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")

xtable(head(X[,1:11]))
```

% latex table generated in R 4.3.0 by xtable 1.8-4 package % Sat Apr 29 18:42:06 2023

	NIR1	NIR2	NIR3	NIR4	NIR5	NIR6	NIR7	NIR8	NIR9	NIR10	NIR11
1	3.07	3.09	3.11	3.10	3.00	2.83	2.62	2.40	2.19	2.01	1.84
2	3.07	3.09	3.10	3.07	2.98	2.84	2.68	2.51	2.35	2.22	2.12
3	3.08	3.10	3.09	3.03	2.88	2.69	2.48	2.27	2.08	1.92	1.77
4	3.08	3.10	3.10	3.07	2.99	2.87	2.74	2.61	2.50	2.42	2.38
5	3.10	3.10	3.08	3.02	2.89	2.72	2.54	2.38	2.24	2.13	2.05
6	3.08	3.08	3.05	2.93	2.73	2.51	2.29	2.10	1.93	1.79	1.67

```
xtable(head(X[,12:21]))
```

% latex table generated in R 4.3.0 by xtable 1.8-4 package % Sat Apr 29 18:42:06 2023

	NIR12	NIR13	NIR14	NIR15	NIR16	NIR17	NIR18	NIR19	NIR20	NIR21
1	1.69	1.58	1.50	1.44	1.34	1.22	1.14	1.12	1.13	1.16
2	2.04	1.98	1.96	1.94	1.89	1.82	1.75	1.71	1.68	1.65
3	1.65	1.55	1.49	1.44	1.35	1.26	1.20	1.18	1.19	1.21
4	2.35	2.35	2.37	2.40	2.40	2.38	2.33	2.28	2.21	2.11
5	1.99	1.95	1.94	1.93	1.90	1.85	1.80	1.76	1.73	1.68
6	1.56	1.48	1.43	1.39	1.32	1.25	1.20	1.19	1.19	1.19

```
xtable(head(X[,22:31]))
```

% latex table generated in R 4.3.0 by xtable 1.8-4 package % Sat Apr 29 18:42:06 2023

	NIR22	NIR23	NIR24	NIR25	NIR26	NIR27	NIR28	NIR29	NIR30	density
1	1.16	1.15	1.15	1.13	1.07	1.02	1.01	1.03	1.08	100.00
2	1.58	1.51	1.45	1.38	1.29	1.20	1.15	1.13	1.14	80.22
3	1.20	1.18	1.17	1.15	1.10	1.07	1.06	1.08	1.12	79.49
4	1.98	1.85	1.75	1.63	1.51	1.40	1.30	1.23	1.20	60.80
5	1.60	1.52	1.46	1.39	1.31	1.24	1.19	1.16	1.17	59.97
6	1.18	1.15	1.14	1.12	1.09	1.06	1.06	1.07	1.11	60.48

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.

```
##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
  lines(lowess(studres(model)~fitted.values(model)), col = "red1")
  abline(h=c(-2,0,2),lty=2)
  qqPlot(model,pch=19,ylab='Residuos Estudentizados',xlab='Cuantiles Teóricos',col=carPalette()[1],col.l
  print('Shapiro Test')
  print(shapiro.test(studres(model)))
  print('Breusch Pagan Test')
  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.firs
    plot(studres(model)[-length(fitted.values(model))],studres(model)[-1],pch=19,panel.first = grid(),co
    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))
}
```

```r
## 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 :

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

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:

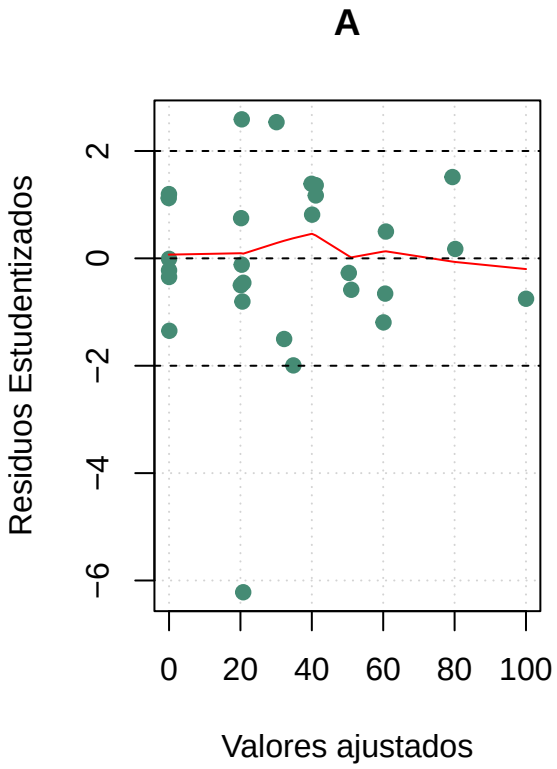
$$S_{lasso}(\beta) = \sum_{i=1}^{n}(y_i - x^T\beta)^2 + \lambda \sum_{j=1}^{p-1} |\beta_j|$$

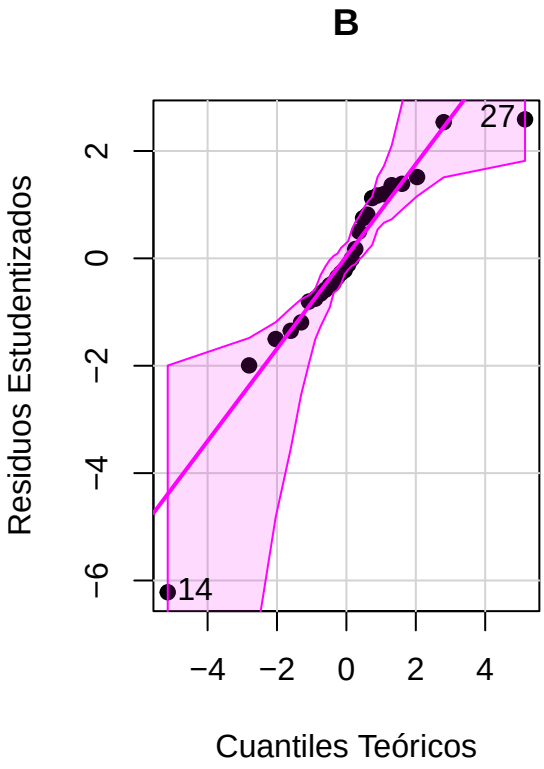
Esta es la suma de cuadrados del estimador por MCO más una penalización (λ), a la suma del valor absoluto de los coeficientes. A medida que λ aumenta la penalización tendrá mas peso sobre la estimación de los coeficientes, es decir que si la penalización es muy grande, todas las estimaciones seran cero. 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: $\text{E(y|x)}=x^T\beta$, y $\text{V(y|x)}=\sigma^2$, 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:

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





[1] “Shapiro Test”

```
Shapiro-Wilk normality test
```

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

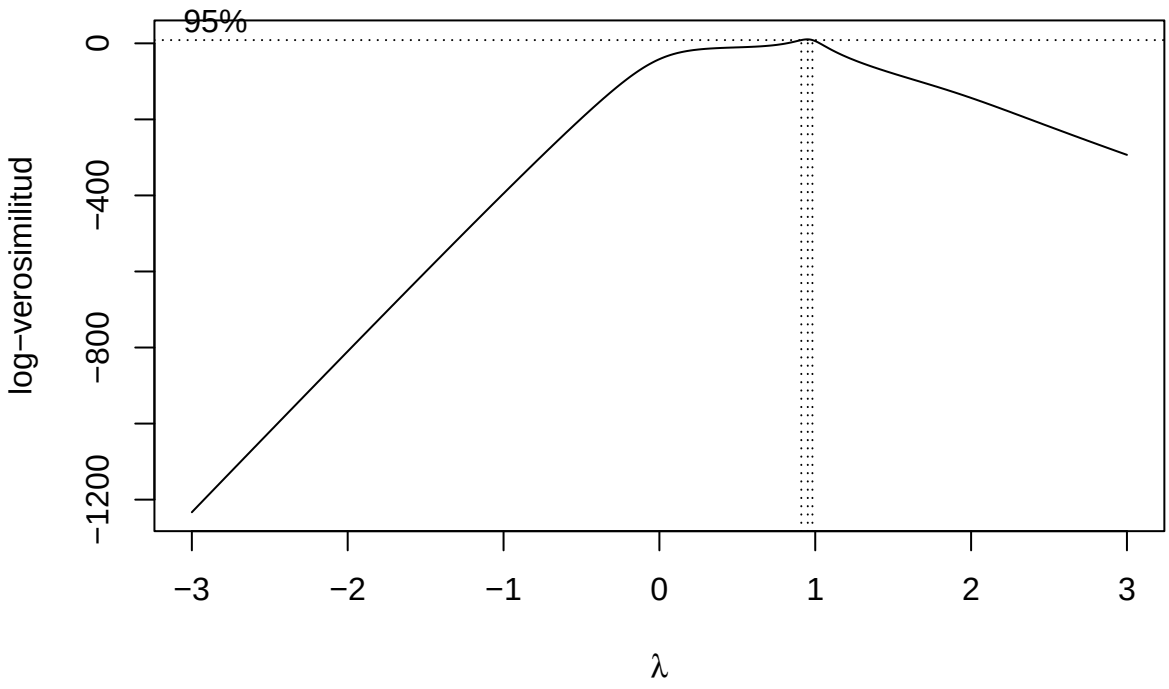
[1] “Breusch Pagan Test”

```
studentized Breusch-Pagan test
```

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

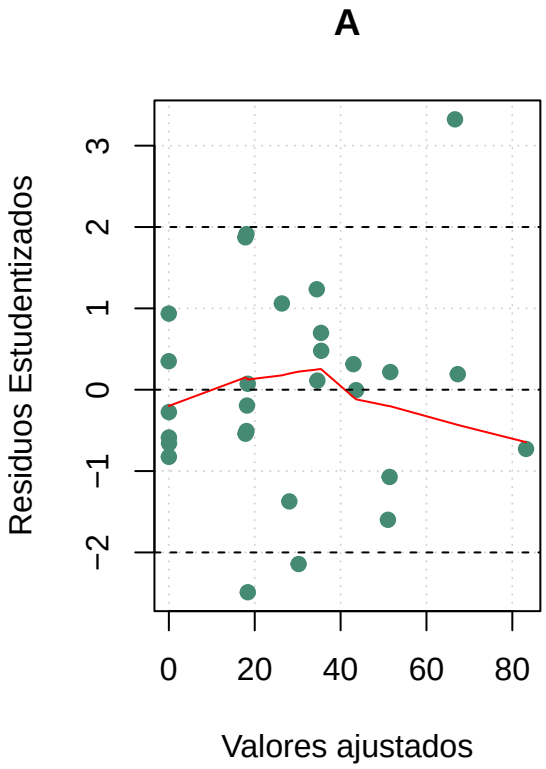
Como no se cumple el supuesto de normalidad se procede a corregir mediante el metodo de BoxCox y se verifica el cumplimiento de los mismos.

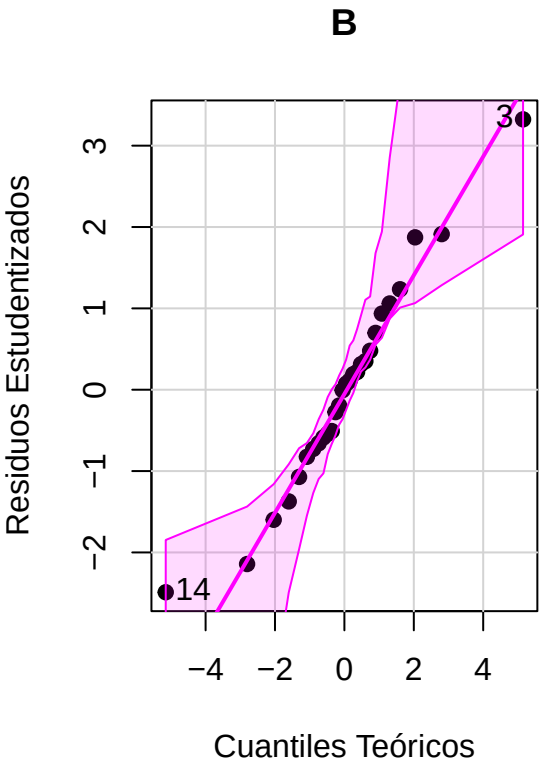
```
model <- lm(density+0.0000001 ~ .-NIR1-NIR8-NIR9-NIR10-NIR11-NIR7, data=X)
lambda(model,-3,3)
```



[1] 0.95

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





[1] “Shapiro Test”

```
Shapiro-Wilk normality test
```

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

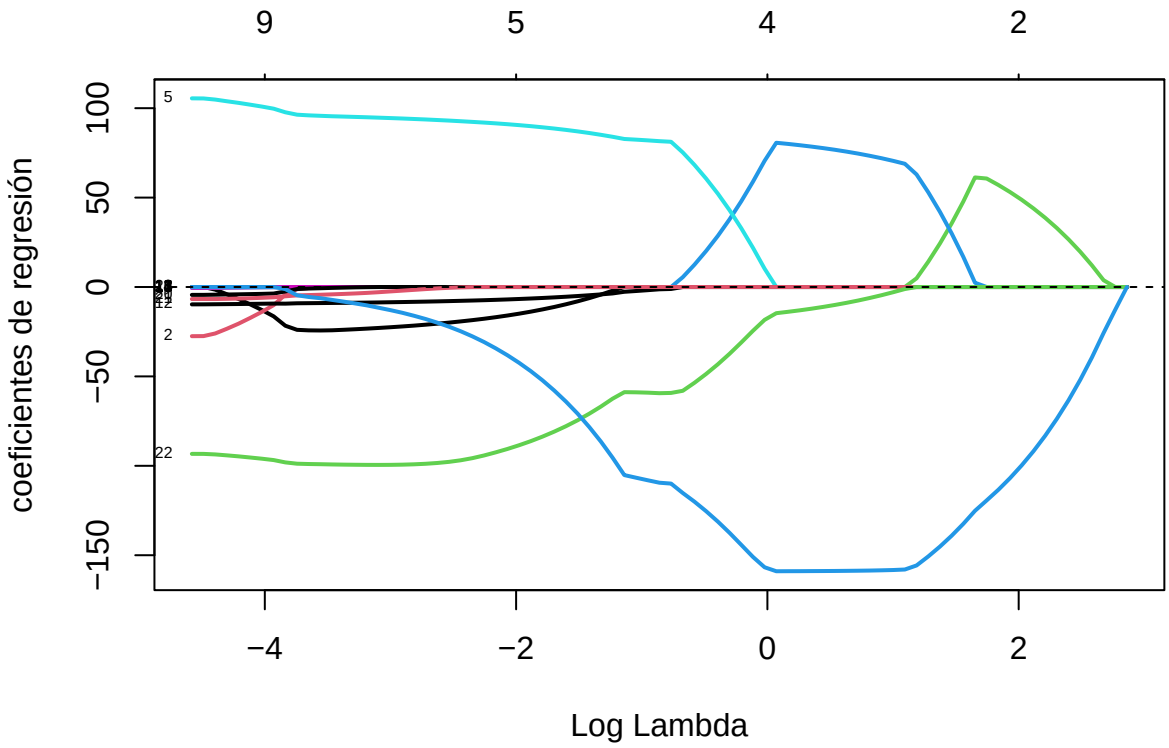
[1] “Breusch Pagan Test”

```
studentized Breusch-Pagan test
```

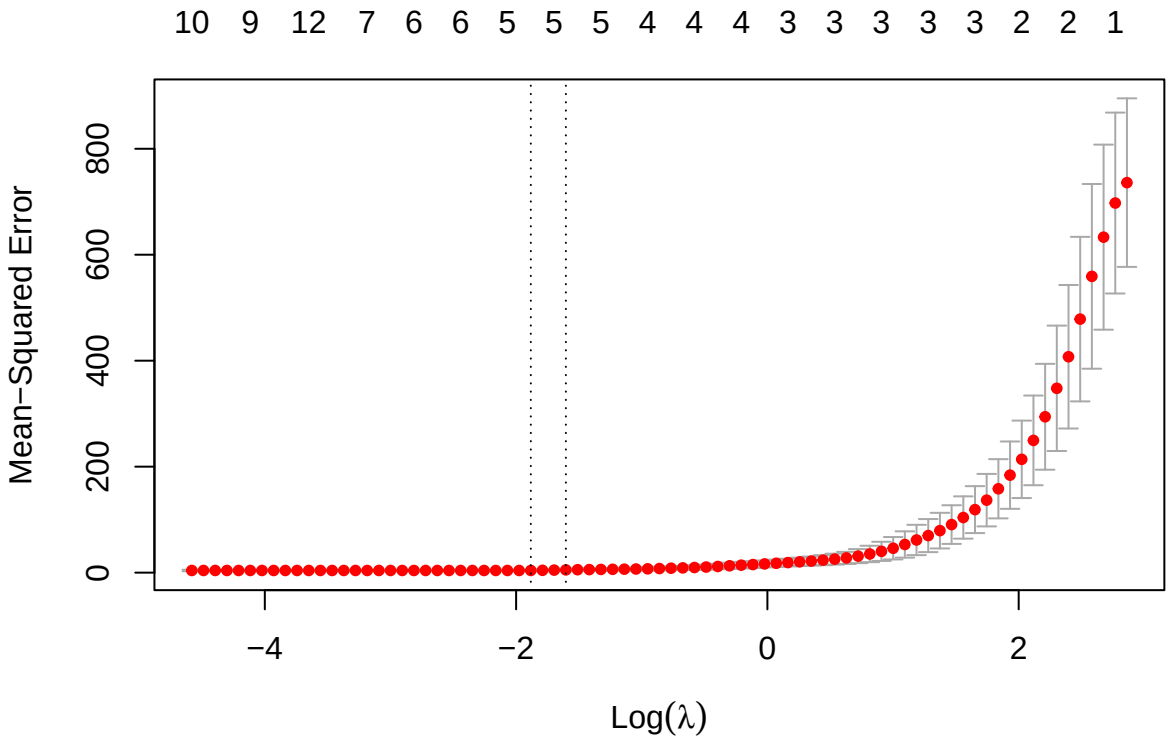
data: model BP = 23.94, df = 24, p-value = 0.4651

3

```
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)
```




```
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 -11.314302 NIR3 .
NIR4 .
NIR5 .
NIR6 88.544252 NIR12 .
NIR13 .
NIR14 .
NIR15 .
NIR16 .
NIR17 .
NIR18 -5.643411 NIR19 .
NIR20 .
NIR21 .
NIR22 .
NIR23 .
NIR24 .
NIR25 .
NIR26 .
NIR27 .
NIR28 -92.202115 NIR29 -40.266344 NIR30 .