STAT 411 Notes

1 Design and Analysis of Experiments

- Replication -
  - More test subjects / data
  - Different from a "repeated measurement" - Measurement of the same experimental unit over time while replication is having many data points
  - Gives power to experiment
  - Reduces variation in measurement process
- Blocking
  - More observations are good - brings more variability, mitigated by blocking
  - Block - group of data subjects who all have the same(or similar) test conditions
- Randomization
  - Helps prevent bias
  - Assignment of subjects to treatments must be done randomly and prior to the start of the experiment
  - All choices about what to do data-wise should be done prior to running the experiment

Planning the Experiment(do this from the beginning):

- Experiment Planning Checklist
  - Define objectives of experiment
    - List the precise questions the experiment will address - what do you want to know?
    - Beware experiment creep - adding more things to measure over time - "fishing expedition" to find significance
    - Dangers of p-hacking - if you run enough experiments you will find a significant p-value
  - ID all sources of variation
    - Anything that could cause observations to differ
    - Include, but don't limit yourself to
      - Treatment factors and levels
      - Experimental units
      - Blocking factors, noise factors, covariates

- Choose an assignment rule
  - The assignment rule specifies which experimental unit receives which treatment
  - The assignment of experimental units to treatment should be done at random, within the constraints imposed by the experimental design - blocking should be random
  - Depending on the experiment, some assignments may not be ethical
- Specify measurements, procedure, and difficulties
  - Measurements need to match what is being tested(values, precision, method, etc.)
  - Procedures need to be the same as possible, as well
  - Difficulties can be hard to anticipate, but do your best(also consider running a pilot)
- Run a pilot
  - Mini-experiment, where procedure is being tested for feasibility - can this be done at a smaller scale and then scaled up?
  - ID any difficulties that didn't occur in the planning, to date
  - Can give you baseline idea for things like measurement error size, effect(rough), cost, etc.
- Specify the model
  - Usually a linear model(response = baseline + treatment + error) - difference in means model
  - Different effects models
    - Fixed effects - variation among a fixed number of groups - can attach effect to each group
    - Random effects - variation among a random number of groups - cannot attach effect to each group
    - Mixed Effects(combo of above two)
- Outline the analysis
  - Sketch out computations, charts that will be produced
  - Depends on above points
  - If experiment takes a while, can spend time coding up the analysis - can just enter the data
- Calculate number of observations needed
  - Helps you stay within resource constraints
  - Depends on experimental design, size of effect you expect to measure, and strength of test - power - if there really is an effect, want to see it $x\%$ of the time
  - Also depends on variability of the data, pilot an be helpful to estimate this

  - Review and revise
    * Review all of above
    * Consider constraints(time, money, other resources)
    * Won't be able to answer question exactly as requested but can do your best

- Standard Designs:
  - Completely Randomized
    * Assign experimental units to treatments completely at random, subject to no other constraints/blocks/etc. except the number of replicants.(A/B testing)
    * Model: Response = baseline + treatment effect + error
    * Simple, requiring lots of replicants
  - Block Designs
    * Partition the EU's into blocks, determine allocation of treatments to blocks, then assigns EU's within a block to treatments at random
    * Model: Response = baseline + block effect + treatment effect + error
    * Complete block: each block gets all treatments, EU's are assigned to those treatments at random(and equally)
      · Also randomized complete block and incomplete block - not all treatments assigned within a block(leave out a different treatment in each block, so treatments are averaged out)
  - Designs with 2+ blocking factors
    * Crossed: row-column design - every combination of blocking factors occurs
    * Nested(Hierarchical) - when one or more blocking factors are grouped within another blocking factor
  - Split-Plot Designs
    * Used when one or more blocks is easy to change, but another is pretty hard.
    * EU's within a block are assigned as usual, but blocks are assigned at random to the levels of another treatment factor.

2 Designs with One Source of Variation

- Randomization - "random assignment rule"
- Assign treatments to units at random
- Subject to any constraints you might have(e.g. blocking)
- Mathematical Model:

$$Y_{it} = \mu + \tau_i + \epsilon_{it}, t = 1, ..., r_i, i = 1, ..., v$$

  - i represents treatments, t is the replicant within a treatment
  - v is the number of treatments
  - r_i is the number of replicants given the i^{th} treatment(same if in a complete random block)
  - Y_{it} is the t^{th} replicant given the i^{th} treatment
  - μ is baseline mean, τ_i are the treatment effects, ϵ_{it} are the errors
  - Errors are iid normal with zero-mean and constant variance

- Parameter(or function of such) is estimable IFF it can be written as an expectaion of a linear combination of the response values

$$E[\sum_i \sum_t a_{it} Y_{it}] = \sum_i \sum_t a_{it} E[Y_{it}] = \sum_i \sum_t a_{it}(\mu + \tau_i) = \sum_i b_i(\mu + \tau_i)$$

- If there exists b_i such that we can get an expression from $\sum_i b_i(\mu + \tau_i)$, that expression is estimable

- Cannot get individual parameters

- Dot Notation:

  - Y_{it} has two subscripts - can sum over either of those to get sums or means
  - $\bar{Y_{i\cdot}}$ is the mean of the ith treatment over the replicants in that group

$$\bar{Y_{i\cdot}} = \frac{1}{r_i} \sum_{t=1}^{r_i} Y_{it}$$

  - $\bar{Y_{\cdot\cdot}}$ is the mean of all values, or the grand mean

$$\bar{Y_{\cdot\cdot}} = \frac{1}{n} \sum_{i=1}^{v} \sum_{t=1}^{r_i} Y_{it}, n = \sum_{i=1}^{v} r_i$$

- Least Squares Estimation - want to find parameters that minimize the sum of squared errors

  - Error $= y_{it} - \mu - \tau_i = y_{it} - (\mu + \tau_i)$
  - Squared Errors: $(y_{it} - (\mu + \tau_i))^2$
  - Sum of squared error: $\sum_i \sum_t (y_i - \mu - \tau_i)^2$
  - Take all partial derivatives($\frac{d}{d\mu}, \frac{d}{d\tau_i}$) and set to 0 - finding μ, τ_i which minimizes sum of squared error - intuitively, due to second order polynomial, will always be a minimum
    - After derivation for $\hat{\mu} + \hat{\tau_i}$:

$$\bar{Y_{i\cdot}} = \hat{\mu} + \hat{\tau_i}$$

  - Gauss-Markov Theorem - Least squares estimator of any estimable function of the parameters is the unique best linear unbiased estimator(BLUE)
  - True for all linear models with independent errors and constant variance
  - Parsing out an estimator that is "unique BLUE":
    * Unique
    * Linear
    * Unbiased
    * Best

- Estimating Variance
  -
$$SSE = \sum_i \sum_t (y_{it} - \hat{\mu} - \hat{\tau_i})^2 = \sum_i \sum_t (y_{it} - \bar{y_i})$$
$$\hat{\sigma^2} = \frac{SSE}{n-v} = MSE$$
  - Where $MSE = \hat{\sigma^2}$ is an unbiased estimate for σ^2

3 ANOVA

- If $v > 2$, multiple treatments or treatment levels
  - Are there significant differences amongst the possible treatment pairs - many different possible pairs
  - Is (atleast) one treatment different from the others? - multiple comparisons
- Null Hypothesis: All treatments equal $\tau_1 = \tau_2 = ...\tau_v$
  - Alternatively: $\tau_1 - \tau_2 = 0, \tau_1 - \tau_3 = 0, ...\tau_1 - \tau_v = 0$
  - $\tau_1 - \bar{\tau_.} = 0, \tau_2 - \bar{\tau_.}, ..., \tau_v - \bar{\tau_.}$ - can only be done up to $v-1$ - if we do it this way, this forces the last v to be the mean - $v-1$ are the treatment degrees of freedom
- Alternative hypothesis: at least one of the t_i's are different from each other
- SSE is a measure of how well the model fits the data(lower = better)
- Comparing SSE under full model(each τ_i is different) and SSE under null hypothesis

model(each τ_i is the same)

$$SSE_{full} = \sum_i \sum_t (y_{it} - \hat{\mu} - \hat{\tau_i})$$
$$SSE_{null} = \sum_i \sum_t (y_{it} - \hat{\mu} - \hat{\tau})^2$$
$$\hat{\mu} = \hat{\tau} = \hat{\mu*}$$
$$SSE_{null} = \sum_i \sum_t (y_{it} - \hat{\mu*}) = \sum_i \sum_t (y_{it} - \bar{y_{..}})$$
$$SSE_{full} = \sum_i \sum_t (y_{it} - \bar{y_{i.}}$$
$$SST(\text{Sum of Squared Errors Treatment}) = SSE_{null} - SSE_{full})$$

- SST cannot be negative, since SSE is minimizing the sum of squared errors given a full model and SSE_{null} is minimizing the sum of squared errors given a null model, where the null model is a subset of the full model - measures how much the model has improved
- $SSE_{null} \geq SSE$ - full model contains null model as a special case
- If we assume data $\sim N(., \sigma^2)$ - then $\frac{SSE}{\sigma^2} \sim \chi^2_{n-\nu}$
- $\frac{SST}{\sigma^2} \sim \chi^2_{\nu-1}$
- Can then compare the ratio which has an $F-$distribution,

$$\frac{\frac{SST}{\sigma^2(\nu-1)}}{\frac{SSE}{\sigma^2(n-\nu)}} \sim F_{\nu-1,n-\nu}$$
$$\frac{SST(n-\nu)}{SSE(\nu-1)} \sim F_{\nu-1,n-\nu}$$

- Bigger F-statistic is more significant - test on the high side

4 Sample Sizes

For simple confidence interval, to calculate sample size:

$$\bar{X} \pm z\frac{\hat{\sigma}}{\sqrt{n}}$$

- What is the sample size for a given CI size - what size do I need to get error down to a certain value?

$$E = z\frac{\hat{\sigma}}{\sqrt{n}}$$

- Might need to "guess" $\hat{\sigma}$, E is determined by self-made assumption - now can just solve for $\sqrt{n}$

$$n = (\frac{z\sigma^2}{E})^2$$

5 Multiple Comparisons

Previous, one-way comparisons, now adjusting for multiple comparisons - most methods are "ad hoc" - doing something that works but no guarantee is best

- A hypothesis test is a probability calculation - multiple tests are a joined probability $P(E_1 \cap E_2 \cap ... \cap E_n)$ - if iid, then $P(E)^n$
- Several Methods:
  - Bonferroni
    - Useful for any set of m preplanned tests
    - Instead of level α, use level $\frac{\alpha}{m}$ - can be used on a few z or t tests, and for ANOVA
    - Very conservative - may miss actually significant things
  - Scheffe
    - Determined by number of treatments and number of total observations, regardless of which comparisons are of interest
    - Use formula to compute confidence interval for any "contrast" - invert the CI for a hypothesis test
    - "contrast" = linear combination of treatment parameters, where sum of coefficients = 0
      - Ex: $\tau_1 - \tau_2$ or $\tau_1 + \tau_2 - 2\tau_3$
    - Primarily will be used for comparing pairs of treatments
  - Tukey
    - Simultaneous comparisons for all difference contrasts(e.g. $\tau_i - \tau_j$
    - Shorter intervals than Bonferroni and Scheffe
    - Uses studentized range distribution for critical values
  - Dunnett
    - Treatment vs. control only
    - τ_1 as control, so $\tau_2 - \tau_1, \tau_3 - \tau_1, ...\tau_\nu - \tau_1$
  - Hsu
    - Ranking, selection, and multiple comparison with best treatment
    - Critical comparison is $\tau_i - max_{i \neq j}(\tau_j)$
    - If τ_i is global max, statistic is positive, otherwise not - includes zero case of multiple equal treatment effects
  - No really best method - more unique situation - most likely to pick Bonferroni, maybe Scheffe or Tukey - can also do computational methods

6 Checking Model Assumptions

- All models are built on assumptions
- If assumptions aren't fully met - conclusions aren't fully valid
- Assumptions are never fully met - just checking to see if assumptions are reasonable
- Need to check:
  - Check model form(form of the means)
  - Check for outliers
  - Check for independence
  - Check for constant variance
  - Check for normality
- Checked in residual plots:
  - Difference in Error and Residual
  - Error = Actual value - True mean value

$$\epsilon_{it} = Y_{it} - E[Y_{it}] = Y_{it} - \mu - \tau_i$$

  - Residual = Actual value - estimated mean value

$$\hat{\epsilon_{it}} = Y_{it} - \hat{Y_{it}}$$

- Can also look at Standardized Residuals:
  -

$$z_{it} = \frac{\hat{\epsilon_{it}}}{\sqrt{SSE/(n-1}}$$

  - Residuals have sum = 0 by properties of least squares
  - Dividing by standard error means now have standard error = 1
  - Should approximately be distributed as $N(0, 1)$
- Can also look at plots - useful for non-technical audience:
  - Plots by treatment - plotting treatment type to residual
  - Plots by run / time order
  - Histogram of residuals
  - QQ plot(normality, outliers)
- Outliers - do not simply discard

  - If not obviously miscoded - may just convey information
  - If true - consider more complicated model
- Cannot say "assumptions are met" - must say assumptions are not violated"
- Cannot prove the assumptions are correct
- If assumptions are not met:
  - Weakens statistical conclusions
  - Depending on the assumption - do different things
- If not independent:
  - If clear pattern - could add some sort of time factor to adjust for the differences
  - Consider re-running experiment using time order as a blocking factor
- Non-equal variance
  - If clear pattern - consider a variable transformation that will settle things out
  - Try log / power transformations - log is helpful for cone distributions in plots between residuals and treatment
- Non-normality
  - Least important of assumptions
  - Typically need to only be approximately satisfied
  - Small deviations are ok - mostly check tails

7 Two-Crossed Treatment Factors

- Often have more than one thing affecting your response
- One experiment with crossed treatments might be cheaper/easier to run than two experiments with single treatments
- Lot of ways that even two treatment factors can interact with each other
- Interactions:
  - Occur all the time - can be a huge determinant in which combination of treatments is optimal
  - If they exist - examining each factor separately is likely to give you incomplete information
- Mathematical Model:

  - $$Y_{ijt} = \mu + \tau_{ij} + \epsilon_{ijt}$$
    $$e_{ijt} \sim N(0, \sigma^2)$$
    $$t = 1, ..., r_{ij}; i = 1, ...a; j = 1, ..., b$$
  - Errors also independent - "cell means model"
  - $$\tau_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij}$$
  - Third term is the "interaction" term
  - Full model:
    $$Y_{ijt} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijt}$$
  - "Two way compete model" or "Two way ANOVA model"
- No interactions - main effects model:
  - $$Y_{ijt} = \mu + \alpha_i + \beta_j + \epsilon_{ijt}$$
  - Can be misleading if there really are interactions - best to use full two-way model
  - If many levels per factor - can create a huge number of parameters
  - More parameters means need more data
- Least Sqares model as
  $$Y_{ijt} = \mu + \tau_{ij} + \epsilon_{ijt}$$
- Identical form to one-way model, implying:
  $$\hat{\mu} + \hat{\tau_{ij}} = \hat{\mu} + \hat{\alpha_i} + \hat{\beta_j} + \hat{(\alpha\beta)_{ij}} = \bar{Y}_{ij.}$$
- Variance is $\frac{\sigma^2}{r_{ij}}$
- Deriving treatment estimators:
  $$\sum_i \sum_j \sum_t (y_{ijt} - \mu - \tau_i j)$$
  $$\frac{d}{d\tau_{ij}} = \sum_t -2(y_{ijt} - \mu - \tau_{ij})$$
  $$\sum_t (y_{ijt} - \hat{\mu} - \hat{\tau_{ij}}) = 0$$
  $$\sum_t^{r_{ij}} y_{ijt} = r_{ij}(\hat{\mu} + \hat{\tau_{ij}})$$
  $$\bar{y_{ij.}} = \hat{\mu} + \hat{\tau_{ij}}$$

- Essentially same estimator - same form - all $\mu+\tau_{ij}$ are estimable and contrasts of those parameters are estimable
- Particularly - differences are estimable $\tau_{ij}-\tau_{kl}$ - can't estimate τ by itself but can do differences or $\tau+\mu$
- Can estimate interaction contrasts:
  - $(\tau_{ij}-\tau_{kj})-(\tau_{il}-\tau_{kl})$

$$\tau_{ij}=\alpha_i+\beta_i+(\alpha\beta)_{ij}$$
$$\tau_{kj}=\alpha_k+\beta_j+(\alpha\beta)_{kj}$$
$$\tau_{il}=\alpha_i+\beta_l+(\alpha\beta)_{il}$$
$$\tau_{kl}=\alpha_k+\alpha_l+(\alpha\beta)_{kl}$$
$$(\tau_{ij}-\tau_{kj})-(\tau_{il}-\tau_{kl})=(\alpha\beta)_{ij}-(\alpha\beta)_{kj}-(\alpha\beta)_{il}+(\alpha\beta)_{kl}$$

  - First factor differs from i to k
  - Second factor differs from j to l
  - Might look at a bunch - finding contrast terms which cause large differences

Other contrasts of interest

- Consider all different treatment levels of one factor within the levels of the other factor - comparing all methods with each teacher
- $\tau_{ih}-\tau_{ij}$, each $i=1,...a$
- If no interactions known/assumed in advance - just use main effect contrasts:

$$\sum_i c_i\bar{\tau_{i.}}=\sum_i c_i\alpha_i,\sum_j d_j\bar{\tau_{j.}}=\sum_j d_j\beta_j,\sum_j c_j=0,\sum_j d_j=0$$

- If interactions are small - may consider the main effect, averaged over its associated interactions:

$$\bar{\tau_{i.}}=\alpha_i+(\bar{\alpha\beta}))_{i.}=\alpha_i^*$$
$$\bar{\tau_{.j}}=\beta_j+(\bar{\alpha\beta}))_{.j}=\alpha_j^*$$
$$(\bar{\alpha\beta})_{i.}=\frac{1}{b}\sum_j(\alpha\beta)_{ij}$$
$$(\bar{\alpha\beta})_{.j}=\frac{1}{a}\sum_i(\alpha\beta)_{ij}$$

- In general - with possibly different sample sizes - main effect estimates are:

$$\hat{\alpha_i}*=\frac{1}{b}\sum_j\bar{Y_{ij}},\hat{\beta}_j^*=\frac{1}{a}\sum_i\bar{Y_{ij}}$$

- a and b are the number of levels in the first and second factor respectively
- If sample sizes are reduced across cells, this reduces to:

$$\hat{\alpha_i^*} = \bar{Y_{i..}}, \hat{\beta_j^*} = \bar{Y_{.j.}}$$

- In equal sample case, difference in levels for main effects reduces to
  - $\hat{\alpha_i^*} - \hat{\alpha_h^*} = \bar{Y_{i..}} - \bar{Y_{h..}}, Var = \frac{2\sigma^2}{br}$
  - $\hat{\beta_j^*} - \hat{\beta_k^*} = \bar{Y_{.j.}} - \bar{Y_{.j.}}, Var = \frac{2\sigma^2}{ar}$
- Estimating σ^2:
  - $SSE = \sum_i \sum_j \sum_t (y_{ijt} - \bar{y_{ij.}})^2$
  - To get MSE, need the right divisor, adjusting for df:

$$MSE = \frac{SSE}{n - a * b}$$

- For confidence intervals / hypothesis test on a set of contrasts(i.e. difference in main effects)

$$\sum_i c_i \bar{\tau_{i.}} \in (\sum_i c_i \bar{y_{i..}} \pm w \sqrt{MSE \sum_i c_i^2 / br})$$

- w depends on which multi-comparison test you're using
- In particular, for two levels of the first main effect factor:

$$\bar{\tau_{i.}} - \bar{\tau_{j.}} \in (\bar{y_{i..}} - \bar{y_{i..}} \pm w \sqrt{MSE * 2/br})$$

- One Observation per Cell / Single Replicant Experiments
  - When $r = 1$
  - CI's wide, hypothesis tests weak
  - σ^2 cannot be estimated
  - Can:
    - Assume a value for σ^2
    - Assume only a few of the contrasts are likely to be non-negligible
    - Analyze orthogonal contrasts

8 Complete Block Design

- Used to reduce effect of unwanted factors in analysis(time of day, operator, machine, geographic) - often present - looks to reduce unwanted variance
- Know it will have an impact but don't care about that impact
- What Kind of effect is this?
  - Block
    - Set of conditions that you can group together in an experiment for the purpose of minimizing variance
    - Use blocks when you want to know the average effect of treatment over a range of conditions
    - Conditions that vary from block to block = blocking factors
    - Blocking factors could be covariates
    - Often, things that aren't conveniently measured
  - Noise
    - Controllable in a lab, but not in the "real world" - like environmental conditions
    - Ideally - want experiment to be minimally affected by differences in noise variables
    - Noise factors are essentially ignored
  - Covariate
    - Cannot be controlled, can be measured prior to or during the experiment
    - Expected to have greater effect than noise variables
    - Could be of interest in their own right, but could also be added into the design to minimize their effect on results
- Type of unwanted effect tells you how to classify it
- Class of effects tells you what to do
- Not always obvious which is which(contextually dependent)
- Complete Block Designs
  - Assign all treatment possibilities to each block, equally so
  - Requires number of replicants per block to be a multiple of the number of treatments
  - Within a block, randomly assign replicant to treatment
  - Randomized complete block design - each treatment occurs once per block
  - General complete block design - each treatment occurs > 1 times per block

  - Model: $Y_{hi} = \mu + \theta_h + \tau_i + \epsilon_{hi}$
  - ϵ are assumed normal nad independent - note similarity to two-way main effects model
  - If we include interaction, we'd run out of degrees of freedom - no interactions is typically reasonable
  - In two-way ANOVA, both effects are randomly assigned - only treatment levels are randomly assigned
  - Blocks are not important thing - there to soak up variance
  - Can run aov() and just ignore the block factor results
  - Two general models: $Y_{hi} + \mu + \theta_h + \tau_i + \epsilon_{hi}$ and $Y_{hi} = \mu + \theta_h + \tau_i + d(\theta\tau)_{hi} + \epsilon_{hi}$ - main effects nad interactions model
  - Only use second if interactions are strong

- Incomplete Block Design
  - When number of replicant per block is NOT equal to a multiple of the treatment levels
  - Very common, often has to do with constraints
  - Assigning with randomness
    * Cannot assign every treatment to every block
    * Assign treatments to each block
    * Assign treatments within block to replicants at random
  - First step is key to having an analyzable experiment
  - Simplest design is cyclic:
    * Start with a set of treatments
    * Then cycle through
    * Each treatment occurs same number of times
    * But not always with each other treatment

- Can treat blocks like treatment factors when blocking

9 The Statistical Bootstrap

- Never give just point estimates
- Always include some measure of uncertainty - quantifying
- Initially done mathematically
  - Suppose normally distributed data - $x_1, x_2, ...x_n$
  - Suppose mean μ and standard deviation σ

  - Sample mean $\bar{x}$ is normally distributed with mean μ and standard deviation $\sigma\sqrt{n}$(sample distribution)
  - Standard error = std dev of sampling distribution
  - Computing CIs: $x \pm z\sigma/\sqrt{n}$
- If not normally distributed
  - Can rely on CLT - sample mean is approximately normally distributed with same mean and standard error from before
- For unknown distributions or statistics other than mean - use bootstrap
  - Non-normal data
  - Any statistic
  - Small-ish data
- Uses compute power to estimate uncertainty
- Start with random sample: $x_1, x_2, ...x_n$
- Take bootstrap sample: $x_1^*, x_2^*, ..., x_n^*$
  - Same sample size
  - From original sample
  - Sampled with replacement
- Compute the statistic
- Do the above 2 steps a lot of times
- Compute confidence intervals from the results
  - Standard deviation of bootstrap statistic is approximate standard error of original statistic
  - Distribution of the bootstrapped statistic is approximately the sampling distribution of the statistic
    - Can use to compute quantile-based CIs or anything else you wish to know
  - Bootstrap could be better than theory if potential violations of normality assumptions
  - Benefits
    - Can use for pretty much any statistic
    - No specific distribution assumptions required
    - No distribution calculations required
    - Only one assumption needed - data is a representative random sample from the population

  - Not great for
    - Extreme values
    - Tail probabilities
    - Extrapolation
    - Very small data
    - Designed for uncertainty computations - don't use for things it wasn't designed for

- Bootstrap Error - comes from fact that we are estimating these quantities - with more "data", gets smaller
  - Taking 10000 or more samples generally gets error low enough not to worry

10 Bootstrap Background

What makes bootstrap work?

- Probability
  - Foundation of statistics - given a distribution's parameters, what can we say abot the events
  - Statistics - given events, what can we say about the data generating distribution's parameters
- Compute
  - Can perform large numbers of simulations in seconds
  - Faster to compute BS distributions than it is to (try and) compute distributions of statistics
- Empirical Distribution
  - Distribution where probability of 1/n is placed on each of the n data points in your sample
  - Distribution of the data
  - For a random sample, empirical distribution approximates the population distribution
- Plug-in Principle
  - Sometimes write paramrters θ as a function t() of the distribution F, $\theta = t(F)$
  - If we want estimate of the paramter $\hat{\theta}$
  - Can compute the function t() of our best estimate of the distribution $\hat{F}$
  - Putting it all together: $\hat{\theta} = t(\hat{f})$

  - Can estimate parameter by computing the function of an estimate of our distribution

- Empiral Distribution + Plug-in
  - If we assume original sample is approximately the population, can resample from it(bootstrap sample)
  - Can compute statistics from these samples(the Bootstrap statistic)
  - We assume this is approximately our population - can sample from this "population" as many times as we want(the BS samples)
  - This mimics what would happen if we had taken samples of size n multiple times
  - Mimic uncertainty of the statistic under sampling
- Say want to compute expectation $E[X]$ - $E[X] = \int xf(x)dx$
- If we think in terms of empirical distribution - integral becomes a sum
- $f(x)dx = \frac{1}{n}$
- xs are data points $x_1, x_2, ..., x_n$
- Above integral becomes $\frac{1}{n}\sum_{i=1}^{n} x_i$
- Sample mean is plug-in estimator for the population mean using the sample data as the empirical distribution
- Bootstrap Samples
  - B samples - let B be 10000 or more - can be smaller for SE's but for CI's nicer to have more
  - Using original sample as the empirical distribution - approximation of the population
  - With replacement mimics independence in sampling
  - Of size n - mimicing sampling a sample just like the one we have
- Calculate Stats
  - Plug-in principle directly
  - For each BS sample - compute statistic of interest, mimics a sample from population
  - Works for complicated statistics
- Calculate SE's or CI's
  - Uses distribution of calculated statistics

  - Another use of empirical distribution - approximating sampling distribution of the statistic

- Computing BS Standard Errors
  - Sample B bootstrap samples(sample with replacement from the original sample)
  - Compute the statistic of interest for each of the B samples
  - Take the sd() of the statistics to estimate the SE

11 More on Standard Errors and Confidence Intervals

- Standard Error of a mean
  - Generate B bootstrap samples from original dataset
  - Compute mean on each bootstrap sample
  - Compute standard deviation of computed means
- Standard Error of a correlation coefficient
  - Generate B bootstrap samples from original dataset
  - Compute correlation coefficient on each bootstrap sample
  - Compute the standard deviation of the computed correlated coefficients
- Statistics - anything that can be computed from data
- Standard Error of any random statistic
  - Generate B bootstrap samples from original dataset
  - Compute random statistic on each bootstrap sample
  - Compute standard deviation of the computed random statistics
  - Should BS sample over entire data set as a whole since original sample is done over whole - BS sample should match original
- Bootstrap Failures
  - Fails for extremes, tail probabilities, anything on the edge
  - Situations where - probability of seeing a value larger than any we've seen, what is value of a 1 in a 100 year insurance industry catatrophe event - if don't have 100 years of data
    * Rely on classical statistics
    * Estimate distributions based on characteristics of what you're estimating - then use distributions to estimate the tail/extreme probabilities/values
    * Or use Extreme Value Theory

12 Two-Sample Data

- Previously done one-sample
- Assume population distribution exists, sample is representative of it
- Resample from this sample to perform the bootstrap
- Two-sample problems:
  - Generalize to multi-sample
  - Two-sample is not the same as two-value single sample(two attributes from one sample)
- Two-Sample vs One-Sample
  - Population: One-sample uses one population, two-sample uses two populations
  - Data Structure: One-sample uses multiple values in a single row of data, two-sample means that each value has its own row
  - Dependence / Independence - one-sample: entire row of data is dependent, for two sample - each data point is independent of each other
  - How to sample? One sample: whole data row is sampled(like in law data), two-sample: each data point is sampled from each population(like incorrect law data alternative sampling scenario)
  - When to use - one-sample uses single t-test or paired t-test while two-sample uses two-sample t-test(extends to multiple comparisons)
- Two-Sampel Algorithm:
  - Sample B bootstrap samples from each population of interest
  - Compute the statistic of interest from the bootstrap samples
  - Compute SE's or quantile-based CI's using these calculated statistics

13 More on Bootstrap and Random Testing

- Randomization Test - non-parametric way to do hypothesis testing
  - Uses resampling, no distributional assumptions
  - Only one assumption
    - Assume null hypothesis(one assumption) - generally "no difference"
    - aka distribution generated under treatment and control are the same
    - aka treatment-control labels are meaningless
    - aka only difference in treatment-control is randomization

- Randomization Test Algorithm
  - Assume null hypothesis
  - Resample / reassign treatment=control labels(WITHOUT replacement)
  - Compute test statistic
  - Compare the test statistic from the original sample with the resampled distribution you generated - this is the p-value
- Randomization vs. Bootstrap
  - Sample - BS: with replacement, RT: - without replacement
  - Key Assumption - BS: data is representative of population, RT: null hypothesis is true, labels are meaningless
- BS Hypothesis Test
  - Assume null hypothesis is true
  - Pull a B bootstrap samples of the full dataset from the combined data(n treatment and m control for n+m total values)
    - Sample length, with replacemejt
  - Assign the first n values to treatment and the last m to control
  - Compute the test statistic
  - Determine the number of times the computed statistic, in absolute value, is larger than or equal to the original test statistic
- BS Test - with replacement vs Randomization Test - without replacement
- Randomization Test Limits
  - Doesn't work in one-sampel case - no labels
  - Doesn't give confidence intervals, just p-values

14 Linear Regression and Bootstrap

- Full data looks like $x_{i1}, ..., x_{ip}, y_i$
- x's are covariates, y's are response
- $\mu = E[y_i|(x_{i1}, ..., x_{ip})]$
- Linear form $\mu_i = \sum_{i=1}^{p} x_i\beta_1$
- Probability-wise: $y_i = \sum_{i=1}^{p} x_i\beta_i + \epsilon_i$ - $\epsilon_i \sim N(0, \sigma^2)$
- Rows of data are independent of each other - want to find inteference on β's

Use when:

- Assumptions aren't fully met
- Regression isn't linear in the β's
- Methods other than the least-squares are used
- Non-normal errors

How to BS linear regression:

- Can BS the data - mimic data pull from the population
  - Pull B bootstrap samples of data from dataset
  - Calculate linear regression on each BS dataset
  - Save the β's form the calculation
  - Use distribution of saved $\beta's$ as an approximation to the sampling distribution
  - Compute CI's and/or SE's
  - SE's are pretty close to SE's from full-assumption - can also compute CI's on coefficients with quantiles
- Can BS the residuals - assume model is "right" but want SE's of the β's
  - By treating covariates as fixed constants can obtain a SE that reflects the precision associated with the sample of covariates actually observed
  - More frequentist while BS'ing the data is more bayesian
  - Compute the linear regression
  - Save the fitted values and residuals
  - Take B bootstrap samples of the residuals
  - Compute new y's as fitted.values + residuals_bs
  - Compute BS linear regression
  - Save BS coefficients
  - Compute standard errors / CI's as before
- Regression BS used for
  - Non-linear model
  - Not using least squares to optimize(least median square residual fitting)
  - Non-normal errors
  - Model assumptions just aren't met

15 Measuring Bias

- Bias
  - Unbiased: $E[\hat{\theta}] = \theta$
  - Biased: $E[\hat{\theta}] = \theta + b$
  - Formally: $bias_F = bias_F(\hat{\theta}, \theta) = E_F[s(x)] - t(F)$
  - Unbiased is generally good but may times trade off bias for variance
- Enter the BS
  - Using empirical distribution and plug-in principle:

$$bias_{\hat{F}} = E_{\hat{F}}[s(x^*)] - t(\hat{F})$$

  - Can compute via:

$$\hat{\theta^*}() = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta^*}(b)$$

  - Then we get $\hat{bias}_B = \hat{\theta^*}() - t(\hat{F})$
- Useful when computing ratios of means and SDs - often with some kind of bias
- Can give improved estimate of bias using resampling vector
  - Vector of probabilities that reflect the resampled dats P*
  - Consider data $x = (2, 4, 6, 8, 10, 12, 14, 16, 18)$ and bootstrapped dats $bs_x = (14, 18, 2, 6, 8, 2, 8, 12$
  - See 2 twice, 4 twice, 8 twice, and 20, 16, and 10 not at all
  - Resampling vector would be ($P^* = (0.2, 0.2, 0.1, 0.2, 0.0, 0.1, 0.1, 0.0, 0.1, 0.0)$
  - Denotes probabilities of the resampled data with respect to the original data
- If you think of orignal data vector, if no repeats then the resampling vector is a vector of 1/ns $P^0 = (1/n, ..., 1/n)$
- Each bootstrap sample i gives rise to a resampling vector P^{*i}
- Let $\bar{P^*} = mean(P^{*i})$
- Can re-write original bias computation as $\hat{bias}_B = \hat{\theta^*}() - t(P^0)$
- Imrpoved bias computation is now $\hat{bias}_B = \hat{\theta^*}() - t(\bar{P^*})$

Least Median Squares

- LMS Regression is more robust to outliers

- Minimizes median square residual

$$MSR(\beta) = median(y_i - x_i\beta)^2$$
$$MSR(\beta) = \hat{} min_\beta[MSR(\beta)]$$

- $\hat{\beta}$ is the value that minimizes MSR(β) - $\hat{\beta}$ is the value that minimizes MSR(β)
- One or several wrong values won't really affect it
- Used when outliers / miscoding / data errors are a concern
- Gives similar values as OLS
- Doesn't have same theoretical computations for SE's or CI's

16 Jackknife Estimator

- Another estimate of bias
- Similar to bootstrap - OG computer-based estimator for bias and SE's, also called "leave one out" estimator
  - Data = $x = (x_1, x_2, ..., x_n)$
  - leave one out: $x_{(i)} = (x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n)$
  - Done for each of $i = 1, ..., n$
  - Also let $\hat{\theta_{(i)}} = s(x_{(i)})$ be the statistic applied to the i-th leave one out sample
  - And let $\hat{\theta}_() = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta_{(i)}}$
  - Jackknife estimate of bias as $\hat{bias}_{jack} = (n-1)(\hat{\theta}_() - \hat{\theta})$
- Very fast - only n jackknife computations
- Still pretty accurate
- Systemic, jackknife influence values, and scaled up
- Bootstrap
  - Sample with replacement is similar to original sample
  - Uses randomization to average out over differences
- Jackknife
  - Leave out out is similar to the original sample
  - Systemically works through the data to get best estimate possible
  - Generally much faster but uses less information(less efficient)

  - JK is the linear approximation to the BS
  - For smooth statistics(like mean), the JK is a good approximation
  - For non-smooth statistics, JK can fail - BS is more general

- Jackknife Failure
  - JK can fail with non smooth data
  - BS computations of median compute different values almost every time - JK doesn t always do this

- Delete-d JK
  - Used to regain consistency for non-smooth statistics
  - Instead of leave-one-out, you delete-d
  - $d > \sqrt{n}, n < d$ - similar to cross-validation
  - $SE = \frac{r}{\binom{n}{d}} \sum(\hat{\theta_{(s)}} - \hat{\theta_{()}}))^{1/2}$
  - $n = r \cdot d$; s indexes the samples with d removed

17 Cross Validation + Prediction Error

17.1 Cross Validation

- Estimated prediction value when not enough for holdout sets
- Error associated with predicting new values
- Residual is bad as
  - Minimized to make the model
  - Always be smaller than prediction errors
  - Need to predict on data the model hasn't seen
- Split into k equal parts
- For kth part, fit on other $k-1$ folds, use to predict kth part
- Repeat for each k, combine all to estimate prediction error
- $CV = \frac{1}{n} \sum(y_i - \hat{y_i}^{k(i)})^2$

17.2 Residual Standard Error Adjusted

- MSE = SSE / n - biased low
- Unbiased $\hat{\sigma} = SSE/(n-p)$
- PE estimate = SSE/(n - 2p)
- Mallows Cp = $SSE/n + 2p\hat{\sigma}^2/n$
- $BIC = SSE/n + log(n)p\hat{\sigma}^2/n$
- Prediction - on new data, inference - info on parameters
  - Importance depends - if want to know why/what - infer
  - Know future - prediction
- CV useful if can't have holdout set
- BS prediction error
  - Fit model on Bs
  - Generate prediction for original
  - Generate prediction for BS
  - Diff of mean error(optimism)
  - Do B times
  - Average over B
  - Mean prediction error + optimism is estimate
- Why does this work?
  - We are getting original MSE and BS vector
  - Dif, we subtract out things that occur at least once
  - Subtracting overseen from underseen - assume we do good on what we see a lot bad on what we don't
- .632 Estimator
  - Average underestimation with overestimation
  - .632 - theoretical argument - approximate probability given observation appears in a BS sample of size n
  - Takes prediction error from original vector on BS models, average with error rate from out of bag values
- Out of Bag

  - Data not in BS
  - These are unseen by BS
  - More likely to be were prediction
  - Average with underestimate, weigh .632, .368
- CV - unbiased large variance
- BS - lower, variance biased low
- 632 - tends to perform best

18 Assessing Bootstrap Error

- Can estimate standard error in our bootstrap estimates - use bootstrap to assess variability in our estimates
- Bootstrap process has error inherently in it
- Error comes from two sources:
  - Sampling variability
  - BS resampling variability
- Process:
  - Step 1: get sample from the population - this is sampling variability
  - Step 2: Run B BS samples with replacement - this is the BS resampling variability
  - No control over step 1 but can reduce step 2
- Can assess error on case by case basis - easy case:
  - Assume s(x) is the sample mean
  - Want to estimate standard error of this statistic, data is normal
  - Want $var(\hat{se}_B) = var(E[\hat{se}_B|x]) + E[var(\hat{se}_B|x)]$
  - Let $\hat{m}_i$ be the ith moment of BS resampling distribution $= var(\sqrt{\hat{m}_2}) + E[\frac{\hat{m}_2}{4B}(\hat{\Delta} + 2), \hat{\Delta} = \frac{\hat{m}_4}{\hat{m}_2{}^2} - 3$

$$var(\hat{se}_B) \approx \frac{\sigma^2}{2n^2} + \frac{\sigma^2}{2nB} = \frac{\sigma^2}{2n}[\frac{1}{n} + \frac{1}{B}]$$

  - True given the "easy" assumptions made on previous slide - shows that variance of estimator is dependent on three things:
    * Variance of underlying population
    * Sample size

    - Number of BS samples
  - As σ increases, variance of estimator increases
  - As n increases, variance of estimator decreases
  - As number of BS samples increases, variance of estimator decreases - only thing that can be controlled

- Sometimes interested in SE w.r.t size of the sample mean
- Can reflect this in coefficient of variation, $CV = \frac{\sigma}{\mu}$, in theory
- In our case, want $cv(\hat{se}_B)$
- By theory, that gives us $cv(\hat{se}_B) = \frac{var[\hat{se}_B]^{1/2}}{E[\hat{se}_B]}$ - plugging in:

$$cv(\hat{se}_B) = \frac{var[\hat{se}_B]^{1/2}}{E[\hat{se}_B]}$$
$$E[\hat{se}_B] = \frac{\sigma}{\sqrt{n}}$$
$$var(\hat{se}_B) \approx \frac{sigma^2}{2n}(\frac{1}{n} + \frac{1}{B})$$
$$cv(\hat{se}_B) = \frac{\sigma/\sqrt{n} * \sqrt{1/2 * (1/n + 1/B))}}{\sigma/\sqrt{n}} = \sqrt{(\frac{1}{2n} + \frac{1}{2B})}$$

- Only depends on n and B - can only control B

Quantile Derivation:

- Using the CV computation:

$$cv(\hat{q_B^{\alpha}} \approx c(\alpha)\sqrt{(\frac{1}{n} + \frac{1}{B})}$$

- $c(\alpha)$ is some constant dependent on alpha
- Get a reduction in variance relative to the quantile value of similar kind to the SE estimate of the mean
- When we crunch the numbers, find that $B > 500$ or $B > 1000$ is typically good

Jackknife-after-bootstrap(JAB)

- Used to estimate uncertainty in the SE estimate - combines two methods previously studied
  - For i in 1,2,...n(number of datapoints)
  - Let $x_{(i)}$ be the original but with the ith data point removed

  - Compute a bootstrap estimate of SE of your statistic using $x_{(i)}$ as your original data
  - Compute uncertainty as $\hat{var}_{jack}(\hat{se}_B) = [(n-1)/n]\sum_i^n (\hat{se}_{B(i)} - \hat{se_{B()}})^2$

- Biggest issue with JAB is the resampling - must do a full BS resampling for each data point you remove from the original data vector
- Alternative - can estimate $\hat{se}_{B(i)}$ from the bootstrap samples that don't have the ith data point
- Works due to the JAB Sampling Lemma - a BS sample drawn with replacement from $x_{(i)}$ has the same distribution as a BS sample drawn from x in which none of the BS values equals x_i
  - Means that we can use the BS samples in a standard BS that don't have the ith value in them, and this is the same sampling distribution as if we had jackknifed the data and BS sampled from that (full JAB)
  - Can estimate JAB using BS values only
- JAB shortcut - but not best estimate - not great as we're using the same BS values over and over again to estimate each JK SE

BS CI's

- CI's combine point estimates with uncertainty - invertible as hypothesis tests
- Rely on assumptions - if not met - CI's can become too small or coverage is not actually as high as stated
- Correctness and Accuracy
  - For parameter θ and coverage level α, would like to consider the one-sided confidence level defined as $P(\theta \leq \hat{\theta}[\alpha]) \approx \alpha$
  - $\hat{\theta}[\alpha]$ is the computed one-sided confidence interval boundary with coverage α
  - Say first order accurate if $P(\theta \leq \hat{\theta}[\alpha]) = \alpha + O(n^{-1/2})$
  - Second order accurate if $P(\theta \leq \hat{\theta}[\alpha]) = \alpha + O(n^{-1})$ - big-O notation, asymptotic behavior of the estimators
  - Say $f(x) = O(g(x))$ if $\exists M, x_0$ such that $\forall x \geq x_0$, $|f(x)| < M|g(x)|$ - function f is bounded by g
- Correctness $\neq$ Accuracy
  - Suppose we have an estimator $\hat{\theta}_{exact}$ that is "exact" - $P(\theta < \hat{\theta}_{exact}[\alpha]) = \alpha$
  - Confidence point is "first order correct" if $\hat{\theta[\alpha]} = \hat{\theta_{exact}} + O(n^{-1})$
  - "Second order correct" if $\hat{\theta[\alpha]} = \hat{\theta_{exact}} + O(n^{-3/2})$
  - Correctness $\implies$ accuracy at the same "order"

- BC_α method
  - Corrected CI using quantiles
  - Second order accurate and second order correct
  - $P(\theta \leq \hat{\theta_{BC_\alpha}}) = \alpha + O(n^{-1})$
  - $\hat{\theta_{BC_\alpha}}[\alpha] = \hat{\theta_{exact}} + O(n^{-3/2})$
  - Typical α-level CI is given by $(\hat{\theta_{lo}}, \hat{\theta_{hi}}) = (\hat{\theta}^{*(\alpha/2)}, \hat{\theta}^{*(1-\alpha/2)})$
  - BC_α is given by $\theta_{lo}, \hat{\theta_{hi}}(\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)}) =$, where we compute α_1, α_2

$$\alpha_1 = \Phi(\hat{z_0} + \frac{\hat{z_0} + z^{(\alpha/2)}}{1 - (\hat{\alpha}/2)(\hat{z_0} + z^{\alpha/2})})$$

$$\alpha_2 = \Phi(\hat{z_0} + \frac{\hat{z_0} + z^{(1-\alpha/2)}}{1 - (\hat{\alpha}/2)(\hat{z_0} + z^{1-\alpha/2})})$$

  - Need z_0 hat and α hat, Φ is normal CDF
  - $\hat{z_0} = \Phi^{-1}(\frac{num(\hat{\theta^*}(b) < \hat{\theta})}{B})$ - called the bias
  - $\hat{\alpha} = \frac{\sum_{i=1}^n (\hat{\theta_{()}} - \hat{\theta_{(i)}})^3}{6(\sum_{i=1}^n (\hat{\theta_{()}} - \hat{\theta_{(i)}})^2)^{3/2}}$ - called the adjustment, higher order error
- $z_0 - hat$ is counts, converted to standard normal
- $\alpha - hat$ is just a variant of the jackknife
- Run bootstrap analysis and then jackknife analysis
- Combine to get the $BC\alpha$ CI
- BC_α - second order correct and accurate while standard BS CI's are only first order accurate
  - Transformation preserving
  - Only method that is both of the above - use BC_α if want better CI's
    - Original BS depends on data
    - Can't predict anything more extreme than what's in the OG data - under-predicts the tails
    - BC_α method adjusts the bias($z_0 - hat$) in the data and second order errors
    - Using quantiles doesn't rely on distributions - so Transformations produce α-equivalent CI's

Summation:

- Bootstrap assumes least yet provides good statistical analysis
- Useful across wiede array of situations
  - Hypothesis tests, CI's, SE's
  - One sample, two sample, ANOVA, regression models
  - For any statistic(function of data) you wish to analyze

19 Bootstrap Review

Basic Bootstrap:

- Used to estimate properties of the sampling distribution of a statistic
  - Standard Errors
  - Confidence Intervals
- Estimates the sampling distribution of the statistic in question - reliant on empirical distribution or plug-in principle
- Basic Algorithm:
  - Start with original data
  - For each B bootstrap iteration
    - Select a sample of the original data with same length and replacement
  - Compute statistic in question
  - Save the value computed
- For SE's - compute standard deviation of the B computed statistics
- For CI's - compute the quantiles of the B computed statistics
  - These work for translations of the data or statistic too
- Take z-score times SE above and below statistic computed from original data
- Only assumption is that data is representative random sample of the population

Bootstrap Hypothesis Testing:

- Method 1:
  - Sample with replacement from each of test and control
  - Calculate test statistic
  - Save those statistics
  - Compute p-value as percentage of times you recorded a statistic as high or higher
  - Assumes that each of test and control comes from own distribution - doesn't assume distributions have different parameter of interest
  - Estimates the difference when repeatedly sampling from each population
- Method 2:
  - Sample with replacement from total vector of responses
  - Assign first m to control and last n to test(if you had m control and n test in original scenario)

  - Calculate the test statistic
  - Save those statistics
  - Compute p-value as percentage of times you recorded a statistic as high or higher
  - Assumes labels have no meaning
  - Assumes the populations for test and control are the same - akin to null hypothesis
  - Stronger assumption

Randomization Test:

- Similar to method 2 of Bootstrap test
- Just pull without replacement instead of with replacement
- Equivalent to reordering the test-control labels to the data values
- Assumes labels are meaningless and the values could have come from either of test or control equally

BS Regression on the data

- Computes regressions on each BS'd dataset
- Saves the coefficients form each BS iteration
- Computes SE's and CI's from those coefficients
- Assumes data is a random sample of the population
- Captures both model uncertainty and sampling uncertainty
- In general, preferred method

BS Regression on residuals:

- Create BS dataset by adding BS'd residuals to the fitted values
- Compute regression on BS datasets
- Save the coefficients from the regression
- Compute SE's and CI's from those coefficients
- Assumes model form is correct
- Doesn't capture model uncertainty like the other method does

BS Estimate of Bias:

- Do usual bootstrap

- Compute mean of BS statistics and subtract the parameter function applied to the original data
- Not so useful for unbiased statistic like sample mean
- Better for things like ratios which can have significant bias

Prediction Error Estimation

- If you care more about prediction than inference
- More important in ML or forecasting
- Several methods:
  - Holdout sets
  - Cross-validation
  - Jackknife (CV subset)
  - Ad hoc methods

Cross-Validation:

- Split data into K groups of roughly equal size
- For each of the K groups (1,...,K)
  - Let group k be the holdout set
  - Build the model on the other k-1 groups
  - Use that model to predict the values for the holdout set
  - Save those predictions(associated with their true values)

Jackknife:

- Subset of cross-validation where K is the number of data points
- For each data row
  - Build the model on the rest of the data
  - Use that model to predict the response value of the data row held out
  - Save those predictions

Ad Hoc prediction error estimation:

- Adjusted SSE(divided by n-2p instead of n-p)
- Mallow's C_p
- BIC

- 632 Bootstrap Estimator
- All of the above are adjustments to the unbiased estimate of standard error(SSE divided by n-p)

Jackknife - Bias estimation

- Can also use JK to estimate bias
- Calculate each JK statistic(on the leave-one-out JK datasets)
- Average those statistics
- Subtract the statistic on the OG data from the JK average
- Multiply by n-1

Better BS confidence intervals

- BC_α method
- Adjusted quantile-based CI's
- Second-order correct and accurate - while simple quantile-based CI's are only first-order correct and accurate
- Standard Errors
- Confidence Interval
- Prediction Error
- Linear Regression
- Bias Estimation

20 Forecasting

- Can forecast seasonality, trends, external impacts, and patterns
- Things that have enough data to make a prediction
- Things where one-off events can be explained away
- Short-term - more accurate, actionable
- Long-term - more forgiving, directional

Forecasting Methods

- Qualitative Forecasting

  - Non-numerical predictions
  - Predicting in the absence of data
- Quantitative Forecasting
  - Useful for predicting numbers
  - Time series modeling
  - ML modeling
  - Anything statistical

Types of data

- Numerical with regularly-spaced intervals
- Historical values of interest
- Additional variables of interest

Types of models

- Explanatory: $y_{t+1} = f(x_t, w_t, z_t, \epsilon_{t+1})$
- Time Series: $y_{t+1} = f(y_t, y_{t-1}, \epsilon_{t+1}$
- Mixed: $y_{t+1} = f(x_t, w_t, z_t, y_t, y_{t-2}, \epsilon_{t+1})$

Time Series

- Relationships with other variables might not be understood
- May need to predict other variables(could be another model in itself)
- Might be reasonable to assume that future values depend on prior values and patterns of prior values

Forecasting Steps

- Problem Definition
- Gathering Information
- Exploratory Analysis
  - Always look at data
  - Can you determine patterns that you can exploit
  - Do you find outliers that need to be explained
  - Is there an actual relationship between explanatory variables and the variable to be predicted

- Model Fitting / Selection
- Model Evaluation
  - What is modeling success criterion
  - Does your model meet it?
  - Is it explainable to your partners / clients?
  - General question - does it solve the real-world problem?

Uncertainty

- Tough to make predictions, especially about future
- Statistic uncertainty is very important in forecasting - include some measure of uncertainty
- Show prediction bands / intervals instead of point estimates
- Can break down patterns into three general categories
  - Trend - long-term movement, typically "increasing" or "decreasing"
  - Seasonality - short-term factors, fixed and known, if know calendar, know effect
  - Cyclic - long-term rising and falling but not fixed duration, often 2 years long or longer - think "business cycle"
- Can also add "outlier" as additional insight - model should adjust for these one-off effects

Seasonal Plot

- Plots seasonal duration as x-axis
- Values of interest on the y-axis
- Shows separate lines for each year
- Handy for seeing seasonal patterns in one plot

Multiple Possible Seasonality

- Might have mutliple seasonality periods - yearly, daily
- Can get this by specifying the period argument in gg_season

Seasonal subseries plot

- Splits up seasonality into separate plot - each season its own plot
- Can look across plots to see patterns - use gg_subseries()

Scatterplots

- Can use scatterplots to investigate how two time series interact with each other
- Correlation coefficient must be interpreted with context, different graphs could yield came correlation

Log Plot

- Plots of data versus data from k periods ago
- Help find seasonal relationships
- Can reinforce what you see in other graphs

Autocorrelation Functions(ACF)

- Time-series-specific plot
- Shows strengths of correlation for the time series vector with itself at various lags
- Correlation between $(y_i, ..., y_{t-k}), (y_{k+1}, ..., y_t)$
- Emphasizes seasonal pattern if it exists