Conrad

School of Entrepreneurship and Business

Agenda

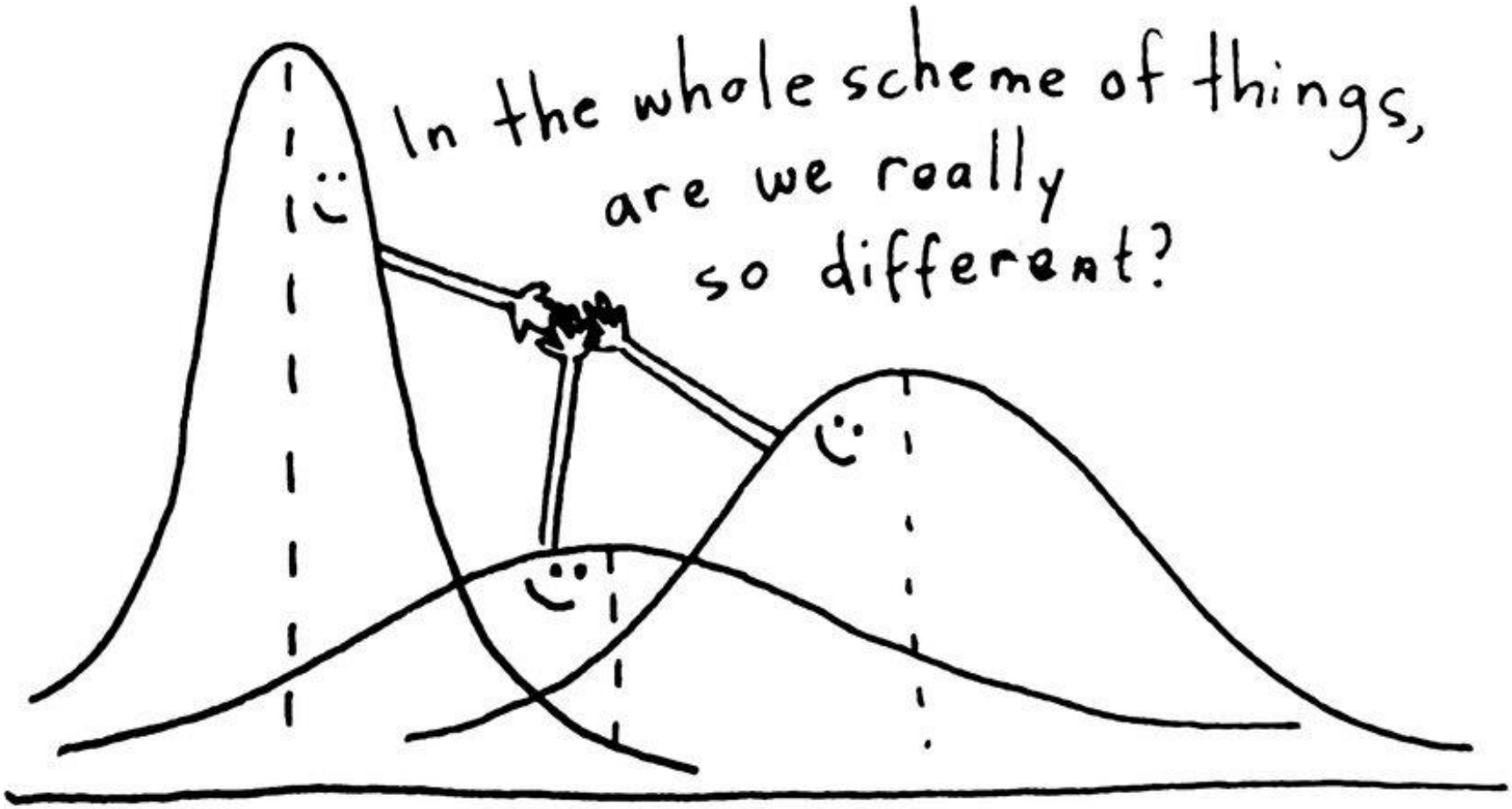
01 Recap

02 Correlation and Regression

03 Hands-on Exercrise

Statistical Inference About Difference in Means of More than Two Populations

ANOVA




ANOVA
One Way ANOVA
Two Way ANOVA

Test for the Equality of *k* Population Means

Hypotheses

H_0: $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$

H_a: Not all population means are equal

Test Statistic

F = MSTR/MSE

One-way ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Treatment	SSTR	k - 1	MSTR	MSTR/MSE
Error	SSE	n_T - k	MSE	
Total	SST	n_T - 1		

SST divided by its degrees of freedom n_T - 1 is simply the overall sample variance that would be obtained if we treated the entire n_T observations as one data set.

Two-way ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Factor A	SSA	$a - 1$	$\text{MSA} = \frac{\text{SSA}}{a-1}$	$\frac{\text{MSA}}{\text{MSE}}$
Factor B	SSB	$b - 1$	$\text{MSB} = \frac{\text{SSB}}{b-1}$	$\frac{\text{MSB}}{\text{MSE}}$
Interaction	SSAB	$(a - 1)(b - 1)$	$\text{MSAB} = \frac{\text{SSAB}}{(a-1)(b-1)}$	$\frac{\text{MSAB}}{\text{MSE}}$
Error	SSE	$ab(n_T - 1)$	$\text{MSE} = \frac{\text{SSE}}{ab(n_T - 1)}$	
Total	SST	$n_T - 1$		

Two-Way ANOVA

- If interaction is significant, then interpret it along with means and plot.
  - This indicates that the IV's are not acting separately from one another in their effect on the DV. Main effect becomes irrelevant.
- If interaction is not significant, interpret main effects.
  - This indicates that IV effects on DV are independent of one another and that there is no significant interaction of the two IV's in the population.

Examining Relationships Among Variables

The Question

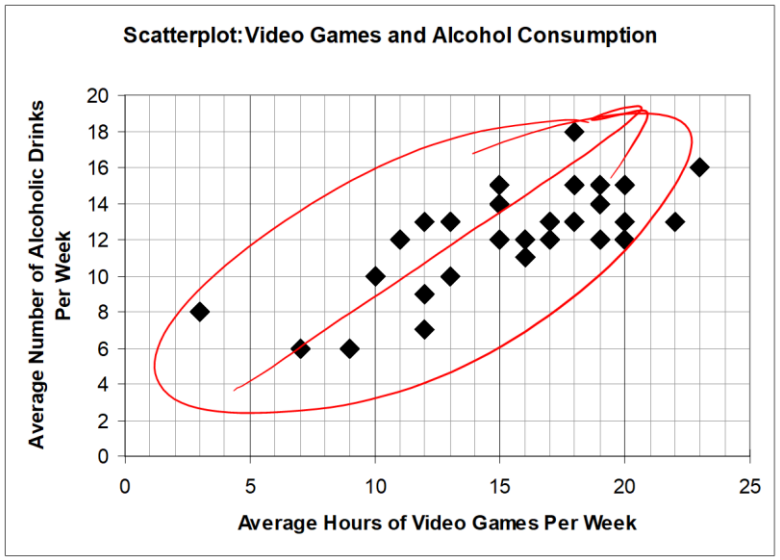
- Are two variables related?
  - Does one increase as the other increases?
    - e. g. skills and income
  - Does one decrease as the other increases?
    - e. g. nutrition and health problems

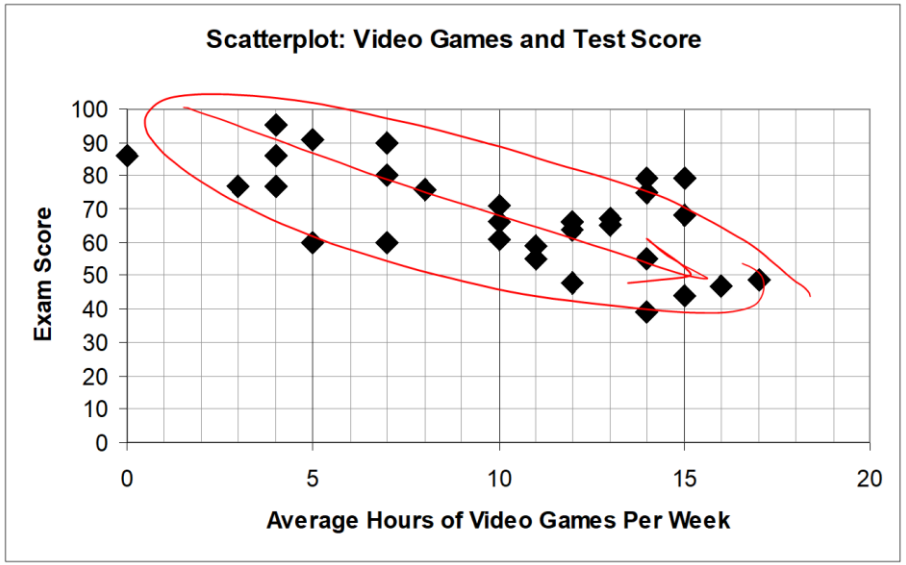
- How can we get a numerical measure of the degree of relationship?

Scatter diagram or scattergram

- Graphically depicts the relationship between two variables in two-dimensional space.

Direct Relationship





Numerical Measure of a Simple Linear Relationship

Correlation

- The relationship between two variables
- Measured with a correlation coefficient
- Most popularly seen **correlation coefficient: Pearson Product-Moment Correlation**
- Symbolized by ***r*** *(sample)* and ρ (*population*)
- A measure of degree of a **linear relationship**
- **Varies between 1 and -1**, where the sign refers to relational direction.
- **Based on covariance**
  - Measure of degree to which large scores on X go with large scores on Y, and small scores on X go with small scores on Y
- **Does not imply causation**

Hypothesis Test for Correlation Coefficient

- It is possible to test whether a correlation coefficient differs significantly from zero:

- The **test statistic** for the correlation coefficient follows a t-distribution when the null hypothesis is true.

- The **significance of the correlation coefficient** will depend on the size of the correlation coefficient and the number of observations in the sample.

- The validity of this test requires that the variables are observed on a random sample of individuals and **variables are continuous.**

Correlation Research Question

Let's revert to the brand tracking questionnaire and data made available for this class on LEARN

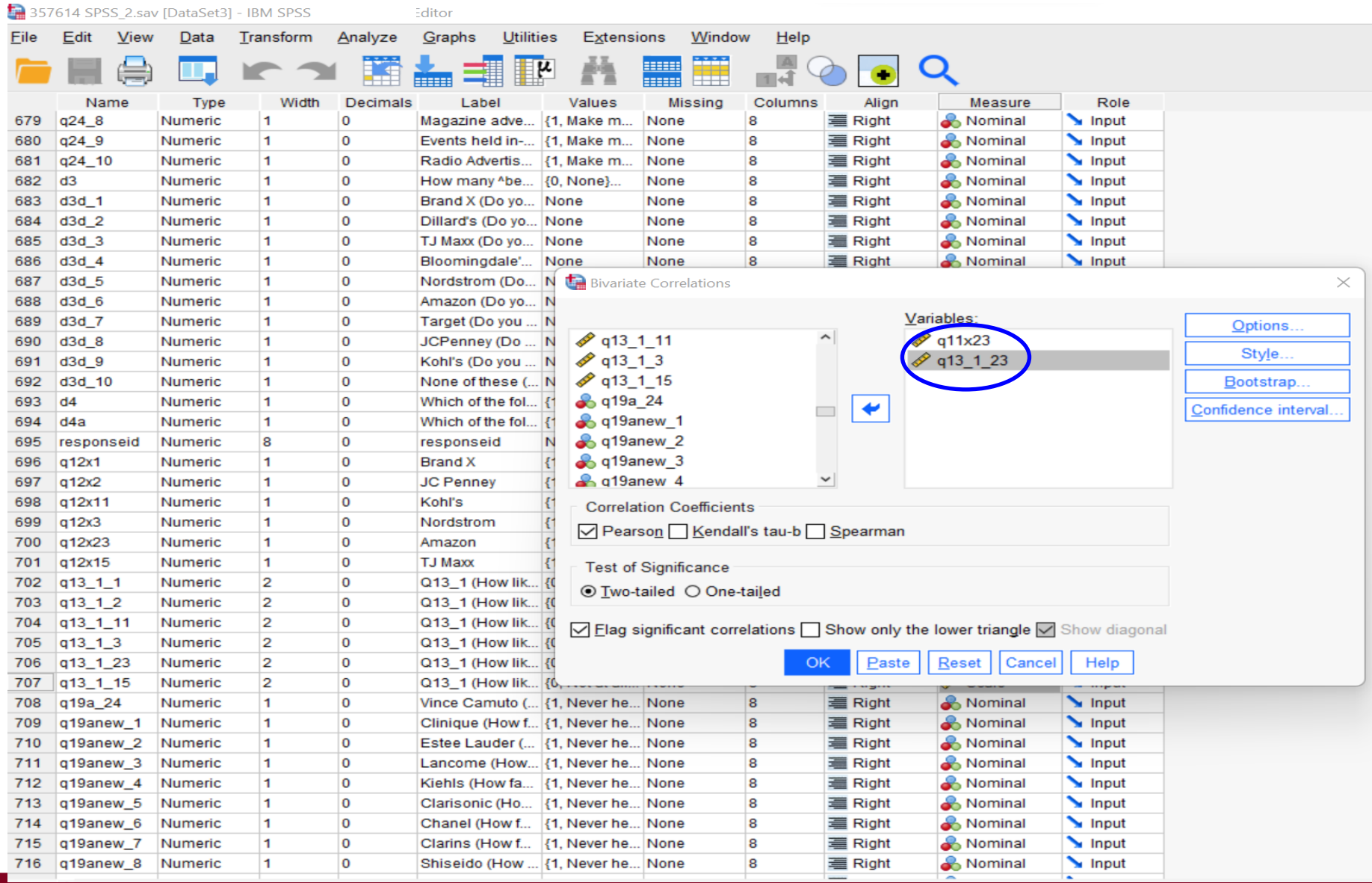
Is there an association between Brand Commitment and Likelihood to Recommend for Amazon ?

Brand Commitment						
Q11 Imagine you had to shop at a retail store, which of these statements best describes how much you would consider shopping at each of these stores? (Please select one answer for each brand) 1 PROGRAMMER: ALLOW ONLY ONE ANSWER PER BRAND. ACCEPT ONLY ONE ANSWER IN FIRST ROW "ONLY STORE WOULD CONSIDER"						
	Brand X	JC Penney	Kohl's	Nordstrom	Amazon	TJ Maxx
Favorite store; only one I consider	○	○	○	○	○	○
Store I prefer and consider highly	○	○	○	○	○	○
Store I consider equally with others	○	○	○	○	○	○
Store I might consider, less so than others	○	○	○	○	○	○
Not a store I usually consider	○	○	○	○	○	○
Store I would never consider	○	○	○	○	○	○

?

Likelihood Recommend						
Q13 How likely are you to recommend the following retailer to friends and family members, on a scale from 0-10, where 0=Not at all likely to recommend, and 10= Will definitely recommend PROGRAMMER: ALLOW ONLY ONE ANSWER PER BRAND. SHOW EACH RETAILER ON SEPARATE SCREENS.						
	Brand X	JC Penney	Kohl's	Nordstrom	Amazon	TJ Maxx
0	○	○	○	○	○	○
1	○	○	○	○	○	○
2	○	○	○	○	○	○
3	○	○	○	○	○	○
4	○	○	○	○	○	○
5	○	○	○	○	○	○
6	○	○	○	○	○	○
7	○	○	○	○	○	○
8	○	○	○	○	○	○
9	○	○	○	○	○	○
10	○	○	○	○	○	○

SPSS *Analyze | Correlate| Bivariate*



Correlation Research Question

Is there an association between Brand Commitment and Likelihood to Recommend for Amazon ?

Correlations

		Q11 Amazon	Q13 Amazon
Q11 Amazon	Pearson Correlation	1	-.497**
	Sig. (2-tailed)		<.001
	N	1608	1608
Q13 Amazon	Pearson Correlation	-.497**	1
	Sig. (2-tailed)	<.001	
	N	1608	1608

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

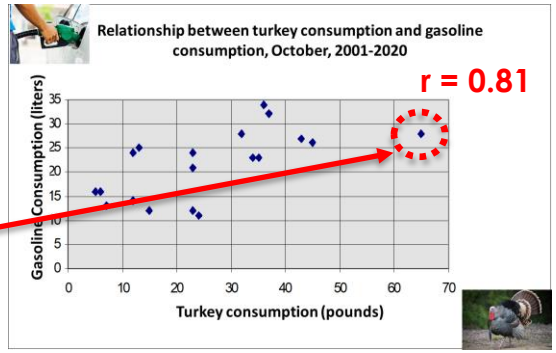
Size of Correlation	Interpretation
± 1	Perfect Positive/Negative Correlation
± .90 to ± .99	Very High Positive/Negative Correlation
± .70 to ± .90	High Positive/Negative Correlation
± .50 to ± .70	Moderate Positive/Negative Correlation
± .30 to ± .50	Low Positive/Negative Correlation
± .10 to ± .30	Very Low Positive/Negative Correlation
± .00 to ± .10	Markedly Low and Negligible Positive/ Negative Correlation

There is a **negative** association between Brand Commitment and Likelihood to Recommend for Amazon

Reject or do not reject null hypothesis.

p < 0.01, **Reject** null hypothesis, there is a significant evidence against null

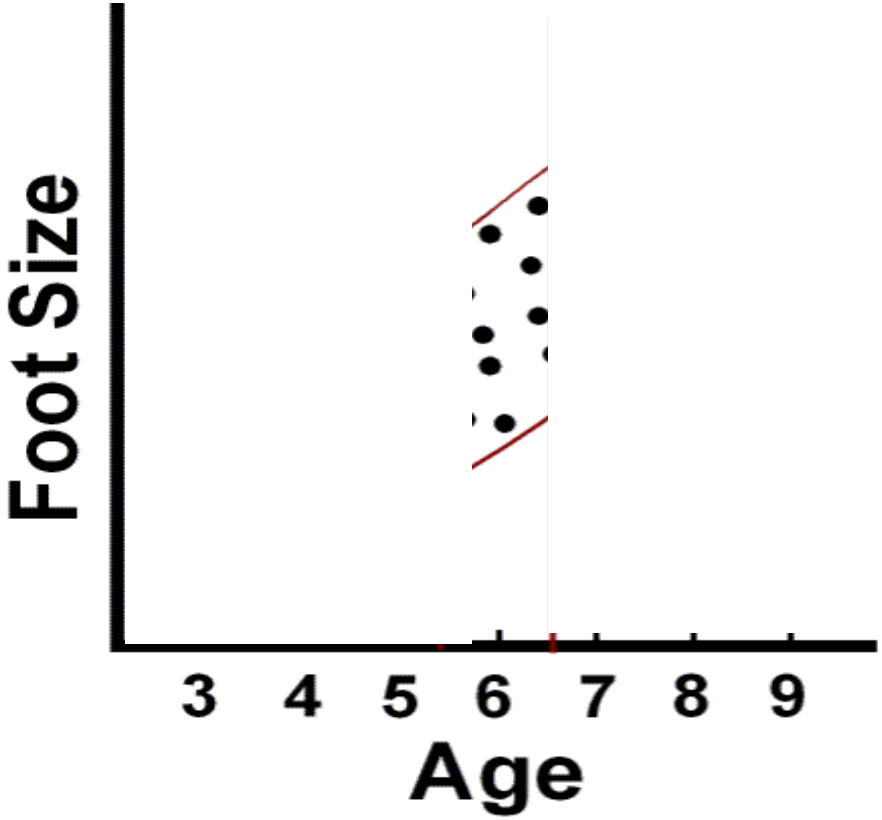
Factors Affecting *r*



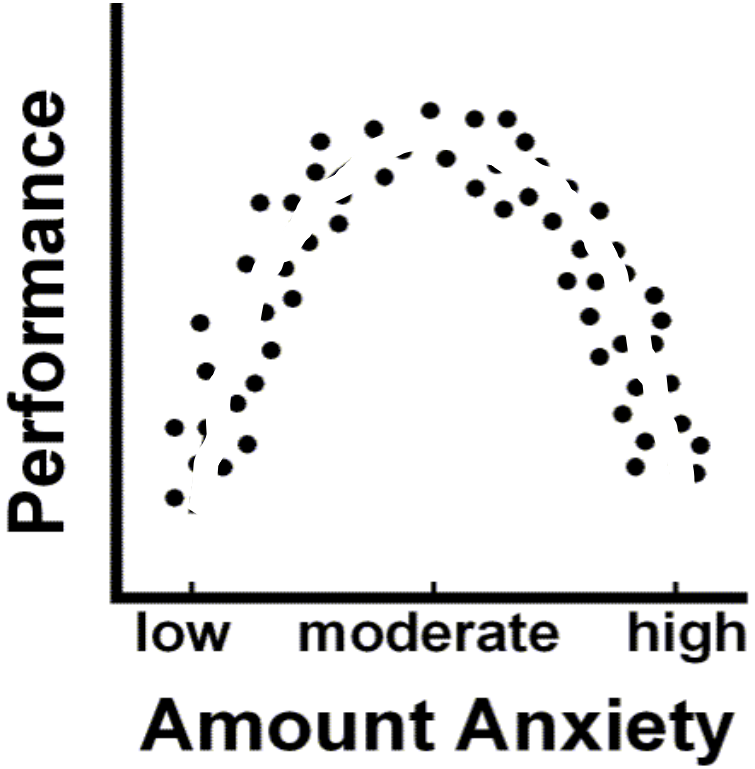
r = 0.61

- **Outliers**
  - Overestimate Correlation
  - Underestimate Correlation
- **Range restrictions**
  - Looking at only a small portion of the total scatter plot (looking at a smaller portion of the scores' variability) **decreases** *r*.
  - Reducing variability reduces *r*
- **Nonlinearity**
  - The Pearson r (and its relatives) measure the degree of **linear** relationship between two variables
  - If a strong non-linear relationship exists, r will provide a low, or at least inaccurate measure of the true relationship

Truncation



Non-linearity



Non-Parametric Correlation

- When data on at least one variable is ordinal, a rank correlation method can be applied (***Spearman's rank correlation***).

- The rank correlation coefficient is calculated in the same way as for Pearson's correlation coefficient, except that it is calculated on the ranks and not the actual values,

- It ranges from -1 to +1 and has the same interpretation,

- No requirement for the data to follow a Normal distribution (non-parametric).

Spearman's Rank Correlation

Let's revert to the brand tracking questionnaire and data made available for this class on LEARN

We need to determine whether there is a correlation between **Employment Status** and **Annual Household Income**

Employment Status is measured on a **6 point ordinal scale** and Annual Household Income measured through a **7 point ordinal scale**

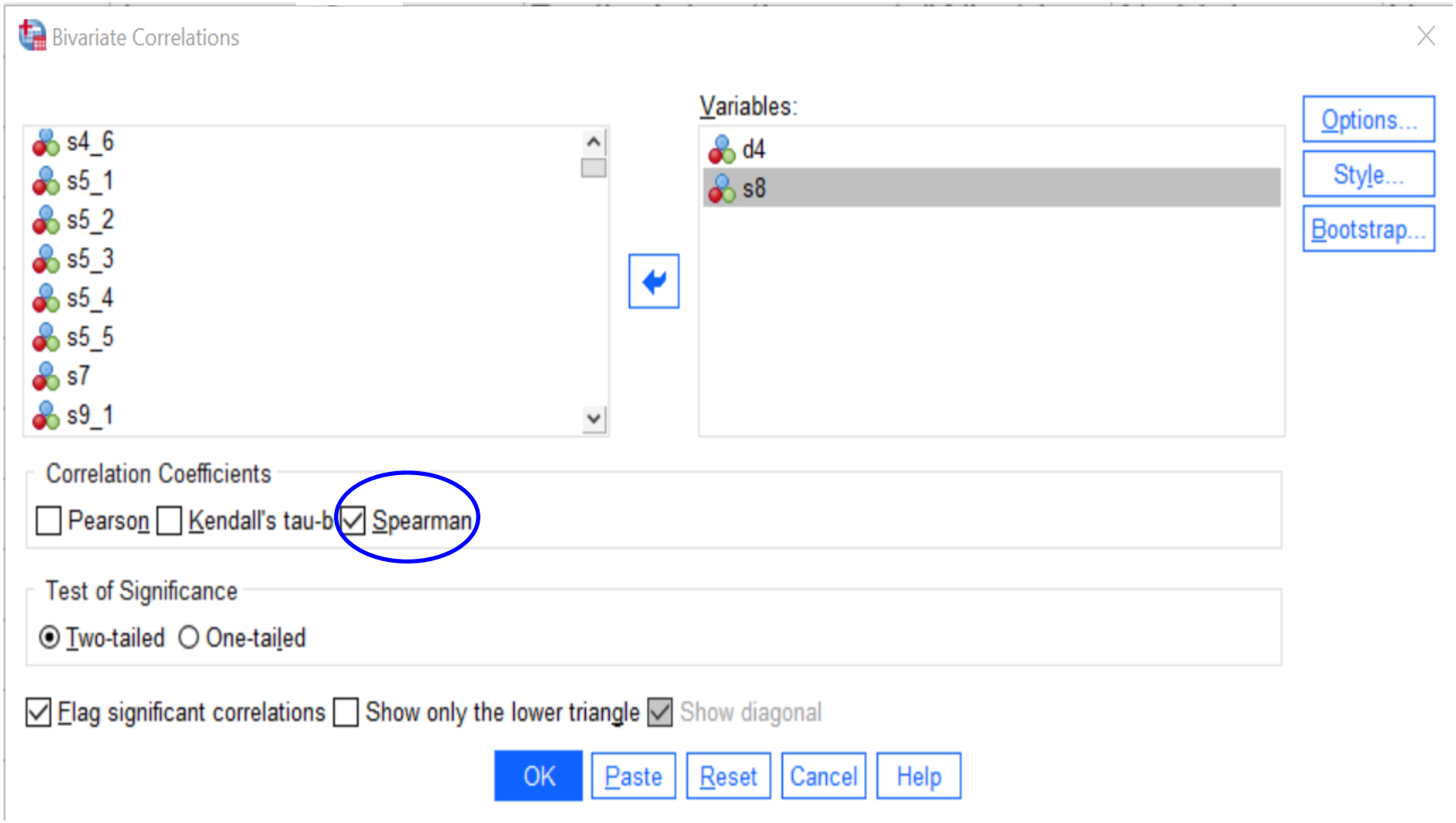
Fix this – Add to Missing Value

Employment Status		
D4	Which of the following statements best describes your current employment status? (*Please select one answer*) **PROGRAMMER: ACCEPT ONLY ONE ANSWER. RADIO BUTTONS.**	
	Employed full time (with an employer or self-employed)	○
	Employed part time (with an employer or self-employed)	○
	Manage household	○
	Full time student	○
	Retired	○
	Not currently employed	○
	Prefer not to say	○

S8	Which of the following best describes your annual household income? (*Please select one answer*)	GO TO
	Under $25,000 $25,000 to less than $35,000	Terminate
	$35,000 to less than $50,000 $50,000 to less than $75,000 $75,000 to less than $100,000 $100,000 to less than $125,000 $125,000 to less than $150,000 $150,000 to less than $200,000 $200,000 or more	S9

H_0 : there is no relationship between Employment Status and Annual Household Income in the population

SPSS *Analyze | Correlate| Bivariate*



Spearman's Rank Correlation

Is there a linear association between Employment Status and Annual H Income?

Size of Correlation	Interpretation
± 1	Perfect Positive/Negative Correlation
± .90 to ± .99	Very High Positive/Negative Correlation
± .70 to ± .90	High Positive/Negative Correlation
± .50 to ± .70	Moderate Positive/Negative Correlation
± .30 to ± .50	Low Positive/Negative Correlation
± .10 to ± .30	Very Low Positive/Negative Correlation
± .00 to ± .10	Markedly Low and Negligible Positive/ Negative Correlation

Correlations

			Which of the following statements best describes your current employment status?	Which of the following best describes your annual household income?
Spearman's rho	Which of the following statements best describes your current employment status?	Correlation Coefficient	1.000	-.108**
		Sig. (2-tailed)	.	.000
		N	4320	4320
	Which of the following best describes your annual household income?	Correlation Coefficient	-.108**	1.000
		Sig. (2-tailed)	.000	.
		N	4320	4331

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

There is a **negative** association between Employment Status and Annual Household Income

Reject or do not reject null hypothesis.

$p < 0.01$, **Reject** null hypothesis, there is a significant evidence against null

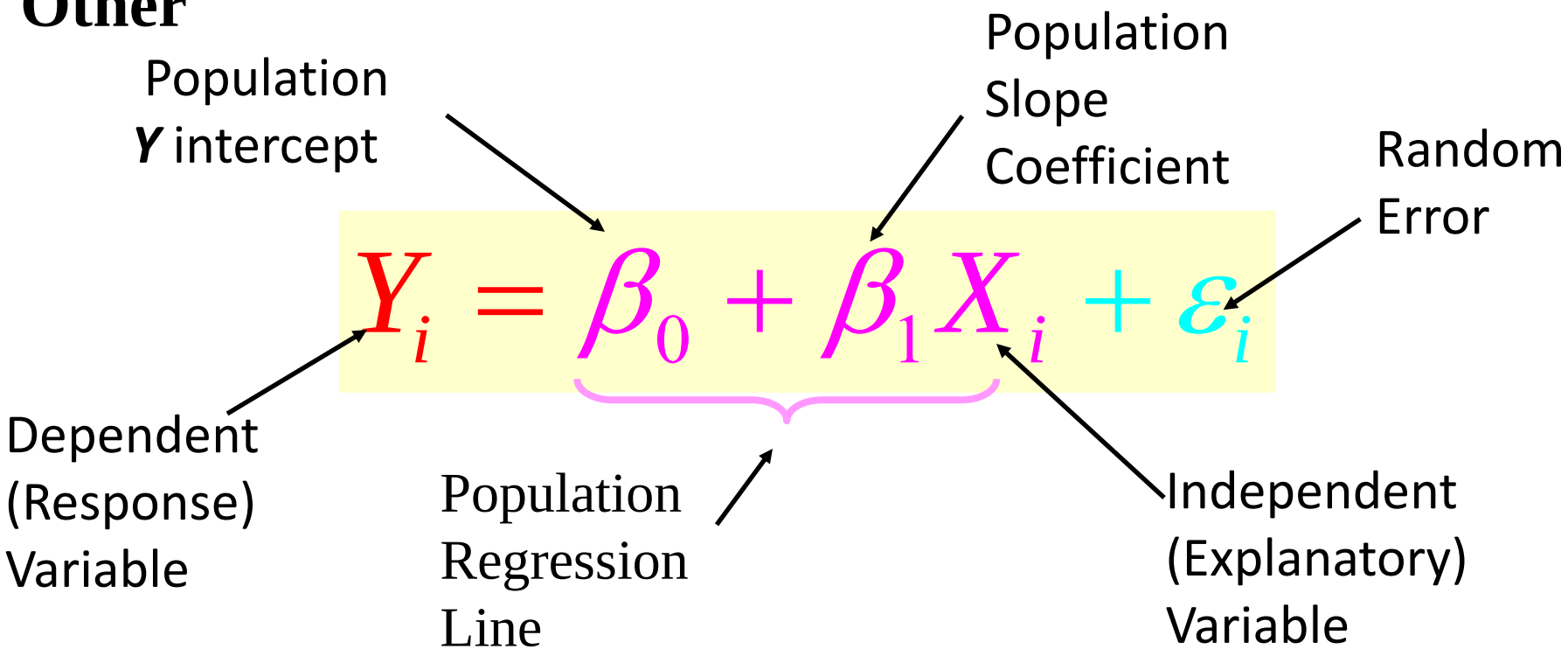
Describing the Linear Relationship

Simple Linear Regression

- Simple linear regression describes the relationship between two continuous variables
- Simple linear regression gives the equation of the straight line that best describes the association between two interval and or ratio variables
- It enables the prediction of one variable using information from another variable
- The dependent variable is the variable to be predicted (i.e. the particular outcome interested in)
  - Also called endogenous or response variable
- The independent variable or explanatory variable is the variable used for predicting the particular outcome.
  - Also called exogenous variable(s)

Population Linear Regression

- **Population Regression Line Is A Straight Line that Describes The Dependence of One Variable on The Other**



Sample Linear Regression

Sample Regression Line Provides an Estimate of The Population Regression Line

Sample Y Intercept

Sample Slope Coefficient

$$Y_i = b_0 + b_1 X_i + e_i$$

Residual

Sample Regression Line

b_0 **provides an estimate of** β_0

b_1 **provides an estimate of** β_1

X	Y
1	4
2	24
4	10
5	32

How is the Best Line Found?

Least-Squares Property

A straight line satisfies this property if the sum of the squares of the residuals is the smallest sum possible.

Testing Hypothesis

H_0: There is no linear relationship between Brand Commitment with Brand X and Likelihood to Recommend for Brand X

H_a: There is a linear relationship between Brand Commitment with Brand X and Likelihood to Recommend for Brand X

A Gut Check – Correlation!

Is there a linear association between Brand Commitment and Likelihood to Recommend for Brand X?

Correlations

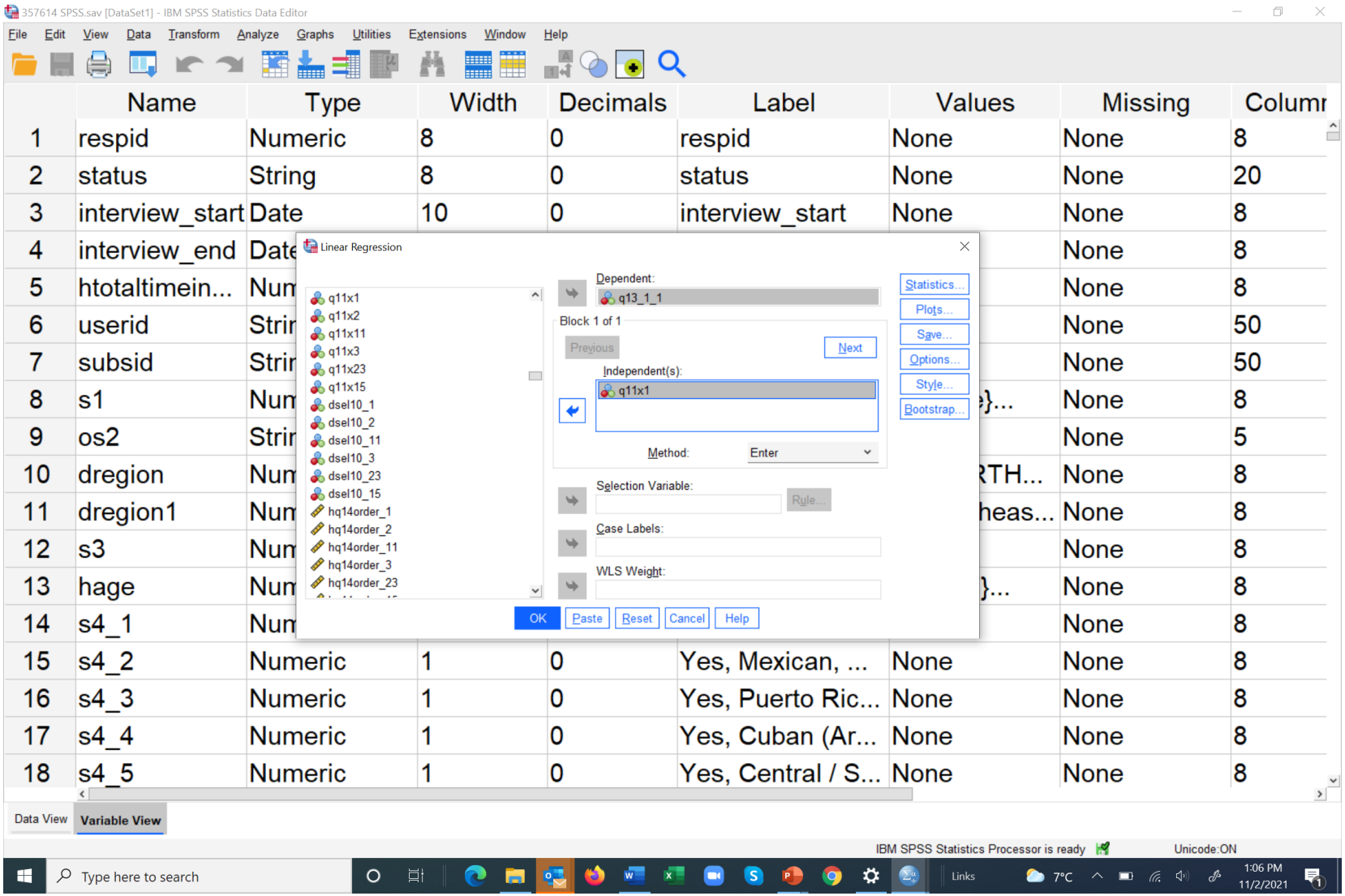
		Brand X	Q13_1 (How likely are you to recommend the following retailer to friends and family members, on a scale from 0-10, where 0=Not at all likely to recommend, and 10= Will definitely recommend. Brand X)
Brand X	Pearson Correlation	1	-.705**
	Sig. (2-tailed)		.000
	N	4331	4331
Q13_1 (How likely are you to recommend the following retailer to friends and family members, on a scale from 0-10, where 0=Not at all likely to recommend, and 10= Will definitely recommend. Brand X)	Pearson Correlation	-.705**	1
	Sig. (2-tailed)	.000	
	N	4331	4331

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

There is a **Significant Negative** association between Brand Commitment and Likelihood to Recommend for Brand X

Describing the relationship – Regression time!

SPSS *Analyze | Regression| Linear*



	Name	Type	Width	Decimals	Label	Values	Missing	Columr
1	respid	Numeric	8	0	respid	None	None	8
2	status	String	8	0	status	None	None	20
3	interview_start	Date	10	0	interview_start	None	None	8
4	interview_end	Date					None	8
5	htotaltimein...	Num					None	8
6	userid	Strir					None	50
7	subsid	Strir					None	50
8	s1	Num				e}...	None	8
9	os2	Strir					None	5
10	dregion	Num				RTH...	None	8
11	dregion1	Num				heas...	None	8
12	s3	Num					None	8
13	hage	Num				}...	None	8
14	s4_1	Num					None	8
15	s4_2	Numeric	1	0	Yes, Mexican, ...	None	None	8
16	s4_3	Numeric	1	0	Yes, Puerto Ric...	None	None	8
17	s4_4	Numeric	1	0	Yes, Cuban (Ar...	None	None	8
18	s4_5	Numeric	1	0	Yes, Central / S...	None	None	8

SPSS: 1. Coefficient of Determination

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.705[a]	.497	.496	1.664

a. Predictors: (Constant), Brand X

- R is the correlation between the two variables

- R square (coefficient of determination) is the proportion of variability in Likelihood to Recommend measurements that can be explained by Brand Commitment with Brand X.

SPSS: 2. Linear Regression

Test Significance of Model

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11819.739	1	11819.739	4269.953	.000[b]
	Residual	11983.187	4329	2.768		
	Total	23802.926	4330			

a. Dependent Variable: Q13_1 (How likely are you to recommend the following retailer to friends and family members, on a scale from 0-10, where 0=Not at all likely to recommend, and 10= Will definitely recommend. Brand X)

b. Predictors: (Constant), Brand X

Interpretation of ANOVA table

- **A statistically significant proportion of the variability in Likelihood to Recommend for Brand X can be attributed to the regression model (P<0.01).**

SPSS: 3. Regression Equation

Test Significance of Coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	11.334	.072		158.308	.000
	Brand X	-1.443	.022	-.705	-65.345	.000

a. Dependent Variable: Q13_1 (How likely are you to recommend the following retailer to friends and family members, on a scale from 0-10, where 0=Not at all likely to

If the *t*-test is significant (say P≤ 0.01) then that predictor has accounted for a significant proportion of the variation of the response variable, in addition to variation of the response variable that has been accounted for by the other predictors in the model.

Predicted Likelihood to Recommend Brand X = constant + B **Brand Commitment with Brand X**

Predicted Likelihood to Recommend Brand X = 11.334 + (-1.443) **Brand Commitment with Brand X**

Prediction

How do you use linear regression for prediction?

The regression equation allows you to predict the value of the dependent variable (Y) for a particular value of the independent variable (X),

Predicted Likelihood to Recommend Brand X = **11.334** + **(-1.443)** **Brand Commitment with Brand X**

Lets check the current mean value of Brand Commitment for Brand X

SPSS *Analyze | Descriptive Statistics| Descriptive*

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Brand X	4331	1	6	3.03	1.145
Valid N (listwise)	4331				

Brand Commitment

Q11 Imagine you had to shop at a retail store, which of these statements best describes how much you would consider shopping at each of these stores? (Please select one answer for each brand)

1 PROGRAMMER: ALLOW ONLY ONE ANSWER PER BRAND.
ACCEPT ONLY ONE ANSWER IN FIRST ROW "ONLY STORE WOULD CONSIDER"

	Brand X	JC ~~Penney~~Target	Kohl's	Nordstrom	Amazon	TJ Maxx
Favorite store; only one I consider	O	O	O	O	O	O
Store I prefer and consider highly	O	O	O	O	O	O
Store I consider equally with others	O	O	O	O	O	O
Store I might consider, less so than others	O	O	O	O	O	O
Not a store I usually consider	O	O	O	O	O	O
Store I would never consider	O	O	O	O	O	O

Prediction

Predicted Likelihood to Recommend Brand X

= 11.334 + (-1.443)*3.03

= 6.96171

= ~ 7

Likelihood Recommend						
Q13 How likely are you to recommend the following retailer to friends and family members, on a scale from 0-10, where 0=Not at all likely to recommend, and 10= Will definitely recommend. PROGRAMMER: ALLOW ONLY ONE ANSWER PER BRAND. SHOW EACH RETAILER ON SEPARATE SCREENS.						
	Brand X	~~JC Penney~~Target	Kohl's	Nordstrom	Amazon	TJ Maxx
0	○	○	○	○	○	○
1	○	○	○	○	○	○
2	○	○	○	○	○	○
3	○	○	○	○	○	○
4	○	○	○	○	○	○
5	○	○	○	○	○	○
6	○	○	○	○	○	○
7	○	○	○	○	○	○
8	○	○	○	○	○	○
9	○	○	○	○	○	○
10	○	○	○	○	○	○

Let's say we are able to move the average Brand Commitment for Brand X from ~ 3 to ~1.5

Predicted Likelihood to Recommend for Brand X

= 11.334 + (-1.443)*1.5

= 9.1695

= ~ 9

Group Work

- Since we have 78 registered students, 20 groups will be created with students being randomly assigned
- Please submit your work no later than 10:00pm EST. There is no drop box, work needs to be submitted via email to harvir.bansal@uwaterloo.ca and sprana@uwaterloo.ca
- Please ensure that you clearly outline the steps taken in your analysis as well as the results.
- As I have said multiple times, data analysis is not a spectator sport so PLAY and have fun

Hands-on Analysis

- **All questions below refer to the brand tracking questionnaire and data made available for this class on LEARN**
  1. What is the nature of association between Brand Commitment and Likelihood to Recommend for the 6 brands included in the data ?
  2. Amazon is interested in knowing whether the satisfaction with their reward program differs by the gender and household income.
  3. JC Penny and Amazon are both interested in knowing whether there is a relationship between annual household income and the respective Likelihood to Purchase (time frame) for their brands.
  4. Does perceptions of Brand Love for JC Penny impact the Likelihood of Recommendation? What is the strength and nature of that relationship?

TILL NEXT TIME