Final Review STAT 505 Fall 2014

Linear model: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim (0, \sigma^2 \boldsymbol{V})$

All the material on the midterm is fair game. In particular, I'm guessing you'll need to interpret output from some sort of linear (perhaps generalized) model. Since then we have covered chapters 3 – 10 in Gelman & Hill.

Chapter 3 Regression Basics

- Advantages of centering and perhaps scaling predictors.
- Compute var-covar of $\boldsymbol{\Lambda}^\mathsf{T}\widehat{\boldsymbol{\beta}}$ and use it for CI's or an F-test.
- The usual assumptions.
- Diagnostic plots.

Chapter 4 More Regression

- More centering and standardizing – changes in interpretation.
- Relationship of slope to correlation in SLR.
- Log transform used to make a skewed response or predictor more symmetric – brings in long right tail. Interpret "additive in log scale" as "multiplicative in original scale".
- Box-Cox transformations
- Problems with identifiability
- Authors recommendations on when to drop predictors from a model.
- Extensive modeling of mesquite data
- Use of the same model multiple times – in this case for different presidential elections from 1972 to 2000.

Chapter 5 Logistic regression

- Use logit transform to map [0,1] to $\Re$, know inverse logit as well.
- Model P(voting Rep) based on income.
- Slope in prob scale is $\widehat{\beta_1}\widehat{p}(1-\widehat{p})$. Crude conversion to $\widehat{\beta_1}/4$ works well for $\widehat{p}$ near .5, not well near 0 or 1 where slopes are flatter.
- Interpretation as additive in log odds or multiplicative in odds.
- Bangladeshi well switching example. Pay attention to interpretations, interactions.
- Can predict response $= 0$ for low probabilities, 1 for higher ones using cutoff c (often .5). Look at the overall error rate of a model. Is it better than always going with the majority?
- Compare models using deviance.
- Logistic regression fails for a predictor with complete separation (use `bayesglm`).

Chapter 6 GLM's

- link, inverse link, and variance functions.
- Poisson, negative binomial and quasi-Poisson use log link.
- Binomial data uses logit, probit, or robit link.

- Poisson example: traffic accidents. Interpret x (speed) effect as multiplier on expected counts.

- Use offset when a predictor scales directly – then we're working with rates. Example: Stop and frisk counts use log(arrests) as an offset.

- Overdispersion: measure as residual deviance over residual df.
  Use negative binomial (or quasi-Poisson adjusts SE's).

- Binomial is like Chapter 4, but grouped together. Also need to check for overdispersion.

- Ordered categories – use multinomial with logit or probit link. Must assume the effect of predictors is the same across levels, just cutpoints change (proportional odds).

- Can also see zero-truncated or zero-inflated models.

- Constructive choice models assume we can put a value on both options.

Chapter 7 Simulation

- Works for complex and non-normal data.

- Births involve singles and 2 types of twins.

- Heights example is odd – combining men and women. I liked doing cattle weights and prices better.

- Sims used for skewed distributions like a max or min.

- Can use `predict` for some cases.

- Often build a function and apply their `sim` function to it.

- Elections example – what did sim do for us?

- Well switching – again why use sim?

- Earnings – two level model: first logit or probit to model binary response: some versus no income, then using only those with some income: model income based on other predictors.

Chapter 8 Simulation to evaluate inference

- Assume true value is known. Generate data, check inference.
  - Check the size of a test (reject α of the time when H_0 is true)
  - Check coverage of CI

- Can the fitted model produce data like the original data?
  What function of the data do we look at? Fish for a good one.
  - min for Newcombe's
  - proportion 0's for roaches.
  - plots and signs of lagged differences for time series data.

Chapter 9 Causal Inference 1

- If treatment assignment is skewed (control group is healthier or sicker than treated group) then treatment effect is biased, may be way off.

- If we adjust for all lurking pre-treatment variables, we can get an unbiased estimate of treatment effect.

- Fundamental Problem of causal inference:
  We can't observe the same unit under more than one treatment condition. The unobserved counterfactual is not available. Work arounds:
    - Use close substitutes
    - Randomize treatment assignment
    - Adjust for lurking variables.

- Randomized Experiments are unbiased. Treatment effect is $\bar{y}^1 - \bar{y}^0$ with SE $= \sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}$.
  Control group (C) acts as counterfactual to treated group (T) and vice versa. Provides internal validity.
  Typically, units are not a random sample from the population, so to infer causation back to the population (external validity) requires further assumptions or comparison of the units used to the population of interest.
  Electric Company example as if it were a CRD, controlling for pretest scores. $y_i = \alpha + \theta T_i + \beta x_i + \epsilon_i$
  Gain score sets $\beta = 1$ and uses $g_i = \alpha + \theta T_i + \epsilon_i$
  Must assume no interference between units or SUTVA "Stable Unit Treatment Assumption"

- Treatment interactions
    - If a pre-treatment variable (pretest, say) interacts with treatment (Elec. Co.), we need to include that in the model. Makes report a bit more cumbersome, may need to compute the average trt effect given the observed pretest scores.
    - Post-stratification uses population means from another source to estimate average treatment effects (accounting for interactions) in the population.

- Observational Studies – often it is impractical or unethical to apply treatments to units of study.
    - Assume **ignorability**: dist'n of units across treatment conditions is essentially random conditional on covariates.

$$y^0, y^1 \perp T | X$$

      Does not hold if treatment is chosen based on unreported covariates (subjects with a bad heart are not assigned to surgery for ovarian cancer, and heart condition is not reported)
    - Given ignorability holds, can we use regression to estimate treatment effect? Yes if we have overlap. Otherwise we have to extrapolate.

- Causal Inference in Observational Studies
    - "Treatment" should be something we could set on a subject. Think about it carefully.
    - Hard to handle more than one treatment factor at a time.
    - Consider: what is an ideal randomized experiment for a given situation to hone in on the definition of the treatment.

- Do not control for post-treatment variables.

- Intermediate outcomes and Causal Paths
    - IQ example: Some parents are poor whether assigned to T or C groups. Others are good parents regardless of their group assignment. Effectiveness of the treatment depends on

* how much improvement we see in the parents who are naturally poor unless given the treatment, and then become good, AND
* what proportion of the population of parents falls into this group.

Controlling for parenting behavior can even reverse the treatment effect's sign.

– Principal stratification: if we can subdivide the population into groups based on potential outcomes for the mediating variables (parenting), then assess treatment effects in each group, we could avoid the problem above. However, these principal strata are unknown. See instrumental variables in chapter 10.

– Intermediate outcomes are even harder to deal with in observational studies. In particular, it's hard to think of the effect of one variable while holding other predictors constant since predictor values are collected as a batch from a randomly sampled unit.

Chapter 10 Causal Inference 2

- Imbalance, lack of overlap
  Assuming ignorability, imbalance means we have different distributions (in means, spreads, skewness, …) of covariates in the C and T groups.
  Lack of complete overlap means that some regions of the covariate space have only T or only C units.
  Must rely more heavily on modeling (extrapolation).
  Childcare example: T group is only low-birthweight babies. C group has mostly babies of higher birthweight, but some low weights as well.

- Subclassification: estimate effects for subpopulations
  Can subdivide by Mom's education and have complete overlap in each group, so they estimate θ in each group. Combine those $\hat{\theta}$'s with weighted average (weight = size of T group) to get overall treatment effect estimate. Hard to do with more than one subclassification variable.
  Weighting by T group makes this "effect of treatment on the treated" – for low bweight babies, not for the population of all babies.

- Matching
  Find controls who are similar to each case in the pre-treatment covariates. Use Mahalanobis distance to adjust for different variances and for correlations in the pre-treatment variables.
  Propensity score is computed using logistic regression to estimate the probability a unit gets treatment based on the pre-treatment covariates. Then match each treated "case" to one "control" unit which is most similar in propensity score. Check for balance. Estimate treatment effect by comparing the T group to the selected controls.
  Reported SE's are biased by the matching algorithm.
  Some use inverse propensity score as a weight on each point in weighted regression.

- Regression Discontinuity – complete separation of T and C in covariate x
  May identify subsets of T and of C which are closest in x and assume continuity of the model in this region. $y_i = \beta_0 + \beta_1 x_i + \theta T_i + \epsilon_i$. Here the jump is of size θ and the line in x is continuous across a neighborhood of interest near $x = .5$.

- Instrumental Variables (IVs)
  Can't access all lurking variables. An IV randomly induces variation in T and is ignorable. Need to assume monotonicity (positive IV does not induce some to avoid T) and exclusion restriction (effect of IV is on the intermediate group which is influenced

toward choosing T, while unaffected groups are always C or always T despite level of IV).

Example: Sesame Street T=regular watching is influenced by IV = encouraged to watch. Induced-watchers are affected by IV and may become regular watchers. IV does not influence never-watchers or always-watchers.

Intent-to-Treat effect (ITT) is based on mean differences, in the induced watchers group.

- Instrumental Variables in regression

  Two equations: one for y given T another for T given z (=IV) reduce to estimating treatment effect $\beta_1 = \delta_1/\gamma_1$ where δ_1 is the slope of IV on y and γ_1 is the slope of IV on T. Can use 2-stage least squares.

  Can assess one assumption: strong relationship between IV and T, but not the ignorability of IV or exclusion restriction.

- Using variation within/between groups

  Want to use repeated measures to control for lurking variables. Only control for pre-treatment variables.

  What are fixed versus random effects?