Review for Exam I

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Preliminaries: Appendix A

- (A.3) Random variable: discrete & continuous.
- Mean and variance.
- Covariance.
- Independent random variables; formulae for mean and variance.
- Sums of independent normal random variables. Why important?
- Central limit theorem.
- (A.4) $N(\mu, \sigma)$, t_ν, F_{ν_1,ν_2}, χ^2_ν distributions. Why important?

One & two sample inference: normal data

- (A.6) $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. CIs and $H_0 : \mu = \mu_0$. Extension to paired data.
- (A.7) Two-sample problem with normal data; equal and unequal variances.
- Checking normality: Q-Q plots, formal tests, histograms, boxplots. Outliers.

One & two sample inference: nonparametric

- Sign test for population median. Assumptions?
- Wilcoxin signed rank test for population median. Assumptions?
- Mann-Whitney-Wilcoxin test for two samples. Assumptions?

Simple linear regression: minimal assumptions

- (1.3) $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Assumptions?
- Interpretation of β_0, β_1, and σ.
- Matrix form of the model.
- (1.6) Least squares. Normal equations. Lots of algebra to get b_0 and b_1.
- Introduction to $\hat{Y}_i = b_0 + x_i b_1$ and $e_i = Y_i - \hat{Y}_i$.
- Estimation: OLS leads to BLUEs (b_0, b_1).
- (1.7) MSE$= \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i' \mathbf{b})^2$ estimates σ^2.

Simple linear regression: normal errors

- $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$. Why?
- (1.8) OLS esimators (b_0, b_1) also MLE under normality.
- (2.1) Both b_0 and b_1 are linear combination of independent normals...
- Inference about b_1: CI & testing.
- (2.3) $\mathbf{b} = (b_0, b_1)$ bivariate normal. Leads to inference about
  1. (2.4) $E(Y_h) = \beta_0 + \beta_1 x_h$. Mean of everyone w/ x_h.
  2. (2.5) $Y_h = \beta_0 + \beta_1 x_h + \epsilon_h$. New obs. at x_h.
- Table of regression effects. Toluca data.

Simple linear regression: ANOVA, SS, tests, & correlation

- (2.7) $SSTO = SSR + SSE$, ANOVA table, F-test for $H_0 : \beta_1 = 0$.
- (2.8) General linear test – "big model / little model".
- (2.9) R^2 and $r = corr(x, Y)$.
- (2.11) Bivariate normal distribution, Pearson correlation between x and Y, Spearman correlation.

Matrices and vectors

- (5.2) Matrix addition, (5.3) matrix multiplication, (5.4) symmetric matrix, transpose, (5.6) inverse of a matrix.
- (5.8) Random vectors.
- (5.9) simple linear regression and two-sample problem using matrices.
- (5.10) $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ are least-squares estimators.

Random vectors (5.8)

Let $\mathbf{Y} \in \mathbb{R}^p$ be random with $E\{\mathbf{Y}\} = \boldsymbol{\mu}$ and $\text{cov}\{\mathbf{Y}\} = \boldsymbol{\Sigma}$. Let $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{A} \in \mathbb{R}^{q \times p}$. Then

$$E\{\mathbf{AY} + \mathbf{a}\} = \mathbf{A}\boldsymbol{\mu} + \mathbf{a},$$

and

$$\text{cov}\{\mathbf{AY} + \mathbf{a}\} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'.$$

If $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

$$\mathbf{AY} + \mathbf{a} \sim N_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{a}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}').$$

Recall $\hat{\mathbf{Y}}$ and $\mathbf{e}$ from multiple regression, the fitted values and residuals. For what $\mathbf{A}$ can we write $\hat{\mathbf{Y}} = \mathbf{AY}$? For what $\mathbf{A}$ can we write $\mathbf{e} = \mathbf{AY}$?

Write $SSTO = SSR + SSE$ in terms of matrices.

Multiple regression

- (6.1) $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + x_{ik}\beta_{ik} + \epsilon_i$. Binary predictors.
- Types of models that fit into this framework. Interpretation of individual regression effects.
- Dwayne Portrait Studios, Inc.
- (6.2) Matrix approach $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- (6.3) Estimation: OLS & MLE.
- (6.4) Fitted values $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$ and residuals $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$.
- (6.5) ANOVA table, F-test for $H_0 : \beta_1 = \cdots = \beta_k = 0$, R^2.
- (6.6) Inference about $\mathbf{b}$ and each b_j. Note $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$. Replace σ^2 by MSE to get $se(b_j)^2$.
- (6.7) Estimating $\mathbf{x}_h'\boldsymbol{\beta}$ and $\mathbf{x}_h'\boldsymbol{\beta} + \epsilon_h$.
- Table of regression effects.

Multiple regression: model checking & transformations

- Assumptions to check: (a) linear mean, (b) constant variance, (c) normal errors. Independence discussed in Chapter 12.
- (3.2–3.3) Residual plots: (a) e_i vs. x_j for $j = 1, \ldots, k$, (b) e_i vs. $\hat{Y}_i$, (c) normal probability plot of $e_1, \ldots, e_n$.
- (6.8) Scatterplot matrix (marginal relationships only).
- (3.9 & 6.8) Transformations in $x_1, \ldots, x_k$ and in Y. Box-Cox family for Y.
- (3.6 & 6.8) Breusch-Pagan test for constant variance.

Extra SS, multicollinearity, coef. partial det., VIFs

- (7.1) Extra sums of squares, how much of $SSTO$ gets eaten up by adding x_3, x_4 to a model with x_1, x_2? Answer: $\text{SSR}(x_3, x_4|x_1, x_2)$. Definition. Sequential SS: $\text{SSR}(x_1)$, $\text{SSR}(x_2|x_1)$,$\text{SSR}(x_3|x_1, x_2)$, etc.
- (7.3) General linear test of $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{m}$, SAS `test` statement. Dropping several predictors at once.
- (7.4) $R^2_{Y23|14} = SSR(x_2, x_3|x_1, x_4)/SSE(x_1, x_4)$, etc.
- (7.6) Multicollinearity: VIF_i's, correlation matrix of predictors. Does multicollinearity necessarily indicate a poor model? How does severe multicollinearity ($VIF_j > 10$) affect interpretation of β_j?

Exam I

- Closed book, closed notes.
- Covers Chapters 1 through 7 plus one and two sample methods from first three lectures.
- Anything in the notes is fair game, but I will not ask you to reproduce long formulas, e.g. the formula for a prediction interval.
- Go over homeworks 1–4.
- Need to know what SAS procs do, e.g. `test` command in `proc reg`. Also `npar1way`, `ttest`, `gplot`, etc.
- Mostly short answer.