Taller 1 Regresión lineal simple

Andrés Felipe Palomino - David Stiven Rojas
Códigos:1922297-1924615
Universidad del Valle

14 de abril de 2023

Universidad del Valle

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). Primeramente, se realiza una descripción de las covariables que se encuentran en la base de datos. Luego, se ajustará un modelo lineal simple para estimar la densidad, mediante una medición NIR seleccionada que garantice el mejor modelo.

1.1. NIR

La espectroscopia de infrarrojo cercano (NIRS) es un método óptico los cuales ofrecen información a tiempo real sobre los diversos procesos fisiológicos y patológicos que ocurren en tejidos y órganos, NIRS involucra un haz de luz que al interactuar con un material produce una radiación electromagnética en forma de ondas en el rango de los 750 a los 2 600 nm dentro del espectro cercano al infrarrojo. Para mayor claridad del espectro electromagnético se presenta una figura en las que se evidencia la luz visible (colores), y los espectros infrarrojos cercanos (NIR).

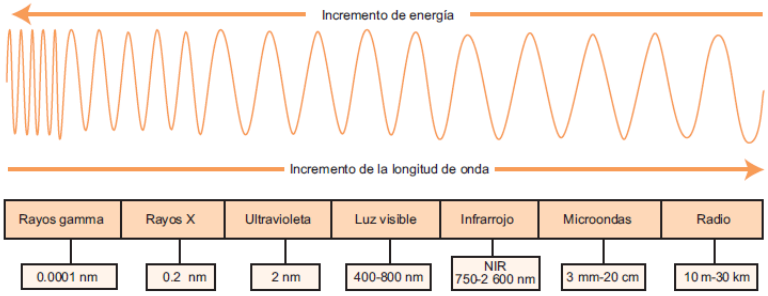


Figura 1. Representaciones esquemáticas de los componentes generales de una onda de un haz de luz señalando su amplitud y su longitud (A) y de la localización del espectro cercano al infrarrojo (NIRS) dentro del espectro general de la luz (B).

De acuerdo a la composición de la muestra, este haz de luz al interactuar con la muestra se verá reflejado (Reflectancia), absorbido (Absorción) o atravesado (Transmitancia). Esta onda se analiza y puede proporcionar información acerca de la muestra como geometría del objeto, tamaño, distribución, composición y densidad. Cabe aclarar que esta metodología para el análisis de cuerpos, o en general cualquier tipo de elemento que presente una densidad, es muy útil, puesto que es un procedimiento no invasivo que permite tener información valiosa a partir de los espectros, dónde más allá de nuestro caso común en el PET también se utiliza para procesos de tomografía para evaluar presencia de anomalías dentro del organismo, tumores, y demás.

1.2. Hilos de PET

Los hilos de PET son hilos fabricados a partir de tereftalato polietileno (PET), el cual es un tipo de plástico fuerte, flexible y 100 % reciclable, este material generalmente es usado en la industria textil, fabricación de envases, bolsos, juguetes y tejidos sintéticos. La variable que se quiere predecir es la densidad, la cual describe la relación entre el peso y el volumen que ocupa el hilo PET.

1.3. 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.

	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

	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

	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

2. Análisis descriptivo

Antes de ajustar el modelo se realiza un análisis descriptivo, en este se observa que la base de datos comprende valores desde 3.19 hasta 1.08, además la media de los NIR va disminuyendo a medida que estos aumentan.

```
Sd <- apply(X,2,sd)
RESU <- rbind(apply(X,2,summary),Sd)
xtable(RESU[,1:11])
```

	NIR1	NIR2	NIR3	NIR4	NIR5	NIR6	NIR7	NIR8	NIR9	NIR10	NIR11
Min.	2.87	2.68	2.45	2.24	2.04	1.86	1.71	1.59	1.52	1.47	1.42
1st Qu.	3.07	3.07	2.98	2.85	2.67	2.49	2.29	2.16	2.07	1.97	1.88
Median	3.10	3.09	3.05	2.97	2.83	2.67	2.50	2.37	2.26	2.18	2.14
Mean	3.09	3.07	3.01	2.90	2.76	2.60	2.46	2.34	2.25	2.19	2.16
3rd Qu.	3.12	3.12	3.10	3.03	2.90	2.77	2.64	2.52	2.46	2.44	2.44
Max.	3.19	3.18	3.13	3.10	3.02	2.93	2.89	2.87	2.88	2.93	2.99
Sd	0.06	0.10	0.15	0.21	0.24	0.26	0.28	0.30	0.32	0.36	0.40

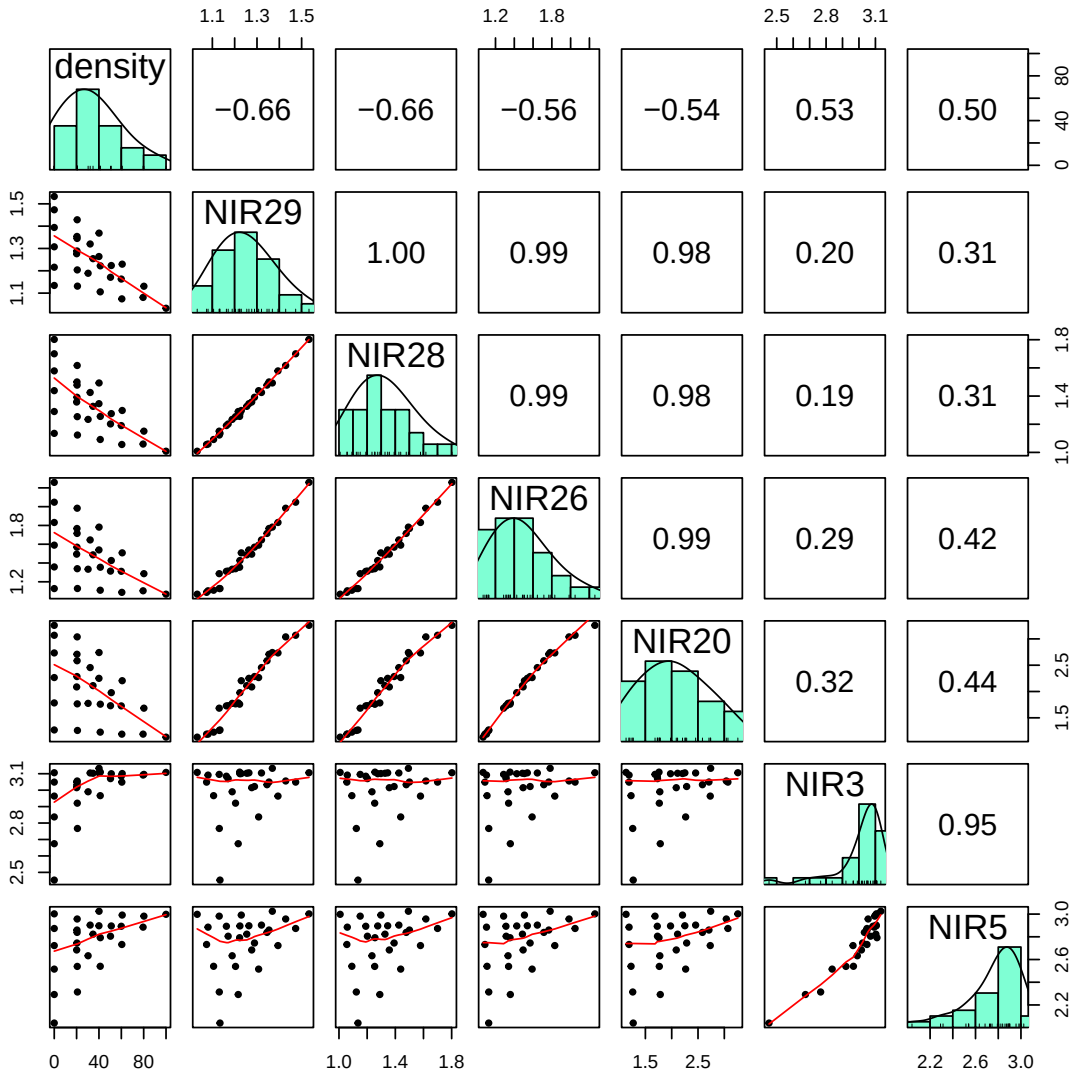
```
xtable(RESU[,12:21])
```

	NIR12	NIR13	NIR14	NIR15	NIR16	NIR17	NIR18	NIR19	NIR20	NIR21
Min.	1.38	1.35	1.33	1.31	1.30	1.22	1.14	1.12	1.13	1.16
1st Qu.	1.85	1.86	1.87	1.88	1.89	1.84	1.79	1.75	1.72	1.67
Median	2.14	2.14	2.16	2.19	2.19	2.18	2.15	2.10	2.03	1.94
Mean	2.14	2.14	2.16	2.18	2.18	2.17	2.14	2.10	2.04	1.96
3rd Qu.	2.46	2.50	2.55	2.61	2.65	2.66	2.64	2.59	2.49	2.35
Max.	3.07	3.15	3.21	3.26	3.29	3.32	3.33	3.31	3.26	3.17
Sd	0.45	0.50	0.54	0.58	0.61	0.64	0.66	0.65	0.62	0.57

```
xtable(RESU[,22:31])
```

	NIR22	NIR23	NIR24	NIR25	NIR26	NIR27	NIR28	NIR29	NIR30	density
Min.	1.16	1.15	1.12	1.12	1.07	1.02	1.01	1.03	1.08	0.00
1st Qu.	1.59	1.50	1.43	1.38	1.31	1.23	1.18	1.16	1.16	20.26
Median	1.82	1.71	1.62	1.54	1.46	1.37	1.29	1.23	1.21	31.22
Mean	1.85	1.74	1.65	1.56	1.48	1.41	1.32	1.25	1.21	33.75
3rd Qu.	2.19	2.03	1.89	1.77	1.66	1.56	1.45	1.33	1.26	50.50
Max.	3.03	2.83	2.62	2.43	2.26	2.08	1.80	1.53	1.36	100.00
Sd	0.52	0.46	0.40	0.35	0.31	0.27	0.20	0.12	0.07	26.96

```
library(psych)
psych::pairs.panels(X[,c(31,29,28,26,20,3,5)],
method = "pearson",hist.col = "aquamarine1",density = TRUE,ellipses = FALSE)
```



En el gráfico se observa que existen relaciones lineales tanto positivas como negativas

entre las covariables y la variable de respuesta (density), dentro de las seleccionadas el NIR29 y NIR28 presenta mayor relación lineal. Cabe destacar que entre más cercanos estén los NIR, estos presentan correlaciones más fuertes entre estas, esto indica problemas de multicolinealidad si se ajusta un modelo múltiple con ellas. También cabe aclarar que la relación entre los NIR_i y densidad presentan diferentes tipos de relaciones, dónde observamos relaciones positivas y negativas de un comportamiento lineal, pero que no cuentan con coeficientes de correlación lineales de pearson muy altos, por lo cual, posiblemente el porcentaje de variabilidad explicada por una de estas para nuestra variable de interés no se espera que sea muy alto.

3. Modelo regresión simple

Para ajustar el modelo se procederá a seleccionar el NIR que presente el mejor R^2 y que además se puedan cumplir o realizar correcciones a los supuestos, el cual es el NIR29, por lo tanto, se ajustara el siguiente modelo:

$$y_i = \beta_0 + \beta_1 NIR29_i + \varepsilon_i$$

Esto lo obtuvimos al calcular la matrix de correlaciones entre las covariables, seleccionando el valor máximo y mínimo entre estas, con las siguientes instrucciones en R, allí seleccionamos el valor de NIR 29, puesto que es la variable con la mayor correlación presentada, en este caso negativa. Luego procedimos a realizar el diagrama de dispersión para ver precisamente esta relación, dónde abajo observaremos una relación lineal negativa, con dispersión notoria entre los puntos, y no se evidencia un patrón tan claro, en general este diagrama nos ayuda a identificar que a medida que aumenta NIR29 disminuye la densidad esperada.

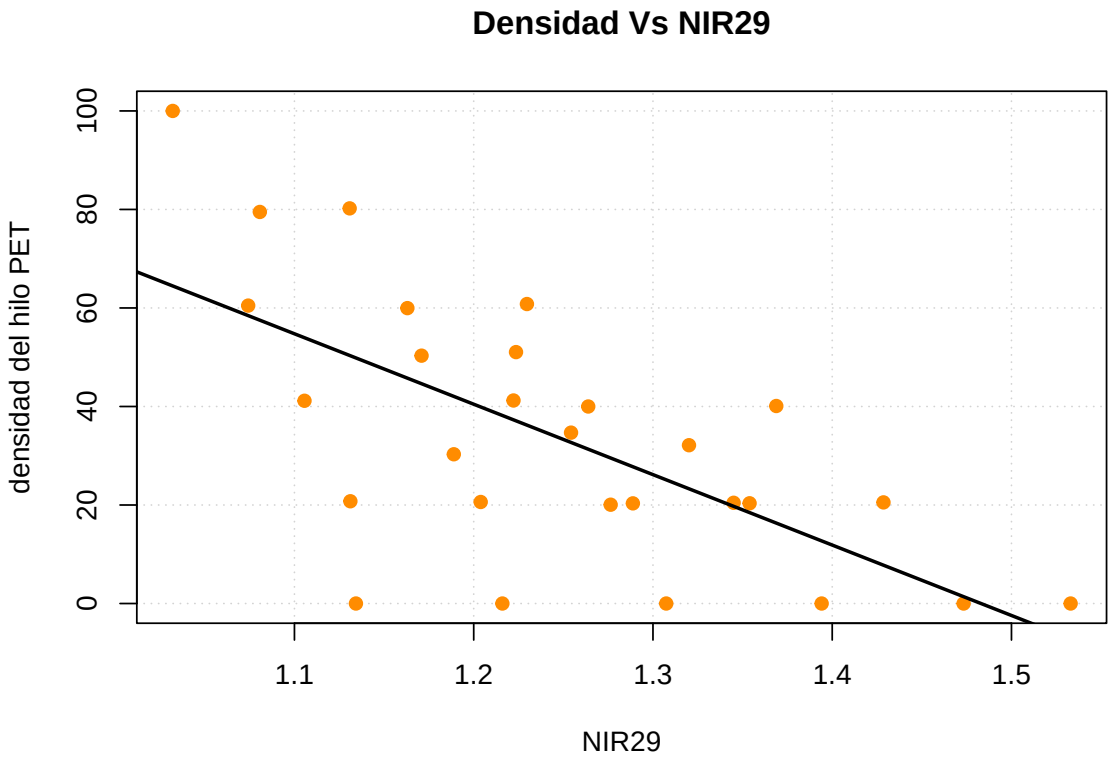
```
Y<-cor(X)
Y<-Y[,-31]
cor<-c(max(Y[31,]),min(Y[31,]))
cor

## [1]  0.5498071 -0.6607189

which(Y[31,]==cor[1]);which(Y[31,]==cor[2])

## NIR4
##    4
## NIR29
##    29
```

```
par(mfrow=c(1,1))
plot(X[,29],X[,31],pch=19,col="#FF8C00",panel.first=grid(),xlab="NIR29",
ylab="densidad del hilo PET",main='Densidad Vs NIR29')
model<- lm(density~NIR29,data=X);abline(model,lwd=2)
```




```
summary(model)

##
## Call:
## lm(formula = density ~ NIR29, data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.856 -11.991   2.011  13.049  35.533
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   212.02      39.91   5.312 1.48e-05 ***
## NIR29        -142.96      31.85  -4.488  0.00013 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.62 on 26 degrees of freedom
## Multiple R-squared:  0.4365,Adjusted R-squared:  0.4149
## F-statistic: 20.14 on 1 and 26 DF,  p-value: 0.0001298
```

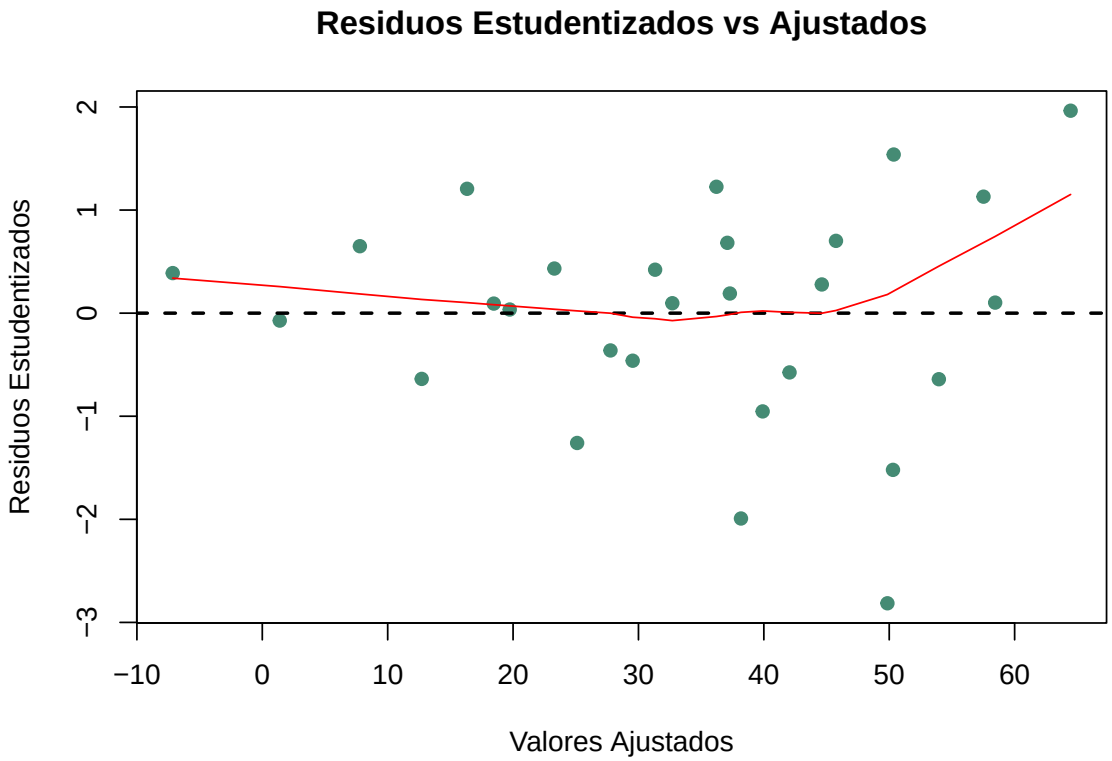
Antes de interpretar los respectivos resultados del modelo se procederá a evaluar supuestos para tener validez en la interpretación,

4. Evaluación Supuestos

Una vez planteado el modelo, se debe verificar si se cumplen los supuestos
En primer lugar, se debe contar con la correcta especificación del modelo y la homogeneidad de las varianzas, es decir, $E(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = \sigma^2$.

```
library(MASS)
library(lmtest)
studenti<- studres(model);ajustados<- fitted.values(model)
```

```
plot(ajustados,studenti, ylab='Residuos Estudentizados',
       xlab='Valores Ajustados',pch=19,col="aquamarine4",
       main="Residuos Estudentizados vs Ajustados")
abline(h=0,lty=2,lwd=2)
lines(lowess(studenti~ajustados), col = "red1")
```



Observando el gráfico se evidencia que a medida que los valores ajustados incrementan, también hay un incremento en los residuos, lo que indicaría problemas de heteroscedasticidad, por ende, se procede a rectificar con la prueba de Breusch Pagan la cual asume que la varianza es una función aditiva de las covariables:

$$\sigma_i^2 = E(\varepsilon_i^2) = \gamma_0 + \gamma_1 x_{i1} + ... + \gamma_{p-1} x_{i,p-1}$$

Por lo tanto, se plantea la siguiente Hipotesis

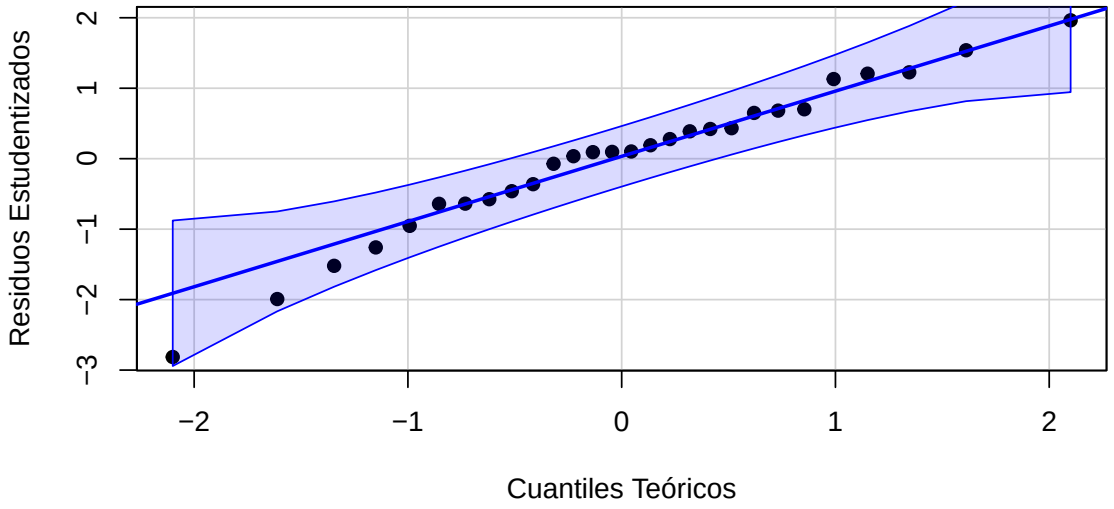
$H_0 : \gamma_1 = \gamma_2 = ... = \gamma_{p-1} = 0$ (Homocedasticidad)
$H_1 : \gamma_j \neq 0$ para algún $j = 1, ..., p-1$ (Heterocedasticidad)

```
bptest(model,~NIR29+I(NIR29^2),data=X)

##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 4.9611, df = 2, p-value = 0.0837
```

En la prueba de Breusch Pagan el valor p se encuentra en el extremo de la cola, el cual no rechazaría la hipótesis si se tiene una confianza del 95 %, sin embargo, teniendo en cuenta el gráfico y que el valor P es muy próximo a la región de rechazo, se concluye que hay problemas de varianza no constante. Por lo tanto, se debe realizar transformaciones para corregirlo.
Ahora probaremos el supuesto de normalidad en los errores, $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. Para esto realizaremos un QQ-Plot y la prueba de normalidad de Shapiro-Wilks

```
qqPlot(studenti,xlab="Cuantiles Teóricos",ylab="Residuos Estudentizados",id=F,pch=19)
```



En el gráfico se observa que todos los puntos caen dentro de las bandas de confianza indicando que los residuos se distribuyen Normal, por ende los errores también, ya que son una función lineal de estos.
El test de Shapiro Wilk es una prueba de bondad de ajuste, entonces, se plantea las siguientes hipótesis:
H_0 : La distribución de los residuos es normal vs H_1 : La distribución de los residuos no es normal

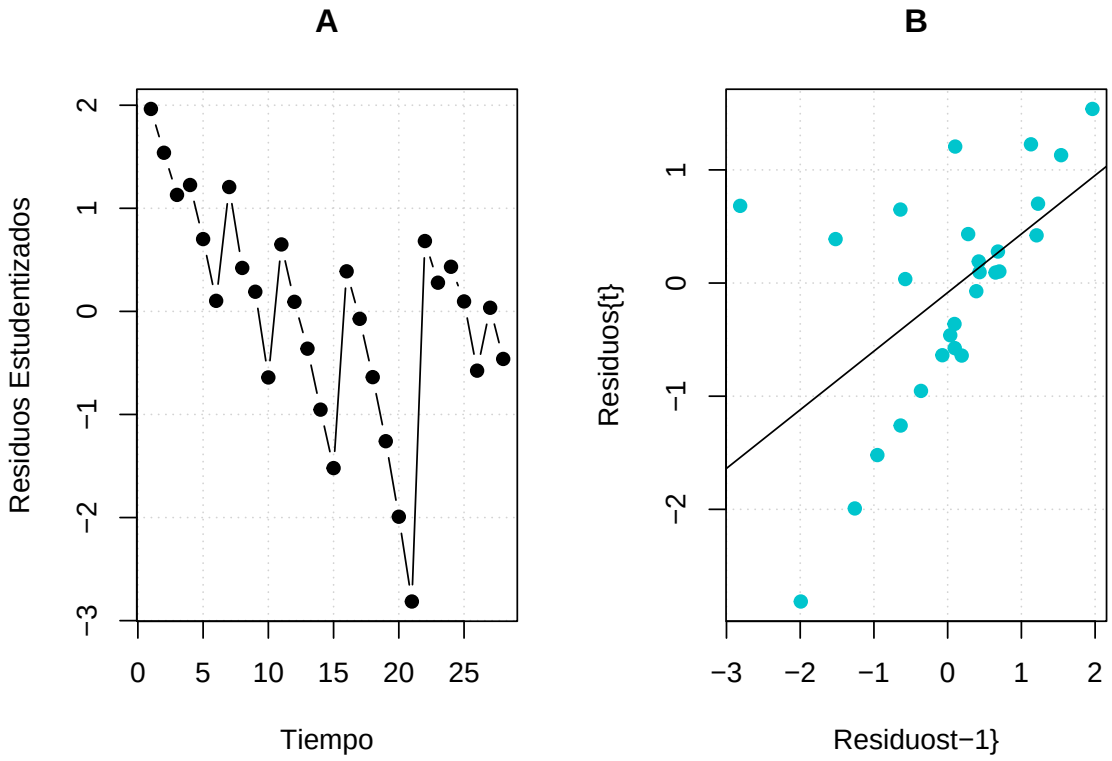
```r
shapiro.test(studenti)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  studenti
## W = 0.97025, p-value = 0.5874
```

Con base en la prueba de ajuste realizada y el gráfico, concluye que los errores del modelo se distribuyen Normal.
Ahora vamos a evaluar el supuesto de independencia entre los errores, al analizar su correlación temporal, donde haremos dos estrategias. Primero el utilizar la prueba de Durbin Watson que define el siguiente esquema de prueba de hipótesis:
$H_0 : \phi = 0$ (Indepedencia) vs $H_1 : \phi \neq 0$ (Correlación temporal)
Junto con un gráfico de los residuos rezagados acompañados de un gráfico de los residuos contra el tiempo. Cabe aclarar que no tenemos un criterio sólido para pensar que los residuos han sido tomados durante el tiempo, es decir, solo lo estamos comprobando, por lo cual, para estar totalmente seguros de la decisión que vayamos a tomar, debemos realizar la prueba con el modelo natural, luego otra desordenando los datos, puesto que si existiera una correlación temporal esta debería mantenerse independiente del orden de los residuos.

```r
#Datos desornedados
Z<- as.data.frame(cbind(X[,31],X[,29]))
colnames(Z)<-c('Density','NIR29')
set.seed(100)
ind<-sample(1:nrow(Z),nrow(Z))
Z<- as.data.frame(Z[ind,])
modelprueba<- lm(Density~NIR29,data=Z)
#Datos ordenados
```

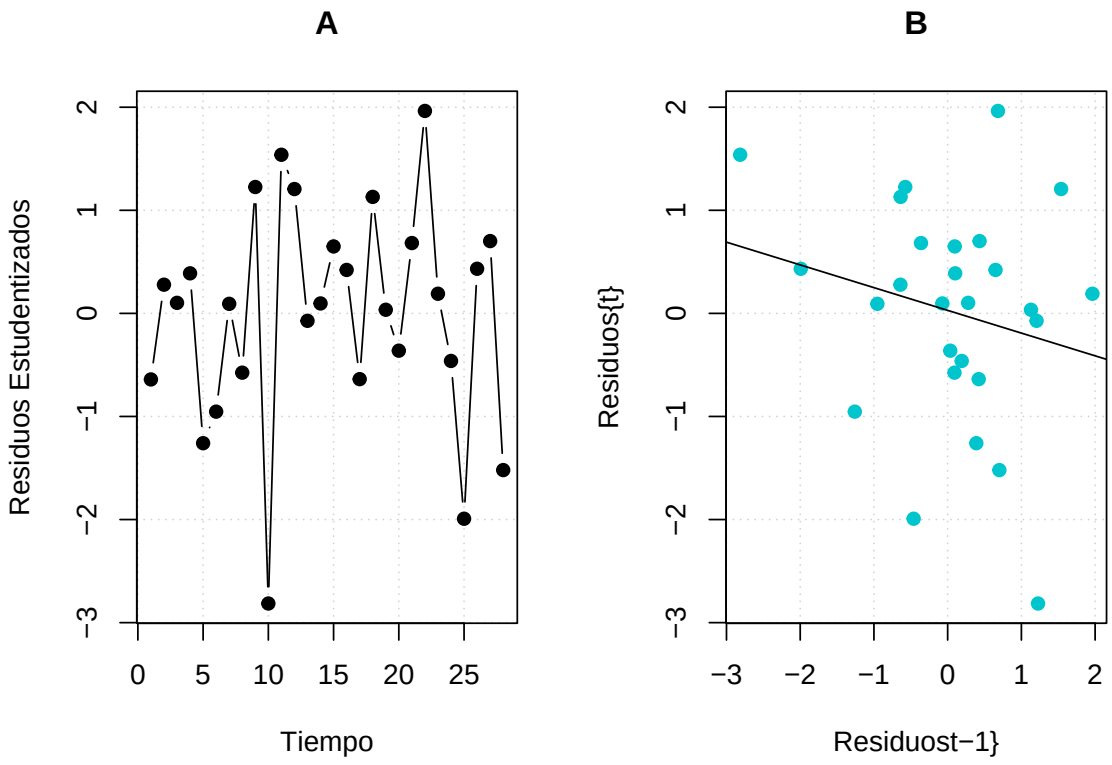
```r
#Validación datos ordenados
par(mfrow=c(1,2))
  plot(studres(model),type="b",xlab="Tiempo",
  ylab="Residuos Estudentizados",main="A",pch=19,panel.first=grid())
  plot(studres(model)[-length(fitted.values(model))],
  studres(model)[-1],pch=19,panel.first = grid(),col="turquoise3",
  xlab="Residuost-1}",ylab="Residuos{t}",main="B")
  abline(lm(studres(model)[-1]~studres(model)[-length(fitted.values(model))]))
```




```
  durbinWatsonTest(model,method='resample',reps=1000)

##  lag Autocorrelation D-W Statistic p-value
##    1       0.5267596     0.8241847       0
##  Alternative hypothesis: rho != 0

par(mfrow=c(1,2))
  plot(studres(modelprueba),type="b",xlab="Tiempo",
  ylab="Residuos Estudentizados",main="A",pch=19,panel.first=grid())
  plot(studres(modelprueba)[-length(fitted.values(modelprueba))],
  studres(modelprueba)[-1],pch=19,panel.first = grid(),col="turquoise3",
  xlab="Residuost-1}",ylab="Residuos{t}",main="B")
  abline(lm(studres(modelprueba)[-1]~studres(modelprueba)
  [-length(fitted.values(modelprueba))]))
```




```
  durbinWatsonTest(modelprueba,method='resample',reps=1000)

##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1942771      2.294844     0.4
##  Alternative hypothesis: rho != 0
```

Dónde en los datos ordenados observamos que el valor p de la prueba nos lleva a pensar que existe correlación temporal, pero al desordenarlos no, por lo cual siempre es pertinente tener la teoría clara, por lo cual para este conjunto de datos asumimos que no existe correlación temporal. En los gráficos de los residuos rezagados nuevamente corroboramos que en los datos ordenados se evidencia una correlación negativa entre los valores que evidenciamos en la velocidad del cambio de las pendientes en el gráfico de los residuos contra el tiempo, pero que no ocurre, cuando se desordenan los datos. Asumiremos independencia, es un supuesto muy fuerte que de cumplirse para este modelo inicial, se deben cumplir para todos independientes de las transformaciones que vayamos a realizar, puesto que son un conjunto de individuos incorrelados temporal, y espacialmente, junto con pertenecer a una muestra aleatoria, nada modificara esto, por lo cual después de evidenciar la independencia temporal asumimos que el supuesto de que $(\epsilon_i, \epsilon_j) = 0$ se cumple para todo $j \neq i$

4.1. Corrección de supuestos por MCP

El método de MCP (minimos cuadrados ponderados) es una alternativa para estimar un modelo lineal que presenta heteroscedasticidad, básicamente calcula las

desviaciones entre las observaciones y_i y los valores ajustados $\hat{y}_i$ usando pesos inversamente proporcionales a la varianza. Se realiza el cálculo de los pesos mediante el modelo del valor absoluto de los residuos en función de las covariables.

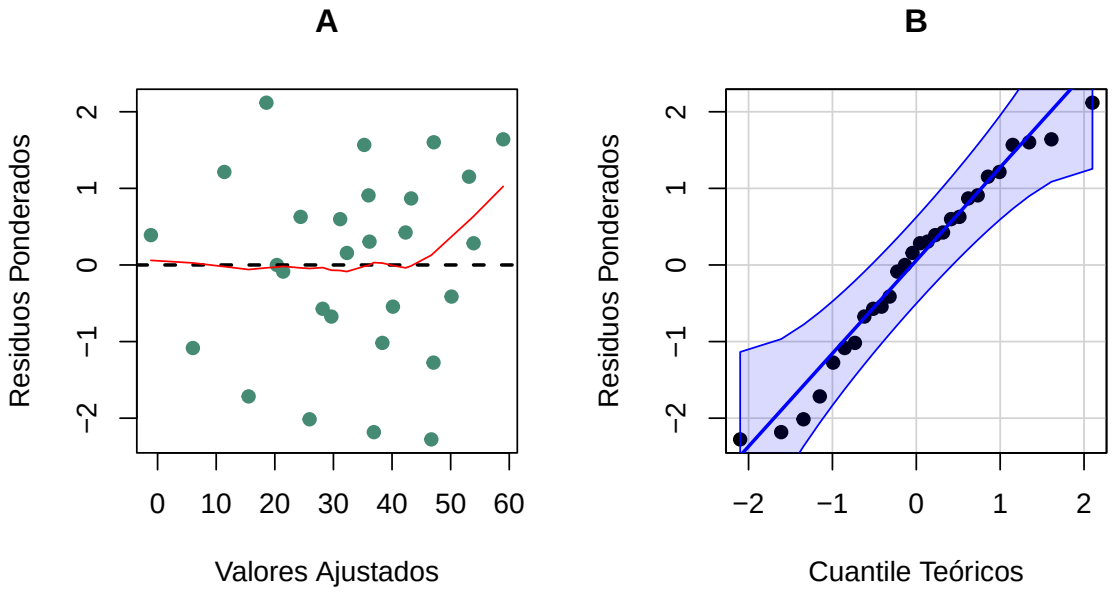
```
res.estu <- residuals(model)
varianza<- lm(abs(res.estu)~NIR29,data=X)
w = 1/(fitted.values(varianza))^2
model.ponderados<- lm(density~NIR29,data=X,weights = w)
```

4.2. Valoración de supuestos del modelo MCP

Se realizan los respectivos gráficos y pruebas mencionadas anteriormente y se observa una mejoría en el gráfico con respecto a la homoscedasticidad, dado que no se evidencia patrones de aumento o disminución de los residuos con respecto a los valores ajustados, por lo tanto, se trabajará con el modelo estimado por MCP, más adelante explicaremos el porqué se usan los residuos ponderados, es pertinente aclarar que para esta estructura de estimación no se realizará la prueba de Breush Pagan, ya que como se mencionó anteriormente, esta asume una estructura de la varianza igual para cada elemento de la diagonal de la matrix de varianzas y covarianzas, y en el método de mínimos cuadrados ponderados esto no ocurre, realizamos estimaciones particulares para cada valor de esta y estipulamos su inverso para generar el equilibrio ponderado, por lo cual el realizar esta prueba nos devolverá como resultado que la varianza es no constante, y claramente es lo que contemplamos. En algún caso remoto puede que nos diga que la varianza es constante, dado que las pruebas de hipótesis cuentan con una potencia que puede fallar, pero con conocimiento teórico sólido sabemos que ocurre, por lo cual solo nos queda la herramienta gráfica para evaluar el comportamiento de los residuos.

```
studenti.ponderados<- residuals(model.ponderados)*sqrt(w)
ajustados.ponderados<- fitted.values(model.ponderados)
```

```
par(mfrow=c(1,2))
plot(ajustados.ponderados,studenti.ponderados, ylab='Residuos Ponderados',
     xlab='Valores Ajustados',pch=19,col="aquamarine4",
     main="A")
abline(h=0,lty=2,lwd=2)
lines(lowess(studenti.ponderados~ajustados.ponderados), col = "red1")
qqPlot(studenti.ponderados,main="B", ylab="Residuos Ponderados",
xlab="Cuantile Teóricos",id=F,pch=19)
```




```
shapiro.test(studenti.ponderados)

##
##  Shapiro-Wilk normality test
##
## data:  studenti.ponderados
## W = 0.96836, p-value = 0.5375
```

5. Modelo por MCP e interpretación

```
summary(model.ponderados)

##
## Call:
## lm(formula = density ~ NIR29, data = X, weights = w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2773 -0.7595  0.2202  0.8785  2.1186
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   182.81      24.97    7.32 8.97e-08 ***
## NIR29        -119.99      17.57   -6.83 3.00e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.239 on 26 degrees of freedom
## Multiple R-squared:  0.6421,Adjusted R-squared:  0.6284
## F-statistic: 46.65 on 1 and 26 DF,  p-value: 2.997e-07
```

En el resumen del modelo contamos con un R^2=0.6679, es decir, el 66.79 % de la variabilidad de la densidad está siendo explicada por el modelo, este R^2 presento un aumento debido a la estimación del modelo por MCP, puesto que anteriormente se contaba con un R^2 de 0.43. Pese a que se espera tener un mayor R^2, es congruente ajustar el modelo debido a que se cuenta con un valor P de 1,114e^{-}07 asociado al estadístico F, que se explicara más adelante.

- Él $\hat{\beta}_0$ indica que la densidad media del Hilo PET es de 181 cuando el NIR29 presenta un valor de 0.
- Él $\hat{\beta}_1$ indica que por cada incremento del NIR29 en una unidad, la densidad media del hilo de PET disminuye en 118.63 unidades. Por lo tanto, existe una relación negativa entre la variable predictora y la variable de respuesta, es decir, a valores bajos del NIR29, el hilo de PET tendrá una mayor densidad.
- Él $\hat{\sigma}$ es el error estándar estimado de la distribución normal de los errores. Si es pequeño quiere decir que se pueden hacer predicciones mas precisas de la densidad media del hilo.

Pruebas t (individuales)
Las pruebas t se construyen a partir de la distribución de los β_j, como este es combinación lineal del y_i, se puede concluir que el $\beta_j \sim N(E(\hat{\beta}_j), v(\hat{\beta}_j)$ y como la varianza es estimada el estadístico de prueba se realiza mediante una t.
Se construyen las siguientes hipótesis:

$$H_0 : \beta_j = 0 \text{ vs } H_1 : \beta j \neq 0 \text{ para j=0,1}$$

- Para él intercepto(β_0) se encontró con un valor t=7,622 y un valor $p = 4{,}33e^{-}08$ lo que indica que él interceptó distinto de 0 tiene un aporte en la estimación del modelo si ya tenemos, incluida la variable NIR19
- Para el β_1 se encontró con un valor t=-7,231 y un valor $p = 1{,}11e^{-}07$. Por lo tanto, rechazamos H_0 y concluimos que la variable NIR29 tiene un aporte en el modelo, es decir, que aunque ya tengamos la media incluida, la variable NIR29 ayuda a predecir la densidad media del hilo PET

Estadístico F
El estadístico de prueba se construye a partir de la partición de la suma de cuadrados totales, también conocida como variabilidad total. La partición se hace sumando y restando los valores ajustados ($\hat{y}_i$), obteniendo la siguiente ecuación.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SST = SS_R + SS_{res}$$

A partir de estos resultados se obtiene la siguiente tabla ANOVA

```
anova(model.ponderados)

## Analysis of Variance Table
##
## Response: density
##           Df Sum Sq Mean Sq F value    Pr(>F)
## NIR29      1 71.604  71.604   46.65 2.997e-07 ***
## Residuals 26 39.908   1.535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Por ende, se dice que la variable β_1 tiene un aporte significativo cuando la suma de cuadrados de la regresión es grande, la cual se evalúa mediante el estadístico F

$$F_0 = \frac{MS_R}{MS_{res}} \sim F_{(1,n-2)}$$

Se plantea la siguiente prueba de hipótesis:

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta 1 \neq 0$$

En el modelo se encuentra un valor de $p = 1{,}114e^{-}07$ asociado al estadístico F, lo que indica que el coeficiente β_1 es distinto de cero y la relación entre la densidad y el NIR29 se puede considerar lineal, por ende, ajustar el modelo con la variable NIR29 es óptimo, en otras palabras, este ayuda predecir la densidad media del hilo PET mediante una recta ajustada.

Residuos estudentizados y estandarizados

El análisis de los residuos es importante tanto para la evaluación de supuestos como para la detección de puntos atípicos, sin embargo, estos residuos se ven influenciados cuando la varianza de los errores no es constante, por eso se realiza la siguiente transformación para estudentizarlos.

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}$$

Al aplicar esto, obtienen varianza 1, por lo que se espera que estos estén alrededor de 0, lo que implicaría que el modelo da buenos valores aproximados a la densidad del hilo PET, comparados con la muestra.

Los residuos estudentizados no se pueden obtener para el cálculo del modelo estimado a partir del modelo por mínimos cuadrado ponderados puedo que como vemos en la expresión anterior se requiere que exista una estimación concreta de varianza, pero contamos con una diferente para cada elemento de la diagonal de la matriz de pesos, por lo cual obtenemos su análogo que serían los residuos ponderados lo cuales se obtienen de la siguiente forma en R-Studio, la idea consiste en multiplicar los residuos por la raíz de los pesos, para poder hacer la comparación de estos mismos, puesto que

si los graficamos en bruto como vienen calculados obtendremos que son no constantes, puesto que está es la estructura obtenida en el modelo, pero pueden ser estudiados, si tuviéramos un modelo realizado por transformaciones en potencia ya sea de manera empírica o por el método de BOX-COX si podríamos estudentizarlos de manera normal.

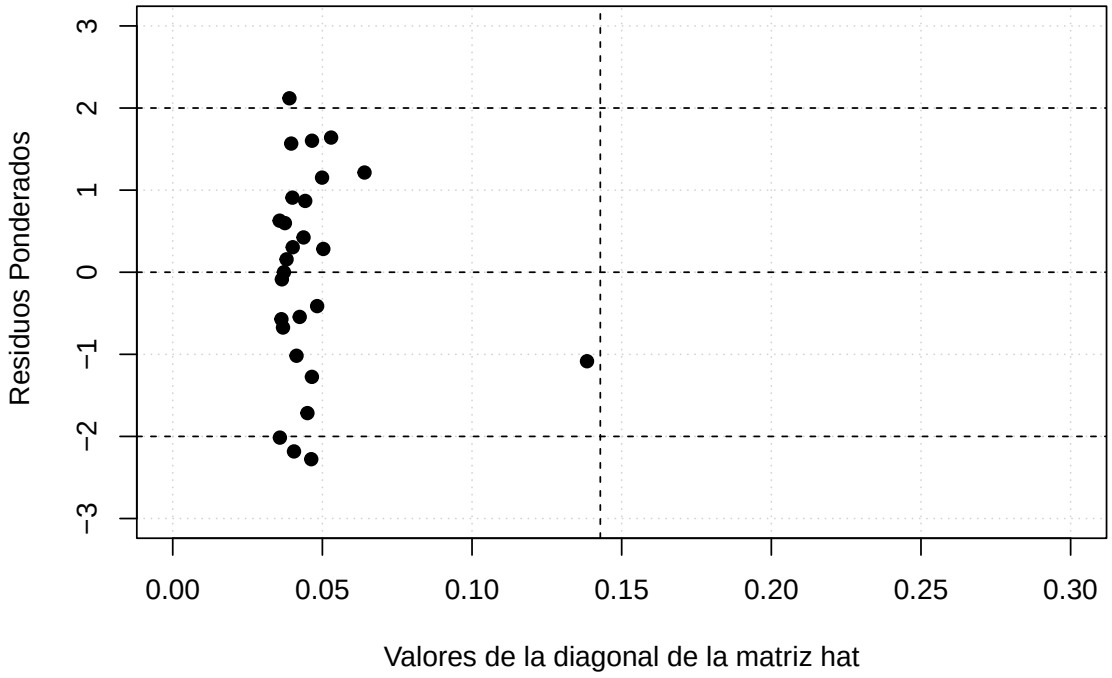
```
studenti.ponderados<- residuals(model.ponderados)*sqrt(w)
```

6. Evaluación de puntos influyentes y atipicos

Para identificar los puntos influyentes y/o atípicos, primeramente se procederá a realizar el gráfico de los residuos ponderados vs los valores de la diagonal de matriz hat, generando para el eje X un punto de corte en el valor de $2\frac{p}{n}$ y en el eje Y de -2 y 2. Esto será un punto de corte arbitrario, puesto que la transformación que nos resultó pertinente para nuestro cálculo asume una estructura de residuos que no se pueden estudentizar, pero que lo hacemos bajo el análogo como si se pudieran realizar. Este criterio gráfico se usa porque una observación no basta que sea solo inusual en su residuo, sino también en su posición de estimación, en la matriz de estimación H y en su respectiva posición h_{ii}, la combinación de esta condición caracteriza un posible punto influyente, un residuo grande una observación atípica, y una con valor grande en la matriz de estimación un punto de balanceo.

```
res.ponderados<- residuals(model.ponderados)*sqrt(w)
library(car)
par(mfrow=c(1,1))
p<- length(coefficients(model.ponderados))
n<- nrow(X)
hii.c<- 2*p/n
hii<- hatvalues(model.ponderados)
hii.ind<- hii[hii>hii.c]
n<- length(residuals(model.ponderados))
p<- length(coefficients(model.ponderados))
hii.c<-2*(p/n)
```

```
plot(hii,res.ponderados,pch=19,xlab="Valores de la diagonal de la matriz hat",
ylab=" Residuos Ponderados",ylim=c(-3,3),xlim=c(0,0.3),panel.first=grid())
abline(h=c(1,0,-1)*2,lty=2,v=hii.c)
```



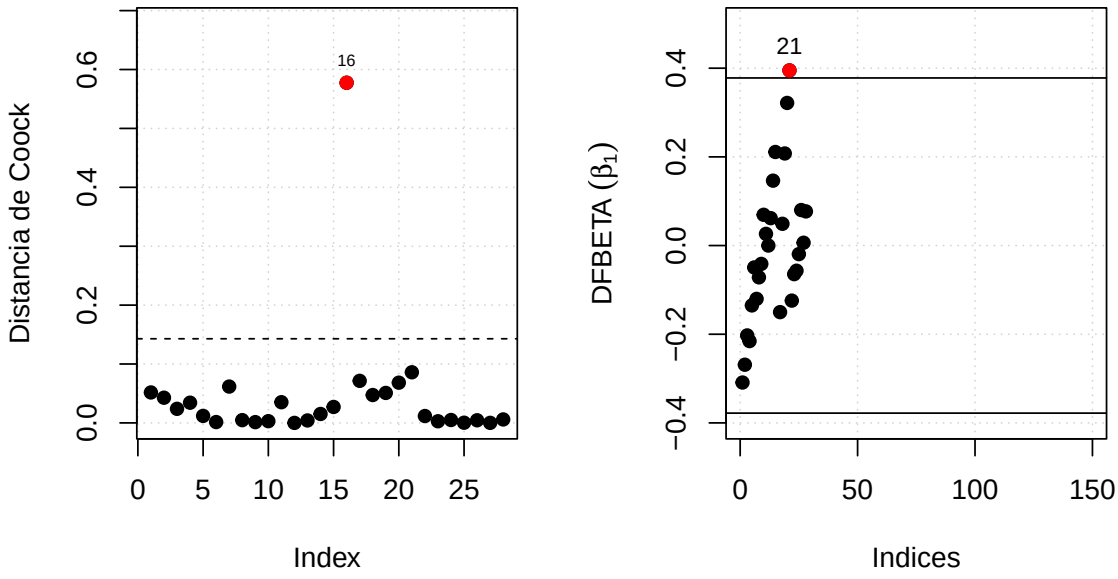
En el gráfico encontramos que no hay ningún punto que sea influyente y/o atípicos. Para confirmar se procederá a ver las medidas de influencia, como lo es la distancia de Cook, DFBETAS, DFFITS, COVRATIO. Estos procedimientos evalúan los cambios que ocurren en el modelo si elimino la i-ésima observación.

6.1. Distancia de Cook - DFBETAS

```
par(mfrow=c(1,2))
ck<- cooks.distance(model.ponderados)
plot(ck,ylab="Distancia de Coock",pch=19,ylim=c(min(ck),max(ck)+0.1),
panel.first=grid())
ck.c<- 4/n
abline(h=ck.c,lty=2)
indices<- (1:nrow(X))[ck>ck.c]
ck<- ck[ck>ck.c]
points(indices,ck,col="red",pch=19)
text(indices,ck,labels=rownames(X)[indices],pos=3,cex=0.6)


DFBETAS = dfbetas(model.ponderados)
plot(DFBETAS[,2],ylab=quote('DFBETA'~(beta[1])),xlab="Indices",
pch=19,ylim=c(-0.4,0.5),xlim=c(0,150),panel.first=grid())
ind = (1:nrow(X))[abs(DFBETAS[,2]) > 2/sqrt(nrow(X))]
```

```
dfb = DFBETAS[abs(DFBETAS[,2]) > 2/sqrt(nrow(X)) ,2]
abline(h=c(1,-1)*2/sqrt(nrow(X)))
text(ind,dfb,rownames(X)[abs(DFBETAS[,2]) > 2/sqrt(nrow(X))],
pos=c(1,3,1,4,3,2,1,4,3,4),
     cex=0.8)
points(ind,dfb,col="red",pch=19)
```



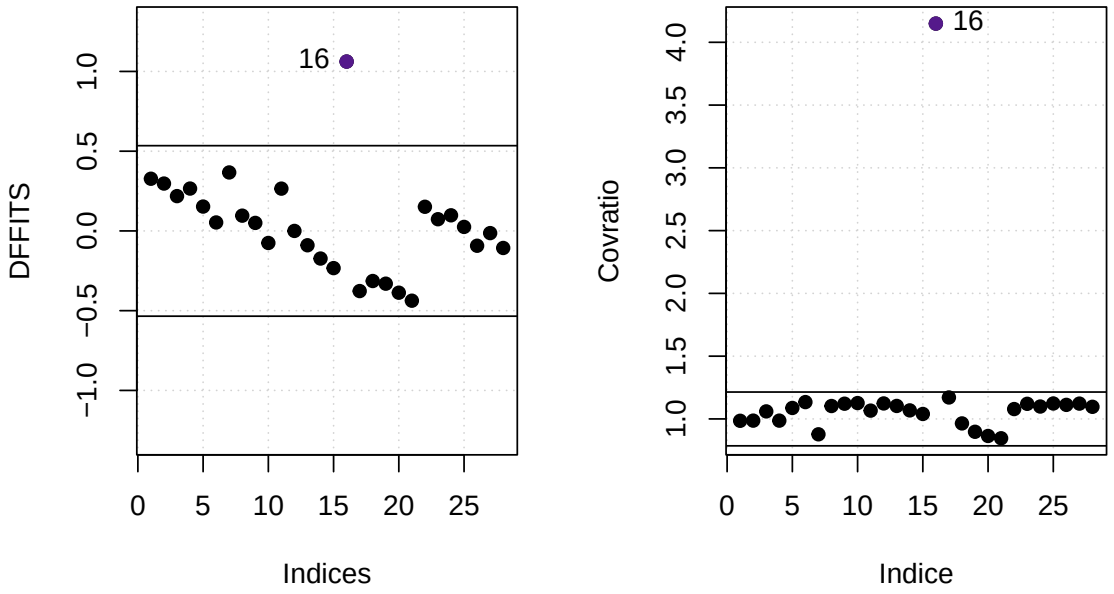
- La distancia de Cook es un indicador global de cuanto cambian las estimaciones de los coeficientes al eliminar la i-esima observación, según Mongotmery.2011 el punto de cohorte para evidenciar un punto influyente es de 4/n, con n igual al número de inviduos. en este caso se evidencia que la observación 16 presenta cambios en estas estimaciones.
- El DFBETAS es un indicador de cuanto cambia la estimación del j-esimo Beta al eliminar la i-esima observación, en este caso se evaluó él β_1, y se evidencia que la observación 21 podría ser influyente en esa estimación. El punto de cohorte según Mongotmery 2011 es $\pm \frac{2}{\sqrt{n}}$.

6.2. DFFITS - COVRATIO

```
par(mfrow=c(1,2))
DFFITS = dffits(model.ponderados)
plot(DFFITS,xlab="Indices",pch=19,ylim=c(-1.3,1.3),panel.first=grid())
abline(h=c(-1,1)*2*sqrt(p/n))
ind = (1:nrow(X))[abs(DFFITS) > 2*sqrt(p/n)]
```

```
dfb = DFFITS[abs(DFFITS) > 2*sqrt(p/n)]
text(ind,dfb,rownames(X)[abs(DFFITS) > 2*sqrt(p/n)],pos=2)
points(ind,dfb,col="purple4",pch=19)

COVR = covratio(model.ponderados)
plot(COVR,pch=19,ylab="Covratio",xlab="Indice",panel.first=grid())
abline(h=1+c(-1,1)*3*(p/n))
covr = COVR[COVR > 1 +3*(p/n) | COVR < 1 -3*(p/n) ]
ind = (1:nrow(X))[COVR > 1 +3*(p/n) | COVR < 1 -3*(p/n) ]
text(ind,covr,rownames(X)[COVR > 1 +3*(p/n) | COVR < 1 -3*(p/n)],pos=4)
points(ind,covr,col="purple4",pch=19)
```



- El DFFITS es indicador de cuanto cambian los valores ajustados ($\hat{y}_i$) al eliminar la i-ésima observación, en este caso se evidencia que la observación 16 presenta cambios en la predicción. El punto de cohorte según Mongotmery 2011 es $\pm 2\sqrt{\frac{n}{p}}$.
- El COVRATIO es un indicador de cuanto cambia la matriz de varianza de los Betas al eliminar la i-ésima observación, es decir, la precisión general de la estimación, evidencia que la observación 16 podría ser influyente en esta matriz. El punto de cohorte según Mongotmery 2011 es $1 \pm 3\frac{p}{n}$.

Como conclusión general de las medidas de influencia y gráficos realizados, es pertinente evaluar el hilo de PET correspondiente a la observación 16. si esta observación fue tomada adecuadamente y en caso de no serlo, evaluar si se elimina o no, con ayuda de los expertos en el tema.

7. Intervalos de confianza

7.1. Intervalo de confianza para β

Como se menciona anteriormente, $\hat{\beta} \sim N(\beta, v(\beta))$, por ende, se pueden construir intervalos de confianza.

```
confint(model.ponderados)
```

```
##                 2.5 %    97.5 %
## (Intercept)  131.4769 234.14711
## NIR29       -156.1064 -83.88172
```

Para β_1, con un nivel de confianza del 95 % podemos decir que cuando el NIR29 aumenta en una unidad, la densidad media del hilo PET disminuirá entre 156.1 y 83.8 unidades.
Para β_0 con un nivel de confianza del 95 % la densidad media del Hilo PET es entre 131.4 y 234.1 cuando el NIR29 presenta un valor de 0
el nivel de confianza hace referencia a que si repito el mismo experimento 100 veces, en 95 de ellos, el parámetro estará en esos rangos mencionados.

7.2. Intervalo de confianza para $E(Y|X = x_0)$

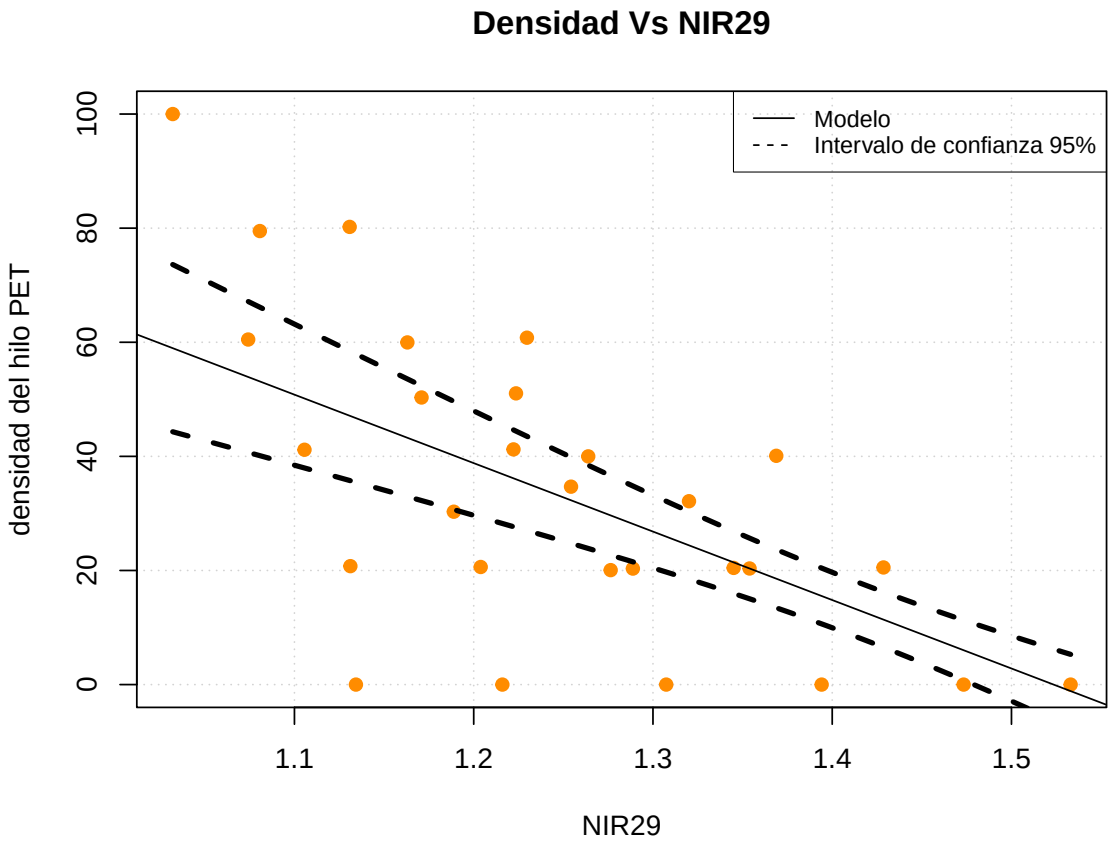
Se realizará un gráfico en el que se evaluaran 100 puntos teniendo en cuenta el no extrapolar y a estos se les obtendrá los intervalos de confianza para el valor esperado.

```
x.nuevo = data.frame(NIR29=seq(min(X[,29]),max(X[,29]),length.out=100))

#Predicción del intervalo de confianza
pred.media = predict(model.ponderados,x.nuevo,interval = "confidence")
```

```
plot(X$NIR29,X$density,pch=19,col="#FF8C00",panel.first=grid(),xlab="NIR29",
ylab="densidad del hilo PET",main='Densidad Vs NIR29')
#Gráficas de las lineas
lines(x.nuevo[,1],pred.media[,2],lty=2,lwd=3)
lines(x.nuevo[,1],pred.media[,3],lty=2,lwd=3)

abline(model.ponderados)
legend(x = "topright",legend=c("Modelo","Intervalo de confianza 95%"),
       lty = c(1, 2,3),pt.cex=1.5,
       box.lwd=0.6,text.font =15,cex=0.8) #Caja de enunciados
```



En el gráfico se observa que el modelo estimado, junto con sus intervalos de confianza, presenta buenas predicciones a la densidad media de los hilos de PET. Se realizará un ejemplo de un valor para entender la interpretación en términos contextuales

```
ejemplo <- data.frame(NIR29=1.3)
pred.media = predict(model.ponderados,ejemplo,interval = "confidence")
pred.media

##        fit     lwr      upr
## 1 26.81972 20.4452 33.19424
```

Esto indica que con un nivel de confianza del 95 %, la densidad media de los hilos de PET de un NIR29 igual a 1.3 está entre 20.44 y 33.19

8. Modelo con variables estandarizadas

Primero estandarizamos las variables de la base de datos.

```
Z <- data.frame(scale(X))
```

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

	NIR1	NIR2	NIR3	NIR4	NIR5	NIR6	NIR7	NIR8	NIR9	NIR10	NIR11
1	-0.43	0.19	0.66	0.96	1.00	0.86	0.59	0.22	-0.17	-0.52	-0.79
2	-0.41	0.18	0.59	0.82	0.93	0.91	0.78	0.56	0.32	0.09	-0.09
3	-0.28	0.29	0.56	0.62	0.53	0.34	0.07	-0.22	-0.51	-0.76	-0.95
4	-0.15	0.30	0.62	0.84	0.97	1.03	1.00	0.91	0.78	0.65	0.55
5	0.20	0.36	0.51	0.59	0.57	0.44	0.29	0.13	-0.03	-0.16	-0.26
6	-0.17	0.18	0.28	0.14	-0.10	-0.34	-0.57	-0.78	-0.97	-1.11	-1.21

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

	NIR12	NIR13	NIR14	NIR15	NIR16	NIR17	NIR18	NIR19	NIR20	NIR21
1	-1.00	-1.13	-1.22	-1.29	-1.38	-1.47	-1.52	-1.52	-1.47	-1.40
2	-0.24	-0.33	-0.38	-0.41	-0.47	-0.54	-0.59	-0.60	-0.58	-0.55
3	-1.09	-1.18	-1.24	-1.29	-1.35	-1.41	-1.43	-1.42	-1.38	-1.32
4	0.46	0.41	0.39	0.39	0.36	0.32	0.29	0.27	0.26	0.26
5	-0.34	-0.38	-0.41	-0.42	-0.45	-0.49	-0.51	-0.52	-0.51	-0.50
6	-1.28	-1.32	-1.35	-1.38	-1.41	-1.43	-1.43	-1.41	-1.38	-1.34

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

	NIR22	NIR23	NIR24	NIR25	NIR26	NIR27	NIR28	NIR29	NIR30	density
1	-1.34	-1.28	-1.24	-1.25	-1.34	-1.45	-1.57	-1.72	-1.85	2.46
2	-0.53	-0.51	-0.49	-0.52	-0.64	-0.76	-0.86	-0.93	-0.96	1.72
3	-1.27	-1.22	-1.18	-1.18	-1.23	-1.27	-1.32	-1.33	-1.24	1.70
4	0.25	0.24	0.25	0.20	0.08	-0.06	-0.13	-0.14	-0.10	1.00
5	-0.49	-0.48	-0.47	-0.49	-0.56	-0.63	-0.67	-0.67	-0.64	0.97
6	-1.31	-1.28	-1.26	-1.26	-1.28	-1.30	-1.33	-1.39	-1.44	0.99

```
modelz<- lm(density~NIR29,data=Z)
summary(modelz)

##
## Call:
## lm(formula = density ~ NIR29, data = Z)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84911 -0.44474  0.07458  0.48396  1.31790
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.874e-16  1.446e-01   0.000  1.00000
## NIR29       -6.607e-01  1.472e-01  -4.488  0.00013 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7649 on 26 degrees of freedom
## Multiple R-squared:  0.4365,Adjusted R-squared:  0.4149
## F-statistic: 20.14 on 1 and 26 DF,  p-value: 0.0001298
```

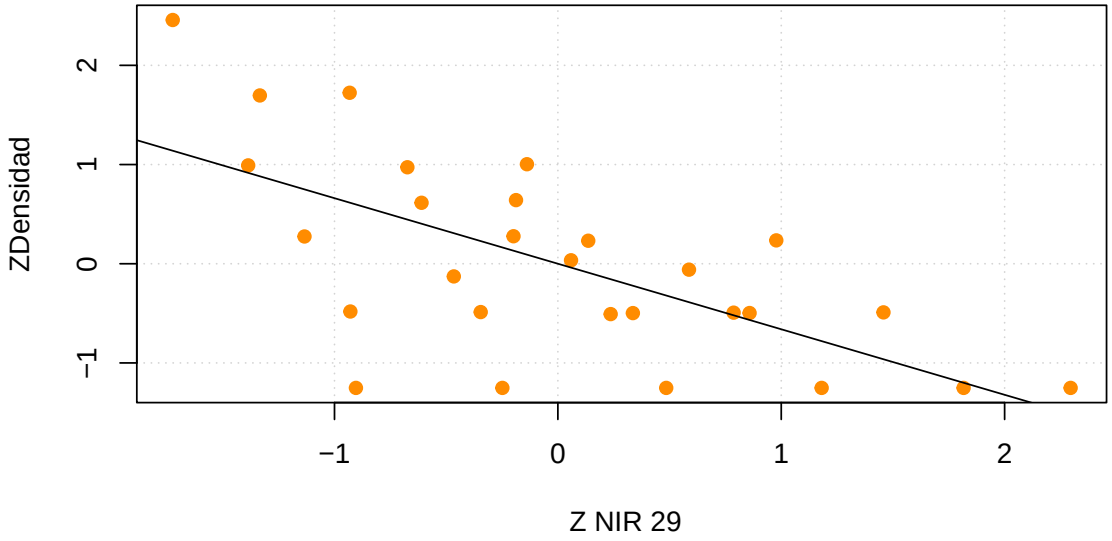
Procederemos a realizar los mismos cálculos, pero con el modelo escalonado cabe aclarar que esta es una transformación lineal 1 a 1 por lo cual ningún supuesto debería arreglarse, sumando a esto que si evaluamos los valores p de las pruebas t asociadas coinciden con las del modelo original, a excepción de la del intercepto que explicaremos su por qué a continuación. El R^2 se mantiene respecto al modelo original. El valor del intercepto es aproximadamente 0, debido a que al realizar un proceso de estandarización la estamos trabajando en una misma unidad de medida, y podemos compararlas en su longitud como si fueran desviaciones estándar, es decir, el proceso análogo a una distribución normal estándar, dónde la media es igual a 0, por lo cual este proceso de centrado nos generará que este se vuelva aproximadamente 0 y cualquier prueba de hipótesis nos debe corroborar que efectivamente se encuentra ubicado aquí. A diferencia de la pendiente asociada a la variable NIR29 escalonada, esta tiene un aporte significativo, es decir, nos ayuda a entender la variable densidad, mantiene la relación negativa.

Nota 1: En regresión lineal múltiple, este procedimiento nos ayudaría a entender que variables dentro de la combinación lineal para generar la predicción de la variable dependiente tiene aportes más significativos.

Nota 2: En general, ninguna gráfica o supuesto debería tener algún cambio más allá de la unidad de medida

En las gráficas que se presentara a continuación vemos que efectivamente las relaciones como en el diagrama de dispersión inicial se mantiene.

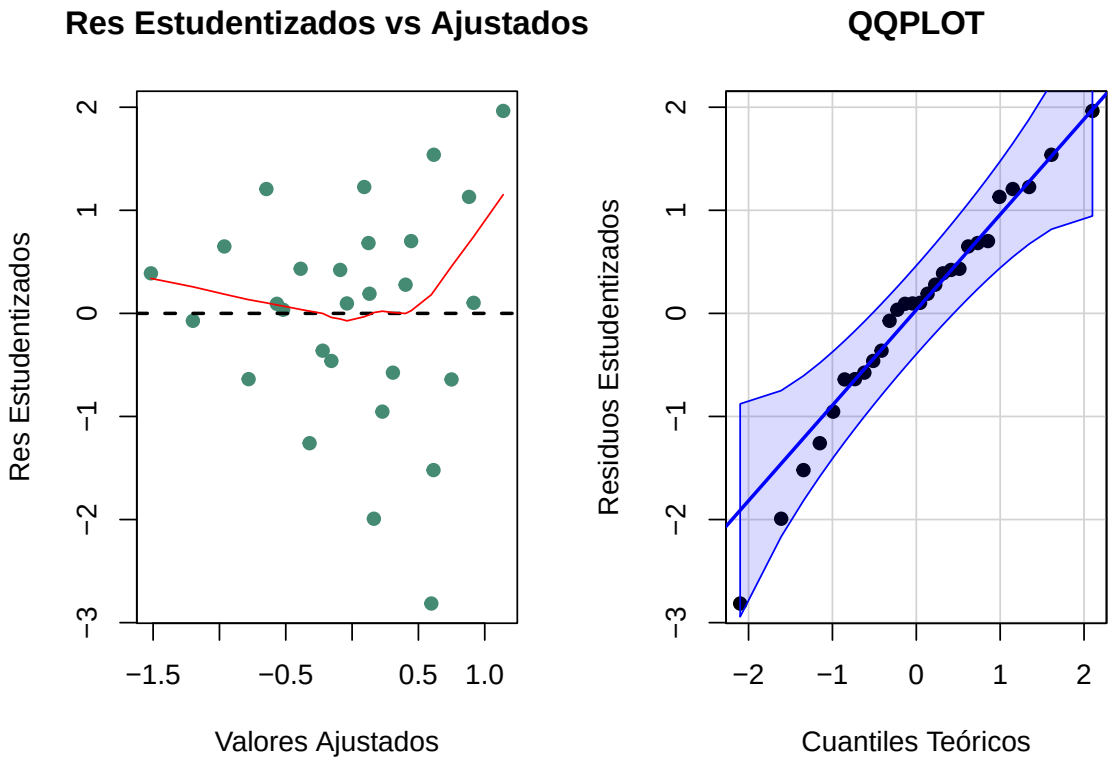
```
plot(Z[,29],Z[,31],ylab='ZDensidad',
xlab=' Z NIR 29',panel.first=grid(),col="#FF8C00",pch=19)
abline(modelz)
```



9. Evaluación de supuestos

```
library(MASS)
library(lmtest)
studenti.<- studres(modelz);ajustados.<- fitted.values(modelz)
```

```
par(mfrow=c(1,2))
plot(ajustados.,studenti., ylab='Res Estudentizados',
       xlab='Valores Ajustados',pch=19,col="aquamarine4",
       main="Res Estudentizados vs Ajustados")
abline(h=0,lty=2,lwd=2)
lines(lowess(studenti.~ajustados.), col = "red1")
qqPlot(studenti.,xlab="Cuantiles Teóricos",ylab="Residuos Estudentizados",id=F,
pch=19,main="QQPLOT")
```




```
bptest(modelz,~NIR29+I(NIR29^2),data=Z)

##
##  studentized Breusch-Pagan test
##
## data:  modelz
## BP = 4.9611, df = 2, p-value = 0.0837

shapiro.test(studenti.)

##
##  Shapiro-Wilk normality test
##
## data:  studenti.
## W = 0.97025, p-value = 0.5874
```

Observamos las mismas conclusiones que en el modelo original, no se evidencia homoscedasticidad y el supuesto de normalidad si se cumple. Procederemos a realizar la transformación de mínimos cuadrados ponderados, realizada con anterioridad.

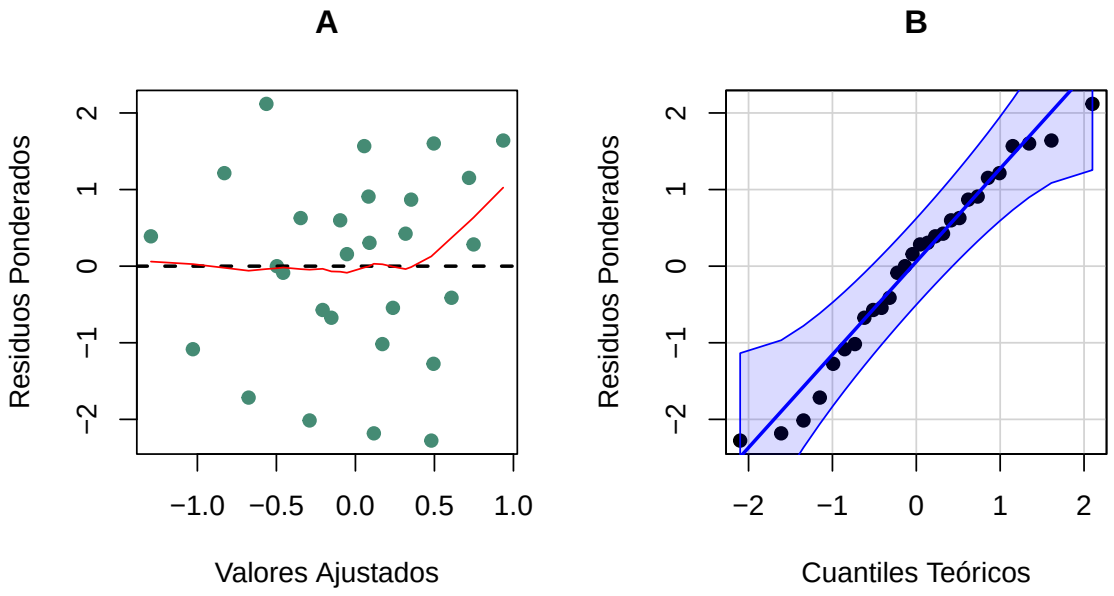
```
res.estuz <- residuals(modelz)
varianzaz<- lm(abs(res.estuz)~NIR29,data=Z)
wz= 1/(fitted.values(varianzaz))^2
model.ponderadosz<- lm(density~NIR29,data=Z,weights = w)
```

9.1. Valoración de supuestos del modelo MCP

Se realizan los respectivos gráficos y pruebas mencionadas anteriormente y se observa una mejoría en el gráfico con respecto a la homoscedasticidad, por lo tanto, se trabajará con el modelo estimado por MCP, no hay diferencias significativas entre este modelo y el original utilizando el mismo método, más allá de la escala de los valores ajustados. El valor p de la prueba de Shapiro Wilk coincide, por lo cual las decisiones son análogas.

```
studenti.ponderadosz<- residuals(model.ponderadosz)*sqrt(wz)
ajustados.ponderadosz<- fitted.values(model.ponderadosz)
```

```
par(mfrow=c(1,2))
plot(ajustados.ponderadosz,studenti.ponderadosz, ylab='Residuos Ponderados',
     xlab='Valores Ajustados',pch=19,col="aquamarine4",
     main="A")
abline(h=0,lty=2,lwd=2)
lines(lowess(studenti.ponderadosz~ajustados.ponderadosz), col = "red1")
qqPlot(studenti.ponderados,main="B", ylab="Residuos Ponderados",
xlab="Cuantiles Teóricos",id=F,pch=19)
```




```
shapiro.test(studenti.ponderados)

##
##  Shapiro-Wilk normality test
##
## data:  studenti.ponderados
## W = 0.96836, p-value = 0.5375
```

10. Modelo por MCP estandarizado e interpretación

```
summary(model.ponderadosz)

##
## Call:
## lm(formula = density ~ NIR29, data = Z, weights = w)
##
## Weighted Residuals:
##       Min        1Q    Median        3Q       Max
## -0.084463 -0.028169  0.008166  0.032582  0.078576
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02098    0.14005   -0.15    0.882
## NIR29       -0.55457    0.08119   -6.83    3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04595 on 26 degrees of freedom
## Multiple R-squared:  0.6421,Adjusted R-squared:  0.6284
## F-statistic: 46.65 on 1 and 26 DF,  p-value: 2.997e-07
```

Vemos que las significancias de las pendientes coinciden, a su vez él R^2, por ende se tienen las mismas interpretaciones del modelo no estandarizado, en el sentido del aporte significativo de la explicación de la variabilidad de la densidad. Para este caso el valor del intercepto es estadísticamente igual a 0. lo cual tiene sentido, ya que al estandarizar las variables, se entiende que se trasladan al origen y carecen de unidad de medida. Se procederá a interpretar las estimaciones y el valor t del intercepto, que son las únicas diferencias entre los modelos. Cabe recalcar que las interpretaciones se hacen en términos de desviaciones estándar.

- El $\hat{\beta}_0$ indica que la densidad media del hilo de PET es de 0.02 desviaciones estándar, cuando el NIR29 presenta 0 desviaciones estándar
- El $\hat{\beta}_1$ índica que por cada aumento en una desviación estándar del NIR29, La densidad media del hilo PET disminuye en 0.55 desviaciones estándar
- Para él intercepto(β_0) se encontró con un valor t=7,622 y un valor p = 4,33e08 lo que indica que es igual a 0, en donde ya mencionamos anteriormente la razón. Esto quiere decir que si ya tenemos incluida la variable NIR29, el intercepto no es necesario, esto se debe a que la variable está en términos de unidades de desviaciones estándar, por lo que técnicamente pasa por el origen.

11. Evaluación de puntos influyentes y atípicos

Como se menciona anteriormente, estandarizar la base de datos no afecta en la manera de como se comportan los mismos, por eso se espera que aparezcan los

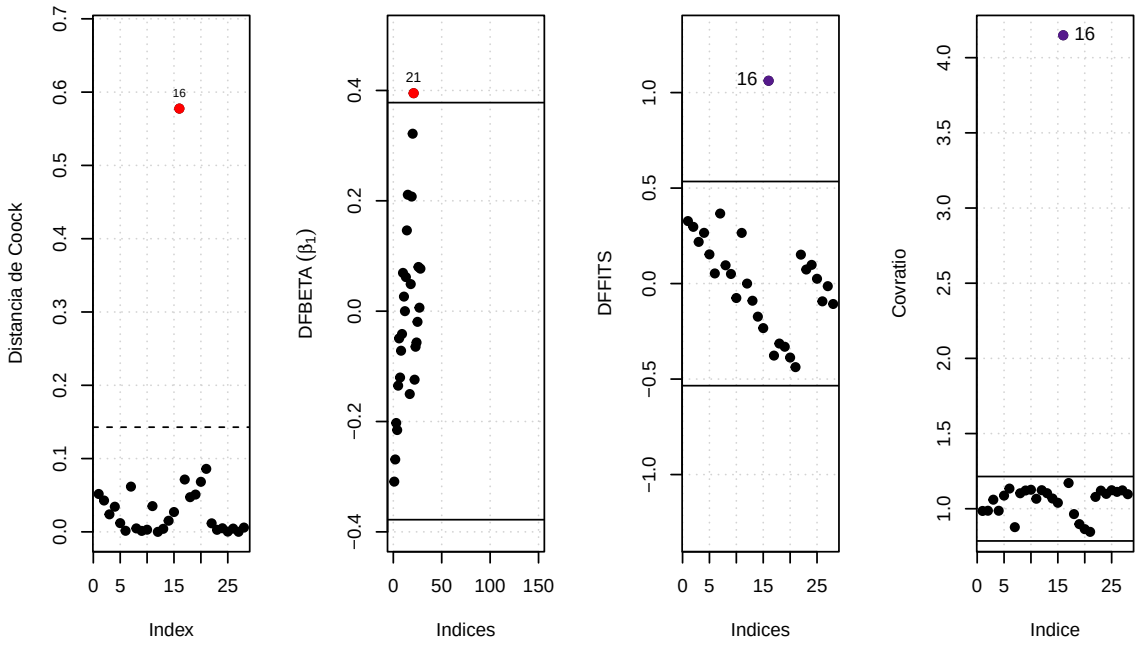
mismos puntos influyentes en las medidas de influencia, lo cual evidenciamos a continuación con las gráficas.

```
par(mfrow=c(1,4))
ck<- cooks.distance(model.ponderadosz)
plot(ck,ylab="Distancia de Coock",pch=19,ylim=c(min(ck),max(ck)+0.1),
panel.first=grid())
ck.c<- 4/n
abline(h=ck.c,lty=2)
indices<- (1:nrow(Z))[ck>ck.c]
ck<- ck[ck>ck.c]
points(indices,ck,col="red",pch=19)
text(indices,ck,labels=rownames(X)[indices],pos=3,cex=0.6)


DFBETAS = dfbetas(model.ponderadosz)
plot(DFBETAS[,2],ylab=quote('DFBETA'~(beta[1])),xlab="Indices",
pch=19,ylim=c(-0.4,0.5),xlim=c(0,150),panel.first=grid())
ind = (1:nrow(Z))[abs(DFBETAS[,2]) > 2/sqrt(nrow(Z))]
dfb = DFBETAS[abs(DFBETAS[,2]) > 2/sqrt(nrow(Z)) ,2]
abline(h=c(1,-1)*2/sqrt(nrow(Z)))
text(ind,dfb,rownames(Z)[abs(DFBETAS[,2]) > 2/sqrt(nrow(Z))],
pos=c(1,3,1,4,3,2,1,4,3,4),
     cex=0.8)
points(ind,dfb,col="red",pch=19)


DFFITS = dffits(model.ponderadosz)
plot(DFFITS,xlab="Indices",pch=19,ylim=c(-1.3,1.3),panel.first=grid())
abline(h=c(-1,1)*2*sqrt(p/n))
ind = (1:nrow(X))[abs(DFFITS) > 2*sqrt(p/n)]
dfb = DFFITS[abs(DFFITS) > 2*sqrt(p/n)]
text(ind,dfb,rownames(X)[abs(DFFITS) > 2*sqrt(p/n)],pos=2)
points(ind,dfb,col="purple4",pch=19)

COVR = covratio(model.ponderadosz)
plot(COVR,pch=19,ylab="Covratio",xlab="Indice",panel.first=grid())
abline(h=1+c(-1,1)*3*(p/n))
covr = COVR[COVR > 1 +3*(p/n) | COVR < 1 -3*(p/n) ]
ind = (1:nrow(X))[COVR > 1 +3*(p/n) | COVR < 1 -3*(p/n) ]
text(ind,covr,rownames(X)[COVR > 1 +3*(p/n) | COVR < 1 -3*(p/n)],pos=4)
points(ind,covr,col="purple4",pch=19)
```



12. Intervalos de confianza

12.1. Intervalo de confianza para β

Se realiza el mismo procedimiento que el modelo sin estandarizar y se obtienen los siguientes resultados

```
confint(model.ponderadosz)

##                  2.5 %     97.5 %
## (Intercept) -0.3088654  0.2669002
## NIR29       -0.7214676 -0.3876711
```

Para β_1, con un nivel de confianza del 95 % podemos decir que cuando el NIR29 aumenta en una desviación estándar, la densidad media del Hilo PET disminuye entre 0.72 y 0.38 desviaciones estándar
Para β_0 con un nivel de confianza del 95 % la densidad media del Hilo PET es entre 0.3 y 0.26 desviaciones estándar cuando el NIR29 presenta 0 desviaciones estándar.

12.2. Intervalo de confianza para $E(Y|Z = z_0)$

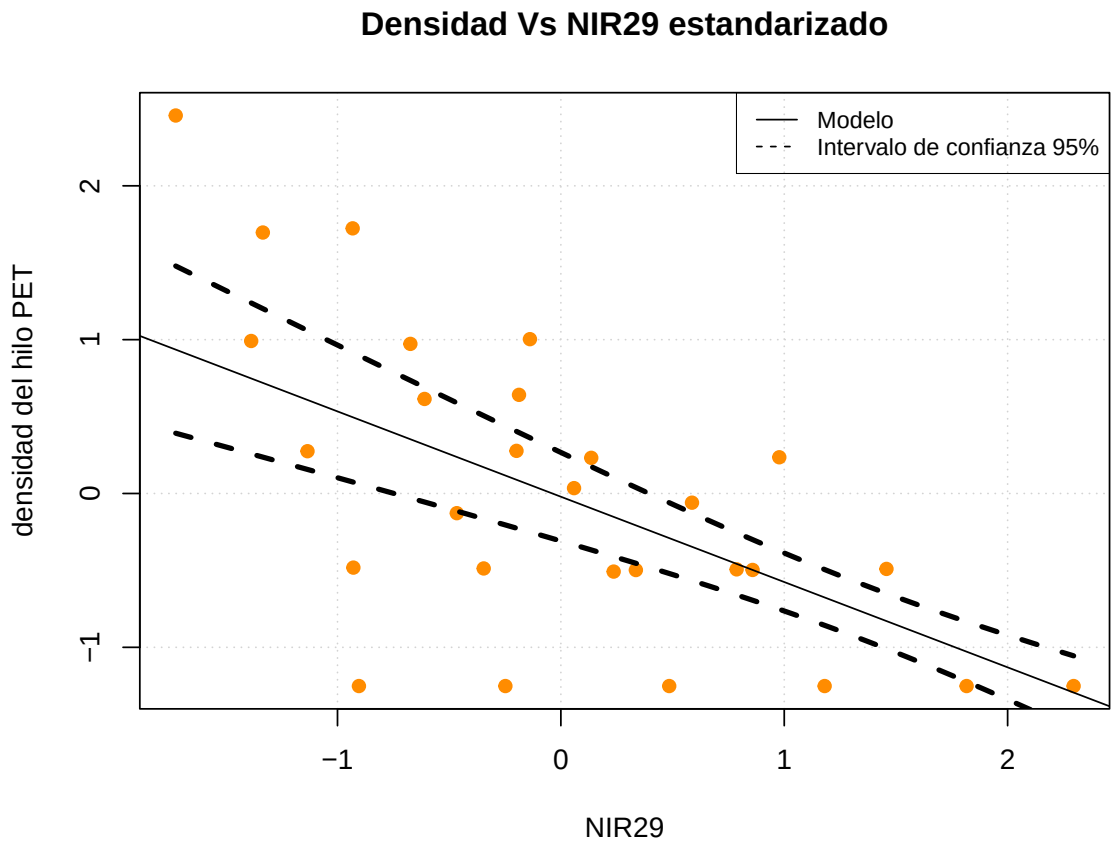
Se realizará un gráfico en el que se evaluaran 100 puntos teniendo en cuenta el no extrapolar y a estos se les obtendrá los intervalos de confianza para el valor esperado.

```r
x.nuevo = data.frame(NIR29=seq(min(Z[,29]),max(Z[,29]),length.out=100))

#Predicción del intervalo de confianza
pred.media = predict(model.ponderadosz,x.nuevo,interval = "confidence")
```

```r
plot(Z$NIR29,Z$density,pch=19,col="#FF8C00",panel.first=grid(),xlab="NIR29",
ylab="densidad del hilo PET",main='Densidad Vs NIR29 estandarizado')
#Gráficas de las lineas
lines(x.nuevo[,1],pred.media[,2],lty=2,lwd=3)
lines(x.nuevo[,1],pred.media[,3],lty=2,lwd=3)

abline(model.ponderadosz)
legend(x = "topright",legend=c("Modelo","Intervalo de confianza 95%"),
       lty = c(1, 2,3),pt.cex=1.5,
       box.lwd=0.6,text.font =15,cex=0.8) #Caja de enunciados
```



En el gráfico se observa que el modelo estimado estandarizado es el mismo que el normal, solo que en unidades de medidas diferentes, por lo que se siguen presentando buenas predicciones a la densidad media de los hilos de PET. Se realizará un ejemplo de un valor para entender la interpretación en términos contextuales

```r
ejemploz <- data.frame(NIR29=1)
pred.media = predict(model.ponderadosz,ejemploz,interval = "confidence")
pred.media
```

```
##          fit        lwr        upr
## 1 -0.5755519 -0.7639564 -0.3871474
```

Esto indica que con un nivel de confianza del 95 %, cuando el NIR29 presente 1 desviación estándar, la densidad media de los hilos de PET está entre -0.76 y -0.38 desviaciones estándar.

13. Modelo polinómico

Se ajustará el siguiente modelo polinómico y se evaluará si mejora estimaciones

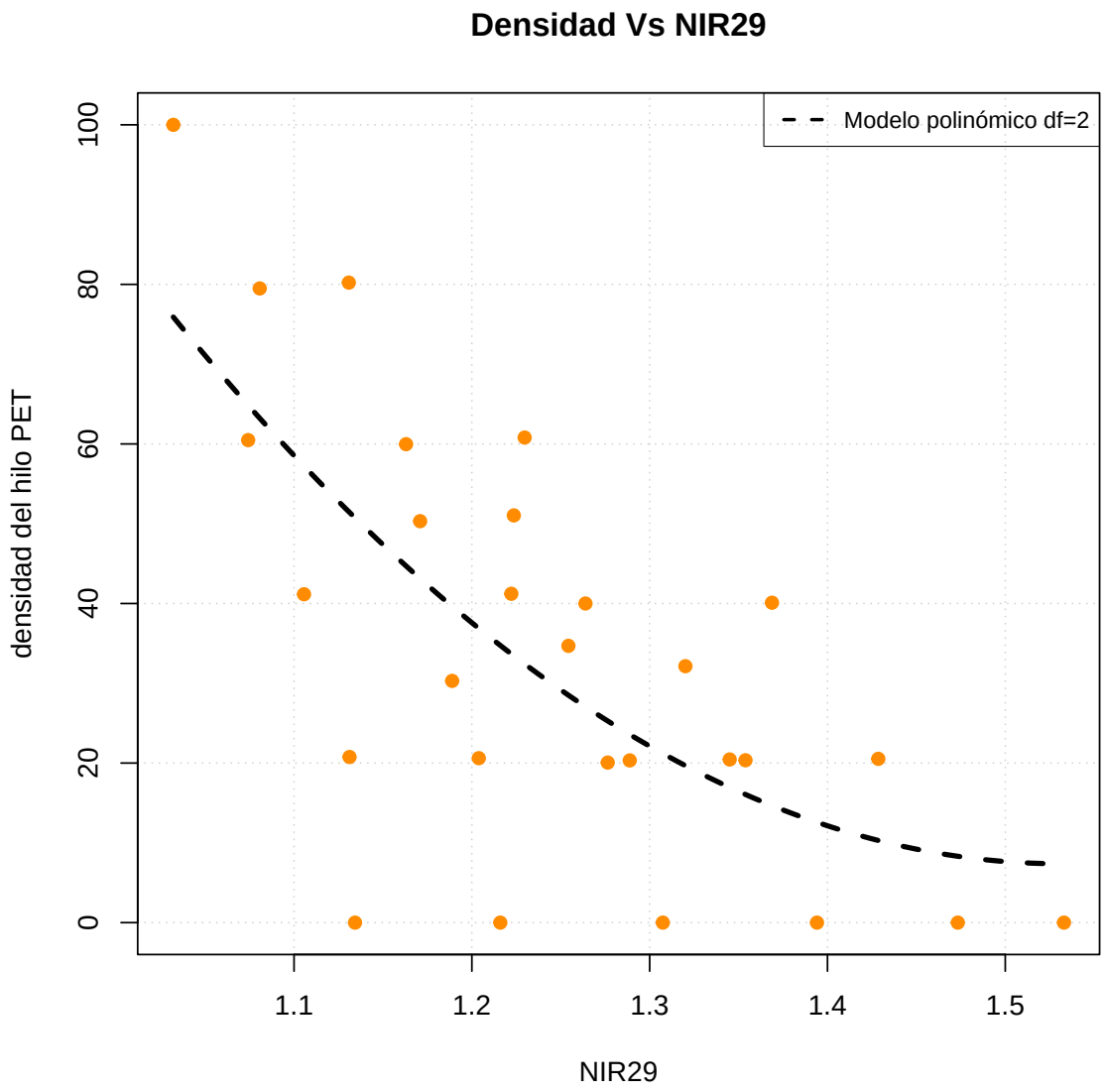
$$y_i = \beta_0 + NIR29_i\beta_1 + NIR29_i^2\beta_2$$

```r
model.<- lm(density~NIR29+I(NIR29^2),data=X)
summary(model.)
```

```
##
## Call:
## lm(formula = density ~ NIR29 + I(NIR29^2), data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.739 -10.152   3.896  13.514  28.714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    651.4      348.2   1.871   0.0731 .
## NIR29         -840.9      550.4  -1.528   0.1391
## I(NIR29^2)     274.5      216.1   1.270   0.2158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.39 on 25 degrees of freedom
## Multiple R-squared:  0.4707,Adjusted R-squared:  0.4284
## F-statistic: 11.12 on 2 and 25 DF,  p-value: 0.0003518
```

Se observa que la estimación de los coeficientes cambiaron, la varianza de los estimadores se incrementaron, esto se debe a la dependencia que tienen las variables, ya que básicamente una es una función de la otra. Si se realiza una comparación con el modelo por lineal simple, el polinómico presenta un R^2 un poco mayor, sin embargo, habría la falta de interpretabilidad en cuanto a los coeficientes, además, los intervalos de confianza serán más amplios debido al aumento de la varianza de los coeficientes estimados. Este aumento de varianza se puede corregir al escalonar las variables.

```r
plot(X[,29],X[,31],pch=19,col="#FF8C00",panel.first=grid(),xlab="NIR29",
ylab="densidad del hilo PET",main='Densidad Vs NIR29')
#Gráficas de las lineas
lines(spline(X[,29],fitted.values(model.)),lty=2,lwd=3)
legend(x = "topright",legend=c("Modelo polinómico df=2"),
       lty = c(2),lwd=2,pt.cex=1.5,
       box.lwd=0.6,text.font =15,cex=0.8) #Caja de enunciados
```

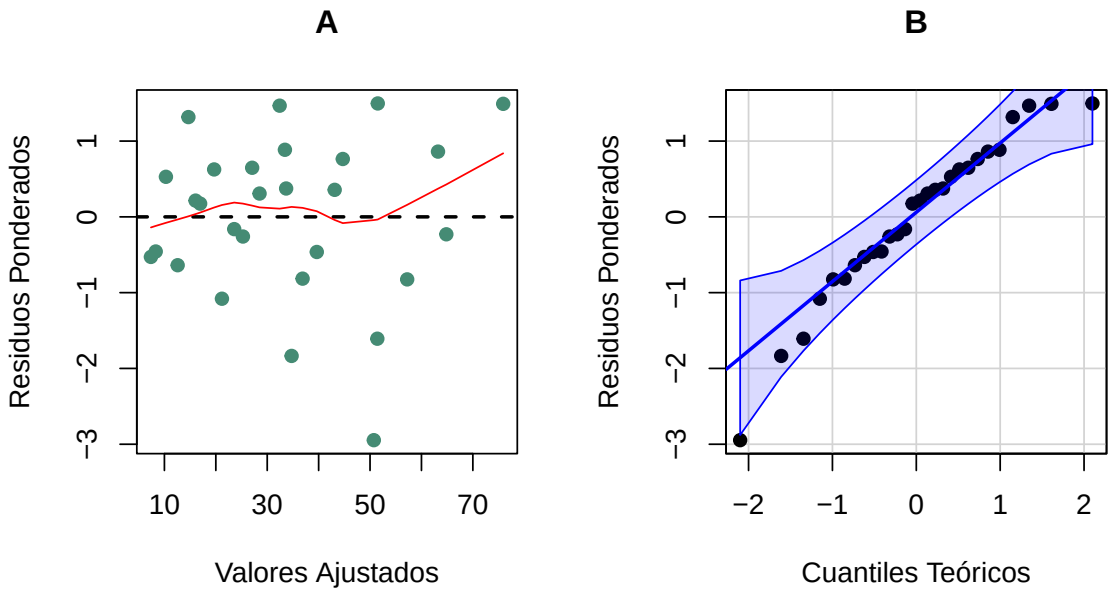


Para realizar la comparación con modelo por MCP, debemos asegurarnos que el modelo polinómico no presente problema en los supuestos.

13.1. Validación Supuestos

```
studenti.z<- studres(model.)
ajustados.z<- fitted.values(model.)
```

```
par(mfrow=c(1,2))
plot(ajustados.z,studenti.z, ylab='Residuos Ponderados',
     xlab='Valores Ajustados',pch=19,col="aquamarine4",
     main="A")
abline(h=0,lty=2,lwd=2)
lines(lowess(studenti.z~ajustados.z), col = "red1")
qqPlot(studenti.z,main="B", ylab="Residuos Ponderados",
xlab="Cuantiles Teóricos",id=F,pch=19)
```




```
shapiro.test(studenti.z)

##
##  Shapiro-Wilk normality test
##
## data:  studenti.z
## W = 0.95226, p-value = 0.2257

bptest(model.,~NIR29+I(NIR29^2),data=Z)

##
##  studentized Breusch-Pagan test
##
## data:  model.
## BP = 3.104, df = 2, p-value = 0.2118
```

Dado los respectivos gráficos y pruebas asociadas a la evaluación de supuestos, se observa el cumplimiento de todos ellos. Por ende se procede a comparar con el modelo por MCP, el MCP presenta un R^2 mucho mejor a comparación con el modelo polinómico, además el modelo por MCP posee la ventaja de realizar conclusiones más sencillas en cuanto a los β estimados. Como conclusión, el modelo propuesto por MCP es el más adecuado al asociar la variable predictora NIR29 con la variable de respuesta densidad.