The Midterm

STA303/STA1002: Methods of Data Analysis II, Summer 2016

Michael Guerzhoy

The Midterm

- 2 hours, timing to be announced on Friday
- Problem(s) similar to the study guide problems
- Problem(s) in the same style as the practice problems
- Problem(s) similar Project 1

What you get with the midterm

- $SST = \sum_{ij}\left(X_{ij} - \bar{X}\right)^2, SSE = \sum_{ij}\left(X_{ij} - \bar{X}_i\right)^2, SSR = \sum_{ij}(\bar{X}_i - \bar{X})^2$
- $Z_1, Z_2, \ldots, Z_k \sim N(0,1)$ iid, $W = Z_1^2 + Z_2^2 + \cdots + Z_k^2 \rightarrow W \sim \chi^2(k)$
- $Z \sim N(0,1), W \sim \chi^2(k)\ indep., T = \frac{Z}{\sqrt{W/k}} \rightarrow T \sim t(k)$
- $AIC = Deviance + 2p$
- $LRT = 2\log(LMAX_{full}) - 2\log(LMAX_{reduced})$
- $Y \sim Bernoulli(\theta) \rightarrow E(Y) = \theta, Var(Y) = \theta(1-\theta)$
- $deviance = const - 2\log \mathrm{P}(\mathrm{y}|\beta)$
- Code of a simulation

Note: SST/SSE/SSR are inconsistently named in various sources. All slides now use SST/SSE/SSR as shown here. You *are* responsible for the interpretation of SSR (Regression sum of Squares) as SSG (Group Sum of Squares)

Sample Practice Problem

- The table below gives the number of observations, mean and standard deviation for the percentage of women for each of the seven Boston judges from the Spock conspiracy trial example. Use these summary statistics to construct the ANOVA (analysis of variance) table

Judge	Number of observations	Mean	Standard deviation
A	5	34.12	11.94
B	6	33.62	6.58
C	9	29.10	4.59
D	2	27.00	3.82
E	6	26.97	9.01
F	9	26.80	5.97
Spock's	9	14.62	5.04

Sample Practice Problem

- Insight: for a particular group Gr

$$MSE_{Gr} = \frac{\sum_{Gr,j}\left(X_{Gr,j} - \bar{X}_{Gr}\right)^2}{N_{gr} - 1} = SD_{Gr}^2$$

$$SSE_{Gr} = \left(N_{gr} - 1\right)SD_{Gr}^2$$

$$SSE = \sum_{Gr} SSE_{Gr}$$

- Insight: The overall mean is the weighted average of the group means
  - $\bar{X} = (\sum_{Gr} N_{Gr}\, \bar{X}_{Gr})/(\sum_{Gr} N_{Gr})$
- $SSR = \sum_{Gr} N_{Gr}(\bar{X}_{Gr} - \bar{X})^2$

Sample Practice Problem

Number of groups is $G = 7$

Total number of observations is $N = 5 + 6 + 9 + 2 + 6 + 9 + 9 = 46$

$\overline{y} = (5 * 34.12 + 6 * 33.62 + \cdots + 9 * 14.62)/46 = 26.583$

Pooled estimate of the error variance (MSE) is $(4 * 11.94^2 + \cdots + 8 * 5.04^2)/(4 + \cdots + 8) = 47.80$

Model Sum of Squares (SSReg) is $5(34.12 - 26.583)^2 + \cdots + 9(14.62 - 26.583)^2 = 1927.856$

This is sufficient to complete the ANOVA table:

Source	df	SS	MS
Model	$7 - 1 = 6$	1927.9	$1927.856/6 = 321.3$
Error	$45 - 6 = 39$	$47.80 * 39 = 1864.1$	47.8
Total	$46 - 1 = 45$	$1927.856 + 1864.1 = 3792.0$	

Study Guide Questions

- Mostly just follow the slides
- Ask for insights similarly to what's done in the practice problems, but without so much context
- Sample:
  - 24. Why is $mean((X_i - \bar{X})^2)$ smaller than (or equal to) $mean\big((X_i - \mu)^2\big)$? Explain intuitively and give a mathematical proof.