Taller 2 Regresión lineal Multiple

Andrés Felipe Palomino - David Stiven Rojas

2023-04-21

1 Introducción

La base de datos "*yarn*" obtenida de la librería (PLS) contiene información sobre espectros NIR y mediciones de densidad de hilos de PET, consta de 28 individuos (hilos de PET), 268 variables predictoras (NIRS) y una variable de respuesta (densidad). Se ajustará un modelo lineal múltiple para estimar la densidad del hilo PET, mediante mediciones NIR

```
#Importación de librerías necesarias
library(car)
library(glmnet)
library(MASS)
library(xtable)
library(lmtest)
library(readxl)
library(lmridge)
library(pls)
library(olsrr)
```

1.1 Base de datos

En la siguiente tabla se encuentra un encabezado de la base de datos que se trabajara, esta consta de 30 covariables predictoras, las cuales estarán desde NIR1 hasta NIR30. De primera mano se observa que los valores de los NIR disminuyen a medida que la covariable aumenta

```
X <- data.frame(matrix(c(yarn$NIR[,1:30],yarn$density),nrow =28, ncol= 31))
colnames(X) <- c(paste("NIR",1:30,sep=""),"density")
```

1.2 Funciones creadas

Antes de empezar con el proceso de seleccionar las variables para ajustar el modelo se crean funciones para optimizar el proceso de validación de supuestos, debido a que constantemente se deben realizar, estas funciones estan diseñadas para objetos lm.

```
##Validacion grafica para homocedasticidad y normalidad y pruebas formales
validaciongrafica<- function(model,cor=F){

  par(mfrow=c(1,2))
  plot(fitted.values(model),studres(model),panel.first=grid(),
```

```
      pch=19,ylab='Residuos Estudentizados',xlab='Valores ajustados',main='A',col='aquamarine4')
  abline(h=c(-2,0,2),lty=2)
  qqPlot(model,pch=19,ylab='Residuos Estudentizados',
         xlab='Cuantiles Teóricos',col=carPalette()[1],
         col.lines=carPalette()[3],main='B')
  print('Shapiro Test; H0: Normalidad vs H1: No Normalidad')
  print(shapiro.test(studres(model)))
  print('Breusch Pagan Test;H0: Homocedasticidad vs H1: No Homocedasticidad')
  print(bptest(model))
  if(cor==T){
    par(mfrow=c(1,2))
    plot(studres(model),type="b",xlab="Tiempo",ylab="Residuos Estudentizados",main="A",
         pch=19,panel.first=grid())
    plot(studres(model)[-length(fitted.values(model))],
         studres(model)[-1],pch=19,panel.first = grid(),col="turquoise3",
         xlab=TeX("$Residuos_{t-1}$"),ylab=TeX("$Residuos_{t}$"),main="B")
    abline(lm(studres(model)[-1]~studres(model)[-length(fitted.values(model))]))
    print('Durbin Watson Test')
    print(durbinWatsonTest(model,
                           method='resample',reps=10000))
  }
  par(mfrow=c(1,1))
}

## Calculo de lambda optimo para boxcox
lambda<- function(model,a,b){
  par(mfrow=c(1,1))
  box.cox<-boxcox(model,lambda=seq(a,b,length.out = 1000),
                  ylab='log-verosimilitud')
  bc<-round(box.cox$x[box.cox$y ==max(box.cox$y)],2)
  print(bc)
}
```

2 Selección de variables

En el proceso de selección de variables se procede a realizar la Regresion de LASSO para identificar las posibles variables que tengan un aporte poco relevante, Por ultimo se ajustara el modelo cuyas variables tengan buenos indicadores y se pueda realizar corrección de supuestos

2.1 Regresión de LASSO

Este es un método de regularización que se implementa cuando se tiene muchas covariables disponibles y se cree que pocas tienen un aporte relevante.

Se asume el modelo de regresión usual, donde :

$$\mathrm{E(y|x)}=X^T\beta, \text{ y } \mathrm{V(y|x)}=\sigma^2$$

Donde se asume que algunos β son cero. El objetivo del estimador es seleccionar los coeficientes que tienen valores diferentes de cero. El cual se obtiene minimizando la siguiente expresión:

$$S_{lasso}(\beta) = \sum_{i=1}^{n}(y_i - x^T\beta)^2 + \lambda \sum_{j=1}^{p-1} |\beta_j|$$

Esta es la suma de cuadrados del estimador por MCO más una penalización (λ), a la suma del valor absoluto de los coeficientes. A medida que λ aumenta la penalización tendrá mas peso sobre la estimación de los coeficientes, es decir que si la penalización es muy grande, todas las estimaciones seran cero. No hay solución analitica para $\hat{\beta}_{lasso}$ por lo que se usan algoritmos para la estimación, como lo es la funcion de glmnet de la libreria glmnet.

2.1.1 Modelo a realizar regresión LASSO

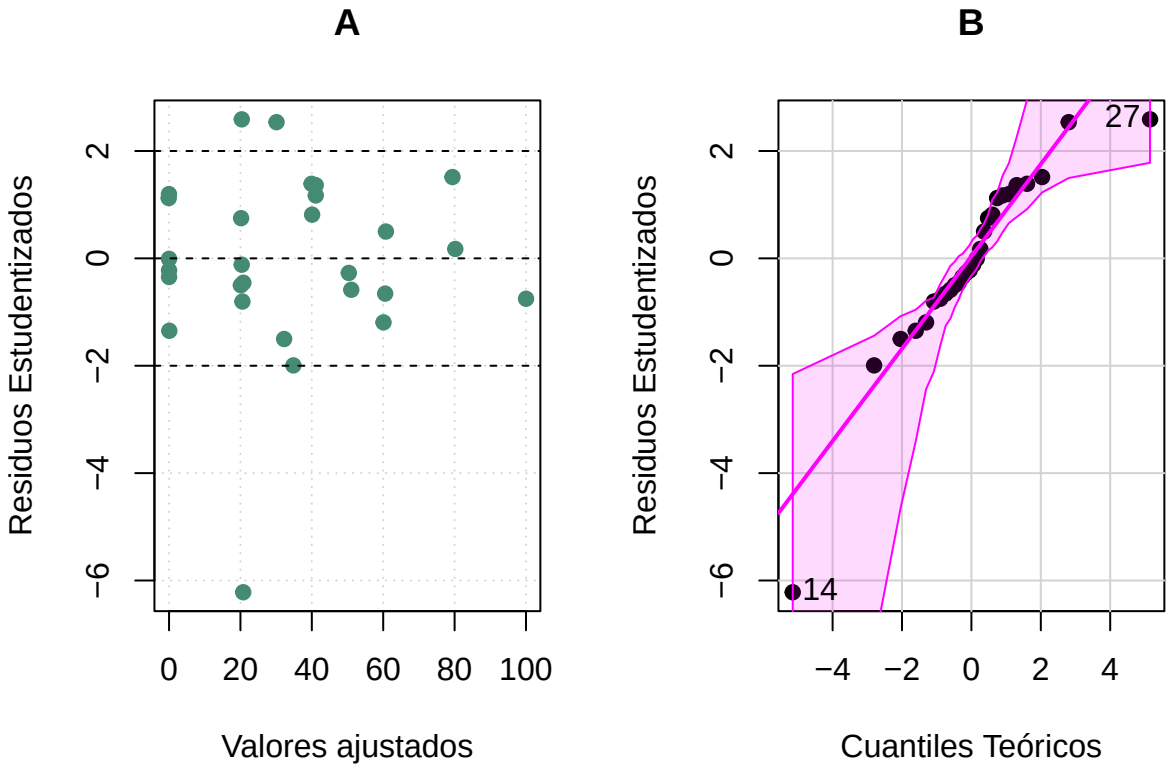
Como se establecio anteriormente, se asume un modelo de regresión usual, el cual debe cumplir los siguientes supuestos: E(y|x)=$x^T\beta$, y V(y|x)=σ^2, es decir, varianza constante y E(ε)=0 . Por ende es necesario proponer un modelo con p<n, en el cual se eliminaran las variables con menor correlación con la variable y. Dicho modelo se expresa acontinuación y se evaluan los supuestos:

```
model <- lm(density ~ .-NIR1-NIR8-NIR9-NIR10-NIR11-NIR7, data=X)
car::vif(model)
```

```
    NIR2         NIR3         NIR4         NIR5         NIR6        NIR12
```

1.664742e+03 3.984131e+04 3.611805e+05 6.232527e+05 2.540141e+05 8.859704e+06 NIR13 NIR14 NIR15 NIR16 NIR17 NIR18 7.628064e+07 7.977960e+07 5.366407e+07 8.067869e+07 9.939894e+07 1.635397e+08 NIR19 NIR20 NIR21 NIR22 NIR23 NIR24 3.087585e+08 3.600363e+08 2.771769e+08 3.693373e+08 4.754762e+08 4.611149e+08 NIR25 NIR26 NIR27 NIR28 NIR29 NIR30 3.850396e+08 2.050074e+08 7.042840e+07 3.712235e+07 2.000184e+07 1.522304e+06

```
validaciongrafica(model)
```



[1] “Shapiro Test; H0: Normalidad vs H1: No Normalidad”

```
Shapiro-Wilk normality test
```

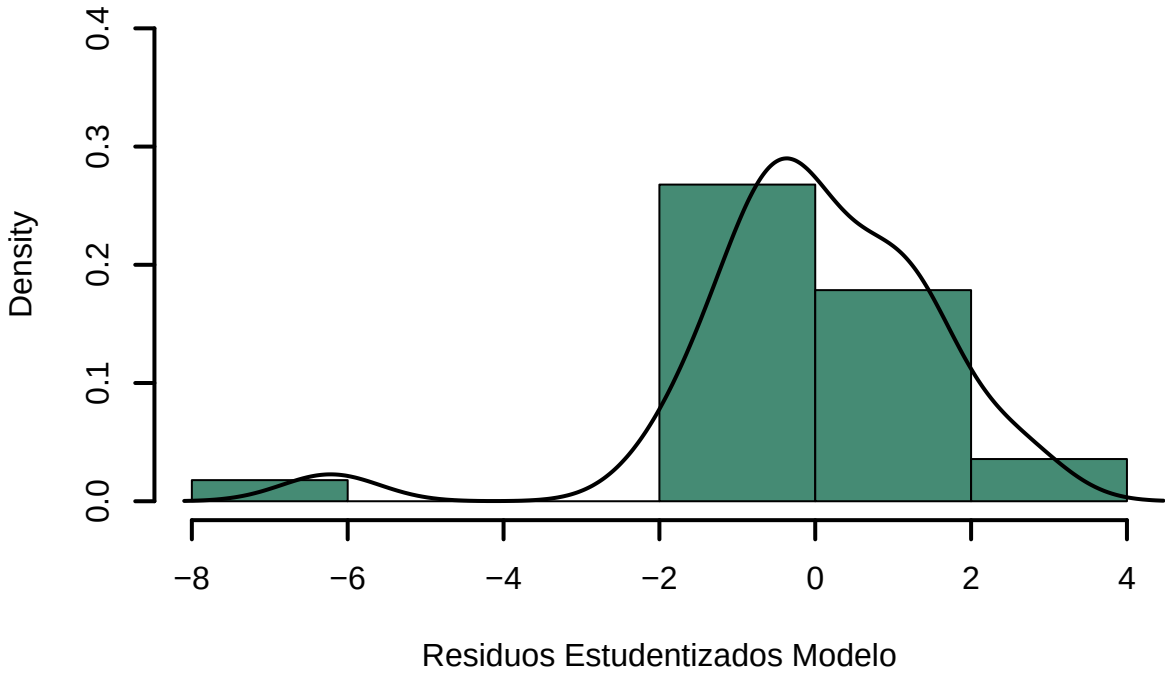
data: studres(model) W = 0.86458, p-value = 0.001868

[1] “Breusch Pagan Test;H0: Homocedasticidad vs H1: No Homocedasticidad”

```
studentized Breusch-Pagan test
```

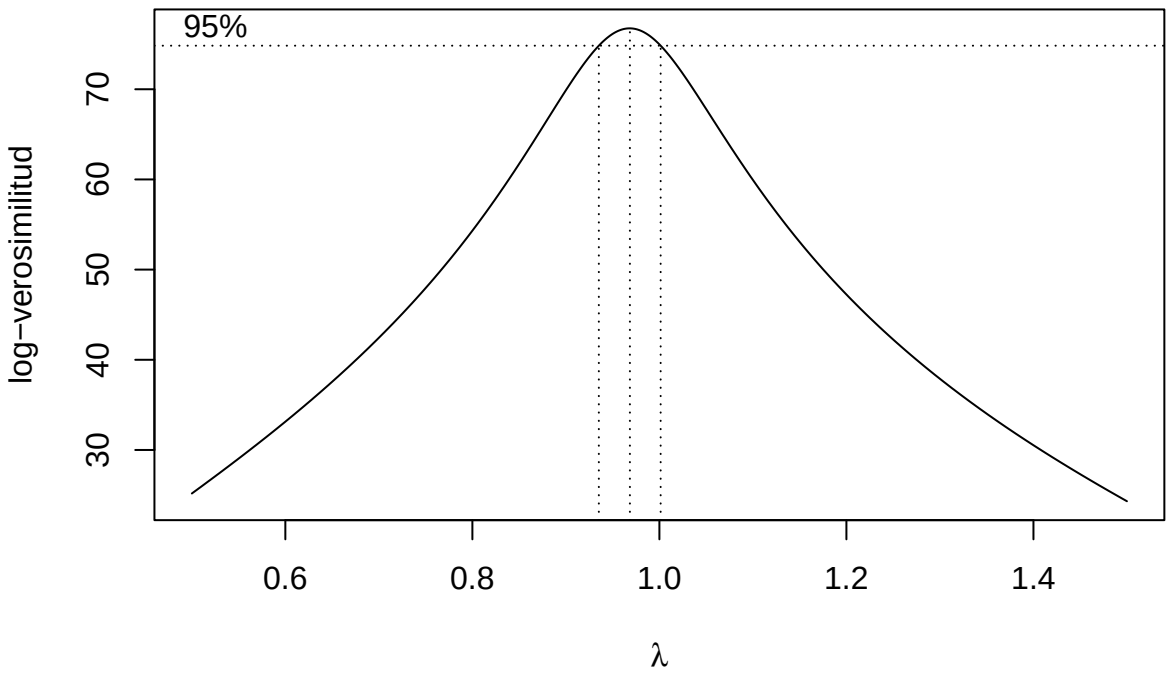
data: model BP = 27.288, df = 24, p-value = 0.2912

```
hist(studres(model),lwd=2,col='aquamarine4',freq=F,ylim=c(0,0.4),
     xlab='Residuos Estudentizados Modelo',main='')
lines(density(studres(model)),lwd=2,col='black')
```



Como no se cumple el supuesto de normalidad se procede a corregir mediante el metodo de BoxCox y se verifica el cumplimiento de los mismos.

```
model <- lm(density+0.01 ~ .-NIR1-NIR8-NIR9-NIR10-NIR11-NIR7, data=X)
lambda(model,0.5,1.5)
```



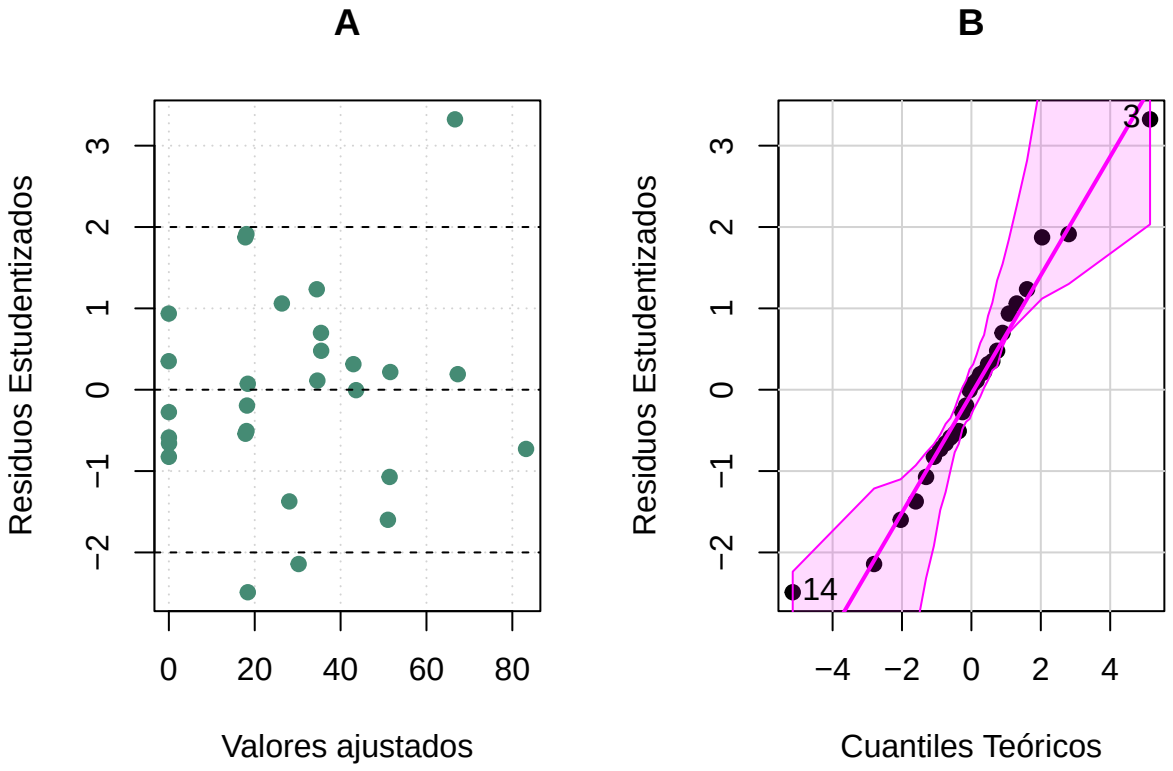
[1] 0.97

```
model.box <- lm(I(density^0.96) ~.-NIR1-NIR8-NIR9-NIR10-NIR11-NIR7,data=X)
car::vif(model.box)
```

```
    NIR2         NIR3         NIR4         NIR5         NIR6         NIR12
```

1.664742e+03 3.984131e+04 3.611804e+05 6.232527e+05 2.540141e+05 8.859703e+06 NIR13 NIR14 NIR15 NIR16 NIR17 NIR18 7.628063e+07 7.977959e+07 5.366406e+07 8.067868e+07 9.939892e+07 1.635397e+08 NIR19 NIR20 NIR21 NIR22 NIR23 NIR24 3.087585e+08 3.600362e+08 2.771768e+08 3.693372e+08 4.754761e+08 4.611148e+08 NIR25 NIR26 NIR27 NIR28 NIR29 NIR30 3.850395e+08 2.050074e+08 7.042839e+07 3.712234e+07 2.000184e+07 1.522303e+06

```
validaciongrafica(model.box)
```



[1] “Shapiro Test; H0: Normalidad vs H1: No Normalidad”

```
Shapiro-Wilk normality test
```

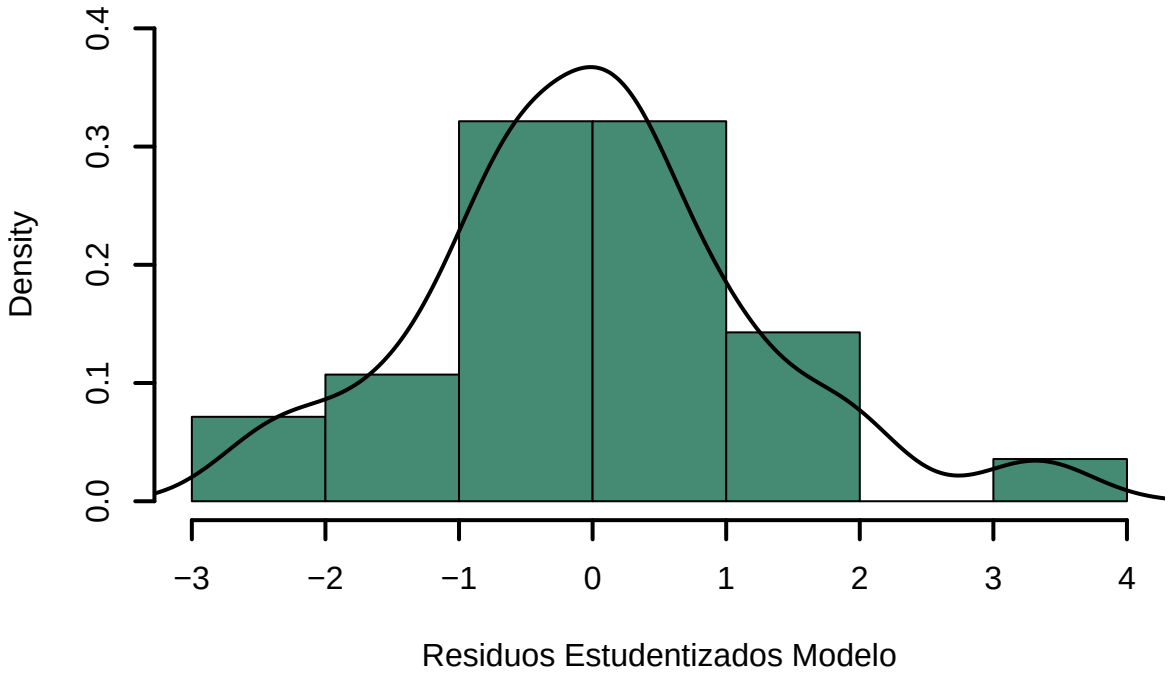
data: studres(model) W = 0.97774, p-value = 0.7934

[1] “Breusch Pagan Test;H0: Homocedasticidad vs H1: No Homocedasticidad”

```
studentized Breusch-Pagan test
```

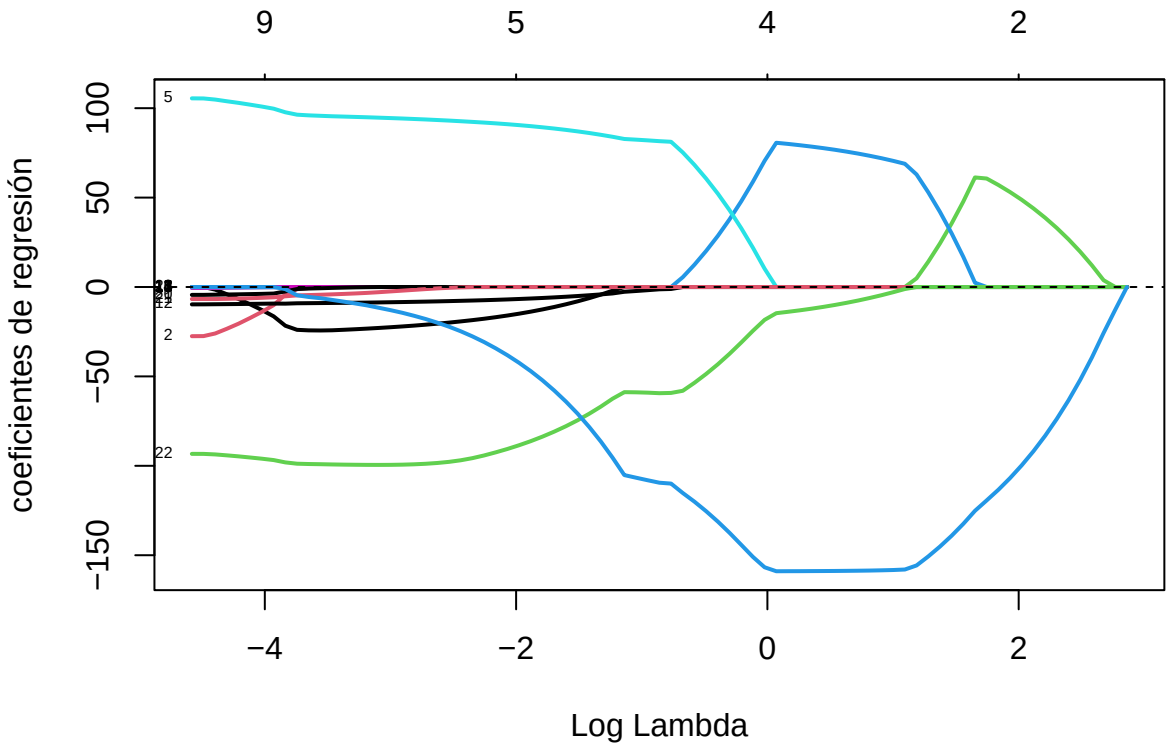
data: model BP = 23.94, df = 24, p-value = 0.4651

```
hist(studres(model.box),lwd=2,col='aquamarine4',
freq=F,ylim=c(0,0.4),xlab='Residuos Estudentizados Modelo',main='')
lines(density(studres(model.box)),lwd=2,col='black')
```



Ya con los requerimentos necesarios para realizar regresión de LASSO, se procede a calcular las estimaciones para distintos valores de λ que se muestran en la siguiente figura:

```
X.<-model.matrix(model.box)[,-1]
lasso.mod <- glmnet(X., X$density, alpha = 1,nlambda = 100)
plot(lasso.mod,xvar='lambda',label=T,lwd=2,ylab='coeficientes de regresión')
abline(h=0,lty=2)
```



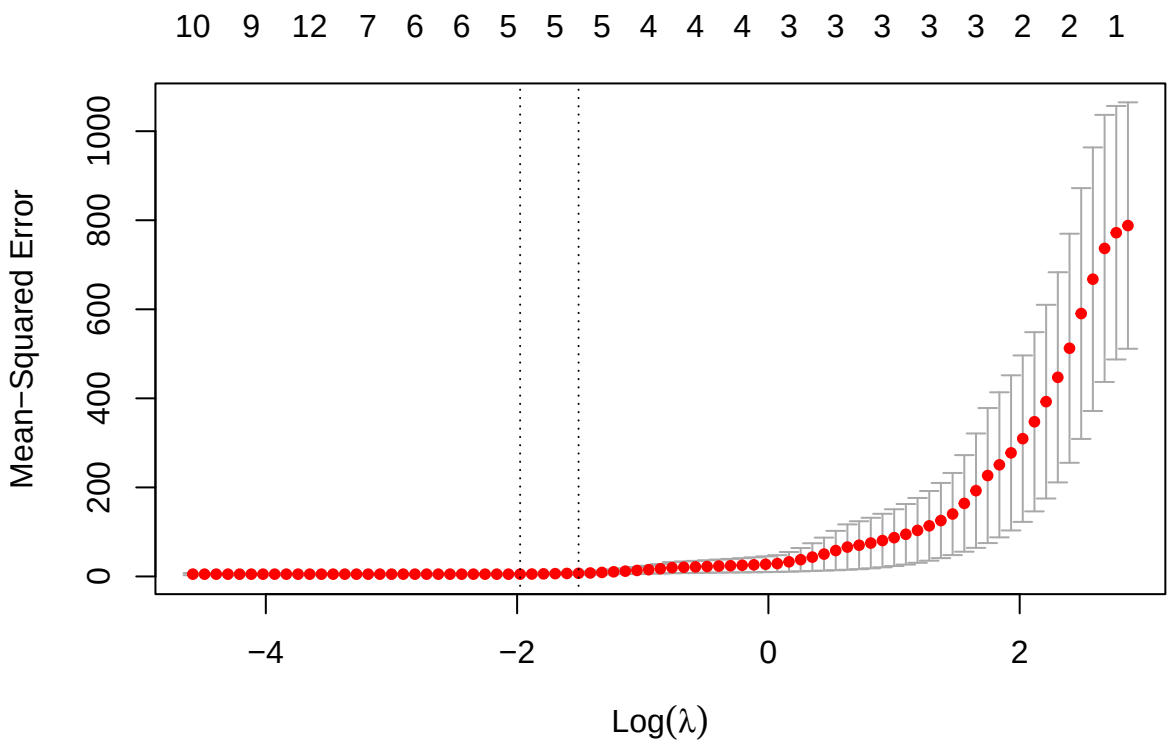
Para identificar el valor de λ optimo se procede a realizar validación cruzada.

2.1.2 Validación cruzada

Es un método para evaluar que tan bueno es un modelo para predecir observaciones futuras de la población objeto de estudio. La muestra se divide en dos grupos:

- Entrenamiento: Se usa para ajustar el modelo.
- Validación: Se utiliza para validar el modelo ajustado.

```
lasso.cv <-cv.glmnet(X., X$density, nfolds = 4, alpha = 1,
                     nlambda = 100)
plot(lasso.cv)
```




```
est = glmnet(X., X$density, alpha = 1,lambda = lasso.cv$lambda.1se)
est$beta
```

```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##               s0
## NIR2   -9.925178
## NIR3    .
## NIR4    .
## NIR5    .
## NIR6   87.822803
## NIR12   .
## NIR13   .
## NIR14   .
## NIR15   .
## NIR16   .
## NIR17   .
## NIR18  -5.289614
## NIR19   .
## NIR20   .
## NIR21   .
## NIR22   .
## NIR23   .
## NIR24   .
## NIR25   .
## NIR26   .
## NIR27   .
```

```
## NIR28 -91.135757
## NIR29 -43.247958
## NIR30   .
```

La selección de variables por medio del estimador LASSO son: NIR2, NIR6, NIR18, NIR28, NIR29. Consiguiente a eso se procede a realizar una suma extra de cuadrados para evaluar si podemos eliminar NIR29 para evitar problemas de multicolinealidad.

2.2 Suma extra de cuadrados

Sirve para probar la significancia de un subconjunto de coeficientes.

Se tiene el siguiente modelo:

$$y = X\beta + \varepsilon$$

donde β=$\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}$

donde β_1 es un vector (p-r)x1 y β_2 es un vector rx1 , se quiere evaluar la siguiente hipotesis:

$$H_0 : \beta_2 = 0 \text{ vs } H_1 : \beta_2 \neq 0$$

Se tienen los siguientes modelos: Modelo completo : $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ Modelo reducido : $y = X_1\beta_1\varepsilon$

```
model.lasso1 <- lm(density~NIR2+NIR6+NIR18+NIR28+NIR29,data=X)
model.lasso2 <- lm(density~NIR2+NIR6+NIR18+NIR28,data=X)
anova(model.lasso2,model.lasso1)
```

```
## Analysis of Variance Table
##
## Model 1: density ~ NIR2 + NIR6 + NIR18 + NIR28
## Model 2: density ~ NIR2 + NIR6 + NIR18 + NIR28 + NIR29
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     23 31.435
## 2     22 30.610  1   0.82493 0.5929 0.4495
```

```
car::vif(model.lasso1)
```

```
##       NIR2       NIR6      NIR18      NIR28      NIR29
##   3.766765   5.643206  36.089199 269.277707 304.968458
```

```
car::vif(model.lasso2)
```

```
##      NIR2      NIR6     NIR18     NIR28
##  2.967327  4.203285 31.085734 26.983026
```

3 Modelo de regresión multiple

Con base en el proceso de selección de variables se ajusta el siguiente modelo y se realiza la respectiva validación de supuestos:

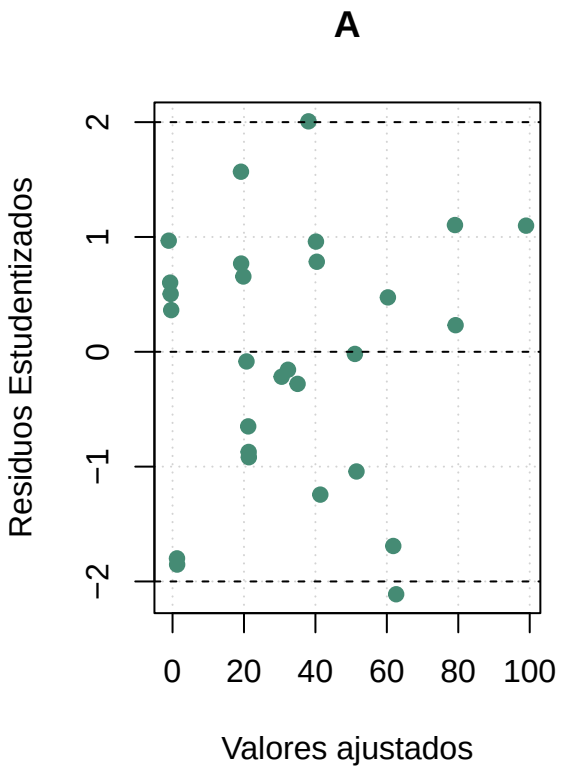
```r
model.lasso1<- lm(density~NIR2+NIR6+NIR18+NIR28,data=X)

summary(model.lasso1)
```

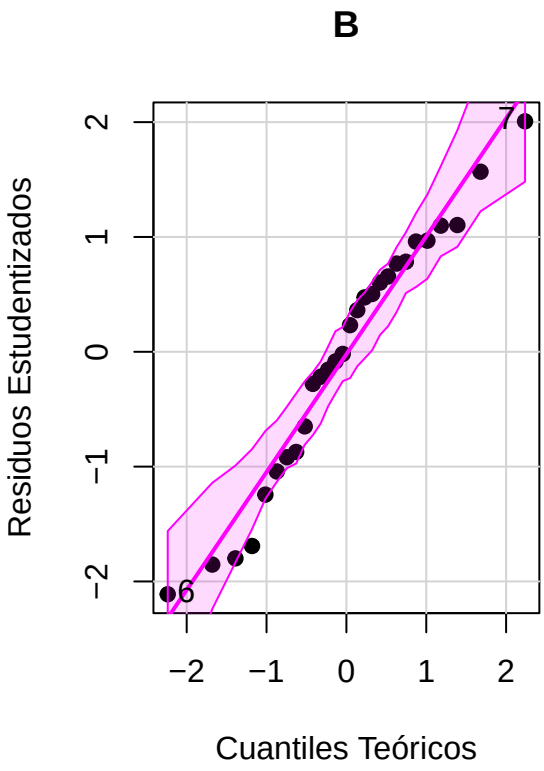
```
##
## Call:
## lm(formula = density ~ NIR2 + NIR6 + NIR18 + NIR28, data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1312 -0.9776  0.1102  0.8381  2.0416
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.389     10.712   2.744   0.0116 *
## NIR2         -26.257      3.892  -6.747 6.99e-07 ***
## NIR6          96.140      1.741  55.211  < 2e-16 ***
## NIR18         -9.055      1.905  -4.753 8.62e-05 ***
## NIR28       -109.939      5.818 -18.896 1.66e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.169 on 23 degrees of freedom
## Multiple R-squared:  0.9984, Adjusted R-squared:  0.9981
## F-statistic:  3584 on 4 and 23 DF,  p-value: < 2.2e-16
```

```r
car::vif(model.lasso1)
```

```
##      NIR2      NIR6     NIR18     NIR28
##  2.967327  4.203285 31.085734 26.983026
```

```r
validaciongrafica(model.lasso1)
```



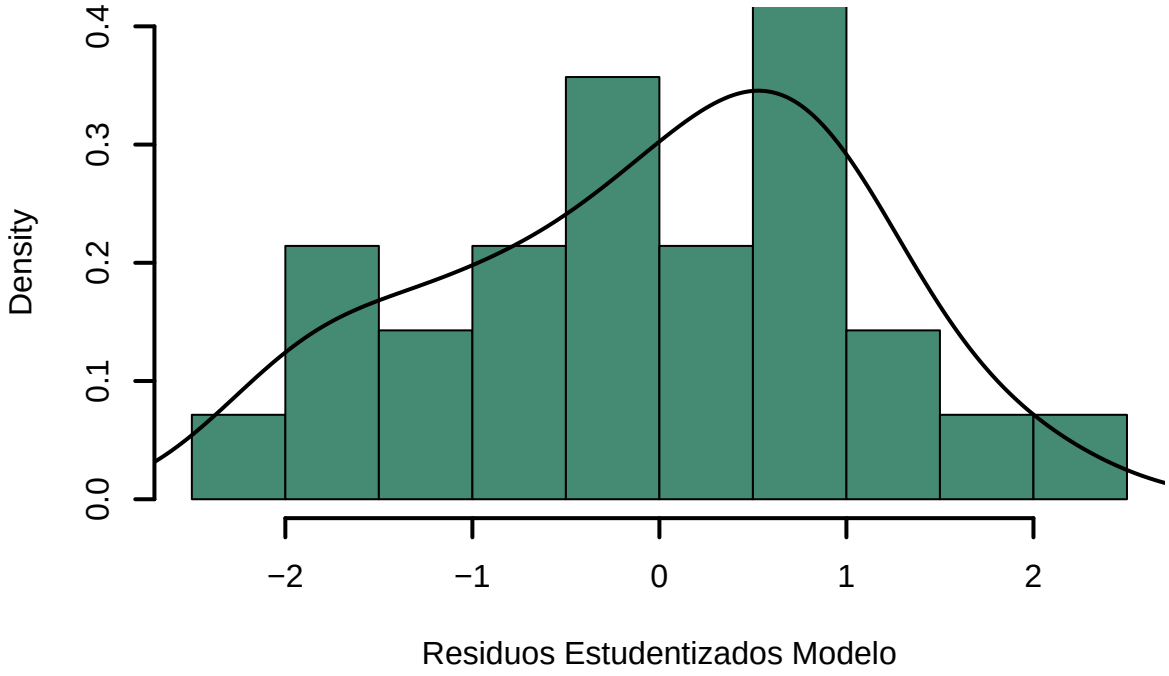



```
## [1] "Shapiro Test; H0: Normalidad vs H1: No Normalidad"
##
##  Shapiro-Wilk normality test
##
## data:  studres(model)
## W = 0.96468, p-value = 0.4471
##
## [1] "Breusch Pagan Test;H0: Homocedasticidad vs H1: No Homocedasticidad"
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 1.6317, df = 4, p-value = 0.8031
```

```
car::vif(model.lasso1)
```

```
##      NIR2      NIR6     NIR18     NIR28
##  2.967327  4.203285 31.085734 26.983026
```

```
hist(studres(model.lasso1),lwd=2,col='aquamarine4',
freq=F,ylim=c(0,0.4),xlab='Residuos Estudentizados Modelo',main='')
lines(density(studres(model.lasso1)),lwd=2,col='black')
```



3.1 Identificación de puntosa atípicos e influyentes

Para esto utilizaremos la función influence.measures()

```
influence.measures(model.lasso1)$infmat[,-1]
```

```
##        dfb.NIR2      dfb.NIR6     dfb.NIR1    dfb.NIR28        dffit     cov.r
## 1  -0.373394566  0.6496848240 -0.331772536  0.184113316  0.828264429 1.4997618
## 2  -0.305492910  0.3720578646  0.073404994 -0.157310694  0.554365040 1.1944235
## 3  -0.019231535  0.0750594731 -0.081707189  0.061293691  0.128782237 1.6143881
## 4  -0.106547754  0.0838881352  0.142356089 -0.158122212  0.238509612 1.4878645
## 5   0.074553155 -0.1793449266 -0.047690711  0.118053390 -0.492162405 0.7352324
## 6  -0.191413380 -0.0858801572  0.403910765 -0.282271360 -0.860714944 0.5788966
## 7  -0.167776877  0.1601914874  0.433679945 -0.401078719  0.820454554 0.6288836
## 8  -0.069030633  0.0884179635 -0.268936383  0.265910872 -0.396733876 0.9794717
## 9   0.306437652 -0.1573248360 -0.132526148  0.132848593  0.377972302 1.3407617
## 10  0.271697306 -0.1981032511 -0.170864580  0.147302866  0.467326908 1.2584910
## 11 -0.157527270  0.1499879881  0.006592832  0.017816423  0.298765048 1.3687873
## 12  0.038781596  0.0435008484 -0.218669684  0.188489125 -0.326187278 1.2009151
## 13 -0.131949786  0.1833351054 -0.155967107  0.129606661 -0.282701090 1.1329332
## 14 -0.016192687  0.0213626824 -0.003712707  0.002586270 -0.028522311 1.3905338
## 15  0.157449597 -0.3162571007 -0.207031055  0.193692160  0.758360671 0.9070889
## 16  0.170444545 -0.9628595573  1.795953591 -1.991921206 -2.336700639 1.5662380
## 17 -0.012587784  0.0453219738 -0.099683412  0.126128085  0.195758464 1.5644498
```

```
## 18 -0.003506908 -0.0381099029  0.003559217  0.032787471  0.209923969 1.2904872
## 19 -0.061830397 -0.1112793632  0.061394764 -0.019920811  0.342151199 1.1407321
## 20 -0.100782895 -0.0516699098  0.034376668 -0.024073454  0.256303217 1.4848636
## 21  1.547939289 -0.3340483842 -0.015071485  0.067136669 -2.264225545 1.6244314
## 22 -0.000733778 -0.0003063477 -0.001200355  0.001528885 -0.004712816 1.3286357
## 23 -0.062384588  0.0207989090 -0.041460498  0.068343148 -0.262229787 1.0436235
## 24 -0.019265307  0.0171498437 -0.026120275  0.022702022 -0.049700844 1.3675029
## 25 -0.065310959  0.0562803062 -0.034684395  0.031194744 -0.092996640 1.3630398
## 26 -0.046421805  0.0530560130 -0.020592018  0.019655727 -0.072195305 1.3720953
## 27  0.009424286  0.0351500667 -0.115781227  0.095327904 -0.200308770 1.2437138
## 28  0.213025445 -0.2025624490  0.018407942  0.004972384  0.274777337 1.2338010
##          cook.d        hat
## 1  1.359813e-01 0.36242057
## 2  6.088590e-02 0.20142292
## 3  3.459292e-03 0.23577992
## 4  1.177436e-02 0.20216121
## 5  4.481989e-02 0.07807664
## 6  1.288001e-01 0.14249406
## 7  1.189678e-01 0.14319400
## 8  3.074773e-02 0.09232509
## 9  2.905892e-02 0.18848578
## 10 4.382906e-02 0.19164343
## 11 1.830555e-02 0.17180913
## 12 2.150305e-02 0.12266013
## 13 1.609381e-02 0.08659086
## 14 1.700455e-04 0.10329971
## 15 1.081585e-01 0.18951329
## 16 9.875246e-01 0.61389878
## 17 7.964918e-03 0.22520021
## 18 9.064674e-03 0.10826245
## 19 2.347849e-02 0.11113378
## 20 1.357895e-02 0.20575666
## 21 9.344561e-01 0.61296861
## 22 4.643968e-06 0.06009151
## 23 1.370158e-02 0.05954171
## 24 5.159177e-04 0.09178965
## 25 1.801908e-03 0.09981384
## 26 1.087484e-03 0.09948725
## 27 8.231598e-03 0.08683178
## 28 1.537417e-02 0.11334702
```

Dónde observamos que las observaciones 16,21 son influyentes a nuestro modelo. Los puntos dentro de la base de datos lucen así y procedemos a ilustrarlos para que cuando un experto en el tema pueda considerarlos y evaluar si fueron errores de mediciones o que ocurre realmente con ellos.

```
X[c(16,21),c(2,6,18,28,31)]
```

```
##      NIR2   NIR6  NIR18  NIR28 density
## 16 3.1229 2.9345 3.3254 1.8021       0
## 21 2.6803 1.8602 1.3031 1.1352       0
```