University of Toronto

Faculty of Arts and Science

Term Test: STA302H1F/STA1001HF

10:10am–12:00pm

Aids Allowed: One 8.5 $\times$ 11in double-sided formula sheet;

Nonprogrammable calculator

Name: ____

Student ID: ____

Read the following instructions carefully:

1. Do not turn the page until told to do so.
2. You must **show your work** to receive full credit.
3. Probability tables needed are attached at the end of the exam paper.
4. If you don't understand a question, or are having some other difficulty, do not hesitate to ask your instructor or TA for clarification.

Questions	Assigned Mks	Earned Mks
Q1	16	
Q2	6	
Q3	8	
Q4	12	
Q5	8	
Total	50	

Please use $\alpha = 0.05$ for all statistical tests, unless specified otherwise

Please keep at least 3 decimal digits in your numerical calculations.

1. [Total 16 marks] The number of thousand pounds of steam used per month at a plant is thought to be related to the average monthly ambient temperature (in Fahrenheit). The past year's usages and temperatures are given below.

Month	Temperature	Usage	Month	Temperature	Usage
Jan	21	186	Jul	68	622
Feb	24	214	Aug	74	675
Mar	32	288	Sep	62	562
Apr	47	425	Oct	50	453
May	50	455	Nov	41	370
Jun	59	539	Dec	30	274

Also computed are $\bar{x} = 46.5, \bar{y} = 421.9167, SXX = 3309, SYY = 280880.9, SXY = 30484.5$.

(a) [3] Fit a simple linear regression model to the data (i.e., calculate $\hat{\beta}_0$ and $\hat{\beta}_1$).
$\hat{\beta}_1 = SXY/SXX = 30484.5/3309 = 9.2126$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 421.9167 - 9.2126 \times 46.5 = -6.4692$.

(b) [5] Calculate the residual sum of squares (RSS) and the sum of squares due to regression (SSreg). Construct the corresponding ANOVA table and test for significance of regression.
$RSS = SYY - \hat{\beta}_1^2 SXX = 39.3$, $SSreg = SYY - RSS = 280841.6$, the ANOVA table is

Source	df	SS	MS	F
Regression	1	280841.6	280841.6	71460.97
Residual	10	39.3	3.93	
Total	11	280880.9		

Since $F = 71460.97 > F(0.05, 1, 10) = 4.9646$, reject $NH : \beta_1 = 0$, i.e., the regression slope is significant.

(c) [4] Plant management believes that an increase in average ambient temperature of 1 degree will increase average monthly steam consumption by less than 10 thousand pounds. Do the data support this statement?
Test $NH : \beta_1 = 10$ vs $AH : \beta_1 < 10$, use t-statistic $t = \frac{\hat{\beta}_1 - 10}{se(\hat{\beta}_1)}$.
First calculate $se(\hat{\beta}_1) = \hat{\sigma}/\sqrt{SXX} = \sqrt{3.93/3309} = 0.03446$, then

$$t = (9.2126 - 10)/0.03446 = -22.8497 < -t(0.05, 10) = -1.8125.$$

Reject NH, i.e, the data support this statement (AH).

(d) [4] Construct a 99% prediction interval on steam usage in a month with ambient temperature of 58 degrees.

First calculate $\tilde{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_* = 527.8616$, the prediction interval with $x_* = 58$ is

$$\begin{aligned}\tilde{y}_* \pm t(0.005, 10)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}} &= 527.8616 \pm 3.169\sqrt{3.93\Big(1 + \frac{1}{12} + \frac{(58-46.5)^2}{3309}\Big)} \\ &= 527.8616 \pm 6.6584 = (521.2032, 534.52).\end{aligned}$$

2. [Total 6 marks] Consider the multiple linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad E(\mathbf{e}) = \mathbf{0}, \quad \mathrm{Var}(\mathbf{e}) = \sigma^2\mathbf{I}.$$

(a) [4] Show that the least-squares estimator $\hat{\boldsymbol{\beta}}$ can be written as $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{R}\mathbf{e}$ where $\mathbf{R} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} = \boldsymbol{\beta} + \mathbf{R}\mathbf{e}$

(b) [2] Using this, or otherwise, show that $\hat{\boldsymbol{\beta}}$ is unbiased.

$E(\hat{\boldsymbol{\beta}}|\mathbf{X}) = E(\boldsymbol{\beta} + \mathbf{R}\mathbf{e}|\mathbf{X}) = \boldsymbol{\beta} + \mathbf{R}E(\mathbf{e}) = \boldsymbol{\beta}$

3. [Total 12 marks] Given data $(x_1, y_1), \ldots, (x_n, y_n)$, consider the simple linear regression model

$$E(Y|X = x) = 3.2 + \beta_1 x, \quad \mathrm{Var}(Y|X = x) = \sigma^2.$$

(a) [5] Find the least-squares estimator for β_1. Denote the answer as $\hat{\beta}_1$.

Take derivative of $RSS(\beta_1) = \sum_i (y_i - 3.2 - \beta_1 x_i)^2$ w.r.t. β_1, we have

$$\sum_i x_i(y_i - 3.2 - \beta_1 x_i) = 0, \text{ i.e., } \quad \beta_1 \sum_i x_i^2 = \sum_i x_i(y_i - 3.2), \quad \hat{\beta}_1 = \frac{\sum_i x_i(y_i - 3.2)}{\sum_i x_i^2}.$$

(b) [3] Calculate $\mathrm{Var}(\hat{\beta}_1)$.

Because y_i's are independent given x_i's,

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sum_i x_i^2 Var(y_i - 3.2)}{(\sum_i x_i^2)^2} = \frac{\sigma^2}{\sum_i x_i^2}.$$

(c) [4] Construct a 95% confidence interval for β_1.

First estimate σ, $RSS = \sum_i (y_i - 3.2 - \hat{\beta}_1 x_i)^2$ and its df is $(n-1)$, thus $\hat{\sigma} = \sqrt{\frac{RSS}{n-1}}$.
Then the 95% confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t(0.025, n-1)\hat{\sigma}/(\sum_i x_i^2).$$

4. [Total 8 marks] A simple linear regression model $y = \beta_0 + \beta_1 x$ is fitted to the following data:

x	1.0	1.0	2.0	3.3	3.3	4.0	4.0	4.0	4.7
y	10.84	9.30	16.35	22.88	24.35	24.56	25.86	29.16	24.59
x	5.0	5.6	5.6	5.6	6.0	6.0	6.5	6.9	
y	22.25	25.90	27.20	25.61	25.45	26.56	21.03	21.46	

The RSS of the fitted model is 250.134. Perform a lack of fit test for this simple linear model.

First calculate the means of replicates at different x_i values,
(10.84+9.3)/2=10.07, (22.88+24.35)/2=23.615, (24.56+25.86+29.16)/3= 26.527,
(25.9+27.2+25.61)/3=26.237, (25.45+26.56)/2=26.005.
Calculate $SS_{pe} = (10.84 - 10.07)^2 + (9.3 - 10.07)^2 + (22.88 - 23.615)^2 + (24.35 - 23.615)^2 + (24.56 - 26.527)^2 + (25.86 - 26.527)^2 + (29.16 - 26.527)^2 + (25.9 - 26.237)^2 + (27.2 - 26.237)^2 + (25.61 - 26.237)^2 + (25.45 - 26.005)^2 + (26.56 - 26.005)^2 = 15.563$
The df of SS_{pe} is $(2-1)+(2-1)+(3-1)+(3-1)+(2-1) = 7$
Also $n = 17$, the df of $RSS = 250.134$ is $17 - 2 = 15$.
The SS due to LOF $SS_{lof} = RSS - SS_{pe} = 234.571$ with $df = 15 - 7 = 8$.
The test statistic is

$$F = \frac{SS_{lof}/df_{lof}}{SS_{pe}/df_{pe}} = \frac{234.571/8}{15.563/7} = 13.1883 > F(0.05, 8, 7) = 3.7257,$$

reject NH, i.e., there is lack of fit by a simple linear model.

5. [Total 8 marks] Suppose the true regression function for a particular data set $(x_{11}, x_{12}, y_1), (x_{21}, x_{22}, y_2), \ldots, (x_{n1}, x_{n2}, y_n)$ is

$$E(Y|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \tag{1}$$

However, the experimenter thinks that the second predictor x_2 does not help explaining Y. Therefore he ignores x_2, fits a straight line regression to $(x_{11}, y_1), (x_{21}, y_2), \ldots, (x_{n1}, y_n)$ and obtains the fitted model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1. \tag{2}$$

Calculate the expectation of the $\hat{\beta}_1$ in (2) when the true model is in fact (1). Is it an unbiased estimator for the β_1 in (1)? Explain.

The OLS estimate is $\hat{\beta}_1 = \frac{SX_1Y}{SX_1X_1} = \frac{\sum_i (x_{i1} - \bar{x}_1) y_i}{SX_1X_1} = \sum_i c_i y_i$, where $c_i = \frac{x_{i1} - \bar{x}_1}{SX_1X_1}$.
The model in fact is $E(Y|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, therefore, conditional on x_{i1}, x_{i2}'s,

$$\begin{aligned} E(\hat{\beta}_1) &= E(\sum_i c_i y_i) = \sum_i c_i E(y_i) = \sum_i c_i (\beta_0 + \beta_1 x_1 + \beta_2 x_2) \\ &= \beta_0 \sum_i \frac{x_{i1} - \bar{x}_1}{SX_1X_1} + \beta_1 \sum_i \frac{(x_{i1} - \bar{x}_1) x_{i1}}{SX_1X_1} + \beta_2 \sum_i \frac{(x_{i1} - \bar{x}_1) x_{i2}}{SX_1X_1}. \end{aligned}$$

Since $\sum_i (x_{i1} - \bar{x}_1) = 0$, $\sum_i (x_{i1} - \bar{x}_1) x_{i1} = \sum_i (x_{i1} - \bar{x}_1)^2 = SX_1X_1$, $\sum_i (x_{i1} - \bar{x}_1) x_{i2} = \sum_i (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = SX_1X_2$, the above becomes

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{SX_1X_2}{SX_1X_1},$$

which is only unbiased when $SX_1X_2 = 0$, i.e., x_{i1}'s and x_{i2}'s are uncorrelated. Otherwise it is biased.

—— CONGRATULATIONS! YOU HAVE SURVIVED HALF WAY OF STA302/1001 ——