Aprendizaje de Maquina práctica 03

Pablo Diaz - 30343 | Kevin Huerta - 30502 | Diego Zuazo - 30046 | Gerardo Hernandez - 29902

***Abstract*— Comprender y utilizar las funciones básicas en el lenguaje de programación R y Python para realizar la regresión lineal.**

I. INTRODUCCIÓN

Este reporte tiene como objetivo explicar las metodologías aplicadas en los ejercicios resueltos de la práctica número 3, a través del uso de dos lenguajes de programación: Python y R. Donde el enfoque principal que tiene esta práctica fue hacer uso de regresión lineal en ambos lenguajes. Antes de entrar a la metodología, explicaremos algunos fundamentos que tuvimos que tener en cuenta para desarrollar nuestros algoritmos de resolución. Será una práctica complicada para python, ya que los ejercicios propuestos están diseñados para el lenguaje de R, pero investigando podemos encontrar algunas librerías que hagan algunas funciones que tiene integrado R.

II. FUNDAMENTOS

A. Regresión lineal

El modelo de pronóstico de regresión lineal permite hallar el valor esperado de una variable aleatoria a cuando b toma un valor específico. La aplicación de este método implica un supuesto de linealidad cuando la demanda presenta un comportamiento creciente o decreciente, por tal razón, se hace indispensable que previo a la selección de este método exista un análisis de regresión que determine la intensidad de las relaciones entre las variables que componen el modelo.

Existen varios tipos de regresión lineal, la primera es regresión lineal simple. Este tipo de regresión se basa en predecir una respuesta o valor cuantitativo. Esta respuesta se representa con la letra Y, y se calcula con base en predictores representados con la letra X. Los predictores son aquellos valores que determinan el comportamiento y la forma de la regresión, por lo tanto, se asume que hay una relación aproximadamente lineal entre X y Y.La formula es la siguiente:

$$Y_i = (a + bX_i) + \epsilon_i$$

Y para calcular a y b se utilizan las siguientes formulas, aunque los lenguajes de programación ya lo hacen automáticamente por nosotros:

$$a = \bar{y} - b\bar{x} = \frac{\sum y}{n} - b\frac{\sum x}{n}$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(yi - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Por otro lado, esta la regresión lineal múltiple y en este tipo de regresión la diferencia es que existen más de 1 predictor para obtener una sola respuesta, también llamada salida. Y es representado matemáticamente de la siguiente forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon.$$

$$y = \beta_0 + \sum_{i=1} \beta_i x_i + \epsilon_i.$$

B. Varianza

La varianza (S^2) mide la dispersión de los datos de una muestra ($x_1, x_2 ... x_n$) respecto a la media (x), calculando la media de los cuadrados de las distancias de todos los datos. La fórmula es la siguiente:

$$S_X^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{x})^2}{N-1}$$

siendo $(X_1, X_2, ..., X_N)$ un conjunto de datos y $\overline{x}$ la media

Al elevar las diferencias al cuadrado se garantiza que las diferencias absolutas respecto a la media no se anulan entre si. Además, resaltan los valores alejados.

Siempre se cumple que la varianza es mayor o igual que cero (S2 > 0). Ésta es cero cuando todos los datos son el mismo (ejemplo: 1,1,1,1,1).

Si en vez de tratarse de una muestra, la varianza se refiere a la población, el denominador será N.

C. Anova

La técnica de análisis de varianza (ANOVA) también conocida como análisis factorial y desarrollada por Fisher en 1930, constituye la herramienta básica para el estudio del efecto de uno o más factores (cada uno con dos o más niveles) sobre la media de una variable continua. Es por lo tanto el test estadístico a emplear cuando se desea comparar las medias de dos o más grupos. Esta técnica puede

generalizarse también para estudiar los posibles efectos de los factores sobre la varianza de una variable.

El funcionamiento básico de un ANOVA consiste en calcular la media de cada uno de los grupos para a continuación comparar la varianza de estas medias (varianza explicada por la variable grupo, intervarianza) frente a la varianza promedio dentro de los grupos (la no explicada por la variable grupo, intravarianza).

D. Factor de inflación de la varianza

El factor de inflación de varianza (vif) es una medida de la cantidad de multicolinealidad en un conjunto de variables de regresión múltiple. matemáticamente, el vif para una variable de modelo de regresión es igual a la razón de la varianza general del modelo a la varianza de un modelo que incluye solo esa variable independiente única. Esta relación se calcula para cada variable independiente. un vif alto indica que la variable independiente asociada es altamente colineal con las otras variables en el modelo.

Un factor de inflación de varianza (vif) proporciona una medida de multicolinealidad entre las variables independientes en un modelo de regresión múltiple. Detectar la multicolinealidad es importante porque si bien no reduce el poder explicativo del modelo, sí reduce la significación estadística de las variables independientes. Una gran vif en una variable independiente indica una relación altamente colineal con las otras variables que deben considerarse o ajustarse en la estructura del modelo y la selección de variables independientes.

III. METODOLOGÍA

La metodología del equipo fue la misma que en veces anteriores, nos dividimos en dos grupos para la implementación del código, un grupo de R y un grupo de Python. Como mencionamos en la práctica anterior, los equipos se alternaron de lenguaje para practicar. Cada ejercicio tuvo su nivel de dificultad, tuvimos que trabajar en equipo completo para encontrar el enfoque adecuada para algunas actividades. De forma mas puntual, la metodología fue la siguiente:

A. Ejercicio número 1

En el ejercicio 1 se hizo uso de una libreria llamada "MASS" la cual contiene un conjunto de datos sobre Boston la cual registra el preico promedio de una casa en 506 vencidades diferentes.Se busca predecir medv lo cual es el precio promedio de una casa usando 13 predictores

B. Ejercicio número 2

Para este ejercicio se hizo una regression lineal multiple usando el metodo de minimos cuadrados de nuevo se uso la funcion de lm(). Con el mismo conjunto de datos de la actividad pasada usamos la siguente funcion lm.fit=lm(medvlstat+age ,data=Boston) y despues usamos summary para obtener los coeficientes de regression para todos los predictores

C. Ejercicio número 3

En el ejercicio 3 incluimos terminos de interaccion a nuestro modelo lineal. Tambien se uso la sintaxis lstat * age que incluye simultáneamente lstat, age,y el término de interacción lstat × edad como predictores; es una taquigrafía para lstat + age + lstat: edad. Y al final se imprime el resultado con el uso de summary()

D. Ejercicio número 4

En el ejercicio 4 se hizo uso de la funcion I() para elevar el predictor lstat al cuadrado de ahi se le aplica la regression para calcular mdev pero ahora incluyendo lstat y lstat2. Y finalmente usamos summary() para imprimir los resultados

E. Ejercicio número 5

Para el ejercicio 5 usamos un conjunto de datos llamado Carseats el cual es parte de la libreria ISLR. Con este conjunto se predijo las ventas y se observo un predictor cualitativo llamado ShelveLoc que es la posicion en los pasillos en el que se encuentra el Carseat.

F. Ejercicio número 6

Para el ultimo ejercicio se creo una función llamada LoadLibraries() lo que hace esta función es que importa las librerías ISLR y MASS , y una vez que las haya importado nos imprima que ya estan importadas. Al llamar la función las librerías son importadas y se imprime lo siguiente : [1] " The libraries have been loaded ."

IV. RESULTADOS

Al igual que en las practicas pasadas, es complicado mostrar los datos de una forma digerible. Lo que haremos es mostrar los resultados mas relevantes obtenidos en las actividades en el orden tal y como lo arroja nuestro programa. En el caso de R, se muestran todos los resultados de R que especifica el libro y en el mismo orden. En el de Python también están en el orden que especifica el libro, pero hay algunos que no se pueden mostrar de la misma forma o con el mismo detalle debido a las restricciones que tiene Python en comparación con R. De igual forma, hay algunos pasos en la actividad que demuestran como se arrojan errores y después se arreglan, por lo que no incluiremos esos resultados ya que no son finales.

A. Ejercicios en R:

1) Actividad 1- Los resultados obtenidos para la Actividad 1 fueron los siguientes :

A data.frame: 506 × 14

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9
8	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.90	19.15	27.1
9	0.21124	12.5	7.87	0	0.524	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9
11	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15.0
12	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.90	13.27	18.9
13	0.09378	12.5	7.87	0	0.524	5.889	39.0	5.4509	5	311	15.2	390.50	15.71	21.7
14	0.62976	0.0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21.0	396.90	8.26	20.4
15	0.63796	0.0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21.0	380.02	10.26	18.2
16	0.62739	0.0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21.0	395.62	8.47	19.9
17	1.05393	0.0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21.0	386.85	6.58	23.1
18	0.78420	0.0	8.14	0	0.538	5.990	81.7	4.2579	4	307	21.0	386.75	14.67	17.5
19	0.80271	0.0	8.14	0	0.538	5.456	36.6	3.7965	4	307	21.0	288.99	11.69	20.2
20	0.72580	0.0	8.14	0	0.538	5.727	69.5	3.7965	4	307	21.0	390.95	11.28	18.2
21	1.25179	0.0	8.14	0	0.538	5.570	98.1	3.7979	4	307	21.0	376.57	21.02	13.6
22	0.85204	0.0	8.14	0	0.538	5.965	89.2	4.0123	4	307	21.0	392.53	13.83	19.6
23	1.23247	0.0	8.14	0	0.538	6.142	91.7	3.9769	4	307	21.0	396.90	18.72	15.2
24	0.98843	0.0	8.14	0	0.538	5.813	100.0	4.0952	4	307	21.0	394.54	19.88	14.5
25	0.75026	0.0	8.14	0	0.538	5.924	94.1	4.3996	4	307	21.0	394.33	16.30	15.6
26	0.84054	0.0	8.14	0	0.538	5.599	85.7	4.4546	4	307	21.0	303.42	16.51	13.9
27	0.67191	0.0	8.14	0	0.538	5.813	90.3	4.6820	4	307	21.0	376.88	14.81	16.6
28	0.95577	0.0	8.14	0	0.538	6.047	88.8	4.4534	4	307	21.0	306.38	17.28	14.8
29	0.77299	0.0	8.14	0	0.538	6.495	94.4	4.4547	4	307	21.0	387.94	12.80	18.4
30	1.00245	0.0	8.14	0	0.538	6.674	87.3	4.2390	4	307	21.0	380.23	11.98	21.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

A matrix: 3 × 3 of type dbl

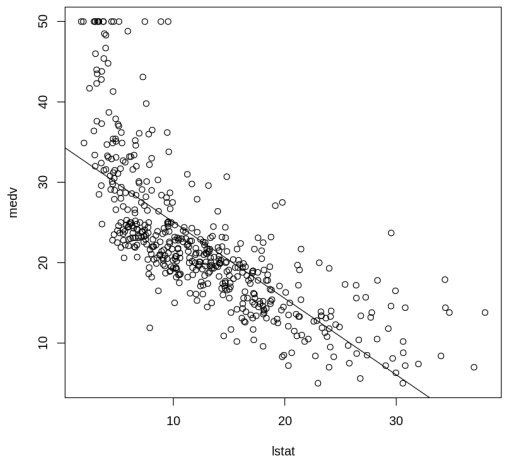
	fit	lwr	upr
1	29.80359	29.00741	30.59978
2	25.05335	24.47413	25.63256
3	20.30310	19.73159	20.87461

A matrix: 3 × 3 of type dbl

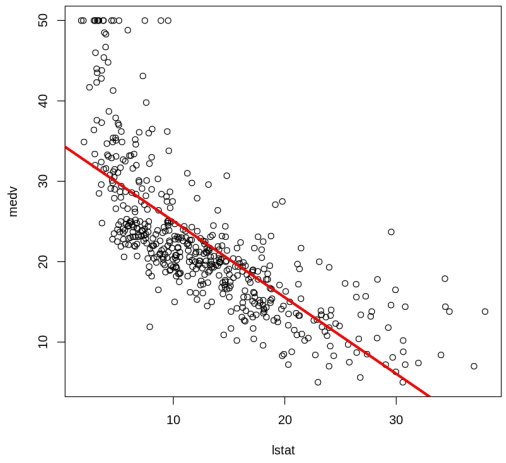
	fit	lwr	upr
1	29.80359	17.565675	42.04151
2	25.05335	12.827626	37.27907
3	20.30310	8.077742	32.52846

```
[1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
[8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

```
Call:
lm(formula = medv ~ lstat)

Coefficients:
(Intercept)        lstat
      34.55         -0.95

Call:
lm(formula = medv ~ lstat)

Residuals:
    Min      1Q  Median      3Q     Max
-15.168  -3.990  -1.318   2.034  24.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,    Adjusted R-squared:  0.5432
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

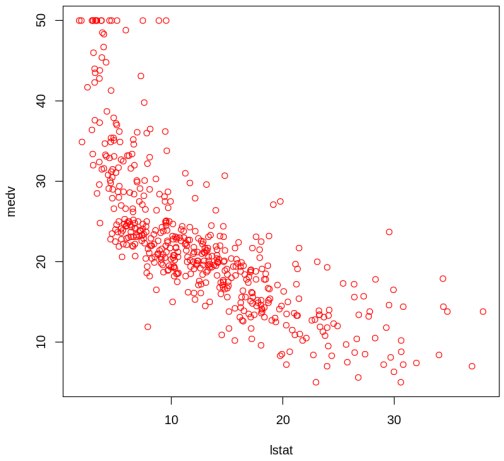


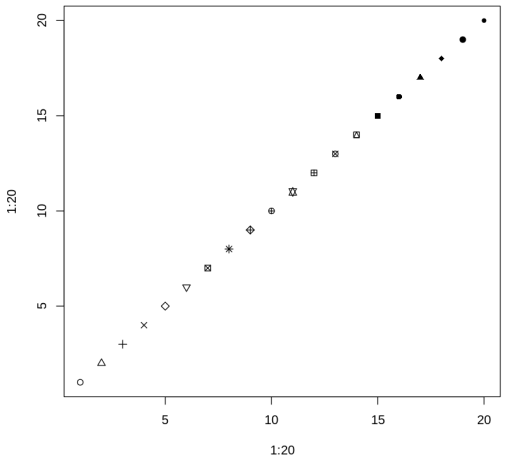

```
'coefficients' · 'residuals' · 'effects' · 'rank' · 'fitted.values' · 'assign' · 'qr' · 'df.residual' · 'xlevels' · 'call' · 'terms' · 'model'
(Intercept):        34.5538408793831 lstat:        -0.950049353757991
```

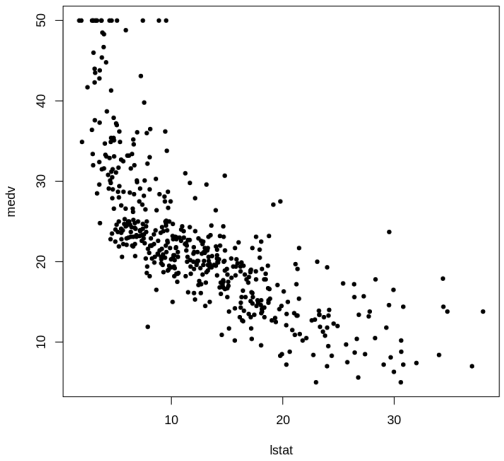


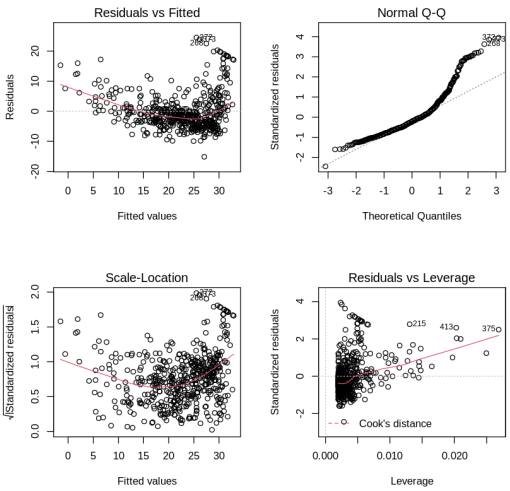
A matrix: 2 × 2 of type dbl

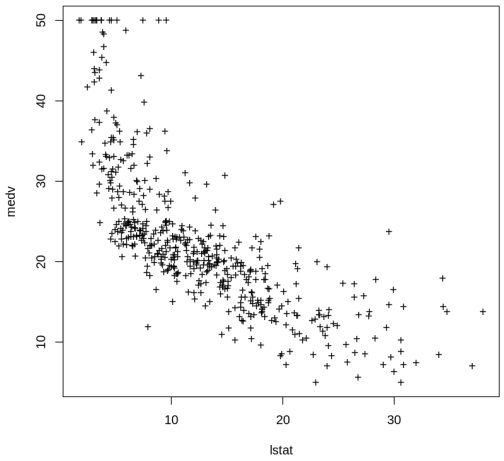
	2.5 %	97.5 %
(Intercept)	33.448457	35.6592247
lstat	-1.026148	-0.8739505

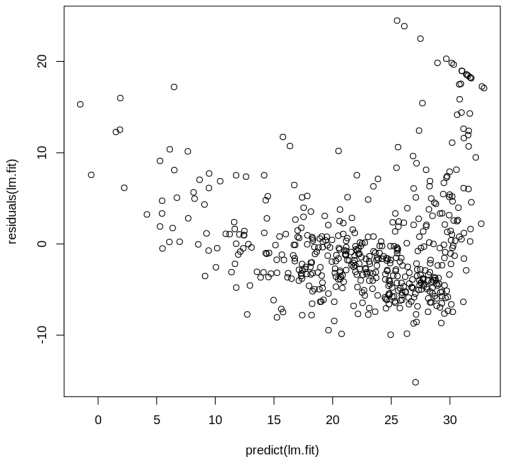

medv
lstat

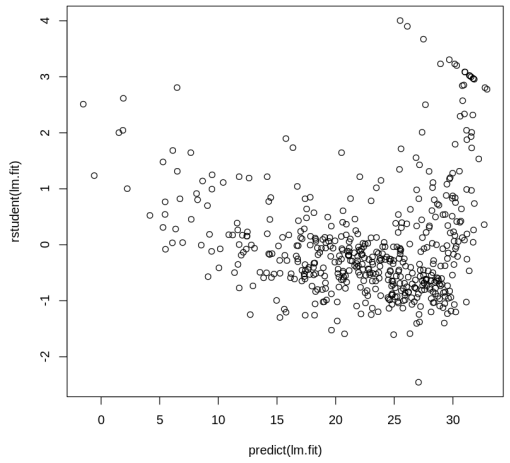

1:20
1:20


medv
lstat


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


medv
lstat

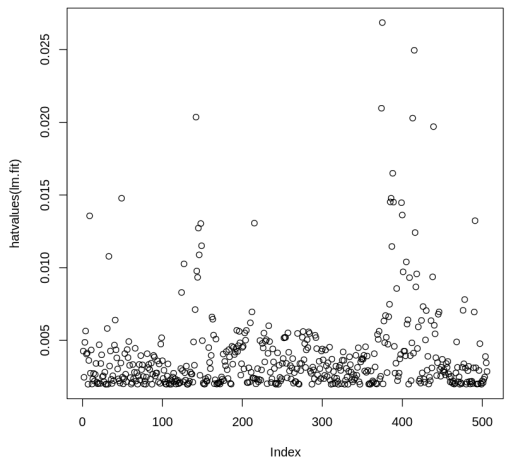

residuals(lm.fit)
predict(lm.fit)




```
Call:
lm(formula = medv ~ lstat + age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.981  -3.978  -1.283   1.968  23.158

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
age          0.03454    0.01223   2.826  0.00491 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.173 on 503 degrees of freedom
Multiple R-squared:  0.5513,    Adjusted R-squared:  0.5495
F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```


```
Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

```
    crim       zn    indus     chas      nox       rm      age      dis
1.792192 2.298758 3.991596 1.073995 4.393720 1.933744 3.100826 3.955945
     rad      tax  ptratio    black    lstat
7.484496 9.008554 1.799084 1.348521 2.941491
```




```
Call:
lm(formula = medv ~ . - age, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6054  -2.7313  -0.5188   1.7601  26.2243

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.436927   5.080119   7.172 2.72e-12 ***
crim         -0.108006   0.032832  -3.290 0.001075 **
zn            0.046334   0.013613   3.404 0.000719 ***
indus         0.020562   0.061433   0.335 0.737989
chas          2.689026   0.859598   3.128 0.001863 **
nox         -17.713540   3.679308  -4.814 1.97e-06 ***
rm            3.814394   0.408480   9.338  < 2e-16 ***
dis          -1.478612   0.190611  -7.757 5.03e-14 ***
rad           0.305786   0.066089   4.627 4.75e-06 ***
tax          -0.012329   0.003755  -3.283 0.001099 **
ptratio      -0.952211   0.130294  -7.308 1.10e-12 ***
black         0.009321   0.002678   3.481 0.000544 ***
lstat        -0.523852   0.047625 -10.999  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.74 on 493 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7343
F-statistic: 117.3 on 12 and 493 DF,  p-value: < 2.2e-16
```

2) Actividad 2- Los resultados obtenidos para la Actividad 2 fueron los siguientes :

3) Actividad 3- Los resultados obtenidos para la Actividad 3 fueron los siguientes :

```
Call:
lm(formula = medv ~ lstat * age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.806  -4.045  -1.333   2.085  27.552

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
age         -0.0007209  0.0198792  -0.036   0.9711
lstat:age    0.0041560  0.0018518   2.244   0.0252 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.149 on 502 degrees of freedom
Multiple R-squared:  0.5557,    Adjusted R-squared:  0.5531
F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = medv ~ poly(lstat, 5))

Residuals:
     Min       1Q   Median       3Q      Max
-13.5433  -3.1039  -0.7052   2.0844  27.1153

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       22.5328     0.2318  97.197  < 2e-16 ***
poly(lstat, 5)1 -152.4595     5.2148 -29.236  < 2e-16 ***
poly(lstat, 5)2   64.2272     5.2148  12.316  < 2e-16 ***
poly(lstat, 5)3  -27.0511     5.2148  -5.187 3.10e-07 ***
poly(lstat, 5)4   25.4517     5.2148   4.881 1.42e-06 ***
poly(lstat, 5)5  -19.2524     5.2148  -3.692 0.000247 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.215 on 500 degrees of freedom
Multiple R-squared:  0.6817,    Adjusted R-squared:  0.6785
F-statistic: 214.2 on 5 and 500 DF,  p-value: < 2.2e-16
```

4) Actividad 4- Los resultados obtenidos para la Actividad 4 fueron los siguientes :

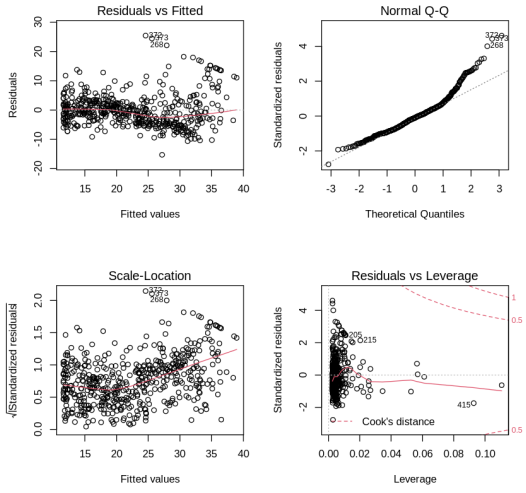
```
Call:
lm(formula = medv ~ lstat + I(lstat^2))

Residuals:
     Min       1Q   Median       3Q      Max
-15.2834  -3.8313  -0.5295   2.3095  25.4148

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.862007   0.872084   49.15   <2e-16 ***
lstat       -2.332821   0.123803  -18.84   <2e-16 ***
I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.524 on 503 degrees of freedom
Multiple R-squared:  0.6407,    Adjusted R-squared:  0.6393
F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = medv ~ log(rm), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-19.487  -2.875  -0.104   2.837  39.816

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -76.488      5.028  -15.21   <2e-16 ***
log(rm)       54.055      2.739   19.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.915 on 504 degrees of freedom
Multiple R-squared:  0.4358,    Adjusted R-squared:  0.4347
F-statistic: 389.3 on 1 and 504 DF,  p-value: < 2.2e-16
```

5) Actividad 5- Los resultados obtenidos para la Actividad 5 fueron los siguientes :

A anova: 2 × 6

Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1 504	19472.38	NA	NA	NA	NA
2 503	15347.24	1	4125.138	135.1998	7.630116e-28



A data.frame: 400 × 11

	Sales <dbl>	CompPrice <dbl>	Income <dbl>	Advertising <dbl>	Population <dbl>	Price <dbl>	ShelveLoc <fct>	Age <dbl>	Education <dbl>	Urban <fct>	US <fct>
1	9.50	138	73	11	276	120	Bad	42	17	Yes	Yes
2	11.22	111	48	16	260	83	Good	65	10	Yes	Yes
3	10.06	113	35	10	269	80	Medium	59	12	Yes	Yes
4	7.40	117	100	4	466	97	Medium	55	14	Yes	Yes
5	4.15	141	64	3	340	128	Bad	38	13	Yes	No
6	10.81	124	113	13	501	72	Bad	78	16	No	Yes
7	6.63	115	105	0	45	108	Medium	71	15	Yes	No
8	11.85	136	81	15	425	120	Good	67	10	Yes	Yes
9	6.54	132	110	0	108	124	Medium	76	10	No	No
10	4.69	132	113	0	131	124	Medium	76	17	No	Yes
11	9.01	121	78	9	150	100	Bad	26	10	No	Yes
12	11.96	117	94	4	503	94	Good	50	13	Yes	Yes
13	3.98	122	35	2	393	136	Medium	62	18	Yes	No
14	10.96	115	28	11	29	86	Good	53	18	Yes	Yes
15	11.17	107	117	11	148	118	Good	52	18	Yes	Yes
16	8.71	149	95	5	400	144	Medium	76	18	No	No
17	7.58	118	32	0	284	110	Good	63	13	Yes	No
18	12.29	147	74	13	251	131	Good	52	10	Yes	Yes
19	13.91	110	110	0	408	68	Good	46	17	No	Yes
20	8.73	129	76	16	58	121	Medium	69	12	Yes	Yes
21	6.41	125	90	2	367	131	Medium	35	18	Yes	Yes
22	12.13	134	29	12	239	109	Good	62	18	No	Yes
23	5.08	128	46	6	497	138	Medium	42	13	Yes	No
24	5.87	121	31	0	292	109	Medium	79	10	Yes	No
25	10.14	145	119	16	294	113	Bad	42	12	Yes	Yes
26	14.90	139	32	0	176	82	Good	54	11	No	No
27	8.33	107	115	11	496	131	Good	50	11	No	Yes
28	5.27	98	118	0	19	107	Medium	64	17	Yes	No
29	2.99	103	74	0	359	97	Bad	55	11	Yes	Yes
30	7.81	104	99	15	226	102	Bad	58	17	Yes	Yes
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

```
 [1] "Sales"       "CompPrice"   "Income"      "Advertising" "Population"
 [6] "Price"       "ShelveLoc"   "Age"         "Education"   "Urban"
[11] "US"
```

```
Call:
lm(formula = Sales ~ . + Income:Advertising + Price:Age, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9208 -0.7503  0.0177  0.6754  3.3413

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.5755654  1.0087470   6.519 2.22e-10 ***
CompPrice           0.0929371  0.0041183  22.567  < 2e-16 ***
Income              0.0108940  0.0026044   4.183 3.57e-05 ***
Advertising         0.0702462  0.0226091   3.107 0.002030 **
Population          0.0001592  0.0003679   0.433 0.665330
Price              -0.1008064  0.0074399 -13.549  < 2e-16 ***
ShelveLocGood       4.8486762  0.1528378  31.724  < 2e-16 ***
ShelveLocMedium     1.9532620  0.1257682  15.531  < 2e-16 ***
Age                -0.0579466  0.0159506  -3.633 0.000318 ***
Education          -0.0208525  0.0196131  -1.063 0.288361
UrbanYes            0.1401597  0.1124019   1.247 0.213171
USYes              -0.1575571  0.1489234  -1.058 0.290729
Income:Advertising  0.0007510  0.0002784   2.698 0.007290 **
Price:Age           0.0001068  0.0001333   0.801 0.423812
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.011 on 386 degrees of freedom
Multiple R-squared:  0.8761,    Adjusted R-squared:  0.8719
F-statistic:   210 on 13 and 386 DF,  p-value: < 2.2e-16
```

```
A matrix: 3 × 2 of type dbl
       Good Medium
Bad    0    0
Good   1    0
Medium 0    1
```

6) Actividad 6- Los resultados obtenidos para la Actividad 6 fueron los siguientes :

```
function ()
{
    library(ISLR)
    library(MASS)
    print(" The libraries have been loaded .")
}
```

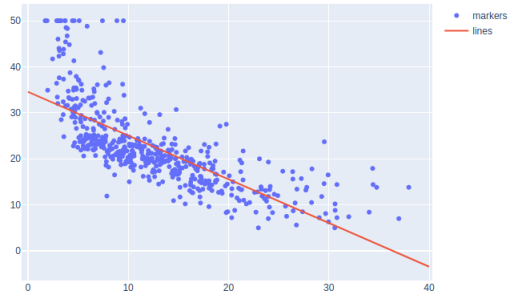
```
[1] " The libraries have been loaded ."
```

B. Ejercicios en Python:

1) Actividad 1- Los resultados obtenidos para la Actividad 1 fueron los siguientes :

```
Index(['crim', 'zn', 'indus', 'chas', 'nox', 'rm', 'age', 'dis', 'rad', 'tax',
       'ptratio', 'black', 'lstat', 'medv'],
      dtype='object')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   medv   R-squared:                       0.544
Model:                            OLS   Adj. R-squared:                  0.543
Method:                 Least Squares   F-statistic:                     601.6
Date:                Wed, 07 Oct 2020   Prob (F-statistic):          5.08e-88
Time:                        19:24:46   Log-Likelihood:                -1641.5
No. Observations:                 506   AIC:                             3287.
Df Residuals:                     504   BIC:                             3295.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     34.5538      0.563     61.415      0.000      33.448      35.659
lstat         -0.9500      0.039    -24.528      0.000      -1.026      -0.874
==============================================================================
Omnibus:                      137.043   Durbin-Watson:                   0.892
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              291.373
Skew:                           1.453   Prob(JB):                     5.36e-64
Kurtosis:                       5.319   Cond. No.                         29.7
==============================================================================
```



2) Actividad 2- Los resultados obtenidos para la Actividad 2 fueron los siguientes :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   medv   R-squared:                       0.551
Model:                            OLS   Adj. R-squared:                  0.549
Method:                 Least Squares   F-statistic:                     309.0
Date:                Wed, 07 Oct 2020   Prob (F-statistic):           2.98e-88
Time:                        19:06:47   Log-Likelihood:                -1637.5
No. Observations:                 506   AIC:                             3281.
Df Residuals:                     503   BIC:                             3294.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     33.2228      0.731     45.458      0.000      31.787      34.659
lstat         -1.0321      0.048    -21.416      0.000      -1.127      -0.937
age            0.0345      0.012      2.826      0.005       0.011       0.059
==============================================================================
Omnibus:                      124.288   Durbin-Watson:                   0.945
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              244.026
Skew:                           1.362   Prob(JB):                     1.02e-53
Kurtosis:                       5.038   Cond. No.                         201.
==============================================================================
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   medv   R-squared:                       0.741
Model:                            OLS   Adj. R-squared:                  0.734
Method:                 Least Squares   F-statistic:                     108.1
Date:                Tue, 06 Oct 2020   Prob (F-statistic):          6.72e-135
Time:                        21:21:40   Log-Likelihood:                -1498.8
No. Observations:                 506   AIC:                             3026.
Df Residuals:                     492   BIC:                             3085.
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     36.4595      5.103      7.144      0.000      26.432      46.487
crim          -0.1080      0.033     -3.287      0.001      -0.173      -0.043
zn             0.0464      0.014      3.382      0.001       0.019       0.073
indus          0.0206      0.061      0.334      0.738      -0.100       0.141
chas           2.6867      0.862      3.118      0.002       0.994       4.380
nox          -17.7666      3.820     -4.651      0.000     -25.272     -10.262
rm             3.8099      0.418      9.116      0.000       2.989       4.631
age            0.0007      0.013      0.052      0.958      -0.025       0.027
dis           -1.4756      0.199     -7.398      0.000      -1.867      -1.084
rad            0.3060      0.066      4.613      0.000       0.176       0.436
tax           -0.0123      0.004     -3.280      0.001      -0.020      -0.005
ptratio       -0.9527      0.131     -7.283      0.000      -1.210      -0.696
black          0.0093      0.003      3.467      0.001       0.004       0.015
lstat         -0.5248      0.051    -10.347      0.000      -0.624      -0.425
==============================================================================
Omnibus:                      178.041   Durbin-Watson:                   1.078
```

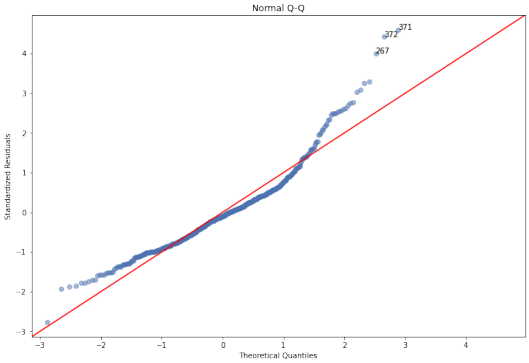
3) Actividad 3- Los resultados obtenidos para la Actividad 3 fueron los siguientes :

OLS Regression Results

Dep. Variable:	medv	R-squared:	0.556
Model:	OLS	Adj. R-squared:	0.553
Method:	Least Squares	F-statistic:	209.3
Date:	Tue, 06 Oct 2020	Prob (F-statistic):	4.86e-88
Time:	21:43:25	Log-Likelihood:	-1635.0
No. Observations:	506	AIC:	3278.
Df Residuals:	502	BIC:	3295.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	36.0885	1.470	24.553	0.000	33.201	38.976
lstat	-1.3921	0.167	-8.313	0.000	-1.721	-1.063
age	-0.0007	0.020	-0.036	0.971	-0.040	0.038
lstat:age	0.0042	0.002	2.244	0.025	0.001	0.008

Omnibus:	135.601	Durbin-Watson:	0.965
Prob(Omnibus):	0.000	Jarque-Bera (JB):	296.955
Skew:	1.417	Prob(JB):	3.29e-65
Kurtosis:	5.461	Cond. No.	6.88e+03



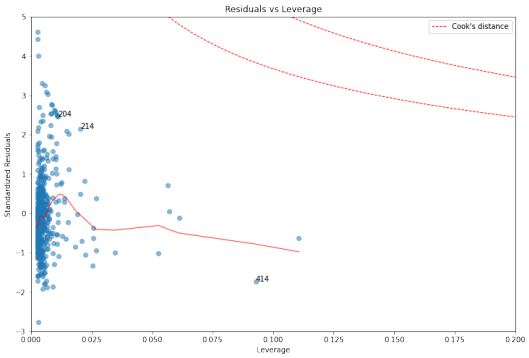
4) Actividad 4- Los resultados obtenidos para la Actividad 4 fueron los siguientes :

OLS Regression Results

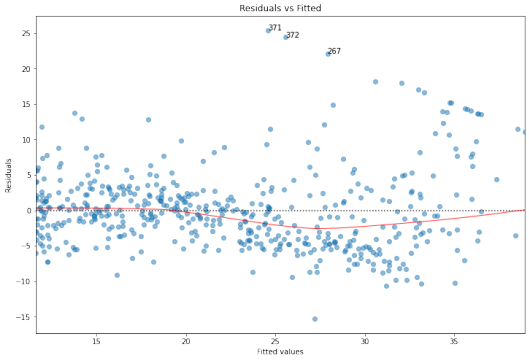
Dep. Variable:	medv	R-squared:	0.641
Model:	OLS	Adj. R-squared:	0.639
Method:	Least Squares	F-statistic:	448.5
Date:	Tue, 06 Oct 2020	Prob (F-statistic):	1.56e-112
Time:	22:57:28	Log-Likelihood:	-1581.3
No. Observations:	506	AIC:	3169.
Df Residuals:	503	BIC:	3181.
Df Model:	2		
Covariance Type:	nonrobust		

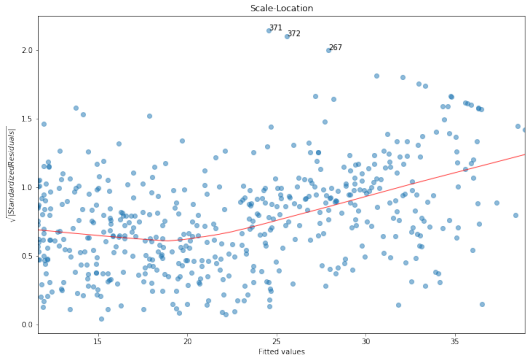
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	42.8620	0.872	49.149	0.000	41.149	44.575
lstat	-2.3328	0.124	-18.843	0.000	-2.576	-2.090
np.power(lstat, 2)	0.0435	0.004	11.628	0.000	0.036	0.051

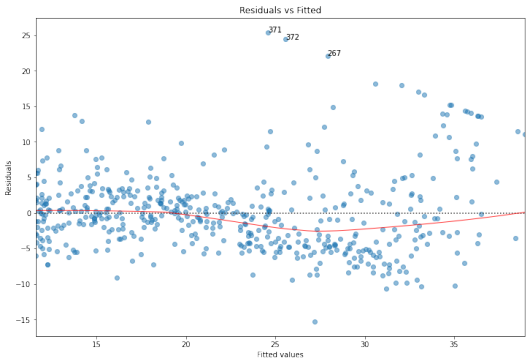
Omnibus:	107.006	Durbin-Watson:	0.921
Prob(Omnibus):	0.000	Jarque-Bera (JB):	228.388
Skew:	1.128	Prob(JB):	2.55e-50
Kurtosis:	5.397	Cond. No.	1.13e+03



	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	504.0	19472.381418	0.0	NaN	NaN	NaN
1	503.0	15347.243158	1.0	4125.13826	135.199822	7.630116e-28







OLS Regression Results

Dep. Variable:	medv	R-squared:	0.682
Model:	OLS	Adj. R-squared:	0.679
Method:	Least Squares	F-statistic:	214.2
Date:	Wed, 07 Oct 2020	Prob (F-statistic):	8.73e-122
Time:	19:40:20	Log-Likelihood:	-1550.6
No. Observations:	506	AIC:	3113.
Df Residuals:	500	BIC:	3139.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	34.3399	0.469	73.205	0.000	33.418	35.262
lstat	-0.9331	0.032	-28.895	0.000	-0.997	-0.870
poly(lstat, 5)[0]	-2.7134	0.037	-73.016	0.000	-2.786	-2.640
poly(lstat, 5)[1]	-64.2272	5.215	-12.316	0.000	-74.473	-53.982
poly(lstat, 5)[2]	-27.0511	5.215	-5.187	0.000	-37.297	-16.805
poly(lstat, 5)[3]	25.4517	5.215	4.881	0.000	15.206	35.697
poly(lstat, 5)[4]	19.2524	5.215	3.692	0.000	9.007	29.498

Omnibus:	144.085	Durbin-Watson:	0.987
Prob(Omnibus):	0.000	Jarque-Bera (JB):	494.545
Skew:	1.292	Prob(JB):	4.08e-108
Kurtosis:	7.096	Cond. No.	1.67e+18

5) Actividad 5- Los resultados obtenidos para la Actividad 5 fueron los siguientes :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.278
Model:                            OLS   Adj. R-squared:                  0.274
Method:                 Least Squares   F-statistic:                     76.45
Date:                Wed, 07 Oct 2020   Prob (F-statistic):           8.21e-29
Time:                        19:17:25   Log-Likelihood:                -917.19
No. Observations:                 400   AIC:                             1840.
Df Residuals:                     397   BIC:                             1852.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     10.2055      0.376     27.177      0.000       9.467      10.944
Income:Advertising     0.0015      0.000      6.779      0.000       0.001       0.002
Price:Age     -0.0006   5.49e-05    -10.130      0.000      -0.001      -0.000
==============================================================================
Omnibus:                        1.974   Durbin-Watson:                   1.918
Prob(Omnibus):                  0.373   Jarque-Bera (JB):                1.726
Skew:                           0.138   Prob(JB):                        0.422
Kurtosis:                       3.167   Cond. No.                     2.04e+04
==============================================================================
```

6) Actividad 6- Los resultados obtenidos para la Actividad 6 fueron los siguientes :

```
The libraries have been loaded .
```

V. CONCLUSIONES

Consideramos que esta práctica es muy importante debido a que son las bases para empezar a trabajar modelos más complejos, dado que introdujo métricas que nos ayudan a lograr nuestro objetivo de construir un modelo estadístico usando como dichas métricas que son elementales en el aprendizaje estadístico. Por lo cual es fundamental debido a que introduce la forma de construir un modelo de regresión múltiple, obteniendo un modelo multidimensional, en el cual que es permitido relacionar variables independientes entre sí, lo que señala que existe una interacción entre ambas, transformando nuestro modelo a un comportamiento no lineal. El cual si vienen al tener un comportamiento no lineal, sigue siendo lineal con respecto a los parámetros del modelo.

VI. REFERENCIAS

- B.Salazar. (2019-Jul-01).Regresión lineal. [Online]. Disponible en:https://www.ingenieriaindustrialonline.com/pronostico-de-la-demanda/regresion-lineal/
- B.Requena. (2014).VARIANZA.[Online].Disponible en: https://www.universoformulas.com/estadistica/descriptiva/varianza/
- P.Vinuesa.(2016-Oct-22). Tema 9 - Regresión lineal simple y polinomial: teoría y práctica. [Online]. Disponible en: https://www.ccg.unam.mx/ vinuesa/R4biosciences/docs/Tema9_regresion.html
- R.Rodrigo. (s.d).Factor de inflación de varianza.[Online]. Disponible en: https://exonegocios.com/factor-de-inflacion-de-varianza/