USF

MATH 372 Fall 2017

Final Exam

Friday December 8th, 2017

10:00am – 12:00pm

First Name: Solutions

Last name:

Instructions:

- **Clearly write your name on this cover page.**
- **This exam consists of 16 pages including this cover page.**
- **If you need extra space please use page 16 labeled "LEFT BLANK" and INDICATE that you have done so. You may remove this page for your convenience.**
- **Page 15 contains tables of quantiles from the $N(0,1)$, $t_{(196)}$ and $F_{(3,196)}$ distributions.**

Question	Points
Q1	/10
Q2	/16
Q3	/18
Q4	/10
Q5	/10
Q6	/11
Total	/75

Question 1 (10 points)

Indicate, by circling T or F, whether the following statements are TRUE or FALSE.

1 (a) [T or [✓ F]] In a simple linear regression of y on x, it is found that the t-statistic for testing $\beta_1 = 0$ was nonsignificant. This implies that there is no relation between y and x.

1 (b) [[✓ T] or F] $y = \beta_0 + \beta_1 \log(x) + \varepsilon$ is a linear regression model in the sense used in class.

1 (c) [T or [✓ F]] Stepwise regression techniques (i.e., forward, backward, hybrid) always lead to the same set of selected predictors.

1 (d) [[✓ T] or F] The log-transformation is appropriate for stabilizing non-constant variability when $SD[y] \propto E[y]$.

1 (e) [T or [✓ F]] The addition of a variable to a regression equation always causes R^2_{adj} to increase.

1 (f) [T or [✓ F]] Observations with large leverage are always influential.

1 (g) [[✓ T] or F] k-fold cross validation tends to provide less variable results than ordinary cross validation.

1 (h) [[✓ T] or F] Multicollinearity exists when an explanatory variable is highly correlated with other explanatory variables.

Use the following scenario in the next two questions: Let $y_i = \mu_i + \varepsilon_i$ where $\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ for $i = 1,2,\ldots,20$, and consider predicting a future response y_p.

1 (i) [[✓ T] or F] A 99% confidence interval for μ_p is narrower than a 99% prediction interval for y_p.

1 (j) [T or [✓ F]] A 95% prediction interval for y_p is wider than a 99% prediction interval for y_p.

Question 2 (16 points)

(a) (4) In the context of multiple linear regression, the relationship between the response variable y and explanatory variables x_1, x_2, …, x_p may be stated as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

for $i = 1,2,\ldots,n$. Define the vectors $\boldsymbol{y}$, $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}$ and the matrix X that allow this system of equations to be written in vector-matrix notation as follows:

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

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \checkmark \quad \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \checkmark \quad \vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \checkmark \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \checkmark$$

4

(b) (1) State the matrix equation for the least squares estimate of $\boldsymbol{\beta}$.

$$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y} \checkmark$$

1

(c) (4) Using the assumption that $y \sim MVN(\overset{X\vec{\beta}}{\cancel{\mathbf{0}}}, \sigma^2 I)$, where I is the $n \times n$ identity matrix, derive the mean vector and the variance-covariance matrix for the least squares estimator $\widehat{\boldsymbol{\beta}}$. In other words, calculate $E[\widehat{\boldsymbol{\beta}}]$ and $Var[\widehat{\boldsymbol{\beta}}]$.

$$E[\hat{\vec{\beta}}] = E[(X^TX)^{-1}X^T\vec{y}]$$

$$= (X^TX)^{-1}X^T E[\vec{y}]$$

$$= (X^TX)^{-1}X^T X\vec{\beta}$$

$$= \vec{\beta} \quad ✓✓$$

4

$$Var[\hat{\vec{\beta}}] = Var[(X^TX)^{-1}X^T\vec{y}]$$

$$= (X^TX)^{-1}X^T Var(\vec{y}] \, [(X^TX)^{-1}X^T]^T$$

$$= (X^TX)^{-1}X^T \sigma^2 I \, X[(X^TX)^{-1}]^T$$

$$= \sigma^2 (X^TX)^{-1}X^TX[(X^TX)^T]^{-1}$$

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

$$= \sigma^2 (X^TX)^{-1} \quad ✓✓$$

(d) (7) Suppose interest lies in investigating the relationship between a response variable y and two explanatory variables x_1 and x_2 and so the following data are collected:

i	y	x_1	x_2
1	103	150	12
2	135	150	24
3	111	190	12
4	138	190	24

Consider fitting the following *mean-corrected* linear regression model to this data:

$$y_i = \beta_0 + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \varepsilon_i$$

where $i = 1,2,3,4$ and where $\bar{x}_1$ and $\bar{x}_2$ are the respective sample means of the two explanatory variables.

i. (1) Write the response vector y associated with the data given in the table above.

$$\vec{y} = \begin{bmatrix} 103 \\ 135 \\ 111 \\ 138 \end{bmatrix} \checkmark$$

1

ii. (2) Write the matrix X of the mean-corrected model based on the data given in the table above.

$$X = \begin{bmatrix} 1 & -20 & -6 \\ 1 & -20 & 6 \\ 1 & 20 & -6 \\ 1 & 20 & 6 \end{bmatrix} \checkmark\checkmark$$

2

iii. (2) Calculate the least squares estimate $\widehat{\boldsymbol{\beta}}$.

$$X^TX = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -20 & -20 & 20 & 20 \\ -6 & 6 & -6 & 6 \end{bmatrix} \begin{bmatrix} 1 & -20 & -6 \\ 1 & -20 & 6 \\ 1 & 20 & -6 \\ 1 & 20 & 6 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1600 & 0 \\ 0 & 0 & 144 \end{bmatrix}$$

$$(X^TX)^{-1} = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/1600 & 0 \\ 0 & 0 & 1/144 \end{bmatrix}$$

$$X^T\vec{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -20 & -20 & 20 & 20 \\ -6 & 6 & -6 & 6 \end{bmatrix} \begin{bmatrix} 103 \\ 135 \\ 111 \\ 138 \end{bmatrix} = \begin{bmatrix} 487 \\ 220 \\ 354 \end{bmatrix}$$

$$\hat{\vec{\beta}} = (X^TX)^{-1} X^T\vec{y} = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/1600 & 0 \\ 0 & 0 & 1/144 \end{bmatrix} \begin{bmatrix} 487 \\ 220 \\ 354 \end{bmatrix} = \begin{bmatrix} 121.75 \\ 0.1375 \\ 2.458 \end{bmatrix}$$ ✓✓

2

iv. (2) Interpret the value of β_0 in this model.

β_0 represents the average response ✓ when x_1 and x_2 are equal to their average values (respectively 170 and 18). In this case the expected response is 121.75. ✓

2

Question 3 (18 points)

A linear regression was fit between a response variable y and three explanatory variables x_1, x_2, x_3 using $n = 200$ data points. Partial Python output from this model is shown below.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       ?????
Model:                            OLS   Adj. R-squared:                  ?????
Method:                 Least Squares   F-statistic:                     ?????
No. Observations:                 200   Prob (F-statistic):              ?????
Df Residuals:                     196   Log-Likelihood:                -386.18
Df Model:                           3   AIC:                             780.4
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      2.9389      0.312      9.422      0.000         2.324     3.554
x1             0.0458      0.001     32.809      0.000         0.043     0.049
x2             0.1885      0.009     21.893      0.000         0.172     0.206
x3            -0.0010      0.006     ??????      ?????         ?????     ?????
==============================================================================
Omnibus:                       60.414   Durbin-Watson:                   2.084
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              151.241
Skew:                          -1.327   Prob(JB):                     1.44e-33
Kurtosis:                       6.332   Cond. No.                         454.
==============================================================================
```

(a) (2) Interpret the value of β_3 in this model.

2 β_3 is the expected change in response associated with a unit increase in x_3, ✓ holding all else fixed. ✓ In this case we expect a decrease of 0.001.

(b) (3) Using the information from the output above, test the following hypothesis at a 5% level of significance. Note that you may find it useful to consult the statistical tables on page 15.

$$H_0: \beta_3 = 0 \text{ vs. } H_A: \beta_3 \neq 0$$

$$t = \frac{\hat{\beta}_3}{SE(\hat{\beta}_3)} = \frac{-0.001}{0.006} = -0.1667 ✓$$

3 $\text{p-value} = 2P(t_{(196)} \geq |-0.1667|)$

$= 2P(t_{(196)} \geq 0.1667)$

Note that $P(t_{(196)} \geq 3.183) = 0.025$

$\Rightarrow 2P(t_{(196)} \geq 3.183) = 0.05$

Since $|t| = 0.1667 < 3.183$, p-value > 0.05 ✓

Therefore we do not reject $\beta_3 = 0$ ✓ at a 5% S.L.

(c) (3) Below is a partially completed ANOVA table. Fill in the blank spaces.

3

Source	df	Sum of Sq.	Mean Sq.	F–Statistic
Regression	3	4860.51	1620.17	570.48
Error	196	556.64	2.84	
Total	199	5417.15		

✓✓✓

(d) (3) State and test (at a 5% level of significance) the hypothesis that the F-Statistic in (c) corresponds to. Note that you may find it useful to consult the statistical tables on page 15.

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs. $H_a: \beta_j \neq 0$ for $j = 1,2,3$ ✓

3

$F_0 = 570.48$

p-value $= P(F_{(3,196)} \geq 570.48)$

Since $P(F_{(3,196)} \geq 2.651) = 0.05$ and $F_0 > 2.651$ p-value < 0.05 ✓ Thus we reject H_0 ✓ and conclude that at least one of x_1, x_2 or x_3 significantly influences y.

(e) (1) What proportion of the variation in y is accounted for by the model?

1

$$R^2 = \frac{SSR}{SST} = \frac{4860.51}{5417.15} = 0.8972 \rightarrow 89.72\%$$ ✓

(f) (2) Calculate R^2_{adj} (i.e., adjusted-R^2) and describe the main advantage of using R^2_{adj} instead of R^2.

$$R^2_{adj} = 1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right)$$

2

$$= 1 - (0.1028)\left(\frac{199}{196}\right)$$

$$= 0.8956$$ ✓

This cannot be arbitrarily inflated by adding more predictors into the model. ✓

USF

(g) (4) Examine the plots in the figure below. Based on these plots answer "Yes" or "No" to the following questions and give a one sentence justification.

i. (1) Do the residuals appear to be normally distributed?

✓Yes, aside from one outlier we see a relatively straight line

ii. (1) Do the residuals appear to have constant variance?

✓Yes, aside from one outlier the residual variation seems to have constant amplitude

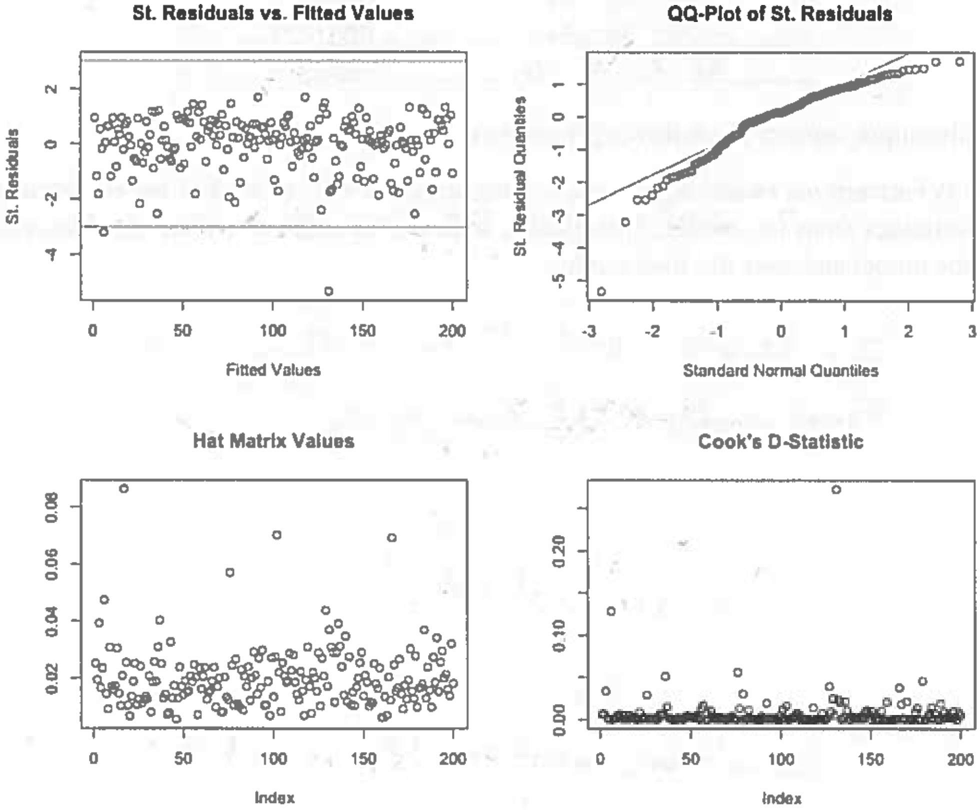
4

iii. (1) Do there appear to be any highly influential data points?

✓Yes, one Cook's D-value is much larger than the others.

iv. (1) Do there appear to be any data points with high leverage?

✓Yes there are several points with high leverage (h values larger than twice their average)



Question 4 (10 points)

Consider the relationship between a response variable y and four explanatory variables x_1, x_2, x_3 and x_4. The output provided gives the model summary of *all possible regressions*, which in this case corresponds to $2^4 = 16$ different models. Note: all models contain an intercept.

Variables in Model	AIC
None (intercept only)	176.9852
x_1	178.8042
x_2	168.5575
− x_3 ←	163.7744
x_4	177.5236
x_1 x_2	170.3156
− x_1 x_3	165.5900
x_1 x_4	179.1512
− x_2 x_3 ←	104.1719
x_2 x_4	170.4609
− x_3 x_4	163.8699
− x_1 x_2 x_3 ←	96.07582
x_1 x_2 x_4	172.1635
x_1 x_3 x_4	165.4374
− x_2 x_3 x_4	106.1541
x_1 x_2 x_3 x_4	97.47674

Using this output, answer the following questions.

(a) (4) Perform *backward stepwise selection* using the AIC as a basis for eliminating variables from the model. Specifically, indicate the order in which variables exit the model and state the final model.

x_4 leaves and then nothing else.

Final model has x_1, x_2, x_3 ✓

4

(b) (4) Perform *forward stepwise selection* using the AIC as a basis for adding variables into the model. Specifically, indicate the order in which variables enter the model and state the final model.

1. → x_3

2. → x_2

3. → x_1 ✓

4

Final model has x_1, x_2, x_3

(c) (2) The best overall model, among *all possible regressions*, is the one with the smallest AIC. Do these stepwise selection techniques choose the best overall model?

✓ Yes, it is the one with x_1, x_2, x_3 ✓

2

Question 5 (10 points)

In the context of a linear regression with two explanatory variables such as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

we may estimate $(\beta_0, \beta_1, \beta_2)$ using *shrinkage methods* such as ridge or LASSO regression. With these methods the estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ are the values of $(\beta_0, \beta_1, \beta_2)$ that minimize the error sum of squares:

$$S(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

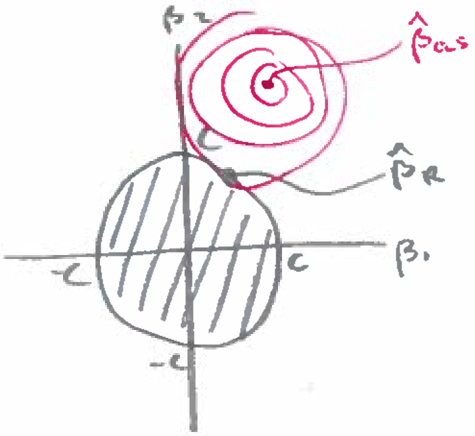
subject to one of the following two shrinkage constraints.

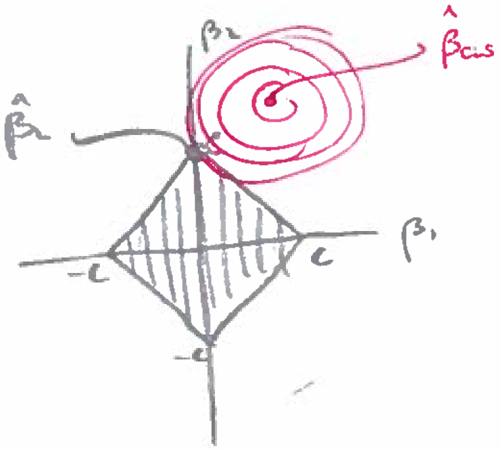
Ridge: $\beta_1^2 + \beta_2^2 \leq c$

LASSO: $|\beta_1| + |\beta_2| \leq c$

Draw these two constraint regions and explain why, in general, LASSO estimates can be 0, but ridge estimates cannot be. You may refer to your diagrams to aid in your explanation.

10





The ridge and LASSO estimates of β are found when the contours of $S(\beta)$ intersect the constraint regions, Since this can happen at an axis in LASSO, but not for ridge, this is why LASSO estimates of βs can be exactly 0.

Question 6 (11 points)

Smoking is known to be associated with lung cancer. To investigate whether smoking status is a good predictor of the onset of lung cancer 500 smokers and 500 non-smokers were observed prospectively to determine whether or not they developed lung cancer. The data are recorded using the following variables:

$$y_i = \begin{cases} 0 \text{ if person } i \text{ does not develop lung cancer} \\ 1 \text{ if person } i \text{ does develop lung cancer} \end{cases}$$

$$x_i = \begin{cases} 0 \text{ if person } i \text{ is a non smoker} \\ 1 \text{ if person } i \text{ is a smoker} \end{cases}$$

for $i = 1,2,\dots,1000$.

Using this data, a logistic regression is performed which relates $\pi_i = P(y_i = 1)$ to x_i via

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i \,.$$

Partial R output from this model is shown below.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9444     0.2052  -14.35   <2e-16 ***
x             3.1370     0.2240   14.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) (4) Interpret e^{β_0} and e^{β_1}.

- e^{β_0} is the odds ✓ that a person develops lung ✓ cancer when a person is a non-smoker. These data estimate this number to be $e^{-2.9444} = 0.053$

2

- e^{β_1} is the odds ratio ✓ that a person develops lung cancer if they were a smoker vs. non smoker. ✓ These data estimate this ratio to be $e^{3.1370} = 23.03$

2 In other words, a smoker is 23 times more likely to develop lung cancer than a non-smoker.

(b) (3) Construct a 95% confidence interval for β_1 and use it to test the hypothesis

$$H_0: \beta_1 = 0 \text{ vs. } H_A: \beta_1 \neq 0$$

Note that you may find it useful to consult the statistical tables on page 15.

95% CI for β_1 is:

$$\hat{\beta}_1 \pm 1.96\ SE(\hat{\beta}_1) \quad ✓$$

3

$$= 3.1370 \pm 1.96 \times 0.2240$$

$$= (2.698, 3.576) \quad ✓$$

Since 0 is not in this interval we reject $\beta_1 = 0$ at a 5% significance level. ✓

(c) (4) The efficacy of this fitted model was evaluated by performing in-sample classification. The results are summarized in the confusion matrix below.

		Truth	
		No Cancer	Cancer
Classification	No Cancer	475	25
	Cancer	225	275

i. (1) Calculate the True Positive rate.

$$TP = \frac{275}{25+275} = 0.92 \quad ✓$$

4

ii. (1) Calculate the False Positive rate.

$$FP = \frac{225}{225+475} = 0.32 \quad ✓$$

iii. (1) Calculate the True Negative rate.

$$TN = \frac{475}{475+225} = 0.68 \quad ✓$$

iv. (1) Calculate the False Negative rate.

$$FN = \frac{25}{25+275} = 0.08 \quad ✓$$

Useful Tables

Quantiles of $X \sim N(0,1)$

For the indicated value of p, the following table provides x^* where $P(X \geq x^*) = p$

p	x^*
0.005	2.576
0.01	2.326
0.025	1.960
0.05	1.645
0.1	1.282

Quantiles of $X \sim t_{(196)}$

For the indicated value of p, the following table provides x^* where $P(X \geq x^*) = p$

p	x^*
0.005	2.601
0.01	2.346
0.025	1.972
0.05	1.653
0.1	1.286

Quantiles of $X \sim F_{(3,196)}$

For the indicated value of p, the following table provides x^* where $P(X \geq x^*) = p$

p	x^*
0.005	4.411
0.01	3.883
0.025	3.183
0.05	2.651
0.1	2.112

LEFT BLANK