ST552 Midterm

Winter 2019

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.
- There is one page of statistical tables at the end of the exam. You may remove the page of tables if you desire.

Question	Points	Score
1	15	
2	15	
3	10	
Total:	40	

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

$$y = X\beta + \varepsilon$$

where $y_{n\times 1}$,

$X_{n\times p}$ is fixed rank p

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

$\varepsilon_{n\times 1}$ i.i.d. with $E(\varepsilon_i) = 0$ &

$Var(\varepsilon_i) = \sigma^2$

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

$$= E\left((X^TX)^{-1}X^T(X\beta + \varepsilon)\right)$$ from regression eqn.

$$= E\left((X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\varepsilon\right)$$ expand

cancel

$$= \beta + (X^TX)^{-1}X^T E(\varepsilon)$$ linearity of expectation

$E(\varepsilon) = 0$ by assumption

$$= \beta + 0$$

$\Rightarrow$ unbiased

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

The estimates will still be unbiased. We still have $E(\hat{\beta}) = \beta$ from previous part (a) since we only used property that $E(\varepsilon) = 0$, not any properties of $\text{Var}(\varepsilon)$.

Aside: $\text{Var}(\varepsilon) = \Sigma$ will affect the variance of our estimates, and in fact the least squares estimates are no longer the "best", i.e. minimum variance. We'll see this again when we talk about GLS

2. As part of the National Longitudinal Study on Youth, cognitive test scores for 434 3- and 4-year old children and their mothers are obtained.

   Consider the following regression model:

   $$\text{child's score}_i = \beta_1 \text{Mother HS}_i + \beta_2 \text{Mother no HS}_i + \beta_3 \text{Mother's score} + \epsilon_i$$

   where "Mother HS" and "Mother no HS" are the indicator variables:

   $$\text{Mother HS}_i = \begin{cases} 1, & \text{if the mother of child } i \text{ completed highschool} \\ 0, & \text{otherwise} \end{cases}$$

   and

   $$\text{Mother no HS}_i = \begin{cases} 1, & \text{if the mother of child } i \text{ did not complete highschool} \\ 0, & \text{otherwise} \end{cases}$$

   Results from the least squares fit are given below.

   $$\hat{\beta} = \begin{pmatrix} 25.7 \\ 31.7 \\ 0.6 \end{pmatrix} \begin{matrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{matrix} \qquad (X^T X)^{-1} = \begin{pmatrix} 0.10 & 0.10 & -0.001 \\ 0.10 & 0.12 & -0.001 \\ -0.001 & -0.001 & 0.00001 \end{pmatrix}$$

   $$\hat{\sigma} = 18.1,$$

   (a) Find a 95% confidence interval for the parameter β_3. (5)

   $\hat{\beta}_3 \pm t_{n-p}(0.975)\, SE(\hat{\beta}_3)$

   $0.6 \pm 1.97(0.057)$

   $(0.487, 0.713)$

   $SE(\hat{\beta}_3) = 18.1\sqrt{0.00001} = 0.057$

   $t_{431}(0.975) \approx 1.97$

   (b) Interpret your confidence interval from part (a) in the context of the study. (3)

   With 95% confidence, a one point increase in the mother's score is associated with an increase in mean child's score of between 0.49 and 0.71 points, after accounting for whether the mother completed high school.

(c) Consider the linear combination $\beta_2 - \beta_1$. Conduct a t-test (at the 5% significance level) on the null hypothesis that the linear combination is zero. (5)

$H_0: \beta_2 - \beta_1 = 0$

t-stat: $\dfrac{\hat{\beta}_2 - \hat{\beta}_1}{SE(\hat{\beta}_2 - \hat{\beta}_1)}$ compare to $t_{431}(0.975)$

$SE(\hat{\beta}_2 - \hat{\beta}_1) = \hat{\sigma}\sqrt{0.1 + 0.12 - 2(0.10)}$

$= 18.1\sqrt{0.02} = 2.56$

$t = \dfrac{31.7 - 25.7}{2.56} = 2.34 \quad >> t_{431}(0.975) \approx 1.97$

Reject H_0

(d) The test in part (c) may also be considered as an F-test. Write down the forms for the two models that would be compared in such an F-test. (2)

Full model: child's score$_i$ = β_1 Mother HS$_i$ + β_2 Mother noHS$_i$ + β_3 Mothers score$_i$ + ε_i

Reduced model: child's score$_i$ = β_0 + β_3 Mothers score$_i$ + ε_i

↑ Get this by substituting $\beta_1 = \beta_2$ into full model, then Mother HS$_i$ + Mother No HS$_i$ = 1 for all i

n =

3. Researchers collected data on student evaluations of instructors' teaching quality and beauty for 463 courses at the University of Texas.

   The first five rows of the data are given below:

evaluation	gender	beauty
4.30	female	0.20
4.50	male	-0.83
3.70	male	-0.66
4.30	female	-0.77
4.40	female	1.42

   Consider the following two models:

   Model 1: $\text{evaluation}_i = \beta_0 + \beta_1 \text{beauty}_i + \beta_2 \text{female}_i + \beta_3 (\text{beauty} \times \text{female})_i + \epsilon_i$

   Model 2: $\text{evaluation}_i = \beta_0 + \beta_1 \text{beauty}_i + \epsilon_i$

   p = 4

   q = 2

   where:

   - evaluation_i is the average score on a teaching evaluation for the instructor of the ith course,
   - beauty_i is an average beauty rating for the instructor of the ith course made by six students who were not in the course, and
   - female_i is an indicator variable for whether the instructor of the ith course was female.

   The sum of squared residuals are 131.9 and 137.2 for the two models respectively.

   (a) Specify the first five rows of the design matrix, X, corresponding to Model 1. (2)

$$X = \begin{pmatrix} 1 & 0.2 & 1 & 0.2 \\ 1 & -0.83 & 0 & 0 \\ 1 & -0.66 & 0 & 0 \\ 1 & -0.77 & 1 & -0.77 \\ 1 & 1.42 & 1 & 1.42 \end{pmatrix}$$

at 5%, significance level

(b) Conduct F-test to compare the two models. (5)

$$F = \frac{(RSS_{reduced} - RSS_{full})/p-q}{RSS_{full}/n-p} \overset{\text{under null}}{\sim} F_{p-q,\,n-p}$$

$$= \frac{(137.2 - 131.9)/2}{131.9/459}$$

$$= 9.26 \gg$$ compare to $F_{2,459}(0.95) \approx F_{2,500} = 3.01$

Reject H_0

(c) Interpret the result of your F-test from part (b) in the context of the study. (3)

Full model: holding beauty constant, different mean eval for male vs female, and different effect of beauty between male & female

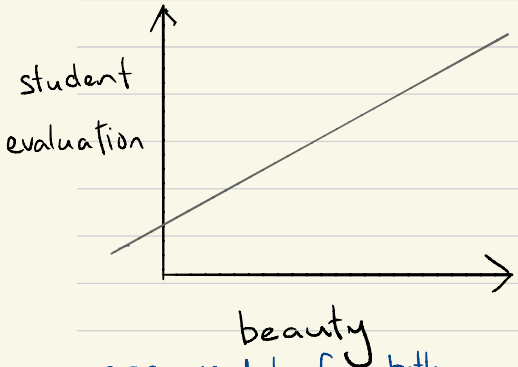
Reduced model: mean eval depends only on beauty, no effect of gender at all.

H_A: atleast

There is convincing evidence, that at least one of gender or the gender beauty interaction is a significant predictor of the mean student evaluation score after accounting for beauty alone.

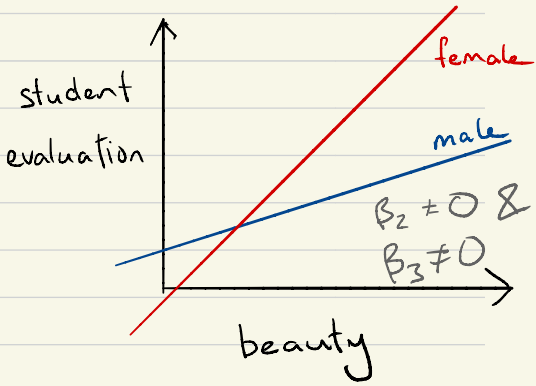
See some alternatives on following pages.

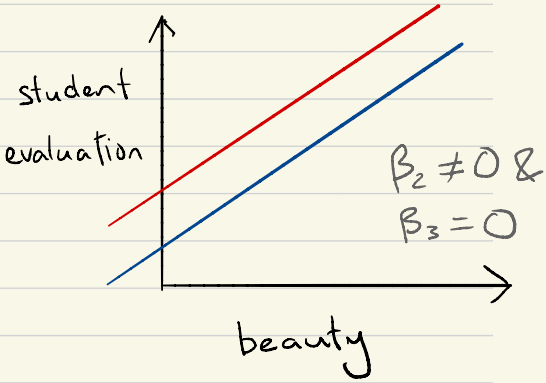
Possibilities under the restricted model

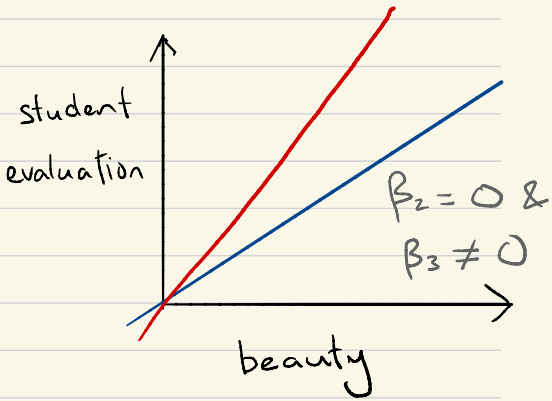


one model for both male and female

Possibilities under the full model







H_A: at least one of $\beta_2 \neq 0$ or $\beta_3 \neq 0$

3(c) Some alternative interpretations:

When considering the relationship between mean student evaluations and the gender and beauty of the instructor, there is convincing evidence the mean student evaluation doesn't depend on beauty alone.

When considering mean student evaluations$^{\text{score}}$ as a straight line function of the instructors beauty$^{\text{score}}$, there is convincing evidence that the line for males is different to that for females.

There is convincing evidence that the mean student evaluation score can not be described by the same straight line function of beauty for male and female instructors.