Simple Linear Regression

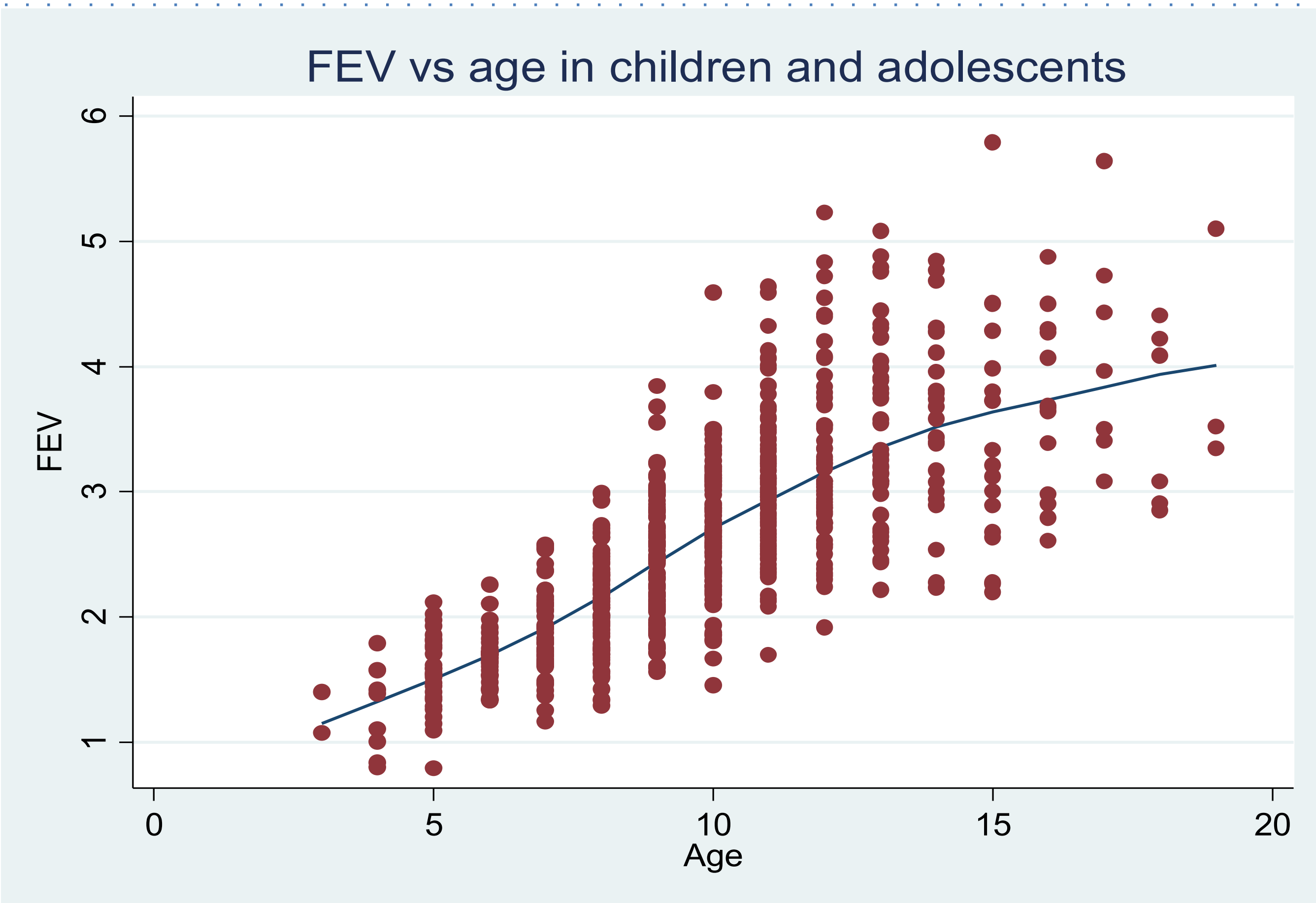
Simple linear regression

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

Example: Forced expiratory volume (FEV)

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

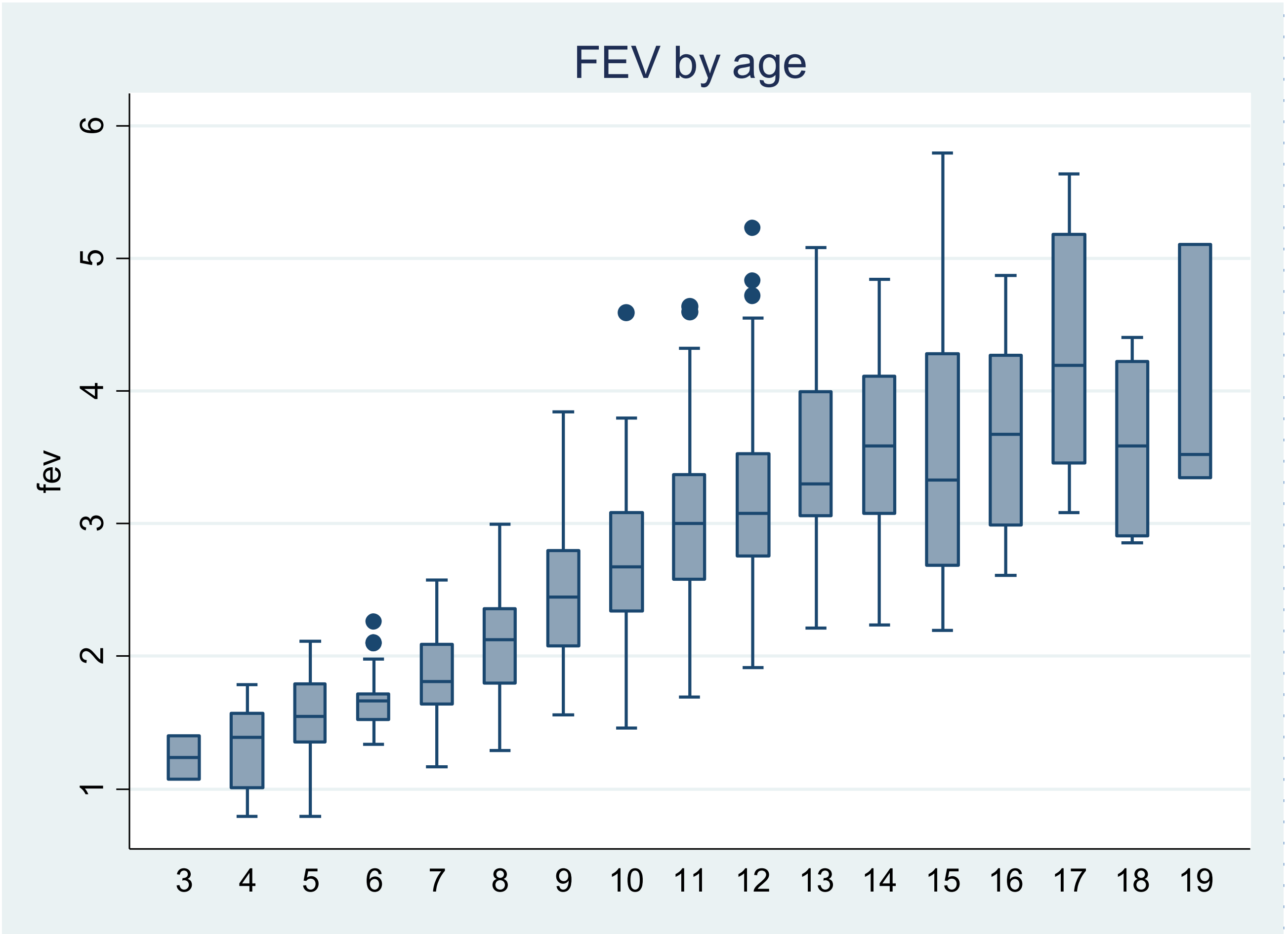
- Tager, I., Weiss, S., Munoz, A., Rosner, B., and Speizer, F. (1983), "Longitudinal Study of the Effects of Maternal Smoking on Pulmonary Function," New England Journal of Medicine, 309(12), 699-703.
- Tager, I., Weiss, S., Rosner, B., and Speizer, F. (1979), "Effect of Parental Cigarette Smoking on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



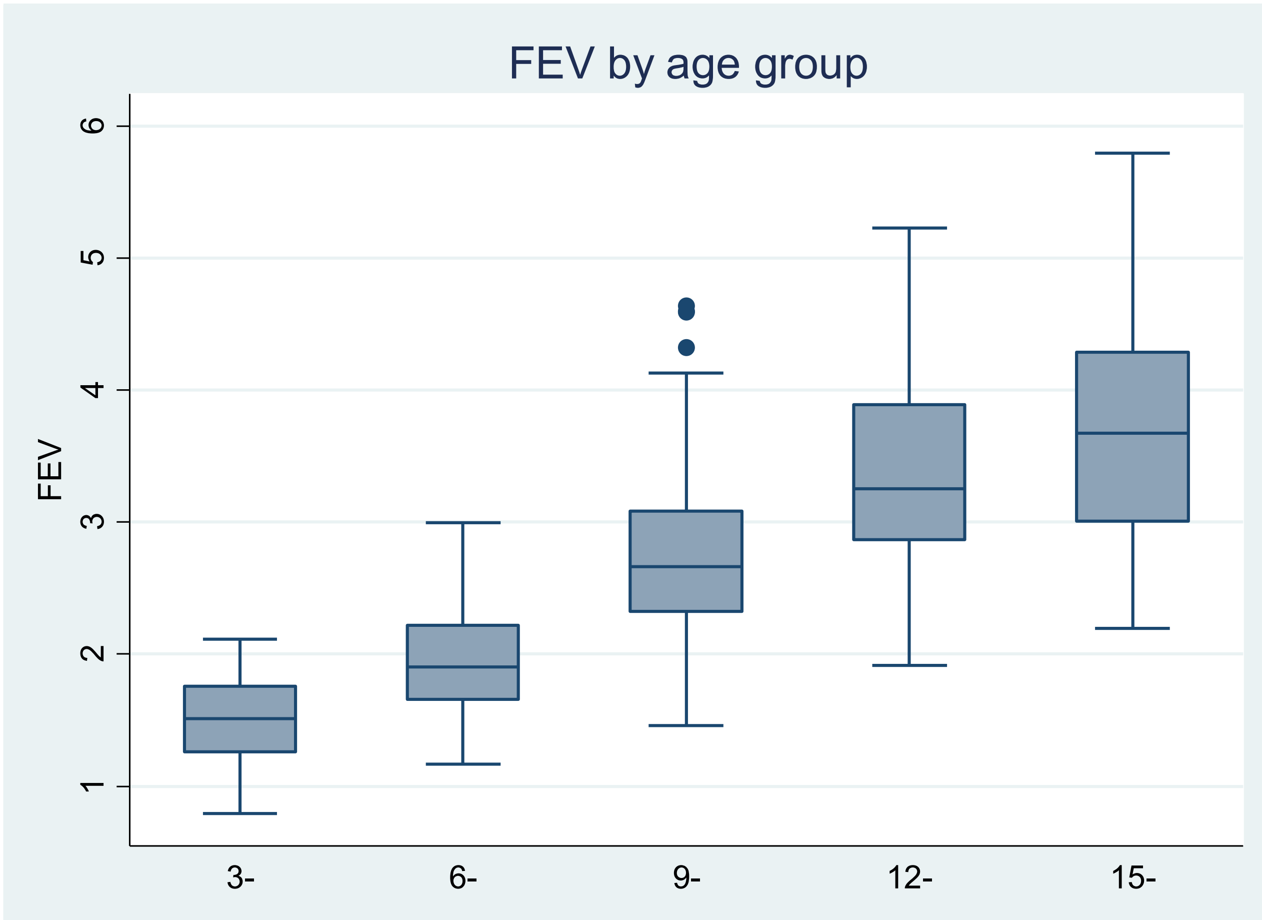

```
twoway (lowess fev age, bwidth(0.8)) (scatter fev age, sort), ytitle(FEV)
xtitle(Age)  legend(off) title(FEV vs age in children and adolescents)
```

Correlation

```
. pwcorr fev age, sig obs

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


. spearman fev age

 Number of obs =     654
Spearman's rho =       0.7984

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




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




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

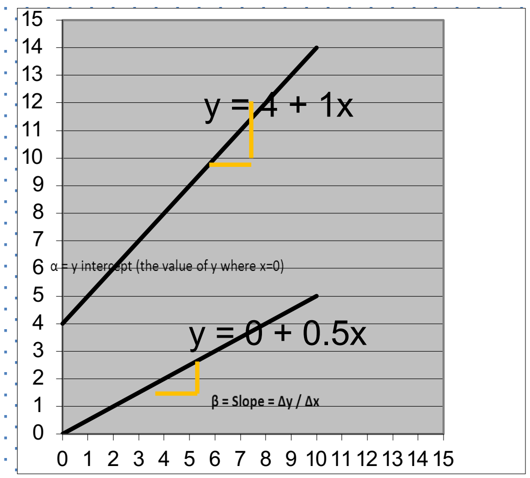
```
. tabstat fev, by(agecat) s(n min median max mean sd)

Summary for variables: fev
     by categories of: agecat

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

Simple linear regression

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

Recall the equation of a straight line

$y = \alpha + \beta x$



Simple linear regression

- Population regression equation is defined as
  $\mu_{y|x} = \alpha + \beta x$
- This is the equation of a straight line
- α and β are constants and are called the coefficients of the equation
- α is the y-intercept of the line $y=\alpha + \beta x$
- Plugging back in to the population regression equation, α is the mean value of Y when X=0 e.g. $\mu_{y|0} = \alpha + \beta X = \alpha$
- β is the slope of the line
- β is the change in the mean value of y that corresponds to a one-unit increase in X
- e.g. X=3 vs. X=2
  $\mu_{y|3} - \mu_{y|2} = (\alpha + \beta*3) - (\alpha + \beta*2) = \beta$

Simple linear regression

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

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

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

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

Simple linear regression

- Statistical assumptions of linear regression
  - X's are measured without error
    - Violations of this cause the coefficients to attenuate toward zero
  - For each value of X, the Y's are normally distributed with mean $\mu_{y|x}$ and standard deviation $\sigma_{y|x}$
  - The regression equation is correct $\mu_{y|x} = \alpha + \beta$
  - Homoscedasticity – the standard deviation of y at each value of X is constant; $\sigma_{y|x}$ the same for all values of X
  - All the y_i 's are independent
    - You can't guess the y value for one person (or observation) based on the outcome of another

- <u>We do not need the X's to be normally distributed, just the Y's at each value of X</u>