Morning - material assessment: 9-12

* 50 questions --> MC, T/F, SA
* Only formula sheet, no outside notes
* Try to get your ability to understand the concepts of what we covered
* Taught in lecture (if not taught in lecture don't worry about it), textbook = additional resources???
* On Moodle, will have a formula sheet

Afternoon: Statistical commuting in R assessment

* 3 questions (5-10 part): code in R and interpret what you see
* Will have to build, assessment, interpret linear regression (will be walked through like in lab)
* Copy paste code in Moodle --> can give partial credit
* Only code what we have coded in class, never ask what is beyond what was in class

REVIEW

Exploring data

Defining Variavkes

* Continuous
* Categorical -->

1. Binary (default- all binary are ordinal): only 2 options
2. Ordinal 2 ways of writing choices
3. Nominal, more than 2 groups and can be ordered in any way

Make categorical variables numeric

* Dummy (reference) coding -->
* Effects coding -->

Distributions of variables

* Summarizing dists
  + Center --> middle of dist, mean, median, mode
  + Location --> where on the dist is the different sets of data, quantile, quartile
  + Spread: IQR, range
  + Shape: skewness, kurtosis, thickness of tail
  + Total: impact on outliers on a single var, points that are far from the rest of the data, number line (really high or low)
* Normal distribution: perfect symmetric, completely defined by mean and sd
  + Inference: empirical rule (68-95-99.7)
  + Evaluate: qq plot:
    - Data on 1 axis and theorical normal distribution on the other axis //
    - Skewness - quadratic looking line
    - Looking for a STRAIGHT LINE = normal dist
* Outliers:
  + Inference (guessing): have an idea of a popu and interested in a parametere
  + Summarize popu ==> sample = best guess that's representative of a popu and make inference on the population
    - Population --> parameter
    - Statistic --> sample // mean, sd, slope
      * Take stats and make prediction of the popu dist
      * Take many sample means/sd from many samples in a popu ==> graph on histogram and it will look like the sample dist
    - CLT --> x bar follows normal dist when mean of mu and sd of sigma over n (don't know sigma, have to estimate it with s)
      * If population is normal than don’t need to sample more???
      * Large sample size = 50
      * Standard error: don't know sigma, and replace with s, SE = s/Srt(n) , not the sd anymore

Hypothesis test

* Know the steps:

1. Stat null and alternative
2. Get test stat (how much evidence do you have)
3. Get p-value (how rare is the data)
4. Decision (is pit rare enough)
5. Make it interpretable

One-Sided Test

* Only care if data leans on one side (of the tails)

Two-Sided Test

* Don;t care which tail it is in

T-Test (hypothesis test for the mean)

* Compare 2 different means to see if one was higher than the other
* Assumption:
  + Equal variance (different test if no equal variance) - run HT for equal variance, then run HT to see if have equal means

ANOVA

* one way- use categorical predictor to predict a continuous response var
  + Do the groups have the same avg
  + Is the categorical var not predictive f the target
* K-levels (F test used for ANOVA - know what F dist looks like):
  + Nulls - all means are the same
  + Alternative - at least ONE mean is different
* ANOVA assumptions:
  + Normality of groups
  + Independence amongst the group
  + Equal variance of group
* Post-hoc test - which mean is different
  + Tukey --> adjust for the experiment wise error rate?
    - As more coin flips occur, more chances mistake/error increases
* Kruskals Wall’s: when assumptions don't hold true

Confidence interval

* allCI = stat +- an error // stat is a guess, add wiggle room above/below it

Linear regression:

* Use continuous var to predict continuous var
* SLR - defines direction and slope of the 2 vars relationship
  + OLS: defines the line
    - Takes the residual (truth-prediction), square them and sum them
    - Minimize it to get the best fit line
    - BLUE - best linear unbiased estimate
  + Know how to write the SLR equation // also know how to interpret slope and intercept
    - Intercept: expected value of y when x is 0
    - Slope: Y increases by beta when X increases by 1??
* Correlation - strength of relationship (not which direction)
* Training and testing split

To-Way ANOVA

* More than 1 categorical variable
* Can have interactions if there are more vars
  + Do not know the relationship between X1 and Y until you know X2
  + Slice data to look at individual interaction between groups
* Assumptions for the groups?:
  + Normality
  + Independence
  + Equal variance

Multiple Linear regression

* Same a two way but more variables
* Still use OLS
* Still have individual global F-test for the slopes
  + Null: all betas are 0 (besides intercept)
  + None of the vars have any predicted power
* Linear combo of vars, structure of the equation (how the Y is written) // can have quad term, interaction term as long is in the linear way
* Assumptions:
  + Residuals centered round 0
  + Independence of residual
  + Normality of residuals
  + Equal variance of residual
  + Multicollinearity --> no PERFECT multicollinearity
  + Assess Muti linear regression
    - % of variance in Y explained by the model
    - R^2 always increase when a var is added
    - Adjusted R^2 will penalize if vars are added
      * Adjusted is not interpretable, but used to when comparing models
  + Dummy coding and effects codding have different interpretation in linear var // know the interpretation of the slopes
    - If you have a category var --> the category is a COMAPRISION of 2 groups
      * dummy group compare grousp to on reference group
      * Effects codding: compare group to overall avg to the avg for ALL groups
  + Best Linear Unbiased Estimator
    - No other way of unbiasedly estimating the slope
  + Regularized regression: LUE
    - Now in unbiased word, don't care about Best
    - No interpretation = better interpretation ?
    - Can find something better --> error of prediction decreased
    - The more we penalize, the closer we get the coefficient to be 0 (don't need to know the equations)
      * Ridge - impact have on coefficients, they get close to 0
      * Lasso - can make coefficient 0
      * Elastic Net - can make coefficient 0

Model building (know differences and impact of removing a var)

* Step-wise selection

1. Forwards - once it is kicked out it is permanent
2. Backwards - permanently added in
3. Step-wise - can come back after getting kicked out, but is kicked out again it is now permanent

* AIC and BIC --> LOWER the better, when comparing other models

Diagnostics

* Evaluate assumptions of linear egression were
* Normality of residual through qq plot
* Constant variance - did residual follow homoscedasticity
* Independence: derbin watson (non-independent in time series)
* Mis specified models --> examined residual plots, build polynomial models
* How to evaluate all of them
* Multicollinearity
* Variance inflation factor, using every single X to predict another X

Polynomials: centering (standardizing) the variables to help address multicollinearity

Categorical data Analysis (predicting categorical vars)

* Nominal v. ordinal
* Look for relationship between 2 categorical variable
* Chi^2 test (Pearson - any var is nomial, likiehood ratio, are interchangeable because serve the same purpose)
  + Null: no association between the vars
  + Alte
* Mentel Henzel chi^2:
  + No linear association between the 2 var
  + Used for ordinal v ordinal relationship // ordinal vars have direction, have potential of directional relationship, and MH accounts for that
  + If fail NH, there is no linear association, does not mean there is NO relationship

Measures of Association

* Test strength of relationship
  + 2 or at least one nominal var : cramers V test
  + 2 ordinal vars: spearman's correlation (can give direction and strength, but the # doesn't mean anything like pearsons), or 2 binary
  + 2 binaries: odds ratio - more interpretable
* Odds v. probability
  + Look at 2 sets of odds // odds divided by other odds
  + Understand how to calculate odds ratio to help interpret R

Logistic Regression

* OLS is bad,
* Interpretation around logistic interpretation - brings back odds ratio (know how to calculate and interpret ratio)
* Concordance (concordance, discordance, tied pair) // combos of 1s and 0s
  + Have higher predictability of 1s - concordance
  + Higher for 0 - discordance

Box-cox - what assumption it helps solve, when to use it?