Agricultural and Applied Economics 637
Applied Econometrics II
Assignment IV
Maximum Likelihood Estimation
(Due: April 6, 2018)

In this assignment, I would like you to apply the theoretical Maximum Likelihood material to some empirical applications. Make sure you hand in your MATLAB code and output files.

Total Points: **150 pts**.

1. (**50 pts**) Let $y_1, y_2, \ldots, y_T$ be a random sample from a population with the following PDF: $f(y_t) = \frac{\lambda^{y_t} e^{-\lambda}}{y_t!}$ where y_t can take any non-negative integer value (e.g., 0, 1, 2, 3,…) . A variable with this PDF is referred to as having a Poisson distribution. One can show that under this distribution, the mean ***and*** variance of this random variable is equal to the λ parameter. The file ***count intro.ppt*** contains a brief overview of what is referred to as a Count Data model which is based on the Poisson distribution. For more detail refer to Greene p. 802-809 or Cameron and Trivedi, 665-682. This modeling framework is often used when the dependent variable is a count of the number of times a particular event occurs or activity undertaken. Given the above distribution function, the total sample log-likelihood function is:

$$L(\lambda|y) = \sum_{t=1}^{T} \left(-\lambda_t + y_t \ln(\lambda_t) - \ln(y_t!)\right) \qquad (1.1)$$

Let's extend the above model to where we allow the mean and variance of the Poisson random variable, λ, to be a function of a set of exogenous variables. That is, like the CRM we would like the mean value of the dependent variable to be conditional on a set of exogenous variables and a set of unknown parameters. The standard Poisson regression model assumption is to use the exponential mean parameterization: $\lambda_t = \exp(Z_t\beta)$, where t=1,…,T, Z is a matrix of exogenous variables and β is the vector of unknown coefficients to be estimated. Using the above likelihood function, one can represent the sample log-likelihood function for the count model to be:

$$L = \sum_{t=1}^{T} \left(-\lambda_t + y_t Z_t\beta - \ln(y_t!)\right) \qquad (1.2)$$

where $\lambda_t=\exp(Z_t\beta)$. The Poisson Maximum Likelihood estimator, β^* is the solution to the K nonlinear equations corresponding to the 1st order condition for maximum likelihood: $\sum_{t=1}^{T} \left(\left(y_t - \exp(Z_t\beta)\right)Z_t'\right) = \underset{(Kx1)}{0}$. If Z_t includes a constant term then the residuals $y_t - \exp(Z_t\beta)$ sum to 0. In addition, one can show that the likelihood function

shown in (1.2) is globally concave, hence solving these equations via iterative algorithms yields unique parameter estimates.

With the increased proliferation of electronic devices such as televisions and computers, one of the contributing factors to electricity consumption is both the number of these devices available and how long each of these devices are being used. Let's examine one piece to this puzzle, the number of devices in the home. Using the above modeling framework, I would like you to examine what determines the number of televisions used in the home. The data file, RECS_Data_assign_4.xlsx is an Excel file that contains a portion of the 2009 Residential Energy Consumption Survey (RECS) data collected by the Energy Information Administration of the U.S. Department of Energy (DOE). There are a total of 12,083 observations in the base dataset.

The table to the right shows the frequency distribution of the number of televisions, computers and printers in the homes of survey respondents. In 2009, 46% of household had more than 2 televisions. Less than 1.5% of the household had no televisions in 2009. Alternatively, 21.6% of the surveyed households had no computers (desktop or portable) in the home. More than 37% of households did not have a printer. Whereas more than 9% had more than 1 printer.

# of Items	No. of TV's		No. of Computers		No. of Printers	
	# of HH	% of HH	# of HH	% of HH	# of HH	% of HH
0	148	1.2%	2614	21.6%	4530	37.5%
1	2431	20.1%	4979	41.2%	6417	53.1%
2	3977	32.9%	2738	22.7%	958	7.9%
3	2892	23.9%	1099	9.1%	143	1.2%
4	1569	13.0%	420	3.5%	30	0.2%
5	678	5.6%	142	1.2%	4	0.0%
6	255	2.1%	58	0.5%	1	0.0%
7	82	0.7%	12	0.1%	0	0.0%
8	34	0.3%	15	0.1%	0	0.0%
9	8	0.1%	3	0.0%	0	0.0%
10	6	0.0%	2	0.0%	0	0.0%
11	0	0.0%	0	0.0%	0	0.0%
12	2	0.0%	0	0.0%	0	0.0%
13	0	0.0%	0	0.0%	0	0.0%
14	1	0.0%	0	0.0%	0	0.0%
15	0	0.0%	1	0.0%	0	0.0%
Total	12083	100.0%	12083	100.0%	12083	100.0%

The file RECS2009_codebook contains a listing of variable definitions.

(a) (15 pts) Estimate the Poisson regression model using the likelihood function that allows for the count variable mean and variance to depend on a set of exogenous variables. Estimate the model where the dependent variable is the number of televisions (TVCOLOR) and its mean and variance are assumed to be determined by the following variables:
Intercept, *F_S*, *N_Rooms*, *StudioD*, *HHINC*, *%<5*, *%5_14*, *%15_19*, *SeniorD*, *ATHOME*, *TELLWORK*, and *PerKWHPR*

where: *F_S* = *STORIES* if the home is a house or *NAPTFLRS* if an apartment (#)

N_Rooms =Total number of rooms excluding bathrooms (#)

$$StudioD = \begin{cases} 1, \text{ if home is studio apartment} \\ 0, \text{ otherwise} \end{cases}$$

HHINC = Household gross income using range mid-points ($). For the last category assume a gross income of $200,000;

%<5 = % of HH members< 5 years old

%5_14 = % of HH members between 5 and 14 years of age

%15_19 = % of HH members between 15 and 19 years of age

$$SeniorD = \begin{cases} 1 \text{ if all household members are more than 59 years of age;} \\ 0 \text{ Otherwise} \end{cases}$$

TELLWORK = 1 if at least one household member telecommutes or teleworks, 0 otherwise

ATHOME = 1 if at least 1 household member at home on typical week day, 0 otherwise

PerKWHPR = average price per KWH ($/kwh).

***Note**: Make sure you include an intercept. Otherwise the model has a difficult time obtaining a solution.*

To estimate these parameters, use the BHHH algorithm code I distributed in class to estimate the ML model in which you use numerical gradients of the log-likelihood function. Present the estimated coefficients, associated standard errors and total sample log-likelihood function value. Evaluate one of the measures of the degree of explanatory power of the regression model outlined in Greene p. 804-805. *(**NOTE**: When implementing the above log-likelihood function you will need to evaluate the natural logarithm of the factorial number of TV's. Without having to actually evaluate the factorial of a large number and then taking the logarithm which could cause problems due to numerical accuracy, use the return from the following: **gammaln(TVCOLOR+1)**.[1])*

(b) (5 pts) Undertake a **single** likelihood ratio test using the above results to examine the null hypothesis that household age compositions of household members as a group of exogenous variables do not impact the number of household televisions.

(c) (10 pts) Using the results from the Poisson regression model, compare the estimated marginal effect of a change in household income, HHINC on the number of televisions when the individual is typically at home (ATHOME=1) versus not at home (ATHOME=0) during the week. Statistically test whether these two marginal

[1] The *Gammaln* function returns the logarithm of the gamma function, which is the continuous version of the factorial.

impacts are equal. (***Note***: *For this question, except for ATHOME, the exogenous variables should be set at their overall sample mean values*)

(d) (10 pts) Using the results from the Poisson regression model, evaluate the elasticity of a change in gross income on the expected number of televisions. Estimate this elasticity using two methods: (i) at the **mean value** of the data; and (ii) the **average** of the elasticity values calculated over all observations. What are the elasticity estimates obtained under both methods? Statistically test whether these elasticities are different from 1.0 **individually**.

(e) (5 pts) Use a **single statistic** to test whether the elasticities calculated in (d) are the **same value**.

(g) (5 pts) At the mean of the data what is the expected number of televisions? Is this number statistically different from the sample mean of televisions per household?

2. (**25 pts**) The assumed equality of conditional mean and variances under the Poisson model is one of its major shortcomings. As Adam will review in lab, the most common extension of the Poisson model is the Negative Binomial (NEGBIN) model. Under the NEGBIN specification, the Poisson model is extended by introducing an observation specific unobserved effect impacting the conditional mean value. That is, let's have: $\ln(\mu_i) = X_i\beta + \varepsilon_i$ where μ_i is the i^{th} observation's mean and the disturbance term, ε_i, is either a specification error or cross-sectional heterogeneity that is characteristic of micro-level data (Greene, p.806).

Following Greene (2016), the count variable's unconditional distribution can be represented by the following: $f(y_i \mid X_i) = \int_0^{\infty} \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} g(u_i)\,du$ where $\lambda_i = \exp(X_i\beta)$.

The distinguishing characteristic of the NEGBIN model is that the distribution has a conditional mean of λ_i and conditional variance, $\text{var}(y_i|X)$, represented via the following: $\text{var}(y_i \mid X_i) = \lambda_i\left(1 + \frac{\lambda_i}{\gamma}\right)$.

What is interesting about this specification is that a test of the Poisson distribution can be obtained by testing whether $(1/\gamma)=0$. In contrast to the Poisson model, the LLF for this type of variable is rather complicated. After some derivation, the i^{th} observation's NEGBIN LLF is:[2]

$$LLF_i(\beta,\gamma) = \sum_{j=0}^{(y_i-1)} \ln(j+\gamma) - \ln y_i! - (y_i+\gamma)\ln\left(1+\frac{\exp(X_i\beta)}{\gamma}\right) - y_i\ln(\gamma) + y_i X_i\beta$$

[2] Alternatively, you can use the mixed-density pdf described at the bottom of page 675 of Cameron and Trivedi to create a likelihood function. In this case, include your dispersion parameter $1/\alpha$ instead of $1/\gamma$. Recall that the ln(Gamma(.)) is given by the gammaln() function in MATLAB.

(a) (10 pts) Estimate the NEGBIN2 count model using the RAND Health Insurance Experiment data ***(newranddata.xls)*** used in lab #7 with Adam. In other words, you are replicating Column NB2-PML of Table 20.5 on page 673 of Cameron and Trivedi – your answers should be close to identical. Use the number of doctor visits as the dependent variable, and the same set of exogenous variables as used for the Poisson regression in lab. The data is already cleaned and should be ready for estimation. All you need to do is generate the appropriate log-likelihood function for maximization. Use the same starting values as the Poisson estimation in the lab file, adding an additional parameter with a reasonable starting value. Present the usual regression results in a table. How many iterations did it take to obtain the optimal solution? How does this compare with the estimation in question #1?

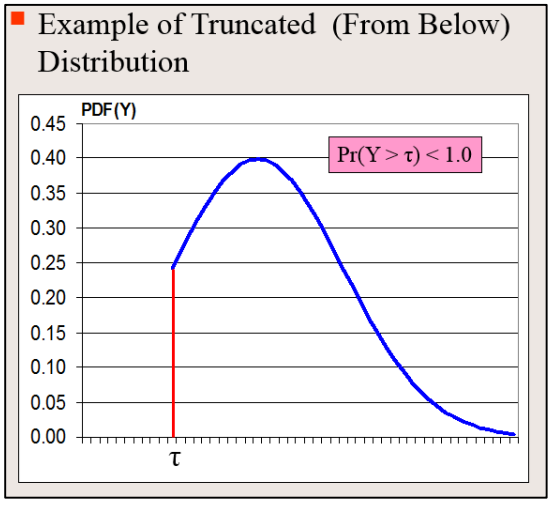
(b) (5 pts) Conduct a likelihood ratio test that the distribution is a Poisson distribution.

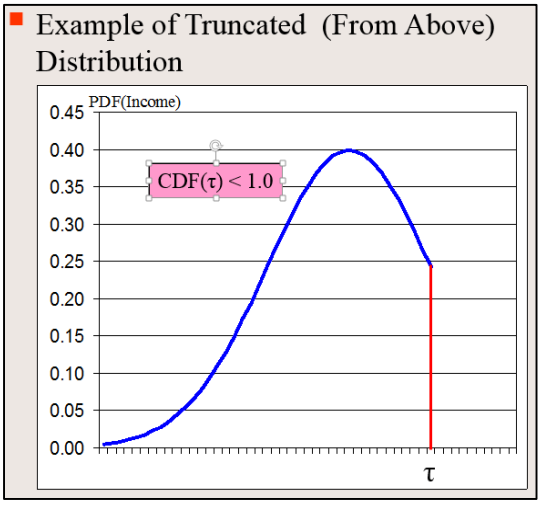
(c) (5 pts) Undertake a Wald test that the distribution is a Poisson distribution.

(d) (5 pts) What is the result of undertaking an LM test that the distribution is Poisson.

3. (**55 pts**) In a few weeks we will be examining how to estimate what is referred to as a truncated regression model. A truncated dependent random variable is a random variable where either an upper portion, bottom portion or both areas of its distribution are omitted from the data used in a regression model. The key concept of random variable truncation is that it is some characteristic of the **dependent variable** is used to define the truncation. That is, the values of exogenous variables are not used to define the sample.

   Below I show two types of random variable truncation: lower and upper. For more detail concerning truncated regression models refer to Greene, p. 837-839. An example of a regression model where the dependent variable (e.g., household income) has an upper (i.e., from above) truncation can be found in Hausman and Wise (1977). I think they do a good job in explaining how they use ML estimation methods to estimate a regression model and how they develop the associated LLF used for parameter estimation.





In this question I would like you to estimate household per capita electricity use (***PC_Use***) but where we only include observations with per capita use is more than 3,000 KWH (i.e., ***PC_Use*** > **τ** where τ = 3,000). That is, we have a lower truncation of per capita electricity use. One can show that the conditional PDF of ***PC_Use*** given this lower truncation, f(PC_Use| PC_Use>τ) can be represented via the conditional PDF shown to the right where we assume the PC_Use is normally distributed. Note that ϕ is the standard normal PDF and Φ is the standard normal CDF, and [1-Φ(•)] is the Prob(PC_Use > τ).[3] Given the above, the LLF for this truncated regression model, $\mathbf{L_{Trunc}}$, can be represented via the following:

$$\mathrm{PC\_Use} = \mathrm{X}\beta + \varepsilon \quad \text{where } \mathrm{PC\_Use} \sim \mathrm{N}\left(\mathrm{X}\beta, \sigma^2\right)$$

$$\rightarrow \mathrm{f}\left(\mathrm{PC\_Use} \mid \mathrm{PC\_Use} > \tau, \mathrm{X}\beta, \sigma^2\right) = \frac{\frac{1}{\sigma}\phi\left(\frac{\mathrm{PC\_Use} - \mathrm{X}\beta}{\sigma}\right)}{1 - \Phi\left(\frac{\tau - \mathrm{X}\beta}{\sigma}\right)}$$

$$\mathrm{L_{Trunc}} = -\frac{T_1}{2}\left[\ln 2\pi + \ln\sigma^2\right] - \frac{1}{2\sigma^2}\sum_{t=1}^{T_1}\left(y_t - \mathrm{X}_t\beta\right)^2 - \sum_{i=1}^{T_1}\ln\left[1 - \Phi\left(\frac{\tau - \mathrm{X}_t\beta}{\sigma}\right)\right]$$

where T_1 represents the number of observations with a value greater than τ.

Later we will also show that given this lower truncation we have

[3] In general we know that: $\mathrm{f}\left(y|\mu, \sigma^2\right) = \mathrm{f}\left(y|y > \tau, \mu, \sigma^2\right)\Pr\left(y > \tau\right) \rightarrow \mathrm{f}\left(y|y > \tau, \mu, \sigma^2\right) = \frac{\mathrm{f}\left(y|\mu, \sigma^2\right)}{\Pr\left(y > \tau\right)}$

$E(y_i | y_i > 0, X_i) = X_i\beta + \sigma\lambda_i$ where $\lambda_i = \dfrac{\phi\left(\dfrac{X_i\beta}{\sigma}\right)}{\Phi\left(\dfrac{X_i\beta+\tau}{\sigma}\right)}$ which implies that

$$\frac{\partial E\left(y_i \mid y_i > 0, X_i\right)}{\partial X_{k,i}} = \beta_k\left(1 - \lambda_i^2 + \left(\frac{X_i\beta}{\sigma}\right)\lambda_i\right).$$

NOTE for the ease in completing this assignment: in parts (a) – (e), before doing any analysis, drop all observations with percapKWH > 10000.[4] Without this edit, upper-end outliers in the dependent variable make the truncated ML estimation by the BHHH algorithm a bit trickier than intended for this course.

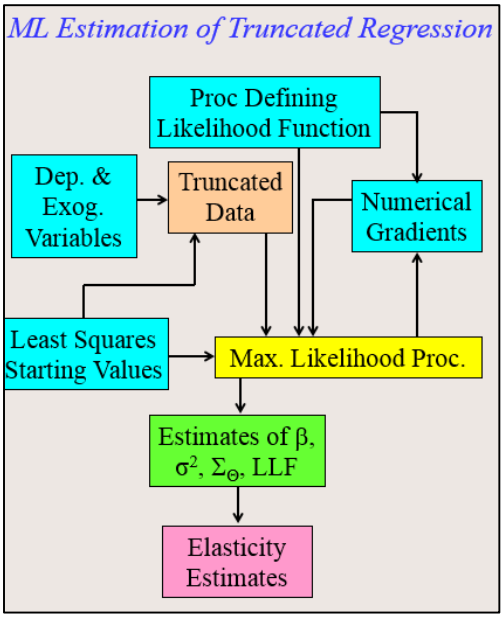
(a) (**10 pts**) Given the above I would like to use another subset of the 2009 RECS survey but this time to you want to estimate the following linear regression model: *PercapKWH* = $X\beta + \varepsilon$ where $\varepsilon \sim N(0,\sigma^2)$ and the exogenous variable matrix X is composed of a vector of ones to generate an intercept term and the variables: ***ln(HDD65)***, ***ln(CDD65)***, ***House_Age***, ***Elec_Pr***, ***Tot_sqFt_H/C***, ***Elec_Stove***, ***EnergyStar***, ***Elec_Water***, ***Air_Cond***, (***Air_Cond*** × ***ln(CDD65)***), ***Elec_Heat*** and (***Elec_Heat*** × ***ln(HDD65)***) . Using your previously developed MATLAB code estimate the linear regression model, present the standard regression output obtained from estimating this model. Are the results consistent with your expectations?

(b) (**5 pts**) At the *mean of the data* used in estimation, what is the price elasticity? Is this elasticity statistically different from 0? Is it statistically different from -1.0?

(c) (**10 pts**) Given the above regression results what is the *average marginal impacts* of a CDD on electricity use for those households with an air conditioning system vs those households without an Air Conditioner present? Are these two *average marginal effects* equal to one another from a statistical perspective?

[4] Unless you've ALREADY managed to obtain convergence with your truncated regression in part (d) with the full dataset.

(d) (**15 pts**) Now estimate the regression model assuming that you only have data on those households with more than 3,000 KWH per capita electricity consumption (i.e., delete those observations with annual KWH's < 3,000. Estimate this truncated regression model using maximum likelihood methods. Present your typical maximum likelihood estimation results. The figure provides a general overview of how you can obtain these truncated regression results:



(e) (**15 pts**) Provide an assessment relative to the results discussed in (c) and (d) as to the implications of asserting that the truncated results are representative of the entire population? Provide statistical evaluations to support your comments.