**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?
3. Plot NovLevel against the lag 1 values. Describe the noteworthy features of the plot.
4. 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?
5. 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.]
6. 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.)
7. (**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.
8. 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:

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.
2. 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.]
3. 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.)
4. 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.]
5. 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?
6. 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.
7. 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.

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.
2. 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.

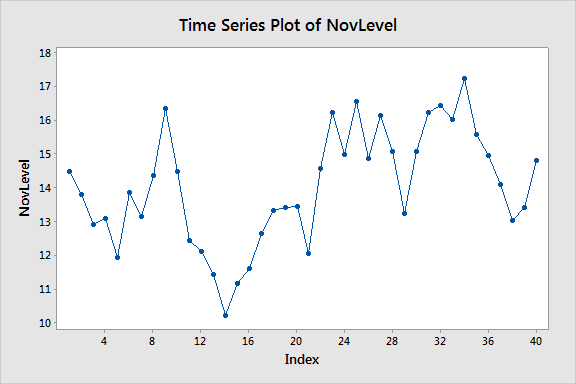
3. (**5x5 = 25 points**) Use the “Geriatric” dataset for a prospective study to investigate the effects of two interventions on the frequency of falls. One hundred subjects were randomly assigned to one of the two interventions: education only (*X1* = 0) and education plus aerobic exercise training (*X1* = 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.
2. 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.
3. 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.
4. 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?
5. 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).

**STAT 501 – Homework 12 Solutions Fall 2015**

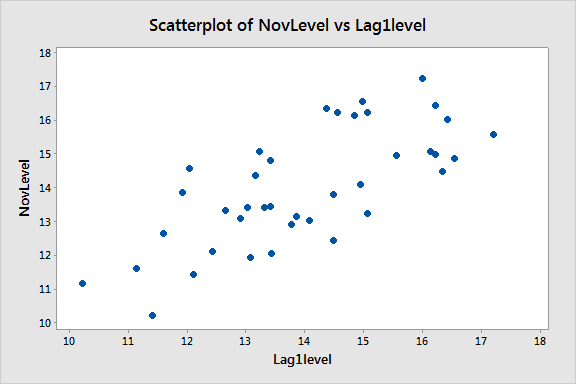
1. (**25 points**) Use the “Lake Erie” dataset. One column of the dataset gives water levels of Lake Erie for *n* = 40 consecutive Novembers and another gives the lag 1 values of the series. The lag 1 value is the value from the previous year.

1. 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?



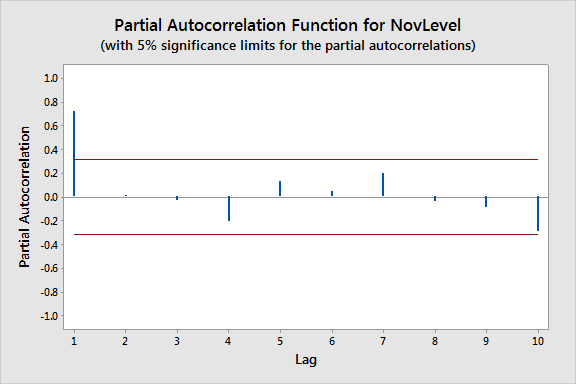
**There’s no overall positive or negative trend. Some students may say that there is a downward trend followed by an upward trend. It is okay to say that, but in fact that is a characteristic of data with a strong first-order autocorrelation. Values may tend to stay on the same side of the average for a while, but then will drift to the other side of the average.**

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

****

**There’s a moderately strong linear association, with no outliers.**

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?



**The lag 1 is high followed by relatively small autocorrelations for other lags. A first order autoregression model is suggested.**

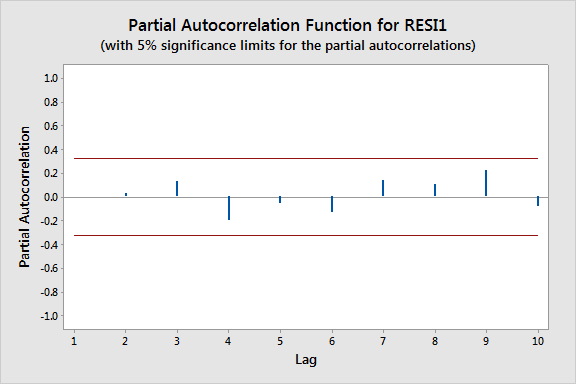
1. Do a simple regression with NovLevel as the response and lag 1 values of NovLevel as the predictor variable. Use the Storage button to store the residuals. (Note: This is a first-order autoregression model for the series.)
   1. Write the estimated regression equation.

**The regression equation is predicted NovLevel = 3.88 + 0.723 × Lag 1.**

* 1. 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.]

**Predicted NovLevel = 3.88 + 0.723 × (14.801) = 14.581. The last observed November level is the lag 1 level for the next time after the series.**

1. 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.)



**There are no large autocorrelations in the residuals. This indicates that these residuals are independent with regard to time order.**

2. (**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.

1. 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:

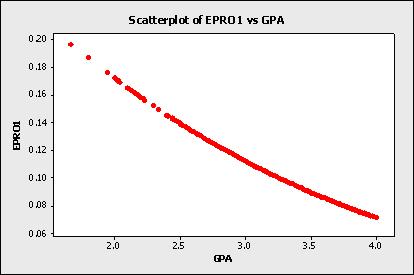
* Enter ChtdExam as Response.
* Enter GPA in the Continuous Predictors box.
* Click Storage and request Fits (event probabilities). Click OK.

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

.

**The sample logistic regression equation is given by**

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 ever having cheated on an exam decreases 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 ratio for GPA is 0.6076. The odds of having cheated on an exam are multiplied by 0.6076 when GPA is increased by one unit.**

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.)

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.]

**At GPA = 4.0, , so the odds of having cheated are 0.07143/(1–0.07143) = 0.0769. Alternatively, using answers for (c) and (d) the odds of having cheated are 0.6076×(0.11235)/(1−0.11235) = 0.0769.**

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 *p*-value for testing for the coefficient that multiplies the variable SkipClass is 0.002. Alternatively, you note that the confidence interval for the odds ratio is entirely above the value 1.**

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.

**Assuming that GPA is held constant, the odds of having cheated are multiplied by 1.4138 for each increase of one skipped class per week.**

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.

**e−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 of having ever cheated on an exam = 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 relevant output for this part is given below. We find the predictors**

* **Statistically significant: FakeID and SmokeCig;**
* **Statistically not significant: CheatedSO and SmokedMJ.**

**Having used a fake ID multiplies the odds of having cheated on an exam by 1.9454 (assuming all other variables are held constant). Being a cigarette smoker multiplies the odds of having cheated on an exam by 2.2769 (assuming all other variables are held constant).**

**Source DF Adj Dev Adj Mean Chi-Square P-Value**

**Regression 6 36.164 6.0273 36.16 0.000**

**GPA 1 0.809 0.8092 0.81 0.368**

**SkipClass 1 4.957 4.9570 4.96 0.026**

**FakeID 1 6.220 6.2197 6.22 0.013**

**ChtdSO 1 2.811 2.8114 2.81 0.094**

**SmokeCig 1 7.199 7.1991 7.20 0.007**

**SmokedMJ 1 0.670 0.6703 0.67 0.413**

**Error 893 562.059 0.6294**

**Total 899 598.223**

**Coefficients**

**Term Coef SE Coef VIF**

**Constant -2.108 0.825**

**GPA -0.222 0.245 1.06**

**SkipClass 0.260 0.111 1.07**

**FakeID**

**Yes 0.665 0.259 1.05**

**ChtdSO**

**Yes 0.414 0.243 1.07**

**SmokeCig**

**Yes 0.823 0.294 1.08**

**SmokedMJ**

**Yes 0.195 0.238 1.12**

**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)**

**Odds ratio for level A relative to level B**

3. (**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.

**H0: population regression coefficient for Gender = 0**

**Ha: population regression coefficient for Gender ≠ 0**

**Deviance Table**

**Source DF Adj Dev Adj Mean Chi-Square P-Value**

**Regression 4 90.404 22.6010 90.40 0.000**

**Inter 1 73.520 73.5202 73.52 0.000**

**Gender 1 0.151 0.1510 0.15 0.698**

**Balance 1 10.282 10.2823 10.28 0.001**

**Strength 1 3.976 3.9761 3.98 0.046**

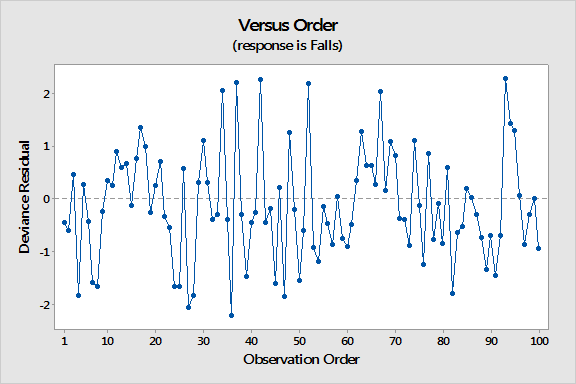
**Error 95 108.790 1.1452**

**Total 99 199.194**

**Test statistic G2 = 0.15 with p-value = 0.698.**

**Since p-value > significance level, fail to reject H0 and conclude that there is insufficient evidence that the population regression coefficient for Gender ≠ 0, i.e., 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.



**There don’t appear to be any 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**

**Yes, aerobic exercise reduces the frequency of falls when controlling for balance and strength since the estimated regression coefficient for Inter is negative.**

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).

**There is no evidence of lack-of-fit based on the Goodness-of-Fit tests since both p-values are greater than 0.05:**

**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 0.24**

**Model Summary**

**Deviance Deviance**

**R-Sq R-Sq(adj) AIC**

**45.31% 43.80% 375.44**

**From the Deviance Table, Pseudo *R2* = 1 – 108.941/199.194 = 45.31%.**