Discrete Choice Modeling
William Greene
Stern School of Business, New York University

Lab Session 7
Models for Count Data

The following exercises and examples use the healthcare.lpj data.

1. **Effect of Censoring** . Use the original variable DOCVIS to fit a Poisson regression model. Then, use DOCVIS10, the censored variable and refit the model. What happens to the estimated coefficients?

```
Sample  ; All $
Namelist; X=One,Age,Educ,Income,hhkids,Married$
Poisson ; Lhs = DocVis ; Rhs = X$
Create  ; DocVis10 = Docvis ~ 10 $
Poisson ; Lhs = DocVis10 ; Rhs = X$
```

2. **Poisson model with sample selection**. We fit a model for the number of visits to the hospital by people who have public insurance. This is a time consuming estimator (uses Hermite quadrature) so we reduce the sample size to speed things up. What do you conclude about the 'selectivity?' Is it significant? Interpret the estimated model.

```
Reject  ; ti < 7 $
Probit  ; Lhs = Public ; Rhs = One,Income,Hhkids ; Hold $
Poisson ; Lhs = HospVis ; Rhs = X ; Selection ; MLE  $
```

3. **Poisson and NB models for doctor visits**. Fit the two models and compare the results. Is there evidence of overdispersion in the results? Fit the Poisson Hurdle model with the same specification. Do the results change significantly?

```
Sample  ; All $
Poisson ; Lhs = DocVis ; Rhs = X $
Poisson ; Lhs = DocVis ; Rhs = X ; Hurdle $
Negbin  ; Lhs = DocVis ; Rhs = X $
Negbin  ; Lhs = DocVis ; Rhs = X ; Model = NB1 $
Negbin  ; Lhs = DocVis ; Rhs = X ; Model = NBP $
```

The last three commands fit three forms of the negative binomial model, NB1, NB2, then NBP which nests NB1 and NB2. Based on the log likelihood functions, does NBP reject NB1 and NB2? Which of the two does NBP suggest is preferred, if one must choose between NB1 and NB2?

4. **Latent class models**. Fit a three class latent class Poisson model to the DOCTOR variable (visited the doctor at least once) variable. Now, fit a three class model to the actual count of visits to the doctor. Fitting LCMs with large numbers of classes is ambitious. Though technically, the model is identified for any number of classes – with the RPM being the limiting model, in practice, identification is fairly weak once the number of classes gets even moderately large. Try fitting a five class model to the DOCVIS variable. Note, we have restricted the sample to those who have a full 7 observations. The last model is a three class model in which age is assumed to influence the class probabilities. Do the results suggest that this is an

appropriate model? (The estimator assumes that the LCM variables are time invariant. Since age is not time invariant, the last year of data on age is used.)

```
Sample  ; All$
Logit   ; Lhs=doctor
        ; Rhs=one,age,educ,income,hhkids,married
        ; LCM ; pts=3 ; pds=ti ; maxit=25 $
Poisson ; Lhs=docvis  ? This model must be fit first for start values.
        ; Rhs=one,age,educ,income,hhkids,married;mar$
Poisson ; Lhs=docvis
        ; Rhs=one,age,educ,income,hhkids,married
        ; LCM ; pts=3 ; pds=ti ; maxit=25 ; mar ; par$
Reject  ; ti < 7 $
Poisson ; Lhs = docvis ? This model must be fit first for start values.
        ; rhs=one,age,educ,income,hhkids,married
        ; LCM ; pts=5 ; pds=7 ; maxit=25 ; mar;par$
Poisson ; Lhs = docvis
        ; rhs=one,age,educ,income,hhkids,married
        ; LCM = age ; pts=3 ; pds=7 ; maxit=25 ; mar;par$
```

These examples are based on the credit data, AmEx.lpj.

1. The count of major derogatory reports in these data is dominated by zeros. Thus, we fit a zero inflation model as an alternative to the standard Poisson model. Do the reported statistics suggest that the extension is needed? Test the hypothesis of the basic Poisson against the alternative ZIP model.

```
Poisson ; Lhs = MajorDrg
        ; Rhs = One,Age,Income,Selfempl
        ; ZIP ; Rh2 = One,Adepcnt,OwnRent $
NegBin  ; Lhs = MajorDrg
        ; Rhs = One,Age,Income,Selfempl
        ; ZIP ; Rh2 = One,Adepcnt,OwnRent $
```

2. The second model is a basic count model fit first for the entire sample, then corrected for sample selection using only the cardholders. Does the selection correction produce a substantial change? In terms of the model, does it appear that the behavior of cardholders differs systematically (on average) from the population (sample) at large?

```
Probit   ; lhs = cardhldr
         ; rhs = one,age,income,ownrent,selfempl,acadmos
         ; hold$
Poisson  ; lhs = majordrg
         ; rhs = one,age,income,incper,expratio
         ; select ; mle ; maxit=25 ;  marginal effects$
Poisson  ; lhs = majordrg
         ; rhs = one,age,income,incper,expratio
         ; marginal effects $
```