Biostat 200
Lecture 9

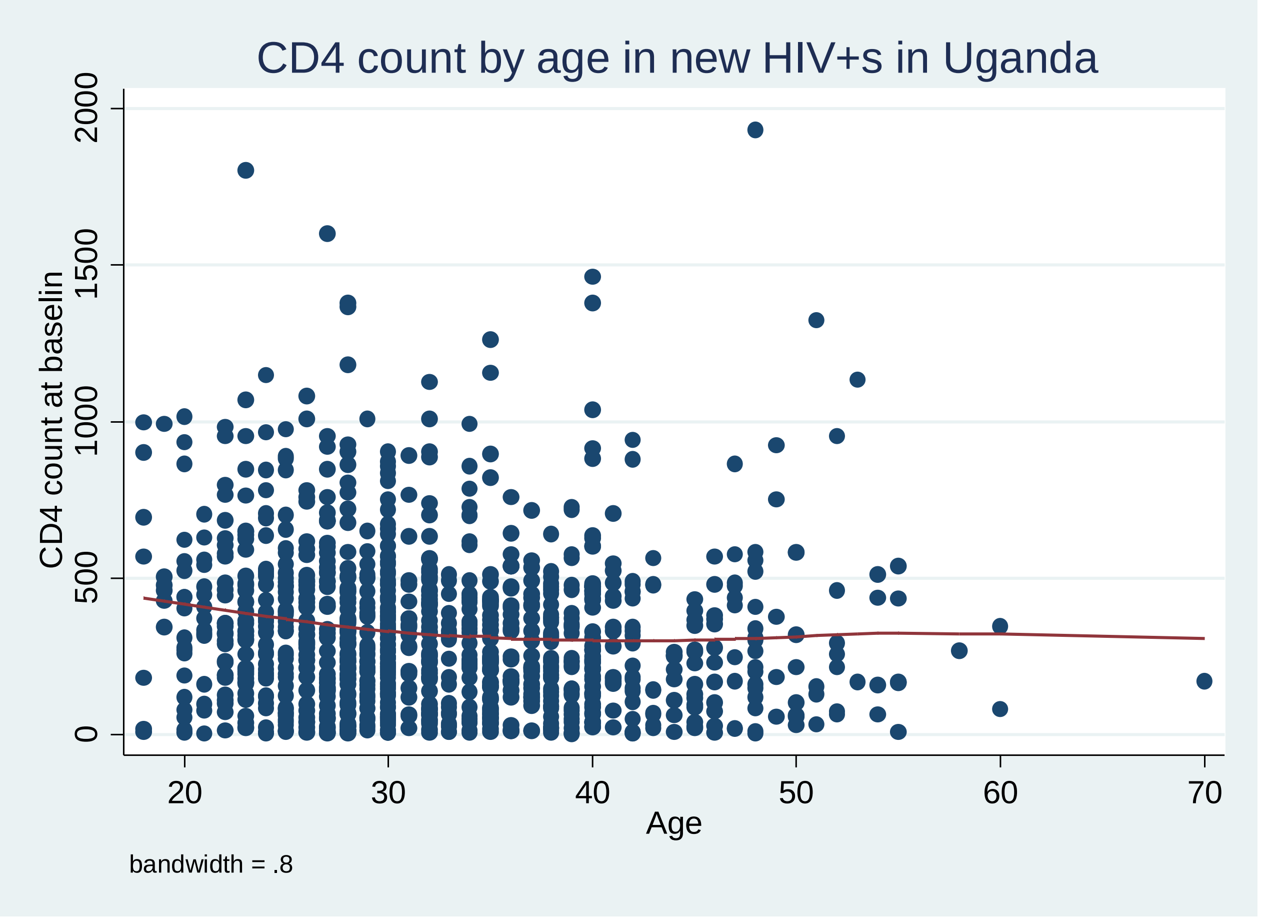
Categorical data analyses

- Another type of paired categorical data
  - Inter-rater agreement, kappa statistic
  - Goes from 0 to 1
  - It uses expected cell frequencies, like we calculated in the previous lecture (Slides 12-13).
  - It compares observed agreement to expected agreement
    - Agreement is defined as # concordant / n
  - In Stata, use kap command (http://www.stata.com/manuals13/rkappa.pdf)

Scatterplot

- Back to continuous outcomes
- T-test, ANOVA, Wilcoxon rank-sum test, Kruskal-Wallis test compare 2 or more independent samples
  - e.g. CD4 cell count by sex or alcohol consumption category
- The scatterplot is a simple method to examine the relationship between 2 continuous variables

Scatter plot




```
lowess cd4count age, xtitle(Age) ytitle(CD4 count at
baselin) title(CD4 count by age in new HIV+s in Uganda)
```

Correlation

- Correlation is a method to examine the relationship between 2 continuous variables
  - Does one increase with the other?
    - E.g. Does CD4 count increase with increasing age?
- Both variables are measured on the same people (or unit of analysis)
- Correlation assumes a *linear* relationship between the two variables
- Correlation is symmetric
  - The correlation of A with B is the same as the correlation of B with A

Correlation

- Correlation is a measure of the relationship between two random variables X and Y
- If the variables increase together (or oppositely), then the average of X*Y will be large (in absolute terms)
- We subtract off the mean and divide by the standard deviation to standardize (i.e. so correlations will be between -1 and 1)

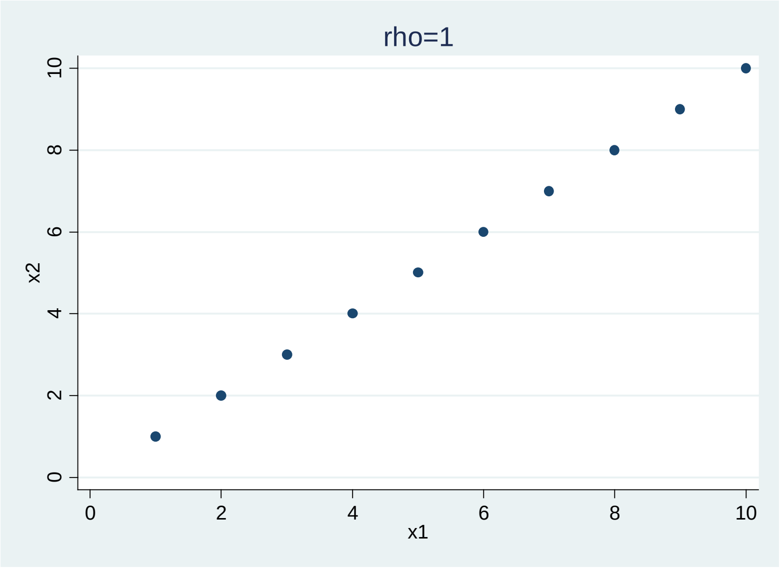
Correlation

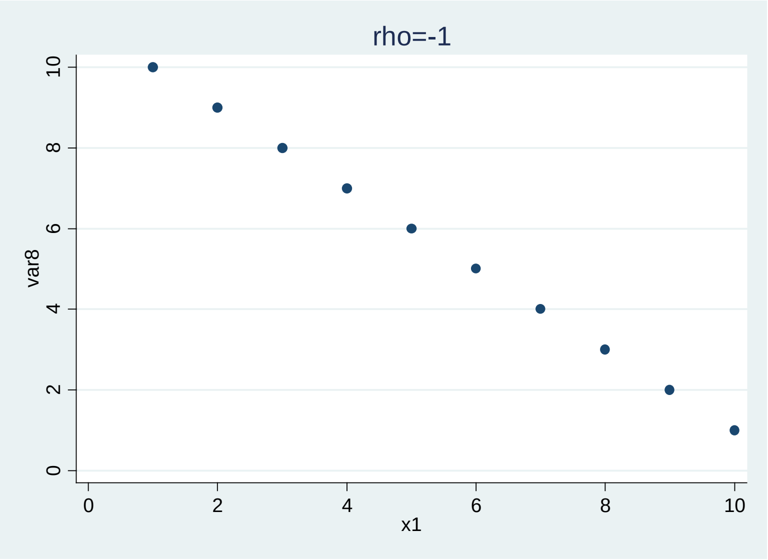
- A correlation is defined as

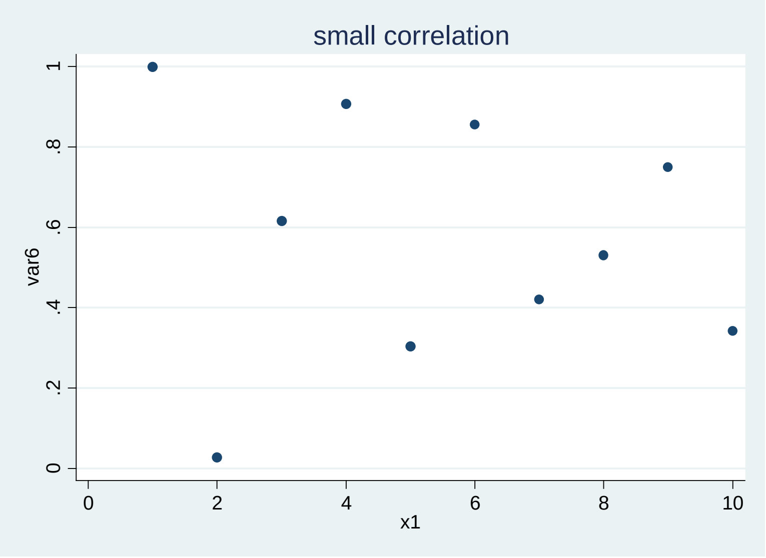
$$\rho = average\left[\left(\frac{X-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_y}{\sigma_y}\right)\right]$$

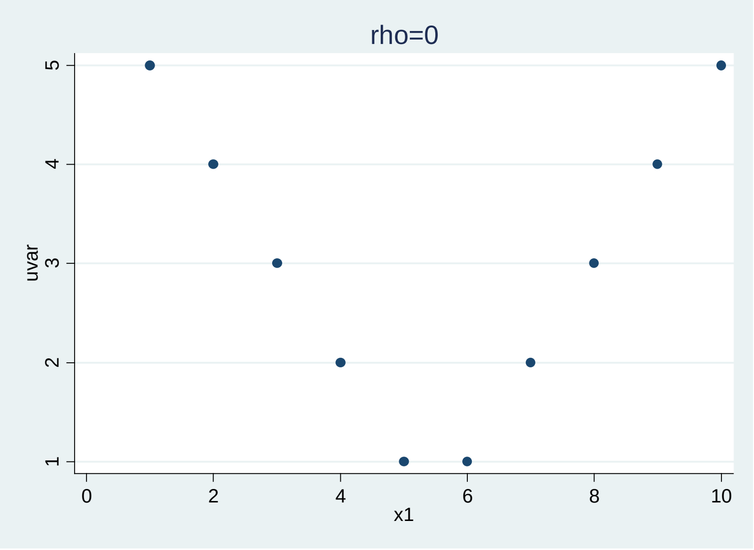
- Correlations can be comparable across variables with different means and variability
- Correlation does not imply causation

Correlation









Correlation

- ρ lies between -1 and 1
- -1 and 1 are perfect correlations, 0 is no correlation
- An estimator of the population correlation ρ is Pearson's correlation coefficient denoted r
- It is estimated by: $$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Correlation

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$= \frac{1}{(n-1)}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right]}\sqrt{\left[\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}\right]}}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]}\sqrt{\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]}}$$

Correlation: hypothesis testing

- To test whether there is a correlation between two variables, our hypotheses are

  $H_0 : \rho=0$ and $H_A : \rho\neq0$

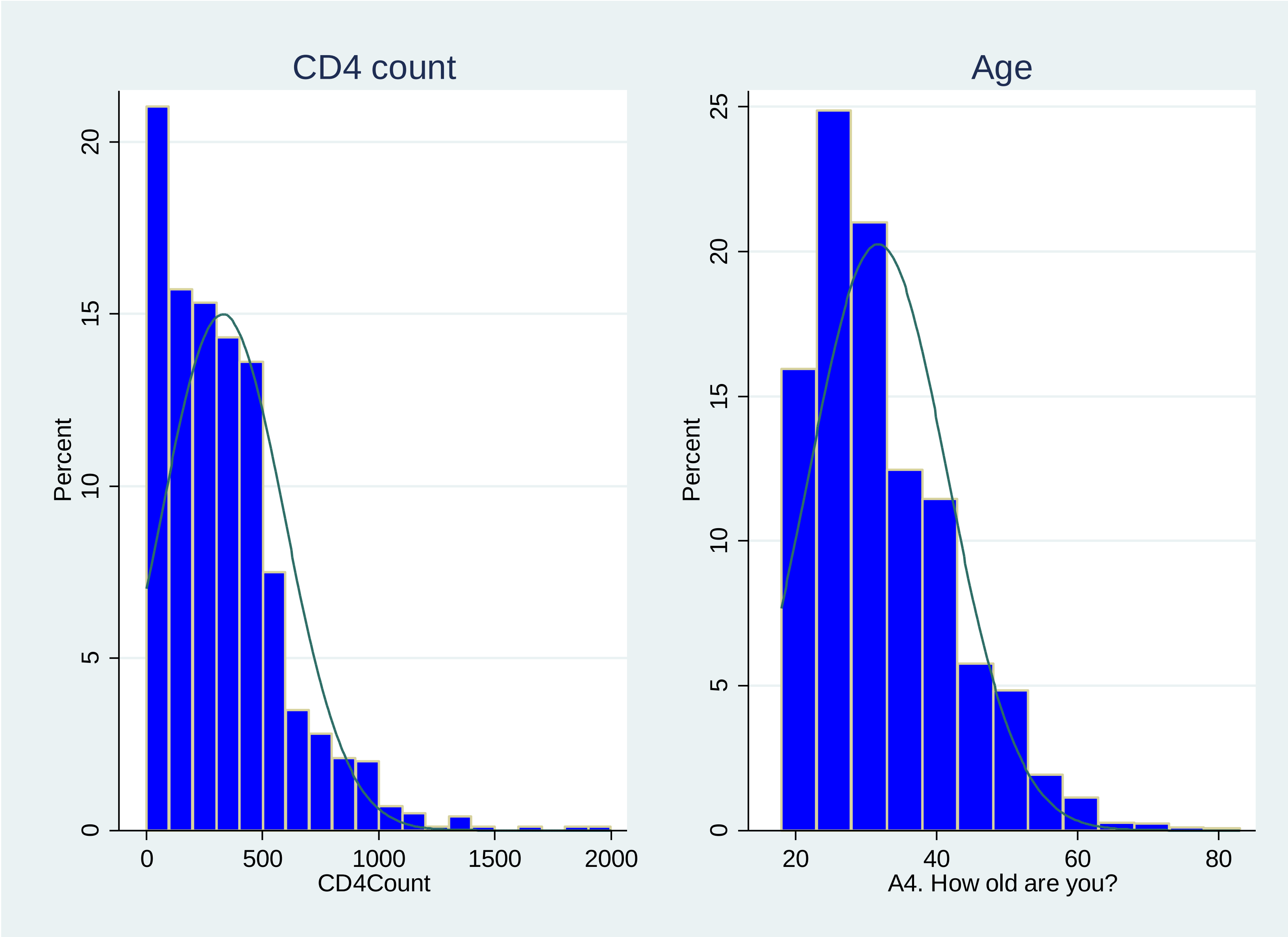
- We need to calculate a test statistic for r
- The test statistic is $t = \dfrac{r-0}{se(r)}$

$$\text{where } se(r) = \sqrt{\frac{1-r^2}{n-2}}$$

$$\text{so } t = r\sqrt{\frac{n-2}{1-r^2}}$$

Correlation: hypothesis testing

- The test statistic follows a t distribution with n-2 degrees of freedom under the null
- The assumptions
  - The pairs of observations (x_i , y_i) were obtained from a random sample
  - X and Y are normally distributed


CD4 count
Percent
0
5
10
15
20
0
500
1000
1500
2000
CD4Count
Age
Percent
0
5
10
15
20
25
20
40
60
80
A4. How old are you?

Correlation example

```
 corr cd4count age
(obs=999)

             | cd4count    age_b
-------------+------------------
    cd4count |   1.0000
       age_b |  -0.1133   1.0000
```

Note that the hypothesis test is only of ρ=0, no other null
Also note that the correlation is the linear relationship only

Correlation example

pwcorr var1 var2, sig obs

```
. pwcorr cd4count age, sig obs

             | cd4count    age_b
-------------+------------------
    cd4count |   1.0000
             |
             |      999
             |
       age_b |  -0.1133   1.0000
             |   0.0003
             |      999     3387
             |
```

Spearman rank correlation (nonparametric)

- Pearson's correlation coefficient is very sensitive to extreme values
- Spearman rank correlation calculates the Pearson correlation on the *ranks* of each variable
- The Pearson correlation coefficient is calculated, but the data values are replaced by the ranks
- The Spearman rank correlation coefficient is

$$r_s = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_{ri} - \bar{x}_r}{s_{rx}} \right) \left(\frac{y_{ri} - \bar{y}_r}{s_{ry}} \right)$$

Spearman rank correlation (nonparametric)

- The Spearman rank correlation ranges between -1 and 1 as does the Pearson correlation
- We can test the null hypothesis that ρ=0
- The test statistic for n>10 is with n-2 degrees of freedom

$$t_s = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

is compared to the t distribution

```
 spearman  cd4count age, stats(rho obs p)

 Number of obs =     999
Spearman's rho =      -0.1319

Test of Ho: cd4count and age_b are independent
    Prob > |t| =       0.0000
```

Matrix of Pearson correlations

- We can calculate a correlation matrix by listing multiple variable names
- Beware of which n's are used (use listwise option to get all n's equal)

```
 pwcorr cd4count age nsupport_b auditc, sig obs

             | cd4count    age_b nsuppo~b   auditc
-------------+------------------------------------
    cd4count |   1.0000
             |
             |      999
             |
       age_b |  -0.1133   1.0000
             |   0.0003
             |      999     3387
             |
  nsupport_b |   0.0067   0.3597   1.0000
             |   0.8340   0.0000
             |      994     3370     3372
             |
      auditc |   0.0670   0.0580   0.0468   1.0000
             |   0.0392   0.0009   0.0078
             |      948     3242     3230     3244
             |
```

Matrix of Spearman correlations

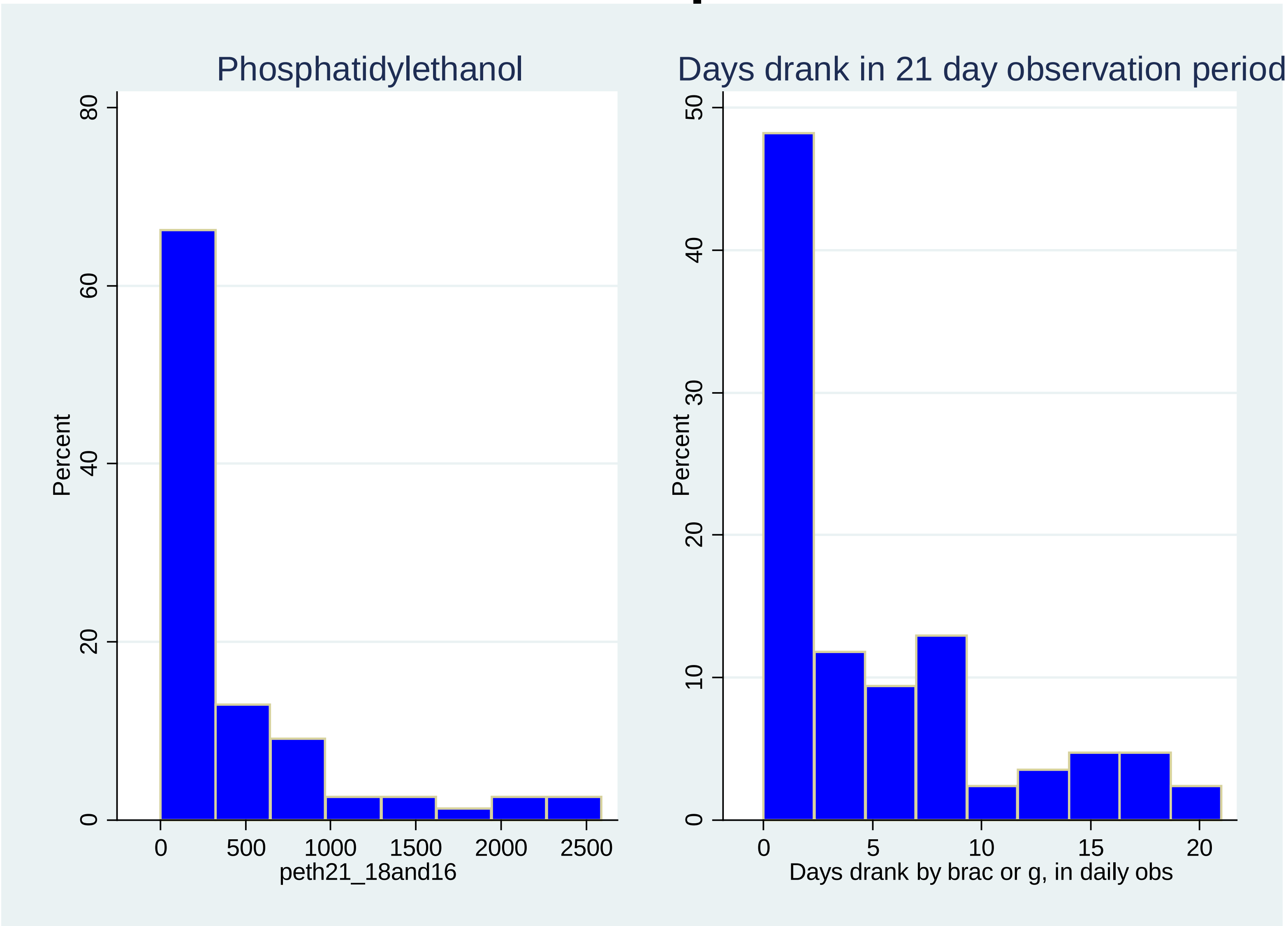
```
spearman cd4count age nsupport_b auditc, pw stats(rho obs p)

+-----------------+
|  Key            |
|-----------------|
|   rho           |
|   Number of obs |
|   Sig. level    |
+-----------------+

             | cd4count    age_b nsuppo~b   auditc
-------------+------------------------------------
    cd4count |   1.0000
             |      999
             |
             |
       age_b |  -0.1319   1.0000
             |      999     3387
             |   0.0000
             |
  nsupport_b |   0.0118   0.4055   1.0000
             |      994     3370     3372
             |   0.7093   0.0000
             |
      auditc |   0.0945   0.0777   0.0447   1.0000
             |      948     3242     3230     3244
             |   0.0036   0.0000   0.0111
             |
```

Here if you drop the “pw” option you get all n’s equal

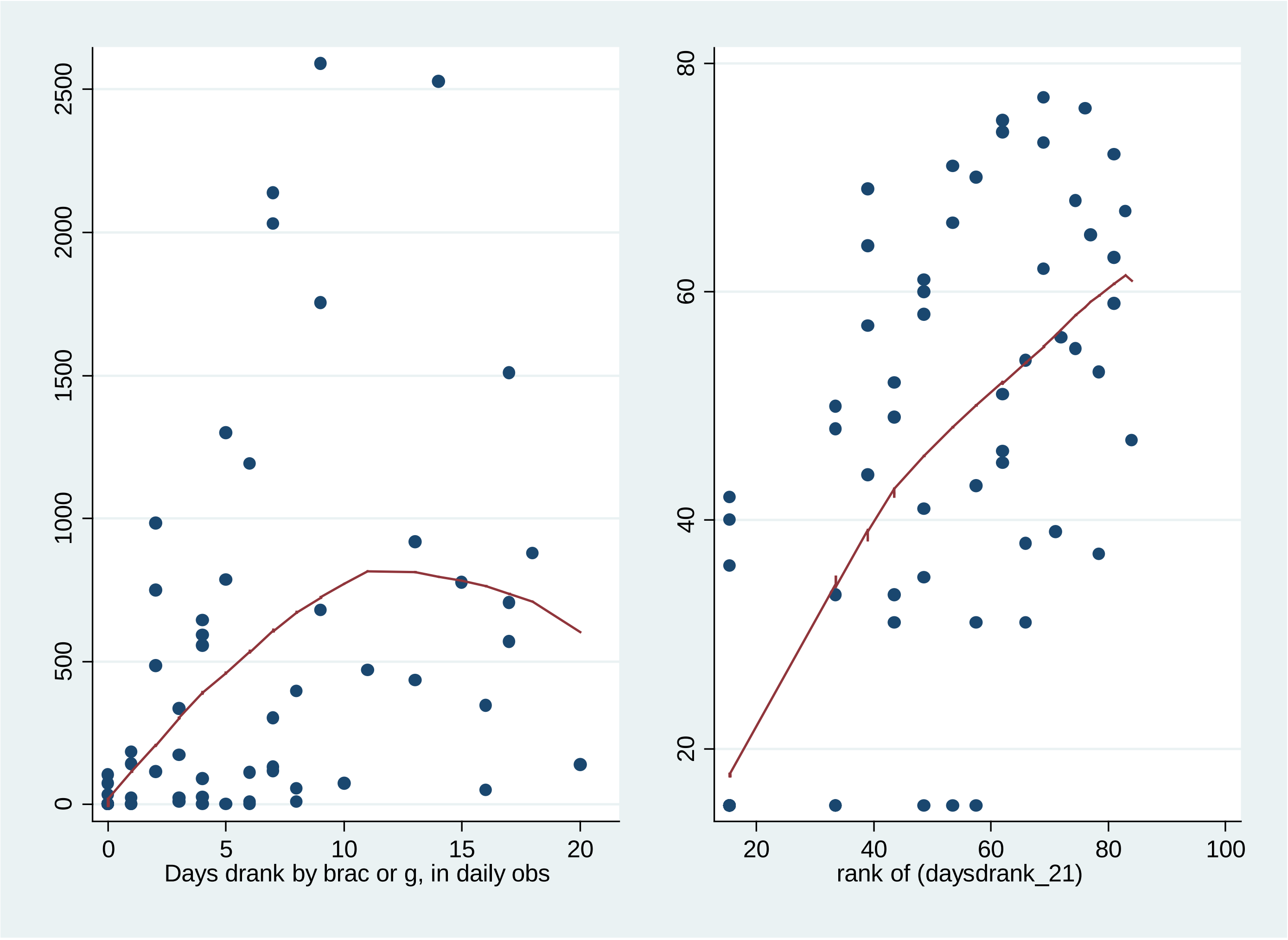
Pearson vs. Spearman

	Pearson	Spearman
Sensitivity to outliers	more sensitive	less sensitive
Can be used with ordinal data	no	yes
Assumes data are from a normal distribution	yes	no
Detects non-linear relationship	no	yes

Biomarker and alcohol consumption



PEth Biomarker of alcohol consumption vs. days drinking (raw data vs. ranks)



Pearson and Spearman correlations

```
. pwcorr peth21_18and16 daysdrank_21, obs sig

             | peth2~16 days~_21
-------------+------------------
peth21_18~16 |   1.0000
             |
             |       77
             |
daysdrank_21 |   0.4717   1.0000
             |   0.0000
             |       77       85
             |
-----------------------------------
. spearman peth21_18and16 daysdrank_21

 Number of obs =      77
Spearman's rho =       0.7413

Test of Ho: peth21_18and16 and daysdrank_21 are independent
    Prob > |t| =       0.0000
```

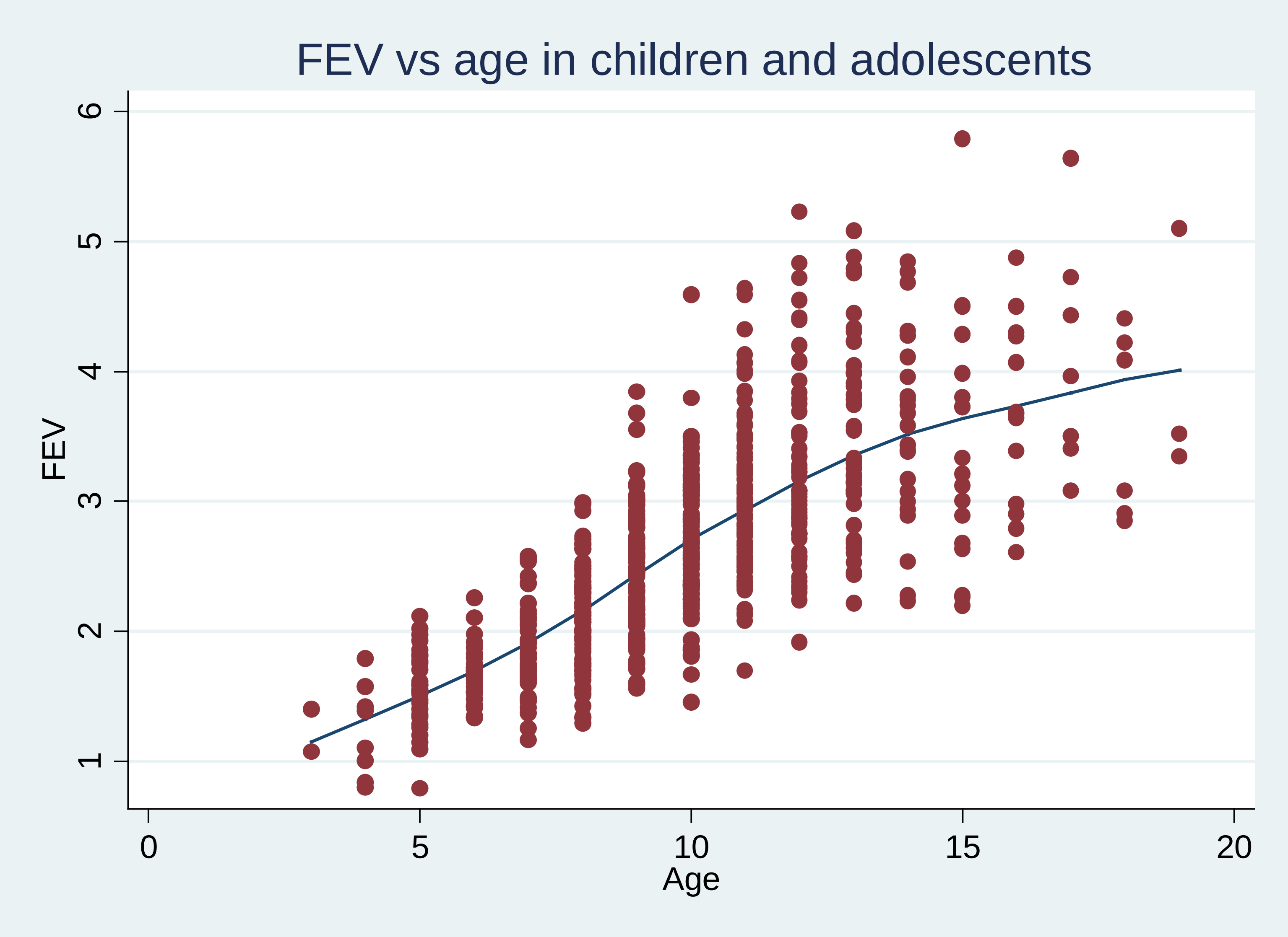
Simple linear regression

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

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.



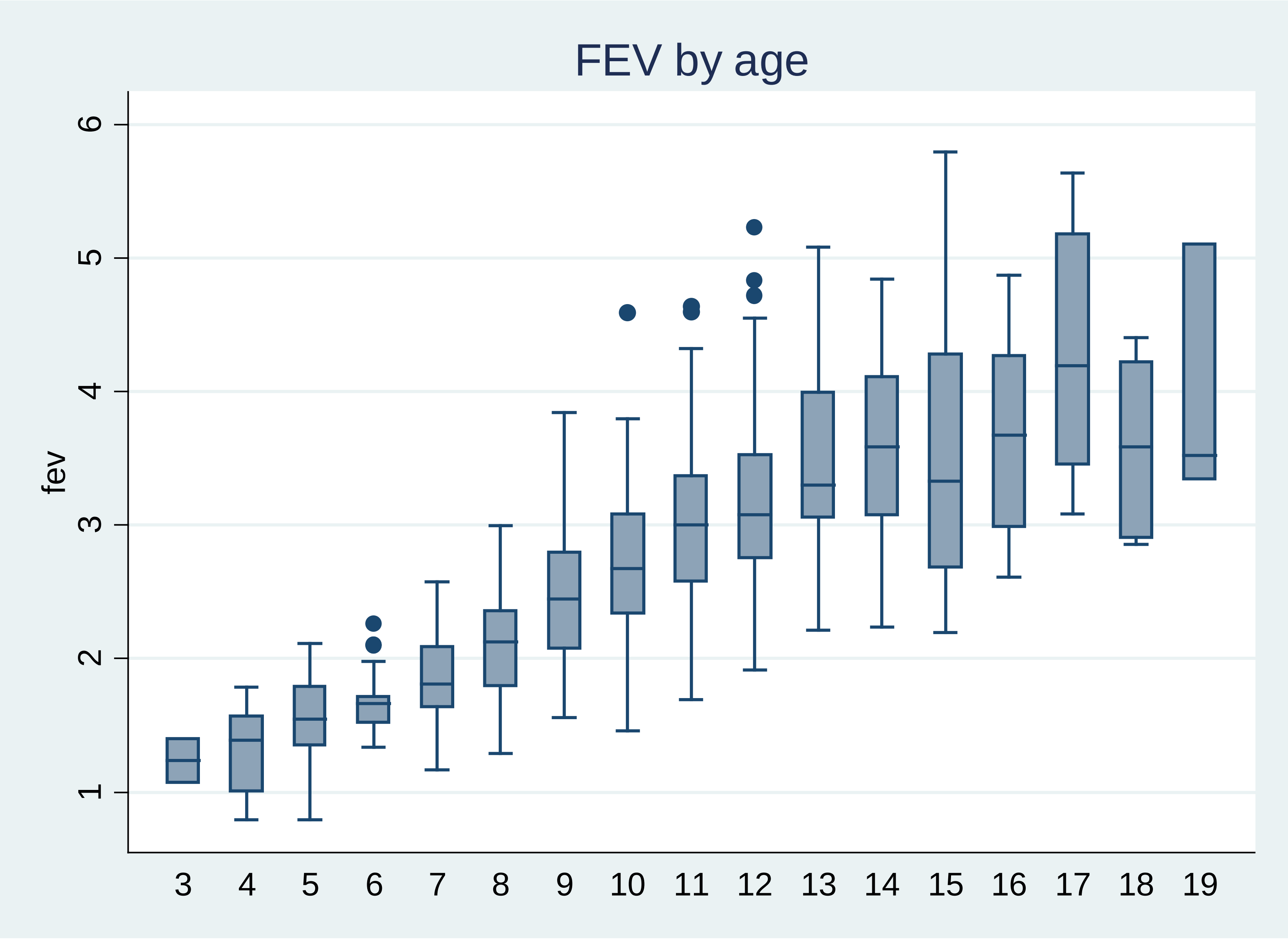

```
twoway (lowess fev age, bwidth(0.8)) (scatter fev age,
sort), ytitle(FEV) xtitle(Age)  legend(off) title(FEV vs
age in children and adolescents)
```

Correlation

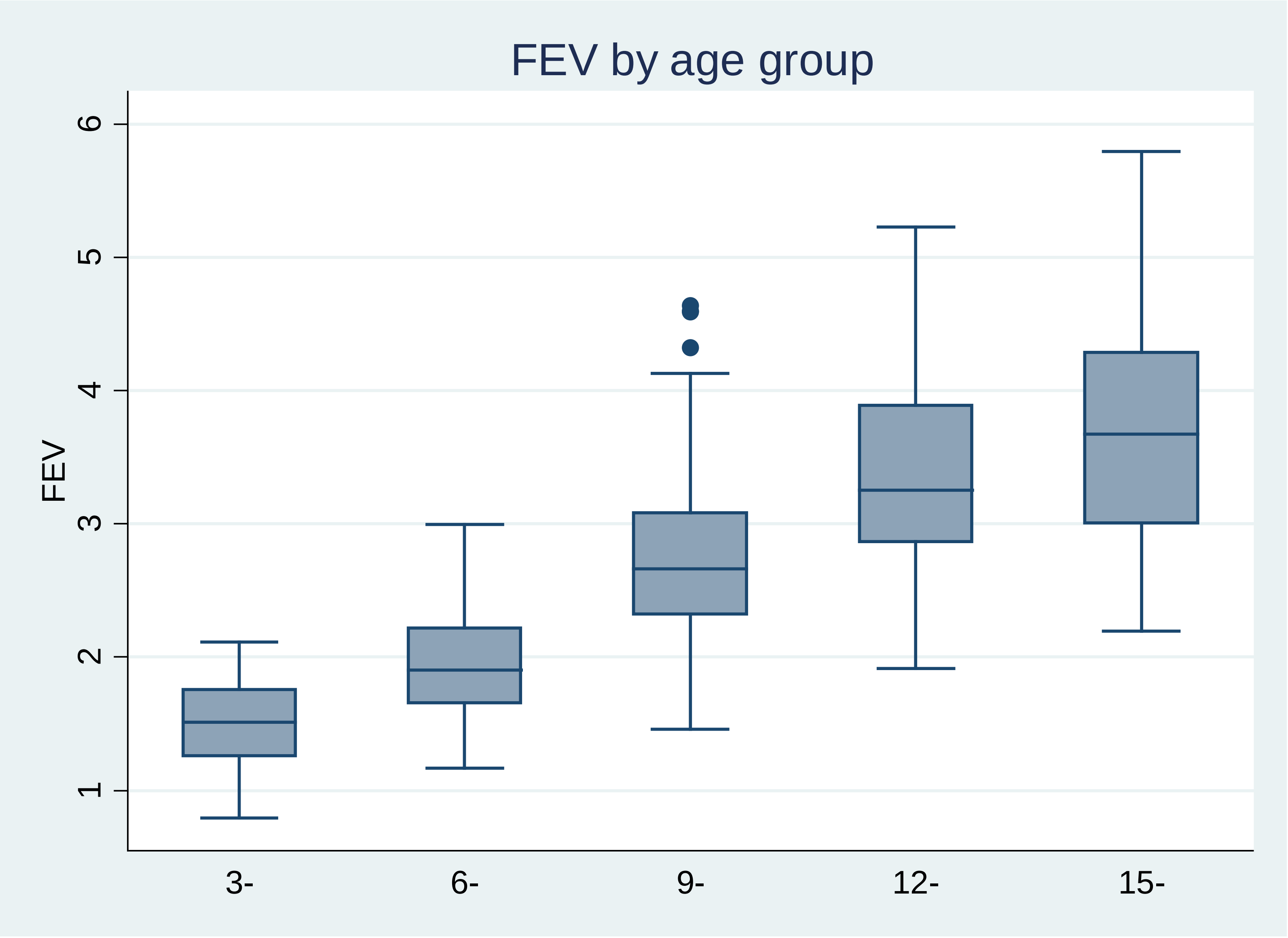
```
. pwcorr fev age, sig obs

             |      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
```

Concept of $\mu_{y|x}$ and $\sigma_{y|x}$

- Consider variables X and Y that are thought to be related
- You want to know how a difference in X translates to a difference in Y
- Plot X versus Y, but instead of using all values of X, categorize X into several categories
- What you get would look like a boxplot of Y by the grouped values of X
- Each of the groups of X has a mean of Y $\mu_{y|x}$ and a standard deviation $\sigma_{y|x}$




```
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)
```

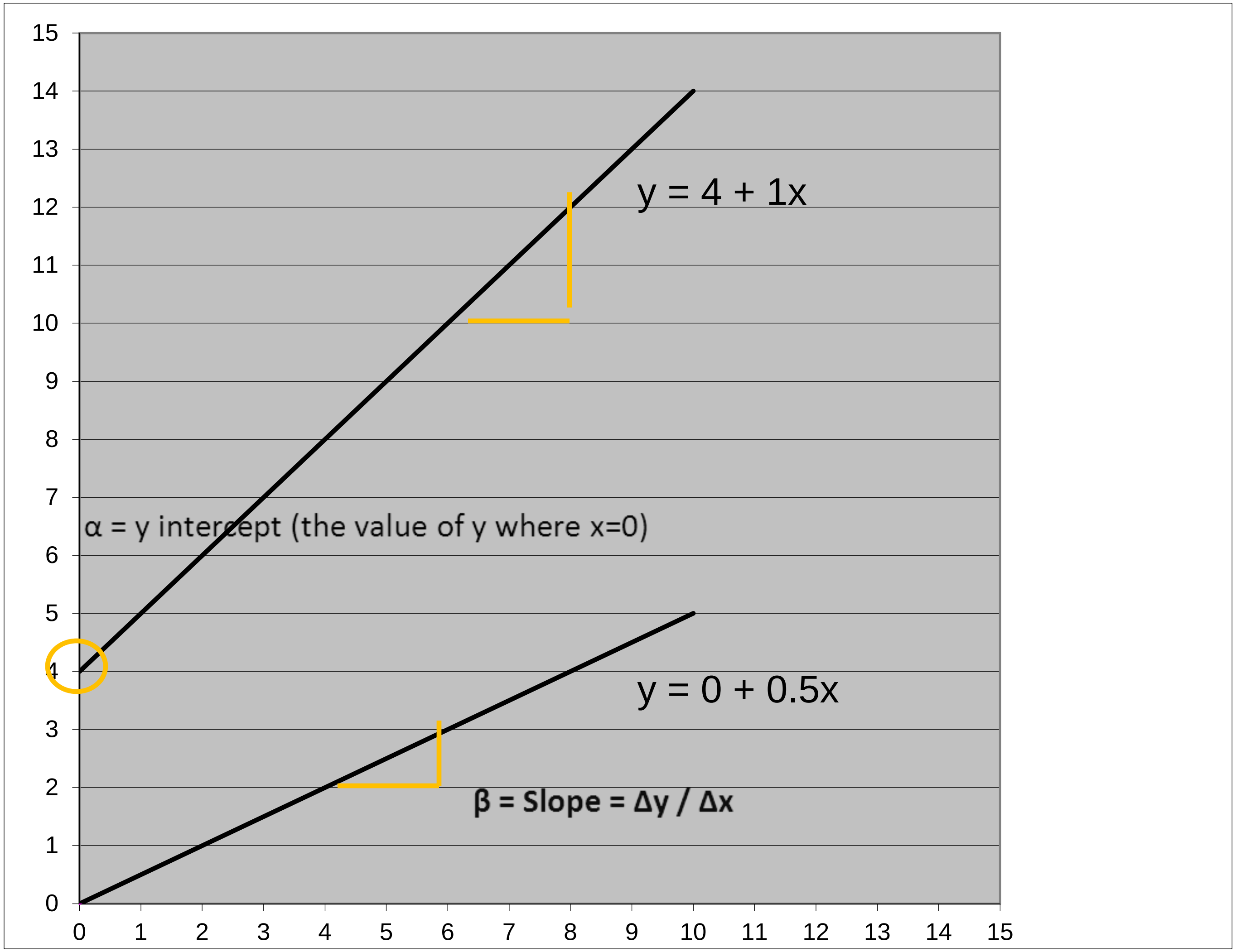
```
. tabstat fev, by(agecat) s(n min median max mean sd)

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)

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

Simple linear regression

- 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

$$\mu_{y|0} = \alpha + \beta X \quad = \alpha$$

Simple linear regression

- 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 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

Simple linear regression

- 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

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

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

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

Simple linear regression

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

Simple linear regression

- Assumptions of linear regression (continued)
  - Homoscedasticity – the standard deviation of y at each value of X is constant; $\sigma_{y/x}$ the same for all values of X
    - The opposite of homoscedasticity is heteroscedasticity
    - This is similar to the equal variance issue that we saw in ttests and ANOVA
  - 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>Note that we do not need the X's to be normally distributed, just the Y's at each value of X</u>

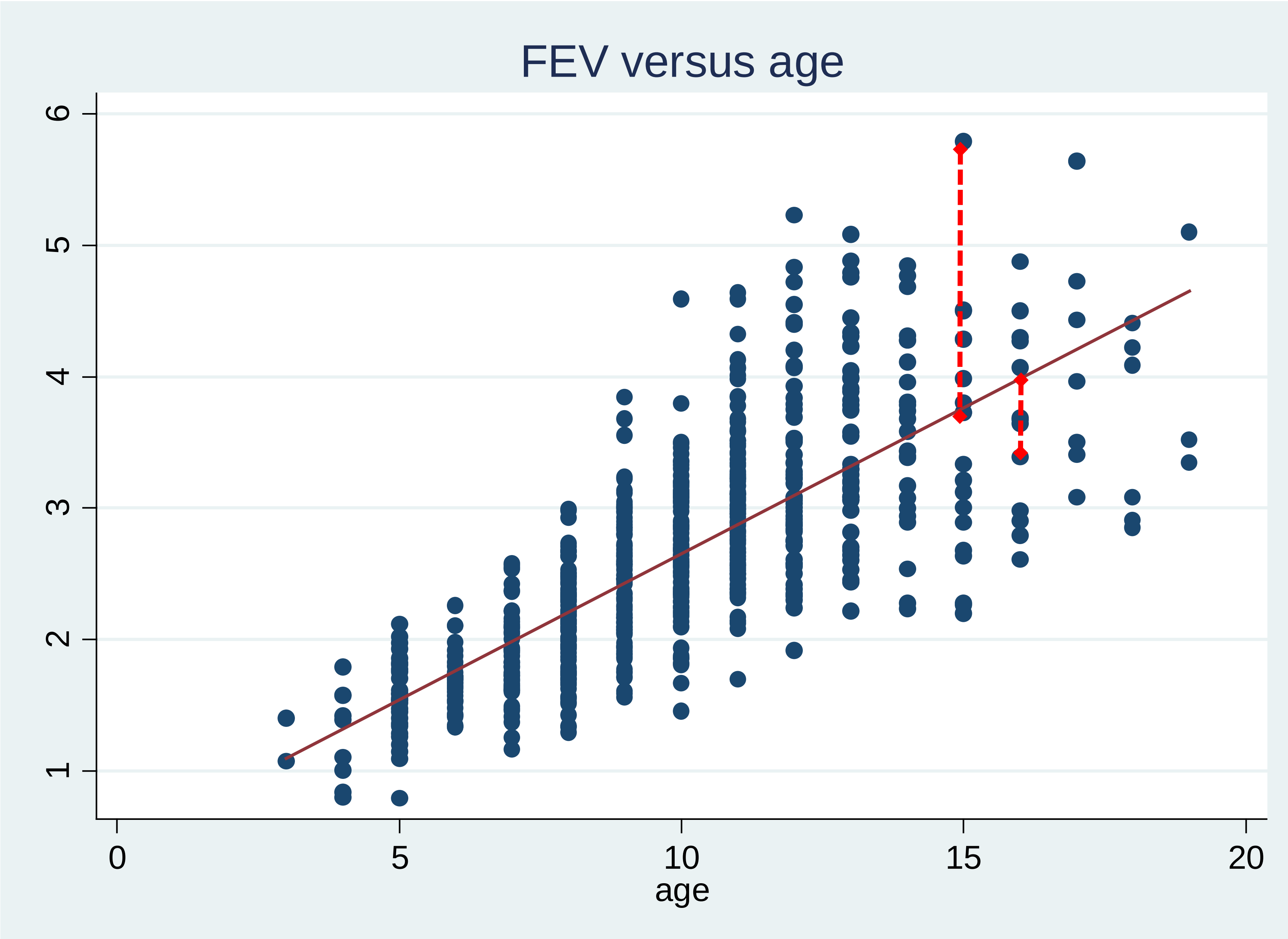
Least squares

- We estimate the coefficients of the population regression line (α and β) using our sample of measurements of Y and X
- We have a set of data, where the points are (y_i, x_i), and we want to put a line through them
- The distance from a data point (x_i, y_i) to the line at x_i is 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

It is the predicted value of y at each x


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

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, 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

- The solution to this minimization is

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{and} \quad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

  - These estimates are calculated directly from the x's and y's

Simple linear regression example: Regression of FEV on age

FEV= $\hat{\alpha} + \beta$ 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)

Interpretation of 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 $\hat{\beta}$

Inference for regression coefficients

- The standard error of the estimates are

$$se(\hat{\beta}) = \frac{\sigma_{y|x}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$$se(\hat{\alpha}) = \sigma_{y|x}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

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 $t = \dfrac{\hat{\beta} - \beta_0}{s\hat{e}(\hat{\beta})}$
- And it follows the t distribution with n-2 degrees of freedom under the null hypothesis

Inference for regression coefficients

- 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

(β - $t_{n-2,.025}$se(β) , β + $t_{n-2,.025}$se(β))

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

R^2

- A summary of the model fit is the coefficient of determination, R^2
- $R^2 = r^2$, i.e. the Pearson correlation coefficient squared
- 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

R^2

- $\sigma^2_{y|x} = (1-\rho^2)\sigma^2_y$

  If correlation=0, then $\sigma^2_{y|x} = \sigma^2_y$

  If correlation=1, then $\sigma^2_{y|x} = 0$

- Substituting in sample values and rearranging:

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

- Looking at this formula illustrates how R^2 represents the portion of the variability that is removed by performing the regression on X

Simple linear regression: evaluating the 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 $= TSS = MSS + RSS$

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

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

- Notation note:
  - Biostat 208 textbook Vittinghoff et al. use slightly different notation
  - The regression line notation we are using is

$$\hat{y}=\hat{\alpha} + \hat{\beta}x$$

Vittinghoff et al. uses

$$\hat{y}= \hat{\beta}_0 + \hat{\beta}_1 x$$

For next time

- Read Pagano and Gauvreau
  - Pagano and Gauvreau Chapter 17-18 (review)
  - Pagano and Gauvreau Chapter 18-19