February 9th, 2022

Homework Assignment 4

Due Date: Wednesday, February 16th, 2022

Note: This is a somewhat shorter assignment, designed to help you practice the material on count models prior to Midterm 2. **Note the turn-in time of 5:00 pm on Wednesday, February 16th to CCLE.** This is designed to let me post the solutions that day in time for evening 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 ACM are adapted from the text *Computer Aided Multivariate Analysis* by Afifi, Clark and May. Problems labeled (AA) are adapted from *An Introduction to Categorical Data Analysis* by Alan Agresti. Problems labeled (ATS) use data sets from UCLA's Academic Technology Services web site.

Warm-up Problems

(1) Poisson and Negative Binomial Regression Basics:

(a) Explain what the distributions, link functions and systematic components are for Poisson and Negative Binomial GLMs.

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

(c) Explain the concepts of overdispersion and zero-inflation, give examples of how each of these situations might arise in a Poisson or Negative Binomial Model, and what you could do to adjust for them.

(2) Munching (Computer) Chips (AA 4.6 and 4.8): An experimenter analyzes imperfection rates for two processes used to fabricate silicon waters for computer chips. Each treatment is applied to 10 chips and the number of imperfections are recorded. The thickness of the silicon coating is also reported (0 = thin and 1 = thick). Use the data to answer the following questions.

(a) Fit a simple Poisson regression using treatment as the predictor. Does there appear to be a significant difference between the two processes? Show two ways you could perform this test.

(b) Give a brief interpretation of the regression coefficients and their confidence intervals, both on the log scale and on the mean scale.

(c) Now add the thickness of the silicon coating as a predictor and rerun the model. Is the new model a significant improvement over the previous model? Describe as carefully as you can the effect of the coating thickness on the number of imperfections. What are you assuming about the joint effect of process and coating thickness by fitting the model this way?

(3) More Sports Fanatics (AA 4.14): British fans are mad (literally!) about their soccer, even more than New Zealanders are about rugby. This data set gives the number of fans in attendance (in thousands) and the total number of arrests in the 1987-1988 season for soccer teams in the second division of the British football league.

(a) Fit a Poisson regression model for number of arrests using attendance as an offset variable. Explain (i) why it is important to use an offset variable here and (ii) what the interpretation of the intercept from the model is. (Note that there are no predictors here other than the offset!)

(b) Plot arrests vs attendance and overlay the prediction equation. Obtain residuals from the model and use them to identify teams that had much larger or much smaller than expected number of arrests.

(4) Camping Data (ATS): This problem goes in depth into the fishing example I used in class. The outcome variable is the number of fish caught (**numfish**) by 250 groups of people who went to a wildlilfe park. The predictors are the number of people in the group (**persons**), the number of children in the group (**children**) and whether or not they were camping (**camper** = 1 for yes and 0 for no.) Use the data to answer the following questions.

(a) Fit a Poisson regression for the number of fish caught with number of people, number of children and whether or not the party was camping as predictors.

(b) Give careful interpretations of the coefficients of persons, children and camper status and their confidence intervals (i) on the log scale and (ii) on the mean scale. Which of these variables appear to be significant predictors?

(c) Plot the Pearson residuals versus the fitted values for this model and also obtain the deviance and Pearson goodness of fit tests. Do you think it is reasonable to use these measures here? Are there any obvious outliers? Does the model seem to be well calibrated? If not can you identify where the misfit is occurring?

(d) Now we evaluate the issue of over-dispersion. Suggest some plausible reasons why there might be over-dispersion in this data set and then check for it in the following ways:

(1) Using basic descriptive statistics obtain the mean and variance of the **numfish** variable. Do this for various subgroups of the data set (e.g. camping or not, with children or not, etc.) What do these statistics suggest about over-dispersion.
(2) Calculate the approximate dispersion factor based on the Pearson chi-squared goodness of fit statistic.
(3) Explain what your plot of residuals from part (c) suggests about over-dispersion.

(e) Rerun the model using a negative binomial regression. Does this model confirm your conclusions about over-dispersion from part (d)? Explain briefly.

(f) Give brief interpretations of the coefficients from the negative binomial regression. Are your conclusions (as to magnitude, direction and significance) of the effects of the predictors any different from those in part (b)?

In the next part of the problem we consider the issue of zero inflation.

(g) Explain intuitively why we might expect zero-inflation in the context of this problem. Do you expect any of these predictors to be especially predictive of who would be a "certain 0?".

(h) Continuing with your idea of descriptive statistics from part (d) calculate the expected number of 0's for various subgroups of the data set assuming a Poisson distribution and then examine the actual number of 0's. Do we seem to have a zero-inflation problem? (Note: You may find it helpful to remember that for a Poisson random variable the probability of k events is $P(Y = k) = \mu^k e^{-\mu}/k!$).

(i) Fit a zero-inflated Poisson model to the data using persons, children and camper as the predictors for the Poisson component and experimenting with various variables the zero-inflation component of the model. Do you think this model is an improvement over the standard Poisson regression? Explain briefly.

(j) Give brief interpretations of the coefficients for each component of the zero-inflated Poisson model. Which variables seem to be significant in each part of the model? Explain as carefully as you can what this model suggests about how these variables affect numbers of fish caught.

(k) Now fit a zero-inflated negative binomial model using the same model set-up as in part (j). Is this model superior to the zero-inflated Poisson model of part (j)? Does it look superior to the plain negative binomial model of part (e)? Explain briefly in each instance. What does that tell you about the issues of overdispersion and zero-inflation? Do any of your conceptual conclusions from part (j) change with this model?

Problems To Turn In

(5) I Wish I Could Play Hookey From 201B (ACM Problems 12.25 and 12.26):
This problem examines factors associated with adolescent students being absent from school without a valid excuse. For each of n = 252 students, the number of times absent in the last month was recorded (abbreviated **nhookey** for number of times the student "played hookey") along with possible predictors including age, sex (1 = male and 0 = female) and the degree to which the student liked school (rated from 1 = liked it very much to 5 = disliked it very much). Our goal in this problem is to build a model for school absences.

(a) Fit the Poisson model with no predictors and explain as carefully as you can the meaning of the estimated intercept. (Note–you may find it helpful to exponentiate it!)

(b) Plot mean number of days absent as a function of (i) likeschool and (ii) age. Do there appear to be significant relationships between these variables and numbers of days absent? Do the relationships appear linear?

(c) Fit a Poisson regression for the number of days the adolescents in the sample were absent from school with sex, age and how much the student liked school as predictors. Obtain two versions–one in which **likeschool** is treated as continuous and one in which it is treated as categorical. Which version of the model do you think fits better and why?

From here on out, for simplicity, treat the likeschool variable as continuous rather than categorical.

(d) Give careful interpretations of the coefficients of the age, sex and likeschool variables and their confidence intervals (i) on the log scale and (ii) on the mean scale. Which of these variables appear to be significant predictors?

(e) Now suppose that the number of absences for some of the students were measured over 1 month and some were measured over 3 months as indicated by the variable **hmonth**. Refit the model for the number of days absent using this variable as the offset. How does this change your answers to part (d)?

Note: For the rest of the problem revert to thinking of the number of absences as all corresponding to one-month intervals as in parts (a)-(d).

(f) In this part of the problem we evaluate the issue of over-dispersion. First suggest some plausible reasons why there might be over-dispersion in this data set and then check for it in the following ways:

(1) Using basic descriptive statistics obtain the mean and variance of the **nhookey** variable. (i) What do these statistics suggest about over-dispersion and (ii) why might this assessment be too crude to evaluate the presence of over-dispersion in the model overall?
(2) Evaluate the goodness of fit of the model using by calculating the Pearson chi-squared goodness of fit statistic and approximating the dispersion factor. Explain what each of these checks tells you.
(3) Obtain the Pearson residuals for the model and plot them as a function of the fitted values. What does this plot suggest about over-dispersion? Do there appear to be any outliers? Briefly justify your answer.

(g) Rerun the model using a negative binomial regression. Does this model confirm your conclusions about over-dispersion from part (f)? Explain briefly.

(h) Give brief interpretations of the coefficients from the negative binomial regression. Are your conclusions (as to magnitude, direction and significance) of the effects of the predictors any different from those in part (d)?

In the next part of the problem we consider the issue of zero inflation.

(i) Explain intuitively why we might expect zero-inflation in the context of this problem. Of our three predictors, which would you expect to be most associated with a "certain zero"? Explain briefly.

(j) Using the estimated rate of unexcused absences from your model in part (a) calculate the expected number of 0's we would see in a sample of 252 students if absences have a Poisson distribution. You may find it helpful to remember that for a Poisson random variable the probability of k events is $P(Y = k) = \mu^k e^{-\mu}/k!$. How many 0's did we actually observe? What does this suggest about whether we have a zero-inflation issue? (Note: This is a fairly crude approximation because we haven't accounted for the covariates but it is rather suggestive. You can of course take this approach further and break it down by sex, age bin and likeschool categories.)

(k) Fit a zero-inflated Poisson model to the data using age, sex and likeschool as the predictors for both the Poisson component and the zero-inflation component of the model. Does this model seem like an improvement over a standard Poisson regression? Explain briefly.

(l) Give brief interpretations of the coefficients for each component of the zero-inflated Poisson model. Which variables seem to be significant in each part of the model? Explain as carefully as you can what this model suggests about how these variables affect school absences.

(m) Now fit a zero-inflated negative binomial model using the same model set-up as in part (k). Is this model superior to the zero-inflated Poisson model of part (k)? Does it appear to be an improvement over the plain negative binomial model of part (g)? Explain briefly in each instance. What does that tell you about the issues of overdispersion and zero-inflation? Do any of your conceptual conclusions from part (l) change with this model?

STATA and SAS Commands

For this assignment you need to be able to fit Poisson/negative binomial models. The necessary commands are given below. To save you time I have given pretty complete details for the commands for the turn-in problems.

Commands in STATA

(1) Poisson Regression: Poisson regression works pretty much the same way as other regression models. You use the **poisson** command followed by the names of your outcome and predictor variables. For example, for warm-up Problem 2 predicting the number of imperfections in computer chips as a function of manufacturing process the command would be

poisson imperfections treatment

If you want to fit a Poisson model with an **offset** you need to be a little bit careful. STATA provides both an **offset** option and an **exposure** option. What I talked about as an offset in class (and what most people mean when they say offset) is what STATA calls **exposure**–namely the amount of time (or other units) over which events could take place. The important thing is to understand how STATA treats these two options. If you use **offset** then STATA includes the offset variable as a predictor in its raw units with a coefficient constrained to 1. This is not really what we usually want since if we are thinking of the offset as the "time" or other "per unit" component of the Poisson rate then our model is really $log(\mu/t) = X\beta$ or $log(\mu) = log(t)+X\beta$, meaning we want the offset in the model on the **log scale** with a coefficient constrained to 1. Thus to use offset we have to first take the log of our "units" variable. STATA's **exposure** option includes the variable in the model on the log scale with the coefficient constrained to 1 which is usually what we want. If we use this option we don't have to transform the offset variable first. For example, in Warm-up Problem 3 where we are looking at arrests at British soccer matches our "offset" variable is attendance–we would express number of arrests to be proportional to the size of the crowd. You could also think of this as "exposure" since each person there had a chance to get arrested. In that problem we are fitting a model with no predictors except the offset so our two versions of the statements become

gen logatt = log(attendance)
poisson arrests, offset(logatt)
poisson arrests, exposure(attendance)

Many of the follow-up commands for the Poisson model are like those for other GLMs. For example, you can use the **predict varname** to obtain the predicted counts for the values in your sample and store them in the new variable varname. This is the default for the predict command with the Poisson model. You can, of course, use other options to get other estimated quantities but the Poisson model doesn't have as many options as logit or probit. In particular, it doesn't automatically create the various residuals. If you want to get those automatically you can use the **glm** command instead of the **poisson** command (see below). Of course it is not hard to create the residuals manually. For example, suppose I want to create the Pearson residuals for a Poisson regression model. The Pearson residuals are the observed minus the predicted values divided by the standard deviation of the prediction. For a Poisson random variable the mean and the variance are equal so the estimated standard deviation for the predicted mean should just be the square root of the predicted mean. For instance, for Warm-up Problem 2, supposed we have saved the predicted number of imperfections in the variable predimp. Then the residuals would be created as

```
gen impresids = (imperfections - predimp)/sqrt(predimp)
```

Various other post-estimation commands also work with **poisson**. For instance, you can store the results using the **estimates store mymodelname** and then compare two nested models using the **lrtest** command to get a likelihood ratio chi-squared test. You can also use **estat gof** to obtain the deviance goodness of fit test and **estat gof, pearson** to get the Pearson chi-squared goodness of fit test.

(2) Negative Binomial Regression: To fit a standard negative binomial regression in STATA you use the **nbreg** command followed by your outcome and predictor variables. For instance for the fishing data of warm-up Problem 4 we would type

```
nbreg numfish camper children persons
```

The other options for nbreg are basically the same as those for Poisson regression so I won't repeat them here. The only other thing of importance to note is that STATA automatically generates the test comparing the negative binomial regression to a Poisson regression. Specifically it gives a likelihood ratio chi-squared test for the over-dispersion parameter it calls alpha at the bottom of the printout. A small p-value means there was overdispersion and the negative binomial model fits better.

(3) Zero-inflated Models: Both Poisson regression and negative binomial regression can be fit with a zero-inflation option. The commands **zip** and **zinb** replace the commands **poisson** and **nbreg** respectively. You also have to use an **inflate** option to specify which variables you want to use to predict the probability of certain zeros. These can be the same or different from the variables used for the count part of the model. If you also include the **zip** option in the negative binomial regression it will give you a comparison between the two models. The commands, applied to the fishing data of Warm-up Problem 4 are

```
zip numfish camper persons children, inflate(camper persons children)
zinb numfish camper persons children, inflate(camper persons children) zip
```

(4) The General GLM Command (Optional): In addition to having specific commands for logistic, Poisson, and other generalized linear models, STATA has a **glm** command in which you can specify the distribution and link function as part of the regression statement. The basic syntax, using Poisson regression with a log link as an example, is

```
glm Y X1 X2 X3, family(poisson) link(log)
```

Note that by default STATA uses the link that is "canonical" or natural for the specified distribution. For Poisson regression this is the log link so you don't in fact have to include the link statement. The advantage to using **glm** for Poisson regression is that it has more post-estimation and other options, including calculation of the residuals–don't ask me why they don't include these with the Poisson command! You can look at the help file on **glm** for additional details.

(5) Review of Useful Old Commands:

(a) Data Summaries: The **summarize** command: To obtain means, standard deviations and the like for a given variable you type "summarize" followed by the variable name. You can obtain this summary information for subgroups of the data set using the **bysort** option, specifying as many grouping variables as you like. For example, to get the mean and standard deviation of the number of fish caught in Warm-up Problem 4, subsetted by the number of children in the party, I would type

```
bysort children: summarize numfish
```

The same command, using **table** instead of **summarize** would allow you to see the frequency tables of numbers of fish caught by subgroup.

(b) Scatter Plots: To obtain a scatterplot of Y vs X you simply type

```
scatter Y X
```

The scatter plot can accept more than one Y variable for the same X variable and will color code the plot so you can tell which outcome is which. This would be useful, for instance, in warm-up Problem 3(b) where I ask you to compare the observed and predicted number of arrests as a function of the number of people attending the game. Assuming my predicted values have been stored in **myarrests** the command would be

```
scatter arrests myarrests attendance
```

(c) Scatterplots of Bin Means: In Problem 5(b) I ask you to plot mean number of absences from school as a function of whether the child likes school and age. You can of course obtain the bin means as described above by typing **bysort likeschool: summarize nhookey** but then you have to enter the 5 means into a new column. An alternative is to use the **egen** command as follows:

```
bysort likeschool: egen likemeans = mean(nhookey)
```

This creates a new variable, likemeans, whose values are the average number of absences for people with the specified rating of liking school. If you then type **scatter likemeans likeschool** you will get the desired plot. Age, of course, works the same way.

(d) Including Categorical Variables in a Model: To do this you can either create your own dummy variables or as part of the model statement in later versions of STATA you can attach the prefix "i." to the variable name. For instance, for Problem 5(c) where you are asked to try treating **likeschool** as categorical you would either need to create 4 dummy variables (since likeschool has 5 categories) or else tell STATA that likeschool was categorical. The resulting two versions of the commands would be

```
poisson nhookey age gender like2 like3 like4 like5
poisson nhookey age gender i.likeschool
```

I already created the dummy variables for you but if you wanted to do it yourself one approach would be the following:

```
gen like2 = 0
replace like2 = 1 if likeschool==2
```

And simmilarly for the other categories.

Commands in SAS

(1) Summaries By Group: Descriptive statistics in SAS are obtained using **proc univariate, proc freq** and **proc means**. You specify the procedure, the data set and the variables you want to summarize using the **var** command. If you want to split the summaries by group you use a **class** statement to specify the grouping variable. As an example I use the fishing data from Warm-up Problem 4. The first set of commands below gives detailed summary statistics for the number of fish caught, the number of people and the number of children. If I wanted to subgroup the number of fish caught by whether or not the people were camping I would add the class statement as shown in the second set of commands. I could also use proc means, specifying which summary statistics I want. The third command below gives a table with the mean, standard deviation minimum and maximum for the number of fish split by the levels of the campter variable. You can include as many variables as you like in the var statement. The final set of commands gives the frequencies for each of the observed values for the camper, persons and children variables using **proc freq**. This procedure and be expanded to give contingency tables by typing var1*var2 in the var statement and you can add the option **/chisq** to get the various contingency table tests as well.

```
proc univariate data = tmp1.hw4 ;
var numfish persons children ;
run;

proc univariate data = tmp1.hw4;
class camper;
var numfish;
run;

proc means data = tmp1.hw4 mean std min max var;
histogram numfish/midpoints 0 to 50 by 1 vscale = numfish;
run;

proc freq data = tmp1.hw4;
tables camper children persons camper*children/chisq;
\run;
```

(2) Including Categorical Variables In A Model: Categorical variables are something SAS has historically handled more easily than STATA. In basically any regression model from OLS to a zero-inflated negative binomial model, all you have to do is add a line with a **class** command and list the variables you want treated as categorical. If the variables are multicategory, SAS will give tests corresponding to whether the group of dummy variables representing the categories significantly improves the model (i.e. a partial F test in OLS and a likelihood ratio chi-squared test in a GLM.) The only tricky issue is what SAS specifies as the reference category. In the sections below I describe some of the referencing options.

(3) Poisson Regression: Poisson regression in SAS is generally fit using **proc genmod** which is shorthand for generalized linear model. In general for proc genmod you have to specify the distribution and link function. If you do not specify the distribution, normality is assumed. If you do not specify the link SAS defaults to the canonical link. For the Poisson distribution this is the log link so you can skip that part. Like other regression models in SAS, proc genmod allows a **class** statement which you can use to tell it which variables are categorical. You can add an option **descending** which tells it that you want to use the lowest value (i.e. 0) as the reference value. Otherwise it defaults to using the highest value as the reference. The model statement has the same form as in standard regression, namely $Y = X_1X_2.....$ After it you can add the option **type3** which will make SAS automatically present global tests for each categorical variable. The model statement is also where you specify the form of the glm using **dist** and **link** options. Below is the syntax we would use to run a Poisson model for the fishing data of Warm-Up Problem 4. The first basic version treats all the variables as continuous (for an indicator variable taking on only the values 0 and 1 it doesn't matter whether you indicate it is categorical or not.) Note that I didn't really need the link option. Also note that SAS automatically gives you the deviance and Pearson goodness of fit statistics along with the log likelihood.

```
proc genmod data = tmp1.hw4;
model numfish = camping persons children/dist = poisson link = log;
run;
```

The second version, below, treats the camping and children variables as categorical for illustrative purposes. Note the descending option. The default test for the option **type3** is likelihood ratio chi-squared tests. If you ad the option **wald** after **type3** you get Wald tests instead.

```
proc genmod data = tmp1.hw4;
class campter children/ descending;
model numfish = camping persons children/type3 dist = poisson link = log;
run;
```

There are many other options you can use. For example, the option **obstat** used after the model statement prints out a table of things like fitted values and residuals for the points in the data set. Beware of doing this with a large data set as you will get MANY pages of printout. The option **offset = variable** is used to include an offset term in the model. Note that like STATA, SAS uses the variable on the raw scale in the model so to get what we were calling offset or "exposure" variable in class you need to take the natural log first. I illustrate these commands below for the soccer arrest data from Warm-Up Problem 3 using the new logattendance variable I created:

```
proc genmod data = tmp1.hw4;
model arrests = / dist = poisson link = log offset = logattendance obstat;
run;
```

(5) Negative Binomial Regression: This works basically exactly the same way as Poisson regression except that for the distribution we use **negbin**. The various other options work the same way. The basic commands for the fishing data would look like

```
proc genmod data = tmp1.hw4;
model numfish = camping persons children/dist = negbin;
run;
```

(5) Zero Inflated Models: To get zero-inflated versions of the Poisson or negative binomial models you can still use **proc genmod** but you have to add a few extra options. For example, for the zero-inflated Poisson you specify the distribution as **zip** instead of **poisson** and you have to tell it what model to fit for the zeros. The syntax for the fishing example of Warm-Up Problem 4 is shown below using all three variables in both parts of the model, treated as continuous. Since the log link is the default I have left that part of the statement off but you are free to specify it. If you wanted to you could add a class statement with the usual options to treat some of the variables as categorical and you could change which variables were involved in which part of the model.

```
proc genmod data = tmp1.hw4;
model numfish = camper persons children/dist = zip;
zeromodel camper persons children/link = logit;
run;
```

The set-up for negative binomial models is similar except our distribution is the zero-inflated negative binomial, **zinb**. The basic code for the fishing data is:

```
proc genmod data = tmp1.hw4;
model numfish = camper persons children/dist = zinb;
zeromodel camper persons children/link = logit;
run;
```

The SAS zero-inflated negative binomial procedure also gives a Wald confidence interval for the dispersion,

There is also a special procedure, **proc countreg** which is more specific to count data and has some extra options. There are several fitting methods. The one that matches the **proc genmod** procedure is call quasi-newton. You can get this with the **method = qn** option in the proc statement. The basic code for the Poisson model is as follows:

```
proc countreg data = tmp1.hw4 method = qn;
model numfish = camper persons children/dist = zip;
zeromodel numfish˜camper persons children;
run;
```

You can even fit these sorts of models using another procedure, **proc nlmixed** (which stands for non-linear mixed models)–not necessary here but useful in some circumstances.