Universidad Nacional de Colombia

Facultad de Ciencias

Análisis de Regresión

Trabajo RLM: Parte 1

Autores:
Santiago Franco Valencia
Juan Pablo Martínez Echavarría
Alejandro Salazar Mejía

Agosto 2021

Análisis Descriptivo

Antes de la solución del taller se realizó un análisis descriptivo para tener un mejor entendimiento de los datos que se están tratando.

Análisis de variables numéricas:
A continuación se muestran una tabla de estadísticos descriptivos por cada variable numérica de la base de datos, donde:

- Min: Mínimo valor registrado de dicha variable.
- $1stQu$: Primer cuartil de los datos (cuantil 0.25), i.e., el 25 % de los datos es menor o igual a dicho valor.
- $Median$: Mediana de los datos, i.e., valor que divide los datos en dos partes iguales.
- $Mean$: Media aritmética de los datos.
- $3stQu$:Tercer cuartil de los datos (cuantil 0.75), i.e., el 25 % de los datos es mayor o igual a dicho valor.
- Max: Máximo valor registrado de dicha variable.
- sd: Desviación estándar muestral de los datos.

Table 1: Estadísticos de resumen para variable DPERM

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
7.08	8.39	9.38	9.72	10.66	19.56	2.03

Table 2: Estadísticos de resumen para variable EDAD

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
38.8	51	53.2	53.25	56.08	65.9	4.59

Table 3: Estadísticos de resumen para variable RINF

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
1.3	3.7	4.4	4.4	5.3	7.8	1.37

Table 4: Estadísticos de resumen para variable RRC

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
1.6	8.4	14.05	16.13	20.75	60.5	10.73

Table 5: Estadísticos de resumen para variable RRX

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
39.6	69.2	85.4	82.24	96.05	133.5	20.1

Table 6: Estadísticos de resumen para variable NCAMAS

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
29	102	184	246.89	305.75	835	185.8

Table 7: Estadísticos de resumen para variable PDP

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
20	66.25	136.5	186.56	247	791	152.75

Table 8: Estadísticos de resumen para variable NENFERM

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
19	66.25	124.5	165.97	208.75	629	129.55

Table 9: Estadísticos de resumen para variable FSD

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
14.3	31.4	41.45	42.16	51.4	74.3	13.47

La figura 1 es una matriz de gráficos que pretende presentar la relación entre parejas de variables. En el margen superior y lateral derecho se indican el nombre de las variables que se ponen a interactuar. En la diagonal principal de esta matriz, se encuentra la gráfica de densidad (no paramétrica) de cada variable numérica. Por encima de esta diagonal, se presenta la correlación de cada par diferente de variables. Por debajo de la diagonal se encuentran los gráficos de dispersión de cada par diferente de variables, junto con una línea recta ajustada y su respectivo intervalo de confianza alrededor de esta.

Similarmente, la figura 2 también es una matriz de gráficos. En este caso, la diagonal de la matriz presenta el histograma correspondiente a cada variable numérica. Por encima de esta diagonal, se muestran graficadas las curvas de nivel de la densidad bivariada (no paramétrica) de las dos variables en cuestión. Por debajo de la diagonal se encuentran los gráficos de disperción de cada par diferente de variables, junto con una curva LOESS ajustada (no paramétrica) y su respectivo intervalo de confianza alrededor de esta.

Figura 1: Figura 1

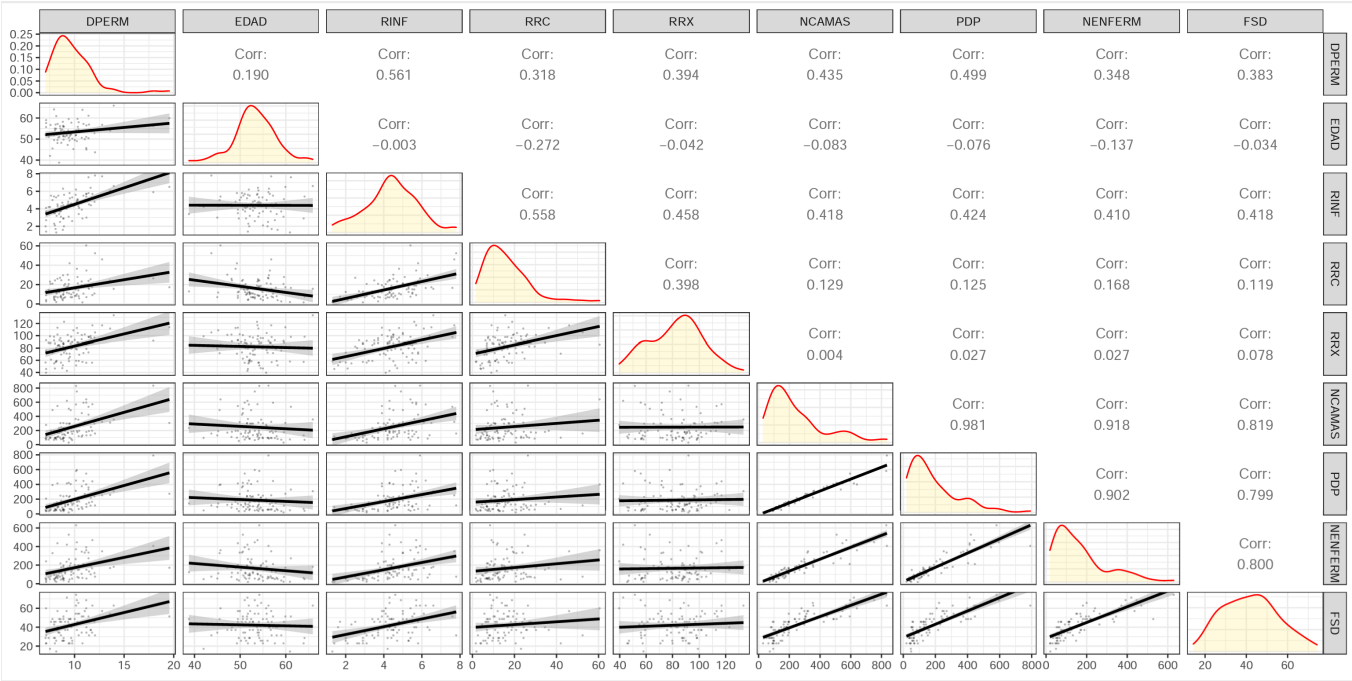
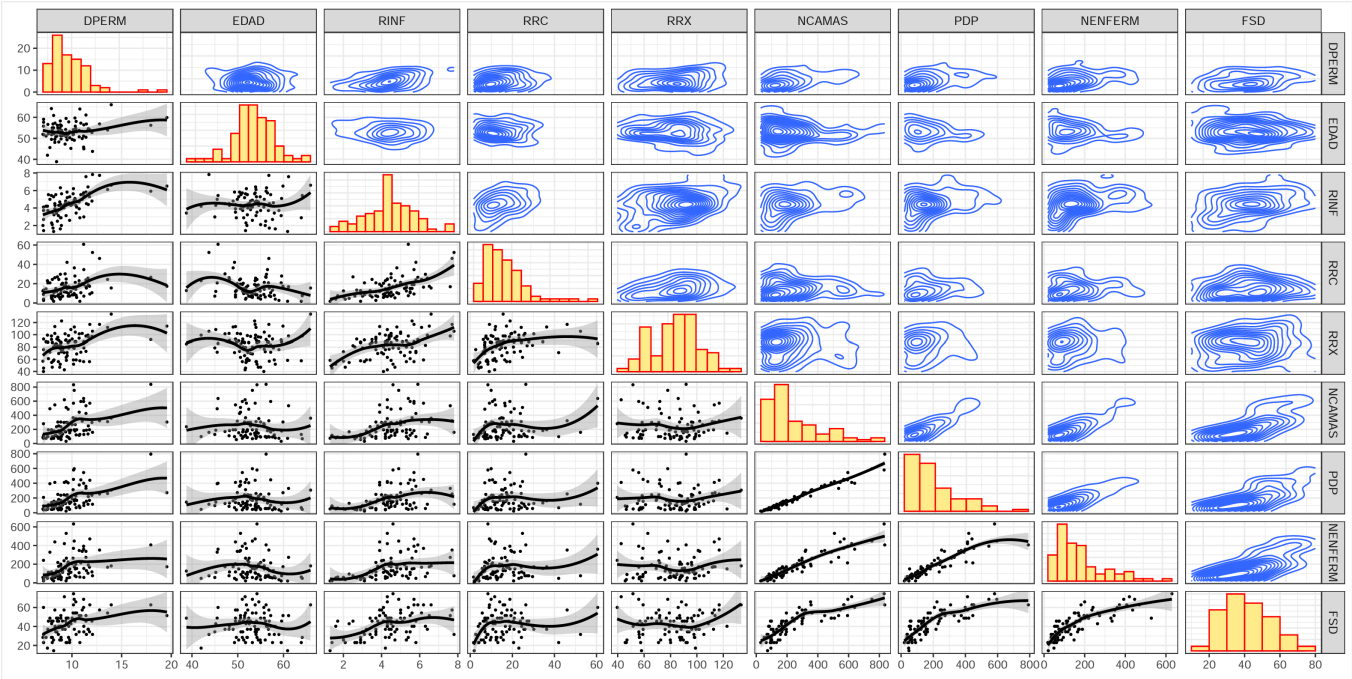


Figura 2: Figura 2



De los gráficos podemos resaltar algunas observaciones interesantes:

1. La varible **Edad** parece tener el efecto más leve sobre la variable respuesta **Longitud de permancencia** (DPERM). En cambio, podría creerse que el resto de variables son significativas para explicar la variabilidad de DPERM, siguiendo una tendencia positiva.

2. Del mismo modo, **Edad** no aparenta estar relacionada con el resto de variables explicatorias.

3. Se observa que las variables **Número de camas**, **Censo promedio diario**, **Número de enfermeras** y **Facilidades y servicios disponibles** están altamente relacionadas positivamente entre sí.

4. La variable **Riesgo de infección** aparenta estar relacionada positivamente con el resto de covariables, excepto **edad**.

5. Las variables **Razón de rutina de rayos X del pecho** y **Razón de rutina de cultivos** no muestran estar relacionadas estadísticamente con las variables **Número de camas**, **Censo promedio diario**, **Número de enfermeras** y **Facilidades y servicios disponibles**.

Análisis de variables categóricas:
A continuación se muestran un par de tablas de frecuencia correspondientes a las variables categóricas de la base de datos.

Table 10: Hospitales afiliados a la escuela de medicina

Afiliados	No Afiliados
13	77

Table 11: Hospitales en regiones geográficas

NE	NC	S	W
23	25	29	13

Análisis de variables numéricas agrupadas por variables categóricas:
Las figuras 3 y 4 son matrices de gráficos que pretenden presentar la relación entre parejas de variables, agrupadas por las variables categóricas **Afiliación a escuela de medicina** y **Región**, respectivamente. En el margen superior y lateral derecho se indican el nombre de las variables que se ponen a interactuar, y en la parte inferior se muestra la leyenda que explica la agrupación de las variables por colores. En la diagonal principal se presenta el diagrama de Cajas y Bigotes de cada variable, discriminado por la variable categórica correspondiente. Por encima de esta diagonal, se presenta la correlación de cada par diferente de variables, de acuerdo a la categoría correspondiente. Por debajo de la diagonal se encuentran los gráficos de dispersión de cada par diferente de variables, donde los colores de cada punto dependen del grupo al que pertenecen.

Figura 3: Figura 3

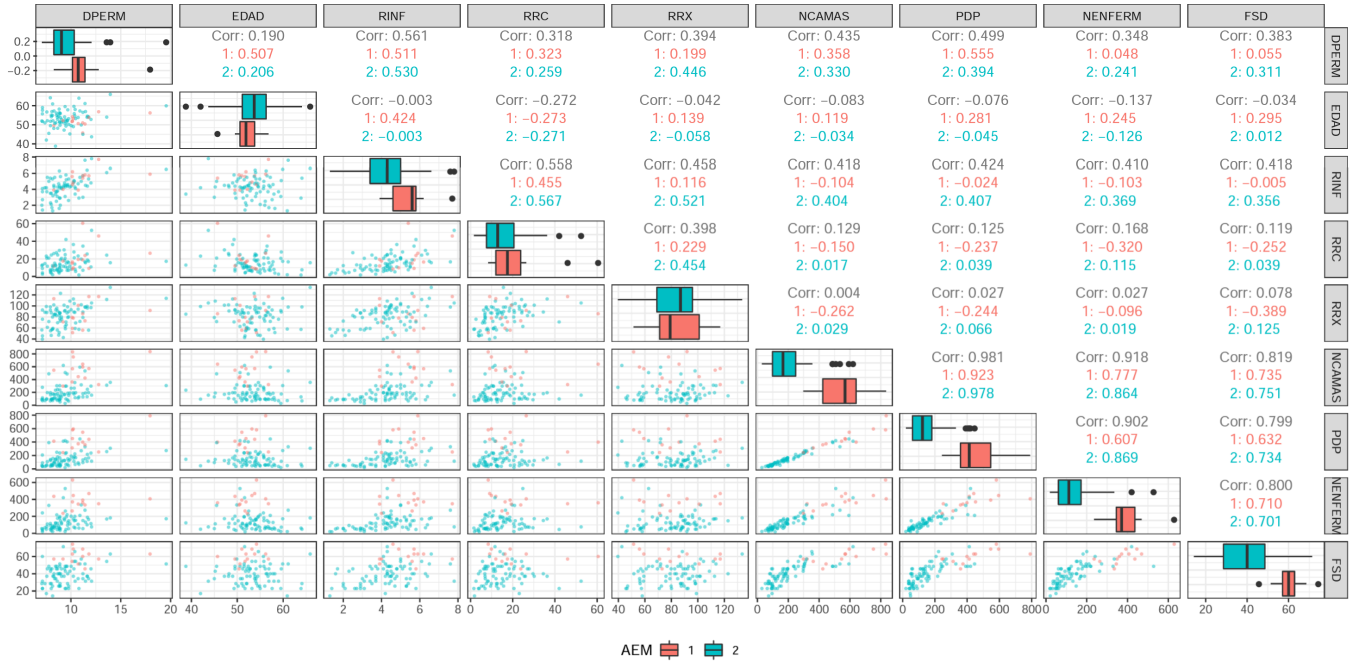
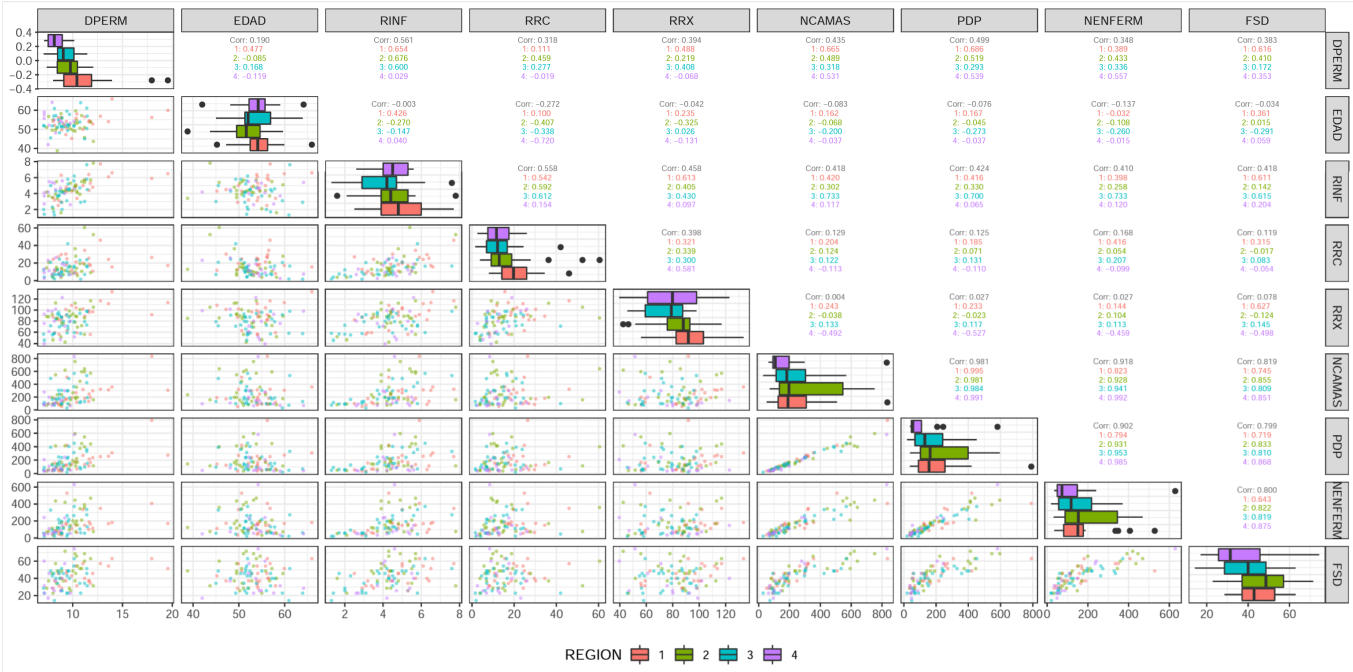


Figura 4: Figura 4



Solución:

1. Ajuste de modelo de regresión lineal múltiple

Considere las siguientes variables:

Y_i : i-ésima observación de la variable respuesta 'Longitud de permanencia' (DPERM).
X_{i1} : i-ésima observación de la variable predictoria 'Edad' (EDAD).
X_{i2} : i-ésima observación de la variable predictoria 'Riesgo de infección' (RINF).
X_{i3} : i-ésima observación de la variable predictoria 'Razón de rutina de cultivos' (RRC).
X_{i4} : i-ésima observación de la variable predictoria 'Razón de rutina de rayos X del pecho' (RRX).
X_{i5} : i-ésima observación de la variable predictoria 'Número de camas' (NCAMAS).
X_{i6} : i-ésima observación de la variable predictoria 'Censo promedio diario' (PDP).
X_{i7} : i-ésima observación de la variable predictoria 'Número de enfermeras' (NENFERM).
X_{i8} : i-ésima observación de la variable predictoria 'Facilidades y servicios disponibles' (FSD).

Se observa que se tienen 90 observaciones y k= 8 variables regresoras.

Se asume que el modelo de regresión lineal múltiple tiene la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \beta_8 X_{i8} + E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), i = 1, ..., 90$$

Se ajusta un modelo de regresión lineal multiple:

Se escribe la ecuación ajustada:

$\hat{Y_i} = -0.2084 + 0.1043X1 + 0.3352X2 + 0.0287X3 + 0.0209X4 - 0.0106X5 + 0.0223X6 - 0.006X7 + 0.0041X8, i = 1, ..., 90.$

Se muestra la tabla de parámetros ajustados:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.208	1.945	-0.107	0.915
EDAD	0.104	0.034	3.069	0.003
RINF	0.335	0.156	2.154	0.034
RRC	0.029	0.018	1.615	0.110
RRX	0.021	0.009	2.466	0.016
NCAMAS	-0.011	0.004	-2.401	0.019
PDP	0.022	0.005	4.582	0.000
NENFERM	-0.006	0.003	-2.055	0.043
FSD	0.004	0.020	0.213	0.832

Se calcula la tabla ANOVA del modelo:

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	217.568	8	27.19606	14.6536	5.0418e-13
Error	150.330	81	1.85592		

Con un p-value casi igual a cero, se concluye que al menos una de las covariable es significativa para explicar la variabilidad de la longitud de permanencia.

Del resumen del modelo se obtiene que el valor de R^2 es 0.5914, es decir que un 59.14% de la variabilidad total de la longitud de permanencia es explicada por el modelo. Se opina que este porcentaje de la variabilidad tan "bajo" puede deberse a que alguna de las covariables no sea significativa, o no es adecuado suponer que existe una relación lineal entre la longitud de permanencia y las covariables númericas presentándose carencia de ajuste.

2. Coeficientes estandarizados:

Se calculan los coeficientes de un modelo de regresión lineal multiple con las variables estándarizadas, ordenados de menor a mayor:

Table 3: Tabla de coeficientes Estandarizados

	Coef.Std
FSD	0.028
RRC	0.151
RRX	0.207
RINF	0.225
EDAD	0.235
NENFERM	0.384
NCAMAS	0.978
PDP	1.680

De la tabla se concluye que la covariable que "más" aporta al modelo cuando los datos se encuentran estándarizados es el "Censo promedio diario", indicando que un aumento unitario en el "Censo promedio diario" estándarizado aumentaría en 1.68 unidades en promedio la longitud de permanencia estándarizada, dado que el resto de covariables están en el modelo.

Además del "Censo promedio diario", se encontró que dos de las covariables que estándarizadas más aportan al modelo son el número de camas y el número de enfermeras, sugiriendo que las primeras tres covariables que estandarizadas más aportan al modelo se relacionan con el tamaño de los hospitales.

3. Significancia individual de los parámetros del modelo:

Cada una de las pruebas t para la significancia individual de los parámetros del modelo tienen la siguiente forma:

$$H_0 : \beta_j = 0 \;\; vs. \;\; H_1 : \beta_j \neq 0 \;, j = 1, ..., 8.$$

$$T_{0j} = \frac{\widehat{\beta}_j}{s.e(\widehat{\beta}_j)} \sim t_{81} \;, j = 1, ..., 8.$$

$$p\text{-}value_j = P(|t_{81}| > |T_{0j}|).$$

Se crea una tabla de coeficientes asociados a cada covariable, que incluye el valor de su estadístico t y el valor p de su prueba de hipótesis.

	Estimación	Estadístico t	Valor p
EDAD	0.104	3.069	0.003

	Estimación	Estadístico t	Valor p
RINF	0.335	2.154	0.034
RRC	0.029	1.615	0.110
RRX	0.021	2.466	0.016
NCAMAS	-0.011	-2.401	0.019
PDP	0.022	4.582	0.000
NENFERM	-0.006	-2.055	0.043
FSD	0.004	0.213	0.832

De la tabla anterior se concluye que utilizando la prueba t los parámetros β_3 y β_8 no son significativos, es decir que la "Razón de rutina de cultivos" y "Facilidades y servicios disponibles" no ayudan a explicar la variabilidad de la "Longitud de permanencia" dado que las demás covariables no se encuentran en el modelo.

Prueba F para dos predictoras:

A partir de la tabla anterior se realiza una prueba F para las covariables "Censo promedio diario", #### Prueba F para test lineal general para el Censo promedio diario Se define el modelo completo:

$$Y_i = \beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}+\beta_4 X_{i4}+\beta_5 X_{i5}+\beta_6 X_{i6}+\beta_7 X_{i7}+\beta_8 X_{i8}+E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2), i = 1, ..., 90.$$

Se define el modelo reducido:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_7 X_{i7} + \beta_8 X_{i8} + E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$$

A partir de los datos, se define:

$$F_{06} = \frac{SSR(X_6|X_1,X_2,X_3,X_4,X_5,X_7,X_8)}{MSE(X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8)} \ , \tag{1}$$

$$F_{06} = \frac{SSE(X_1,X_2,X_3,X_4,X_5,X_7,X_8) - SSE(X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8)}{MSE(X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8)} \ , \tag{2}$$

$$F_{06} \sim f_{1,81} \tag{3}$$

Cálculo de valor p:

$$p\text{-}value = P(f_{1,81} > F_{06})$$
$$g.l.(SSE(X_1,X_2,X_3,X_4,X_5,X_7,X_8)) = 82$$
$$g.l.(SSE(X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8)) = 81$$
$$SSE(X_1,X_2,X_3,X_4,X_5,X_7,X_8) = 189.30$$
$$SSE(X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8) = 150.33$$
$$g.l(SSR(X_6|X_1,X_2,X_3,X_4,X_5,X_7,X_8)) = 82-81 = 1$$
$$SSR(X_6|X_1,X_2,X_3,X_4,X_5,X_7,X_8) = 189.30-150.33 = 38.969$$

A partir de los valores previos se obtiene la siguiente tabla:

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
82	189				
81	150	1	38.9694680543685	20.9973358998904	1.6484475391443e-05

La columna 'F' presenta el valor de F_{06} y la columna Pr(>F) su respectivo p-value.

Del resultado de la prueba anterior se concluye que el "Censo promedio diario" ayuda a explicar la "Longitud de permenancia" dado que el resto de las covariables se encuentran en el modelo.

Prueba F para test lineal general para la Edad

Modelo completo: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \beta_8 X_{i8} + E_i$, $E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Modelo reducido: $Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \beta_8 X_{i8} + E_i$, $E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

A partir de los datos, se define:

$$F_{01} = \frac{SSR(X_1|X_2, X_3, X_4, X_5, X_6, X_7, X_8)}{MSE(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)} , \tag{4}$$

$$F_{01} = \frac{SSE(X_2, X_3, X_4, X_5, X_6, X_7, X_8) - SSE(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)}{MSE(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)} , \tag{5}$$

$$F_{01} \sim f_{1,81} \tag{6}$$

Se calcula el valor p:

$$p\text{-}value = P(f_{1,81} > F_{01})$$

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
82	168				
81	150	1	17.4778182328601	9.41731151476297	0.00292351905114798

La columna 'F' presenta el valor de F_{06} y la columna Pr(>F) su respectivo p-value.

Del resultado de la prueba anterior se concluye que la Edad ayuda a explicar la "Longitud de permenancia" dado que el resto de las covariables se encuentran en el modelo.

4. Sumas de cuadrados de tipo 1 y de tipo 2:

Sumas de cuadrados de tipo 1:

Table 7: Sumas de cuadrados tipo I y SSE

	Suma de cuadrados	g.l
EDAD	13.285	1
RINF	116.224	1
RRC	1.936	1
RRX	8.443	1
NCAMAS	31.852	1
PDP	37.873	1
NENFERM	7.870	1
FSD	0.084	1
Residuals	150.330	81

Sumas de cuadrados de tipo II:

Table 8: Sumas de cuadrados tipo II y SSE

	Suma de cuadrados	g.l
FSD	0.084	1
RRC	4.841	1
NENFERM	7.839	1
RINF	8.614	1
NCAMAS	10.700	1
RRX	11.289	1
EDAD	17.478	1
PDP	38.969	1
Residuals	150.330	81

Se encuentra que en ambas tablas, la regresora que tiene menor valor en las sumas de cuadrados de tipo I y II es Facilidades y Servicios disponibles".

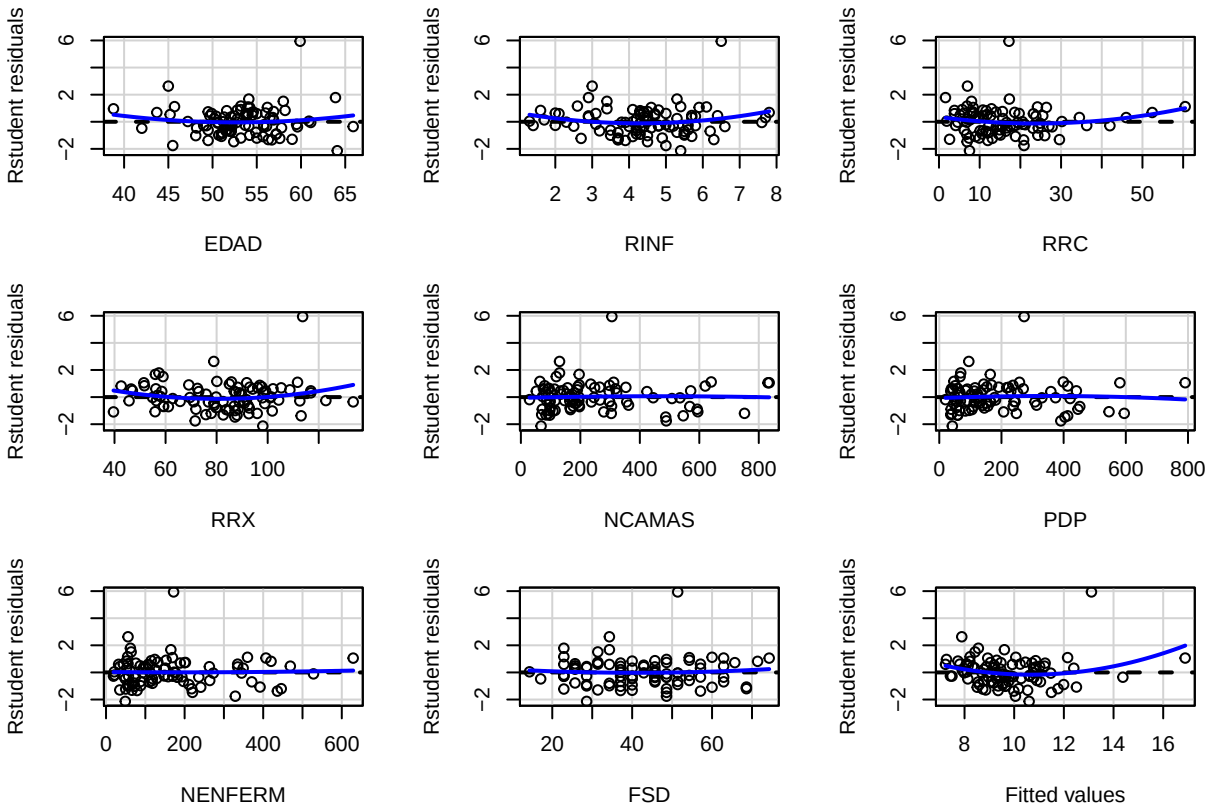
Para la suma de cuadrados parciales este bajo valor conlleva a un estadístico F también de bajo valor, implicando un valor p alto, dando indicios de que las Facilidades y Servicios disponibles no ayudan a explicar la Longitud de permanencia, dado que todas las demás covariables están en el modelo, lo cuál se comprobó en el numeral anterior.

Para la suma de cuadrados secuenciales se obtuvo el mismo valor que en la suma de cuadrados parciales ya que fue la última variable que se ingresó en la secuencia. Este valor significa el aumento marginal en la suma de cuadrados de la regresión que se obtiene al agregar Facilidades y Servicios disponibles dado que las demás variables se encuentran en el modelo.

En ambos casos su interpretación es la misma.

5. Gráfico de residuales estudentizados vs Valores ajustados

De los gráficos studentizados buscamos evaluar el supuesto de varianza constante y carencia de ajuste en el modelo.

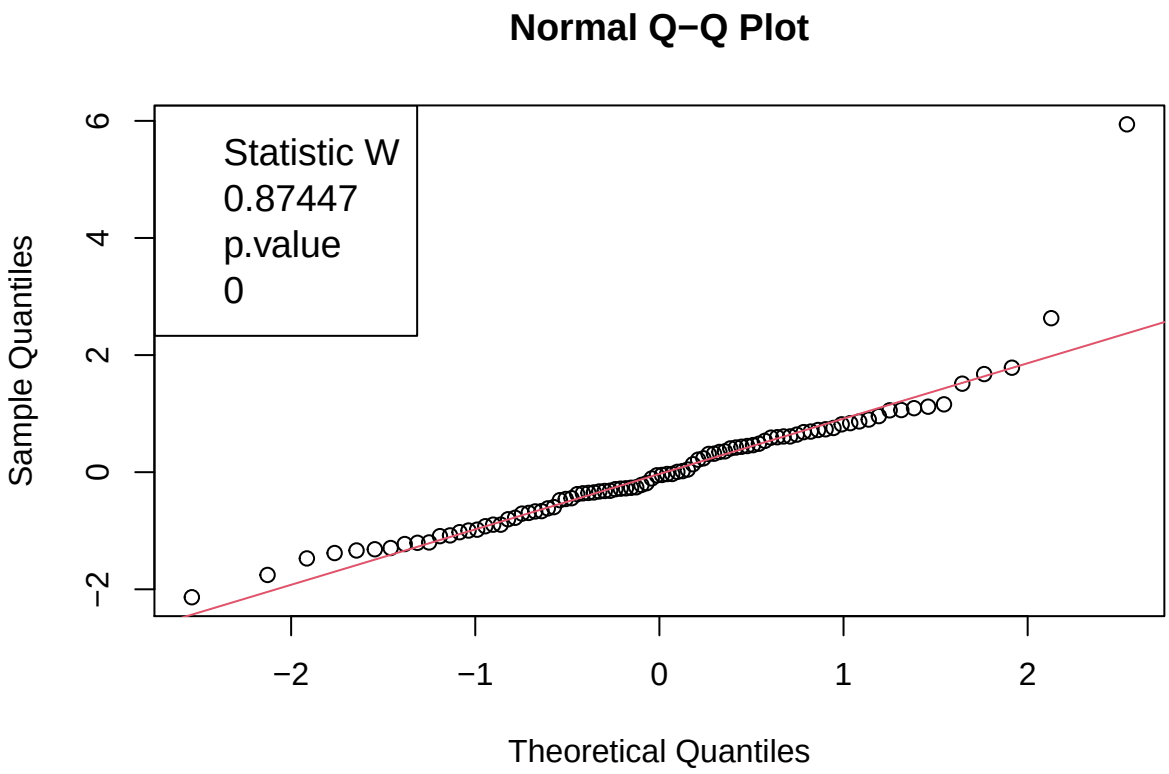


En las gráficas de los residuos vs. covariables no se nota algún patrón en contra del supuesto de varianza constante. Por el otro lado, del gráfico de Residuales vs. variable respuesta (Longitud de permanecia), aunque se cumple el supuesto de varianza constante, hay motivos para creer que se presencia una carencia de ajuste debido al patrón cuadrático.

6. Gráfica de probabilidad normal para los residuales estudentizados.

Se encuentra el estadístico W de la prueba de Shapiro-Wilk asociada a la normalidad de los datos definida a continuación:

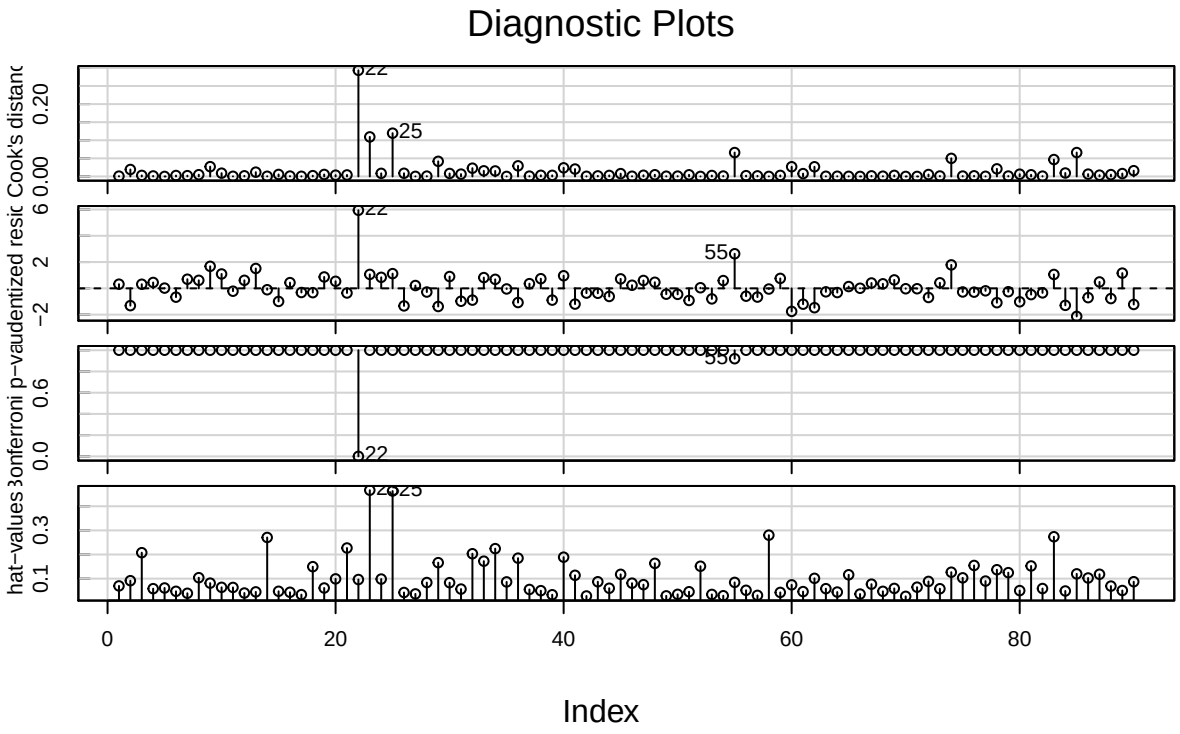
H_o : Los datos son normales. vs H_1 : Los datos son no normales.

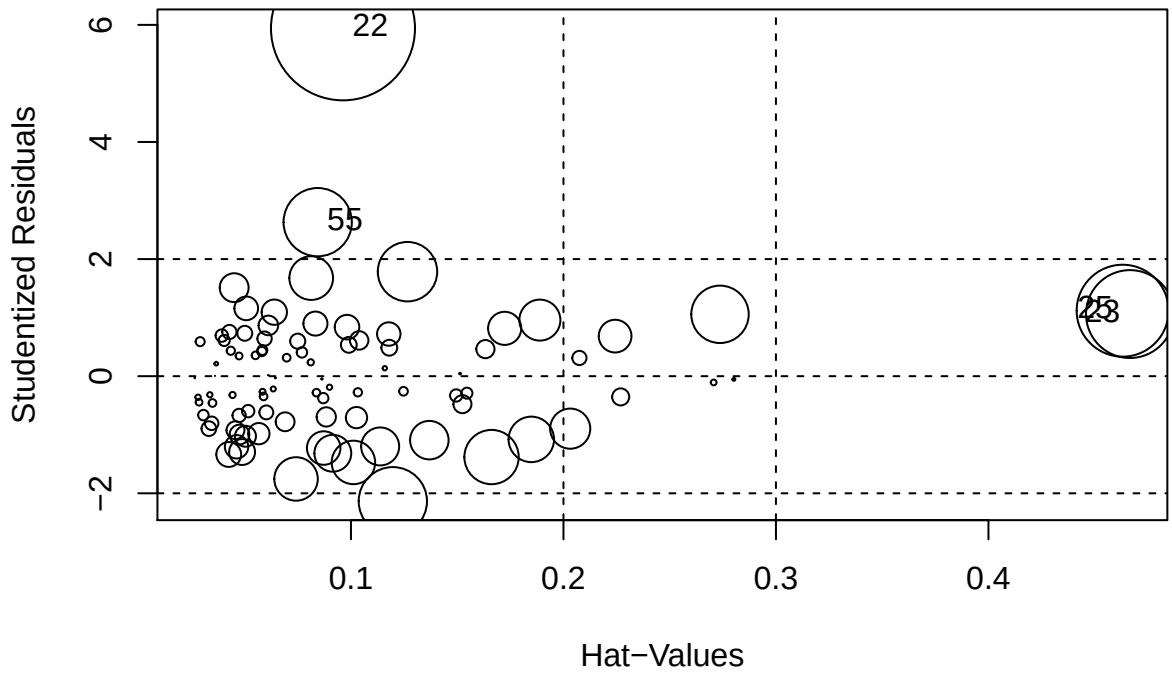


Se observa un desajuste al inicio y al final de la gráfica que permite dudar sobre la hipótesis de normalidad sobre los errores del modelo. De acuerdo a la prueba de bondad de ajuste Shapito-Wilks, los errores studentizados no se distribuyen normalmente. Sin embargo, esta prueba es muy sensible a datos atípicos los cuales son observados en la gráfica.

7. Diagnóstico de la presencia de observaciones atípicas, de balanceo y/o influenciales.

Se crea una tabla con las medidas de influencia:


Diagnostic Plots
Cook's distanc
0.00
0.20
22
25
p−vaudentized resic
−2
2
22
55
onferroni p
0.0
0.6
55
22
hat−values3
0.1
0.3
22
25
0
20
40
60
80
Index




```
[1] 47
```

Como lo ilustran las gráficas y la tabla correspondiente, la observación con índice 22, i.e. el hospital con ID 47, es una observación atípica.

```
[1]  13  66 104 112   8  48  54  53  46
```

Según el criterio de los hat-values: $h_{ii} > 0.2$, las observaciones de balanceo son aquellos hospitales con ID 13, 66, 104, 112, 8, 48, 54, 53 y 46.

Observaciones influyentes según las distancias de cook:

```
named integer(0)
```

Según el criterio de las distancias de cook no se tienen observaciones influyentes.

Observaciones influyentes según los DFBetas:

Según el criterio de los DFBetas las siguientes observaciones son cándidatos a influyentes:

	dfb.1_	dfb.EDAD	dfb.RINF	dfb.RRC	dfb.RRX	dfb.NCAM	dfb.PDP	dfb.NENF	dfb.FSD
22	-1.055	0.702	0.600	-0.348	0.627	-0.541	0.702	-0.418	0.346

	dfb.1_	dfb.EDAD	dfb.RINF	dfb.RRC	dfb.RRX	dfb.NCAM	dfb.PDP	dfb.NENF	dfb.FSD
40	0.332	-0.390	-0.026	-0.106	0.061	0.120	-0.111	-0.148	0.150
55	0.642	-0.637	-0.037	-0.297	0.090	0.027	0.043	-0.263	0.065
60	-0.333	0.287	-0.065	0.015	0.117	-0.007	-0.069	-0.021	0.169
74	-0.250	0.442	-0.037	-0.045	-0.101	0.200	-0.165	0.053	-0.330
85	0.438	-0.451	-0.320	0.272	-0.145	-0.022	0.083	-0.112	0.215
9	0.090	0.018	0.290	-0.021	-0.363	-0.169	0.122	0.138	-0.136
25	-0.008	0.001	-0.352	0.714	-0.028	0.626	-0.504	-0.231	-0.031
29	-0.011	0.061	0.246	0.120	-0.410	-0.099	0.120	-0.270	0.126
23	-0.098	0.148	-0.173	0.229	0.033	-0.301	0.678	-0.374	-0.156
36	-0.085	0.065	-0.076	-0.039	0.254	0.310	-0.364	0.039	-0.138
78	0.021	-0.046	-0.064	0.000	0.294	0.123	-0.017	-0.093	-0.256
32	0.024	-0.038	-0.077	0.053	0.032	-0.247	0.068	0.319	0.095
83	-0.050	0.088	-0.037	-0.071	-0.086	0.278	-0.331	0.283	-0.097
62	-0.098	0.098	0.004	0.160	-0.118	-0.081	-0.120	0.286	0.101

```
 [1]  47  75  43  65 106  63  33   8  26 112  62   1  48  46  69
```

Estas corresponden a los hospitales con los IDs 47, 75, 43, 65, 106, 63, 33, 8, 26 ,112, 62, 1, 48, 46 y 69.

Observaciones influenciales según los DFFITS:

Según los dffits se tiene que las siguientes observaciones son candidatas a influenciales:

influencias_dffits	
22	1.939
23	0.992
25	1.039
55	0.798
74	0.680
83	0.648
85	-0.788

```
[1] "47, 112, 8, 43, 106, 46, 63"
```

Estas corresponden a los hospitales con los IDs 47, 112, 8, 43, 106, 46, 63.

Puntos influenciales según el COVRATIO:

Se verifica que $90 > 3(9)$, luego se puede concluir que una observación será cándidata a ser influencial sí $|COVRATIO_i - 1| > \frac{3(9)}{90}$, así, las siguientes observaciones son candidatas a ser influenciales:

	x
3	1.396
14	1.531
21	1.426
22	0.046
23	1.849

	x
25	1.811
34	1.368
48	1.305
52	1.317
55	0.579
58	1.553
76	1.310
83	1.359

```
[1] "13, 66, 104, 47, 112, 8, 54, 109, 40, 43, 53, 110, 46"
```

Estas corresponden a los hospitales con los IDs 13, 66, 104, 47, 112, 8, 54, 109, 40, 43, 53, 110, 46.

Como se mencionó al inicio, el hospital con $ID = 47$ es una observación atípica. Además es considerada como candidata a influenciable por los críterios de DFBeta, para al menos un β_i, DFFITS Y CovRatio. Por otro lado, el hospital con $ID = 112$ es considerado un punto de balanceo. Además, es cándidato a ser influenciable según los criterios de DFBeta, para al menos un β_i, DFFITS Y CovRatio.

Se ajusta un modelo para los datos sin incluir las observaciones cuyo ID no es 47 ni 112:

Se presenta la tabla de parámetros ajustados:

Table 12: Tabla de Parámetros estandarizados sin obs. con ID 47 y 112

	Estimate	Std. Error	t value	Pr(>|t|)
EDAD	0.226	0.085	2.657	0.010
RINF	0.266	0.115	2.311	0.023
RRC	0.195	0.106	1.846	0.069
RRX	0.209	0.092	2.268	0.026
NCAMAS	-0.742	0.444	-1.671	0.099
PDP	1.279	0.442	2.896	0.005
NENFERM	-0.266	0.216	-1.229	0.223
FSD	0.029	0.143	0.203	0.840

Table 13: Tabla de parámetros con todas las observaciones estandarizadas

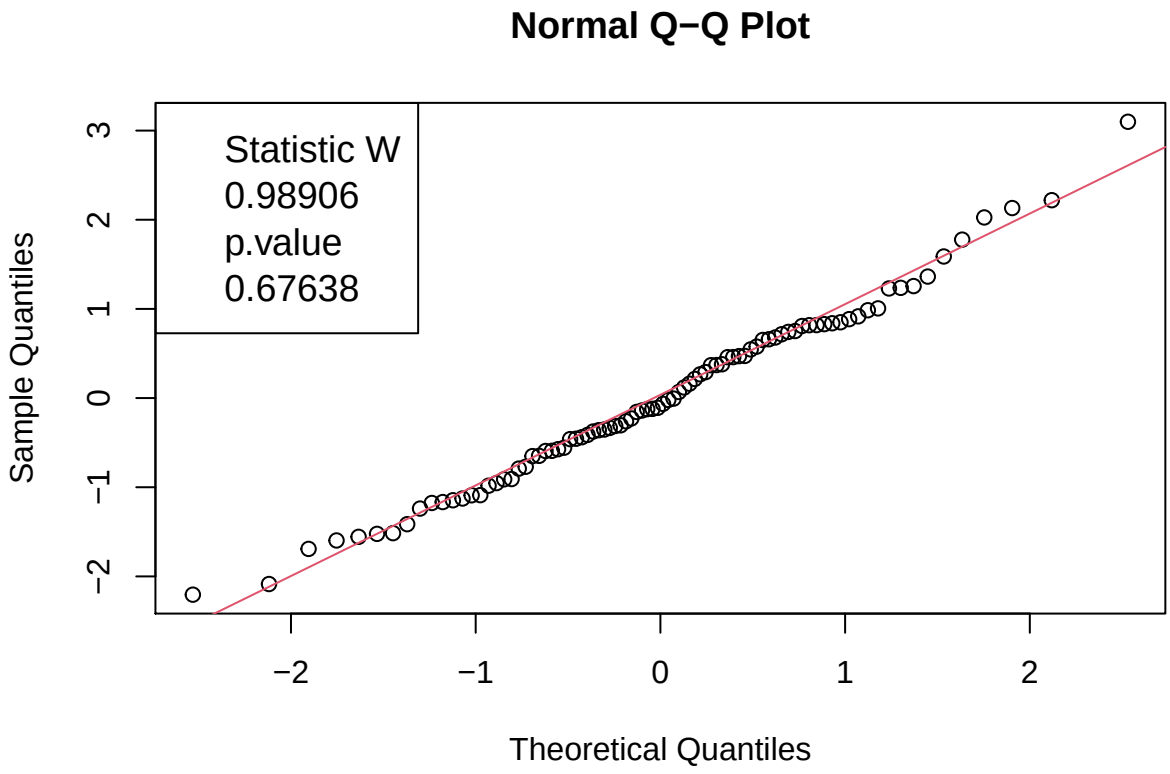
	Estimate	Std. Error	t value	Pr(>|t|)
EDAD	0.235	0.076	3.088	0.003
RINF	0.225	0.104	2.168	0.033
RRC	0.151	0.093	1.625	0.108
RRX	0.207	0.084	2.481	0.015
NCAMAS	-0.978	0.405	-2.416	0.018
PDP	1.680	0.364	4.610	0.000
NENFERM	-0.384	0.186	-2.068	0.042
FSD	0.028	0.129	0.214	0.831

En las siguiente tabla se muestra el error relativo del modelo sin las observciones con ID = 47 y 112 con respecto al modelo con todos las observaciones.

	Estimate	Std. Error	t value	Pr(>|t|)
EDAD	0.038	0.118	0.140	2.459
RINF	0.181	0.108	0.066	0.293
RRC	0.287	0.133	0.136	0.365
RRX	0.008	0.103	0.086	0.721
NCAMAS	-0.241	0.097	-0.308	4.504
PDP	0.238	0.212	0.372	332.388
NENFERM	-0.308	0.164	-0.406	4.325
FSD	0.048	0.107	0.054	0.011

Se observa que la estimación de $\beta_3, \beta_5, \beta_6, \beta_7$ cambia en más del 20% con respecto a la estimación original con todas las observaciones en unidades estándar. Solo el error estándar de la estimación de β_6 aumenta en más de un 20% con respecto a la estimación original con todas las observaciones en unidades estándar. Con un nivel de significancia del 5%, el modelo sin las observaciones $ID = 47 y ID = 112$ muestra que las co-variables "Número de camas" y "Número de enfermeras" ya no son significativa, dado que el resto de variables explicatorias están en el modelo. Es decir, en el modelo sin las dichas observaciones, el número de camas y enfermeras no ayuda a explicar la longitud de permanencia.

Gráfico de normalidad para los residuales estudentizados:



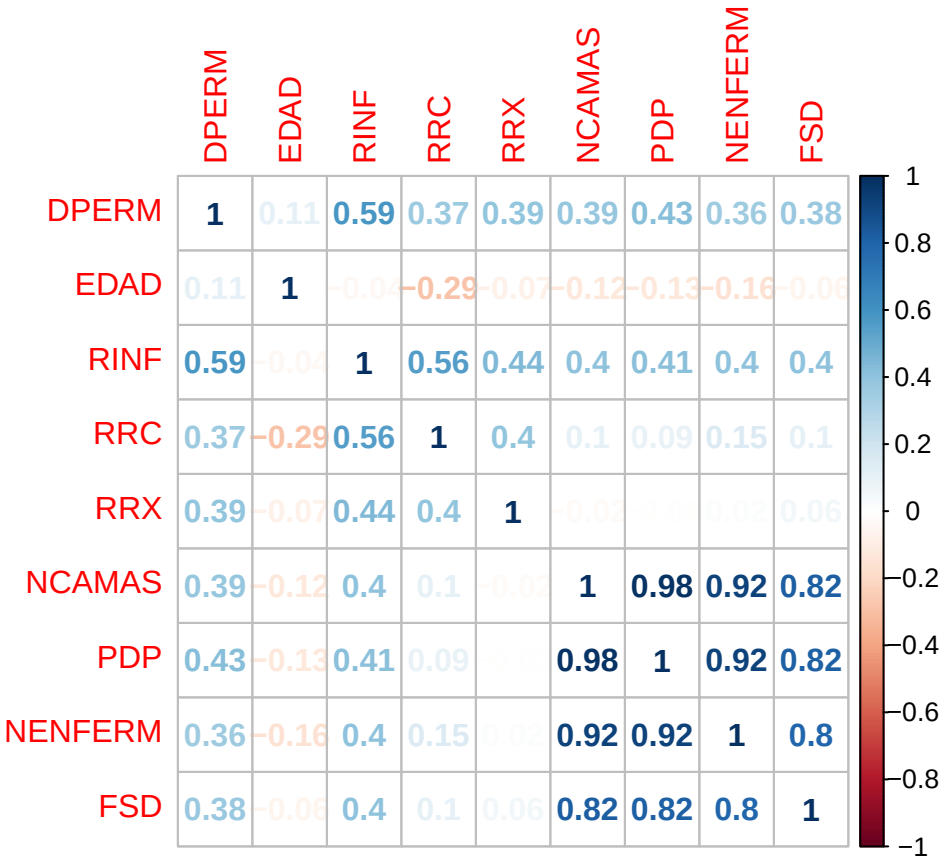
Sin las observaciones mencionadas, la prueba de bondad de ajuste para normalidad de shapiro-wilks indica que los errores estudentizados se distribuyen normalmente. Esto se puede rectificar con el gráfico.

8. Diagnósticos de multicolinealidad:

Diagnósticos de multicolinealidad mediante la Matriz de correlación de las variables predictoras:

	DPERM	EDAD	RINF	RRC	RRX	NCAMAS	PDP	NENFERM	FSD
DPERM	1.000	0.110	0.589	0.366	0.392	0.389	0.425	0.358	0.376
EDAD	0.110	1.000	-0.038	-0.286	-0.073	-0.120	-0.129	-0.157	-0.059
RINF	0.589	-0.038	1.000	0.558	0.440	0.404	0.410	0.402	0.401
RRC	0.366	-0.286	0.558	1.000	0.399	0.101	0.090	0.152	0.103
RRX	0.392	-0.073	0.440	0.399	1.000	-0.021	-0.006	0.016	0.059
NCAMAS	0.389	-0.120	0.404	0.101	-0.021	1.000	0.983	0.923	0.823
PDP	0.425	-0.129	0.410	0.090	-0.006	0.983	1.000	0.923	0.815
NENFERM	0.358	-0.157	0.402	0.152	0.016	0.923	0.923	1.000	0.796
FSD	0.376	-0.059	0.401	0.103	0.059	0.823	0.815	0.796	1.000

Buscamos los pares de variables cuya correlación sea superior 0.9, para facilitar esto se crea un mapa de calor de correlaciones:



Según la matriz de correlaciones, se observan problemas de multicolinealidad al asociar el Número de camas, el Censo promedio diario, el número de enfermas y las Facilidades y servicios disponibles entre sí.

Diagnósticos de multicolinealidad mediante los VIF's:

```
Warning in vif.default(modelo2): No intercept: vifs may not be sensible.
```

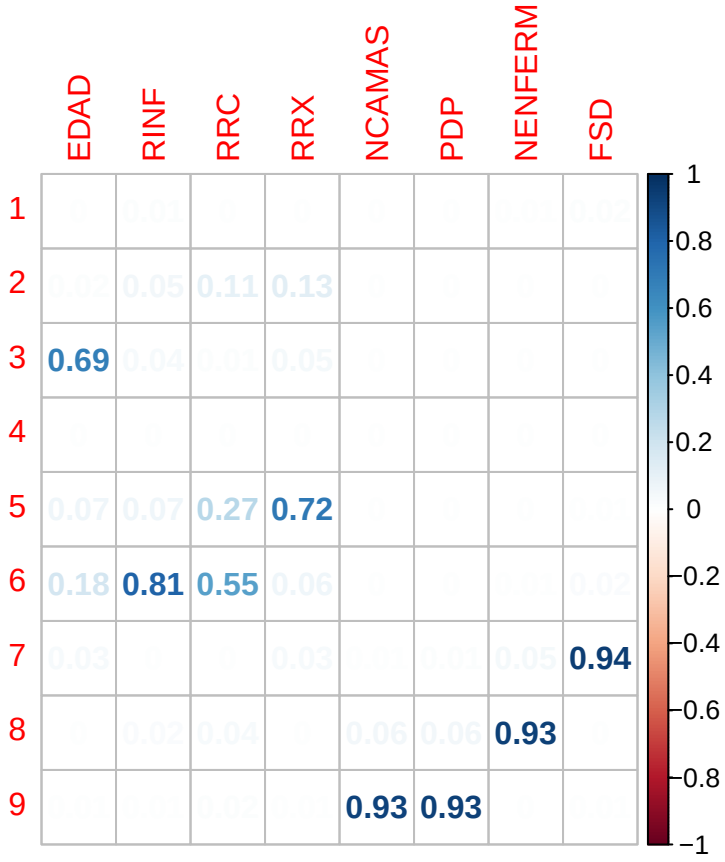
EDAD	RINF	RRC	RRX	NCAMAS	PDP	NENFERM	FSD
1.18	2.15	1.81	1.38	31.9	31.6	7.56	3.29

Se observan factores de inflación de varianza VIF_j superiores a 10 asociados al número de camas y al censo promedio , diágnosticando problemas serios de multicolinealidad en al menos dos variables.

Diagnósticos de multicolinealidad mediante las proporciones de varianza con los datos estandarizados:

Se muestran las proporcioens de varianza para los datos estandarizados:

EDAD	RINF	RRC	RRX	NCAMAS	PDP	NENFERM	FSD
0.002	0.010	0.003	0.001	0.002	0.002	0.007	0.015
0.020	0.051	0.110	0.130	0.001	0.000	0.001	0.003
0.690	0.036	0.011	0.047	0.000	0.000	0.000	0.002
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.068	0.066	0.272	0.716	0.000	0.000	0.000	0.006
0.184	0.806	0.549	0.061	0.000	0.000	0.008	0.022
0.030	0.004	0.000	0.032	0.008	0.011	0.048	0.943
0.000	0.021	0.036	0.005	0.059	0.056	0.934	0.001
0.006	0.006	0.020	0.009	0.930	0.931	0.001	0.008



Según las proporciones de descomposición de varianza, existen problemas de multicolinealidad entre el

riesgo de infección y la Razón de rutina de cultivos, además de el Número de camas y el Censo promedio diario.

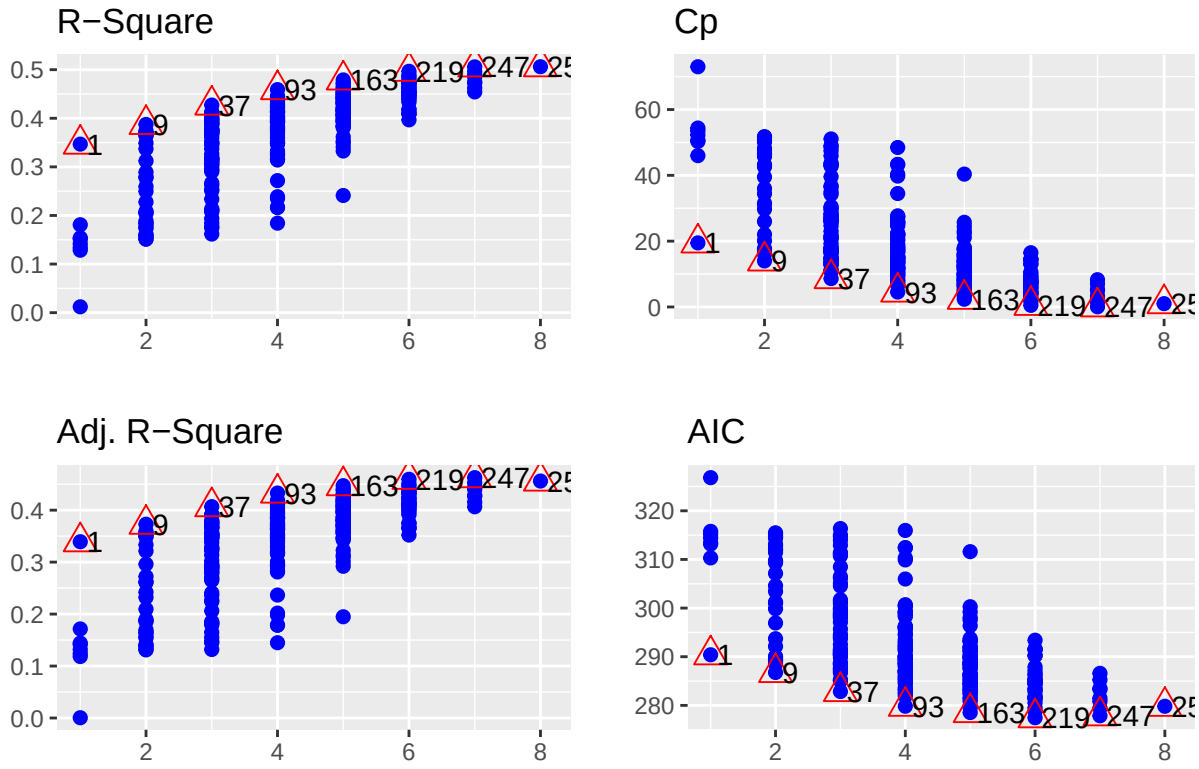
9. Selección de modelo:

Tabla de todas las regresiones posibles, con los datos sin centrar y sin las observaciones con ID igual a 47 y 112:

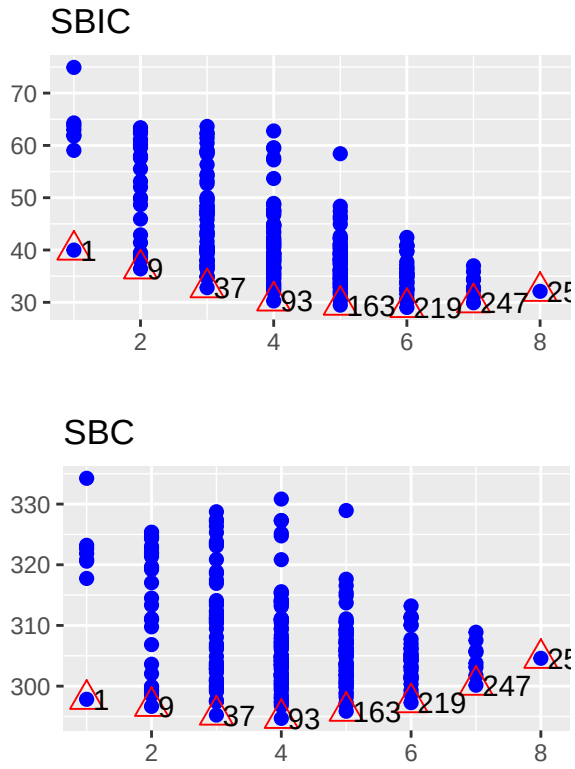
	mindex	n	predictors	rsquare	adjr	predrsq	cp
2	1	1	RINF	0.347	0.339	0.317	20.5
6	2	1	PDP	0.181	0.171	0.145	47.0
4	3	1	RRX	0.154	0.144	0.106	51.3
5	4	1	NCAMAS	0.152	0.142	0.116	51.7
8	5	1	FSD	0.142	0.132	0.102	53.3
3	6	1	RRC	0.134	0.124	0.099	54.5
7	7	1	NENFERM	0.129	0.118	0.087	55.4
1	8	1	EDAD	0.012	0.001	-0.050	74.0
19	9	2	RINF PDP	0.387	0.373	0.347	16.0
18	10	2	RINF NCAMAS	0.374	0.360	0.334	18.1

Gráfica de los modelos que más destacan acorde a su criterio de selección

page 1 of 2



page 2 of 2



Selección de modelo según el R^2_{adj} :

Según el criterio del R^2_{adj} y el principio de parsimonia, se elige el modelo con indice 163 en la tabla de todos los modelos de regresión, ya que el crecimiento en el R^2_{adj} no es tan grande al añadir más de 5 variables.

El modelo con 5 covariables que tiene el R^2_{adj} más alto, es el modelo que está compuesto de las covariables Edad, riesgo de infección, número de camas, censo promedio diario y la Razón de rutina de rayos X del pecho.

Resumen númerico del modelo seleccionado usando el R^2_{adj}

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.552	1.572	1.62	0.108
EDAD	0.061	0.027	2.28	0.025
RINF	0.416	0.112	3.71	0.000
RRX	0.018	0.007	2.53	0.013
NCAMAS	-0.007	0.004	-1.77	0.080
PDP	0.011	0.005	2.43	0.017

R^2 ajustado
0.447

ANOVA del modelo seleccionado usando el R^2_{adj}

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	95.667	5	19.13340	15.0705	1.76486e-10
Error	104.106	82	1.26959		

Selección de modelo según el estadístico C_p

	predictors	n	cp	abs
2	RINF	1	20.46	19.464
19	RINF PDP	2	15.98	13.978
64	RINF RRX PDP	3	11.61	8.611
99	EDAD RINF RRX PDP	4	8.53	4.529
173	EDAD RINF RRX NCAMAS PDP	5	7.34	2.343
219	EDAD RINF RRC RRX NCAMAS PDP	6	6.49	0.492
247	EDAD RINF RRC RRX NCAMAS PDP NENFERM	7	7.04	0.041
255	EDAD RINF RRC RRX NCAMAS PDP NENFERM FSD	8	9.00	1.000

Según el criterio de los C_p el modelo con el menor valor posible en C_p, tal que acorde al principio de parcimonia $|C_p - p|$ es minimo es el modelo que es explicado por la edad, el riesgo de infección, la razón de rutina de cultivos, el número de camas, la razón de rutina de rayos X del pecho y el censo promedio diario. Se muestra una tabla de resumen de parámetros y su respectiva tabla ANOVA.

Resumen númerico del modelo seleccionado usando el criterio de C_p

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.864	1.607	1.16	0.249
EDAD	0.078	0.028	2.75	0.007
RINF	0.305	0.129	2.37	0.020
RRC	0.025	0.015	1.69	0.094
RRX	0.016	0.007	2.26	0.026
NCAMAS	-0.007	0.004	-1.96	0.053
PDP	0.013	0.005	2.68	0.009

R^2 ajustado
0.459

ANOVA del modelo seleccionado usando el C_p

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	99.2286	6	16.53810	13.3233	1.91678e-10
Error	100.5447	81	1.24129		

Selección de modelo según el metodo Stepwise:

Resumen del algoritmo Stepwise:

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	RINF	addition	0.347	0.339	20.4640	290.4006	1.2318
2	PDP	addition	0.387	0.373	15.9780	286.7598	1.1999
3	RRX	addition	0.427	0.407	11.6110	282.8468	1.1672
4	EDAD	addition	0.459	0.433	8.5290	279.8244	1.1412
5	NCAMAS	addition	0.479	0.447	7.3430	278.5239	1.1268
6	RRC	addition	0.497	0.459	6.4920	277.4606	1.1141

Resumen númerico del modelo seleccionado usando el algoritmo stepwise:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.864	1.607	1.16	0.249
RINF	0.305	0.129	2.37	0.020
PDP	0.013	0.005	2.68	0.009
RRX	0.016	0.007	2.26	0.026
EDAD	0.078	0.028	2.75	0.007
NCAMAS	-0.007	0.004	-1.96	0.053

	Estimate	Std. Error	t value	Pr(>|t|)
RRC	0.025	0.015	1.69	0.094

R^2 ajustado
0.459

ANOVA del modelo seleccionado usando el algoritmo stepwise:

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	99.2286	6	16.53810	13.3233	1.91678e-10
Error	100.5447	81	1.24129		

Selección mediante el metodo forward:

Resumen del algoritmo forward:

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	RINF	0.3468	0.3392	20.4638	290.4006	1.2318
2	PDP	0.3874	0.3730	15.9778	286.7598	1.1999
3	RRX	0.4272	0.4067	11.6106	282.8468	1.1672
4	EDAD	0.4590	0.4329	8.5286	279.8244	1.1412
5	NCAMAS	0.4789	0.4471	7.3433	278.5239	1.1268
6	RRC	0.4967	0.4594	6.4920	277.4606	1.1141
7	NENFERM	0.5058	0.4625	7.0406	277.8592	1.1109

Resumen númerico del modelo seleccionado usando el algoritmo forward:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.959	1.604	1.22	0.226
RINF	0.300	0.129	2.33	0.022
PDP	0.014	0.005	2.90	0.005
RRX	0.016	0.007	2.29	0.024
EDAD	0.076	0.028	2.68	0.009
NCAMAS	-0.006	0.004	-1.66	0.101
RRC	0.027	0.015	1.84	0.070
NENFERM	-0.003	0.003	-1.21	0.229

R^2 ajustado
0.463

ANOVA del modelo seleccionado usando el algoritmo forward:

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	101.0417	7	14.43453	11.696	3.73752e-10
Error	98.7316	80	1.23415		

Selección mediante el método backward:

Resumen del algoritmo backward:

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	RINF	0.3468	0.3392	20.4638	290.4006	1.2318
2	PDP	0.3874	0.3730	15.9778	286.7598	1.1999
3	RRX	0.4272	0.4067	11.6106	282.8468	1.1672
4	EDAD	0.4590	0.4329	8.5286	279.8244	1.1412
5	NCAMAS	0.4789	0.4471	7.3433	278.5239	1.1268
6	RRC	0.4967	0.4594	6.4920	277.4606	1.1141
7	NENFERM	0.5058	0.4625	7.0406	277.8592	1.1109

Resumen númerico del modelo seleccionado usando el algoritmo backward:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.959	1.604	1.22	0.226
EDAD	0.076	0.028	2.68	0.009
RINF	0.300	0.129	2.33	0.022
RRC	0.027	0.015	1.84	0.070
RRX	0.016	0.007	2.29	0.024
NCAMAS	-0.006	0.004	-1.66	0.101
PDP	0.014	0.005	2.90	0.005
NENFERM	-0.003	0.003	-1.21	0.229

R^2 ajustado
0.463

ANOVA del modelo seleccionado usando el algoritmo backward:

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	101.0417	7	14.43453	11.696	3.73752e-10
Error	98.7316	80	1.23415		

10. ¿Cuál modelo sugiere para la variable respuesta?

Al analizar el número de variables que arrojan los métodos de selección automáticos se observa un alto número de predictoras para este contexto, por lo que se descartan los modelos sugeridos por estos métodos.

Se observan valores similares para el R^2 ajustado tanto en el modelo propuesto por C_p, cómo para el modelo propuesto por R^2_{adj}, así que por principio de parsimonia se decanta por el modelo propuesto por el R^2_{adj}. Aún así, se encuentran incorformidades con el modelo propuesto, pues en un análisis previo se encontraron problemas de colinealidad asociados a las variables censo promedio diario y número de camas.

Se realiza un ajuste del modelo propuesto y la verificación de sus supuestos:

Ajuste de modelo de regresión lineal multiple para el modelo explicado por la Edad, el riesgo de infección, ratio de rayos X en el pecho, número de camas y censo promedio diario:

Se ajusta un modelo de regresión lineal múltiple con las variables propuestas y se muestra su resumen númerico:

$$\hat{y}_i = 2.5521 + 0.0610 \cdot x_{i1} + 0.4158 \cdot x_{i2} + 0.0176 \cdot x_{i4} - 0.0065 \cdot x_{i5} + 0.0114 \cdot x_{i6}.i = 1, ..., 90.$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.552	1.572	1.62	0.108
EDAD	0.061	0.027	2.28	0.025
RINF	0.416	0.112	3.71	0.000
RRX	0.018	0.007	2.53	0.013
NCAMAS	-0.007	0.004	-1.77	0.080
PDP	0.011	0.005	2.43	0.017

R^2 ajustado
0.447

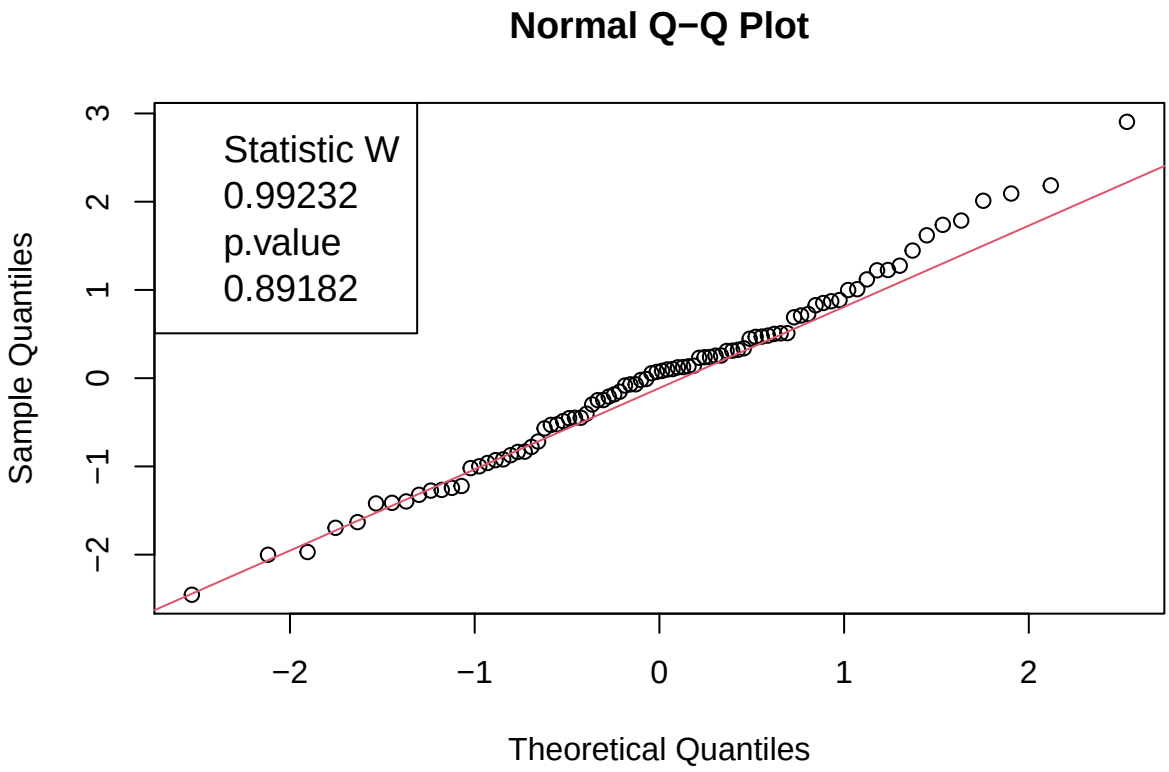
Prueba de hipótesis para la significancia de la regresión del modelo propuesto:

Se evalúa la significancia de la regresión a partir de su tabla ANOVA:

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	95.667	5	19.13340	15.0705	1.76486e-10
Error	104.106	82	1.26959		

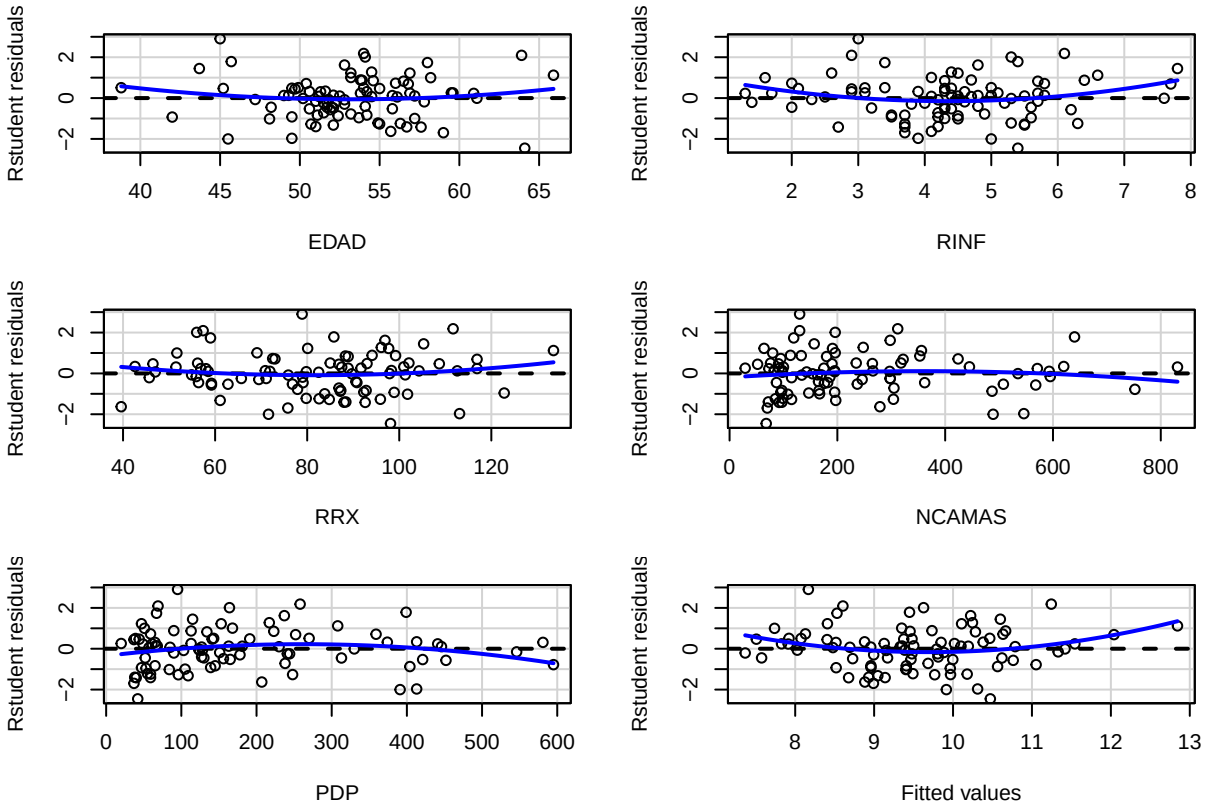
Según el valor p de la prueba F se observa que la regresión sí es significativa.

Verificación del supuesto de normalidad en los residuales:



Se observa un mejor ajuste en la normalidad de los datos, sin embargo, al final de la linea se sigue observando un ligero desajuste. Aún así, se observa un mejor valor en el estadístico de la prueba de Shapiro-Wilk, por lo que se considera que en este caso sí se cumple el supuesto de normalidad en los residuales.

Verificación del supuesto de varianza constante:



Se observa que se cumple aparentemente con el supuesto de varianza constante.