Choice models using Stata

Joerg Luedicke

StataCorp LLC

October 29, 2019

What are choice models? (1)

- With discrete choice models, we describe decision makers' choices among a given set of alternatives
- Discrete choice models are used across disciplines to analyze choice behavior, for example:
  - Companies choose whether to use TV, print, or Internet advertising
  - Individuals choose their favorite breakfast cereal
  - Voters choose their favorite candidate or party
  - Individuals choose among long-term care options such as nursing home, assisted living, moving in with family
  - Individuals choose among educational programs
- In all of these cases, we observe decision making entities that are faced with a set of alternatives to choose from

What are choice models? (2)

- The set of alternatives is called a choice set, which ...
  - ... consists of mutually exclusive alternatives
  - ... includes all possible alternatives, i.e. is exhaustive
  - ... contains a finite number of alternatives
- If we have a discrete choice model that allows for including variables that can vary both over decision makers as well as alternatives, we speak of **discrete choice models with alternative-specific variables**. This is what we are referring to when we speak of choice models in Stata.
- In other words, we can incorporate attributes of the decision maker as well as attributes of the alternatives into our analysis

Discrete choice analysis with alternative-specific variables (cross-sectional data)

```
. webuse transport
(Transportation choice data)
. list id alt choice trcost trtime age income if t==1 & id < 3, sepby(id)
```

id	alt	choice	trcost	trtime	age	income
1	Car	1	4.14	0.13	3.0	3
1	Public	0	4.74	0.42	3.0	3
1	Bicycle	0	2.76	0.36	3.0	3
1	Walk	0	0.92	0.13	3.0	3
2	Car	0	4.36	0.23	3.0	2
2	Public	0	4.43	0.43	3.0	2
2	Bicycle	0	1.25	1.23	3.0	2
2	Walk	1	0.89	0.12	3.0	2

Discrete choice analysis with alternative-specific variables (panel data)

```
. list id t alt choice trcost trtime age income in 1/12, sepby(t) noobs
```

id	t	alt	choice	trcost	trtime	age	income
1	1	Car	1	4.14	0.13	3.0	3
1	1	Public	0	4.74	0.42	3.0	3
1	1	Bicycle	0	2.76	0.36	3.0	3
1	1	Walk	0	0.92	0.13	3.0	3
1	2	Car	1	8.00	0.14	3.2	5
1	2	Public	0	3.14	0.12	3.2	5
1	2	Bicycle	0	2.56	0.18	3.2	5
1	2	Walk	0	0.64	0.39	3.2	5
1	3	Car	1	1.76	0.18	3.4	5
1	3	Public	0	2.25	0.50	3.4	5
1	3	Bicycle	0	0.92	1.05	3.4	5
1	3	Walk	0	0.58	0.59	3.4	5

Examples of things we want to learn from discrete choice analyses

- How does the probability of choosing public transportation change if yearly income increases from $30,000 to $40,000?
- How does travel time and cost affect the probability of choosing each transportation mode?
- If travel cost related to car travel increases, how does that affect the probability of using a car?
- If travel time is increasing for public transportation, how does that affect the probability of choosing car travel?
- If we are public administrators and wish to reduce car travel in our metropolitan area during rush hours by ten percentage points, how strong do we need to incentivize public transportation?

Some estimation results from a discrete choice model

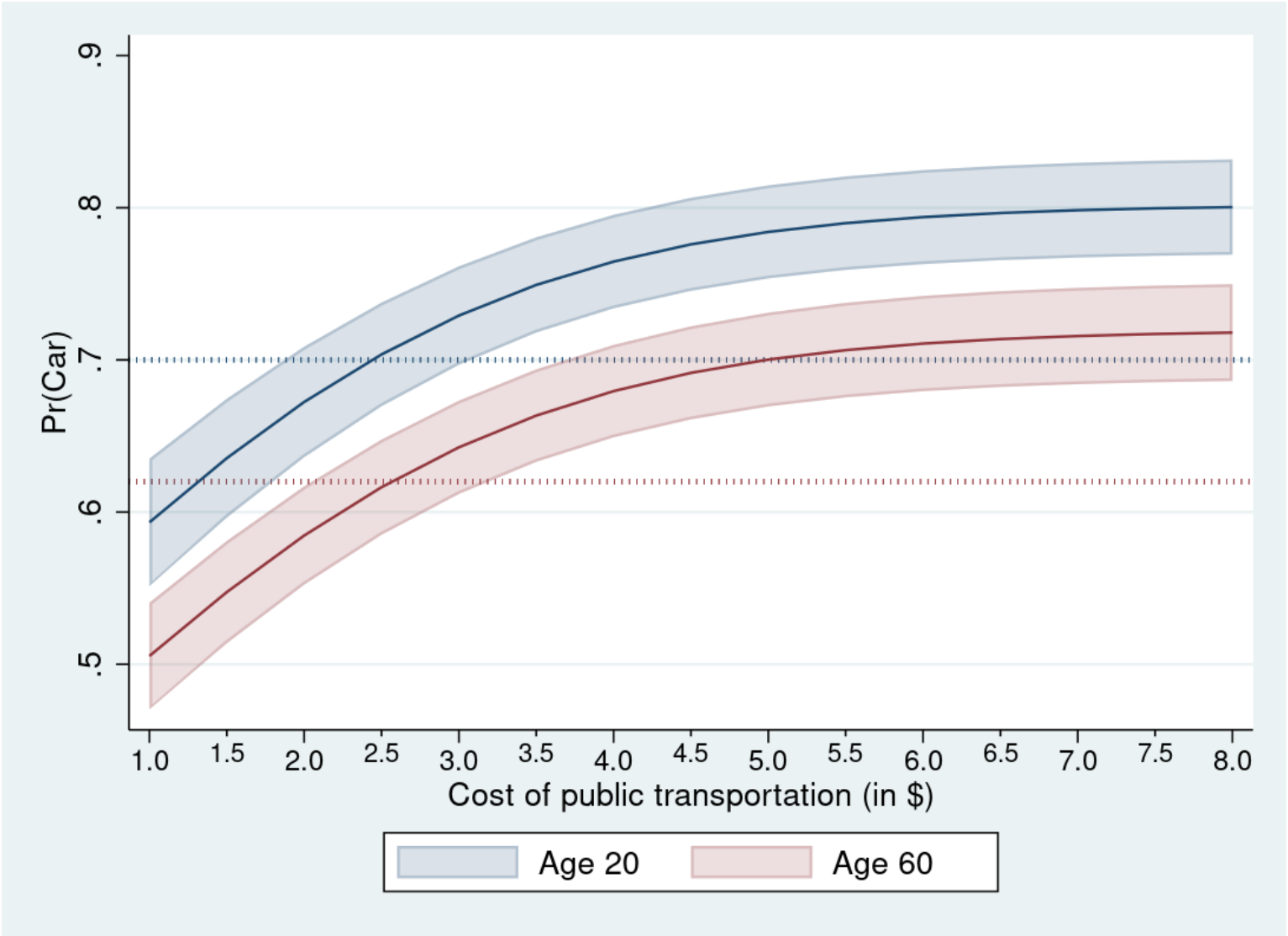
```
<snip>
```

choice	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
alt						
trcost	-.8388216	.0438587	-19.13	0.000	-.9247829	-.7528602
trtime	-1.508756	.2641554	-5.71	0.000	-2.026492	-.9910212

```
<snip>
```

- If cost and time of travel increases for a given alternative, the probability of choosing that alternative decreases.
- While a result like this may provide some information, it is not a lot!
- In Stata 16, we can now use **`margins`** not only to discover more interesting results, but also results that are easier to quantify.

Probability of choosing car as a function of public transportation cost



Discrete choice estimators in Stata 16

cm commands in Stata 16:

- **cmclogit** (formerly `asclogit`)
- **cmmprobit** (formerly `asmprobit`)
- **cmroprobit** (formerly `asroprobit`)
- **cmrologit** (formerly `rologit`)
- **cmmixlogit** (formerly `asmixlogit`)
- **cmxtmixlogit** (new in Stata 16)

All **cm** commands now support **margins**

New **[CM]** manual

Set-up and utility commands: `cmset`, `cmchoiceset`, `cmsample`, `cmsummarize`, `cmtab`

Other discrete choice estimators:

- `nlogit`, `mlogit`, `mprobit`, `logit`, `probit`, ...

cmxtmixlogit

- Works with **margins**
- Random coefficient distributions $f(\beta)$:
  - (multivariate) normal
  - lognormal
  - truncated normal
  - uniform
  - triangle
- Estimates the parameters of the mixed logit model by **maximum simulated likelihood**
- Halton, Hammersley, and pseudo-random draws with uni- and multidimensional **antithetics**
- Full support of **factor variables** and **time-series operators**
- Support of complex **survey** data
- Case-specific variables

cmset – declaring cm data

```
. cmset id t alt
panel data: panels id and time t
note: case identifier _caseid generated from id t
note: panel by alternatives identifier _panelaltid generated from id alt

                     caseid variable:  _caseid
               alternatives variable:  alt
      panel by alternatives variable:  _panelaltid (strongly balanced)
                       time variable:  t, 1 to 3
                               delta:  1 unit

note: data have been xtset
```

Panel-data mixed logit model using `cmxtmixlogit` (1)

```
. cmxtmixlogit choice trcost, random(trtime) casevars(age income) nolog
Mixed logit choice model                    Number of obs        =      6,000
                                            Number of cases      =      1,500
Panel variable: id                          Number of panels     =        500
Time variable: t                            Cases per panel: min =          3
                                                             avg =        3.0
                                                             max =          3
Alternatives variable: alt                  Alts per case:   min =          4
                                                             avg =        4.0
                                                             max =          4
Integration sequence:      Hammersley
Integration points:              594        Wald chi2(8)         =     432.68
Log simulated likelihood = -1005.9899       Prob > chi2          =     0.0000
------------------------------------------------------------------------------
      choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
<snip>
```

Panel-data mixed logit model using `cmxtmixlogit` (2)

```
<snip>
```

choice	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
alt						
trcost	-.8388216	.0438587	-19.13	0.000	-.9247829	-.7528602
trtime	-1.508756	.2641554	-5.71	0.000	-2.026492	-.9910212
/Normal						
sd(trtime)	1.945596	.2594145			1.498161	2.526661
Car	(base alternative)					

```
<snip>
```

Panel-data mixed logit model using `cmxtmixlogit` (3)

<snip>

Car	(base alternative)					
Public						
age	.1538915	.0672638	2.29	0.022	.0220569	.2857261
income	-.3815444	.0347459	-10.98	0.000	-.4496451	-.3134437
_cons	-.5756547	.3515763	-1.64	0.102	-1.264732	.1134222
Bicycle						
age	.20638	.0847655	2.43	0.015	.0402426	.3725174
income	-.5225054	.0463235	-11.28	0.000	-.6132978	-.4317131
_cons	-1.137393	.4461318	-2.55	0.011	-2.011795	-.2629909
Walk						
age	.3097417	.1069941	2.89	0.004	.1000372	.5194463
income	-.9016697	.0686042	-13.14	0.000	-1.036132	-.7672078
_cons	-.4183279	.5607111	-0.75	0.456	-1.517302	.6806458

What would be the expected choice probabilities if every person in the population had a yearly income of $30,000?

```
. margins, at(income=3)
Predictive margins                              Number of obs     =      6,000
Model VCE    : OIM
Expression   : Pr(alt), predict()
at           : income          =           3
```

	Delta-method				
	Margin	Std. Err.	z	P>|z|	[95% Conf. Interval]
_outcome					
Car	.3331611	.0196734	16.93	0.000	.294602 .3717203
Public	.2210964	.0184285	12.00	0.000	.1849772 .2572156
Bicycle	.1676081	.0181511	9.23	0.000	.1320325 .2031837
Walk	.2781343	.0243791	11.41	0.000	.2303521 .3259166

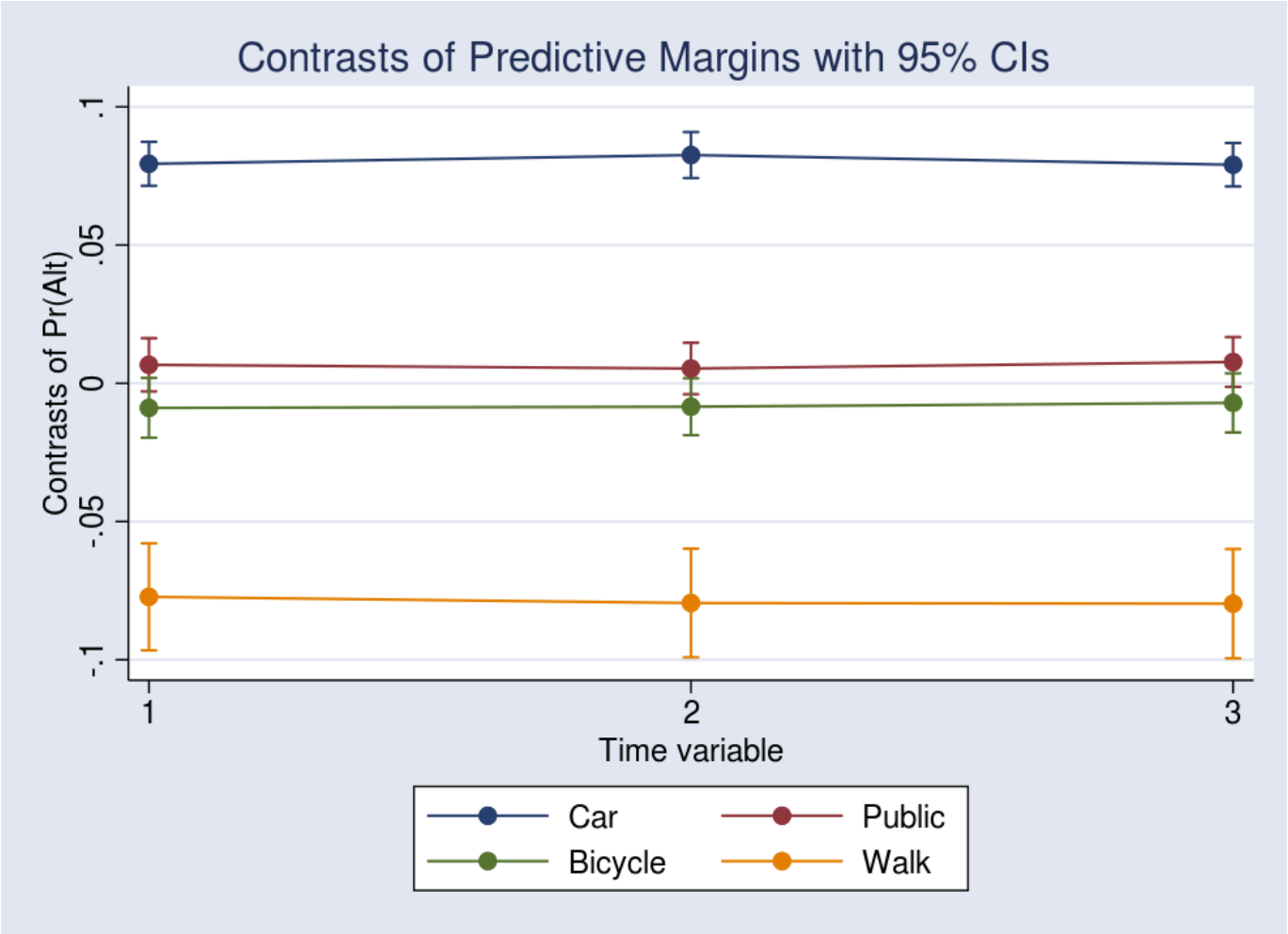
What would be the differences between an income of $40,000 and $30,000 over time?

```
. margins, at(income=(3 4)) contrast(at(r) nowald) over(t)
Contrasts of predictive margins                 Number of obs     =      6,000
Model VCE    : OIM

Expression   : Pr(alt), predict()
over         : t
1._at        : 1.t
                    income          =           3
1._at        : 2.t
                    income          =           3
1._at        : 3.t
                    income          =           3
2._at        : 1.t
                    income          =           4
2._at        : 2.t
                    income          =           4
2._at        : 3.t
                    income          =           4
```

	Contrast	Delta-method Std. Err.	[95% Conf.	Interval]
_at@_outcome#t				
(2 vs 1) Car#1	.0793997	.0040536	.0714548	.0873446
(2 vs 1) Car#2	.0825786	.0042477	.0742532	.090904
(2 vs 1) Car#3	.0790618	.0040101	.0712022	.0869214
(2 vs 1) Public#1	.0066981	.0049098	-.002925	.0163212
(2 vs 1) Public#2	.0053644	.00474	-.0039258	.0146547
(2 vs 1) Public#3	.0077187	.0046076	-.0013121	.0167495
(2 vs 1) Bicycle#1	-.0088805	.0055205	-.0197005	.0019396
(2 vs 1) Bicycle#2	-.0084672	.0052449	-.018747	.0018126
(2 vs 1) Bicycle#3	-.0070729	.0054537	-.017762	.0036161
(2 vs 1) Walk#1	-.0772173	.0098791	-.09658	-.0578546
(2 vs 1) Walk#2	-.0794758	.0100246	-.0991236	-.059828
(2 vs 1) Walk#3	-.0797076	.0100757	-.0994556	-.0599596

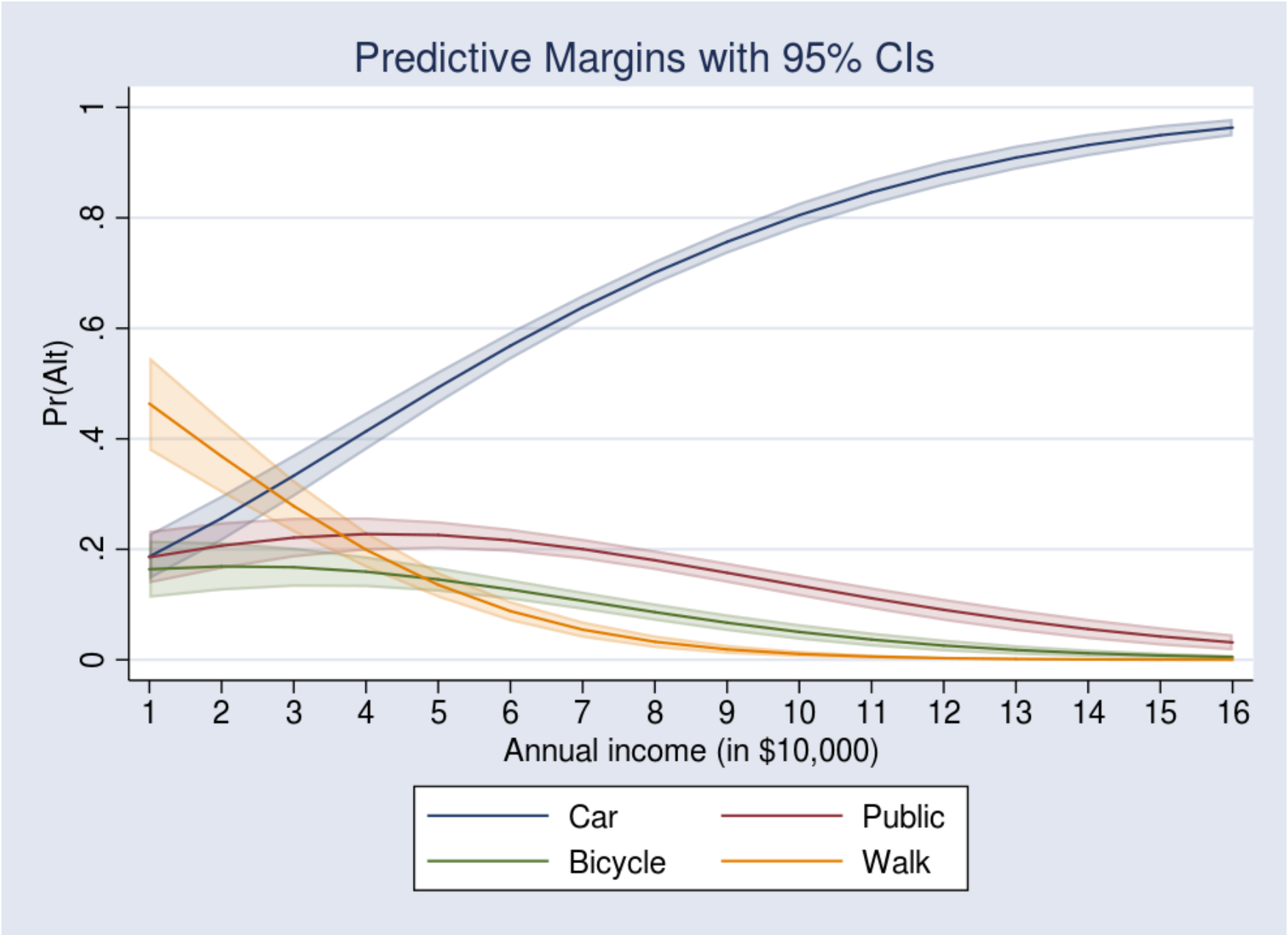
We better plot these:

```
. marginsplot
  Variables that uniquely identify margins: t _outcome
```



What are the averaged choice probabilities over the entire income range?

```
. margins, at(income=(1(1)16))
<output omitted>

. marginsplot, recast(line) ciopts(recast(rarea) color(%20))
  Variables that uniquely identify margins: income _outcome
```


Predictive Margins with 95% CIs
Pr(Alt)
0
.2
.4
.6
.8
1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Annual income (in $10,000)
Car
Public
Bicycle
Walk

Marginal predictions with alternative-specific variables

- Direct and indirect effects
- If travel costs related to cars increased by 25%, how would that affect the probability of choosing a car?
- How would that increase affect the probability of choosing any of the other transportation modes?

`margins` specification

```
. margins, alternative(Car)                          ///
>          at(trcost = generate(trcost))             ///
>          at(trcost = generate(1.25*trcost))        ///
>          subpop(if t==1)
```

Applying the counterfactual

```
. webuse transport
(Transportation choice data)
. generate trcost_cf = trcost
. qui replace trcost_cf = 1.25*trcost if alt == 1
. format trcost_cf %3.2f
. list id t alt choice trcost trcost_cf in 1/12, sepby(t) noobs
```

id	t	alt	choice	trcost	trcost_cf
1	1	Car	1	4.14	5.17
1	1	Public	0	4.74	4.74
1	1	Bicycle	0	2.76	2.76
1	1	Walk	0	0.92	0.92
1	2	Car	1	8.00	10.00
1	2	Public	0	3.14	3.14
1	2	Bicycle	0	2.56	2.56
1	2	Walk	0	0.64	0.64
1	3	Car	1	1.76	2.20
1	3	Public	0	2.25	2.25
1	3	Bicycle	0	0.92	0.92
1	3	Walk	0	0.58	0.58

`margins` output

```
Predictive margins                              Number of obs     =      6,000
Model VCE    : OIM                              Subpop. no. obs   =      2,000

Expression   : Pr(alt), predict()
Alternative  : Car
1._at        : trcost          = trcost
2._at        : trcost          = 1.25*trcost
```

	Delta-method					
	Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_outcome#_at						
Car#1	.5439062	.0113994	47.71	0.000	.5215638	.5662486
Car#2	.4405694	.0101017	43.61	0.000	.4207704	.4603683
Public#1	.2010082	.0104382	19.26	0.000	.1805497	.2214668
Public#2	.2548516	.0117988	21.60	0.000	.2317264	.2779769
Bicycle#1	.1255662	.0095539	13.14	0.000	.1068409	.1442914
Bicycle#2	.1566796	.0110237	14.21	0.000	.1350736	.1782856
Walk#1	.1295194	.0101536	12.76	0.000	.1096187	.1494201
Walk#2	.1478994	.0110109	13.43	0.000	.1263185	.1694803

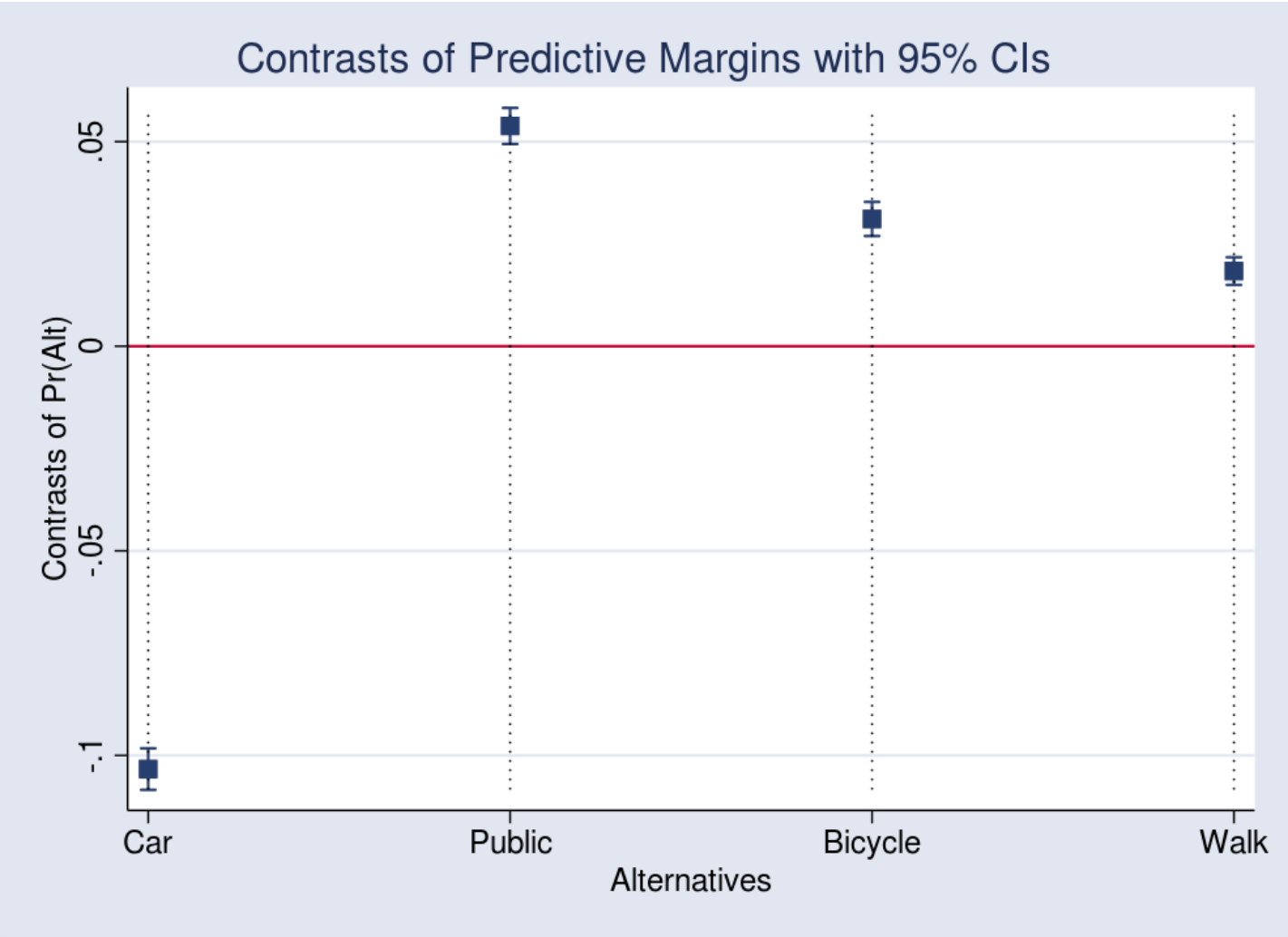
Contrasts with alternative-specific variables

```
. margins, alternative(Car)                        ///
>          at(trcost = generate(trcost))           ///
>          at(trcost = generate(1.25*trcost))      ///
>          contrast(at(r) nowald)                  ///
>          subpop(if t==1)

Contrasts of predictive margins                    Number of obs     =      6,000
Model VCE    : OIM                                 Subpop. no. obs   =      2,000

Expression   : Pr(alt), predict()
Alternative  : Car

1._at        : trcost          = trcost

2._at        : trcost          = 1.25*trcost
```

	Contrast	Delta-method Std. Err.	[95% Conf.	Interval]
_at@_outcome				
(2 vs 1) Car	-.1033369	.0025876	-.1084084	-.0982653
(2 vs 1) Public	.0538434	.0022563	.0494212	.0582656
(2 vs 1) Bicycle	.0311134	.0021237	.0269511	.0352757
(2 vs 1) Walk	.01838	.0017167	.0150153	.0217448

Plotting contrasts

```
. marginsplot, recast(dot) yline(0) plotopts(msymbol(square))
<output omitted>
```



Average marginal effects: how does the probability of choosing a car change with car travel time?

```
. margins, dydx(trtime) outcome(Car) alternative(Car)
Average marginal effects                        Number of obs     =      6,000
Model VCE    : OIM
Expression   : Pr(alt), predict()
Alternative  : Car
Outcome      : Car
dy/dx w.r.t. : trtime
```

	Delta-method					
	dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
trtime						
_cons	-.1581844	.0269102	-5.88	0.000	-.2109275	-.1054414

Average marginal effects: how does the probability of choosing public transportation change with travel time related to car use?

```
. margins, dydx(trtime) outcome(Public) alternative(Car)
Average marginal effects                        Number of obs     =      6,000
Model VCE    : OIM
Expression   : Pr(alt), predict()
Alternative  : Car
Outcome      : Public
dy/dx w.r.t. : trtime
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
trtime						
_cons	.1055447	.0171745	6.15	0.000	.0718834	.139206

Average direct & indirect marginal effects

```
. margins, dydx(trtime) outcome(Car)
Average marginal effects                        Number of obs     =      6,000
Model VCE    : OIM
Expression   : Pr(alt), predict()
Outcome      : Car
dy/dx w.r.t. : trtime
```

	Delta-method					
	dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
trtime						
alt						
Car	-.1581844	.0269102	-5.88	0.000	-.2109275	-.1054414
Public	.1055447	.0171745	6.15	0.000	.0718834	.139206
Bicycle	.0374872	.0073318	5.11	0.000	.0231171	.0518573
Walk	.0151526	.0043034	3.52	0.000	.006718	.0235871

Theoretical motivation of discrete choice models

- Random utility models
- $U_{ijt} = V_{ijt} + \epsilon_{ijt}$
  - U_{ijt} $\rightarrow$ Utility of person i for the jth alternative at time t
  - V_{ijt} $\rightarrow$ Observed component of utility
  - ϵ_{ijt} $\rightarrow$ Unobserved component of utility
- Decision makers choose alternative j if $U_{ijt} > U_{ikt} \quad \forall \quad k \neq j$
- Specification of V_{ijt} and assumptions about ϵ_{ijt} constitute different discrete choice estimators (e.g., logit or probit)
- New estimation command in Stata 16: **`cmxtmixlogit`** for fitting panel-data mixed logit models

The mixed logit model

- With mixed logit, for the random utility model $U_{ijt} = V_{ijt} + \epsilon_{ijt}$ we have:
  - $V_{ijt} = x_{ijt}\beta_i$
  - $\epsilon_{ijt} \sim$ iid type I extreme value
- The random coefficients β_i induce correlation across the alternatives
- We estimate the parameters of a specified distribution for β_i
- The mixed multinomial logit model uses random coefficients to model the correlation of choices across alternatives, thereby relaxing IIA

cmchoiceset – exploring choice sets

```
. cmchoiceset
Tabulation of choice-set possibilities
```

Choice set	Freq.	Percent	Cum.
1 2 3 4	1,053	70.20	70.20
1 2 3 5	210	14.00	84.20
1 2 5 6	90	6.00	90.20
2 3 4 7	147	9.80	100.00
Total	1,500	100.00	

```
Total is number of cases.
```

cmsample – reasons for sample exclusion

```
. preserve
. webuse transport, clear
(Transportation choice data)
. replace trcost = . in 5
(1 real change made, 1 to missing)
. replace alt = . in 2
(1 real change made, 1 to missing)
. replace choice = 0 if t==3 & id==1
(1 real change made)
. replace income = 1 in 1
(1 real change made)
```

cmsample – reasons for sample exclusion

```
. cmset id t alt
panel data: panels id and time t
note: case identifier _caseid generated from id t
note: panel by alternatives identifier _panelaltid generated from id alt
note: alternatives are unbalanced across choice sets; choice sets of
      different sizes found
                      caseid variable:  _caseid
                alternatives variable:  alt
      panel by alternatives variable:  _panelaltid (unbalanced)
                        time variable:  t, 1 to 3
                                delta:  1 unit
note: data have been xtset
```

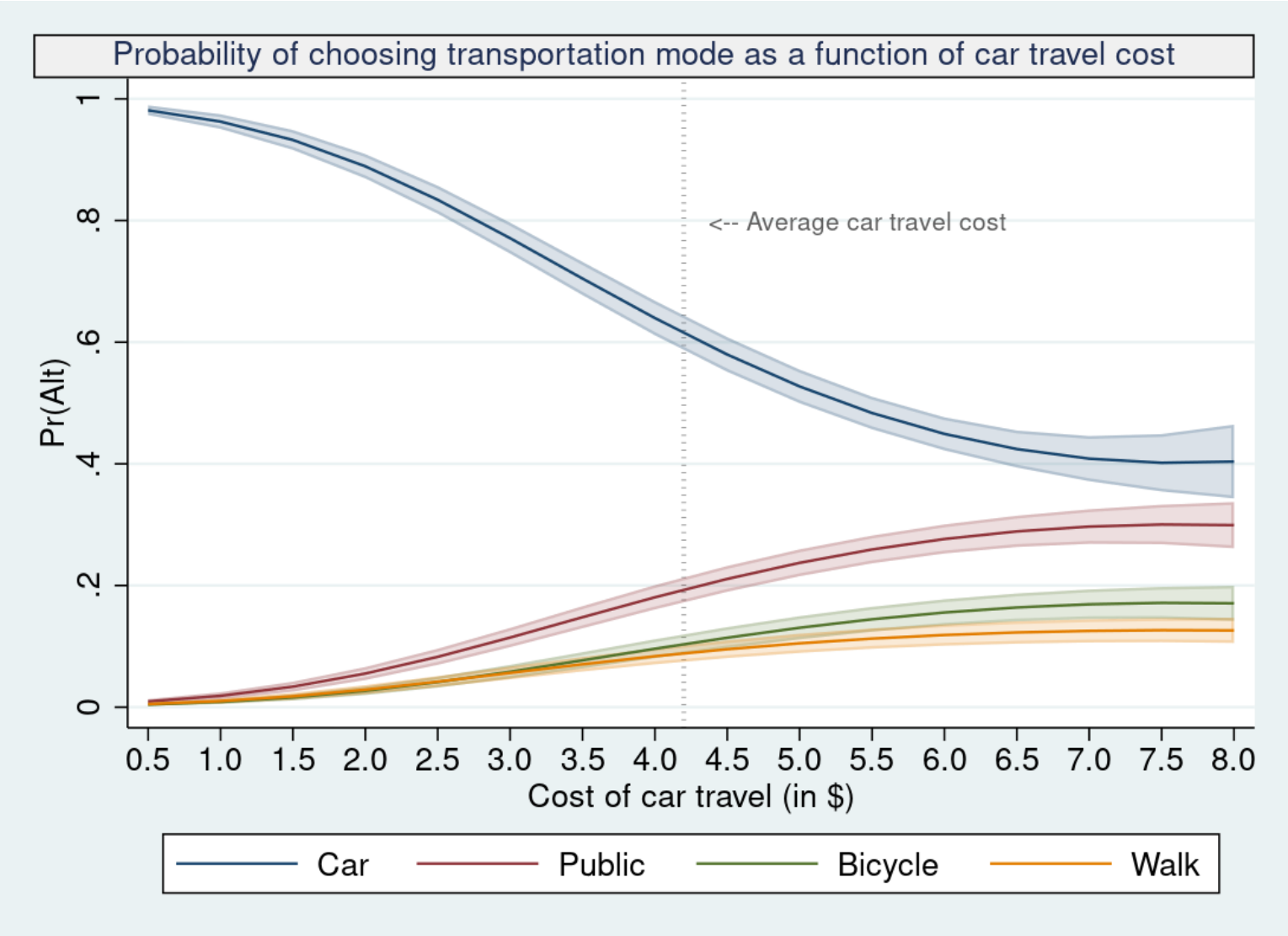
cmsample – reasons for sample exclusion

```
. cmsample trcost trtime, choice(choice) casevars(age income)
```

Reason for exclusion	Freq.	Percent	Cum.
observations included	5,988	99.80	99.80
caseid variable missing	1	0.02	99.82
varlist missing	4	0.07	99.88
choice variable all 0	4	0.07	99.95
casevars not constant within case*	3	0.05	100.00
Total	6,000	100.00	

```
* indicates an error
. restore
```

Thank You!


Probability of choosing transportation mode as a function of car travel cost
Pr(Alt)
0
.2
.4
.6
.8
1
<-- Average car travel cost
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
Cost of car travel (in $)
Car
Public
Bicycle
Walk