Regresión Lineal simple taller#1

Julián Camilo Riaño Moreno

16/3/2020

Actividad taller 1 regresión lineal simple

Considere los datos del tiempo de espera entre erupciones y la duración de la erupción para el géiser Old Faithful en el Parque Nacional de Yellowstone, Wyoming, EE.UU.

1. Realice un análisis exploratorio de la base de datos.

Resumen de los datos del archivo geiser.csv

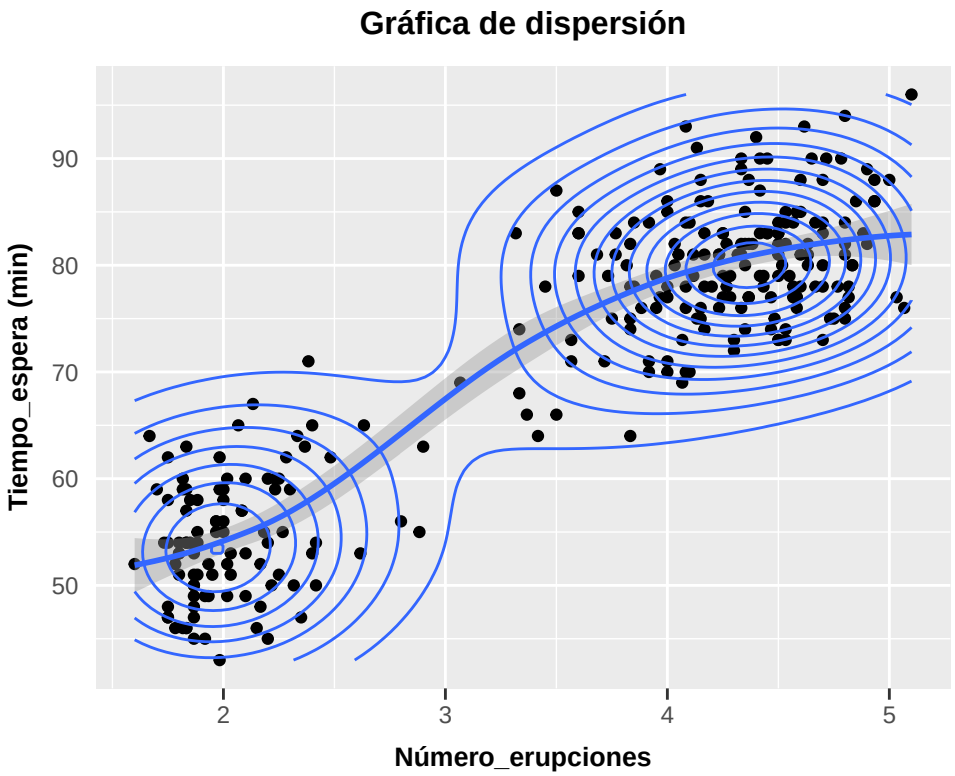
Table 1: Resumen datos data.frame

Número_erupciones	Tiempo_espera (min)
Min. :1.600	Min. :43.0
1st Qu.:2.163	1st Qu.:58.0
Median :4.000	Median :76.0
Mean :3.488	Mean :70.9
3rd Qu.:4.454	3rd Qu.:82.0
Max. :5.100	Max. :96.0

Se observa en la base de datos entregada, 272 observaciones acerca de dos variables, a saber: variable "Número de erupciones" y la "tiempo de espera". La primera es una varible cuantitativa discreta y la segunda corresponde a una variable cuantitativa continua. La descripción de las variables se puede encontrar en la tabla 1. En dicha tabla se observa una media aproximada de 3.5 erupciones con una mediana de 4, por otra parte, la media de tiempo de espera es de 70.9 minutos con una mediana de 16 minutos. Para determinar si existe alguna correlación entre dichas variables se decide realizar una gráfica de dispesión, que será descrita a continuación.

Gráfica de dispersión de los datos.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



La gráfica de dispersión acá mostrada muestra que existe un agrupamiento entre las observaciones hacia los extremos, encontrandose que los "números de erupciones" menores se correlaciona con "tiempo de espera" menores; lo mismo ocurre para los "número de erupciones" mayores con "tiempo de espera mayor". Sin embargo esta correlación no puede explicarse para los valores intermedios. Por esta razón, pareciera que no hay un correlación entre las dos variables, para evaluar esto se aplica un coeficiente de correlación de Pearson.

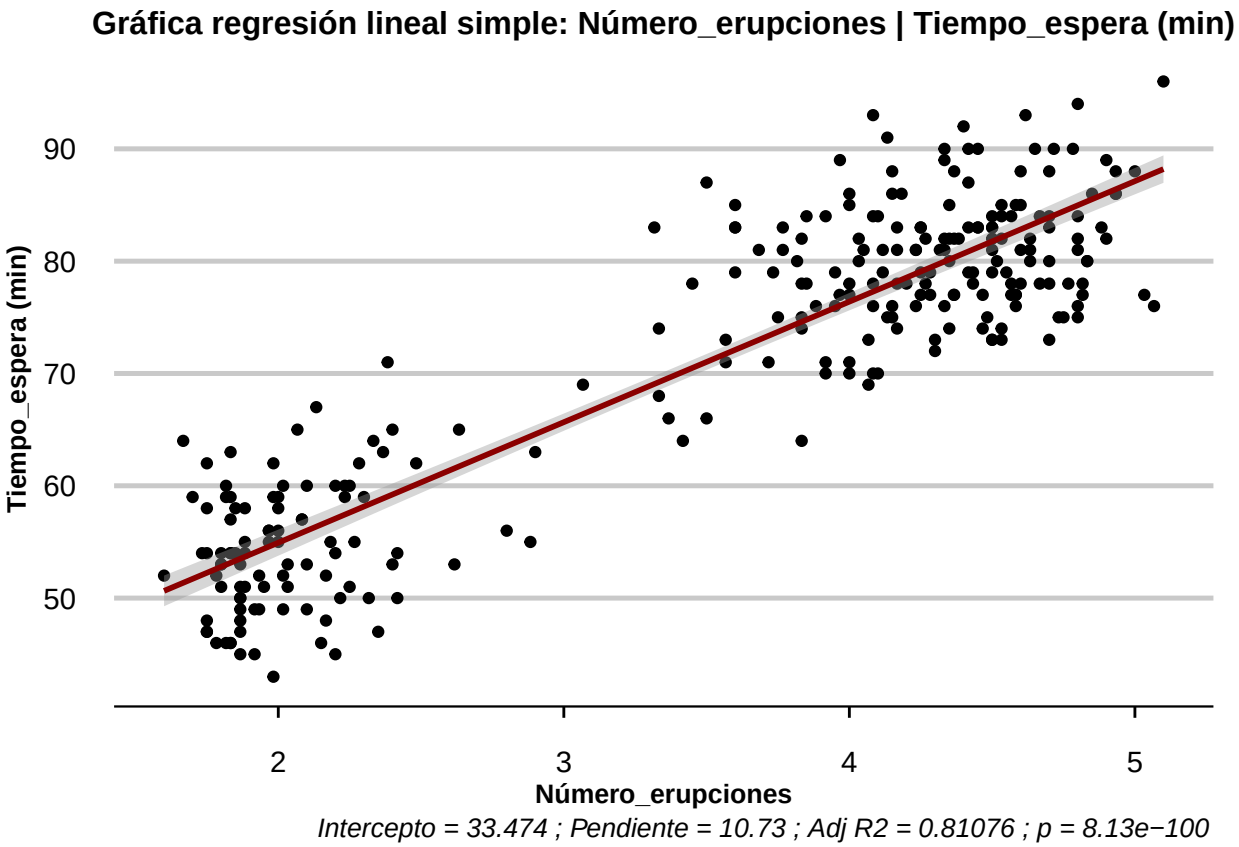
```
##
##  Pearson's product-moment correlation
##
## data:  geiserdat1$`Tiempo_espera (min)` and geiserdat1$Número_erupciones
## t = 34.089, df = 270, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8756964 0.9210652
## sample estimates:
##       cor
## 0.9008112
```

A través del indice de correlación de Pearson obtenido (0,9), con un valro de p (<0.05) de manera que es posible determinar una correlación positiva (≈ 1) entre las variables.

2. Construya un gráfico de dispersión que relacione la variable eruption y la variable waiting. ¿Es razonable suponer que existe una relación de dependencia lineal entre estas variables?

Gráfica de regresión lineal simple (utilizando formula desarrollo propio: `RLsimple_f`)

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



Se realizó un gráfica siguiendo ajustada el modelo lineal utilizado por R (modelo por mínimos cuadrados). Dónde a partir de la linea regresora (en rojo) resultante es posible, establecer que existe una dependencia lineal entre las variables, esto comprendiendo que los errores de las observaciones no están alejadas de la linea regresora y su desviación estandar.

3. Ajuste un modelo de regresión lineal que relacione el tiempo de espera con la duración de la erupción. Interprete los parámetros del modelo.

##Ajuste del modelo de regresión lineal

Table 2: Estimaciones de parametros de la regresión

	Estimado	ErrorStand
β_0	33.474	1.155
β_1	10.730	0.315

Como se puede observar en la tabla 2. A partir del ajuste del modelo lineal y la identificación de los parametros: β_0 (entendida como el intercepto de la lineal regresora) y β_1 (entendida como la pendiente de la linea regresora). encontrando que β_0 es 33, 474 y β_1 es 10.730, lo que al asumir una correlación lineal supone

que el incremento en una unidad de el "tiempo de espera" (un minuto) se puede incrementar el número de erupciones aproximadamente 10 veces.

4. Utilice las pruebas t para evaluar la contribución de cada variable regresora al modelo. Discuta sus hallazgos.

Table 3: t-values para los parametros dados

	t-value	p-value
β_0	28.985	7.136015e-85
β_1	34.089	8.000000e-100

La tabla 3 por su parte los valores de t y su respectiva signifcancia en el modelo lineal ajustado. A través de una prueba de hipotesis dónde:

$$H_o = \beta_0 = 0$$
$$H_1 = \beta_0 \neq 0$$

En este primer escenario se rechaza la hipotesis nula, ya que, β_0 (33,474), es un valor mayor a 0, con un valor de p significativo.

$$H_o = \beta_1 = 0$$
$$H_1 = \beta_1 \neq 0$$

En este segundo escenario se rechaza la hipotesis nula, ya que, β_0 (10.730), es un valor mayor a 0, con un valor de p significativo.

Así pues es posible realizar una correlación entre las variables estudiadas dado que la pendiente y el intercepto son aceptados en la prueba de hipotesis, de manera que la variable regresora tiene un efecto importante en la variable respuesta.

5. ¿Tiene el modelo obtenido un buen ajuste?. Justifique su respuesta.

Finalmente al revisar el valor de r^2 obtenido:

```
## [1] 0.8107625
```

Es posible inferir que el modelo lineal utilizado para este caso puede explicar que aproximadamente el 81% de la varianza de la variable respuesta "número de erupciones" es consecuencia a la variable regresora "tiempo de espera". De manera que sería posible afirmar que el modelo está bien ajustado. El resto de la varianza resultante puede ser explicado por otras variables que no se han tenido encuenta en el modelo o por azar.