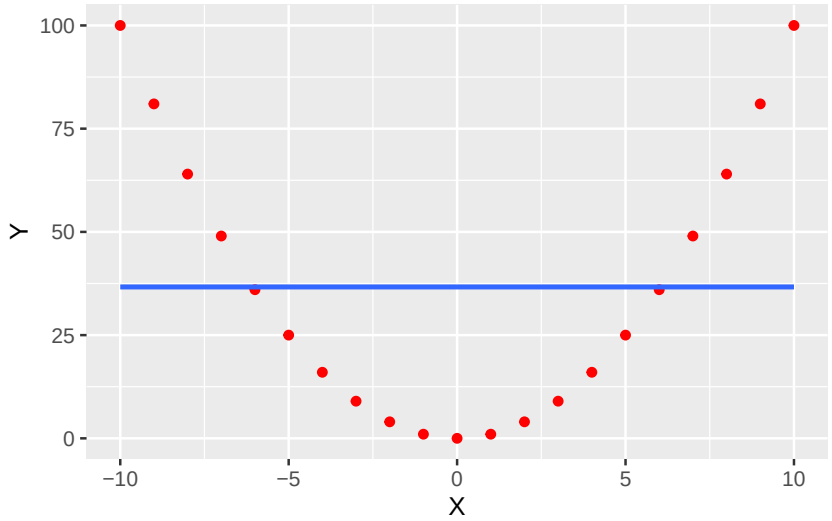
Este coeficiente necesariamente se encuentra entre -1 y 1. Es decir

$$-1 \leq \rho \leq 1$$

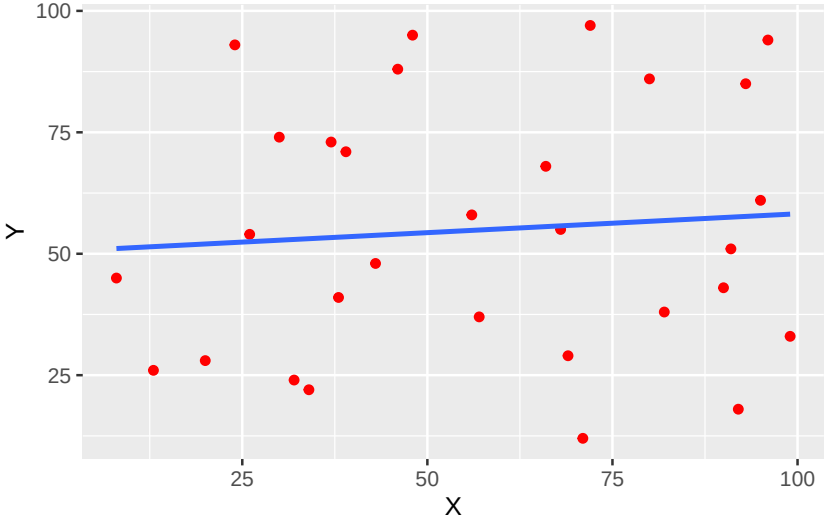
A continuación se ilustran tres tipos de diferentes de valores extremos de asociación entre variables.

Análisis de Correlación

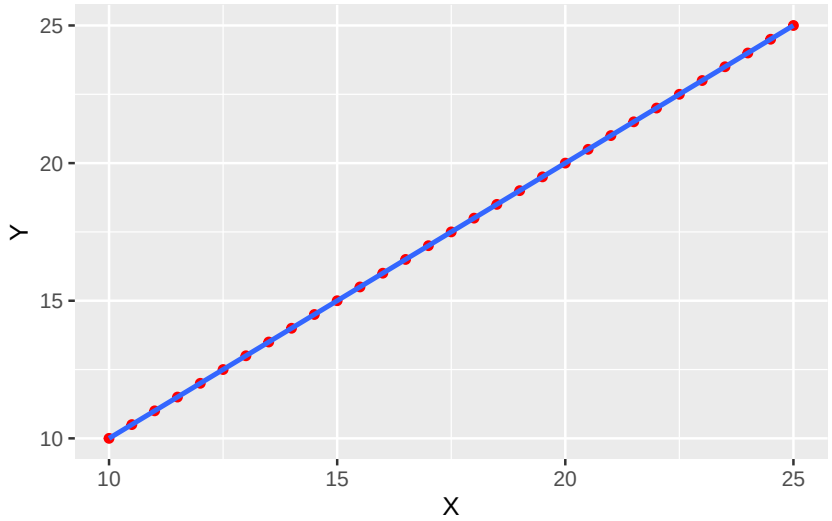
Eq 1: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \rho = 0$



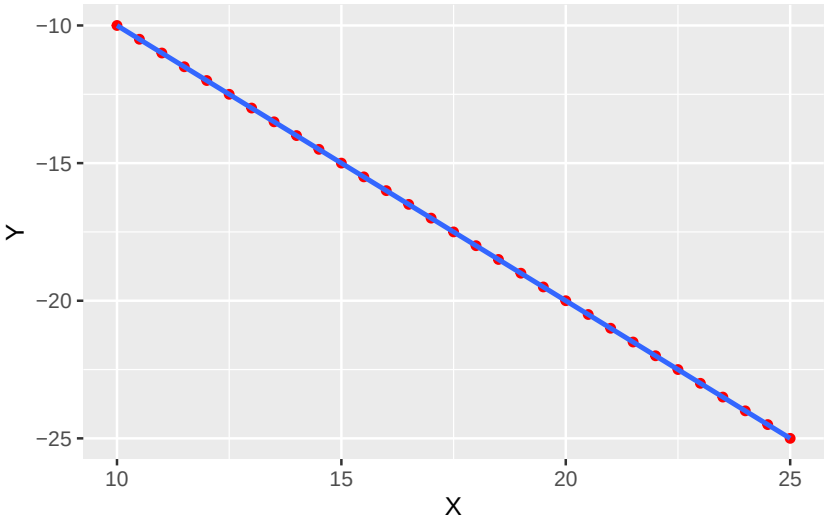
Eq 2: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \rho = 0.08348675$



Eq 3: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \rho = 1$



Eq 4: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \rho = -1$



Análisis de Correlación

El parámetro ρ, coeficiente de correlación lineal poblacional, generalmente es desconocido, por lo tanto, hay que estimarlo.

El coeficiente de correlación lineal poblacional $\hat{\rho}$ es el estimador del parámetro ρ. En términos formales,

$$\hat{\rho} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}}$$

Si $\rho = 0$ entonces se cumple que el estadígrafo

$$T^e = \frac{\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}} \sqrt{n-2} \sim t_{n-2} \ gl$$

tiene distribución t de Student con **n-2** grados de libertad (degrees of freedom).

Si $\rho \neq 0$ entonces para obtener una distribución que explique el comportamiento probabilístico de $\hat{\rho}$, hay que hacer una transformación logarítmica del estimador.

Análisis de Correlación

$$\mathfrak{z}(\hat{\rho}) = \frac{1}{2} ln \frac{(1+\hat{\rho})}{(1-\hat{\rho})}$$

Esta transformación tiene distribución asintóticamente Normal con los siguientes valores de parámetros,

$$E[\mathfrak{z}(\hat{\rho})] = \mathfrak{z}(\rho) = \frac{1}{2} ln \frac{(1+\rho)}{(1-\rho)}$$

$$Var[\mathfrak{z}(\hat{\rho})] = \frac{1}{n-3}$$

Entonces, la variable estandarizada tiene Distribución Normal,

$$\frac{\mathfrak{z}(\hat{\rho}) - \mathfrak{z}(\rho)}{\sqrt{\frac{1}{n-3}}} \sim N(0,1)$$

Pruebas de Significancia

Este test consiste en probar si ρ es significativo o no, es decir si bajo la evidencia proporcionada por la muestra existen razones suficientes para rechazar o no la hipótesis nula de que sea igual a cero. En términos formales,

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Si se rechaza la hipótesis nula, entonces el coeficiente de correlación lineal poblacional es significativo, es decir es distinto de cero bajo la evidencia que proporciona la muestra.

El estadígrafo de prueba que se utiliza para este constraste es,

$$T^e = \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\sqrt{n-2} \sim t_{n-2} \ gl$$

Esta es una prueba bilateral (a dos colas), donde los puntos críticos son,

$$t_c \ = \left| \pm \ t_{n-2;1-\frac{\alpha}{2}} \right|$$

Pruebas de Significancia

La regla de decisión que se utiliza para tomar la decisión estadística es,

$$Si|T^e| \geq t_c \Rightarrow \mathrm{RH}_0$$

Donde $t_{n-2;1-\frac{\alpha}{2}}$ es el Percentil de orden $1-\frac{\alpha}{2}$ de la Distribución t de Student con $n-2$ grados libertad. Es posible construir un punto crítico en los términos del coeficiente de correlación lineal muestral,

$$r_c = \left| \pm \sqrt{\frac{t_{n-2;1-\frac{\alpha}{2}}}{n-2+t_{n-2;1-\frac{\alpha}{2}}}} \right|$$

Si el módulo del coeficiente de correlación muestral mediante la muestra de tamaño n es mayor o igual al punto crítico, r_c, se rechaza H_0, esto quiere decir que el coeficiente de correlación poblacional es distinto de cero, o sea, **significativo**.

$$Si|r| \geq r_c \Rightarrow \mathrm{RH}_0 \Rightarrow \rho \neq 0$$

Los valores de r_c están tabulados para distintos valores, tanto de α, como de n.

Prueba Condicional a la Significacia del ρ

Si el ρ es significativo, puede ser necesario saber cuál o cuáles son los posibles valores que puede asumir. Para ello es necesario plantear **alguno** de las siguientes hipótesis,

- Prueba Bilateral (Contra Distinto)

$$H_0 : \rho = \rho_0$$

$$H_1 : \rho \neq \rho_0$$

- Prueba Unilateral (Contra Menor)

$$H_0 : \rho \geq \rho_0$$

$$H_1 : \rho < \rho_0$$

- Prueba Unilateral (Contra Mayor)

$$H_0 : \rho \leq \rho_0$$

$$H_1 : \rho > \rho_0$$

Prueba Condicional a la Significacia del ρ

El estadigrafo de prueba que se utiliza para realizar una prueba de hipótesis acerca del ρ cuando se sabe que es distinto de cero es,

$$\frac{\mathfrak{z}(\hat{\rho}) - \mathfrak{z}(\rho)}{\sqrt{\frac{1}{n-3}}} \sim N(0,1)$$

El valor numérico de $\mathfrak{z}(\hat{\rho})$ y de $\mathfrak{z}(\rho)$ se calculan con las correspondientes transformaciones logarítmicas, las cuales se encuentran tabuladas.

Aplicación

```
##
##  Pearson's product-moment correlation
##
## data:  TV and sales
## t = 17.668, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7218201 0.8308014
## sample estimates:
##       cor
## 0.7822244
```

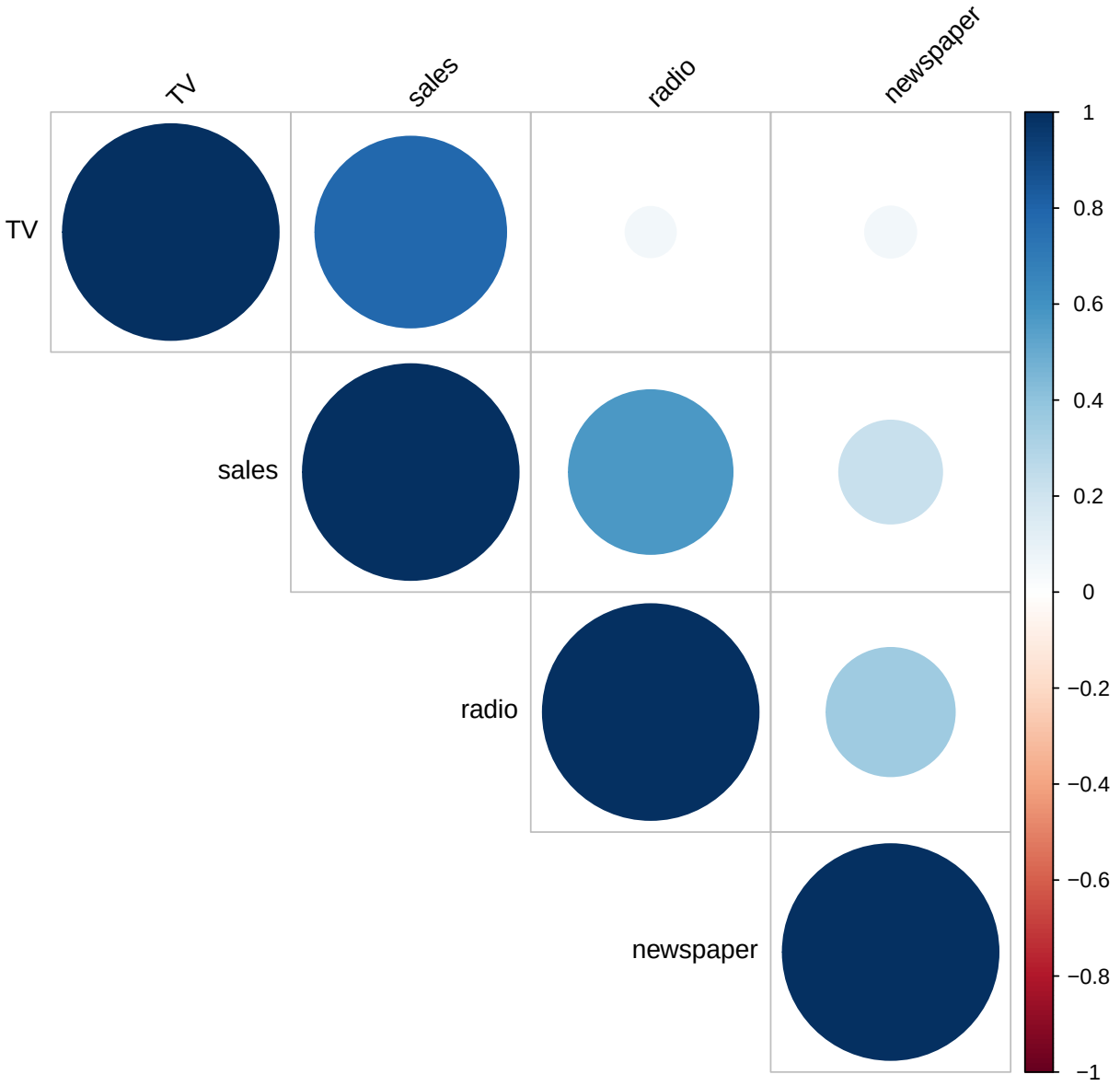
Aplicación

```
##
##  Pearson's product-moment correlation
##
## data:  radio and sales
## t = 9.9208, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4754954 0.6620366
## sample estimates:
##       cor
## 0.5762226
```

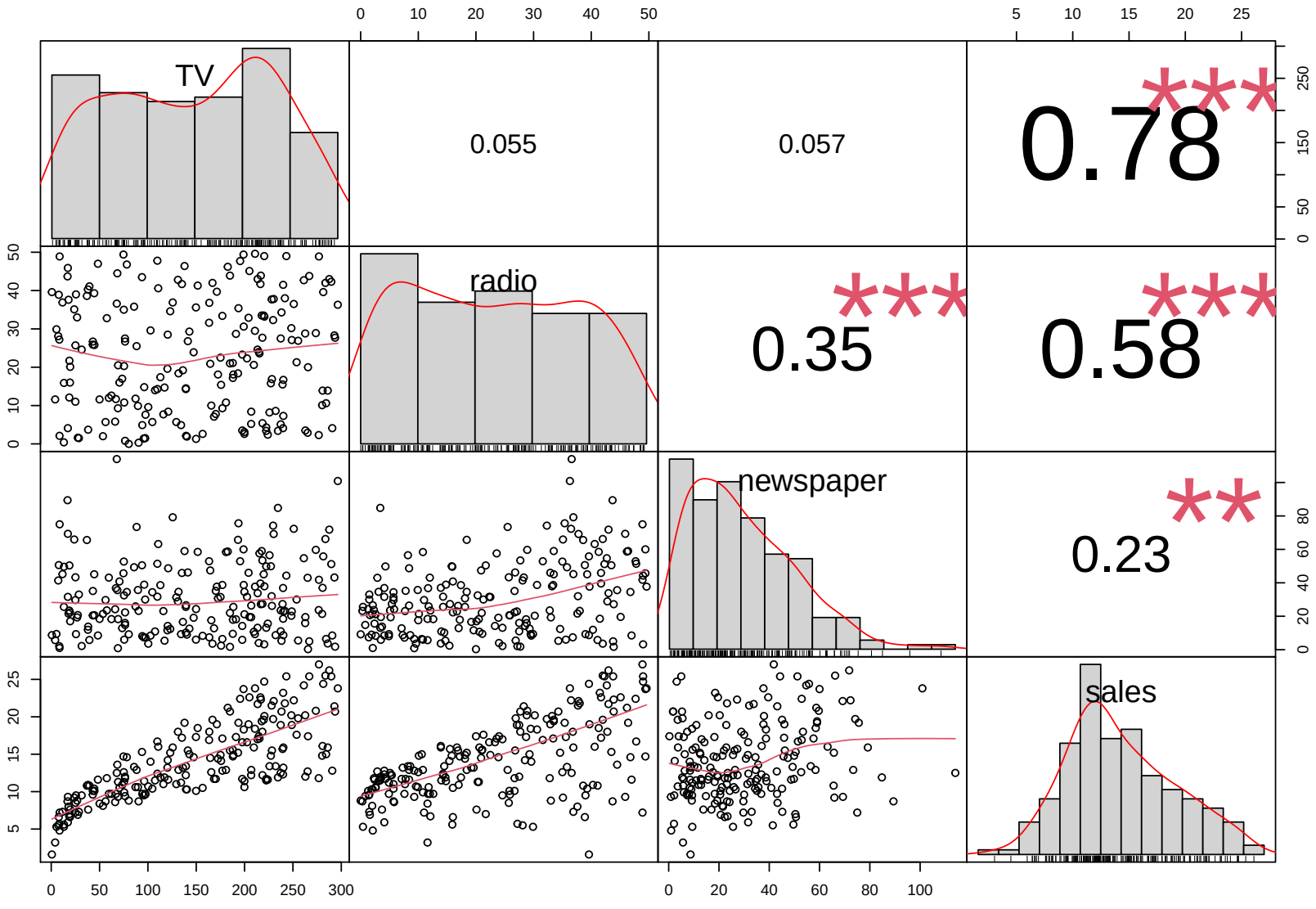
Aplicación

```
##
##  Pearson's product-moment correlation
##
## data:  newspaper and sales
## t = 3.2996, df = 198, p-value = 0.001148
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0924875 0.3557712
## sample estimates:
##      cor
## 0.228299
```

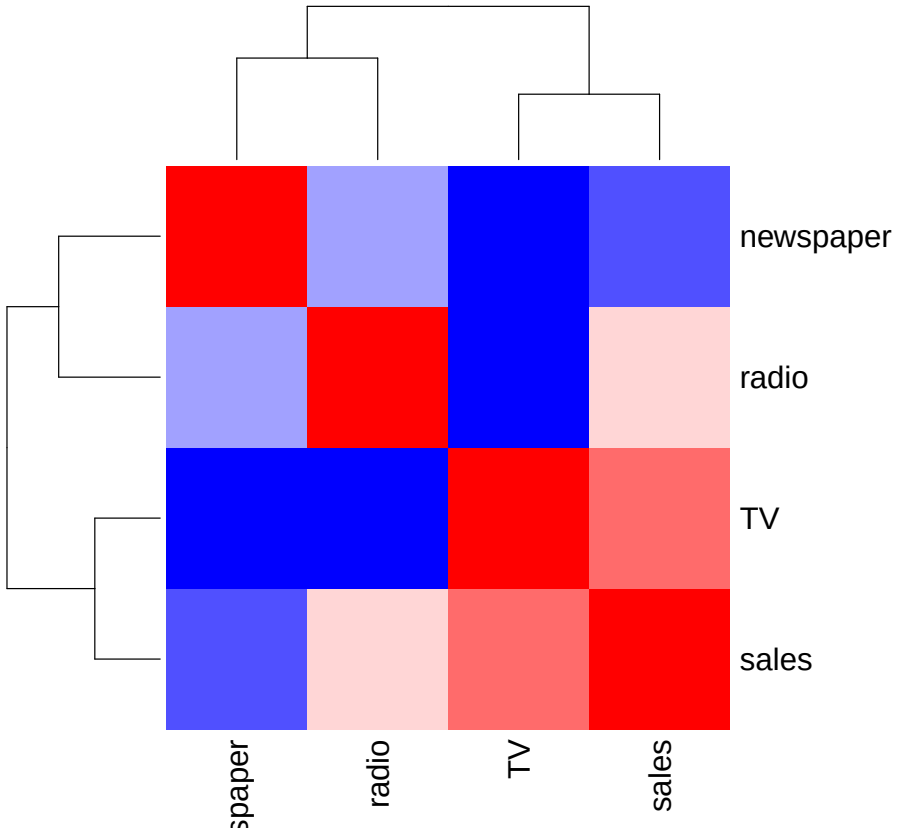
Aplicación


TV
sales
radio
newspaper
TV
sales
radio
newspaper
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Aplicación



Aplicación



Fin