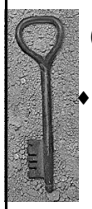
Scatterplots & Linear Association

Review

- Association between two categorical variables
  - Cross-tabulation
  - Chi-square test of independence
  - Measures of association

Comparing Two Continuous Variables

- Three questions:
- 1) Is there an association?
  - Hypothesis test of independence
- 2) How strong is the association?
  - Measure the association
- 3) What is the form of the relationship?
  - A formula to predict values



Comparing Continuous Variables

- Tools to examine the relationship between two continuous variables
  - Visual representation: scatterplot
  - Measure of linear association: correlation
  - Predictive model: regression
  - Can be elaborated to address more than one independent variable

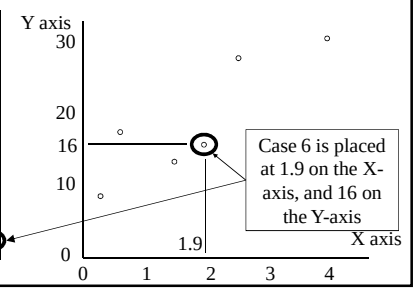
Scatterplots

- Visually represent the relationship between two continuous variables
- Like histograms, they have an X and Y axis
  - But, histograms indicate the frequency of each X value along the Y axis
- Scatterplots use X and Y axes to locate a case along two different variables
- Usually use X axis for the independent variable, Y axis for the dependent variable (based on your theory and hypothesis)

Scatterplots

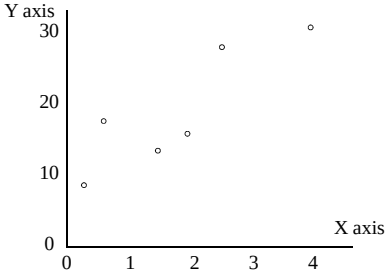
- Example: Study time and student achievement.
  - X variable: Average # hours spent studying per day
  - Y variable: Score on reading test

Case	X	Y
1	2.6	28
2	1.4	13
3	.65	17
4	4.1	31
5	.25	8
6	1.9	16



Scatterplots

- Question: Do you think that study time and test score are associated?

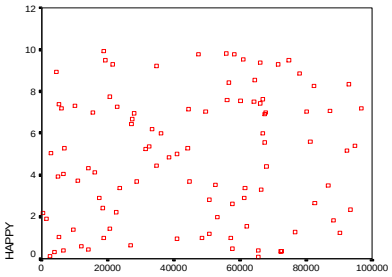


Interpreting Scatterplots

- Scatterplots tell you a great deal about the relationship between two variables
  - But, like histograms, it takes practice to read them
- Advice/Tips:
  - Set up the axes so cases aren't too tightly packed
  - Focus on the general trend, not extreme cases

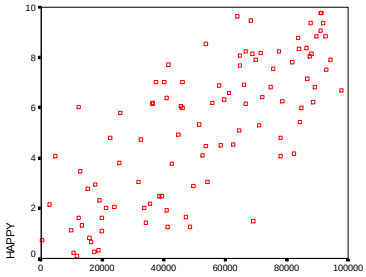
Interpreting Scatterplots

- E.g. Low association between variables
  - Annual Income vs. Happiness Index (1-10 scale)



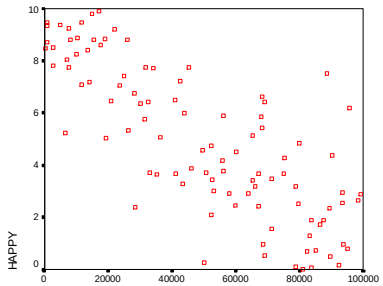
Interpreting Scatterplots

- E.g. A more materialistic sample
  - Positive association between income & happiness



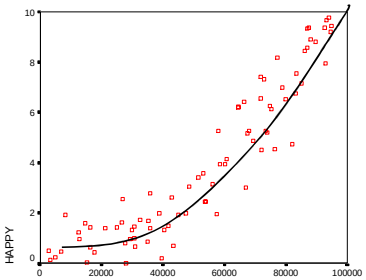
Interpreting Scatterplots

- E.g.: Can you interpret this sample?



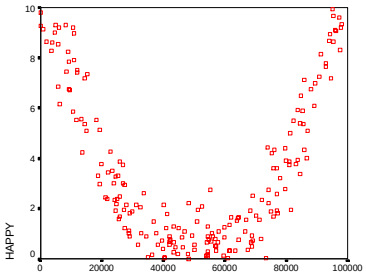
Interpreting Scatterplots

- The relationships can be curved (non-linear)



Interpreting Scatterplots

- What would this society be like?

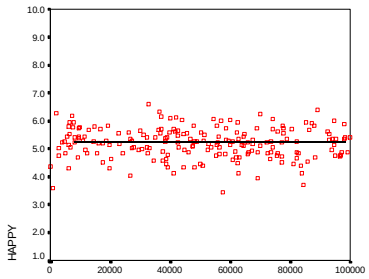


Interpreting Scatterplots

- No relationship is represented by a "cloud" of evenly distributed points
- Strong linear relationships are reflected by visible "diagonal lines" on the graph
- Non-linear (curved) relationships are reflected by various curved patterns
  - U-shaped, upside down U-shaped
  - S-shaped, J-shaped

Interpreting Scatterplots

- What if you have a clear horizontal line?
  - Is there a strong association?



NO! Happy does not change consistently as income increases. It stays the same

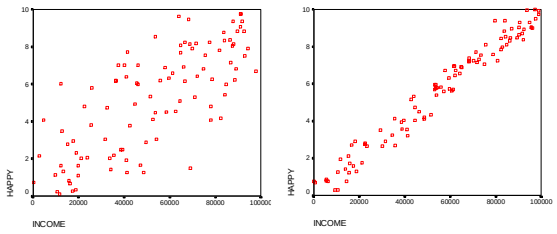
The data points are linear, but the line is "constant" (flat), indicating no association

Linear Association

- Linear Association = covariation among two variables along a straight line
  - Values change in proportion to each other.
- Linear association can be positive or negative
  - Positive: Increase in X = proportional increase in Y
  - Negative: Increase in X = proportional decrease in Y
  - Positive: line slopes up, Negative: Line slopes down
- Absence of "linear" association does not mean that variables are unrelated
  - the U-shaped example

Linear Association Examples

- The closer points fall to a single line, the higher the linear association

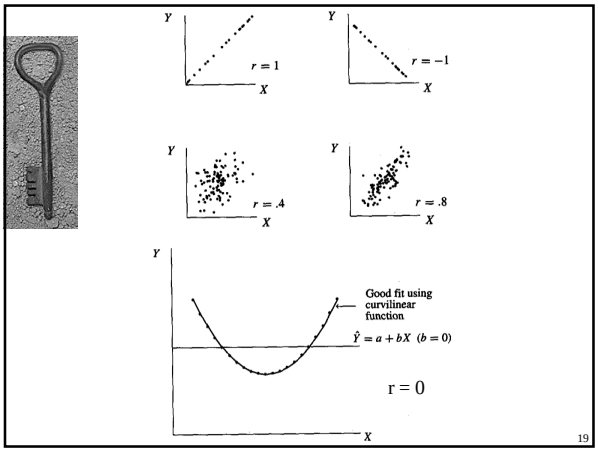


Measuring Linear Association

- Linear association is measured by the correlation coefficient (r) ("Pearson's Correlation Coefficient")

$$r = \frac{S_{xy}}{S_x S_y}$$

- Measure both strength and direction
- -1<= r <=1
- r = 0 indicates no linear association
- r = 1 indicates perfect positive linear association
  - All point fall exactly on a single line
  - Line goes from lower-left to upper-right of scatterplot
- r = -1 indicates perfect negative linear association
  - Line goes from upper-left to lower-right
- r is determined by how close the points fall to an exact line, not the slope (steepness) of the line



Linear Functions

- If two variables are perfectly linearly associated (r = 1), we can powerfully summarize the data
  - Simply write down the formula for the line that all points fall onto
  - Formula: Y = a + bX
- The formula for a line is a mathematical function
  - It specifies a relationship between two variables (X,Y)
  - Such that for any value of X, Y can be calculated, and vice versa

Linear Functions: Y = a + bX

- As a pure mathematical function, X and Y are not conceived of as a set of specific points
  - We don't have a dataset: Y_i, where i goes from 1 to N
  - Instead, a function relates every possible X to a corresponding Y
  - Later, we will link this formula to real data: X_i, Y_i

Linear Functions: Y = a + bX

- It is a family of functions, like the normal distributions
  - For any chosen value of μ, σ, you get a normal curve
  - For any value of a and b, you get a particular line
- Jargon:
  - a is referred to as the "constant" or "intercept"
  - b is referred to as the "slope"

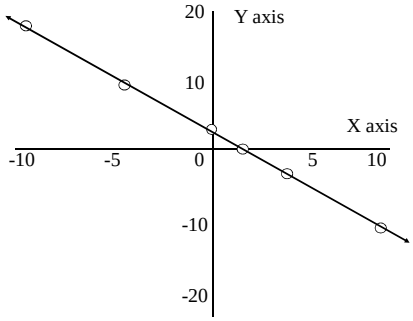
Linear Functions: Y = a + bX

- To graph a linear function:
- 1. You must have known values for a, b
- 2. Pick several values for X (at least 2)
  - Zero is the easiest point to start with, because Y = a
- 3. Calculate corresponding values of Y for each value of X
- 4. Draw X and Y axes, with relevant units
- 5. Mark down your X,Y points on the graph
- 6. Draw a line connecting them (and continuing on in either direction)

Linear Functions: Y = a + bX

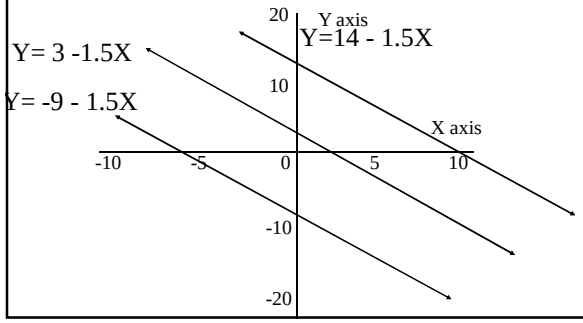
- Graphing example: Y = 3 - 1.5X

X	Y
0	3
2	0
3	-1.5
10	-12
- 4	9
-10	18



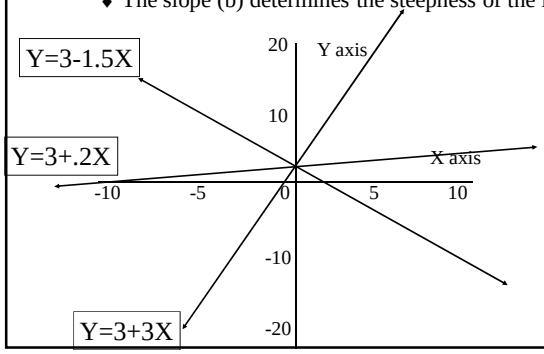
Linear Function: Y = a + bX

- **a: determines where the line intersects the Y-axis**



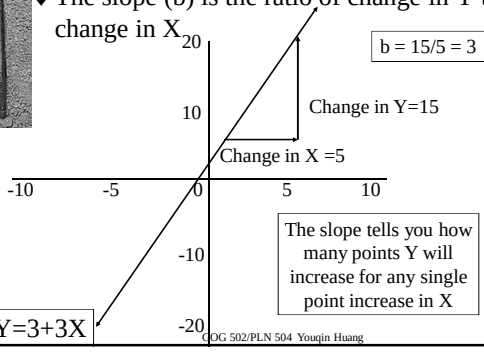
Linear Functions: Y = a + bX

- The slope (b) determines the steepness of the line



Linear Functions: Slopes

- The slope (b) is the ratio of change in Y to change in X



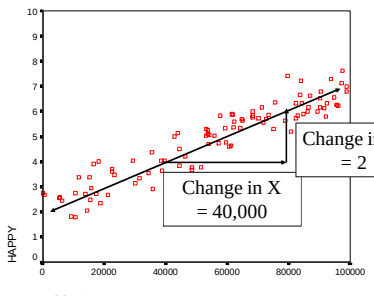
The slope tells you how many points Y will increase for any single point increase in X

Linear Functions: Slopes

- If slope b = 1, any change in X is accompanied by an equal change in Y
- b > 1 indicates that Y increases more than 1 point for each increase in X
- b < 1 indicates that Y increases by less than 1 point for each increase in X

Linear Functions as Summaries

- A linear function can be used to summarize the relationship between two variables:



Slope:

b = 2 / 40,000 = .00005 pts/$

Change in Y = 2

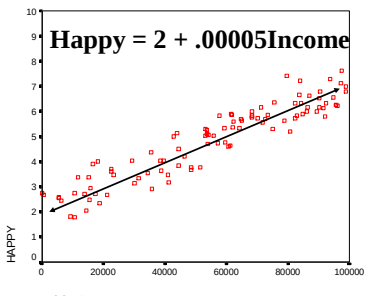
Change in X = 40,000

If you change units:

b = 2 / 40K
b = .5 pts/$10K
b = 5 pts/$100K

Linear Functions as Summaries

- Slope and constant can be "eyeballed" to approximate a formula:



Slope (b):

b = 2 / 40,000 = .00005 pts/$

Constant (a) = Value where line hits Y axis

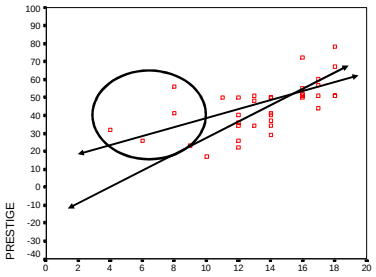
a = 2

Linear Functions as Summaries

- Linear functions can powerfully summarize data:
  - Formula: Happy = 2 + .00005Income
- Gives a sense of how the two variables are related
  - Namely, people get a .00005 increase in happiness for every extra dollar of income (or 5 pts per $100K)
- Also lets you "predict" values. What if someone earns $150,000?
  - Happy = 2 + .00005($150,000) = 7.5
- But be careful… You shouldn't assume that a relationship remains linear indefinitely
  - Also, negative income or happiness make no sense…

Linear Functions as Summaries

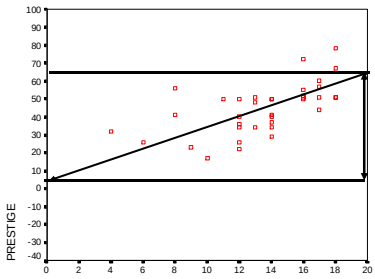
- Come up with a linear function that summarizes this real data: years of education vs. job prestige



It isn't always easy! The line you choose depends on how much you "weight" these points.

Linear Functions as Summaries

- One estimate of the linear function



The line meets the Y-axis at Y=5. Thus a = 5

The line increases to about 60 as X reaches 20. The increase is 60 in Y per 20 in X. Thus: b = 60/20 = 3

Formula:

Y = 5 + 3X

Linear Functions as Summaries

- Q: How much additional job prestige do you get by going to college (an extra 4 years of education)?
  - Formula: Prestige = 5 + 3*Education
- Answer: About 12 points of job prestige
  - Change in X is 4… Slope is 3. 3 x 4 = 12 points
- What is the interpretation of the constant?
  - It is the predicted job prestige of someone with zero years of education… (Prestige = 5)

Linear Functions as Summaries

- What do you think happens to the relationship between education and job prestige when education exceeds 20?
  - Would it remain linear?
  - Or would the effect taper off?
- Answer: Some would argue that the returns from education diminish beyond a certain point

Interpreting Linear Functions

- Example: In a society, the relationship between education (years) and income (in 1000s of dollars per year) can be summarized by: Income = 10 + 3(Education)
- Q: What is the general range of salaries?
  - 0 education = 10k, 20 yrs education = 70K
- What is the economic benefit of college?
- Would you encourage your child to attend school?
  - What if it were: Income = 30 + .2(Education) ?

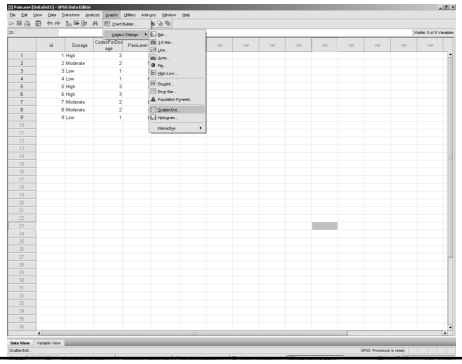
Interpreting Linear Functions

- Example: Income = 10 + 3(Education)
- How would the society be different if the constant was 0? Provide a possible social interpretation.
- How would the society be different if the constant was 30?
- How would the society be different if the slope was zero? If it was negative?
- How would the society be different if the slope was 8?

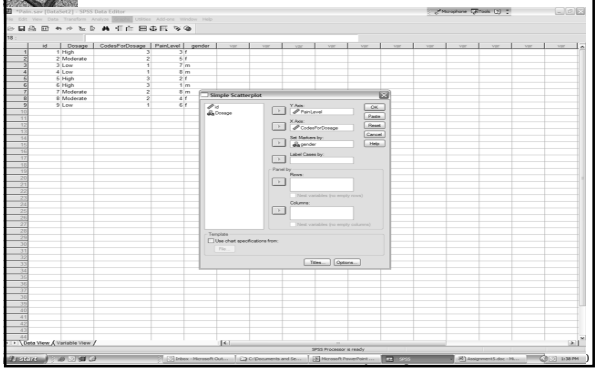
Linear Functions

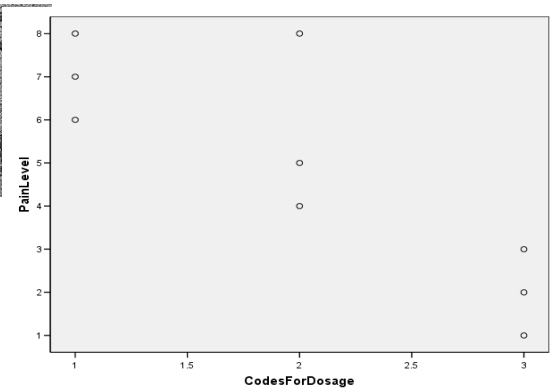
- Many issues remain:
- 1. How to test the "independence" among two continuous measures (like a chi-square test)?
  - In order to know if a linear relationship exists
- 2. How to calculate correlation coefficients (r) to measure linear association?
- 3. How to calculate the linear formula that best summarizes the relationship between two real variables (i.e., based on actual data)?
- 4. What kinds of hypothesis tests can be done?

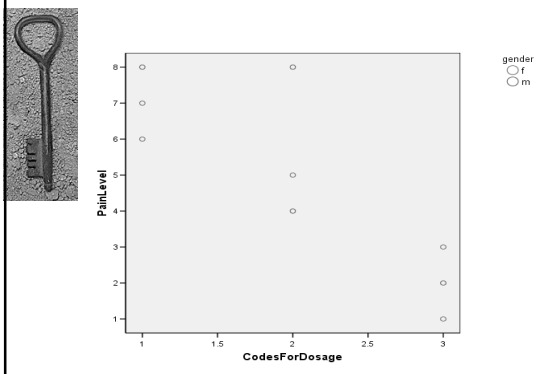
Create Scatterplot in SPSS

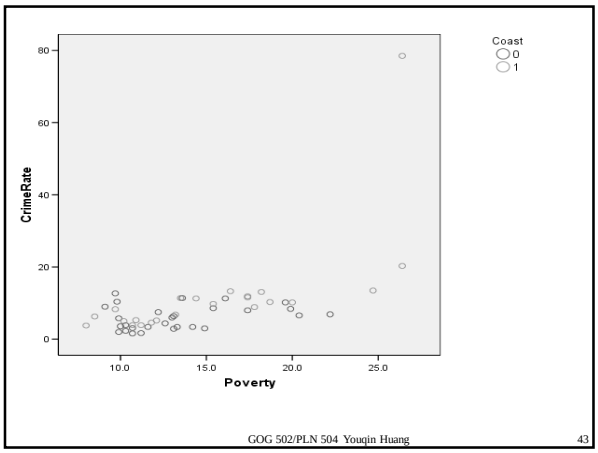


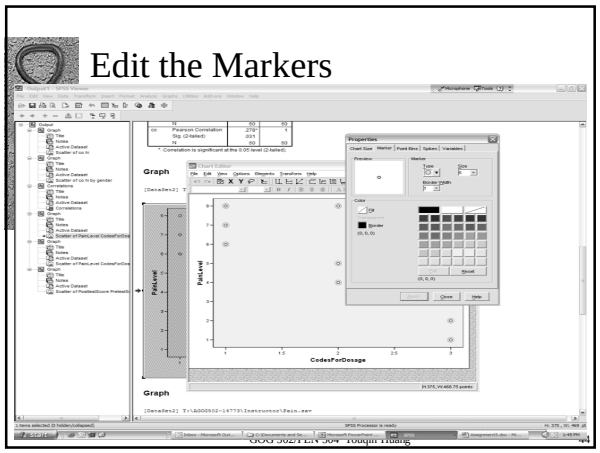
Create Scatterplot in SPSS



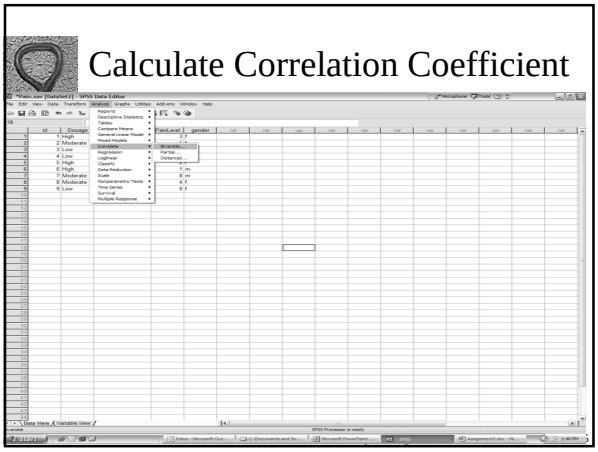




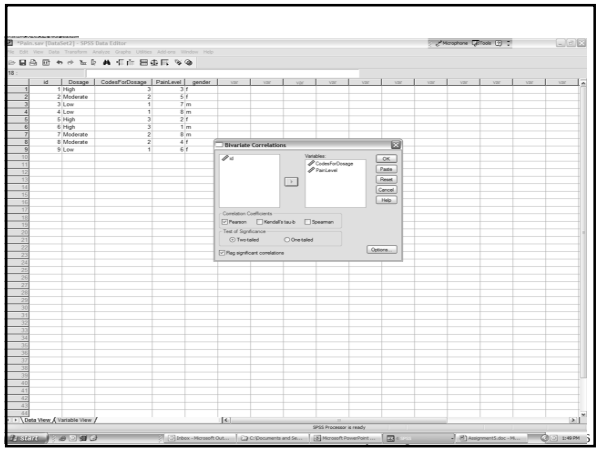




Edit the Markers

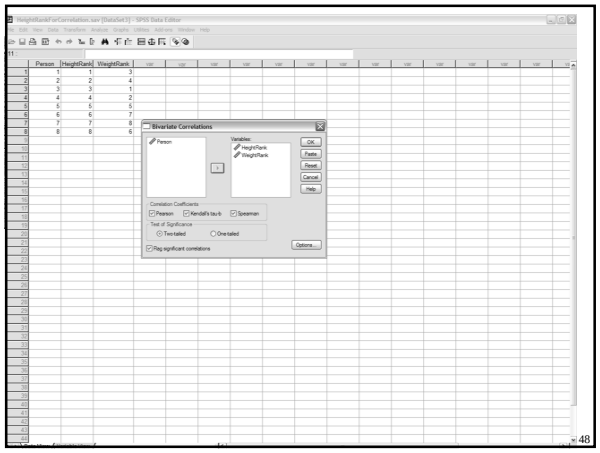


Calculate Correlation Coefficient



Correlation Coefficient

- Pearson's:
  - Linear association between two interval/ratio variable
- Kendall's Tau-b
  - Association between two ordinal variables, adjusted for ties, and most suitable for squared tables
  - Kendall's Tau-c is suitable for rectangle tables
  - Kendall's Tau-a is not adjusted for ties
- Spearman's rho
  - Association between two ordinal variables



Correlations

		HeightRank	WeightRank
HeightRank	Pearson Correlation	1	.738*
	Sig. (2-tailed)		.037
	N	8	8
WeightRank	Pearson Correlation	.738*	1
	Sig. (2-tailed)	.037	
	N	8	8

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

			HeightRank	WeightRank
Kendall's tau_b	HeightRank	Correlation Coefficient	1.000	.571*
		Sig. (2-tailed)	.	.048
		N	8	8
	WeightRank	Correlation Coefficient	.571*	1.000
		Sig. (2-tailed)	.048	.
		N	8	8
Spearman's rho	HeightRank	Correlation Coefficient	1.000	.738*
		Sig. (2-tailed)	.	.037
		N	8	8
	WeightRank	Correlation Coefficient	.738*	1.000
		Sig. (2-tailed)	.037	.
		N	8	8

*. Correlation is significant at the 0.05 level (2-tailed).

Summary

- Scatterplot
- Correlation coefficient
  - Linear association
- Linear function
  - Intercept
  - Slope
- SPSS application