Monday March 23, 2015

Monday, March 23, 2015 9:28 AM

Welcome back! I hope everybody enjoyed their spring break.

- Go over midterm at the end
- 1st hr: tie up loose ends

Today's topics:

- 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.

Project:

- Look at sports statistics to who's following who on Twitter. Come chat with Scott for project ideas.
- Posing the Q and finding the right data set is 75% of the project.
- Need to use the material we've learned in an intellectual and engaging way.
- Due April 6, 2015 - Monday
- Write up + R-script (or whatever software program) + data set due - email to dropbox email
    - Analysis must be reproducible
- Groups - no more the 4 ppl,
    - one write up with everyone's name

Continue Regression for the next 2 weeks

- Step wise selection on Wed
- Next week:
    - Time series and forecasting
    - Wed: Logistic Regression

Pep talk:

- Analogy: Rowing race : 2 km
    - Endurance sports are painful
    - Coach: Every rowing race is won in the 3rd 500 m block
- Apply it to this course. You still have time to catch up. Don't leave it to the end. If you're ahead, don't back off.

A few loose ends:

- ANOVA tables in multiple regression
- t-tests
- f-tests

See: Hourly wages in the Current Population Survey walkthrough

Goal : understand ANOVA table in MR

- Start 2 var: educ + sector as dependent variables
- Standard errors: calculated under assuptions of normal linear regression model
    - NLRM: yi = yhati + Ei
        - Yhati = predictor
        - Ei -> Ei~N(O,o^2)
- Switch effects position in model.
    - Fitted values are the same.
    - You can see that order of effects doesn’t matter - addition is commutative.
- One thing does change

ANOVA Tables in multiple regression

- ANOVA Table changes based on order of effects
- ANOVA - attempting to take decompisition of variance
    - TV = PV + UV
    - Total sum of squares = regression sum of squares + error sum of squares
    - TSS = RSS + ESS
        - RSS: breaks up into parts
    - ANOVA table - way to assign credit
        - What's a stronger predictor (What matters more)
        - What contributes a greater number to the sum of squares
- When effects order is switched
    - RSS is the same
    - Assigned credit is different
        - Educ is first -> more contribution
        - Sector first -> then more contribtuion
- Anova table depends on order, but order is arbitrary
    - Therefore attribution of credit is somewhat arbitrary
    - Then not really objective way to partition credit
- Basketball teams
    - Hard to partition teams to assign credit
    - Total team effort transcends total sum of parts
    - Same for MR model
    - Side note: only if one member, contributes a lot independently
        - See baseball - at the plate, everyone is independent
        - **Regression modeling : if variable is independent, uniquely attribute credit to individual variables
- Anova table numbers are calculated sequentially
    - Gives credit like so
    - See hypothetical draft
        - Each additional person to the team supposedly accounts for the additional points made
        - Totally ignores team work or correlation of variables
    - Still arbitrary
- Usefulness: reasoning with regard to change
    - See walkthrough
        - Can say if education is useful, some of it might be redundant
    - How useful is this predictor in the context that we have it
        - Should we add or remove something?
        - Low SS - not necessarily bad predictor
    - Looks at the marginal benefit of the predictor on the model

T- tests and F-tests

- See exercise 6 R-script (line 81) permutation
    - Shuffled equipment to test the effect
    - See histogram of permutation of r squared
    - .27 is in the rejection region so equipment matters
- $F = \frac{\Delta R^2}{1 - R_F^2} * \frac{n\ - P_F\ - 1}{P_F\ - P_R}$
- $\Delta R^2 =\ R_F^2\ -\ R_R^2$
    - $P_F : number\ of\ total\ parameters$
    - $P_R : number\ parametersin\ reduced\ model$
    - $n :\ \ number\ of\ rows\ or\ points\ in\ data\ set$

- $F = \frac{0.27 - 0.2}{1 - 0.27} * \frac{159 - 7 - 1}{7 - 4}$

- F-test: same as t-test, just different scaling factors
    - Anova table does this for us
- 2 ways to calculate Standard error
    - Bootstrapping
    - Normal assumptions
    - Boot strapping are to permutations as NLRM are to f-test
- $t - statistic = \frac{esitmate}{standard\ error}$
    - t-stat - signal to noise ratio
    - How big is the estimate as a signal of how precisely we measure it
- f-test - perm test using R-squared
- t-test - perm test using the coefficient itself as a test statistic

Midterm review

- Part 1
    - Study 1 - uses stat analysis but no designed experiment
    - Study 2 is superior has a control
        - Need to recognize importance of experimental vs. control groups
            - Assigning groups on a random basis
- Part 2
    - Samp dist - compute estimator from population
        - Assess stability of estimates
    - Bootstrapping
        - Resampling with replacement from original sample
        - To replicate or approximate process of sampling from population
    - Frequentist coverage property
        - Procedure used to construct confidence intervals
        - Should fall into confidence interval of .95 95% of the time
    - Interaction term: joint effect of two variables that's different from the sum of its parts
    - Y = Kx^beta
        - Take log transformation of both sides
        - Logy = log (Kx^beta)
        - Logy = logK + beta log x
- Part 3
    a. What's the group mean? What's the confidence interval for the number?
        i. It was lm2. it was 168-178 (false)
    b. What is the diff b/w final exam period and summer controlling for temp?
        i. Needed to make decision between lm3 and lm4 - judge which one you choose
        ii. Needed lm3
            1) Offset -69 and-59 (false)
        iii. Needed lm 4
            1) Offset -76 and -65 (false)
    c. False. Fit linear model, the slope is this, see lm 4 but can't use that - quadratic
    d. Needed to see histogram of residuals in order to judge if d was true - undecided
    e. See anova table lm 5 (False)
        i. Temp has a lower effect than period does
        ii. Sum of squares - period has a greater effect