Trabajo 1

3,4

Estudiantes

Salomé Marín Brun
Juan Sebastián Zapata Turizo
José Luis Guerrero Crespo
Valeria Mejía Urrego

Equipo 46

Docente

Carlos Mario Lopera

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín

5 de octubre de 2023

Estadística II - Trabajo 01 **Grupo 46** *Profesor* :

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.

Se analizará una muestra aleatoria de 74 hospitales; a continuación se presenta una vista previa de la base de datos:

Y	X1	X2	X3	X4	X5
3.1	9.41	59.5	20.6	91.7	29
5.2	9.84	53	17.7	72.6	210
4.1	7.13	55.7	9	39.6	279
5.7	11.18	51	18.8	55.9	595
6.4	11.62	53.9	25.5	99.2	133
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2.9	10.79	44.2	2.6	56.6	461
1.6	8.82	58.2	3.8	51.7	80
3.2	8.19	52.1	10.8	59.2	176
4.3	9.42	50.6	24.8	62.8	508
4.1	10.47	53.2	5.7	69.1	196

Donde,

- **Y**: Riesgo de infección. *(Probabilidad promedio estimada de adquirir infección en el hospital en porcentaje).*
- **X1**: Duración de la estadía. *(Duración promedia de la estadía de todos los pacientes en el hospital en días).*
- **X2**: Rutina de cultivos. *(Razón del número de cultivos realizados en pacientes sin síntomas de infección, por cada 100).*
- **X3**: Número de camas. *(Número promedio de camas en el hospital durante el periodo del estudio).*
- **X4**: Censo promedio diario. *(Número promedio de pacientes en el hospital por día durante el periodo del estudio).*
- **X5**: Número de enfermeras. *(Número promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio).*

Ejercicio 1

18pt

Teniendo en cuenta la base de datos y las variables predictoras a estudiar se plantea un modelo de RLM para el problema:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_5 X_{i5} + \varepsilon_i, \quad i = 1, 2, \ldots, 74, \quad donde\ \varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \quad \forall_{i=1,2,\ldots,74}$$

Además se pretende analizar la estimación de cada parámetro, para lo cual se presenta la siguiente tabla obedeciendo a las restricciones y definiciones vistas en clase:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5951679	1.6773172	0.9510234	0.3449606
X1	0.2081023	0.0718301	2.8971467	0.0050623
X2	-0.0249455	0.0316100	-0.7891647	0.4327587
X3	0.0534200	0.0133552	3.9999504	0.0001588
X4	0.0103573	0.0066980	1.5463189	0.1266692
X5	0.0011721	0.0006857	1.7094794	0.0919224

3pt

Teniendo en cuenta esta información, se plantea la ecuación de regresión ajustada que está dada por:

$$\widehat{Y}_i = 1.5952 + 0.2081X_{i1} - 0.0249X_{i2} + 0.0534X_{i3} + 0.0104X_{i4} + 0.0012X_{i5}, \quad i = 1, 2, \ldots, 74$$

Significancia de los parámetros del modelo.

Si tenemos en cuenta el siguiente juego de hipótesis para medir la significancia de los parámetros:

$$\begin{array}{l} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{array} \quad \text{para} \quad j = 0, 1, \ldots, 5.$$

6pt

y tomamos los valores del estadístico de prueba y el valor-P para la prueba de la tabla anterior, se puede concluir a un nivel de significancia de $\alpha = 0.05$ que los parámetros individuales β_1 y β_3 son significativos en presencia de los demás parámetros, por otro lado individualmente $\beta_0, \beta_2, \beta_4$ y β_5 no lo son cuando los demás parámetros del modelos están presentes.

Interpretaciones:

- β_0, primero observemos los valores máximos y mínimos para cada variable.

Tabla 3. Máximos y mínimos del modelo

	Y	X1	X2	X3	X4	X5
Min	1.3	6.70	43.7	1.6	39.6	29
Max	7.8	19.56	65.9	60.5	133.5	835

3pt

como $X_j = 0 \notin [X_{j,\min}, X_{j,\max}]\ \forall_j$, entonces β_0 no es interpretable.

- Teniendo en cuenta la información anterior $\beta_0, \beta_2, \beta_4$ y β_5 no son interpretables al no ser significativos.
- $\widehat{\beta}_1 = 0.2081023$ indica que por cada unidad que aumente la duración de la estadia (X_1) el riesgo de infección aumenta en 0.2081023 unidades, cuando las demás predictoras se mantienen fijas.
- $\widehat{\beta}_3 = 0.0534200$ se puede ver como el aumento en 0.0534200 unidades que tiene el riesgo de infección por cada unidad que aumente el número promedio de camas (X_3), cuando las demas predictoras toman un valor constante.

5 pt

Significancia de la regresión

Para estudiar la significancia de la regresión se plantea:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_5 = 0, \quad \text{vs.}$$
$$H_1: \text{algún } \beta_j \neq 0, j = 1, \ldots, 5.$$

Además se presenta la tabla de análisis de varianza del modelo:

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	79.4605	5	15.89211	15.4126	4.10875e-10
Error	70.1157	68	1.03111		

Teniendo en cuenta el estadístico de prueba $F_0 = 15.4126$ y su valor-P correspondiente $V_p = 4.10875e^{-10}$, como $V_p < \alpha = 0.05$ se rechaza H_0 concluyendo que el modelo es significativo, en otras palabras, el riesgo de infección depende de al menos una de las predictorias del modelo.

Coeficiente de determinación R^2

1 pt

Sabemos que $R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$, de manera que se puede calcular de la tabla ANOVA de la siguiente froma:

¿Qué R^2 es apto?

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{79.4605}{70.1157 + 79.4605} = 0.5312376$$

Eso no dice el R^2

El 53.12% de la variabilidad total en el riesgo de infección es explicado por el modelo de RLM propuesto. Este coeficiente de determinación múltiple suguiere que el modelo nos es el más apto para poder tomar decisiones precisas por lo que en la práctica se debería replantear el modelo ampliando la base de datos o eliminando información redundante e innecesaria.

Ejercicio 2

2,5 pt

esos son parámetros

ojo, no es así.

Se toman los tres variables con menor valor p del punto anterior, con el fin de estudiar la significancia del subconjunto formado con estas variables. Teniendo eso en mente se analizarán el subconjunto formado por las variables β_1, β_3 y β_5, para esto se muestran únicamente las filas de interes de la tabla de todas las regresiones posibles que nos da la información de modelo reducido y del modelo completo.

	GL	R^2	${R^2}_{adj}$	SSE	Cp	Variables
14	2	0.197	0.174	120.108	48.484	X2 X4
16	3	0.511	0.490	73.215	5.005	X1 X3 X5
31	5	0.531	0.497	70.116	6.000	X1 X2 X3 X4 X5

→ innecesario

Se quiere probar que:

$$H_0: \; \beta_1 = \beta_3 = \beta_5 = 0 \quad \text{vs.} \quad H_1: \; \text{Algún } \beta_j \neq 0, \; j = 1, 3, 5.$$

Para esto tenemos que el estadístico de prueba está dado por:

$$
\begin{aligned}
F_0 &= \frac{\text{MSextra}}{\text{MSE}} = \frac{\text{MSR}(\beta_1, \beta_3, \beta_5 | \beta_0, \beta_2, \beta_4)}{\text{MSE}} = \frac{[\text{SSR}(\beta_1, \beta_3, \beta_5 | \beta_0, \beta_2, \beta_4)]/3}{\text{MSE}} \\
&= \frac{[\text{SSE}(\beta_0, \beta_2, \beta_4) - \text{SSE}(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)]/3}{\text{MSE}} \\
&= \frac{[120.108 - 70.116]/3}{1.03111} = 16.1612243
\end{aligned}
$$

1,5pt

Para el criterio de decisión se requiere obtener el valor crítico de una distribución $f_{3,74-6} = f_{3,68}$, a un nivel de significancia $\alpha = 0.05$, esto es, $f_{0.05,3,68} = 2.7395023$.

1pt

Como $F_0 = 16.16122 > f_{0.05,3,68} = 2.7481909$, entonces se rechaza H_0 y se concluye que el conjunto de predictoras individualmente son significativas.

se descartan?

Ejercicio 3

2pt

Supongamos que se quiere estudiar si el efecto de el número de camas y el censo promedio diario es igual, además se quiere observar si la duración de la estadía y la rutina de cultivos presentan similitudes en sus efectos, esto con el fin de analizar la disposición del hospital para afrontar un incremento en los niveles de infección (*caso hipotético*).

Podemos verlo como:

$$H_0: \ \beta_1 = \beta_2, \ \beta_3 = \beta_4 \quad \text{vs.} \quad H_1: \ \beta_1 \neq \beta_2 \text{ ó } \beta_3 \neq \beta_4$$

O equivalentemente,

$$H_o: \begin{cases} \beta_1 - \beta_2 = 0 \\ \beta_3 - \beta_4 = 0 \end{cases}$$

Además, se puede representar matricialmente de la siguiente forma:

$$H_o: \mathbf{L}\underline{\beta} = 0$$

Donde $\mathbf{L} = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix}$ y $\underline{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \end{bmatrix}^T$

2pt

Así el modelo reducido es

$$RM: Y_i = \beta_0 + \beta_1(X_1 + X_2) + \beta_3(X_3 + X_4) + \varepsilon$$

supuestos 0pt

Luego, se tiene que el estadístico de prueba está dado por:

$$F_0 = \frac{\frac{SSE(RM) - SSE(FM)}{\text{g.l}_{SSE(RM)} - \text{g.l}_{SSE(FM)}}}{MSE} = \cancel{\frac{\frac{102.878 - 70.116}{71 - 68}}{1.03111}} = \cancel{10.5911752}$$

0pt

$\sim F \ldots$

Ejercicio 4

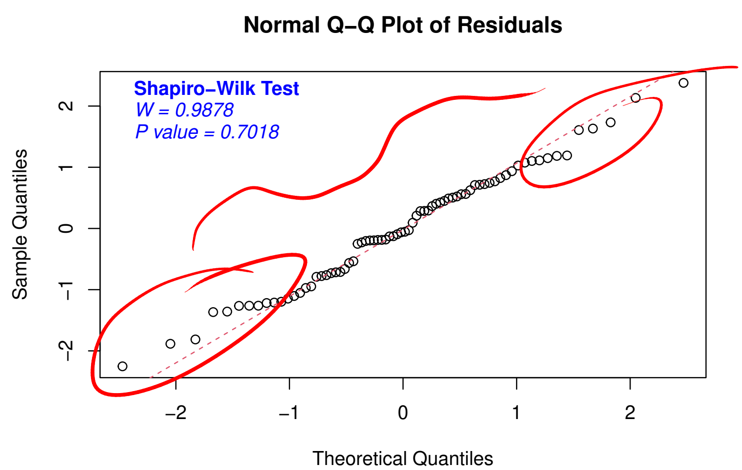
11pt

Validación de los supuestos y observaciones extremas.

Procedemos a validar los supuestos de normalidad y varianza constante de los errores del modelo.

El supuesto de normalidad lo validaremos con la gráfica de normalidad y la prueba de Shapiro-Wilk, en primer lugar se plantean las siguientes hipótesis

$$H_0: \varepsilon_i \sim \text{ Normal. vs. } H_1: \varepsilon_i \not\sim \text{ Normal}$$

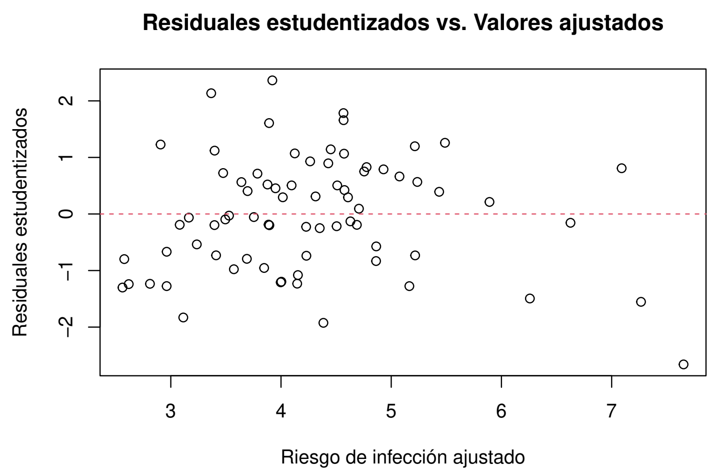


2pt

Analisis muy deficientes, no observan los patrones

Teniendo en cuenta que la mayoria de los valores de los residuales sigue el patrón que indica la línea roja y apoyados en el resultado de la prueba de Shapiro-Wilk ($V_p > \alpha = 0.05$ podemos concluir que el supuesto de normalidad se cumple.

Para el supuesto de varianza constante veamos el gráfico de residuales estudentisados vs valores ajustados



Se quiere probar:

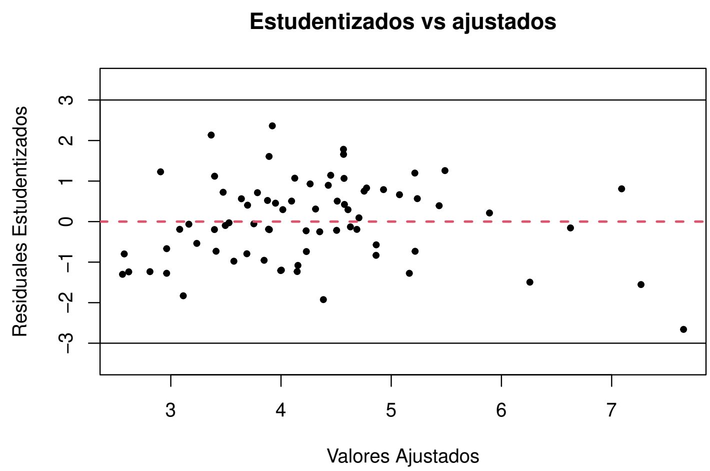
$$H_0: V[\varepsilon_i] = \sigma^2 \quad \text{vs.} \quad H_1: V[\varepsilon_i] \neq \sigma^2$$

3pt

Se observa una nube de puntos en los primeros dos tercios de la gráfica que puede indicar que el supuesto se cumple, pero en el resto de la gráfica se observan valores extremos alejados de la nube, en conclusión, se dice que el supuesto se cumple pero se advierte de la existencia de valores extremos alejados de la nube principal de datos.

Análisis de observaciones extremas.

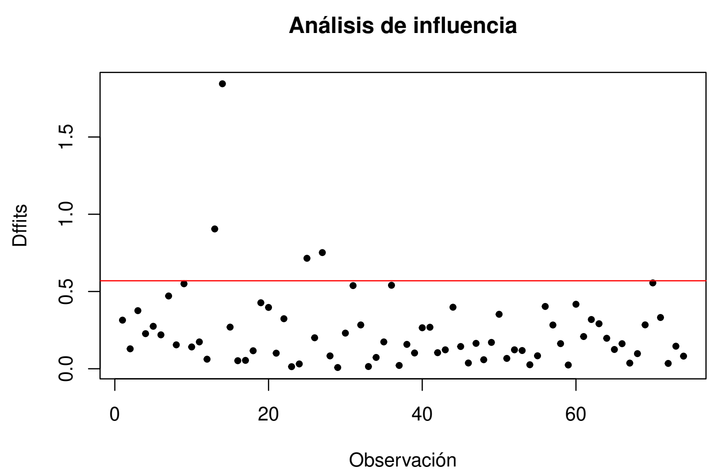
Datos Atípicos.



5pt

Podemos observar que ningún $|r_i| > 3$, por lo tanto se concluye que no hay valores atípicos en los datos.

Observaciones Influenciales.



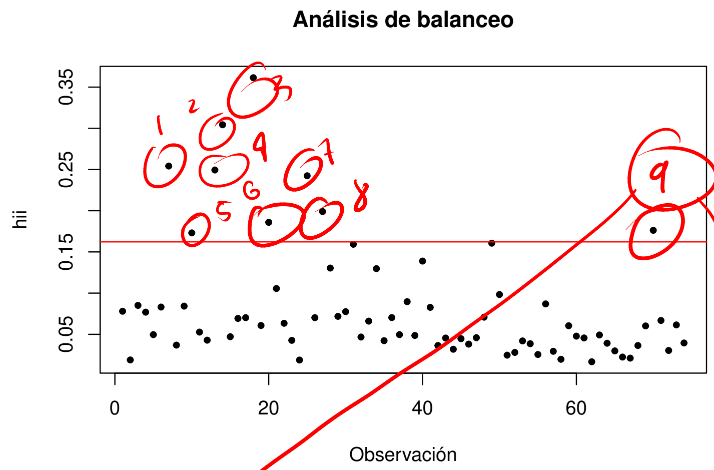
2pt

¿causan...?

¿Dónde demuestran eso?

En primer lugar podemos observar que por el criterio de la distancia de Cook ($D_i > 1$) no se observan datos influenciales, sin embargo, el criterio del diagnóstico DFFITS ($|\mathbf{DFFITS}_i| > 2\sqrt{(\frac{6}{74})} = 0.5694948$) podemos identificar que las observaciones 13, 14, 25 y 27 son influenciales. y se aclara que hay observaciones muy cerca del límite que seria de interés analizar posteriormente.

Puntos de balanceo.



7, 8 o 9 ?

No se deciden...

huh ?

De acuerdo con el Diagnóstico DFBETAS ($h_{ii} > 2 \times \frac{6}{74} = 0.1621622$) se concluye que las observaciones 7, 10, 13, 14, 18, 20, 25, 27, y 70 son puntos de balanceo.

1 2 3 4 5 6 7 8

causan...?

1pt

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

- No se observan datos atípicos.
- Las observaciones 13, 14, 25 y 27 son puntos de balanceo.
- Las observaciones 7, 10, 13, 14, 18, 20, 25, 27, y 70 son influenciales.

Se detecta la presencia de observaciones extremas que deben ser estudiadas antes de usar el modelo como predictor o estimador de valores de la respuesta.

Válido o no? 0pt