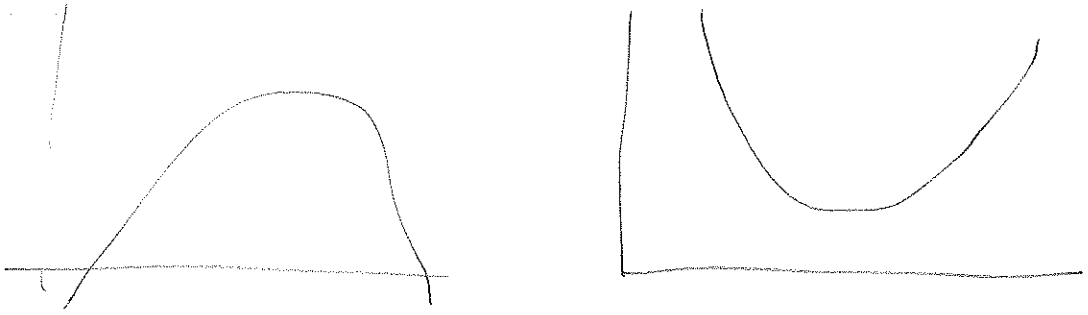
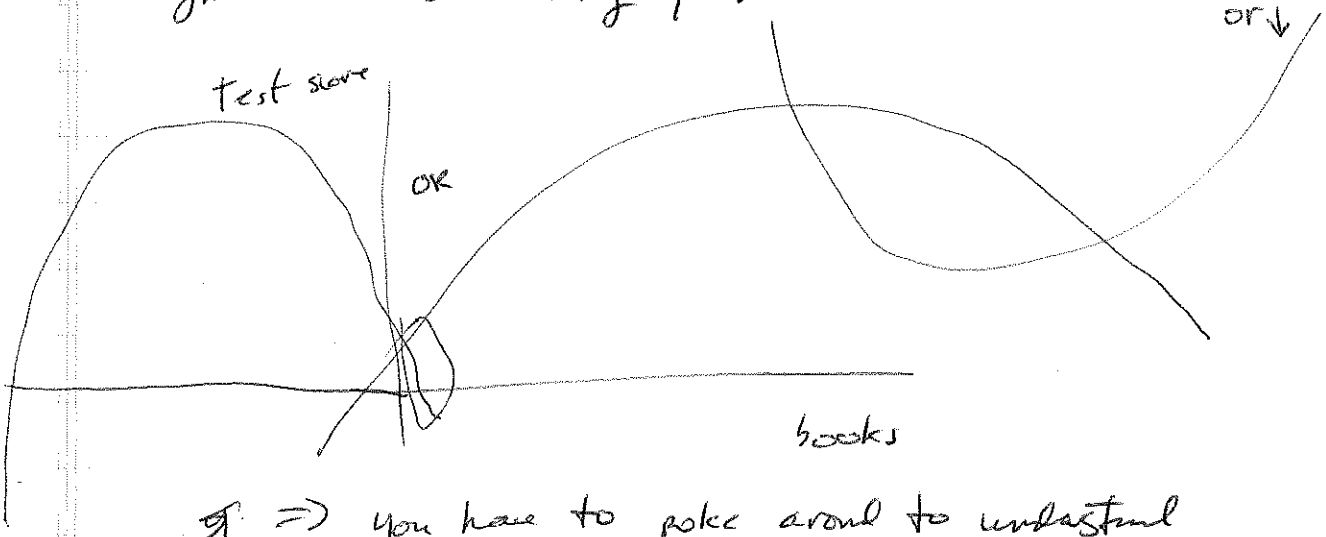
4th class Part I

Quadratics -- annoying but necessary

Formula for the quadric: $y = a + b_1 X + b_2 X^2$



symmetrical BUT any position.



☹ ⇒ you have to poke around to understand

☺ It is very useful to consider the formula for the slope

$$\frac{dy}{dx} = b_1 + 2b_2 X$$ Formula for the slope

4th class Part I

Formula for the quadric ⇒ plug in values for X to get points on the curve

" " " slope of the quadric ⇓ plug in values for X to get slope at any point.

Slope

1. Every parabola has a maximum or minimum point

where is it? It is the X where slope is zero

if $\text{slope} = b_1 + 2b_2X$

then $0 = b_1 + 2b_2X \Rightarrow X = \frac{-b_1}{2b_2}$

but slope formula is more generally useful -- e.g.

what is slope at $X=0$?

$\text{slope} = b_1$ so b_1 tells you slope at $X=0$

sign of X^2 tells you "$\cup$" or "$\cap$"

$b_2 < 0 \Rightarrow \cap$

$b_2 > 0 \Rightarrow \cup$

Table 2. Estimated Racial Achievement Gap over the First Four Years of School, Math

	(1) Fall-K	(2) Spring-K	(3) Spring-1st	(4) Spring-3rd	(5) Fall-K	(6) Spring-K	(7) Spring-1st	(8) Spring-3rd
Black	−0.663 (0.025)	−0.724 (0.027)	−0.758 (0.029)	−0.882 (0.031)	−0.099 (0.026)	−0.209 (0.028)	−0.279 (0.031)	−0.382 (0.033)
Hispanic	−0.738 (0.024)	−0.681 (0.025)	−0.568 (0.026)	−0.539 (0.026)	−0.197 (0.024)	−0.189 (0.026)	−0.122 (0.027)	−0.078 (0.028)
Asian	0.11 (0.058)	0.088 (0.056)	−0.025 (0.052)	0.066 (0.054)	0.258 (0.050)	0.226 (0.050)	0.092 (0.047)	0.163 (0.049)
Other race	−0.495 (0.047)	−0.481 (0.048)	−0.497 (0.050)	−0.541 (0.050)	−0.158 (0.040)	−0.175 (0.043)	−0.21 (0.046)	−0.244 (0.046)
Age (in months)					0.058 (0.002)	0.053 (0.002)	0.037 (0.002)	0.019 (0.002)
Birth weight (in ounces)					0.003 (0.000)	0.003 (0.000)	0.003 (0.000)	0.003 (0.000)
Female					0.005 (0.017)	−0.005 (0.017)	−0.044 (0.018)	−0.175 (0.018)
Number of children's books					0.006 (0.001)	0.006 (0.001)	0.005 (0.001)	0.006 (0.001)
Number of children's books (squared) (*1000)					−0.021 (0.002)	−0.02 (0.003)	−0.019 (0.003)	−0.020 (0.003)
Mother over 30 at first birth					0.165 (0.026)	0.107 (0.025)	0.086 (0.022)	0.083 (0.024)
Socioeconomic status measure					0.306 (0.016)	0.282 (0.015)	0.256 (0.015)	0.288 (0.015)
Mother receives Wic benefits					−0.212 (0.021)	−0.191 (0.022)	−0.19 (0.023)	−0.208 (0.024)
Mother a teenager at first birth					−0.114 (0.021)	−0.118 (0.022)	−0.131 (0.025)	−0.132 (0.025)
Constant	0.307 (0.013)	0.304 (0.013)	0.286 (0.012)	0.275 (0.012)	−4.357 (0.154)	−3.952 (0.160)	−2.795 (0.168)	−1.576 (0.168)
Observations	11201	11201	11201	11201	11201	11201	11201	11201
R-squared	0.11	0.11	0.1	0.12	0.32	0.29	0.24	0.26

Notes: The dependent variable is the math test score at various points in a student's career. Test scores are IRT scores, normalized to have a mean of 0 and a standard deviation of 1 in the full, unweighted sample. Non-Hispanic Whites are the omitted race category, so all of the race coefficients are relative to that group. The unit of observation is a student. Standard errors in parentheses. Estimation is done using weighted least squares, using sample weights provided in the data set.

4th class Part ~~II~~ I

Levitt and Fryer

Fall-K math:

$$T = a + .006\,\text{books} - .021\,\text{books}^2$$
$$(.001) \qquad (.002)$$

"*1000" means that $-.021$ is really $-.000021$

$.002$ " " $(.000002)$

NEVER enter $.000$ into a table
$(.000)$ ⇒ completely
uninformative.

Figure out how to provide useful numbers

First class

$$Ed = a + .14\,\text{Family Income}$$
$$(.02)$$

Income in $10,000s = $ $.14$ $(.02)$

Income in dollars = $.000014$ $(.000002)$

$.14\ (\times 10^{-4})$

$(.02)\ (\times 10^{-4})$

return to

$$T = a + .006B - .021B^2$$

① negative coeff on squared term tells you 

② when is max?

$$\frac{-b_1}{2b_2} = \frac{-.006}{.000042} = 142 \quad (\text{mean} \sim 60)$$

③ what is slope at zero?

unit coeff on B $\Rightarrow$ $\pm .006$

$\Rightarrow$ first book is associated with a .006 sd increase in test score

first 10 books $\sim$ .06 sd

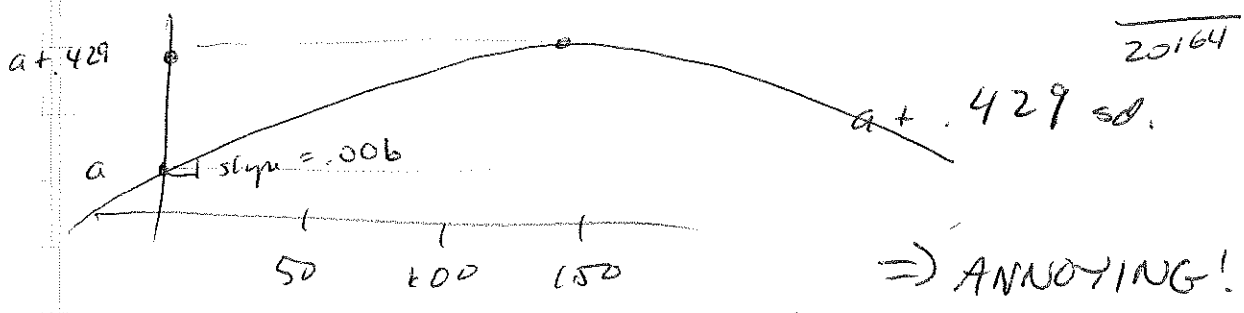
④ what is value of function at B = 0?

$T = a$

.852 −.423 −.000021×

at max? $B = 142 \Rightarrow a + .006(142) - (142)^2$

20164



$\Rightarrow$ ANNOYING!

Latent growth models -- often quadratic

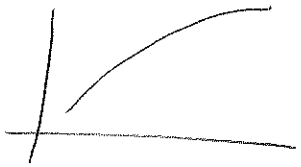
Level 1 (child) $T = a + b_1 Age + b_2 Age^2$

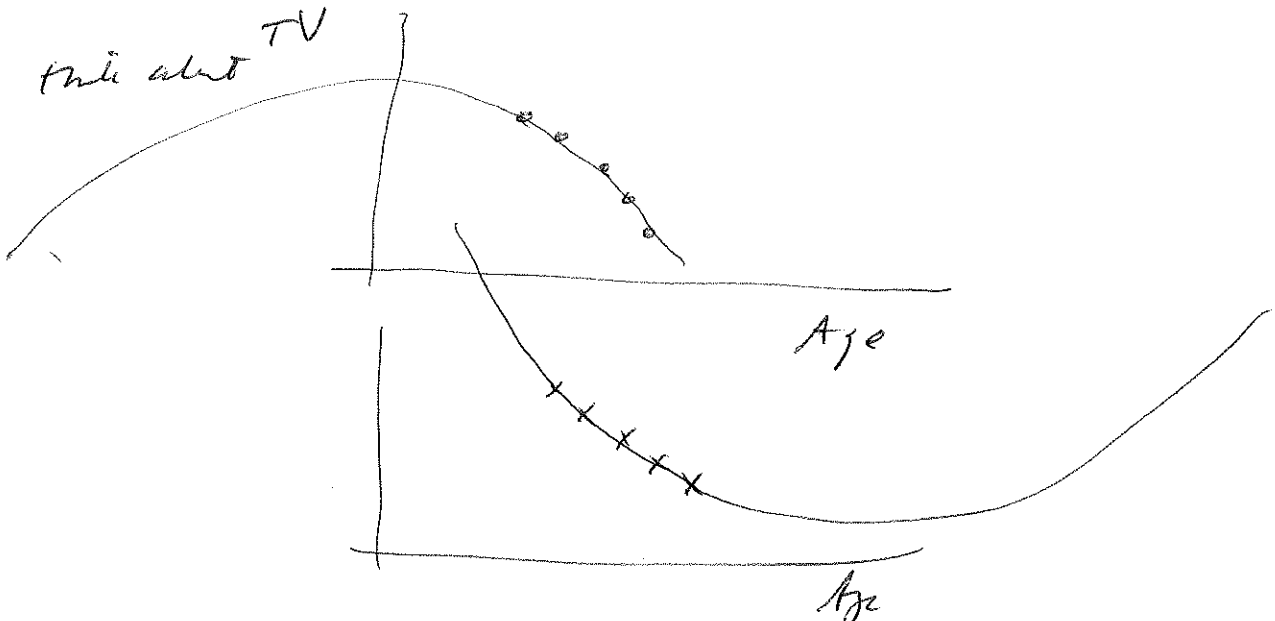
⇒ fit a quadratic to each child

$a = a + b_1 Demog$

$b_1 = c + d_1 Demog$

$b_2 = e + f_1 Demog$

even though  in the aggregate, each person's change is curvilinear



Offset b_1 (slope at Age = 0)
opposite sign b_2

] from very similar trajectories

4th class Part II

move on **Dummy variables -- interaction**

Suppose two groups Blacks (B) Whites
→ interest Black -- B

is Fam Inc less advantage for blacks?

Whites: $Ed = a + b_1 Inc$

Blacks $Ed = c + d_1 Inc$ Is $d_1 < b_1$?

Investigate with dummy variable interaction

Data set: only Blacks & Whites

Include both groups all races:

$$Ed = e + f_1 Inc + f_2 B + f_3 B * Inc$$

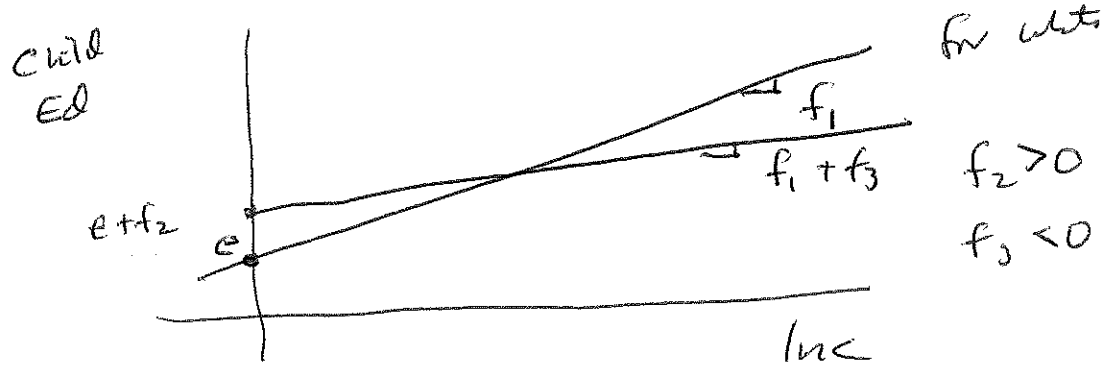
why is that the same?

For Whites: $Ed = e + f_1 Inc + 0$

$e = a \quad f_1 = b_1$

For blacks: $Ed = (e + f_2) + (f_1 + f_3) Inc$

$= c \qquad = d$



4th class Part II

key is that f_3 provides exactly the sig test that you want.

Breakout group:

Suppose that you pooled groups and ran

$Ed = c + f_1 Inc + f_3 B * Inc$

(i.e. left out the main effect of B)

What would the graph look like?

Suppose 3 group whites, Black(B) and Hispanic(H)

Pool and estimate

$Child\text{-}Ed = g + h_1 FamIncome + h_2 B + h_3 H + h_4 B * FamInc + h_5 H * FamInc$

For whites: $Ed = g + h_1 FamIncome$

Blacks $Ed = (g + h_2) + (h_1 + h_4) FamIncome$

Hisp $Ed = (g + h_3) + (h_1 + h_5) FamIncome$

h_4 and h_5 test the interaction

Also might test ~~Fam~~

$B = H = 0$

Use dummy variables for missing data?

Suppose Race/Ethnicity = White (B), Black, Hispanic (H), missing (M)

~~Gl Lath~~

Fryer-Levitt

$$Test = a + b_1 B + b_2 H + b_3 M$$

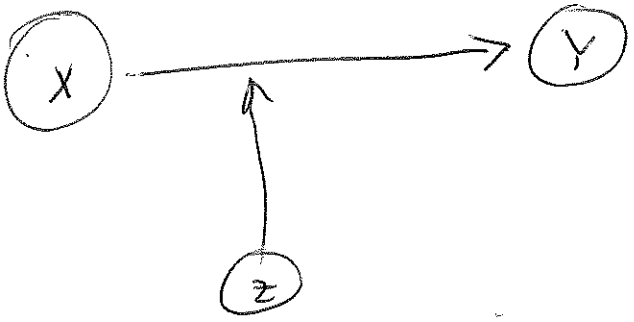
b_3 = do missing cases differ from whites?

Useful for minor variables

4th class Part III (asynchronous)

Mediation and Moderation

Moderation -- interactions. Does the effect of X on Y depend on Z?



Z should be an exogenous variable and it often a demographic measure

Does the effect of family income (X) on children's schooling (Y) depend by race?

Misusing moderation Suppose treatment T and outcome Y

and suppose we run $Y = a + b_1 T + \text{other controls}$

and find that b_1 is insignificant - nice try!

b_1 is the "main effect" of the experiment.

Oftentimes null effects launch people off on a search for subgroups for whom T does matter.

Two problems.

1. If you don't preregister experiments then your chance of finding a fluke interaction is high.

2. More serious -- Suppose two subgroups A & B. If overall effect is null but effect for subgroup A is positive and significant then effect for subgroup B is negative

   B is almost always swept under the rug.

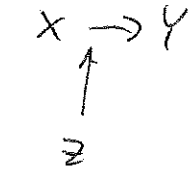
   Honesty demands as much theorizing about negative effects for B as positive effects for A.

   e.g. ECE and disadvantaged children

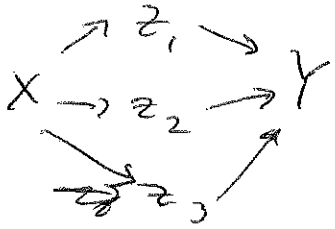
Mediation mediation is about exogenous variables that differentiate effects of some X on Y

mediator is about endogenous variable that accounts for the effect of X on Y

Moderation



Mediation



A mediated model is sometimes called a "causal model" in statistics, a focus on understanding Y in the presence and absence of X for individuals

Causal model

accounts for

Conceptual sense -- what ~~are effects~~ the ~~map~~ effects of X on Y

$X \longrightarrow Y$ "total effect"

$X \rightarrow Z \rightarrow Y$ direct effect + indirect effect operating through Z

Ideally, you want to know all of the Z_i's that account for the effect and not have any indirect effect.

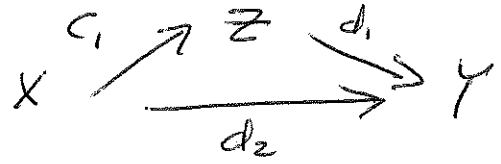
Mediated models are typically estimated with SEM

SEM handles measurement issues well but at its heart SEM models are just regression models

suppose $X \xrightarrow{b_1} Y$ "reduced form" "total effects"

$$Y = a_1 + b_1 X$$

if



then regression models treat everything with arrows running to them as dependent variables

(1) $Z = a_2 + c_1 X$

(2) $Y = a_3 + d_1 Z + d_2 X$

If you substitute (1) into (2)

Compare

$$Y = a_3 + d_1[a_2 + c_1 X] + d_2 X$$

$$= [a_3 + d_1 a_2] + d_1 c_1 X + d_2 X$$

$$= \text{constant} + [d_2 + d_1 c_1] X$$

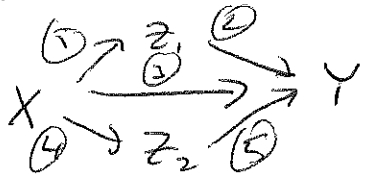
In other words, the total effect of X on Y is the sum of the

direct effect (d_3) that operates independently of the mediator

indirect effect ($c_i d_1$) that operates through the mediator

Try it -- it's exactly the same.

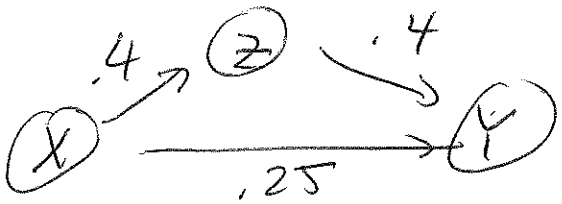
If more than one mediator -- Z_1 & Z_2 -- then



Still the case that the total effect of X on Y is the sum of the direct effect (3) and the two indirect effects (1×2) & (4×5)

Important things to note:

1. Indirect effect involve multiplication of path coefficients



looks like a really important indirect effect
= but it's .4 × .4 = .16

Still holding model

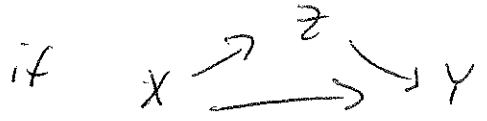
if only we can boost S_1 the kids' math trajectories are saved.

But what do the effects of S_1 on S_4

$b_1 \times b_2 \times b_3 \times b_4$

if b's are .6, then $.6^4$ = .13 not so big
if .4, then .03

2. Just because this is a causal model in a structural sense doesn't mean you shouldn't pay attention to omitted variable bias



then indirect effect attributed to Z might rightly belong to other mediators

more important: do not control should you add. There should be just as many.

And even if X is validly causing, path X → Z → Y is not caused.