STATISTICS 608 Linear Models -Final Exam
May 2, 2014

Student's Name: ____

Student's Email Address: ____

INSTRUCTIONS FOR STUDENTS:

1. There are **13** 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 **two** 8.5" X 11" sheets of notes (front and back) 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 75 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. When the log odds of an event are positive, what does that tell you about the probability of the event? Be as specific as possible.

   (a) The probability is greater than 0.5.

   (b) The probability is smaller than 0.5.

   (c) The probability is greater than 0.

   (d) The probability is greater than 1.

   (e) The probability is smaller than 1.

2. Why does R^2_{adj} have an adjustment?

   (a) As n increases, the value of R^2 increases, on average. The adjustment involves the sample size.

   (b) Adding irrelevant predictor variables often increases R^2. The adjustment involves the number of predictors.

   (c) R^2 is a biased estimate of the proportion of total variability in the $Y's$ explained by the regression model. The adjustment corrects for that bias.

   (d) R^2 is a biased estimate of the variance of the errors. The adjustment corrects for that bias.

3. When error terms have positive autocorrelation, for example in data collected across time, but the usual least squares models are fit assuming independent errors, what happens to the calculated p-values for slopes?

   (a) The p-values are too large, due to variance of the sample slopes being larger than calculated.

   (b) The p-values are too large, due to variance of the sample slopes being smaller than calculated.

   (c) The p-values are too small, due to variance of the sample slopes being larger than calculated.

   (d) The p-values are too small, due to variance of the sample slopes being smaller than calculated.

4. Transformations could be used for which of the following?

   (a) Overcoming problems due to nonconstant variance

   (b) Estimating percentage effects

   (c) Overcoming problems due to nonlinearity

   (d) All of the above are potential reasons to use transformations

5. Suppose that a predictor variable x is normally distributed. For a logistic regression model shown below, which of the following is true?

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

where $\theta(x) = P(Y = 1|X = x)$.

   (a) The log odds are always a linear function of x.

   (b) The log odds are a linear function of x if the means of x are equal when $Y = 1$ and when $Y = 0$.

   (c) The log odds are a linear function of x if the variances of x are equal when $Y = 1$ and when $Y = 0$.

   (d) The log odds are a linear function of x if the means and variances of x are equal when $Y = 1$ and when $Y = 0$.

6. Which two variable selection techniques will always result in the same model chosen?

   (a) Backward selection and stepwise selection based on the same criterion (R^2_{adj}, AIC, $AICC$, or BIC)

   (b) Backward selection and forward selection based on the same criterion

   (c) AIC and $AICC$, as long as $n > 30$

   (d) AIC and BIC, as long as $n > 30$

   (e) The F-test for model reduction for only one variable dropped and one step of a backward selection based on p-value

7. We often have transformed predictors from their original distribution to a normal distribution, for example, using the Box-Cox method to choose the transformation. Of the following, which is the most important benefit of transforming a predictor variable to a normal distribution for a simple linear regression model?

   (a) To ensure that residuals are normally distributed.

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

   (c) To ensure that there is a linear relationship between the predictor and the response variable.

   (d) To ensure that the residuals are independent of the response variable.

Part II: Short Answer

8. Explain, as if to someone with very little experience with regression, the need for the logit function for fitting regression models when the response variable is binary.

   First, if we tried using simple linear regression models, we might estimate probabilites that are greater than 1 or smaller than 0 if we fit a straight line between X and Y.

   Second, one assumption of a linear regression model is that variance of the response is constant across different values of the predictor variable. But if the probability of a success is different for different values of X, the variance of the sample proportions estimating those probabilities will also be different for different values of X.

9. Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where the errors are assumed to have mean zero but known variance-covariance matrix Σ. The generalized least squares estimator of the parameter vector is given by:

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y}$$

   where Σ is a symmetric $(n \times n)$ matrix with (i,j) element equal to $\text{Cov}(e_i, e_j)$. Show that the variance of $\hat{\boldsymbol{\beta}}_{GLS}$ is equal to $(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}$.

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}_{GLS}|\mathbf{X}) &= \text{Var}((\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y}|\mathbf{X}) \\
&= (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\text{Var}(\mathbf{Y}|\mathbf{X})\Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\Sigma\Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}
\end{aligned}$$

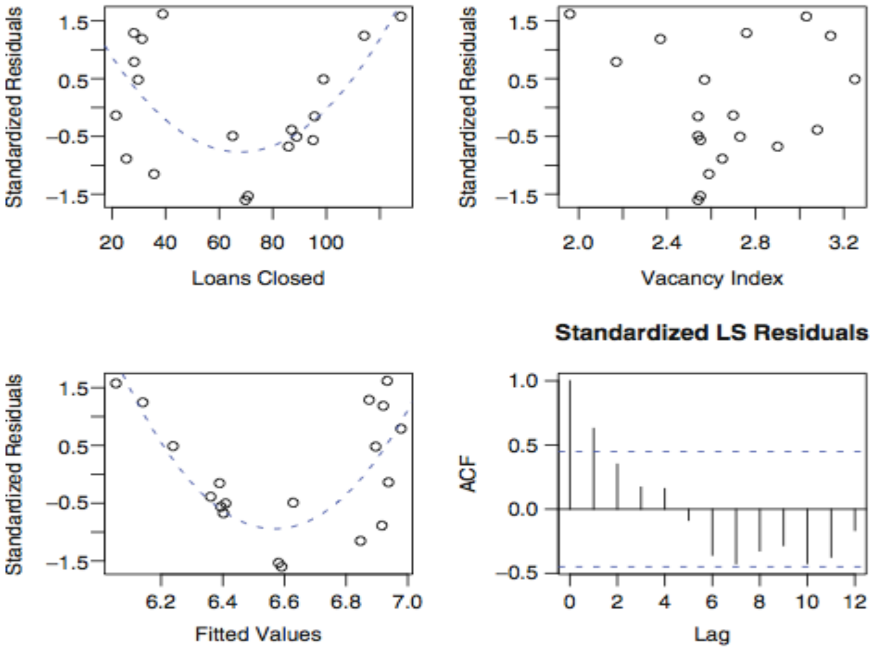
10. Consider a multiple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$. An added variable plot is used to assess the effect of x_2 on y, after having adjusted for the effect of x_1. The added variable plot will plot the residuals from regressing y onto x_1 on the y-axis against the residuals from regressing x_2 onto x_1 on the x-axis. Now suppose that the sample correlation between x_1 and x_2 is exactly equal to 1, while the correlation between y and x_1 is between 0 and 1. Describe what the added variable plot will look like, and why.

    If the correlation between x_1 and x_2 is exactly equal to 1, then the residuals from regressing x_2 onto x_1 will all be equal to zero. But that is what is plotted on the x-axis of the added variable plot. Since the correlation between y and x_1 is between 0 and 1, the residuals will be non-zero, so we expect some spread on the vertical axis. So with all 0's on the x-axis and a spread of values on the y-axis, we expect a single column of dots. (Sketches are great for this question.)

11. In a case study from Tryfos (1998), the savings and loan associations in San Francisco, California held almost all of the residential real estate loans in the 1990s. Interest centers on developing a regression model to predict interest rates (Y) from x_1, the amount of loans closed (in millions of dollars) and x_2, the vacancy index. Data from the city are available on each of these variables over a consecutive 19-month period in the 1990s. A model considered was

$$InterestRate_i = \beta_0 + \beta_1 LoansClosed_i + \beta_2 VacancyIndex_i + e$$

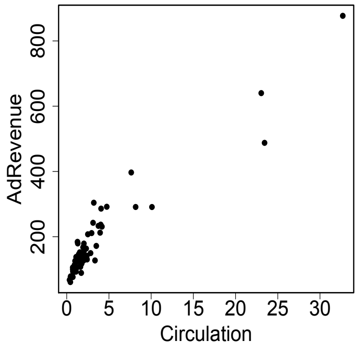
Below are shown a few plots from fitting the model above. Describe the best strategy for addressing the quadratic pattern in the residuals.



While it is possible that a quadratic term is necessary in LoansClosed, the correlation in the errors is the most important reason the model is not valid. It is possible that this autocorrelation is causing the quadratic pattern in the residuals, and once the autocorrelation is addressed, the model will be valid.

Part II: Long Answer

12. Below is shown a plot of the circulation of popular magazines (in millions) against revenue from advertising (in millions of dollars).



(a) Suppose we fit a simple linear regression model to the data. What would be a problem with resulting prediction intervals for advertising revenue using this model? Describe in particular the difference between intervals when circulation = 2 million as compared to when circulation = 25 million.

Notice that the variability of AdRevenue is different for different values of Circulation. That is, we have a problem with heteroskedasticity. Because of that assumption, a prediction interval calculated assuming equal variances in AdRevenue across the Circulation values will create similar-width intervals when Circulation = 2 and when Circulation = 25 million. However, we notice the observed variability in the data is much smaller when Circulation = 2, so that prediction interval should be much narrower.

(b) Instead of a simple linear regression model, researchers decided to transform both circulation and advertising revenue using the log transformation. A plot and output from this transformed model are found in the appendix. Assume the model is valid. A resulting 95% prediction interval for log(AdRevenue) was found to be (4.686, 5.397). Use this information to calculate an appropriate prediction interval for AdRevenue.

$$\log(AdRevenue) = \beta_0 + \beta_1 \log(Circulation) + e$$

Remember the correction factor! We want to add MSE/2 before back-transforming. Or, if you don't, state that you don't. The general principal of the back-transformation from log units to original units is the most important.

$$(exp(4.686 + 0.0313/2), exp(5.397 + 0.0313/2))$$
$$(110.13, 224.23)$$

And if you forgot the correction factor on this particular example, you get pretty close: (108.42, 220.74). That's not the end of the world!

13. A study on the Pacific Brant[1] (a small migratory goose) considered the effects of air traffic such as helicopters on Pacific Brants' flights. In this study, helicopters were flown at several different altitudes (vertical levels) and lateral (side-to-side) distances from flocks of Brant. The response variable *Flight* takes the value 1 if more than 10% of the flock flew away in response to the helicopter, and 0 otherwise.

    (a) A preliminary model considered both altitude and lateral distance to predict whether the flock would fly away, as shown below. Interpret the parameter estimate for β_1 from this model in context. Output from Model 1 is in the appendix.

$$\text{Model 1: } \log\left(\frac{\theta}{1-\theta}\right) = \beta_0 + \beta_1\,\text{altitude} + \beta_2\,\text{lateral} + e,$$

where θ = probability of Flight = 1.

Remember to transform, and remember to discuss odds, not probability! When altitude of the helicopter increases by one unit (perhaps feet), the odds of the birds flying away is multiplied by exp(0.19652) = 1.23. That is, they increase by 23%.

    (b) Researchers were also considering adding an interaction term between altitude and lateral distance. Explain what this would mean in context, as if to someone with no statistical experience.

If it is substantial, an interaction would mean that the effect of lateral distance to the birds on whether they fly away is substantially different depending on the altitude of the helicopter.

[1] Peck, Roxy; Haugh, Larry; and Goodman, Arnold (1998). Statistical Case Studies: A Collaboration Between Academe and Industry. Washington DC: SIAM and ASA.

(c) Plots and output from a second model with an interaction term, along with a plot of altitude by lateral distance for different values of *Flight* are shown in the appendix. Should the interaction term be included? Explain carefully why or why not.

Yes, I would recommend including the interaction term. Notice that the relationship between lateral distance and altitude is substantially different for the two different values of Y (whether or not the flock flew away). That implies that the relationship between lateral distance and whether or not the flock flew away is different for different altitudes.

14. Your research project is to predict y = the total gross earnings (in dollars) of the top movies released in the U.S. in 2013. Predictors of interest are found in the table below, along with transformations for those predictors.

Predictor	Value	Transformation
x_1	number of theaters movie is played in	Square
x_2	gross earnings on opening weekend	Log
x_3	number of days in theaters	Log

The total gross earnings in dollars was also transformed using log. A preliminary model fitted was:

$$\log(y) = \beta_0 + \beta_1 x_1^2 + \beta_2 \log(x_2) + \beta_3 \log(x_3) + e$$

(a) A colleague is interested in using the stepwise variable selection procedure to choose variables to include in the model. Why might all possible subsets be preferable?

Since we have so few variables, it will require very little computational time to fit all 7 possible subset models and calculate our selection statistic on all the models. We would usually only consider forward, backward, or stepwise selection methods when we had a very large number of variables to select.

(b) Below is found a table of R^2_{adj}, AIC, AIC_C, and BIC. Which models are selected by each of thefour model selection criteria?

Subset Size	Predictors	R^2_{adj}	AIC	AIC_C	BIC
1	$log(x_2)$	0.7777	-211.96	-211.71	-206.75
2	$log(x_2)$, $log(x_3)$	0.8925	-284.68	-284.26	-276.87
3	x_1^2, $log(x_2)$, $log(x_3)$	0.8957	-286.72	-286.09	-276.30

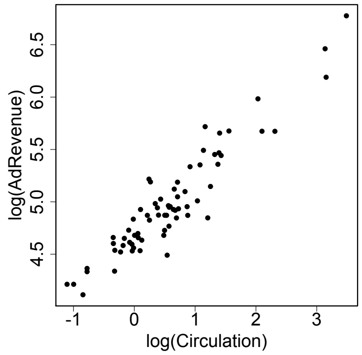
R^2_{adj} chooses the model with the largest value of R^2_{adj}, while the other three selection methods choose the smallest value. We notice model 3 is chosen by R^2_{adj}, AIC, and AIC_C, while model 2 is chosen by BIC, though the BIC values are not substantially different between model 2 and model 3. If the goal of the researchers were interpretation, I would prefer the simpler model, while if the goal were prediction, I would prefer the more complex model, assuming both models were valid.

(c) Now your colleague wants to use the p-values from the selected model above to test for significance of the predictors in that model. Explain the issues regarding using p-values after variable selection.

Since we have chosen the model that best fits this particular data set, it may also be true that this particular data set fits this model better than other data sets. That means that p-values will be, on average, smaller after model selection than they would be on a separately generated data set. If the goal of the researchers is to use the model for statistical inference (hypothesis testing and confidence intervals), we should use this model on a separate data set to find out how well it actually works.

Appendix

Ad Circulation: Transformed Model



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.67473    0.02525  185.16   <2e-16 ***
logcir       0.52876    0.02356   22.44   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1768 on 68 degrees of freedom
Multiple R-squared:  0.881,Adjusted R-squared:  0.8793
F-statistic: 503.6 on 1 and 68 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: logads
          Df  Sum Sq Mean Sq F value    Pr(>F)
logcir     1 15.7427 15.7427  503.62 < 2.2e-16 ***
Residuals 68  2.1256  0.0313
```

Pacific Brant Geese: Model 1

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  2.39541    0.30551   7.841 4.48e-15 ***
Altitude     0.19652    0.06745   2.914  0.00357 ** 
Lateral     -0.23883    0.02248 -10.625  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 618.81  on 463  degrees of freedom
Residual deviance: 359.63  on 461  degrees of freedom
AIC: 365.63
```

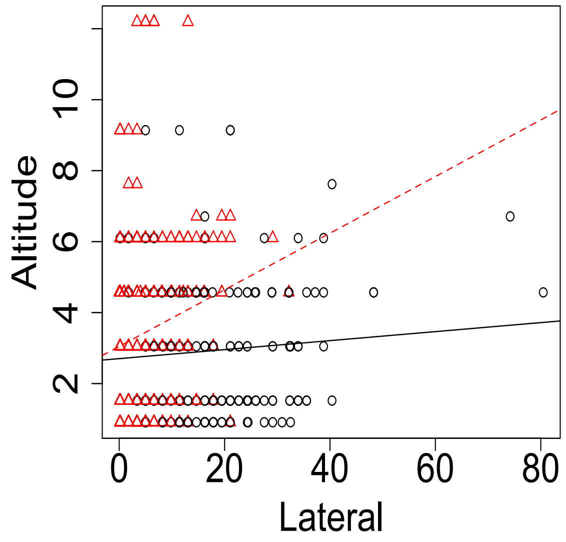
Pacific Brant Geese: Model 2

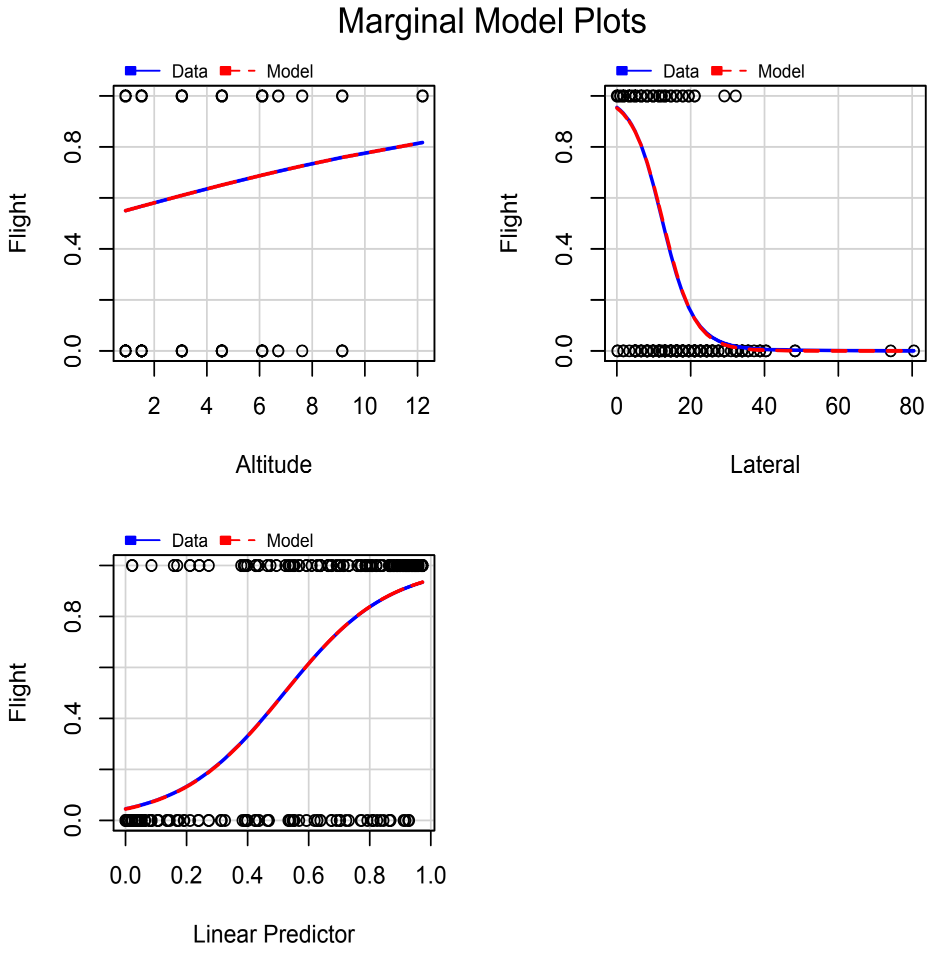
```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)       3.825039   0.499216   7.662 1.83e-14 ***
Altitude         -0.203013   0.105580  -1.923   0.0545 .  
Lateral          -0.390944   0.048391  -8.079 6.54e-16 ***
Altitude:Lateral  0.040137   0.009487   4.231 2.33e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 618.81  on 463  degrees of freedom
Residual deviance: 342.36  on 460  degrees of freedom
AIC: 350.36
```

Pacific Brant Geese: Model 2


Altitude
Lateral


Marginal Model Plots
Data
Model
Flight
Altitude
Data
Model
Flight
Lateral
Data
Model
Flight
Linear Predictor