Soc 73994, Homework #10
Count Models

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised November 18, 2024

All answers should be submitted via Canvas. Be sure your response includes your name, the date, and a clear title, e.g. Homework # 10. If there is a huge amount of output for any analyses you run yourself, you may want to be selective in what you copy and paste into your assignment (but make sure you include enough so it is clear what commands you executed, e.g. you might show all the commands but only parts of the output).

You will answer each of the following questions *twice*. The first time, you will use the American National Election Survey (ANES) 2008 data. Former TA Chris Quiroz has already developed code for this, so you can just run it and interpret the results (although you are welcome to tweak the code if you wish). The second time, you will analyze a data set of your choice. You'll have to develop your own code for this, although Chris's code may serve as a template.

This assignment has multiple parts. First, you will run some descriptive statistics just to get a feel for what your dependent variable is like. Then, you will estimate different count models and indicate which you think is most appropriate given your data, and why. Then, you will interpret your preferred model's parameters, using only the information reported by the estimation command, e.g. you can interpret the results by noting the sign and significance of coefficients and also by looking at the IRRs (Incident Rate Ratios). Finally, you will use post-estimation commands like `margins` and `mcp` and/or others to further interpret your results. Use the handout on Count Models and/or the ANES example to provide you with a template if you need one.

Ideally you will use your own data for this. Even if you haven't been planning to estimate a count model, see if your data set includes something that could legitimately be used as a dependent variable in a count model. However, if you don't have anything, you can use one of the count data sets that are included with Stata. These include

```
webuse airline
webuse dollhill3
webuse rod93
webuse runshoes
webuse medpar
```

You will probably want to look these up in the Stata manuals to find out more about them. But, you will want to go beyond whatever analysis that is already presented. If you can come up with one, you could also use a different dependent variable in the ANES data.

1. Run some descriptive statistics on the dependent variable (as well as the others that may get included in your model if you have not worked with these data before). Be sure to note whether the mean and variance of the dependent variable appear to differ from each other. Also, if there are no zeros in the data set, you should note that (and do your best to explain why the

zeros are missing – they may be missing because of the way the data were collected, or they may be missing simply because 0s are very unlikely to occur).

2. Run a series of count models on your data. In most cases, you will want to run a Poisson Model and an NBREG model. You should also try a hurdle model and/or a zero inflated model. Long and Freese's `countfit` command may be able to make things a little easier for you. If, however, your dependent variable does not have any zeros because of the way the data were collected, you will want to try out a couple of zero-truncated models.

After estimating different models, indicate which model you think is best, and why. The decision may just be based on which model fits the best. But, theory may also lead you to, say, argue in favor of a hurdle model or else a simpler model that fits almost as well as the seemingly best model.

3. Using only the output from the estimation command for your preferred model, interpret the results. Indicate the sign and significance of coefficients. Also discuss what the IRRs tell you. If you estimate something like a hurdle model or a zip model, explain what the coefficients from both parts of the analysis tell you.

4. Now do additional analyses to make your results easier to understand. You can use commands like `margins`; `listcoef`; `mtable`; `mcp`; and others if you feel they are helpful. Be sure to make clear what these commands tell you, e.g. maybe they show you how expected counts differ by gender or across age.

Here is code you can use for the first half of the assignment using the ANES data. Again, you can tweak it a bit if you wish, and/or use it to give you ideas when analyzing the data set of your choice. The dependent variable is numorgs, which is a version of V085127 where numeric missing data codes have been recoded to missing. The codebook says

> Here is a list of some organizations people can belong to. There are:
> labor unions, associations of people who do the same kinds of work, fraternal groups such as Lions or Kiwanis, hobby clubs or sports teams, groups working on political issues, community groups, and school groups. Of course, there are lots of other types of organizations, too.

Incidentally not all variables have statistically significant effects. You may still want to comment on one or more of them if you think the lack of significance is itself an interesting finding.

```
clear all
use https://www3.nd.edu/~rwilliam/statafiles/anes_codeddata, clear

*************************************************************************;
*Descriptives
sum numorgs female politics yearseduc age
fre politics female
hist numorgs, discrete freq scheme(sj) name(histogram)

*************************************************************************;
*Poison Regression Model
```

```
poisson numorgs i.female politics yearseduc age

*Post Estimation Goodness of Fit
est store poisson
mgen, pr(0/13) meanpred stub(psn)
rename psnobeq ObsProp
rename psnpreq PsnPrdct
rename psnval NumbOrgs
list NumbOrgs ObsProp PsnPrdct in 1/13
graph twoway connected ObsProp PsnPrdct NumbOrgs, ///
ytitle("Probability") ylabel(0(.1).6) xlabel(0/13) msym(O Th) name(poisson)

*******************************************************************************;
*Negative Binomial Model

nbreg numorgs i.female politics yearseduc age

*Post Estimation Goodness of Fit
est store nbr
drop NumbOrgs ObsProp PsnPrdct psn*
mgen, pr(0/13) meanpred stub(psn)
rename psnobeq ObsProp
rename psnpreq PsnPrdct
rename psnval NumbOrgs
list NumbOrgs ObsProp PsnPrdct in 1/13
graph twoway connected ObsProp PsnPrdct NumbOrgs, ///
ytitle("Probability") ylabel(0(.1).6) xlabel(0/13) msym(O Th) name(nbreg)

*******************************************************************************;
*Hurdle Model

logit numorgs female politics yearseduc age, or nolog
est store Hlogit
ztnb numorgs female politics yearseduc age if numorgs>0, nolog irr
est store Hztnb
suest Hlogit Hztnb, vce(robust) eform(expB)
est store hurdle

*******************************************************************************;
*zero-inflated Poisson Model

zip numorgs i.female politics yearseduc age, inflate(i.female politics yearseduc age)
est store zip

*zero-inflated Model With Negative Binomial Regression

zinb numorgs i.female politics yearseduc age, inflate(i.female politics yearseduc age)
est store zinb

*******************************************************************************;
*Compare Goodness of Fit Across Models

est stats poisson nbr hurdle nbhurdle zip zinb
```

```
*************************************************************************;
*Best Model - Zinb may have a slight edge empirically but we will follow Allison's
*advice and go with the more parsimonious nbreg. But feel free to change this
*if you disagree with this choice.

nbreg numorgs i.female politics yearseduc age, irr
mcp age yearseduc, at2(8 12 16) plotopts(name(nbregmcp))
```