Regresión, modelos y métodos Prueba de evaluación continua 1

Demetrio Muñoz Alvarez

2023-05-07

R Markdown

Antes de comenzar con el primer ejercicio, hemos establecido el directorio de trabajo y cargado los datos del conjunto "**cicindela**". También hemos llevado a cabo un análisis descriptivo de los datos para familiarizarnos con ellos:

```
setwd("C:/Users/Deme/Desktop/Master/Regresion modelos y metodos/PEC 1/Pec 1 Modelos") #Est
ablecemos el directorio de trabajo.
library(readxl)
cicindela <- read_excel("cicindela.xlsx") #Cargarmos los datos de cicindela.
head(cicindela) #Mostramos las primeras entradas del conjunto de datos.
```

```
## # A tibble: 6 × 5
##   BeetleDensity `Wave exposure` Sandparticlesize `Beach steepness`
##           <dbl>           <dbl>            <dbl>              <dbl>
## 1            13               5                2                 17
## 2            12               4                2                  8
## 3            54              10                7                 15
## 4            19               9                4                  7
## 5            37               8                6                  6
## 6             2               4                1                  9
## # ℹ 1 more variable: AmphipodDensity <dbl>
```

```
summary(cicindela) #Resumen de los datos de cicindela.
```

```
##  BeetleDensity   Wave exposure   Sandparticlesize Beach steepness
##  Min.   : 1.00   Min.   : 4.00   Min.   :1.000    Min.   : 6.0
##  1st Qu.:12.75   1st Qu.: 4.75   1st Qu.:1.750    1st Qu.: 8.0
##  Median :18.50   Median : 8.00   Median :3.000    Median :10.5
##  Mean   :23.17   Mean   : 7.25   Mean   :3.333    Mean   :10.5
##  3rd Qu.:33.25   3rd Qu.: 8.25   3rd Qu.:4.500    3rd Qu.:12.0
##  Max.   :54.00   Max.   :11.00   Max.   :7.000    Max.   :17.0
##  AmphipodDensity
##  Min.   : 5.00
##  1st Qu.: 7.50
##  Median :12.50
##  Mean   :11.67
##  3rd Qu.:14.50
##  Max.   :19.00
```

```
str(cicindela) #Estructura de los datos de cicindela.
```

```
## tibble [12 × 5] (S3: tbl_df/tbl/data.frame)
##  $ BeetleDensity   : num [1:12] 13 12 54 19 37 2 18 1 53 16 ...
##  $ Wave exposure   : num [1:12] 5 4 10 9 8 4 8 4 11 8 ...
##  $ Sandparticlesize: num [1:12] 2 2 7 4 6 1 4 1 6 1 ...
##  $ Beach steepness : num [1:12] 17 8 15 7 6 9 8 11 12 10 ...
##  $ AmphipodDensity : num [1:12] 14 16 6 14 8 17 12 19 5 13 ...
```

Ejercicio 1

Ajustar un modelo de regresión lineal múltiple que estime todos los coeficientes de regresión parciales referentes a todas las variables regresoras y el intercepto.

Con el conjunto de datos “cicindela”, hemos llevado a cabo un análisis de regresión lineal múltiple para examinar el efecto de varias variables predictoras (“Wave exposure”, “Sandparticlesize”, “Beach steepness”, “Amphipoddensity”) sobre la variable respuesta “Beetledensity” (densidad de escarabajos) de la especie *cicindela dorsalis dorsalis* en 12 playas del Atlántico Norte, donde se han realizado muestreos.

Hemos ajustado el modelo de regresión utilizando estas variables predictoras:

```
modelo_1 <- lm(BeetleDensity ~ `Wave exposure` + Sandparticlesize + `Beach steepness` + Am
phipodDensity, data = cicindela) #Ajustamos el modelo de regresión linel multiple con la f
unción lm().
summary(modelo_1) #Mostramos el resumen del modelo.
```

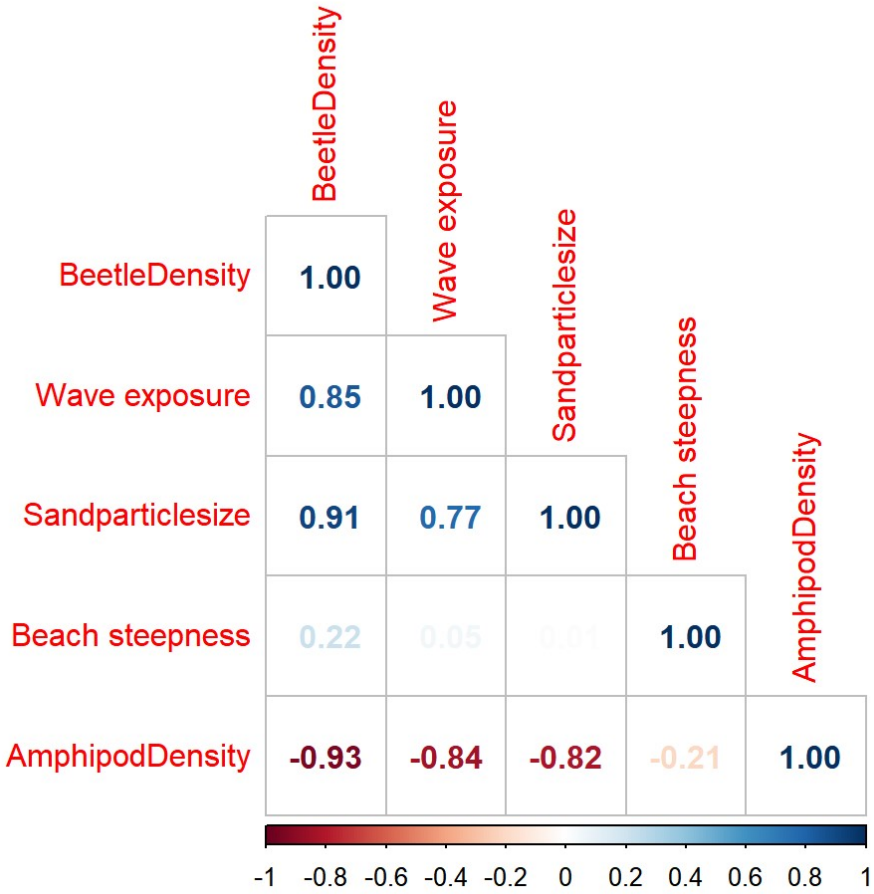
```
##
## Call:
## lm(formula = BeetleDensity ~ `Wave exposure` + Sandparticlesize +
##     `Beach steepness` + AmphipodDensity, data = cicindela)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3004 -2.7038  0.0795  2.6017  5.3924
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        14.9531    17.2661   0.866   0.4152
## `Wave exposure`     0.9123     1.0935   0.834   0.4317
## Sandparticlesize    3.8970     1.1690   3.334   0.0125 *
## `Beach steepness`   0.6511     0.4530   1.437   0.1938
## AmphipodDensity    -1.5624     0.6610  -2.364   0.0501 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.513 on 7 degrees of freedom
## Multiple R-squared:  0.9578, Adjusted R-squared:  0.9337
## F-statistic: 39.71 on 4 and 7 DF,  p-value: 6.727e-05
```

A parte del ajuste del modelo, hemos generado una matriz de correlaciones para ver las relaciones de nuestras variables:

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
matriz_cor <- cor(cicindela) #Creamos la matriz de correlación.
corrplot(matriz_cor, method = "number",type = "lower") #Mostramos de forma gráfica la matr
iz de correlación con el paquete corrplot().
```



Con el modelo ajustado y la matriz de correlación, podemos observar la relación entre las variables predictoras y la variable respuesta (intercepto). Hemos notado una relación positiva entre la variable “Sandparticlesize” y la densidad de escarabajos Cicindela (“BeetleDensity”), así como una relación negativa para la variable “AmphipodDensity”. Además, el nivel de significación de estas dos variables es menor o muy cercano a α = 0.05, lo que indica que estas variables tienen un impacto en la densidad de escarabajos de la especie cicindela.

¿Es significativo el modelo obtenido? ¿Qué test estadístico se emplea para contestar a esta pregunta?. Plantear la hipótesis nula y la alternativa del test.

Se ha empleado el test F para determinar la significación global del modelo ajustado (modelo_1). Los resultados del test indican que el modelo es significativo ($F = 39.71$, $p < 0.0001$); es decir, al menos una de las variables predictoras está relacionada de manera significativa con la variable respuesta (densidad de escarabajos, “BeetleDensity”). Además, el valor del R-cuadrado ajustado, es de 0.9337, indica que el modelo explica aproximadamente el 93.37% de la variabilidad en la variable respuesta.

La hipótesis nula (H_0) es que todos los coeficientes de regresión en el modelo son iguales a cero, mientras que la hipótesis alternativa (H_1) es que al menos uno de los coeficientes de regresión es diferente de cero.

¿Qué variables han salido significativas para un nivel de significación α = 0.10?

Para un nivel de significación α = 0.10, las variables significativas son aquellas con un pvalor < 0.10. De acuerdo con esto, las únicas variables que cumple con este criterio son “Sandparticlesize”, con un pvalor = 0.0125, y “AmphipodDensity” con un pvalor = 0.0501.

Calcular los intervalos de confianza al 90 y 95% para el parámetro que acompaña a la variable AmphipodDensity. Utilizando sólo estos intervalos, ¿qué podríamos haber deducido sobre el pvalor para la densidad de los anfípodos depredadores en el resumen del modelo de regresión? ¿Qué interpretación práctica tiene este parámetro β4?

```
confint(modelo_1, "AmphipodDensity", level = 0.90) #Con la función confint() podemos calcu
lar el intervalo de confianza de una variable específica de nuestro modelo.
```

```
##                       5 %       95 %
## AmphipodDensity -2.814699 -0.3100058
```

```
confint(modelo_1, "AmphipodDensity", level = 0.95)
```

```
##                      2.5 %       97.5 %
## AmphipodDensity -3.125407 0.0007019125
```

Para el intervalo de confianza del 90% (**-2.814699, -0.3100058**) de la variable “AmphipodDensity” , podemos decir que el valor del coeficiente AmphipodDensity es significativamente diferente de cero, ya que el intervalo no incluye el valor cero. Por lo tanto, podemos afirmar que esta variable es significativa.

Sin embargo, el intervalo de confianza del 95% (**-3.125407, 0.00070**) para el coeficiente de AmphipodDensity incluye el valor cero, lo que sugiere que no podemos rechazar la hipótesis nula de que el coeficiente es igual a cero. Esto significa que no hay suficiente evidencia para demostrar que el coeficiente de “AmphipodDensity” es estadísticamente significativo en nuestro modelo. Sin embargo, su p-valor es de 0.0501, que está justo por encima del nivel de significación $\alpha = 0.05$. Por lo tanto, podemos interpretar que aunque no hay suficiente evidencia para rechazar la hipótesis nula, el coeficiente de AmphipodDensity podría tener un efecto significativo en la variable dependiente a un nivel de significación del $\alpha = 0.05$. Relacion negativa.

Estudiar la posible multicolinealidad del modelo con todas las regresoras calculando los VIFs.

```
library(car)
```

```
## Loading required package: carData
```

```
vif(modelo_1) #Para calcular la multicolinealidad usamos la funcion vif() del paquete "ca
r".
```

```
##   `Wave exposure`  Sandparticlesize `Beach steepness`   AmphipodDensity
##          3.771652          3.398998          1.158425          5.119632
```

Un VIF mayor que 1 indica la presencia de alguna multicolinealidad en el modelo. En general, un VIF mayor a 5 sugiere que la multicolinealidad puede estar afectando significativamente los resultados de nuestro modelo. La variable “AmphipodDensity” tiene un VIF mayor a 5, lo que indica que puede haber multicolinealidad. Es importante tener en cuenta que los VIF también pueden ser influenciados por el tamaño de la muestra y la distribución de las variables en el modelo, por lo que es necesario realizar una evaluación cuidadosa del problema.

La multicolinealidad puede provocar que los resultados del modelo pueden ser poco fiables. Para minimizar el efecto de la multicolinealidad, se pueden tomar varias acciones, como eliminar algunas de las variables

explicativas del modelo o combinarlas en una sola variable.

Considerar el modelo más reducido que no incluye las variables exposición a las olas y la pendiente de la playa y decidir si nos podemos quedar con este modelo reducido mediante un contraste de modelos con el test F para un α = 0.05. Escribir en forma paramétrica las hipótesis H0 y H1 de este contraste. Comparar el ajuste de ambos modelos.

```
modelo_2 <- lm(BeetleDensity ~  Sandparticlesize + AmphipodDensity, data=cicindela) #Ajust
amos otro modelo mas reducido que el primero modelo_1.
summary(modelo_2)
```

```
##
## Call:
## lm(formula = BeetleDensity ~ Sandparticlesize + AmphipodDensity,
##     data = cicindela)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.933 -2.226 -0.512  3.315  5.787
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       35.5651     9.4259   3.773  0.00440 **
## Sandparticlesize   3.7103     1.1215   3.308  0.00911 **
## AmphipodDensity   -2.1228     0.5167  -4.108  0.00264 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.621 on 9 degrees of freedom
## Multiple R-squared:  0.9431, Adjusted R-squared:  0.9305
## F-statistic: 74.58 on 2 and 9 DF,  p-value: 2.501e-06
```

```
anova(modelo_1, modelo_2)#Usamos la funcion anova() para comparar los dos modelos. El resu
men de esta función nos permite comparar lo modelos y ver si el modelo inicial explica mej
or la variabilidad de la densidad de escarabajos.
```

```
## Analysis of Variance Table
##
## Model 1: BeetleDensity ~ `Wave exposure` + Sandparticlesize + `Beach steepness` +
##     AmphipodDensity
## Model 2: BeetleDensity ~ Sandparticlesize + AmphipodDensity
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      7 142.59
## 2      9 192.19 -2    -49.61 1.2178 0.3517
```

La función anova() realiza un análisis de varianza (ANOVA) para comparar los modelos “modelo_1” y “modelo_2”. La hipótesis nula es que los dos modelos no son significativamente diferentes y la hipótesis alternativa es que hay una diferencia significativa entre los modelos. Podemos formularlo de la siguiente manera:

- $\mathbf{H_0}$: modelo_1 = modelo_2
- $\mathbf{H_1}$: modelo_1 ≠ modelo_2

La prueba F se basa en la comparación de la suma residual de cuadrados (RSS) de cada modelo. En este caso, el “modelo _1” tiene una RSS de 142.59 y el “modelo_2” tiene una RSS de 192.19. La diferencia en la RSS entre ambos modelos se puede utilizar para calcular el estadístico F y su correspondiente pvalor. En la tabla se puede observar que el valor de F es 1.2178 y su pvalor es 0.3517. Como el valorp es mayor que el nivel de significanción ($\alpha = 0.05$), no podemos rechazar la hipótesis nula y podemos concluir que no hay evidencia suficiente para afirmar que el “modelo 1” es significativamente mejor que el “modelo_2”.

Calcular y dibujar una región de confianza conjunta al 95% para los parámetros asociados con Sandparticlesize y AmphipodDensity con el modelo que resulta del apartado anterior. Dibujar el origen de coordenadas. La ubicación del origen respecto a la región de confianza nos indica el resultado de una determinada prueba de hipótesis. Enunciar dicha prueba y su resultado.

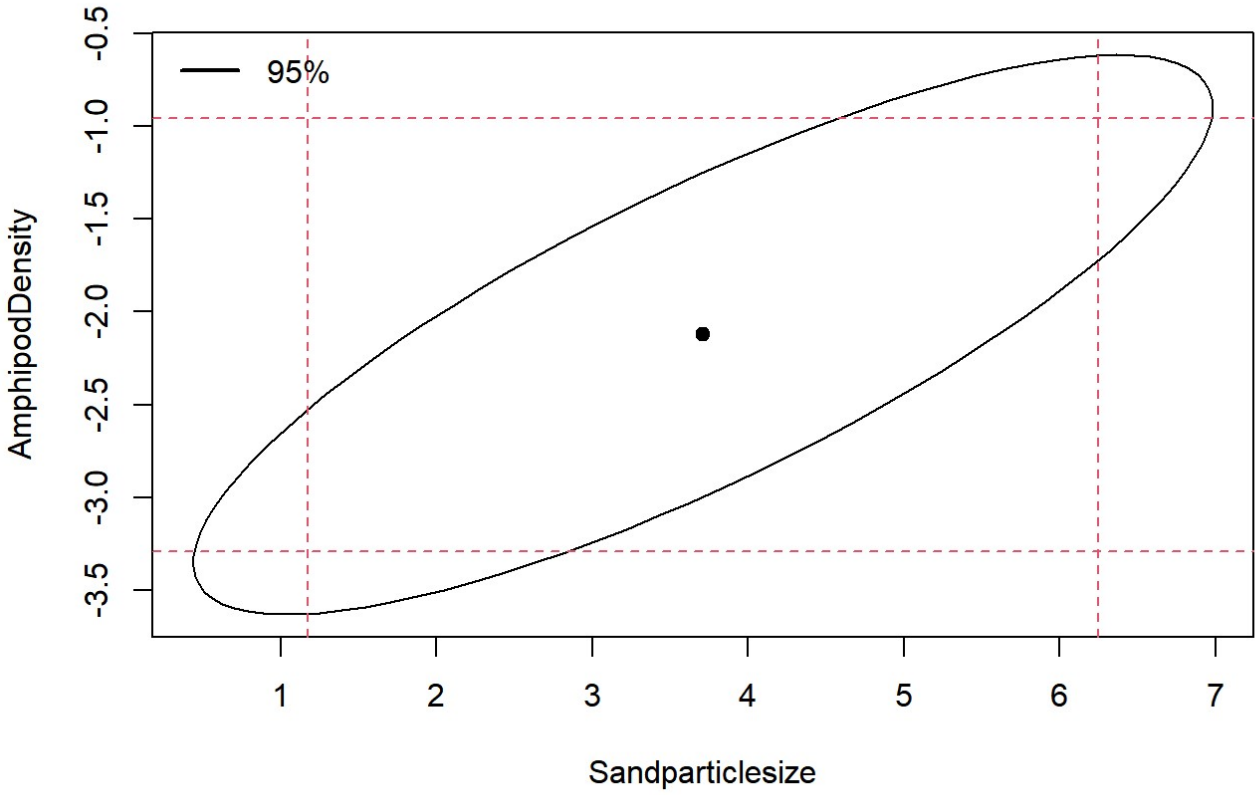
```
library(ellipse)#Cargamos la funcion ellipse() del paquete car().
```

```
##
## Attaching package: 'ellipse'
```

```
## The following object is masked from 'package:car':
##
##     ellipse
```

```
## The following object is masked from 'package:graphics':
##
##     pairs
```

```
plot(ellipse(modelo_2,2:3),type="l") #Dibujamos una elipse que representa el intervalo de
confianza del 95% de las variables "Sandparticlesize" (2) y "AmphipodDensity" (3) del mode
lo_2.
points(coef(modelo_2)[2], coef(modelo_2)[3], pch=19)#Agregamos un punto en el gráfico que
representa las estimaciones puntuales de los coeficientes de regresión de las variables 2
y 3.
abline(v=confint(modelo_2, level = 0.95)[2,],lty=2,col=2)#Agregamos una línea vertical que
representa los límites superior e inferior del intervalo de confianza del 95% de la variab
le predictora 2.
abline(h=confint(modelo_2, level = 0.95)[3,],lty=2,col=2)#Agregamos una línea vertical que
representa los límites superior e inferior del intervalo de confianza del 95% de la variab
le predictora 3.
points(0, 0, col = "red", pch = 19) #Agregamos un punto rojo en el origen de coordenadas
()
legend('topleft', col=c('black'), lwd=2, bty='n', legend=c('95%'))
```



La región de confianza no incluye al origen, lo que indica que ambas variables son significativas para predecir la densidad de escarabajos. Esto coincide con los resultados del "modelo_2", donde ambas variables tienen un pvalor significativo menor a α =0.05. La prueba de hipótesis correspondiente a cada variable en el modelo lineal es una prueba t individual con la hipótesis nula de que el coeficiente de regresión es igual a cero.

En nuestro modelo, las hipótesis nulas son:

H_0: El coeficiente de regresión de "Sandparticlesize" y el coeficiente de regresión de "AmphipodDensity" es igual a cero. O lo que es lo mismo:

- $H_{0\text{-}\text{"Sandparticlesize"}}$: $\beta 1 = 0$
- $H_{0\text{-}\text{"AmphipodDensity"}}$: $\beta 2 = 0$

H_1: La alternativa para ambas pruebas es que los coeficientes de regresión son diferentes de cero:

- $H_{1\text{-}\text{"Sandparticlesize"}}$: $\beta 1 \neq 0$
- $H_{1\text{-}\text{"AmphipodDensity"}}$: $\beta 2 \neq 0$

Por lo tanto, podemos rechazar ambas hipótesis nulas y concluir que hay suficiente evidencia para afirmar que ambos predictores son significativamente diferentes de cero, es decir, contribuyen a la variación en la densidad de escarabajos.

Con el modelo reducido del apartado (d), predecir en forma de intervalo de confianza al 95% la densidad de los escarabajos tigre previsible para una playa cercana a un conocido hotel donde el tamaño de partícula de arena es 5 y la densidad de anfípodos depredadores es 11. Comprobar previamente que los valores observados no suponen una extrapolación.

```
range(cicindela$Sandparticlesize) #Comprobamos los valores observados en el conjunto de datos "cicindela".
```

```
## [1] 1 7
```

```
range(cicindela$AmphipodDensity) #En ambas comporbaciones vemos que los valores dados estan dentro del conjunto de datos "cicindela", lo que no supondría una extrapolación.
```

```
## [1]  5 19
```

```
densidad_ejercicio <- data.frame(Sandparticlesize = 5, AmphipodDensity = 11) #Establecemos un dataframe con los valores del ejercicio.
prediccion_escarabajos <- predict(modelo_2, densidad_ejercicio, interval = "prediction", level = 0.95) #Usamos la función predict() con el modelo_2 y el nuevo dataframe con un intervalo de confianza del 95%.
prediccion_escarabajos #Mostramos la prediccion con los datos del ejercio.
```

```
##        fit      lwr      upr
## 1 30.76569 19.29834 42.23304
```

Podemos afirmar con un nivel de confianza del 95% que la densidad de escarabajos tigre para una playa con un tamaño de partícula de arena 5 y densidad de anfípodos depredadores 11 estará entre 19 y 42 individuos aproximadamente de la especie *cicindela dorsalis dorsalis*.

Ejercicio 2

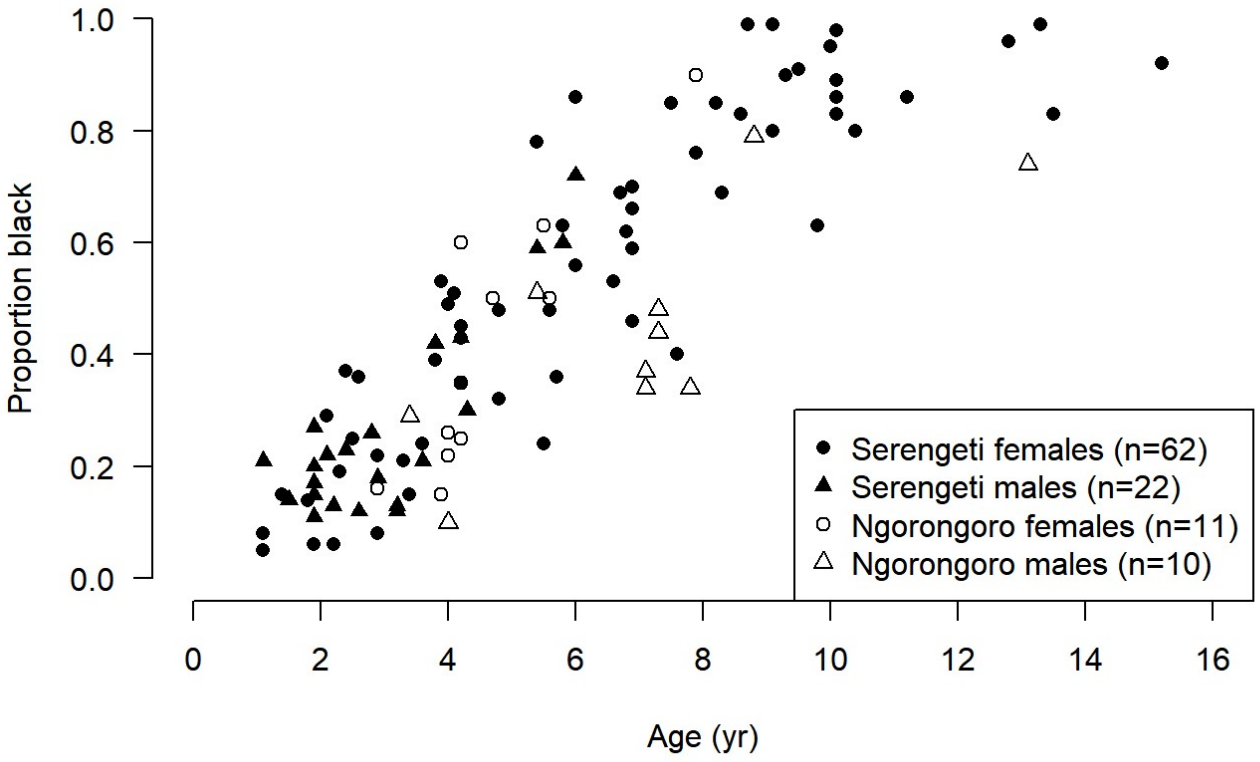
Reproducir el gráfico de dispersión de la figura 1 (figura 4d del artículo) lo más fielmente posible al original, ya que se trata de una exigencia de los editores de la revista.

```
library(readr)
lions <- read_csv("lions.csv") #Cargamos el conjunto de datos para el ejercicio 2.
```

```
## Rows: 105 Columns: 4
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): sex, area
## dbl (2): prop.black, age
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
pch_vals <- ifelse(lions$area == "S" & lions$sex == "M", 17,
                   ifelse(lions$area == "S" & lions$sex == "F", 16,
                   ifelse(lions$area == "N" & lions$sex == "M", 2, 1))) #Creamos un vector
para los valores pch que vamos autilizar para representar y diferenciar a los leones.

#Creamos el grafico de dispersion lo mas paracido a la figura del articulo adjunto.
plot(prop.black ~ age, data = lions, xlim = c(0,16), ylim = c(0, 1), axes= FALSE, pch = pc
h_vals, col = "black", xlab = "Age (yr)", ylab = "Proportion black")
axis(side = 1, at = c(0,2,4,6,8,10,12,14,16)) #Creamos el eje de x, con valores de 0 al 16
en intervalos de 2.
axis(side = 2, at = c(0,0.2,0.4,0.6,0.8,1), las=1) #Creamos el eje y, con valores del 0 al
1 en intervalos de 2. Incluimos la caracteristica "las=1" para poner en horizontal los val
ores.
#Incluimos una leyenda en la esquina inferior derecha para los datos que se representan en
el gráfico.
legend("bottomright", legend = c("Serengeti females (n=62)", "Serengeti males (n=22)", "Ng
orongoro females (n=11)", "Ngorongoro males (n=10)"), col = c("black"), pch = c(16, 17, 1,
2))
```



En el artículo se destacan los siguientes resultados:

"After controlling for age, there was no effect of sex on nose colour in the Serengeti, but Ngorongoro males had lighter noses than Ngorongoro females."

Ajustar un primer modelo sin considerar la posible interacción entre el sexo y las áreas y contrastar si el sexo es significativo en el modelo así ajustado y en los modelos separados según el área.

```
modelo_lions <- lm(prop.black ~ age + sex + area, data = lions) #Modelo sin tener en cuent
a la interacion del sexo y la areas. El pvalor de la variable sex = 0.0279, menor a 0.05,
lo que implica que es significativo.
summary(modelo_lions)
```

```
##
## Call:
## lm(formula = prop.black ~ age + sex + area, data = lions)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30265 -0.09116  0.00592  0.10049  0.32242
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.023324   0.044314   0.526   0.5998
## age          0.074464   0.004396  16.939   <2e-16 ***
## sexM        -0.068416   0.030662  -2.231   0.0279 *
## areaS        0.067473   0.034106   1.978   0.0506 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1367 on 101 degrees of freedom
## Multiple R-squared:  0.7713, Adjusted R-squared:  0.7645
## F-statistic: 113.5 on 3 and 101 DF,  p-value: < 2.2e-16
```

```
#Modelos separados segun el area:

modelo_areaS <- lm(prop.black ~ age + sex, data = subset(lions, area=="S")) #El efecto del
sexo no es significativo pvalor = 0.4098
modelo_areaN <- lm(prop.black ~ age + sex, data = subset(lions, area=="N")) #El efecto del
sexo es significativo pvalor = 0.047776

summary(modelo_areaS)
```

```
##
## Call:
## lm(formula = prop.black ~ age + sex, data = subset(lions, area ==
##     "S"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32208 -0.08310  0.00054  0.09561  0.33087
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.064161   0.034787   1.844   0.0688 .
## age          0.077495   0.004805  16.127   <2e-16 ***
## sexM        -0.030123   0.036358  -0.829   0.4098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1316 on 81 degrees of freedom
## Multiple R-squared:  0.8065, Adjusted R-squared:  0.8017
## F-statistic: 168.8 on 2 and 81 DF,  p-value: < 2.2e-16
```

```
summary(modelo_areaN)
```

```
##
## Call:
## lm(formula = prop.black ~ age + sex, data = subset(lions, area ==
##     "N"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20193 -0.11281 -0.02567  0.14511  0.23160
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.04337    0.09071   0.478 0.638321
## age          0.07912    0.01681   4.707 0.000176 ***
## sexM        -0.16748    0.07885  -2.124 0.047776 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1531 on 18 degrees of freedom
## Multiple R-squared:  0.5538, Adjusted R-squared:  0.5042
## F-statistic: 11.17 on 2 and 18 DF,  p-value: 0.000701
```

Podemos observar que para el área S, la variable sexo no es significativa (pvalor = 0.4098), mientras que para el área N sí lo es (pvalor = 0.0478). Esto sugiere que el efecto del sexo en la proporción de manchas negras puede depender del área en la que se encuentra el león. Por lo tanto, se puede concluir que la influencia del sexo en la proporción de manchas negras depende del área en la que se encuentran los leones y la interacción entre el sexo y el área puede ser un factor importante a considerar en el análisis.

Otro resultado destacado es que para los machos hay diferencias según el área. Contrastar este resultado y dibujar las rectas de regresión para las dos áreas que se obtienen del modelo.

```
#Crear un subconjunto del dataframe solo para leones machos.

lions_machos <- subset(lions, sex == "M")

#En el modelo ajustado solo para los leones machos, no incluimos la variable "sex" ni su i
nteracción con la variable "area", ya que solo estamos interesados en analizar las diferen
cias entre las áreas en los leones machos.

modelo_lionsM<- lm(prop.black ~ age + area, data = lions_machos)
summary(modelo_lionsM)
```

```
##
## Call:
## lm(formula = prop.black ~ age + area, data = lions_machos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16192 -0.08356 -0.01158  0.08842  0.22278
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.10826    0.08831  -1.226   0.2301
## age          0.07689    0.01126   6.827  1.7e-07 ***
## areaS        0.14411    0.06402   2.251   0.0321 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1162 on 29 degrees of freedom
## Multiple R-squared:  0.6798, Adjusted R-squared:  0.6577
## F-statistic: 30.78 on 2 and 29 DF,  p-value: 6.745e-08
```

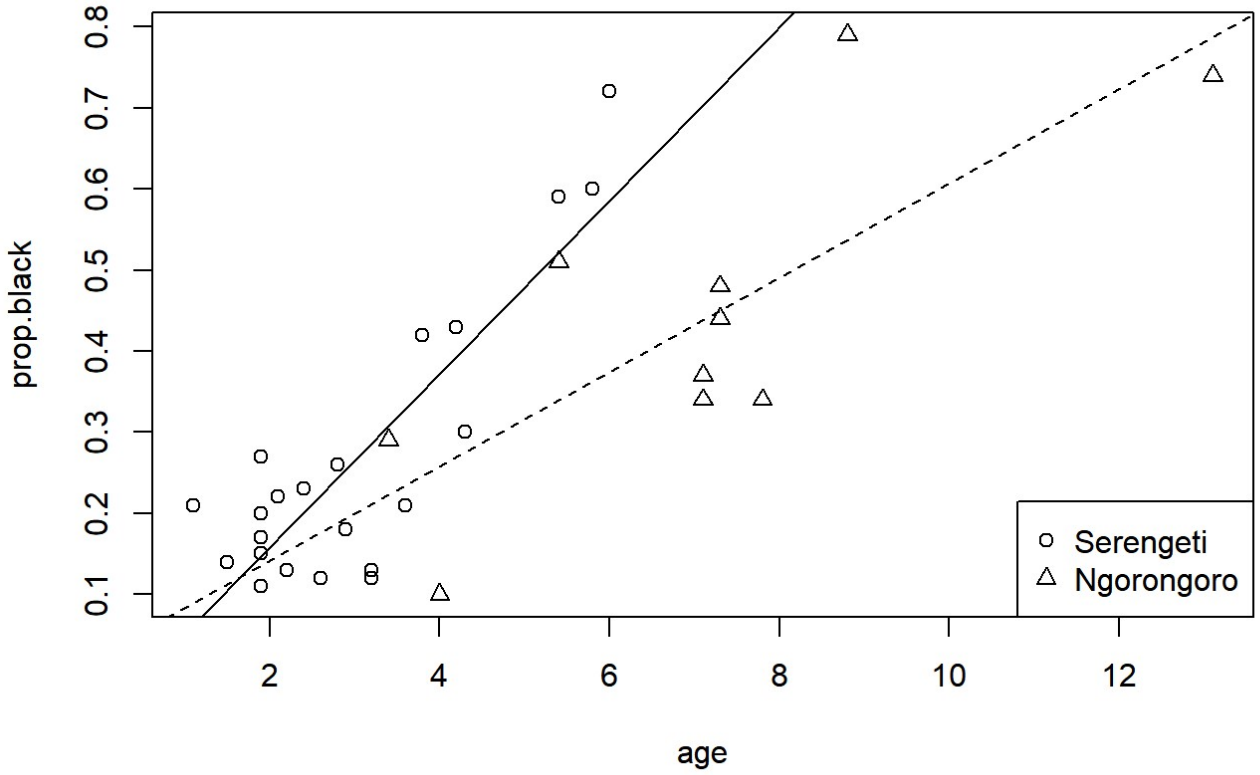
```
machos_S <- lm(prop.black ~ age, data = subset(lions_machos, area == "S"))
machos_N <- lm(prop.black ~ age, data = subset(lions_machos, area == "N"))
summary(machos_S)
```

```
##
## Call:
## lm(formula = prop.black ~ age, data = subset(lions_machos, area ==
##     "S"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16711 -0.06881  0.02520  0.05159  0.14751
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.05516    0.04828  -1.143    0.267
## age          0.10696    0.01455   7.353 4.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09283 on 20 degrees of freedom
## Multiple R-squared:   0.73,  Adjusted R-squared:  0.7165
## F-statistic: 54.06 on 1 and 20 DF,  p-value: 4.185e-07
```

```
summary(machos_N)
```

```
##
## Call:
## lm(formula = prop.black ~ age, data = subset(lions_machos, area ==
##     "N"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15771 -0.09075 -0.02880  0.05795  0.25274
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02474    0.13198   0.187   0.8560
## age          0.05824    0.01742   3.343   0.0102 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1411 on 8 degrees of freedom
## Multiple R-squared:  0.5829, Adjusted R-squared:  0.5307
## F-statistic: 11.18 on 1 and 8 DF,  p-value: 0.01018
```

```
plot(prop.black ~ age, pch = ifelse(area=="S", 1,2), data = lions_machos) #Generemos el gr
afico de dispersión segun la proporción de manchas negras y la edad.
abline(machos_S, lty=1) #Incluimos la recta de regresion del modelo machos_S para los mach
os del área S.
abline(machos_N, lty=2) #Incluimos la recta de regresion del modelo machos_N para los mach
os del área N.
legend("bottomright", legend = c("Serengeti", "Ngorongoro"), col = c("black"), pch = c(1,
2))
```



Los resultados indican que hay una diferencia significativa en la proporción de manchas negras entre las áreas S y N para los leones machos, y que la edad tiene un efecto significativo en la proporción de manchas negras en ambas áreas. Además, el impacto de la edad en la proporción de manchas negras es mayor para los leones machos en el área S en comparación con los de área N, como se puede observar en las pendiente de las rectas del gráfico anterior.

En la tabla 1 del artículo de Whitman et al. se dan los intervalos de confianza al 95 %, al 75% y al 50% para predecir la edad de una leona de 10 años o menos según su proporción de pigmentación oscura en la nariz. La primera cuestión es: ¿sirven para esto los modelos estudiados en los apartados anteriores? Reproducir la fila de la tabla 1 para una proporción del 0.50 según el modelo que proponen en el artículo. Aclarar un detalle: lo que en la tabla 1 del artículo se llama s.e., standard error ¿qué es exactamente?

Los modelos anteriores no son adecuados para predecir la edad de una leona a partir de la proporción de manchas negras. Sería necesario utilizar la variable “age” del conjunto de datos como variable respuesta y examinar el efecto que tiene la variable “prop.black” en un modelode regresión lineal, como se muestra en el artículo de Withman et al. En el modelo del artículo usa la variable “age” como variable respuesta y hace una transformación de la proporción de manchas negras “prop.black” con la función arcoseno para mejorar la simetría y la distribución de los residuos en el modelo de regresión.

A lo que denomina el árticulo “**s.e., standard error**” es la medida de la variabilidad de los residuos en el modelo o la desviación estandar de la distribución. En este caso, se utiliza para indicar la precisión de la edad de las leonas en función de la proporción de manchas negras.

```
lions_hembras <- subset(lions, sex == "F" & age <=10) #Generamos un nuevo conjunto de dato
s para separar a las hembras con edades inferiores a 10 años.
model_hembras <- lm(age ~ asin(prop.black) , data = lions_hembras) #El modelo utilizado lo
extraemos del articulo original para reproducir la tabla que nos pide el ejercicio.
summary(model_hembras)
```

```
##
## Call:
## lm(formula = age ~ asin(prop.black), data = lions_hembras)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1423 -0.9404  0.0587  0.6691  3.7460
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.0302     0.2824    7.19 1.09e-09 ***
## asin(prop.black)   5.9039     0.4400   13.42  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.224 on 61 degrees of freedom
## Multiple R-squared:  0.7469, Adjusted R-squared:  0.7428
## F-statistic:   180 on 1 and 61 DF,  p-value: < 2.2e-16
```

```
#Para cada intervalo de confianza usamos la función predict() para predecir la edad de una
leona cuando la proporción de manchas es igual a 0.5:

propporcion_manchas <- data.frame(prop.black=0.5)

#Con este código replicamos la fila de la tabla que nos pide el ejercicio:

predict(model_hembras,propporcion_manchas,interval="prediction",level=0.95)
```

```
##        fit     lwr     upr
## 1 5.121485 2.65542 7.58755
```

```
predict(model_hembras,propporcion_manchas,interval="prediction",level=0.75)
```

```
##        fit     lwr     upr
## 1 5.121485 3.68916 6.55381
```

```
predict(model_hembras,propporcion_manchas,interval="prediction",level=0.50)
```

```
##        fit      lwr      upr
## 1 5.121485 4.284674 5.958296
```

Ejercicio 3

Verificar las hipótesis de Gauss-Markov y la normalidad de los residuos del modelo completo del

apartado (b) del ejercicio 2. Realizar una completa diagnosis del modelo para ver si se cumplen las condiciones del modelo de regresión: normalidad, homocedasticidad,. . . y estudiar la presencia de valores atípicos, de alto leverage y/o puntos influyentes. Construir los gráficos correspondientes y justificar su interpretación. ¿Podemos considerar el modelo ajustado como fiable?

Para verificar las hipótesis de Gauss-Markov para el modelo propuesto, se debe ajustar el modelo a los datos y comprobar la linealidad, homocedasticidad, normalidad e independencia de los residuos:

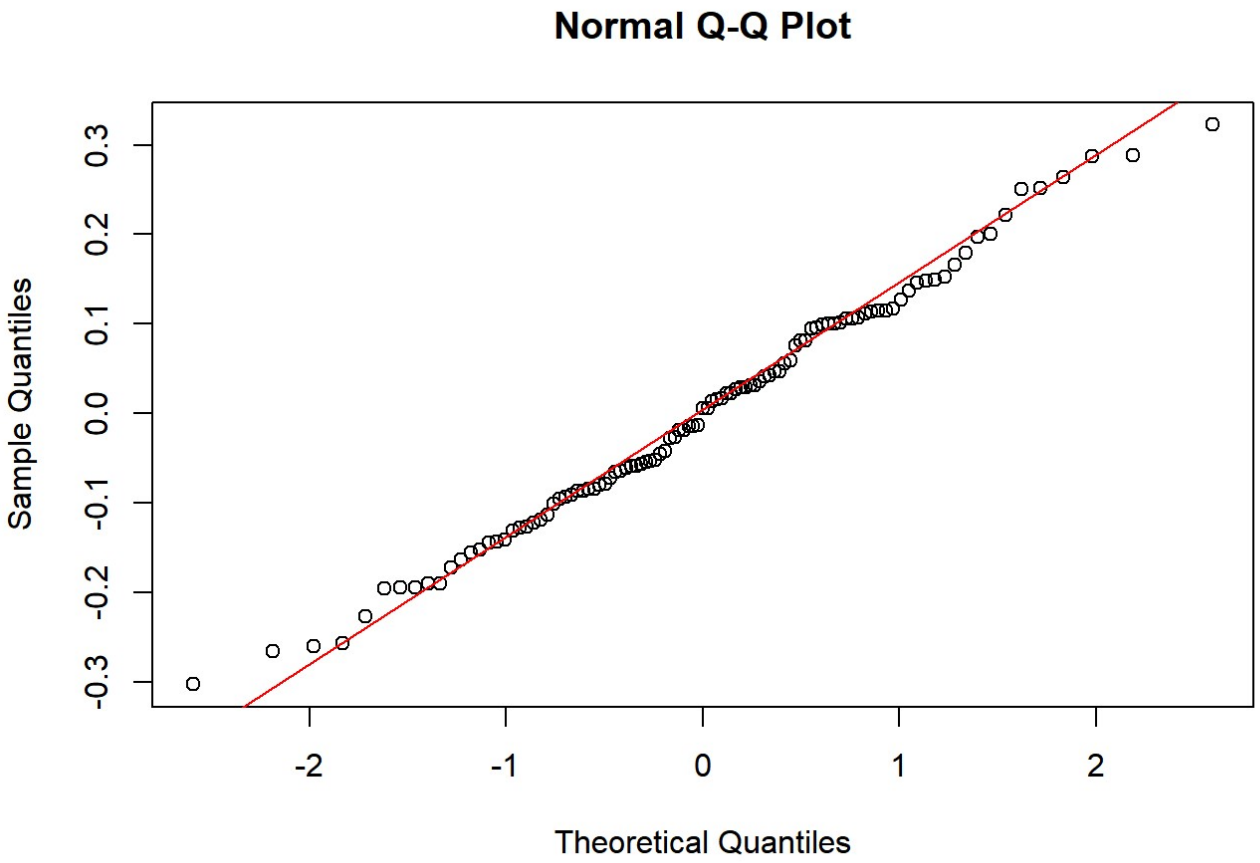
```
#Modelo completo del apartado 2(b): "modelo_lions"

residuos <- residuals(modelo_lions) #Calculamos los residuos del modelo.

#Gauss-Markov

#Primero vamos a comprobar la normalidad de los dato:

qqnorm(residuos) #Calculamos el grafico de los qq de los residuos.
qqline(residuos, col ="red") #Añadimos la linea de la distribucion de los residuos.
```




```
shapiro.test(residuos) #El test de Shaprio-Wilk no es significatico (pvalor < 0.05).
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.9909, p-value = 0.7072
```

Con el gráfico y la prueba anterior podemos asumir que el supuesto de normalidad se cumple para nuestro conjunto de datos.

```
# Comprobamos la homocedasticidad gráficamente:
library(lmtest)
```

```
## Loading required package: zoo
```

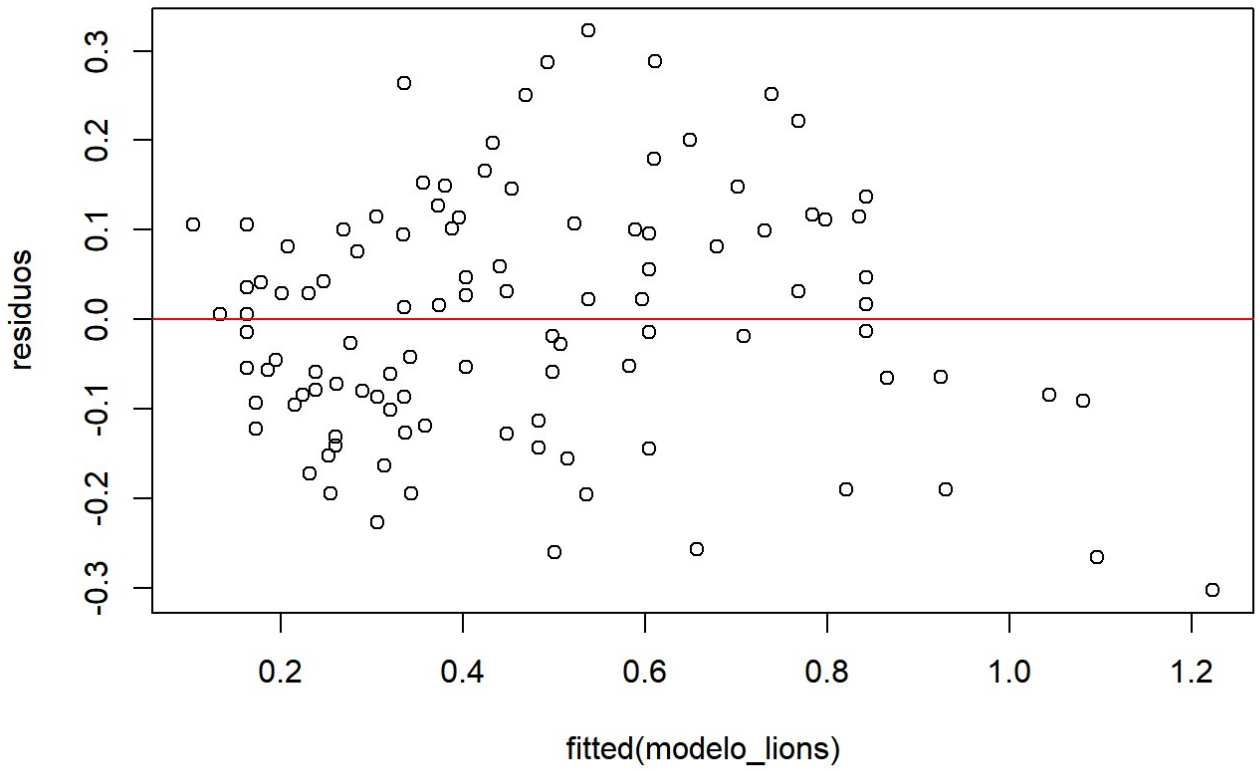
```
##
## Attaching package: 'zoo'
```

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

```
bptest(modelo_lions)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_lions
## BP = 10.171, df = 3, p-value = 0.01717
```

```
plot(fitted(modelo_lions), residuos)
abline(h=0, col="red")
```

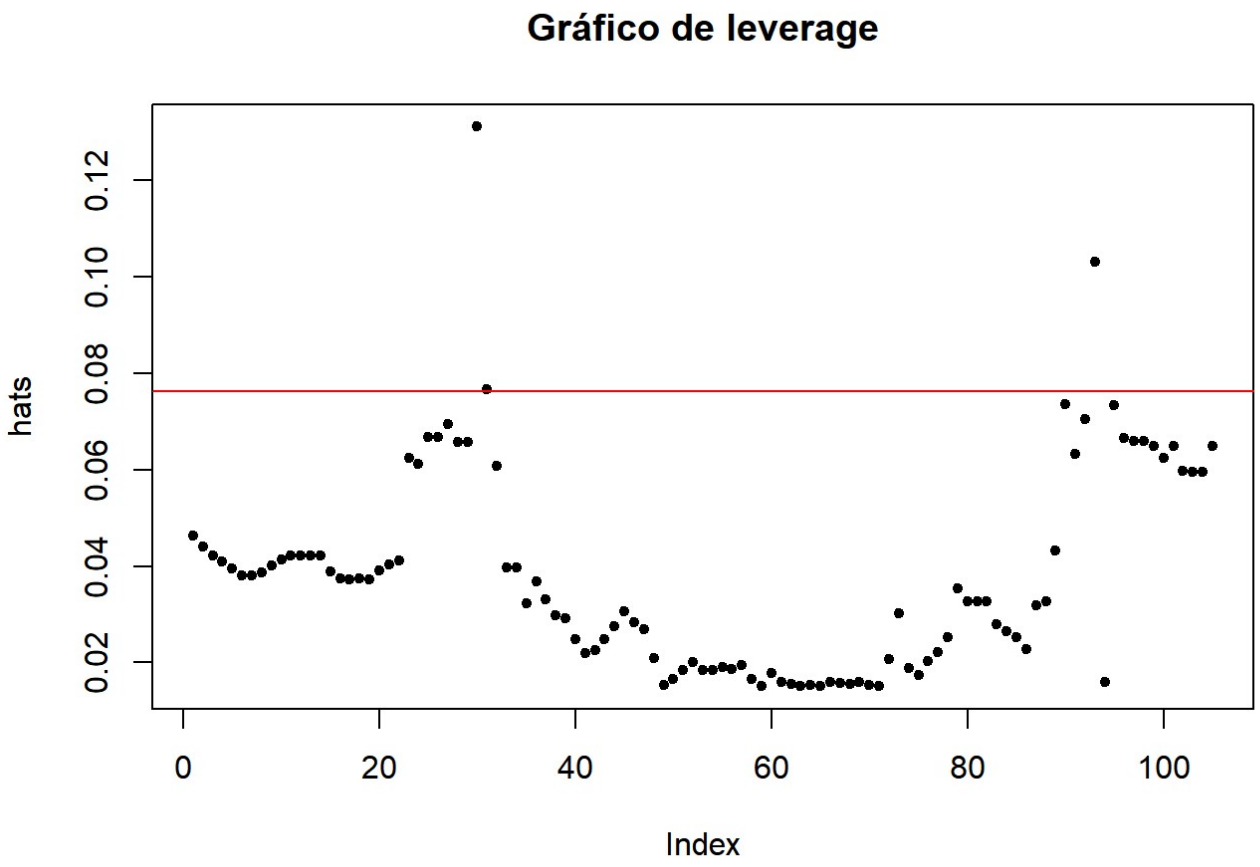


El resultado de la prueba de Breusch-Pagan indica que hay evidencia significativa de heterocedasticidad en los residuos del modelo, lo que incumple el supuesto de homocedasticidad en el modelo de regresión.

```
sum(modelo_lions$residuals) #El supuesto de la suma de residuos igual a 0 sí se cumple.
```

```
## [1] 1.19349e-15
```

```
hats <- hatvalues(modelo_lions) #Valores de leverage del modelo.
plot(hats, pch = 20, main = "Gráfico de leverage") #Gráfico de dispersión de los valores d
e leverage.
umbral <- 2 * mean(hats) #Umbral de leverage.
abline(h = umbral, col = "red")
```



Los puntos que están por encima del umbral de leverage son aquellos que tienen un valor de leverage que es mayor que dos veces el promedio de leverage de todas las observaciones. Estos puntos pueden ser considerados como observaciones con alto leverage, lo que significa que tienen un valor extremo en una o más de las variables independientes. Es importante tener en cuenta que no todos los puntos que tienen un alto leverage son puntos influyentes. Los puntos con alto leverage pueden afectar al ajuste del modelo y deben ser evaluados en conjunto con otros métodos de diagnóstico de residuos y de detección de puntos influyentes.

Si alguno de los supuestos de Gauss-Markov no se cumple, entonces el modelo ajustado no puede considerarse totalmente fiable. Cada uno de los supuestos de Gauss-Markov tiene un propósito específico en la inferencia estadística, y si uno de ellos no se cumple, la inferencia basada en el modelo ajustado puede ser erronea. En nuestro caso el supuesto de homocedasticidad no se cumple.

Teniendo en cuenta que la variable respuesta de la regresión del apartado (b) del ejercicio 2 es una proporción, ¿presenta algún problema este modelo? ¿Qué alternativas nos podemos plantear para mejorar el ajuste de los datos? Atendiendo a la naturaleza de la variable respuesta, ¿hay alguna transformación adecuada?

Cuando una varaible respuesta es una proporción puede presetnar un problema ya que la proporción toma valores en el rango 0 y 1 y esto puede alterar las suposiciones de normalidad y homocedasticidad del modelo.

Una posible alternativa para solucionar los problemas que nos puede causar una proporción en nuestro modelo seria la transformación de la variable proporción a otra escala mas adecuada usando la transformación de la variable proporción mediante la función arcoseno. Otras altenativas serian utilizar modelos no lineales, modelos de regresión específicas para manejar proporciones o considerar el estudio de otra variable.

Para nuestro caso lo opción más adecuada sería la transformación arcoseno que transforma los valores de

la proporción a un rango “-π/2, π/2”, lo que puede ayudar a cumplir las suposiciones del modelo lineal, como la normalidad y la homocedasticidad.

Aplicar la transformación más adecuada a la variable respuesta del modelo considerado. Comparar los dos modelos: con y sin la transformación. ¿Qué modelo es mejor? Justificar la respuesta.

```
summary(modelo_lions) #Modelo sin transformacion.
```

```
##
## Call:
## lm(formula = prop.black ~ age + sex + area, data = lions)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30265 -0.09116  0.00592  0.10049  0.32242
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.023324   0.044314   0.526   0.5998
## age          0.074464   0.004396  16.939   <2e-16 ***
## sexM        -0.068416   0.030662  -2.231   0.0279 *
## areaS        0.067473   0.034106   1.978   0.0506 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1367 on 101 degrees of freedom
## Multiple R-squared:  0.7713, Adjusted R-squared:  0.7645
## F-statistic: 113.5 on 3 and 101 DF,  p-value: < 2.2e-16
```

```
modelo_lions_t <- lm(formula = asin(sqrt(prop.black)) ~ age + sex + area, data = lions)#Ap
licamos la transformación en la proporción con la función asin(sqrt(prop.blac)).
summary(modelo_lions_t)
```

```
##
## Call:
## lm(formula = asin(sqrt(prop.black)) ~ age + sex + area, data = lions)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34496 -0.09694  0.00157  0.11161  0.40186
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.238919   0.051858   4.607 1.19e-05 ***
## age          0.086189   0.005144  16.755  < 2e-16 ***
## sexM        -0.074900   0.035882  -2.087   0.0394 *
## areaS        0.079999   0.039912   2.004   0.0477 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.16 on 101 degrees of freedom
## Multiple R-squared:  0.7665, Adjusted R-squared:  0.7596
## F-statistic: 110.5 on 3 and 101 DF,  p-value: < 2.2e-16
```

Para comparar los modelos “modelo_lions” y “modelo_lions_t”, podemos utilizar diferentes criterios de ajuste y evaluación de modelos. Como los más comunes: R-cuadrado ajustado, el (Akaike’s Information Criterion) y el BIC (Bayesian Information Criterion). En general, un modelo se considera mejor que otro si tiene un valor más alto de R-cuadrado ajustado y un valor más bajo de AIC y BIC.

```
#Valores de R-cuadrado ajustado:
r2_modelo_lions <- summary(modelo_lions)$adj.r.squared
r2_modelo_lions_t <- summary(modelo_lions_t)$adj.r.squared

#Guardamos los valores de AIC y BIC:
AIC1 <- AIC(modelo_lions)
AIC2 <- AIC(modelo_lions_t)
BIC1 <- BIC(modelo_lions)
BIC2 <- BIC(modelo_lions_t)

#Comparamos los valores de R2, BIC y AIC de los dos modelos:
comparacion <- data.frame(Modelo = c("modelo_lions", "modelo_lions_t"),r2 = c(r2_modelo_li
ons, r2_modelo_lions_t),AIC = c(AIC1, AIC2), BIC = c(BIC1, BIC2))
print(comparacion)
```

```
##           Modelo        r2        AIC        BIC
## 1   modelo_lions 0.7644874 -114.00625 -100.73645
## 2 modelo_lions_t 0.7595642  -80.99363  -67.72383
```

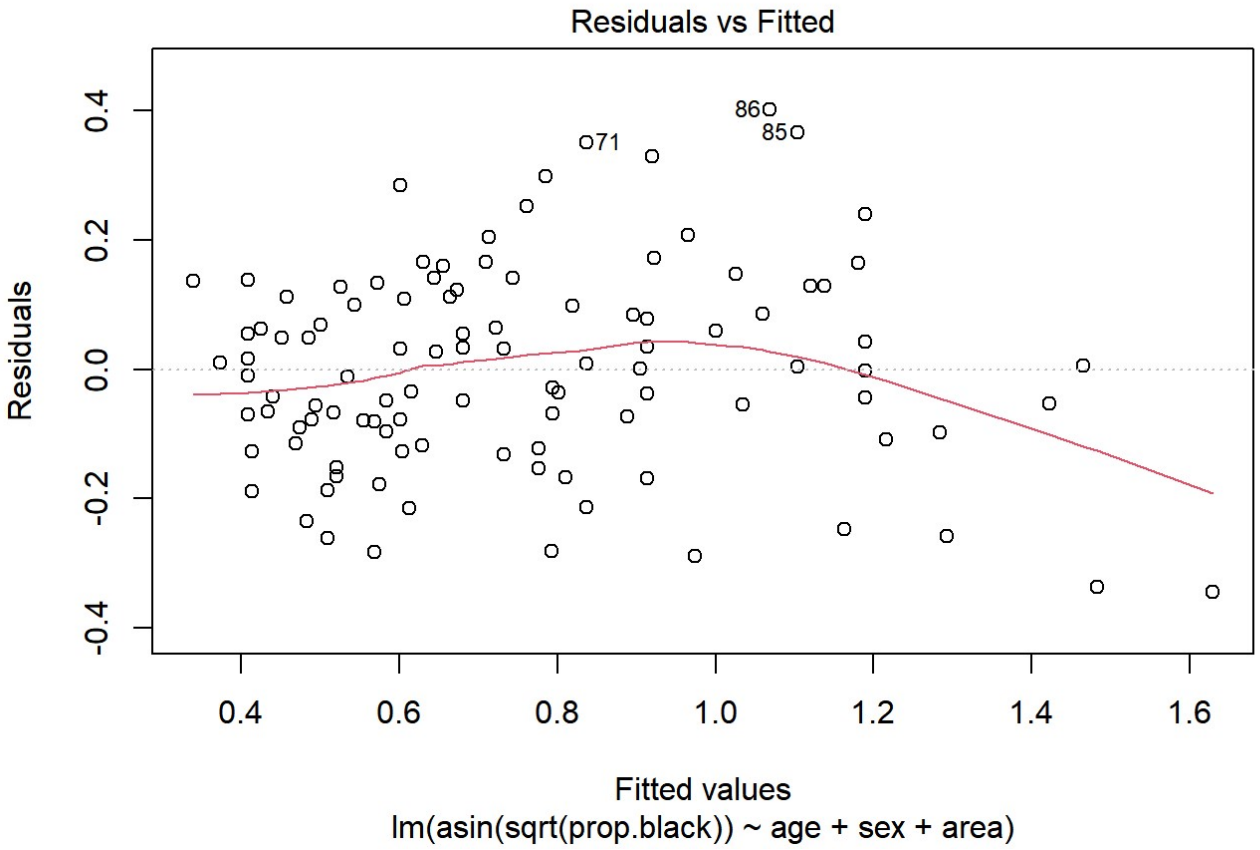
En conclusión, el modelo “modelo_lions_t” parece ser mejor que el modelo “modelo_lions” ya que tiene valores más bajos de AIC y BIC y proporciona una buena calidad de ajuste, aunque la diferencia en los valores de AIC y BIC no es muy grande.

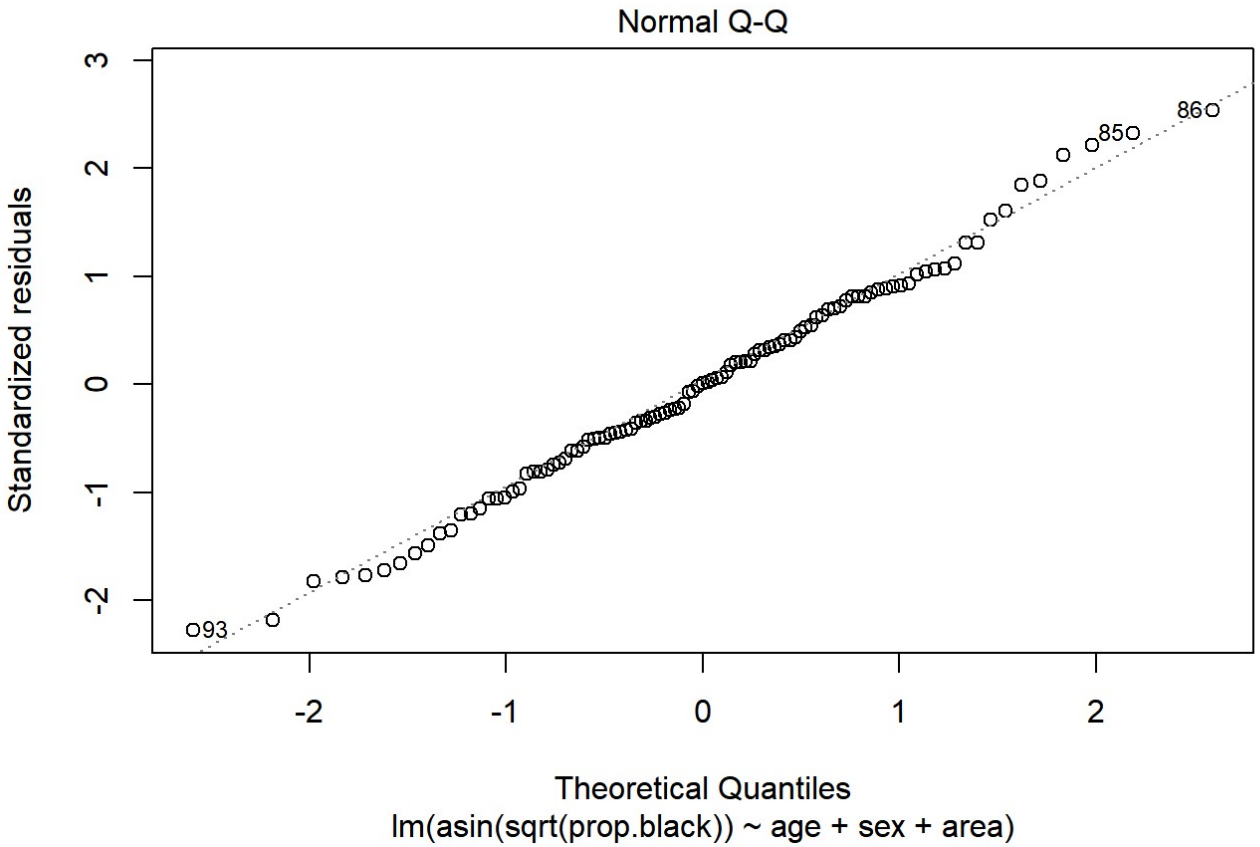
Realizar una rápida diagnosis del modelo transformado. ¿Estamos satisfechos con este nuevo modelo? ¿Qué otro ajuste nos podemos plantear para mejorar el modelo?

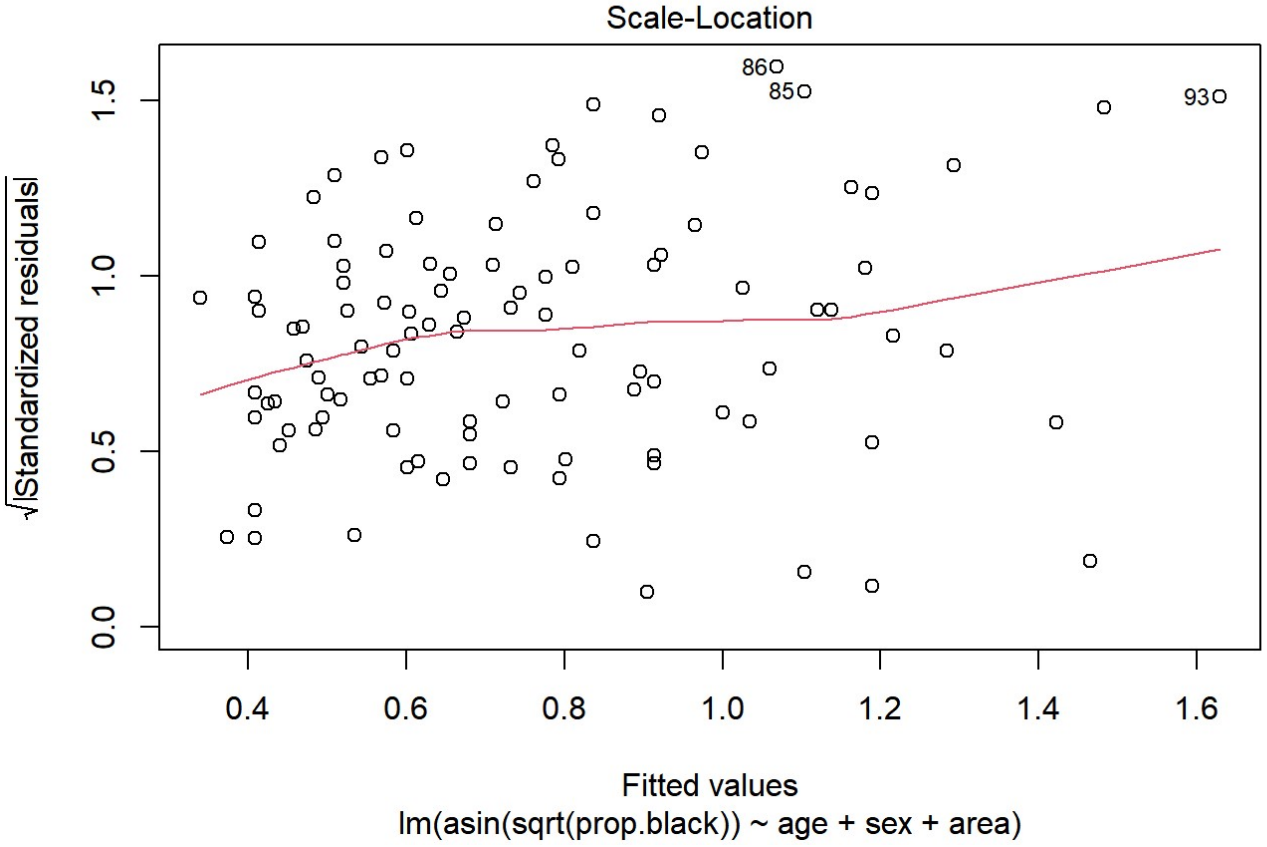
Para realizar un analisi rapido del modelo ajustado vamso a presentar la graficas para ver los supuestos de

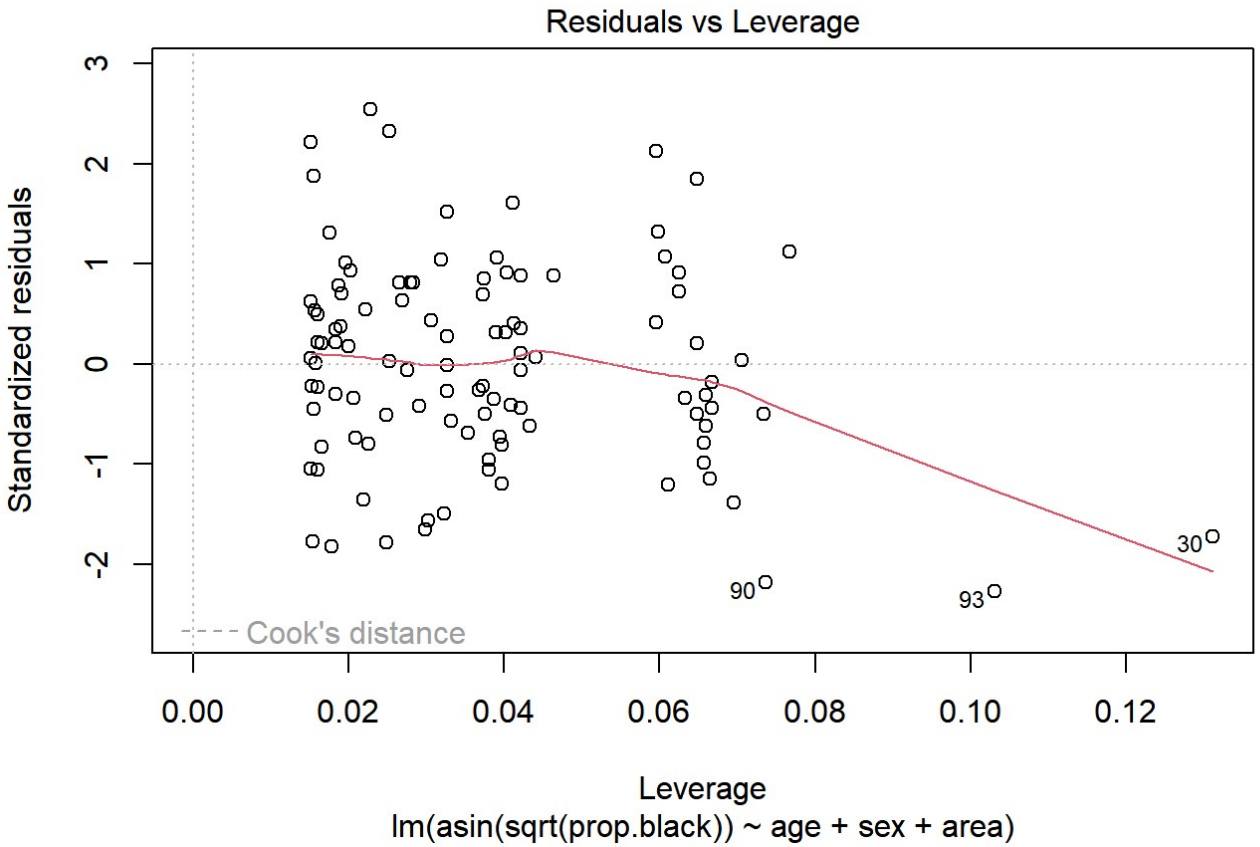
Guass-Markov:

```
plot(modelo_lions_t)
```




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(asin(sqrt(prop.black)) ~ age + sex + area)
93
85
86


Scale-Location
√|Standardized residuals|
Fitted values
lm(asin(sqrt(prop.black)) ~ age + sex + area)
86
85
93



Como hemos visto anteriormen, el modelo es muy similar al modelo original fallando tambien en algunos de los supuestos de Gauss-Markov (supuesto de homocedasticidad, como se aprecia en la grafica “Residual vs Fitted”).

Para corregir el modelo podríamos plantear utilizar un modelo de regresión lineal generalizada si se detecta heterocedasticidad. También podríamos considerar incluir interacciones entre las variables predictoras si hay alguna evidencia de que las interacciones son importantes en la relación entre la variable dependiente y las variables predictoras.

Discutir la utilización de la transformación arcoseno en el modelo del apartado (d) del ejercicio 2.

La transformación arcoseno se utiliza en este modelo para ajustar una relación lineal entre la variable independiente transformada y la variable dependiente. Esto puede ser útil cuando la relación entrelas variables no es lineal y ayuda a hacer que la interpretación del modelo sea más manejable.