Linear regression

October 16, 2019

1. Metabolism of a fish according to salinity

The dataset `sardinella.csv` comes from a study by Wohlschlag (1957), "Differences in metabolic rates of migratory and resident freshwater forms of an Arctic Whitefish". It contains weight (*log_weight*) and oxygen consumption (*log_O2*) measurements for individuals of *Coregonus sardinella* caught in freshwater or marine environments.

```
sardinella <- read.csv("sardinella.csv")
str(sardinella)
```

```
## 'data.frame':    22 obs. of  3 variables:
##  $ environment: Factor w/ 2 levels "freshwater","marine": 2 2 2 2 2 2 2 2 1 1 ...
##  $ log_O2     : num  1.59 1.4 1.47 1.66 1.55 ...
##  $ log_weight : num  2.5 2.04 2.15 2.35 2.24 ...
```

a) Estimate the additive effects of environment and weight on the oxygen consumption of this fish. How do you interpret each of the parameters of the model?

b) Repeat the model in (a) with a standardized version of the predictor *log_weight* (*norm_weight*). What is the interpretation of the coefficients now?

c) Repeat the model in (b) by adding the interaction between the weight (normalized) and the environment. Is this interaction meaningful? What is the interpretation of the coefficients?

2. Diversity of plants on British Isles

The dataset `britain_species.csv` comes from the study of Johnson and Simberloff (1974), "Environmental determinants of island species numbers in the British Isles". These data indicate the number of vascular plant species (*species*) for 42 British isles according to different predictors: area in km^2, elevation in m, number of soil types, latitude and distance from Great Britain in km (*dist_britain*).

```
iles <- read.csv("britain_species.csv")
str(iles)
```

```
## 'data.frame':    42 obs. of  7 variables:
##  $ island      : Factor w/ 42 levels "Ailsa","Anglesey",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ area        : num  0.8 712.5 429.4 18.4 31.1 ...
##  $ elevation   : int  340 127 874 384 226 1343 210 103 143 393 ...
##  $ soil_types  : int  1 3 4 2 1 16 1 3 1 1 ...
##  $ latitude    : num  55.3 53.3 55.6 57 60.1 54.3 57.1 56.6 56.1 56.9 ...
##  $ dist_britain: num  14 0.2 5.2 77.4 201.6 ...
##  $ species     : int  75 855 577 409 177 1666 300 443 482 453 ...
```

a) Suppose that a theory predicts that the number of species (S) depends on the area of an island (A) according to the following equation, where c and z are parameters to be determined:

$$S = cA^z$$

Use a linear model to test the hypothesis that the number of vascular plant species follows that equation with an exponent $z = 0.25$ (one quarter).

Hint: Assume that the estimated value of z follows a normal distribution. From the estimated value of z and its standard error, use the formula seen during the second class to calculate the confidence interval.

$$(\hat{z} + t_{df,\alpha/2} SE, \hat{z} + t_{df,1-\alpha/2} SE)$$

In this formula, SE is the standard error, α is the significance threshold you choose (ex: 0.05) and df is the number of degrees of freedom of the t distribution, that you can determine from the summary of the regression.

b) Now estimate the following model, where the number of species depends both on the area of the island and its distance from Great Britain, on a logarithmic scale. You will first have to exclude the island of Britain from the dataset.

$$\log(species) \sim \log(area) + \log(dist\_britain)$$

c) Using the model in (b), give a 90% prediction interval for the number of species for (i) a 1-km^2 island at a distance of 5 km and (ii) an 40-km^2 island located at a distance of 20 km.

Note:

- Change the % of the prediction interval with the `level` argument of `predict`.
- Since the response of the model is `log(species)`, the result of `predict` will be on a logarithmic scale.