Homework Assignment 2

Due Date: Wednesday, January 26th, 2022

Note: There are 5 problems on this assignment. The first 3 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 4-5 together with the corresponding STATA/SAS printouts to receive full credit. The assignment is due Wednesday, January 26th. You may turn it in on CCLE any time before 5:00 p.m. with no penalty. **NOTE THE UNUSUAL TURN-IN TIME!** I would like to get the solutions up before the end of the day to assist with midterm studying.

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 (HL) are adapted from the text *Applied Logistic Regression, 2nd edition* by David W. Hosmer and Stanley Lemeshow. Problems labeled (ACM) are adapted from *Computer-Aided Multivariate Analysis, 4th edition* by Abdelmonem Afifi, Virginia A. Clark and Susanne May.

Warm-up Problems

(1) Don't Drink and Derive (Interactions and Transformations): Interpreting interactions and transformations in the logistic regression context can be tricky. This problem highlights some of those issues using data from an old 201A final (the topic ordering was a bit different that year so logistic regression basics had been covered by the end of the first quarter). I'll include the complete problem for you in the midterm study set but here I've just highlighted the particular issues around the interaction and transformation terms. A summary of the original problem statement and the relevant STATA printouts are shown below.

Dr. Caroline R. Ash, a community health scientist and public health policy expert at our favorite school, the University of Calculationally Literate Adults, studies traffic accidents. In this particular study she focuses on the factors associated with traffic fatalities using data from the last 300 serious traffic accidents in the city of Los Seraphim. She has fit a logistic regression model with whether or not the accidents involved fatalities as the outcome ($Y = 1$ for yes and $Y = 0$ for no) and the predictors listed below. Note that all personal characteristics (e.g. gender, drinking status) refer to the at fault driver or vehicle.

X_1: Drinking status. $X_1 = 1$ if the driver was legally intoxicated and 0 if not.
X_2: Gender. $X_2 = 1$ for female and 0 for male.
X_3: Age, in years.
X_4: Age squared.
X_5: Speed, in miles per hour.
X_6: Dark. $X_6 = 1$ if the accident took place after sunset and 0 otherwise.
X_7: Speed by Dark interaction. $X_7 = X_5 * X_6$.
X_8, X_9: Indicators for the number of cars involved in the accident. This is divided into three categories: solo (1 car), pair (2 cars) or multi (3 or more cars). $X_8 = 1$ if it was a "solo" accident and 0 otherwise and $X_9 = 1$ if it was a "multi-car" (3 or more vehicles) accident.

```
. logit fatal drunk gender age agesq speed dark speedbydark solo multi
Logistic regression                               Number of obs   =        300
                                                  LR chi2(9)      =     138.55
                                                  Prob > chi2     =     0.0000
Log likelihood = -93.708786                       Pseudo R2       =     0.4250

------------------------------------------------------------------------------
       fatal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       drunk |   .8964695   .3912408     2.29   0.022     .1296516    1.663287
      gender |  -.9672396   .4247317    -2.28   0.023    -1.799698   -.1347807
         age |  -.1507639   .0577549    -2.61   0.009    -.2639614   -.0375664
       agesq |   .0015265   .0005447     2.80   0.005     .0004589     .002594
       speed |   .0261336   .0288708     0.91   0.365     -.030452    .0827193
        dark |   2.726086   2.314635     1.18   0.239    -1.810514    7.262686
 speedbydark |   .0183128   .0079621     2.30   0.021      .016752     .019873
        solo |   1.178641   .4137755     2.85   0.004     .3676563    1.989626
       multi |   .9519267   .3399738     2.80   0.005     .2855781    1.618275
       _cons |  -3.367394   2.605579    -1.29   0.196    -8.474236    1.739448
------------------------------------------------------------------------------
*******************************************************************************
. logistic fatal drunk gender age agesq dark speedbydark solo multi
Logistic regression                               Number of obs   =        300
                                                  LR chi2(9)      =     138.55
                                                  Prob > chi2     =     0.0000
Log likelihood = -93.708786                       Pseudo R2       =     0.4250

------------------------------------------------------------------------------
       fatal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       drunk |   2.450935   .9589057     2.29   0.022     1.138432    5.276629
      gender |   .3801309   .1614537    -2.28   0.023     .1653488    .8739075
         age |   .8600507   .0496721    -2.61   0.009     .7680032    .9631305
       agesq |   1.001528   .0005455     2.80   0.005     1.000459    1.002597
       speed |   1.026478   .0296352     0.91   0.365      .970007    1.086237
        dark |   15.27299    35.3514     1.18   0.239       .16357    1426.083
 speedbydark |   1.018481   .4218738     2.30   0.021     1.016893    1.020072
        solo |   3.249956   1.344752     2.85   0.004     1.444346    7.312801
       multi |   2.590696   1.021375     2.80   0.005     1.330531    5.044381
------------------------------------------------------------------------------
```

(a) Interactions: Give as precise an interpretation as you can of the speed by dark interaction. In particular you should carefully explain the meanings of the regression coefficients b_5, b_6 and b_7 and their corresponding odds ratios and create a rough sketch of the log odds as a function of speed and whether or not it is after dark. Explain how you chose the values you are using for the other variables and how that affects your picture, and say what your plot should look like if there is or is not an interaction. Based on this printout does it seem as if there is a significant interaction? Justify your answer with an appropriate hypothesis test and explain what the results tell you in real-world terms.

(b) Transformations: The model contains both linear and quadratic terms in age. Is there evidence that the quadratic transformation improves the fit of the model? Explain using an appropriate test and describe briefly how you could have checked for this graphically before fitting the model. What do the coefficients of age and age-squared tell you about the relationship between age and the likelihood of a fatal accident? Can you provide reasonable interpretations of the odds ratios corresponding to the age and age-squared variables? Discuss.

(2) Goodness of Fit Basics:

(a) Explain the distinction between **calibration** and **predictive accuracy**.

(b) Define **sensitivity** and **specificity** and explain what an **ROC curve** is. What do these assorted concepts have to do with evaluating goodness of fit in logistic regression?

(3) Still Gasping For Breath: This problem uses the asthma data from Warm-up Problem 5 of assignment 1. Recall that Professor Urtha Green is studying risk factors 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. Use them to answer the following questions about the logistic model using all six of the predictors.

(a) Perform an appropriate graphical check to determine whether a linear relationship adequately describes the effect of pollution on the log odds of getting asthma.

(b) Briefly describe the Hosmer-Lemeshow test for overall goodness of fit and use it to evaluate this model. Does the model appear to fit adequately?

(c) Obtain one or more pseudo-R^2 values for this data set and explain as carefully as you can what they tell you about the model performance.

(d) Compute the true positive rate (sensitivity), true negative rate (specificity) and overall fraction predicted correctly for this data set for a range of threshold values and plot them to identify a good cutoff for predicting that a child will get asthma. What are your sensitivity, specificity and overall error rate at this cutoff? Do you think the model performs well in this sense? Obtain the corresponding ROC curve for these data and compute the AUC value. Does the model perform well by this standard?

Problems To Turn In

Problems 4 and 5 involve variables from the depression data set that I used as an example in class, courtesy of the text *Practical Multivariate Analysis* (formerly *Computer-Aided Multivariate Analysis*) by Afifi, Clark and May. The complete data set is available on line in multiple formats on the UCLA IDRE web site at

https://stats.idre.ucla.edu/other/examples/pma5/

In fact, IDRE has nice worked out examples from the text in several statistical packages (including Chapter 12 on logistic regression) and you can get all the data sets for the book as text files (see the Appendix link at the bottom of the table). I have provided a subset of the data, recoded to make things easier for the models we will be fitting, as part of this week's homework data file. The variables included in this reduced version are **sex** (1 = female, 0 = male), **drink** (1 for a regular drinker and 0 for not), **depressed** (1 = yes and 0 = no), **CESD** which is a continuous measure of depression on a scale of 0-60 with higher being worse, **income** in thousands of dollars, **age** in years and **health status** which is a four level variable with 0 = Excellent, 1 = Good, 2 = Fair and 3 = Poor.

(4) Drink is a Downer: (ACM 12.9) For this problem we use drinking status as the outcome and are interested in sex (treated as binary) and depression (either categorical or continuous) as predictors.

(a) Fit a simple logistic regression of drink on sexr and use it to find the odds of being a regular drinker for women and men and the corresponding odds ratio. Does there appear to be a sex difference?

(b) Repeat part (a) but do the calculations separately for people who are depressed and people who are not depressed and compare the odds ratios for the two groups. (Note: for this part you do not need to give the individual odds by sex; just focus on the odds ratios.)

(c) What do the results from part (b) suggest about whether there is an interaction between sex and depression? Explain carefully what such an interaction would mean. Then fit an appropriate model and test whether the interaction is significant. (The depbysex interaction variable has been provided for your convenience.) Do the results confirm your theory? If not, what do you think might have happened?

(d) Instead of using the case indicator one could instead use the continuous depression rating, CESD. Fit a logistic regression of drink on sex, CESD and their interaction. (the cesdbysex interaction has been provided for your convenience.) Give as precise an interpretation as you can of this interaction. In particular you should carefully explain the meanings of the regression coefficients b_1, b_2 and b_3 and their corresponding odds ratios and create a rough sketch of the log odds as a function of sex and depression score. Based on your model does it seem as if there is a significant interaction? Are the results more or less significant than those in part (c)? Explain what the results tell you in real-world terms and give a possible explanation for why the test has come out the way it has.

(5) Homework is Depressing: (ACM) For this problem we will model depression status (yes or no) as a function of sex, age, income and general health (but not how much statistics homework one has!)

(a) Fit the logistic regression model with all 4 predictors, treating health as a continuous variable. Which of these variables appear to be associated with depression status? Are all the relationships in the expected direction? Discuss briefly.

(b) Perform an appropriate graphical check to determine whether a linear term adequately describes the relationship between health status and depression. Note: Because the health status variable only has four values you do not need to go through the tedious process of calculating quartiles–you simply need to rerun the model using an appropriate set of indicator variables (which I have provided for you!) In addition to

creating the appropriate plot, discuss whether the model using the indicators appears superior to the original model with the linear term. Is it legitimate to formally test this? If the relationship appears non-linear suggest and test an appropriate transformation of the health variable.

Note: For the rest of the problem use the model from part (a).

(c) Briefly describe the Hosmer-Lemeshow test for overall goodness of fit and use it to evaluate this model. Does the model appear to fit adequately?

(d) Obtain one or more pseudo-R^2 values for this data set and explain as carefully as you can what they tell you about the model performance.

(e) Compute (i) the true positive rate (sensitivity) and (ii) the true negative rate (specificity)–or if you prefer the false negative rate for this data set for a range of threshold values and plot them to identify a good cutoff for predicting that a person will be depressed. What are your sensitivity, specificity and overall error rate at this cutoff? Do you think the model performs well in this sense?

(f) Obtain the ROC curve to go with your calculations from part (e) and the corresponding AUC value. Does the model perform well by this standard?

STATA and SAS Commands

For this assignment you need to be able to run standard logistic regression models and obtain assorted follow-up statistics and tests. The key commands are defined below and as usual Nadia will go over them in lab.

Commands in STATA

Logistic Regression: As we saw on the last assignment, the basic logistic regression command is **logit Y X1 X2 X3...**. The model fit can be followed by a number of additional commands:

Goodness of Fit: Specifically, the major thing you need to know for this assignment is how to obtain various goodness of fit statistics. These include sensitivity and specificity values and ROC curves as well as running the Hosmer-Lemeshow test. Here are some of the corresponding commands which can be typed after fitting the main model:

estat class, cutoff(threshold) produces the classification table (predicted values versus observed values) for the threshold specified using the cutoff option. The default if you don't include the cutoff option is cutoff(.5).

estat gof, group(k) table performs a Hosmer-Lemeshow goodness of fit test dividing the data set using k quantiles. The usual choice, as we learned in class, is group(10). If you do not use the group option then STATA will perform the Pearson goodness of fit test which is fine if there are only a small number of possible combinations of predictors (e.g. no continuous X variables) but is otherwise not a good idea. If you want to see the table of obeserved and expected counts that go with the test you can also add the table option after the group option.

lroc produces the ROC curve and calculates the area under the curve.

lsens graphs sensitivity and specificity versus probability cutoff. You can also use it to store the grid of thresholds, the sensitivity values and the specificity values by using the options **genprob(varnameprob)**, **gensens(varnamesens)** and **genspec(varnamespec)** after a comma.

Used after fitting the model of interest, **predict** produces predicted probabilities. (You can also use it to get residuals, influence statistics and the like. I have not covered these in class as they are generally harder to interpret than in OLS regression but I've included the commands in case anyone is interested.) You type "predict" followed by the name of the variable where you want to store the results, then a comma and the code for the type of thing you want to predict (probability, log odds, etc.) The default, if you don't specify an option after the comma, is the predicted rpobabilities. For example:

predict myprobs puts predicted probabilities in the variable myprobs
predict myfits, xb produces the fitted values–i.e. log odds
predict myses, stdp produces standard errors for the fitted values. (This is how you would get a CI for a predicted probability.)
predict mydeltabeta, dbeta produces delta beta influence values.
predict mydevianceresids, dev produces deviance residuals.
predict mydeltachisquared, dx2 produces the delta chi-squared fit influence values.
predict myresids, residuals produces the Pearson residuals.

If you want some nice examples of all of the various logistic models in STATA, UCLA's academic technology services has great on-line STATA pages. Go to

https://stats.idre.ucla.edu/stata/

Commands in SAS

Logistic Regression: As we learned on HW1, logistic regression models are fit in SAS using **proc logistic** which works very much the same way as OLS regression using **proc reg**. You specify your data set, the class (i.e. categorical) variables (which can now includes your outcome though this is optional–SAS knows if you are doing a logistic model that the outcome is categorical), and a model statement connecting the outcome to the predictors. You can also have contrast or test statements that perform additional tests about your parameters. **Very Important Note:** SAS proc logistic for some reason defaults to modeling the probability that your outcome is 0 rather than 1. To fix this you need to tell it to treat your outcome in **descending order**, so that the outcome value 1 is the most important. You can adjust this using the **desc** option in the data statement.

Add-On Options: There are a number of options you can specify as part of the model statement in **proc logistic** after a /. For instance, **ctable** produces the classification table using a variety of cutoffs. The default is to disply the classification for a range of probabilities from the smallest to the largest estimated probability incrementing by .02. If you add the **pprob=** command you can specify the thresholds you'd like to see. The option **Rsquare** displays generalized r-squared values. The option **clparm=wald** produces Wald confidence intervals for the regression parameters. For example, using the depression data of Problem 5 our syntax would be

```
proc logistic data = tmp1.hw2 desc;
model depressed= sex age income health/ctable pprob = (.3, .5 to .8 by .1)
                                        Rsquare
                                        clparm=wald;
run;
```

This will give you the table for cutoffs of .3, .5 , .6, .7 and .8., an R-squared value and the Wald CIs.

If you want some nice examples of all of the various logistic models in SAS, UCLA's academic technology services has great on-line SAS pages. Go to

https://stats.idre.ucla.edu/sas/

and see the section on logistic regression for examples relevant to this assignment.

Goodness of Fit: There are additional options you can add to the model statement to produce goodness of fit statistics. For example, **lackfit** performs the Hosmer-Lemeshow goodness of fit test using 10 groups. To produce the sensitivity, specificity and an ROC curve you need to save the necessary information to another SAS data file. The model statement option for doing this is **outroc=**. Similarly, you can use the **output=** and **score** statements to save the predicted probabilities, their standard errors and so on. SAS will also save various regression diagnostics. See the SAS help files for more details. The syntax below will produce the H-L test and save the sensitivity/specificity/ROC information to a SAS data file called myroc for the depression model of Problem 5:

```
proc logistic data = tmp1.hw2 desc;
model depressed= sex age income health/lackfit outroc=myroc;
run;
```

Note that the file "myroc" will be in the termporary work directory. If you want to keep it permanently you can instead say outcoc = tmp1.myroc and it will put it in your current directory. The "myroc" file will contain, among other things, the variables _SENSIT_ and _1MSPEC_ which are respectively the sensitivity and one minus the specificity at various cutoffs, specified in the variable _PROB_. These are what you need to produce the ROC curve and sensitivity/specificity plots. For instance you could use the **gplot command** procedure as follows (I have added a few mild formatting options):

```
proc gplot data=roc1;
   title 'ROC Curve';
   plot _sensit_*_1mspec_=1 / vaxis=0 to 1 by .1 cframe=ligr;
run;
```