Quantitative Data Analysis II

SOC 781

Descriptive statistics and statistical modeling

Today we’re going to…

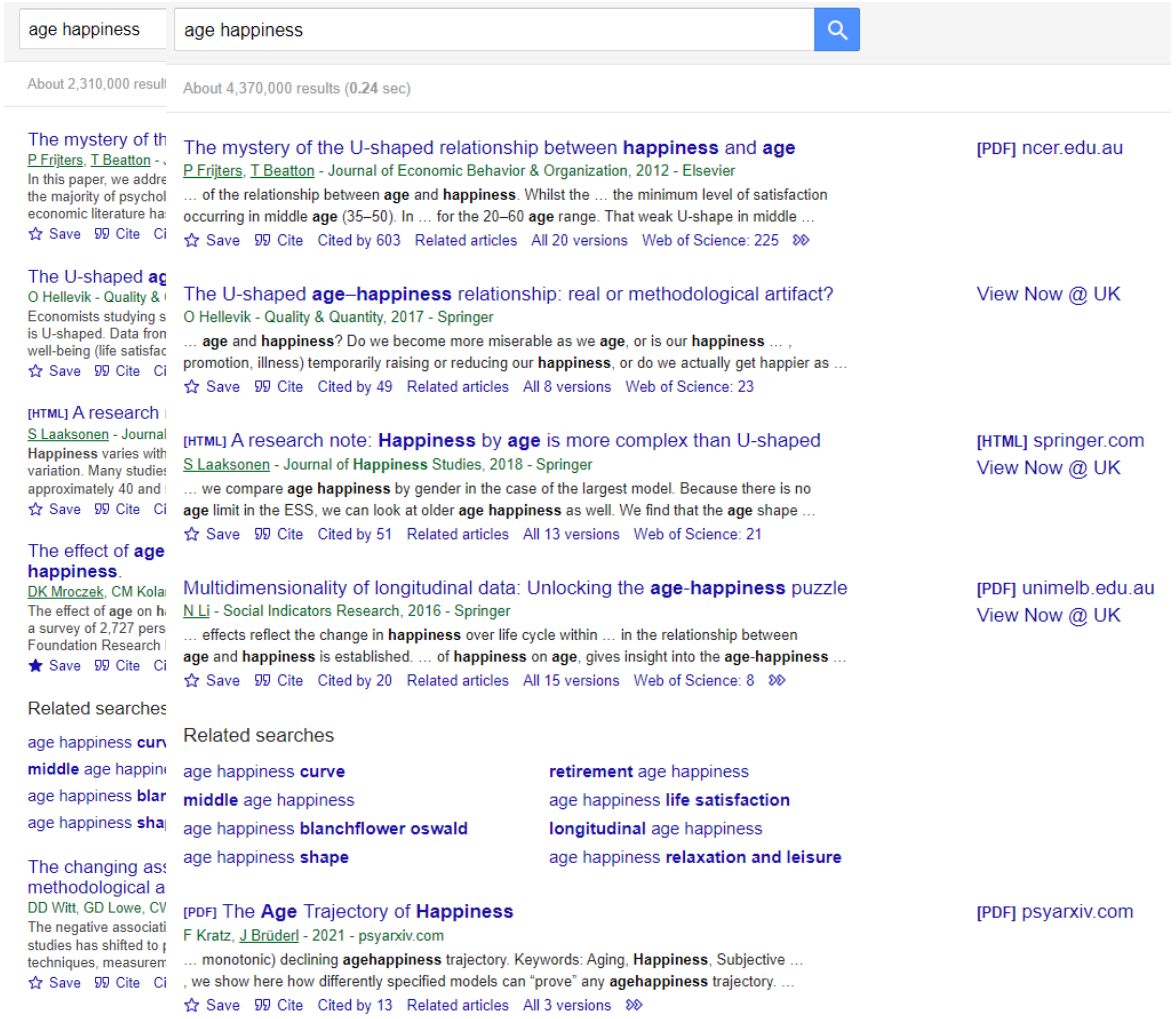
- Discuss descriptives and modeling considerations
  - work with graphics and a little in R

Lies, damn lies, and statistics

- Descriptive statistics are an important first step,
  - but they don’t tell the whole story
- That’s why we need inferential statistics,
  - but only make sense with theory
- Stata output not always easily interpretable,
  - graphing is a very helpful tool
- Today’s examples: methods need to match theory
  - don’t get overwhelmed; don’t try to follow along in Stata; take it in
  - focus on big picture; play with examples in Stata later; apply to own data

Graphing descriptives and regression results

- There's a huge literature on the age-happiness pattern
  - most research treats happiness as a continuous variable



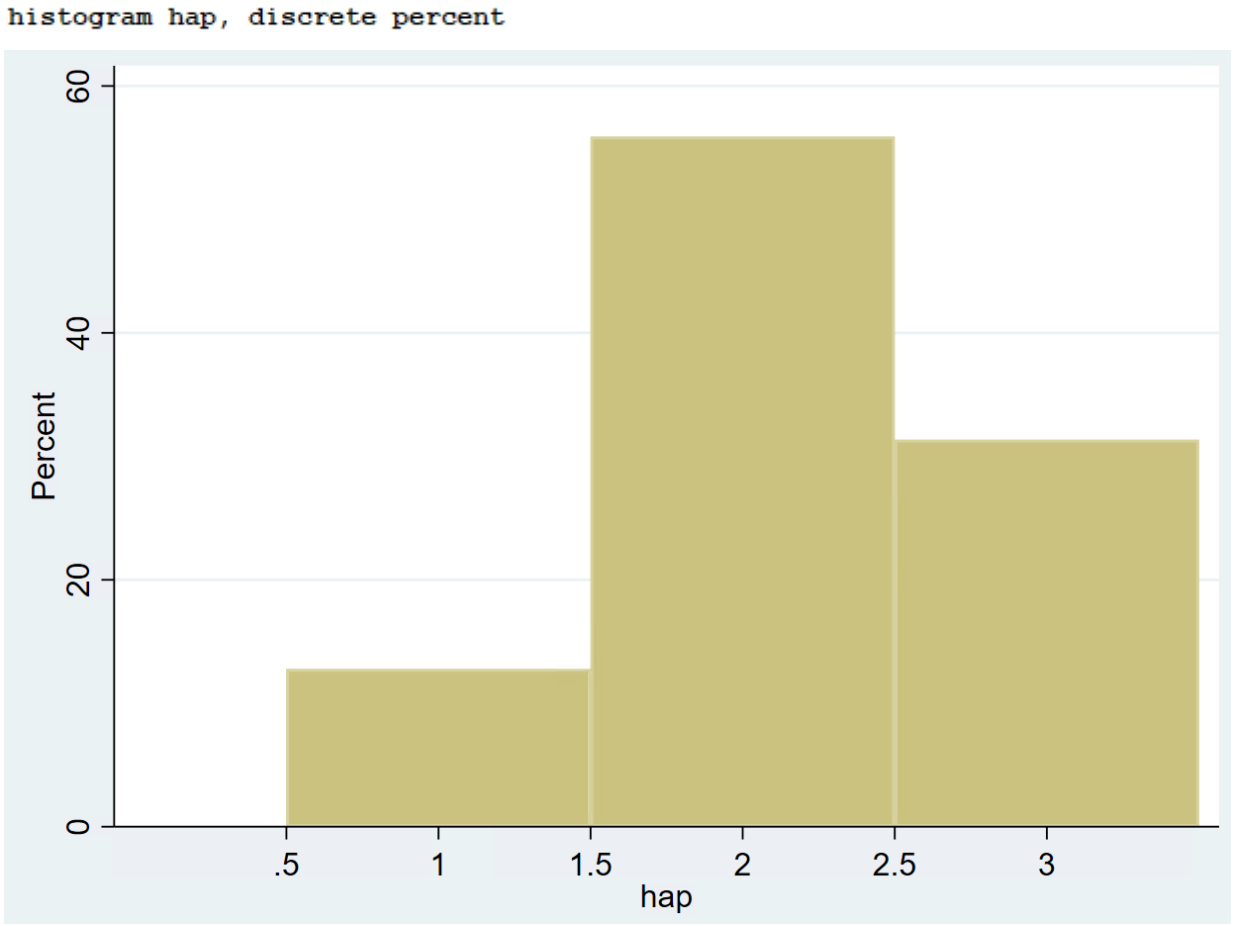
- What's the age pattern in happiness actually look like?
- Do we get more or less happy with age?
- Let's look at the raw data first

Graphing data

- Here's a histogram of overall happiness, but we're interested in age

- In, general how would you say things are these days…
(1) very happy,
(2) pretty happy,
(3) not so happy

reverse coded to make positive

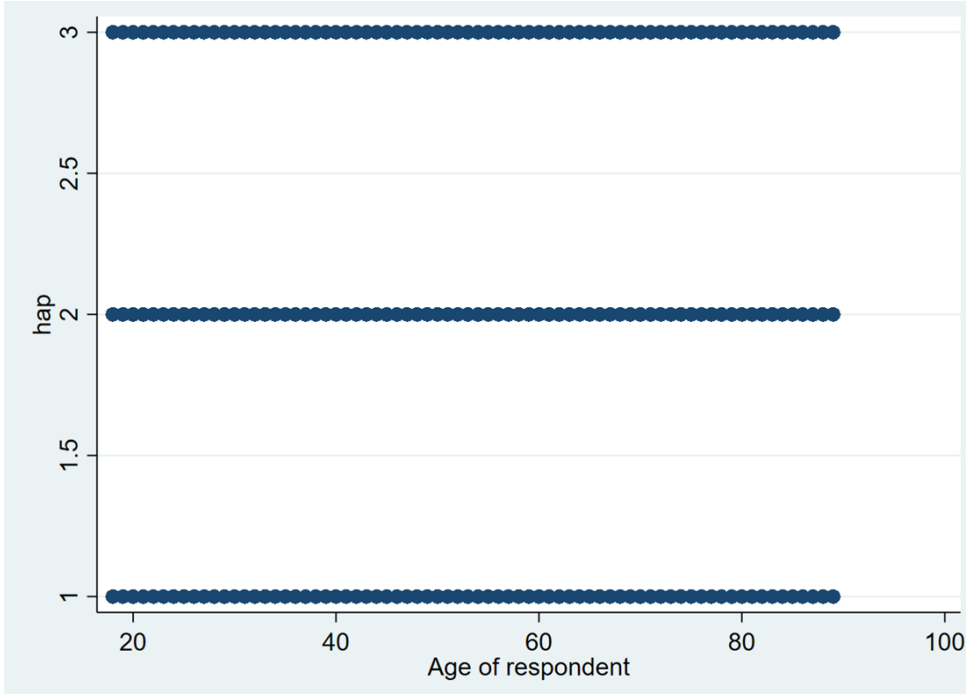


- One of these for each year of age between 18 and 89+…um, no

Graphing data

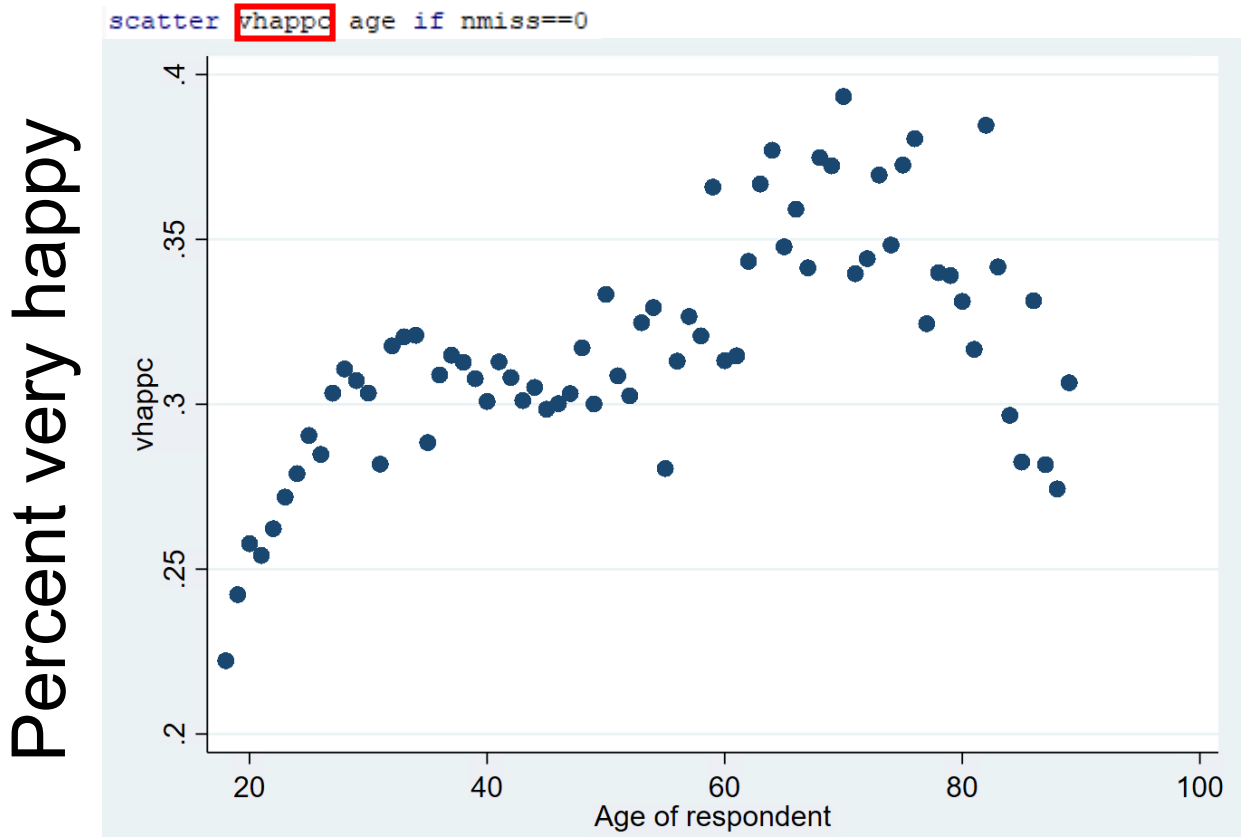
- Scatter plots can be useful
- Is this helpful?
  - Why not?
  - How can we show happiness across age?

```
scatter hap age if nmiss==0
```



Graphing data

- We can treat happiness like the ordinal measure it is, and
- show percentages for each category by age
  - What are we missing?



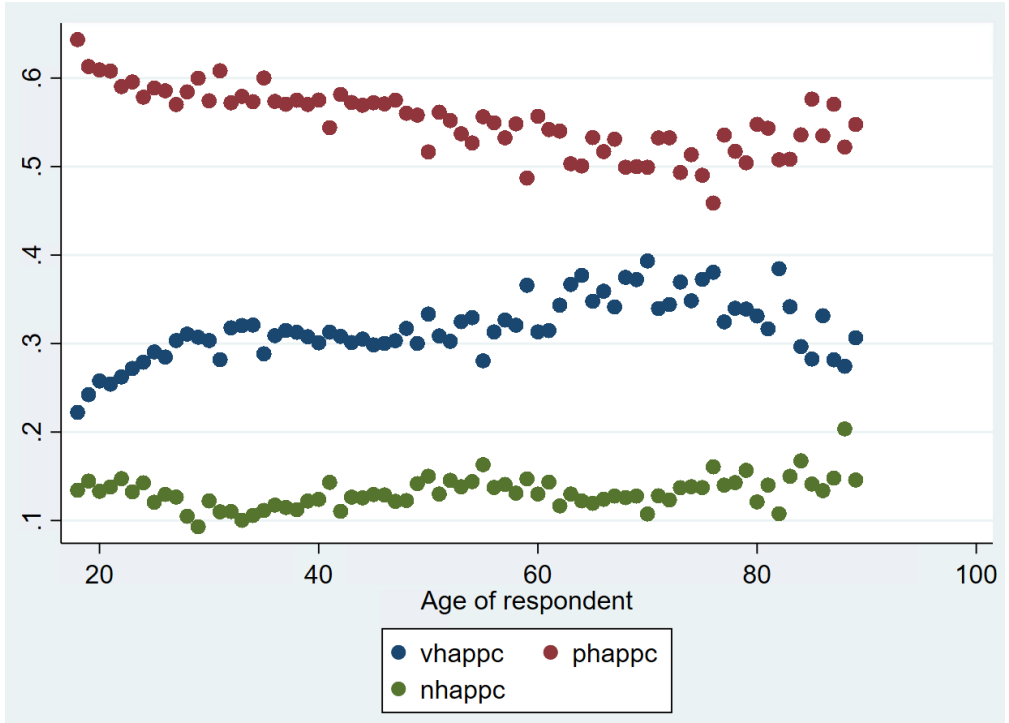
Graphing data

- What's the age pattern in happiness?
  - Not easy to interpret, even with only 3 categories
- Note how the scale can influence interpretation
  - compare blue dots to previous graph

```
/*Rather than mean, consider percent in each response category*/
/*Create indicators for each of the 3 happiness response categories*/
gen vhap=1 if hap==3 & hap!=. /*very happy*/
replace vhap=0 if hap!=3 & hap!=.
gen phap=1 if hap==2 & hap!=. /*pretty happy*/
replace phap=0 if hap!=2 & hap!=.
gen nhap=1 if hap==1 & hap!=. /*not too happy*/
replace nhap=0 if hap!=1 & hap!=.

gen vhappc=.
gen phappc=.
gen nhappc=.
quietly foreach i of numlist 18/89{
egen vhappc`i'=mean(vhap) if age==`i'
replace vhappc=vhappc`i' if vhappc`i'!=.
egen phappc`i'=mean(phap) if age==`i'
replace phappc=phappc`i' if phappc`i'!=.
egen nhappc`i'=mean(nhap) if age==`i'
replace nhappc=nhappc`i' if nhappc`i'!=.
}
```

```
scatter vhappc phappc nhappc age if nmiss==0
```



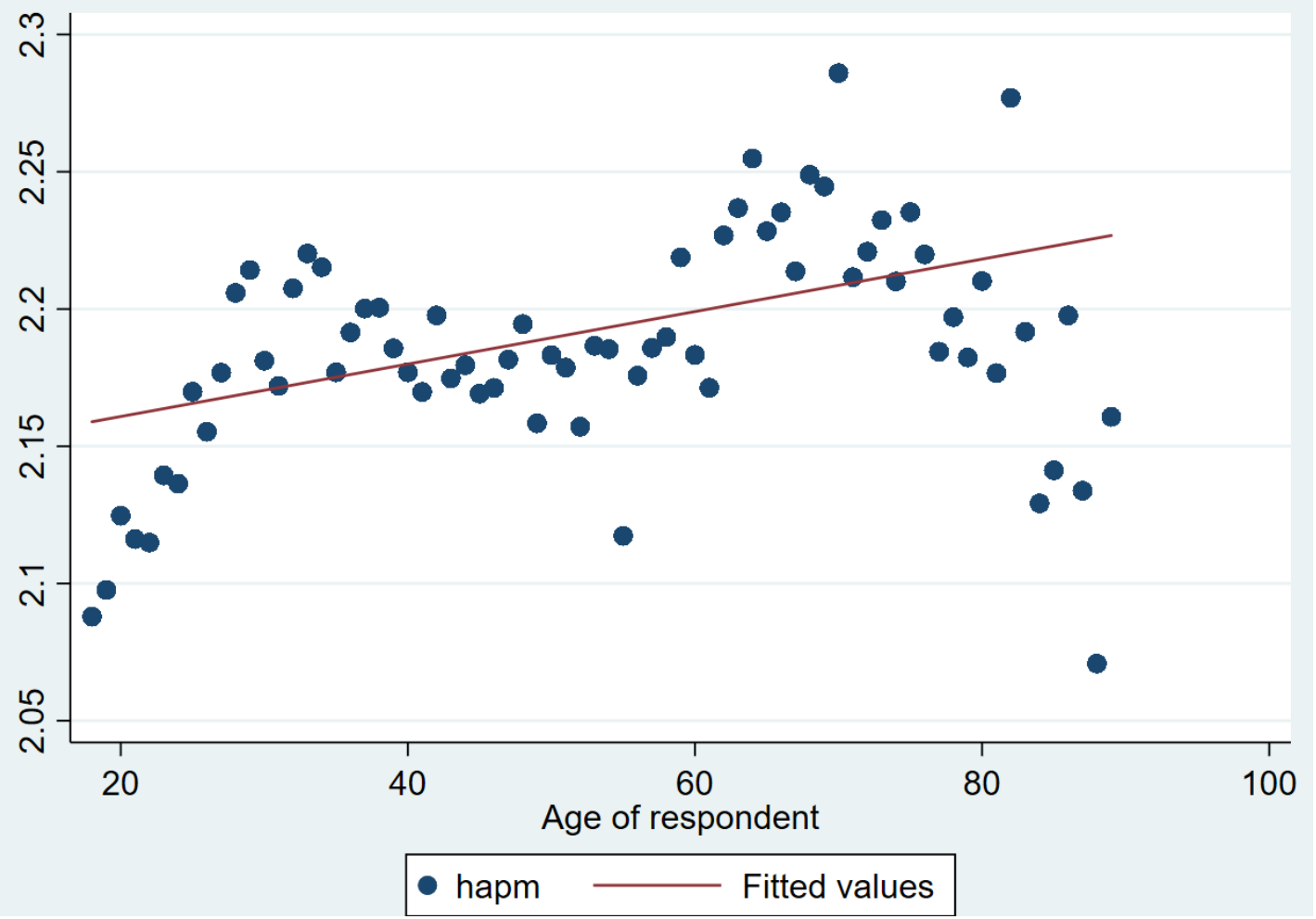
Mean happiness by age

- Often ordinal measures are treated as interval/ratio
  - make compromises for sake of interpretation

```
*compute mean for each year of age, respectively
gen hapm=.
foreach i of numlist 18/89{
egen hapm`i'=mean(hap) if age==`i'
replace hapm=hapm`i' if hapm`i'!=.
}

graph twoway (scatter hapm age) (lfit hapm age)
```

- As you are starting to see, and will continue to see, summarizing and interpreting non-continuous variables can be complex

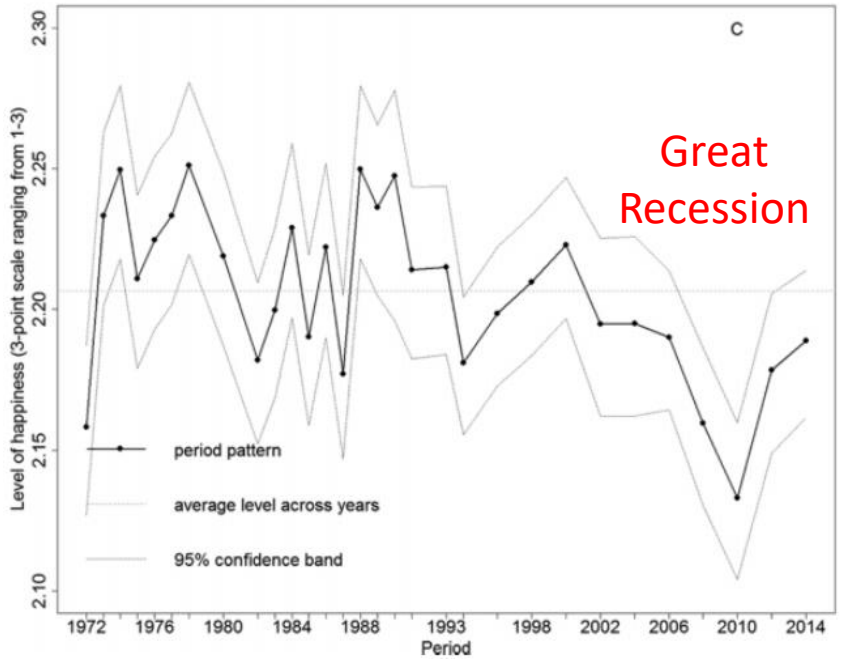


So, what does the age pattern look like?

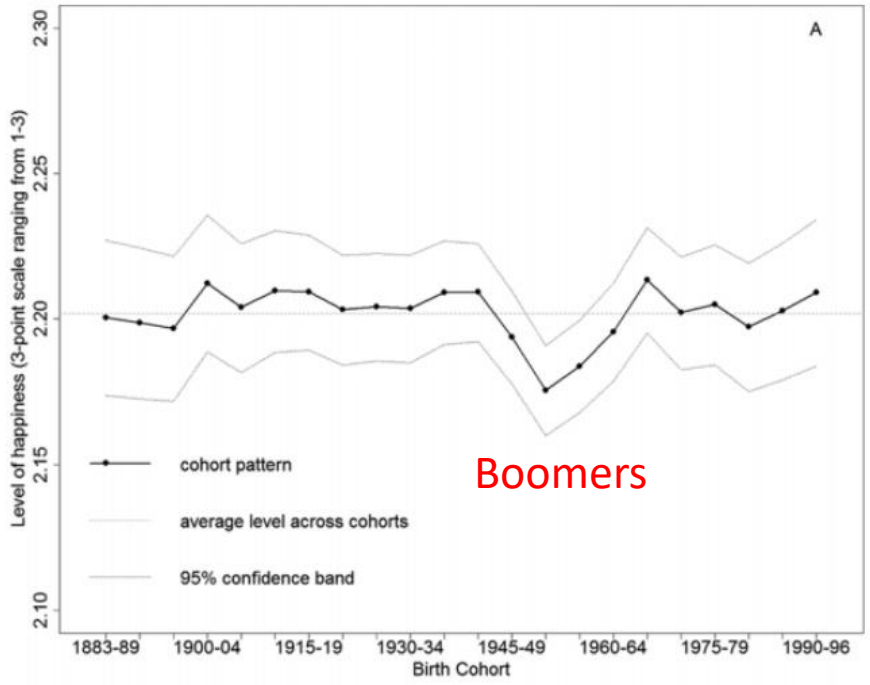
- It's not simply the underlying pattern in happiness by age
  - We need THEORY and inferential statistics!

- Given that these are cross-section data that span 1972-2018, and theory
- Need to account for fluctuations at certain points in time
  - e.g., Great Recession
- Need to account for shared experiences among birth cohorts
  - Boomers least happy cohort
- Need statistical technique to capture change in happiness (DV) given change in age (IV) controlling for period and cohort (Xs)

Hierarchical Cross-Classified Random Effects



- Identification issue
- A + C = P
- Need special methods to break this



- Don't worry about what a HCCRE model is

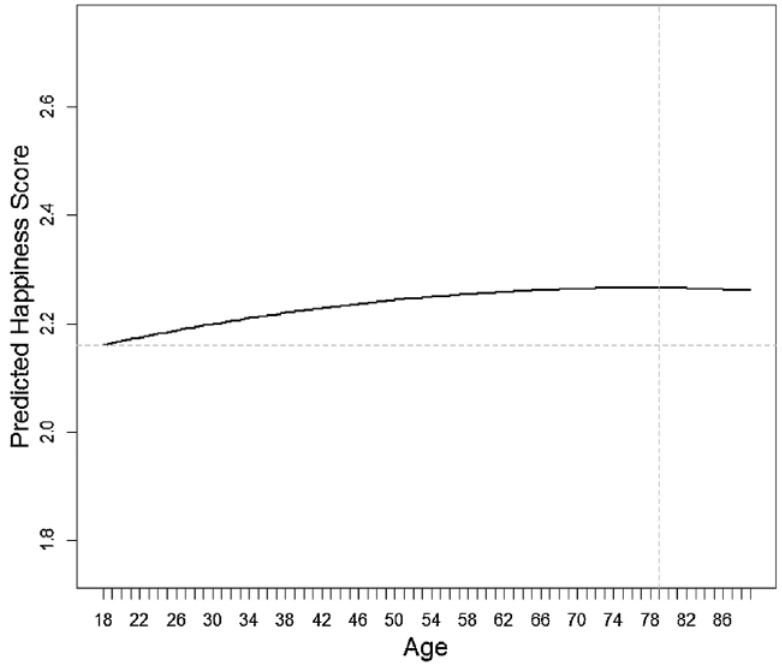


Fig. 3. Age pattern in happiness net of period and cohort.

- Does this age pattern look like the raw data?
- Why not?

Is the pattern consistent across time?

- Previous age pattern reflects average
  - across all time periods (1972-2014) and cohorts

More Happiness for Young People and Less for Mature Adults: Time Period Differences in Subjective Well-Being in the United States, 1972–2014

Jean M. Twenge, Ryne A. Sherman, Sonja Lyubomirsky

First Published November 5, 2015 | Research Article | Check for updates
https://doi.org/10.1177/1948550615602933

Article information

Altmetric 404

Abstract

Are Americans happier, or less happy, than they used to be? The answer may depend on life stage. We examined indicators of subjective well-being (SWB) in four nationally representative samples of U.S. adolescents (aged 13–18 years, n = 1.27 million) and adults (aged 18–96 years, n = 54,172). Recent adolescents reported greater happiness and life satisfaction than their predecessors, and adults over age 30 were less happy in recent years. Among adults, the previously established positive correlation between age and happiness has dwindled, disappearing by the early 2010s. Mixed-effects analyses primarily demonstrated time period rather than generational effects. The effect of time period on SWB is about d = |.13| in most age groups, about the size of reported links between SWB and objective health, marital status, being a parent, and volunteering.

- They've got a theory and statistics to back it up
- Does that mean it's true?

Social Psychological and Personality Science
1-11

DOI: 10.1177/1948550615602933
spps.sagepub.com
SAGE

More Happiness for Young People and Less for Mature Adults: Time Period Differences in Subjective Well-Being in the United States, 1972–2014

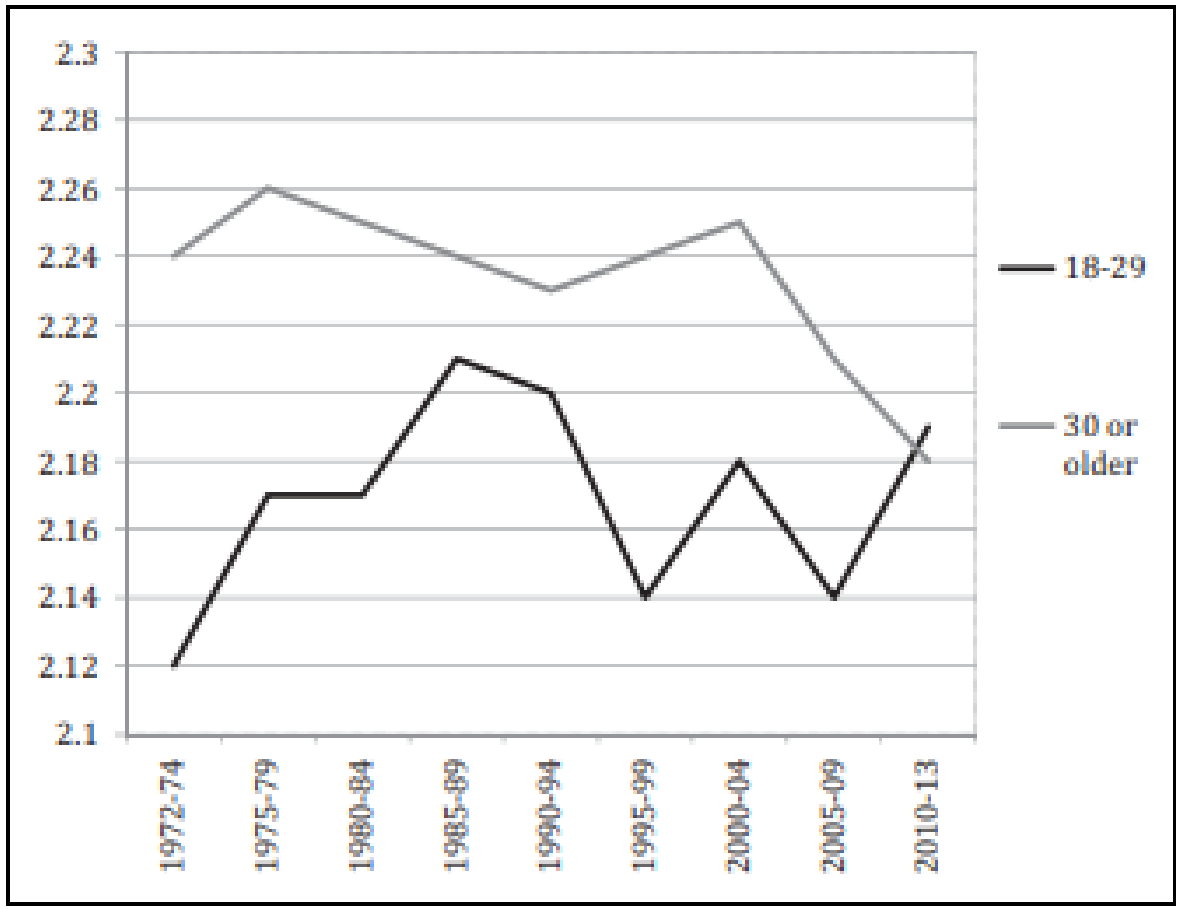


Figure 2. Happiness among 18- to 29-year-old adults and adults 30 or over by time period, U.S. General Social Survey.

- Look at the years. What do we know about the GSS?
- What did they do?
- What happened during the steep decline in 30-plus happiness?
- When did Boomers start making up majority of 30-plus?

Social Psychological and Personality Science
2017, Vol. 8(3) 341-350

DOI: 10.1177/1948550616673874
journals.sagepub.com/home/spp
SAGE

The Importance of the Baby Boom Cohort and the Great Recession in Understanding Age, Period, and Cohort Patterns in Happiness

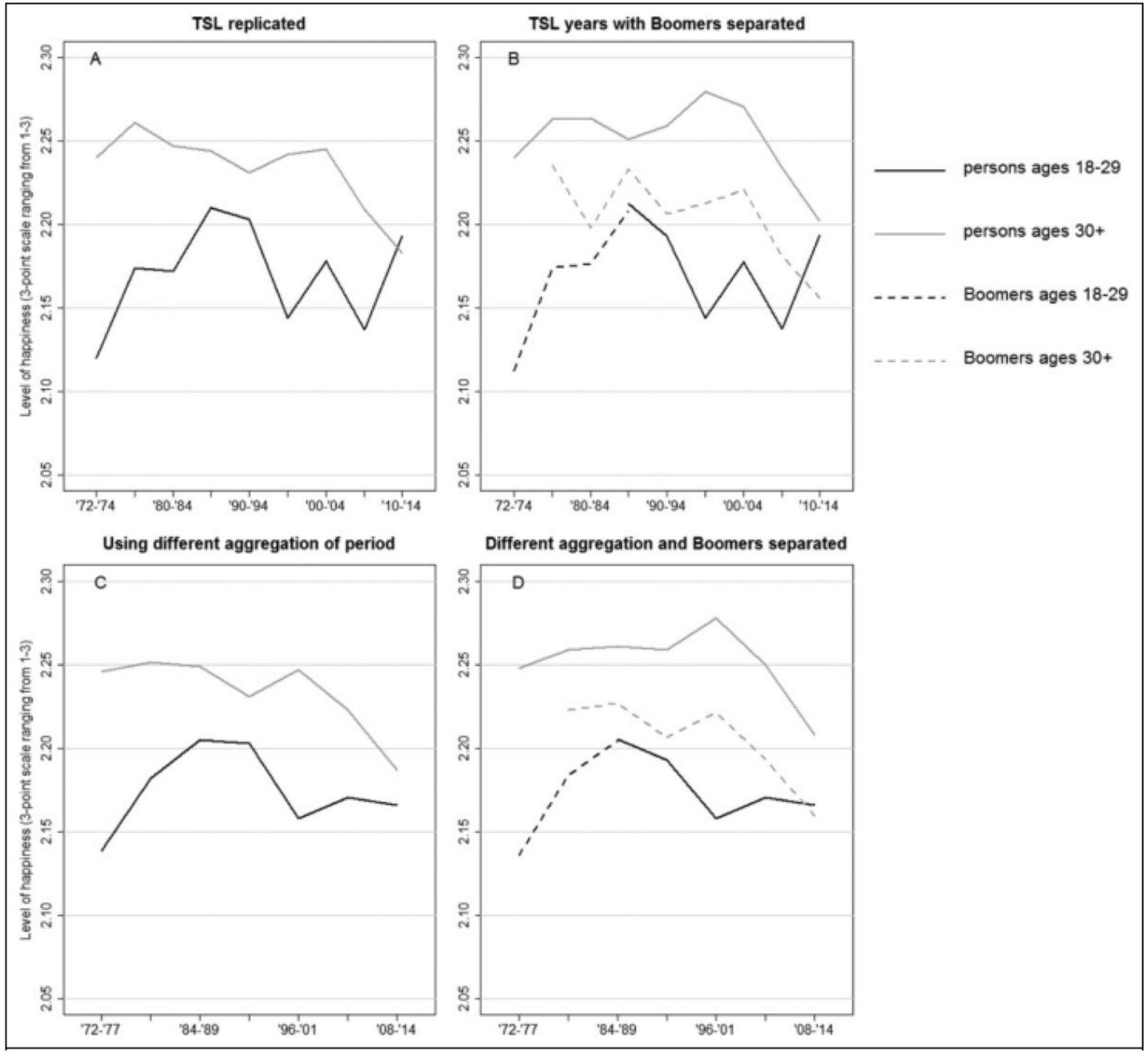


Figure 1. Average levels of happiness among 18- to 29-year-old adults and adults aged 30 and over by time period. The 1982 and 1987 Black oversamples were removed, and the WTSSALL weight variable was used; Boomers include those born between 1945 and 1964, which includes an additional year (i.e., 1945) of birth than normally recognized by demographers, but this year range was selected to match TSL's 5-year birth cohorts used in mixed regression analyses; see Table 1 for 95% confidence bounds.

Social Psychological and Personality Science
1-11

DOI: 10.1177/1948550615602933
spps.sagepub.com
SAGE

More Happiness for Young People and Less for Mature Adults: Time Period Differences in Subjective Well-Being in the United States, 1972–2014

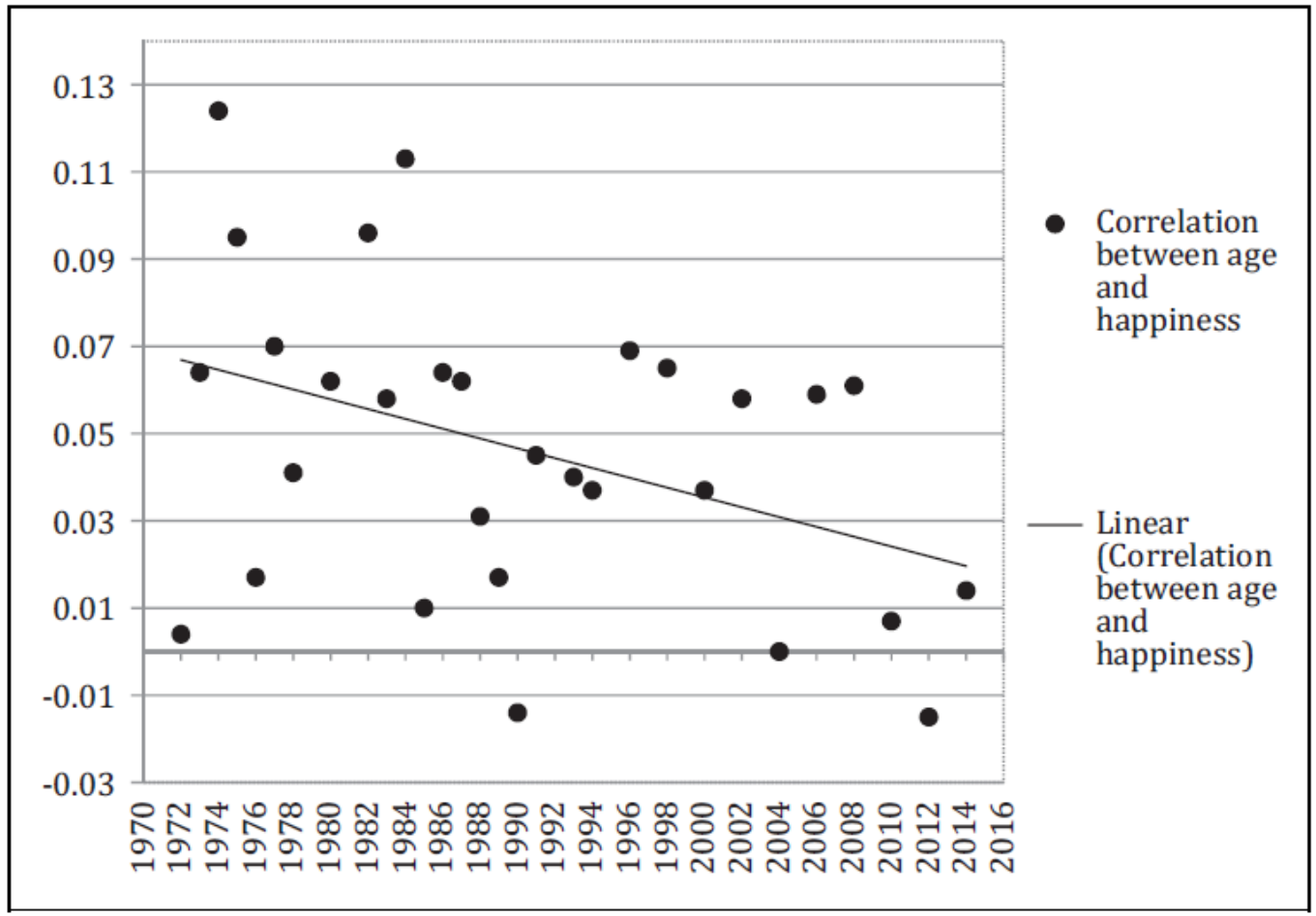


Figure 3. Year-by-year trends in the correlation between age and happiness, scatterplot and linear regression line, General Social Survey of U.S. adults, 1972–2014.

- At least years are separated, but that doesn't help the cohort (Boomer) issue
- Consider bivariate statistical test
  - Pearson's R
- What assumptions violated?

Social Psychological and
Personality Science
2017, Vol. 8(3) 341-350

DOI: 10.1177/1948550616673874
journals.sagepub.com/home/spp
SAGE

The Importance of the Baby Boom Cohort and the Great Recession in Understanding Age, Period, and Cohort Patterns in Happiness

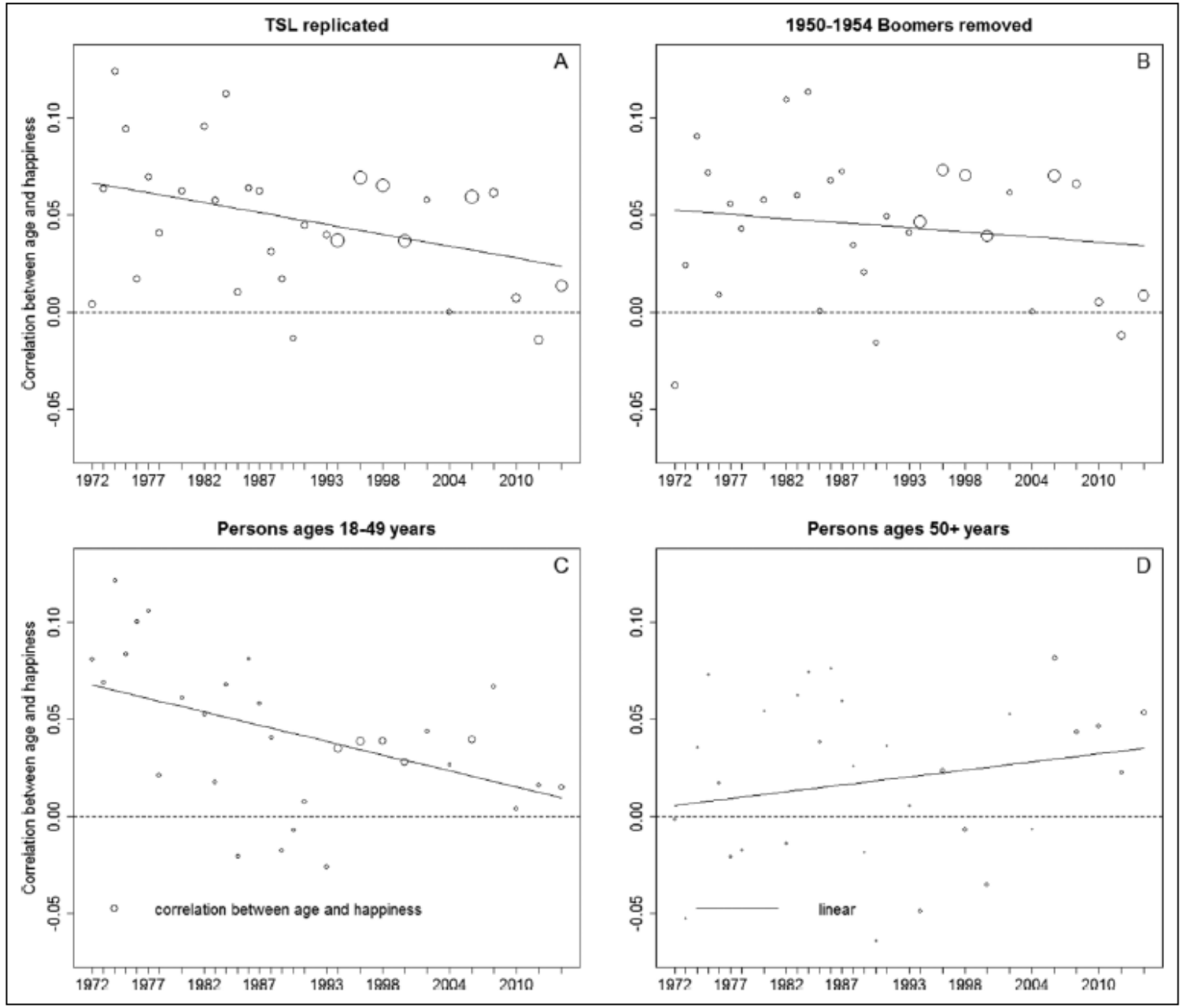
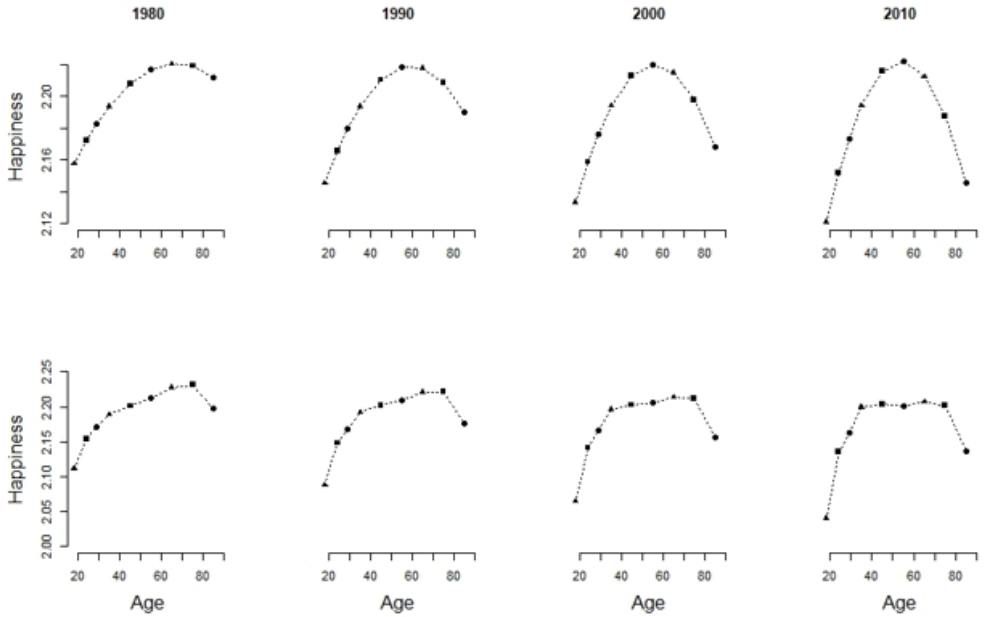


Figure 3. Trends in the correlation between age and happiness across years: scatterplots and linear regression lines. The 1982 and 1987 Black oversamples were removed, and the WTSSALL weight variable was used; circle size reflects relative sample size.

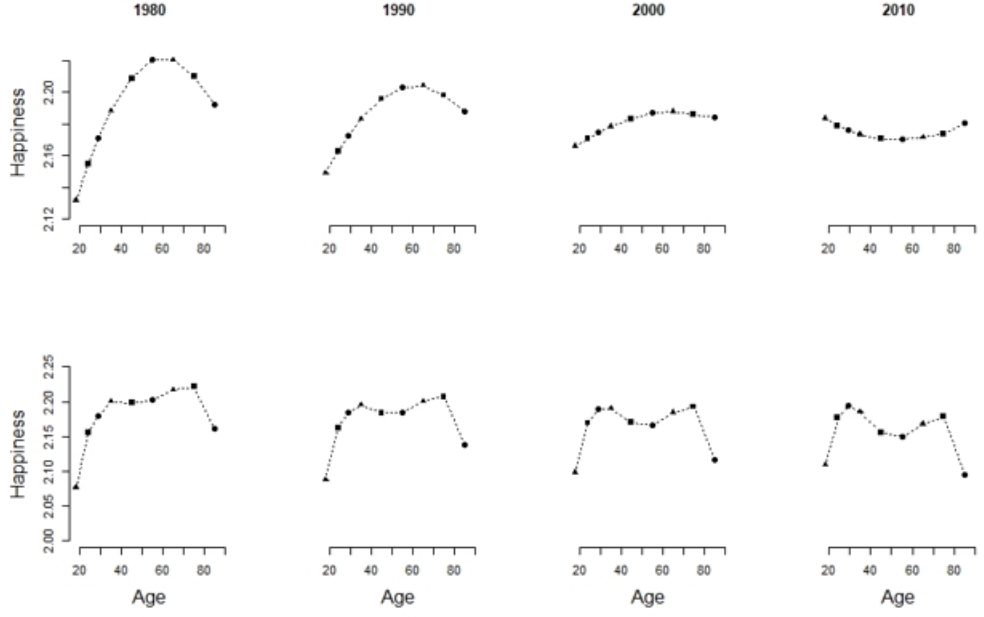
So, has the age pattern in happiness shifted over time?

Figure 4. Trend in the age pattern in happiness: Dummy variable model with Boomers excluded from the sample



Note: the top row depicts results from models that used a quadratic age term; the bottom row depicts results from models that used a fourth order polynomial age term

Figure 5. Trend in the age pattern in happiness: Dummy variable model with Boomers included in the sample



Note: the top row depicts results from models that used a quadratic age term; the bottom row depicts results from models that used a fourth order polynomial age term

Techniques used to obtain these results

- Polynomial transformations of x
  - in this case, age
- Predicted values of y at specific age in each year, respectively
  - in this case, happiness
- Let’s practice interpreting output
  - NOT same approach as previous slide – not considering change over time
- This can get complex
  - graphics help

Higher order terms capture complex features

```
. reg hap age
```

Source	SS	df	MS	Number of obs	=	59,860
				F(1, 59858)	=	42.06
Model	17.0766225	1	17.0766225	Prob > F	=	0.0000
Residual	24304.5682	59,858	.406037091	R-squared	=	0.0007
				Adj R-squared	=	0.0007
Total	24321.6448	59,859	.406315589	Root MSE	=	.63721

hap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0009603	.0001481	6.49	0.000	.0006701	.0012506
_cons	2.14151	.0073021	293.27	0.000	2.127198	2.155822

```
. reg hap age age2
```

Source	SS	df	MS	Number of obs	=	59,860
				F(2, 59857)	=	27.34
Model	22.2014984	2	11.1007492	Prob > F	=	0.0000
Residual	24299.4433	59,857	.405958256	R-squared	=	0.0009
				Adj R-squared	=	0.0009
Total	24321.6448	59,859	.406315589	Root MSE	=	.63715

hap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0038745	.0008335	4.65	0.000	.002241	.0055081
age2	-.0000293	8.25e-06	-3.55	0.000	-.0000455	-.0000131
_cons	2.078526	.0191716	108.42	0.000	2.040949	2.116102

```
. reg hap age age2 age3
```

Source	SS	df	MS	Number of obs	=	59,860
				F(3, 59856)	=	18.33
Model	22.3271692	3	7.44238974	Prob > F	=	0.0000
Residual	24299.3177	59,856	.405962939	R-squared	=	0.0009
				Adj R-squared	=	0.0009
Total	24321.6448	59,859	.406315589	Root MSE	=	.63715

hap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0056597	.0033151	1.71	0.088	-.0008378	.0121573
age2	-.0000672	.0000687	-0.98	0.328	-.0002019	.0000674
age3	2.47e-07	4.43e-07	0.56	0.578	-6.22e-07	1.11e-06
_cons	2.053334	.0491695	41.76	0.000	1.956961	2.149706

```
. reg hap age age2 age3 age4
```

Source	SS	df	MS	Number of obs	=	59,860
				F(4, 59855)	=	28.77
Model	46.6771709	4	11.6692927	Prob > F	=	0.0000
Residual	24274.9677	59,855	.405562905	R-squared	=	0.0019
				Adj R-squared	=	0.0019
Total	24321.6448	59,859	.406315589	Root MSE	=	.63684

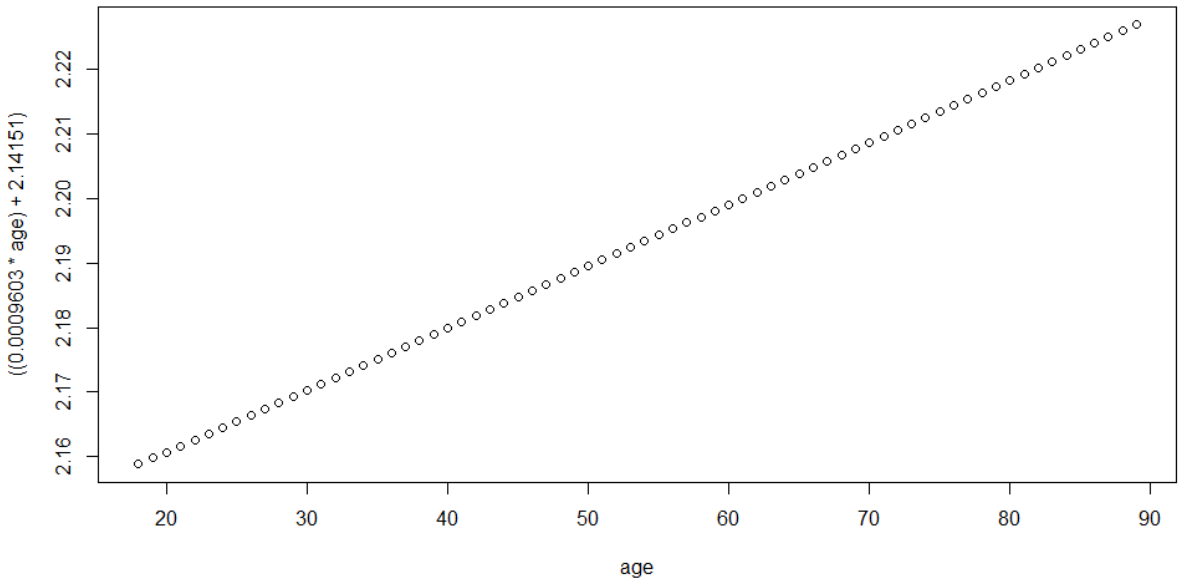
hap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0921653	.0116454	7.91	0.000	.0693402	.1149903
age2	-.0029242	.000375	-7.80	0.000	-.0036593	-.0021891
age3	.0000392	5.04e-06	7.77	0.000	.0000293	.0000491
age4	-1.87e-07	2.41e-08	-7.75	0.000	-2.34e-07	-1.39e-07
_cons	1.150145	.1264991	9.09	0.000	.9022063	1.398084

```
*age
replace age=. if age>89
tab age,m
*age quadratic
gen age2=age * age
*age cubic
gen age3=age * age * age
*age quartic
gen age4=age * age * age * age
```

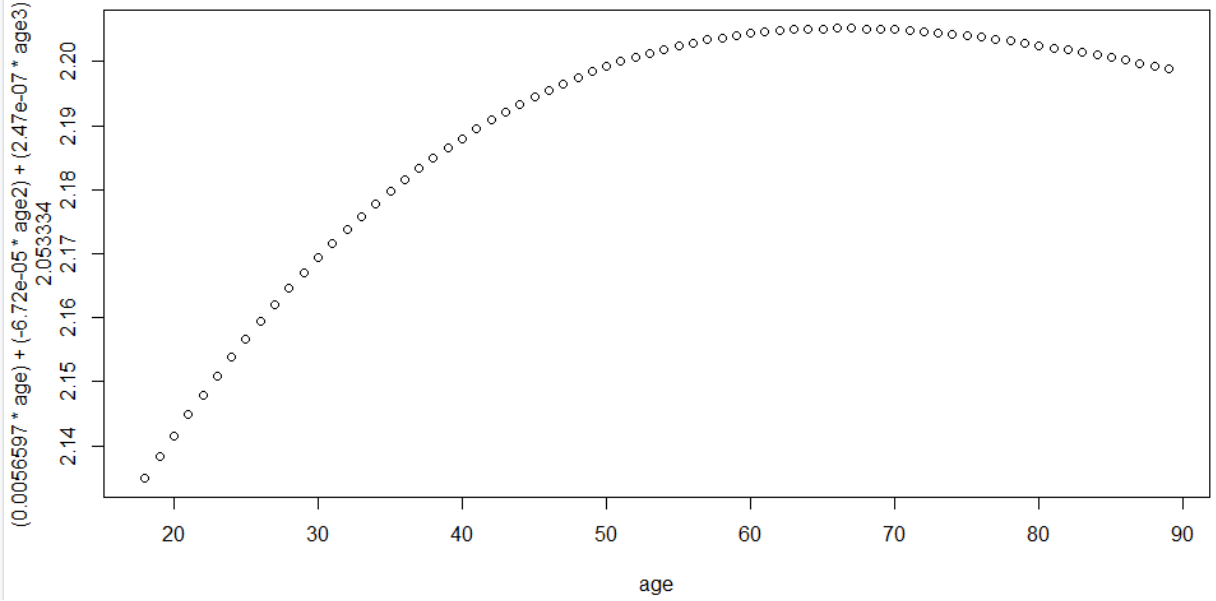
- Second order (quadratic)
  - one inflection point
- Third order (cubic)
  - two inflection points
- Fourth order (quartic)
  - three inflection points

- Interpret these results

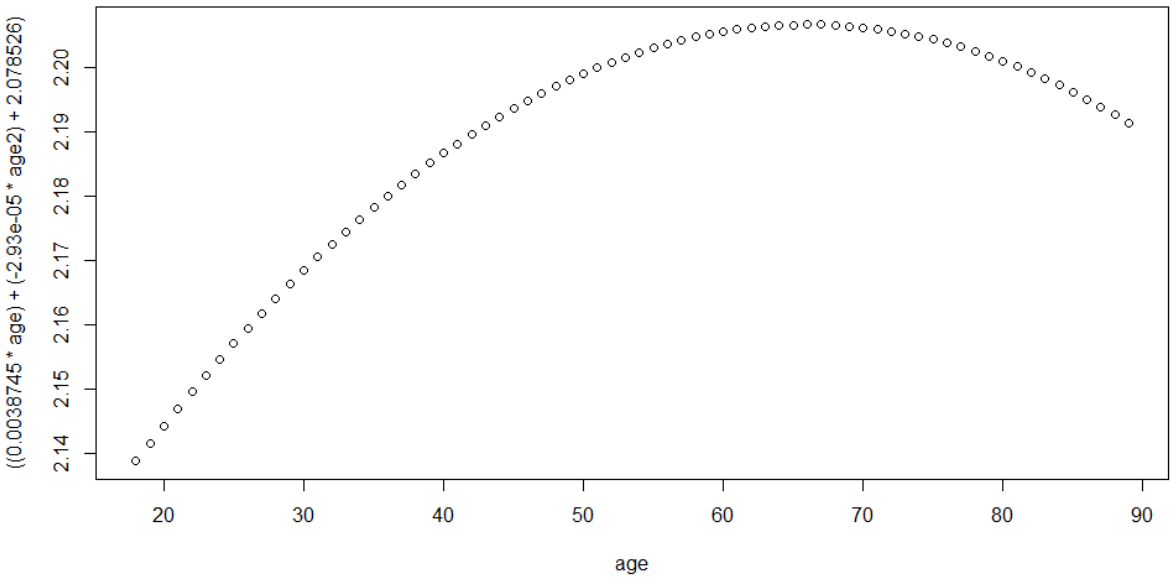
1st order



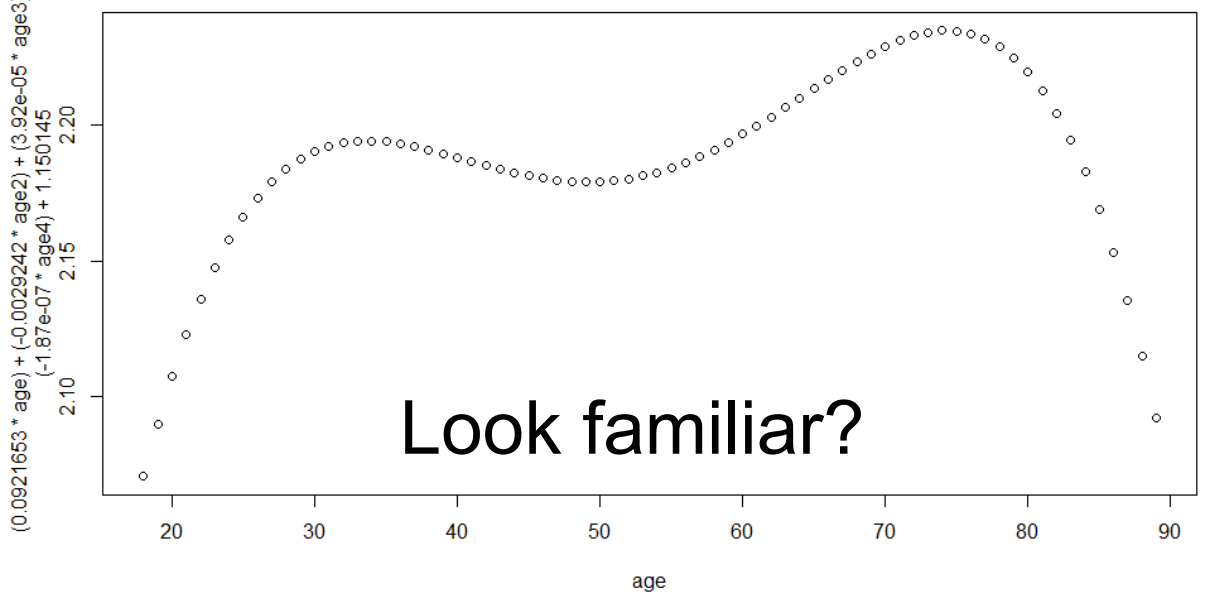
3rd order

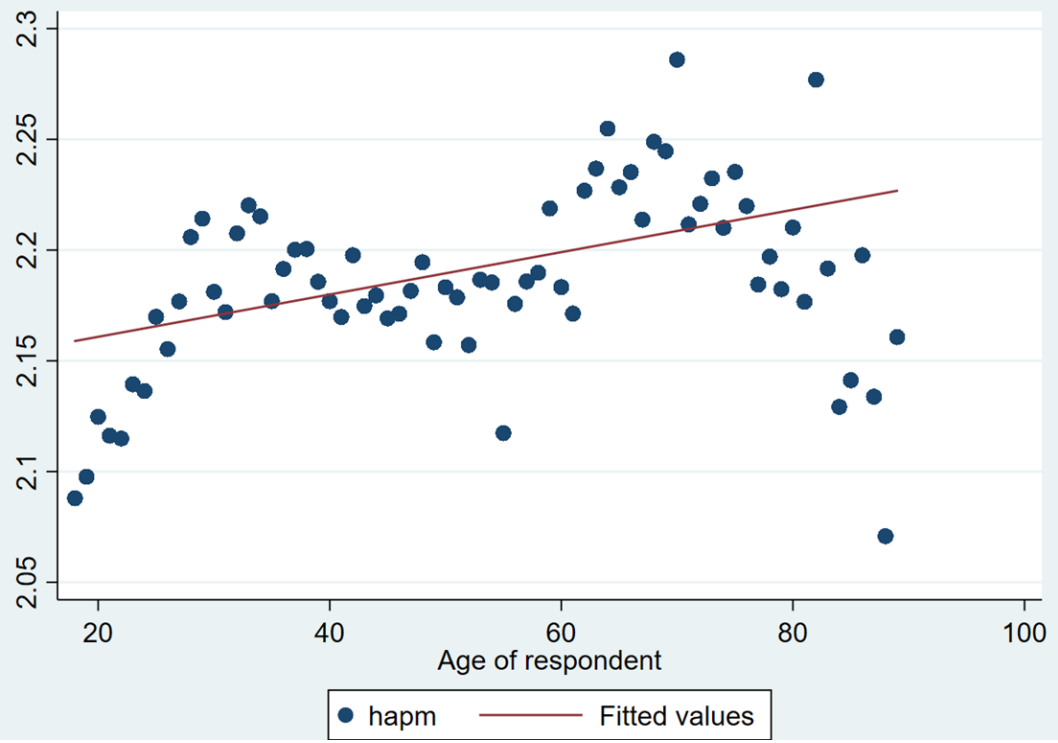


2nd order

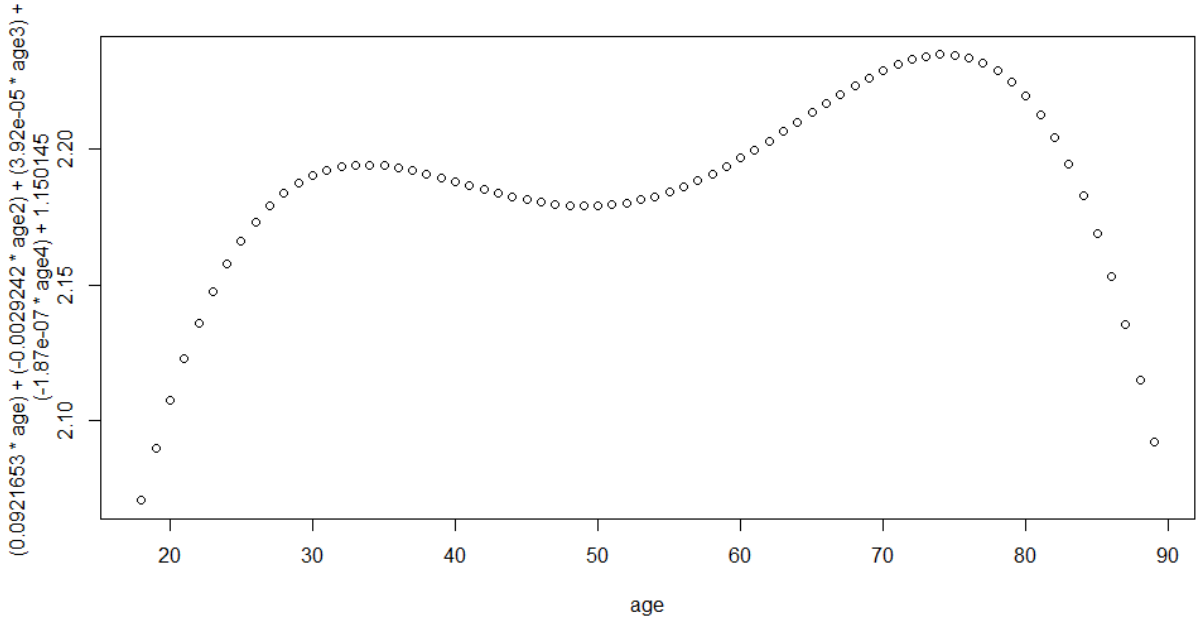


4th order





4th order



- Where did these numbers come from, and how did we use them?

Coefficients and constant

```
. reg hap age
```

Source	SS	df	MS	Number of obs	=	59,860
				F(1, 59858)	=	42.06
Model	17.0766225	1	17.0766225	Prob > F	=	0.0000
Residual	24304.5682	59,858	.406037091	R-squared	=	0.0007
				Adj R-squared	=	0.0007
Total	24321.6448	59,859	.406315589	Root MSE	=	.63721

hap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0009603	.0001481	6.49	0.000	.0006701	.0012506
_cons	2.14151	.0073021	293.27	0.000	2.127198	2.155822

```
. reg hap age age2
```

Source	SS	df	MS	Number of obs	=	59,860
				F(2, 59857)	=	27.34
Model	22.2014984	2	11.1007492	Prob > F	=	0.0000
Residual	24299.4433	59,857	.405958256	R-squared	=	0.0009
				Adj R-squared	=	0.0009
Total	24321.6448	59,859	.406315589	Root MSE	=	.63715

hap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0038745	.0008335	4.65	0.000	.002241	.0055081
age2	-.0000293	8.25e-06	-3.55	0.000	-.0000455	-.0000131
_cons	2.078526	.0191716	108.42	0.000	2.040949	2.116102

```
. reg hap age age2 age3
```

Source	SS	df	MS	Number of obs	=	59,860
				F(3, 59856)	=	18.33
Model	22.3271692	3	7.44238974	Prob > F	=	0.0000
Residual	24299.3177	59,856	.405962939	R-squared	=	0.0009
				Adj R-squared	=	0.0009
Total	24321.6448	59,859	.406315589	Root MSE	=	.63715

hap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0056597	.0033151	1.71	0.088	-.0008378	.0121573
age2	-.0000672	.0000687	-0.98	0.328	-.0002019	.0000674
age3	2.47e-07	4.43e-07	0.56	0.578	-6.22e-07	1.11e-06
_cons	2.053334	.0491695	41.76	0.000	1.956961	2.149706

```
. reg hap age age2 age3 age4
```

Source	SS	df	MS	Number of obs	=	59,860
				F(4, 59855)	=	28.77
Model	46.6771709	4	11.6692927	Prob > F	=	0.0000
Residual	24274.9677	59,855	.405562905	R-squared	=	0.0019
				Adj R-squared	=	0.0019
Total	24321.6448	59,859	.406315589	Root MSE	=	.63684

hap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0921653	.0116454	7.91	0.000	.0693402	.1149903
age2	-.0029242	.000375	-7.80	0.000	-.0036593	-.0021891
age3	.0000392	5.04e-06	7.77	0.000	.0000293	.0000491
age4	-1.87e-07	2.41e-08	-7.75	0.000	-2.34e-07	-1.39e-07
_cons	1.150145	.1264991	9.09	0.000	.9022063	1.398084

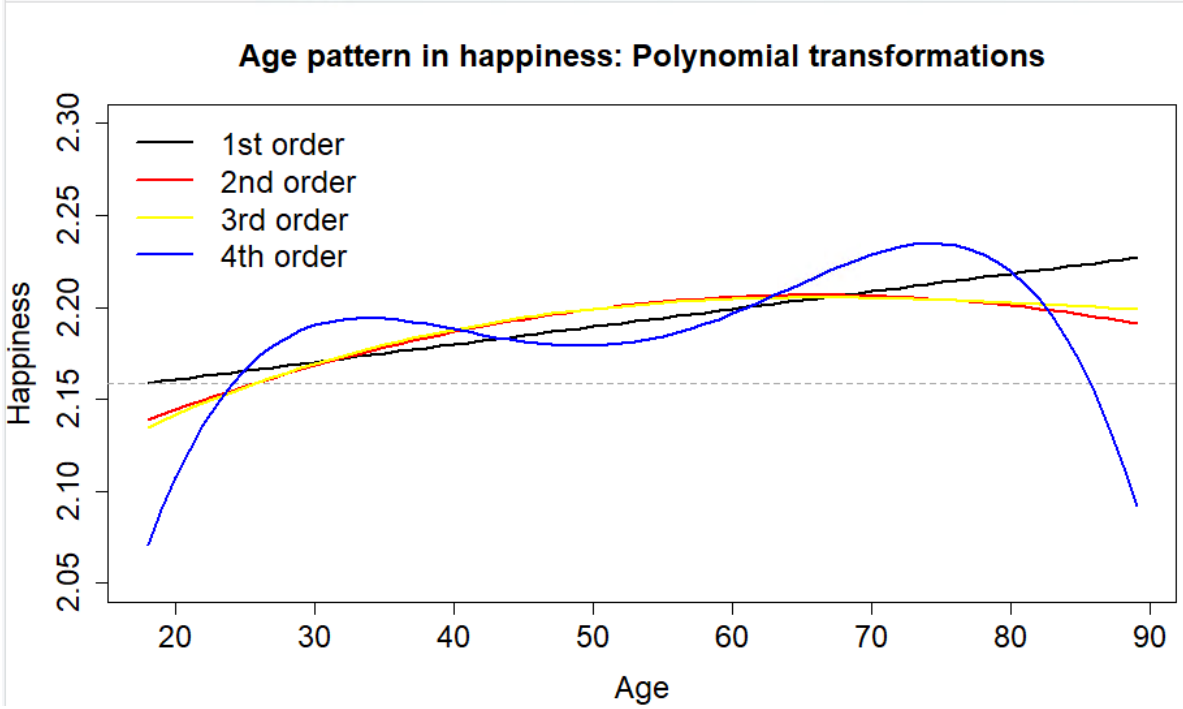
```
*age
replace age=. if age>89
tab age,m
*age quadratic
gen age2=age * age
*age cubic
gen age3=age * age * age
*age quartic
gen age4=age * age * age * age
```

- Coefficients denote effect magnitude
  - corresponds to one unit Δ in x

- Constant = intercept
  - use to plot expected mean value of y
    - for x=0 (reference group)
    - mean center & hypotheticals
      - more later

Let's pretty this up: R plot example

```
plot(age,((0.0009603*age)+2.14151),ylim=c(2.05,2.3),ylab="Happiness",xlab="Age",cex.lab=1.5,
     cex.axis=1.5,type="l",lwd=2,main="Age pattern in happiness: Polynomial transformations",cex.main=1.5)
abline(h=(0.0009603*18)+2.14151,col="dark grey",lty=2)
lines(age,((0.0038745*age)+(-0.0000293*age2)+2.078526),col="red",lwd=2)
lines(age,((0.0056597*age)+(-0.0000672*age2)+(2.47e-07*age3)+2.053334),col="yellow",lwd=2)
lines(age,((0.0921653*age)+(-0.0029242*age2)+(0.0000392*age3)+(-1.87e-07*age4)+1.150145),col="blue",lwd=2)
legend("topleft",c("1st order","2nd order","3rd order","4 order"),lty=c(1,1,1,1),lwd=c(2,2,2,2),
       col=c("black","red","yellow","blue"),bty="n",cex=1.5)
```



That's why we use multiple regression

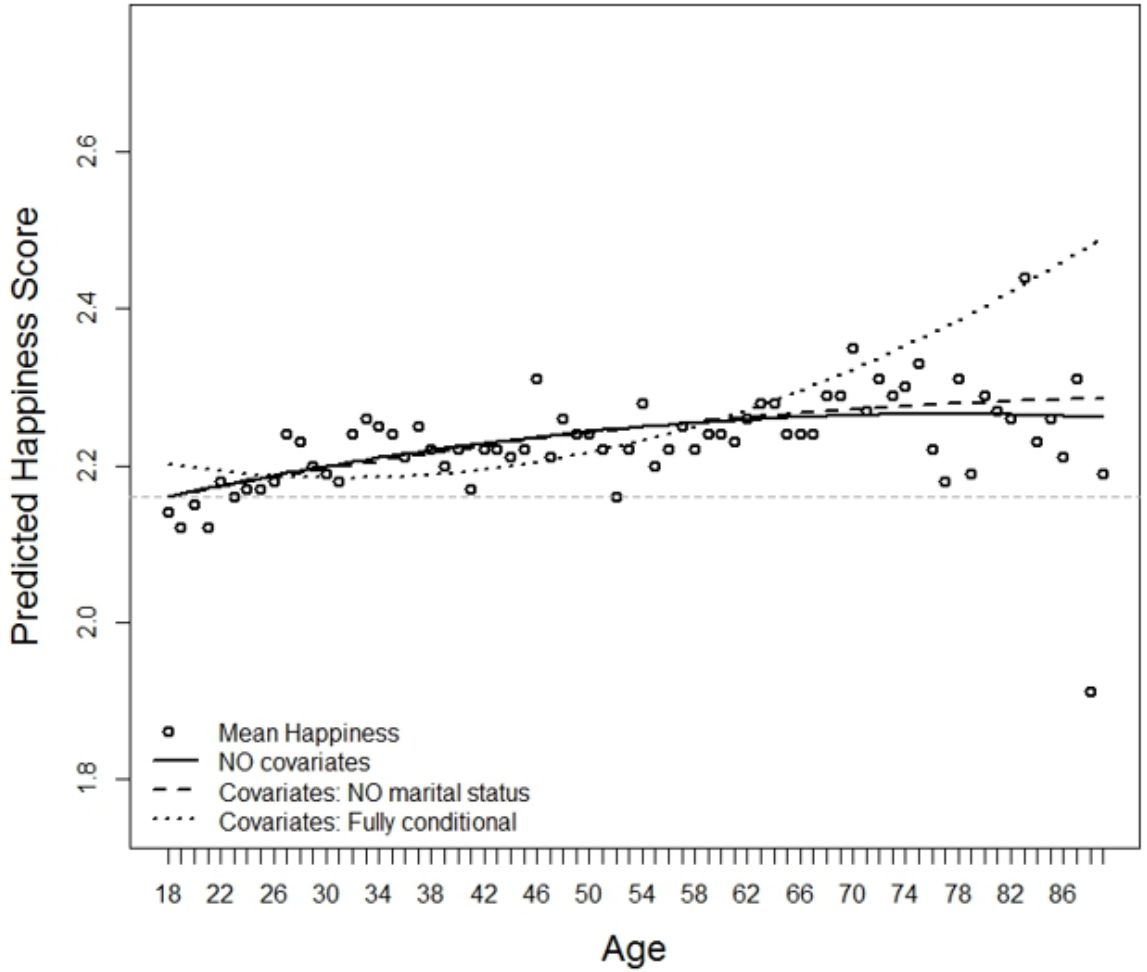
- To estimate the Δ in DV(y) given a Δ in IV(x) net of controls
  - in this case, cohort and period
- How do you decide what controls to include?
  - THEORY!
- Explain as much variance in y (R^2) as possible
  - but only as much as theoretically justified
    - don't over control
- Explain away relationship between DV and IV
  - reject or fail to reject hypotheses

Modeling issues

- Generally good idea to run step-wise models
  - insert controls in theoretically meaningful order
  - and examine Δ in IV and other controls

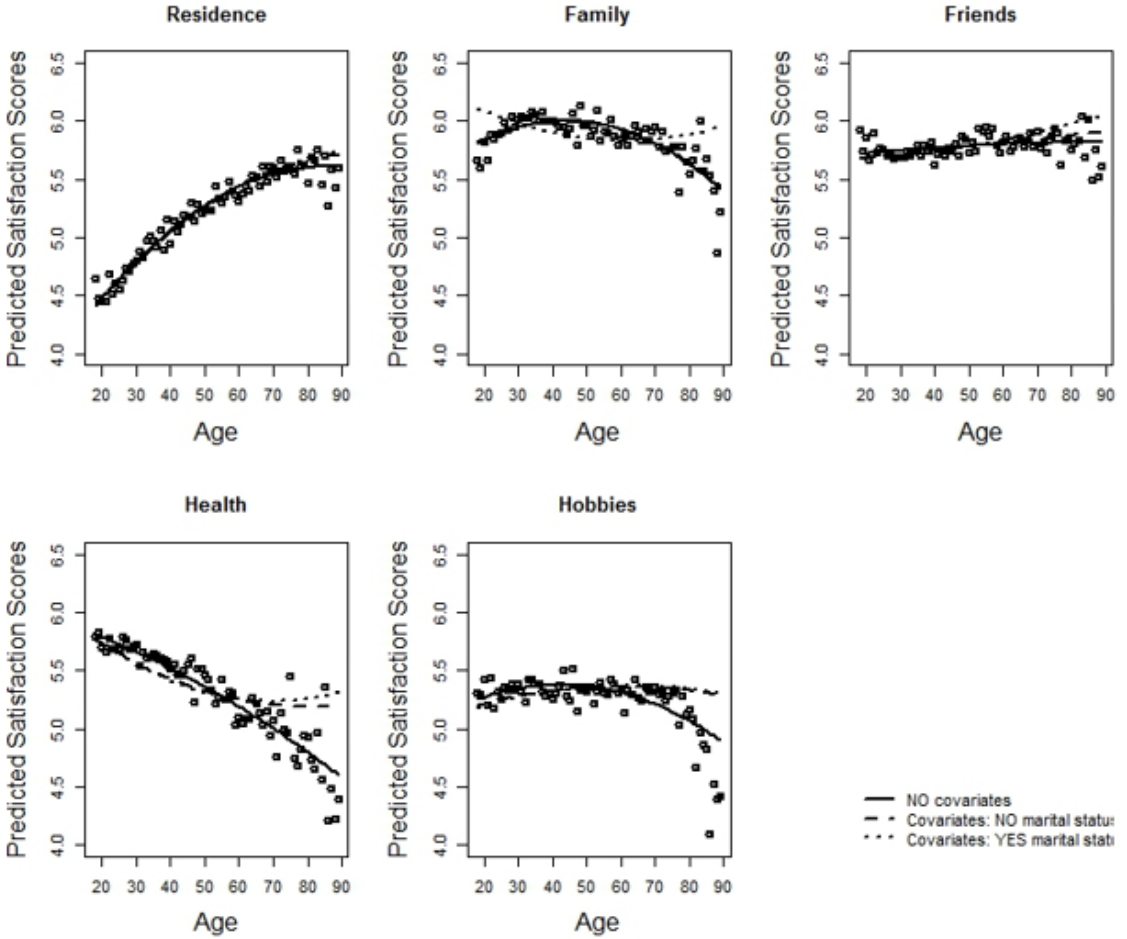
- The inclusion of some covariates can have unintended consequences
  - e.g., linear combinations (e.g., APC: C = P – A)
  - e.g., distinctly patterned combinations (e.g., martial status and age)

Figure 1R. Empirical evidence for how control variables can change main results on age effects



- Results from HCCRM
- But we can play with them using linear regression

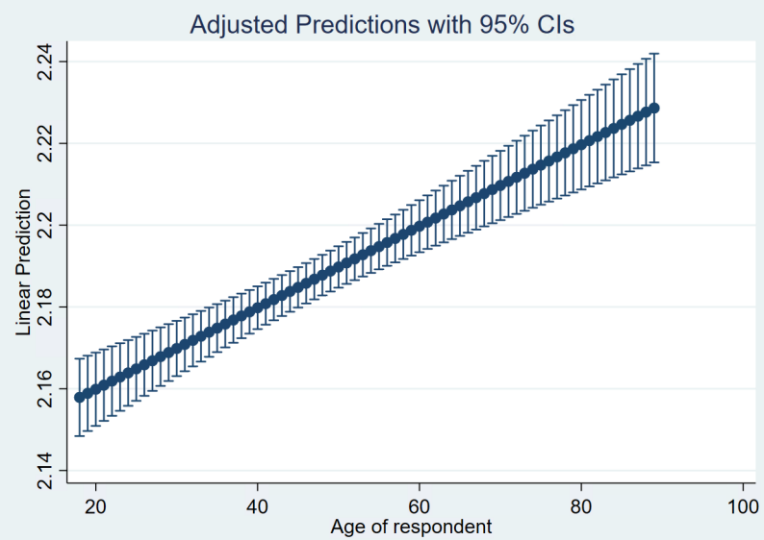
Figure 2R. Empirical evidence for controls continued: Domain-specific satisfaction



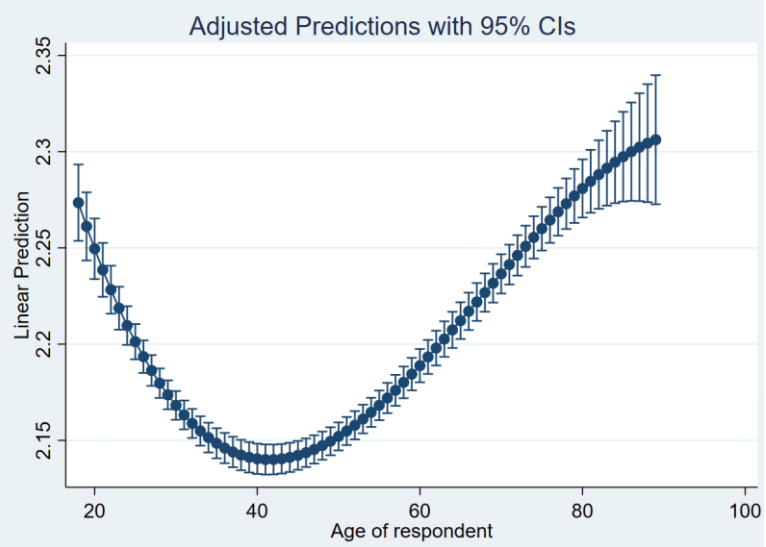
First, the inclusion of marital status in these regressions does not distort the underlying age patterns in domain-specific satisfaction, except for satisfaction with family. In terms of satisfaction with family, when sex, race, education, and labor force status are held constant, the adjusted age pattern maps directly onto the unadjusted age pattern. However, when marital status was entered into the equation, the age pattern in satisfaction with family flipped from concave down to concave up. This finding shows that the U-shaped age pattern in happiness, obtained only when marital status was held constant, is driven by underlying life course patterns in family structures.

Our data: age-pattern with covariates

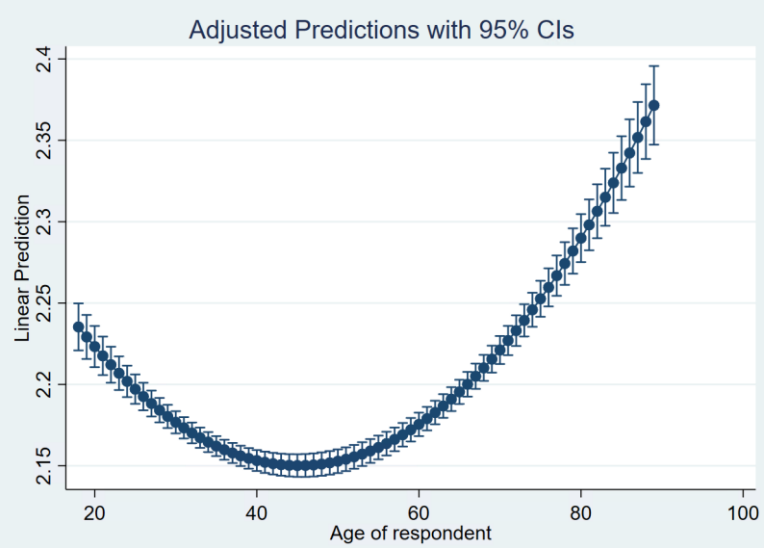
1st order



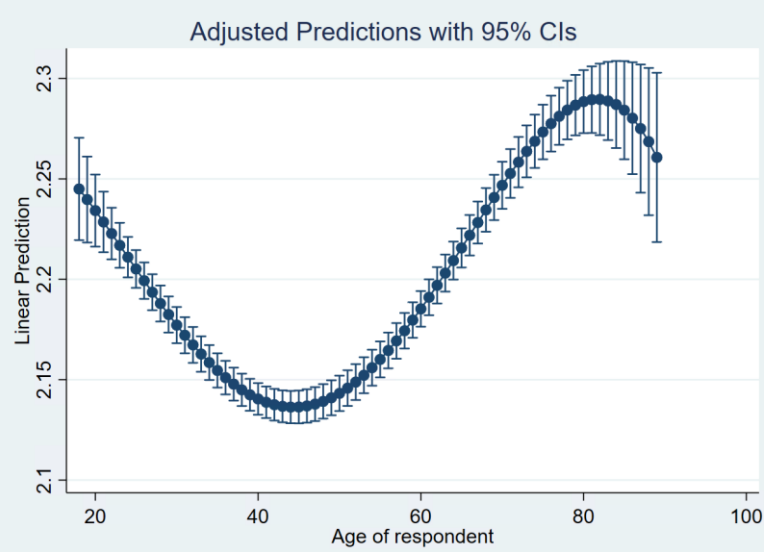
3rd order



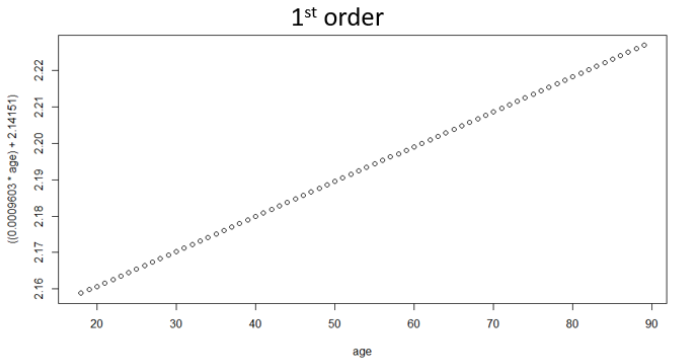
2nd order

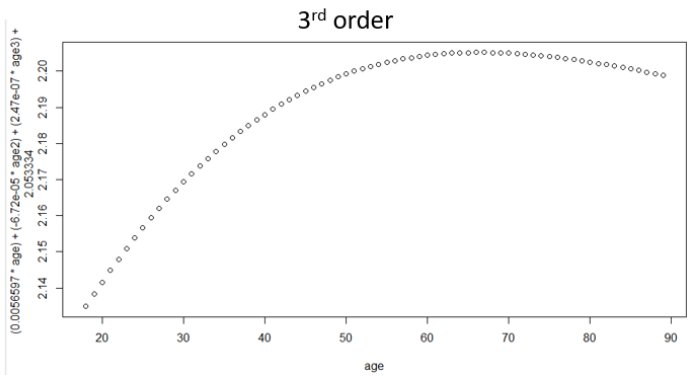


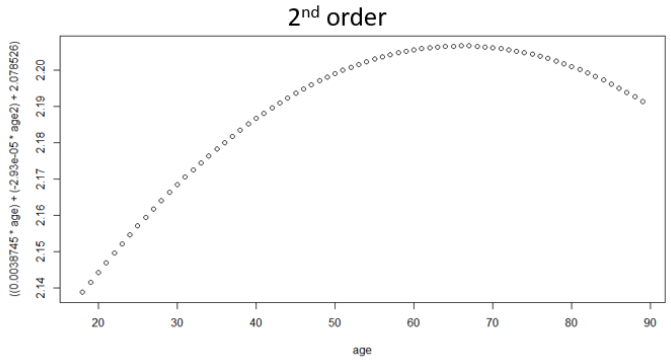
4th order

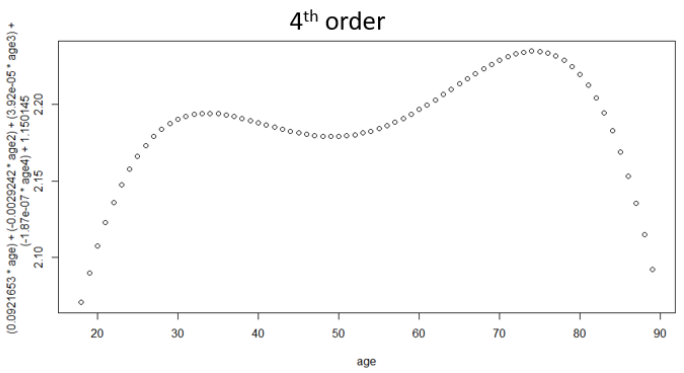


No covariates

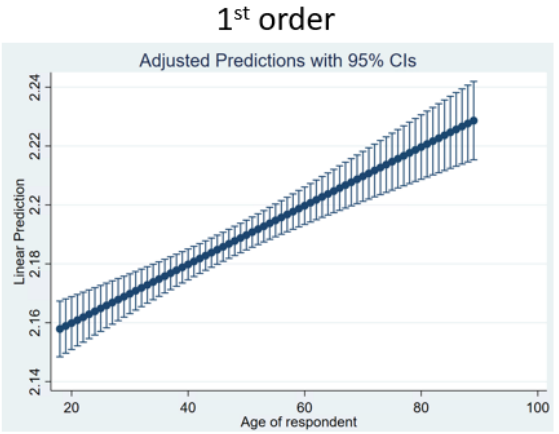


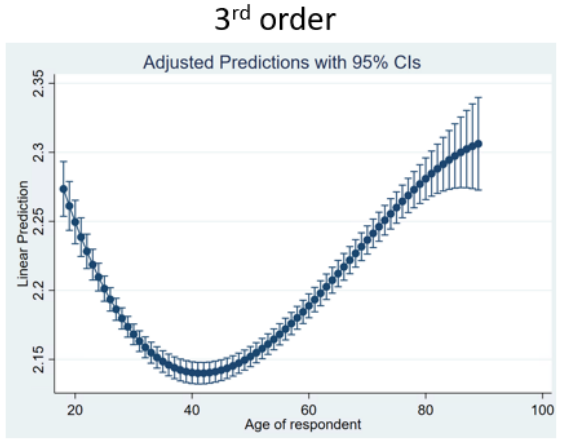


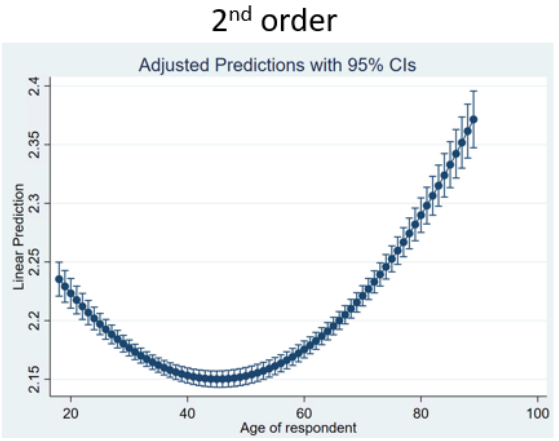


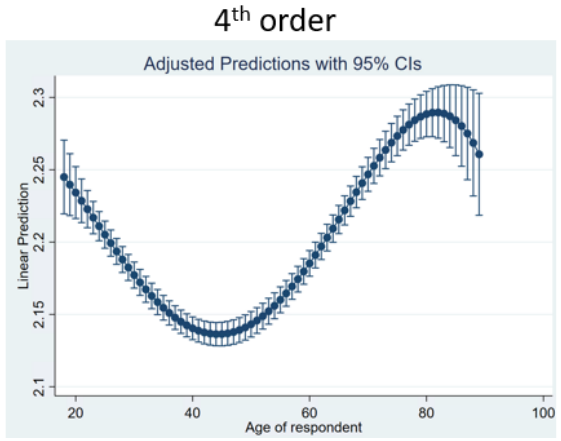


Yes covariates









- What covariate is flipping the age pattern in happiness?
- Play with data using stepwise techniques
  - insert covariates individually

Takeaway

- It’s easy to get output from Stata
  - doesn’t mean its right
- Decisions should be guided by theory
  - first, get dirty with raw data and descriptives
- Know whether you’re violating assumptions
  - more on this later
- Understand the output
  - graphics can help

Next class we will…

- review linear regression
- read Hoffman CH1 before class