Actividad 3: Regresión Lineal Simple

Especialidad en Métodos EStadísticos - CIMAT

Edgar Anuar Sánchez Hernández

2022-12-10

Pregunta de investigación

Un profesor de quinto grado quería determinar empíricamente si existe una relación entre la cantidad de libros que leen los estudiantes y la extensión de su vocabulario. Para lo anterior, el profesor aplicó una encuesta a una muestra aleatoria de alumnos de quinto grado en su escuela, les preguntó cuántos libros leen al mes y luego les dio una prueba de vocabulario (calificaciones en una escala de 0 a 10). Las puntuaciones se muestran a continuación.

```
library(ggplot2)
library("ggthemes")

num_libr <- c(12, 11, 11, 10, 10, 8, 7, 7, 6, 6, 5, 5, 3, 2, 1, 0)
calif <- c(10, 9, 10, 6, 8, 8, 6, 9, 6, 7, 8, 7, 5, 6, 3, 4)
mi_df <- data.frame(num_libr, calif)
```

Preguntas

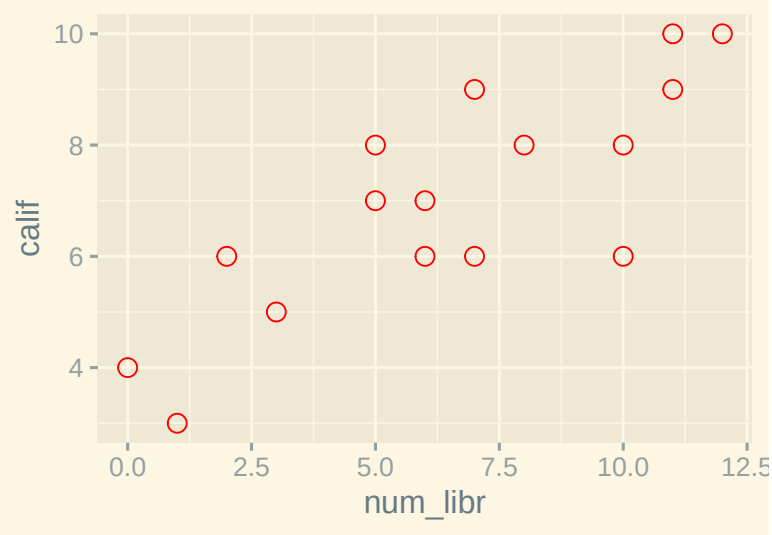
a) Identifique cuál es la variable dependiente y cuál la independiente

Variable independiente x_i: num. de libros leídos.

Variable dependiente y_i: calificación en examen de vocabulario (depende del número de libros leídos).

b) Gráfico de dispersión y cálculo de correlación e interpretación

```
#gráfica
ggplot(data = mi_df) +
  geom_point(mapping = aes(x = num_libr, y = calif),
             color = "red", size = 3, shape = 1) +
  theme_solarized_2()
```



Cálculo del coeficiente de correlación lineal:

```
cor(mi_df$num_libr,mi_df$calif)
```

```
## [1] 0.8189463
```

Existe una asociación aproximadamente lineal positiva entre la variable número de libros y calificación en el examen. Esta asociación se corrobora con un valor del coeficiente de correlación lineal aproximadamente igual a 0.82.

c) Ajuste de modelo de regresión, recta estimada e interpretación de coeficientes.

```
reg <- lm(calif~num_libr, data = mi_df)
summary(reg)
```

```
##
## Call:
## lm(formula = calif ~ num_libr, data = mi_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5649 -0.6178  0.1058  0.7500  1.7764
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.09375    0.62241   6.577 1.23e-05 ***
## num_libr     0.44712    0.08374   5.340 0.000104 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.208 on 14 degrees of freedom
## Multiple R-squared:  0.6707, Adjusted R-squared:  0.6471
## F-statistic: 28.51 on 1 and 14 DF,  p-value: 0.0001044
```

Los valores obtenidos para los coeficientes son $a = 4.09$ y $b = 0.45$. Por lo tanto la línea de regresión estimada es:

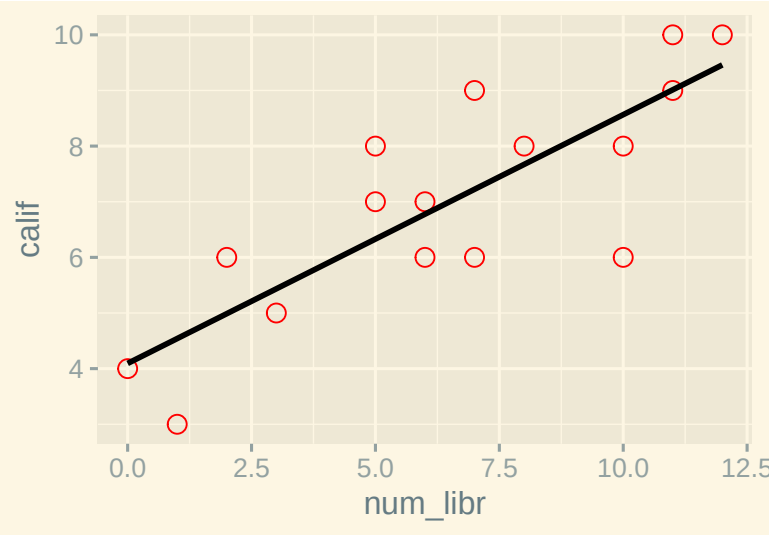
$$\hat{y}_i = 4.09 + 0.45x_i$$

La interpretación de estos coeficientes es la siguiente: por cada libro adicional que lean los estudiantes, su calificación promedio aumentará en 0.45. Mientras que 4.09, representa la calificación media cuando no se lee ningún libro.

A continuación se presenta el ajuste junto a los puntos.

```
ggplot(mi_df, aes(num_libr, calif)) +
  geom_point(color = "red", size = 3, shape = 1) +
  geom_smooth(method = lm, se = F, color = "black") +
  theme_solarized_2()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



d) R^2 e interpretación

El valor de R^2 se obtiene del código del inciso c) El "Multiple R-squared" tiene un valor de 0.67, lo cual mide la bondad del ajuste de la recta a los datos. En este caso el valor de 0.67, no es tan cercano a 1 por lo cual concluimos que el modelo es regularmente bueno.

e) Pruebas de hipótesis para los coeficientes con $\alpha = 0.05$.

Utilizando el código del inciso c) podemos realizar las siguientes pruebas de hipótesis.

Primero una prueba de hipótesis para la ordenada al origen.

$$H_0 : \alpha = 0$$

$$H_1 : \alpha \neq 0$$

Si se considera un nivel de significancia de 0.05, entonces H_0 se rechaza pues el p-value=$1.23\times10^{-5} < 0.05$.

Por otro lado, la prueba de hipótesis para la pendiente.

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

Considerando un nivel de significancia de 0.05, se rechaza H_0, pues el p-value=0.000104 < 0.05.

f) Estimar media de calificaciones para $x = 13, 5, 7,\ y\ 9$

```
predict(reg, newdata=data.frame(num_libr=13),interval='confidence',level=0.95)
```

```
##       fit      lwr     upr
## 1 9.90625 8.571303 11.2412
```

```
predict(reg, newdata=data.frame(num_libr=5),interval='confidence',level=0.95)
```

```
##        fit     lwr      upr
## 1 6.329327 5.62798 7.030674
```

```
predict(reg, newdata=data.frame(num_libr=7),interval='confidence',level=0.95)
```

```
##        fit      lwr      upr
## 1 7.223558 6.569817 7.877299
```

```
predict(reg, newdata=data.frame(num_libr=9),interval='confidence',level=0.95)
```

```
##        fit      lwr      upr
## 1 8.117788 7.329812 8.905765
```

Se obtienen estimaciones de la calificación para los siguientes valores de x=número de libros leídos:

Para $x = 13$ la media de calificaciones se estima entre 8.57 y 11.24 con un nivel de confianza del 95%.

Para $x = 5$ la media de calificaciones se estima entre 5.63 y 7.03 con un nivel de confianza del 95%.

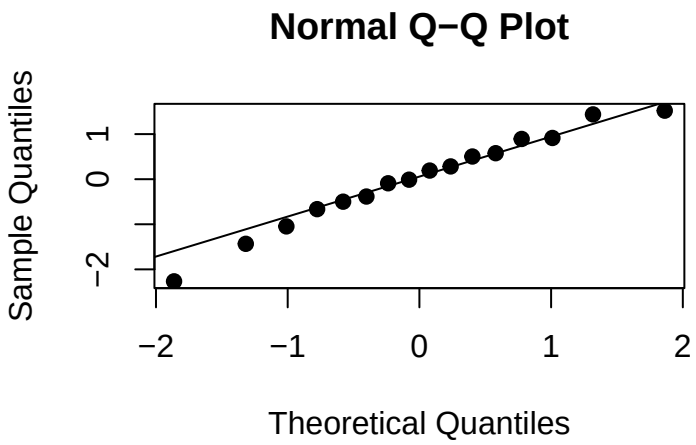
Para $x = 7$ la media de calificaciones se estima entre 6.57 y 7.88 con un nivel de confianza del 95%.

Para $x = 9$ la media de calificaciones se estima entre 7.33 y 8.91 con un nivel de confianza del 95%.

g) Verificación de los supuestos del modelo

```
residuales_stand<-rstandard(reg)
qqnorm(residuales_stand,pch=19)
qqline(residuales_stand)
```

1. Normaldad de los residuales: para esto realizamos la siguiente gráfica y observamos que los puntos se distribuyen aproximadamente cerca de la recta.



Además podemos realizar la prueba de normalidad Shapiro-Wilk para realizar la siguiente prueba de hipótesis:

H_0 : Los residuales estandarizados provienen de una distribución normal

H_1 : Los residuales estandarizados no provienen de una distribución normal

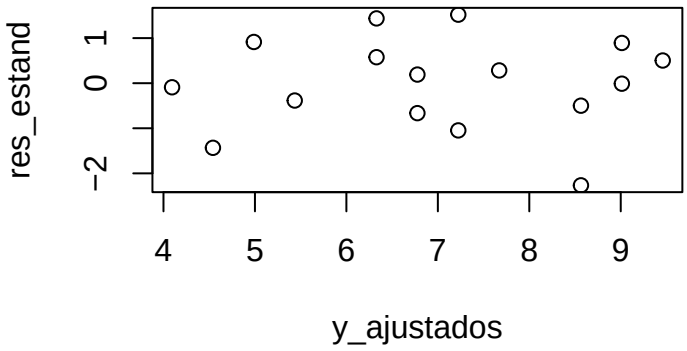
```
shapiro.test(residuales_stand)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuales_stand
## W = 0.97303, p-value = 0.885
```

Se observa un p-value de 0.885>0.05, por lo cual no se rechaza la hipótesis nula y concluimos que los residuales estandarizados provienen de una distribución normal.

2. Variabilidad constante de los residuales. Para esto se graficarán los valores predichos $\hat{y}_i$ por el modelo de regresión ajustado contra los residuales estandarizados d_i:

```
plot(reg$fitted.values,residuales_stand,xlab = "y_ajustados",ylab ="res_estand")
```



Visualmente se observa la uniformidad de los puntos, es decir la variabilidad constante de los residuales.

3. Los residuales no deben estar autocorrelacionados. Por la manera en que se realiza el levantamiento de la muestra, es decir de forma aleatoria, se garantiza la no autocorrelación de los residuales.

4. Detección de Outliers: Detectando si hay algún o algunos residuales con valor más extremo que 3 o -3. Para conocer la distribución de los residuales se expone un resumen de los residuales.

```
summary(residuales_stand)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -2.265837 -0.540388  0.090226 -0.004815  0.655809  1.520193
```

De acuerdo al resumen de los residuales, el valor mínimo fue -2.26, mientras que el mayor fue 1.52, entonces se descarta la presencia de un valor outlier.