Make sure that this examination has 3 numbered pages

University of Regina
Department of Mathematics & Statistics
Final Examination
201830
(December 21, 2018)

Statistics 354
Linear Statistical Methods

Name: ______________________________ **Student Number:** ______________

Instructor: Michael Kozdron **Time:** 3 hours

Read all of the following information before starting the exam.

*You have **3** hours to complete this exam. Please read all instructions carefully, and check your answers. Show all work neatly and in order, and clearly indicate your final answers. Answers must be justified whenever possible in order to earn full credit. **Unless otherwise specified, no credit will be given for unsupported answers, even if your final answer is correct.** Points will be deducted for incoherent, incorrect, and/or irrelevant statements. Unless otherwise noted, you must answer all questions in the test booklets provided.*

*You are permitted to have **TWO** 8.5×11 page of handwritten notes (double-sided) for your personal use, as well as a non-programmable calculator. No other aids are allowed. The order of the test questions is essentially random; they are not intentionally written easiest-to-hardest.*

*This test has **3** numbered pages with **7** questions totalling **150** points. The number of points per question is indicated. For questions with multiple parts, all parts are equally weighted.*

1. (*16 points*) Consider the multiple linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$. Let $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, $\hat{\boldsymbol{\mu}} = H\mathbf{y}$, $\mathbf{e} = (1 - H)\mathbf{y}$ where $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

(a) Carefully determine the distribution of $\hat{\boldsymbol{\mu}}$. (Simplify your answer as much as possible.)

(b) Carefully determine the distribution of $\mathbf{e}$. (Simplify your answer as much as possible.)

2. (*24 points*) In a study on the effect of coffee consumption on blood pressure, 30 patients are selected at random from among the patients of a medical clinic. A questionnaire is administered to each patient to get the following information:

x_1: number of cups of coffee consumed per day

x_2: number of minutes of daily exercise

x_3: age

x_4: sex ($x_4 = 0$ for males, $x_4 = 1$ for females)

y: systolic blood pressure during last visit to the medical clinic

A linear model of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$ is proposed, where the errors $\epsilon_1, \ldots, \epsilon_{30}$ are independent and identically distributed with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all i

(a) Carefully explain the meaning of the parameter β_4.

(b) If β_1 is very large, can we conclude from this study that increased coffee consumption **causes** increased blood pressure? Discuss.

(c) Another model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_4 + \epsilon$ is fit to the data. Explain the meaning of the hypothesis $\beta_5 = 0$.

3. (*14 points*) Consider a regression through the origin

$$y_i = \beta x_i + \epsilon_i$$

where $x_i > 0$ for $i = 1, \ldots 8$, and $\epsilon_1, \ldots, \epsilon_8$ are independent with $\epsilon_i \sim \mathcal{N}(0, \sigma^2 x_i^{-1})$. Derive the generalized least squares estimate of β and obtain its variance.

4. (*16 points*) Percentage yields from a chemical reaction for changing temperature (factor 1) and concentration of a certain ingredient (factor 2) are as follows:

x_1	x_2	Percentage yield (y)
-1	-1	79
1	-1	74
-1	1	76
1	1	70

The regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

is proposed, where $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ are independent and identically distributed with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all i.

(a) Estimate the coefficients in the regression model. That is, determine numerical values for $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$.

(b) The mean square error is found to be $s^2 = 0.25$. Estimate the squared standard errors in the regression model. That is, determine numerical values for $\hat{V}(\hat{\beta}_0)$, $\hat{V}(\hat{\beta}_1)$, $\hat{V}(\hat{\beta}_2)$.

5. (*32 points*) Suppose that we observe data (x_i, y_i), $i = 1, \ldots, n$, and postulate that it is appropriate to describe the relationship between x and y by the regression model

$$y_i = \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

where β_1 and β_2 are parameters and $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all i.

(a) Carefully determine

(i) a 2×1 vector $\boldsymbol{\beta}$,

(ii) $n \times 1$ vectors $\mathbf{y}$, $\boldsymbol{\epsilon}$, and

(iii) an $n \times 2$ matrix $\mathbf{X}$

so that this model can be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

Let $\hat{\beta}_1$, $\hat{\beta}_2$ denote the least squares estimators of β_1, β_2, respectively. It is a fact that the least squares result derived in class applies to this model: if $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1, \hat{\beta}_2]'$, then $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

(b) Determine the distribution of $\hat{\beta}_1$.

(c) Determine the distribution of $\hat{\beta}_2$.

(d) Determine $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$.

6. (*16 points*) Suppose that we have data (x_i, y_i), $i = 1, \ldots, n$, and determine the fitted values for a simple linear regression to be

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n.$$

The least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$ are given by

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

However, suppose that the true model is actually a quadratic model so that

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all i. In particular, this implies that

$$\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2, \quad i = 1, \ldots, n,$$

or, in vector notation,

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + \beta_2 \mathbf{z}$$

where

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \text{and} \quad \mathbf{z} = \begin{bmatrix} x_1^2 \\ \vdots \\ x_n^2 \end{bmatrix}.$$

(a) Compute $\mathbb{E}(\hat{\boldsymbol{\beta}})$. (Note that this shows $\hat{\boldsymbol{\beta}}$ is a *biased* estimator of $\boldsymbol{\beta}$.)

(b) Conclude that if $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ denotes the vector of residuals, then $\mathbb{E}(\mathbf{e}) = \beta_2(I - H)\mathbf{z}$ where $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

7. (*32 points*) Accident rate data $y_1, \ldots, y_{12}$ were collected over 12 consecutive years $t = 1, \ldots, 12$. At the end of the sixth year, a change in safety regulations occurred. For each of the following situations, set up a linear model of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Define $\mathbf{X}$ and $\boldsymbol{\beta}$ appropriately.

(a) The accident rate y is a linear function of t with the new safety regulations having no effect.

(b) The accident rate y is a quadratic function of t with the new safety regulations having no effect.

(c) The accident rate y is a linear function of t. The slope for $t \geq 7$ is the same as for $t < 7$. However, there is a discrete jump in the function at $t = 7$.

(d) The accident rate y is a linear function of t. After $t = 7$, the slope changes, with the two lines intersecting at $t = 7$.

The End. Happy Holidays.