Introduction to Statistical Methods
SOC-GA 2332

Lecture 10: Categorical Data Analysis

Siwei Cheng

NEW YORK UNIVERSITY

Association between categorical data (based on contingency tables)

Contingency Table

- When we are interested in the association between two categorical variables, we describe their distribution using a **contingency table**.

Education	Attitude toward premarital sex		
	Disapproval	Approval	Total
High school or less	800	1200	2000
College and above	500	1000	1500
Total	1300	2200	3500

Contingency Table - Three Types of Distributions

- Joint distribution
- Marginal distribution
- Conditional distribution

Contingency Table - Joint Distribution

Education	Attitude toward premarital sex		
	Disapproval	Approval	Total
High school or less	800 **(22.86%)**	1200 **(34.29%)**	2000
College and above	500 **(14.29%)**	1000 **(28.57%)**	1500
Total	1300	2200	3500

Contingency Table - Marginal Distribution

Education	Attitude toward premarital sex		
	Disapproval	Approval	Total
High school or less	800	1200	2000 **(57.14%)**
College and above	500	1000	1500 **(42.86%)**
Total	1300	2200	3500
	37.14%	**62.86%**	**100%**

Contingency Table - Conditional Distribution

Education	Attitude toward premarital sex: Disapproval	Approval	Total
High school or less	800 **(40%)**	1200 **(60%)**	2000 **(100%)**
College and above	500 **(33.33%)**	1000 **(66.67%)**	1500 **(100%)**
Total	1300	2200	3500

Education	Attitude toward premarital sex: Disapproval	Approval	Total
High school or less	800 **(61.54%)**	1200 **(54.55%)**	2000
College and above	500 **(38.46%)**	1000 **(45.45%)**	1500
Total	1300 **(100%)**	2200 **(100%)**	3500

Contingency Table - Conditional Distribution

- Is the attitude towards premarital sex associated with educational attainment?
- When we look at the conditional probability of "approval" in each educational group, we will see that those with a college degree are more likely to hold a positive attitude towards premarital sex than those with high school or less.
- But, is this a statistically significant association?

	Attitude toward premarital sex		
Education	Disapproval	Approval	Total
High school or less	800 **(40%)**	1200 **(60%)**	2000 **(100%)**
College and above	500 **(33.33%)**	1000 **(66.67%)**	1500 **(100%)**
Total	1300	2200	3500

Statistical Independence and Chi-squared Test

- Two categorical variables are statistically independent if the population conditional distributions on one of them are identical at each category of the other.
- Chi-squared Test of Independence tests the following:
- H_0 The variables are independent;
- H_1 The variables are not independent.

Observed and Expected Frequencies

Education	Attitude toward premarital sex: Disapproval	Approval	Total
High school or less	800	1200	2000 **(57.14%)**
College and above	500	1000	1500 **(42.86%)**
Total	1300	2200	3500
	37.14%	**62.86%**	**100%**

- f^o is the *observed* frequency.
- f^e is the *expected* frequency - the count expected in a cell if the variables were independent (i.e. if H_0 is true).

Observed and Expected Frequencies

Education	Attitude toward premarital sex: Disapproval	Approval	Total
High school or less	f_{11}	f_{12}	$f_{1,}$
College and above	f_{21}	f_{22}	$f_{2,}$
Total	$f_{,1}$	$f_{,2}$	N=3500

- $f_{11}^o = 800$
- $f_{11}^e = N \cdot f_{1,} \cdot f_{,1} = 3500 \cdot 57.14\% \cdot 37.14\%$

Statistical Independence and Chi-squared Test

- χ^2 statistic:

$$\chi^2 = \Sigma \frac{(f^o - f^e)^2}{f^e}$$

Education	Attitude toward premarital sex: Disapproval	Approval	Total
High school or less	800 742.76	1200 1257.14	2000 **(57.14%)**
College and above	500 557.14	1000 942.96	1500 **(42.86%)**
Total	1300 **37.14%**	2200 **62.86%**	3500 **100%**

Statistical Independence and Chi-squared Test

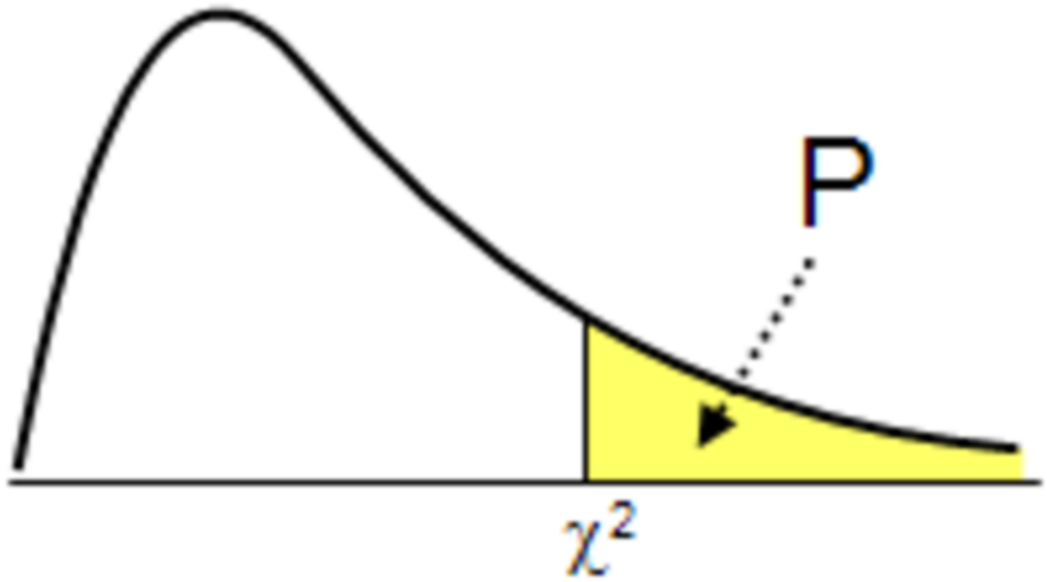
- χ^2 statistic:

$$\chi^2 = \Sigma \frac{(f^o - f^e)^2}{f^e} = \mathbf{16.32}$$

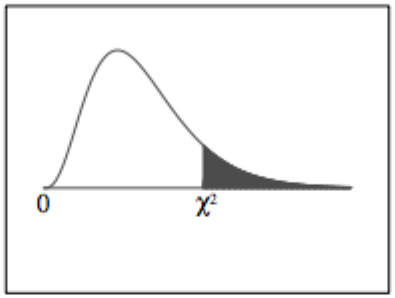
Education	Attitude toward premarital sex: Disapproval	Approval	Total
High school or less	800 742.76	1200 1257.14	2000 **(57.14%)**
College and above	500 557.14	1000 942.96	1500 **(42.86%)**
Total	1300 **37.14%**	2200 **62.86%**	3500 **100%**

Chi-squared Statistics

- χ^2 degree of freedom: $df = (r-1) \cdot (c-1)$



Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156

Regression Models with Categorical Dependent Variables

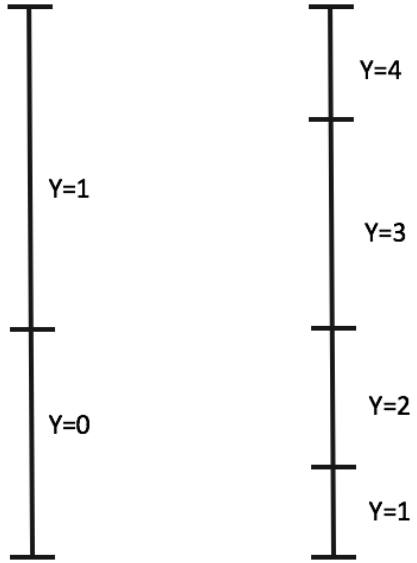
Regression with Ordinal Dependent Variables

- A quick review on logistic regression model we discussed last week:

$$logit(p_i) = log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \quad (1)$$

- Many social phenomena of interest are measured by discrete variables with more than two categories. If the categories are **ordered**, the variable is called an *ordinal* dependent variable; if the categories are **unordered**, the variable is called a *nominal* dependent variable.

Regression with Ordinal Dependent Variables



- Note that an ordinal variable requires a rank order of the categories, but it does not require that the *distance* between adjacent categories to be the same.
- Examples include the Likert scale (strongly disagree, disagree, neutral, agree, strongly agree) or education level (less than 8 years, 9-11 years, 12 years, and 13 or more years).

Scoring Methods

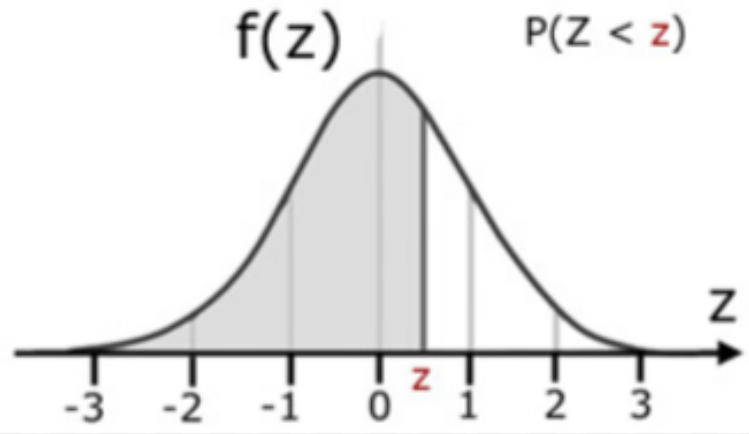
- How to deal with ordinal dependent variables?
- Let's start with **Scoring Methods**, which assign scores to ordinal categories, and then treat the scores as continuous variables.
- Integer Scoring: we can assign integers to represent the rank order. e.g. Likert-scaled (1=strongly disagree, 2= disagree, 3=neutral, 4=agree, 5=strongly agree). (what's the underlying assumption?)

Scoring Methods

- Midpoint Scoring: some ordinal variables are based on categorical measures of variables that are conceptually continuous.
- For example, we may collapse the years of schooling variable into a discrete variable with four categories: Less than high school (0-11 years); High school (exactly 12 years); Some College (13-15 years); and college and beyond (16+ years). We can use the "midpoints" 5.5, 12, 14, and 18 as the score assigned to each category respectively. (what's the underlying assumption?)

Scoring Methods

- Normal Score Transformation: We assume that the categorical response is based on a normally-distributed response variable. This method takes four steps: (1) calculate the sample proportion in each category; (2) accumulate proportions across categories; (3) for each category, find the cumulative proportion that corresponse to its midpoint (50th percentile within category); and (4) transform the cumulative proportion of the midpoint to a z-score based on the standardized normal distribution.



Scoring Methods

- Let's look at an example in Chapter 7 (Table 7.1) in Powers and Xie (2000), which creates z-scores based on the attitudes towards premarital sex:

	Aways	**Almost Always**	**Sometimes**	**Not at all**
Frequency	1020	386	825	1573
Proportion	0.268	0.101	0.217	0.414
Cumulative prop	0.268	0.370	0.586	1
Mid-point	0.134	0.319	0.478	0.793
Z-score	-1.107	-0.471	-0.055	0.818

Scoring Methods

- Let's look at an example in Chapter 7 (Table 7.1) in Powers and Xie (2000), which creates z-scores based on the attitudes towards premarital sex:

	Aways	**Almost Always**	**Sometimes**	**Not at all**
Frequency	1020	386	825	1573
Proportion	0.268	0.101	0.217	0.414
Cumulative prop	0.268	0.370	0.586	1
Mid-point	0.134	0.319	0.478	0.793
Z-score	-1.107	-0.471	-0.055	0.818

- Scaling with Additional Information: e.g. construct socio-economic index for different occupational/class categories. The index may be based on earnings/educational level/prestige.

Ordered Logit Regression Models

Ordered Logit Regression Models

- Another important way to deal with ordinal dependent variables is the **Ordered Logit Regression Models**.
- Let's begin by introducing the cumulative probability for individual i's choice up to response level j:

$$C_{i,j} = Pr(y_i \leq j) = \Sigma_{k=1}^{j} Pr(y_i = k) = \frac{1}{1 + exp(-\phi_j + x_i\beta)}$$
$$j = 1, 2, ...J. \tag{2}$$

- By definition, the cumulative probabilities must sum to one when j=J, meaning that $C_{i,J}$=1.0 for all i. This constraint implies that only $J - 1$ cumulative probabilities are uniquely identified.
- Also note that defined this way, $C_{i.j} > C_{i,j-1}$.

Ordered Logit Regression Models

- In Equation (2), ϕ_1, ϕ_2,...ϕ_J are **cutpoints** corresponding to each response category. We also set ϕ_0 to be $-\infty$ and ϕ_J to be ∞. If ϕ_j is very small, then $C_{i,j}$ approaches zero, and thus it is easy to get past the "threshold" if being in this category or higher.
- Importantly, β captures the effect of person-level attribute X on the likelihood of moving to the "higher rank."
- To see this, compare the **odds** of having response $Pr(y_i > j)$

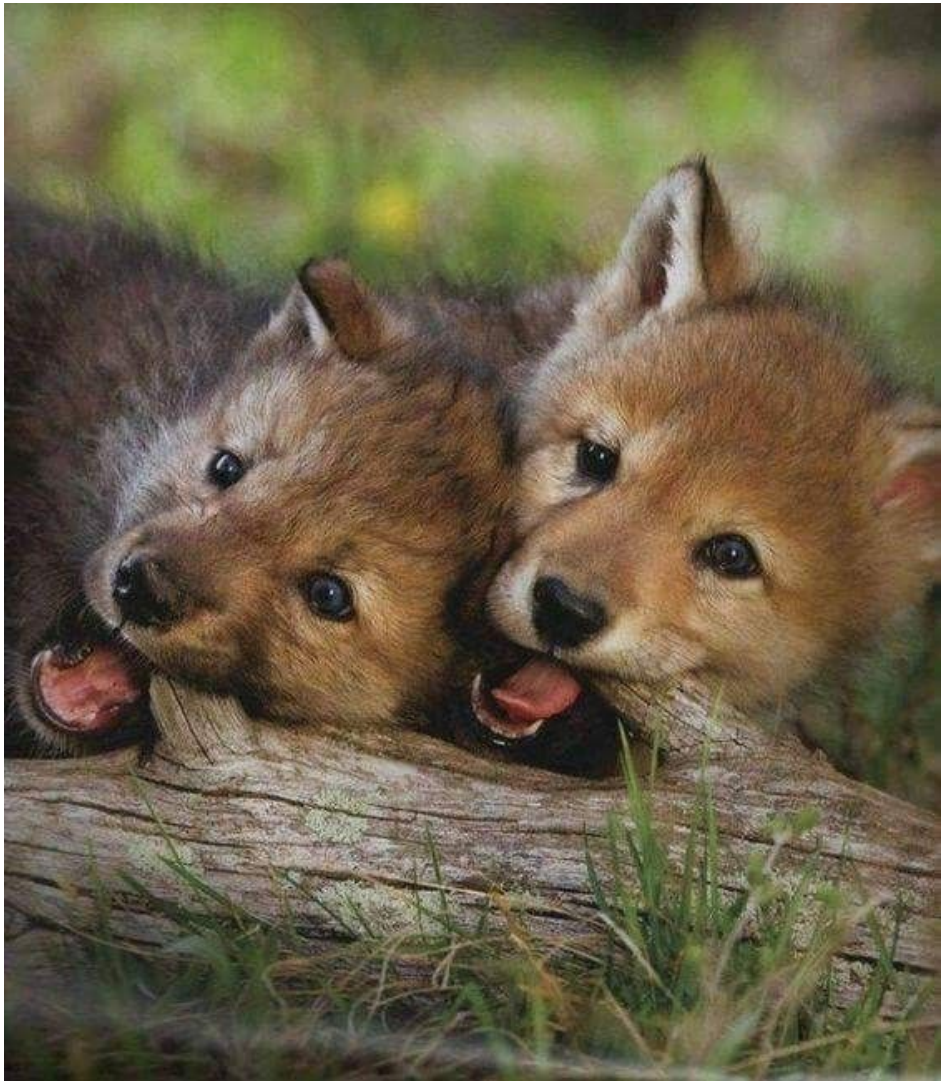
$$\frac{Pr(y_i > j|x_1)/Pr(y_i \leq j|x_1)}{Pr(y_i > j|x_2)/Pr(y_i \leq j|x_2)} = \frac{exp(-\phi_j + x_1\beta)}{exp(-\phi_j + x_2\beta)} \quad (3)$$
$$= exp\{(x_1 - x_2)\beta\}$$

Ordered Logit Regression Models

- Given Equation (2), we can write the probability of being in response category j as:

$$C_{i,j} = Pr(y_i = j) = \frac{1}{1 + exp(-\phi_j + x_i\beta)} - \frac{1}{1 + exp(-\phi_{j-1} + x_i\beta)} \quad (4)$$

- The estimation of the ordered logit model is based on the probability implied by Equation (4).



Regression with Nominal Dependent Variables

Regression with Nominal Dependent Variables

- In the case of a multi-category discrete dependent variable with unordered categories, we can use the **Multinomial Logit Model** to predict **the probability of a response falling into a certain category**.
- Example research questions: How does education level affect a person's party affiliation (Republican, Democratic, and Independent)? How does gender affect workers' choice of employment sector (public, private, other)? How does one's income affect one's mode of transportation (train, bus, car, other)?

Multinomial Logit Model

- Let's consider an example with three response categories.
- Because the three probabilities should sum to 1.0, only two of the probabilities will be identified. So we need to set one category as the "baseline" or "reference" category.
- The Multinomial Logit Model writes the probability of falling into each category as:

$$P_{i1} = Pr(y_i = 1|x_i) = \frac{1}{1 + exp(\beta_2 x_i) + exp(\beta_3 x_i)} \quad (5)$$

$$P_{i2} = Pr(y_i = 2|x_i) = \frac{exp(\beta_2 x_i)}{1 + exp(\beta_2 x_i) + exp(\beta_3 x_i)} \quad (6)$$

$$P_{i3} = Pr(y_i = 3|x_i) = \frac{exp(\beta_3 x_i)}{1 + exp(\beta_2 x_i) + exp(\beta_3 x_i)} \quad (7)$$

Multinomial Logit Model

$$P_{i1} = Pr(y_i = 1|x_i) = \frac{1}{1 + exp(\beta_2 x_i) + exp(\beta_3 x_i)} \quad (8)$$

$$P_{i2} = Pr(y_i = 2|x_i) = \frac{exp(\beta_2 x_i)}{1 + exp(\beta_2 x_i) + exp(\beta_3 x_i)} \quad (9)$$

$$P_{i3} = Pr(y_i = 3|x_i) = \frac{exp(\beta_3 x_i)}{1 + exp(\beta_2 x_i) + exp(\beta_3 x_i)} \quad (10)$$

- How to interpret β_2 and β_3?
- β_2 represents the effect of variable X on the likelihood of choosing category 2 as opposed to category 1; β_3 represents the effect of variable X on the likelihood of choosing category 3 as opposed to category 1.

Reference

Multinomial Logit Model

- In fact, β_2 is the log odds ratio between category 2 and category 1. To see this, take the ratio between Equation (6) and (5):

$$\frac{Pr(y_i = 2|x_i)}{Pr(y_i = 1|x_i)} = exp(\beta_2 x_i) \tag{11}$$

$$\text{Hence:} \quad \beta_2 x_i = log \frac{Pr(y_i = 2|x_i)}{Pr(y_i = 1|x_i)} \tag{12}$$

- For example, if our dependent variable Y is employment sector, with Y=1 denoting public sector, Y=2 denoting private sector, and Y=3 denoting other. And let our key independent variable be X years of schooling. Then, β_2 gives us the effect of years of schooling on the log odds ratio between working in the private sector and working in the public sector.

Multinomial Logit: Example

Table 4 Multilevel multinomial logit model of cohabitation and marriage at ages 40–49: Men in 25 European countries in 2002–2010

	Cohabiting[a]				Married[a]			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Intercept	.428	.362	.405	.370	.672*	.664*	.698*	.668*
	(.573)	(.530)	(.527)	(.538)	(.320)	(.334)	(.336)	(.333)
× Gender roles		.576*	.678*	.385*		−.216*	−.138*	−.210*
		(.095)	(.116)	(.098)		(.049)	(.058)	(.080)
× Inequality		−.008	−.100	−.043		−.135*	−.217*	−.134†
		(.108)	(.119)	(.083)		(.064)	(.083)	(.067)
× GDP			−.206				−.168	
			(.119)				(.102)	
× Divorce rate				.278*				−.012
				(.066)				(.092)
Age	−.036*	−.036*	−.037*	−.036*	−.001	−.001	−.001	−.001
	(.012)	(.012)	(.012)	(.012)	(.007)	(.007)	(.007)	(.007)
Year of Birth	.039	.045	.036	.043	−.186*	−.186*	−.194*	−.187*
	(.069)	(.070)	(.069)	(.070)	(.045)	(.044)	(.043)	(.044)
Immigrant	−.143	−.148	−.144	−.147	.206	.207	.211	.208
	(.126)	(.125)	(.127)	(.126)	(.107)	(.108)	(.109)	(.108)
Religious Background	−.265*	−.263*	−.251*	−.253*	.330*	.331*	.340*	.332*

NOTE: reference category: **single**. Table from Kalmijn 2013

Conditional Logit Model

- The multinomial logit model introduced above focuses on the characteristics of the individual.
- What if we are interested in the effect of the **characteristics of the potential choices (i.e. response categories)**, rather than the effect of the attributes of the individuals? Another set of models, namely the **Conditional Logit Model**, addresses this type of questions.
- In the Conditional Logit Model, explanatory variables can be the characteristics of the response categories, and the effect of these characteristics on the choice probability can also depend on the individual's characteristics.

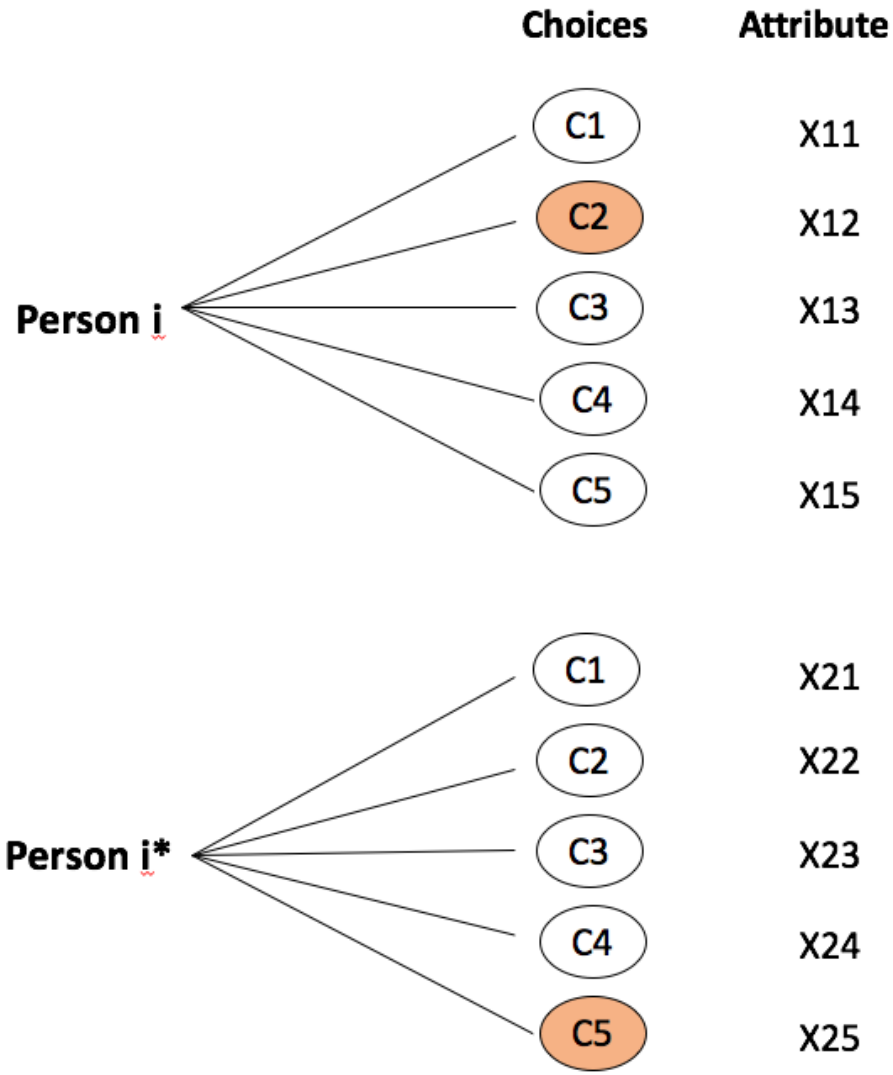
Conditional Logit Model

- For subject i and response choice j. Suppose there are Q possible choices. Let X_{ij} denote the explanatory variables (which may depend on characteristics of both i and j.). The probability of person i selecting option j is:

$$\pi_{ij} = \frac{exp(\beta^T X_{ij})}{\sum_{q=1}^{Q} exp(\beta^T X_{iq})} \tag{13}$$

- For example, what factors individuals consider in making their neighborhood choices? Here, all candidate neighborhoods are the possible choices, and characteristics of these neighborhoods (poverty level, racial composition, etc) can be explanatory variables in the Conditional Logit Models.
- We can also specify the model so that the effect of these neighborhood characteristics on choice probability depends on individual-level characteristics. That is, there can be interactions between the characteristics of the **chooser** and those of the **choices**.

Conditional Logit Model



Independence from Irrelevant Alternatives Assumption in CLM

- The coniditional logit model requires an important assumption (Independence from Irrelevant Alternatives or IIA), which states that the odds of choosing alternative j over alternative k should be independent of the choice set for all pairs j,k.
- Whether or not this assumption is reasonable depends on the context.
- For example, suppose individuals are choosing between red bus and train for transportation, with the odds being 1:1. Now, suppose another alternative is added: blue bus. Suppose that individuals who choose bus are indifferent to the color of the bus, and that individuals' preference for bus versus train is unchanged. In this case, adding in blue bus will make the odds of choosing a red bus and train 1:2, which is equal to the odds of choosing a blue bus and train.

Quick Recap on What We have Covered So Far