# ECON 300 – Econometrics

University of Illinois at Chicago
Summer session II 2017

Lecture 3 – Matching and regression

# Overview of today's lecture

- Discuss matching as a way to improve means comparisons.
- Review/summarize properties of covariance.
- Go over the basics of Ordinary Least Square regression using both a geometric and intuitive mathematical approach.
- Focus on the various elements of a simple, two-variable ("bivariate") regression model.
- Relate the workings of regression back to the matching idea we began with in the context of the Dale and Krueger study in MM.

# Matching I

Random assignment is the only econometric technique we will study that uses an explicitly random assignment of individuals to treatment.

When we cannot randomize, what we are trying to do is find situations where treatment assignment approximates randomization.

Matching on visible characteristics is a crude but sometimes effective way of doing this. We can do this “by hand” or with regression.

*Exact matching* is a method whereby we choose a list of visible characteristics (observables) that we think may be well correlated with things we cannot measure (unobservables), put people into matched groups according to observables, and compare them.

# Matching II

Following Mastering ‘Metrics (MM), say we are interested in the earnings premium from attending a private versus a public college. Say we only have gender, and high school GPA. Let’s group GPA into

- below 2.0
- 2.01-3.0
- 3.01-3.5
- 3.51-4.0+

How many group comparisons can we make?

What sorts of things might these two variables stand in for?

Is this comparison better / worse than an unconditional means comparison?

# Matching III

What sort of additional characteristics might help our comparison? What might they tell us?

- Family income
- Parental educational attainment
- Avg number of students in HS attending public/private schools

These covariates should help reduce bias in our estimates because they are both important characteristics on their own and they are likely correlated with other unobservable characteristics we think matter.

What are the two really important variables that the Dale & Krueger study in MM focuses on to make a *ceteris paribus* comparison?

# Matching IV

(go to next slide)

Table 2.1 - The college matching matrix

| Applicant group | Student | Private | | | Public | | | 1996 earnings |
|---|---|---|---|---|---|---|---|---|
| | | Ivy | Leafy | Smart | All State | Tall State | Altered State | |
| | 1 | | Reject | **Admit** | | Admit | | 110,000 |
| A | 2 | | Reject | **Admit** | | Admit | | 100,000 |
| | 3 | | Reject | Admit | | **Admit** | | 110,000 |
| B | 4 | **Admit** | | | Admit | | Admit | 60,000 |
| | 5 | Admit | | | Admit | | **Admit** | 30,000 |
| C | 6 | | **Admit** | | | | | 115,000 |
| | 7 | | **Admit** | | | | | 75,000 |
| D | 8 | Reject | | | **Admit** | Admit | | 90,000 |
| | 9 | Reject | | | Admit | **Admit** | | 60,000 |

* Enrollment decisions highlighted in gray.

# Matching V

Let’s make a naïve comparison from table 2.1 and see what it suggests about the earnings differential. (amounts in $1000s)

- Those who attended private: (110+100+60+115+75)/5=92
- Those who attended public: (110+30+90+60)/4=72.5
- This suggests the earnings premium is 92-72.5= $19,500/year.

# Matching VI

(go to table 2.1)

Now let's generate a comparison of earnings by admissions groupings and see how the estimated premium changes. Which groups do we compare?

- Group A contains students who were admitted to the same public and the same private school. Use them.
- Group B contains students admitted to the same private and the same two public schools. Use them.
- Group C contains students admitted to the same private school. There is no *public* "road not taken" for these students, so we won't use them. Why?
- Group D students were rejected from one private and admitted to the same two public schools. No *private* "road not taken" so we won't use them either.

# Matching VII

How do we generate a single number from the remaining students in groups A & B? Just add them up and average like we did before?

No. We will need to generate a *weighted average* that reflects the number of students contributing to each exact match comparison.

We have 3 students in group A and 2 students in group B so group A will contribute 3/5$^{ths}$ to the estimate and group B 1/5$^{th}$. What about *within* group A?

In group A there are two students attending private and one student attending public. The *private* students can be averaged and the single public student will be used alone. Here is the math…

# Matching VIII

- Within-group-A comparison: (110 + 100)/2 – 110 = -5
- Group B comparison: 60 – 30 = 30
- Weighted avg of two: 3/5(-5) + 2/5(30) = 9,000

This is less than half the size of the premium estimated by the naïve comparison.

# Matching IX

The key insight to the Dale and Krueger's 2002 study is that comparing earnings outcomes among those who applied to and were admitted to similar colleges can correct for most of the selection that biases the naïve comparison. Why is this?

- You must have thought it was a *potentially* good idea to go to a college if you applied to it.
- You must have demonstrated some level of *potential* ability to have been accepted. Colleges get much more data than we could and their "treatment assignment" reflects this detailed information

Once we condition on these factors, what kinds of things influence where a student actually goes?

# Matching X

The remaining factors are likely related to financial costs (tuition net of aid differences), family structure (those with tight family bonds might like to stay closer to home), the choices of peers (want to go to college with your peers, or maybe escape them!).

These decisions may not be quite “as good as randomly assigned”, but they are likely a lot closer to that ideal since the part of the college decision that is correlated with underlying academic ability and other such factors that influence earnings directly has been matched on.

Dale and Krueger implement their matching algorithm using regression. Let’s explore the basic concepts behind regression and we will return to their estimates afterwards.

# Covariance I

Recall our definition of variance (at the population level):

$$V(Y_i) = E[(Y_i - E[Y_i])^2] = E[(Y_i - E[Y_i])(Y_i - E[Y_i])].$$

When we are considering the relationship between two variables, we are interested not only in the dispersion, but the *nature* of the dispersion. Do the variables display a positive, negative, or zero relationship?

We can modify our variance formula to accommodate two variables.

$$Cov(Y_i, X_i) = E[(Y_i - E[Y_i])(X_i - E[X_i])]$$

The covariance relationship will be positive if the variables tend to move together, and negative if they tend to move in opposition.

# Covariance II

It may also help to consider the correlation coefficient (this is what we used in our data exploration in lab last week).

$$\frac{Cov(Y_i, X_i)}{S(Y_i)\, S(X_i)}$$

What is the denominator?

This is a normalized measure of covariance. It will always give a value on the interval [-1, 1] with -1 meaning perfect negative covariance, 1 meaning perfect positive covariance, and 0 meaning no covariance at all.

Let's work through why this is the case on the board.

# Covariance III

Covariance has three important properties we will make use of in thinking about regression.

1. The covariance of a variable with itself is its variance.

$$Cov(X_i, X_i) = E[(X_i - E[X_i])(X_i - E[X_i])] = E[(X_i - E[X_i])^2]$$

2. If expectation of $X_i$ or $Y_i$ =0, then $Cov(Y_i, X_i) = E[Y_i X_i]$. Let's work through this on the board.

3. The covariance between linear functions of $W_i = a + bX_i$ and $Z_i = c + dY_i$, (equations defining a line) where *a, b, c, d* are constants is given by

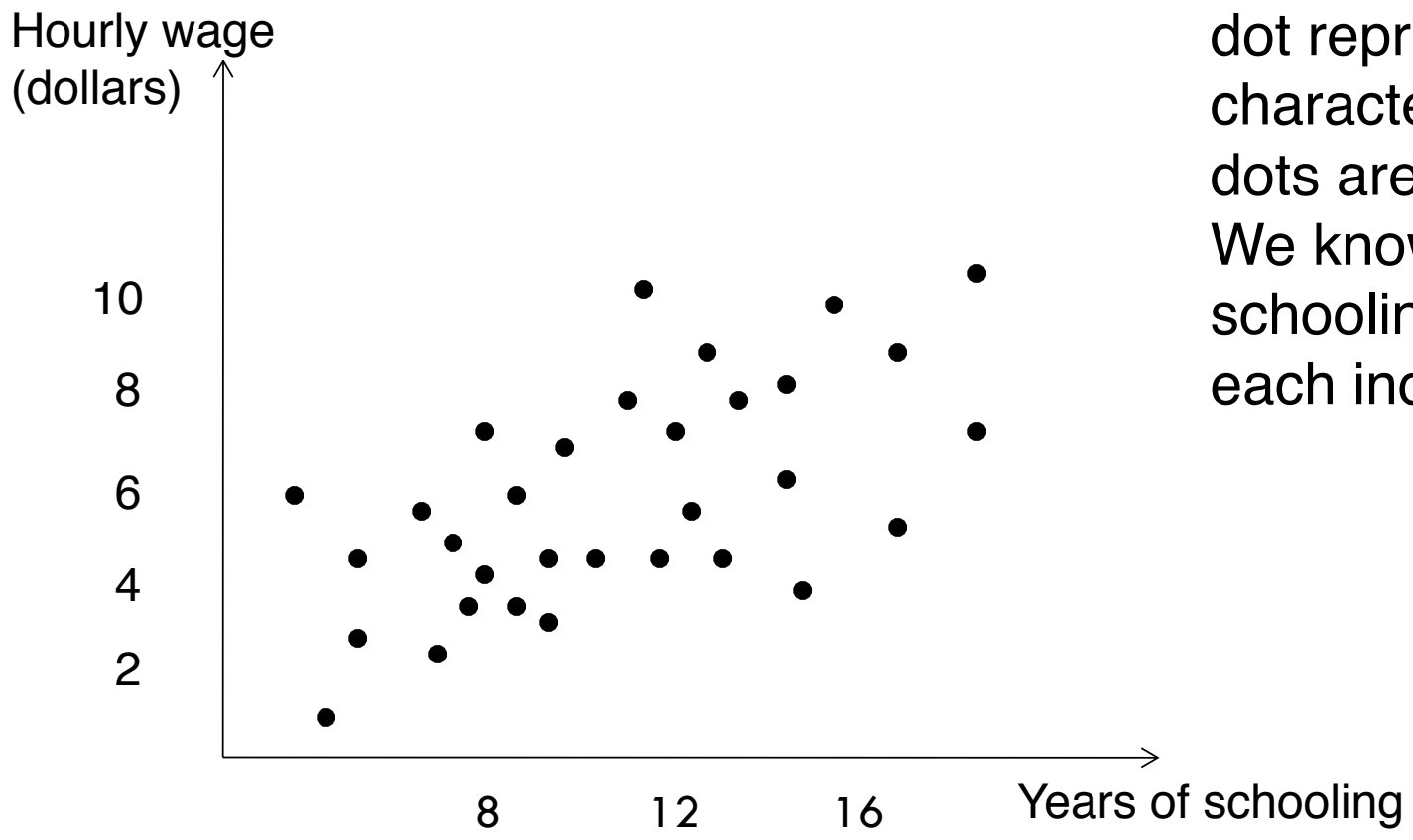
$$Cov(W_i, Z_i) = bd\ Cov(X_i, Y_i)$$

Let's go over this last property on the board too.

# Regression I

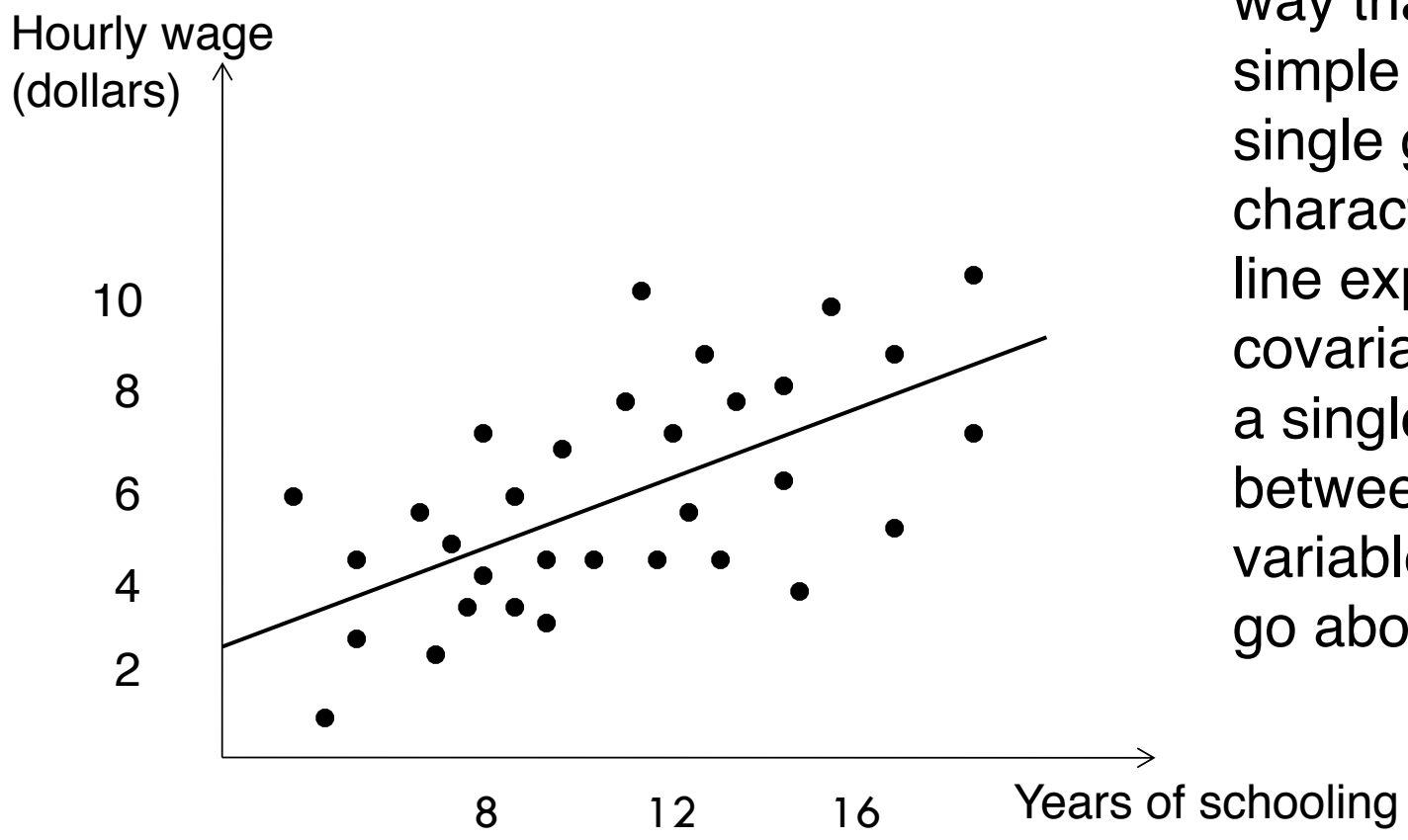
Let’s start our discussion of regression with a geometric approach. To examine a simple covariance relationship with two characteristics (variables), we can create a scatter plot of data where each observation has a pair of characteristics (x, y). Following basic Euclidean geometry, we can plot these two points on a graph.

# Regression II




In a scatter plot, each dot represents a pair of characteristics. Here, dots are individuals. We know years of schooling and wage for each individual.

# Regression III




To simplify this relationship in the same way that a mean is a simple expression of a single group characteristic, we can fit a line expressing the covariance relationship as a single number (a slope) between these two variables. How should we go about this?

# Regression IV

A line in two-dimensions assumes the form *y = a + bx*. What are *a* and *b*?

- **The intercept, *a*, gives the point at which our line crosses the y-axis** (the value of y when *x=0*).
- **The slope, *b*, gives the amount that *y* changes as you increase *x.***

The equation of a line is analogous to the *bivariate regression model,*

$$Y_i = \alpha + \beta X_i + e_i.$$

What is the additional term?

The term $e_i$ is the "error term" representing the part of outcomes that does not covary with *X.* In the mathematical regression we will "run", it represents the residual distance from whatever line we fit to the actual data point. We generically call $e_i$ "the residual."

# Regression V

The criterion we choose in fitting a line to the data is an *objective function*. This defines our goal in fitting a line through the data.

The objective function that regression analysis seeks to satisfy is to minimize the sum of squared residuals. This is an “optimization problem” in math terms.

We generate the value we want to minimize by fitting a line across the scatter plot, measuring the *vertical* distance remaining from our line to each data point, squaring these distances, and adding them all up. If we do this for every possible line, one of these values will be the lowest. This is the regression line we will choose.

Mathematically, since $Y_i = \alpha + \beta X_i + e_i \Longrightarrow e_i = Y_i - \alpha - \beta X_i$, our minimization problem is

$$\min RSS\,(\alpha, \beta) = \sum_{i=1}^{N} (Y_i - \alpha - \beta X_i)^2 .$$

# Regression VI

"RSS" is "residual sum of squares." This optimization problem,

$$\min RSS\,(\alpha,\beta) = \sum_{i=1}^{N} (Y_i - \alpha - \beta X_i)^2\,.$$

Has the following solutions:

$$\beta = \frac{Cov(Y_i, X_i)}{V(X_i)}$$

$$\alpha = E[Y_i] - \beta E[X_i]$$

In our regression figure on the board, what do $\alpha$ and $\beta$ represent?

# Regression VII

So we just need the following ingredients to find $\alpha$ and $\beta$:

- The covariance between $X_i$ and $Y_i$ .
- The variance of $X_i$ .
- The expectation (mean values) of $Y_i$ and $X_i$ .

With these ingredients, we could manually plot a regression line and the intercept.

I will walk through an example of this with you.

Be sure you can do it on your own.

# Regression VIII

**Regression interpretation example 1:**

Say a regression of the simple model,

$$wage = \alpha + \beta * educ + e_i,$$

has the following "ingredients":

- Cov(*wage, educ)* = .2
- *Var(educ)* = 2
- *E* [*wage*] = $12.00
- *E* [*educ*] = 12.

Let's solve this step-by-step.

$\beta$ = .2/12 **= .1**; $\alpha$ = $12 - .1*12 = $12.00 - $1.20 **= $10.80**

# Regression IX

**Regression interpretation example 2:**

Say a regression of the simple model,

$$health\ index = \alpha + \beta * \log(income) + e_i,$$

gives the following result:

$$health\ index = 2 + .1 * \log(income) + e_i$$

Suppose income is in $1000s of dollars. Interpret the results. What does the intercept term mean? The slope term? What is the predicted value at log income = 10 (this is equal to about $23,000)?

The regression results predict that those with log income equal to 10 will, on average, have a health index of 3. This is the result of $\alpha + \beta * 10$.

Hint: Become *very comfortable* with these two exercises.

# Regression X

Let's focus on the bivariate regression formula,

$$\beta = \frac{Cov(Y_i, X_i)}{V(X_i)}.$$

What is $\beta$ equal to if $Cov(Y_i, X_i)$=0?

What happens to $\beta$ if $Cov(Y_i, X_i)$ is large in magnitude and overall $V(X_i)$ is small in magnitude?

# Regression XI

A few terms that you will encounter in discussing regression:

Both our “treatment” variable (the one we are especially interested in) and other control variables (things we want to make sure we match people on) are often referred to as *regressors* or *independent variables* (in using regression, when we put things on the right hand side of the regression model—make them *Xs*—we *assume* that they do not depend on *Y*).

Our “outcome” variable will also be referred to as the *regressand* or the *dependent variable* (where we are making the related assumption, that *Y depends on X*).

In a regression involving a dependent variable *Y* and an independent variable *X*, we say that we “*regress Y on X.”* (dependant var comes first)

# Regression XII

Regression fits a line across data. This line plots the relationship between *X* and *Y* for each value of *X*.

If you generate a regression line then take the *X* values in your data set and put them back into the regression function you estimated, they will give you the corresponding *Y* values on the line. Let me show you this graphically.

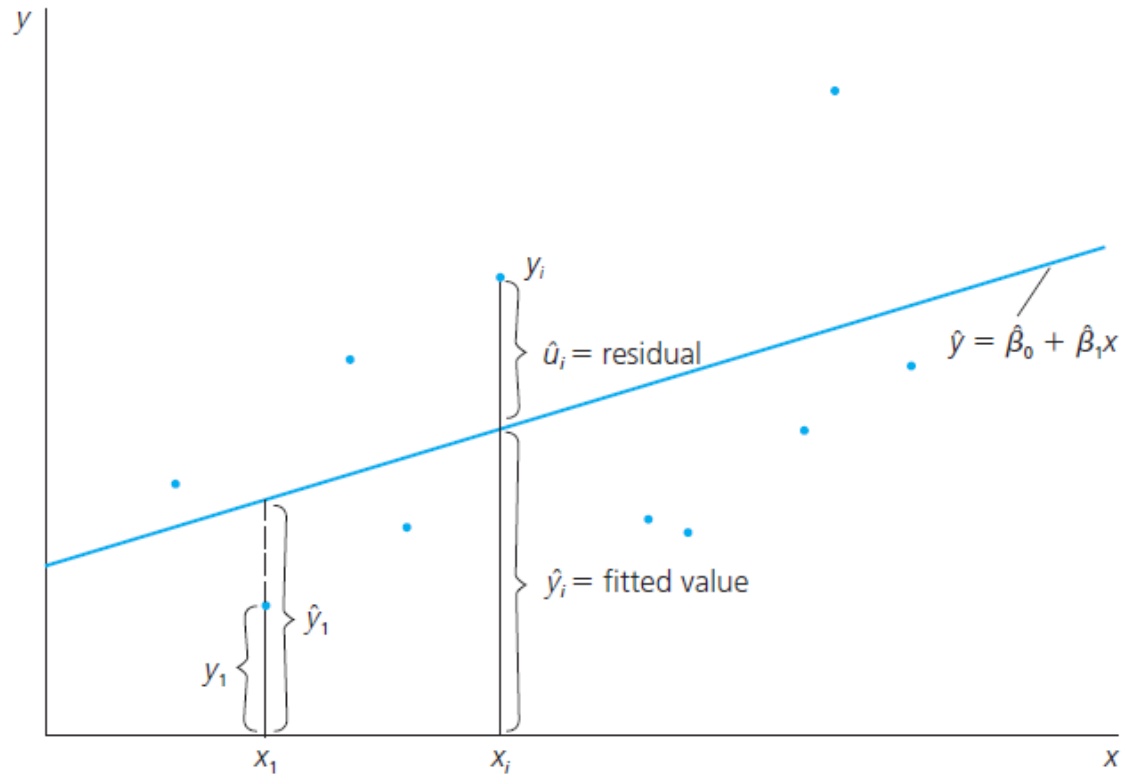
These values that the regression function "gives back" are estimated *Y* values. We call these values $\hat{Y}$ ("Y hat", also *predicted values*).

Each $Y_i$ in our data can be represented as $\hat{Y}_i + e_i$. That is, we can represent each $Y_i$ value as the value our fitted line predicts plus the residual distance from the point on our regression line (the discrepancy between what we actually observe for person *i* and her predicted value).

# Regression XIII

FIGURE 2.4

Fitted values and residuals.



# Regression XIV

Our predicted values, $\hat{Y}_i$, are equal to $\alpha + \beta X_i$.

Since we also know that $Y_i = \hat{Y}_i + e_i$,

$$e_i = Y_i - \hat{Y}_i = Y_i - (\alpha + \beta X_i).$$

The solution to the regression problem guarantees that the residuals are completely uncorrelated with the fitted values (if you ever take linear algebra, you will probably have to prove this "orthogonality" property).

But for our purposes, you need only know *three key properties of regression residuals.*

# Regression XV

Regression residuals

1) have both population and sample mean equal to zero,

$$E[e_i] = \frac{1}{n}\sum_{i=1}^{n} e_i = 0\,,$$

2) are uncorrelated in both sample and population with all of the regressors used to calculate them (that's why they are "residual"),

$$E[X_i e_i] = \frac{1}{n}\sum_{i=1}^{n} X_i e_i = 0,$$

3) are uncorrelated with the fitted values (b/c residuals are the "leftovers"),

$$E\left[\hat{Y}_i e_i\right] = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i e_i = 0.$$

# Conditional expectation function I

As we discussed in lecture 2, conditional expectation is the value of the population mean conditional on a set of characteristics. Regression estimates the *conditional expectation function (CEF)* or a best *linear* approximation to it.

For a regression that considers the mean values of *Y* as a function of some treatment *D* with only two values, 1 and 0, regression "finds" the *CEF.*

$$E[Y_i|D_i] = \alpha + \beta D_i$$

What happened to the residual?

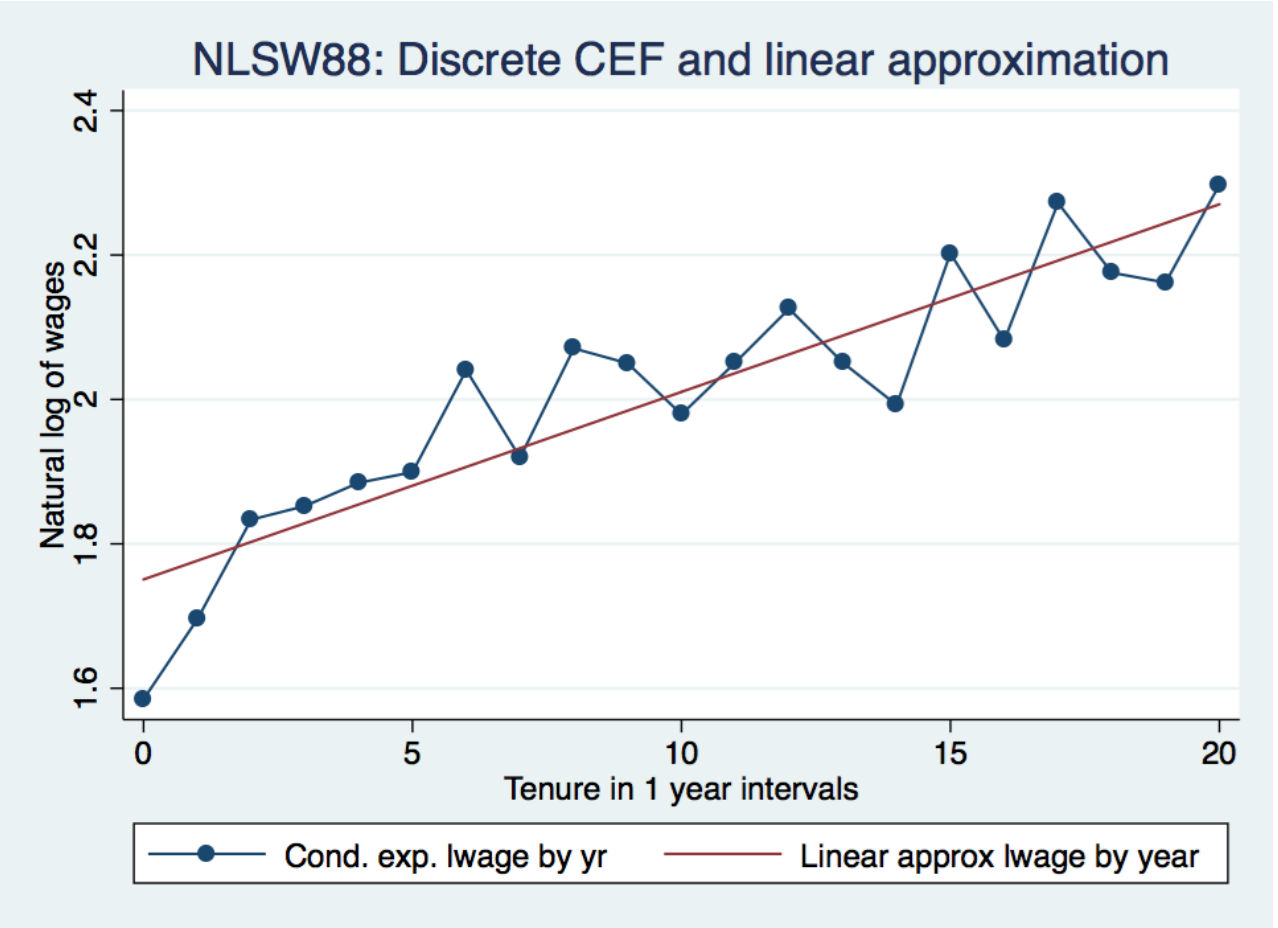
The CEF for individual *i* equals $\hat{Y}_i$, so $Y_i = \hat{Y}_i + e_i = E[Y_i|D_i] + e_i$.

From this regression, we get that $E[Y_i|D_i = 1] = \alpha + \beta$ and $E[Y_i|D_i = 0] = \alpha$

# Conditional expectation function II

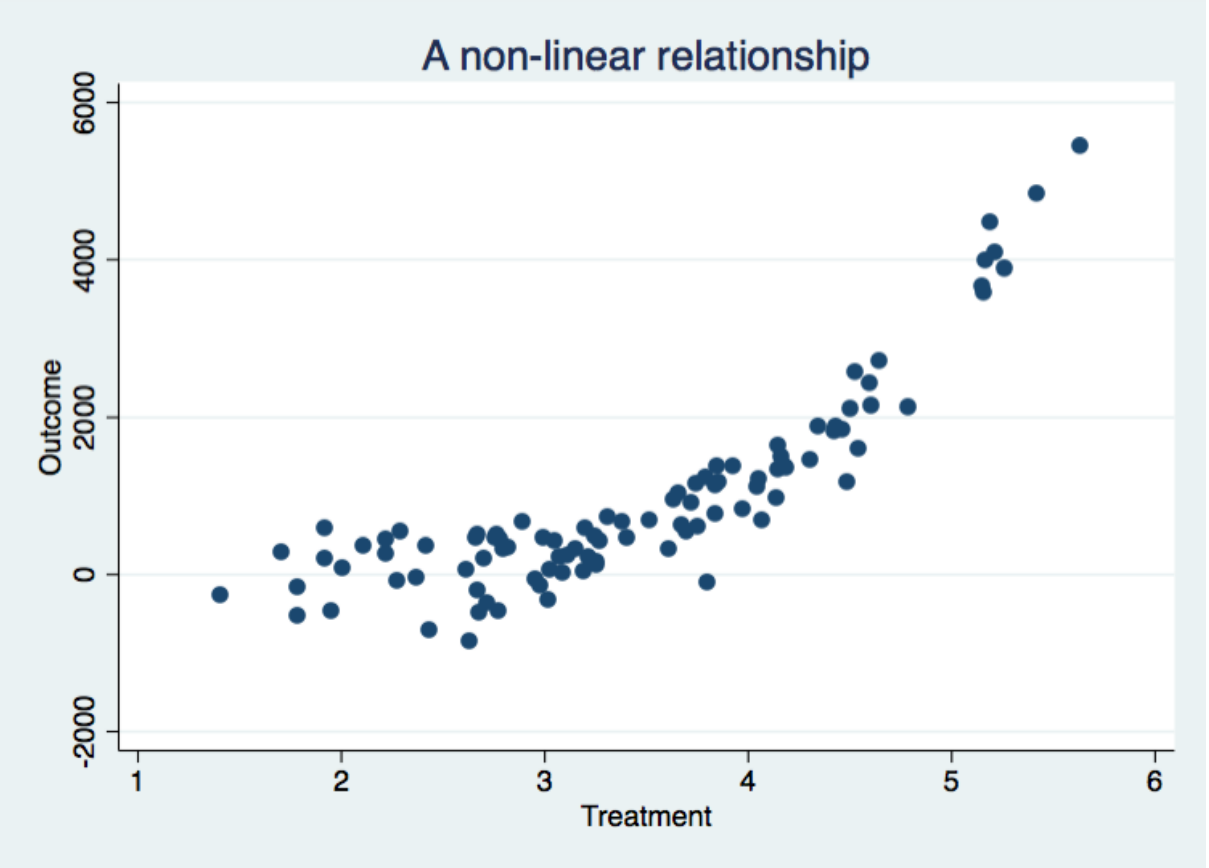
If a regressor has discrete values, then regression exactly plots the CEF for the relationship between the regressor and the regressand.

This graph plots both the exact CEF for the natural logarithm of wages among a group of workers by years of tenure on the job, and a linear approximation to it.



# Conditional expectation function III

Now consider the relationship in this scatter plot. Does the relationship look linear?



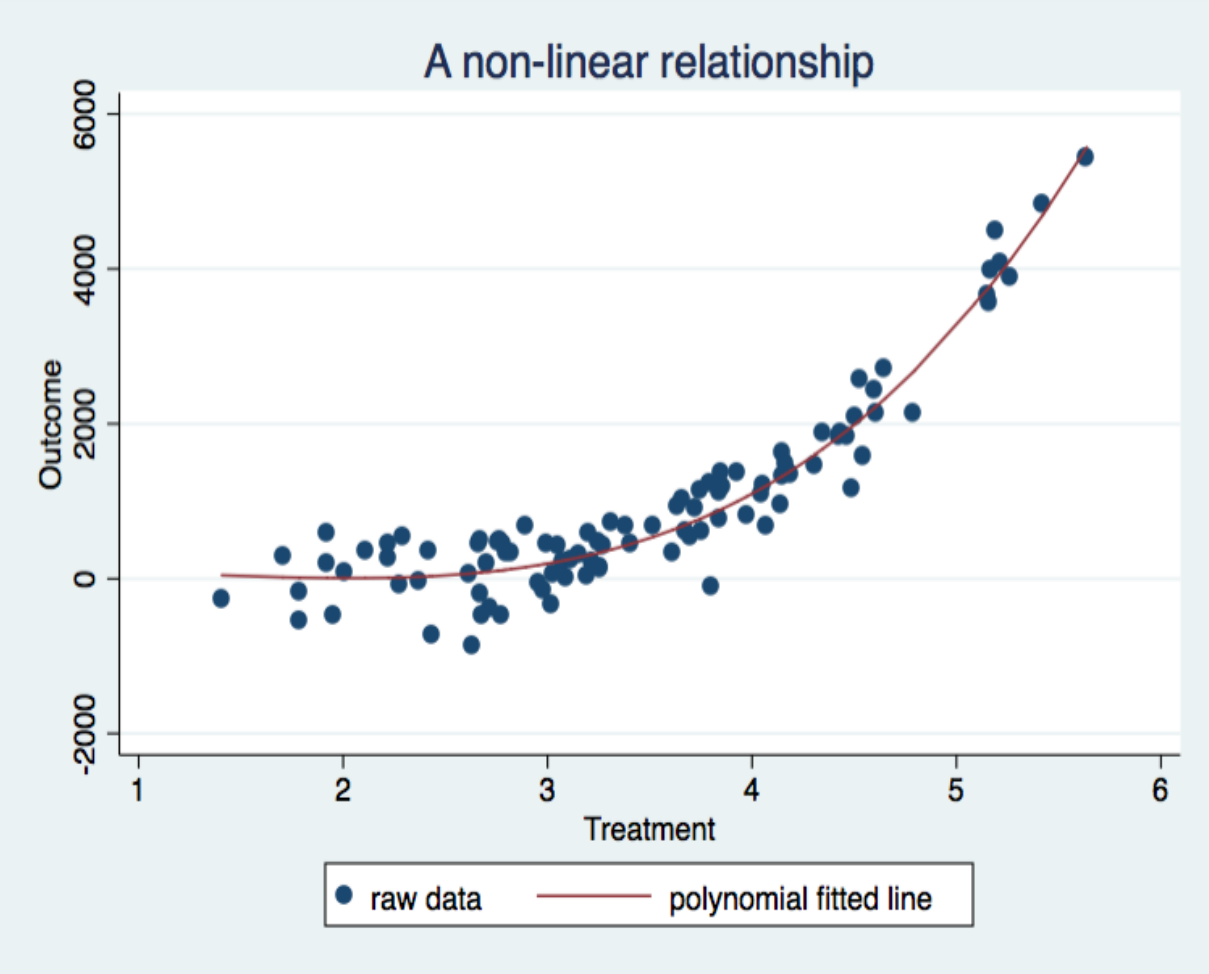
# Conditional expectation function IV

Now consider the relationship in this scatter plot. Does the relationship look linear?

No. Let's fit a polynomial line to it (a line that can curve freely to fit the data accurately).

There is a lot of curvature to this relationship suggesting that treatment has more of an exponential relationship with the outcome.

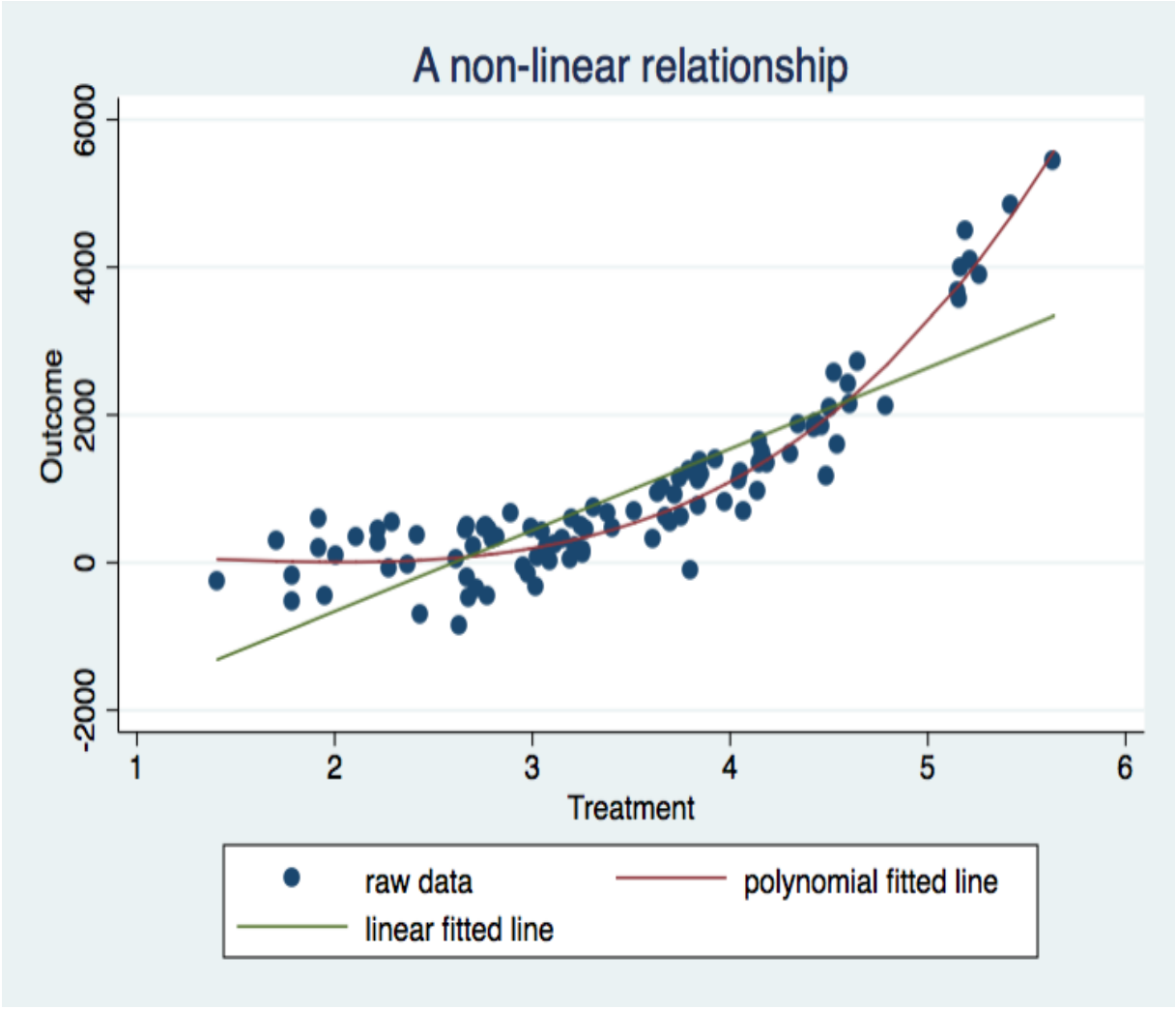
What will linear regression do?



# Conditional expectation function V

It will approximate this relationship as closely as possible with a line with no curvature.

This is what it means when our text says that when the CEF is not linear, regression “finds a good approximation to it.” (p. 85)



# Dale and Krueger (2002) I

Exact matching tends to throw away a lot of observations that don't match exactly and this problem grows as you add regressors.

Think of a group with 2 states (treated, non-treated), then think of a second group with two states (male, female). Now you have 4 groups.

Now think of a third group with two states (employed, unemployed). Now you have 8 groups.

If another variable has four states, then we go to 2 x 2 x 2 x 4 groups = 32.

Regression, however, can treat variables you might like to match on as a linear set of values, allowing it to "match" on observations that don't exist by imputing a value from nearby observations.

# Dale and Krueger (2002) II

Suppose a simplified Dale and Krueger study only matched on two criteria, the selectivity of colleges a student applied to and was admitted to, and SAT score (0, 1600) that we group into 100 point “bins” (so there are 16 of them).

Suppose one person who was admitted to the same two types of colleges (say highly competitive and most competitive) had an SAT score of 1400 and chose a public college.

Suppose two other people who were admitted to the same school groups had SAT scores of 1300 and 1500, respectively, and chose a private college.

Exact matching would not compare these people as they are not in the same SAT score “bins.” Regression *would* compare them, using an average of the two private college students with scores on either side of the public college student’s score.

# Dale and Krueger (2002) III

D&K use a multi-variate (multiple independent variables) regression model.

Their main "matching" strategy is to group people into 151 bins of selectivity for schools they were admitted to (each bin includes both public and private schools). They also include *controls* for SAT score, log of parental income, as they want to make a comparison *holding these factors fixed*.

The "treatment" they are interested in is private school attendance *conditional on school selectivity, SAT score, and parental income.*

$$\ln Y_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i + e_i$$

* Our text leaves out some additional demographic controls for the sake of simplicity (race, gender, student athlete, being in top 10% of your HS).

# Dale and Krueger (2002) IV

They compare this to a regression model without the selectivity groups. It is the estimates generated by this type of regression that they suspect are biased, so they include these results to compare with their model.

$$\ln Y_i = \theta + \rho P_i + \pi_1 SAT_i + \pi_2 \ln PI_i + e_i$$

This is often referred to as a “baseline” model and many papers include one.

Don’t be put off by the different Greek characters. When one study uses two different models, we don’t re-use the parameter symbols to avoid confusion about, for instance, which $\beta$ we are talking about.

What we want to do is compare $\beta$ from the previous slide with $\rho$ above.

# Dale and Krueger (2002) V

Table 2.2 (excerpt) - Private school effects: Barron's matches

| | No selection controls | | | Selection controls | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Private school | .135 | .095 | .086 | .007 | .003 | .013 |
| | (.055) | (.052) | (.034) | (.038) | (.039) | (.025) |
| Own SAT score/100 | | .048 | .016 | | .033 | .001 |
| | | (.009) | (.007) | | (.007) | (.007) |
| Log parental income | | | .219 | | | .190 |
| | | | (.022) | | | (.023) |
| Selectivity dummies | No | No | No | Yes | Yes | Yes |

# Dale and Krueger (2002) VII

The moral of the story of the D&K study is that even with fairly detailed controls for personal and family characteristics, estimates that lacked the controls for applying to and being admitted to similar types of schools suffered from bias.

The bias that results from the absence of important explanatory variables is

***OMITTED VARIABLE BIAS (OVB)***

and we will spend Thursday's lecture learning about properties of it. This is probably the single most important concept in the course.

# Identification strategies and regression I

- An “identification strategy” is a set of assumptions about the nature of the problem you are considering that allow you to use regression to identify causal effects.
- The D&K study uses a “selection-on-observables” identification strategy.
- This strategy uses an assumption that some observable variable (their “selection controls”) is sufficiently correlated with unobservable characteristics that would otherwise introduce bias into estimates of the effect of interest (the earnings return to private colleges) that using this control allows for a regression model to identify causal effects.
- Most of what they do in the paper is show many pieces of evidence supporting their hypothesis that they have such a variable.

# Identification strategies and regression II

- Everything else we are going to do in class is learning about additional identification strategies and what assumptions are necessary to try to use them for causal estimation.
- Regression is NOT CAUSAL without some identification strategy to support what you are using it for.
- **Without an identification strategy, regression is only comparing means and covariance, ways of describing numerical relationships.**

# Regression odds and ends I

- Every regression generates a statistic called $R^2$ ("R squared"). This is a ratio of the variation in *Y* that is correlated with variation in the included *X* variables over the total variation in *Y.*
- $R^2$ was a significant focus in more traditional econometrics courses, but this focus has decreased with a shift to interest in causality—as opposed to "explaining" as much variation as possible with many controls—since good causal research designs often involve regressions that have a very low $R^2$.
- $R^2$ is reported in every regression you will run in Stata and can still be a useful measure to allow you to think about how powerful the overall relationship between your *Y* and *X*s are.

# Regression odds and ends II

- By default, Stata calculates standard errors that are often biased because they use an assumption about the distribution of variables that is commonly not satisfied in the real world (”homoskedasticity”). In the past, you may have spent a few weeks on this issue in econometrics class.
- In the real world, most variables are “heteroskedastic,” meaning their variance changes with their level (rather than remaining constant).
- To compensate for this, always use the option “, robust” at the end of your regressions (type a comma then “robust”). This will approximately solve this problem for our purposes.
- There are other important situations where you need to use different sorts of SE estimators, but these are beyond the scope of the course. I can provide more info for those who may be interested.

# Matching / Regression takeaways

- Exact matching creates “bins” of people with identical characteristics and compares these groups (by forming a weighted average of comparisons).
- Regression efficiently automates matching.
- The bivariate regression estimator is $\beta = \frac{Cov(Y_i, X_i)}{V(X_i)}$.
- Regression minimizes the sum of squared residuals.
- Regression estimates the *conditional expectation function* exactly if variables all have discrete values, or estimates the best linear approximation to the *CEF* otherwise.
- Be able to construct and interpret regression coefficients in bivariate and multivariate regressions.
- Regression is a mathematical tool for measuring correlation. It only estimates causal relationships when coupled with the assumptions of an identification strategy (that should be supported with as much evidence as possible)!

# For Thursday

- Required reading: Mastering ‘Metrics pp. 68-78.
- Optional reading: Chapter 2 in Introductory Econometrics
- Watch this short Stata tutorial on basic regression: https://www.youtube.com/watch?v=HafqFSB9x70 (The host uses a bunch of dropdown menu commands you should not waste time using, but he discusses how regression results look and some relevant data visualizations that can be helpful to think about when running regressions. I will put this link up on BB too.

# Multivariate regression figures (go back)