ST552 Midterm

Winter 2015

Answer the questions in the spaces provided on this exam.

Name: SOLUTIONS

- You have 50 minutes to complete the exam.
- There are 3 questions. Answer all of the questions.
- Please
  - do not look at the exam until I tell you and
  - stop writing when I announce that the exam is over.

Question	Points	Score
1	15	
2	10	
3	20	
Total:	45	

1. (a) Show that the least squares estimates are unbiased. You should begin by stating the multiple linear regression model in matrix form, along with any assumptions you require. (10)

Model: $Y = X\beta + \varepsilon$ where $E(\varepsilon) = 0$ (1) (1)

$X_{n\times p}$ fixed with rank p (1)

$Y_{n\times 1}$ observed response

$\beta_{p\times 1}$ unknown parameters

(1) some indication of dimensions

Least squares estimates: $\hat{\beta} = (X^TX)^{-1}X^TY$ (1)

$$E(\hat{\beta}) = E\left((X^TX)^{-1}X^TY\right)$$

$$= E\left((X^TX)^{-1}X^T(X\beta + \varepsilon)\right) \quad \text{from model}$$

$$= (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^TE(\varepsilon) \quad \text{linearity of expectation}$$

$$= \beta \quad \text{since } E(\varepsilon) = 0$$

$\Rightarrow$ unbiased

(4) for correct steps

(1) for reasoning on each step

(b) Imagine the errors are correlated. For example, that $\text{Var}(\epsilon) = \Sigma$, where Σ is a symmetric $n \times n$ matrix. Are the least squares estimates still unbiased? Justify your answer. (5)

Yes, since derivation for 1(a) did not rely on $\text{Var}(\varepsilon)$. In fact, only assumption required for the errors was $E(\varepsilon) = 0$

3. An experiment was conducted to explore the relationship between the *lifetime* (measured in days) and sexual activity of fruitflys.

125 fruit flys were divided randomly into 5 treatment groups, each of 25 flys. Each treatment was designed to simulate a different level of sexual activity, with levels: *none*, *one*, *many*, *low* and *high*.

The *thorax length* of each male was also measured as this was known to affect longevity.

One observation in the *many* group was lost.

The following models were fit:

$$\text{Lifetime}_i = \beta_0 + \beta_1 \text{Thorax Length}_i + \beta_2 \text{one}_i + \beta_3 \text{many}_i + \beta_4 \text{low}_i + \beta_5 \text{high}_i + \epsilon_i$$
$$\text{Lifetime}_i = \beta_0 + \beta_1 \text{Thorax Length}_i + \epsilon_i$$

where *one*, *many*, *low*, and *high* are indicator variables for the respective treatment groups. (6)

The two models have **residual sum squares** of 13107 and 22742 respectively.

(a) Conduct an F-test to compare the two models.

$$F = \frac{(RSS_R - RSS_F)/(df_R - df_F)}{RSS_F / df_F}$$

$$= \frac{(22742 - 13107)/4}{13107/118}$$

$$= \frac{2408.75}{111.0763}$$

$= 21.7$ (2) compare to $F^{0.05}_{4,118} = 2.44$ (1)

reject null (1) at 5% level.

prefer more complicated model (1)

lost

$n = 125 - 1 = 124$

$p_{full} = 6$

$p_{reduced} = 2$

$df_{full} = 118$

$df_{reduced} = 122$ (1)

(b) Under what condition would the estimate for β_1 be the same for both models?

(4)

$X = (1 \; TL \; one \; many \; low \; high)$

$= (X_0 \; X_1 \; X_2 \; X_3 \; X_4 \; X_5)$

if X_1 is orthogonal to X_2, X_3, X_4 & X_5

or if thorax length " " other explanatories

3. 2. The following regression model is fit to a subset of Galton's data on the heights of parents and their children:

$$\text{Child's Height}_i = \beta_0 + \beta_1 \text{Father's Height}_i + \beta_2 \text{Mother's Height}_i + \epsilon_i \quad i = 1, \ldots, n$$

where the heights are measured in inches. The subset consists of one male child from each family, for a total of 179 children. Results from the least squares fit are given below.

$$\hat{\beta} = \begin{pmatrix} 20.6 \\ 0.43 \\ 0.29 \end{pmatrix} \begin{matrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{matrix} \qquad \hat{\sigma} = 2.21, \qquad (X^T X)^{-1} = \begin{pmatrix} 7.4 & -0.1 & -0.1 \\ -0.1 & 0.0009 & -0.0001 \\ -0.1 & -0.0001 & 0.0010 \end{pmatrix}$$

(a) Conduct a t-test of the null hypothesis that $\beta_2 = 0$. 6

$t\text{-statistic} = \dfrac{\hat{\beta}_2}{SE\,\hat{\beta}_2}$

$SE\,\hat{\beta}_2 = \hat{\sigma}\sqrt{(X^T X)^{-1}_{33}}$

$= 2.21\sqrt{0.0010}$

$= 0.699$

$= \dfrac{0.29}{0.699}$

$= 4.14$ compare to $t^{(\alpha/2)}_{n-p} = t^{\alpha/2}_{176} = 1.97$

$n = 179$

$p = 3$

$n - p = 176$

Reject null at 5% level

(b) Write a sentence interpreting your result from (a) in context of the study. 4

(i)

There is convincing evidence that the mean child's height is associated with their mother's height even after accounting for their father's height. (1)

(c) ~~(c)~~ Construct a 95% confidence interval for $\beta_1 - \beta_2$.

95% CI : $\hat{\beta}_1 - \hat{\beta}_2 \pm t^{(\alpha/2)}_{176} SE(\hat{\beta}_1 - \hat{\beta}_2)$ (1) (1) 6

either $c^T\beta$ $c = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$ (1) or $Var(\hat{\beta}_1 - \hat{\beta}_2) = Var(\hat{\beta}_1)$ or (1)
$+ Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2)$

$0.43 - 0.29 \pm 1.97(0.091)$ $= \sigma^2(0.0009 + 0.0010 - 2(-0.0001))$
(1) $= \sigma^2(0.0021)$ (1)
$= 0.14 \pm 0.20$
$= (-0.06, 0.34)$ $SE(\hat{\beta}_1 - \hat{\beta}_2) = 2.21\sqrt{0.0021}$
$= 0.101$ (1)

(c) ~~(d)~~ What is the predicted value for a future child's height when the father is 68 inches tall and the mother is 64 inches tall?

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 68 + \hat{\beta}_2 64$ (1) 2
$= 20.6 + 0.43(68) + 0.29(64)$
$= 68.4$ inches (1)

(d) ~~(e)~~ How would you find a standard error for the estimate in (d)? You need only state the calculation you would do, **do not do the calculation.**

$x_0 = \begin{pmatrix} 1 \\ 68 \\ 64 \end{pmatrix}$ (1) $SE(pred) = \hat{\sigma}\sqrt{1 + x_0^T(X^TX)^{-1}x_0}$ 2
(1)