

Regresión Avanzada

Universidad Austral

PhD. Débora Chan

Junio-Julio de 2023

Facultad de Ingeniería

Organización

- 1 Asociación entre variables
- 2 Regresión Lineal Simple
- 3 Estimación e Interpretación del modelo

Facultad de Ingeniería

Diagrama de dispersión

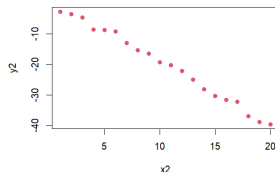
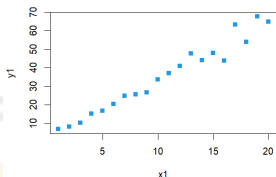
Es útil para analizar el grado de asociación entre variables cuantitativas, la intensidad de dicha asociación y la forma de la misma. Cuanto más cercanos estén los puntos del diagrama a una configuración rectilínea tanto más fuerte será la asociación lineal que vincula los valores de estas dos variables.

Si a medida que los valores de la variable X aumentan:

- ☐ En promedio también lo hacen los valores de la variable Y, esto señala la presencia de una **asociación positiva** entre las variables X e Y.
- ☐ Ee en promedio los valores de la variable Y decrecen esto señala la presencia de una **asociación negativa** entre las variables X e Y.
- ☐ No se aprecia un comportamiento, esto señala la no existencia de asociación entre las variables.

Formalizando

- a) Hay asociación positiva cuando valores superiores al promedio de una de ellas se presentan generalmente con valores superiores al promedio de la otra. Análogamente valores inferiores al promedio de una de ellas se presentan generalmente con valores inferiores al promedio de la otra.
- b) Hay asociación negativa cuando valores superiores al promedio de una de ellas se presentan generalmente con valores inferiores al promedio de la otra y viceversa



Cuantificando la Asociación Lineal

Necesitamos un coeficiente que:

- ⊖ Resulte positivo si la asociación lineal es positiva.
- ⊖ Resulte negativo si la asociación lineal es negativa.
- ⊖ Crezca en valor absoluto, a medida que la distribución de las observaciones se aproxima a una configuración rectilínea.
- ⊖ Su valor no dependa de las unidades utilizadas para expresar las variables.

Facultad de Ingeniería

Coefficiente de Correlación de Pearson

Karl Pearson (1837 –1936; estadístico y pensador británico)

propuso el Coeficiente de Correlación de Pearson (ρ) cuya expresión simbólica es la siguiente:



$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Donde:

σ_{XY} es la covarianza de (X, Y)

σ_X es la desviación estándar de la variable X

σ_Y es la desviación estándar de la variable Y

Estimación del CCLP

CCLP Muestral







En general disponemos de información muestral podemos estimar este coeficiente con un estadístico muestral, denotado como r_{xy} definido de la siguiente forma:

$$r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{(n-1) s_x s_y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

El numerador de esta expresión es la covarianza muestral, es muy sensible a las unidades de medición, por eso se elimina el efecto de las mismas mediante una estandarización.

Características del CCLP

Observaciones de r (coeficiente de correlación muestral)

-  Toma valores comprendidos entre -1 y 1 inclusive.
-  No distingue entre variables explicativas y variables respuestas.
-  Valores absolutos altos, del coeficiente indican grado de asociación lineal fuerte.
-  Si el valor de r es cercano a 0 , indica que no existe una tendencia creciente o decreciente entre las variables estudiadas.
-  El valor 1 (-1) ocurre sólo cuando todos los puntos yacen sobre una recta de pendiente positiva (negativa).
-  El signo de r depende de su numerador, puesto que su denominador resulta siempre positivo.

Inferencia

Supuestos para la validez de la Inferencia

Para realizar inferencias acerca de la población es necesario que las variables X e Y satisfagan ciertos supuestos.

- a) Los n pares han sido seleccionados aleatoriamente.
- b) Los pares de observaciones de la muestra son independientes.
- c) Ambas variables X y Y tienen distribución conjunta normal bivariada.

Facultad de Ingeniería

Ejemplo 1: Publicidad-Ventas

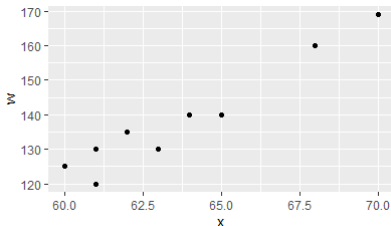
Interesa estudiar la existencia de correlación lineal entre la inversión en publicidad (X) y los ingresos de la empresa (W) con este objetivo se recogieron los datos correspondientes a ambas variables en un grupo de empresas nacionales con locales en shoppings.



Los Datos

Tabla: Gasto de Publicidad vs Ventas

X	60	61	61	62	63	64	65	68	70
W	125	130	120	135	130	140	140	160	169



Se aprecia asociación lineal positiva!

El código

analizar con cuidado que la distribución conjunta de las variables

sea normal bivariada, caso contrario la correlación debe estimarse en forma no paramétrica mediante el coeficiente de correlación de Spearman

```
library(MVN)
# incorporamos los datos
X=c(60, 61, 61, 62, 63, 64, 65, 68, 70)
W=c( 125, 130, 120, 135, 130, 140, 140, 160, 169)
data=data.frame(X,W)
# guardamos los resultados del Test Henze-Zirkler en el objeto result
result <- mvn(data , mvnTest = "hz")
# mostramos los resultados
result$multivariateNormality
```

La Salida

Test	HZ statistic	p value	MVN
Henze-Zirkler	0.4992	0.11982	YES

Puede sostenerse el supuesto distribucional normal bivariado para estas variables.

`cor(X,W)`

0.9665763

Limitaciones

En la Interpretación

- 1 La existencia de correlación entre dos variables implica solamente que las dos variables comparten variabilidad, no puede establecerse a partir de ella existencia de causalidad.
- 2 Puede ocurrir que exista una tercer variable que es la causal de esta correlación entre ambas.
- 3 También puede ocurrir que las variables no estén linealmente correlacionadas realmente sino que la correlación haya aparecido en forma casual en esta muestra seleccionada.

Las Hipótesis a Testear

$$H_0 : \rho = 0$$

versus

$$H_1 : \rho \neq 0$$

Como hemos planteado un test bilateral con estas hipótesis, rechazaremos la hipótesis de nulidad si el coeficiente de correlación lineal muestral, r , es suficientemente lejano de cero.

El estadístico del test es:



$$t_{obs} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

cuando

H_0

es verdadera

$r \sim t_{n-2}$

Ejemplo 1: continuación

Para nuestro ejemplo:

$$t_{obs} = \frac{0.966\sqrt{9-2}}{\sqrt{1-0.966^2}} = 9.97$$

El p- valor se calcula :

$$p - valor = 2 \times P(t_{9-2} > 9.97) = 2 \times 1.0875e - 05 = 2.175e - 05$$

Concluimos que el coeficiente de correlación poblacional entre las variables X (inversión en publicidad) y W (ventas de la empresa) es significativamente distinto de cero.

Ejemplo 1: el código y la salida

```
cor.test(X,W)
```

```
t = 9.9748, df = 7, p-value = 2.175e-05
```

```
95 percent confidence interval:
```

```
0.8446666 0.9931627
```

```
sample estimates:
```

```
cor
```

```
0.9665763
```

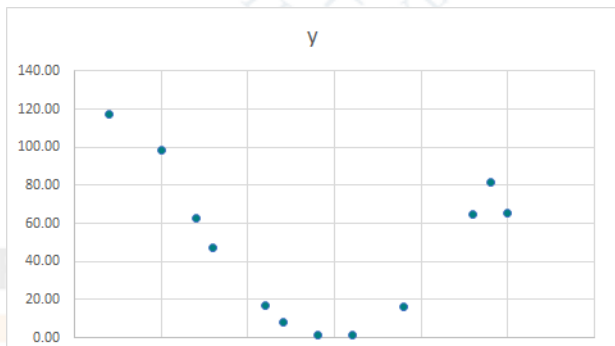
Nota: éste estadístico sólo es válido utilizarlo para testear hipótesis para Rho con respecto a cero (sea la alternativa: distinto, menor o mayor) pero no corresponde utilizarlo para hipótesis donde Rho es diferente de cero.

Siempre es lineal la relación??

Consideremos el siguiente ejemplo:

Para el siguiente conjunto de pares de observaciones:

x	2	5	7	8	11	12	14	16	19	
y	117.40	98.50	62.58	47.45	16.58	8.39	1.11	1.18	16.16	6



Ejemplo 2: Relación no Lineal

Cuál es el problema?

Si bien es claro que las dos variables están relacionadas, el coeficiente de correlación muestral $r = -0.23$ y el p valor del contraste de la prueba de Pearson es 0.4608.

Que la relación entre las dos variables no es lineal, es funcional, en este caso deberíamos realizar alguna transformación de las variables para poder asociar linealmente a las variables transformadas.

Conclusión


La inspección del diagrama de dispersión es fundamental para analizar si debe realizarse alguna transformación de la variable dependiente para linealizar la asociación entre las variables.

Coefficiente de correlación de Spearman

- Cuando no se satisfacen los supuestos no puede aplicarse el test de correlación lineal de Pearson, sin embargo disponemos de una alternativa no paramétrica para este coeficiente; el coeficiente de correlación por rangos de Spearman.
- Este coeficiente es una medida de asociación lineal que utiliza los rangos, números de orden, de cada grupo de sujetos y compara dichos rangos.
- El valor del coeficiente de correlación ρ_S de Spearman es el mismo que el coeficiente de correlación de Pearson, pero calculado sobre el conjunto de pares de rangos de las observaciones.
- El CCS se puede aplicar cuando las mediciones son de nivel al menos ordinal, cuando no puede sostenerse el supuesto de normalidad de ambas variables o cuando los datos presentan valores extremos ya que dichos valores afectan mucho el CCLP.

Coefficiente de Correlación de Spearman

El cálculo del coeficiente está dado por:


$$\hat{\rho}_S = r_S = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n(n+1)(n-1)}$$

en donde $d_i = \text{rango}(x_i) - \text{rango}(w_i)$ es la diferencia entre los rangos de los valores observados de X y W.

Facultad de Ingeniería

Ejemplo 3: Nicotina

Se realiza un estudio para determinar la asociación entre la concentración de nicotina en sangre (nmol/l) de un individuo y el contenido en nicotina de un cigarrillo (mg).

X	196.5	199.1	199.9	204.2	204.2	207.4	234.1	181.7	183
Y	0.76	1.11	1.66	0.96	1.21	1.14	1.53	1.51	1.28



Orden por Rangos

X(Nicot en Sangre)	Y(Nicot por cigarrillo)	d_i	d_i^2
181.7 (1)	1.51 (8)	-7	49
183 (2)	1.28 (7)	-5	25
192.8 (3)	0.84 (2)	1	1
196.5 (4)	0.76 (1)	3	9
199.1 (5)	1.11 (4)	1	1
199.9 (6)	1.66 (10)	-4	16
204.2 (7.5)	0.96 (3)	4.5	20.25
204.2 (7.5)	1.21 (6)	1.5	2.25
207.4 (9)	1.14 (5)	4	16
224.1 (10)	1.52 (9)	1	1

El Procesamiento

En este caso $\sum_{i=1}^n d_i^2 = 140.5$, luego el estadístico de contraste es:

$$r_S = 1 - \frac{6 \times 140.5}{10 \times 9 \times 11} = 0.145897$$

```
x=c(196.5, 199.1, 199.9, 204.2 ,204.2, 207.4, 234.1, 181.7 ,183, 192.8)
y=c(0.76, 1.11, 1.66, 0.96, 1.21, 1.14, 1.53, 1.51, 1.28, 0.84)
cor.test(x,y,method="spearman")
```

Spearman's rank correlation rho

$S = 140.93$, p-value = 0.6876

alternative hypothesis: true rho is not equal to 0

sample estimates: rho 0.145897

Interpretación del Coeficiente de Sperman

Interpretación del Coeficiente ρ_s de Spearman

La interpretación de este coeficiente es similar a la Pearson.

- Valores próximos a 1 indican una correlación fuerte y positiva.
- Valores próximos a -1 indican una correlación fuerte y negativa.
- Valores próximos a cero indican que no hay correlación lineal.

negativa
fuerte

-1 a -0.7

-0.7 a -0.4

-0.4 a 0.4

0.4 a 0.7

0.7 a 1

baja

positiva
fuerte

Correlación dentro de una base

Para analizar la correlación de a pares entre varias variables de una base disponemos del recurso del correlograma que hace un poco más visual este análisis.

Datos de calificación de estudiantes disponibles en shorturl.at/juE69.

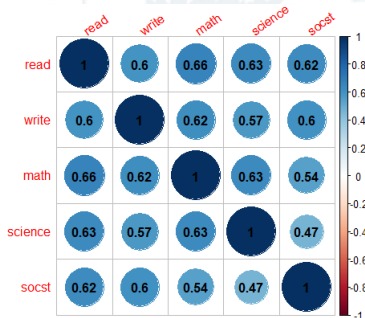
```
library(corrplot)
```

```
base_est=estud %>%select(read, write, math, science, socst)
```

```
M=cor(base_est)
```

```
corrplot(M, tl.col = "red", bg = "White", tl.srt = 35, addCoef.col  
= "black", type = "full")
```

Correlograma



En todos los casos la asociación entre las variables es positiva en algunos casos es más fuerte que en otros, por ejemplo es muy bajo en ciencia con estudios sociales y más alto en matemática con lectura.

Organización

- 1 Asociación entre variables
- 2 Regresión Lineal Simple
 - Modelo de regresión lineal simple
 - Bondad de Ajuste del Modelo Lineal Simple
- 3 Estimación e Interpretación del modelo

Facultad de Ingeniería

Modelo

Modelo General

Cuando los valores de ciertas variables que denominaremos regresoras o explicativas o predictoras nos permite aproximar el valor de una variable de interés que denominaremos variable objetivo o respuesta pero, aún así, no nos permite determinarlo con exactitud tendrá sentido definir un **modelo de regresión** :



$$Y = f(X_1, X_2, \dots, X_k) + \varepsilon$$

La expresión del modelo tiene una parte funcional y un término de error (ε) al modelo. La magnitud del término del error indica que tan cerca se encuentran los valores de la variable respuesta de la componente funcional

Ejemplo 4: Damascos

La siguiente tabla contiene información acerca de 18 variedades de damascos. Las variables registradas son las siguientes:

Superficie de la hoja en cm^2 (**SUPHOJA**).

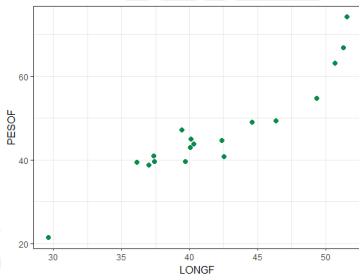
Peso del fruto en gramos (**PESOF**).

Longitud del fruto en centímetros (**LONGF**).

SUPHOJA	44.09	36.67	51.72	36.04	38.97	41.28	42.06	53.3
PESOF	49.29	49	43.04	66.79	63.11	43.8	39.63	44.9
LONGF	46.34	44.58	40.06	51.3	50.7	40.29	39.71	40.
SUPHOJA	40.14	39.31	33.53	36.88	36.94	34.13	42.03	41.5
PESOF	21.44	38.75	40.96	39.39	54.7	44.65	39.65	47.1
LONGF	29.63	37	37.38	36.14	49.33	42.37	37.4	39.4

Ejemplo 4: Damascos

Interesara analizar si el la longitud del fruto puede explicarse a través del peso del mismo y si esta relación entre ambos puede suponerse lineal.



Las preguntas son ahora:

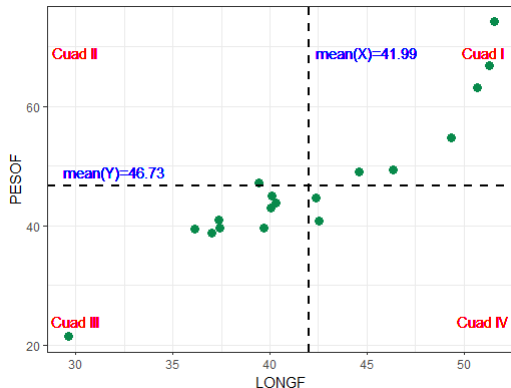
Diagrama de Dispersión

- ♠ ¿Existe alguna relación funcional que pueda explicar el peso del fruto a partir de la longitud del mismo?
- ♠ ¿Al aumentar una de estas variables qué sucede con la otra? y al disminuir?

Para visualizar la relación entre dos variables cuantitativas se usa el diagrama de dispersión. Cada punto corresponde a un par de valores (uno para cada variable), observados sobre la misma unidad de análisis.

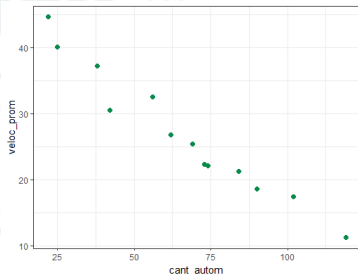
En general, la variable explicativa se representa en el eje horizontal o de abscisas (eje x) y a la variable respuesta en el eje vertical o de ordenadas (eje y).

Ejemplo 4: Diagrama de dispersión



En este gráfico los puntos se aprecian cercanos a una recta de pendiente positiva.

Ejemplo 5: Embotellamiento



Facultad de Ingeniería

Cuándo es adecuado el MLS?

Relación Funcional

Cuando un diagrama de dispersión muestra un patrón lineal es deseable resumir ese patrón mediante la ecuación de una recta. Esa recta debe representar a la mayoría de los puntos del diagrama, aunque ningún punto de los observados caiga sobre ella. La relación entre datos reales rara vez es tan simple como la expresada por la ecuación funcional.

Un modelo más realista plantea que la media poblacional de Y , más que los valores individuales, cambia linealmente con X .

Simbólicamente lo podemos expresar:



$$\mu_Y(x) = E(Y/X = x) = \alpha + \beta x$$

Término del error del Modelo

Lo que falta explicar

Otras variables, además de X , hacen que los valores individuales Y varíen alrededor de la media $\mu_{(X)}$ cuando X toma el valor x .

En el ejemplo, además de la cantidad de autos de la cuadra, la presencia o no de una ambulancia, el horario, el tamaño de los autos, el carácter de los conductores, la proximidad a una escuela u hospital, etc.

Todas esas 'otras variables' están representadas en el término del error, ε (épsilon) como la diferencia entre un valor individual y la media de la variable Y en la población, para un valor fijo de la variable explicativa X .

Errores del Modelo Lineal



$$\varepsilon_i = Y_i - \alpha - \beta X_i$$

Es decir:



$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Facultad de Ingeniería

Supuestos del ML relativos a los Errores

Supuestos del Modelo Lineal

- ☺ Los errores son independientes: $cor(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$.
- ☺ Los errores tienen distribución normal $\varepsilon_i \sim N \quad \forall i$.
- ☺ Los errores tienen media cero $E(\varepsilon_i) = 0$.
- ☺ Los errores son homocedásticos $Var(\varepsilon_i) = \sigma^2 \quad \forall i$.
- ☺ Sintéticamente: $\varepsilon_i \sim N(0, \sigma^2) \quad \forall i$ y $cor(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$.

El modelo de regresión lineal permite que los valores individuales de la variable respuesta se encuentren alrededor de la recta de regresión y no necesariamente sobre ella.

Parametrización del Modelo Lineal

Parámetros del Modelo Lineal

α : ordenada al origen de la recta.

β : pendiente de la recta.

σ^2 : varianza de los errores.

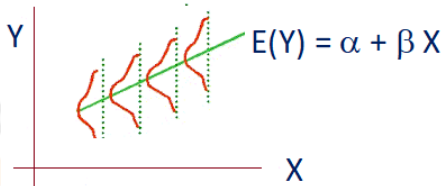
$$\overbrace{Y = \alpha + \beta x}^{\text{Recta}} + \underbrace{\varepsilon}_{\substack{\text{Otras cosas} \\ \text{además de } x}}$$

Supuestos Distribucionales de los Errores del ML

Lo que falta explicar

El término de error permite describir la variabilidad de las observaciones alrededor de la media poblacional $\mu(X)$.

Utilizamos la curva Normal para describir la variabilidad de las observaciones alrededor de la media y tenemos una media diferente para cada valor de la variable explicativa. Sobre la recta de regresión, el valor medio de y está determinado por $(x; \alpha + \beta x)$.






Método de Mínimos Cuadrados

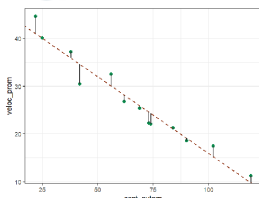
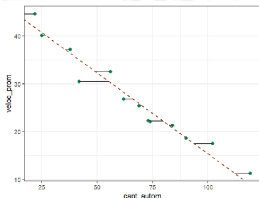
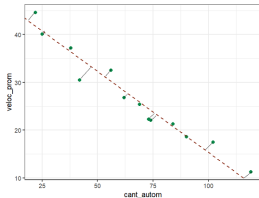
La recta de regresión es desconocida, se conoce la recta de regresión estimada.

El método de cuadrados mínimos propone elegir la recta que minimiza la suma de los cuadrados de las distancias de los puntos observados a la recta ajustada.

La pregunta es cuáles distancias queremos minimizar y por qué?

-  Las distancias verticales.
-  Las distancias ortogonales.
-  Las distancias horizontales.

Qué distancias interesa minimizar?



Los mismos puntos están en los tres diagramas de dispersión

El primero de los diagramas minimiza las distancias ortogonales, el segundo las distancias horizontales y el tercero las distancias verticales. Utilizaremos el método de mínimos cuadrados para minimizar las distancias verticales dado que lo que pretendemos es minimizar el error cometido en la estimación de la variable respuestas Y .

Visualizamos los Residuos

Sean:

(x_i, y_i) coordenadas de un punto del plano representando al dato i

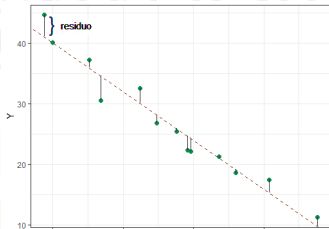
(x_i, \hat{y}_i) coordenadas de un punto sobre la recta estimada con $x = x_i$.

e_i residuo correspondiente a la i -sima observación.

$\hat{\alpha} = a$ estimación de la ordenada al origen de la recta.

$\hat{\beta} = b$ estimación de la pendiente de la recta.

Figure: Residuo



Residuos

La distancia vertical de un punto (x_i, y_i) a la recta es llamada residuo y se obtiene de la siguiente manera:



$$e_i = \text{valor observado}_i - \text{valor ajustado}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}x_i$$

Algunos residuos son positivos, la respuesta observada está por encima de la recta, y otros son negativos, la respuesta observada está por debajo de la recta estimada.

Cuando un punto se encuentra por encima de la recta ($y_i > \hat{y}_i$) luego $e_i = (y_i - \hat{y}_i) > 0$.

Cuando un punto se encuentra por debajo de la recta ($y_i < \hat{y}_i$) luego $e_i = (y_i - \hat{y}_i) < 0$.

Estimación de los coeficientes del Modelo

Una estrategia sencilla es el método de cuadrados mínimos (CM) que minimiza la suma de los cuadrados de los residuos sobre el conjunto de todas las observaciones. A esta forma de estimación se la conoce como **O.L.S.** (ordinary least squares).



$$SCR = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

La suma de cuadrados residuales, puede pensarse como una función a optimizar que depende de los parámetros a y b .

Para minimizarla se deben realizar las derivadas parciales de la misma, buscar el o los puntos estacionarios y verificar mediante el criterio de la derivada segunda o Hessiano, que se trate de un mínimo.

Sistema de Ecuaciones Normales




$$\begin{cases} \frac{\partial(SCR)}{\partial a} = -2 \sum_{i=1}^n (y_i - a - bx_i) = 0 \\ \frac{\partial(SCR)}{\partial b} = -2 \sum_{i=1}^n (y_i - a - bx_i)x_i = 0 \end{cases}$$

Este sistema de ecuaciones recibe el nombre de **sistema de ecuaciones normales**.

Facultad de Ingeniería

Siempre admite solución el Sistema de Ecuaciones Normales?

El Hessiano correspondiente tiene la siguiente expresión:


$$H = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} = n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 = n^2 V(x) > 0$$

El valor positivo del hessiano nos garantiza la existencia de un mínimo local.

Expresión de la Solución del Sistema

Además, este sistema admite siempre solución y la solución del sistema puede expresarse de la siguiente manera:



$$\begin{cases} \hat{\beta} = b = \frac{S_y}{S_x} r \\ \hat{\alpha} = a = \bar{y} - \hat{\beta} \cdot \bar{x} \end{cases}$$

Siendo:

S_x y S_y las desviaciones standard de x e y respectivamente.

r el coeficiente de correlación lineal de Pearson estimado entre x e y .

Cabe destacar que las estimaciones de los coeficientes del modelo lineal realizadas mediante cuadrados mínimos ordinarios coinciden con las estimaciones realizadas por máxima verosimilitud.

Expresión Matricial del Modelo Lineal

La expresión matricial del modelo lineal simple puede expresarse:



$$Y = X\beta + \varepsilon$$

donde:



$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \dots & \dots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

Expresión Matricial de la Estimación

Procedimiento

Siendo X^t la matriz traspuesta de X , premultiplicando la expresión anterior por X^t , se tiene:



$$X^t Y = X^t X \beta + X^t \varepsilon$$

La matriz $X^t X$ es una matriz cuadrada y no siempre admite inversa. Vamos a trabajar con el caso en que sí la admite y luego se pueden generalizar los resultados al caso en que no admite inversa utilizando la matriz inversa generalizada.

Forma Matricial de la Solución

Procedimiento

Como trabajamos el caso de $X^t X$ inversible podemos premultiplicar la expresión por $(X^t X)^{-1}$, obteniendo:



$$(X^t X)^{-1} X^t Y = \beta + (X^t X)^{-1} X^t \varepsilon$$

Pero como $E(\varepsilon) = 0$, entonces $E((X^t X)^{-1} X^t \varepsilon) = 0$ y resulta razonable la estimación de los coeficientes:



$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

Características de la solución

El estimador presentado es un estimador insesgado.



$E(Y/X) = X\beta$ (varianza nula de los errores).



$V(Y/X) = \sigma^2 I$ (homocedasticidad de los errores).



$E(\hat{\beta}) = E((X^t X)^{-1} X^t Y) = (X^t X)^{-1} X^t X \beta = \beta$



Como los errores tienen distribución normal, por definición del modelo, Y también tiene distribución normal (estamos considerando X constante, en principio). Luego los coeficientes estimados son una combinación lineal de Y , por lo tanto también tienen distribución normal.

Estimador B.L.U.E.

Optimalidad

El Teorema de Gauss-Markov asegura la **optimalidad** del estimador de mínimos cuadrados ordinarios, bajo el cumplimiento de:

- 1 buena especificación del modelo
- 2 independencia de las observaciones
- 3 esperanza de los errores nula, condicionada a X
- 4 homocedasticidad de los errores
- 5 matriz X de rango completo.

Optimalidad se refiere a la condición de BLUE (best linear unbiased estimator).


Estimación de la Varianza Residual

Los errores son homocedásticos con varianza común igual a σ^2 y media 0.

El problema es que a los errores no los podemos observar y en su lugar tenemos los residuos.


Tanto los errores como los residuos son variables aleatorias; sin embargo los residuos **no heredan todas las propiedades de los errores**. Por ejemplo los errores son independientes y los residuos, no. Además, los errores tienen todos la misma varianza, pero los residuos no.


Para estimar la varianza residual utilizaremos la expresión:


$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Variabilidad de los Coeficientes estimados

Designamos con $\hat{\beta}_{\sim}$ al vector de coeficientes estimados:


$$\hat{\beta}_{\sim} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}$$


$$Var(\hat{\beta}_{\sim}) = \underbrace{(X^t X)^{-1} X^t X (X^t X)^{-1}}_I \sigma^2 I = (X^t X)^{-1} \sigma^2$$

Expresión para el caso de Modelo Lineal Simple

$$(X^t X) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \dots & \dots \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

$$(X^t X)^{-1} = \frac{1}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

Facultad de Ingeniería

Variabilidad de los Coeficientes Estimados

De la expresión matricial se desprende que:

$$\begin{aligned} \text{Var}(\hat{\alpha}) &= \\ &\frac{\frac{1}{n} \sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2 - n\bar{X}^2} (\sigma^2) \end{aligned}$$

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \\ &\frac{1}{\sum_{i=1}^n x_i^2 - n\bar{X}^2} (\sigma^2) \end{aligned}$$

$$\begin{aligned} \text{Cov}(\hat{\alpha}, \hat{\beta}) &= \\ &\frac{-\bar{X}}{\sum_{i=1}^n x_i^2 - n\bar{X}^2} (\sigma^2) \end{aligned}$$

Test de Wald

Objetivo

Testear la significación de los coeficientes del modelo lineal. En la prueba estadística Wald, la estimación de máxima verosimilitud del parámetro de interés $\hat{\beta}$ se compara con el valor propuesto β_0 , suponiendo que la diferencia tipificada entre ambos tendrá una distribución normal aproximada. El cuadrado de esta diferencia se compara con una distribución de chi-cuadrado. En el caso univariado, las hipótesis del test de Wald son:



$$H_0 : \beta = 0$$

versus

$$H_1 : \beta \neq 0$$

Estadístico de Contraste

Expresión



$$U_{obs} = \frac{(\hat{\beta} - \beta_{H_0})^2}{\widehat{var}(\hat{\beta})} \sim \chi_1^2$$

Otra opción es comparar la diferencia con la distribución t de Student.




$$T_{obs} = \frac{\hat{\beta} - \beta_{H_0}}{\widehat{sd}(\hat{\beta})} \sim t_1$$

$\widehat{sd}(\hat{\beta})$ es el error estándar estimado de la estimación de Max.verosim.

Intervalos de Confianza

También pueden construirse intervalos de confianza para la estimación de los valores verdaderos de los parámetros de la regresión.

La expresión del Intervalo de Confianza de nivel $1 - \alpha$ para el coeficiente β del modelo de regresión lineal es:


$$\hat{\beta} \pm t_{n-2, 1-\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}$$

Facultad de Ingeniería

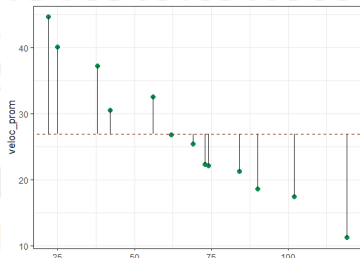
Coeficiente de determinación(R^2)

Cuantifica la proporción en que se reduce el error entre el que se comete al estimar la Y con su promedio muestral \bar{Y} y estimarla con la recta de regresión $Y_i = a + bX_i$.

Suma de Cuadrados Totales: distancias verticales a \bar{Y}



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



Coefficiente de determinación(R^2)

La **suma de Cuadrados Residuales** es la suma de las distancias verticales entre valores observados y sus correspondientes estimaciones sobre la recta.



$$SCR = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

y su diferencia relativa; denominada **coeficiente de determinación**:



$$R^2 = \frac{SCT - SCR}{SCT}$$

En síntesis, el coeficiente de determinación indica la proporción de la variabilidad de la variable respuesta que logra ser explicada por la variable explicativa a través de la recta de regresión lineal estimada.

Limitaciones del Coeficiente de Determinación

Necesidad de otras cuantificaciones de bondad de ajuste



A veces no resulta una buena medida de ajuste .



Al agregar nuevas variables al modelo suele mejorar pero podría conducir a modelos poco parsimoniosos o sobreajustados.



R^2 no puede determinar si las estimaciones y predicciones de los coeficientes están sesgadas, y es por eso que se deben examinar las gráficas de residuos.

Facultad de Ingeniería

Cuarteto de Anscombe: Los Datos

Cuarteto de Anscombe

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

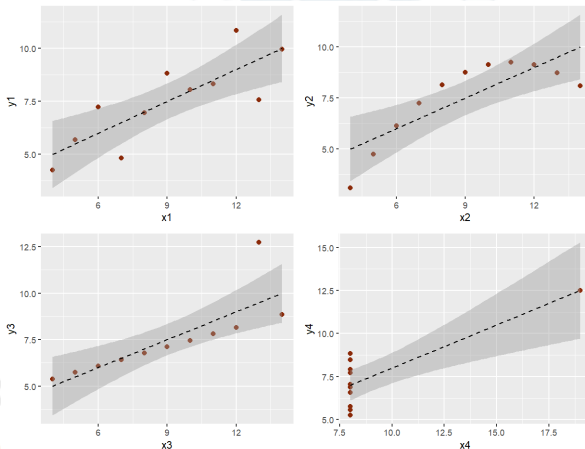
Cuarteto de Anscombe: Las propiedades

El cuarteto de Anscombe consta de cuatro conjuntos de datos que comparten las mismas propiedades estadísticas, ponen de relieve la importancia de la inspección gráfica un conjunto de datos antes de analizarlos.

Propiedad	Valor
Media de la variable X	9
Varianza de la variable X	11
Media de la variable Y	7.5
Varianza de la variable Y	4.12
Correlación entre cada una de las variables X e Y	0.816
Recta de regresión estimada	$\hat{Y} = 3 + 0.5X$

Cuarteto de Anscombe: Configuración de los Datos

La configuración de los cuatro conjuntos:



Coeficiente de Determinación Corregido R^2_{adj}

Necesidad

El R^2 ajustado se justifica pues a medida que añadimos variables a una regresión, el R^2 tiende a aumentar. Incluso cuando la contribución de las nuevas variables añadidas no tenga relevancia estadística. Por ende añadir variables podría conducir al 'sobreajuste del mismo'.

Para solucionar este problema muchos investigadores sugieren ajustar el coeficiente determinación mediante la siguiente fórmula:



$$R^2_{ajust} = 1 - \left[\frac{n-1}{n-k-1} \right] (1 - R^2)$$

con n es la cant. de observaciones y k es la cant. de variables predictoras

Relación entre R^2 y R^2_{ajust}

R^2_{ajust} , penaliza la incorporación de nuevas variables. En la fórmula se puede observar que al encontrarse el parámetro k en el denominador, disminuye el valor alcanzado por R^2_{ajust} a medida que se incorpore una nueva variable. Algebraicamente hablando, si la cantidad de parámetros (k) fuera cero, se lograría una igualdad entre R^2 y R^2_{ajust} .

Descomposición de la SCT

$$SCTot = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \underbrace{2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})}_{=0}$$

El sumando del doble producto puede probarse fácilmente que es nulo puesto que si se aplica la propiedad distributiva se ve que es la suma de los productos de residuos por ajustados (que son ortogonales) menos la suma de los residuos por la media de la variable respuesta que también es nula. De este modo queda que la suma de cuadrados totales es igual a la suma de los cuadrados residuales y la suma de los cuadrados de la regresión (lo que la regresión logra explicar).



$$SCTot = SCRes + SCReg$$

Salida ANOVA

La salida típica del modelo de regresión lineal simple es una tabla como la siguiente:

Fuente	S de Cuad	Df	Cuad Medios	est F	p-valor
Reg	SCReg	1	CMReg=SCReg/1	$\frac{CMReg}{CMRes}$	$P(F_{1,n-2})$
Resid	SCRes	n-2	CMRes=SCRes/(n-2)		
Tot	SCTot	n-1			

Este test corresponde al test de la regresión. En regresión lineal simple es el mismo test que el test de Wald para el coeficiente β , para varias variables predictoras este test tiene como hipótesis nula que todas las regresoras tienen simultáneamente valor nulo y como alternativa la negación de esta afirmación. Esta salida también se conoce como salida de ANOVA del modelo de regresión.

Bandas de Confianza

Para estimar el valor de Y_0 para $X = x_0$, puede hacerse a partir de la recta ajustada tendríamos en principio una estimación puntual del mismo.



$$\widehat{Y}_0 = \widehat{\alpha} + \widehat{\beta}x_0$$

Pero si queremos una estimación por intervalo del valor esperado para Y_0 , debemos recordar:

$E(Y/X = x_0) = \alpha + \beta x_0$ luego su estimación puntual es

$$\widehat{E}(Y/X = x_0) = \widehat{\alpha} + \widehat{\beta}x_0$$

$$V(Y/X = x_0) = V(\widehat{\alpha} + \widehat{\beta}x_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

IC para el valor esperado de una nueva observación

Este segundo resultado se deduce de las estimaciones de las varianzas de los coeficientes halladas previamente. Entonces la expresión del intervalo de confianza de nivel $1 - \alpha$ para la esperanza de la variable respuesta dada una nueva observación x_0 es:

$$\hat{Y}_0 \pm t_{n-2, 1-\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

La amplitud de este IC varía con la distancia de la nueva observación x_0 al valor medio de las observaciones \bar{x} . Cuando se construye este intervalo sobre todo el recorrido de las observaciones se obtienen las **bandas de confianza**.

Bandas de Predicción




Puede una nueva observación caer fuera de las bandas de confianza?

Consideremos ahora la necesidad de predecir una nueva observación Y_0 correspondiente a un nivel de $X = x_0$ nuevo. Este nuevo valor debe ser independiente de los valores muestrales $(x_1; y_1), (x_2; y_2), \dots (x_n, y_n)$ usados para estimar de la recta de regresión.

En el caso del IP predecimos el resultado individual de Y , tenemos dos fuentes de variabilidad: la variabilidad de $E(Y)$ como en el caso del IC y la variabilidad propia de la distribución de Y .

Intervalo de Predicción

La expresión del IP es similar a la del IC pero con una amplitud un poco mayor debida a la variabilidad de Y:


$$\hat{Y}_0 \pm t_{n-2, 1-\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Cuando se construyen las bandas de predicción la amplitud es variable y crece con la distancia de las observaciones respecto de la media de la variable predictora.

Organización

- 1 Asociación entre variables
- 2 Regresión Lineal Simple
- 3 Estimación e Interpretación del modelo

Facultad de Ingeniería

Estimación e Interpretación del modelo

Retomamos el ejemplo de los frutos de damasco. Vamos a estimar los coeficientes del modelo utilizando R y vamos a graficar la recta de regresión ajustada para este ejemplo.

```
mod_hojas=lm(PESOF~LONGF,data=hojas)
summary(mod_hojas)
coef=mod_hojas$coefficients
ggplot(hojas,aes(x=LONGF,y=PESOF))+
  geom_point(size=3,col="#088A4B")+ theme_bw()+
  geom_abline(intercept=coef[1] , slope=coef[2], col='#0B3B2E'
,linetype="dotted")+
  geom_text(x=45,y=60,label='y=32.2798+1.8818x' )
anova(mod_hojas)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-32.2798	7.3606	-4.385	0.000461	***
LONGF	1.8818	0.1736	10.838	8.87E-09	***

Residual standard error: 4.289 on 16 degrees of freedom

Multiple R-squared: 0.8801, Adjusted R-squared: 0.8726

F-statistic: 117.5 on 1 and 16 DF, p-value: 8.867e-09

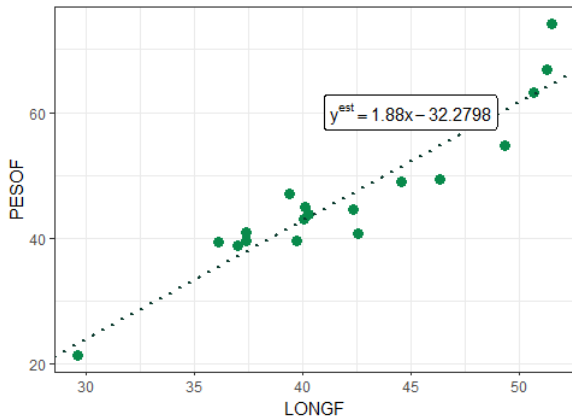
Analysis of Variance Table

Response: PESOF

	Df	Sum Sq	Mean Sq	F value	Pr(> t)	
LONGF	1	2160.33	2160.33	117.46	8.87E-09	***
Residuals	16	294.28	18.39			

Ajuste por OLS

La ecuación de la recta ajustada por mínimos cuadrados es:



Ajuste por OLS



$$\hat{Y} = -32.2798 + 1.8818X$$

La estimación de los parámetros del modelo es:

$$\hat{\alpha} = a = -32.2798$$

$$\hat{\beta} = b = 1.8818$$

$$\hat{\sigma}^2 = s_e^2 = 18.39$$



Cómo debemos interpretar la pendiente estimada?

La pendiente ajustada es $b = 1.8818$, es decir que, un aumento unitario en la variable X (longitud del fruto) produce un aumento de 1.8818 unidades en el valor esperado de la variable Y (peso del fruto).

Interpretación de la Salida

Además en la misma salida se nos informa que:

$$R^2 = 0.8801$$

$$R^2_{adj} = 0.8736$$

Esto indica que el 88% de la variabilidad de Y queda explicado por la variable de X a través del modelo estimado.

Coincidencia?

El análisis de la varianza del modelo señala que es significativo. Notemos que el p valor del test de Wald correspondiente al coeficiente β coincide con el p valor del test de la regresión. Esto no es casual, esto será siempre así cuando se trate de regresión lineal simple (una única variable explicativa), dado que ambos test evalúan la significación de la misma variable.

Hallamos el intervalo de confianza para una nueva observación con longitud del fruto $LONGF = -34.3$ de nivel 95%

```
nuevo <- data.frame(LONGF=34.3)
predict(object=mod_hojas, newdata=nuevo, interval=' confidence' ,
level=0.95)
```

	fit	lwr	upr
1	32.26443	28.71394	35.81492

Facultad de Ingeniería

Bandas de Predicción

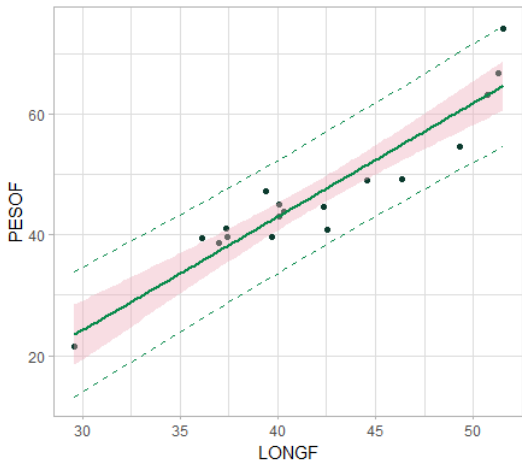
Realizamos esto sobre el recorrido de la variable predictora:

```
predichos <- predict(object=mod_hojas, interval='prediction',  
level=0.95)
```

```
nuevos_datos=data.frame(hojas,predichos)
```

```
ggplot(nuevos_datos, aes(x=LONGF, y=PESOF)) +  
geom_point(col=' #0B3B2E' ) +  
geom_line(aes(y=lwr), color=' #088A4B' , linetype=' dashed' ) +  
geom_line(aes(y=upr), color=' #088A4B' , linetype=' dashed' ) +  
geom_smooth(method=lm, formula=y~x, se=TRUE, level=0.95,  
col=' #088A4B' , fill='pink2') +theme_light()
```

Bandas de Confianza y Predicción



Seguimos la próxima

