Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Review of Correlation
(From Lecture 7, Biost 517, Fall 2012)

January 12, 2015

1

Correlation

Correlation Coefficient

- A measure of the tendency of the largest measurements for one variable to be associated with the largest measurements of the other variable
  - Dimensionless
  - The sample correlation r estimates the population correlation ρ (rho)

3

Pearson's Correlation Coefficient

- Definition of correlation between X and Y:

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

$$= \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}\sqrt{\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2}}$$

4

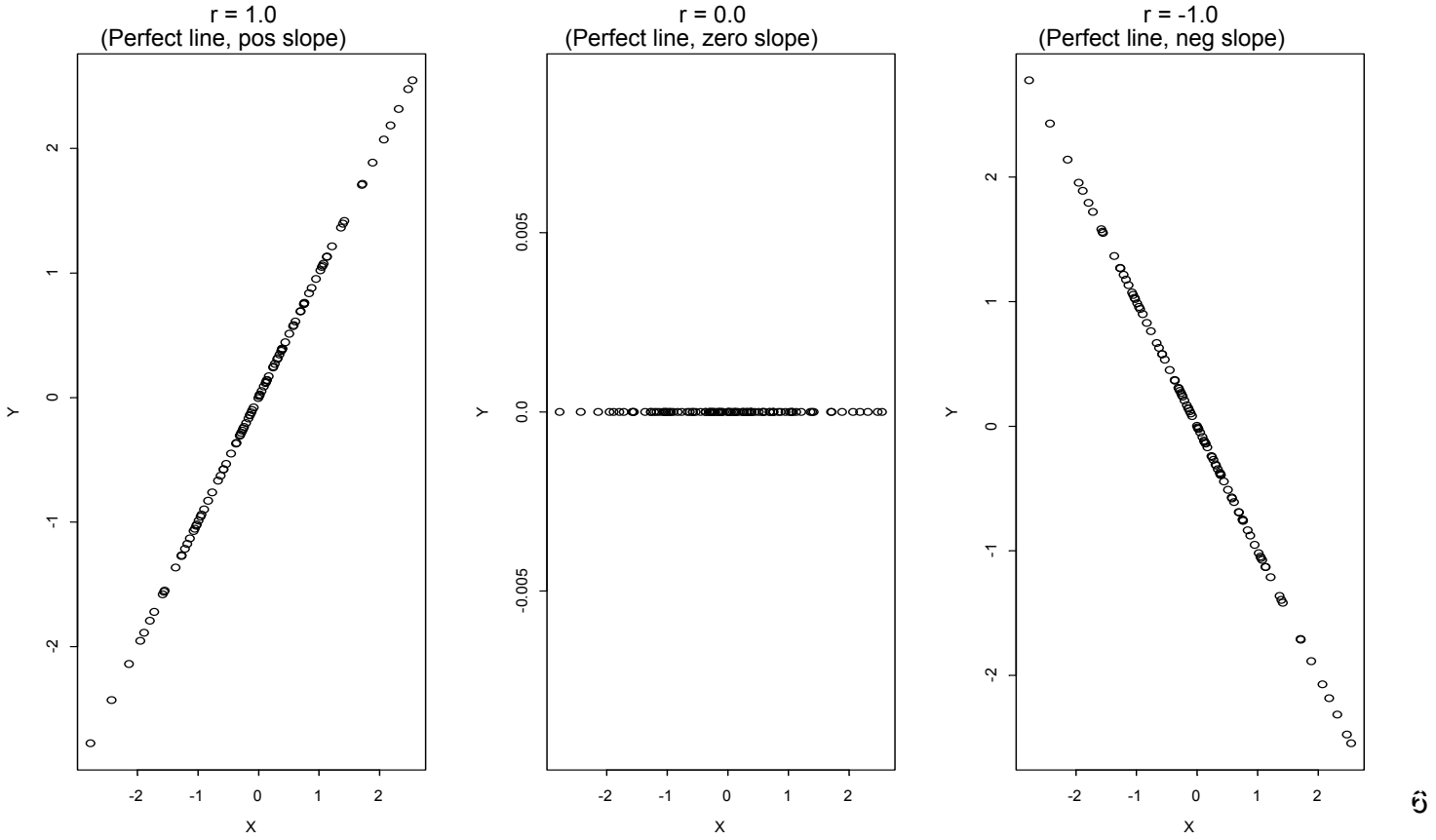
Possible Values of r

- Range of r : $-1 \leq r \leq 1$
  - r = 1 : perfect positive correlation
    - a graph of X vs Y will be a straight line with positive slope
  - r = -1 : perfect negative correlation
    - a graph of X vs Y will be a straight line with negative slope
  - r = 0 : no correlation

5

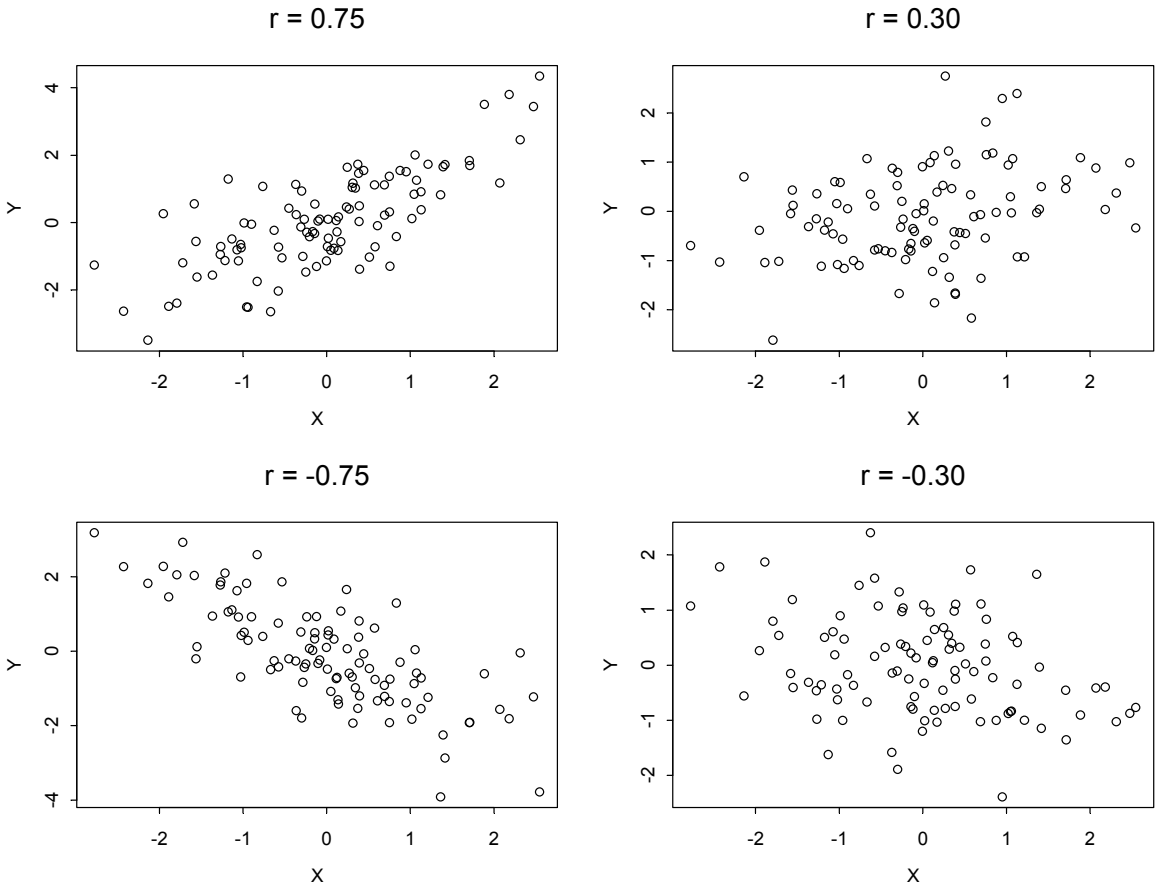
Straight Line Relationships

- Pearson's correlation coefficient with linear data



Linear Trends in Data

- Pearson's correlation coefficient with variable data



7

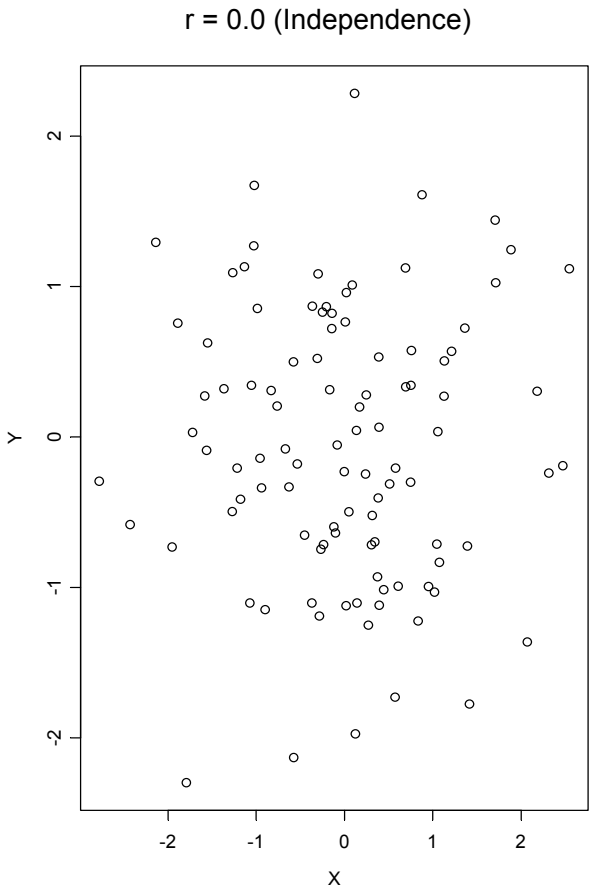
Correlation and Independence

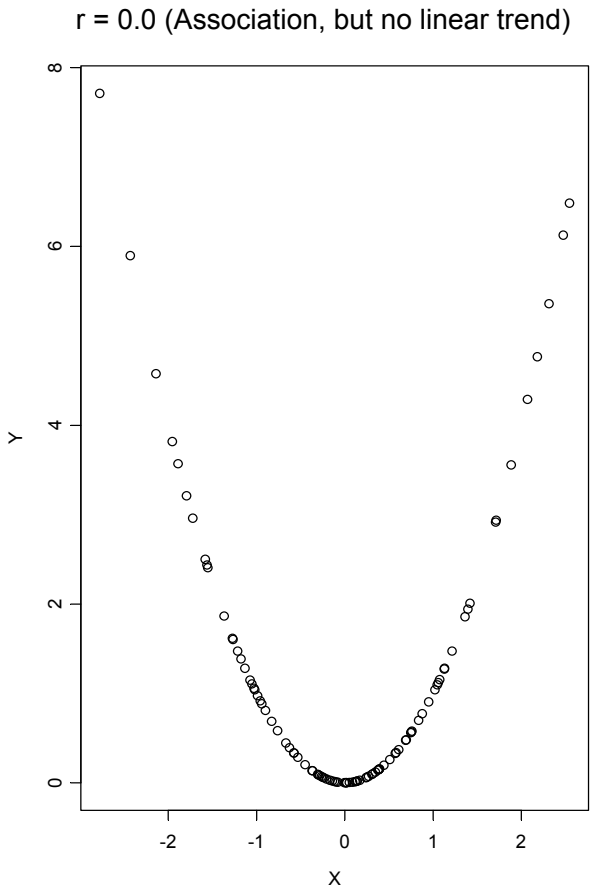
- Independent variables will have ρ = 0
  - (and r tending to be close to 0)

- However, uncorrelated variables are not necessarily independent
  - Correlation measures linear trend in the mean of one variable in groups defined by the other
  - It is possible that a nonlinear association exists between two variables, and that the first order trend is a zero slope

Uncorrelated Variables

- No linear trend between the variables





Software Commands

- Stata
  - "correlate *varlist*"
    - Correlation of all pairs of variables
    - Missing data deleted on a casewise basis
  - "pwcorr *varlist*"
    - Correlation of all pairs of variables
    - Missing data deleted on a pairwise basis
- R
  - `FEV <- read.table("fev.txt",header=TRUE)`
  - `correlate(FEV)`
    - Correlation of all pairs of variables
    - Missing data deleted on a pairwise basis by default

10

Ex: Correlation in FEV Data

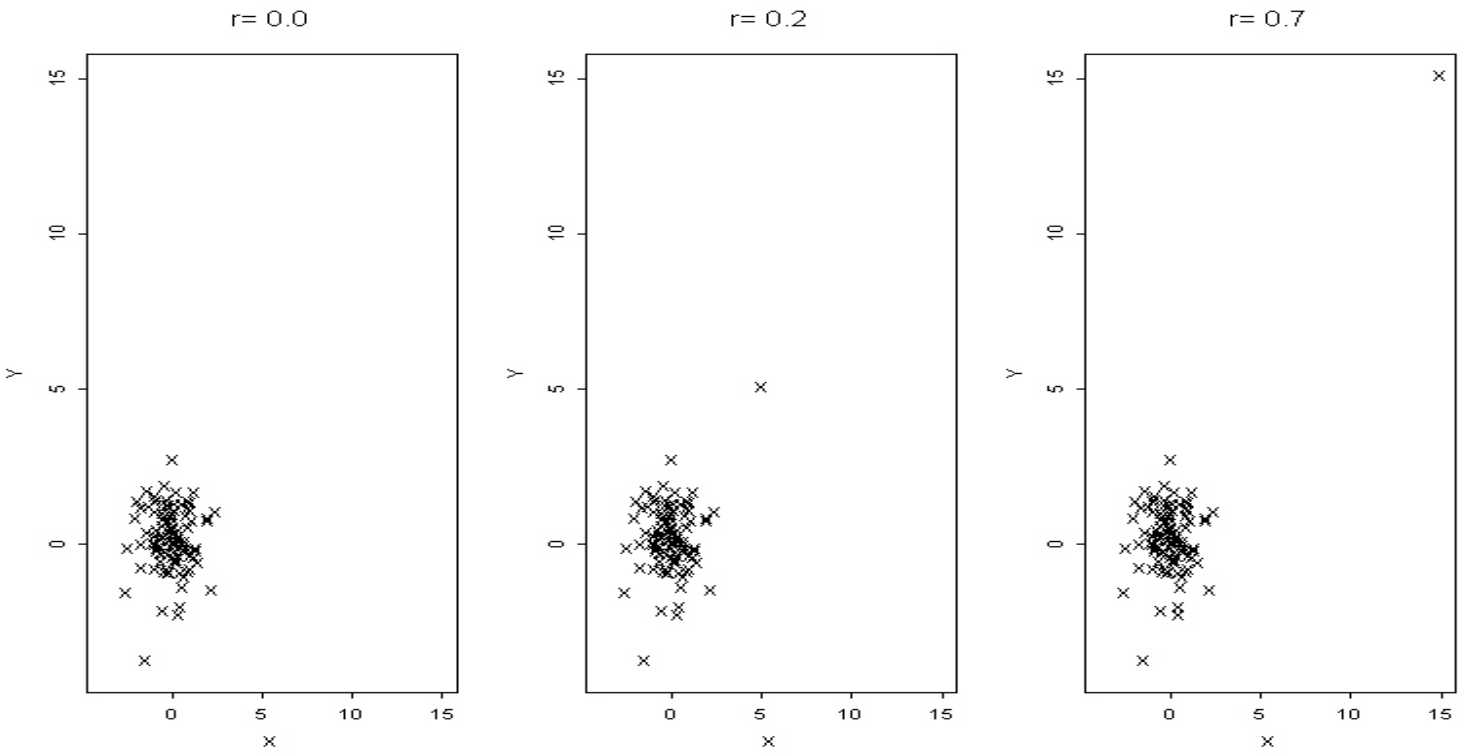
```
. corr subjid age fev height sex smoke
(obs=654)
       | subjid     age     fev  height     sex   smoke
-------+------------------------------------------------
subjid |  1.0000
   age | -0.0112  1.0000
   fev | -0.0147  0.7565  1.0000
height | -0.0317  0.7919  0.8681  1.0000
   sex |  0.0407 -0.0291 -0.2084 -0.1590  1.0000
 smoke | -0.0601 -0.4043 -0.2454 -0.2804 -0.0756  1.0000
```

- Some of these correlations don't make much sense
  - subjid is a nominal variable
  - sex, smoke are binary variables

11

Effect of Outliers on r

- Pearson's correlation coefficient can be greatly affected by outliers



Spearman's Rank Correlation

- To decrease the influence of outliers, Spearman's rank correlation coefficient computes the correlation of the ranks of the data

- In the previous example, the rank correlation is always the same: approximately 0.07

13

Software: Spearman's Correlation

- Stata: "**`spearman var1 var2`**"
  - Correlation of one pair of variables
  - Cases with missing data for either variable are deleted, and then ranks are computed
- R:
  - `FEV <- read.table("fev.txt",header=TRUE)`
  - `correlate(FEV,method="spearman")`
    - Correlation of all pairs of variables
    - Missing data deleted on a pairwise basis by default

14

Ex: Correlation in PSA Data

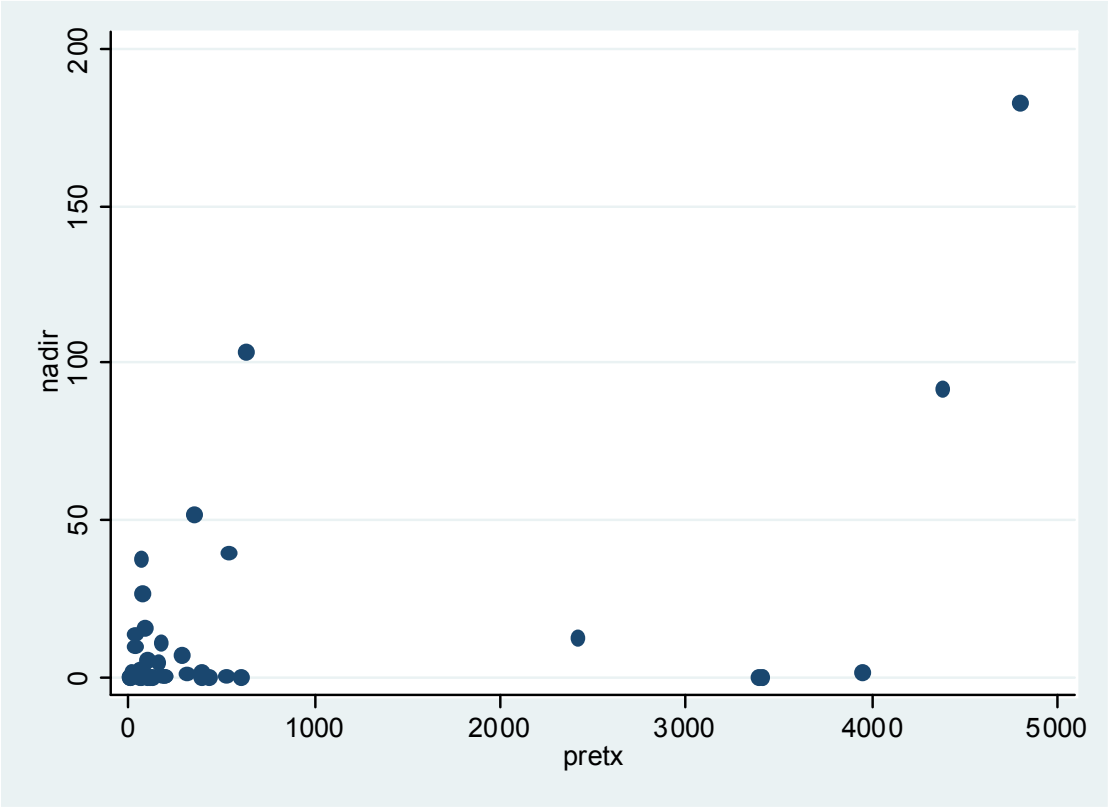
```
corr nadir pretx
(obs=43)
        |    nadir    pretx
--------+------------------
   nadir|   1.0000
   pretx|   0.5371   1.0000

spearman nadir pretx
Number of obs =      43
Spearman's rho =   0.1489
```

15

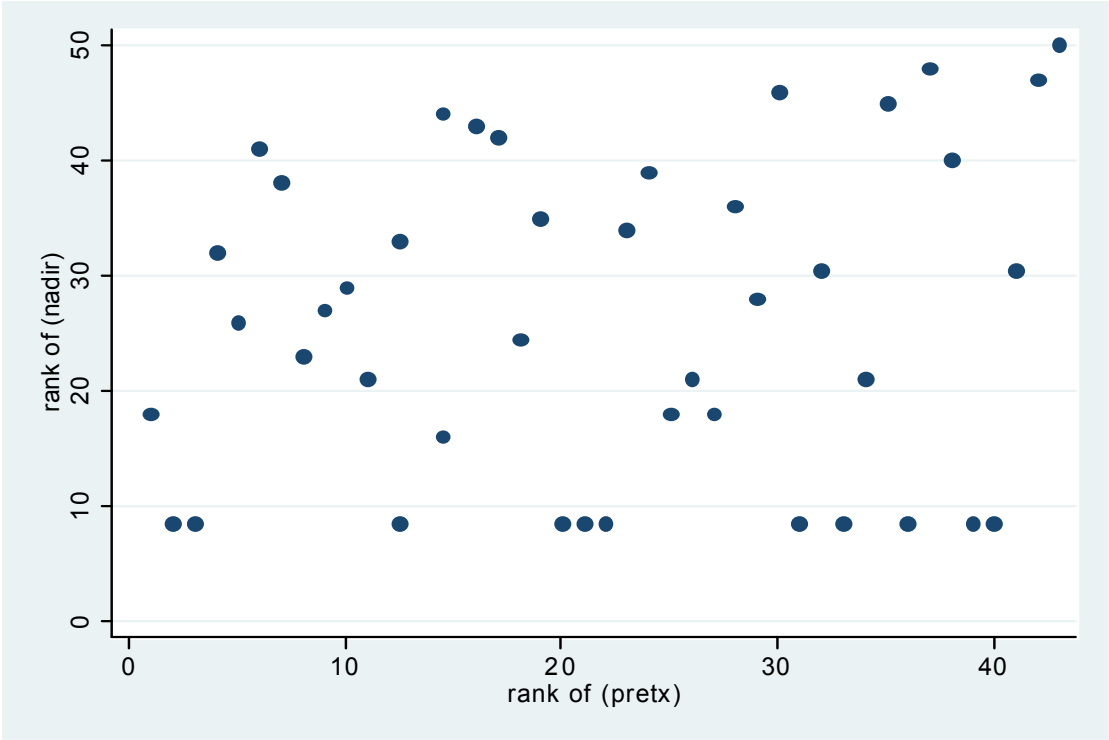
Ex: Nadir vs Pretreatment PSA

- Scatterplot of nadir versus pretx
  - `scatter nadir pretx`



Ex: Nadir vs Pretx Ranks

- `egen rnknadir = rank(nadir)`
- `egen rnkpretx = rank(pretx)`
- `scatter rnknadir rnkpretx`



17

Ex: Spearman's Corr vs r

- Possible explanation for lower rank correlation with Spearman's
  - Perhaps outliers in distribution of nadir and/or pretx unduly inflate r
  - Perhaps transforming to ranks masks true linear association in skewed variables

18

Uses of Correlation

- By type of variable
  - Correlation is a mean, thus only makes sense when a mean does
    - Limited interpretability with categorical data
    - Of no scientific relevance with censored data
- By scientific question
  - Greatest relevance when looking for associations between variables
    - But not particularly generalizable across studies

19

Correlation and Regression

More Interpretable Formula for r

$$r \approx \beta \sqrt{\frac{Var(X)}{\beta^2 Var(X) + Var(Y \mid X = x)}}$$

β = (LS) slope between Y and X

$Var(X)$ = variance of X in sample

$Var(Y \mid X = x)$ = variance of Y in groups that have same value of X

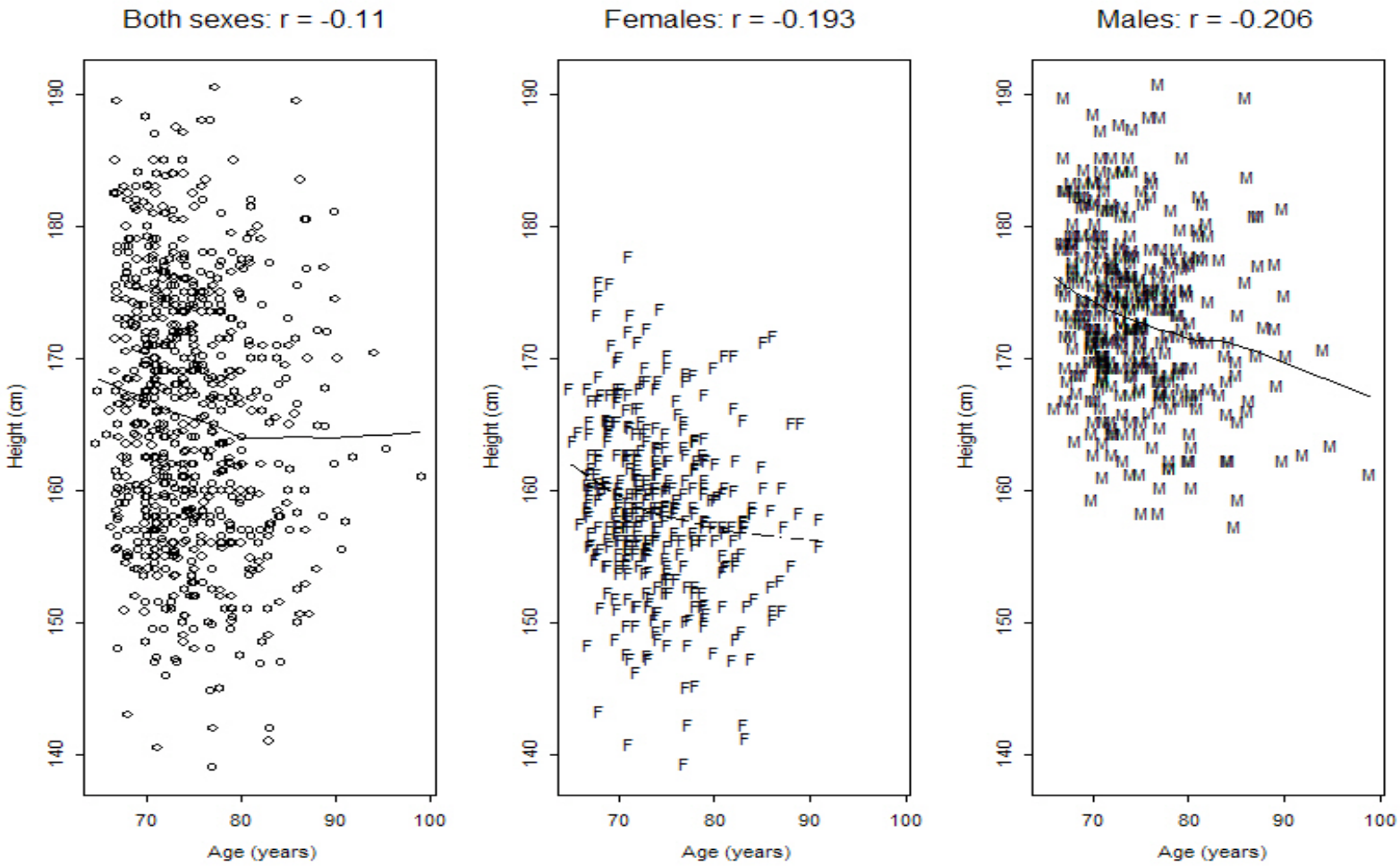
21

Properties of Correlation

- Correlation tends to increase in absolute value as
  - The absolute value of the slope of the line increases
    - More negative the slope, the more negative the correlation
    - More positive the slope, the more positive the correlation
  - The variance of data decreases within groups that share a common value of X
  - The variance of X increases
  - (Sample size is unimportant in tendencies toward lower or higher correlation)

22

Ex: Height vs Age (by Sex)

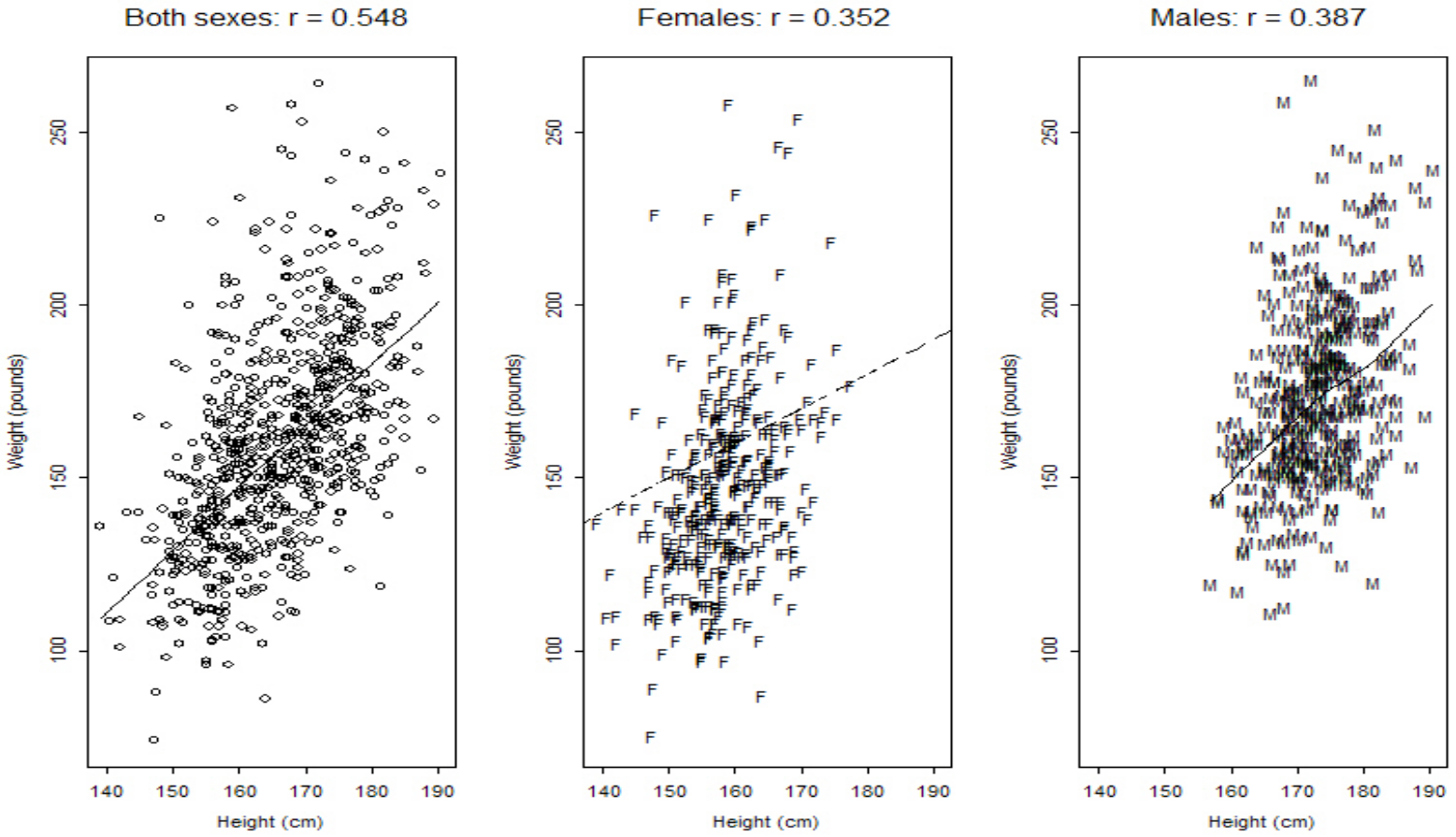


Ex: Height vs Age (by Sex)

- Correlation between Height and Age
  - Males: r = -0.206; Females: r = -0.193
  - Combined: r = -0.110

- Less extreme r in combined sexes
  - Approximately same slope in each sex and overall
  - Approximately same variance of age in each sex and overall
  - Combined group has higher within group variance of height by age (due to sex effect)

Ex: Weight vs Height (by Sex)



Ex: Weight vs Height (by Sex)

- Correlation between Height and Weight
  - Males: r = .387; Females: r = 0.352
  - Combined: r = 0.548

- More extreme r in combined sexes
  - Approximately same slope in each sex and overall
  - Approximately same within group variance (by height) for each sex and overall
  - Combined group has higher variance of height

26

Scientific Relevance of r

- It should be noted that
  - the slope between X and Y is of scientific interest
  - the variance of Y|X=x is partly of scientific interest, but it can be affected by restricting sampling to certain values of another variable
    - E.g., var (Height | Age) is less in males than when both sexes are included
  - the variance of X is often set by study design
    - This is often not of scientific interest

27