TA Session 3: Discrete Choice

Microeconometrics with Hanna Wang

IDEA, Fall 2022

TA: Conghan Zheng

Overview

Binary Outcome Models

Introduction

- **Data** (`TA3_1.dta`): US individual data on labor force participation from the Current Population Survey (CPS). 2010 cross-section, 16-64 years-old women.
- **Research question**: We are going to study the determinants of the decision to participate in the labor market for women. This choice is recorded by dummy `lfp` (denoted by y).

$$y = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

- *Limited dependent variable*: y has support $\{0, 1\}$, and this restriction has consequences for econometric modeling.
- In regression analysis, we want to measure how response probability p varies across individuals as a function of regressors X: $\mathbb{P}(y = 1|X) = p(X)$.
- A traditional approach is parametric modelling with MLE. Two parametric forms for $p(X)$: *logit* and *probit*.

Latent Variable Interpretation

Random Utility Formulation

- A decision-maker chooses between alternatives 0 and 1 according to which has the higher utility. Outcome variable y indicates which alternative is chosen.
- The additive random utility model (ARUM) specifies the utilities of alternatives:

$$U_0 = V_0(X) + \varepsilon_0$$
$$U_1 = V_1(X) + \varepsilon_1$$

- where Vs are deterministic components of utility (deterministic function of data) and εs are random components of utility.
- It follows that

$$y = \begin{cases} 1 & \text{if } U_1 \geq U_0 \\ 0 & \text{otherwise} \end{cases}$$

Latent Variable Interpretation

Random Utility Formulation

$$
\begin{aligned}
\mathbb{P}(y=1|X) &= \mathbb{P}(U_1 \geq U_0) \\
&= \mathbb{P}\left[V_1(X)+\varepsilon_1 \geq V_0(X)+\varepsilon_0\right] \\
&= \mathbb{P}\left[\varepsilon_0-\varepsilon_1 \leq V_1(X)-V_0(X)\right] \\
&= F\left[V_1(X)-V_0(X)\right]
\end{aligned}
$$

where $\varepsilon_0-\varepsilon_1 \sim F$.

- Notice when we model the reponse probability on regressors:

$$\mathbb{P}(y=1|X)=F(X\beta) \Leftarrow X\beta = V_1(X)-V_0(X)$$

- The outcome probabilities depend on the difference in errors, only $m-1$ errors (m is the number of alternatives, here $m=2$) are free to vary, and similarly, only $m-1$ of the $\beta^{(1)}, \ldots, \beta^{(m)}$ are free to vary.
- Therefore the model identification requires a scale normalization on $Var(\varepsilon_0-\varepsilon_1)$, or on $Var(\varepsilon_0)$ and $Var(\varepsilon_1)$ separately.

Models for the Response Probability

- **Linear Probability Model**: where $F(X\beta) = X\beta$, has the advantage that it's simple to interpret. But it has two problems: (1) some of the OLS fitted values $\hat{y}$ could be outside the unit interval – larger than 1 or smaller than 0; (2) heteroskedasticity is present unless all of the slope coeffcients β are zero (recall Bernoulli distribution), and we can't apply WLS to fix this if (1) is true. Overall, LPM is a poor choice for modelling probabilities.
- *Index Models* restrict the way in which the response probability depends on X.
  - **Probit Probability Model**: where $F(X\beta) = \Phi(X\beta)$, Φ is the standard normal CDF.
  - **Logit Probability Model**: where $F(X\beta) = \Lambda(X\beta)$, Λ is the logistic CDF. The logistic and normal distribution (appropriately scaled) have similar shapes so they typically produce similar estimates for the response probabilities and marginal effects. One advantage of logit: its distribution function is available in closed form which speeds computation.
- For binary models other than the LPM, estimation is done by ML. The MLE is obtained by iterative methods and is asymptotically normally distributed. Consistent estimates are obtained if $F(\cdot)$ is correctly specified.

Interpreting estimates

Partial effects

- **Partial effects**
  - Continuous regressor:

$$\frac{\partial p}{\partial X_j} = \frac{\partial F(X\beta)}{\partial X_j} = f(X\beta)\cdot\beta_j, \quad \text{where } \underbrace{f(X\beta)}_{F'(\cdot)>0} = \left.\frac{\partial F(u)}{\partial u}\right|_{X\beta}$$

    The effect of one regressor on the response probability depends on the values of all other regressors.

    And the relative effects doesn't depend on X: $\frac{\frac{\partial F(X\beta)}{\partial X_j}}{\frac{\partial F(X\beta)}{\partial X_h}} = \frac{\beta_j}{\beta_h}$.
  - Discrete regressor: the partial effect from X_j changing one unit is

$$\begin{aligned}\Delta p =& F\left[\beta_0+\beta_1X_1+\cdots+\beta_{j-1}X_{j-1}+\beta_j(X_j+1)+\beta_{j+1}X_{j+1}+\cdots+\beta_KX_K\right] \\ &-F\left[\beta_0+\beta_1X_1+\cdots+\beta_{j-1}X_{j-1}+\beta_jX_j+\beta_{j+1}X_{j+1}+\cdots+\beta_KX_K\right]\end{aligned}$$

  - The estimated $\hat{\beta}_{MLE}$ is not comparable across different specifications of $F(\cdot)$.

Binary Outcome Models

Logit

```
. logit lfp age age2 married educ black nchild citiz

Logistic regression                                     Number of obs = 169,588
                                                        LR chi2(7)    = 18561.70
                                                        Prob > chi2   =   0.0000
Log likelihood = -94992.85                              Pseudo R2     =   0.0890
```

lfp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
age	.2603412	.0029713	87.62	0.000	.2545176	.2661648
age2	-.0032281	.0000368	-87.83	0.000	-.0033002	-.0031561
married	-.2690702	.0131599	-20.45	0.000	-.2948632	-.2432772
educ	.0181616	.0002605	69.71	0.000	.017651	.0186723
black	-.153129	.0173409	-8.83	0.000	-.1871166	-.1191414
nchild	-.1586691	.0055978	-28.35	0.000	-.1696405	-.1476978
citiz	.3888647	.0204338	19.03	0.000	.3488153	.4289142
_cons	-5.307922	.0539313	-98.42	0.000	-5.413625	-5.202218

Interpreting estimates

Odds Ratio

- For *ordered categorical regressors*, many researchers prefer odds ratio from Logit. In this way, β_j can be interpreted as semi-elasticity.
- Recall in logit we have $P(y=1|x) = F(x\beta) = \frac{e^{x\beta}}{1+e^{x\beta}}$.
- **odds ratio/relative risk**: $\frac{p}{1-p} = \frac{\frac{e^{x\beta}}{1+e^{x\beta}}}{\frac{1}{1+e^{x\beta}}} = e^{x\beta}$.
- Consider x_1 (e.g. income quantile) increases for one unit, $\delta = (0, 1, 0, \ldots, 0)$, it follows that

$$\frac{odds[(x+\delta)\beta]}{odds(x\beta)} = \frac{e^{\beta_0+(x_1+1)\beta_1+x_2\beta_2+x_3\beta_3+\cdots}}{e^{\beta_0+x_1\beta_1+x_2\beta_2+x_3\beta_3+\cdots}} = e^{\beta_1}$$

- The interpretation on odds ratio is meaningless when x_1 is unordered, and is questionable if x_1 is not coded with consecutive numbers. Then you could run `logit y i.x, or` in Stata to deliver the odds ratio for each category of x_1 and interprete on them.
- For Probit model, we can't have this interpretation on $\hat{\beta}_{MLE}$.

Interpreting estimates

Odds Ratio

```
. logit lfp age age2 married educ black nchild citiz, or
```

```
Logistic regression                                     Number of obs =  169,588
                                                        LR chi2(7)    = 18561.70
                                                        Prob > chi2   =   0.0000
Log likelihood = -94992.85                              Pseudo R2     =   0.0890
```

lfp	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
age	1.297373	.0038549	87.62	0.000	1.289839	1.30495
age2	.9967771	.0000366	-87.83	0.000	.9967053	.9968489
married	.7640896	.0100554	-20.45	0.000	.7446334	.7840542
educ	1.018328	.0002653	69.71	0.000	1.017808	1.018848
black	.858019	.0148789	-8.83	0.000	.829347	.8876823
nchild	.8532786	.0047764	-28.35	0.000	.8439681	.8626918
citiz	1.475305	.030146	19.03	0.000	1.417387	1.535589
_cons	.0049522	.0002671	-98.42	0.000	.0044555	.0055043

Note: **_cons** estimates baseline odds.

Interpreting estimates

Odds Ratio

- Consider binary variable `married`.

$$\begin{aligned}
\text{odds ratio}_{married} &= \frac{\text{odds}_{married}}{\text{odds}_{not\ married}} = \frac{p}{1-p} \approx 0.76 \\
\text{coefficient } b_{married} &= \ln \text{odds}_{married} - \ln \text{odds}_{not\ married} \\
&= \ln \frac{\text{odds}_{married}}{\text{odds}_{not\ married}} = \ln \frac{p}{1-p} \approx -0.27 \\
\text{odds ratio} &= \exp(\text{coefficient})
\end{aligned}$$

  $e^{-0.27} \approx 0.76$ implies that the odds of participating versus not participating for the married is 0.76 times that of non-married (relative probability decreases), that is to say, the married are less likely to participate.
- For continuous variables, where the odds ratios could be very confusing, we better choose to interpret marginal effects.

Interpreting estimates

Marginal effects

- Marginal effects are measured in the probability scale which is often the scale of interest.
- In a nonlinear model (e.g. Logit and Probit), marginal effects are more informative than coefficients.
- Three variants of **Marginal effects**:
  - Marginal effects at the mean (MEM)
  - Marginal effects at a representative value (MER)
  - Average marginal effects (AME)

Model	Probability $p = P(y = 1 \mid x)$	Marginal effect $\frac{\partial p}{\partial x_j}$
LPM	$F(x\beta) = x\beta$	β_j
Logit	$\Lambda(x\beta) = \frac{e^{x\beta}}{1+e^{x\beta}}$	$\Lambda(x\beta)(1 - \Lambda(x\beta))\beta_j$
Probit	$\Phi(x\beta) = \int_{-\infty}^{x\beta} \phi(z)dz$	$\phi(x\beta)\beta_j$

Interpreting estimates

Marginal effect at the mean (MEM)

- Marginal effect at the mean: covariates are fixed at their means. Marginal effects are interpreted in terms of expected probabilities of a person with average characteristics.

```
. margins, dydx(*) atmeans
```

```
Conditional marginal effects                    Number of obs     =    169,588
Model VCE    : OIM

Expression   : Pr(lfp), predict()
dy/dx w.r.t. : age age2 married educ black nchild citiz
at           : age             =    39.78121 (mean)
               age2            =    1776.929 (mean)
               married         =    .5147652 (mean)
               educ            =    84.80023 (mean)
               black           =    .1168007 (mean)
               nchild          =     .866783 (mean)
               citiz           =    .9211619 (mean)
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0531153	.0006006	88.43	0.000	.051938	.0542925
age2	-.0006586	7.43e-06	-88.69	0.000	-.0006732	-.0006441
married	-.0548962	.0026818	-20.47	0.000	-.0601524	-.04964
educ	.0037054	.0000527	70.28	0.000	.003602	.0038087
black	-.0312417	.0035372	-8.83	0.000	-.0381744	-.024309
nchild	-.032372	.0011398	-28.40	0.000	-.034606	-.0301379
citiz	.0793369	.0041705	19.02	0.000	.0711629	.0875109

Interpreting estimates

Marginal effect at a representative value (MER)

- Marginal effect at a representative value: covariates are fixed at a vector chosen by the economist.
- A chosen benchmark: a 20-year-old married black female citizen with two children ...

```
. margins, dydx(*) at(age=20 age2=400 married=1 educ=4
black=1 nchild=2 citiz=1)
```

```
Conditional marginal effects                    Number of obs     =    169,588
Model VCE    : OIM

Expression   : Pr(lfp), predict()
dy/dx w.r.t. : age age2 married educ black nchild citiz
at           : age             =          20
               age2            =         400
               married         =           1
               educ            =           4
               black           =           1
               nchild          =           2
               citiz           =           1
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0347012	.0007019	49.44	0.000	.0333256	.0360769
age2	-.0004303	8.88e-06	-48.45	0.000	-.0004477	-.0004129
married	-.0358647	.0015513	-23.12	0.000	-.0389052	-.0328242
educ	.0024208	.0000391	61.87	0.000	.0023441	.0024975
black	-.0204108	.0021195	-9.63	0.000	-.0245649	-.0162566
nchild	-.0211492	.0008047	-26.28	0.000	-.0227264	-.019572
citiz	.0518323	.0031036	16.70	0.000	.0457494	.0579152

Interpreting estimates

Average Marginal Effect (AME)

- Average marginal effect: $AME = \frac{\partial F(X\beta)}{X} = \beta\mathbb{E}[f(X\beta)]$, the average of marginal effects for each individual.

```
. margins, dydx(*)
```

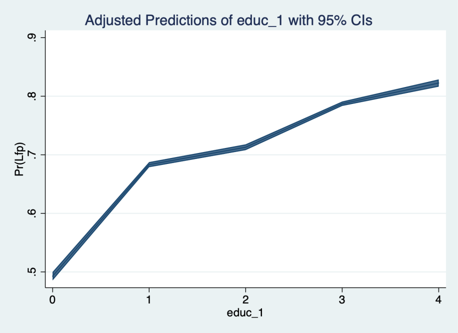
```
Average marginal effects                        Number of obs     =    169,588
Model VCE    : OIM

Expression   : Pr(lfp), predict()
dy/dx w.r.t. : age age2 married educ black nchild citiz
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0490896	.0005149	95.33	0.000	.0480804	.0500989
age2	-.0006087	6.37e-06	-95.58	0.000	-.0006212	-.0005962
married	-.0507356	.0024718	-20.53	0.000	-.0555802	-.0458909
educ	.0034245	.0000468	73.24	0.000	.0033329	.0035162
black	-.0288738	.0032674	-8.84	0.000	-.0352779	-.0224698
nchild	-.0299185	.0010475	-28.56	0.000	-.0319715	-.0278655
citiz	.0733239	.0038377	19.11	0.000	.0658022	.0808456

Interpreting estimates

Marginal Effects

- When we calculate at-means marginal effects, for categorical variables, they are set to their sample averages, which are not meaningful (e.g., avg(educ) = 84). Instead, we can either create a benchmark value or calculate the marginal effect at each of the categories.
- Example: Margins by education. After simplifying the education categories (educ_1), we plot the margins:



Interpreting estimates
Iteration Log

```
Iteration 0:   log likelihood =  -104273.7
Iteration 1:   log likelihood = -95138.212
Iteration 2:   log likelihood = -94993.011
Iteration 3:   log likelihood =  -94992.85
Iteration 4:   log likelihood =  -94992.85

Logistic regression                             Number of obs =   169,588
                                                LR chi2(7)    =  18561.70
                                                Prob > chi2   =    0.0000
Log likelihood = -94992.85                      Pseudo R2     =    0.0890
```

- The iteration log shows fast convergence in four iterations. In practice, a large number of iterations may signal a high degree of multicollinearity (which may lead to a ridge instead of a peak).

Goodness of Fit

- **Goodness of Fit** is interpreted as closeness of fitted values to sample values of the dependent variable.
- Measures of Goodness of fit:
  1. Predicted outcomes
  2. Predicted frequencies
  3. Pseudo-R^2

Goodness of Fit

Predicted Outcomes

Classification:

- If we want to predict the outcome variable ($y = 0, 1$) and assume a symmetric loss function, it's natural to set

$$\tilde{y} = 1 \text{ if } F(x\beta) \geq 0.5,$$
$$\tilde{y} = 0 \text{ if } F(x\beta) < 0.5$$

- One measure of goodness of fit is the percentage of correctly classified obervations. Four possible cases:
  1. $(y, \tilde{y}) = (0, 0)$
  2. $(y, \tilde{y}) = (1, 1)$
  3. $(y, \tilde{y}) = (1, 0)$
  4. $(y, \tilde{y}) = (0, 1)$
- Problem: If we have 100 observations, 70 of them are zeros, and we predict all of them are zero. We still correctly predict 70% of all outcomes even if none of the $y = 1$ values are correctly predicted.
- Solution: Set the overall percent correctly predicted as the weighted average of the percent correctly predicted for $y = 0$ and $y = 1$.

Goodness of Fit

Predicted Outcomes

Threshold:

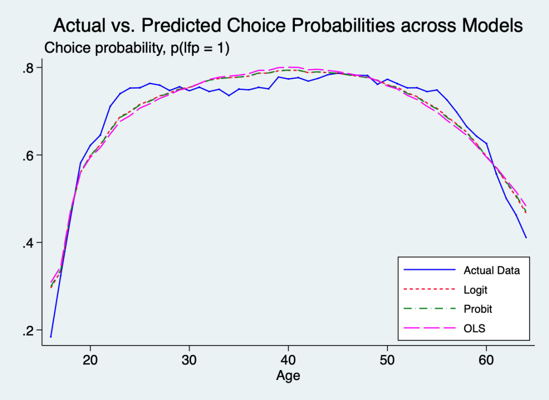
1. The 0.5 threshold
   - Some have criticized the prediction rule for always using a threshold value of 0.5, especially when one of the outcomes is unlikely.
2. One alternative is to use the fraction of successes in the sample ($\bar{y}$) as the threshold.
   - If $\bar{y} < 0.5$ (> 0.5), using this rule will certainly increase the number of predicted successes (failures), but not without cost: we necessarily make more mistakes in predicting the failures (successes).
   - In terms of the overall percent correctly predicted, we may actually do worse than when using the traditional 0.5 threshold.
3. A third possibility is to choose the threshold such that the fraction of above threshold values $\tilde{y}_i = 1$ in the sample is the same (or very close) to $\bar{y}$:

$$\alpha = \arg\min_{\alpha} \left\{ \sum_i \mathbb{1}\left(F(X_i'\beta) \geq \alpha\right) - \sum_i y_i \right\}$$

Goodness of Fit

Predicted Probabilities

- Problem: average predicted probabilities $\frac{1}{N}\sum_i \hat{p}_i$ = sample frequency $\bar{y}$.
  ▸ ML FOC
- Solution: use subsamples (e.g., cohort, income decile).



Goodness of Fit

Pseudo-R^2

- A pseudo-R^2 is an extension of R^2 to nonlinear regression model.
- Pseudo-R^2 measure proposed by McFadden (1974):

$$\tilde{R}^2 = \frac{\ln \mathcal{L}_N(\hat{\beta}) - \ln \mathcal{L}_N(\bar{y})}{\ln 1 - \ln \mathcal{L}_N(\bar{y})} = \frac{\ln \mathcal{L}_N(\hat{\beta}) - \ln \mathcal{L}_N(\bar{y})}{0 - \mathcal{L}_N(\bar{y})} = 1 - \frac{\mathcal{L}_N(\hat{\beta})}{\mathcal{L}_N(\bar{y})}$$
$$= 1 - \frac{\sum_i \left\{ y_i \ln F(X_i'\hat{\beta}) + (1 - y_i) \ln[1 - F(X_i'\hat{\beta})] \right\}}{N\left[\bar{y} \ln \bar{y} + (1 - \bar{y}) \ln(1 - \bar{y})\right]}$$

  where $\ln 1$ is the maximum value in the support of a log-likelihood $\mathcal{L}_N(\beta)$; $\ln 1 - \mathcal{L}_N(\bar{y})$ is the maximum possible improvement from the likelihood of a intercept-only model (only includes the constant term as regressor, $\bar{y}$ estimated); and $\ln \mathcal{L}_N(\hat{\beta}) - \ln \mathcal{L}_N(\bar{y})$ is the improvement in likelihood achieved by the estimated $\hat{\beta}$ from the intercept-only model.
- $\tilde{R}^2$ is the proportion of the actual increase in the likelihood to the maximum possible increase of the likelihood, it increases as more regressors are added.
- Because the log likelihood for a binary response model is always negative ($p \in (0,1) \Rightarrow \ln p < 0$), $0 > \mathcal{L}_N(\hat{\beta}) \geq \mathcal{L}_N(\bar{y})$, and so the pseudo-$R^2$ is always between zero and one.

Model Specification tests

Examples

- Wald Test: add regressors $(X_{K+1}, \cdots, X_{K+l})$ to the regression, test $H_0 : (\beta_{K+1}, \cdots, \beta_{K+l}) = 0$.
- Likelihood-ratio test: add regressors $(X_{K+1}, \cdots, X_{K+l})$ to the regression, test $H_0 \Leftrightarrow \ln \mathcal{L} = \ln \mathcal{L}_{+l}$.
- Lagrange multiplier test: add regressor $(X\hat{\beta})^2$ to the regression, test on its coefficient $H_0 : \beta_{K+1} = 0$.
    - If the null is rejected, it means that the departure from $X\beta$ in the direction of an asymmetric form provides us a better model.

Comparison of Estimates

- Logit and Probit models have similar shapes for central values of $F(\cdot)$ but differ in the tails.
- According to Amemiya (1981), coefficients can be compare across models using the rough conversion factors

$$\hat{\beta}_{Logit} \approx 4\hat{\beta}_{OLS}$$
$$\hat{\beta}_{Probit} \approx 2.5\hat{\beta}_{OLS}$$
$$\hat{\beta}_{Logit} \approx 1.6\hat{\beta}_{Probit}$$

This can be derived from the marginal effects across models.

Comparison of Estimates

	(1) Logit	(2) Logit r	(3) Probit	(4) Probit r	(5) OLS	(6) OLS r
main						
age	0.242***	0.242***	0.145***	0.145***	0.0487***	0.0487***
	(0.00303)	(0.00309)	(0.00180)	(0.00183)	(0.000576)	(0.000605)
age2	-0.00303***	-0.00303***	-0.00181***	-0.00181***	-0.000609***	-0.000609***
	(0.0000373)	(0.0000380)	(0.0000221)	(0.0000225)	(0.00000710)	(0.00000746)
married	-0.279***	-0.279***	-0.165***	-0.165***	-0.0507***	-0.0507***
	(0.0132)	(0.0131)	(0.00778)	(0.00774)	(0.00243)	(0.00237)

- The estimates from the models tell a consistent story about the impact of a regressor on $\mathbb{P}(lfp = 1)$.
- In binary outcome models, by adopting the Logit or Probit model, the distribution of the error term and the independence of observations over i are assumed. Since the variance of a binary variable is always $p(1-p)$, if the model is correctly specified, there is no need to use the `vce(robust)` option in Stata or the `sandwich` package in R.
- The only need for robust variance is when there is clustering.
- But if the model is mis-specified (on $F(\cdot)$ or on $X\beta$), the estimates are not even consistent, and the quasi-ML theory applies.

Multinomial Models

Additive random utitlity model

Conditional Logit

- Let's consider the useful *additive random utitlity model* we have seen before, now we have $J > 2$:

$$U_j = X\beta_j + Z_j\gamma + \varepsilon_j,\;\; j \in \{1, \dots, J\}$$

- The response probability:

$$\begin{aligned} p_j(x,z) &\equiv \mathbb{P}(y = j | X = x, Z = z) \\ &= \mathbb{P}(U_j \geq U_k, \forall k \neq j) \\ &= \mathbb{P}(\varepsilon_k - \varepsilon_j \leq x(\beta_j - \beta_k) + (z_j - z_k)\gamma, \forall k \neq j) \end{aligned}$$

- Under the assumption that $\{\varepsilon_1, \dots, \varepsilon_J\}$ are jointly Type-I Extreme Value distributed, it follows that $p_j = \frac{e^{x\beta_j + z_j\gamma}}{\sum_{l=\mathbf{1}}^{J} e^{x\beta_l + z_l\gamma}}$.
- Only $J-1$ errors of $\{\varepsilon_1, \dots, \varepsilon_J\}$ are free to vary, and similarly, only $J-1$ of $\{\beta_1, \dots, \beta_J\}$ are free to vary, while γ is identified. We have $J-1$ differences to solve for J parameters, one of the errors need to be normalized.

Multinomial Models

- Multinomial Logit (MNL)
  - Response utility: $p_j(x) = \frac{e^{x\beta_j}}{\sum_{l=1}^{J} e^{x\beta_l}}$; latent utility: $U_j = X\beta_j + \varepsilon_j$.
  - Regressors (e.g., age and income) are alternative-invariant: $x_j = x$ for all $j = 1, ..., J$, which means, regressors are specific to the individual but not the alternative (they do not have a j subscript)
- Conditional Logit (CL)
  - Response utility: $p_j(x) = \frac{e^{z_j\gamma}}{\sum_{l=1}^{J} e^{z_j\gamma}}$; latent utility: $U_j = Z_j\gamma + \varepsilon_j$.
  - Regressors vary across alternatives (e.g. price or time cost of each alternative). These alternative-specific regressors only affect an individual's utility if that specific alternative is selected, so they have a j subscript (the regressor varies across j while the coefficient γ are common).
- Or more generally, a conditional Logit:
  - Response utility: $p_j(x) = \frac{e^{x\beta_j + z_j\gamma}}{\sum_{l=1}^{J} e^{x\beta_j + z_j\gamma}}$; latent utility: $U_j = X\beta_j + Z_j\gamma + \varepsilon_j$.
- The MNL model can be reexpressed as a CL model.

Multinomial Logit

Multinomial logistic regression	Number of obs = **147,843**
	LR chi2(**20**) = **20095.69**
	Prob > chi2 = **0.0000**
Log likelihood = **-113748.54**	Pseudo R2 = **0.0812**

sector	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
Not_participating	(base outcome)					
Self_employed						
age	.3625662	.0081511	44.48	0.000	.3465904	.3785421
age2	-.004007	.0000939	-42.69	0.000	-.0041909	-.003823
married	.1393773	.0279609	4.98	0.000	.0845749	.1941797

- $\mathbb{P}(\texttt{sector} = j) = \frac{e^{x\beta_j}}{\sum_{l=1}^{J} e^{x\beta_l}}$
- **Coefficient interpretation**:
  - Coefficients in a multinomial model can be interpreted in the same way as binary logit model parameters are interpreted, with comparison being to the base category.
  - $\hat{\beta}_j$ can be viewed as parameters of a binary logit model between alternative j and the base alternative (the omitted category).
  - A positve coefficient from `mlogit` means that as the regressor increases, we are more likely to choose alternative j than the base.

Multinomial Logit

```
Multinomial logistic regression                 Number of obs =  147,843
                                                LR chi2(20)   = 20095.69
                                                Prob > chi2   =   0.0000
Log likelihood = -113748.54                     Pseudo R2     =   0.0812
```

sector	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
Not_participating	(base outcome)					
Self_employed						
age	.3625662	.0081511	44.48	0.000	.3465904	.3785421
age2	-.004007	.0000939	-42.69	0.000	-.0041909	-.003823
married	.1393773	.0279609	4.98	0.000	.0845749	.1941797

Coefficient interpretation (*continued*):

- For multinomial models, Stata reports the pseudo-R^2 we've seen in the binary model: $\tilde{R}^2 = 1 - \frac{\ln L_{fit}}{\ln L_0}$, where $\ln L_0$ is the log likelihood of an intercept-only model, and $\ln L_{fit}$ is the likelihood of the fitted model. And again, for discrete dependent variables, $\tilde{R}^2$ has desirable properties including that it increases as regressors are added for models fitted by ML.
- The model fit is quite poor with pseudo-R^2 equal to 0.0812.
- The LR chi-squred is super large (20095.69), hence the regressors are jointly statistically significant at the 0.05 level.

Multinomial Logit

Relative-risk ratio

Relative-risk ratio (odds ratio as in the binary case):

- The relative risk ratio of choosing alternative j rather than alternative 0 is given by

$$\frac{sector_i = j}{sector_i = 0} = e^{x_i \beta_j}$$

where e^{β_j} gives the proportionate change in the relative risk of choosing j over 0 when x_i changes by one unit.

Multinomial Logit

Relative-risk ratio

sector	RRR	Std. Err.
Not_participating	(base outcome)	
Self_employed		
age	1.437012	.0117132
age2	.9960011	.0000935
married	1.149558	.0321427

- A one-year increase in age leads to relative odds of choosing to be self-employed (dependent variable, `sector=1`) rather than not participating (`sector=0`) that are 1.437 times what they were before the change (one-year younger).
- The original coefficient of age for the alternative self-employed is 0.363, and we have $e^{0.363} \approx 1.437$.

Multinomial Logit

- For an unordered multinomial model, there is no single conditional mean of the dependent variable. Instead, insterest lies in how the probabilities of alternatives change as regressors change.
- For the multinomial model ($p_j(x) = \frac{e^{x\beta_j}}{\sum_{l=1}^{J} e^{x\beta_l}}$), the marginal effects can be shown to be

$$\frac{\partial p_{ij}}{\partial x_i} = p_{ij}(\beta_j - \bar{\beta}_i)$$

  where $\bar{\beta}_i = \sum_k p_{ik}\beta_k$ is a probability weighted average of β_i.
- **The sign of the regression coefficients do not give the signs of the marginal effects.**

Multinomial Logit

- For example, table below gives part of the marginal effects on $\mathbb{P}(sector = 2)$ of a change in the regressors evaluated at the sample mean of them.

```
Average marginal effects                        Number of obs     =    147,843
Model VCE    : OIM

Expression   : Pr(sector==Private_sector_employee), predict(outcome(2))
dy/dx w.r.t. : age age2 married 1.educ_1 2.educ_1 3.educ_1 4.educ_1 black nchild citiz
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0399803	.0006508	61.43	0.000	.0387047	.0412559
age2	-.0005343	7.89e-06	-67.72	0.000	-.0005497	-.0005188
married	-.0796828	.002815	-28.31	0.000	-.0852001	-.0741656

- Being `married` decreases by 0.797 the probability of being in a private sector (2) rather than not participating (0) or being self-employed (1). But if we check the regression output, the parameter estimate for `married` is positve (0.7102, not inlucded in this slides).

Conditional Logit

Data

`TA3_2.dta` (Herriges and Kling, 1999):

- Individuals choose between fishing using one of four possible modes: (1) from the beach, (2) the pier, (3) a private boat, or (4) a charter boat;
- Case-specific regressor: `income`;
- Alternative-specific regressor: price `p` and catch rate `c`.

Conditional Logit

Reshaping data

- In our original wide data, each obervation refers to one individual.

id	mode	pbeach	ppier	pboat	pcharter	cbeach	cpier	cboat	ccharter	income	dbeach	dpier	dboat	dcharter
1	4	157.93	157.93	157.93	182.93	.0678	.0503	.2601	.5391	7083.3317	0	0	0	1
2	4	15.114	15.114	10.534	34.534	.1049	.0451	.1574	.4671	1249.9998	0	0	0	1
3	3	161.874	161.874	24.334	59.334	.5333	.4522	.2413	1.0266	3749.9999	0	0	1	0
4	2	15.134	15.134	55.93	84.93	.0678	.0789	.1643	.5391	2083.3332	0	1	0	0
5	3	106.93	106.93	41.514	71.014	.0678	.0503	.1082	.324	4583.332	0	0	1	0
6	4	192.474	192.474	28.934	63.934	.5333	.4522	.1665	.3975	4583.332	0	0	0	1
7	1	51.934	51.934	191.93	220.93	.0678	.0789	.1643	.5391	8750.001	1	0	0	0
8	4	15.134	15.134	21.714	56.714	.0678	.0789	.0102	.0209	2083.3332	0	0	0	1
9	3	34.914	34.914	34.914	53.414	.2537	.1498	.0233	.0219	3749.9999	0	0	1	0
10	3	28.314	28.314	28.314	46.814	.2537	.1498	.0233	.0219	2916.6666	0	0	1	0
11	2	34.914	34.914	24.334	48.334	.1049	.0451	.1574	.4671	3749.9999	0	1	0	0

- The parameters of conditional logit are estimated with commands that require the data to be in long form, with one observation providing the data for just one alternative for an individual.

Conditional Logit

Reshaping data

- After reshaping, there are now four observations for each individual. One is chosen for that individual ($\mathtt{d} = 1$), the other three alternatives are not chosen but we still have the price and catch rate information of them.
- Price (`p`) and catch rate (`c`) are the two alternative-specific variables, they have different values for different alternatives.
- All case-specific variables appear as a single variable that takes on the same value for the four outcomes. We only have one case-specific variable here: `income`.

id	fishmode	p	c	income	d
1	beach	157.93	.0678	7083.3317	0
1	boat	157.93	.2601	7083.3317	0
1	charter	182.93	.5391	7083.3317	1
1	pier	157.93	.0503	7083.3317	0
2	beach	15.114	.1049	1249.9998	0
2	boat	10.534	.1574	1249.9998	0
2	charter	34.534	.4671	1249.9998	1
2	pier	15.114	.0451	1249.9998	0
3	beach	161.874	.5333	3749.9999	0
3	boat	24.334	.2413	3749.9999	1
3	charter	59.334	1.0266	3749.9999	0
3	pier	161.874	.4522	3749.9999	0

Conditional Logit

Coefficient interpretation

d	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
fishmode						
p	-.0228473	.0018051	-12.66	0.000	-.0263853	-.0193093
c	.2801154	.1352867	2.07	0.038	.0149583	.5452725
beach	(base alternative)					
boat						
income	.0000766	.0000521	1.47	0.142	-.0000256	.0001787
_cons	.5084965	.2421028	2.10	0.036	.0339837	.9830093
charter						
income	-.0000471	.0000523	-0.90	0.368	-.0001496	.0000554
_cons	1.71245	.2400162	7.13	0.000	1.242027	2.182873
pier						
income	-.0001183	.0000536	-2.21	0.027	-.0002234	-.0000133
_cons	.7648191	.2455243	3.12	0.002	.2836003	1.246038

- *Alternative-specific regressors*: The negative coefficient of -0.023 for `p` means that if the price of one mode of fishing increases, then the demand (total number of choices or probability of choosing) for that mode decreases and demand for all other modes increases, as expected.
- *Case-specific regressor*: The three `income` coefficients mean that, relative to the probability of beach fishing (base category), an increase in income has nearly no effect on the probability of choosing other three alternatives.

Conditional Logit

Marginal effects

Pr(choice = beach|1 selected) = .05193131

variable	dp/dx	Std. err.	z	P>|z|	[95% C.I.]		X
p							
beach	-.001125	.000117	-9.62	0.000	-.001354	-.000896	108.02
boat	.000489	.000053	9.17	0.000	.000385	.000594	52.559
charter	.000557	.00006	9.28	0.000	.000439	.000674	81.779
pier	.000079	.000015	5.16	0.000	.000049	.000109	108.02

Pr(choice = charter|1 selected) = .46919337

variable	dp/dx	Std. err.	z	P>|z|	[95% C.I.]		X
p							
beach	.000557	.00006	9.28	0.000	.000439	.000674	108.02
boat	.00442	.000466	9.48	0.000	.003506	.005334	52.559
charter	-.00569	.000459	-12.41	0.000	-.006589	-.004791	81.779
pier	.000713	.00007	10.15	0.000	.000576	.000851	108.02

Pr(choice = boat|1 selected) = .41233945

variable	dp/dx	Std. err.	z	P>|z|	[95% C.I.]		X
p							
beach	.000489	.000053	9.17	0.000	.000385	.000594	108.02
boat	-.005536	.000461	-12.00	0.000	-.00644	-.004632	52.559
charter	.00442	.000466	9.48	0.000	.003506	.005334	81.779
pier	.000627	.000063	10.01	0.000	.000504	.00075	108.02

Pr(choice = pier|1 selected) = .06653588

variable	dp/dx	Std. err.	z	P>|z|	[95% C.I.]		X
p							
beach	.000079	.000015	5.16	0.000	.000049	.000109	108.02
boat	.000627	.000063	10.01	0.000	.000504	.00075	52.559
charter	.000713	.00007	10.15	0.000	.000576	.000851	81.779
pier	-.001419	.000133	-10.66	0.000	-.00168	-.001158	108.02

- For each regressor (here we take p for example), 16 marginal effects are reported (response probabilities for four modes $\times$ p for four modes).
- All own effects are negative and all cross effects are positive (we have just explained the reason: demand).

Conditional Logit

Marginal effects

Pr(choice = beach|1 selected) = .05193131

variable	dp/dx	Std. err.	z	P>|z|	[95% C.I.	]	X
p							
beach	-.001125	.000117	-9.62	0.000	-.001354	-.000896	108.02
boat	.000489	.000053	9.17	0.000	.000385	.000594	52.559
charter	.000557	.00006	9.28	0.000	.000439	.000674	81.779
pier	.000079	.000015	5.16	0.000	.000049	.000109	108.02

- The first effect value given in the output is - 0.001125, a one dollar increase in the price of beach fishing decreases the probability of beach fishing by 0.001125, with price and income set to sample means.

Nested Logit

Independence of Irrelevant Alternatives (IIA)

- The IIA condition means that the ratio of the probability of selecting train to that of selecting car is unaffected by the price of an airplane ticket.
- This may make sense if individuals view the set of choices as similarly substitutable, but does not make sense if train and air are close substitutes.
- The multinomial logit (MNL) and conditional logit (CL) models have the IIA property, they impose the restriction that the choice between any two pairs of alternatives is simply a binary logit model (errors ε_{ij} in their random utility models are i.i.d).
- Try to think about: is the odds ratio still informative if IIA is violated?
- Nested logit (NL) is one of the most tractable models that allow for correlated errors.

Nested Logit

Tree structure

- The NL model requires a nesting structure that splits the alternatives into groups, where errors are correlated within group but uncorrelated across group.
- In our fishing example, we specify a two-level NL model, assume a fundamental distinction between shore and boat fishing:

```
                Mode
              /      \
          Shore      boat            (level 1)
         /     \    /     \
    Beach    Pier  Charter   Pivate   (level 2)
```

- level 1 (a limb): shore/boat contrast; level 2 (a branch): the next level.
- NL model permits correlation of errors within each of the level 2 groupings, whereas the two pairs $(\varepsilon_{i,beach}, \varepsilon_{i,pier})$ and $(\varepsilon_{i,private}, \varepsilon_{i,charter})$ are independent.
- The CL model is a special case of NL, while the MNL and nested logit are special cases of CL.

Nested Logit

Tree structure

- Tree stucture in Stata:

```
tree structure specified for the nested logit model

 type    N        fishmode  N     k
----------------------------------------
 coast 2000 ─┬─ beach     1000  107
             └─ pier      1000  143
 water 2000 ─┬─ boat      1000  355
             └─ charter   1000  395
----------------------------------------
                  total   4000 1000

k = number of times alternative is chosen
N = number of observations at each level
```

- **Predicted probabilities**:

fishmode	Summary of Pr(fishmode alternatives) Mean	Std. Dev.	Freq.
beach	.12074824	.14489275	1,000
boat	.3469483	.14423437	1,000
charter	.40304971	.16979586	1,000
pier	.12925375	.15864406	1,000
Total	.25	.19990564	4,000

- The average predicted probabilities for NL are quite close to the sample probabilities.

Nested Logit

Marginal effects of $\texttt{p}$ on $\mathbb{P}(\texttt{fishmode} == \texttt{beach})$:

fishmode	Summary of dpdbeach Mean	Std. Dev.	Freq.
beach	-.00089689	.00088494	1,000
boat	.00081157	.00081603	1,000
charter	.00091984	.00091461	1,000
pier	-.00083453	.00084101	1,000
Total	-2.747e-09	.00122443	4,000

Figure 1: ME from NL

Pr(choice = beach|1 selected) = .05193131

variable	dp/dx	Std. err.	z	P>|z|	[95% C.I.	]	X
p							
beach	-.001125	.000117	-9.62	0.000	-.001354	-.000896	108.02
boat	.000489	.000053	9.17	0.000	.000385	.000594	52.559
charter	.000557	.00006	9.28	0.000	.000439	.000674	81.779
pier	.000079	.000015	5.16	0.000	.000049	.000109	108.02

Figure 2: ME from CL

- Compare to CL, the probability of pier fishing falls in addition to the probability of beach fishing (due to the correlated errors within one limb).

Comparison of Multinomial models

Variable	MNL	CL	NL
p		-0.025	-0.027
q		0.358	1.347
N	1182	4728	4728
ll	-1477	-1215	-1192
aic	2966	2446	2405
bic	2997	2498	2469

- For the information criteria, low values are preferred (lower AIC or BIC means higher log-likelihood with penalties for model size being considered). MNL is least preferred, and NL is most preferred.

Commands

Model	Stata Commands	R packages[1]
Multinomial logit	mlogit	mlogit, nnet
Conditional logit	clogit, asclogit, cmclogit	survival
Nested logit	nlogit	mlogit
Mixed logit	mixlogit, asclogit	mlogit
Multinomial probit	mprobit, asmprobit	mlogit
Ordered outcome models	ologit, oprobit	MASS, erer, oglmx
Marginal effects	Margins, mfx	margins, mfx

[1] These are just as far as I know and may not be the best options, please check before using. You can always call Stata from R (RStata, or Statamarkdown if R Markdown), Python (PyStata, works with IPyhon or Python shell). Students who use open-source languages to solve the problem sets may sometimes have to put in extra effort and for this reason will get a small bonus from me (in the grade for that particular PS, and I need to see the extra effort from your solution).

Appendix

Binary outcome

MLE

▸ Back to Predicted Probabilities

$$\mathcal{L}_N = \sum_i \{y_i \ln F(X\beta) + (1-y_i)\ln[1-F(X\beta)]\}$$

$$\text{FOC wrt } \beta: \quad \frac{\partial \mathcal{L}_N}{\partial \beta} = \sum_i \left\{ y_i \frac{f(X\beta)}{F(X\beta)} X_i + (1-y_i)\frac{-f(X\beta)}{1-F(X\beta)} X_i \right\}$$

$$= \sum_i \left\{ \frac{y_i f(X\beta)[1-F(X\beta)] - (1-y_i) f(X\beta) F(X\beta)}{F(X\beta)[1-F(X\beta)]} X_i \right\}$$

$$= \sum_i \left\{ \frac{[y_i - F(X\beta)] f(X\beta)}{F(X\beta)[1-F(X\beta)]} X_i \right\}$$

$$= \frac{f(X\beta)}{F(X\beta)[1-F(X\beta)]} \sum_i \{[y_i - F(X\beta)] X_i\}$$

$$= 0$$

$$\Rightarrow \sum \{[y_i - F(X\beta)] X_i\} = 0$$

References

- Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics: methods and applications. Cambridge university press. Chapter 14-15.
- Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press. Chapters 15-16.
- Hansen, B. E. (2022). Econometrics. Chapter 25-26.
- Cameron, A. C., & Trivedi, P. K. (2022). Microeconometrics using stata (Second Edition). Stata press. Chapters 17-18.