STA302/1001 Section L0101- Midterm Exam

June 6, 2017, 2:10pm- 3:40pm at EX200

Solution

U of T e-mail: ______________________@mail.utoronto.ca

Surname (Last name):

Given name (First name):

Student ID:

UTORID: (e.g. lihao8)

Instructions:

- You have 90 minutes for 3 questions with multiple parts. Keep these papers closed on your desk until the start of the test is announced.
- Use a benchmark of $\alpha = 5\%$ for all inference, unless otherwise indicated
- You may use a calculator. For numerical answer, please round it off to 4 decimal digits.
- Full mark: 50. Total pages (include the cover): 7.
- Write your answers in the given space only. You cannot use blank space for other questions nor can you write answers on the back. **Your entire answer must fit in the designated space provided immediately after each question.**

Some formulae:

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2} \qquad b_0 = \bar{Y} - b_1\bar{X}$$

$$Var\{b_1\} = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \qquad Var\{b_0\} = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

$$Cov\{b_0, b_1\} = -\frac{\sigma^2 \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \qquad SSTO = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \qquad SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\sigma^2\{\hat{Y}_h\} = Var\{\hat{Y}_h\} = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \qquad r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[\sum_{i=1}^n (X_i - \bar{X})^2\right]\left[\sum_{i=1}^n (Y_i - \bar{Y})^2\right]}}$$

$$\sigma^2\{pred\} = Var\{Y_h - \hat{Y}_h\} = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

Q1 **(12 pts)** Short answer questions. (SLR stands for Simple Linear Regression: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$)

(a) (1 pt) Suppose I have **X=10**. Running **class(X)** on R console, what is the output?

"numeric" 1

(b) (2 pts) There is a generic function "tapply()" in R. You can get help information about it by running a command on R console, what is it?

?tapply
?tapply()
help(tapply)

any one of them 2

(c) (2 pts) Suppose I have two vectors **Y = c(3, 5, 1, 10, 12, 6)**; **X=c(0,1,1,0,0,1)** and I want to print out all the Y values where X takes 1. What R code achieves this?

Y[X==1] 2

(d) (2 pts) True or false and justify your answer: "in lecture we shown that $\sum_{i=1}^{n} e_i = 0$ if a SLR is fitted to a set of n cases by method of Least squares. We then have $\sum_{i=1}^{n} \epsilon_i = 0$ where ϵ_i is the error term in SLR model."

False. 1

ϵ_i's are RVs with mean 0. We can't actually observe them. There is no reason that they should sum to 0. 1

(e) (2 pts) True or false and justify your answer: $Var(b_0+b_1\bar{X}) = \sigma^2/n$ under the Gauss-Markov conditions.

1 True.

1 since $\hat{Y}_i = b_0 + b_1 X_i$ always goes through $(\bar{X}, \bar{Y})$, so we have $\bar{Y} = b_0 + b_1 \bar{X}$

$\Rightarrow Var(b_0 + b_1 \bar{X}) = Var(\bar{Y}) = Var(\frac{1}{n}\sum_{i=1}^n Y_i) = \frac{\sigma^2}{n}$.

(f) (3 pts) Under the normal error model assumption, what is the distribution of b1 and why? (Please specify the distribution and the model parameters).

2 $b_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$, $S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$

1 why: $b_1 = \frac{\sum (X_i - \bar{X}) Y_i}{S_{xx}} = \sum_{i=1}^n \boxed{\frac{X_i - \bar{X}}{S_{xx}}} Y_i = \sum_{i=1}^n k_i Y_i$

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$\varepsilon_i \sim$ iid $N(0, \sigma^2)$

$\Rightarrow Y_i \overset{indep.}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$

$\Rightarrow b_1 = \sum_{i=1}^n k_i Y_i$ is also normal distributed.

$E(b_1) = \beta_1 \Leftarrow b_1$ is BLUE.

$Var(b_1) = \sigma^2 / S_{xx}$

Q2 **(15 pts)** A simple linear regression model is fit on n observed data points. Assume Gauss-Markov conditions hold, coefficients are estimated by least squares method.

(2.a) (3 pts) In class, we show $b_1 = \sum_i k_i Y_i, b_0 = \sum_i w_i Y_i$ where $k_i = \frac{X_i - \bar{X}}{S_{xx}}$, and $w_i = \frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{xx}}$. Now find h_{ij} s.t. $\hat{Y}_i = \sum_{j=1}^n h_{ij} Y_j$ and show that $\sum_{j=1}^n h_{ij} = 1$

$\hat{Y}_i = b_0 + b_1 X_i$, $b_1 = \sum_{j=1}^n k_j Y_j$. $b_0 = \sum_{j=1}^n w_j Y_j$

$$= \sum_{j=1}^n \left(\frac{1}{n} - \frac{(X_j - \bar{X})\bar{X}}{S_{XX}}\right) Y_j + X_i \sum_{j=1}^n \frac{X_j - \bar{X}}{S_{XX}} Y_j$$

(2)

$$= \sum_{j=1}^n \left(\frac{1}{n} - \frac{(X_j - \bar{X})\bar{X}}{S_{XX}} + \frac{(X_j - \bar{X})X_i}{S_{XX}}\right) Y_j$$

$$= \sum_{j=1}^n \left(\frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{S_{XX}}\right) Y_j = \sum_{j=1}^n h_{ij} Y_j$$

where $h_{ij} = \frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{S_{XX}}$

$$\sum_{j=1}^n h_{ij} = \sum_{j=1}^n \left(\frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{S_{XX}}\right)$$

(1)

$$= \sum_{j=1}^n \left(\frac{1}{n}\right) + \frac{(X_i - \bar{X})}{S_{XX}} \underbrace{\sum_{j=1}^n (X_j - \bar{X})}_{=0}$$

$$= 1$$

(2.b) (2 pts) Explain why the result in (a) implies that if h_{ii} is close to 1 then Y_i is close to $\hat{Y}_i$.

2.a implies that

$$\hat{Y}_i = h_{i1} Y_1 + h_{i2} Y_2 + \cdots + h_{ii} Y_i + \cdots + h_{in} Y_n$$

$$\text{and } h_{i1} + h_{i2} + \cdots + h_{ii} + \cdots + h_{in} = 1$$ (1)

(1) so if $h_{ii} \approx 1$ then $\hat{Y}_i \approx Y_i$

it follows that Y_i is close to $\hat{Y}_i$.

(2.c) (5 pts) Show $cov(b_1, \bar{Y}) = 0$.

Soln ①

$$cov(b_1, \bar{Y}) = cov(b_1, b_0 + b_1\bar{X}) \quad ①$$
$$= cov(b_1, b_0) + \bar{X}\, Var(b_1) \quad ①$$
$$= -\frac{\sigma^2 \bar{X}}{S_{xx}} + \bar{X}\frac{\sigma^2}{S_{xx}} \quad ②$$
$$= 0 \quad ①$$

($cov(b_0, b_1)$, $var(b_1)$ given in cover page)

$$b_1 = \sum_{i=1}^n \boxed{\frac{X_i - \bar{X}}{S_{xx}}} Y_i$$
$$= \sum_{i=1}^n k_i Y_i$$

Soln 2

$$cov(b_1, \bar{Y}) = cov\left(\sum_{i=1}^n k_i Y_i, \sum_{j=1}^n \tfrac{1}{n} Y_j\right) \quad ①$$
$$= \sum_{i=1}^n \sum_{j=1}^n \tfrac{1}{n} k_i\, cov(Y_i, Y_j) \quad ①$$
$$= \sum_{i=1}^n \tfrac{1}{n} k_i\, Var(Y_i) + \sum_{i \neq j}\sum \tfrac{1}{n} k_i \underbrace{cov(Y_i, Y_j)}_{=0} \quad ①$$
$$= \frac{\sigma^2}{n}\sum_{i=1}^n k_i = \frac{\sigma^2}{n}\frac{1}{S_{xx}}\underbrace{\sum(X_i - \bar{X})}_{=0} \quad ①$$
$$= 0 \quad ①$$

(2.d) (5 pts) Show that $Cov(e_i, \hat{Y}_j) = 0$. $(i \neq j)$

$$cov(Y_i - \hat{Y}_i, \hat{Y}_j) = cov(Y_i, \hat{Y}_j) - cov(\hat{Y}_i, \hat{Y}_j) \quad ①$$
$$= cov\left(Y_i, \sum_{m=1}^n h_{jm} Y_m\right) - cov(b_0 + b_1 X_i, b_0 + b_1 X_j) \quad ①$$
$$= cov(Y_i, h_{ji} Y_i) - Var(b_0) - X_j\, cov(b_0, b_1) - X_i\, cov(b_1, b_0) - X_i X_j\, Var(b_1)$$

$h_{ij} = h_{ji}$ $\longrightarrow$

$$= h_{ij}\, Var(Y_i) - \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right) + (X_i + X_j)\frac{\sigma^2 \bar{X}}{S_{xx}} - X_i X_j \frac{\sigma^2}{S_{xx}} \quad ②$$
$$= \sigma^2 h_{ij} - \sigma^2\left[\frac{1}{n} + \frac{1}{S_{xx}}\left(\bar{X}^2 - \bar{X}(X_i + X_j) + X_i X_j\right)\right]$$
$$= \sigma^2 h_{ij} - \sigma^2 \underbrace{\left(\frac{1}{n} + \frac{1}{S_{xx}}(X_i - \bar{X})(X_j - \bar{X})\right]}_{= h_{ij}} \quad ①$$
$$= \sigma^2 h_{ij} - \sigma^2 h_{ij}$$
$$= 0$$

Q3 (23 pts) Analysis of Handspan Data
A simple linear regression model is fitted to the data where y = handspan (cm), X = Height (inch), for n = 167 students.

```
> summary(mod)
Call:
lm(formula = HandSpan ~ Height, data = HH)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.00161    1.69394     [A]   0.0782 .
Height       0.35057        [B]     [C]   <2e-16 ***

Residual standard error: [D] on [E] degrees of freedom
Multiple R-squared:  [F],Adjusted R-squared:  0.5442
F-statistic: 199.2 on 1 and [G] DF,  p-value: < 2.2e-16

> anova(mod)
Analysis of Variance Table

Response: HandSpan
           Df Sum Sq Mean Sq F value    Pr(>F)
Height      1  [H]     [I]     199.2   < 2.2e-16 ***
Residuals 165  [J]     1.69
```

3.a) (10 pts) Find the 10 missing values (A through H). Give mark for correct value only.

$A = \frac{-3.0016}{1.6939} = -1.772$ $C = \sqrt{199.2} = 14.1138$

$B = 0.35056/C = 0.0248$ $D = \sqrt{1.69} = 1.3$

$E = n-2 = 165$ $F = H/(H+J) = 0.5470$

$G = 165$ $H = I = 336.648$

$I = 1.69 \times 199.2 = 336.648$ $J = 165 \times 1.69 = 278.85$

3.b) (1 pt) What is the total sum of squares, $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$?

SST = $H + J = 615.498$

3.c) (2 pts) What is the fitted regression line? (define all terms in your answer)

$\xrightarrow{} (1)$

$\widehat{\text{Handspan}} = -3.00161 + 0.35057\ \text{Height}$ (1)

3.d) (3 pts) In the summary output with F value =199.2, what are the null and alternative hypotheses? And what do you conclude (in plain language instead of rejecting or failing to reject H_0)?

$H_0: \beta_1 = 0 \quad H_a: \beta_1 \neq 0.$ 1

The observed F-value is 199.2 with p value < 0.001. 1

1 we reject Ho. We have very strong evidence that there is a positive correlation between handspan and height.

3.e) (2+2 pts) Find a 95% prediction interval of handspan when height is 20 inches. Also find the 95% confidence interval for the mean response at the same given height. Here are some quantiles from t-distributions which may be useful.

$t_{0.95,1} = 1.6542,\ t_{0.95,165} = 1.6541,\ t_{0.95,166} = 1.6541;\ t_{0.95,167} = 1.6540$

$t_{0.975,1} = 1.9745,\ t_{0.975,165} = 1.9744,\ t_{0.975,166} = 1.9744;\ t_{0.975,167} = 1.9743$

$\bar{X}$ = height $= 68.0719$

$X = 20$

$\hat{Y} = -3.00161 + 0.35057 \times 20 = 4.0098$ 1

$s^2(b_1) = 0.0248^2 = 0.0006 = \frac{MSE}{S_{xx}} = \frac{1.69}{S_{xx}}$

$\Rightarrow S_{xx} = \frac{1.69}{0.0006} = 2742.763$ 1

Denote $Y_n = \beta_0 + \beta_1 20 + \varepsilon_n$, 95% PI for Y_n is

$(\hat{Y}_n \pm t_{0.975,165} \cdot s_{pred}) = (4.0098 \pm 1.9744 \cdot \sqrt{1.69(1+\frac{1}{167} + \frac{(20-68.0719)^2}{2742.763})}$

1 $= (4.0098 \pm 1.9744\sqrt{3.1222}) = (4.0098 \pm 1.9744 \times 1.770)$

$= \boxed{(0.5210, 7.4986)} \leftarrow PI$

95% CI for $E(Y)$ at $X=20$: $(\hat{Y} \pm 1.9744 \cdot s(\hat{Y}))$

1 $= (4.0098 \pm 1.9744\sqrt{1.69(\frac{1}{167} + \frac{(20-68.0719)^2}{2742.763})})$

$= (4.0098 \pm 1.9744 \times 1.1968) = \boxed{(1.6449, 6.3747)} \leftarrow CI$.

3.f) (2 pts) If we assume height is a random variable, what is the estimate of the correlation between handspan and height? Are they positively or negatively correlated and why?

$\widehat{corr}(Y, X) = r = \text{sign}(b_1)\sqrt{R^2} = +\sqrt{0.5470} = 0.7396$

1 1

3.g) (1 pts) Answer only True or False for this statement: "For the F-test in ANOVA, it can also be used for testing $H_0 : \beta_1 = 0,\ vs\ H_a : \beta_1 < 0$"

False. 1