Homework_1

DUE: Tuesday, February 4th, at 7pm via Canvas Submission.

Please submit the solution in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word).

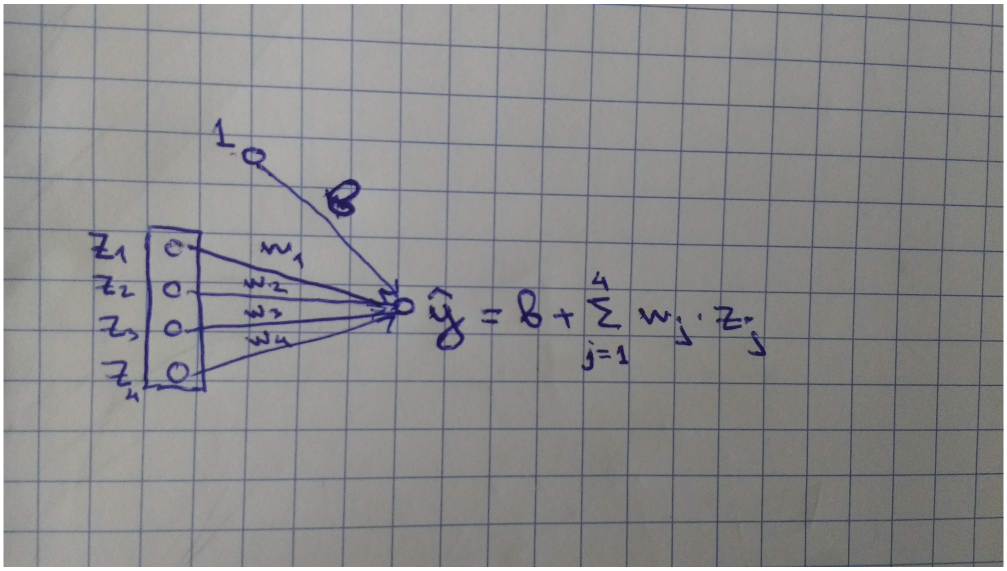
Goal of this homework is to have you 1) solidify your understanding of observational vs experimental studies; 2) work on simple linear regression concepts; and 3) lightly practice your R skills.

Whenever asked to "derive" something, which includes heavy usage of formulas and algebraic notation, you may either:

1. **Type it up in LaTeX mode (check this source file for some examples), like this**

$$\sum_i e_i x_i = \sum_i (\hat{y}_i - y_i) x_i = \sum_i (\hat{\alpha} + \hat{\beta} x_i - y_i) x_i = \dots$$

2. **Write the solutions by hand, take a picture, insert it into this R markdown document (DON'T supply it separately), like this**



Please check the ***.Rmd*** **source file for examples, and also, see https://stackoverflow.com/questions/25166624/insert-picture-table-in-r-markdown for reference on how to insert images and manipulate image size in R markdown.**

Problem #1

Imagine that students in an introductory statistics course complete 20 assignments during two semesters. Each assignment is worth 1% of a student's final grade, and students get credit for assignments that are turned in on time and that show reasonable effort. The instructor of the course is interested in whether doing the homework contributes to learning, and she observes a linear, moderately strong, and highly statistically significant relationship between the students' grades on the final exam in the course and the number of homework assignments that they completed. For concreteness, imagine that for each additional assignment completed, the students' grades on average were 1.5 higher (so that, e.g., students completing all of the assignments on average scored 30 points higher on the exam than those who completed none of the assignments).

a. Can this result be taken as evidence that completing homework assignments causes higher grades on the final exam? Why or why not?

b. Is it possible to design an experimental study that could provide more convincing evidence that completing homework assignments causes higher exam grades? If not, why not? If so, how might such an experiment be designed?

c. Is it possible to marshal stronger observational evidence that completing homework assignments causes higher exam grades? If not, why not? If so, how?

Problem #2

1. Presume for data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we have fitted simple linear regression equation

   $$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i,\ i = 1, \ldots, n$$

   where $\hat{\alpha}$ and $\hat{\beta}$ are least squares estimates. Proceed to derive that

   (a) $\sum_i e_i = 0$, where $e_i = \hat{y}_i - y_i$
   (b) $\sum_i e_i x_i = 0$

2. Besides the derivation provided in part $1(a, b)$, also confirm these results on the example of simple regression of *prestige* on *education* in the *Prestige* data set. Do it **in** *R*: fit the model, extract the vector of residuals e and that of explanatory variable x, and apply appropriate operations to calculate the required quantities.

3. For the "null model" that predicts a constant for every observation:

   $$\hat{y}_i = \hat{\alpha},\ i = 1, \ldots, n,$$

   derive that the formula for the least squares estimate is

   $$\hat{\alpha} = \bar{y}$$