He93

Seat

Please print in pen: STC 1012

Waterloo Student ID Number:

WatIAM/Quest Login Userid:

UNIVERSITY OF WATERLOO

Examination Midterm Spring 2019 STAT 331

Times: Wednesday 2019-06-19 at 14:30 to 15:50 (2:30 to 3:50PM)
Duration: 1 hour 20 minutes (80 minutes)
Exam ID: 4124522
Sections: STAT 331 LEC 001
Instructors: Nathaniel Stevens

SOLUTIONS

Special Materials

Candidates may bring only the listed aids.

- Calculator - Pink Tie
- Study Notes - Double-Sided 8.5x11

Instructions:

- This test consists of 12 pages including this cover page.
- Page 10 contains additional space for rough work. If you use this page for work that you would like to have marked, you must clearly indicate this.
- Page 11 contains tables of quantiles from the $t_{(8)}$, $t_{(9)}$, $t_{(24)}$, $t_{(25)}$, $t_{(45)}$ and $t_{(52)}$ distributions.
- Page 12 is left blank. For your convenience you may remove pages 11/12.
- All numeric answers should be rounded to four decimal places (unless the answer is exact to fewer than four decimal places).
- Incorrect answers may receive partial credit if your work is shown. An incorrect answer with no work shown will receive 0 points.

Question	Points
Q1	16 /16
Q2	9 /9
Q3	16 /16
Q4	10 /10
Total	51 /51

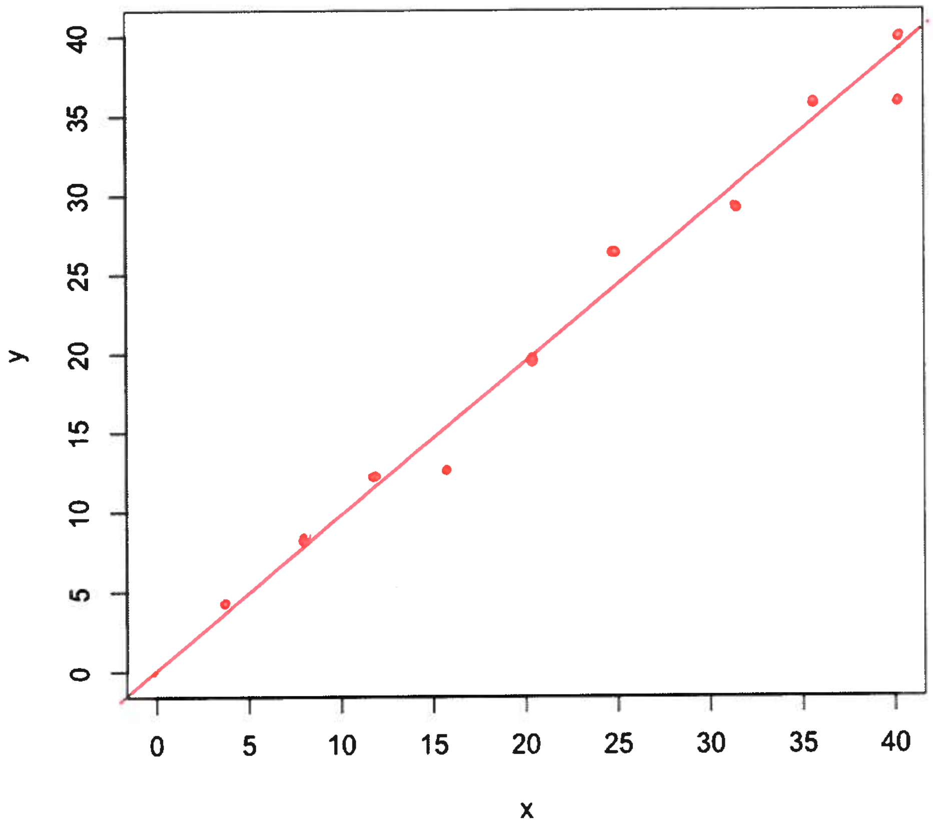
Signature:

- Please identify yourself by signing here: ______________________

Question 1 [16 points]
Two weight scales are being compared, a new weight scale (y) and an old weight scale (x). To determine whether the measurements by the two scales agree I randomly select $n = 10$ objects and weight each of them with both scales. The data available in the following table:

x	4	8	13	16	20	25	31	36	40	40
y	4	8	12	13	20	26	29	36	36	39

(a) [2] Plot this data on the axes provided.



(b) [2] The correlation coefficient for these data is $\hat{\rho} = 0.9933$. Based on this value, the direction of the linear relationship between y and x may be described as (circle one):

[✓ POSTIVE] NEGATIVE

And the strength of the linear relationship between y and x may be described as (circle one):

[✓ STRONG] WEAK

(c) [2] A simple linear regression was fit between y and x. Partial R output of this model is shown below.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02798    1.03980    ???      ???
x            0.95828    0.03934    ???      ???
```

State the equation of the line of best fit and, as accurately as you can, draw it on the scatter plot from part (a).

$\hat{\mu} = -0.02798 + 0.95828x$

(d) If the weight scales agree, then we would expect $\beta_0 \approx 0$ and $\beta_1 \approx 1$.

i. [3] Using the output from part (c), test the following hypothesis at a 5% level of significance. (N.B. You must calculate the test statistic and draw your conclusion by referring to the relevant quantiles found on page 11).

$$H_0: \beta_0 = 0 \text{ vs. } H_A: \beta_0 \neq 0$$

$$t = \frac{\hat{\beta}_0 - 0}{SE[\hat{\beta}_0]} = \frac{-0.02798}{1.03980} = -0.0269$$

The rejection region for this test is $R = \{t \mid t \geq 2.306 \text{ or } t \leq -2.306\}$.

Since $t \notin R$ we do not reject H_0 at a 5% significance level.

ii. [3] Using the output from part (c), test the following hypothesis at a 5% level of significance. (N.B. You must calculate the test statistic and draw your conclusion by referring to the relevant quantiles found on page 11).

$$H_0: \beta_1 = 1 \text{ vs. } H_A: \beta_1 \neq 1$$

$$t = \frac{\hat{\beta}_1 - 1}{SE[\hat{\beta}_1]} = \frac{0.95828 - 1}{0.03934} = -1.0605$$

The rejection region for this test is the same as above: $R = \{t \mid t \geq 2.306 \text{ or } t \leq -2.306\}$.

Because $t \notin R$ we do not reject H_0 at a 5% level of significance

iii. [1] Based on your findings in parts i. and ii., draw a conclusion about the agreement of the two weight scales.

Because we fail to reject H_0 in the preceding two tests we have no evidence against $\beta_0 = 0$ and $\beta_1 = 1$ and so we conclude that the weight scales agree

(e) In the context of weighing objects, when nothing is being weighed, both scales should read 0. In other words, when $x = 0$ then $y = 0$. This phenomenon suggests that β_0 should be 0 and a more appropriate model would be one that does not include an intercept term, such as the following:

$$y_i = \beta x_i + \varepsilon_i$$

for $i = 1,2,\dots,n$. As usual, we assume the errors ε_i are independent and identically distributed $\mathrm{N}(0,\sigma^2)$ random variables.

i. [2] Derive an expression for $\hat{\beta}$, the least squares estimate of β.

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

$$\frac{dS(\beta)}{d\beta} = \sum_{i=1}^{n} (2)(y_i - \beta x_i)(-x_i)$$

$$= -2 \sum_{i=1}^{n} (x_i)(y_i - \beta x_i)$$

$$\frac{dS(\beta)}{d\beta} = 0 \Rightarrow \sum_{i=1}^{n} (x_i)(y_i - \beta x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \beta x_i^2$$

$$\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

ii. [1] Provide an expression for $\hat{\sigma}$, the least squares estimate of σ.

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{\beta} x_i)^2}{n-1}}$$

Question 2 [9 points]
Suppose the following linear regression model is used to relate a response variable y to four explanatory variables x_1, x_2, x_3, x_4:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

for $i = 1,2,\ldots,30$. Relevant data summaries are shown below.

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 177.23 \\ 2.17 \\ 3.54 \\ -22.16 \\ 0.20 \end{bmatrix} \qquad \hat{\sigma}^2(X^TX)^{-1} = \begin{bmatrix} 77.22 & -2.59 & -0.52 & -3.27 & -0.65 \\ -2.59 & 0.45 & -0.01 & 0.02 & 0.03 \\ -0.52 & -0.01 & 0.01 & 0.01 & -0.01 \\ -3.27 & 0.02 & 0.01 & 0.30 & 0.01 \\ -0.65 & 0.03 & -0.01 & 0.01 & 0.10 \end{bmatrix}$$

(a) [1] Interpret the value 177.23.

The expected response when $x_1 = x_2 = x_3 = x_4 = 0$

(b) [1] If x_2 is increased by 1 unit and x_1, x_3 and x_4 are held fixed, how much do we expect y to change? Be sure to state whether this change is an *increase* or a *decrease* in y.

We expect y to increase by 3.54 units

(c) [1] If x_3 is increased by 2 units and x_1, x_2 and x_4 are held fixed, how much do we expect y to change? Be sure to state whether this change is an *increase* or a *decrease* in y.

We expect y to decrease by $2 \times (22.16) = 44.32$ units

(d) [1] What does the value 0.3 represent?

$\text{Var}[\hat\beta_3]$

(e) [1] State the value of $\hat{\beta}_4$.

0.2

(f) [1] State the value of $\text{Cov}[\hat{\beta}_1, \hat{\beta}_3]$.

0.02

(g) [1] What is $\text{Corr}[\hat{\beta}_1, \hat{\beta}_3]$?

$$\text{Corr}[\hat\beta_1, \hat\beta_3] = \frac{\text{Cov}[\hat\beta_1, \hat\beta_3]}{SD[\hat\beta_1]\, SD[\hat\beta_3]} = \frac{0.02}{\sqrt{0.45 \times 0.3}} = 0.0544$$

(h) [2] Calculate a 95% confidence interval for β_1.

$\hat\beta_1 \pm t_{(25)}(0.975) \times SE[\hat\beta_1]$

$= 2.17 \pm 2.0595 \times \sqrt{0.45}$

$= (0.7884, 3.5516)$

Question 3 [16 points]

Recently in academia, universities have been trying to eliminate systemic pay inequity between males and females. The Biology Department at a Canadian university collected the following information on each of their $n = 52$ professors to determine whether there was a significant difference between male and female salaries.

- `sex`: takes on values `1` or `0` indicating whether the professor is a female or male, respectively
- `rank`: takes on values `assistant`, `associate` or `full` depending on the rank of the professor
- `degree`: takes on values `1` or `0` indicating whether the professor's highest degree is a Master's or a PhD degree, respectively
- `yrs_in_rank`: the number of years the professor has worked in their current rank
- `yrs_since_degree`: the number of years since the professor earned their highest degree
- `salary`: the professor's annual salary (in $1000s)

The linear regression model that relates `salary` (y) to the other variables may be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon$$

where x_1 is the `sex` indicator, x_2 and x_3 are `rank` indicators (`assistant` is treated as the baseline), x_4 is the `degree` indicator and x_5 and x_6 are the `yrs_in_rank` and `yrs_since_degree` variables, respectively. R output for this model is shown below. You may refer to this in the questions that follow.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        69.2071     5.2677  13.138  < 2e-16 ***
sex                -4.4108     3.5001  -1.260    0.214
rankassociate      20.0136     4.3314   4.621 3.22e-05 ***
rankfull           42.0467     5.1119   8.225 1.62e-10 ***
degree             -5.2512     3.8525  -1.363    0.180
yrs_since_degree   -0.4711     0.2930  -1.608    0.115
yrs_in_rank         1.8012     0.3589   5.018 8.65e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.07 on 45 degrees of freedom
Multiple R-squared:  0.855,   Adjusted R-squared:  0.8357
F-statistic: 44.24 on 6 and 45 DF,  p-value: < 2.2e-16
```

(a) [2] Consider two professors with the same rank, same years in rank, same degree, same number of years since their degree but one is a female and the other is a male. On average, does the female get paid more or less than the male? Circle one.

MORE [✓ LESS]

If the university was going to adjust the salaries for females so that on average, they were paid the same as their male colleagues, what would that adjustment be?

Females should receive an increase of $4410.80

(b) [2] Construct a 95% confidence interval for the expected difference in female vs. male salaries.

$\hat{\beta}_1 \pm t_{(45)}(0.975) \times SE[\hat{\beta}_1]$

$= -4.4108 \pm (2.0141)(3.5001)$

$= (-11.4604, 2.6388)$

(c) One might expect that professors with PhDs get paid more, on average, than professors with only Master's degrees.

i. [1] How much larger (on average) is the salary of a professor with a PhD relative to the salary of a professor with a Master's degree (all else equal)?

\$ 5251.20

ii. [1] State the hypothesis that would be tested to determine whether professors with PhDs and professors with Master's degrees have significantly different salaries.

$H_0: \beta_4 = 0$ vs. $H_A: \beta_4 \neq 0$

iii. [1] State the p-value associated with the hypothesis in i.

0.180

iv. [1] At a 5% level of significance, do you conclude that the salaries are significantly different? Circle one.

YES [✓ NO]

(d) [1] Predict the salary of a male assistant professor that earned his PhD 4 years ago and who has been an assistant professor for 4 years.

$69.2071 - 4.4108(0) + 20.0136(0) + 42.0467(0) - 5.2512(0) + 1.8012(4)$
$- 0.4711(4)$

$= 74.5275$

$\therefore$ \$ 74,527.50

(e) [1] Consider the prediction interval that might have accompanied the *predicted* salary in part (d), as well as the associated confidence interval for the *expected* salary of a male assistant professor that earned his PhD 4 years ago and who has been an assistant professor for 4 years. True or false, the prediction interval would be narrower than the confidence interval. Circle one.

TRUE [✓ FALSE]

(f) [1] The coefficient of determination for this model is $R^2 = 0.855$. Interpret this value.

85.5% of the response variation is explained by the model

(g) [3] Using the output above and the fact that the variance of observed salaries is $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 500.72$, complete the following ANOVA table.

Source	SS	df	MS	F
Regression	21833.8956	6	3638.9826	44.2241
Error	3702.8244	45	82.2850	
Total	25536.72	51		

(h) [2] In the ANOVA table above, circle the least squares estimate of σ^2, $\hat{\sigma}^2$. In the R output on page 6, circle the least squares estimate of σ, $\hat{\sigma}$.

Question 4 [10 points]
In the context of the linear regression model

$$y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2 I)$ we've seen that the least squares estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$ and the residual vector is defined as $\boldsymbol{e} = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}} = (I - H)\boldsymbol{y}$, where I is the $n \times n$ identity matrix and $H = X(X^T X)^{-1} X^T$ is the "hat" matrix. In the questions that follow, you will consider the $2n \times 1$ random vector $\boldsymbol{v}$ which is obtained by stacking the vectors $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$ on top of each other:

$$\boldsymbol{v} \equiv \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ -- \\ \boldsymbol{e} \end{bmatrix} \equiv A\boldsymbol{y} \qquad \text{where } A = \begin{bmatrix} (X^T X)^{-1} X^T \\ ---- \\ I - H \end{bmatrix}$$

(a) [1] State the distribution of the random vector $\boldsymbol{y}$. (N.B. You must name the distribution and state both the expected value vector and the variance-covariance matrix.)

$$\vec{y} \sim MVN(\vec{\mu} = X\vec{\beta},\ \sigma^2 I)$$

(b) [1] True or false, the random vector $\boldsymbol{v}$ follows a multivariate normal distribution. Circle one.

[✓ TRUE] FALSE

(c) [2] Find $\text{E}[\boldsymbol{v}]$.

$$E[\vec{v}] = \begin{bmatrix} E[\hat{\vec{\beta}}] \\ ---- \\ E[\vec{e}] \end{bmatrix} = \begin{bmatrix} (X^T X)^{-1} X^T E[\vec{y}] \\ ------ \\ (I - H) E[\vec{y}] \end{bmatrix}$$

$$= \begin{bmatrix} (X^T X)^{-1} X^T X \vec{\beta} \\ ------ \\ (I - H) X \vec{\beta} \end{bmatrix}$$

$$= \begin{bmatrix} \vec{\beta} \\ ---- \\ \vec{0} \end{bmatrix}$$

(d) [2] Is $\widehat{\boldsymbol{\beta}}$ an unbiased estimator of $\boldsymbol{\beta}$?

Circle one: [✓ YES] NO

Explain: $E[\hat{\vec{\beta}}] = \vec{\beta}$

(e) [2] Find Var[$\boldsymbol{v}$].

$$\text{Var}[\vec{v}] = A\,\text{Var}[\vec{y}]\,A^T$$

$$= A\,\sigma^2 I\,A^T$$

$$= \sigma^2 A A^T$$

$$= \sigma^2 \left[\begin{array}{c} (X^TX)^{-1}X^T \\ \hline I-H \end{array}\right] \left[\begin{array}{c:c} ((X^TX)^{-1}X^T)^T & (I-H)^T \end{array}\right]$$

$$= \sigma^2 \left[\begin{array}{c:c} (X^TX)^{-1}X^TX(X^TX)^{-1} & (X^TX)^{-1}X^T(I-H) \\ \hdashline (I-H)X(X^TX)^{-1} & (I-H)(I-H) \end{array}\right]$$

$$= \sigma^2 \left[\begin{array}{c:c} (X^TX)^{-1} & 0_{(p+1)\times n} \\ \hdashline 0_{n\times(p+1)} & I-H \end{array}\right]$$

(f) [2] Are $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$ independent?

Circle one: [✓ YES] NO

Explain:

The off-diagonal blocks of Var[$\vec{v}$] (i.e., the covariances between $\vec{\beta}$ and $\vec{e}$) are all zero.

BONUS [1 point]

Which brewery did "*Student*" work for? Circle one.

(a) Block Three Brewing Co.

(b) Stella Artois

(c) Elora Brewing Co.

[✓ (d) Guinness]

You may use this page for rough work

For the indicated value of p, the following tables provide x^* where $\mathrm{P}(X \geq x^*) = p$

$\boldsymbol{X \sim t_{(8)}}$

p	x^*
→ 0.005	3.3554
0.01	2.8965
→ 0.025	2.3060
0.05	1.8595
0.1	1.3968

$\boldsymbol{X \sim t_{(9)}}$

p	x^*
0.005	3.2498
0.01	2.8214
0.025	2.2622
0.05	1.8331
0.1	1.3830

$\boldsymbol{X \sim t_{(24)}}$

p	x^*
0.005	2.7969
0.01	2.4922
0.025	2.0639
0.05	1.7109
0.1	1.3178

$\boldsymbol{X \sim t_{(25)}}$

p	x^*
0.005	2.7874
0.01	2.4851
→ 0.025	2.0595
0.05	1.7081
0.1	1.3163

$\boldsymbol{X \sim t_{(45)}}$

p	x^*
0.005	2.6896
0.01	2.4121
→ 0.025	2.0141
0.05	1.6794
0.1	1.3006

$\boldsymbol{X \sim t_{(52)}}$

p	x^*
0.005	2.6737
0.01	2.4002
0.025	2.0066
0.05	1.6747
0.1	1.2980