Chapter 4a:

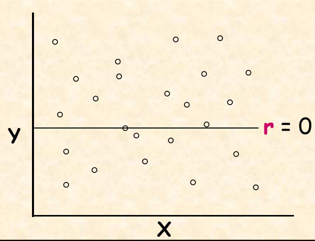
Bivariate Correlation: Review

Review of Terms

Covariation	a relationship or association between two variables
Pearson Product-moment coefficient	*r* – measure of association, indexes the extent to which a linear relationship exists between 2 variables
Perfect Positive Bivariate Correlation	*r* = 1.0, all the points on a scatterplot fall on a straight line

Does this show a perfect linear relationship?

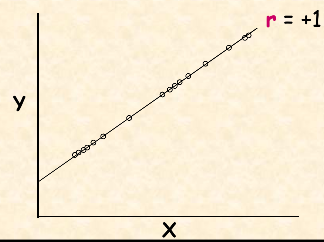
r = 0 means that there is no association between the two variables.



Does this show a perfect linear relationship?

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

r = +1 means a perfect positive correlation.



Review of Terms

Non-perfect bivariate correlation	points on a scatterplot fall around a straight line
Statistical significance	tells us how confident we can be (the probability) that an obtained correlation is different from zero
Substantive significance	tells us the importance of the correlation found (a relatively small correlation could be very important)

Review

Statistical significance and sample size	the larger the sample the more likely the significance
Shared variance between two variables	r^2 - the extent to which two or more variables share the same variability (when one goes up or down the other goes up or down) -the amount of variance in the Y variable explained by the X variable

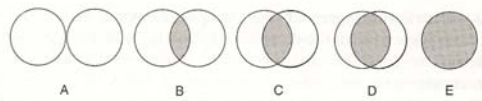


Figure 4a.6 Venn Diagrams Showing a Continuum of Shared Variance From No Shared Variance (A) to Completely Shared Variance (E)

Review

Sample selection & lack of variability

if a sample lacks variability for a variable, covariation that includes the variable will not be possible

Group differentiation problem

if there is a lack of difference between a group's values (eg., all the values in the group score "5" on satisfaction) analyses between the values will show no difference

Review

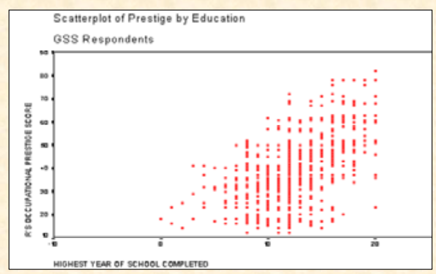
Value of a scatterplot

-provides a visual of the degree of relationship between two variables

-can identify outliers

-can identify curvilinear relationships

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



Review

If two variables have a curvilinear relationship, can we include them in a regression?

-a curvilinear relationship is a quadratic rather than linear relationship.

-squaring a variable will result in measuring the distance between the values and a curved line rather than straight line (cubing a variable works for a relationship with two curves).

How would you describe this relationship?

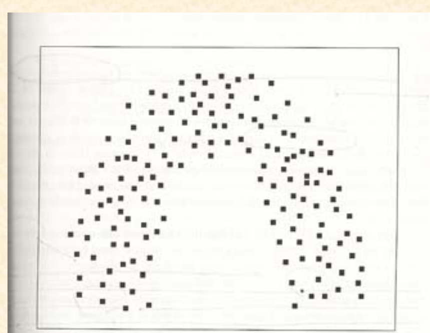


Figure 4a.11 Stylized Curvilinear Relationship

Review

How does the regression process allow us to examine the strength of relationship between two variables?

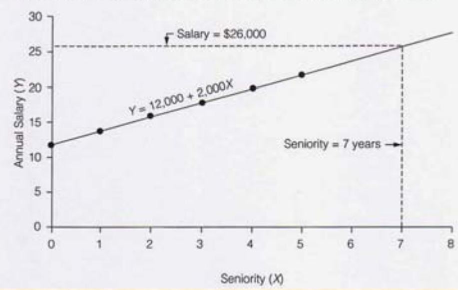
-the dependent variable is plotted on the Y axis and independent variable on X axis

-a best fitting line is calculated and drawn on the graph

-the closer the values are to the line the stronger the relationship between the two variables

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)**



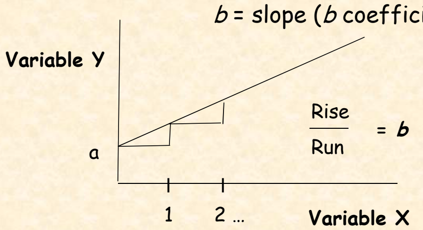
How do we determine the best fitting line?

The best fitting line can be drawn on a set of axes once the Y intercept and the slope are computed.

Equation for a Straight Line

$Y = a + bX$

where: Y = dependent variable
X = independent variable
a = intercept (when x = 0)
b = slope (*b* coefficient)



Bivariate Linear Regression Equation

$\hat{Y} = a + bX$

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

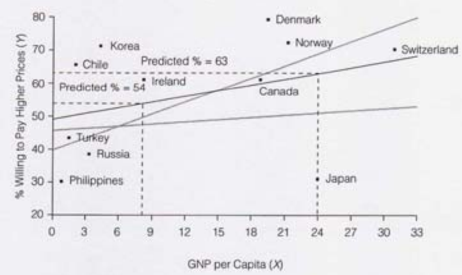
Bivariate Linear Regression Equation

$$\hat{Y} = a + bX$$

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)**

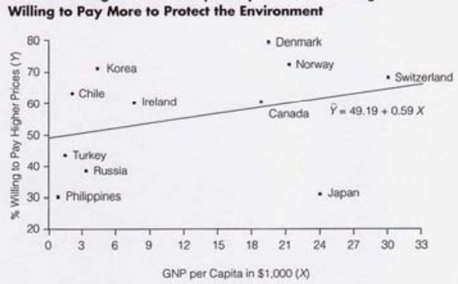
Other Representative Lines

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 (best fitting) Line!

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



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. The method is based on identifying the line where there is the least amount of error between the line and each case.

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. That is, the co-variance divided by the variance or:

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

We can use the data from the Seniority and Salary graph to estimate a "best fitting" line

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

Estimating the $b = \dfrac{\text{sum}(x - \bar{x})(y - \bar{y})}{\text{sum}(x-\bar{x})^2}$

$\text{sum}(x - \bar{x})(y - \bar{y})$		$\text{sum}(\bar{x} - x)^2$	
$(0-2.5)(12-17) = (-2.5)(-5) =$	12.5	$(0-2.5)^2 =$	6.25
$(1-2.5)(14-17) = (-1.5)(-3) =$	4.5	$(1-2.5)^2 =$	2.25
$(2-2.5)(16-17) = (-.5)(-1) =$	.5	$(2-2.5)^2 =$	.25
$(3-2.5)(18-17) = (.5)(1) =$	.5	$(3-2.5)^2 =$	.25
$(4-2.5)(20-17) = (1.5)(3) =$	4.5	$(4-2.5)^2 =$	2.25
$(5-2.5)(22-17) = (2.5)(5) =$	12.5	$(5-2.5)^2 =$	6.25
	35.0		17.50

$$\frac{35.0}{17.5} = 2.0 = b$$

Estimating the Y axis: a

The point at which the regression line crosses the Y axis is determined by:

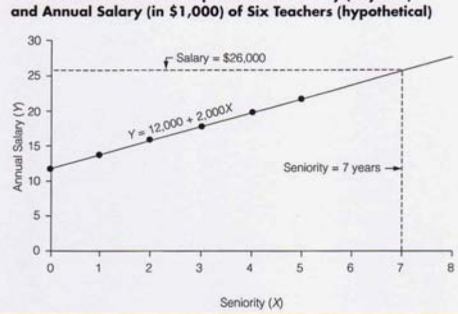
$$a = \bar{Y} - b\bar{X}$$

or

$$a = 17 - (2)(2.5) = 12$$

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)



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 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.

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 Regression Equation

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	$\boxed{6.120}$	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:

$$Y = 6.120 + 2.762(X)$$

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
	$\boxed{\text{HIGHEST YEAR OF SCHOOL COMPLETED}}$	$\boxed{2.762}$	.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 values for Y for any and all values 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 _____ points.
- For each additional year of education, our model predicts a _____ point increase in occupational prestige.

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 _____ point increase in occupational prestige.

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.

Review

What is the difference between the b coefficient and the beta?

-the b coefficient reports the raw score, that is, it is in the units of the dependent variable

-the beta is in standardized form. This allows us to see the effects of one independent variable relative to another.

Go to Chapter 4b and walk through exercises?

-bivariate correlation
-simple linear regression