# ESTADISTICA II

3,3

## PRIMER TRABAJO

**PRESENTADO POR :**


**JUAN SIMÒN ZAPATA MONSALVE**
**HAROLD JHOAN SANCHEZ VIVAS**
**DIEGO ANDRES CASTAÑEDA URANGO**
**JHON ALEXANDER VALENZUELA**


**DOCENTE:**
**JULIETH VERONICA GUARIN ESCUDERO**



**INSTITUTO DE EDUCACIÒN EN INGENIERÍA**
**FACULTAD DE MINAS – SEDE MEDELLÍN**
**UNIVERSIDAD NACIONAL DE COLOMBIA**

**2023**

**Grupo 17 - Solución primer trabajo**

De una total de 113 hospitales en estados unidos se tiene una muestra aleatoria de 64 hospitales, se muestra a continuación los primeros 10 registros

Table 1: Primeras observaciones base de datos

| Y | X1 | X2 | X3 | X4 | X5 |
|---|---|---|---|---|---|
| 2.9 | 10.80 | 63.9 | 1.6 | 57.4 | 130 |
| 3.9 | 11.15 | 56.5 | 7.7 | 73.9 | 281 |
| 1.4 | 7.14 | 51.7 | 4.1 | 45.7 | 115 |
| 5.8 | 11.41 | 50.4 | 23.8 | 73.0 | 424 |
| 4.8 | 9.84 | 62.2 | 12.0 | 82.3 | 600 |
| 5.2 | 9.84 | 53.0 | 17.7 | 72.6 | 210 |
| 4.3 | 7.65 | 47.1 | 16.4 | 65.7 | 318 |
| 3.7 | 7.58 | 56.7 | 20.8 | 88.0 | 97 |
| 4.7 | 8.77 | 54.5 | 5.2 | 47.0 | 143 |
| 3.0 | 11.20 | 45.0 | 7.0 | 78.9 | 130 |

Tenenmos como variable de interés o variable respuesta $\boldsymbol{Y}$ Riesgo de infección, y como variables predictoras $\mathbf{X_1}$ Duración de la estadía, $\mathbf{X_2}$ Rutina de cultivos, $\mathbf{X_3}$ Número de camas, $\mathbf{X_4}$ Censo promedio diario y $\mathbf{X_5}$ Número de enfermeras.

***Punto 1.*** Regresión Lineal Múltiple: 15 pt

Con la información anterior se quiere modelar a $\boldsymbol{Y}$ en función de las demás variables, de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

supuestos.

planteando el modelo tenemos la siguiente tabla que resume la información para cada parámetro:

Table 2: Resúmen del modelo

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -0.1650397 | 1.5436881 | -0.1069126 | 0.9152273 |
| X1 | 0.1424980 | 0.0885128 | 1.6099136 | 0.1128465 |
| X2 | 0.0103015 | 0.0275102 | 0.3744615 | 0.7094266 |
| X3 | 0.0469583 | 0.0138310 | 3.3951331 | 0.0012433 |
| X4 | 0.0194602 | 0.0076270 | 2.5514789 | 0.0133855 |
| X5 | 0.0008251 | 0.0006345 | 1.3004436 | 0.1985927 |

✓ 2pt

Así, el modelo con los coeficientes estimados es:

$$\widehat{Y} = -0.1650397 + 0.1424980 X_1 + 0.0103015 X_2 + 0.0469583 X_3 + 0.0194602 X_4 + 0.0008251 X_5$$

✓

## Análisis de significancia de los parámetros.

Deseamos ver que parámetros son significativos individualmente en el modelo, así que se quiere probar el siguiente juego de hipótesis:

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

Teniendo en cuenta los valores-p de la primer tabla, podemos notar que las únicas variables significativas dentro del modelo son $\beta_3$ y $\beta_4$ son significativas cuando los demás parámetros están presentes.

- Sabiendo que $\beta_1$, $\beta_2$ y $\beta_5$ no son significativos, tampoco serán interpretables.
- $\widehat{\beta}_3 = 0.0469583$, El riesgo de infección aumentará en 0.0469583 unidades por cada unidad que aumente el número promedio de camas.
- $\widehat{\beta}_4 = 0.0194602$, El riesgo de infección aumentará en 0.0194602 unidades por cada unidad que aumente el número promedio de pacientes.
- $\widehat{\beta}_0$ : Dado que ninguna variable tiene al cero dentro de su dominio, $\beta_0$ no es interpretable.

## Significancia del modelo de regresión

Para estudiar la significancia de la regresión se plantea:

$$\begin{array}{ll} H_0 : \beta_1 = \beta_2 = \cdots = \beta_5 = 0, & \text{vs.} \\ H_1 : \text{Algún } \beta_j \neq 0, j = 1, \ldots, 5. & \end{array}$$

Si tomamos en cuenta la información de la tabla ANOVA con ayuda de la función myAnova() se obtiene:

Table 3: Tabla ANOVA.

| | Sum_of_Squares | DF | Mean_Square | F_Value | P_value |
|---|---|---|---|---|---|
| Model | 54.6752 | 5 | 10.935038 | 12.6923 | 2.42589e-08 |
| Error | 49.9697 | 58 | 0.861546 | | |

Como $V_p = 2.42589e^{-08} < \alpha = 0.05$ entonces se rechaza $H_0$ y se puede concluir que el modelo es significativo, por lo tanto, el riesgo de infección depende de al menos una de las cinco variables predictoras.

## Interpretación del $R^2$:

Del modelo obtenemos que $\mathbf{R^2 = 0.5225}$, por lo tanto, el modelo planteado explica el 52.25% de la variabilidad total.

***Punto 2.*** Todas las regresiones posibles.

Para este punto se eligen las tres variables con el menor valor-p, por lo tanto se estudiará la significancia de los parámetros $\beta_1$, $\beta_3$ y $\beta_4$.Por lo tanto se quiere probar que:

$$H_0: \ \beta_1 = \beta_3 = \beta_4 = 0$$
$$H_1: \ \beta_j \neq 0, \ \text{para algún } j = 1, 3, 4.$$

Observando la tabla de todas las regresiones posibles, tenemos como filas de interés:

Table 4: Todas las regresiones posibles.

| GL | $R^2$ | $R^2{}_{adj}$ | SSE | Cp | Variables |
|---|---|---|---|---|---|
| 1 | 0.370 | 0.359 | 65.970 | 16.572 | X3 |
| 1 | 0.288 | 0.277 | 74.495 | 26.466 | X4 |
| 1 | 0.262 | 0.251 | 77.186 | 29.590 | X1 |
| 1 | 0.145 | 0.131 | 89.458 | 43.834 | X5 |
| 1 | 0.007 | -0.009 | 103.891 | 60.587 | X2 |
| 2 | 0.459 | 0.441 | 56.606 | 7.703 | X1 X3 |
| 2 | 0.453 | 0.435 | 57.251 | 8.452 | X3 X4 |
| 2 | 0.415 | 0.396 | 61.168 | 12.998 | X3 X5 |
| 2 | 0.399 | 0.379 | 62.920 | 15.032 | X1 X4 |
| 2 | 0.376 | 0.355 | 65.333 | 17.833 | X2 X3 |
| 2 | 0.375 | 0.354 | 65.405 | 17.916 | X4 X5 |
| 2 | 0.288 | 0.265 | 74.484 | 28.454 | X1 X2 |
| 2 | 0.288 | 0.265 | 74.493 | 28.465 | X2 X4 |
| 2 | 0.286 | 0.263 | 74.686 | 28.689 | X1 X5 |
| 2 | 0.149 | 0.121 | 89.067 | 45.381 | X2 X5 |
| 3 | 0.508 | 0.484 | 51.466 | 3.737 | X1 X3 X4 |
| 3 | 0.494 | 0.469 | 52.959 | 5.469 | X3 X4 X5 |
| 3 | 0.469 | 0.442 | 55.582 | 8.514 | X1 X3 X5 |
| 3 | 0.461 | 0.434 | 56.436 | 9.506 | X2 X3 X4 |
| 3 | 0.459 | 0.432 | 56.603 | 9.699 | X1 X2 X3 |
| 3 | 0.425 | 0.396 | 60.206 | 13.882 | X1 X4 X5 |
| 3 | 0.421 | 0.392 | 60.581 | 14.317 | X2 X3 X5 |
| 3 | 0.405 | 0.375 | 62.265 | 16.271 | X1 X2 X4 |
| 3 | 0.375 | 0.344 | 65.398 | 19.908 | X2 X4 X5 |
| 3 | 0.306 | 0.271 | 72.670 | 28.349 | X1 X2 X5 |
| 4 | 0.521 | 0.489 | 50.090 | 4.140 | X1 X3 X4 X5 |
| 4 | 0.509 | 0.475 | 51.427 | 5.691 | X1 X2 X3 X4 |
| 4 | 0.501 | 0.467 | 52.203 | 6.592 | X2 X3 X4 X5 |
| 4 | 0.469 | 0.433 | 55.578 | 10.510 | X1 X2 X3 X5 |
| 4 | 0.428 | 0.389 | 59.901 | 15.527 | X1 X2 X4 X5 |
| 5 | 0.522 | 0.481 | 49.970 | 6.000 | X1 X2 X3 X4 X5 |

De la tabla de todas las regresiones posibles, nos interesan las filas 15, 16, y 31 que corresponden a la información del modelo reducido.

$$Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Y el modelo completo en especial sus SSE, necesarios para calcular el estadístico de prueba $F_0$:

$$F_0 = \frac{[\text{SSE}(\beta_0, \beta_2, \beta_5) - \text{SSE}(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)]/3}{\text{MSE}}$$
$$= \frac{[89.067 - 49.970]/3}{0.861546} = 15.1266831$$

Dado que $F_0 = 15.12668 > f_{0.05,3,58} = 2.7635518$, entonces rechazamos $H_0$ y concluimos que el conjunto de predictoras son significativamente distinto de cero.

***Punto 3.*** Prueba de hipótesis lineal general

En un caso hipotético, es de interés para el hospital saber si el censo promedio diario es igual al número promedio de enfermeras y si la duración de la estadía y la rutina de cultivos presentan diferencias importantes en sus efectos:

$$H_0 : \begin{cases} \beta_1 - \beta_2 = 0 \\ \beta_3 - \beta_5 = 0 \end{cases}$$

si se analiza de la forma $H_0 : \mathbf{L}\underline{\beta} = 0$ se tiene que:

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

Así el modelo nulo es

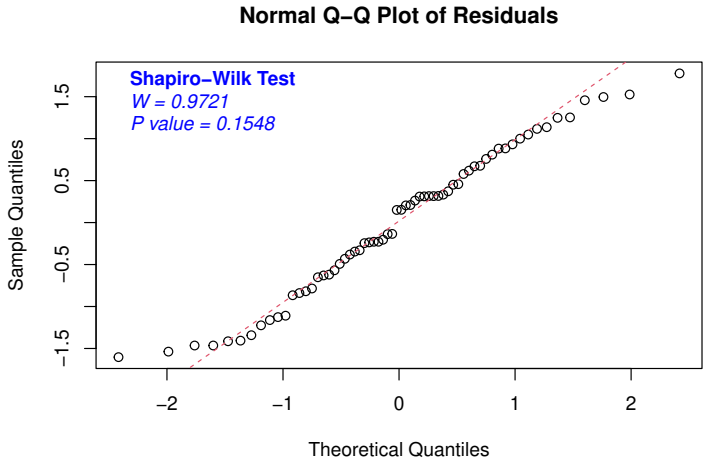
$$Y = \beta_0 + \beta_1(X_1 + X_2) + \beta_3(X_3 + X_5) + \varepsilon$$

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

$$F_0 = \frac{\frac{SSE(ModeloReducido) - SSE(ModeloCompleto)}{g.l(ModeloReducido) - g.l(ModeloCompleto)}}{MSE} = \frac{\frac{SSE(RM) - 49.9697}{61-58}}{0.861546}$$

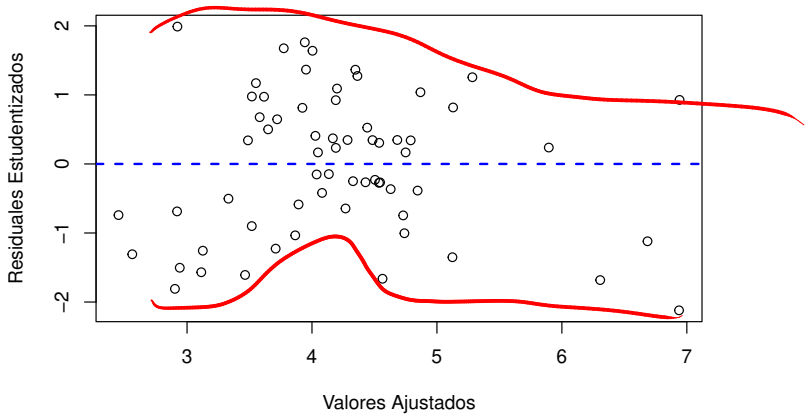
***Punto 4.1 Supuestos del modelo***

Primero se desea validar el supuesto de normalidad en los errores, veamos el siguiente gráfico:



No hacen el análisis gráfico que es más importante 2pt

Como $V_p > \alpha = 0.05$ se concluye que el supuesto de normalidad se cumple. Por otro lado, para el supuesto de varianza constante observemos el gráfico de los Residuales estudentizados vs. Valores ajustados.




2pt

Como no podemos observar un patrón notable en los residuales, podemos decir que los residuales tienen varianza constante. → Análisis muy incompleto.

***Punto 4.2 Observaciones Extremas***

Teniendo en cuenta los distintos métodos para identificar observaciones extremas, como se ve en la siguiente tabla:

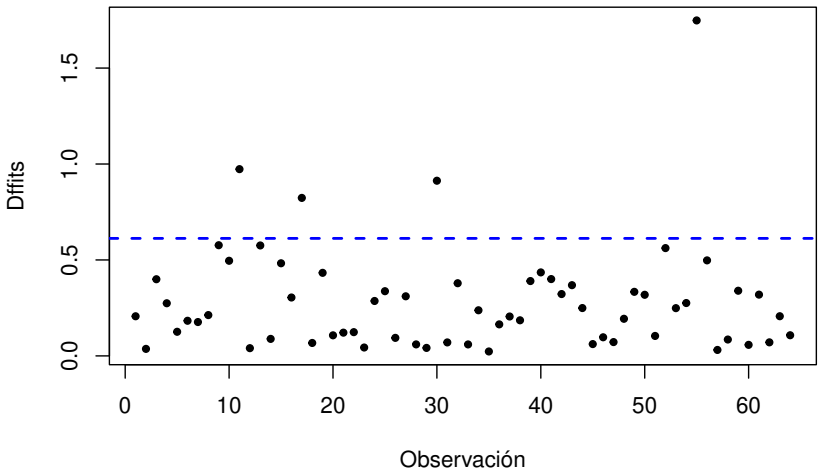
Table 5: Tabla de diagnósticos.

| base.Y | res.stud | Cooks.D | hii | Dffits |
|---|---|---|---|---|
| 2.9 | -0.5031 | 0.0072 | 0.1458 | -0.2065 |
| 3.9 | -0.1523 | 0.0002 | 0.0554 | -0.0366 |
| 1.4 | -1.3080 | 0.0263 | 0.0844 | -0.3995 |
| 5.8 | 1.0384 | 0.0125 | 0.0652 | 0.2743 |
| 4.8 | 0.3055 | 0.0027 | 0.1467 | 0.1257 |

## Observaciones Influenciales.

El criterio del diagnóstico DFFITS ($|\mathbf{DFFITS}_i| > 2\sqrt{(\frac{6}{64})} = 0.6123724$),

¿Distancias de cook? 0pt

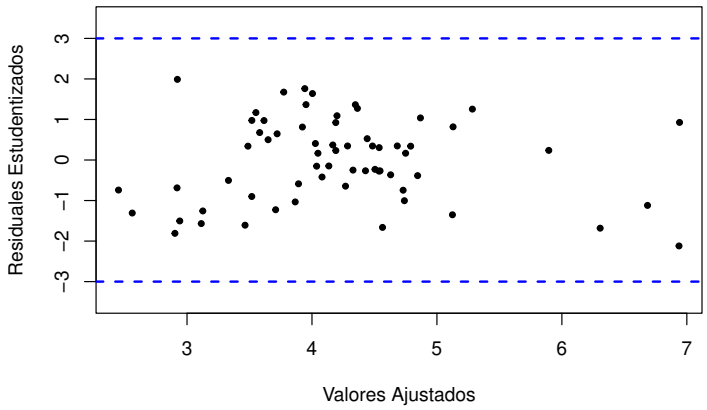



1pt

De acuerdo a este criterio podemos ver que las observaciones 11, 17, 30 y 55 se consideran observaciones influenciales.

¿Qué causan?

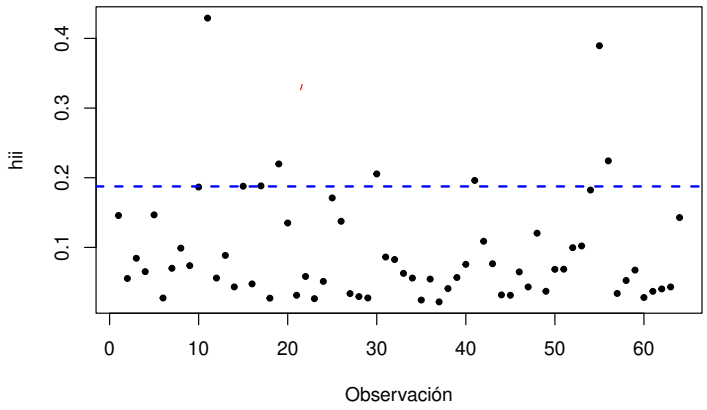
**Datos Atípicos.**




Un valor se considera atípico si $-3 < residualesestudentizados < 3$, por lo tanto según el gráfico no hay observaciones atípicas a considerar.

✓ 3pt

**Puntos de balanceo.**




1 pt

De acuerdo con el Diagnóstico DFBETAS ($h_{ii} > 0.1875$) se sabe que las observaciones 11, 15, 17, 19, 30, 41, 55 y 56 se consideran puntos de balanceo.

¿Causan?

## Conclusiones.

2,5 pt

- Como a la luz de los residuales, el supuesto de normalidad y varianza constante se cumplen, se puede concluir que el modelo propuesto es apto para hacer estimaciones sobre la variable respuesta, sin embargo si observamos con detalle el coeficiente de determinación $R^2$, El modelo propuesto solo explica el 52.25% de la variabilidad total y por lo tanto se podría ampliar la base de datos, estudiar la multicolinealidad o eliminar información redundante.
- Como se presenciaron observaciones influenciales y puntos de balanceo, se sugiere investigar dichas observaciones con el fin de determinar que tanto afectan al modelo y sus parámetros. Y si al corregirlas o eliminarlas el modelo podría mejorar.

No hablan de validez