Summary

Finally, the end of the semester! We made it!

Topic 1: Introduction and Overview

1. Why linear regression, why not t-test?
2. Basic concepts: Population, Sample, parameter, statistic
3. Statistical Activities: Parameter Estimation, Inference

Topic 2: Linear Algebra Review

1. Matrix operation, matrix addition, matrix multiplication ...
2. An *orthogonal matrix* is a **square matrix** with $\mathbf{A}' = \mathbf{A}^{-1}$.
3. Rules of Matrix Operation.
4. Linear Dependence and Rank, matrix determinant
5. Positive Definite and Semi-positive Definite Matrices
6. Inverse and Generalized Inverse
7. Eigenvalues, Eigenvectors. Suppose $\mathbf{A}$ is an symmetric matrix. Then there exists an orthogonal (column orthonormal) matrix $\mathbf{V}$ such that $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$.

8. Random Vectors and Matrices

$$E(\mathbf{AY} + \mathbf{b}) = \mathbf{A}E(\mathbf{Y}) + \mathbf{b} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

$$\text{Cov}(\mathbf{AY} + \mathbf{b}) = \mathbf{A}\text{Cov}(\mathbf{Y})\mathbf{A}' = \mathbf{A\Sigma A}'.$$

9. Important Distributions for Linear Models. If $Z \sim N(0, 1)$, $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$, and X_1 and X_2 are independent. Construct random variables following t-distribution and F distribution.

10. Maximum Likelihood Estimates (MLE)

Topics 3 and 4: Simple Linear Regression and the General Linear Model: Estimation and Testing

1. $\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\ \boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$
2. Least Squares Estimation: $\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.
   $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and Cov$(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
3. HILE Gauss
   - Existence Assumption
   - Linearity Assumption
   - Independence Assumption
   - Homogeneity Assumption
   - Gaussian Errors Assumption
4. $\boldsymbol{\beta}$ is the vector of primary parameters, and $\boldsymbol{\theta}_{a\times 1} = \mathbf{C}_{a\times p}\ \boldsymbol{\beta}_{p\times 1}$ is

a vector of secondary parameters, defined by $\mathbf{C}$, the *contrast matrix*. Each row of $\mathbf{C}$ defines a new scalar parameter in terms of the $\boldsymbol{\beta}$'s, e.g.,$\beta_1 - \beta_2$. The general linear hypothesis is

$$\begin{aligned} H_0 : \boldsymbol{\theta}_{a\times 1} &= \boldsymbol{\theta}_0 \\ H_A : \boldsymbol{\theta}_{a\times 1} &\neq \boldsymbol{\theta}_0. \end{aligned}$$

5. Estimability and Testability of a Parameter. If $\mathbf{X}$ is full rank , then $\widehat{\boldsymbol{\beta}}$ exists (uniquely), $\boldsymbol{\beta}$ is estimable, and any (nonzero) $\boldsymbol{C}$ gives estimable $\boldsymbol{\theta}$. If $\boldsymbol{C}$ is full rank, $\boldsymbol{\beta}$ is testable.

6. Computation of Test Statistic and p-value. Let $\mathbf{M}_{a\times a} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$ and $SSH_{1\times 1} = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\boldsymbol{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. The test-statistic is

$$F_{obs} = \frac{SSH/a}{SSE/(n-p)} = \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\boldsymbol{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/a}{\widehat{\sigma}^2} = \frac{MSH}{MSE}$$

Topic 5: Some Distributional Results for the GLM

- If $\mathbf{X}$ is full rank, $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- $\boldsymbol{\theta} = \mathbf{C}_{a\times p}\boldsymbol{\beta}$, then $\widehat{\boldsymbol{\theta}} \sim N_a(\boldsymbol{\theta}, \sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')$.
- Predicted Values: Conditional Means and Future Observations
  - $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y} = \mathbf{H}\mathbf{y}$,
  - $E(\widehat{\mathbf{y}}) = \mathbf{X}\boldsymbol{\beta}$,
  - $\text{cov}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
- Definitions and Properties of Residuals
- Residual Variance $\widehat{\sigma}^2 = \frac{SSE}{n-p} = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n-p} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-p}$

Topic 6: Multiple Regression: General Consideration

- Basic Sum Squares:
  $USS(\text{total}) = USS(\text{model}) + SSE, \quad \mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{H}\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}.$

$$
\begin{aligned}
CSS(\text{total}) &= CSS(\text{model}) + SSE \\
\mathbf{y}'\left[\mathbf{I} - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right]\mathbf{y} &= \mathbf{y}'\left[\mathbf{H} - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right]\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}.
\end{aligned}
$$

- $F_{obs} = \dfrac{MS(\text{hypothesis})}{MSE} = \dfrac{SSH/dfH}{SSE/dfE}$

$$
= \frac{[SSE(\text{reduced}) - SSE(\text{full})]/[dfE(\text{reduced}) - dfE(\text{full})]}{SSE(\text{full})/dfE(\text{full})}
$$

$$
= \frac{CSS(\text{Regression})/(p-1)}{SSE(full)/(n-p)}.
$$

Reject the hypothesis if $F_{obs} \geq F_F^{-1}(1-\alpha,\, p-1,\, n-p) = f_{crit}$. The usual test of overall regression assumes model spans an intercept and excludes the intercept from the test.

- ANOVA table.
- Usual "Corrected" $\boldsymbol{R^2}$:
  $R_{\mathsf{c}}^2 = \frac{CSS(\text{Regression})}{CSS(\text{Regression})+SSE(\text{full})} = \frac{CSS(\text{Regression})}{CSS(\text{total})}$. R_c^2 estimates ρ_{c}^2, the population ratio of model to total variance, with $0 \leq \rho_{\mathsf{c}}^2 \leq 1$ and $0 \leq R_{\mathsf{c}}^2 \leq 1$.
- The corrected test for overall regression,

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

holds if and only if $H_0 : \rho_{\mathsf{c}}^2 = 0$

Topic 7: Testing Hypotheses in Multiple Regression

- All tests compare two models: the full model and the reduced model (this is the basic idea of likelihood ratio tests, called the *likelihood ratio principle*).
- Overalltest: $F_{obs} = \frac{CSS(\beta_1,\ldots,\beta_{p-1})\big/(p-1)}{SSE(\beta_0,\ldots,\beta_{p-1})\big/(n-p)}$.
- Added-Last Test: the *added-last test* seeks to assess the usefulness of one predictor, above and beyond all others. Coefficient Estimates/t-test table, Type III table. The F statistic is

$$F_{obs} = \frac{\dfrac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{SSE(\text{full})/dfE(\text{full})} = \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big/dfH}{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}\big/dfE},$$

where $\mathbf{C} = \begin{bmatrix} 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \end{bmatrix}_{1\times p}$

- Added-in-Order Test: the *added-in-order test* seeks to assess the contribution of predictor j above and beyond all of the preceding $j-1$ predictors (without the $j+1$, $j+2$, etc.predictors in the model).

- Group Added-Last Tests
- Group Added-in-order Tests

Topic 8: Correlations

$$
\begin{aligned}
\rho &= \mathsf{Corr}(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}} \\
R &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)\left(\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right)}}.
\end{aligned}
$$

Partial correlations describe the strength of the linear relationship between two variables, Y and X, after controlling for the effects of other variables $\mathbf{Z}$.

Topic 9: GLM Assumption Diagnostics

- The First Step: Get to Know Your Data
- Homogeneity: violations seen in the pattern of residuals.
- Independence: assessed through logic of sampling scheme.
- Linearity: examine pattern of residuals.
- Existence: (finite sample...).
- Gaussian distribution: distributional assessment involves box plot of residuals, histogram of residuals, and test of Gaussian distribution of residuals. (The discrepancy between T and Gaussian random variables somewhat inflates the probability of rejecting the null...why?)
- Outliers: leverage, Influence: Cook's Distance

Topic 10: Computation Diagnostics

- Colinearity
- Eigenanalysis
- Condition Number and Condition Index: the *condition index* for the kth eigenvalue equals $\sqrt{\lambda_1/\lambda_k}$. The maximum condition index, called the *condition number*
- $\boldsymbol{R_j^2}$, Tolerance, and VIF
  $R_j^2 = R^2(X_j, \{X_1, \ldots X_{j-1}, X_{j+1}, \ldots X_{p-1}\})$

$$\text{VIF}_j = \frac{1}{1 - R_j^2} = \frac{1}{\text{tolerance}}.$$

- Leverage
- Cook's distance

Topic 11: Selecting the Best Model

1. Specify the maximum model under consideration.

2. Specify a criterion for model selection.

3. Specify a strategy for applying the criterion.

4. Conduct the analysis.

Topic 12: ANOVA

- Coding schemes

$$\text{Es}(\mathbf{X}_{ref}) = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}_{3\times 3} \qquad \text{Es}(\mathbf{X}_{cell}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{3\times 3}$$

$$\text{Es}(\mathbf{X}_{anova}) = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}_{3\times 4}$$

$$\text{Es}(\mathbf{X}_{effect}) = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}_{3\times 3}$$

- Step down test

Topic 13: Coding Schemes for Regression

- (Regression) $\boldsymbol{y} = \begin{bmatrix} \mathbf{1} & \mathbf{x} \\ \mathbf{1} & \mathbf{x} \\ \mathbf{1} & \mathbf{x} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \boldsymbol{\varepsilon}$

- (ANOVA) $\boldsymbol{y} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \boldsymbol{\varepsilon}$

- (Intercept Only) $\boldsymbol{y} = \begin{bmatrix} \mathbf{1} \\ \mathbf{1} \\ \mathbf{1} \end{bmatrix} [\beta_0] + \boldsymbol{\varepsilon}$

- (Null) $\boldsymbol{y} = \boldsymbol{\varepsilon}$

Topic 14: Logistic Regression

- Definition of odds, and odds ratio
- The general logistic regression model is given by

$$\begin{aligned}\text{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1}\end{aligned}$$

  with $y_i \sim \text{Bernoulli}(p_i)$, $i = 1, \ldots, n$, and the y's independent of each other.

- Interpretation of regression coefficients in terms of odds ratio.
- Model comparison by likelihood ratio test
- Logistic regression with categorical covariates and their interactions.
- Goodness of fit test

Topic 15: Mixed Effects Model

- When data are correlated and the independence assumption does not hold, mixed effects models are one way to adjust for the non-independence of observations
- Random effects may be introduced to account for the fact that observations within one subject (or more generally, within one cluster) may be more alike than observations from different clusters
- Forms of covariance matrices for clustered and repeated measurements
- Parameter interpretation of models for longitudinal data