# Problem Set 7 <br> Economics 60303 <br> (Due Friday, April 11, 2013) 

## Bill Evans

Spring 2014

1. Blood tests to detect cholesterol levels are a common component of routine medical exams. Patients with high cholesterol levels are defined as those with 240 mg or more of cholesterol per deciliter of blood. Given a high cholesterol reading, a doctor may then write a prescription for cholesterol lowering medication such as Lipitor. There is a concern that doctors are being more lax at providing Lipitor and other cholesterol lowering drugs to patients. Suppose one has a sample of $n$ adults currently not using cholesterol lowering medication and each of these patients gets a new cholesterol test generating a cholesterol level reading labeled as $\mathrm{c}_{\mathrm{i}}(\mathrm{in} \mathrm{mg} / \mathrm{dl})$. Suppose that in response to the test results, a doctor can decide to write a prescription for cholesterol $\left(\mathrm{y}_{\mathrm{i}}=1\right)$ or not $\left(\mathrm{y}_{\mathrm{i}}=0\right)$. A researcher is interested in econometrically uncovering at what level for $\mathrm{c}_{\mathrm{i}}$ must a patient have for the doctor to write a prescription. Suppose there is a test level c* above which a prescription is written. Therefore

$$
\begin{aligned}
& y_{i}=1 \text { if } c_{i} \geq c^{*} \\
& y_{i}=0 \text { if } c_{i}<c^{*}
\end{aligned}
$$

a. Show how one can use a probit model to uncover information about the distribution of $c^{*}$ across doctors in the country. In your model, what are $\operatorname{Pr}\left(\mathrm{y}_{\mathrm{i}}=1\right)$ and $\operatorname{Pr}\left(\mathrm{y}_{\mathrm{i}}=0\right)$. Be explicit about what assumptions must be made for your model.
b. Suppose that the data set contains a vector of characteristics for the patient $\left(x_{i}\right)$. Suppose that one suspects that the threshold used by a doctor varies by observed characteristics of the patient. How can one use this information in the analysis?
2. The Pareto distribution is used to describe the right hand tail of income distributions. If $y_{i}$ is income for person i and $y_{i}>y_{m}$, the pdf of income is defined as $f\left(y_{i}\right)=\alpha \frac{y_{m}^{\alpha}}{y_{i}^{\alpha+1}}$ for $\infty \geq y_{i}>y_{m}$ and $\alpha>0$. Graphing the pdf, notice that it is downward sloping for all y which pretty much is the right hand tail of the distribution for income.
a. Show that $\operatorname{Pr}\left(y_{i}>a\right)=\left(\frac{y_{m}}{a}\right)^{\alpha}$ for $y_{m}<a$
b. What is $E\left[y_{i}\right]$ given $\infty \geq y_{i}>y_{m}$
c. Suppose one has a sample of n individuals whose income is in excess of $y_{m}$. What is the log likelihood function that describes this sample?
d. What is the MLE for $\hat{\alpha}$ ?
3. Suppose you want to estimate a probit model where the only covariate, $X_{i}$, which is a dummy variable and therefore, $\operatorname{Pr}\left(\mathrm{Y}_{\mathrm{i}}=1\right)=\Phi\left(\beta_{0}+\mathrm{X}_{\mathrm{i}} \beta_{1}\right)$. In this case, there are only four states of the world:

$$
\begin{aligned}
& \mathrm{P}_{11}=\operatorname{Prob}\left(\mathrm{Y}_{\mathrm{i}}=1 \mid \mathrm{X}_{\mathrm{i}}=1\right)=\Phi\left(\beta_{0}+\beta_{1}\right)=\mathrm{P}_{1} \\
& \mathrm{P}_{10}=\operatorname{Prob}\left(\mathrm{Y}_{\mathrm{i}}=1 \mid \mathrm{X}_{\mathrm{i}}=0\right)=\Phi\left(\beta_{0}\right)=\mathrm{P}_{0} \\
& \mathrm{P}_{01}=\operatorname{Prob}\left(\mathrm{Y}_{\mathrm{i}}=0 \mid \mathrm{X}_{\mathrm{i}}=1\right)=1-\Phi\left(\beta_{0}+\beta_{1}\right)=1-\mathrm{P}_{1} \\
& \mathrm{P}_{00}=\operatorname{Prob}\left(\mathrm{Y}_{\mathrm{i}}=0 \mid \mathrm{X}_{\mathrm{i}}=0\right)=1-\Phi\left(\beta_{0}\right)=1-\mathrm{P}_{0}
\end{aligned}
$$

Suppose there are $\mathrm{N}_{11}$ observations for $\mathrm{P}_{11}, \mathrm{~N}_{00}$ for $\mathrm{P}_{00}$, etc. Write a log likelihood function solely as a function of $\mathrm{N}_{11}, \mathrm{~N}_{10}, \mathrm{~N}_{01}, \mathrm{~N}_{00}, \mathrm{P}_{1}$ and $\mathrm{P}_{0}$. In the general case, find an expression for the maximum likelihood estimates for $\hat{P}_{1}$ and $\hat{P}_{0}$

In this case, what is the estimate of the marginal effect $\Phi\left(\beta_{0}+\beta_{1}\right)-\Phi\left(\beta_{0}\right)$ ?
4. Continue with problem 3. The $2 \times 2$ table below provides the number of observations in each pair of ( $\mathrm{X}, \mathrm{Y}$ ) combinations:

|  | $\mathrm{X}_{\mathrm{i}}=0$ | $\mathrm{X}_{\mathrm{i}}=1$ |
| :--- | :--- | :--- |
| $\mathrm{Y}_{\mathrm{i}}=0$ | 35 | 15 |
| $\mathrm{Y}_{\mathrm{i}}=1$ | 20 | 30 |

What are the maximum likelihood estimates for $\hat{P}_{1}$ and $\hat{P}_{0}$ in this case? Given these values, what are the implied ML estimates for $\hat{\beta}_{0}$ and $\hat{\beta}_{1}$ ?
5. Continuing with problem 3, suppose one were to estimate the relationship between $Y$ and $X$ via a linear probability model, $\mathrm{Y}_{\mathrm{i}}=\beta_{0}+\mathrm{X}_{\mathrm{i}} \beta_{1}+\varepsilon_{\mathrm{i}}$. What would be the OLS estimate of $\beta_{1}$ (write as a function of $\mathrm{N}_{11}, \mathrm{~N}_{01}$, etc.)?
6. Continuing with problem 3 , instead of using a probit specification, suppose one were to estimate a logit model. In this case

$$
\begin{aligned}
& \mathrm{P}_{11}=\operatorname{Prob}\left(\mathrm{Y}_{\mathrm{i}}=1 \mid \mathrm{X}_{\mathrm{i}}=1\right)=\mathrm{P}_{1}=\frac{e}{1+e^{\beta_{0}+\beta_{1}}}=F\left(\beta_{0}+\beta_{1}\right) \\
& \mathrm{P}_{10}=\operatorname{Prob}\left(\mathrm{Y}_{\mathrm{i}}=1 \mid \mathrm{X}_{\mathrm{i}}=0\right)=\mathrm{P}_{0}=\frac{e^{\beta_{0}}}{1+e^{\beta_{0}}}=F\left(\beta_{0}\right)
\end{aligned}
$$

In this case, what are the MLE of $\hat{P}_{1}$ and $\hat{P}_{0}$ ? What is the estimate of the marginal effect $F\left(\beta_{0}+\beta_{1}\right)-F\left(\beta_{0}\right)$ ?
7. In the data set workplace1.dta, run a linear probability of smoker on worka and use robust to correct for heteroskedasticity. Next, run a probit model and get the marginal effects for the worka variable using dprobit. Finally, use a logistic regression, estimate a logit model and use mfx compute to get the marginal effects. Fill in the table below.

| OLS coefficient or marginal effect on worka variable |  |  |
| :--- | :--- | :--- |
| (standard error) |  |  | \left\lvert\, \(\left.\begin{array}{l}Marginal effect on <br>

worka from probit\end{array} \quad \begin{array}{l}Marginal effect on <br>

worka from logit\end{array}\right.\right]\)| OLS (Linear |
| :--- |
| probability) |

8. Continue with problem 3. In problem set 3, we examined the impact of children on labor supply using twins on the $1^{\text {st }}$ birth as an instrument for the presence of at least a second child in the house. The first stage in the Wald estimate was a regression of SECOND on TWIN1ST where SECOND is a dummy that equals 1 if there are more than 1 kids in the household. Using the data from the previous problem set, estimate the first stage model using a probit.

Now, you are to explain the results you get. Write an equation for the log likelihood function that is a function of $\mathrm{N}_{11}, \mathrm{~N}_{10}, \mathrm{~N}_{01}, \mathrm{~N}_{00}, \Phi\left(\beta_{0}+\beta_{1}\right)$ and $\Phi\left(\beta_{0}\right)$. Next, take the first order conditions with respect to $\beta_{0}$ and $\beta_{1}$. Finally, what are the values for $\mathrm{N}_{11}, \mathrm{~N}_{10}, \mathrm{~N}_{01}, \mathrm{~N}_{00}$ in this context? With these results, why can you NOT get ML estimates in this context?

