USF

MATH 372 Fall 2018

Final Exam

Wednesday December 12th, 2018

10:00am – 12:00pm

First Name: Solutions

Last name:

Instructions:

- Clearly write your name on this cover page.
- This exam consists of 17 pages including this cover page.
- If you need extra space, please use pages 14 and 15 labeled "LEFT BLANK" and INDICATE that you have done so.
- Page 16 contains tables of quantiles from the $N(0,1)$, $t_{(25)}$, $F_{(2,23)}$ and $X \sim \chi_{(4)}$ distributions.
- Page 17 contains Table 1 and Figure 1 for easy reference.
- You may remove pages 15-17 for your convenience.
- Where appropriate, round all numeric final answers to 2 decimal places.

Question	Points
Q1	/5
Q2	/50
Q3	/10
Q4	/15
Total	/80

USF

Question 1 [5 points]

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

(a) [T or F] Stepwise regression techniques (i.e., forward, backward, hybrid) never lead to the same set of selected predictors.

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

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

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

(e) [T or F] The null hypothesis for the test of overall significance of a regression model (with p explanatory variables) is $H_0: \beta_0 = \beta_1 = \cdots = \beta_p = 0$.

Question 2 [50 points]

Consumer Reports tested $n = 28$ different point-and-shoot digital cameras. Based upon factors such as the number of megapixels, weight, image quality and ease of use, they developed a quality score for each camera such that higher scores indicate better overall quality. Specifically, for each camera the following information was recorded:

- y = quality score
- x_1 = brand, where 0 indicates Canon and 1 indicates Nikon
- x_2 = price (in US dollars)
- x_3 = number of megapixels
- x_4 = weight (in ounces)

(a) Interest lies in building a model which relates a camera's quality score to these other factors. The information provided in Table 1 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. Using the information in this table, answer the following questions.

USF

i. [3] Perform *backward elimination* 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.

Drop x_4

Drop x_3

Stop.

Final Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

ii. [3] Perform *forward 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.

Add x_2

Add x_1

Stop

Final Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

iii. [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 NO (circle one)

iv. [5] Compare the full model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ to the "best overall" model from iii. using the additional sum of squares principle. Be sure to state the hypothesis being tested and the conclusion you draw at a 5% level of significance.

$H_0: \beta_3 = \beta_4 = 0$ vs. $H_A: \beta_j \neq 0$ for $j = 3$ or 4

$$F_0 = \frac{(SSE_{Red} - SSE_{Full})/\ell}{SSE_{Full}/(n-p-1)}$$

$$= \frac{(484.79 - 478.86)/2}{478.86/23}$$

$$= 0.14$$

Note $P(F_{(2,23)} \geq 3.422) = 0.05$ so we know p-value $= P(F_{(2,23)} \geq 0.14) > 0.05$ and so we do not reject H_0. Thus weight and megapixels do not significantly influence quality score

v. [3] Using appropriate sums of squares from Table 1, complete the following ANOVA table for the "best overall" model from iii.

Source	df	Sum of Sq.	Mean Sq.	F–Statistic
Regression	2	725.64	362.82	18.71
Error	25	484.79	19.39	
Total	27	1210.43		

USF

vi. [3] Calculate both R^2 and R^2_{adj} (i.e., adjusted-R^2) for the "best overall" model from iii. Describe the main advantage of using R^2_{adj} instead of R^2.

$$R^2 = \frac{SSR}{SST} = \frac{725.64}{1210.43} = 0.5995$$

$$R^2_{adj} = 1-(1-R^2)\left(\frac{n-1}{n-p-1}\right) = 1-(0.4005)\left(\frac{27}{25}\right) = 0.5675$$

R^2_{adj} does not become arbitrarily inflated simply adding extra predicters into a model

(b) [4] Consider the reduced model which contains only x_1 (brand) and x_2 (price). This model may be stated as

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

for $i = 1, 2, \ldots, 28$. 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:

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

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{28} \end{bmatrix} \quad \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad \vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{28} \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{1,28} & x_{1,28} \end{bmatrix}$$

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

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

USF

(d) [3] The vector $X^T y$ and the matrix $(X^T X)^{-1}$ are given below. Use these to show that $\hat{\beta}_0 = 49.01$, $\hat{\beta}_1 = -6.17$ and $\hat{\beta}_2 = 1.37$.

$$X^T y = \begin{bmatrix} 1578 \\ 813 \\ 286940 \end{bmatrix} \text{and } (X^T X)^{-1} = \begin{bmatrix} 0.24025 & -0.07427 & -0.00094 \\ -0.07427 & 0.14363 & -0.00002 \\ -0.00094 & -0.00002 & 0.00001 \end{bmatrix}$$

$$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y} = \begin{bmatrix} 0.24025 & -0.07427 & -0.00094 \\ -0.07427 & 0.14363 & -0.00002 \\ -0.00094 & -0.00002 & 0.00001 \end{bmatrix} \begin{bmatrix} 1578 \\ 813 \\ 28694 \end{bmatrix}$$

$$= \begin{bmatrix} 49.01 \\ -6.17 \\ 1.37 \end{bmatrix}$$

(e) [3] Interpret the values of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ computed in part (d).

- $\hat{\beta}_0 = 49.01$ is the expected quality score for a Canon camera that is free.
- $\hat{\beta}_1 = -6.17$ is the amount by which we expect quality score to change for a Nikon vs. a Canon camera. Thus Canons on average have a larger score (by 6.17 points) than Nikons, controlling for price.
- $\hat{\beta}_2 = 1.37$ is the amount by which we expect quality score to increase for every additional dollar increase in price, controlling for brand.

USF

(f) [4] Using the assumption that $y \sim MVN(X\boldsymbol{\beta}, \sigma^2 I)$, where I is the $n \times n$ identity matrix, derive the mean vector and the variance-covariance matrix for the least squares estimator $\widehat{\boldsymbol{\beta}}$. Specifically, show that

$$E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta} \text{ and } Var[\widehat{\boldsymbol{\beta}}] = \sigma^2(X^TX)^{-1}$$

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

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

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

$$= \vec{\beta}$$

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

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

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

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

USF

(g) [2] Using $(X^TX)^{-1}$ from part (d) and the fact that $\hat{\sigma} = 4.404$, show that the standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ are respectively, 1.6691 and 0.0139.

$$SE(\hat{\beta}_1) = 4.404 \times \sqrt{0.14363} = 1.6691$$

$$SE(\hat{\beta}_2) = 4.404 \times \sqrt{0.00001} = 0.0139$$

(h) [4] Calculate and interpret a 99% confidence interval for β_1.

$$\hat{\beta}_1 \pm t_{(25)}(0.995)\, SE(\hat{\beta}_1)$$

$$= -6.17 \pm 2.787 \times 1.6691$$

$$= (-10.82, -1.52)$$

Thus, we're 95% confident that the true difference between Canon and Nikon quality scores is with -10.82 and -1.52.

(i) [4] Test the following hypothesis at a 5% level of significance. Be sure to state your conclusion in the context of the data.

$$H_0: \beta_2 = 0 \text{ versus } H_A: \beta_2 \neq 0$$

$$t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{1.37}{0.0139} = 98.56$$

Note that $P(t_{(25)} \geq 2.060) = 0.025 \rightarrow 2P(t_{(25)} \geq 2.06) = 0$

Since $t = 98.56 > 2.060$ p-value $= 2P(t_{(25)} \geq 98.56) < 0.0$

Therefore we reject H_0 and conclude that a camera's price significantly influences its quality score

USF

(j) [1] Predict the quality score of a Canon camera that costs $150.00.

$\hat{y} = 49.01 - 6.17(0) + 1.37(150) = 254.51$

(k) A variety of diagnostic plots are provided in Figure 1. Refer to these plots in the following questions.

i. [1] Do the residuals appear to be normally distributed? State YES or NO and provide a one-sentence justification.

No – as evidenced by the QQ-plot and histogram, they are not bell-shaped and symmetric.

ii. [1] Do there appear to be any highly influential observations? State YES or NO and provide a one-sentence justification.

Yes – One observation has a Cook's-D value much bigger than 0.5 and all other values.

iii. [2] Calculate twice the average leverage, $2\bar{h}$. Do there appear to be any observations with high leverage? State YES or NO and provide a one-sentence justification.

$2\bar{h} = \frac{2(p+1)}{n} = \frac{2 \times 3}{28} = 0.21$

Yes – there is at least one observation with leverage greater than $2\bar{h}$.

iv. [1] Briefly (in 1-2 sentences) explain how you would determine whether the *constant variance* assumption in a linear regression is satisfied.

You could plot the residuals vs. fitted values and look for evidence of increasing / decreasing variation in the residuals as a function of fitted values.

USF

Question 3 [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:

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

subject to one of the following two shrinkage constraints.

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

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

(a) [5] Describe the relationship between $\widehat{\boldsymbol{\beta}}_{OLS}$, $\widehat{\boldsymbol{\beta}}_{LASSO}$ and $\widehat{\boldsymbol{\beta}}_{Ridge}$ as $c \to \infty$ and as $c \to 0$.

- As $c \to \infty$ the constraint region becomes large and $\hat{\vec{\beta}}_{LASSO} \to \hat{\vec{\beta}}_{OLS}$ and $\hat{\vec{\beta}}_{Ridge} \to \hat{\vec{\beta}}_{OLS}$
- As $c \to 0$ the constraint region becomes very small and the LASSO and ridge estimates are shrunken toward zero: $\hat{\vec{\beta}}_{LASSO} \to \vec{0}$ and $\hat{\vec{\beta}}_{Ridge} \to \vec{0}$

(b) [5] Explain why, in general, LASSO estimates can be 0, but ridge estimates cannot be.

The Ridge and LASSO estimates of $\vec{\beta}$ are found at the intersection of the contours of $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ and their respective constraint regions. Because the L_1 constraint region of LASSO is "pointy" this intersection can happen on an axis, whereas this will not happen for the "smooth" L_2 constraint region of Ridge.

USF

Question 4 [15 points]

A variety of different factors influence a person's annual salary. Here we consider the relationship between a person's salary and their age and level of education. In particular, we have information on the following variables for $n = 3000$ individuals.

- $y_i = \begin{cases} 1 & \text{if person } i \text{ earns more than \$100K per year} \\ 0 & \text{if person } i \text{ earns less than or exactly \$100K per year} \end{cases}$

- x_{i1} = age (in years) of person i

- A categorical variable with five levels that indicate the maximum level of education for person i (either "Some_HS", "HS_Diploma", "Some_College", "College_Diploma" or "Advanced_Degree"). For purposes of modeling, this education variable is represented by four indicator variables:

  - $x_{i2} = \begin{cases} 1 & \text{if person } i \text{ has "HS\_Diploma"} \\ 0 & \text{otherwise} \end{cases}$

  - $x_{i3} = \begin{cases} 1 & \text{if person } i \text{ has "Some\_College"} \\ 0 & \text{otherwise} \end{cases}$

  - $x_{i4} = \begin{cases} 1 & \text{if person } i \text{ has "College\_Diploma"} \\ 0 & \text{otherwise} \end{cases}$

  - $x_{i5} = \begin{cases} 1 & \text{if person } i \text{ has "Advanced\_Degree"} \\ 0 & \text{otherwise} \end{cases}$

  - The category "Some_HS" is the baseline category.

Using this data, a logistic regression is performed which relates $\pi_i = P(y_i = 1)$ to $\{x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}\}$ via

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}.$$

Partial R output from this model is shown below.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.862500   0.222568 -12.861  < 2e-16 ***
x1           0.037334   0.003677  10.153  < 2e-16 ***
x2           0.741927   0.164223   4.518 6.25e-06 ***
x3           1.589230   0.170714   9.309  < 2e-16 ***
x4           2.297414   0.173770  13.221  < 2e-16 ***
x5           3.143868   0.208860  15.053  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) [4] Interpret $e^{\hat{\beta}_1}$ and $e^{\hat{\beta}_4}$.

$e^{\hat{\beta}_1} = 1.04$ → Thus odds earning more than \$100K per year increases by a factor of 1.04 for additional year someone ages.

$e^{\hat{\beta}_4} = 9.95$ → Thus the odds of earning more than \$100K p- year increases by a factor of 9.95 when you have a college diploma vs. when you've only had "some" high school.

(b) [3] Using the estimates shown in the output above, calculate $\hat{\pi}_i$ for each of the three observations shown below, and use these values (and the threshold $c = 0.5$) to classify each person i as either making more than \$100K per year or not.

i	Age	Education
1	20	HS_Diploma
2	30	Some_College
3	30	College_Diploma

$$\hat{\pi}_1 = \frac{e^{\hat{\beta}_0 + 20\hat{\beta}_1 + \hat{\beta}_2}}{1 + e^{\hat{\beta}_0 + 20\hat{\beta}_1 + \hat{\beta}_2}} = 0.2$$ → Not making > \$100K /yea

$$\hat{\pi}_2 = \frac{e^{\hat{\beta}_0 + 30\hat{\beta}_1 + \hat{\beta}_3}}{1 + e^{\hat{\beta}_0 + 30\hat{\beta}_1 + \hat{\beta}_3}} = 0.46$$ → Not making > \$100K /year

$$\hat{\pi}_3 = \frac{e^{\hat{\beta}_0 + 30\hat{\beta}_1 + \hat{\beta}_4}}{1 + e^{\hat{\beta}_0 + 30\hat{\beta}_1 + \hat{\beta}_4}} = 0.64$$ → Making > \$100K /year.

USF

(c) The efficacy of this fitted model was evaluated by performing out-of-sample classification on a held-out test set. The results are summarized in the confusion matrix below.

		Truth	
		≤\$100K	>\$100K
Classification	≤\$100K	892	450
	>\$100K	451	1207

i. [1] Calculate the overall correct classification rate.

$$\frac{892+1207}{3000} = 0.70$$

ii. [1] Calculate the overall misclassification rate.

$$1 - 0.70 = 0.30$$

iii. [1] What percentage of people earning more than \$100K per year were correctly classified?

$$\frac{1207}{450+1207} = 0.73$$

iv. [1] What percentage of people earning less than or equal to \$100K per year were misclassified?

$$\frac{451}{892+451} = 0.34$$

(d) [4] Interest lies in fitting a reduced model which ignores the potential influence of education. To determine whether a person's education significantly influences whether they earn more than \$100K per year, use a likelihood ratio test to test the following hypothesis at a 1% level of significance.

$$H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \text{ versus } H_A: \beta_j \neq 0 \text{ for } j = 2,3,4 \text{ or } 5$$

Note that the maximized log-likelihood value for the full model is -1738.7 and the maximized log-likelihood value for the reduced model is -2001.8.

$$t = -2(\ell_{red} - \ell_{full})$$

$$= -2(-2001.8 + 1738.7)$$

$$= 526.2$$

Note $P(\chi^2_{(4)} \geq 13.277) = 0.01$ and so p-value $= P(\chi^2_{(4)} \geq 526.2) < 0.01$. Thus we reject H_0 and conclude education significantly influences one's earnings.

USF

Useful Tables of Quantiles

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 $X \sim t_{(25)}$
For the indicated value of p, the following table provides x^* where $P(X \geq x^*) = p$

p	x^*
0.005	2.787
0.01	2.485
0.025	2.060
0.05	1.708
0.1	1.316

Quantiles of $X \sim F_{(2,23)}$
For the indicated value of p, the following table provides x^* where $P(X \geq x^*) = p$

p	x^*
0.005	6.730
0.01	5.664
0.025	4.349
0.05	3.422
0.1	2.549

Quantiles of $X \sim \chi_{(4)}$
For the indicated value of p, the following table provides x^* where $P(X \geq x^*) = p$

p	x^*
0.005	14.860
0.01	13.277
0.025	11.143
0.05	9.488
0.1	7.779

Variables in Model	AIC	SSE
None (intercept only)	188.92	1210.43
x_1	187.21	1060.09
x_2	173.32	645.43
x_3	190.92	1210.36
x_4	188.54	1111.64
x_1 x_2	167.30	484.79
x_1 x_3	188.33	1027.20
x_1 x_4	187.10	983.00
x_2 x_3	174.75	632.43
x_2 x_4	175.18	642.33
x_3 x_4	190.46	1108.60
x_1 x_2 x_3	169.01	479.73
x_1 x_2 x_4	169.30	484.66
x_1 x_3 x_4	187.50	938.58
x_2 x_3 x_4	176.72	631.74
x_1 x_2 x_3 x_4	170.96	478.86

Table 1: AIC and SSEs for various models

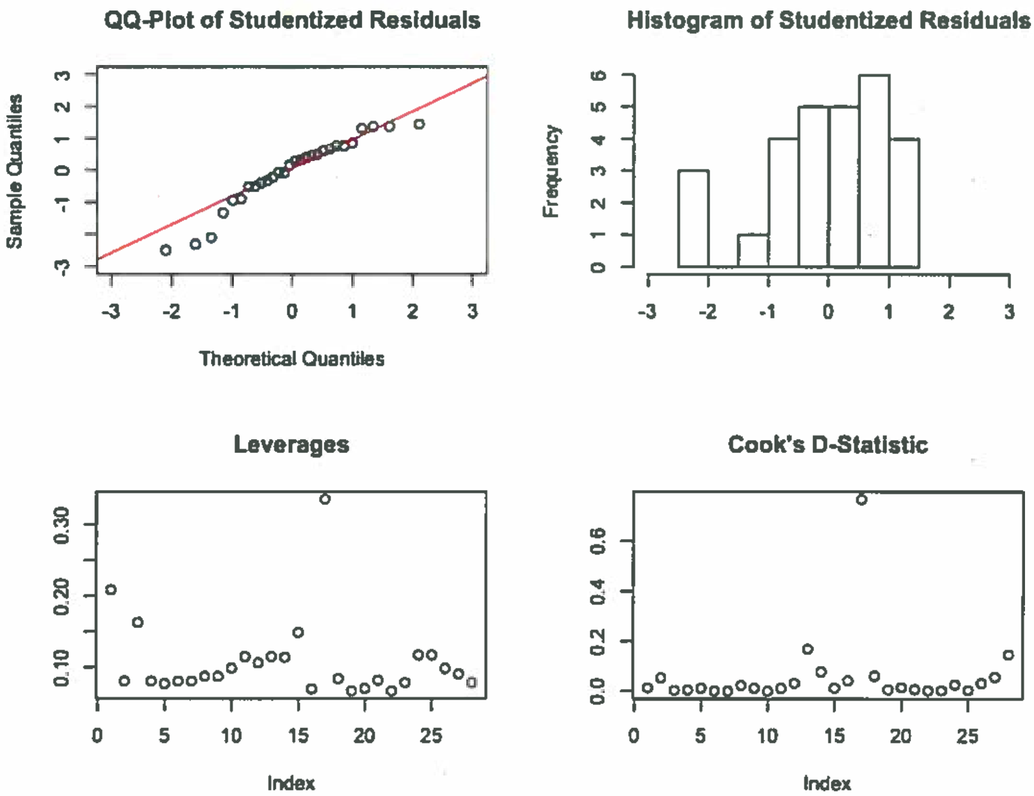

Figure 1: Diagnostic Plots