ISTA 370: Homework 2

Due by Friday, 11:59 PM, September 25th

September 18, 2015

The goal of this homework is to further review the concept of explaining variability and to successfully perform the analysis of variance. For simplicity we will be using the available R datasets.

1 Explaining variability

Review the example of explaining variability in Lecture 5, slide 38. If you look carefully, the example shows you have much of the total variability in heights is explained by gender. We will do something similar but with a different dataset. The mtcars dataset you used in HW1 is a really good example to review some of these concepts. How about you load and review that dataset again. Look through all the variables. You can use for following R code:

```
attach(mtcars) # load the dataset
names(mtcars) # look at different variables

##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"

View(mtcars) # view the entire dataset
```

1. (10 Points) Using the similar analogy from the example in Lecture 5 (slide 38), find out (using R) what *percentage* of the total variance in mpg is explained by the number of cylinders (cyl) in the mtcars dataset.

You should write a function in R, which will return the percentage of the total variance in mpg, which is explained by cylinders. Also, remember you have to define the popvar function before you use it in your code. It is not an in-built R function. *Hint*: The difference between the total mpg variance and the average mpg variance after separating by number of cylinders is the amount of variance explained by the factor cyl.

2. (10 Points) Now find out what percentage of the total variance in mpg is explained by the number of gears (gear). Again, write a function in R, which will return the percentage of the total variance in mpg, which is explained by gears. Use the above hint here as well.

2 Analysis of variance

Review the lecture on the analysis of variance and recall the steps involved in that analysis. It is called "one way" because it involves just a single X or independent variable (measured or controlled) and a single y or dependent variable. In class (Lecture 7: Analysis of Variance) we worked through an example by hand and we also looked at the R example where the independent variable was the number of cylinders a car has and the dependent variable was the car's mileage. Analysis of variance is an enormously valuable statistical tool, but it is also a way to think about research. The researcher is always looking for factors that explain a lot of the variance in a dependent variable, relative to background variation or noise. The analysis of variance, and particularly the sums of squares, literally divide the grand variance into two parts: that due to the independent variable, and that due to everything other than the independent variable. Dividing variance into the part due to the factor you measure/control and the part due to all other factors is such an important idea that you must really feel these quantities and how they are extracted from data, rather than blindly learning to run an analysis of variance procedure in R. For this reason, your assignment is to implement in R the formulae for SS_{total}; $SS_{between}$ and SS_{within} as well as $MS_{between}$; MS_{within} and F, and run your code on two different analysis of variance problems and explain your results. Now, I've already given you code in the lecture notes, but it is ugly code (for example, it doesn't sum over groups) and besides, the point is for you to write your own. **Also, for each question below, run R's aov function and report the summary; use this to**

check that your sums of squares are calculated accurately, and note the reported significance (the p-value) level of the F-score. For each question below, submit your answers in a R file. Clearly label each question and provide your answers and explanations as comments. We will test your R code by running it in R.

3. (15 points) Write different functions in R (with your own efforts, don't copy the code from the slides) that calculates different elements in the analysis of variance. Run your code on the mtcars example from class to make sure you get the same result. Be sure to include all your code! If we can't run your code from what you submit, you won't get credit! Show the output (in your comments) of running your code in the R console for computing SS_{total}; $SS_{between}$, SS_{within}, $MS_{between}$;MS_{within} and F.

4. (5 points) Run it again with mpg as the dependent variable but gear as the independent variable. Report SS_{total}; $SS_{between}$ and SS_{within} as well as $MS_{between}$;MS_{within} and F. Compare these results with the original analysis. Which explains more of the variance in mpg? Is it cyl or gear? Explain!

5. (10 points) Do the same analysis on the ChickWeight data that comes with R. The dependent variable will be ChickWeight$weight and the independent variable will be ChickWeight$Diet.Report the same statistics as in the previous analyses. Does diet matter? Explain your answer.