Solutions To Homework Assignment 4 Warm-Up Problems

General Comments:

- The solutions given below are (quite a bit) more extensive than would have been necessary to get full credit. I use the answer key as an opportunity to make important points, or mention commonly made mistakes. Nonetheless, the answer key should give you an idea of the type of solutions I would like to receive.
- I have included the graphics in a separate file since they don't import super easily into my mathematical word processing program.

Warmup Problems

(1) Poisson and Negative Binomial Regression Basics:

(a) In Poisson regression the outcome, Y is assumed to be a count variable with a Poisson distribution with a mean number of events of μ (or more generally this can be stated as a mean rate of events per unit of measurement in which case it may be denoted λ rather than μ.) The link function is the natural logarithm and the systematic component is our usual linear combination of the X variables, $\beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$. For negative binomial regression everything is the same except that we assume that the count variable, Y has a negative binomial distribution. This distribution can be thought of as the number of independent trials, each with success probability p, which occur before the first failure (e.g. number of times you flip a coin and get heads (success) before the first tails (failure). The mean of this distribution is $p/(1-p)$. We specifically model the mean as the log of a linear combination of the X variables but of course you can also transform this to get estimates of the success probability for the individual trials if conceptually your Y variable fits this framework. The negative binomial is often used as an alternative to the Poisson distribution if the variance in the observed data is larger than what would be true for a Poisson (which has variance = mean). This can be done without really believing that the values arose from a sequence of success/failure trials.

(b) Extending some of our earlier homework problems, a Poisson model could be used for the number of ear infections an infant has per year, with the rate of infections depending on factors like gender, age, whether or not the child is breast-fed, general health and so forth. The actual number of infections will be relatively small, implying a skewed distribution, so a normal approximation model or a square-root transformation would not be appropriate. (Note that if the mean is large a Poisson distribution is approximately normal while if the mean is moderate a square-root transformation will often normalize it. In both cases this allows us to use ordinary least squares regression. However, if the mean is small, OLS will work rather badly and in particular may produce negative predictions for some X values.) A negative binomial regression model can be used in most situations where one would apply a Poisson model if there is concern about over-dispersion (see part (c)). Conceptually, an example where the negative binomial framework would make sense could be modeling the number of treatment sessions or doses of a medication that are needed before a patient responds to the intervention. The mean number of sessions/doses could depend on factors like age, gender, illness severity, type of treatment, dose, etc.

(c) The term **overdispersion** refers to the situation when the observed variance of a variable is greater than what would be expected for the assumed distribution. You can also have under-dispersion if there is

less than the expected variability but this is a lot rarer. For example, for a Poisson distributed random variable the mean and the variance are assumed equal. However it is often the case in Poisson regression settings that the observed variance is substantially larger than the mean. The phenomenon can occur with other distributions as well (e.g. binomial, negative binomial, etc.) but is not an issue in OLS regression because for the normal distribution the variance does not depend on the mean and we estimate it separately (the standard deviation, σ, is estimated by RMSE.) Over-dispersion can occur for a variety of reasons in Poisson and negative binomiral regression including heterogeneity of the observed subjects/failure to include important predictors in the model, cluster sampling, because the events being measured tend to occur in bunches or because the distribution of events is really a mixture of people who will never have the event (and hence always will have Y=0) and people whose number of events do follow the specified distribution. The term **zero-inflation** refers to this last situation. We will see examples below where it would not be surprising to have over-dispersion, zero-inflation, or both.

Over-dispersion can be tested for and if it is present one can either use a distribution for Y with a larger variance (e.g. using negative binomial instead of Poisson regression) or if one doesn't have a good idea of the ideal distribution one can estimate the degree of over-dispersion and try to adjust the standard error estimates of the parameters directly. In a Poisson regression, if there is no over-dispersion the Pearson chi-squared goodness of fit statistic should have an expected value of roughly n-m where n is the number of data points and m is the number of parameters in the regression model. However if there is over-dispersion then the goodness of fit statistic (which essentially looks at the squared difference between the data points and their predicted values–i.e. it's a variance estimate!) will be much bigger. Dividing the Pearson statistic by n-m thus provides an estimate of the variance inflation. One of the biggest problems with over-dispersion is that the estimated standard errors of the regression coefficients are too small (they are based on the fact that the Poisson variance is the same as the mean) and thus the significance of effects is over-stated. Multiplying the standard error of the parameter estimates by the square-root of the variance inflation factor provides a rough adjustment to correct for this problem with over-dispersion.

To adjust for zero-inflation one fits a two part model in which one first assesses whether the subject would ever have an event (specifically it models the probability of being a "certain zero") using a logistic model and then one models the **number** events for the people who could have them using Poisson, negative binomial or some other regression count model. One can interpret the two pieces of the model separately in the standard way. Most computer packages have implementations of both **zip** (zero inflated Poisson) and **zinb** (zero-inflated negative binomial) models.

(2) Munching (Computer) Chips:

(a) The printouts for the simple Poisson regression with treatment process as the predictor are shown below. The p-value for the Wald test of the treatment effect is .001 and the p-value for the likelihood ratio chi-squared test for the significance of the model is .0007. (Since this is a simple model with one predictor we get the likleihood ratio chi-squared statistic and p-value directly. We could also have obtained it by fitting the model with no predictors and taking -2 times the difference in log likelihoods.) From either of these we conclude that treatment is a significant predictor of the number of imperfections in the computer chips. Since treatment is an indicator variable we interpret this as meaning there is a significant difference in the mean number of imperfections between the groups. Since the coefficient of the treatment variable is positive we know that the number of imperfections is higher for the treatment = 1 process than the treatment = 0 process.

```
IN STATA:
. poisson imperfections treatment

Poisson regression                              Number of obs   =         20
                                                LR chi2(1)      =      11.59
```

```
                                                Prob > chi2     =     0.0007
Log likelihood = -45.174552                     Pseudo R2       =     0.1137

------------------------------------------------------------------------------
imperfecti~s |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |   .5877867   .1763834     3.33   0.001     .2420815    .9334918
       _cons |   1.609438   .1414214    11.38   0.000     1.332257    1.886619
------------------------------------------------------------------------------
```

****

IN SAS:

```
proc genmod data = tmp1.hw5;
model imperfections = treatment/dist = poisson link = log type3;
run;
```

The GENMOD Procedure

Model Information

Data Set	TMP1.HW5	
Distribution	Poisson	
Link Function	Log	
Dependent Variable	imperfections	imperfections

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	18	16.2676	0.9038
Scaled Deviance	18	16.2676	0.9038
Pearson Chi-Square	18	16.0444	0.8914
Scaled Pearson X2	18	16.0444	0.8914
Log Likelihood		138.2221	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	1.6094	0.1414	1.3323	1.8866	129.51	<.0001
treatment	1	0.5878	0.1764	0.2421	0.9335	11.11	0.0009
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
treatment	1	11.59	0.0007

(b) Here our model is $ln(\mu) = \beta_0 + \beta_1 X$ where X is the treatment indicator. When $X = 0$ so we are using

the reference treatment process we get $ln(\mu) = \beta_0$. Thus β_0 represents the log mean number of imperfections for the reference treatment process which here is 1.61. The corresponding confidence interval for the log mean number of imperfections is 1.33 to 1.89. These numbers are hard to interpret numerically although they are all positive which means that the average number of imperfections is above 1. (The log of 1 is 0; log values above 0 correspond to raw values above 1.) If we exponentiate we get that the mean number of imperfectins for the treatment 0 process is $e^{1.61} = 5.00$ or 5 imperfections per chip. The corresponding confidence interval is $[e^{1.33}, e^{1.89}] = [3.78, 6.62]$. We are 95% sure that the average number of imperfections per chip for treatment process 0 is between 3.78 and 6.62. The coefficient for the treatment variable gives the difference in log mean number of imperfections between treatment process 1 and treatment process 0:

$$ln(\mu_1) - \ln(\mu_0) = \beta_0 + \beta_1(1) - (\beta_0 + \beta_1(0)) = \beta_1$$

Here we see that the the log mean for the treatment 1 process is .588 units higher than the log mean for the treatment 0 process. The corresponding confidence interval says that the difference in log means could be anywhere from .242 to .933 units. When we exponentiate the difference in the log mean scale turns into a multiplicative factor on the mean scale:

$$\beta_1 = ln(\mu_1) - \ln(\mu_0) = ln(\frac{\mu_1}{\mu_0}$$

so we have

$$e^{\beta_1} = \frac{\mu_1}{\mu_0}$$

Here the estimated ratio of the means is $e^{.588} = 1.80$, which suggests that the mean number of imperfections using treatment process 1 is 1.8 times as high as for treatment process 0. Since we know the estimated mean for treatment process 0 was 5 imperfections, the implication is that the mean number of imperfections using treatment process 1 is $5 * 1.8 = 9$ imperfections per chip. Of course we could have gotten this directly as $e^{\beta_0+\beta_1(1)} \approx e^{1.61+.588} = 9$. Exponentiating the confidence interval for β_1 yields $[e^{.242}, e^{.933}] = [1.27, 2.54]$. We are 95% sure that the rate of imperfections using process 1 is between 27% and 154% higher or between 1.27 to 2.54 times higher than with process 0. Note that in this particular case since we only have the single indicator variable it would have been just as easy to calculate the sample group means directly! The printout is

```
. bysort treatment: summarize imperfections

-------------------------------------------------------------------------------
-> treatment = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
imperfecti~s |        10           5    2.054805          2          8

-------------------------------------------------------------------------------
-> treatment = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
imperfecti~s |        10           9    2.905933          5         14

-------------------------------------------------------------------------------
```

One interesting thing to note here: If we square the standard deviations within the two strata to obtain the sample variances we get 4.22 and 8.47 which are very close to (and in fact slightly below) the sample means.

We do not have any indication of overdispersion in these data.

(c) The printouts with thickness added are shown below. We can check whether the model is significantly improved by either performing a Wald test for the thickness variable (which is insignificant with p-value .177) or by performing a likelihood ratio chi-squared test. We could do this manually by looking at -2 times the difference in log likelihoods but I got STATA to save the results of the simple Poisson model and the expanded model and used the lrtest command. The resulting p-value for the likelihood ratio chi-squared test is .176, also not close to significance. It doesn't appear that the thickness of the chips adds any information about the mean number of imperfections beyond what was explainned by the treatment process. It doesn't really make sense to interpret the coefficient when it isn't significant, but for the sake of illustration I'll say what the interpretation would have been. Basically we coded thickness=1 for the thicker chips and 0 for the thinner chips. The negative coefficient on the thickness variable means that all else equal (i.e. if the treatment process is the same) the log mean number of imperfections is lower by .23 units for thicker chips than for thinner chips which seems reasonable–the thicker chips might be more durable. Exponentiating gives $e^{-.23} = .795$ meaning that all else equal we would expect only 80% as many imperfections per chip on thicker chips as on thinner chips. Note that because we do not have an interaction we are assuming that thickening the chips leads to the same reduction in the rate of imperfections regardless of which treatment process we are using.

```
IN STATA:
. poisson imperfections treatment thickness

Poisson regression                                Number of obs   =         20
                                                  LR chi2(2)      =      13.42
                                                  Prob > chi2     =     0.0012
Log likelihood = -44.258266                       Pseudo R2       =     0.1317

------------------------------------------------------------------------------
imperfecti~s |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |   .5877867   .1763834     3.33   0.001     .2420815    .9334918
   thickness |  -.2295744   .1701457    -1.35   0.177    -.5630538    .1039049
       _cons |   1.717651   .1602425    10.72   0.000     1.403582    2.031721
------------------------------------------------------------------------------

. estimates store txthick

. lrtest tx txthick

Likelihood-ratio test                                 LR chi2(1)  =      1.83
(Assumption: tx nested in txthick)                    Prob > chi2 =    0.1758
*******************************************************************************
IN SAS:

proc genmod data = tmp1.hw5;
model imperfections = treatment thickness/dist = poisson link = log type3;
run;

                                The GENMOD Procedure

                                Model Information
```

Data Set	TMP1.HW5	
Distribution	Poisson	
Link Function	Log	
Dependent Variable	imperfections	imperfections

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	17	14.4351	0.8491
Scaled Deviance	17	14.4351	0.8491
Pearson Chi-Square	17	14.6871	0.8639
Scaled Pearson X2	17	14.6871	0.8639
Log Likelihood		139.1384	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	1.7177	0.1602	1.4036	2.0317	114.90	<.0001
treatment	1	0.5878	0.1764	0.2421	0.9335	11.11	0.0009
thickness	1	-0.2296	0.1701	-0.5631	0.1039	1.82	0.1772
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
treatment	1	11.59	0.0007
thickness	1	1.83	0.1758

(3) More Sports Fanatics:

(a) The Poisson model fitting the number of arrests with no predictors except attendance as an offset variable is shown below. In STATA there are two choices of how to deal with an offset variable–using the **offset** option or the **exposure** option. 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 since 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. I fit the model both ways below, having first used the **generate** command to create a log attendance variable. Reassuringly I get the same answer either way! Similar cautions apply in SAS where the offset variable needs to be on the log scale.

It was important to use an offset variable here because the numbers of fans at the different games were very

different. Having 110 people get arrested when 321,000 fans were present (Middlesbro) is far less impressive (or unimpressive depending on your point of view!) than having 101 fans get arrested when 189,000 were present (Birmingham). Using the offset gives us a rate of arrests per 1000 fans which makes the different games comparable.

Since there are no predictors other than the adjustment for attendance, our only parameter is the intercept which just gives us the log of the overall rate of arrests. In other words, we are assuming that the **rate** of arrests is constant across all the games. If we exponentiate the intercept we get the rate per unit attendance. here $e^{-.91} = .40$ meaning approximately .4 people get arrested per 1000 fans in attendance or, more meaningfully, that we expect 4 people to get arrested per 10,000 fans attending the match.

```
IN STATA:
. poisson arrests, offset(logatt)

Poisson regression                                Number of obs   =         23
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -405.30989                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
     arrests |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -.9102802   .0216371   -42.07   0.000    -.9526882   -.8678722
      logatt |   (offset)
------------------------------------------------------------------------------

*******************************************************************************
. poisson arrests, exposure(attendance)

Poisson regression                                Number of obs   =         23
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -405.30985                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
     arrests |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -.9102802   .0216371   -42.07   0.000    -.9526882   -.8678722
  attendance | (exposure)
------------------------------------------------------------------------------
```

(b) In the accompanying graphics file I show a plot of arrests–both actual and predicted by the model–as a function of attendance. Natrually the predicted numbers of arrests follow a straight line since we have a constant rate of arrests per thousand people. There is one match where the number of arrests seems rather high for the number attending, namely the Aston Villa match where there were 308 arrests with an attendance of 404,000. (The prediction was for only 163 arrests.) There is also one match with a surprisingly small number of arrests, namely Manchester City with only 35 arrests and 429,000 people attending (predicted number of arrests 173). Of course since the variance gets larger as the mean gets larger and both of these observations occurred when the attendance was very high) these results may not be as surprising as they first look. (Here

even though the *rate* is fixed the mean number of events, and therefore the variance in number of events, grows as the attendance grows.) I also created a plot of Pearson residuals (actual minus predicted number of arrests divided by the square root of the predicted number of arrests) vs attendance. On this plot the two games cited previously do have the largest residuals but they no longer stand out nearly as much from the other points, indicating that they may not really be outliers. The commands I used to create the plots and modified variables are shown below.

```
. predict myarrests
(option n assumed; predicted number of events)

. scatter arrests myarrests attendance

. gen resids = (arrests-myarrests)/sqrt(myarrests)

. scatter resids attendance
```

(4) Camping Data:

(a) The printout for the basic Poisson regression is shown below.

```
. poisson numfish camper persons children

Poisson regression                                Number of obs   =        250
                                                  LR chi2(3)      =    1621.29
                                                  Prob > chi2     =     0.0000
Log likelihood = -837.07248                       Pseudo R2       =     0.4920

------------------------------------------------------------------------------
     numfish |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      camper |   .9309359   .0890869    10.45   0.000     .7563289    1.105543
     persons |   1.091262   .0392553    27.80   0.000     1.014323    1.168201
    children |  -1.689957   .0809922   -20.87   0.000    -1.848699   -1.531215
       _cons |  -1.981827    .152263   -13.02   0.000    -2.280257   -1.683397
------------------------------------------------------------------------------
```

(b) Based on their Wald test p-values all three of the predictors, number of people in the group, number of children in the group and whether or not the group was camping, appear to be significant predictors of the number of fish caught. On the log scale we see that all else equal, groups that camped had an estimated .93 higher log mean number of fish caught than groups that didn't camp, with the confidence interval showing the increase could have been anywhere from .76 to 1.11. Exponentiating we get that the groups who camped caught an estimated $e^{.93} = 2.52$ times as many fish on average as the groups who didn't camp (possible range 2.13 to 3.03 times as many fish caught). For the other two variables we are looking at the effect on number of fish caught per additional person or child present. The log mean number of fish caught goes up 1.09 per additional person in the group (confidence interval of 1.01 to 1.17 increase in log mean per person). It makes sense that the more people there are in the group, the more fish they might catch. Exponentiating ($e^{1.09} = 2.97$) we get that the mean number of fish caught goes up by a factor of almost three for each additional person in the party. This seems a little surprising if we imagine that each person catches the

same number of fish...apparently there is some sort of synergistic effect or else groups that are bigger are likely to contain a larger *proportion* rather than just a larger *number* of fishers. The confidence interval tells us that the number of fish caught could be anywhere from 2.75 to 3.25 times as high per additional person in the party. For children the effect goes in the other direction–hardly a surprise since (a) the children are less likely to fish and (b) may make a lot of noise and scare the fish whether they are fishing themselves or not. We see that our best estimate is that the log mean number of fish caught goes down 1.69 per child (range anywhere from 1.53 to 1.85 decrease.). Exponentiating, $e^{-1.69} = .18$ tells us that our best estimate is that there is an 80% reduction in the number of fish caught for each additional child in the group. There is something important to remember here which is that the children are also people in the group so these two variables are correlated and their effects may be cancelling each other out. This is perhaps part of the reason that the gain in catch per person appears so high and the reduction per child appears so big.

(c) The plot of residuals versus fitted values is shown in the accompanying graphics file. There are two points with very big Pearson residuals. The largest, which has a Pearson residual of 37, corresponds to group that was predicted to catch .67 of a fish. They weren't camping, had a child, and had 3 people total so wouldn't be expected to catch much but actually caught 31 fish! Note that the small predicted value in the denominator helps to inflate the residual. The second point, with a residual of 23, corresponds to a group that was predicted to catch 27 fish. They were camping, had 4 people and no kids so it seems reasonable they would catch a lot–but they caught more than a lot–they caught 149 fish, by far the most in the sample! Either they were really good fishers, they stayed for a really long time, or they got really lucky! These two points shrink the scale of the rest of the plot so I regraphed the data without them. Even on the Pearson scale there seems to be a bit of fanning, suggesting that we may have some over-dispersion. The deviance and Pearson goodness of fit tests for this model are shown below. They are probably fairly appropriate here since there are a limited number of possible combinations of the predictor variables. The tests are hugely significant, meaning that a standard Poisson model is probably not appropriate for these data.

```
. predict fishpred
(option n assumed; predicted number of events)

gen fishresids = (numfish - fishpred)/sqrt(fishpred)

. scatter fishresids fishpred

scatter fishresids fishpred if fishresids < 20
**************************************************
. estat gof

         Goodness-of-fit chi2  =   1337.08
         Prob > chi2(246)      =    0.0000
**************************************************
. estat gof, pearson

         Goodness-of-fit chi2  =  2910.627
         Prob > chi2(246)      =    0.0000
```

(d) There are several possible reasons for over-dispersion in this data set. One is that there are probably some people visiting the state park who don't try to fish at all and so are certain zeros–in other words we have zero-inflation because we will have a lot of zeros–and then a lot of high values for people who actually fish and these will be mixed together in calculating one overall rate. We may also have real heterogeneity in our groups in other ways–how long they stayed, whether they were good fishers, what the weather was

like, how much time they spent actually fishing, and so on. If we don't include these variables in the model we may again be trying to calculate our means by merging together groups of subjects who are really very different. Finally, it may well be that fish are caught in batches–they swim in schools, get hungry at the same time, are attracted by a particular kind of lure, and so on–which may mean that a Poisson distribution which assumes events happen evenly at a fixed rate–may not be correct. Below I illustrate a number of methods for checking for over-dispersion.

(1) One method is to look at the mean and variance for different subgroups of the data. If the variance is much bigger than the mean then we may have over-dispersion. Here natural ways to group the data are by camping status, number of people and number of children. I give the printouts for those below. To be really thorough and match what our model is doing we'd need to look at all camping/people/children combinations. There are a total of 32 combinations–2 (camping yes/no) x 4 (persons 1,2,3,4) x 4 (children 0,1,2,3). I give the complete printout below for illustrative purposes. Note that although this matches our model it is probably overkill since some of these subgroups have very few observations. It is probably enough just to subdivide by camping status and children for instance. In basically all of these (except the subsets where no one caught any fish so the mean and sd are both 0) we see that the mean is much smaller than the variance (what is shown is the standard deviation–when the sd is over 1 then this is even less than the variance; for the ones where the sd is less than 1 you can tell by squaring.) Thus we have a strong suggestion of overdispersion.

```
. bysort camper person children: summarize numfish
-------------------------------------------------------------------------------> camper = 0, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |        22    .4545455    .9116846          0          3
-------------------------------------------------------------------------------> camper = 0, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |        17    1.294118    2.229482          0          9
-------------------------------------------------------------------------------> camper = 0, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |        12    .3333333    .6513389          0          2
-------------------------------------------------------------------------------> camper = 0, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |         8        4.25    5.284749          0         15
-------------------------------------------------------------------------------> camper = 0, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |         8           4    10.91526          0         31
-------------------------------------------------------------------------------> camper = 0, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |         5           0           0          0          0
-------------------------------------------------------------------------------> camper = 0, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |         6    8.166667    10.34247          2         29
-------------------------------------------------------------------------------> camper = 0, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
```

```
     numfish |        10          .6    1.074968          0          3
-------------------------------------------------------------------------------> camper = 0, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |        11           0           0          0          0
-------------------------------------------------------------------------------> camper = 0, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |         4           0           0          0          0
-------------------------------------------------------------------------------> camper = 1, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |        35    .9142857    1.578745          0          7
-------------------------------------------------------------------------------> camper = 1, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |        18    3.388889    6.509169          0         21
-------------------------------------------------------------------------------> camper = 1, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |        23    .6956522    1.362977          0          5
-------------------------------------------------------------------------------> camper = 1, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |        13           7    8.041559          0         30
-------------------------------------------------------------------------------> camper = 1, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |        11    .8181818    1.401298          0          4
-------------------------------------------------------------------------------> camper = 1, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |        12    .0833333    .2886751          0          1
-------------------------------------------------------------------------------> camper = 1, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |        13    29.69231    40.17956          1        149
-------------------------------------------------------------------------------> camper = 1, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |        11    5.909091    5.088311          0         16
-------------------------------------------------------------------------------> camper = 1, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |         5         1.2    2.167948          0          5
-------------------------------------------------------------------------------> camper = 1, persons =
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |         6           0           0          0          0
-------------------------------------------------------------------------------
```

(2) From part (c) we obtained the Pearson goodness of fit statistic which was $\chi^2(246) = 2910.627$. When

we divide the value by the degrees of freedom we get $2910.627/246 = 11.83$. This value is supposed to be 1 if there is no over-dispersion! It seems clear we've got a big problem....In fact, if we wanted to use this to adjust our standard errors we see that we would need to multiply them by $\sqrt{11.83} = 3.44$ or more than triple them! Incidentally, the Pearson and deviance goodness of fit tests were in and of themselves an indication of possible over-dispersion since they suggested that the model wasn't well calibrated. However there are many ways to be badly calibrated, not just over-dispersion. The Pearson goodness of fit test because of the way it standardizes does correspond to a test of whether there is more variance than there should be but we have to look at the individual points to tell if there are just a few combinations of X values that are over-dispersed (e.g. some outliers or a particular area of bad fit) or whether it occurs across the board. Here by calculating the means and SDs for all the possible X values we can see that the problem occurs across the board.

(3) Our residual plot from (c) should look like it has equal spread if the model is well calibrated and we would also hope that most of the residuals would be small and even about 0. From the graphics in part (c) there is still a suggesting of fanning out (residuals getting bigger with fitted values) even after removing the most extreme outliers and lots of the residuals are rather large (we're used to 2 or 3 being a big value; here we have many over 5 though we can't exactly use a normal distribution to guide us unless we further adjust the residuals by their leverage values) and there are more big positive ones than negative ones. All of this also suggests an overdispersion problem.

(e) The printout for the negative binomial model is shown below. The likelihood ratio test for the over-dispersion parameter, alpha, is highly significant (p-value 0) meaning that the negative binomial model is a significant improvement over the standard Poisson model. This suggests that over-dispersion was a significant problem in the original model as we had already seen many other ways.

```
. nbreg numfish camper persons children
Fitting Poisson model:
Fitting constant-only model:
Fitting full model:

Negative binomial regression                    Number of obs   =        250
                                                LR chi2(3)      =     118.43
Dispersion     = mean                           Prob > chi2     =     0.0000
Log likelihood = -405.222                       Pseudo R2       =     0.1275

------------------------------------------------------------------------------
     numfish |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      camper |   .6211286   .2358072     2.63   0.008      .158955    1.083302
     persons |     1.0608   .1174733     9.03   0.000     .8305564    1.291043
    children |   -1.78052   .1920379    -9.27   0.000    -2.156907   -1.404132
       _cons |   -1.62499   .3294006    -4.93   0.000    -2.270603   -.9793765
-------------+----------------------------------------------------------------
    /lnalpha |   .7688868   .1538497                      .4673469    1.070427
-------------+----------------------------------------------------------------
       alpha |   2.157363   .3319098                      1.595755    2.916624
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) =  863.70 Prob>=chibar2 = 0.000
```

(f) The coefficients are easiest to interpret on the mean scale (after exponentiating) rather than on the log scale. For the camping variable we have $e^{.62} = 1.86$ meaning people who are camping are expected to

catch 86% more or 1.86 times as many fish as people who do not camp. For the persons variable we have $e^{1.06} = 2.89$ which means the group is predicted to catch 2.89 times more fish for each additional person in the party. For the children variable we have $e^{-1.62} = .168$ meaning there is an 83.2% reduction in the mean number of fish caught for each additional child in the party. The effects are all (not surprisingly) in the same direction as before. Larger groups who camp catch more; people who bring children catch less. The magnitudes of the effects are fairly similar to those in the Poisson model–the effect of camping is a bit smaller, the effect of additional people is about the same as before, and the effect of children is a bit bigger. The variables are still all highly significant but in fact the Z statistics have gotten a lot smaller. This is hardly a surprise–in our Poisson model we know our standard error estimates were a lot smaller than they should have been so the resulting Z scores were larger than they should have been and the p-values smaller than they should have been. It is only because in both models the variables are so significant that most of the p-values in the negative binomial model are still at .000.

(g) As noted above, zero-inflation happens when we get more 0's than we would expect from our basic model and is one factor that can lead to over-dispersion. Here zero-inflation means that we get more people who catch no fish than would be expected from a Poisson or negative binomial model. This could occur because there are some people who are simply not fishing. It seems that people who are not camping (i.e. are just staying for the day) or who have brought children are more likely not to be fishing so these are the variables that I would intuitively expect to be important for the inflation factor part of the model that predicts which observations are "certain zeros." However if you consider that any given person is equally likely to be a fisher then groups with fewer people would be more likely to have no fishers in them than large groups and thus could also be more likely to be certain 0's. We won't really be able to tell until we experiment with the models. In fact the ATS web site which demos this data set ends up using the "person" variable as the predictor for the inflation part of the model and the camping and children variables as the predictors of the actual catch size. You can make an argument that you don't want the person and children variables in the same part of the model since they are likely to be related and cause multicollinearity problems.

(h) We could get the actual number of zeros for each of the 32 subgroups considered in part (d) but that would end up being many pages of printouts. I picked out a few of the largest groups to include here to illustrate the point. For instance, the biggest category is single people (no children) camping. There are 35 such groups. 21 of these groups caught no fish. Another large group is campers with 1 parent and 1 child. There are 23 such groups and 16 of them caught no fish. There were also several subsets with 2 or 3 children in which none of the groups caught any fish. Our model predicts that single campers should catch about 1 fish on average.

$$\mu = e^{-1.62+.62(1)+1.06(1)-1.78(0)} = e^{.06} = 1.06$$

Using the Poisson probability formula the chance such group catches no fish should be

$$(1.06)^0 e^{-1.06}/0! = .35$$

Thus we expect just over a third of such groups to catch 0 fish. Instead 21/35 or nearly 2/3rds of these groups caught no fish–this is definitely more than we would have expected. The calculation is pretty much the same if we use the actual mean for this subgroup rather than the model predicted mean but the latter makes more sense since we are trying to see if the model doesn't adequately account for the 0's. Similar calculations can be done for the other groups. They all suggest a lot of zero inflation. In fact, if we look at the data set overall, 142 or 56.8% of the 250 groups caught no fish. The mean number of fish caught was 3.3 (see below). With this mean we would expect the fraction of 0's to be only

$$(3.30)^0 e^{-3.30}/0! = .037$$

or under 4%!! We clearly have many more 0's than that and it is not completely accounted for by adjusting for the number of children, people and camping status in our Poisson model. A pure Poisson distribution just doesn't make sense here.

```
. table numfish if camper==1 & person==1
----------------------
  numfish |      Freq.
----------+-----------
        0 |         21
        1 |          7
        2 |          2
        3 |          2
        4 |          2
        7 |          1
----------------------

. table numfish if camper==1 & person==2 & children==1
----------------------
  numfish |      Freq.
----------+-----------
        0 |         16
        1 |          3
        2 |          2
        4 |          1
        5 |          1
----------------------

. table numfish
----------------------
  numfish |      Freq.
----------+-----------
        0 |        142
        1 |         31
        2 |         20
        3 |         12
        4 |          6
        5 |         10
        6 |          4
        7 |          3
        8 |          2
        9 |          2
       10 |          1
       11 |          1
       13 |          1
       14 |          1
       15 |          2
       16 |          1
       21 |          2
       22 |          1
       29 |          1
       30 |          1
```

```
       31 |          1
       32 |          2
       38 |          1
       65 |          1
      149 |          1
----------------------
```

```
. summarize numfish

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     numfish |       250       3.296    11.63503          0        149
```

(i) The printout for the zero-inflated Poisson model is shown below. Consistent with our earlier summary statistics, it suggests that zero-inflation may be a significant issue in this data set. The log likelihood is a lot higher than in the plain Poisson model, suggesting the zero inflated model fits better (although as we discussed in class, there are some technical issues associated with this comparison so we shouldn't view this as a formal test.) Moreover, all the terms in the inflation component of the model are highly significant which again suggests that the zero-inflated Poisson model is doing something useful and in particular the predictors are giving us some extra leverage on who is likely to catch no fish. (Note that the ZIP model fitting better might not solely reflect zero-inflation/be the "right" model; there could be other things like general overdispersion that make it a better "match" to the data/underlying distribution. However, it certainly seems like a better choice than the plain Poisson, regardless of the reason.)

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

Fitting constant-only model:
Fitting full model:
Zero-inflated Poisson regression                  Number of obs   =        250
                                                  Nonzero obs     =        108
                                                  Zero obs        =        142

Inflation model = logit                           LR chi2(3)      =     658.63
Log likelihood  = -752.7315                       Prob > chi2     =     0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
numfish      |
      camper |   .7242542    .093144     7.78   0.000     .5416953     .906813
     persons |   .8290424   .0439535    18.86   0.000      .742895    .9151897
    children |   -1.13666    .092988   -12.22   0.000    -1.318913   -.9544065
       _cons |  -.7982616   .1708072    -4.67   0.000    -1.133038   -.4634856
-------------+----------------------------------------------------------------
inflate      |
      camper |  -.8336283   .3526521    -2.36   0.018    -1.524814   -.1424428
     persons |  -.9227879    .199216    -4.63   0.000    -1.313244   -.5323317
    children |   1.904574    .326105     5.84   0.000      1.26542    2.543728
       _cons |   1.663579   .5155314     3.23   0.001     .6531564    2.674002
------------------------------------------------------------------------------
```

(j) We can interpret the coefficients for the first part of the model (in the box labeled "numfish") as if they

are coefficients for a Poisson model, restricted to the people who are fishing/could ever catch fish. As usual the interpretations are easier if we exponentiate. We see that among people who are fishing, camping is associated with a factor of $e^{.72} = 2.05$ increase in the average number of fish caught all else equal. Similarly, for each additional person in the group the number of fish caught goes up by a factor of $e^{.83} = 2.29$ and for each additional child the multiplicative factor is $e^{-1.13} = .32$, corresponding to a 67.7% reduction in the number of fish, all else equal. All three variables are significant in this component of the model.

The coefficients for the "inflation" portion of the model correspond to a logistic model for whether or not the group would ever catch fish. Specifically, it models the probability of being a "certain 0." We see that camping and having more people are associated with a lower probability of being a certain 0 (negative coefficients) while having more children is associated with a higher probability of being a certain 0 (positive coefficient.) This is exactly what we would expect and indeed all three variables are significant in this portion of the model. To interpret the coefficients more precisely it helps to exponentiate them to obtain odds ratios. The odds ratio for the camper variable is $e^{-.83} = .436$. This means that all else equal a group that is camping has 56.4% lower odds of being a "certain 0 fish" group than a group that does not camp. Similarly, each additional member of the party is associated with a 60% reduction in the odds of being a certain 0 (odds ratio $e^{-.92} = .399$) and each additional child is associated with 6.7 times higher odds of being a certain 0 (odds ratio $e^{1.90} = 6.69$.)

The implication of all this is that groups that don't camp and are small or have lots of children are both more likely to never catch fish and even when they could catch fish tend to catch fewer of them then groups that camp, have more people or fewer children.

(k) The printouts for three versions of the zero-inflated negative binomial model are shown below. The first uses all the predictors in both parts of the model. This model has the best log likelihood but turns out to be a little unstable which you see in the fitting iterations STATA printouts out. Moreover, even though the log likelihood looks a lot better. none of the individual components of the inflation part of the model were significant which makes it a bit hard to interpret. As we discussed earlier there may be some multicollinearity among these variables and in this model, unlike the zip model in the previous parts, once we use a negative binomial model, the variables do not seem to each provide unique information about the likelihood of being a certain 0. A bit of experimenting suggests that the second model with persons as the inflation predictor and camper and children as the number of fish predictors works fairly well. In both of these models both the likelihood ratio chi-squared test of the overdispersion factor, alpha, is significant, indicating that this model is superior to the various versions we have fit previously. From the second version of the model we get the interpretation that the more people there are the less likely it is that we have a non-fishing group and that among people who could catch fish camping and smaller numbers of children are associated with increased catch. This is a little different from our interpretation above in which all three variables contributed to both components. It seems that in the zip model some of the variables were trying to compensate for the over-dispersion while in the negative binomial model they are not all needed in both components. However given the correlation of .55 between the persons and children variables it is hard to too strongly say which effect is responsible for which component of the model. The model reversing the roles of children and persons also fits quite well–in fact it's likelihood is somewhat better than the model where the persons variable is the inflation variable (look at the log likelihood) but it also has a few signs of instability. A certain amount of playing around is necessary to get to a good model and there may be more than one that gives similar answers. Here the overall picture is pretty clear. We have zero-inflation and overdispersion so we want to use a zinb model and the direction of all the effects is clear.

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

Fitting constant-only model:
Fitting full model:
```

```
Zero-inflated negative binomial regression        Number of obs   =        250
                                                  Nonzero obs     =        108
                                                  Zero obs        =        142

Inflation model = logit                           LR chi2(3)      =      78.84
Log likelihood  = -395.5392                       Prob > chi2     =     0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
numfish      |
      camper |   .3855611   .2461125     1.57   0.117    -.0968105    .8679328
     persons |   1.090075   .1116773     9.76   0.000     .8711915    1.308958
    children |  -1.261222    .247326    -5.10   0.000    -1.745972   -.7764725
       _cons |   -1.61765   .3202037    -5.05   0.000    -2.245238   -.9900624
-------------+----------------------------------------------------------------
inflate      |
      camper |  -15.23543   599.7906    -0.03   0.980    -1190.803    1160.333
     persons |   .2907057   .7314388     0.40   0.691    -1.142888    1.724299
    children |   15.41647   599.7889     0.03   0.979    -1160.148    1190.981
       _cons |  -16.45837    599.798    -0.03   0.978    -1192.041    1159.124
-------------+----------------------------------------------------------------
    /lnalpha |   .5928722   .1579517     3.75   0.000     .2832926    .9024518
-------------+----------------------------------------------------------------
       alpha |   1.809177   .2857626                      1.327494    2.465641
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0: chibar2(01) =   714.38 Pr>=chibar2 =  0.0000
********************************************************************************
. zinb numfish camper children, inflate(persons) zip

Fitting constant-only model:

Fitting full model:

Zero-inflated negative binomial regression        Number of obs   =        250
                                                  Nonzero obs     =        108
                                                  Zero obs        =        142

Inflation model = logit                           LR chi2(2)      =      61.72
Log likelihood  = -432.8909                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
numfish      |
      camper |   .8790514   .2692731     3.26   0.001     .3512857    1.406817
    children |  -1.515255   .1955912    -7.75   0.000    -1.898606   -1.131903
       _cons |   1.371048   .2561131     5.35   0.000     .8690758    1.873021
-------------+----------------------------------------------------------------
inflate      |
     persons |  -1.666563   .6792833    -2.45   0.014    -2.997934   -.3351922
       _cons |   1.603104   .8365065     1.92   0.055     -.036419    3.242626
```

```
-------------+----------------------------------------------------------------
    /lnalpha |   .9853533     .17595     5.60   0.000     .6404975    1.330209
-------------+----------------------------------------------------------------
       alpha |   2.678758   .4713275                      1.897425    3.781834
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0: chibar2(01) =  1197.43 Pr>=chibar2 =  0.0000
******************************************************************************
 zinb numfish camper persons, inflate(children) zip
Zero-inflated negative binomial regression        Number of obs   =        250
                                                  Nonzero obs     =        108
                                                  Zero obs        =        142

Inflation model = logit                           LR chi2(2)      =      75.10
Log likelihood  = -408.8739                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
numfish      |
      camper |   .6526337   .2465466     2.65   0.008     .1694113    1.135856
     persons |   .9544212   .1061983     8.99   0.000     .7462764    1.162566
       _cons |  -1.682536   .3253132    -5.17   0.000    -2.320138   -1.044934
-------------+----------------------------------------------------------------
inflate      |
    children |   2.980199   .6092847     4.89   0.000     1.786023    4.174375
       _cons |  -3.564681   .7997577    -4.46   0.000    -5.132177   -1.997185
-------------+----------------------------------------------------------------
    /lnalpha |   .5788034   .1730646     3.34   0.001     .2396029    .9180038
-------------+----------------------------------------------------------------
       alpha |   1.783903   .3087304                      1.270744    2.504286
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0: chibar2(01) =   928.99 Pr>=chibar2 =  0.0000
******************************************************************************
. cor persons children camper
(obs=250)

             |  persons children   camper
-------------+---------------------------
     persons |   1.0000
    children |   0.5463   1.0000
      camper |  -0.0484  -0.0340   1.0000
```