Laboratorio #3: Modelos de Regresión en R

Ejercicio #1: utilizando R realice una función que dado un dataframe cualquiera de dos columnas, donde la primera (índice 1) sea el valor de la variable independiente (X) y la segunda sea el valor de una variable dependiente (Y), devuelva una lista con los siguientes elementos:

1) Un arreglo con los valores de los estimadores para β_0 y β_1.
2) El valor del coeficiente de determinación r^2 del modelo.
3) El coeficiente de correlación r (raíz cuadrada de r^2).
4) Un arreglo con los valores de los residuos.
5) Una gráfica con la nube de puntos y la recta de regresión del modelo.

Nota: Para este ejercicio **NO** está permitido utilizar la función lm()para calcular ninguno de los elementos solicitados (incisos 1 al 4), sin embargo puede utilizar ggplot para realizar la gráfica del inciso 5.

Recuerde de su curso de Econometria que:

$$\beta_1 = \frac{\sum x \sum y - n\sum xy}{(\sum x)^2 - n\sum x^2}$$

$$\beta_0 = \frac{\sum y - \beta_1 \sum x}{n}$$

$$r^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

$$r = \pm\sqrt{r^2}$$

Recuerde que n representa la cantidad de filas en el dataset

Ejercicio #2: Para este ejercicio se le solicita que desarrolle las siguientes actividades utilizando RStudio Con el dataset Admissions adjunto a este laboratorio realice lo siguiente:

1. Realice un análisis estadístico sobre todas las variables del dataset, recuerde que pude usar la función summary().
2. Realice una gráfica de densidad para cada una de las variables numéricas en el dataset: GRE.Score, TOEFEL.Score, CGPA y Chance of Admit.
3. Realice una gráfica de correlación entre las variables del inciso anterior.
4. Realice comentarios sobre el análisis estadístico de las variables numéricas y la gráfica de correlación.
5. Realice un scatter plot (nube de puntos) de todas las variables numéricas contra la variable Chance of Admit.
6. Utilizando la función train y trainControl para crear un cross-validation y le permita evaluar los siguientes modelos:
   - Chance of Admit ~ TOEFEL.Score.
   - Chance of Admit ~ CGPA.
   - Chance of Admit ~ GRE.Score.
   - Chance of Admit ~ TOEFEL.Score + CGPA.
   - Chance of Admit ~ TOEFEL.Score + GRE.Score.
   - Chance of Admit ~ GRE.Score + CGPA.
   - Chance of Admit ~ TOEFEL.Score + CGPA + GRE.Score.

   Posteriormente cree una lista ordenando de mejor a peor cual es el mejor modelo en predicción, recuerde que es necesario caclular el RMSE para poder armar correctamente la lista.

Ejercicio #3: A continuación se le muestran tres imágenes que muestran los resultados obtenidos de correr la función summary() a dos modelos de regresión lineal, para este ejercicio se le solicita que realice la interpretación de las tablas resultantes. Recuerde tomar en cuenta la signficancia de los parámetros (signfícancia local), la signficancia del modelo (signficancia global), el valor del r^2: y cualquier observación que considere relevante para determinar si el modelo estructuralmente es adecuado o no.

Modelo #1:

```
Call:
lm(formula = ROLL ~ UNEM, data = datavar)

Residuals:
    Min      1Q  Median      3Q     Max
-7640.0 -1046.5   602.8  1934.3  4187.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3957.0     4000.1   0.989   0.3313
UNEM          1133.8      513.1   2.210   0.0358 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3049 on 27 degrees of freedom
Multiple R-squared: 0.1531, Adjusted R-squared: 0.1218
F-statistic: 4.883 on 1 and 27 DF,  p-value: 0.03579
```

En este modelo la formula es ROLL-UNEM, la cual toma como la variable salida es ROLL y la predictora es UNEM.

El valor intercepto de estos es 3957 de estimado, con un error hacia 4000.1, lo cual muestra que la desviación es muy baja. Dado que este es el valor intercepto cuando UNEM es 0 ROLL debe ser 3957. El valor p de este es 0.3313 lo cual quiere decir que no es cerca del valor de significancia de 0.05.

El modelo tienen un R cuadrado de 0.1531 sea un 15.31% de variabilidad con la variable ROLL

En resumen: el modelo es significativo dado estas condiciones, el R cuadrado es bajo así que la variabilidad de los datos es baja y la variable UNEM es significativa. Aun así dado que el intercepto no es significativo puede ser que se necesiten mas variables para la determinación del modelo.

Modelo #2:

```
Call:
lm(formula = ROLL ~ UNEM + HGRAD + INC, data = datavar)

Residuals:
     Min       1Q   Median       3Q      Max
-1148.840  -489.712    -1.876   387.400  1425.753

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.153e+03  1.053e+03  -8.691 5.02e-09 ***
UNEM         4.501e+02  1.182e+02   3.809 0.000807 ***
HGRAD        4.065e-01  7.602e-02   5.347 1.52e-05 ***
INC          4.275e+00  4.947e-01   8.642 5.59e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 670.4 on 25 degrees of freedom
Multiple R-squared: 0.9621, Adjusted R-squared: 0.9576
F-statistic: 211.5 on 3 and 25 DF,  p-value: < 2.2e-16
```

El intercepto en este caso es de -9.153c+03, esta seria la varianza de la variable ROLL, el valor asociado p es muy pequeño: 5.02c-09 asi que este es significativo.

El coeficiente de UNEM nos indica que este hará que de forma directamente proporcional crezca en una proporción de 4.501c+02, suponiendo que se mantengan constantes las demás. El coeficiente p es también muy pequeño así que es significativo

El coeficiente HGRAD nos indica que este hará que de forma directamente proporcional crezca en una proporción de 4.065c-01, suponiendo que se mantengan constantes las demás. El coeficiente p es también muy pequeño así que es significativo

El coeficiente INC nos indica que este hará que de forma directamente proporcional crezca en una proporción de 4.275, suponiendo que se mantengan constantes las demás. El coeficiente p es también muy pequeño así que es significativo

El R cuadrado del modelo es de 0.9621 lo cual indica un 96.21% de la variabilidad en la respuesta ROLL. El ajustado seria de 0.9576

El valor F es de 211.5 con un valor p bastante pequeño, lo que indica el modelo es significativo

En resumen se ve que todos los índices estadísticos son significativos y el R cuadrado es alto. Aun así siempre debemos de verificar con las suposiciones de la regresión lineal.

Modelo #3:

```
Call:
lm(formula = Cab.Price ~ Months, data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max
-11.034  -2.305  -1.034   2.764   9.241

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.6826     3.2377   22.45 6.92e-10 ***
Months        4.8626     0.3495   13.91 7.18e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.657 on 10 degrees of freedom
Multiple R-squared:  0.9509,    Adjusted R-squared:  0.946
F-statistic: 193.6 on 1 and 10 DF,  p-value: 7.181e-08
```

El valor del intercepto es de 72.6826 o sea que cabPrice al ser la x 0 esta seria 72.6826. Su valor P es de 6.92e-10 asi que es estadísticamente significativo.

El coeficiente Months nos indica que este hará que de forma directamente proporcional crezca en una proporción de 4.8626, suponiendo que se mantengan constantes las demás. El coeficiente p es también muy pequeño así que es significativo

El R cuadrado del modelo es de 0.9509 lo cual indica un 95.09% de la variabilidad en la respuesta CabPrice. El ajustado seria de 0.9576

El valor F es de 193.6 con un valor p bastante pequeño, lo que indica el modelo es significativo

En resumen se ve que todos los índices estadísticos son significativos y el R cuadrado es alto. Aun así siempre debemos de verificar con las suposiciones de la regresión lineal