ExpEcon Methods:
Binary Dependent Variables: Logit & Probit

ECON 8877

P.J. Healy

First version thanks to Zexin Ye

Updated 2026-01-28

Discrete Choice Model: Overview

The Choice Set

- The set of options that are available to the decision maker.

Derivation

- Define choice probabilities and derive them from the utility-maximizing behavior.
- Derive logit model.

Discrete Choice Model: The Choice Set

Discrete choice models describe decision makers' choices among alternatives.

The set of alternatives, called the choice set, needs to exhibit three characteristics.

- The alternatives must be mutually exclusive
  - Choosing one alternative necessarily implies not choosing any of the other alternatives.
- The choice set must be exhaustive
  - All possible alternatives are included.
- The number of alternatives must be finite

Discrete Choice Model: The Choice Set

The third condition, namely, that the number of alternatives is finite, is actually restrictive.

Main difference from regression models.

With regression models, the dependent variable is continuous - an infinite number of possible outcomes.

Discrete Choice Model: Derivation of Choice Probabilities

Discrete choice models are usually derived under an assumption of utility-maximizing behavior by the decision maker.

- Thurstone (1927) originally developed the concepts in terms of psychological stimuli, leading to a binary probit model of whether respondents can differentiate the level of stimulus.
- Marschak (1960) interpreted the stimuli as utility and provided a derivation from utility maximization.

Following Marschak, models that can be derived in this way are called random utility models (RUMs).

Discrete Choice Model: RUMs

Random utility models (RUMs) are derived as follows.

- A decision maker, labeled n, faces a choice among J alternatives.
- The decision maker would obtain a certain level of utility (or profit) from each alternative.
- The utility that decision maker n obtains from alternative j is $U_{nj}, \quad j = 1, \ldots, J$.
- This utility is known to the decision maker but not by the researcher.
- The decision maker chooses the alternative that provides the greatest utility.
- The behavioral model is therefore: decision maker n chooses alternative i if and only if $U_{ni} > U_{nj}, \forall j \neq i$.

Discrete Choice Model: RUMs

Consider now the researcher.

- The researcher does not observe the decision maker's utility.
- The researcher observes
  - some attributes of the alternatives as faced by the decision maker, labeled $x_{nj}, \forall j$,
  - some attributes of the decision maker, labeled s_n,
- The researcher can specify a function that relates these observed factors to the decision maker's utility.
- The function is denoted $V_{nj} = V\left(x_{nj}, s_n\right), \forall j$ and is often called representative utility.

Discrete Choice Model: RUMs

- Since there are aspects of utility that the researcher does not or cannot observe, $V_{nj} \neq U_{nj}$.
- Utility is decomposed as $U_{nj} = V_{nj} + \varepsilon_{nj}$, where ε_{nj} captures the factors that affect utility but are not included in V_{nj}.
- The distribution of ε_{nj} depends critically on the researcher's specification of V_{nj}.

Discrete Choice Model: RUMs

- The researcher does not know ε_{nj} $\forall j$ and therefore treats these terms as random.
- The joint density of the random vector $\varepsilon_n' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle$ is denoted $f(\varepsilon_n)$.
- With this density, the researcher can make probabilistic statements about the decision maker's choice.
- The probability that decision maker n chooses alternative i is

$$\begin{aligned} P_{ni} &= \text{Prob}\left(U_{ni} > U_{nj} \ \forall j \neq i\right) \\ &= \text{Prob}\left(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \forall j \neq i\right) \\ &= \text{Prob}\left(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \ \forall j \neq i\right). \end{aligned}$$

Discrete Choice Model: RUMs

Using the density $f(\varepsilon_n)$, this cumulative probability can be rewritten as

$$\begin{aligned} P_{ni} &= \text{Prob}\left(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \; \forall j \neq i\right) \\ &= \int_{\varepsilon} I\left(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \; \forall j \neq i\right) f(\varepsilon_n) \, d\varepsilon_n, \end{aligned}$$

where $I(\cdot)$ is the indicator function, equaling 1 when the expression in parentheses is true and 0 otherwise.

This is a multidimensional integral over the density of the unobserved portion of utility, $f(\varepsilon_n)$.

Discrete Choice Model: RUMs

- Different discrete choice models are obtained from different specifications of this density $f(\varepsilon_n)$, that is, from different assumptions about the distribution of the unobserved portion of utility.
- The integral takes a closed form only for certain specifications of $f(\cdot)$.
- Logit has closed-form expressions for this integral. They are derived under the assumption that the unobserved portion of utility is distributed iid extreme value.

Discrete Choice Model: Identification of Choice Models

Only Differences in Utility Matter

- The choice probability is $P_{ni} = \text{Prob}\left(U_{ni} > U_{nj}\ \forall j \neq i\right) = \text{Prob}\left(U_{ni} - U_{nj} > 0\ \forall j \neq i\right)$, which depends only on the difference in utility, not its absolute level.
- Adding a constant to the utility of all alternatives does not change the decision maker's choice.

The Overall Scale of Utility Is Irrelevant

- The model $U^0_{nj} = V_{nj} + \varepsilon_{nj}\ \forall j$ is equivalent to $U^1_{nj} = \lambda V_{nj} + \lambda\varepsilon_{nj}\ \forall j$ for any $\lambda > 0$.
- Multiplying each alternative's utility by a constant also does not change the decision maker's choice.

Discrete Choice Model: Logit

- By far the easiest and most widely used discrete choice model is logit. Its popularity is due to the fact that the formula for the choice probabilities takes a closed form and is readily interpretable.
- Originally, the logit formula was derived by Luce (1959) from assumptions about the characteristics of choice probabilities, namely the independence from irrelevant alternatives (IIA).
- Marschak (1960) showed that these axioms implied that the model is consistent with utility maximization.
- The relation of the logit formula to the distribution of unobserved utility was developed by Marley, as cited by Luce and Suppes (1965), who showed that the extreme value distribution leads to the logit formula.
- McFadden (1974) completed the analysis by showing the converse: that the logit formula for the choice probabilities necessarily implies that unobserved utility is distributed extreme value.

Discrete Choice Model: Logit

- A decision maker, labeled n, faces J alternatives.
- The utility that the decision maker obtains from alternative j is decomposed as $U_{nj} = V_{nj} + \varepsilon_{nj}\ \forall j$.
- The logit model is obtained by assuming that each ε_{nj} is independently, identically distributed extreme value. The distribution is also called Gumbel and type I extreme value. The density for each unobserved component of utility is

$$f\left(\varepsilon_{nj}\right) = e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}}$$

- The cumulative distribution is

$$F\left(\varepsilon_{nj}\right) = e^{-e^{-\varepsilon_{nj}}}$$

Discrete Choice Model: Logit

- The variance of this distribution is $\pi^2/6$. By assuming the variance is $\pi^2/6$, we are implicitly normalizing the scale of utility.
- The mean of the extreme value distribution is not zero; however, the mean is immaterial, since only differences in utility matter, and the difference between two random terms that have the same mean has itself a mean of zero.
- The difference between two extreme value variables is distributed logistic. That is, if ε_{nj} and ε_{ni} are iid extreme value, then $\varepsilon^*_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$ follows the logistic distribution

$$F\left(\varepsilon^*_{nji}\right) = \frac{e^{\varepsilon^*_{nji}}}{1 + e^{\varepsilon^*_{nji}}}$$

Discrete Choice Model: Logit

We now derive the logit choice probabilities, following McFadden (1974). The probability that decision maker n chooses alternative i is

$$\begin{aligned} P_{ni} &= \text{Prob}\left(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \; \forall j \neq i\right) \\ &= \text{Prob}\left(\varepsilon_{nj} < \varepsilon_{ni} + V_{ni} - V_{nj} \; \forall j \neq i\right). \end{aligned}$$

If ε_{ni} is considered given, this expression is the cumulative distribution for each ε_{nj} evaluated at $\varepsilon_{ni} + V_{ni} - V_{nj}$, which is $\exp\left(-\exp\left(-\left(\varepsilon_{ni} + V_{ni} - V_{nj}\right)\right)\right)$.

Since the ε 's are independent, this cumulative distribution over all $j \neq i$ is the product of the individual cumulative distributions:

$$P_{ni} \mid \varepsilon_{ni} = \prod_{j \neq i} e^{-e^{-\left(\varepsilon_{ni} + V_{ni} - V_{nj}\right)}}$$

Discrete Choice Model: Logit

Of course, ε_{ni} is not given, and so the choice probability is the integral of $P_{ni} \mid \varepsilon_{ni}$ over all values of ε_{ni} weighted by its density:

$$P_{ni} = \int \left(\prod_{j \neq i} e^{-e^{-\left(\varepsilon_{ni} + V_{ni} - V_{nj}\right)}} \right) e^{-\varepsilon_{ni}} e^{-e^{-\varepsilon_{ni}}} d\varepsilon_{ni}.$$

Some algebraic manipulation of this integral results in a succinct, closed form expression:

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}$$

Discrete Choice Model: Logit

Representative utility is usually specified to be linear in parameters: $V_{nj} = \beta' x_{nj}$, where x_{nj} is a vector of observed variables relating to alternative j. With this specification, the logit probabilities become

$$P_{ni} = \frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}}.$$

Under fairly general conditions, any function can be approximated arbitrarily closely by one that is linear in parameters.

Importantly, McFadden (1974) demonstrated that the log-likelihood function with these choice probabilities is globally concave in parameters β, which helps in the numerical maximization procedures.

Discrete Choice Model: Logit

The logit probabilities exhibit several desirable properties.

- First, P_{ni} is necessarily between zero and one, as required for a probability.
- Second, the choice probabilities for all alternatives sum to one: $\sum_{i=1}^{J} P_{ni} = \sum_i \exp(V_{ni}) / \sum_j \exp(V_{nj}) = 1$.
- Third, the relation of the logit probability to representative utility is sigmoid, or S-shaped, as shown in Figure below. This shape has implications for the impact of changes in explanatory variables.

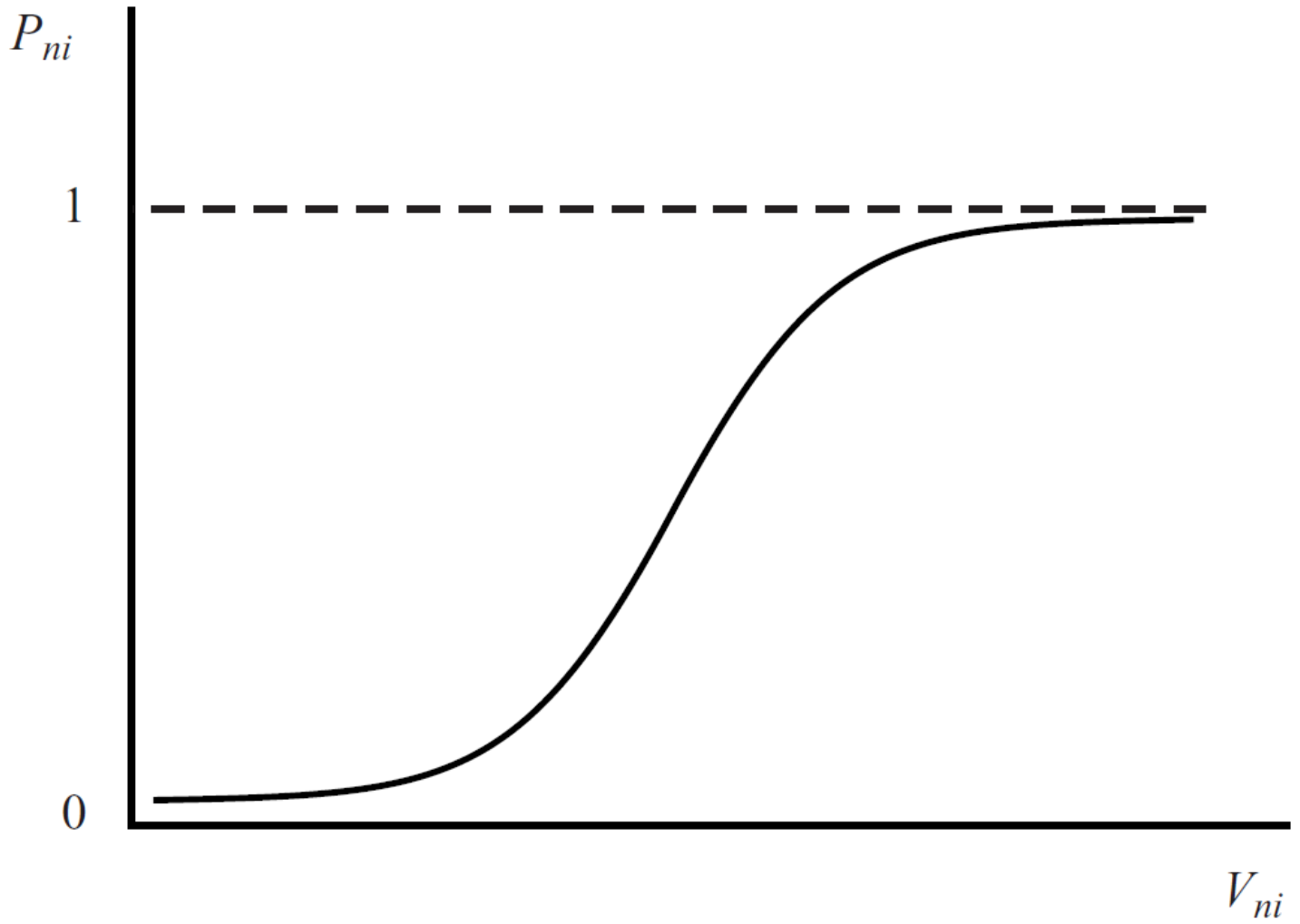


Figure 3.1. Graph of logit curve.

Discrete Choice Model: Logit

The logit model exhibits independence from irrelevant alternatives, or IIA.

- For any two alternatives i and k, the ratio of the logit probabilities is

$$\frac{P_{ni}}{P_{nk}} = \frac{e^{V_{ni}} / \sum_j e^{V_{nj}}}{e^{V_{nk}} / \sum_j e^{V_{nj}}}$$
$$= \frac{e^{V_{ni}}}{e^{V_{nk}}} = e^{V_{ni} - V_{nk}}$$

- This ratio does not depend on any alternatives other than i and k. That is, the relative odds of choosing i over k are the same no matter what other alternatives are available or what the attributes of the other alternatives are.
- Since the ratio is independent from alternatives other than i and k, it is said to be independent from irrelevant alternatives.

Logit: Estimation

- Consider first the situation in which the sample is exogenously drawn, that is, is either random or stratified random with the strata defined on factors that are exogenous to the choice being analyzed.
- We also assume that the explanatory variables are exogenous to the choice situation. That is, the variables entering representative utility are independent of the unobserved component of utility.
- A sample of N decision makers is obtained for the purpose of estimation. Since the logit probabilities take a closed form, the traditional maximum-likelihood procedures can be applied.

Logit: Estimation

- The probability of person n choosing the alternative that he was actually observed to choose can be expressed as

$$\prod_i (P_{ni})^{y_{ni}}$$

  where $y_{ni} = 1$ if person n chose i and zero otherwise. Note that since $y_{ni} = 0$ for all nonchosen alternatives and P_{ni} raised to the power of zero is 1, this term is simply the probability of the chosen alternative.
- Assuming that each decision maker's choice is independent of that of other decision makers, the probability of each person in the sample choosing the alternative that he was observed actually to choose is

$$L(\beta) = \prod_{n=1}^{N} \prod_i (P_{ni})^{y_{ni}},$$

Logit: Estimation

- The loglikelihood function is then

$$\mathrm{LL}(\beta) = \sum_{n=1}^{N} \sum_{i} y_{ni} \ln P_{ni}$$

- and the estimator is the value of β that maximizes this function.
- McFadden (1974) shows that $\mathrm{LL}(\beta)$ is globally concave for linear-inparameters utility, and many statistical packages are available for estimation of these models.
- At the maximum of the likelihood function, its derivative with respect to each of the parameters is zero:

$$\frac{d\mathrm{LL}(\beta)}{d\beta} = 0$$

The maximum likelihood estimates are therefore the values of β that satisfy this first-order condition.

Marginal Effect

Consider the binary logit model,

$$p_i = \Pr(y_i = 1 \mid \boldsymbol{x}) = \frac{e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}}} = \frac{1}{1 + e^{-\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\right)}}$$

With some algebraic transformations,

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

The marginal effect for x_1 is given by the expression:

$$\frac{\partial \Pr(y_i = 1 \mid \boldsymbol{x})}{\partial x_1} = \beta_1 \frac{e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}}}{\left(1 + e^{-\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\right)}\right)^2}$$

Marginal Effect

- Nonlinear - as it has to be since the outcome must be bounded between 0 and 1.
- The direction of the change is given by the sign of β_1.
- The effect of x_1 depends on the value of all other covariates in the model even if the underlying model does not include interactions.

Marginal Effect: Numerical Derivative

One-sided derivative

$$f'(x = x_0) \approx \frac{f(x_0 + h) - f(x_0)}{h}$$

Two-sided derivative

$$\begin{aligned} f'(x = x_0) &\approx \frac{f(x_0 + h) - f(x_0) - [f(x_0 - h) - f(x_0)]}{2h} \\ &= \frac{f(x_0 + h) - f(x_0 - h)}{2h} \end{aligned}$$

Marginal Effect: Average Marginal Effect

- Use birth weight data from Wooldridge (bcuse bwght)
- Create an indicator for low birth weight.

Variable	Obs	Mean	Std. Dev.	Min	Max
lw	1,388	.1491354	.3563503	0	1
faminc	1,388	29.02666	18.73928	.5	65
motheduc	1,387	12.93583	2.376728	2	18
cigs	1,388	2.087176	5.972688	0	50

Marginal Effect: Average Marginal Effect

$$\log\left(\frac{lw_i}{1-lw_i}\right) = \beta_0 + \beta_1 \text{cigs}_i + \beta_2 \text{faminc}_i + \beta_3 \text{motheduc}_i$$

```
. logit lw cigs faminc motheduc

Iteration 0:   log likelihood = -584.47305
Iteration 1:   log likelihood =  -573.6873
Iteration 2:   log likelihood = -572.15915
Iteration 3:   log likelihood = -572.15891
Iteration 4:   log likelihood = -572.15891

Logistic regression                             Number of obs     =      1,387
                                                LR chi2(3)        =      24.63
                                                Prob > chi2       =     0.0000
Log likelihood = -572.15891                     Pseudo R2         =     0.0211

------------------------------------------------------------------------------
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |   .0449006   .0104436     4.30   0.000     .0244316    .0653696
      faminc |  -.0080855     .004801    -1.68   0.092    -.0174953    .0013243
    motheduc |   .0031552   .037153      0.08   0.932    -.0696634    .0759738
       _cons |  -1.678173   .4497551    -3.73   0.000    -2.559676   -.7966687
------------------------------------------------------------------------------
```

Marginal Effect: Average Marginal Effect

- Estimate the logit model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

- Increase the value of the variable x_1 by a "small" amount $h : x_1 = x_1 + h$.
  For each observation i, calculate predictions $\hat{y}_{1i}$ in the probability scale
  keeping all other covariate values $(x_{2i}, \ldots, x_{pi})$ as observed.
- Repeat for $x_0 = x_0 - h$
  For each observation i, calculate predictions $\hat{y}_{0i}$ in the probability scale
- For each observation i, calculate the difference of the two predictions divided by $2h$: $(\hat{y}_{i1} - \hat{y}_{i0})/2h$
- The average of this difference is the numerical derivative: $E\left[\frac{\hat{y}_{1i} - \hat{y}_{0i}}{2h}\right] \approx \frac{\partial \Pr(y_i=1|x;\beta)}{\partial x_1}$

Marginal Effect: Average Marginal Effect

```
. sum dydx /* by hand */
```

Variable	Obs	Mean	Std. Dev.	Min	Max
dydx	1,387	.0055771	.001245	.0040404	.0112832

```
. margins, dydx(cigs)

Average marginal effects                        Number of obs     =      1,387
Model VCE    : OIM

Expression   : Pr(lw), predict()
dy/dx w.r.t. : cigs
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
cigs	.0055782	.0012814	4.35	0.000	.0030666	.0080898

- Stata uses an algorithm to ensure numerical precision

Marginal Effect: Average Marginal Effect

- What about 10 extra cigarettes?

```
* 10 units change
preserve
qui logit lw cigs faminc motheduc
predict double lw_0 if e(sample)

replace cigs = cigs + 10
predict double lw_1 if e(sample)

gen double dydx = (lw_1-lw_0)/10
sum dydx
restore
```

```
. sum dydx
```

Variable	Obs	Mean	Std. Dev.	Min	Max
dydx	1,387	.0064608	.0012196	.0048265	.0111532

Marginal Effect: Marginal Effect at the Mean (MEM)

Marginal Effect at the Mean (MEM)

- We can also calculate marginal effects at the mean (of each covariate)
- There is some discussion about which way is better (see Williams, 2012)
- The difference will be so small that it is better to spend mental resources somewhere else.

Marginal Effect: Marginal Effect at the Mean (MEM)

```
preserve
qui sum cigs
scalar h = (abs(r(mean))+0.0001)*0.0001
qui logit lw cigs faminc motheduc, nolog
margins, dydx(cigs) at((mean) faminc motheduc)
margins, dydx(cigs) atmeans

clonevar cigs_c = cigs
* At mean
replace faminc = 29.02666
replace motheduc = 12.93583
* Small negative change
replace cigs = cigs_c - scalar(h)
predict double lw_0 if e(sample)
* Small positive change change
replace cigs = cigs_c + scalar(h)
predict double lw_1 if e(sample)
gen double dydx = (lw_1-lw_0)/(2*scalar(h))
sum dydx
restore
```

```
. sum dydx
```

Variable	Obs	Mean	Std. Dev.	Min	Max
dydx	1,387	.0055624	.0010396	.0051876	.011267

Marginal Effect: Marginal Effect at the Mean (MEM)

```
. margins, dydx(cigs) at((mean) faminc motheduc)

Average marginal effects                        Number of obs     =      1,387
Model VCE    : OIM

Expression   : Pr(lw), predict()
dy/dx w.r.t. : cigs
at           : faminc          =     29.04218 (mean)
               motheduc        =     12.93583 (mean)
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
cigs	.005563	.0012843	4.33	0.000	.0030458	.0080801

```
. margins, dydx(cigs) atmeans

Conditional marginal effects                    Number of obs     =      1,387
Model VCE    : OIM

Expression   : Pr(lw), predict()
dy/dx w.r.t. : cigs
at           : cigs            =     2.088681 (mean)
               faminc          =     29.04218 (mean)
               motheduc        =     12.93583 (mean)
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
cigs	.0055506	.0012879	4.31	0.000	.0030264	.0080749

Interaction Term: Logit

Logit Model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{edu} + \beta_2 \text{male} + \beta_3 \text{edu} * \text{male} + \varepsilon$$

1. Difference male - female for educated:

$$\log\left(\frac{p_{me}}{1-p_{me}}\right) - \log\left(\frac{p_{fe}}{1-p_{fe}}\right) = \beta_2 + \beta_3$$

2. Difference male - female for uneducated:

$$\log\left(\frac{p_{mu}}{1-p_{mu}}\right) - \log\left(\frac{p_{fu}}{1-p_{fu}}\right) = \beta_2$$

3. The difference of differences (2)-(3) is:

$$\left[\log\left(\frac{p_{me}}{1-p_{me}}\right) - \log\left(\frac{p_{fe}}{1-p_{fe}}\right)\right] - \left[\log\left(\frac{p_{mu}}{1-p_{mu}}\right) - \log\left(\frac{p_{fu}}{1-p_{fu}}\right)\right] = \beta_3$$

Difference-in-difference in the log-odds scale.

Interaction Term: Marginal Effect

Marginal Effects

```
. margins, dydx(*)

Average marginal effects                        Number of obs     =      1,388
Model VCE    : OIM

Expression   : Pr(lw), predict()
dy/dx w.r.t. : cigs 1.inc
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
cigs	.0046941	.0018482	2.54	0.011	.0010717	.0083165
1.inc	-.0288422	.0225222	-1.28	0.200	-.0729849	.0153006

Note: dy/dx for factor levels is the discrete change from the base level.

Only two effects? The model has three coefficients. Where is the interaction?

Interaction Term: Marginal Effect

How Stata calculates marginal effects?

- For cigs, a continuous variable, it's using the two-sided derivative increasing cigs by a little bit and calculating predictions.
  It's increasing cigs in both the main effect and the interaction.
  Then it takes an average so the marginal effect of cigs is the numerical derivative for both inc=1 and inc=0 combined.
- For the marginal effect of inc, it's doing the same going from 0 to 1, averaging over the values of cigs

To get the marginal effect of cigs separately for inc=1 and inc=0, we have to be more specific.

Interaction Term: More Specific

```
. margins, dydx(cigs) at(inc=(0 1)) vsquish

Average marginal effects                        Number of obs     =      1,388
Model VCE    : OIM

Expression   : Pr(lw), predict()
dy/dx w.r.t. : cigs
1._at        : inc             =           0
2._at        : inc             =           1
```

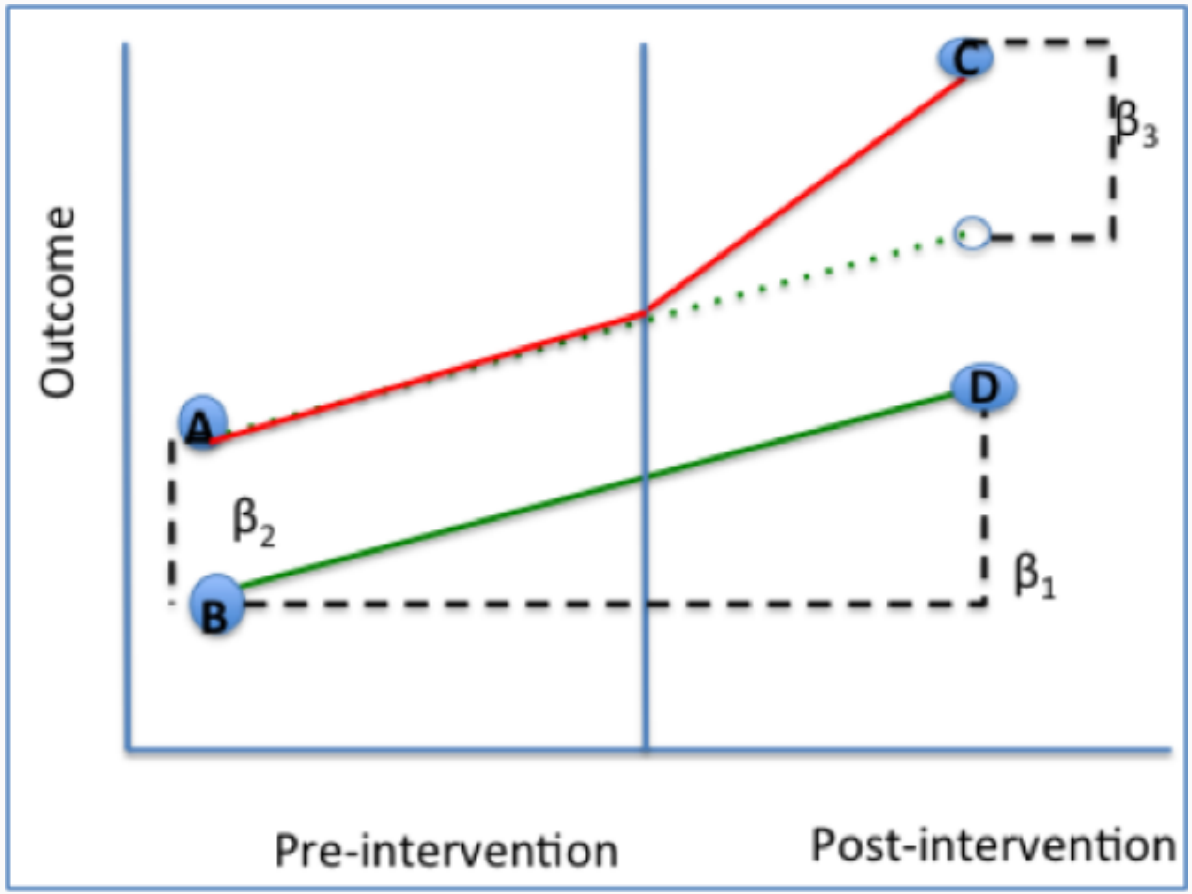
	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cigs						
_at						
1	.0062867	.0012881	4.88	0.000	.0037621	.0088113
2	-.0004394	.0062301	-0.07	0.944	-.0126501	.0117713

- A small increase in cigs increases the probability of low birth weight by 0.6 percentage points for low income type.
- While increase in cigs has no significant effect for high income type.

Interaction Term: DID in Linear

Regression Model

$$y = \beta_0 + \beta_1 \text{time} + \beta_2 \text{treated} + \beta_3 \text{time} * \text{treated} + \varepsilon$$



Interaction Term: DID in Non-Linear

Let $u = F(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + X\beta)$

- When the interacted variables are both continuous

$$\begin{aligned}\frac{\partial^2 F(u)}{\partial x_1 \partial x_2} &= \frac{\partial \{(\beta_1 + \beta_{12} x_2) f(u)\}}{\partial x_2} \\ &= \beta_{12} f(u) + (\beta_1 + \beta_{12} x_2)(\beta_2 + \beta_{12} x_1) f'(u)\end{aligned}$$

  where $f(u) = F'(u)$ and $f'(u) = F''(u)$.

- When the interacted variables are both dummy variables

$$\begin{aligned}\frac{\Delta^2 F(u)}{\Delta x_1 \Delta x_2} =& \frac{\Delta \{F(\beta_1 + \beta_2 x_2 + \beta_{12} x_2 + X\beta) - F(\beta_2 x_2 + X\beta)\}}{\Delta x_2} \\ =& F(\beta_1 + \beta_2 + \beta_{12} + X\beta) \\ & - F(\beta_1 + X\beta) - F(\beta_2 + X\beta) + F(X\beta)\end{aligned}$$

Interaction Term: Logit Formula

For the logit model,

$$F(u) = \frac{1}{1 + e^{-(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + X\beta)}}$$

When the interacted variables are both dummy variables, the interaction effect is the discrete double difference:

$$\begin{aligned}\frac{\Delta^2 F(u)}{\Delta x_1 \Delta x_2} = &\frac{1}{1 + e^{-(\beta_1 + \beta_2 + \beta_{12} + X\beta)}} \\ &- \frac{1}{1 + e^{-(\beta_1 + X\beta)}} - \frac{1}{1 + e^{-(\beta_2 + X\beta)}} + \frac{1}{1 + e^{-X\beta}}\end{aligned}$$

Interaction Term: SEs

Ai and Norton (2003) derive the standard errors for the interaction effect in logit and probit models, applying the Delta method.

For the case of two dummy variables, the asymptotic variance of the estimated interaction effect is estimated consistently by

$$\frac{\partial}{\partial \beta'} \left\{ \frac{\Delta^2 F(u)}{\Delta x_1 \Delta x_2} \right\} \widehat{\Omega}_\beta \frac{\partial}{\partial \beta} \left\{ \frac{\Delta^2 F(u)}{\Delta x_1 \Delta x_2} \right\}$$

where $\widehat{\Omega}_\beta$ is a consistent covariance estimator of $\widehat{\beta}$.

For continuous variables, we replace the discrete difference operator Δ with the partial derivative operator.

Interaction Term: Command in Stata

```
. qui logit outcome treated  time did, nolog

. inteff outcome treated  time did
Logit with two dummy variables interacted
(0 observations deleted)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
_logit_ie	3,000	.14106	0	.14106	.14106
_logit_se	3,000	.0391754	0	.0391754	.0391754
_logit_z	3,000	3.600726	0	3.600726	3.600726

Interaction Term: Check

```
/* manual estimation */
logit outcome treated time did, nolog

replace treated = 0
replace time = 0
replace did = 0
predict double y00 if e(sample)

(repeated steps omitted)

gen ie = y11-y01-y10+y00
sum ie
```

LPM vs Logit

We have often used binary ("dummy") variables as explanatory variables in regressions.

- It's possible to use OLS:

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

  where y is the dummy variable. This is called the linear probability model (LPM).

- Estimating the equation:

$$\hat{P}(y = 1 \mid x) = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

  $\hat{y}$ is the predicted probability of having $y = 1$ for the given values of $x_1 \ldots x_k$.

LPM vs Logit: Problems in LPM

First Problems with LPM:

- Possible to get $\hat{y} < 0$ or $\hat{y} > 1$. This makes no sense-we can't have a probability below 0 or above 1 .
- This is a fundamental problem with the LPM that we can't patch up.

LPM vs Logit: Problems in LPM

Second Problem with LPMd: SEs are not right

- Recall that in the linear model we assume $Y \sim N(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \sigma^2)$ or equivalently, $\epsilon_i \sim N(0, \sigma^2)$
- That is, Y distributes normal conditional on **Xs** or the error distributes normal with mean 0
- Obviously, a 1/0 variable can't distribute normal, and ϵ_i can't be normally distributed either

LPM vs Logit: Heteroskedasticity

- The variance of a 1/0 (binary) depends on the values of X so there is always heteroskedasticity: $\text{var}(y \mid \mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})]$
- We can correct SEs in LPMs using the robust option (Huber-White SEs; aka sandwich estimator)
- Still, we do know that SEs are not totally correct because they do not distribute normal either, even is we somehow correct for heteroskedasticity

But at the very least, use the robust option by default.

LPM is the wrong but super useful model because changes can be interpreted in the probability scale.

LPM vs Logit: Solution

Solution: Use the logit or probit model.

- These models are specifically made for binary dependent variables and always result in $0 < \hat{y} < 1$.
- This is the main feature of a logit/probit that distinguishes it from the LPM - predicted probability of $y = 1$ is never below 0 or above 1.

LPM vs Logit: Solution

Another feature for the logit or probit model.

- The relation of the logit probability to representative utility is sigmoid, or S-shaped.
- When the representative utility of an alternative is very low, a small increase in the utility of the alternative has little effect on the probability of its being chosen.
- The same when the representative utility of an alternative is very high.

- When the probability is close to 0.5, meaning a 50-50 chance of the alternative being chosen, the increase in representative utility has the greatest effect on the probability of its being chosen.
- In this case, a small improvement tips the balance in people's choices, inducing a large change in probability.

Reference

- Train, Kenneth E. "Discrete choice methods with simulation", 2009.
- Perraillon, M., R. Lindrooth, and D. Hedeker. "Health services research and program evaluation: causal inference and estimation", 2022.
- Karaca-Mandic, Pinar, Edward C. Norton, and Bryan Dowd. "Interaction terms in nonlinear models." Health services research 47.1pt1 (2012): 255-274.
- Norton, Edward C., Hua Wang, and Chunrong Ai. "Computing interaction effects and standard errors in logit and probit models." The Stata Journal 4.2 (2004): 154-167.

Thanks!

Derivation of Logit

We have

$$P_{ni} = \int_{s=-\infty}^{\infty} \left(\prod_{j \neq i} e^{-e^{-\left(s+V_{ni}-V_{nj}\right)}} \right) e^{-s} e^{-e^{-s}} ds,$$

where s is ε_{ni}. Our task is to evaluate this integral. Noting that $V_{ni} - V_{ni} = 0$ and then collecting terms in the exponent of e, we have

$$\begin{aligned} P_{ni} &= \int_{s=-\infty}^{\infty} \left(\prod_{j} e^{-e^{-\left(s+V_{ni}-V_{nj}\right)}} \right) e^{-s} ds \\ &= \int_{s=-\infty}^{\infty} \exp\left(-\sum_{j} e^{-\left(s+V_{ni}-V_{nj}\right)} \right) e^{-s} ds \\ &= \int_{s=-\infty}^{\infty} \exp\left(-e^{-s} \sum_{j} e^{-\left(V_{ni}-V_{nj}\right)} \right) e^{-s} ds. \end{aligned}$$

Derivation of Logit

Define $t = \exp(-s)$ such that $-\exp(-s)ds = dt$. Note that as s approaches infinity, t approaches zero, and as s approaches negative infinity, t becomes infinitely large. Using this new term,

$$
\begin{aligned}
P_{ni} &= \int_{\infty}^{0} \exp\left(-t \sum_j e^{-\left(V_{ni}-V_{nj}\right)}\right)(-dt) \\
&= \int_{0}^{\infty} \exp\left(-t \sum_j e^{-\left(V_{ni}-V_{nj}\right)}\right) dt \\
&= \left.\frac{\exp\left(-t \sum_j e^{-\left(V_{ni}-V_{nj}\right)}\right)}{-\sum_j e^{-\left(V_{ni}-V_{nj}\right)}}\right|_{0}^{\infty} \\
&= \frac{1}{\sum_j e^{-\left(V_{ni}-V_{nj}\right)}} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}
\end{aligned}
$$

Go back