Trabajo regresión lineal múltiple

Estudiantes

Juan Daniel Bula Isaza
José Daniel Bustamante Arango
Marlon Calle Areiza
Santiago Carvajal Torres

Docente

Raul Alberto Perez Agamez

Asignatura

Estadística II

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
Noviembre de 2021

Índice

Índice de figuras

Índice de cuadros

1. Ejercicio 1

1.1. Análisis de las variables

Es de nuestro interés ajustar un modelo de regresión lineal múltiple acorde a las 72 observaciones de la base de datos, cuyas filas están compuestas por las variables: y: Riesgo de infección, x_1: Duración de la estadía, x_2: Rutina de cultivos, x_3: Número de camas, x_4: Censo promedio diario y x_5: Número de enfermeras. Para ello comenzaremos con un análisis a la matriz de correlaciones entre ellas:

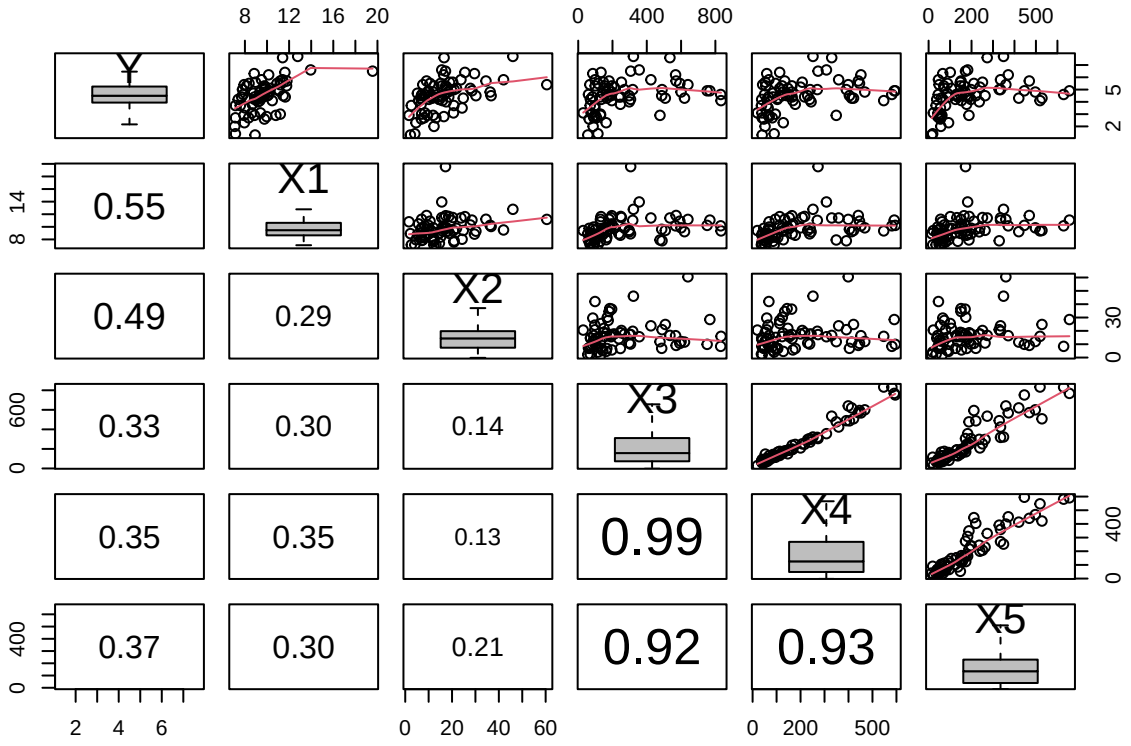


Figura 1: Matriz de correlaciones

Se puede notar una alta dependencia lineal entre X_3 y X_4, X_3 y X_5, y, X_4 y X_5. Además, notar que las correlaciones entre la variable respuesta Y y las variables X_1 X_2 son las más altas en relación a la primera, con respectivos valores de 0.55 y 0.49, suponiendo una relación lineal débil entre las dos variables en los dos casos.

1.2. Ajuste del modelo

Ahora, con las observaciones de la base de datos, procederemos a hacer el ajuste del siguiente modelo de regresión lineal múltiple:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 72$$

Así, el ajuste, redondeando los coeficientes a cuatro cifras decimales, nos resultará en la ecuación:

$$\hat{y}_i = 0.9586 + 0.2682 x_{1i} + 0.0409 x_{2i} + -0.0007 x_{3i} + 0.0009 x_{4i} + 0.001 x_{5i};\ 1 \leq i \leq 72$$

1.3. Análisis de varianza

Se desea verificar la significancia de la regresión usando la tabla ANOVA, mediante la cual compararemos el siguiente juego de hipotesis:

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

Para la cual obtendremos lo siguiente:

Cuadro 1: Tabla ANOVA para el modelo

	Suma de cuadrados	gl	Cuadrado Medio	F_0	Valor P
Regresión	52.3209	5	10.464176	10.7672	1.35475e-07
Error	64.1423	66	0.971853		

Donde $F_0 = \frac{MSR}{MSE} \sim F_{5,66}$ bajo H_0

Podemos concluir así que con un nivel de significancia del 5 % al menos uno de los coeficientes es significativo, esto del rechazo de H_0 por el Valor P.

1.4. Análisis de varianza

Ahora procederemos hacer un análisis marginal de cada uno de los coeficientes usando la siguiente prueba para $j = 1, \cdots, 5$:

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

con $T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{66}$ bajo H_0

Para lo que usaremos la siguiente tabla:

Cuadro 2: Resumen de los coeficientes

	Estimación	Error estándar	T_0	Valor P
β_0	0.9586	0.6599	1.4527	0.1511
β_1	0.2682	0.0742	3.6135	0.0006
β_2	0.0409	0.0120	3.4091	0.0011
β_3	-0.0007	0.0035	-0.2036	0.8393
β_4	0.0009	0.0050	0.1850	0.8538
β_5	0.0015	0.0021	0.7108	0.4797

Se puede observar que con un nivel de significancia de $\alpha = 0.05$, a nivel marginal las únicas variables que tienen un efecto significativo en la respuesta son x_1 y x_2 que representan la duración de la estadía y la rutina de cultivos respectivamente.

Notar que: $\hat{\beta}_1 = 0.2682$ indica que por un aumento unitario en la duración de la estadía x_1, la media de la variable respuesta aumentará en 0.2682 unidades; de la misma manera, $\hat{\beta}_2 = 0.0409$ indica que por cada aumento en una unidad en la rutina de cultivos, la media de la variable respuesta aumentará en 0.0409 unidades; estos dos casos siempre y cuando las otras variables se mantengan constantes.

1.5. Cálculo y análisis del coeficiente de determinación múltiple

Tenemos que el coeficiente de determinación multiple, denotado por R^2 se define como:

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

con SSR: Suma de cuadrados de la regresión , SSE: Suma de cuadrados de los residuales, y $SST = SSR + SSE$. Así tenemos, con los datos calculados anterioremente en la tabla ANOVA, que:

$$R^2 = \frac{52.3209}{52.3209 + 64.1423} = 0.4492483$$

Tal R^2 nos dice que un 44.92 % de la variabilidad total en el Riesgo de infección (Variable respuesta) es explicado por el modelo RLM propuesto.

2. Ejercicio 2

De la tabla resumen de coeficientes que se mostró anteriormente, obtenemos que las 3 variables con el mayor valor p (y mayor a 0.05) son β_3, β_4 y β_5, que equivalen a el número de camas, censo promedio diario y número de enfermeras respectivamente.

Para probar la significancia simultánea de las anteriores variables se usará el siguiente juego de hipotesis:

$$\begin{cases} H_0 : \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : Algún\ \beta_j \neq 0,\ j = 3, 4, 5, \end{cases}$$

De la tabla de todas las regresiones posibles se extraen los siguientes modelos de interés:

	Numero de variables	R^2	R^2 ajustado	SSE	CP	Variables del modelo
Reducido	2	0.421	0.404	67.452	3.405	X1 X2
Completo	5	0.449	0.408	64.142	6.000	X1 X2 X3 X4 X5

Para medir la importancia de β_3, β_4 y β_5, se usará la suma de cuadrados extra, de la anterior tabla, con el fin de obtener el siguiente estadístico de prueba para la prueba de hipótesis:

$$F_0 = \frac{MSextra}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} = \frac{MSR(\beta_3, \beta_4, \beta_5 | \beta_0, \beta_1, \beta_2)}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} = \frac{SSR(\beta_3, \beta_4, \beta_5 | \beta_0, \beta_1, \beta_2)/3}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)}$$

$$= \frac{[SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)]/3}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} = \frac{[SSE(Reducido) - SSE(Completo)]/3}{MSE(Completo)}$$

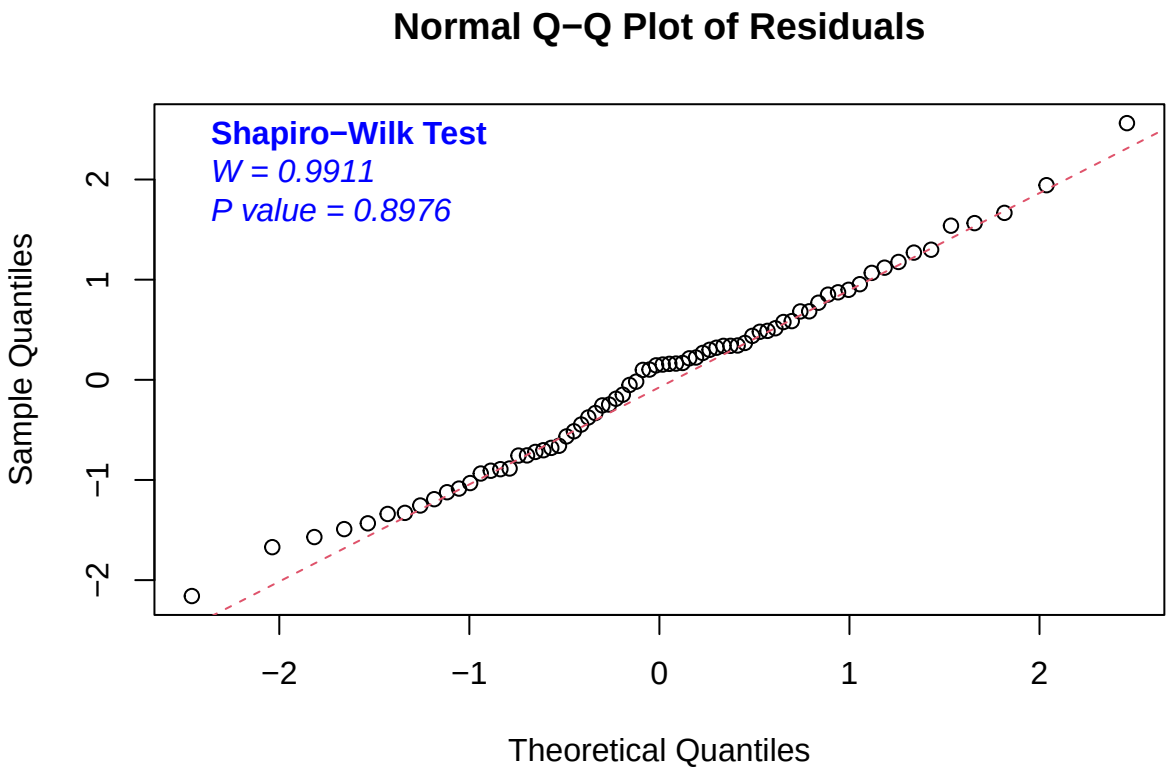
$$= \frac{[67.452 - 64.142]/3}{0.971853} = 1.1353$$

El criterio de rechazo de la hipótesis nula es $F_0 > f_{0.05,3,66}$. Como $F_0 = 1.1353$ y $f_{0.05,3,66} = 2.7437$ luego $F_0 \not> f_{0.05,3,66}$. Esto quiere decir que no se puede rechazar la hipótesis nula, es decir, la prueba concluye que es posible descartar el número de camas, el censo promedio diario y el número de enfermeras.

3. Ejercicio 3

4. Ejercicio 4

4.1. Validación de los supuestos

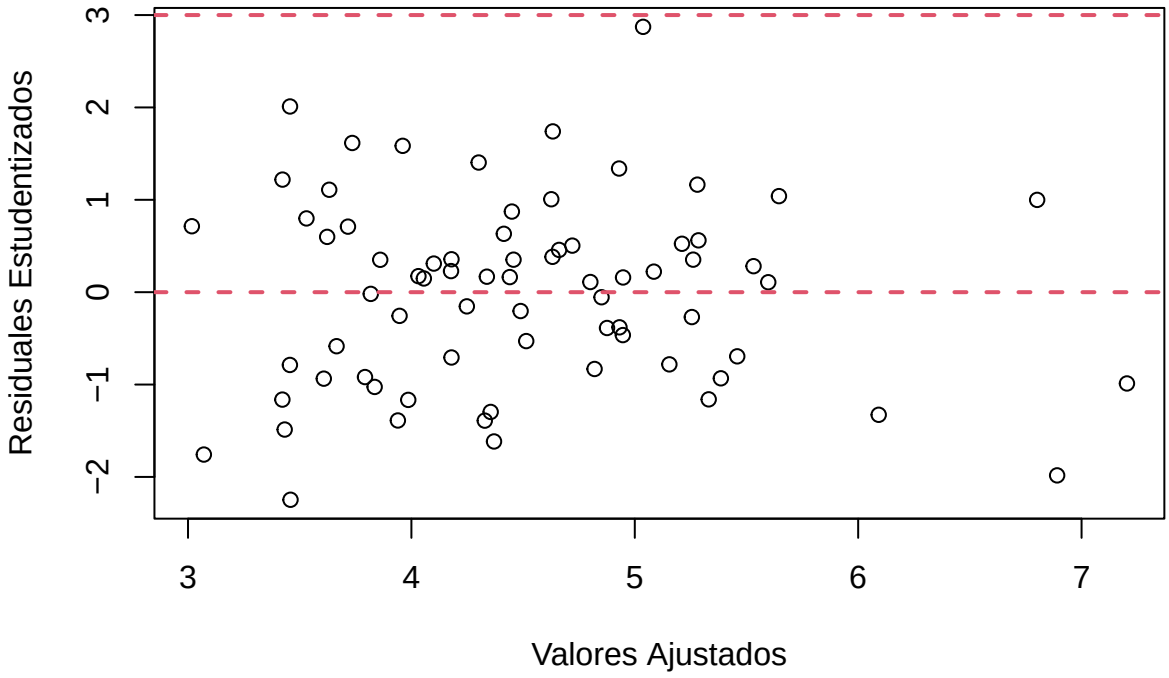


Como podemos ver, el gráfico de normalidad se comporta de una buena manera ya que los puntos están bastante cercanos a la recta de normalidad. Además si nos fijamos en la prueba de hipótesis que sería:

$$H_0 : r \sim N(\mu, \sigma^2)$$
$$H_1 : r \neq N(\mu, \sigma^2)$$

Podemos notar que el p-value es 0.8976 por lo cual si tomamos un $\alpha = 0.05$ es claro que $0.8976 > \alpha$ por lo cual no negamos la hipótesis nula y concluimos que los residuales se distribuyen de manera normal.

4.2. Gráfico de los residuales estudentizados vs valores ajustados



Podemos ver que la varianza de los errores se ve constante dado que no se ve ninguna forma en particular, además se ve que ningún punto sobre pasa el límite $|r_i| > 3$ por lo cual no hay datos atípicos.

4.3. Puntos de balanceo

Para hallar los puntos de balanceo realizamos el cálculo:

$$h_{ii} > \frac{2 * p}{n} = \frac{2 * 6}{72} = \frac{1}{6}$$

Ahora lo comparamos con los hat values y los siguientes son los puntos mayores al hii:

Cuadro 3: Primeros y ultimos valores hat values

	x
1	0.0353
2	0.0230
3	0.2238
4	0.0497
5	0.0992
6	0.0744
67	0.0404
68	0.0478
69	0.0339
70	0.2185
71	0.1440
72	0.0302

Este sería un resumen de la tabla completa la cual reducimos para no ver una tabla tan enorme y solo sacamos los valores que nos interesan:

Cuadro 4: Hat-values >0.166

	x
3	0.2238
13	0.1684
28	0.4187
31	0.1671
34	0.1674
41	0.1876
46	0.1703
49	0.2285
53	0.4789
63	0.1808
70	0.2185

Como podemos ver tenemos 11 datos de balanceo los cuales son los datos: 3,13,28,31,34,41,46,49,53,63,70 los cuales son mayores a $\frac{1}{6}$

4.4. Puntos influenciales

Para los puntos influenciables nos ayudaremos del diagnóstico DFFITS el cual nos dice que una observación será influenciable si $|DFFITS_i| > 2 * \sqrt{\frac{p}{n}}$ con el cual si reemplazamos $|DFFITS_i| > 2 * \sqrt{\frac{6}{72}} = 0.577$.

Nuestra tabla de DFFITS es algo como:

Cuadro 5: Primeros y ultimos valores DFFITS

	x
1	0.0318
2	0.0536
3	0.5402
4	-0.1056
5	-0.0180
6	-0.0759
67	-0.1598
68	0.0852
69	-0.0722
70	-0.1999
71	-0.4208
72	-0.0929

De la tabla completa sacamos los valores los cuales nos cumplan la condición que estamos buscando con la cual nos queda la siguiente tabla:

Cuadro 6: |DFFITS| >0.577

	x
28	-1.7225
46	-0.6049
53	-0.9470
63	1.4313

Como podemos ver solo tenemos 4 datos influenciales, los cuales son el 28,46,53,63.

4.5. Conclusión

Como conclusión podemos decir que el modelo no es del todo acertado dado que los datos 28,46,53,63 son datos influenciales y a la vez son datos de balanceo. Lo cual hace que esos datos halen el modelo en esta dirección y son observaciones que tienen valores inusuales por lo cual se recomendaría quitar dichas observaciones.