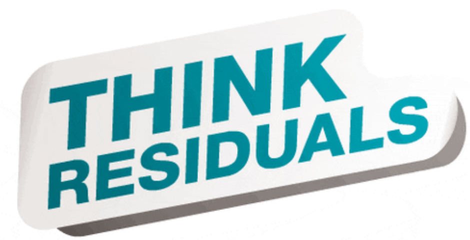
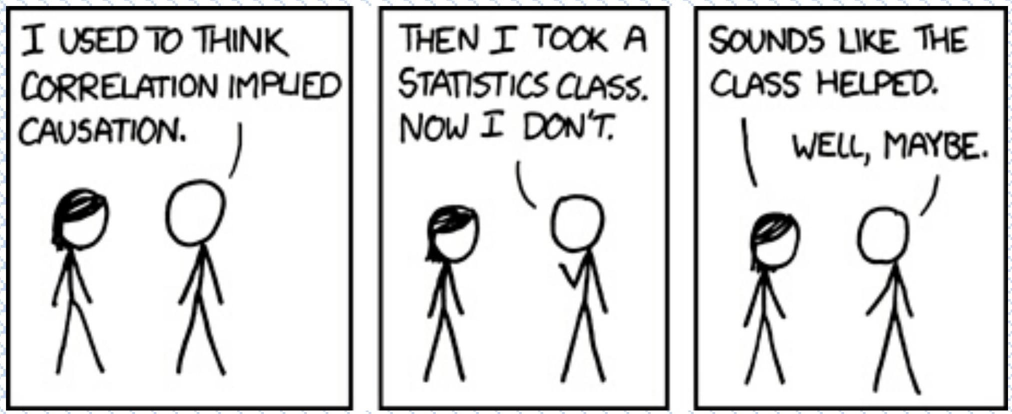
Biostatistics 200

Linear and Logistic Models

Residuals





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 ($\hat{y}$)
  - 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:

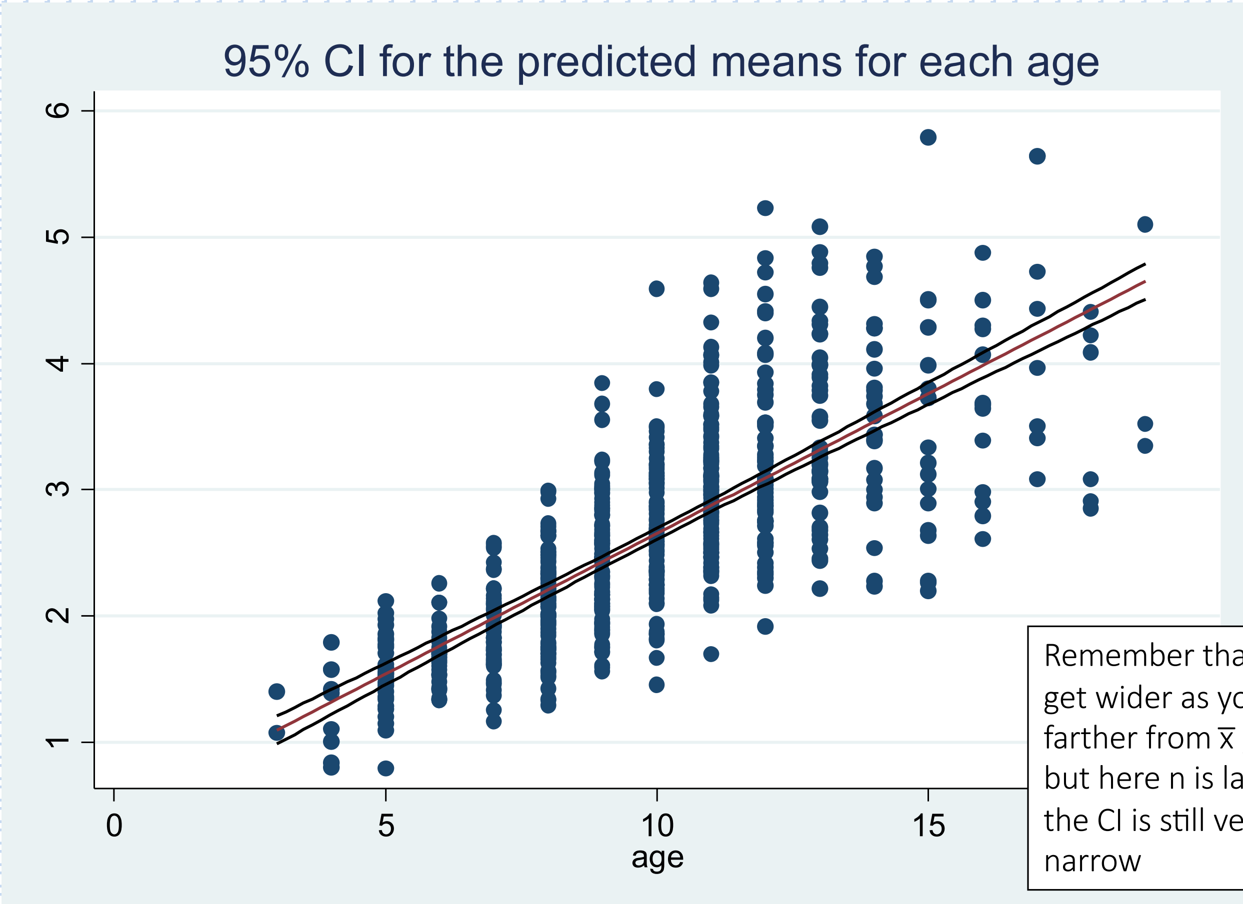
twoway (lfitci fev age)

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 $\overline{x}$;
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

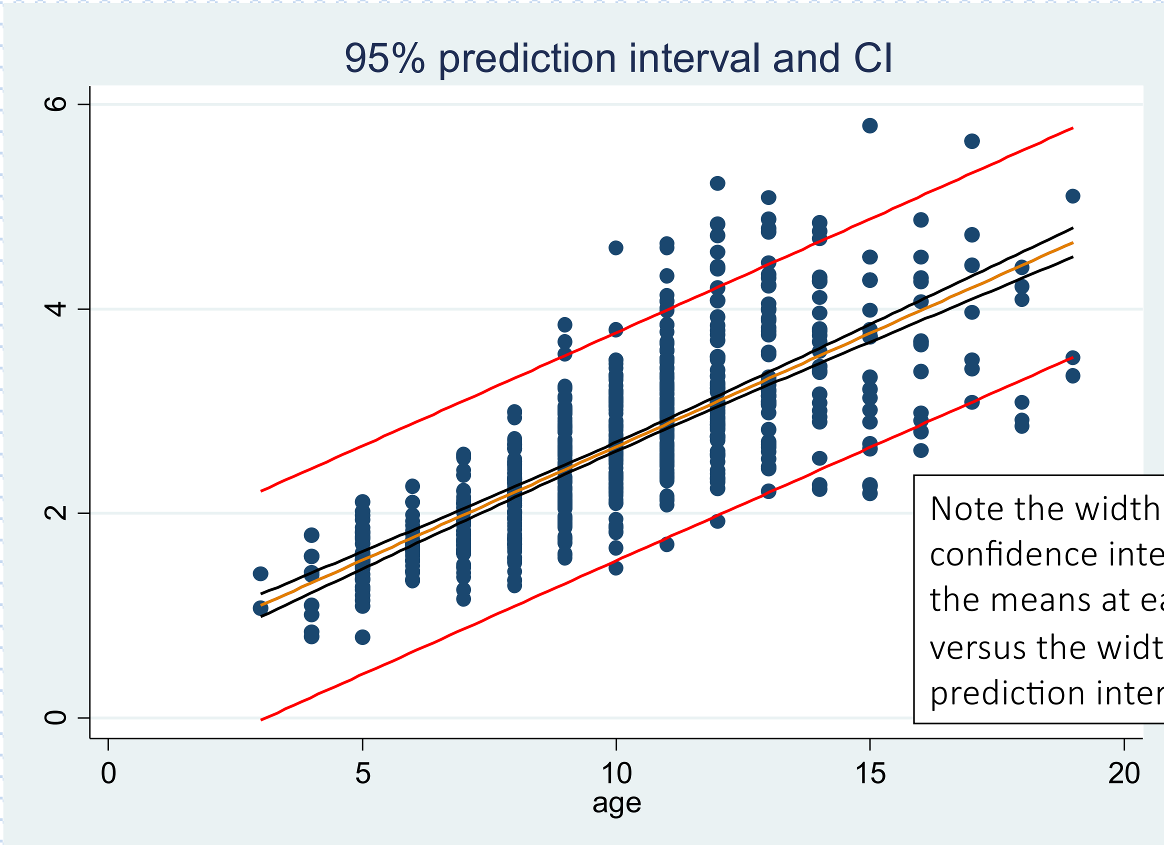
•Stata will calculate these using *predict fev_predse_ind, stdf*

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 |
```

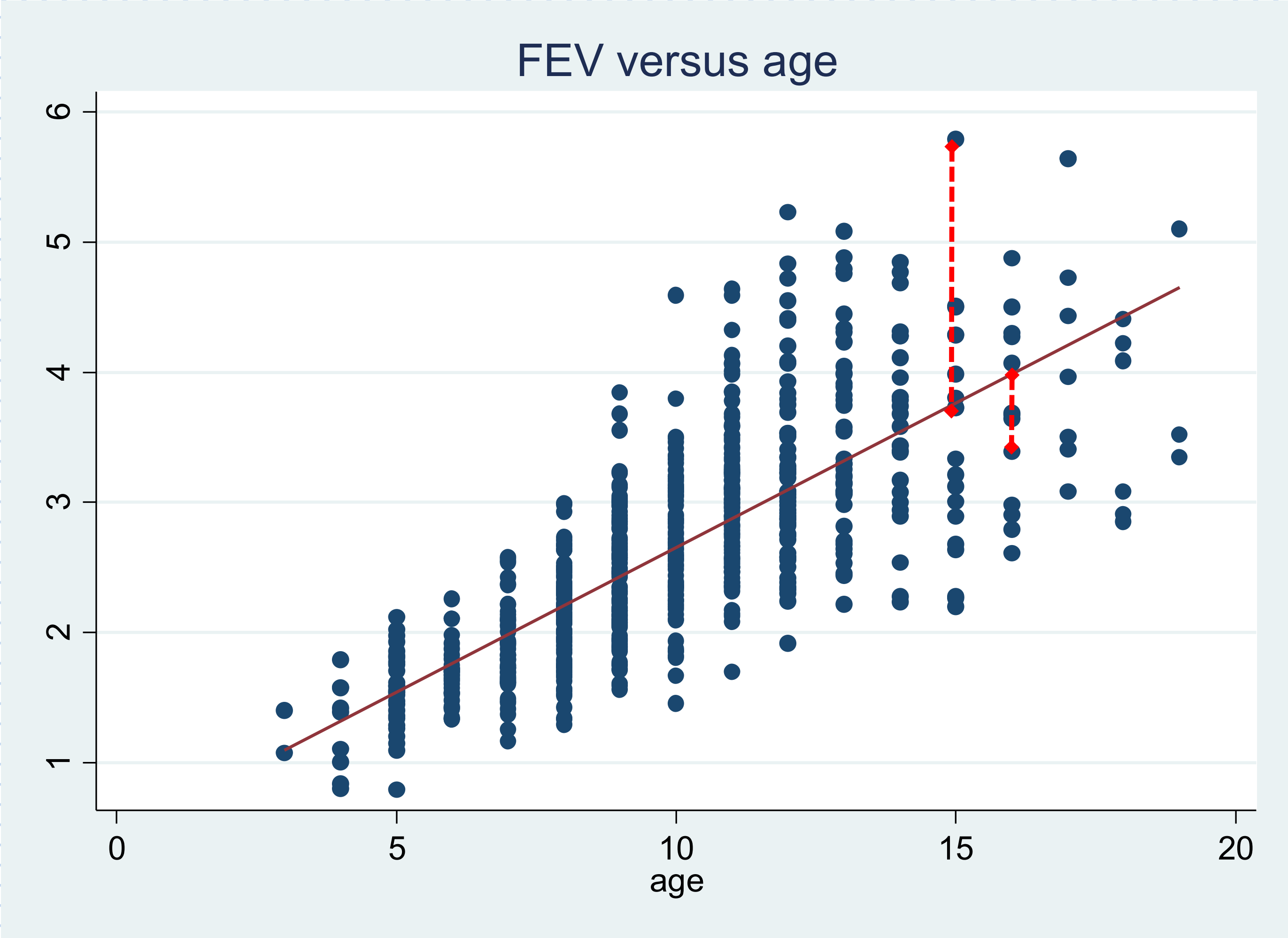


Note the width of the confidence intervals for the means at each x versus the width of the prediction intervals

```
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

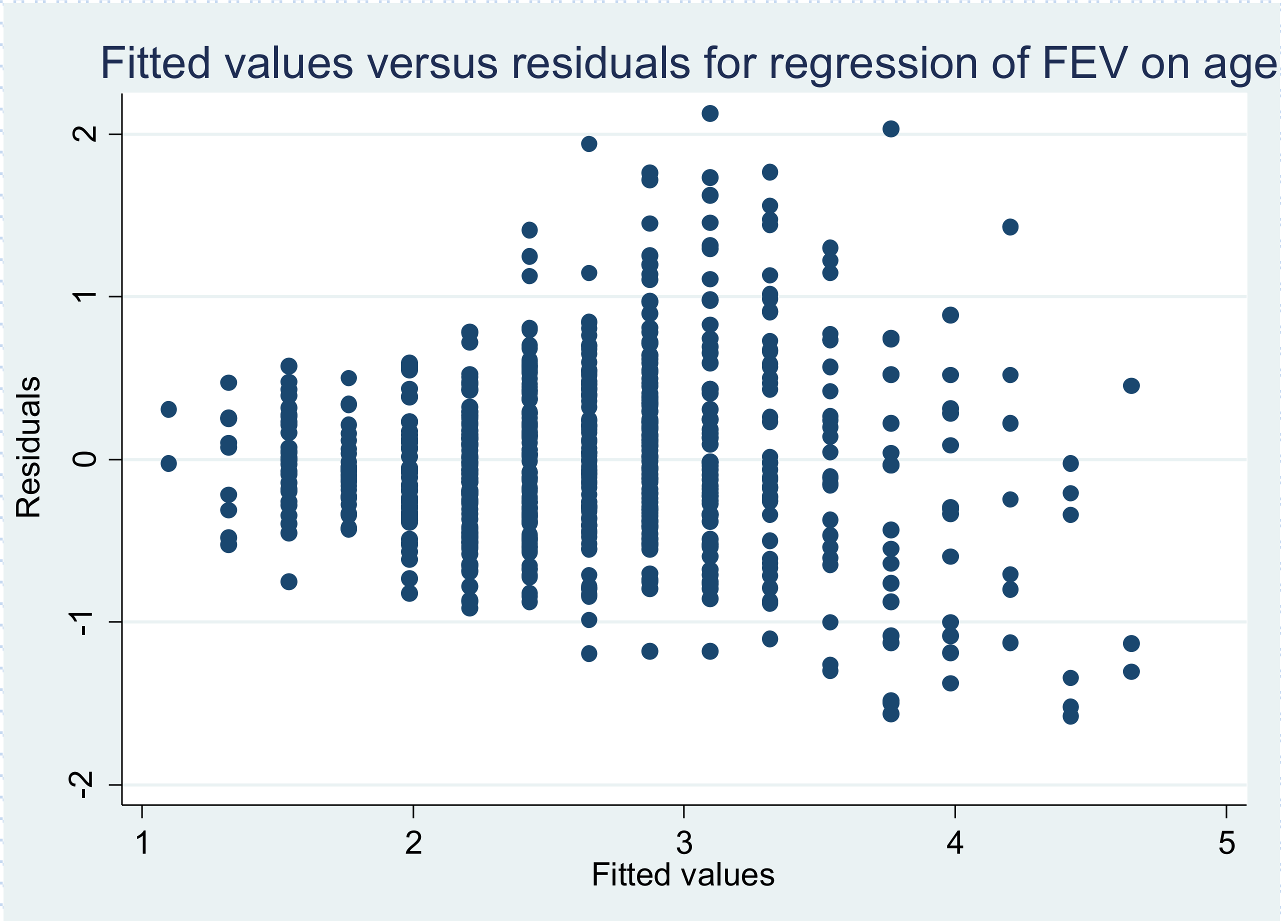

FEV versus age
1
2
3
4
5
6
0
5
10
15
20
age

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

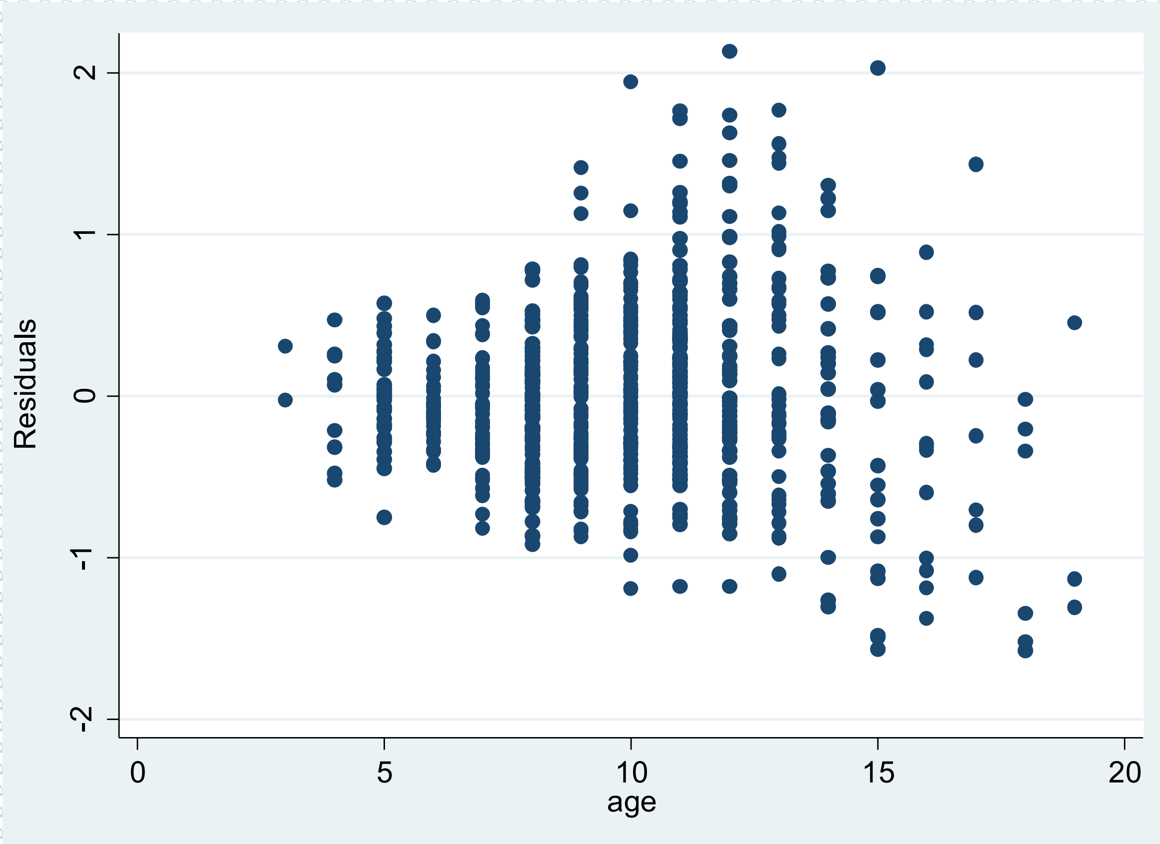


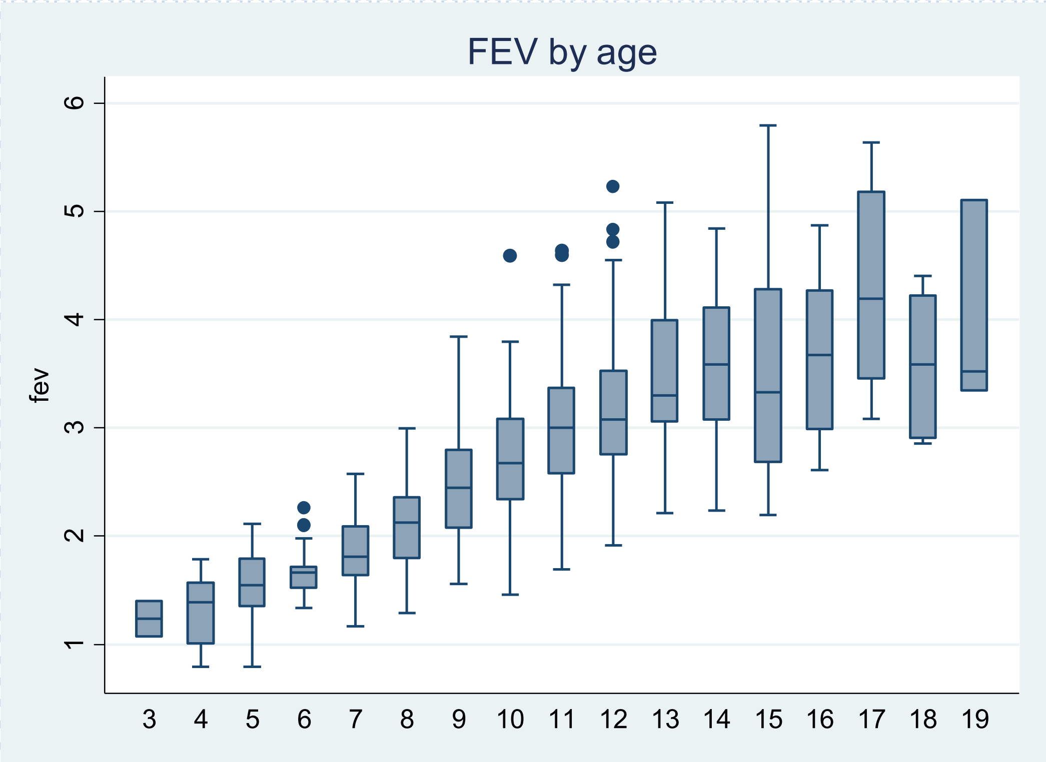

```
rvfplot, title(Fitted values versus residuals for regression of FEV on age)
```

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)