Bivariate Regression & Correlation

- Overview
- The Scatter Diagram
- Two Examples: Education & Prestige
- Correlation Coefficient
- Bivariate Linear Regression Line
- SPSS Output
- Interpretation
- Covariance

Overview

You already know how to deal with two nominal variables

Independent Variables

Dependent Variable	Nominal	Interval
Nominal	**Lambda**	Considers how a change in a variable affects a discrete outcome
Interval	Considers the difference between the mean of one group on a variable with another group	Considers the degree to which a change in one variable results in a change in another

Overview

You already know how to deal with two nominal variables

TODAY!

Independent Variables

Dependent Variable	Nominal	Interval
Nominal	**Lambda**	Considers how a change in a variable affects a discrete outcome
Interval	**Confidence Intervals T-Test**	Considers the degree to which a change in one variable results in a change in another

You have already learned this (e.g., compare two groups)

Overview

You already know how to deal with two nominal variables

This cell is not covered in this course

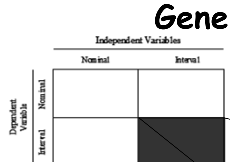
Independent Variables

Dependent Variable	Nominal	Interval
Nominal	**Lambda**	~~**Logistic Regression**~~
Interval	**Confidence Intervals T-Test**	**Regression Correlation**

You know how to do this.

TODAY!

General Examples

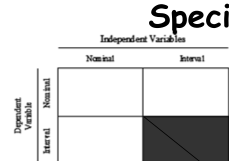


Does the independent variable reduce the error when trying to predict the cases of the dependent variable (treated as a PRE measure).

Does the independent variable significantly affect the dependent variable?

Is the independent variable highly associated with the dependent variable?

Specific Examples



Does a person's salary level increase with years of work experience?

Does marital satisfaction increase with length of marriage?

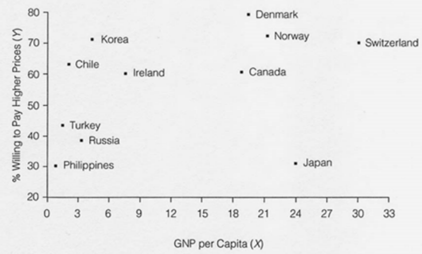
How does an additional year of education affect one's job prestige score?

Scatter Diagrams

- **Scatter Diagram (scatterplot)**—a visual method used to display a relationship between two interval-ratio variables (it wouldn't work to use a table because each interval level variable has too many categories to fit into a table.

Scatter Diagram Example of a Positive Relationship

Figure 8.1 **Scatter Diagram of GNP per Capita (in $1,000) and Percentage Willing to Pay More to Protect the Environment**



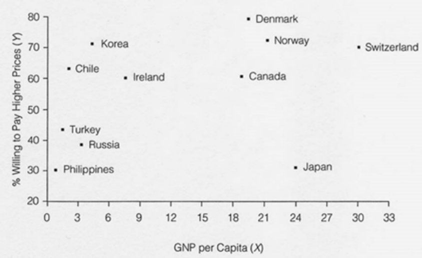
Source: Adapted from Steven R. Brechin and Willett Kempton, "Global Environmentalism: A Challenge to the Postmaterialism Thesis?" *Social Science Quarterly 75*, no. 2 (June 1994): 245–266.

Scatter Diagrams

Typically, the independent variable is placed on the X-axis (horizontal axis), while the dependent variable is placed on the Y-axis (vertical axis.)

Scatter Diagram Example of a Positive Relationship

Figure 8.1 **Scatter Diagram of GNP per Capita (in $1,000) and Percentage Willing to Pay More to Protect the Environment**



Source: Adapted from Steven R. Brechin and Willett Kempton, "Global Environmentalism: A Challenge to the Postmaterialism Thesis?" *Social Science Quarterly 75*, no. 2 (June 1994): 245–266.

Scatter Diagram Example

The actual data from the previous scatterplot…

Table 8.2 **GNP per Capita Recorded for Eleven Countries in 1992 (in $1,000)**

Country	GNP per Capita
Denmark	20.0
Norway	22.0
Korea	4.4
Switzerland	30.3
Chile	2.0
Canada	19.0
Ireland	8.0
Turkey	1.4
Russia	3.6
Japan	24.0
Philippines	0.7

$$\overline{X} = \frac{\sum X}{N} = \frac{135.4}{11} = 12.31$$

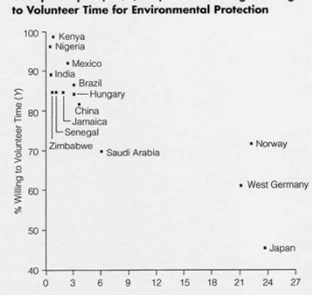
$$\text{Variance } X = S_X^2 = \frac{\sum (X - \overline{X})^2}{N-1} = \frac{1{,}175.3}{10} = 117.52$$

Range X = \$30.3 – \$0.7 = \$29.6

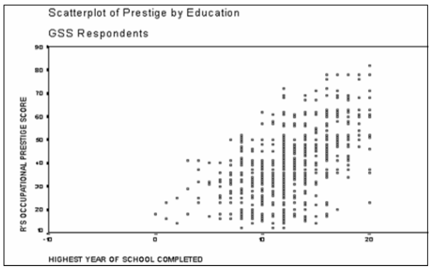
Source: Adapted from Steven R. Brechin and Willett Kempton, "Global Environmentalism: A Challenge to the Postmaterialism Thesis?" *Social Science Quarterly 75*, no. 2 (June 1994): 245–266.

A Scatter Diagram Example of a Negative Relationship

Figure 8.2 **GNP per Capita (in $1,000) and Percentage Willing to Volunteer Time for Environmental Protection**



How would you describe this relationship between occupational prestige and education?

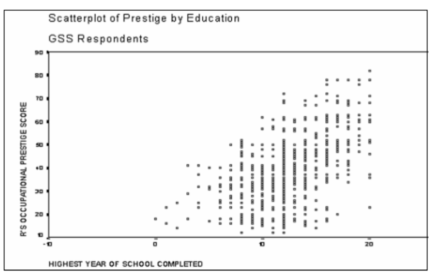


Based on the scatterplot for Education & Prestige

Does education have a positive or negative effect on occupational prestige?

Positive, when we examine a scatterplot of these two variables, we see that as the respondent's level of education increases, the respondent's occupational prestige score also increases (for more cases than not).

Scatterplot of Prestige by Education



Example: Education & Prestige

- The previous scatter diagram can be **represented by a straight line** since it generally appears that, as years of occupational prestige goes up education increases.
- In addition, because occupational prestige goes up when years of education goes up, we can say that the **relationship is a positive one**.

Take your best guess?

If you know nothing else about a person, except that he or she has a job and I asked you to guess the prestige score for his or her occupation, what would you guess?

The <u>mean</u> prestige score for occupations.

Take your best guess?

Now if I tell you that this person has a PhD, would you change your guess?

With quantitative analyses we are generally **trying to predict** or take our best guess at the value of the dependent variable.

One way to assess the relationship between two variables is to consider the degree to which the **extra information** of the second variable **makes your guess better**.

Take your best guess?

By creating a scatter diagram, we can **draw the most accurate line possible** through the data points and use this regression line to help us predict the dependent variable for any value of the independent variable.

So, we can make a much better guess at someone's **occupational prestige**, if we have information about her/his years or level of education and use this information to **create a regression line.**

The Regression Line

The Regression line allows us to reduce the amount of error that would be made if we tried to predict scores with no help from an independent variable.

The higher the association between the independent variable and the dependent variable the more error that is reduced by creating a regression line between the two.

Linear Relationships

- **Linear relationship** – A relationship between two interval-ratio variables in which the observations displayed in a scatter diagram can be approximated with a straight line.
- **Deterministic** (*perfect*) **linear relationship** – A relationship between two interval-ratio variables in which all the observations (the dots) fall along a straight line. The line provides a predicted value of Y (the vertical axis) for any value of X (the horizontal axis).

In class assignment: Graph the data below and examine the relationship (similar idea as the education and prestige graph):

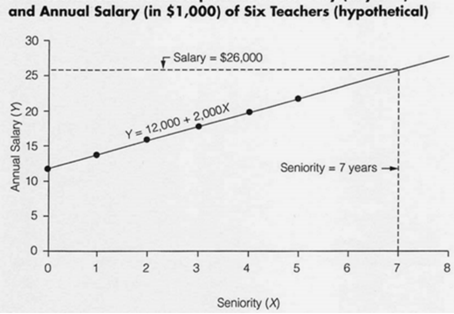
Table 8.3 **Seniority and Salary of Six Teachers (hypothetical data)**

Seniority (in years) X	Salary (in dollars) Y
0	12,000
1	14,000
2	16,000
3	18,000
4	20,000
5	22,000

Draw a straight line through the dots. If you were asked to predict the salary of a teacher with 6 years of seniority what would you guess? How about 7 years? What is the substantive explanation for this relationship?

The Seniority-Salary Relationship

Figure 8.5 **A Perfect Linear Relationship Between Seniority (in years) and Annual Salary (in $1,000) of Six Teachers (hypothetical)**



Take your best guess?

How can we draw a regression line that provides the best prediction of the dependent variable?

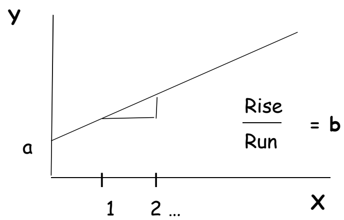
There is an equation that we can follow that will help us to draw the regression line.

The equation **allows us to determine the specific line where the least error occurs**.

Equation for a Straight Line

Y= a + bX

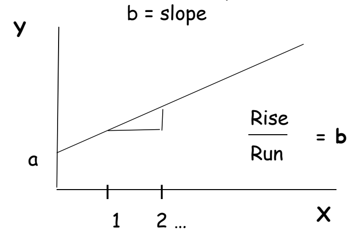
where: Y = dependent variable
X = independent variable



Equation for a Straight Line

Y= a + bX

where: Y = dependent variable
X = independent variable
a = intercept
b = slope



Bivariate Linear Regression Equation

$\hat{Y} = a + bX$

- **Y-intercept (a)**—The point where the regression line crosses the Y-axis, or the value of Y when X=0.
- **Slope (b)**—The change in variable Y (the dependent variable) with a unit change in X (the independent variable.)

Bivariate Linear Regression Equation

$\hat{Y} = a + bX$

The **slope ("b")** is determined by identifying that line where the **sum of the distances** between the line and each case is at a minimum **(that is, the sum of the errors are least)**

Bivariate Linear Regression Equation

$\hat{Y} = a + bX$

The formula for the **slope ("b")** includes the variance of X (squared errors from the mean).

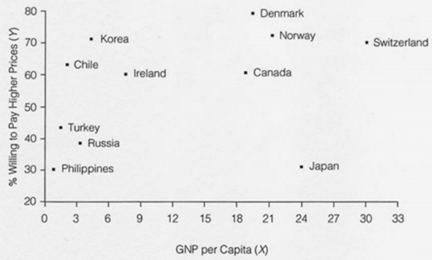
The regression line (or best fitting line) is the one where the sum of all the squared errors is the least possible (least squares line).

Calculating the Regression Line: Using the Least Squares Method

- **Least-squares line** (also referred to as the *regression line* and the *best fitting line*) – A line where the errors are at a minimum.
- **Least-squares method** – The technique that produces the least squares line (you will not be responsible for using this method to calculate the least squares line. Just be aware that the method is based on identifying the line where there is the least amount of error between the line and each case.

Back to the Scatterplot of GNP Per Capita and Willingness to Pay:

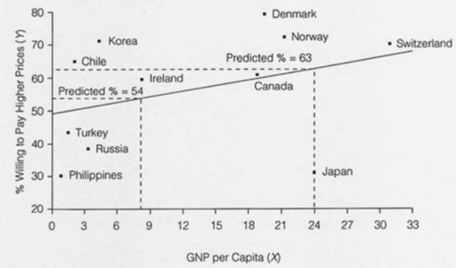
Figure 8.1 **Scatter Diagram of GNP per Capita (in $1,000) and Percentage Willing to Pay More to Protect the Environment**



Source: Adapted from Steven R. Brechin and Willett Kempton, "Global Environmentalism: A Challenge to the Postmaterialism Thesis?" *Social Science Quarterly* 75, no. 2 (June 1994): 245–266.

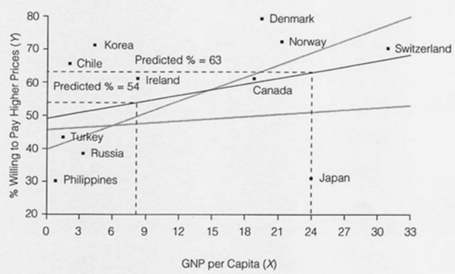
A (best-fit) Regression Line

Figure 8.3 **A Straight-Line Graph for GNP per Capita (in $1,000) and Percentage Willing to Pay More to Protect the Environment**



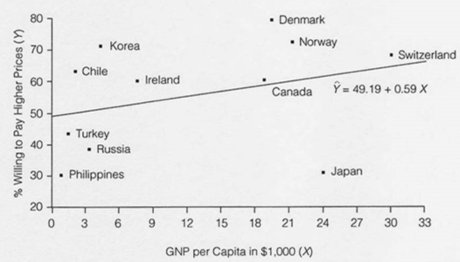
Other Possible Regression Lines

Figure 8.4 **Alternative Straight-Line Graphs for GNP per Capita (in $1,000) and Percentage Willing to Pay More to Protect the Environment**



The Least Squares (error) Line!

Figure 8.6 **The Best-Fitting Line for GNP per Capita and Percentage Willing to Pay More to Protect the Environment**



Summary: Properties of the Regression Line

- **Represents the predicted values for Y** for any and all values of X.
- It is the **best fitting line** in that it minimizes the error (sum of the squared errors or deviations).
- Has a **slope that can be positive or negative**; null hypothesis is that the slope is zero.
- Provides us with two statistics: the **coefficient of determination (r^2)** and the **correlation coefficient (r)**.

Interpreting the Coefficient of Determination

- **R^2** tells us how accurate a prediction our regression equation provides.
- By examining the regression line we can determine the **amount of error that is reduced** (the higher the R^2 the more error is reduced).
- The extent to which prediction error is reduced by taking into account one or more independent variables.

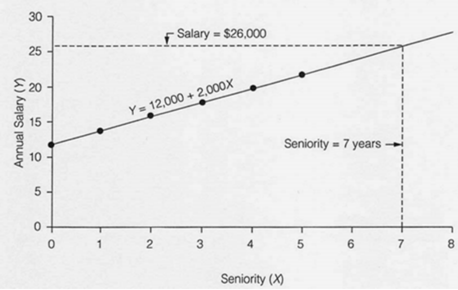
Interpreting the Coefficient of Determination

The r^2 is a **PRE measure** reflecting the proportional reduction of error that results from using the linear regression model.

Still another way of viewing r^2 is to say that it reflects the **proportion of the total variation (or change)** in the dependent variable, **explained** by the independent variable.

The Seniority-Salary Relationship (Coefficient of Determination = 1.0)

Figure 8.5 **A Perfect Linear Relationship Between Seniority (in years) and Annual Salary (in $1,000) of Six Teachers (hypothetical)**

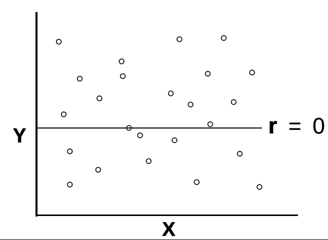


Interpreting Pearson's Correlation Coefficient (r)

- It is a **measure of association** between two interval-ratio variables. The square root of r^2.
- **Symmetrical measure**—No specification of independent or dependent variables.
- **Ranges from –1.0 to +1.0**. The sign (±) indicates direction. The closer the number is to ±1.0 the stronger the association between X and Y.

For example: Here is the Correlation Coefficient for two variables with no association

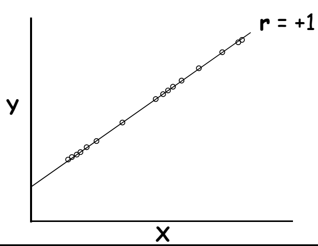
r = 0 means that there is no association between the two variables.



The Correlation Coefficient

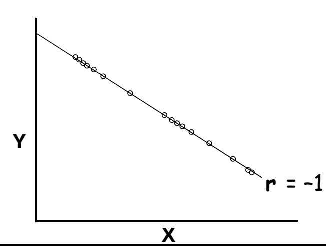
r = 0 means that there is no association between the two variables.

r = +1 means a perfect positive correlation.



The Correlation Coefficient

r = 0 means that there is no association between the two variables.
r = +1 means a perfect positive correlation.
r = –1 means a perfect negative correlation.



The Coefficient of Determination (r^2) and the The Correlation Coefficient (r)

are both determined by calculating the "best fit" regression line (as noted earlier)

SPSS Regression Output: 1998 GSS
Education & Occupational Prestige

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.559[a]	.313	.312	11.90

a. Predictors: (Constant), HIGHEST YEAR OF SCHOOL COMPLETED

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	87633.680	1	87633.680	619.305	.000[a]
	Residual	192727.378	1362	141.503		
	Total	280361.058	1363			

a. Predictors: (Constant), HIGHEST YEAR OF SCHOOL COMPLETED

b. Dependent Variable: RS OCCUPATIONAL PRESTIGE SCORE (1980)

SPSS Regression Output: 1998 GSS
Education & Occupational Prestige

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.120	1.531		3.997	.000
	HIGHEST YEAR OF SCHOOL COMPLETED	2.762	.111	.559	24.886	.000

a. Dependent Variable: RS OCCUPATIONAL PRESTIGE SCORE (1980)

Now let's interpret the SPSS output...

The Regression Equation

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.120	1.531		3.997	.000
	HIGHEST YEAR OF SCHOOL COMPLETED	2.762	.111	.559	24.886	.000

a. Dependent Variable: RS OCCUPATIONAL PRESTIGE SCORE (1980)

Prediction Equation:

$$\hat{Y} = 6.120 + 2.762(X)$$

This line represents the predicted value for Y when X = 0.

The Regression Equation

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.120	1.531		3.997	.000
	HIGHEST YEAR OF SCHOOL COMPLETED	2.762	.111	.559	24.886	.000

a. Dependent Variable: RS OCCUPATIONAL PRESTIGE SCORE (1980)

Prediction Equation:

$$\hat{Y} = 6.120 + 2.762(X)$$

This line represents the predicted value change for Y for a one unit change of X

Interpreting the regression equation

$$Y = a + bX$$

$$\hat{Y} = 6.120 + 2.762(X)$$

- If a respondent had zero years of schooling (if X = 0), this model predicts that his occupational prestige score (Y) would be 6.120 points.
- For each additional year of education, our model predicts a 2.762 point increase in occupational prestige.

www.ruf.rice.edu/~lane/stat_sim/reg_by_eye

Will provide simulations for regression

- guess your own reg. line
- Notice amount of error reduced
- Guess the size of the r (correlation coefficient)

Estimating the slope: b

- The bivariate regression coefficient or the ***slope*** of the regression line can be obtained from the observed X and Y scores (you don't need to learn this).

$$b = \frac{S_{YX}}{S_X^2} = \frac{\dfrac{\sum (X-\overline{X})(Y-\overline{Y})}{N-1}}{\dfrac{\sum (X-\overline{X})^2}{N-1}} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (X-\overline{X})^2}$$