Homework Assignment 3

Due Date: Wednesday, February 9th, 2022

Note: There are 7 problems on this assignment. The first 4 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 5-7 together with the corresponding STATA/SAS printouts to receive full credit. The assignment is due Wedbesday, Febnruary 9th by the end of the day on CCLE.

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 (ACM) are adapted from *Computer-Aided Multivariate Analysis, 4th edition* by Abdelmonem Afifi, Virginia A. Clark and Susanne May.

Warm-up Problems

1) Probit Regression Basics:

(a) Explain what the distributions, link functions and systematic components are for the probit model.

(b) Give examples of circumstances where you might want to use this models.

(2) Multinomial and Ordinal Logistic Regression Basics:

(a) Explain what the three basic GLM components are for a multinomial logistic regression.

(b) Explain how the ideas of odds and odds ratios are extended for multinomial logistic regression.

(c) Explain how multinomial logistic regression can be adapted if there is a natural ordering to the outcome categories and what the proportional odds assumption means.

(3) Gasping For Breath Some More: This problem continues the Asthma warm-up problem from assignments 1 and 2. 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, sex (1 = Female, 0 = Male). The data are given in the accompanying file.

(a) On HWs 1 and 2 you fit a standard logistic model using all the predictors. Suppose Professor Green informs you that the study actually oversampled children living in urban environments and/or having a family history to make sure that there were enough asthma cases to get good estimates of the effects of the various predictors. She says the true rate of childhood asthma cases in the population is closer to 5%. Explain what you would have to do to adjust the model from Assignment 3 to get good predicted probabilities under this assumption.

(b) Suppose Professor Green had told you that the data in her sample came from 500 sibling pairs. Explain how this would affect your analyses. (You do not need to actually run a model–just explain how you would do so, what variables you would include, and why it would be important to take account of the pairing.) For parts (c)-(d) go back to assuming that the data are a representative sample of children from this region.

(c) Fit a probit regression model to these data and provide an interpretation of the coefficients for the family history and pollution variables on the probit scale. Is there an easy interpretation on the probability scale?

(d) Find the predicted probability of developing asthma for a boy born in the city of Los Seraphim where the pollution level is 35 thousand cm^3 to parents who had asthma and an SES index score of 50 and who was breastfed for 6 months based on your probit model. Compare your answer to what you found in Problem 5k on HW1.

(e) Compare the log likelihoods for your probit model and the standard logistic model from HWs 1 and 2. Does one appear to be better than the other? Do you think the difference is substantial?

(4) The Accidental Statistician: The problem continues with the data set from the 2011 201a final on drinking and traffic accidents which we saw as Warm-up Problem 1 on HW2. Recall that there were three types of accidents: single car (solo) accidents, two-car accidents and a three-or-more (multi) car accidents. We now want to understand what factors differentiate among the different accident types. I have recoded the data so that our outcome variable is "type" which has three levels: 0 = a two-car accident, 1 = a solo accident and 2 = a multi-car accident. As predictors we consider the following subset of the original variables: whether or not the accident involved a fatality (fatal = 1 for yes and 0 for no), whether the accident took place at night (dark = 1 for yes and 0 for no), the sex of the driver (1=Female and 0=Male) and the speed at which the cars were travelling (in miles per hour.) The data are given in the accompanying file.

(a) Fit a multinomial logistic regression for these data using the two-car accident as the reference category. Is this model overall significant? Explain briefly.

(b) Give a careful interpretation of the coefficients for the coefficients of "fatal" and "speed" for each of the two model components (solo versus two-car and multi versus two-car).

(c) Which factors appear to be important for differentiating among the three accident types? You should explain both in terms of the p-values in the individual model components and by testing for the overall significance of the variables. Make sure you are able to write down the null and alternative hypotheses you are testing, both matehmatically and in words.

(d) For the variables that are significant, does it appear that their coefficients differ across the two compo-

nents of the model? Check by performing an appropriate set of tests. Again, be sure you know how to write out the formal hypotheses.

(e) Based on your answers to (a)-(d) describe as carefully as you can what sort of conditions are more likely to lead to what sort of accident.

(f) Find the predicted probability of each type of accident given you know that the at fault car was driven by a man at 90 miles per hour after dark and that there were no fatalities involved.

For the remainder of the problem, suppose that you consider the outcome categories to be ordered in terms of seriousness (two-car crash is the least serious, then a solo crash, then a multi-car crash).

(g) Fit an appropriate model for this situation using the same set of predictors as in parts (a)-(e). Is the model overall significant?

(h) Give careful interpretations of the coefficients of fatal, dark and speed and their corresponding odds ratios. Which of these variables appear to be significant predictors?

(i) Find the predicted probability of each type of accident based on this model given you know that the at fault car was driven by a man at 90 miles per hour after dark and that there were no fatalities involved.

(j) Compare and contrast your results with those from the multinomial logistic model.

Problems To Turn In

(5) Cancer Conundrums: At a chemical plant workers are exposed to compounds which may cause skin cancer. In order to evaluate the effect of exposure on risk of cancer, 50 plant workers who developed skin cancer are identified and are matched on age and sex with workers who were employed during the same time period but did not develop skin cancer. The variables for this problem are cancer (1 = yes, 0 = no), exposure (a continuous score with higher being more exposure) and pairid which identifies how the employees were matched up. Our goal is to evaluate the impact of exposure.

(a) Fit a conditional logistic regression to these data. Does there appear to be a significant effect of exposure? Give a careful interpretation of the odds ratio for the exposure variable.

(b) Now rerun the model ignoring the pairing–i.e supposing that we were simply performing a case control study–and obtain the odds ratio for exposure. Does the effect of exposure seem stronger or weaker than in the conditional logistic model? Is the variable more or less significant? Does this fit with your expectations of how these models work? Why or why not?

(c) Suppose we really had been doing a case-control rather than a matched pair study but were told that the actual rate of skin cancer in the general population was only 5%. How would we adjust the model from part (b) so that it would give appropriate estimates of the predicted probabilities? How much would this change the predicted probability for a person with an exposure score of 20?

(6) I Await Your Response: Researchers are pioneering a new medication and are interested in knowing how high the dose needs to be before patients will respond. There is also some concern that too high a dose may actually start to be toxic and decrease the chance of a response. They have tried the medication at different doses on a sample of 500 patients and recorded whether or not the patients had a positive response to treatment. The data are given in the accompanying file.

(a) Fit a probit regression model with response as the outcome and dose as the predictor and say whether there is a significant dose-response relationship. Explain why a probit model may be a natural choice for these data.

(b) Give an interpretation of the intercept in this model. (Note–you may find this easier to do if you transform back to the probability scale. Recall that you can use the **display normal(z)** command to get the probability that a standard normal random variable is less than a particular value, z.)

(c) Find the dose at which the probability of a response is 50%. Explain your reasoning.

(d) Perform a Hosmer-Lemeshow goodness of fit test for this model. Does the model appear to be well-calibrated? Explain briefly.

(e) Now rerun the model including both dose and dose squared as predictors. (The variable dosesq has been included for your convenience.) Is the curvilinear model in dose an improvement? Explain briefly. Explain what the signs of the dose and dose-squared terms tell you about the dose-response relationship. Do they confirm the researchers' theories? Is the calibration of this model adequate?

(f) Based on the fitted model from part (e) identify the optimal dose of this medication and the response rate at that optimal dose. (Note: You may find it useful to recall that the peak of the parabola $ax^2 + bx + c$ occurs at $x = -b/2a$.)

(g) Obtain the fitted probabilities for your model from part (e). Then rerun the model using a logit link and obtain the corresponding fitted probabilities. Do the two models seem very different? Discuss this by

(i) Comparing the predicted probabilities of response when the person receives no medication.
(ii) Comparing the optimal doses.
(iii) Comparing the log likelihoods for the two models (make sure you understand why this is a reasonable comparison!) Say which model (if either) appears better by this criterion.
(iv) Calculating the correlation between the predicted probabilities for the two models.
(v) Obtaining a scatterplot showing the relationship between the two sets of predicted probabilities.

(7) Healthy and Happy? (ACM 12.22): This problem uses the depression data set which was featured in HW2, Problems 4 and 5. However, we are now going to reverse our set-up and evaluate health status as a function of age, income and depression.

(a) Fit a multinomial logistic regression for these data using the "excellent health" category as the reference. Is this model overall significant? Explain briefly.

(b) Give a careful interpretation of the coefficients for the age, income and depression variables for each of the three model components (good, fair or poor vs excellent health). What would be the effect of an increase in age of 10 years on the odds of poor health? What would be the effect of an increase in income of $5000 on the odds of fair health? Show your work.

(c) Which factors appear to be important for differentiating among the levels of health status? You should explain both in terms of the p-values in the individual model components and by testing for the overall significance of the variables. For one of the latter tests write the hypotheses mathematically and in words and give the test statistic. For the other tests you may simply give the p-values and your conclusions.

(d) For the variables that are significant, does it appear that their coefficients differ across the three components of the model? Check by performing an appropriate set of tests. For one of these tests write out

the details of the hypotheses and the test statistic. For the others you may simply give your p-values and conclusions.

(e) Based on your answers to (a)-(d) describe as carefully as you can the relationships among the variables in this model.

(f) Find the predicted probability of each level of health status for a 50 year old who makes $50,000 per year and is depressed.

(g) The values for the health status outcome have a natural ordering which our multinomial logit model ignored. Fit an appropriate model taking this into account using the same set of predictors as in parts (a)-(f). Is the model overall significant?

(h) Give careful interpretations of the coefficients of age, income and depression and their corresponding odds ratios for the model in part (g). Which of these variables appear to be significant predictors?

(i) Find the predicted probability of each level of health status for a 50 year old who makes $50,000 per year and is depressed based on the model from part (g).

(j) Compare and contrast your results with those from the multinomial logistic model. Do you think one of the models fits better than the other? Explain.

(k) Do you think the proportional odds assumption has been met? Justify your answer conceptually and with an appropriate test.

STATA and SAS Commands

For this assignment you need to be able to fit conditional, multinomial and ordinal logistic models as well as probit regression models and obtain assorted follow-up statistics and plots. The necessary commands are given below.

Commands in STATA

For this assignment you need to be able to run variants of the standard logistic model and obtain assorted follow-up statistics and tests. Fortunately all the different versions of logistic regression are fit basically the same way. The key commands are defined below.

(1) Logistic Regression Variants: As we saw on previous assignments, the basic logistic regression command is **logit Y X1 X2 X3...**. To get other logistic models we just add their starting letter in front of "logit", producing **clogit** for conditional logistic regression, **mlogit** for multinomial logistic regression and **ologit** for ordinal logistic regression. Each of these functions differs slightly in terms of how it expects the data to be formatted and what follow-up commands you use to perform the tests of interest.

Conditional Logistic Regression: The basic command is **clogit Y X1 X2..., group(matchvar)** where Y is the outcome variable, X1, X2, etc. are the predictor variables (things on which the subjects are not matched, if you have perfectly matched covariate values) and matchvar is the variable that tells you which subjects are matched with each other–it has a separate value for each matched pair or group. For example, for Problem 5 our outcome is cancer, our predictor is exposure, and we have a grouping variable which I named pairid. It takes on the values 1-50. Each of these values occurs twice: once for a case of cancer and once for that case's matched control. The command is

clogit cancer exposure, group(pairid)

Multinomial Logistic: STATA will accept almost any sort of variable as the outcome for a multinomial logistic regression. It takes as the number of categories the number of different values it finds in the variable. This is fine as long as you don't mistakenly feed it a continuous variable where most of the values only occur once! By default, STATA uses the value that occurs most frequently as the reference category. If you want to select the reference category yourself you use the **baseoutcome(value)** option. For example, suppose my Y variable is **health** as in turn-in problem 7. The most common value is 0, excellent health. This would be STATA's default and probably makes sense since we are trying to figure out what are risk factors for bad health. To fit the model with this reference we would type

mlogit health age income depressed

However, if we wanted to use poor health, which is coded as 3, as the reference we would type

mlogit health age income depressed, baseoutcome(3)

With multinomial logistic regression there are two major kinds of follow-up tests you might wish to perform that are not given on the main printout. One is to test whether a predictor variable, **overall** makes a significant contribution to the model. There are two basic ways to do this. One is to perform a likelihood ratio chi-squared test using the **lrtest** command as we did on HW3 for standard logistic. You fit the model with and without the variable of interest, saving the results each time, and then you use lrtest to compare the two models. (Of course you could also do this by hand using the two log likelihoods!) For example, to check whether the variable age is important in the model for problem 7 we would use the sequence:

mlogit health age income depressed
estimates store FULL
mlogit health income depressed
estimates store NOAGE
lrtest NOAGE FULL

The other option is to use the **test** command after fitting the full model to obtain a Wald test. This is really easy. You just type **test** followed by the variable name, e.g.

mlogit health age income depressed
test age

Finally, you might wish to test whether the effects of a particular variable were the same across levels of the model. For this you also use the **test** command but you need specify both which variable and which levels you want to compare. The basic syntax is

test [1 = 2 =...]: varname

For instance, if we wanted to test whether the effect of age was the same across all the levels in the model above after fitting the full model we would type

test [1=2=3]: age

The UCLA Academic Technology Services website has a nice example of a multinomia logistic regression in STATA at "https://stats.idre.ucla.edu/stata/dae/multinomiallogistic-regression/"

Ordinal Logistic Regression: The basic command is **ologit Y X1 X2....** For ordinal logistic regression the outcome variable needs to be numeric but the actual values don't matter. STATA simply assumes that the larger values correspond to the "higher" outcomes. There is something you need to be careful about though with interpreting the coefficients. In the conventional parameterization for ordinal logistic regresison, the model looks at the odds of Y being lower to Y being higher. Thus you would think that a positive regression coefficient would correspond to being more likely to have a lower value of the outcome. However STATA parameterizes its log odds as

$\beta_0 - \beta_1 X_1 - \beta_2 X_2 - ...$

As a result a positive coefficient actually corresponds to a higher probability of a higher value. In the end this may seem like the natural thing but it is important to be aware of how the package you are using parameterizes the model you are fitting! All the usual test commands from ordinary logistic regression apply to ordinal logistic regression.

Finally, the usual tests for the propotional odds assumption are not part of the default ologit post-estimation commands but there are several user-written functions that will do them which are very easy to import. To do so, go into STATA and select help/stata command. Type in either "oparallel" or "spost13" (these are two different useful functions) and it will pop up a search list of things that might be relevant. It should be easy to spot the ones that represent the actual packages. Click on the link to the respective site and when you are there click "install." It will give you a message saying the function has been loaded. Then, after you have run your ordinal model you type (respectively) as a command line

oparallel–this will give you an output table with a bunch of different tests (of which the Brant and Score tests are probably the most common)

brant, detail–this one does only the Brant test but it has areally nice feature which is that it gives you the coefficients for the predictors that you would get if you did the logistic models corresponding to each of the splits, side by side, so you can see how well they match each other.

The UCLA Academic Technology Services website has a nice example of ordinal logistic regression in STATA at "https://stats.idre.ucla.edu/stata/dae/ordinal-logistic-regression/"

(2) Probit Regression: To fit a probit regression in STATA you use the **probit** command. Pretty much everything about the probit platform works the same way as for logistic regression. The basic command is

probit Y X1 X2 X3

For instance, for warm-up problem 3 the command would be

probit asthma urban pollution ses breastfed famhist sexQ3.

To obtain the predicted probabilities based on the probit model you follow the main fit with the **predict** command, providing the name for the variable where you would like to store the results. For warm-up problem 3 if we want to store the results in the variable asthmaprobs we would type

predict asthmaprobs

immediately after fitting the probit model. The default for the predict command is to produce estimated probabilities. However, just as was described for HW2 for the logit command you can also obtain fitted values, residuals, influence statistics and so on. See the HW2 commands section for details.

You can also obtain all of the various goodness of fit statistics for the probit model. Typing **lroc** produces the ROC curve, typiing **lsens** produces the sensitivity and specificity plots and typing **estat gof, group(k)** produces the Hosmer-Lemeshow goodness of fit test based on k groups (usually k=10).

(3) Miscellaneous Useful Commands: There are some old commands you may find it helpful to recall for this assignment.

(a) Correlation: To obtain correlations we use the **cor** command. For instance, in Problem 6g I ask you to obtain the correlation between the predicted probabilities from a probit model and a logit model using the same data. Assuming you have stored the predicted probabilities in the variables probitprobs and logitprobs you would type

cor probitprobs logitprobs

(b) Normal Probabilities: To obtain the probability of a standard normal probability being less than a particular value, z^* we use the following command:

display normal(z^***)**

Commands in SAS

Conditional logistic (for problem 5): All you need to do in SAS to fit a conditional logistic regression is to add a **strata** command where you tell it what variable identifies the pairs or matched groups. The syntax for Problem 5 becomes

```
proc logistic data = tmp1.hw3 desc;
model cancer = exposure;
strata = pairid;
run;
```

In general, SAS is pretty good about telling you what ordering or reference categories it is using. You can use the "descending" option (in the proc line) as I did above to reverse the ordering of the categories but if you don't the printout should tell you what it did. There are also **ref=** and **event=** options which would let you control which is the reference group explicitly as part of the model statement. For example the above syntax could be rewritten as

```
proc logistic data = tmp1.hw3;
model cancer(event='1') = exposure;
strata pairid;
run;
```

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. For the conditional logistic model, go to the FAQ page below and go to the logistic regression section and click on the appropriate example:

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

(2) Multinomial and Ordinal Logistic Regression:

For multinomial and ordinal logistic regression, SAS continues to use **proc logistic** but you have to add a few extra options to tell it which sort of model you are fitting:

Multinomial Logistic: The special add-ons here are that you have to tell SAS that your outcome is a categorical variable (done in the **class** statement) and what category you want to use as the reference (also done within the class statement using **(ref = "**")**). You also have to specify that the link function is a generalized logit, done by adding the option **link = glogit** to the model statement. The commands for turn-in Problem 7(a), with "Excellent Health" (0) as the reference category would be

```
proc logistic data = tmp1.hw3;
class health (reference = ''0'');
model health = age income depressed/link = glogit;
run;
```

SAS automatically generates tests for each of the variables overall (as part of it's "Type III Analysis of Effects" table) as well as parameter estimates, Wald tests and confidence intervals for each variable at each level of the model (as part of its "Analysis of Maximum Likelihood Estimates" table) and of course the corresponding odds ratios and confidence intervals. You can use follow-up **test** statements to compare coefficients across levels of the model. To do this you need to know how SAS is internally naming these different objects. To get this information you can include the option **outest = filename** as part of your proc logistic statement, telling it to store the results of the model fit in "filename". Then to look at the parameter labels you use **proc transpose** and **proc print**. For the above model we would type

```
proc logistic data = tmp1.hw3 outest = mlogit_health;
class health (reference = ''0'');
model health = age income depressed/link = glogit;
run;

proc transpose data = mlogit_health;
```

```
run;
proc print noobs;
run;
```

Using "noobs" in the print statement supresses printing lots of numerical values that you don't want! From this you get a list of SAS' internal labels. For instance, in the above model we would find that the coefficient of age at the level comparing good health to excellent health is labeled age_1 while the coefficient of age comparing fair health to excellent health is age_2 and for poor health age_3. To test whether these coefficients are equal we then rerun our logistic model as follows:

```
proc logistic data = tmp1.hw3;
class health (reference = ''0'');
model health = age income depressed/link = glogit;
agecoeffeq: test age_1 = age_2 = age_3;
run;
```

The expression "agecoeffeq" is just a label we give to this particular test so that SAS can list it on the printout. This is useful if you include multiple test statements.

Note: If you have categorical predictors with more than 2 levels, you should include them in the class statement too with the reference category specified and you should also tell SAS as part of the class statement that you want what is called "dummy coding," meaning that you want indicators for all the non-reference categories. This is done by adding **param = ref;** at the end of your class statement. Dummy coding is the coding we have generally used. SAS defaults to something called "effects coding" which makes it a little harder to interpret the coefficients. If you have a multicategory predictor SAS will generate the dummy variables for you based on the commands above and will also automatically give you a test of whether those categories as a group are useful which can save time. However of course you can also create the dummy variables yourself. This isn't needed for any of the problems on this assignment but may be handy for your projects. We could have used this technique for treating health as a categorical predictor of depression or accident type as a categorical predictor of fatality on HW3.

The UCLA Academic Technology Services website has a nice example of a multinomial logistic regression in SAS at "https://stats.idre.ucla.edu/sas/dae/multinomiallogistic-regression"

Ordinal Logistic Regression: For ordinal logistic regression things are even easier as you don't need the glogit link. You just need to tell SAS that your outcome variable is categorical and which order you want the categories in. The default is numerical lowest to highest (with lowest as the reference) or alphabetical if the variable is written as characters rather than numbers. You can as usual use the **descending** option to change the order. For the problem above since we want Excellent Health = 0 as the reference we would just have

```
proc logistic data = tmp1.hw3;
class health;
model health = age income depressed;
run;
```

The UCLA Academic Technology Services website has a nice example of ordinal logistic regression in SAS at "https://stats.idre.ucla.edu/sas/dae/ordinal-logistic-regression/"

(3) Probit Regression: There are several ways to fit probit models in SAS, using **proc logistic, proc probit** and **proc genmod**. I illustrate the procedure using **proc logistic** below which handles the category ordering a little more easily. Proc probit has some follow-up graphics that proc logistic doesn't have. The important thing is that in proc logistic the default link is the logit but we want a probit link so we have to

specify this in the model statement. The **descending** option is used to force SAS to treat 0 as the reference outcome for the model. Otherwise it computes the probability of 0 instead of the probability of a 1 which is the reverse of most programs. The commands for the asthma data of warm-up problem 3 are as follows:

```
proc logistic data = tmp1.hw3 descending;
model asthma =urban pollution ses breastfed famhistQ3 sexQ3/link = probit;
run;
```

(4) Correlations: Correlations in SAS are obtained using **proc corr**. The basic command structure is

```
proc corr data = tmp1.hw3;
var var1 var2 var3;
with var4 var5 var6;
run;
```

This produces all pairwise correlations of the variables in the **var** command line with the variables in the **with** command line. If you do not have a **with** command SAS just produces the complete set of pairwise correlations between all variables in the **var** line.