Trabajo 2

David Cardona Duque

9/3/2021

```
datos <- read.csv("datos.csv",
                  encoding = "UTF-8")
source("Functions.R")
library(leaps)
library(perturb)
library(car)
```

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, los datos se encuentran en publicados junto con el este archivo (datos2.txt). La base de datos contiene las siguientes columnas (variables):

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.

Puntos

1. Estime un modelo de regresión lineal múltiple que explique el Riesgo de Infección en términos de todas las variables 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. Comente los resultados.

```
mod=lm(y~x1+x2+x3+x4+x5,data=datos)
```

Tabla resumen regresión

```
summary(mod)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.03944 -0.80043 -0.00266  0.60450  2.23292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.5986009  1.5159559  -0.395 0.694365
## x1           0.2106683  0.0785765   2.681 0.009501 **
## x2           0.0197512  0.0277108   0.713 0.478803
## x3           0.0470925  0.0132888   3.544 0.000779 ***
## x4           0.0105604  0.0073166   1.443 0.154213
## x5           0.0008996  0.0007379   1.219 0.227679
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.035 on 59 degrees of freedom
## Multiple R-squared:  0.4468, Adjusted R-squared:  0.3999
## F-statistic:  9.53 on 5 and 59 DF,  p-value: 1.058e-06
```

Tabla ANOVA regresión

```
myAnova(mod)
```

```
##       Sum_of_Squares DF Mean_Square F_Value     P_value
## Model       51.0453  5    10.20905  9.5301 1.05836e-06
## Error       63.2033 59     1.07124
```

Significancia de la regresión.

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ vs H_1: $\beta_1 \neq 0$ o $\beta_2 \neq$ o $\beta_3 \neq$ o $\beta_4 \neq 0$ o $\beta_5 \neq 0$

$$F_0 = \frac{MSR}{MSE}$$

$$F_0 = \frac{10.20905}{1.07124} = 9.5301$$

$f_{0.05,5,59} = 2.37098$

Note que $F_0 > f_{0.05,5,59}$ por lo tanto con un 95% de significancia al menos un $\beta_j \neq 0$. Lo que significa que por lo menos uno de los parámetros si es significativo en presencia de los otros a la hora de explicar el riesgo de infección.

Significancia de los parámetros individuales e interpretación.

Significancia β_1

H_0: $\beta_1 = 0$ Vs H_1: $\beta_1 \neq 0$

$$t_0 = \frac{\hat{\beta_1}}{Se(\hat{\beta_1})}$$

$$t_0 = \frac{0.2106683}{0.0785765} = 2.68106$$

$t_{0.025,59} = 2.000995$

Note que $t_0 > t_{0.025,59}$ por lo tanto con una significancia de 95% (he inclusive con un 99% si observamos la tabla resumen) podemos afirmar que β_1 es significativo en presencia de los otros parámetros. Su interpretación es que por cada día de más en la variable “Duración de la estadía” la probabilidad promedio estimada de adquirir una infección en el hospital aumenta en promedio 0.2106683%, siempre que las demás predictoras permanezcan constantes.

En adelante todas las hipótesis de significancia se resolverán con la tabla resumen y la información brindada por de la misma

Significancia β_0

Con una significancia de 95% podemos afirmar que β_0 no es significativo en presencia de los otros parámetros. Este coeficiente no es interpretable debido a lo anterior.

Significancia β_2

Con una significancia de 95% podemos afirmar que β_2 no es significativo en presencia de los otros parámetros. Este coeficiente no es interpretable debido a lo anterior.

Significancia β_3

Con una significancia de 99.9% podemos afirmar que β_3 es significativo en presencia de los otros parámetros. La interpretación de este coeficiente es que por un aumento de una cama en la variable “Número de camas” la probabilidad promedio estimada de adquirir una infección en el hospital crece en promedio 0.0470925%, siempre que las demás predictoras permanezcan constantes.

Significancia β_4

Con una significancia de 95% podemos afirmar que β_4 no es significativo en presencia de los otros parametros. Este coeficiente no es interpretable debido a lo anterior.

Significancia β_5

Con una significancia de 95% podemos afirmar que β_5 no es significativo en presencia de los otros parámetros. Este coeficiente no es interpretable debido a lo anterior.

Coeficiente de determinación múltiple R^2

El coeficiente R^2=0.4468 significa que el modelo explica el 44.68% de la variabilidad del “Riesgo de infección”, por lo tanto las variables predictoras elegidas no son las mejores en cuanto a la explicación de la variable de respuesta. Si hacemos una comparación con el coeficiente R^2 **ajustado** el cual es igual a 0.3999 y nos ayuda a vislumbrar como tenemos una penalización por utilizar variables poco significativas de sobra como lo vimos en el punto pasado.

Comentario acerca de los resultados

Desde los análisis anteriores podemos encontrar varias cosas, entre ellas que aunque el modelo es significativo en su generalidad no es un buen modelo en cuanto a predicción porque incluye 3 variables no significativas dentro del mismo además

que el coeficiente múltiple R^2 el cual no penaliza por adherir tales variables es aún muy bajo, por lo tanto esto puede darce debido a que las variables significativas del modelo no tienen una relación lineal muy adecuada con la variable de respuesta "Riesgo de infección" lo cual se puede evidenciar en el poco aumento marginal que aportan estas variables cuando el resto esta constante, esto debido posiblemente a que "Riesgo de infección" depende de otros factores.

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

Las variables con valores p mayores del punto anterior fueron: x2,x4,x5

```
mod1=lm(y~x2+x4+x5,data=datos)
myAnova(mod1)
```

```
##       Sum_of_Squares DF Mean_Square F_Value     P_value
## Model        26.9277  3     8.97590 6.27031 0.000884258
## Error        87.3209 61     1.43149
```

Significancia simultánea del subconjunto como modelo

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ vs H_1: $\beta_1 \neq 0$ o $\beta_2 \neq$ o $\beta_3 \neq 0$

$$F_0 = \frac{MSR}{MSE}$$

$$F_0 = \frac{8.97590}{1.43149} = 6.27031974$$

$f_{0.05,3,61} = 2.75548$

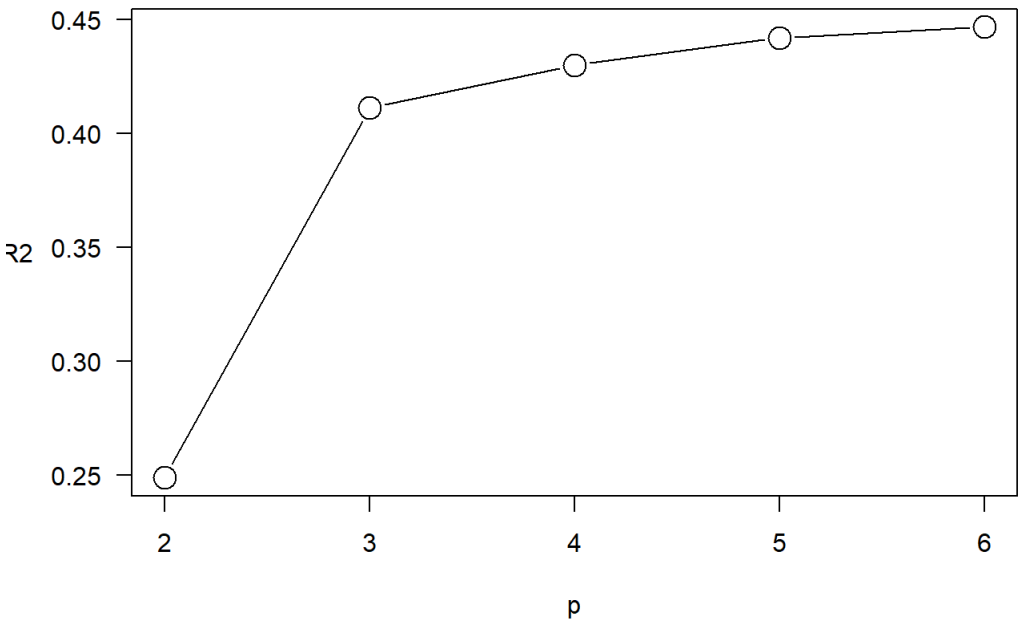
Análisis comparativo mediante todas las regresiones posibles

```
myAllRegTable(mod)
```

```
##    k  R_sq adj_R_sq     SSE     Cp Variables_in_model
## 1  1 0.249    0.237  85.813 19.106                 x3
## 2  1 0.244    0.232  86.376 19.632                 x1
## 3  1 0.182    0.169  93.468 26.252                 x4
## 4  1 0.068    0.053 106.508 38.425                 x5
## 5  1 0.001   -0.014 114.090 45.503                 x2
## 6  2 0.411    0.392  67.257  3.784              x1 x3
## 7  2 0.319    0.297  77.804 13.629              x3 x4
## 8  2 0.311    0.289  78.739 14.503              x1 x4
## 9  2 0.293    0.270  80.764 16.392              x3 x5
## 10 2 0.276    0.253  82.667 18.169              x2 x3
## 11 2 0.258    0.235  84.721 20.087              x1 x5
## 12 2 0.248    0.224  85.877 21.166              x1 x2
## 13 2 0.234    0.209  87.519 22.698              x4 x5
## 14 2 0.182    0.156  93.416 28.203              x2 x4
## 15 2 0.071    0.041 106.102 40.046              x2 x5
## 16 3 0.430    0.402  65.108  3.778           x1 x3 x4
## 17 3 0.422    0.393  66.088  4.693           x1 x3 x5
## 18 3 0.415    0.386  66.850  5.404           x1 x2 x3
## 19 3 0.359    0.327  73.279 11.405           x3 x4 x5
## 20 3 0.336    0.303  75.868 13.822           x2 x3 x4
## 21 3 0.328    0.295  76.791 14.684           x1 x4 x5
## 22 3 0.325    0.292  77.090 14.963           x2 x3 x5
## 23 3 0.314    0.280  78.400 16.186           x1 x2 x4
## 24 3 0.261    0.224  84.459 21.842           x1 x2 x5
## 25 3 0.236    0.198  87.321 24.514           x2 x4 x5
## 26 4 0.442    0.405  63.748  4.508        x1 x3 x4 x5
## 27 4 0.433    0.395  64.795  5.486        x1 x2 x3 x4
## 28 4 0.427    0.389  65.435  6.083        x1 x2 x3 x5
## 29 4 0.379    0.338  70.904 11.188        x2 x3 x4 x5
## 30 4 0.329    0.284  76.656 16.558        x1 x2 x4 x5
## 31 5 0.447    0.400  63.203  6.000     x1 x2 x3 x4 x5
```

Criterio R^2

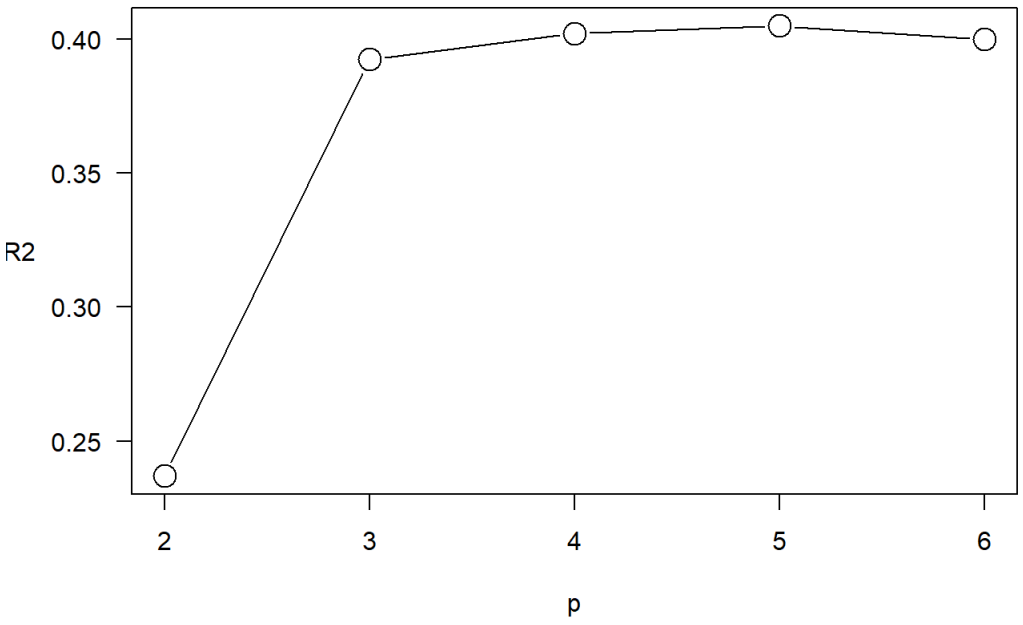
```
myR2_criterion(mod)
```




```
## Models are Indexed in rows
##  k p       R2 Variables.in.model
##  1 2 0.2488942               x3
##  2 3 0.4113095            x1 x3
##  3 4 0.4301181         x1 x3 x4
##  4 5 0.4420276       x1 x3 x4 x5
##  5 6 0.4467911     x1 x2 x3 x4 x5
```

Criterio R^2 ajustado

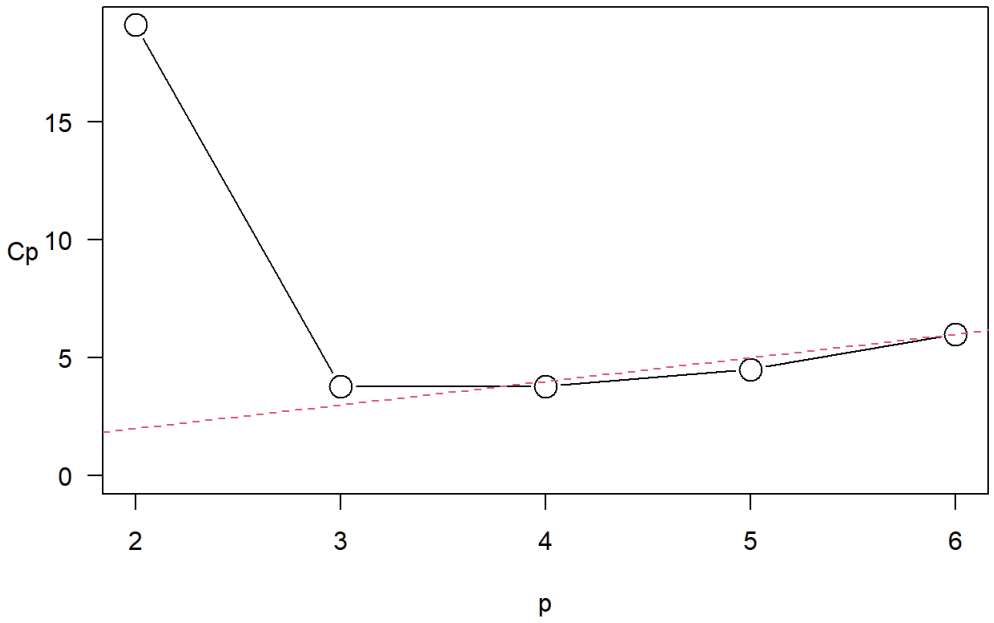
```
myAdj_R2_criterion(mod)
```




```
## Models are Indexed in rows
##  k p     adjR2 Variables.in.model
##  1 2 0.2369719               x3
##  2 3 0.3923195            x1 x3
##  3 4 0.4020911         x1 x3 x4
##  4 5 0.4048294       x1 x3 x4 x5
##  5 6 0.3999090     x1 x2 x3 x4 x5
```

Criterio C_p

```
myCp_criterion(mod)
```




```
## Models are Indexed in rows
##  k p       Cp Variables.in.model
##  1 2 19.105797               x3
##  2 3  3.784129            x1 x3
##  3 4  3.778185         x1 x3 x4
##  4 5  4.508030      x1 x3 x4 x5
##  5 6  6.000000   x1 x2 x3 x4 x5
```

Podemos notar que sin la existencia de otras variables al menos una de las 3 es significativa con una seguridad del 95%, ya que $F_0 > f_{0.05,3,61}$, ahora bien si hacemos un análisis comparativo desde la tabla de todas las regresiones podemos observar que bajo el criterio de R^2, R^2 **ajustado** y C_p en todos existen por lo menos una variable del subconjunto que debería entrar en el modelo, por ejemplo observamos en la tabla de todas las regresiones entendemos que el mejor modelo basado en el criterio C_p sería el que involucra a las variables x1 x3 x4, lo mismo al analizar bajo el criterio R^2. Por lo tanto no es posible descartar todas las variables del subconjunto, aunque si es posible descartar x2 y x5 con base a los criterios anteriores pues su aporte al R^2 y al R^2 **ajustado** es mínimo además que su impacto en el $|C_p - p|$ es aumentarlo.

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

Son los coeficientes $\beta_1 = \beta_4$, $\beta_3 = \beta_2$, $\beta_5 = 0$?

H_0: $\beta_1 = \beta_4$, $\beta_3 = \beta_2$, $\beta_5 = 0$ vs H_1: $\beta_1 \neq \beta_4$ o $\beta_3 \neq \beta_2$ o $\beta_5 \neq 0$

```
b0=c(0,0,0)
b1=c(1,0,0)
b2=c(0,-1,0)
b3=c(0,1,0)
b4=c(-1,0,0)
b5=c(0,0,1)
matriz_L <-cbind(b0,b1,b2,b3,b4,b5)
```

Matriz **L**

```
matriz_L
```

```
##      b0 b1 b2 b3 b4 b5
## [1,]  0  1  0  0 -1  0
## [2,]  0  0 -1  1  0  0
## [3,]  0  0  0  0  0  1
```

Modelo reducido

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i2} + \beta_3 X_{i3} + \beta_1 X_{i4} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 (X_{i1} + X_{i4}) + \beta_3 (X_{i3} + X_{i2}) + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 (Z_{i1,4}) + \beta_3 (Z_{i2,3}) + \varepsilon_i$$

```
attach(datos)
Z14= x1 + x4
Z23= x2+x3
modR=lm(y~Z14+Z23)
```

Tabla ANOVA modelo reducido

```
anova(modR)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Z14        1 23.705 23.7053  19.647 3.872e-05 ***
## Z23        1 15.737 15.7370  13.043 0.0006104 ***
## Residuals 62 74.806  1.2066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tabla ANOVA modelo completo

```
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 27.872 27.8721 26.0185 3.770e-06 ***
## x2         1  0.499  0.4995  0.4662    0.4974
## x3         1 19.028 19.0275 17.7621 8.703e-05 ***
## x4         1  2.054  2.0542  1.9176    0.1713
## x5         1  1.592  1.5919  1.4861    0.2277
## Residuals 59 63.203  1.0712
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Expresión para el estadístico de prueba.

$$F_0 = \frac{\frac{SSE(M.R) - SSE(M.C)}{m}}{MSE(M.C)}$$

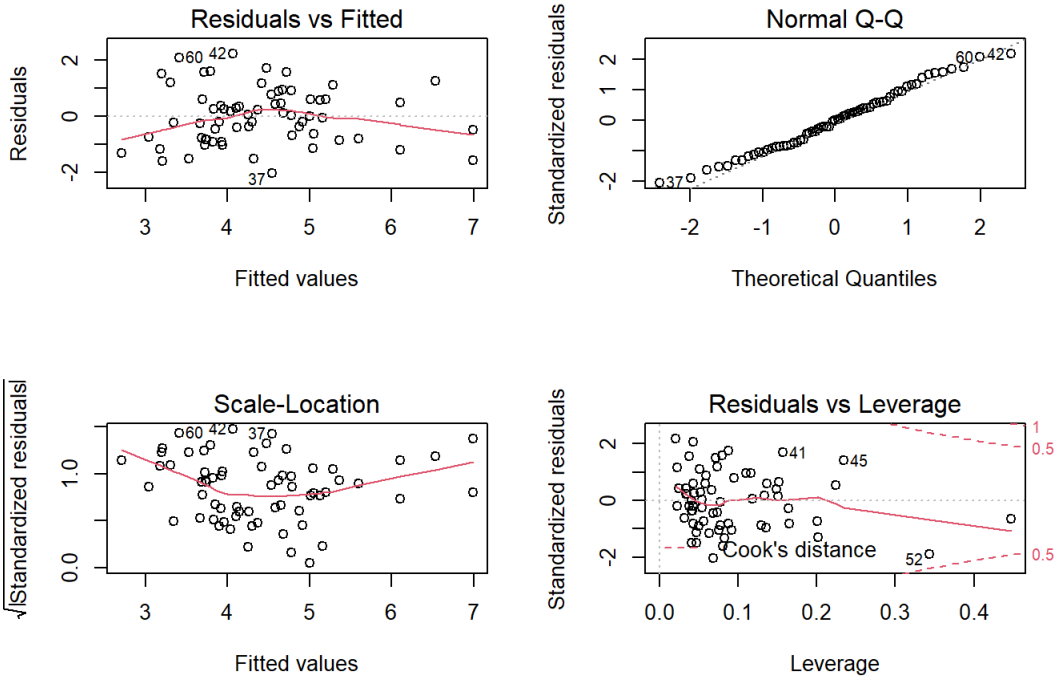
$$F_0 = \frac{\frac{74.806 - 63.203}{3}}{1.0712} = 3.61059$$

$$f_{0.05,3,59} = 2.76077$$

Note que $F_0 > f_{0.05,3,59}$, entonces se rechaza H_0 con un nivel de significancia del 95%, lo cual nos lleva a que por lo menos una hipótesis alternativa es cierta.

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 este modelo?. Argumente.

```
par(mfrow=c(2,2))
plot(mod)
```




```
##Validacion media 0
ei=mod$residuals
round(mean(ei),0)
```

```
## [1] 0
```

Validación varianza constante.

En la grafica “Residuals vs Fitted” se puede observar como la varianza tiene una tendencia no constante ni lineal esto debido a que los datos tienen una variación distinta respecto a su media por cada observación.

Validación normalidad

```
shapiro.test(ei)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ei
## W = 0.98462, p-value = 0.5979
```

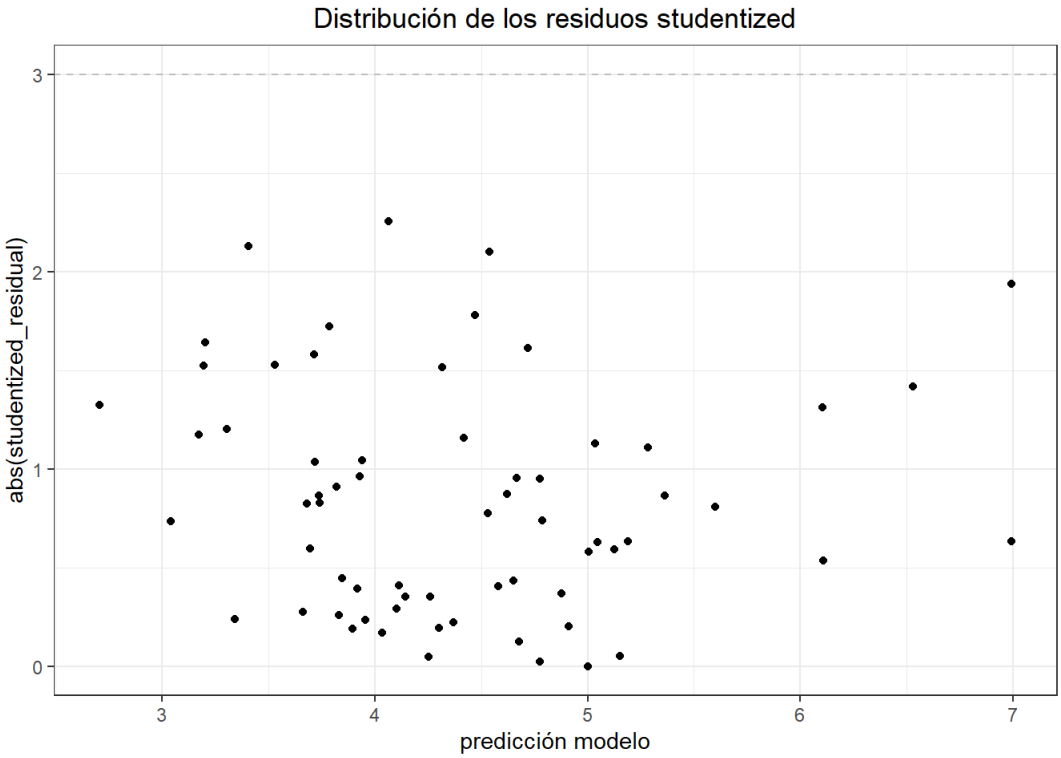
En la grafica “Normal Q-Q” se observa como los residuales tienen una tendencia normal, pero se tiene una desviación en las partes extremas de la gráfica. Luego gracias al test de Shapiro Wilk se concluye que no hay evidencia para decir que no existe normalidad la distribución de los residuales, puesto que el valor p de la prueba es considerablemente grande por lo tanto no se rechaza la hipótesis inicial de que se distribuían normal

Independencia de los errores

Dado que estos registros no corresponden a datos en el tiempo no se tiene un orden temporal para realizar la validación de este supuesto. Se valida por definición del tipo de datos de corte transversal.

Observaciones atípicas

```
library(dplyr)
library(ggplot2)
datos$studentized_residual <- rstudent(mod)
ggplot(data = datos, aes(x = predict(mod), y = abs(studentized_residual))) +
geom_hline(yintercept = 3, color = "grey", linetype = "dashed") +
# se identifican en rojo observaciones con residuos estandarizados absolutos > 3
geom_point(aes(color = ifelse(abs(studentized_residual) > 3, 'red', 'black'))) +
scale_color_identity() +
labs(title = "Distribución de los residuos studentized",
    x = "predicción modelo") +
theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```




```
which(abs(datos$studentized_residual) > 3)
```

```
## integer(0)
```

Si $|r_i| > 3$ entonces i es un punto atípico

No se observa ninguna observación atípica en el modelo.

Puntos influyentes y de balanceo

```
t1<-predict(mod,se.fit=T)
t2<-round(residuals(mod),4)
t3<-round(cooks.distance(mod),4)
t4<-round(hatvalues(mod),4)
t5<-round(dffits(mod),4)
restud<-round(rstudent(mod),4)
est_salida <- data.frame(datos$y,yhat=round(t1$fit,4),
se.yhat=round(t1$se.fit,6),residuals=t2
,res.estud=restud,Cooks.D=t3,hii.value=t4,Dffits=t5)
(est_salida)
```

```
##    datos.y   yhat  se.yhat residuals res.estud Cooks.D hii.value  Dffits
## 1      3.7 4.1143 0.281627   -0.4143   -0.4130  0.0023    0.0740 -0.1168
## 2      2.8 4.3171 0.223854   -1.5171   -1.5179  0.0184    0.0468 -0.3362
## 3      4.2 3.9570 0.193758    0.2430    0.2371  0.0003    0.0350  0.0452
## 4      6.2 4.4732 0.306929    1.7268    1.7787  0.0490    0.0879  0.5523
## 5      5.7 4.7749 0.352663    0.9251    0.9499  0.0198    0.1161  0.3443
## 6      4.5 4.1448 0.265863    0.3552    0.3525  0.0015    0.0660  0.0937
## 7      1.6 3.2059 0.293548   -1.6059   -1.6411  0.0382    0.0804 -0.4854
## 8      5.1 5.1539 0.287489   -0.0539   -0.0537  0.0000    0.0772 -0.0155
## 9      4.1 3.8323 0.218296    0.2677    0.2625  0.0005    0.0445  0.0566
## 10     4.4 4.1014 0.237148    0.2986    0.2941  0.0008    0.0525  0.0692
## 11     5.0 4.5792 0.162106    0.4208    0.4088  0.0007    0.0245  0.0648
## 12     4.3 3.6955 0.249986    0.6045    0.5986  0.0037    0.0583  0.1490
## 13     5.3 3.7137 0.201230    1.5863    1.5822  0.0160    0.0378  0.3136
## 14     4.8 4.6769 0.402158    0.1231    0.1280  0.0005    0.1510  0.0540
## 15     4.4 5.0460 0.185804   -0.6460   -0.6312  0.0022    0.0322 -0.1152
## 16     5.3 4.5319 0.318533    0.7681    0.7774  0.0106    0.0947  0.2514
## 17     2.9 3.7425 0.213885   -0.8425   -0.8298  0.0051    0.0427 -0.1753
## 18     4.3 4.2522 0.356599    0.0478    0.0488  0.0001    0.1187  0.0179
## 19     2.0 3.1731 0.259816   -1.1731   -1.1747  0.0154    0.0630 -0.3046
## 20     2.7 3.7208 0.314166   -1.0208   -1.0358  0.0181    0.0921 -0.3300
## 21     5.6 4.6666 0.343458    0.9334    0.9553  0.0188    0.1101  0.3360
## 22     4.1 4.7862 0.463640   -0.6862   -0.7387  0.0230    0.2007 -0.3701
## 23     6.6 6.1065 0.490120    0.4935    0.5380  0.0141    0.2242  0.2893
## 24     5.1 4.6535 0.189494    0.4465    0.4358  0.0011    0.0335  0.0812
## 25     4.5 5.3635 0.287627   -0.8635   -0.8667  0.0105    0.0772 -0.2507
## 26     4.3 3.9189 0.398923    0.3811    0.3962  0.0046    0.1486  0.1655
## 27     6.5 6.9911 0.692102   -0.4911   -0.6350  0.0549    0.4471 -0.5711
## 28     2.9 3.6824 0.421061   -0.7824   -0.8252  0.0226    0.1655 -0.3675
## 29     4.5 3.3066 0.280221    1.1934    1.2022  0.0189    0.0733  0.3381
## 30     4.9 6.1056 0.464484   -1.2056   -1.3114  0.0714    0.2014 -0.6586
## 31     5.6 5.0062 0.213415    0.5938    0.5831  0.0025    0.0425  0.1229
## 32     3.0 3.9295 0.380628   -0.9295   -0.9652  0.0243    0.1352 -0.3817
## 33     5.7 5.1266 0.381952    0.5734    0.5928  0.0093    0.1362  0.2354
## 34     5.0 5.0027 0.208808   -0.0027   -0.0026  0.0000    0.0407 -0.0005
## 35     2.9 3.8205 0.230729   -0.9205   -0.9110  0.0073    0.0497 -0.2083
## 36     4.5 4.8777 0.208361   -0.3777   -0.3698  0.0010    0.0405 -0.0760
## 37     2.5 4.5394 0.270130   -2.0394   -2.0993  0.0508    0.0681 -0.5676
## 38     3.4 3.8490 0.271107   -0.4490   -0.4464  0.0025    0.0686 -0.1212
## 39     5.8 5.1913 0.403037    0.6087    0.6353  0.0121    0.1516  0.2686
## 40     4.8 5.6004 0.307301   -0.8004   -0.8075  0.0106    0.0882 -0.2511
## 41     5.4 3.7884 0.409678    1.6116    1.7237  0.0890    0.1567  0.7429
## 42     6.3 4.0671 0.146814    2.2329    2.2535  0.0163    0.0201  0.3229
## 43     6.3 4.7184 0.291469    1.5816    1.6141  0.0364    0.0793  0.4737
## 44     3.4 3.6644 0.418980   -0.2644   -0.2772  0.0025    0.1639 -0.1227
## 45     7.8 6.5282 0.501587    1.2718    1.4167  0.1010    0.2349  0.7849
## 46     6.4 5.2820 0.234490    1.1180    1.1112  0.0111    0.0513  0.2585
## 47     4.6 4.3703 0.191635    0.2297    0.2240  0.0003    0.0343  0.0422
## 48     3.1 3.3435 0.241952   -0.2435   -0.2400  0.0006    0.0546 -0.0577
## 49     4.1 4.3027 0.156387   -0.2027   -0.1965  0.0002    0.0228 -0.0300
## 50     2.9 3.9397 0.285121   -1.0397   -1.0458  0.0149    0.0759 -0.2997
## 51     4.7 3.1953 0.276387    1.5047    1.5255  0.0291    0.0713  0.4227
## 52     5.4 6.9902 0.606346   -1.5902   -1.9397  0.3130    0.3432 -1.4021
## 53     4.8 4.7744 0.240466    0.0256    0.0252  0.0000    0.0540  0.0060
## 54     2.3 3.0423 0.246426   -0.7423   -0.7356  0.0055    0.0567 -0.1803
## 55     2.0 3.5318 0.209103   -1.5318   -1.5281  0.0162    0.0408 -0.3152
## 56     5.5 4.6228 0.252506    0.8772    0.8721  0.0081    0.0595  0.2194
## 57     1.4 2.7058 0.298558   -1.3058   -1.3261  0.0263    0.0832 -0.3995
## 58     4.7 4.9103 0.201498   -0.2103   -0.2055  0.0003    0.0379 -0.0408
## 59     3.9 4.2621 0.211224   -0.3621   -0.3547  0.0009    0.0416 -0.0739
## 60     5.5 3.4066 0.215746    2.0934    2.1291  0.0324    0.0435  0.4538
## 61     3.7 3.8963 0.213522   -0.1963   -0.1922  0.0003    0.0426 -0.0405
## 62     3.9 5.0371 0.224410   -1.1371   -1.1280  0.0104    0.0470 -0.2505
## 63     4.2 4.0350 0.378101    0.1650    0.1698  0.0008    0.1335  0.0667
## 64     2.9 3.7372 0.372875   -0.8372   -0.8652  0.0187    0.1298 -0.3342
## 65     5.6 4.4172 0.156936    1.1828    1.1596  0.0052    0.0230  0.1779
```

Si $h_{ii} > \frac{2p}{n}$ =0.1846 entonces i es un punto de balanceo

Si $|DFFTITS_i| > 2\sqrt{\frac{2p}{n}}$ =0.6076 entonces i es un punto influyente

Puntos de balanceo

```
which((est_salida$hii.value) > 0.1846)
```

```
## [1] 22 23 27 30 45 52
```

Las observaciones en estas posiciones superan los valores aceptables de h_{ii} y por lo tanto son observaciones de balanceo

Puntos influyentes

```
which(abs(est_salida$Dffits) > 0.6076)
```

```
## [1] 30 41 45 52
```

Las observaciones en estas posiciones superan los valores aceptables de

$$|DFFITS_i|$$

y por lo tanto son observaciones influyentes

Comentario acerca de la validez del modelo

Desde la validación de supuestos el único que el modelo no cumple es el de varianza constante en los residuales, lo cual se ocasiona posiblemente por la presencia de puntos influyentes y hace que se pierda eficiencia y confiabilidad en el modelo debido a que se desvanece la efectividad en el estimador mínimo cuadrático. Por otro lado se encontraron varios puntos problemáticos de 2 tipos, los cuales generan situaciones no deseadas en el modelo, entre ellas, los puntos de balanceo (de los cuales se encontraron 6) afectan los resultados del coeficiente R^2 generando una falsa ilusión de explicación por parte de las variables predictoras al "Riesgo de infección", además los puntos influyentes (de los cuales se encontraron 4) jalan el modelo en su dirección y tienen un mayor efecto sobre la recta de la regresión originando errores en las predicciones sobre "Riesgo de infección". Por lo cual la validez general del modelo es dudosa gracias a los problemas anteriores, debido a esto las predicciones y resultados del mismo no deben tomarse como válidas, estudios más exhaustivos consistirían en rehacer el modelo sin las observaciones problemáticas y garantizando una varianza constante para ver el impacto.

5. Verificar la presencia de multicolinealidad usando gráficos y/o indicadores apropiados.

Con base al indicador numérico VIF

```
myCoefficients(mod,datos)
```

```
## Estimated and standardized coefficients, their 95% CI's and VIF's
```

```
##               Estimation   Coef.Std   Limit_2.5% Limit_97.5%      Vif
## (Intercept) -0.5986008691 0.00000000 -3.632021689 2.434819951 0.000000
## x1           0.2106683091 0.30449605  0.053437116 0.367899502 1.375666
## x2           0.0197512078 0.07487653 -0.035697988 0.075200403 1.176970
## x3           0.0470924905 0.38685413  0.020501581 0.073683400 1.270949
## x4           0.0105603543 0.15949429 -0.004080114 0.025200822 1.302309
## x5           0.0008995893 0.12518477 -0.000577038 0.002376217 1.124675
```

Según los indicadores VIF del modelo se puede concluir que no existen problemas de multicolinealidad, puesto que todos son menores a 5.

Con base a los indicadores numéricos Índices de condición y Proporción de descomposición de varianza

```
myCollinDiag(mod)
```

```
## Collinearity Diagnostics
##                         Variance Decomposition Proportions
```

```
##   Eigen_Value Condition_Index Intercept      x1       x2       x3       x4
## 1   5.4018256        1.000000  0.000239 0.000882 0.000250 0.006425 0.001403
## 2   0.3025937        4.225132  0.000186 0.000115 0.000186 0.155729 0.003353
## 3   0.2365402        4.778789  0.002390 0.004999 0.003835 0.640772 0.004296
## 4   0.0343352       12.542973  0.022346 0.003719 0.028546 0.114363 0.920851
## 5   0.0207865       16.120536  0.039926 0.976147 0.029334 0.004189 0.065410
## 6   0.0039187       37.127569  0.934914 0.014139 0.937849 0.078522 0.004686
##         x5
## 1 0.007901
## 2 0.776606
## 3 0.124325
## 4 0.002073
## 5 0.067694
## 6 0.021401
```

```
myCollinDiag(mod,center = T)
```

```
## Collinearity Diagnostics (intercept adjusted)
##                       Variance Decomposition Proportions
```

```
##   Eigen_Value Condition_Index      x1      x2      x3      x4      x5
## 1     1.73237        1.000000 0.132674 0.000165 0.107417 0.142121 0.064845
## 2     1.23968        1.182127 0.084396 0.443369 0.131654 0.001073 0.000002
## 3     0.98749        1.324506 0.009577 0.028267 0.056622 0.136118 0.637512
## 4     0.54613        1.781035 0.013604 0.206505 0.645686 0.572581 0.002663
## 5     0.49433        1.872015 0.759749 0.321693 0.058621 0.148107 0.294978
```

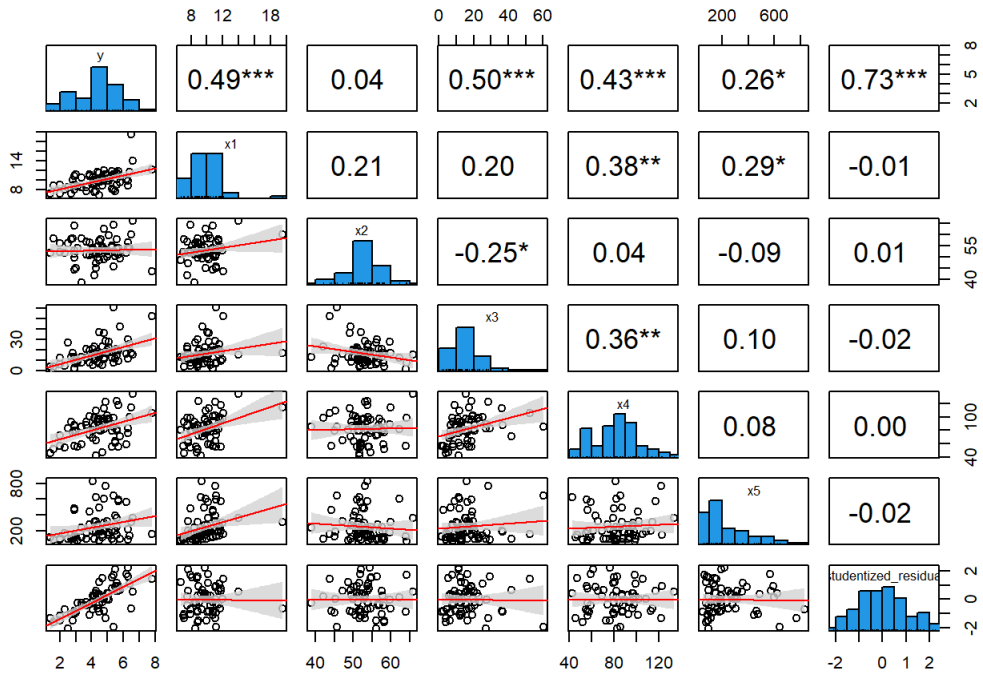
Con base a los Índices de condición en el modelo se puede concluir que existen 2 problemas de multicolinealidad moderados y un problema severo.

Con base a la Proporción de descomposición de varianza se puede concluir que existe multicolinealidad entre las variables "Número de camas"(x3) y "Censo promedio diario"(x4) puesto que sus π_{ij} son mayores a 0.5 y están asociados a un mismo valor propio, más específicamente $\pi_{3,5}$ y $\pi_{3,6}$.

Con base a la matriz de correlación

```
library(psych)

pairs.panels(datos,
         smooth = FALSE,      # Si TRUE, dibuja ajuste suavizados de tipo loess
         scale = FALSE,      # Si TRUE, escala la fuente al grado de correlación
         density = FALSE,     # Si TRUE, añade histogramas y curvas de densidad
         ellipses = FALSE,    # Si TRUE, dibuja elipses
         method = "pearson", # Método de correlación (también "spearman" o "kendall")
         pch = 21,           # Símbolo pch
         lm = TRUE,          # Si TRUE, dibuja un ajuste lineal en lugar de un ajuste LOESS
         cor = TRUE,         # Si TRUE, agrega correlaciones
         jiggle = FALSE,     # Si TRUE, se añade ruido a los datos
         factor = 2,         # Nivel de ruido añadido a los datos
         hist.col = 4,       # Color de los histogramas
         stars = TRUE,       # Si TRUE, agrega el nivel de significación con estrellas
         ci = TRUE)          # Si TRUE, añade intervalos de confianza a los ajustes
```



Con base a la matriz de correlación se puede concluir que no existen indicios de problemas de multicolinealidad pues ningún valor entre las variables predictoras es mayor o igual a 0.5.