Associations, Correlations, and Causation

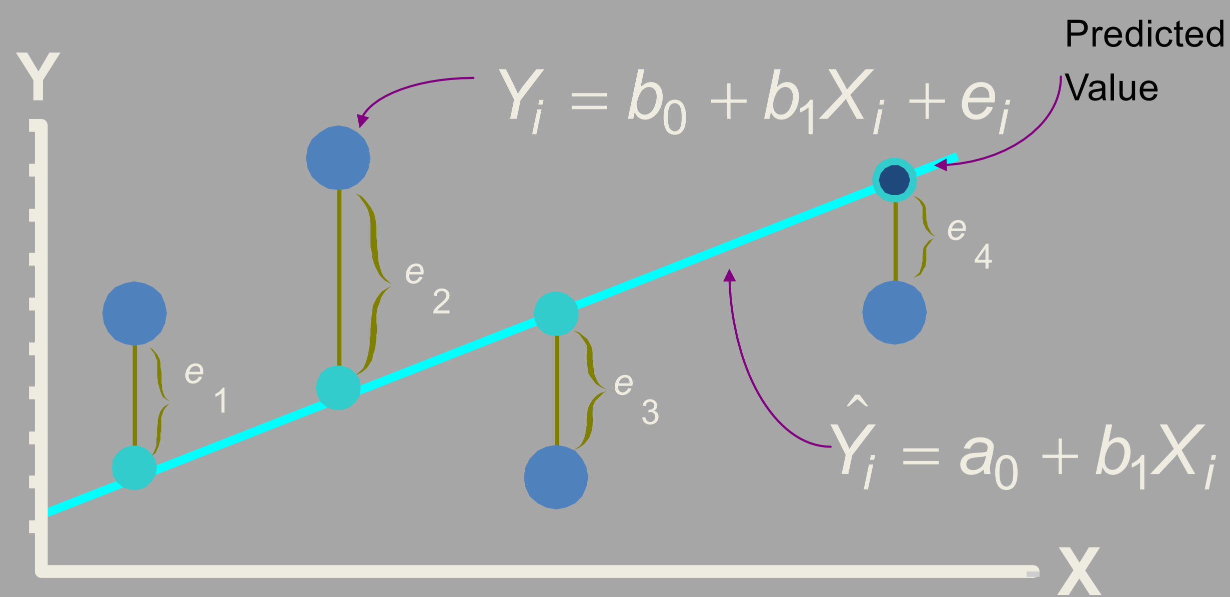
Ordinary Least Squares Regression

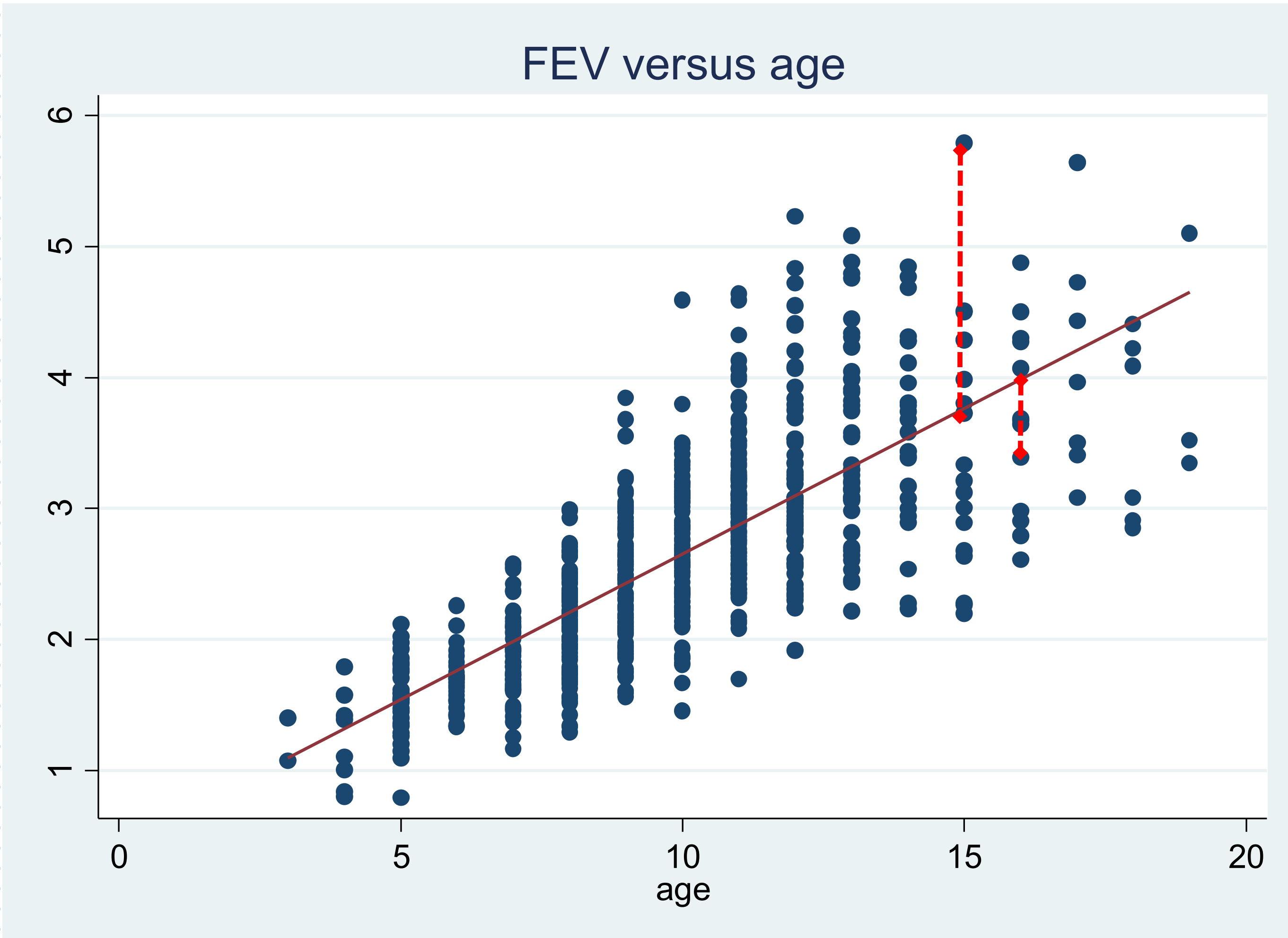
- We estimate the coefficients of the population regression line (α and β) using our sample of measurements of y and x
- The distance from a data point (x_i, y_i) to the line at x_i is the error and called the residual, e_i

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

  - $\hat{y}_i$ 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 error sum of squares or the residual sum of squares

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

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

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

```
 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 $= TSS = MSS + RSS$

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

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

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

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

```
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

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

- 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:

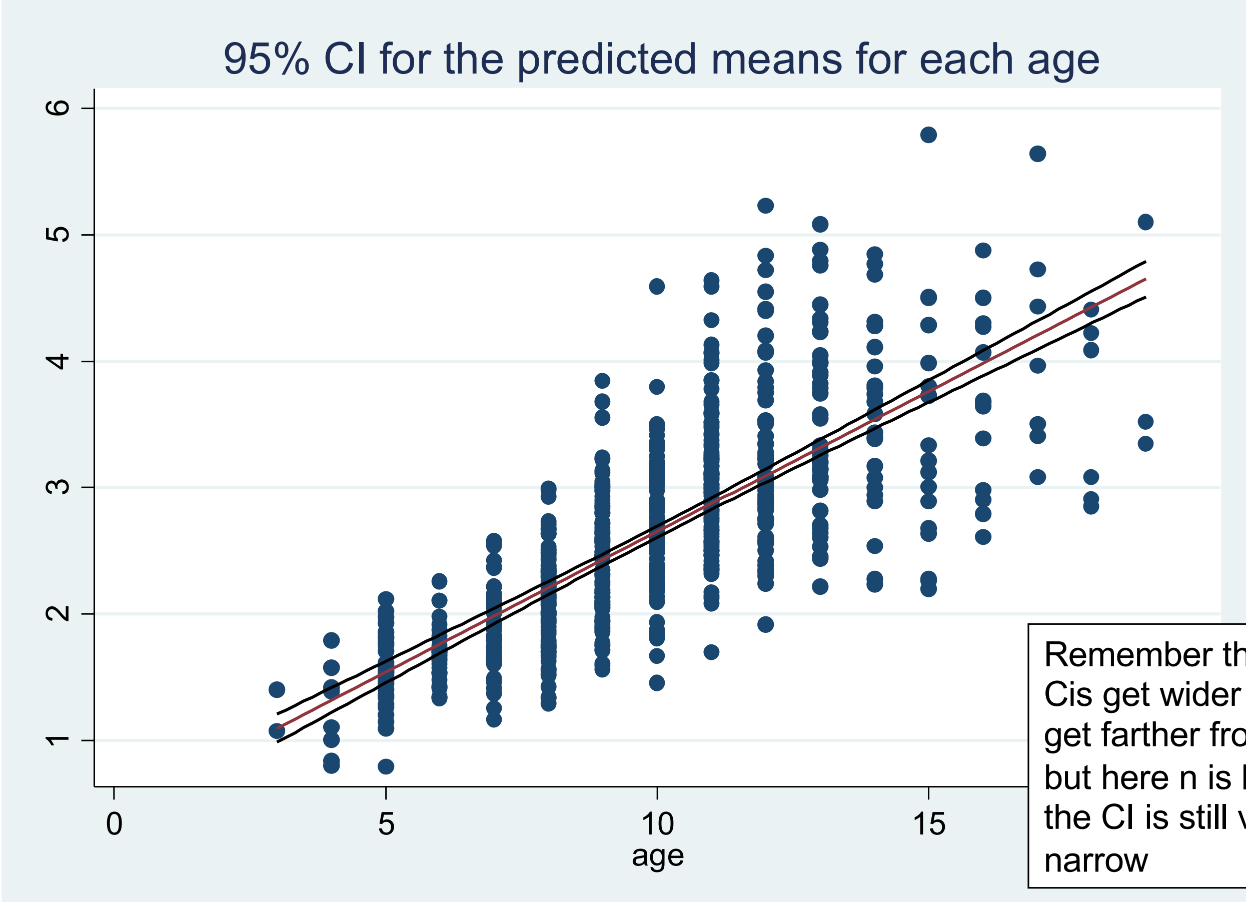
```
twoway (lfitci fev age)
```

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 )
```