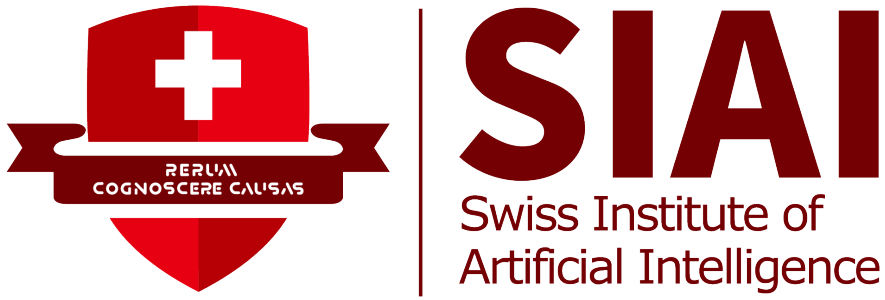
Data-based Decision Making
Lecture 5
Probability models



Keith Lee

October 7, 2021

Binary Dependent Variable: The Linear Probability Model

Introduction - Binary Choice Model

- Binary or dummy explanatory variables are often used in regressions.
- In binary choice models, we want to consider the use of dummy variable as the dependent variable.
  - We want to explain a qualitative outcome of $y = 0/1$ (=no/yes)
    - study the question of married women's labor force participation (in the labor force or out), or
    - whether a student is accepted to SIAI.
- How do we interpret population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

when y is binary? (also called "Boolean" for True/False values in computer science, named after George Boole, a mathematician)

  - y can only change between 0 and 1.
  - Suppose $\beta_1 = .035$ and $x_1 = educ$. What does it mean for one year increase in $educ$ to increase y by 0.35?

The Linear Probability Model I

- Recall that the standard linear regression model, $\mathbb{E}(u|x) = 0$, yields

$$\mathbb{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

  - We interpret β_j as

$$\Delta\mathbb{E}(y|\mathbf{x}) = \beta_j \Delta x_j \text{ holding other regressors fixed}$$

- With y taking only the values 0 and 1,

$$\begin{aligned}\mathbb{E}(y|\mathbf{x}) &= 0 \cdot Pr(y = 0|\mathbf{x}) + 1 \cdot Pr(y = 1|\mathbf{x}) \\ &= Pr(y = 1|\mathbf{x})\end{aligned}$$

  - We will therefore interpret β_j in this case as the effect a regressor has on the probability that $y = 1$, holding everything else fixed.

The Linear Probability Model II

- The use of the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \;\; with \;\; \mathbb{E}(u|\mathbf{x}) = 0$$

in the setting where y is binary (and simply apply OLS), is referred to as using a **linear probability model**

  - The reason is that it assumes $Pr(y = 1|\mathbf{x})$ is linearly related to the regressors:
  - In other words,

$$\begin{aligned} Pr(y = 1|\mathbf{x}) &= \mathbb{E}(y|\mathbf{x}) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \mathbb{E}(u|\mathbf{x}) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \;\; \leftarrow \;\; \text{linear in } \beta \end{aligned}$$

The Linear Probability Model III

- $Pr(y = 1|\mathbf{x}) = p(\mathbf{x})$ yields the **response probability** ($\mathbf{x}$ denotes all explanatory variables).
  - In binary choice models, our interest is to evaluate how explanatory variables can affect the response probability

$$Pr(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

- Conclude: In the LPM, the parameters have a nice interpretation

$$\frac{\Delta Pr(y|\mathbf{x})}{\Delta x_j} = \beta_j$$

  - "How does the probability change when x change, ceteris paribus?"
  - The partial (marginal) effects are constant (in real world?)
    - This means, e.g., in the labour participation example, every extra year of education affects the probability of participating in the labor market with the same amount.

The Linear Probability Model IV

- Using a sample, we can obtain OLS parameter estimates and the regression line

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + ... + \widehat{\beta}_k x_k,$$

  - $\widehat{y}$ is the **predicted probability** that $y = 1 : \widehat{Pr(y = 1|\mathbf{x})}$.
  - $\widehat{\beta}_0$ is the predicted probability when each x_j is set to zero (may not make sense).
  - $\widehat{\beta}_j$ measures the **change in the estimated probability** of $y = 1$ when $\Delta x_j = 1$ other factors held fixed.

EXAMPLE: Married Women's Labor Force Participation

The variable *inlf* is one if a woman worked for a wage during a certain year, and zero if not. (Note: What about married $= 0/1$ and woman $= 0/1$?)

The Linear Probability Model V

The estimated LPM: $n = 753$, $R^2 = .264$

$$\widehat{inlf} = \underset{(.152)}{.586} - \underset{(.0015)}{.0034}\,nwifeinc + \underset{(.007)}{.038}\,educ + \underset{(.006)}{.039}\,exper$$
$$- \underset{(.00019)}{.0006}\,exper^2 - \underset{(.002)}{.016}\,age - \underset{(.032)}{.262}\,kidslt6 + \underset{(.014)}{.013}\,kidsge6$$

Heteroskedasticity-robust standard errors in brackets

- Each year of education increases the probability by an estimated .038, or 3.8 percentage points.
- Having young children has a very large negative effect: being in the labor force falls by .262 for each young child. Reasonable?
- The coefficient on *nwifeinc* (other sources of income): modest effect
  - If it increases by 20 ($20,000, about one std deviation), the probability of being in the labor force falls by .068 (6.8 percentage points).
- Past workforce experience has a positive but diminishing effect.
  - Years of work experience have diminishing effect, as the squared term has negative coefficient

Shortcomings of the LPM

- While the LPM is convenient because estimation and interpretation is easy, it does have some shortcomings.
  1. The **fitted values** from an OLS regression **are never guaranteed to lie between zero and one**, yet they represent estimated probabilities.
  2. The **estimated partial effects are constant** ; may lead to silly estimated effects for large changes.
     - For example, take a woman who has no other source of income, 25 years of prior work experience, no children, who is 48 years old. As a function of *educ* the equation looks like

$$\widehat{inlf} = .417 + .038educ$$

     - At $educ = 12$, the predicted probability is .873, at $educ = 14$ it is .949, and at $educ = 16$, $\widehat{inlf} = 1.025 > 1$.
     - For the estimated model to truly represent a probability, the effect of education should be diminishing [Quadratics are typically limited]
  3. The LPM exhibits **heteroskedasticity** - A4 violation (not efficient)

Shortcomings of the LPM - technical details (Optional for MBA)

- Why does the LPM exhibit heteroskedasticity?
  - Recall with y a binary (0,1) - variable

$$\mathbb{E}(y|\mathbf{x}) = 0 \cdot Pr(y = 0|\mathbf{x}) + 1 \cdot Pr(y = 1|\mathbf{x}) = p(x)$$

- What can we say about $Var(y|\mathbf{x})$?
  - Recall: $Var(y|x) = \mathbb{E}(y^2|\mathbf{x}) - [\mathbb{E}(y|\mathbf{x})]^2$
  - $\mathbb{E}(y^2|\mathbf{x}) = 0^2 \cdot Pr(y = 0|\mathbf{x}) + 1^2 \cdot Pr(y = 1|\mathbf{x}) = p(\mathbf{x})$
- Hence

$$Var(y|\mathbf{x}) = p(\mathbf{x}) \cdot (1 - p(\mathbf{x}))$$

  - $Var(y|\mathbf{x}) = Var(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u|\mathbf{x}) = Var(u|\mathbf{x})$?
  - Unless $\beta_1 = ... = \beta_k$, we get heteroskedasticity, as the variance of y depends on $\mathbf{x}$! (related to over-confidence in Machine Learning)
  - Thus, OLS will not be BLUE

Logit and Probit Models for Binary Choice

Binary Choice: Functional Specification I

- Rather than assuming $p(\mathbf{x})$ is linear, we may prefer instead to model the probability directly as

$$Pr(y = 1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)$$

for **some function G that takes values between zero and one.**

  - A natural choice for $G(\cdot)$ is to use a **cumulative distribution function.**
  - When $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ is large, the probability of $y = 1$ is close to one.
    - Two most used cases are

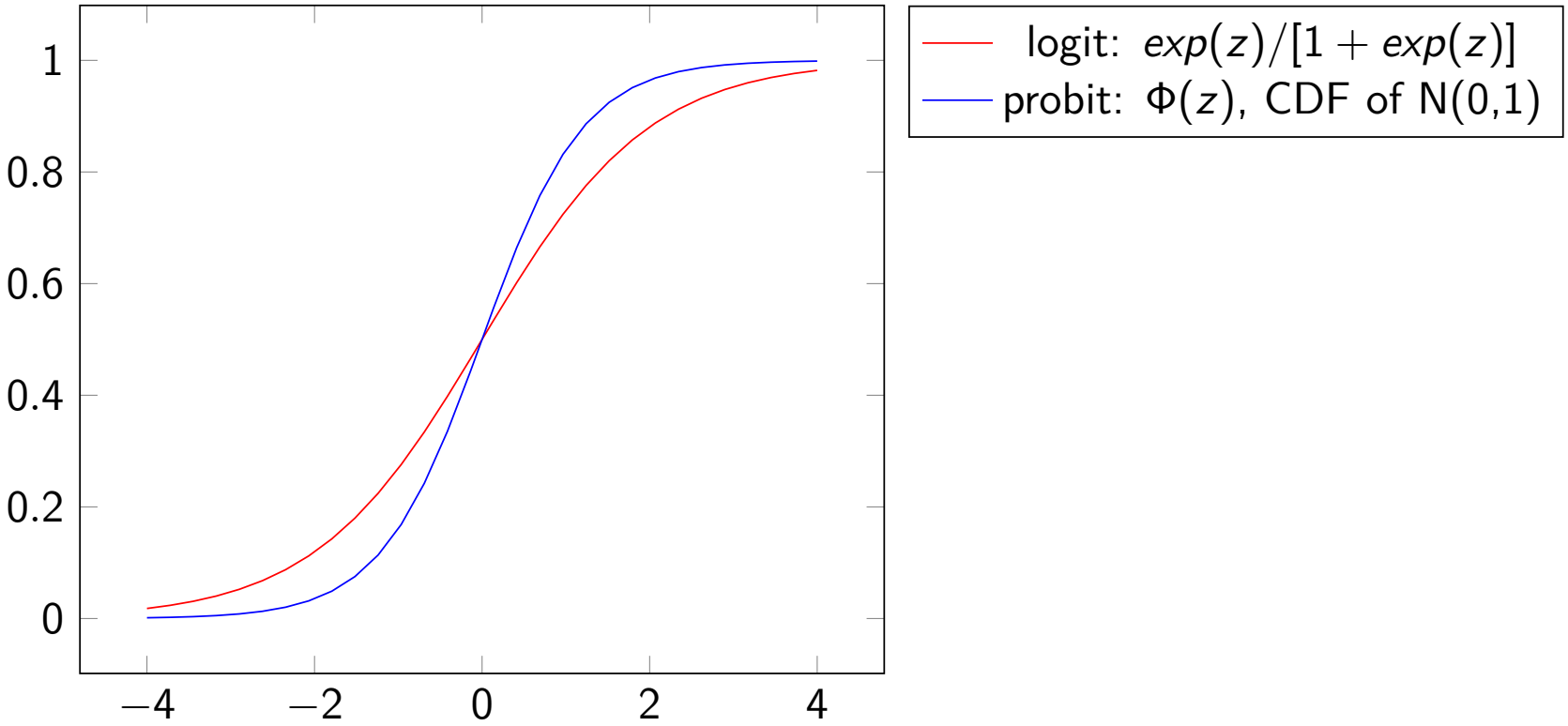
$$G(z) = Logistic\ CDF = \Lambda(z) = \frac{exp(z)}{[1 + exp(z)]}\ \textbf{(logit)}$$

$$G(z) = Normal(0,1)\ CDF = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}u^2)du\ \textbf{(probit)}$$

Binary Choice: Functional Specification II

- Both G functions have similar shapes but the logistic is more spread out.

Probit and Logit Response Functions



Binary Choice: Functional Specification III

- We may consider the following, non-linear, regression model

$$y = G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k) + u$$

where $\mathbb{E}(u|\mathbf{x}) = 0$.

  - We could then minimize the residual sum of squares (non-linear)

$$\min_b \sum_{i=1}^{n} (y_i - G(b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik}))^2$$

  - Predicted values will always lie between zero and one.

$$\widehat{y}_i = G(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + ... + \widehat{\beta}_k x_{ik})$$

  - Would **not be efficient**, as we still have the **heteroskedasticity** in the model - variance changes as x changes

$$Var(y|\mathbf{x}) = G(\mathbf{x})(1 - G(\mathbf{x})) \equiv Pr(y = 1|\mathbf{x})(1 - Pr(y = 1|\mathbf{x}))$$

Logit and Probit - Empirical example I

EXAMPLE: Married Women's Labor Force Participation The variable *inlf* is one if a woman worked for a wage during a certain year, and zero if not.

```
. probit inlf nwifeinc educ exper expersq 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
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	-.7267473	1.266901

Logit and Probit - Estimating Partial Effects I

- Important: The parameters estimates provided by probit/logit, $\widehat{\beta}$, are not the partial effects.
- Recall: In binary choice models, the partial effect should explain how each explanatory variable affect the probability that $y = 1$ holding everything else constant
  - In LPM, where we specified $Pr(y = 1|\mathbf{x})$ linearly

$$Pr(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

β_j does denotes the partial effect of interest.

Logit and Probit - Estimating Partial Effects II

- In probit/logit model, we specified $Pr(y = 1|\mathbf{x})$ non-linearly (ensuring restricted between 0 and 1)

$$Pr(y = 1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)$$

  - For **continuous explanatory variables**, the partial effect is given by

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = \frac{\partial G(\mathbf{x}\beta)}{\partial x_j} = \beta_j g(\mathbf{x}\beta) \quad (chain - rule)$$

  where $\mathbf{x}\beta = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ and $g(z) = dG(z)/dz$

  - For **explanatory variables that are dummy variables**, the partial effect evaluates the difference in probability of participation when the dummy variable switches from 0 to 1. Say x_1 is a dummy variable, then

$$\begin{aligned}\frac{\Delta Pr(y = 1|x)}{\Delta x_1} =& G(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k) \\ & - G(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)\end{aligned}$$

Logit and Probit - Estimating Partial Effects III

- Unlike in the LPM β_j **does not have an easy interpretation**.
  - The **sign of β_j does tell us whether the partial effect is positive or negative** (because $g(z) > 0$), but the magnitude of the partial effect depends on $g(\mathbf{x}\beta)$.

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = \frac{\partial G(\mathbf{x}\beta)}{\partial x_j} = \beta_j g(\mathbf{x}\beta)$$

- **The partial effect is not constant! Depends on x.**
  - For reporting a partial effect, we consider
    1. Partial effect of the average individual (PEA)
    2. Average partial effect of all individuals (APE)
    3. Partial effect of an individual with specific characteristics

Logit and Probit - Estimating Partial Effects IV

- Partial effect of the average individual (PEA)

$$\widehat{PEA}_j = g(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + ... + \hat{\beta}_k\bar{x}_k) \cdot \hat{\beta}_j$$

where $\bar{x}_j$ denotes the average of the j^{th} explanatory variable.

- Average partial effect of all individuals (APE)

$$\widehat{APE}_j = \frac{1}{n}\sum_{i=1}^{n} g(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_{i1} + \hat{\beta}_2\bar{x}_{i2} + ... + \hat{\beta}_k\bar{x}_{ik}) \cdot \hat{\beta}_j$$

- Partial effect of an individual with specific characteristics

$$g(\hat{\beta}_0 + \hat{\beta}_1\tilde{x}_1 + \hat{\beta}_2\tilde{x}_2 + ... + \hat{\beta}_k\tilde{x}_k) \cdot \hat{\beta}_j$$

where $\tilde{x}_j$ denotes a particular value for the j^{th} explanatory variable.

Logit and Probit - Estimating Partial Effects V

- If x_1 is a dummy variable, we can estimate our partial effects using:
-

$$\begin{aligned}\widehat{PEA}_1 =& G(\hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 \bar{x}_2 + ... + \hat{\beta}_k \bar{x}_k) \\ & - G(\hat{\beta}_0 + \hat{\beta}_1 \times 0 + \hat{\beta}_2 \bar{x}_2 + ... + \hat{\beta}_k \bar{x}_k)\end{aligned}$$

-

$$\begin{aligned}\widehat{APE}_1 =& \frac{1}{n} \sum_{i=1}^{n} G(\hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik}) \\ & - G(\hat{\beta}_0 + \hat{\beta}_1 \times 0 + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik})\end{aligned}$$

  - The APE represents an average treatment effect (ATE). (The "treatment", x1, is binary.)

Logit and Probit - Estimating Partial Effects VI

- Evaluating marginal effects for the average person has the following potential problems.
  - If some explanatory variables are discrete, the averages of them represent no one in the sample (even population)
    - E.g., consider dummy variable rural (60% of our sample is rural). What sense does it make use of 0.6 for $\overline{rural}$?
  - If a continuous explanatory variable appears as a nonlinear function, how should the averages be obtained?
    - E.g., consider variable $log(sales)$. Should we use $\overline{log(sales)}$ or $log(\overline{sales})$?

Logit and Probit - Empirical example I

- The Partial Effect of the Average individual (PEA)

$$\widehat{APE}_j = g(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + ... + \hat{\beta}_k\bar{x}_k) \cdot \hat{\beta}_j$$

```
. margins, dydx(*) atmeans

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

Expression   : Pr(inlf), predict()
dy/dx w.r.t. : nwifeinc educ exper expersq age kidslt6 kidsge6
at           : nwifeinc        =    20.12896 (mean)
               educ            =    12.28685 (mean)
               exper           =    10.63081 (mean)
               expersq         =    178.0385 (mean)
               age             =    42.53785 (mean)
               kidslt6         =    .2377158 (mean)
               kidsge6         =    1.353254 (mean)
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
nwifeinc	-.0046962	.0018903	-2.48	0.013	-.0084012	-.0009913
educ	.0511287	.0098592	5.19	0.000	.0318051	.0704523
exper	.0481771	.0073278	6.57	0.000	.0338149	.0625392
expersq	-.0007371	.0002347	-3.14	0.002	-.001197	-.0002771
age	-.0206432	.0033079	-6.24	0.000	-.0271265	-.0141598
kidslt6	-.3391514	.0463581	-7.32	0.000	-.4300117	-.2482911
kidsge6	.0140628	.0169852	0.83	0.408	-.0192275	.0473531

Logit and Probit - Empirical example II

- The Average Partial Effect of all the individuals (APE)

$$\widehat{APE_j} = \frac{1}{N}\sum_{i=1}^{N} g(\widehat{\beta}_0 + x_{i1}\widehat{\beta}_1 + x_{i2}\widehat{\beta}_2 + ... + x_{ik}\widehat{\beta}_k) \cdot \widehat{\beta}_j$$

```
. margins, dydx(*)

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

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

	dy/dx	Delta-method 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	.0370974	.0051522	7.20	0.000	.0269993	.0471956
expersq	-.0005675	.0001771	-3.20	0.001	-.0009146	-.0002204
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

Logit and Probit - Empirical example III

- Recall, in Linear Probability Model (LPM)

$$\widehat{inlf} = .586 - .0034 nwifeinc + .038 educ + ... - .262 kidslt6 + ..$$

  - For every additional young child, labor force (LF) participation decreases with 26.2 percentage points.
  - Reasonable?
- As discused, in probit/logit model, the marginal effects are not constant
- Indeed, they do permit the largest effect on LF participation to be associated with first child.

Logit and Probit - Empirical example IV

- Compare the predicted LF participation for the average person with different numbers of young children:

$$
\begin{aligned}
&Pr(infl = 1|\bar{x}, kidslt6 = 0) \\
&\quad = \Phi(.27 - .012\overline{nwifeinc} + .131\overline{educ} + \ldots - .868 \cdot 0 + ..) = .707 \\
&Pr(infl = 1|\bar{x}, kidslt6 = 1) \\
&\quad = \Phi(.27 - .012\overline{nwifeinc} + .131\overline{educ} + \ldots - .868 \cdot 1 + ..) = .373 \\
&Pr(infl = 1|\bar{x}, kidslt6 = 2) \\
&\quad = \Phi(.27 - .012\overline{nwifeinc} + .131\overline{educ} + \ldots - .868 \cdot 2 + ..) = .117
\end{aligned}
$$

  - The first young child reduces the LF participation with 33.4 percentage points.
    - $\widehat{Pr}(infl = 1|\bar{x}, kidslt6 = 1) - \widehat{Pr}(infl = 1|\bar{x}, kidslt6 = 0) = -.334$
  - The second young child reduces the LF partipation with 25.6 percentage points.
    - $\widehat{Pr}(infl = 1|\bar{x}, kidslt6 = 2) - \widehat{Pr}(infl = 1|\bar{x}, kidslt6 = 1) = -.256$
- Effect is not constant: The biggest effect of having young children is the first one!

Hypothesis Testing: Probit/Logit Model

Testing single linear restriction: z-test

- How do we test a **null hypothesis**?
  - Example: $H_0 : \beta_2 = 1$, against $H_A : \beta_2 \neq 1$
  - Your regression output provides parameter estimates $(\widehat{\beta}_1, \widehat{\beta}_2, ...)$ and their SE's.
  - Test statistic:

$$z = \frac{\widehat{\beta}_2 - 1}{SE(\widehat{\beta}_2)} \overset{a}{\sim} N(0,1) \ \textit{under} \ H_0$$

  - Reject H_0 if $|z| > 1.96$ at the 5% level of significance.

Testing multiple linear restriction: F-test

- How do we test multiple linear restrictions?
  - Example: $H_0 : \beta_2 = 0$ and $\beta_3 = 0$, against $H_A : \beta_2 \neq 0$ and/or $\beta_3 \neq 0$
  - OLS: Recall we used the F test for multiple linear restrictions
    - In this test, you compared the restricted residual sum of squares ($RRSS$) with the unrestricted residual sum of squares ($URSS$) (or equivalently R^2_R with R^2_{UR}).
    - Made sense, because the OLS attempts to minimize the residual sum of squares (test loss of fit)
  - Goes back to F-test

Example "Chow" Test

- Extension: How can we test whether the labor market participation for women is the same for urban ($city = 1$) as it is for rural women ($city = 0$).
- We want to test:

$$H_0 : \beta_j^{rural} = \beta_j^{urban} \text{ for all } j = 0, 1, ..., k$$
$$H_1 : \text{At least one } \beta_j^{rural} \neq \beta_j^{urban}$$

  - We have k + 1 restrictions we want to test (intercept + slopes)
  - F-test stat

$$F = \frac{(RRSS - URSS)/(k+1)}{URSS/(n-2(k+1))}$$

  - Reject H_0 if F-test stat crosses the critical value

Example "Chow" Test

- Restricted model
  - The LF participation decision is the same for urban and rural women
  - To obtain *RRSS* , we simply perform probit (logit) using all observations
- Unrestricted model
  - The LF participation for rural and urban women are different
  - To obtain *URSS* we can run separate probit (logit) regressions for the urban and rural sample
  - From here we compute $URSS = RSS_{urban} + RSS_{rural}$