STA302/1001 - Midterm Exam

Oct 20, 2016, 6:10 pm- 7:40 pm

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 4 questions with multiple parts. Keep these papers closed on your desk until the start of the test is announced.
- You may use a calculator. For numerical answer, please round it off to 3 decimal digits.
- 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 **(10 pts)** Answer the following questions.

(1.a) (4 pts) State the simple linear regression model for dependent variables Y and independent variable X and the Gauss-Markov conditions.

Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, 2, \ldots, n.$
Gauss-Markov conditions:

(1) $E(\epsilon_i) = 0;$

(2) $Var(\epsilon_i) = \sigma^2;$

(3) $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$, or the errors are mutually uncorrelated.

(1.b) (3 pts) In order to do inference about the slope (such as testing whether or not the slope is 0), we need to make one more assumption about the model in (1.a). What is the usual assumption and why it is necessary?

One more assumption: the errors have a normal distribution.

We need a distribution assumption for the errors in order to have a sampling distribution for the estimators which is used for critical values and p-values.

(1.c) (3 pts) An estimate is more precise than another if it has smaller variance. The estimate of the $E(Y_h)$ varies with the value of X_h. At what value of X_h will there be the most precise estimate of the E(Y)? Justify your answer.

The estimate of $E(Y_h)$ is $\hat{Y}_h = b_0 + b_1 X_h$.
The variance of $\hat{Y}_h$ at X_h is

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

This is smallest when X_h is at the sample mean, $\bar{X}$, of the X's.

Q2 **(15 pts)** Consider the regression through the origin model $Y_i = \beta X_i + \epsilon_i$ where $i = 1, 2, ..., n$ and where the errors are independent and normally distributed with mean 0 and variance σ^2.

(2.a) (5 points) Show that the least squares estimator for β is

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

The least squares estimator for β is obtained by minimizing

$$Q = \sum_1^n (Y_i - b_1 X_i)^2$$

The first two derivatives gives

$$\frac{\partial Q}{\partial b_1} = -2\sum_1^n X_i(Y_i - b_1 X_i) = -2(\sum_1^n (X_i Y_i) - b_1 \sum_1^n X_i^2) \quad (1)$$

$$\frac{\partial^2 Q}{\partial b_1^2} = 2\sum_1^n X_i^2 \quad (2)$$

Set (1)=0, we have the LS estimator of β, $b_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$. The second derivative is non-negative, Q attains its minimum at b_1.

(2.b) (2 points) Show that b_1 is unbiased estimator.

$$E(b_1) = E\{\frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}\} = E\{\sum_{i=1}^n k_i Y_i\}, k_i = \frac{X_i}{\sum_{i=1}^n X_i^2} \quad (3)$$

$$= \sum_{i=1}^n k_i E(Y_i) = \sum_{i=1}^n k_i \beta X_i \quad (4)$$

$$= \beta \sum_{i=1}^n k_i X_i \quad (5)$$

$$= \beta \sum_{i=1}^n \frac{X_i}{\sum_{i=1}^n X_i^2} X_i \quad (6)$$

$$= \beta \frac{\sum_{i=1}^n X_i^2}{\sum_{i=1}^n X_i^2} \quad (7)$$

$$= \beta \quad (8)$$

(2.c) (4 points) Show that the variance of b_1 is $\sigma^2/(\sum_1^n X_i^2)$.

Note that ϵ_i's are independent, so are Y_i's and $Y_i \sim N(\beta X_i, \sigma^2)$.

$$\begin{aligned}
V(b_1) &= V\{\frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}\} = V\{\sum_{i=1}^n k_i Y_i\}, k_i = \frac{X_i}{\sum_{i=1}^n X_i^2} \tag{9}\\
&= \sum_{i=1}^n k_i^2 V(Y_i) = \sum_{i=1}^n k_i^2 \sigma^2 \tag{10}\\
&= \sigma^2 \sum_{i=1}^n \frac{X_i^2}{(\sum_{i=1}^n X_i^2)^2} \tag{11}\\
&= \sigma^2 \frac{1}{(\sum_{i=1}^n X_i^2)^2} \sum_{i=1}^n X_i^2 \tag{12}\\
&= \frac{\sigma^2}{\sum_{i=1}^n X_i^2} \tag{13}
\end{aligned}$$

(2.d) (4 points) Show $Cov(e_i, \hat{Y}_i) = 0$ where $e_i = Y_i - \hat{Y}_i = Y_i - b_1 X_i$

Note that ϵ_i's are independent, so are Y_i's and $Y_i \sim N(\beta X_i, \sigma^2)$.

$$\begin{aligned}
Cov(e_i, \hat{Y}_i) &= Cov(Y_i - \hat{Y}_i, \hat{Y}_i) \tag{14}\\
&= Cov(Y_i, \hat{Y}_i) - V(\hat{Y}_i) \tag{15}\\
&= Cov(Y_i, b_1 X_i) - V(b_1 X_i) \tag{16}\\
&= X_i Cov(Y_i, b_1) - X_i^2 V(b_1) \tag{17}
\end{aligned}$$

by uncorrelatedness of y_i and y_j

$$\begin{aligned}
&= X_i Cov(Y_i, \sum_{j=1}^n k_j Y_j) - X_i^2 V(b_1), k_j = \frac{X_j}{\sum_j^n X_j^2} \tag{18}\\
&= X_i Cov(Y_i, k_i Y_i) - X_i^2 \frac{\sigma^2}{\sum_i^n X_i^2} \tag{19}\\
&= X_i k_i V(Y_i) - \sigma^2 \frac{X_i^2}{\sum_i^n X_i^2} \tag{20}\\
&= \sigma^2 \frac{X_i^2}{\sum_i^n X_i^2} - \sigma^2 \frac{X_i^2}{\sum_i^n X_i^2} \tag{21}\\
&= 0 \tag{22}
\end{aligned}$$

Q3 **(10 pts)** The Fisher's Iris data set is a multivariate data set introduced by Ronald Fisher in his 1936 paper. A subset of it with two species is analysed. In this question, we will only consider how species can be used to predict the petal width. Some output from R is given below. Note that some numbers have been replaced by letters.

```
> tapply(subIris$Petal.Width, subIris$Species, mean)
    setosa versicolor
     0.246      1.326

> summary(fit)
Call: lm(formula = Petal.Width ~ factor(Species), data = subIris)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 (A)     0.0224     10.98    <2e-16 ***
factor(Species)versicolor   (B)     0.03169      (C)      (P1)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1584 on 98 degrees of freedom
Multiple R-squared:  0.9222,Adjusted R-squared:  0.9214
F-statistic:  1161 on (D) and (E) DF,  p-value: (P2)

> anova(fit)  # Analysis of Variance Table

Response: Petal.Width
                Df  Sum Sq Mean Sq F value    Pr(>F)
factor(Species)  1 29.1600    (F)     (G)      (P3) ***
Residuals       98     (H)    (I)
```

3.a) (9 pts) Find the 9 missing values (A through H) in above R output.

$A = 0.246$ $\qquad$ $B = 1.326 - 0.246 = 1.08$

$C = \sqrt{1161} = 34.073$ $\qquad$ $D = 1$

$E = 98$ $\qquad$ $F = 29.160$

$G = 1161$ $\qquad$ $H = I * 98 = 2.459$

$I = 0.1584^2 = 0.0251$

3.b) (1pt) True or false: The p-values at P1, P2 and P3 are the same.

True

Q4 **(15 pts)** In this question, we analyzed the Old Faithful Geyser. Two variables in this data, one is the waiting time (mins) between eruptions and the duration of the eruption for the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA.

```
> summary(faithful)
    eruption        waiting    
 Min.   :1.600   Min.   :43.0  
 1st Qu.:2.163   1st Qu.:58.0  
 Median :4.000   Median :76.0  
 Mean   :3.488   Mean   :70.9  
 3rd Qu.:4.454   3rd Qu.:82.0  
 Max.   :5.100   Max.   :96.0  

> summary(mod)

Call:
lm(formula = eruption ~ waiting, data = faithful)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
waiting      0.075628   0.002219   34.09   <2e-16 ***
---

Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,Adjusted R-squared:  0.8108 
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

4.a) (3 pts) For the regression, find sample size n, R^2 and explain what R^2 measures.

n= 272 R^2= 0.8115

Almost 81.25% of the variation in the durations of eruptions is explained by its linear relationship with the waiting time between two eruptions.

4.b) (1+2 pts) For t value with 34.09 in the output, what are the null and alternative hypothesis? And what do you conclude?

$$H_0 : \beta_1 = 0 \qquad vs \qquad H_a : \beta_1 \neq 0$$

The two-side test gives p-value is less than 2e-16. We have strong evidence that the slope is significantly different from zero.

Here are some quantiles from t-distributions which may be useful to answer (d) and (4.c, 4.d)

$$t_{0.925,268} \approx t_{0.925,270} \approx t_{0.925,271} \approx t_{0.925,272} = 1.444$$
$$t_{0.95,268} \approx t_{0.95,270} = 1.651;\ t_{0.95,271} \approx t_{0.95,272} = 1.651$$
$$t_{0.975,268} \approx t_{0.975,270} \approx t_{0.975,271} \approx t_{0.975,272} = 1.969$$

4.c) (4 pts) Calculate 95% confidence interval for the slope. How is it related to your answer in (4.b)?

The 95% confidence interval for the slope,

$$b_1 \pm \approx t_{0.975,270} s(b_1) = 0.075628 \pm 1.969 * 0.002219 = (0.071, 0.080)$$

This interval excludes 0 which agrees with part (4.b) that the slope is significantly different from zero.

4.d) (3 pts) Use the fitted model to predict the duration of eruption when waiting time between eruptions is 10 minutes and construct the 95% prediction interval for it.

$$\hat{Y} = -1.874016 + 0.075628 * 10 = -1.117736, \bar{X} = 70.9$$

$$S_{xx} = MSE/S(b_1)^2 = 0.4965^2/(0.002219^2) = 50063.81,$$

The 95% prediction interval is

$$-1.117736 \pm 1.969 * 0.4965 \sqrt{1 + \frac{1}{272} + \frac{(10 - 70.9)^2}{50063.81}} = (-2.133, -0.103)$$

4.e) (2 pts) Do you feel confident that the actual duration of eruption at 10 minutes waiting time is in the interval you calculated in part (4.d)? Why or why not?

No. Because 10 waiting time is outside the range of the data and we can't be sure that the linear model is still appropriate below the minimum waiting time in the data (43 minutes).