Simple linear regression

October 6, 2020

Answers for this lab must be submitted before **October 13th at 5pm on Moodle**. In your answer for each question, please include the R code used (if applicable) and the results obtained.

1. Growth of a pine species

The `Loblolly` data frame included with R indicates the height in feet (*height*) measured at six values of age (3 to 25 years) for 14 individuals of the species *Pinus taeda*. Each individual is indicated by a different value in the *Seed* column.

```
head(Loblolly)
```

```
##    height age Seed
## 1    4.51   3  301
## 15  10.89   5  301
## 29  28.72  10  301
## 43  41.74  15  301
## 57  52.70  20  301
## 71  60.92  25  301
```

a) Estimate and interpret the parameters of a linear model of height versus age for these pines. Does the intercept make biological sense for this model?

b) Add a new column in the data frame containing the residuals of the model in (a), obtained with the `residuals` function, then make a scatterplot of those residuals as a function of the tree (*Seed*). Based on this result, which assumption of linear regression could be invalid here?

c) From the diagnostic graphs of the model in (a), indicate whether the assumptions of linearity, equality of variances and normality appear to be met.

d) Explain how each of the unmet assumptions identified in (b) and (c) affects the validity of the model's conclusions. In other words, how might the model's estimates and predictions differ from reality?

2. Plant diversity in the British Isles

The data table britain_species.csv comes from the study by Johnson and Simberloff (1974), "Environmental determinants of island species numbers in the British Isles". These data indicate the number of vascular plant species for 42 British Isles according to different predictors, including the area of the island in km^2.

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

```
## 'data.frame':    42 obs. of  7 variables:
##  $ island      : chr  "Ailsa" "Anglesey" "Arran" "Barra" ...
##  $ 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 island area (A) according to the following equation, where c and z are parameters to be esimated:

$$S = cA^z$$

How can you transform this equation into a linear model?

b) Fit the model in (a) to the data and check the diagnostic graphs. What is the estimate of z?

c) Suppose the theory predicts that $z = 0.25$. Calculate a confidence interval for z and determine whether or not this assumption can be rejected at a significance level of $\alpha = 0.05$.

d) From the model in (b), give a 90% prediction interval of the number of species for (i) an island of 0.5 km^2 and (ii) an island of 20 km^2.

Notes:

- Change the % of the prediction interval with the `level` argument of `predict`.
- If the model response is a transformation of the `species` variable, you can apply the inverse transformation to the lower and upper bounds of the intervals obtained with `predict` to find the corresponding number of species.