ECN140: Handout 6

Takuya Ura

February 26, 2019

The objective of this handout is to consider the situation in which the dependent variable is binary (i.e., y takes either 0 or 1). It follows Ch.7-5 & 17 of Wooldridge's textbook, but it is slightly different. To illustrate the concepts, we are going to use MROZ.DTA.

When y is binary, it is adequate to model the conditional probability of y given x. That is,

$$P(y = 1 \mid x_1, \dots, x_k) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k), \tag{1}$$

where G is one of the following three functions

$$\begin{aligned} &\text{(LPM)} && G(z) = z \\ &\text{(Logit)} && G(z) = \frac{\exp(z)}{1+\exp(z)} \\ &\text{(Probit)} && G(z) = \Phi(z) \text{ where } \Phi \text{ is the standard normal cumulative distribution function.} \end{aligned}$$

7.5 Linear Probability Model

In the linear probability model, we have $G(z) = z$ in Eq. (1) so

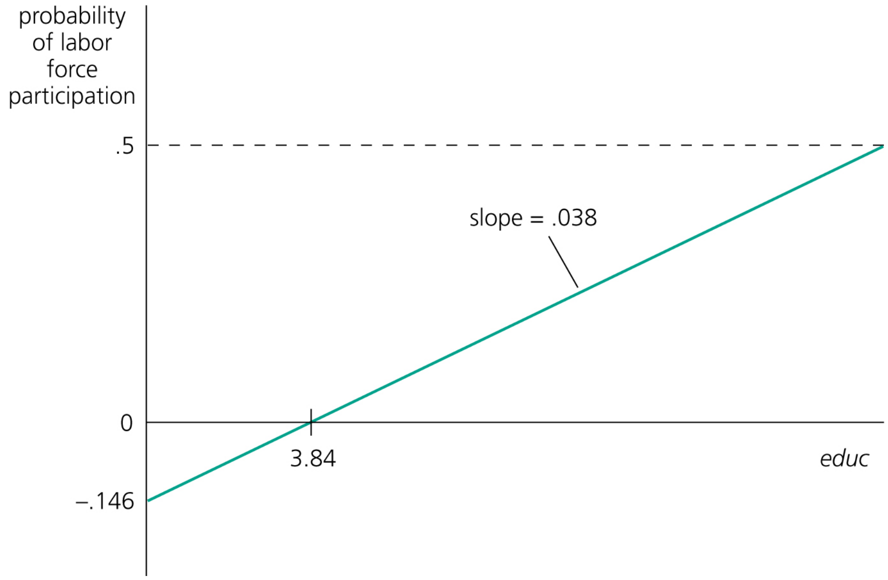
$$P(y = 1 \mid x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Here the interpretation of β_j is straightforward: β_j measures the change in the probability of $y = 1$ when x_j increases by one unit, holding other factors fixed. That is,

$$\Delta P(y = 1|x) = \beta_j \Delta x_j.$$

7.5.1 Shortcoming for the Linear Probability Model

The predicted probabilities may be less than 0 or greater than 1.



7.5.2 Heteroskedasticity for Regressing y on $x_1, \ldots, x_k$

The parameters, $\beta_0, \beta_1, \ldots, \beta_k$ are estimated by the OLS. Define

$$u = y - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k).$$

Then we can write down a linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u.$$

Furthermore, ZCM holds (i.e., $E[u \mid x_1, \ldots, x_k] = 0$), because

$$\begin{aligned}
E[u \mid x_1, \ldots, x_k] &= E[y - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \mid x_1, \ldots, x_k] \\
&= E[y \mid x_1, \ldots, x_k] - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \\
&= 1 \cdot P(y = 1 \mid x_1, \ldots, x_k) - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \\
&= (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \\
&= 0.
\end{aligned}$$

One remark is that the homoscedasticity assumption does not hold. It is because

$$\begin{aligned}
V(u \mid x_1, \ldots, x_k) &= V(y \mid x_1, \ldots, x_k) \\
&= E[y^2 \mid x_1, \ldots, x_k] - E[y \mid x_1, \ldots, x_k]^2 \\
&= P(y = 1 \mid x_1, \ldots, x_k) - P(y = 1 \mid x_1, \ldots, x_k)^2 \\
&= (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)^2,
\end{aligned}$$

and then $V(u \mid x_1, \ldots, x_k)$ depends on $(x_1, \ldots, x_k)$.

When the error is heteroskedastic, the usual formula for the OLS standard errors is wrong. We need to use "robust" option in STATA "regress."

17.1 Logit and Probit Models for Binary Respose

The logit/probit specifications force the predicted probability to be in $(0, 1)$.

In these models, the interpretation of β_j is not straightforward. Consider x_1 is a continuous variable. The partial effect of x_1 on $P(y = 1 \mid x_1, \ldots, x_k)$ holding the other independent variables fixed is captured by

$$g(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)\beta_1 \tag{2}$$

where g is the derivative of G, i.e., $g(z) = \frac{\partial}{\partial z} G(z)$.[1] Since $(x_1, \ldots, x_k)$ are random variables, we want to construct a summary statistic of Eq. (2). The sample mean of Eq. (2) is called the average partial effect:

$$\frac{1}{n}\sum_{i=1}^{n} g(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})\beta_1.$$

Stata's "margins, dydx(*)" produce this value for each independent variable.

The parameters, $\beta_0, \beta_1, \ldots, \beta_k$ are estimated by the maximum likelihood estimation. In the next three sections, we are going to see how to implement the MLE.

Binary Response Models with No Independent Variable

Before going to the general binary response models, we are going to assume that we have no independent variable. The probability of $y = 1$ is denoted by

$$\rho = P(y = 1).$$

(In this section, I will use ρ for the true parameter value and p for a generic value of the parameter.)

We do not know ρ, so we want to estimate it from a dataset. In this case, we are going to use the maximum likelihood estimation. For every parameter value p, the probability of $y = 0$ is evaluated as

$$1 - p$$

and the conditional probability of $y = 1$ is evaluated as

$$p.$$

For every observation i, we can observe the value of y and then we can evaluate the probability of that value being realized:

$$\begin{cases} 1 - p & \text{if } y_i = 0 \\ p & \text{if } y_i = 0. \end{cases}$$

More simply, for every parameter value p, we can evaluate the probability of y_i being realized as

$$p^{y_i}(1 - p)^{1 - y_i}.$$

[1] If x_j is a binary variable, the approximation via differentiation is not appropriate. The partial effect of x_1 on $P(y = 1 \mid x_1, \ldots, x_k)$ holding the other independent variables fixed is captured by

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \cdots + \beta_k x_k).$$

We observe the values of $\{y_i : i = 1, \ldots, n\}$, and, for every parameter value p, we can evaluate the probability of those values being realized as

$$L(p) = p^{y_1}(1-p)^{1-y_1} \times \cdots \times (p)^{y_n}(1-p)^{1-y_n}.$$

The function $L(p)$ is called the likelihood function. Taking the log of the likelihood function, then we have the new function

$$\mathscr{L}(p) = \sum_{i=1}^{n} \log\left(p^{y_i}(1-p)^{1-y_i}\right).$$

It is called the log-likelihood function. The maximum $\hat{\rho}$ of the (log-)likelihood function is called the maximum likelihood estimator.

In this simple case, we can show

$$\hat{\rho} = \frac{\text{(the number of observations with } y_i = 1)}{n}.$$

We will skip the proof but we demonstrate it by the first order condition. The log-likelihood function can be simplified to

$$\mathscr{L}(p) = \sum_{i=1}^{n} \left(y_i \log(p) + (1-y_i)\log(1-p)\right).$$

Taking the derivative, we have

$$\frac{\partial}{\partial p}\mathscr{L}(p) = \sum_{i=1}^{n}\left(y_i \frac{1}{p} + (1-y_i)\frac{-1}{1-p}\right). = \sum_{i=1}^{n} \frac{y_i - p}{p(1-p)} = \frac{\sum_{i=1}^{n}(y_i - p)}{p(1-p)} = \frac{n(\bar{y}-p)}{p(1-p)}.$$

The first order condition $\frac{\partial}{\partial p}\mathscr{L}(p) = 0$ implies

$$p = \bar{y}.$$

Binary Response Models with One Binary Independent Variable

Then, we are going to assume that we have only one independent variable and it is binary. The probability of $y = 1$ given $x = 0$ is denoted by

$$\rho_0 = P(y = 1 \mid x = 0)$$

and the probability of $y = 1$ given $x = 1$ is denoted by

$$\rho_1 = P(y = 1 \mid x = 1).$$

The difference $\rho_1 - \rho_0$ can be interpreted as the effect of x on $P(y = 1 \mid x)$.

We do not know ρ_0 and ρ_1, so we want to estimate them from a dataset. In this case, we are going to use the maximum likelihood estimation. For every parameter value (p_0, p_1), the conditional probability of $y = 0$ given x is evaluated as

$$\begin{cases} 1 - p_0 & \text{if } x = 0 \\ 1 - p_1 & \text{if } x = 1 \end{cases}$$

and the conditional probability of $y = 1$ given x is evaluated as

$$\begin{cases} p_0 & \text{if } x = 0 \\ p_1 & \text{if } x = 1. \end{cases}$$

For every observation i, we can observe the value of (y, x) and then we can evaluate the conditional probability of that value being realized:

$$\begin{cases} 1 - p_0 & \text{if } (y_i, x_i) = (0, 0) \\ p_0 & \text{if } (y_i, x_i) = (1, 0) \\ 1 - p_1 & \text{if } (y_i, x_i) = (0, 1) \\ p_1 & \text{if } (y_i, x_i) = (1, 1). \end{cases}$$

More simply, for every parameter value (p_0, p_1), we can evaluate the conditional probability of (y_i, x_i) being realized as

$$(p_{x_i})^{y_i}(1 - p_{x_i})^{1-y_i}.$$

We observe the values of $\{(y_i, x_i) : i = 1, \dots, n\}$, and, for every parameter value (p_0, p_1), we can evaluate the conditional probability of those values being realized as

$$L(p_0, p_1) = (p_{x_1})^{y_1}(1 - p_{x_1})^{1-y_1} \times \cdots \times (p_{x_n})^{y_n}(1 - p_{x_n})^{1-y_n}.$$

The function $L(p_0, p_1)$ is called the likelihood function. Taking the log of the likelihood function, then we have the new function

$$\mathscr{L}(p_0, p_1) = \sum_{i=1}^{n} \log\left((p_{x_i})^{y_i}(1 - p_{x_i})^{1-y_i}\right).$$

It is called the log-likelihood function. The maximum $(\hat{\rho}_0, \hat{\rho}_1)$ of the (log-)likelihood function is called the maximum likelihood estimator. In this case, it turns out that

$$\hat{\rho}_0 = \frac{(\text{the number of observations with } y_i = 1 \text{ and } x_i = 0)}{(\text{the number of observations with } x_i = 0)}$$

$$\hat{\rho}_1 = \frac{(\text{the number of observations with } y_i = 1 \text{ and } x_i = 1)}{(\text{the number of observations with } x_i = 1)}.$$

General Binary Response Models

The probability of $y = 1$ given $x_1, \dots, x_k$ is denoted by

$$G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k).$$

We are going to use the maximum likelihood estimation to estimate $(\beta_0, \dots, \beta_k)$. For every parameter value $(b_0, b_1, \dots, b_k)$, we can observe the i's value of (y, x) and then we can evaluate the conditional probability of that value being realized:

$$\begin{cases} 1 - G(b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}) & \text{if } y_i = 0 \\ G(b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}) & \text{if } y_i = 1. \end{cases}$$

More simply, for every parameter value $(b_0, b_1, \ldots, b_k)$, we can evaluate the conditional probability of (y_i, x_i) being realized as

$$G(b_0 + b_1 x_{i1} + \cdots + b_k x_{ik})^{y_i}(1 - G(b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}))^{1-y_i}.$$

We observe the values of $\{(y_i, x_i) : i = 1, \ldots, n\}$, and, for every parameter value $(b_0, b_1, \ldots, b_k)$, we can evaluate the conditional probability of those values being realized as

$$L(b_0, b_1, \ldots, b_k) = \prod_{i=1}^{n} \left(G(b_0 + b_1 x_{i1} + \cdots + b_k x_{ik})^{y_i}(1 - G(b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}))^{1-y_i} \right).$$

The function $L(b_0, b_1, \ldots, b_k)$ is called the likelihood function. Taking the log of the likelihood function, then we have the new function

$$\mathscr{L}(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n} \log \left(G(b_0 + b_1 x_{11} + \cdots + b_k x_{1k})^{y_1}(1 - G(b_0 + b_1 x_{11} + \cdots + b_k x_{1k}))^{1-y_1} \right).$$

It is called the log-likelihood function. The maximum $(\hat{\beta}_0, \ldots, \hat{\beta}_k)$ of the (log-)likelihood function is called the maximum likelihood estimator.

```
. use "MROZ.DTA"

. regress inlf nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6, robust

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

------------------------------------------------------------------------------
             |               Robust
        inlf |      Coef.   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
------------------------------------------------------------------------------

. margins, dydx(*)
```

```
Average marginal effects                        Number of obs     =        753
Model VCE    : Robust

Expression   : Linear prediction, predict()
dy/dx w.r.t. : nwifeinc educ exper age kidslt6 kidsge6

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   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 |   .0268138   .0024535    10.93   0.000     .0219973    .0316304
         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
------------------------------------------------------------------------------

.

. logit inlf nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6

Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -402.38502
Iteration 2:   log likelihood = -401.76569
Iteration 3:   log likelihood = -401.76515
Iteration 4:   log likelihood = -401.76515

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
------------------------------------------------------------------------------

. margins, dydx(*)

Average marginal effects                        Number of obs     =        753
Model VCE    : OIM

Expression   : Pr(inlf), predict()
dy/dx w.r.t. : nwifeinc educ exper age kidslt6 kidsge6

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0038118   .0014824    -2.57   0.010    -.0067172   -.0009064
        educ |   .0394965   .0072947     5.41   0.000     .0251992    .0537939
       exper |   .0254254   .0022364    11.37   0.000     .0210421    .0298088
         age |  -.0157194   .0023808    -6.60   0.000    -.0203856   -.0110532
     kidslt6 |  -.2577537   .0319416    -8.07   0.000    -.3203581   -.1951492
     kidsge6 |   .0107348    .013333     0.81   0.421    -.0153974    .0368671
------------------------------------------------------------------------------


.  probit inlf nwifeinc educ exper c.exper#c.exper age kidslt6 kidsge6

Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -402.06651
Iteration 2:   log likelihood = -401.30273
Iteration 3:   log likelihood = -401.30219
Iteration 4:   log likelihood = -401.30219

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
                |
c.exper#c.exper |  -.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    -.7267473    1.266901
---------------------------------------------------------------------------------

. margins, dydx(*)

Average marginal effects                        Number of obs     =        753
Model VCE    : OIM

Expression   : Pr(inlf), predict()
dy/dx w.r.t. : nwifeinc educ exper age kidslt6 kidsge6

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0036162   .0014414    -2.51   0.012    -.0064413   -.0007911
        educ |   .0393703   .0072216     5.45   0.000     .0252161    .0535244
       exper |   .0255825   .0022272    11.49   0.000     .0212172    .0299478
         age |  -.0158957   .0023587    -6.74   0.000    -.0205186   -.0112728
     kidslt6 |  -.2611542   .0318597    -8.20   0.000    -.3235982   -.1987103
     kidsge6 |   .0108287   .0130584     0.83   0.407    -.0147654    .0364227
------------------------------------------------------------------------------
```