Parte 1 según Alejo

Alejandro Salazar Mejía

12/5/2021

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.2
```

```
## Loading required package: carData
```

```r
library(perturb)
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.0.4
```

```r
library(olsrr)
```

```
## Warning: package 'olsrr' was built under R version 4.0.5
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
library(knitr)
library(rsm)
```

```
## Warning: package 'rsm' was built under R version 4.0.4
```

Lectura de datos

```r
datos <- read.table("APC1modifm3.csv", header = T, sep = ";", dec = ",",
                    colClasses = c(rep("numeric",7),
                                   "factor",rep("numeric",3),"factor"))
attach(datos)
```

Parte I:

Antes de comenzar 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).

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

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

```r
cond <- names(datos) != c("ID", "AEM", "REGION")
numvar <- names(datos)[cond]
datosNumericos <- datos[numvar]
```

```r
miscoeficientes <- function(modeloreg, datosreg){
    coefi <- coef(modeloreg)
    datos2 <- as.data.frame(scale(datosreg))
    coef.std <- c(0, coef( lm( update( formula(modeloreg) ,~.+0 ),datos2 ) ) )
    limites <- confint(modeloreg,level = 0.95)
    vifs <- c(0,vif(modeloreg))
    resul <- data.frame(
        Estimación = coefi,Limites = limites,Vif = vifs,Coef.Std = coef.std)
    resul
}
```

Punto 1:

```
# Ajuste del modelo de regresión lineal múltiple
modelo <- lm(DPERM ~ EDAD+RINF+RRC+RRX+NCAMAS+PDP+NENFERM+FSD)

miscoefs <- miscoeficientes(modelo, datosNumericos)
summaryModelo <- summary(modelo)

kable(miscoefs["Estimación"], caption = "Tabla de parámetros ajustados")
```

Table 1: Tabla de parámetros ajustados

	Estimación
(Intercept)	-0.2084670
EDAD	0.1043806
RINF	0.3352223
RRC	0.0287063
RRX	0.0209817
NCAMAS	-0.0106992
PDP	0.0223642
NENFERM	-0.0060256
FSD	0.0041605

$$
\begin{aligned}
&\textit{Ecuación ajustada}: \\
\widehat{y}_i = &-0.208467 + 0.104381 \cdot x_{i1} \\
&+0.335222 \cdot x_{i2} \\
&+0.028706 \cdot x_{i3} \\
&+0.020982 \cdot x_{i4} \\
&-0.010699 \cdot x_{i5} \\
&+0.022364 \cdot x_{i6} \\
&-0.006026 \cdot x_{i7} \\
&+0.004160 \cdot x_{i8}
\end{aligned}
$$

```r
# Anova del modelo
# Se puede obtener de summaryModelo
summaryModelo
```

```
##
## Call:
## lm(formula = DPERM ~ EDAD + RINF + RRC + RRX + NCAMAS + PDP +
##     NENFERM + FSD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6716 -0.8945 -0.0621  0.7941  6.4488
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.208467   1.945083  -0.107  0.91491
## EDAD         0.104381   0.034014   3.069  0.00292 **
## RINF         0.335222   0.155597   2.154  0.03418 *
## RRC          0.028706   0.017775   1.615  0.11020
## RRX          0.020982   0.008507   2.466  0.01576 *
## NCAMAS      -0.010699   0.004456  -2.401  0.01863 *
## PDP          0.022364   0.004881   4.582 1.65e-05 ***
## NENFERM     -0.006026   0.002932  -2.055  0.04308 *
## FSD          0.004160   0.019548   0.213  0.83199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.362 on 81 degrees of freedom
## Multiple R-squared:  0.5914, Adjusted R-squared:  0.551
## F-statistic: 14.65 on 8 and 81 DF,  p-value: 5.042e-13
```

De los resultados arrojados por R, rescatamos la siguiente línea:

F-statistic: 14.65 on 8 and 81 DF, p-value: 5.042e-13

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

```r
kable(data.frame("R squared" = summaryModelo$r.squared))
```

R.squared
0.591382

Aproximadamente el 60% de la variabilidad de la longitud de permanencia es explicada por el modelo. Este porcentaje no es tan alto, lo que podría indicar una carencia de ajuste, es decir, las covariables involucradas hacen que el modelo no se ajuste lo suficiente a los datos reales.

Punto 2:

Se presenta la tabla de coeficientes estandarizados y además el valor absoluto de estos coeficientes ordenados de menor a mayor.

```
kable(miscoefs["Coef.Std"], caption = "Tabla de coeficientes Estandarizados")
```

Table 3: Tabla de coeficientes Estandarizados

	Coef.Std
(Intercept)	0.0000000
EDAD	0.2354129
RINF	0.2254678
RRC	0.1514452
RRX	0.2074015
NCAMAS	-0.9777531
PDP	1.6801779
NENFERM	-0.3839389
FSD	0.0275649

```
stCoefs <- miscoefs$Coef.Std
names(stCoefs) <- row.names(miscoefs)
sort(abs(stCoefs))
```

```
## (Intercept)         FSD         RRC         RRX        RINF        EDAD
##   0.0000000   0.0275649   0.1514452   0.2074015   0.2254678   0.2354129
##     NENFERM      NCAMAS         PDP
##   0.3839389   0.9777531   1.6801779
```

De esta última lista ordenada podemos conluir que las variables que más aportan a la longitud de permanencia son Censo promedio diario, Número de camas y tal vez Número de enfermeras. El resto de variables parece no aportar mucho

Punto 3:

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

$$H_0 : \beta_j = 0 \ \ vs. \ \ H_1 : \beta_j \neq 0 \ ,$$

$$T_0 = \frac{\widehat{\beta}_j}{s.e(\widehat{\beta}_j)} \sim t_{81} \ ,$$

$$p\text{-}value = P(|t_{81}| > |T_0|)).$$

```
pruebast <- summaryModelo$coefficients[-1,-2]
kable(pruebast)
```

	Estimate	t value	Pr(>|t|)
EDAD	0.1043806	3.0687638	0.0029235
RINF	0.3352223	2.1544291	0.0341775
RRC	0.0287063	1.6150188	0.1101955
RRX	0.0209817	2.4662932	0.0157625
NCAMAS	-0.0106992	-2.4011396	0.0186350
PDP	0.0223642	4.5822850	0.0000165
NENFERM	-0.0060256	-2.0552029	0.0430819
FSD	0.0041605	0.2128384	0.8319878

En la anterior tabla, los nombres de las filas corresponden al de la variable explicatoria asociada a cada parámetro. La columna Estimate corresponde a la estimación del parámetro, $\widehat{\beta}_j$; t value al valor del estadítico T_0 y Pr(>|t|) al p-value de la respectiva prueba de hipótesis.

- **Prueba F para test lineal general de la variable PDP**

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

Estadístico de prueba y distribución:

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

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

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

Cálculo de valor p:

$$p\text{-}value = P(f_{1,81} > F_{06})$$

```r
lh1 <- linearHypothesis(modelo, "PDP = 0")
lh1[1, 3:6] <- c(" ", " ", " ", " ")
kable(lh1)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
82	189.2993				
81	150.3299	1	38.9694680543685	20.9973358998904	1.6484475391443e-05

La tabla anterior nos brinda todo lo necesario para construir nuestro estadítico de prueba F_{06}.

$$g.l.(SSE(X_1, X_2, X_3, X_4, X_5, X_7, X_8)) = 82$$
$$g.l.(SSE(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)) = 81$$
$$SSE(X_1, X_2, X_3, X_4, X_5, X_7, X_8) = 189.30$$
$$SSE(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = 150.33$$
$$g.l(SSR(X_6|X_1, X_2, X_3, X_4, X_5, X_7, X_8)) = 82 - 81 = 1$$
$$SSR(X_6|X_1, X_2, X_3, X_4, X_5, X_7, X_8) = 189.30 - 150.33 = 38.969$$

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

- **Prueba F para test lineal general de la variable 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:
$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \beta_8 X_{i8} + E_i \ , \ E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
Estadístico de prueba y distribución:

$$F_{01} = \frac{SSR(X_1|X_2, X_3, X_4, X_5, X_6, X_7, X_8)}{MSE(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)} \ , \tag{4}$$
$$F_{01} = \frac{SSE(X_2, X_3, X_4, X_5, X_6, X_7, X_8) - SSE(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)}{MSE(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)} \ , \tag{5}$$
$$F_{01} \sim f_{1,81} \tag{6}$$

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

```r
lh2 <- linearHypothesis(modelo, "EDAD = 0")
lh2[1, 3:6] <- c(" ", " ", " ", " ")
kable(lh2)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
82	167.8077				
81	150.3299	1	17.4778182328601	9.41731151476297	0.00292351905114798

La tabla anterior nos brinda todo lo necesario para construir nuestro estadítico de prueba F_{01}.

$$g.l.(SSE(X_2, X_3, X_4, X_5, X_6, X_7, X_8)) = 82$$
$$g.l.(SSE(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)) = 81$$
$$SSE(X_2, X_3, X_4, X_5, X_6, X_7, X_8) = 167.81$$
$$SSE(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = 150.33$$
$$g.l(SSR(X_1|X_2, X_3, X_4, X_5, X_6, X_7, X_8)) = 82 - 81 = 1$$
$$SSR(X_1|X_2, X_3, X_4, X_5, X_6, X_7, X_8) = 167.81 - 150.33 = 17.478$$

La columna 'F' presenta el valor de F_{01} y la columna Pr(>F) su respectivo p-value.

Punto 4:

En la siguiente tabla se muestra la suma de cuadrados secuencial, y en la última fila la suma de cuadrados del error.

```
SS1 <- anova(modelo)["Sum Sq"]
kable(SS1, caption = "Sumas de cuadrados tipo I y SSE")
```

Table 7: Sumas de cuadrados tipo I y SSE

	Sum Sq
EDAD	13.2851764
RINF	116.2244592
RRC	1.9362640
RRX	8.4433232
NCAMAS	31.8519622
PDP	37.8732896
NENFERM	7.8699352
FSD	0.0840737
Residuals	150.3298765

Análogamente, en la siguiente tabla se muestra la suma de cuadrados parciales, y en la última fila la suma de cuadrados del error.

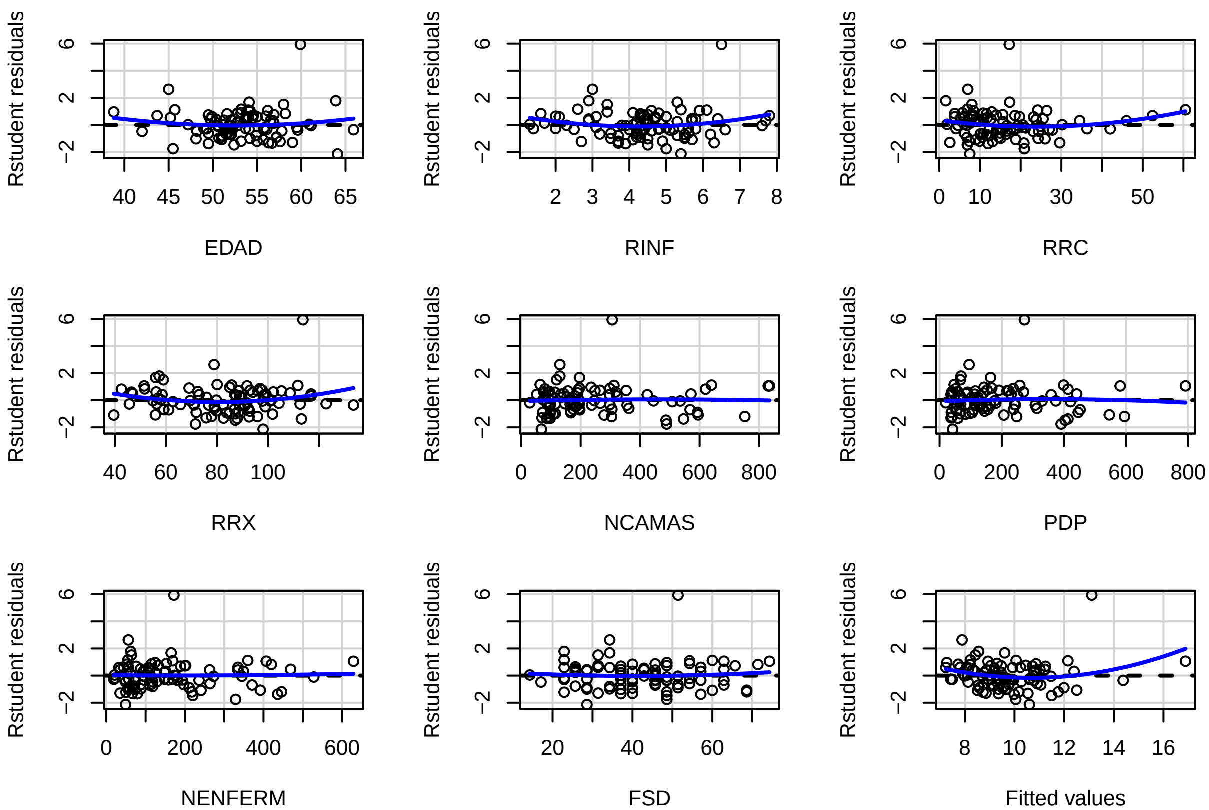
```
SS2 <- Anova(modelo)["Sum Sq"]
kable(SS2, caption = "Sumas de cuadrados tipo II y SSE")
```

Table 8: Sumas de cuadrados tipo II y SSE

	Sum Sq
EDAD	17.4778182
RINF	8.6143936
RRC	4.8407814
RRX	11.2888497
NCAMAS	10.7002791
PDP	38.9694681
NENFERM	7.8391631
FSD	0.0840737
Residuals	150.3298765

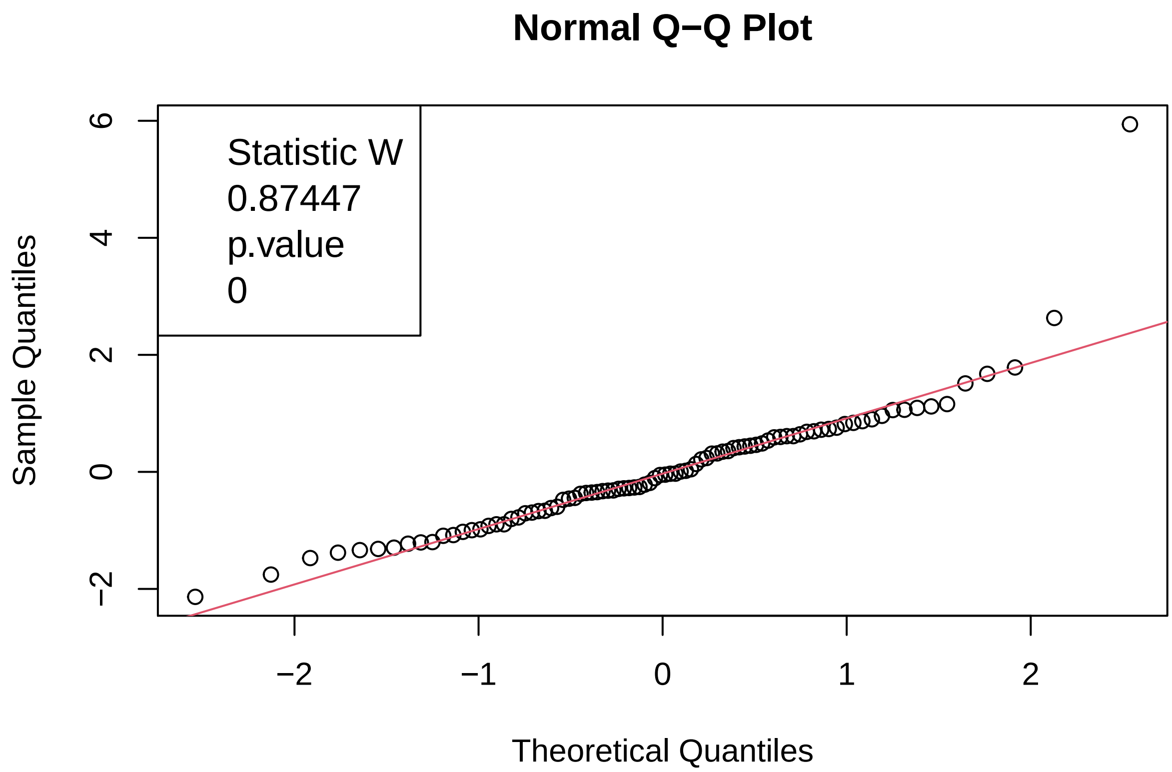
Punto 5:

```r
win.graph()
residualPlots(modelo, tests = FALSE ,type="rstudent")
```



Punto 6:

```r
test <- shapiro.test(rstudent(modelo))
qqnorm(rstudent(modelo),cex=1)
qqline(rstudent(modelo),col=2)
legend("topleft",legend=rbind(c("Statistic W","p.value"),round(c(test$statistic,test$p.value),digits=5))
```



Normal Q−Q Plot
Statistic W
0.87447
p.value
0
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
6
4
2
0
−2