265

First Class Part I -- $\beta$ -1-

Intro to class - walk through syllabus

•

How to think about regression coeffiecients

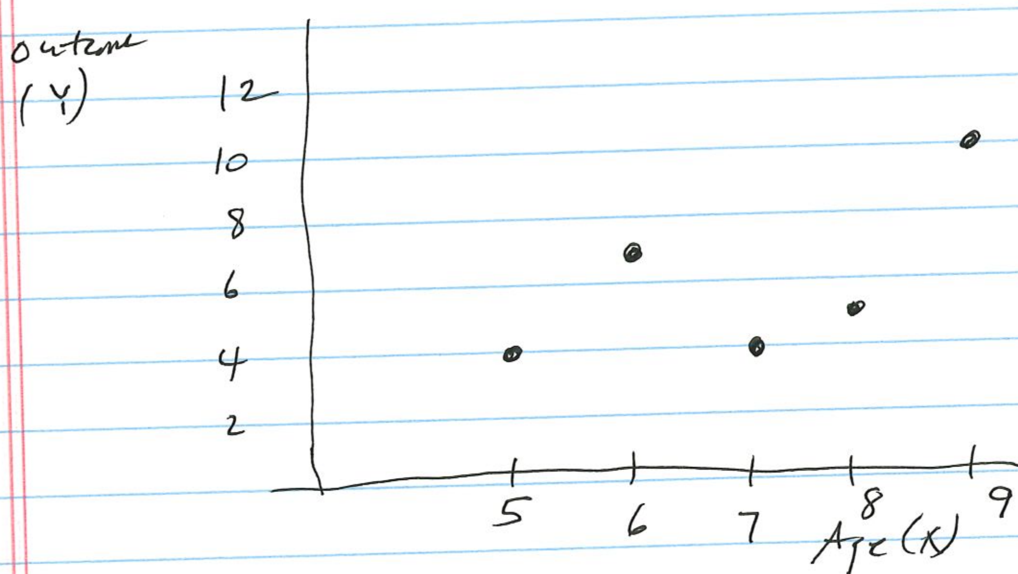
Data

| Age (X) | School outcome (Y) |
|---|---|
| 5 | 4 |
| 6 | 7 |
| 7 | 4 |
| 8 | 5 |
| 9 | 10 |

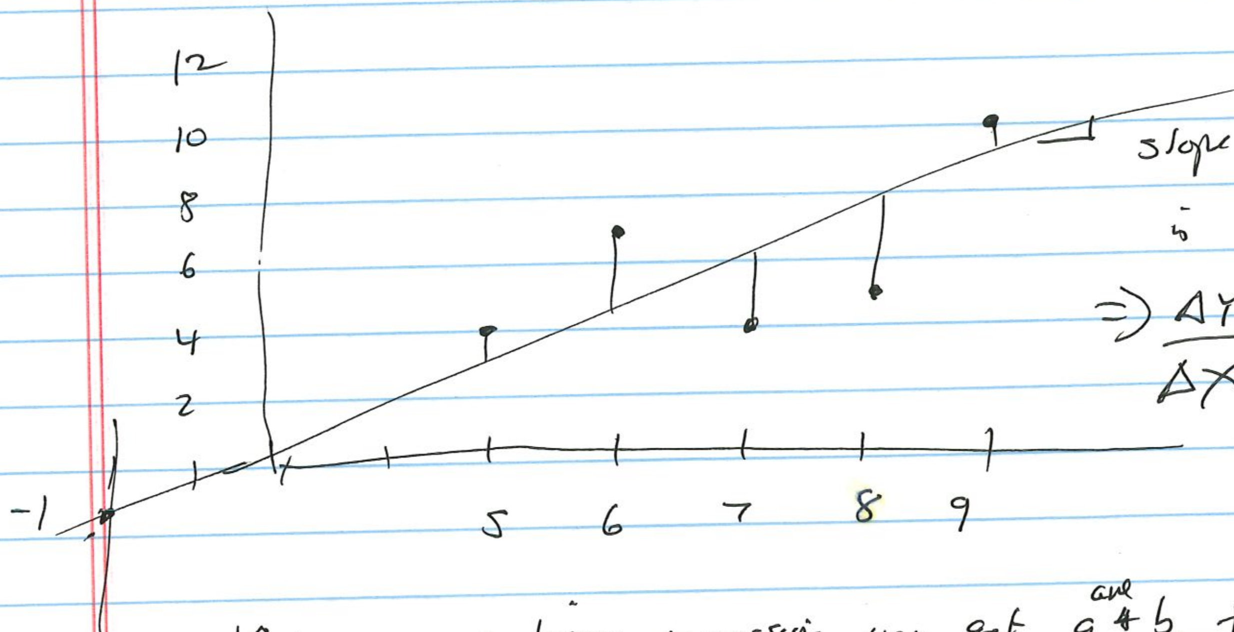
$\bar{x} = 7$
$\bar{y} = 6 \left(= 30/5\right)$

A ~~simple~~ bivariate regression tells you the "best fitting" linear relationship between X : Y

Plot it:




if I hand you a thin rod, how are you going to place it to best fit it to these data?




slope of regression is $b_1$

$\Rightarrow \frac{\Delta Y}{\Delta X}$

If you run a linear regression, you get $a$ and $b_1$ that best fits the data

$$Y = a + b_1 X$$

"Best fits" is based on a "least squares" criterion –

take all of the vertical distances, square them and make that as small as possible

It turns out that "best fitting" equation is

$a = -1$
$b = +1$

$\cancel{X = 0.8 + .8\, Age}$

$Y = -1 + 1 .\, Age$

how to interpret $\cancel{.8}$ 1.0 ? slope?

Each ~~add~~ additional unit of X is associated with a $\cancel{.8}$ 1.0 unit increase in Y

How do interpret the a? -1

$$Y = a + b_1 X$$

$\Rightarrow$ $a = Y$ when $X = 0$

$\Rightarrow$ predicted test score when age $= 0$!

The constant term (a) rarely tells you anything useful

All of our attention is focused on the slope coefficients

Slope coeffient

5 observations

Formula?

$$b = \frac{\sum_{i=1}^{5} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{5} (x_i - \bar{x})^2}$$

That doesn't help very much!

Apply it to our data. First off, "center" the data by subtracting $\bar{x}$ from $x_i$ s

| i | New X | Age (x) | Y |
|---|---|---|---|
| 1 | -2 | 5 | 4 |
| 2 | -1 | 6 | 7 |
| 3 | 0 | 7 | 4 |
| 4 | 1 | 8 | 5 |
| 5 | 2 | 9 | 10 |

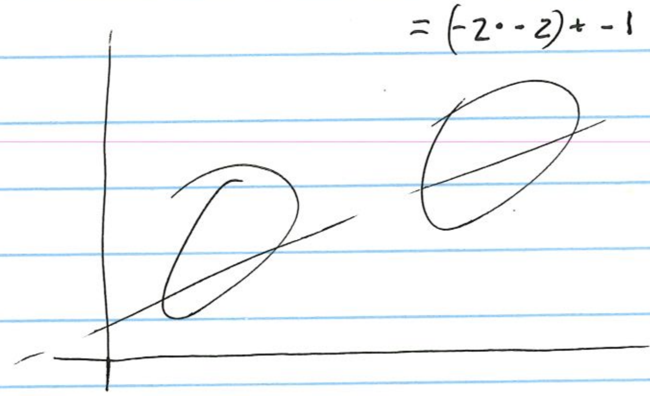
Let's first do the denominator:

$$\sum_{i=-2}^{2} (x_i - \bar{x})^2 = (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2$$
$$4 + 1 + 0 + 1 + 4$$
$$= 10$$

$\Rightarrow$ a number that is used to scale the numerator

So $b = \dfrac{\text{bunch of stuff}}{10}$

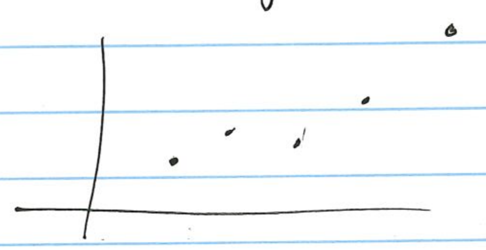
Numerator: $-2 \cdot (4-6) + (-1)(7-6) + 0(4-6)$
$(+1)(5-6) + 2(10-6)$
$= (-2 \cdot -2) + -1 - 1 + 8 = 10$



Really just a bunch of stuff on the right half minus a bunch of stuff on the left

⇒ a weighted average ~~of high~~

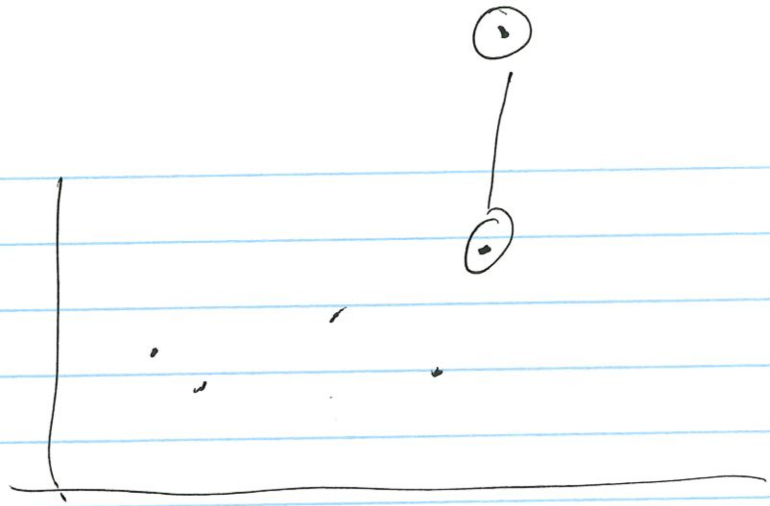
Note: more extreme Xs get more weight in determining the slope



⇒ why not just sample very high and very low X's

⇒ measure Y at ages 5 and 9 but not in between

⇒ worry about outliers associated with very large or very small X values.



huge influence

What about the influence of outliers at $\bar{X}$?

BREAKOUT ROOM

$\Rightarrow$ outliers at $\bar{X}$ have no effect on the slope

why measure anything at age 7 ??

So what is $b_1$ and $a$

$$\text{Numerator} = \overset{4}{(-2 \cdot -2)} \overset{-1}{- 1(7-6)} + \overset{-1}{1(-1)} + \overset{+8}{2 \cdot 4} = 10$$

Denominator $= 10$ slope $= 1$

$\Rightarrow y = a + X$

intercept

First class Part II - Table 3

Now lets take a look at Table 3

Regression:

$$Ed = a + b_1 \, Inc + b_2 \, Controls$$

in years

Have open:
1. Hurray 101
2. Table 3
3. Syllabus

Controls = child race and gender, race, maternal schooling, age of mother at birth, region, single-parent family structure, maternal employment.

$b_1$ is .14 (.02) $\Rightarrow$ a one unit change in Income is associated with a .14 unit change in education ($1/7^{th}$ of a year)

What is a one-unit change in income?

$\boxed{\$10,000}$!

So, all else constant, an additional \$10,000/year $\boxed{\text{for 15 years}}$ is associated with a $1/7^{th}$ of a year increase in schooling

$\Rightarrow$ small!!

Is it causal?

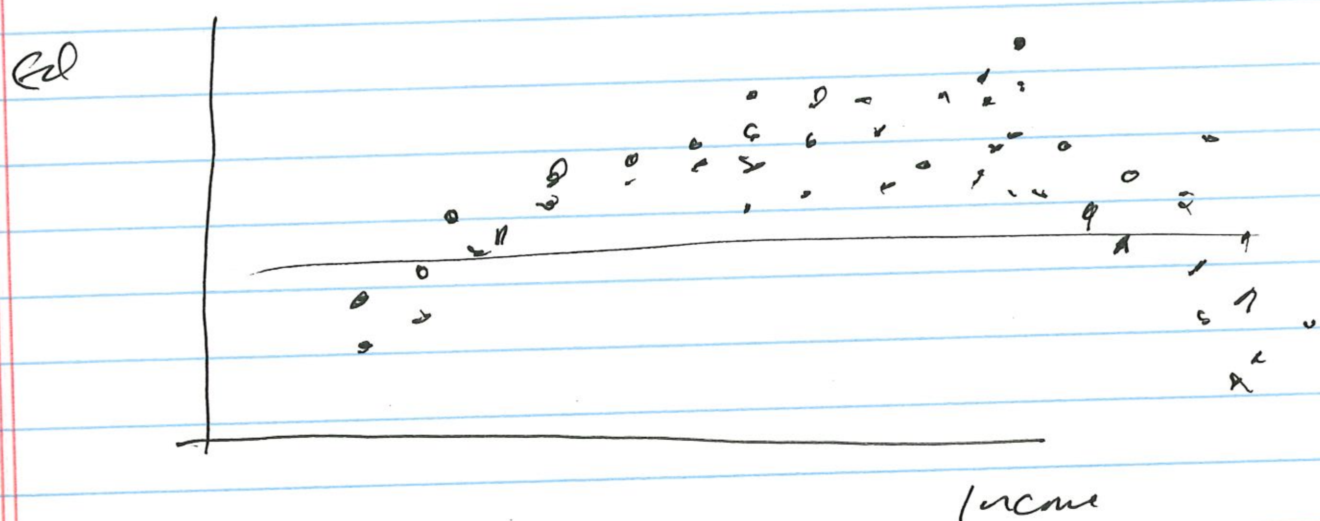
GREAT QUESTION!

(1) ~~ex~~ identification; (2) functional form

First class Part II Table 3

Don't stop with just $Y = a + bX$
and $b = .14$

What if non-linear relationship?




In this case, best fitting line is flat
$\Rightarrow$ look like no relationship

What fits better? A parabola

$$Y = a + b_1 \, Inc + b_2 \, Inc^2 + \text{Controls}$$

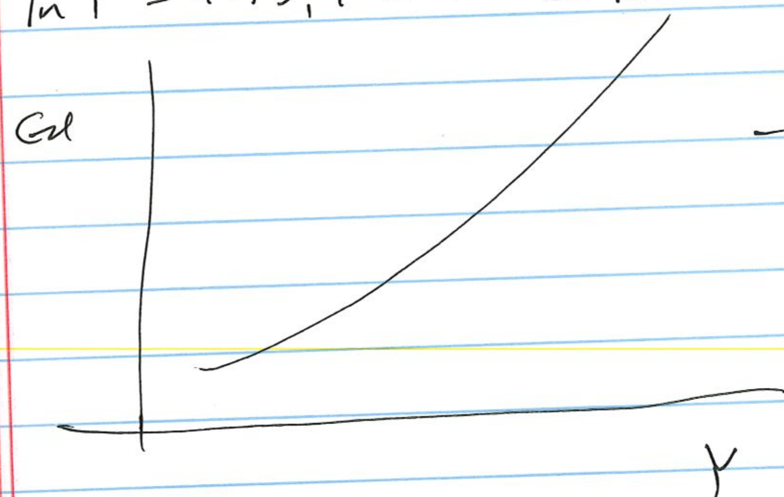
$1^{st}$ p II Table 3

In the article we do 3 things

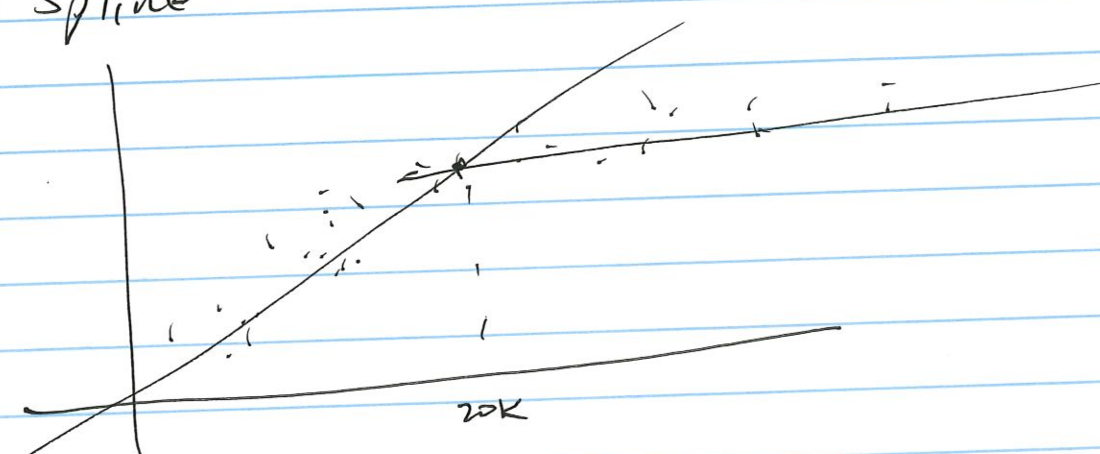
(1) Fit a log income function

$$Y = a + b_1 \log \text{income} + \text{controls} \downarrow$$

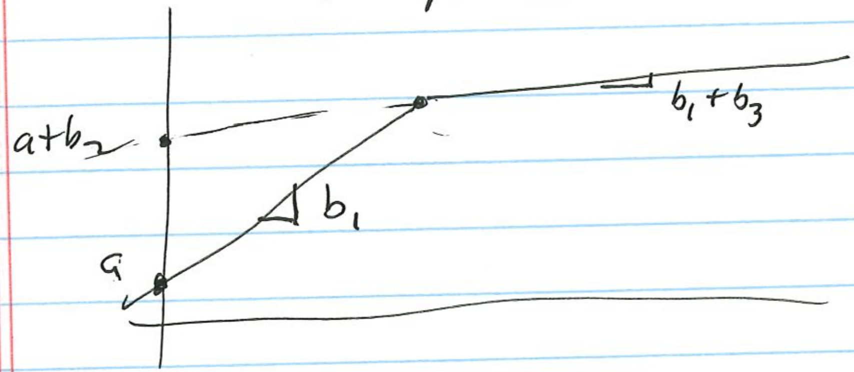

If $\ln Y = a + b_1 \text{income} + \text{controls}$




(2) Fit a spline



1ST part II Table 3




how to do this?

(1) $Y = a + b_1 \text{ Income} + b_2 (1,0) \text{ whether income} >^{\$} 20K$

$+ b_3 (1,0) \text{ income} >^{\$} 20K \cdot \text{Income}$

How to make sense out of this?

Suppose income is $< 20K$, then (1) gives you

$$\underbrace{Y = a + b_1 \text{ Income}} + 0 + 0$$

if income $> 20K$ then

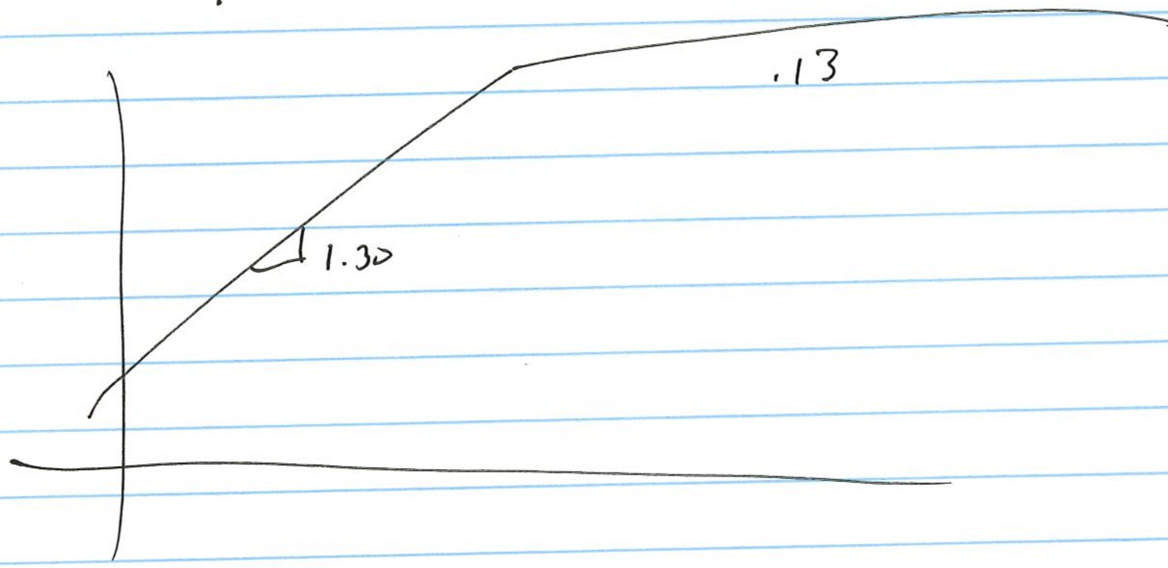
$$Y = a + b_1 \text{ Income} + b_2 \cdot 1 + b_3 \cdot \text{Income}$$

$$\Rightarrow Y = (a + b_2) + (b_1 + b_3) \text{ Income}$$

1st part II Table 3

look at Table $b_1 = 1.30$ ! much bigger than .14

$$b_1 + b_3 = 1.30 - 1.17 = .13$$



1ST Part II



What about ~~dummy~~ column 4?

$$Y = a + b_1\ 1{,}0\ \text{Inc}(15\text{–}25) + b_2\ 1{,}0\ \text{inc}(25\text{–}35)$$
$$+ b_3\ (1{,}0)\ \text{income}\ (35\text{–}50) + b_4\ (1{,}0)\ \text{inc}\ 50+ \ + \text{controls}$$

Hmm...

For someone with a 20K income:

$$Y = a + b_1 \cdot 1 + 0 + 0 + 0 + \text{controls}$$

$b_1$: .82

... 30K

$$Y = a + 0 + b_2 + 0 + 0 + \text{controls}$$

a + 1.41

a + 1.69

a + 2.35

Someone with a 10K income?

$$Y = a + 0 + 0 + 0 + 0 + \text{controls}$$

a is the ed level associated with very low income

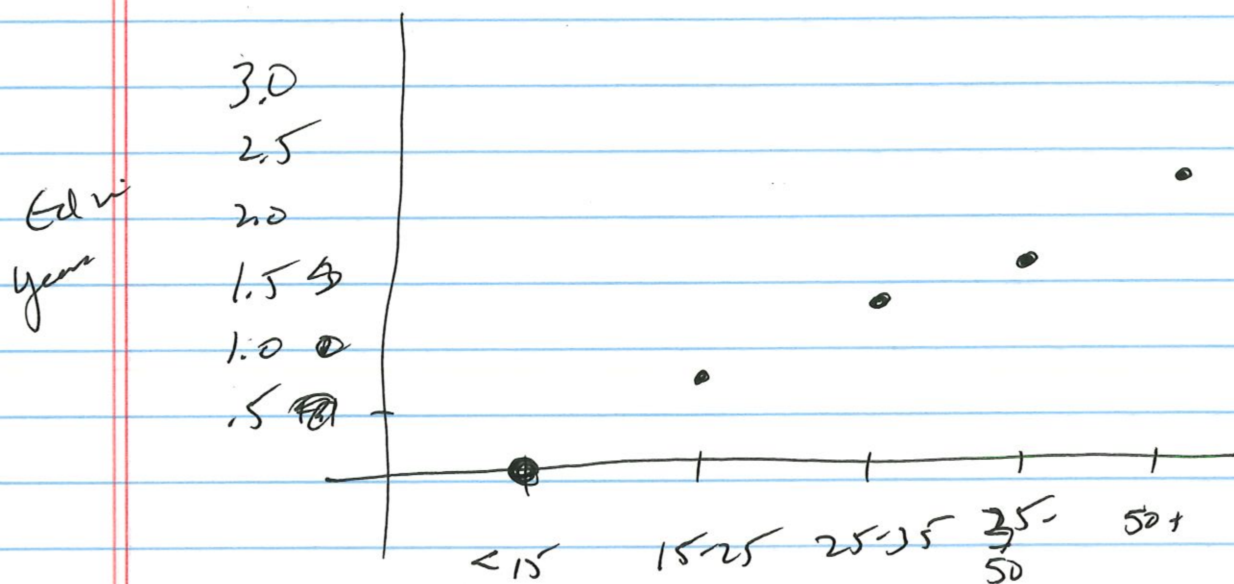
$b_1$ shows <u>difference</u> between 20K and 10K

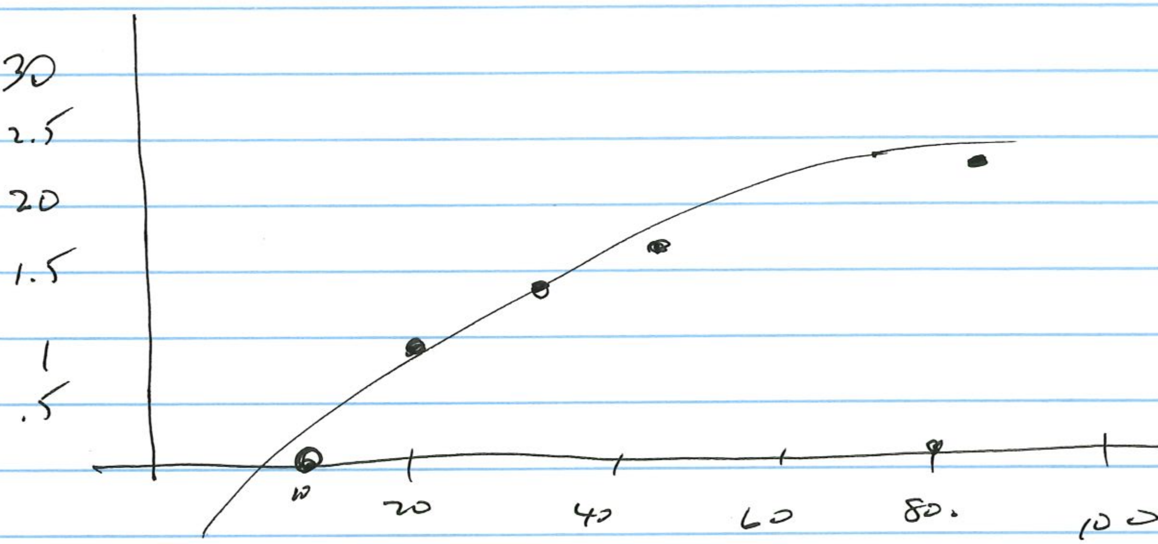
.82 years

$b_2$ shows difference between 30K and low income

1.41

1st Part II Table 3




most flexible

Other columns: (1,0) dependent variable → logit regression
1,0 dependent variable with censoring
event history