STAT 525 FALL 2018

Chapter 4
Miscellaneous Topics

Professor Min Zhang

Simultaneous Inference

- Consider a collection of CIs / hypothesis tests
- Each interval has $(1-\alpha)\%$ confidence level
- What about the overall confidence level?
  - Level of confidence that all constructed intervals contain their true parameter values
  - Often much lower than individual $(1-\alpha)\%$ level
- Will discuss methods that adjust individual confidence levels
- Recall confidence band $\longrightarrow$ widened individual intervals

Joint Estimation of β_0 and β_1

- Will focus here on rectangular region formed by individual CIs
- If estimates were independent
  - Overall confidence level of rectangle is $(1-\alpha)^2$
  - Could set equal to 0.95 and solve for α
- Estimates (b_0,b_1) are not independent so how do we handle this?
- Given normal error terms, can show (b_0, b_1) multivariate normal
- Can show natural (i.e., smallest) confidence region defined by an ellipse (STAT 524)

Bonferroni Correction

- Let A_1 denote event that CI excludes β_0
- Let A_2 denote event that CI excludes β_1
- By construction $\Pr(A_1) = \Pr(A_2) = \alpha$
- What is the probability that both events don't occur?

$$\begin{aligned}
\Pr(\overline{A}_1 \cap \overline{A}_2) &= 1 - \Pr(A_1 \cup A_2) \\
\Pr(A_1 \cup A_2) &= \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2) \\
&\leq \Pr(A_1) + \Pr(A_2) \\
&\downarrow \\
\Pr(\overline{A}_1 \cap \overline{A}_2) &\geq 1 - (\Pr(A_1) + \Pr(A_2))
\end{aligned}$$

- If $\Pr(A_1) + \Pr(A_2) = .05$, then $\Pr(\overline{A}_1 \cap \overline{A}_2) \geq 0.95$

- Want to have family confidence level $1 - \alpha$
- Consider g tests or CIs each using α^*

$$\Pr\left(\bigcap_{i=1}^{g} \overline{A}_i\right) \geq 1 - g\alpha^*$$

- Use level $1 - \alpha/g$ for each test ($\alpha^* = \alpha/g$)
- Provides lower bound for confidence level
- Increasingly conservative as g increases
- True confidence level often much higher than $1 - \alpha$ so larger family-wise α used

Mean Response CIs

- Could apply Bonferroni correction
  - Want to know $E(Y|X)$ for g X's
  - Construct CIs using $\alpha^* = \alpha/g$
  - Reasonable approach when g small

$$\hat{Y}_h \pm B \times s(\hat{Y}_h) \text{ where } B = t(1 - \alpha/(2g), n-2)$$

- Previously discussed Working-Hotelling
  - Uses F distribution instead of t distribution
  - Coefficient W does not change as g increases

$$\hat{Y}_h \pm W \times s(\hat{Y}_h) \text{ where } W^2 = 2F(1 - \alpha, 2, n-2)$$

Prediction Intervals

- Could apply Bonferroni correction
  - Want to know $Y_{h(new)}$ for g X's
  - Construct PIs using $\alpha^* = \alpha/g$
  - Reasonable approach when g small

$$\widehat{Y}_h \pm B \times s(\text{pred}) \text{ where } B = t(1 - \alpha/(2g), n - 2)$$

- Can also use Scheffé procedure
  - Uses F distribution instead of t distribution
  - Coefficient S increases as g increases

$$\widehat{Y}_h \pm S \times s(\text{pred}) \text{ where } S^2 = gF(1 - \alpha, g, n - 2)$$

Regression through the Origin

- Many instances where the line is known to go through the origin
- Statistical model is

$$Y_i = \beta_1 X_i + \varepsilon_i \text{ where } \varepsilon_i \sim N(0, \sigma^2)$$

  - Can show $b_1 = \sum X_i Y_i / \sum X_i^2$
  - Can fit using option `NOINT` for `MODEL` in `PROC REG`
  - Degrees of freedoms for some statistics are changed!
  - Problems with R^2 and other statistics
- If the line does go through the origin, little is lost fitting a line with both intercept and slope
  - **Note:** If there is truly no intercept, no adjustment is necessary for the family of tests

Measurement Error

- Measurement Error in Y
  - Generally not a problem if the measurement error is random with mean zero
  - Error term in model represents unexplained variation which is often a combination of many factors not considered

- Measurement Error in X
  - Does cause problems
  - Often results in biased estimators
  - Tends to reduce strength of association
  - Berkson model: special case where predictor variable is set at a target level, see p. 167-168.

Inverse Predictions

- Given Y_h, predict corresponding X, $\widehat{X}_h$
- Given fitted equation this is

$$\widehat{X}_h = \frac{Y_h - b_0}{b_1} \qquad\qquad b_1 \neq 0$$

- This is the MLE (i.e., function of b_0, b_1)
- Approximate CI can be constructed using inverse mapping of CI for new observation Y_h

$$\widehat{Y}_h - t(1-\alpha/2, n-2)s\{pred\} \leq Y_h \leq \widehat{Y}_h + t(1-\alpha/2, n-2)s\{pred\})$$

$$b_0 + b_1 X_h - t(1-\alpha/2, n-2)s\{pred\} \leq Y_h \leq b_0 + b_1 X_h + t(1-\alpha/2, n-2)s\{pred\})$$

$$Y_h - t(1-\alpha/2, n-2)s\{pred\} \leq b_0 + b_1 X_h \leq Y_h + t(1-\alpha/2, n-2)s\{pred\})$$

$$\Longrightarrow \widehat{X}_h \pm t(1-\alpha/2, n-2)s\{pred\}/b_1$$

Chapter Review

- Simultaneous Inference / Multiplicity
- Regression through the origin
- Measurement Error
- Inverse predictions