Stats 203V: Introduction to Regression Models and ANOVA Summer 2020

Homework 1

Due date : July 5, 11:59 PM (PT)

For full credit, solutions (.pdf file, either typed or handwritten and scanned) must be uploaded to `Canvas` by **11:59 pm PT on on Sunday July 5**. Any changes to the Homework will be posted on `Canvas`.

Collaboration on homework problems is fine, but your write up should be your own and the write up should mention the names of your collaborators.

For questions calling for `R`, if you are familiar with `R Markdown`, by all means use it. It is not required for now (though it will be strongly encouraged later in the course).

1. (15 points) Freedman, Ch. 2. Exercises B15, B16.

2. (5 points) Freedman, Ch. 3. Exercise B11.

3. (20 points) Let X be a matrix of dimension $n \times p$, with $rank(X) = k \leq p$. Then show that $rank(X'X) = k$.

   (Note that this problem slightly differs from the proposition proved in the lecture which shows that X is of rank p if and only if $X'X$ has rank p.)

   Using the above show that

   $$\text{Column space of } X = \text{Column space of } XX'.$$

   (Hint : Use the following fact from Matrix Algebra.

   $$rank(X) + \text{dimension of null space of } X = \text{Number of columns of } X.)$$

4. (20 points) This is Lab 1. from Freedman (p. 295), retyped to incorporate some small changes (mostly relating to using `R`) :

   "In this lab, you will calculate some descriptive statistics for Yule's data and do a simple regression. The data are in table 1.3, and in the file `yule.txt` in the HW folder on `Canvas`. You will need to subtract 100 from each entry to get the percent change. Refer to Ch. 1 for more information, note that the data in `yule.txt` are ratios, not changes.

   (i) Compute the means and SDs of ΔPaup, ΔOut, ΔPop and ΔOld.

   (ii) Compute all 6 correlations between ΔPaup, ΔOut, ΔPop and ΔOld.

   (iii) Make a scatter plot of ΔPaup against ΔOut.

   (iv) Run a regression of ΔPaup on ΔOut, i.e. find the slope and intercept of the regression line "by hand". Also use `lm()` and compare your results.

   Useful `R` commands : `read.table, cov, sd, cor, plot, lm`.

5. (20 points) This exercise aims to use a little theory (the square-root law for variability) and a little `R` to better understand Wainer's Figure 3 about Kidney cancer rates for some 3100 US counties.

   Let Y be the number of kidney cancer deaths in a year in a county with population n (for US counties, n will vary from below 10^4 to 10^7). Suppose that Y has a $Binomial(n, p)$ distribution, so that the number of cancer deaths is modeled as being like the outcome of n independent tosses of a (very biased) coin with small $p = \mathbb{P}$(death from kidney cancer).

   Recall from probability that if $Y \sim Bin(n, p)$, then $\mathbb{E}(Y) = np$ and $Var(Y) = np(1-p) \approx np$ (since p is assumed small).

   (a) Let Z be the death rate per $100,000$, so that with $n_0 = 100,000$, we have $Z = (n_0/n)Y$. Calculate $\mu(n) = \mathbb{E}(Z)$, $Var(Z)$ and $\sigma(n) = SD(Z)$, and express the answers in terms of n and $\lambda = n_0 p$. Explain how the square root law is manifested here.

   (b) The cancer death rate per $100,000$ that is k standard deviation above the mean will be a function of n:

   $$q_k(n) = \mu(n) + k\sigma(n).$$

   Using `R`, make a plot of $q_k(n)$ versus $\log_{10}(n)$ for $k = 2, 3, -2$. Make an educated guess at a value for λ (or p) from Figure 3. What conclusions do you draw about Wainer's Figure 3?

   (Suggestion : Choose $\log_{10}(n)$ values equally spaced from 4 to 8, and restrict the y-axis to $[0, 20]$, using `ylim=c(0,20)`.)

   (c) If a county with population n has 1 death in a given year, then the corresponding rate per $100,000$ is n_0/n. Add a plot of n_0/n versus $\log_{10}(n)$ to your graph from (b) and say what it explains about Wainer's Figure 3.

   (d) Suppose that there are about 200 counties having a population between $200,000$ and $400,000$. About how many of these counties would you expect to have cancer rates lying above the $2SD$ line, namely $q_2(n)$? (Hint : Use normal approximation)