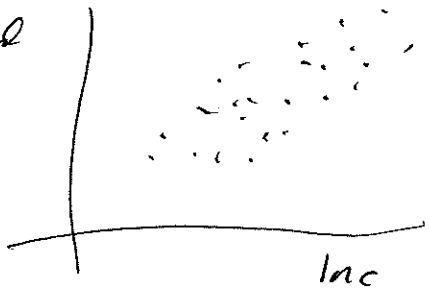
Logistic regression -- very annoying but necessary

Thus far, our dependent variables have been continuous



$\rightarrow Y = a + b_1 X$

What if dichotomous?
(two valued)

First point: variation in a dependent variable is a terrible thing to waste!

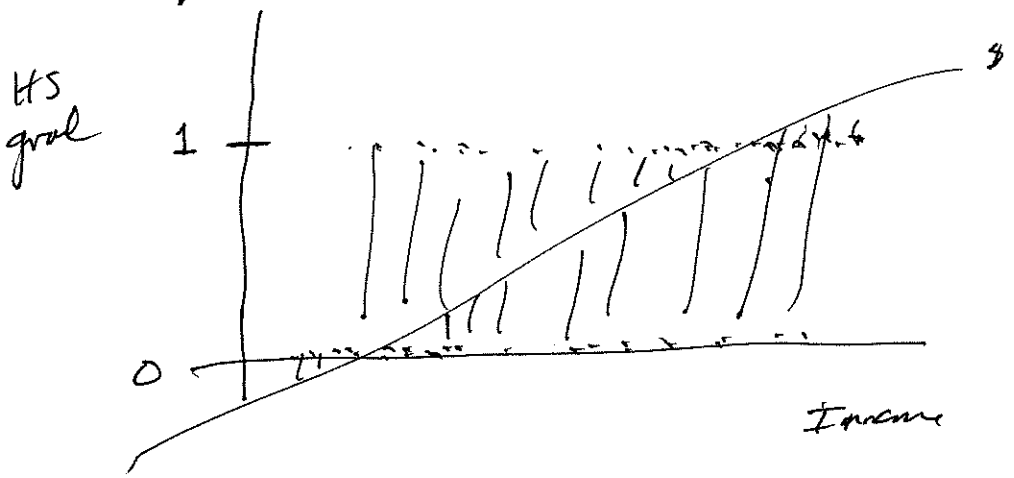
Duncan et al. analyzed years of Ed but also whether hs grad (1,0)

In most cases, go with the continuous variable even if it is not well distributed

If points of the distribution are of particular interest. (1,0) or quantile regression

But sometimes you are stuck with a 1,0 variable.

Suppose (1,0) = a + b, Family Income
HS grad



what does ~~scatter~~ scatter look like

OLS: do your job! minimize squared residuals

It gives you the best fitting slope -- say .03 (.01)

How do interpret .03

Second big point: think of the 0, 1 scale as a 0 → 100 percent chance of being a 1

If 1/2 of people (exactly) like me finished high school, then my chance of finishing are 50%

If 1/4 → 25%

3/4 → 75%

Now the .03 makes a lot of sense
(.01)

-- it is the increase in probability of graduating hs school associated with a $10,000 increase in income
So if I am at 50% at the .03 is a causal estimate, then my chances will be 53% with more money

Very straightforward! 1,0 = a + β_1 Fam Inc
is a linear probability model

⇒ change in ~~inco~~ probability as a linear function of income .25 → .28
.75 → .78

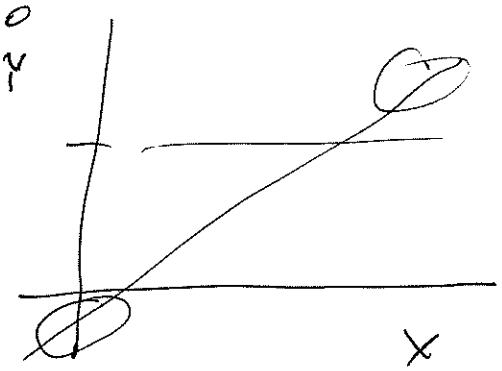
Lot of people use LPMs because ~~you can do~~ they aren't bad with $\bar{Y}$ is between .2 and .8

But LPMs offend the sensibilities of statisticians!

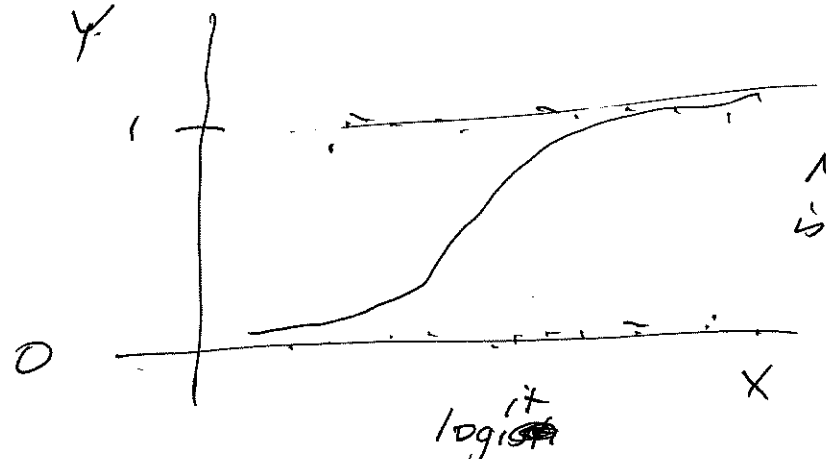
Horrors!

LPM problems:

1. Predictions below 0 & above 1
2. Non constant variance



Want a nonlinear model that doesn't exceed bounds



Note that slope is constantly changing

logit

two ~~bivariate~~ models -- ~~logistic~~ and probit

Almost interchangable. Logit is more popular

Instead of $P = a + b_1 X_{1i}$

$$\ln\left(\frac{P}{1-P}\right) = a + b_1 X_{1i}$$

Ugh!

5th class Part I

$$\ln\left(\frac{p}{1-p}\right)$$

natural logarithm ↑ odds ratio

	p	odds
Odds $\frac{p}{1-p}$	.5	$\frac{.5}{.5} = 1$ even odds
	.67	$\frac{.67}{.33} = 2$ 2 to 1
	.33	$\frac{.33}{.67} = \frac{1}{2}$ 1 to 2

If we run $\ln\frac{p}{1-p} = a + b_1 \text{Fam Inc}$

what to make of b_1

literally change in log odds associated with a one unit $10K change in Fam income

Still unclear -- how to talk about change in log odds ??

$$\ln\left(\frac{p}{1-p}\right) = a + b_1 \text{Fam Inc}$$

because $e^{\ln x} = x$

$$e^{\ln \frac{p}{1-p}} = \frac{p}{1-p} = e^{a + b_1 \text{Fam Inc}}$$

if X change by 1 unit $\Rightarrow \Delta \frac{p}{1-p} = e^{b_1}$

so take logit coeff and exponentiate to get change in log odds

5th class Part I

In Duncan et al. .23 (07)

$.00 \Rightarrow e^0 = 1$

1 = no effect on odds

literally Δ in log odds of completing hs associated with an extra $10K in income

If log odds change by .23 the odds ratio changes by $e^{.23} = 1.259$

so odds change by 25.9%

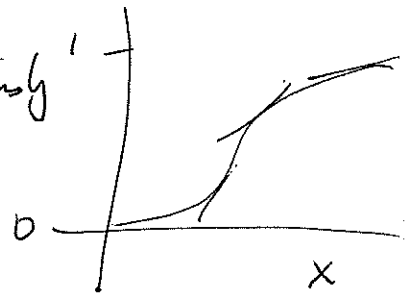
if −.15 then $e^{-.15} = .86$

odds drop by 14%

most medical articles focus on odds ratios

Get me back to changes in probabilities!!

Key challenge: Δ pr is changing continuously



Need some way of evaluating slope at some relevant point on the curve

"Margins" give you a probability interpretation at mean of X's or any other place you want.

5th class Part I

My preference is to have regression table with coefficients and standard errors that have probability interpretation

Key takeaways

1. ~~Ethan~~ LPM's are usually fine for exploration and econ journals

2. Logit with "margins" is good because of its probability interpretation

3. log-odds and odds ratios if you must!

Compare Prospecting = a + b, TV

	coefficient	s.e.	t or z stat
OLS (LPM)	.066	.015	4.52
log odds	.405	.142	4.27
odds ratio ($e^{.405}$)	1.500	.094	4.27
marginal at mean	.089	.021	4.27

$\beta(P)(1-P)$ .096

↑ plug in mean

5th Class Part II [Asynchronous] -1-

This week marks a turning point in the class

Up to Now: basic tools of regression analysis

Next 2+ weeks: using regression for causal analysis

"Every regression is an experiment". Q: is it a convincing causal experiment?

How to identify causal effects?
aka "identification strategies"

use example of effect of income on child ed. "$5K more

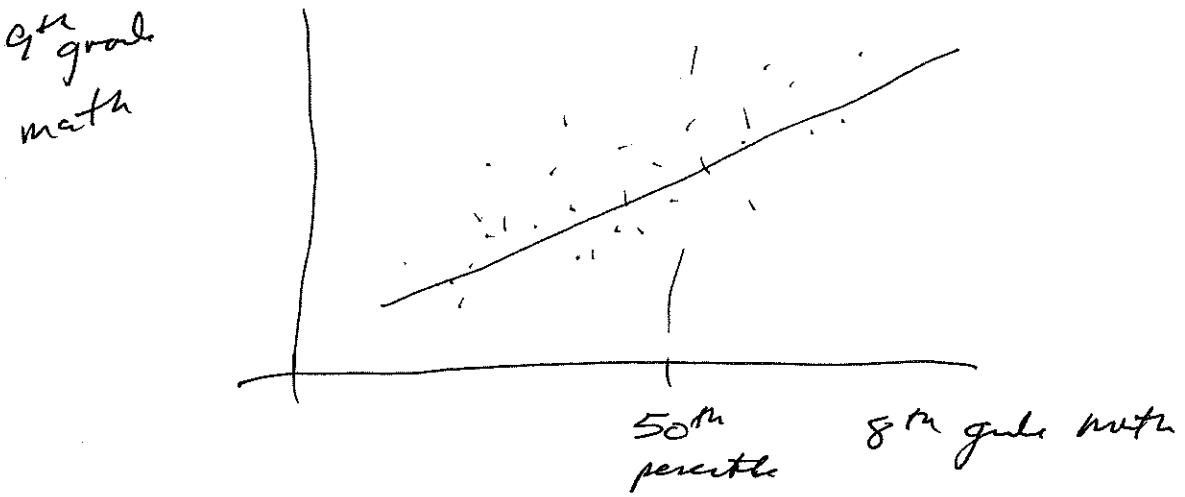
Procedure	Strength of identification strategy	Problems
1. Random assignment (of $5K)	Very strong ~~criteria~~	Imbalance, generalizability
2. Regression with controls	Often weak	omitted variable bias
3. Quasi-experimental approaches "natural experiments"		

- ~~case~~ regression discontinuity -- Wong + Cortes
- differences in differences -- Akee, Dahl & Lochner
  based on some exogenous change in variable of interest
- → sibling models

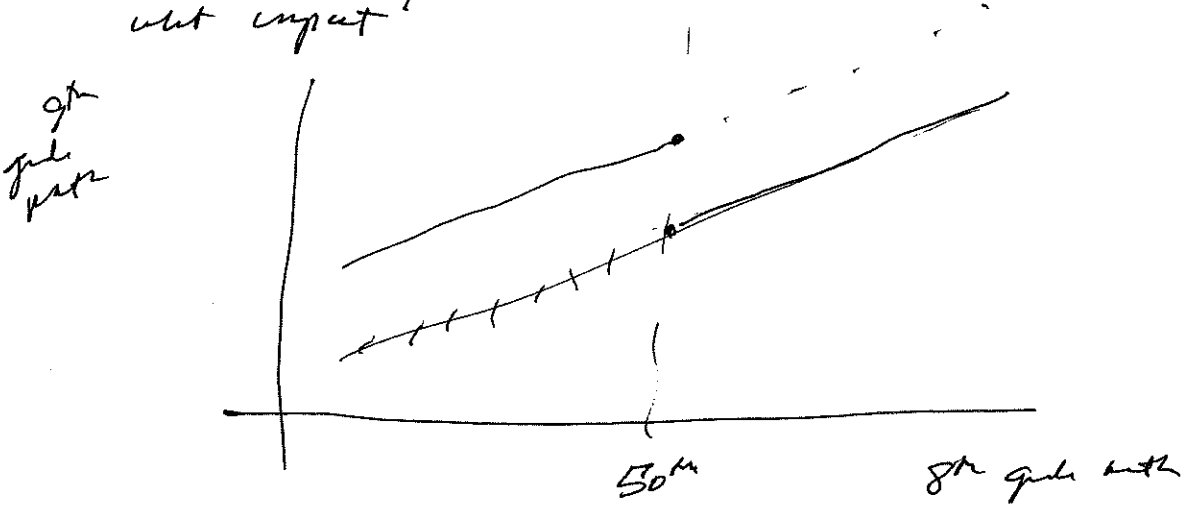
Week 6 is about regression discontinuity
- Did some arbitrary event disrupt what would normally be an unremarkable correlation?

e.g. Double dose algebra for ninth-grade student

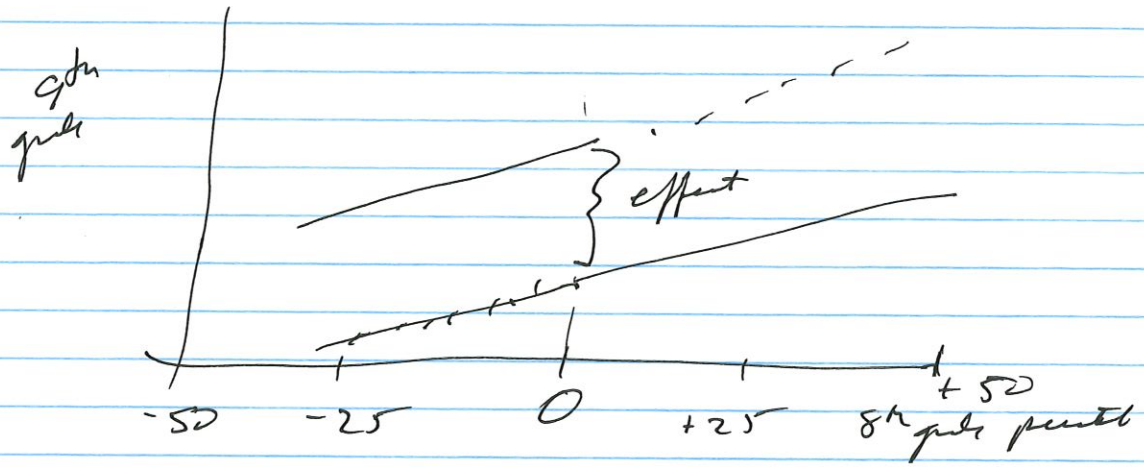
In the absence of double dose --



DD targets student below the 50th percentile in 8th grade
what impact?



How to model DD effects. First, center data on the 50th percentile



9^{th} grade math = $a + b_1$ 8th grade percentile
$+ b_2$ (where < 50th ~~b_3 (where 50th 8th grade percentile~~

For above 50th percentiles:

9^{th} grade = $a + b_1$ 8^{th} grade percentile

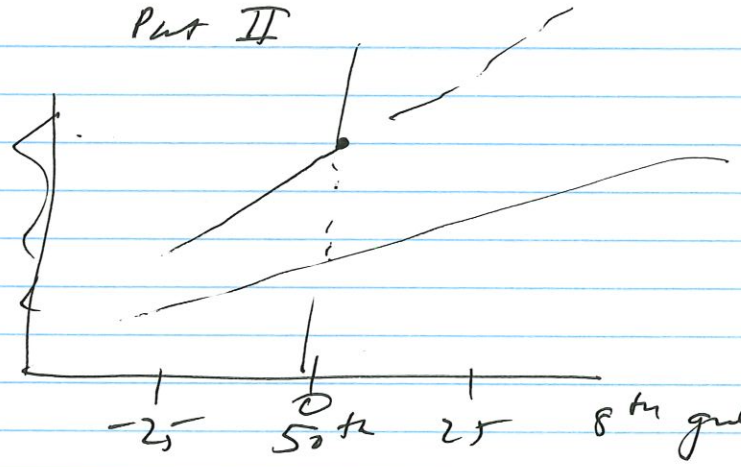
for below 50th percentiles:

9^{th} grade = $(a + b_2)$ + 8^{th} grade math

b_2 is the input

because intercept is evaluated at 50th percentile

5th class Part II



9^{th} math $= a + b_1\, 8^{th}$ grade $+ b_2$ indicator $<50^{th}$ $+ b_3$ (what $<50^{th}$ & 8^{th} grade)

for $>50^{th}$ percentile $\quad 9^{th} = a + b_1\, 8^{th}$

for <50 percentile $\quad 9^{th} = (a + b_2) + (b_1 + b_3)\, 8^{th}$ grade math

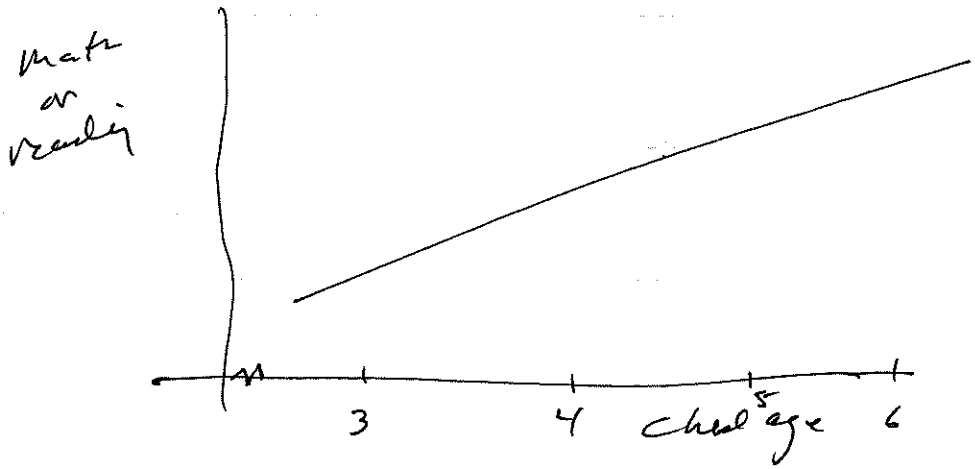
Allowing for diffit slopes enables you to better nail down the jump at the 50th percentile.

5th Class Part II

Week 6 is about regression discontinuity --

Did some intensity event disrupt what would normally be a ~~[illegible]~~ unremarkable correlation?

e.g. Preschool offer disrupts the relationship b/w child age and, say, math scores



Now offer preschool to ~~[illegible]~~ on 4th birthday and rescale