Bivariate Regression and Correlation

Scatter Diagram

- A scatter diagram or scatterplot is a quick visual method used to display a relationship between two interval-ratio variables.
- It is often used as a first step in regression analysis.
- It can suggest whether two variables are associated.

A Scatterplot

- Typically, in a scatterplot the independent variable, (X) is arranged along the horizontal axis and the dependent variable, (Y) is arranged along the vertical axis.
- Example, let us look at *GNP per capita and percentage willing to pay more to protect the environment.

Scatter Diagram Example

- The data…

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

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

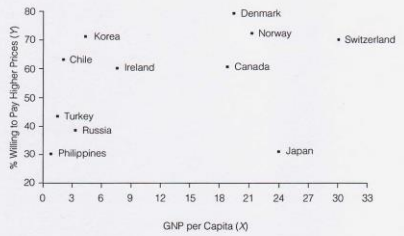
Variance $X = S_X^2 = \frac{\sum (X - \bar{X})^2}{N - 1} = \frac{1,175.3}{10} = 117.52$

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. Copyright © 1994 by the University of Texas Press. All rights reserved.

Scatter Diagram Example

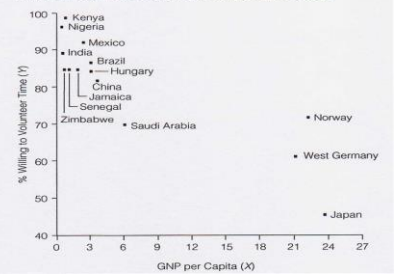
Figure 8.1 **Scatter Diagram of GNP per Capita (in $1,000) and Percentage Willing to Pay More to Protect the Environment**



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. Copyright © 1994 by the University of Texas Press. All rights reserved.

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**



Linear Relationships

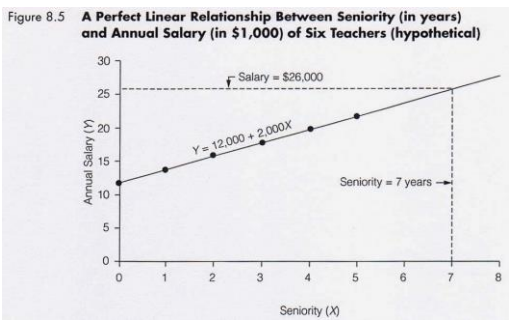
- **Linear relationship** – A relationship between two interval-ratio variables in which the observations displayed in a scatter diagram can be approximated with a straight line.
- **Deterministic** (*perfect*) **linear relationship** – A relationship between two interval-ratio variables in which all 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.

Seniority and Salary of Six Teachers

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

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?

If you know nothing else about a person, except that he or she lives in United States and I asked you to his or her age, what would you guess?

The mean age for U.S. residents.

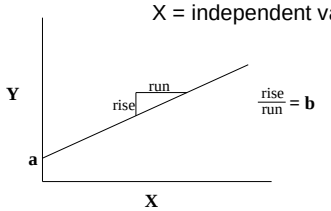
Now if I tell you that this person owns a skateboard, would you change your guess? (Of course!)

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. If someone owns a skateboard, that is likely to indicate to us that s/he is younger and we may be able to guess closer to the actual value.

Equation for a Straight Line

Y= a + bX

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



Straight Line Graphs

- This graph allows us to obtain a predicted salary value for any value of seniority level simply by using a formula.

Y = a + bX

Y = Salary (dependent variable)
X = Seniority (independent variable)
a = the Y intercept
b = the slope

Finding the Best-Fitting Line

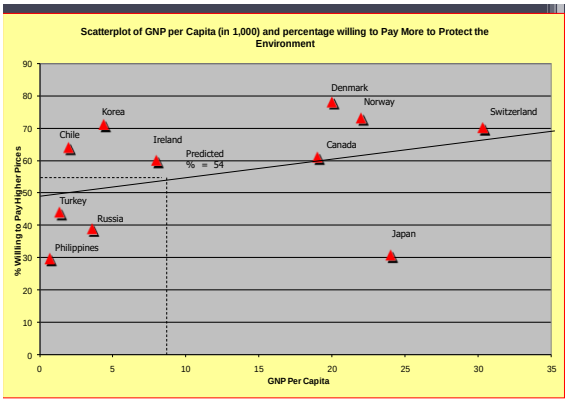
- Unfortunately, most relationships we study in the social sciences are not deterministic.
- In reality teachers salaries are not completely determined by seniority.
- When the dependent variable (Y) is not completely determined by the independent variable (X) not all the observations will lie exactly on the line.

Finding the Best-Fitting Line

- Given that none of the lines are perfect, our task is to choose one line – the best-fitting line.
- The best-fitting line is the one that produces the least amount of error.

Defining Error

- Looking at the next graph, the line predicts a value of Y. For example, for Ireland, with a GNP of $8,000 gives us a predicted value for Y of 54%, but Irelands actual value is 60%.
- Thus we have two values for Y.



Defining Error

- The two values of Y:

1. A predicted Y that is symbolized as $\hat{Y}$ and which is generated by the predication equation, also called the linear regression equation.
2. The observed Y that is symbolized as Y.

- Error is the difference between the two Y's.

The Sum of Squared Errors

- We want a line that will minimize error for each observation. However, any line we choose will minimize the error for some observations and maximize errors for others.
- We want to find a prediction equation that minimizes errors over all observations.

The Sum of Squared Errors

- There are many mathematical ways of defining errors.
- Statisticians prefer to use the sum of the absolute errors or the sum of squared errors.
- We simply square and sum the errors over all observations.

$$\sum e^2 = \sum (Y - \hat{Y})^2$$

The Least-Squares Line

- The best fitting regression line is that line where the sum of the squared errors, or $\sum e^2$ is at a minimum.
- This line is called the least-squares line and the technique that produces this line is called the least-squares method.
- The technique involves choosing *a* and *b* for the equation $\hat{Y} = a+bX$ such that $\sum e^2$ will have the smallest possible value.

Prediction Equation

- Computing *a* and *b* for the prediction equation.
- To find the values of a and b that minimizes the sum of squared errors.

$$b = \frac{S_{yx}}{S_x^2}$$

$$a = \overline{Y} - b(\overline{X})$$

How to Figure the Variance and Covariance

- The variance = $S^2_x = \frac{\Sigma (X-\bar{X})^2}{N-1}$
- The Covariance (X,Y) = $S_{yx} = \frac{\Sigma (X-\bar{X})(Y-\bar{Y})}{N-1}$

Covariance

- The covariance is a measure of how X and Y vary together.
- Basically, the covariance tells us to what extent one variable goes together with the second variable.
- The value reflects both strength and the direction of the relationship.

GNP per Capita (in 1,000) and Percentage Willing to Pay More to Protect the Environment

Country	GNP per Capital (X)	% Willing to Pay (Y)
Denmark	20.0	78%
Norway	22.0	73%
Korea	4.4	71%

Worksheet for Calculating a and b for the Regression Equation

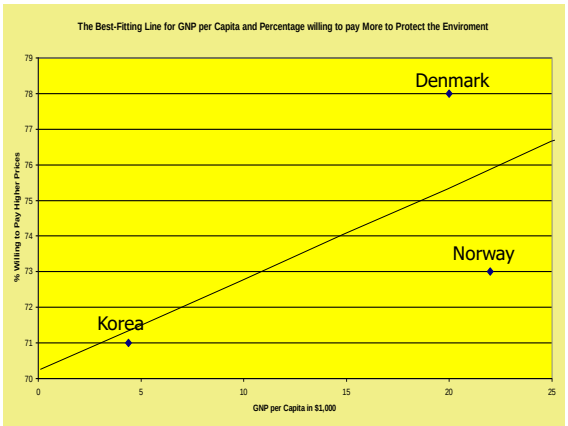
Country	GNP Capita (X)	% willing to Pay (Y)	$(X-\bar{X})$	$(X-\bar{X})^2$	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
Denmark	20.0	78%					
Norway	22.0	73%					
Korea	4.4	71%					
	ΣX = 46.4	ΣY= 222					

Let's Calculate r using the formula for r

Country	GNP Capita (X)	% willing to Pay (Y)	$(X-\bar{X})$	$(X-\bar{X})^2$	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
Denmark	20.0	78%	4.5	20.25	4	16	18
Norway	22.0	73%	6.5	42.25	-1	1	-6.5
Korea	4.4	71%	- 11.1	123.21	-3	9	33.3
	ΣX = 46.4	ΣY= 222	0.00	185.71	0.00	26	44.8

Plotting a Straight line

- Now we can plot the straight line corresponding to the regression equation.
- How do you think we would do this?



Summary: Properties of the Regression Line

- Represents the *predicted* values for Y for any and all values of X.
- Always goes through the point corresponding to the mean of both X and Y.
- It is the *best fitting line* in that it minimizes the sum of the squared deviations.
- Has a slope that can be positive or negative; null hypothesis is that the slope is zero.

The Coefficient of Determination or r^2

- r^2 measures the proportional reduction of error that results from using the linear regression model.
- It represents the proportion of the total variation in the dependent variable Y, explained by the independent variable X.
- r^2 ranges form 0.0 to 1.00.

Calculating r^2

- The following equation is used for figuring r^2

$$r^2 = \frac{[\text{Covariance } (X,Y)]^2}{[\text{Variance } (X)][\text{Variance } (Y)]}$$

OR

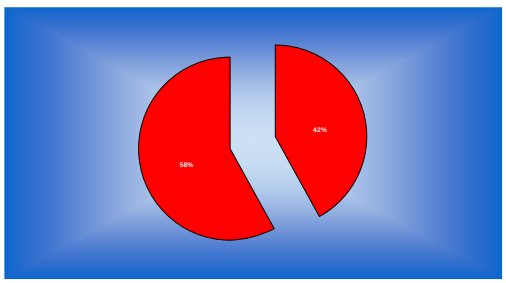
$$r^2 = \frac{S^2_{xy}}{S^2_x S^2_y}$$

Figure r^2 for Our Three Countries

- Covariance = 22.4
- Variance (X) = 92.855
- Variance (Y) = ?

$$r^2 = \frac{(22.4)^2}{(92.855)(?)}$$

A Pie Graph Approach to r^2



The Correlation Coefficient

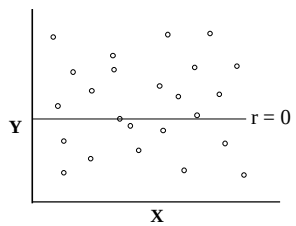
- **Pearson's Correlation Coefficient (r)** — The square root of r^2. It is a measure of association between two interval-ratio variables.

$$r = \frac{\text{cov}(X,Y)}{[s.d.(X)][s.d.(Y)]}$$

- 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.

The Correlation Coefficient

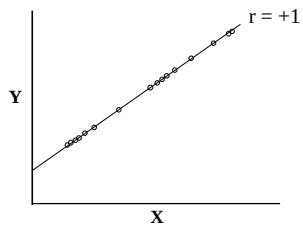
r = 0 means that there is no association between the two variables.



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.