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 <u>90 minutes</u> for <u>3 questions</u> with multiple parts. Keep these papers closed on your desk until the start of the test is announced.

- Use a benchmark of <u>$\alpha = 5\%$</u> for all inference, unless otherwise indicated

- You may use a calculator. For numerical answer, please round it off to <u>4 decimal digits</u>.

- 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:

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

$$b_0 = \bar{Y} - b_1\bar{X}$$

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

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

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

$$SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

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

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

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]}}$$

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

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?

$$\begin{cases} ? \, tapply \\ ? \, tapply() \\ help (tapply) \end{cases}$$ 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.

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

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

why: $b_1 = \frac{\sum(X_i - \bar{X})Y_i}{S_{XX}} = \sum_{i=1}^{n}\left[\frac{X_i - \bar{X}}{S_{XX}}\right]Y_i = \sum_{i=1}^{n} k_i Y_i$

① $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ 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.

$E(b_1) = \beta_1$ ← 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 $\sum_{j=1}^{n} h_{ij} = 1$

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

$$= \sum_{j=1}^{n} \left(\frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{S_{xx}} \right) Y_j + X_i \sum_{j=1}^{n} \frac{x_j - \bar{x}}{S_{xx}} Y_j$$

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 $\quad h_{ij} = \frac{1}{n} + \frac{(X_i - \bar{x})(X_j - \bar{x})}{S_{xx}}$

①

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

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

$$= 1$$

(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

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

and $\quad h_{i1} + h_{i2} + \cdots + h_{ii} + \cdots + h_{in} = 1 \quad \Big\} \; ①$

① $\Big\{$ 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 $cov(b_1, \bar{Y}) = 0$.

$\underline{Soln ①}$

$cov(b_1, \bar{Y}) = cov(b_1, b_0 + b_1\bar{x})$ ①

$\qquad = cov(b_1, b_0) + \bar{x} \, Var(b_1)$ ①

$\qquad = -\dfrac{\sigma^2 \bar{x}}{S_{xx}} + \bar{x} \dfrac{\sigma^2}{S_{xx}}$ ②

$\qquad = 0$ ①

($cov(b_0, b_1)$, $Var(b_1)$ given

in cover page)

$b_1 = \sum_{i=1}^{n} \boxed{\dfrac{x_i - \bar{x}}{S_{xx}}} Y_i$

$\quad = \sum_{i=1}^{n} k_i Y_i$

$\underline{Soln \ 2}$

$cov(b_1, \bar{Y}) = cov(\sum_{i=1}^{n} k_i Y_i, \sum_{j=1}^{n} \tfrac{1}{n} Y_j)$ ①

$\quad = \sum_{i=1}^{n} \sum_{j=1}^{n} \tfrac{1}{n} k_i \, cov(Y_i, Y_j)$ ①

$\quad = \sum_{i=1}^{n} \tfrac{1}{n} k_i \, Var(Y_i) + \sum_{i \neq j} \tfrac{1}{n} k_i \underline{cov(Y_i, Y_j)}$ ①

$\quad = \dfrac{\sigma^2}{n} \sum_{i=1}^{n} k_i = \dfrac{\sigma^2}{n} \dfrac{1}{S_{xx}} \underline{\sum(x_i - \bar{x})}$ ①

$\quad = 0$ ①

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

dunno

$cov(Y_i - \hat{Y}_i, \hat{Y}_j) = cov(Y_i, \hat{Y}_j) - cov(\hat{Y}_i, \hat{Y}_j)$ ①

$\quad = cov(Y_i, \sum_{m=1}^{n} h_{jm} Y_m) - cov(b_0 + b_1 x_i, b_0 + b_1 x_j)$ ①

$\quad = cov(Y_i, h_{ji} Y_i) - Var(b_0) - x_j \, cov(b_0, b_1) - x_i \, cov(b_1, b_0)$

$h_{ij} = h_{ji} \longrightarrow = h_{ij} Var(Y_i) - \sigma^2(\tfrac{1}{n} + \tfrac{\bar{x}^2}{S_{xx}}) + (x_i + x_j)\tfrac{\sigma^2 \bar{x}}{S_{xx}} - x_i x_j \, Var(b_1)$ ②

$\quad = \sigma^2 h_{ij} - \sigma^2 [\tfrac{1}{n} + \tfrac{1}{S_{xx}}(\bar{x}^2 - \bar{x}(x_i + x_j) + x_i x_j)]$

$\quad = \sigma^2 h_{ij} - \sigma^2 (\tfrac{1}{n} + \tfrac{1}{S_{xx}} \underline{(x_i - \bar{x})(x_j - \bar{x})})$ ①

$\qquad\qquad\qquad\qquad\qquad\qquad = h_{ij}$

$\quad = \sigma^2 h_{ij} - \sigma^2 h_{ij}$

$\quad = 0$

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
```

I MSReg

H is SSReg   J is RSS

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

$A = \dfrac{-3.0016}{1.6934} = -1.772$ *(estimate - 0)/s.e. = t*

$C = \sqrt{199.2} = 14.1138$ *F = t^2*

$B = 0.35061/C = 0.0248$

$D = \sqrt{1.69} = 1.3$ *residual standard error = sqrt(MSE)*

$E = n-2 = 165$

$F = H/(H+J) = 0.5470$ *R^2 = SSReg / SST = SSReg / (RSS + SSReg)*

$G = 165$

$H = I = 336.648$ *MSReg = SSReg / 1*

$I = 1.69 \times 199.2 = 336.648$ *MSReg = MSE * Fvalue*

$J = 165 \times 1.69 = 278.85$ *SSE = MSE * (n-2)*

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)

$\widehat{HandSpan} = -3.0016 + 0.35057 \; Height$

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)?

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

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

① $\{$ we reject H_0. 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

$S^2(b_1) = 0.0248^2 = 0.0006$

$= \dfrac{MSE}{S_{xx}} = \dfrac{1.69}{S_{xx}}$

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

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

$(\hat{Y}_n \pm t_{0.975,165} \cdot Spred) = (4.0098 \pm 1.9744 \cdot \sqrt{1.69(1 + \frac{1}{167} + \frac{(20-68.07)^2}{2742.763})}$ 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 $X=20$: $(\hat{Y} \pm 1.9744 \cdot S(\hat{Y}))$

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

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

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?

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

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   no 2 sided only