Count Data Analysis Geog210B Winter 2018

Kostas Goulias

2/21/2018

In this document you will find examples of Count Data regression models. Poisson and Negative Binomial are the two key paradigms and their zero inflation counterparts.

These models are good for data that are counts (positive integers with the value zero having a meaning).

Preliminary tasks

We use the same database as your assignments 1 and 2

```
HHfile <- read.csv("~/Desktop/geog210b/SmallHHfile.csv", header=TRUE)
```

```
library(stargazer)
```

```
## Warning: package 'stargazer' was built under R version 3.4.3
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary St
atistics Tables.
```

```
##  R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(HHfile, type = "text", title="Descriptive statistics", median=TRUE,
digits=2, out="table1.txt")
```

```
##
## Descriptive statistics
## ========================================================================
## Statistic   N        Mean       St. Dev.      Min     Median      Max
## ------------------------------------------------------------------------
## SAMPN     42,431 2,588,379.00 1,641,345.00 1,031,985 1,971,814 7,212,388
## INCOM     42,431    13.18         26.29         1         5         99
## HHSIZ     42,431     2.57          1.37         1         2          8
## HHEMP     42,431     1.22          0.88         0         1          6
## HHSTU     42,431     0.64          1.02         0         0          8
## HHLIC     42,431     1.86          0.85         0         2          8
## DOW       42,431     4.02          1.99         1         4          7
## HTRIPS    42,431     8.29          7.78         0         6         99
## Mon       42,431     0.14          0.34         0         0          1
## Tue       42,431     0.14          0.35         0         0          1
## Wed       42,431     0.14          0.35         0         0          1
## Thu       42,431     0.15          0.35         0         0          1
## Fri       42,431     0.14          0.35         0         0          1
## Sat       42,431     0.14          0.35         0         0          1
## Sun       42,431     0.15          0.35         0         0          1
## TotDist   42,431    68.09         118.52      0.00      33.89   5,838.26
## center    42,431     0.28          0.45         0         0          1
## suburb    42,431     0.29          0.45         0         0          1
## exurb     42,431     0.23          0.42         0         0          1
## rural     42,431     0.20          0.40         0         0          1
## other     42,431     0.00          0.00         0         0          0
## highinc   42,431     0.41          0.49         0         0          1
## HHVEH     42,431     1.86          1.00         0         2          8
## HHBIC     42,431     1.58          3.79         0         1         99
## VEHNEW    42,431     2.15          2.02         1         2          9
## OWN       42,431     1.24          0.56         1         1          9
## CarBuy    42,431     0.45          0.50         0         0          1
## snglhm    42,431     0.82          0.39         0         1          1
## ownhm     42,431     0.77          0.42         0         1          1
## MilesPr   42,431    27.12         43.46       0.00      14.50   1,167.65
## TrpPrs    42,431     3.28          2.58       0.00      3.00      32.00
## ------------------------------------------------------------------------
```

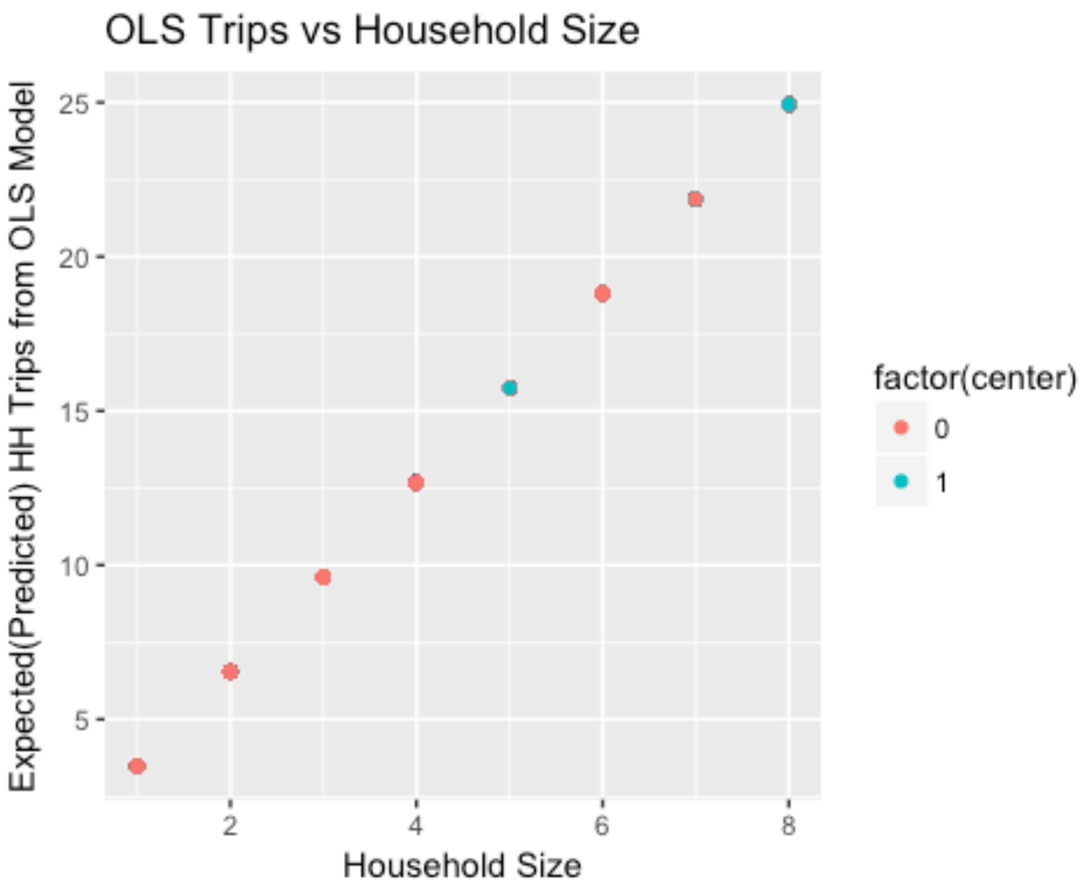
The count variable we want to analyze is HTRIPS

Let's run an ordinary least squares regression model to use as reference

```
OLS1 = lm(HTRIPS ~ HHSIZ, data=HHfile)
summary(OLS1)
```

```
##
## Call:
## lm(formula = HTRIPS ~ HHSIZ, data = HHfile)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.933  -3.604  -0.538   3.330  79.264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.40683    0.06735   6.041 1.55e-09 ***
## HHSIZ        3.06574    0.02310 132.720  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.537 on 42429 degrees of freedom
## Multiple R-squared:  0.2934, Adjusted R-squared:  0.2933
## F-statistic: 1.761e+04 on 1 and 42429 DF,  p-value: < 2.2e-16
```

```
library(ggplot2)
OLS1fit = fitted(OLS1)
Plot0 <- ggplot(data = HHfile, aes(x = HHSIZ, y = OLS1fit, col=factor(center)
))
Plot0 <- Plot0 + geom_point()
Plot0 <- Plot0 + xlab("Household Size") + ylab("Expected(Predicted) HH Trips
from OLS Model") + ggtitle("OLS Trips vs Household Size")
Plot0
```



The "nature" of the data HTRIPS is a positive integer that has a clear meaning at zero (no travel = stay home all day). These are also called episodes = something happened or an occurrence.

Poisson Models

One possible model is the Poisson Regression Model

Poisson (from book chapter on gauchospace)

$$P(y_i) = \frac{EXP(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

$$\lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i) \text{ or, equivalently } LN(\lambda_i) = \boldsymbol{\beta}\mathbf{X}_i,$$

Lambda is the mean and variance of y_i per unit time

$$L(\boldsymbol{\beta}) = \prod_i \frac{EXP\left[-EXP(\boldsymbol{\beta}\mathbf{X}_i)\right]\left[EXP(\boldsymbol{\beta}\mathbf{X}_i)\right]^{y_i}}{y_i!}. \qquad LL(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[-EXP(\boldsymbol{\beta}\mathbf{X}_i) + y_i\boldsymbol{\beta}\mathbf{X}_i - LN(y_i!)\right]$$

Poisson Regression Equations

I will need some added librarries for these models

```
library(car)

## Warning: package 'car' was built under R version 3.4.3

library(stats)
library(MASS)
library(Zelig)

## Warning: package 'Zelig' was built under R version 3.4.2

## Loading required package: survival

library(margins)
```

POISSON MODEL

(glm)

```
pmodel1 = glm(HTRIPS ~ HHSIZ , family=poisson, data=HHfile)
summary(pmodel1)
```

```
##
## Call:
## glm(formula = HTRIPS ~ HHSIZ, family = poisson, data = HHfile)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.7463  -1.7383  -0.3292   1.0758  13.5939
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.246830   0.003700   337.0   <2e-16 ***
## HHSIZ       0.299662   0.001007   297.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 289764  on 42430  degrees of freedom
## Residual deviance: 211355  on 42429  degrees of freedom
## AIC: 352426
##
## Number of Fisher Scoring iterations: 5
```

```
anova(pmodel1)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: HTRIPS
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev
## NULL                  42430     289764
## HHSIZ  1    78409     42429     211355
```

The Null deviance is in essence - Twice the log likelihood of the model with a constant only

The Residual Deviance of the model is -Twice the log likelihood at convergence (this means when the iterations reached the maximum of the log likelihood function)

The name deviance is coming from its derivation that compares a model to one that uses all the degrees of freedom to fit the data perfectly (on Gauchospace there is a short note on this).

The most important thing here is to compute in the model above the difference between the Null Deviance (=289764) which is the deviance value we get when we run a Poisson model with only a constant and the deviance of pmodel1 which is 211355. The difference between these two is: 78409 (this is also reported by the Anova table). The pmodel1 has one additional regression coefficient than the null model. This means it uses one additional degree of freedom to estimate a parameter.

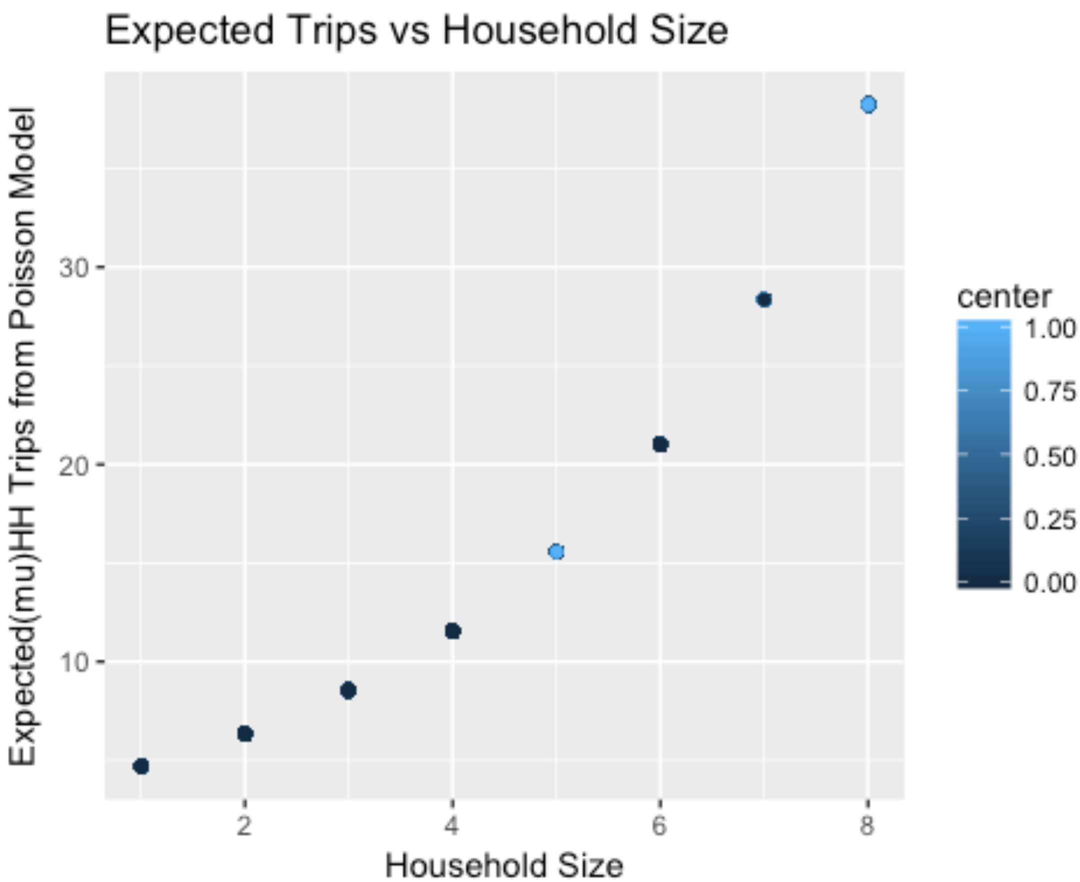
It has been shown that the difference of deviances between models that are relatives (pmodel1 is in essence the null model with an added regression coefficient) is Chi-square distributed with degrees of freedom the difference in the number of estimated coefficients between the Null model and pmodel1 (which in this case is 1 because we estimated the coefficient for HHSIZ). A chi-square statistic value of 78409 is a huge number when compared to the chi-square critical value.

Bottom line: just adding one explanatory variable in this Poisson model makes a HUGE difference in fitting the data we are given.

We can compare models using the deviance reported in the R libraries.

The comparison based on difference of deviances and difference on degrees of freedom is called the Likelihood Ratio test.

```
PoiMean1 <-fitted.values(pmodel1)
Plot2 <- ggplot(data = HHfile, aes(x = HHSIZ, y = PoiMean1, col=center))
Plot2 <- Plot2 + geom_point()
Plot2 <- Plot2 + xlab("Household Size") + ylab("Expected(mu)HH Trips from Poi
sson Model") + ggtitle("Expected Trips vs Household Size")
Plot2
```



The derivative of the number of trips of each household with respect to its household size changes with the household size. This means that the difference in number of trips between two households that one has household size 2 and the other household size 3 is different than the difference between two households that one has household size 4 and the other has household size 5.

In equations this is:

$$\frac{\partial E[y_i | x_i]}{\partial x_i} = \lambda_i \beta$$

$$\lambda_i = \exp(\beta' x_i)$$

Poisson Marginal Effects Equations

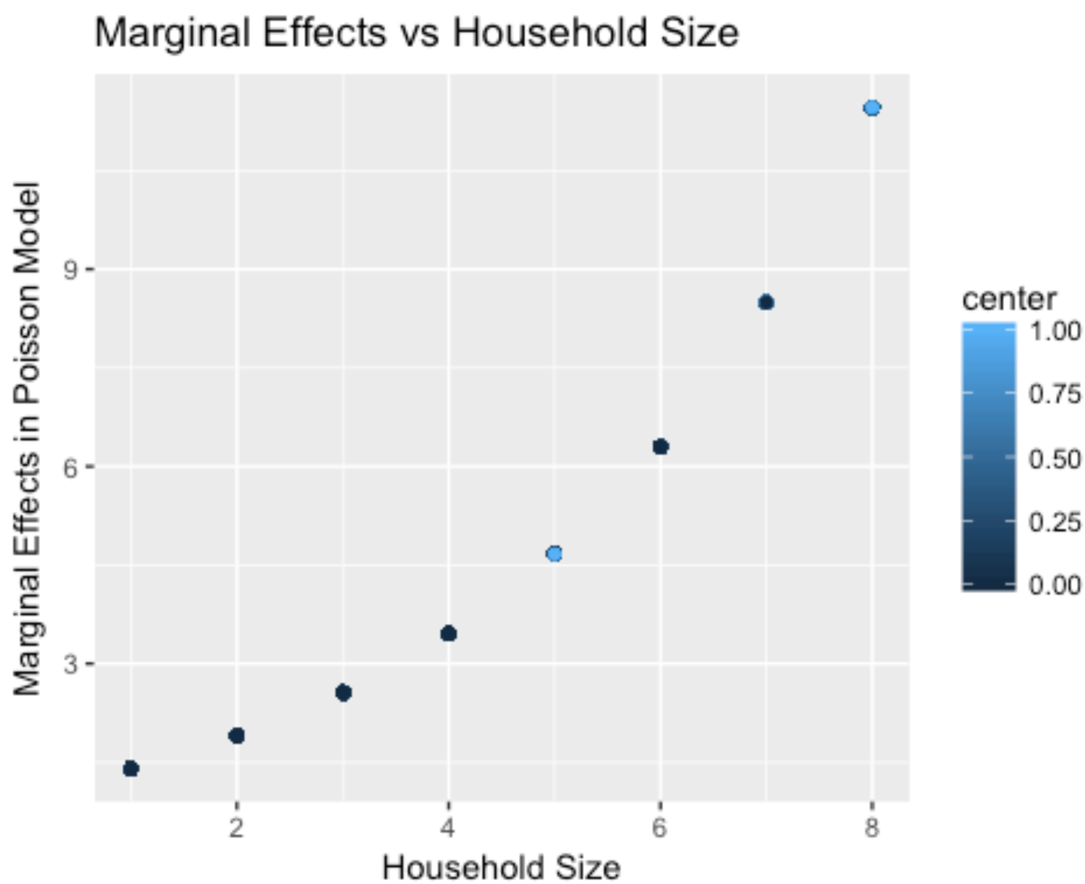
We can compute the derivative for each observation using mfx and margins libaries in R.

```
marginalEffectspmodel1 <- marginal_effects(pmodel1)
summary(marginalEffectspmodel1)
```

```
##    dydx_HHSIZ
##  Min.   : 1.407
##  1st Qu.: 1.898
##  Median : 1.898
##  Mean   : 2.484
##  3rd Qu.: 2.562
##  Max.   :11.462
```

```
Plot3 <- ggplot(data = HHfile, aes(x = HHSIZ, y = marginalEffectspmodel1, col
=center))
Plot3 <- Plot3 + geom_point()
Plot3 <- Plot3 + xlab("Household Size") + ylab("Marginal Effects in Poisson M
odel") + ggtitle("Marginal Effects vs Household Size")
Plot3

## Don't know how to automatically pick scale for object of type data.frame.
Defaulting to continuous.
```



The Poisson model assumes its mean and variance are the same. This is too restrictive and one way to "release" this restriction is to use a Negative Binomial model.

This non-linear regression models is defined by:

Negative Binomial Model

$$\lambda_i = EXP(\boldsymbol{\beta}X_i + \varepsilon_i),$$

$$VAR[y_i] = E[y_i][1 + \alpha E[y_i]] = E[y_i] + \alpha E[y_i]^2.$$

$$P(y_i) = \frac{\Gamma((1/\alpha) + y_i)}{\Gamma(1/\alpha) y_i!} \left(\frac{1/\alpha}{(1/\alpha) + \lambda_i} \right)^{1/\alpha} \left(\frac{\lambda_i}{(1/\alpha) + \lambda_i} \right)^{y_i}$$

$$L(\lambda_i) = \prod_i \frac{\Gamma((1/\alpha) + y_i)}{\Gamma(1/\alpha) y_i!} \left(\frac{1/\alpha}{(1/\alpha) + \lambda_i} \right)^{1/\alpha} \left(\frac{\lambda_i}{(1/\alpha) + \lambda_i} \right)^{y_i} .$$

6

NegBin Model Equations

Upper case gamma is the Gamma function.

The important item in these equations is the Var(y) which is a function of the mean and a function of the square of the mean.

```r
pmodel2 = glm.nb(HTRIPS ~ 1 +HHSIZ , data=HHfile)
summary(pmodel2)
```

```
##
## Call:
## glm.nb(formula = HTRIPS ~ 1 + HHSIZ, data = HHfile, init.theta = 1.7138964
95,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4093  -0.7893  -0.0834   0.4273   3.9895
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.102985   0.008895   124.0   <2e-16 ***
## HHSIZ       0.349064   0.002936   118.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.7139) family taken to be 1)
##
##     Null deviance: 64996  on 42430  degrees of freedom
## Residual deviance: 51119  on 42429  degrees of freedom
## AIC: 257183
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.7139
##           Std. Err.:  0.0160
##
##  2 x log-likelihood:  -257176.8800
```

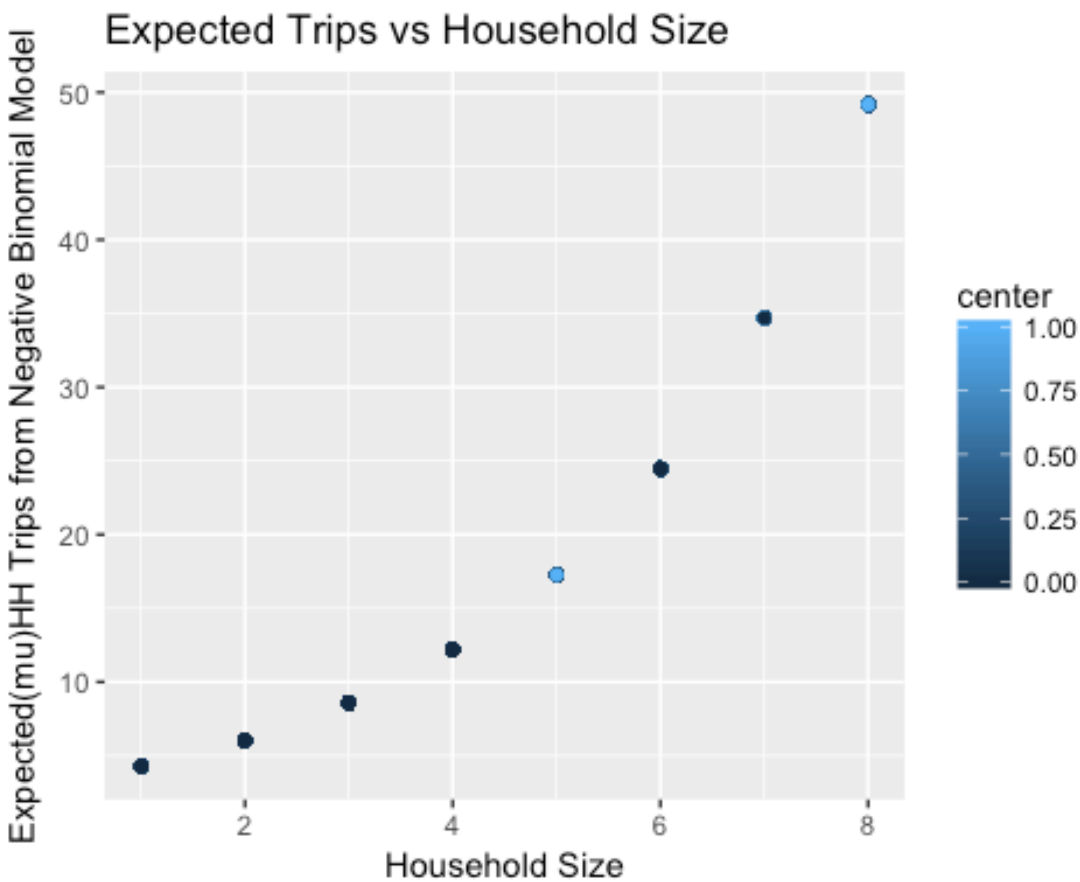
The reported Theta is the parameter indicated as alpha in the negative binomial equations. It is 1.7139 and if we take the ratio of 1.7139 over its standard error (0.0160) we get a big number telling us theta is significantly different than zero and we have overdispersion. This means the mean and variance are not the same and the negative binomial is a better representation f the data we have.

```
stargazer(pmodel1, pmodel2,  type="text", title="Regression Results",
          dep.var.labels=c("Number of Trips per Household"),
          covariate.labels=c("Household Size"), out="output1.txt")

##
## Regression Results
## =================================================
##                          Dependent variable:
##                    ------------------------------
##                    Number of Trips per Household
##                       Poisson        negative
##                                      binomial
##                         (1)            (2)
## -------------------------------------------------
## Household Size      0.300***       0.349***
##                      (0.001)        (0.003)
##
## Constant            1.247***       1.103***
##                      (0.004)        (0.009)
##
## -------------------------------------------------
## Observations         42,431         42,431
## Log Likelihood    -176,211.100   -128,589.400
## theta                          1.714*** (0.016)
## Akaike Inf. Crit.  352,426.200    257,182.900
## =================================================
## Note:                 *p<0.1; **p<0.05; ***p<0.01
```

So after all this work does the best model fit the data?

```
NegMean1 <-fitted.values(pmodel2)
Plot3 <- ggplot(data = HHfile, aes(x = HHSIZ, y = NegMean1, col=center))
Plot3 <- Plot3 + geom_point()
Plot3 <- Plot3 + xlab("Household Size") + ylab("Expected(mu)HH Trips from Neg
ative Binomial Model") + ggtitle("Expected Trips vs Household Size")
Plot3
```



The negative binomial has one parameter more than the poisson (the theta) and has a residual deviance of 51119. Recall the Poisson model has a residual deviance of 211355. So, the negative binomial is an improvement of 211355-51119 in its deviance. By far better model than the Poisson.

```
pmodel3 = glm.nb(HTRIPS ~ 1 +HHSIZ + HHVEH + center , data=HHfile)
summary(pmodel3)

##
## Call:
## glm.nb(formula = HTRIPS ~ 1 + HHSIZ + HHVEH + center, data = HHfile,
##     init.theta = 1.732289339, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4971  -0.8061  -0.1545   0.4561   4.1865
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.014079   0.011014  92.074  < 2e-16 ***
## HHSIZ       0.346010   0.003162 109.414  < 2e-16 ***
## HHVEH       0.025602   0.004504   5.685 1.31e-08 ***
## center      0.164955   0.009274  17.786  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.7323) family taken to be 1)
##
##     Null deviance: 65459  on 42430  degrees of freedom
## Residual deviance: 51138  on 42427  degrees of freedom
## AIC: 256865
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.7323
##           Std. Err.:  0.0162
##
##  2 x log-likelihood:  -256855.3150

anova(pmodel3)

## Warning in anova.negbin(pmodel3): tests made without re-estimating 'theta'

## Analysis of Deviance Table
##
## Model: Negative Binomial(1.7323), link: log
##
## Response: HTRIPS
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                   42430      65459
## HHSIZ   1  13998.4     42429      51461  < 2e-16 ***
```

```
## HHVEH   1      5.4     42428      51456  0.02045 *
## center  1    317.5     42427      51138  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that we can also compute the contribution of every variable in helping decrease the deviance of the model.

Let's estimate a couple of big models

```
OLS2 = lm(HTRIPS ~ HHSIZ + HHVEH + highinc + Mon + Tue + Wed + Thu + Fri + Sa
t + center + suburb + exurb +HHEMP + HHSTU + HHLIC, data=HHfile)
summary(OLS2)
```

```
##
## Call:
## lm(formula = HTRIPS ~ HHSIZ + HHVEH + highinc + Mon + Tue + Wed +
##     Thu + Fri + Sat + center + suburb + exurb + HHEMP + HHSTU +
##     HHLIC, data = HHfile)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.934  -3.686  -0.795   2.940  79.351
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.14274    0.12898  -8.860  < 2e-16 ***
## HHSIZ        2.10067    0.04609  45.579  < 2e-16 ***
## HHVEH       -0.23910    0.04399  -5.436 5.48e-08 ***
## highinc      1.31432    0.06755  19.457  < 2e-16 ***
## Mon          1.28193    0.11509  11.138  < 2e-16 ***
## Tue          2.41685    0.11354  21.287  < 2e-16 ***
## Wed          2.32913    0.11343  20.534  < 2e-16 ***
## Thu          2.26158    0.11292  20.027  < 2e-16 ***
## Fri          2.21778    0.11442  19.382  < 2e-16 ***
## Sat          0.84541    0.11414   7.407 1.32e-13 ***
## center       1.62999    0.09122  17.868  < 2e-16 ***
## suburb       1.06261    0.08958  11.862  < 2e-16 ***
## exurb        0.68753    0.09386   7.325 2.43e-13 ***
## HHEMP        0.71303    0.04391  16.239  < 2e-16 ***
## HHSTU        1.46971    0.05207  28.227  < 2e-16 ***
## HHLIC       -0.22974    0.06068  -3.786 0.000153 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.305 on 42415 degrees of freedom
## Multiple R-squared:  0.3429, Adjusted R-squared:  0.3426
## F-statistic:  1475 on 15 and 42415 DF,  p-value: < 2.2e-16
```

```r
pmodel4 = glm.nb(HTRIPS ~ HHSIZ + HHVEH + highinc + Mon + Tue + Wed + Thu + F
ri + Sat + center + suburb + exurb +HHEMP + HHSTU + HHLIC , data=HHfile)
summary(pmodel4)
```

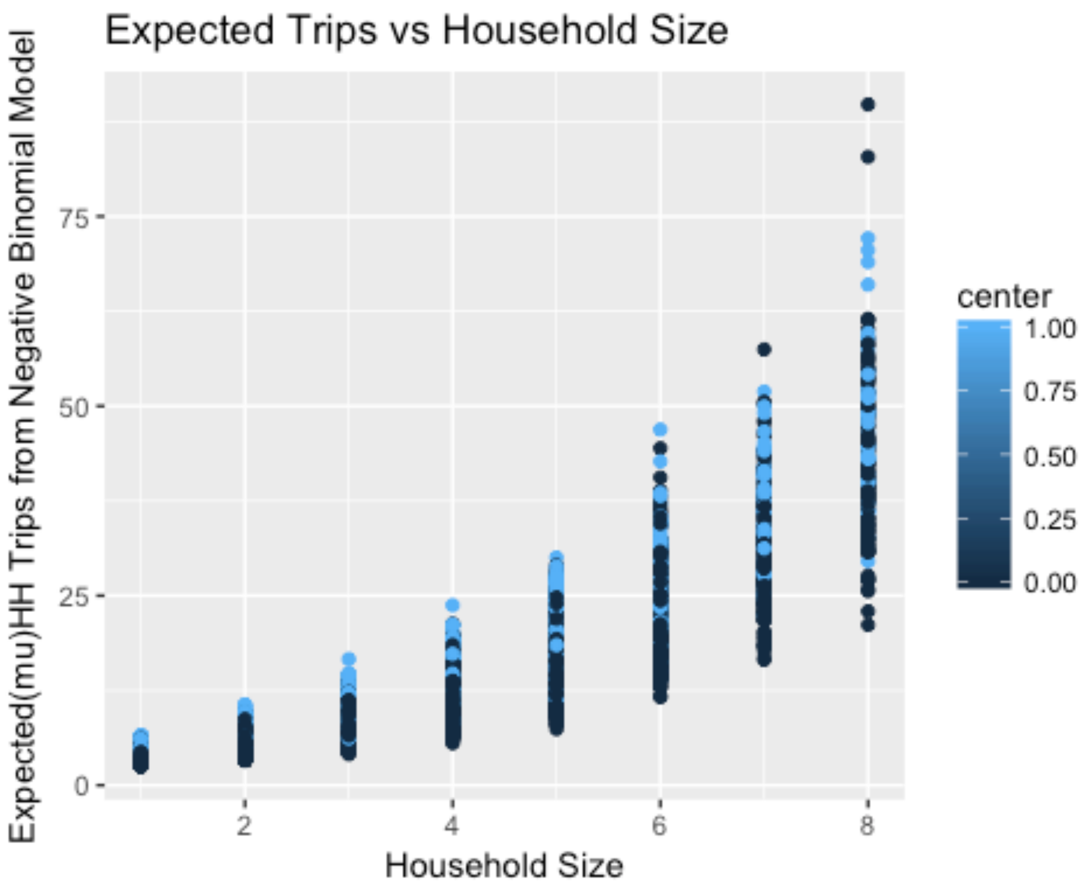
```
##
## Call:
## glm.nb(formula = HTRIPS ~ HHSIZ + HHVEH + highinc + Mon + Tue +
##     Wed + Thu + Fri + Sat + center + suburb + exurb + HHEMP +
##     HHSTU + HHLIC, data = HHfile, init.theta = 1.873602678, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6024  -0.8165  -0.1482   0.4432   4.5365
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.731805   0.017280  42.350  < 2e-16 ***
## HHSIZ        0.253943   0.005841  43.474  < 2e-16 ***
## HHVEH       -0.019778   0.005752  -3.438 0.000585 ***
## highinc      0.168640   0.008777  19.214  < 2e-16 ***
## Mon          0.186770   0.015280  12.224  < 2e-16 ***
## Tue          0.304132   0.014977  20.307  < 2e-16 ***
## Wed          0.293657   0.014974  19.611  < 2e-16 ***
## Thu          0.286035   0.014914  19.179  < 2e-16 ***
## Fri          0.284701   0.015115  18.836  < 2e-16 ***
## Sat          0.122100   0.015210   8.027 9.95e-16 ***
## center       0.230827   0.012038  19.175  < 2e-16 ***
## suburb       0.152304   0.011839  12.865  < 2e-16 ***
## exurb        0.110151   0.012422   8.868  < 2e-16 ***
## HHEMP        0.113573   0.005703  19.916  < 2e-16 ***
## HHSTU        0.098768   0.006566  15.041  < 2e-16 ***
## HHLIC        0.008202   0.007765   1.056 0.290847
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.8736) family taken to be 1)
##
##     Null deviance: 68935  on 42430  degrees of freedom
## Residual deviance: 51437  on 42415  degrees of freedom
## AIC: 254720
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.8736
##           Std. Err.:  0.0181
##
##  2 x log-likelihood:  -254685.9980
```

```r
NegMean4 <-fitted.values(pmodel4)

marginalEffectspmodel4 <- marginal_effects(pmodel4)
summary(marginalEffectspmodel4)

##    dydx_HHSIZ         dydx_HHVEH         dydx_highinc       dydx_Mon
##  Min.   : 0.6339   Min.   :-1.77517   Min.   : 0.421   Min.   : 0.4662
##  1st Qu.: 1.2619   1st Qu.:-0.20028   1st Qu.: 0.838   1st Qu.: 0.9281
##  Median : 1.6658   Median :-0.12973   Median : 1.106   Median : 1.2251
##  Mean   : 2.1450   Mean   :-0.16706   Mean   : 1.424   Mean   : 1.5776
##  3rd Qu.: 2.5715   3rd Qu.:-0.09828   3rd Qu.: 1.708   3rd Qu.: 1.8913
##  Max.   :22.7928   Max.   :-0.04937   Max.   :15.136   Max.   :16.7637
##     dydx_Tue          dydx_Wed          dydx_Thu          dydx_Fri
##  Min.   : 0.7592   Min.   : 0.7331   Min.   : 0.714   Min.   : 0.7107
##  1st Qu.: 1.5113   1st Qu.: 1.4593   1st Qu.: 1.421   1st Qu.: 1.4148
##  Median : 1.9950   Median : 1.9263   Median : 1.876   Median : 1.8675
##  Mean   : 2.5689   Mean   : 2.4804   Mean   : 2.416   Mean   : 2.4048
##  3rd Qu.: 3.0798   3rd Qu.: 2.9737   3rd Qu.: 2.897   3rd Qu.: 2.8830
##  Max.   :27.2975   Max.   :26.3574   Max.   :25.673   Max.   :25.5535
##     dydx_Sat         dydx_center       dydx_suburb        dydx_exurb
##  Min.   : 0.3048   Min.   : 0.5762   Min.   : 0.3802   Min.   :0.2750
##  1st Qu.: 0.6067   1st Qu.: 1.1470   1st Qu.: 0.7568   1st Qu.:0.5474
##  Median : 0.8009   Median : 1.5141   Median : 0.9991   Median :0.7225
##  Mean   : 1.0313   Mean   : 1.9497   Mean   : 1.2865   Mean   :0.9304
##  3rd Qu.: 1.2364   3rd Qu.: 2.3375   3rd Qu.: 1.5423   3rd Qu.:1.1154
##  Max.   :10.9591   Max.   :20.7180   Max.   :13.6702   Max.   :9.8867
##    dydx_HHEMP         dydx_HHSTU        dydx_HHLIC
##  Min.   : 0.2835   Min.   :0.2466   Min.   :0.02047
##  1st Qu.: 0.5644   1st Qu.:0.4908   1st Qu.:0.04076
##  Median : 0.7450   Median :0.6479   Median :0.05380
##  Mean   : 0.9593   Mean   :0.8343   Mean   :0.06928
##  3rd Qu.: 1.1501   3rd Qu.:1.0002   3rd Qu.:0.08305
##  Max.   :10.1938   Max.   :8.8650   Max.   :0.73615

NegMean4 <-fitted.values(pmodel4)
Plot4 <- ggplot(data = HHfile, aes(x = HHSIZ, y = NegMean4, col=center))
Plot4 <- Plot4 + geom_point()
Plot4 <- Plot4 + xlab("Household Size") + ylab("Expected(mu)HH Trips from Neg
ative Binomial Model") + ggtitle("Expected Trips vs Household Size")
Plot4
```



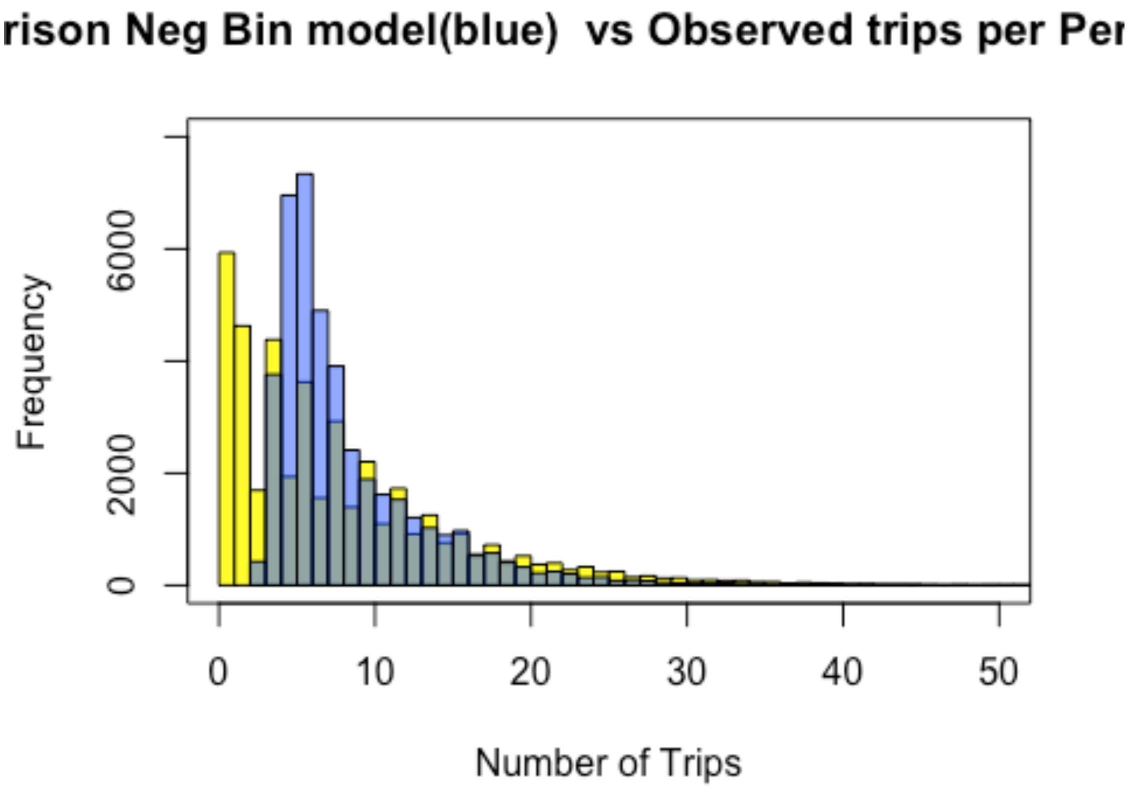

```
stargazer(pmodel2, pmodel3,pmodel4, type="text", title="Regression Results",
          dep.var.labels=c("Number of Trips per Household"),
          covariate.labels=c("Household Size","Household Cars","High Income", "Monday
",
                              "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"
,
                              "Residence in Center", "Residence in Suburb", "Residence
in Exurb",
                              " Number of Employed", "Number of Students", "Number of
Driver Licenses"), out="output1.txt")
```

```
##
## Regression Results
## ===========================================================================
==
##                                           Dependent variable:
##                              -----------------------------------------------
--
##                                   Number of Trips per Household
##                                 (1)              (2)              (3)
## ---------------------------------------------------------------------------
--
## Household Size                0.349***         0.346***         0.254***
##                               (0.003)          (0.003)          (0.006)
##
## Household Cars                                 0.026***        -0.020***
##                                                (0.005)          (0.006)
##
## High Income                                                     0.169***
##                                                                 (0.009)
##
## Monday                                                          0.187***
##                                                                 (0.015)
##
## Tuesday                                                         0.304***
##                                                                 (0.015)
##
## Wednesday                                                       0.294***
##                                                                 (0.015)
##
## Thursday                                                        0.286***
##                                                                 (0.015)
##
## Friday                                                          0.285***
##                                                                 (0.015)
##
## Saturday                                                        0.122***
##                                                                 (0.015)
##
## Residence in Center                            0.165***         0.231***
##                                                (0.009)          (0.012)
##
## Residence in Suburb                                             0.152***
##                                                                 (0.012)
##
## Residence in Exurb                                              0.110***
##                                                                 (0.012)
##
## Number of Employed                                              0.114***
##                                                                 (0.006)
##
```

```
## Number of Students                                              0.099***
##                                                                 (0.007)
##
## Number of Driver Licenses                                        0.008
##                                                                 (0.008)
##
## Constant                      1.103***         1.014***         0.732***
##                               (0.009)          (0.011)          (0.017)
##
## ----------------------------------------------------------------------------
--
## Observations                   42,431           42,431           42,431
## Log Likelihood              -128,589.400     -128,428.700     -127,344.000
## theta                     1.714*** (0.016) 1.732*** (0.016) 1.874*** (0.01
8)
## Akaike Inf. Crit.           257,182.900      256,865.300      254,720.000
## ============================================================================
==
## Note:                                              *p<0.1; **p<0.05; ***p<0.
01
```

Let's see if our best model does a good job in replicating observed values

```
hist(HHfile$HTRIPS, col=rgb(1,1,0,0.9),breaks=100, xlim=c(0,50),
      ylim=c(0,8000), xlab="Number of Trips", main="Comparison Neg Bin model(
blue)  vs Observed trips per Person (Gold) ")
hist(NegMean4, col=rgb(0,0.3,1,0.5),breaks=100, add=T)
 box()
```



This shows that maybe we have two distributions that are "mixed" in the same data.

Maybe some households are consistently staying home all day. Maybe we interviewed many people during vacation days (we actually did in CHTS). The models that can account for this type of issue are called Zero Inflated.

We need a new library called pscl

```
library(pscl)

## Warning: package 'pscl' was built under R version 3.4.2

## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

The model below has two component. One component is the same as the negative binomial above. The second component estimates a binary model that classifies observations based on their having a zero or not having a zero in the HTRIPS. In this models specification we include Sat, Sun, and rural as explanatory variables because we think that maybe people stay home during weekend days and people that live in rural environment might combine all their errands in one day and then home the next.

Look at the explanatory variables. They are separated by a vertical line.

```
zinbTrips <- zeroinfl(HTRIPS~HHSIZ + HHVEH + highinc +
                        Mon + Tue + Wed + Thu + Fri + Sat + center + suburb +
exurb +
                        HHEMP + HHSTU + HHLIC | Sat + Sun + rural, dist="negb
in", data=HHfile)
```

```r
summary(zinbTrips)
```

```
##
## Call:
## zeroinfl(formula = HTRIPS ~ HHSIZ + HHVEH + highinc + Mon + Tue +
##     Wed + Thu + Fri + Sat + center + suburb + exurb + HHEMP + HHSTU +
##     HHLIC | Sat + Sun + rural, data = HHfile, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.5732 -0.7312 -0.1520  0.5505 10.9014
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.076307   0.015419  69.803  < 2e-16 ***
## HHSIZ        0.252101   0.004843  52.054  < 2e-16 ***
## HHVEH       -0.018877   0.004787  -3.943 8.04e-05 ***
## highinc      0.143268   0.007152  20.033  < 2e-16 ***
## Mon          0.091292   0.012989   7.028 2.09e-12 ***
## Tue          0.195154   0.012607  15.479  < 2e-16 ***
## Wed          0.190967   0.012626  15.125  < 2e-16 ***
## Thu          0.188539   0.012600  14.963  < 2e-16 ***
## Fri          0.179998   0.012766  14.100  < 2e-16 ***
## Sat          0.102160   0.013121   7.786 6.93e-15 ***
## center       0.177444   0.010177  17.435  < 2e-16 ***
## suburb       0.109746   0.009989  10.986  < 2e-16 ***
## exurb        0.069717   0.010487   6.648 2.98e-11 ***
## HHEMP        0.065482   0.004736  13.827  < 2e-16 ***
## HHSTU        0.079195   0.005340  14.829  < 2e-16 ***
## HHLIC       -0.006799   0.006448  -1.054    0.292
## Log(theta)   1.289845   0.011897 108.414  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.47467    0.02642  -93.66   <2e-16 ***
## Sat          0.71173    0.04508   15.79   <2e-16 ***
## Sun          0.84055    0.04342   19.36   <2e-16 ***
## rural        0.44541    0.03986   11.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 3.6322
## Number of iterations in BFGS optimization: 28
## Log-likelihood: -1.243e+05 on 21 Df
```