STATISTICS 608 Linear Models - Final Exam
May 8, 2013

Student's Name: _____

Student's Email Address: _____

INSTRUCTIONS FOR STUDENTS:

1. There are **12** pages including this cover page.
2. You have exactly 2 hours 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 2 hours 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:

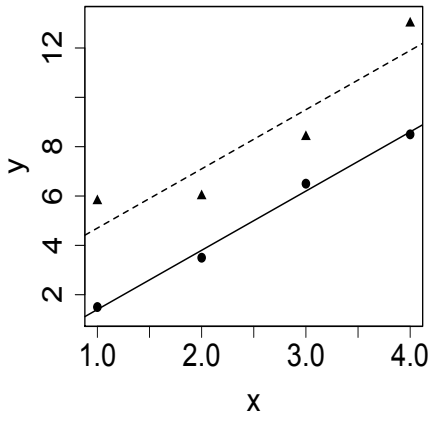
Immediately after the student completes the exam scan it to a pdf file and have student upload to Webassign.

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 (5 points each)

1. Two linear models were fit to two data sets, shown on the plot below (drawn to scale). Model 1, the dashed line fit to the triangular points, and Model 2, the solid line fit to the circle points, had the same slope, and were fit to data sets with the same mean for x, 2.5. The total sum of squares for the first model was greater. Which of the following statements is true?



   (a) RSS (residual sum of squares) for Model 1 > RSS for Model 2

   (b) RSS for Model 1 < RSS for Model 2

   (c) RSS for Model 1 = RSS for Model 2

   (d) The answer cannot be determined for this data set.

2. In a logistic regression model $\log(\theta(x_i)/(1-\theta(x_i)) = \beta_0 + \beta_1 x_i$ with multiple observations at each value of x_i (that is, $m_i > 1$), what does deviance measure?

   (a) The difference between the logistic regression model and another model with as many parameters as values of x_i. (This is a description of the saturated model.)

   (b) The difference between the logistic regression model and a straight line regression model.

   (c) The difference between the logistic regression model and a horizontal line fitted to the data at the y-value equal to the proportion of successes in the entire data set.

   (d) The difference between the logistic regression model and another logistic model with only one parameter, the proportion of successes in the data set.

3. Suppose a plot of standardized residuals against x_1 for a model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$ has a parabolic shape, where variables x_1 and x_2 are quantitative (numeric). What does that tell you about the model?

   (a) The residuals will be randomly scattered after adding the term x_1^2 to the model.

   (b) Predictors x_1 and x_2 are not independent of one another.

   (c) Predictors x_1 and x_2 do not have a linear association.

   (d) The errors from the model are autocorrelated.

   (e) Any of the above are possible.

4. A model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 x_2$ was fit to a data set, where x_1 and y were continuous variables, and x_2 was an indicator (or dummy) variable. If all parameters are non-zero, what is the geometric interpretation of the model?

   (a) A single regression line

   (b) Parallel lines

   (c) One y-intercept with separate slopes

   (d) Separate slopes and separate y-intercepts

   (e) None of the above

5. A researcher notices that there are 3 insignificant predictor variables in a valid multiple linear regression model with 7 predictors total. Why would you recommend that the researcher not drop all three predictors from the model at once?

   (a) Because some of the predictors may be correlated.

   (b) Because some of the predictors may need to be transformed first.

   (c) Because leverage points may need to be removed from the model instead.

   (d) Because the variances of the residuals may not be constant.

6. A straight line was fit by least squares to 50 pairs of points using the standard model $y = \beta_0 + \beta_1 x + e$. Assume that the usual least squares regression assumptions are met. Statistical output showed that the sample variance, s_y^2, of the 50 response values was equal to 100, and that the estimate of the error variance, $MSE = RSS/(n-2)$ was equal to 10. The two estimates s_y^2 and MSE are so very different because:

   (a) s_y^2 was calculated incorrectly.

   (b) MSE was calculated incorrectly.

   (c) s_y^2 is not an estimate of the variance of the residuals.

   (d) $s_y^2 = 100$ implies that there is a great deal of statistical variation, so the discrepancy is to be expected.

   (e) A weighted least squares model should be fit.

7. For least squares regression models which have autocorrelated errors according to the AR(1) model, we studied a method to transform the data such that the transformed errors were not correlated. What is the main benefit of conducting such a transformation?

   (a) To diagnose whether the model is valid using residual diagnostic plots.

   (b) To enable simpler interpretations of the slope and intercept parameter estimates.

   (c) To correctly calculate p-values for parameters (since generalized least squares methods didn't calculate p-values correctly)

   (d) To control for lurking variables in the model.

8. We considered transformations of predictor variables in multiple linear regression models to normality using methods like the Box-Cox transformation. Which of the following is the most important reason to transform predictor variables to a normal distribution?

   (a) To ensure the residuals are normally distributed, one of the assumptions for regression.

   (b) To ensure the predictor variables are independent of one another.

   (c) To ensure the predictor variables are independent of the response variable.

   (d) To ensure a linear relationship between the predictors and the response variable.

Part II: Short Answer

9. Consider the regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where the n errors in the error vector $\mathbf{e}$ are independent and identically distributed. That is, $\text{Var}(\mathbf{e}) = \sigma^2\mathbf{I}$ where $\mathbf{I}$ is the identity matrix and σ^2 is a constant. Calculate the variance matrix of the least squares estimate of $\boldsymbol{\beta}$, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. (7 points)

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= \text{Var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

10. Consider regression through the origin (i.e. straight line regression with population intercept known to be zero) with $\text{Var}(e_i|x_i) = x_i\sigma^2$. The corresponding regression model is $y_i = \beta x_i + e_i$, $(i = 1 \ldots n)$. Show that the explicit expression for the weighted least squares estimate of β is $\bar{y}/\bar{x}$. (7 points)

We are looking for a weight such that $\text{Var}(e_i|x_i)$ is constant across the values of x. More specifically, we need $\text{Var}(\sqrt{w_i}e_i|x_i)$ to be a constant. Since we know $\text{Var}(e_i|x_i) = x_i\sigma^2$, take $w_i = 1/x_i$:

$$\text{Var}\left(\frac{1}{\sqrt{x_i}}e_i|x_i\right) = \frac{1}{x_i}\text{Var}(e_i|x_i) = \frac{1}{x_i}x_i\sigma^2 = \sigma^2$$

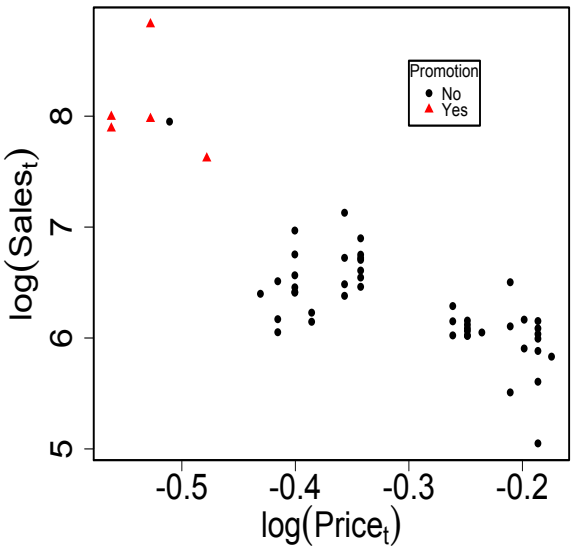
One method to find the weighted least squares estimator is to use $\hat{\beta} = (X'WX)^{-1}(X'WY)$. First, it is critical to notice that this regression is through the origin, so that the design matrix is $[x_1\, x_2 \ldots x_n]'$. Some students tried to pretend that $\hat{\beta}_0 = 0$, which is not what we're doing. The weight matrix is as follows:

$$W = \begin{bmatrix} 1/x_1 & 0 & 0 & \ldots & 0 \\ 0 & 1/x_2 & 0 & \ldots & 0 \\ 0 & 0 & 1/x_3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1/x_n \end{bmatrix}$$

And then we have:

$$
\begin{aligned}
X'W &= \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \\
X'WX &= \sum x_i \\
X'WY &= \sum y_i \\
\hat{\beta} &= (\sum x_i)^{-1} \sum y_i = \frac{\sum y_i/n}{\sum x_i/n} = \frac{\bar{y}}{\bar{x}}
\end{aligned}
$$

11. A large grocery store chain is considering weekly sales of Brand 1 over the course of a year, for a total of $n = 52$ records from the year's sales. A plot of the log of sales by the log of price of Brand 1 can be seen below. Also denoted on the plot is whether a promotion like an advertisement occurred that week.



As a preliminary model, a statistician decided to fit the following:

$$\log(Sales_t) = \beta_0 + \beta_1 \log(Price_t) + e_t, t = 1, \ldots, n$$

assuming that the errors were independent and normallly distributed with constant variance. The t-value found was -10.10. This t-value is inappropriate for this scenario. Explain the most important reason why not, as if to someone with very little statistical experience. (6 points)

The answer I was looking for was that weekly sales over 52 weeks will almost certainly be autocorrelated, meaning the variance of the estimate of β_1 and thus the denominator of this t-value will be incorrectly calculated.

However, lots of people mentioned that we may have problems with the group of weeks when there was a promotion may need a separate line fit to it, so we may want to add a dummy variable and an interaction between that dummy variable and $\log(Price)$ to fit a separate line. While potentially important, the promotion observations aren't far from the general pattern of the rest of the points - they seem to fit reasonably close to the regression line that would pass through the rest of the points.

Part III: Long Answer

12. As cheddar cheese matures, a variety of chemical processes take place. In one study of cheddar chesse from Victoria, Australia, cheese samples were chemically analyzed and tasted by professional taste testers. We consider predicting the average taste score from several tasters using three potential predictors:

$$x_1 = \text{natural log hydrogen sulfide concentration}$$
$$x_2 = \text{lactic acid concentration}$$
$$x_3 = \text{natural log acetic acid concentration}$$

The first (full) model we consider involves all three predictor variables:

$$taste = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

The second (reduced) model uses only hydrogen sulfide and lactic acid:

$$taste = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

Output from both models can be found in the appendix.

(a) Show that the F-statistic for reducing the model from Model 1 to Model 2 is equal to 0.0057. (6 points)

$$F = \frac{(2669 - 2668.41)/1}{2668.41/26}$$

(b) As can be seen from the output in the appendix, the first model using all three predictors has a **lower** adjusted R-squared than the model using only the first two. Explain why, as if to someone with very little statistical knowledge. (7 points)

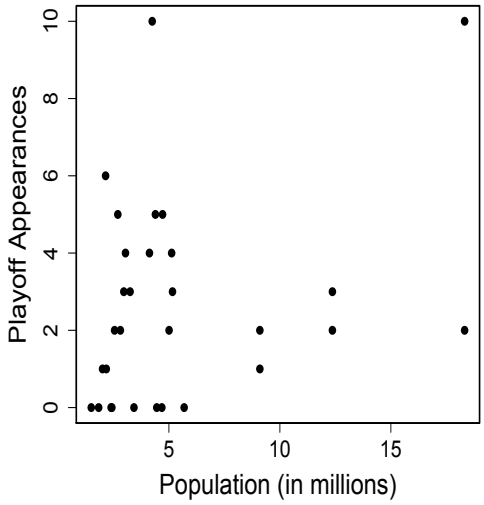
We are indeed surprised that a model with more predictors would have a lower R^2 value, as residual sum of squares are either reduced or (in very lucky cases) maintained by adding a predictor variable. The difference between R^2 and R^2_{adj}, however, is a correction factor for adding predictors that don't help explain a large amount of residual variability in the model, after the other variables are included. We saw from our very small F-value in part (a) that adding the variable x_3 doesn't add a significant amount of explanatory power to our model (p-value = 0.94198 from output for corresponding t-test), so perhaps we shouldn't be surprised that since the reduction in residual sums of squares is so tiny, after taking into account this penalty for adding an extra variable, our adjusted R-squared value is lower for the model using all three predictors than for the one using only two.

(c) The table below shows values of R^2_{adj}, AIC_C, and BIC for the best models with one, two, and three predictors. Which model(s) do the different criteria select? (6 points)

Subset Size	Predictors	R^2_{adj}	AIC_C	BIC
1	x_1	0.5558	145.81	147.69
2	x_2x_3	0.6259	141.57	144.85
3	$x_1x_2x_3$	0.6116	143.56	148.25

All three criteria choose the model with two predictors: the R^2 value is highest and AICC and BIC are lowest for that model.

13. A book on baseball[1] used regression analysis to conclude that "it is hard to find much correlation between market size and... success in making the playoffs. The relationship... is quite weak." The data from the 30 Major League Baseball teams from the 10 seasons between 1995 and 2004 is plotted below, followed by output implied by the author's comments. The correlation between Y_i, the number of times team i made the post-season playoffs in the 10 seasons under consideration and x_i = population in millions of the city associated with team i was 0.28.



(a) Describe in detail two major concerns that potentially threaten the validity of the analysis implied by the author's comments. (7 points)

Notice first that it is not possible to appear in the playoffs fewer than 0 or more than 10 times in 10 seasons. Linear correlation, then, is not appropriate because the mean function may fit values that are less than 0 or more than 10 for observed values of population. We should instead use logistic regression, as in the next section.

Second, it appears there may be a bad leverage point: that large city (around 18 million people) that made 10 playoff appearances will probably be both a large residual and is somewhat far away from the other points in the x-direction. We'll make another plot to double-check, but linear models are not valid in the presence of bad leverage points. (Some people also talked about the non-constant variance, which is also quite reasonable.)

[1] Bradbury, JC (2007) *The Baseball Economist.* Dutton, New York.

(b) Because of your concerns, you decide to instead fit the model below. Output for this model is found in the appendix. Is there strong evidence of a relationship between city population and making it into the playoffs? Conduct the appropriate hypothesis test, assuming your model is valid. (7 points)

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x$$

$H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

z = 2.838, p-value = 0.0045
OR $124.10 - 116.22 = 7.88$ from χ^2_1, implying $\sqrt{7.88} = 2.81$ from N(0,1), implying tiny p-value.

Because our p-value is so tiny, we indeed have strong evidence of a relationship between city population and making it into the playoffs. (Note: strong evidence of a relationship is not the same thing as a strong relationship. The evidence is strong when the p-value is small, and p-value depends not only on the slope but also variability of the slope from sample to sample.)

(c) Interpret the estimate for the parameter β_1 in your model in the context of this problem. (7 points)

Our model predicts that if the population of a city increases by one million, the odds of that city's major league baseball tame making it into the playoffs are multiplied by $e^{0.07807} = 1.0811$.

Cheddar Cheese

Model 1:

```
Analysis of Variance Table
Response: taste
          Df  Sum Sq Mean Sq F value    Pr(>F)
Acetic     1 2314.14 2314.14 22.5481 6.528e-05 ***
H2S        1 2147.02 2147.02 20.9197 0.0001035 ***
Lactic     1  533.32  533.32  5.1964 0.0310795 *
Residuals 26 2668.41  102.63

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
Acetic        0.3277     4.4598   0.073  0.94198
H2S           3.9118     1.2484   3.133  0.00425 **
Lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared: 0.6518,Adjusted R-squared: 0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

Model 2:

```
Analysis of Variance Table
Response: taste
          Df Sum Sq Mean Sq F value    Pr(>F)
H2S        1 4376.7  4376.7 44.2764 3.851e-07 ***
Lactic     1  617.2   617.2  6.2435   0.01885 *
Residuals 27 2669.0    98.9


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27.592      8.982  -3.072  0.00481 **
H2S            3.946      1.136   3.475  0.00174 **
Lactic        19.887      7.959   2.499  0.01885 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.942 on 27 degrees of freedom
Multiple R-squared: 0.6517,Adjusted R-squared: 0.6259
F-statistic: 25.26 on 2 and 27 DF,  p-value: 6.551e-07
```

Baseball

```
Call:
glm(formula = cbind(PlayoffAppearances, fails) ~ Population,
    family = binomial())

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.45843    0.21102  -6.911  4.8e-12 ***
Population   0.07807    0.02751   2.838  0.00455 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.10  on 29  degrees of freedom
Residual deviance: 116.22  on 28  degrees of freedom
AIC: 170.33
```