BST 232 Methods: Final Review
December 3rd, 2023
Christian Testa

Course Outline:

1. Overview
2. Linear Regression
3. Diagnostics
4. Model Selection
    - Substantive Knowledge
        - Model Building (Specification, Interpretation, Validating Assumptions)
            - Causal Diagrams (Confounding; Colliders; Mediators)
        - Parsimony
        - Hierarchically Well Formulated Models
    - Statistical Criteria
        - R^2 and Adjusted R^2
        - AIC
    - Predictive Performance
        - Train vs. Test Error
        - Data Splitting and Cross Validation
        - the PRESS Statistic
    - Bias Variance Tradeoff
    - LASSO and Ridge
5. Bootstrap/Permutation Testing
    - Bootstrap
        - Resampling the (Nonparametric) Empirical Distribution
        - the Bootstrap Algorithm
        - Robust Regression & Least Median Squares
    - Permutation Methods
        - Constructing the null distribution under permutation
        - Testing / Test Statistics
6. Heteroscedastic Errors
    - Where do homoscedastic assumptions come into Linear Regression
    - Properties of Linear Regression under homoscedasticity
    - BLUE of Linear Regression under Assumptions
    - Theoretical and Simulation Results on Impacts of Heteroscedasticity
    - Remedial Measures
        - Transformations
        - Robust Variance Estimation
        - Generalized Least Squares
            - The estimator and its properties
            - BLUE
            - Estimation using IRLS
7. Proportions & Contingency Tables
    - Inference for Risk Differences
        - Review of Binomial Distribution
        - Wald Test for Binomial Proportions
        - Score Test for Binomial Proportions
        - Exact Inference
        - Testing for Risk Differences
    - Inference for ORs, log ORs
    - Contingency Tables and Pearson χ^2 Tests
    - Delta Method
    - Fisher Exact Tests
8. Likelihood Theory
    - Likelihood & Log Likelihood
    - Fisher Information
    - MLE, Variance in MLE
    - Wald vs. Score vs. LRT
9. GLMs
    - Motivations
    - Exponential Dispersion Families
    - Estimation and Inference (using Likelihoods, Score Equations)
    - Fisher Information
    - Testing and CI
    - IRLS Algorithm (in Practice, in R)
    - Residuals
        - Challenges with Interpretation
10. Binary Outcomes
    - Specifying GLMs, link functions
    - Non-collapsibility of the Odds Ratio
    - Case Control Studies
        - Matched Case Control Studies
11. Count Outcomes
    - Binomial Regression
        - For already 'collapsed' contingency tables
    - Poisson Regression
        - For when there's no fixed n of trials

Questions for Review

1. What are the comparative advantages / disadvantages of the Wald / Score / Likelihood Ratio Tests?
2. If the setting were different, how would I do something analagous to what we did to derive the *retrospective likelihood function* and *conditional logistic regression*?
3. How does GLS compare to other methods we learned (GLMs / Bootstrap / Permutations / Least Median Squares)?
4. Going off the most recent homework, how would I implement risk ratio regression? Would it just be a Poisson model? What about for fixed n of trials, like a binomial setting?
5. What are some other examples of GLMs that might come up on exams?
6. What is the difference between GLS and robust variance estimation again?
7. What is goodness of fit, why is it so commonly discussed, and why have we not talked about it?
8. How exactly was it that a GLM enforces a mean-variance relationship?

Morals of the Class

- How to derive a "best, least unbiased estimator (BLUE)"
- How to assess model performance
- Analytic Intractability $\longrightarrow$ Nonparametric Methods like Bootstrap or Permutation Testing
- Robust Variance (Huber-White/Sandwich) Estimators allow valid inference
- Allowing for ε to depend on $X \longrightarrow$ GLS
  - GLS is BLUE given the assumption $\text{Var}[\varepsilon] = \Sigma$
- Exact tests can be conservative, especially in small samples
- There are multiple perspectives on how to construct statistical tests and confidence intervals (e.g., the Wald / Score / Likelihood Ratio approaches)
- We can account for study designs that deliberately violate the assumptions of Y being random if we redesign the likelihood to be retrospective
- Note the difference between $\mathbb{E}[Y|X] = X\beta$, $\mathbb{E}[Y|X] = g(X)\beta$, $\mathbb{E}[g(Y)|X] = X\beta$ and $g(\mathbb{E}[Y|X]) = X\beta$.
- Most of the standard toolkit relies on statements about asymptotic variance
- The non-collapsibility of the odds ratio is an example where our intuition from linear models doesn't work for GLMs — in this case, precision variables aren't necessarily helpful.

Recommended Problems

Recommended problems by topic from the Agresti textbook (3rd edition):

Likelihood theory:

- 1.6, 1.21, 1.22, 4.33, 5.33

GLMs with binary outcomes:

- 4.22, 4.32, 4.35, 5.6, 5.12, 5.17, 5.26, 5.27, 7.13, 7.20

GLMs with count outcomes:

- 4.7, 4.19, 4.31

Note the availability of some solutions from Agresti here: https://users.stat.ufl.edu/~aa/cda/solutions-part.pdf

For some more applied exercises, see the ones provided in the Poisson and logistic regression chapters of this online book: https://bookdown.org/roback/bookdown-BeyondMLR/

Recommended Reading

On Sandwich Estimators: https://www.stat.berkeley.edu/~census/mlesan.pdf

Heteroskedasticity-Consistent Standard Errors: https://en.wikipedia.org/wiki/Heteroskedasticity-consistent_standard_errors

Review of Material

§6 How did GLS and Robust Variance Estimation differ?

In GLS, we allow for the variance of each ε_i to vary:

$$\text{Var}(\varepsilon_i) = \sigma_i^2 \implies \text{Var}(\varepsilon) = \text{diag}(\sigma_1^2, ..., \sigma_n^2) = \Sigma$$
$$\text{and} \quad \hat{\beta}_{GLS} = (X^\intercal \Sigma^{-1} X)^{-1} X^\intercal \Sigma^{-1} Y.$$

This actually works for any arbitrary positive definite variance matrix Σ.

A *more general* approach is given by weighted least squares (WLS) or Robust Variance Estimation or Huber-White Sandwich estimators or whatever you want to call it:

$$\hat{\beta}_{WLS} = (X^\intercal W X)^{-1} X^T W Y \text{ for an arbitrary } W$$

This yields the following variance estimator:

$$\widehat{\text{Var}}_{HW}(\hat{\beta}_{OLS}) = (X^\intercal X)^{-1} X^\intercal \hat{\Sigma} X (X^\intercal X)^{-1}.$$

Why do we need a more general weighting approach than GLS?

- Correlated outcomes; survey data with given sampling weights; using inverse probability weighting.

Only linear regression is valid and BLUE under homoskedasticity; under heteroscedastic assumptions, GLS is BLUE; and under heteroscedastic scenarios, the Sandwich/Huber-White variance estimators are valid but inefficient.

Also note that while the $\hat{\beta}_{GLS}$ estimates are different from the $\hat{\beta}_{OLS}$ estimates, there isn't any special $\hat{\beta}_{HW}$ — in fact, I've read in Agresti that they say that one is supposed to treat the $\hat{\beta}_{OLS}$ as just wrong but that inference on the variance is valid.

Why is it called the "Sandwich variance estimator"? The estimator "sandwiches" the standard OLS estimate of the variance between two matrices, which adjust for heteroskedasticity and autocorrelation. This is done post-estimation, after obtaining the OLS estimates.

Purpose: GLS modifies the estimation process to handle specific known forms of heteroskedasticity and correlation, whereas the Sandwich estimator provides a robust variance estimate post-estimation, irrespective of the specific form of heteroskedasticity or autocorrelation.

Assumptions: GLS requires more specific assumptions about the error structure and often relies on correct specification, while the Sandwich estimator is more flexible and robust to model misspecifications.

Efficiency: GLS can be more efficient if the error structure is correctly specified, but the Sandwich estimator is more robust in the face of unknown or complex error structures.

§4a GLS vs GLMs/Boot./Permutations/Least Median Squares

One obvious similarity between GLS and GLMs is that both use iteratively reweighted least squares to fit.

What was Least Median Squares? ⇒ Take $\hat{\beta}$ as the values that minimize the median squares:

$$\text{median}(y_i - x_i'\beta)^2.$$

No closed form exists for the variances of these estimators, so we have to resort to bootstrap.

Assumptions:
GLS: Assumes a linear relationship with known error structure
GLMs: Suitable for non-normal response distributions assuming a specific link function
Bootstrap: Few assumptions (mostly that we can approximate the sampling distribution), resamples the observed data.
Permutation testing: Non-parametric, uses permutations to construct an empirical estimate of the null distribution
LMS: Assumes a linear relationship, minimizes influence of outliers

§8 Compare/contrast the Wald/Score/Likelihood Ratio Test?

From lab 07:

Wald:

$$W_n = (\hat{\theta}_{MLE} - \theta_0)^\intercal \mathcal{I}_n(\hat{\theta}_{MLE})(\hat{\theta}_{MLE} - \theta_0) \approx \chi^2(p)$$

- Advantages: Simple to compute, easy to construct confidence interval
- Disadvantages: Requires MLE, approximation not as accurate in small sample

Score:

$$S_n = U(\theta_0|y)^\intercal \mathcal{I}_n(\theta_0) U(\theta_0|y) \approx \chi^2(p),$$
$$\text{and} \quad U(\theta_0|y) = \frac{\partial}{\partial\theta}\ell(\theta|y)|_{\theta=\theta_0}$$

- Advantages: Do not need to compute MLE, more computationally efficient
- Disadvantage: Hard to construct confidence interval since inverting this equation is difficul

LRT:

$$Q_n = -2(\ell(\hat{\theta}_{MLE} - \ell(\theta_0))) \approx \chi^2(p)$$

- Advantages: No derivatives needed, most accurate approximation
- Disadvantage: Requires both MLE and knowledge of log-likelihood under H0.

§8 How does a GLM enforce a mean-variance relation?

Basically it comes in through the specification of the likelihood being from an exponential dispersion family, and how the IRLS algorithm uses the variance.

$$\mathcal{L} = \exp\left\{ \frac{y_i\theta_i - b(\theta)}{a_i(\phi)} + c(y,\phi) \right\},$$

where θ_i is the canonical parameter and ϕ is the dispersion parameter.

Recall that $\mu_i = b'(\theta_i)$, $Var(Y) = b''(\theta_i)a(\phi)$, and $V(\mu_i) = b''(\theta(\mu_i))$ is the variance function, which is used in IRLS to enforce the mean-variance relationship if there is any. In R, one can find this by looking at something like `binomial()$variance` or `Gamma()$variance` or `inverse.gaussian()$variance`, though it looks like the mean-variance relation for quasipoisson and quasibinomial families are handled differently.

To be a little bit more precise, we first looked at solving the following equation using Newton Raphson, then with Fisher Scoring, and then IRLS:

Suppose the current estimate of β is $\hat{\beta}^{(r)}$:

1. $\eta_i^{(r)} = x_i'\hat{\beta}^{(r)}$
2. $\mu_i^{(r)} = g^{-1}(\eta_i^{(r)})$
3. $W_i^{(r)} = \left(\frac{\partial \mu_i}{\partial \eta_i}|_{\eta_i^{(r)}}\right)^2 \frac{1}{V(\mu_i^{(r)})}$
4. $z_i^{(r)} = \eta_i^{(r)} + (y_i - \mu_i^{(r)})\frac{\partial \eta_i}{\partial \mu_i}|_{\mu_i^{(r)}}$ Where W_i is called the "working weight."

The updated value of $\hat{\beta}$ is obtained as the WLS estimate to

$$\hat{\beta}^{(r+1)} = (X'W^{(r)}X)^{-1}(X'W^{(r)}Z^{(r)})$$

where X is the design matrix as usual and W is diagonal with the W_i values and Z is the n-vector $(z_1^{(r)}, ..., z_n^{(r)})$.

§9 How and why do we construct the retrospective likelihood?

Let's start with the usual, prospective likelihood. Because (Y, X) are jointly random and we presume that X does not depend on β, we can write that

$$\mathcal{L}_{\text{joint}} = \prod P(Y_i, X_i) = \prod P(Y_i|X_i)P(X_i)$$

and simplify this to

$$\mathcal{L} = \prod P(Y_i|X_i).$$

Instead, the only thing that's random in a case-control study is the exposure since participants are chosen based on their established/observed outcomes.

$$\mathcal{L}_R = \prod P(X_i|Y_i).$$

We introduce the variable S for selection and condition on $S = 1$ to denote our inference on the observed data:

$$\mathcal{L}_R = \prod P(X_i|Y_i, S = 1),$$

and we apply Bayes' rule to get:

$$\mathcal{L}_R = \prod P(Y_i|X_i, S = 1)\frac{P(X_i|S = 1)}{P(Y_i|S = 1)}.$$

We then apply Bayes' rule again to the quantity $P(Y|X, S = 1)$ and make an assumption about the true form of the prospective likelihood/association in the *target* population to be able to make comparison with it.

Ultimately we find that

$$P(Y|X, S = 1) = \frac{\exp(\log\left(\frac{\pi_1}{\pi_0}\right) + X\beta)}{1 + \frac{\pi_1}{\pi_0}\exp(X\beta)}.$$

The intuition for this is two-fold:

- The only "messed up" part of performing a case-control study is that it messes up the baseline prevalence of the outcome; this study-design doesn't affect the association between exposure and outcome within the sampled population.
- We can see that work out in the math with the β_0 term having this extra $\log(\frac{\pi_1}{\pi_0})$ term added to it.