Quantitative Data Analysis II

SOC 781

Linear regression

Today we're going to…

- Review linear regression
  - assumptions and diagnostics

Regression

- Whether outcome associated with another (or set of other) variable(s)
  - simple vs. (multiple regression)

- How outcome associated with other variables
  - strength and direction
  - net of other variables (holding all else constant)

- Different types of regression models
  - specify form of association

Linear regression

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$

- **The outcome $\boldsymbol{Y}$ is a function of a**
  - dependent variable

- **combination of other variables $\boldsymbol{X^s}$ and**
  - independent variable and controls
  - all “independent variables,” but
  - IV is the key interest hypothesized to “cause” Y

- **an error term $\boldsymbol{\varepsilon}$**
  - what’s not explained by the X^s

Linear regression

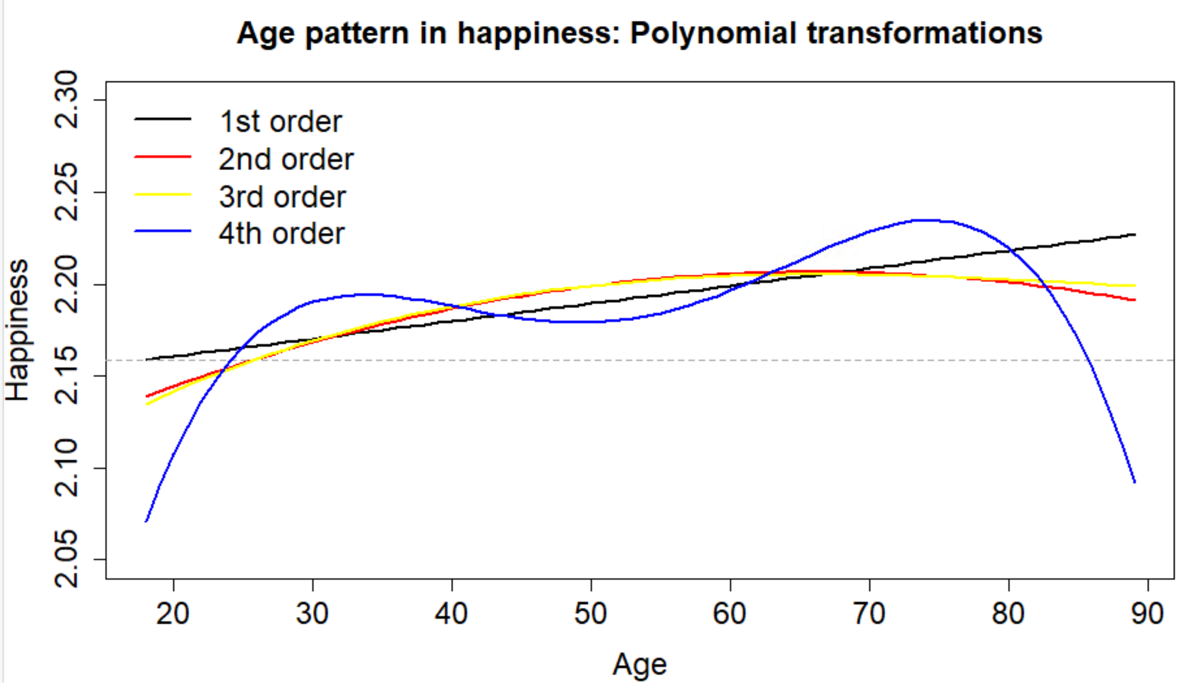
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$

- β_0 is the intercept (or constant)
  - the value of Y when all X^s = 0

- Recall our age-happiness example
  - added the constant to the hap. coef. value at age 18 to compute a reference line because there was no age 0

$$\beta_1 \text{ x } X_1 + \beta_0$$

```
abline(h=(0.0009603*18)+2.14151
```




```
reg hap age
```

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

Linear regression

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$

- X^s are independent variables

- β^s are regression coefficients
  - average estimated Δ in Y given a one unit Δ in X
    - holding all other X^s constant

- What is the predicted value of happiness at age 50?

```
reg hap age
```

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

Linear regression

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$

- We just interpreted a simple linear regression model
  - not yet held other X^s constant

- What would model look like if added education groups?
  - less than high school; high school; greater than high school

- Why might we not want to combine the edu. categories in one variable?
  - e.g., LHS=0, HS=1, GHS=2

Linear regression

- $Hap = \beta_0 + \beta(Age) + \beta(LTHS) + \beta(GTHS) + \varepsilon$
  - HS = reference group

M1

```
reg hap age if nmiss==0
```

hap	Coef.
age	.0009619
_cons	2.141542

M2

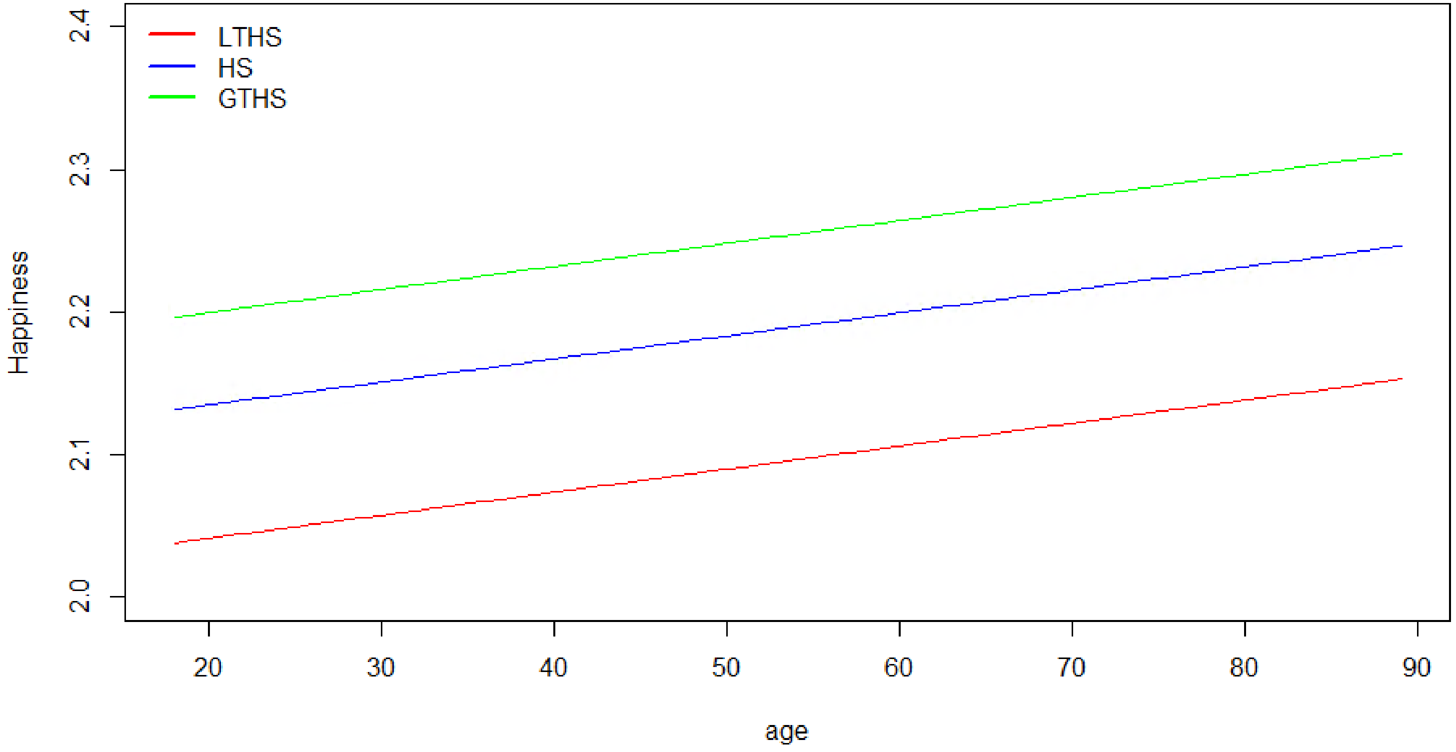
```
reg hap age lhs ghs if nmiss==0
```

hap	Coef.
age	.0016144
lhs	-.0933821
ghs	.0648564
_cons	2.10264

- Why is age coef. different between models?
- What does constant reflect in M1? What about M2?
- Does LHS or GHS have a greater impact on happiness?
- Why can't we compare age vs. edu. coef. strength?
- If we plotted age patterns by edu. what would they look like?

Linear regression

```
#less than high school
plot(age,(0.0016144*age)+(2.10264-0.0933821),type="l",col="red",ylim=c(2.0,2.4),ylab="Happiness")
#high school
lines(age,(0.0016144*age)+(2.10264),type="l",col="blue")
#greater than high school
lines(age,(0.0016144*age)+(2.10264+0.0648564),type="l",col="green")
legend("topleft",c("LTHS","HS","GTHS"),lty=c(1,1,1),lwd=c(2,2,2),
       col=c("red","blue","green"),bty="n")
```

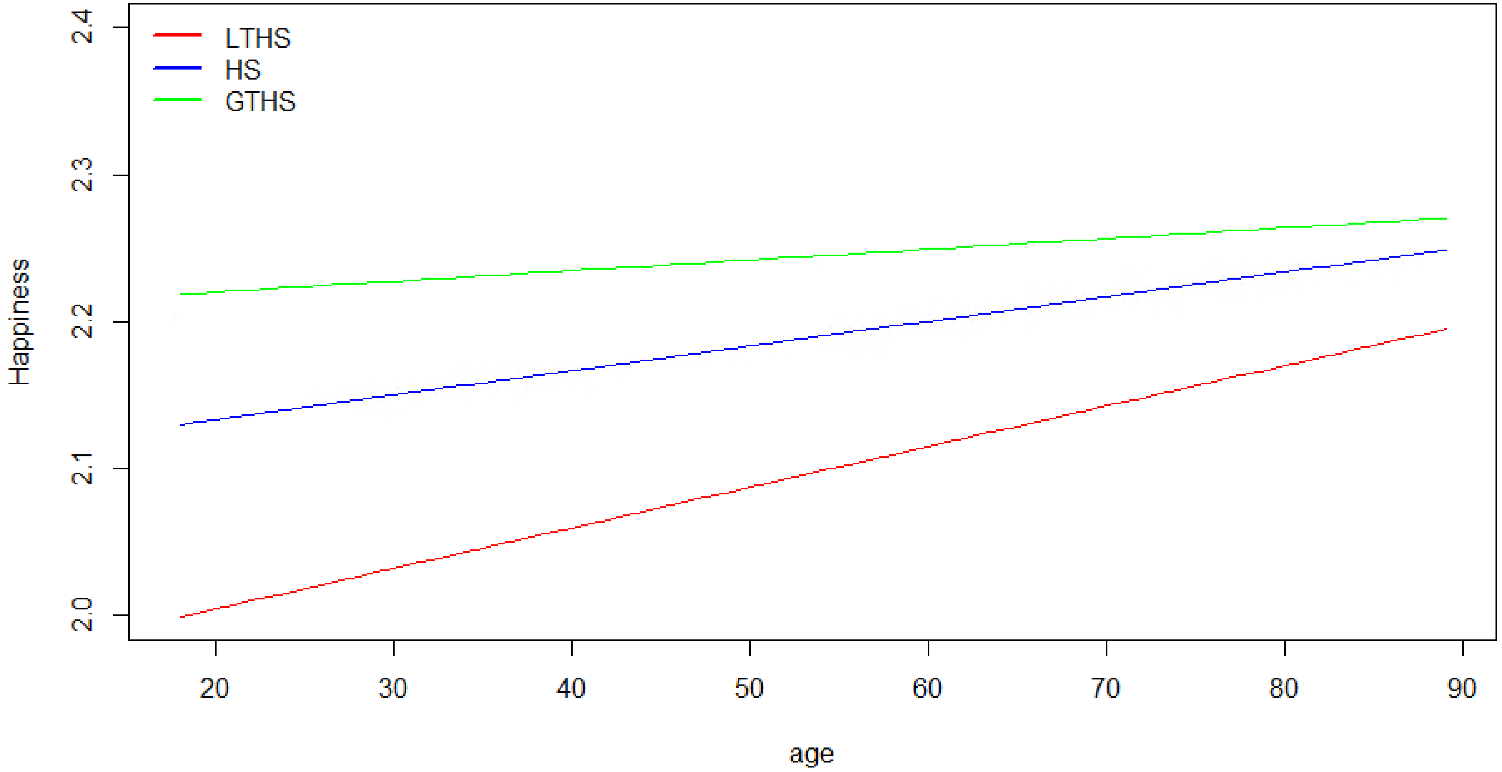


- Why are lines parallel?
- Wouldn't you expect happiness to follow a different pattern across age by education?
- To see whether age slopes differ by education need interaction terms

Linear regression

```
reg hap age lhs agelhs ghs ageghs if nmiss==0
```

hap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0016754	.0002699	6.21	0.000	.0011464	.0022043
lhs	-.1503634	.0203191	-7.40	0.000	-.190189	-.1105378
agelhs	.0010849	.0003889	2.79	0.005	.0003226	.0018472
ghs	.1058946	.017075	6.20	0.000	.0724276	.1393616
ageghs	-.0009485	.0003588	-2.64	0.008	-.0016518	-.0002453
_cons	2.099889	.0130507	160.90	0.000	2.07431	2.125469



- Note the slopes are no longer parallel
- Interaction terms
  - Age * LHS
  - Age * GHS
- To see if age had a different effect on happiness at different levels of education
- More on this next week
  - Later factor notation

OLS assumptions

1. Linearity
2. Mean independence
3. Constant error variance
4. Uncorrelated errors
5. Normal distribution of errors

Examples

- A lot of problems with happiness
  - violates almost every assumption because ordinal measure

- Domsat = satfam + satfrnd + sathlt + sathob + satres
  - (1) not satisfied to (7) very satisfied
  - summed and recoded to create satisfaction index
  - (1) not satisfied with any domain to (31) very satisfied with all domains

```
/*domain satisfaction index: 5 items 1-7 very great deal to none: reverse code*/
gen satfamily=8-satfam
replace satfamily=. if satfamily>7
gen satfriends=8-satfrnd
replace satfriends=. if satfriends>7
gen sathealth=8-sathealt
replace sathealth=. if sathealth>7
gen sathobbies=8-sathobby
replace sathobbies=. if sathobbies>7
gen satresidence=8-satcity
replace satresidence=. if satresidence>7
/*sum all 5 to create index*/
gen domsat=satfamily + satfriends + sathealth + sathobbies + satresidence
/*subtract 4 so range=1 to 31*/
replace domsat=domsat - 4
```

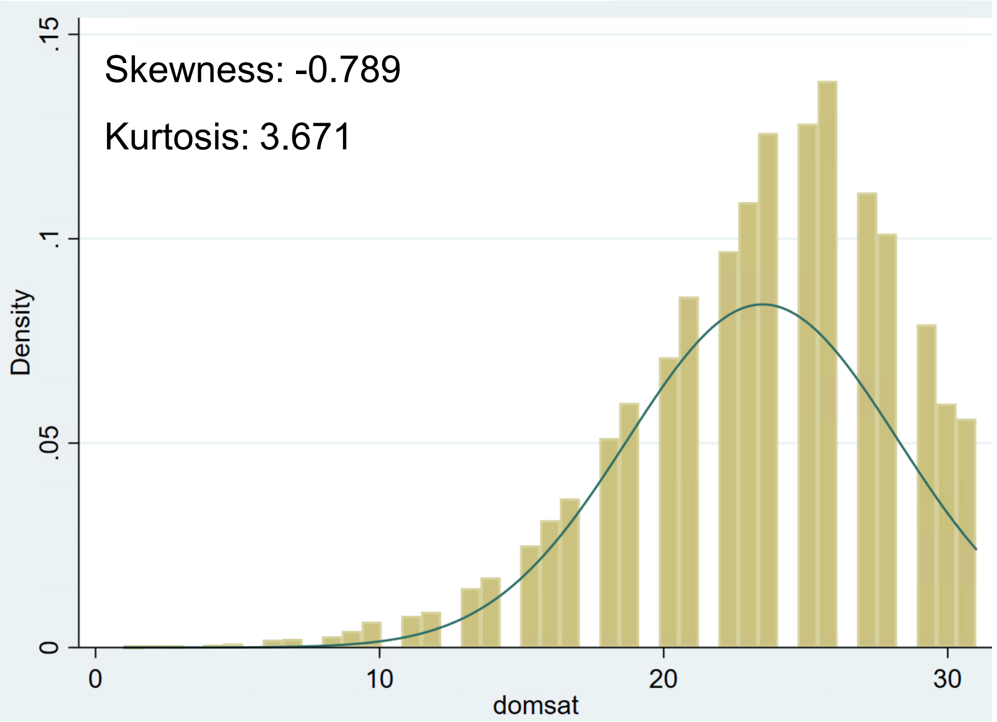
```
. sum domsat if nmiss==0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      domsat |     23,549    23.48677    4.752527          1         31

.
```

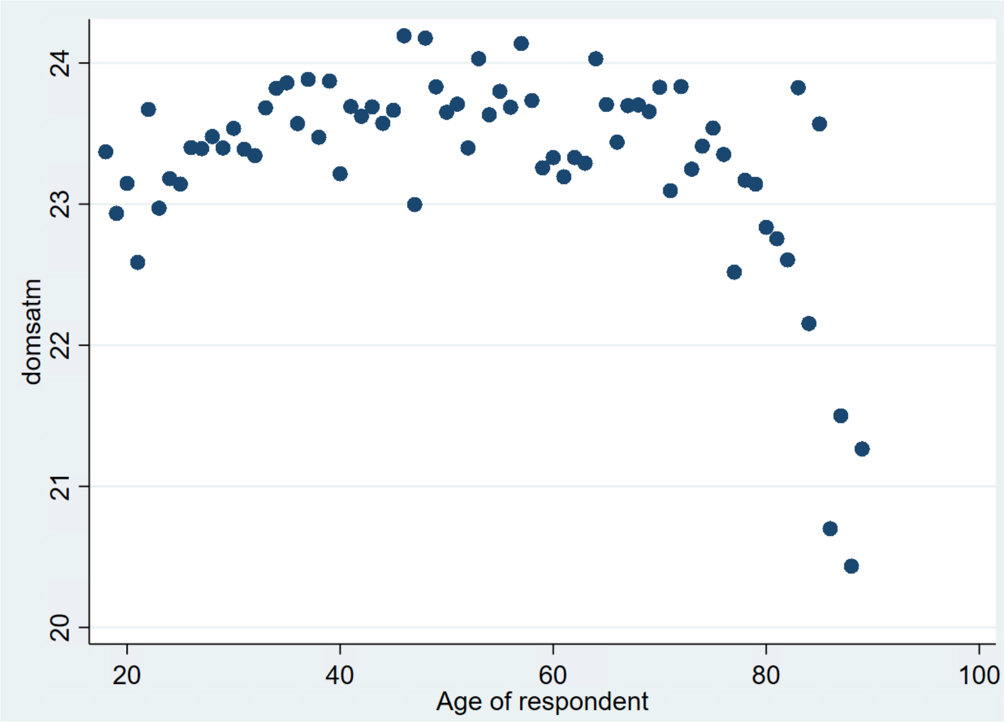
Domain satisfaction index

```
histogram domsat if nmiss==0, normal
```



- Skewed?
  - Right or left?

```
*egen domsatm=mean(domsat), by(age)
scatter domsatm age
```



- How many inflection points?
  - What should reg include?

Domain satisfaction index

```
reg domsat age if nmiss==0
```

Source	SS	df	MS
Model	.26910782	1	.26910782
Residual	531866.86	23,547	22.5874574
Total	531867.13	23,548	22.5865097

Number of obs	=	23,549
F(1, 23547)	=	0.01
Prob > F	=	0.9131
R-squared	=	0.0000
Adj R-squared	=	-0.0000
Root MSE	=	4.7526

domsat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.0001925	.0017637	-0.11	0.913	-.0036496	.0032645
_cons	23.49537	.0846016	277.72	0.000	23.32954	23.66119

```
reg domsat age age2 if nmiss==0
```

Source	SS	df	MS
Model	1768.95676	2	884.478378
Residual	530098.173	23,546	22.5133005
Total	531867.13	23,548	22.5865097

Number of obs	=	23,549
F(2, 23546)	=	39.29
Prob > F	=	0.0000
R-squared	=	0.0033
Adj R-squared	=	0.0032
Root MSE	=	4.7448

domsat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0887041	.0101829	8.71	0.000	.068745	.1086633
age2	-.0009127	.000103	-8.86	0.000	-.0011146	-.0007109
_cons	21.62726	.2270576	95.25	0.000	21.18221	22.07231

OLS assumptions: linearity

- Y is a linear function of Xs plus a random error term
- Diagnosis: examine the residuals (deviations from the fitted line to the observed values) against the explanatory variable
  - residual plot and smoothing line

OLS assumptions: linearity

```
reg domsat age c.age#c.age female nonwhite if nmiss==0
```

domsat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0872315	.0100797	8.65	0.000	.0674746	.1069885
c.age#c.age	-.0009242	.0001019	-9.07	0.000	-.001124	-.0007244
female	.1904473	.0618171	3.08	0.002	.0692817	.3116129
nonwhite	-1.865783	.0844855	-22.08	0.000	-2.03138	-1.700186
_cons	21.90364	.2282358	95.97	0.000	21.45629	22.351

```
acprplot age, lowess
```

- Does the smooth line approximate the regression line?
- Is the entire pattern uniform?
- We can do better

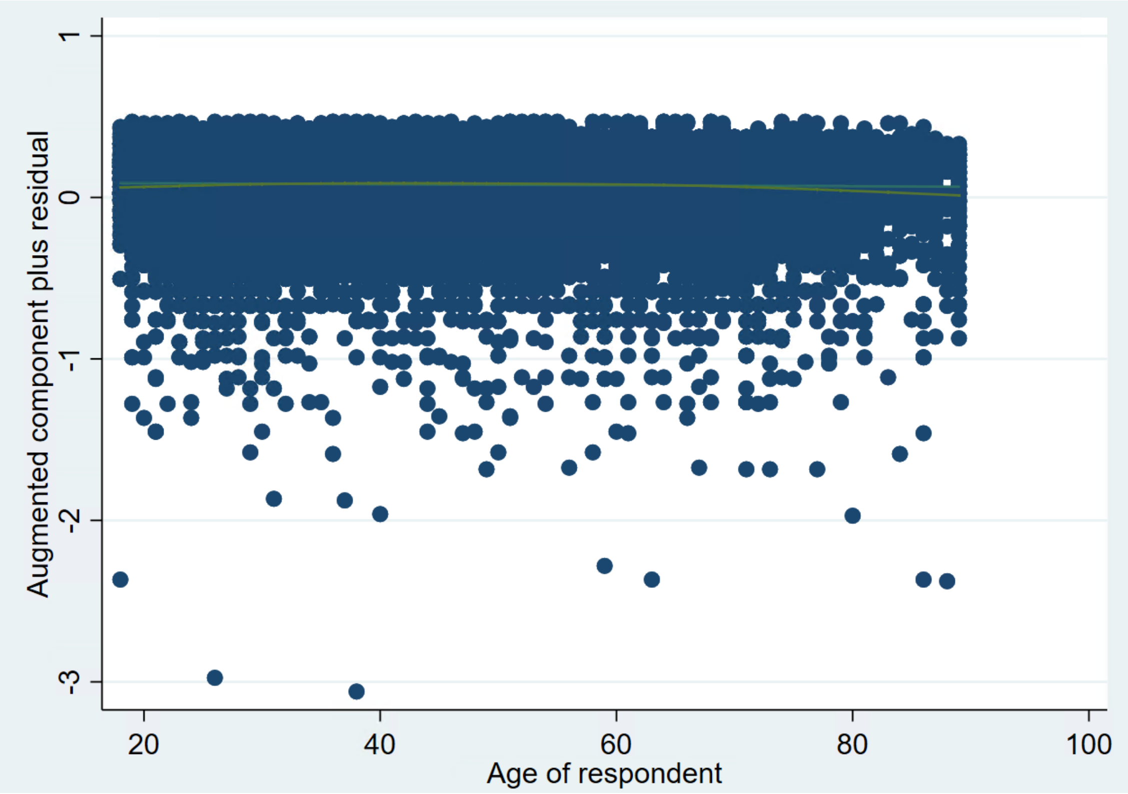
OLS assumptions: linearity

- If y is entirely positive: log transformation
  - adjusts for skewed DV (see also sktest)

```
gen domsatlog=ln(domsat)
reg domsatlog age c.age#c.age female nonwhite if nmiss==0
```

domsatlog	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0041634	.0005183	8.03	0.000	.0031475	.0051793
c.age#c.age	-.0000459	5.24e-06	-8.76	0.000	-.0000562	-.0000356
female	.0099845	.0031786	3.14	0.002	.0037542	.0162149
nonwhite	-.094928	.0043442	-21.85	0.000	-.103443	-.086413
_cons	3.059741	.0117359	260.72	0.000	3.036738	3.082744

```
acprplot age, lowess
```



- The residuals are more uniform, and the smoothed line is a little closer
  - Trade off: [interpretation of log scale]

OLS assumptions: mean independence

- Error term mean is zero, and not dependent on value of Xs

- Violated if…

1. Measurement error: systematic misreporting
2. Omitted Xs
3. Reverse causation: Y has causal effect on X

- Diagnosis: theory, previous literature, sensitivity analyses
    - strive for parsimony vs. kitchen sink

OLS assumptions: constant error variance

- Error term variance is not dependent on the Xs' value
  - homoscedasticity: variance of error is same across all levels of X
    - → inefficiency and/or biased standard errors

- Diagnosis: plot residuals against fitted values (rvfplot)
  - values NOT cluster in even width → heteroscedasticity

- Also, Breusch-Pagan test
  - whether variance of residuals is homogenous
    - large chi-square and p-value < 0.05 → heteroscedasticity

OLS assumptions: constant error variance

```
reg domsat age c.age#c.age female nonwhite if nmiss==0
```

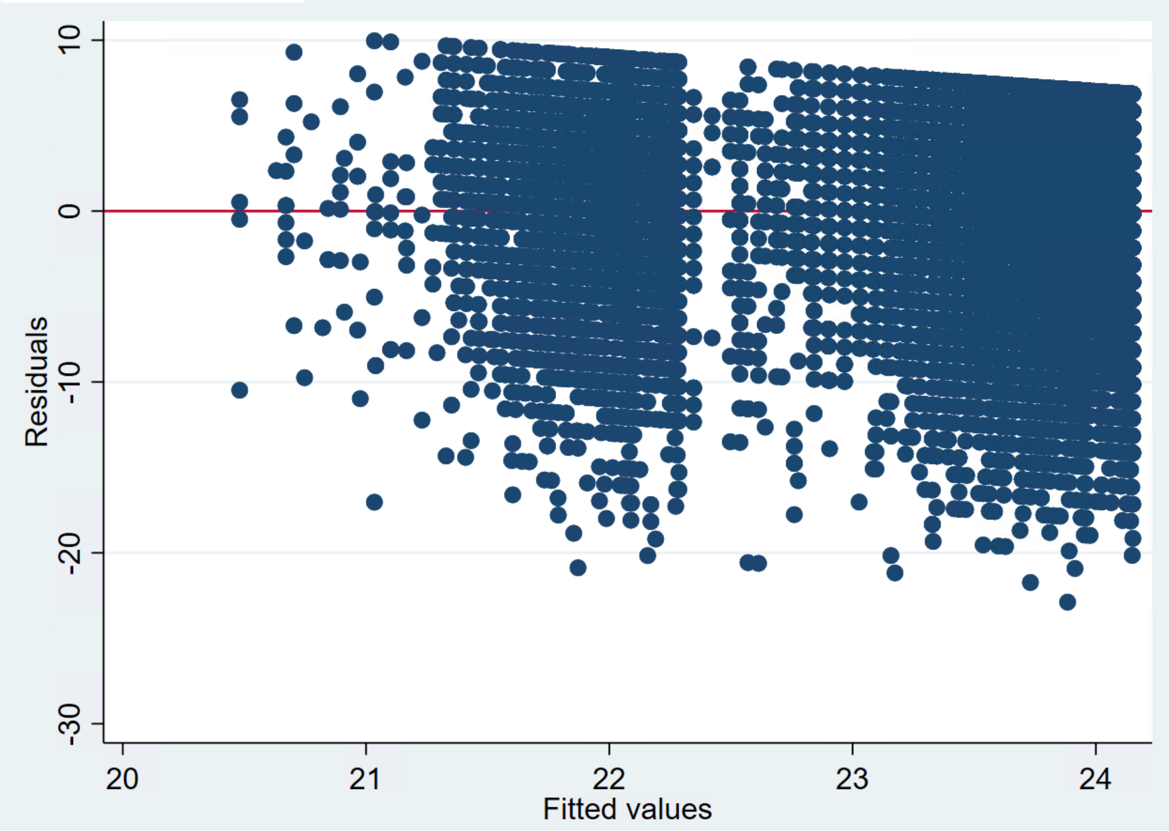
domsat	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
age	.0872315	.0100797	8.65	0.000	.0674746	.1069885
c.age#c.age	-.0009242	.0001019	-9.07	0.000	-.001124	-.0007244
female	.1904473	.0618171	3.08	0.002	.0692817	.3116129
nonwhite	-1.865783	.0844855	-22.08	0.000	-2.03138	-1.700186
_cons	21.90364	.2282358	95.97	0.000	21.45629	22.351

```
estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of domsat

         chi2(1)      =   138.28
         Prob > chi2  =   0.0000
```

```
rvfplot, yline(0)
```



- Reject null hypothesis
- note pattern in plot
  - not uniform around fitted line

OLS assumptions: constant error variance

```
reg domsat age c.age#c.age female nonwhite if nmiss==0
```

domsat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0872315	.0100797	8.65	0.000	.0674746	.1069885
c.age#c.age	-.0009242	.0001019	-9.07	0.000	-.001124	-.0007244
female	.1904473	.0618171	3.08	0.002	.0692817	.3116129
nonwhite	-1.865783	.0844855	-22.08	0.000	-2.03138	-1.700186
_cons	21.90364	.2282358	95.97	0.000	21.45629	22.351

```
reg domsat age c.age#c.age female nonwhite if nmiss==0, robust
```

domsat	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0872315	.0104239	8.37	0.000	.0667999	.1076631
c.age#c.age	-.0009242	.000107	-8.64	0.000	-.0011339	-.0007145
female	.1904473	.062005	3.07	0.002	.0689134	.3119812
nonwhite	-1.865783	.0916303	-20.36	0.000	-2.045385	-1.686182
_cons	21.90364	.2316908	94.54	0.000	21.44951	22.35777

- Robust SE
- Not impact coef.
  - compare models
- Also, can use weighted least squares, but
  - have to know weights
- Resource link

OLS assumptions: uncorrelated errors

- Error term for any obs. is uncorrelated with error term for any other obs.

- Why would we expect error terms for any obs. to be correlated with error term for any other obs.?
  - shared environment and/or repeated observation
    - multilevel modeling, longitudinal design

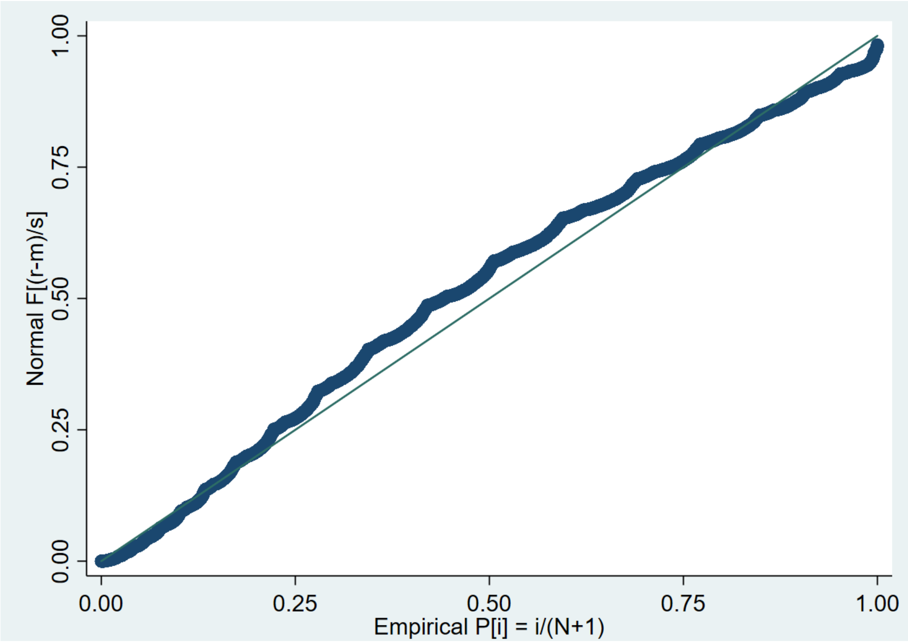
- Diagnosis: No good way to examine unless your data are clustered

OLS assumptions: normal distribution of errors

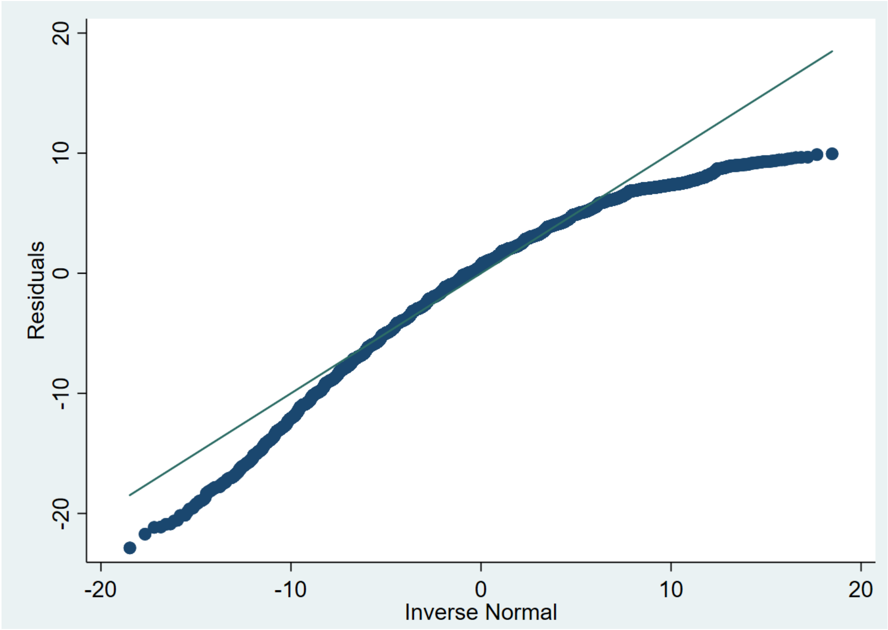
- Diagnosis: pnorm (middle range) qnorm (tails)
  - save residuals

```
reg domsat age c.age#c.age female nonwhite if nmiss==0
predict resid, resid
```

```
pnorm resid
```




```
qnorm resid
```



- Not a problem if N > 100 and Xs < 5
  - CLT: N increases approximate parameters even if errors not normally distributed

Regression diagnostics

- Multicollinearity
  - if goal is to isolate effect of IV on DV, then don't want variation in IV to map too closely onto variation in other x variables
    - inflates SE

- Variance Inflation Factors (VIFs): test for multicollinearity
  - Can one x be held constant while another varies?

- VIF = 1: provides completely independent information on y
- VIF = 2.5: problematic threshold
- VIF = 10: should probably be dealt with

Multicollinearity: solutions

- Ignore: used for control purposes only and effects will not be reported

- Drop one or more from the model
  - keep whichever is most theoretically meaningful
    - if no difference, then choose largest VIF

- Combine into an index
  - based on previous research, or own theoretical justification

```
reg domsat age c.age#c.age female nonwhite if nmiss==0
vif
```

Variable	VIF	1/VIF
age	33.45	0.029894
c.age#c.age	33.46	0.029889
female	1.00	0.996906
nonwhite	1.01	0.994657
Mean VIF	17.23	

Regression diagnostics

- Influential observations (outliers)
    - dispersion of variables

- Thoroughly examine dispersion of all variables
    - range, SD, scatter plot, box plot

- Make sure:

a) no coding errors
b) missing values handled properly
c) run model with and without outliers

Influential observations: solutions

- Ignore: if not substantively influence results
  - observations are distinct, but theoretically expect such cases in this population
- Transform: if substantively influence results, but expected to do so
  - some variables we expect to have outliers
- Exclude: no reason to expect outliers, or such an extreme
  - possibly an error, or due to some random circumstance
- Robust regression techniques
  - places less weight on outliers, and removes extreme cases

Next class we will…

- discuss mediation and moderation