**STAT 501 – Homework 12 (covering Lessons 14 & 15)**

**Due date Dec 9 Wednesday**

**Instructions**: Use Word to type your answers within this document. Then, submit your answers in the appropriate dropbox in ANGEL by the due date. The point distribution is located next to each question. If there are multiple parts, then the points are divided equally over the subparts.

1. (**5x5 = 25 points**) Use the “Lake Erie” dataset. One column of the dataset gives water levels of Lake Erie for *n* = 40 consecutive Novembers and another gives the lag 1 values of the series. The lag 1 value is the value from the previous year.
2. Do a time series plot of the variable NovLevel. That is, plot the NovLevel in time order. [In Minitab, use Graph **>** Time Series Plot.] Is there any obvious upward or downward trend to the Lake Erie levels?



There is no obvious upward or downward trend in this plot.

1. Plot NovLevel against the lag 1 values. Describe the noteworthy features of the plot.



The noteworthy features are:

* There appears to be a moderate linear pattern, suggesting that the first-order autoregression model  could be useful.

1. Determine the partial autocorrelations for the NovLevel series. [In Minitab, use Stat **>** Time Series **>** Partial Autocorrelation, enter NovLevel as the variable (or series) and click OK.] What do the results indicate about the autoregression order for a model that describes November water levels of Lake Erie?



**Partial Autocorrelation Function: NovLevel**

Lag PACF T

1 0.719189 4.55

2 0.012461 0.08

3 -0.025954 -0.16

4 -0.204191 -1.29

5 0.129207 0.82

6 0.044124 0.28

7 0.198148 1.25

8 -0.034274 -0.22

9 -0.087917 -0.56

10 -0.292095 -1.85

It is obvious from the relatively high value of the lag 1 partial autocorrelation that an AR(1) model would likely be feasible for this data set.

1. Do a simple regression with NovLevel as the response and lag 1 values of NovLevel as the predictor variable. Use the Storage button to store the residuals. (Note: This is a first-order autoregression model for the series.)
   1. Write the estimated regression equation.
   2. Use the regression equation to find the fitted value of the November water level of Lake Erie in the next year after the final year of the data series. [For this question you *don’t* need to use the method at the bottom of Section 14.3.]
2. NovLevelHat = 3.88 + 0.723 Lag1level
3. NovLevelHat = 3.88 + (0.723 \* 14.801) = 14.581
4. Determine partial autocorrelations for the residuals from the regression that you did in part (d). [In Minitab, use Stat **>** Time Series **>** Partial Autocorrelation, enter RESI1 as the variable (or series) and click OK.] What do the results indicate? (Note: The residuals ideally are a random sample, which means that they should have no time series structure to them. That is, ideally, any partial autocorrelations should have values near 0.)



**Partial Autocorrelation Function: RESI1**

Lag PACF T

1 0.004891 0.03

2 0.031685 0.20

3 0.127993 0.80

4 -0.199285 -1.24

5 -0.056746 -0.35

6 -0.127988 -0.80

7 0.141765 0.89

8 0.109857 0.69

9 0.221144 1.38

10 -0.078279 -0.49

It is obvious from the plot that none of the partial autocorrelation are higher than the boundaries. Here we see that the residuals are a random sample and don’t have any time series structure to them.

1. (**10x5 = 50 points**) Use the “Penn State Student” dataset. The data are from *n* = 900 students at Penn State, surveyed in February of 2007. We will use logistic regression to predict the probability that a student says they have ever cheated on a college exam. The *y*-variable is ChtdExam and possible responses are Yes and No.
2. Do a logistic regression that relates the probability of having ever cheated on an exam to GPA. In Minitab, use Stat **>** Regression **>** Binary Logistic Regression > (Fit) Binary Logistic Model. Take the following actions:

* Enter ChtdExam as Response.
* Enter GPA in the Continuous Predictors box.
* Click Storage and request Fits (event probabilities). Click OK.

Write the sample logistic regression equation for this situation by filling in values for the coefficients b0 and b1 in the equation:

b0=-0.573

b1=-0.498

1. Plot the stored predicted probabilities of having ever cheated on an exam versus GPA. [In Minitab, use Graph **>** Scatterplot and select “With Connect Line.” Use the column Fits1 as the *y*-variable and GPA as the *x*-variable.] Copy and paste the plot as part of your answer AND briefly describe what the plot shows about how the probability of having ever cheated on an exam is related to GPA.



The probability of having ever cheated decreases non linearly as GPA increases.

1. Examine the results that were generated in part (a) in order to find the odds ratio for GPA. Write a sentence that gives the value and interprets it in this situation. [If you're using software that does not give the odds ratio, calculate it using the formula *eb*1.]

The odds ration from minitab output = 0.6076

Also, exp(-0.498) = 0.6078

The interpretation of the odds ratio is that for every increase of 1 unit in GPA, the estimated odds of student ever cheated on a college exam are multiplied by 0.6076 (reduce). Since the odds ratio is less than 1, the odds of cheating are less for higher levels of GPA.

1. Refer to the equation that you wrote in part (a). Use it to estimate the probability of having ever cheated on an exam for a student with a GPA = 3.0. (You can use the plot in part (b) to see if your answer is in the neighborhood of being correct.)

p^ =EXP(-0.573-0.498\*3)/(1+EXP(-0.573-0.498\*3)) = 0.1123

This aligns with the plot in part b. We can also get the odds:

Odds = 0.1123 / (1-0.1123) = 0.1265 equivalent to = EXP(-0.573-0.498\*3)

1. Calculate the odds of having ever cheated on an exam for a student with a GPA = 4.0. [Hint: You can calculate this two different ways. Use the equation from part (a) and the fact that the odds of an event are p/(1–p). Alternatively, use the answers to parts (c) and (d) and the fact that the odds ratio multiplies the odds each time *x* is increased by one unit.]

p^ =EXP(-0.573-0.498\*4)/(1+EXP(-0.573-0.498\*4)) = 0.0714

Odds = 0.0714 / (1-0.0714) = 0.07689 equivalent to = EXP(-0.573-0.498\*4)

Using the 2nd mechanism = 0.1265 \* 0.6076 = 0.07686

1. Now add the variable SkipClass as a predictor variable in the model (along with GPA). SkipClass is the number of classes student says he or she misses in a typical week. What is the evidence in the output that SkipClass is related to the probability of having ever cheated on an exam?

The output:

Coefficients

Term Coef SE Coef 95% CI Z-Value P-Value VIF

Constant -1.190 0.762 (-2.683, 0.303) -1.56 0.118

GPA -0.375 0.234 (-0.835, 0.084) -1.60 0.109 1.02

SkipClass 0.346 0.104 ( 0.142, 0.550) 3.33 0.001 1.02

The p-values for SkipClass is less than 0.05. This is evidence that SkipClass is useful for predicting the probability of having ever cheated on an exam.

1. For the model with two *x*-variables, write a sentence that gives the odds ratio for SkipClass and interprets it in the context of this situation.

For SkipClass, the odds ratio is 1.4138 (calculated as e^0.346). It is interpreted as: the odds of ever having cheated on an exam increases by 1.4138 times for every additional skipped class, assuming GPA is held constant.

1. In the logistic model with two *x* variables, the odds of an event can be computed directly from the equation:

Use this equation to estimate the odds of having ever cheated on an exam for students with a 3.5 GPA who typically misses one class per week.

Odds (GPA = 3.5 and SkipClass=1) = exp(-1.190-(0.375\*3.5)+(0.346\*1)) = 0.1157

1. In part (g) you wrote the odds ratio for SkipClass and in part (h) you found the odds of having ever cheated on an exam for students with a 3.5 GPA who typically skip one class per week. Use only the answers to parts (g) and (h) to determine the odds of having ever cheated on an exam for students with a 3.5 GPA who typically skip two classes per week.

Odds (GPA = 3.5 and SkipClass=2) = 1.4138 \* 0.1157 = 0.1636

1. Four more variables in the dataset are the following:

* FakeID = whether student has ever used a fake id in order to be served alcohol.
* ChtdSO = whether student has ever cheated on another person with whom they were having a romantic relationship.
* SmokeCig = whether student smokes cigarettes.
* SmokeMJ = whether student has ever smoked marijuana.

All four of these variables are coded as either Yes or No. In the multiple logistic regression model for predicting the probability of ever having cheated on an exam, add the four variables just listed to the model along with GPA and SkipClass.

[In Minitab17, list these variables in the “Categorical Predictors.” In Minitab16, list these variables in the Model box and also in the Factors box. Minitab will create indicator variables for variables in the Factors box. In this case, each indicator variable created will equal 1 for Yes and 0 for No.]

Discuss the results with respect to these added variables only. Which variables are significant, which are not? Interpret the significant odds ratios.

The output is:

Coefficients

Term Coef SE Coef 95% CI Z-Value P-Value VIF

Constant -2.108 0.825 (-3.725, -0.490) -2.55 0.011

GPA -0.222 0.245 (-0.703, 0.259) -0.90 0.366 1.06

SkipClass 0.260 0.111 ( 0.041, 0.478) 2.33 0.020 1.07

FakeID

Yes 0.665 0.259 ( 0.158, 1.173) 2.57 0.010 1.05

ChtdSO

Yes 0.414 0.243 (-0.063, 0.891) 1.70 0.089 1.07

SmokeCig

Yes 0.823 0.294 ( 0.247, 1.399) 2.80 0.005 1.08

SmokedMJ

Yes 0.195 0.238 (-0.271, 0.661) 0.82 0.412 1.12

The p-values for SkipClass, FakeID, SmokeCig is less than 0.05. Therefore there is evidence that these are useful for predicting the probability of having ever cheated on an exam.

Odds Ratios for Categorical Predictors

Level A Level B Odds Ratio 95% CI

FakeID

Yes No 1.9454 (1.1716, 3.2305)

ChtdSO

Yes No 1.5131 (0.9389, 2.4384)

SmokeCig

Yes No 2.2769 (1.2797, 4.0509)

SmokedMJ

Yes No 1.2154 (0.7624, 1.9376)

The interpretation of various OddRatio is:

* For FakeID, the odds ratio is 1.9454 (calculated as exp(0.665)). For students who have used FakeId, the odds of ever having cheated on an exam are 1.85 times greater than the students who have not, assuming all the other variables in the model are held constant.
* For ChtdSO, the odds ratio is 1.5131. For students who have cheated on another person with whom they were having a romantic relationship, the odds of ever having cheated on an exam are 1.5131 times greater than the students who have not, assuming all the other variables in the model are held constant.
* For SmokeCig, the odds ratio is 2.2769. For students who have smoked cigarettes, the odds of ever having cheated on an exam are 2.2769 times greater than the students who have not, assuming all the other variables in the model are held constant.
* For SmokedMJ, the odds ratio is 1.2154. For students who have smoked marijuana, the odds of ever having cheated on an exam are 1.2154 times greater than the students who have not, assuming all the other variables in the model are held constant.

3. (**5x5 = 25 points**) Use the “Geriatric” dataset for a prospective study to investigate the effects of two interventions on the frequency of falls. One hundred subjects were randomly assigned to one of the two interventions: education only (*X1* = 0) and education plus aerobic exercise training (*X1* = 1). Subjects were at least 65 years of age and in reasonably good health. Three variables considered to be important as control variables were gender (*X2*: 0 = female; 1 = male), a balance index (*X3*), and a strength index (*X4*). The lower the balance index, the more stable is the subject; and the lower the strength index, the stronger is the subject. Each subject kept a diary recording the number of falls (*Y*) during the six months of the study.

1. Fit a Poisson regression model. In Minitab, use Stat **>** Regression **>** Poisson Regression > Fit Poisson Model. Enter Falls as the Response and Inter, Gender, Balance, and Strength in the Continuous Predictors box. Report the estimated regression equation.

Regression Equation

Falls = exp(Y')

Y' = 0.489 - 1.069 Inter - 0.047 Gender + 0.00947 Balance + 0.00857 Strength

1. Use a deviance (G2) test to determine whether Gender can be dropped from the model. State the null and alternative hypotheses, the test statistic and p-value from the Deviance Table, and draw a conclusion based on a significance level of 0.05.

The output is:

Deviance Table

Source DF Seq Dev Seq Mean Chi-Square P-Value

Regression 4 90.404 22.6010 90.40 0.000

Inter 1 75.209 75.2090 75.21 0.000

Balance 1 10.444 10.4438 10.44 0.001

Strength 1 4.600 4.6003 4.60 0.032

Gender 1 0.151 0.1510 0.15 0.698

Error 95 108.790 1.1452

Total 99 199.194

The test statistic for testing the interaction terms is G2=0.151 which is compared to a chi-square distribution with 5−4=1 degrees of freedom to find the p-value

We have from minitab:

**Cumulative Distribution Function**

Chi-Square with 1 DF

x P( X ≤ x )

0.151 0.302418

p-value = 1-0.302418 = 0.6976

Since p-value > alpha (0.05), we fail to reject the null hypothesis and conclude that Gender can be dropped from the model.

1. Fit a Poisson regression model using just Inter, Balance, and Strength as predictors. Click the Graphs button and select “Residuals versus order” before clicking OK. Do there appear to be any outlying cases? Include the graph in your answer.



Above is the plot of the deviance residuals vs order. There are no alarming patterns in these plots to suggest a major problem with the model.

There appear to be a couple of outlying cases.

1. Report the estimated regression equation for the model you fit in part (c). Based on the estimated regression equation, does aerobic exercise reduce the frequency of falls when controlling for balance and strength?

Regression Equation

Falls = exp(Y')

Y' = 0.444 - 1.078 Inter + 0.00947 Balance + 0.00898 Strength

Now we have (education only (X1 = 0) and education plus aerobic exercise training (X1 = 1)) i.e. the variable Inter = 1 for aerobic exercise training. We also know that the coef for Inter is negative. Therefore aerobic exercise **reduces** the frequency of falls when controlling for balance and strength.

1. Use the Minitab output to assess the overall fit of the model. Is there any evidence of lack-of-fit? Use values in the Deviance Table to derive the value of Pseudo *R2* (Deviance R-sq).

Deviance Table

Source DF Seq Dev Seq Mean Chi-Square P-Value

Regression 3 90.253 30.084 90.25 0.000

Inter 1 75.209 75.209 75.21 0.000

Balance 1 10.444 10.444 10.44 0.001

Strength 1 4.600 4.600 4.60 0.032

Error 96 108.941 1.135

Total 99 199.194

Goodness-of-Fit Tests

Test DF Estimate Mean Chi-Square P-Value

Deviance 96 108.94089 1.13480 108.94 0.173

Pearson 96 105.28778 1.09675 105.29 0.243

* The null model in this case has no predictors. The deviance for the null model is −2ℓ(β^(0))= 199.194, which is shown in the "Total" row in the Deviance Table.
* The deviance for the fitted model is −2ℓ(β^)=108.941, which is shown in the "Error" row in the Deviance Table.
* The deviance test statistic is therefore G2=199.194−108.941 = 90.253
* The p-value comes from a χ2 distribution with 4−1=3 degrees of freedom

We can interpret the high p-value as indicating no evidence of lack-of-fit.



R2 =1-(108.941/199.194) = 0.4531