Categorical Dependent Variable Models

Categorical dependent variables include binary, ordinal, nominal or event count data. In the categorical dependent variable models, the left-hand side (LHS) variable is neither interval nor ratio, while right-hand side (RHS) is a linear function of independent variables. Under these circumstances, ordinary least squares (OLS) method can no longer produce the best linear unbiased estimator (BLUE); that is, OLS is biased and inefficient. Categorical dependent variable models (CDVMs) provide better ways of estimating parameters using the maximum likelihood (ML) method. Logit models use the logistic probability distribution, while probit models assume the cumulated standard normal distribution. The following table summarizes the model in terms of the data scales to be used.

Table 1. Comparison between OLS and CDVMs

<table>
<thead>
<tr>
<th>Model</th>
<th>Dependent (LHS)</th>
<th>Method</th>
<th>Independent (RHS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>Ordinary least squares</td>
<td>Interval or ratio</td>
<td>Moment based method</td>
</tr>
<tr>
<td>CDVMs</td>
<td>Binary response</td>
<td>Binary (0 or 1)</td>
<td>Maximum likelihood method</td>
</tr>
<tr>
<td></td>
<td>Ordinal response</td>
<td>Ordinal (1st, 2nd, 3rd...)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nominal response</td>
<td>Nominal (A, B, C ...)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Event count data</td>
<td>Count (0, 1, 2, 3...)</td>
<td></td>
</tr>
</tbody>
</table>

SAS, STATA, and SPSS have procedures or commands for CDVMs (see the following table). SAS provides various procedures for CDVMs, whereas STATA supports fascinating features for computing marginal effects and predicted probabilities in an easy manner. SPSS does not provide any command for count data models.

Table 2. Comparison of the Procedures and Commands for CDVMs

<table>
<thead>
<tr>
<th>Model</th>
<th>SAS 8.2</th>
<th>STATA 7.0 SE</th>
<th>SPSS 11.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>Ordinary least squares</td>
<td>REG</td>
<td>.regress</td>
</tr>
<tr>
<td>Binary</td>
<td>Binary logit</td>
<td>PROBIT, LOGISTIC, GENMOD, CATMOD</td>
<td>.logit; logistic ****</td>
</tr>
<tr>
<td></td>
<td>Binary probit</td>
<td>PROBIT, LOGISTIC, GENMOD</td>
<td>.probit</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Ordinal logit</td>
<td>PROBIT, LOGISTIC</td>
<td>.ologit *****</td>
</tr>
<tr>
<td></td>
<td>Generalized logit</td>
<td>-</td>
<td>.gologit</td>
</tr>
<tr>
<td></td>
<td>Ordinal probit</td>
<td>PROBIT, LOGISTIC</td>
<td>.oprobit</td>
</tr>
<tr>
<td>Nominal</td>
<td>Multinomial logit</td>
<td>CATMOD</td>
<td>.mlogit</td>
</tr>
<tr>
<td></td>
<td>Conditional logit</td>
<td>MDC***, (PHREG)</td>
<td>.clogit</td>
</tr>
<tr>
<td></td>
<td>Multinomial probit*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Count</td>
<td>Poisson</td>
<td>GENMOD</td>
<td>.poisson</td>
</tr>
<tr>
<td></td>
<td>Negative Binomial</td>
<td>GENMOD</td>
<td>.nbreg</td>
</tr>
<tr>
<td></td>
<td>Zero-Inflated Poisson</td>
<td>-</td>
<td>.zip</td>
</tr>
<tr>
<td></td>
<td>Zero-inflated NB**</td>
<td>-</td>
<td>.znb</td>
</tr>
</tbody>
</table>

* The multinomial probit model is rarely used due to the estimation problem.
** Zero-inflated negative binomial regression model.
*** The MDC procedure is available in SAS 8.0 and later version.
**** The “logistic” produces estimators in terms of the odd ratio instead of logit
***** An add-on command written by Fu (1998)
****** The plum and nomreg commands are not available in SPSS 7.5 or older version.
I. Binary Logit Regression Model

Suppose we want to know whether budget (dollars), age, and gender (1 for male) cause car ownership. The “owncar” is coded 1 when a respondent owns a car and 0 otherwise. The binary logit model will be \( \log \frac{P}{1-P} = \beta_0 + \beta_1 \text{budget} + \beta_2 \text{age} + \beta_3 \text{gender} \). In general, the binary logit model is represented as \( \Pr(\text{ob}(Y=1)) = \frac{\exp(\beta'x)}{1+\exp(\beta'x)} = \Lambda(\beta'x) \), where \( \Lambda \) indicates the logistic probability distribution function. The marginal effect of \( i \)th variable is computed as,

\[
\frac{\partial \Lambda(\beta'x)}{\partial x_i} = \frac{e^{\beta'x}}{(1+e^{\beta'x})^2} \Lambda(\beta'x)[1-\Lambda(\beta'x)] \beta_i
\]

1. Binary Logit Model in STATA

STATA provides two commands for the binary logit model: .logit and .logistic. The .logit presents the results (coefficients) in terms of logit, while the .logistic produces coefficients with respect to the odd ratio. Although they are equivalent, .logit is more commonly used than .logistic. The both commands are followed by a dependent variable, a set of independent variables, and a series of options after a comma.

```
.logistic owncar budget age gender
```

Logit estimates                                   Number of obs =       1000
LR chi2(3)      =      43.07
Prob > chi2     =     0.0000
Log likelihood = -567.60271
Pseudo R2       =     0.0366

|                | Odds Ratio | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|----------------|------------|-----------|------|------|----------------------|
| owncar         |            |           |      |      |                      |
| budget         | 1.001857   | 0.0003946 | 4.71 | 0.000| 1.001084 1.002631    |
| age            | 1.23444    | 0.0613009 | 4.24 | 0.000| 1.119954 1.360629    |
| gender         | 1.007803   | 0.1460882 | 0.05 | 0.957| 0.7585566 1.338947   |

```

```
.logit owncar budget age gender
```

Iteration 0:  log likelihood = -589.13567
Iteration 1:  log likelihood = -568.08472
Iteration 2:  log likelihood = -567.60345
Iteration 3:  log likelihood = -567.60271

Logit estimates                                   Number of obs =       1000
LR chi2(3)      =      43.07
Prob > chi2     =     0.0000
Log likelihood = -567.60271
Pseudo R2       =     0.0366

http://www.indiana.edu/~statmath
Categorical DV Models

### Owncar Coefficients

| Coef.     | Std. Err. | z   | P>|z|  | [95% Conf. Interval] |
|-----------|-----------|-----|------|----------------------|
| budget    | 0.0018557 | 0.0003939 | 4.71 | 0.000 | 0.0010837 - 0.0026277 |
| age       | 0.2106171 | 0.0496589 | 4.24 | 0.000 | 0.1132875 - 0.3079467 |
| gender    | 0.0077728 | 0.1449571 | 0.05 | 0.957 | -0.2763379 - 0.2918836 |
| _cons     | -4.567904 | 1.064209 | -4.29 | 0.000 | -6.653715 - -2.482093 |

---

Note that a coefficient of the `.logit` is equivalent to the corresponding estimator of the `.logistic` when it is logarithmic transformed. For example, `.0018557 = \log(1.001857)` and so on. STATA allows researchers to compute marginal effects and discrete effects in a simple manner. The following is an example of the `.prchange` command, which is included in J. Scott Long’s SPost module (http://www.indiana.edu/~jsl650/).

```
. prchange

logit: Changes in Predicted Probabilities for owncar

<table>
<thead>
<tr>
<th></th>
<th>min-&gt;max</th>
<th>0-&gt;1</th>
<th>-+1/2</th>
<th>-+sd/2</th>
<th>MargEfct</th>
</tr>
</thead>
<tbody>
<tr>
<td>budget</td>
<td>0.2796</td>
<td>0.0005</td>
<td>0.0004</td>
<td>0.0728</td>
<td>0.0004</td>
</tr>
<tr>
<td>age</td>
<td>0.3332</td>
<td>0.0076</td>
<td>0.0410</td>
<td>0.0653</td>
<td>0.0410</td>
</tr>
<tr>
<td>gender</td>
<td>0.0015</td>
<td>0.0015</td>
<td>0.0015</td>
<td>0.0008</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

Pr(y|x) 0.2648 0.7352

budget  age  gender
x=  650.126  20.789   0.54
sd(x)  =  201.844  1.59469  0.498647
```

### Binary Logit Model In SAS

SAS provides four different procedures: PROBIT, LOGISTIC, GENMOD, and CATMOD. The probit and logit models can be estimated in either the PROBIT or LOGISTIC procedure. The CATMOD procedure is designed to fit the logit model to the functions of categorical response variables, while the GENMOD provides the methods of analyzing generalized linear model.

```
PROC LOGISTIC DESCENDING DATA = binary.car;
MODEL owncar = budget age gender;
RUN;
```

Note that the DESCENDING option forces SAS to use a larger value (e.g., 1) in the dependent variable as success. Otherwise, the coefficients have opposite signs to those of STATA and SPSS.

```
PROC PROBIT DATA = binary.car;
CLASS owncar;
MODEL owncar = budget age gender / DIST=LOGISTIC;
RUN;
```

Unlike the LOGISTIC, the PROBIT does not have the DESCENDING option. It requires categorical variables to be explicitly specified in the CLASS statement. Note that the
/DIST=LOGISTIC option specifies the probability distribution to be used in maximum likelihood estimation.

The GENMOD procedure provides higher flexibility than other procedures. For instance, the procedure allows users to use categorical variables in the right-hand side without creating dummy variables (see the second example). The following two procedures are equivalent:

```plaintext
PROC GENMOD DATA = binary.car DESC;
MODEL owncar = budget age gender /DIST=BINOMIAL LINK=LOGIT;
RUN;

PROC GENMOD DATA = binary.car DESC;
CLASS gender;
MODEL owncar = budget age gender /DIST=BINOMIAL LINK=LOGIT;
RUN;
```

Note that the LINK=LOGIT option specifies the link function. Alternative way is to write explicitly the link function using the FWDLINK and INVLINK statements without the LINK=LOGIT option:

```plaintext
PROC GENMOD DATA = binary.car DESC;
FWDLINK link=LOG(_MEAN_/(_MEAN_-1));
INVLINK invlink=1/(1+EXP(-1*_XBETA_));
MODEL owncar = budget age gender /DIST=BINOMIAL;
RUN;
```

The following example uses the CATMOD procedure, which produces a little bit different estimators. Interval or ratio variables should be specified in the DIRECT statement. Note that the /NOPROFILE suppresses the display of the population profiles and the response profiles:

```plaintext
PROC CATMOD DATA = binary.car;
DIRECT budget age;
MODEL owncar = budget age gender /NOPROFILE;
RUN;
```

3. Binary Logit Model In SPSS

The following is an example of the binary logit in SPSS:

```plaintext
LOGISTIC REGRESSION VAR=owncar
/METHOD=ENTER budget age gender
/CRI TERIA P Y( .05) POUt( .10) ITERATE(20) CUT(.5).
```
II. Binary Probit Regression Model

The probit model is represented as $\Pr(Y = 1) = \int_{-\infty}^{x} \phi(t)dt = \Phi(\beta'x)$, where $\Phi$ indicates the cumulative standard normal distribution function. The marginal effect of $i$th variable is computed as $\frac{\partial E(y|x)}{\partial x_i} = \phi(\beta'x)\beta_i$

1. Binary Probit In STATA

STATA has the .probit command with the similar usage as .logit.

```
.probit owncar budget age gender
Iteration 0:   log likelihood = -589.13567
Iteration 1:   log likelihood = -567.89453
Iteration 2:   log likelihood = -567.71705
Iteration 3:   log likelihood = -567.71702

Probit estimates                                  Number of obs   =       1000
LR chi2(3)      =      42.84
Prob > chi2     =     0.0000
Log likelihood = -567.71702                       Pseudo R2       =     0.0364

------------------------------------------------------------------------------
  owncar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    budget |    .001092   .0002277     4.79   0.000     .0006456    .0015384
     age |   .1214482   .0284631     4.27   0.000     .0656614    .1772349
   gender |   .0034862   .0862954     0.04   0.968    -.1656498    .1726221
    _cons |  -2.614903   .6110651    -4.28   0.000    -3.812568   -1.417237
------------------------------------------------------------------------------
```

2. Binary Probit In SAS

You can use the PROBIT, the LOGISTIC, or GENMOD procedure to estimate the binary probit model. Keep in mind that the coefficients of the PROBIT has opposite signs.

```
PROC PROBIT DATA = binary.car;
CLASS owncar;
MODEL owncar = budget age gender;
RUN;

PROC LOGISTIC DATA = binary.car DESC;
MODEL owncar = budget age gender /LINK=NORMIT;
RUN;

PROC GENMOD DATA = binary.car DESC;
MODEL owncar = budget age gender /DI=BLNOMAL LI=PROBIT T;
RUN;
```
Note that `/LINK=NORMIT` or `/LINK=PROBIT` in the PROC LOGISTIC indicate the standard normal probability distribution, while the `/LINK=PROBIT` in the PROC GENMOD specifies PROBIT as the link function.

### 3. Binary Probit in SPSS

The following is an example of the binary probit in SPSS. Note that the variable n with constant 1 is created to be used in the probit command.

```plaintext
COMPUTE n=1.
PROBIT owncar OF n WITH budget age sex
/LOG NONE /MODEL PROBIT /PRINT FREQ /CRITERIA ITERATE(20) STEPLIMIT(.1).
```
III. Ordinal Response Regression Model

Suppose we want to know whether budget (dollars), age, and gender (1 for male) cause degree of illegal parking. The “park” is scaled as “none (0),” “Sometimes (1),” and “Often (2).” The general form of the model is \( \Pr(y = m | x) = F(\tau_m - x\beta) - F(\tau_{m-1} - x\beta) \), where \( \tau \) (tau) is a cutpoint.

1. Ordinal Response Models in STATA

STATA has the .ologit and .oprobit command to conduct the ordinal logit and probit model, respectively. In addition, the .gologit, an add-on command, is developed to estimate the generalized ordered logit model, which is appropriate when the parallel regression assumption is violated.

```
. ologit park budget age gender
```

```
Iteration 0:  log likelihood = -246.46009
Iteration 1:  log likelihood = -218.15141
Iteration 2:  log likelihood = -211.25248
Iteration 3:  log likelihood = -210.67863
Iteration 4:  log likelihood = -210.67135
Iteration 5:  log likelihood = -210.67135

Ordered logit estimates
Number of obs = 1000
LR chi2(3) = 71.58
Prob > chi2 = 0.0000
Log likelihood = -210.67135
Pseudo R2 = 0.1452

|       | Coef.     | Std. Err. |      z  | P>|z|    |   [95% Conf. Interval] |
|-------|-----------|-----------|--------|-------|------------------------|
| park  |          |           |        |       |                        |
| budget| -.0008171 | .0007604  | -1.07  | 0.283 | [.0023073, .006732]    |
| age   | -.8669616 | .1310896  | -6.61  | 0.000 | [-1.123893, -0.6100307]|
| gender| -.9708374 | .2972005  | -3.27  | 0.001 | [-1.55334, -0.3883352] |
| _cut1 | -15.55553 | 2.623819  |       |       |                        |
| _cut2 | -13.56126 | 2.631907  |       |       | (Ancillary parameters) |
```

```
. oprobit park budget age gender
```

```
Iteration 0:  log likelihood = -246.46009
Iteration 1:  log likelihood = -213.74164
Iteration 2:  log likelihood = -208.77727
Iteration 3:  log likelihood = -208.45573
Iteration 4:  log likelihood = -208.45391

Ordered probit estimates
Number of obs = 1000
LR chi2(3) = 76.01
Prob > chi2 = 0.0000

```

Log likelihood = -208.45391                       Pseudo R2       =     0.1542

| park  | Coef.  | Std. Err. | z     | P>|z|   | 95% Conf. Interval |
|-------|--------|-----------|-------|-------|-------------------|
| budget| -0.00374 | 0.003671  | -1.02 | 0.308 | -0.0010936 - 0.0003456 |
| age   | -0.44545 | 0.0661963 | -6.73 | 0.000 | -0.575169 - -0.3157121 |
| gender| -0.46943 | 0.143538  | -3.27 | 0.001 | -0.750754 - -0.1880962 |

_ cut1 | -7.836069    1.33797          (Ancillary parameters)  
_ cut2 | -6.900104   1.331292


```
. gologit park budget age gender
Iteration 0:  Log Likelihood = -246.46009
Iteration 1:  Log Likelihood = -223.07329
Iteration 2:  Log Likelihood = -209.7263
Iteration 3:  Log Likelihood = -208.55942
Iteration 4:  Log Likelihood = -208.55656
Iteration 5:  Log Likelihood = -208.55656

Generalized Ordered Logit Estimates
Number of obs    =    1000
Model      chi2(6)    =   75.81
Prob > chi2      =  0.0000
Log Likelihood =   -208.5565590                     Pseudo R2        =  0.1538

| park  | Coef.  | Std. Err. | z     | P>|z|   | 95% Conf. Interval |
|-------|--------|-----------|-------|-------|-------------------|
| mleq1 | budget| -0.008042 | 0.007607 | -1.06 | 0.290 | -0.0022952 - 0.006868 |
|       | age   | -0.8566991 | 0.130381 | -6.57 | 0.000 | -1.112241 - -0.601157 |
|       | gender| -0.9497461 | 0.2971743 | -3.20 | 0.001 | -1.532197 - -0.3672952 |
|       | _cons | 15.33158   2.608365 | 5.88 | 0.000 | 10.21928 - 20.44388 |

| mleq2 | budget| -0.00678  | 0.018914  | -0.36 | 0.720 | -0.0043852 - 0.0030291 |
|       | age   | -1.451909 | 0.4340185 | -3.35 | 0.001 | -2.302569 - -0.6012479 |
|       | gender| -2.110354 | 1.068173 | -1.98 | 0.048 | -4.203935 - -0.0167732 |
|       | _cons | 24.91948   8.221891 | 3.03 | 0.002 | 8.804871 - 41.03409 |
```

2. Ordinal Response Models in SAS
SAS provides the LOGISTIC and PROBIT procedures to conduct the ordinal response regression model. In the LOGISTIC, the signs of intercepts are opposite to corresponding cut points in STATA when the DESCENDING option is used. The usages of the LOGISTIC and PROBIT procedures remain unchanged in the ordinal response models. The following two procedures are for the ordinal logit model.
Keep in mind that the signs of coefficients are opposite in the PROBIT procedure. The following two procedures are for the ordinal probit model.

```plaintext
PROC LOGISTIC DATA = ordinal.park DESC;
MODEL park = budget age gender /LINK=NORMIT;
RUN;

PROC PROBIT DATA = ordinal.park;
CLASS park;
MODEL park = budget age gender /DIST=NORMAL;
RUN;
```

3. Ordinal Response Models in SPSS

The followings are the examples of ordinal logit and probit models in SPSS.

```plaintext
PLUM park WITH budget age gender
/CRITERIA = CI (.95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5)
PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
/LINK = LOGIT /PRINT = FIT PARAMETER SUMMARY .

PLUM park WITH budget age gender
/CRITERIA = CI(.95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5)
PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
/LINK = PROBIT /PRINT = FIT PARAMETER SUMMARY .
```
IV. Nominal Response Regression Model

Suppose we want to know how budget, age, and gender affect the modes of transportation (walk, bike, bus, and car). The multinomial logit and conditional logit models are commonly used; the multinomial probit model is rarely used mainly due to the practical reason of estimation difficulties.

In the multinomial logit model, the independent variables contain characteristics of individuals, while they are the attributes of the choices in the conditional logit model. In other words, the conditional logit estimates how alternative-specific, not individual-specific, variables affect the likelihood of observing a given outcome (Long 2001). Therefore, data need to be appropriately arranged to conduct two models. The multinomial logit model is described as

\[ \text{Prob}(y = m | x_i) = \frac{e^{x_i' \beta_m}}{\sum_{j=1}^{J} e^{x_i' \beta_j}}, \text{ while the conditional logit is modeled as} \]

\[ \text{Prob}(y = m | z_i) = \frac{\exp(z_i' \gamma_m)}{\sum_{j=1}^{J} \exp(z_i' \gamma_j)}, \text{ where m ranges from 1 to J.} \]

1. Multinomial Logit Model in STATA

STATA has the .mlogit for the multinomial logit model. In the following example, the “base(3)” option indicates the value to be used as the base of the estimation. The default is zero, “walk” in this example.

.mlogit mode budget age gender, base(3)

Iteration 0: log likelihood = -1308.7916
Iteration 1: log likelihood = -1171.4576
Iteration 2: log likelihood = -1162.1648
Iteration 3: log likelihood = -1162.0211
Iteration 4: log likelihood =  -1162.021

Multinomial regression                            Number of obs   =       1000
LR chi2(9)      =     293.54
Prob > chi2     =     0.0000
Log likelihood =  -1162.021                       Pseudo R2       =     0.1121

------------------------------------------------------------------------------
mode |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0            |
  budget | -2.4153   .005906    -4.09   0.000    -.0035729   -.0012577
  age | -3.1958   .055531    -0.58   0.565     -.1397967   .0768804
  gender | -2.4364   .174419    -1.40   0.162     -.5854967   .0982145
  _cons |  1.5059   1.211733     1.24   0.214     -.8690574    3.880848
-------------+----------------------------------------------------------------
1            |
  budget | -2.0318   .005880     0.54   0.588    -.0083360   .0014715

http://www.indiana.edu/~statmath
(Outcome mode==3 is the comparison group)

2. **Multinomial Logit Model in SAS**

SAS has the CATMOD procedures for the multinomial logit. In the CATMOD procedure, the RESPONSE statement is used to specify the functions of response probabilities.

```
PROC CATMOD DATA = nominal.trans;
DIRECT budget age gender;
RESPONSE LOGITS;
MODEL mode = budget age gender /NOPROFILE;
RUN;
```

Note that “gender” is also specified in the DIRECT statement. Keep in mind that the above STATA example used the largest value of the dependent variable, 3 (“car”), as a baseline. Otherwise, the results must be different.

3. **Multinomial Logit Model in SPSS**

SPSS has the nomreg command to estimate the multinomial logit model. The following is the example of the nomreg command.

```
NOMREG mode WITH budget age gender
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001) SINGULAR(0.00000001)
/MODEL /INTERCEPT INCLUDE /PRINT PARAMETER SUMMARY LRT .
```

4. **Conditional Logit Model in STATA**

Suppose the choice of mode of transportation is affected by time (minutes) and cost (dollars) spent on the four alternatives of walk, bike, bus, and car. There are 210 subjects, each of whom has four choices; thus total 850 cases are included in the data set. The dependent variable is set 1 if the subject chooses the mode of transportation and zero otherwise.

STATA has the .clogit command to estimate the condition logit model. Following is an example of the model. Note that walk, bike, and bus are dummy variables for flagging the mode of transportation. For example, the “bus” is set 1 if an observation contains information about taking a bus. The group() option specifies the variable that identify unique subjects.

```
.clogit choice walk bike bus time cost, group(subject)
Iteration 0:  log likelihood = -262.77756
```
Iteration 1:   log likelihood = -209.50653
Iteration 2:   log likelihood = -205.72682
Iteration 3:   log likelihood = -205.5967
Iteration 4:   log likelihood = -205.59646

Conditional (fixed-effects) logistic regression   Number of obs   =        840
LR chi2(5)      =     171.05
Prob > chi2     =     0.0000
Log likelihood = -205.59646                       Pseudo R2       =     0.2938

------------------------------------------------------------------------------
choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
walk |    6.54371   .7815754     8.37   0.000     5.011851     8.07557
bike |   3.808292   .4600735     8.28   0.000     2.906564    4.710019
bus |   3.202466   .4537241     7.06   0.000     2.313183    4.091749
time |  -.1000569   .0104929    -9.54   0.000    -.1206226   -.0794912
cost |  -.0097447   .0062539    -1.56   0.119    -.0220022    .0025128
------------------------------------------------------------------------------

listcoef

clogit (N=840): Factor Change in Odds

Odds of: 1 vs 0

choice |      b         z     P>|z|    e^b
-------------+------------------------------------
walk |   6.54371    8.372   0.000 694.8600
bike |   3.80829    8.278   0.000  45.0734
bus |   3.20247    7.058   0.000  24.5931
time |  -0.10006   -9.536   0.000   0.9048
cost |  -0.00974   -1.558   0.119   0.9903

The .listcoef provides an easy way of interpret the coefficients by transforming the estimators into odds ratios. The e^b column, .9048 for example, is interpreted as follow. For one minute increase in the time of travel for a given mode of transportation, we can expect decrease in the odds of using that mode of transportation by 10 percent (a factor of .9048), holding other variables constant.

5. Conditional Logit Model in SAS
SAS has the MDC procedure available in the 8.0 or later version to estimate the conditional logit model. You may also use the PHREG procedure though, which is mainly designed to conduct the survival analysis. The PHREG is based on the Cox proportional hazards model, which is known to be similar to the conditional logit model.

PROC MDC DATA=clogit.travel;
MODEL choice = walk bike bus time cost /TYPE=CLOGIT NCHOICE=4;
ID subject;
RUN;

Note that the NCHOICE=4 option specifies the number of choices.

In order to use the PHREG procedure, a failure time variable needs to be created as f_time=1 - choice so that data arrangement is consistent with the survival analysis data. The STRATA statement specifies the variable for the subjects.

PROC PHREG DATA=clogit.travel 2;
STRATA subject;
MODEL f_time*choice(0)=walk bike bus time cost;
RUN;

Note that the PHREG presents the hazard ratio at the last column of the output, which is equivalent to the e^b column of the output in the .listcoef of STATA.

6. Conditional Logit Model in SPSS
Unlike SAS and STATA, SPSS does not have the right command for the conditional logit model. Instead, SPSS provides the coxreg command of the survival analysis as a backdoor way of estimating the conditional logit model. Compared to SAS and STATA, SPSS produces slightly different estimators and associated statistics.

COXREG f_time WITH walk bike bus time cost
/STATUS=choice(1)
/STRATA=subject.

Note that the Exp(B) column of the output is equivalent to the hazard ratio of the PHREG and the e^b column of the .listconf.
V. Count Data Model

Event count data models include Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial regression models. The example is how waste quota (“emps”) and strictness of policy implementation (“strict”) affect the frequency of waste spill accidents of plants (“accident”). Unlike SAS and STATA, SPSS does not provide any command for the count data models.

1. Poisson Regression Model

Poisson regression model, the simplest one, is described as 
\[ P(y_i \mid X_i \beta) = \frac{e^{(X_i \beta)}}{y_i!} \] . The following is a STATA example using the .poisson command.

```
. poisson accident emps strict
Iteration 0:   log likelihood = -1821.5112
Iteration 1:   log likelihood = -1821.5101
Iteration 2:   log likelihood = -1821.5101
Poisson regression                                Number of obs   =        778
LR chi2(2)      =     124.82
Prob > chi2     =     0.0000
Log likelihood = -1821.5101                       Pseudo R2       =     0.0331
------------------------------------------------------------------------------
accident |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
emps |   .0054186   .0007434     7.29   0.000     .0039615    .0068757
strict |  -.7041664   .0667619   -10.55   0.000    -.8350174   -.5733154
_cons |   .3900961   .0466787     8.36   0.000     .2986076    .4815846
------------------------------------------------------------------------------
```

The following GENMOD produces the same results as the above.

```
PROC GENMOD DATA = count.waste;
MODEL accident = emps strict / DIST=POISSON LINK=LOG;
RUN;
```

2. Negative Binomial Regression Model

Negative binomial regression model is able to cope with overdispersion problem. It is modeled as 
\[ \text{Pr}(y_i \mid x_i) = \frac{\Gamma(y_i + v_i)}{y_i! \Gamma(v_i)} \left( \frac{v_i}{v_i + \mu_i} \right)^{v_i} \left( \frac{\mu_i}{v_i + \mu_i} \right)^{y_i} \] where 1/v determines the degree of dispersion in the predictions. In STATA, the .nbreg command is used to estimate the model.

```
. nbreg accident emps strict
```

Fitting comparison Poisson model:
Iteration 0: log likelihood = -1821.5112
Iteration 1: log likelihood = -1821.5101
Iteration 2: log likelihood = -1821.5101

Fitting constant-only model:

Iteration 0: log likelihood = -1256.6761
Iteration 1: log likelihood = -1152.6155
Iteration 2: log likelihood = -1125.6643
Iteration 3: log likelihood = -1125.4183
Iteration 4: log likelihood = -1125.4183

Fitting full model:

Iteration 0: log likelihood = -1117.1731
Iteration 1: log likelihood = -1116.7201
Iteration 2: log likelihood = -1116.7182
Iteration 3: log likelihood = -1116.7182

Negative binomial regression

Number of obs = 778
LR chi2(2) = 17.40
Prob > chi2 = 0.0002
Log likelihood = -1116.7182
Pseudo R2 = 0.0077

| accident | Coef. | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|----------|-------|-----------|------|------|----------------------|
| emps     | 0.0051981 | 0.0022595 | 2.30 | 0.021 | 0.0007694 - 0.0096267 |
| strict   | -0.6702548 | 0.1671191 | -4.01 | 0.000 | -0.9978021 - 0.3427074 |
| _cons    | 0.3851111 | 0.1278468 | 3.01 | 0.003 | 0.134536 - 0.6356861 |

/lnalpha | 1.37509 | 0.0885176 | 1.201599 | 1.548582 |

alpha | 3.955434 | 0.3501257 | 3.32543 | 4.704793 |

Likelihood ratio test of alpha=0: chi-bar2(01) = 1409.58 Prob>=chi-bar2 = 0.000

Note the the last line shows the test of overdispersion with the null hypothesis of alpha=0. There is statistically significant evidence of overdispersion (p<.000), which indicates the negative binomial regression model is better than the Poisson regression model in the example.

The following GENMOD produces the equivalent results to the above.

```plaintext
PROC GENMOD DATA = count.waste;
MODEL accident = emps strict / DIST=NEGBIN LINK=LOG;
RUN;
```

3. Zero-Inflated Poisson Regression Model (ZIP)
Zero-inflated models are designed to figure out the overdispersion problem that has many zero count in the dependent variable. Zero-inflated Poisson regression model is described as

$$
\Pr(y_i | x_i) = \left(1 - \psi_i\right) e^{-x_i \beta} \left(\frac{x_i \beta}{y_i!}\right)^{y_i} .
$$

STATA has the .zip command to estimate zero-inflated Poisson model as following. Note that “inflate()” option specifies a variable whose zero count is assumed inflated. In SAS, there is no built-in procedure or option equivalent to the .zip.

```
.zip accident emps strict, inflate(accident)
```

**Fitting constant-only model:**

- Iteration 0: log likelihood = -1627.0779
- Iteration 1: log likelihood = -1373.1619
- Iteration 2: log likelihood = -1035.447
- Iteration 3: log likelihood = -859.09673
- Iteration 4: log likelihood = -803.64624
- Iteration 5: log likelihood = -794.98612
- Iteration 6: log likelihood = -792.98612
- Iteration 7: log likelihood = -792.60554
- Iteration 8: log likelihood = -792.04565
- Iteration 9: log likelihood = -792.49254
- Iteration 10: log likelihood = -792.49239
- Iteration 11: log likelihood = -792.49236
- Iteration 12: log likelihood = -792.49236
- Iteration 13: log likelihood = -792.49236

**Fitting full model:**

- Iteration 0: log likelihood = -792.49236
- Iteration 1: log likelihood = -791.38398
- Iteration 2: log likelihood = -791.3833
- Iteration 3: log likelihood = -791.3833

Zero-inflated poisson regression

| accident | Coef. | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|----------|-------|-----------|-------|-------|----------------------|
| emps     | 0.000203 | 0.0008072 | -0.25 | 0.801 | -0.0017853 to 0.0013789 |
| strict   | 0.08557 | 0.0688976 | -1.24 | 0.214 | -0.2206068 to 0.3917311 |
| _cons    | 1.381173 | 0.0470788 | 29.34 | 0.000 | 1.2889 to 1.473445 |

| inflate | Coef. | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|---------|-------|-----------|-------|-------|----------------------|
| accident | -49.01636 | 21848.63 | -0.00 | 0.999 | -42871.55 to 42773.52 |
| _cons   | 25.49139 | 15557.63 | 0.00 | 0.999 | -30466.91 to 30517.9 |

http://www.indiana.edu/~statmath
4. Zero-Inflated Negative Binomial Regression Model (ZINB)

Zero-inflated negative binomial regression model is also developed to deal with overdispersion problem. It is modeled as:

\[ \text{Pr}(y_i | x_i, G_i = 0) = \frac{\Gamma(y_i + v_i)}{y_i! \Gamma(v_i)} \left( \frac{\mu_i}{v_i + \mu_i} \right)^{y_i} \left( \frac{v_i}{v_i + \mu_i} \right)^{\mu_i} \]

In STATA, .zinb is the command for the zero-inflated negative binomial model. In SAS, however, there is no built-in procedure or option equivalent to the .zinb.

```
.zinb accident emps strict, inflate(accident)
```

Fitting constant-only model:

Iteration 0: log likelihood = -1190.5117  (not concave)
Iteration 1: log likelihood = -1018.1736  (not concave)
Iteration 2: log likelihood = -835.55703
Iteration 3: log likelihood = -722.09514
Iteration 4: log likelihood = -677.39366
Iteration 5: log likelihood = -667.06546
Iteration 6: log likelihood = -664.68626
Iteration 7: log likelihood = -664.22538
Iteration 8: log likelihood = -664.14999
Iteration 9: log likelihood = -664.13182
Iteration 10: log likelihood = -664.12783
Iteration 11: log likelihood = -664.12679
Iteration 12: log likelihood = -664.12677
Iteration 13: log likelihood = -664.12677

Fitting full model:

Iteration 0: log likelihood = -664.12677
Iteration 1: log likelihood = -663.69106
Iteration 2: log likelihood = -663.69038
Iteration 3: log likelihood = -663.69038

Zero-inflated negative binomial regression

|                  | Coef. | Std. Err. | z    | P>|z| [95% Conf. Interval] |
|------------------|-------|-----------|------|-----------------------|
| accident emps    | -0.0002947 | 0.0012568 | -0.23 | 0.815 [-0.0027579, 0.0021685] |
| strict           | -0.0847249 | 0.1072602 | -0.79 | 0.430 [-0.294951, 0.1255012] |
| _cons            | 1.385103  | 0.0783557 | 17.68 | 0.000 [1.231529, 1.538678] |
5. Comparison of Three Count Data Models

The following plot compares the negative binomial regression model, zero-inflated Poisson model, and zero-inflated negative binomial model. The negative binomial regression model shows best fit to the observed data of this example with overdispersion.

- Observed
- Predicted by NBRM
- Predicted by ZIP
- Predicted by ZINB
References


