Biostat 200
Lecture 10

Announcements

- Lectures
  - No lecture 11/26
  - Last lecture 12/3 (logistic regression)
- Assignment 8 – due Tuesday 11/26
  - Covers lectures 9+10
- Final
  - Will be posted 12/5
  - Will be due 12/12
  - Take home format – no collaboration
  - Material through linear regression

Simple linear regression

- Population regression equation $\mu_{y|x} = \alpha + \beta x$
- α and β are constants and are called the coefficients of the equation
- α is the y-intercept and which is the mean value of Y when X=0, which is $\mu_{y|0}$
- The slope 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

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

Simple linear regression

- There is variability in populations, so the linear regression equation is $y = \alpha + \beta x + \varepsilon$

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

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

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

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

Simple linear regression

- Assumptions of linear regression
  - X’s are measured without error
  - For each value of x, the y’s are normally distributed with mean $\mu_{y|x}$ and standard deviation $\sigma_{y|x}$
  - $\mu_{y|x} = \alpha + \beta x$
  - Homoscedasticity
  - All the y_i ‘s are independent

Simple linear regression

- 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 (hence the name "least squares")
- We are minimizing the sum of the squares of the residuals

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

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

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)

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

```
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
------------------------------------------------------------------------------
```

$=.7565^2$

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

$$\text{total sum of squares} = TSS = MSS + RSS$$

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

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

```
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
------------------------------------------------------------------------------
```

$=.7565^2$

$s^2_{y|x}$

$s_{y|x}$

s^2_y

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

$$\text{total sum of squares} = TSS = MSS + RSS$$

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

Inference for regression coefficients

- We can use these to test the null hypothesis H_0: $\beta = \beta_0$ (=0)
- The test statistic for this is $t = \frac{\hat{\beta} - 0}{s\hat{e}(\hat{\beta})}$
- It follows the t distribution with n-2 degrees of freedom under the null hypothesis

- 95% confidence intervals for β

$$(\widehat{\beta} - t_{\text{n-2,0.25}}\,\text{se}(\widehat{\beta}), \widehat{\beta} - t_{\text{n-2,0.25}}\,\text{se}(\widehat{\beta}))$$

Inference for predicted values

- We might want to estimate the mean value of y at a particular value of x
- E.g. what is the mean FEV for children who are 10 years old?

$$\hat{y} = .432 + .222*x = .432 + .222*10 = 2.643 \text{ liters}$$

Inference for predicted values

- 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

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

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

- 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:

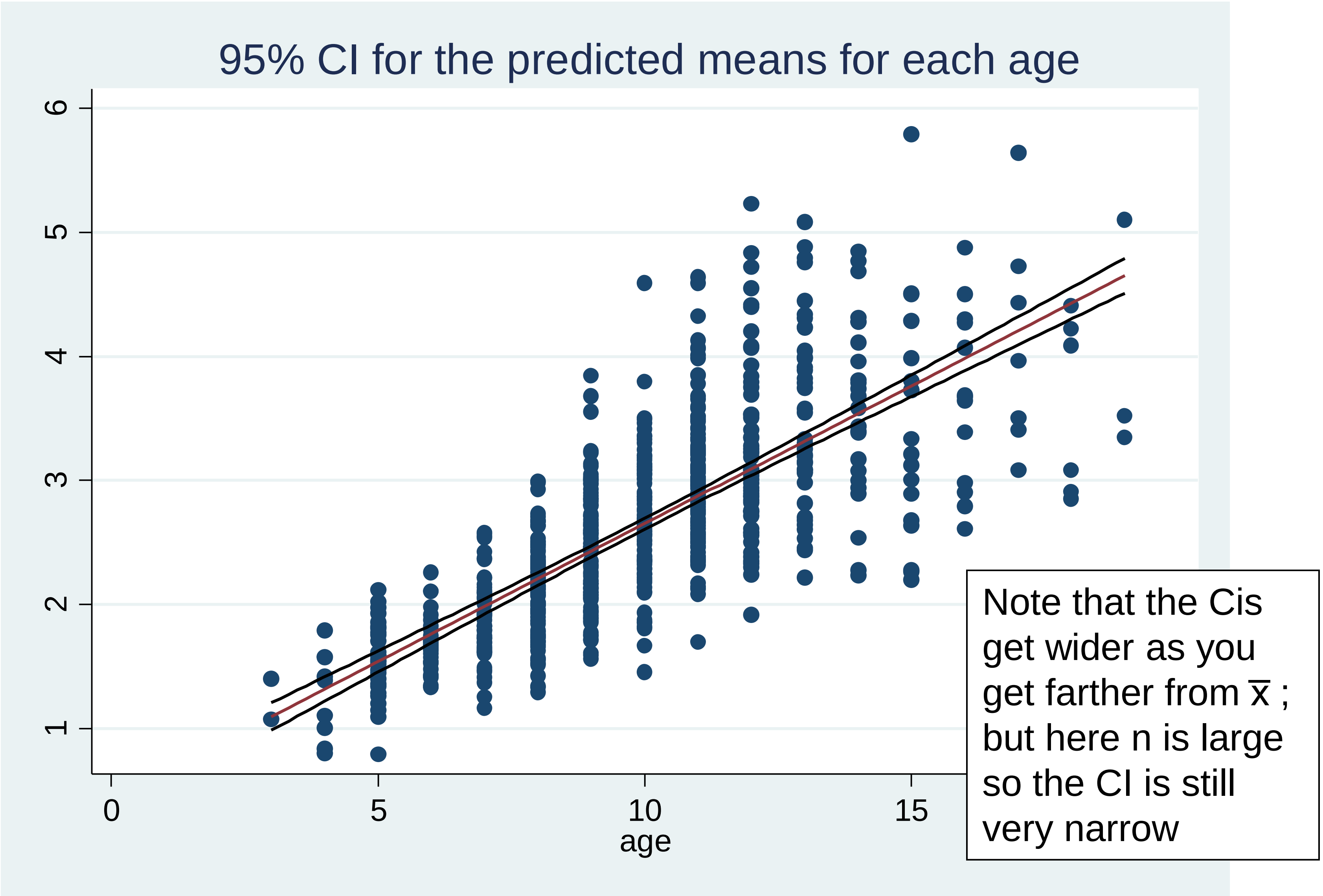
```
twoway (lfitci fev age)
```

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

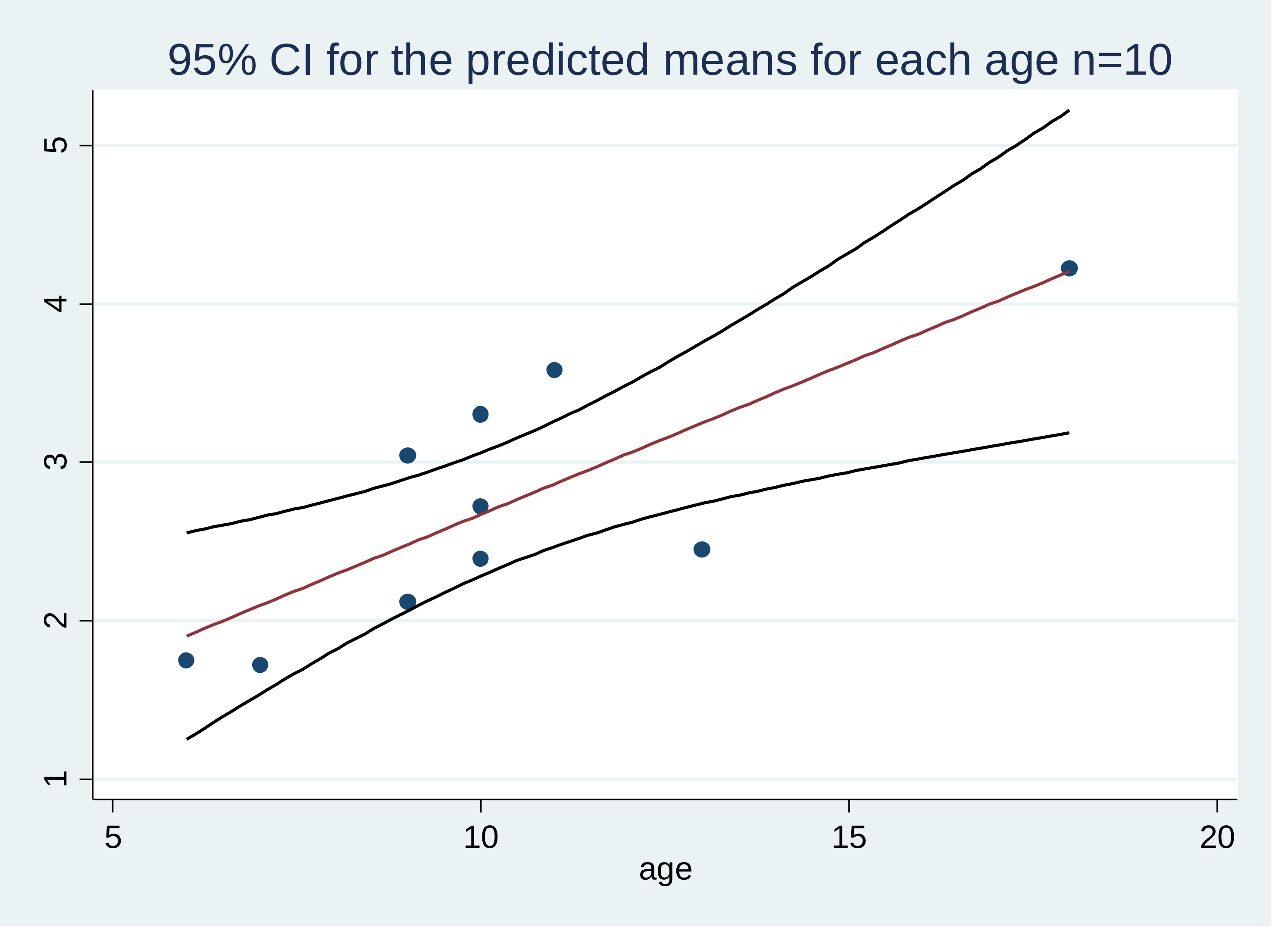
```
. list fev age fev_pred fev_predse

     +-----------------------------------+
     |   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 |
```




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



The 95% confidence intervals get much wider with a small sample size

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?
- Same thing – 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

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

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

- 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

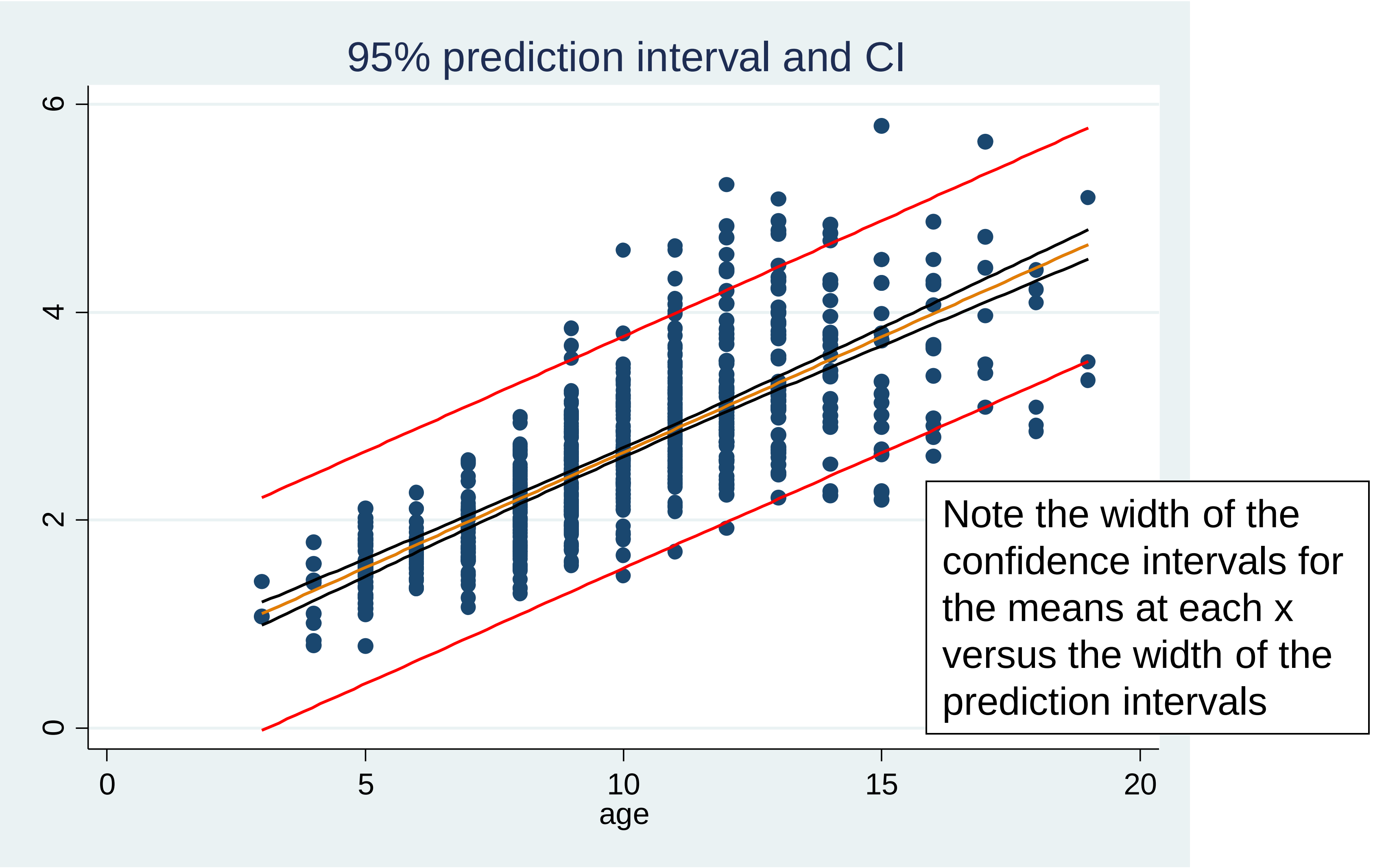
```
predict fev_predse_ind, stdf
```

f is for forecast

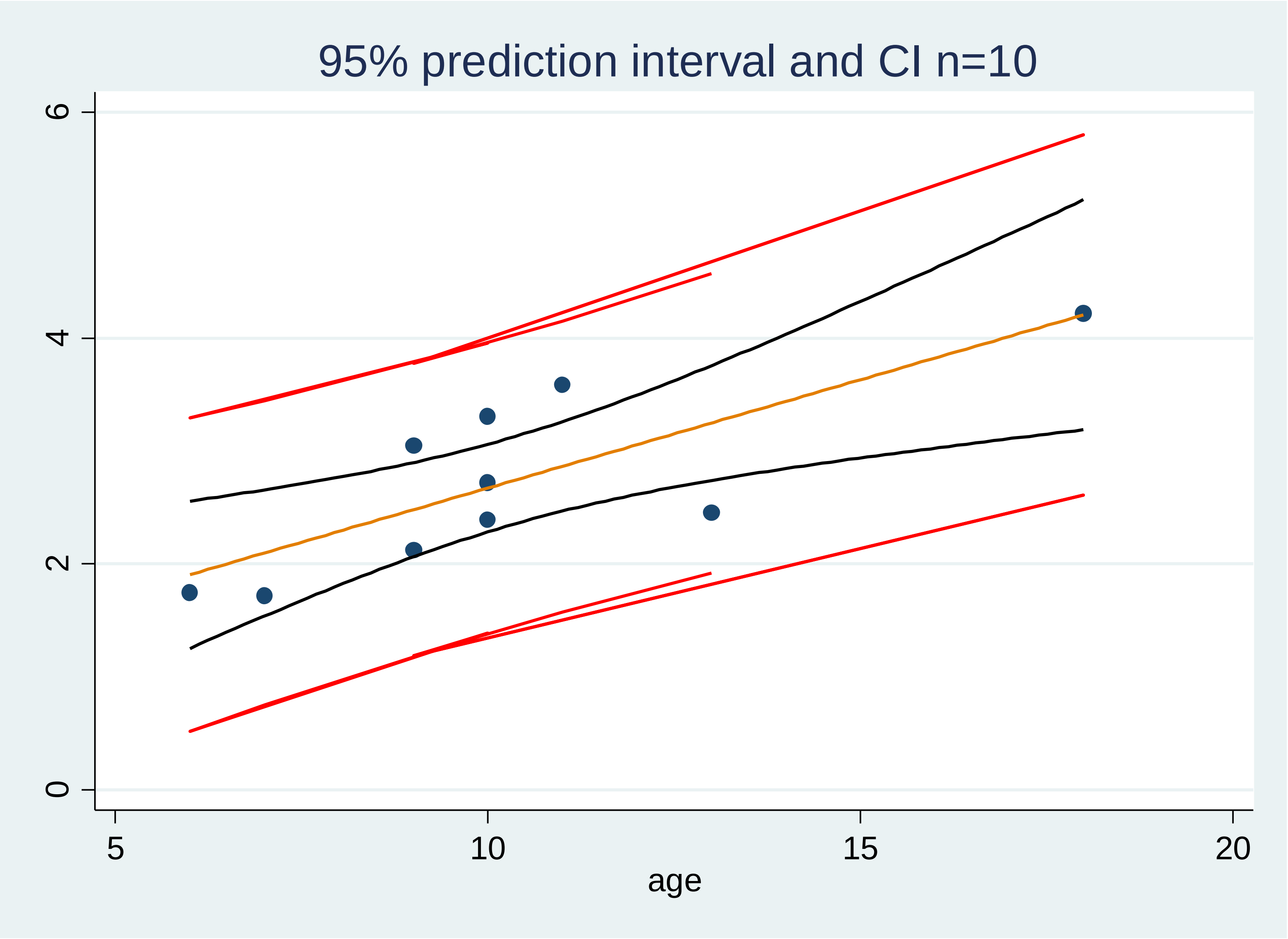
```
. list fev age fev_pred fev_predse fev_pred_ind

     +------------------------------------------------+
     |   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 |
```




```
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 )
```



The intervals are wider farther from $\overline{x}$, but that is only apparent for small n because most of the width is due to the added $s_{y|x}$

Model fit

- A summary of the model fit is the coefficient of determination, R^2

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

- R^2 represents the portion of the variability that is removed by performing the regression on X
- R^2 is calculated from the regression with 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
------------------------------------------------------------------------------
```

$=.7565^2$

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

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

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

Model fit

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

- The statistic is

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

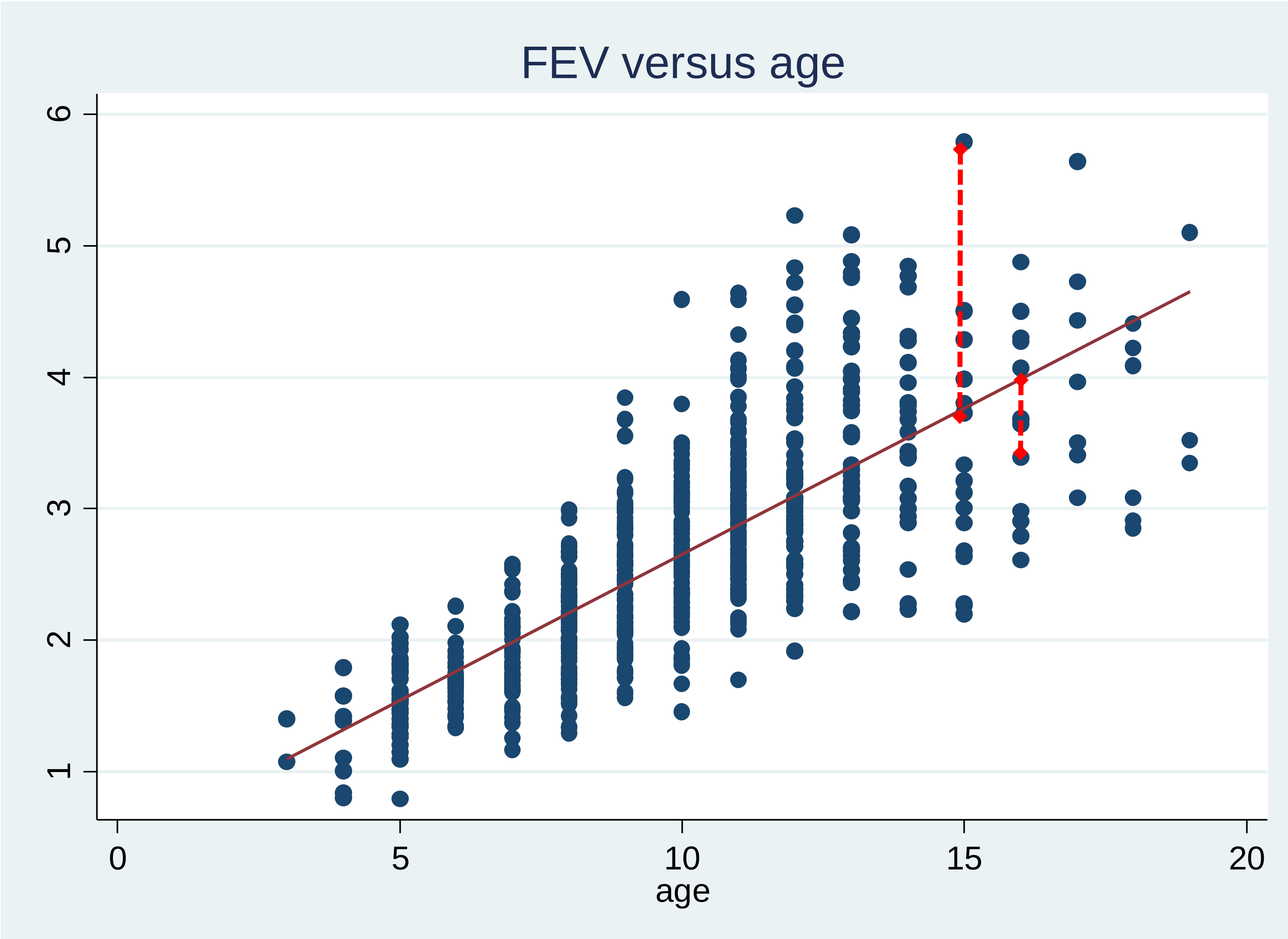
Model fit

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 test that β=0
- The test that r=0 (Pearson correlation)

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



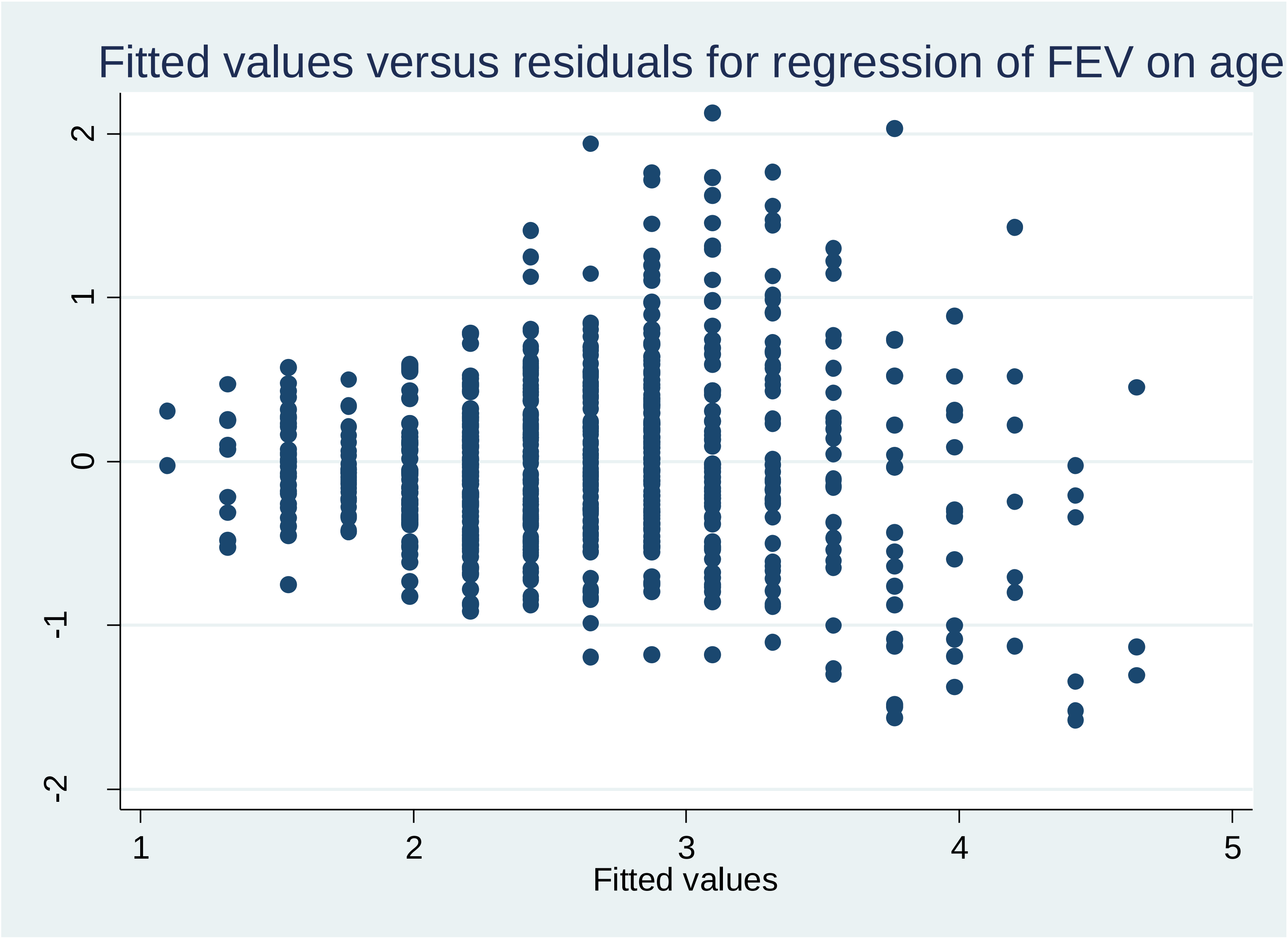
FEV versus age
6
5
4
3
2
1
0
5
10
15
20
age

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

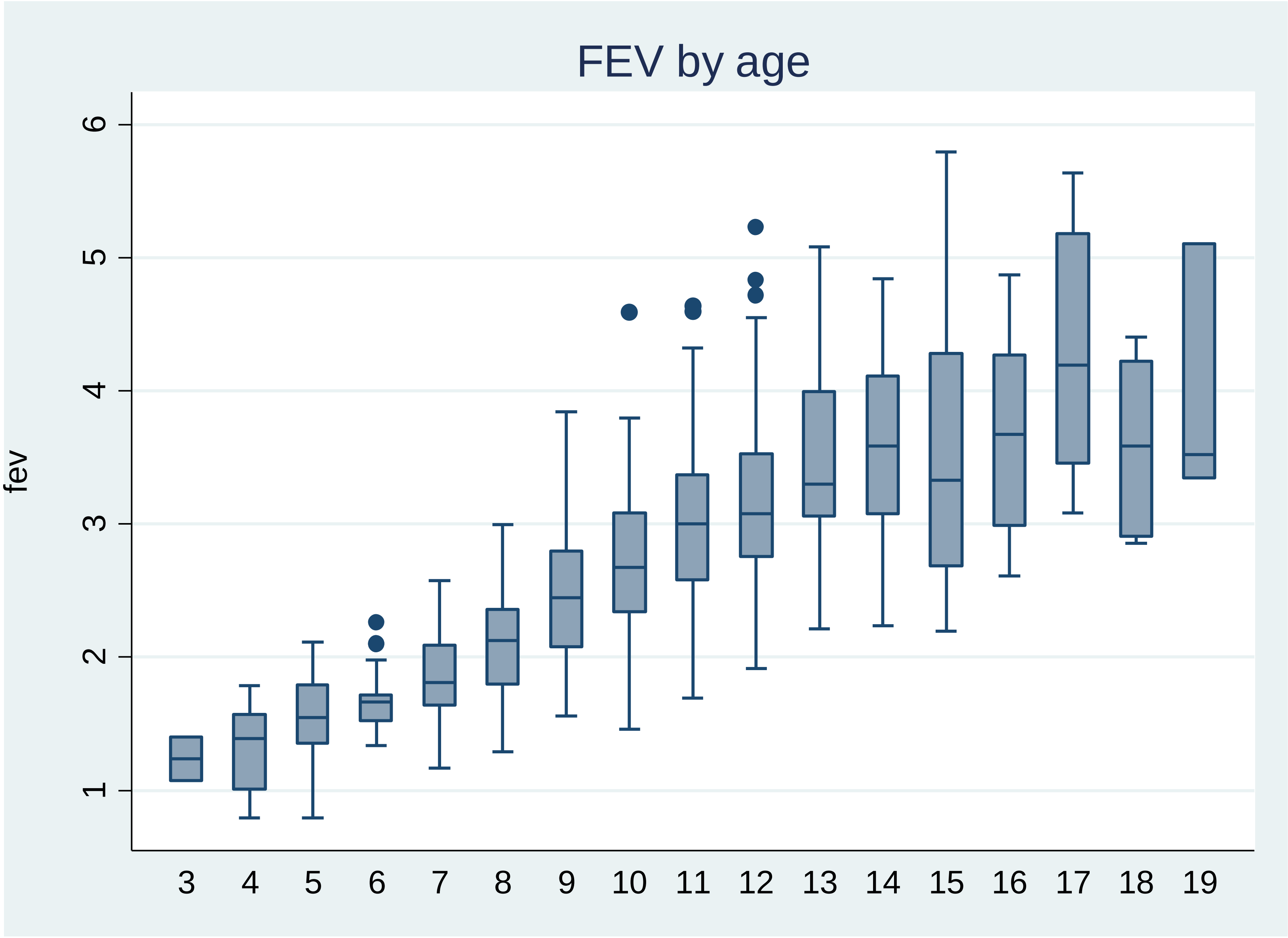



```
rvfplot, title(Fitted values versus residuals for
regression of FEV on age)
```

- 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




```
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 function refresher

- Log_{10}
  - $\text{Log}_{10}(x) = y$ means that $x=10^y$
  - So if $x=1000$ $\log_{10}(x) = 3$ because $1000=10^3$
  - $\text{Log}_{10}(103) = 2.01$ because $103=10^{2.01}$
  - $\text{Log}_{10}(1)=0$ because $10^0 =1$
  - $\text{Log}_{10}(0)=-\infty$ because $10^{-\infty} =0$

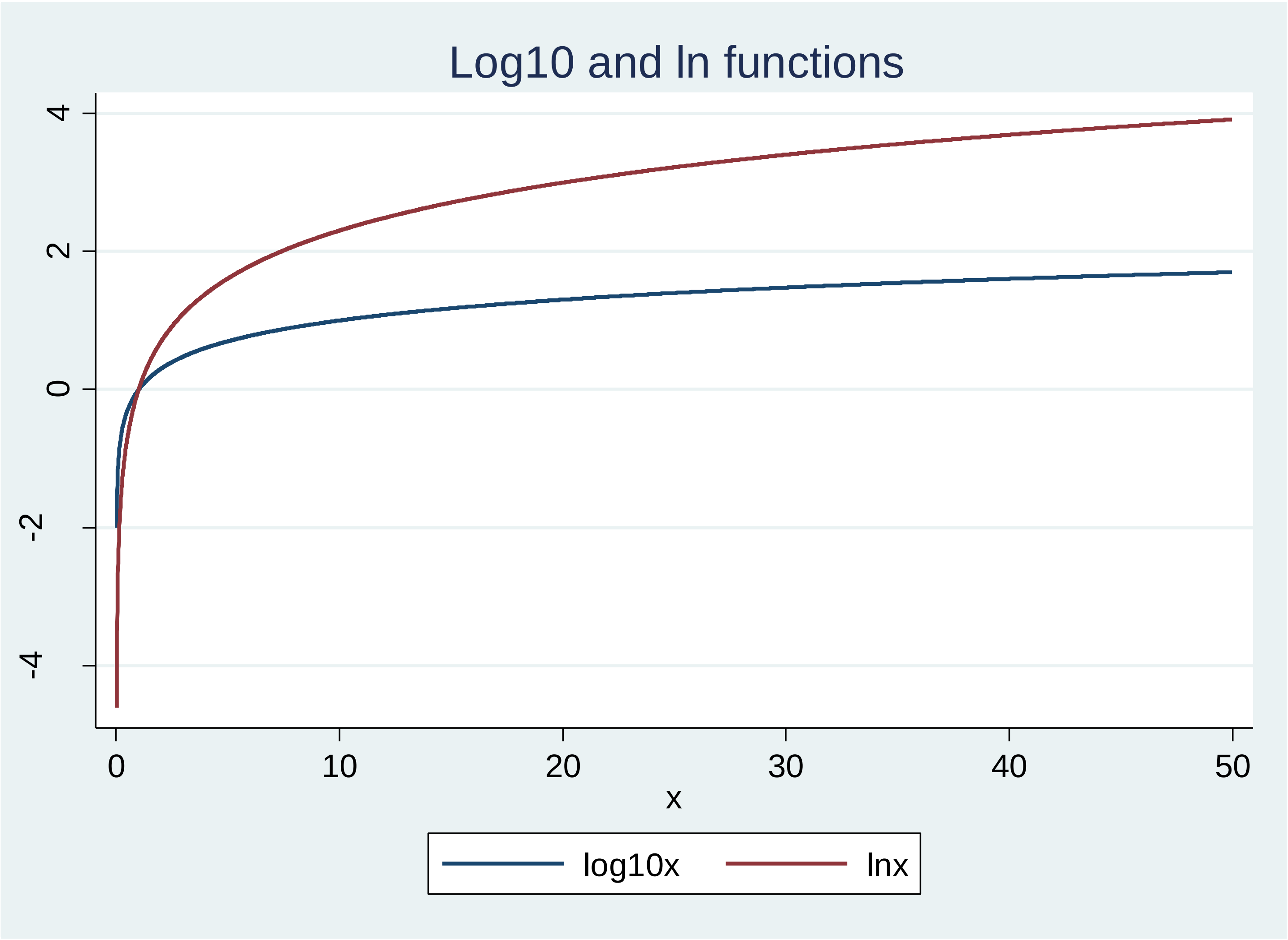
Log function refresher

- Log_e or ln
  - e is a constant approximately equal to 2.718281828
  - $\ln(1) = 0$ because $e^0 = 1$
  - $\ln(e) = 1$ because $e^1 = e$
  - $\ln(103) = 4.63$ because $103 = e^{4.63}$
  - $\text{Ln}(0) = -\infty$ because $e^{-\infty} = 0$

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()


Log10 and ln functions
4
2
0
-2
-4
0
10
20
30
40
50
x
log10x
lnx

- Let’s try transforming 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)
```

Interpretation of regression coefficients for transformed y value

- The regression equation is:

$$\ln(\text{FEV}) = \widehat{\alpha} + \hat{\beta}\, age$$

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

- 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
- $e^{0.087} = 1.09$ – so a one year change in age corresponds to a 9% increase in FEV

- $\ln(FEV) = 0.051 + 0.087\ age$

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

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

Now using height as the independent variable

1. Make a scatter plot of FEV by height
2. Run a regression of FEV on height and examine the output
3. Construct a plot of the residuals vs. the fitted values

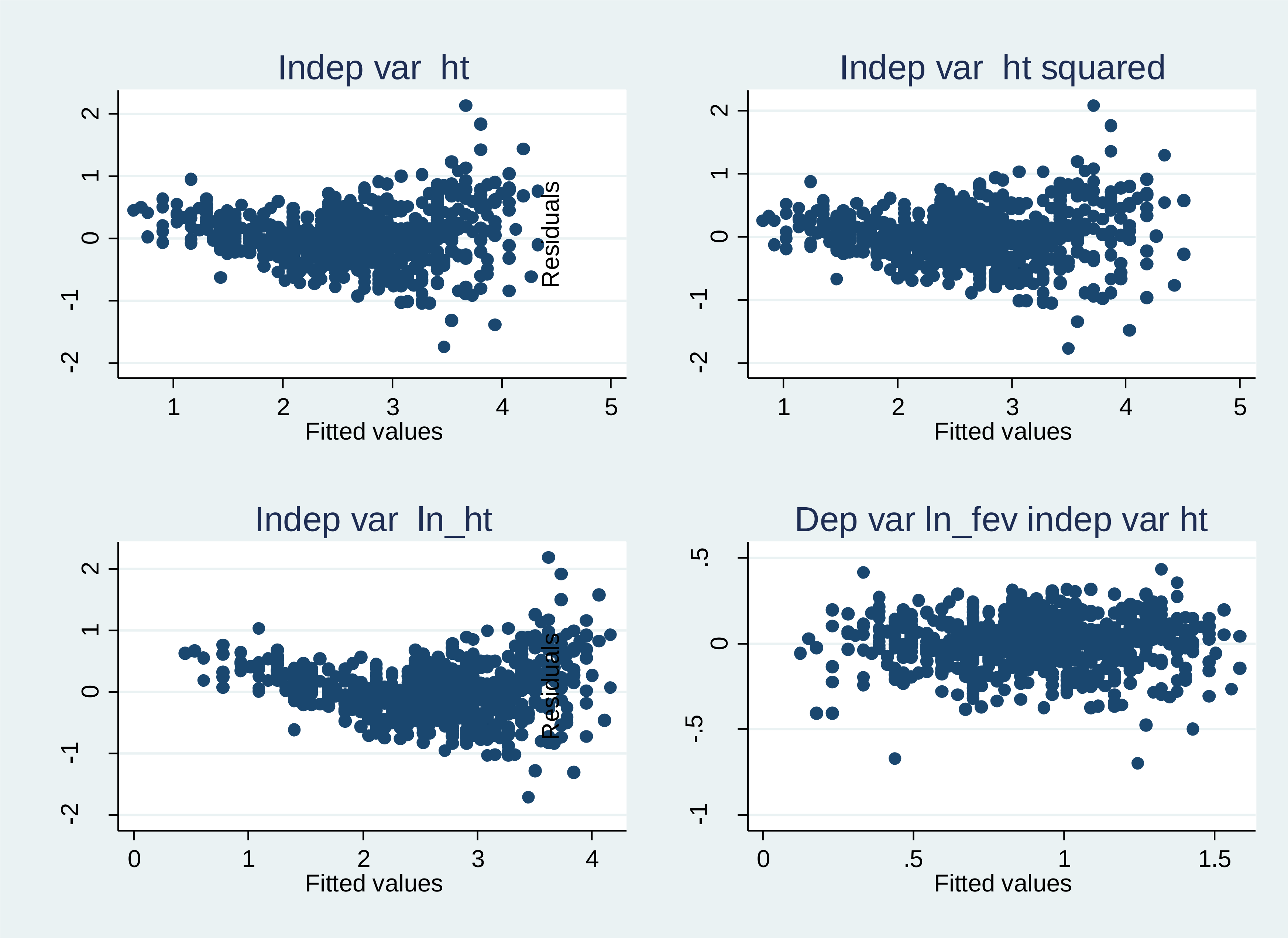
Now using height as the independent variable

4. Consider transformation that might be a better fit

   -You can transform either x or y (or both)

   -The interpretation easier (still additive) if you transform x

1. Run the regression and examine the output
2. Examine the residuals


Indep var ht
Residuals
-2
-1
0
1
2
1
2
3
4
5
Fitted values
Indep var ht squared
-2
-1
0
1
2
1
2
3
4
5
Fitted values
Indep var ln_ht
Residuals
-2
-1
0
1
2
0
1
2
3
4
Fitted values
Dep var ln_fev indep var ht
-1
-.5
0
.5
0
.5
1
1.5
Fitted values

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

1. Using the FEV data, run the regression with FEV as the dependent variable and sex as the independent variable
2. What is the estimate for beta? How is it interpreted?
3. What is the estimate for alpha? How is it interpreted?
4. What hypothesis is tested where it says P>|t|?
5. What is the result of this test?
6. How much of the variance in FEV is explained by sex?

Categorical independent variable

- Remember that the regression equation is

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

- The only variables x can take are 0 and 1
- $\mu_{y|0} = \alpha$ $\quad\quad$ $\mu_{y|1} = \alpha + \beta$
- So the estimated mean FEV for females is $\hat{\alpha}$ and the estimated mean FEV for males is $\hat{\alpha} + \hat{\beta}$
- When we conduct the hypothesis test of the null hypothesis $\beta=0$ what are we testing?
- What other test have we learned that tests the same thing? Run that test.

Categorical independent variables

- In general, you need k-1 dummy or indicator variables (0-1) for a categorical variable with k levels
- One level is chosen as the reference value
- Indicator variables are set to one for each category for only one of the dummy variables, they are set to 0 otherwise

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:

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

- For race group=White

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

- For race group=Asian/PI

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

- For race group=other

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

- You actually don’t have to make the dummy variables yourself (when I was a girl we did have to do)
- All you have to do is tell Stata that a variable is categorical using i. before a variable name
- Run the regression equation for the regression of FEV regressed on age group

```
regress fev i.agecat
```

1. What is the estimated mean FEV for age group = 3-6?
2. What is the estimated mean FEV for age group = 6-9?
3. What is the estimated mean FEV for age group = 15-19?
4. What other test looks at the same thing? Run that test.

- A Stata trick allows you to specify the reference group with the prefix b# where # is the number value of the group that you want to be the reference group.

1. Try out `regress fev b3.agecat`

1. Now the reference category is age group= 9-12
2. Interpret that parameter estimates
3. Note if other output is changed

Multiple regression

- Additional explanatory variables might add to 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

- Because there is natural variation in the response variable, the model we fit is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \varepsilon$$

- 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

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_q x_q$$

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
- Try `regress fev age ht`

- We can test hypotheses about individual slopes
- The null hypothesis is H_0: $\beta_i = \beta_{i0}$ *(=0)* assuming that the values of the other explanatory variables are held constant
- The test statistic $t = \dfrac{\hat{\beta}_i - \beta_{i0}}{s\hat{e}(\hat{\beta}_i)}$

follows a t distribution with <u>n-q-1</u> degrees of freedom (q explanatory variables)

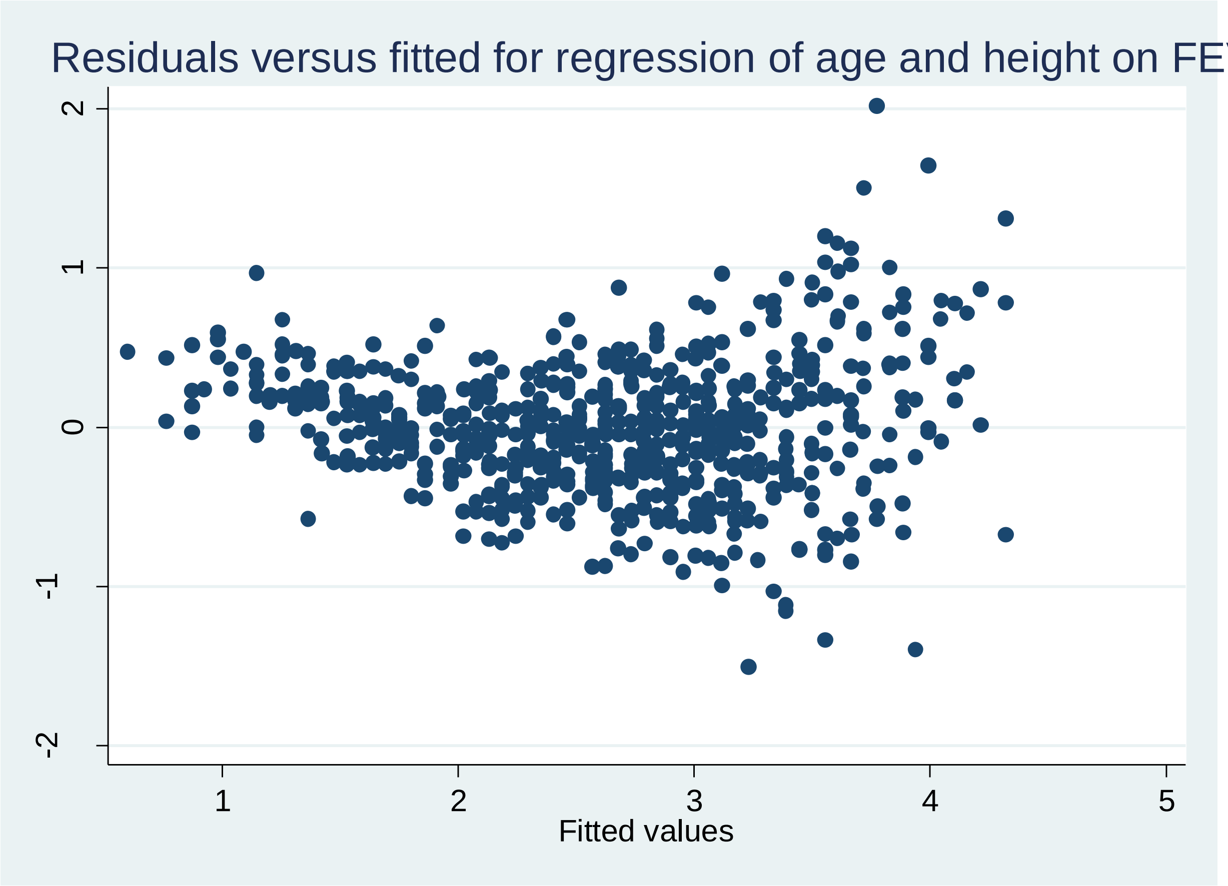
```
. regress fev age ht

      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
------------------------------------------------------------------------------
```

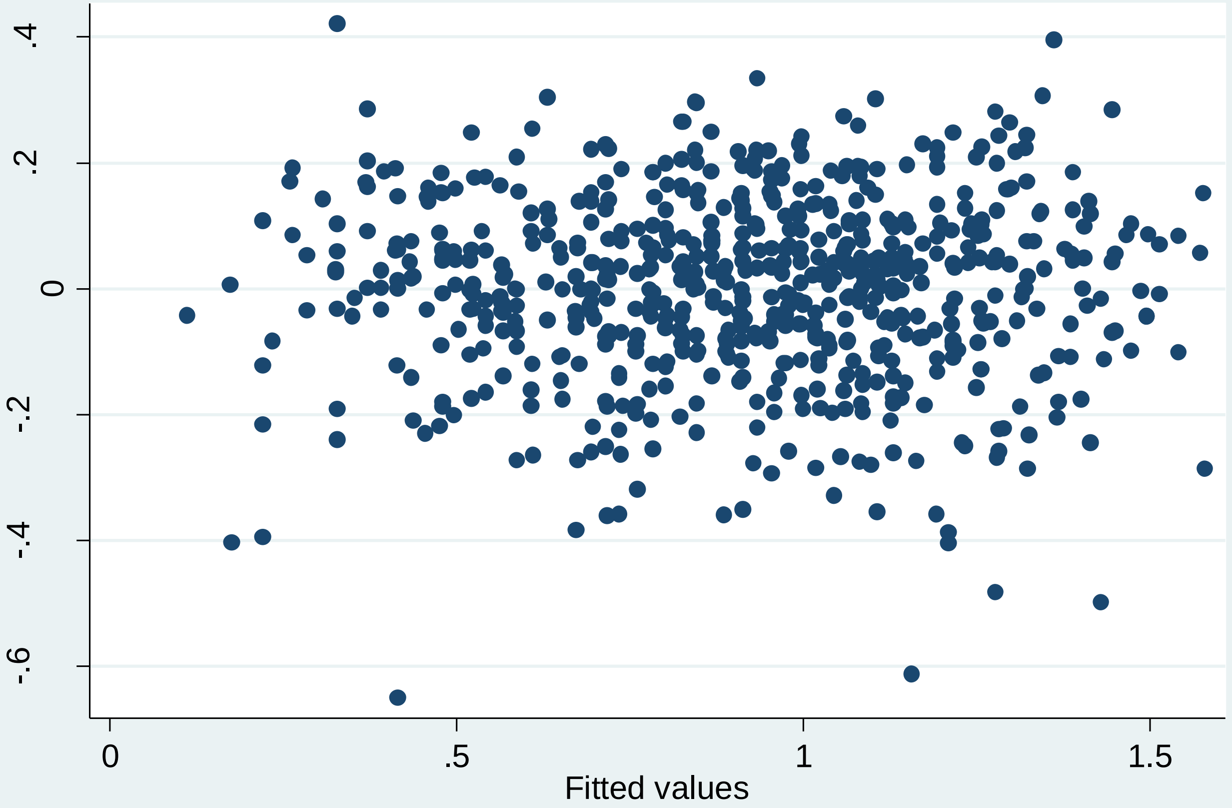
- Now the F-test has 2 degrees of freedom in the numerator because there are 2 explanatory variables
- R^2 will always increase as you add more variables into the model
- The Adj R-squared accounts for the addition of variables and is comparable across models with different numbers of parameters
- Note that the beta for age decreased

Examine the residuals…




```
rvfplot, title(Residuals versus fitted for regression
of age and height on FEV)
```

Residuals versus fitted for regression of age and height on lnFEV



For next time

- Read Pagano and Gauvreau
  - Pagano and Gauvreau Chapters 18-19 (review)
  - Pagano and Gauvreau Chapter 20