3,55

Trabajo Corto 01 – Estadística II.

Equipo 10

Andrés Alexis Galvis Herrera

Said Alejandro Durán Rodríguez

Juan Camilo Miranda Paz

Juan José Zapata Cadavid

Universidad Nacional de Colombia

Sede Medellín

Estadística II

2023-1S

Julieth Verónica Guarín Escudero

Preguntas a resolver.

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 R2.

2. Use la tabla de todas las regresiones posibles, para probar la significancia simultánea del subconjunto de tres variables con los valores p más grandes del punto anterior. Según el resultado de la prueba es posible descartar del modelo las variables del subconjunto? Explique su respuesta.

3. Plantee una pregunta donde su solución implique el uso exclusivo de una prueba de hipótesis lineal general de la forma $H0 : L\beta = 0$ (solo se puede usar este procedimiento y no SSextra). Especifique claramente la matriz L, el modelo reducido y la expresión para el estadístico de prueba (no hay que calcularlo).

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.

16pt

Solución.

1. El modelo de regresión lineal múltiple que explica el riesgo de infección en términos de las variables predictoras es: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i$ con

¿?

$$\varepsilon_i \overset{iid}{\simeq} N(0, \sigma^2)\ \forall\, i = 1, 2, \ldots, n$$ ¿Quién es n?

¿Qué formato de tabla van a usar? No copien y peguen cosas.

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.

Anexo 1. Variables del modelo.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercepto	0,8283489490	2,2331691681	0,370929800	0,712469815
X1	0,0762053230	0,13176289550	0,578351900	0,565975215
X2	0,0094780000	0,03247160780	0,291885800	0,771745779
X3	0,0772306790	0,01889796830	4,086718600	0,000182573
X4	0,0092838860	0,00929898590	0,998376200	0,323556102
X5	0,0015235260	0,00075399480	2,020606200	0,049432786

✓

Anexo 2. Resumen del modelo.

¿de dónde a dónde?

El modelo de regresión ajustado es $\widehat{Y}_i = 0.828 + 0.076X_{i1} + 0.009X_{i2} + 0.077X_{i3} + 0.009X_{i4} + 0.001X_{i5}$ ✓ 2pt

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	42,23550	5	8,447091	9,04559	5,69E-06
Error	41,0887	44	0,933835		

Anexo 3. Tabla ANOVA.

Con el siguiente juego de hipótesis → ¿De qué o qué?

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

y considerando un alpha $\alpha = 0.05$ tenemos que: como el valor P = 0,0000056946l es menor a alpha entonces rechazamos H_0, por lo tanto, la regresión es significativa. ✓ 4pt

Con el siguiente juego de hipótesis → ¿De qué? ¿individual?

$$H_0\!: \beta_j = 0\ vs\ H_1\!: \beta_j \neq 0\ para\ j = 1, 2, 3, 4, 5$$

¿y β_0?

		Pr(>|t|)
X1	β_1	0,565975215
X2	β_2	0,771745779
X3	β_3	0,000182573
X4	β_4	0,323556102
X5	β_5	0,049432786

¿β_0?

5pt

considerando un alpha $\alpha = 0.05$, a partir del anexo 2 y observando los diferentes valores P para la prueba de significancia de los parámetros individuales se evidencia que β_3, β_5 son significativos.

$\widehat{\beta_3} =$ 0.077 implica que por cada aumento unitario en el número promedio de camas se incrementa en 0.077 el riesgo de infección promedio, cuando las demás variables predictoras permanecen fijas.

$\widehat{\beta_5} =$ 0.001 implica que por cada aumento unitario en el número promedio de enfermeras se incrementa en 0.001 el ~~riesgo de infección promedio~~, cuando las demás variables predictoras permanecen fijas. → porcentaje promedio del riesgo de infección.

2pt

De la Tabla ANOVA tenemos que:

$$SSR = 42.2355; SSE = 41.0887$$

de esa manera, obtenemos que $R^2 = \frac{SSR}{SST} = \frac{42.2355}{42.2355+41.0887} = 0.5069$

3pt

No peguen salidas de R

~~Multiple R-squared: 0,5069~~

Anexo 4. Coeficiente de determinación múltiple R2.

El 50.69% de la variabilidad total observada en la respuesta es explicada por el modelo. de regresión

2. En el Anexo 2 podemos ver que las tres variables con valores P más grandes son X_1, X_2, X_4; procedemos a probar la significancia simultánea de dicho subconjunto:

2pt

$$H_0: \beta_1 = \beta_2 = \beta_4 = 0 \ vs \ H_1: algún\ \beta_j \neq 0, j = 1, 2, 4.$$
$$A = \{\beta_1, \beta_2, \beta_4\},\ B = \{\beta_0, \beta_3, \beta_5\}.$$

→ ¿para qué esto?

Esto es, $SSR(\beta_1, \beta_2, \beta_4 | \beta_0, \beta_3, \beta_5) = SSR(\beta_0, \beta_1, \beta_2, \beta_3\ \beta_4, \beta_5) - SSR(\beta_0, \beta_3, \beta_5) = SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \beta_1, \beta_2, \beta_3\ \beta_4, \beta_5)$

→ ¿Dónde está? Deben reportarla con número de tabla

A partir de la tabla de todas las regresiones obtenemos que

$$SSE(\beta_0, \beta_3, \beta_5) = 43 \text{ y } SSE(\beta_0, \beta_1, \beta_2, \beta_3\ \beta_4, \beta_5) = 41.089$$

→ 42,505

obteniendo así que

$$SSR(\beta_1, \beta_2, \beta_4 | \beta_0, \beta_3, \beta_5) = \cancel{43} - 41.089 = \cancel{1.914.} = 1,416$$

Los grados de libertad de $SSR(\beta_1, \beta_2, \beta_4 | \beta_0, \beta_3, \beta_5)$ son 3 (tamaño del subconjunto A).

$$F_0 = \frac{MSR(\beta_1, \beta_2, \beta_4 | \beta_0, \beta_3, \beta_5)}{MSE(\beta_0, \beta_1, \beta_2, \beta_3\ \beta_4, \beta_5)} = \cancel{\frac{\frac{1.914}{3}}{0.933}} = \cancel{0.683}$$

0,505

1pt

Con un nivel de significancia $\alpha = 0.05$, n = 50 tenemos que $f_{0.05,3,50-6} = 2.816$

~~=DISTR.F.INV(0,05;3;44) 2,816465817~~

No peguen salidas así.

Como 0.683 $<$ 2.816 entonces no se rechaza H_0 lo que significa que $\beta_1, \beta_2\ y\ \beta_4$ no son significativos para el modelo. y se pueden descartar

1pt

Para qué todo eso, Sólo interesa línea 6 y 31

	k	R_sq	adj_R_sq	SSE	Cp	Variables_in_model
1	1	0,419	0,407	48	6	X3
2	1	0,182	0,165	68	27	X1
3	1	0,16	0,143	70	29	X4
4	1	0,094	0,075	76	35	X5
5	1	0,034	0,014	81	40	X2
6	2	0,49	0,468	43	2	X3 X5
7	2	0,454	0,431	45	5	X1 X3
8	2	0,426	0,402	48	7	X3 X4
9	2	0,419	0,395	48	8	X2 X3
10	2	0,281	0,25	60	20	X1 X4
11	2	0,27	0,239	60,798	21,106	X4 X5
12	2	0,197	0,163	66,886	27,625	X1 X5
13	2	0,191	0,156	67,427	28,204	X1 X2
14	2	0,178	0,143	68,514	29,368	X2 X4
15	2	0,12	0,082	73,351	34,548	X2 X5
16	3	0,503	0,47	41,452	2,389	X3 X4 X5
17	3	0,495	0,462	42,116	3,1	X1 X3 X5
18	3	0,491	0,457	42,444	3,451	X2 X3 X5
19	3	0,46	0,425	44,98	6,166	X1 X3 X4
20	3	0,455	0,42	45,388	6,604	X1 X2 X3
21	3	0,426	0,389	47,819	9,207	X2 X3 X4
22	3	0,315	0,27	57,07	19,114	X1 X4 X5
23	3	0,285	0,238	59,592	21,814	X1 X2 X4
24	3	0,281	0,234	59,881	22,124	X2 X4 X5
25	3	0,206	0,155	66,12	28,805	X1 X2 X5
26	4	0,506	0,462	41,168	4,085	X1 X3 X4 X5
27	4	0,503	0,459	41,401	4,334	X2 X3 X4 X5
28	4	0,496	0,451	42,02	4,997	X1 X2 X3 X5
29	4	0,461	0,413	44,901	8,083	X1 X2 X3 X4
30	4	0,32	0,259	56,685	20,701	X1 X2 X4 X5
31	5	0,507	0,451	41,089	6	X1 X2 X3 X4 X5

Anexo 5. Tabla de todas las regresiones posibles.

3. Se quiere probar: → No es B, es β   3,5 pt

$$H_0: B_3 = B_5; B_1 = B_4 \text{ } vs \text{ } H_1: Alguna\ de\ las\ igualdades\ no\ se\ cumple$$

Matricialmente se representa como:

$$H_0: \mathbb{L}\beta = 0 \text{ } vs \text{ } H_1: \mathbb{L}\beta \neq 0$$ ✓

Con $\mathbb{L}$ dada por:

MF = Modelo full ¿

$$\mathbb{L} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & 0 & -1 & 0 \end{bmatrix}$$ ✓ 2 pt

El modelo reducido (MF) está dado por:

$$Y_i = B_0 + B_1(X_{i1} + X_{i4}) + B_2 X_{i2} + B_3(X_{i3} + X_{i5}) + \varepsilon_i, \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)\ \forall\ i = 1, 2, \dots, n$$ → n = ¿?

0,5 pt

El estadístico de prueba F_0 para la prueba de hipótesis está dado por:

1pt

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{\alpha,2,49}$$

44 ; Sin α

$$F_0 = \frac{(SSE(MR) - 41.0887)/2}{0,933835} \overset{H_0}{\sim} f_{\alpha,2,49}$$

44

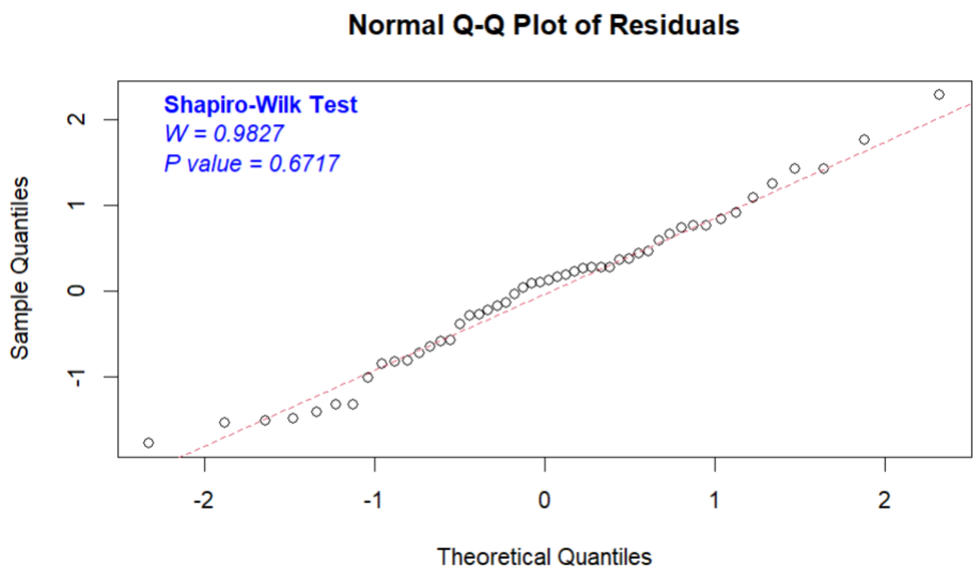
4. Validación de los supuestos sobre los errores 14pt

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

3pt

Se quiere probar:

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



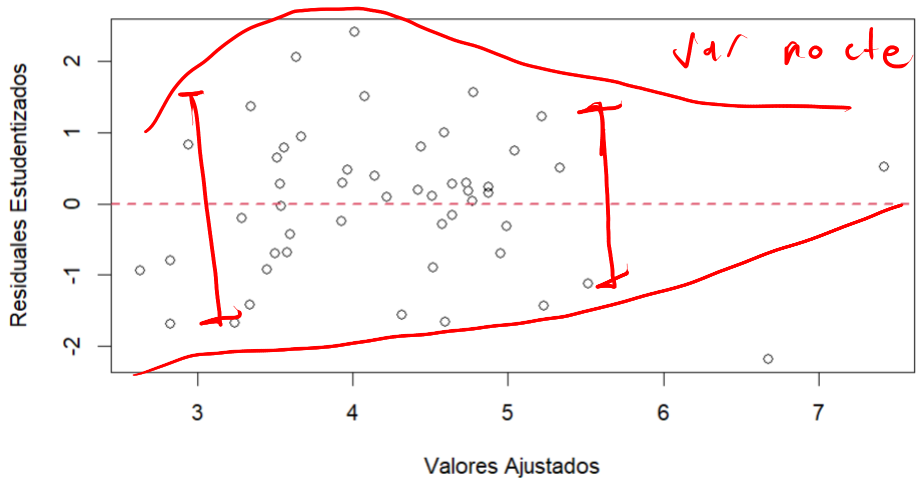
Con el patrón de los residuales (98% de los valores del centro) siguen la línea roja que representan el ajuste de la distribución de los residuales a una distribución normal. Se concluye que el supuesto de normalidad se cumple ya que observando el valor P, como este es mayor a $\alpha = 0.05$, no se rechaza H_0

Análisis muy poco profundo. No hicieron realmente un análisis gráfico de ese patrón

Supuesto de varianza constante - Gráfica de residuales vs. valores ajustados 3pt

Se quiere probar:

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



Observando la gráfica, podemos concluir que no se acepta el supuesto de varianza constante ya que estos presentan una forma más semejante a un polinomio que a una figura uniforme, es decir, podemos observar como la varianza fluctúa bruscamente en la mayoría de los residuales por lo que no se cumple el supuesto de $H_0: V[\varepsilon_i] = \sigma^2$

¿?

meh... Pero sí es var no cte.

Análisis de la presencia de observaciones extremas

Se deben calcular los estadísticos que nos permiten identificar si en el modelo hay observaciones extremas, los cuales incluyen: residuales estudentizados, los valores de la diagonal de la matriz $H(los\ h_{ii})$, la distancia de Cook (Di) y los $DFFITS$.

Residuales estudentizados 3 pt

Para identificar las observaciones atípicas por medio de los residuales estudentizados, debemos cumplir la siguiente restricción: $|r_i| > 3$, debemos observar en los residuales estudentizados aquellos que superen esta cota para poder afirmar si existen observaciones atípicas.

Luego de observar nuestros residuales estudentizados podemos concluir que ninguna de las observaciones es atípica utilizando $|r_i| > 3$ ← esto no es un buen análisis descriptivo, muestren con una gráfica esto, no una tabla con 50 datos

Los valores de la diagonal de la matriz $\boldsymbol{H(los\ h_{ii})}$ 2 pt

Para obtener los puntos de balanceo, tenemos que identificar los valores de la diagonal principal $\boldsymbol{h_{ii}}$ de la matriz H, los cuales cumplan la siguiente condición: $\boldsymbol{h_{ii}} > \frac{\boldsymbol{2p}}{\boldsymbol{n}}$ entonces, aquellos que sean mayores que: $\boldsymbol{h_{ii}} > \frac{\boldsymbol{2*6}}{\boldsymbol{50}} = \boldsymbol{0.24}$ ✓

Observando los valores proporcionados por R, tenemos que las observaciones 17 y 37 son puntos de balanceo ya que cumplen la cota establecida, con valores de $\boldsymbol{h_{ii}}$ 0.4363 y 0.2844 respectivamente. ✓

Lo mismo del punto anterior y los siguientes, tabla más descriptiva y sin tanto dato que satura el trabajo, además, ¿Qué causan estos puntos en el modelo?

La distancia de Cook (Di) 2pt

Para obtener las observaciones influenciales realizamos la distancia de Cook, la cual tiene la siguiente condición: $D_i > 1$

Observando de los valores proporcionados por R, tenemos que ninguna de las observaciones es una observación influencial. según este criterio

Los *DFFITS* 1 pt

Este método al igual que la distancia de Cooks, ayuda a la detección de puntos influenciables, la cual tiene la siguiente condición: $|\boldsymbol{DFFITS}| > 2\sqrt{\frac{p}{n}}$

Entonces, aquellos que sean mayores que $|\boldsymbol{DFFITS}| > 2\sqrt{\frac{6}{50}} = 0.6928$ ✓

Observando de los valores proporcionados por R concluimos que las observaciones 19 y 37 con valores $|\boldsymbol{DFFITS}|$1.0771 y 1.4366 respectivamente, cumplen con la cota para ser puntos influenciales. según este criterio, y ¿qué causan estos puntos?

Filtren y presenten sólo lo que necesita el lector. Además hagan gráficas.

	Y	X1	X2	X3	X4	X5	yhat	se.yhat	residuals	res.stud	Cooks.D	hii.value	Dffits
1	4,5	9,61	52,4	6,9	87,2	487	4,1417	0,274	0,3583	0,3866	0,0022	0,0804	0,1132
2	1,8	7,67	51,7	2,5	40,4	106	2,6325	0,3848	-0,8325	-0,9392	0,0277	0,1586	-0,4072
3	6,2	10,15	51,9	16,4	59,2	568	4,7753	0,3154	1,4247	1,5597	0,0483	0,1065	0,5477
4	4,9	10,23	53,2	9,9	77,9	752	4,7456	0,3636	0,1544	0,1724	0,0008	0,1416	0,0692
5	4,3	8,3	57,2	6,8	83,8	167	3,5606	0,2302	0,7394	0,7878	0,0062	0,0567	0,1924
6	3,1	8,63	54	8,4	56,2	76	3,2841	0,2753	-0,1841	-0,1987	0,0006	0,0812	-0,0584
7	4,5	11,46	56,9	15,6	97,7	191	4,6438	0,3797	-0,1438	-0,1618	0,0008	0,1544	-0,0684
8	4,6	9,68	57,8	16,7	79	186	4,4204	0,2442	0,1796	0,1921	0,0004	0,0639	0,0496
9	4,7	10,72	53,8	23,2	94,1	113	4,9927	0,2922	-0,2927	-0,3178	0,0017	0,0914	-0,0998
10	1,3	8,16	60,9	1,9	58	73	2,8238	0,3408	-1,5238	-1,6851	0,0672	0,1244	-0,6491
11	2,9	7,91	52,8	11,9	79,5	477	4,3154	0,3255	-1,4154	-1,5556	0,0516	0,1135	-0,566
12	1,7	8,09	56,9	7,6	56,9	92	3,2395	0,2826	-1,5395	-1,666	0,0433	0,0855	-0,5204
13	4,5	6,7	48,6	13	80,8	76	3,6695	0,374	0,8305	0,9321	0,0255	0,1498	0,3907
14	4,9	9,89	50,5	17,7	103,6	167	4,6439	0,2515	0,2561	0,2745	0,0009	0,0677	0,0732
15	6,3	9,74	54,4	11,4	76,1	221	4,0098	0,1745	2,2902	2,4095	0,0326	0,0326	0,4693
16	6,3	8,84	56,3	29,6	82,6	85	5,218	0,3841	1,082	1,2202	0,0466	0,158	0,5316
17	7,8	12,07	43,7	52,4	103,3	157	7,426	0,6383	0,374	0,5155	0,0343	0,4363	0,4497
18	4,3	10,39	54,6	14	88,3	353	4,5764	0,2057	-0,2764	-0,2928	0,0007	0,0453	-0,0631
19	5,4	7,93	64,1	7,5	98,1	68	3,6338	0,4362	1,7662	2,0483	0,179	0,2038	1,0771
20	5,1	9,76	50,9	21,9	97	150	4,875	0,2158	0,225	0,2389	0,0005	0,0499	0,0541
21	5	9,78	52,3	17,6	95,9	270	4,7303	0,1844	0,2697	0,2843	0,0005	0,0364	0,0547
22	2	8,93	56	6,2	72,5	95	3,3363	0,2236	-1,3363	-1,4214	0,019	0,0535	-0,3421
23	4,6	10,16	54,2	8,4	51,5	831	4,5092	0,4526	0,0908	0,1063	0,0005	0,2193	0,0557
24	3,1	9,41	59,5	20,6	91,7	29	4,5958	0,3384	-1,4958	-1,6526	0,0636	0,1226	-0,6306
25	5,7	11,18	51	18,8	55,9	595	5,0411	0,3802	0,6589	0,7417	0,0168	0,1548	0,3158
26	4,6	7,84	49,1	7,1	87,9	60	3,347	0,3087	1,253	1,3684	0,0355	0,1021	0,4661
27	5,5	10,9	57,2	10,6	71,9	593	4,5907	0,3249	0,9093	0,9991	0,0212	0,113	0,3566
28	5,8	11,41	50,4	23,8	73	424	5,3373	0,3123	0,4627	0,5059	0,005	0,1045	0,1713
29	4,3	9,89	45,2	11,8	108,7	190	4,2204	0,413	0,0796	0,0911	0,0003	0,1826	0,0426
30	4,4	10,02	49,5	8,3	93	265	3,9692	0,2849	0,4308	0,4665	0,0035	0,0869	0,1427
31	5,2	9,84	53	17,7	72,6	210	4,4415	0,2035	0,7585	0,8029	0,005	0,0444	0,1723
32	4,5	9,31	47,2	30,2	101,3	170	5,517	0,3275	-1,017	-1,1186	0,0271	0,1149	-0,4042
33	4,8	9,84	62,2	12	82,3	600	4,7727	0,4078	0,0273	0,0312	0	0,1781	0,0143
34	4,2	7,39	51	14,6	88,4	72	3,9328	0,2888	0,2672	0,2897	0,0014	0,0893	0,0898
35	5,5	8,37	50,7	15,1	84,8	115	4,0754	0,2053	1,4246	1,5087	0,0179	0,0452	0,3331
36	3,7	7,58	56,7	20,8	88	97	4,5146	0,3262	-0,8146	-0,8955	0,0172	0,1139	-0,3203
37	4,9	11,07	53,2	28,5	122	768	6,6799	0,5153	-1,7799	-2,1774	0,314	0,2844	-1,4366
38	3,5	7,94	49,5	6,2	92,3	195	3,5354	0,3236	-0,0354	-0,0389	0	0,1122	-0,0137
39	4,3	8,67	48,2	24,4	90,8	182	4,9506	0,2802	-0,6506	-0,7034	0,0076	0,0841	-0,2118
40	2,6	9,76	53,2	6,9	80,1	64	3,4504	0,2735	-0,8504	-0,9175	0,0122	0,0801	-0,2703
41	3,7	7,14	59	2,6	75,8	70	2,9428	0,3134	0,7572	0,8283	0,0134	0,1052	0,2829
42	3,8	8,66	52,8	6,8	69,5	246	3,5339	0,184	0,2661	0,2805	0,0005	0,0363	0,0538
43	3	11,2	45	7	78,9	130	3,5795	0,4643	-0,5795	-0,6838	0,0234	0,2308	-0,3723
44	2,1	8,02	55	3,8	46,5	91	2,8246	0,3299	-0,7246	-0,7978	0,014	0,1166	-0,2886
45	3,7	8,58	55	7,4	95,9	304	3,9285	0,2808	-0,2285	-0,2471	0,0009	0,0844	-0,0742
46	3,9	10,73	50,6	19,3	101	445	5,2318	0,2609	-1,3318	-1,4313	0,0268	0,0729	-0,4062
47	4,1	10,47	53,2	5,7	69,1	196	3,5108	0,2985	0,5892	0,6411	0,0072	0,0954	0,2068
48	5	7,78	45,5	20,9	71,6	489	4,8763	0,4722	0,1237	0,1467	0,0011	0,2388	0,0812
49	2,9	10,79	44,2	2,6	56,6	461	3,4981	0,4608	-0,5981	-0,7042	0,0243	0,2274	-0,3798
50	3,2	8,19	52,1	10,8	59,2	176	3,5981	0,2531	-0,3981	-0,4269	0,0022	0,0686	-0,1148

Anexo 6. Tabla de diagnóstico.

No concluyen si el modelo es válido o no: 0pt