ECO 6375
Prof. Millimet
Problem Set #2: Binary Choice & Count Models

1. Download the *Stata* data set *ej.dta* from the course web site. Estimate a probit model where the dependent variable is a binary indicator if the wife works (*laborw>0*) and the regressors are her education, *educw*, age, *agew*, and a dummy variable for white, *white*.

   (a) Program yourself to compute the sample average of the individual-specific marginal effects for *educw* and *agew.*

   (b) Compare your answers to those from -*dprobit*- and -*probit*- with -*margin*-. Account for the differences between the two commands, as well as the commands and your answer in (a) where there are differences.

   [Note: You may need to install -margin- in Stata. Type -ssc install margin-.]

2. Suppose that a linear probability model is used to estimate the model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

   where y takes on the values zero and one and the single regressor, x, is continuous. Obtain an expression for the OLS estimator of β in terms of the mean of x in the sub-sample of $y = 1$ and $y = 0$, and the variance of x. Interpret the result.

3. Consider the following probit estimation results

Variable	**β (std error)**
Experience	0.031
	(0.009)
Male	0.174
	(0.059)
Constant	0.632
	(0.200)

   where the dependent variable equals one for individuals in the labor force and zero otherwise.

   (a) Write down an expression for the probability that a male individual with ten years of experience is in the labor market.

   (b) Write down an expression for the probability that a female individual with five years of experience is in the labor market.

   (c) Write down an expression for the marginal effect of an additional year of experience for a male individual. Is the marginal effect constant, or does it vary across males?

   (d) Write down an expression for the marginal effect of an additional year of experience for a female individual.

   (e) Are the marginal effects in (c) and (d) equal for a male and female individual with the identical years of experience? Justify your answer.

4. A researcher interested in the relationship between parenting, age, and schooling has data for the year 2000 for a sample of 1,167 married males and 870 married females aged 35 to 42. In particular,

she is interested in how the presence of young children in the household is related to the age and education of the respondent. She defines $CHILDL6$ to be 1 if there is a child less than 6 years old in the household and 0 otherwise. She regresses $CHILDL6$ on age and years of schooling, S, for males and females separately using probit models. Defining the probability of having a child less than 6 in the household to be $p = F(Z)$ where

$$Z = \beta_1 + \beta_2 AGE + \beta_3 S$$

she obtains the results shown in the table (standard errors in parentheses).

	Males	**Females**
AGE	-0.137 (0.018)	-0.154 (0.023)
S	0.132 (0.015)	0.094 (0.020)
Constant	0.194 (0.358)	0.547 (0.492)
$\bar{Z}$	-0.399	-0.874
$f(\bar{Z})$	0.368	0.272

For males and females separately, she calculates

$$\bar{Z} = b_1 + b_2 \overline{AGE} + b_3 \overline{S}$$

where $\overline{AGE}$ and $\bar{S}$ are the mean values of age and S and b_1, b_2, and b_3 are the probit coefficients in the corresponding regression, and she further calculates

$$f\left(\bar{Z}\right) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\bar{Z}^2\right\}$$

where $f(Z) = \frac{dF}{dZ}$. The values of $\bar{Z}$ and $f(\bar{Z})$ are shown in the table.

(a) Explain in general terms why the researcher did not regress $CHILDL6$ on AGE and S using OLS.

(b) Explain why probit analysis avoids the problems associated with OLS.

(c) Explain how one may derive the marginal effects of the explanatory variables on the probability of having a child less than 6 in the household, and calculate for both males and females the marginal effects at the means of AGE and S.

(d) At a seminar someone asks the researcher whether the marginal effect of S is significantly different for males and females. The researcher does not know how to test whether the difference is significant and asks you for advice. What would you say?

5. Consider the Poisson distribution

$$f(y) = \frac{\exp\{-\lambda\}\lambda^y}{y!}$$

where y is the number of children in a household. You have data on N households, however the sample only includes households with at least one child. Construct the appropriate log-likelihood function.

6. As a consultant, you are hired to analyze a company's data to assess the role of various inputs on the number of new patents obtained by the firm in a particular year. You are provided annual data on the number of new patents obtained in each year, P_t, as well as data on the firm's annual R&D expenditures, RD_t, the number firm employees engaged in research-related activities, R_t, and the number of firm employees engaged in non-research-related activities, N_t (i.e., administrative work). The data cover the past 30 years, $t = 1, ..., 30$, and the number of patents obtained in any given year ranges from zero to four.

   (a) You decide to proceed using a poisson model, where the form of the Poisson pdf is

   $$\Pr(P_t|x_t, \theta) = \frac{\exp\{-\lambda_t\}\lambda_t^{P_t}}{P_t!}$$

   where λ_t =exp$\{\beta_1 RD_t + \beta_2 R_t + \beta_3 N_t\}$ is the mean (i.e., $\mathrm{E}[P_t|RD_t, R_t, N_t]$) Derive the log likelihood function used to obtain the effects of the three firm characteristics on the number of new patents obtained per year.

   (b) Discuss two concerns you may have about the appropriateness of the model utilized in (a), as well as how you might go about deciding if these concerns are valid.

   (c) Assuming you can ignore the concerns in (b), you report your estimates of $\widehat{\beta}_k$, $k = 1, 2, 3$, to the firm. How do you explain the meaning of these coefficients?

   (d) The firm then asks you to assess the effect of adding 10 additional research employees on the *expected* number of patents in a year. Describe three methods for quantifying the impact. Be as detailed as possible (within reason). [*Hint: This refers to the marginal effect of a increase in* R_t *by ten.*]