Universidad Simón Bolívar

Departamento de Cómputo Científico y Estadística

Estadística Para Ingenieros CO-3321

Prof. Desiree Villalta

Informe de Laboratorio: Laboratorio 4. Modelos lineales y Anova

Estudiantes:

Carlos Sivira 15-11377

José Barrera 15-10123

1. **La variable Y en el archivo de texto "calificaciones.txt" indica la calificación final alcanzada por estudiantes en una evaluación final. Así mismo, las variables E1, E2, E3, E4, E5 y E6 son las calificaciones alcanzadas en evaluaciones pasadas que se cree pueden explicar el rendimiento del estudiante en la evaluación final.**

   a. Realice un análisis descriptivo de los datos (histograma, gráfico de cajas, número de la muestra, mínimo, cuartiles, media y desviación).

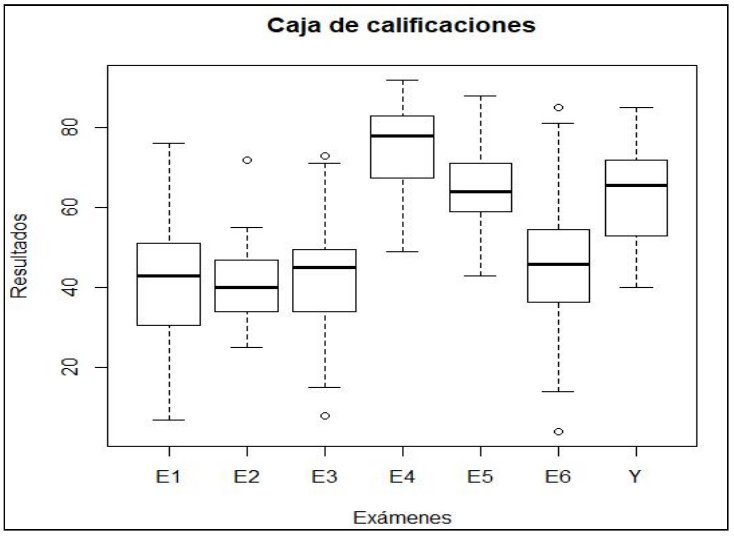


Diagrama de Caja de las Variables

E6	E5	E4	E3
"Min. : 4.00 "	"Min. :43.0 "	"Min. :49.00 "	"Min. : 8.00 "
"1st Qu.:37.75 "	"1st Qu.:59.0 "	"1st Qu.:67.75 "	"1st Qu.:34.00 "
"Median :46.00 "	"Median :64.0 "	"Median :78.00 "	"Median :45.00 "
"Mean :45.18 "	"Mean :64.5 "	"Mean :75.10 "	"Mean :42.25 "
"3rd Qu.:54.25 "	"3rd Qu.:71.0 "	"3rd Qu.:83.00 "	"3rd Qu.:49.25 "
"Max. :85.00 "	"Max. :88.0 "	"Max. :92.00 "	"Max. :73.00 "
sd "18.71"	"9.28"	"10.21"	"16.72"

E2	E1	Y
"Min. :25.00 "	"Min. : 7.00 "	"Min. :40.00 "
"1st Qu.:34.00 "	"1st Qu.:30.75 "	"1st Qu.:53.00 "
"Median :40.00 "	"Median :43.00 "	"Median :65.50 "
"Mean :40.45 "	"Mean :40.80 "	"Mean :63.77 "
"3rd Qu.:47.00 "	"3rd Qu.:51.00 "	"3rd Qu.:72.00 "
"Max. :72.00 "	"Max. :76.00 "	"Max. :85.00 "
sd "8.41"	"15.67"	"12.00"

Summary y Desviación Estándar de las Variables

Se observa que las cajas correspondientes a los resultados de los exámenes 2, 3 y 6 presentan datos atípicos. Además en las cajas 4 y 5, la agrupación de la mayoría de los datos se encuentran por encima del resto. Con lo cual, el conjunto de valores de Y hereda esta tendencia. También se destaca que E1 y E6 abarca un rango de valores mayor al resto, y E2 lo contrario. Por último, los valores en Y no se encuentran normalmente distribuidos, a pesar de que los valores de los demás si (excepto E1, E2).

b. <u>Realice un gráfico de dispersión y una matriz de correlación de las variables independientes respecto a Y. Interprete los resultados.</u>

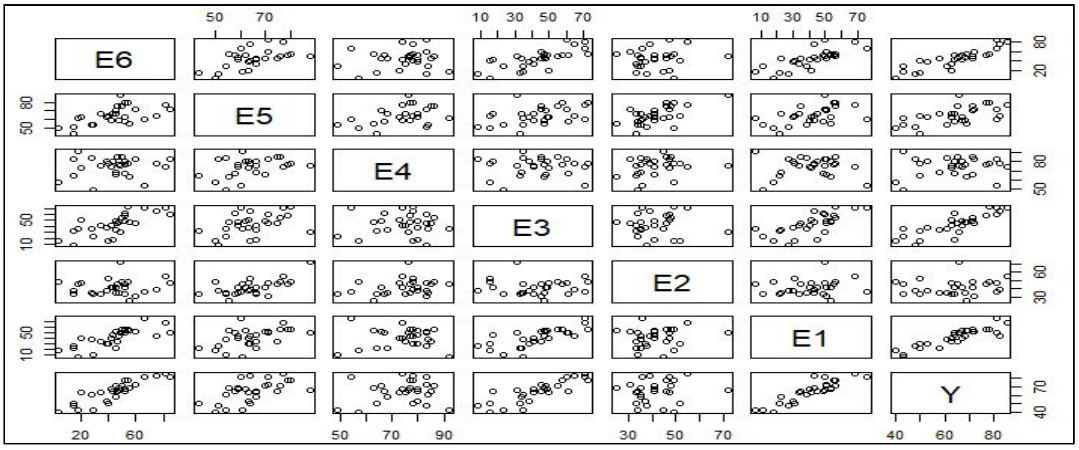


Gráfico de Dispersión

	E6	E5	E4	E3	E2	E1	Y
E6	1.0000000	0.25345678	-0.55196859	0.90344430	-0.8035864	0.91439288	0.9661236
E5	0.2534568	1.00000000	0.05160964	0.07844537	0.1309178	0.09050384	0.1807454
E4	-0.5519686	0.05160964	1.00000000	-0.66545216	0.2159270	-0.68641138	-0.5949314
E3	0.9034443	0.07844537	-0.66545216	1.00000000	-0.8265635	0.93092228	0.9625431
E2	-0.8035864	0.13091777	0.21592697	-0.82656353	1.0000000	-0.78654034	-0.8454770
E1	0.9143929	0.09050384	-0.68641138	0.93092228	-0.7865403	1.00000000	0.9724744
Y	0.9661236	0.18074536	-0.59493142	0.96254314	-0.8454770	0.97247443	1.0000000

Matriz de Correlación

Se observa en la matriz de correlación que la variable independiente menos relacionada con la variable dependiente (Y) es E5, mientras que E1, E3 y E6, son las más relacionadas con esta. Esto se corresponde con lo observado en gráfico de dispersión, donde E1, E3 y E6 son las más lineales y E5 la que menos sigue el patrón. Adicionalmente cabe destacar que E1, E3 y E6 también se encuentran fuertemente relacionadas entre ellas.

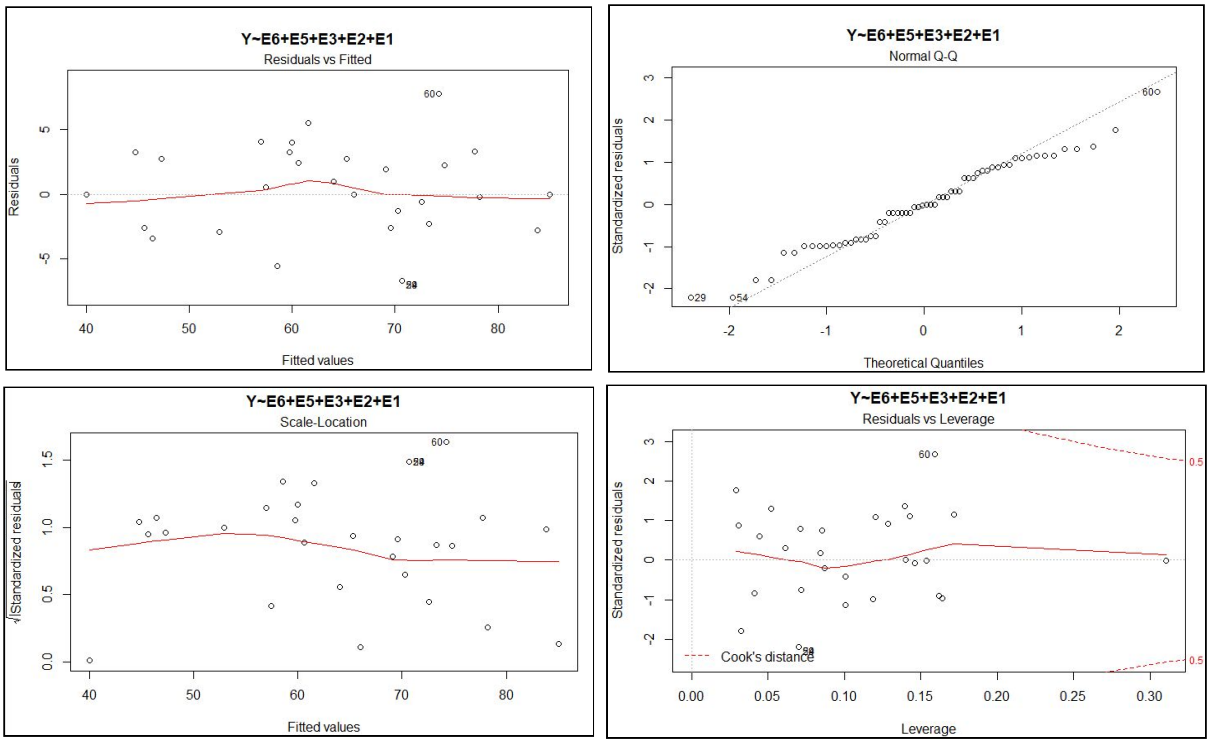
c. ¿Cuál es el modelo que explica mejor la variabilidad de Y ? Incluya todas las pruebas necesarias para llegar a este modelo. Utilice un nivel de significancia de 0.05.

Para calcular el mejor modelo, se calculan y analizan los modelos lineales entre la variable dependiente Y contra las independientes (E1, E2, E3, E4, E5, E6). Luego se aplica el método de regresión hacia atrás paso a paso, comenzando con el modelo múltiple que tiene todas las variables independientes para luego eliminar la variable menos significativa y así hallar el modelo que mejor se ajusta con significancia 0.05. Al comparar todos los modelos obtenidos, el que mejor se ajusta y corresponde con la significancia *0.05* es el modelo lineal múltiple Y~E6+E5+E3+E2+E1, con un error de *3.153*, un ajuste de *0.931*, y posee una simetría aceptable de los cuartiles, máximo y mínimo respecto a la mediana. Cabe destacar que todos los modelos lineales simples, obtuvieron un error más elevado y un ajuste menor que el mejor modelo. Aunque el modelo múltiple anterior al escogido posee un mejor ajuste, no cumple con la significancia requerida. Los modelos múltiples posteriores desmejoraron el ajuste e incrementaron el error. Para crear los modelos se usa la función *lm*, y para su análisis *summary*.

d. Realice un análisis de residuos al modelo ganador.

Se puede observar que para este modelo los datos son independientes de los residuales pues, en la gráfica Residual vs Fitted, no se aprecia ningún patrón, la forma de los residuales parece razonablemente normal, y además es

homocedástico, ya que en la gráfica Residuals vs Leverage no se aprecia ningún patrón.



Gráficos de Análisis De Residuos

e. Con los datos "calificaciones_prediccion.txt" realice una predicción de la variable Y (con el mejor de los modelos) y haga un histograma y boxplot de los residuos de predicción (valor observado - predicción del modelo) para concluir con relación al poder predictivo del modelo.

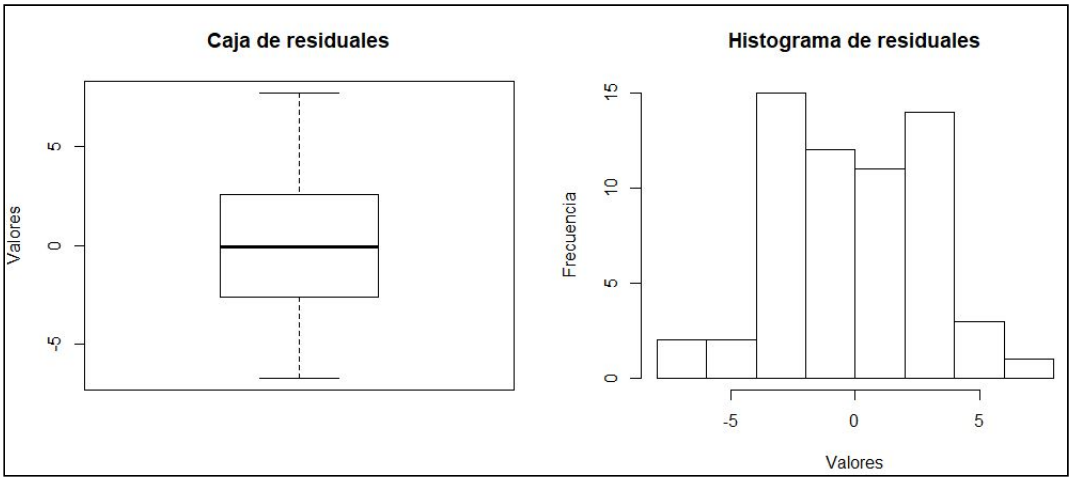


Diagrama de Caja de los residuos e histograma de residuos

Se observa que los datos son simétricos respecto al origen y no hay datos atípicos. Adicionalmente el *summary* indica que la media es 0, mientras que la mediana es cercana a 0 con lo cual el modelo en promedio hace una buena estimación de los datos, pues, las predicciones en su mayoría se encuentran a 2.57 puntos por debajo o 2.48 puntos por arriba del valor real (primer cuartil -2.57261, tercer cuartil 2.48407).

2. **Los miembros de un equipo ciclista se dividen al azar en tres grupos que entrenan con métodos diferentes. El primer grupo realiza largos recorridos a ritmo pausado, el segundo grupo realiza series cortas de alta intensidad y el tercero trabaja en el gimnasio con pesas y se ejercita en el pedaleo de alta frecuencia. Después de un mes de entrenamiento se realiza un test de rendimiento consistente en un recorrido cronometrado de 9 Km. Los tiempos empleados fueron los siguientes:**

   a. A un nivel de confianza del 95% ¿Puede considerarse que los tres métodos producen resultados equivalentes? O por el contrario ¿Hay algún método superior a los demás?

      Para ver si los tres métodos producen resultados equivalentes es necesario usar una prueba de hipótesis de la igualdad de las medias sobre cada grupo de resultados por método.

      - Se toma como hipótesis nula Ho: La media de todos los métodos son iguales.
      - Se toma como hipótesis alternativa Ha: La media de alguno de los métodos difiere.

      Se crea un factor sobre los datos suministrados mediante la función *factor()*. Luego se genera un modelo lineal con dicho factor usando la función *lm()*. Este modelo es usado para generar la tabla ANDEVA mediante la función *anova()*. Se obtuvieron los siguiente resultados:

```
Analysis of Variance Table

Response: dat
          Df Sum Sq Mean Sq F value   Pr(>F)
fact       2   26.8 13.4000  9.3488 0.003568 **
Residuals 12   17.2  1.4333
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se observa que el p-valor (*0.003568*) obtenido es lo suficientemente pequeño para rechazar la hipótesis nula, lo que indica que no existe suficiente evidencia para afirmar que las medias de cada uno de los métodos son iguales.

Para determinar si alguno de los método es superior a los demás, se utiliza la función *pairwise.t.test()* para realizar múltiples pruebas de medias sobre los métodos. Se obtuvieron los siguiente resultados:

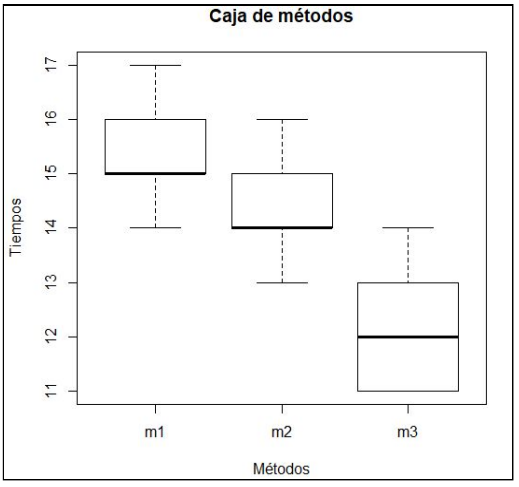
```
        Pairwise comparisons using t tests with pooled SD

data:  dat and fact

   m1     m2
m2 0.2112 -
m3 0.0035 0.0264

P value adjustment method: holm
```

Se evidencia de que no es posible rechazar la hipótesis nula de que las medias del método 1 y 2 sean iguales. Sin embargo, existe suficiente evidencia para rechazar la hipótesis nula sobre las medias de los métodos 1 y 3 y sobre los métodos 2 y 3, ya que, el p-valor obtenido es menor al valor de alpha = 0.05. Por lo tanto, no es posible determinar si un método es mejor que otro, pero si es posible decir que el método 3 es el de peor desempeño.



Es posible confirmar este resultado observando el diagrama de caja generado por el modelo lineal del problema. Por otro lado, el rango de valores que tiene la caja del método 3 es mucho menor que el rango de valores de las cajas 1 y 2.