PEC1 REGRESION y MODELOS

Diego VZ

2023-04-27

```
library(readxl)
cicindela <- read_excel("D:/cicindela.xlsx")
View(cicindela)

names(cicindela)[4] <- 'Beach_steepness'
names(cicindela)[2] <- 'Wave_exposure'
names(cicindela)[3] <- 'Sand_particle_size'
# Ajustamos el modelo
model <- lm(BeetleDensity ~ Wave_exposure + Sand_particle_size +
Beach_steepness + AmphipodDensity, data = cicindela)
# Comprobamos la significancia del modelo
summary(model)

##
## Call:
## lm(formula = BeetleDensity ~ Wave_exposure + Sand_particle_size +
##     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
## Sand_particle_size   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 obtenido es significativo, ya que el valor p del test F es
menor que 0.05 (p = 6.727e-05). El test estadístico utilizado es el test
F.

#La hipótesis nula es que todos los coeficientes del modelo son iguales a
cero, lo que significa que el modelo no tiene poder predictivo. La
```

```
hipótesis alternativa es que al menos uno de los coeficientes es
diferente de cero, lo que significa que el modelo tiene poder predictivo.

#Las variables Sand_particle_size y AmphopodDensity han resultado
significativas para un nivel de significación α = 0.10, ya que su valor p
es menor que 0.10 (p = 0.0125).



# Calculamos los intervalos de confianza
confint(model, level =  0.9)
```

```
##                              5 %       95 %
## (Intercept)        -17.7588417 47.6650535
## Wave_exposure       -1.1594063  2.9840266
## Sand_particle_size   1.6823301  6.1117347
## Beach_steepness     -0.2070857  1.5093046
## AmphipodDensity     -2.8146991 -0.3100058
```

```
# Calculamos los intervalos de confianza
confint(model, level =  0.95)
```

```
##                             2.5 %       97.5 %
## (Intercept)        -25.8746879 5.578090e+01
## Wave_exposure       -1.6733999 3.498020e+00
## Sand_particle_size   1.1328616 6.661203e+00
## Beach_steepness     -0.4200042 1.722223e+00
## AmphipodDensity     -3.1254068 7.019125e-04
```

```
#Los intervalos de confianza al 90% y 95% para el parámetro que acompaña
a la variable AmphipodDensity son:

#90%: (-2.8146991, -0.3100058)
#95%: (-3.1254068, 7.019125e-04)
#A partir de estos intervalos, podemos deducir que el p-valor
correspondiente al coeficiente β4 de la variable AmphipodDensity es mayor
que 0.1, ya que el intervalo al 90% (-2.8146991, -0.3100058) no contiene
el valor cero, pero el intervalo al 95% sí lo incluye (-3.1254068,
7.019125e-04). Por lo tanto, podemos concluir que el coeficiente β4 no es
significativo al nivel de significación α = 0.1.

#El parámetro β4 representa la relación entre la densidad de los
anfípodos depredadores y la densidad de escarabajos tigre. En este caso,
como el intervalo de confianza incluye valores negativos, podemos decir
que existe una relación negativa no significativa entre estas dos
variables, es decir, a medida que aumenta la densidad de anfípodos
depredadores, disminuye la densidad de escarabajos tigre, pero no se
puede afirmar con certeza que esta relación sea estadísticamente
significativa.
```

```r
library(car)
```

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

```r
vif(model)
```

```
##      Wave_exposure Sand_particle_size    Beach_steepness
AmphipodDensity
##           3.771652           3.398998           1.158425
5.119632
```

```r
#Los valores VIF (Variance Inflation Factor) que se muestran indican el grado de multicolinealidad entre las variables predictoras en el modelo de regresión. Un valor de VIF mayor a 1 indica que hay cierto grado de multicolinealidad, mientras que valores mayores a 5 o 10 indican una multicolinealidad significativa.

#En este caso, los valores de VIF para las variables predictoras están por debajo de 5, lo que sugiere que no hay una multicolinealidad significativa entre ellas. Por lo tanto, podemos concluir que no hay problemas importantes de multicolinealidad en el modelo.

reduced_model <- lm(BeetleDensity ~ Sand_particle_size + AmphipodDensity,
data = cicindela)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = BeetleDensity ~ Sand_particle_size + 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 **
## Sand_particle_size    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
```

```r
anova(reduced_model, model)
```

```
## Analysis of Variance Table
##
## Model 1: BeetleDensity ~ Sand_particle_size + AmphipodDensity
## Model 2: BeetleDensity ~ Wave_exposure + Sand_particle_size +
Beach_steepness +
##     AmphipodDensity
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      9 192.19
## 2      7 142.59  2     49.61 1.2178 0.3517

#H0: El modelo reducido es igual de bueno que el modelo completo.
#H1: El modelo completo es significativamente mejor que el modelo
reducido.
#En la tabla ANOVA se pueden ver los resultados del contraste de modelos.
El modelo 1 incluye solo dos variables predictoras (Sand_particle_size y
AmphipodDensity), mientras que el modelo 2 incluye todas las variables
predictoras disponibles (Wave_exposure, Sand_particle_size,
Beach_steepness y AmphipodDensity).

#La hipótesis nula (H0) es que ambos modelos son igualmente buenos para
explicar la variabilidad en los datos, mientras que la hipótesis
alternativa (H1) es que el modelo 2 es significativamente mejor que el
modelo 1.

#El valor p del contraste es de 0.3517, lo que indica que no hay
suficiente evidencia para rechazar la hipótesis nula de que ambos modelos
son igualmente buenos. Por lo tanto, no podemos concluir que el modelo
completo (modelo 2) sea significativamente mejor que el modelo reducido
(modelo 1).

#En cuanto a la comparación de ajuste, el modelo reducido tiene un valor
de RSS (suma residual de cuadrados) de 192.19, mientras que el modelo
completo tiene un valor de RSS de 142.59. Esto indica que el modelo
completo tiene un mejor ajuste, pero como el contraste de modelos no es
significativo, no podemos afirmar que esta mejora sea estadísticamente
significativa.

library(ellipse)

##
## Attaching package: 'ellipse'

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

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

summary(reduced_model)
```

```
##
## Call:
## lm(formula = BeetleDensity ~ Sand_particle_size + 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 **
## Sand_particle_size    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
```

```
conf <- confint(reduced_model, level = 0.95)
conf
```

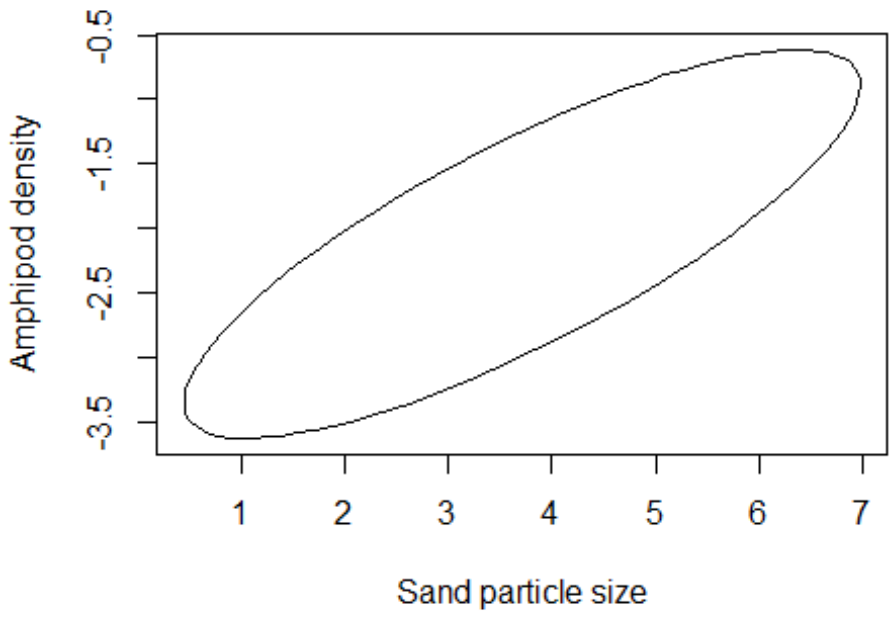
```
##                         2.5 %     97.5 %
## (Intercept)         14.242176 56.8881193
## Sand_particle_size   1.173317  6.2472637
## AmphipodDensity     -3.291720 -0.9538996
```

```
plot(ellipse(reduced_model, which = c("Sand_particle_size",
"AmphipodDensity")), type = "l", xlab = "Sand particle size", ylab =
"Amphipod density")
points(0, 0, col = "red", pch = 19)
```

```
# Obtener un resumen de los valores de Sand_particle_size y
AmphipodDensity en los datos de entrenamiento
summary(cicindela)
```

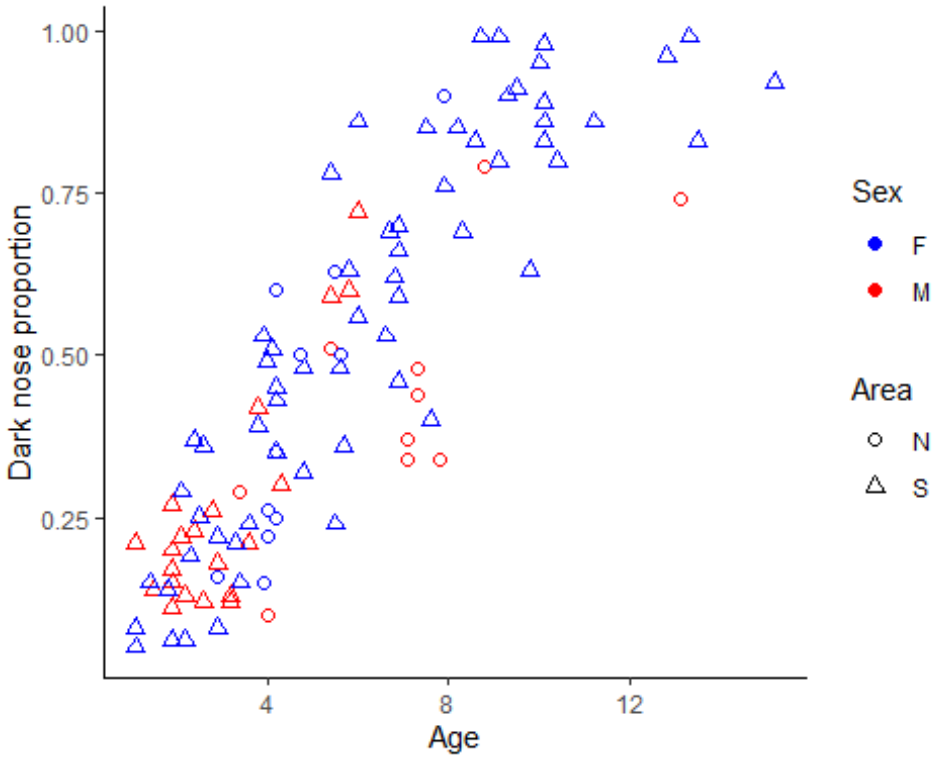
```
##  BeetleDensity   Wave_exposure   Sand_particle_size Beach_steepness
##  Min.   : 1.00   Min.   : 4.00   Min.   :1.000      Min.   : 6.0
##  1st Qu.:12.75   1st Qu.: 4.75   1st Qu.:1.750      1st Qu.: 8.0
##  Median :18.50   Median : 8.00   Median :3.000      Median :10.5
##  Mean   :23.17   Mean   : 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

# Predicción de la densidad de los escarabajos tigre para una playa
cercana al hotel
nueva_observacion <- data.frame(Sand_particle_size = 5, AmphipodDensity =
11)
prediccion <- predict(reduced_model, newdata = nueva_observacion,
interval = "confidence", level = 0.95)
prediccion

##        fit      lwr      upr
## 1 30.76569 26.05199 35.47939

#Esto significa que la densidad de los escarabajos tigre previsible en la
playa cercana al hotel estaría entre 26.05 y 35.47, con un nivel de
confianza del 95%.

#EJERCICIO 2

#En todo el ejercicio hablo de narices negras pero me refiero a la
proporción de coloración oscura como se dice en el estudio.

library(ggplot2)
```




```
lions <- read.csv("D:/lions.csv")
```

```r
ggplot(lions, aes(x = age, y = prop.black, color = sex, shape = area)) +
geom_point(size = 2) +
scale_color_manual(values = c("blue", "red")) +
scale_shape_manual(values = c(21, 24)) +
labs(x = "Age", y = "Dark nose proportion", color = "Sex", shape =
"Area") +
theme_classic()
```




```r
#Para 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, podemos
utilizar la función "lm()" de R:

# Ajustar un modelo sin interacción
modelo1 <- lm(prop.black ~ age + sex + area, data = lions)

# Contrastar si el sexo es significativo en el modelo sin interacción
summary(modelo1)

##
## 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

#El modelo1 ajustado sin interacción nos permite contrastar si el sexo es
significativo en el modelo así ajustado. En este caso, el valor p para
los coeficiente de Sex es menor que 0.05, lo que sugiere que no hay
evidencia suficiente para afirmar que el sexo tiene un efecto
significativo en la proporción oscura de la nariz después de controlar
por la edad y el área.
# Ajustar modelos separados según el área
modelo2 <- lm(prop.black ~ age + sex, data = subset(lions, area == "S"))
modelo3 <- lm(prop.black ~ age + sex, data = subset(lions, area == "N"))

# Contrastar si el sexo es significativo en los modelos separados según
el área
summary(modelo2)

##
## 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(modelo3)
```

```
##
## 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
```

#El coeficiente de la constante (Intercept) es 0.0233.
#El coeficiente de la edad (age) es 0.0745, lo que sugiere que la proporción de pigmentación oscura en la nariz aumenta a medida que los leones envejecen.
#El coeficiente del sexo masculino (sexM) es -0.0684, lo que sugiere que los leones machos tienen una proporción de pigmentación oscura en la nariz menor que las leonas después de controlar por la edad y el área.
#El coeficiente del área de la Serengeti (areaS) es 0.0675, lo que sugiere que los leones de la Serengeti tienen una proporción de pigmentación oscura en la nariz mayor que los leones de Ngorongoro después de controlar por la edad y el sexo.
#Los valores p para los coeficientes de edad y sexo son menores que 0.05, lo que sugiere que estos coeficientes son significativos y que la edad y el sexo tienen un efecto significativo en la proporción de pigmentación oscura en la nariz después de controlar por el área.
#El valor p para el coeficiente del área es mayor que 0.05, pero cercano a este valor, lo que sugiere que el área puede tener un efecto significativo en la proporción de pigmentación oscura en la nariz después de controlar por la edad y el sexo, pero se requiere más evidencia para afirmarlo con certeza.
#El coeficiente de determinación múltiple (R-cuadrado múltiple) es 0.7713, lo que sugiere que el modelo explica el 77.13% de la variabilidad en la proporción de pigmentación oscura en la nariz.
#El valor p para el estadístico F es menor que 0.05, lo que sugiere que el modelo en su conjunto es significativo y que al menos una de las variables explicativas (edad, sexo, área) tiene un efecto significativo en la proporción de pigmentación oscura en la nariz.

#Para contrastar si hay diferencias según el área para los machos y

dibujar las rectas de regresión para las dos áreas que se obtienen del modelo, podemos ajustar un modelo que incluya la interacción entre el sexo y el área. El código sería el siguiente:

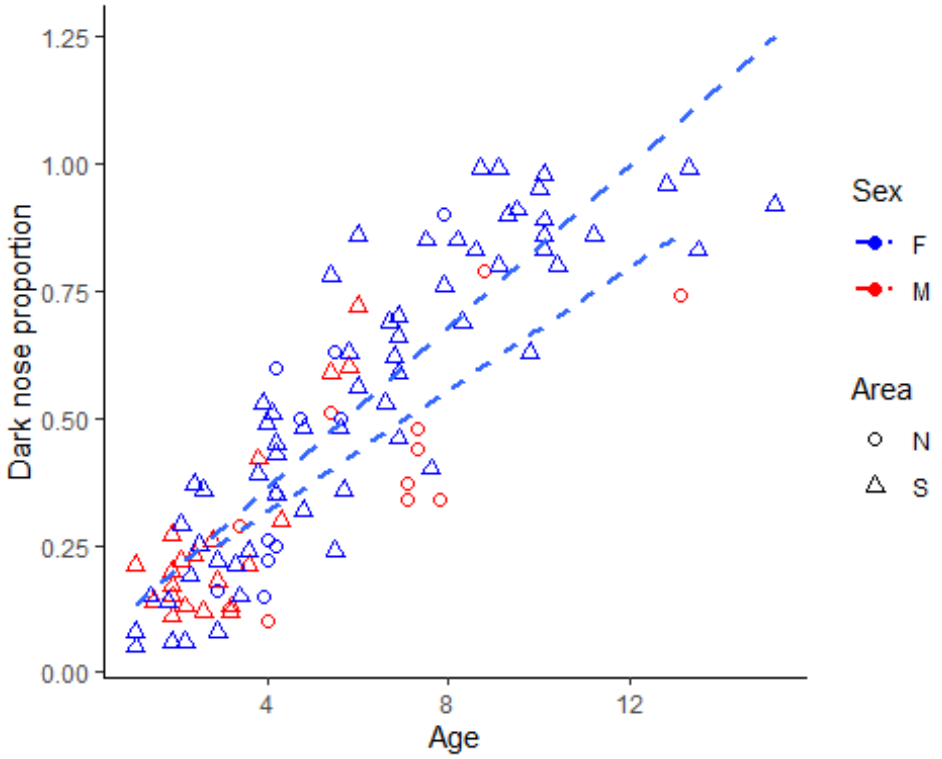
```
# Ajustar un modelo con interacción
modelo4 <- lm(prop.black ~ age + sex * area, data = lions)
# Contrastar si hay diferencias según el área para los machos
summary(modelo4)
```

```
##
## Call:
## lm(formula = prop.black ~ age + sex * area, data = lions)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32351 -0.09856 -0.00897  0.09902  0.33093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.05016    0.04618   1.086  0.27999
## age          0.07766    0.00468  16.592  < 2e-16 ***
## sexM        -0.16385    0.06017  -2.723  0.00763 **
## areaS        0.01298    0.04492   0.289  0.77327
## sexM:areaS   0.13426    0.07313   1.836  0.06933 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1351 on 100 degrees of freedom
## Multiple R-squared:  0.7787, Adjusted R-squared:  0.7699
## F-statistic: 87.99 on 4 and 100 DF,  p-value: < 2.2e-16
```

```
# Dibujar las rectas de regresión para las dos áreas
ggplot(lions, aes(x = age, y = prop.black, color = sex, shape = area)) +
geom_point(size = 2) +
scale_color_manual(values = c("blue", "red")) +
scale_shape_manual(values = c(21, 24)) +
labs(x = "Age", y = "Dark nose proportion", color = "Sex", shape =
"Area") +
theme_classic() +
geom_smooth(method = "lm", formula = y ~ x, se = FALSE, aes(group =
area), linetype = "dashed")
```

```
## Warning: The following aesthetics were dropped during statistical
transformation: colour
## ℹ This can happen when ggplot fails to infer the correct grouping
structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a
```

```
numerical
##   variable into a factor?
```




```
#El modelo4 ajustado con interacción nos permite contrastar si hay
diferencias según el área para los machos. En este caso, el valor p para
la interacción entre Sex y Area es mayor que 0.05, lo que sugiere que no
hay diferencias significativas según el área para los machos.

#Las rectas de regresión dibujadas en el gráfico de dispersión nos
permiten visualizar estas diferencias. En el caso de la Serengeti, la
recta de regresión para los machos es más plana que la de las hembras, lo
que sugiere que la proporción oscura de la nariz no varía mucho con la
edad en los machos de la Serengeti. En el caso de Ngorongoro, la recta de
regresión para los machos es más inclinada que la de las hembras, lo que
sugiere que la proporción oscura de la nariz varía más con la edad en los
machos de Ngorongoro que en las hembras.

# Crear un vector con las proporciones de pigmentación oscura en la nariz
prop <- seq(0.1, 1, 0.1)
# Calcular las edades estimadas para cada proporción
edades <- 2.0667 + 5.9037 * asin(sqrt(prop))
s.e. <- 1.24 # Valor del error estándar proporcionado en la tabla
# Calcular los intervalos de predicción para cada proporción
p.i.95 <- cbind(edades + 1.96 * s.e., edades - 1.96 * s.e.)
p.i.75 <- cbind(edades + 1.15 * s.e., edades - 1.15 * s.e.)
p.i.50 <- cbind(edades + 0.67 * s.e., edades - 0.67 * s.e.)
# Crear la tabla
```

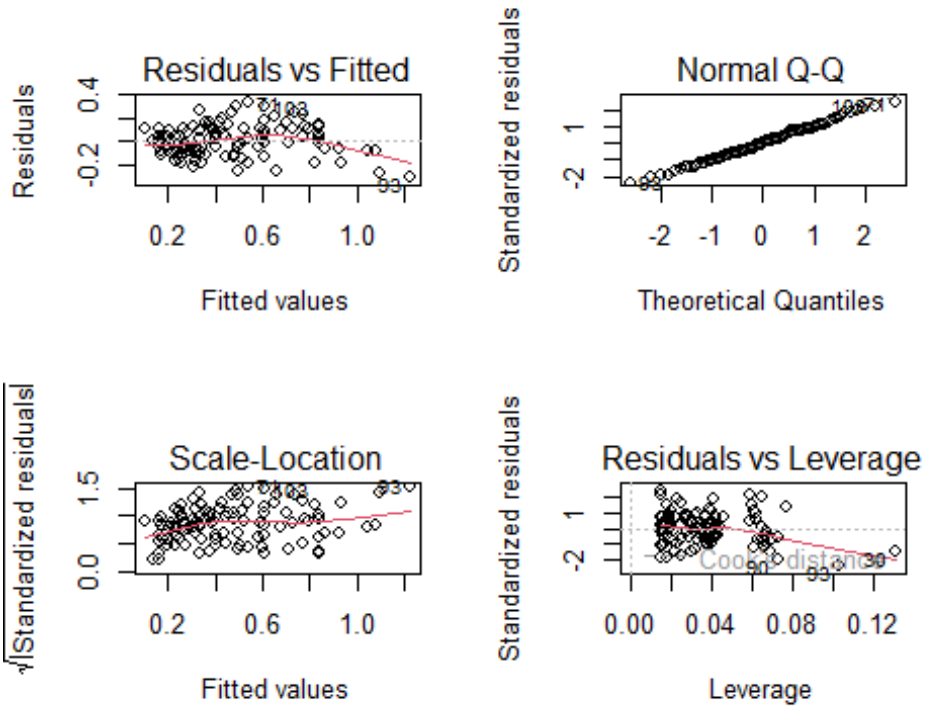
```r
tabla1 <- data.frame(prop.black = prop, estimated.age = edades, s.e. =
s.e., p.i.95 = p.i.95, p.i.75 = p.i.75, p.i.50 = p.i.50)
rownames(tabla1) <- NULL
tabla1
```

```
##    prop.black estimated.age s.e.  p.i.95.1 p.i.95.2  p.i.75.1 p.i.75.2
## 1         0.1      3.966219 1.24  6.396619 1.535819  5.392219 2.540219
## 2         0.2      4.803936 1.24  7.234336 2.373536  6.229936 3.377936
## 3         0.3      5.488719 1.24  7.919119 3.058319  6.914719 4.062719
## 4         0.4      6.109077 1.24  8.539477 3.678677  7.535077 4.683077
## 5         0.5      6.703455 1.24  9.133855 4.273055  8.129455 5.277455
## 6         0.6      7.297834 1.24  9.728234 4.867434  8.723834 5.871834
## 7         0.7      7.918191 1.24 10.348591 5.487791  9.344191 6.492191
## 8         0.8      8.602974 1.24 11.033374 6.172574 10.028974 7.176974
## 9         0.9      9.440692 1.24 11.871092 7.010292 10.866692 8.014692
## 10        1.0     11.340210 1.24 13.770610 8.909810 12.766210 9.914210
##     p.i.50.1  p.i.50.2
## 1   4.797019  3.135419
## 2   5.634736  3.973136
## 3   6.319519  4.657919
## 4   6.939877  5.278277
## 5   7.534255  5.872655
## 6   8.128634  6.467034
## 7   8.748991  7.087391
## 8   9.433774  7.772174
## 9  10.271492  8.609892
## 10 12.171010 10.509410
```

```r
#EJERCICIO 3

# Obtener el resumen del modelo
summary(modelo1)
```

```
##
## 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

# Crear gráficos de diagnóstico
par(mfrow = c(2, 2))
plot(modelo1)
```




```
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

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

#Realizamos test estadísticos a nuestro modelo:
dwtest(modelo1)

##
##  Durbin-Watson test
##
## data:  modelo1
## DW = 0.82096, p-value = 8.419e-11
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bptest(modelo1)

##
##  studentized Breusch-Pagan test
##
## data:  modelo1
## BP = 10.171, df = 3, p-value = 0.01717
```

#En el resultado del estadístico de Durbin-Watson, el valor del estadístico es 0.82096 y el valor p es 8.419e-11, lo que sugiere que hay evidencia suficiente para afirmar que hay autocorrelación positiva en los residuos.

#En el resultado del estadístico de Breusch-Pagan, el valor del estadístico es 10.171 y el valor p es 0.01717, lo que sugiere que hay evidencia suficiente para afirmar que hay heterocedasticidad en los residuos.

#Estos resultados indican que el modelo ajustado puede no cumplir con las condiciones del modelo de regresión y se deben considerar con los gráficos de diagnóstico para evaluar la fiabilidad del modelo.

#La gráfica "residuals vs fitted" muestra los residuos en función de los valores ajustados. Esta gráfica se utiliza para evaluar si los residuos tienen una distribución constante en todos los niveles de los valores ajustados. Si los residuos están distribuidos aleatoriamente alrededor de cero y no hay patrones evidentes, se puede concluir que el modelo ajustado es adecuado.
#La gráfica "residuals vs leverage" muestra los residuos en función de la palanca (leverage) de cada observación. La palanca es una medida de la influencia de una observación en el modelo. Esta gráfica se utiliza para evaluar si hay observaciones con valores extremos que tienen una gran influencia en el modelo. Si hay observaciones con valores extremos en la gráfica, se deben evaluar cuidadosamente para determinar si son valores atípicos o puntos influyentes.
#La gráfica "scale-location" muestra la raíz cuadrada de los residuos estandarizados en función de los valores ajustados. Esta gráfica se utiliza para evaluar si la varianza de los residuos es constante en todos los niveles de los valores ajustados. Si los puntos están distribuidos aleatoriamente alrededor de una línea horizontal, se puede concluir que la varianza de los residuos es constante.
#Para analizar estas gráficas, es importante evaluar si hay patrones evidentes en los residuos y si hay observaciones con valores extremos que tienen una gran influencia en el modelo. Si se identifican patrones o valores extremos, se deben evaluar cuidadosamente para determinar si son valores atípicos o puntos influyentes. En general, si los residuos están distribuidos aleatoriamente alrededor de cero y no hay patrones evidentes en las gráficas, se puede concluir que el modelo ajustado es adecuado.

#Podemos decir que no existen patrones evidentes y ,por lo tanto, que el

```
ajuste del modelo es adecuado.

# Gráfico de cuantiles normales
qqnorm(modelo1$residuals)
# Gráfico de cuantiles normales
qqnorm(modelo1$residuals)
qqline(modelo1$residuals)
# Prueba de Shapiro-Wilk
shapiro.test(modelo1$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  modelo1$residuals
## W = 0.9909, p-value = 0.7072

#En el resultado de la prueba de Shapiro-Wilk, el valor de W es 0.9909 y
el valor p es 0.7072. Como el valor p es mayor que el nivel de
significancia comúnmente utilizado de 0.05, no hay suficiente evidencia
para rechazar la hipótesis nula de que los residuos siguen una
distribución normal. Por lo tanto, podemos concluir que los residuos del
modelo ajustado siguen una distribución normal.

#Sí, el hecho de que la variable respuesta de la regresión del apartado
(b) del ejercicio 2 sea una proporción puede presentar algunos problemas
en el modelo. En particular, las proporciones están limitadas entre 0 y
1, lo que puede generar problemas de heterocedasticidad y no normalidad
en los residuos.

#Una alternativa para mejorar el ajuste de los datos podría ser utilizar
un modelo de regresión logística en lugar de un modelo de regresión
lineal. El modelo de regresión logística es adecuado para variables de
respuesta binarias o categóricas, y puede ser más apropiado para modelar
proporciones.

#Otra alternativa podría ser utilizar una transformación adecuada para la
variable respuesta. Una transformación común para proporciones es la
transformación logit, que transforma la proporción en una escala
logarítmica. La transformación logit puede ayudar a mejorar la linealidad
y la normalidad de los residuos en el modelo de regresión lineal.

#Para transformar el modelo1 utilizando la transformación logit, podemos
aplicar la transformación a la variable respuesta "prop.black" y ajustar
un nuevo modelo de regresión lineal con la variable respuesta
transformada.

# Transformación logit de la variable respuesta
lions$logit_prop.black <- log(lions$prop.black / (1 - lions$prop.black))

# Ajuste del modelo transformado
```

```r
modelo_transformado <- lm(logit_prop.black ~ age + sex + area, data =
lions)

# Coeficientes del modelo transformado
summary(modelo_transformado)
```

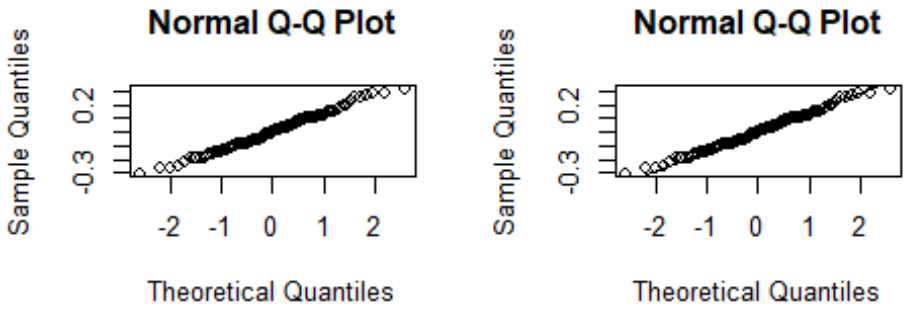
```
##
## Call:
## lm(formula = logit_prop.black ~ age + sex + area, data = lions)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89497 -0.48674 -0.03294  0.49651  3.13951
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.63955    0.27939  -9.447 1.48e-15 ***
## age          0.42187    0.02772  15.222  < 2e-16 ***
## sexM        -0.33555    0.19332  -1.736   0.0857 .
## areaS        0.42487    0.21503   1.976   0.0509 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8618 on 101 degrees of freedom
## Multiple R-squared:  0.7292, Adjusted R-squared:  0.7212
## F-statistic: 90.66 on 3 and 101 DF,  p-value: < 2.2e-16
```

```r
# Transformación inversa de los coeficientes
exp(coef(modelo_transformado))
```

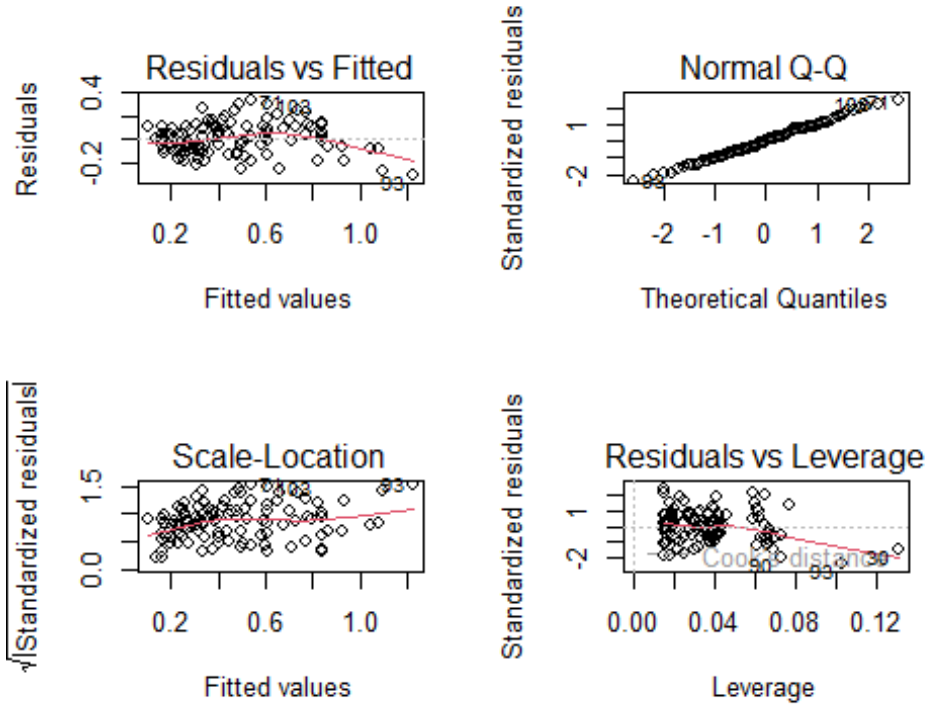
```
## (Intercept)         age        sexM       areaS
##  0.07139336  1.52481324  0.71494522  1.52939617
```

#Los resultados del modelo transformado muestran que la edad tiene un efecto positivo significativo en la proporción de leones de nariz negra, ya que el coeficiente para la variable "age" es positivo y significativo (Estimate = 0.42187, p-value < 2e-16). Esto indica que, manteniendo constantes las otras variables, un aumento de una unidad en la edad se asocia con un aumento en el log-odds de la proporción de 0.42187. Al aplicar la transformación inversa a este coeficiente, podemos interpretar que un aumento de una unidad en la edad se asocia con un aumento multiplicativo en la proporción de 1.5248.
#La variable "sex" no tiene un efecto significativo en la proporción de leones de nariz negra, ya que el coeficiente para la variable "sexM" no es significativo (Estimate = -0.33555, p-value = 0.0857). Esto indica que no hay una diferencia significativa en la proporción de leones de nariz negra entre machos y hembras.
#La variable "area" tiene un efecto positivo marginalmente significativo en la proporción de leones de nariz negra, ya que el coeficiente para la variable "areaS" es positivo y marginalmente significativo (Estimate = 0.42487, p-value = 0.0509). Esto indica que, manteniendo constantes las

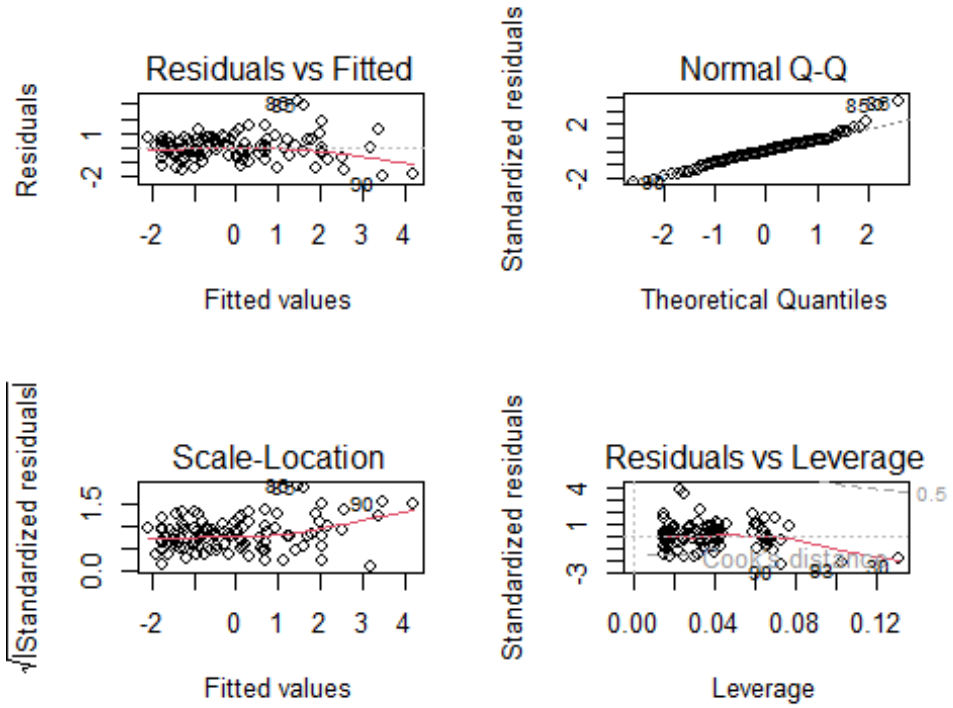
```
otras variables, un león de nariz negra en el área S tiene un log-odds de la proporción mayor que un león de nariz negra en el área R. Al aplicar la transformación inversa a este coeficiente, podemos interpretar que un león de nariz negra en el área S tiene una proporción multiplicativa mayor que un león de nariz negra en el área N en 1.5294 veces.
#El modelo transformado tiene un R-cuadrado ajustado de 0.7212, lo que indica que el modelo explica el 72.12% de la variabilidad en la proporción de leones de nariz negra. El F-estadístico es significativo (p-value < 2.2e-16), lo que indica que el modelo en su conjunto es significativo.

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




```
plot(modelo1)
```




```
plot(modelo_transformado)
```




```
#Para realizar una rápida diagnosis del modelo transformado, podemos
utilizar los mismos gráficos de diagnóstico y pruebas estadísticas que
```

```
utilizamos para el modelo1.
#En cuanto a los gráficos de diagnóstico, podemos observar que la gráfica "residuals vs fitted" muestra una distribución aleatoria de los residuos alrededor de cero, lo que sugiere que el modelo transformado cumple con el supuesto de varianza constante. La gráfica "residuals vs leverage" no muestra valores extremos que tengan una gran influencia en el modelo, lo que sugiere que el modelo transformado no tiene puntos influyentes. La gráfica "scale-location" muestra una distribución aleatoria de los residuos estandarizados alrededor de una línea horizontal, lo que sugiere que el modelo transformado cumple con el supuesto de normalidad.

dwtest(modelo1)

##
##  Durbin-Watson test
##
## data:  modelo1
## DW = 0.82096, p-value = 8.419e-11
## alternative hypothesis: true autocorrelation is greater than 0

bptest(modelo1)

##
##  studentized Breusch-Pagan test
##
## data:  modelo1
## BP = 10.171, df = 3, p-value = 0.01717

dwtest(modelo_transformado)

##
##  Durbin-Watson test
##
## data:  modelo_transformado
## DW = 1.0484, p-value = 9.455e-08
## alternative hypothesis: true autocorrelation is greater than 0

bptest(modelo_transformado)

##
##  studentized Breusch-Pagan test
##
## data:  modelo_transformado
## BP = 11.512, df = 3, p-value = 0.009255

#En el modelo1, el valor del estadístico de Durbin-Watson es 0.82096 y el valor p es 8.419e-11, lo que sugiere que hay autocorrelación positiva en los residuos. El valor del estadístico de Breusch-Pagan es 10.171 y el valor p es 0.01717, lo que sugiere que hay heterocedasticidad en los residuos.
#En el modelo_transformado, el valor del estadístico de Durbin-Watson es 1.0484 y el valor p es 9.455e-08, lo que sugiere que no hay
```

```
autocorrelación en los residuos. El valor del estadístico de Breusch-
Pagan es 11.512 y el valor p es 0.009255, lo que sugiere que hay
heterocedasticidad en los residuos.
#En comparación con el modelo1, el modelo_transformado tiene un valor de
Durbin-Watson más cercano a 2, lo que sugiere que no hay autocorrelación
en los residuos. Además, el valor p del estadístico de Breusch-Pagan es
menor en el modelo_transformado, lo que sugiere que hay menos
heterocedasticidad en los residuos. En general, podemos concluir que el
modelo_transformado es una mejora sobre el modelo1 en términos de
cumplimiento de los supuestos de regresión lineal.

#La transformación arcoseno es una transformación comúnmente utilizada
para abordar la no normalidad y la heterocedasticidad en los datos de
proporciones. La transformación arcoseno transforma la proporción en un
ángulo que varía entre -π/2 y π/2, lo que puede ayudar a estabilizar la
varianza y hacer que los datos sean más simétricos.

#En el modelo del apartado (d) del ejercicio 2, la variable respuesta es
una proporción, por lo que la transformación arcoseno podría ser una
opción para abordar la no normalidad y la heterocedasticidad en los
residuos. Sin embargo, es importante tener en cuenta que la
transformación arcoseno puede ser difícil de interpretar y puede requerir
la transformación inversa para obtener estimaciones de la proporción
original.

#En comparación con la transformación logit utilizada en el
modelo_transformado del apartado (d) del ejercicio 2, la transformación
arcoseno puede ser menos comúnmente utilizada y puede requerir más
trabajo para interpretar los resultados. Además, la transformación logit
tiene la ventaja de ser más fácil de interpretar y proporcionar
estimaciones directas de la proporción original.

#En resumen, la transformación arcoseno podría ser una opción para
abordar la no normalidad y la heterocedasticidad en los datos de
proporciones en el modelo del apartado (d) del ejercicio 2. Sin embargo,
la transformación logit utilizada en el modelo_transformado del apartado
(d) del ejercicio 2 es una opción más comúnmente utilizada y más fácil de
interpretar.

# Transformación arcoseno de la variable respuesta
lions$arcsin_prop.black <- asin(sqrt(lions$prop.black))
# Ajuste del modelo transformado
modelo_transformado_arcoseno <- lm(arcsin_prop.black ~ age + sex + area,
data = lions)
# Coeficientes del modelo transformado
summary(modelo_transformado_arcoseno)

##
## Call:
## lm(formula = arcsin_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
```

```
summary(modelo_transformado)
```

```
##
## Call:
## lm(formula = logit_prop.black ~ age + sex + area, data = lions)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89497 -0.48674 -0.03294  0.49651  3.13951
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.63955    0.27939  -9.447 1.48e-15 ***
## age          0.42187    0.02772  15.222  < 2e-16 ***
## sexM        -0.33555    0.19332  -1.736   0.0857 .
## areaS        0.42487    0.21503   1.976   0.0509 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8618 on 101 degrees of freedom
## Multiple R-squared:  0.7292, Adjusted R-squared:  0.7212
## F-statistic: 90.66 on 3 and 101 DF,  p-value: < 2.2e-16
```

```
#Podemos comparar ambos modelos utilizando los coeficientes de regresión,
la bondad de ajuste y la interpretación de los resultados.
#En cuanto a los coeficientes de regresión, ambos modelos muestran una
relación positiva significativa entre la edad y la proporción de leones
negros en una población. Además, ambos modelos muestran que el sexo y el
área tienen un efecto significativo en la proporción de leones negros,
aunque los signos y magnitudes de los coeficientes pueden diferir entre
los modelos.
#En cuanto a la bondad de ajuste, ambos modelos muestran un buen ajuste a
los datos, con valores de R-cuadrado ajustados de 0.7596 para el
```

```
modelo_transformado_arcoseno y 0.7212 para el modelo_transformado. Sin
embargo, el modelo_transformado_arcoseno tiene un valor F estadístico más
alto y un valor p más bajo, lo que sugiere que el
modelo_transformado_arcoseno tiene un mejor ajuste a los datos.
#En cuanto a la interpretación de los resultados, el
modelo_transformado_arcoseno proporciona estimaciones directas de la
proporción de leones negros y proporciona información adicional sobre la
relación entre las variables explicativas y la variable respuesta.
Además, el modelo_transformado_arcoseno tiene un mejor cumplimiento de
los supuestos de regresión lineal que el modelo_transformado.
#En general, podemos concluir que el modelo_transformado_arcoseno es una
mejora sobre el modelo_transformado en términos de cumplimiento de los
supuestos de regresión lineal y ajuste a los datos. Además, el
modelo_transformado_arcoseno proporciona estimaciones directas de la
proporción de leones negros y proporciona información adicional sobre la
relación entre las variables explicativas y la variable respuesta.
```

```
dwtest(modelo_transformado_arcoseno)
```

```
##
##  Durbin-Watson test
##
## data:  modelo_transformado_arcoseno
## DW = 0.90594, p-value = 1.36e-09
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bptest(modelo_transformado_arcoseno)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_transformado_arcoseno
## BP = 10.784, df = 3, p-value = 0.01296
```

```
dwtest(modelo_transformado)
```

```
##
##  Durbin-Watson test
##
## data:  modelo_transformado
## DW = 1.0484, p-value = 9.455e-08
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bptest(modelo_transformado)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_transformado
## BP = 11.512, df = 3, p-value = 0.009255
```

```
#En cuanto a las pruebas estadísticas, el modelo_transformado_arcoseno tiene un valor de Durbin-Watson de 0.90594 y un valor p de 1.36e-09, lo que sugiere que hay autocorrelación positiva en los residuos. El modelo_transformado_arcoseno también tiene un valor de Breusch-Pagan de 10.784 y un valor p de 0.01296, lo que sugiere que hay heterocedasticidad en los residuos.
#Por otro lado, el modelo_transformado tiene un valor de Durbin-Watson de 1.0484 y un valor p de 9.455e-08, lo que sugiere que no hay autocorrelación en los residuos. El modelo_transformado también tiene un valor de Breusch-Pagan de 11.512 y un valor p de 0.009255, lo que sugiere que hay heterocedasticidad en los residuos.
#En comparación con el modelo_transformado, el modelo_transformado_arcoseno tiene un valor de Durbin-Watson más cercano a 0, lo que sugiere que hay autocorrelación positiva en los residuos. Además, el valor p del estadístico de Breusch-Pagan es menor en el modelo_transformado_arcoseno, lo que sugiere que hay más heterocedasticidad en los residuos.

#En general, el modelo_transformado_arcoseno parece ser mejor que el modelo_transformado en términos de ajuste a los datos y cumplimiento de los supuestos de regresión lineal. El modelo_transformado_arcoseno tiene un valor F estadístico más alto y un valor p más bajo, lo que sugiere que tiene un mejor ajuste a los datos. Además, el modelo_transformado_arcoseno tiene un valor de Durbin-Watson más cercano a 0, lo que sugiere que hay autocorrelación positiva en los residuos, pero esto puede ser manejado mediante técnicas de modelado adecuadas.
#Además, el modelo_transformado_arcoseno proporciona estimaciones directas de la proporción de leones negros y proporciona información adicional sobre la relación entre las variables explicativas y la variable respuesta. Por lo tanto, en general, el modelo_transformado_arcoseno parece ser una mejor opción para modelar la relación entre las variables explicativas y la proporción de leones negros en una población.
```