Diagnóstico en Regresión Lineal

15 de noviembre de 2018

Recordemos que en el modelo lineal asumimos que

$$Y_i = x_i^T \beta_{\mathbf{0}} + \varepsilon_i$$

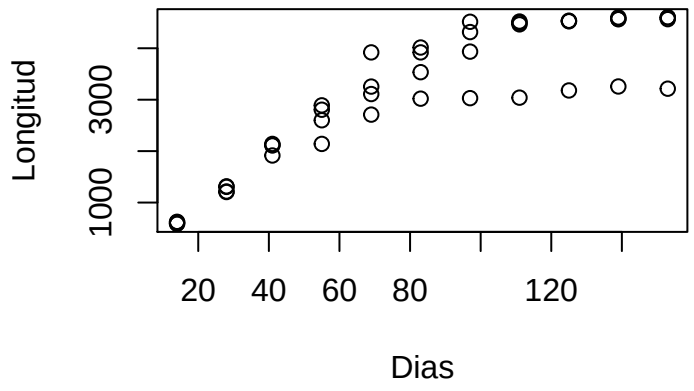
donde $x_i \in \mathbb{R}^p$, $\beta_{\mathbf{0}} \in \mathbb{R}^p$ y $\varepsilon_i \sim N(0, \sigma^2)$ y son independientes. Los supuestos, entonces, se pueden reescribir de la siguiente manera:

1. Las variables Y_i son independientes.
2. Existe un $\beta_{\mathbf{0}}$ tal que $\mathbb{E}(Y_i) = x_i^T \beta_{\mathbf{0}}$.
3. $V(Y_i) = \sigma^2$ para todo i (homocedasticidad).
4. Las variables Y_i tienen distribución normal.

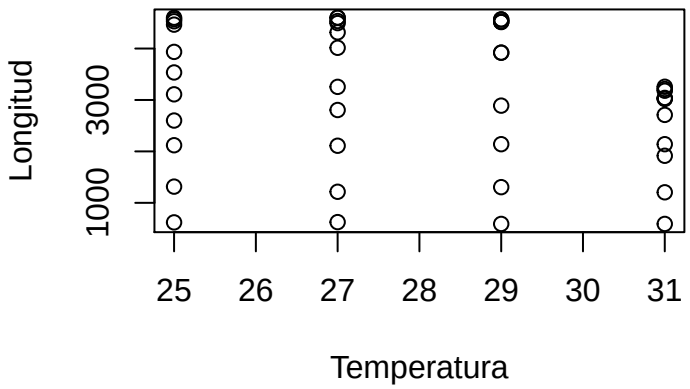
En esta clase vamos a dar algunas herramientas básicas para validar estos los supuestos 2, 3 y 4. El supuesto 1 en general se puede asumir según el contexto del problema.

Ejemplo 1: en el archivo `fish.txt` se encuentran los datos de 44 peces. Se tiene información de la edad en días de cada uno, la temperatura del agua en donde vive y su longitud. Se quiere explicar la longitud en función de las otras dos variables.

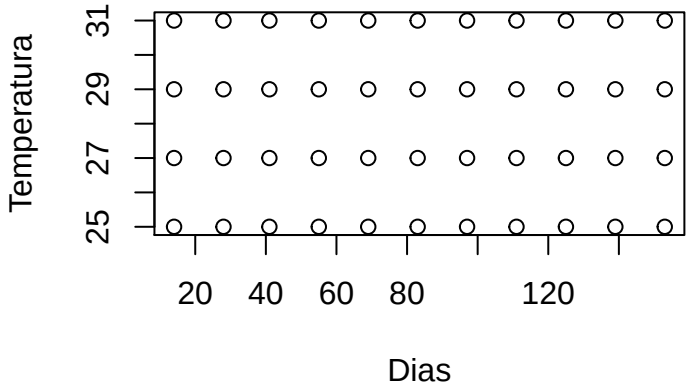
```
# Cargamos los datos
peces = read.table("fish.txt")[,2:4]
colnames(peces) = c("Dias","Temperatura","Longitud")
attach(peces)

#Miramos un poco los plots 1 contra 1 de las 3 variables consideradas:
plot(Dias, Longitud)
```




```
plot(Temperatura, Longitud)
```




```
plot(Dias, Temperatura)
```



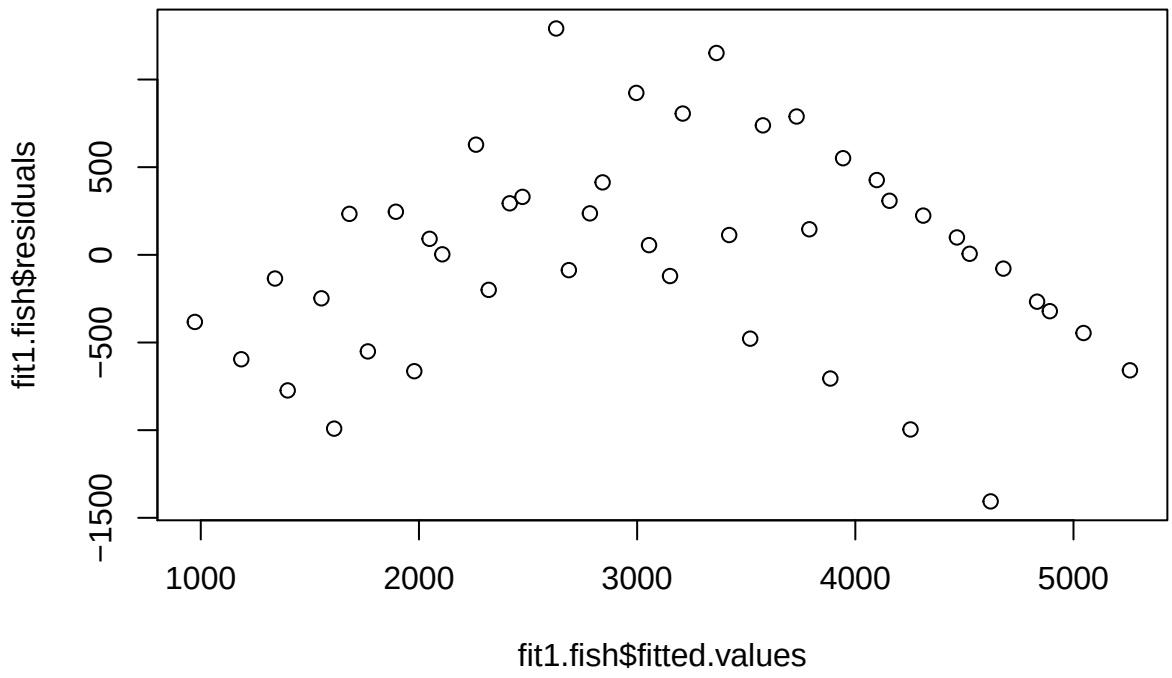

```
# Ajustamos el modelo
fit1.fish = lm(Longitud ~ Dias + Temperatura)
```

```r
summary(fit1.fish)
```

```
##
## Call:
## lm(formula = Longitud ~ Dias + Temperatura)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1406.27  -398.59    30.73   313.94  1291.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3904.266   1149.044   3.398  0.00152 **
## Dias          26.241      2.055  12.769 7.11e-16 ***
## Temperatura -106.414     40.452  -2.631  0.01195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 600 on 41 degrees of freedom
## Multiple R-squared:  0.8056, Adjusted R-squared:  0.7962
## F-statistic: 84.98 on 2 and 41 DF,  p-value: 2.607e-15
```

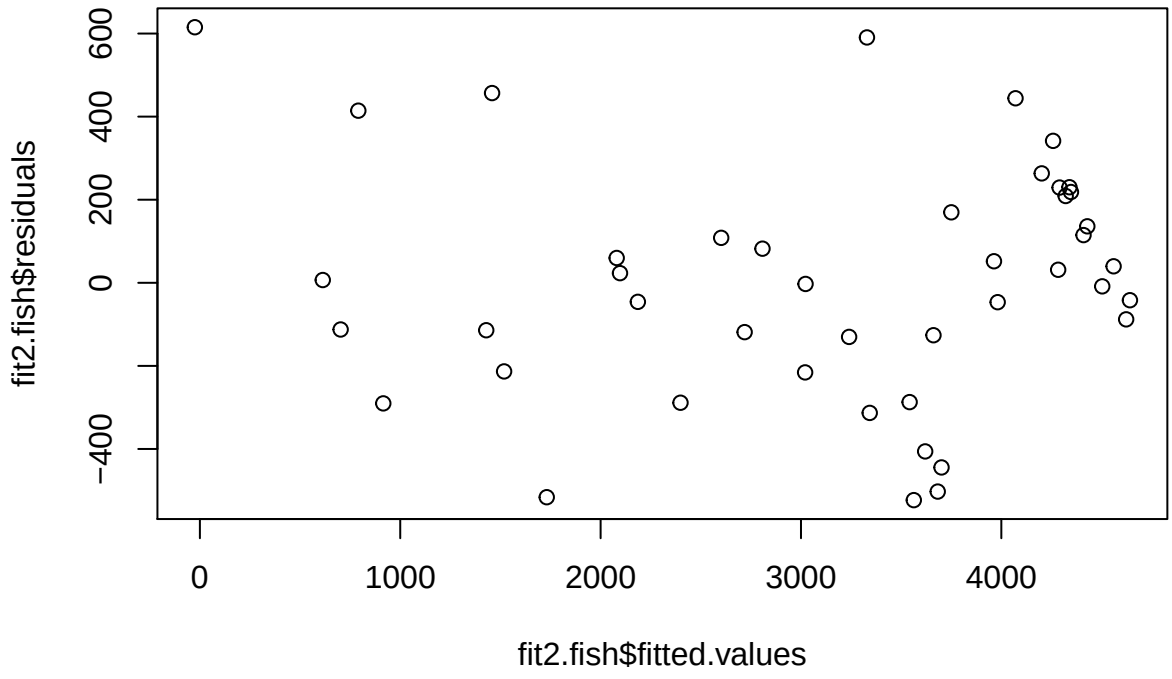
Una forma de ver si los supuestos 2 y 3 se cumplen es haciendo el gráfico de $Y-$es predichos (fitted values) contra los residuos (o sea, $r_i = Y_i - \widehat{Y}_i$). Veamos como da en este caso.

```r
plot(fit1.fish$fitted.values, fit1.fish$residuals)
```



Claramente parece haber una estructura cuadrática. Esto nos da un indicio de que el supuesto 2 no se cumple, porque es la esperanza de los residuos parece ser distinta de cero para algunos valores de $\widehat{Y}_i$. Al ver una esctuctura de esta forma, suele ser razonable agregar las variables explicativas al cuadrado en el modelo:

```
fit2.fish = lm(Longitud ~ Dias + Temperatura + I(Dias**2) + I(Temperatura**2))
plot(fit2.fish$fitted.values, fit2.fish$residuals)
```



Ahora la estructura cuadrática parece no estar más. En general, para que los supuestos 2 y 3 se cumplan, uno busca que el gráfico de $Y-$es predichos vs. residuos no tenga ninguna estructura.

Por último, para ver si el supuesto 4 se cumple, aplicamos el test de normalidad de Shapiro-Wilk en los residuos. En este test, la hipótesis nula es que los residuos son normales y la alternativa es que no lo son. Notar que en este caso, en el fondo queremos NO rechazar la hipótesis nula (al contrario de lo que veníamos haciendo en tests de hipótesis). Nos vamos a conformar si el $p-$valor de este test es mayor a 0.20 (si todo esto le hace ruido, va por buen camino. Es un buen ejercicio pensar cuándo tiene sentido trabajar con un test de hipótesis de esta manera).

```
shapiro.test(fit2.fish$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit2.fish$residuals
## W = 0.98092, p-value = 0.6702
```

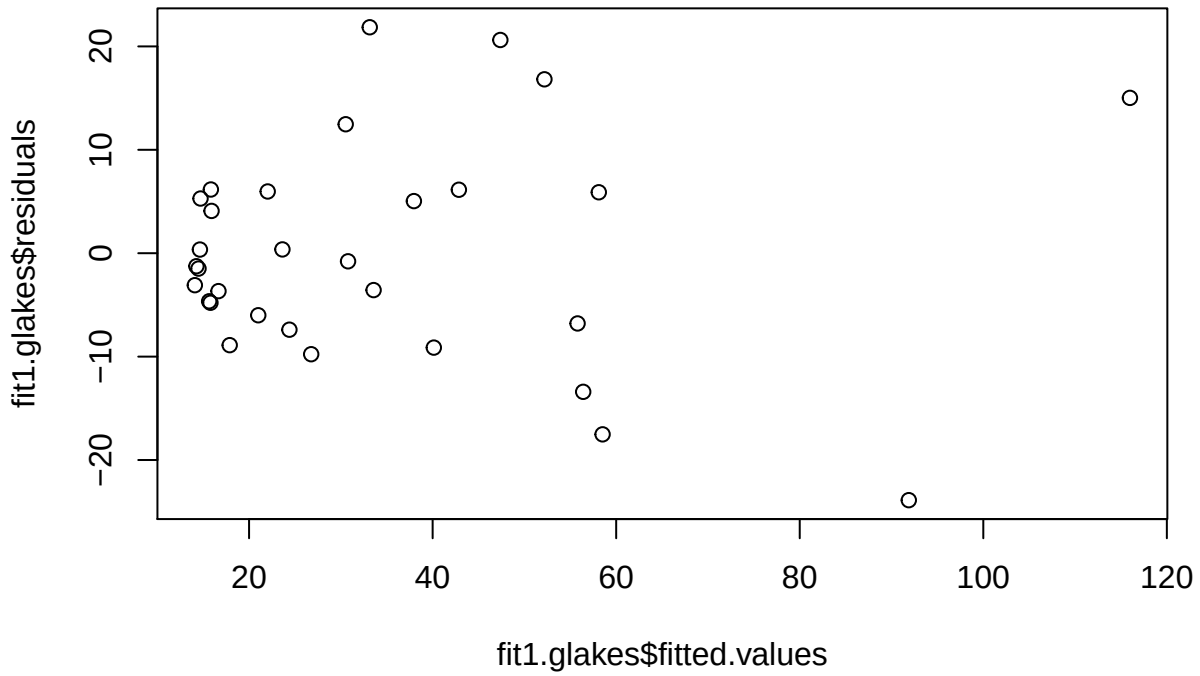
Es mucho mayor que 0.2, entonces no rechazamos la hipótesis de normalidad.

veamos ahora un ejemplo en donde sospechamos del supuesto 3 (homocedasticidad).

Ejemplo 2: en el archivo `glakes.csv` se encuentra la información del tiempo de descarga (variable `Time`) y

el peso (variable `Tonnage`) de 31 containers en la ciudad de Grand Lakes, en Canadá. Veamos qué pasa si hacemos el mismo gráfico que hicimos en el ejemplo anterior.

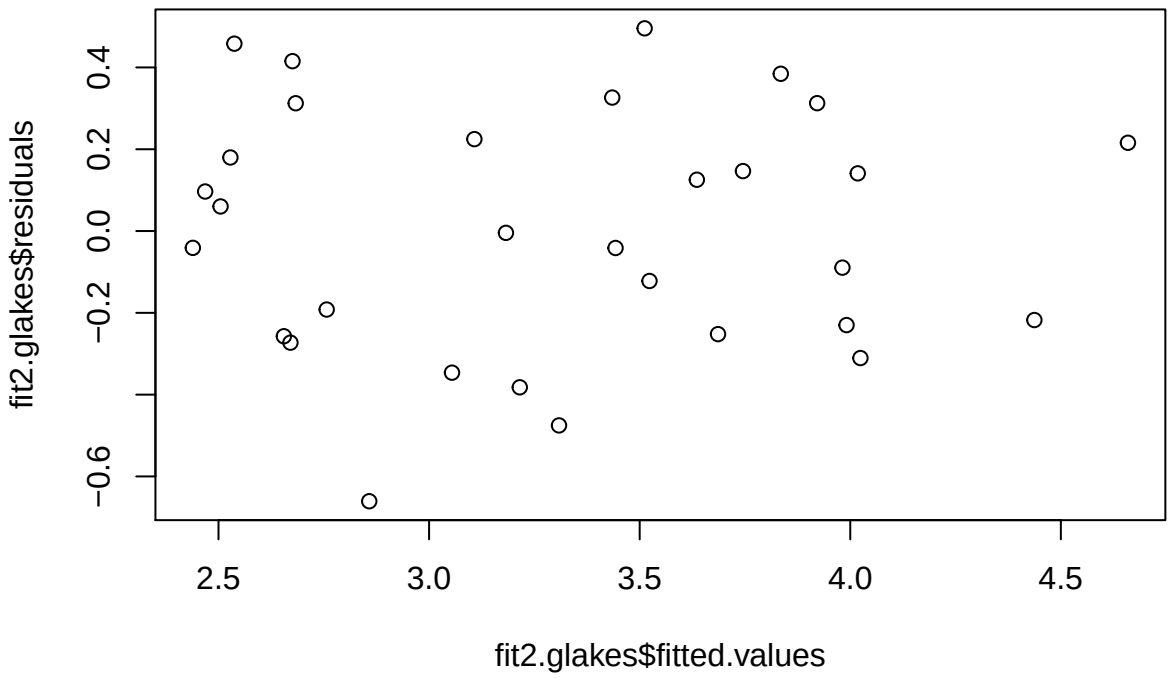
```
glakes = read.table("glakes.csv", sep = ",", header = T)[,2:3]
attach(glakes)
fit1.glakes = lm(Time ~ Tonnage)
plot(fit1.glakes$fitted.values, fit1.glakes$residuals)
```



En este caso podemos ver que si bien los residuos parecen estar centrados en cero, su dispersión aumenta a medida que el $\widehat{Y}_i$ es mayor. Esto probablemente se deba que el supuesto 3 (homocedasticidad) no se cumple. Existen alternativas a mínimos cuadrados para modelos heterocedásticos (por ejemplo mínimos cuadrados pesados) pero acá vamos a intentar tomar transformaciones de los datos para llegar a un modelo homocedástico.

Luego de buscar transformaciones adecuadas, veamos que pasa cuando tomamos el logaritmo de la variable respuesta y la raíz cuarta de la variable explicativa:

```
fit2.glakes = lm(log(Time) ~ I(Tonnage^{0.25}))
plot(fit2.glakes$fitted.values, fit2.glakes$residuals)
```



Ahora mágicamente llegamos a la homocedasticidad. También con este gráfico podemos ver que el supuesto 2 parece cumplirse. Veamos que pasa con el supuesto 4.

```
shapiro.test(fit2.glakes$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit2.glakes$residuals
## W = 0.97356, p-value = 0.6216
```

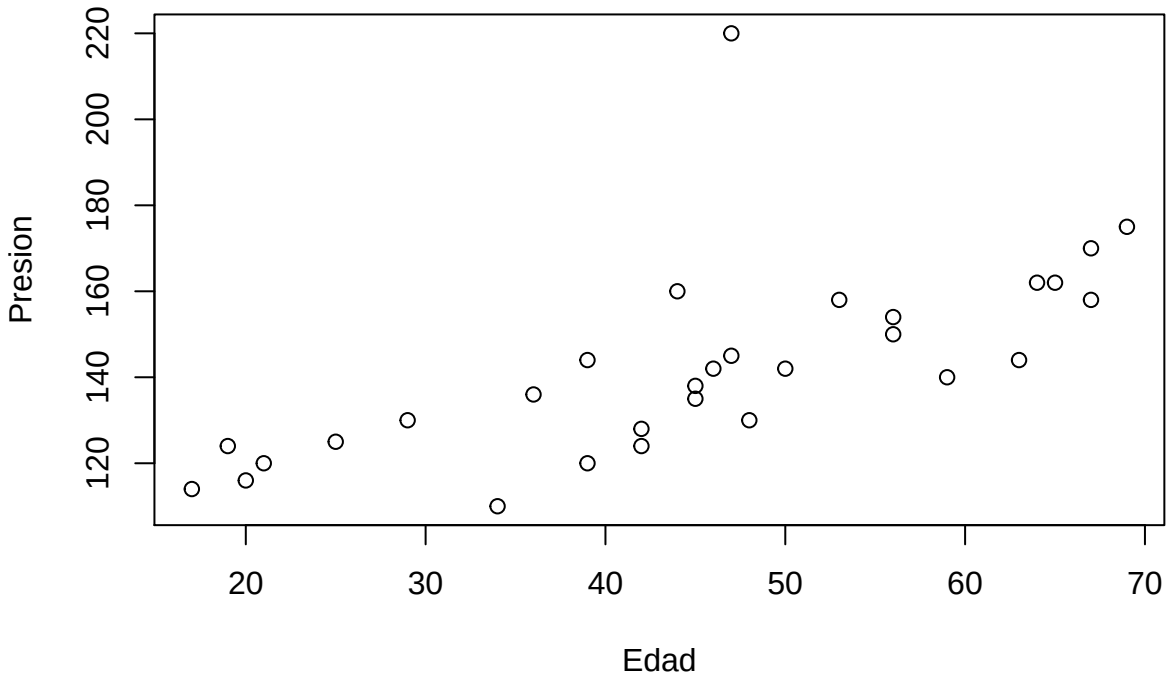
Por lo tanto no rechazamos normalidad.

Ejemplo 3: en el archivo `presion_sistolica.txt` se encuentran la edad y presión sistólica de 30 pacientes.

```
presion = read.table("presion_sistolica.txt")[,3:4]
colnames(presion) = c("Edad", "Presion")
attach(presion)
```

Hagamos primero un plot de Edad vs Presión:

```
plot(Edad, Presion)
```



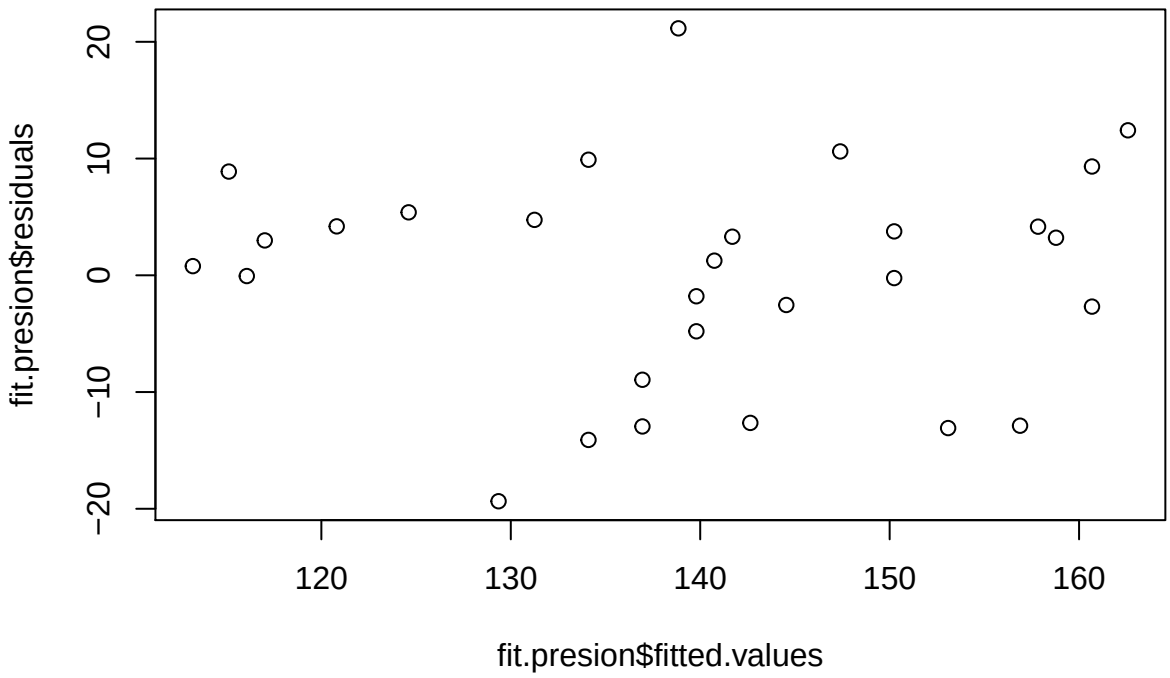
En este caso parece haber una observación atípica. Veamos cuál es

```
which(Presion == max(Presion))
```

```
## [1] 2
```

Vamos a hacer el análisis sin la observación 2. Dejo a ustedes como ejercicio mirar qué pasa si dejamos esta observación. Tambien es interesante pensar qué pasaría si tuvieramos una observación con Edad 20 y Presión 220.

```
fit.presion = lm(Presion~Edad, data = presion[-2,])
plot(fit.presion$fitted.values, fit.presion$residuals)
```




```
shapiro.test(fit.presion$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit.presion$residuals
## W = 0.96258, p-value = 0.38
```

En este caso no sospechamos de ningún supuesto.