STA302/1001 Section L5101- Midterm Exam

June 5, 2017, 6:10pm- 7:40pm 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.

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

logical. 1

(b) (2 pts) Suppose I have **X=27, Y=6**. Running **X%%Y** on R console, what does it give me?

3 2

(c) (2 pts) Suppose I have a vector **x = c(17, 14, ... 5, 13, 12)** and I want to set all elements of this vector that are greater than 10 to be equal to 7. What R code achieves this?

X[X>10] =7 2

(d) (2 pts) True or false and justify your answer: "for the least squares method to be fully valid, it is required that the distribution of Y be normal".

False 1

For the LS method to be valid, we need only the Gauss-Markov conditions which doesn't require any distribution assumption of Y. 1

(e) (2 pts) True or false and justify your answer: "The sum of residuals weighted by observation is zero, i.e. $\sum Y_i e_i = 0$.".

False. 1

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

$$= \sum_{i=1}^n e_i^2 + \underbrace{\sum_{i=1}^n e_i \hat{Y}_i}_{=0}$$

1

clearly $\sum e_i Y_i = \sum e_i^2$ is not 0 in general.

(f) (3 pts) True or false and justify your answer: Two observations on Y were obtained at each of three X levels, namely, X=5, X=10 and X=15. Then we claim that the least squares regression line fitted to the 6 data points is the same the fitted line to the three points: $(5, \bar{Y}_1), (10, \bar{Y}_2), (15, \bar{Y}_3)$ where $\bar{Y}_1, \bar{Y}_2, \bar{Y}_3$ denote the means of the Y observations at the three X levels.

True. 1

Assume $\bar{Y}_1 = \frac{1}{2}(Y_{11}+Y_{12})$ $\bar{Y}_2 = \frac{1}{2}(Y_{21}+Y_{22})$, $\bar{Y}_3 = \frac{1}{2}(Y_{31}+Y_{32})$

1

clearly overall sample mean $\bar{X} = \frac{1}{6}(5\times 2 + 10\times 2 + 15\times 2)$
$= \frac{1}{3}(5+10+15)$

and $\bar{Y} = \frac{1}{6}(Y_{11}+Y_{12}+Y_{21}+Y_{22}+Y_{31}+Y_{32}) = \frac{1}{3}(\bar{Y}_1+\bar{Y}_2+\bar{Y}_3)$

so it suffice to show b_1 is same for both data sets since $b_0 = \bar{Y} - b_1\bar{X}$.

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

1

For original 6 data pts

$$b_1 = \frac{(-5)(Y_{11}-\bar{Y})^2 + (-5)(Y_{12}-\bar{Y}) + 0 + 0 + 5(Y_{31}-\bar{Y}) + 5(Y_{32}-\bar{Y})}{25+25+0+0+25+25} = \frac{\bar{Y}_3 - \bar{Y}_1}{10}$$

For $(5, \bar{Y}_1), (10, \bar{Y}_2), (15, \bar{Y}_3)$

$$b_1 = \frac{(-5)(\bar{Y}_1 - \bar{Y}) + 5(\bar{Y}_3 - \bar{Y})}{25+0+25} = \frac{\bar{Y}_3 - \bar{Y}_1}{10}$$

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 the lecture we showed $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} e_i \hat{Y}_i = 0$. **Using both results** to prove that $\sum_{i=1}^{n} e_i X_i = 0$.

$$\sum_{i=1}^{n} e_i \hat{Y}_i = \sum_{i=1}^{n} e_i (b_0 + b_1 X_i) \quad ①$$

$$= b_0 \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} e_i X_i \quad ①$$

$$= 0$$

since $\sum_{i=1}^{n} e_i = 0$, so we have $\boxed{\sum_{i=1}^{n} e_i X_i = 0}$ ①

(2.b) (2 pts) Explain why the result in (a) implies that residuals and predictor values are uncorrelated and why this is useful?

This implies that the sample Pearson correlation is 0.

$$r = \frac{\overset{\to 0}{\sum e_i X_i} - \overset{\to 0}{(\sum e_i)}(\sum X_i)}{\sqrt{\sum (e_i - \bar{e})^2 \sum (X_i - \bar{X})^2}} = 0 \quad ①$$

This is useful for residual plots since we then don't expect a pattern in the plot of e_i's versus X_i's. ①

(2.c) (5 pts) Find the w_i s.t. $b_0 = \sum_{i=1}^n w_i Y_i$. Also show $\sum_{i=1}^n w_i = 1$.

$$b_1 = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \sum_{i=1}^n \boxed{\frac{X_i-\bar{X}}{S_{XX}}} Y_i = \sum_{i=1}^n k_i Y_i$$

$$b_0 = \bar{Y} - b_1\bar{X}$$
$$= \sum_{i=1}^n \frac{1}{n} Y_i - \sum_{i=1}^n k_i Y_i$$
$$= \sum_{i=1}^n \boxed{\left(\frac{1}{n} - \bar{X}\frac{X_i-\bar{X}}{S_{XX}}\right)} Y_i = \sum_{i=1}^n w_i Y_i$$ 3

$$\sum_{i=1}^n w_i = \sum_{i=1}^n \left(\frac{1}{n} - \bar{X}\frac{X_i-\bar{X}}{S_{XX}}\right)$$
$$= \sum_{i=1}^n \frac{1}{n} - \bar{X}\frac{1}{S_{XX}} \underbrace{\sum(X_i-\bar{X})}_{=0}$$
$$= 1$$ 2

(2.d) (5 pts) Show that $Cov(b_0, \bar{Y}) = \sigma^2/n$

From 2.C. $b_0 = \sum_{i=1}^n w_i Y_i$ and $\sum w_i = 1$

$i \neq j$
$cov(\epsilon_i, \epsilon_j) = 0$
$\Downarrow$
$cov(Y_i, Y_j) = 0$

$$cov(b_0, \bar{Y}) = cov\left(\sum_{i=1}^n w_i Y_i, \sum_{j=1}^n \frac{1}{n} Y_j\right)$$ 1
$$= \sum_{i=1}^n w_i \sum_{j=1}^n \frac{1}{n} cov(Y_i, Y_j)$$ 1
$$= \underbrace{\sum_{i=1}^n \frac{1}{n} w_i\, cov(Y_i, Y_i)}_{j=i} + \underbrace{\sum_i \sum_{j\neq i} \frac{1}{n} w_i\, cov(Y_i, Y_j)}_{i\neq j}$$ $\shortparallel 0$ 1
$$= \frac{1}{n} \sum_{i=1}^n w_i\, Var(Y_i) + 0$$
$$= \frac{\sigma^2}{n} \sum_{i=1}^n w_i$$ 1
$$= \frac{\sigma^2}{n}$$ 1 since $\sum_{i=1}^n w_i = 1$

or using fact that
$Y_i = \bar{Y} + b_1(X_i - \bar{X}) = b_0 + b_1 X_i$
always passes $(\bar{X}, \bar{Y})$:
$\boxed{\bar{Y} = b_0 + b_1\bar{X}}$ 1

$$cov(b_0, \bar{Y}) = cov(b_0, b_0 + b_1\bar{X})$$
$$= var(b_0) + \bar{X}\, cov(b_0, b_1)$$ 2
$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right) - \sigma^2\frac{\bar{X}^2}{S_{XX}}$$ 1
$$= \frac{\sigma^2}{n}$$ 1

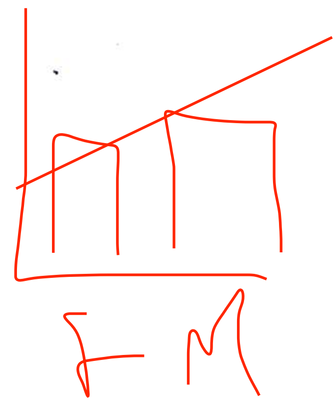
($var(b_0)$, $cov(b_0, b_1)$: cover page)

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

```
> with(HH, tapply(HandSpan,Sex,mean)) # average of HandSpan for F/M
  Female     Male
19.60112 22.30128

> summary(mod)
Call:
lm(formula = HandSpan ~ factor(Sex), data = HH)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       [A]      0.1461     [B]    <2e-16 ***
factor(Sex)Male   [C]      0.2137     [D]    <2e-16 ***

Residual standard error: [E] on [F] degrees of freedom
Multiple R-squared:  0.4917,Adjusted R-squared:  0.4887
F-statistic: 159.6 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)
factor(Sex)   1  [I]     [J]   159.63   < 2.2e-16 ***
Residuals   165  [H]    1.899
```

residual std.err. RSE = sqrt(MSE)

x df_x SSReg MSReg
r df_r RSS MSE



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

$A =$ 19.6011 $B =$ A/0.1461 = 134.1622 ① X 10

$C =$ 22.3013 − 19.6011 = 2.7002 $D =$ C/0.2137 = 12.6355

$E =$ $\sqrt{1.899}$ = 1.3780 $F =$ n−2 = 165

RSS = MSE (n-2)

$G =$ n−2 = 165 $H =$ MSE×df = 1.899×165 = 313.335

SSReg = MSReg

F = MSReg / MSE

$I =$ J = 303.1374 $J =$ 1.899×159.63 = 303.1374

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

SST = I + H = 303.1374 + 313.335 = 616.4724 ①

F = t^2

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

dummy variable regression

$\widehat{Handspan} = 19.6011 + 2.7002 I_M$, where $I_M = \begin{cases} 1, & Male \\ 0, & Female. \end{cases}$

① lose 1 without hat. ①

3.d) (3 pts) In the summary output with F value =159.6, what are the null and alternative hypotheses? And what do you conclude?

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

observed F-value = 159.6, p.value < 0.001, reject H_0. ①

① We have very strong evidence that the means of handspan are different for male and female.

same conclusion

3.e) (2+2 pts) If we assume the true mean of handspan for female and male groups are μ_F and μ_M respectively. Find a 95% confidence interval for $\mu_M - \mu_F$. Here are some quantiles from t-distributions which may be useful. From the confidence you obtained, what conclusion do you have for the test of $H_0 : \mu_M - \mu_F = 0, H_a : \mu_M - \mu_F \neq 0$? Why?

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

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

$\beta_1 = E(Y|M) - E(Y|F) = \mu_M - \mu_F$

95% CI for β_1 is $b_1 \pm t_{0.975,165}\, s.e(b_1)$

② $(2.7002 - 1.9744 \times 0.2147,\ 2.7002 + 1.9744 \times 0.2147)$

$= (2.2763, 3.1241)$

② The 95% CI doesn't include 0, so we have evidence that $\mu_M - \mu_F \neq 0$, the mean of handspan for F and M are different.

3.f) (3 pts) If we calculate a 95% confidence interval for the mean response $E(Y_h)$ and the 95% prediction interval at the same level of sex (same X_h). Compare both intervals, which one is wider? and why?

PI is wider. ① The S.e. of E(Y) only takes into account the variance of the estimation of $\beta_0 + \beta_1 X_h$ while the S.E. of the estimate of Y has this source of variance plus the model error variance. ①