Regresión, modelos y métodos: Prueba de evaluación continua 1

Aràntzazu Alonso Carrasco

2023-05-03

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:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i$$

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$= 6.7269394 \times 10^{-5}$. 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

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

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.

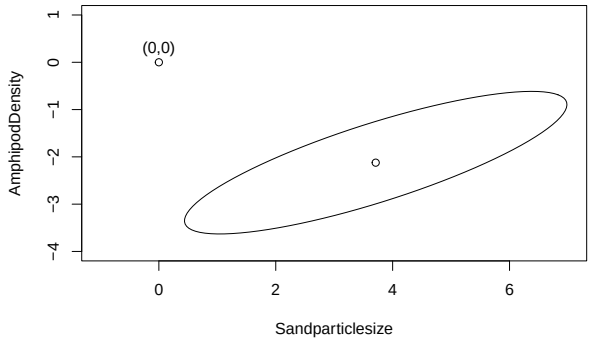
Las hipótesis escritas en forma paramétrica son las siguientes:

$$H_0 : \beta_1 = \beta_3 = 0$$
$$H_1 : \beta_i \neq 0 \text{ para algún } i = 1, 3$$

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:

$$H_0 : \beta_{Sandparticlesize} = \beta_{AmphipodDensity} = 0$$
$$H_1 : \beta_{Sandparticlesize} \neq 0 \text{ o } \beta_{AmphipodDensity} \neq 0$$

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

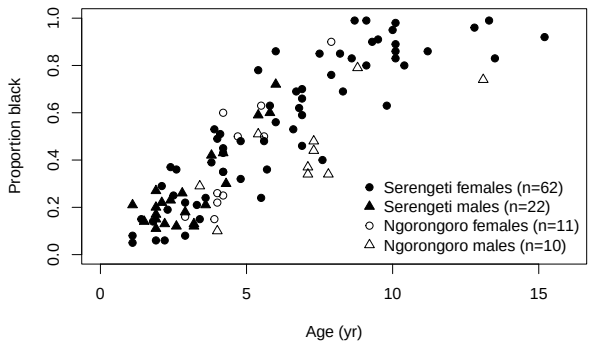
En el trabajo de Whitman et al. (2004) se estudia, entre otras cosas, la relación entre la edad de los leones y la proporción oscura en la coloración de sus narices. En el archivo `lions.csv` disponemos de los datos de 105 leones machos y hembras de dos áreas de Tanzania, el parque nacional de Serengueti 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.**

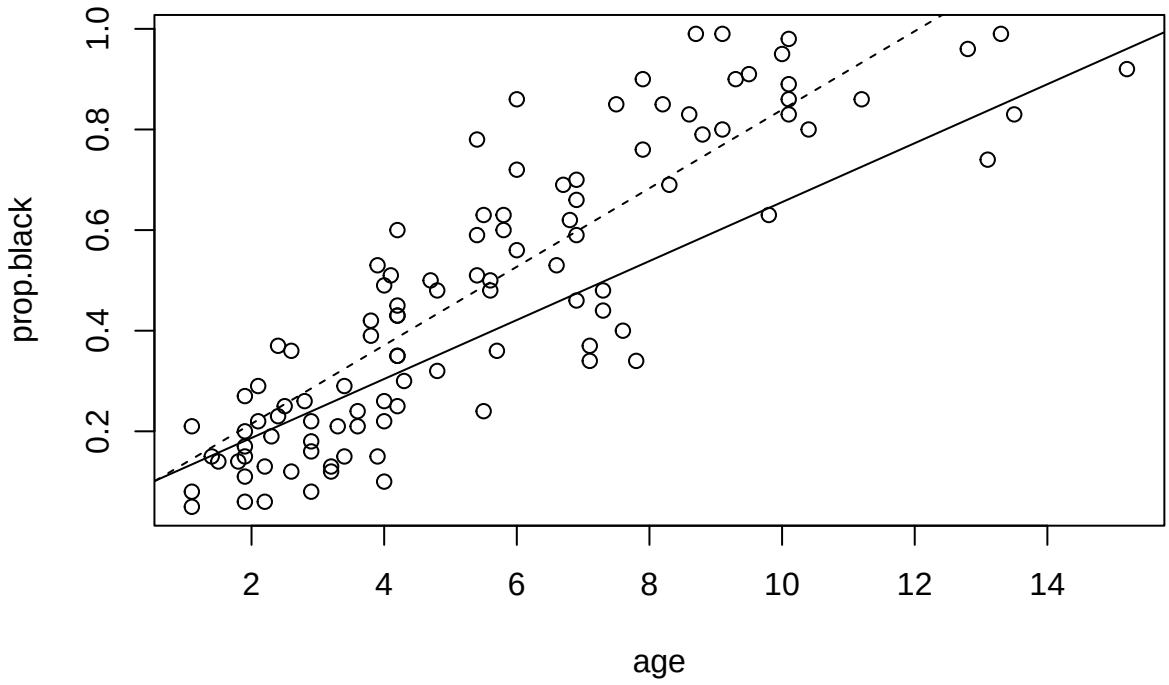
```
male.lm<-lm(prop.black~age,data=lions,subset = sex=="M")
female.lm<-lm(prop.black~age,data=lions,subset = sex=="F")
summary(male.lm)
```

```
##
## Call:
## lm(formula = prop.black ~ age, data = lions, subset = sex ==
##     "M")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20406 -0.07758 -0.01750  0.07913  0.29876
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.069696   0.041956   1.661    0.107
## age         0.058591   0.008307   7.053 7.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1238 on 30 degrees of freedom
## Multiple R-squared:  0.6238, Adjusted R-squared:  0.6113
## F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
```

```
summary(female.lm)
```

```
##
## Call:
## lm(formula = prop.black ~ age, data = lions, subset = sex ==
##     "F")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32503 -0.10638 -0.00395  0.10829  0.33292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.058858   0.035665    1.65    0.103
## age         0.078038   0.005178   15.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1422 on 71 degrees of freedom
## Multiple R-squared:  0.7618, Adjusted R-squared:  0.7585
## F-statistic: 227.1 on 1 and 71 DF,  p-value: < 2.2e-16
```

```
plot(prop.black~age,data=lions)
abline(male.lm,lty=1)
abline(female.lm,lty=2)
```



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
```

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

```
##        value
## 6.726939e-05
```

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

```
## [1] 0.013
```

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

```
## [1] 0.05
```

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

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

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

```
##         2.5 %        97.5 %
## -3.1254068143  0.0007019125
```

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

```
## [1] -1.5624
```

```r
# Calculamos los VIF
vif(cici.lm)
```

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

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

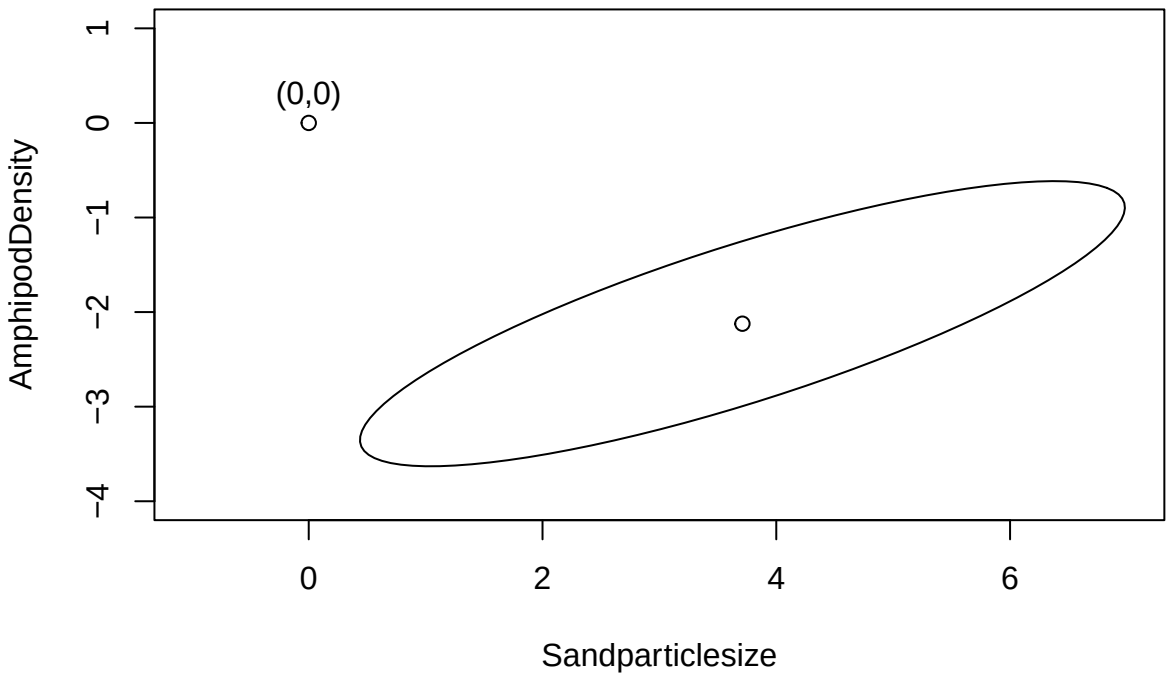
```
##
## Call:
## lm(formula = BeetleDensity ~ Sandparticlesize + AmphipodDensity,
##     data = cici)
##
```

```
## 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")
```