

Correlation and Regression

PSY 5101: Advanced Statistics for Psychological and Behavioral Research I

Correlation and Regression

- ◎ Both examine linear (straight line) relationships
- ◎ Correlation works with a pair of scores
 - One score on each of two variables (X and Y)
- ◎ Correlation:
 - Defined as the **degree of linear relationship between X and Y**
 - Is measured/described by the statistic r
- ◎ Regression:
 - Describes the form or function of the linear relationship between X & Y
 - Is concerned with the prediction of Y from X
 - Forms a prediction equation to predict Y from X

Why do we care?

- ◎ Central tendency and dispersion are critical for DESCRIBING the characteristics of a distribution
- ◎ However, sometimes we are interested in the relationship between variables (i.e., how the value of one variable changes when the value of another variable changes)
 - Correlation and regression help us understand these relationships

Correlation

- The aspect of the data that we want to describe/measure is the degree of linear relationship between X and Y
- The statistic *r* describes/measures the degree of linear relationship between X and Y
 - Pearson product moment correlation coefficient
- $$r = \frac{\sum(z_x \cdot z_y)}{N-1} = \frac{Cov(X,Y)}{\sqrt{Var X} \sqrt{Var Y}}$$
- The average product of z scores for X and Y
- Works with two variables (X and Y)
- $-1 \leq r \leq 1$
- *r* measures positive or negative relationships
- Measures only the degree of **linear** relationship
- r^2 = proportion of variability in Y that is explained by X
- *r* is undefined if X or Y has zero spread

Correlation Matrix

Correlations

		npi_tot	mach_tot	srps_tot	new_ssis_tot
npi_tot	Pearson Correlation	1	.116	.421**	.236*
	Sig. (2-tailed)		.165	.000	.004
	N	145	145	145	145
mach_tot	Pearson Correlation	.116	1	.477**	.434**
	Sig. (2-tailed)	.165		.000	.000
	N	145	145	145	145
srps_tot	Pearson Correlation	.421**	.477**	1	.530**
	Sig. (2-tailed)	.000	.000		.000
	N	145	145	145	145
new_ssis_tot	Pearson Correlation	.236*	.434**	.530**	1
	Sig. (2-tailed)	.004	.000	.000	
	N	145	145	145	145

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

Correlation: $-1 \leq r \leq 1$

- The sign of *r* shows the type of linear relationship between X and Y
- We can use the definitional formula for *r* and these scatterplots to see positive, negative, and zero relationships

$r = 1$

$r = -1$

$r = 0$

The Formula for Correlation

- ◉ $\frac{Cov(X,Y)}{\sqrt{Var X} \sqrt{Var Y}}$
- ◉ Formula has two parts
 1. Measure of the association between two variables
 2. Standardizing process
- ◉ Numerator deals with the linear association
- ◉ Denominator deals with standardization

Covariance

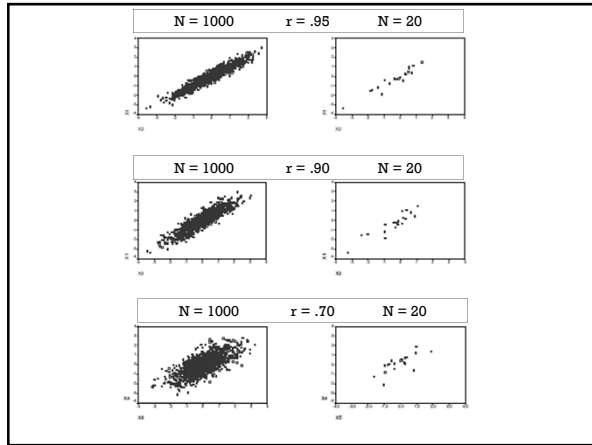
- ◉ Covariance between two variables captures the linear association between two variables
- ◉ Covariance is similar to the variance except that the covariance deals with two variables whereas the variance only deals with a single variable
- ◉ Covariance is a measure of the direction and magnitude of the linear association between X and Y
- ◉ Covariance depends on the scale of the variable
 - Variables on a 1-7 scale will have a smaller covariance than variables on a 1-100 scale with the same association

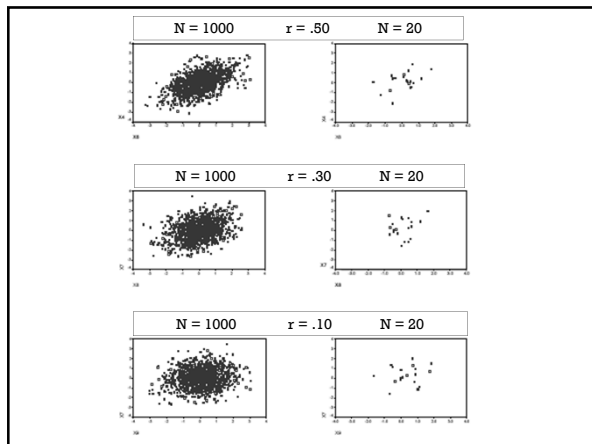
Standardizing Process

- ◉ Dividing the covariance by the standard deviation of each variable serves to standardize the covariance
- ◉ Correlation ranges between -1 and +1 regardless of the scales of the variables involved
- ◉ r is a measure of effect size

Correlation: $-1 < r < 1$

What happens to Variable X	What happens to Variable Y	Type of Correlation	Value	Example
X increases in value	Y increases in value	Positive	Positive, ranging from 0 to +1	The more time you spend studying, the higher your test score will be
X decreases in value	Y decreases in value	Positive	Positive, ranging from 0 to +1	The less money you put in the bank, the less interest you will earn
X increases in value	Y decreases in value	Negative	Negative, ranging from 0 to -1	The more you exercise, the less you will weigh
X decreases in value	Y increases in value	Negative	Negative, ranging from 0 to -1	The less time you take to complete a test, the more you'll get wrong





1 0.8 0.4 0 -0.4 -0.8 -1
 1 1 1 -1 -1 -1
 0 0 0 0 0 0 0

- The correlation reflects the degree and direction of linear relationship (top row)
- Correlation does not reflect the slope of the relationship (middle row)
- Correlation does not reflect many aspects of nonlinear relationships (bottom row)

Correlation: Linear

- If there is a curvilinear relationship between X and Y, then r will not detect it
- The value of r will be zero if there is no linear relationship between X and Y

r = 0




r = 0




Interpreting Correlation Coefficients

<u>Size of correlation</u>	<u>Interpretation</u>
.8 to 1.0	very strong relationship
.6 to .8	strong relationship
.4 to .6	moderate relationship
.2 to .4	weak relationship
.0 to .2	weak or no relationship

Correlation: r^2

- r^2 = proportion of variability in Y that is explained by X
 - Also known as coefficient of determination
 - If $r = .5$, $r^2 = .25$, so the proportion of variability in Y that is explained by X is .25 (25% explained by X, 75% unexplained)
- Scatterplots:

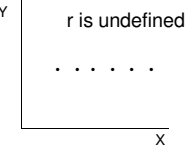
$r = .5, r^2 = .25$	$r = .7, r^2 = .49$	$r = .9, r^2 = .81$
		
- Venn Diagrams: r^2 is represented by the proportion of overlap.

$r = .5, r^2 = .25$	$r = .7, r^2 = .49$	$r = .9, r^2 = .81$
		

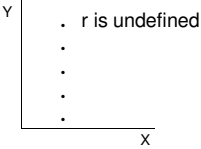
Correlation: Undefined

- If there is no spread in X or Y, then r is undefined. Note that any z is undefined if the standard deviation is zero
 - This is a problem because $r = \frac{\sum z_x z_y}{N-1}$

$s_y = 0$



$s_x = 0$



Correlation and Causation

- Correlation as a statistical technique vs. correlational research design
- Correlational research is non-experimental
 - This is what leads to the common idea of “correlation does not imply causation”
- Correlation as a statistical technique is simply a way to analyze data
 - This has nothing to do with issues of causation
- Common misunderstanding
 - Using particular analyses (e.g., ANOVA instead of Pearson’s r) does not allow researchers to make assumptions about causation

Correlation and Causation

- ◉ Example of correlation:
 - Murder rates and ice cream sales are positively correlated
 - As murder rates increase, ice cream sales also increase
 - Why?
- ◉ Possible explanations for the causal connections underlying correlational findings
 - Murders may cause increases in ice cream sales
 - Ice cream sales may cause more murders
 - Some other variable may cause both murders and ice cream sales

Correlation

- ◉ Things to remember:
 - Correlations can range from -1 to +1
 - The absolute value of the correlation coefficient reflects the strength of the correlation
 - A correlation of -.7 is stronger than a correlation of +.5
 - Do not assign a value judgment to the sign of the correlation
 - Negative correlations are not "bad" and positive correlations are not "good"
 - Population correlation coefficient is ρ (rho)
 - Impact on r:
 - Restriction of range
 - Extreme scores (outliers)

Regression

- Not only can we compute the degree to which two variables are related (correlation coefficient) but we can use these correlations as the basis for predicting the value of one variable from the value of the other
- Prediction is an activity that estimates future outcomes from present ones
 - When we want to predict one variable from another, we need to first compute the correlation between the two variables

Regression

Total High School GPA and First-Year College GPA are Correlated

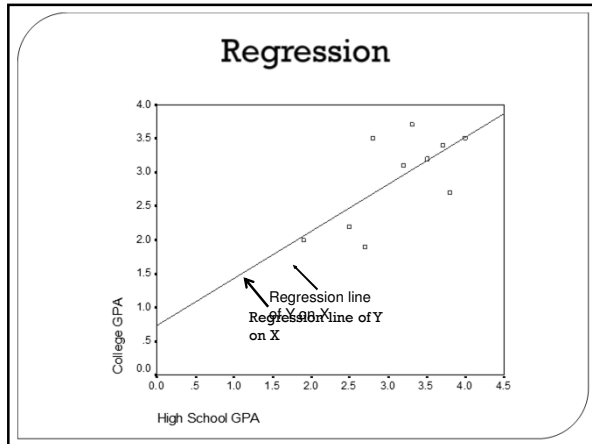
High School GPA	First-Year College GPA
3.5	3.2
2.5	2.2
4.0	3.5
3.8	2.7
2.8	3.5
1.9	2.0
3.2	3.1
3.7	3.4
2.7	1.9
3.3	3.7

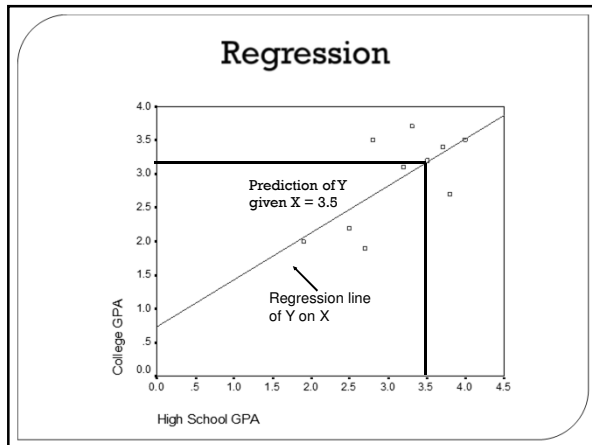
Regression

$r = .68$ for these two variables

Regression

- ⊙ Regression is concerned with forming a prediction equation to predict Y from X
- ⊙ Uses the formula for a straight line: $Y' = bX + a$
 - Y' is the predicted Y score on the criterion variable
 - b is the slope, $b = \Delta Y / \Delta X = \text{rise/run}$
 - X is a score on the predictor variable
 - a is the Y-intercept, where the line crosses the Y axis, the value of Y' when X=0
 - Example: if $b = .695$, $a = .739$, and $X = 3.5$,
 - then $Y' = .695(3.5) + .739 = 3.17$





- ### Regression
-
- Linear only
 - Generalize only for X values in your sample
 - Actual observed Y is different from Y' by an amount called error (e)
 - That is, $Y = Y' + e$
 - Error in regression is $e = Y - Y'$
 - Many different potential regression lines

Regression: Best-Fitting Line

- There are many different potential regression lines, but only one "best-fitting" line

- The statistics b and a are computed so as to minimize the sum of squared errors
 - $\sum e^2 = \sum (Y - Y')^2$ is a minimum which is called the Least Squares Criterion
 - This means that it minimizes the distance between each individual point and the regression line

Regression

• Error in prediction: the distance between each individual data point and the regression line (a direct reflection of the correlation between two variables)

Regression: $s_{y \cdot x}$

- Standard error of estimate is a statistic that measures/describes spread of errors or Y scores in regression
- $s_{y \cdot x}$ is the standard deviation of errors in regression

$$s_{y \cdot x} = \sqrt{\frac{\sum e^2}{(N-2)}} = \sqrt{\frac{\sum (Y - Y')^2}{(N-2)}}$$

- As r^2 increases, $s_{y \cdot x}$ decreases
- For example, if $N=100$ and $s_y=4$

r^2	$s_{y \cdot x}$
.2	3.94
.4	3.68
.6	3.22
.8	2.41
.9	1.75

$s_{y \cdot x}$ is the standard deviation of Y around the regression line Y'

Regression: Partitioning

◎ Partitioning total variability

- Total = Explained + Not Explained
- This is true for proportion of spread and amount of spread
 - Proportion: $1 = r^2 + (1-r^2)$
 - Amount: $s_y^2 = r^2*s_y^2 + (1-r^2)s_y^2$

	Total	Explained	Not Explained
Proportion	1	r^2	$1-r^2$
Amount	s_y^2	$r^2*s_y^2$	$(1-r^2)s_y^2$

Regression: Partitioning

◎ Example:

- $r=.7, s_y^2=150$

	Total	Explained	Not Explained
Proportion	1	$r^2 = .49$	$1-r^2 = .51$
Amount	$s_y^2 = 150$	$r^2*s_y^2 = 73.5$	$(1-r^2)s_y^2 = 76.5$
