KING FAHD UNIVERSITY OF PETROLEUM & MINERALS, DHAHRAN, SAUDI ARABIA
DEPARTMENT OF MATHEMATICS

STAT 310: Regression Analysis

Term 211, First Major Exam
Saturday October 16, 2021, 07:00 PM

Name: _____ ID #: _____

Question No	Full Marks	Marks Obtained
1	**07**	
2	**10**	
3	**08**	
4 (Bonus)	**08**	
5	**17**	
6	**8**	
Total	**50**	

Instructions:

1. Formula sheet will be provided to you in exam. You are not allowed to bring, with you, formula sheet or any other printed/written paper.
2. Mobiles are not allowed in exam. If you have your **mobile** with you, **turn it off** and put it **under your seat** so that it is visible to proctor.
3. Show all the calculation steps. There are points for the steps so if your miss them, you lose points.
4. Derive every result that you use in your solution, unless mentioned otherwise.
5. Anything bold in a question indicates that it is a vector or matrix.

Q1: (7 points) Consider a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where β_0, β_1 are unknown parameters and x_i's are fixed. The OLS estimates of β_0 and β_1 are given as $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ where $S_{XY} = \sum XY - \frac{\sum X \sum Y}{n}$ and $S_{XX} = \sum X^2 - \frac{(\sum X)^2}{n}$.

Starting with $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, mathematically derive the $\text{Var}(\hat{\beta}_0)$.

Note: You can use $\sum c = 0$, $\sum cx = 1$ and $\sum c^2 = \frac{1}{S_{XX}}$ without deriving, where $c = \frac{x - \bar{x}}{S_{XX}}$.

Now, for the sampling variance of these estimators:

$$V\left(\hat{\beta}_1\right) = V\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2 V\left(y_i\right) + \sum_{i \neq}^{n} \sum_{j}^{n} c_i c_i Cov\left(y_i y_j\right)$$

$$= \sum_{i=1}^{n} c_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \left(\frac{1}{S_{XX}}\right) = \frac{\sigma^2}{S_{XX}}$$

These covariance terms are equal to zero because the values of y (i.e. y_i and y_j) are independent of each other.

Similarly, $V\left(\hat{\beta}_0\right) = V\left(\sum \left(\frac{1}{n} - c_i \bar{x}\right) y_i\right)$

$$= \sum_{i=1}^{n} \left(\frac{1}{n} - c_i \bar{x}\right)^2 V\left(y_i\right) + \sum_{i \neq}^{n} \sum_{j}^{n} \left(\frac{1}{n} - c_i \bar{x}\right)\left(\frac{1}{n} - c_j \bar{x}\right) Cov\left(y_i y_j\right)$$

$$= \sum_{i=1}^{n} \left(\frac{1}{n} - c_i \bar{x}\right)^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} + c_i^2 \bar{x}^2 - \frac{2}{n} c_i \bar{x}\right)$$

$$= \sigma^2 \left(\frac{n}{n^2} + \bar{x}^2 \sum_{i=1}^{n} c_i^2 - \frac{2}{n} \bar{x} \sum_{i=1}^{n} c_i\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)$$

Q2: (4+3+3 = 10 points) For a multiple linear regression model $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times(k+1)}\boldsymbol{\beta}_{(K+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$,

a) Derive the ordinary least square (OLS) estimates of $\boldsymbol{\beta}$ vector.

The objective is to derive $\hat{\beta}$ subject to minimizing $\epsilon'\epsilon$.

$$\epsilon'\epsilon = (y - x\beta)'(y - x\beta)$$

$$= y'y - \beta'x'y - y'x\beta + \beta'x'x\beta$$

$$= y'y - 2\beta'x'y + \beta'x'x\beta$$

Differentiating w.r.t. β and putting $= 0$

$$\frac{\partial}{\partial\beta}\epsilon'\epsilon\Big|_{\hat{\beta}} \quad -2x'y + 2x'x\hat{\beta} = 0$$

$$\Rightarrow (x'x)\hat{\beta} = x'y$$

$$\Rightarrow \hat{\beta} = (x'x)^{-1}x'y$$

b) Show that $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.

1- $E(\hat{\beta}) = \beta$

$E(\hat{\beta}) = E\left[(X'X)^{-1} X'y\right]$

$= (X'X)^{-1} X' E(X\beta + \underline{\epsilon})$

$= (X'X)^{-1} X' (X\beta)$

$= (X'X)^{-1} (X'X) \beta$

$= I\beta$

$E(\hat{\beta}) = \beta$

$V(\epsilon) = \sigma^2$

$V(2\epsilon) = 2^2 \, Var(\epsilon)$

c) Also derive the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$.

$Var\text{-}Cov(\hat{\beta}) = V\text{-}C\left[(X'X)^{-1} X'y\right]$

$= (X'X)^{-1} X' \, Var(y) \left[(X'X)^{-1} X'\right]'$

$= (X'X)^{-1} X' \sigma^2 X (X'X)^{-1}$

$= \sigma^2 (X'X)^{-1} X'X (X'X)^{-1}$

$Var\text{-}Cov(\hat{\beta}) = \sigma^2 (X'X)^{-1}$

Q3: (8 points) Consider a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where β_0, β_1 are unknown parameters and x_i's are fixed. The OLS estimates of β_0 and β_1 are given as:

$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ where the formulas of S_{XY} and S_{XX} are provided in formula sheet. Mathematically derive the $E(\hat{\beta}_0 \hat{\beta}_1)$.

Hint: The covariance between 2 variables X and Y is $Cov(X,Y) = E(XY) - E(X)E(Y)$

Note: You can use the unbiasedness of $\hat{\beta}_0$ and $\hat{\beta}_1$ without deriving it (if needed anywhere.)

We start with $Cov(\bar{Y}, \hat{\beta}_1) = Cov\left[\frac{\sum Y_i}{n}, \sum c_i Y_i\right]$

$= Cov\left[\sum \left\{\frac{1}{n}(Y_i, c_i Y_i)\right\}\right]$

$= \frac{1}{n} \sum Cov(Y_i, Y_i) c_i = \frac{Var(Y_i) \sum c_i}{n}$

$= \frac{\sigma^2}{n} \sum c_i = \frac{\sigma^2}{n}(0) = 0$

Here $c_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$ with $\sum c_i = 0$

Now $Cov(\hat{\beta}_0, \hat{\beta}_1) = Cov[\bar{Y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1]$

$= Cov(\bar{Y}, \hat{\beta}_1) - \bar{x}\, Cov(\hat{\beta}_1, \hat{\beta}_1)$

$= 0 - \bar{x}\, Var(\hat{\beta}_1) = -\bar{x} \frac{\sigma^2}{S_{XX}}$

Finally $Cov(\hat{\beta}_0, \hat{\beta}_1) = E(\hat{\beta}_0 \hat{\beta}_1) - E(\hat{\beta}_0) E(\hat{\beta}_1)$

$\Rightarrow -\bar{x} \frac{\sigma^2}{S_{XX}} = E(\hat{\beta}_0 \hat{\beta}_1) - \beta_0 \beta_1$

$\Rightarrow E(\hat{\beta}_0 \hat{\beta}_1) = \beta_0 \beta_1 - \bar{x} \frac{\sigma^2}{S_{XX}}$

Q4: (6 bonus points) Show that MSE is an unbiased estimator of σ^2 for a simple linear regression model. Note: No partial credit for this question.

Good Luck

Q5: Code 1

Download the dataset from Blackboard and write down the code number in above blank.

(2+5+1+5+3+1 = 17 points) An accountant for a large department store would like to develop a model to predict the amount of time (y) it takes to process invoices (x). Data are collected from the past 26 working days, and the number of invoices processed and completion time (in hours) are stored.

a) Estimate the covariance between the number of invoices and completion time. Interpret its meaning (if any)

Cov(x, y) = 73.63

This implies a positive linear association b/w no. of invoices and time to process

b) Fit the following two models to given dataset:

Model 1: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\hat{\beta}_1 = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Model 2: $y_i = \beta_1 x_i + \epsilon_i$, $\hat{\beta}_1 = \frac{\sum XY}{\sum X^2}$.

Which model has smaller *SSE*? Explain why.

SSE(model 1) = 3.421 *SSE*(model 2) = 5.103

Reason: Forcing the line to pass through the origin increases the magnitude of residuals.

c) For model 1, what percent of the variation in completion time is not explained by number of invoices?

13.3 percent.

d) For model 1, test the hypothesis that *"the completion time increases by more than one hour due to an increase of one unit in number of invoices, and vice versa"* i.e. $\beta_1 > 1$.

H_0: $\beta_1 \leq 1$ H_1: $\beta_1 > 0$

Test Statistic = −1019.71

p-value = 1

Decision and conclusion: Fail to reject H_0 & conclude that the completion time does not increase by more than one hour due to an increase of one unit in no. of invoices.

e) For model 1, construct a 98% interval estimate for the average completion time of 3 invoices. Interpret the interval.

Lower limit = 0.1817 Upper Limit = 0.9048

Interpretation: We are 98% confident that the average time needed to complete 3 invoices is between (0.1817, 0.9048)

f) Is prediction done in part e) interpolation or extrapolation? Justify your answer with a valid reason.

Extrapolation. We are predicting outside the domain of available data.

Q6: Code 1

Download the dataset from Blackboard and write down the code number in above blank.

(5+3 = 8 points) The state of Texas in US is divided into 254 counties. A sample of 13 counties is selected. For each county, the "Average January High Temperature", the "Latitude", the "Longitude" and the "Elevation" are measured. Fit a multiple regression line for predicting the temperature based on longitude, latitude and elevation of the county.

a) Fill in the following blanks:

$\hat{\boldsymbol{\beta}} =$

$$\begin{bmatrix} 155.11 \\ -1.965 \\ -0.432 \\ -0.0092 \end{bmatrix}$$

$\boldsymbol{Var-\widehat{Cov}(\hat{\beta})} =$

$$\begin{bmatrix} 690.24 & -2.83 & -6.26 & 0.015 \\ -2.83 & 0.02 & 0.023 & -7\times10^{-5} \\ -6.26 & 0.023 & 0.058 & -0.00013 \\ 0.015 & -7\times10^{-5} & -0.00013 & 3.59\times10^{-7} \end{bmatrix}$$

b) There is another county named Houston in Texas. The latitude, the longitude and the elevation of Houston are 29.7604, 95.3698 and 49, respectively. Predict the average January High Temperature of Houston and also construct an interval estimate using $\alpha = 0.06$.

Predicted Average High Temperature of Houston = 55.33

Lower limit: 53.26 Upper Limit: 57.41

[Good Luck]