STATISTICS 608 Linear Models -EXAM II
March 27, 2013

Student's Name: _____

Student's Email Address: _____

INSTRUCTIONS FOR STUDENTS:

1. There are **10** pages including this cover page.
2. You have exactly 50 minutes to complete the exam.
3. There may be more than one correct answer; choose the best answer.
4. You will not be penalized for submitting too much detail in your answers, but you may be penalized for not providing enough detail.
5. Do not discuss or provide any information to any one concerning any of the questions on this exam or your solutions until I post the solutions next week.
6. You may use one 8.5" X 11" sheet of notes and a calculator.
7. At the end of the exam, leave your sheet of notes with your proctor along with the exam.

I attest that I spent no more than 50 minutes to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature: _____

INSTRUCTIONS FOR PROCTOR:

Immediately after the student completes the exam scan it to a pdf file and have student upload to Webassign.

1. I certify that the time at which the student started the exam was _____ and the time at which the student completed the exam was _____.
2. I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
3. I certify that the exam was scanned in to a pdf and uploaded to Webassign in my presence.
4. I certify that the student has left the exam and sheet of notes with me, to be returned to the student no less than one week after the exam or shredded.

Proctor's Signature: _____

Part I: Multiple choice

1. A model with five variables $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + e$ was fit to a data set. Researchers were interested in selecting a model with fewer variables. Would you suggest using all possible subsets, or a selection procedure using stepwise? Why?

   (a) All possible subsets, because all combinations of the five variables are considered.

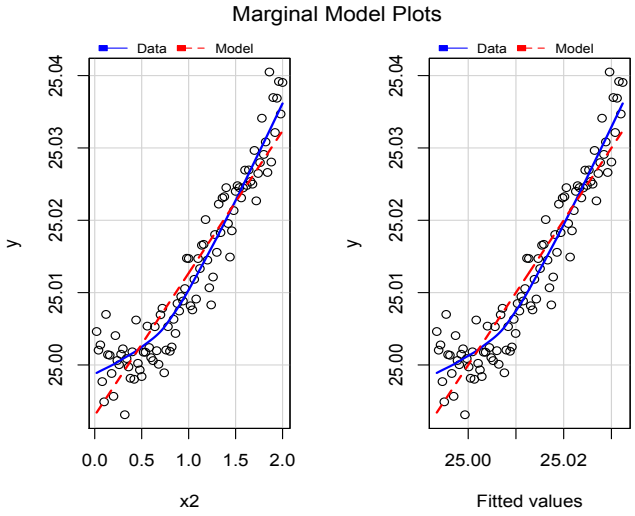
   (b) All possible subsets, because it considers fewer models.

   (c) Stepwise, because variables that definitely explain some variation in the response stay in the model.

   (d) Stepwise, because variables are allowed to both enter and leave the model.

   (e) Stepwise, because it usually chooses fewer variables, resulting in more power.

2. A linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$, with all three variables quantitative (numeric), has been fit using several predictor variable, and the following marginal model plots were observed for variable x_2 and the fitted values. What do they tell you about the appropriateness of the linear model?



   (a) Because the lines on the plots are not straight, the model is not valid.

   (b) Because the solid and dotted lines do not match each other, the model is not valid.

   (c) Because the plot for variable x_2 has increasing slope, it explains additional variation in y after x_1 has already been added to the model.

   (d) Because the plot for variable x_2 has increasing slope, the model is valid.

   (e) Because the solid and dotted lines do not match each other, variable x_2 explains additional variation in y after x_1 has already been added to the model.

3. A researcher has two models that are being compared (assume n is large enough):

$$\text{Model 1: } \mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \mathbf{e}_{n\times 1}, \text{ with } R^2 = 0.8$$
$$\text{Model 2: } \log(\mathbf{Y}_{n\times 1}) = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \mathbf{e}_{n\times 1}, \text{ with } R^2 = 0.7$$

The researcher states that because R^2 is greater for Model 1 than Model 2, Model 1 is preferable. Do you agree? Why or why not?

(a) Yes, because the proportion of variability in Y explained is greater for Model 1.

(b) Yes, because Model 1 is more linear than Model 2.

(c) No, because the scales of the response variables are different.

(d) No, because R^2 can never be used to compare two models.

4. A researcher has fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$, and notices that the correlation between x_1 and x_2 is 0.1. What does that tell us about the variance of the parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$?

(a) They will be large; the variance inflation factors will be high.

(b) They will be large; the variance inflation factors will be low.

(c) They will be small; the variance inflation factors will be high.

(d) They will be small; the variance inflation factors will be low.

5. Suppose we ran a backward selection procedure using p-values and the criterion that p-values had to be less than 0.07 to stay in the model. If we began with 100 variables and a sample size of $n = 1000$ observations, how many variables would we expect to be in the final model, on average, if in reality $\beta_1 = \beta_2 = \ldots \beta_{100} = 0$?

(a) 0

(b) 7

(c) 10

(d) 70

(e) 100

Part II: Short Answer

6. Two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ are orthogonal if $\mathbf{AB} = \mathbf{BA} = \mathbf{0}$. We have shown that where $\mathbf{H}$ is the projection matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{H}$ and $(\mathbf{I} - \mathbf{H})$ are orthogonal. Use this result to show that the residual vector $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$ and the fitted vector $\hat{\mathbf{y}} = \mathbf{Hy}$ are orthogonal.

7. A researcher wants to use the statistic $y_1 + y_2 - 3y_3$ to estimate the parameter α in the linear model $y_i = \alpha + e_i$, $i = 1, 2, 3$ with errors independent and identically distributed with mean zero. Is this statistic the Best Linear Unbiased Estimator? Why or why not?

Part III: Long Answer

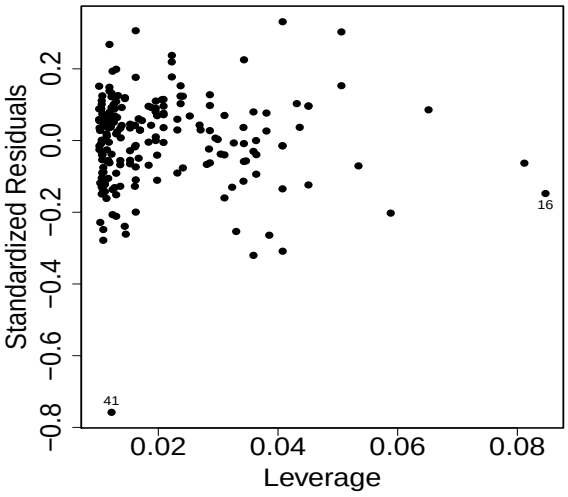
8. Graduation: Kiplinger's Magazine released data on admissions and graduation rates (both expressed as proportions) for public and private colleges. The scatterplot below on the right shows the data; private universities are shown as a circle on the scatterplot, and public universities are shown as a square. Private universities tend to admit fewer applicants and graduate more, while public universities tend to admit more applicants and graduate fewer. The model fit to the data was:

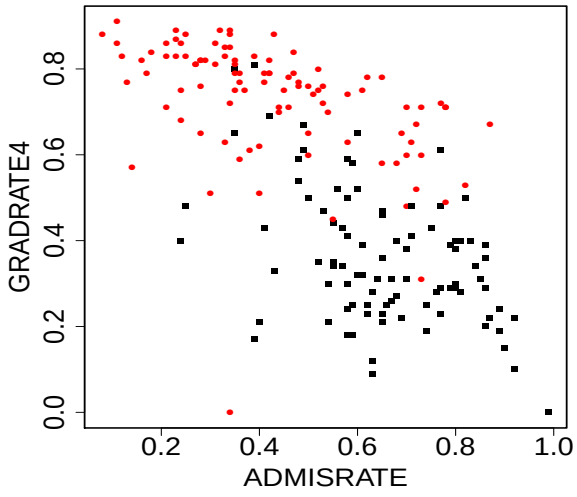
$$GraduationRate = \beta_0 + \beta_1 AdmissionsRate + \beta_2 iPrivate + e,$$

where iPrivate was an indicator variable taking the value 1 for private colleges and the value 0 for public colleges. Output for the model is found at the end of the exam.

(a) Interpret the parameter estimate $\hat{\beta}_2$ for the model in the context of the problem.

(b) Before the model above was fit, two problematic points were identified and corrected, as shown on the plot below on the left of Leverage vs. Standardized Residuals as points 41 (in the bottom left corner) and 16 (on the far right).





Circle your best guess for the locataions of those two points on the scatterplot of Admissions Rate vs. Graduation Rate on the right. Label clearly which point is which. Explain why you identified those as points 41 and 16.

9. Marketing: A coroporation is interested in predicting sales based on money spent on advertising, the number of accounts held, and several other variables listed below.

   (a) The first model including all variables was:

$$Sales = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_8 x_8 + e,$$

   where the variables were as follows:
   x_1 = (TIME) length of time salesperson has been with the company
   x_2 = (POTENT) industry sales in units for the area
   x_3 = (ADV) dollar expenditures on advertising
   x_4 = (SHARE) market share
   x_5 = (SHARECHG) change in market share from last year
   x_6 = (ACCTS) total number of accounts assigned to salesperson
   x_7 = (WORKLOAD) average workload per account
   x_8 = (RATING) rating of performance on a 1 - 7 scale

   Are there problems with multicollinearity in this model? Explain why or why not.

   (b) A manager suggests that only variables POTENT, ADV, SHARE, and ACCTS should be important for predicting sales. Conduct an F-test of model reduction, assuming appropriate assumptions have been met. Be sure to state your hypotheses, give a test statistic, and write a conclusion in the context of the problem. The p-value for the test is less than 0.01.

Graduation Rate:

```
> summary(lm1)

Call:
lm(formula = GRADRATE4 ~ ADMISRATE + PRIVATE)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31343 -0.06220  0.01155  0.07635  0.35060

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.60583    0.03369  17.984  < 2e-16 ***
ADMISRATE   -0.37546    0.04810  -7.806 3.86e-13 ***
PRIVATE      0.29169    0.02004  14.559  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1165 on 190 degrees of freedom
Multiple R-squared: 0.7487,     Adjusted R-squared: 0.7461
F-statistic:   283 on 2 and 190 DF,  p-value: < 2.2e-16
```

Marketing – Full Model:

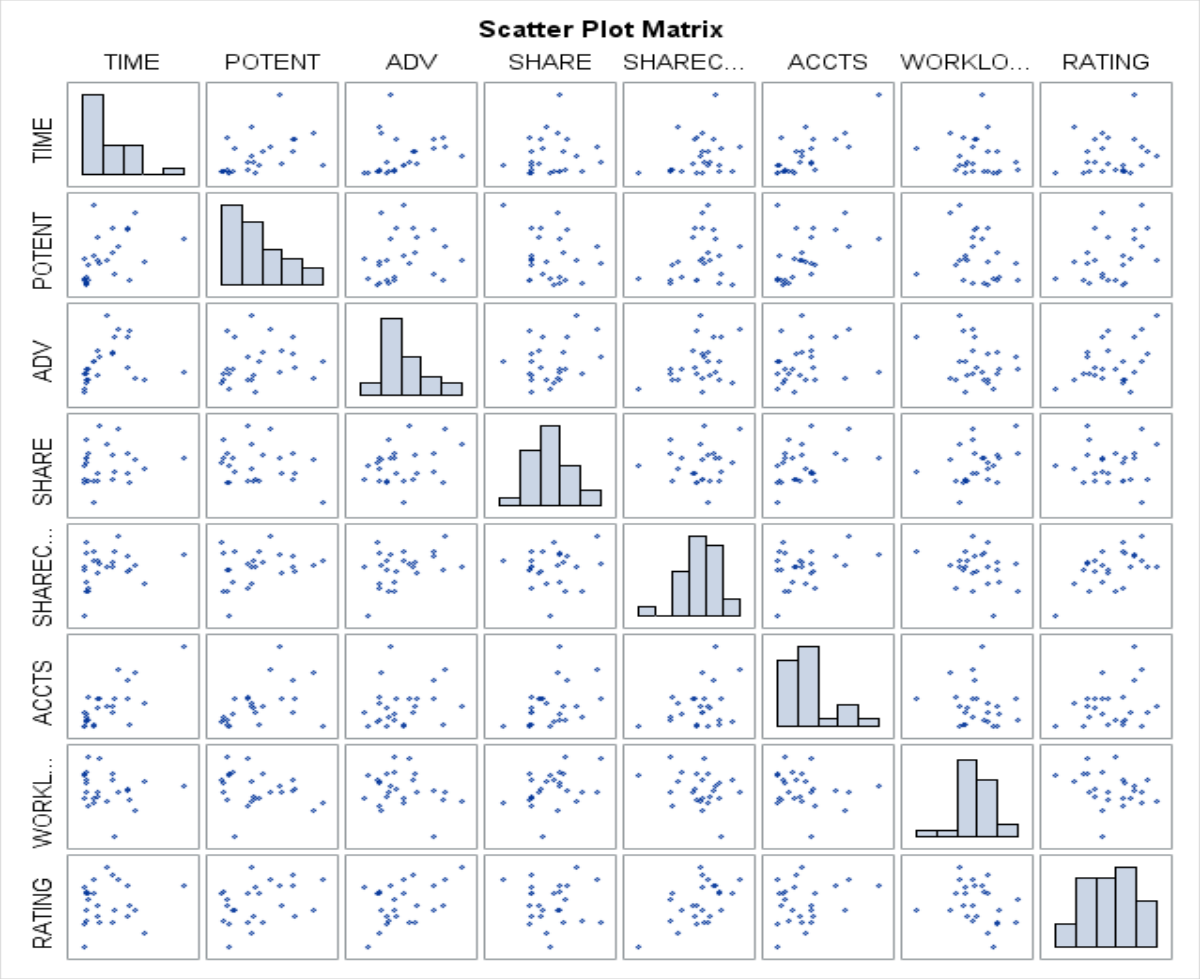
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	38153558	4769195	23.65	<.0001
Error	16	3225991	201624		
Corrected Total	24	41379549			

Root MSE	449.02607	**R-Square**	0.9220
Dependent Mean	3374.56760	**Adj R-Sq**	0.8831
Coeff Var	13.30618		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	**1**	-1507.83390	778.63693	-1.94	0.0707	0
TIME	**1**	2.00952	1.93066	1.04	0.3134	3.34263
POTENT	**1**	0.03720	0.00820	4.54	0.0003	1.97763
ADV	**1**	0.15099	0.04711	3.21	0.0055	1.91022
SHARE	**1**	199.02263	67.02809	2.97	0.0090	3.23577
SHARECHG	**1**	290.85260	186.78224	1.56	0.1390	1.60173
ACCTS	**1**	5.55111	4.77557	1.16	0.2621	5.63936
WORKLOAD	**1**	19.79446	33.67678	0.59	0.5649	1.81835
RATING	**1**	8.19043	128.50560	0.06	0.9500	1.80855


Scatter Plot Matrix
TIME
POTENT
ADV
SHARE
SHAREC...
ACCTS
WORKLO...
RATING
TIME
POTENT
ADV
SHARE
SHAREC...
ACCTS
WORKL...
RATING

Marketing – Reduced Model:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	37260212	9315053	45.23	<.0001
Error	20	4119337	205967		
Corrected Total	24	41379549			

Root MSE	453.83573	**R-Square**	0.9004
Dependent Mean	3374.56760	**Adj R-Sq**	0.8805
Coeff Var	13.44871		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	**1**	-1441.93759	423.58122	-3.40	0.0028	0
POTENT	**1**	0.03822	0.00798	4.79	0.0001	1.83101
ADV	**1**	0.17499	0.03691	4.74	0.0001	1.14772
SHARE	**1**	190.14437	49.74410	3.82	0.0011	1.74459
ACCTS	**1**	9.21396	2.86521	3.22	0.0043	1.98718