Quantitative Data Analysis II

SOC 781

GLM basics

Today we're going to…

- More info on descriptive and reg tables
- Discuss the basics of GLM
    - compare LRM and GLM
- Briefly introduce unique models for different outcomes
    - briefly discuss MLE
- Briefly cover techniques for assessing
    - and interpreting GLM output

Descriptive tables

- Descriptive tables
  - single table for ALL basics
    - additional for more than basics
  - only report on analytic sample
    - in this case, stratified
  - Only report (1) for binary measure
    - (0) deduced
  - Means and percentages can be reported in the same column
    - SD in (SD)
- Source: 12_Hua

Table 1. Descriptive statistics of the analytic sample

	Non-Hispanic blacks (n = 393; 24.90%)	Non-Hispanic whites (n = 1185; 75.10%)	Black vs White χ^2/t
Predisposing factors			
Female	0.69	0.61	−2.77**
Age	59.53 (12.83)	61.60 (14.64)	2.50**
Less than High School	0.34	0.16	−7.88***
Need factors			
Self-Rated Health	3.22 (1.03)	3.53 (1.01)	5.31***
Functional Limitations	1.56 (1.03)	1.41 (0.87)	−2.74**
Health Conditions Index	1.46 (1.30)	1.35 (1.24)	−1.58†
Enabling factors			
Employed	0.73	0.69	−1.18
Income	4.19 (2.60)	5.73 (2.40)	10.84***
Insured	0.94	0.96	1.21
Married	0.46	0.70	8.79***
Rural	0.22	0.35	5.04***
ED Used	0.32	0.24	−2.98***
PC Used	0.88	0.91	1.55
Trust in physicians			
Mistrust	1.44	1.49	1.29
Life course factors			211.72***
South-Never	0.28	0.64	12.79***
South-After 16	0.03	0.08	3.14***
South-Left After 16	0.17	0.04	−8.84***
Health care utilization			
PC Utilization	5.65 (8.99)	5.41 (8.89)	−0.46
ED Utilization	0.59 (1.13)	0.40 (1.08)	−2.90**

Note: †$p \le 0.10$; *$p \le 0.05$; **$p \le 0.01$; ***$p \le 0.001$.

Descriptive tables

- Descriptive tables
  - have some logical organization
    - in this case, theory
  - informative names
    - female, not gender
    - <HS, not education
    - married, not marital status
  - adequate spacing
    - make sure easy to read
    - I don't like lines
- Source: 12_Hua

Table 1. Descriptive statistics of the analytic sample.

	Non-Hispanic blacks (n = 393; 24.90%)	Non-Hispanic whites (n = 1185; 75.10%)	Black vs White χ^2/t
Predisposing factors			
Female	0.69	0.61	−2.77**
Age	59.53 (12.83)	61.60 (14.64)	2.50**
Less than High Schoo	0.34	0.16	−7.88***
Need factors			
Self-Rated Health	3.22 (1.03)	3.53 (1.01)	5.31***
Functional Limitations	1.56 (1.03)	1.41 (0.87)	−2.74**
Health Conditions Index	1.46 (1.30)	1.35 (1.24)	−1.58†
Enabling factors			
Employed	0.73	0.69	−1.18
Income	4.19 (2.60)	5.73 (2.40)	10.84***
Insured	0.94	0.96	1.21
Married	0.46	0.70	8.79***
Rural	0.22	0.35	5.04***
ED Used	0.32	0.24	−2.98***
PC Used	0.88	0.91	1.55
Trust in physicians			
Mistrust	1.44	1.49	1.29
Life course factors			211.72***
South-Never	0.28	0.64	12.79***
South-After 16	0.03	0.08	3.14***
South-Left After 16	0.17	0.04	−8.84***
Health care utilization			
PC Utilization	5.65 (8.99)	5.41 (8.89)	−0.46
ED Utilization	0.59 (1.13)	0.40 (1.08)	−2.90**

Note: †$p \leq 0.10$; *$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$.

- Descriptive tables
  - no single best approach
  - include most pertinent information
    - here, missing data was a major concern
  - measures section provides details
- Source: 15_ Zang_OSS

Table 1 Variable mean before and after imputation by country

	China			Japan			South Korea		
	Percentage missing (%)	Mean before imputation	Mean after imputation	Percentage missing (%)	Mean before imputation	Mean after imputation	Percentage missing (%)	Mean before imputation	Mean after imputation
Health outcomes									
Self-rated health	0.12	0.79	0.79	0.13	0.70	0.70	0.08	0.73	0.73
Depression	0.39	0.37	0.37	0.66	0.29	0.29	0.16	0.34	0.34
Chronic diseases	0.00	0.39	0.39	0.00	0.49	0.49	0.08	0.35	0.35
Objective social status	37.71			42.43			40.12		
Lower status		0.50	0.39		0.33	0.28		0.30	0.31
Lower-middle status		0.06	0.14		0.27	0.25		0.30	0.23
Upper-middle status		0.24	0.24		0.21	0.24		0.16	0.23
Upper status		0.20	0.23		0.18	0.23		0.24	0.23
Subjective social status	0.45			1.19			1.01		
Lower status		0.36	0.36		0.17	0.17		0.29	0.29
Lower-middle status		0.19	0.19		0.14	0.14		0.15	0.16
Upper-middle status		0.30	0.30		0.15	0.15		0.29	0.29
Upper status		0.14	0.14		0.54	0.54		0.26	0.26
Age	0.03	51.06	51.06	0.00	56.73	56.73		50.07	50.01
Marital status	0.54			0.04			0.31		
Married or cohabiting		0.86	0.86		0.78	0.78		0.76	0.76
Widowed, separated or divorced		0.12	0.12		0.13	0.13		0.16	0.15
Never married		0.02	0.02		0.09	0.09		0.09	0.09
Community type	0.00			0.22			0.62		
A big city		0.18	0.18		0.05	0.05		0.28	0.28
Suburbs or outskirts of a big city		0.07	0.07		0.16	0.16		0.27	0.27
A town or a small city		0.35	0.36		0.45	0.45		0.29	0.30
Country		0.39	0.39		0.35	0.35		0.15	0.15
Female	0.00	0.51	0.51	0.00	0.54	0.54	0.00	0.54	0.53
Have insurance	3.08	0.90	0.90	3.19	0.97	0.97	0.78	0.94	0.94
N		3307			2260			1281	

Table 1 (continued)

Regression Tables

- coefficients and p-values
  - are minimum
- often include SE or 95% CI
  - focus on readability
    - additional info in other, or supplementary, tables
- N should match descriptive table
  - naming convention too
- informative title
  - include notes to aid interpretation

Table 2. PC utilization: Relative risk ratios from Poisson regression N = 1578.

Variables	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Predisposing factors										
Black	1.04	1.28*	0.92	0.62**	0.76	0.68*	0.70*	0.67*	0.69	0.60**
Female	1.28***	1.27***	1.11***	1.11***	1.16***	1.17***	1.17***	1.17***	1.17***	1.16***
Age	1.01***	1.01***	0.99***	0.99*	0.99**	0.99*	0.99**	0.99**	0.99**	0.99**
Age × Black		0.99	1.00	0.99	0.99*	0.99	0.99	0.99	0.99	1.00
Less than HS	1.05	1.05	0.83***	0.83***	0.93**	0.92**	0.92**	0.92**	0.93*	0.94*
Need factors										
SRH			0.77***	0.75***	0.74***	0.75***	0.74***	0.74***	0.74***	0.74***
Functional Limit.			1.19***	1.18***	1.18***	1.17***	1.17***	1.17***	1.18***	1.18***
Health Conditions			1.18***	1.15***	1.14***	1.13***	1.13***	1.13***	1.13***	1.13***
SRH × Black				1.10***	1.07*	1.07*	1.06*	1.06*	1.06*	1.04
Func. Lim. × Black				1.02	1.03	1.04	1.03	1.03	1.02	1.03
Health Con. × Black				1.10***	1.10***	1.10***	1.10***	1.10***	1.10***	1.10***
Enabling factors										
Employed					1.01	1.02	1.02	1.02	1.03	1.02
Income					1.04***	1.04***	1.04***	1.04***	1.04***	1.04***
Insured					2.08***	2.09***	2.05***	2.05***	2.02***	1.95***
Married					0.94*	0.95*	0.94*	0.94*	0.96	0.96
Rural					0.89***	0.89***	0.89***	0.89***	0.91***	0.91***
ED Use						1.21***	1.21***	1.21***	1.21***	1.21***
Trust in physicians										
Mistrust							0.93***	0.93***	0.93***	0.93***
Mistrust × Black								1.02	1.00	0.97
Life course factors										
South-Never									1.15***	1.07*
South-After 16									1.39***	1.17**
South-Left After 16									0.97	0.87
South-Never × Black										1.25***
South-After 16 × Black										2.25***
South-Left After 16 × Black										1.26**

*Note: 95% CIs provided in Supplementary Table 1; †p ≤ 0.10; *p ≤ 0.05; **p ≤ 0.01; ***p ≤ 0.001.*

Today's class...

- Focus on big picture
  - But stop me if have question: I may move on if in weeds

- Some probability background to justify
- GLM vs LRM: model choice depends on measurement of Y
- Different GLM models (Link functions)
  - Just need to know what model for what outcome
- MLE vs OLS: just different ways under hood to fit models
- GLM model fit
- GLM interpretation of results
- Play with my examples afterwards, then mess with your data.
  - More applied next week: logit

Generalized linear models (GLM)

- accommodate many types of outcomes
  - binary, ordinal, nominal, count

- by computing most likely values of β given the observed data
  - less restrictive than linear regression model (LRM)
    - can fit nonlinear models w/o special transformations of Y

- usually by using Maximum Likelihood Estimation (MLE)
  - rather than minimizing the sum of square residuals (OLS)

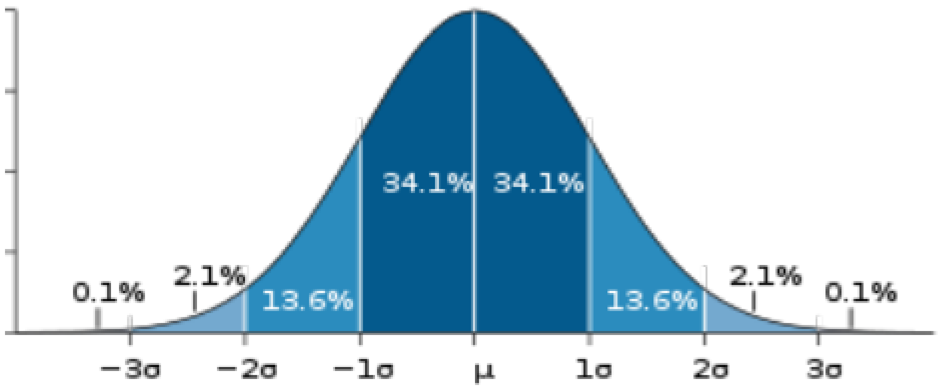
Probability: basics

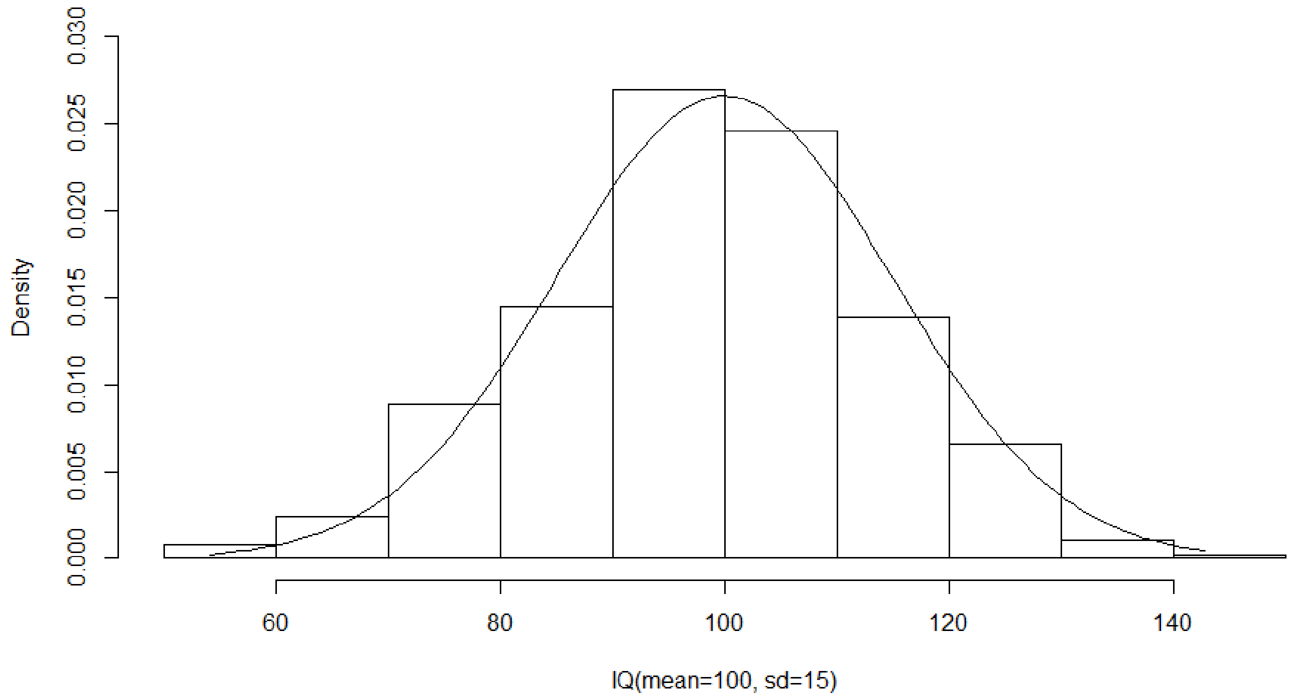
- Continuous outcomes (normal distribution)
  - closer to mean → greater likelihood to observe
- Peak of the function (mean of the distribution) most likely observed value
- IQ distributed normally, with mean 100 and SD 15
  - sample random individual
- What is more likely:
  - IQ = 100 or IQ = 80?

The 68-95-99.7 Rule

In the Normal distribution with mean μ and standard deviation σ:

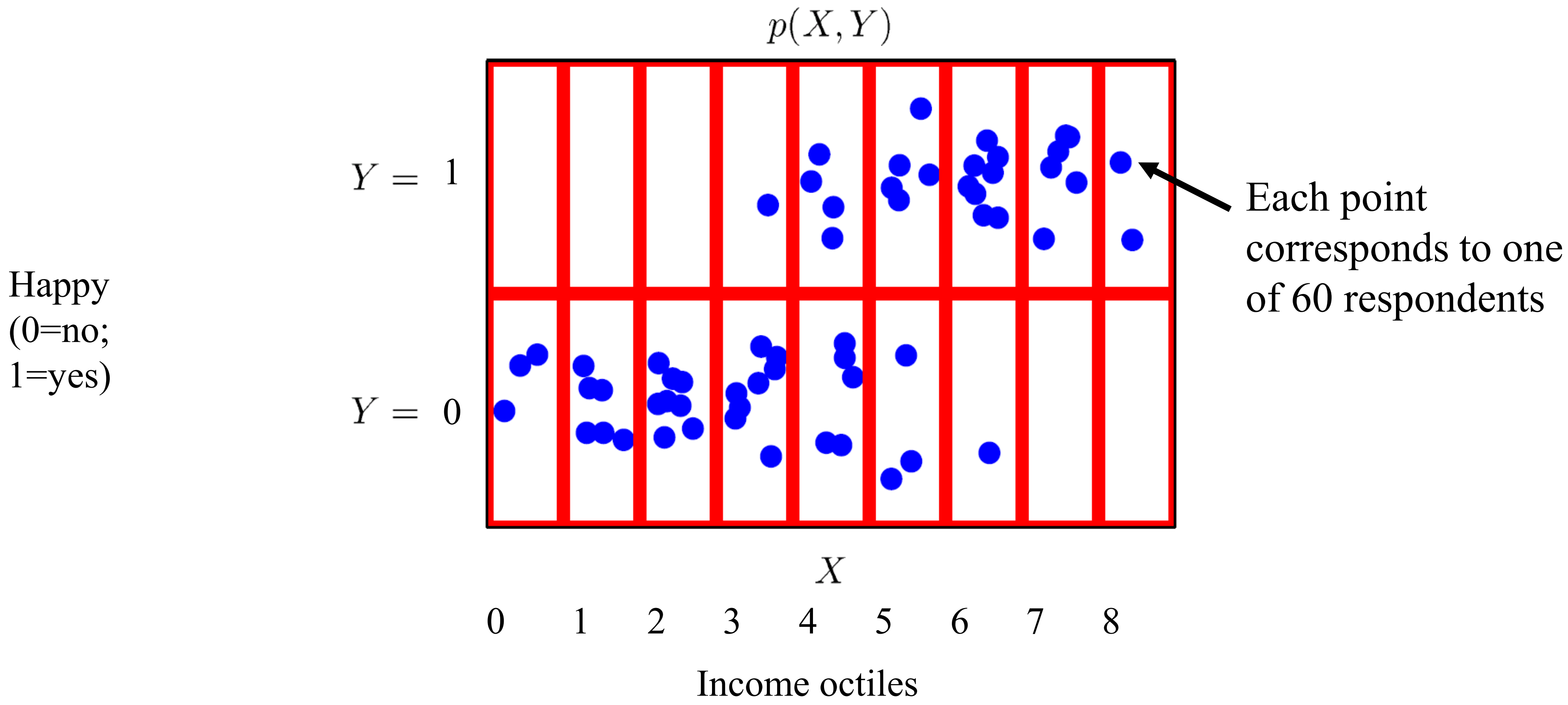
- Approximately **68%** of the observations fall within σ of μ.
- Approximately **95%** of the observations fall within 2σ of μ.
- Approximately **99.7%** of the observations fall within 3σ of μ.





- Let’s suppose that:
  - We don’t know the mean
  - We pick 2 random observations: IQ=[80,100]
  - We assume IQ is normally distributed
- What’s best guess about value of the mean?
- MLE designed to address this issue
  - we know the data, we assume a distribution, and we estimate the parameters

Why GLM: ask 60 ppl. if happy (1) or not (0)



Probability: happy (0,1)

- $P(X{=}i, Y{=}j)$: probability (relative frequency) of observing $X = i$ and $Y = j$
- Properties: $0 \leq P \leq 1, \;\; \sum P = 1$

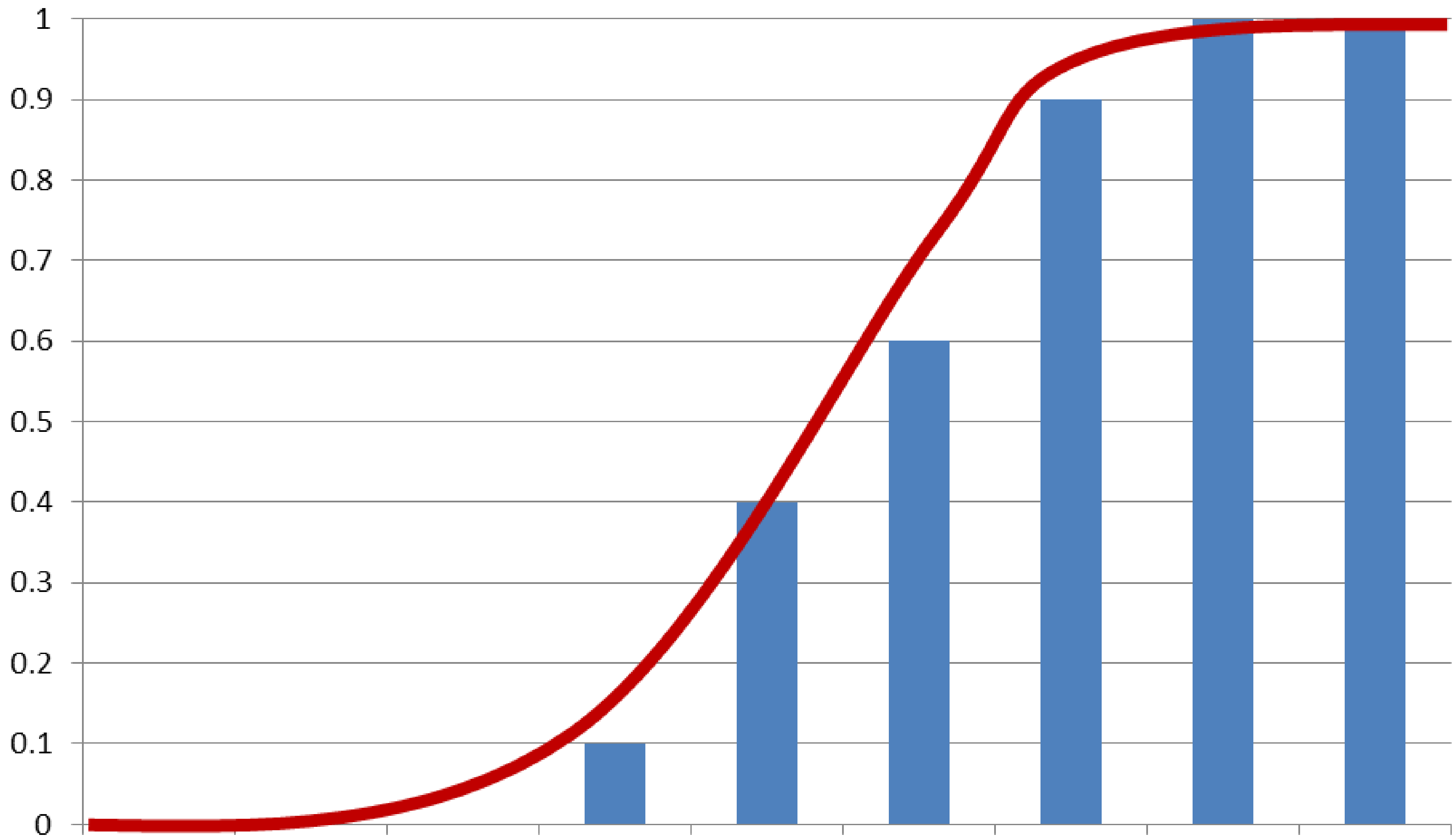
	0	1	2	3	4	5	6	7	8
Y=1	$\frac{0}{60}$	$\frac{0}{60}$	$\frac{0}{60}$	$\frac{1}{60}$	$\frac{4}{60}$	$\frac{5}{60}$	$\frac{8}{60}$	$\frac{6}{60}$	$\frac{2}{60}$
Y=0	$\frac{3}{60}$	$\frac{6}{60}$	$\frac{8}{60}$	$\frac{8}{60}$	$\frac{5}{60}$	$\frac{3}{60}$	$\frac{1}{60}$	$\frac{0}{60}$	$\frac{0}{60}$

Conditional probability: happy by income

- *Conditional* probability $P(Y=j|X=i)$
- What is the probability of being happy if income is 4?
- General pattern: Probability of being happy increases as income increases

Y=1	$\frac{0}{3}$ 0.0	$\frac{0}{6}$ 0.0	$\frac{0}{8}$ 0.0	$\frac{1}{9}$ 0.1	$\frac{4}{9}$ 0.4	$\frac{5}{8}$ 0.6	$\frac{8}{9}$ 0.9	$\frac{6}{6}$ 1.0	$\frac{2}{2}$ 1.0
Y=0	$\frac{3}{3}$ 1.0	$\frac{6}{6}$ 1.0	$\frac{8}{8}$ 1.0	$\frac{8}{9}$ 0.9	$\frac{5}{9}$ 0.6	$\frac{3}{8}$ 0.4	$\frac{1}{9}$ 0.1	$\frac{0}{6}$ 0.0	$\frac{0}{2}$ 0.0
	0	1	2	3	4	5	6	7	8

P of being happy conditional on income



Conditional probability

- Becomes extremely complex with multiple covariates (just like LRM), but still trying to describe effects of X on Y
  - Now, it's the probability of Y given X expressed in terms of odds

- Interpretation is relative, so odds are somewhat meaningless
  - → rely heavily on postestimation techniques

- Unique models for certain outcomes

Outcome	Model	
Binary	Logit and probit	← Next week
Ordinal	Ordered logit and probit	← FEB 26
Nominal	Multinomial logit	← MAR 18
Count	Poisson and negative binomial regression	← MAR 25

LRM vs GLM

- LRM: the expected value of Y: or $E(Y)$
- can also be expressed as the conditional mean: or μ
- which equals the sum of the intercept β_0, the reg. coefs. $\beta_{1\ldots i}$, and the respective independent variables $X_{1\ldots i}$
- Condensed using linear predictor eta: η

$$\eta = \sum_{k=1}^{K} \beta_K X_K$$

- When conditional mean makes no sense
  - Y is not a continuous normally distributed variable
- use GLM to “link” η and μ

LRM vs GLM: example

- Now we are going to work with categorical outcomes
  - binary, ordinal, nominal, count

- The first outcomes we will work with are binary
  - only two categories; dichotomous (0, 1)

- We can operationalize ordinal happiness measure as dichotomy by collapsing the three response categories into two
  - happy (1) vs. unhappy (0)

- Let's first examine this binary measure using LRM with OLS

Happy (1) vs. Unhappy (0): LRM

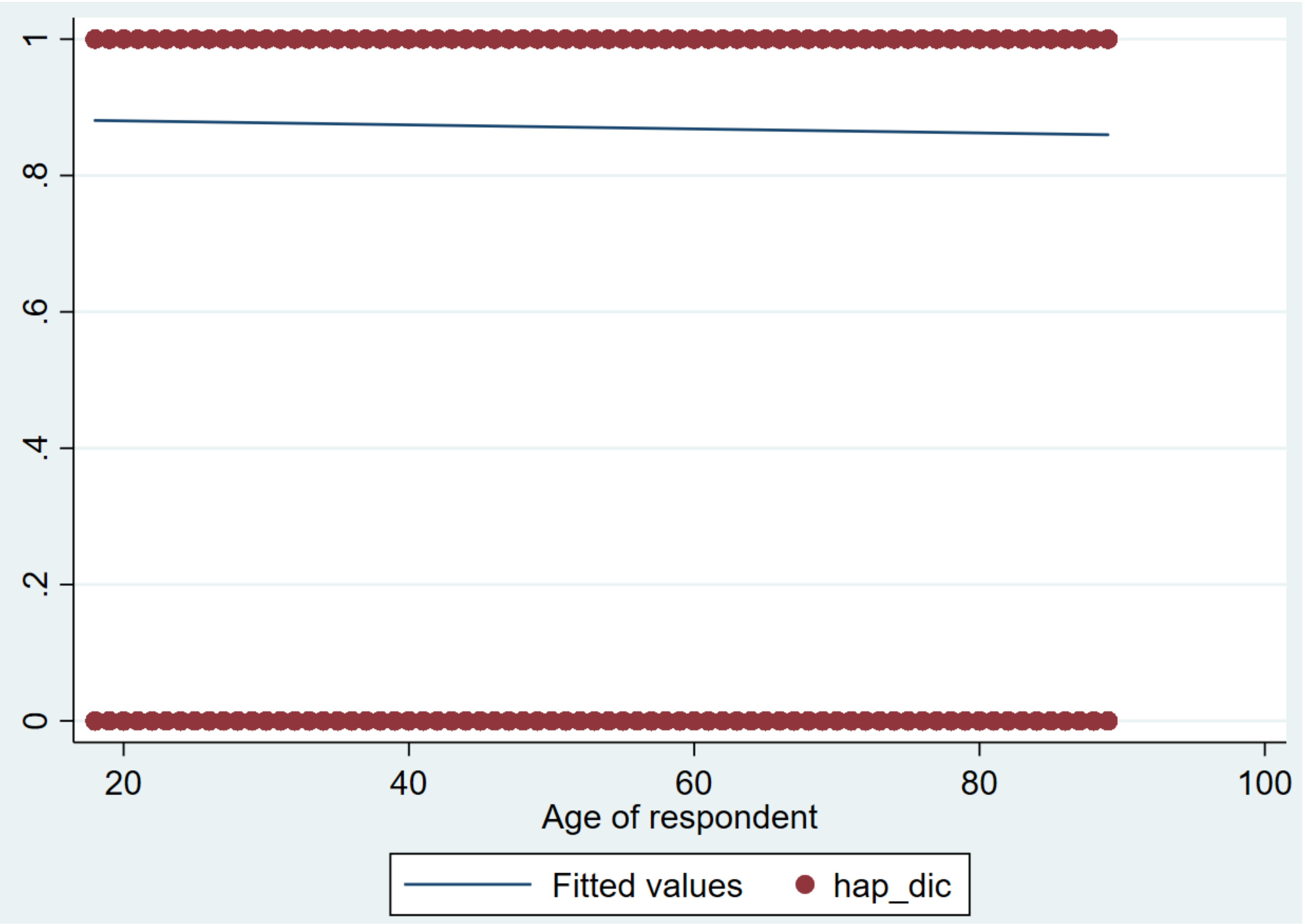
- Stata will give you results
- Hey, look! They're even significant!!!
- But what do they mean…
  - Why NOT use binary in Sobel mediation

```
reg hap_dic age if nmiss==0

      Source |       SS           df       MS      Number of obs   =    59,725
-------------+----------------------------------   F(1, 59723)     =     14.08
       Model |  1.56594153         1  1.56594153   Prob > F        =    0.0002
    Residual |  6641.76858    59,723  .11120956    R-squared       =    0.0002
-------------+----------------------------------   Adj R-squared   =    0.0002
       Total |  6643.33452    59,724  .111233918   Root MSE        =    .33348

------------------------------------------------------------------------------
     hap_dic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0002914   .0000776    -3.75   0.000    -.0004435   -.0001392
       _cons |   .8859318   .0038268   231.51   0.000     .8784312    .8934324
------------------------------------------------------------------------------
```

Happy (1) vs. Unhappy (0): LRM

- Interpret the slope
- What is the predicted value of happiness at age ___?

```
graph twoway (scatter hap_dic age) (lfit hap_dic age) if nmiss==0
```

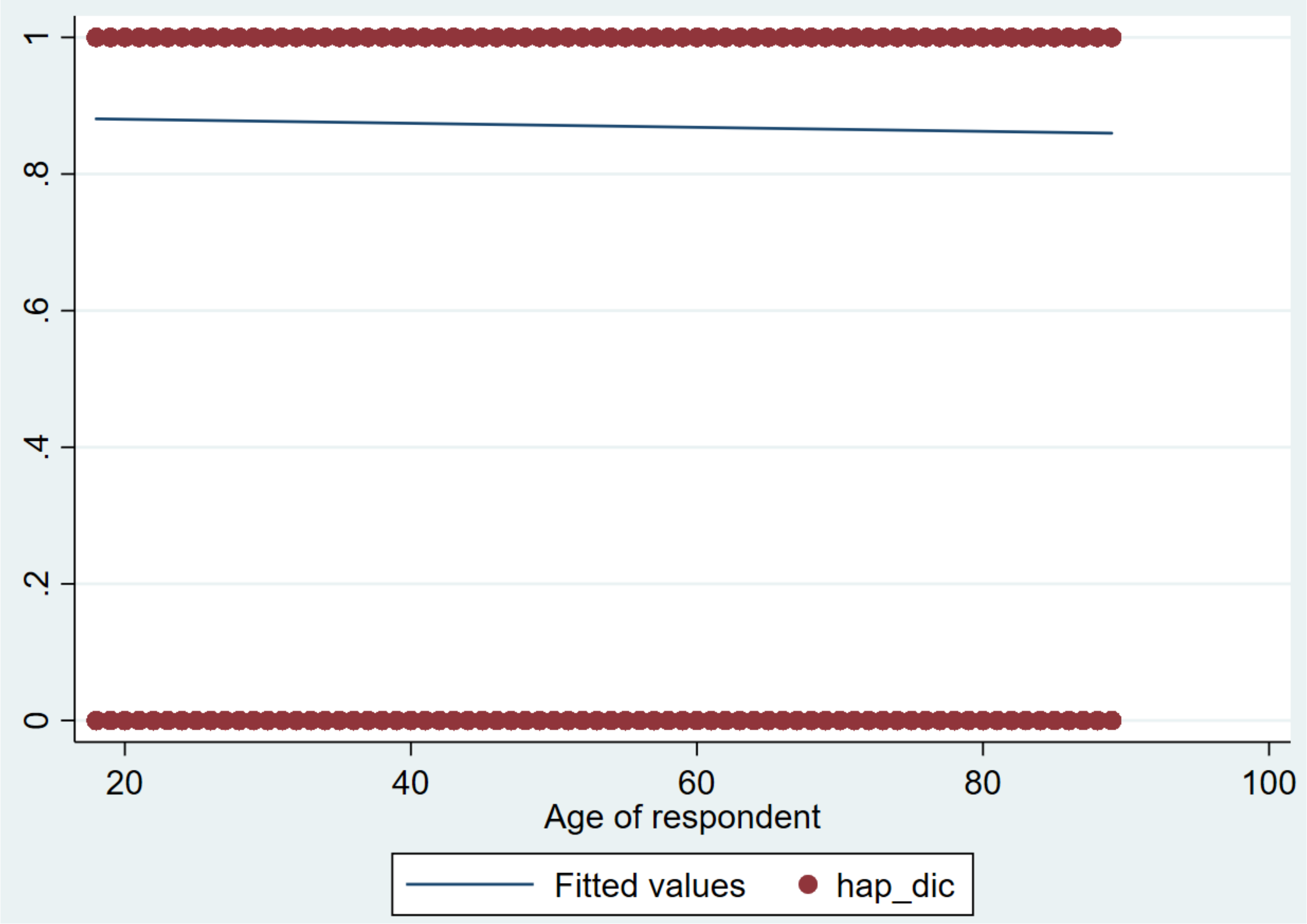


Happy (1) vs. Unhappy (0): LRM

- Linearity

- Is the linear functional form reasonable?

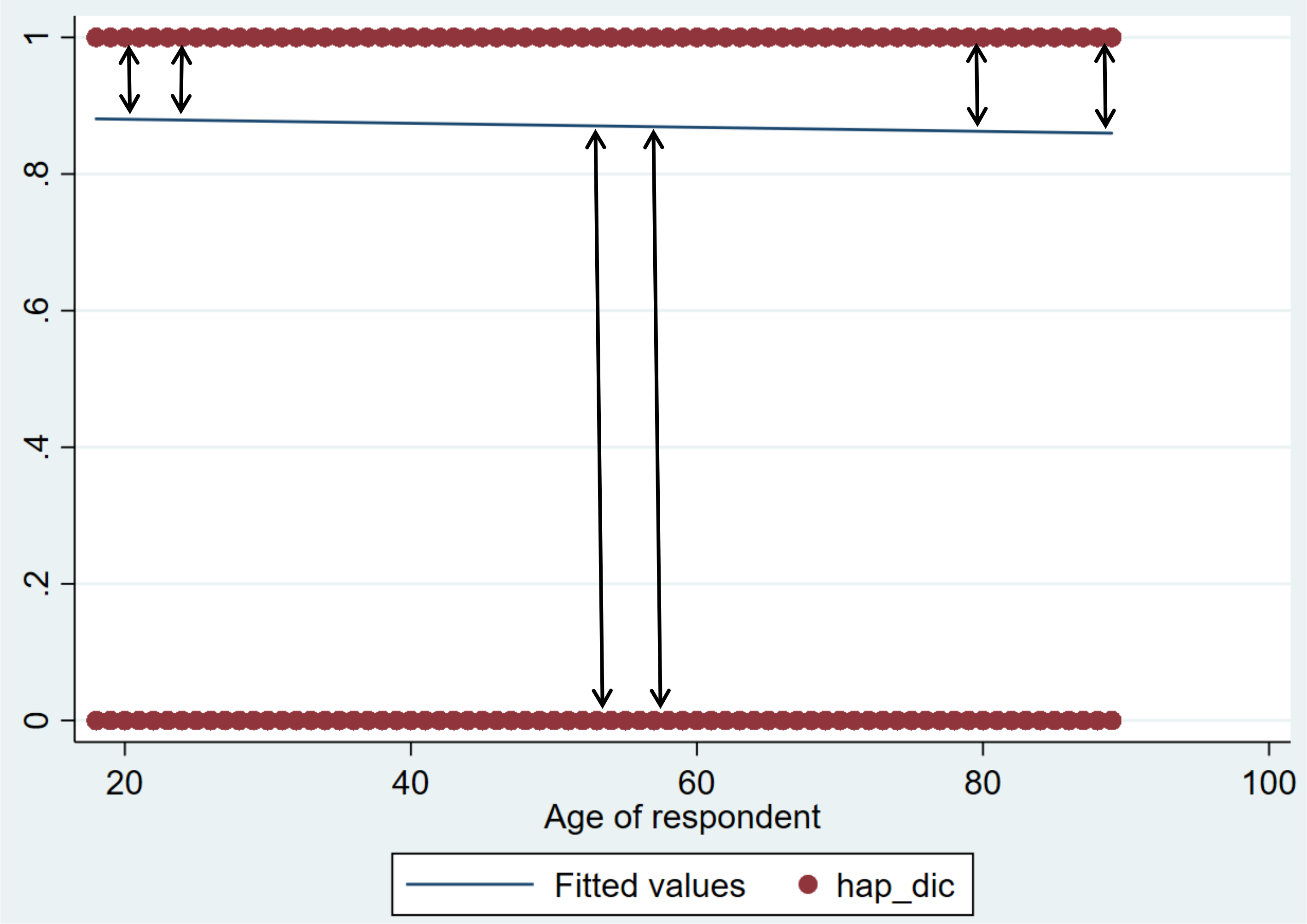
```
graph twoway (scatter hap_dic age) (lfit hap_dic age) if nmiss==0
```



Happy (1) vs. Unhappy (0): LRM

- Heteroscedasticity
- Is the error variance constant?
  - distance between observed and fitted values

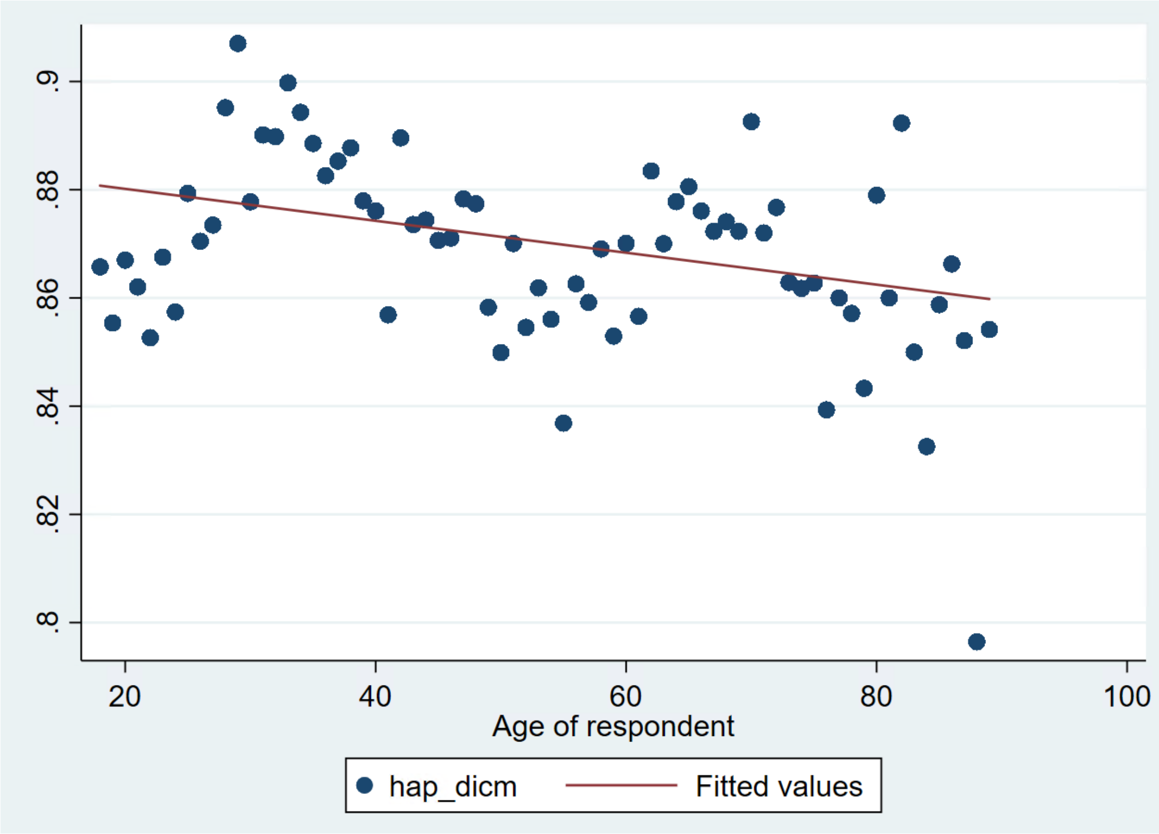
```
graph twoway (scatter hap_dic age) (lfit hap_dic age) if nmiss==0
```



Happy (1) vs. Unhappy (0): LRM

- What do these results tell us?
  - linear slope of age between mean values of happiness
- What are mean values of hap_dic?
  - percent happy at each year of age
- The identity function isn't appropriate for modeling dichotomous outcomes
  - need to use another function

```
graph twoway (scatter hap dicm age) (lfit hap dicm age) if nmiss==0
```



GLM

- Not required to understand how link functions work under the hood

- Know which model is appropriate for different types of outcomes
  - and why

- Be able to use these models, diagnose fit, and interpret outcomes

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
- Or, an inverse normal link: η = $\Phi^{-1}(\mu)$
  - probit model

Ordered outcomes

- The logit and inverse normal link also used for ordered outcomes
  - ordered logistic and ordered probit models
- Consider the original GSS happy measure
  - (0) not too happy, (1) pretty happy, (2) very happy
- As you'll see, these variables have "cut-points"
  - we have to make assumptions about

Nominal outcomes

- When there are more than two categories but NO meaningful order
  - need to account for unordered nature
- Thus, need to use another link function
- Generalized logit link: $\eta_j = \log_e\left(\frac{\mu_j}{\mu_J}\right)$
  - Multinomial Logistic Regression

Count outcomes

- Count measures can not fall below zero
  - unlike continuous measures
- e.g., the number of children one expects to have
- Thus, need to use another link function
- Natural logarithm link: η = $\log_e \mu$
  - Poisson and negative binomial models

GLM

- Not required to understand how link functions work under the hood
- Know which model is appropriate for different types of outcomes
  - and why
- Be able to use these models, diagnose fit, and interpret outcomes

Modeling technique

- Typically uses maximum likelihood estimation (MLE)

- Finds unknown parameters that make observed combination of X^s and Y^s most likely to occur

- Calculates how likely a set of outcome values are if a set of parameter estimates were true

- Keeps doing this (iterations) until the maximum of the likelihood function is found (convergence)
  - luckily, Stata does most of the work

MLE in Stata

- The iteration process is shown at the top of the Stata log

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

Iteration 0:   log likelihood = -22789.647
Iteration 1:   log likelihood = -21556.708
Iteration 2:   log likelihood = -21477.531
Iteration 3:   log likelihood = -21477.296
Iteration 4:   log likelihood = -21477.296
```

- Convergence was reached in fourth iteration

MLE in Stata: factor notation

- Thus far, not consistent with factor notation
  - just getting familiar with it
- Now, need to be more consistent
  - post estimation techniques rely on it
- As you'll see, we'll be using a lot of post estimation
- This can change variable name
  - use coeflegend

```
logit, coeflegend
```

hap_dic	Coef.	Legend
age	-.0413751	_b[age]
c.age#c.age	.0004016	_b[c.age#c.age]
1.female	.0937882	_b[1.female]
1.nonwhite	-.4425555	_b[1.nonwhite]
educat		
0	-.4118941	_b[0.educat]
2	.3627807	_b[2.educat]
1.married	1.026049	_b[1.married]
_cons	2.415824	_b[_cons]

MLE model fit

- Likelihood Ratio (LR) test: constrained model [intercept only ($Y = \alpha$)] vs. unconstrained model [includes all independent variables]
  - Basically: do the independent variables “explain” Y better than nothing
    - more useful for comparing nested models

- McFadden (pseudo) adjusted R^2
  - not a very interpretable, or agreed upon, statistic

- Information measures: comparing models
  - Akaike’s Information Criterion (AIC): smaller is better
  - Bayesian Information Criterion (BIC): smaller is better

MLE in Stata: model fit

```
logit hap_dic c.age##c.age i.female i.nonwhite i.married if nmiss==0
fitstat, saving(noedu)
logit hap_dic c.age##c.age i.female i.nonwhite ib1.educat i.married if nmiss==0
fitstat, using(noedu)
```

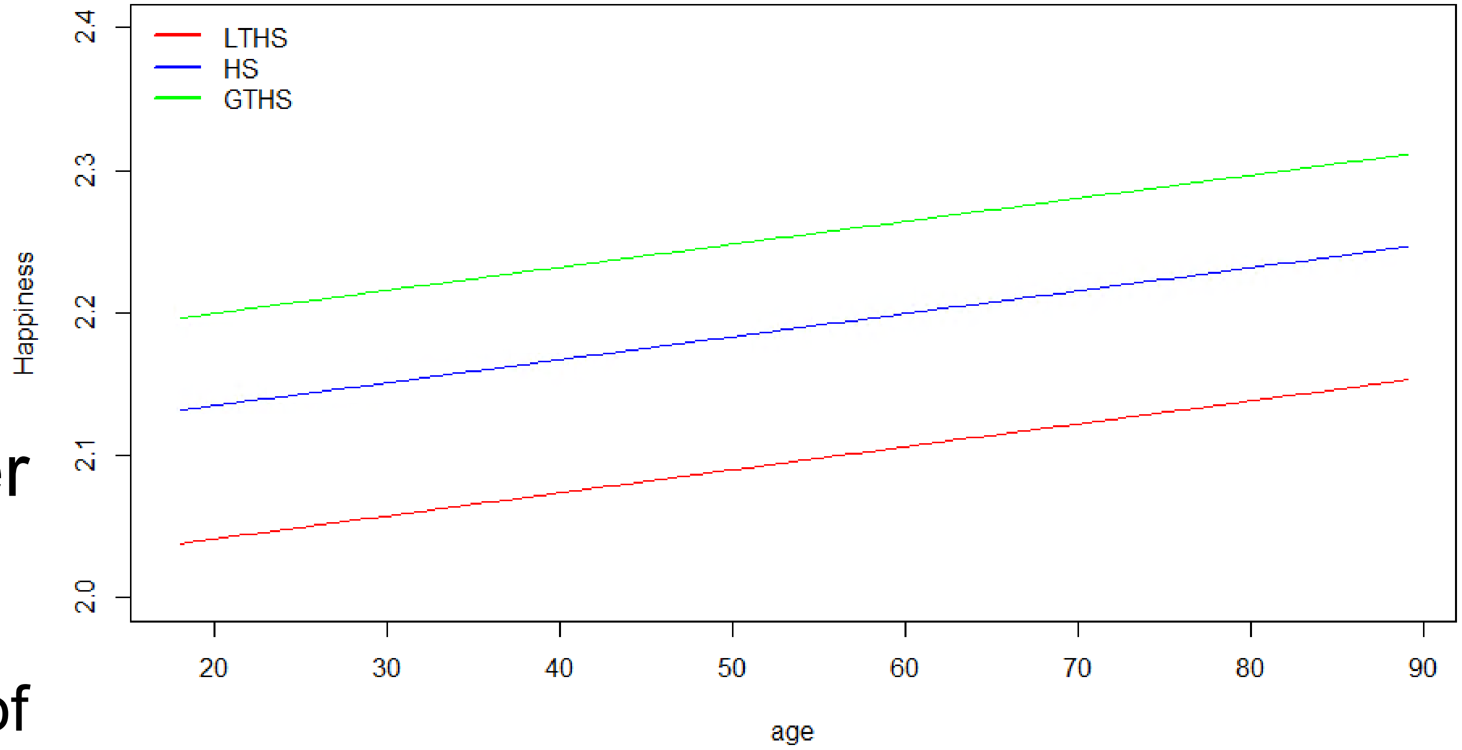
- Save fit statistics and compare to other models
  - Which is the better fitting model?

	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

Note: Likelihood-ratio test assumes saved model nested in current model.

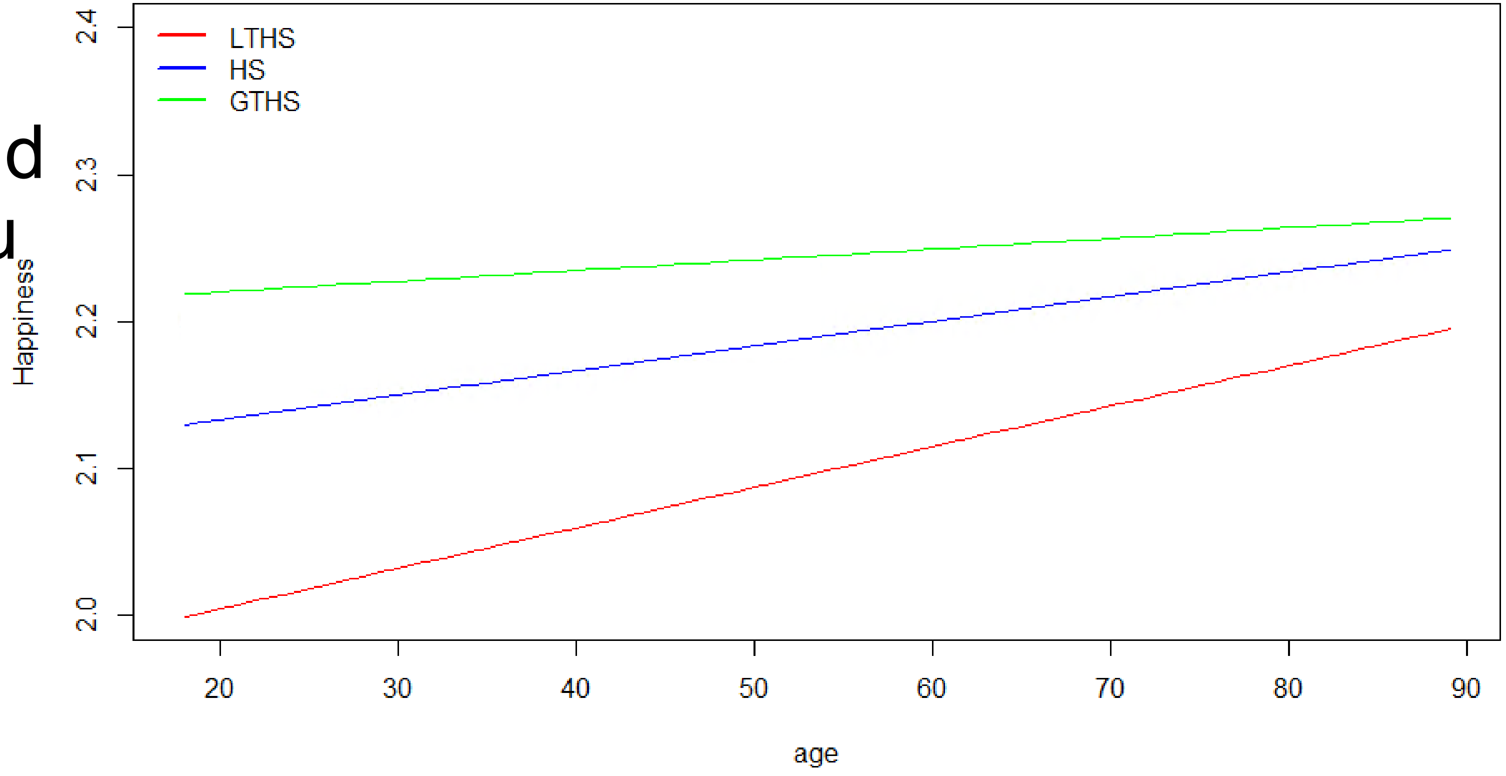
LRM vs GLM interpretation

- Recall our happiness-age by education example
- B_{age} = Δ in happiness for every one-year Δ in age
  - holding education constant
- Linear interpretation no longer applies with GLM
  - because the effect of the Δ in one X depends on the values of all other X^s



OLS vs GLM interpretation

- This is sort of like what we did when we introduced age*edu interaction terms
  - making model non-linear



GLM interpretation

- Estimated parameters typically don’t make much sense
  - used to make predictions at meaningful values of the Xs via postestimation
- Certain techniques are more useful to address specific issues
  - that’s why we’ll learn many different techniques
- Some of this may not make sense right now – that’s okay
  - let’s become familiar with the techniques and Stata programming first

GLM interpretation

- Predictions for each observation
  - a good starting point

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

Variable	Obs	Mean	Std. Dev.	Min	Max
prob	59,725	.8725157	.0722409	.6214635	.967612

- Mean probability of happy is 0.87
  - range: 0.62 to 0.97

- Try this with a different combination of IVs
  - Same or different? Why?

GLM interpretation

- Predictions at specified values: substantively interesting combinations of X values → profiles or ideal types
  - e.g., the average respondent (ALL Xs set at mean)

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

```
Expression   : Pr(hap_dic), predict()
at           : age             =    46.04745 (mean)
               0.female        =    .4422101 (mean)
               1.female        =    .5577899 (mean)
               0.nonwhite      =    .8068648 (mean)
               1.nonwhite      =    .1931352 (mean)
               0.educat        =    .2285308 (mean)
               1.educat        =    .3058853 (mean)
               2.educat        =    .4655839 (mean)
               0.married       =    .4702721 (mean)
               1.married       =    .5297279 (mean)
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.8751479	.0019592	446.69	0.000	.871308	.8789879

- Probability of being happy vs. not happy is 0.875 (CI 0.871-0.879), on average

- Often doesn't make too much sense
  - e.g., what's 0.55 female?

GLM interpretation

- Profiles or ideal types: set Xs at meaningful values
  - factor notation makes this easy when including interactions or polynomials

```
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married if nmiss==0
margins, at(age=30 female=1 nonwhite=0 educat=2 married=1)
```

	Delta-method Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.9534082	.0015001	635.55	0.000	.9504679	.9563484

```
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married if nmiss==0
margins, at(age=30 female=1 nonwhite=0 educat=0 married=1)
```

	Delta-method Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.9041268	.0031965	282.85	0.000	.8978617	.9103918

- Probability of being happy vs. not happy is 0.953 (CI 0.950-0.956) for age 30, female, white, college educated, and married
- What if we change education to <HS?
  - Whose more likely to be happy?

GLM interpretation

- Profiles or ideal types: set Xs at meaningful values
  - can simplify syntax for hypothetical comparisons

```
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married if nmiss==0
margins, at(age=30 female=1 nonwhite=0 educat=0 educat=1 educat=2 married=1)
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	.9041268	.0031965	282.85	0.000	.8978617	.9103918
2	.9343698	.0020958	445.83	0.000	.9302621	.9384776
3	.9534082	.0015001	635.55	0.000	.9504679	.9563484

GLM interpretation

- Profiles or ideal types: set Xs at meaningful values
  - things can get messy quick, so make sure you know what you're looking at

```
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married if nmiss==0
margins, at(age=(20(10)90) female=1 nonwhite=0 educat=0 educat=1 educat=2 ///
married=1) noatlegend
```

_at	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
1	.9210645	.0034122	269.93	0.000	.9143767	.9277524
2	.9462821	.0022886	413.48	0.000	.9417965	.9507676
3	.9620053	.001652	582.33	0.000	.9587675	.9652432
4	.9041268	.0031965	282.85	0.000	.8978617	.9103918
5	.9343698	.0020958	445.83	0.000	.9302621	.9384776
6	.9534082	.0015001	635.55	0.000	.9504679	.9563484
7	.891997	.0033388	267.16	0.000	.8854531	.8985409
8	.9257522	.0022075	419.36	0.000	.9214255	.9300789
9	.9471489	.0015711	602.87	0.000	.9440697	.9502282
10	.8868522	.0034209	259.24	0.000	.8801473	.893557
11	.9220748	.0023318	395.44	0.000	.9175046	.926645
12	.9444678	.0016687	566.00	0.000	.9411972	.9477383
13	.8896312	.0032111	277.05	0.000	.8833376	.8959249
14	.9240628	.0022839	404.60	0.000	.9195865	.9285391
15	.945918	.0016564	571.05	0.000	.9426714	.9491646
16	.8998262	.0030433	295.67	0.000	.8938614	.9057909
17	.9313228	.0022766	409.09	0.000	.9268608	.9357848
18	.9511988	.0016804	566.05	0.000	.9479053	.9544924
19	.9155934	.0034393	266.22	0.000	.9088525	.9223343
20	.9424496	.0025916	363.66	0.000	.9373702	.947529
21	.9592463	.0019174	500.29	0.000	.9554884	.9630043
22	.9341881	.0041649	224.30	0.000	.926025	.9423512
23	.955416	.0030437	313.90	0.000	.9494505	.9613814
24	.9685546	.0022238	435.53	0.000	.9641959	.9729133

mlistat

_at	age	educat
1	20	0
2	20	1
3	20	2
4	30	0
5	30	1
6	30	2
7	40	0
8	40	1
9	40	2
10	50	0
11	50	1
12	50	2
13	60	0
14	60	1
15	60	2
16	70	0
17	70	1
18	70	2
19	80	0
20	80	1
21	80	2
22	90	0
23	90	1
24	90	2

- Interpretation?

GLM interpretation

- Margins is a base Stata command for postestimation
  - mtable is a powerful user-created package that simplifies margins

```
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married if nmiss==0
mtable, at(age=(20(10)90) female=1 nonwhite=0 educat=0 educat=1 educat=2 ///
married=1) statistics(ci)
```

	age	educat	Pr(y)	ll	ul
1	20	0	0.921	0.914	0.928
2	20	1	0.946	0.942	0.951
3	20	2	0.962	0.959	0.965
4	30	0	0.904	0.898	0.910
5	30	1	0.934	0.930	0.938
6	30	2	0.953	0.950	0.956
7	40	0	0.892	0.885	0.899
8	40	1	0.926	0.921	0.930
9	40	2	0.947	0.944	0.950
10	50	0	0.887	0.880	0.894
11	50	1	0.922	0.918	0.927
12	50	2	0.944	0.941	0.948
13	60	0	0.890	0.883	0.896
14	60	1	0.924	0.920	0.929
15	60	2	0.946	0.943	0.949
16	70	0	0.900	0.894	0.906
17	70	1	0.931	0.927	0.936
18	70	2	0.951	0.948	0.954
19	80	0	0.916	0.909	0.922
20	80	1	0.942	0.937	0.948
21	80	2	0.959	0.955	0.963
22	90	0	0.934	0.926	0.942
23	90	1	0.955	0.949	0.961
24	90	2	0.969	0.964	0.973

- Same results as previous slide

- A little more condensed
  - easier to interpret

- mtable is also useful when dealing with categorical and count outcomes
  - as you'll see things get more complex when Y is not dichotomous

GLM interpretation

- Marginal effects: how changes in one variable are associated with changes in the outcomes, holding all else constant

```
logit hap_dic c.age##c.age i.female i.nonwhite ibl.educat i.married if nmiss==0
margins, dydx(*)
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.000418	.0000691	-6.05	0.000	-.0005533	-.0002826
1.female	.0099826	.0027294	3.66	0.000	.0046331	.0153322
1.nonwhite	-.0511805	.0036188	-14.14	0.000	-.0582732	-.0440878
educat						
0	-.0521551	.0041551	-12.55	0.000	-.0602989	-.0440113
2	.0352867	.0030513	11.56	0.000	.0293062	.0412671
1.married	.1085001	.0028608	37.93	0.000	.102893	.1141071

- On average, the probability of happiness decreases by 0.052 moving from a HS education to <HS education, holding all else constant

GLM interpretation

- Marginal effects: can get the same with mchange
  - will be more useful as we progress

```
logit hap_dic c.age##c.age i.female i.nonwhite ib1.educat i.married if nmiss==0
mchange, statistics(ci)
```

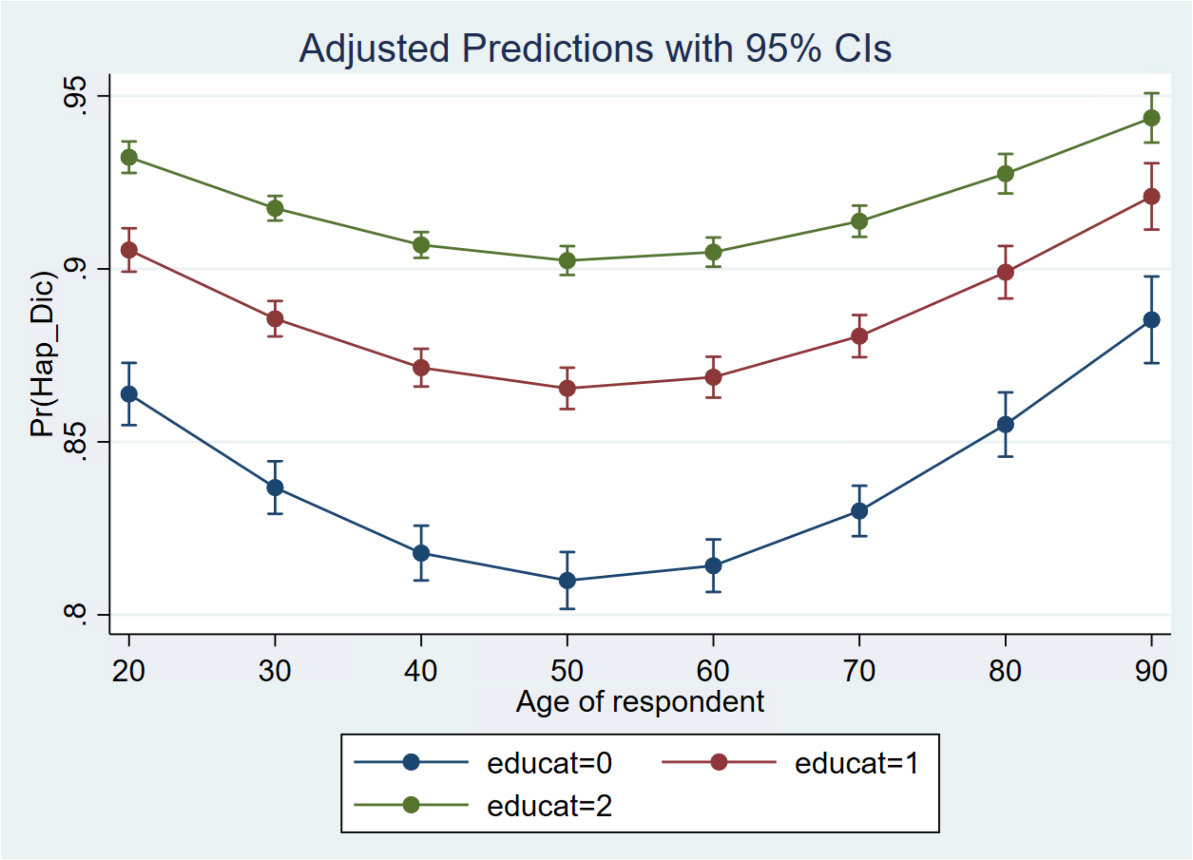
	Change	LL	UL
age			
+1	-0.000	-0.001	-0.000
+SD	0.003	0.001	0.006
Marginal	-0.000	-0.001	-0.000
female			
1 vs 0	0.010	0.005	0.015
nonwhite			
1 vs 0	-0.051	-0.058	-0.044
educat			
1 vs 0	0.052	0.044	0.060
2 vs 0	0.087	0.080	0.095
2 vs 1	0.035	0.029	0.041
married			
1 vs 0	0.109	0.103	0.114

- Can anyone figure out how to display the estimates for age?

GLM interpretation

- Plotting probabilities useful for continuous Xs

```
logit hap_dic c.age##c.age i.female i.nonwhite ib1.educat i.married if nmiss==0
margins, at(age=(20(10)90) educat=(0 1 2)) atmeans
marginsplot
```



Next Monday we will…

- discuss logit and probit models (binary outcomes)

- Read Hoffman CH3 and Long & Freese CH 5 & 6 before class