Sketch of Solutions for WRITTEN PAPER III (ECON 4135)

Nov,7th, 2006

1 . A student suggests to run an OLS regression of Y on emply, rs, cc and q, including a constant term. Estimate this linear model. Explain briefly what are the main problems with linear regression for this problem.

```
. reg exit  emply rs cc q,robust
```

```
Regression with robust standard errors          Number of obs =      964
                                                F(  4,   959) =     2.47
                                                Prob > F      =   0.0431
                                                R-squared     =   0.0132
                                                Root MSE      =   .22295


------------------------------------------------------------------------------
             |               Robust
      exit |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      emply |  -.0107117   .0043944    -2.44   0.015    -.0193354    -.002088
         rs |  -.0043711   .0016678    -2.62   0.009     -.007644   -.0010982
         cc |  -.0245635   .0253191    -0.97   0.332    -.0742507    .0251237
          q |  -.0000489   .0010215    -0.05   0.962    -.0020536    .0019558
      _cons |    .057686   .0085782     6.72   0.000     .0408518    .0745201
------------------------------------------------------------------------------
```

Main problems for linear regression:

1. The predicted probabilities are not bounded between [0,1],

2. Disturbances are always heteroskedastic, and not normally distributed.

2. Another student suggests instead to apply a probit model. Specify a probit model for this problem, i.e. write down an expression for the conditional probability of exit given the explanatory variables, i.e. $P(Y = 1|emply, rs, cc, q)$.

$P(Y = 1|\text{emply, rs, cc, q}) = \Phi(\beta_0 + \beta_1\text{emply} + \beta_2\text{rs} + \beta_3\text{cc} + \beta_4\text{q})$, where $\Phi()$ is the standard normal CDF.

3. With the help of STATA, estimate the parameters of the probit model. Predict the exit probabilities of the firms in the sample, i.e., calculate $P(Y = 1|\text{emply, rs, cc, q})$ for each firm. (Hint: Use the predict command).

Define the "residual" of the model as the difference between the binary variable Y_i and the conditional expectation, $E(Y_i|\text{emply,rs,cc,q})$:

$$\varepsilon_i = Y_i - E(Y_i|\text{emply,rs,cc,q}) \tag{1}$$

Show that

$$\varepsilon_i = \begin{cases} 1 - P(Y = 1|\text{emply,rs,cc,q}) & \text{if } Y_i = 1 \\ -P(Y = 1|\text{emply,rs,cc,q}) & \text{if } Y_i = 0 \end{cases}$$

Calculate R^2 defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2},$$

where n is the number of firms and $\overline{Y}$, as usual, denotes the mean of the Y_i.

```
 . probit exit  emply rs cc q
```

```
Probit estimates                               Number of obs   =        964
                                               LR chi2(4)      =      12.08
                                               Prob > chi2     =     0.0168
Log likelihood = -193.49119                    Pseudo R2       =     0.0303


------------------------------------------------------------------------------
        exit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       emply |  -.1018457   .0361245    -2.82   0.005    -.1726485   -.0310429
          rs |  -1.512777   1.712408    -0.88   0.377    -4.869034    1.843481
          cc |  -.1651089    .119873    -1.38   0.168    -.4000557    .0698379
           q |  -.0005965   .0081891    -0.07   0.942    -.0166468    .0154539
       _cons |  -1.591897   .0779181   -20.43   0.000    -1.744614    -1.43918
------------------------------------------------------------------------------
```

```
note: 3 failures and 0 successes completely determined.
```

It is easy to see that

$$
\begin{aligned}
\varepsilon_i &= Y_i - E(Y_i|\text{emply,rs,cc,q}) \\
&= Y_i - (1 \cdot P(Y = 1|\text{emply,rs,cc,q}) + 0 \cdot P(Y = 0|\text{emply,rs,cc,q})) \\
&= \begin{cases} 1 - P(Y = 1|\text{emply,rs,cc,q}) & \text{if } Y_i = 1 \\ -P(Y = 1|\text{emply,rs,cc,q}) & \text{if } Y_i = 0 \end{cases}
\end{aligned}
$$

Calculation of R^2 is staightforward $R^2 = 0.0166$ for this dataset.

4. Test the hypothesis that R&D does not influence the probability of exit.

$H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$
As we can see from the printout from Stata. The 95% CI for β_2 is (-4.87,1.84) which includes 0. So we cannot reject the null hypothesis at 5% LOS.

5. Explain how you will calculate the marginal effect of Tobin's q on the probability of exit for the Probit model. Calculate the marginal effect of Tobin's q for the following to firms:

	emply	cc	rs	Tobin's q
Firm 1	3	20	0	1
Firm 2	-5	-7	10	1

The marginal effect is defined as

$$
\begin{aligned}
\frac{\partial P(Y = 1|emply, rs, cc, q)}{\partial q} &= \frac{\partial \Phi(\beta_0 + \beta_1 emply + \beta_2 rs + \beta_3 cc + \beta_4 q)}{\partial q} \\
&= \beta_4 \phi(\beta_0 + \beta_1 emply + \beta_2 rs + \beta_3 cc + \beta_4 q).
\end{aligned}
$$

Where $\phi()$ is the standard Normal PDF.
The marginal effect of Tobin's q for both firms given is close to 0.

6. What is the difference between the Probit model and the Logit model? Estimate a Logit model with the same regressors as you used in the Probit model. What are the ratios between the estimated regression coefficients in the Logit model and in the Probit model (calculate one ratio for each coefficient)? Also predict the probability of exit for each firm in the data set, and compare with the predicted exit probabilities for the Probit model by graphing the two predicted exit probabilities against each other. Why does the estimated regression coefficients in the two regression models differ so much, whereas the predicted exit probabilities tend to be very similar ?

```
. logit exit  emply rs cc q
```

```
Logit estimates                             Number of obs   =          964
                                            LR chi2(4)      =        12.27
                                            Prob > chi2     =       0.0154
Log likelihood = -193.39214                 Pseudo R2       =       0.0308


------------------------------------------------------------------------------
        exit |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       emply |  -.2227231   .0763379    -2.92   0.004    -.3723427   -.0731035
          rs |  -3.070176   3.454662    -0.89   0.374     -9.84119    3.700838
          cc |  -.3184826   .2344954    -1.36   0.174     -.778085    .1411199
           q |  -.0036512   .0183396    -0.20   0.842    -.0395961    .0322938
       _cons |  -2.847477    .168336   -16.92   0.000     -3.17741   -2.517545
------------------------------------------------------------------------------

note: 1 failure and 0 successes completely determined.
```

Ratios between the two regressions' coefficients can be calculated accordingly.

The following figure shows the two predicted probabilities graphed against each other. The red line is the 45 degree line.
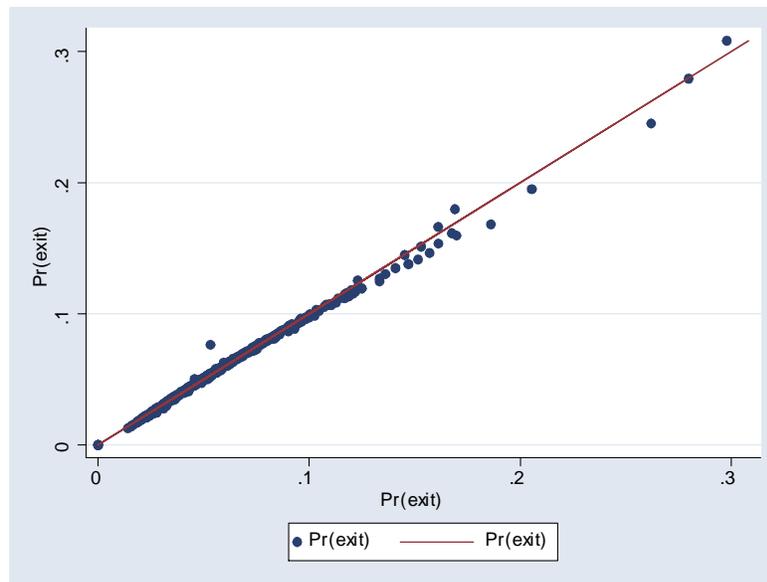

Figure 1: The Predicted Probabilities (Probit vs Logit)

In probit model we have

$P(Y = 1|\text{emply, rs, cc, q}) = \Phi(\beta_0 + \beta_1\text{emply} + \beta_2\text{rs} + \beta_3\text{cc} + \beta_4\text{q})$, where $\Phi(x)$ is the standard normal CDF.

In Logit, we have

$P(Y = 1|\text{emply, rs, cc, q}) = F(\gamma_0 + \gamma_1\text{emply}+\gamma_2\text{rs}+\gamma_3\text{cc}+\gamma_4\text{q})$, where $F(x) = 1/(1 + \exp(-x))$ is the CDF for Logistic distribution.

As we can see from the following diagram, Logistic and Probit CDFs are quite comparable (dash line is a transformed Logitistic distribution with mean 0 and variance 1). The main difference is that the logistic distribution has flatter tails. So in general these two models qualitatively give similar results.
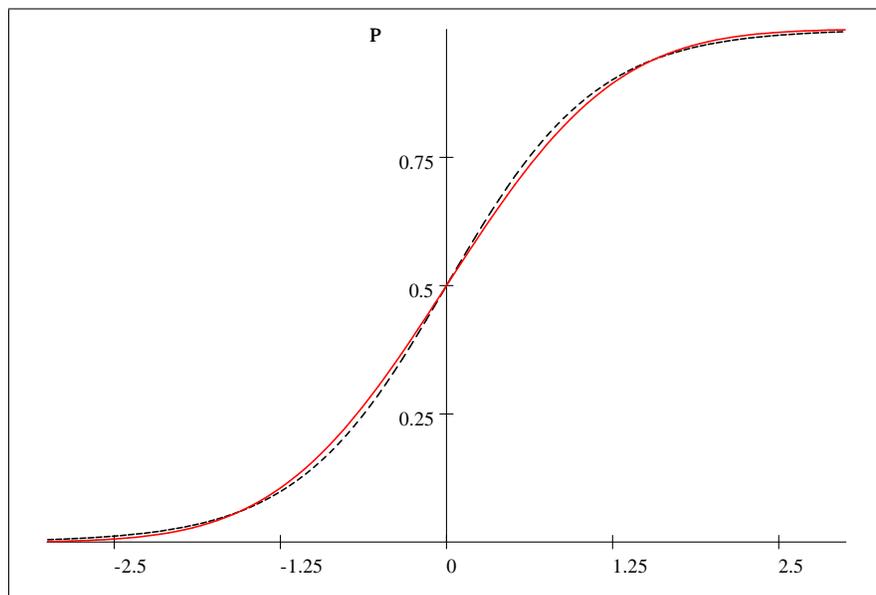
Logit vs Probit CDF

However, as we can see from the Stata printout, the estimates of the parameters of the two models are not directly comparable. The main reason for this is that the variance of the standard normal distribution is one, while the variance of logistic distribution is $\pi^2/3$.