Práctica dirigida 2

Samuel Huamaní Flores

2023-03-30

Ejemplo práctico de regresión lineal simple

Base de datos

La base de datos que vamos a trabajar es el `dataset` Boston del paquete `MASS` recoge la mediana del valor de la vivienda en 506 áreas residenciales de Boston. Junto con el precio, se han registrado 13 variables adicionales.

- `crim`: ratio de criminalidad per cápita de cada ciudad.
- `zn`: Proporción de zonas residenciales con edificaciones de más de 25.000 pies cuadrados.
- `indus`: proporción de zona industrializada.
- `chas`: Si hay río en la ciudad (= 1 si hay río; 0 no hay).
- `nox`: Concentración de óxidos de nitrógeno (partes per 10 millón).
- `rm`: promedio de habitaciones por vivienda.
- `age`: Proporción de viviendas ocupadas por el propietario construidas antes de 1940.
- `dis`: Media ponderada de la distancias a cinco centros de empleo de Boston.
- `rad`: Índice de accesibilidad a las autopistas radiales.
- `tax`: Tasa de impuesto a la propiedad en unidades de $10,000.
- `ptratio`: ratio de alumnos/profesor por ciudad.
- `black`: 1000(Bk - 0.63)^2 donde Bk es la proporción de gente de color por ciudad.
- `lstat`: porcentaje de población en condición de pobreza.
- `medv`: Valor mediano de las casas ocupadas por el dueño en unidades de $1000s.

Regresión lineal simple

Estimación de los coeficientes de regresión

Se pretende predecir el valor de la vivienda en función del porcentaje de pobreza de la población.

```
library(readxl)
datos <- read_excel("base de datos.xlsx")
```

Como se trata de regresión lineal simple, entonces vamos escoger solo dos variables:

Variable independiente X: porcentaje de población en condición de pobreza

Variable dependiente Y: Valor mediano de las casas ocupadas por el dueño en unidades de $1000s.

```
X<-datos$lstat
Y<-datos$medv
data<-data.frame(X,Y)
```

Empleando la función `lm()` se genera un modelo de regresión lineal por mínimos cuadrados en el que la variable respuesta es medv (Y) y el predictor lstat (X).

La sintaxis de la función `lm` requiere indicar primero la ecuación que se desea estimar, indicando primero la variable dependiente y después del signo $\sim$ las variables independientes, en este caso al ser una regresión simple, solo se indica una variable, finalmente se debe indicar en la opción data el nombre del objeto donde tomará los datos en este caso el objeto que llamamos *data* :

```
modelo <- lm(Y ~ X, data=data)
modelo$coefficients
```

```
## (Intercept)           X
##  34.5538409  -0.9500494
```

De donde: $\widehat{\beta}_o = 34.56$ y $\widehat{\beta}_1 = -0.95$ Con estos datos, podemos decir que la ecuación de regresión estimada es:

$y = 34.55 - 0.95x$

El $\widehat{\beta}_1 = -0.95$ significa que si el porcentaje de la pobreza de la población se incremente en 1%, entonces en promedio el valor mediano de las casas ocupadas disminuye en 0.95 unidades.

Intervalo de confianza para los coeficientes de regresión

Los coeficientes estimados de una regresión son estadísticos y siguen una distribución t, por ello podemos calcular intervalos de confianza, los cuales pueden complementar el análisis en vez de solo utilizar las estimaciones puntuales.

Una forma sencilla de calcular los intervalos de confianza para los coeficientes es utilizar la función `coefci()`, de la paquetería `lmtest`, solo debemos indicar el objeto donde se encuentra la regresión y el nivel de confianza que sea desea:

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
coefci(modelo, level=.95)
```

```
##                 2.5 %     97.5 %
## (Intercept) 33.448457 35.6592247
## X           -1.026148 -0.8739505
```

Luego el intervalo de confianza para β_1 es $-1.026 \leq \beta_1 \leq -0.874$

La interpretación de este intervalo de confianza es: Dado el coeficiente de confianza de 95%, en 95 de cada 100 casos, los intervalos como $-1.026 \leq \beta_1 \leq -0.874$ contendrán al verdadero valor de β_1

Otra forma de obtener los coeficientes de regresión es mediante el siguiente código:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.7      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(printr) # Para presentar en tabla
```

```
## Registered S3 method overwritten by 'printr':
##   method                from
##   knit_print.data.frame rmarkdown
```

```
library(broom)
resultado<-modelo %>%
  tidy()
names(resultado)<-c("Variable", "beta", "Error_estandar","t","p_value")
resultado
```

Variable	beta	Error_estandar	t	p_value
(Intercept)	34.5538409	0.5626274	61.41515	0
X	-0.9500494	0.0387334	-24.52790	0

Prueba de hipótesis de la regresión

H_o: $\beta_1 = 0$ (El porcentaje de población en condición de pobreza no influye en el valor mediano de las casas ocupadas por el dueño.)

H_a: $\beta_1 \neq 0$ (El porcentaje de población en condición de pobreza influye en el valor mediano de las casas ocupadas por el dueño.)

```
modelo %>%  aov() %>%  summary()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## X             1  23244   23244   601.6 <2e-16 ***
## Residuals   504  19472      39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El punto crítico para esta prueba es $F_{0.95,1,504} = 3.86$ y dado que $F_{cal} = 601.6 > F_{0.95,1,504} = 3.86$, se rechaza la H_o; por lo tanto; al nivel de 0.05 de significancia, existe suficiente evidencia estadística para indicar que el valor mediano de las casas ocupadas por el dueño depende del porcentaje de población que se encuentra en condición de pobreza.

$p - valor < 0.05$, se rechaza la H_o.

Otra forma de describir la tabla de ANOVA es:

El estadístico F es de 601.6 y su valor-p es prácticamente cero, por lo tanto podemos rechazar la hipótesis nula de que los coeficientes (β_0 y β_1) son iguales cero, concluyendo que el modelo en general es significativo.

Estimación de la variabilidad

El valor de σ^2 es la varianza de la regresión y se estima con cuadrado medio del error.

```
summary(modelo)$sigma^2 # Varianza de la regresión
```

```
## [1] 38.63568
```

```
summary(modelo)$sigma # desviación estandar de regresión o error estándar de regresión.
```

```
## [1] 6.21576
```

Intervalo de confianza para la varianza de la regresión con $\alpha = 0.10$

```
LI.var = sum(modelo$residuals^2)/qchisq(0.95,df=504)
LI.var
```

```
## [1] 34.93841
```

```
LS.var = sum(modelo$residuals^2)/qchisq(0.05,df=504)
LS.var
```

```
## [1] 42.99118
```

Pruebas de hipótesis específicas

¿Un incremento de 1% porcentaje de población en condición de pobreza provocará un decaimiento del valor mediano de las casas ocupadas por el dueño en más de 0.90 unidades?

H_o: $\beta_1 \geq -0.90$

H_a: $\beta_1 < -0.90$

$\alpha = 0.05$

$t_{cal} = \frac{\widehat{\beta_1} - \beta_o}{S_{\widehat{\beta_1}}} = \frac{-0.95005-(-0.90)}{0.03873} = -1.2923$

En este caso no se rechaza la H_o, pues $t_{cal} = -1.2923 > t_{0.05,504} = -1.6479$

Ajuste del modelo

Coeficiente de determinación

```
summary(modelo)$r.squared
```

```
## [1] 0.5441463
```

$R^2 = 54.41$ lo que significa que el predictor X (porcentaje de población en condición de pobreza) empleado en el modelo es capaz de explicar el 54.44% de la variabilidad observada en el precio de las viviendas.

Estimación y predicción

Estimación de la media μ_{Y/X_o}

```
predConf<-predict(modelo,interval="confidence",level=0.95)
ICMeanY<-cbind(X, Y,predConf)
head(ICMeanY)
```

X	Y	fit	lwr	upr
4.98	24.0	29.82260	29.02530	30.61989
9.14	21.6	25.87039	25.26525	26.47553
4.03	34.7	30.72514	29.87348	31.57681
2.94	33.4	31.76070	30.84359	32.67780
5.33	36.2	29.49008	28.71208	30.26808
5.21	28.7	29.60408	28.81952	30.38865

Por ejemplo para $x_o = 4.98$ se obtiene el intervalo de confianza $29.03 \leq \mu_{Y/x_0} \leq 30.62$. Esto es, si el porcentaje de población en condición de pobreza es de 4.98%, se estima que en promedio el valor mediano de las casas ocupadas por el dueño se encuentre entre 29.03 y 30.62

Tambien para un valor específico de x_o se puede hacer la estimación; por ejemplo para $x_o = 10.5$

```
modelo %>% predict(data.frame(X=10.5),
interval = "confidence",
level = 0.95)
```

fit	lwr	upr
24.57832	24.01125	25.1454

Lo que significa que para $x_o = 10.5$ se obtiene el intervalo de confianza $24.01 \leq \mu_{Y/x_0} \leq 25.15$. Esto es, si el porcentaje de población en condición de pobreza es de 10.5%, se estima que en promedio el valor mediano de las casas ocupadas por el dueño se encuentre entre 24.04 y 25.15.

Para ver la gráfica de dispersión ejecute las siguientes lineas de código.

```
ggplot(data = data, aes(x = X, y = Y)) +  geom_point() +  geom_smooth(formula = y ~ x,
method = "lm", se = TRUE, color = "firebrick") +  theme_bw() + labs(x = "Porcentaje de
población en condición de pobreza", y = "Valor mediano de las casas ocupadas")
```

El resultado se muestra an la última página.

Predicción para una nueva observación y_o

```
S<-data
head(S)
```

X	Y
4.98	24.0
9.14	21.6
4.03	34.7
2.94	33.4

X	Y
5.33	36.2
5.21	28.7

```
predY<-predict(modelo,S,interval="prediction",level=0.95)
IPY<-cbind(X, Y,predY)
head(IPY)
```

X	Y	fit	lwr	upr
4.98	24.0	29.82260	17.58460	42.06059
9.14	21.6	25.87039	13.64341	38.09737
4.03	34.7	30.72514	18.48349	42.96680
2.94	33.4	31.76070	19.51432	44.00708
5.33	36.2	29.49008	17.25333	41.72683
5.21	28.7	29.60408	17.36691	41.84125

Tambien para un valor específico de x_o se puede hacer la predicción; por ejemplo para $x_o = 10.5$

```
xo<-data.frame(X=10.5)
predY<-predict(modelo,xo,interval="prediction",level=0.95)
cbind(xo, predY)
```

X	fit	lwr	upr
10.5	24.57832	12.35317	36.80347

Para $x_o = 10.5$ se obtiene el intervalo de confianza $12.35 \leq \mu_{Y/x_0} \leq 36.80$. Esto es, si el porcentaje de población en condición de pobreza es de 10.5%, se estima que el valor mediano de las casas ocupadas por el dueño se encuentre entre 12.35 y 36.80.

ACTIVIDAD.

Ejecute nuevamente los procedimientos vistos en clase con las variables: medv y rm.

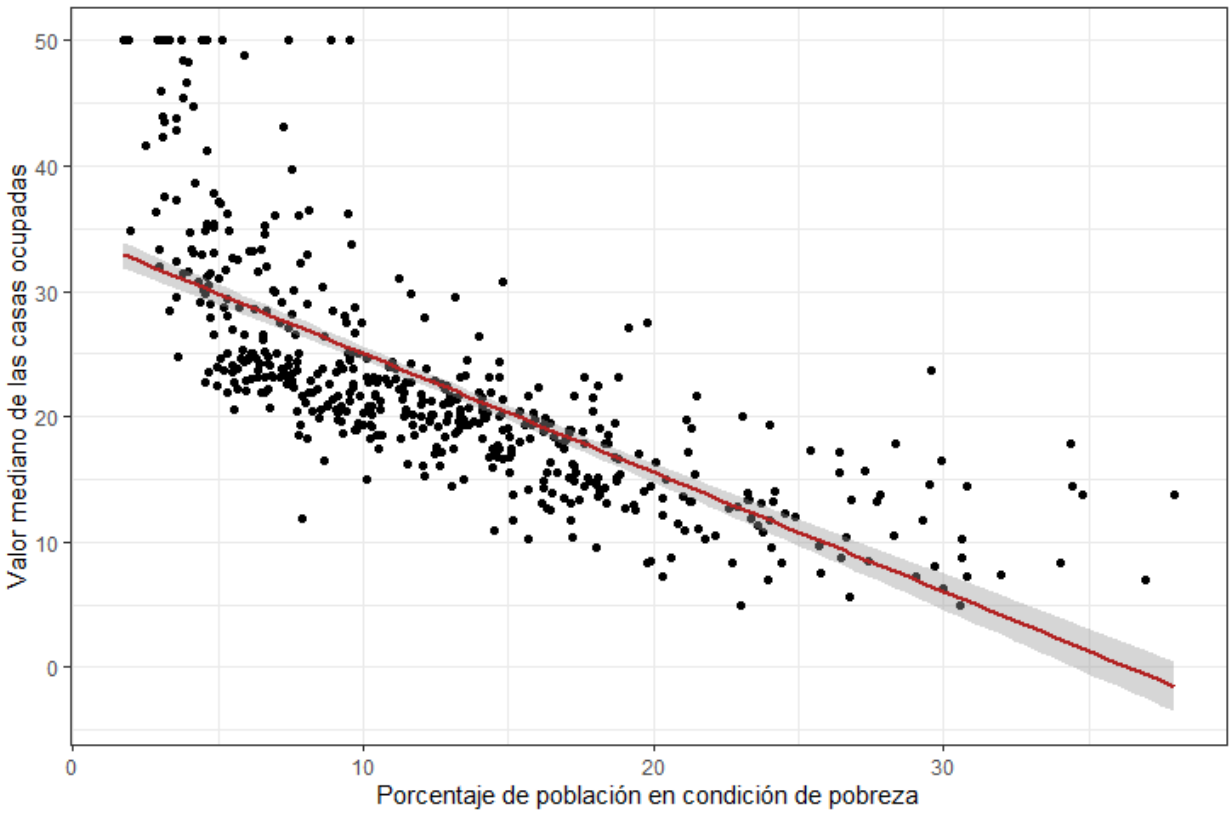

Figure 1: Gráfica de dispersión