Regresión Lineal Simple en R

*Estudiar la relación entre dos variables*

*Abril 2018*

FUENTE: Toda la información usada en este tutorial corresponde al equipo *[00RTeam](http://gauss.inf.um.es/00rteam.html)* de la Universidad de Murcia y está disponible en <http://gauss.inf.um.es/>

*Tarea 2. [0011] DEFAD. Métodos de contraste de hipótesis y diseño de experimentos.*

Apartado 2

Una empresa quiere saber si existe relación entre el salario de un trabajador y las ausencias del mismo al trabajo. Para el estudio se dividió el salario en distintas categorías y se eligió aleatoriamente un grupo de trabajadores para determinar el número de días que habían faltado en los últimos tres años.

¿Se puede construir un modelo que relacione la categoría del salario y las ausencias en el trabajo? (Trabaja con el fichero “william.csv”)

Nuestros datos

Cargamos los datos, y comprobamos nuestros datos antes de empezar con los análisis. Tenemos una muestra de 15 observaciones y 2 variables cuantitativas (salarioy ausencias). Este tamaño de muestra pequeño se deberá tener en cuenta en los anaálisis y en la interpretación de los resultados.

df <- read.table(file = "william.csv",

header = TRUE,

sep = ";",

dec = ".",

encoding = "UTF-8",

stringsAsFactors = FALSE) *# cargamos los datos*

head( df ) *# comprobamos que se han leido bien*

## salario ausencias

## 1 11 18

## 2 10 17

## 3 8 29

## 4 5 36

## 5 9 11

## 6 7 28

str ( df ) *# detalle de las variables*

## 'data.frame': 15 obs. of 2 variables:

## $ salario : int 11 10 8 5 9 7 3 11 8 7 ...

## $ ausencias: int 18 17 29 36 11 28 35 14 20 32 ...

**a)** Comprueba la normalidad de la variable explicativa (salario) y calcula la correlación entre salario y ausencias.

**Comprobar la normalidad** Se realizar con un *contraste para normalidad*. En este contraste *la hipótesis nula es la hipótesis de normalidad*, esto es, no hay diferencias entre nuestra distribución y una distribución normal con esa media y esa desviación típica. Para contrastar la normalidad usamos el test de Shapiro-Wilk, con la función shapiro.test( ), que funciona bien con muestras pequeñas (menores a 50).

shapiro.test( df$salario )

##

## Shapiro-Wilk normality test

##

## data: df$salario

## W = 0.93541, p-value = 0.3281

Interpretación:

Se obtiene un p-value mayor de 0.05 (0.3281), no podemos rechazar la hipótesis nula (hipótesis de normalidad). Por lo tanto, podemos concluir que nuestros datos cumplen el supuesto de normalidad.

**Estudiar la correlación** Previo a estimar el modelo de regresión lineal simple, se examina la correlación entre las variables. La *correlación* mide el grado de asociación entre las dos variables. Por su parte el *coeficiente de correlación lineal de Pearson* contrasta la correlación lineal entre las dos variables.

cor.test( df$salario, df$ausencias )

##

## Pearson's product-moment correlation

##

## data: df$salario and df$ausencias

## t = -7.4737, df = 13, p-value = 4.672e-06

## alternative hypothesis: true correlation is not equal to 0

## 95 percent confidence interval:

## -0.9668476 -0.7211085

## sample estimates:

## cor

## -0.9006674

Interpretación:

Las variables presentan un alto grado de correlación (-0.9006674), la correlación negativa indica que están asociadas de forma inversa, esto es, valores altos de una de las variables se corresponden con valores bajos de la otra. Con un coeficiente de Pearson significativo (menor de 0.05) podemos asumir correlacion lineal entre las variables `salario` y `ausencias`.

**Comprobar gráficamente la relación lineal entre las variables** El diagrama de dispersión representa los valores de la variable independiente (x), junto con los de la variable dependiente (y). Permite determinar de una forma visual si hay una relación lineal entre ambas.

Interpretación:

En el diagrama se aprecia un patrón lineal entre las variables `salario` y `ausencias`.

plot( df$salario, df$ausencias ) *# ( x, y )*

**EN CONCLUSIÓN**: Tras el estudio de la correlación se puede asumir una relación lineal entre ambas variables. Es adecuado plantear el modelo de regresión lineal simple.

**b)** Realiza el ajuste del modelo.

Con el análisis de **Regresión Lineal Simple** queremos construir un modelo que relacione la categoría del salario y las ausencias en el trabajo.

* Nuestra variable de interés es ausencias (variable explicada) y queremos conocer su comportamiento a partir del salario (variable explicativa).
* La recta de regresión será la función que mejor explique la relación entre ambas variables.

Hacemos la regresión lineal la función lm( y ~ x, data). Examinamos los coeficientes del modelo de regresión con la función summary( ).

modeloR <- lm( ausencias ~ salario, *# ( y ~ x )*

data = df )

summary( modeloR ) *# para ver el resumen del modelo*

##

## Call:

## lm(formula = ausencias ~ salario, data = df)

##

## Residuals:

## Min 1Q Median 3Q Max

## -9.516 -3.053 1.428 2.961 5.475

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 47.6002 3.0789 15.460 9.50e-10 \*\*\*

## salario -3.0094 0.4027 -7.474 4.67e-06 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 4.294 on 13 degrees of freedom

## Multiple R-squared: 0.8112, Adjusted R-squared: 0.7967

## F-statistic: 55.86 on 1 and 13 DF, p-value: 4.672e-06

**Interpretación**

El modelo ajustado es significativo: Los coeficientes de regresión son 47.6002 y -3.0094, estos parámetros son significativos, con p-valor menor de 0.05 (9.50e-10, 4.67e-06). El error estándar para cada parámetro es 3.0789 y 0.4027 respectivamente. La R2�2 ajustada es 0.7967, que indica un buen ajuste del modelo (próximo a 1),

El modelo tiene la forma de:

ausencias=47.60−3.01×salario���������=47.60−3.01×�������

* Los coeficientes de regresión son 47.60 y -3.01.
* 47.60 es el valor medio de la variable dependiente (ausencias) cuando la predictora es cero (salario).
* 3.01 es el efecto medio (negativo) sobre la variable dependiente Y (ausencias) al aumentar en una unidad el valor de la predictora X (salario). Esto es, variación que se produce en Y (-3.01) por cada unidad de incremento en X.

**EN CONCLUSIÓN**: Existe una relación lineal negativa entre las variables: cuando aumentamos en una unidad el salario, las ausencias disminuyen en 3.01 unidades. De forma que por cada aumento de la categoría del salario, las ausencias de los trabajadores disminuyen en 3.01 días.

Por último, representamos la recta del modelo de regresión: que es la ecuación de una recta y = mx + n, con ordenada en el origen 47.60 (punto de corte con el eje Y), y con pendiente negativa -3.01.

plot( df$salario,

df$ausencias,

xlim = c( 0, 14 ),

ylim = c( 0, 70 ),

xlab = "Salario",

ylab = "Ausencias" )

abline( modeloR, col = "#e74c3c" ) *# añadir la recta*

**d)** Estudia la bondad de ajuste con la función anova( ) y explica el significado de los coeficientes obtenidos en la recta de regresión.

anova( modeloR )

## Analysis of Variance Table

##

## Response: ausencias

## Df Sum Sq Mean Sq F value Pr(>F)

## salario 1 1030.01 1030.01 55.857 4.672e-06 \*\*\*

## Residuals 13 239.72 18.44

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Interpretación**

La variabilidad del modelo se puede descomponer como SST = SSM + SSR. SST es la variación total. SSM es la variación explicada por el modelo. SSR es la variación no explicada por el modelo. Para nuestros datos tenemos: SSM 1030.01, SSR 239.72. Por lo tanto la variabilidad explica por el modelo de regresión es mayor que la que queda sin explicar (residuos).

Un estadístico F mayor de 1 indica un buen ajuste del modelo. El estadístico F (55.857) contrasta si el modelo tiene significativa capacidad predictiva. En el contraste la hipotesis nula es F = 1, con un p-valor menor de 0.05 (4.672e-06) se rechaza la hipótesis nula. Por lo tanto concluimos que el modelo tiene una capacidad predictiva significativa.

**d)** Realiza el diagnóstico del modelo. Normalidad de los residuos, homogeneidad de varianzas e incorrelación de los residuos.

**Obtenemos:**

* Valores ajustados: valores ajustados (valores de la variable respuesta) para las observaciones originales de la predictora.
* Residuos: diferencia entre valor observado de la respuesta y valor ajustado por el modelo.
* Estadísticos: residuos estudentizados del modelo ajustado.

df$fitted.modeloR <- fitted( modeloR ) *# valores ajustados*

df$residuals.modeloR <- residuals( modeloR ) *# residuos*

df$rstudent.modeloR <- rstudent( modeloR ) *# residuos estudentizados*

**Normalidad de los residuos** Comprobar el contraste de normalidad para los residuos estudentizados del modelo ajustado. Para contrastar la normalidad usamos el test de Shapiro-Wilk, con la función shapiro.test( ), que funciona bien con muestras pequeñas (menores a 50).

shapiro.test( df$rstudent.modeloR )

##

## Shapiro-Wilk normality test

##

## data: df$rstudent.modeloR

## W = 0.91538, p-value = 0.1637

Interpretación:

Se obtiene un p-value mayor de 0.05 (0.1637), no podemos rechazar la hipótesis nula (hipótesis de normalidad). Por lo tanto, podemos suponer normalidad de los residuos.

**Homogeneidad de varianzas** Comprobar la homocedasticidad para los residuos estudentizados del modelo ajustado. Para contrastarla utilizamos la función bptest( ),

**library**(lmtest)

## Loading required package: zoo

##

## Attaching package: 'zoo'

## The following objects are masked from 'package:base':

##

## as.Date, as.Date.numeric

bptest( modeloR )

##

## studentized Breusch-Pagan test

##

## data: modeloR

## BP = 0.52916, df = 1, p-value = 0.467

Interpretación:

Con un p-value = 0.467, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto suponemos homogeneidad de varianzas.

**Incorrelación de los residuos** Comprobar la incorrelación para los residuos estudentizados del modelo ajustado. Se realiza a través del Test de Durbin-Watson (asume bajo la hipótesis nula que no existe correlación). La función dwtest( ) realiza este contraste directamente sobre los residuos estudentizados.

dwtest( ausencias ~ salario,

alternative = "two.sided",

data = df)

##

## Durbin-Watson test

##

## data: ausencias ~ salario

## DW = 2.452, p-value = 0.3763

## alternative hypothesis: true autocorrelation is not 0

Interpretación:

Con un p-value = 0.3763, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto suponemos incorrelación para los residuos estudentizados del modelo ajustado.