Trabajo regresión lineal múltiple

Estudiantes

Juan Daniel Bula Isaza
José Daniel Bustamante Arango
Marlon Calle Areiza
Santiago Carvajal Torres

Docente

Raul Alberto Perez Agamez

Asignatura

Estadística II

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
Noviembre de 2021

Índice

Índice de figuras

Índice de cuadros

1. Ejercicio 1

1.1. Análisis de las variables

Es de nuestro interés ajustar un modelo de regresión lineal múltiple acorde a las 72 observaciones de la base de datos, cuyas filas están compuestas por las variables: 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 y x_5: Número de enfermeras. Para ello comenzaremos con un análisis a la matriz de correlaciones entre ellas:

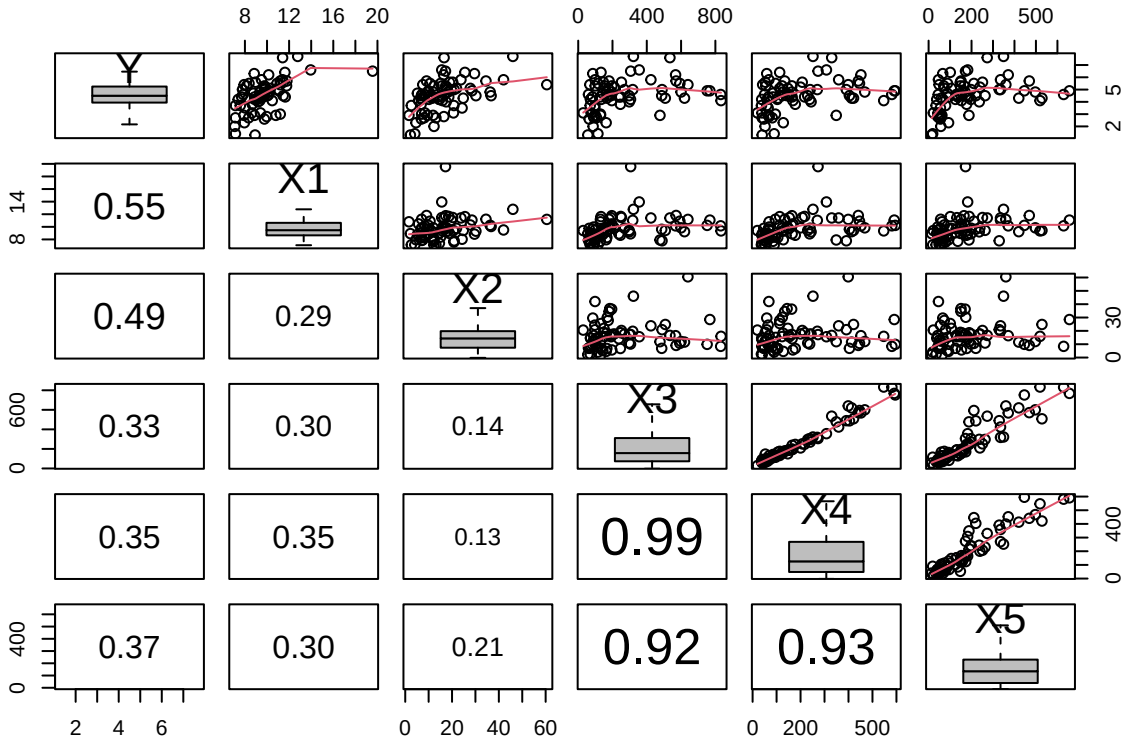

Figura 1: Matriz de correlaciones

Se puede notar una alta dependencia lineal entre X_3 y X_4, X_3 y X_5, y, X_4 y X_5. Además, notar que las correlaciones entre la variable respuesta Y y las variables X_1 X_2 son las más altas en relación a la primera, con respectivos valores de 0.55 y 0.49, suponiendo una relación lineal débil entre las dos variables en los dos casos.

1.2. Ajuste del modelo

Ahora, con las observaciones de la base de datos, procederemos a hacer el ajuste del siguiente modelo de regresión lineal múltiple:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 72$$

Así, el ajuste, redondeando los coeficientes a cuatro cifras decimales, nos resultará en la ecuación:

$$\hat{y}_i = 0.9586 + 0.2682 x_{1i} + 0.0409 x_{2i} + -0.0007 x_{3i} + 0.0009 x_{4i} + 0.001 x_{5i};\ 1 \leq i \leq 72$$

1.3. Análisis de varianza

Se desea verificar la significancia de la regresión usando la tabla ANOVA, mediante la cual compararemos el siguiente juego de hipotesis:

$$\begin{cases} H_0 : \beta_1 = \cdots = \beta_5 = 0 \\ H_1 : \text{Al menos un } \beta_j \neq 0 \end{cases}$$

Para la cual obtendremos lo siguiente:

Cuadro 1: Tabla ANOVA para el modelo

	Suma de cuadrados	gl	Cuadrado Medio	F_0	Valor P
Regresión	52.3209	5	10.464176	10.7672	1.35475e-07
Error	64.1423	66	0.971853		

Donde $F_0 = \frac{MSR}{MSE} \sim F_{5,66}$ bajo H_0

Podemos concluir así que con un nivel de significancia del 5 % al menos uno de los coeficientes es significativo, esto del rechazo de H_0 por el Valor P.

1.4. Análisis de varianza

Ahora procederemos hacer un análisis marginal de cada uno de los coeficientes usando la siguiente prueba para $j = 1, \cdots, 5$:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

con $T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{66}$ bajo H_0

Para lo que usaremos la siguiente tabla:

Cuadro 2: Resumen de los coeficientes

	Estimación	Error estándar	T_0	Valor P
β_0	0.9586	0.6599	1.4527	0.1511
β_1	0.2682	0.0742	3.6135	0.0006
β_2	0.0409	0.0120	3.4091	0.0011
β_3	-0.0007	0.0035	-0.2036	0.8393
β_4	0.0009	0.0050	0.1850	0.8538
β_5	0.0015	0.0021	0.7108	0.4797

Se puede observar que con un nivel de significancia de $\alpha = 0.05$, a nivel marginal las únicas variables que tienen un efecto significativo en la respuesta son x_1 y x_2 que representan la duración de la estadía y la rutina de cultivos respectivamente.

Notar que: $\hat{\beta}_1 = 0.2682$ indica que por un aumento unitario en la duración de la estadía x_1, la media de la variable respuesta aumentará en 0.2682 unidades; de la misma manera, $\hat{\beta}_2 = 0.0409$ indica que por cada aumento en una unidad en la rutina de cultivos, la media de la variable respuesta aumentará en 0.0409 unidades; estos dos casos siempre y cuando las otras variables se mantengan constantes.

1.5. Cálculo y análisis del coeficiente de determinación múltiple

Tenemos que el coeficiente de determinación multiple, denotado por R^2 se define como:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

con SSR: Suma de cuadrados de la regresión , SSE: Suma de cuadrados de los residuales, y $SST = SSR + SSE$. Así tenemos, con los datos calculados anterioremente en la tabla ANOVA, que:

$$R^2 = \frac{52.3209}{52.3209 + 64.1423} = 0.4492483$$

Tal R^2 nos dice que un 44.92 % de la variabilidad total en el Riesgo de infección (Variable respuesta) es explicado por el modelo RLM propuesto.

2. Ejercicio 2

3. Ejercicio 3

4. Ejercicio 4