Contents

11.5 Regression

The regression model is a statistical procedure that allows a researcher to estimate the linear, or straight line, relationship that relates two or more variables. This linear relationship summarizes the amount of change in one variable that is associated with change in another variable or variables. The model can also be tested for statistical significance, to test whether the observed linear relationship could have emerged by chance or not. In this section, the two variable linear regression model is discussed. In a second course in statistical methods, multivariate regression with relationships among several variables, is examined.

The two variable regression model assigns one of the variables the status of an **independent** variable, and the other variable the status of a **dependent** variable. The independent variable may be regarded as causing changes in the dependent variable, or the independent variable may occur prior in time to the dependent variable. It will be seen that the researcher cannot be certain of a causal relationship, even with the regression model. However, if the researcher has reason to make one of the variables an independent variable, then the manner in which this independent variable is associated with changes in the dependent variable can be estimated.

In order to use the regression model, the expression for a straight line is examined first. This is given in the next section. Following this is the formula for determining the regression line from the observed data. Following that, some examples of regression lines, and their interpretation, are given.

11.5.1 Linear Relationships

In the regression model, the independent variable is labelled the X variable, and the dependent variable the Y variable. The relationship between X and Y can be shown on a graph, with the independent variable X along the horizontal axis, and the dependent variable Y along the vertical axis. The aim of the regression model is to determine the straight line relationship that connects X and Y.

The straight line connecting any two variables X and Y can be stated algebraically as

$$Y = a + bX$$

where a is called the Y **intercept**, or simply the intercept, and b is the **slope** of the line. If the intercept and slope for the line can be determined, then this entirely determines the straight line.

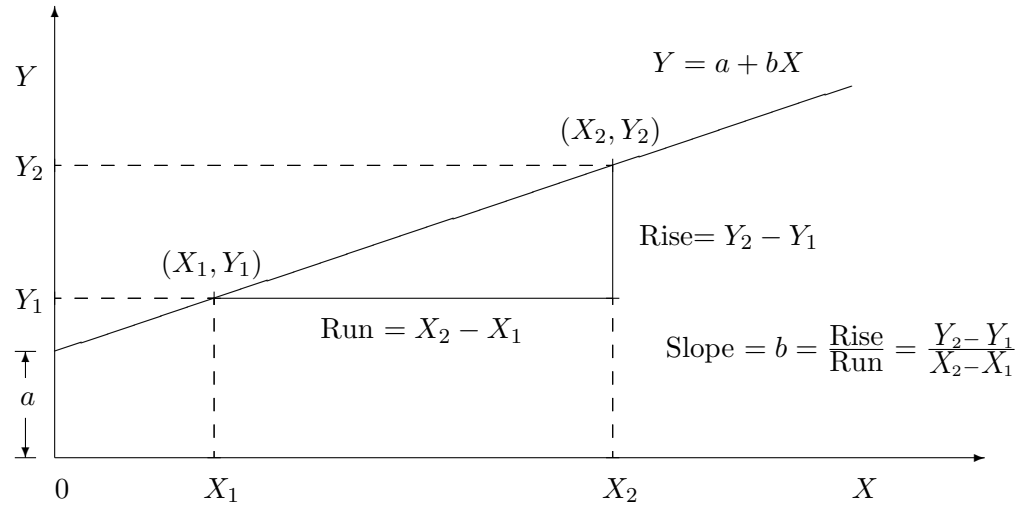


Figure 11.6: Diagrammatic Representation of a Straight Line

Figure 11.6 gives a diagrammatic presentation of a straight line, showing the meaning of the slope and the intercept. The solid line that goes from the lower left to the upper right of the diagram has the equation $Y = a+bX$. The intercept for the line is the point where the line crosses the Y axis. This occurs at $X = 0$, where

$$Y = a + bX = a + b(0) = a + 0 = a$$

and this means that the intercept for the line is a.

The slope of the line is b and this refers to the steepness of the line, whether the line rises sharply, or is fairly flat. Suppose that two points on the line are (X_1, Y_1) and (X_2, Y_2). The horizontal and vertical distances between these two points form the basis for the slope of the line. In order to determine this slope, begin with point (X_1, Y_1), and draw a horizontal line

as far to the right at X_2. This is the solid line that goes to the right from point (X_1, Y_1). Then draw a vertical line going from point (X_2, Y_2) down as far as Y_1. Together these produce the right angled triangle that lies below the line. The base of this triangle is referred to as the *run* and is of distance $X_2 - X_1$. The height of the triangle is called the *rise*, and this height is $Y_2 - Y_1$. The slope of the line is the ratio of the rise to the run. This is

$$\text{Slope of the line} = \frac{\text{rise}}{\text{run}}$$

or

$$\text{Slope} = b = \frac{\text{rise}}{\text{run}} = \frac{Y_2 - Y_1}{X_2 - X_1}.$$

If a line is fairly flat, then the rise is small relative to the run, and the line has a small slope. In the extreme case of a horizontal line, there is no rise, and $b = 0$. When the line is more steeply sloped, then for any given run, the rise is greater so that the slope is a larger number. In the extreme case of a vertical line, there is no run, and the slope is infinitely large.

The slope is negative if the line goes from the upper left to the bottom right. If the line is sloped in this way, $Y_2 < Y_1$ when $X_2 > X_1$.

$$\text{Slope} = b = \frac{Y_2 - Y_1}{X_2 - X_1} < 0.$$

That is, the run has a positive value, and the run has a negative value, making the ratio of the rise to the run a negative number.

Once the slope and the intecept have been determined, then this completely determines the straight line. The line can be extended towards infinity in either direction.

The aim of the regression model is to find a slope and intercept so that the straight line with that slope and intercept fits the points in the scatter diagram as closely as possible. Also note that only two points are necessary to determine a straight line. If only one point is given, then there are many straight lines that could pass through this point, but when two points are given, this uniquely defines the straight line that passes through these two points. The following section shows how a straight line that provides the best fit to the points of the scatter diagram can be found.

11.5.2 The Least Squares Regression Line

Suppose that a researcher decides that variable X is an independent variable that has some influence on a dependent variable Y. This need not imply

that Y is directly caused by X, but the researcher should have some reason for considering X to be the independent variable. It may be that X has occurred before Y or that other researchers have generally found that X influences Y. In doing this, the aim of the researcher is twofold, to attempt to find out whether or not there is a relationship between X and Y, and also to determine the nature of the relationship. If the researcher can show that X and Y have a linear relationship with each other, then the slope of the line relating X and Y gives the researcher a good idea of how much the dependent variable Y changes, for any given change in X.

If there are n observations on each of X and Y, these can be plotted in a scatter diagram, as in Section 11.4.2. The independent variable X is on the horizontal axis, and the dependent variable Y along the vertical axis. Using the scatter diagram, the researcher can observe the scatter of points, and decide whether there is a straight line relationship connecting the two variables. By sight, the researcher can make this judgment, and he or she could also draw the straight line that appears to fit the points the best. This provides a rough and ready way to estimate the regression line. This is not a systematic procedure, and another person examining the same data might produce a different line, making a different judgment concerning whether or not there is a straight line relationship between the two variable.

In order to provide a systematic estimate of the line, statisticians have devised procedures to obtain an estimate of the line that fits the points better than other possible lines. The procedure most commonly used is the **least squares criterion**, and the regression line that results from this is called the **least squares regression line**. While not all steps in the derivation of this line are shown here, the following explanation should provide an intuitive idea of the rationale for the derivation.

Begin with the scatter diagram and the line shown in Figure 11.7. The asterisks in the diagram represent the various combinations of values of X and Y that are observed. It is likely that there are many variables that affect or influence the dependent variable Y. Even if X is the single most important factor that affects Y, these other influences are likely to have different effects on each of the observed values of Y. It is this multiplicity of influences and effects of various factors on Y that produces the observed scatter of points. A single straight line cannot possibly connect all these points. But if there is a strong effect of X on Y, the X and Y values may fall more or less along a straight line. It is this general pattern that the researcher is attempting to find.

The regression line is given the expression

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

where X represents the observed values of the independent variable, and $\hat{Y}$ represents the values of the dependent variable Y that are on the regression line. These are the predicted values of the dependent variable. That is, for each value of X, the predicted values of the dependent variable Y are those that lie on the line. The observed values of Y may or may not lie on the line. Because a straight line cannot pass through all the possible points in a scatter diagram, most of the observed values of Y do not lie on the line. While the line may be useful at predicting values of Y for the various values of X, there will always be errors of prediction. (The $\hat{}$ on top of Y means that these values are estimates of Y. The expression $\hat{Y}$ is called *Y hat*).

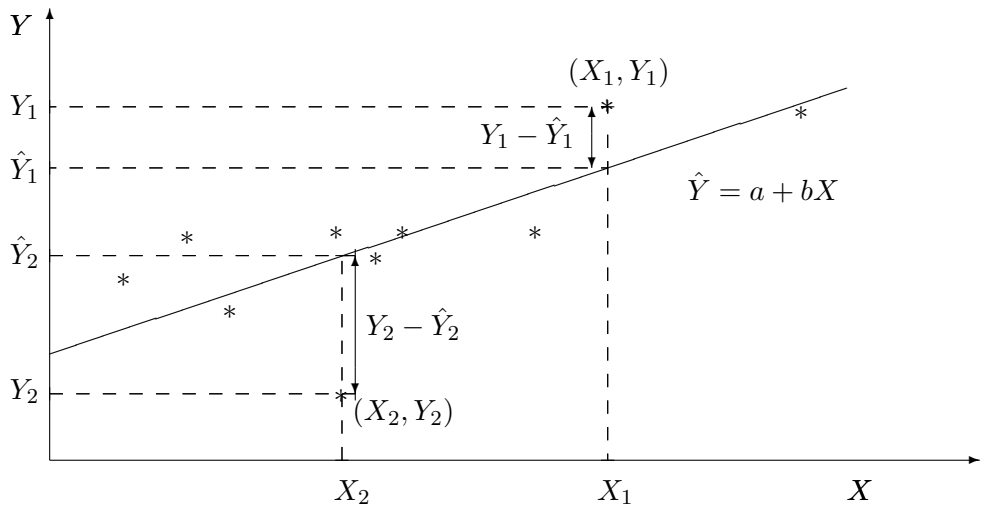


Figure 11.7: Error of Estimate in Regression Line

Now consider point (X_1, Y_1) near the upper right of the diagram. For this point, the observed value of the independent variable is X_1 and the observed value of the dependent variable is Y_1. If the regression line had been used to predict the value of the dependent variable, the value $\hat{Y}_1$ would have been predicted. As can be seen by examining the dashed line that lies at height $\hat{Y}_1$, the point $(X_1, \hat{Y}_1)$ lies on the regression line. This value of the dependent variable was obtained by putting X_1 in the equation, and

$$\hat{Y}_1 = a + bX_1.$$

The error of prediction for X_1 is $Y_1 - \hat{Y}_1$. By entering X_1, the observed value of X, in the equation, it is possible to come close to predicting the value of the dependent variable, but there is always some error of prediction. The aim of the least squares regression line is to minimize these errors of prediction. Let the error of prediction associated with X_1 be e_1, so that

$$e_1 = Y_1 - \hat{Y}_1.$$

Consider the asterisk labelled (X_2, Y_2) near the lower left of Figure 11.7. This point lies a considerable distance from the line, and has error of prediction of $Y_2 - \hat{Y}_2$. That is, when $X = X_2$, the observed value of the dependent variable Y is Y_2 and the predicted value of Y is

$$\hat{Y} = a + bX_2 = \hat{Y}_2.$$

The error of prediction associated with X_2 is

$$e_2 = Y_2 - \hat{Y}_2.$$

A similar error of prediction could be obtained for each of the observed data points.

Now imagine that there could be many possibles line that could be drawn. Each of these has associated with it a set of errors of prediction for the Y values. Some of these lines fit the points of the scatter diagram better than do other lines. Better fitting lines have smaller errors of prediction than do lines that do not fit the points so well. One of the aims of the regression model is to find the line that fits the points best of all. In order to find this line, statisticians use the **least squares criterion**. This criterion involves attempting to minimize the sums of the squares of the errors of prediction. It is this minimization that produces the line that fits the observed points best of all.

The least squares criterion can be written in algebraic form as follows. Suppose there are n observed points (X_i, Y_i), where $i = 1, 2, \ldots, n$. Now consider a line that is drawn in the plane. For each observed X value, a predicted Y value of $\hat{Y}$ can be obtained by putting the X value in the equation for the line. These predicted values of Y will differ from the observed values of Y. For each of the i observations the difference between the observed and predicted value can be written

$$e_i = Y_i - \hat{Y}_i.$$

One criterion for determing the line that fits the points best is that the positive and negative errors cancel out so that

$$\sum e_i = 0.$$

The squares of these error terms are

$$e_i^2 = (Y_i - \hat{Y}_i)^2$$

and the sum of the squares of these errors is

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

The least squares regression line is the straight line that has the minimum possible value for this sum.

It can be shown mathematically that there is only one line that satisfies the criterion

$$\sum e_i = 0$$

and that produces

$$\text{Minimum } \sum e_i^2.$$

It can be shown with some algebra and calculus that this occurs when a and b take the following values:

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

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

This intercept a and slope b are the statistics that produce the line that fits the points the best. All other possible lines result in larger values for the sums of the squares of the errors of prediction, $\sum e_i^2$.

The values of a and b can be computed as shown in the above formulas, but computationally it is more straightforward to use the formulas that were developed when determining the correlation coefficient. In Section 11.4.3, the values S_{XX} and S_{XY} were used. These were defined as

$$S_{XX} = \sum X^2 - \frac{(\sum X)2}{n}$$

$$S_{XY} = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$

These expressions can be calculated from the observed values of X and Y in the same manner as in Section 11.4.3. Based on these expressions, the slope and intercept can be shown to equal

$$b = \frac{S_{XY}}{S_{XX}}$$

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

The steps involved in determining a and b are as follows. First compute $\sum X$, the sum of the X values, and $\bar{X}$, the mean of X. Do the same for Y, computing $\sum Y$ and $\bar{Y}$. Also compute the squares of the X values, and sum these to obtain $\sum X^2$. Then take the products of each X times each Y. The sum of these is $\sum XY$. From the summations $\sum X$, $\sum Y$, $\sum X^2$ and $\sum XY$, compute S_{XX} and S_{XY}. These are then used to obtain a and b. The resulting least squares regression line is written

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

where $\hat{Y}$ is the predicted value of Y.

An example of how a regression line can be obtained is contained in the following example. After that, a test for the statistical significance of the regression line is given.

Example 11.5.1 Regression of Alcohol Consumption on Income for the Provinces of Canada, 1985-86

Alcohol consumption per capita varies considerably across the provinces of Canada, with consumption in the province having the highest level of consumption averaging 50% greater than in the province having the lowest consumption level. There are many variables that might affect alcohol consumption, factors such as different laws, different price levels for alcohol, different types of stores where alcohol can be purchased, and so on. One of the main factors that is likely to affect alcohol consumption is the income of consumers. Economists generally consider alcohol to be a superior good, one whose consumption level increases as incomes rise. This example contains data concerning alcohol consumption in each province of Canada, and income per household for each province. Use the data in Table 11.1 to obtain the regression line relating income and alcohol consumption.

Province	Income	Alcohol
Newfoundland	26.8	8.7
Prince Edward Island	27.1	8.4
Nova Scotia	29.5	8.8
New Brunswick	28.4	7.6
Quebec	30.8	8.9
Ontario	36.4	10.0
Manitoba	30.4	9.7
Saskatchewan	29.8	8.9
Alberta	35.1	11.1
British Columbia	32.5	10.9

Table 11.1: Income and Alcohol Consumption, Provinces of Canada

Table 11.1 gives family income per capita in 1986 for each of the provinces of Canada, and alcohol consumption in litres per person 15 years and older

in 1985-86. The income data comes from Statistics Canada, Economic Families - 1986 Income [machine-readable data file]. 1988 Edition. The data concerning alcohol consumption is taken from Saskatchewan Alcohol and Drug Abuse Commission, **Fast Factsheet** *(Regina, 1988).*

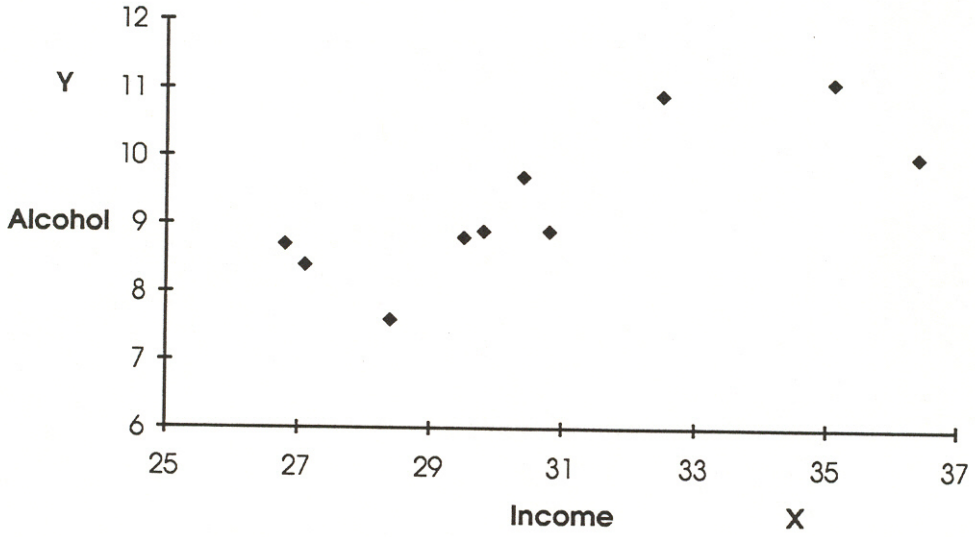
Figure 11.8: Scatter Diagram of Income and Alcohol Consumption

Solution.

The first step in obtaining the regression equation is to decide which of the two variables is the independent variable and which is the dependent variable. The suspicion is that differences in income levels are a factor in explaining differences in alcohol consumption per capita across the provinces. This means that income is being considered the independent variable, affecting alcohol consumption, the dependent variable.

While there are many other factors that are likely to affect alcohol consumption, if different income levels lead to different levels of alcohol consumption, the relationship between the two variables may be apparent in a scatter diagram. In Figure 11.8, the values of income are plotted along the horizontal axis. The values of the dependent variable, alcohol consumption, are plotted on the vertical axis.

Figure 11.8



Based on the scatter diagram of Figure 11.8, there appears to be a definite relationship between income and alcohol consumption. The three provinces with the lowest alcohol consumption are also those that have the lowest income. For the three highest income provinces, alcohol consumption per capita is also the greatest. The other provinces are between these two extremes on both variables. While the relationship is not perfect, it appears that a straight line that starts near the lower left of the diagram, and goes toward the upper right, can show the relationship between the two variables.

The regression equation regresses alcohol consumption on income, that is, income is the independent variable and alcohol consumption is the dependent variable. For the regression, income is represented by X and alcohol consumption by Y. The calculations for the regression are shown in Table 11.2. Note that the fourth column, the squares of Y are not required in order to determine the regression equation. However, these values have been included because they are required in order to calculate the correlation coefficient, and in Section 11.5.4, to conduct an hypothesis test for the statistical significance of the regression line.

X	Y	X^2	Y^2	XY
26.8	8.7	718.24	75.690	233.16
27.1	8.4	734.41	70.560	227.64
29.5	8.8	870.25	77.440	259.60
28.4	7.6	806.56	57.760	215.84
30.8	8.9	948.64	79.210	274.12
36.4	10.0	1324.96	100.000	364.00
30.4	9.7	924.16	94.090	294.88
29.8	8.9	888.04	79.210	265.22
35.1	11.1	1232.01	123.210	389.61
32.5	10.9	1056.25	118.810	354.25
306.8	93.0	9503.52	875.98	2878.32

Table 11.2: Calculations for Regression of Alcohol Consumption on Income

From Table 11.2, the following values are obtained:

$$\sum X = 306.8$$

$$\sum Y = 93.0$$

$$\sum X^2 = 9,503.52$$

$$\sum Y^2 = 875.98$$

$$\sum XY = 2,878.32$$

and these can then be used to determine the following:

$$\begin{aligned} S_{XY} &= \sum XY - \frac{(\sum X)(\sum Y)}{n} \\ &= 2,878.32 - \frac{(306.8)(93.0)}{10} \\ &= 2,878.32 - 2,853.24 \\ &= 25.08 \end{aligned}$$

$$\begin{aligned} S_{XX} &= \sum XX - \frac{(\sum X)^2}{n} \\ &= 9,503.52 - \frac{306.8^2}{10} \\ &= 9,503.52 - 9,412.644 \\ &= 90.896 \end{aligned}$$

$$\begin{aligned} S_{YY} &= \sum Y^2 - \frac{(\sum Y)^2}{n} \\ &= 875.98 - \frac{93.0^2}{10} \\ &= 875.98 - 864.9 \\ &= 11.08 \end{aligned}$$

The value for the slope of the regression line is

$$b = \frac{S_{XY}}{S_{XX}} = \frac{25.08}{90.896} = 0.276$$

The intercept is

$$\begin{aligned} a &= \bar{Y} - b\bar{X} \\ &= \frac{93.0}{10} - 0.276\frac{306.8}{10} \\ &= 9.30 - (0.276 \times 30.68) \\ &= 9.30 - 8.68 = 0.832 \end{aligned}$$

The least squares regression line is thus

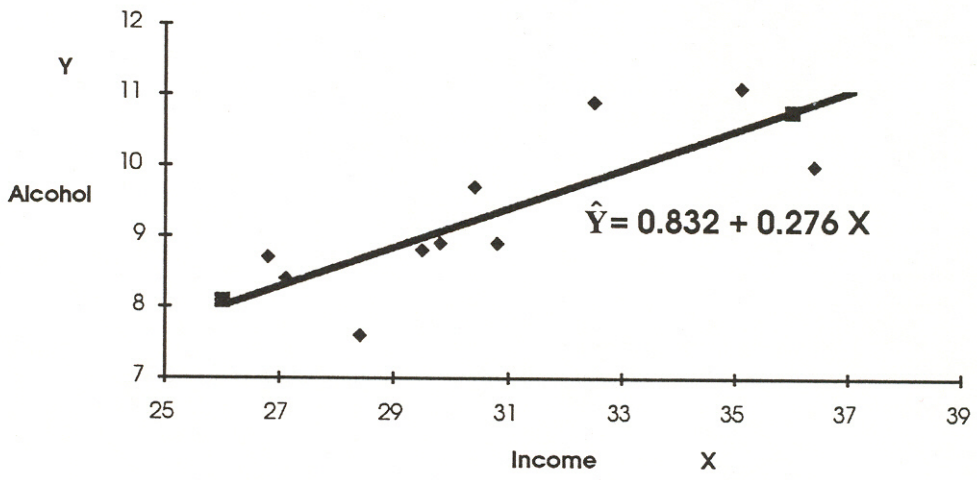
$$\hat{Y} = 0.832 + 0.276X.$$

Figure 11.9 shows the scatter diagram along with the regression line.

Figure 11.9: Scatter Diagram and Regression Line for Regression of Alcohol Consumption on Income

While the intercept $a = 0.832$ has little real meaning, the slope of the line can be interpreted meaningfully. The slope $b = 0.276$ is positive, indicating that as income increases, alcohol consumption also increases. The size of the slope can be interpreted by noting that for each increase of one unit in

Figure 11.9


Y
Alcohol
12
11
10
9
8
7
25
27
29
31
33
35
37
39
Income
X
$\hat{Y}$= 0.832 + 0.276 X

X, the dependent variable Y changes by 0.276 units. For this example, X is in units of one thousand dollars of income, and Y is in units of litres of alcohol consumed annually per capita. This means that each $1,000 increase in family income is associated with an increase of a little over one quarter of a litre of alcohol consumption per capita annually. This relationship has been estimated across the ten provinces of Canada, so that this gives an idea of the manner in which income and alcohol consumption are related across the provinces.

Province	X	Y	$\hat{Y}$	$e_i = Y_i - \hat{Y}_i$
Newfoundland	26.8	8.700	8.229	0.471
PEI	27.1	8.400	8.312	0.088
Nova Scotia	29.5	8.800	8.974	-0.174
New Brunswick	28.4	7.600	8.671	-1.071
Quebec	30.8	8.900	9.333	-0.433
Ontario	36.4	10.000	10.878	-0.878
Manitoba	30.4	9.700	9.223	0.477
Saskatchewan	29.8	8.900	9.057	-0.157
Alberta	35.1	11.100	10.520	0.580
British Columbia	32.5	10.900	9.802	1.098

Table 11.3: Actual and Fitted Values and Prediction Errors

Finally, the equation can be used to determine the predicted values of alcohol consumption, and these can in turn be used to obtain the prediction errors. These are not usually calculated in the detail shown here. The errors are contained in Table 11.3. For example, for Newfoundland, income per capita is $26,800, so that $X = 26.8$. Putting this value in the equation gives

$$\hat{Y} = 0.832 + 0.276X = 0.832 + (0.276 \times 26.8) = 0.832 + 7.397 = 8.229.$$

That is, the equation predicts that alcohol consumption for Newfoundland is 8.229, while in fact it is 8.700. This means that there is a prediction error of

$$Y - \hat{Y} = 8.700 - 8.229 = 0.471$$

for Newfoundland. The equation underpredicts the level of alcohol consumption for Newfoundland. The errors of prediction for each of the other

provinces can be seen in the table. For all but two of the provinces, the equation comes within one litre per capita of predicting the actual value of alcohol consumption. For only New Brunswick and British Columbia does the equation not predict so well, with alcohol consumption for New Brunswick being overpredicted by 1.1 litres and British Columbia being underpredicted by 1.1 litres. Also note that the sum of the errors of prediction is 0, with the negative errors of prediction being balanced overall by the positive errors of prediction.

11.5.3 Using the Regression Line

There are various ways that the regression line can be used. Table 11.3 used the regression line of alcohol consumption on income to predict values of alcohol consumption for each province. The line could also be used to predict values in other regions where the income per capita is known. These predictions are of two sorts, interpolation and extrapolation. Before discussing these, a short note concerning how to draw the line is given.

Drawing the Line. In order to draw a line in the $X - Y$ plane, it is necessary to obtain only two points on the line. The regression equation $\hat{Y} = a + bX$ is used to do this. Pick two values of X, one value near the lower end of the observed X values, and one near the upper end. These need not be values actually observed, but could be any values of X. Then obtain the predicted Y values for these X values. The resulting combinations of X and Y give two points in the plane, and these are joined to produce the line.

In Figure 11.9, the values $X = 26$ and $X = 36$ are the two values selected. The predicted values of Y for these values of X are

$$\hat{Y} = 0.832 + 0.276X = 0.832 + (0.276 \times 26) = 0.832 + 7.176 = 8.088$$

$$\hat{Y} = 0.832 + 0.276X = 0.832 + (0.276 \times 36) = 0.832 + 9.936 = 10.768$$

These two values are shown as the squares in Figure 11.9

Interpolation. When the value for a particular X is within the range of X values used to determine the regression line, the use of this X to predict a Y value is referred to as interpolation. In Example 11.5.1, suppose it is known that a particular city has a per capita income of 31 thousand dollars.

Then the line can be used to predict the level of alcohol consumption in this city. Putting $X = 31$ into the equation for the regression line gives

$$\hat{Y} = 0.832 + 0.276X = 0.832 + (0.276 \times 31) = 0.832 + 8.556 = 9.388$$

The predicted alcohol consumption level for this city would be 9.4 litres per capita annually.

When using the regression equation to predict values of Y, it is unlikely that the prediction will be a perfectly accurate prediction of the dependent variable Y. There are many factors that are likely to affect the value of Y, and the independent variable X is only one of these. Even if the regression line fits the points of the scatter diagram quite closely, there will be errors of prediction as a result of random variation, or the effect on the variable of these other factors.

For example, suppose that the regression line of Example 11.5.1 is used to predict alcohol consumption for a city in the Yukon territory or in Alaska. If the per capita income for this city in the Yukon territory or Alaska is known, and this value falls between 26 and 36 thousand dollars, then the equation can be used to predict alchol consumption. But for this city, the equation may not provide a very close estimate of alcohol consumption per capita. The Yukon may have a different pattern of alcohol consumption than does the rest of Canada. In the case of Alaska, the city is not even in Canada, and there may be different laws and quite different patterns of alcohol consumption.

Extrapolation. The regression line can be extended indefinitely in either direction. The dependent variable Y can be predicted for values of X that lie considerably outside the range of the values over which the straight line was estimated. In Example 11.5.1, the 10 provinces of Canada had incomes that ranged between approximately 26 and 36 thousand dollars. Over this range of incomes, the relationship between income and alcohol consumption appears to be linear. But outside this range, the relationship may no longer be linear. While values of X outside this range may be used in the equation, extrapolation of this sort may lead to considerable errors of prediction.

Suppose that a city has a per capita income of 50 thousand dollars. Using the regression line from Example 11.5.1, the predicted value of alcohol consumption for this city is

$$\hat{Y} = 0.832 + 0.276X = 0.832 + (0.276 \times 50) = 0.832 + 13.8 = 14.632.$$

The equation predicts alcohol consumption of 14.6 litres per capita for this city, well above the level of alcohol consumption for any province shown. While this may well be close to the level of alcohol consumption for this city, it seems unlikely that alcohol consumption would be this high. It may be that the relationship between income and alcohol consumption levels off after reaching a certain income level, and reaches an upper limit. The straight line relationship may turn into a curved relationship beyond a certain point. Without observations on units that have larger values of X, the researcher has no way of determining whether this is the case or not.

When extrapolating, the researcher must be quite careful to realize that the prediction error may be quite considerable for values of X that lie outside the range of values over which the straight line relationship has been obtained. If this caution is taken, extrapolation can provide a useful approach, allowing the researcher to examine the linear effect of X on Y for values outside the observed range.

Changes in X and Y. The slope of the regression line can be used to estimate the effect on Y of a change in the values of X. Recall that the slope b is

$$b = \frac{\text{Change in Y}}{\text{Change in X}}.$$

This expression can be rearranged so that

$$\text{Change in Y} = b \times \text{Change in X}.$$

Using the slope for the regression line of Example 11.5.1, an increase of 10 thousand dollars in per capita income could be expected to increase alcohol consumption by

$$\text{Change in Y} = 0.276 \times 10 = 2.76$$

or 2.8 litres of alcohol per capita annually.

An estimate of the change in Y obtained from the regression line is one of the reasons regression may be preferred to correlation when investigating the relationship between two variables. While a correlation coefficient tells whether two variables are related or not, the slope of the regression line tells how the two variables are related. If the researcher has a reason for making one of the variables an independent, and the other a dependent variable, then the slope of the estimated regression line provides an estimate of how one variable might change as the other variable is changed. This estimate is, of course, subject to some errors, and is an estimate of the average relationship. For any specific set of two values of X, the relationship may not hold.

11.5.4 Hypothesis Test for the Line

Just as there is a test to determine whether the correlation coefficient is sufficiently different from 0 to provide evidence for a relationship between two variables, there is also a test for the slope of the regression line. In order to test whether the line can be regarded as showing a slope that is different than 0, it is first necessary to introduce some other concepts concerning the regression line. These are the standard error of estimate, and the idea of the true regression line.

Parameters for the True Regression Line. The values of a and b obtained when estimating the regression line are statistics, obtained by using values of the observed data. This results in a particular regression line, one that fits the observed data points better than any other regression line. It is possible to imagine that there is a true regression line, where all members of the population are observed, and where there are no measurement or data production errors. Let this regression line have the intercept α, the Greek *alpha*, and the slope β, the Greek *beta*. The true regression line is then

$$\hat{Y} = \alpha + \beta X + \epsilon$$

The slope β and the intercept α are parameters, and their point estimators are a and b, respectively. The letter ϵ is the Greek letter *epsilon*. This is used to denote the error or unexplained part of Y. Even if X does have a linear or straight line influence on Y, X by itself cannot explain the exact Y value. The effect of random factors, and other variables that may affect Y in a systematic fashion, are all included as unexplained influences on Y in ϵ.

It is next necessary to imagine many different samples of data. Each sample provides a different set of values of X and Y. For each set of these pairs of X and Y values, there is a different scatter diagram, and a different regression line. Each regression line is a least squares regression line, fitting the observed points better than any other line. But becuase each scatter diagram is different, each regression line differs. This means that each sample has associated with it a value of a and b, defining the best fitting regression line for that sample.

If the samples are random samples, these values of a and b vary in a systematic fashion. The mean of a is α, the true intercept, and the mean of b is β. Each of a and b has a standard deviation. For now, these are defined as s_a for a, with s_b representing the standard deviation of b. In addition,

under certain conditions that can be specified for the error term ϵ, each of the distributions of a and b can be approximated by a t distribution with the means and standard deviations as just given. These t distributions have $n-2$ degrees of freedom, where n is the size of the sample. This can be written as follows:

$$a \text{ is } t_{n-2}(\alpha, s_a)$$

$$b \text{ is } t_{n-2}(\beta, s_b)$$

These distributions can then be used to test the regression line for statistical significance.

For an hypothesis test, the researcher is usually interested only in the slope. The intercept is a necessary component of the regression line, telling where the line will be placed in the vertical direction. But there are relatively few examples of situations where the researcher must test for the significance of the intercept.

The hypothesis test for the slope β of the regression line begins with the null hypothesis that there is no relationship between the variables. Ordinarily, the researcher can anticipate whether the direction of the relationship is positive or negative, so that the alternative hypothesis will be a one directional one, that $\beta > 0$ or that $\beta < 0$. The null and alternative hypotheses are thus

$$H_0 : \beta = 0$$

meaning that there is no relationship between X and Y. If the relationship is expected to be a positive one between X and Y, the alternative hypothesis is

$$H_1 : \beta > 0$$

and if a negative relationship is expected, then

$$H_1 : \beta < 0$$

While it is possible to conduct a two tailed test of the form $\beta \neq 0$, this is not commonly done with regression, because the researcher is interested in the direction of the relationship.

The test statistic and its distribution are given above, and the standardized t statistic is the variable b minus its mean β, divided by its standard deviation s_b. This is

$$t = \frac{b - \beta}{s_b}$$

and the t statistic has $n-2$ degrees of freedom. The researcher selects a significance level, and if the t statistic falls in the region of rejection of the null hypothesis, then the assumption of no relationship is rejected, and the alternative hypothesis accepted. The formula for determing s_b is given in the next section. Here an example of the hypothesis test for the relationship between alcohol consumption and income is given.

Example 11.5.2 Hypothesis Test for the Relationship between Alcohol Consumption and Income

In Example 11.5.1, X represents per capita income, and Y represents alcohol consumption per capita. Let α be the true intercept and β be the true slope of the regression line relating income and alcohol consumption. The sample of 10 provinces can be regarded as one means of obtaining data concerning this relationship. For this sample, $n = 10$, $a = 0.832$ and $b = 0.276$, so that the least squares regression line is

$$\hat{Y} = a + bX = 0.832 + 0.276X.$$

When this data is entered into a computer program, the program gives the standard deviation for the slope as $s_b = 0.0756$. This can be used to test for significance as follows.

The null hypothesis is that there is no relationship between family income and alcohol consumption per capita. If this is the case, then $\beta = 0$. The researcher suspects that alcohol consumption per capita is positively related to income. If this is the case, then $\beta > 0$. The value of b obtained from the sample is $b = 0.276$, and while this is positive, the question is whether this is enough greater than 0, to reject the null hypothesis and accept the alternative hypothesis.

The null and alternative hypotheses are

$$H_0 : \beta = 0$$

$$H_1 : \beta > 0$$

The test statistic is the sample slope b and this has a t distribution with mean β, standard deviation $s_b = 0.0756$, and $n - 2 = 10 - 2 = 8$ degrees of freedom. Pick the 0.01 level of significance, and for a one tailed test, $t_{0.01;8} = 2.897$. The region of rejection of H_0 is all t values of 2.897 or more. If $t < 2.897$, then the null hypothesis cannot be rejected.

From the sample

$$\begin{aligned} t &= \frac{b-\beta}{s_b} \\ &= \frac{0.276-0}{0.0756} \\ &= 3.651 > 2.897 \end{aligned}$$

This means that the slope and the t value are in the region of rejection for the null hypothesis. At the 0.01 level of significance, there is evidence that the relationship between income and alcohol consumption is indeed a positive one. That is, under the assumption of no relationship between the two variables, the probability of obtaining a value of b of 0.276 or larger, is less than 0.01. Since this is quite a small probability, most researchers would conclude that this provides quite strong evidence that there is a positive relationship between income and alcohol consumption.

11.5.5 Goodness of Fit

Another way of examining whether the regression line shows a relationship between the two variables is to determine how well the line fits the observed data points. This was examined earlier, at the end of Example 11.5.1, where the errors of prediction $Y - \hat{Y}$ were given. But to list all the errors of prediction is an awkward way of examining the fit of the line. Fortunately, these errors of prediction can be summarized into a single statistic called **R squared** and written R^2. This statistic is also called the **goodness of fit** of the regression line. A formula for this statistic is provided next, followed by an explanation of the statistic.

The most straightforward formula for R^2 is based on the values S_{XX}, S_{XY} and S_{XY} calculated earlier in connection with the correlation coefficient and the regression line. Based on these expressions

$$R^2 = \frac{S_{XY}^2}{S_{XX}S_{YY}}.$$

The minimum value for R^2 is 0. This would occur when there is no relationship between the two variables, so that X does not help at all in explaining the differences in values of Y. The maximum possible value for R^2 is 1. This would occur when the two variables are perfectly related, so that the observed values of Y exactly correspond with the predicted values from the regression line, and there are no prediction errors. This would mean a perfect goodness of fit.

By comparing this formula with the formula for the correlation coefficient r, it can be seen that $R^2 = r^2$. However, the goodness of fit has a different interpretation than does the correlation coefficient.

For Example 11.5.1, it was shown that

$$S_{XY} = 25.08$$

$$S_{YY} = 11.08$$

$$S_{XX} = 90.896$$

Based on these,

$$\begin{aligned} R^2 &= \frac{S_{XY}^2}{S_{XX}S_{YY}} \\ &= \frac{25.08^2}{11.08 \times 90.896} \\ &= \frac{629.0064}{1007.12768} \\ &= 0.625 \end{aligned}$$

This shows that the fit of the regression line to the points is fairly good, above one half, but still considerably less than 1. An R^2 of 0.625 means that 0.625 or 62.5% of the variation in the values of Y can be explained on the basis of the regression line. That is, alcohol consumption varies across the provinces of Canada, and income differences among provinces explain 62.5% of these differences in alcohol consumption. This does not necessarily mean that income differences are a cause of different levels of alcohol consumption, although the evidence does point in that direction. The explanation is a statistical one, meaning that 62.5% of the differences among alcohol consumption in the different provinces are explained statistically by differences in income per capital among the provinces. The following paragraphs should assist in explaining this.

The Basis for R^2. The R^2 or goodness of fit statistic can be explained as follows. Being with the expression

$$\sum (Y - \bar{Y})^2.$$

This expression is called the **variation in** Y, and is a means of considering how different from each other the values of Y are. Note that the variation of Y is the same as the numerator of the expression for the variance of Y. The variation in Y tells how varied the values of Y are by taking the deviations about the mean of Y, squaring these deviations, and adding them. If the values of Y are all concentrated around the mean $\bar{Y}$, then the variation is a small number, but if the values of Y are more spread out, then the variation is a larger number. The aim of the goodness of fit statistic is to tell what proportion of this variation can be explained statistically on the basis of the independent variable X.

Note that $Y - \bar{Y}$ can be written

$$Y - \bar{Y} = Y - \hat{Y} + \hat{Y} - \bar{Y} = (Y - \hat{Y}) + (\hat{Y} - \bar{Y})$$

The first part of the above expression, $(Y - \hat{Y})$ is the same as the error of prediction, encountered earlier. This can be regarded as the unexplained portion of the deviation of Y about the mean. This is a measure of the extent to which factors other than X cause the value of Y to differ from what is predicted from the regression line. The latter part of the expression $(\hat{Y} - \bar{Y})$, can be considered to be the explained portion of the deviation of Y about the mean. This is explained in the sense that the regression line predicts $\hat{Y}$, so that the difference $\hat{Y} - \bar{Y}$ is expected on the basis of the regression line.

Since the goodness of fit is based on the variation, it is necessary to square the above expression for $Y - \bar{Y}$, and sum these squares. The squares of these deviations about the mean, and the sums of these squares are

$$(Y - \bar{Y})^2 = (Y - \hat{Y})^2 + (\hat{Y} - \bar{Y})^2 + 2(Y - \hat{Y})(\hat{Y} - \bar{Y})$$

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

The latter expression seems rather intimidating, but it turns out that it can be simplified because it can be shown that the last summation equals zero. The sum

$$\sum (Y - \hat{Y})(\hat{Y} - \bar{Y}) = 0$$

and this then means that

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

and it should be possible to make sense out of this last expression. On the left is the variation in Y, often called the **total variation**, or the **total sum of squares**. On the right are two other forms of variation, which together sum to the total variation. The first expression on the right is termed the **unexplained variation** or the **error sum of squares**. This is the sum of the squares of the error terms, or the sum of squares of the unexplained differences of the Y from their predicted values. The last term on the right is called the **explained variation** or the **regression sum of squares**. This is the sum of the squares of the predicted values about the mean. This can be regarded as explained variation in the sense that the regression line explains these differences. This can be summarized as follows.

Expression for Variation	Source of Variation
$\sum(Y_i - \bar{Y})^2$	Total Variation or Total Sum of Squares
$\sum(Y_i - \hat{Y})^2$	Unexplained Variation or Error Sum of Squares
$\sum(\hat{Y}_i - \bar{Y})^2$	Explained Variation or Regression Sum of Squares

The goodness of fit is defined as the explained variation divided by the total variation.

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

It can now be seen why the limits on R^2 are 0 and 1. If the regression line explains none of the variation in the Y values, then the explained variation is 0, with the unexplained variation being equal to the total variation. In contrast, if there are no errors of prediction, and all the variation is explained, then $R^2 = 1$ because the total variation equals the explained variation.

While it is unlikely to be apparent that R^2 defined in the manner it is here, and the earlier expression used for computing R^2 are equal, the two expressions can be shown to be equal.

In the regression line of Example 11.5.1, it can be shown that the parts of the variation are as follows.

Source of Variation	Amount of Variation
Explained Variation	$\sum(\hat{Y}_i - \bar{Y})^2 = 6.92$
Unexplained Variation	$\sum(Y_i - \hat{Y})^2 = 4.16$
Total Variation	$\sum(Y_i - \bar{Y})^2 = 11.08$

By breaking the variation into these parts, the goodness of fit is

$$\begin{aligned} R^2 &= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \\ &= \frac{6.92}{11.08} \\ &= 0.625 \end{aligned}$$

and this is the same value of R^2 as obtained earlier.

11.5.6 Standard Errors

The last set of statistics to be introduced here is the standard error of estimate, and the standard deviation of the slope. From each observed value of X, the regression line gives a predicted value $\hat{Y}$ that may differ from the observed value of Y. This difference is the error of estimate or error of prediction and can be given the sumbol e. For observation i,

$$e_i = Y_i - \hat{Y}_i$$

As noted earlier, in Example 11.5.1, when all these errors of estimate are added, they total 0. That is,

$$\sum e_i = 0$$

These values of e_i represent deviations of the observed about the predicted values of Y. Just as deviations about the mean can be summarized into the standard deviation, so these errors can be summarized into a standard error. The standard error of estimate is often given the symbol s_e and is

$$s_e = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}}$$

It can be seen that this is very similar to the standard deviation, with two exceptions. In the numerator, the deviations of the predicted values about

the regression line, rather than the deviations of the observed values about the mean are used. The second difference is that the denominator is $n-2$ rather than $n-1$. In the case of s_e, this occurs because the deviations are about the regression line, and two values are required to fix a regression line. In the case of the standard deviation, the deviation is about the mean, and a given value for the mean fixes one of the n sample values.

In terms of computation a different approach is used. It can be shown that

$$s_e = \sqrt{\frac{\sum Y^2 - a\sum XY - b\sum XY}{n-2}}$$

All of the terms on the right are obtained in the process of calculating the regression line, so that s_e can be computed from the sums of the table used to determine a and b.

Finally, the standard deviation for s_b can be determined from s_e. The standard deviation of the sampling distribution of b is

$$s_b = \frac{s_e}{\sqrt{S_{XX}}}$$

Once the standard error of estimate is calculated, then the above formula can be used to determine s_b.

The following example calculates these for the alcohol and income example, and also shows how the standard error can be interpreted.

Example 11.5.3 Standard Errors for Regression of Alcohol Consumption on Income

From Table 11.2 and following, $n = 10$ *and*

$$\sum X = 306.8$$

$$\sum Y = 93.0$$

$$\sum X^2 = 9,503.52$$

$$\sum Y^2 = 875.98$$

$$\sum XY = 2,878.32$$

$$S_{XY} = 25.08$$

$$S_{YY} = 11.08$$

$$S_{XX} = 90.896$$

These can be used to determine s_e and s_b.

$$\begin{aligned} s_e &= \sqrt{\frac{\sum Y^2 - a\sum XY - b\sum XY}{n-2}} \\ &= \sqrt{\frac{875.98 - (0.832 \times 93.0) - (0.276 \times 2,878.32}{10-2}} \\ &= \sqrt{\frac{875.98 - 77.376 - 794.416}{8}} \\ &= \sqrt{\frac{4.188}{8}} \\ &= \sqrt{0.523} \\ &= 0.724 \end{aligned}$$

and the standard error of estimate is 0.724.

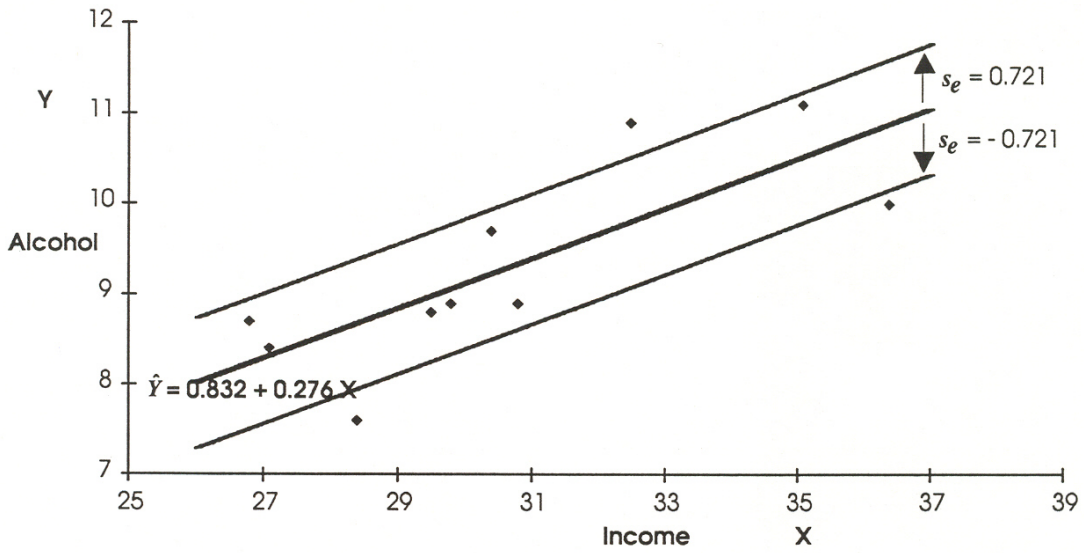
The standard deviation of the sampling distribution of b, sometimes referred to as the standard error of b is

$$\begin{aligned} s_b &= \frac{s_e}{\sqrt{S_{XX}}} \\ &= \frac{0.724}{\sqrt{90.896}} \\ &= \frac{0.724}{9.534} \\ &= 0.0759 \end{aligned}$$

Because of rounding differences, this value for s_b differs slightly from the earlier value obtained from the computer output. Note that s_b is used when conducting the hypothesis test for the statistical significance of the regression coefficient.

Interpretation of the Standard Error of Estimate. The first point to note concerning the standard error of estimate, s_e is that it is measured in units of the dependent variable Y. In the above example, this means that the standard error of estimate is measured in litres of alcohol per capita, the unit used to measure Y. This happens because the calculation of the standard error of estimate is based on the errors of estimate. These are differences between the observed and predicted values of the dependent variable Y, so that these are measured in units of the dependent variable.

Figure 11.10


12
11
10
9
8
7
Y
Alcohol
s_e = 0.721
s_e = - 0.721
$\hat{Y}$ = 0.832 + 0.276 X
25
27
29
31
33
35
37
39
Income
X

The standard error of estimate can be interpreted in much the same manner as the standard deviation. Recall that in Section 5.9.3 of Chapter 5, there were various rules concerning the percentage of cases that were within one, two or three standard deviations of the mean. These same rules can be used for the standard error of estimate, with the exception that these are distances around the line, rather than around $\bar{Y}$.

Figure 11.10: Standard Error and the Regression Line

Figure 11.10 gives the scatter diagram for the alcohol and income example, with the regression line $\hat{Y} = 0.832 + 0.276X$ shown as the heavily shaded line in the centre. The standard error for this example is $s_e = 0.721$. Two lines are drawn parallel to the regression line, one line a distance of 0.721 above the regression line, and the other line a distnace of 0.721 below the regression line. These two lines define a band around the regression line that is within one standard error of the line. As a rough rule of thumb, it is to be expected that about two thirds of all the points in the scatter diagram lie within this band. By counting the points, it can be seen that 7 of the 10 points of the scatter diagram lie within this band, so that 3 of the 10 points lie farther than one standard error from the regression line.

A line parallel to the regression line, but lying two standard errors on each side of the regression line could also be drawn. The rule of thumb concerning this is that 95% or more of the points in the scatter diagram lie within two standard errors of the least squares regression line. While this line is not drawn in Figure 11.10, it is easy to see that all 10 of the points in the diagram do lie within two standard errors of the regression line.

The standard error provides an idea of how well the line fits the points of the scatter diagram. If the standard error is very small, then the fit of the line to the points is very good. If the standard error is quite large, then the line does not fit the points very well. Exactly what is a large or small standard error depends on the problem involved, and the extent to which an accurate estimate of the dependent variable Y is required. In the alcohol and income example, the standard error is 0.721 litres of alcohol. This provides a rough estimate of the relationship between income and alcohol consumption, meaning that estimates of alcohol consumption on the basis of knowledge of family income can be provided to within about three quarters of a litre, about two thirds of the time.

The rules of thumb concerning the standard error are that

$$\hat{Y} \pm s_e \text{ contains about two thirds of the cases.}$$
$$\hat{Y} \pm 2s_e \text{ contains roughly 95\% of the cases.}$$
$$\hat{Y} \pm 3s_e \text{ contains over 99\% of the cases.}$$

While these are very rough rules, they provide another way of thinking about the fit of the regression line. R^2 provides a statistic that describes the goodness of fit of the line. The standard error of estimate, s_e is a statistic that gives an idea of the average distance from the line that the observed values of X and Y lie.

11.5.7 Example of Regression Using Time Series Data

The data in Table F.1 of Appendix F allows a researcher to examine the relationship between various labour force variables in more detail than in Example 11.4.4. The earlier example gave a negative correlation between the crude birth rate and the female labour force participation rate in Canada from the 1950s through the mid 1970s. In Appendix F, annual data for Canada for the years 1953 through 1982 is provided. While many connections among these variable could be hypothesized, here regression models that examine two hypotheses are considered.

The labour force participation rate for females aged 20-24 either fell slightly, or stayed constant through most of the 1950s, but rose continuously and dramatically from 1959 through 1982. Various explanations for these changes have been given by researchers. This section focusses on two of these. Some researchers have argued that the increased participation of young females in the labour force was a result of the increase in wages that these females could receive. A second explanation is that young people and young families experienced a decline in relative economic status beginning in the 1950s, and this led to increased labour force participation of young females in the labour force. This section examines these two models, providing two regression models. Following the calculation of the regression lines and the tests of significance, there are some comments on the usefulness of these models.

The decline in relative economic status argument is examined first. Economic status is measured by variable RIYF, the relative income of young males. As noted in Appendix F, this variable is a ratio of the income of young males to the income of middle aged families. This variable increases from 1953 to 1957, and declines continuously after that. While the absolute income of young families increased through much of this period, in relative terms young families experienced a decline in economic status. That is, relative to the living standards of middle aged families, younger males had lower incomes. This could have led to a decline in the economic status of young families, had not more women entered the labour force, and contributed to family income. According to the relative economic status argument, more young females entered the labour force in an attempt to earn income that could maintain or improve the economic status of young families. In order to test whether this is a reasonable explanation for the increase in female labour force participation, FPR is the dependent variable, and RIYF is the independent variable. Table 11.4 gives the calculations for the regression, with Y representing the female labour force participation rate FPR, and X representing the relative economic status of young families.

$$\sum X = 8.58147$$

$$\sum Y = 1726.0$$

$$\sum X^2 = 2.597183$$

$$\sum Y^2 = 101,926.30$$

Year	RIYF X	FPR Y	X^2	Y^2	XY
1953	0.37866	47.2	0.143383	2227.84	17.8728
1954	0.38799	46.6	0.150536	2171.56	18.0803
1955	0.39685	46.3	0.157490	2143.69	18.3742
1956	0.40498	47.1	0.164009	2218.41	19.0746
1957	0.41285	46.5	0.170445	2162.25	19.1975
1958	0.38780	47.4	0.150389	2246.76	18.3817
1959	0.36281	46.5	0.131631	2162.25	16.8707
1960	0.34889	47.9	0.121724	2294.41	16.7118
1961	0.33593	48.7	0.112849	2371.69	16.3598
1962	0.31499	49.7	0.099219	2470.09	15.6550
1963	0.29566	50.3	0.087415	2530.09	14.8717
1964	0.27796	51.0	0.077262	2601.00	14.1760
1965	0.26175	52.6	0.068513	2766.76	13.7681
1966	0.28461	55.6	0.081003	3091.36	15.8243
1967	0.30635	56.6	0.093850	3203.56	17.3394
1968	0.25556	58.4	0.065311	3410.56	14.9247
1969	0.20925	59.3	0.043786	3516.49	12.4085
1970	0.23535	58.5	0.055390	3422.25	13.7680
1971	0.24629	59.9	0.060659	3588.01	14.7528
1972	0.24630	60.5	0.060664	3660.25	14.9011
1973	0.23382	62.5	0.054672	3906.25	14.6138
1974	0.23966	63.0	0.057437	3969.00	15.0986
1975	0.23923	67.0	0.057231	4489.00	16.0284
1976	0.22450	67.4	0.050400	4542.76	15.1313
1977	0.23961	68.9	0.057413	4747.21	16.5091
1978	0.21401	70.3	0.045800	4942.09	15.0449
1979	0.23843	71.3	0.056849	5083.69	17.0001
1980	0.20914	73.0	0.043740	5329.00	15.2672
1981	0.22050	72.9	0.048620	5314.41	16.0745
1982	0.17174	73.1	0.029495	5343.61	12.5542
Total	8.58147	1726.0	2.597183	101926.30	476.6349

Table 11.4: Calculations for Regression of FPR on RIYF

$$\sum XY = 476.6349$$
$$n = 30$$

$$\begin{aligned} S_{XY} &= \sum XY - \frac{(\sum X)(\sum Y)}{n} \\ &= 476.6349 - \frac{(8.58147)(1726.0)}{30} \\ &= 476.6349 - 493.7206 \\ &= -17.0857 \end{aligned}$$

$$\begin{aligned} S_{XX} &= \sum XX - \frac{(\sum X)^2}{n} \\ &= 2.597183 - \frac{8.58147^2}{30} \\ &= 2.597183 - 2.454721 \\ &= 0.142462 \end{aligned}$$

$$\begin{aligned} S_{YY} &= \sum Y^2 - \frac{(\sum Y)^2}{n} \\ &= 101,926.30 - \frac{1,726.0^2}{30} \\ &= 101,926.30 - 99,302.53 \\ &= 2,623.77 \end{aligned}$$

The value for the slope of the regression line is

$$b = \frac{S_{XY}}{S_{XX}} = \frac{-17.0857}{0.142462} = -119.932$$

The intercept is

$$\begin{aligned} a &= \bar{Y} - b\bar{X} \\ &= \frac{1,726.0}{30} - (-119.932\frac{8.58147}{30}) \\ &= 57.533 - (-119.932 \times 0.286049) \\ &= 57.533 + 34.306 = 91.839 \end{aligned}$$

The least squares regression line is thus

$$\hat{Y} = 91.839 - 119.93X.$$

Note that the slope of the line is negative, indicating that as the relative economic status of the young males and the female labour force participation rate move in opposite directions from each other. For most of the 1953-1982 period, the relative economic status of young males declined while the female labour force participation rate increased.

The fit of the line is given by the goodness of fit statistic R^2. This is

$$\begin{aligned} R^2 &= \frac{S_{XY}^2}{S_{XX}S_{YY}} \\ &= \frac{(-17.0857)^2}{0.142462 \times 2,623.77} \\ &= \frac{291.9211}{373.7875} \\ &= 0.7810 \end{aligned}$$

This means that 0.7810 of the total variation in the female labour force participation rate can be explained statistically on the basis of variation in the relative economic status of young males. Also note that the correlation coefficient r is the square root of R^2, so that

$$r = \sqrt{R^2} = \sqrt{0.7810} = -0.8837.$$

The correlation coefficient is the negative square root in this case, because $S_{XY} < 0$, and the relationship between X and Y is a negative one.

The standard error of estimate is

$$\begin{aligned} s_e &= \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}} \\ &= \sqrt{\frac{101,926.30 - (91.839 \times 1,726.0) - (-119.93 \times 476.6349)}{30-2}} \\ &= \sqrt{\frac{101,926.30 - 158,514.114 + 57,162.83256}{28}} \\ &= \sqrt{\frac{575.00956}{28}} \\ &= \sqrt{20.53606} \\ &= 4.532 \end{aligned}$$

and the standard error of estimate is 4.532.

The standard deviation of the sampling distribution of b is

$$\begin{aligned} s_b &= \frac{s_e}{\sqrt{S_{XX}}} \\ &= \frac{4.532}{\sqrt{0.142462}} \\ &= \frac{4.532}{0.377441} \\ &= 12.007 \end{aligned}$$

This value can now be used to test for the significance of the slope of the regression line. The null hypothesis is that there is no relationship between X and Y, that is, that $\beta = 0$, where β represents the true slope of the relationship between X and Y. The alternative hypothesis is that $\beta < 0$ since the claim is that the economic status of young families (X) is negatively related to female labour force participation (Y). The null and alternative hypotheses are

$$H_0 : \beta = 0$$

$$H_1 : \beta < 0$$

The statistic used to test the relationship is b, and it has mean β and standard deviation s_b. The distribution of b is a t distribution with $n - 2 = 30 - 2 = 28$ degrees of freedom. At the 0.01 level of significance, for a one tailed t test, the critical t value is -2.468. The region of rejection of H_0 is all t values of -2.468 or lower. The t statistic is

$$t = \frac{b}{s_b} = \frac{-119.93}{12.007} = -9.988 < -2.468.$$

The t value is in the region of rejection, so that b is enough different from 0 to reject the hypothesis of no relationship between X and Y. There is very strong evidence that $\beta < 0$ and that the relationship between the relative economic status of young males and female labour force participation is a negative relationship.

Wages and Female Labour Force Participation. For the second hypothesis, the claim is that increased wages were an important reason why female labour force participation increased. For this regression model, the independent variable X is AWW, the index of average weekly wages. The dependent variable Y is again the female labour force participation rate, FPR. The detailed calculations for this regression model are not given here,

but only the summations and the statistics required to describe and test the regression model are provided. You can use the values of AWW as X and FPR as Y, to verify the following. You need not carry all the decimals shown below, although you should carry at least 4 or 5 significant figures throughout the calculations. If you carry fewer significant figures than this, you will come close when making the calculations, but your answers will differ because of rounding.

$$\sum X = 2,757.230$$

$$\sum Y = 1,726.0$$

$$\sum X^2 = 262,272.8960$$

$$\sum Y^2 = 101,926.30$$

$$\sum XY = 163,279.0784$$

$$n = 30$$

$$\begin{aligned} S_{XY} &= \sum XY - \frac{(\sum X)(\sum Y)}{n} \\ &= 163,279.0784 - \frac{(2,757.230)(1,726.0)}{30} \\ &= 163,279.0784 - 158,632.6327 \\ &= 4,646.4457 \end{aligned}$$

$$\begin{aligned} S_{XX} &= \sum X^2 - \frac{(\sum X)^2}{n} \\ &= 262,272.8960 - \frac{2,757.230^2}{30} \\ &= 262,272.8960 - 253,410.5758 \\ &= 8,862.3202 \end{aligned}$$

$$\begin{aligned} S_{YY} &= \sum Y^2 - \frac{(\sum Y)^2}{n} \\ &= 101,926.30 - \frac{1,726.0^2}{30} \\ &= 101,926.30 - 99,302.53 \\ &= 2,623.77 \end{aligned}$$

The value for the slope of the regression line is

$$b = \frac{S_{XY}}{S_{XX}} = \frac{4,646.4457}{8,862.3202} = 0.52429$$

The intercept is

$$\begin{aligned} a &= \bar{Y} - b\bar{X} \\ &= \frac{1,726.0}{30} - (0.52429\frac{2,757.230}{30}) \\ &= 57.533 - (0.52429 \times 91.9077) \\ &= 57.533 - 48.186 = 9.347 \end{aligned}$$

The least squares regression line is thus

$$\hat{Y} = 9.347 + 0.52429X.$$

The slope of the line in this example is positive, indicating that as wages increase, the female labour force participation rate increases. For most of the 1953-1982 period, both wages and female labour force participation rates rose, so that the two variables move together in a positive manner.

The fit of the line is given by the goodness of fit statistic R^2. This is

$$\begin{aligned} R^2 &= \frac{S_{XY}^2}{S_{XX}S_{YY}} \\ &= \frac{4,646.4457^2}{8,862.3202 \times 2,623.77} \\ &= \frac{21,589,457.64}{23,252,689.87} \\ &= 0.9285 \end{aligned}$$

This means that 0.9285, or 92.85% of the total variation in the female labour force participation rate can be explained statistically on the basis of variation in the average wage. The correlation coefficient r is

$$r = \sqrt{R^2} = \sqrt{0.9285} = -0.9636.$$

The standard error of estimate is $s_e = 2.589$ and $s_b = 0.0275$. For the test of statistical significance,

$$H_0 : \beta = 0$$

$$H_1 : \beta > 0$$

and the t statistic is

$$t = \frac{b}{s_b} = \frac{0.52429}{0.0275} = 19.065.$$

There are 28 degress of freedom, and from the t table in Appendix I,

$$t_{0.0005,28} = 3.674$$

The t value of 19.065 is much greater than this, so that the null hypothesis can be rejected very decisively. This data provides strong evidence for a positive relationship between wages and female labour force participation.

Comments on the Results. The first point to note for the above regressions are the high values for the goodness of fit R^2 and the very significant t statistics. This is not unusual when working with time series data. This same feature was noted for correlation coefficients in Example 11.4.6. For other types of data, especially for survey data, such high values for the goodness of fit or for the test of significance cannot ordinarily be expected.

The scatter diagrams for each of the relationships is given in Figures 11.11 and 11.12. As can be seen there, both of the scatter diagrams show a fairly close relationship with FPR. The scatter of wages and FPR appears to more of a straight line relationship, but the scatter of relative economic status and FPR also shows that the two variables are highly related statistically.

The next point to note is that both of the explanations have been supported by the respective regression models. In each case, the dependent variable that was being explained was the female labour force participation rate. The first model used the relative income of young males as the independent or explanatory variable, and showed that there was a negative relationship between the relative economic status of young males, and the female labour force participation rate. Based on the model, the initial increase in the relative economic status of young males, followed by the long and continuous decline, was first associated with little change in female labour force participation, and then a long increase in this participation rate. This explanation makes some sense, and also is strongly supported by the model.

The other explanation is the argument that increased female labour force participation was caused by increases in average wages. Over the thirty year period from 1953 to 1982, the increase in wages made it more attractive for

Figure 11.11: Scatter Diagram of Relative Economic Status and Female Labour Force Participation

females to participate in the labour force. Looked at in another way, females who did not enter the labour force, gave up more potential income by not participating in the labour force. This explanation appears to be a reasonable one, and this explanation is strongly supported by the regression model constructed. The line relating the average weekly wage as the explanatory variable, and the female labour force participation rate as the dependent variable, has a positive slope. The goodness of fit is above 90%, and the slope is very significantly positive.

Using the results here, it is difficult to decide which of the two models provides a better explanation for the variation in the female labour force participation rate over these years. The goodness of fit for the model with wages as the independent variable is greater than for the model with relative economic status as the independent variable, and this might lead a researcher to favour the model using wages as the independent variable. But the other model is also very significant statistically, and the two variables have a close relationship with each other.

Given the above considerations, it might make sense to include both wages and relative economic status as independent variables. This would

Figure 11.12

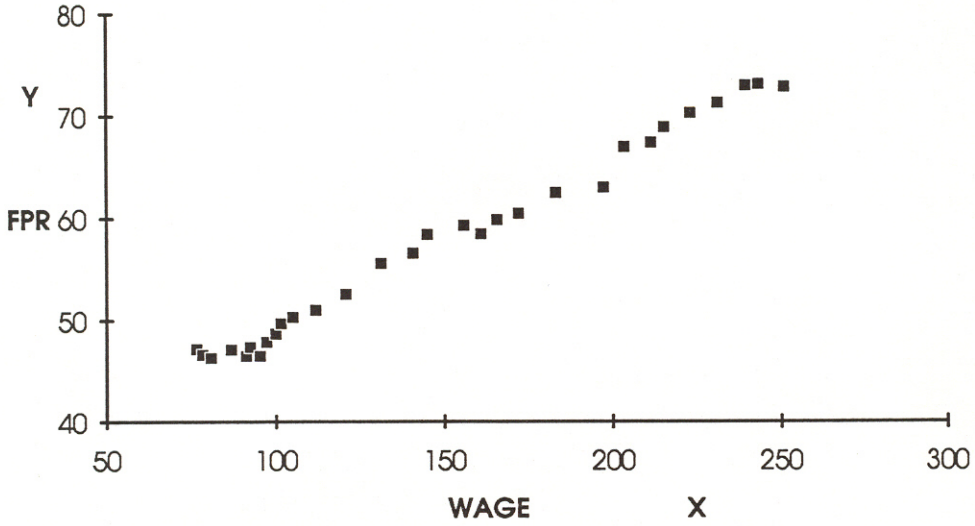


Figure 11.12: Scatter Diagram of Relative Economic Status and Female Labour Force Participation

make the model a multivariate one, and this moves the model beyond the scope of this introductory textbook. However, such a multivariate model might show that both increased wages and the decline in relative economic status of young males contributed to the increase in female labour force participation rates. Each of the contributory variables could be independent factors that encouraged or pushed women into the labour force. In addition, there are likely to be many other factors that led in this same direction. Changed views concerning women's roles, the desire of women for independent sources of income, and the growth in the number of jobs that employed women over these years, all might be contributory factors.

In conclusion, the two regression models given here are both useful means of testing some fairly simple explanations. A more detailed analysis would require a multivariate model, simultaneously examining the influence of a number of different factors on female labour force participation rates.

11.5.8 Regression Line for Data from a Survey

This section gives an example of a regression line being fitted to cross sectional data obtained from a sample survey. With diverse data from a survey, the fit of the regression line is usually much smaller than when using other types of data. But this is partly compensated by the larger sample size that is often available from a survey. As a result, the regression line often has a statistically significant slope, even though the goodness of fit statistic is very low.

This example examines the relationship between the age of the respondent and the annual income of the respondent. Ordinarily it is expected that earnings increase with age for those members of the labour force who are regularly employed. This occurs for various reasons, including factors such as experience gained on the job and seniority. These patterns differ considerably by sex, occupation, industry and educational background, so that a cross sectional survey of all labour force members may not allow this relationship to become visible.

In order to reduce the diversity of respondents, a set of Saskatchewan males that are married, in husband-wife families, between 30 and 60 years of age, and employed is selected. In addition, all of these males are teachers, that is, from the same occupational category, and an occupation where pay does generally increase with years on the job. The data for this example is taken from Statistics Canada's 1989 Survey of Consumer Finances. The income data obtained refers to annual income for the whole of 1988. The age of the respondent is given as the X value in Table 11.5. The earnings of the respondents are given as the Y values in Table 11.5. These are shown in thousands of dollars. For example, the first respondent was 34 years old in 1989, and earned \$24,900 in 1988.

The data in Table 11.5 is used to determine the relationship between age and earnings. Since it is hypothesized that earnings increase with age, earnings are the dependent variable and age is the dependent variable. While age alone does not cause earnings to increase, age can be considered to be a proxy for other variables such as seniority or experience. It is to be expected that there will be a positive relationship between age and earnings.

The scatter diagram for the data in Table 11.5 is given in Figure 11.13. The independent variable, age, is given along the X axis. The dependent variable is earnings, and this is given on the Y axis. While a generally positive relationship between age and earnings may be visible in the diagram, it is not very clear, and the extensive scatter may make it seem as if there

X	Y	X	Y	X	Y
34	24.900	46	46.193	35	40.257
57	69.000	48	52.884	44	48.936
53	42.384	47	48.853	36	29.434
40	60.000	59	40.000	56	42.374
55	56.657	34	30.306	38	38.000
35	42.000	49	36.480	45	40.500
57	64.748	57	52.884	56	67.620
42	57.824	37	50.188	58	42.398
44	42.388	51	40.446	32	42.136
34	13.884	49	40.069	54	30.995
42	42.346	36	38.081	38	22.776
45	45.000	43	42.388	38	45.000
51	20.103	44	28.786	40	51.844
48	52.000	50	46.630	38	52.000
44	22.932	50	40.308	40	39.575
43	35.214	39	39.500	43	41.962
34	29.597				

Table 11.5: Age (X) and Earnings in thousands of dollars (Y) for 49 Saskatchewan Male Teachers

is no relationship between the two variables.

From Table 11.5, the following summations can be obtained.

$$\sum X = 2188.0$$

$$\sum Y = 2070.8$$

$$\sum X^2 = 100576$$

$$\sum Y^2 = 94134$$

$$\sum XY = 94247$$

$$n = 49$$

These summations provide sufficient information to obtain S_{XX}, S_{YY}, S_{XY}, and from these the various statistics required for the regression line

Figure 11.13

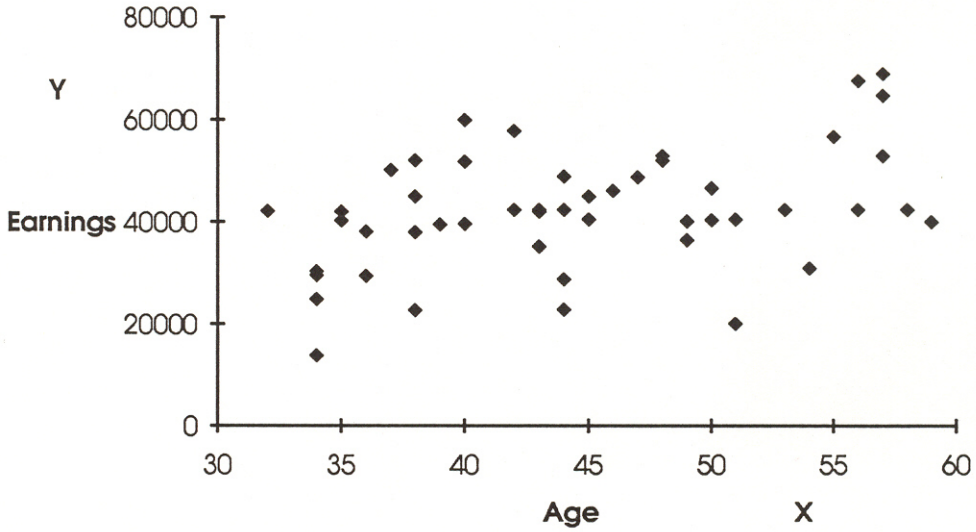


Figure 11.13: Scatter Diagram of Age and Earnings

and the test of significance can be obtained. These are given as follows. (These statistics were obtained from the MINITAB computer program. If you do the calculations yourself, the values may differ slightly because of different rounding.)

$$a = 14.612$$

$$b = 0.6192$$

$$s_b = 0.2021$$

$$s_e = 10.84$$

$$R^2 = 0.166$$

$$r = 0.408$$

The regression equation is

$$\hat{Y} = 14.612 + 0.6192X$$

The slope of 0.6192 means that a unit change in X is associated with a change in Y of 0.619. That is, each increase in age (X) of one year is associated with an increased of 0.619 thousand dollars. This is the estimate of the effect of X on Y, using this sample of 49 male teachers in Saskatchewan. The regression line shows that for this sample, each extra year of age is associated with an increase of \$619 in earnings. This estimate may not apply to any particular individual or time period, but based on this cross section of cases, this is an estimate of the average effect of X on Y.

The significance of the slope of the line can be tested using the t test. The null hypothesis is that the true slope for the regression line between age and income, β, is zero, and the alternative hypothesis is that β exceeds zero. The t statistic is

$$t = \frac{b - \beta}{s_b} = \frac{b - 0}{s_b} = \frac{0.6192}{0.2021} = 3.064.$$

For $n - 2 = 49 - 2 = 47$ degrees of freedom, the t value for a one tailed test at the 0.005 level of significance is just over 2.678. The observed value of 3.064 exceeds this, so that the null hypothesis of no relationship between between age and earnings can be rejected. At the 0.005 level of significance, there is evidence that the relationship between age and earnings is positive.

Figure 11.14 shows that scatter diagram, along with the least squares regression line. Around the line, a band that lies within one standard deviation is also drawn. As copuld be noted in the earlier diagram, the regression line does not fit the points all that closely. However, the line does have a positive slope, so that the relationship between X and Y is generally a positive one. It can be seen that a considerable number of the points fall within one standard error of the regression line. The standard error $s_e = 10.8$, or \$10,800, so that about two thirds of all the respondents have earnings that are within \$10,800 of the earnings predicted from the regression line.

Finally, note that the goodness of fit of the regression line is only $R^2 = 0.166$. After working with the previous examples, where the goodness of fit was over 0.6 in all cases, this might seem too small a goodness of fit to make the regression model at all worthwhile here. But the slope is statistically significant, so that the assumption of no relationship between age and earnings can be rejected. When survey data is used, it is not uncommon to encounter this situation. The scatter diagram seems so dispersed, and the goodness of fit statistic seems so low, that there appears to be no relationship between X and Y. But if the slope is significant statistically, there is evidence of some relationship between X and Y. This situation

Figure 11.14

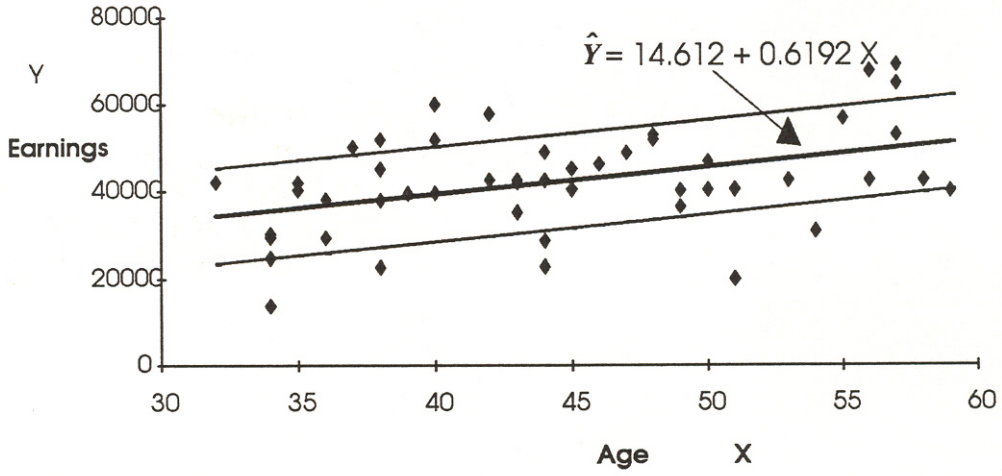

$\hat{Y}$ = 14.612 + 0.6192 X
Y
Earnings
80000
60000
40000
20000
0
30
35
40
45
50
55
60
Age X

Figure 11.14: Scatter Diagram and Regression Line for Age and Earnings

cannot be fully resolved on the basis of a two variable regression model, because this most likely means that there are various other factors that also affect Y. What the two variable regression model shows is that age is a significant factor in explaining earnings, so that as age increases, earnings generally increase. At the same time, age by itself can only explain about 0.166, or 16.6% of the variation in earnings. There are many other factors that influence earnings, of which age is only one. Age itself can explain a certain portion of the variation in earnings, but the other 80% or more of the variation in earnings is left unexplained. While this model gives no indication of what these other factors are, among the variables that might be investigated are the amount of schooling of the respondent, the length of time in the job, the continuity of employment, who the employer is, and so on. In addition, some characteristics of individuals such as the age at which the respondent began full time work, the abilities of the individual, and the level of ambition of the individual, might also be investigated. Some of these factors may be diffficult to quantify and investigate. As a result, even with a multivariate model of earnings, it is difficult to explain over one half of the

variation in individual earnings in a large cross sectional survey of respondents. The two variable regression model is still very useful though, because it does allow the researcher to determined what are some of the factors that do affect earnings. Based on bivariate regression models, the researcher can then build a multivariate model, with several explanatory variables, that can be used to explain variation in earnings.

11.5.9 Additional Comments on Regression

Before leaving the two variable regression model, a few additional comments concerning various aspects of the regression model are made here.

Reporting Results. When reporting the results of a regression equation it is also common to report, in brackets, the standard errors of each coefficent. The results The equation for the regression of earnings on age in Section 11.5.8 could be reported as follows:

$$\begin{array}{rccc} Y = & 14.612 & + & 0.6192X \\ & (9.1554) & & (0.2021) \end{array}$$

where the standard deviation of a is 9.1554 and $s_b = 0.2021$. Hypotheses tests for the intercept or the slope of the line could be easily conducted using the data from this equation as it is reported here.

A Quick Test of Significance. A quick test of significance for the slope of the regression line can be made by comparing the value of b and s_b. As a very rough rule of thumb, if b is at least twice the size of s_b, the slope can be regarded as being significant statistically. This is based on the following considerations.

The null hypothesis is that the slope of the true regression line is $\beta = 0$. The t statistic is

$$t = \frac{b - \beta}{s_b}$$

so that under the assumption that $\beta = 0$,

$$t = \frac{b}{s_b}$$

If t exceeds 2, and if $n > 8$, the t table shows that the null hypothesis that $\beta = 0$ can be rejected at the 0.05 level of significance. (For $n = 8$, meaning

$n - 2 = 6$, the t value for a one tailed test at the 0.05 level of signficance is $t = 1.943$, so that a t value of 2 or more would lead to rejection of the null hypothesis. Since the $\alpha = 0.05$ level of significance is often the level chosen, and since the sample size usually exceeds 8, this provides a quick test for whether or not the slope is significantly greater than or less than zero.

From the reported regression result on the previous page, $b = 0.6192$ and $s_b = 0.2021$ so that

$$t = \frac{b}{s_b} = \frac{0.6192}{0.2021} = 3.1 > 2$$

so that the slope of the regression line in that equation can be considered to be significantly greater than 0.

If the result is to be more soundly based, the formal test of significance should be conducted. But if you have several regression slopes to check, the rule that the slope should be about double the value of the standard deviation of the slope gives a quick way of providing a check for the statistical significance of the slope of the line.

Assumptions for the Regression Model. None of the assumptions for the regression model have been given in this section, except to note that both the independent and the dependent variable should have an interval or ratio level of measurement. In order to obtain the estimates of a and b, no other assumptions are required. However, in order to conduct the hypothesis test for the line, and to ensure that the regression model is meaningful, there are several other assumptions involved in the model. These assumptions are outlined here, but there full importance is not discussed in any detail.

The assumptions concerning the regression model relate to the behaviour of the error term. In the true regression model,

$$\hat{Y} = \alpha + \beta X + \epsilon.$$

The term ϵ covers all the unexplained factors, variables other than the variable X that might be considered to affect Y. It is the behaviour of this term ϵ that is of concern when conducting tests of significance for the regression line, and deciding whether the line has any real meaning.

The assumption concerning ϵ is that the error term has a normal distribution with mean 0 and the same variance for each of the possible values of X. In addition, for different values of X, the assumption is that the error terms ϵ are uncorrelated with each other.

While these assumptions concerning ϵ are fairly simply stated, their meaning and the consequences of violating these assumptions are not nearly so clear. Much of a second course in statistics may be devoted to examining the validity of the assumptions when working with data, and deciding how to work with the data when the assumptions are violated. Here only a few comments are made concerning these assumptions.

1. The error term ϵ includes all the unexplained factors, that is, the effect of all those variables other than X, that may influence Y. A short consideration of this should allow you to realize that this makes it quite unlikely that this term will be normally distributed, much less having the other assumptions given. If a variable that has a major effect on Y has been excluded from the equation, then the effect of this variable on Y will be systematic, meaning that when this effect is included in ϵ, the distribution of ϵ is very unlikely to be normal.

2. The assumption that the variance of ϵ will be the same for all values of X is often violated. As noted when discussing the coefficient of relative variation in Chapter 5, the standard deviation is often larger when values of the variable are larger. If this is the case, then the variance also differs for different values of X and this may create misleading hypothesis tests.

3. One of the assumptions was that the values of ϵ are unrelated to each other for different X values. With time series data, as in Section 11.5.7, this assumption is almost always violated. One of the reasons for this is that annual observations constitute an arbitrary unit of time. Social and economic variables do not stop having their effect on December 31, to start completely anew on January 1. Factors that are measured over time tend to have an influence that extends across time in a manner that is not neatly divided into units such as days, months or years. This means that influences on Y that occur in one time period are likely to be felt in the next time period as well. If some of these factors are included in ϵ, then this makes the values of ϵ for one time period correlated with those of other time periods.

The above list gives only a few examples of the way in which the assumptions concerning the construction or testing of a regression line might be violated. Economists have spent considerable time and effort examining the consequences of violation of these assumptions, and the subject of econometrics is largely devoted to dealing with these.

Summary. In spite of these problems, the regression model has proven very useful for examining the structure of the relationship between two variables. If done with care, this model can also prove very useful for purposes of prediction. By adding more explanatory variables, and constructing multivariate regressions, the model can often illustrate many aspects of socioeconomic structure and behaviour which might not otherwise be apparent.

11.6 Conclusion

This chapter has examined the relationship between two variables in several different ways. Before summarizing these, a few comments concerning the way in which analysis of these relationships can be extended are made.

Multivariate Models. The main defect of the bivariate models is that the effect of variables that have not been included in the models cannot easily be considered. In addition, where two or more variables interact to cause an effect on one or more dependent variables, bivariate models may be unable to explain this, or may mislead a researcher concerning the true relationships among variables. One way of dealing with these problems is to create multivariate models, where the effect of all relevant variables is considered.

When dealing with many variables, all of which are measured at no more than the nominal level, a multivariate model produces tables of 3, 4, 5 or more dimensions. These are very difficult to analyze, although some researchers use **loglinear models** to examine these. The other difficulty of these models is that even where relationships among variables are found, it may be difficult to describe them in an understandable manner.

Where the variables have at least an ordinal level of measurement, researchers generally move well beyond cross classification tables, and are able to examine the correlation among the rankings of members of a population on several variables. The same is possible with variables that have interval or ratio scales. For variables having these latter scales, the correlation coefficients measure the manner in which distances between values of the variable are related. Using these correlation coefficients, it is possible to construct various types of multivariate models. **Factor analysis** is one of the most common of these. In factor analysis, a researcher takes a large number of variables, and attempts to group these variables into common types of variables, or factors. **Cluster analysis** is another multivariate method that

can be used. Cluster analysis can be used to group variables together, but is more commonly be used to provide clusters of cases that are reasonably similar to each other.

If all the variables have interval or ratio level scales, then multivariate regression models are commonly used. One advantage of a regression model over factor or cluster analysis is that the regression model can be used to obtain an estimate of the actual amount of change in a dependent variable that occurs as a result of a change in an independent variable. Where the model includes several independent variables, both the individual and the combined effect of these on the dependent variable can be estimated. For example, in the example of the female labour force participation rate, it is possible to obtain a model that examines the effect of both increased wages, and declining economic status, on female labour force participation rates.

It is also possible to produce simultaneous equation estimates, where variables may be simultaneously independent and dependent variables. In the labour force participation example, the increased entry of females into the labour force helps improve the economic status of families. Thus the relative economic status of young families cannot be regarded as a variable that is completely independent of other labour force variables. The two are simultaneously determined.

Multivariate models go considerably beyond what can be introduced in an introductory textbook. The methods introduced in this chapter provide a way of beginning to examine the relationship between two variables. Once you understand these basic principles, it should be possible to begin working with multivariate models. These are likely to be encountered in a second course in statistics.

Summary. This chapter examines measures of association, a means of providing a summary statistic to explain the relationship between two or more variables. In this chapter, only the two variable, or bivariate model, was discussed.

The first methods used in the chapter allow a researcher to examine the relationship among any two variables. The chi square based measures of association and the proportional reduction in error methods begin with a cross classification table. Since a cross classification table can be produced for variables that have no more than a nominal scale of measurement, these first measures can always be used.

When the variables are measured with a least ordinal scales, then the

different values of the variable can be ranked. This ranking information can be used to determine correlation coefficients. These coefficients provide more information concerning the nature of the relationship between variables, and if they can legitimately be calculated, are to be preferred to the first set of measures.

Finally, if the variables have at least an interval level of measurement, then the method of regression may be more appropriate. The regression model requires that the researcher has some idea of the direction of causation or of influence. If this is known, then the regression model provides considerably more information than the earlier methods. A regression line allows the researcher to tell the influence of a change in the values for one variable on another variable. The line can be used to determine this relationship as well as predict values for the independent variable.

All of the measures in this chapter are widely used. In Sociology, Psychology and Political Science, many of the scales used to measure variables have only a nominal or ordinal level of measurement. If this is the case, then regression models are less commonly used. Where all the variables are measured using interval and ratio level scales, then the more powerful methods of regression can be used. In Economics and Administration, many of the variables are measured in monetary units, and this allows researchers there to concentrate on regression models.