SML201 Test 2, Spring 2020

Test rules

During the test, you may use Rstudio, and consult the course website, the course textbook, and the course Piazza. No other aids are allowed. You must not communicate about the test with anyone except the course instructor. You may communicate with the course instructor by Zoom at https://princeton.zoom.us/j/95501753393], by email, or, only if your internet connection is unreliable, by cell phone at +1-609-375-7283. You may not communicate with anyone else regarding the test and SML201 content.

Submit the file `test2.R` containing all the answers on Gradescope at https://www.gradescope.com/courses/88811/assignments/472336/. All your answers must be in `test2.R`.

Announcements regrarding errata may be made by email during the test.

Unless you have an approved time extension, you have 4.5 hours to work on the test. Submit the test on Gradescope no later than 5:30 p.m. Eastern Time. If you have a time extension, *submit the test by email.*

Answer as many questions as possible. Partial credit may be given for answers where substantial progress was made toward a correct solution.

Note that the problems are not necessarily arranged in order of difficulty. Read over the test, and be strategic about which problems to work on.

Some of the questions will ask for R code as well as English and math. Below is an example of how you should answer those. Note that any English or math is commented out (using the pound sign), but R code is not commented out. You must separate the answers to the different problems using lines of pound signs.

```
##################################################################################
# PROBLEM 1

sq <- function(x){
  x ** 2
}

x <- c(5, 6, 7)
sum(sq(x)) # Output: 110

# I used the sq function to compute 5^2, 6^2, and 7^2,
# summed the results up, and got 110
###################################################################################
# Problem 2
# .........
```

Include the following text, and add your name to it to confirm that you have have not violated the rules for this test.

```
# I pledge my honor that I have not violated the Honor Code during this examination.
# Specifically, I have taken the no more than the amount of time alotted to me to
# write the test, I have not used any unauthorized aids, and I have not communicated
# with anyone apart from the course instructor regarding the test or the course material
# Digital siganture: <your name here>
```

Good luck!

Question 1 (20 pts)

Read in the following dataset:

```
brains <- read.csv("http://guerzhoy.princeton.edu/201s20/brains.csv")
```

The dataset contains information about the weight of the brain, the weight of the body, the gestation period, and the litter size for a large number of mammals.

You should make a model that would do as good a job as possible (using the tools we use in the course) of predicting the weight of the brain of a new mammal, and estimate how well you'd be able to predict the weight of the brain of a new mammal. Follow the instructions below.

Question 1(a) (7 pts)

Split the dataset into training/test/validation sets. You may use a 70%/15%/15% Training/Test/Validation split. Explain why it is necessary to split the data into three sets. Explain why would you want the training set to be large. Explain why you would want the test and validation sets to be large.

Question 1(b) (6 pts)

Use at least three graphs to illustrate how you would use data visualization to make the best model possible. Use `ggplot`. Include the code you used in the R file that you submit. State what you see in the graphs and how you would use that.

Question 1(c) (7 pts)

Use what you found in 1(b), as well as other techniques we learned in the course, to make the best possible model you can to predict the weight of the brain. How well do you expect your model to predict the weight of the brain of a new mammal species? Give as precise and clear an answer as possible.

Question 2 (10 pts)

Suppose the average amount of sleep per night in the student population is 7 hours. (That is, if we ask everyone for the average amount of sleep they get per night and then average the response, we'd get 7 hours).

Question 2(a) (2 pts)

What is a plausible value of the standard deviation of the amount of sleep per night in the student population? Explain your reasoning as precisely as possible.

Question 2(b) (2 pts)

We intend to survey 200 randomly-selected students and got each student's average amount of sleep per night, and compute the average. What is the probability that this average is greater than 7.2 hours? Use `pnorm(..., mean = 0, sd = 1)` to obtain the answer. Show your work.

Question 2(c) (3 pts)

Answer Problem 2(b) again, but now you must use `rnorm(..., mean = 0, sd = 1)`.

Question 2(d) (3 pts)

There are at least two quite different ways to answer 2(c). Answer 2(c) again, using `rnorm(..., mean = 0, sd = 1)` again, but now you must use it in a way that's substantially different from the way you used `rnorm(..., mean = 0, sd = 1)` in 2(c). Show your work. Explain how your approaches in 2(c) and 2(d) are different.

Question 3 (20 pts)

Consider the following model and predictions:

```r
titanic <- read.csv("http://guerzhoy.princeton.edu/201s20/titanic.csv")
fit <- glm(Survived ~ Sex + Age + Pclass, data = titanic )
titanic$pred <- predict(fit, newdata = titanic) > 0.5
```

```r
titanic.male <- titanic %>% filter(Sex == "male")
titanic.female <- titanic %>% filter(Sex == "female")
mean(titanic.male$pred == titanic.male$Survived)       # 0.8097731
```

```
## [1] 0.8097731
```

```r
mean(titanic.female$pred == titanic.female$Survived)   # 0.7579618
```

```
## [1] 0.7579618
```

Question 3(a) (8 pts)

Test the null hypothesis that this classifier does not violate accuracy parity on the training set `titanic`. State your conclusions. You may use either fake-data simulation or an analytic method.

Question 3(b) (12 pts) (Challenge)

The accuracies for male and female passengers are different (though you may have concluded in 3(a) that there is not enough evidence to reject the null hypothesis that they are the same and the differences are due to chance).

Write a function with the signature `predict.fair.acc <- function(fit, sex, age, pclass, male.acc, female.acc)` which takes in the model `fit` (which is the same as what's displayed above), the sex, age, and class, of a new passenger, and the accuracy for `fit` for male and female passengers (so 0.81 and 0.76 in our case), and does the best job of outputting predictions that would satisfy accuracy parity while also being fairly accurate.

Use simulation to check to what extent your function achieves the objectives set out above. Explain your work.

Question 4 (15 pts)

Suppose we are exploring a dataset similar to the `finches` dataset. Use `ggplot` to illustrate how the probability of a type S error depends the magnitude of the difference between the population means. Explain your work.

Question 5 (10 pts)

Suppose that the results of a poll of likely voters is that 51% intend to vote DEM and 49% intend to vote REP. What is the smallest sample size needed with this kind of result (a 51%/49% split) in order to reject the null-hypothesis that the true probability of intending to vote DEM is 50%?

Show your work to the extent that a grader can understand your thinking exactly. It is not necessary for your work to consist just of code in order to get full credit for this problem – some manual exploration is allowed.

Question 6 (10 pts)

A CDC survey from 2015 found that people who sleep less have lower incomes. One possible causal story is that sleeping more causes higher incomes because job performance improves with more sleep. Outline other hypotheses that explain the data, and briefly explain each one. Your hypotheses should demonstrate four different possible causal structures. The causal structures must be substantively different, rather than be closely analoguous.