Stat 021 Homework 7

Suzanne Thornton

Due: Saturday, Nov. 16, 12:00pm

Instructions: A **pdf** version of your homework must be submitted to Gradescope by **noon** on the due date. The course passcode is **MPKJ4Z**. If you are having trouble getting your *.Rmd* file to compile, you need to get help with this **before** the due date.

You are allowed to hand in **only one** late homework assignment throughout the semester. If you need to hand in this particular assignment late, you must let me know via email by noon on the due date.

You are encouraged to study with your peers to help complete the homework assignments but no copying is allowed. If I see that two or more homework assignments are copied, all students involved will receive a grade of 0 on that assignment and will forfeit (perhaps retroactively) the opportunity to hand in a late homework.

Q1) Read the article, "Scientists rise up against statistical significance" at https://www.nature.com/articles/d41586-019-00857-9.

(a) The article claims, "…researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment on some measured outcome)." Explain why failing to reject the null hypothesis does not prove that there is no effect. What does failing to reject the null hypothesis really mean instead? (1 point)

(b) In the graphic "Beware false conclusions", results are shown from two studies: one that found "significant" results, and another that found "non-significant" results. The article claims that it is "ludicrous" to say that the second study found "no association." Briefly explain why this is the case. (1 point)

(c) Regarding the same two studies in part (b), the article claims that it is "absurd" to say that the two studies are in conflict, even though one was "significant" and the other was "not significant". Briefly explain why this is the case. (1 point)

(d) In the section titled "Quit categorizing", the article claims that, "Statistically significant estimates are biased… Consequently, any discussion that focuses on estimates chosen for their significance will be biased." Briefly explain why this is the case. (1 point)

(e) Now that you've read this article, going back to Q1 of HW 6, redo your answer for part (e) and explain if you would change your final model suggestion or not and why. (1 point)

Q 2) Create an R function that takes two numerical vectors as it's input and fits a SLR model using the second vector to predict the first one. The output of this function will be a phrase that either says "Good fit" or "Bad fit" depending on the R-squared value of the model. Let's say, for simplicity, that a model is a good fit if the R-squared value is 0.60 or higher. To get you started, you can use the following code as a template:

```r
my.SLR.fun <- function(vec1, vec2){

  ## add you own code here

  if(){return("Good fit")}
  ##figure out what logical test needs to go into the if statment

  else{return("Bad fit")}
}
```

Then test your function:

```r
#Should return "Bad fit"
x <- rnorm(10,2,1.3)
y <- rnorm(10,0,2)
my.SLR.fun(y,x)

#Should return "Good fit"
x <- rnorm(10,2,1.3)
y <- x + rnorm(10,0,0.8)
my.SLR.fun(y,x)
```

Q 3) The dataset uploaded to Moodle called "airplanes.csv" was collected from national publication advertising the sale of used aircraft in the early 1990s. The variables included in this dataset include the year of the aircraft, TT (total flight time in hours), SMOH (hours since major overhaul), DME (distance measuring equipment), LORAN (long-range navigation based on satellite communication), HP (engine horsepower), paint (new or recent paint job), and price. The variables DME, LORAN, and paint are each binary categorical variables that indicate whether the corresponding item was mentioned as being present in the ad. The price is given in thousands of dollars.

(a) Read in the dataset from Moodle and be sure to double check that each of the variable types are what you want them to be. (Note: Please do **not** print out the entire data set in your final pdf document, instead just double check the data import process on the side on your own. Also, double check the data after reading it into R, get rid of any rows of only NA values and make sure each variable is the correct variable type.) Make a scatterplot matrix using only the numerical variables. Comment on any notable features or patterns (or lack of thereof). What is notable about the variable HP? What do you think you should do with this variable? (2 points)

(b) Which data point appears to be an outlier? Can you guess why it might be an outlier? For the purposes of this assignment, let's delete this point from any further analyses since we don't know its correct value. Delete this outlier observational unit from the data and make a new scatterplot matrix of the quantitative variables. (1 point)

(c) Fit a model with all of the predictor variables except SMOH. What is the value of R-squared? What is the estimated error variance? What is the interpretation of the latter? Which variables appear significant, and which do not? (2 points)

(d) It seems plausible that TT and year might be collinear: older planes may have been flown more. Based on the scatterplot matrix and the estimated correlation between these variables, does this appear to be the case? (1 point)

(e) For our current model, do the regression assumptions appear to be satisfied? Make a residual plot (residuals vs predicted) and an Normal probability plot and comment on whether you think the assumptions are satisfied or whether there may be cause for concern. (2 points)

(f) Summarize your findings from this model. In particular, what do the regression coefficients mean? What is the estimated error variance and what does it represent? What is the R-squared? (2 points)