

Regresión, modelos y métodos: Prueba de evaluación continua 1

Aràntzazu Alonso Carrasco

2023-05-04

Ejercicio 1

Un grupo de científicos norteamericanos están interesados en encontrar un hábitat adecuado para reintroducir una especie rara de escarabajos 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 factores bióticos y abióticos tales como la exposición a las olas, tamaño de la partícula de arena, pendiente de la playa y densidad de los anfípodos depredadores.

Los datos se hallan en la hoja de cálculo `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.**

Dado que estamos interesados en ver cómo se relaciona la densidad de escarabajos tigre con el resto de factores ambientales que se tienen en consideración, el modelo de regresión lineal a considerar debe tener la densidad de escarabajos (`BeetleDensity`) como variable respuesta y el resto de factores ambientales como variables regresoras. La ecuación del modelo será:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

siendo 1, 2, 3 y 4 los números que denominan las variables `Wave exposure`, `Sandparticlesize`, `Beach steepness` y `AmphipodDensity`, respectivamente.

Los coeficientes de regresión parciales se muestran a continuación (para ver el modelo completo, ver el apéndice):

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	14.9531059	17.2660810	0.8660394	0.41516279
## 'Wave exposure'	0.9123102	1.0934972	0.8343050	0.43165598
## Sandparticlesize	3.8970324	1.1689683	3.3337366	0.01252673
## 'Beach steepness'	0.6511095	0.4529741	1.4374098	0.19375861
## AmphipodDensity	-1.5623525	0.6610160	-2.3635624	0.05007816

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

El modelo obtenido es significativo (al 5%), ya que el p -valor = 6.7269394×10^{-5} . Para contestar a esta pregunta se usa un test F que se sitúa en la última fila del resumen del modelo que acabamos de ajustar. Concretamente, la hipótesis nula es

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

mientras que la hipótesis alternativa es

$$H_1 : \beta_i \neq 0 \text{ para algún } i = 1, \dots, 4$$

Por tanto, dado que hemos rechazado la hipótesis nula, aceptamos que al menos uno de los coeficientes de regresión es distinto de cero.

¿Qué variables han salido significativas para un nivel de significación $\alpha = 0.10$?

Para un nivel de significación $\alpha = 0.10$, las variables que han salido significativas son `Sandparticlesize`, que presenta un p -valor de 0.013, y `AmphipodDensity`, con un p -valor de 0.05.

- (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 β_4 ?**

El intervalo de confianza al 90% es (-2.8146991, -0.3100058), mientras que el intervalo de confianza al 95% es (-3.1254068, 7.0191249×10^{-4}). Dado que el intervalo de confianza al 95% incluye el cero pero el del 90% no, podíamos haber deducido que el parámetro de `AmphipodDensity` es significativamente diferente de cero para una significación del 10% pero no para una significación del 5%. Y, de hecho, es exactamente la misma conclusión que extraemos de mirar el p -valor = 0.05.

El parámetro correspondiente al predictor `AmphipodDensity` es -1.5624. Esto quiere decir que por cada unidad de aumento de la densidad de los anfípodos depredadores, la densidad de los escarabajos tigre disminuirá en -1.5624 unidades.

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

Para estudiar la multicolinealidad en un modelo de regresión lineal múltiple se utiliza el factor de inflación de la varianza (VIF, *variance inflation factor*). Un VIF = 1 indica que no hay multicolinealidad entre la variable predictora y las demás variables predictoras del modelo. Cuanto mayor sea el VIF, mayor será la inflación de la varianza y, por tanto, mayor será el grado de multicolinealidad. En general, se acepta que un VIF ≥ 5 indica una multicolinealidad significativa.

Los valores de VIF del modelo son los que se muestran a continuación:

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

Como podemos ver, el VIF para la variable `AmphipodDensity` es 5.12. Por tanto, vemos que hay un problema de multicolinealidad con esta variable, hecho que puede afectar a la interpretación de los coeficientes de regresión y a la precisión de las predicciones.

- (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 $\alpha = 0.05$. Escribir en forma paramétrica las hipótesis H_0 y H_1 de este contraste. Comparar el ajuste de ambos modelos.**

Consideramos el modelo siguiente:

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_4 x_{i4} + \epsilon_i$$

donde 2 y 4 representan **Sandparticlesize** y **AmphipodDensity**, respectivamente. En este caso, el p -valor que obtenemos mediante un contraste de modelos con un test F es 0.35. Con este p -valor aceptamos la hipótesis nula de que los parámetros de las variables no incluídas en el modelo simple son cero. Por tanto, la simplificación es perfectamente justificable.

Las hipótesis escritas en forma paramétrica son las siguientes:

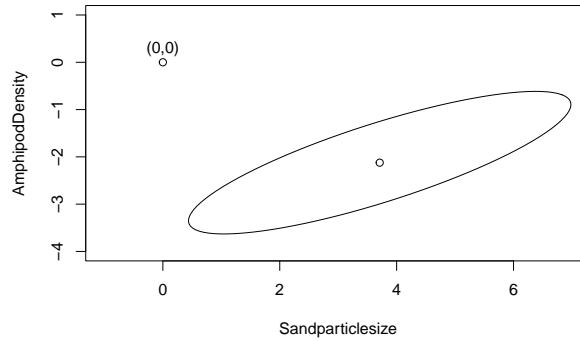
$$H_0 : \beta_1 = \beta_3 = 0$$

$$H_1 : \beta_i \neq 0 \text{ para algún } i = 1, 3$$

Para comparar el ajuste de cada uno de los modelos, nos podemos fijar en el coeficiente de determinación R^2 . El modelo sin simplificar tiene una $R^2 = 0.9336635$, mientras que el modelo simplificado tiene una $R^2 = 0.9304535$. Como vemos, ambos modelos tienen un ajuste muy parecido y muy cercano a 1, por lo que ambos ajustan bien los datos. Este hecho ya lo podíamos intuir al aceptar la hipótesis nula del contraste anterior ya que, dado que podemos aceptarla, significa que el modelo más simple explica bien los datos.

- (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.**

La región de confianza conjunta al 95% para los parámetros asociados con **Sandparticlesize** y **AmphipodDensity** para el modelo simplificado es la siguiente:



La elipse limita la región de confianza conjunta, mientras que el centro de la elipse corresponde a la estimación puntual de los dos parámetros.

Ver dónde queda el origen respecto a esta región de confianza equivale a hacer el siguiente contraste de hipótesis:

$$H_0 : \beta_{\text{Sandparticlesize}} = \beta_{\text{AmphipodDensity}} = 0$$

$$H_1 : \beta_{\text{Sandparticlesize}} \neq 0 \text{ o } \beta_{\text{AmphipodDensity}} \neq 0$$

Dado que el punto $(0, 0)$ se encuentra fuera de la región de confianza, rechazamos la hipótesis nula y aceptamos la hipótesis alternativa de que al menos un parámetro es significativamente distinto de cero si se considera conjuntamente.

- (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.**

Primero, miramos el rango de valores entre los que se mueven el tamaño de la partícula de arena y la densidad de anfípodos depredadores de los datos con los que hemos generado el modelo. Para el tamaño de la partícula de arena es $(1, 7)$. Dado que 5 se encuentra dentro del intervalo, no supone una extrapolación. Para la densidad de anfípodos depredadores es $(5, 19)$. Como 11 también se encuentra dentro de este rango, tampoco supone una extrapolación.

Por tanto, podemos usar el modelo reducido para obtener la predicción. El intervalo de confianza al 95% de la densidad de los escarabajos tigre es $(19.3, 42.23)$.

Ejercicio 2

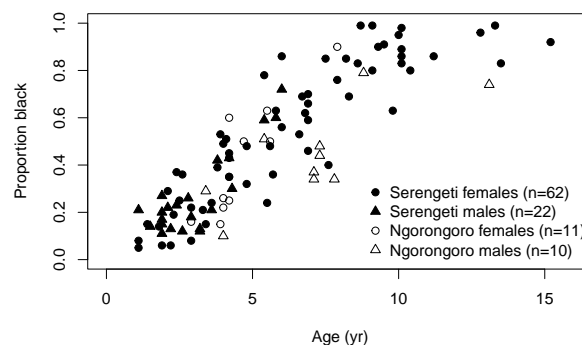
En el trabajo de Whitman et al. (2004) se estudia, entre otras cosas, la relación entre la edad de los leones y la proporción oscura en la coloración de sus narices. En el archivo `lions.csv` disponemos de los datos de 105 leones machos y hembras de dos áreas de Tanzania, el parque nacional de Serengueti y el cráter del Ngorongoro, entre 1999 y 2002. Las variables registradas son la edad conocida de cada animal y la proporción oscura de su nariz a partir de fotografías tratadas digitalmente.

En la figura 1 se reproduce el gráfico de dispersión 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 paquetes 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 numéricos de vuestro análisis pueden ser ligeramente distintos a los del trabajo original.

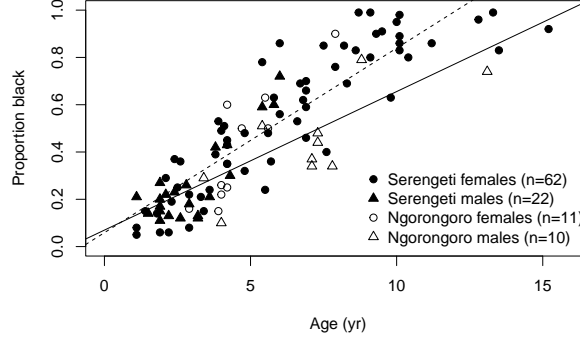
- (a) **Reproducir el gráfico de dispersión de la figura 1 (figura 4d del artículo) lo más fielmente posible al original, ya que se trata de una exigencia de los editores de la revista.**

La representación del gráfico es la siguiente:



- (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.

Primero, podemos representar las rectas correspondientes a machos y hembras sin hacer distinciones según la región (la línea sólida representa los machos y la línea discontinua representa las hembras):



Como podemos ver, las rectas parten del mismo lugar, pero las pendientes parecen distintas. Los modelos asociados a cada una de las rectas son:

$$y_{1i} = \alpha_1 + \beta_1 x_{1i} + \epsilon_{1i} \quad i = 1, \dots, n_1$$

$$y_{2i} = \alpha_2 + \beta_2 x_{2i} + \epsilon_{2i} \quad i = 1, \dots, n_2$$

donde los errores ϵ_{1i} y ϵ_{2i} verifican todas las hipótesis de un modelo lineal normal. Entonces, se puede plantear el siguiente contraste:

$$H_0 : \beta_1 = \beta_2$$

$$H_1 : \beta_1 \neq \beta_2$$

Para hacer este contraste, hay que considerar un modelo general único con el siguiente aspecto:

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_1} & 0 & 0 \\ 0 & 0 & 1 & x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{2n_2} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}$$

No obstante, si aceptamos la hipótesis nula, el modelo se simplifica de la manera siguiente:

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_1} & 0 & 0 \\ 0 & 0 & 1 & x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{2n_2} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}$$

Es decir, si las rectas son paralelas, las pendientes serán iguales y, por tanto, ambas β también serán iguales. Tras ajustar ambos modelos, debemos hacer un test F para resolver el contraste:

```
## Analysis of Variance Table
##
## Model 1: y ~ 0 + x0
## Model 2: y ~ 0 + x
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     102 1.9603
## 2     101 1.8954   1   0.06488 3.4573 0.06589 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como podemos ver, para un nivel de significación del 5%, aceptamos la hipótesis nula de que ambas rectas tienen la misma pendiente. Por tanto, no vemos diferencias entre los sexos. Cabe destacar, y esto es válido para todos los contrastes posteriores que realicemos, que tras aceptar la hipótesis de paralelismo, debemos estudiar si las hipótesis de Gauss-Markov se cumplen para dar por válidas las conclusiones.

Sin embargo, si tenemos en cuenta el área, debemos hacer el contraste entre machos y hembras para cada una de las regiones. Empecemos, por ejemplo, por Serengeti. Si repetimos el test F solo para los leones que forman parte de la región de Serengeti, obtenemos el resultado siguiente:

```
## Analysis of Variance Table
##
## Model 1: y ~ 0 + x0
## Model 2: y ~ 0 + x
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1       81 1.4037
## 2       80 1.3663   1   0.037385 2.189 0.1429
```

Como podemos ver, aceptamos la hipótesis nula de igualdad de pendientes. Por tanto, asumimos que no hay diferencias significativas entre la proporción de negro en la nariz de leones y leonas de la región de Serengeti.

Ahora, procedemos a repetir el contraste para Ngorongoro:

```
## Analysis of Variance Table
##
## Model 1: y ~ 0 + x0
## Model 2: y ~ 0 + x
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1       18 0.42171
## 2       17 0.28434   1   0.13737 8.2129 0.01071 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como podemos ver, para un nivel de significación del 5%, machos y hembras de la región de Ngorongoro sí presentan diferencias significativas en la proporción de negro de la nariz, tal como concluían en el estudio original.

- (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.

Para este contraste, seguiremos exactamente los mismos contrastes del apartado (b). Los modelos asociados a cada una de las rectas son:

$$y_{1i} = \alpha_1 + \beta_1 x_{1i} + \epsilon_{1i} \quad i = 1, \dots, n_1$$

$$y_{2i} = \alpha_2 + \beta_2 x_{2i} + \epsilon_{2i} \quad i = 1, \dots, n_2$$

donde los errores ϵ_{1i} y ϵ_{2i} verifican todas las hipótesis de un modelo lineal normal. Entonces, se puede plantear el siguiente contraste:

$$H_0 : \beta_1 = \beta_2$$

$$H_1 : \beta_1 \neq \beta_2$$

Para hacer este contraste, hay que considerar un modelo general único con el siguiente aspecto:

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_1} & 0 & 0 \\ 0 & 0 & 1 & x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{2n_2} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}$$

No obstante, si aceptamos la hipótesis nula, el modelo se simplifica de la manera siguiente:

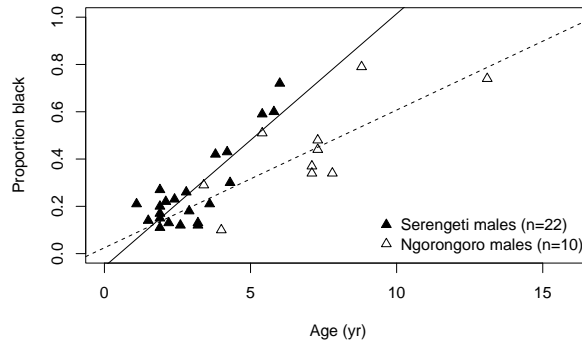
$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_1} & 0 & 0 \\ 0 & 0 & 1 & x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{2n_2} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}$$

Es decir, si las rectas son paralelas, las pendientes serán iguales y, por tanto, ambas β también serán iguales.

Tras ajustar ambos modelos, debemos hacer un test F para resolver el contraste:

```
## Analysis of Variance Table
##
## Model 1: y ~ 0 + x0
## Model 2: y ~ 0 + x
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      29 0.39136
## 2      28 0.33170   1  0.059652 5.0354 0.03292 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como podemos observar, hay diferencias significativas entre los leones de Serengeti y los leones de Ngorongoro. Para verlo gráficamente, podemos hacer la siguiente representación, donde la línea continua representa los leones de Serengeti y la línea discontinua representa los leones 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?

Los modelos estudiados en los apartados anteriores no sirven para hacer estas predicciones. Deberíamos generar un modelo en el que solo tuvieramos a las hembras de ambas regiones y, a partir de éste, hacer las predicciones.

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

En el artículo especifican que el modelo usado es *least-squares regression of a truncated data set for 63 known-aged females in the Serengeti and Ngorongoro aged ≤ 10 yr* ($y = 2.0667 + 5.9037 \arcsin(x)$).

Es decir, debemos generar el modelo siguiente:

$$age_i = \beta_0 + \beta_1 \arcsin(prop.black_i) + \epsilon_i$$

Una vez ajustamos el modelo y generamos las predicciones, obtenemos la tabla que se muestra a continuación, en la que tenemos la edad estimada (en años) y los distintos intervalos de predicción según el nivel de confianza (95, 75 y 50%, respectivamente):

prop.black	age	lower95	upper95	lower75	upper75	lower50	upper50
0.5	5.34	2.95	7.74	3.96	6.73	4.53	6.16

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

El error estándar de la estimación de la edad es:

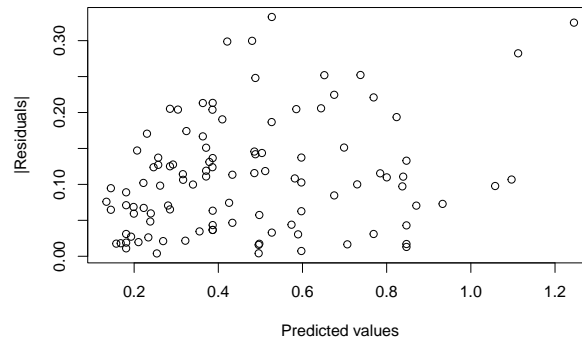
$$Var(age) = Var(\beta_0 + \arcsin(prop.black)\beta_1)$$

Ejercicio 3

- (a) 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?

Empezaremos estudiando si la varianza de los errores es constante. Para ello, representamos los valores absolutos de los residuos para ver la magnitud de la varianza:



La forma del gráfico nos hace sospechar que no hay homocedasticidad en la varianza, ya que los residuos de los valores predichos pequeños son muy inferiores a los residuos de los valores predichos mayores. Para comprobarlo, también podemos ajustar una recta a las raíces cuadradas de los valores absolutos de los residuos y los valores predichos. El resumen de dicho ajuste es el siguiente:

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.252599   0.025145 10.0456 < 2e-16
## fitted(general.complete.lm) 0.117504   0.047430  2.4774  0.01486
##
## n = 105, p = 2, Residual SE = 0.11957, R-Squared = 0.06
```

El p -valor de la pendiente obtenido es estadísticamente significativo para un nivel de significación del 5%. Por tanto, confirmamos las sospechas de la falta de homocedasticidad.

A continuación, estudiaremos la asunción de normalidad. Para ello haremos un qqplot:

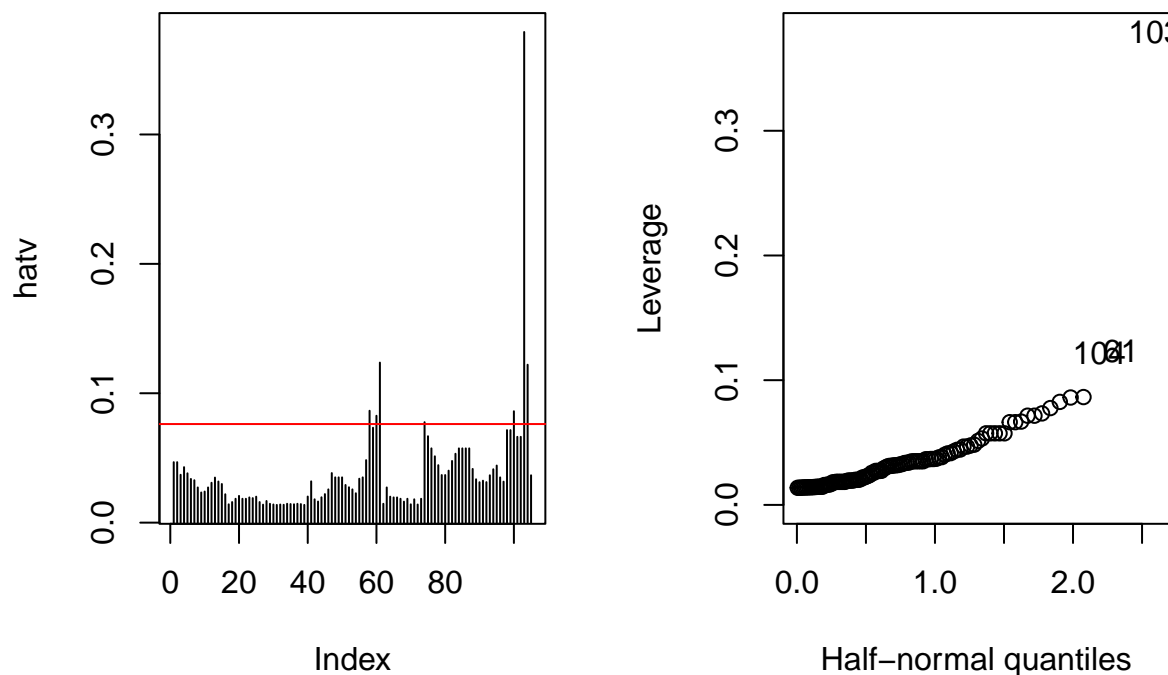


No vemos que la distribución de los residuos se aparte mucho de la normalidad. Además, podemos hacer el test de Shapiro-Wilk para ver si hay una desviación de la normalidad significativa:

```
##
## Shapiro-Wilk normality test
##
## data: residuals(general.complete.lm)
## W = 0.99354, p-value = 0.9043
```

Aceptamos, por tanto, la hipótesis de normalidad.

A continuación, buscaremos los puntos con alto *leverage*. El valor medio de los *leverages* es p/n (siendo p el número de parámetros y n el número de observaciones) y como referencia podemos considerar anómalas las observaciones con *leverages* del doble de este valor (representado con una línea horizontal roja en el gráfico). A continuación, mostramos un gráfico donde podemos visualizar si hay valores por encima del umbral que hemos determinado, es decir, si hay valores con elevado *leverage*. Para identificar los valores más grandes, también podemos representar los *leverages* contra los cuantiles de una distribución seminormal (como se muestra a la derecha).



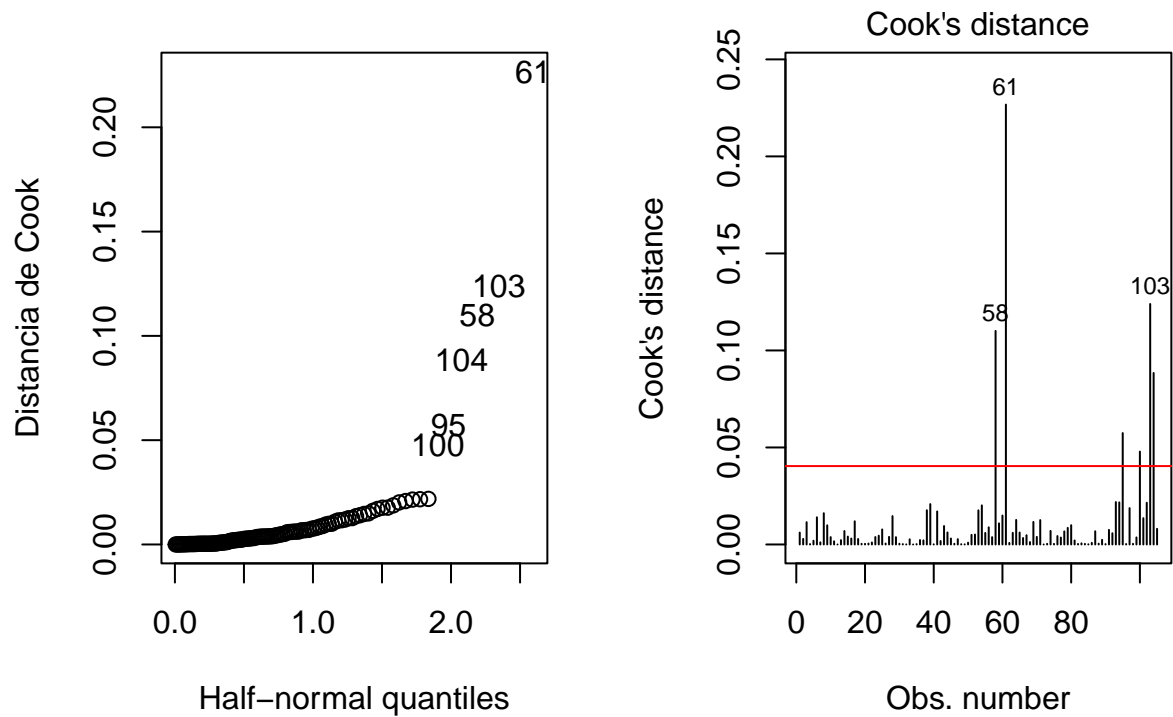
Como vemos en el gráfico de la izquierda, sí que hay observaciones con un elevado *leverage*. En el gráfico de la derecha podemos ver que la observación del león 103 tiene un *leverage* unusually muy elevado, seguido de los leones 61 y 104.

Seguidamente, buscamos los *outliers* que pueda haber. Para ello, usaremos el valor crítico de la t de Student y la corrección de Bonferroni para comparaciones múltiples. El número de *outliers* según este criterio es:

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

Vemos que con este criterio todos los residuos quedan por debajo del valor crítico y no hallamos ningún valor atípico. Queda la duda de si puede haber grupos de valores atípicos que no hayamos sabido encontrar.

Para buscar los valores influyentes, calculamos la distancia de Cook y la representamos contra los cuartiles de una distribución seminormal (gráfico de la izquierda). Posteriormente, podemos generar un gráfico con el criterio de selección (gráfico de la derecha).



Como vemos, los leones 61, 103, 58, 104, 95 y 100 sobrepasan el umbral que hemos establecido.

Con todo, no podemos considerar el modelo ajustado como fiable porque no se cumple la asunción de homocedasticidad.

- (b) 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?

Apéndice

En este apéndice se muestra todo el código que se ha ido empleando en la resolución de los ejercicios.

Ejercicio 1

```
# Cargamos los paquetes necesarios
library("readxl")
library("faraway")
# Creamos un dataset a partir de los datos del fichero cicindela.xlsx
cici<-read_excel("cicindela.xlsx")
# Ajustamos la regresión lineal múltiple con la función lm()
cici.lm<-lm(BeetleDensity~.,data=cici)
# Guardamos el summary del modelo
sum.cici.lm<-summary(cici.lm)
sum.cici.lm
```

```
##
## Call:
## lm(formula = BeetleDensity ~ ., data = cici)
##
## 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
```

```
# Obtenemos el estadístico f
f<-summary(cici.lm)$fstatistic
# Imprimimos los coeficientes del modelo
sum.cici.lm$coefficients
```

```
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  14.9531059 17.2660810  0.8660394 0.41516279
## 'Wave exposure'  0.9123102 1.0934972  0.8343050 0.43165598
## Sandparticlesize 3.8970324 1.1689683  3.3337366 0.01252673
## 'Beach steepness' 0.6511095 0.4529741  1.4374098 0.19375861
## AmphipodDensity -1.5623525 0.6610160 -2.3635624 0.05007816
```

```

# Obtenemos el p-valor del modelo
pf(f[1],f[2],f[3],lower.tail=F)

##          value
## 6.726939e-05

# Extraemos el p-valor de las variables significativas
round(sum.cici.lm$coefficients[3,4],3) # Sandparticlesize

## [1] 0.013

round(sum.cici.lm$coefficients[5,4],3) # AmphipodDensity

## [1] 0.05

# Calculamos los intervalos de confianza de AmphipodDensity
# al 90%
confint(cici.lm,level=0.9)[5,]

##          5 %          95 %
## -2.8146991 -0.3100058

# al 95%
confint(cici.lm,level=0.95)[5,]

##          2.5 %          97.5 %
## -3.1254068143  0.0007019125

# Extraemos el parámetro de AmphipodDensity
round(sum.cici.lm$coefficients[5],4)

## [1] -1.5624

# Calculamos los VIF
vif(cici.lm)

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

# Generamos el modelo reducido
cici.lm.i<-lm(BeetleDensity~Sandparticlesize+AmphipodDensity,data=cici)
summary(cici.lm.i)

##
## Call:
## lm(formula = BeetleDensity ~ Sandparticlesize + AmphipodDensity,
##     data = cici)
##

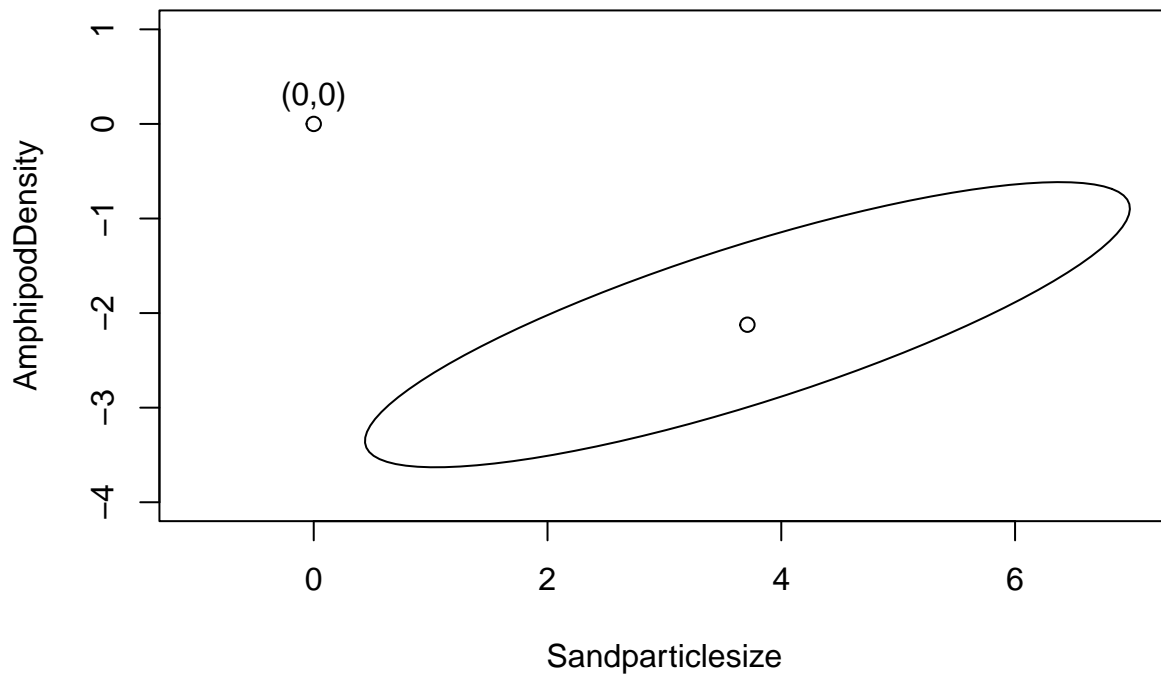
```

```
## 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

# Comparamos ambos modelos
anova(cici.lm,cici.lm.i)

## 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

# Generamos la región de confianza de Sandparticlesize y AmphipodDensity
require(ellipse)
plot(ellipse(cici.lm.i,c(2,3)),type="l",xlim=c(-1,7),ylim=c(-4,1))
points(coef(cici.lm.i)[2],coef(cici.lm.i)[3])
# Añadimos el origen
points(0,0)
text(0,0,labels="(0,0)",pos=3)
```



```
# Comprobamos si es una extrapolación
range_sand<-c(min(cici$Sandparticlesize),max(cici$Sandparticlesize))
range_amphi<-c(min(cici$AmphipodDensity),max(cici$AmphipodDensity))
print(c(range_sand,range_amphi))
```

```
## [1]  1  7  5 19
```

```
x0<-data.frame(Sandparticlesize=5,AmphipodDensity=11)
predict(cici.lm.i,x0,interval = "prediction",level=.95)
```

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

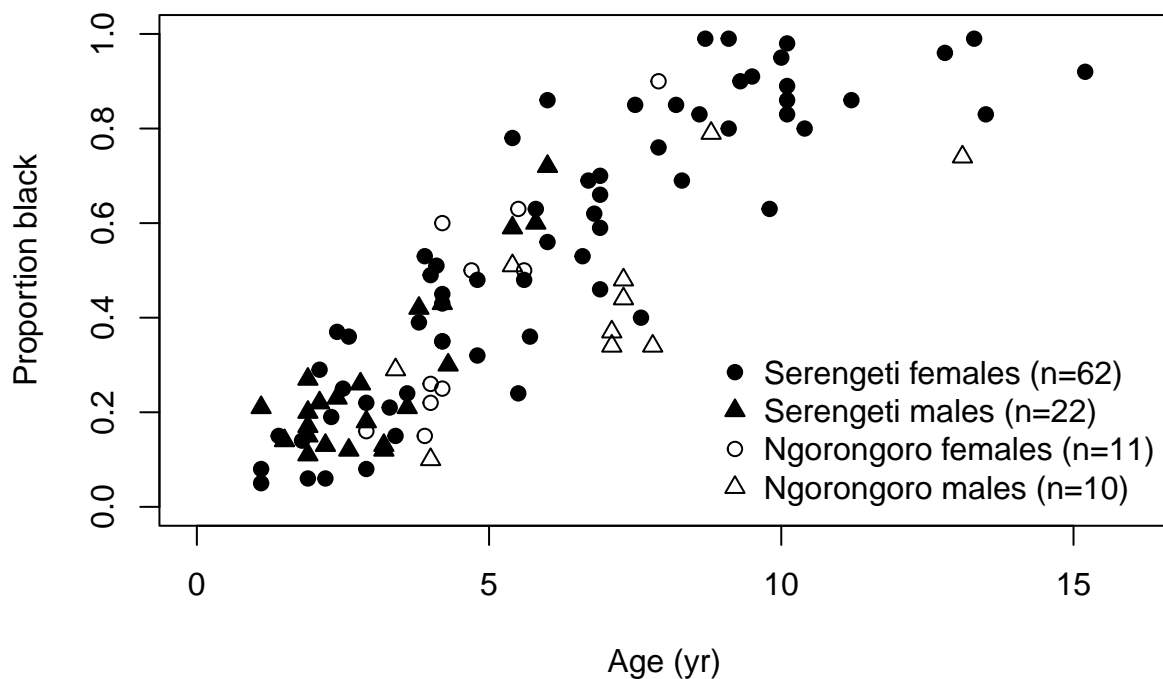
Ejercicio 2

```
# Cargamos los datos
lions<-read.csv("lions.csv")
# Generamos el gráfico
plot(lions$prop.black ~ lions$age,
     pch = ifelse(lions$sex == "F", ifelse(lions$area == "S", 21, 1),
                  ifelse(lions$area == "S", 24, 2)),
     col = ifelse(lions$area == "S", "black", "black"),
```

```

bg = ifelse(lions$area == "S", "black", "white"),
data = lions,
ylab = "Proportion black",
xlab = "Age (yr)",
ylim = c(0,1),
xlim = c(0,16))
legend("bottomright",
      legend = c("Serengeti females (n=62)",
                  "Serengeti males (n=22)",
                  "Ngorongoro females (n=11)",
                  "Ngorongoro males (n=10)"),
      col = c(rep("black",4)),
      pch = c(21,24,1,2),
      pt.bg = c("black","black","white", "white"),
      bty = "n")

```



```

# Generamos el gráfico con las rectas según el sexo
male.lm<-lm(prop.black~age,data=lions,subset = sex=="M")
female.lm<-lm(prop.black~age,data=lions,subset = sex=="F")
plot(lions$prop.black ~ lions$age,
     pch = ifelse(lions$sex == "F", ifelse(lions$area == "S", 21, 1),
                  ifelse(lions$area == "S", 24, 2)),
     col = ifelse(lions$area == "S", "black", "black"),
     bg = ifelse(lions$area == "S", "black", "white"),
     data = lions,

```

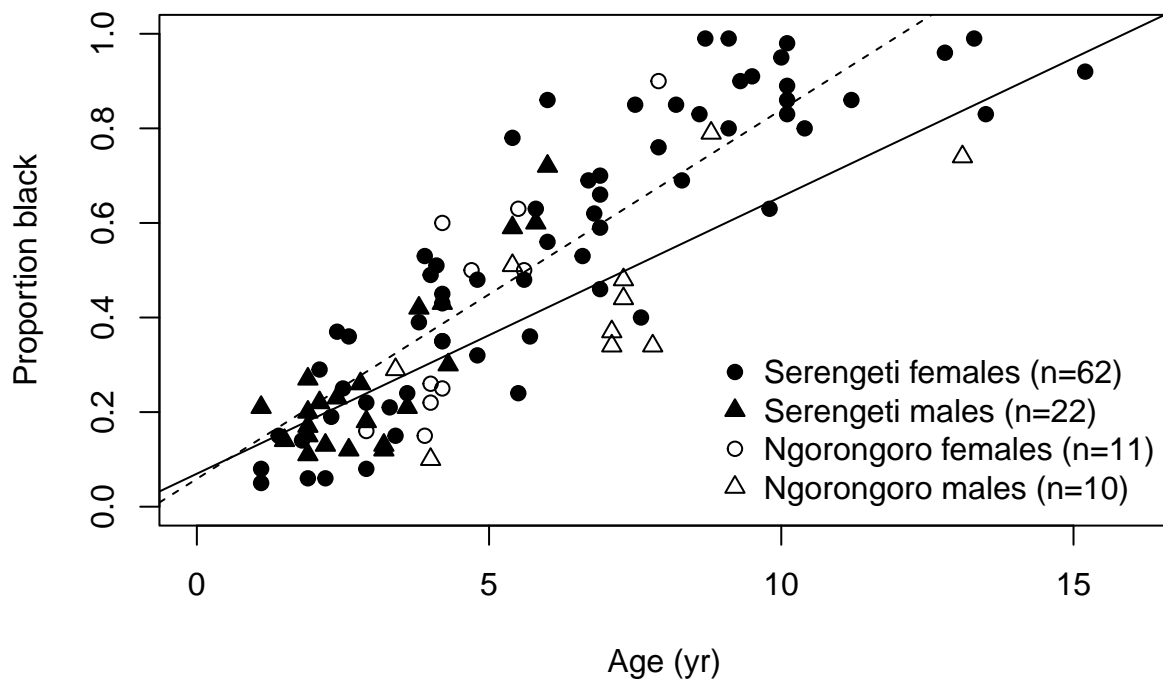


```

ylab = "Proportion black",
xlab = "Age (yr)",
ylim = c(0,1),
xlim = c(0,16))
legend("bottomright",
      legend = c("Serengeti females (n=62)",
                  "Serengeti males (n=22)",
                  "Ngorongoro females (n=11)",
                  "Ngorongoro males (n=10)"),
      col = c(rep("black",4)),
      pch = c(21,24,1,2),
      pt.bg = c("black","black","white", "white"),
      bty = "n")

abline(male.lm,lty=1)
abline(female.lm,lty=2)

```



```

# Contraste sin tener en cuenta el área
attach(lions)

```

```

## The following objects are masked from lions (pos = 3):
##
##   age, area, prop.black, sex

```

```

n1<-sum(sex=="F")
n2<-sum(sex=="M")
y<-c(prop.black[sex=="F"],prop.black[sex=="M"])
x<-matrix(numeric(4*(n1+n2)),ncol=4)
male.lm<-lm(prop.black~age,data=lions,subset = sex=="M")
female.lm<-lm(prop.black~age,data=lions,subset = sex=="F")
x[1:n1,1:2]<-model.matrix(female.lm)
x[(n1+1):(n1+n2),3:4]<-model.matrix(male.lm)
general.lm<-lm(y~0+x)

x0<-matrix(numeric(3*(n1+n2)),ncol=3)
x0[,1]<-x[,1]
x0[,2]<-x[,3]
x0[,3]<-x[,2]+x[,4]
h0.lm<-lm(y~0+x0)

anova(h0.lm,general.lm)

```

```

## Analysis of Variance Table
##
## Model 1: y ~ 0 + x0
## Model 2: y ~ 0 + x
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     102 1.9603
## 2     101 1.8954  1   0.06488 3.4573 0.06589 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

# Contraste para Serengeti
serengeti<-subset(lions,area=="S")

n1<-sum(serengeti$sex=="F")
n2<-sum(serengeti$sex=="M")
y<-c(serengeti$prop.black[serengeti$sex=="F"],serengeti$prop.black[serengeti$sex=="M"])
x<-matrix(numeric(4*(n1+n2)),ncol=4)
male.s.lm<-lm(prop.black~age,data=serengeti,subset = sex=="M")
female.s.lm<-lm(prop.black~age,data=serengeti,subset = sex=="F")
x[1:n1,1:2]<-model.matrix(female.s.lm)
x[(n1+1):(n1+n2),3:4]<-model.matrix(male.s.lm)
general.lm<-lm(y~0+x)

x0<-matrix(numeric(3*(n1+n2)),ncol=3)
x0[,1]<-x[,1]
x0[,2]<-x[,3]
x0[,3]<-x[,2]+x[,4]
h0.lm<-lm(y~0+x0)

anova(h0.lm,general.lm)

```

```

## Analysis of Variance Table
##
## Model 1: y ~ 0 + x0

```

```
## Model 2: y ~ 0 + x
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      81 1.4037
## 2      80 1.3663  1  0.037385 2.189 0.1429

# Contraste para Ngorongoro
ngoro<-subset(lions,area=="N")

n1<-sum(ngoro$sex=="F")
n2<-sum(ngoro$sex=="M")
y<-c(ngoro$prop.black[ngoro$sex=="F"],ngoro$prop.black[ngoro$sex=="M"])
x<-matrix(numeric(4*(n1+n2)),ncol=4)
male.n.lm<-lm(prop.black~age,data=ngoro,subset = sex=="M")
female.n.lm<-lm(prop.black~age,data=ngoro,subset = sex=="F")
x[1:n1,1:2]<-model.matrix(female.n.lm)
x[(n1+1):(n1+n2),3:4]<-model.matrix(male.n.lm)
general.lm<-lm(y~0+x)

x0<-matrix(numeric(3*(n1+n2)),ncol=3)
x0[,1]<-x[,1]
x0[,2]<-x[,3]
x0[,3]<-x[,2]+x[,4]
h0.lm<-lm(y~0+x0)

anova(h0.lm,general.lm)

## Analysis of Variance Table
##
## Model 1: y ~ 0 + x0
## Model 2: y ~ 0 + x
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      18 0.42171
## 2      17 0.28434  1  0.13737 8.2129 0.01071 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Contraste para machos
males<-subset(lions,sex=="M")

n1<-sum(males$area=="S")
n2<-sum(males$area=="N")
y<-c(males$prop.black[males$area=="S"],males$prop.black[males$area=="N"])
x<-matrix(numeric(4*(n1+n2)),ncol=4)
sere.lm<-lm(prop.black~age,data=males,subset = area=="S")
ngoro.lm<-lm(prop.black~age,data=males,subset = area=="N")
x[1:n1,1:2]<-model.matrix(sere.lm)
x[(n1+1):(n1+n2),3:4]<-model.matrix(ngoro.lm)
general.lm<-lm(y~0+x)

x0<-matrix(numeric(3*(n1+n2)),ncol=3)
x0[,1]<-x[,1]
x0[,2]<-x[,3]
x0[,3]<-x[,2]+x[,4]
```

```
h0.lm<-lm(y~0+x0)
```

```
anova(h0.lm,general.lm)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Model 1: y ~ 0 + x0
```

```
## Model 2: y ~ 0 + x
```

```
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
```

```
## 1      29 0.39136
```

```
## 2      28 0.33170  1  0.059652 5.0354 0.03292 *
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

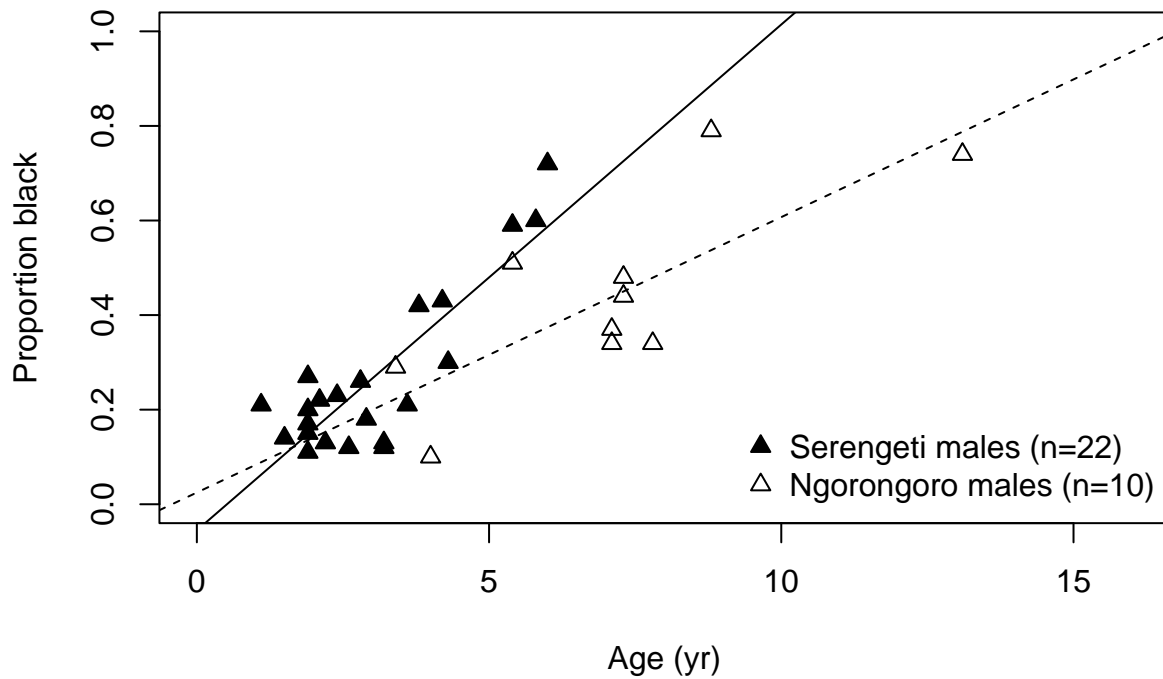
```
# Representación de los machos según área
```

```
plot(males$prop.black ~ males$age,  
     pch = ifelse(males$area == "S", 24, 2),  
     col = ifelse(males$area == "S", "black", "black"),  
     bg = ifelse(males$area == "S", "black", "white"),  
     data = males,  
     ylab = "Proportion black",  
     xlab = "Age (yr)",  
     ylim = c(0,1),  
     xlim = c(0,16))
```

```
legend("bottomright",  
      legend = c("Serengeti males (n=22)",  
                  "Ngorongoro males (n=10)"),  
      col = c(rep("black",4)),  
      pch = c(24,2),  
      pt.bg = c("black", "white"),  
      bty = "n")
```

```
abline(sere.lm,lty=1)
```

```
abline(ngoro.lm,lty=2)
```



```
# Modelo para las predicciones de las hembras
females<-subset(subset(lions,age<=10),sex=="F")
females.lm<-lm(age~asin(sqrt(prop.black)),females)
lion0<-data.frame(prop.black=0.5)
PI95<-predict(females.lm,lion0,interval = "prediction",level=0.95)
PI75<-predict(females.lm,lion0,interval = "prediction",level=0.75)
PI50<-predict(females.lm,lion0,interval = "prediction",level=0.5)

PI_df <- round(data.frame(
  prop.black = 0.5,
  age = PI95[1],
  lower95 = PI95[2],
  upper95 = PI95[3],
  lower75 = PI75[2],
  upper75 = PI75[3],
  lower50 = PI50[2],
  upper50 = PI50[3]
),2)
knitr::kable(PI_df)
```

prop.black	age	lower95	upper95	lower75	upper75	lower50	upper50
0.5	5.34	2.95	7.74	3.96	6.73	4.53	6.16