Análisis de Datos con el Sistema Estadístico R

Lic. Patricia Vásquez Sotero

Correlación y Análisis de regresión lineal

Contenidos

- **Medidas de asociación lineal**
  - Covarianza
  - Correlación
- **Análisis de Regresión Lineal**
  - Introducción
  - Formulación del modelo
  - Estimación de parámetros
  - Inferencia del modelo
  - Contrastes en el modelo
  - Bondad de ajuste
  - Diagnosis

Bibliografía recomendada

Altman DG. Practical statistics for medical research. Boca Ratón, Chapman & Hall/ CRC; 1991.

Peña, Daniel (2002). Regresión y diseño de experimentos". Editores Alianza Editorial.

Peña, D. y Romo, J. (1997), Introducción a la Estadística para las Ciencias Sociales, Editorial McGraw Hill, Madrid.

ASOCIACIÓN LINEAL EN VARIABLES CUANTITATIVAS

Medidas de asociación

MEDIDAS DE ASOCIACIÓN LINEAL

- **Covarianza**
- Correlación

→ **Datos Cuantitativos**

x
x_1
x_2
⋮
x_n

Recordemos que: Hasta ahora hemos estudiado las **medidas tendencia central** (Media, Mediana, Moda) **y dispersión** (Varianza y Desviación Estándar) para **una Variable Cuantitativa** (x).

Covarianza: Es una medida de **Variabilidad Conjunta** entre **dos** variables (**x_1, x_2**) o bien (**x**, **y**)

x	y
$x_{(1)}$	$y_{(1)}$
$x_{(2)}$	$y_{(2)}$
⋮	⋮
$x_{(n)}$	$y_{(n)}$

$$\text{Cov}(x, y) = s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

donde x_i e y_i son los valores observados, $\bar{x}$ e $\bar{y}$ son las medias muestrales y n es el tamaño de la muestra.

- Si Cov(x,y) es positiva: la asociación entre **x** e **y** es directamente proporcional, es decir que cuando **x** aumenta **y** también aumenta; y viceversa.
- Si Cov(x,y) es negativa: la asociación entre **x** e **y** es inversamente proporcional, es decir que cuando **x** aumenta **y** disminuye; y viceversa.
- Si Cov(x,y) es cero: no existe asociación entre **x** e **y**.

Medidas de asociación

MEDIDAS DE ASOCIACIÓN LINEAL

- **Covarianza**
- Correlación

→ **Datos Cuantitativos**

x
x_1
x_2
$\vdots$
x_n

$$\text{Cov}(x, y) = s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

donde x_i e y_i son los valores observados, $\bar{x}$ e $\bar{y}$ son las medias muestrales y n es el tamaño de la muestra.

Inconvenientes de la Covarianza:

- No está acotada ni superior ni inferiormente. Por lo tanto no se sabe cuándo es s_{xy} suficientemente grande o pequeña.
- Depende de las unidades de medida de las variables:
  Si s_{xy} es la covarianza de X e Y, y $a, b \in \mathbb{R}$, $b \neq 0$ y $T = a + bY$, entonces:
  $s_{xt} = b s_{xy}$

Correlación lineal simple

MEDIDAS DE ASOCIACIÓN LINEAL

- Covarianza
- **Correlación**

→ **Datos Cuantitativos**

Correlación: Se refiere al grado de asociación entre **dos** variables (x_1 , x_2) o bien (x, y).

Coeficiente de Correlación de Pearson (r): Mide el **grado** de asociación lineal entre dos variables Cuantitativas.

x	y
$x_{(1)}$	$y_{(1)}$
$x_{(2)}$	$y_{(2)}$
⋮	⋮
$x_{(n)}$	$y_{(n)}$

$$r = \frac{cov(x,y)}{s_x s_y} \quad \cdots\cdots\blacktriangleright \quad r = \frac{\sum_{i=1}^{n} x_i y_i - nxy}{(n-1)s_x s_y}$$

$$-1 \le r \le 1$$

- Es una medida acotada
- Es adimensional

- Si r **es positivo:** la asociación entre x e y es directamente proporcional, es decir que cuando x aumenta y también aumenta; y viceversa. Si $r = 1$**:** la asociación lineal es perfecta.
- Si r **es negativo:** la asociación entre x e y es inversamente proporcional, es decir que cuando **x** aumenta **y** disminuye; y viceversa. Si $r = -1$**:** la asociación lineal es perfecta.
- Si r **es cero:** no existe asociación entre x e y.

Correlación lineal simple

MEDIDAS DE ASOCIACIÓN LINEAL

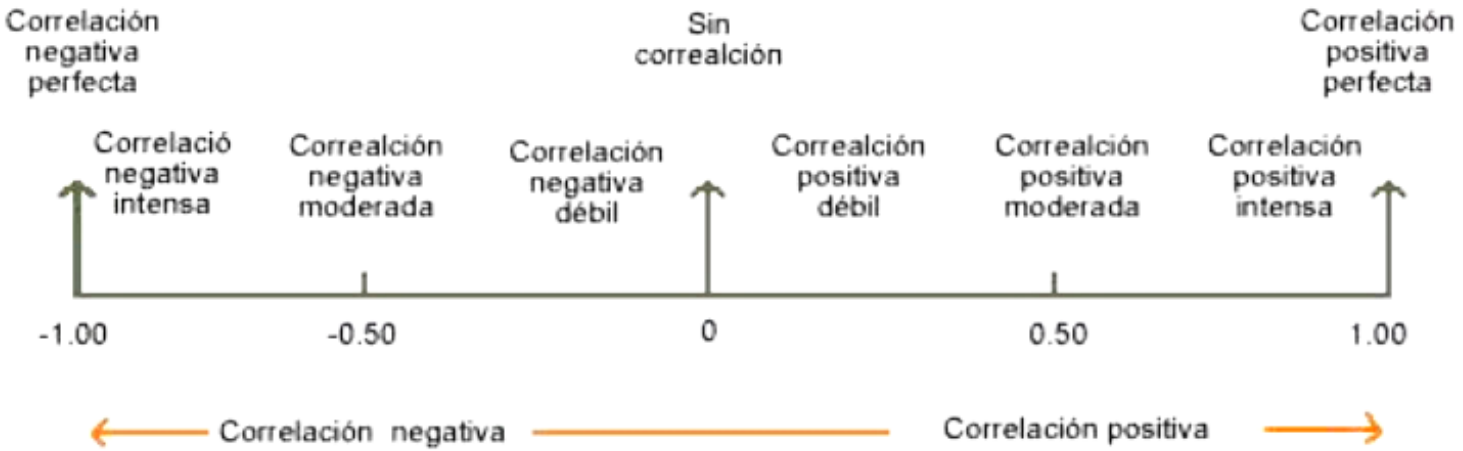
- Covarianza
- **Correlación**

→ **Datos Cuantitativos**

Correlación: Se refiere al grado de asociación entre **dos** variables (x_1 , x_2) o bien (x, y).

Coeficiente de Correlación de Pearson (r): Mide el **grado** de asociación lineal entre dos variables Cuantitativas.

Otras formas referenciales para interpretar r:



Correlación lineal simple

Diagrama de dispersión

- La representación gráfica más común para dos variables cuantitativas es el diagrama de dispersión

Ejemplo m^2 habitables y Precio de 15 viviendas.

m2	Precio
107	62657
114	65554
91	54506
100	62103
96	58271
107	66925
104	61917
100	61149
80	52263
81	51878
105	65678
111	66696
108	65387
97	61806
106	63824

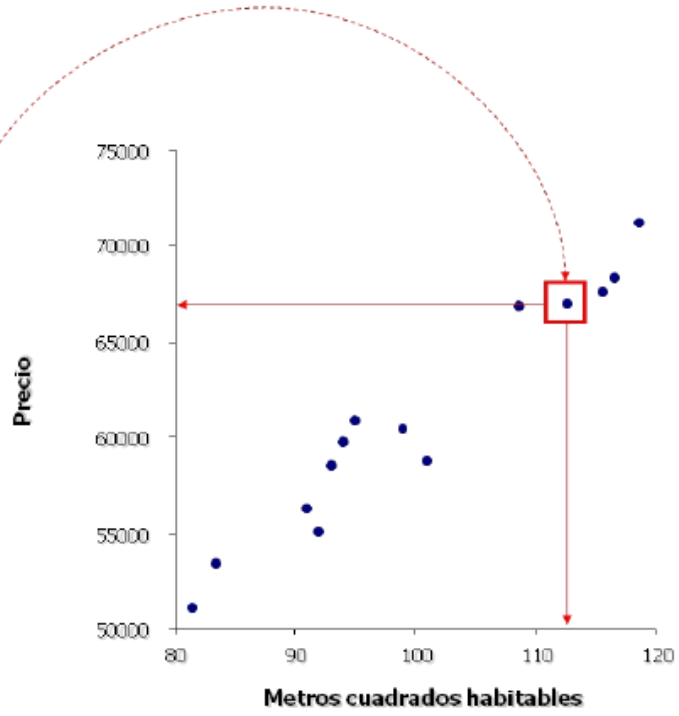


Diagrama de dispersión

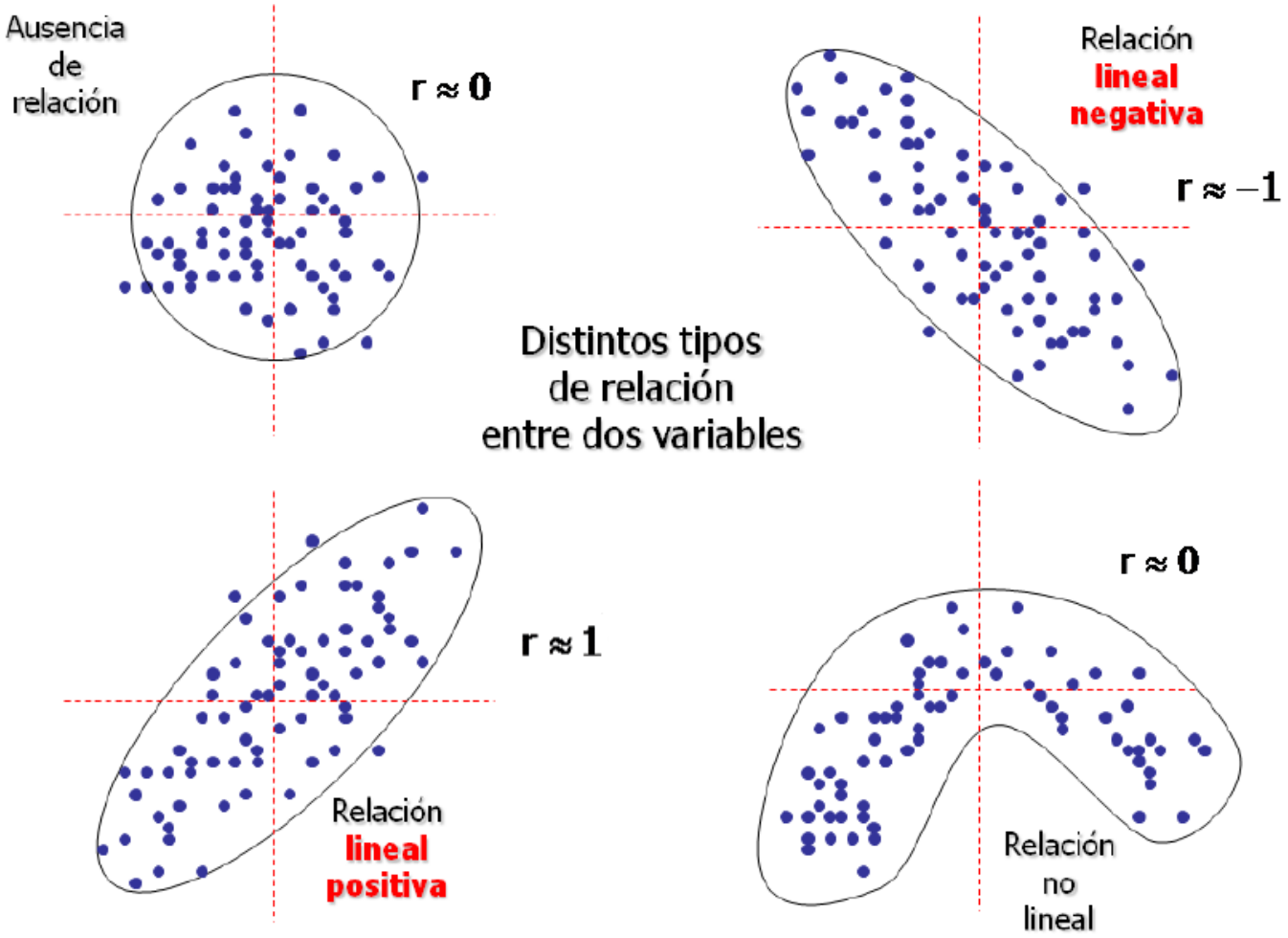
Correlación lineal simple

Interpretación de un diagrama de dispersión

- Es importante fijarse en las unidades de cada eje.
- ¿Se observa alguna asociación entre las variables?
- ¿Cómo es de estrecha la asociación entre las variables?
- ¿Cuál es la "dirección" de la asociación entre las variables?
- ¿Hay algún punto o colección de puntos que no siga el patrón general del resto?
- Si hay una tercera variable cualitativa, resulta conveniente utilizar símbolos o colores diferentes para cada valor de esta tercera variable.

Correlación lineal simple

Tipos de relación entre variables cuantitativas



Test de correlación lineal simple

Test de correlación

- Contraste sobre la independencia (correlación cero) de dos variables
- Datos normales
- Si no lo son, utilizamos la opción `method="kendall"`

Contraste

- $H_0 : \rho = 0$
- $H_A : \rho \neq 0$

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

Con distribución t de Student con n-2 grados de libertad cuando $\rho=0$

Ejemplo en código R

```
load("glucosa.Rdata"); attach(glucosa)
# Analizar distribución de variables ¿normal?
shapiro.test(g2antes); shapiro.test(g2des)
# Calculamos el coeficiente de correlación mediante la función cor
cor(g2antes, g2des) # Coef. de correlación entre el tiempo2 antes y después de
        haber realizado la prueba de glucosa en mujeres
cor.test(g2antes, g2des) # O simplemente con el Test de correlación
```

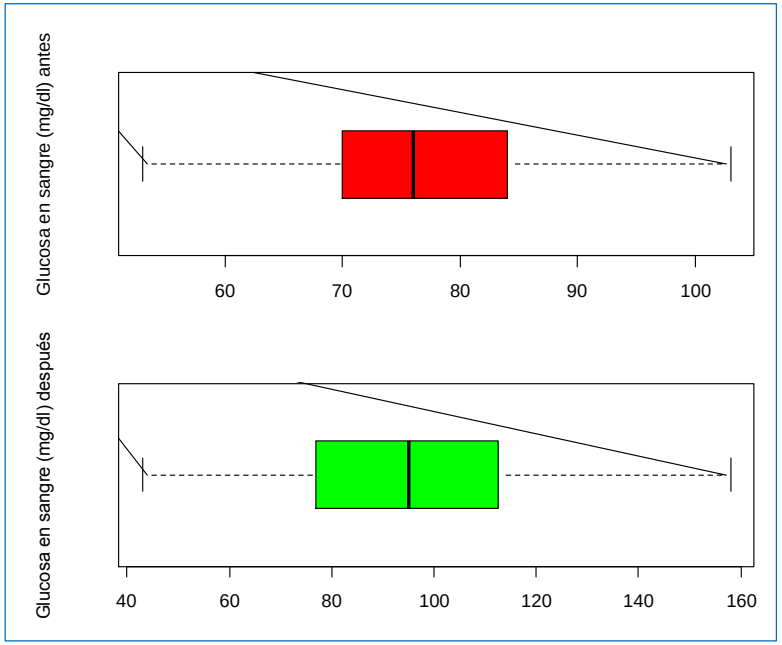
Test de correlación lineal simple

Resultados

```
        Shapiro-Wilk normality test

data:  g2antes
W = 0.98846, p-value = 0.695
```

```
        Shapiro-Wilk normality test

data:  g2des
W = 0.98763, p-value = 0.6402
```

El análisis gráfico y el contraste de normalidad muestran que se puede asumir normalidad en las variables




```
        Pearson's product-moment correlation

data:  g2antes and g2des
t = 1.7727, df = 78, p-value = 0.08019
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.02396395  0.39924354
sample estimates:
      cor
0.1967891
```

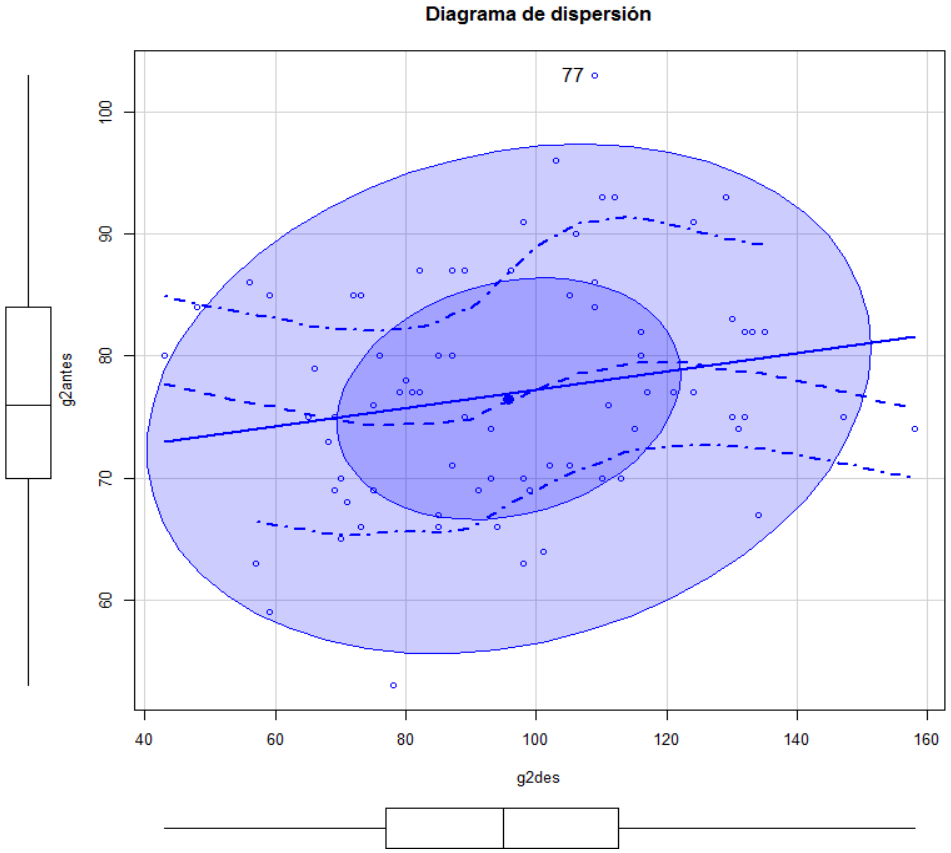
Existe una correlación significativa entre el tiempo2 antes y después de haber realizado la prueba de glucosa en mujeres.

Test de correlación lineal simple

Además es recomendable observar la forma de la relación entre las variables de análisis así como la presencia de valores extremos, utilizando un diagrama de dispersión.

Ejemplo en código R

```
scatterplot(g2antes ~ g2des,
    data=glucosa, ellipse=TRUE,
    main="Diagrama de
    dispersión",
    legend=list(coords="topleft"),
    id=list(method="identify"))
```



Test de correlación lineal simple

Ejemplo. Se dispone de un data set con información sobre diferentes automóviles. Se quiere estudiar si existe una correlación entre el peso de un vehículo (Weight) y la potencia de su motor (Horsepower).

Ejemplo en código R

```
require(MASS)
require(ggplot2)
data("Cars93")
```

En primer lugar se representan las dos variables mediante un diagrama de dispersión (+) para intuir si existe relación lineal o monotónica. Si no la hay, no tiene sentido calcular este tipo de correlaciones.

Ejemplo en código R

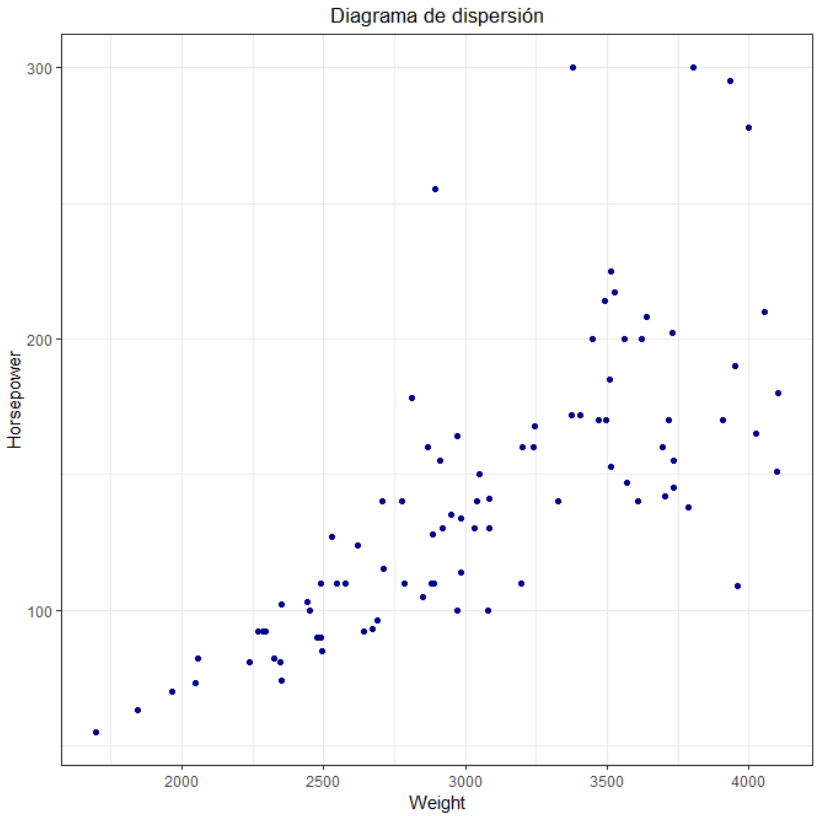
```
ggplot(data = Cars93, aes(x = Weight, y = Horsepower)) +
  geom_point(colour = "red4") +
  ggtitle("Diagrama de dispersión") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

Test de correlación lineal simple

Ejemplo. Cont.

El diagrama de dispersión parece indicar una posible relación lineal positiva entre ambas variables.

Para poder elegir el coeficiente de correlación adecuado, se tiene que analizar el tipo de variables y la distribución que presentan. En este caso, ambas variables son cuantitativas continuas y pueden transformarse en rangos para ordenarlas, por lo que a priori se podrían aplicar los coeficientes de kendall y de Pearson. La elección se hará en función de la distribución que presenten las observaciones.



Test de correlación lineal simple

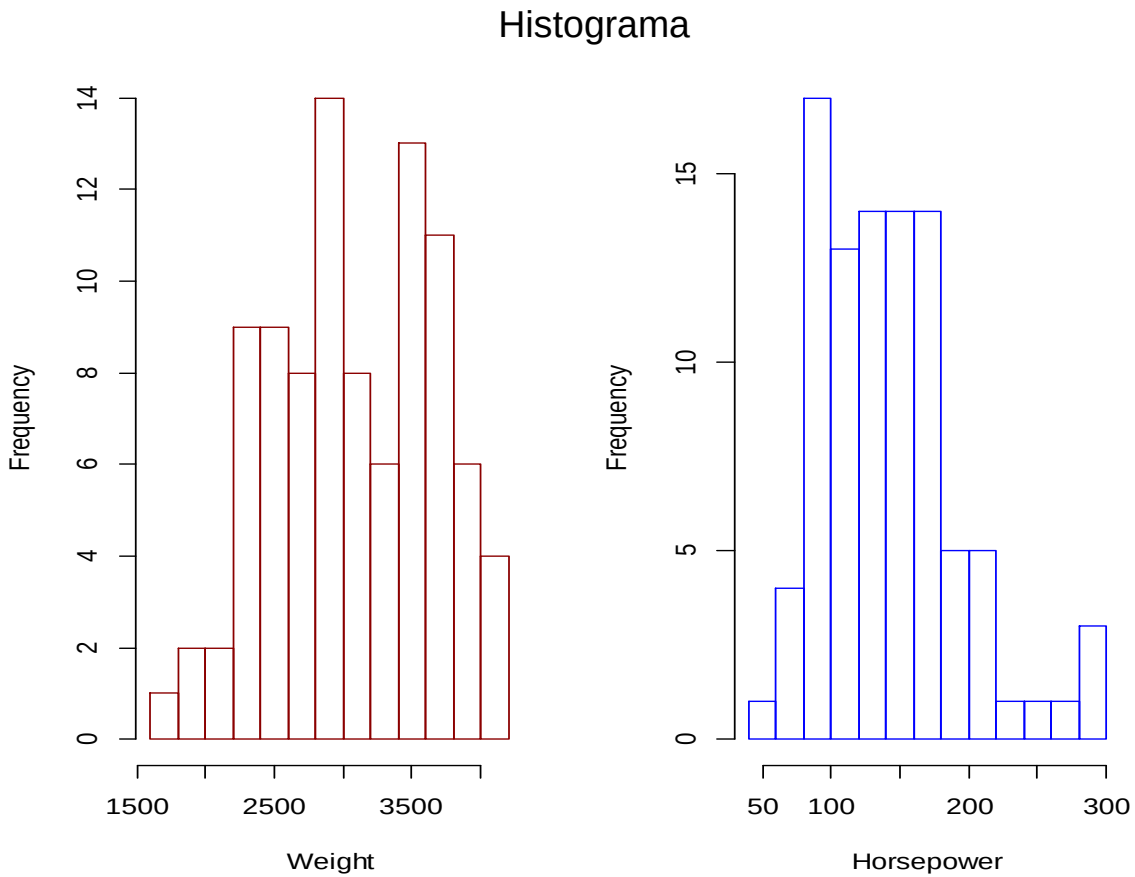
Ejemplo. Cont.

Ejemplo en código R - Análisis de la normalidad

```
# Representación gráfica
par(mfrow = c(1, 2))
hist(Cars93$Weight, breaks = 10, main = "", xlab = "Weight", border = "darkred")
hist(Cars93$Horsepower, breaks = 10, main = "", xlab = "Horsepower",
     border = "blue")

par(mfrow = c(1, 2))
qqnorm(Cars93$Weight, main = "Weight", col = "darkred")
qqline(Cars93$Weight)
qqnorm(Cars93$Horsepower, main = "Horsepower", col = "blue")
qqline(Cars93$Horsepower)

# Test de hipótesis para el análisis de normalidad
shapiro.test(Cars93$Horsepower)
shapiro.test(Cars93$Weight)
```

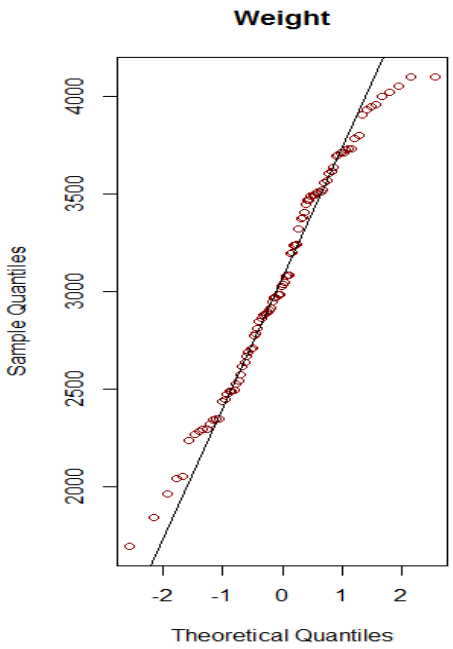
Test de correlación lineal simple

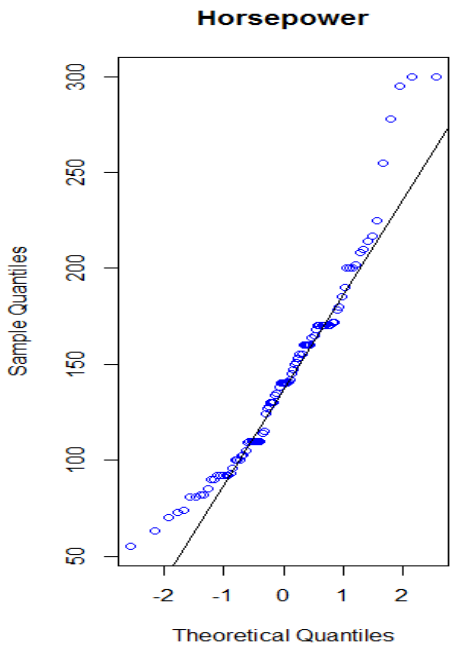
Ejemplo. Cont.



Test de correlación lineal simple

Ejemplo. Cont.





Test de normalidad

```
        Shapiro-Wilk normality test

data:  Cars93$Horsepower
W = 0.93581, p-value = 0.0001916


        Shapiro-Wilk normality test

data:  Cars93$Weight
W = 0.97432, p-value = 0.06337
```

El análisis gráfico y el contraste de normalidad muestran que para la variable Horsepower no se puede asumir normalidad y que la variable Weight está en el límite. Siendo estrictos, este hecho excluye la posibilidad de utilizar el coeficiente de Pearson, dejando como alternativas el de Spearman o Kendall.

Test de correlación lineal simple

Ejemplo. Cont.

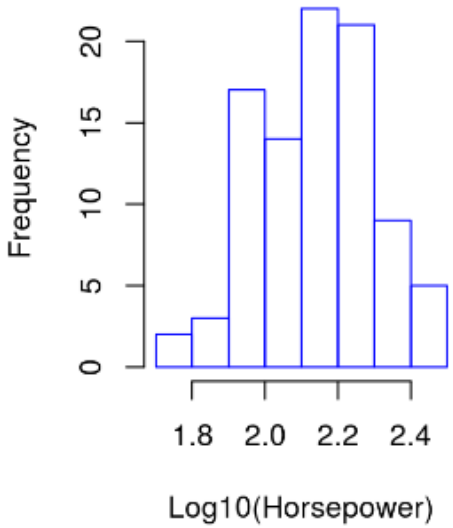
Sin embargo, dado que la distribución no se aleja mucho de la normalidad y de que el coeficiente de Pearson tiene cierta robustez, a fines prácticos sí que se podría utilizar siempre y cuando se tenga en cuenta este hecho en los resultados. Otra posibilidad es tratar de transformar las variables para mejorar su distribución.

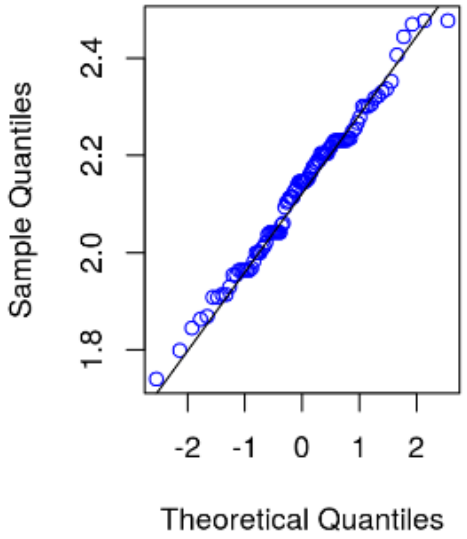
Ejemplo en código R - Análisis de la normalidad

```
# Representación gráfica
par(mfrow = c(1, 2))
hist(log10(Cars93$Horsepower), breaks = 10, main = "", xlab =
"Log10(Horsepower)", border = "blue")

par(mfrow = c(1, 2))
qqnorm(log10(Cars93$Horsepower), main = "", col = "blue")
qqline(log10(Cars93$Horsepower))

# Test de hipótesis para el análisis de normalidad
par(mfrow = c(1, 1))
shapiro.test(log10(Cars93$Horsepower))
```

Test de correlación lineal simple

Ejemplo. Cont.






```
        Shapiro-Wilk normality test

data:  log10(Cars93$Horsepower)
W = 0.98761, p-value = 0.5333
```

La trasformación logarítmica de la variable Horsepower consigue una distribución de tipo normal.

Test de correlación lineal simple

Ejemplo. Cont.

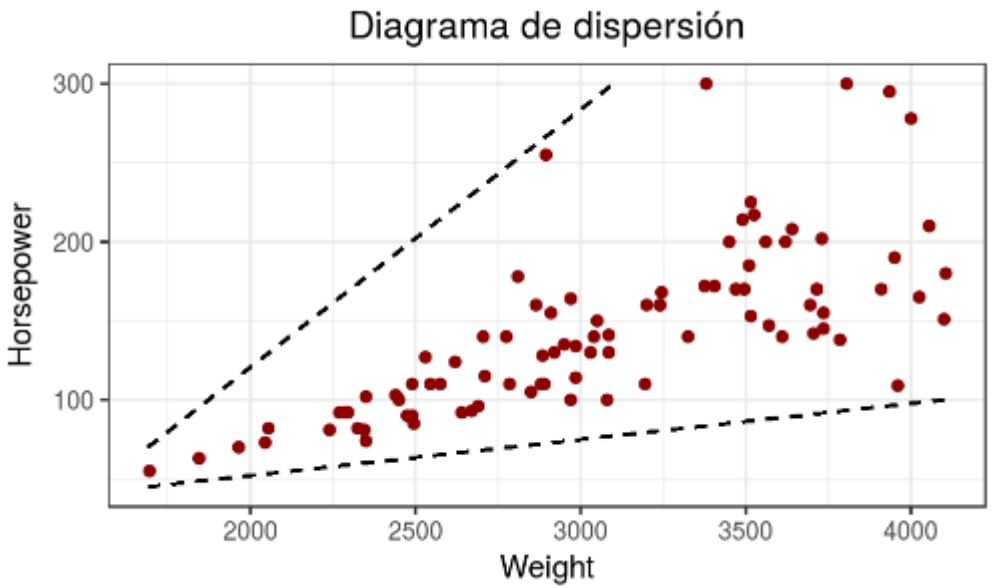
La homocedasticidad implica que la varianza se mantenga constante. Puede analizarse de forma gráfica representando las observaciones en un diagrama de dispersión y viendo si mantiene una homogeneidad en su dispersión a lo largo del eje X. Una forma cónica es un claro indicativo de falta de homocedasticidad.

Ejemplo en código R - Análisis de la homocedasticidad

```
# Representación gráfica
ggplot(data = Cars93, aes(x = Weight, y = Horsepower)) +
    geom_point(colour = "red4") +
    geom_segment(aes(x = 1690, y = 70, xend = 3100, yend =
    300),linetype="dashed") +
    geom_segment(aes(x = 1690, y = 45, xend = 4100, yend =
    100),linetype="dashed") +
    ggtitle("Diagrama de dispersión") +
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5))
```

Test de correlación lineal simple

Ejemplo. Cont.



Tal como muestra el diagrama de dispersión generado, sí hay un patrón cónico. Esto debe de tenerse en cuenta si se utiliza el coeficiente de Pearson puesto que viola una de sus condiciones.

Test de correlación lineal simple

Ejemplo. Cont.

Cálculo de la correlación y el test respectivo.

Debido a la falta de homocedasticidad, los resultados generados por Pearson no son precisos, desde el punto de vista teórico Spearman o Kendall son más adecuados. Sin embargo, en la bibliografía emplean Pearson, así que se van a calcular tanto Pearson como Spearman.

Ejemplo en código R - Correlación

```
# Correlación de Pearson
cor.test(x = Cars93$Weight, y = log10(Cars93$Horsepower),
        alternative = "two.sided", conf.level = 0.95, method = "pearson")
# Correlación de Spearman
cor.test(x = Cars93$Weight, y = log10(Cars93$Horsepower),
        alternative = "two.sided", conf.level = 0.95, method = "spearman")
```

Test de correlación lineal simple

Ejemplo. Cont.

Cálculo de la correlación y el test respectivo.

```
        Pearson's product-moment correlation

data:  Cars93$Weight and log10(Cars93$Horsepower)
t = 13.161, df = 91, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7256502 0.8699014
sample estimates:
      cor
0.809672


        Spearman's rank correlation rho

data:  Cars93$Weight and log10(Cars93$Horsepower)
S = 26239, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8042527

Warning message:
In cor.test.default(x = Cars93$Weight, y = log10(Cars93$Horsepower),  :
  Cannot compute exact p-value with ties
```

Por muy alto que sea un coeficiente de correlación, si no es significativa se ha de considerar inexistente.

Ambos coeficientes de correlación son significativos (p_value ≈ 0).

EL MODELO DE REGRESIÓN LINEAL SIMPLE

Introducción

- El Análisis de Regresión tiene como objetivo estudiar la relación entre variables.
- Permite expresar dicha relación en términos de una ecuación que conecta una variable de respuesta Y, con una variable explicativa X.
- Finalidad:
  - Determinación explícita del funcional que relaciona las variables. (Predicción)
  - Comprensión por parte del analista de las interrelaciones entre las variables que intervienen en el análisis.

Introducción

Datos Cuantitativos

REGRESION LINEAL SIMPLE

Objetivo 1
Determinar si dos variables están asociadas y en qué sentido se da la asociación

Determinar si existe relación entre las variables x e y:
Análisis de Correlación

x	y
$x_{(1)}$	$y_{(1)}$
$x_{(2)}$	$y_{(2)}$
⋮	⋮
$x_{(n)}$	$y_{(n)}$

Objetivo 2
Estudiar si los valores de una variable pueden ser utilizados para predecir el valor de la otra

Estudiar la dependencia de una variable respecto de la otra:
Modelo de Regresión

Términos clave

Variable Respuesta (= variable dependiente o de respuesta)
Variable Explicativa (= variable independiente o explicativa)
Relación Lineal (modelo lineal)
Parámetros (intercepto y pendiente)
Intercepto (respuesta media)
Pendiente (efecto de la variable explicativa sobre la respuesta)
Error (residuo)

Metodología

1. Formular el modelo: Identificar las variables respuesta y explicativa
2. "Comprobar" si son ciertas las hipótesis de linealidad y homocedasticidad
   - Correlación: Diagrama de dispersión
   - Transformación de los datos
3. Estimar los parámetros del modelo
4. Hacer el contraste de la regresión

$$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$$

H_0 : **No** existe **relación lineal** entre Y y X

H_0 : El **modelo no sirve** para explicar la respuesta

5. Bondad de ajuste del modelo: Coeficiente de determinación
6. Diagnosis del modelo con los residuos. ¿Se cumple la hipótesis de normalidad?
7. Hay alguna otra variable explicativa que pueda ser relevante y que podamos medir en los individuos de la muestra?

   SI ⟶ Regresión lineal múltiple

8. Hacer predicciones con el modelo

Especificación o formulación del modelo

Las técnicas de Regresión lineal simple parten de una teoría sólida para relacionar dos variables cuantitativas:

La variable explicativa (X)
La variable dependiente a explicar (Y)

Y tratan de explicar la **Y** mediante una función lineal de los valores de la **X** representada por la recta

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

Para ello dispondremos:

De un modelo de probabilidad (la Normal)
y de n pares de datos (x_i, y_i) que suponemos que provienen del modelo establecido y que se representan como una nube de puntos

Especificación o formulación del modelo

Muestra aleatoria

Se obtendrán n parejas de observaciones

$$Y_i = \beta_0 + \beta_1 x_i + U_i$$

donde

$$U_i \longrightarrow N(0, \sigma)$$

i = 1,2,…,n v.a. independientes

Especificación o formulación del modelo

Parámetros del modelo

β_0 Representa el valor medio de la respuesta (y) cuando la variable explicativa (x) vale cero (intersección de la recta con el eje y)

β_1 Representa el incremento de la respuesta media (y) cuando la variable explicativa (x) aumenta en una unidad (pendiente de la recta)

σ^2 Representa la variabilidad respecto a la recta

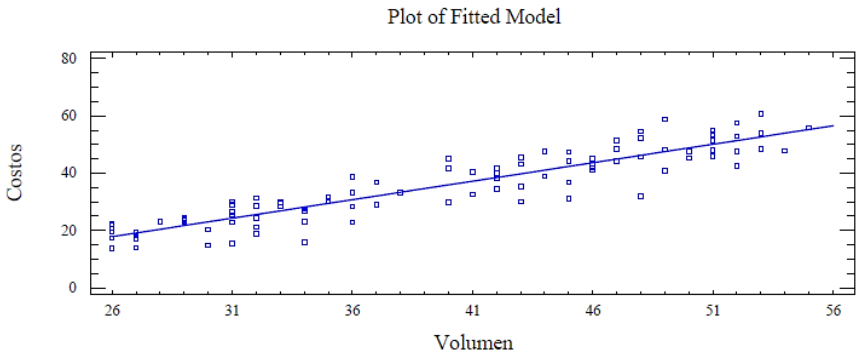
Especificación o formulación del modelo

Hipótesis básicas del modelo

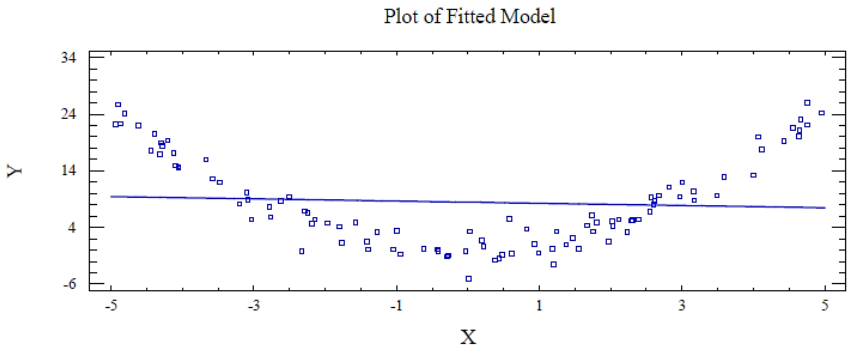
Linealidad

La relación que existente entre X e Y es lineal,
$y = \beta_0 + \beta_1 x$

Los datos deben ser razonablemente rectos



Si no, la recta de regresión no representa la estructura de los datos



Especificación o formulación del modelo

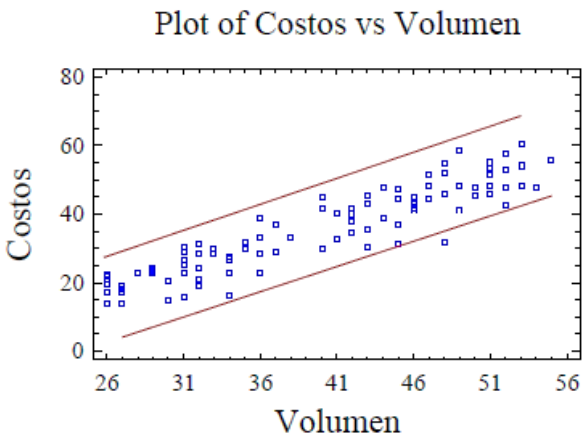
Hipótesis básicas del modelo

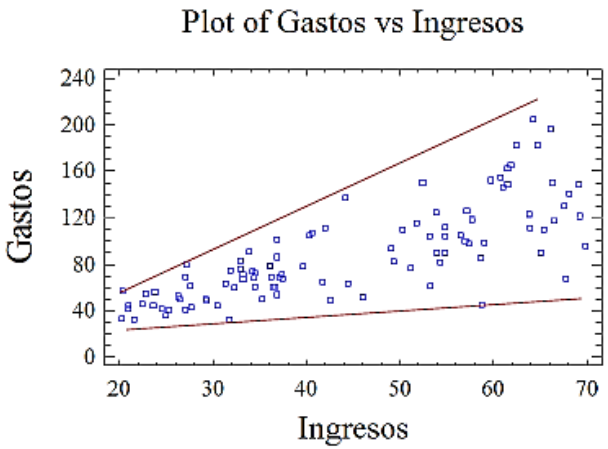
Homocedasticidad

La varianza de los errores es constante,
$Var[\mu_i] = \sigma^2$

La dispersión de los datos debe ser constante para que los datos sean homocedásticos

Si no se cumple, los datos son heterocedásticos





Especificación o formulación del modelo

Hipótesis básicas del modelo

Independencia	Las observaciones son independientes, $E[\mu_i \mu_j] = 0$

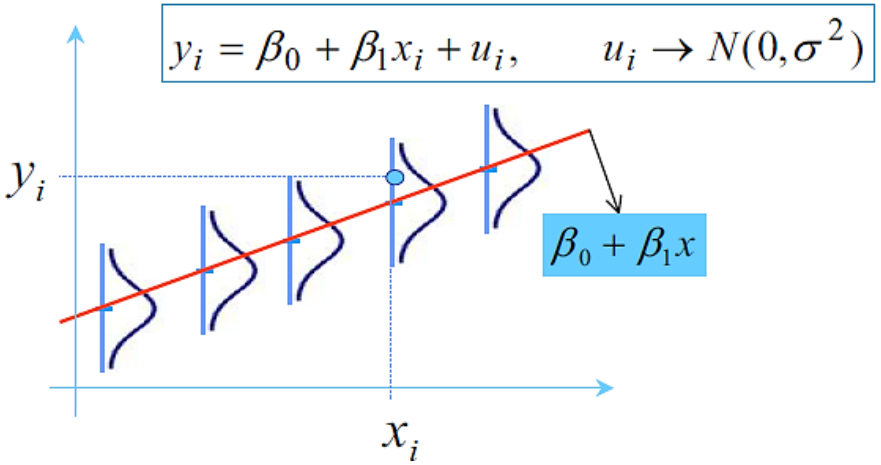
- Los datos deben ser independientes.
- Una observación no debe dar información sobre las demás.
- Habitualmente, se sabe por el tipo de datos si son adecuados o no para el análisis.
- En general, las series temporales no cumplen la hipótesis de independencia.

Especificación o formulación del modelo

Hipótesis básicas del modelo

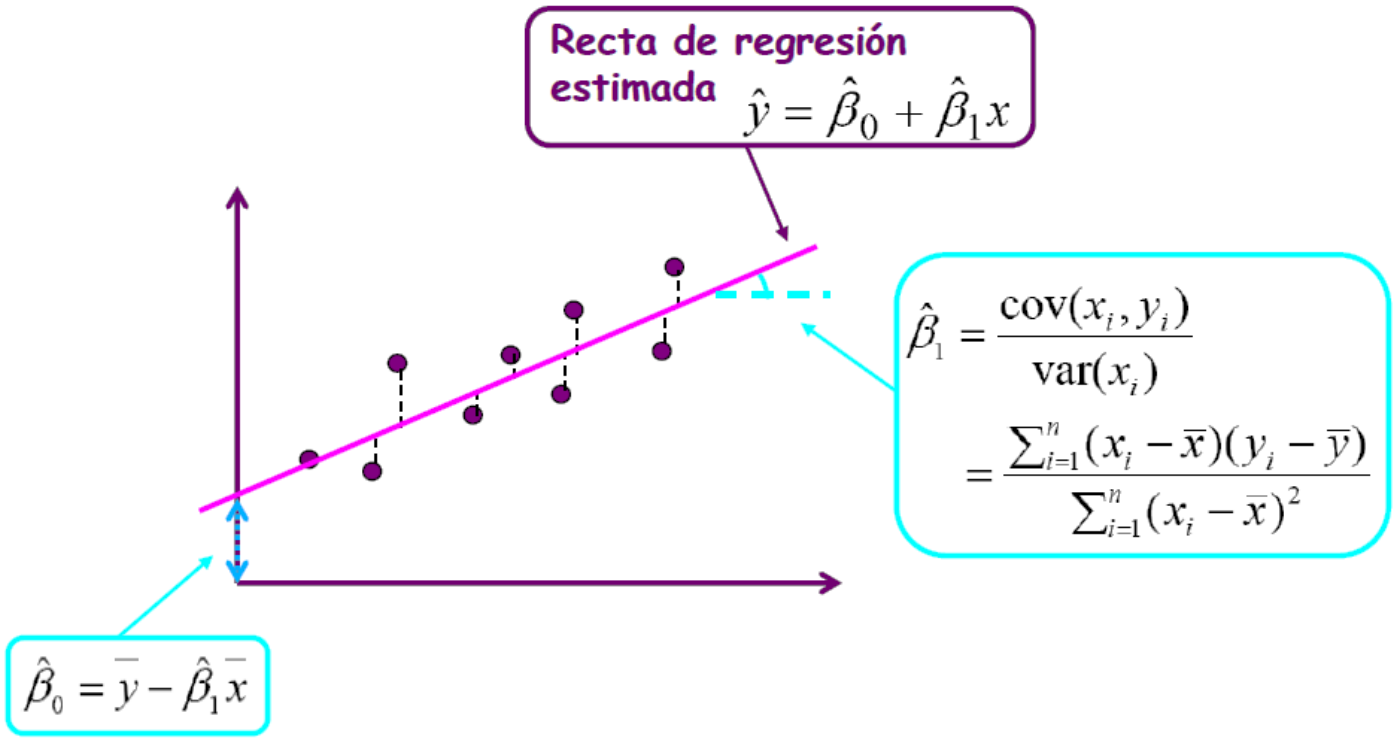
Normalidad	Los errores siguen una distribución normal, $\mu_i \sim N(0, \sigma^2)$

Se asume que los datos son normales a priori. El componente aleatorio de error *μ* explica las desviaciones de los puntos alrededor de la recta. Una respuesta particular *y* se describe usando el modelo probabilístico.



Estimadores de Mínimos Cuadrados

Ajuste de una recta a n pares de datos (x_i, y_i)
Estimación de los coeficientes de la recta



Estimadores de Mínimos Cuadrados

Ejemplo. Evaluar la relación de dependencia entre la tensión arterial sistólica y la edad, a partir de una muestra de 69 pacientes de cierto hospital.

Ejemplo en código R - Objetivo 1 del análisis

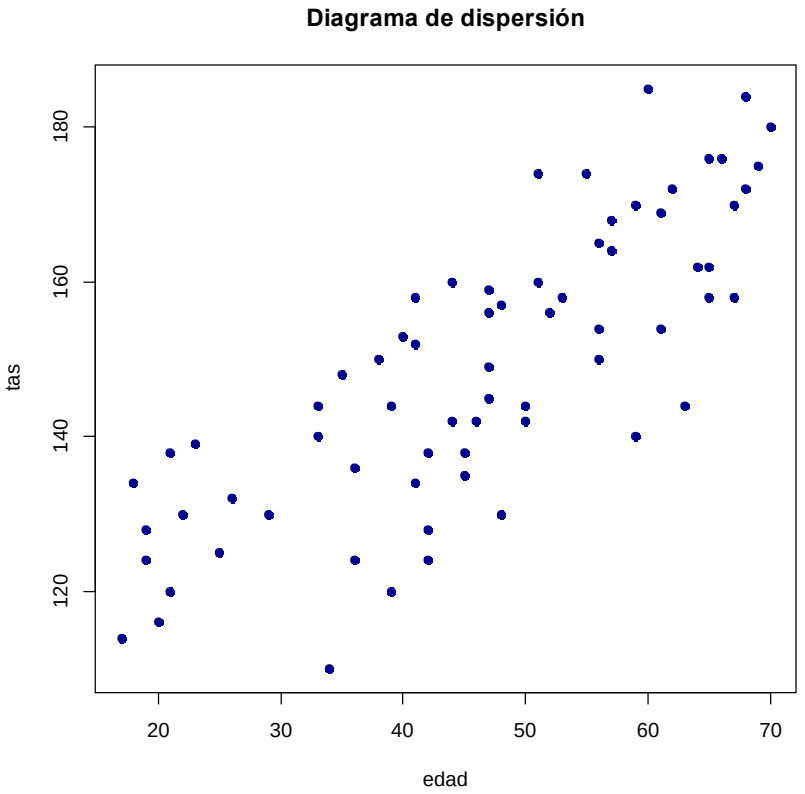
```
library(foreign)
pacientes <- read.spss(file="pacientes.sav", to.data.frame=TRUE)
attach(pacientes)

# Análisis univariado
summary(pacientes)

# Correlación, prueba de independencia y diagrama de dispersión
cor.test(tas, edad, alternative = "two.sided", method = "pearson")
plot(edad, tas, col="blue4", main="Diagrama de dispersión", pch = 19)

# Formulación modelo. Y=tensión, X=edad
# Prueba de normalidad en la variable dependiente
shapiro.test(tas)
```

Estimadores de Mínimos Cuadrados

Ejemplo. Cont.

- Resultados del análisis univariado y de correlación

```
      tas              edad
 Min.   :110.0   Min.   :17.00
 1st Qu.:135.0   1st Qu.:36.00
 Median :149.0   Median :47.00
 Mean   :148.7   Mean   :46.13
 3rd Qu.:162.0   3rd Qu.:59.00
 Max.   :185.0   Max.   :70.00

        Pearson's product-moment correlation

data:  tas and edad
t = 11.027, df = 67, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6991356 0.8736082
sample estimates:
      cor
0.8029505
```



Como se observa una relación lineal, entonces el modelo a estimar es:

$$tensión = \beta_0 + \beta_1 edad + \mu$$

```
        Shapiro-Wilk normality test

data:  tas
W = 0.98207, p-value = 0.4252
```

Estimadores de Mínimos Cuadrados

El objeto lm

Algunos elementos importantes

- `coefficients`: Valores de $\hat{\beta}_0$ y $\hat{\beta}_1$
- `fitted.values`: Valores de $\hat{y}_i$
- `residuals`: Valores de los residuos (no tipificados)
- `call`: Llamada a la función `lm` que generó el objeto
- `model`: Información sobre el modelo (datos, etc.)

Funciones para acceder a estos elementos

R tiene una serie de funciones que permiten acceder a estos elementos individualmente: `coef(lm, ...)`, `fitted(lm, ...)`, `residuals(lm, ...)`, `vcov(lm, ...)`.

Inferencia sobre el modelo

- ☑ Inicialmente se observan las estimaciones puntuales de los coeficientes de regresión.
- ☑ Usando intervalos de confianza podemos obtener una medida de la precisión de dichas estimaciones.
- ☑ Usando contrastes de hipótesis podemos comprobar si un determinado valor puede ser el auténtico valor del parámetro. Significación individual.
- ☑ Usando contrastes de hipótesis podemos comprobar la significación global del modelo. Contraste de la regresión ANOVA.

Inferencia sobre el modelo

Estimación puntual de los parámetros del modelo

$$\hat{\beta}_0 = \bar{y} - \frac{cov}{v_x}\bar{x} \qquad \hat{\beta}_1 = \frac{cov}{v_x}$$

$$\hat{\sigma}^2 = S_R^2 = \frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Estimación por intervalos de los parámetros del modelo

$$IC_{1-\alpha}(\beta_0) = \left(\hat{\beta}_0 \pm t_{n-2;\alpha/2}\, S_R \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n v_x}}\right)$$

$$IC_{1-\alpha}(\beta_1) = \left(\hat{\beta}_1 \pm t_{n-2;\alpha/2}\, S_R \sqrt{\frac{1}{n v_x}}\right)$$

$$IC_{1-\alpha}(\sigma^2) = \left(\frac{(n-2)S_R^2}{\chi^2_{n-2;\alpha/2}} ; \frac{(n-2)S_R^2}{\chi^2_{n-2;1-\alpha/2}}\right)$$

Contrastes de la regresión: t

$H_0 : \beta_1 = 0$ (Los valores de la X no influyen en los valores de Y en una relación lineal)

$H_1 : \beta_1 \neq 0$ (El modelo es válido)

Con nivel de significación α rechazamos H_0 si el cero no está en el intervalo de confianza:

Error típico

$$IC_{1-\alpha}(\beta_1) = \left(\hat{\beta}_1 \pm t_{n-2;\alpha/2} S_R \sqrt{\frac{1}{n v_x}}\right)$$

Contrastes de la regresión: ANOVA

Descomposición de la variabilidad en regresión

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$$y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i} + \underbrace{e_i}_{y_i - \hat{y}_i}$$

$$y_i = \hat{y}_i + (y_i - \hat{y}_i) \text{ (restando } \bar{y})$$

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \text{ (elevando al cuadrado y sumando)}$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SCT}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SCE}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SCR}}$$

SCT Suma de cuadrados total
(variabilidad total de la y)

SCE Suma de cuadrados explicada
(variabilidad de y debida a su relación lineal con la x)

SCR Suma de cuadrados residual
(variabilidad de y respecto a la recta ajustada)

Contrastes de la regresión: ANOVA

Tabla ANOVA: Contraste global del modelo

Suma de cuadrados	G.l.	Varianza	Estadístico	p-valor
$SCE = \sum_i (\hat{y}_i - \bar{y})^2$	1	$\frac{SCE}{1}$	$F = \frac{SCE/1}{SCR/(n-2)}$	¿?
$SCR = \sum_i (y_i - \hat{y}_i)^2$	$n-2$	$\frac{SCR}{n-2}$		
$SCT = \sum_i (y_i - \bar{y})^2$	$n-1$			

$$SCT = nv_y \qquad SCR = nv_y(1 - r^2)$$

H_0 : **El modelo de regresión lineal NO sirve para explicar la respuesta**
H_1 : **El modelo de regresión lineal SI sirve para explicar la respuesta**

A nivel de significación α, rechazamos cuando

$$F > F_{1,n-2,\alpha}$$

Bondad de ajuste del modelo

Coeficiente de determinación – R^2

Valoración de cuánto se ajustan los puntos a la recta.

El **COEFICIENTE DE DETERMINACIÓN** es la proporción de variabilidad explicada por la regresión

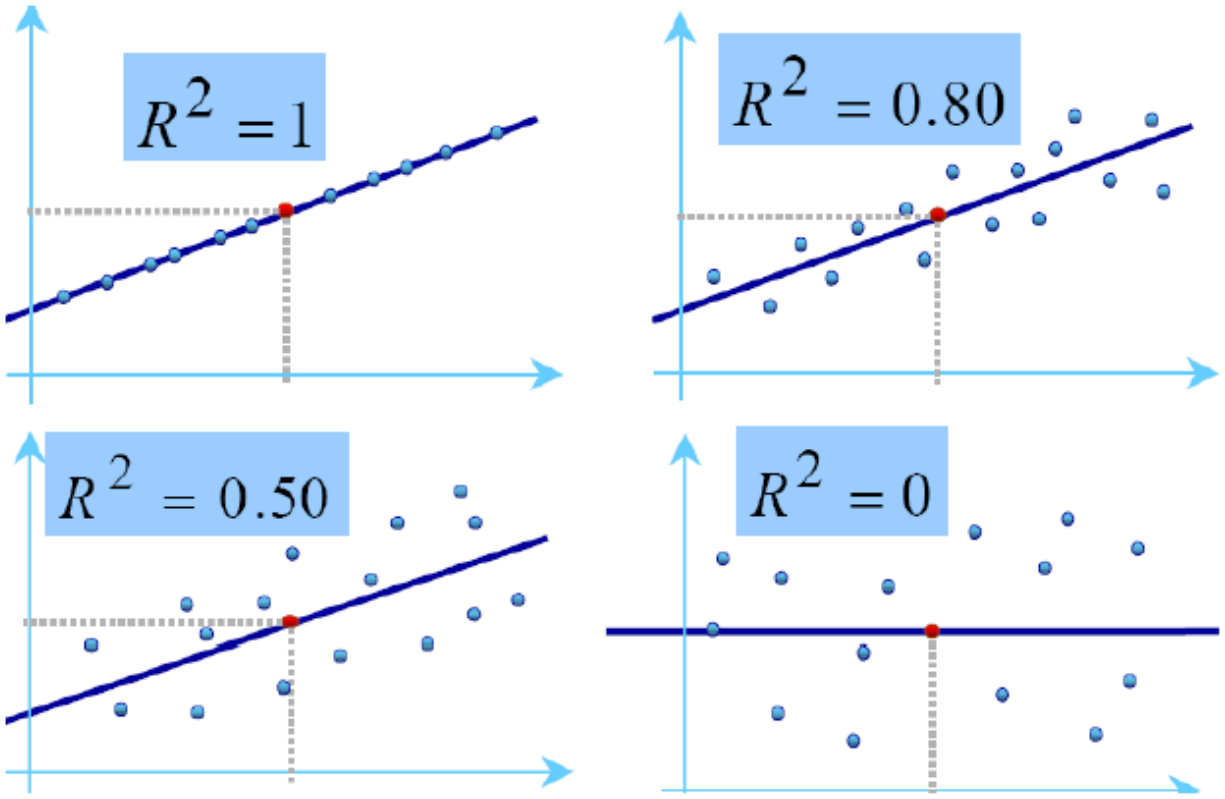
$$R^2 = SCE / SCT$$

En regresión simple el COEFICIENTE DE DETERMINACIÓN coincide con el COEFICIENTE DE CORRELACIÓN AL CUADRADO

$$R^2 = r^2$$

Bondad de ajuste del modelo

Coeficiente de determinación – R^2



Bondad de ajuste del modelo

Coeficiente de determinación ajustado R^2-Adj

- R^2 presenta el inconveniente de que a medida que vamos incrementando el número de variables que participan en el modelo (será el caso propio del análisis multivariable) mayor es su valor, de ahí que puede que sobrestime el verdadero valor R de la población.

- Por esta razón, algunos autores recomiendan utilizar el Coeficiente de Determinación Ajustado $\overline{R}^2$ pues éste no aumenta, necesariamente, a medida que añadimos variables a la ecuación. Viene dado por:

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

- Se debe tener en cuenta que tanto R^2 como R^2-Ajustado son estadísticas de muestra, y que no debemos depender únicamente de sus valores para decidir si un modelo es útil o no para predecir la variable respuesta.

Modelo de regresión lineal simple

Comentarios

- El contraste de la regresión supone que la relación (más o menos fuerte) es LINEAL. Por tanto, si no rechazamos la hipótesis nula lo único que podemos decir es que no hemos encontrado evidencia de que exista una relación lineal, puede existir una relación no lineal…

- En REGRESIÓN SIMPLE el contraste F coincide exactamente con el contraste de la t para el coeficiente de la variable regresora.

Modelo de regresión lineal simple

La función lm( ) toma la siguiente forma general,

lm(dependiente ~ predictora(s), data = dataFrame, na.action = "acción")

donde na.action es opcional, puede ser útil si tenemos valores perdidos.

Ejemplo. Continuamos con el ejemplo anterior de los pacientes.

Ejemplo en código R - Objetivo 2 del análisis

```
# Ajuste del modelo
regresion <- lm(tas ~ edad)
summary(regresion) # Muestra resultados
confint(regresion, level = 0.95) # Intervalos de confianza
aov(regresión) # ANOVA

## Otras opciones del análisis
regre2 <- lm(tas ~ edad - 1) # Omite el intercepto
summary(regre2)$coefficients
```

Modelo de regresión lineal simple

Ejemplo. Cont. Pacientes.

- Resultados

```
> regresion <- lm(tas ~ edad)
> summary(regresion) # Muestra resultados

Call:
lm(formula = tas ~ edad)

Residuals:
    Min      1Q  Median      3Q     Max
-26.794  -7.024   1.960   8.190  22.634

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.3527     4.3261   23.89   <2e-16 ***
edad          0.9836     0.0892   11.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.09 on 67 degrees of freedom
Multiple R-squared:  0.6447,    Adjusted R-squared:  0.6394
F-statistic: 121.6 on 1 and 67 DF,  p-value: < 2.2e-16
```

Los parámetros de la ecuación de la recta de mínimos cuadrados que relaciona la presión arterial sistólica en función de la edad de los pacientes vienen dados por ´Coefficients´ de la tabla de la salida anterior. se obtiene la recta, Presión = 0.9836 * Edad + 103.3527.

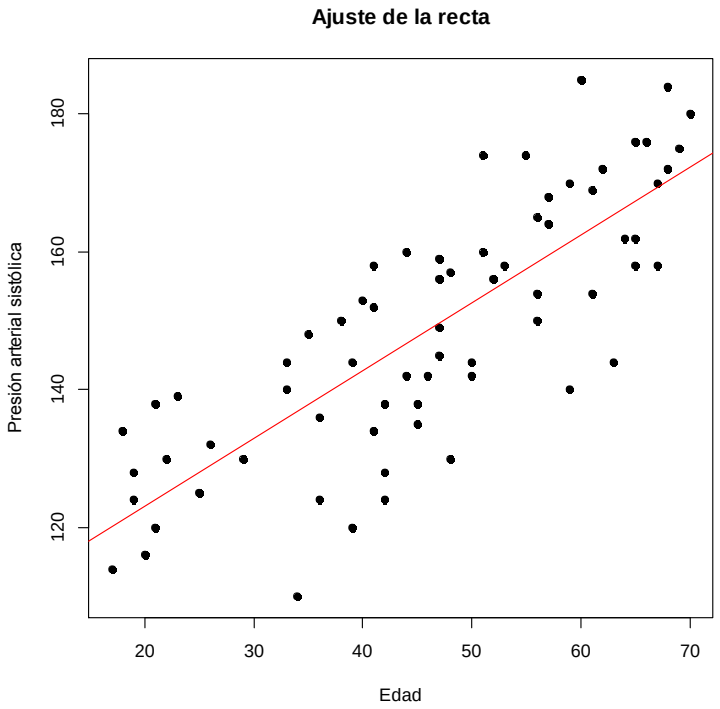
Modelo de regresión lineal simple

Ejemplo. Cont. Pacientes.

Los siguientes comandos representan la nube de puntos (comando plot) y añaden la representación gráfica de la recta de mínimos cuadrados (comando abline aplicado al objeto generado por lm):

Ejemplo en código R - Objetivo 2

```
# Ajuste de la recta de regresión
plot(tas ~ edad, main= "Ajuste de la
        recta", xlab='Edad', ylab='Presión
        arterial sistólica', pch = 19)
abline(regresion, col="red")
```



Comunicación constante con la Escuela del INEI

Correo de la Dirección Técnica de la ENEI
Sr. Eduardo Villa Morocho (Eduardo.villa@inei.gob.pe)

Coordinación Académica
Sra. María Elena Quirós Cubillas (Maria.Quiros@inei.gob.pe)

Correo de la Escuela del INEI
enei@inei.gob.pe

Área de Educación Virtual
Sr. Gonzalo Anchante (gonzalo.anchante@inei.gob.pe)