Exam 2, Take Home

Your Name

11/8/2021

You are allowed to use your book, other books, notes and the internet for coding help to complete this exam. You are not allowed to use someone else's solution or data analysis with this same data set. You may NOT talk to anyone in person or over any type of media about this exam. You may not ask questions of any other human. You may not help anyone else. You must do your own work, if there are signs of academic dishonesty I will report them.

Honor Pledge

I affirm that I did not give or receive any unauthorized help on this exam, and that all work is my own.

retype the pledge:

you can type your name and the date below to indicate agreement:

Signature:

You will be graded on:

1. Correctness and completeness.

Problem completed, proper methods used and numbers correct will result in full credit.

2. Presentation and quality of writing and explanations.

Nice graphics and correct and clear explanations will result in full credit.

Problem 1

40 points

Gender discrimination in bank salaries. In the 1970's, Harris Trust was sued for gender discrimination in the salaries it paid its employees. One approach to addressing this issue was to examine the starting salaries of all skilled, entry-level clerical workers between 1965 and 1975. The following variables, which can be found in `banksalary.csv`, were collected for each worker.

- `bsal` = beginning salary (annual salary at time of hire)
- `sal77` = annual salary in 1977
- `sex` = MALE or FEMALE
- `senior` = months since hired
- `age` = age in months
- `educ` = years of education
- `exper` = months of prior work experience

```
library(readr)
banksal <- read_csv("~/SharedProjects/Kapitula/STA631/exam/banksalary.csv")
```

a. Identify observational units, the response variable, and explanatory variables.

b. The mean starting salary of male workers ($5957) was 16% higher than the mean starting salary of female workers ($5139). Confirm these mean salaries. Compute the standard error and 95% confidence interval for this comparison. Discuss any possible sources of bias or unmodeled uncertainty. Is this enough evidence to conclude gender discrimination exists? Give reasons why or why not.
c. How would you expect age, experience, and education to be related to starting salary? Generate appropriate exploratory plots; are the relationships as you expected? Do you see any patterns of concern for modeling?
d. Why might it be important to control for seniority (number of years with the bank) if we are only concerned with the salary when the worker started?
e. By referring to exploratory plots and summary statistics, are any explanatory variables (including sex) closely related to each other? What implications does this have for modeling?
f. Fit a simple linear regression model with starting salary as the response and experience as the sole explanatory variable (Model 1). Interpret the intercept and slope of this model; also interpret the R-squared value. Is there evidence of a relationship between experience and starting salary?
g. Does Model 1 meet all simple linear regression assumptions? List each assumption and how you decided if it was met or not.
h. Fit a model using `sex`, and `exper` as explanatory variables. You might need to make an indicator variable for `sex`. Discuss your fit model, including an interpretation of all parameters in the model.

i. What conclusions can be drawn about gender discrimination at Harris Trust based on your work above? Do these conclusions have to be qualified at all, or are they pretty clear cut?
j. Add an interaction term to the model in h. Give an interpretation of the new model and give a graphic representation of the data and fit model.
k. Often salary data is logged before analysis. Would you recommend logging starting salary in this study? Support your decision.

Problem 2

35 points

Sitting and MTL thickness. @Siddarth2018 researched relations between time spent sitting (sedentary behavior) and the thickness of a participant's medial temporal lobe (MTL) in a 2018 paper entitled, "Sedentary behavior associated with reduced medial temporal lobe thickness in middle-aged and older adults". MTL volume is negatively associated with Alzheimer's disease and memory impairment. Their data on 35 adults can be found in `sitting.csv`. Key variables include:

- `MTL` = Medial temporal lobe thickness in mm
- `sitting` = Reported hours/day spent sitting
- `MET` = Reported metabolic equivalent unit minutes per week
- `age` = Age in years
- `sex` = Sex (`M` = Male, `F` = Female)
- `education` = Years of education completed

```
library(readr)
sitting <- read_csv("~/SharedProjects/Kapitula/STA631/exam/sitting.csv")
```

a. In their article's introduction, Siddarth et al. differentiate their analysis on sedentary behavior from an analysis on active behavior by citing evidence supporting the claim that, "one can be highly active yet still be sedentary for most of the day." Fit your own linear model with `MET` and `sitting` as your explanatory and response variables, respectively. Using R^2, how much of the variability in hours/day spent sitting can be explained by MET minutes per week? Does this support the claim that sedentary behaviors may be independent from physical activity?
b. In the paper's section, "Statistical analysis", the authors report that, "Due to the skewed distribution of physical activity levels, we used log-transformed values in all analyses using continuous physical activity measures." Generate both a histogram of `MET` values and log–transformed `MET` values. Do you agree with the paper's decision to use a log-transformation here?

c. Fit a model with `MTL` as the response and `sitting` as the sole explanatory variable. Are the linear regression conditions satisfied? (give reasons)
d. One model fit in @Siddarth2018 includes `sitting`, log–transformed `MET`, and `age` as explanatory variables. Fit this model and get parameter estimates and SEs.
e. Based on your results from the previous part, do you support the paper's claim that, "it is possible that sedentary behavior is a more significant predictor of brain structure, specifically MTL thickness [than physical activity]"? Why or why not?
f. A *New York Times* article was published discussing @Siddarth2018 with the title "Standing Up at Your Desk Could Make You Smarter" [@Friedman2018]. Do you agree with this headline choice? Why or why not?

Problem 3

25 points

Simulate two Normally distributed random variables X1 and X2 with correlation 0.8, both should have a mean value of 70 and a standard deviation of 8. See https://www.probabilitycourse.com/chapter5/5_3_2_bivariate_normal_dist.php equation 5.23 for formulas that will help you see how to do the simulation. There are other ways to do this as well. Think of X2 as a score on the second exam in a class, and X1 as the score on the first exam.

a. Analytically find the expected value and variance of `Change = X2 - X1`.

b. Use a simulation with n=1000 to find the mean and variance of `Change` using simulation.

c. Plot your simulated data on a scatterplot.

d. Also plot X1 vs. Change.

e. How might regression to the mean cause issues in assessing whether or not a student improved or did worse on the second exam compared to the first?