Biostat 200: Lecture 9

Associations, Correlations, and Causation

Visualization ReCap

- http://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/
- https://www.gapminder.org/tools/#_chart-type=bubbles
- www.Theyrule.net
- http://www.informationisbeautiful.net/
- www.bewitched.com

Association: Peeling the Onion

I See Dead People

What's wrong with this statement?

Associations are not Causations

Everyone who ate pickles in the year 1743 is now dead.

Therefore, pickles are fatal?

True statement:

Palm size correlates with your life expectancy

The larger your palm, the less you will live, on average.

Try it out - look at your neighbors and you’ll see who is expected to live longer.

Associations are ??? Causations

Why?

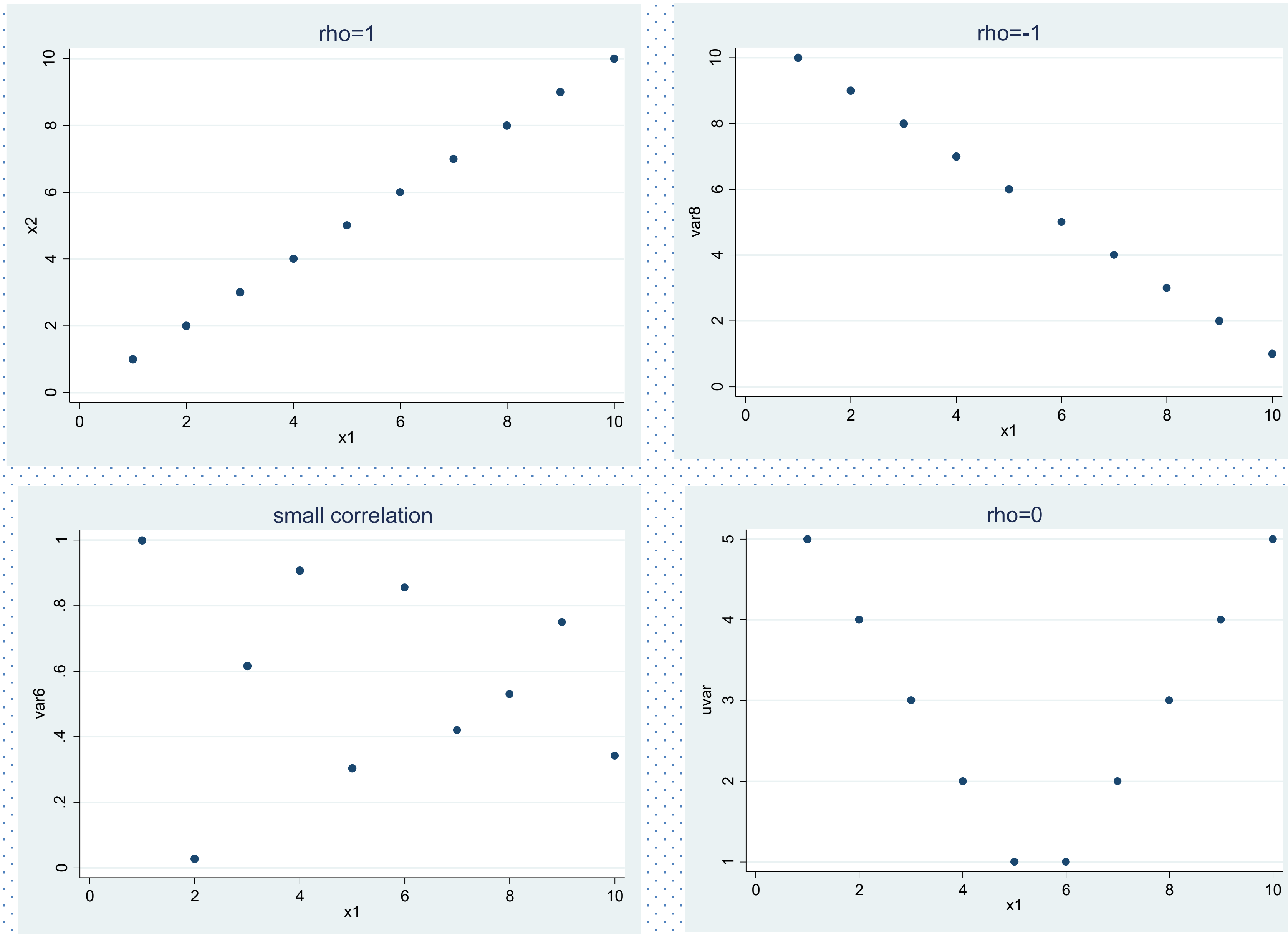
Women have smaller palms and live 6 years longer on average

Don't stop at the first association.
Ask yourself (and the data) WHY?

Correlation

- Correlation is a examines the relationship between 2 (usually) continuous variables
  - Does one increase with the other?
    - e.g. Does CD4 count increase with increasing age?
- Both variables are measured on the same people
- Correlation assumes a ***linear*** relationship between the two variables
- Correlation is symmetric
  - The correlation of A with B is the same as the correlation of B with A
  - Designated by ρ (rho)

Correlation



Correlation

- ρ lies between -1 and 1
- -1 and 1 are perfect correlations, 0 is no correlation
- An *estimator* of the population correlation ρ is Pearson's correlation coefficient denoted r
- It is estimated by: $$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Correlation

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$= \frac{1}{(n-1)}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right]}\sqrt{\left[\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}\right]}}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]}\sqrt{\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]}}$$

Correlation: hypothesis testing

- To test whether there is a correlation between two variables, our hypotheses are

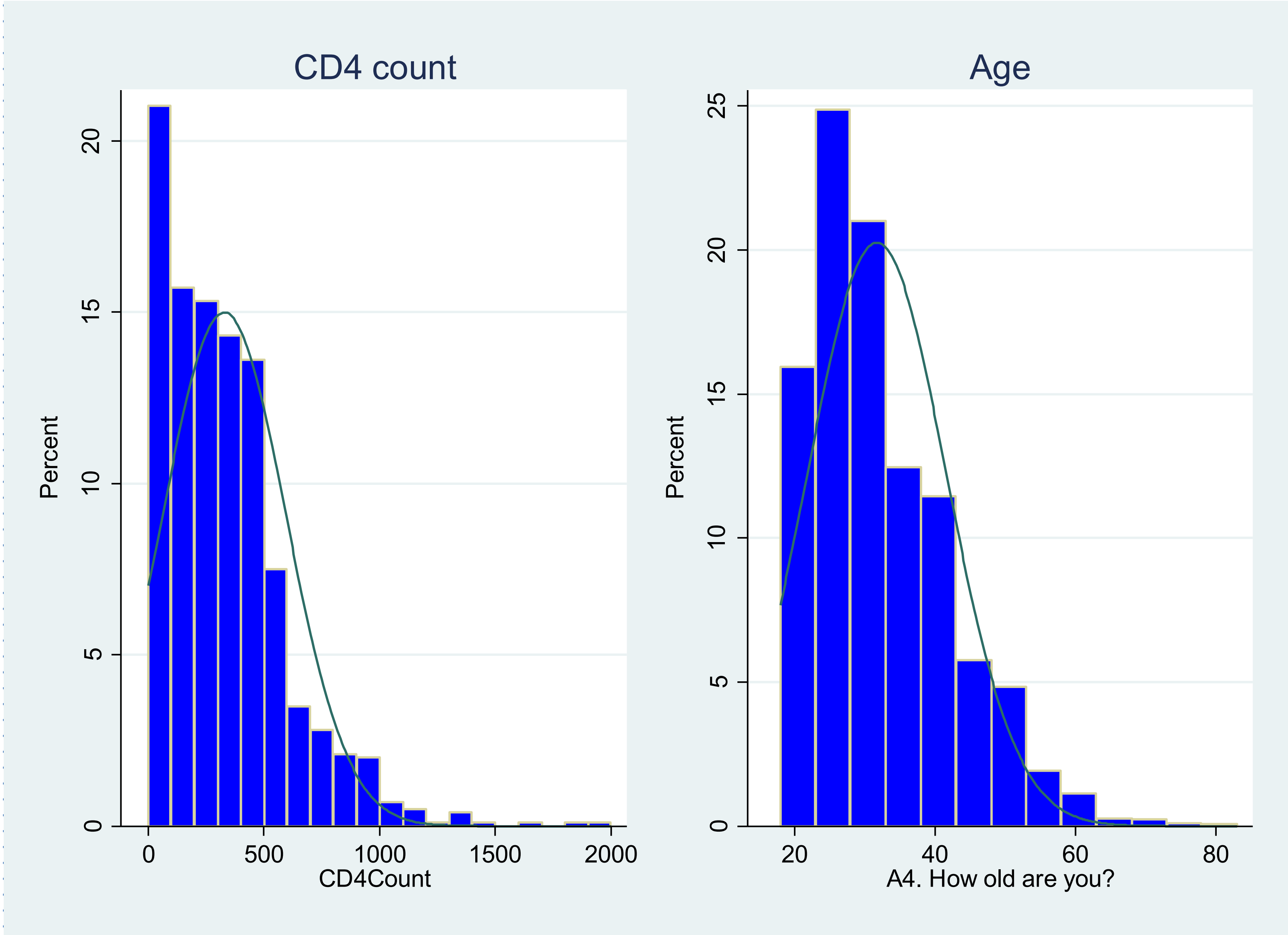
  $H_0 : \rho=0$ and $H_A : \rho\neq0$
- We need to calculate a test statistic for r
- The test statistic is $t = \dfrac{r-0}{se(r)}$

$$\text{where } se(r) = \sqrt{\frac{1-r^2}{n-2}}$$

$$\text{so } t = r\sqrt{\frac{n-2}{1-r^2}}$$

Correlation: hypothesis testing

- The test statistic follows a t distribution with n-2 degrees of freedom under the null
- The assumptions
  - The pairs of observations (x_i , y_i) were obtained from a random sample
  - X and Y are normally distributed


CD4 count
Percent
0
5
10
15
20
0
500
1000
1500
2000
CD4Count
Age
Percent
0
5
10
15
20
25
20
40
60
80
A4. How old are you?

Correlation example

```
 corr cd4count age
(obs=999)

             | cd4count    age_b
-------------+------------------
    cd4count |   1.0000
       age_b |  -0.1133   1.0000
```

Note that the hypothesis test is only of ρ=0, no other null
Also note that the correlation is the linear relationship only

Correlation example

pwcorr var1 var2, sig obs

```
. pwcorr cd4count age, sig obs

             | cd4count    age_b
-------------+------------------
    cd4count |   1.0000
             |
             |      999
             |
       age_b |  -0.1133   1.0000
             |   0.0003
             |      999     3387
             |
```

Spearman rank correlation (Nonparametric correlation)

- Pearson’s correlation coefficient is very sensitive to extreme values

- Spearman rank correlation calculates the Pearson correlation on the *ranks* of each variable

- The Pearson correlation coefficient is calculated, but the data values are replaced by the ranks

Spearman rank correlation (nonparametric)

- The Spearman rank correlation ranges between -1 and 1 as does the Pearson correlation
- We can test the null hypothesis that $\rho=0$
- The test statistic for n>10 is with n-2 degrees of freedom

is compared to the t distribution

$$t_s = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

```
spearman  cd4count age, stats(rho obs p)

 Number of obs =      999
Spearman's rho =       -0.1319

Test of Ho: cd4count and age_b are independent
    Prob > |t| =        0.0000
```

Matrix of Pearson correlations

- We can calculate a correlation matrix by listing multiple variable names
- Beware of which n's are used (use listwise option to get all n's equal so you are testing similar relationships)

```
pwcorr cd4count age nsupport_b auditc, sig obs

             | cd4count    age_b nsuppo~b   auditc
-------------+------------------------------------
    cd4count |   1.0000
             |
             |      999
             |
       age_b |  -0.1133   1.0000
             |   0.0003
             |      999     3387
             |
  nsupport_b |   0.0067   0.3597   1.0000
             |   0.8340   0.0000
             |      994     3370     3372
             |
      auditc |   0.0670   0.0580   0.0468   1.0000
             |   0.0392   0.0009   0.0078
             |      948     3242     3230     3244
             |
```

Matrix of Pearson correlations

- Using listwise option to get all n's equal

```
pwcorr cd4count age nsupport_b auditc, sig obs listwise

             | cd4count    age_b nsuppo~b   auditc
-------------+------------------------------------
    cd4count |   1.0000
             |
             |      944
             |
       age_b |  -0.1053   1.0000
             |   0.0012
             |      944      944
             |
  nsupport_b |   0.0107   0.3209   1.0000
             |   0.7416   0.0000
             |      944      944      944
             |
      auditc |   0.0668  -0.0075   0.0475   1.0000
             |   0.0401   0.8181   0.1449
             |      944      944      944      944

               |
 |
```

Matrix of Spearman correlations

```
spearman cd4count age nsupport_b auditc, pw stats(rho obs p)

+-----------------+
|  Key            |
|-----------------|
|   rho           |
|   Number of obs |
|   Sig. level    |
+-----------------+

             | cd4count    age_b nsuppo~b   auditc
-------------+------------------------------------
    cd4count |   1.0000
             |      999
             |
             |
       age_b |  -0.1319   1.0000
             |      999     3387
             |   0.0000
             |
  nsupport_b |   0.0118   0.4055   1.0000
             |      994     3370     3372
             |   0.7093   0.0000
             |
      auditc |   0.0945   0.0777   0.0447   1.0000
             |      948     3242     3230     3244
             |   0.0036   0.0000   0.0111
             |
```

Matrix of Spearman correlations

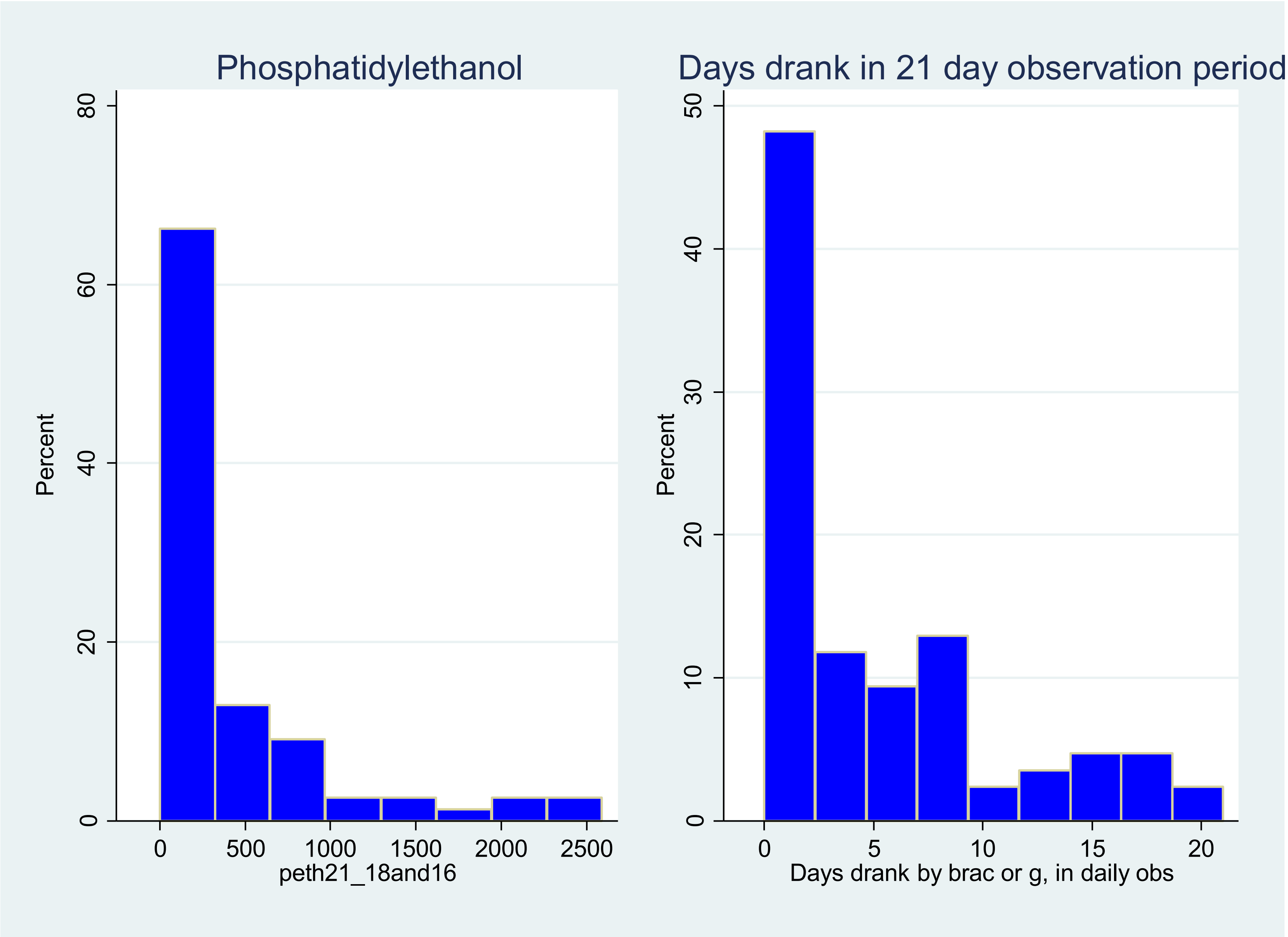
```
spearman cd4count age nsupport_b auditc, stats(rho obs p)

+-----------------+
|  Key            |
|-----------------|
|   rho           |
|   Number of obs |
|   Sig. level    |
+-----------------+

             | cd4count    age_b nsuppo~b   auditc
-------------+------------------------------------
    cd4count |   1.0000
             |      944
             |
             |
       age_b |  -0.1261   1.0000
             |      944      944
             |   0.0001
             |
  nsupport_b |   0.0127   0.3546   1.0000
             |      944      944      944
             |   0.6971   0.0000
             |
      auditc |   0.0940  -0.0116   0.0431   1.0000
             |      944      944      944      944
             |   0.0039   0.7221   0.1859
             |

|
```

Here if you drop the "pw" option you get all n's equal

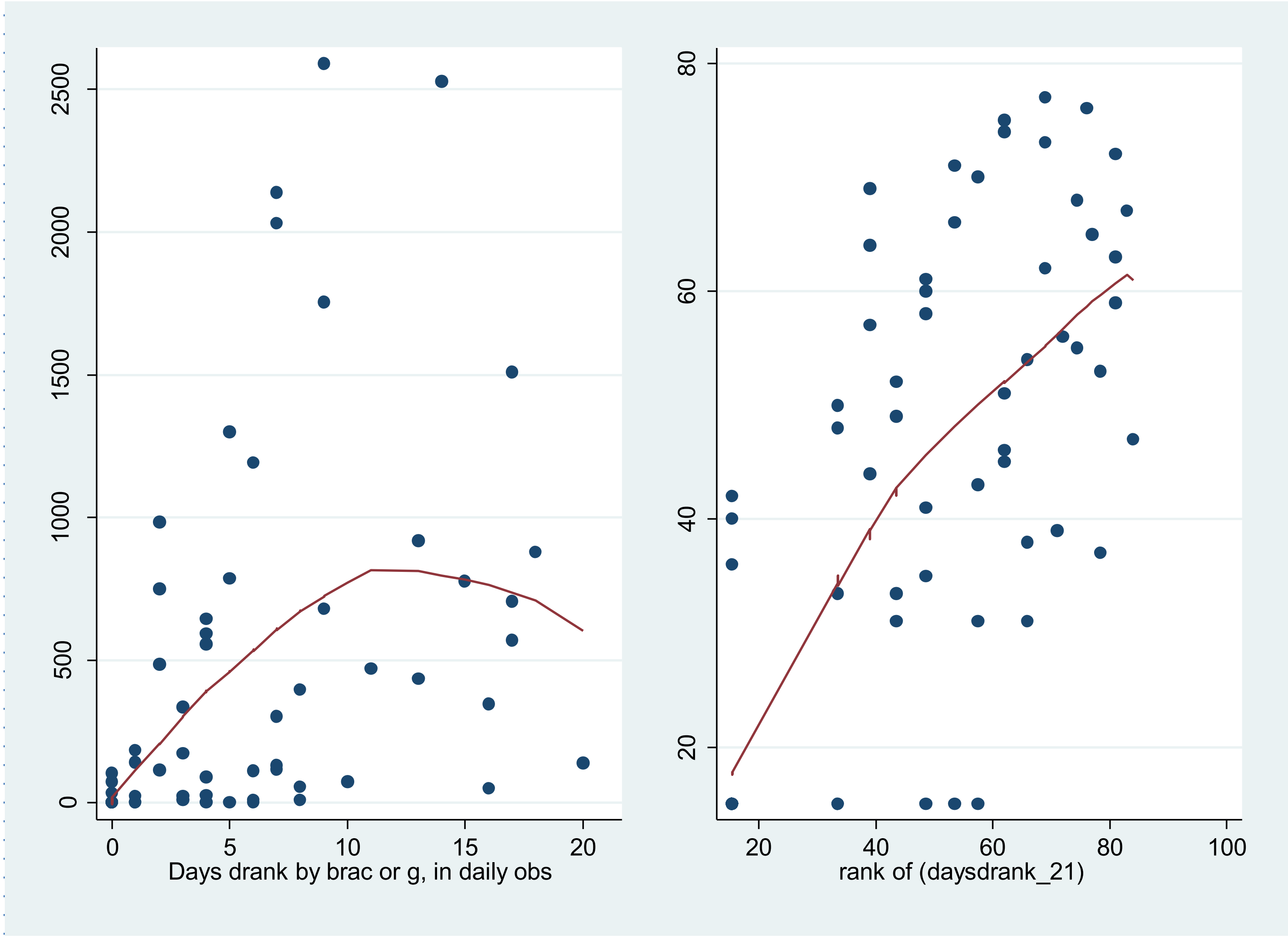
Pearson vs. Spearman

	Pearson	Spearman
Sensitivity to outliers	more sensitive	less sensitive
Can be used with ordinal data	no	yes
Assumes data are from a normal distribution	yes	no
Detects non-linear relationship	no	yes

Biomarker and alcohol consumption



PEth Biomarker of alcohol consumption vs. days drinking (raw data vs. ranks)



Pearson and Spearman correlations

```
. pwcorr peth21_18and16 daysdrank_21, obs sig

             | peth2~16 days~_21
-------------+------------------
peth21_18~16 |   1.0000
             |
             |       77
             |
daysdrank_21 |   0.4717   1.0000
             |   0.0000
             |       77       85
             |
-----------------------------------
. spearman peth21_18and16 daysdrank_21

 Number of obs =      77
Spearman's rho =       0.7413

Test of Ho: peth21_18and16 and daysdrank_21 are independent
    Prob > |t| =       0.0000
```

Association Identification

Which scatterplot below shows

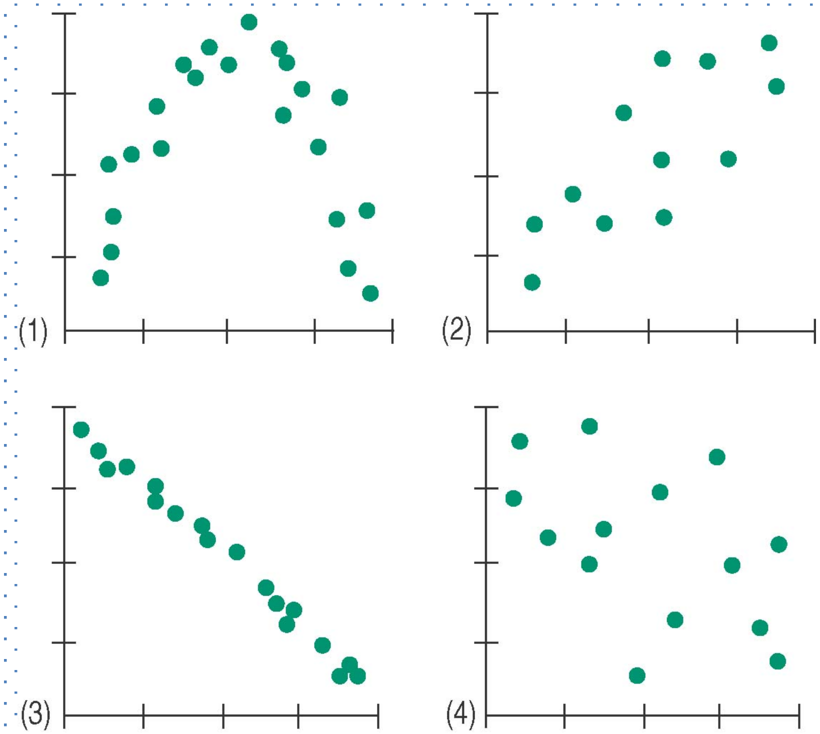
a) **little association?**

b) **negative association?**

c) **linear association?**

d) **moderately strong association?**

e) **strong association?**



Association Identification

Matching.

Here are several scatterplots.

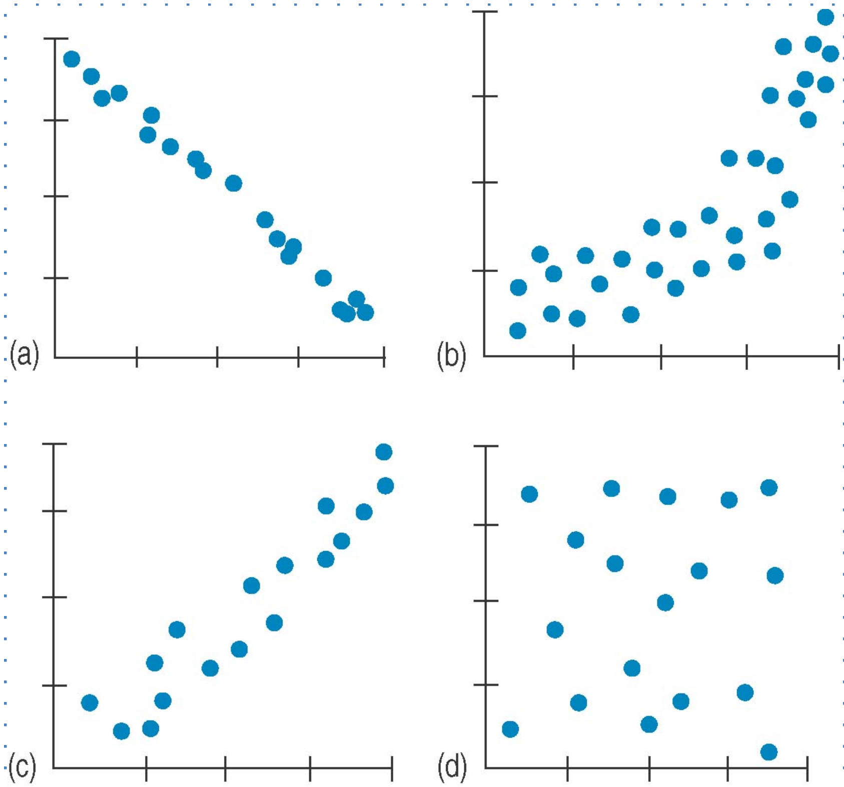
The calculated correlations are:

a) –0.977

b) –0.021

c) 0.736

d) 0.951.



Which is which?

http://www.tylervigen.com/spurious-correlations

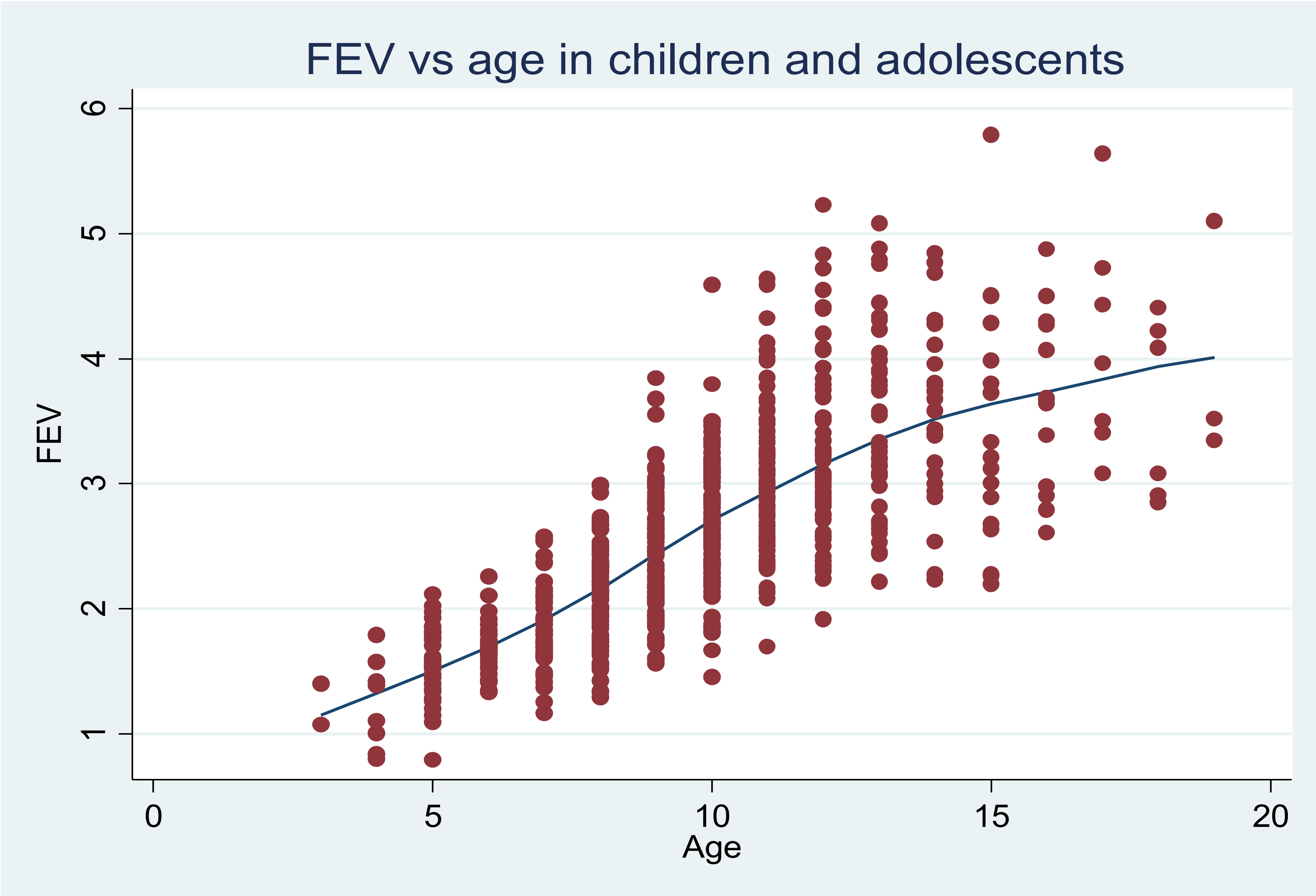
Simple linear regression

- Correlation allows us to quantify a linear relationship between two variables
- Linear regression allows us to additionally estimate how a change in a random variable X corresponds to a change in random variable Y

Forced expiratory volume (FEV)

- Studies in the 1970's of children and adolescent's pulmonary function, examining their own smoking and secondhand smoke
- FEV is the amount of air expelled in the first second of exhalation; forced expiratory volume
- The data are cross-sectional data from a larger prospective study

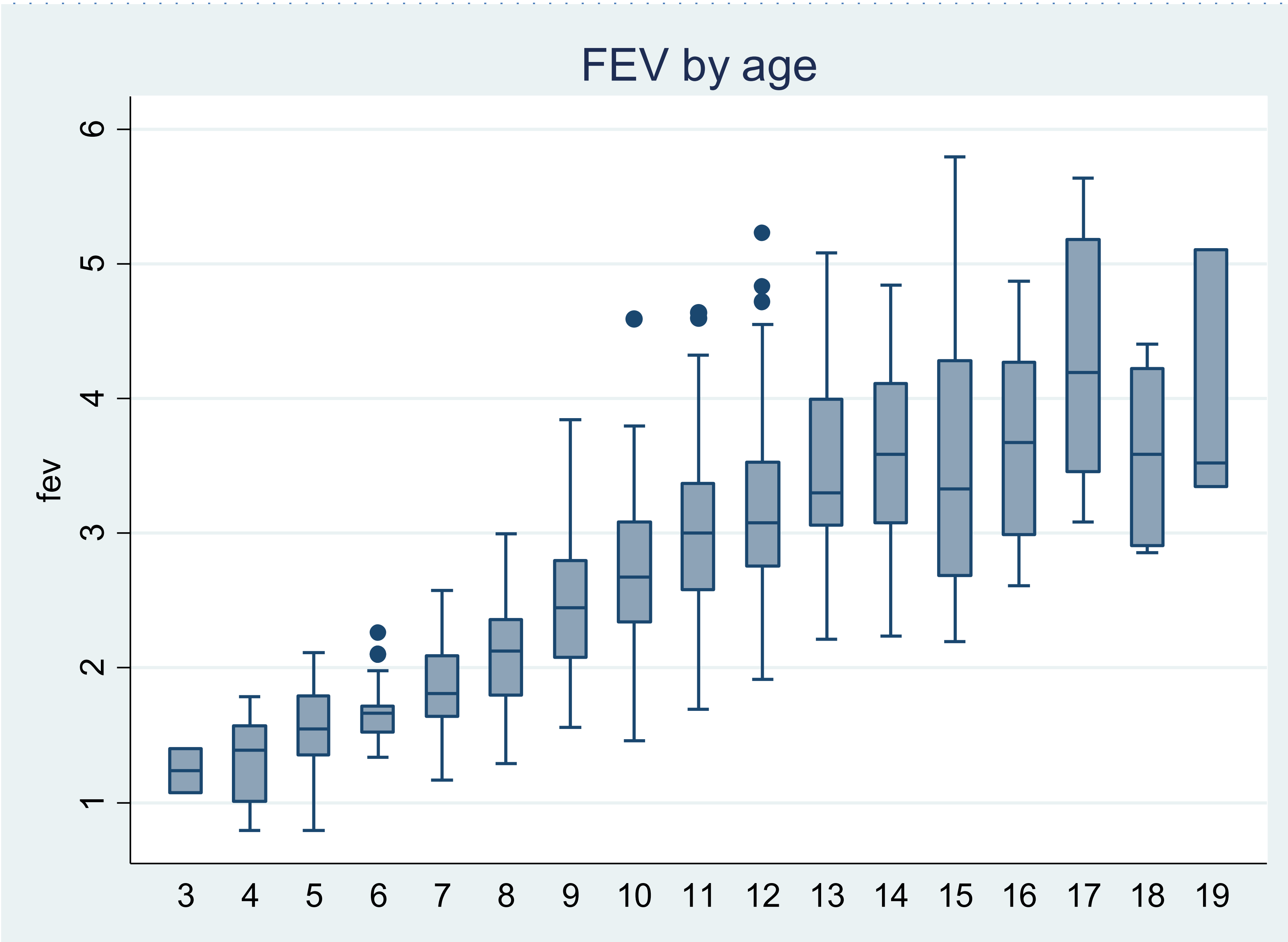
- Tager, I., Weiss, S., Munoz, A., Rosner, B., and Speizer, F. (1983), "Longitudinal Study of the Effects of Maternal Smoking on Pulmonary Function," *New England Journal of Medicine*, 309(12), 699-703.
- Tager, I., Weiss, S., Rosner, B., and Speizer, F. (1979), "Effect of Parental Cigarette Smoking on the Pulmonary Function of Children," *American Journal of Epidemiology*, 110(1), 15-26.



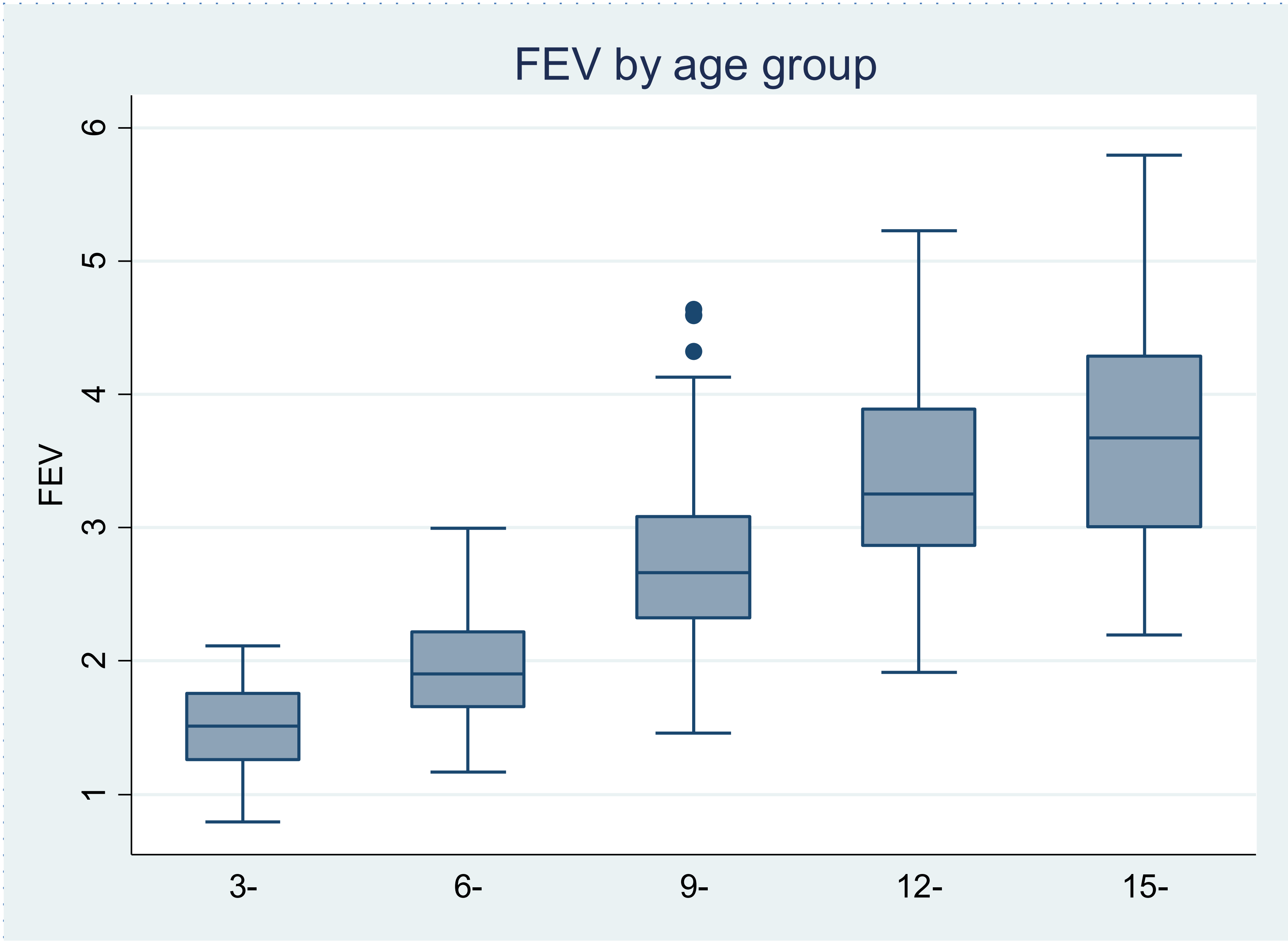

```
twoway (lowess fev age, bwidth(0.8)) (scatter fev age, sort), ytitle(FEV)
xtitle(Age)  legend(off) title(FEV vs age in children and adolescents)
```

Correlation

```
. pwcorr fev age, sig obs

             |      fev      age
-------------+------------------
         fev |   1.0000
             |
             |      654
             |
         age |   0.7565   1.0000
             |   0.0000
             |      654      654
             |


. spearman fev age

 Number of obs =     654
Spearman's rho =       0.7984

Test of Ho: fev and age are independent
    Prob > |t| =       0.0000
```




```
graph box fev, over(age) title(FEV by age)
```




```
egen agecat = cut(age), at(0,3,6,9,12,15,20) label
graph box fev, over(agecat) title(FEV by age group) ytitle(FEV)
```

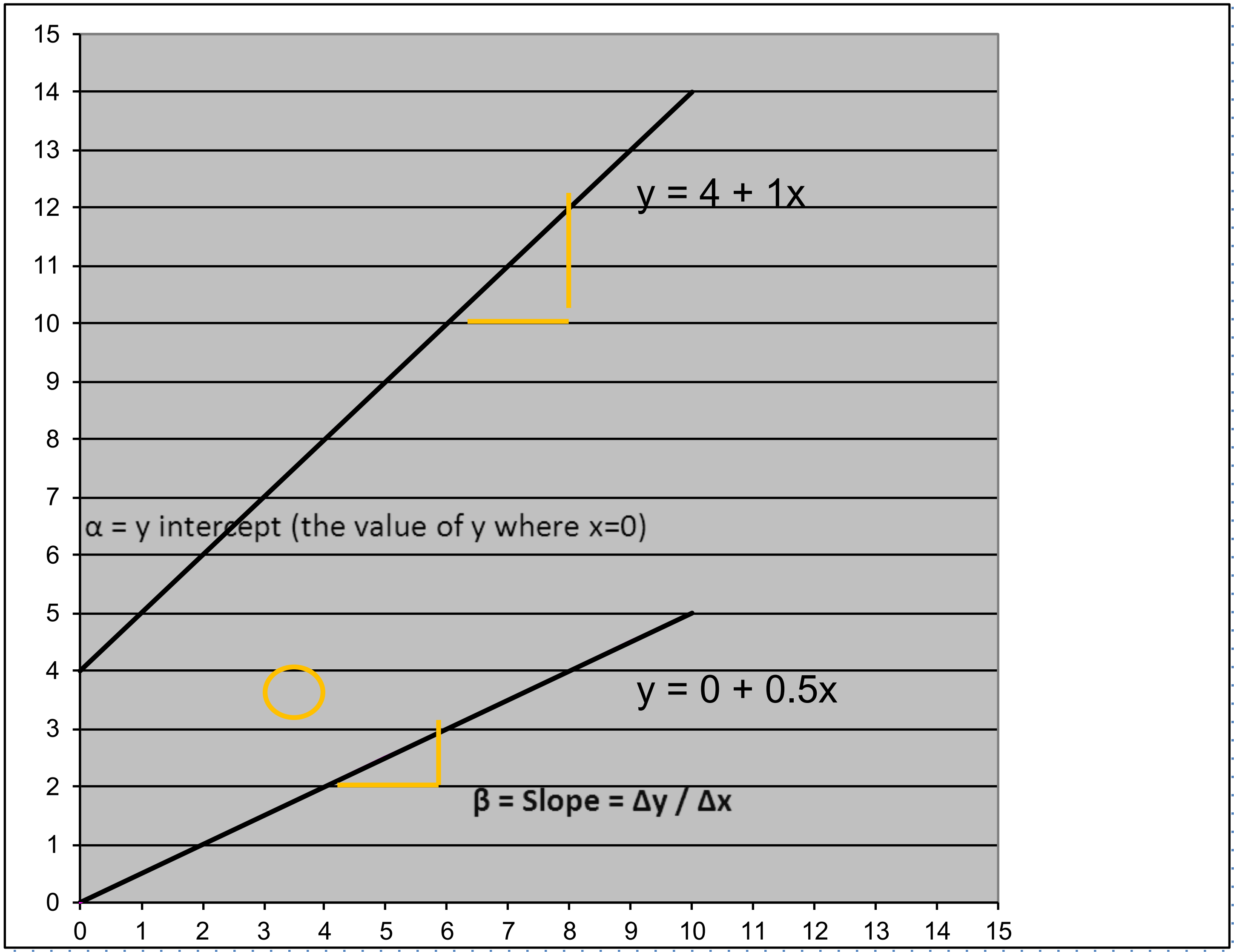
```
. tabstat fev, by(agecat) s(n min median max mean sd)

Summary for variables: fev
     by categories of: agecat

agecat |         N       min       p50       max      mean        sd
-------+------------------------------------------------------------
    3- |        39      .791     1.514     2.115  1.472385  .3346982
    6- |       176     1.165     1.901     2.993  1.943727  .3885005
    9- |       265     1.458     2.665     4.637   2.71723  .5866867
   12- |       125     1.916     3.255     5.224  3.384576  .7326963
   15- |        49     2.198     3.674     5.793  3.710143  .8818795
-------+------------------------------------------------------------
 Total |       654      .791    2.5475     5.793   2.63678  .8670591
--------------------------------------------------------------------
```

Simple linear regression

- The method allows us to investigate the effect of a difference in the *explanatory* variable on the *response* variable.
- Equivalent terms:
  - Response variable, dependent variable, outcome variable, Y
  - Explanatory variable, independent variable, predictor variable, correlate, X
- Here it matters which variable is X and which variable is Y
- Y is the variable that you want to predict, or better understand with X (we regress Y on X)

Recall the equation of a straight line

$$y = \alpha + \beta x$$



Simple linear regression

- Population regression equation is defined as $\mu_{y|x} = \alpha + \beta x$
- This is the equation of a straight line
- α and β are constants and are called the coefficients of the equation

Simple linear regression

- α is the y-intercept of the line y=$\alpha + \beta x$
- Plugging back in to the population regression equation, α is the mean value of Y when X=0

  e.g. $\mu_{y|0} = \alpha + \beta X \quad = \alpha$

- β is the slope of the line
- β is the change in the mean value of y that corresponds to a one-unit increase in X
- e.g. X=3 vs. X=2

  $\mu_{y|3} - \mu_{y|2} \quad = \quad (\alpha + \beta*3) - (\alpha + \beta*2) = \beta$

Simple linear regression

- Even if there is a linear relationship between Y and X in theory, there will (usually) be some variability in the population
- At each value of X, there is a range of Y values, with a mean $\mu_{y|x}$ and a standard deviation $\sigma_{y|x}$
- We note this by including an error term, ε, in our regression equation
- The linear regression equation is $y = \alpha + \beta x + \varepsilon$

Simple linear regression

- The linear regression equation is $y = \alpha + \beta x + \varepsilon$

- The error, ε, is the distance a sample value y has from the population regression line

  $y = \alpha + \beta x + \varepsilon$

  $\mu_{y|x} = \alpha + \beta x$ *(population regression line)*

  so $y - \mu_{y|x} = \varepsilon$ is the error

Simple linear regression

- Statistical assumptions of linear regression
  – X's are measured without error
    - Violations of this cause the coefficients to attenuate toward zero
  – For each value of X, the Y's are normally distributed with mean $\mu_{y|x}$ and standard deviation $\sigma_{y|x}$
  – The regression equation is correct $\mu_{y|x} = \alpha + \beta x$

Simple linear regression

- More assumptions of linear regression
  - Homoscedasticity – the standard deviation of y at each value of X is constant; $\sigma_{y/x}$ the same for all values of X
  - All the y_i 's are independent
    - You can't guess the y value for one person (or observation) based on the outcome of another
- We do not need the X's to be normally distributed, just the Y's at each value of X

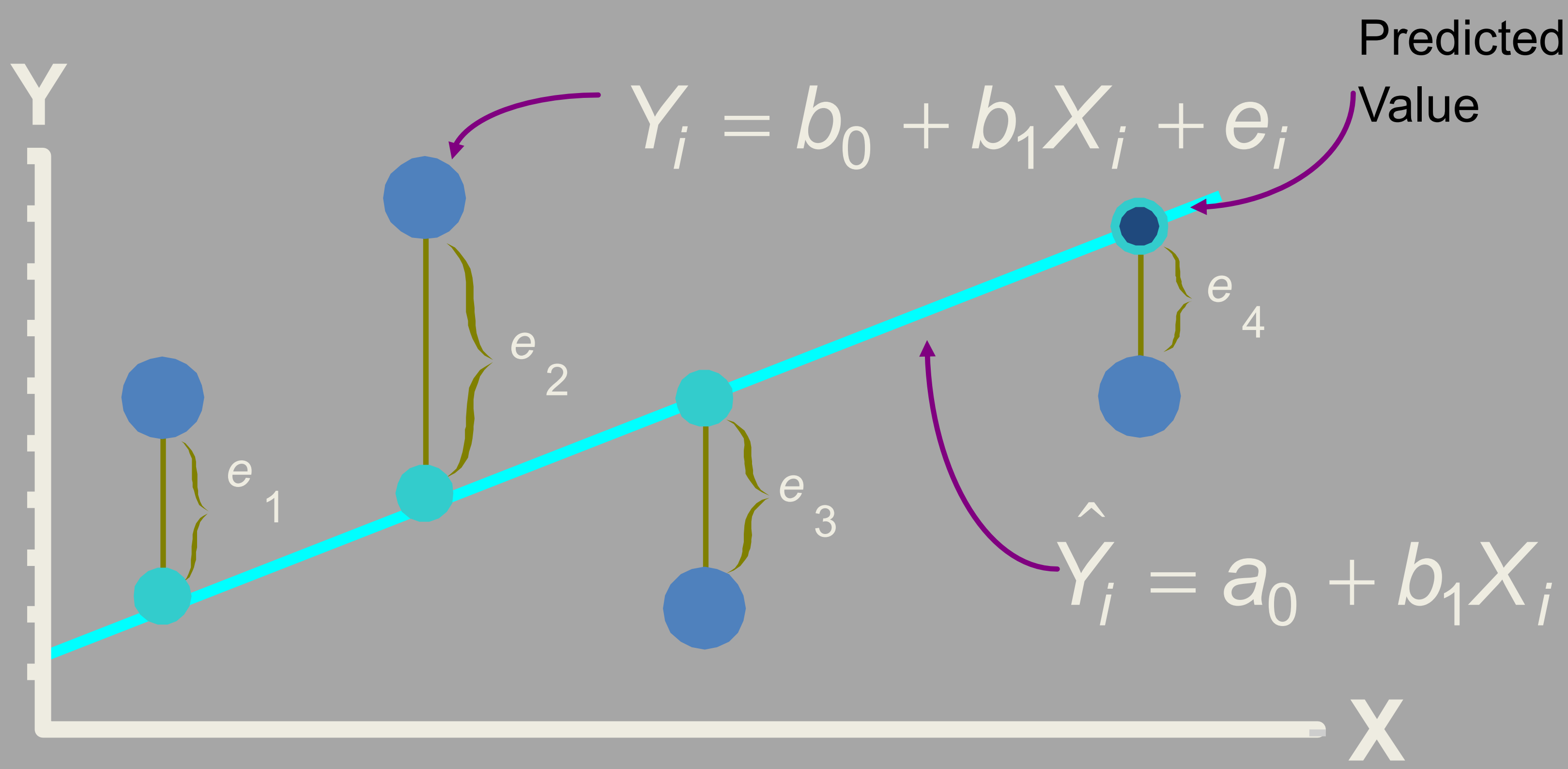
Ordinary Least Squares Regression

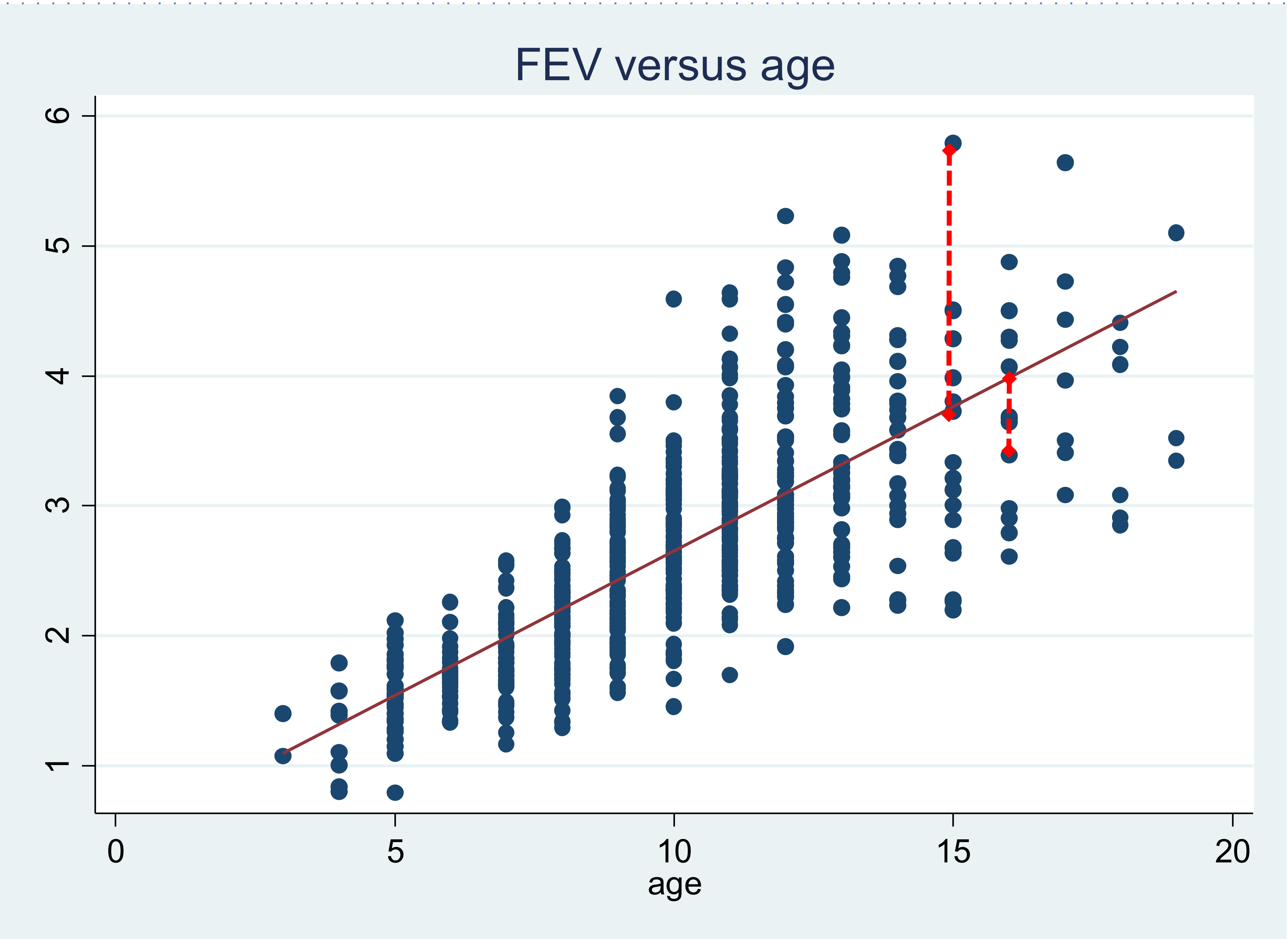
- We estimate the *coefficients* of the population regression line (α and β) using our sample of measurements of y and x
- The distance from a data point (x_i, y_i) to the line at x_i is the error and called the *residual*, e_i

$$e_i = y_i - \hat{y}_i$$

  - $\hat{y}_i$ is y-value of the regression line at x_i and is the predicted value of y at each x

Linear Regression Graphically




FEV versus age
6
5
4
3
2
1
0
5
10
15
20
age

Simple Linear Regression or Least Squares

- The regression line equation is $$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

- The "best" line is the one that finds the α and β that minimize the sum of the squared residuals Σe_i^2

- We are minimizing the sum of the squares of the residuals, called the <u>error sum of squares</u> or the <u>residual sum of squares</u>

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta}x_i)]^2$$

Simple linear regression example: Regression of FEV on age

$$FEV = \hat{\alpha} + \hat{\beta}\, age$$

regress yvar xvar

```
. regress fev age

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  1,   652) =  872.18
       Model |  280.919154     1  280.919154           Prob > F      =  0.0000
    Residual |  210.000679   652  .322086931           R-squared     =  0.5722
-------------+------------------------------           Adj R-squared =  0.5716
       Total |  490.919833   653  .751791475           Root MSE      =  .56753

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .222041   .0075185    29.53   0.000     .2072777    .2368043
       _cons |   .4316481   .0778954     5.54   0.000      .278692    .5846042
------------------------------------------------------------------------------
```

$\hat{\beta}$ = Coef for age

$\hat{\alpha}$ = _cons (short for constant)

Interpretating the parameter estimates

- The least squares estimate is

  $\hat{y} = 0.432 + 0.222\ x$

- The intercept, 0.432 is the fitted value of y (FEV) for x (age) = 0

- The slope, 0.222 is the change in FEV corresponding to a change of 1 year in age. So a child with age=10 would have an FEV that is (on average) 0.222 higher than someone age 9. And the same for age 7 vs. 6, etc.

Simple linear regression – hypothesis testing

- We want to know if there is a relationship between X and Y.
  - If there is no relationship then the value of Y does not change with the value of X, and $\beta=0$.
  - Therefore $\beta=0$ is our null hypothesis.

- This is mathematically equivalent to the null hypothesis that the correlation $\rho=0$.

- We can also calculate a 95% confidence interval for β

Inference for regression coefficients

- We want to use the least squares regression line $\hat{y}=\hat{\alpha} + \hat{\beta}x$ to make inference about the population regression line $\mu_{y|x} = \alpha + \beta x$
- If we took repeated samples in which we measured x and y together and calculated the least squares estimates, we would have a distribution for the estimates $\hat{\alpha}$ and $\hat{\beta}$
- The standard error of the estimates are

$$se(\hat{\beta}) = \frac{\sigma_{y|x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$se(\hat{\alpha}) = \sigma_{y|x}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where we estimate $\sigma_{y|x}$ with $s_{y|x}$

and $$s_{y|x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

Inference for regression coefficients

- We can use these to test the null hypothesis H_0: $\beta = \beta_0$ against the alternative H_0: $\beta \neq \beta_0$

- The test statistic for this is $$t = \frac{\hat{\beta} - \beta_0}{s\hat{e}(\hat{\beta})}$$

- And it follows the t distribution with n-2 degrees of freedom under the null hypothesis

Inference for regression coefficients

- When β_0=0 , i.e. testing H_0: β =0 this is equivalent to testing $\mu_{y|x} = \alpha + 0*x = \alpha$
- This is the same as testing the null hypothesis H_0: ρ=0
- The regression slope and the correlation coefficient are related: $\hat{\beta} = r\left(\frac{s_y}{s_x}\right)$
- 95% confidence intervals for β

$(\hat{\beta} - t_{n-2,.025} se(\hat{\beta}) , \hat{\beta} + t_{n-2,.025} se(\hat{\beta}))$

Simple linear regression example: Regression of FEV on age

$$FEV = \hat{\alpha} + \hat{\beta}\ age$$

```
regress yvar xvar

. regress fev age

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  1,   652) =  872.18
       Model |  280.919154     1  280.919154           Prob > F      =  0.0000
    Residual |  210.000679   652  .322086931           R-squared     =  0.5722
-------------+------------------------------           Adj R-squared =  0.5716
       Total |  490.919833   653  .751791475           Root MSE      =  .56753

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .222041   .0075185    29.53   0.000     .2072777    .2368043
       _cons |   .4316481   .0778954     5.54   0.000      .278692    .5846042
------------------------------------------------------------------------------
```

Coefficient of Determination R^2

- $R^2 = r^2$, i.e. just the Pearson correlation coefficient squared
- R^2 ranges from 0 to 1, and measures the proportion of the variability in y that is explained by the regression of y on x

$$R^2 = \frac{s_y^2 - s_{y|x}^2}{s_y^2}$$

Simple linear regression example: Regression of FEV on age

FEV= $\hat{\alpha} + \hat{\beta}$ age

```
regress yvar xvar

. regress fev age

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  1,   652) =  872.18
       Model |  280.919154     1  280.919154           Prob > F      =  0.0000
    Residual |  210.000679   652  .322086931           R-squared     =  0.5722
-------------+------------------------------           Adj R-squared =  0.5716
       Total |  490.919833   653  .751791475           Root MSE      =  .56753

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .222041   .0075185    29.53   0.000     .2072777    .2368043
       _cons |   .4316481   .0778954     5.54   0.000      .278692    .5846042
------------------------------------------------------------------------------
```

Evaluating the regression model

model sum of squares $= MSS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

```
regress fev age

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  1,   652) =  872.18
       Model |  280.919154     1  280.919154           Prob > F      =  0.0000
    Residual |  210.000679   652  .322086931           R-squared     =  0.5722
-------------+------------------------------           Adj R-squared =  0.5716
       Total |  490.919833   653  .751791475           Root MSE      =  .56753

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .222041   .0075185    29.53   0.000     .2072777    .2368043
       _cons |   .4316481   .0778954     5.54   0.000      .278692    .5846042
------------------------------------------------------------------------------
```

$=.7565^2$

residual sum of squares $= RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

total sum of squares $= TSS = MSS + RSS$

$= \sum_{i=1}^{n} (y_i - \bar{y})^2$

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{MSS}{TSS}$$

$$F_{statistic}(\text{MSS df, RSS df}) = \frac{\left(\frac{\text{MSS}}{\text{MSS df}}\right)}{\left(\frac{\text{RSS}}{\text{RSS df}}\right)}$$

$$MSS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

```
regress fev age

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  1,   652) =  872.18
       Model |  280.919154     1  280.919154           Prob > F      =  0.0000
    Residual |  210.000679   652  .322086931           R-squared     =  0.5722
-------------+------------------------------           Adj R-squared =  0.5716
       Total |  490.919833   653  .751791475           Root MSE      =  .56753

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .222041   .0075185    29.53   0.000     .2072777    .2368043
       _cons |   .4316481   .0778954     5.54   0.000      .278692    .5846042
------------------------------------------------------------------------------
```

$=.7565^2$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{MSS}{TSS}$$

Overall Model Fit Using the F-Statistic

- The F statistic compares the model to a model with just $\bar{y}$ (the null model)
- The statistic is

$$F_{stat} = \frac{\left(\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\#\text{ indep vars}}\right)}{\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\text{n - \# indep vars}}\right)} = \frac{\left(\frac{MSS}{\text{MSS df}}\right)}{\left(\frac{RSS}{\text{RSS df}}\right)}$$

- When there is only one independent variable in the model, these are equivalent tests
  - F test that compares the model fit to the null model
  - The ttest that β=0
  - The ttest that r=0 (Pearson correlation)

Inference for predicted values

- We might want to estimate the mean value of y at a particular value of x
- What is the mean FEV for children who are 10 years old? $\hat{y} = .432 + .222*x = .432 + .222*10 = 2.643$ liters
- We can construct a 95% confidence interval for the estimated mean
  - $(\hat{y} - t_{n-2,.025}se(\hat{y}), \hat{y} + t_{n-2,.025}se(\hat{y}))$, where

$$s\hat{e}(\hat{y}) = s_{y|x}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$\text{where } s_{y|x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{n-2}} = \sqrt{\frac{RSS}{n-2}}$$

Use Stata for your prediction

- Stata will calculate the fitted regression values and the standard errors
  - `regress fev age`
  - `predict fev_pred, xb` -> predicted mean values ($\hat{y}$)
  - `predict fev_predse, stdp` -> se of $\hat{y}$ values

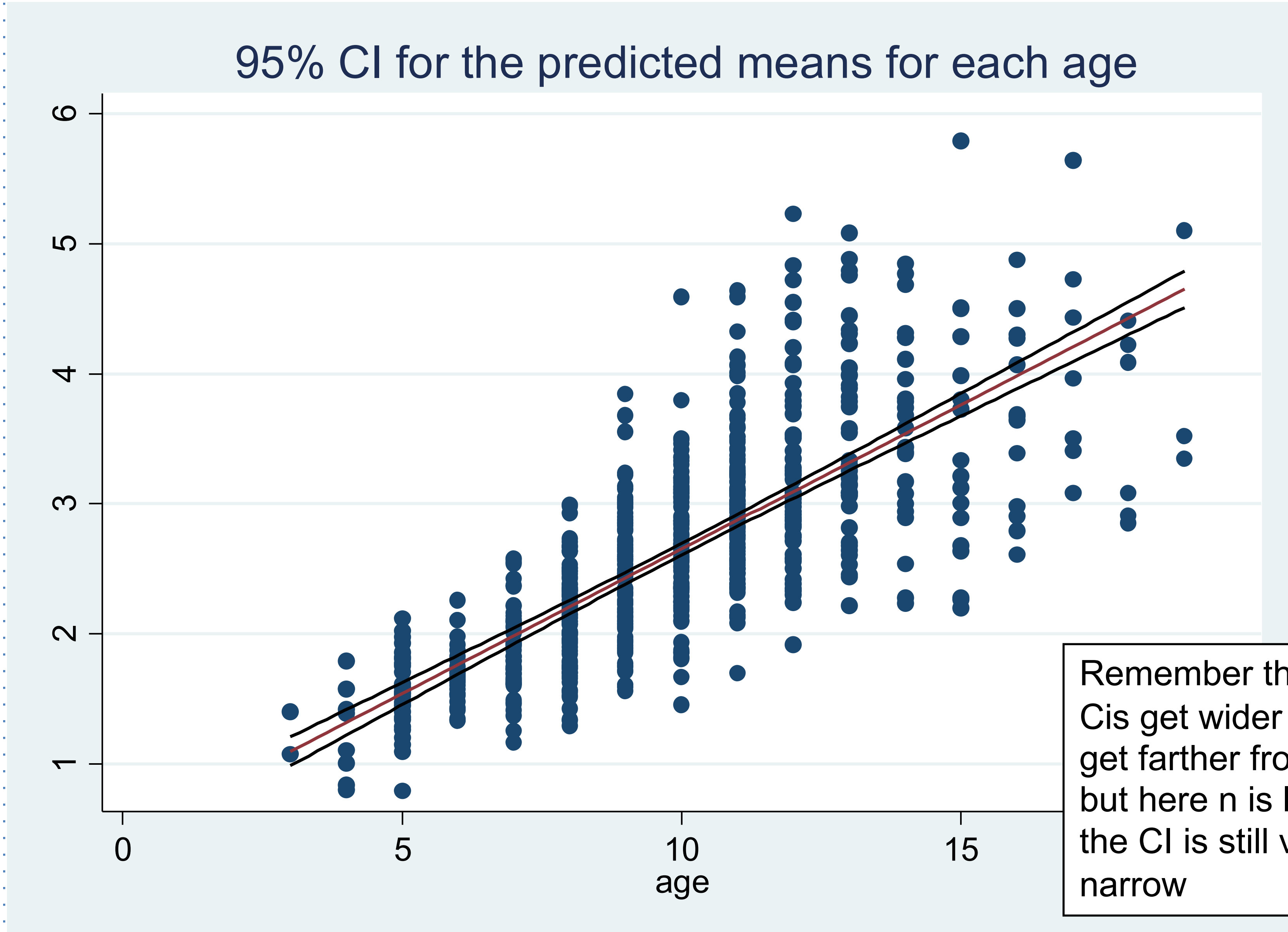
New variable names that I made up

You don't have to calculate these to get a plot with the 95% CI:

```
twoway (lfitci fev age)
```

Stands for linear fit for the line, CI for the confidence interval

```
. list fev age fev_pred fev_predse

     +-----------------------------------------+
     |   fev   age   fev_pred   fev_pr~e |
     |-----------------------------------------|
  1. | 1.708     9   2.430017   .0232702 |
  2. | 1.724     8   2.207976   .0265199 |
  3. |  1.72     7   1.985935   .0312756 |
  4. | 1.558     9   2.430017   .0232702 |
  5. | 1.895     9   2.430017   .0232702 |
     |-----------------------------------------|
  6. | 2.336     8   2.207976   .0265199 |
  7. | 1.919     6   1.763894   .0369605 |
  8. | 1.415     6   1.763894   .0369605 |
  9. | 1.987     8   2.207976   .0265199 |
 10. | 1.942     9   2.430017   .0232702 |
     |-----------------------------------------|
 11. | 1.602     6   1.763894   .0369605 |
 12. | 1.735     8   2.207976   .0265199 |
 13. | 2.193     8   2.207976   .0265199 |
 14. | 2.118     8   2.207976   .0265199 |
 15. | 2.258     8   2.207976   .0265199 |

336. | 3.147    13   3.318181   .0320131 |
337. |  2.52    10   2.652058   .0221981 |
338. | 2.292    10   2.652058   .0221981 |
```



Remember that the Cis get wider as you get farther from $\bar{x}$; but here n is large so the CI is still very narrow

```
twoway (scatter fev age)  (lfitci fev age, ciplot(rline)
blcolor(black)), legend(off) title(95% CI for the predicted
means for each age )
```

Next Time: More and More Modeling

- More linear regression
- More prediction
- Some non-linear regression
- Logistic regression