Estadística II, semestre 01 de 2023
Equipo de trabajo No. 16
Docente: Julieth Veronica Guarin Escudero.

3,8

Trabajo 1

Margy Karina Mateus, Nataly García Osorio, Daniel Hernández, Nicolas García
Universidad Nacional de Colombia, Sede Medellín
Fecha de entrega: 30 de marzo de 2023

Descripción del problema: En un estudio a gran escala realizado en EE.UU sobre la eficacia en el control de infecciones hospitalarias se recogió información en 113 hospitales con el fin de hallar un modelo que explique el riesgo de adquirir una infección hospitalaria dependiendo de la duración de la estadía del paciente, la rutina de cultivos, número de camas, censo promedio diario y número de enfermeras.

Punto 1 (1,5)

Estimación del modelo

Se propone el siguiente modelo de regresión lineal múltiple:

¿Quiénes son esas variables? No dicen que es X_i sea nada arriba ✓

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i, \quad \varepsilon_i \sim iid\ N(0, \sigma^2)\ ; \quad 1 \leq i \leq 55$$

Ahora, con los parámetros estimados se obtiene el siguiente modelo ajustado:

$$Y_i = 0.143592081 + 0.228535756 X_{i1} - 0.010908210 X_{i2} + 0.079624508 X_{i3} + 0.009309859 X_{i4} + 0.001768553 X_{i5}$$ ✓

$i = 1, 2, \ldots, 55$

3 pt

Significancia de la regresión:

Planteando el siguiente juego de hipótesis se analiza la significancia de la regresión: 4 pt

$$H_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \qquad vs. \qquad H_1 = \ algún\ \beta_j \neq 0, \qquad para\ j = 1,2,3,4,5$$ ✓

Para tomar una decisión frente al anterior juego de hipótesis, se tiene la siguiente tabla ANOVA:

¿Estadístico de prueba y región de rechazo?

Tabla 1: Tabla ANOVA para el modelo

	Suma de cuadrados	g.l	Cuadrado Medio	fo	P-valor
Modelo	84.8228	5	16.964568	24.6508	2.55966e-12
Error	33.7215	49	0.688194		

✓

De la tabla ANOVA, se observa un valor P muy pequeño aproximado a cero, teniendo en cuenta además un nivel de significancia del 5%, es posible rechazar la hipótesis nula, aceptando entonces la hipótesis alternativa en donde hay al menos un $\beta_j \neq 0, para\ j = 1,2,3,4,5$ ✓

Significancia de los parámetros individuales

6 pt

Se presenta el siguiente juego de hipótesis

$$H_0: \beta_j = 0 \qquad vs. \qquad H_1: \beta_j \neq 0 \quad j = 0,1,2,3,4,5.$$ ✓

¿Estadístico de prueba y RR?

Con la siguiente tabla de parámetros estimados:

Tabla 2: Resumen de los parámetros estimados

Parámetros	Estimación	Error Std	Valor T	Valor P
β_0	0.143592081	1.451355849	0.09893651	0.9215
β_1	0.228535756	0.0690061423	3.30916664	0.001759
β_2	-0.010908210	0.026960886	-0.40459392	0.68753
β_3	0.079624508	0.013274476	5.99831636	0,0000002354
β_4	0.009309859	0.007198780	1.29325509	0.20198
β_5	0.001768553	0.000634679	2.78653210	0.0075576

De la tabla anterior (*Tabla 2*), con un nivel de significancia de $\alpha = 0.05$ se concluye que los parámetros individuales $\beta_1, \beta_3, \beta_5$ son significativos en presencia de los demás parámetros y por el contrario, los parámetros $\beta_0, \beta_2, \beta_4$ no son significativos en presencia de los demás parámetros. Cabe aclarar que la significancia de β_0 es irrelevante en nuestro caso para una creación de un nuevo de modelo.

Interpretación individual de los parámetros ajustados

Así se concluye con los datos de la *Tabla 2: Resumen de los parámetros estimados*, que:

(β_1) Por cada día que aumente la duración en la estadía de los pacientes del hospital, el promedio del riesgo de infección aumenta significativamente un 0.228535% mientras que las demás predictoras se mantienen constantes.

(β_3) Por cada incremento unitario del número de camas, aumenta el promedio de contraer una infección hospitalaria significativamente en un 0.07962% cuando las demás predictoras se mantienen constantes.

(β_5) Por cada incremento unitario en el número de enfermeras, equivalentes a tiempo completo, aumenta significativamente en promedio el riesgo de contraer una infección hospitalaria en un 0.001768% cuando las demás predictoras se mantienen constantes.

Coeficiente de determinación múltiple R^2

El coeficiente de determinación múltiple R^2está dado por : $R^2 = \frac{SSR}{SST}$

Los resultados del SSE y SST, fueron obtenidos de la tabla ANOVA anteriormente, siendo así:

$$R^2 = \frac{84.8228}{84.8228 + 33.7215} = 0.7155$$

Esto se traduce en que el 71,55% de la variabilidad total en el riesgo promedio estimado de adquirir una infección hospitalaria es explicado por el modelo propuesto. Si bien todas las variables no son significativas individualmente, el modelo está siendo bien explicado por las variables propuestas por lo que se concluye que hay un buen ajuste.

Punto 2

Significancia simultánea del subconjunto

Para calcular la significancia simultánea del subconjunto, se tienen las tres variables predictoras con los valores p más grandes $\beta_2, \beta_4, \beta_5$, y por consiguiente se utilizará el siguiente juego de hipótesis:

$$H_0: \beta_2 = \beta_4 = \beta_5 = 0 \quad vs. \quad H_1: algún\ \beta_j \neq 0,\ \ j = 2,4,5$$

Cuadro 3: Resumen de todas las regresiones

	SSE	Covariables en el modelo
Modelo Completo	33.722	X1 X2 X3 X4 X5
Modelo Reducido	41.085	X1 X3

Un modelo reducido para la prueba de la significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_{i1}X_{i1} + \beta_{i3}X_{i3} + \varepsilon_i \ \ \varepsilon_i \sim iid\ N(0, \sigma^2);\ \ 1 \leq i \leq 55$$

Estadístico de prueba

Para hallar el estadístico de prueba, se utilizan sumas de cuadrados extras usando la tabla de todas las regresiones posibles brindadas. El estadístico de prueba para este juego de hipótesis está dado por:

$$F_0 = \frac{(SSE(\beta_1, \beta_3) - SSE(\beta_1, \beta_2\beta_3, \beta_4, \beta_5)/3}{MSE(\beta_0, \beta_1, \beta_2\beta_3, \beta_4, \beta_5)} \sim f_{3,49}$$

$$F_0 = \frac{[41.085 - 33.722]/3}{0.6882040816} = 3.566287093$$

Como $F_0 = 3.566287093 > f_{0.05,3,49} = 0.0149586$, por lo que se rechaza H_0 y el subconjunto es significativo, por lo que se concluye que la probabilidad promedio estimada de adquirir una infección en el hospital depende de al menos una de las variables asociadas $(\beta_2, \beta_4, \beta_5)$, así que **no** es posible descartar el modelo.

Punto 3

Preguntas propuestas:

1. ¿Está relacionado el número promedio de camas con el promedio de pacientes en el hospital por día durante el periodo de estudio?,
2. ¿Está relacionada la duración promedio de la estadía de los pacientes en el hospital con el número promedio de enfermeras, equivalentes a tiempo completo, en el periodo de estudio?.

Se quiere probar si $\beta_3 = \beta_4\ y\ \beta_1 = \beta_5$, con el siguiente juego de hipótesis

$$H_0: \beta_3 = \beta_4, \quad \beta_1 = \beta_5 \quad vs. \quad H_1: \beta_3 \neq \beta_4, \beta_1 \neq \beta_5$$

Reescrito con la hipótesis nula

$$H_0: \beta_1 - \beta_5 = 0, \beta_3 - \beta_4 = 0 \quad vs. \quad H_1: \beta_1 - \beta_5 \neq 0, \beta_3 - \beta_4 \neq 0$$

Teniendo entonces la hipótesis nula *m=2* ecuaciones y escribiendo la hipótesis nula de la forma $H_0: LB = 0$, se tiene que

$$H_0: \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

¿Quién es la matriz L?

En esta prueba se deben establecer los modelos:

Modelo full

$$MF = Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i \text{ ; } i \text{ supuestos}$$

Modelo Reducido

$$MR = Y_i = \beta_0 + \beta_1 (X_1 + X_5) + \beta_3 (X_{i3} + X_{i4}) + \beta_5 X_{i5} + \varepsilon_i \text{ ; } i \text{ supuestos}$$

Así, la expresión para el estadístico de prueba está dada por:

$$F_0 = \frac{MSH}{MSE} = \frac{SSH/2}{MSE} = \frac{[SSE(MR) - SSE(MF)]/2}{MSE} = \frac{[SSE(MR) - 33.722]/2}{0.6882040816}$$

En este caso, para calcular dicho estadístico de prueba, es necesario conocer el SSE(MR), sin embargo, esto no es posible, pues en la tabla de regresiones totales no se admiten las sumas de variables, por lo que se deja expresado.

Punto 4 15,5 pt

Validación de los supuestos:

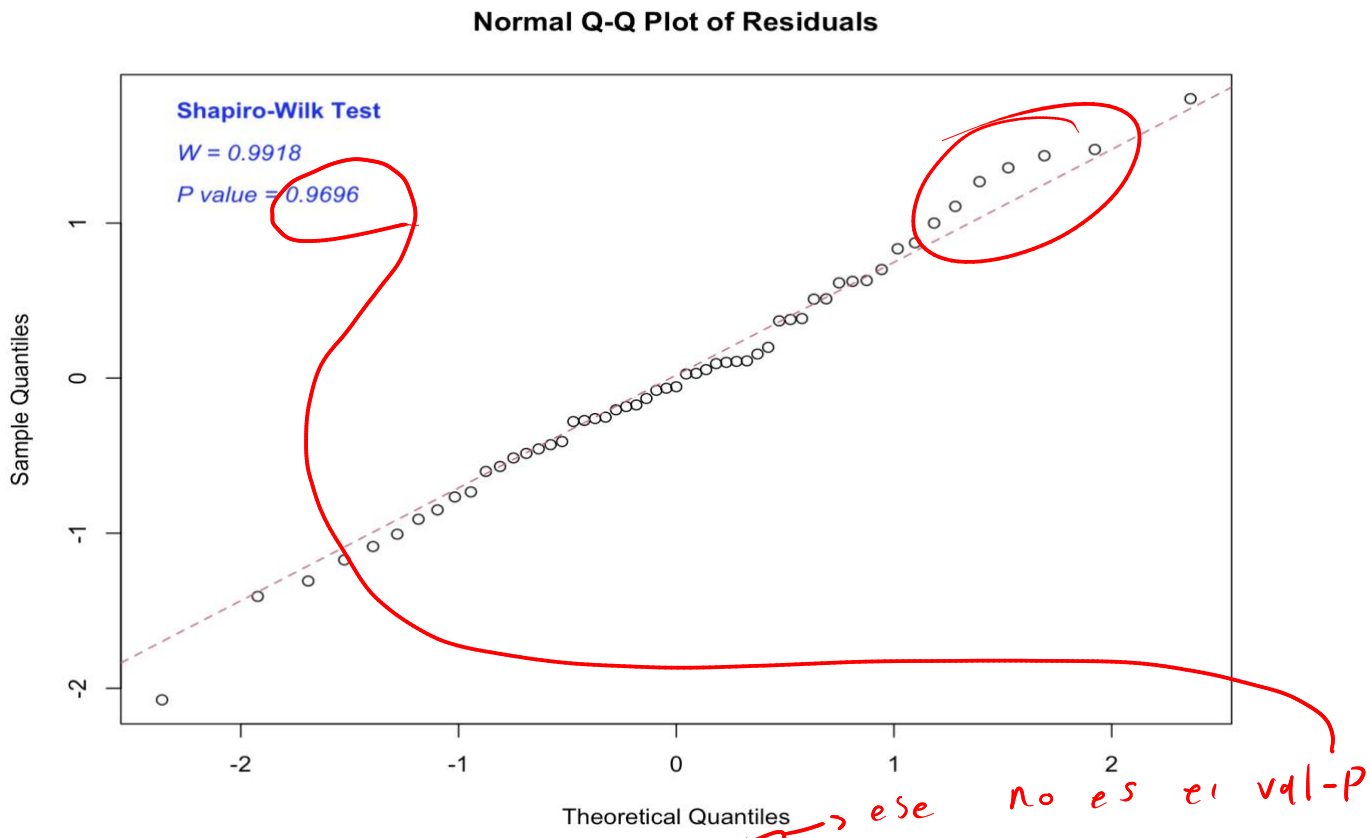
Para la validación de supuestos, se examinará más a detalle si hay valores atípicos, de balanceo e influenciales.

Normalidad de los residuos: 3,5 pt

Para validar que el modelo cumple con el criterio de normalidad, se realiza la siguiente prueba de hipótesis junto al gráfico cuantil-cuantil. ✓

$$H_0: \varepsilon_i \sim Normal \quad vs \quad H_1: \varepsilon_i \nsim Normal$$ ✓

Grafica 1: Cuantil-Cuantil



Haciendo la prueba de Shapiro-Wilk, se tiene que el valor p=0.9918 es un valor muy cercano a 1 y como este $VP > 0.05$, se aceptaría la hipótesis nula concluyendo que los errores se distribuyen normales, sin embargo, al observar más a detalle el *gráfico 1 cuantil-cuantil* se evidencian colas muy pesadas y patrones irregulares, por lo que al tener más peso el análisis gráfico, se termina **rechazando** la hipótesis nula y se acepta la alterna, pues la tendencia de puntos no logra explicar la recta ajustada.

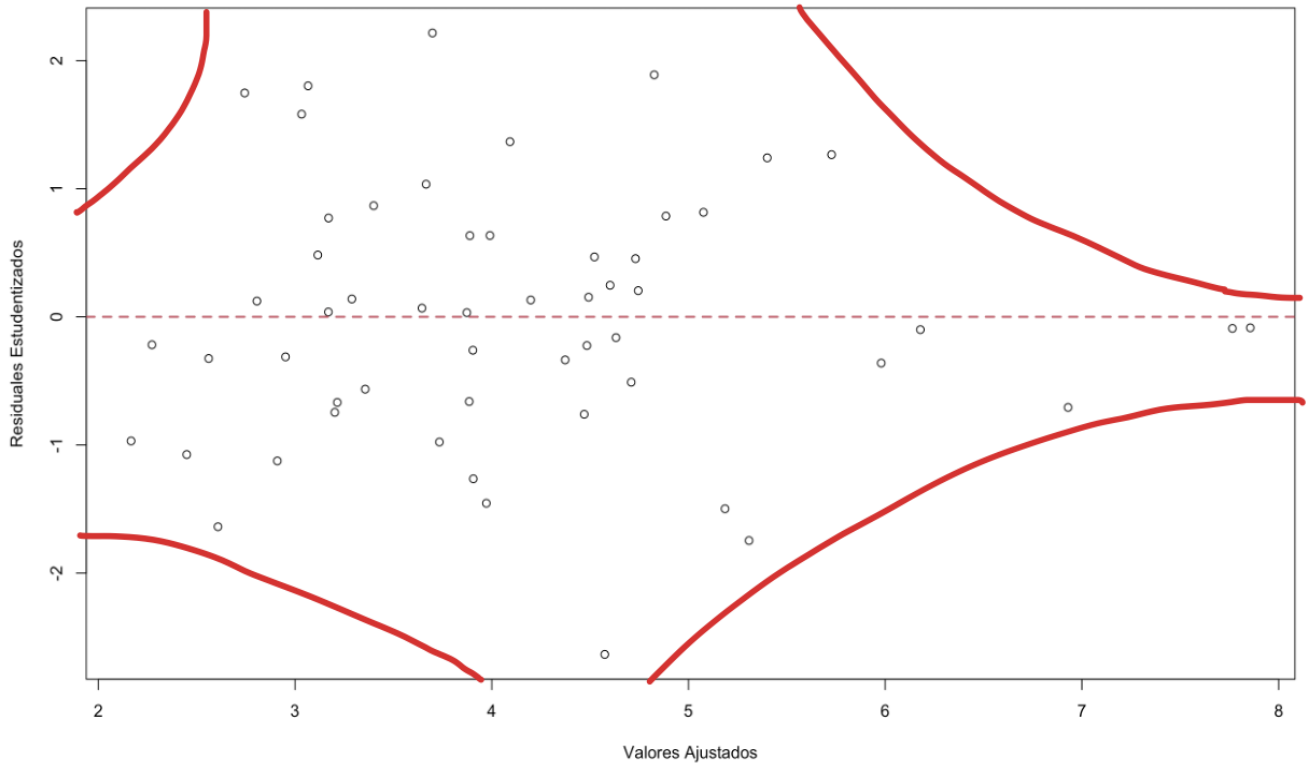
-> ese no es el val-P

-> Esta base de ddos no tiene tanta evidenáa en contra del supuesto, sin embargo les valgo el análisis,

Supuestos con Media 0 y varianza constante:

Se puede evidenciar con el siguiente gráfico de Residuales vs Valores ajustados que los supuestos de varianza constante y media cero no se cumplen a cabalidad, pues en el gráfico se observa claramente que la varianza aumenta y luego disminuye. Ahora, podemos observar unos puntos nivelados alrededor de cero pero con la varianza no constante No podemos asegurar con firmeza que la media es cero.

Gráfico 2: de los residuales vs los valores ajustados



Detección de outliers, puntos de balanceo o influenciales.

No pongan tanto dato innecesario, para qué yhat por ejemplo.

Tabla de diagnósticos

	Y	X1	X2	X3	X4	X5	yhat	se.yhat	residuals	res.stud	Cooks.D	hii.value	Dffits
1	4.3	10.39	54.6	14.0	88.3	353	4.4836	0.1364	-0.1836	-0.2244	0.0002	0.0271	-0.0370
2	4.8	12.01	52.8	10.8	96.9	298	4.6015	0.1984	0.1985	0.2465	0.0006	0.0572	0.0601
3	3.4	10.42	58.0	8.0	59.0	119	3.2890	0.2369	0.1110	0.1396	0.0003	0.0816	0.0412
4	3.7	8.48	51.1	12.1	92.8	166	3.6452	0.1823	0.0548	0.0678	0.0000	0.0483	0.0151
5	4.5	6.70	48.6	13.0	80.8	76	3.0664	0.2371	1.4336	1.8033	0.0482	0.0817	0.5508
6	3.7	7.58	56.7	20.8	88.0	97	3.9044	0.2709	-0.2044	-0.2607	0.0014	0.1066	-0.0892
7	2.9	7.91	52.8	11.9	79.5	477	3.9066	0.2333	-1.0066	-1.2644	0.0229	0.0791	-0.3729
8	1.4	7.14	51.7	4.1	45.7	115	2.1667	0.2489	-0.7667	-0.9688	0.0155	0.0900	-0.3045
9	2.9	10.80	63.9	1.6	57.4	130	2.8064	0.3346	0.0936	0.1233	0.0005	0.1627	0.0538
10	2.8	9.97	58.2	16.5	76.5	90	3.9724	0.1975	-1.1724	-1.4551	0.0212	0.0567	-0.3609
11	5.7	11.80	53.8	9.1	116.9	571	5.0762	0.3216	0.6238	0.8157	0.0196	0.1503	0.3419
12	4.3	8.67	48.2	24.4	90.8	182	4.7093	0.2122	-0.4093	-0.5103	0.0030	0.0654	-0.1340
13	6.6	13.95	65.9	15.6	133.5	356	5.7274	0.4618	0.8726	1.2662	0.1200	0.3099	0.8540
14	2.7	8.34	56.9	8.1	74.0	107	2.9520	0.1919	-0.2520	-0.3123	0.0009	0.0535	-0.0736
15	4.9	9.89	50.5	17.7	103.6	167	4.5221	0.1890	0.3779	0.4678	0.0020	0.0519	0.1085
16	2.6	9.76	53.2	6.9	80.1	64	3.2021	0.1912	-0.6021	-0.7459	0.0052	0.0531	-0.1759
17	4.5	11.46	56.9	15.6	97.7	191	4.6314	0.1784	-0.1314	-0.1622	0.0002	0.0462	-0.0354
18	4.1	9.05	51.2	20.5	79.8	195	4.3734	0.1567	-0.2734	-0.3357	0.0007	0.0357	-0.0640
19	5.1	9.76	50.9	21.9	97.0	150	4.7310	0.1716	0.3690	0.4547	0.0015	0.0428	0.0954
20	2.1	8.02	55.0	3.8	46.5	91	2.2729	0.2457	-0.1729	-0.2182	0.0008	0.0877	-0.0670
21	6.3	8.84	56.3	29.6	82.6	85	4.8259	0.2819	1.4741	1.8893	0.0777	0.1155	0.7016
22	4.1	9.35	53.8	15.9	80.9	833	5.1859	0.4035	-1.0859	-1.4982	0.1159	0.2366	-0.8450
23	4.5	9.61	52.4	6.9	87.2	487	3.9907	0.2122	0.5093	0.6350	0.0047	0.0655	0.1670
24	1.3	8.92	53.9	2.2	79.5	56	2.6085	0.2239	-1.3085	-1.6381	0.0351	0.0729	-0.4675
25	6.1	13.59	54.0	24.2	111.7	312	6.1790	0.2430	-0.0790	-0.0996	0.0002	0.0858	-0.0302
26	3.9	8.28	49.5	12.0	113.1	546	4.4700	0.3565	-0.5700	-0.7609	0.0219	0.1847	-0.3606
27	3.9	11.15	56.5	7.7	73.9	281	3.8735	0.1833	0.0265	0.0327	0.0000	0.0488	0.0073
28	1.6	8.82	58.2	3.8	51.7	80	2.4498	0.2515	-0.8498	-1.0750	0.0195	0.0919	-0.3426
29	2.3	7.95	51.8	4.6	54.9	163	2.5611	0.2051	-0.2611	-0.3248	0.0011	0.0612	-0.0821
30	4.3	8.30	57.2	6.8	83.8	167	3.0335	0.2172	1.2665	1.5819	0.0307	0.0686	0.4360
31	3.9	10.73	50.6	19.3	101.0	445	5.3079	0.1958	-1.4079	-1.7464	0.0300	0.0557	-0.4335
32	5.2	9.53	51.5	15.0	65.7	298	4.0928	0.1799	1.1072	1.3672	0.0154	0.0470	0.3064
33	3.5	7.94	49.5	6.2	92.3	195	3.1160	0.2357	0.3840	0.4827	0.0034	0.0807	0.1419
34	4.3	9.89	45.2	11.8	108.7	190	4.1983	0.3042	0.1017	0.1317	0.0004	0.1345	0.0514
35	2.0	8.93	56.0	6.2	72.5	95	2.9102	0.1806	-0.9102	-1.1242	0.0105	0.0474	-0.2514
36	6.4	11.62	53.9	25.5	99.2	133	5.4004	0.1993	0.9996	1.2413	0.0157	0.0577	0.3090
37	4.6	10.16	54.2	8.4	51.5	831	4.4923	0.4432	0.1077	0.1536	0.0016	0.2854	0.0961
38	5.5	8.37	50.7	15.1	84.8	115	3.6986	0.1659	1.8014	2.2163	0.0341	0.0400	0.4721
39	3.4	8.45	38.8	12.9	85.0	235	3.8856	0.3851	-0.4856	-0.6609	0.0200	0.2155	-0.3444
40	5.7	11.20	56.5	34.5	88.9	180	5.9799	0.2926	-0.2799	-0.3606	0.0031	0.1244	-0.1347
41	3.0	11.20	45.0	7.0	78.9	130	3.7342	0.3516	-0.7342	-0.9771	0.0348	0.1796	-0.4570
42	4.1	7.13	55.7	9.0	39.6	279	2.7442	0.2942	1.3558	1.7479	0.0732	0.1257	0.6776
43	4.1	10.47	53.2	5.7	69.1	196	3.3999	0.1930	0.7001	0.8678	0.0072	0.0542	0.2071
44	5.5	11.08	50.2	18.6	63.6	387	4.8857	0.2796	0.6143	0.7865	0.0132	0.1136	0.2804
45	6.5	19.56	59.9	17.2	113.7	306	6.9296	0.5648	-0.4296	-0.7070	0.0720	0.4635	-0.6538
46	7.8	12.07	43.7	52.4	105.3	157	7.8556	0.5113	-0.0556	-0.0852	0.0007	0.3798	-0.0660
47	2.9	8.86	51.3	9.5	87.5	100	3.3567	0.1817	-0.4567	-0.5643	0.0027	0.0480	-0.1258
48	4.5	9.44	52.5	10.9	58.5	297	3.6661	0.1979	0.8339	1.0351	0.0108	0.0569	0.2544
49	4.9	10.23	53.2	9.9	77.9	752	4.7447	0.3362	0.1553	0.2048	0.0014	0.1642	0.0899
50	3.2	8.19	52.1	10.8	59.2	176	3.1693	0.1811	0.0307	0.0379	0.0000	0.0477	0.0084
51	2.5	8.54	56.1	27.0	82.5	98	4.5746	0.2612	-2.0746	-2.6348	0.1273	0.0991	-0.9337
52	3.8	8.66	52.8	6.8	69.5	246	3.1703	0.1473	0.6297	0.7713	0.0032	0.0315	0.1386
53	7.7	12.78	56.8	46.0	116.9	322	7.7652	0.4034	-0.0652	-0.0900	0.0004	0.2365	-0.0496
54	2.7	7.14	57.6	13.1	92.6	92	3.2149	0.3064	-0.5149	-0.6679	0.0117	0.1364	-0.2639
55	4.4	10.02	49.5	8.3	93.0	265	3.8889	0.1963	0.5111	0.6341	0.0040	0.0560	0.1535

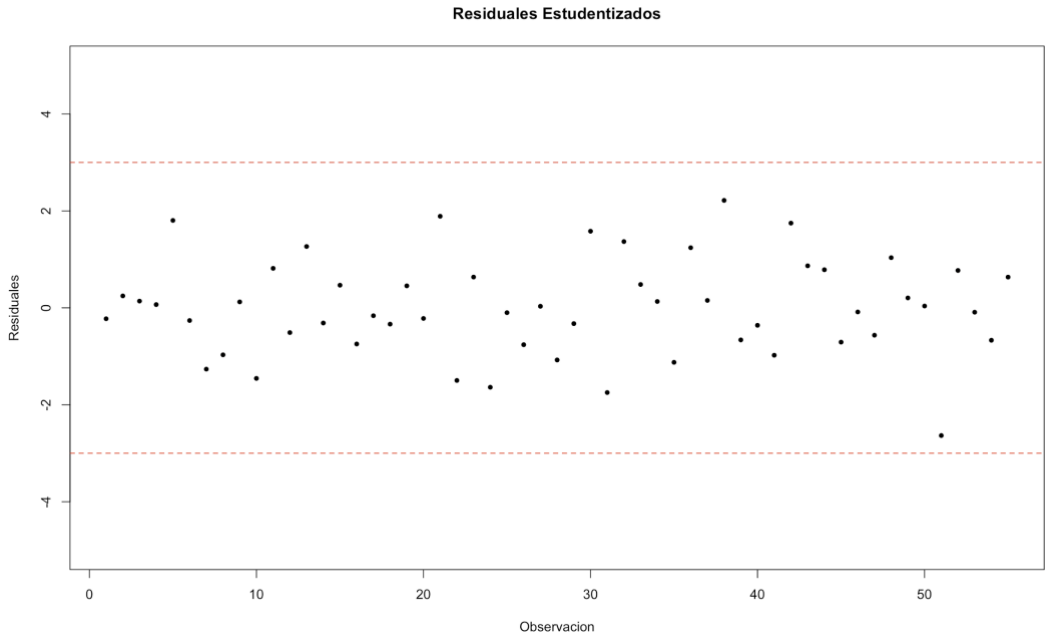
copiaron y pegaron esta tabla como imagen, no

Siguen un formato general.

Datos atípicos: 3 pt

Pese a que en el gráfico de normalidad las observaciones se ven agrupadas, se puede notar la presencia de dos candidatos a ser outlier pero al verificar en la tabla se observa que **no** existen datos atípicos, pues una observación es atípica cuando su residual estudentizado r_i, es tal que: $|r_i| > 3$, caso que no se cumple en la columna "res estud" de la tabla anterior. ✓

Gráfico 3: Análisis a los Residuales estudentizados:



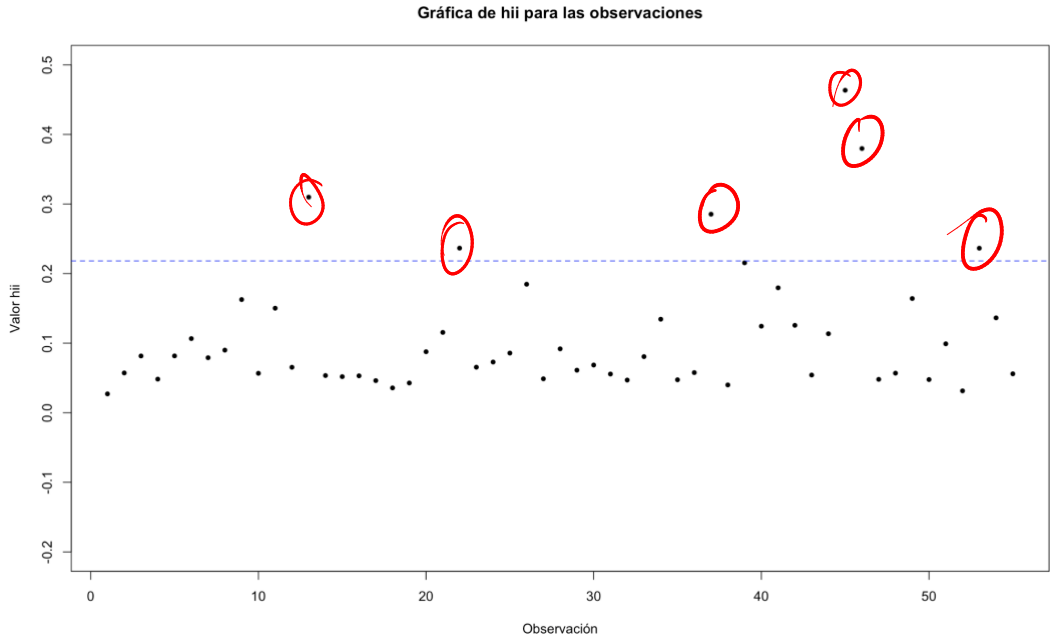
Según el gráfico de Análisis a los Residuales estudentizados, se puede confirmar que en el modelo no existen datos atípicos pues no hay datos fuera de la región acotada. ✓

Puntos de balanceo 2 pt

$hii > 2 * p/n \quad = \quad hii > 2 * 6/55 \quad = \quad hii > 0.21818$

Así, teniendo en cuenta la tabla diagnóstica, las observaciones que cumplen está desigualdad son: 13, 22, 37, 45, 46, 53.

Gráfico 4: Hii para las observaciones



Según el grafico 4, se confirma que al igual que en la tabla diagnostica, se observan los 6 puntos de balanceo.

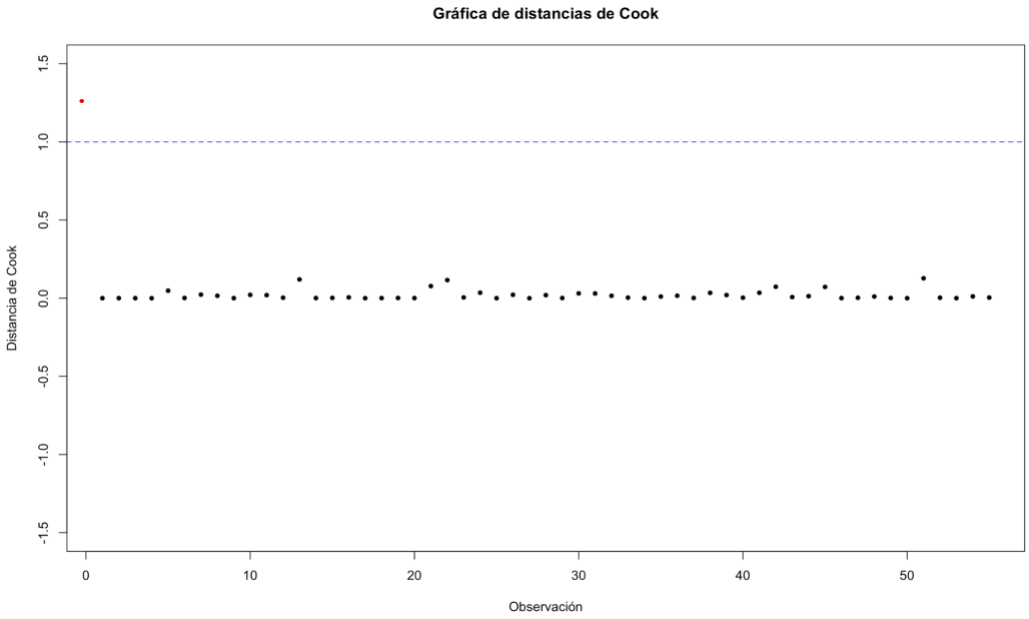
¿Qué causan estos puntos?

2 o 7

Puntos influenciales
Por último, se identifican
los puntos influenciales usando los siguientes métodos:

Distancia de Cook (Di>1) , en la tabla diagnóstica, no se encuentra ningún valor superior a 1, así que se considera que no hay ningún punto influencial mediante este método de evaluación.

Gráfico 5: Distancia de Cook

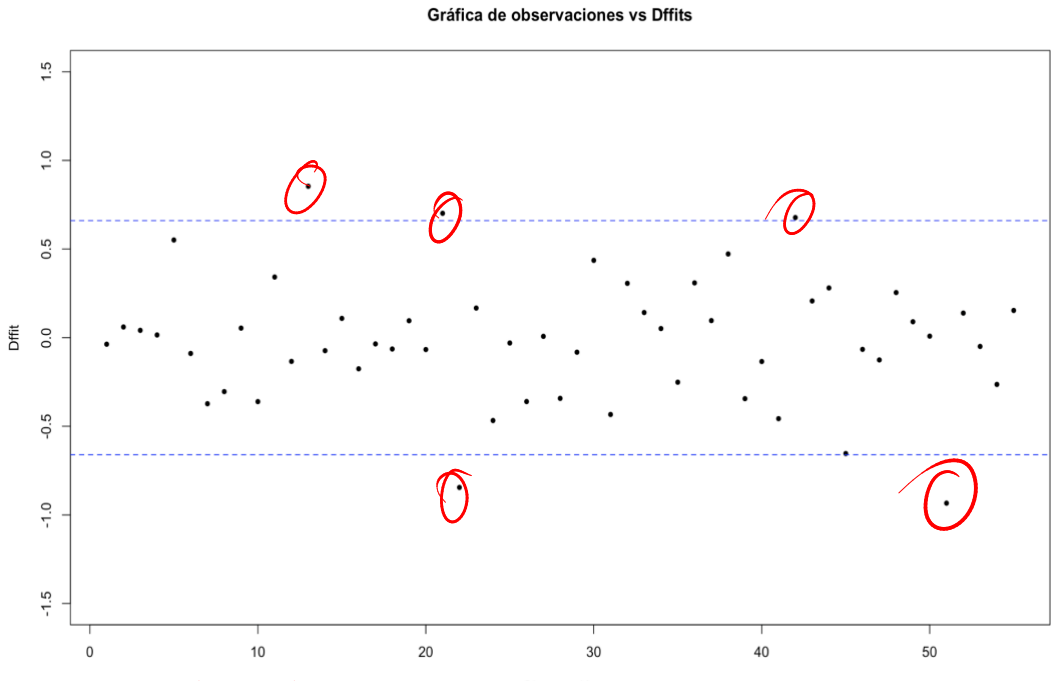


Con el análisis gráfico de la Distancia de Cook es posible reafirmar lo anteriormente mencionado en la tabla diagnostica

$|DFFITSi| > 2*\sqrt{p/n} = |DFFITSi| > 2*\sqrt{6/55} = |DFFITSi| > 0.6605$

En este criterio y teniendo en cuenta la tabla diagnostica, las observaciones : 13, 21, 22, 42 y 51 se consideran puntos influénciales del modelo de regresión

Gráfico 6: Observaciones vs Dffits



Conclusión

Con base al análisis previos, es posible concluir que el modelo no es lo suficientemente confiable como para arrojar una variable de respuesta optima en el estudio sobre la eficacia del control de enfermedades hospitalarias, pues a pesar de que no posee datos atípicos, de balanceo e influénciales el modelo sigue teniendo un mal comportamiento dadas las siguientes problemáticas:

No cumple con los supuestos de normalidad y varianza constante, por lo que de entrada demuestra que no es un estudio funcional para este caso. Puede influir el hecho de contar con un tamaño de muestra pequeño, pues 55 datos no son suficientes para sacar resultados óptimos para el estudio. Se necesitan otras variables que puedan proporcionar más información para desarrollar y analizar el caso de investigación.

Por lo anterior, el modelo se rechaza.