homework2

Trevor Caughlin

January 20, 2018

For Questions 1 and 2, use `glm` to run binomial regressions on each of the two data sets. As part of the ouput, include:

a) a `plot` of raw data with the best-fit regression line (use `curve`) overlaid on the plot (note: as long as it has raw data and a curve I'm fine if you use ggplot or another package besides base graphics).

b) Point estimates for slope and intercept parameters, including a verbal description of the baseline and effect size for these parameters

c) Confidence intervals for slope and intercept

Question 1: Dataset: "SEEDLING_SURVIVAL.csv"

These data represent the annual survival rates of 1435 tagged tree seedlings in Huai Kha Khaeng Wildlife Sanctuary, Thailand. Run separate glm models for:

a) effect of height on seedling survival
b) effect of light on seedlings survival

Is height or light a stronger predictor of seedling survival?

Question 2: Dataset: "seeds.csv"

These data represent results from a seed addition experiment where 5, 15, or 45 seeds were added to plots in the Huai Kha Khaeng Wildlife Sanctuary. The column "seeds" represents the number of seeds added. The column "recruits" represents the number of germinating seedlings that emerged from the added seeds. Choose one of the following predictor variables:

a) DBH: total amount of adult conspecific trees in a 10 m radius around plots

b) Seedlings: total number of conspecific seedlings in plots (prior to seed addition)

c) Grass: amount of grass in plots

d) Light: light measured with hemispherical photographs

Run a binomial regression with proportion of recruits as a response variable and DBH, Seedlings, Grass, or Light as a predictor variable.

Do your results support the hypothesis that your selected predictor variable has a significant effect on seedling germination?

Hints:

(1) to plot proportional data, use the number of successes divided by the total number of trials as the variable on the y-axis

(2) to do a proportional glm in R, you must first create a proportional variable, with one column representing successes, and one column representing failures:

```
response<-cbind(recruits, seeds-recruits)
```

This new variable is your input to the binomial regression:

```
glm(response~predictor,family=binomial)
```

Note: data from question 1 and 2 is published in following paper:

Caughlin, T. T., J. M. Ferguson, J. W. Lichstein, S. Bunyavejchewin, and D. J. Levey. 2013. The importance of long-distance seed dispersal for the demography and distribution of a canopy tree species. Ecology 95:952–962.

Question 3: "mosquito_data.csv"

These data represent observations of how many adult mosquitoes emerged from mosquito eggs in pools of water in a tropical urban area. This species of mosquito carries dengue fever. The objective of the study is to predict habitat quality for larval mosquitos, with organic detritus in pools of water as a predictor variable.

a. Plot the data. (Again, note that plotting the proportion of eggs is necessary, since the number of eggs varies)

b. Add curves for the following equations to predict mosquito emergence to your plot:

Polynomial: $1.44 - 0.19x - 0.21x^2 + 0.04x^3$

Ricker: $10xe^{-2x}$

c. How are the biological implications of the polynomial model different from the Ricker model?
d. use `dbinom` to calculate the likelihood of both models (with given parameter values). Hint: use `log=T` as an argument in `dbinom` before taking `sum` of dbinom output

e. According to `dbinom, the` likelihood of the data is higher for which model?

For more details on likelihood, see Chapter 6 and Chapter 8 in Ben Bolker's excellent "Ecological Models and Data in R" book:
https://ms.mcmaster.ca/~bolker/emdbook/book.pdf

For further reading on power analyses (question below) see the section in Bolker's book titled "Power Analysis" beginning on page 209.

Question 4: Power analysis

Power analysis asks how much data is required to detect an effect. In this question, you will conduct two power analyses using stochastic simulation. The first power analysis will be for a linear regression (assume normally distributed data) and the second power analysis will be for binary data (assume logit-link function).

We will use for loops to simulate data and run our analysis. Here is an example of a for loop:

```
empty_vector<-rep(NA,times=100)

for(i in 1:100) {
    empty_vector[i]<-rnorm(n=1,mean=0,sd=1)
  }
```

We initialize the for loop with an empty vector for the for loop to fill in Here, the object `empty vector` is just `NA` repeated 100 times. We then fill the vector with draws from a normal distribution.

Here are the steps for the power analysis:

1. Decide on values for the true slope, intercept, and standard deviation (for the Normal values). If you can get these values from the literature, that would be particularly awesome.
2. Create a predictor variable, using either `runif` or `seq` functions.
3. Create a vector of sample size. The easiest way to do this is simply to create a vector from 3 to a maximum sample size:

   ```
   sample_size<- c(3:10000)
   ```

   Note that in this example, the maximum sample size is 10000. You should adjust the sample size to be realistic for whatever data you are simulating. N=10000 might be realistic for Landsat pixels, but not realistic for monitored bird nests.

4. Create an empty vector to fill in with results from the power analysis. Note that the number of elements in the vector needs to be equal to the length of the sample size vector (in this case, 10000-3).

```
power_vector<-rep(NA,times=9997)
```

5. Write a for loop that fills in the empty vector with results from simulated analyses. This for loop will need to do three things:
    1. Simulate data from either a Normal or a Binomial distribution (use `rnorm` or `rbinom`), with specified intercept and slope parameters.
    2. Analyze simulated data with `glm`
    3. Extract output of interest (p-value of slope, or slope estimate).
6. Plot results, with the filled-in vector on the y-axis and sample size on the x-axis.

Using this framework, answer the following questions:

a. How many samples do you need to accurately estimate the slope parameter in a binomial vs. linear regression? (remember: you know the true value of the slope parameter) Use MSE to calculate the accuracy and precision of your estimate vs. the real value: $\left(slope - \widehat{slope}\right)^2$

b. How many samples do you need to ensure a p_value<0.05 for binomial vs. linear regression? Note: you can extract p-values for the slope using the following code:

```
summary(model)$coefficients[2,4]
```

c. In general, why is statistical power generally higher for continuous than discrete response variables?