Trabajo sin integrantes -1

3,1

Universidad Nacional de Colombia – Sede Medellín
Escuela de Estadística – Semestre 2023-2S

Trabajo Corto 01 – Estadística II - Grupo 11

Objetivo: Usar de manera eficiente las herramientas del análisis de regresión para resolver un problema práctico.

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. A su equipo de trabajo le corresponde analizar una muestra aleatoria de 54 hospitales. Cada base de datos contiene las siguientes columnas (variables):

Variable	***Descripción***
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

Preguntas a resolver. 17pt

1. Estime un modelo de regresión lineal múltiple que explique el riesgo de infección en términos de las variables restantes (actuando como predictoras) Analice la significancia de la regresión y de los parámetros individuales. Interprete los parámetros estimados. Calcule e interprete el coeficiente de determinación múltiple R^2.

¿Modelo teórico y supuestos?

Contruyendo el modelo de regresión lineal múltiple de la variable Riesgo de infección en función de las demás variables, tenemos:

$$\widehat{Y} = -1.8856 + 0.1712X_1 + 0.0299X_2 + 0.0429X_3 + 0.0248X_4 + 0.0010X_5$$

2pt

Para interpretar estos coeficientes, veamos si los parámetros son significativos con ayuda de la siguiente tabla:

Table 2: Tabla de parámetros estimados

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.8855582	1.6351236	-1.153159	0.2545551
X1	0.1712486	0.0988163	1.732998	0.0895169
X2	0.0299898	0.0292446	1.025480	0.3102772
X3	0.0429832	0.0149588	2.873439	0.0060317
X4	0.0247918	0.0084952	2.918346	0.0053416
X5	0.0010103	0.0007551	1.337999	0.1872014

Donde se quiere probar de forma individual:

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

De acuerdo a los valores p de cada una de las pruebas, con un nivel de significancia $\alpha = \mathbf{0.05}$, Los únicos parámetros significativos son β_3 y β_4 por lo tanto solo los efectos de estos parámetros son interpretables:

- $\widehat{\beta}_3 = 0.0429832$, Indica que por cada unidad que aumente el número de camas (X_3) el riesgo de infección aumenta en 0.0429832 unidades, cuando las demás predictoras se mantienen constantes.
- $\widehat{\beta}_4 = 0.0247918$, Indica que por cada unidad que aumente el censo promedio diario (X_4) el riesgo de infección aumenta en 0.0247918 unidades, cuando las demás predictoras se mantienen constantes.
- $\widehat{\beta}_0$ Tampoco puede ser interpretable ya que $0 \notin \mathbf{R}$ de ninguna de las variables.

Ahora, si deseamos ver la significancia de todo el mmodelo, es decir, todas las variables en conjunto, se plantea el siguiente juego de hipótesis:

$$\begin{array}{l} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0, \\ H_1 : \beta_i \neq 0, \text{ para algún i} = 1, \ldots, 5. \end{array}$$

Table 3: Tabla ANOVA.

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	58.1542	5	11.63085	11.8496	1.71594e-07
Error	47.1139	48	0.98154		

De la tabla ANOVA podemos ver que $V_p < 0.05$, por lo tanto se rechaza H_0 y se concluye que el modelo es significativo, es decir que al menos una predictora del modelos es significativamente distinta de cero.

Coeficiente de determinación R^2

$$\mathbf{R^2} = \frac{\text{SSR}}{\text{SST}} = \mathbf{1} - \frac{\text{SSE}}{\text{SST}} = \mathbf{1} - \frac{\mathbf{47.1139}}{\mathbf{47.1139 + 58.1542}} = \mathbf{0.552439}$$

por lo tanto el modelo planteado, explica el 55.24% de la variabilidad total

9 pt

2. Use la tabla de todas las regresiones posibles, para probar la significancia simultánea del subconjunto de tres variables con los valores p más pequeños del punto anterior (β_1, β_3 y β_4). Según el resultado de la prueba este subconjunto de parámetros son todos significativos? Explique su respuesta.

Table 4: Tabla de todas las regresiones posibles.

GL	R^2	$R^2{}_{adj}$	SSE	Cp	Variables
1	0.337	0.324	69.769	21.081	X3
1	0.314	0.301	72.228	23.587	X1
1	0.296	0.282	74.135	25.529	X4
1	0.129	0.112	91.696	43.421	X5
1	0.005	-0.014	104.736	56.706	X2
2	0.458	0.437	57.025	10.098	X1 X3
2	0.457	0.436	57.136	10.211	X1 X4
2	0.446	0.424	58.344	11.441	X3 X4
2	0.390	0.366	64.176	17.383	X4 X5
2	0.390	0.366	64.199	17.407	X3 X5
2	0.361	0.336	67.306	20.572	X2 X3
2	0.323	0.297	71.233	24.573	X1 X5
2	0.314	0.287	72.175	25.532	X1 X2
2	0.304	0.276	73.294	26.672	X2 X4
2	0.140	0.107	90.496	44.198	X2 X5
3	0.530	0.502	49.441	4.371	X1 X3 X4
3	0.499	0.469	52.755	7.747	X3 X4 X5
3	0.474	0.442	55.411	10.453	X1 X4 X5
3	0.466	0.434	56.165	11.222	X2 X3 X4
3	0.465	0.433	56.276	11.334	X1 X3 X5
3	0.463	0.431	56.539	11.602	X1 X2 X3
3	0.457	0.425	57.116	12.190	X1 X2 X4
3	0.419	0.384	61.192	16.343	X2 X3 X5
3	0.405	0.369	62.680	17.858	X2 X4 X5
3	0.323	0.283	71.233	26.572	X1 X2 X5
4	0.543	0.505	48.146	5.052	X1 X3 X4 X5
4	0.536	0.498	48.871	5.790	X1 X2 X3 X4
4	0.524	0.486	50.062	7.003	X2 X3 X4 X5
4	0.475	0.433	55.218	12.257	X1 X2 X4 X5
4	0.473	0.430	55.473	12.517	X1 X2 X3 X5
5	0.552	0.506	47.114	6.000	X1 X2 X3 X4 X5

-7 Sólo datos necesarios

3 pt

Se quiere probar que: $\mathbf{H_0}: \ \beta_1 = \beta_3 = \beta_4 = \mathbf{0}$ vs. $\mathbf{H_1}: \ \beta_j \neq \mathbf{0}$, para algún $\mathbf{i = 1, 3, 4}$.

Para esta prueba, el estadístico de prueba es:

$$F_0 = \frac{[\text{SSE}(\beta_0, \beta_2, \beta_5) - \text{SSE}(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)]/3}{\text{MSE}}$$

$$F_0 = \frac{[90.496 - 47.1139]/3}{0.98154}$$

$$F_0 = 14.732665$$

Podemos observar que $F_0 = 14.73266 > 2.7980606 = f_{0.05,3,48}$, entonces se rechaza H_0 y se concluye que el conjunto de predictoras si son significativas.

1 pt

¿Se descartan o no?

2 pt

3. Plantee una pregunta donde su solución implique el uso exclusivo de una prueba de hipótesis lineal general de la forma $H_0 : L\beta = 0$ (solo se puede usar este procedimiento y no SSextra). Especifique claramente la matriz $\mathbf{L}$, el modelo reducido y la expresión para el estadístico de prueba (no hay que calcularlo).

Planteamiento del problema: Dentro del estudio que se realizó en los hospitales, también es de interés comparar si los efectos entre la duración de la estadía y el censo promedio diario son iguales, y simultáneamente, si se observan diferencias significativas entre el número de camas y enfermeras. Y como interés adicional si la Rutina de cultivos es distinta de cero:

Podemos ver este problema en el siguiente juego de hipótesis:

$$H_0 : \ \beta_1 = \beta_4, \ \beta_3 = \beta_5, \ \beta_2 = 0$$
$$H_1 : \ \beta_1 \neq \beta_4 \text{ ó } \beta_3 \neq \beta_5 \text{ ó } \beta_3 \neq 0$$

También podemos reescribir la hipótesis nula de la forma:

$$H_o : \begin{cases} \beta_1 - \beta_4 = 0 \\ \beta_3 - \beta_5 = 0 \\ \beta_2 = 0 \end{cases}$$

si se analiza de la forma $H_o : \mathbf{L}\underline{\beta} = 0$ se tiene

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

✓ 2 pt

Así el modelo reducido tendría la forma:

$$Y_i = \beta_0 + \beta_1(X_1 + X_4) + \beta_3(X_3 + X_5) + \varepsilon$$

✗ opt supuestos

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

$$F_0 = \frac{\frac{SSE(RM) - SSE(FM)}{51-48}}{MSE} = \frac{\frac{60.965 - 47.1139}{3}}{0.98154} = \frac{60.965 - 47.1139}{2.94462} = 4.7038667$$

0 pt

Table 5: Todas las regresiones posibles - Modelo Reducido

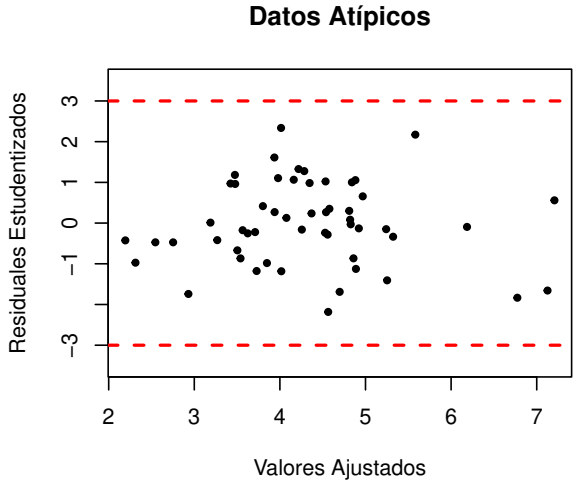
k	R_sq	adj_R_sq	SSE	Cp	Variables_in_model
1	0.335	0.322	70.053	8.603	X14
1	0.147	0.131	89.766	25.094	X35
2	0.421	0.398	60.965	3.000	X14 X35

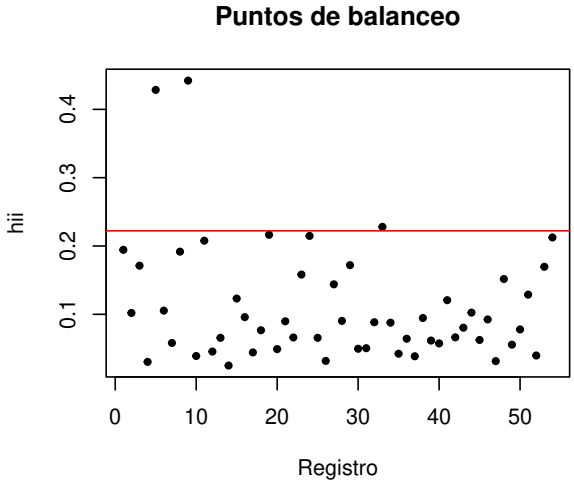
no se calculaba

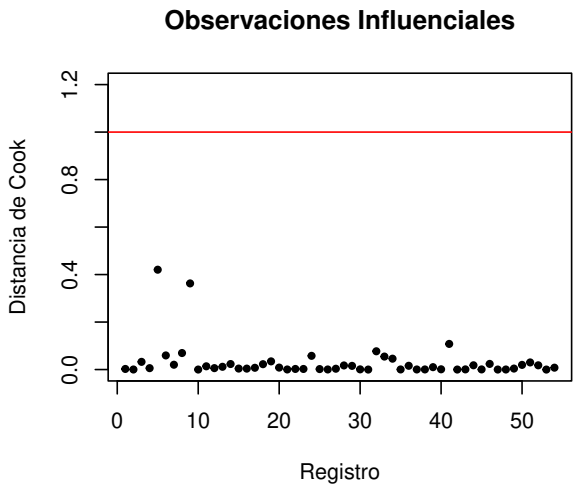
4. Realice una validación de los supuestos en los errores y examine si hay valores atípicos, de balanceo e influenciales. Qué puede decir acerca de la validez de éste modelo?. Argumente su respuesta.

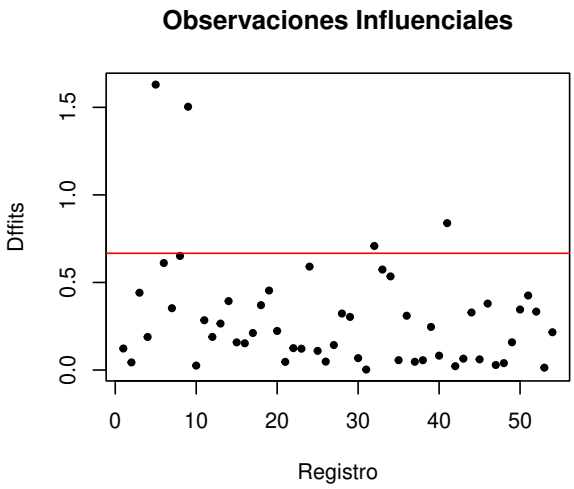
9 pt

4.1 Observaciones extremas.









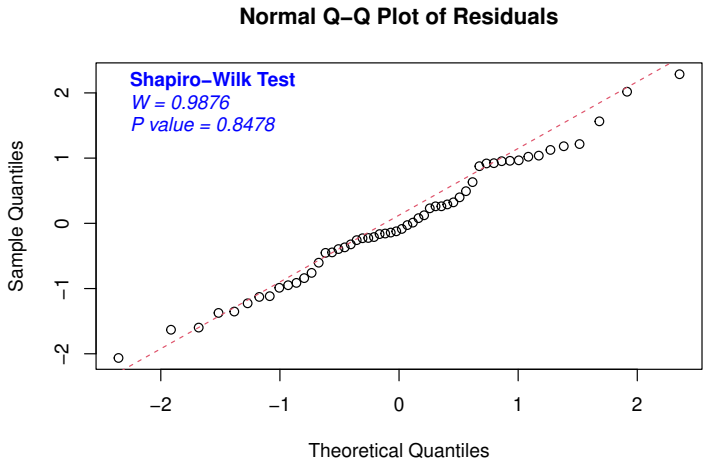
4 por ↗ debajo?

- Un valor se considera atípico si $-3 < r_i < 3$, por lo tanto según el gráfico no hay observaciones atípicas. 3 pt
- De acuerdo con el Diagnóstico DFBETAS ($h_{ii} > 0.2222$) se sabe que las observaciones 5, 9 y 33 se consideran puntos de balanceo. ✗ → error conceptual. 1 pt
- Por el criterio de la distancia de Cook ($D_i > 1$) no se observan datos influenciales 2 pt
- Por el método Dffits ($|\mathbf{DFFITS}_i| > 0.6667$) las observaciones 5, 9, 32 y 41 son influenciales. 1 pt

4.2 Validación de Supuestos.

Dentro de los supuestos del modelo a validar tenemos la normalidad y la varianza constante en los residuales.

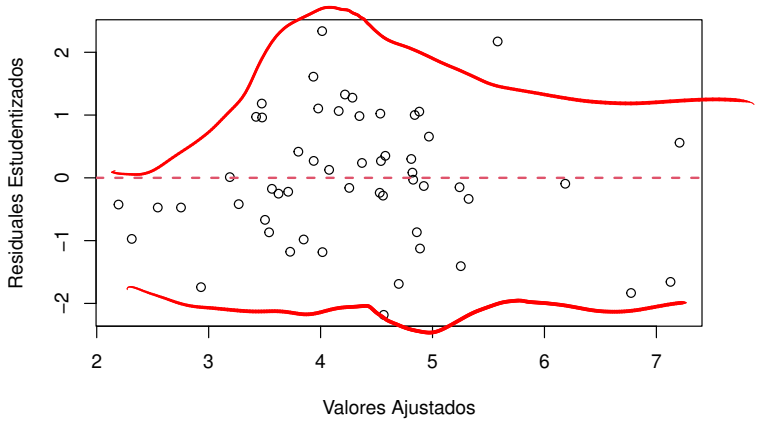
$$H_0: \varepsilon_i \sim \text{Normal vs. } H_1: \varepsilon_i \not\sim \text{Normal}$$



No analizas gráfico, que es más importante.

2pt

Apoyados en la prueba de Shapiro-Wilk como $V_p > \alpha = 0.05$ podemos concluir que el supuesto de normalidad se cumple al no rechazar H_0. Para el supuesto de varianza constante veamos el gráfico de residuales estudentizados vs valores ajustados



No es tan claro ver un patrón como los vistos en clase en la distribucion de los residuales, sin embargo es importante, aclarar que si se ve una agrupación en el centro de las observaciones. Sería interesante ver como se comportan los residuales al extraer las observaciones influenciales y los puntos de balanceo.

8pt

Constante o no?

Válido el modelo o no? 0pt