Análisis de Regresión Lineal y Pruebas de Hipótesis

2025-05-13

Javier Ramirez Cervantes

Utilizando R:

Creamos nuestro DataFrame con los datos requeridos:

```
Ventas_Y = c(200000, 210000, 215000, 220000, 225000, 230000, 235000, 240000, 245000, 250000,
             255000, 260000, 265000, 270000, 275000, 280000, 285000, 290000, 295000, 300000,
             305000, 310000, 315000, 320000, 325000, 330000, 335000, 340000, 345000, 350000)

Gasto_Publicidad_X1 = c(20000, 22000, 23000, 25000, 26000, 28000, 29000, 31000, 32000, 33000,
                        35000, 36000, 37000, 39000, 40000, 42000, 43000, 45000, 46000, 48000,
                        49000, 51000, 52000, 54000, 55000, 57000, 58000, 60000, 61000, 63000)

Num_Empleados_X2 = c(50, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
                     66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80)

df = data.frame(Ventas_Y, Gasto_Publicidad_X1, Num_Empleados_X2)
df
```

```
##    Ventas_Y Gasto_Publicidad_X1 Num_Empleados_X2
## 1    200000               20000               50
## 2    210000               22000               52
## 3    215000               23000               53
## 4    220000               25000               54
## 5    225000               26000               55
## 6    230000               28000               56
## 7    235000               29000               57
## 8    240000               31000               58
## 9    245000               32000               59
## 10   250000               33000               60
## 11   255000               35000               61
## 12   260000               36000               62
## 13   265000               37000               63
## 14   270000               39000               64
## 15   275000               40000               65
## 16   280000               42000               66
## 17   285000               43000               67
## 18   290000               45000               68
## 19   295000               46000               69
## 20   300000               48000               70
## 21   305000               49000               71
## 22   310000               51000               72
## 23   315000               52000               73
```

```
## 24   320000              54000              74
## 25   325000              55000              75
## 26   330000              57000              76
## 27   335000              58000              77
## 28   340000              60000              78
## 29   345000              61000              79
## 30   350000              63000              80
```

Paso 1: Análisis Descriptivo de las variables: Analisis descriptivo de las variables proporcionadas

Se realizaran pruebas a cada variable para conocer el comportamiento de estas.

Primero se llaman las librerias necesarias para realizar el analisis

```
library(reticulate)
library(dplyr)
```

```
##
## Adjuntando el paquete: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(gapminder)
```

Medidas de Tendencia central y dispersión de las variables.

- Variable de Ventas_Y

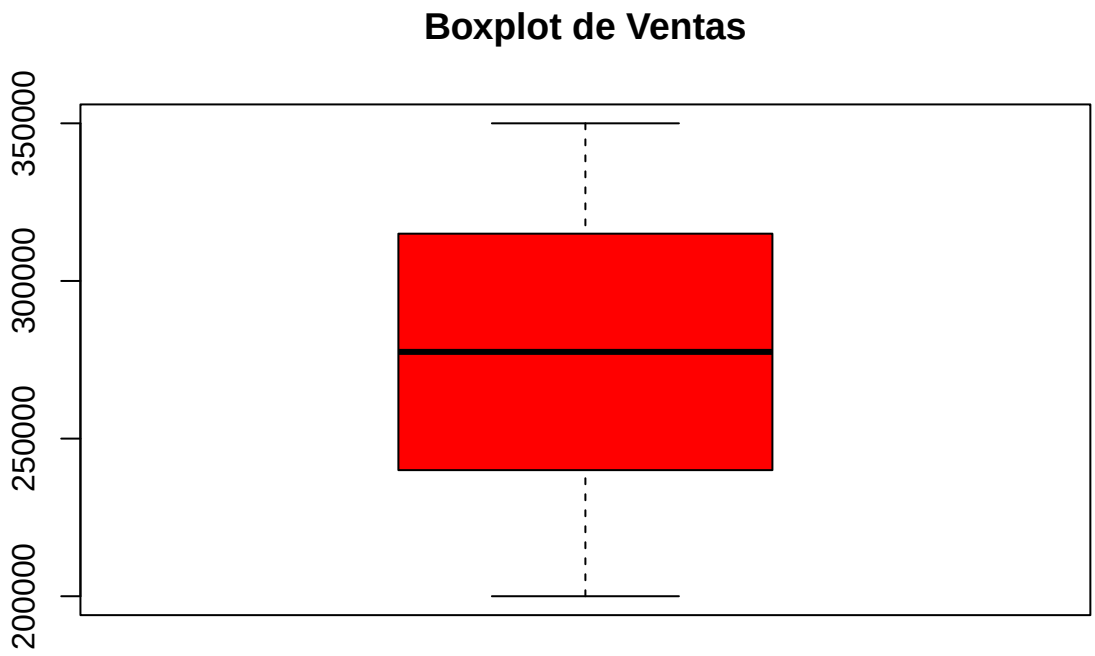
```
VY <-df %>%
  summarise(Media = mean(Ventas_Y),
            Mediana = median(Ventas_Y),
            Desviacion = sd(Ventas_Y) )

Rango = range(Ventas_Y)

VY
```

```
##      Media Mediana Desviacion
## 1 277333.3  277500   44309.52
```

```
Rango
```

```
## [1] 200000 350000
```

```r
boxplot(Ventas_Y, main="Boxplot de Ventas", col = "red")
```



De acuerdo con los datos anteriores, el valor de la media y la mediana es bastante cercano, lo cual puede indicar un distribucion casi normal de los datos, dato que se contrasta con el grafico de caja el cual parece indicar una distribucion no sesgada. La desviación estandar es normal dado que si se le suma o resta a la mediana, los resultados son similares a los que muestra el boxplot en los limites de la caja.

- Variable Gasto_en_Publicidad_X1

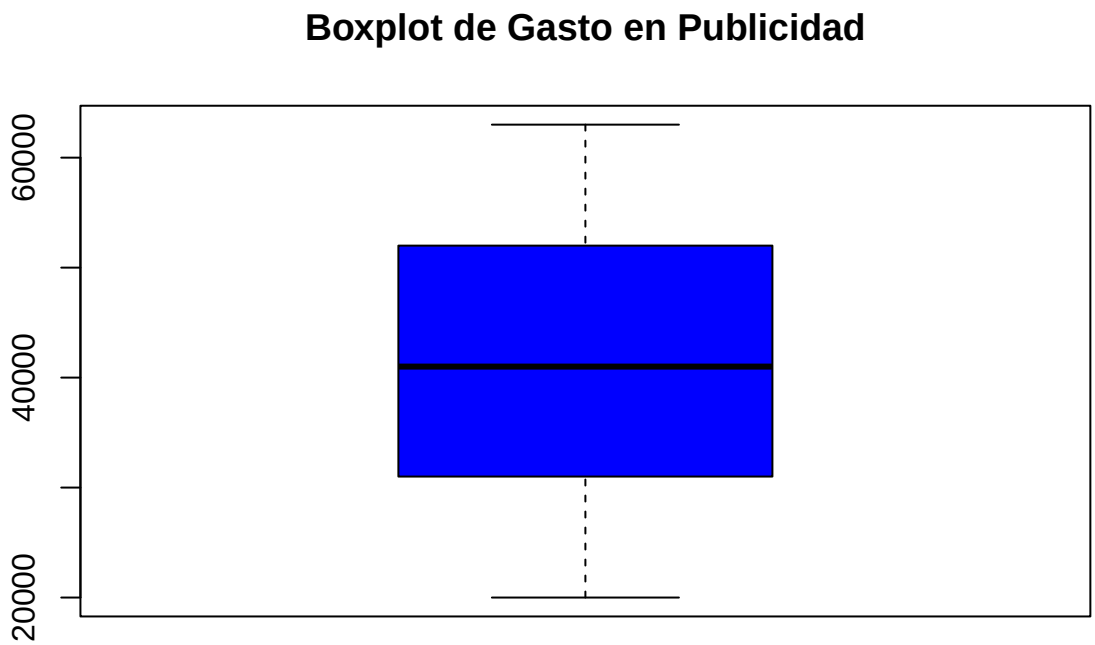
```r
GP <-df %>%
  summarise(Media = mean(Gasto_Publicidad_X1),
            Mediana = median(Gasto_Publicidad_X1),
            Desviacion = sd(Gasto_Publicidad_X1) )

Rango = range(Gasto_Publicidad_X1)

GP
```

```
##      Media Mediana Desviacion
## 1 41333.33   41000   12836.73
```

```r
Rango
```

```
## [1] 20000 63000
```

```r
boxplot(Gasto_Publicidad_X1, main="Boxplot de Gasto en Publicidad", col = "blue")
```



De acuerdo con los datos anteriores, el valor de la media y la mediana es bastante cercano,lo cual puede indicar un distribucion casi normal de los datos, dato que se contrasta con el grafico de caja el cual parece indicar una distribucion ligeramente segada hacia la parte superior del grafico de caja, es decir, un sesgo a la derecha. La desviación estandar es un poco alta dado que si se le suma o resta a la mediana, los resultados son cercanos a los que muestra el boxplot en los limites de la caja.

- Variable Numero de Empleados X2

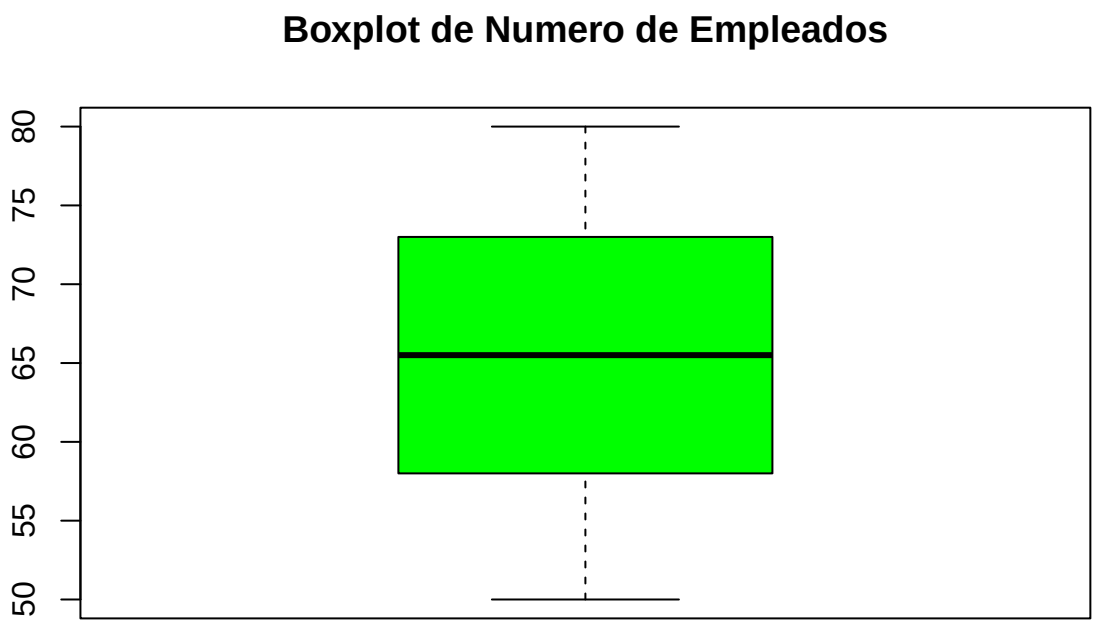
```r
NE <-df %>%
  summarise(Media = mean(Num_Empleados_X2),
            Mediana = median(Num_Empleados_X2),
            Desviacion = sd(Num_Empleados_X2) )

Rango = range(Num_Empleados_X2)

NE
```

```
##      Media Mediana Desviacion
## 1 65.46667    65.5   8.861903
```

```r
Rango
```

```
## [1] 50 80
```

```r
boxplot(Num_Empleados_X2, main="Boxplot de Numero de Empleados", col = "green")
```



De acuerdo con los datos anteriores, el valor de la media y la mediana es bastante cercano, lo cual puede indicar un distribucion casi normal de los datos, dato que se contrasta con el grafico de caja el cual parece indicar una distribucion no sesgada. La desviación estandar es normal dado que si se le suma o resta a la mediana, los resultados son similares a los que muestra el boxplot en los limites de la caja.

Paso 2: Construcción del Modelo de Regresión Lineal

Elaboracion del modelo

Se crea un modelo de regresion lineal multiple que explique el comportamiento de la variable Ventas dependiendo de los cambios en las variables de Gasto en Publicidad y Numero de empleados

```r
x_train = df[, c("Gasto_Publicidad_X1","Num_Empleados_X2")]
y_train = df['Ventas_Y']
lm <- lm(y_train$Ventas_Y ~ x_train$Gasto_Publicidad_X1 + x_train$Num_Empleados_X2)
print(summary(lm))
```

```
## Warning in summary.lm(lm): essentially perfect fit: summary may be unreliable

##
## Call:
## lm(formula = y_train$Ventas_Y ~ x_train$Gasto_Publicidad_X1 +
```

```
##     x_train$Num_Empleados_X2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.016e-10 -1.054e-11  2.388e-12  1.188e-11  3.564e-11
##
## Coefficients:
##                               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)                 -5.000e+04  6.113e-10 -8.180e+13  < 2e-16 ***
## x_train$Gasto_Publicidad_X1 -3.831e-14  1.142e-14 -3.355e+00  0.00236 **
## x_train$Num_Empleados_X2     5.000e+03  1.654e-11  3.023e+14  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.492e-11 on 27 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 4.585e+31 on 2 and 27 DF,  p-value: < 2.2e-16
```

Reportando Resultados de los coeficientes

Una vez que se estimo el modelo, los resultados de los coeficientes de determinacion son los siguientes:

R2 = 1 y R2 ajustado = 1

Estos resultados muestran un problema de sobreajuste del modelo.

Paso 3: Pruebas de Hipótesis sobre los Coeficientes

En esta parte del modelo se verifica la significancia de las variables para la estimacion del modelo.

Resultados del modelo:

```
print(summary(lm))
```

```
## Warning in summary.lm(lm): essentially perfect fit: summary may be unreliable
```

```
##
## Call:
## lm(formula = y_train$Ventas_Y ~ x_train$Gasto_Publicidad_X1 +
##     x_train$Num_Empleados_X2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.016e-10 -1.054e-11  2.388e-12  1.188e-11  3.564e-11
##
## Coefficients:
##                               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)                 -5.000e+04  6.113e-10 -8.180e+13  < 2e-16 ***
## x_train$Gasto_Publicidad_X1 -3.831e-14  1.142e-14 -3.355e+00  0.00236 **
## x_train$Num_Empleados_X2     5.000e+03  1.654e-11  3.023e+14  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.492e-11 on 27 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 4.585e+31 on 2 and 27 DF,  p-value: < 2.2e-16
```

Hipotesis nula: H0 -> El coeficiente no es significativo
Hipotesis alternativa: Ha -> El coeficiente es significativo

Reglas de decision: p-value > 0.05 No se rechaza H0 // p-value < 0.05 Se rechaza H0

Segun los resultados de los p-values de cada uno de los coeficientes, los resultados son los siguientes:

- Coeficiente del intercepto = 2e-16 < 0.05 , por lo cual se rechaza la H0 y el coeficiente es significativo
- Coeficiente del Gasto en Publicidad = 0.00236 < 0.05 , por lo cual se rechaza la H0 y el coeficiente es significativo
- Coeficiente del Numero de Empleados = 2e-16 < 0.05 , por lo cual se rechaza la H0 y el coeficiente es significativo

Paso 4: Análisis de los Residuos

En esta parte del analisis se realizan las pruebas sobre los residuos del modelo estimado

Prueba de normalidad, Jarque-Bera

```
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
jarque.bera.test(lm$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  lm$residuals
## X-squared = 125.44, df = 2, p-value < 2.2e-16
```

Hipotesis nula: H0 -> Sesgo = 0 y Kurtosis = 3 -> Los residuos se distribuyen como una normal

Hipotesis alternativa: Ha -> Sesgo =/ 0 y/o Kurtosis =/ 3 -> Los residuos no se distribuyen como una normal

Reglas de decision: p-value > 0.05 No se rechaza H0 p-value < 0.05 Se rechaza H0

- Jarque-Bera p-value = 2.2e-16 < 0.05 , por lo cual se rechaza la H0 y el los residuos no se distribuyen como una normal

Prueba de Homocedasticidad, Breusch-Pagan

```r
library(lmtest)
```

```
## Cargando paquete requerido: zoo
```

```
##
## Adjuntando el paquete: 'zoo'
```

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

```r
bptest(lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm
## BP = 12.805, df = 2, p-value = 0.001657
```

Hipotesis nula: H0 -> Homocedasticidad - La varianza de los residuos es constante
Hipotesis alternativa: Ha -> Heterocedasticidad - La varianza de los residuos no es constante

Reglas de decision: p-value > 0.05 No se rechaza H0 p-value < 0.05 Se rechaza H0

- Breusch-Pagan p-value $= 0.001657 < 0.05$, por lo cual se rechaza la H0 y la varianza de los residuos no es constante.

Prueba de autocorrelacion, Durbin-Warson:

```r
library(car)
```

```
## Cargando paquete requerido: carData
```

```
##
## Adjuntando el paquete: 'car'
```

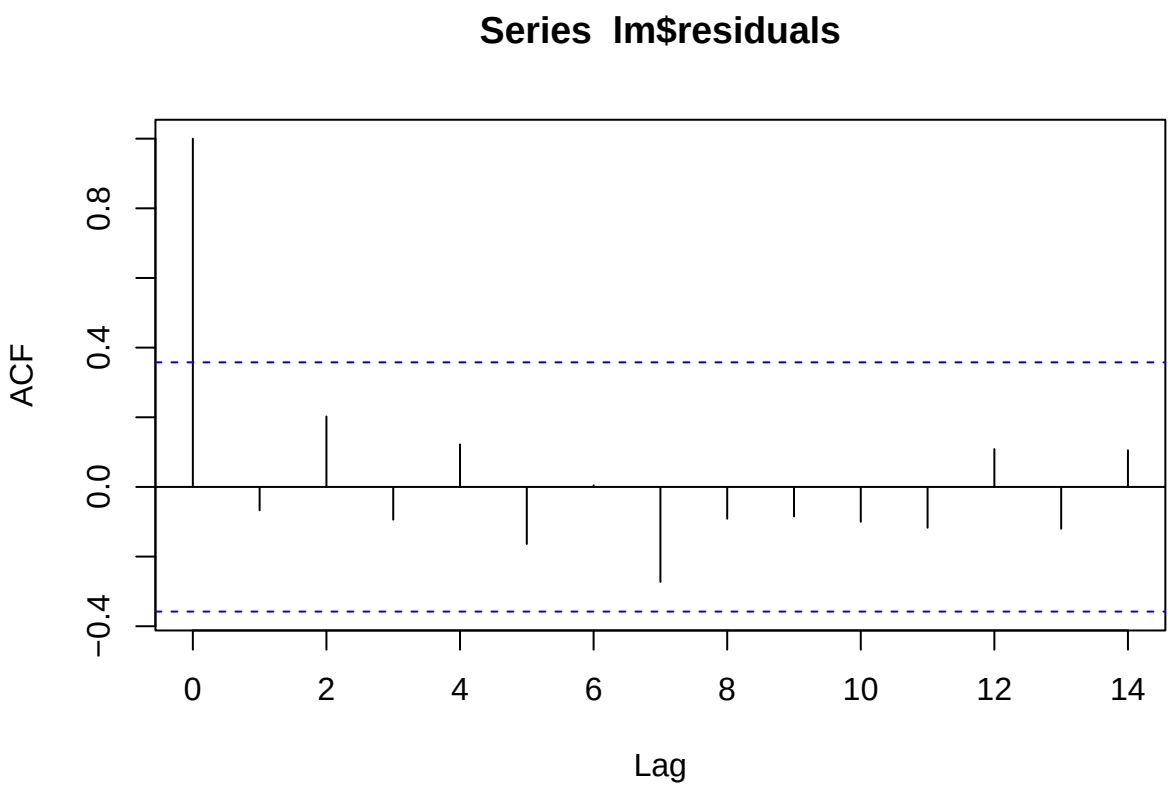
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
durbinWatsonTest(lm)
```

```
## Warning in summary.lm(model): essentially perfect fit: summary may be
## unreliable
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.06706091      1.508906   0.228
##  Alternative hypothesis: rho != 0
```

```r
## Correlogramas
library(lmtest)
acf(lm$residuals)
```



Hipotesis nula: H0 -> No autocorrelacion - los residuos no estan autocorrelacionados
Hipotesis alternativa: Ha -> Autocorrelacion - los residuos estan autocorrelacionados

Reglas de decision: p-value > 0.05 No se rechaza H0 p-value < 0.05 Se rechaza H0

- Breusch-Pagan p-value $= 0.23 > 0.05$, por lo cual no se rechaza la H0 y los residuos no estan autocorrelacionados

Asimismo el correlograma muestra que no hay problemas de autocorrelacion.

Paso 5: Conclusiones

Resumiendo los resultados de la pruebas del modelo, en la primera parte de la descripcion de las variables, la mayoria de ellas muestran una distribucion casi normal, salvo la variable de Gasto en Publicidad que si muestra cierto sesgo a la derecha. En la elaboración del modelo de regresion lineal multiple, los resultados de los coeficientes de determinacion de R2 y R2 ajustado tienen un valor de 1 en ambos casos, lo cual muestra un problema de sobreajuste del modelo. En la siguiente prueba de significancia de los coeficientes, de acuerdo con los resultados, el intercepto, y la variables de Gasto en Publicidad y Numero de Empleados son significativos para la estimacion del modelo. En las siguientes pruebas sobre el analisis de los residuos del modelo, los resultados arrojaron mas irregularidades. El modelo no tiene una distribucion normal de

sus residuos, y la varianza de estos no es constante. A pesar de los resultados, los residuos no muestran problemas de autocorrelacion

Para buscar una razon a los problemas del modelo, se realiza la prueba de multicolinealidad:

Prueba del VIF (Factor de inflacion de la varianza)

```
library(car)
vif(lm)
```

```
## Warning in summary.lm(object, ...): essentially perfect fit: summary may be
## unreliable
```

```
## x_train$Gasto_Publicidad_X1    x_train$Num_Empleados_X2
##                    1003.348                    1003.348
```

Criterio de decision:

VIF $<$ 5 -> Baja Multicolinealidad
VIF $>$ 5 -> Alta Multicolinealidad

De acuerdo con los resultados, ambas variables explicativas presentan altos problemas de multicolinealidad, lo cual podria ser la causada de que los residuos no tengan una distribucion normal, el problema de la heterocedasticidad y sobreajuste del modelo. Para verificar que esta sea la causa de dichos problemas, se tendria que hacer un nuevo modelo eliminando algunas de las variables debido a que existe multicolinealidad en ambos casos.