binscatter: Binned Scatterplots in Stata

Michael Stepner

MIT

August 1, 2014

Motivation

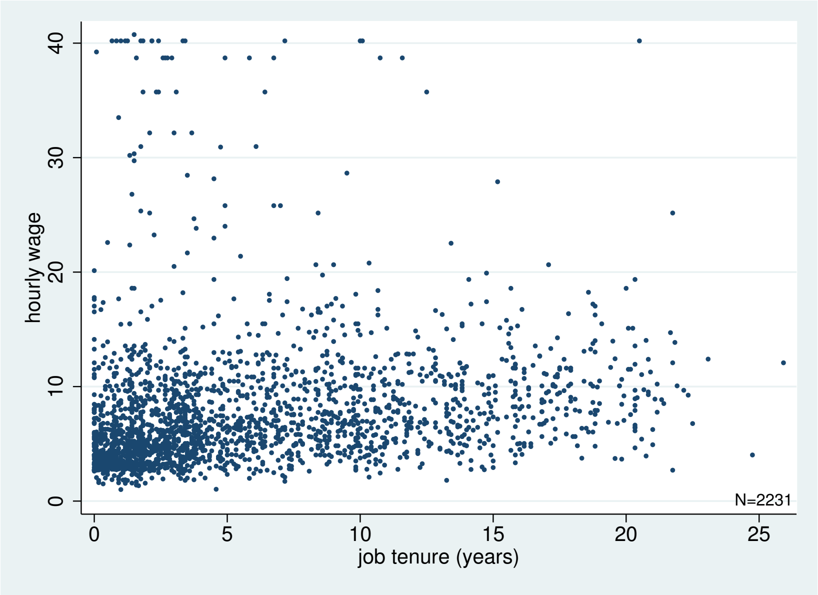
- Binned scatterplots are an informative and versatile way of visualizing relationships between variables
- They are useful for:
  - Exploring your data
  - Communicating your results
- Intimately related to regression
  - Any coefficient of interest from an OLS regression can be visualized with a binned scatterplot
- Can graphically depict modern identification strategies
  - RD, RK, event studies

Familiar Ground

Scatter Plots

Scatterplots:

- Are the most basic way of visually representing the relationship between two variables
- Show every data point
- Become crowded when you have lots of observations
  - Very informative in small samples
  - Not so useful with big datasets



Source: National Longitudinal Survey of Women 1988 (nlsw88)

OLS Regression

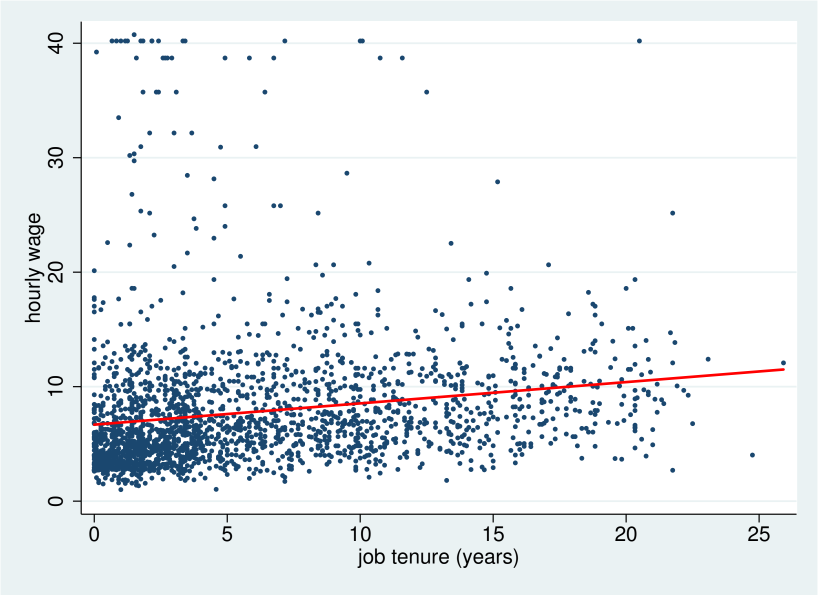
Linear regression:

- Gives a number (coefficient) that describes the observed association
  - "On average, 1 extra year of job tenure is associated with an \$m higher wage"
- Gives us a framework for inference about the relationship (statistical significance, confidence intervals, etc.)

```
. reg wage tenure

      Source |       SS       df       MS              Number of obs =    2231
-------------+------------------------------           F(  1,  2229) =   72.66
       Model |  2339.38077     1  2339.38077           Prob > F      =  0.0000
    Residual |  71762.4469  2229  32.1949066           R-squared     =  0.0316
-------------+------------------------------           Adj R-squared =  0.0311
       Total |  74101.8276  2230  33.2295191           Root MSE      =  5.6741

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |   .1858747   .0218054     8.52   0.000     .1431138    .2286357
       _cons |   6.681316   .1772615    37.69   0.000     6.333702    7.028931
------------------------------------------------------------------------------
```

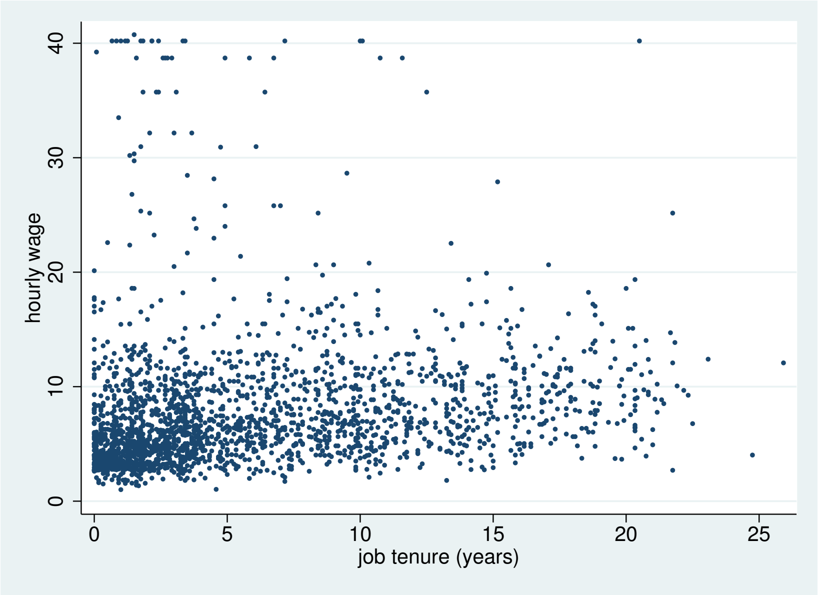

hourly wage
0
10
20
30
40
0
5
10
15
20
25
job tenure (years)

binscatter: step-by-step introduction

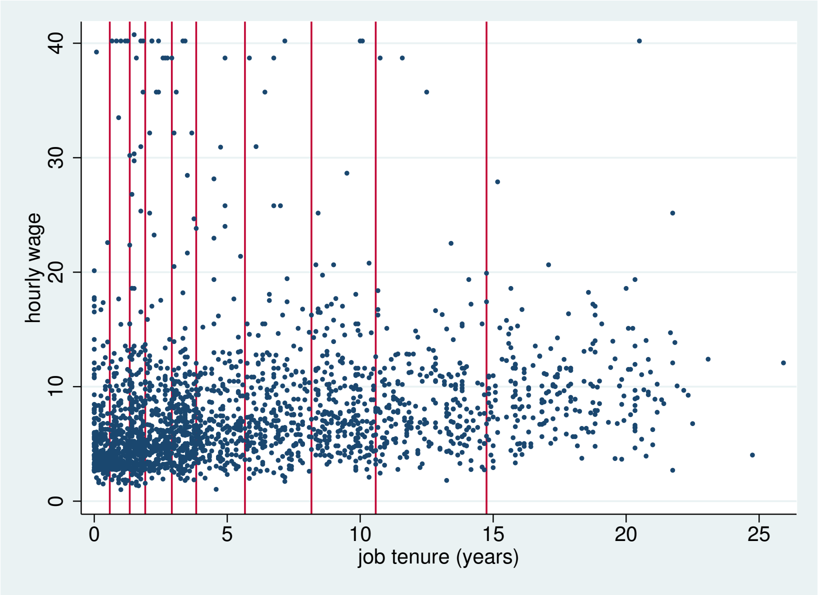
binscatter: step-by-step introduction

Let's walk through what happens when you type:

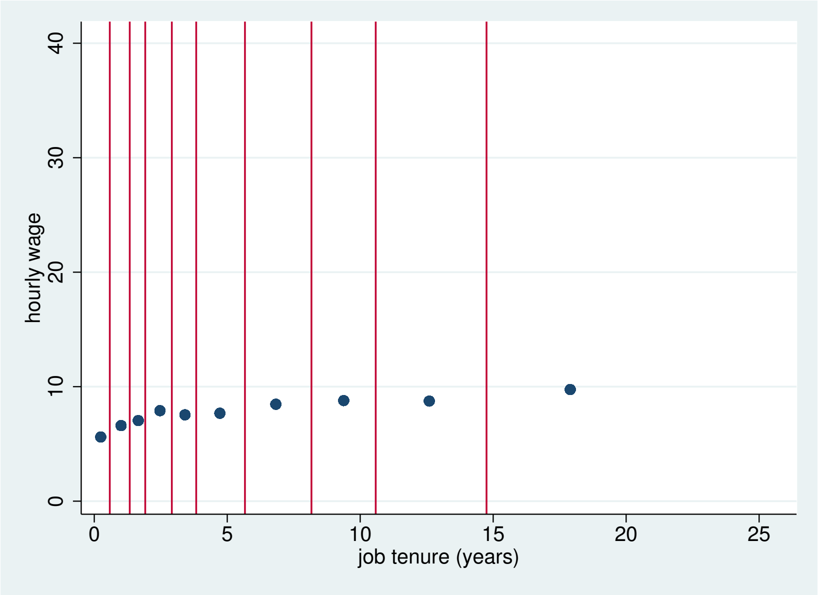
```
.  binscatter wage tenure
```

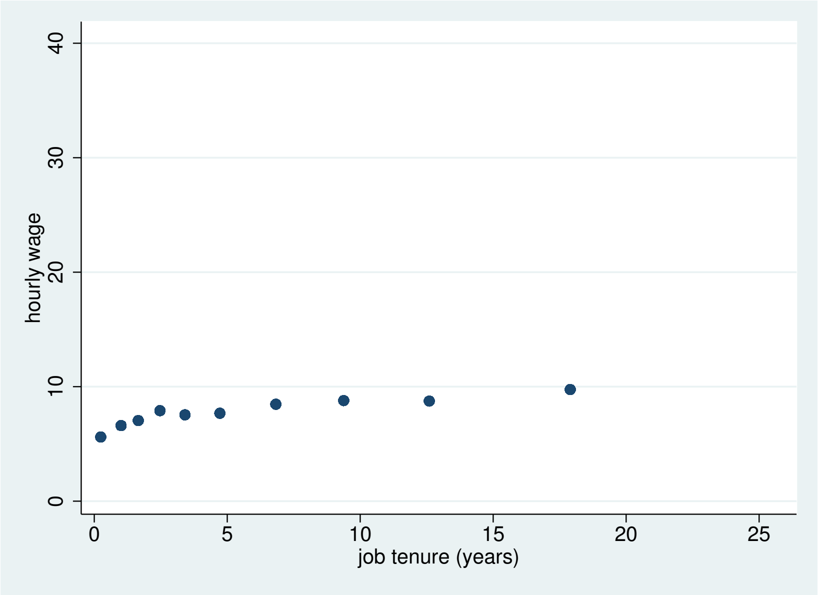


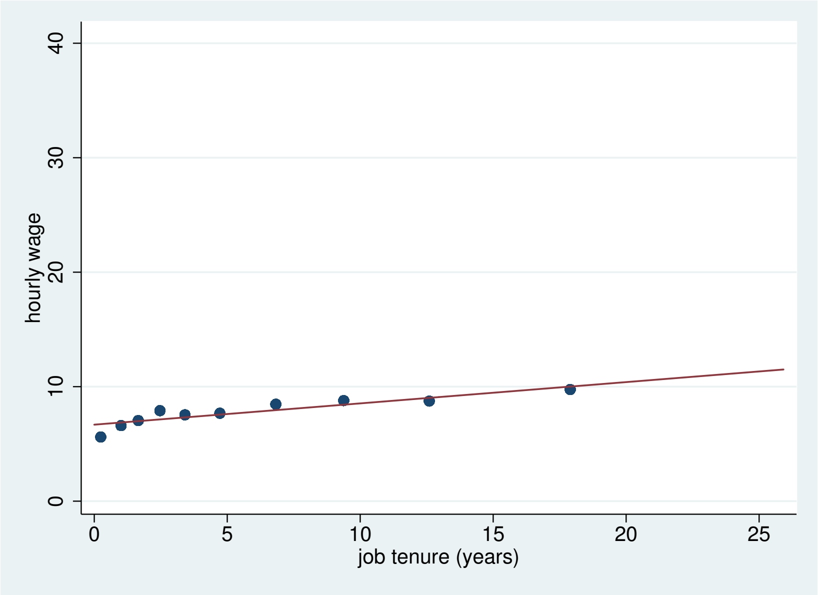
hourly wage
0
10
20
30
40
0
5
10
15
20
25
job tenure (years)



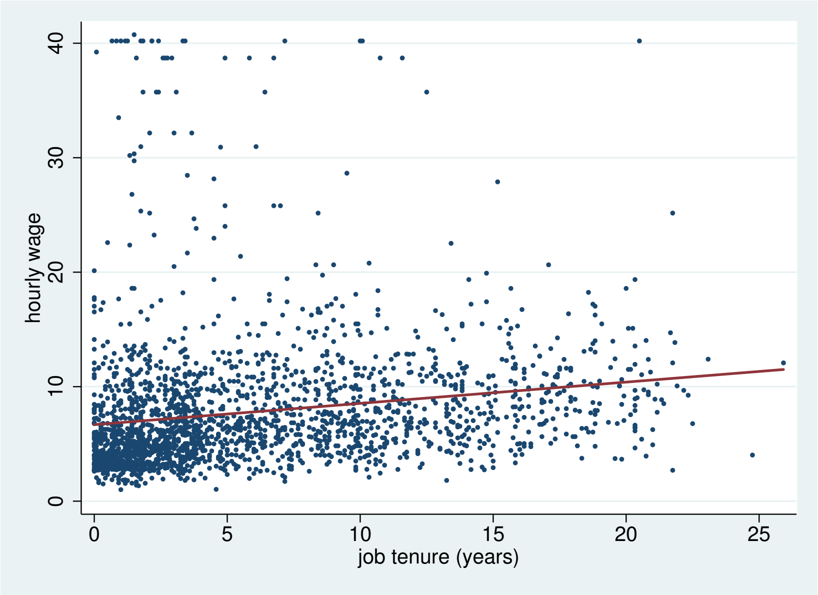
hourly wage
0
10
20
30
40
0
5
10
15
20
25
job tenure (years)


hourly wage
0
10
20
30
40
0
5
10
15
20
25
job tenure (years)

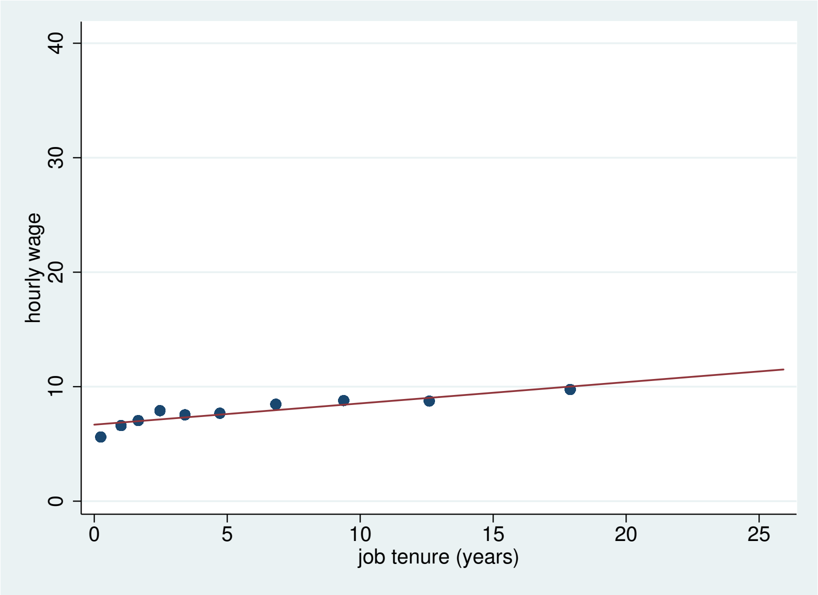

hourly wage
0
10
20
30
40
0
5
10
15
20
25
job tenure (years)



hourly wage
0
10
20
30
40
0
5
10
15
20
25
job tenure (years)

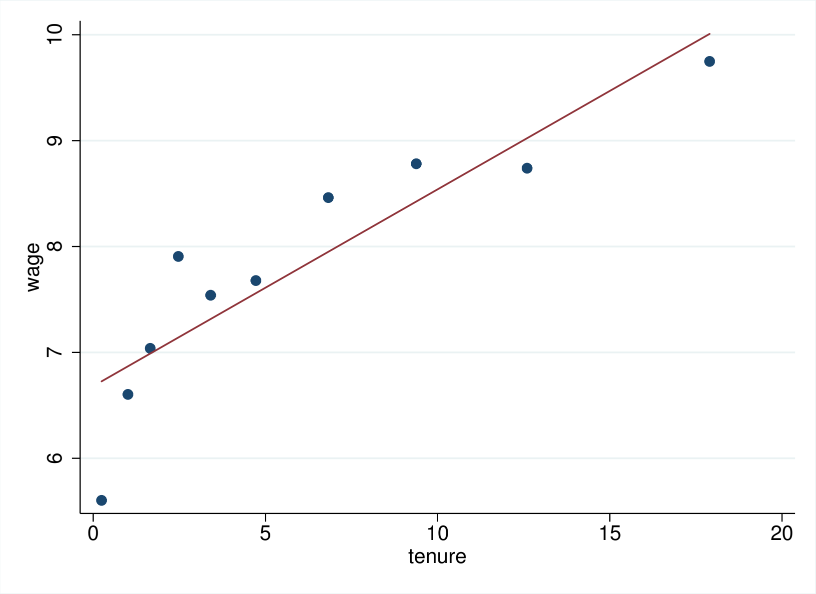


hourly wage
0
10
20
30
40
0
5
10
15
20
25
job tenure (years)



hourly wage
0
10
20
30
40
0
5
10
15
20
25
job tenure (years)

```
.  binscatter wage tenure
```



binscatter: Summary

- To create a binned scatterplot, binscatter
  1. Groups the x-axis variable into equal-sized bins
  2. Computes the mean of the x-axis and y-axis variables within each bin
  3. Creates a scatterplot of these data points
  4. Draws the population regression line
- binscatter supports weights
  - weighted bins
  - weighted means
  - weighted regression line

Binscatter and Regression: intimately linked

Conditional Expectation Function

- Consider two random variables: Y_i and X_i
- The conditional expectation function (CEF) is

$$\mathbb{E}[Y_i|X_i = x] \equiv h(x)$$

- The CEF tells us the mean value of Y_i when we see $X_i = x$
- The CEF is the best predictor of Y_i given X_i
  - in the sense that it minimizes Mean Squared Error

Regression CEF Theorem

- Suppose we run an OLS regression:

$$Y_i = \alpha + \beta X_i + \epsilon$$

- We obtain the estimated coefficients $\hat{\alpha}$, $\hat{\beta}$
  - Regression fit line: $\hat{h}(x) = \hat{\alpha} + \hat{\beta}x$

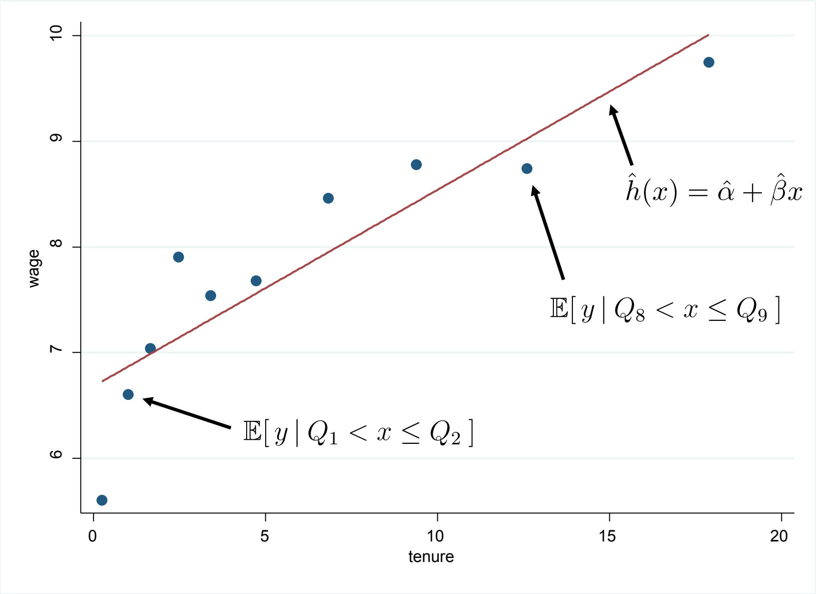
Regression CEF Theorem:

- The regression fit line $\hat{h}(x) = \hat{\alpha} + \hat{\beta}x$ is the best linear approximation to the CEF, $h(x) = \mathbb{E}[Y_i|X_i = x]$
  - in the sense that it minimizes Mean Squared Error

binscatter: CEF and regression fit line

A typical binned scatterplot shows two related objects:

- a non-parametric estimate of the CEF
  - the binned scatter points
- the best linear estimate of the CEF
  - the regression fit line



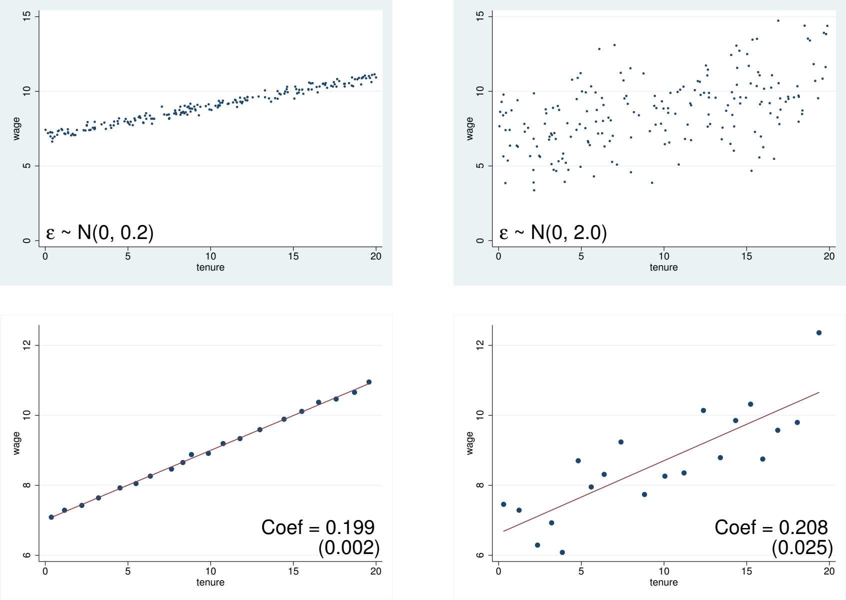
10
9
8
7
6
wage
0
5
10
15
20
tenure
$\hat{h}(x) = \hat{\alpha} + \hat{\beta}x$
$\mathbb{E}[\, y \mid Q_8 < x \leq Q_9 \,]$
$\mathbb{E}[\, y \mid Q_1 < x \leq Q_2 \,]$

Interpreting binscatters

binscatters: informative about standard errors

- If the binned scatterpoints are tight to the regression line, the slope is precisely estimated
  - regression standard error is small
- If the binned scatterpoints are dispersed around the regression line, the slope is imprecisely estimated
  - regression standard error is large

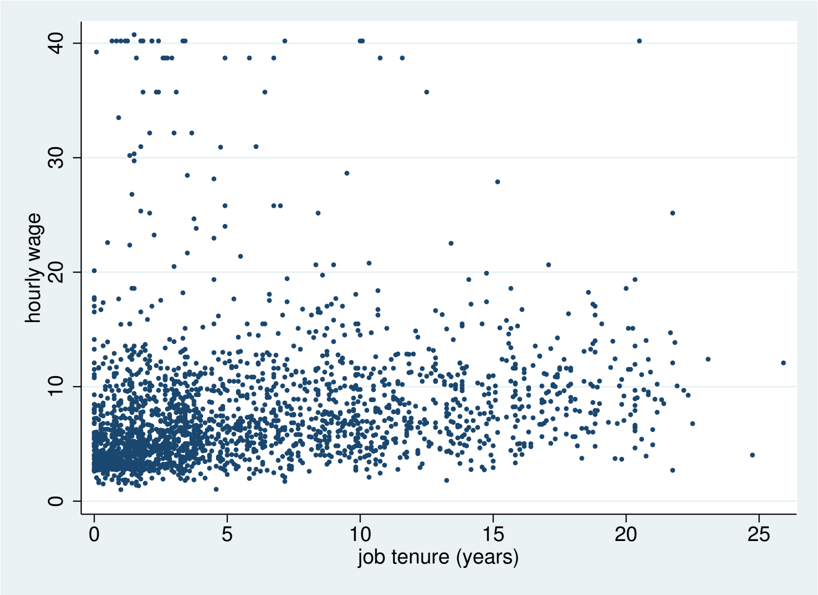
- Dispersion of binned scatterpoints around the regression line indicates statistical significance



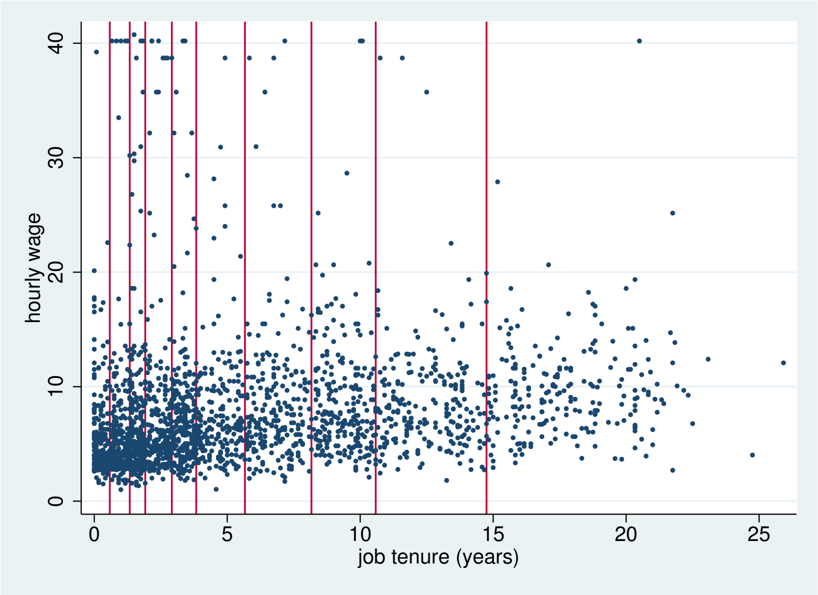
ε ~ N(0, 0.2)
ε ~ N(0, 2.0)
wage
tenure
0
5
10
15
20
6
8
12
Coef = 0.199
(0.002)
Coef = 0.208
(0.025)

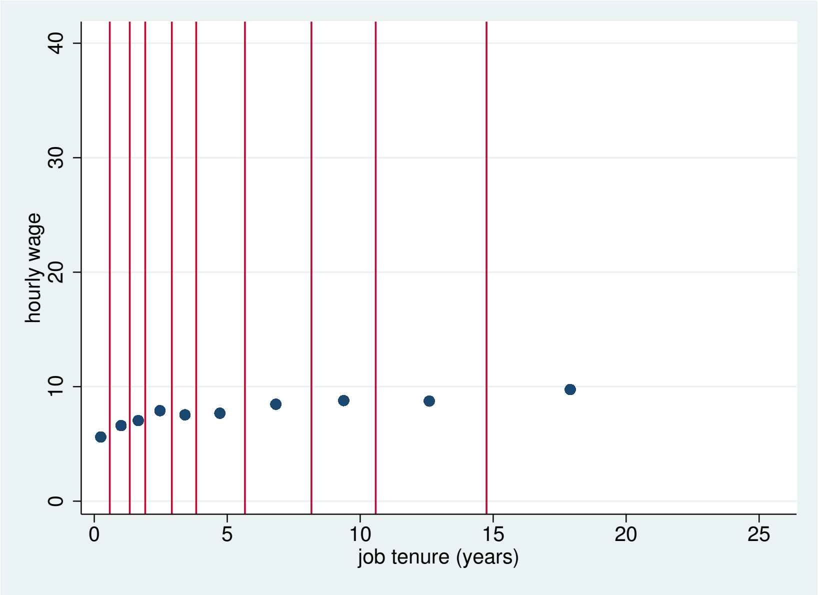
binscatters: not informative about R^2

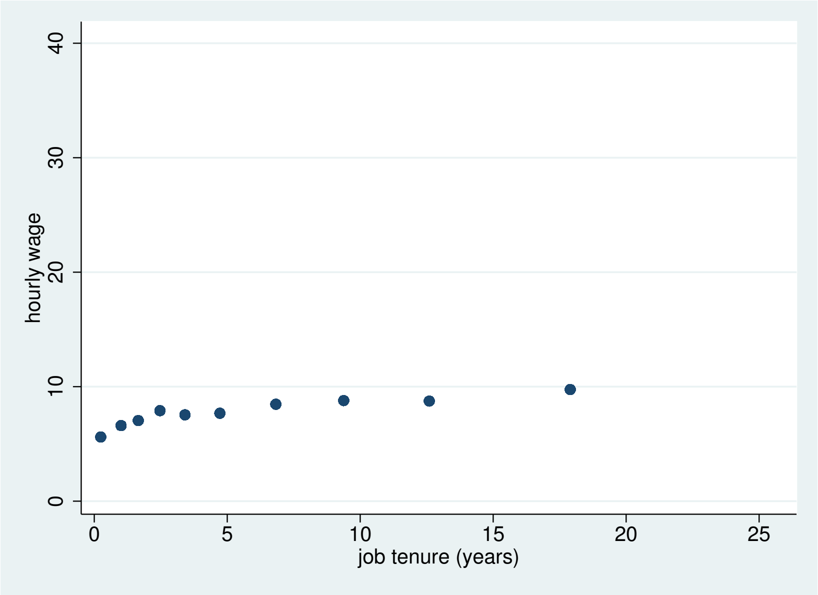
- R^2 tells you what fraction of the *individual* variation in Y is explained by the regressors
- A binned scatterplot collapses all the individual variation, showing only the mean within each bin



hourly wage
0
10
20
30
40
0
5
10
15
20
25
job tenure (years)


hourly wage
0
10
20
30
40
0
5
10
15
20
25
job tenure (years)


hourly wage
0
10
20
30
40
0
5
10
15
20
25
job tenure (years)

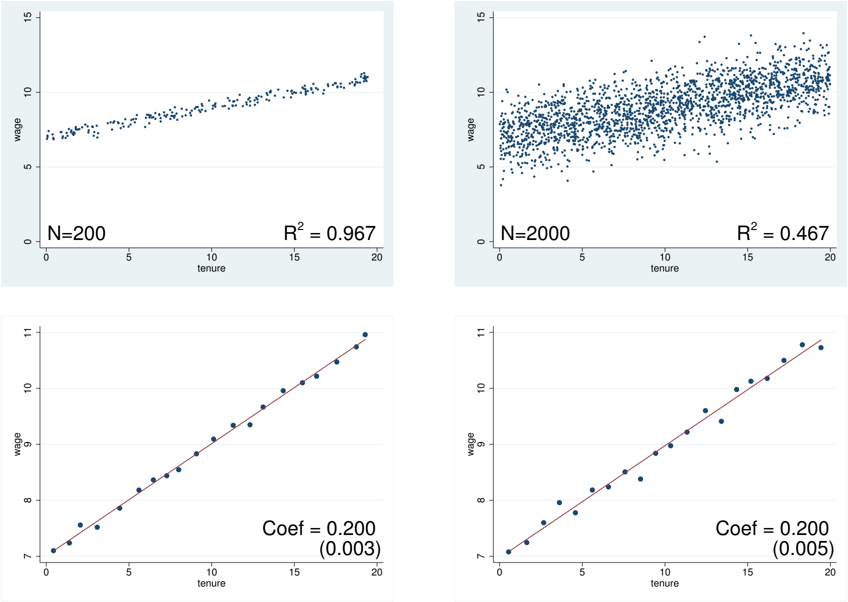


hourly wage
0
10
20
30
40
0
5
10
15
20
25
job tenure (years)

binscatters: not informative about R^2

- The same binscatter can be generated with:
  - enormous variance in $Y|X = x$
  - or almost no individual variance

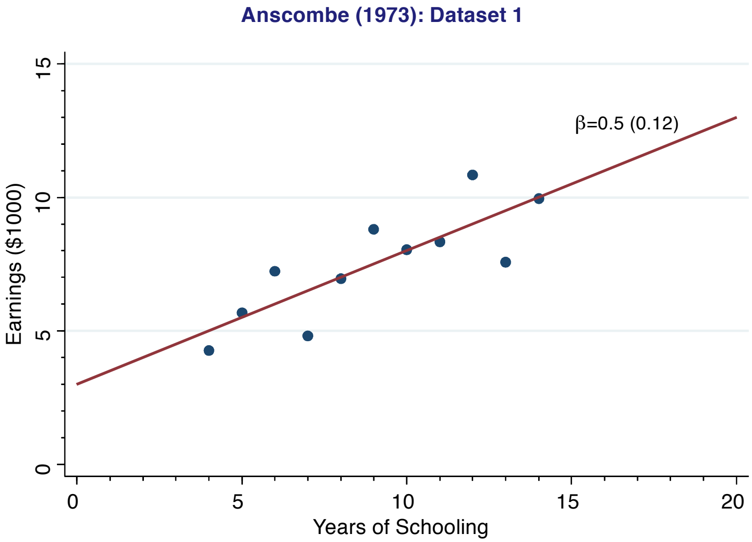
- because binscatter only shows $\mathbb{E}[Y|X = x]$

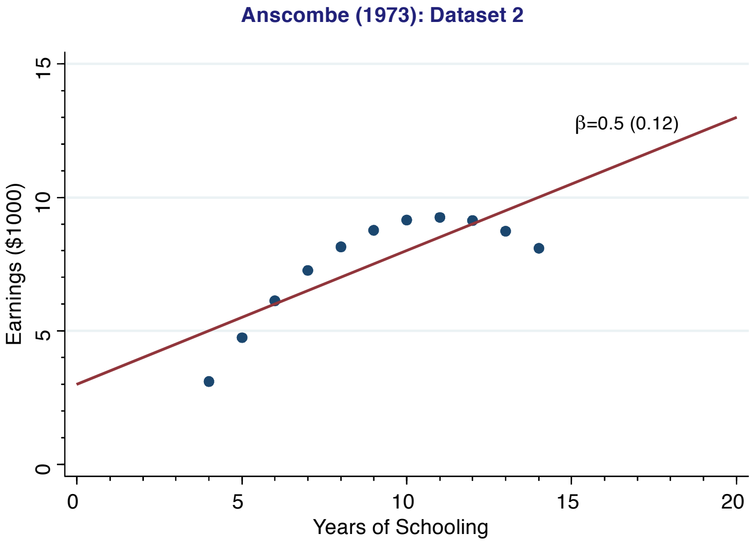


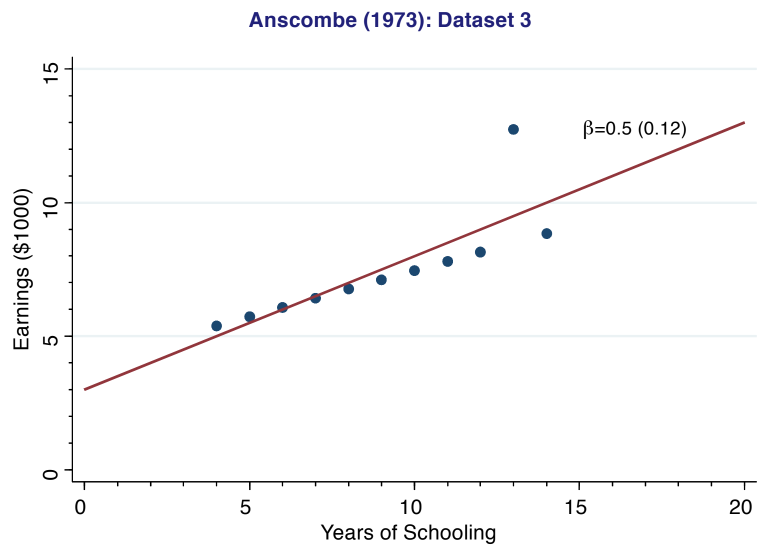
N=200
R^2 = 0.967
N=2000
R^2 = 0.467
Coef = 0.200
(0.003)
Coef = 0.200
(0.005)
wage
tenure

binscatters: informative about functional form

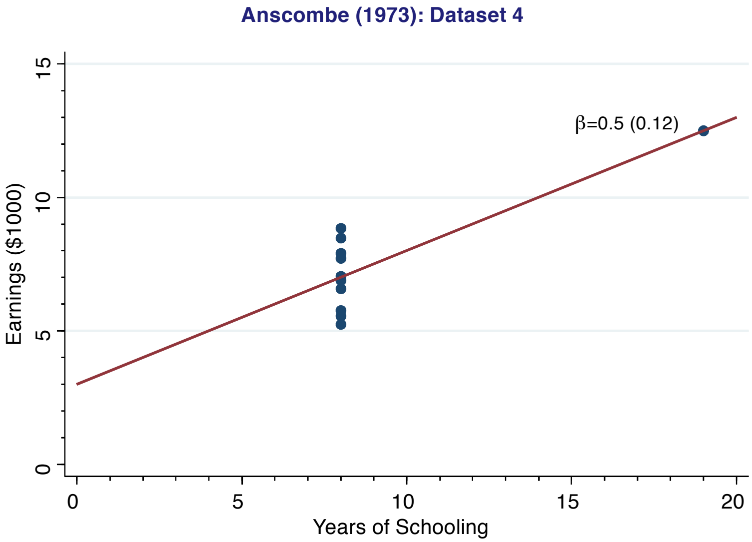
- Many different forms of underlying data can give the same regression results
  - Some examples from Anscombe (1973)...


Anscombe (1973): Dataset 1
β=0.5 (0.12)
Earnings ($1000)
Years of Schooling
0
5
10
15
20


Anscombe (1973): Dataset 2
β=0.5 (0.12)
Earnings ($1000)
Years of Schooling
0
5
10
15
20



Anscombe (1973): Dataset 3
β=0.5 (0.12)
Earnings ($1000)
0
5
10
15
0
5
10
15
20
Years of Schooling


Anscombe (1973): Dataset 4
β=0.5 (0.12)
Earnings ($1000)
Years of Schooling
0
5
10
15
20

binscatters: informative about functional form

- Suppose the true data generating process is logarithmic

$$wage_i = 10 + log(tenure_i) + \epsilon_i$$

binscatters: informative about functional form

Now forget that I ever told you that...

You're just handed the data.

- Run a linear regression:

$$wage_i = \alpha + \beta tenure_i + \epsilon_i$$

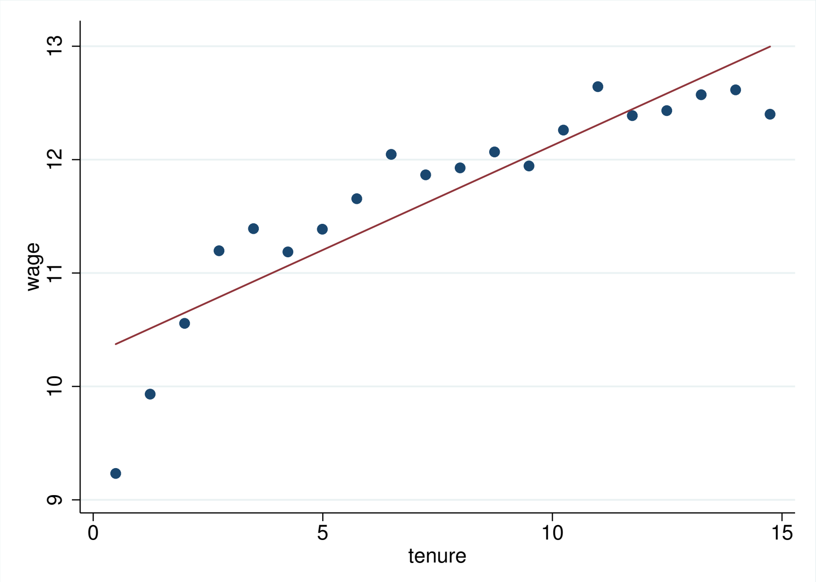
```
. reg wage tenure

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  281.25
       Model |  317.940139     1  317.940139           Prob > F      =  0.0000
    Residual |  562.975924   498  1.13047374           R-squared     =  0.3609
-------------+------------------------------           Adj R-squared =  0.3596
       Total |  880.916063   499  1.76536285           Root MSE      =  1.0632

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |   .1841569   .0109811    16.77   0.000     .1625819    .2057318
       _cons |   10.28268   .0961947   106.89   0.000     10.09369    10.47168
------------------------------------------------------------------------------
```

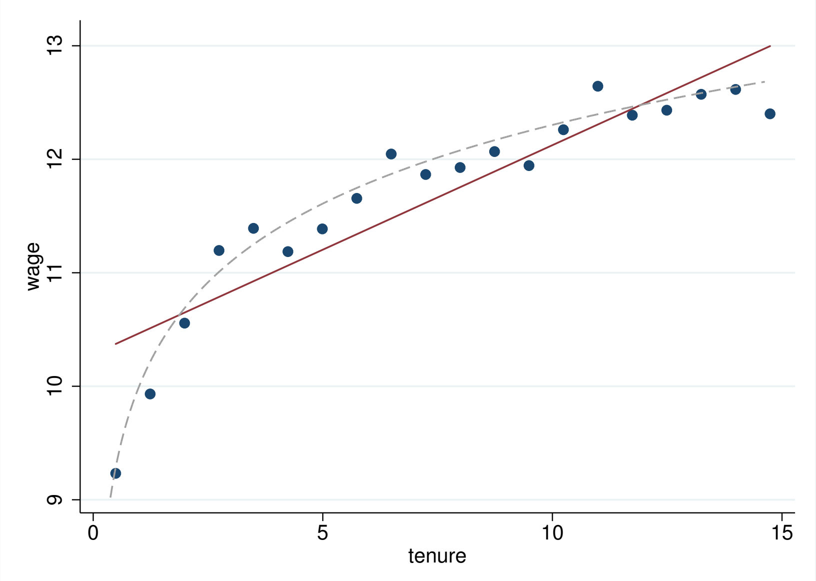
binscatters: informative about functional form

```
. binscatter wage tenure
```



binscatters: informative about functional form

```
.  binscatter wage tenure
```

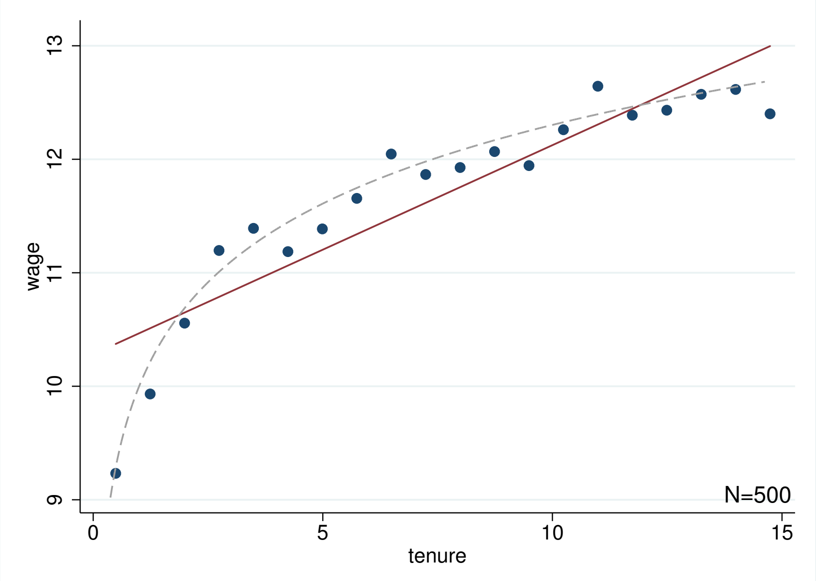


binscatters: informative about functional form

- If the underlying CEF is smooth, binscatter provides a consistent estimate of the CEF
  - As N gets large, holding the number of quantiles constant, each binned scatter point approaches the true conditional expectation

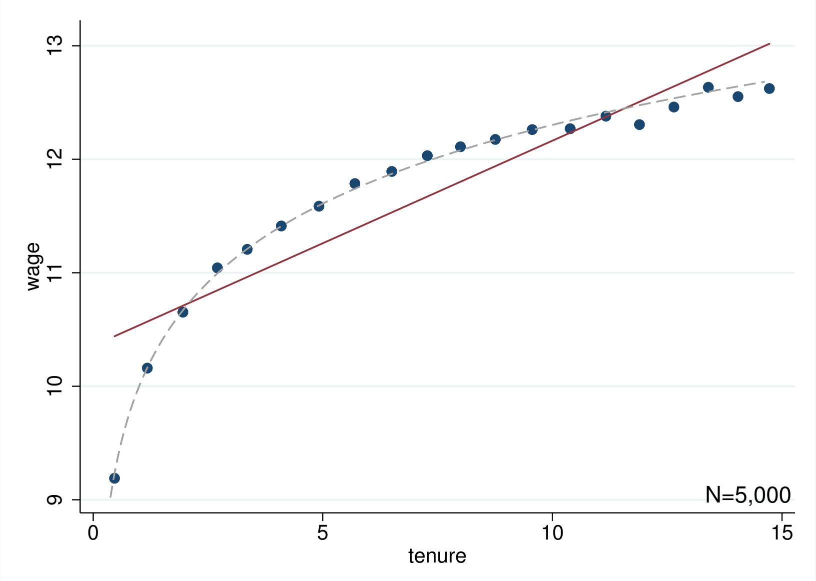
binscatters: informative about functional form

```
. binscatter wage tenure in 1/500
```



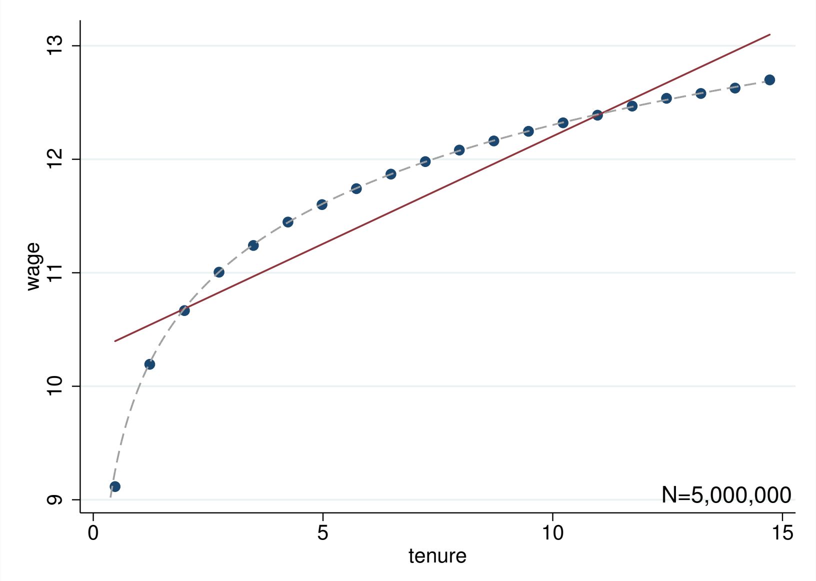
binscatters: informative about functional form

```
.  binscatter wage tenure in 1/5000
```



binscatters: informative about functional form

```
. binscatter wage tenure in 1/5000000
```



Interpreting binscatters: moral of the story

1. Binned scatterplots are informative about standard errors
2. Binned scatterplots are not informative about R^2
3. And binned scatterplots are informative about functional form

How many bins?

How many bins?

What is the "best" number of bins to use?

- Default in binscatter is 20
  - in my personal experience, this default works very well

- Optimal number of bins to accurately represent the CEF depends on curvature of the underlying CEF
  - which is unknown (that's why we're approximating it!)

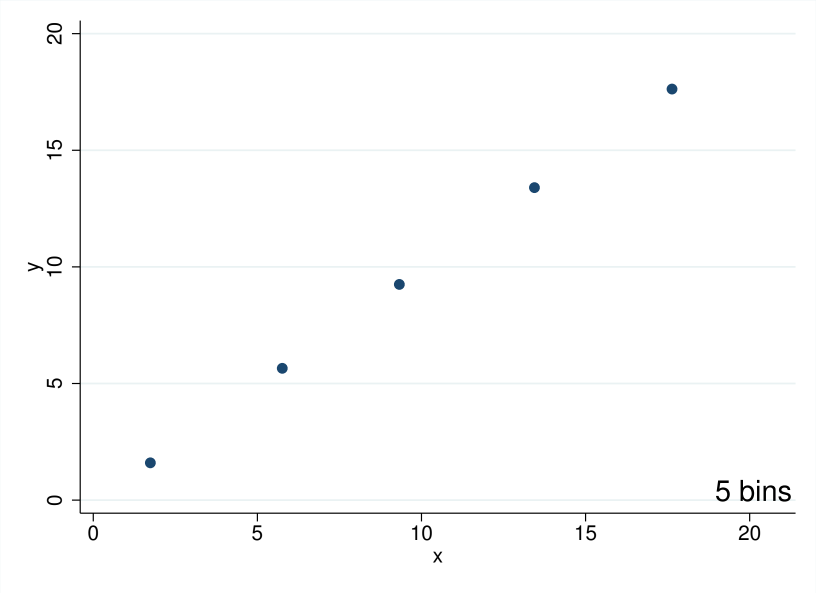
  - a smooth function can be well approximated with few points

  - a function with complex local behaviour requires many points to approximate its shape

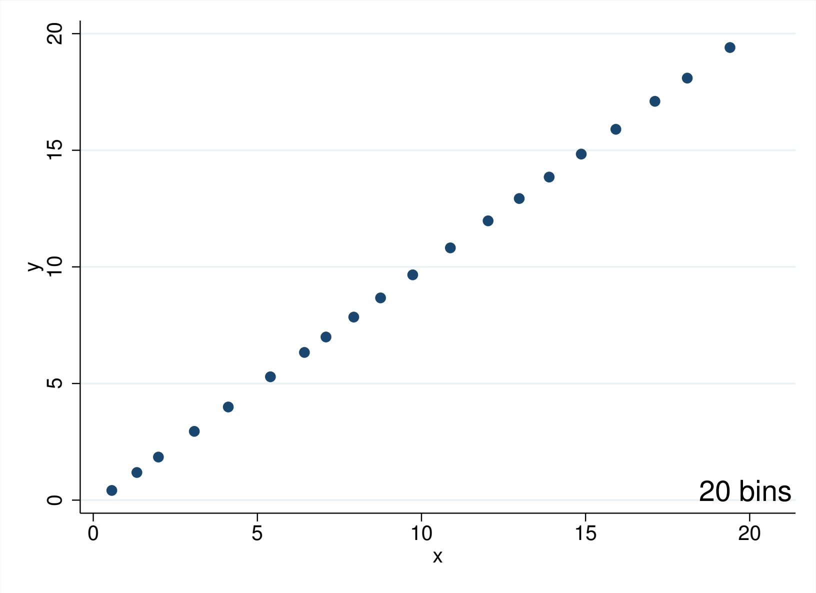
Let's play a quick game of...

What function is it?

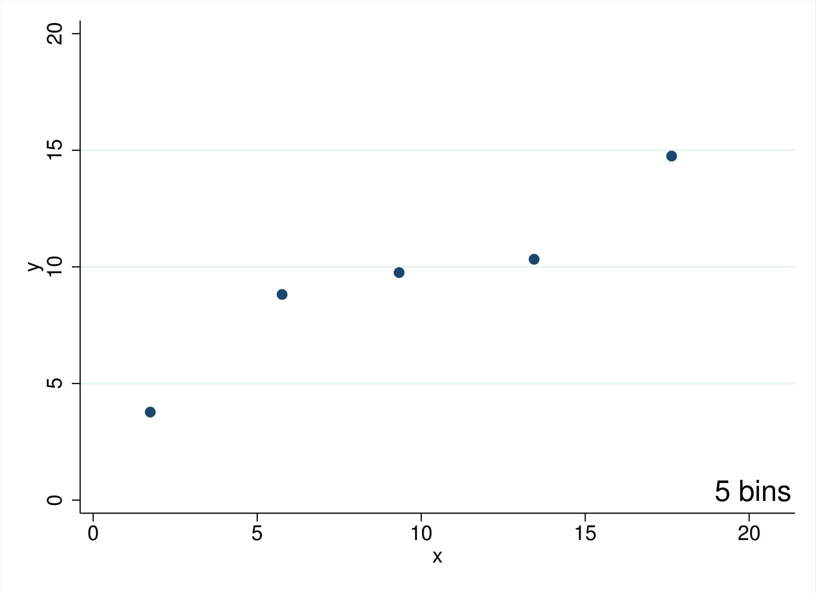
Round 1:



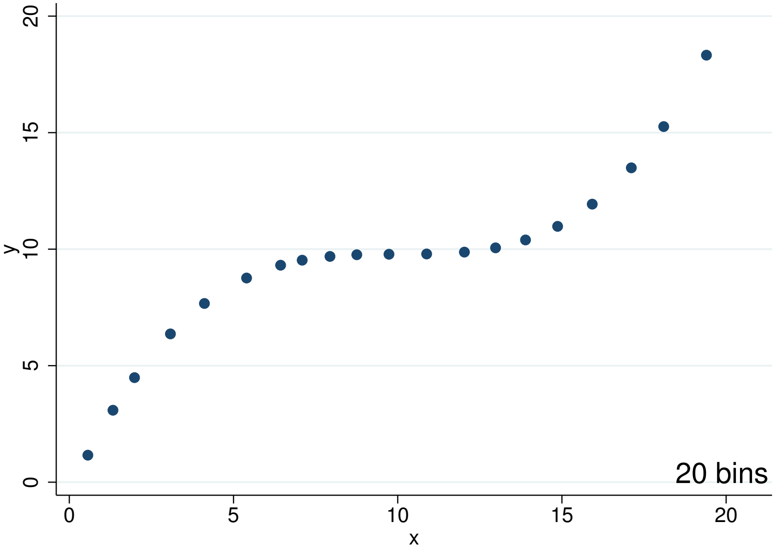
Round 1: Linear



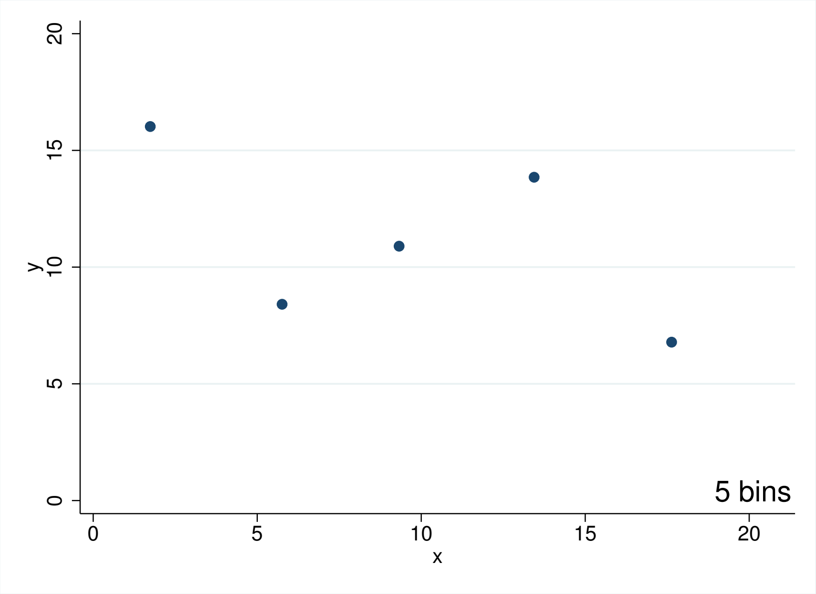
Round 2:



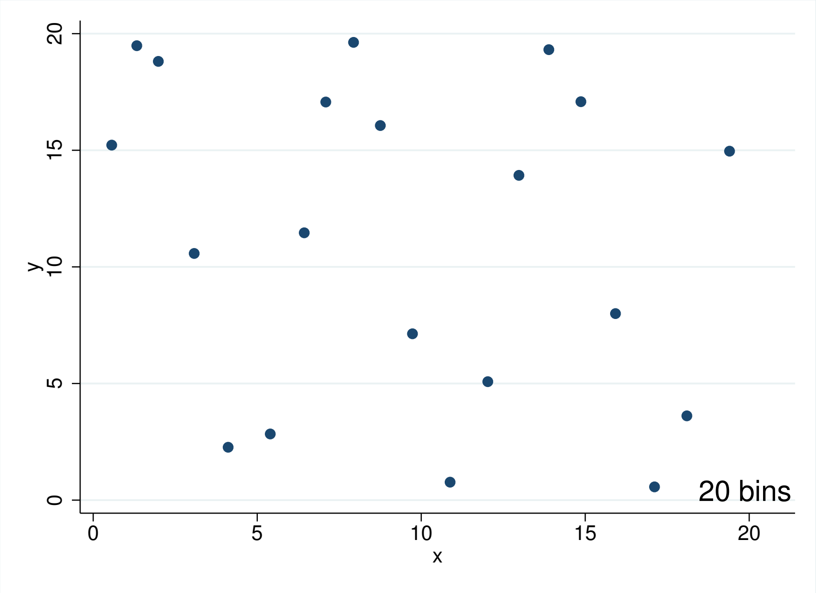
Round 2: Cubic



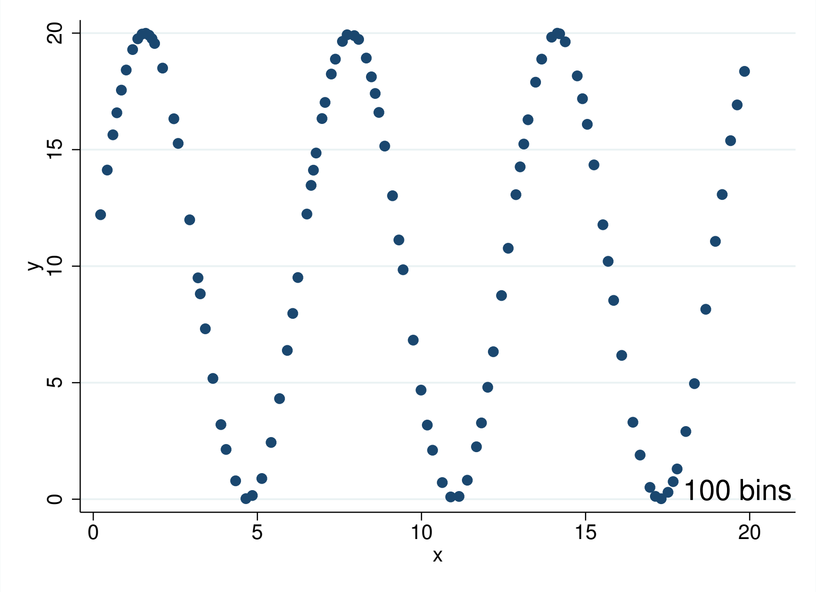
Round 3:



Round 3:



Round 3: Sinusoidal



binscatter: Multivariate Regression

- The use of binned scatterplots is not restricted to studying simple relationships with one x-variable
- binscatter can use partitioned regression to illustrate the relationship between two variables while controlling for other regressors

Partitioned regression: FWL theorem

- Suppose we're interested in the relationship between y and x in the following multivariate regression:

$$y = \alpha + \beta x + \Gamma Z + \epsilon$$

- **Option 1:** Run the full regression with all regressors, obtain $\hat{\beta}$
- **Option 2:** Partitioned regression
  1. Regress y on Z $\quad \Rightarrow \quad$ residuals $\equiv \widetilde{y}$
  2. Regress x on Z $\quad \Rightarrow \quad$ residuals $\equiv \widetilde{x}$
  3. Regress $\widetilde{y}$ on $\widetilde{x}$ $\quad \Rightarrow \quad$ coefficient $= \hat{\beta}$

▶ The $\hat{\beta}$ obtained using full regression and partitioned regression are identical

binscatter: Applying partitioned regression

- We're interested in the relationship between wage and tenure, but want to control for total work experience:

$$wage = \alpha + \beta\ tenure + \gamma\ experience + \epsilon$$

- Could directly apply partitioned regression:

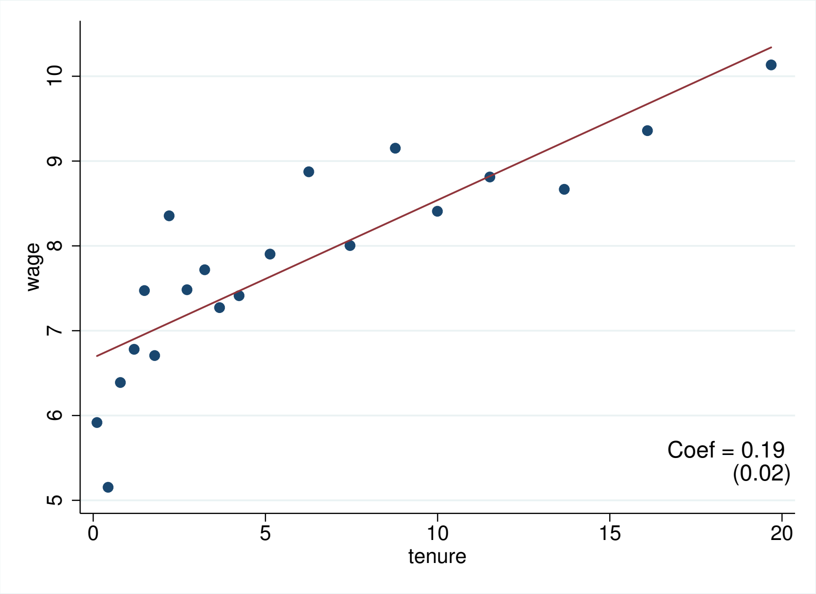
```
. reg wage experience
. predict wage_r, residuals
. reg tenure experience
. predict tenure_r, residuals

. binscatter wage_r tenure_r
```

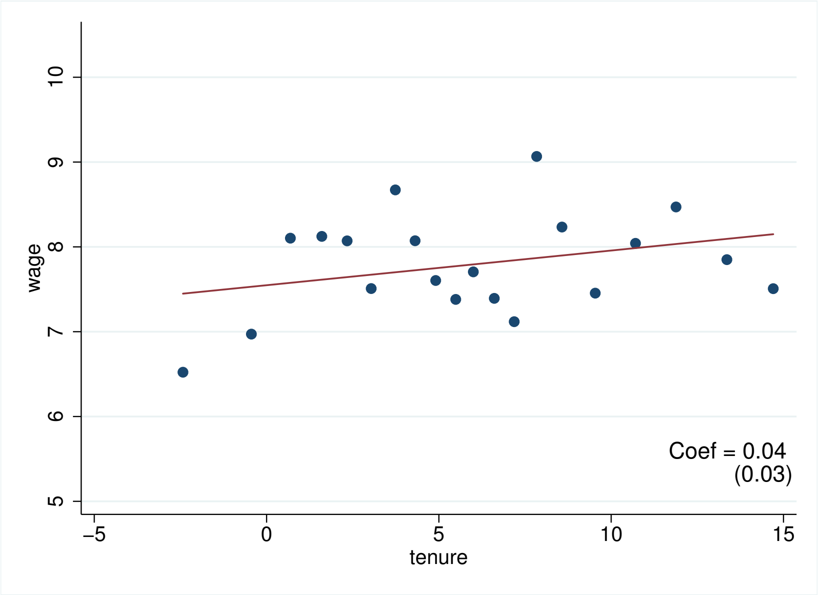
- The procedure is built into binscatter:

```
. binscatter wage tenure, controls(experience)
```

```
. binscatter wage tenure
```



```
.  binscatter wage tenure, controls(experience)
```

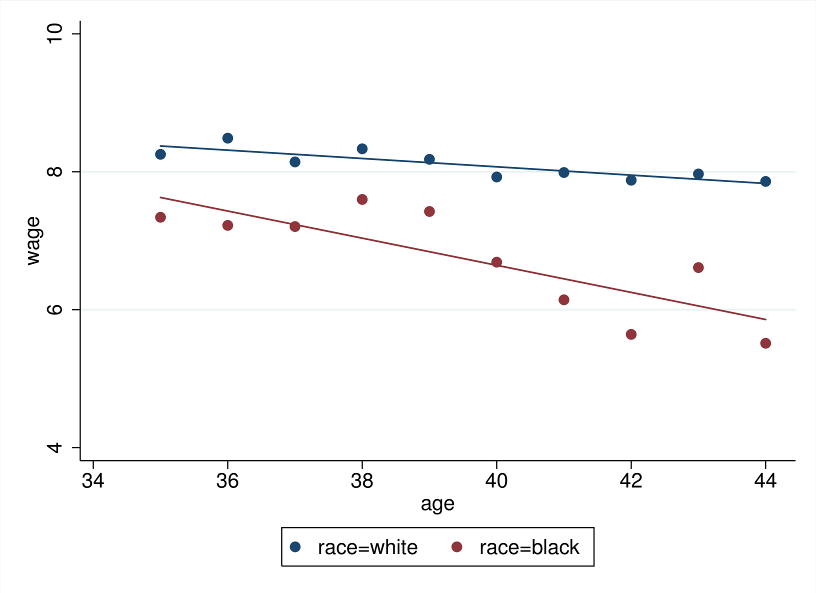


by-variables

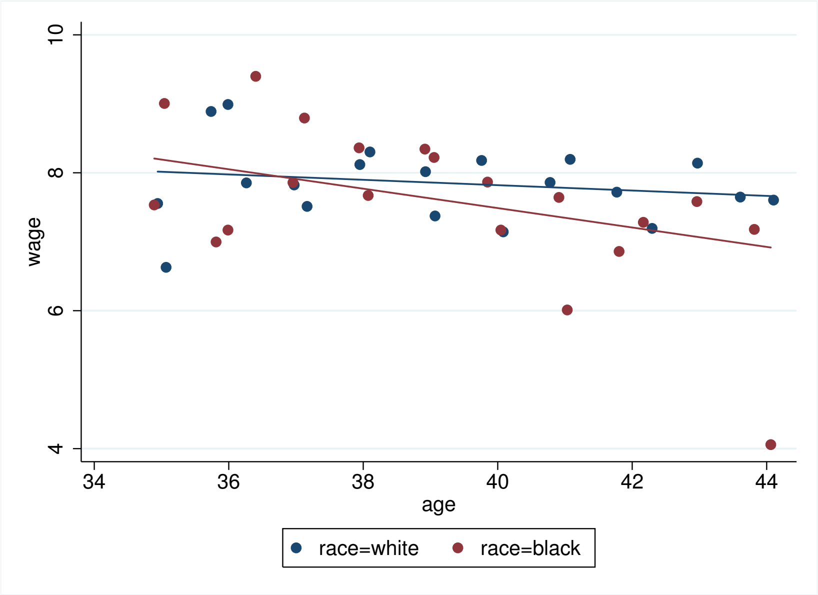
Plotting multiple series using a by-variable

- binscatter will plot a separate series for each group
  - each by-value has its own scatterpoints and regression line
  - the by-values share a common set of bins
    - constructed from the unconditional quantiles of the x-variable

```
.  binscatter wage age, by(race)
```



```
.  binscatter wage age, by(race) absorb(occupation)
```

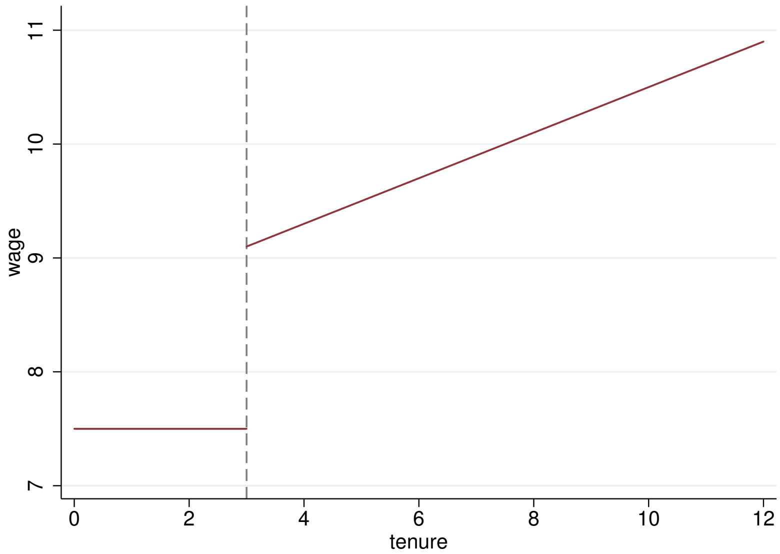


RD and RK designs

RD and RK designs

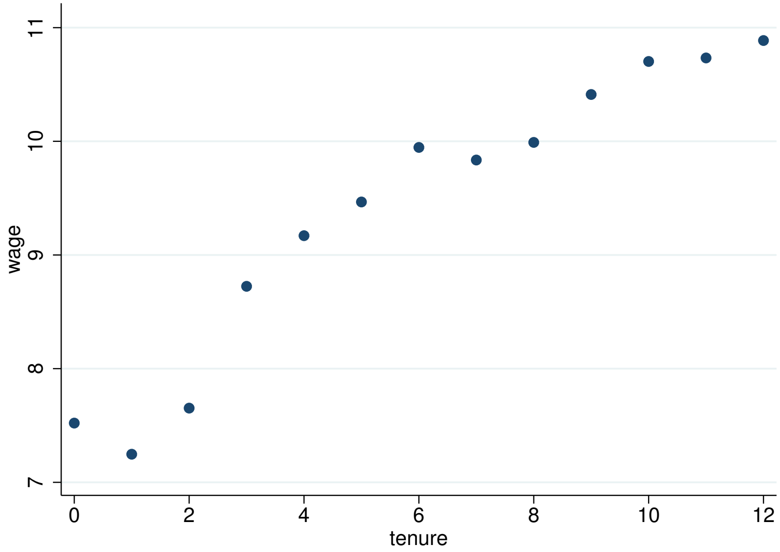
- Binned scatterplots are very useful for illustrating regression discontinuities (RD) or regression kinks (RK)
- Consider a wage schedule where the first 3 years are probationary
  - After 3 years, receive a salary bump
  - After 3 years, steady increase in salary for each additional year

RD design



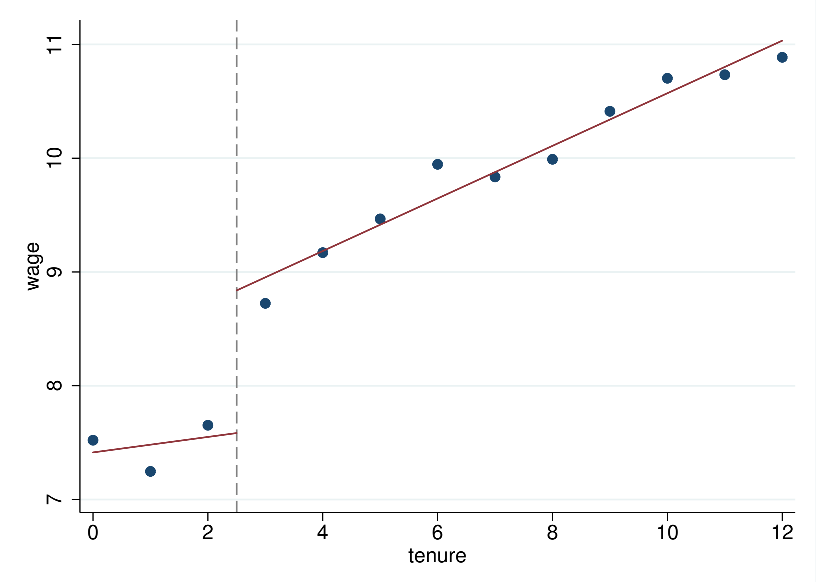
RD design

```
.  binscatter wage tenure, discrete line(none)
```



RD design

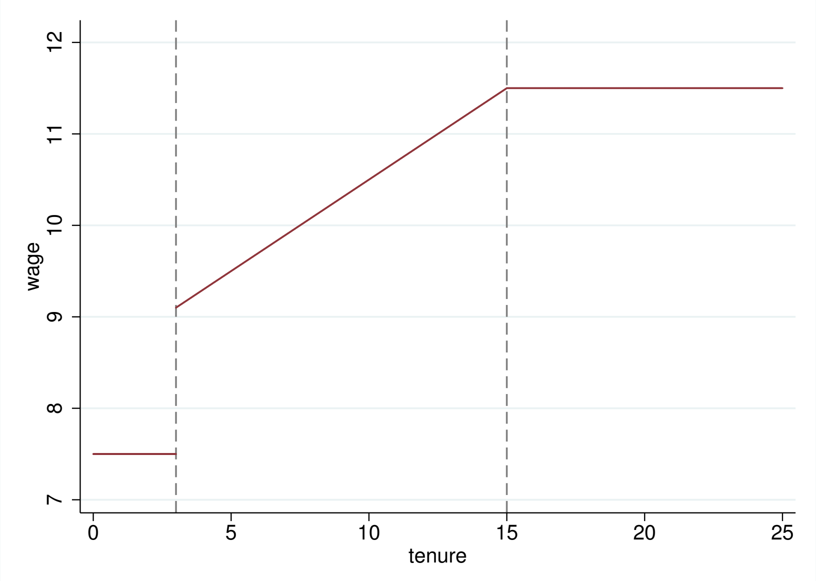
```
. binscatter wage tenure, discrete rd(2.5)
```



RK design

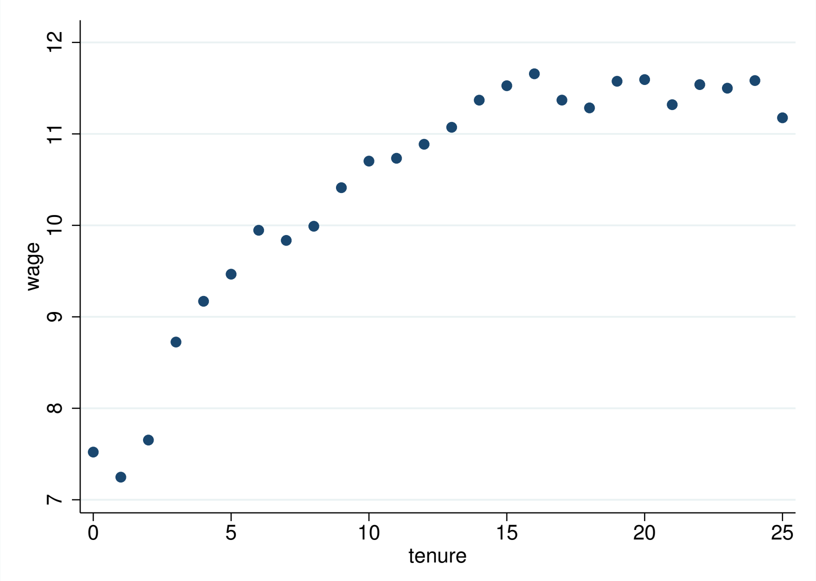
- The firm decides to cap the wage schedule after 15 years of tenure
  - No more salary increases past 15 years

RK design



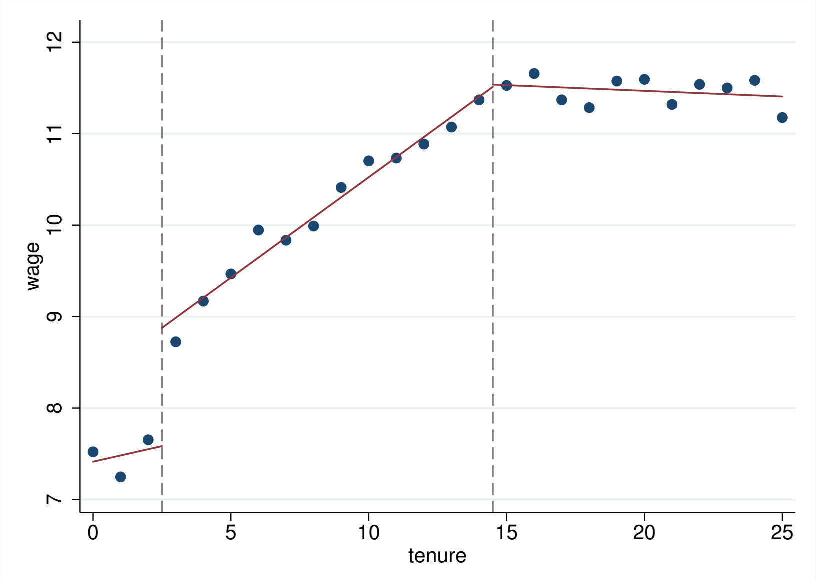
RK design

```
. binscatter wage tenure, discrete line(none)
```



RK design

```
.  binscatter wage tenure, discrete rd(2.5 14.5)
```



RD and RK designs

Important caution:

- The `rd()` option in binscatter only affects the regression lines
  - It does not affect the binning procedure
  - A bin could contain observations on both sides of the discontinuity, and average them together

Implications:

- Doesn't matter with discrete x-variable and option `discrete`
  - No binning is performed, each x-value is its own bin
- With continuous x-variable, need to manually create bins
  - Use `xq()` to specify variable with correctly constructed bins
  - A future version of binscatter respect RDs when binning

Event Studies

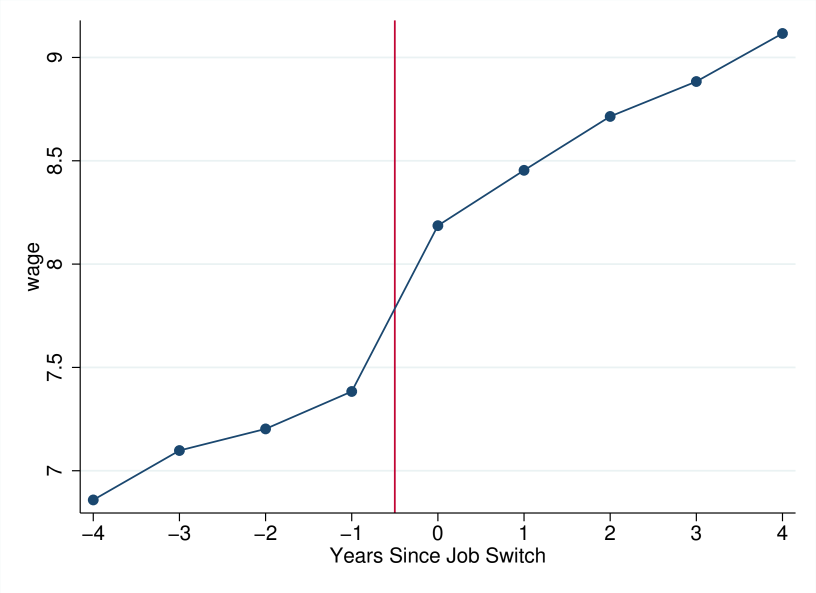
Event Studies

binscatter makes it easy to create event study plots.

Suppose we have a panel of people, with yearly observations of their wage and employer:

- We observe when people change employers
- For each person with a job switch
  - Define year 0 as the year they start a new job
  - So year -1 is the year before a job switch
  - Year 1 is the year after a job switch

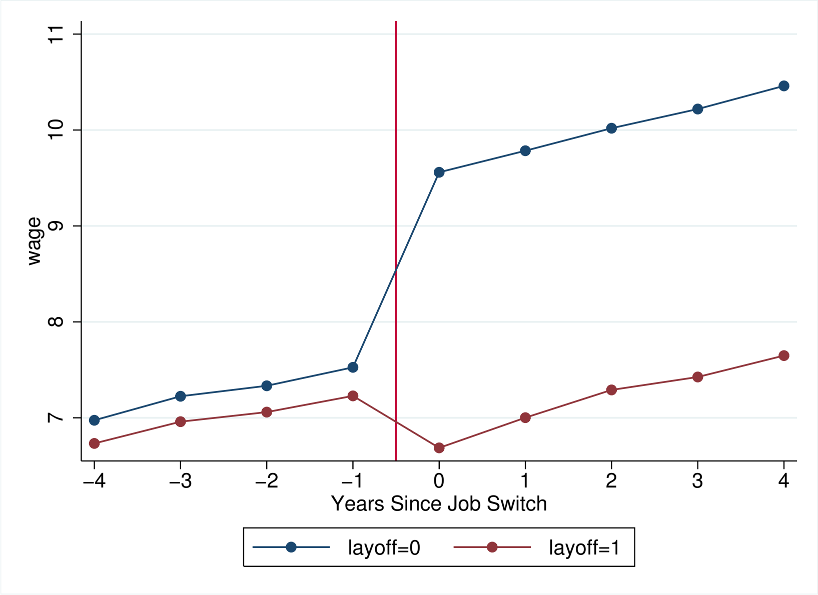
```
.  binscatter wage eventyear, line(connect) xline(-0.5)
```



Now suppose we also know whether they were laid off at their previous job.

- Does the wage experience of people who are laid off differ from those who quit voluntarily?

```
.  binscatter wage eventyear, line(connect) xline(-0.5)
>    by(layoff)
```



Final Remarks

Final Remarks

- binscatter is optimized to run quickly and efficiently in large datasets
- It can be installed from the Stata SSC repository
  - `ssc install binscatter`
- These slides and other documentation is posted on the binscatter website:

  www.michaelstepner.com/binscatter

References

Examples of binscatter used in research

Chetty, Raj, John N Friedman, and Emmanuel Saez. 2013. "Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings." *American Economic Review*, *103* (7): 2683–2721.

Chetty, Raj, John N. Friedman, and Jonah Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review*, *forthcoming.*

Chetty, Raj, John N. Friedman, and Jonah Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review*, *forthcoming.*

Chetty, Raj, John N. Friedman, Soren Leth-Petersen, Torben Nielsen, and Tore Olsen. 2013. "Active vs. Passive Decisions and Crowdout in Retirement Savings Accounts: Evidence from Denmark." *Quarterly Journal of Economics*, *forthcoming.*

References for this talk

Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, NJ: Princeton University Press.

Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *The American Statistician*, *27* (1): 17.

Chetty, Raj. 2012. "Econ 2450a: Public Economics Lectures." Lecture Slides, Harvard University. http://www.rajchetty.com/index.php/lecture-videos.