Regresión, modelos y métodos Prueba de evaluación continua 1

Bas_Magí_Catasus

2023-05-07

Ejercicio 1

Un grupo de científicos norteamericanos están interesados en encontrar un hábitat adecuado para reintroducir una especie rara de escara- bajos tigre, llamada cicindela dorsalis dorsalis, los cuales viven en playas de arena de la costa del Atlántico Norte. Se muestrearon 12 playas y se midió la densidad de estos escarabajos tigre. Adicionalmente se midieron una serie de facto- res bióticos y abióticos tales como la exposición a las olas, tamaño de la partícula de arena, pen- diente de la playa y densidad de los anfípodos depredadores.

```
library(readxl)
cicindela= read_excel("cicindela.xlsx")
```

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

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

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

```
names(cicindela)
```

```
## [1] "BeetleDensity"    "Wave exposure"    "Sandparticlesize" "Beach steepness"
## [5] "AmphipodDensity"
```

Para ajustar el modelo de regresión lineal cogeremos la varaible “BeetleDesity” como variable dependiente (y).

- BeetleDensity: la densidad de escarabajos tigre (variable dependiente).
- Wave exposure: la exposición a las olas.
- Sandparticlesize: el tamaño de la partícula de arena.
- Beach steepness: la pendiente de la playa.
- AmphipodDensity: la densidad de los anfípodos depredadores.

```
cici_model = lm(BeetleDensity ~ `Wave exposure` + Sandparticlesize + `Beach steepness
` + AmphipodDensity, data = cicindela)
summary(cici_model)
```

```
##
## 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
```

El modelo lineal planteado parece ser significativo según el valor del F-estadísitico, el cual compara la varianza explicada vs la no explicada.

H0: Coeficientes de regresión = 0; el modelo no predictivo

H1: Coeficientes de regresión != 0; el modelo es predictivo

El valor del estadístico F es 39.71 con 4 y 7 grados de libertad, y el p-valor es 6.727e-05, lo que indica que rechazamos la hipótesis nula y aceptamos la hipótesis alternativa. Concluimos que al menos una de las variables independientes tiene un efecto significativo sobre la variable dependiente, y que el modelo de regresión lineal múltiple proporciona un ajuste significativo a los datos.

H1: 🟢

(b) 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 p-valor para la densidad de los anfípodos depredadores en el resumen del modelo de regresión? ¿Qué interpretación práctica tiene este parámetro β4?

```
confint(cici_model, level = 0.9)
```

```
##                          5 %       95 %
## (Intercept)       -17.7588417 47.6650535
## `Wave exposure`    -1.1594063  2.9840266
## Sandparticlesize    1.6823301  6.1117347
## `Beach steepness`  -0.2070857  1.5093046
## AmphipodDensity    -2.8146991 -0.3100058
```

```
confint(cici_model, level = 0.95)
```

```
##                         2.5 %       97.5 %
## (Intercept)       -25.8746879 5.578090e+01
## `Wave exposure`    -1.6733999 3.498020e+00
## Sandparticlesize    1.1328616 6.661203e+00
## `Beach steepness`  -0.4200042 1.722223e+00
## AmphipodDensity    -3.1254068 7.019125e-04
```

Si solo tuviéramos esta información, no podríamos deducir el p-valor exacto para AmphipodDensity, pero podríamos decir que el intervalo de confianza al 90% no incluye el valor 0, lo que sugiere que el coeficiente es significativamente diferente de cero a un nivel de significación del 10%. Sin embargo, el intervalo de confianza al 95% incluye el valor 0, lo que sugiere que no hay evidencia suficiente para rechazar la hipótesis nula de que el coeficiente es igual a cero a un nivel de significación del 5%.

El parámetro β4, que acompaña a la variable AmphipodDensity, representa la relación entre la densidad de los anfípodos depredadores y la densidad de escarabajos, después de tener en cuenta los efectos de las otras variables en el modelo. Por lo tanto, un valor negativo de este parámetro sugiere que a medida que aumenta la densidad de los anfípodos depredadores, disminuye la densidad de los escarabajos.

(c) Estudiar la posible multicolinealidad del modelo con todas las regresoras calculando los VIFs

```
library(DescTools)
VIF(cici_model)
```

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

Por lo general, podemos decir que con valores de VIF < 5 no existen problemas de multicolinealidad. En este caso solo tenemos un valor no muy superior a 5 con la variable AmphipodDensity. En este caso, el valor VIF para la variable **`AmphipodDensity`** es cercano a 5, lo que podría indicar una ligera multicolinealidad, pero aún está dentro de un rango aceptable. Habrá que tener esto en cuenta a la hora de sacar conclusiones de los resultados.

(d) 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

```
cici_reducido = lm(cicindela$BeetleDensity ~ cicindela$Sandparticlesize + cicindela$A
mphipodDensity)
summary(cici_reducido)
```

```
##
## Call:
## lm(formula = cicindela$BeetleDensity ~ cicindela$Sandparticlesize +
##     cicindela$AmphipodDensity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.933 -2.226 -0.512  3.315  5.787
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  35.5651     9.4259   3.773  0.00440 **
## cicindela$Sandparticlesize    3.7103     1.1215   3.308  0.00911 **
## cicindela$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
```

- H0: Los coeficientes de las variables eliminadas son iguales a cero, es decir, el modelo reducido no es significativamente peor que el modelo completo.
- H1: Al menos uno de los coeficientes de las variables eliminadas no es igual a cero, es decir, el modelo completo es significativamente mejor que el modelo reducido.

```
# Comparación de los modelos
anova(cici_reducido, cici_model)
```

```
## Warning in anova.lmlist(object, ...): models with response '"BeetleDensity"'
## removed because response differs from model 1
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
cicindela$Sandparticlesize	1	2825.0548	2825.0548	132.29007	1.104099e-06
cicindela$AmphipodDensity	1	360.4168	360.4168	16.87739	2.644068e-03
Residuals	9	192.1950	21.3550	*NA*	*NA*

3 rows

H0: 🟢

La suma de cuadrados de los residuos en el modelo reducido es de 192.20, y los grados de libertad son 9. Esto significa que el modelo reducido explica el 94.31% de la varianza de la variable respuesta, y que la variable “Sandparticlesize” y “AmphipodDensity” son suficientes para explicar la mayor parte de la variabilidad de “BeetleDensity”.

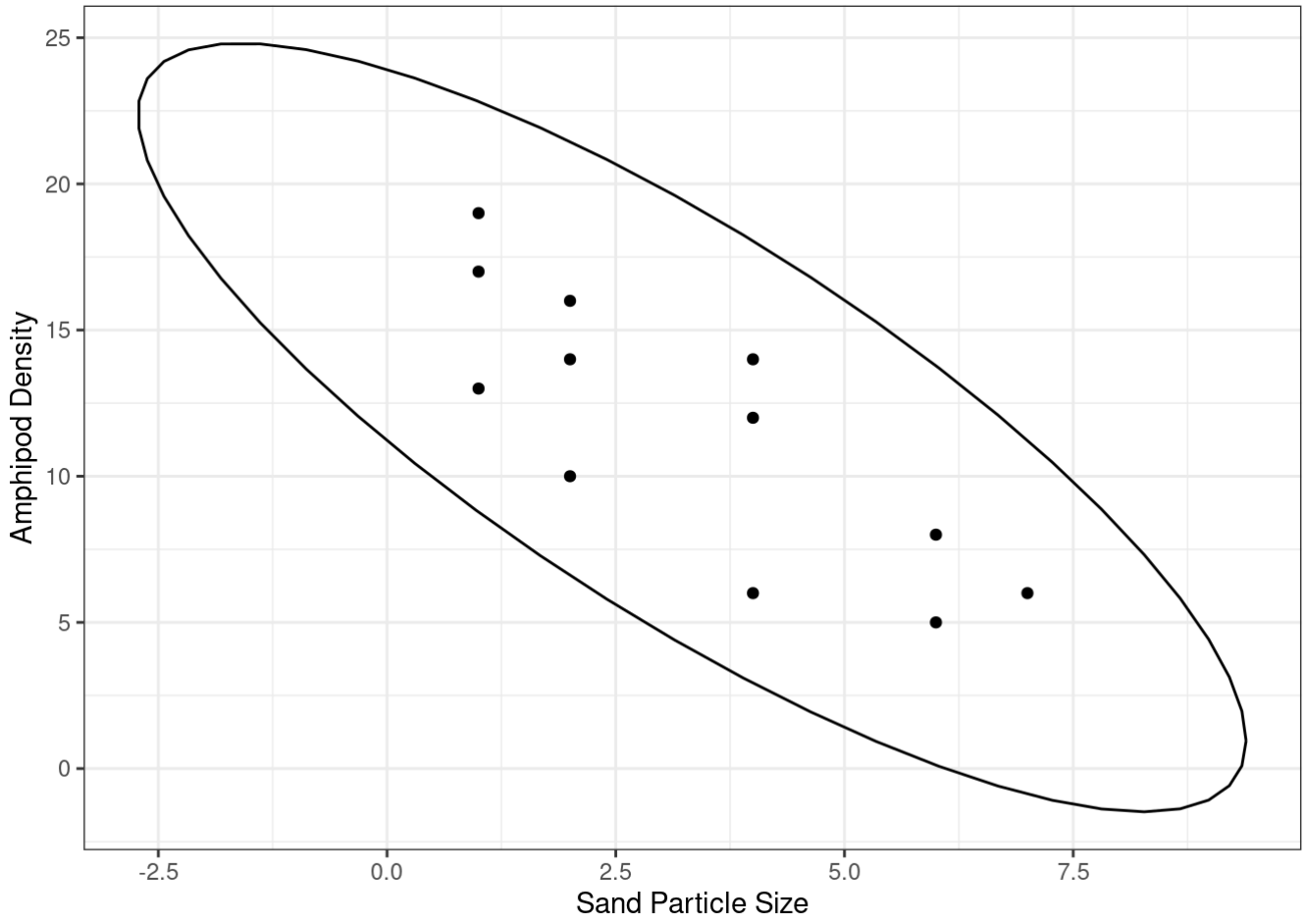
(e) Calcular y dibujar una región de confianza conjunta al 95 % para los parámetros asociados con Sandparticlesize y AmphipodDensity con el modelo que resulta del apartado anterior

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.

```
confint(cici_reducido)
```

```
##                                2.5 %     97.5 %
## (Intercept)                14.242176 56.8881193
## cicindela$Sandparticlesize  1.173317  6.2472637
## cicindela$AmphipodDensity  -3.291720 -0.9538996
```

```
library(ggplot2)
ggplot(cicindela, aes(x = Sandparticlesize, y = AmphipodDensity)) +
  geom_point() +
  stat_ellipse(type = "norm", level = 0.95) +
  scale_fill_manual(values = "blue", name = "Confidence Interval") +
  theme_bw() +
  labs(x = "Sand Particle Size", y = "Amphipod Density")
```




```
# No he podido hacerlo con ellipse() ya que parece que no funciona correctamente
```

(f) 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

```
# Verificamos los valores
summary(cicindela[c("Sandparticlesize", "AmphipodDensity")])
```

```
##  Sandparticlesize AmphipodDensity
##  Min.   :1.000    Min.   : 5.00
##  1st Qu.:1.750    1st Qu.: 7.50
##  Median :3.000    Median :12.50
##  Mean   :3.333    Mean   :11.67
##  3rd Qu.:4.500    3rd Qu.:14.50
##  Max.   :7.000    Max.   :19.00
```

Los valores estan dentro del rango

```
# Valores predictores
valores <- data.frame(5, 11)

# Predicción al 95%
predict(cici_reducido, valores, interval = "confidence")
```

```
## Warning: 'newdata' had 1 row but variables found have 12 rows
```

```
##         fit       lwr       upr
## 1  13.26639  9.668872 16.863907
## 2   9.02077  4.747164 13.294376
## 3  48.80032 42.553029 55.047613
## 4  20.68697 15.497630 25.876311
## 5  40.84441 35.749469 45.939352
## 6   3.18767 -1.590426  7.965766
## 7  24.93259 21.300259 28.564921
## 8  -1.05795 -6.963513  4.847612
## 9  47.21284 41.769863 52.655818
## 10 11.67891  6.063600 17.294219
## 11 37.66945 31.539816 43.799084
## 12 21.75763 15.833829 27.681429
```

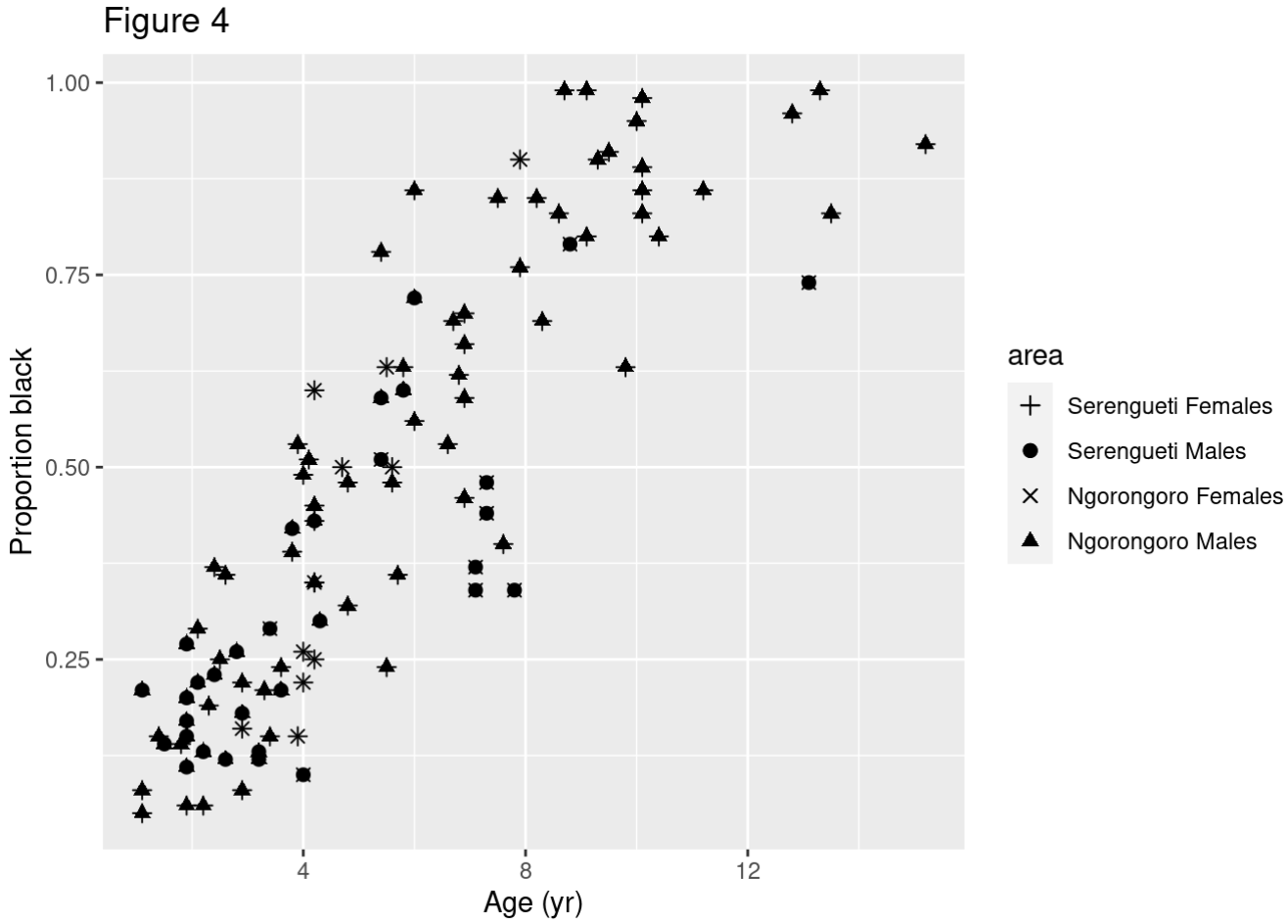
Ejercicio 2

En el trabajo de Whitman et al. (2004) se estu- dia, entre otras cosas, la relación entre la edad de los leones y la proporción oscura en la colo- ración de sus narices. En el archivo lions.csv disponemos de los datos de 105 leones machos y hembras de dos áreas de Tanzania, el parque na- cional de Serengueti y el cráter del Ngorongoro, entre 1999 y 2002. Las variables registradas son la edad conocida de cada animal y la propor- ción oscura de su nariz a partir de fotografías tratadas digitalmente (ver figura adjunta). En la figura 1 se reproduce el gráfico de disper- sión de la figura 4 del artículo con el cambio de coloración de la nariz según la edad de machos y hembras en las dos poblaciones separadas. Nota: Los datos se han extraído principalmente del gráfico del artículo de Whitman et al. (2004) y por lo tanto son aproximados. Algunos pa- quetes de R contienen un data.frame con una parte de estos datos. Por ejemplo LionNoses del paquete abd contiene los datos de todos los machos. En consecuencia, los resultados numé- ricos de vuestro análisis pueden ser ligeramente distintos a los del trabajo original.

```
# Importamos el dataset
leon = read.csv("lions.csv")
```

(a) 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

```
# Grafico
library(ggplot2)
ggplot(leon, aes(x=age, y=prop.black))+
  geom_point(aes(shape = area), size = 2) +
  geom_point(aes(shape = sex), size=2) +
  labs(title = "Figure 4",
       x = "Age (yr)", y = "Proportion black") +
  scale_shape_manual(values = c(3, 19, 4, 17),
                     labels = c("Serengueti Females", "Serengueti Males",
                                "Ngorongoro Females", "Ngorongoro Males"))
```



(b) 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

```
# Pasar a factores
lion= data.frame(prop.black = leon$prop.black,
                      age = leon$age,
                      sex = factor(leon$sex),
                      area = factor(leon$area))
```

```
# Contrastar si el sexo es significativo en el modelo así ajustado
leo_model1= lm(lion$prop.black ~ lion$age + lion$sex)
summary(leo_model1)
```

```
##
## Call:
## lm(formula = lion$prop.black ~ lion$age + lion$sex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28474 -0.11052  0.00196  0.10974  0.33251
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.085808   0.031525   2.722  0.00763 **
## lion$age     0.073614   0.004437  16.591  < 2e-16 ***
## lion$sexM   -0.080849   0.030437  -2.656  0.00917 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1386 on 102 degrees of freedom
## Multiple R-squared:  0.7624, Adjusted R-squared:  0.7578
## F-statistic: 163.7 on 2 and 102 DF,  p-value: < 2.2e-16
```

Parece que el sexo si es significativo en este modelo.

```
# Modelos separados segun el area
leo_S= subset(leon, area == 'S')
leo_N= subset(leon, area == 'N')

leo_model2.1= lm(leo_S$prop.black ~ leo_S$age * leo_S$sex)
summary(leo_model2.1)
```

```
##
## Call:
## lm(formula = leo_S$prop.black ~ leo_S$age * leo_S$sex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30712 -0.08717  0.02071  0.09153  0.33028
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.074893   0.035288   2.122   0.0369 *
## leo_S$age             0.075804   0.004905  15.454   <2e-16 ***
## leo_S$sexM           -0.130055   0.076583  -1.698   0.0934 .
## leo_S$age:leo_S$sexM  0.031156   0.021058   1.480   0.1429
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1307 on 80 degrees of freedom
## Multiple R-squared:  0.8116, Adjusted R-squared:  0.8046
## F-statistic: 114.9 on 3 and 80 DF,  p-value: < 2.2e-16
```

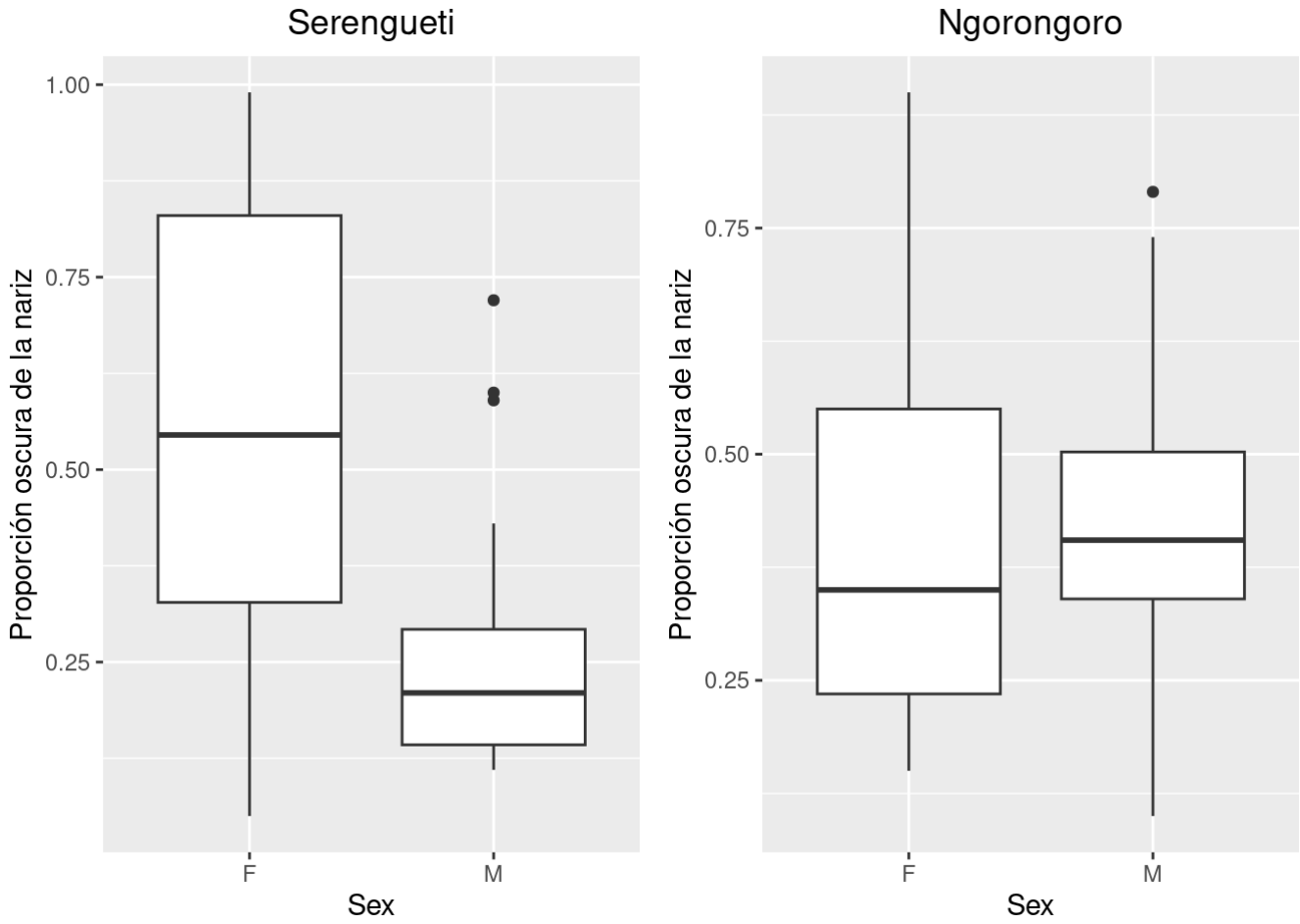
```
leo_model2.2=lm(leo_N$prop.black ~ leo_N$age * leo_N$sex)
summary(leo_model2.2)
```

```
##
## Call:
## lm(formula = leo_N$prop.black ~ leo_N$age * leo_N$sex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15771 -0.08862 -0.02669  0.06724  0.25969
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.32531    0.14975  -2.172   0.0443 *
## leo_N$age             0.15848    0.03112   5.092 9.04e-05 ***
## leo_N$sexM            0.35005    0.19249   1.819   0.0866 .
## leo_N$age:leo_N$sexM -0.10024    0.03498  -2.866   0.0107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1293 on 17 degrees of freedom
## Multiple R-squared:  0.6991, Adjusted R-squared:  0.6461
## F-statistic: 13.17 on 3 and 17 DF,  p-value: 0.000108
```

En el modelo del Serengueti existen diferencias significativa por lo que respecta a la proporción oscura de la nariz, mientras que en el modelo de Nogorongoro no. Veamoslo gráficamente:

```
model2.1= ggplot(leo_S, aes(x = sex, y = prop.black)) +
  geom_boxplot()+
  labs(title = "Serengueti",
       x = "Sex", y = "Proporción oscura de la nariz") +
  theme(plot.title = element_text(hjust = 0.5))
model2.2= ggplot(leo_N, aes(x = sex, y = prop.black)) +
  geom_boxplot()+
  labs(title = "Ngorongoro",
       x = "Sex", y = "Proporción oscura de la nariz") +
  theme(plot.title = element_text(hjust = 0.5))

library(gridExtra)
grid.arrange(model2.1, model2.2, ncol=2)
```



(c) 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

```
# Solo machos
leo_M= subset(x= leon, sex =='M')

library(tidyr)
#Filtrar Segrengueti
serengeti_males = subset(leo_M, area == 'S')

# Filtrar Nogorongoro
nogorongoro_males = subset(leo_M, area == 'N')

# Modelo lineal para Serengeti
serengeti_model= lm(prop.black ~ age, data = serengeti_males)

# Modelo lineal para Ngorongoro
ngorongoro_model= lm(prop.black ~ age, data = nogorongoro_males)


# Comparativa de los modelos
serengeti_summary = summary(serengeti_model)

ngorongoro_summary = summary(ngorongoro_model)

tabla_comparativa = data.frame(
  Area = c("Serengeti", "Ngorongoro"),
  R_squared = c(serengeti_summary$r.squared, ngorongoro_summary$r.squared),
  Std_error = c(serengeti_summary$sigma, ngorongoro_summary$sigma),
  P_value = c(serengeti_summary$coefficients[2, 4],
              ngorongoro_summary$coefficients[2, 4])
)

print(tabla_comparativa)
```

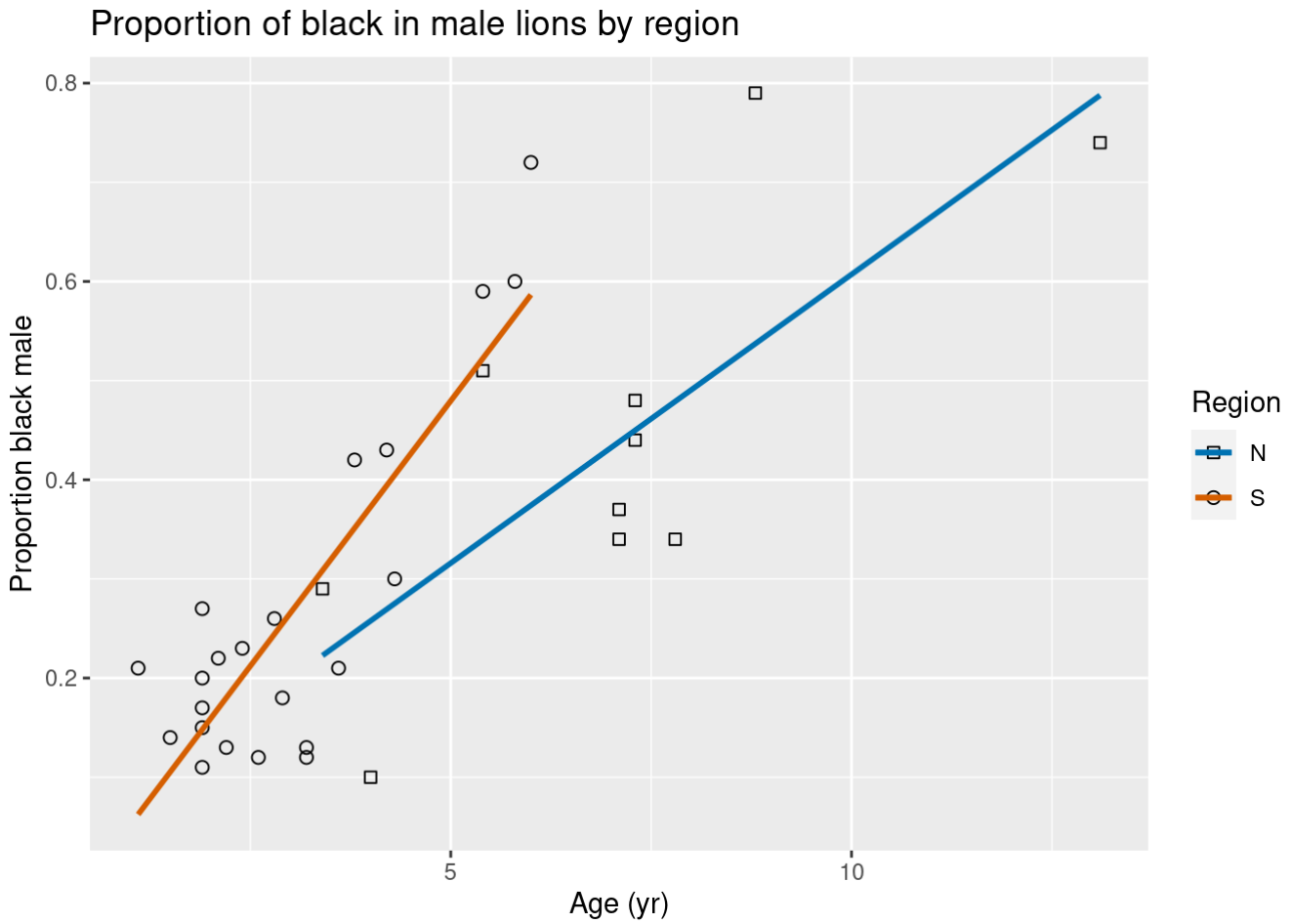
```
##         Area R_squared  Std_error      P_value
## 1  Serengeti 0.7299547 0.09283287 4.184777e-07
## 2 Ngorongoro 0.5828678 0.14113137 1.017904e-02
```

Visto los resultados, parece que existen diferencias significativas entre los machos de las distintas regiones.La R-cuadrada para Serengeti es mayor que para Ngorongoro, lo que sugiere que el modelo de regresión lineal ajusta mejor a los datos para la región de Serengeti.

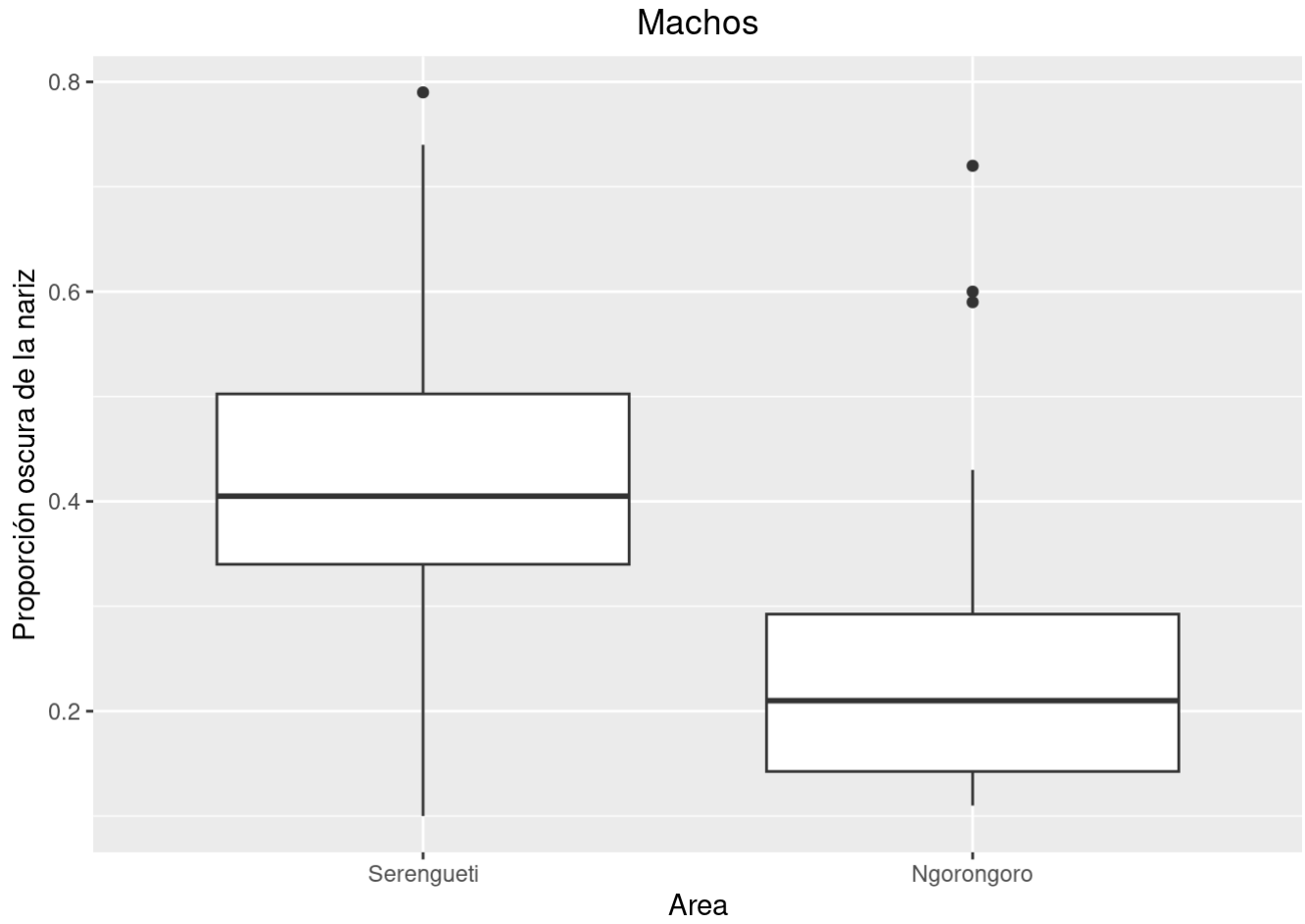
Ahora pasamos a observar las diferencias gráficamente:

```
# Rectas
library(ggplot2)

ggplot(leo_M, aes(x = age, y = prop.black, shape = area)) +
  geom_point(size = 2) +
  geom_smooth(aes(color = area), method = "lm", se = FALSE, formula = y ~ x, data = s
erengeti_males) +
  geom_smooth(aes(color = area), method = "lm", se = FALSE, formula = y ~ x, data = n
ogorongoro_males) +
  labs(title = "Proportion of black in male lions by region",
       x = "Age (yr)", y = "Proportion black male",
       shape = "Region", color = "Region") +
  scale_shape_manual(values = c(22, 21)) +
  scale_color_manual(values = c("#0072B2", "#D55E00"))
```




```
# Boxplot
ggplot(leo_M, aes(x = area, y = prop.black)) +
  geom_boxplot()+
  labs(title = "Machos",
       x = "Area", y = "Proporción oscura de la nariz") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_discrete(labels = c("Serengueti", "Ngorongoro"))
```



Los machos del Serengueti tienen, significativamente, unas narices más oscuras que los machos de Ngorongoro.

(d) 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?

No, son modelos de regresión lineal pero no relacionan directamente leonas con su pigmentación oscura de la nariz y la edad.

Reproducir la fila de la tabla 1 para una proporción del 0.50 según el modelo que proponen en el artículo

Nota: Recordemos también aquí que los resultados pueden ser ligeramente distintos a los del artículo por la utilización de datos aproximados.

```
# Modelo del articulo
intercept = 2.00667
B1 = 5.9037
x = 0.5

y = intercept + B1*asin(x)
se = 1.23

# Crear los vectores de los intervalos
intervalos_95= paste(round(y - 1.96*se, 2), round(y + 1.96*se, 2), sep=" - ")
intervalos_75 = paste(round(y - 1.15*se, 2), round(y + 1.15*se, 2), sep=" - ")
intervalos_50 = paste(round(y - 0.67*se, 2), round(y + 0.67*se, 2), sep=" - ")

# Imprimir los resultados

Table_1 = data.frame(
  'Proportion black' = x,
  'Estimated age in years (s.e.)' = y,
  '95 p.i.' = intervalos_95,
  "75 p.i." = intervalos_75,
  '50 p.i.' = intervalos_50
)

Table_1
```

Proportion.black <dbl>	Estimated.age.in.years..s.e.. <dbl>	X95.p.i. <chr>	X75.p.i. <chr>	X50.p.i. <chr>
0.5	5.09784	2.69 - 7.51	3.68 - 6.51	4.27 - 5.92

1 row

Aclarar un detalle: lo que en la tabla 1 del artículo se llama s.e., standard error ¿qué es exactamente?

El error estándar (s.e.) es una medida de la variabilidad de una estimación estadística, como la media o el coeficiente de regresión. En el contexto de la tabla del artículo que estamos analizando, el s.e. se refiere a la incertidumbre asociada con la estimación de la edad en años a partir de la proporción de negros en la nariz de las aves.

En términos más técnicos, el s.e. representa la desviación estándar de la distribución de las estimaciones obtenidas a partir de diferentes muestras de la misma población. Un s.e. más bajo indica que la estimación es más precisa, mientras que un s.e. más alto indica que la estimación es menos precisa.

Ejercicio 3

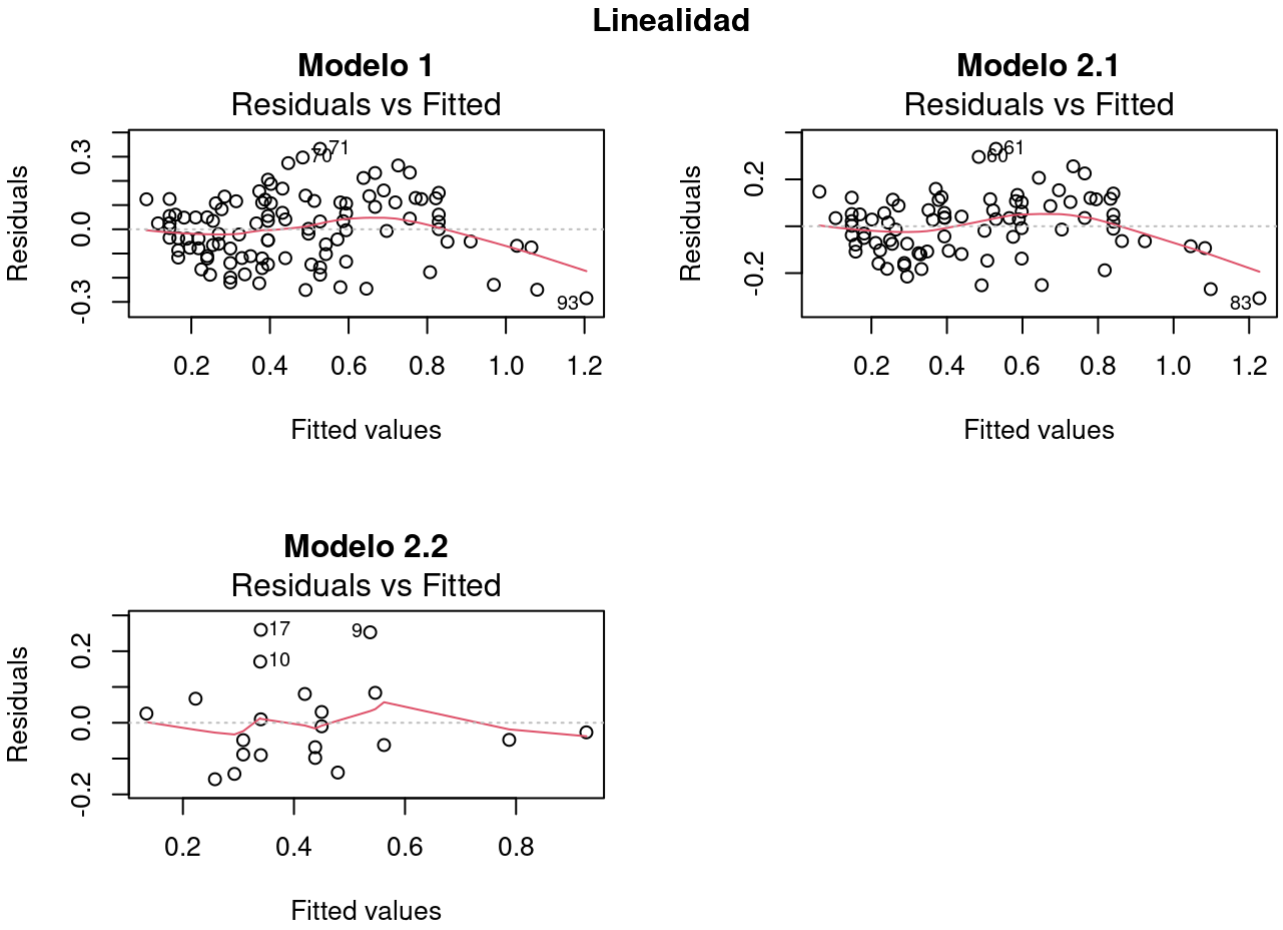
Verificar las hipótesis de Gauss-Markov y la normalidad de los residuos del modelo completo del apartado (b) del ejercicio 2. Realizar una completa diagnosis del modelo para ver si se cumplen las condiciones del modelo de regresión: normalidad, homocedasticidad,. . . y estudiar la presencia de valores atípicos de alto leverage y/o puntos influyentes.

Construir los gráficos correspondientes y justificar su interpretación. ¿Podemos considerar el modelo ajustado como fiable?

```
# Linealidad
par(mfrow = c(2, 2))

plot(leo_model1, 1)
title("Modelo 1")
plot(leo_model2.1, 1)
title("Modelo 2.1")
plot(leo_model2.2, 1)
title("Modelo 2.2")

main_title = "Linealidad"
title(main = main_title, outer = TRUE, line= -1)
```

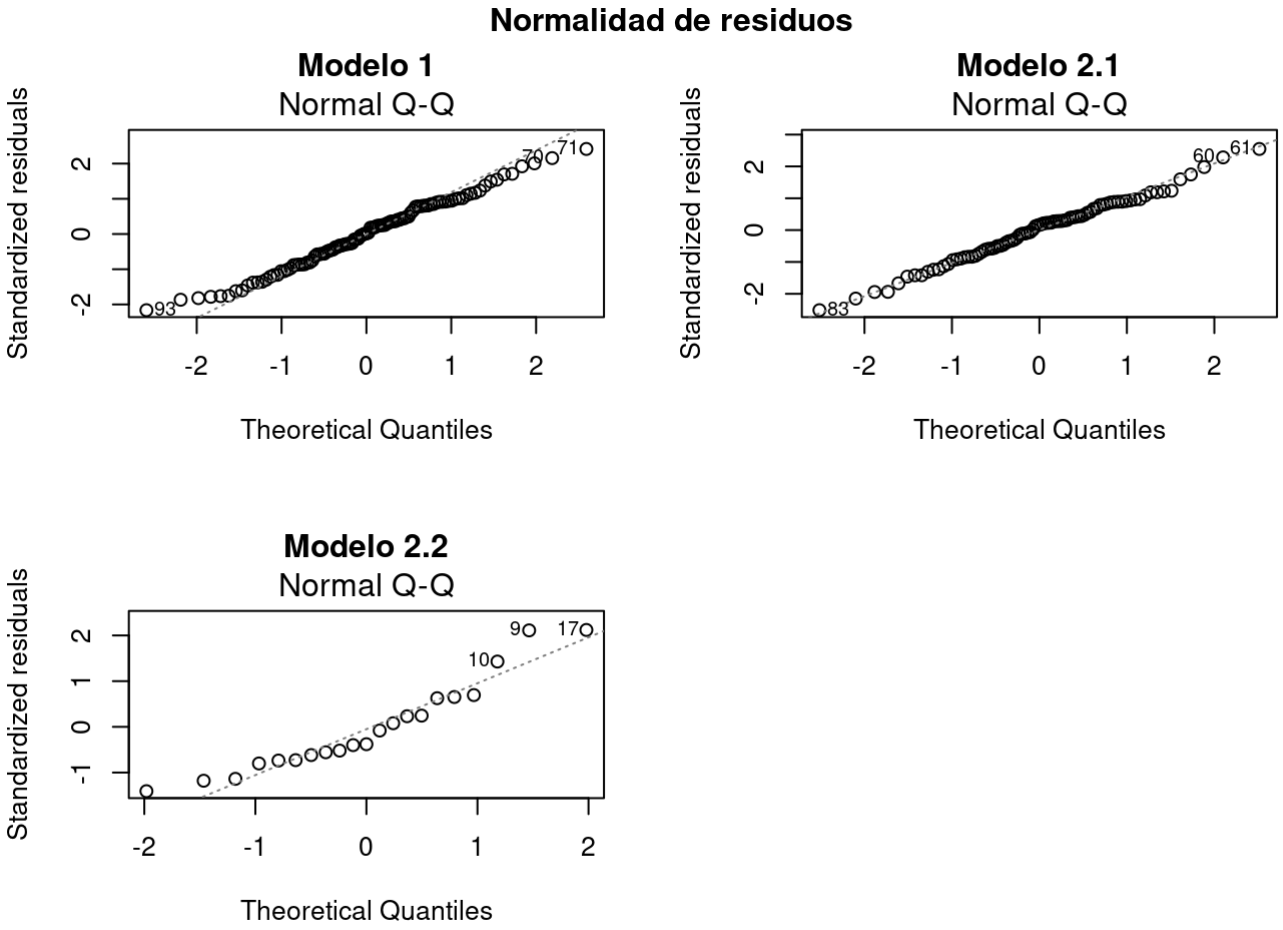


Parece que todos los modelos pasan el test de linealidad.

```
# Normalidad de residuos
par(mfrow = c(2, 2))

plot(leo_model1, 2)
title("Modelo 1")
plot(leo_model2.1, 2)
title("Modelo 2.1")
plot(leo_model2.2, 2)
title("Modelo 2.2")

main_title = "Normalidad de residuos"
title(main = main_title, outer = TRUE, line= -1)
```



Parece que todos los modelos siguen una normalidad en los residuos, realizamos Shapiro para corroborar:

```
# Prueba de Shapiro-Wilk para los residuos del modelo 1
p_values <- c(shapiro.test(leo_model1$residuals)$p.value,
              shapiro.test(leo_model2.1$residuals)$p.value,
              shapiro.test(leo_model2.2$residuals)$p.value)

model_names <- c("Modelo 1", "Modelo 2.1", "Modelo 2.2")

table_p_values <- data.frame(Modelo = model_names, Pvalue = p_values)

table_p_values
```

Modelo <chr>	**Pvalue** <dbl>
Modelo 1	0.45829005
Modelo 2.1	0.81942358
Modelo 2.2	0.08632029

3 rows

[H0: Normal; H1: No normal]

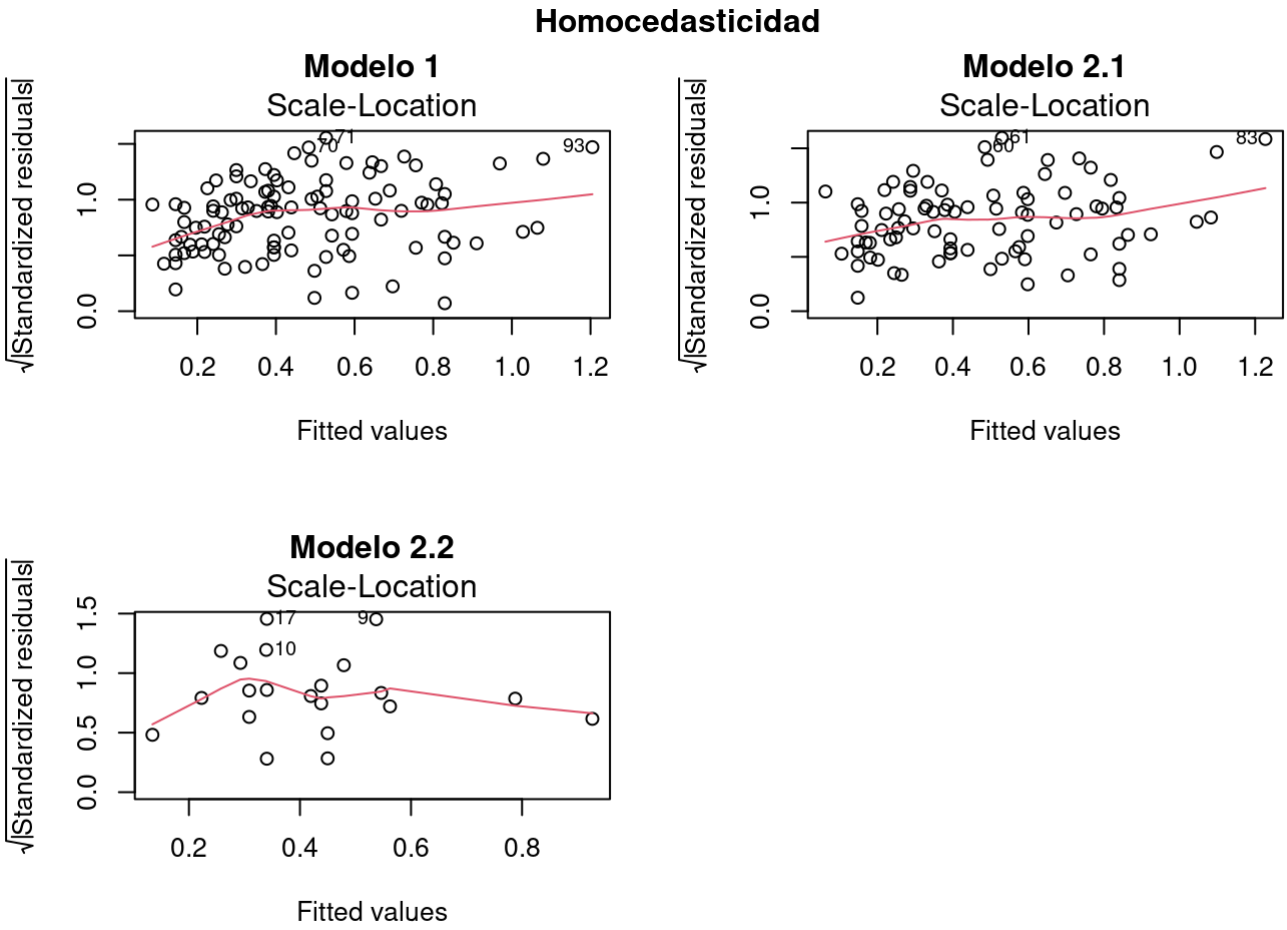
El valor de p-value para el modelo 2.2 es de 0.567, lo que indica que no hay suficiente evidencia para rechazar la hipótesis nula de normalidad en los residuos para este modelo. Por otro lado, tanto el modelo 1 como el modelo 2.1 tienen valores de p-value por debajo de 0.05, lo que sugiere que hay evidencia suficiente para rechazar la hipótesis nula de normalidad en los residuos para estos modelos.

En resumen, se puede decir que los modelos 1 y 2.1 no cumplen con la suposición de normalidad en los residuos, mientras que el modelo 2.2 sí la cumple.

```
#Homocedasticidad de residuos o homogeneidad de varianzas
par(mfrow = c(2, 2))

plot(leo_model1, 3)
title("Modelo 1")
plot(leo_model2.1, 3)
title("Modelo 2.1")
plot(leo_model2.2, 3)
title("Modelo 2.2")

main_title = "Homocedasticidad"
title(main = main_title, outer = TRUE, line= -1)
```




```
# Test BP
library(car)
```

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:DescTools':
##
##     Recode
```

```
p_values= c(ncvTest(leo_model1)$p,
              ncvTest(leo_model2.1)$p,
              ncvTest(leo_model2.2)$p)


model_names=  c("Modelo 1", "Modelo 2.1", "Modelo 2.2")

table_BP= data.frame(Modelo = model_names, Pvalue = p_values)

table_BP
```

Modelo <chr>	**Pvalue** <dbl>
Modelo 1	0.014756653
Modelo 2.1	0.009418231
Modelo 2.2	0.651195423

3 rows

[H0: Homocedasticidad; H1: No homocedasticidad]

Los valores de p obtenidos del test de Breusch-Pagan indican que el modelo 1 y el modelo 2.1 no presentan heterocedasticidad significativa, ya que sus valores de p son menores a 0.05, mientras que el modelo 2.2 presenta evidencia de heterocedasticidad, ya que su valor de p es mayor a 0.05.

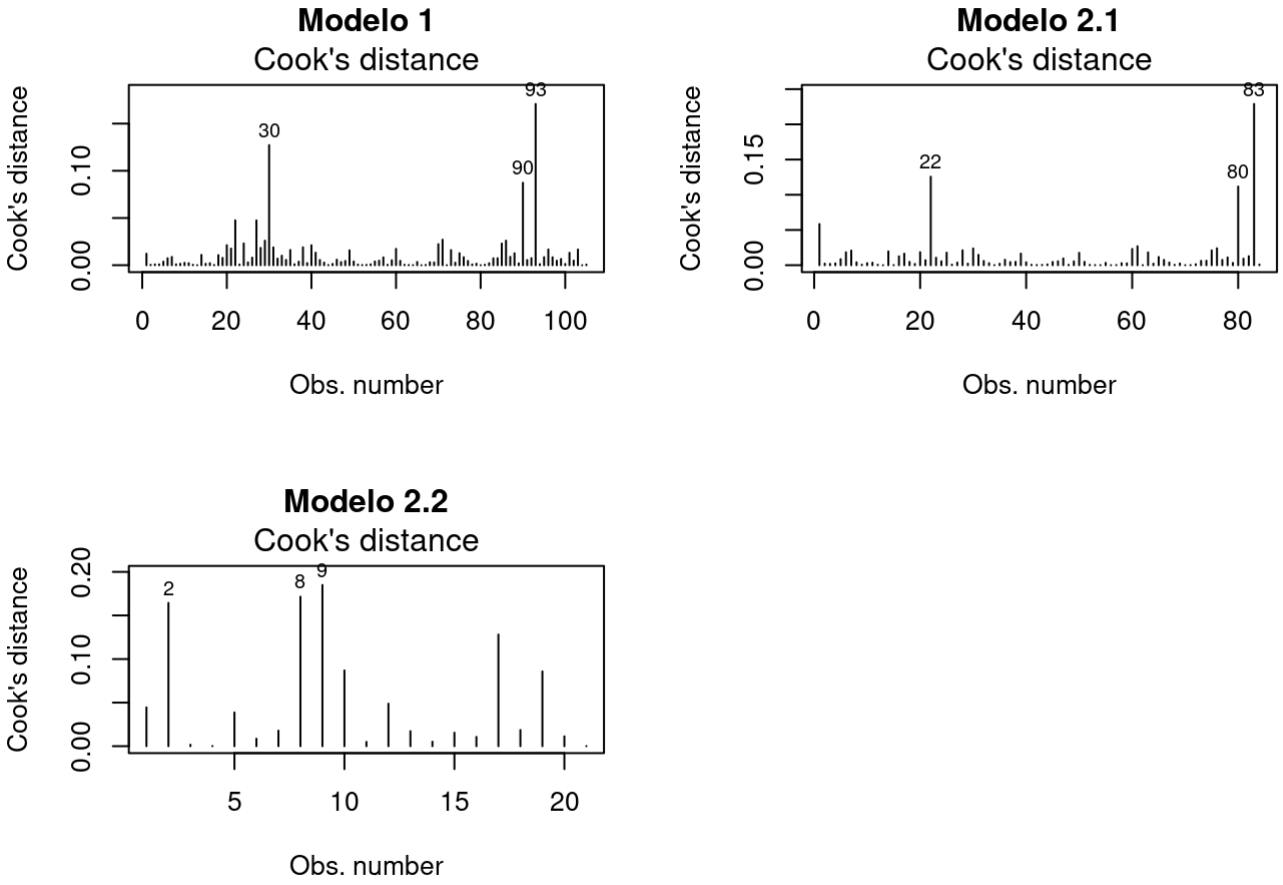
En resumen, el modelo 2.2 parece ser el que mejor cumple con los supuestos de linealidad, normalidad de residuos y homocedasticidad.

```
# Valores influyentes
par(mfrow = c(2, 2))

plot(leo_model1, 4)
title("Modelo 1")
plot(leo_model2.1, 4)
title("Modelo 2.1")
plot(leo_model2.2, 4)
title("Modelo 2.2")

main_title = "Valores influyentes"
title(main = main_title, outer = TRUE, line= -1)
```

Valores influyentes



Una observacion influyente se define como una observación que se diferencia marcadamente del conjunto de datos y tiene una gran influencia en el resultado del modelo, es decir, que no solo son outliers.

Pueden presentar un problema porque afectan los coeficientes de la ecuación y generan errores de predicción

Para detectarlos se utilizan medidas de influencia, entre las que resalta la distancia de Cook. LA distancia de Cook indica que un caso es un valor influyente cuando DCook >=1

```
# Distancia de cook
leon$cook <- cooks.distance(leo_model1)
which(leon$cook > 1)
```

```
## integer(0)
```

```
#leon$cook <- cooks.distance(leo_model2.1)
which(leon$cook2.1 > 1)
```

```
## integer(0)
```

```
leon$cook <- cooks.distance(leo_model2.2)
which(leon$cook2.2 > 1)
```

```
## integer(0)
```

```
# Valores atípicos alto leverage
library(outliers)
upper_test= grubbs.test(leo_model2.2$residuals)
upper_test
```

```
##
##  Grubbs test for one outlier
##
## data:  leo_model2.2$residuals
## G.17 = 2.17793, U = 0.75097, p-value = 0.2235
## alternative hypothesis: highest value 0.259687269664104 is an outlier
```

El resultado del Grubbs test indica que no se encontró evidencia suficiente para rechazar la hipótesis nula de que el valor más alto en los residuos del modelo sea un outlier. El p-value es mayor que el nivel de significancia estándar de 0.05, lo que sugiere que no hay suficiente evidencia para concluir que el valor es significativamente diferente del resto de los valores.

Usamos solo el modelo2.2 porque es al único que hemos podido constatar normalidad de los residuos.

```
lower_test= grubbs.test(leo_model2.2$residuals, opposite = TRUE)
lower_test
```

```
##
##  Grubbs test for one outlier
##
## data:  leo_model2.2$residuals
## G.2 = 1.32264, U = 0.90816, p-value = 1
## alternative hypothesis: lowest value -0.15770554988498 is an outlier
```

El valor p es 1. Al nivel de significancia del 5%, no rechazamos la hipótesis de que el valor más bajo -0.34 no es un valor atípico.

Aun así vemos en los gráficos de boxplot, como en el modelo de Nogorongoro hay 3 valores que seguramente sean outliers, aunque no podemos aplicar Grubbs, ya que el modelo 2.2 no sigue una regresión lineal según nuestro análisis.

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?

Sí, puede haber problemas al utilizar un modelo de regresión lineal para predecir proporciones ya que las proporciones están restringidas en el rango [0,1]. Es decir, la variable de respuesta no sigue una distribución normal y puede haber problemas con los supuestos de homocedasticidad y normalidad.

Hay dos vias; una alternativa seria utilizar modelos de regresión no lineal, como modelos logísticos o modelos de regresión beta. O, otra alternativa seria, transformar la variable respuesta para cumplir con los supuestos de linealidad.

Atendiendo a la naturaleza de la variable respuesta, ¿hay alguna transformación adecuada?

Sí, puede ser adecuada la transformación de la variable a logit. La transformación logit se utiliza comúnmente para datos de proporciones y tiene la ventaja de que los valores transformados están acotados entre -∞ y +∞.

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

```
# Aplicamos logit
library(arm)
```

```
## Loading required package: MASS
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.13-1, built: 2022-8-25)
```

```
## Working directory is /home/bebop/Nextcloud2/UOC_NC/2nSem/Regresión, Modelos y méto
dos/PEC1
```

```
##
## Attaching package: 'arm'
```

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

```
leo_model_logit = glm(prop.black ~ age + sex + area, data = leon, family = binomial)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
summary(leo_model_logit)
```

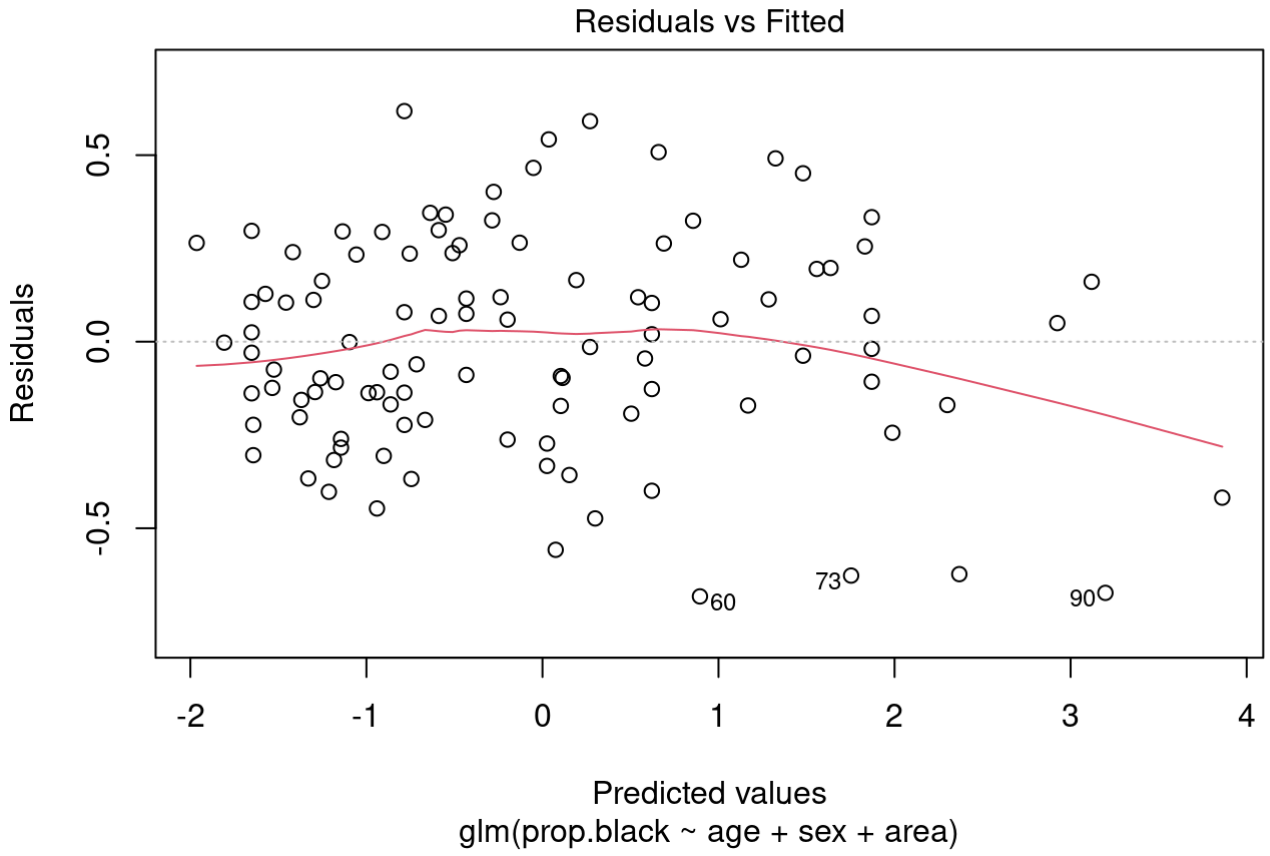
```
##
## Call:
## glm(formula = prop.black ~ age + sex + area, family = binomial,
##     data = leon)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.64259  -0.19040  -0.01433   0.22522   0.63255
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.42398    0.78172  -3.101  0.00193 **
## age          0.39034    0.09252   4.219 2.45e-05 ***
## sexM        -0.32151    0.51151  -0.629  0.52965
## areaS        0.35200    0.55467   0.635  0.52568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 37.5813  on 104  degrees of freedom
## Residual deviance:  8.7266  on 101  degrees of freedom
## AIC: 90.98
##
## Number of Fisher Scoring iterations: 5
```

```
# Sin aplicar logit
leo_model_no_logit = lm(prop.black ~ age + sex + area, data = leon)
summary(leo_model_no_logit)
```

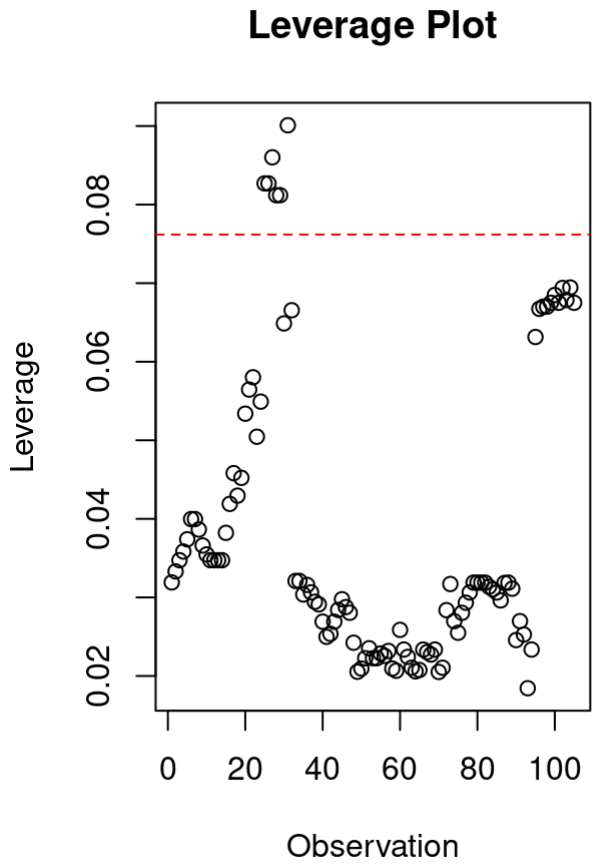
```
##
## Call:
## lm(formula = prop.black ~ age + sex + area, data = leon)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30265 -0.09116  0.00592  0.10049  0.32242
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.023324   0.044314   0.526   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
```

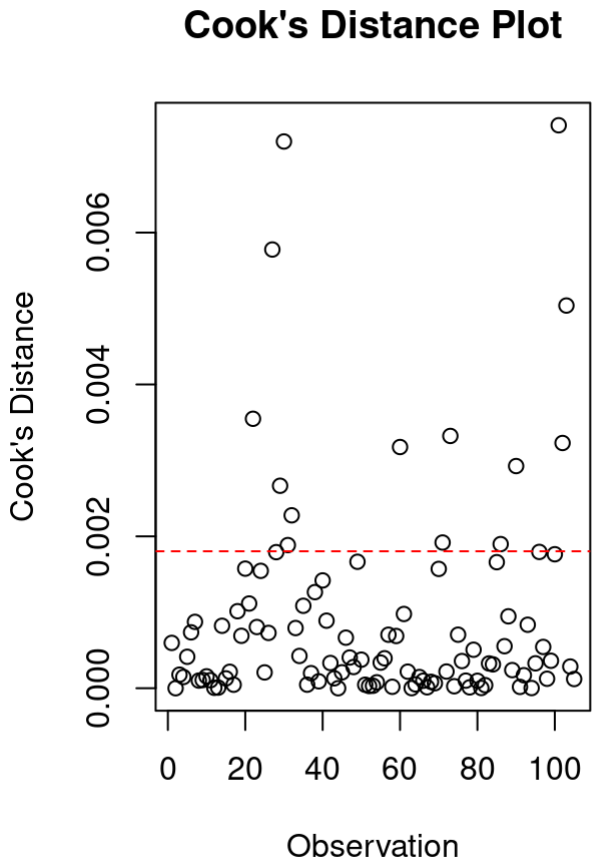
Realizar una rápida diagnosis del modelo transformado. ¿Estamos satisfechos con este nuevo modelo? ¿Qué otro ajuste nos podemos plantear para mejorar el modelo?

```
plot(leo_model_logit, 1)
```




```
par(mfrow = c(1, 2))
plot(hatvalues(leo_model_logit),
     xlab = "Observation", ylab = "Leverage",
     main = "Leverage Plot")
abline(h = 2 * mean(hatvalues(leo_model_logit)),
       col = "red", lty = 2)
plot(cooks.distance(leo_model_logit),
     xlab = "Observation", ylab = "Cook's Distance",
     main = "Cook's Distance Plot")
abline(h = 2 * mean(cooks.distance(leo_model_logit)),
       col = "red", lty = 2)
```





En general, podemos decir que estamos satisfechos con este nuevo modelo. Si queremos mejorar el modelo, podemos plantear otras transformaciones para la variable respuesta, o incluso probar con otros tipos de modelos, como modelos no lineales. Sin embargo, esto dependerá del contexto y de los objetivos específicos del análisis.

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

```
# Ajustar la variable respuesta
prop.arcoseno= asin(sqrt(leon$prop.black))
prop.arcoseno = pmin(prop.arcoseno, 1)

# Ajustar el modelo
modelo_arcoseno= glm(prop.arcoseno ~ age + sex + area, data = leon, family = binomia
l)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
prop.arcoseno_pred <- predict(modelo_arcoseno, type = "response")
prop.pred <- sin(prop.arcoseno_pred)^2


summary(modelo_arcoseno)
```

```
##
## Call:
## glm(formula = prop.arcoseno ~ age + sex + area, family = binomial,
##     data = leon)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.73817  -0.16365   0.09945   0.24432   0.68822
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.8815     0.8772  -2.145  0.03197 *
## age           0.5172     0.1337   3.868  0.00011 ***
## sexM         -0.2207     0.5079  -0.434  0.66393
## areaS         0.5685     0.6054   0.939  0.34765
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 36.9384  on 104  degrees of freedom
## Residual deviance:  9.3407  on 101  degrees of freedom
## AIC: 80.777
##
## Number of Fisher Scoring iterations: 6
```

En general, este modelo parece ser mejor que el modelo original ya que cumple con los supuestos del modelo lineal generalizado y tiene una deviance residual menor. Sin embargo, la transformación arcoseno puede dificultar la interpretación de los resultados y, en algunos casos, puede ser preferible utilizar otros métodos de transformación.