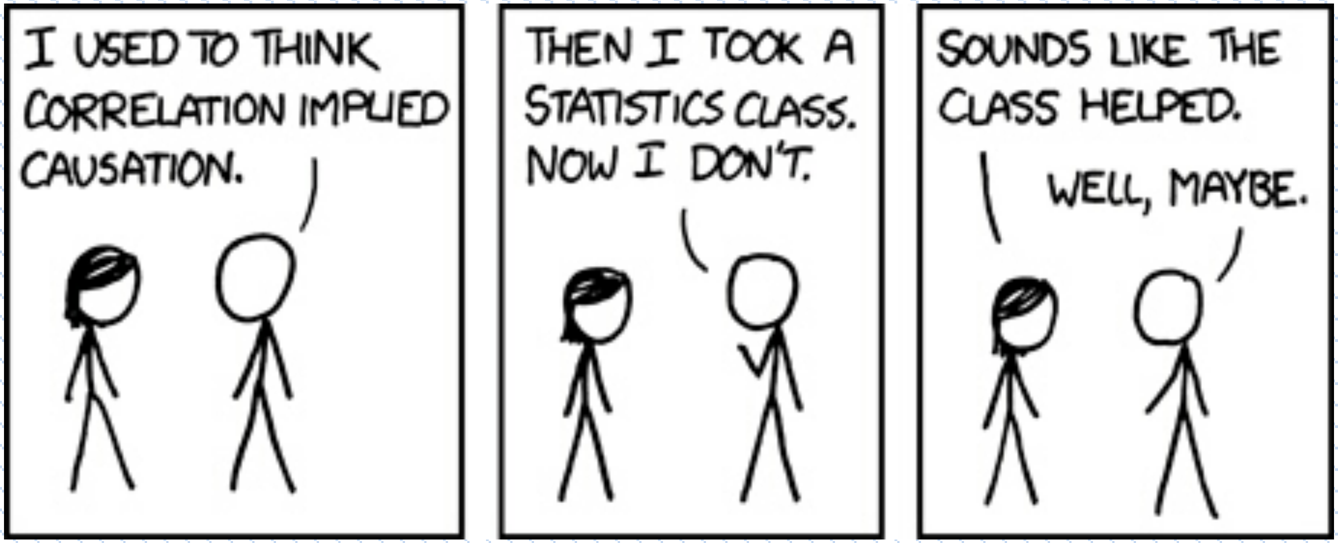
Biostat 200

Linear and Logistic Models



Correlation is not Causation

If she loves you more each and every day,
by linear regression she hated you before you met.

Review: 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 (ŷ)
  - `predict fev_predse, stdp` -> se of ŷ values

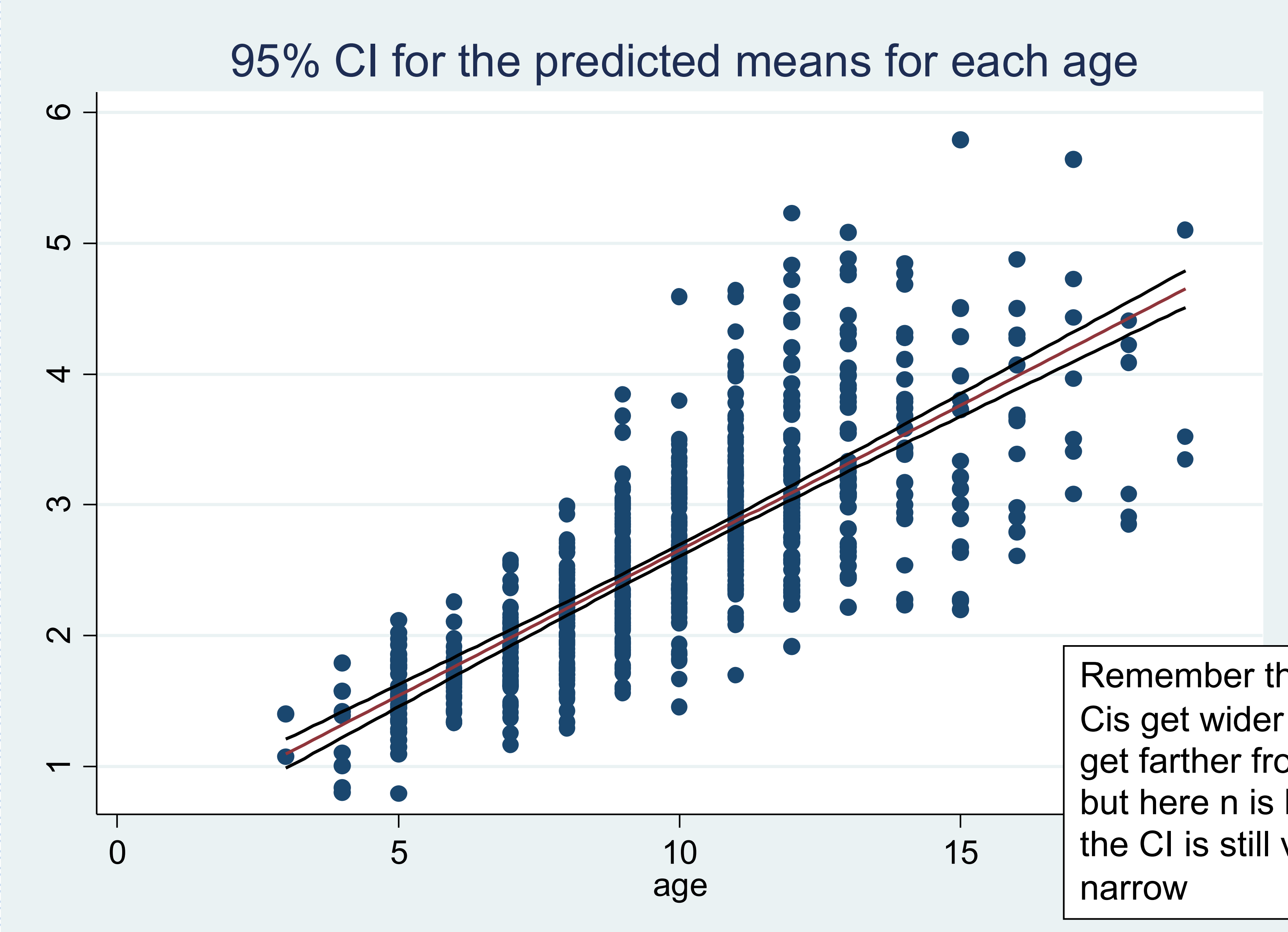
New variable names that I made up

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 )
```

Prediction intervals

- The intervals we just made were for means of y at particular values of x
- What if we want to predict the FEV value for an *individual child* at age 10?
- The prediction is the same – plug into the regression equation: $\tilde{y} = .432 + .222*10 = 2.643$ liters
- But the standard error of $\tilde{y}$ is not the same as the standard error of $\hat{y}$

Prediction intervals

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

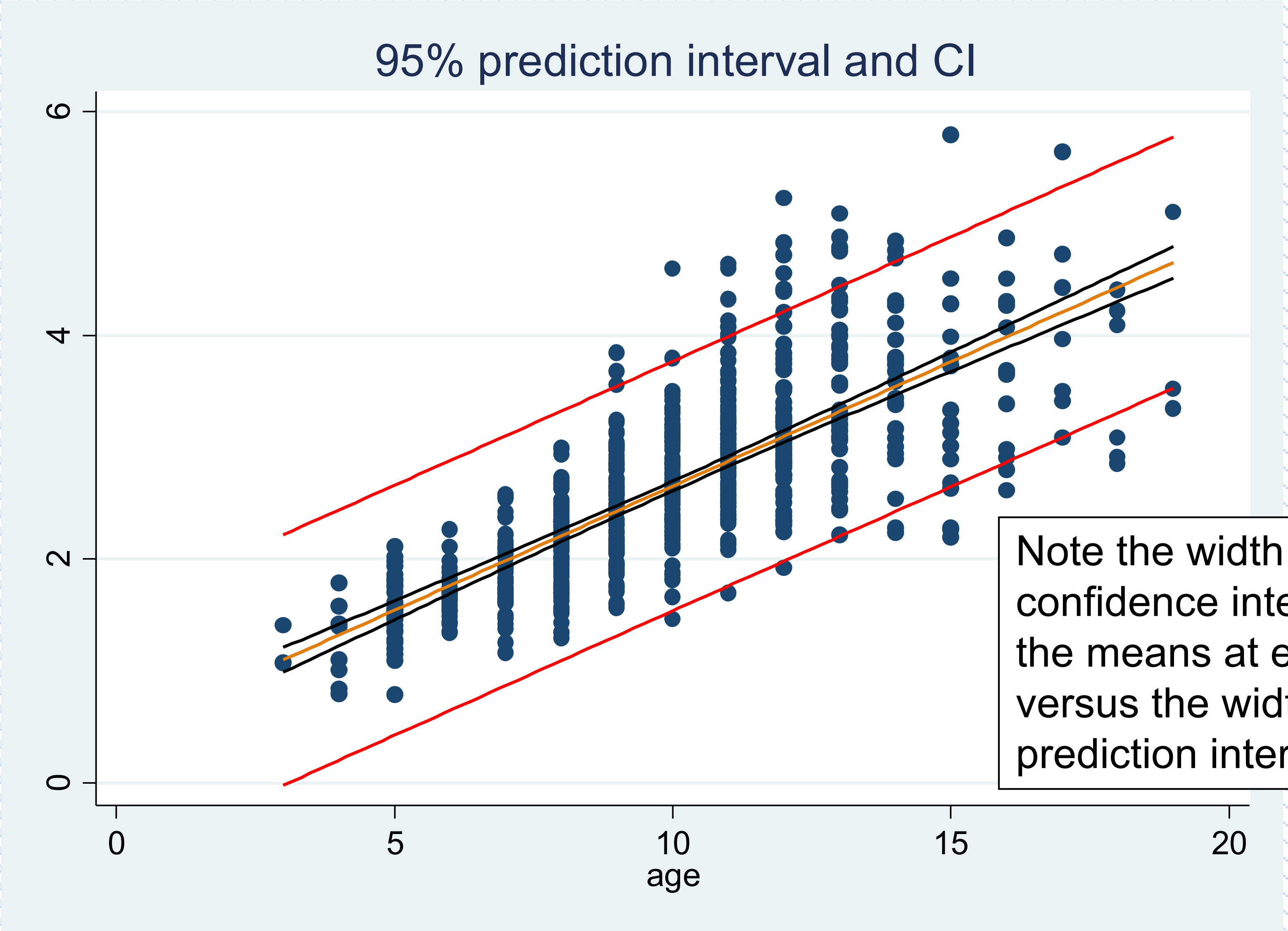
$$= \sqrt{s_{y|x}^2 + \frac{s_{y|x}^2}{n} + \frac{s_{y|x}^2(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

- This differs from the se(ŷ) only by the extra variance of y in the formula but it makes a big difference
- There is much more uncertainty in predicting a future value versus predicting a mean
- Stata will calculate these using

```
predict fev_predse_ind, stdf
```

Where the **f** is for forecast

```
. list fev age fev_pred fev_predse fev_pred_ind

     +--------------------------------------------------+
     |   fev   age   fev_pred   fev~edse    fev~ndse |
     |--------------------------------------------------|
  1. | 1.708     9   2.430017   .0232702    .5680039 |
  2. | 1.724     8   2.207976   .0265199    .5681463 |
  3. |  1.72     7   1.985935   .0312756    .5683882 |
  4. | 1.558     9   2.430017   .0232702    .5680039 |
  5. | 1.895     9   2.430017   .0232702    .5680039 |
     |--------------------------------------------------|
  6. | 2.336     8   2.207976   .0265199    .5681463 |
  7. | 1.919     6   1.763894   .0369605    .5687293 |
  8. | 1.415     6   1.763894   .0369605    .5687293 |
  9. | 1.987     8   2.207976   .0265199    .5681463 |
 10. | 1.942     9   2.430017   .0232702    .5680039 |
     |--------------------------------------------------|
 11. | 1.602     6   1.763894   .0369605    .5687293 |
 12. | 1.735     8   2.207976   .0265199    .5681463 |
 13. | 2.193     8   2.207976   .0265199    .5681463 |
 14. | 2.118     8   2.207976   .0265199    .5681463 |
 15. | 2.258     8   2.207976   .0265199    .5681463 |

336. | 3.147    13   3.318181   .0320131    .5684292 |
337. |  2.52    10   2.652058   .0221981     .567961 |
338. | 2.292    10   2.652058   .0221981     .567961 |
```

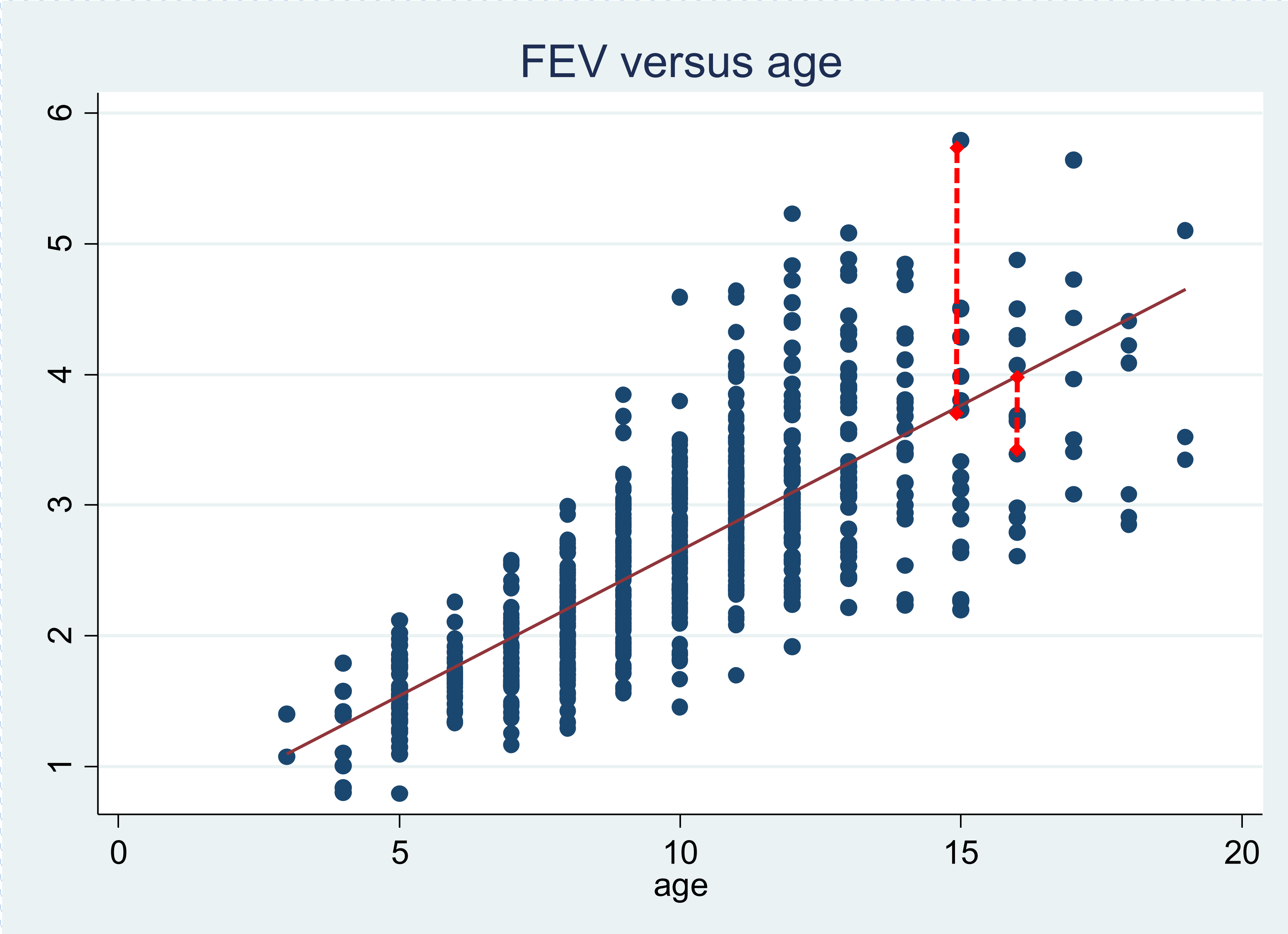


Note the width of the confidence intervals for the means at each x versus the width of the prediction intervals

```
twoway (scatter fev age)   (lfitci fev age, ciplot(rline)
blcolor(black) ) (lfitci fev age, stdf ciplot(rline)
blcolor(red) ), legend(off) title(95% prediction interval and
CI )
```

Model fit -- Residuals

- Residuals are the difference between the observed y values and the regression line for each value of x: $y_i - \hat{y}_i$
- If all the points lie along a straight line, the residuals are all 0
- If there is a lot of variability at each level of x, the residuals are large
- The sum of the squared residuals is what was minimized in the least squares method of fitting the line

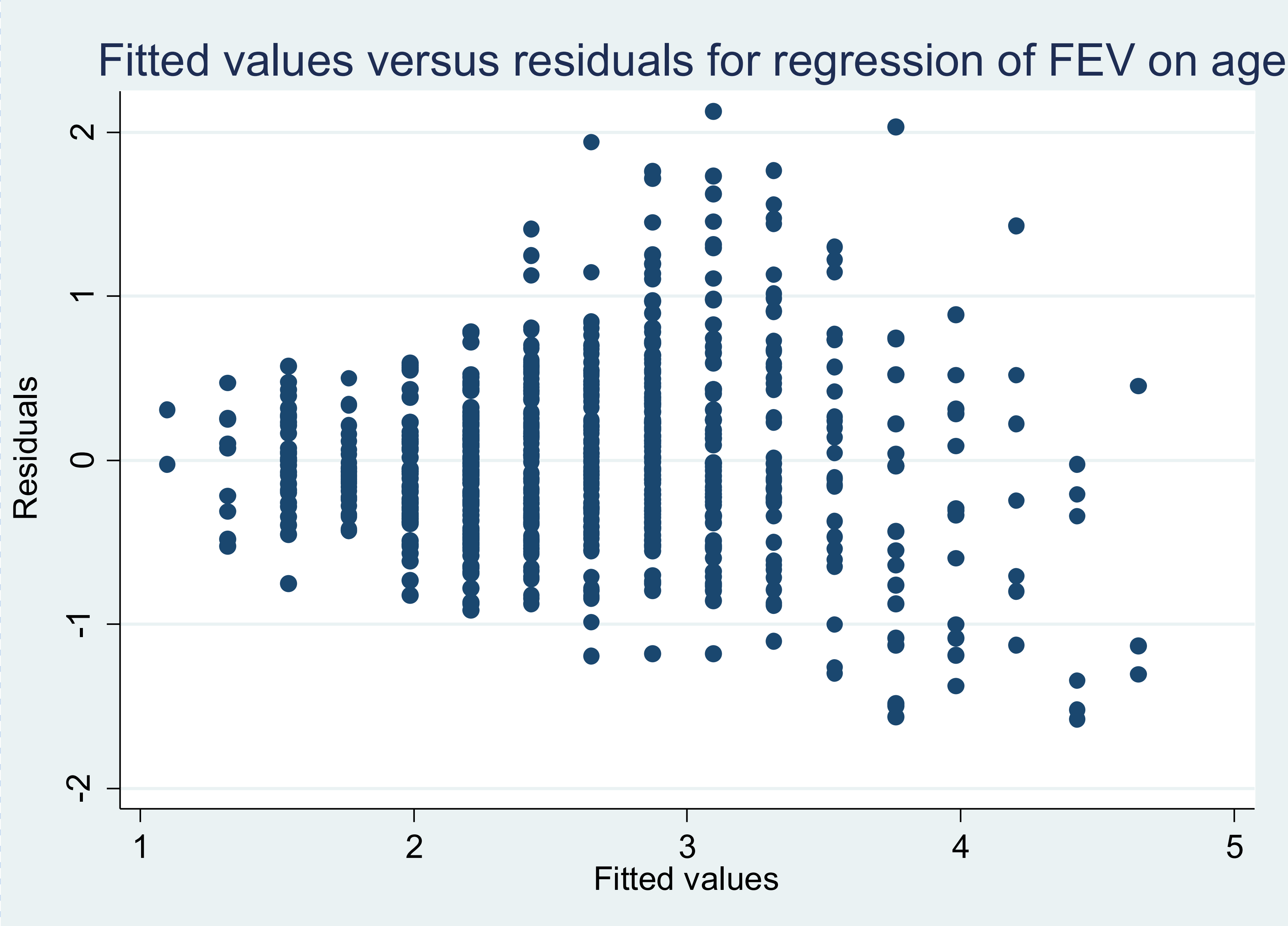

FEV versus age
6
5
4
3
2
1
0
5
10
15
20
age

Residuals

- We examine the residuals using scatter plots
- We plot the fitted values $\hat{y}_i$ on the x-axis and the residuals y_i-$\hat{y}_i$ on the y-axis
- We use the fitted values because they have the effect of the independent variable removed
- To calculate the residuals and the fitted values Stata:

```
regress fev age
rvfplot
```

- RVF = residuals versus fitted



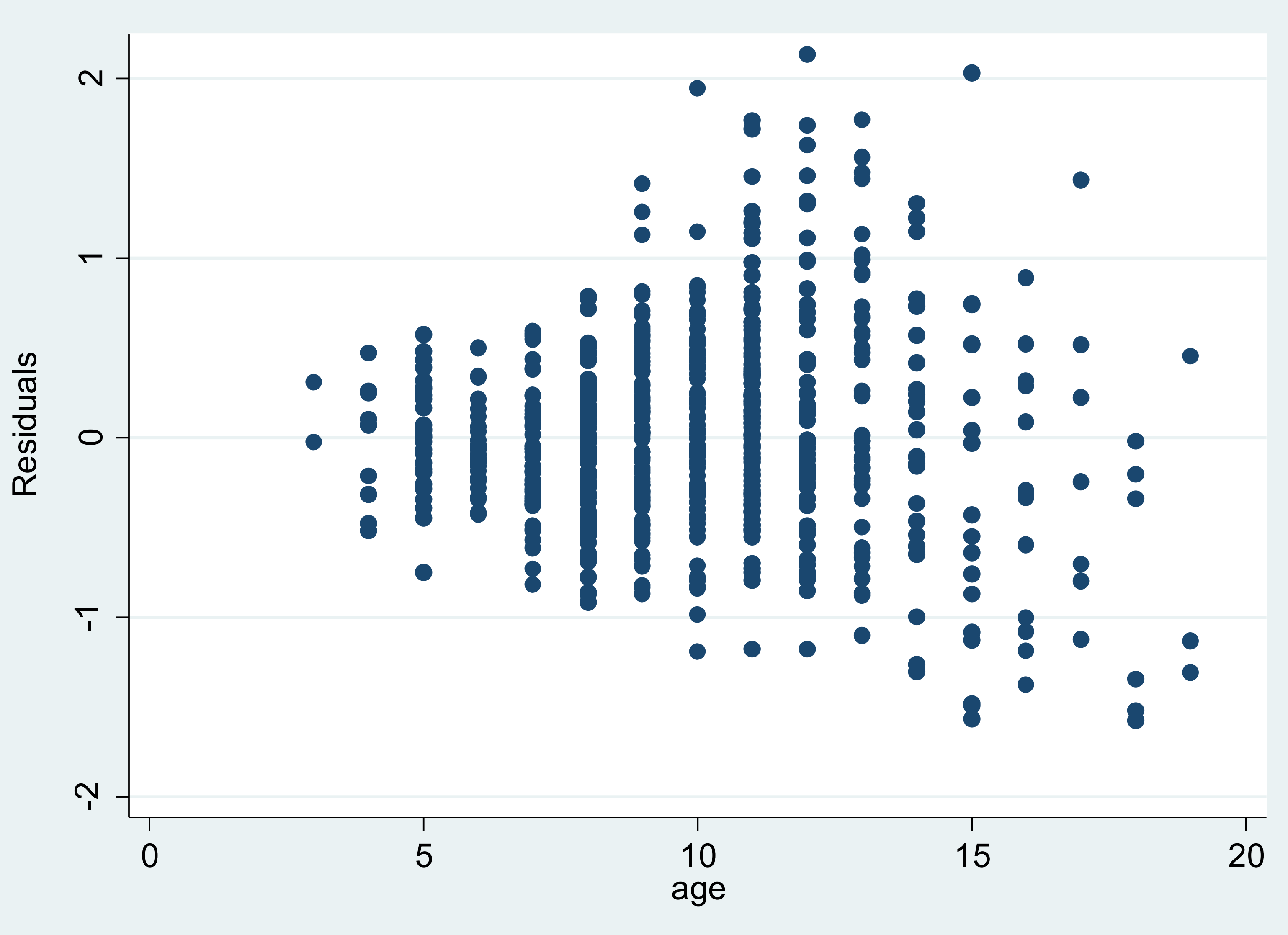

```
rvfplot, title(Fitted values versus residuals for
regression of FEV on age)
```

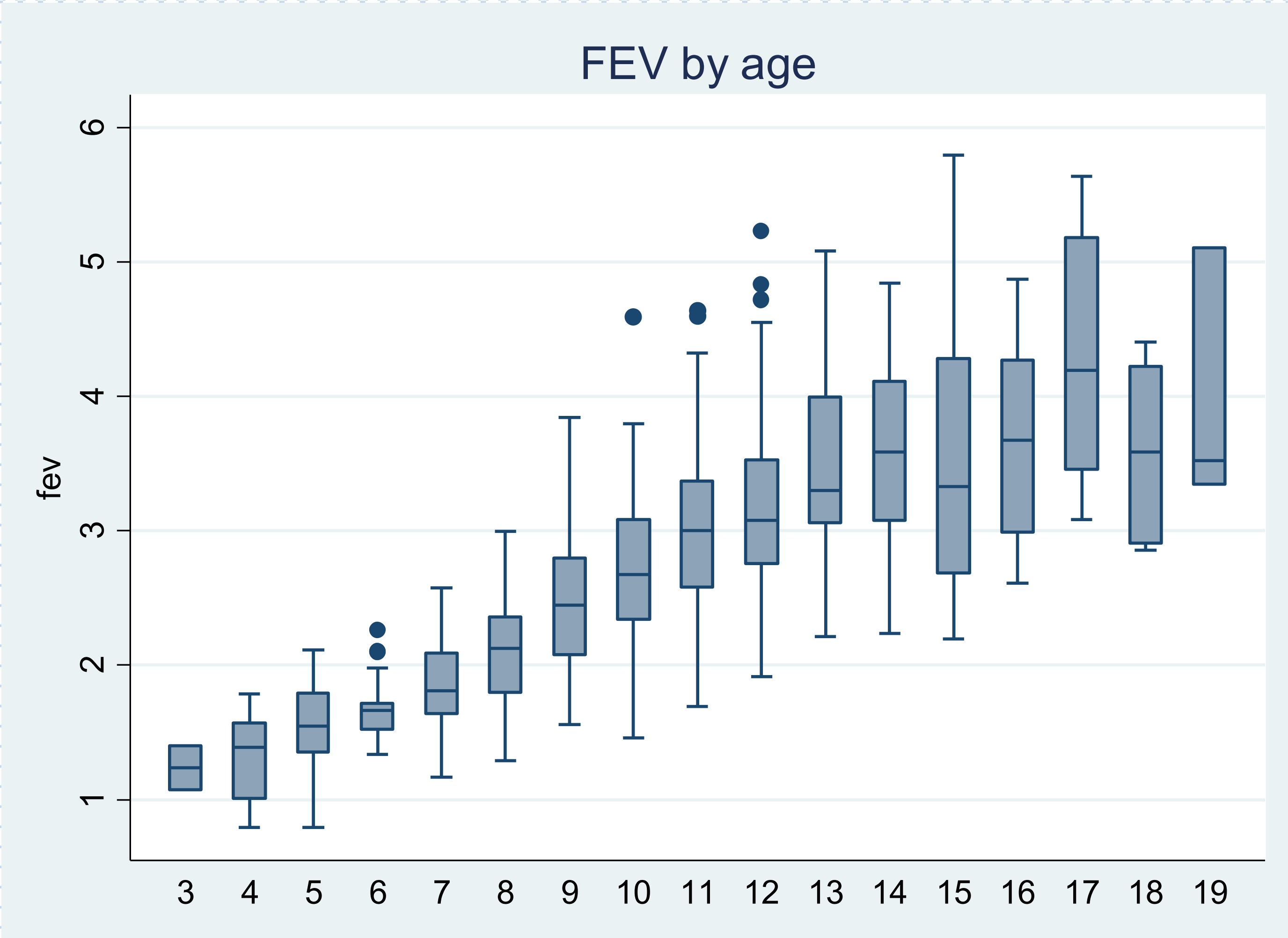
Residuals

- This plot shows that as the fitted value of FEV increases, the spread of the residuals increase – this suggests heteroscedasticity
- We would get a similar plot if we plotted age on the x-axis

```
rvpplot age, name(res_v_age)
```

- RVP=residuals versus predictor
- We had a hint of this when looking at the box plots of FEV by age groups in the previous lecture


Residuals
-2
-1
0
1
2
0
5
10
15
20
age




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

Note that heteroscedasticity does not bias the estimates of the parameters, but it does reduce the precision of the estimates (and therefore reduces power)

Transformations

- One way to deal with this is to transform either x or y or both
- A common transformation is the log transformation
- Log transformations bring large values closer to the rest of the data
- There are methods to correct the standard errors for heteroscedasticity other than transformations

Log transformations

Value	Ln	Log_{10}
0	$-\infty$	$-\infty$
0.001	-6.91	-3.00
0.05	-3.00	-1.30
1	0.00	0.00
5	1.61	0.70
10	2.30	1.00
50	3.91	1.70
103	4.63	2.01

Be careful of log(0) or ln(0)

Be sure you know which log base your computer program is using

- In Stata log() will give you ln()
- If you want log10 use log10()

Transformations

- Transform FEV to ln(FEV)

```
. gen fev_ln=log(fev)
. summ fev fev_ln

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         fev |        654     2.63678    .8670591       .791      5.793
      fev_ln |        654     .915437    .3332652  -.2344573    1.75665
```

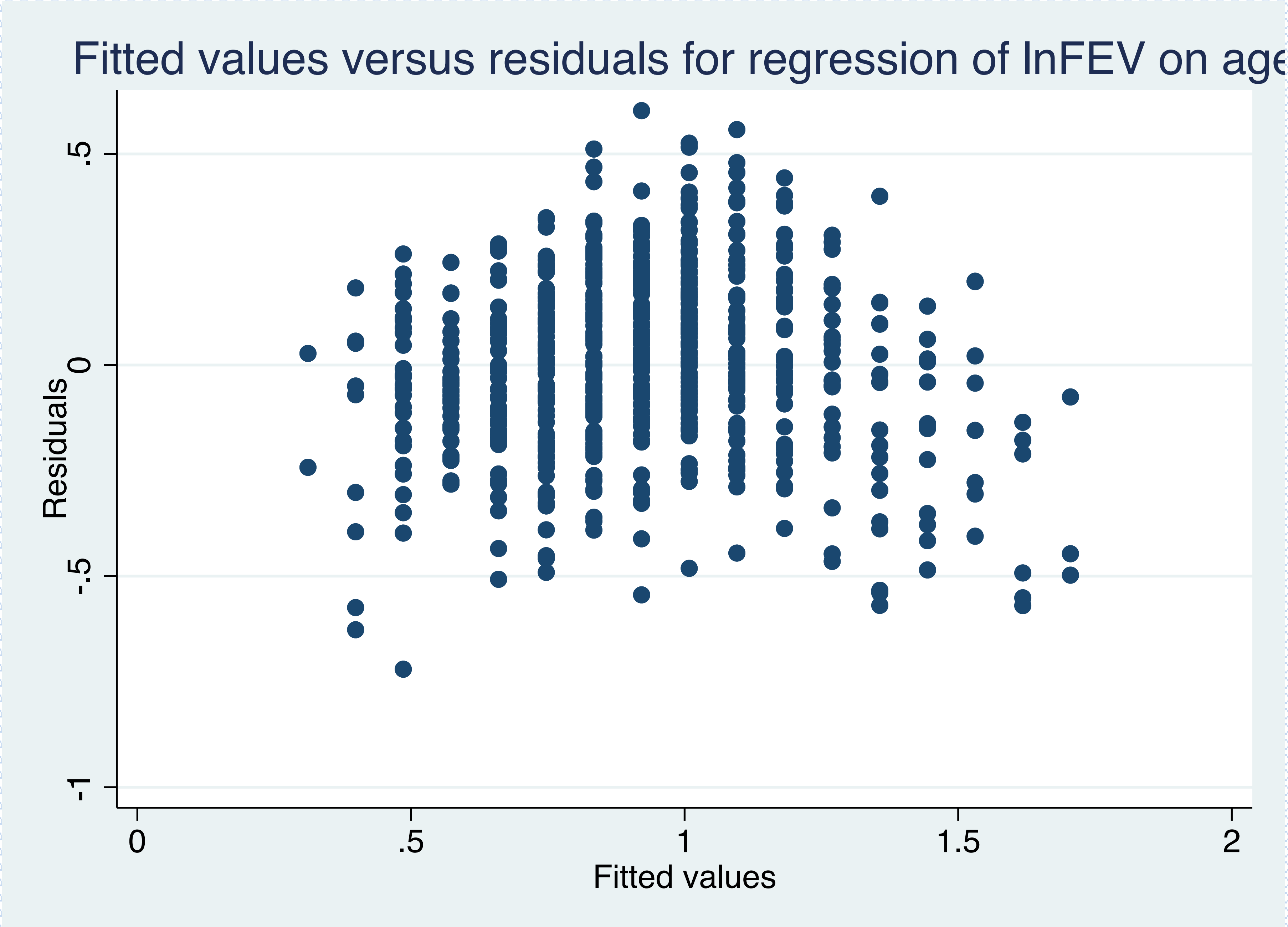
- Examine the histograms
- Run the regression of ln(FEV) on age and examine the residuals

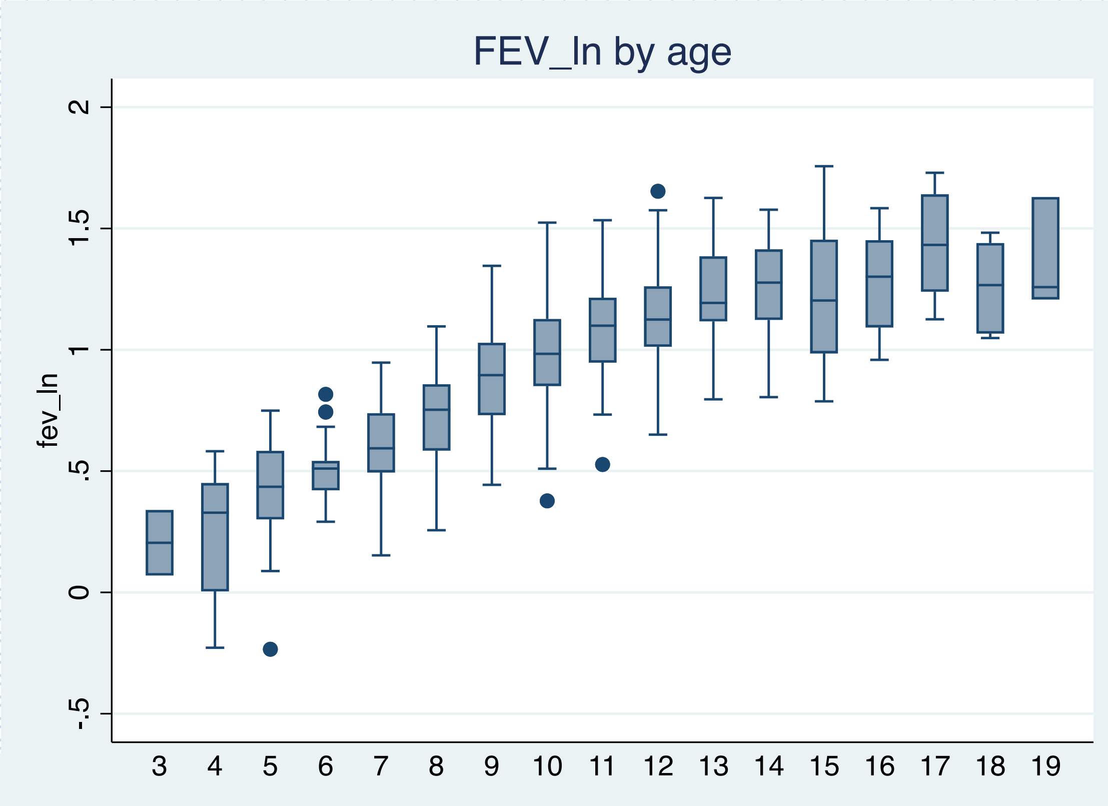
```
regress fev_ln age
rvfplot, title(Fitted values versus residuals for
   regression of lnFEV on age)
```

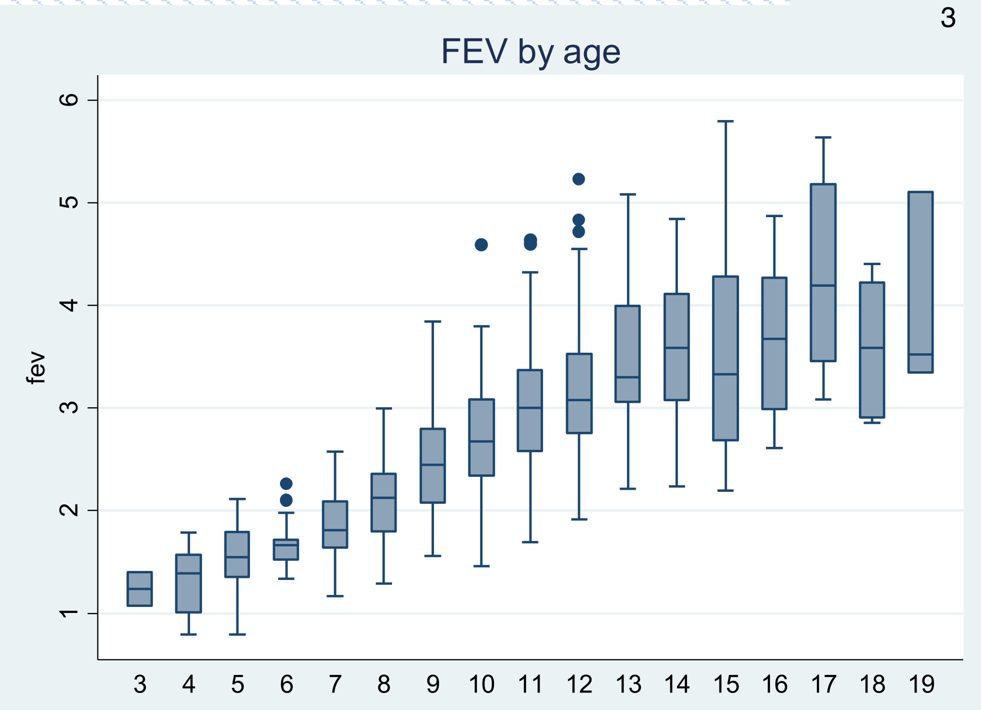
Regressions: fev & fev_ln

```
. regress fev_ln age

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(1, 652)       =    961.01
       Model |  43.2100544         1  43.2100544   Prob > F        =    0.0000
    Residual |  29.3158601       652  .044962976   R-squared       =    0.5958
-------------+----------------------------------   Adj R-squared   =    0.5952
       Total |  72.5259145       653  .111065719   Root MSE        =    .21204

------------------------------------------------------------------------------
      fev_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0870833   .0028091    31.00   0.000     .0815673    .0925993
       _cons |    .050596    .029104     1.74   0.083    -.0065529    .1077449
------------------------------------------------------------------------------
```

```
. 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
------------------------------------------------------------------------------
```


Fitted values versus residuals for regression of lnFEV on age
Residuals
.5
0
-.5
-1
0
.5
1
1.5
2
Fitted values


FEV_ln by age
fev_ln
-.5
0
.5
1
1.5
2
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19


FEV by age
fev
1
2
3
4
5
6
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Interpretation of regression coefficients for transformed y value

- The regression equation is:

$$\ln(FEV) = \hat{\alpha} + \hat{\beta} * age$$

$$= 0.051 + 0.087 \text{ age}$$

- So a one year change in age corresponds to a .087 change in ln(FEV)
- The change is on a multiplicative scale, so if you exponentiate, you get a percent change in y
- $e^{0.087} = 1.091$ or a 9.1% change in FEV

For example…

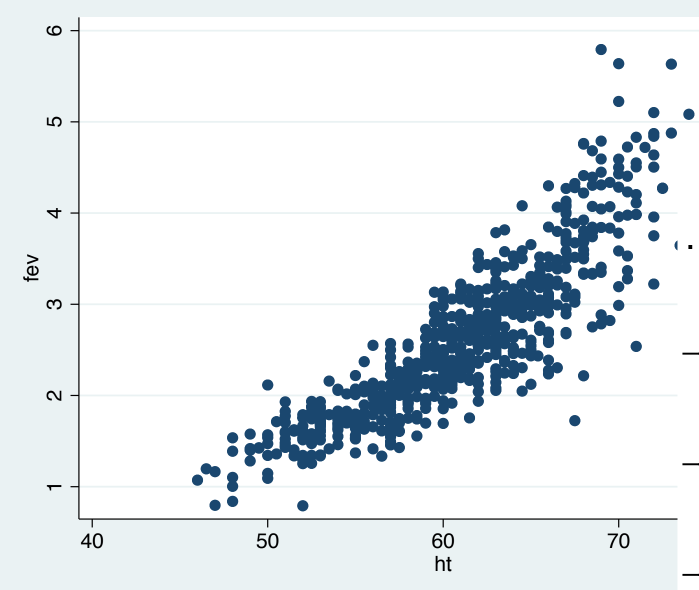
- $\text{Ln}(FEV) = 0.051 + 0.087 \text{ age}$

- $\text{Ln}(FEV_{age21}) = 0.051 + 0.087*21$
- $\text{Ln}(FEV_{age20}) = 0.051 + 0.087*20$
- $\text{Ln}(FEV_{age21}) - \ln(FEV_{age20}) = 0.087$

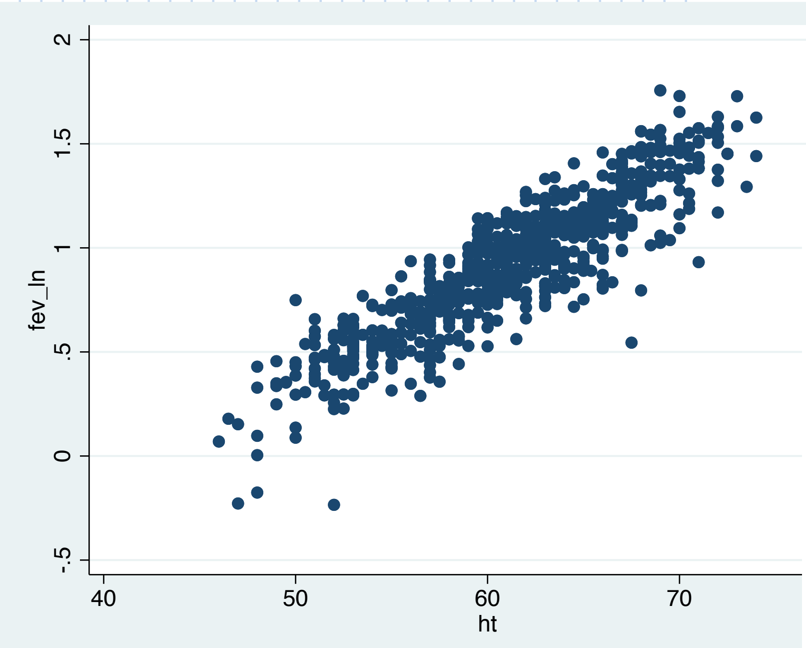
  Remember $\ln(a)-\ln(b) = \ln(a/b)$
- $\text{Ln}(FEV_{age21} / FEV_{age20}) = 0.087$
- $FEV_{age21}/FEV_{age20} = e^{0.087} = 1.09$

Now using height as the independent variable

1. Make a scatter plot of FEV by height and FEV_ln by height
2. Run a regression of FEV on height and FEV_ln on height
3. What does this tell you?




```
. regress fev ht

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(1, 652)       =   1994.73
       Model |  369.985854         1  369.985854   Prob > F        =    0.0000
    Residual |  120.933979       652  .185481563   R-squared       =    0.7537
-------------+----------------------------------   Adj R-squared   =    0.7533
       Total |  490.919833       653  .751791475   Root MSE        =    .43068

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ht |   .1319756    .002955    44.66   0.000     .1261732     .137778
       _cons |  -5.432679   .1814599   -29.94   0.000    -5.788995   -5.076363
------------------------------------------------------------------------------
```




```
. regress fev_ln ht

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(1, 652)       =   2537.94
       Model |  57.7021467         1  57.7021467   Prob > F        =    0.0000
    Residual |  14.8237679       652  .02273584    R-squared       =    0.7956
-------------+----------------------------------   Adj R-squared   =    0.7953
       Total |  72.5259145       653  .111065719   Root MSE        =    .15078

------------------------------------------------------------------------------
      fev_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ht |   .0521191   .0010346    50.38   0.000     .0500876    .0541506
       _cons |  -2.271312    .063531   -35.75   0.000    -2.396062   -2.146562
------------------------------------------------------------------------------
```

Categorical independent variables

- We previously noted that the independent variable (the X variable) does not need to be normally distributed
- In fact, this variable can be categorical

Categorical independent variables

- Coding: Dichotomous variables in regression models

  1 to represent the level of interest

  0 to represent the comparison or reference group.

  These 0-1 variables are called indicator or dummy variables.

- The regression model is the same
- The interpretation is slightly different
  - β is the change in y that corresponds to being in the group of interest vs. the reference category

Categorical independent variables

- Example sex: female x_{sex}=0, for male x_{sex} =1
- Regression of FEV and sex
- $$\hat{fev} = \hat{\alpha} + \hat{\beta}\, x_{sex}$$
- For male: $\hat{fev}_{male} = \hat{\alpha} + \hat{\beta}*1 = \hat{\alpha} + \hat{\beta}$
- For female: $\hat{fev}_{female} = \hat{\alpha} + \hat{\beta}*0 = \hat{\alpha}$

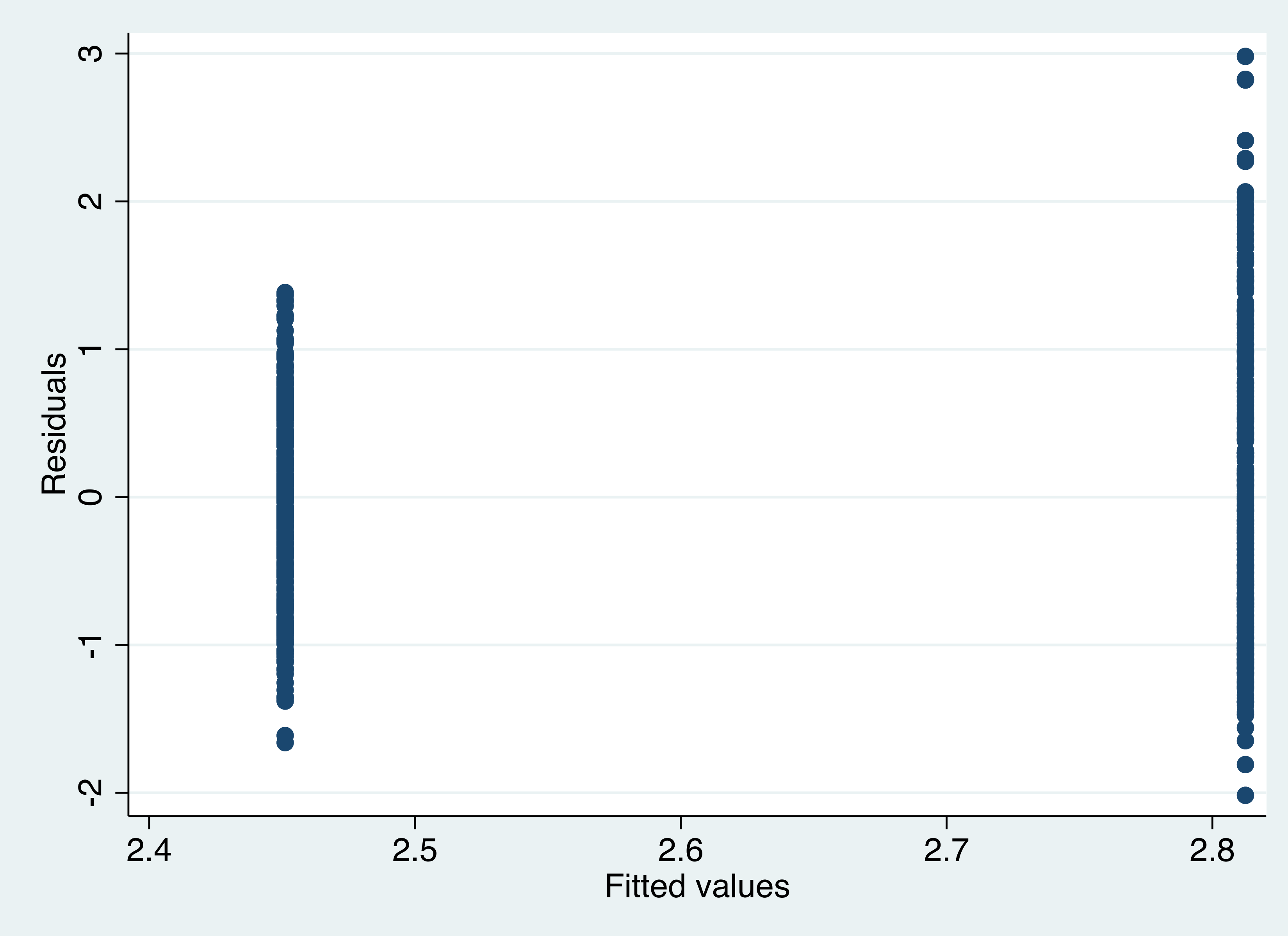
  So $\hat{fev}_{male} - \hat{fev}_{female} = \hat{\alpha} + \hat{\beta} - \hat{\alpha} = \hat{\beta}$

- Remember, $\hat{\alpha}$ is the mean value of y when x=0

  So here it is the mean FEV for sex=female

```
. regress fev sex

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(1, 652)     =   29.61
       Model |  21.3239848     1  21.3239848           Prob > F      =  0.0000
    Residual |  469.595849   652  .720239032           R-squared     =  0.0434
-------------+------------------------------           Adj R-squared =  0.0420
       Total |  490.919833   653  .751791475           Root MSE      =  .84867

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .3612766   .0663963     5.44   0.000     .2309002     .491653
       _cons |    2.45117    .047591    51.50   0.000      2.35772     2.54462
------------------------------------------------------------------------------
```

1. What is the estimate for beta? How is it interpreted?
2. What is the estimate for alpha? How is it interpreted?
3. What hypothesis is tested where it says P>|t|?
4. How much of the variance in FEV is explained by sex?
5. What does the rvfplot look like?



Residuals
3
2
1
0
-1
-2
2.4
2.5
2.6
2.7
2.8
Fitted values

Categorical independent variable

- Remember that the regression equation is

  $\mu_{y|x} = \alpha + \beta x$

- The only variables x can take are 0 and $\mu_{y|0} = \alpha$
  $\mu_{y|1} = \alpha + \beta$

- The estimated mean FEV for females is $\hat{\alpha}$ and the estimated mean FEV for males is $\hat{\alpha} + \widehat{\beta}$

- What other test have we learned that tests the same thing?

```
. ttest fev, by(sex)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     318     2.45117    .0362111     .645736    2.379925    2.522414
       1 |     336    2.812446    .0547507    1.003598    2.704748    2.920145
---------+--------------------------------------------------------------------
combined |     654     2.63678    .0339047    .8670591    2.570204    2.703355
---------+--------------------------------------------------------------------
    diff |           -.3612766    .0663963                -.491653   -.2309002
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -5.4412
Ho: diff = 0                                     degrees of freedom =      652

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

```
. regress fev sex

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(1, 652)       =     29.61
       Model |  21.3239848         1  21.3239848   Prob > F        =    0.0000
    Residual |  469.595849       652  .720239032   R-squared       =    0.0434
-------------+----------------------------------   Adj R-squared   =    0.0420
       Total |  490.919833       653  .751791475   Root MSE        =    .84867

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .3612766   .0663963     5.44   0.000     .2309002     .491653
       _cons |    2.45117    .047591    51.50   0.000      2.35772     2.54462
------------------------------------------------------------------------------
```

Categorical independent variables

- e.g. Race group = White, Asian/PI, Other
- If Race=White is set as reference category, dummy variables look like:

	$x_{Asian/PI}$	x_{Other}
White	0	0
Asian/PI	1	0
Other	0	1

Categorical independent variables

- Then the regression equation is:

  $y = \alpha + \beta_1 x_{Asian/PI} + \beta_2 x_{Other} + \varepsilon$

- For race group=White

  $\hat{y} = \hat{\alpha} + \widehat{\beta_1}0 + \widehat{\beta_2}0 = \hat{\alpha}$

- For race group=Asian/PI

  $\hat{y} = \hat{\alpha} + \widehat{\beta_1}1 + \widehat{\beta_2}0 = \hat{\alpha} + \widehat{\beta_1}$

- For race group=other

  $\hat{y} = \hat{\alpha} + \widehat{\beta_1}0 + \widehat{\beta_2}1 = \hat{\alpha} + \widehat{\beta_2}$

Stata indicator variables

- You actually don't have to make the dummy variables yourself
- All you have to do is tell Stata that a variable is categorical using i. before a variable name
- The regression of FEV regressed on age group

```
regress fev i.agecat
```

```
. regress fev i.agecat

      Source |       SS           df       MS      Number of obs   =       654
-------------+----------------------------------   F(4, 649)       =    191.07
       Model |  265.481813         4  66.3704533   Prob > F        =    0.0000
    Residual |   225.43802       649  .347362127   R-squared       =    0.5408
-------------+----------------------------------   Adj R-squared   =    0.5380
       Total |  490.919833       653  .751791475   Root MSE        =    .58937

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
         6-  |   .4713427    .104309     4.52   0.000     .2665188    .6761665
         9-  |   1.244846   .1010818    12.32   0.000     1.046359    1.443332
        12-  |   1.912191      .1081    17.69   0.000     1.699923    2.124459
        15-  |   2.237758   .1264743    17.69   0.000      1.98941    2.486106
             |
       _cons |   1.472385   .0943754    15.60   0.000     1.287067    1.657703
------------------------------------------------------------------------------
```

Multiple regression

- Additional explanatory variables might increase our understanding of a dependent variable
- We can posit the population equation

$$\mu_{y|x1,x2,\ldots,xq} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q$$

- α is the mean of y when all the explanatory variables are 0
- β_i is the change in the mean value of y the corresponds to a 1 unit change in x_i when all the other explanatory variables are held constant

Multiple regression

- $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \varepsilon$
- Assumptions
  - $x_1, x_2, \ldots, x_q$ are measured without error
  - The distribution of y is normal with mean $\mu_{y|x1,x2,\ldots,xq}$ and standard deviation $\sigma_{y|x1,x2,\ldots,xq}$
  - The population regression model holds
  - For any set of values of the explanatory variables, $x_1, x_2, \ldots, x_q$, $\sigma_{y|x1,x2,\ldots,xq}$ is constant – homoscedasticity
  - The y outcomes are independent

Multiple regression – Least Squares

- We estimate the regression line

$$\hat{y} = \hat{\alpha} + \widehat{\beta_1}\, x_1 + \widehat{\beta_2}\, x_2 + \ldots \widehat{\beta_q}\, x_q$$

using the method of least squares to minimize

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

$$= \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_q x_{qi})]^2$$

Multiple regression

- For one explanatory variable – the regression model represents a straight line through a cloud of points -- in 2 dimensions
- With 2 explanatory variables, the model is a plane in 3 dimensional space (one for each variable)
- etc.
- In Stata we just add explanatory variables to the regress statement
- `regress fev age ht`

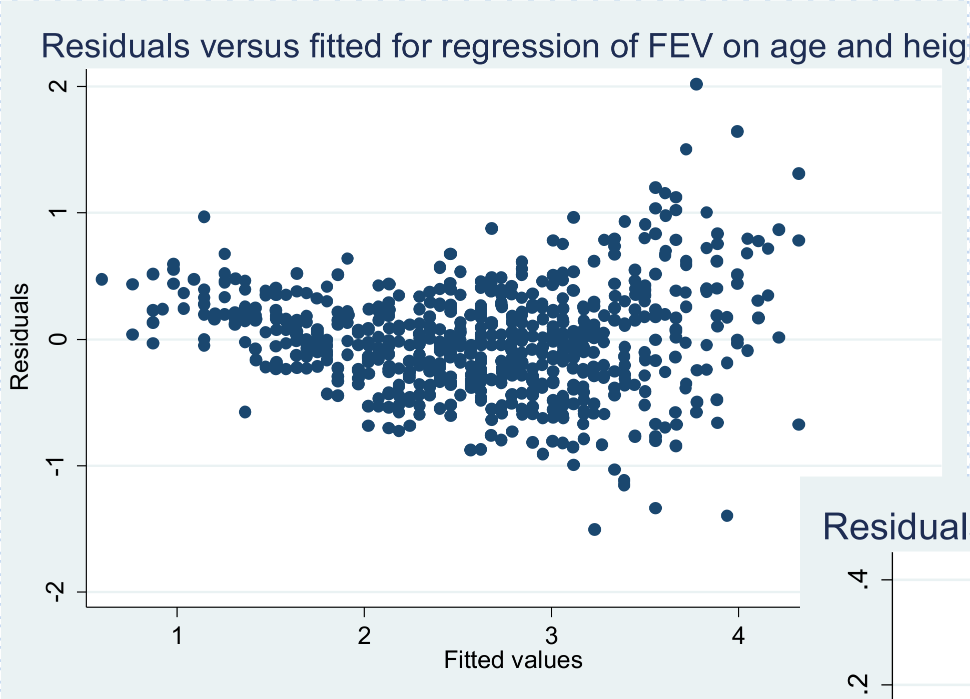
```
. regress fev age ht

      Source |       SS       df       MS              Number of obs =     654
-------------+------------------------------           F(  2,   651) = 1067.96
       Model |  376.244941     2  188.122471           Prob > F      =  0.0000
    Residual |  114.674892   651  .176151908           R-squared     =  0.7664
-------------+------------------------------           Adj R-squared =  0.7657
       Total |  490.919833   653  .751791475           Root MSE      =   .4197

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0542807   .0091061     5.96   0.000     .0363998    .0721616
          ht |   .1097118   .0047162    23.26   0.000      .100451    .1189726
       _cons |  -4.610466   .2242706   -20.56   0.000    -5.050847   -4.170085
------------------------------------------------------------------------------
```

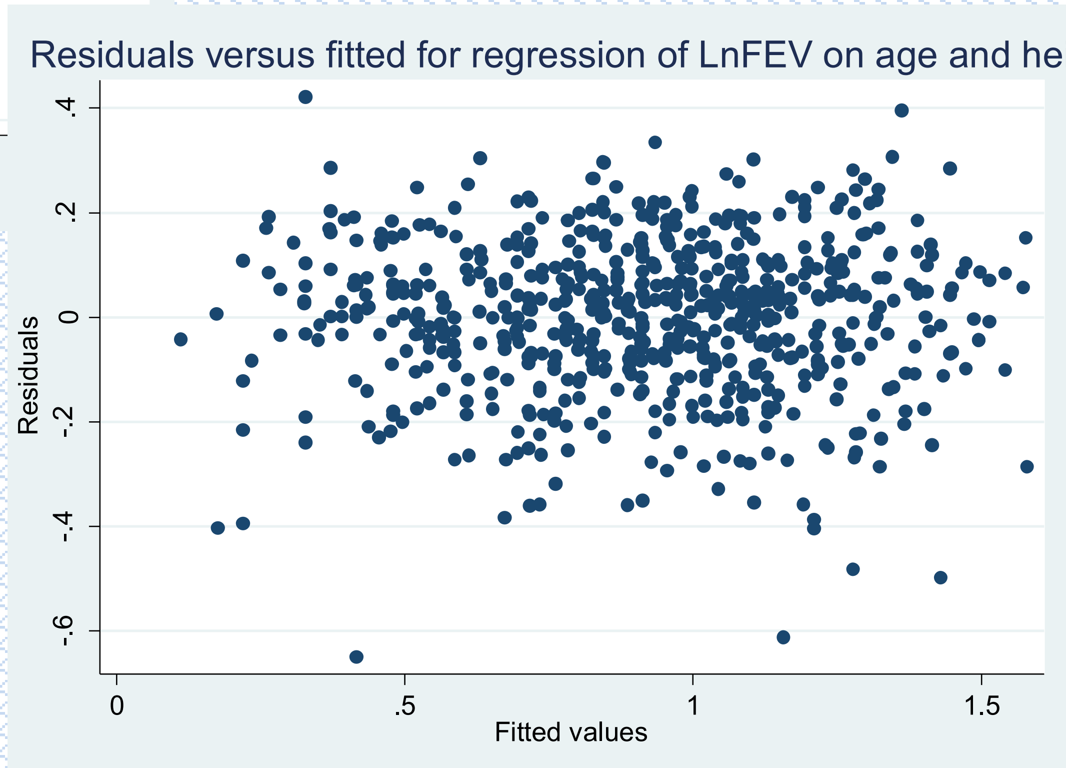
•Now the F-test has 2 degrees of freedom in the numerator because there are 2 explanatory variables

•R^2 will always increase as you add more variables into the model

•The Adj R-squared accounts for the addition of variables and is comparable across models with different numbers of parameters

Examine the residuals...





Logistic regression

- For linear regression
  - The predicted values can be from $-\infty$ to $+\infty$
  - Y values are assumed to follow a normal distribution

- For many situations we want to model a dichotomous outcome, such as disease/no disease and cannot use linear regression

- We want to model p, the probability of disease

Logistic regression

- p=P(Y=1)   i.e. P(outcome=disease)

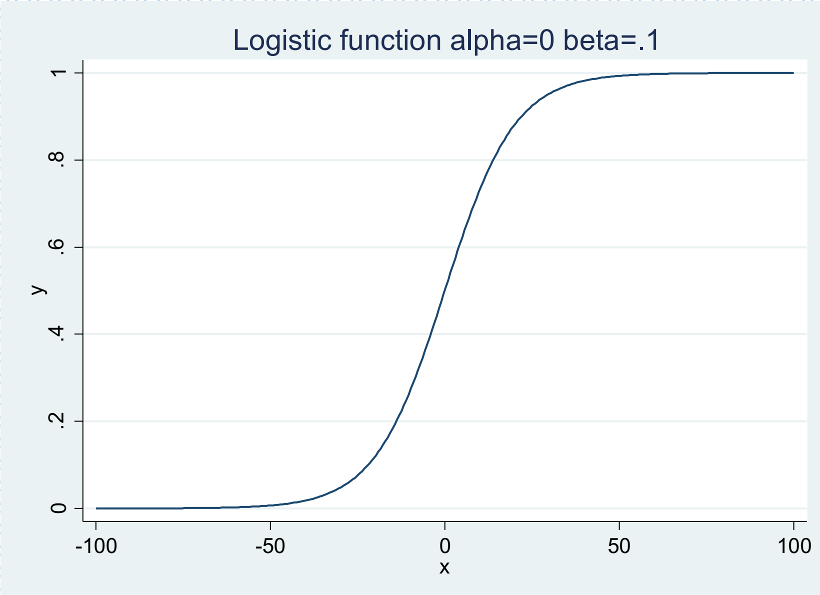
- A model of the form p= $\alpha + \beta x$ would be able to take on negative values or values more than 1

- $p=e^{\alpha + \beta x}$ is an improvement because it cannot be negative, but it still could be greater than 1

Logistic regression

- How about the function?

$$p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$



- When $\alpha + \beta x = 0$, p=1/(1+1) = 0.5
- When x=big negative number, p nears 0/(1+0) = 0
- When x=big positive number, p nears (big number)/(1+big number) = 1
- The function models the probability slowly increasing over the values of X, until there is a steep rise, and another leveling off

Logistic regression

- The logistic function $$p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$
- Now check this out:

$$1-p = 1-\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1+e^{\alpha+\beta x}-e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{\alpha+\beta x}}$$

- So odds of disease/success are

$$\frac{p}{1-p} = \frac{\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}{\frac{1}{1+e^{\alpha+\beta x}}} = e^{\alpha+\beta x}$$

- And therefore $\ln(p/(1-p)) = \ln(e^{\alpha+\beta x}) = \alpha + bx$

Logistic regression

- So instead of assuming that the relationship between p and X is linear (as in linear regression), we are assuming that the relationship between ln(p/(1-p)) and X is linear.
- ln(p/(1-p)) is called the <u>logit</u> function
- Logit(p) = ln(p/(1-p))
- In generalized linear models, some function of the outcome variable is assumed to be linear

  e.g. $f(y) = \alpha + \beta x$ or $f(p) = \alpha + \beta x$

  e.g. $\text{logit}(p) = \alpha + \beta x$

Logistic regression

- Assumptions
  - Y_i follows a bernoulli distribution (0,1)
  - The mean of Y at every X, $P(Y=1|x)$ is given by the logistic function
  - The values of the outcomes are independent
- Estimation
  - The parameters are estimated via a procedure called *maximum likelihood*
  - There is not a closed form solution of the maximization (i.e. we cannot write down a solution) – it is an iterative procedure

Interpretation of coefficients

- One dichotomous explanatory variable

- For x=0 logit(p_0) = ln(p_0/(1-p_0))
  = ln(P(Y=1|X=0)/(1- P(Y=1|X=0)))
  = ln(odds of the outcome when x=0)
  = $\alpha + \beta * x = \alpha + \beta * 0 = \boldsymbol{\alpha}$

- For x=1 logit(p_1) = ln(p_1/(1-p_1))
  = ln(P(Y=1|X=1)/(1- P(Y=1|X=1)))
  = odds of the outcome when x=1
  = $\alpha + \beta * 1 = \boldsymbol{\alpha + \beta}$

Difference between treatment arms in PEth outcome (>50 ng/ml)*

```
logistic peth_audit_50 i.studyarm_n

Logistic regression                               Number of obs   =        324
                                                  LR chi2(1)      =       1.14
                                                  Prob > chi2     =     0.2858
Log likelihood = -223.91123                       Pseudo R2       =     0.0025

------------------------------------------------------------------------------
peth_audit_50 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
 1.studyarm_n |   .7869577   .1768227    -1.07   0.286     .5066332    1.222388
        _cons |    1.05618   .1562084     0.37   0.712     .7903976    1.411335
------------------------------------------------------------------------------
```

*BREATH cohort and comparisons_v3.dta

Difference between treatment arms in PEth outcome (>50 ng/ml)

logistic depvar indepvar

```
logistic peth_audit_50 i.studyarm_n

Logistic regression                             Number of obs   =        324
                                                LR chi2(1)      =       1.14
                                                Prob > chi2     =     0.2858
Log likelihood = -223.91123                     Pseudo R2       =     0.0025

------------------------------------------------------------------------------
peth_audit_50 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
 1.studyarm_n |   .7869577   .1768227    -1.07   0.286     .5066332    1.222388
        _cons |    1.05618   .1562084     0.37   0.712     .7903976    1.411335
------------------------------------------------------------------------------
```

$=e^{\beta}$

Test of whether the model fit differs from the model with only α and no covariates

This is used in comparing nested models of the same data set

The Pseudo R^2 is an attempt to be analogous with linear regression but it is not

Stata logistic notes

- logistic depvar indepvars
  - gives you odds ratios
- logistic depvar indepvars, coef
  - gives you coefficients (alphas and betas)
- logit depvar indepvars
  - gives you coefficients
- logit depvar indepvars, or
  - gives you odds ratios

```
logit peth_50  i.studyarm_n

Iteration 0:   log likelihood = -198.01065
Iteration 1:   log likelihood = -197.99359
Iteration 2:   log likelihood = -197.99359

Logistic regression                               Number of obs     =        287
                                                  LR chi2(1)        =       0.03
                                                  Prob > chi2       =     0.8534
Log likelihood = -197.99359                       Pseudo R2         =     0.0001

------------------------------------------------------------------------------
     peth_50 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.studyarm_n |  -.0453478   .2455377    -0.18   0.853    -.5265928    .4358972
       _cons |  -.1438942   .1490438    -0.97   0.334    -.4360146    .1482262
------------------------------------------------------------------------------
```

Odds ratio of PEth>=50 for comparison vs. standard group

```
tabodds peth_audit_50 studyarm_n, or

---------------------------------------------------------------------------
  studyarm_n |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
           0 |    1.000000          .            .             .          .
           1 |    0.786958       1.13       0.2867      0.505895   1.224172
---------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =     1.13
                                  Pr>chi2  =   0.2867

Score test for trend of odds:     chi2(1)  =     1.13
                                  Pr>chi2  =   0.2867
```

For <u>one independent variable that is categorical</u>, the odds ratio estimate is the same for logistic regression as for tabular methods.

What is the ***<u>odds ratio</u>***?

- The odds ratio represents the change in the odds of membership in the “target” category (e.g., the category coded as 1.0), given a 1-unit increase in the predictor variable.
- Odds ratio is the increase (or decrease if negative) in odds of being in one outcome category when the value of the predictor variable increases by one unit.
- The odds ratio is calculated by the formula: e^g (where e = 2.718, and g = the b coefficient from the logistic regression output).

What is the ***odds ratio***?

- For example:
  - An odds ratio of 1.5 indicates that that outcome of "1" is 1.5 times as likely (or 50% more likely) with a 1-unit increase in the predictor. The odds increase by 50%.
  - An odds ratio of 0.80 indicates that an outcome of "1" is 0.8 times as likely (or 20% less likely [1 – 0.8 = 0.2] with a 1-unit increase in the predictor; the odds are decreased by 20%.

How do you compute the probability of being in the "target" group, for a specific individual?

- Multiply each predictor value by its "b" coefficient, and add them to the constant. That sum is called "g."
- Use "g" in the following formula:
- **$e^g / 1 + e^g$**

Continuous explanatory variable

```
.
. . logistic peth_audit_50 age_b

Logistic regression                               Number of obs   =        324
                                                  LR chi2(1)      =       5.32
                                                  Prob > chi2     =     0.0210
Log likelihood = -221.81862                       Pseudo R2       =     0.0119

------------------------------------------------------------------------------
peth_audit_50 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
        age_b |   1.029989   .0133876     2.27   0.023     1.004081    1.056565
        _cons |   .3782632   .1590054    -2.31   0.021     .1659534    .8621885
------------------------------------------------------------------------------
```

- Interpretation of the coefficients: The odds ratio is for a one unit change in the predictor
- For this example 1.03 is the odds ratio for a one-year difference in age

Continuous explanatory variable

- When the explanatory variable is a continuous variable, there is no equivalent tabular method without dividing the variable into groups (e.g. age groups)

- You need to use logistic regression when you want to model the relationship between a dichotomous outcome and a continuous explanatory variable

Continuous explanatory variable

- To find the OR for a 10-year change in age

```
.
. logistic peth_audit_50 age_b, coef

Logistic regression                               Number of obs   =        324
                                                  LR chi2(1)      =       5.32
                                                  Prob > chi2     =     0.0210
Log likelihood = -221.81862                       Pseudo R2       =     0.0119

------------------------------------------------------------------------------
peth_audit_50 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
        age_b |   .0295481   .0129978     2.27   0.023     .0040729    .0550234
        _cons |   -.972165   .4203565    -2.31   0.021    -1.796049   -.1482814
------------------------------------------------------------------------------
```

OR for a 10-year change in age = exp(10*0.029) = 1.34

.

Continuous explanatory variable

- To find the OR for a 10-year change in age

```
.
. lincom age_b*10

 ( 1)  10*[peth_audit_50]age_b = 0

------------------------------------------------------------------------------
peth_audi~50 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.343773   .1746612     2.27   0.023      1.04157    1.733658
------------------------------------------------------------------------------

.
```

Continuous explanatory variable

- Another way to find the OR for a 10-year change in age

```
gen age_10 = age_b/10

. logistic peth_audit_50 age_10

Logistic regression                               Number of obs   =        324
                                                  LR chi2(1)      =       5.32
                                                  Prob > chi2     =     0.0210
Log likelihood = -221.81862                       Pseudo R2       =     0.0119

------------------------------------------------------------------------------
peth_audit_50 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_10 |   1.343773   .1746612     2.27   0.023      1.04157    1.733658
        _cons |   .3782632   .1590054    -2.31   0.021     .1659534    .8621885
------------------------------------------------------------------------------
```

Checking logit-linearity

```
. tab agecat

     agecat |      Freq.     Percent        Cum.
------------+-----------------------------------
         15 |         80       24.69       24.69
         25 |        136       41.98       66.67
         35 |         80       24.69       91.36
         45 |         28        8.64      100.00
------------+-----------------------------------
      Total |        324      100.00
```

```
 logistic peth_audit_50 i.agecat

Logistic regression                             Number of obs   =        324
                                                LR chi2(3)      =       8.22
                                                Prob > chi2     =     0.0416
Log likelihood =  -220.3693                     Pseudo R2       =     0.0183

-------------------------------------------------------------------------------
peth_audit_50 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
       agecat |
          25  |   1.525822   .4390615     1.47   0.142     .8680944    2.681887
          35  |   2.254902    .728875     2.52   0.012     1.196699    4.248838
          45  |   2.575758    1.16071     2.10   0.036      1.06495    6.229893
              |
        _cons |         .6   .1385641    -2.21   0.027     .3815704    .9434694
-------------------------------------------------------------------------------
```

Logistic regression with >1 explanatory variable

- In multiple logistic regression, the probability of success depends on more than one explanatory variable

$$p = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}$$

$$\text{and } \frac{p}{1-p} = e^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}$$

- Therefore

$ln(p/(1-p)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$

Interpretation of the coefficients

- Example: 2 independent variables, X_1 a numerical variable, X_2 a dichotomous exposure variable
- $\text{logit}(p) = \log(p/(1-p)) = \alpha + \beta_1 x_1 + \beta_2 x_2$

- When $x_2=1$ (exposed), holding x_1 constant,

$$\text{logit}(p_1) = \alpha + \beta_1 x_1 + \beta_2$$

- When $x_2=0$ (not exposed), holding x_1 constant,

$$\text{logit}(p_0) = \alpha + \beta_1 x_1$$

```
. logistic peth_audit_50 age_10 i.studyarm_n

Logistic regression                             Number of obs   =        324
                                                LR chi2(2)      =       6.00
                                                Prob > chi2     =     0.0498
Log likelihood = -221.48068                     Pseudo R2       =     0.0134

------------------------------------------------------------------------------
peth_audit_50 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |   1.328877   .1737573     2.17   0.030     1.028457    1.717053
 1.studyarm_n |   .8295213   .1886524    -0.82   0.411     .5311835     1.29542
       _cons |   .4248297    .188248    -1.93   0.053     .1782524    1.012498
------------------------------------------------------------------------------
```

0.83 is the odds ratio for PEth>=50 for being in the minimally assessed vs. standard study arm when you hold age constant

To sum up…

<u>OLS</u>

- Response variable = Y
- Link function between the response variable and what is being modeled =1
- Eq. Y=a+bX
- Distr. = Normal
- Var. = Constant

<u>Logistic</u>

- Response variable = Y
- Link function between the response variable and what is being modeled= logit
- $\mathrm{Ln}(\pi/1-\pi)=a+bX$
- Distr. = Binomial
- Var = Not Constant

To sum up…

<u>OLS</u>

- $Y=a+bX$
- Interpretation of b
- Omnibus test =F
- Testing of b = t
- Goodness of fit=R^2

<u>Logistic</u>

- $Ln(\pi/1-\pi)=a+bX$
- Interpretation of b
- Omnibus test =-2LL
- Testing of b = Wald
- Goodness of fit= -2LL, pseudo-R^2

To sum up… assumptions

<u>OLS</u>

- Linearity
- Additivity
- Normal distribution
- Errors terms are independent

<u>Logistic</u>

- Non-Linearity … however assumes a linear relationship between the logit of the dependent and the independent variables.
- Additivity
- Non-normal distribution
- Errors terms are independent

To sum up… assumptions

OLS

- OLS relies on the central limit theorem which states that 'as we increase the sample size of the statistics that make up the sampling distribution,.., the form of the distribution becomes increasingly symmetrical and bell-shaped

Logistic

- 'Maximum likelihood estimation (MLE) relies on large-sample asymptotic normality which means that reliability of estimates decline when there are few cases for each observed combination of X variables.'

- Next week – more multiple regression and multivariate logistic regression
- Review of methods this term
- Good website for Stata statistical methods:

    http://www.ats.ucla.edu/stat/stata/webbooks/