Trabajo 1

4,3

Integrantes:
María Angélica Gracia Tamayo
Julian Camilo Hincapie López
Juan Pablo Marín Montoya
Felipe Soto González

Equipo #39

Estadística II

Sede Medellín
30 de marzo 2023

Descripción 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.

Variable Descripción Y: Riesgo de infección Probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje).

X1: Duración de la estadía: Duración promedio 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 hospitalaria, 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.

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.

Resolución de preguntas:

20 pt

PREGUNTA 1
Modelo de regresión:
Teniendo en cuenta la base de datos 39, en la cual encontramos 5 variables regresoras:

Y: Riesgo de infección
X_1: Duración de la estadía
X_2: Rutina de cultivos
X_3: Número de camas
X_4: Censo promedio diario
X_5: Número de enfermeras

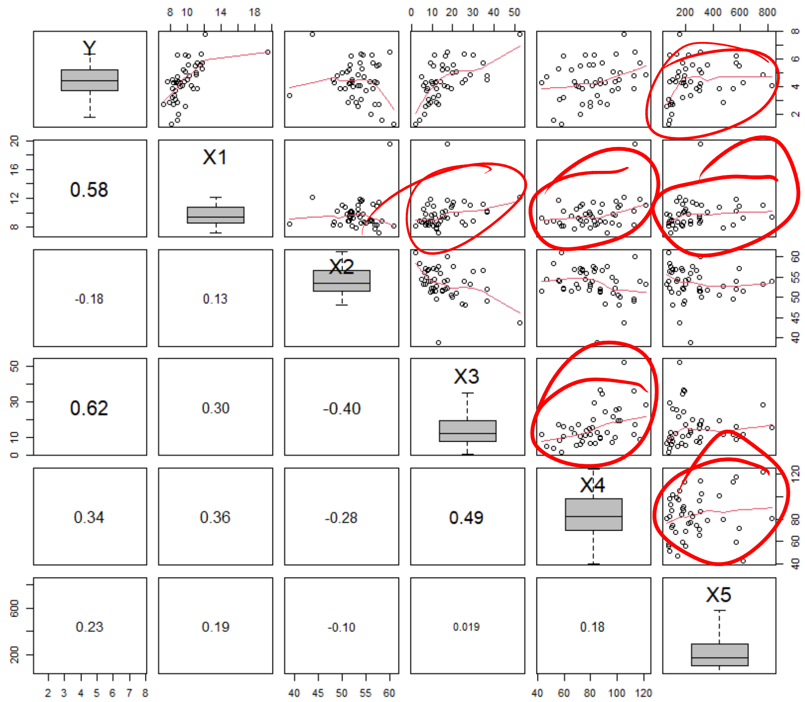
Entonces, se plantea el modelo de regresión lineal múltiple:

$$Y_i = \beta_0 + \beta_1(X_{1i}) + \beta_2(X_{2i}) + \beta_3(X_{3i}) + \beta_4(X_{4i}) + \beta_5(X_{5i}) + \varepsilon_i$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 45$$

Se verificará que no haya ningún tipo de colinealidad entre las variables regresoras para poder determinar un modelo de regresión ajustado que no contenga información redundante.

Matriz de gráficas de dispersión con boxplots y correlaciones de las variables
Figura 1



No hay ningún tipo de relación fuerte entre alguna de las variables regresoras por lo que no tenemos algún problema de colinealidad. Sí se ven.

En R obtuvimos los siguientes coeficientes al ajustar el modelo:

Tabla de parámetros estimados
Figura 2

Valor del parámetro	
β_0	1.44363
β_1	0.31207
β_2	-0.01503
β_3	0.06998
β_4	-0.00851
β_5	0.00108

3 pt

El modelo de regresión ajustado sería:

$$\widehat{Y}_i = 1.44363 + 0.31207(X_{1i}) - 0.01503(X_{2i}) + 0.06998(X_{3i}) - 0.00851(X_{4i}) + 0.00108(X_{5i})$$

5 pt

Significancia de la regresión:
Se plantean las siguientes hipótesis para realizar un análisis sobre la significancia de la regresión:

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

$H_a: Algún\ \beta_j \neq 0;\ j = 1, 2, 3, 4, 5$

Utilizando el estadístico de prueba $F_0 = \frac{MSR}{MSE} \sim f_{5,39}$

Tabla ANOVA del modelo
Figura 3

	Sum of Squares	DF	Mean Square	F_0	Valor P
Model	48.3999	5	9.679975	10.8389	1.39296e-06
Error	34.8299	39	0.893074		

En la Figura 3 observamos un valor P cercano a 0, con esto rechazamos H_0 y aceptamos H_a en la que existe algún $\beta_j \neq 0;\ j = 1, 2, 3, 4, 5$. Concluimos que la regresión es significativa.

Significancia de los parámetros 6 pt

Figura 4

	Estimate	Std. Error	t-value	Valor P
β_0	1.44363	2.35678	0.61254	0.54373
β_1	0.31207	0.08584	3.63546	0.00080
β_2	-0.01503	0.04187	-0.35906	0.72148
β_3	0.06998	0.01762	3.97088	0.00029
β_4	-0.00851	0.00896	-0.9504	0.34772
β_5	0.00108	0.00076	1.42119	0.16320

En la Figura 4 utilizando un nivel de significancia $\alpha = 0.05$, observamos que los parámetros β_1 y β_3 son significativos ya que sus P-valores son menores a α.

Interpretando los parámetros: 3 pt

- Por cada día que aumenta la estadía de los pacientes en el hospital, aumenta en un 31.207% (valor de $\beta_1 = 0.31207$ de *Figura 4*) la probabilidad (promedio) de que el paciente adquiera una infección en el hospital, mientras las otras variables permanecen constantes.
- Por cada cama que aumenta en el hospital, aumenta en un 6.998% (valor de $\beta_3 = 0.06998$ de *Figura 4*) la probabilidad (promedio) de que el paciente adquiera una infección en el hospital, mientras las otras variables permanecen constantes.

Coeficiente R^2: 3 pt

Sabemos que $R^2 = \frac{SSR}{SST}$ y $SST = SSR + SSE$, utilizando los valores obtenidos para el SSR y SST en la Figura 3, obtenemos un $R^2 = 0.58152$ Con esto podemos decir que aproximadamente el 58.152% de la variabilidad total observada en ~~los datos observados~~ la variable respuesta es explicada por el modelo de regresión.

PREGUNTA 2

Pruebas de hipótesis y modelo reducido: 5 pt

Las covariables con el valor-p más alto en el modelo planteado en la pregunta 1 fueron X_2, X_4, X_5, por lo tanto, a través de la tabla de todas las regresiones posibles, se plantea la siguiente prueba de hipótesis:

$H_0: \beta_2 = \beta_4 = \beta_5 = 0$
$H_a: Algún\ \beta_j \neq 0\ para\ j = 2, 4, 5$

Tabla posibles regresiones
Figura 5

	SSE	Covariables en el modelo
Modelo completo	34.830	X_1, X_2, X_3, X_4, X_5
Modelo reducido	37.344	X_1, X_3

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

$Y_i = \beta_0 + \beta_1(X_{1i}) + \beta_3(X_{3i}) + \varepsilon_i; \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 45$
Modelo reducido:
$\widehat{Y}_i = 1.44363 + 0.31207(X_{1i}) + 0.06998(X_{3i})$

Estadístico de prueba y conclusión :
Se construye el estadístico de prueba:

3 pt

$$F_0 = \frac{(SSE(\beta_0, \beta_1, \beta_3) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5))/3}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} \sim f_{3,39}$$

2 pt

$$F_0 = \frac{(37.344 - 34.830)/3}{(34.830/39)} = ¿?$$

$F_{3,39}$

El F_0 es igual a 0.93832 y el $\boxed{f}$ es igual a 2.845068, por lo que no rechazamos la hipótesis nula al ser F_0 menor que f, entonces concluimos que se descartan las variables del subconjunto debido a que no son significativas, y el modelo reducido logra explicar el riesgo de adquirir la infección en el hospital, mediante la precesncia de alguna de las covariables X_1 que es la duración promedio de la estadía de todos los pacientes en el hospital (en días), y X_3 que representa el número promedio de camas en el hospital durante el periodo del estudio.

PREGUNTA 3 3,5 pt
¿El número de camas es igual al número de enfermeras?

Se comparan los efectos, no las covariables.

¿El tiempo de estancia es igual al censo promedio diario? X

Para resolver esto, se plantea la siguiente prueba de hipótesis:

$H_0 : \boldsymbol{\beta}_3 = \boldsymbol{\beta}_5 , \boldsymbol{\beta}_1 = \boldsymbol{\beta}_4$

$H_1 : \boxed{\boldsymbol{\beta}_3 \neq \boldsymbol{\beta}_5 , \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_4}$ ó

Cuando se escribe la prueba de hipótesis de forma matricial:

$H_0 : L\boldsymbol{\beta} \ = 0$

$H_1 : L\boldsymbol{\beta} \ \neq 0$

Con L igual a: 1,5 pt

$$L = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & 0 & -1 & 0 \end{pmatrix}$$

Con el siguiente modelo reducido:

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1(X_{1i} + X_{4i}) + \boldsymbol{\beta}_2 X_{2i} + \boldsymbol{\beta}_3(X_{3i} + X_{5i}) + \varepsilon_i ; \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 45$$ 1 pt

Estadístico de prueba:

El estadístico de prueba dado para esta hipótesis es:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \sim f_{3,39}$$ $F_{2,39}$ 1 pt

Mediante este cálculo se puede comprobar la prueba de hipótesis lineal general planteada para la solución de las preguntas. Reemplazar lo que conocen

14,5

PREGUNTA 4

Para la validación de los supuestos del modelo:

- **Se probará la normalidad:**

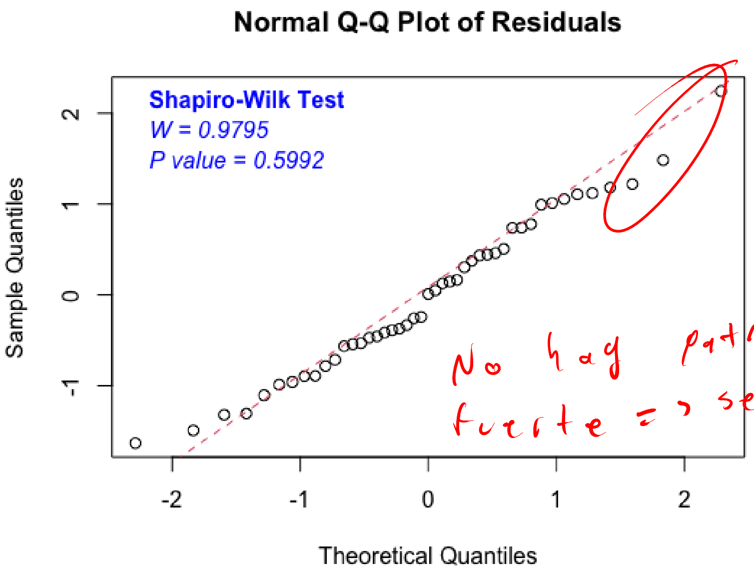
  Para la validación de este supuesto, se planteará la siguiente prueba de hipótesis ~~shapiro-wilk~~, acompañada de un gráfico cuantil-cuantil:

  $H_0 : \varepsilon_i \sim \text{Normal}$

  $H_1 : \varepsilon_i \nsim \text{Normal}$

  Gráfica y prueba de normalidad de Shapiro-Wilk 2 pt

  Figura 6



cola pesada

No hay patrón irregular fuerte => se puede considerar normal

En la Figura 6 vemos que mediante este test se concluye que como el valor-p es igual a 0.5992 y como el nivel de significancia es α = 0.05, el valor p es mucho más grande, por lo que no rechazamos la hipótesis nula y decimos que los datos se distribuyen normal.

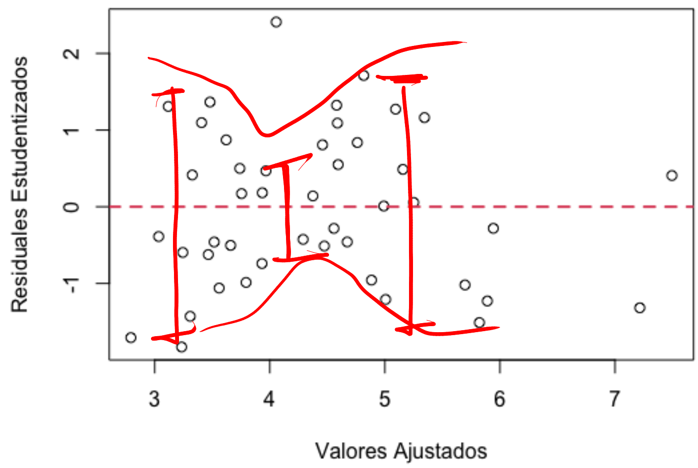
No hacen análisis gráfico, que es más importante

- **Se probará media = 0 y varianza constante:**
  Para esto analizaremos la siguiente gráfica:

Gráfica de Residuales Estudentizados vs. Valores ajustados

3 pt

Figura 7



Con la Figura 7 podemos concluir que la media se aproxima a 0 y de la varianza podemos decir que no es constante, porque se da una forma aproximadamente cónica.

- Se probará si hay datos atípicos:

Para esto se utilizará el la siguiente tabla de diagnosticos:

Tabla de diagnósticos de las observaciones

Para qué ?

Figura 8

Saturan el reporte con esa tabla, sólo muestren lo que en concepto del reporte interesa.

	Y	X1	X2	X3	X4	X5	yhat	se.yhat	residual	res.stud	Cooks.D	hii.value	Dffits
1	5	11.03	49.9	19.7	102.1	318	4.9911	0.237	0.0089	0.0097	0	0.0631	0.0025
2	6.3	9.74	54.4	11.4	76.1	221	4.0557	0.161	2.2443	2.4098	0.0288	0.0289	0.4444
3	4.2	8.88	51.5	10.1	86.9	305	3.7395	0.209	0.4605	0.4997	0.0021	0.0491	0.1125
4	5.1	9.76	50.9	21.9	97	150	4.594	0.216	0.506	0.5499	0.0028	0.0521	0.1277
5	2.7	7.14	57.6	13.1	92.6	92	3.0341	0.393	-0.334	-0.389	0.0053	0.173	-0.176
6	6.2	10.15	51.9	16.4	59.2	568	5.0933	0.37	1.1067	1.2726	0.0488	0.1531	0.5456
7	5.2	9.53	51.5	15	65.7	298	4.4582	0.223	0.7418	0.8078	0.0064	0.0558	0.1955
8	4.7	8.77	54.5	5.2	47	143	3.4805	0.309	1.2195	1.3657	0.0373	0.1071	0.4785
9	2.9	8.86	51.3	9.5	87.5	100	3.4658	0.265	-0.566	-0.624	0.0055	0.0787	-0.181
10	2	8.93	56	6.2	72.5	95	3.3083	0.235	-1.308	-1.429	0.0224	0.0618	-0.372
11	3.9	8.28	49.5	12	113.1	546	3.7548	0.418	0.1452	0.1714	0.0012	0.1958	0.0835
12	3.9	11.15	56.5	7.7	73.9	281	4.2894	0.237	-0.389	-0.426	0.002	0.0626	-0.109
13	4.4	9.66	52.1	9.9	98.3	83	3.6209	0.309	0.7791	0.8722	0.0151	0.1066	0.3004
14	4	9.2	52.2	17.5	71.1	298	4.4737	0.19	-0.474	-0.512	0.0018	0.0404	-0.104
15	3.7	8.58	55	7.4	95.9	304	3.3266	0.288	0.3734	0.4148	0.0029	0.0926	0.1311
16	7	19.6	60	17	114	306	7.216	0.773	-0.716	-1.318	0.5872	0.6697	-1.896
17	6	10.9	57	11	71.9	593	4.761	0.338	0.7392	0.8376	0.0171	0.1278	0.3194
18	4	9.35	54	16	80.9	833	4.884	0.471	-0.784	-0.957	0.0503	0.2481	-0.549
19	1	8.16	61	1.9	58	73	2.793	0.358	-1.493	-1.707	0.0814	0.1435	-0.717
20	6	11.5	58	20	82	252	5.157	0.258	0.443	0.4873	0.0032	0.0746	0.137
21	6	11.8	54	9.1	117	571	4.581	0.423	1.1194	1.3243	0.0731	0.1999	0.6687
22	4	8.3	57	6.8	83.8	167	3.118	0.277	1.1821	1.3084	0.0269	0.086	0.4052
23	3	8.34	57	8.1	74	107	3.244	0.247	-0.544	-0.596	0.0043	0.0683	-0.16
24	6	10.1	52	15	79.1	362	4.588	0.181	1.012	1.0912	0.0076	0.0367	0.2137
25	5	8.28	48	26	102	108	4.375	0.319	0.1254	0.141	0.0004	0.1142	0.05
26	4	8.67	48	24	90.8	182	4.557	0.257	-0.257	-0.283	0.0011	0.0738	-0.079
27	4	9.23	52	12	42.6	620	4.673	0.482	-0.373	-0.459	0.0124	0.2604	-0.269
28	4	10.7	51	19	101	445	5.007	0.234	-1.107	-1.209	0.016	0.0615	-0.311
29	5	11.8	54	17	56	196	5.251	0.356	0.0494	0.0564	0.0001	0.142	0.0227
30	3	8.19	52	11	59.2	176	3.66	0.246	-0.46	-0.504	0.0031	0.0679	-0.135
31	2	8.82	58	3.8	51.7	80	3.234	0.309	-1.634	-1.829	0.0665	0.1065	-0.652
32	6	11.2	57	35	88.9	180	5.943	0.388	-0.243	-0.282	0.0027	0.1685	-0.125
33	3	9.76	53	6.9	80.1	64	3.56	0.272	-0.96	-1.061	0.0169	0.0829	-0.319
34	5	11.1	53	29	122	768	5.891	0.493	-0.991	-1.229	0.0942	0.2723	-0.757
35	4	10.5	53	5.7	69.1	196	3.935	0.268	0.1649	0.182	0.0005	0.0805	0.0532
36	6	8.84	56	30	82.6	85	4.816	0.378	1.4836	1.7129	0.0932	0.1601	0.7675
37	5	10.1	52	37	87.5	184	5.822	0.357	-1.322	-1.511	0.0634	0.1429	-0.628
38	6	11.6	54	26	99.2	133	5.344	0.261	1.0559	1.1627	0.0187	0.0765	0.3362
39	5	10.2	49	36	113	195	5.696	0.347	-0.896	-1.02	0.027	0.1347	-0.403
40	8	12.1	44	52	105	157	7.495	0.576	0.3053	0.4074	0.0163	0.3712	0.3096
41	3	8.45	39	13	85	235	3.932	0.614	-0.532	-0.741	0.0669	0.4225	-0.63
42	4	7.7	57	12	67.9	129	3.407	0.272	0.9929	1.0971	0.0181	0.0829	0.3307
43	3	8.63	54	8.4	56.2	76	3.517	0.268	-0.417	-0.46	0.0031	0.0801	-0.134
44	3	7.91	53	12	79.5	477	3.794	0.271	-0.894	-0.987	0.0146	0.0825	-0.296
45	4	8.88	56	14	76.8	237	3.967	0.191	0.4333	0.4682	0.0016	0.041	0.0958

Observaciones atípicas: 3pt

Recordamos que una observación es atípica cuando su residual estudentizado r_i cumple: $|r_i| > 3$. En la columna "res.stud" podemos observar que no existen valores que cumplan esta condición, esto podemos decir que no hay observaciones atípicas.

Puntos de balanceo: 2 pt
Consideramos que una observación i es un punto de balanceo si:

Se cumple que $h_{ii} > 2p/n$.

Evaluando obtenemos que $h_{ii} > 2(6)/(45) = h_{ii} > 0.266667$, por lo que se cumple que $2p/n < 1$ para poder que el criterio sea válido.

Al observar la columna “hii.value” notamos que las observaciones **16, 34, 40** y **41** son puntos de balanceo.

¿Qué causan?

Identificación de observaciones influenciales:
Los siguientes criterios nos ayudan a identificar las observaciones influenciales:

Diagnóstico de Cook:

Se dice que la observación i es un punto influencial si:

$D_i > 1$

Observando la columna “Cooks.D” vemos que no existen observaciones que cumplan este criterio. 2 pt

Diagnóstico DFFITS:

Se dice que la observación i es un punto influencial si:

$$\left|DFFITS_i\right| > 2\sqrt{\frac{p}{n}}$$

Teniendo p = 6, n = 45, se tiene que i es influencial si:

$$\left|DFFITS_i\right| > 0.7303$$

Observando la columna “Dffits” vemos que las observaciones **16, 34** y **36** son influenciales.

¿Qué causan? 1 pt

Se puede concluir que:

- No hay observaciones atípicas teniendo en cuenta el criterio de los residuales estudentizados.
- Las observaciones **16, 34** y **35** son influenciales tomando el criterio del diagnóstico DFFITS.
- Las observaciones **16, 34, 40** y **41** son puntos de balanceo por el criterio de los hii.value.

Conclusiones 1,5 pt

- La base de datos tomada resulta ser útil para el análisis, pues no se muestran datos atípicos y se muestra una relación entre las variables regresoras y la variable descripción, sin tener problemas de colinealidad entre las regresoras.
- Debido a una varianza no constante de los errores para un mejor ajuste del modelo se recomienda hacer como medida remedial transformaciones en la variable respuesta.

¿Según qué? No han dicho si quiera si es válido.

Así, el modelo ajustado es bueno para el tratamiento de la variable de descripción, que responde según las variables regresoras del fenómeno estudiado sobre la eficacia en el control de infecciones hospitalarias. Mientras no se hagan estos ajustes el modelo queda desestimado por no cumplir con el supuesto de varianza constante de los errores.

¿Qué se supone que implica eso?

- Teniendo en cuenta que las variables duración de la estadía y número de camas están relacionadas con el riesgo epidemiológico de contraer la infección, podemos decir que cada día que aumenta la estadía de un paciente, y cada que aumenta una cama, se hace mayor la exposición a la infección y por tanto la posibilidad de contraerla. ✓

No dijeron si era válido a pesar que hablaron del cumplimiento de supuestos