COMPROBACIÓN DE ADECUACIÓN DEL MODELO

Asencios Menacho Soledad

Análisis de Residuales

El análisis de los residuales es una forma eficaz de descubrir diversos tipos de inadecuación del modelo.

$$CM_{Res} = \frac{SC_{Res}}{n-p}$$

Métodos para escalar residuales

- Residuales Estandarizados
- Residuales Estudentizados
- Residuales PRESS

Gráfico de residuales y Pruebas Estadísticas

1) Normalidad

Gráfica de probabilidad normal

H0: Los errores se distribuyen normalmente
H1: Los errores no se distribuyen normalmente
En la literatura en estadística existen muchas pruebas de normalidad, una de ellas es la Prueba de Shapiro Wilk
Otras pruebas de normalidad son: Anderson-Darling, Jarque-Bera, Shapiro- Francia, Cramer von Mises, Pearson, entre otras. El paquete nortest del R ofrece varias pruebas de normalidad.
shapiro.test()

Interpretación

- **W** cercano a1 y p-valor mayor que 0.05: Estos valores indican que es probable que los datos sigan una distribución normal.
- **W** lejos de 1 y p-valor menor que 0.05: Estos valores sugieren que los datos no se distribuyen normalmente.

2) Homogeneidad de varianzas

Homocedasticidad

Gráfica de residuales en función de los valores ajustados

Prueba de Breusch-Pagan
H0: Los errores son homocedasticos
H1: Los errores no son homocedasticos
Consiste en ajustar un modelo de regresión lineal con variable respuesta dada por residuales del modelo original al cuadrado e_i^2 y como covariables las variables del modelo original. Otras pruebas de homogeneidad son: White, Prueba de Score (ncvtest del paquete car), Goldfeld-Quandt (gqtest del paquete lmtest) , Harrison-McCabe (hmctest del paquete lmtest), entre otras.

```
library(lmtest)
bptest()
```

Interpretación

- p-valor **bajo** (menor a 0.05): Se rechaza la hipótesis nula, lo que indica que hay evidencia de heterocedasticidad. La varianza de los errores no es constante y se deben tomar medidas correctivas.
- p-valor **alto** (mayor a 0.05): No se rechaza la hipótesis nula, lo que sugiere que la varianza de los errores es constante y se cumple el supuesto de homocedasticidad.

BP: Este es el estadístico de prueba de Breusch-Pagan. Un valor alto indica una mayor probabilidad de heterocedasticidad.
df: Son los grados de libertad, que en este caso dependen de la especificación del modelo.
p-value: Este es el valor p.

Prueba de Durbin-Watson

H0: Los errores son independientes
H1: Los errores no son independientes
Permite ver si los valores presentan algún tipo de dependencia en cuanto al orden de obtención Es un estadístico que varía entre 0 y 4.
Otra prueba de independencia (autocorrelación) es: Breusch-Godfrey que se encuentra en R en la función gbtest del paquete lmtest.

```
library(lmtest)
dwtest(modelo2,alternative="t")
```

Interpretación

La prueba de Durbin-Watson proporciona un estadístico, denotado como d, que puede tomar valores entre 0 y 4. La interpretación de este valor es crucial para determinar la presencia y el tipo de autocorrelación:

- d ≈ 2: Indica que no hay autocorrelación significativa en los residuos.

- d < 2: Sugiere autocorrelación positiva, es decir, los residuos tienden a ser similares a los residuos de las observaciones anteriores.

- d > 2: Indica autocorrelación negativa, lo que significa que los residuos tienden a ser opuestos a los residuos de las observaciones anteriores.

- **La estadística PRESS**

Se considera que PRESS es una medida de lo bien que funciona un modelo de regresión para predecir nuevos datos. Lo deseable es tener un modelo con valor pequeño de PRESS

```
rpress <- rstandard(modelo2, type = "pred")
scp <- sum(rpress^2)
sct <- sum(anova(modelo2)$'Sum Sq')
(R2_pred <- 1 - scp/sct)
```

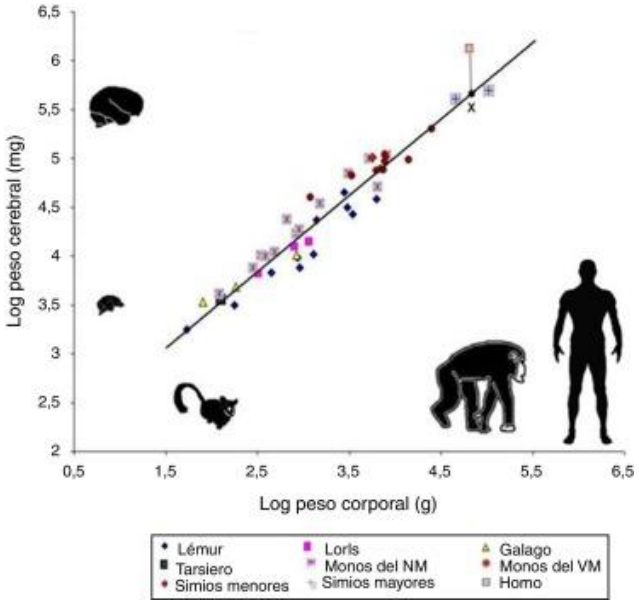
Interpretación

- PRESS bajo: Indica que el modelo generaliza bien a nuevos datos y es menos propenso a sobreajuste.

- PRESS alto: Sugiere que el modelo puede estar sobreajustado a los datos de entrenamiento y puede no ser muy preciso al predecir nuevos datos.

PRACTICA DIRIGIDA 3

Caso 1: Relación peso corporal y peso del cerebro



El tamaño neto del cerebro esta positivamente relacionado con el tamaño de un animal. Sin embargo, **esta relación no es lineal;** los animales pequeños como los ratones, tienen una relación cerebro/cuerpo similar a la de los humanos, mientras los elefantes tienen esta relación cerebro/cuerpo mucho más reducida, pero se trata de animales con evidente inteligencia. El dataset **mammalsleep** ubicado dentro del paquete **mice** tiene datos de diferentes variables de 62 especies de mamíferos. Se desea obtener un modelo de regresión que permita predecir el peso del cerebro brw (en gramos) usando el peso del cuerpo bw (en kilogramos).

a) Usar un diagrama de dispersión para identificar el tipo de relación existente entre las variables.

```
library(mice)
```

```
##
## Adjuntando el paquete: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
data("mammalsleep")
head(mammalsleep)
```

```
##                      species       bw    brw sws  ps   ts  mls  gt pi
sei odi
## 1          African elephant 6654.000 5712.0  NA  NA  3.3 38.6 645  3
5   3
## 2 African giant pouched rat    1.000    6.6 6.3 2.0  8.3  4.5  42  3
1   3
## 3                Arctic Fox    3.385   44.5  NA  NA 12.5 14.0  60  1
1   1
## 4    Arctic ground squirrel    0.920    5.7  NA  NA 16.5   NA  25  5
2   3
## 5            Asian elephant 2547.000 4603.0 2.1 1.8  3.9 69.0 624  3
5   4
## 6                    Baboon   10.550  179.5 9.1 0.7  9.8 27.0 180  4
4   4
```

```
windows()
```

```
plot(brw~bw,xlab="Peso del cuerpo", ylab="Peso del cerebro",
     col="blue",main="Gráfico de dispersión",
     data=mammalsleep)
```

Interpretación:
Relación positiva: Existe una tendencia general a que a mayor peso corporal mayor

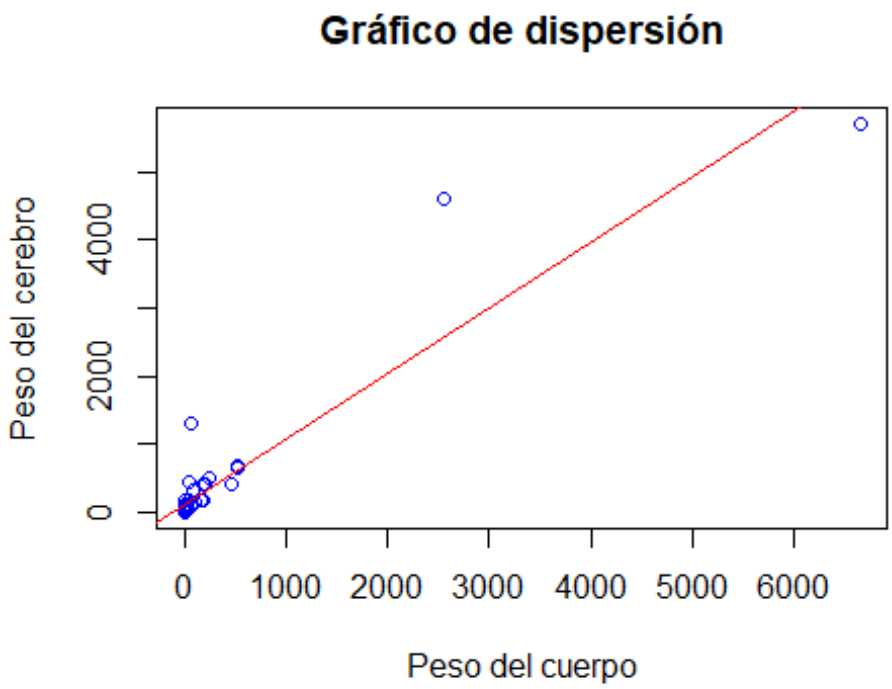
sera el peso del cerebro. Esto sugiere que hay una correlación positiva entre ambas variables.

No linealidad: El aumento en el peso del cerebro no es proporcional al aumento en el peso corporal.

Valores atipicos: Hay algunos puntos que se alejan considerablemente de la tendencia general.

b) Obtener el modelo de regresión lineal y graficar la recta obtenida sobre el diagrama de dispersión.

```
mls1<-lm(brw~bw,data=mammalsleep)
mls1

##
## Call:
## lm(formula = brw ~ bw, data = mammalsleep)
##
## Coefficients:
## (Intercept)           bw
##     91.0044       0.9665

plot(brw~bw,xlab="Peso del cuerpo", ylab="Peso del cerebro",
     col="blue",main="Gráfico de dispersión",
     data=mammalsleep)
abline(mls1,col="red") # linea pendiente en la grafica
```

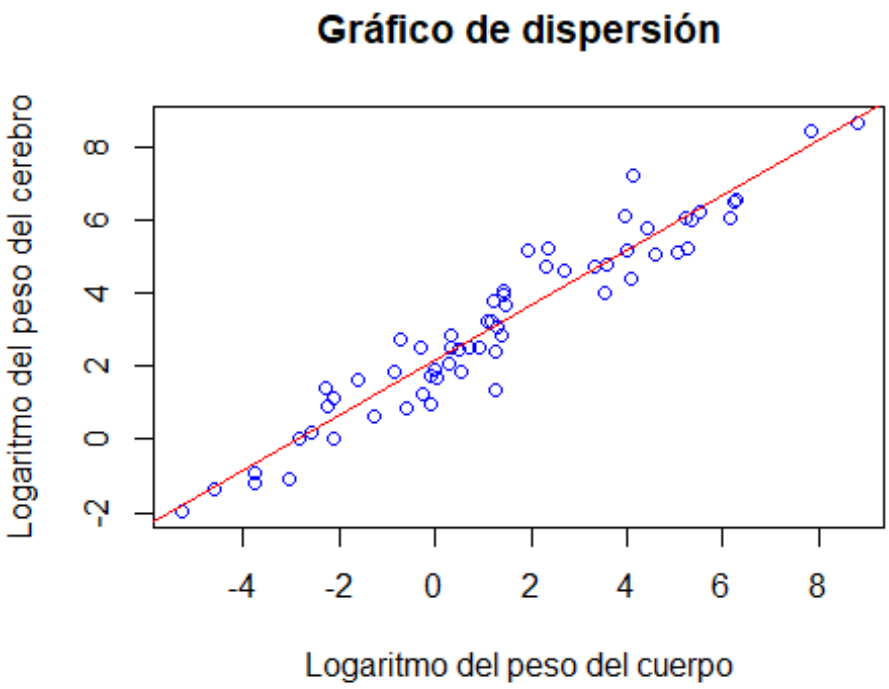


```r
summary(mls1)
```

```
##
## Call:
## lm(formula = brw ~ bw, data = mammalsleep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -810.07  -88.52  -79.64  -13.02 2050.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 91.00440   43.55258    2.09   0.0409 *
## bw           0.96650    0.04766   20.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 334.7 on 60 degrees of freedom
## Multiple R-squared:  0.8727, Adjusted R-squared:  0.8705
## F-statistic: 411.2 on 1 and 60 DF,  p-value: < 2.2e-16
```

c) Aplicar la transformación logaritmo a la variable respuesta y la variable predictora. Obtener el diagrama de dispersión y el modelo de regresión para las variables transformadas.

```r
mp1<-lm(log(brw)~log(bw),data=mammalsleep)

plot(log(brw)~log(bw),xlab="Logaritmo del peso del cuerpo",
ylab="Logaritmo del peso del cerebro",
     col="blue",main="Gráfico de dispersión",
     data=mammalsleep)
abline(mp1,col="red")
```




```
summary(mp1)
```

```
##
## Call:
## lm(formula = log(brw) ~ log(bw), data = mammalsleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71550 -0.49228 -0.06162  0.43597  1.94829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.13479    0.09604   22.23   <2e-16 ***
## log(bw)      0.75169    0.02846   26.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6943 on 60 degrees of freedom
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9195
## F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

d) Repetir la pregunta anterior usando las transformaciones raíz cuadrada, raíz cúbica e inversa sobre la variable predictora.

Raiz cuadrada

```
library(car)
```

```
## Cargando paquete requerido: carData
```

```
m1<-lm(brw~bcPower(bw,lambda=1/2),data=mammalsleep)
summary(m1)
```

```
##
## Call:
## lm(formula = brw ~ bcPower(bw, lambda = 1/2), data = mammalsleep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -905.85   -8.91   72.48  118.83 1297.30
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -77.595     45.492  -1.706   0.0932 .
## bcPower(bw, lambda = 1/2)   34.197      1.662  20.573   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 330.5 on 60 degrees of freedom
## Multiple R-squared:  0.8758, Adjusted R-squared:  0.8738
## F-statistic: 423.2 on 1 and 60 DF,  p-value: < 2.2e-16
```

```
summary(m1)$r.sq*100
```

```
## [1] 87.58353
```

Raiz cúbica

```
m2<-lm(brw~bcPower(bw,lambda=1/3),data=mammalsleep)
summary(m2)
```

```
##
## Call:
## lm(formula = brw ~ bcPower(bw, lambda = 1/3), data = mammalsleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1105.43  -125.97    79.76   208.24  1625.58
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -126.044     67.811  -1.859    0.068 .
## bcPower(bw, lambda = 1/3)   81.735      6.194  13.196   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 474.8 on 60 degrees of freedom
## Multiple R-squared:  0.7437, Adjusted R-squared:  0.7395
## F-statistic: 174.1 on 1 and 60 DF,  p-value: < 2.2e-16
```

```r
summary(m2)$r.sq*100
```

```
## [1] 74.37242
```

Inversa

```r
m3<-lm(brw~bcPower(bw,lambda=-1),data=mammalsleep)
summary(m3)
```

```
##
## Call:
## lm(formula = brw ~ bcPower(bw, lambda = -1), data = mammalsleep)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -301.0 -290.8 -259.4 -127.4 5406.3
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 302.927    122.278   2.477   0.0161 *
## bcPower(bw, lambda = -1)      2.782      4.139   0.672   0.5040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 934.5 on 60 degrees of freedom
## Multiple R-squared:  0.007476,   Adjusted R-squared:  -0.009067
## F-statistic: 0.4519 on 1 and 60 DF,  p-value: 0.504
```

```r
summary(m3)$r.sq*100
```

```
## [1] 0.74755
```

Transformation	$r^2 \uparrow$	$\sqrt{CME} \downarrow$
Sin transformación (Modelo lineal)	87.27%	334.7
Con trans. logaritmo a la variable respuesta y predictora	92.08%	0.6943
Raiz cuadrada	87.58%	330.5
Raiz cúbica	74.37%	474.8
Inversa	74.75%	0.007476

La transformación logarítmica mejora significativamente el ajuste del modelo:

El modelo con la transformación logarítmica a la variable respuesta y predictora presenta el valor más alto de R^2 (92.08%) y el valor más bajo de √CME (0.6943). Esto sugiere que la transformación logarítmica linealiza la relación entre las variables, mejorando la capacidad del modelo para explicar la variabilidad en los datos.

Las otras transformaciones no ofrecen una mejora sustancial:

Las otras transformaciones (raíz cuadrada, raíz cúbica e inversa) no logran superar el rendimiento del modelo con la transformación logarítmica. Esto indica que la transformación logarítmica es la más adecuada para estos datos en particular.

El modelo lineal sin transformación presenta un buen ajuste:

Incluso sin ninguna transformación, el modelo lineal base muestra un R^2 bastante alto (87.27%), lo que sugiere que la relación entre las variables es relativamente lineal.

Caso 2: Árboles de cedro

El dataset ufc ubicado dentro del paquete **alr4** tiene datos sobre la altura y el diámetro de 372 árboles de cedro. Se desea obtener un modelo de regresión que permita predecir el diámetro del árbol Dbh (en mm) en función de su altura Height (en dm).

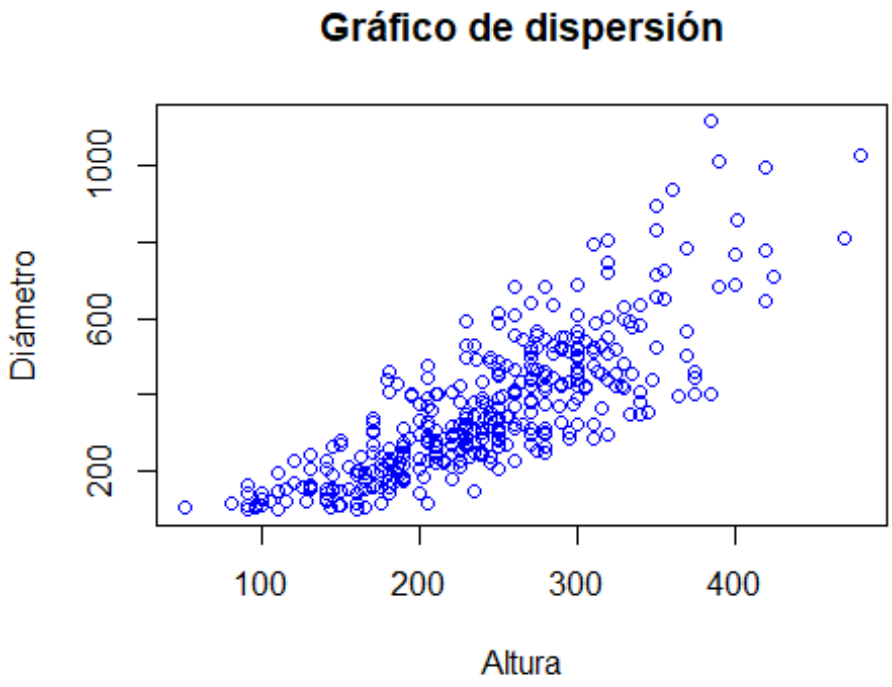
```
library(alr4)

## Cargando paquete requerido: effects

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```r
data(ufc)
head(ufc)
```

```
##   Plot Tree Species Dbh Height
## 1    2    1      DF 390    205
## 2    2    2      WL 480    330
## 3    3    2      GF 520    300
## 4    3    5      WC 360    207
## 5    3    8      WC 380    225
## 6    4    1      WC 460    180
```

a) Usar un diagrama de dispersión para identificar el tipo de relación existente entre las variables.

```r
plot(Dbh~Height,xlab="Altura", ylab="Diámetro",
     col="blue",main="Gráfico de dispersión",
     data=ufc)
```



b) Aplicar las transformaciones inversa y raíz cuadrada sobre la variable predictora.

c) Elegir la mejor transformación usando el coeficiente de determinación y la suma de cuadrados del residual.

Transformación 1/x

Inversa

```r
mi<-lm(Dbh~bcPower(Height,lambda=-1),data=ufc)
summary(mi)
```

```
##
## Call:
## lm(formula = Dbh ~ bcPower(Height, lambda = -1), data = ufc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -242.61  -97.79  -26.15   69.84  698.94
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -61751       3752  -16.46   <2e-16 ***
## bcPower(Height, lambda = -1)    62405       3769   16.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 137.8 on 370 degrees of freedom
## Multiple R-squared:  0.4256, Adjusted R-squared:  0.424
## F-statistic: 274.1 on 1 and 370 DF,  p-value: < 2.2e-16
```

```
c(summary(mi)$r.sq*100,summary(mi)$sigma)
```

```
## [1]  42.5575 137.8117
```

```
summary(mi)$sigma
```

```
## [1] 137.8117
```

Raiz cuadrada

```
library(car)
mi1<-lm(Dbh~bcPower(Height,lambda=1/2),data=ufc)

summary(mi1)
```

```
##
## Call:
## lm(formula = Dbh ~ bcPower(Height, lambda = 1/2), data = ufc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -217.42  -74.56  -15.31   70.28  507.50
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -473.744     34.995  -13.54   <2e-16 ***
## bcPower(Height, lambda = 1/2)    29.167      1.198   24.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 112.7 on 370 degrees of freedom
```

```
## Multiple R-squared:  0.6157, Adjusted R-squared:  0.6146
## F-statistic: 592.7 on 1 and 370 DF,  p-value: < 2.2e-16
```

```
shapiro.test(mi1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mi1$residuals
## W = 0.96654, p-value = 1.597e-07
```

Resultado:

- W = 0.96654: Este valor es el estadístico de prueba. Se acerca a 1 cuando los datos se distribuyen normalmente.

- p-value = 1.597e-07: Este valor es muy cercano a 0.

Conclusión:

Dado que el valor p es mucho menor que el nivel de significancia convencional de 0.05, **se rechaza la hipótesis nula** de que los datos se distribuyen normalmente. Esto significa que los residuos del modelo ml1 (ml1$residuals) no siguen una distribución normal.

```
library(lmtest)
```

```
## Cargando paquete requerido: zoo
```

```
##
## Adjuntando el paquete: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(mi1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mi1
## BP = 34.63, df = 1, p-value = 3.987e-09
```

Interpretación

p-valor es mucho menor que el nivel de significancia convencional de 0.05, se rechaza la hipótesis nula de homocedasticidad. Esto significa que hay evidencia sólida de que la varianza de los errores no es constante en el modelo, es decir, existe heterocedasticidad.

Transformation	$r^2 \uparrow$	$\sqrt{CME} \downarrow$
Inversa	42.56%	137.8
Raiz cuadrada	61.57%	112.7

La transformación de raíz cuadrada presenta un mejor ajuste que la transformación inversa:

- El modelo con la transformación de raíz cuadrada tiene un R^2 más alto (61.57%) en comparación con el modelo con la transformación inversa (42.56%). Esto significa que el modelo con la transformación de raíz cuadrada explica una mayor proporción de la variabilidad en los datos.
- Además, el √CME es menor para la transformación de raíz cuadrada (112.7) en comparación con la transformación inversa (137.8), lo que indica que el modelo con la transformación de raíz cuadrada se ajusta mejor a los datos y tiene una menor cantidad de error.

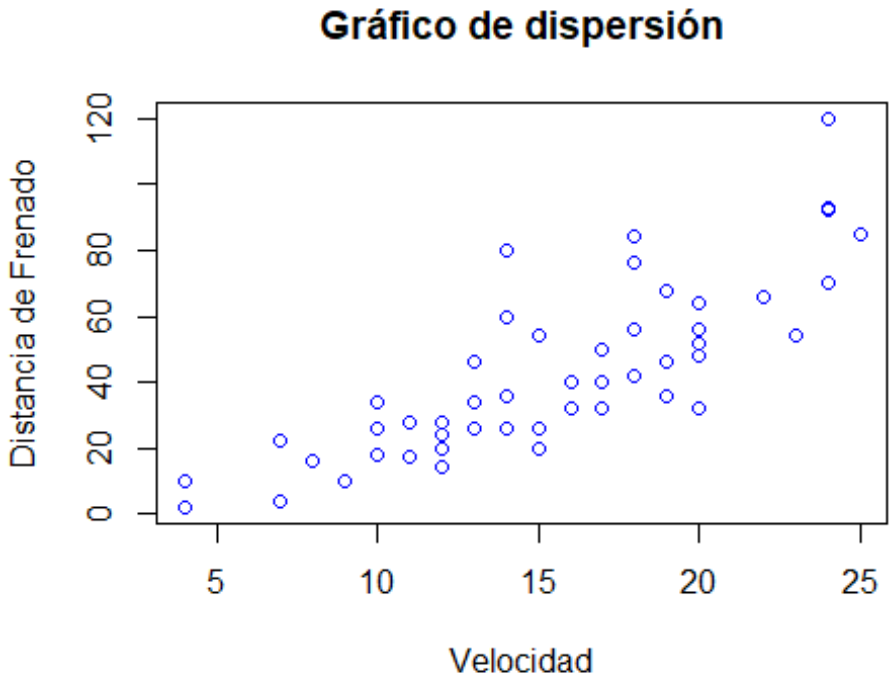
Caso 3: Vehículos

El dataset cars tiene datos sobre la velocidad y la distancia de frenado de 50 autos. Se desea obtener un modelo de regresión que permita predecir la distancia de frenado dist (en pies) en función de la velocidad speed (en millas por hora).

a) Usar un diagrama de dispersión para identificar el tipo de relación existente entre las variables.

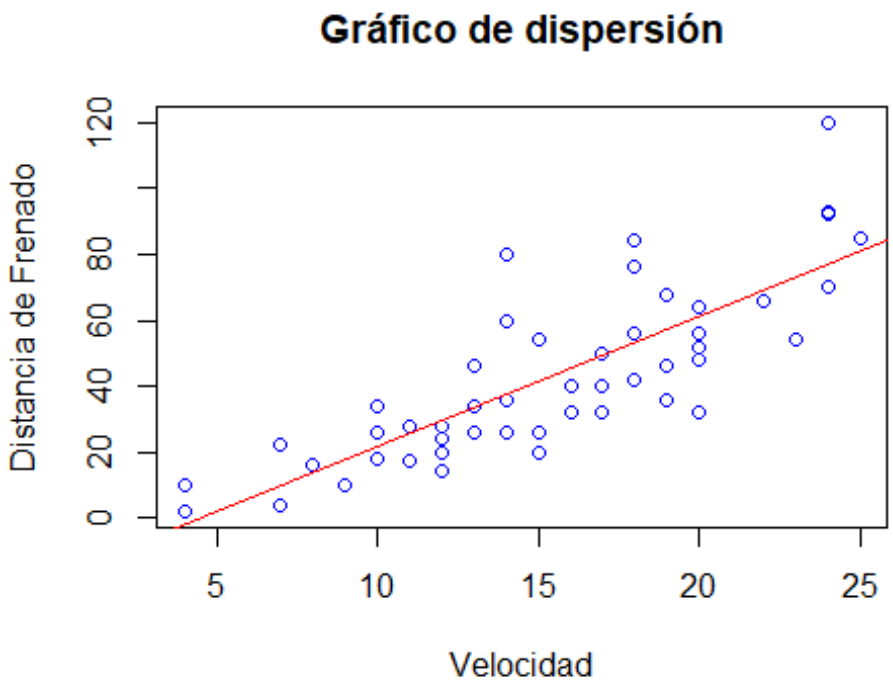
```
data(cars)

plot(dist~speed,xlab="Velocidad", ylab="Distancia de Frenado",
```

```
     col="blue",main="Gráfico de dispersión",
     data=cars)
```



b) Obtener un modelo de regresión lineal y evaluar el cumplimiento del supuesto de normalidad de los errores usando un nivel de significación del 5%. Use la prueba de Anderson-Darling

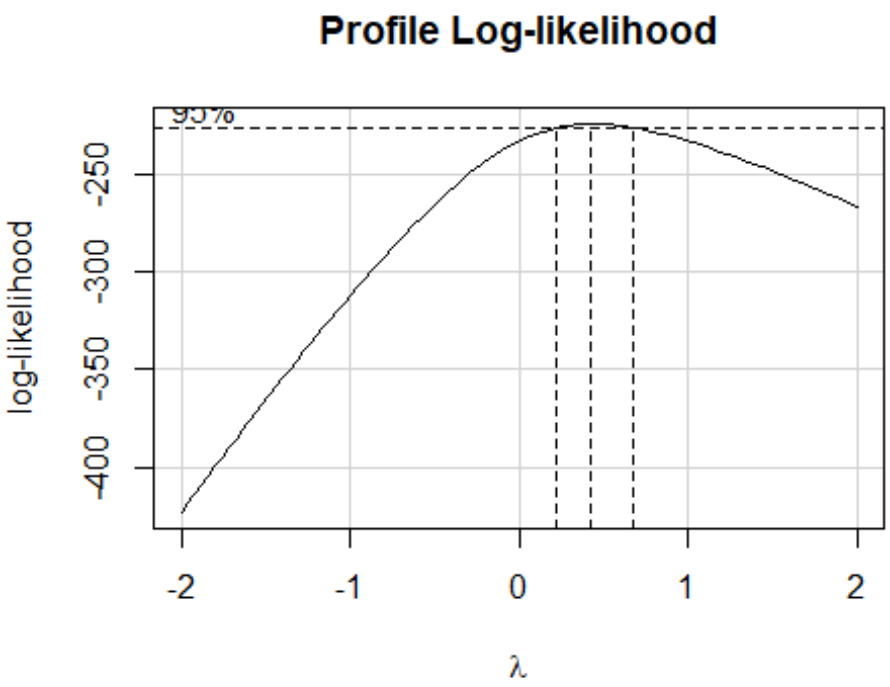
```
ml2<-lm(dist~speed,data=cars)

plot(dist~speed,xlab="Velocidad", ylab="Distancia de Frenado",
     col="blue",main="Gráfico de dispersión",
     data=cars)

abline(ml2,col="red")
```




```
library(nortest)
library(lmtest)
ad.test(ml2$residuals)

##
##  Anderson-Darling normality test
##
## data:  ml2$residuals
## A = 0.79406, p-value = 0.0369

ad.test(rstandard(ml2))

##
##  Anderson-Darling normality test
##
## data:  rstandard(ml2)
## A = 0.8005, p-value = 0.03555

bptest(ml2)

##
##  studentized Breusch-Pagan test
##
## data:  ml2
## BP = 3.2149, df = 1, p-value = 0.07297
```

c) Usar la familia de transformaciones de Box-Cox para elegir un valor apropiado del parámetro λ.

```
library(car)
box<-boxCox(ml2,lambda=seq(-2,2,by=0.1))
```




```
lambda<-box$x[which.max(box$y)] # which.max buscar en el vector el valor maximo.Nos entrega la ubicacion del valor maximo
lambda

## [1] 0.4242424
```

Box-Cox:Probar diferentes valores de lambda de tal manera que se logre minimizar el cuadrado medio del error.
Recomendable (-2,2, by=0.1)

d) Aplicar la transformación propuesta por Box-Cox. Obtener el nuevo modelo de regresión y evaluar el supuesto de normalidad de los errores. Use la prueba de Anderson-Darling

```
ml3<-lm(dist^lambda~speed,data=cars)
ad.test(ml3$residuals)

##
##  Anderson-Darling normality test
##
## data:  ml3$residuals
## A = 0.34822, p-value = 0.4636
```

```r
ad.test(rstandard(ml3))
```

```
##
##  Anderson-Darling normality test
##
## data:  rstandard(ml3)
## A = 0.33505, p-value = 0.4972
```

```r
bptest(ml3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  ml3
## BP = 0.13933, df = 1, p-value = 0.709
```

Interpretacióon

- Prueba de Normalidad de Anderson-Darling

Ambos p-valores son mayores a 0.05: Esto indica que no se puede rechazar la hipótesis nula de que los residuos se distribuyen normalmente. En otras palabras, no hay evidencia suficiente para afirmar que los residuos no siguen una distribución normal.

- Prueba de Breusch-Pagan Estudiantizada

p-value = 0.709: Este valor es mucho mayor que el nivel de significancia convencional de 0.05. Esto significa que no se puede rechazar la hipótesis nula de homocedasticidad. En otras palabras, no hay evidencia de que la varianza de los errores varíe significativamente con los valores de las variables independientes.