Stat 021 Final Project Instructions

Suzanne Thornton

Due: Friday, Dec. 13, 11:59pm

Instructions for the final project

These step by step instructions follow closely the guide to writing an R package found here: https://hilaryparker.com/2014/04/29/writing-an-r-package-from-scratch/

First

Install and call (with the *library()* function) all required packages as shown below.

```
install.packages("devtools")
library("devtools")
devtools::install_github("klutometis/roxygen")
library(roxygen2)

#change to wherever you want to save your R package:
setwd("/Users/m/Google Drive Swat/Swat docs/Stat 21/Final project")
create("Package.Name") #name your package whatever you want
```

Navigate to the location of you R package. Once you run the code above, R will automatically create a description file for your package. Edit this file to include the author's name, affiliation, and email; list any dependent packages (e.g. *tidyverse*); a description of the included data set including variable types and units etc. You will be assigned a unique data set that includes at least one numerical response and one numerical predictor and one categorical predictor. If you need assistance getting the data into a usable format, please let me know and I'll help you with this step. Also if you have a different data set of your own that you'd like to use that fine too as long as it meet the previously mentioned specifications.

Second

Include your data set in your R package by following the guide here: https://kbroman.org/pkg_primer/pages/data.html. To follow these instructions you will need to have a sub-folder within your package called "Data". This is where you will create and save the new .Rdata data file. Note that in order to produce documentation for your data set you must create an R script file containing only comments for the description.

Create a R script file for each of the required functions listed below. This will be the part of this assignment that takes the longest since I am asking you to create several functions, each with a different purpose.

Third

The *roxygen2* package makes creating all the documentation for your R package really easy. All you need to do is go back to each R script document for each function and add some comments to the beginning of these files using the following format.

```
#' <Name of the function>
#'
#' <Description of what this function does>.
#' @param example_parameter what is the first parameter
```

```
#' @param example_parameter what is the second parameter... and so on
#' @return describe the output of your function here
#' @keywords what are some keywords for this function? (e.g. regression, linear model, etc)
#' @export
#' @examples
#' code for a simple example to test the function for example...
#' x <- rnorm(500, 0, 2)
#' y <- x + rnorm(500,0,5)
#' name.of.the.function(x,y,alpha=0.05)
```

Note each of your six functions should contain all of the items listed in above.

For each function, these comments should describe the purpose of the function, the expected input and output. You are allowed to rely on any functions that exist in base R or in the *tidyverse* package.

Fourth

The way your project will be graded is by me making sure that it works on your data set and on another (arbitrary) data set that includes at least two numerical variables.

To hand in

Ultimately you have one item to email to me: a compressed folder containing everything needed to install and run your R package. The way I will be grading your projects is installing your package on my R instance, calling it to my R library, testing out each of the functions in your package on the data set included in your package, and finally, testing out each of the functions on another data set (of my choice). You may want to do the same to make sure that everything works.} Please make sure that you send your R package to me as a zipped folder and that in your emailed submission the subject heading is "Final Project Stat 21".

Some tips for this assignment

Note that for this assignment you will be writing R script files (that end in .R) and **not** writing R Markdown files (that end in .Rmd). If you're having trouble understanding my instructions, see if you can follow along with the example package demonstrated in the link at the top of this document. You are allowed to work with your classmates on this project but you must show me that your package works on your assigned data set as well as on another data set.

A strategy I recommend for completing this assignment is to start by writing the code for your functions first, and testing them out on some data set we've used in class (for example) to make sure they work as you intend. Then once you have six R script files for each of the required functions, test them out on your assigned data set, debugging as necessary. Then follow the Third and Fourth steps above to combine everything into an R package.

Also, don't wait until the week this assignment is due to get started. I will **not** have extra office hours that week and you will likely find yourself up all night debugging code and getting very frustrated. If your submitted package does not run, you will automatically lose 10 points on this project (worth 100 points total and worth 25% of your final grade).

Required functions

- *check.varb.types* - this function will take a vector input and will output the vector type (e.g. double, integer, character, factor).

In all of the subsequent functions you must check that the input consists of one data.frame and one numerical vector whose length matches the length of the columns in the data frame.

- *good.fit1* - this function will take a data frame of predictor variables, a vector of the response variable, and a significance level (α) as the input and will output the p-value of the overall F test for the regression model fit. Specifically the output should be a phrase such as "At the α =whatever significance level we find that the overall goodness of fit of this regression model is satisfactory/unsatisfactory with a p-value of whatever." (Hint: use the pf() function in R to calculate the p-value associated with the F-statistic output from the lm() function.)
- *good.fit2* - this function will take a data frame of predictor variables and a response vector as the input and will output the adjusted R squared value for the regression model fit. Specifically the output should be a phrase such as "Whatever % of the variability in the response can be explained by the predictors."
- *estimated.lm* - this function will take a data frame of predictor variables and a response vector as the input and will output the estimated coefficients and their standard errors for all predictor variables (**not** including the intercept).
- *sig.predictor.test* - this function will take a data frame of predictor variables, a response vector, an integer referencing which numerical predictor variable you want to test the significance of, and a confidence level as the input. The output will be the result of the individual t-test for the statistical significance of that predictor variable. (Note I am NOT asking you to produce the test results for a categorical variable.) Specifically the output should be a phrase such as "At the α =whatever confidence level we find that the the individual effects of this predictor variable are/are not significantly contributing to the regression model".
- *resid.plot* - this function will take a data frame of the predictor variables and a vector of the response variable as the input and will output a simple residual plot for the estimated regression equation. The residual plot must be labeled (with a main title, and with the horizontal and vertical axes labeled) and must include a horizontal line at zero for reference.