Quantitative Data Analysis II

SOC 781

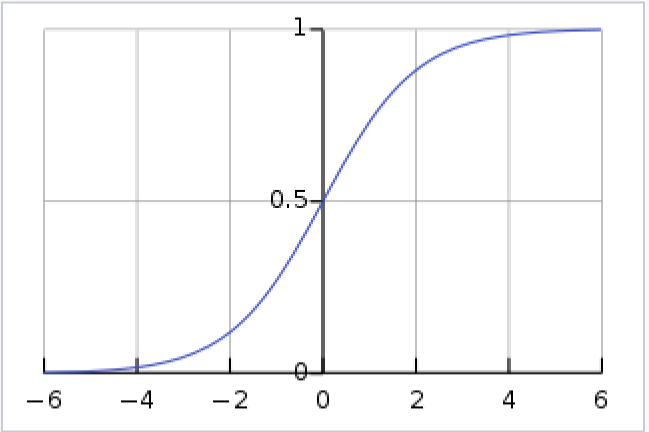
Binary outcomes: logit and probit models

Today we will…

- regression for binary outcomes
  - postestimation techniques for interpretation
- compare logit and probit models
  - model diagnostics
- graphing output

Binary (0,1) outcomes: logistic regression

- Objective: predict membership into one of two categories given values of Xs
- We know the outcomes (number of 0s and 1s)
  - and the conditions under which they occur (corresponding Xs)
- Logistic function (S-shaped curve) and MLE fit model



Standard logistic sigmoid function i.e.
$L = 1, k = 1, x_0 = 0$

Recall: binary outcomes

- How can η link to μ?

- 0,1 implies a binomial distribution

- Thus, we can use a logit link: η = $\log_e\left[\frac{\mu}{1-\mu}\right]$
  - logistic model

- To estimate probability that Y = 1

Binary (0,1) outcomes: logistic regression

$$\ln[p/(1-p)] = \boldsymbol{a} + \boldsymbol{B}X$$

- ln is the natural logarithm ($\log_{exp}$)
  - exp=2.71828…

- p is the probability that the event Y occurs: p(Y=1)

- Thus, ln[p/(1-p)] is the log odds or "logit"

Do you speak in log odds?

Logit example: log odds

- The base Stata output for the logit command provides log odds
- Interpretation?

```
logit hap_dic c.age##c.age i.female i.nonwhite ib1.educat i.married if nmiss==0
```

hap_dic	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0413751	.0040548	-10.20	0.000	-.0493223	-.0334279
c.age#c.age	.0004016	.0000403	9.96	0.000	.0003225	.0004806
1.female	.0937882	.0255448	3.67	0.000	.0437212	.1438551
1.nonwhite	-.4425555	.0288965	-15.32	0.000	-.4991916	-.3859194
educat						
0	-.4118941	.0323797	-12.72	0.000	-.4753572	-.348431
2	.3627807	.0306636	11.83	0.000	.3026812	.4228802
1.married	1.026049	.0274623	37.36	0.000	.9722234	1.079874
_cons	2.415824	.0940427	25.69	0.000	2.231504	2.600144

Odds ratios

- Therefore, it is useful to convert log odds into odds ratios
  - What's an odds ratio?
- First, what's an odd? The ratio of two probabilities
- $\frac{P_{success}}{P_{failure}}$ where $P_{success}$ = $\frac{\#events}{\#_{obs}}$, $P_{failure}$ = $\frac{\#non\text{−}events}{\#_{obs}}$ or $\frac{P_{success}}{(1 - P_{success})}$
  - events = 1, non-events = 0
- Odds ratio = odds of one group divided by the odds of another group

Odds ratios

- A bivariate example

```
tab hap_dic female if nmiss==0
```

hap_dic	female 0	1	Total
0	3,339	4,275	7,614
1	23,072	29,039	52,111
Total	26,411	33,314	59,725

- Probability males happy = 23,072 / 26,411 = 0.87
- Probability females happy = 29,039 / 33,314 = 0.87
- $\mathrm{OR}_{females\ vs.\ males}$ = [0.87/(1-0.87)] / [0.87/(1-0.87)] = 1.00
- OR = 1
    - sex is not associated with the odds of being happy

Odds ratios

- What if the OR >1?
- Can also compute as
  - ad/bc

a	b
c	d

```
tab hap_dic married if nmiss==0

           |        married
   hap_dic |         0          1 |     Total
-----------+----------------------+----------
         0 |     5,197      2,417 |     7,614
         1 |    22,890     29,221 |    52,111
-----------+----------------------+----------
     Total |    28,087     31,638 |    59,725
```

- $OR_{\text{married vs. unmarried}}$ = [(5,197) (29,221)] / [(2,417) (22,890)] = 2.74
- The odds of being happy are 2.74 times larger for married vs. unmarried
- The odds of being happy are 174% greater for married vs. unmarried
  - 2.74 – 1 = 174

Odds ratios

- What if the OR < 1?
- Let's transform married into unmarried

```
gen unmarried=.
replace unmarried=0 if married==1
replace unmarried=1 if married==0
```

a	b
c	d

OR = ad/bc

hap_dic	unmarried 0	unmarried 1	Total
0	2,431	5,237	7,668
1	29,342	23,044	52,386
Total	31,773	28,281	60,054

- $OR_{\text{unmarried vs. married}}$ = [(2,431) (23,044)] / [(5,237) (29,342)] = 0.36
- The odds of being happy are 0.36 lower for unmarried vs. married
- The odds of being happy are 64% lower for unmarried vs. married
  - 1 – 0.36 = 0.64

Odds ratios: interpretation

- If there is no change in odds associated with a unit change in x: OR = 1
- If the odds increase with a unit change in x: OR>1
- If the odds decrease with a unit change in x: OR<1
- In other words: “positive” effects are greater than one, while “negative” effects are between zero and one

Binary (0,1) outcomes: logistic regression

- This gets much more complicated when X is continuous
  - or there is more than one X
- Luckily, Stata does this for us with “or” command

$$\ln[p/(1-p)] = \boldsymbol{a} + \boldsymbol{B}X$$

$$[p/(1-p)] = \exp(\boldsymbol{a} + \boldsymbol{B}X)$$

- ln is the natural logarithm, $\log_{exp}$, where exp=2.71828…
- p is the probability that the event Y occurs: p(Y=1)
- ln[p/(1-p)] is the log odds ratio or "logit"
- p/(1-p) is the "odds ratio"

Logit example: odds ratios

```
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married if nmiss==0
```

hap_dic	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0413751	.0040548	-10.20	0.000	-.0493223	-.0334279
c.age#c.age	.0004016	.0000403	9.96	0.000	.0003225	.0004806
1.female	.0937882	.0255448	3.67	0.000	.0437212	.1438551
1.nonwhite	-.4425555	.0288965	-15.32	0.000	-.4991916	-.3859194
educat						
0	-.4118941	.0323797	-12.72	0.000	-.4753572	-.348431
2	.3627807	.0306636	11.83	0.000	.3026812	.4228802
1.married	1.026049	.0274623	37.36	0.000	.9722234	1.079874
_cons	2.415824	.0940427	25.69	0.000	2.231504	2.600144

```
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married ///
if nmiss==0, or
```

hap_dic	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.9594691	.0038904	-10.20	0.000	.9518743	.9671246
c.age#c.age	1.000402	.0000404	9.96	0.000	1.000323	1.000481
1.female	1.098327	.0280566	3.67	0.000	1.044691	1.154717
1.nonwhite	.6423927	.0185629	-15.32	0.000	.6070212	.6798253
educat						
0	.6623944	.0214481	-12.72	0.000	.621663	.7057946
2	1.437321	.0440734	11.83	0.000	1.353483	1.526351
1.married	2.79002	.0766205	37.36	0.000	2.643816	2.944308
_cons	11.19899	1.053183	25.69	0.000	9.31386	13.46568

- Note from equation in previous slide
  - ORs are simply exponentiated log odds
    - exp(-0.0413751) = 0.9594691

- Interpretation depends on level of Xs

Interpretation: OR >1 (dummies)

```
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married ///
if nmiss==0, or
```

hap_dic	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.9594691	.0038904	-10.20	0.000	.9518743	.9671246
c.age#c.age	1.000402	.0000404	9.96	0.000	1.000323	1.000481
1.female	$\boxed{1.098327}$	.0280566	3.67	0.000	1.044691	1.154717
1.nonwhite	.6423927	.0185629	-15.32	0.000	.6070212	.6798253
educat						
0	.6623944	.0214481	-12.72	0.000	.621663	.7057946
2	1.437321	.0440734	11.83	0.000	1.353483	1.526351
1.married	2.79002	.0766205	37.36	0.000	2.643816	2.944308
_cons	11.19899	1.053183	25.69	0.000	9.31386	13.46568

- The odds of being happy are 1.10 times greater for females vs. males
  - all else equal
- The odds of being happy are 10% higher for females vs. males
  - all else equal

Interpretation: OR <1 (dummies)

```
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married ///
if nmiss==0, or
```

hap_dic	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.9594691	.0038904	-10.20	0.000	.9518743	.9671246
c.age#c.age	1.000402	.0000404	9.96	0.000	1.000323	1.000481
1.female	1.098327	.0280566	3.67	0.000	1.044691	1.154717
1.nonwhite	.6423927	.0185629	-15.32	0.000	.6070212	.6798253
educat						
0	.6623944	.0214481	-12.72	0.000	.621663	.7057946
2	1.437321	.0440734	11.83	0.000	1.353483	1.526351
1.married	2.79002	.0766205	37.36	0.000	2.643816	2.944308
_cons	11.19899	1.053183	25.69	0.000	9.31386	13.46568

- The odds of being happy are 36% lower for nonwhites vs. whites
  - all else equal

Odds ratios: relative to base group

```
logit hap_dic c.age##c.age i.female i.nonwhite ib1.educat i.married ///
if nmiss==0, or
```

hap_dic	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.9594691	.0038904	-10.20	0.000	.9518743	.9671246
c.age#c.age	1.000402	.0000404	9.96	0.000	1.000323	1.000481
1.female	1.098327	.0280566	3.67	0.000	1.044691	1.154717
1.nonwhite	.6423927	.0185629	-15.32	0.000	.6070212	.6798253
educat						
0	.6623944	.0214481	-12.72	0.000	.621663	.7057946
2	1.437321	.0440734	11.83	0.000	1.353483	1.526351
1.married	2.79002	.0766205	37.36	0.000	2.643816	2.944308
_cons	11.19899	1.053183	25.69	0.000	9.31386	13.46568

base group is (1) HS

- This would be (0) LTHS w/o ib1
  - by default

- may want to compare to another educational group

- Interpretation?

```
pwcompare educat, effects eform
```

	exp(b)	Std. Err.	Unadjusted z	Unadjusted P>|z|	Unadjusted [95% Conf.	Interval]
hap_dic						
educat						
1 vs 0	1.509675	.0488828	12.72	0.000	1.416843	1.608589
2 vs 0	2.169886	.0680512	24.70	0.000	2.040525	2.307449
2 vs 1	1.437321	.0440734	11.83	0.000	1.353483	1.526351

Interpretation: OR (continuous)

- To avoid the age polynomial (for simplicity) let’s look at edu years

```
logit hap_dic c.age##c.age i.female i.nonwhite c.educ i.married ///
if nmiss==0, or
```

hap_dic	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.9589154	.0038924	-10.34	0.000	.9513167	.9665748
c.age#c.age	1.000413	.0000405	10.22	0.000	1.000334	1.000493
1.female	1.093578	.0279077	3.51	0.000	1.040225	1.149667
1.nonwhite	.6566978	.0190268	-14.51	0.000	.6204451	.6950688
educ	1.10252	.0044311	24.28	0.000	1.09387	1.111239
1.married	2.787257	.0765109	37.34	0.000	2.641261	2.941323
_cons	3.436189	.3575823	11.86	0.000	2.802193	4.213628

- The odds of being happy increase by 1.10 with each additional year of edu
  - all else equal

- One additional year of edu increases odds of being happy by 10%
  - all else equal

Odds ratios: limitations

- Don’t speak toward absolute magnitude
- Substantive meaning depends on value of odds before they change
  - which depend on the predicted probability
- Predicted probability depends on values of all Xs

Predicted probabilities

- ORs are informative, but they are relative
- Can use predicted probability to assess magnitude
  - we'll start with predicted probabilities for all combinations of Xs
- When using "predict" make sure to limit to analytic sample

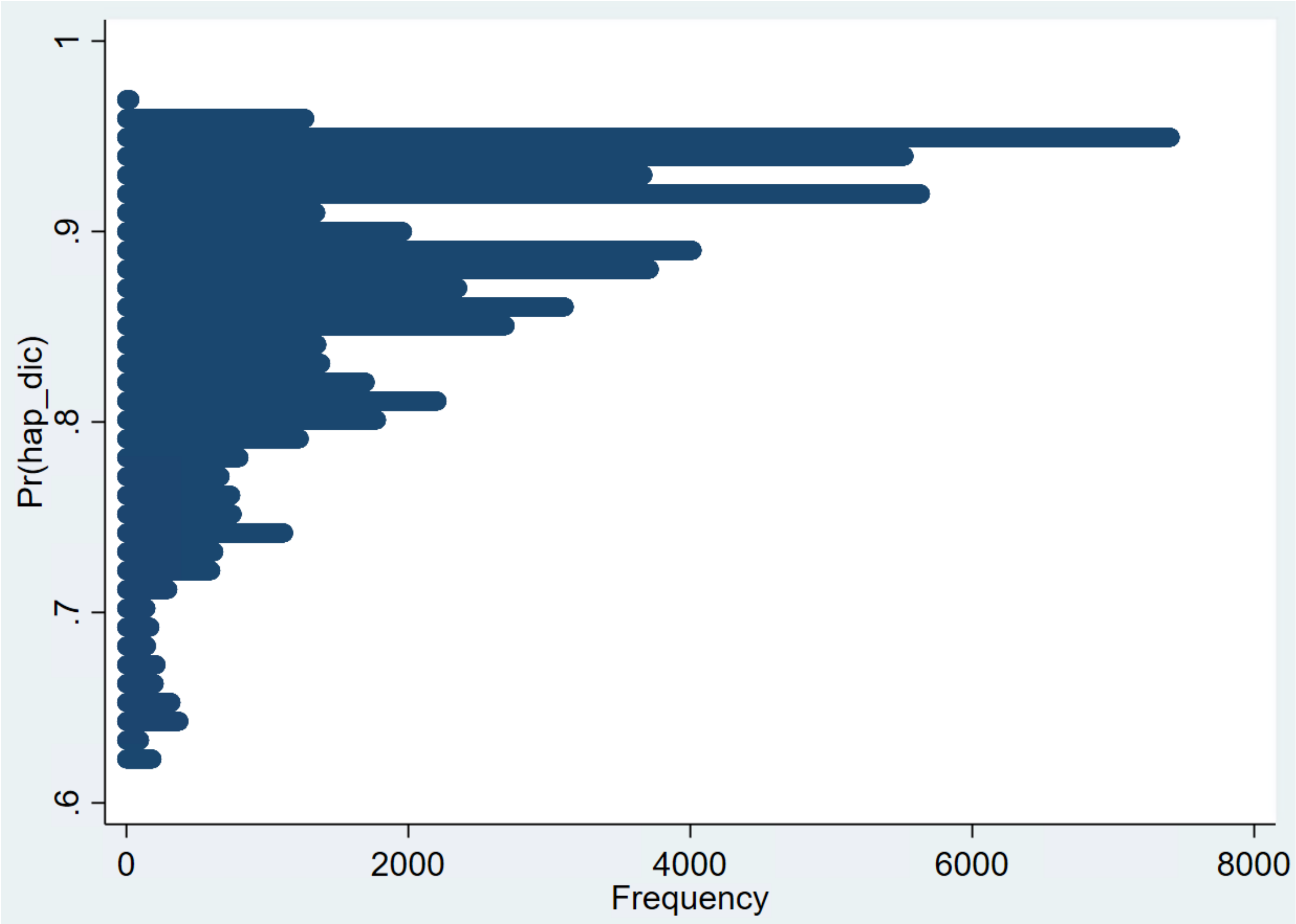
```
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married ///
if nmiss==0, or
predict prlogit if nmiss==0
predict prlogit2 if e(sample)==1
predict prlogitwrong
codebook prlogit prlogit2 prlogitwrong, compact
```

Variable	Obs	Unique	Mean	Min	Max	Label
prlogit	59725	1654	.8725157	.6214635	.967612	Pr(hap_dic)
prlogit2	59725	1654	.8725157	.6214635	.967612	Pr(hap_dic)
prlogitwrong	64586	1663	.8724231	.6214635	.967612	Pr(hap_dic)

- Predicted probabilities range from 0.62 to 0.97 with a mean of 0.87

Predicted probabilities

- Plot the predicted probabilities to examine the distribution



Predicted probabilities: Marginal effects

- Marginal effect: Δ in the predicted probability given a Δ in X
  - holding all other Xs constant
    - Is there a meaningful way to hold all other Xs constant?
- Average marginal effect (AME): the average of the marginal effect for all observations
  - Likely, no one is "average." What about underrepresented groups?
- Marginal effect at the mean (MEM): all other Xs held at their means
  - Many mean values are often meaningless (e.g., dummy Xs)
- Marginal effect at representative values (MER): all other Xs held at substantively meaningful values
  - What are "meaningful" values? Can become quickly overwhelmed with details

Average marginal effect (AME): continuous

- Avg. Δ in probability for Δ in education (years), holding all else constant

```
logit hap_dic c.age##c.age i.female i.nonwhite c.educ i.married ///
if nmiss==0, or

mchange educ, decimals(5)
```

	Change	p-value
educ		
+1	0.01001	0.00000
+SD	0.02955	0.00000
Marginal	0.01035	0.00000

- On average, one additional year of edu. is associated with a 0.01 increase in the probability of being happy, all else equal

- Consider this effect across the range of edu: [0 to 20 years]

```
mchange educ, amount(range) statistics(change from to pvalue)
```

	Change	From	To	p-value
educ				
Range	0.251	0.683	0.933	0.000

- On average, increasing education from 0 to 20-years is associated with a 0.25 increase in the probability of happiness, all else equal

Average marginal effect (AME): categorical

- Avg. Δ in probability for Δ in education (groups), holding all else constant

```
logit hap_dic c.age##c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0, or
mchange educat, statistics(change from to pvalue)
```

	Change	From	To	p-value
educat				
1 vs 0	0.052	0.816	0.868	0.000
2 vs 0	0.087	0.816	0.904	0.000
2 vs 1	0.035	0.868	0.904	0.000

- On average, having a HS education versus <HS increases the probability of happiness by 0.05, all else equal…

- by 0.09 for a college education vs. <HS, and by 0.04 for college vs. HS

Marginal effect at the mean (MEM)

- Summary table for all Xs

```
logit hap_dic c.age##c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0, or
mchange, atmeans statistics(ci) decimals(4)
```

	Change	LL	UL
age			
+1	-0.0004	-0.0006	-0.0003
+SD	0.0050	0.0019	0.0082
Marginal	-0.0005	-0.0006	-0.0003
female			
1 vs 0	0.0103	0.0048	0.0158
nonwhite			
1 vs 0	-0.0536	-0.0611	-0.0460
educat			
1 vs 0	0.0551	0.0464	0.0639
2 vs 0	0.0917	0.0836	0.0999
2 vs 1	0.0366	0.0303	0.0429
married			
1 vs 0	0.1163	0.1095	0.1231

- For respondents average on all characteristics, a one-year increase in age is associated with a 0.0004 decrease in the probability of happiness

- Females have a 0.01 greater probability of happiness compared to males, holding other covariates at their means

Base values of regressors

	age	1. female	1. nonwhite	1. educat	2. educat	1. married
at	46.05	.5578	.1931	.3059	.4656	.5297

Marginal effect at representative values (MER)

```
logit hap_dic age age2 i.female i.nonwhite i.educat i.married ///
if nmiss==0, or
mchange married, at(married=0 age=40 female=1 nonwhite=0 educat=1)
```

	Change	p-value
married		
1 vs 0	0.093	0.000

Base values of regressors

	age	female	nonwhite	educat	married
at	40	1	0	1	0

```
mchange married, at(married=0 age=40 female=0 nonwhite=0 educat=1)
```

	Change	p-value
married		
1 vs 0	0.100	0.000

Base values of regressors

	age	female	nonwhite	educat	married
at	40	0	0	1	0

- For HS educated, white, 40-year-old, females the probability of happiness is 0.093 greater among those who are married compared to those who are not married

- For males with the same characteristics the probability of happiness is 0.100 higher among those who are married versus those who are not married

- Note how marginal effects depend on values of Xs

Postestimation group differences

- When comparing postestimation statistics across groups
  - the CIs are conservative estimates
- Because ignores the covariance of the estimators
- See: Schenker & Gentleman (2001)

Marginal effects: interactions

```
logit hap_dic c.age##c.age i.female##i.nonwhite c.educ i.married ///
if nmiss==0, or
mtable, dydx(female) over (nonwhite) stat(ci) post
```

	d Pr(y)	ll	ul
0	0.011	0.006	0.017
1	0.001	-0.013	0.016

```
mlincom 1-2
```

	lincom	pvalue	ll	ul
1	0.010	0.204	-0.005	0.026

- Although the average effect of female is greater for whites, this difference is not statistically significant

Ideal types

- Often it makes sense to compute predicted probabilities for substantively meaningful groups to make comparisons
  - set values of X to create hypothetical observation
    - e.g., age 40 whites vs non-whites

```
logit hap_dic c.age##c.age i.female i.nonwhite c.educ i.married ///
if nmiss==0, or
mtable, at(age=40 nonwhite=0) atmeans ci
mtable, at(age==40 nonwhite==1) atmeans ci below
```

	Pr(y)	ll	ul
estimate	0.887	0.884	0.891
estimate	0.838	0.831	0.845

- The probability of being happy at age forty is 0.89 for whites and 0.84 for non-whites, holding all else at global means
  - but whites and non-whites differ on female, educ, and married

Ideal types

- Use subgroup means
  - all Xs in model not specified with atspec held at subgroup means

```
mtable if _sel40W==1, rowname(1 40yr whites) atmeans ci
mtable if _sel40N==1, rowname(2 40yr non-whites)atmeans ci below
```

	Pr(y)	ll	ul
1 40yr whites	0.909	0.905	0.912
2 40yr non-whites	0.822	0.814	0.829

	age	1. female	nonwhite	educ	1. married
Set 1	40	.523	0	13.8	.67
Current	40	.5	1	12.8	.422

- The probability of being happy at age forty is 0.91 for whites and 0.82 for non-whites, when all other variables held at subgroup means
  - But does 0.523 female or 0.67 married make sense?

Ideal types: comparison

- Test whether difference is statistically significant

```
logit hap_dic c.age##c.age i.female i.nonwhite c.educ i.married ///
if nmiss==0, or
estimates store base /*store estimates*/
mtable, post at(age=40 nonwhite=0 female=0 educ=12 married=0) at(age=40 ///
nonwhite=1 female=0 educ=12 married=0)
mlincom 1-2
estimates restore base /*restore estimates*/
```

	nonwhite	Pr(y)	ll	ul
1	0	0.800	0.792	0.809
2	1	0.725	0.713	0.737

	lincom	pvalue	ll	ul
1	0.076	0.000	0.065	0.086

- For 40-year-old, males, not married, with 12-years of education whites are significantly more likely to be happy than non-white counterparts
  - gets complex when many different hypothetical groups included
    - for this course, I’m okay with relying on 95% CIs

Logit vs. probit

- Recall: How can η link to μ?
- Logit link: η = $\log_e \left[\frac{\mu}{1-\mu}\right]$
- Can also use inverse normal link: η = $\Phi^{-1}(\mu)$
  - probit model
- Different assumptions than logit
  - but produces substantively comparable results, typically

Logit vs. probit

- Probit: assumes unobserved continuous scale underlies binary outcome
- Differs from logit in how it deals with the error term
- The coefficients differ substantially
  - but the predicted probabilities almost identical, typically

Logit vs. probit

- Probit coefficients reflect Δ in a standard deviation increase in the predicted probit index
  - can't be converted into ORs

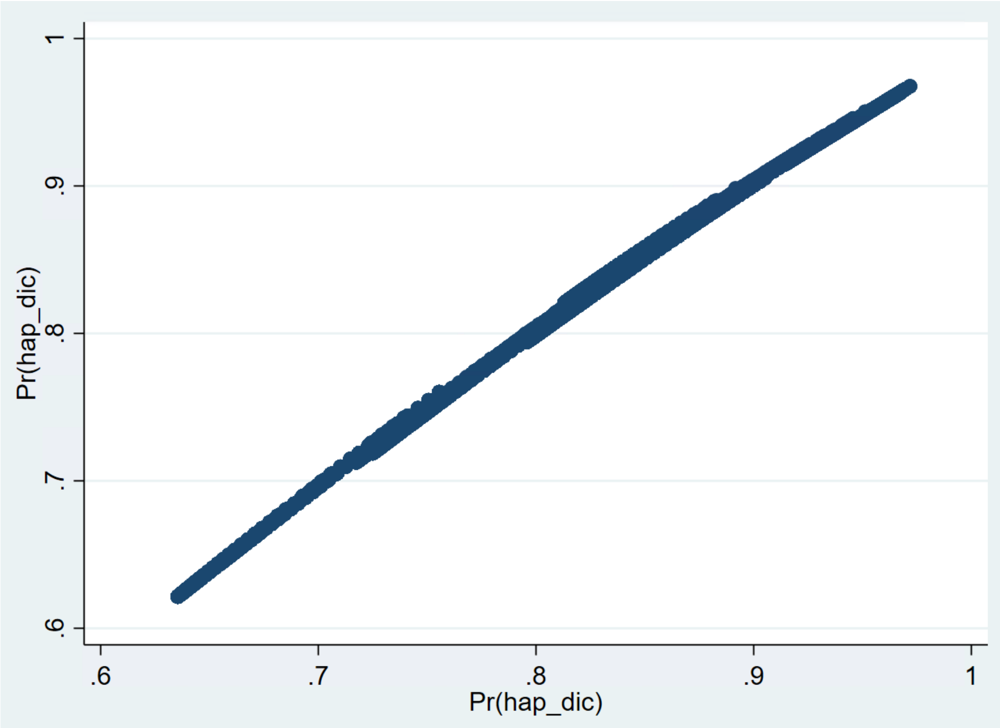
```
logit hap_dic c.age##c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0
estimates store Alogit
probit hap_dic c.age##c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0
estimates store Aprobit
estimates table Alogit Aprobit
```

Variable	Alogit	Aprobit
age	-.04137513	-.02201129
c.age#c.age	.00040159	.00021265
female		
1	.09378816	.0509238
nonwhite		
1	-.44255549	-.25161174
educat		
1	.41189408	.22887691
2	.77467474	.4211535
married		
1	1.0260486	.54320436
_cons	2.0039298	1.1674772

Logit vs. probit

- Need to rely on predicted probabilities
- Should be almost identical to those from logit
  - if robust

```
logit hap_dic c.age##c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0
predict prlogit if nmiss==0
probit hap_dic c.age##c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0
predict prprobit if nmiss==0
scatter prlogit prprobit
```



Logit vs. probit

- Can use same postestimation techniques
  - except "or"

AME logit

	Change	p-value
age		
+1	-0.0004	0.0000
+SD	0.0035	0.0034
Marginal	-0.0004	0.0000
female		
1 vs 0	0.0100	0.0003
nonwhite		
1 vs 0	-0.0512	0.0000
educat		
1 vs 0	0.0522	0.0000
2 vs 0	0.0874	0.0000
2 vs 1	0.0353	0.0000
married		
1 vs 0	0.1085	0.0000

AME probit

	Change	p-value
age		
+1	-0.0004	0.0000
+SD	0.0034	0.0066
Marginal	-0.0004	0.0000
female		
1 vs 0	0.0101	0.0002
nonwhite		
1 vs 0	-0.0537	0.0000
educat		
1 vs 0	0.0526	0.0000
2 vs 0	0.0879	0.0000
2 vs 1	0.0353	0.0000
married		
1 vs 0	0.1082	0.0000

Binary logit & probit: diagnostics

- Residuals and influential observations
- LR chi-square test: overall test of model fit
- Pseudo-R^2
- BIC & AIC: information criteria measures
- Mostly the same for ologit

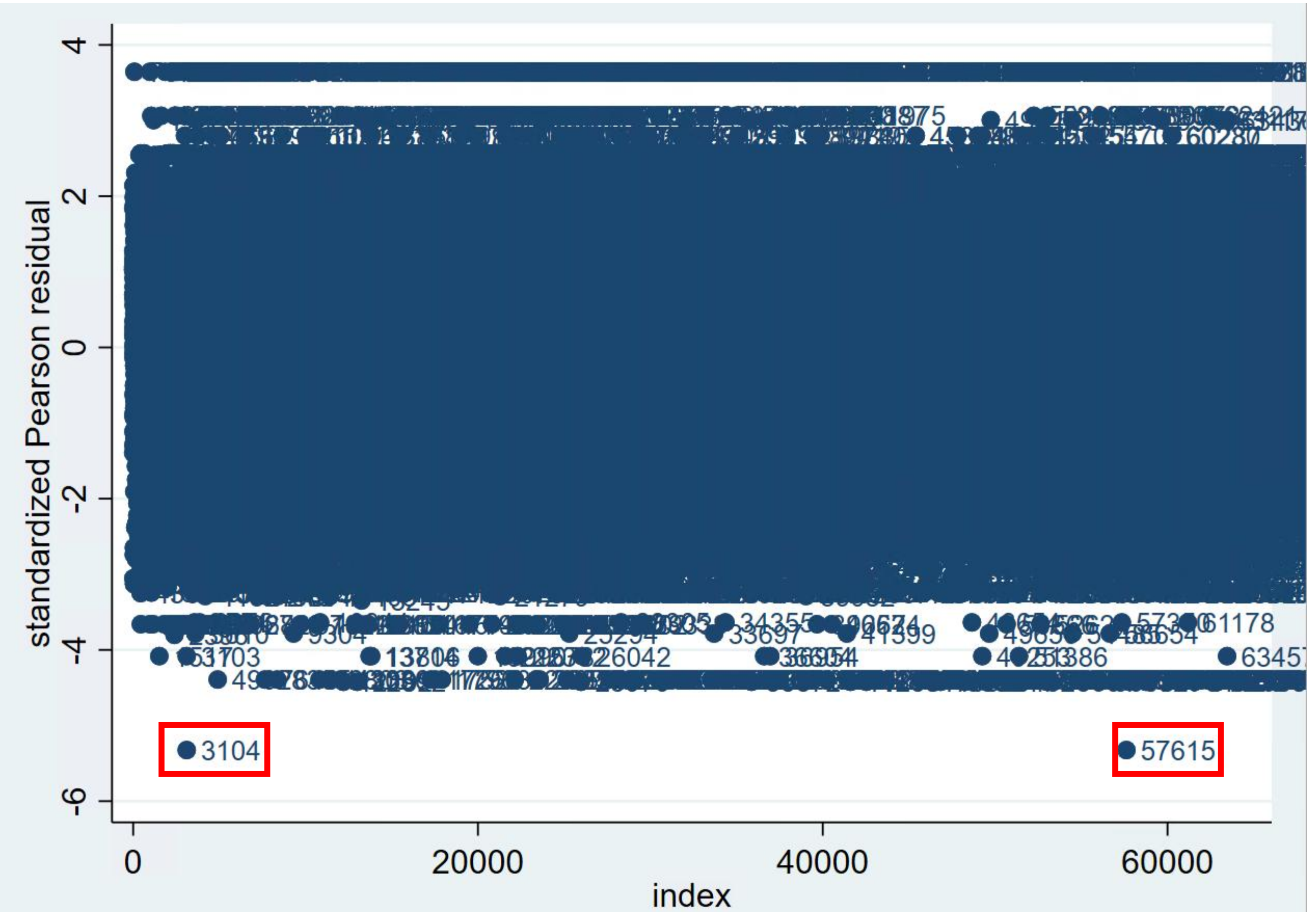
Binary logit & probit: residuals and influential

- Plot residuals against index of observations
  - and check out any possibly influential cases

```
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married ///
if nmiss==0, or
predict rstd if nmiss==0, rstandard
generate index = _n if nmiss==0
graph twoway scatter rstd index, mlabel(index)
```

```
list rstd index age female nonwhite educat married if rstd<-4.5
```

	rstd	index	age	female	nonwhite	educat	married
3104.	-5.324152	3104	77	1	1	2	1
57615.	-5.324152	57615	77	1	1	2	1



Binary logit & probit: least likely observations

- Identify any possible patterns in least likely observations
- Among those who are not happy, the lowest predicted probability of not being happy typically occurs among married, white, females with a college education
- Among those who are happy, the lowest predicted probability of being happy occurs among unmarried, black, males, with less than HS education

```
leastlikely age female nonwhite educat married
```

Outcome: 0

	Prob	age	female	nonwhite	educat	married
3860.	.0389157	21	1	0	2	1
8950.	.0389157	21	1	0	2	1
25270.	.0375198	87	0	0	2	1
40264.	.0342759	87	1	0	2	1
48629.	.038021	83	1	0	2	1
55052.	.0389157	21	1	0	2	1
59214.	.0364898	88	0	0	2	1

Outcome: 1

	Prob	age	female	nonwhite	educat	married
2465.	.6214635	52	0	1	0	0
19969.	.6214635	52	0	1	0	0
31374.	.6214635	52	0	1	0	0
37050.	.6214635	52	0	1	0	0
40448.	.6214635	52	0	1	0	0
40614.	.6214635	52	0	1	0	0
54965.	.6214635	52	0	1	0	0
64771.	.6214635	52	0	1	0	0

Likelihood ratio (LR) chi-square test

- Test for overall model fit
  - contrasts to model w/ no IVs (constant only)
- Not super informative
  - somewhat useful for nested models

```
Logistic regression                             Number of obs     =     59,725
                                                LR chi2(5)        =    2017.58
                                                Prob > chi2       =     0.0000
Log likelihood = -21780.858                     Pseudo R2         =     0.0443
```

hap_dic	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.9671035	.0038916	-8.31	0.000	.9595061	.974761
c.age#c.age	1.000285	.0000397	7.18	0.000	1.000207	1.000363
1.female	1.089552	.0276389	3.38	0.001	1.036706	1.145093
1.nonwhite	.5941199	.0169221	-18.28	0.000	.5618619	.6282298
1.married	2.72853	.0743733	36.83	0.000	2.586587	2.878263
_cons	11.09686	1.02477	26.06	0.000	9.259624	13.29862

```
Logistic regression                             Number of obs     =     59,725
                                                LR chi2(7)        =    2624.70
                                                Prob > chi2       =     0.0000
Log likelihood = -21477.296                     Pseudo R2         =     0.0576
```

hap_dic	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.9594691	.0038904	-10.20	0.000	.9518743	.9671246
c.age#c.age	1.000402	.0000404	9.96	0.000	1.000323	1.000481
1.female	1.098327	.0280566	3.67	0.000	1.044691	1.154717
1.nonwhite	.6423927	.0185629	-15.32	0.000	.6070212	.6798253
educat						
0	.6623944	.0214481	-12.72	0.000	.621663	.7057946
2	1.437321	.0440734	11.83	0.000	1.353483	1.526351
1.married	2.79002	.0766205	37.36	0.000	2.643816	2.944308
_cons	11.19899	1.053183	25.69	0.000	9.31386	13.46568

Pseudo-R^2

- Not same as OLS R^2: proportion of explained variance
  - improves likelihood of the model by __% vs. constant-only model

Logistic regression

Number of obs = 59,725
LR chi2(5) = 2017.58
Prob > chi2 = 0.0000
Log likelihood = -21780.858
Pseudo R2 = 0.0443

hap_dic	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.9671035	.0038916	-8.31	0.000	.9595061	.974761
c.age#c.age	1.000285	.0000397	7.18	0.000	1.000207	1.000363
1.female	1.089552	.0276389	3.38	0.001	1.036706	1.145093
1.nonwhite	.5941199	.0169221	-18.28	0.000	.5618619	.6282298
1.married	2.72853	.0743733	36.83	0.000	2.586587	2.878263
_cons	11.09686	1.02477	26.06	0.000	9.259624	13.29862

Logistic regression

Number of obs = 59,725
LR chi2(7) = 2624.70
Prob > chi2 = 0.0000
Log likelihood = -21477.296
Pseudo R2 = 0.0576

hap_dic	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.9594691	.0038904	-10.20	0.000	.9518743	.9671246
c.age#c.age	1.000402	.0000404	9.96	0.000	1.000323	1.000481
1.female	1.098327	.0280566	3.67	0.000	1.044691	1.154717
1.nonwhite	.6423927	.0185629	-15.32	0.000	.6070212	.6798253
educat						
0	.6623944	.0214481	-12.72	0.000	.621663	.7057946
2	1.437321	.0440734	11.83	0.000	1.353483	1.526351
1.married	2.79002	.0766205	37.36	0.000	2.643816	2.944308
_cons	11.19899	1.053183	25.69	0.000	9.31386	13.46568

Information criteria measures

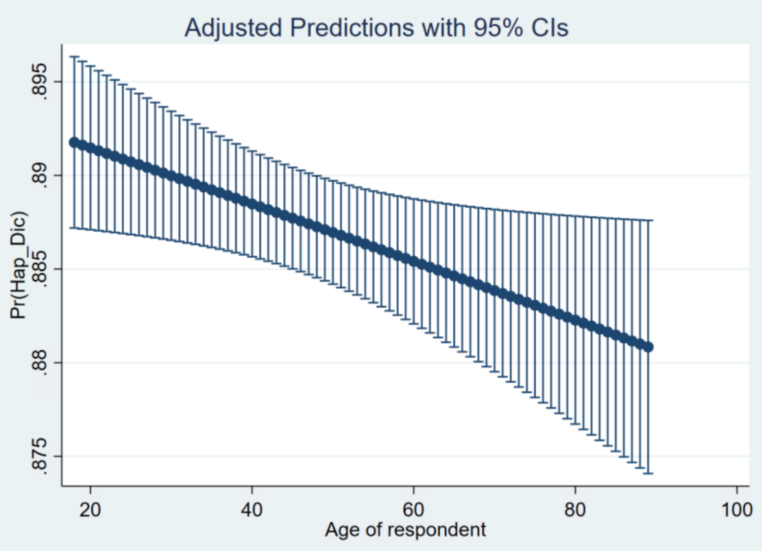
- AIC: Akaike's Information Criteria
- BIC: Bayesian Information Criteria
  - Doesn't matter which one you use, just be consistent
- Smaller → better model fit: BIC rule of thumb
  - 0-2 = no difference between models
  - 2-6 = positive support for model 1
  - 6-10 = strong support
  - > 10 = very strong support

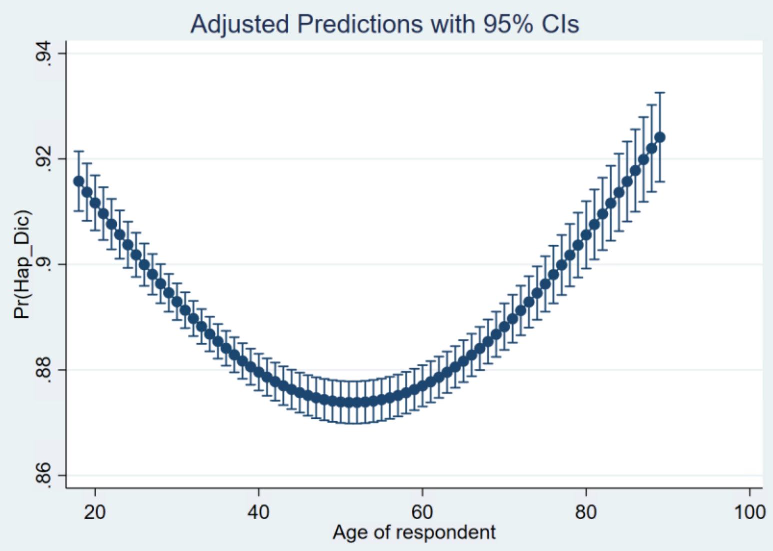
```
logit hap_dic c.age##c.age i.female i.nonwhite i.married ///
if nmiss==0, or
quietly fitstat, save
/*see how AIC & BIC decreases after adding educ.?*/
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married ///
if nmiss==0, or
fitstat, dif
```

	Current	Saved	Difference
Log-likelihood			
Model	-21477.296	-21780.858	303.562
Intercept-only	-22789.647	-22789.647	0.000
Chi-square			
D(df=59717/59719/-2)	42954.591	43561.715	-607.124
LR(df=7/5/2)	2624.702	2017.578	607.124
p-value	0.000	0.000	0.000
R2			
McFadden	0.058	0.044	0.013
McFadden(adjusted)	0.057	0.044	0.013
McKelvey & Zavoina	0.109	0.086	0.023
Cox-Snell/ML	0.043	0.033	0.010
Cragg-Uhler/Nagelkerke	0.081	0.062	0.018
Efron	0.045	0.034	0.011
Tjur's D	0.046	0.035	0.011
Count	0.873	0.873	0.000
Count(adjusted)	0.000	0.000	0.000
IC			
AIC	42970.591	43573.715	-603.124
AIC divided by N	0.719	0.730	-0.010
BIC(df=8/6/2)	43042.571	43627.700	-585.129
Variance of			
e	3.290	3.290	0.000
y-star	3.692	3.599	0.092

Graphing predicted probabilities: margins

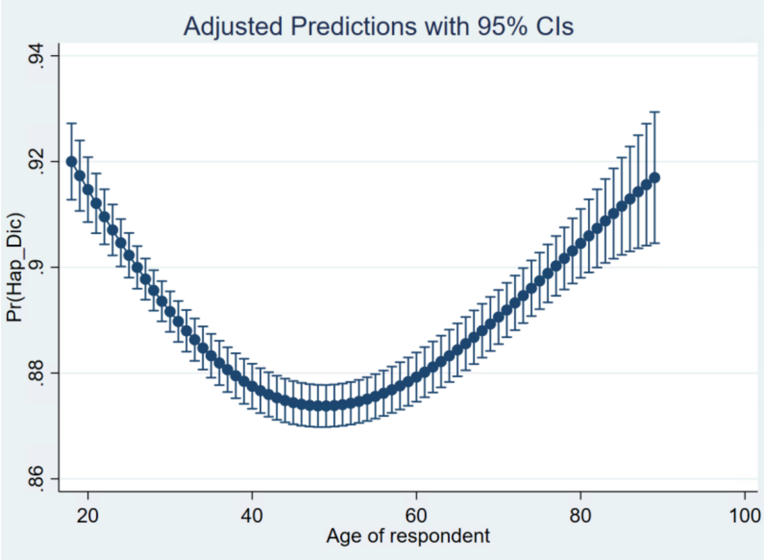
```
logit hap_dic c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0, or
estimates store base
margins, at(age=(18(1)89)) atmeans
marginsplot
```






```
logit hap_dic c.age##c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0, or
estimates store base
margins, at(age=(18(1)89)) atmeans
marginsplot
```

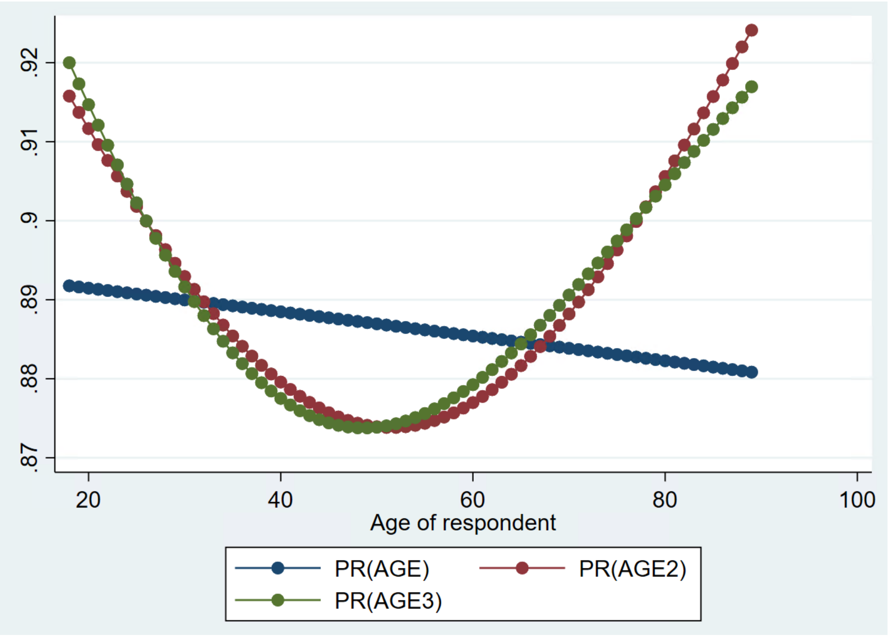
```
logit hap_dic c.age##c.age##c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0, or
estimates store base
margins, at(age=(18(1)89)) atmeans
marginsplot
```



Graphing predicted probabilities: mgen

```
logit hap_dic c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0, or
mgen, at(age=(18(1)89)) atmeans replace stub(PLT) predlabel(PR(AGE))
logit hap_dic c.age##c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0, or
mgen, at(age=(18(1)89)) atmeans replace stub(PLT2) predlabel(PR(AGE2))
logit hap_dic c.age##c.age##c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0, or
mgen, at(age=(18(1)89)) atmeans replace stub(PLT3) predlabel(PR(AGE3))
```

```
twoway (connected PLTpr1 PLT2pr1 PLT3pr1 PLTage)
```

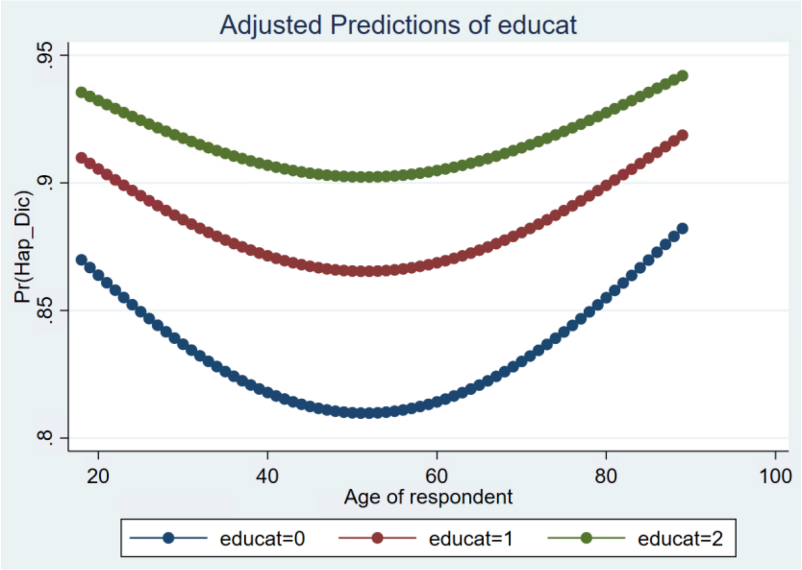


- Useful for combining graphs

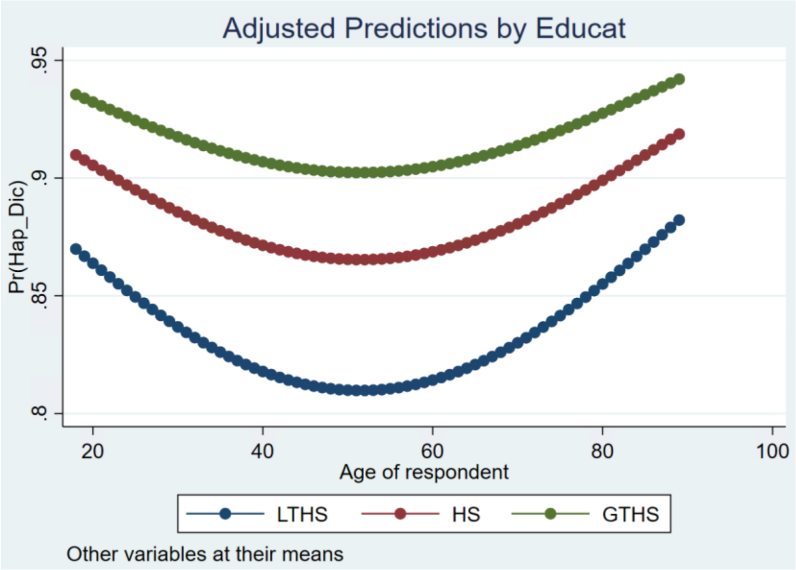
- Will become even more useful with other glm techniques
  - e.g., when we examine ordinal and nominal outcomes

Graphing predicted probs: by groups

```
logit hap_dic c.age##c.age i.female i.nonwhite i.educat i.married ///
if nmiss==0, or
margins educat, at(age=(18(1)89)) atmeans
marginsplot, noci legend(cols(3))
```




```
mgen, at(age=(18(1)89) educat=0) atmeans replace stub(PLT1) predlab(LTHS)
mgen, at(age=(18(1)89) educat=1) atmeans replace stub(PLT2) predlab(HS)
mgen, at(age=(18(1)89) educat=2) atmeans replace stub(PLT3) predlab(GTHS)
twoway connected PLT1pr1 PLT2pr1 PLT3pr1 PLT1age, ///
title("Adjusted Predictions by Educat") ///
caption("Other variables at their means") ///
ytitle("Pr(Hap_Dic)") legend(cols(3))
```

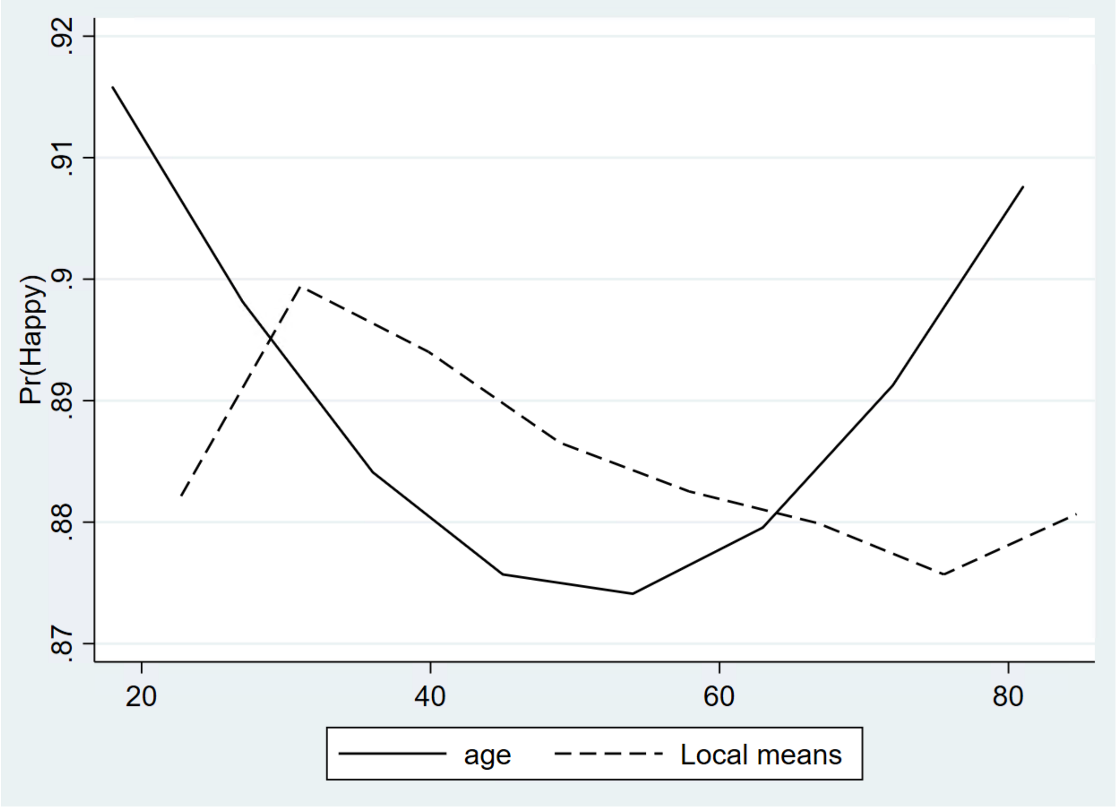


- There is often more than one way to get the same results
  - become familiar with mgen and other Spost commands
    - often more difficult than base Stata now, but it will be necessary later

Postestimation group differences

- Recall: when comparing postestimation statistics across groups
    - the CIs are conservative estimates
- Because ignores the covariance of the estimators
- Consider only hold few variables at specific values when plotting
    - rest set at global means
- Is this a reasonable assumption?
    - e.g., How might education, marriage, female, and nonwhite differ by age?
        - What else does age capture in cross-sectional data that span from 1972-2018?

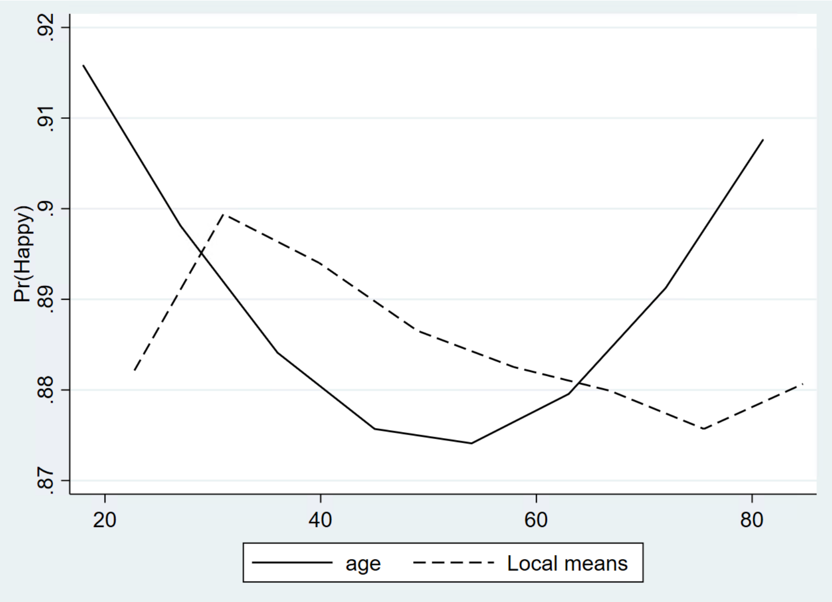
Postestimation group differences

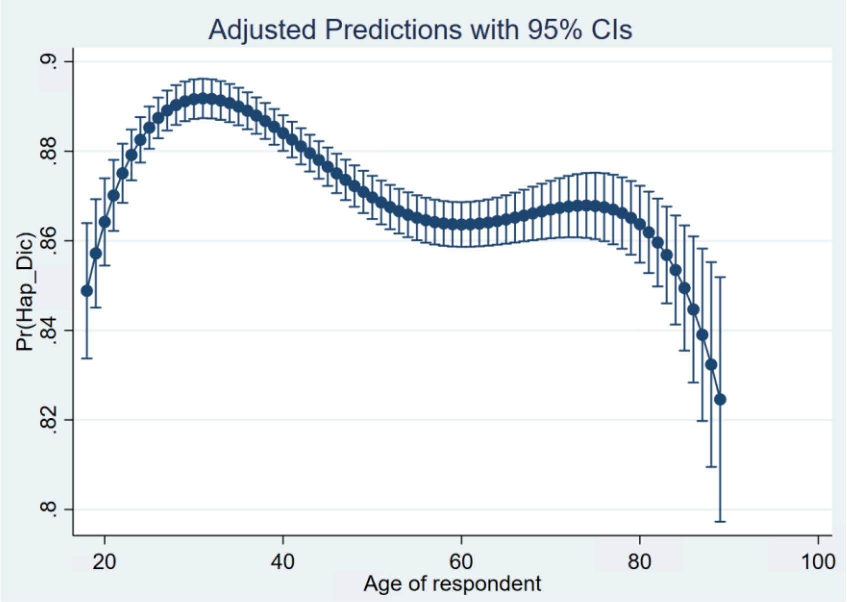


- Interpretation?
- There's no "one size fits all" approach for examining and interpreting glm results – theory

Postestimation group differences

Age quartic without covariates





- It looks like holding covariates constant at global means distorts the underlying age pattern in happiness
- Need theory!
  - and proper statistical techniques – does theory match assumptions

Next Monday we will…

- discuss ordinal outcomes
- read Hoffmann CH 4 and Long & Freese CH 7 before class