HW2 Econometrics 3

Matthew Aaron Looney

10/08/2017

Problem 2

Consumer's Behavior: The accompanying data describes the weekly buying behavior of consumers in Pittsfield, Massachusetts and Eau Claire, Wisconsin for coffee across different supermarket chains in each city from June 18, 2001 to July 30, 2006. Your task is to analyze the effect of the factors that affect consumer's brand and chain choice. To start your analysis, you decide to pick only one city and one week (pick a week that gives you at least 200 observations). The data (Homescan1 data from Information Resource infoscan, IRIconsist of panelist id, panelist income (category), race, and household size; the week, the number of units purchased by the panelist, the dollar amount spent (this, is not the price. The price can be obtained using the dollar amount, the units, and the volume equivalent), universal product code (COLUPC), the city, the chain, the volume equivalent,2 and the brand. In what follows, we assume that brand j purchased in chain k is different than bran j purchased in chain l, i.e., there is differentiation at the brand and chain level. Perform the following analysis:

(1.a)

Estimate a multinomial logit model to explain the brand/chain choice using demographic variables. Choose a brand and estimate the effect of moving from income category .

Odds Ratio Estimates				
Effect	choice	Point Estimate	95% Wald Confidence Limits	
income	2	1.180	0.749	1.858
income	3	1.146	0.631	2.083
income	4	1.410	0.873	2.278
income	5	1.378	0.637	2.981
income	6	1.084	0.625	1.880
income	7	1.149	0.729	1.810
income	8	1.847	0.919	3.712
size	2	0.734	0.247	2.179
size	3	0.447	0.076	2.621
size	4	0.733	0.231	2.327
size	5	2.306	0.467	11.386
size	6	1.947	0.560	6.771
size	7	1.048	0.355	3.091
size	8	0.374	0.048	2.884
race	2	387.370	<0.001	>999.999
race	3	1.449	<0.001	>999.999
race	4	>999.999	<0.001	>999.999
race	5	1.299	<0.001	>999.999
race	6	1.016	<0.001	>999.999
race	7	360.533	<0.001	>999.999
race	8	2.330	<0.001	>999.999

Figure 1: SAS Output of Multinomial Logit

(1.b)

Estimate a conditional logit model to explain the brand/chain choice using the price and the brand and chain dummy variables as product characteristics. Choose a brand and estimate the effect of decreasing the price by $1.25.

The MDC Procedure
Conditional Logit Estimates

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
price	1	-1.57E-11	0.0553	-0.00	1.0000
d1	1	-1.0156	0.6644	-1.53	0.1264
d2	1	2.5009	0.1462	17.11	<.0001
d3	1	-0.7279	0.5217	-1.40	0.1630
d4	1	1.1439	0.2397	4.77	<.0001
d5	1	-2.1142	1.0070	-2.10	0.0358
d6	1	-0.1683	0.3924	-0.43	0.6680
d7	0	2.3967	.	.	.
d8	1	-1.0156	0.5911	-1.72	0.0858

Figure 2: SAS Output of Conditional Logit

(1.c)

Estimate a mixed logit model to explain the brand/chain choice using demographic variables and the price.

The MDC Procedure
Mixed Multinomial Logit Estimates

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
d1inc	1	-0.2551	0.4219	-0.60	0.5454
d2inc	1	-0.0517	0.0648	-0.80	0.4254
d3inc	1	-0.0588	0.4548	-0.13	0.8971
d3inc	0	-0.0588	.	.	.
d5inc	1	-0.1597	1.1642	-0.14	0.8909
d6inc	1	-0.1871	0.1778	-1.05	0.2925
d7inc	1	-0.1041	0.0695	-1.50	0.1344
d1size	1	0.6432	7.2157	0.09	0.9290
d2size	1	0.3839	6.9578	0.06	0.9560
d3size	1	-0.1696	7.8509	-0.02	0.9828
d4size	1	0.2571	6.9539	0.04	0.9705
d5size	1	1.3785	9.6996	0.14	0.8870
d6size	1	1.3976	6.9864	0.20	0.8414
d7size	1	0.7128	6.9429	0.10	0.9182
d1race	1	0.3459	13.9537	0.02	0.9802
d2race	1	3.0292	13.9261	0.22	0.8278
d3race	1	1.3840	14.4543	0.10	0.9237
d4race	1	1.7450	13.9026	0.13	0.9001
d5race	1	-3.5598	29.0015	-0.12	0.9023
d6race	1	-1.6421	14.1788	-0.12	0.9078
d7race	1	2.5561	13.8957	0.18	0.8541
price_cl	1	-0.006409	0.0356	-0.18	0.8573

Figure 3: SAS Output of Mixed Logit

(1.d)

Estimate a nested logit model with the upper level being the chain (use chain dummy variables as explanatory variables) and the lower level the brands (use the price as the explanatory variable). Test the validity of IIA.

The MDC Procedure
Nested Logit Estimates

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
price_L1G6	1	-0.6919	0.0833	-8.31	<.0001
price_L1G7	1	-0.5384	0.1208	-4.46	<.0001
price_L1G8	1	-0.5936	0.1297	-4.58	<.0001
c6_L2G1	1	4.6420	2.4842	1.87	0.0617
c7_L2G1	1	12.5756	6.0840	2.07	0.0387
INC_L2G1C6	1	3.6829	0.8792	4.19	<.0001
INC_L2G1C7	1	9.9739	2.4123	4.13	<.0001
INC_L2G1C8	1	3.4437	1.0820	3.18	0.0015

Test Results				
Test	Type	Statistic	Pr > ChiSq	Label
Test0	L.R.	51.26	<.0001	INC_L2G1C6 = 1 , INC_L2G1C7 = 1 , INC_L2G1C8 = 1

Figure 4: SAS Output of Nested Logit

- Sadly I am not able to get the IIA macro working in SAS so I am unable to test for IIA using the Hausman's specification test for IIA.

Problem 2

Problem 2. Censoring/Truncation. Greene (2007) analyzed the de- fault behavior and monthly behavior of a large sample of credit card users (13,444).

(2.1)

Estimate the following model

$$\log spend = \beta_1 + \beta_2 \ln income + \beta_3 Age + \beta_4 Adepcnt + \beta_5 ownrent + \varepsilon$$

Table 1: Summary Statistics: Problem 2

Statistic	N	Mean	St. Dev.	Min	Max
CARDHLDR	13,444	0.781	0.414	0	1
DEFAULT	13,444	0.074	0.262	0	1
AGE	13,444	33.472	10.226	0.000	88.667
ACADMOS	13,444	55.319	63.090	0	576
ADEPCNT	13,444	1.017	1.279	0	9
MAJORDRG	13,444	0.463	1.433	0	22
MINORDRG	13,444	0.291	0.768	0	11
OWNRENT	13,444	0.456	0.498	0	1
INCOME	13,444	2,509.528	1,252.947	50.000	8,333.250
SELFEMPL	13,444	0.058	0.234	0	1
INCPER	13,444	21,719.680	13,591.210	362.500	150,000.000
EXP_INC	13,444	0.071	0.104	0.0001	2.038
SPENDING	10,499	226.983	294.101	0.111	4,810.309
LOGSPEND	10,499	4.729	1.405	−2.197	8.479
Ln_income	13,444	7.725	0.450	3.912	9.028

(2.1.a)

Using OLS. What is the effect of 10% increase in income on credit card expenditure?

- Since we are dealing with log-log we can simply multiply the paramater estimate on income by ten, which gives 11.2120776. So a 10 percent increase in income is estimated to increase credit card spending by 11.2120776 percent.

(2.1.b)

Using Censored regression. What is the effect of 10% increase in income on credit card expenditure?

We will need to employ a Censored (Tobit) Regression and calculate the Partial Effects.

The general formulation for the Tobit Model (Greene 7th. ed., pg 848):

$$y_i{}^* = x_i{}'\beta + \varepsilon_i$$

$$\begin{array}{r} y_i = 0 \\ y_i = y_i{}^* \end{array} \begin{cases} \text{if } y_i{}^* \leqslant 0 \\ \text{if } y_i{}^* \geqslant 0 \end{cases}$$

Table 2: Regression output used to answer Problem 2.1

	Dependent variable:		
	LOGSPEND		NA
	OLS	*censored regression*	*Heckman selection*
	(1)	(2)	(3)
Ln_income	1.121^{***}	1.117^{***}	0.907^{***}
	(0.033)	(0.033)	(0.162)
AGE	-0.015^{***}	-0.014^{***}	-0.014^{***}
	(0.001)	(0.001)	(0.002)
ADEPCNT	-0.027^{**}	-0.027^{**}	0.016
	(0.011)	(0.011)	(0.034)
OWNRENT	-0.203^{***}	-0.201^{***}	-0.281^{***}
	(0.030)	(0.030)	(0.065)
logSigma		0.296^{***}	
		(0.007)	
Constant	-3.363^{***}	-3.340^{***}	-1.419
	(0.243)	(0.246)	(1.458)
Observations	10,499	10,499	13,444
R^2	0.105		0.105
Adjusted R^2	0.104		0.104
Log Likelihood		$-18{,}012.210$	
Akaike Inf. Crit.		36,036.430	
Bayesian Inf. Crit.		36,079.980	
ρ			-0.608
Inverse Mills Ratio			-0.878 (0.646)
Residual Std. Error	1.330 (df = 10494)		
F Statistic	306.358^{***} (df = 4; 10494)		
Note:	*p<0.1; **p<0.05; ***p<0.01		

The censored regression model is a generalisation of the standard Tobit model. The dependent variable can be either left-censored, right-censored, or both left-censored and right-censored, where the lower and/or upper limit of the dependent variable can be any number:

$$y_i^* = x_i'\beta + \varepsilon_i \tag{1}$$
$$y_i = \begin{cases} a & \text{if } y_i^* \leq a \\ y_i^* & \text{if } a < y_i^* < b \\ b & \text{if } y_i^* \geq b \end{cases} \tag{2}$$

Here a is the lower limit and b is the upper limit of the dependent variable. If $a = -\infty$ or $b = \infty$, the dependent variable is not left-censored or right-censored, respectively.

Censored regression models (including the standard Tobit model) are usually estimated by the Maximum Likelihood (ML) method. Assuming that the disturbance term ε follows a normal distribution with mean 0 and variance σ^2, the log-likelihood function is

$$\begin{aligned} \log L = \sum_{i=1}^{N} \Bigg[& I_i^a \log \Phi\left(\frac{a - x_i'\beta}{\sigma}\right) + I_i^b \log \Phi\left(\frac{x_i'\beta - b}{\sigma}\right) \\ & + \left(1 - I_i^a - I_i^b\right)\left(\log \phi\left(\frac{y_i - x_i'\beta}{\sigma}\right) - \log \sigma\right)\Bigg], \end{aligned} \tag{3}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and the cumulative distribution function, respectively, of the standard normal distribution, and I_i^a and I_i^b are indicator functions with

$$I_i^a = \begin{cases} 1 & \text{if } y_i = a \\ 0 & \text{if } y_i > a \end{cases} \tag{4}$$
$$I_i^b = \begin{cases} 1 & \text{if } y_i = b \\ 0 & \text{if } y_i < b \end{cases} \tag{5}$$

The log-likelihood function of the censored regression model~(3) can be maximised with respect to the parameter vector $(\beta', \sigma)'$ using standard non-linear optimisation algorithms.

The proper Marginal (Partial) Effects formula:

$$\frac{\partial E[y|x]}{\partial x} = \beta \Pr ob[a < y^* < b]$$

The marginal effects of an explanatory variable on the expected value of the dependent variable is (Greene 7th. ed., pg 849):

$$ME_j = \frac{\partial E[y|x]}{\partial x_j} = \beta_j \left[\Phi\left(\frac{b - x'\beta}{\sigma}\right) - \Phi\left(\frac{a - x'\beta}{\sigma}\right)\right] \tag{6}$$

In order to compute the approximate variance covariance matrix of these marginal effects using the Delta method, we need to obtain the Jacobian matrix of these marginal effects with respect to all estimated parameters (including~σ):

$$\frac{\partial ME_j}{\partial \beta_k} = \Delta_{jk} \left[\Phi\left(\frac{b - x'\beta}{\sigma}\right) - \Phi\left(\frac{a - x'\beta}{\sigma}\right) \right] - \frac{\beta_j\ x_k}{\sigma}\ \left[\phi\left(\frac{b - x'\beta}{\sigma}\right) - \phi\left(\frac{a - x'\beta}{\sigma}\right) \right] \tag{7}$$

and

$$\frac{\partial ME_j}{\partial \sigma} = -\beta_j\ \left[\phi\left(\frac{b - x'\beta}{\sigma}\right)\ \frac{b - x'\beta}{\sigma^2} - \phi\left(\frac{a - x'\beta}{\sigma}\right)\ \frac{a - x'\beta}{\sigma^2} \right], \tag{8}$$

where Δ_{jk} is "Kronecker's Delta"

with $\Delta_{jk} = 1$ for $j = k$ and $\Delta_{jk} = 0$ for $j \neq k$. If the upper limit of the censored dependent variable (b) is infinity or the lower limit of the censored dependent variable (a) is minus infinity, the terms in the square brackets in equation~(8) that include b or a, respectively, have to be removed.

- Where I compute the partial effect at each observation and then compute the mean.
  The parginal effect of Ln_income on LOGSPEND is 1.1169911. Therefore, a 10 percent increase of income is estiimated to increase credit card spending by 11.169911.

(2.1.c)

Using Heckman Two-Step Estimator. What the is effect of 10% increase in income on credit card expenditure?

Heckman's standard sample selection model is also called "Tobit-2" model (Amemiya 1984, Amemiya 1985). It consists of the following (unobserved) structural process:

$$y_i^{S*} = \vec{\beta}^{S\prime} \vec{x}_i^S + \varepsilon_i^S \tag{9}$$
$$y_i^{O*} = \vec{\beta}^{O\prime} \vec{x}_i^O + \varepsilon_i^O, \tag{10}$$

where y_i^{S*} is the realisation of the the latent value of the selection "tendency" for the individual i, and y_i^{O*} is the latent outcome. $\vec{x}_i^S$ and $\vec{x}_i^O$ are explanatory variables for the selection and outcome equation, respectively. $\vec{x}^S$ and $\vec{x}^O$ may or may not be equal. We observe

$$y_i^S = \begin{cases} 0 & \text{if } y_i^{S*} < 0 \\ 1 & \text{otherwise} \end{cases} \tag{11}$$
$$y_i^O = \begin{cases} 0 & \text{if } y_i^S = 0 \\ y_i^{O*} & \text{otherwise,} \end{cases} \tag{12}$$

i.e. we observe the outcome only if the latent selection variable y^{S*} is positive. The observed dependence between y^O and x^O can now be written as

$$E[y^O | \vec{x}^O = \vec{x}_i^O, \vec{x}^S = \vec{x}_i^S, y^S = 1] = \vec{\beta}^{O\prime} \vec{x}_i^O + E[\varepsilon^O | \varepsilon^S \geq -\vec{\beta}^{S\prime} \vec{x}_i^S]. \tag{13}$$

Estimating the model above by OLS gives in general biased results, as $E[\varepsilon^O | \varepsilon^S \geq -\vec{\beta}^{S\prime} \vec{x}_i^S] \neq 0$, unless ε^O and ε^S are mean independent (in this case $\varrho = 0$).

Assuming the error terms follow a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon^S \\ \varepsilon^O \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \varrho \\ \varrho & \sigma^2 \end{pmatrix} \right), \tag{14}$$

we may employ the following simple strategy: find the expectations $E[\varepsilon^O|\varepsilon^S \geq -\vec{\beta}^{S\prime}\vec{x}_i^S]$, also called the *control function*, by estimating the selection equations and by probit, and thereafter insert these expectations as additional covariates (see Greene 2002 for details). Accordingly, we may write:

$$y_i^O = \vec{\beta}^{O\prime}\vec{x}_i^O + E[\varepsilon^O|\varepsilon^S \geq -\vec{\beta}^{S\prime}\vec{x}_i^S] + \eta_i \equiv \vec{\beta}^{O\prime}\vec{x}_i^O + \varrho\sigma\lambda(\vec{\beta}^{S\prime}\vec{x}_i^S) + \eta_i \tag{15}$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is commonly referred to as inverse Mill's ratio, $\phi(\cdot)$ and $\Phi(\cdot)$ are standard normal density and cumulative distribution functions and η is a new disturbance term, independent of $\vec{x}^O$ and $\vec{x}^S$. The unknown multiplicator $\varrho\sigma$ can be estimated by OLS ($\hat{\beta}^\lambda$). Essentially, we describe the selection problem as an omitted variable problem, with $\lambda(\cdot)$ as the omitted variable. Since the true $\lambda(\cdot)$s are generally unknown, they are replaced by estimated values based on the probit estimation in the first step.

The relations also reveal the interpretation of ϱ. If $\varrho > 0$, the third term in the right hand side is positive as the observable observations tend to have above average realizations of ε^O. This is usually referred to as positive selection in a sense that the observed outcomes are better than the average. In this case, the OLS estimates are upward biased.

An estimator of the variance of ε^O can be obtained by

$$\hat{\sigma}^2 = \frac{\hat{\vec{\eta}}'\hat{\vec{\eta}}}{n^O} + \frac{\sum_i \hat{\delta}_i}{n^O}\,\hat{\beta}^{\lambda\,2} \tag{16}$$

where $\hat{\vec{\eta}}$ is the vector of residuals from the OLS estimation, n^O is the number of observations in this estimation, and $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i + \hat{\vec{\beta}}^{S\prime}\vec{x}_i^S)$. Finally, an estimator of the correlation between ε^S and ε^O can be obtained by $\hat{\varrho} = \hat{\beta}^\lambda/\hat{\sigma}$. Note that $\hat{\varrho}$ can be outside of the $[-1, 1]$ interval.

Since the estimation is not based on the true but on estimated values of $\lambda(\cdot)$, the standard OLS formula for the coefficient variance-covariance matrix is not appropriate [p.~157]{heckman79}. A consistent estimate of the variance-covariance matrix can be obtained by

$$\widehat{VAR}\left[\hat{\vec{\beta}}^O, \hat{\vec{\beta}}^\lambda\right] = \hat{\sigma}^2\left[\mathbf{X}_\lambda^{\mathbf{O}\prime}\mathbf{X}_\lambda^{\mathbf{O}}\right]^{-1}\left[\mathbf{X}_\lambda^{\mathbf{O}\prime}\left(\mathbf{I} - \hat{\varrho}^{\mathbf{2}}\hat{\mathbf{\Delta}}\right)\mathbf{X}_\lambda^{\mathbf{O}} + \mathbf{Q}\right]\left[\mathbf{X}_\lambda^{\mathbf{O}\prime}\mathbf{X}_\lambda^{\mathbf{O}}\right]^{-1} \tag{17}$$

where

$$\mathbf{Q} = \hat{\varrho}^{\mathbf{2}}\left(\mathbf{X}_\lambda^{\mathbf{O}\prime}\hat{\mathbf{\Delta}}\mathbf{X}^{\mathbf{S}}\right)\widehat{\mathbf{VAR}}\left[\hat{\vec{\beta}}^{\mathbf{S}}\right]\left(\mathbf{X}^{\mathbf{S}\prime}\hat{\mathbf{\Delta}}\mathbf{X}_\lambda^{\mathbf{O}}\right), \tag{18}$$

$\mathbf{X^S}$ is the matrix of all observations of $\vec{x}^S$, $\mathbf{X}_\lambda^{\mathbf{O}}$ is the matrix of all observations of $\vec{x}^O$ and $\hat{\lambda}$, $\mathbf{I}$ is an identity matrix, $\hat{\mathbf{\Delta}}$ is a diagonal matrix with all $\hat{\delta}_i$ on its diagonal, and $\widehat{VAR}\left[\hat{\vec{\beta}}^S\right]$ is the estimated variance covariance matrix of the probit estimate (Greene 1981, Greene 2002).

This is the original idea by (Heckman 1976). As the model is fully parametric, it is straightforward to construct a more efficient maximum likelihood (ML) estimator. Using the properties of a bivariate normal distribution, it is easy to show that the log-likelihood can be written as

$$L = \sum_{\{i:y_i^S=0\}} \log\Phi(-\vec{\beta}^{S\prime}\vec{x}_i^S) + \tag{19}$$

$$+ \sum_{\{i:y_i^S=1\}}\left[\log\Phi\left(\frac{\vec{\beta}^{S\prime}\vec{x}_i^S + \dfrac{\varrho}{\sigma}(y_i^O - \vec{\beta}^{O\prime}\vec{x}_i^O)}{\sqrt{1-\varrho^2}}\right) - \frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2}\frac{(y_i^O - \vec{\beta}^{O\prime}\vec{x}_i^O)^2}{\sigma^2}\right]. \tag{20}$$

The original article suggests using the two-step solution for exploratory work and as initial values for ML estimation. This was a result of the high costs of estimation. Nowadays, costs are no longer an issue, however, the two-step solution allows certain generalisations more easily than ML, and is more robust in certain circumstances.

This model and its derivations were introduced in the 1970s and 1980s. The model is well identified if the exclusion restriction is fulfilled, i.e. if $\vec{x}^S$ includes a component with a substantial explanatory power but which is not present in $\vec{x}^O$. This means essentially that we have a valid instrument. If this is not the case, the identification is related to the non-linearity of the inverse Mill's ratio $\lambda(\cdot)$. The exact form of it stems from the distributional assumptions. During the recent decades, various semiparametric estimation techniques have been increasingly used in addition to the Heckman model.

- Having run the Heckman two-step estimation procedure and calculated the marginal effect of income on credi card spending we see that a 10 percent increase in income is estimated to increase credit card spending by 11.240879 percent.

(2.2)

Create a subsample where only credit cardholders appear and do the following

(2.2.a)

Estimate the above model using OLS. What is the difference in credit card spending between home owner and renter?

Table 3: Regression output used to answer Problem 2.2.a

	Dependent variable:
	LOGSPEND
Ln_income	1.121^{***}
	(0.033)
AGE	-0.015^{***}
	(0.001)
ADEPCNT	-0.027^{**}
	(0.011)
OWNRENT	-0.203^{***}
	(0.030)
Constant	-3.363^{***}
	(0.243)
Observations	10,499
R^2	0.105
Adjusted R^2	0.104
Residual Std. Error	1.330 (df = 10494)
F Statistic	306.358^{***} (df = 4; 10494)
Note:	*p<0.1; **p<0.05; ***p<0.01

- When an individual moves from renting to owning a house we estimate a decrease in credit card spending by 18.37218 percent.

(2.2.b)

Estimate the above model using truncated regression. What is the difference in credit card spending between home owner and renter?

Following Greene (Greene 7th. ed., pg 833–839) and Davidson and MacKinnon (1993, 534–537) provide introductions to the truncated regression model.

Let $y = \mathbf{x}\beta + \varepsilon$ the model. y represents continuous outcomes either observed or not observed. Our model assumes that ε $N(o, \sigma^2 I))$.

Let a be the lower limit and b be the upper limit. The log likelihood is

$$\ln L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - x_j\beta)^2 - \sum_{j=1}^{n}\log\left\{\Phi\left(\frac{b - x_j\beta}{\sigma}\right) - \Phi\left(\frac{a - x_j\beta}{\sigma}\right)\right\}$$

- The marginal effect of an individual who moves from renting to owning a house is estimate to decrease credit card spending by 15.35483 percent.

(2.3)

Now we are interested in explaining the number of major derogatory reports as function of log income, age, the number of dependents, home ownership status and ratio of monthly credit card expenditure to yearly income.

New Model:

$$Majordrg = \beta_1 + \beta_2 \ln Income + \beta_3 Age + \beta_4 Adepcnt + \beta_5 Ownrent + \beta_6 Exp_I nc + \varepsilon$$

(2.2.a)

Estimate this model using Poisson regression for credit cardholders only. What is the effect of 10% increase in income on the expected value (mean) of the number of major derogatory reports? Is Poisson regression a good specification for the data at hand?

- The poisson regression is not a bad model for this data but we need to be aware of overdispersion issues. When we test for over overdispersion we see there is overdispersion and we can use a quaipoisson regression to allow our Dispersion parameter to be estimated or we could use a negative binomial regression to deal with the overdispersion, which is the next regression we run.

(2.2.b)

Estimate this model using negative binomial regression for credit cardholders only. What is the effect of 10% increase in income on the expected value (mean) of the number of major derogatory reports?

- Estimating with the negative binomial take into account the overdispersion problems from the poisson model and allows mean and variance to be different. We obtain different paramater estimates while using this negative binomial. It appears the negative binomial regression model is a more appropriate model to employ given this specific data set.

(2.2.c)

Estimate the two models taking into account the truncation.What is the effect of 10% increase in income on the expected value (mean) of the number of major derogatory reports?

- Figures 1 and 2 below show the SAS regression output from using the Truncated Poisson and the Truncated Negative Binomial models. It is generally advisable to employ a censored model over a truncated model because the censored model will preserve valuable data while the truncated model throws away data. We can clearly see that while using a truncated model we obtain paramater estimates that are different from previous results, in fact, in several situations we see a sign flip while employing a truncated model. This is alarming.

Table 4: Summary Statistics: Problem 2.3

Statistic	N	Mean	St. Dev.	Min	Max
CARDHLDR	10,499	1.000	0.000	1	1
DEFAULT	10,499	0.095	0.293	0	1
AGE	10,499	33.675	10.291	0.000	88.667
ACADMOS	10,499	55.904	64.127	0	564
ADEPCNT	10,499	0.990	1.274	0	9
MAJORDRG	10,499	0.143	0.462	0	6
MINORDRG	10,499	0.221	0.637	0	7
OWNRENT	10,499	0.479	0.500	0	1
INCOME	10,499	2,606.126	1,287.983	50.000	8,333.250
SELFEMPL	10,499	0.054	0.225	0	1
INCPER	10,499	22,581.360	13,754.970	700.000	150,000.000
EXP_INC	10,499	0.091	0.110	0.0001	2.038
SPENDING	10,499	226.983	294.101	0.111	4,810.309
LOGSPEND	10,499	4.729	1.405	−2.197	8.479
Ln_income	10,499	7.766	0.440	3.912	9.028

Table 5: Regression output used to answer Problem 2.3

	Dependent variable:	
	MAJORDRG	
	Poisson	*negative binomial*
	(1)	(2)
Ln_income	0.697***	0.736***
	(0.063)	(0.078)
AGE	0.021***	0.024***
	(0.003)	(0.003)
ADEPCNT	0.045**	0.042
	(0.020)	(0.026)
OWNRENT	−0.093	−0.084
	(0.060)	(0.073)
EXP_INC	1.303***	1.489***
	(0.163)	(0.242)
Constant	−8.269***	−8.709***
	(0.476)	(0.595)
Observations	10,499	10,499
Log Likelihood	−4,557.087	−4,325.300
θ		0.335*** (0.027)
Akaike Inf. Crit.	9,126.173	8,662.599
Note:	*p<0.1; **p<0.05; ***p<0.01	

Parameter Estimates for Truncated Poisson Model				
Effect	Estimate	Standard Error	z Value	Pr > |z|
Intercept	-4.7200	0.9597	-4.92	<.0001
lnincome	0.4331	0.1231	3.52	0.0004
AGE	0.02104	0.005257	4.00	<.0001
ADEPCNT	0.03131	0.03630	0.86	0.3884
OWNRENT	-0.2423	0.1141	-2.12	0.0337
EXP_INC	0.3833	0.3150	1.22	0.2237

Figure 5: SAS Output of Truncated Poisson

Parameter Estimates for Truncated Negative Binomial Model				
Effect	Estimate	Standard Error	z Value	Pr > |z|
Intercept	-5.3033	1.1269	-4.71	<.0001
lnincome	0.4602	0.1396	3.30	0.0010
AGE	0.02322	0.006134	3.79	0.0002
ADEPCNT	0.03246	0.04085	0.79	0.4268
OWNRENT	-0.2522	0.1277	-1.97	0.0483
EXP_INC	0.4439	0.3939	1.13	0.2598
Scale Parameter	0.4017	0.2564		

Figure 6: SAS Output of Truncated Negative Binomial