36-726: Statistical Practice

Comments on Progress Reports

Brian Junker

132E Baker Hall

brian@stat.cmu.edu

Outline

- (Very) brief comments on progress reports

- Data Status

- Methods

- Second progress reports

- End of Semester Written Reports

- End of Semester Final Presentations

- Keeping Track of Your Work

(Very) brief comments on progress reports

- Generally great!
  - Most followed standard formula, which is fine and appropriate

- Most (but not all) slide decks had slide numbers
  - As the material becomes more complicated, slide numbers are important for Q&A
  - Please include them!

Data Status

- **HCI learning discontinuities**
  - Downloaded data from data shop
  - Need to clean data and get something we can work on
  - General search for learning discontinuities
- **HCI perquisite discovery**
  - Received data from 2019-2020 from Mathia software
  - Have data and transaction values; some data cleaning required
  - Some prior knowledge about prereq's, maybe will get from client

Data Status

- **IFDA understanding manpower market**
  - Interviewed all the companies
  - Meet today about public data
  - Get data from companies this week
- **NBA plus/minus**
  - Have scraped 18-19 contract data; older data costs $!
    - Hoping to come up with a team rating for prior also
  - Play by play data from Kostas (8 seasons)
- **NHL understanding player paths to NHL**
  - Scraped data from NHL NCAA, USHL, 20 seasons
  - Hard to scape biometric info (individ player links…)
  - Expensive API, talk to client about it

Data Status

- **NPD anomaly discovery**
  - ❑ Waiting to get credentials to access real data
  - ❑ Anomaly and changepoint methods (Valerie reading)
- **PHIGHT COVID effect of policy on pandemic**
  - ❑ New data, new objective for project
  - ❑ Cases and county level data from Ohio sch. districts

Data Status

- **PPS use of Promise Scholarships**
  - ❑ Have 9 data sets, not many obs's per data set (<10K)
  - ❑ Data seems pretty clean; have questions for client
    - ■ Meaning of fields; what range of data to focus on… etc.
  - ❑ Can start EDA after that
- **SIMON effect of COVID on teaching and learning**
  - ❑ Received actual data last Fri in sql format
    - ■ 3 semesters of data
  - ❑ Explore relats among 40 data tables; expl variables etc
  - ❑ Identifying variables for analysis

Data Status

- ## CMU Africa AFRINIC

  - (later)

- ## CMU Africa JP MORGAN

  - Getting data from "gdelt" (sp?)

  - Working with newscatcher and google news api's to get news more oriented to business

  - Will be using all open source data

Methods

- What is common among the projects?
- Part of being a good statistician is seeing when apparently different projects actually pose similar data analysis problems
- The specific methods may be
  - the same,
  - slightly different, or
  - very different,

  but it is good to see the commonalities
- Talk to other project teams about methods
  - (without breaking data confidentiality of course!)

Some possible links...

- **Detecting unusual patterns**
  - ❑ HCI learning discontinuities
  - ❑ NPD anomaly detection
- **Missing data in Human Resources**
  - ❑ IFDA understanding labor market
  - ❑ NHL understanding player movement to NHL
- **Developing or improving "indices"**
  - ❑ AFRINIC network resilience indices
  - ❑ NBA improving plus/minus index

These are not the only or perhaps even the major methods for each project, but they illustrate some common features of the projects!

More possible links…

- **Discovering networks, graphical models**
  - ❑ HCI prerequisite discovery
  - ❑ JP MORGAN discovering social networks
- **Logistic regression, causal analysis**
  - ❑ PPS use of Promise scholarships
  - ❑ SIMON effect of COVID on teaching and learning
- **Interrupted time series**
  - ❑ PHIGHT COVID effect of policy on public health
  - ❑ SIMON effect of COVID on teaching and learning

These are not the only or perhaps even the major methods for each project, but they illustrate some common features of the projects!

Second Progress Reports

- If you are supposed to) have data:
  - Data description, EDA
  - Conjectures about analysis methods
- For everyone:
  - What are you doing?
  - What have you learned so far?
- Give me a clearer picture of what the final product may look like
- These are supposed to start next week would you like another week for this?

End of Semester

- **Written report**
  - ❑ I think that every project has some sort of written report for the client (in addition to dashboard, shiny app, etc. perhaps)
  - ❑ I would like either an IDMRAD paper or the written report you are writing for your client

- **Final Presentation**
  - ❑ Accommodate client schedule
  - ❑ To the extent possible, I want to attend
  - ❑ Would be good for faculty mentor to attend also

Keeping track of your work

- You should already be keeping a common "Dear diary" of everything you and your team are doing
  - May be the raw materials for a "technical appendix"
  - For many purposes Rstudio/rmarkdown will be good for this;
  - Use another tool if more appropriate (e.g. recording results of meetings, interviews, etc.)
- Keep in common file space that everyone can access
  - At least one common document for the whole team
  - Indivdiual documents for each team member are OK as long as common document is kept up to date