- Multiple regression tied up
  collinearity and the ANOVA table in multiple regression.
  * t statistics and t tests.
  * F tests and their connection with permutation tests based on R-squared.
- Go over midterm
- Work on project for the next two weeks. No homework
  - Anything possible for project, from analyzing startups to sport statistics, who's following whom on twitter, what predicts how many followers someone has
  - Spend this week finding the right dataset.
  - Use in some way the material that we have learned in this class
  - Project is due two weeks from today.
  - In addition to writeup, provide the datasets and r scripts as well
  - Want hard copy of written report, and electronic copy of everything else
  - Can work in groups, one writeup for everyone
- Plan for near future
  - Multiple regression
  - Step-wise selection
  - Time series of forcasting
  - Logistic regression
  - Decision analysis
- Pep talk
  - Every rowing race is won in the third 500m
  - Endure!
  - Catch up
- A few loose ends:
- ANOVA tables in multiple regression
  - http://jgscott.github.io/teaching/r/cps/cps.html
  - Know the sector will be necessary
  - SE calculated under normal linear regression model
    - Y(i)=Y-hat(i)+E(i)
  - Baseline is clerical
  - lm1 = lm(wage ~ educ + sector, data=cps)
    - Educ and sector were placed in that order arbitrarily, but the numbers are the same in whichever order. This shows that we are just adding the numbers together. Nothing will change if the variables are placed in different orders
    - ANOVA tables, are, however, different
  - TV=PV+UV, or TSS=RSS+ESS
    - The analysis variance takes the RSS/PV and breaks it into its constituents. This allows to give credit to separate parts.
    - So while the models themselves don't change based on the order, the ANOVA table does depend on order. But if the order is arbitrary, then the attribution of credit is also somewhat arbitrary.
    - So, there is no objective way to partition credit among the variables
      - Think of it as, all the variables are working as a team to create the model, and it is very difficult to assign credit
    - The only way to assign credit is if there is almost no interaction between each separate variable. So if the variables are independent, then we can partition credit objectively. But if the variables are correlated, then much harder to give credit.
  - How does the ANOVA table give credit?
    - Start with a model with only education
    - How much does PV go up and UV go down?
    - Then add another variable, and ask the question again

        - This ignores any teamwork effects, any correlation. It's an arbitrary way to assign credit.
    - So why is it useful?
        - Still useful with reasoning in regards to change
        - ANOVA useful for answering the question, how useful is this as a predictor, **considering what we've already got?**
        - Allows you to reason about marginal benefit of adding an X to a regression model
- T tests and F tests
    - Go back to georgia2000, exercise 6 (line81)
        - Do a shuffle to test the effect
    - The F-statistic
        - F=$\frac{\Delta R^2}{1-R_F^2}\ X\ \frac{n-P_F-1}{P_F-P_R}$
        - $\Delta R^2 = R_F^2 - R_R^2$
    - The F test is the same as the t test, just with different scaling factors
    - ANOVA table does this for us
- Remember two resampling ways
    - Permutation tests are to bootstrapping as f tests are to standard errors of the NLRM
    - F test is like Permutation test using R2 as test statistic, while t test is like permutation test using the coefficient itself as test statistic
- A t statistic is like a signal to noise ratio:
    - $t - statistic = \frac{estimate}{standard\ error}$

- Mid-term
    - Question 1:
        - Although number one is big study, no evidence of experimental design, while number 2 had a control group versus experimental group. But you could have argued it the other way, if you had mentioned number two did not mention randomized assignments
    - Question 2:
        - Sampling distribution: what I get when I compute an estimator after the next. This is sampling from the population, and computing the thing we care about. If the answers do not change much from sample to sample, then can reasonably trust the answer
        - Bootstrapping- why? Replicate process of taking real samples from a population
        - Frequentist coverage principle- property of a procedure to construct confidence intervals. Refers to property where ex. 95% confidence intervals over and over again, then they should contain the true variable 95% of the time
        - Somehow the joint of two variables is even more than just the two on its own
        - y=KX^{β} $- take\ the \log on\ both\ sides$
    - Question 3:
        - False. The question is asking, what is the group mean, the average number of coffee cups sold per day. Look at ln2, and look at the baseline. The statement was false, the confidence interval should have been the intercept
        - False. The numbers quoted in the question said model 3, should have mentioned model 4 as well.
        - False, could have quoted ln3, when those numbers came from ln4, which has a quadratic effect, so can't simply read off the linear terms
        - Not determinable. Would have needed something like a histogram of residuals.
        - Temperature has a lower effect than period does. Could have reasoned about the anova table, or the actual estimated values of the coefficients