PRÁCTICA PUNTUABLE EN R (BIOESTADÍSTICA/(WDBC))

Leonardo Madsen

6/1/2021

Cargamos librerías

```
library(dplyr)
library(ggplot2)
library(DescTools)
library(lmtest)
```

Leemos la tabla de datos

```
WDBC <- read.table(file.path("./WDBC.dat"), header=TRUE, sep = "\t")
```

Eliminamos las columnas "id", "..._se" y "..._worst"

```
WDBC <- WDBC[, -grep("id|_se|_worst", colnames(WDBC))]
```

```
colnames(WDBC) <- gsub('_mean', '', colnames(WDBC), fixed=TRUE)
```

```
WDBC$diagnosis <- as.factor(recode(WDBC$diagnosis, B = "Benigno", M = "Maligno"))
```

Sumario de los datos que utilizaremos

```
summary(WDBC)
```

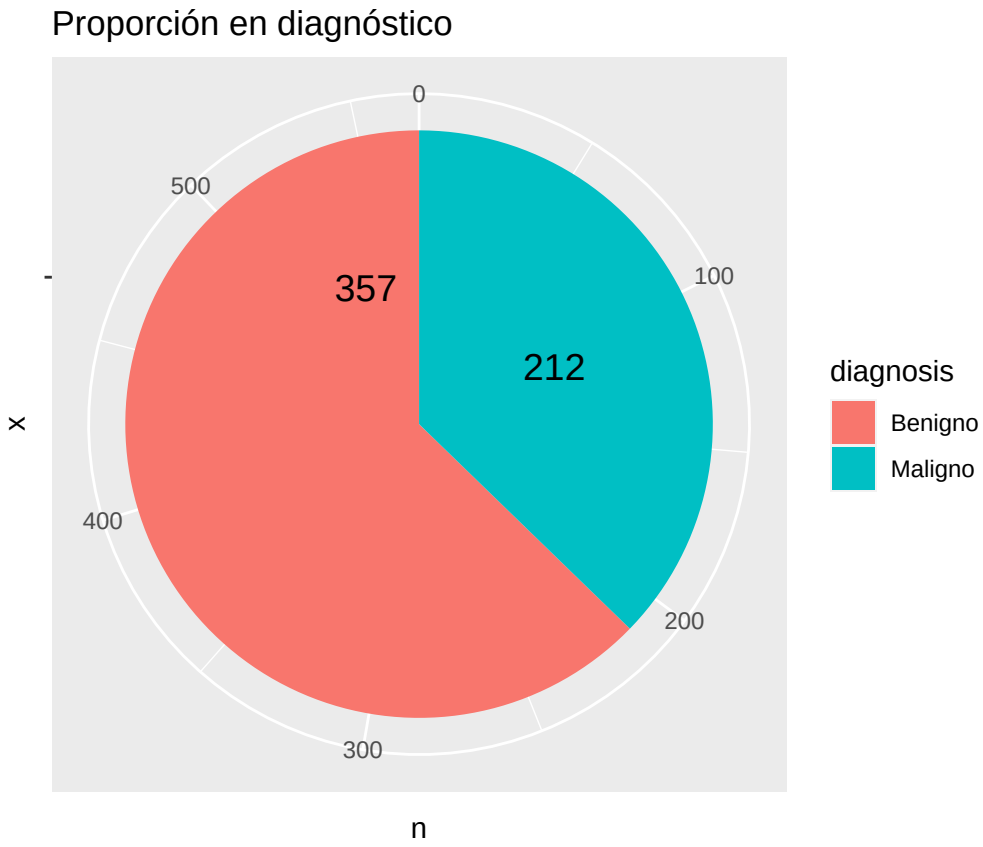
```
##    diagnosis        age            radius          texture
##  Benigno:357   Min.   :30.00   Min.   : 6.981   Min.   : 9.71
##  Maligno:212   1st Qu.:44.00   1st Qu.:11.700   1st Qu.:16.17
##                Median :58.00   Median :13.370   Median :18.84
##                Mean   :56.29   Mean   :14.127   Mean   :19.29
##                3rd Qu.:69.00   3rd Qu.:15.780   3rd Qu.:21.80
##                Max.   :80.00   Max.   :28.110   Max.   :39.28
##    perimeter           area          smoothness       compactness
##  Min.   : 43.79   Min.   : 143.5   Min.   :0.05263   Min.   :0.01938
##  1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492
##  Median : 86.24   Median : 551.1   Median :0.09587   Median :0.09263
##  Mean   : 91.97   Mean   : 654.9   Mean   :0.09636   Mean   :0.10434
##  3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040
##  Max.   :188.50   Max.   :2501.0   Max.   :0.16340   Max.   :0.34540
##    concavity       concave.points       symmetry      fractal_dimension
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.1060   Min.   :0.04996
##  1st Qu.:0.02956   1st Qu.:0.02031   1st Qu.:0.1619   1st Qu.:0.05770
##  Median :0.06154   Median :0.03350   Median :0.1792   Median :0.06154
##  Mean   :0.08880   Mean   :0.04892   Mean   :0.1812   Mean   :0.06280
##  3rd Qu.:0.13070   3rd Qu.:0.07400   3rd Qu.:0.1957   3rd Qu.:0.06612
##  Max.   :0.42680   Max.   :0.20120   Max.   :0.3040   Max.   :0.09744
```

Recuento de diagnosis

```
count.diag <- count(WDBC, diagnosis)
count.diag
```

```
##   diagnosis   n
## 1   Benigno 357
## 2   Maligno 212
```

```
pie<- ggplot(count.diag, aes(x="", y=n, fill=diagnosis)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  geom_text(aes(y = n/2 + c(cumsum(n)[-length(n)], 0),
            label = n), size=5) +
  ggtitle("Proporción en diagnóstico")
pie
```



1. Seleccionar aleatoriamente una muestra sin reemplazamiento de 50 individuos de cada diagnóstico.

```
set.seed(1832)
```

Selección aleatoria

```
df <- WDBC %>%
  group_by(diagnosis) %>%
  slice_sample(n = 50, replace = FALSE)
df
```

```
## # A tibble: 100 x 12
## # Groups:   diagnosis [2]
##    diagnosis   age radius texture perimeter  area smoothness compactness
##    <fct>     <int>  <dbl>   <dbl>     <dbl> <dbl>      <dbl>       <dbl>
##  1 Benigno      59  15.1     16.4      99.6  674.     0.115       0.181
##  2 Benigno      36   9.33    21.9      59.0  264      0.0924      0.0560
##  3 Benigno      61  13.8     19.6      88.7  593.     0.0868      0.0633
##  4 Benigno      54  13.8     15.2      89.0  587.     0.0952      0.0769
##  5 Benigno      51  12.3     17.9      78.4  466.     0.0868      0.0653
##  6 Benigno      49  13.0     18.6      85.1  512      0.108       0.130
##  7 Benigno      55  10.5     19.3      67.4  336.     0.0999      0.0858
##  8 Benigno      78  14.7     25.4      94.7  669.     0.0828      0.0721
##  9 Benigno      79  12.8     16.7      82.5  494.     0.112       0.112
## 10 Benigno      56  12.2     17.8      77.8  451.     0.104       0.0706
## # ... with 90 more rows, and 4 more variables: concavity <dbl>,
## #   concave.points <dbl>, symmetry <dbl>, fractal_dimension <dbl>
```

Comprobamos resultado de la selección aleatoria

```
df %>% count(diagnosis)
```

```
## # A tibble: 2 x 2
## # Groups:   diagnosis [2]
##   diagnosis     n
##   <fct>     <int>
## 1 Benigno      50
## 2 Maligno      50
```

```
quant.03area = quantile(df$area,0.333)
print(quant.03area)
```

```
##    33.3%
## 537.6591
```

```
quant.06area = quantile(df$area,0.666)
print(quant.06area)
```

```
##    66.6%
## 731.9446
```

Categorizamos la variable "area" en tres categorías cada una de un tercio de la cantidad de los datos

```
df[,"area.categorica"] = cut(df$area, breaks = c(min(df$area), quant.03area, quant.06area, max(df$area))
    labels = c("Pequeña", "Media", "Grande"),
    include.lowest = TRUE)

df$area.categorica <- as.factor(df$area.categorica)

table(df$area.categorica)
```

```
##
## Pequeña   Media  Grande
##      33      33      34
```

Categorizamos la variable "texture" en dos categorías divididas según la media

```
df[,"textura.categorica"] = cut(df$texture,
                      breaks = c(min(df$texture), mean(df$texture), max(df$texture)),
                      labels = c("Claro", "Oscuro"),
```

```
                      include.lowest = TRUE)

df$textura.categorica <- as.factor(df$textura.categorica)

table(df$textura.categorica)
```

```
##
##  Claro Oscuro
##     49     51
```

Análisis inferencial.

En todos los apartados que siguen especificar con detalle las hipótesis de los test, los cálculos hechos y las conclusiones. Considerar una de las variables continuas del conjunto de datos en la que sea posible asumir normalidad. Asumiendo normalidad:

a. ¿Entre qué valores se mueve la media de la distribución con una confianza del 95%? Suponer varianza desconocida. Interpretar los resultados.

Shapiro

```
shapiro.test(df$texture)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$texture
## W = 0.99016, p-value = 0.6773
```

Nuestro p-value = 0.6773 es mayor al nivel de confianza fijado (0.05) por lo cual no hay evidencias para

```
IC=t.test(df$texture, conf.level = 0.95)
IC$conf.int
```

```
## [1] 18.63809 20.15371
## attr(,"conf.level")
## [1] 0.95
```

La media de la distribución de la variable "texture" se mueve entre los valores 18.63809 y 20.15371 con

```
shapiro.test(df$perimeter)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$perimeter
## W = 0.96958, p-value = 0.02054
```

Nuestro p-value = 0.02054 es menor al nivel de confianza fijado (0.05) por lo cual sí hay evidencias pa

b. Hacer un contraste sobre la media de nivel 0.1 tomando como hipótesis nula el extremo superior del intervalo bilateral del apartado a. Dar el p-valor e interpretar el resultado.

```
limite_sup= IC$conf.int[2]
t.test(df$texture, alternative='two.sided',
       conf.level=0.9, mu=limite_sup)
```

```
##
##  One Sample t-test
##
## data:  df$texture
## t = -1.9842, df = 99, p-value = 0.05
## alternative hypothesis: true mean is not equal to 20.15371
## 90 percent confidence interval:
##  18.76177 20.03003
## sample estimates:
## mean of x
##   19.3959
```

H0: mu=20.15371 y dado que el p-value = 0.05 es menor al nuestro nivel de confianza (0.1) sí hay evidenc

Considerar esa misma variable cuantitativa y una variable cualitativa con dos niveles. Suponiendo normalidad:

c.Hacer un contraste de igualdad de varianzas a nivel 0.05. Dar el p-valor e interpretar el resultado.

HOMOCEDASTICIDAD Es la homogeneidad de varianza de la variable dependiente entre los grupos.

```
bartlett.test(df$texture~df$area.categorica)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  df$texture by df$area.categorica
## Bartlett's K-squared = 1.3903, df = 2, p-value = 0.499
```

Interpretación:
Con un p-value = 0.499, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto suponemos hor

```
bartlett.test(df$texture~df$diagnosis)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  df$texture by df$diagnosis
## Bartlett's K-squared = 0.11539, df = 1, p-value = 0.7341
```

Interpretación:
Con un p-value = 0.7341, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto suponemos h

d. Hacer un contraste de comparación de medias con varianzas desconocidas, dar el p-valor e interpretar el resultado

```
t.test(df$texture~df$diagnosis)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  df$texture by df$diagnosis
## t = -5.1513, df = 97.768, p-value = 1.344e-06
## alternative hypothesis: true difference in means between group Benigno and group Maligno is not equal
## 95 percent confidence interval:
##  -4.859731 -2.156669
## sample estimates:
## mean in group Benigno mean in group Maligno
##               17.6418               21.1500
```

Hay evidencias para rechazar la hipótesis de igualdad de medias debido a que p-value = 1.344e-06 es meno

f. Hacer un contraste chi-cuadrado. Dar el p-valor e interpretar el resultado.

```
table(df$diagnosis,df$area.categorica)
```

```
##
##           Pequeña Media Grande
##   Benigno      31    18      1
##   Maligno       2    15     33
```

Comprobamos las siguientes hipótesis
H0: No existe relación entre diagnosis y textura.categorica
H1: Si existe relación entre diagnosis y textura.categorica

```
chisq.test(table(df$diagnosis,df$textura.categorica))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(df$diagnosis, df$textura.categorica)
## X-squared = 19.368, df = 1, p-value = 1.078e-05
```

Puesto que p-value = 1.078e-05 es menor a 0.05, no hay evidencias para aceptar H0, por lo que concluímos

Comprobamos las siguientes hipótesis
H_0: No existe relación entre diagnosis y area.categorica
H_1: Si existe relación entre diagnosis y area.categorica

```
chisq.test(table(df$diagnosis,df$area.categorica))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$diagnosis, df$area.categorica)
## X-squared = 55.875, df = 2, p-value = 7.36e-13
```

Puesto que p-value = 7.36e-13 es menor a 0.05, no hay evidencias para aceptar H0, por lo que concluímos

g. Calcular una medida de asociación junto con un intervalo de confianza del 95%. Interpretar los resultados.

```
diag_tex_tb <- table(df$textura.categorica,df$diagnosis)
diag_tex_tb
```

```
##
##          Benigno Maligno
##   Claro       36      13
```

```
##   Oscuro      14      37
```

El odds ratio (OR) expresa si la probabilidad de ocurrencia del evento Benigno/Maligno difiere o no en l

```
res <- OddsRatio(diag_tex_tb, conf.level=0.95)
res
```

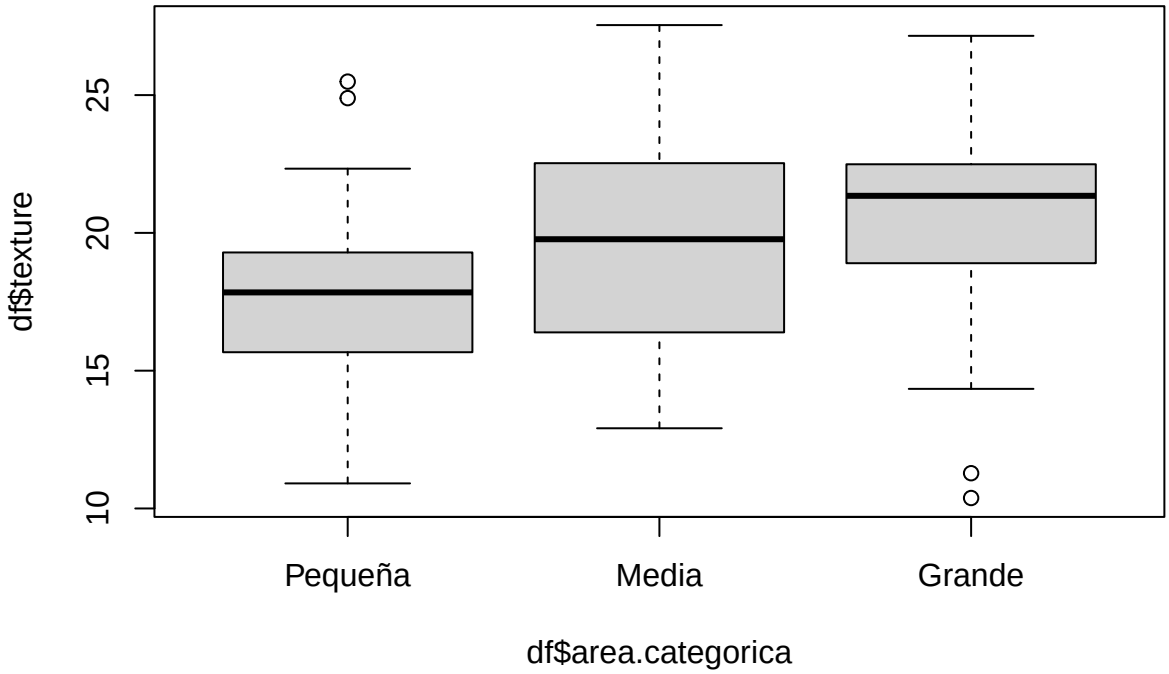
```
## odds ratio     lwr.ci     upr.ci
##   7.318681   3.025323  17.704917
```

Vemos que la fuerza de asociación es alta.
Entre los casos de "textura.categorica" "Claro" hay 7 veces más "diagnosis Benigno" por cada "diagnosi
El intervalo de confianza al 95% es (3.025323, 17.704917)

4. ANOVA y Regresión:

a. Categorizar, en tres grupos, una de las variables cuantitativas medidas en las imágenes y llevar a cabo un ANOVA.

```
boxplot(df$texture ~ df$area.categorica)
```



A simple vista podríamos intuir que la media de "texture" con iguales en las tres categorías de área.

Procedemos con una prueba paramétrica ANOVA

```
aov1 <-aov(df$texture ~ df$area.categorica)
summary(aov1)
```

```
##                    Df Sum Sq Mean Sq F value  Pr(>F)
## df$area.categorica  2  154.5   77.27   5.812 0.00413 **
## Residuals          97 1289.5   13.29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El nivel de significancia Pr(>F)=0.00413 es menor que "0.05" por lo que hay evidencias para rechazar "H

Aun no sabemos cuál media es diferente de cuál. aunque podemos intuir que son las medias de "area pequei

Realizamos la prueba de Tukey

```
TukeyHSD(aov1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = df$texture ~ df$area.categorica)
##
## $`df$area.categorica`
##                       diff       lwr      upr     p adj
## Media-Pequeña  2.0690909 -0.067380 4.205562 0.0598061
## Grande-Pequeña 2.9681194  0.847416 5.088823 0.0034765
## Grande-Media   0.8990285 -1.221675 3.019732 0.5729622
```

Puesto que "Grande-Pequeña" p adj = 0.0034 es menor a 0.05 podemos corroborar que las medias de Grande y

b. Efectuar un análisis de regresión lineal simple de dos de las variables continuas del conjunto de datos.

I. Dar la ecuación del modelo. Interpretar el modelo.

```
model0 <- lm(formula =  area ~ texture, data = df)
summary(model0)
```

```
##
## Call:
## lm(formula = area ~ texture, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -688.64 -227.33  -81.82  181.66 1024.77
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  242.972    173.417   1.401   0.1643
## texture       24.498      8.774   2.792   0.0063 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 333.4 on 98 degrees of freedom
## Multiple R-squared:  0.07369,    Adjusted R-squared:  0.06424
## F-statistic: 7.796 on 1 and 98 DF,  p-value: 0.006296
```

Interpretación:

El área está linealmente relacionada con la textura según la siguiente fórmula:

$area = 242.972 + 24.498 * texture$

La ordenada en el origen (Intercept) es 242.972 por lo que valores de textura igual a 0 estimarán valor
La pendiente es 24.498 por lo que cada aumento medio en una unidad de textura producirá un aumento de 2
El estadístico F (7.796) contrasta si el modelo tiene significativa capacidad predictiva.
En el contraste la hipótesis nula es F = 1, con un p-valor menor de 0.05 (p-value: 0.006296) se rechaza

II. Estudiar la significación del modelo y la bondad de ajuste.

Se muestra un valor del estadístico de contraste F de 7.796 con un p_valor = 0.006296. Deduciendo qu
Respecto a la bondad del ajuste, el coeficiente de determinación R^2 tiene un valor de 0.07369, indica

III. Hacer un análisis residual, incluir los gráficos apropiados y estudiar la adecuación del modelo.

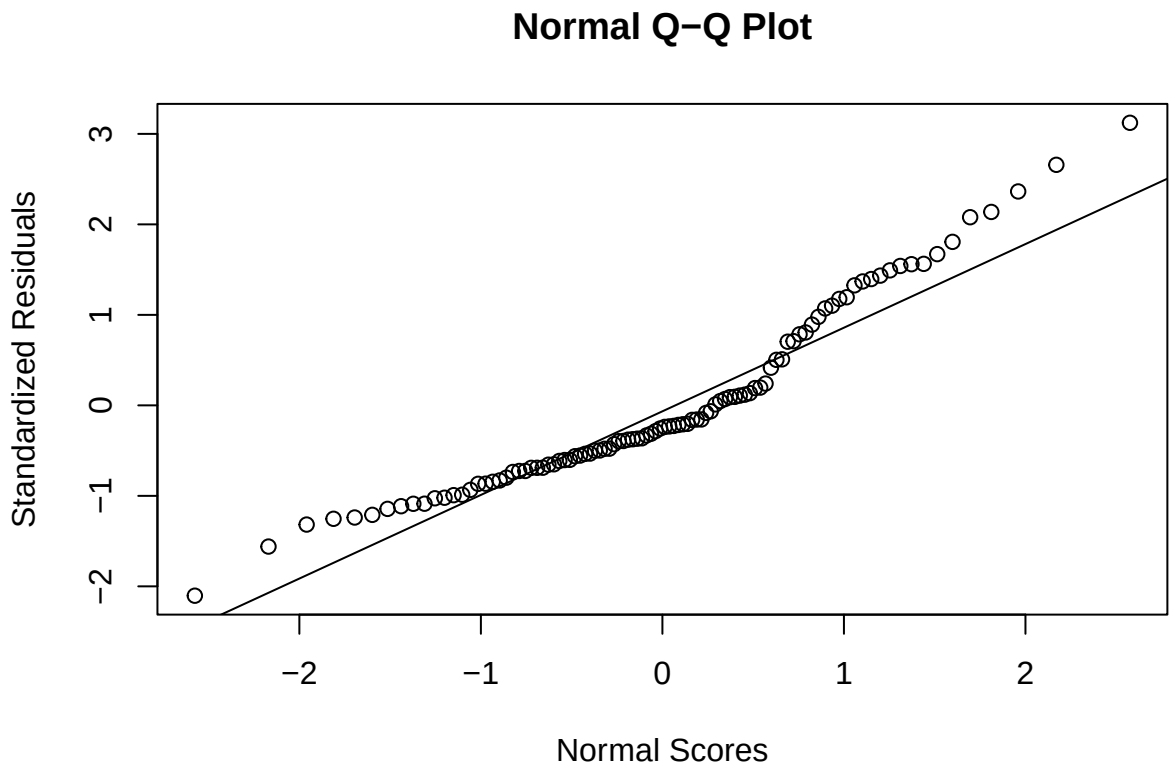
Realizaremos el diagnóstico de los residuos. Normalidad de los residuos, homogeneidad de varianzas e in

```
shapiro.test(model0$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model0$residuals
## W = 0.93329, p-value = 7.683e-05
```

El Shapiro-Wilk normality test nos indica que no hay evidencias para aceptar la hipótesis nula (H0:lo

```
model0.stdres <- rstandard(model0)
qqnorm(model0.stdres,  ylab="Standardized Residuals", xlab="Normal Scores")
qqline(model0.stdres)
```




```
bptest(model0)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model0
## BP = 2.6195, df = 1, p-value = 0.1056
```

Interpretación:

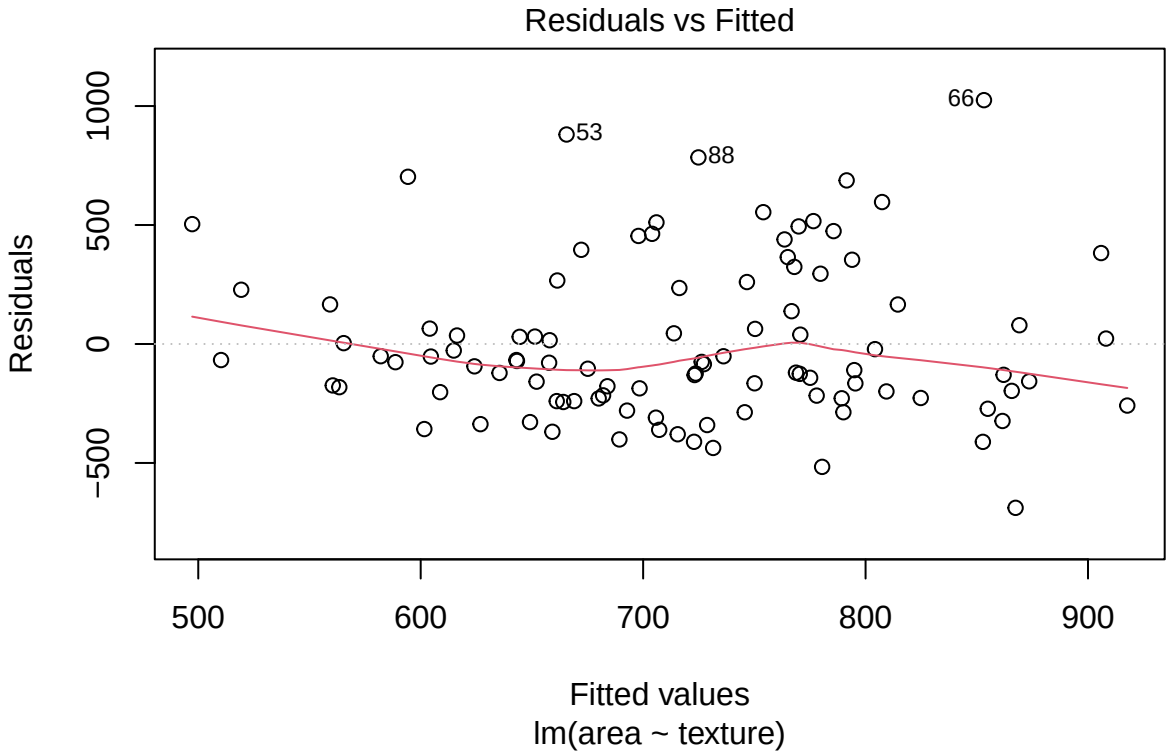
Con un p-value = 0.1056, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto suponemos l

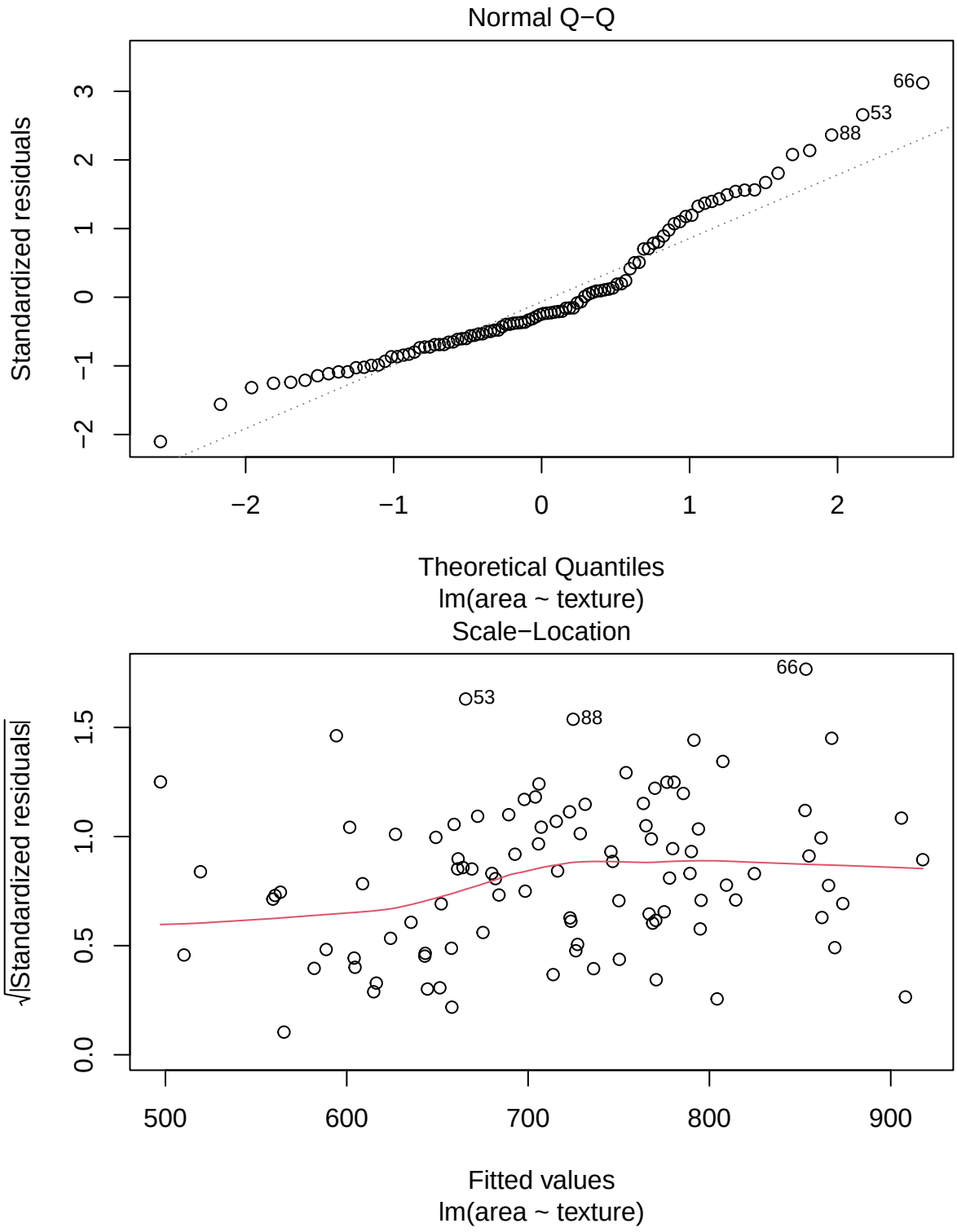
```
dwtest(model0)
```

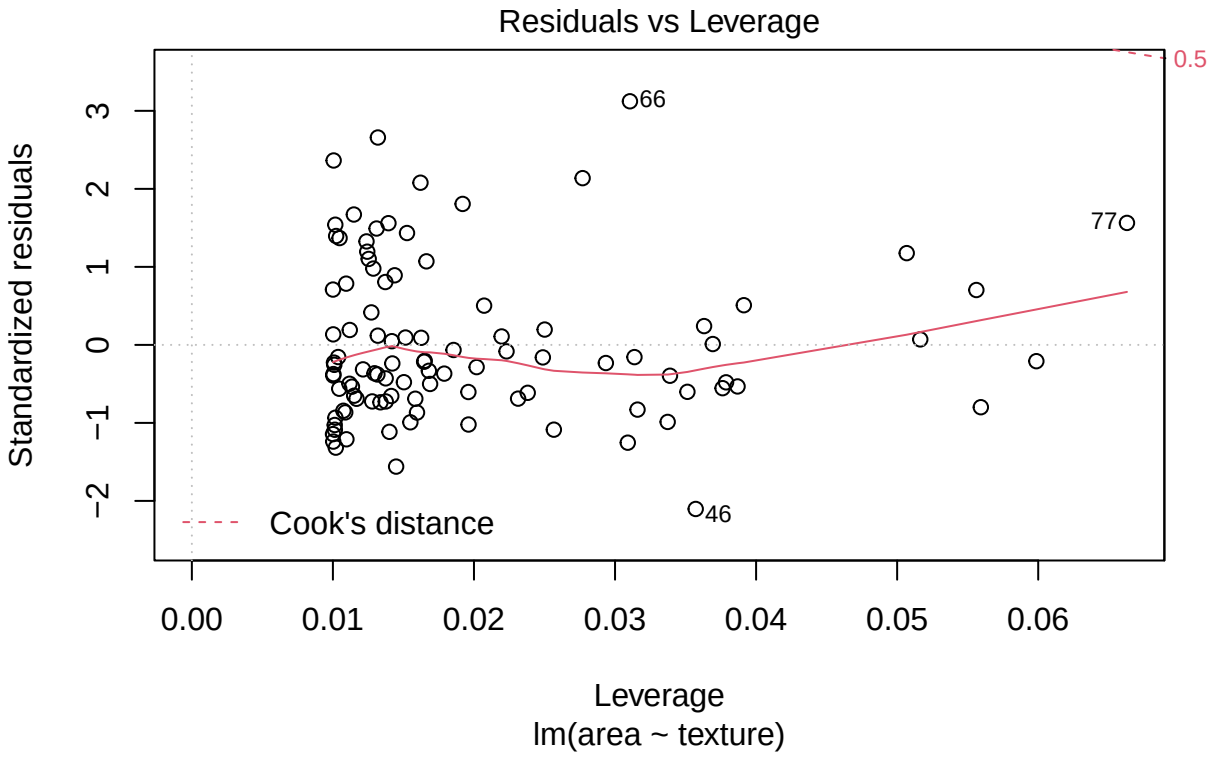
```
##
##  Durbin-Watson test
##
## data:  model0
## DW = 1.2234, p-value = 3.698e-05
## alternative hypothesis: true autocorrelation is greater than 0
```

Interpretación:
Con un p-value = 3.698e-05, menor de 0.05, hay indicios para rechazar la hipótesis nula. Por lo tanto no

```
plot(model0)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(area ~ texture)
66
53
88
−2
−1
0
1
2
3
Scale−Location
√|Standardized residuals|
Fitted values
lm(area ~ texture)
0.0
0.5
1.0
1.5
500
600
700
800
900



IV.Si el modelo no es apropiado tratar de encontrar transformaciones que corrijan el modelo obteniendo un modelo aceptable.

```
model1 <- lm(formula =  area ~ I(texture^2), data = df)
model2 <- lm(formula =  area ~ log(texture), data = df)
model3 <- lm(formula =  area ~ I(texture^(-1)), data = df)
model_cuadratico <- lm(formula =  area ~ poly(texture, 1), data = df)
model_cuadratico2 <- lm(formula =  area ~ poly(texture, 2), data = df)
```

```
 anova(model0, model1)
```

```
## Analysis of Variance Table
##
## Model 1: area ~ texture
## Model 2: area ~ I(texture^2)
##   Res.Df      RSS Df Sum of Sq F Pr(>F)
## 1     98 10894349
## 2     98 10890032  0    4317.3
```

```
 anova(model0, model3)
```

```
## Analysis of Variance Table
##
## Model 1: area ~ texture
## Model 2: area ~ I(texture^(-1))
##   Res.Df      RSS Df Sum of Sq F Pr(>F)
## 1     98 10894349
## 2     98 11072194  0   -177845
```

```
anova(model0, model_cuadratico)
```

```
## Analysis of Variance Table
##
## Model 1: area ~ texture
## Model 2: area ~ poly(texture, 1)
##   Res.Df      RSS Df  Sum of Sq F Pr(>F)
## 1     98 10894349
## 2     98 10894349  0 1.8626e-09
```

```
anova(model0, model_cuadratico2)
```

```
## Analysis of Variance Table
##
## Model 1: area ~ texture
## Model 2: area ~ poly(texture, 2)
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     98 10894349
## 2     97 10888586  1    5763.1 0.0513 0.8212
```

Tratando de encontrar transformaciones que mejore el modelo no hemos obteniendo un modelo aceptable.

Comparamos los modelos siguiendo el criterio de información de Akaike

```
cbind(model0=AIC(model0), model1=AIC(model1), model2=AIC(model2), model3=AIC(model3), model_cuadratico=A
```

```
##        model0   model1   model2   model3 model_cuadratico model_cuadratico2
## [1,] 1449.646 1449.607 1450.205 1451.265         1449.646          1451.593
```

Según este criterio el mejor modelo es el que contiene la transformación logarítmica de la variable tex

c. A partir de la variable diagnosis , ajustar un modelo que nos permita predecir la probabilidad de tener un tumor maligno. El modelo debe incluir al menos 5 variables independientes. Identifica e interpreta factores de riesgo/protección.

Convertimos la variable "diagnosis" a numérico (Benigno = 0, Maligno = 1)

```
df$diagnosis.fac <- c(rep(0,50),rep(1,50))
```

```
model.m <- lm(formula =  diagnosis.fac ~ I(perimeter^2) + radius + texture  + concavity + area, data = d
summary(model.m)
```

```
##
## Call:
## lm(formula = diagnosis.fac ~ I(perimeter^2) + radius + texture +
##     concavity + area, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60550 -0.14438 -0.00133  0.17162  0.68138
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.811e+00  4.575e-01  -3.958 0.000147 ***
## I(perimeter^2) -3.130e-04  6.753e-05  -4.635 1.15e-05 ***
## radius          1.223e-01  5.633e-02   2.170 0.032500 *
## texture         3.378e-02  6.958e-03   4.855 4.78e-06 ***
## concavity       4.679e+00  6.148e-01   7.611 2.07e-11 ***
## area            3.436e-03  1.106e-03   3.106 0.002508 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.253 on 94 degrees of freedom
## Multiple R-squared:  0.7593, Adjusted R-squared:  0.7465
## F-statistic: 59.32 on 5 and 94 DF,  p-value: < 2.2e-16
```

Interpretación:
diagnosis está linealmente relacionada con perimeter, radius, texture, concavity y area según la siguie

$diagnosis = -1.811 - 3.130e^{-4} * perimeter^2 + 1.223e^{-1} * radius + 3.378e^{-2} * texture + 4.679 * concavity + 3.436e^{-3} * area$

El estadístico F (59.32) contrasta si el modelo tiene significativa capacidad predictiva.
En el contraste la hipótesis nula es F = 1, con un p-valor menor de 0.05 (p-value: < 2.2e-16) se rechaza
El R2 ajustado = 0.7465 , lo que significa que "perimeter, radius, texture, concavity y area pueden pre
Los coeficiente de las variables radius, texture, concavity y area son positivos por tanto corresponde
Por el contrario, el coeficiente de perimeter^2 es negativo por lo que se trata de un factor de protecc