

Correlation

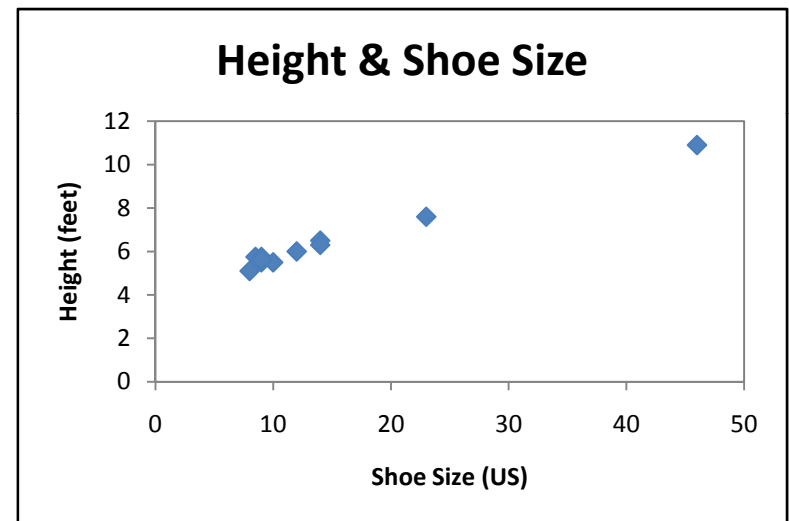
Relationships between variables

What question correlation asks?

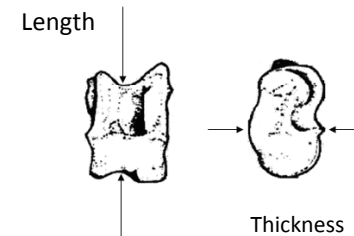
- How related are two variables to one another?
- Considers paired data on two variables for the same individuals (e.g., height and shoe size for multiple people)
- Determines the strength of a linear relationship between two variables.
- Does shoe size increase or decrease with height?
- Does test achievement increase or decrease with IQ?
- These are bivariate questions; they consider two variables (paired), not one (univariate).

Paired Data

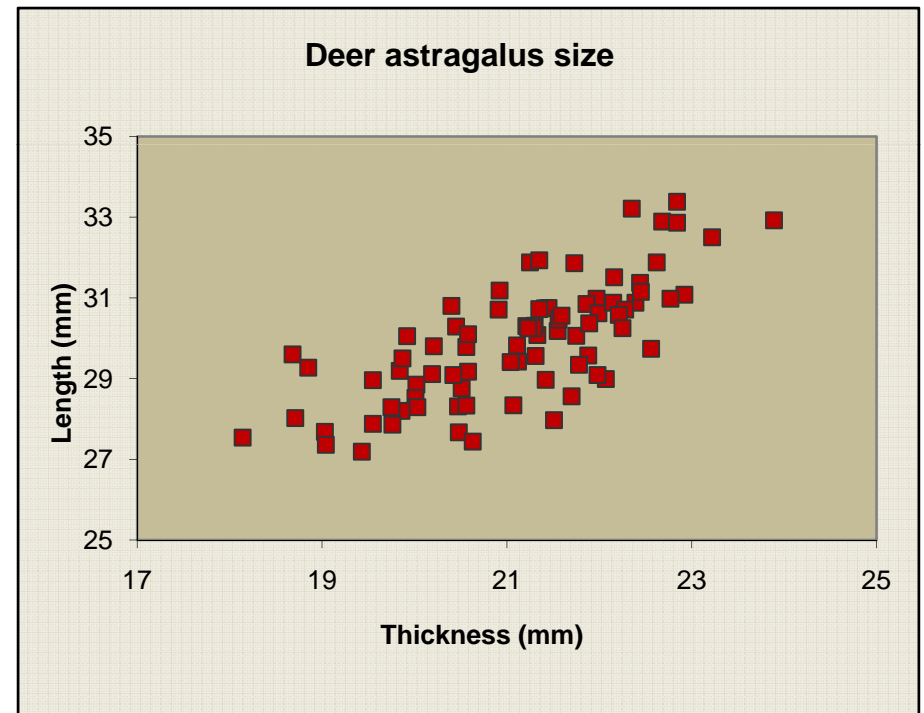
Person of esteem	Shoe size (US)	Height (feet)
Matthew	12	6
Mark	9	5.5
Luke	8.5	5.75
John	9	5.75
Buddha	10	5.5
Gandhi	8	5.1
Zeus	23	7.6
Dr. Oppong	14	6.5
Dr. Lyons	14	6.3
Sampson	9	5.6
Goliath	46	10.9



Scatterplots



- Scatterplots visually represent bivariate relationships in Cartesian space (X & Y axes)
- Here we would say that length positively correlates to thickness & vice versa
- But we do not know how strong the relationship is
- To find out we need to calculate a correlation coefficient

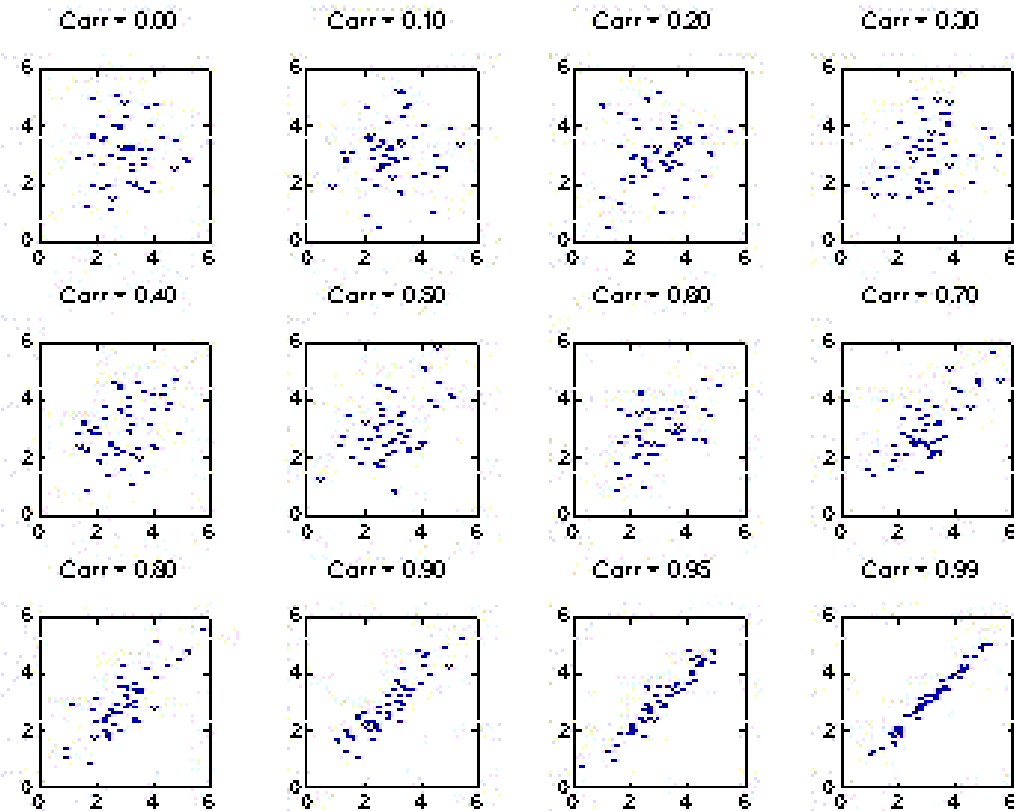


Correlation coefficient

- The correlation coefficient (r) varies from -1 to 0 to 1.
- $r = 1$ means that there is a perfect positive correlation
- If IQ & achievement were perfectly correlated an *increase* in one would produce the same magnitude of *increase* in the other.
- $r = -1$ is a perfect negative correlation
- If cholesterol level & lifespan were perfectly correlated, an *increase* in one would result in a *decrease* in the other at the same magnitude.
- $r = 0$ means there is no correlation and that two variables do not covary.

Scatterplots & Correlation

- It is easiest to see correlation in scatterplots



The more directional and the tighter the scatter, the more highly correlated two variables are

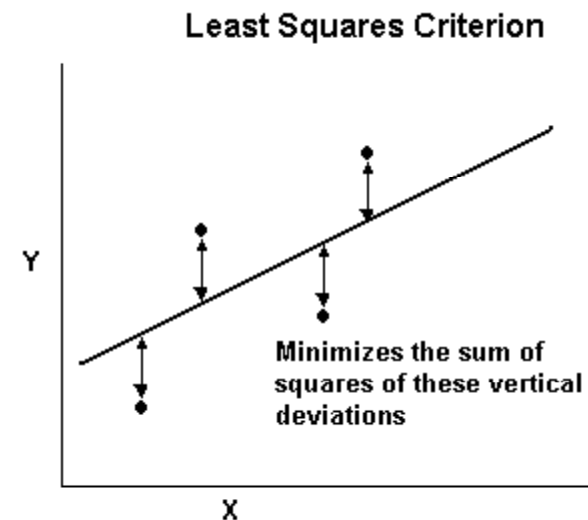
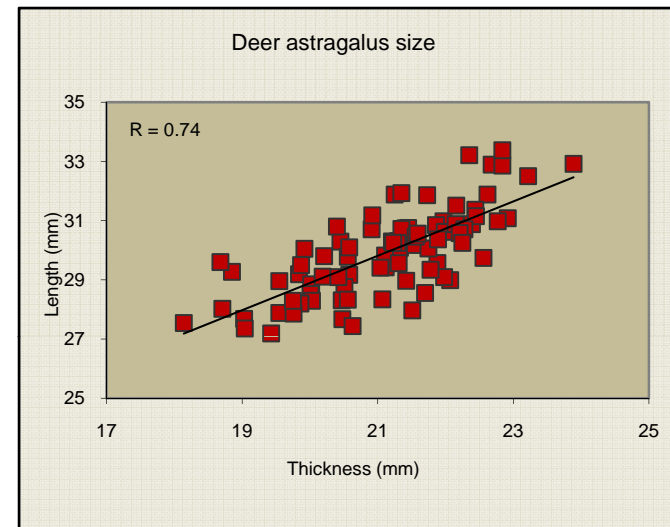
Calculating r

- r summarizes deviation of each point from the mean
- Below is the formula for **Pearson's r** , which is parametric

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y},$$

Least Squares Criterion

- “The straight line that best fits a set of data points is the one having the smallest possible sum of the squared errors”
- Simply a plot of a line through the scatter at the point that represents the least squared distance to a point on Y given X
- Thus, the line is a *best fit model*
- The tighter the scatter, the less error there is, the higher is r



Probability & r

- H_0 is $r = 0$
- H_a is $r \neq 0$ beyond that which can be explained by chance alone.
- We use a probability distribution (actually the t -distribution) to assess the significance of r
- Larger r values (negative or positive) are more likely to be significant
- Significant relationships easier to attain with larger samples.

What does r reflect?

- The slope of the scatter
- Magnitude of r = strength of linear relationship
- Sign of r reflects the type of relationship
- Significance of r is gauged on the t-distribution

$$t = \frac{r}{S_r} \quad \text{where } S_r \text{ is the standard error estimate of } r \quad S_r = \sqrt{\frac{1-r^2}{(n-2)}}$$

There is a t-critical for α ; if your test t is greater than α then r is significant

Nonparametric correlation

- Pearson's r assumes normality & must have interval/ratio scale data ($n \geq 30$, population normal)
- Spearman's ρ is correlation of ranks ($n < 30$, non-normal)
 - Data for each variable are ranked; then the ranks are correlated.
 - ρ and r_s are the same
 - A p-value is associated to determine significance

Problems with correlation

- Test power: large samples will provide significant tests, even with low r
 - This because, it is easier to “find covariance” with large samples
- Generally speaking
 - Significance matters most when $n < 70$
 - In all cases
 - $r \leq 0.30$ is weak correlation
 - $r > 0.30, \leq 0.70$ is moderate correlation
 - $r \geq 0.70$ is strong correlation

Coefficient of Determination

- r^2 = the coefficient of determination
- r^2 determines the variability in the dependent variable (Y) predicted by the independent variable (X) as a percent.
- $r = 0.977$, $r^2 = 0.95$ for shoe size predicting height; means that 95% of the variability in height can be accounted for by differences in shoe size
 - That is, shoe size is a great predictor of height!
 - State this as 5% of the variability in Y cannot be explained by difference in X
- A high r does not mean that X causes Y, just that they covary tightly
 - One is a good measure of the other

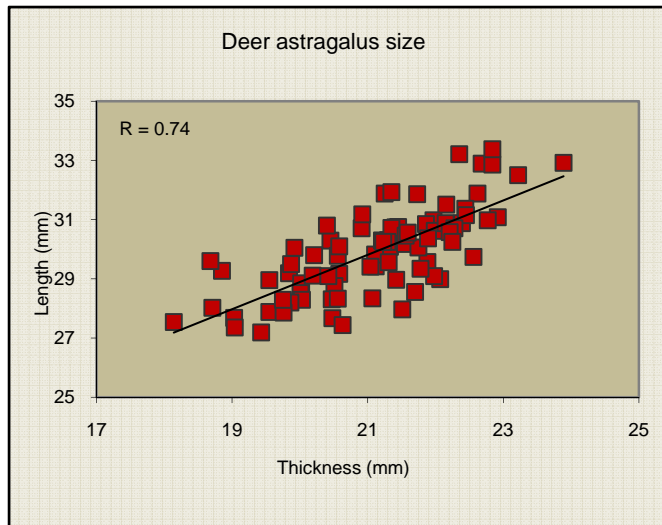
Causation: 3 Philosophical Rules

- 1) X has to occur temporally before Y
- 2) X and Y must be correlated
- 3) All other possible causes must be ruled out
 - Internal validity and science

Example Problem

Question: do astragalus length and thickness covary significantly as measures of size?

Implications: if they covary, I might be able to use both to make size comparisons between samples, more variables = less error



*Untitled1 [DataSet0] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

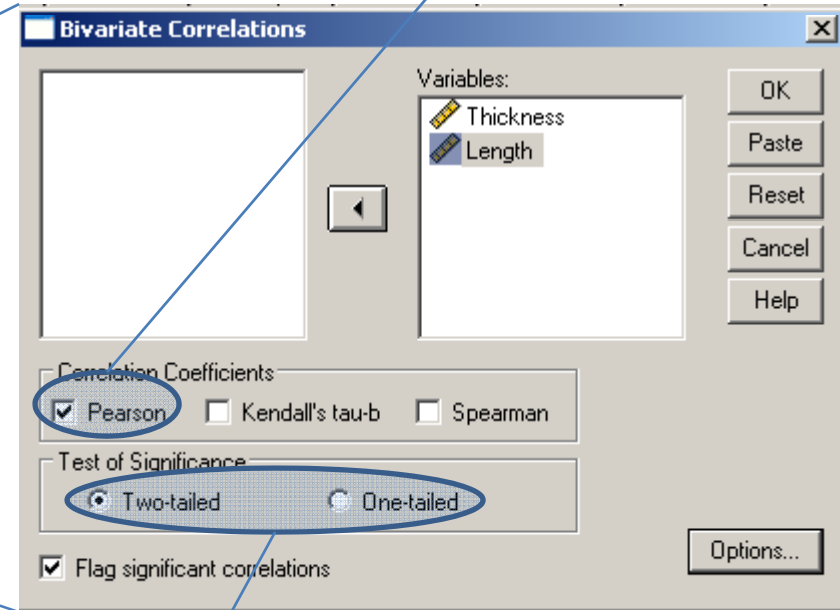
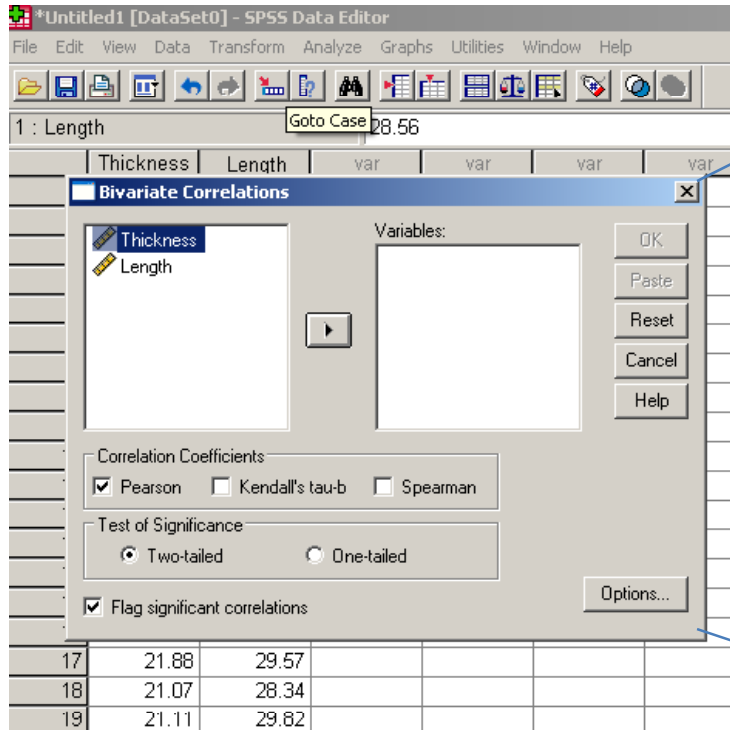
1 : Length

	Thickness	Length
1	21.70	28.5
2	22.62	31.8
3	21.97	30.9
4	20.45	30.2
5	21.86	30.8
6	21.12	29.4
7	22.16	31.5
8	22.44	31.3
9	20.56	29.7
10	19.84	29.1
11	21.33	30.0
12	22.92	31.0
13	22.07	28.9
14	20.02	28.85
15	20.19	29.11

Correlate > Bivariate...

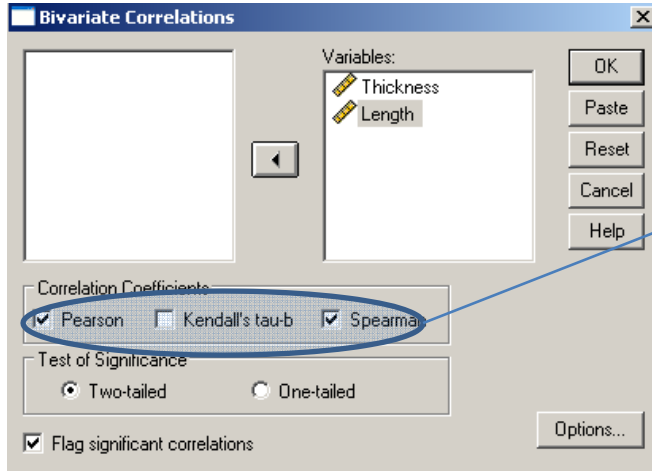
SPSS example

Pearson's = parametric



Irrelevant, the sign of r gives direction, always use two-tailed

SPSS Output



I chose Pearson's and Spearman's to demonstrate output for both.

Pearson's (parametric)

Correlations

		Thickness	Length
Thickness	Pearson Correlation	1	.740**
	Sig. (2-tailed)		.000
	N	82	82
Length	Pearson Correlation	.740**	1
	Sig. (2-tailed)	.000	
	N	82	82

** . Correlation is significant at the 0.01 level (2-tailed).

Spearman's (non-parametric)

Correlations

		Thickness	Length
Spearman's rho	Thickness	1.000	.730**
	Correlation Coefficient		.000
	Sig. (2-tailed)		.000
Length	Length	.730**	1.000
	Correlation Coefficient		.000
	Sig. (2-tailed)		.000
N		82	82

** . Correlation is significant at the 0.01 level (2-tailed).

Simple linear regression

- Uses basically the same mathematics as correlation
- Asks an additional question: how well can the value of one variable be used to *predict* the value of another, different variable?
- For example, how well can height be *predicted* from shoe size?
- How well can astragalus length be predicted from thickness?
- *Uses the least squares line as a predictive model*
 - $Y = a + bX$ ($a = y$ intercept; $b =$ slope)
 - a = the expected value of Y when $X = 0$
 - b = the change in Y with an increase of 1 in X

Important concepts

- **Independent variable**: the variable creating the influence or effect, the predictor (shoe size)
 - Always on the X axis
- **Dependent variable**: the variable receiving influence of effect, the predicted (height)
 - Always on the Y axis

What do you need to know?

- What paired data are
- How scatterplots work and how they relate to correlation
- What r stands for and how it can vary (-1 to +1)
- What the H_0 is for correlation
- What the least squares line represents
- How Spearman's differs from Pearson's and when to use each
- The power problem and strength/weakness criteria
- Questions correlation & regression answer
- What the coefficient of determination (r^2) is
- How to use SPSS to analyze correlation
- The 3 rules of causation

HW 11

HW11.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : Thick1 20.45

	Thick1	Length1	Length2	Width	Length3	Thick2	Weight	Age
1	20.45	34.54	30.29	22.83	37.43	20.69	72.00	1.5
2	21.12	34.83	29.42	23.46	38.77	21.60	83.00	1.5
3	22.16	36.15	31.51	24.29	38.80	21.40	83.00	1.5
4	20.19							4.5
5	20.91							3.5
6	22.39							2.5
7	20.58							3.5
8	23.89							8.5
9	22.45							2.5
10	21.73							2.5
11	20.48							2.5
12	21.98							1.5
13	21.45							3.5
14	20.51							1.5
15	20.47							4.5
16	20.21							2.5
17	21.78							3.5
18	20.92							1.5
19	22.84							1.5
20	21.35	35.30	31.93	23.73	36.96	21.14	117.00	3.5

Bivariate Correlations

Variables:

- Width
- Length3
- Thick2
- Age
- Thick1
- Length1
- Length2
- Weight

Correlation Coefficients

Pearson Kendall's tau-b Spearman

Test of Significance

Two-tailed One-tailed

Flag significant correlations

Options...

I have entered four variables because I want to find out how they each correlate to one another

Output

Correlations

		Thick1	Length1	Length2	Weight
Thick1	Pearson Correlation	1	.796**	.785**	.595**
	Sig. (2-tailed)		.000	.000	.000
	N	136	136	136	136
Length1	Pearson Correlation	.796**	1	.924**	.547**
	Sig. (2-tailed)	.000		.000	.000
	N	136	136	136	136
Length2	Pearson Correlation	.785**	.924**	1	.597**
	Sig. (2-tailed)	.000	.000		.000
	N	136	136	136	136
Weight	Pearson Correlation	.595**	.547**	.597**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	136	136	136	136

** . Correlation is significant at the 0.01 level (2-tailed).

Multiple Regression

- Incorporates more than one independent variable to predict the dependent variable

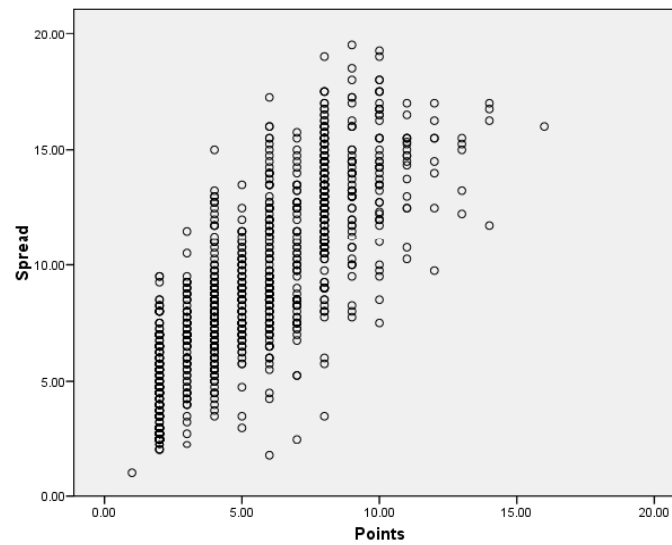
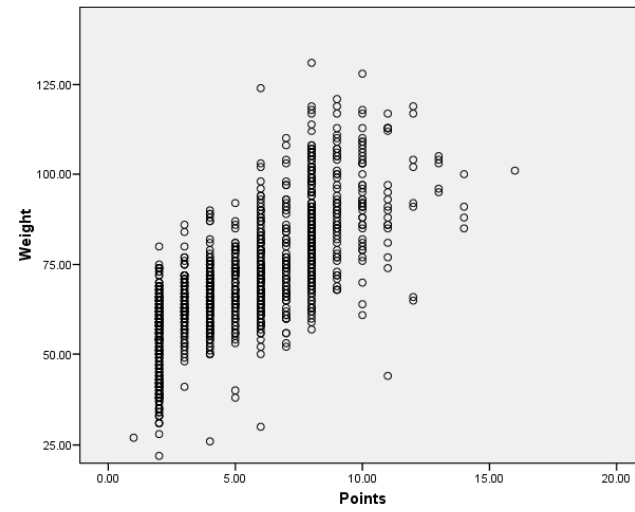
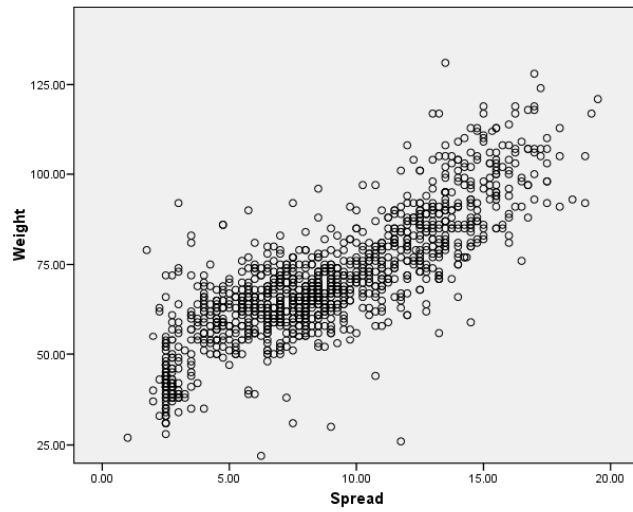
$$Y = a + b_1X_1 + b_2X_2 + b_3X_3$$

- Ability to predict increases because more variability can be accounted for

Multicollinearity

- Relates to multiple regression
- Occurs when independent variables measure the same thing
 - Adding new multi-collinear variables does not add much predictive power

Example: predicting weight from antler size



Example: predicting weight from antler size

Spread only

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.827 ^a	.683	.683	9.83043

a. Predictors: (Constant), Spread

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	238039.0	1	238039.032	2463.222	.000 ^a
	Residual	110359.8	1142	96.637		
	Total	348398.8	1143			

a. Predictors: (Constant), Spread

b. Dependent Variable: Weight

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	38.766	.706		54.891	.000
	Spread	3.595	.072	.827	49.631	.000

a. Dependent Variable: Weight

Spread & points

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.834 ^a	.695	.694	9.63392

a. Predictors: (Constant), Spread, Points

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	241014.3	2	120507.131	1298.394	.000 ^a
	Residual	105806.2	1140	92.812		
	Total	346820.4	1142			

a. Predictors: (Constant), Spread, Points

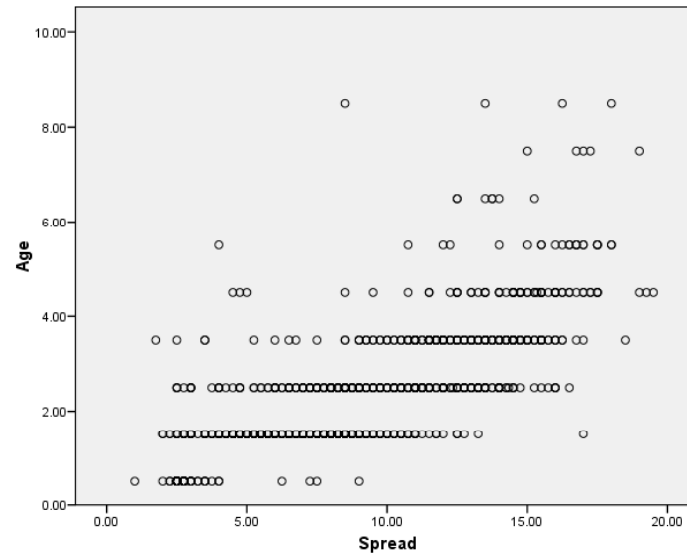
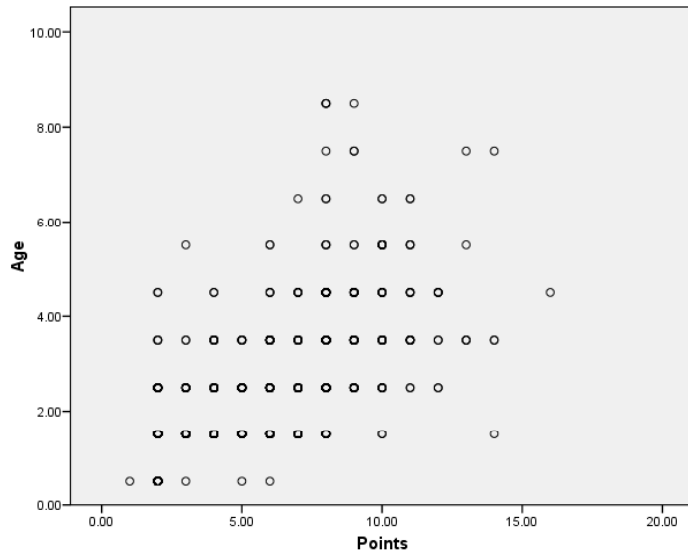
b. Dependent Variable: Weight

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	38.499	.694		55.447	.000
	Points	1.069	.178	.170	6.005	.000
	Spread	2.989	.123	.689	24.291	.000

a. Dependent Variable: Weight

Example: predicting age from antler size



Example: predicting age from antler size

Spread only

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.752 ^a	.566	.566	.83469

a. Predictors: (Constant), Spread

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1037.808	1	1037.808	1489.582	.000 ^a
	Residual	795.644	1142	.697		
	Total	1833.451	1143			

a. Predictors: (Constant), Spread

b. Dependent Variable: Age

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.207	.060		3.454	.001
	Spread	.237	.006	.752	38.595	.000

a. Dependent Variable: Age

Spread & points

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.764 ^a	.584	.583	.81721

a. Predictors: (Constant), Points, Spread

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1068.827	2	534.414	800.229	.000 ^a
	Residual	761.322	1140	.668		
	Total	1830.149	1142			

a. Predictors: (Constant), Points, Spread

b. Dependent Variable: Age

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.178	.059		3.022	.003
	Spread	.178	.010	.565	17.069	.000
	Points	.105	.015	.230	6.934	.000

a. Dependent Variable: Age

In class exercise

- For the simple linear regression, predicting weight from inner spread
 - Write up the results
 - Explain the correlation in the scatterplot
- The do the same with the multiple regression predicting age