

# STATISTICS 608 Linear Models -EXAM II

## March 27, 2013

Student's Name: \_\_\_\_\_

Student's Email Address: \_\_\_\_\_

### INSTRUCTIONS FOR STUDENTS:

1. There are **13** 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:

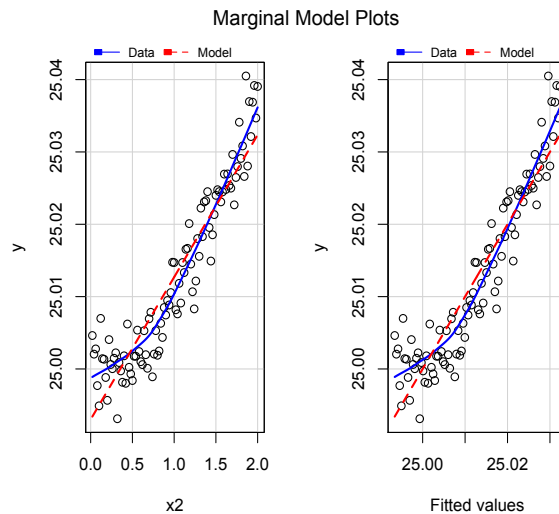
**Immediately** after the student completes the exam scan it to a pdf file and have student upload to Webassign.

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  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + e$  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  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + e$ , 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 solid 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 mean function has not been appropriately specified.
- (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 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):

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

Model 2:  $\log(\mathbf{Y}_{n \times 1}) = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \mathbf{e}_{n \times 1}$ , 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) It will be large; the variance inflation factor will be high.
  - (b) It will be large; the variance inflation factor will be low.
  - (c) It will be small; the variance inflation factor will be high.
  - (d) It will be small; the variance inflation factor will be low.  $VIF = 1/(1 - 1/100) = 1.0101 \approx 1$
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  $\beta_1 = \beta_2 = \dots \beta_{100} = 0$ ?
- (a) 0
  - (b)  $7 = 0.07(100)$
  - (c) 10
  - (d) 70
  - (e) 100

## Part II: Short Answer

6. Two  $n \times n$  matrices  $\mathbf{A}$  and  $\mathbf{B}$  are orthogonal if  $\mathbf{AB} = \mathbf{BA} = \mathbf{0}$ . Show 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.

$$\hat{\mathbf{y}}'\hat{\mathbf{e}} = (\mathbf{Hy})'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{y}'\mathbf{H}'(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}'\mathbf{0}\mathbf{y} = 0$$

7. A researcher wants to use the statistic  $y_1 + y_2 - 3y_3$  to estimate the parameter  $\alpha$  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?

No; it's not unbiased.

$$\begin{aligned} E[y_1 + y_2 - 3y_3] &= E[\alpha + e_1 + \alpha + e_2 - 3(\alpha + e_3)] \\ &= E[-\alpha + (e_1 + e_2 - 3e_3)] \\ &= -\alpha + 0(\text{errors mean } 0) \\ &\neq \alpha \end{aligned}$$

So it can't be an unbiased estimator if it is biased.

### 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:

$$\text{GraduationRate} = \beta_0 + \beta_1 \text{AdmissionsRate} + \beta_2 i\text{Private} + e,$$

where  $i\text{Private}$  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.

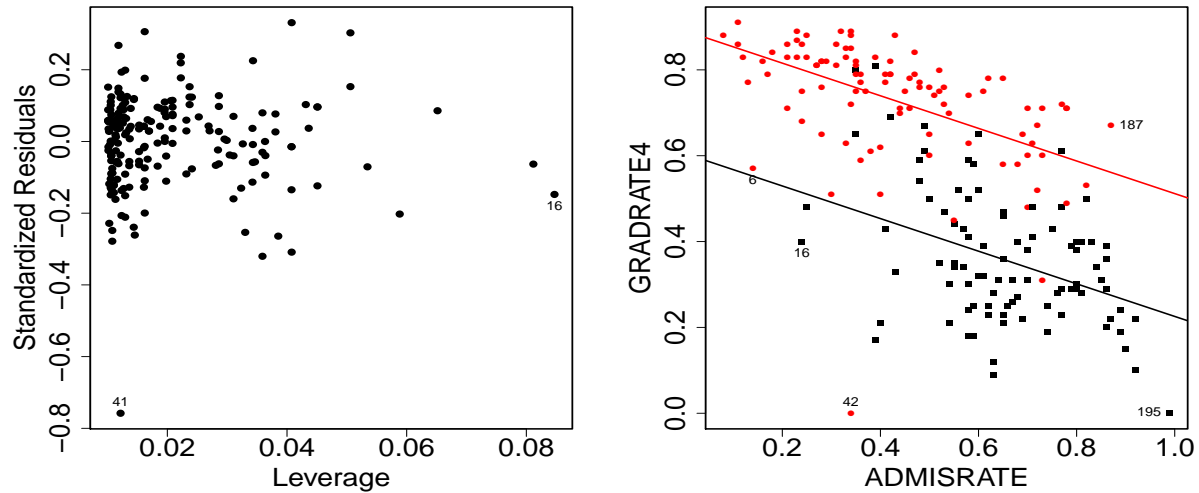
- (a) Interpret the parameter estimate  $\hat{\beta}_2$  for the model in the context of the problem.

For a particular admissions rate (e.g. 0%), private schools' graduation rate is estimated to be 29 percentage points higher than public schools, on average.

Notes: First, the graduation rate is not 29 percent higher for private schools. Remember that if we go from 50% to 75%, that's an increase of 25 percentage points, but an increase of 50%. An increase from 2% to 3% is an increase of 1 percentage point, but an increase of 50%. I know that's probably a piece of math you've forgotten; I didn't count off for this.

Second, this is not the difference in the two mean graduation rates; that number is much higher for this model, as private schools admit far fewer students on average. We have to say something about "for a particular admissions rate." Many students opted to say "when the admissions rate is 0%," which is acceptable.

- (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 locations 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.

Point 16 has the highest leverage because it is farthest from the average admissions rate in the x-direction. It has a small standardized residual, so it is very close to the model. Point 41 has the highest negative residual, so it is farthest below the model; however, its low leverage value means it must be close to the average admissions rate in the x-direction.

Many students chose the point in the bottom right hand corner; to be honest, that's what I would have chosen. It has a small standardized residual as well (it is close to the model), and it has pretty high leverage, being far from the other values in the x-direction. Apparently the point marked 16 on this plot is slightly farther away from the average admissions rate for public schools. Its residual is also somewhat smaller, something you don't pick up on without very carefully graphing the two regression lines.

9. A corporation 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 the following.

$$Sales = \beta_0 + \beta_1 x_1 + \dots + \beta_8 x_8 + e, \quad (1)$$

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.

Yes; one of the VIF's is  $5.6 > 5$ . You can also see some relationships between some of the variables, but most of them weren't strong. But remember that even if individual correlations are weak, it is still possible that some linear combination of the other predictors explains a predictor, so we have to check the VIF's.

- (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.

$H_0 : \beta_1 = \beta_5 = \beta_7 = \beta_8 = 0$  vs.  $H_a : \text{At least one of } \beta_1, \beta_5, \beta_7, \beta_8 \text{ not equal to } 0$ .

$$\begin{aligned} F &= \frac{(RSS(reduced) - RSS(full)) / (df_{reduced} - df_{full})}{RSS(full) / df_{full}} \\ &= \frac{(4119337 - 3225991) / (20 - 16)}{3225991 / 16} = 1.1077 \end{aligned}$$

If the p-value is really less than 0.01 (it should have been 0.39), we should reject the null hypothesis and conclude that variables TIME, SHARECHG, WORKLOAD, and RATING should not be dropped from the model; they explain variation in sales that the other variables don't explain.

Note: Some students wrote out a hypothesis using statistics. It's fine to write out hypotheses in words; mathematical symbols are a simple shortcut for ideas that take longer to explain in English. But remember that we form hypotheses about parameters rather than statistics because we want to conclude something about a population. We don't need to hypothesize anything about the data at hand; we know those numbers.

Also, some students had a little trouble with remembering that when the p-value is low we should be rejecting the null hypothesis, meaning we have evidence in favor of the alternative. When the p-value is high, we fail to reject the null hypothesis, meaning we don't have evidence that the alternative is false.

Finally, remember to state your conclusion in context. Without this final step, someone who is not a statistician has no idea which model to use. The point of conducting a hypothesis test is to make a conclusion; explain clearly what conclusion was made.



## 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

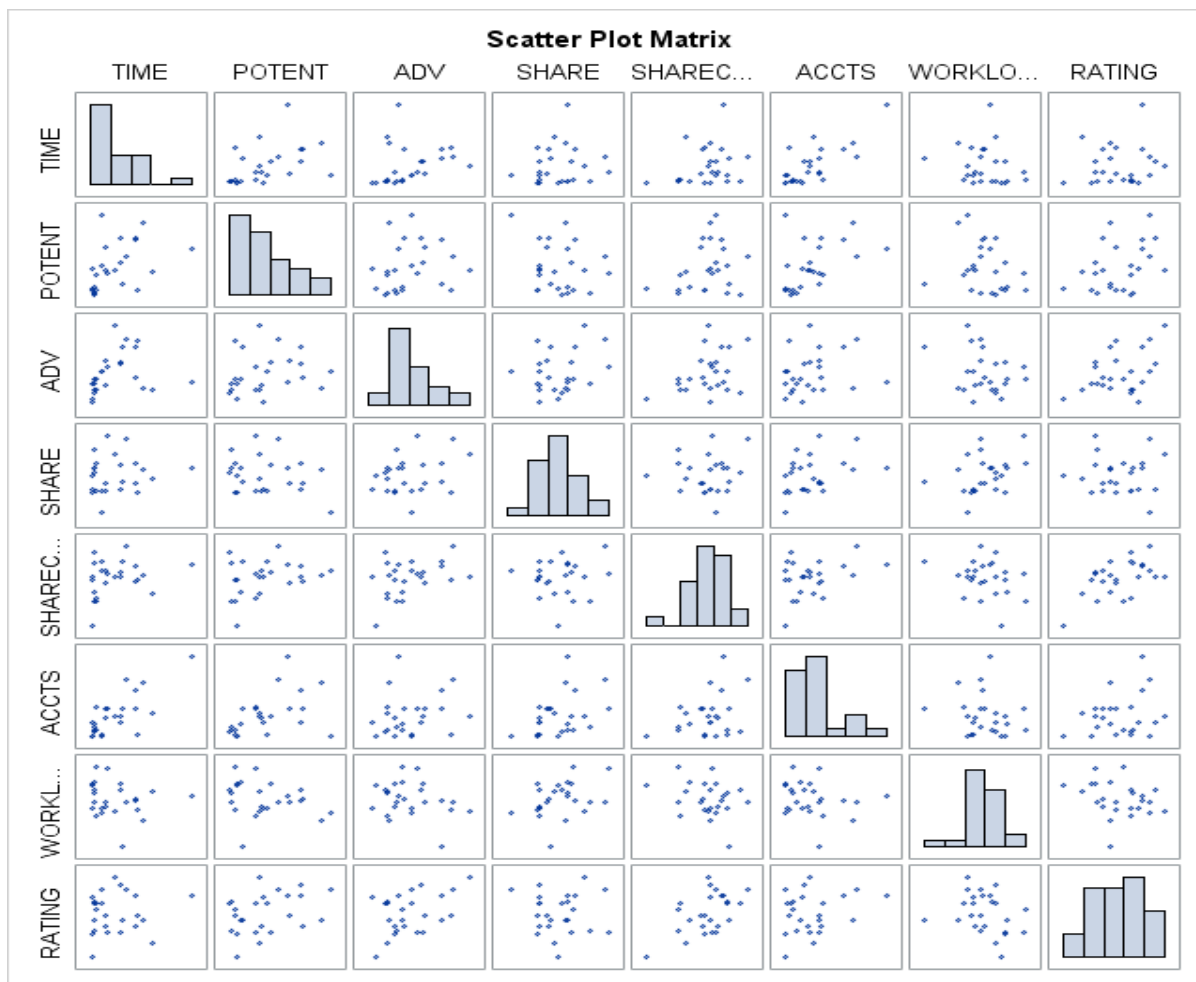
F-statistic: 283 on 2 and 190 DF, p-value: < 2.2e-16

Marketing – Full Model:

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

<b>Root MSE</b>	449.02607	<b>R-Square</b>	0.9220
<b>Dependent Mean</b>	3374.56760	<b>Adj R-Sq</b>	0.8831
<b>Coeff Var</b>	13.30618		

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



Marketing – Reduced Model:

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

<b>Root MSE</b>	453.83573	<b>R-Square</b>	0.9004
<b>Dependent Mean</b>	3374.56760	<b>Adj R-Sq</b>	0.8805
<b>Coeff Var</b>	13.44871		

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