MAT 3375, Regression Analysis

Assignment 1 (Total = 50)
Professor: Termeh Kousha
Deadline: **Friday, Sep 23, 2016 SHARP 3pm (Math department (585 KED) drop boxes)**

Late assignments will NOT be accepted; nor will unstapled assignments. Professors in the math department will not lend you a stapler.
You should complete ALL the questions in the assignments. It is possible, however, that not all the questions will be marked. In that case, the same questions will be marked in all assignments. You will not be informed beforehand which questions will be marked.

Student Name _____
Student Number _____
By signing below, you declare that this work was your own and that you have not copied from any other individual or other source.

Signature _____

8 points

1. Question 2.25

4 Points

part (b):

$$\text{Cov}(\bar{Y}, \hat{\beta}_1) = \text{Cov}\left(\frac{\sum Y_i}{n}, \frac{S_{xY}}{S_{xx}}\right) = \frac{1}{n S_{xx}} \text{Cov}\left(\sum Y_i, \sum (x_i - \bar{x}) Y_i\right)$$

$$= \frac{1}{n S_{xx}} \sum \text{Cov}\left(Y_i, \underbrace{(x_i - \bar{x})}_{\text{constant}} Y_i\right) =$$

$$= \frac{1}{n S_{xx}} \sum (x_i - \bar{x}) \underbrace{\text{Cov}(Y_i, Y_i)}_{\sigma^2} = \frac{\sigma^2}{n S_{xx}} \sum (x_i - \bar{x}) = 0$$

(the $\sum (x_i - \bar{x})$ term is 0)

4 points

Part (a) $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{Cov}(\bar{Y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) =$

$$\underbrace{\text{Cov}(\bar{Y}, \hat{\beta}_1)}_{=0 \text{ by part(b)}} - \text{Cov}(\hat{\beta}_1 \bar{x}, \hat{\beta}_1) = -\bar{x}\, \text{Cov}(\hat{\beta}_1, \hat{\beta}_1)$$

$$= -\bar{x}\, \text{Var}(\hat{\beta}_1)$$

$$= -\frac{\bar{x} \sigma^2}{S_{xx}}$$

2. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent, $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$. Show that

$$Var[\widehat{\beta}_0] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right].$$

Hint: Use question 1. ($Cov(\hat{Y}_i, \hat{\beta}_0] = 0$.

$$Var(\hat{\beta}_0) = Var(\bar{Y} - \hat{\beta}_1 \bar{x}) = Var(\bar{Y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\,Cov(\bar{Y}, \hat{\beta}_1)$$

$$= \frac{\sigma^2}{n} + Var(\hat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\bar{x}^2}{S_{xx}}\sigma^2$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right).$$

3. Consider the following sum of cross-products:

$$S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

Show that S_{XY} can be written as $S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})Y_i$ and $S_{XY} = (\sum_{i=1}^{n} X_i Y_i) - n\bar{X}\bar{Y}$. Consequently, by plugging-in $Y_i = X_i$ we get

$$S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \left(\sum_{i=1}^{n} X_i^2\right) - n\bar{X}\bar{X}.$$

$$S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})X_i.$$

$$S_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum(x_i - \bar{x})Y_i - \underbrace{\sum(x_i - \bar{x})\bar{Y}}_{\bar{Y}\sum(x_i - \bar{x}) = 0}$$

$$= \sum(x_i - \bar{x})Y_i = \sum x_i Y_i - \bar{x}\sum Y_i$$

$$= \sum x_i Y_i - n\bar{x}\bar{Y}$$

20 points

4. Simple linear regression method was used to analyze the data from a study investigating the relationship between roadway surface temperature in Fahrenheit (X) and pavement deflection (Y). Summary statistics were $n = 20$,

$$\sum_{i=1}^{n} Y_i = 12.575, \quad \sum_{i=1}^{n} Y_i^2 = 8.86, \quad \sum_{i=1}^{n} X_i = 1478, \quad \sum_{i=1}^{n} X_i^2 = 143,215.8, \quad \sum_{i=1}^{n} X_i Y_i = 1083.67.$$

Note: Alternatively, SSCP table can be given.

(a) Verify that $S_{XY} = 154.3775$, $S_{XX} = 33991.6$, $S_{YY} = 0.9534688$. (*Hint:* Use Question 3)

(b) Obtain the least squares estimates of the slope and intercept.

(c) Give an estimate of the mean pavement deflection when the surface temperature is 85F.

(d) Give a point estimate of the pavement deflection when the surface temperature is 90F.

(e) Compute the coefficient of correlation.

(f) Compute the coefficient of determination and interpret it.

(g) Estimate σ^2. *Hint:* The alternative formula for SSE is $SSE = S_{YY} - \hat{\beta}_1 S_{XY}$.

(h) Use t-test to check for significance of regression using $\alpha = 0.05$. Find the p-value for this test. What conclusion can you draw?

(i) Use F-test to check for significance of regression using $\alpha = 0.05$

(j) Find a 99% confidence interval for slope. *Hint:* $t_{0.01/2,18} = 2.878$.

(k) Find a 99% confidence interval for intercept.

(l) Find a 99% confidence interval for the mean deflection when temperature is 85F.

(m) Find a 99% prediction interval on pavement deflection when the temperature is 90F.

(a) verfiy.

(2.5) ← b) $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

(1P) (1P)

So $\hat{\beta}_1 = 0.004542$ $\hat{\beta}_0 = 0.293123$

$\hat{y}_i = 0.293123 + 0.004542(x_i)$

(0.5) c) $x_i = 85$ $\hat{y}_i = 0.68$ (0.5) d) $x_i = 90 \rightarrow \hat{y}_i = 0.7$

(1) e) $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = 0.857$

(1) (f) $R^2 = r^2 = 0.735$ → That means 73.5% →

Continue

← of the variation in pavement defection is explained by the linear relationship. with the surface temperature.

(2) (8) $\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{SYY - \hat{\beta}_1 SXY}{n-2} = 0.01402$

(2) h) $H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} = 7.072.$$

$t_{20-2} = t_{18}$ $\quad 2\left(P(t_{18} > 7.072)\right) < 0.001$

Reject the Null hypo. → there is a linear relation ship between x and y.

(3) (i) $F^* = \frac{SSR}{MSE} = \frac{SST - SSE}{0.01402} = \frac{0.9534688 - 18(0.01402)}{0.01402}$

$$= \frac{0.7011}{0.01402} = 50.007$$

$F^* > \underbrace{F_{0.05,1,18}}_{4.41}$ → we can Reject.

(2) (j) $SD[\hat{\beta}_1] = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$ $\quad \hat{\beta}_1 \in (0.0027, 0.0064)$

(2) (K) $SD[\hat{\beta}_0] = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$ $\quad \hat{\beta}_0 \in (0.137, 0.45)$

(2) (l) $X_h = 85$ $\quad E(Y_h) = \beta_0 + \beta_1 X_h$

$$\hat{Y}_h \pm t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right]}$$

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = 0.68$$

$$\hat{Y}_h \in (0.601, 0.761).$$

(m) $\quad X_{h(new)} = 90.$

(2) $\quad \hat{Y}_{h(new)} = \hat{\beta}_0 + \hat{\beta}_1 (90) = 0.7$

$$\hat{Y}_{h(new)} \pm t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{x_h - \bar{x})^2}{S_{xx}}\right.}$$

$$Y_{h(new)} \in (0.3513, 1.0524).$$

12 points.

5. Consider the simple regression model with $\beta_0 = 0$ (regression passing through the origin). Thus, the regression model is

$$Y_i = \beta_1 X_i + \epsilon_i,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent, $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$.

4 Ps (a) Find the least-squares estimator of β_1. Is it unbiased?

3 Ps (b) Compute $Var[\widehat{\beta}_1]$.

5 Ps (c) Let $\widehat{Y}_i = \widehat{\beta}_1 X_i$. Show that

$$\widehat{\sigma^2} := \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

is an unbiased estimator of σ^2.

(a). we want to minimize

$$Q(\beta_1) = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2$$

$$\frac{dQ}{d\beta_1} = -2\sum (Y_i - \beta_1 X_i) X_i = 0$$

2 points

$$\sum Y_i X_i - \beta_1 \sum_i X_i^2 = 0 \qquad \hat{\beta}_1 = \frac{\sum Y_i X_i}{\sum_i X_i^2} \quad *$$

$$E(\hat{\beta}_1) = E\left(\frac{\sum_i Y_i X_i}{\sum_i X_i^2}\right) = \frac{\sum_i X_i E(Y_i)}{\sum X_i^2} = \frac{\sum X_i \beta_1 X_i}{\sum X_i^2} = \beta_1 \frac{\sum X_i^2}{\sum X_i^2} = \beta_1$$

2 points. $\hat{\beta}_1$ is an unbiased estimator for β_1.

(b) $$var[\hat{\beta}_1] = var\left[\frac{\sum_i Y_i X_i}{\sum_i X_i^2}\right] = \frac{1}{(\sum_i X_i^2)^2} \sum_{i=1}^{n} X_i^2 var(Y_i)$$

$$= \frac{1}{(\sum X_i^2)^2} \sigma^2 \sum_{i=1}^{n} X_i^2 = \frac{\sigma^2}{\sum_i X_i^2}$$

(C) we have $\sum (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - \hat{\beta}_1 X_i)^2 = \sum Y_i^2 + \hat{\beta}_1^2 \sum X_i^2$

$- 2\hat{\beta}_1 \sum_i X_i Y_i = \sum Y_i^2 + \hat{\beta}_1^2 \sum_i X_i^2 - 2\hat{\beta}_1 (\sum X_i^2 \hat{\beta}_1)$ from *

$$= \sum Y_i^2 - \hat{\beta}_1^2 \sum_i X_i^2$$

→

Note that

$$E(Y_i^2) = \text{var}(Y_i) + E(Y_i)^2 = \sigma^2 + \beta_1^2 x_i^2$$

and

$$E(\hat{\beta}_1^2) = \text{var}(\hat{\beta}_1) + E(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum x_i^2} + \beta_1^2$$

Thus

$$E\left(\sum (Y_i - \hat{Y}_i)^2\right) = E\left(\sum Y_i^2 - \hat{\beta}_1^2 \sum x_i^2\right)$$

$$= \sum E(Y_i^2) - E(\hat{\beta}_1^2) \sum x_i^2$$

$$= \sum (\sigma^2 + \beta_1^2 x_i^2) - \left(\frac{\sigma^2}{\sum x_i^2} + \beta_1^2\right) \sum x_i^2$$

$$= n\sigma^2 + \sum \beta_1^2 x_i^2 - \sigma^2 - \beta_1^2 \sum x_i^2 =$$

$$n\sigma^2 - \sigma^2 = (n-1)\sigma^2.$$

Therefore:

$$E(\hat{\sigma}^2) = E\left(\frac{\sum (Y_i - \hat{Y}_i)^2}{n-1}\right) = \sigma^2$$

which shows that $\hat{\sigma}^2$ is an unbiased estimator for σ^2.

5 points.

6. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent, $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$ Consider the test

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0.$$

Show that t-test and F-test are equivalent.
Hint: Use the following result: if $Z \sim t(\nu)$, then $Z^2 \sim F(1, \nu)$.

The t-test : $t^* = \dfrac{\hat{\beta}_1}{\frac{s}{\sqrt{S_{xx}}}} \sim t_{(n-2)}$.

$T \sim t_{n-2}$ P-value is $2P(|T| > |t^*|)$

for the F-test $F^* = \dfrac{MSR}{MSE} = \dfrac{\hat{\beta}_1^2 S_{xx}}{\hat{\sigma}^2} =$

$s^2 = MSE$

$$\frac{\hat{\beta}_1^2}{\frac{s^2}{S_{xx}}} = (t^*)^2 \sim F_{1,n-2}$$

The P-value is

$$2P(F_{1,n-2} > F^*) =$$

$$2P((T)^2 > F^* = t^{*2}) =$$

$T \sim t_{n-2}$ $2P(|T| > |t^*|) = 2P(T > t^*)$

two test are equivalent.

7 Always positive

5 points.

7. Let r_{XY} ne the sample correlation coefficient between X and Y:

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}.$$

Let $Z = aX + b$, $a > 0$. Show that $r_{ZY} = r_{XY}$.

We first show $S_{ZY} = a S_{XY}$

$$S_{ZY} = \sum_i z_i y_i - n\bar{z}\bar{y} = \sum_i (ax_i + b) y_i - n(a\bar{x} + b)\bar{y}$$

$$= a\sum_i x_i y_i + b\sum_i y_i - na\bar{x}\bar{y} - nb\bar{y} =$$

$$= a\sum_i x_i y_i - na\bar{x}\bar{y} = a\left(\sum_i x_i y_i - \bar{x}\bar{y}\right).$$

Also you can show simmary $S_{ZZ} = a^2 S_{XX}$.

Therefore:

$$r_{ZY} = \frac{S_{ZY}}{\sqrt{S_{ZZ}\, S_{YY}}} = \frac{a\, S_{XY}}{\sqrt{a^2 S_{XX}\, S_{YY}}} =$$

$$\frac{S_{XY}}{\sqrt{S_{XX}\, S_{YY}}}.$$