4,2

Universidad Nacional de Colombia sede Medellín

Facultad de ciencias

Escuela de Estadística

Taller 1

Estadística II

David Julián Taimal Poso
1144091406

Juan Esteban Rodriguez Ramirez

1000869132

Grupo 37

Profesor

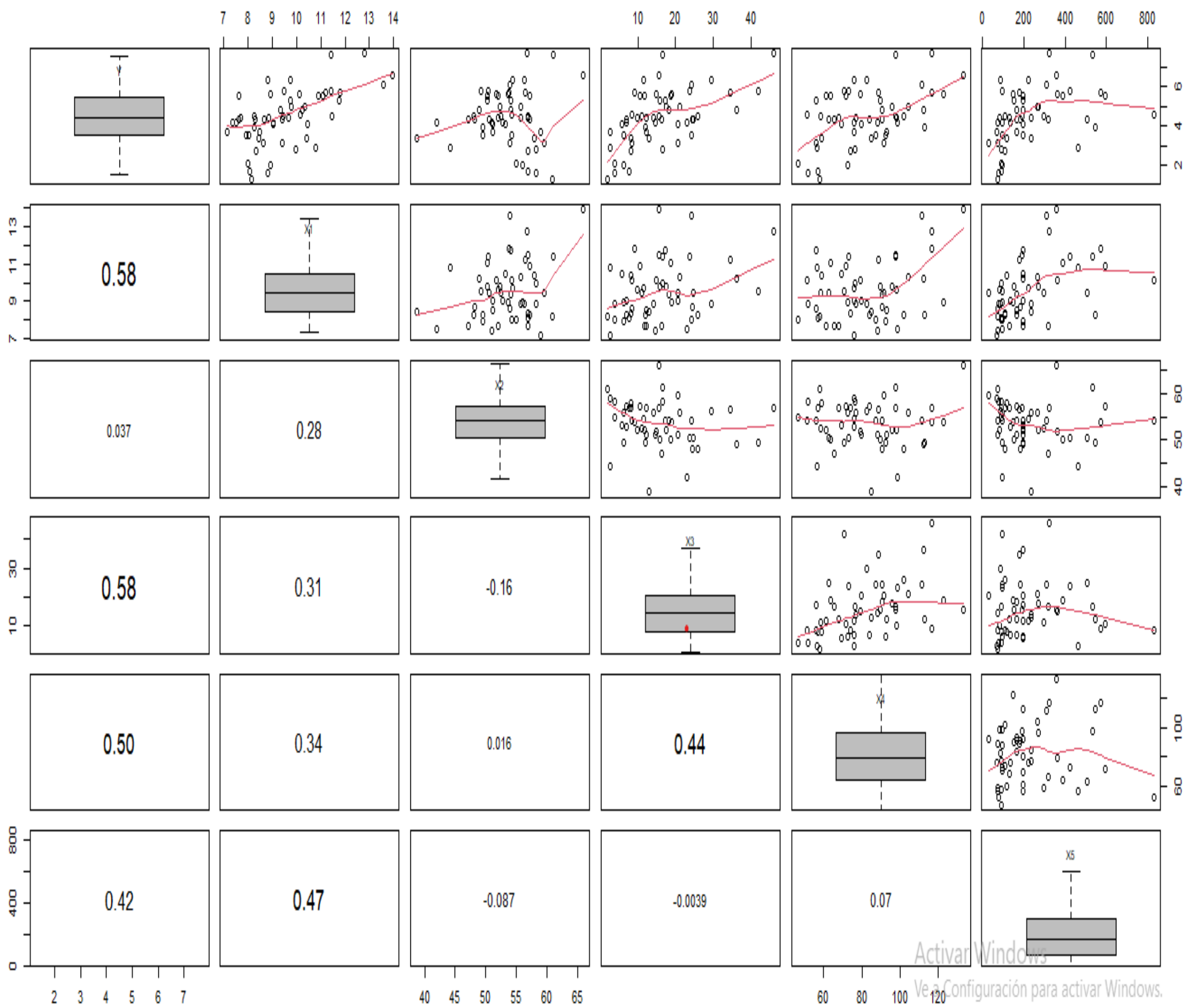
Mateo Ochoa Medina

Medellín

Octubre de 2023

Variable	Descripción
Y: **Riesgo de infección**	Probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje).
X_1: **Duración de la estadía**	Duración promedio de la estadía de todos los pacientes en el hospital (en días).
X_2: **Rutina de cultivos**	Razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100.
X_3: **Número de camas**	Número promedio de camas en el hospital durante el periodo del estudio.
X_4: **Censo promedio diario**	Número promedio de pacientes en el hospital por día durante el periodo del estudio.
X_5: **Número de enfermeras**	Número promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio.

Matriz de gráficas de dispersión con boxplots y correlaciones:



Nota: Para el presente trabajo, cada una de las pruebas de hipótesis que se desarrollan se realizaron con un nivel de significancia estadística de $\alpha = 0.05$.

1. Con base en el análisis de la matriz de gráficos de dispersión se plantea un modelo de RLM para el problema:

20pt

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i,$$
$i=1,2, \ldots, 54$

que tiene como supuestos lo siguiente:
$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \forall i=1,2,\ldots,54$

- **Tabla de parámetros estimados:**

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.322543552	1.5518288972	-0.8522483	0.3983086910
X1	0.181592982	0.1129991330	1.6070299	0.1146069881
X2	0.022387928	0.0304891252	0.7342922	0.4663427008
X3	0.063135560	0.0157336347	4.0127765	0.0002093704
X4	0.014734466	0.0070648808	2.0855930	0.0423556146
X5	0.002689781	0.0009152381	2.9388872	0.0050511224

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

3pt

$$\hat{y}_i = -1.322543552 + 0.181592982X_{i1} + 0.022387928X_{i2} + 0.063135560X_{i3} + 0.014734466X_{i4} + 0.002689781X_{i5}$$
$i = 1, 2, 3, \ldots, 54$

- **Significancia de la regresión:**

Para analizar la significancia de la regresión, se plantea el siguiente juego de hipótesis:

$H0: \beta 1 = \beta 2 = \beta 3 = \beta 4 = \beta 5 = 0$, vs.
$H1:$ algún $\beta j \neq 0$, $j = 1, 2, 3, 4, 5$.

Para poder validar H0 o no, se usa la tabla ANOVA.

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	64.3022	5	12.860430	15.3061	5.5202e-09
Error	40.3304	48	0.840218		

5pt

De la tabla ANOVA anterior se obtienen los valores del estadístico de prueba F0 = 15.3061 y su correspondiente valor-P vp = 5.5202 x 10^09, a un nivel de significancia α = 0.05.
Como vp < 0.05 = α se rechaza H0 concluyendo que el modelo de RLM propuesto es significativo. Esto quiere decir, que el **riesgo de infección** (probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje)) depende significativamente de al menos una de las predictoras del modelo.

- **Significancia de los parámetros individuales estimados:** 6pt

Se puede observar de la **Tabla de parámetros estimados** que β0, β1, β2 son individualmente no significativos en presencia de los demás parámetros.

También concluimos que sólo se podrán interpretar parámetros que resultaron significativos individualmente, en este caso son: β3, β4 y β5, pues sus P-valores son menores a α.
Por otro lado se podría añadir que β^0=-1.322543552 y cómo Xj = 0 Ɇ [Xj,min, Xj,max] ∀j entonces este valor no es interpretable.

	Y	X1	X2	X3	X4	X5
min	1.3	7.14	38.8	1.9	46.5	29
max	7.7	13.95	65.9	46.0	133.5	831

Ahora pasamos a interpretar cada parámetro que resultó individualmente significativo:

3pt

β^3 = 0.063135560, indica que por cada unidad que se aumente el **número de camas** (número promedio de camas en el hospital durante el periodo del estudio) (X3), la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje), aumenta (signo '+' de β^3) en 0.063135560 unidades, cuando las demás predictoras se mantienen fijas.

β^4 = 0.014734466, indica que por cada unidad que se aumente el **censo promedio diario** (número promedio de pacientes en el hospital por día durante el periodo del estudio) (X4), la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje), aumenta (signo '+' de β^4) en 0.014734466 unidades, cuando las demás predictoras se mantienen fijas.

β^5 = 0.002689781, indica que por cada unidad que se aumente el **número de enfermeras** (número promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio)(X5), la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje), aumenta (signo '+' de β^5) en 0.002689781 unidades, cuando las demás predictoras se mantienen fijas.

- Cálculo e interpretación de R^2: 3pt

  Sabemos que $R^2 = \frac{SSR}{SST}$ de manera que se puede calcular de la tabla ANOVA.

$$R^2 = \frac{SSR}{SST} = \frac{64.3022}{64.3022 + 40.3304} = 0.6145522$$

A partir del cálculo del R^2 se concluye que 61.46% de la variabilidad total en el **riesgo de infección** (probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje)) es explicado por el modelo de RLM propuesto.
En sentido inverso el 38.54% de la variabilidad total del **riesgo de infección** (probabilidad promedio de adquirir infección en el hospital(en porcentaje)), es explicado por el error.

Por otro lado, se puede calcular el R2 ajustado como una medida de bondad de ajuste, así:

$$R^2 adj = 1 - \frac{(n-1)MSE}{SST} = 1 - \frac{(54-1)0.840218}{64.3022 + 40.3304} = 0.5744007$$

De lo anterior podemos concluir que el valor de $R^2 adj$ = 0.5744007 es menor que R^2 = 0.6145522, lo que indica que en el modelo puede haber variables que no aporten significativamente. En otras palabras, se puede depurar el modelo (quitar variables que no aporten).

2. Tabla de todas las regresiones posibles: 4 pt

k	R_sq	adj_R_sq	SSE	Cp	Variables_in_model
5	0.615	0.574	40.330	6.000	X1 X2 X3 X4 X5
2	0.357	0.332	67.268	32.060	X1 X2

Son los datos que usaremos de la tabla.

El número de modelos posibles con k predictoras es 2^k-1. Para este taller, tenemos $2^5 - 1 = 31$ modelos posibles.

- Las 3 covariables con el P-valor más pequeño en el modelo fueron X3, X4, X5 ahora se debe probar la significancia simultánea de las variables asociadas al número de camas (X3), censo promedio diario (X4), Número de enfermeras (X5), que equivale a la siguiente prueba de hipótesis.

  $H0: \beta 3 = \beta 4 = \beta 5 = 0$ vs. $H1:$ Algún $\beta j \neq 0, j = 3, 4, 5.$

  Para probar esta hipótesis se usan sumas de cuadrados extra y nos apoyaremos en la tabla de todas las regresiones posibles.

  Luego un modelo reducido para la prueba de significancia del subconjunto es:

  $Y_i = \beta 0 + \beta 1 X_{i1} + \beta 2 X_{i2} + \varepsilon i;\ i=1,2,\ldots,54.$

  $\varepsilon_i \text{ iid} \sim N(0, \sigma^2), \forall i = 1, 2, \ldots, 54$

  Para esta prueba de hipótesis se tiene como estadístico de prueba a:

$$F0 = \frac{MSextra}{MSE} = \frac{\frac{SSR(\beta 3, \beta 4, \beta 5 | \beta 0, \beta 1, \beta 2)}{3}}{MSE}$$

$$= \frac{(SSE(\beta 0, \beta 1, \beta 2) - SSE(\beta 0, \beta 1, \beta 2, \beta 3, \beta 4, \beta 5))/3}{MSE}$$

2 pt

$$= \frac{[67.268 - 40.330\ ]/3}{0.840218} = 10.686909$$

Para el criterio de decisión se requiere obtener el valor crítico de una distribución $f_{3,54-6} = f_{3,48}$ a un nivel de significancia $\alpha = 0.05$, esto es, $f_{0.05,3,48} = 2.798061$. Como $F_0 = 10.6869 > f_{0.05,3,48} = 2.798061$, entonces se rechaza H_0 y se concluye que la probabilidad promedio estimada de adquirir infección en el hospital depende de al menos una de las variables asociadas al número de camas (X3), censo promedio diario (X4), número de enfermeras (X5). Por tanto, no es posible descartar las variables de este subconjunto (X3, X4, X5) del modelo. 2 pt

3. **Prueba de la hipótesis lineal general.** 4 pt

Teniendo en cuenta el contexto de la regresión, se plantea la siguiente pregunta: ¿Si se presenta un aumento de una unidad en el número promedio de enfermeras, equivalentes a tiempo completo (X5), durante el periodo del estudio, la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje) incrementará el doble que si aumentara en una unidad el número promedio de pacientes en el hospital por día durante el periodo del estudio (X4)?

Adicionalmente podríamos preguntarnos si ¿el efecto en la probabilidad promedio estimada de adquirir infección en el hospital que causa el aumentar en un día la duración promedio de la estadía de todos los pacientes en el hospital (X1), es igual al efecto ocasionado por aumentar en dos el número promedio de camas en el hospital(X3) durante el mismo periodo?

$$\begin{cases} \mathrm{H}_0 : \beta_4 = 2\beta_5;\ \beta_1 = 2\beta_3 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

Podemos reescribir la hipótesis nula de la siguiente manera:

$$H_0 : \begin{cases} \beta_4 - 2\beta_5 = 0 \\ \beta_1 - 2\beta_3 = 0 \end{cases}$$

De manera que la hipótesis nula tiene *r=2* ecuaciones.

Y de forma matricial se puede ver de la siguiente forma:

$$\text{H0}: \; L = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & -2 \\ 0 & 1 & 0 & -2 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta 0 \\ \beta 1 \\ \beta 2 \\ \beta 3 \\ \beta 4 \\ \beta 5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

2pt

De esta manera tenemos una matriz ***L*** de dimensiones 2*5 con *r=2* filas independientes y que nos dejaría un modelo reducido RM:

2Xi + X3i   2X4i + X5i

$$\begin{aligned} Y &= \beta_0 + \beta_1 (X_{1i} + 2X_{3i}) + \beta_2 X_2 + \beta_4 (X_{4i} + 2X_{5i}) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \\ &= \beta_0 + \beta_1 X_{1,3,i} + \beta_2 X_2 + \beta_4 X_{4,5,i} + \varepsilon_i \end{aligned}$$

X 0pt

En este modelo se tiene una suma de cuadrados del error SSE(RM) = SSE ($\beta_0, \beta_1, \beta_2, \beta_4$) con n-6 grados de libertad. Luego el estadístico de prueba F0 dado por:

$$F_0 = \frac{MSH}{MSE} = \frac{[SSE(MR) - SSE(MF)]/GL}{MSE} = \frac{[SSE(MR) - 40.3304]/2}{0.840218}$$

Con región de rechazo: $F_0 > f_{0.05,2,54}$

2pt

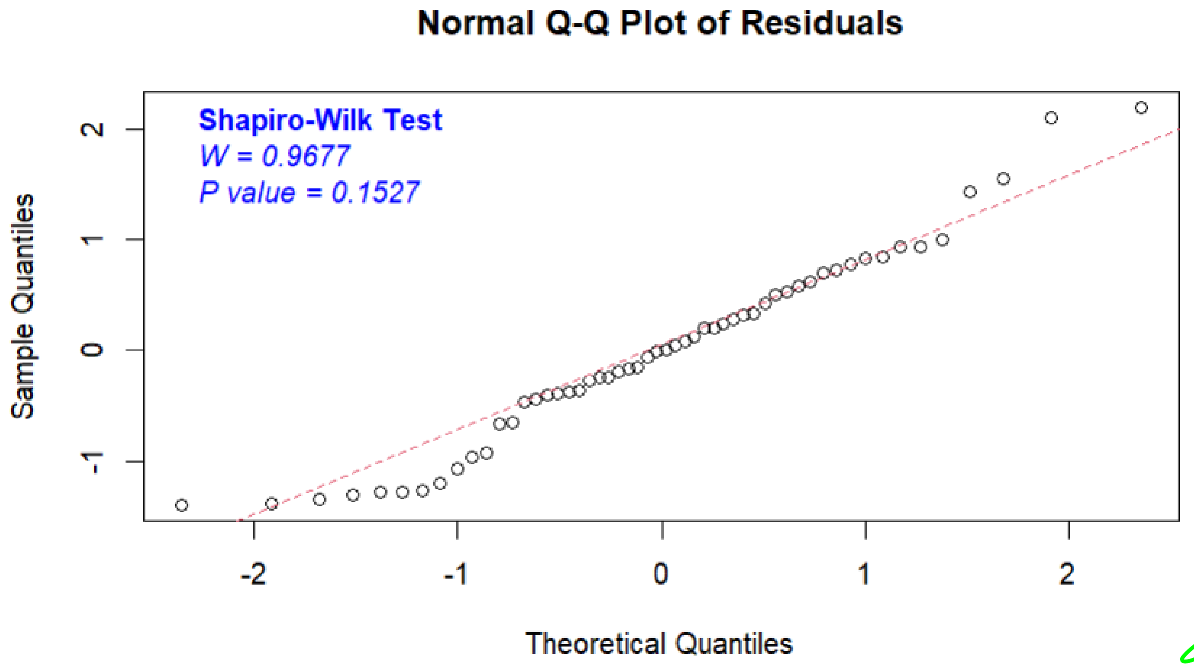
Solo resta establecer el valor SSE(MR), el cual no se puede obtener de la tabla de todas las regresiones posibles, ya que ésta no admite sumas de variables entre sus opciones.

4. **Validación de los supuestos sobre los errores.** 13,5pt

Queremos probar:

H_0: $\varepsilon_i \sim$ Normal vs. H_1: $\varepsilon_i \sim$ No Normal

Supuesto de normalidad – Gráfica de normalidad y prueba de Shapiro-Wilk



4 pt

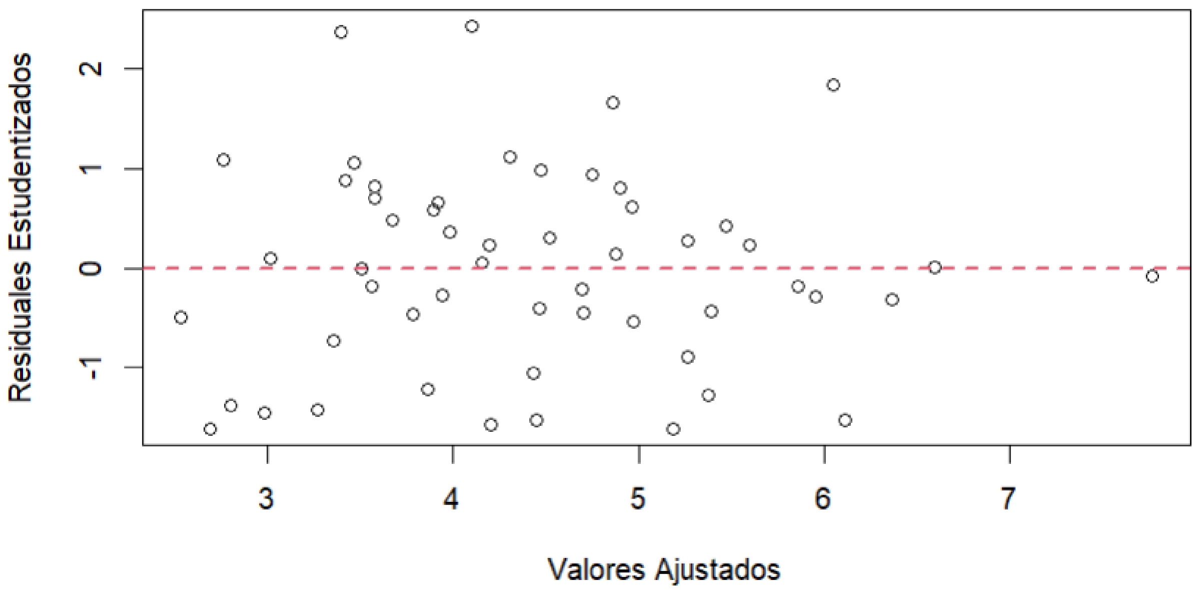
Como el patrón de los residuales no sigue la línea roja que representa el ajuste de la distribución de los residuales a una distribución normal, quizás eso pueda deberse a la → y balanceo presencia de observaciones influencia~~b~~les que pueden aumentar la variabilidad del modelo, entonces se concluye que, debido a este motivo, el supuesto de normalidad no se cumple. **Esto se concluye a pesar de que la prueba de normalidad S-W indica que los errores son normales (valor-P = 0.1527 mayor a 0.05).** Habría que corregir los datos que están generando esa alta variabilidad.

Supuesto de varianza constante - Gráfica de residuales vs. Valores ajustados

Se quiere probar:

$$H_0\text{: } V[\varepsilon_i] = \sigma^2 \quad \text{VS.} \quad H_1\text{: } V[\varepsilon_i] \neq \sigma^2$$

Para ello visualizamos la gráfica de residuales estudentizados vs los valores ajustados de la regresión:



3 pt

En la visualización de la gráfica se observa que el patrón formado por la nube de puntos no se aleja mucho de un patrón rectangular.
Lo anterior nos lleva a pensar que el supuesto de varianza constante de los errores si se cumple.
Es posible que algunas observaciones extremas estén afectando este análisis.

Análisis de la presencia de observaciones extremas:

	Y	X1	X2	X3	X4	X5	yhat	se.yhat	residuals	res.stud	Cooks.D	hii.value	Dffits
1	6.1	13.59	54.0	24.2	111.7	312	6.3672	0.3939	-0.2672	-0.3228	0.0039	0.1847	-0.1522
2	4.6	10.16	54.2	8.4	51.5	831	5.2602	0.5471	-0.6602	-0.8977	0.0743	0.3562	-0.6664
3	5.6	8.95	53.7	18.9	122.8	147	4.9030	0.3201	0.6970	0.8115	0.0152	0.1220	0.3014
4	3.4	8.45	38.8	12.9	85.0	235	3.7795	0.4402	-0.3795	-0.4721	0.0111	0.2307	-0.2564
5	4.8	10.24	49.0	36.3	112.6	195	6.1094	0.3192	-1.3094	-1.5239	0.0534	0.1213	-0.5742
6	4.1	9.05	51.2	20.5	79.8	195	4.4617	0.1587	-0.3617	-0.4007	0.0008	0.0300	-0.0698
7	6.3	8.84	56.3	29.6	82.6	85	4.8577	0.3008	1.4423	1.6657	0.0558	0.1077	0.5899
8	6.3	9.74	54.4	11.4	76.1	221	4.0996	0.1465	2.2004	2.4318	0.0258	0.0255	0.4159
9	4.3	7.65	47.1	16.4	65.7	318	3.9799	0.2849	0.3201	0.3674	0.0024	0.0966	0.1190
10	4.4	8.88	55.8	14.1	76.8	237	4.1985	0.1714	0.2015	0.2237	0.0003	0.0350	0.0422
11	2.8	9.97	58.2	16.5	76.5	90	4.2019	0.2160	-1.4019	-1.5737	0.0243	0.0555	-0.3877
12	5.7	11.20	56.5	34.5	88.9	180	5.9484	0.3166	-0.2484	-0.2888	0.0019	0.1193	-0.1053
13	4.8	9.36	54.1	18.3	90.6	165	4.5225	0.1476	0.2775	0.3068	0.0004	0.0259	0.0496
14	5.8	9.50	49.3	42.0	70.9	98	5.4663	0.4455	0.3337	0.4166	0.0089	0.2362	0.2296
15	4.3	8.67	48.2	24.4	90.8	182	4.6989	0.2232	-0.3989	-0.4487	0.0021	0.0593	-0.1117
16	4.4	7.70	56.9	12.2	67.9	129	3.4673	0.2492	0.9327	1.0573	0.0149	0.0739	0.2991
17	2.7	8.34	56.9	8.1	74.0	107	3.3554	0.2072	-0.6554	-0.7340	0.0048	0.0511	-0.1695
18	3.7	8.48	51.1	12.1	92.8	166	3.9392	0.2014	-0.2392	-0.2675	0.0006	0.0483	-0.0597
19	1.3	8.16	60.9	1.9	58.0	73	2.6936	0.3152	-1.3936	-1.6190	0.0586	0.1182	-0.6033
20	4.1	10.47	53.2	5.7	69.1	196	3.6750	0.2735	0.4250	0.4858	0.0038	0.0891	0.1507
21	5.6	10.12	51.7	14.9	79.1	362	4.7526	0.1690	0.8474	0.9406	0.0052	0.0340	0.1762
22	4.5	8.28	48.1	26.0	101.8	108	4.6899	0.2748	-0.1899	-0.2172	0.0008	0.0898	-0.0675
23	2.0	8.93	56.0	6.2	72.5	95	3.2680	0.2093	-1.2680	-1.4209	0.0185	0.0521	-0.3369
24	1.7	8.09	56.9	7.6	56.9	92	2.9861	0.2446	-1.2861	-1.4558	0.0271	0.0712	-0.4079
25	7.7	12.78	56.8	46.0	116.9	322	7.7627	0.4521	-0.0627	-0.0786	0.0003	0.2433	-0.0441

26	3.7	7.14	59.0	2.6	75.8	70	2.7642	0.3267	0.9358	1.0926	0.0290	0.1270	0.4176
27	5.5	11.08	50.2	18.6	63.6	387	4.9658	0.2830	0.5342	0.6128	0.0066	0.0953	0.1976
28	3.5	8.03	54.2	24.3	87.3	97	4.4305	0.2575	-0.9305	-1.0577	0.0160	0.0789	-0.3100
29	3.4	10.42	58.0	8.0	59.0	119	3.5627	0.2897	-0.1627	-0.1870	0.0006	0.0999	-0.0617
30	4.5	11.46	56.9	15.6	97.7	191	4.9706	0.2513	-0.4706	-0.5339	0.0039	0.0752	-0.1511
31	5.5	7.63	52.1	11.6	61.1	197	3.3920	0.2234	2.1080	2.3713	0.0592	0.0594	0.6276
32	5.0	9.78	52.3	17.6	95.9	270	4.8748	0.1578	0.1252	0.1387	0.0001	0.0296	0.0240
33	4.2	9.06	52.8	6.9	75.9	134	3.4192	0.2012	0.7808	0.8731	0.0064	0.0482	0.1960
34	5.8	11.41	50.4	23.8	73.0	424	5.5965	0.2830	0.2035	0.2334	0.0010	0.0953	0.0750
35	4.3	8.30	57.2	6.8	83.8	167	3.5785	0.2327	0.7215	0.8138	0.0076	0.0645	0.2128
36	5.0	10.33	55.8	21.2	104.3	266	5.3933	0.1988	-0.3933	-0.4396	0.0016	0.0470	-0.0968
37	7.6	11.41	61.1	16.6	97.9	535	6.0469	0.3606	1.5531	1.8429	0.1036	0.1548	0.8095
38	5.3	9.77	50.2	15.7	89.7	154	4.3026	0.2056	0.9974	1.1165	0.0110	0.0503	0.2577
39	2.9	10.79	44.2	2.6	56.6	461	3.8645	0.4663	-0.9645	-1.2222	0.0869	0.2588	-0.7259
40	5.7	11.80	53.8	9.1	116.9	571	5.8576	0.4013	-0.1576	-0.1912	0.0014	0.1917	-0.0922
41	4.4	9.66	52.1	9.9	98.3	83	3.8947	0.2798	0.5053	0.5788	0.0057	0.0932	0.1843
42	3.5	7.94	49.5	6.2	92.3	195	3.5034	0.2752	-0.0034	-0.0039	0.0000	0.0902	-0.0012
43	3.1	9.41	59.5	20.6	91.7	29	4.4481	0.2638	-1.3481	-1.5357	0.0355	0.0828	-0.4683
44	3.9	8.28	49.5	12.0	113.1	546	5.1820	0.4596	-1.2820	-1.6165	0.1463	0.2514	-0.9533
45	6.6	13.95	65.9	15.6	133.5	356	6.5956	0.5255	0.0044	0.0059	0.0000	0.3287	0.0041
46	4.3	9.42	50.6	24.8	62.8	508	5.3784	0.3681	-1.0784	-1.2846	0.0529	0.1612	-0.5671
47	5.5	10.90	57.2	10.6	71.9	593	5.2611	0.3375	0.2389	0.2803	0.0021	0.1356	0.1099
48	1.6	8.82	58.2	3.8	51.7	80	2.7990	0.2835	-1.1990	-1.3754	0.0334	0.0957	-0.4517
49	2.1	8.02	55.0	3.8	46.5	91	2.5350	0.2771	-0.4350	-0.4979	0.0042	0.0914	-0.1567
50	4.2	7.39	51.0	14.6	88.4	72	3.5792	0.2408	0.6208	0.7019	0.0061	0.0690	0.1901
51	4.2	7.53	42.0	23.1	98.9	95	4.1563	0.3786	0.0437	0.0523	0.0001	0.1706	0.0235
52	3.1	8.63	54.0	8.4	56.2	76	3.0164	0.2346	0.0836	0.0944	0.0001	0.0655	0.0247
53	5.3	11.77	54.1	17.3	56.0	196	4.4706	0.3699	0.8294	0.9890	0.0317	0.1628	0.4361
54	4.5	9.44	52.5	10.9	58.5	297	3.9161	0.2035	0.5839	0.6533	0.0037	0.0493	0.[illegible]479

Sólo presentar lo importante y necesario

Observaciones atípicas: consideraremos que hay observaciones atípicas cuando el residual estudentizado r_i, es tal que $ABS(r_i)>3$. De acuerdo a la columna **res.stud** de residuales estudentizados se tiene que no hay observaciones atípicas. 3 pt

Identificación de puntos de balanceo: asumimos que una observación *i* es un punto de balanceo si $h_{ii} > 2p/n$, siendo $2p/n = 2(6/54) \approx 0.22$. Acorde a esto, tenemos 7 puntos de balanceo que corresponden a las observaciones 2,4,14,25,39,44,45. → gráfica, qué causan? 1,5 pt

Identificación de observaciones influencias:

1.Se dice que la observación i será influencial si Di>1. gráfica

2.Una observación será influencial si $|(DFFITS_i)| > 2\sqrt{\frac{6}{54}} \approx 0.666$.

gráfica 2 pt

De acuerdo a la columna Cooks.D de distancias de Cook no hay ninguna observación influencial.
De acuerdo a la columna Dffits de valores DFFITS tenemos que las observaciones 37, 39 y 44 son influenciales.

En conclusión, de acuerdo a los dos criterios tenemos que las observaciones 37, 39 y 44 son influenciales.

causan?

En resumen, para el análisis de observaciones extremas se tiene que:

- No hay valores atípicos.
- Las observaciones 2,4,14,25,39,44,45 son puntos de balanceo.
- Las observaciones 37, 39 y 44 son influenciales.

- **Validez del Modelo:** Opt

Con todos los estudios del modelo realizado anteriormente se logra concluir que el modelo es válido con estas características:

1. la regresión múltiple es válida con una confianza de 0,05
2. Las variables X3,X4,X5 son significativas individualmente en presencia de las demás.
3. Aproximadamente el 61.46% de la variabilidad total de la probabilidad promedio de adquirir infección en el hospital,es explicada por la RLM propuesta.
4. Se concluye que la probabilidad promedio estimada de adquirir infección en el hospital depende de al menos una de las variables asociadas al número de camas (X3), censo promedio diario (X4), así como del número de enfermeras (X5).
5. Los Ei no se distribuyen de forma normal pero si con varianza constante.
6. Se encuentran 0 datos atípicos, 7 puntos de balanceo y 3 puntos influenciales.

Si un supuesto no se cumple el modelo deja de ser válido.