Trabajo regresión lineal múltiple

Estudiantes

Rojas Martinez, Ivan Santiago
Hernandez Ruiz, Juan Sebastian
Londoño Montoya, Wilson Duván
Perez Garcia, Pablo

Docente

Isabel Cristina Ramirez Guevara

Asignatura

Análisis de Regresión

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
Enero de 2022

Índice

Índice de figuras

Índice de cuadros

Se realizará una análisis de regresión lineal múltiple(RLM):

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i,\ \varepsilon \overset{iid}{\sim} N(0, \sigma^2)$$

Con la intención de validar si dicho modelo es adecuado para explicar la posibilidad de ser admitido a una carrera de postgrado en la india teniendo en cuenta determinadas pruebas de aptitud.

1. Base de datos

1.1. Breve Descripción de los Datos

La base de datos disponible en Kaggle corresponde a puntajes de admision creados para la predicción de las admisiones de posgrado en La India. Cuenta con 400 observaciones y 9 variables. De las cuales se consideran los primeros 100 estudiantes y 6 variables de interes por indicación de la docente.

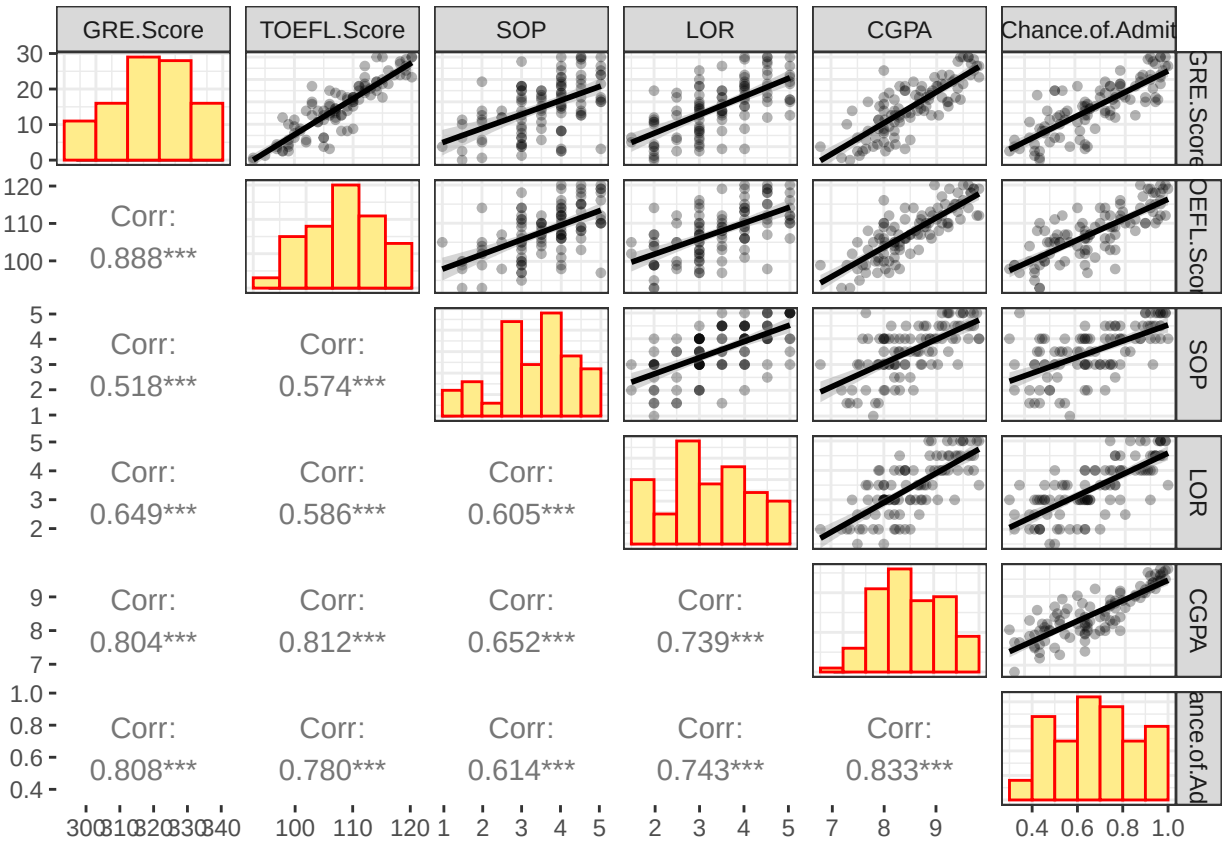
Variables	**Descripción**
Chance.of.Admit:	Posibilidad de ser admitido. Variable numérica continua de 0-1.
GRE Score:	Puntaje de Examen que proporciona a las escuelas una medida común para la comparación de la capacidad de razonamiento verbal, razonamiento cuantitativo, y habilidades para pensar y escribir de forma analítica. Variable numérica que toma valores de 294 - 340.
TOEFL Score:	Puntaje en prueba estandarizada de dominio del idioma inglés. Variable numérica que toma valores del 93 - 120.
SOP:	Puntaje en Ensayo de admisión o solicitud de postgrado. Variable numérica que toma valores del 1 - 5, tomando el valor medio entre cada par de enteros en el intervalo.
LOR:	Puntaje en Carta de recomendación. Variable numérica que toma valores del 1.5 - 5, tomando el valor medio entre cada par de enteros en el intervalo.
CGPA:	Promedio general acumulado en el pregrado. Variable numérica que toma valores del 6.8 - 9.8.

1.2. Renombrando las variables:

- **GRE Score** $= X_1$
- **TOEFL Score** $= X_2$
- **SOP** $= X_3$
- **LOR** $= X_4$
- **CGPA** $= X_5$

2. Análisis descriptivo

2.1. Grafico de dispersión con Matriz de Correlaciones y conclusiones



- Se observan relaciones de interés.
- La variable **Chance.of.Admit** (Posibilidad de ser admitido) se encuentran altamente correlacionada con las variables **GRE.Score**, **TOFL.Score**, **SOP**, **LOR** y **CGPA**

con correlaciones de **0.808**, **0.780**, **0.614**, **0.743** y **0.833** respectivamente. Con relaciones del tipo lineales positivas.

- La variable **CGPA** (Promedio general acumulado en el pregrado) se encuentran altamente correlaciona con las variables **GRE.Score**, **TOFL.Score**, **SOP** y **LOR** con correlaciones de **0.804**, **0.812**,**0.652**y **0.739** respectivamente. Con relaciones del tipo lineales positivas. Esto nos puede indicar redundancia en el modelo o multicolinealidad lo cual validaremos más adelante.
- La variable **GRE.Score** se encuentra altamente correlacionadas con las variables **TOFL.Score** y **CGPA**. Y moderadamente con las variables **SOP** y **LOR**.

3. Modelo Ajustado de Regresion Lineal múltiple (MRLM)

3.1. Tabla de parámetros ajustados

Cuadro 2: Resumen de los coeficientes

	Estimación	Error estándar	T_0	Valor P
β_0	-1.7723	0.3007	-5.8939	0.0000
β_1	0.0041	0.0017	2.4400	0.0166
β_2	0.0029	0.0031	0.9417	0.3488
β_3	0.0120	0.0119	1.0098	0.3152
β_4	0.0428	0.0143	3.0023	0.0034
β_5	0.0757	0.0263	2.8756	0.0050

3.2. Ecuación Ajustada

Con base en la tabla de parámetros estimados se obtiene la ecuación de regresión ajustada:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2} + \cdots + \widehat{\beta}_5 X_{i5}, \quad i = 1, 2, \ldots, 100$$

$$\widehat{Y}_i = -1.7723 + 0.0041 X_{i1} + 0.0029 X_{i2} - 0.0120 X_{i3} + 0.0428 X_{i4} + 0.0757 X_{i5}, \quad i = 1, 2, \ldots, 100$$

3.3. Tabla ANOVA

Donde F-value $= F_0 = \frac{\text{MSR}}{MSE} \sim F_{5,94}$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FO(GRE.Score, TOEFL.Score, SOP, LOR, CGPA)	5	2.3674478	0.4734896	65.99381	0
Residuals	94	0.6744272	0.0071748	NA	NA

3.4. Prueba de significancia del Modelo

$$\begin{cases} H_0 : \beta_1 = \cdots = \beta_5 = 0 \\ H_1 : \text{Al menos un } \beta_j \neq 0 \end{cases}$$

Analizando el **p-valor = 2.2e-16 = 0** de la tabla ANOVA y con una confianza del **95 %** hay evidencia suficiente para rechazar la **hipótesis nula**. Esto quiere decir que el modelo es globalmente significativo y por lo tanto al menos una de las pruebas de aptitud ayuda a explicar la variabilidad de ser admitido a un curso de postgrado.

3.5. Coeficiente de determinación R^2: proporción de la variabilidad total de la respuesta explicada por el modelo y opiniones al respecto

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

$$R^2 = \frac{2.3674478}{2.3674478 + 0.6744272} = 0.7782857$$

El **77.83 %** de la variabilidad de la posibilidad de ser admitido es explicada por la relación con las variables GRE.Score, TOEFL.Score, SOP, LOR y CGPA.

4. Coeficientes de regresión estandarizados

4.1. Tabla de coeficientes estandarizados

	Estimación	Limites.2.5..	Limites.97.5..	Vif	Coef.Std
(Intercept)	-1.7722651	-2.3693009	-1.1752294	0.000000	0.0000000
GRE.Score	0.0041091	0.0007654	0.0074528	5.691210	0.0495542
TOEFL.Score	0.0029116	-0.0032277	0.0090508	5.858052	0.0194023
SOP	0.0120402	-0.0116343	0.0357148	1.928844	0.0119389
LOR	0.0428307	0.0145058	0.0711556	2.519579	0.0405706
CGPA	0.0757081	0.0234330	0.1279833	4.615227	0.0525903

Gracias a esta tabla, se puede deducir con una diferencia en el valor muy pequeña que, las variables que más aportan según el valor de sus coeficientes estandarizados son **CGPA** y **GRE.Score**

5. Significancia individual de los parámetros del modelo

5.1. Tabla de la significancia individual de los parámetros

Cuadro 3: Resumen de los coeficientes

	Estimación	Error estándar	T_0	Valor P
β_0	-1.7723	0.3007	-5.8939	0.0000
β_1	0.0041	0.0017	2.4400	0.0166
β_2	0.0029	0.0031	0.9417	0.3488
β_3	0.0120	0.0119	1.0098	0.3152
β_4	0.0428	0.0143	3.0023	0.0034
β_5	0.0757	0.0263	2.8756	0.0050

De la tabla anterior, se puede observar que a nivel marginal, las variables **GRE Score**(β_1), **LOR**(β_4) y **CGPA**(β_5) son significativas en la respuesta, con un nivel de significancia de $\alpha = 0.05$.

Con el estadístico de prueba $T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{94}$

Dicha afirmaciones serán contrastadas con las siguientes pruebas de hipótesis y analizando el **p-valor**.

5.2. Pruebas de hipótesis

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Analizando el **valor-p = 0.0166** del parámetro β_1 y con una confianza del **95 %** hay evidencia para rechazar la hipótesis nula. Luego el parámetro β_1 es significativo.Esto quiere decir que **GRE Score** ayuda a explicar la posibilidad de ser admitido a una carrera de postgrado dado que las demas pruebas de aptitud se encuentran en el modelo.

$$\begin{cases} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{cases}$$

Analizando el **valor-p = 0.3488** del parámetro β_2 y con una confianza del **95 %** no hay evidencia para rechazar la hipótesis nula. Luego el parámetro β_2 no es significativo.Esto quiere decir que **TOEFL.Score** no ayuda a explicar la posibilidad de ser admitido a una carrera de postgrado dado que las demas pruebas de aptitud se encuentran en el modelo.

$$\begin{cases} H_0 : \beta_3 = 0 \\ H_1 : \beta_3 \neq 0 \end{cases}$$

Analizando el **valor-p = 0.3152** del parámetro β_3 y con una confianza del **95 %** no hay evidencia para rechazar la hipótesis nula. Luego el parámetro β_3 no es significativo. Esto quiere decir que **SOP** no ayuda a explicar la posibilidad de ser admitido a una carrera de postgrado dado que las demas pruebas de aptitud se encuentran en el modelo.

$$\begin{cases} H_0 : \beta_4 = 0 \\ H_1 : \beta_4 \neq 0 \end{cases}$$

Analizando el **valor-p = 0.0034** del parámetro β_4 y con una confianza del **95 %** no hay evidencia para rechazar la hipótesis nula. Luego el parámetro β_4 es significativo. Esto quiere decir que **LOR** ayuda a explicar la posibilidad de ser admitido a una carrera de postgrado dado que las demas pruebas de aptitud se encuentran en el modelo.

$$\begin{cases} H_0 : \beta_5 = 0 \\ H_1 : \beta_5 \neq 0 \end{cases}$$

Analizando el **valor-p = 0.0050** del parámetro β_5 y con una confianza del **95 %** hay evidencia para rechazar la hipótesis nula. Luego el parámetro β_5 es significativo.Esto quiere decir que **CGPA** ayuda a explicar la posibilidad de ser admitido a una carrera de postgrado dado que las demas pruebas de aptitud se encuentran en el modelo.

6. Sumas de cuadrados extras

Teniendo en cuenta los resultados anteriores, realice una prueba con sumas de cuadrados extras con test lineal general; especifique claramente el modelo reducido y completo, estadístico de la prueba, su distribución, cálculo de valor P, decisión y conclusión a la luz de los datos. Justifique la hipótesis que desea probar en este numeral.

$$SSR(X_1, X_4, X_5 | X_2, X_3) = SSR(X_1, X_2, X_3, X_4, X_5) - SSR(X_2, X_3)$$

6.1. Prueba de hipótesis

$$\begin{cases} H_0 : \beta_2 = 0, \beta_3 = 0 \\ H_1 : \beta_2 \neq 0 \vee \beta_3 \neq 0 \end{cases}$$

6.2. Modelo completo y reducido

$$MF : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_5 X_{i5} + \varepsilon_i, \quad i = 1, 2, \ldots, 100$$

$$MR : Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i, \quad i = 1, 2, \ldots, 100$$

6.3. Estadístico de prueba

$$F_0 = \frac{[SSR(X_1, X_4, X_5 | X_2, X_3)]/2}{MSE}$$
$$= \frac{[SSE(X_2, X_3) - SSE(X_1, X_2, X_3, X_4, X_5)]/2}{\frac{SSE(MF)}{n-k-1}} = \frac{[0.6915854 - 0.6744272]/2}{0.6744272/94}$$
$$= 1.1957338$$

6.4. Tabla del Test lineal general

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
96	0.6915854	NA	NA	NA	NA
94	0.6744272	2	0.0171582	1.195733	0.3070405

Con un nivel de significancia de $\alpha = 0.05$ el valor crítico es $f_{0.05,2,94} = 3.093266$.

Como $F_0 = 1.1957338 < f_{0.05,2,94} = 3.093266$, No hay evidencia para rechazar la **hipótesis nula**. por lo tanto X_2(**TOEFL.Score**) y X_3(**SOP**) no ayudan a explicar la posibilidad de ser admitido a una carrera de postgrado dado que en el modelo estan presentes **GRE.Score**, **LOR** y **CGPA**.

Tomamos esta prueba de hipótesis con la finalidad de mirar si dicho modelo era significativo globalmente. Con las pruebas de significancia individual de los parámetros y mirando la magnitud de los parámetros estandarizados sabíamos que X_2(**TOEFL.Score**) y X_3(**SOP**) no eran significativos y no ayudaban a explicar la posibilidad de ser admitidos a una carrera de postgrado. Dicha hipótesis nos permitió descartar este modelo y de esta manera poder continuar en búsqueda del modelo más adecuado.

7. Sumas de cuadrados tipo I y tipo II

Calcule las sumas de cuadrados tipo I (secuenciales) y tipo II (parciales) ¿Cuál de las variables tienen menor valor en tales sumas? ¿Qué puede significar ello?

7.1. Sumas de cuadrados secuenciales

7.2. Tabla anova

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
GRE.Score	1	1.9853655	1.9853655	276.715340	0.000000
TOEFL.Score	1	0.0559996	0.0559996	7.805092	0.006314
SOP	1	0.1181923	0.1181923	16.473354	0.000102
LOR	1	0.1485633	0.1485633	20.706392	0.000016
CGPA	1	0.0593270	0.0593270	8.268855	0.004989
Residuals	94	0.6744272	0.0071748	NA	NA

SS1

- X_1 $SSR(X_1) = 1.98537$ Dicho modelo tiene mayor **aumento** del **SSR**
- $X_2|X_1$ $SSR(X_2|X_1) = 0.05600$
- $X_3|X_1, X_2$ $SSR(X_3|X_1, X_2) = 0.1181923$
- $X_4|X_1, X_2, X_3$ $(X_4|X_1, X_2, X_3) = 0.14856$
- $X_5|X_1, X_2, X_3, X_4$ $SSR(X_5|X_1, X_2, X_3, X_4) = 0.05933$ Dicho modelo tiene la mayor **disminución** del **SSR**.

7.3. Sumas de cuadrados parciales

7.4. Tabla Anova

	Sum Sq	Df	F value	Pr(>F)
GRE.Score	0.0427161	1	5.9536667	0.0165615
TOEFL.Score	0.0063619	1	0.8867053	0.3487854
SOP	0.0073158	1	1.0196631	0.3151912
LOR	0.0646740	1	9.0141001	0.0034318
CGPA	0.0593270	1	8.2688553	0.0049890
Residuals	0.6744272	94	NA	NA

$SS2$

- $X_1|X_2, X_3, X_4, X_5$ $SSR(X_1|X_2, X_3, X_4, X_5)$
- $X_2|X_1, X_3, X_4, X_5$ $SSR(X_2|X_1, X_3, X_4, X_5)$
- $X_3|X_1, X_2, X_4, X_5$ $SSR(X_3|X_1, X_2, X_4, X_5)$
- $X_4|X_1, X_2, X_3, X_5$ $SSR(X_4|X_1, X_2, X_3, X_5)$
- $X_5|X_1, X_2, X_3, X_4$ $SSR(X_5|X_1, X_2, X_3, X_4)$

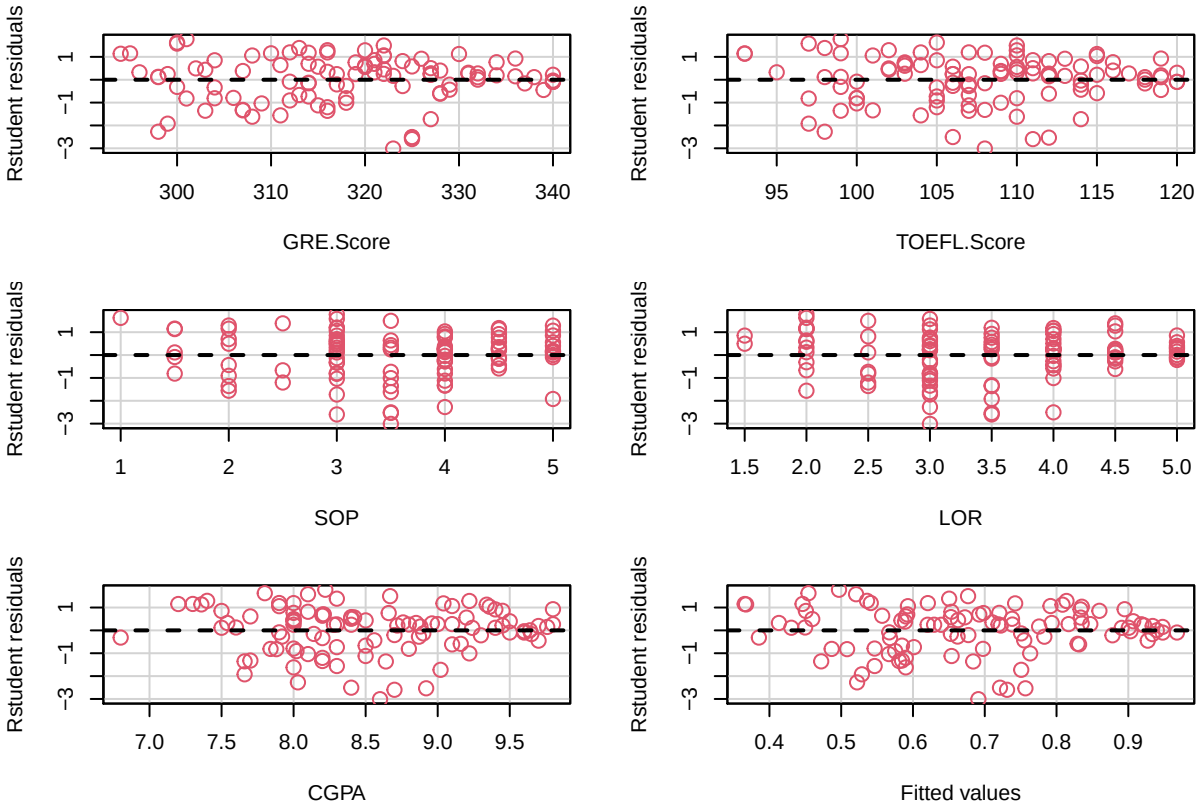
7.5. Prueba de hipótesis

$$\begin{cases} H_0 : \beta_j = 0 conj = 1, \ldots, 5 \\ H_1 : \beta_j \neq 0 \end{cases}$$

- Analizando el **p-valor** se puede concluir que el efecto parcial de incluir X_1(GRE.Score) dado que en modelo se encuentra X_2, X_3, X_4, X_5 es significativa de esta manera aumentando el SSR = 0.0427161.
- Analizando el **p-valor** se puede concluir que el efecto parcial de incluir X_4(LOR) dado que en modelo se encuentra X_1, X_2, X_3, X_5 es significativa de esta manera aumentando el SSR = 0.0646740.
- Analizando el **p-valor** se puede concluir que el efecto parcial de incluir X_5(CGPA) dado que en modelo se encuentra X_1, X_2, X_3, X_4 es significativa de esta manera aumentando el SSR = 0.0593270.
- Se observa que X_4(LOR) tiene el efecto parcial mas grande con un **SSR = 0.0646740**.
- Analizando el **p-valor** se puede concluir que el efecto parcial de incluir X_2(TOEFL.Score) dado que en modelo se encuentra X_1, X_3, X_4, X_5 no es significativa de esta manera disminuye el SSR = 0.0063619.
- Analizando el **p-valor** se puede concluir que el efecto parcial de incluir X_3(SOP) dado que en modelo se encuentra X_1, X_2, X_4, X_5 es significativa de esta manera aumentando el SSR = 0.0073158.

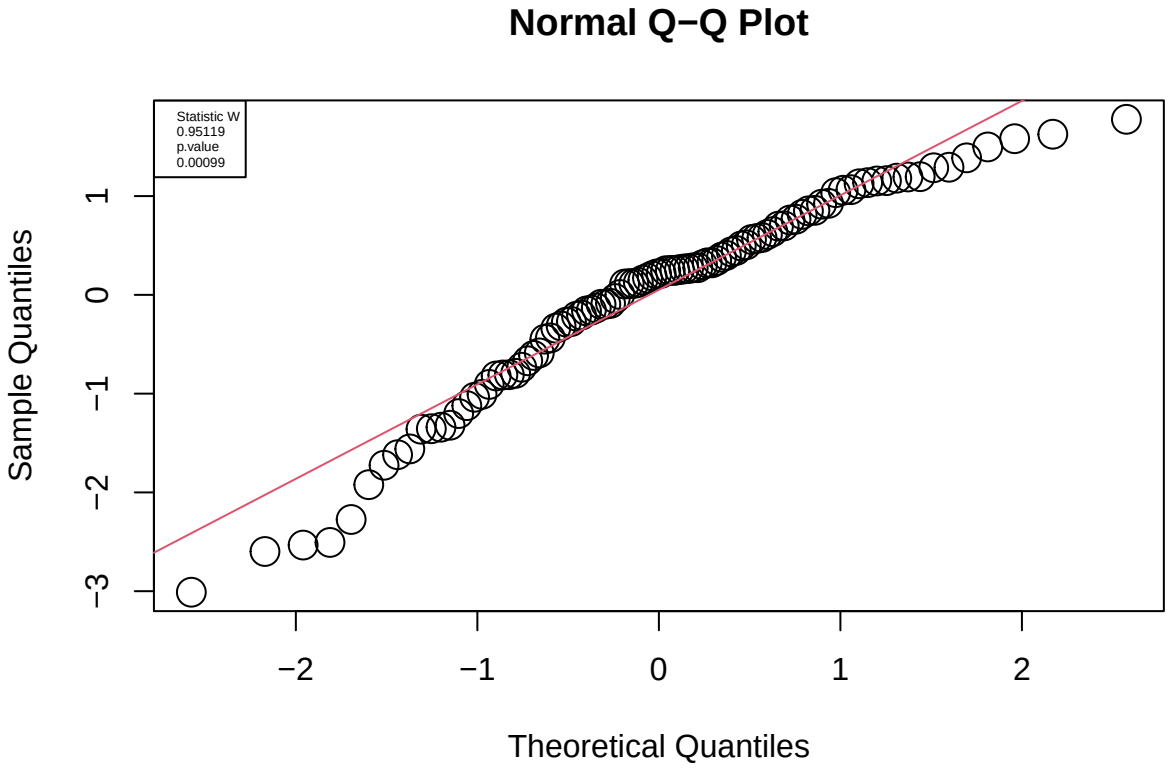
8. Residuales estudentizados vs. Valores ajustados

8.1. Gráfico de los residuales estudentizados vs. Valores ajustados



- En los gráficos de las variables **GRE.Score**, **TOEFL.Score**, **SOP**, **LOR**, **CGPA** y además de la gráfica de los valores ajustados no se observa ningún tipo de patrón, por lo tanto se cumple el supuesto de varianza constante. Se aprecian algunos valores **atípicos**, información que se verificará más adelante.

9. Gráfico q-norm residuales estudentizados



9.1. Pruebas de hipótesis

$$\begin{cases} H_0 : \varepsilon \sim Normal \\ H_1 : \varepsilon \not\sim Normal \end{cases}$$

Aunque muchos residuales se concentren cerca de la recta ajustada, se puede observar cantidad considerable que se aleja de esta generando una asimetría hacia la derecha, además al realizar la prueba de **Shapiro-Wilk** tenemos un **p valor de 0.00099** por lo que podemos rechazar la hipótesis nula, concluyendo de esta manera que hay evidencia para decir que no tienen un comportamiento normal.

	dfb.1_	dfb.GRE.	dfb.TOEF	dfb.SOP	dfb.LOR	dfb.CGPA	dffit	cov.r	cook.d	hat
10	0.3312129	-0.3365468	0.2641299	-0.0776585	0.3234579	-0.1008349	-0.5199544	0.6285422	0.0414993	0.0289676
11	0.2748995	-0.4352307	0.3325789	-0.0384741	-0.1568373	0.2077023	-0.5823847	0.7592033	0.0535208	0.0512194
32	-0.0514907	0.0853945	-0.0761287	0.0342908	0.0053108	-0.0281974	0.1009979	1.2189200	0.0017171	0.1291027
37	0.0883736	-0.0895248	0.0584789	-0.0014274	0.0466126	-0.0006190	0.0990952	1.2318929	0.0016531	0.1378627
53	-0.1601062	0.1422062	0.0820887	0.0995052	-0.0341875	-0.2546427	0.3279626	1.2150773	0.0180058	0.1549517
65	0.1757763	-0.0875596	-0.0480967	0.2399782	0.0003671	-0.0104244	-0.4079735	0.7175194	0.0261410	0.0240603
66	0.1292378	-0.0190079	-0.0466489	0.1141425	0.1003505	-0.1116136	-0.3533659	0.7283787	0.0196772	0.0190858
92	-0.3421799	0.0334767	0.1701959	-0.5818706	-0.1450394	0.2592579	-0.7764099	0.9818877	0.0976705	0.1403247

10. Diagnostico sobre la presencia de observaciones atípicas, de balanceo y/o influenciales y conclusiones

10.1. Valores ajustados VS Residuales Estudentizados

	Ajustados	Errores
10	0.6911442	-3.010410
65	0.7310632	-2.598320
66	0.7566507	-2.533292
11	0.7212283	-2.506545
93	0.5221676	-2.274321
92	0.5288087	-1.921725

10.2. Valores influenciables

	dffit	cov.r	cook.d	hat
10	FALSO	VERDADERO	FALSO	FALSO
11	FALSO	VERDADERO	FALSO	FALSO
32	FALSO	VERDADERO	FALSO	FALSO
37	FALSO	VERDADERO	FALSO	FALSO
53	FALSO	VERDADERO	FALSO	FALSO
65	FALSO	VERDADERO	FALSO	FALSO
66	FALSO	VERDADERO	FALSO	FALSO
92	VERDADERO	FALSO	FALSO	FALSO

- La observación **10** es un **outlier** ya que el valor absoluto de su residual estudentizado es **3.01**, mayor a **3**.
- De acuerdo al **COVRATIO** y el **DFFITS**, las observaciones **10**,**11**, **32**, **37**, **53**, **65**, **66** y **92** son **influenciables**.
- Para evaluar las observaciones de **balanceo** miramos las que superan la cota de $\frac{2(k+1)}{n} = \frac{2(5+1)}{100} = 0.12$. Las observaciones **32**, **37**, **38**, **53** y **92** superan dicha cota y por lo tanto, son de balanceo.

- En conclusión, las observaciones **10**, **11**, **53**, **65** y **66** son **influenciables**; la observación **38** es de **balanceo** y las observaciones **32**, **37**, **53** y **92** son **influenciables** y de **balanceo**.

11. Ejercicio11

Cuadro 4: Tabla de parámetros ajustados resultantes

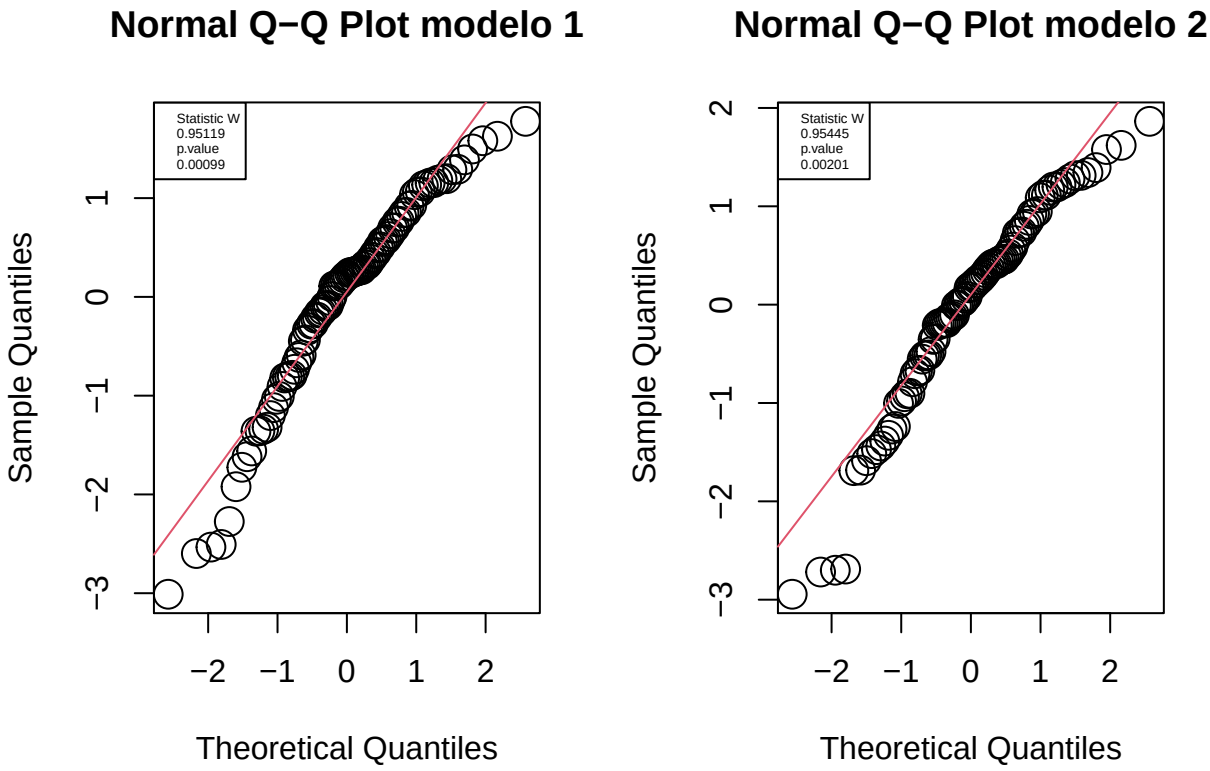
	Estimación	Error estándar	T_0	Valor P
β_0	-1.8565	0.2897	-6.4072	0.0000
β_1	0.0054	0.0016	3.3092	0.0013
β_2	0.0001	0.0030	0.0389	0.9690
β_3	0.0267	0.0122	2.2006	0.0303
β_4	0.0401	0.0134	2.9901	0.0036
β_5	0.0682	0.0248	2.7479	0.0072

Para validar si los cambios fueron notorios en la estimación de los parámetros y los errores estándar, debido a que la diferencia entre ambos modelos consiste en la exclusión en el segundo modelo de las medidas que se supone que fueron errores de digitación, se presenta una tabla de diferencias relativas respecto al primer modelo. Se evidencia que para los parámetros β_1, β_2 y β_3 los cambios fueron muy notorios; β_2 se estimó en un valor 96 % menor al inicialmente calculado, β_3 se estimó en un valor 122 % mayor al calculado en el primer modelo, y β_1 en el segundo modelo excedió al valor del primer modelo en 31.4 %. Los errores estándar no cambiaron tanto, el mayor cambio relativo fue de cerca del 6 % para β_4, cuyo valor en el segundo modelo fue 9×10^{-4} unidades menos que el calculado en el primer modelo.

Cuadro 5: Diferencias relativas respecto al primer modelo

	Estimación	Error estándar
β_0	0.0475	-0.0364
β_1	0.3138	-0.0313
β_2	-0.9599	-0.0291
β_3	1.2214	0.0193
β_4	-0.0648	-0.0610
β_5	-0.0996	-0.0578

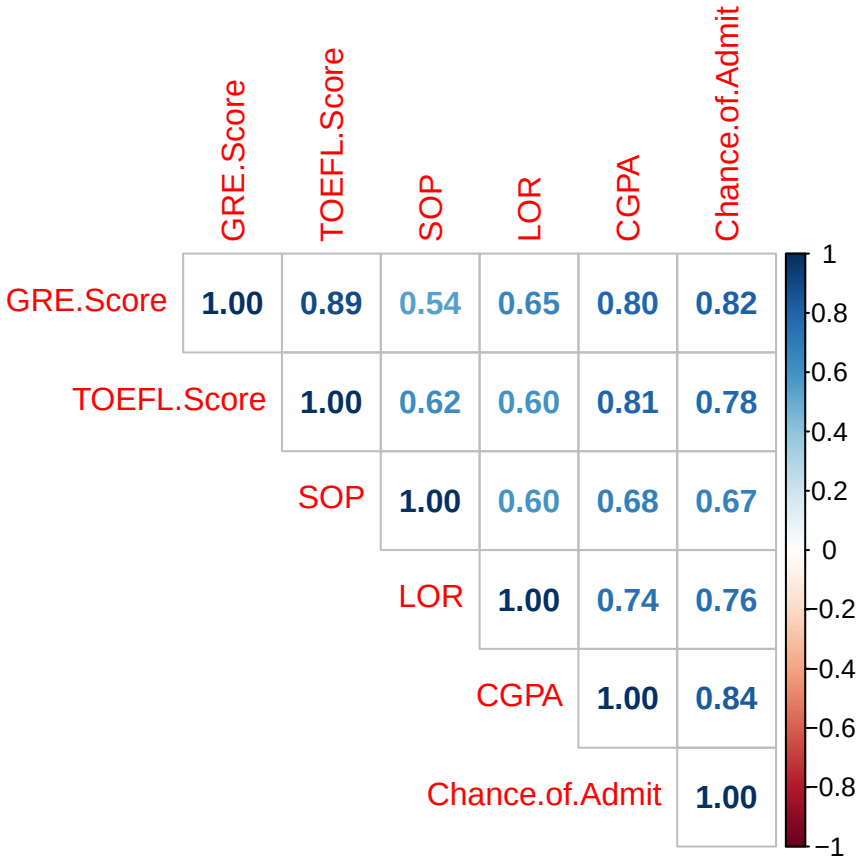
Con un $\alpha = 0.05$ hubo diferencias para el parámetro β_3 que no era significativo en el primer modelo y en el segundo ya es significativo (con valore-p de 0.0303 en el segundo modelo y 0.3152 en el primero).



Se logra observar gráficamente que la normalidad mejoró aunque no es lo suficientemente significativa para que dicho modelo cumpla los supuestos de normalidad.

12. Ejercicio 12

12.1. Matriz de correlación de las variables predictoras



Entre las pruebas **GRE - TOEFL**, **GRE - CGPA**, **TOEFL - CGPA** y finalmente **LOR - CGPA** se observan **correlaciones fuertes**, esto puede indicar problemas de **multicolinealidad**.

Se observa que entre **GRE - SOP**, **GRE - LOR**, **TOEFL - SOP**, **TOEFL - LOR** se tienen **correlaciones moderadas**.

12.2. VIF's

GRE.Score	TOEFL.Score	SOP	LOR	CGPA
5.859487	6.190994	2.108352	2.497704	4.599604

- En los factores de **inflación** de **varianza** no se concluye que existan problemas de **multicolineadlidad**, pues nos indica que ninguna estimación **supera** el valor de **10**.

12.3. Proporciones de varianza

```
## Condition
## Index    Variance Decomposition Proportions
##              intercept GRE.Score TOEFL.Score SOP   LOR   CGPA
## 1     1.000 0.000     0.000     0.000       0.001 0.001 0.000
## 2    10.083 0.002     0.000     0.001       0.173 0.161 0.001
## 3    14.645 0.000     0.000     0.000       0.597 0.490 0.000
## 4    60.699 0.197     0.003     0.036       0.179 0.235 0.399
## 5    83.509 0.092     0.009     0.368       0.001 0.025 0.593
## 6   189.536 0.710     0.988     0.596       0.049 0.088 0.007
```

- La raíz del número condición es de **189**. Lo cual nos indica que se tienen problemas graves de **multicolinealidad**.
- Examinando la descomposión de varianza se visualiza que existe problemas de **multicolinealidad** entre las pruebas **GRE** - **TOEFL** y las pruebas **SOP** - **LOR**

13. Ejericio13

# de covariables	modelo	R2_adj
1	(1) y~CGPA	0.75
2	(6) y~GRE.Score+CGPA	0.81
3	(16) y~GRE.Score+LOR+CGPA	0.83
4	(26) y~GRE.Score+SOP+LOR+CGPA	0.83
5	(31) y~GRE.Score+TOEFL.Score+SOP+LOR+CGPA	0.83

- De acuerdo al **principio de parsimonia** un buen modelo bajo el criterio del **R2_adj** es el modelo **(6) y~GRE.Score+CGPA**

13.1. Selección según el R^2_{adj}

Cuadro 7: Resumen de los coeficientes

	Estimación	Error estándar	T_0	Valor P
Intercepto	-2.290156730	0.264865653	-8.646484	1.000000e-13
GRE.Score	0.005941449	0.001205231	4.929719	3.533574e-06
CGPA	0.127527762	0.020684755	6.165302	1.747250e-08

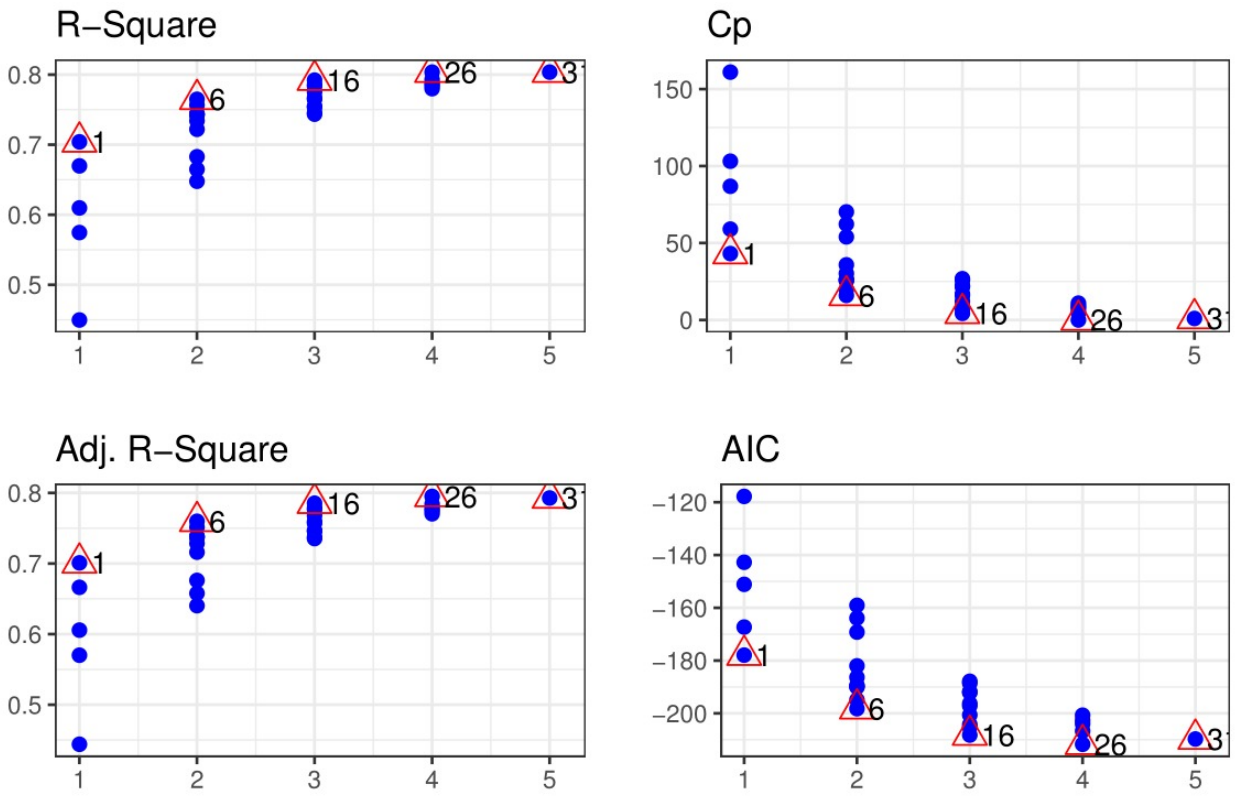


Figura 1: Metodo de seleccion R Ajustado y Cp.

Ecuacion Ajustada

$$\hat{Y}i = -2.290157 + 0.005941X_i1 + 0.127528X_i5$$

Tabla anova

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FO(GRE.Score, CGPA)	2	2.2040778	1.1020389	152.8941	0
Residuals	94	0.6775387	0.0072079	NA	NA

Como se observan en la tabla anova se tiene un **p-valor** $< \alpha = 0.05$ por lo tanto hay evidencia para rechazar H_0, Esto quiere decir que al menos una de las pruebas(**GRE.Score, CGPA**) ayuda a explicar la variabilidad de ser admitido a un curso de postgrado.

13.2. Selección según el estadístico C_p

# de covariables	**modelo**	**abs(Cp - p)**
1	(1) y~CGPA	45.00-1 =44
2	(6) y~GRE.Score+CGPA	15.39-2 =13.39
3	(16) y~GRE.Score+LOR+CGPA	4.17-3 =1.17
4	(26) y~GRE.Score+SOP+LOR+CGPA	4.00-4 =0.00
5	(31) y~GRE.Score+TOEFL.Score+SOP+LOR+CGPA	6-5 =1

- De acuerdo al **Cp** el mejor modelo es el **(26) y~GRE.Score+SOP+LOR+CGPA**

Cuadro 9: Resumen de los coeficientes

	Estimación	Error estándar	T_0	Valor P
Intercepto	-1.86080061	0.26614162	-6.991769	4.235000e-10
GRE.Score	0.00544423	0.00112664	4.832273	5.374931e-06
LOR	0.03993901	0.01299522	3.073362	2.785099e-03
CGPA	0.06843760	0.02367305	2.890951	4.791586e-03
SOP	0.02687834	0.01160322	2.316455	2.275417e-02

Ecuacion ajustada

$$\hat{Y}i = -1.860801 + 0.005444X_{i1} + 0.026878X_{i3} + 0.039939X_{i4} + 0.068438X_i5$$

Tabla anova

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FO(GRE.Score, LOR, CGPA, SOP)	4	2.3158387	0.5789597	94.14348	0
Residuals	92	0.5657778	0.0061498	NA	NA

Bajo el metodo de seleccion **Cp** el modelo escogido vs modelo completo **MF** no hay un cambio significante en la estimacion de los parametros.

13.3. Stepwise

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	2.316	4	0.579	94.143	0.0000
Residual	0.566	92	0.006		
Total	2.882	96			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-1.861	0.266		-6.992	0.000	-2.389	-1.332
CGPA	0.068	0.024	0.275	2.891	0.005	0.021	0.115
GRE.Score	0.005	0.001	0.375	4.832	0.000	0.003	0.008
LOR	0.040	0.013	0.219	3.073	0.003	0.014	0.066
SOP	0.027	0.012	0.149	2.316	0.023	0.004	0.050

13.4. Forward

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	2.316	4	0.579	94.143	0.0000
Residual	0.566	92	0.006		
Total	2.882	96			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-1.861	0.266		-6.992	0.000	-2.389	-1.332
CGPA	0.068	0.024	0.275	2.891	0.005	0.021	0.115
GRE.Score	0.005	0.001	0.375	4.832	0.000	0.003	0.008
LOR	0.040	0.013	0.219	3.073	0.003	0.014	0.066
SOP	0.027	0.012	0.149	2.316	0.023	0.004	0.050

13.5. Backward

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	2.316	4	0.579	94.143	0.0000
Residual	0.566	92	0.006		
Total	2.882	96			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-1.861	0.266		-6.992	0.000	-2.389	-1.332
GRE.Score	0.005	0.001	0.375	4.832	0.000	0.003	0.008
SOP	0.027	0.012	0.149	2.316	0.023	0.004	0.050
LOR	0.040	0.013	0.219	3.073	0.003	0.014	0.066
CGPA	0.068	0.024	0.275	2.891	0.005	0.021	0.115

De los tres métodos de selección automática se observa que el análisis de varianza se muestra que al menos una de las covariables aporta a el modelo. Finalmente también se concluye que en los tres métodos de selección las estimaciones no cambian significativamente respecto al modelo completo.

14. Selección del modelo

Modelos	**MSE**
(6)y~GRE.Score+CGPA	0.0072079
(26)y~GRE.Score+SOP+LOR+CGPA	0.0061498
(MF)y~GRE.Score+TOEFL.Score+SOP+LOR+CGPA	0.0071748

Como se observa en la tabla anterior de los modelos propuestos por medio de los metodos de seleccion el modelo **26** cuenta con un **Error Cuadratico Medio** menor a los recomendados, por lo cual se sugiere ese modelo **y~GRE.Score+SOP+LOR+CGPA**.