Cole Brookson - R Software Manual Developer Application Materials

Cole B. Brookson

2022-04-14

Dear Drs. Sztepanacz & Riskin,

I am writing to apply to the position of R software manual developer within the Department of Ecology and Evolutionary Biology at the University of Toronto. As a previous undergraduate in the EEB department, I am accutely aware of the need in the undergraduate curriculum for this type of product, and having spent the past 4 years working intensely in R both in research and teaching capacities, I believe I have the skills and aptitude to build a dynamic, engaging, and adaptable product, usable in the department for years to come.

I have significant experience teaching R to students at beginner to advanced levels. Along with colleagues at the University of Alberta, I developed and delivered free, open-source workshop series introducing over 150 Biological Sciences students to both R and Python (https://colebrookson.github.io/r-for-biology/). I have developed and delivered workshops in reproducible research and coding practices for graduate students at the University of Alberta, and as the TA for Marine Population Ecology & Dynamics (An intensive field course held at Bamfield Marine Sciences Centre), I completely revamped the statistical and programming components of the courses, and built an open-source website to host the course online. I am also a certified Data/Software Carpentries instructor, and completed graduate teaching and learning training through the University of Alberta's Faculty of Graduate Studies and Research. I have significant experience with a variety of other programming languages which lends context to my teaching approach to R (Julia, Python, C++, HTML, SQL, and JavaScript), and I am an expert in reproducible technologies such as version control with Git/GitHub, containerization with Docker & Singularity, and workflow implementation with bash, Make, Snakemake and others.

I am incredibly passionate about computational education, and having struggled myself when first learning R as an EEB undergraduate, the opportunity to contribute to this type of product would be the ultimate full-circle experience for me, allowing me to give back to the EEB community that formed me as a student and offered me my first exposure to the wonderful world of R.

As I am currently completing my MSc and will be on contract as an Ecological Data Analyst for the Kwikwasut'inuxw Haxwa'mis, 'Namgis, and Mamalilikulla First Nations until August 31, 2022, I would not be able to start the position until Aug 01 at earliest, and would prefer a September 01 start date, however, I understand that might not fit with the timeline of this project, and would be very open to discussing possible solutions (i.e. part time earlier start dates).

Thank you for your consideration, Cole Brookson —

A brief foray into regression analysis

Let's now explore using some regression techniques in R as a way to investigate some common ecological problems. As usual, we will separate the domain problem from the implementation.

Regression as Concept

To refresh our memories, **regression** is a method of analysis we can use for hypothesis testing. At the simplest level, this idea of regression is simply *a measure of the relation between the mean value of one variable, and the corresponding values of other variables.* It is most common to relate these to the ideas of explanatory variables (denoted X) and the independent variable, Y. Since, as the name implies, linear regression means simply fitting a straight line through some points, what we are actually doing in practice most of the time, is finding the slope of that straight line, along with the y-intercept. Assuming for a moment the y-intercept isn't of interest, we can think of the simplest regression framework as being

$$Y \sim \beta X,$$

where out goal is to ask how X explains Y, by estimating or "fitting" some value of β.

Implementation in R

So as usual, we will start by loading the required packages for this activity. To perform a simple linear regression, we can use the `stats` package which comes pre-loaded. We will want two additional packages, we'll use the `tidyverse` package to organize our data and plot, then we'll also use the `lterdatasampler` package to load in some data for this task.

```
# install packages if needed
#install.packages("tidyverse")
#remotes::install_github("lter/lterdatasampler")

# load in libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lterdatasampler)
```

Our example dataset we are using today is the `pie_crab` dataset, which includes data on the size of fiddler crabs, alongside environmental variables. Let's first take a quick look at our dataset.

```
head(pie_crab)
```

```
## # A tibble: 6 x 9
##   date       latitude site   size air_temp air_temp_sd water_temp water_temp_sd
##   <date>        <dbl> <chr> <dbl>    <dbl>       <dbl>      <dbl>         <dbl>
## 1 2016-07-24       30 GTM    12.4     21.8        6.39       24.5          6.12
## 2 2016-07-24       30 GTM    14.2     21.8        6.39       24.5          6.12
```

```
## 3 2016-07-24       30 GTM    14.5     21.8        6.39       24.5          6.12
## 4 2016-07-24       30 GTM    12.9     21.8        6.39       24.5          6.12
## 5 2016-07-24       30 GTM    12.4     21.8        6.39       24.5          6.12
## 6 2016-07-24       30 GTM    13.0     21.8        6.39       24.5          6.12
## # ... with 1 more variable: name <chr>
```

To avoid p-hacking ourselves (see Appendix 1 for a review of p-hacking), we need to first develop a hypothesis that we're going to use our linear regression to test. A classic theory in ecology is Bergmann's Rule. Given the premise of the rule, we might expect this to happen within species, so let's test this with our fiddler crabs!

> ***NULL HYPOTHESIS:*** To ensure we are going about this properly, let's state our biological null hypothesis. *Our null hypothesis is that there is NO significant positive relationship between latitude and the size of fiddler crabs.*. Now, let's think about what this means statistically. In our regression framework, remember we're thinking about **slope** of a line. A lack of a relationship is denoted by a slope of zero, but to reject our null hypothesis, we need a *positive* relationship since Bergmann's Rule states that body size *increases* as temperature (and therefore latitude) decreases! We can *reject our null hypothesis* (recall we can never accept a hypothesis, only reject the null!!) if the slope of our regression line is significantly different than zero in the positive direction.

Null Hypothesis To ensure we are going about this properly, let's state our biological null hypothesis. *Our null hypothesis is that there is NO significant positive relationship between latitude and the size of fiddler crabs.*. Now, let's think about what this means statistically. In our regression framework, remember we're thinking about **slope** of a line. A lack of a relationship is denoted by a slope of zero, but to reject our null hypothesis, we need a *positive* relationship since Bergmann's Rule states that body size *increases* as temperature (and therefore latitude) decreases! We can *reject our null hypothesis* (recall we can never accept a hypothesis, only reject the null!!) if the slope of our regression line is significantly different than zero in the positive direction.