Nonlinear models

CONTENTS:

- **Binary response models: Linear probability Model, Probit and Logit**
- **Multinomial response models: Multinomial Logit**
- **Censored models: Tobit and Selection models**

Binary response models: Linear Probability Model, Probit and Logit

In qualitative response models, the variable to be explained, *y, is* a random variable taking on a finite number of outcomes; in practice, the number of outcomes is usually small. The leading case occurs where *y* is a binary response, taking on the values zero and one, which indicate whether or not certain event has occurred. For example, $y = 1$ if a person is employed, $y = 0$ otherwise.

In binary response models, interest lies primarily on the (conditional) **response probability:**

$$p(\mathbf{x}) \equiv \mathrm{P}(y = 1|\mathbf{x}) = \mathrm{P}(y = 1|x_1, x_2, \dots, x_K) \tag{1}$$

for various values of **x**.

Linear probability model

The linear probability model (LPM) for binary response y is specified as

$$\mathrm{P}(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K \equiv \mathbf{x}\boldsymbol{\beta} \tag{2}$$

$$\mathrm{P}(y = 0|\mathbf{x}) = 1 - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K) \equiv 1 - \mathbf{x}\boldsymbol{\beta} \tag{3}$$

Assuming that x_1 is not functionally related to the other explanatory variables $\beta_1 = \partial \mathrm{P}(y = 1|\mathbf{x})/\partial x_1$. Therefore, β_1 is the change in probability of success given a one-unit increase in x_1. If x_1 is a binary explanatory variable, β_1 is just the difference in the probability of success when x_1=1 and x_1=0, holding the other x_j constant. The important point is that the β_i now measures the effects of the explanatory variables x_i on a particular probability.

Given a random sample, the OLS regression of y on 1, $x_1, x_2, \dots, x_K$ produce consistent and unbiased estimators of the β_j. Heteroskedasticity will be present in this model so it is often convenient to use heteroskedasticity-robust standard errors.[1]

[1] Since y_i can only take two values (0 and 1), given x_i, the error term u_i can only take on two values also $(1\text{-}\beta x_i)$ when y_i=1 and $-\beta x_i$ when y_i=0. This implies that $\mathrm{Var}(u_i) = \beta x_i(1\text{-}\beta x_i)$ meaning that the variance of the error term depends on x_i and therefore we have heteroskedasticity.

EXAMPLE 1 (Married Women's Labor Force Participation)

We can estimate a linear probability model for labor force participation (*inlf*) for married women using the MROZ.RAW data. On the 753 women in the sample, 428 report working non-zero hours during the year. The variables we use to explain labor force participation are age, education, experience, nonwife income in thousands (*nwifeinc*), number of children less than six years of age (*kidslt6*), number of kids between 6 and 18 inclusive (*kidsge6*). We estimate the following model:

$$inlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \beta_3 exper + \beta_4 expersq + \beta_5 age + \beta_6 kidslt6 + \beta_7 kidsge6 \qquad (4)$$

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/mroz

. regress inlf nwifeinc educ exper expersq age kidslt6 kidsge6,r

Linear regression                               Number of obs =     753
                                                F(  7,   745) =   62.48
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2642
                                                Root MSE      =  .42713
```

inlf	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
nwifeinc	-.0034052	.0015249	-2.23	0.026	-.0063988	-.0004115
educ	.0379953	.007266	5.23	0.000	.023731	.0522596
exper	.0394924	.00581	6.80	0.000	.0280864	.0508983
expersq	-.0005963	.00019	-3.14	0.002	-.0009693	-.0002233
age	-.0160908	.002399	-6.71	0.000	-.0208004	-.0113812
kidslt6	-.2618105	.0317832	-8.24	0.000	-.3242058	-.1994152
kidsge6	.0130122	.0135329	0.96	0.337	-.013555	.0395795
_cons	.5855192	.1522599	3.85	0.000	.2866098	.8844287

With the exception of *kidsge6,* all coefficients have reasonable signs and are statistically significant; *kidsge6* is neither statistically significant. The coefficient of *nwifeinc* means that if nonwife income increases by 10 ($10,000), the probability of being in the labor force is predicted to fall by 0.034. Having one more small kid is estimated to reduce the probability of labor force participation by about 0.262, which is fairly a large effect.

Probit and Logit model

We now study binary response models of the form

$$\mathrm{P}(y = 1|\mathbf{x}) = G(\mathbf{x}\beta) \equiv p(\mathbf{x}) \qquad (5)$$

where **x** is 1xK, β is Kx1 and we take the first element of **x** to be unity (constant). For the linear probability model, G(z)=z is the identity function, which means that the response probabilities are not necessarily between 0 and 1 for all z. Now, we will assume that G (.) takes values between the unit interval: 0<G (z)>1 for all z$\epsilon\mathbb{R}$.

The **probit model** is a special case of equation (5) with:

$$G(z) \equiv \Phi(z) \equiv \int_{-\infty}^{z} \emptyset(v)dv \qquad (6)$$

where $\Phi(z)$ is the cumulative distribution function of a standard normal distribution and $\emptyset(z)$ is the standard normal density is $\emptyset(z) = (2\pi)^{-1/2}\exp(-z^2/2)$.

The **logit model** is a special case of equation (5) with:

$$G(z) = \Lambda(z) \equiv \exp(z)/[1 + \exp(z)] \quad (7)$$

where $\Lambda(z)$ is the logistic function.

In order to implement the probit and logit models, it is important to know how to interpret the β_j. Take into account that in the classical linear regression model, β_j give us the effect on y of a one-unit change in x_j holding all the others x's constant but note that this is not true for the β_j that is estimated in the probit or logit model. First note that the sign of the effect is given by the sign of β_j as in the case of the classical lineal regression model. Now, for the change in the j-th regressor x_j, the marginal change of the conditional probability is given as:

$$\frac{\partial \Pr(y=1|\mathbf{x})}{\partial x_j} = \beta_j f(\mathbf{x}\boldsymbol{\beta}) \quad (8)$$

where *f* is the probability density function. This is equal to the standard normal probability density function in the probit model and the logistic distribution probability density function in the logit model.

The coefficients of the probit and the logit model are not directly comparable since they are scaled differently. However, signs and significance are comparable.

Both the logit and the probit model are estimated by **Maximum Likelihood Estimation (MLE)**

EXAMPLE 1 (Married Women's Labor Force Participation-Continuation)

We now estimate the probit and logit models for women labor force participation. The signs of the coefficients are consistent across the models and the same variables are statistically significant in each model (including the LPM above) but, as already commented, the magnitudes of the coefficients are not directly comparable across models.

```
. probit inlf nwifeinc educ exper expersq age kidslt6 kidsge6

Probit regression                                 Number of obs   =        753
                                                  LR chi2(7)      =     227.14
                                                  Prob > chi2     =     0.0000
Log likelihood = -401.30219                       Pseudo R2       =     0.2206
```

inlf	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nwifeinc	-.0120237	.0048398	-2.48	0.013	-.0215096	-.0025378
educ	.1309047	.0252542	5.18	0.000	.0814074	.180402
exper	.1233476	.0187164	6.59	0.000	.0866641	.1600311
expersq	-.0018871	.0006	-3.15	0.002	-.003063	-.0007111
age	-.0528527	.0084772	-6.23	0.000	-.0694678	-.0362376
kidslt6	-.8683285	.1185223	-7.33	0.000	-1.100628	-.636029
kidsge6	.036005	.0434768	0.83	0.408	-.049208	.1212179
_cons	.2700768	.508593	0.53	0.595	-.7267472	1.266901

```
. logit inlf nwifeinc educ exper expersq age kidslt6 kidsge6

Logistic regression                             Number of obs   =        753
                                                LR chi2(7)      =     226.22
                                                Prob > chi2     =     0.0000
Log likelihood = -401.76515                     Pseudo R2       =     0.2197
```

inlf	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nwifeinc	-.0213452	.0084214	-2.53	0.011	-.0378509	-.0048394
educ	.2211704	.0434396	5.09	0.000	.1360303	.3063105
exper	.2058695	.0320569	6.42	0.000	.1430391	.2686999
expersq	-.0031541	.0010161	-3.10	0.002	-.0051456	-.0011626
age	-.0880244	.014573	-6.04	0.000	-.116587	-.0594618
kidslt6	-1.443354	.2035849	-7.09	0.000	-1.842373	-1.044335
kidsge6	.0601122	.0747897	0.80	0.422	-.086473	.2066974
_cons	.4254524	.8603697	0.49	0.621	-1.260841	2.111746

We can use a rough rule of thumb to compare the estimates between models. In particular, we can divide the logit estimates by four and the probit estimates by 2.5 to make all estimates comparable to the LPM estimates. For example, for the coefficients on *kidslt6,* the scaled logit estimate is about -0.361 and the scaled probit estimate is about -0.347.

Why? If we evaluate the standard normal probability density function, $\emptyset(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)$, at the average values of the independent variables in the sample, we obtain about 0.391; this value is close enough to 0.4 to make the rule of thumb for scaling the probit coefficients useful in obtaining the effects on the response probability (in the case of the logit model we obtain a value of 0.243 which is close to 0.25). In other words, to estimate the change in the response probability given a one-unit increase in any independent variable, we multiply the corresponding coefficients by 0.4 in the probit model and by 0.25 in the logit model.

How do we formally get the marginal effects (ME) when using a probit or logit model? One easy way to get the marginal effects and their standard errors in Stata is using the postestimation command *mfx.* The command can compute several different types of MEs, evaluated at different values, for all regressors or a subset of regressors. The default is to evaluate at the sample mean $x = \bar{x}$. Note that for binary variables, *mfx* computes discrete changes of the variable from 0 to 1. Getting the marginal effects in the example above:

```
. quietly probit inlf nwifeinc educ exper expersq age kidslt6 kidsge6

. mfx

Marginal effects after probit
      y  = Pr(inlf) (predict)
         =  .58154201
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.]		X
nwifeinc	-.0046962	.00189	-2.48	0.013	-.008401	-.000991	20.129
educ	.0511287	.00986	5.19	0.000	.031805	.070452	12.2869
exper	.0481771	.00733	6.57	0.000	.033815	.062539	10.6308
expersq	-.0007371	.00023	-3.14	0.002	-.001197	-.000277	178.039
age	-.0206432	.00331	-6.24	0.000	-.027127	-.01416	42.5378
kidslt6	-.3391514	.04636	-7.32	0.000	-.430012	-.248291	.237716
kidsge6	.0140628	.01699	0.83	0.408	-.019228	.047353	1.35325

First note that the predicted probability of participating in the labor force is 0.5815 for a married woman at the average age of 42.5 years, with 12 years of education, 10 years of working experience, having one kid more than 6 years old and no kids less than 6 years old, and whose family nonwife income is $20,139 in average.

From the results we can see that one year more of education increases the probability of married woman of participating in the labor force by 0.05 and having one more kid less than 6 years old decreases the probability of married woman of participating in the labor force by 0.34.

The results are similar when using a logit model to estimate the same model above. The choice between logit and probit models will depend on which model performs better than the other.

```
. quietly logit inlf nwifeinc educ exper expersq age kidslt6 kidsge6

. mfx

Marginal effects after logit
      y  = Pr(inlf) (predict)
         =  .58277201
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.]		X
nwifeinc	-.0051901	.00205	-2.53	0.011	-.009204	-.001176	20.129
educ	.0537773	.01056	5.09	0.000	.033078	.074476	12.2869
exper	.0500569	.00782	6.40	0.000	.034721	.065393	10.6308
expersq	-.0007669	.00025	-3.10	0.002	-.001252	-.000281	178.039
age	-.021403	.00354	-6.05	0.000	-.028341	-.014465	42.5378
kidslt6	-.3509498	.04964	-7.07	0.000	-.448241	-.253658	.237716
kidsge6	.0146162	.01819	0.80	0.422	-.021032	.050265	1.35325

Multinomial response models: Multinomial logit

The logit model for binary outcomes can be extended to the case where the **unordered response** has more than two outcomes. Examples of unordered multinomial response include occupational choice, choice of health plan and transportation mode for commuting to work. In each case, an individual chooses one alternative from a group of choices and the labeling of the choices is arbitrary.

As in the binary response case, we are interested in how changes in **x** affect the response probabilities $\mathrm{P}(y = j|\mathbf{x}), j = 0,1,2, \dots, J$ holding everything else constant. Since the probabilities must sum up to unity, $\mathrm{P}(y = 0|\mathbf{x})$ is determined once we know the probabilities for $j = 1,2, \dots, J$. The coefficients are interpreted with respect to that category with is called the **base category**.

Let x be a 1xK vector with first-element unity. The **multinomial logit (MNL)** model has response probabilities:

$$\mathrm{P}(y = j|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta}_j) / \left[1 + \sum_{h=1}^{J} \exp(\mathbf{x}\boldsymbol{\beta}_h)\right], \quad j = 1,2, \dots, J \tag{9}$$

where $\boldsymbol{\beta}_j$ is Kx1 , $j = 1,2, \dots, J$.

When *J*=1, the model reduces to the binary logit model. The multinomial logit model is estimated by maximum likelihood.

EXAMPLE 2 (School and Employment Decisions for Young Men) The data KEANE.RAW contains employment and schooling history for a sample of men for the years 1981 to 1987. We will use the data for the year 1987. The three possible outcomes are enrolled in school (*status*=1), not in school and not working (*status*=2), and working (*status*=3). The explanatory variables are education (*educ*), past work experience (*exper*), and a black race indicator (*black*). We estimate the following model:

$$status = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 black \tag{10}$$

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/keane

. keep if year==87
(10985 observations deleted)

. mlogit status educ exper expersq black, baseoutcome (1)
```

```
Multinomial logistic regression                 Number of obs   =       1717
                                                LR chi2(8)      =     583.72
                                                Prob > chi2     =     0.0000
Log likelihood = -907.85723                     Pseudo R2       =     0.2433
```

status	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1	(base outcome)					
2						
educ	-.6736313	.0698999	-9.64	0.000	-.8106325	-.53663
exper	-.1062149	.173282	-0.61	0.540	-.4458414	.2334116
expersq	-.0125152	.0252291	-0.50	0.620	-.0619633	.036933
black	.8130166	.3027231	2.69	0.007	.2196902	1.406343
_cons	10.27787	1.133336	9.07	0.000	8.056578	12.49917
3						
educ	-.3146573	.0651096	-4.83	0.000	-.4422699	-.1870448
exper	.8487367	.1569856	5.41	0.000	.5410507	1.156423
expersq	-.0773003	.0229217	-3.37	0.001	-.1222261	-.0323746
black	.3113612	.2815339	1.11	0.269	-.240435	.8631574
_cons	5.543798	1.086409	5.10	0.000	3.414475	7.673121

.

A positive coefficient from *mlogit* means that as the regressor increases, we are more likely to choose alternative j than alternative 1 (where alternative 1 the base category). Thus, another year of education reduces the odds of being at home than enrolled in school, while the odds of being at home than enrolled in school is higher for black men. The magnitudes of these coefficients are difficult to interpret. Instead, we can compute relative-risk ratios (odds ratio) fo a unit-change in each corresponding variable by taking exponentials of the multinomial logit coefficients $e^{\widehat{\beta}}$ or by specifying the option *rrr*:

```
. mlogit status educ exper expersq black, rrr baseoutcome (1)

Multinomial logistic regression                 Number of obs   =       1717
                                                LR chi2(8)      =     583.72
                                                Prob > chi2     =     0.0000
Log likelihood = -907.85723                     Pseudo R2       =     0.2433
```

status	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1	(base outcome)					
2						
educ	.5098538	.0356387	-9.64	0.000	.4445768	.5847154
exper	.8992314	.1558206	-0.61	0.540	.6402853	1.262901
expersq	.9875628	.0249153	-0.50	0.620	.9399174	1.037624
black	2.254699	.6825496	2.69	0.007	1.245691	4.081004
_cons	29082.01	32959.68	9.07	0.000	3154.477	268115.2
3						
educ	.730039	.0475326	-4.83	0.000	.6425762	.8294066
exper	2.336693	.3668271	5.41	0.000	1.717811	3.178543
expersq	.9256118	.0212166	-3.37	0.001	.8849483	.9681438
black	1.365282	.3843732	1.11	0.269	.7862857	2.370634
_cons	255.6471	277.7374	5.10	0.000	30.40098	2149.781

.

Thus, one unit increase in education leads to a relative odd ratio of working instead of attending school that is 0.73 times what the ratio was before the change. Also, for black people relative to non-black, the relative risk for working relative to stay in school would be expected to increase by a factor of 1.36, holding all other variables in the model constant.

Also, as in the bivariate case we could use the *mfx* post-estimation command. In this case we will obtain the change in probability of the *j* alternative rather relative to the other J-1 alternatives by specifying for which of the alternatives we are obtaining the marginal effect:

```
. mfx, predict (pr outcome(3))

Marginal effects after mlogit
      y  = Pr(status==3) (predict, pr outcome(3))
         =  .82625034
```

variable	dy/dx	Std. Err.	z	P>|z|	[	95% C.I.	]	X
educ	.034636	.00428	8.09	0.000	.026246	.043026		12.5492
exper	.1344291	.01354	9.93	0.000	.107887	.160971		3.4403
expersq	-.0096146	.00173	-5.57	0.000	-.013	-.006229		17.1992
black*	-.0548704	.02091	-2.62	0.009	-.095851	-.01389		.379732

(*) dy/dx is for discrete change of dummy variable from 0 to 1

One year more of education increases by 0.034 the probability of being working rather than being in school or being at home.

Censored models: Tobit and Selection models

Sometimes we have **incompletely observed data**. Causes of this incompleteness are truncation and censoring. In **truncated data**, some observations on both the dependent variable and the regressors are lost. For example, income may be the dependent variable and only low-income people are included in the sample. In **censored data**, information on the dependent variable is lost but not the data on the regressors. For example, people of all income levels may be included in the sample, but for confidentiality reasons the income of high-income people may be top-coded and reported only as exceeding, say, $100,000 per year.

Truncation involves greater information loss than does censoring.

Tobit model

Suppose that our data consists of $(y_i, \mathbf{x}_i)$, i=1,..., N. Assume that $\mathbf{x}_i$ is fully observed but y_i is not always observed. Specifically some y_i are zero. One interpretation is that zero is a censored observation. Suppose that a household has a latent (unobserved) demand for goods, denoted by y^*, and it is not expressed as a purchase until some known constant threshold, denoted by L, is achieved. We observe y^* only when $y^*>L$. Then the zero expenditure can be interpreted as a left-censored variable that equals zero when $y^* \leq L$. Thus, the observed sample consists of censored and uncensored observations. Observations can be left-censored or right-censored.

The regression of interest is specified as an unobserved **latent variable**, y^*:

$$y_i^* = \mathbf{x}_i'\beta + \varepsilon_i\,, \quad i=1,..., N \tag{11}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ and $\mathbf{x}_i$ denotes the Kx1 vector of exogenous and fully observed regressors. If y^* were observed, we would estimate (β, σ^2) by OLS in the usual way.

The observed variable y_i is related to the latent variable y^*, through the observation rule

$$y = \begin{cases} y^* & \text{if } y^* > L \\ L & \text{if } y^* \leq L \end{cases}$$

EXAMPLE 3 (Annual Hours Equation for Married Women) We use the MROZ.RAW dataset to estimate the annual hours equation for married women. Of the 753 women in the sample, 428 worked for a wage outside home during the year; 325 of the women worked zero hours. Thus, annual hours worked is a reasonable candidate to implement a Tobit model. The regressors are the same as in EXAMPLE 1 (age, education, experience, nonwife income in thousands, number of children less than six years of age, number of kids between 6 and 18).

Since this data is subject to left censoring at zero, the option *ll(0)* is required when estimating the model.

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/mroz

. tobit hours nwifeinc educ exper expersq age kidslt6 kidsge6, ll(0)

Tobit regression                                  Number of obs   =        753
                                                  LR chi2(7)      =     271.59
                                                  Prob > chi2     =     0.0000
Log likelihood = -3819.0946                       Pseudo R2       =     0.0343

------------------------------------------------------------------------------
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -8.814243   4.459096    -1.98   0.048    -17.56811   -.0603723
        educ |   80.64561   21.58322     3.74   0.000     38.27453    123.0167
       exper |   131.5643   17.27938     7.61   0.000     97.64231    165.4863
     expersq |  -1.864158   .5376615    -3.47   0.001    -2.919667   -.8086479
         age |  -54.40501   7.418496    -7.33   0.000    -68.96862    -39.8414
     kidslt6 |  -894.0217   111.8779    -7.99   0.000    -1113.655   -674.3887
     kidsge6 |    -16.218   38.64136    -0.42   0.675    -92.07675    59.64075
       _cons |   965.3053   446.4358     2.16   0.031     88.88529    1841.725
-------------+----------------------------------------------------------------
       /sigma |   1122.022   41.57903                      1040.396    1203.648
------------------------------------------------------------------------------

  Obs. summary:        325  left-censored observations at hours<=0
                       428     uncensored observations
                         0 right-censored observations
```

All the regressors aside from *kidsge6* are statistically significant at 0.05 level. Tobit regression coefficients are interpreted in the similar manner to OLS regression coefficients; however, the linear effect is on the uncensored latent variable, not the observed outcome. For example, for a one unit increase in education, there is an 80.64 point increase in the predicted value of hours, and having one more kid less than 6, is associated with an 894 unit decrease in the predicted value of hours.

Selection models

Sometimes we may have nonrandom samples or **selected samples.** Selection may be due to **self-selection,** where the outcome of interest determined in part by individual choice of whether or not to participate in the activity of interest. Selection may also result from **sample selection,** with those who participate in the activity of interest may be deliberately oversampled, an extreme case being sampling only participants. In either case, similar issues arise and it is recommended to use selection models (or sample selection models) to account for potential selection bias.

One of the leading examples in selection models are the **bivariate sample selection model.**

EXAMPLE 4 (Wage Offer Equation for Married Women) We use the data in MROZ.RAW to estimate a wage supply function for married women accounting for potential selectivity bias into the workforce.

Let y_2^* denote the outcome of interest, in this case wage. We introduce a second latent variable y_1^* which in this case represents the participation in the labor force. Thus, the outcome y_2^* is observed if $y_1^* > 0$.

The maximum likelihood estimation of the bivariate sample-selection model with the *heckman* command is straightforward.

```
. heckman lwage educ exper expersq, select (inlf=nwifeinc educ exper expersq age kidslt6 kidsge6) twostep

Heckman selection model -- two-step estimates   Number of obs      =       753
(regression model with sample selection)        Censored obs       =       325
                                                Uncensored obs     =       428

                                                Wald chi2(3)       =     51.53
                                                Prob > chi2        =    0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwage						
educ	.1090655	.015523	7.03	0.000	.0786411	.13949
exper	.0438873	.0162611	2.70	0.007	.0120163	.0757584
expersq	-.0008591	.0004389	-1.96	0.050	-.0017194	1.15e-06
_cons	-.5781033	.3050062	-1.90	0.058	-1.175904	.0196979
inlf						
nwifeinc	-.0120237	.0048398	-2.48	0.013	-.0215096	-.0025378
educ	.1309047	.0252542	5.18	0.000	.0814074	.180402
exper	.1233476	.0187164	6.59	0.000	.0866641	.1600311
expersq	-.0018871	.0006	-3.15	0.002	-.003063	-.0007111
age	-.0528527	.0084772	-6.23	0.000	-.0694678	-.0362376
kidslt6	-.8683285	.1185223	-7.33	0.000	-1.100628	-.636029
kidsge6	.036005	.0434768	0.83	0.408	-.049208	.1212179
_cons	.2700768	.508593	0.53	0.595	-.7267472	1.266901
mills						
lambda	.0322619	.1336246	0.24	0.809	-.2296376	.2941613
rho	0.04861					
sigma	.66362876					

The *select* option is used to specify the selection equation which in this example is the same labor force participation equation as in EXAMPLE 1. The coefficients in the wage equation are interpreted as if we observed wage data for all women in the sample.

For comparison, we present the OLS results below. The differences between the OLS and Heckman estimates are practically small as the "selectivity effect" (lambda) to correct for potential selection bias is not statistically significant.

```
. regress lwage educ exper expersq
```

Source	SS	df	MS	Number of obs =	428
Model	35.0223023	3	11.6741008	F(3, 424) =	26.29
Residual	188.305149	424	.444115917	Prob > F =	0.0000
				R-squared =	0.1568
Total	223.327451	427	.523015108	Adj R-squared =	0.1509
				Root MSE =	.66642

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1074896	.0141465	7.60	0.000	.0796837	.1352956
exper	.0415665	.0131752	3.15	0.002	.0156697	.0674633
expersq	-.0008112	.0003932	-2.06	0.040	-.0015841	-.0000382
_cons	-.5220407	.1986321	-2.63	0.009	-.9124668	-.1316145

EXERCISE 1 (Linear probability model and probit model)

Use the data in GROGGER.RAW ("use http://fmwww.bc.edu/ec-p/data/wooldridge/grogger") for this question.

a) Define a binary variable, say *arr86,* equal to unity if a man was arrested at least once during 1986, and zero otherwise. Estimate a linear probability model relating *arr86* to *pcnv*, *avgsen*, *tottime*, *inc86*, *black*, *hispan*, and *born60.* Report the usual heteroskedasticity-robust standard errors. What is the estimated effect on the probability of arrest if pcnv goes from 0.25 to 0.75?
b) Test the joint significance of *avgsen* and *tottime,* using a nonrobust and robust test.
c) Now estimate the model by probit. At the average values of *avgsen*, *tottime, inc86*, and *ptime86* in the sample, and with *black*=1, *hispan*=0 and *born60*=1, what is the estimated effect on the probability of *arresr* if *pcnv* goes from 0.25 to 0.75? Compare the results with the answer from part a)
d) For the probit model estimated in part c), obtain the percent correctly predicted. What is the percent correctly predicted when *narr86*=0? When *narr86*=1? What do you make of these findings?

EXERCISE 2 (Tobit model)

Use the data in FRINGE.RAW ("use http://fmwww.bc.edu/ec-p/data/wooldridge/fringe") for this question.

a) Estimate the linear model by OLS relating hrbens to *exper, age, educ, tenure, married, male, white, northeast, nrthcen, south*, and *union*.
b) Estimate a Tobit model relating the same variables from part a), Why do you suppose the OLS and Tobit estimates are so similar?
c) Add $exper^2$ and $tenure^2$ to the Tobit model from part b). Should these be included?
d) Are there significant differences in hourly benefits across industry, holding the other factors fixed?