Homework Assignment 1

Due Date: Tuesday, January18th

Note: There are 7 problems on this assignment. The first 5 provide examples and extra practice. They are relevant for the exams but do not need to be turned in. Solutions for them are available on the class web site. You must turn in Problems 6-7 together with the corresponding STATA/SAS printouts to receive full credit. The assignment is due Tuesday, January 18th to give you time to ask any last questions after the holiday weekend. It may be uploaded to CCLE any time before midnight on that date.

Note: Output from any calculations done in STATA or SAS MUST be included with your assignment for full credit. If I do not specify which way to do a problem you may chose whether to do it by hand or on the computer. All the STATA/SAS commands needed to complete this homework are given at the end of the assignment and will be reviewed in the lab. You do not need to hand in a separate lab report–simply turn in the relevant output as part of your homework.

Note: You are encouraged to work with fellow students, as necessary, on these problems. However, each of you MUST write up your solution ON YOUR OWN and IN YOUR OWN WORDS. The style of your write-up is as important as getting the correct answer. Your solutions should be easy to follow, and contain English explanations of what you are doing and why. You do not have to write an essay for each problem, but you should give enough comments so that someone who has not seen the problem statement can understand your work. You do not have to type your assignments. However, if they are too sloppy to read, too hard to understand, or give just numbers with no comments, you WILL lose points. Problems labeled (GS) are adapted from our optional text, *Primer of Applied Regression & Analysis of Variance* by Stanton Glantz and Bryan Slinker. Problems labeled BR are adapted from the 100A/B text, *Fundamentals of Biostatistics*, 6th edition by Bernard Rosner.

Warm-up Problems

(1) Maximum Likelihood Basics:

(a) Explain what a likelihood function is.

(b) Explain the basic principle behind **maximum likelihood estimation**.

(c) You are collecting data on a random variable, X, which is assumed to be normally distributed with unknown mean μ and variance σ^2. Suppose you take a sample of size n=1 and find $X = 10$. Explain what the MLE for μ would be and why. Can you get an MLE for σ^2 based on this sample? Explain briefly.

(d) Optional Extra: Under the assumptions of part (c), show that the MLE for μ based on a sample of size n is $\bar{X}$.

(2) Generalized Linear Model Basics:

(a) Describe the 3 key components of the generalized linear model and the basic approach to estimating the coefficients.

(b) Explain what the terms **deviance, null model, full model** and **saturated** model refer to and how they relate to our evaluation of how well a GLM performs.

(3) Logistic Regression Basics:

(a) Explain what the three basic GLM components are for a logistic regression and why they are chosen that way.

(b) Explain what the coefficients in a logistic regression tell us (i) for a continuous predictor variable and (ii) for an indicator variable.

(c) Give the basic definition of an odds ratio and explain how this relates to your answers in part (b).

(4): Sports Fanatics My husband, Gareth, is from New Zealand where the national sports passion is rugby (sort of like American football only much better!) The national rugby team is called the All Blacks (they wear black) and their main rivals are Australia (the Wallabies) and South Africa (the Springboks). Gareth is interested in understanding what factors are related to the likelihood of an All Blacks victory. He therefore decides to perform a logistic regression with the response variable, Y, being whether or not the All Blacks win ($Y = 1$ if they win and 0 if they lose). The predictors are

AB Win% (the percentage of the previous ten games that the All Blacks had won going into the game in question, ranging from 0 to 100.)

OppWin% (same definition for the opponent's last 10 games.)

Home? (an indicator variable with 1 corresponding to an All Blacks home game and 0 an away game.)

Temperature (the temperature at which the game was played.)

Australia? (a dummy variable with 1 corresponding to a game against archrival Australia and 0 a game against another team.)

Below are the p-value for the overall likelihood ratio chi-square test along with a table of coefficients, standard errors, Z scores and p-values for the Wald tests corresponding to the various predictors. Use them to answer the questions below.

```
LR chi2 p-Value < 0.0001
```

	Coef	SE	Z	p-value
Constant	-25.30	10.54	-2.40	0.0163
AB Win %	0.466	0.176	2.65	0.0082
Opp Win %	-0.170	0.643	-2.65	0.0081
Home?	1.45	0.66	2.20	0.0278
Temperature	0.115	0.045	2.55	0.0108
Australia?	-0.245	1.89	-0.13	0.8969

(a) Is there evidence that at least one of the variables is a statistically significant predictor of whether the All Blacks win? Justify your answer.

(b) What does the coefficient for Temperature tell us about the relationship between Temperature and the probability that the All Blacks win? Compute the corresponding odds ratio for a 10 degree increase in

temperature and explain what it means. Give a confidence interval for this odds ratio.

(c) Which variables are statistically significant? Justify your answer. Do the signs of the various coefficients make sense?

(d) Estimate the probability of the New Zealand All Blacks winning a game against South Africa played in South Africa at 50 degree temperatures where both teams have a winning percentage of 70%. (Note: Use 70 (not 0.7) in the calculation!)

(e) Find a confidence interval for the coefficient of the Home? variable and give a brief interpretation. Find the corresponding odds ratio its 95% confidence interval and interpret those results.

(f) The coefficient for the Home? variable seems to indicate that the All Blacks are more likely to win at home than on the road. However, somewhat surprisingly, the All Blacks turn out to win more games on the road than at home. One of my husband's MBA students (from that school on the wrong side of town) looks at these results and states that this indicates that there must be some mistake in the analysis. However, you tell them that in fact this apparent inconsistency is entirely possible even if the model is correct. Assuming that the model is correct (i.e. there are no important variables missing from the model or violations of the basic assumptions etc.) and the coefficient estimates are exactly correct, how could the coefficient for Home? be positive even though the All Blacks win more games on the road?

(5) Asthma As Math : Professor Urtha Green, an environmental health scientist at my favorite school, the University of Calculationally Literate Adults, is interested in the role of air quality as a risk factor for childhood asthma. She has participated in a study that followed 1000 children from birth to age 10, recording whether or not they developed asthma during that period ($Y = 1$ for yes and $Y = 0$ for no). The study also collected information on potential risk factors and protective effects from the child's first year of life including X_1, an indicator for whether the child's family lived in an urban setting (Yes = 1, No = 0), X_2, the average annual pollution level in thousands of particles per cm^3 for the county in which the child lived, X_3, an index of socio-economic status for the child's family (higher is better), X_4, the number of months for which the child was breast-fed, X_5, an indicator for whether there was a family history of asthma (1 = Yes, 0 = No), and X_6, gender (1 = Female, 0 = Male). The data are given in the accompanying file. Note that the family history variable for this problem is labeled famhist to distinguish it from a similar variable in turn-in Problem 7. Use the data to answer the following questions.

First Dr. Green wants to know whether living in an urban environment is associated with increased risk of developing asthma. We will analyze this three different ways (and hopefully get the same answer from all three!):

(a) First, use a two-sample test of proportions to compare the probability of developing asthma in children who did and did not live in an urban setting as infants. Carefully state the null and alternative hypotheses mathematically and in words, use STATA or SAS to get the sample proportions for each group, obtain the test statistic and p-value, and explain your real-world conclusions. Show by hand how you can obtain the odds ratio for comparing children who were and were not brought up in an urban envoronment from the sample proportions.

(b) Contingency tables are a standard method for examining the relationship between two categorical variables. When the outcome variable and the predictor are both binary a contingency table is equivalent to a two-sample test of proportions and to a logistic regression. Redo part (a) using STATA or SAS to perform a contingency table (chi-squared) test. You do not need to restate the hypotheses or conclusions. Simply get the printout and confirm that your chi-squared test statistic is just the square of the Z statistic in part (a) and that your p-values match. What do you have to do to the p-value on the output to make it match the

test of interst to Dr. Green?

(c) Now view this problem as a logistic regression with whether or not the child develops asthma as the outcome and urban as the predictor. Write the hypothesis of interest in terms of the logistic regression coefficients and then fit the model in STATA or SAS. There is a simple relationship between the regression coefficients and various quantities you computed in part (a). Say what those relationships are and verify that the regression coefficients from your printout match them.

(d) Next Dr. Green wants to look at the relationship between pollution and asthma. To help her visualize the relationship, bin the pollution data into intervals of length 5 (the data values range from 0 to 40 thousands of particles per cm^3), compute the proportion of children in each pollution range who got asthma, and plot the results. Provide a brief clinical interpretation of what you see. Do you think there will be a significant relationship? Verify your answer by running the corresponding logistic regression in STATA or SAS. (I have included the bins and counts in the data set for your reference in case you want to skip the computation.)

(e) Now fit a logistic model with both the urban indicator and pollution included as predictors. Which variable(s) are significant now? Is this consistent with your answers to parts (a)-(d)? If not, explain what you think has happened.

For the remainder of the problem we will focus on a logistic model that includes all 6 of the predictor variables listed in the problem statement.

(f) Using STATA or SAS fit (i) the null model (the model with no predictors) and (ii) the full model (the model with all 6 predictors). Is the full model significantly better than the null model? Check this by performing the likelihood ratio chi-squared test, writing out all the details. You should be able to do this either by reading off the test results from the full-model printout or by computing the test statistic from the log likelihoods for the two models.

(g) The model in part (e) contains only environmental factors while the model in part (f) adds a set of child/family characteristics. Is the model with the personal characteristics a significant improvement over the model with just the environmental characteristics? Carry out an appropriate test (i) by hand given information from the printouts in (e) and (f) and (ii) by using a follow-up test command to the model in (f).

(h) Which conditions/characteristics appear to be risk factors for and which appear to be protective against developing childhood asthma? Briefly justify your answers.

(i) Give a brief interpretation of the odds ratio estimate for the family history variable and the corresponding confidence interval. Show how you would compute the odds ratio and its CI from the table of estimated regression coefficients and vice versa.

(j) Give a brief interpretation of the odds ratio and corresponding confidence interval for the SES variable. Show how to obtain a confidence interval for the odds ratio corresponding to a 10 point change in SES.

(k) In the city of Los Seraphim in 2008 the average pollution level was 35 thousand particles per cm^3. Find the predicted probability of developing asthma for a boy born during that year to parents who had asthma and an SES index score of 50 and who was breastfed for 6 months.

(l) The boy in part (k) is going to have a little sister born next year. Assuming the family hasn't moved, the air pollution levels in Los Seraphim have gone down by 5 thousand particles per cm^3 and that the girl is also breastfed for 6 months, find the odds ratio for her risk of developing asthma compared to her brother. You should be able to do this WITHOUT calculating the full log odds for the girl.

Problems To Turn In

(6) Ear Infections (Based on Rosner 13.66): In this problem we assess the impact of two different antibiotics on the chances a child will be cured of an ear infection after adjusting for age and whether one or both ears were infected. The variables are **clear** which indicates whether or not the infection has been cleared from both ears after 14 days of treatment, **antibiotic** which indicates which medication the child was given (1 = Ceftriaxone, 0 = Amoxicillin), **numears** which says how many ears were infected (either 1 or 2), and **age** which is divided into three categories: under 2 years old, 2-5 years old and 6 years or older. This variable is provided in two forms in the data set: first as **agegroup** with 1 = under two years old, 2 = two to five years old and 3 = six years or older; and second as a set of indicator variables for the three categories, **undertwo, twotofive** and **sixplus**.

(a) MLE Basics: Our reference point in logistic regression (as indeed in any regression!) is a model with no predictor variables:

$$\ln(\frac{p}{1-p}) = \beta_0$$

(i) Explain briefly what the interpretation of this model is and in particular what the maximum likelihood estimates of p and β_0 ought to be intuitively for this data set. (Note: It may help to get the frequency table for the **clear** variable.)

(ii) Fit the model with no predictors in STATA or SAS and check that the estimated value of β_0 matches your prediction from part (i).

(iii) (Optional Bonus) Write down the general expression for the likelihood corresponding to this model and evaluate it for the value of p that you suggested as the MLE in (i). (You can get even more credit for actually deriving the MLE by using calculus to perform the relevant maximization!) Then check that the likelihood matches the value on the STATA or SAS printout from part (ii).

Note: Parts (b)-(d) focus on a simple comparison of the clearance rates in the two medication groups without adjusting for age or severity of infection.

(b) Test of Proportions: Use a two-sample test of proportions to compare the rate of infection clearance for the two medication groups. Carefully state the null and alternative hypotheses mathematically and in words, use STATA or SAS to get the sample proportions for each group, obtain the test statistic and p-value, and explain your real-world conclusions. Show by hand how you can obtain the odds ratio for comparing children who were on Ceftriaxone to those on Amoxicillin from the sample proportions.

(c) Contingency Table Approach: Now use a contingency table (chi-squared) test to perform the same test in STATA or SAS. Confirm that your chi-squared test statistic is just the square of the Z statistic in part (b) and that your p-values match. (If you are feeling really brave you can compute the chi-squared statistic by hand too and derive the fact that the Z test and the chi-squared test are equivalent!)

(d) Now suppose that the researcher analyzes the data with a logistic regression model with Y being whether or not the ear infection cleared ($Y = 1$ for yes and $Y = 0$ for no) and X being the indicator for the antibiotic with which the child was treated ($X = 1$ for Ceftriaxone and $X = 0$ for Amoxicillin). Figure out what the estimated regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$ must be by hand based on the various values you computed in part (b) and verify your results by fitting the logistic model in STATA or SAS.

Note: For the remainder of the problem we focus on the model involving all of the predictors.

(e) Fit the model with all of the predictors in STATA or SAS, obtaining the estimates on both the logit scale and the odds ratio scale. Overall do these variables help explain how likely a child is to have their ear infections cleared in 14 days? Carefully write the null and alternative hypotheses mathematically and in words, obtain the test statistic and p-value and give your real-world conclusions using $\alpha = .05$. Verify the test statistic given in the printout for the likelihood-ratio chi-squared test by calculating it by hand from the log-likelihoods for the null model (fit in part (a)) and the full model (fit here).

(f) Do these variables explain a lot of the "variability" in how likely an ear infection is to clear? Explain briefly. What are the practical implications of this statement for treating ear infections in small children with antibiotics?

(g) Give a brief interpretation of the odds ratio for the **antibiotic** variable and its confidence interval and show how you would compute these values from the parameter estimates table (i.e. the output on the logit scale.) After adjusting for age and number of infected ears does it seem that the type of antibiotic matters and if so which one is superior? Explain briefly.

(h) Does our model show whether either antibiotic helps cure ear infections? Explain briefly.

(i) After adjusting for the other factors, does age impact the likelihood of an infection clearing within 14 days? Perform an appropriate test, writing out the null and alternative hypotheses mathematically and in words, obtain the likelihood ratio chi-squared statistic and give your real-world conclusions using $\alpha = .05$. (Note: There are several different ways to do this. If you use the indicator versions of the age group variables, picking one as the reference, you can either fit the models with and without the indicators and manually compute the likelihood ratio test statistic or else you can do the test as a follow-up contrast to the full model fit from part (e). If you use **agegroup** and specify it as a class variable you will get the test for free as part of your output. You only need to show one of the versions but make sure you understand how to do each of them!)

(7) Arteriostatistics?: A cardiologist at my favorite school, the University of Calculationally Literate Adults, is interested in the factors that lead to arteriosclerosis (hardening or blockage of the arteries, often due to the build-up of fatty plaques on the artery walls.) She is also studying medications for lowering cholesterol levels since high cholesterol is a risk factor for this disease. Her response variable is whether or not a person has arteriosclerosis ($Y = 1$ for yes and $Y = 0$ for no). Her possible predictor variables are age (in years), weight (in pounds), blood cholesterol level (measured in mg/dL) and whether the person has a family history of coronary artery disease (1 = yes, 0=no).

(a) First the investigator wants to look at the relationship between cholesterol level and arteriosclerosis. To help her visualize the relationship, bin the cholesterol variable by intervals of length 20 (i.e. 150-170 mg/dL, 179-190 mg/dL, etc.), find the proportion of subjects in each bin with arteriosclerosis, and plot your results. (Note that I have calculated the bin percentages for you in the data set but the commands are shown below in case you are interested in how I got them from the raw data. You can do the plot by hand or on the computer, whichever you prefer.) Based on your plot does it appear that there is a significant relationship in the expected direction?

(b) Formally check your conclusions from part (a) by fitting a logistic regression of disease status on cholesterol level and performing an appropriate test. Write the null and altermative hypotheses mathematically and in words and give your real-world conclusions using $\alpha = .05$. Note that you can get two different test statistics from your printout–the likelihood ratio chi-squared statistic and the Wald test Z statistic. In general the Wald test is somewhat less stable (and tends to be more conservative) than the likelihood ratio test. Is that the case here and does it make any practical difference to your conclusions? (Note: To get an adequate number of decimal places on the p-values to check this you may need to use a distribution

calculator–see the command instructions below.)

(c) Now fit the full model including age, weight, cholesterol level and family history. Does cholesterol remain a significant predictor? Explain what you think has happened and confirm your suspicions by (i) obtaining the correlations among the various predictors and (ii) refitting the model without the potential confounder.

Note: For the remainder of the problem use the second model from part (c) with the confounder variable removed!

(d) Find the probability that a 50 year old with a cholesterol level of 250 and no family history of coronary artery disease would have arteriosclerosis. You may do this either using the computer package or by hand. You only need to include one method with your homework but make sure you know how to do it both ways!

(e) Give a brief interpretation of the confidence interval for the odds ratio of the age variable. What does this interval tell you about the usefulness of the age variable in this model?

(f) Give your best estimate of and a 95% confidence interval for the odds ratio comparing the likelihood of arteriosclerosis for a person with high cholesterol (250 mg/dL) to an otherwise equivalent person with normal cholesterol (200 mg/dL). Show your work.

(g) Suppose a medication could lower your cholesterol by 50 mg/dL. The manufacturer would like to claim that this would cut your odds of arteriosclerosis in half. Based on your answer to (f) is this a reasonable claim? If not, what is the strongest claim they could make with 95% (really 97.5%) confidence?

(h) Optional Bonus: In ordinary least squares regression, the estimated change in Y associated with a 1 unit change in X is constant (this is what it means for the relationship to be linear). Similarly, if you have an indicator variable for a characteristic, the difference between people who do and do not have the characteristic is constant, regardless of the levels of the other variables (at least as long as there are no interactions!) However this is not the case for the predicted probabilities in a logistic regression. To illustrate this point use the code provided in the commands section below to create predicted probabilities of arteriosclerosis as a function of cholesterol for people who do and do not have a family history of the disease, assuming age is fixed at 50 years old, and plot these predicted probabilities on a common graph. Give a brief clinical description based on your graph of how these variables jointly affect the predicted probability of the disease.