5. Linear Regression with Single and Multiple Regressors

Management decisions are often based on the relationship between two or more variables. For example, a marketing manager may try to forecast sales at a certain level of advertising expenditure after examining the relationship between that expenditure and sales.

In the second case, the public undertaking may use the ratio between the daily maximum value temperature and electricity demand to predict electricity consumption. Sometimes the manager relies on intuition. Intuitively, he judges how the two variables are related. However, if it is possible to obtain the data, it makes sense to use a statistical procedure called regression analysis to show how the two variables are related to each other.

In regression terminology, the predicted variable is called the dependent variable.

The variable or variables used to predict the value of the dependent variable are called independent variables.

In analysing the effect of advertising expenditure on sales, sales would thus be the dependent variable. Advertising expenditure would be the independent variable. In statistical notation y denotes the dependent variable, and x denotes the independent variable.

In this section, we will look at the simplest type of regression analysis, which involves one independent variable and one dependent variable. The relationship between the two variables will be approximated by a straight line. It is called simple linear regression. Regression analysis involving two or more independent variables is called multiple regression analysis.

5.1 Simple linear regression model

Best Burger is a chain of fast-food restaurants located in a multi-state area. Best Burger locations are located near university campuses. Managers believe that the quarterly sales of these restaurants (indicated by y) is positively correlated with the size of the student population (denoted by x). Restaurants near campuses with a large number of students tend to generate more sales than those near campuses with a small number of students. Using regression analysis, we can develop an equation that shows how the dependent variable y is related to the independent variable x.

5.2 Regression model and regression equation

In the case of Best Burger, the population is all Best Burger restaurants. For each restaurant in the population there is a value x (student population) and a corresponding value y (quarterly sales). The equation describing how the y is related to x is called a regression model.

$$y = \beta_0 + \beta_1 x + \epsilon$$

β_0 and $\beta_1$are called the model parameters, ϵ (Greek letter epsilon) is a random variable called the model error. The error represents the variability y which cannot be explained by a linear relationship between x in y.

The population of all Best Burger restaurants can also be seen as a collection of sub-populations, one for each separate value x. For example, one subpopulation consists of all Best Burger restaurants near university campuses with 8000 students. The second subpopulation consists of all Best Burger restaurants located near university campuses with 9000 students and so on. Each subpopulation has a corresponding distribution of values y. Each value distribution y has its mean or expected value. The equation describing what the expected value is y, denoted by $E(y)$, which is related to x is called the regression equation. The regression equation for a simple linear regression is as follows

$$E(y) = \beta_0 + \beta_1 x$$

The graph of a simple linear regression equation is a straight line. β_0 represents the initial value of the regression line, β_1 is the direction coefficient of the line, and $E(y)$ the mean value or expected value y for a given value of x.

Examples of possible regression lines are shown in the figure 5.1 below. The regression line in case A shows that the value of y is positively correlated with x. As the values increase x, the values also increase $E(y)$. The regression line in Panel B shows the value of y, which is negatively correlated with x. Where smaller values are $E(y)$ are associated with higher values x. The regression line in Panel C shows the case where the value of y is not associated with x. This means that the value y is the same for each value x.

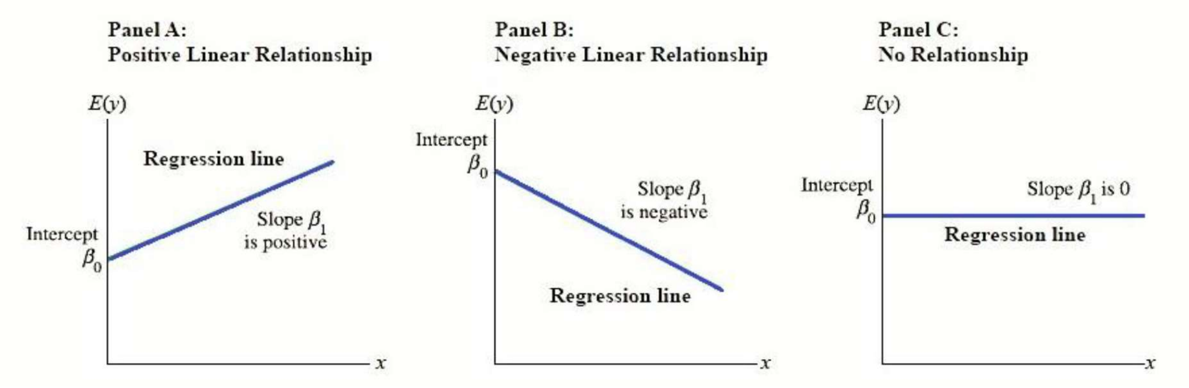


Figure 5.1 Graph examples of Linear Relationship.

5.3 Estimated regression equation

If the values of the population parameters were known β_0 and β_1, we could use the above equation to calculate the values of y for a given value x. In practice, these parameters are difficult to access, so they are simply estimated using sample data. The sample statistics (denoted by b_0 and b_1) are calculated as estimates of the population parameters β_0 and β_1. Replacing the values of the sample statistics b_0 and b_1 instead of $\beta_0$and $\beta_1$in the regression equation gives us a new, estimated regression equation. The estimated regression equation for a simple linear regression is as follows

$$\hat{y} = b_0 + b_1 x$$

The graph of an estimated simple linear regression is called the estimated regression line. b_0 represents the initial value of the regression line, b_1 is the direction coefficient of the line.

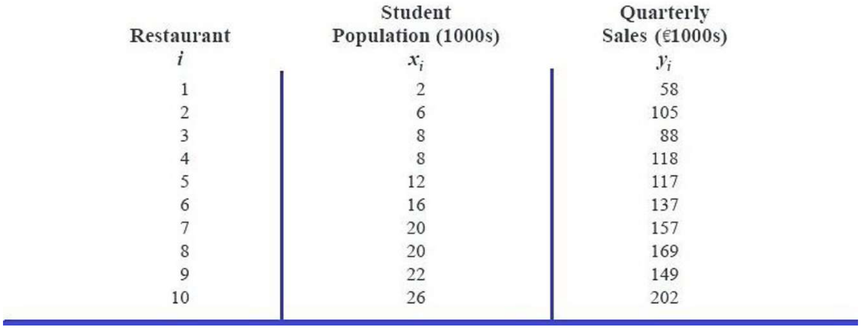
Below we show how to use the least squares method to calculate the values of b_0 and b_1 in the estimated regression equation.

In general $\hat{y}$ (score for $E(y)$) average value y for a given value x. If we now wanted to estimate the expected value of quarterly sales for all Best Burger restaurants located close to campuses with 10000 students, the value would be x would be replaced by the value 10000 in the last equation. In some cases, however, we may be more interested in forecasting sales for only one specific restaurant. For example, suppose you wanted to forecast quarterly sales for a restaurant that you plan to build near a college with 10000 students. As it turns out, even in this case, the best predictor of the value of the y for a given value x value $\hat{y}$.

5.4 Least squares method

The least squares method is a procedure where, using sample data, we find the equation of the estimated regression line. To illustrate the least squares method, let us assume that the data were collected from a sample of 10 Best Burger restaurants near university campuses. With x_i will denote the size of the student population (in thousands) and by y_i the size of quarterly sales (in thousands of EUR). $x_i \; in \; y_i$ for the 10 sample restaurants are summarised in the table below. We see that restaurant 1, z x_1 = 2 and y_1 = 58, is close to a campus with 2000 students and has quarterly sales of € 58,000. Restaurant 2, with x_2 = 6 and y_2 = 105, is close to a campus with 6000 students and has quarterly sales of 105.000 €. The restaurant with the highest sales value is restaurant 10, which is close to the campus with 26,000 students and has quarterly sales of € 202,000.

The following is a scatter plot of the data in the figure 5.2 below. Student population is shown on the horizontal axis and quarterly sales on the vertical axis. The scatter diagrams for the regression analysis are constructed with the independent variable x on the horizontal axis and the dependent variable y on the vertical axis. The scatter diagram thus allows us to draw preliminary conclusions about the possible relationship between the variables.



Restaurant i	Student Population (1000s) x_i	Quarterly Sales (€1000s) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Figure 5.2 Scatter plot of data.

What preliminary conclusions can be drawn from the figure below 5.3? Higher quarterly sales occur in campuses with a larger student population. In addition, there is a constant relationship between the size of the student population and quarterly sales, which can be described by a straight line. Between $x \; in \; y$ a positive linear relationship is indeed implied. Therefore, we have chosen a simple linear regression model to represent the relationship between quarterly sales and the student population. Given this choice, our next task is to use the sample data table to determine the values of b_0 and b_1, which are important parameters in the estimation of a simple linear regression equation. For the i-th restaurant, the estimated regression equation is

$$\hat{y}_i = b_0 + b_1 x_i$$

Where

$\hat{y}_i$ _ estimated value of quarterly sales (€1000) for the i-th restaurant

b_0 _ initial value of the estimated regression line

b_1 _ direction coefficient of the estimated regression line

x_i _size of the student population (1000) for the i-th restaurant

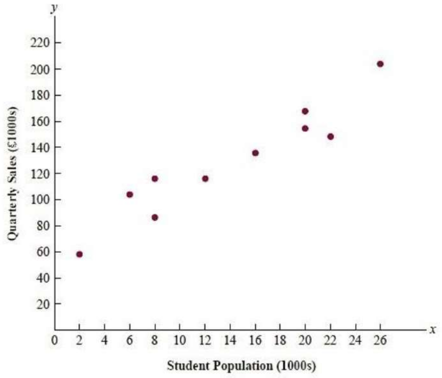


Figure 5.3 Scatter plot graph.

y_i denotes the observed (actual) sales for the restaurant i and $\hat{y}_i$, representing the estimated value of sales for the restaurant i, each restaurant in the sample will have an observed sales value of y_i and the predicted sales value $\hat{y}_i$. For the estimated regression line to ensure a good fit to the data, we want the differences between the observed sales values and the predicted sales values to be as small as possible.

The least squares method uses sample data to provide values b_0 and b_1.

Minimise the sum of the squares of the deviations between the observed values of the dependent variable y_i and the predicted value of the dependent variable $\hat{y}_i$. The starting point for calculating the minimum sum by the least squares method is given by the expression

Minimum Sum Criterion: $\min \sum (\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2$

Where

$\mathrm{y_i}$ =observed value of the dependent variable for the i-th observation

$\hat{\mathrm{y}}_\mathrm{i}$ =predicted value of the dependent variable for the i-th observation

The directional coefficient of the regression line and the initial value:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

x_i _ value of the independent variable for the i-th observation

y_i _ value of the dependent variable for the i-th observation

$\bar{x}$ _ average value for the independent variable

$\bar{y}$ _ average value for the dependent variable

n _total number of observations

Some of the calculations needed to develop the estimated least squares regression line are shown below. With a sample of 10 restaurants, we have n=10 observations. The above equations first require the calculation of the mean value of x and the average value y.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{10} = 14, \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{1300}{10} = 130$$

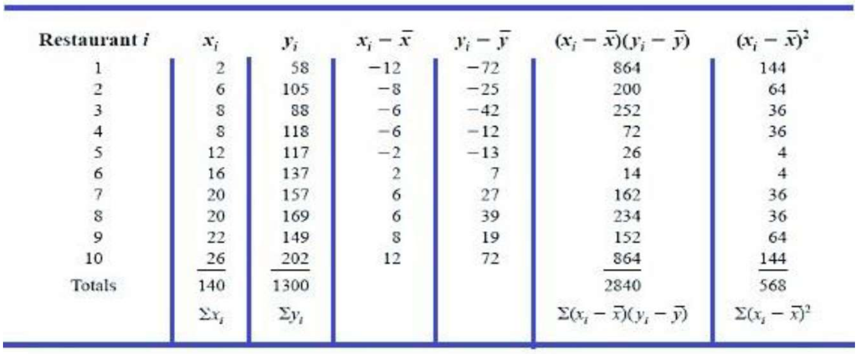
Alternative calculation equation b_1:

$$b_1 = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}$$

Using the last equations and the information in Figure 5.4, we can calculate the directional coefficient of the regression line for the Best Burger restaurants example. Calculating the slope (b_1) is as follows.

Figure 5.5 shows a plot of this equation on a scatter diagram.

The slope of the estimated regression equation or the directional coefficient of the equation ($b_1 = 5$) is positive.



Restaurant i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

Figure 5.4 Plot of equation on a scatter diagram.

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

This is followed by the calculation of the initial value (b_0).

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

This is how the regression equation is estimated:

$$\hat{y} = 60 + 5x$$

Figure shows a plot of this equation on a scatter plot.

The slope of the estimated regression equation (b_1 = 5) is positive, which means that as a student population increases, sales increase. In fact, we can infer (based on measured sales in the 1000s and student population in the 1000s), meaning an increase in the student population of 1000 is associated with an increase in expected sales of 5000; i.e. quarterly sales are expected to increase by 5€ per student.

If we believe that the regression equation, estimated by least squares, adequately describes the relationship between $x\ in\ y$, it seems reasonable to use the estimated regression equation predict the value y for a given value x. For example, if you wanted to predict quarterly sales for a restaurant located near a campus of 16,000 students would calculated by

$$\hat{y} = 60 + 5(16) = 140$$

Therefore, we would assume quarterly sales of 140,000 for this restaurant. In the following sections we discuss methods for assessing the appropriateness of using the estimated regression equation for estimation and forecasting.

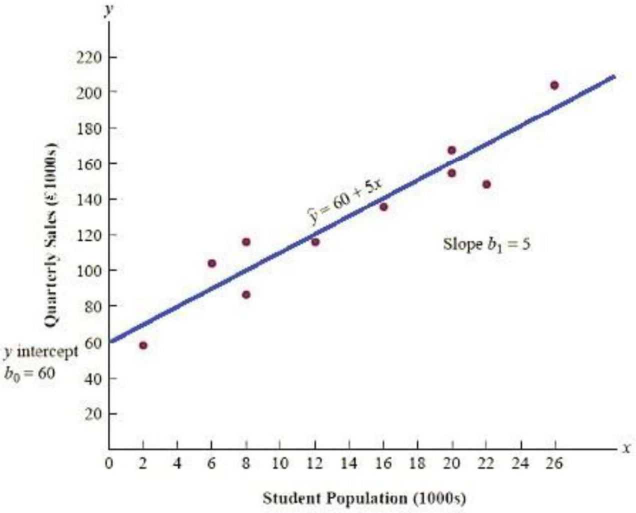


Figure 5.5 Scatter plot of student population and quarterly sales.

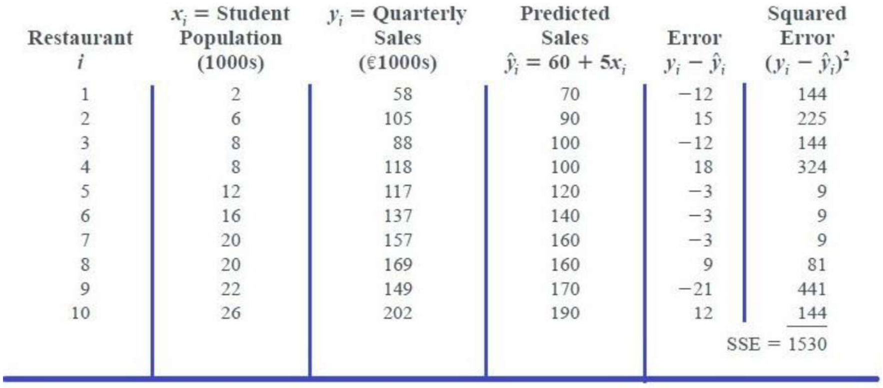
5.5 Coefficient of determination

For the Best Burger restaurants example, we developed an estimated regression equation $y = 60 + 5x$ for an approximately linear relationship between the size of the student population x and quarterly sales y. Now the question is: how well does the estimated regression equation fit the data? In this section, we show that the coefficient of determination provides a goodness-of-fit measure for the estimated regression equation. For the i-th observation, the difference between the observed value of the dependent variable $y_i$and the predicted value of the dependent variable is called the i-th residual.

The sum of the squares of these residuals or errors is the quantity that is minimised by the least squares method. This quantity, also known as the sum of squares squared to the error, is denoted by SSE.

$$SSE = \sum (y_i - \hat{y}_i)^2$$

The SSE value is a measure of the error in using the estimated regression equation to predict the values of the dependent variable in the sample. Figure 5.6 shows the calculations needed to calculate the sum of squares due to the error for the Best Burger case.

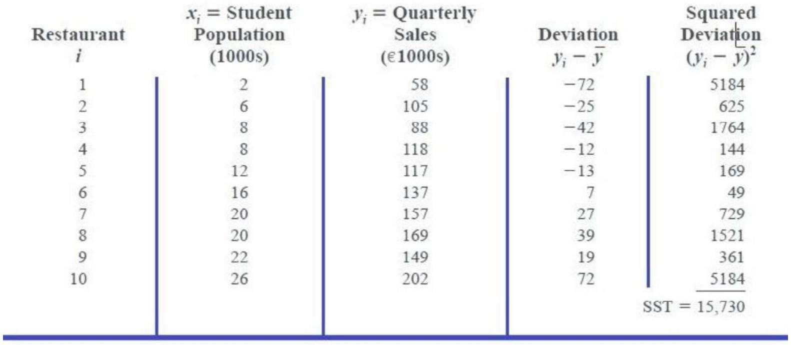


Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales (€1000s)	Predicted Sales $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Squared Error $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SSE = 1530

Figure 5.6 Squares of errors in Best Burger case.

Suppose we are asked to produce an estimate of quarterly sales without knowing the size of the student population. Without knowing any associated variables, we would use the sample average as an estimate of quarterly sales at any restaurant. Table in Figure 5.6 showed that for sales data y_i =1300. Therefore, the average quarterly sales value for a sample of 10 Best Burger restaurants is y_i/n = 1300/10 = 130. In Table 14.4 we show the sum of squares of the deviations obtained by using the sample mean of 130 to predict the value of quarterly sales for each restaurant in the sample. For the i-th restaurant in the sample, the difference y_i provides a measure of the error that is included in the application for sales forecasting. The corresponding sum of squares, called the total sum of squares, is denoted by SST.

$$SST = \sum (y_i - \bar{y})^2$$



Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales (€1000s)	Deviation $y_i - \bar{y}$	Squared Deviation $(y_i - \bar{y})^2$
1	2	58	−72	5184
2	6	105	−25	625
3	8	88	−42	1764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1521
9	22	149	19	361
10	26	202	72	5184
				SST = 15,730

Figure 5.7 Sum of squares.

The sum at the bottom of the last column in Figure 5.7 is the total sum of squares for BestBurger's restaurants SST = 15,730. In Figure 5.8 we show the estimated regression line $y = 60 + 5x$ and the line corresponding to y = 130. Note that the points cluster more closely around the estimated regression line than about the line y = 130. For example, for the 10th restaurant in the sample, we see that the error is much larger when 130 is used to predict y = 10 than when 130 is used $y = 60 + 5x$ and is 190. We can think of SST as a measure of how well the observations cluster around the line and SSE as a measure of how well the observations cluster around the line.

To measure how much the values on the estimated regression line deviate from the following, another sum of squares is calculated. This sum of squares, called the sum of squares due to regression, is denoted as SSR.

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

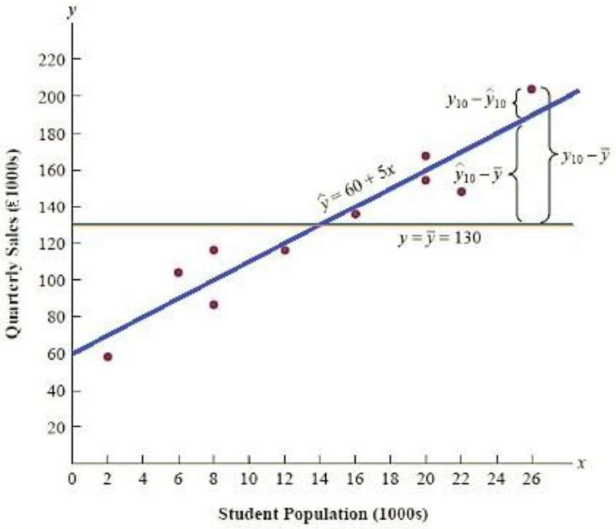


Figure 5.8 Regression line for Best Burger case.

From the previous discussion, we should expect that SST, SSR and SSE are linked. In fact, the relationship between these three sums of squares is one of the most important results in statistics.

5.6 The relationship between SST, SSR and SSE:

$$SST = SSR + SSE$$

Where it is:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

Equation ($SST = SSR + SSE$) shows that the total sum of squares can be divided into two components, sum of squares due to regression and sum of squares due to error. So if the values of any two of these sums of squares are known, the third sum of squares can be known easily by calculation. For example, in the case of Best Burger restaurants, we already know that SSE = 1530 and SST = 15,730; therefore, by solving for the SSR in equation above, we find that the sum of squares due to regression is

$$SSR = SST - SSE = 15730 - 1530 = 14200$$

Now let's see how we can use the three sums of squares, SST, SSR and SSE, to provide

a goodness-of-fit criterion for the estimated regression equation. The estimated regression equation would provide a perfect fit if each value of the dependent variable $yi\_$ would lie randomly on the estimated regression line. In this case it would be zero for each observation, resulting in SSE =0. Since SST = SSR + SSE, we see that for a perfect fit SSR must equal SST and the ratio (SSR/SST) must equal one. A worse fit will result in larger values for SSE. Solving for SSE in equation (14.11), we see that SSE = SST - SSR. Therefore, the largest value for SSE (and hence the worst fit) occurs when SSR = 0 and SSE = SST.

The SSR/SST ratio is used for the estimation, which has values between zero and one fit to the estimated regression equation.

This ratio is called the coefficient of determination and is denoted by r^2 .

$$r^2 = \frac{SSR}{SST}$$

For the example of Best Burger restaurants, the value of the coefficient of determination is

$$r^2 = \frac{SSR}{SST} = \frac{14200}{15730} = 0.9027$$

When the coefficient of determination is expressed as a percentage, we can r^2 can be interpreted as the percentage of the total sum of squares that can be explained using the estimated regression equation. For Best Burger Restaurants we can conclude that 90.27% of the total sum of squares can be explained using the estimated regression equation $y = 60 + 5x$ to predict quarterly sales. In other words, 90.27% of the variability in sales can be explained by a linear relationship between the size of the student population and sales. We should be pleased to see that it fits the estimated regression equation so well.

5.7 Correlation coefficient

The correlation coefficient can be thought of as a descriptive measure of the strength of the linear relationship between two variables, x and y. The values of the correlation coefficient are always between -1 and +1. A value of +1 means that the two variables $x\ in\ y$ are perfectly correlated in a positive linear sense. This means that all data points on a are a straight line with a positive slope. A value of -1 means that $x\ in\ y$ perfectly related in a negative linear sense, with all data points on a straight line having a negative slope. Correlation coefficient values close to zero mean that x and y are not linearly related.

If a regression analysis has already been carried out and the coefficient of determination r^2 has been calculated, the sample correlation coefficient can be calculated as follows.

$$r_{xy} = (sign\ of\ b_1)\sqrt{coefficient\ of\ determination}$$

$$r_{xy} = (sign\ of\ b_1)\sqrt{r^2}$$

PEARSON CORRELATION COEFFICIENT: SAMPLE DATA

$$rxy = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

$$rxy = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Where they are:

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}\ ,\ s_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}},\ s_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$$

The sign for the sample correlation coefficient is positive if the estimated regression equation has a positive slope ($b_1 > 0$) and negative if the estimated regression equation has

negative slope ($b_1 < 0$).

For the Best Burger case, the value of the coefficient of determination corresponding to the estimated regression equation $y = 60 + 5x$ is 0.9027. Because the slope of the estimated regression equation is positive, equation (14.13) shows that the sample correlation coefficient is By the sample correlation coefficient

Rxy=0.9501, we would conclude that a strong positive linear relationship exists between $x\ in\ y$.

In the case of a linear relationship between two variables, both coefficients of determination and the sample correlation coefficient provide a measure of the strength of the relationship.

The coefficient of determination provides a measure between zero and one, while the sample correlation coefficient provides a measure between -1 and +1. Although the sample correlation coefficient is limited to a linear relationship between two variables, the coefficient of determination can be applied to non-linear relationships and to relationships that have two or more independent variables. Thus, the coefficient of determination provides a wider range of applicability.

5.8 Multiple Regression Model

In the following sections, we continue our study of regression analysis by considering situations involving two or more independent variables. This subject area, called multiple regression analysis, allows us to take more factors into account and thus obtain better predictions than are possible with simple linear regression.

Multiple regression analysis is the study of how the dependent variable y is related to two or more independent variables. In the general case, we will denote by p the number of independent variables.

5.9 Regression model and regression equation

The concepts of regression model and regression equation introduced in the previous section apply in the case of multiple regression. The equation describing how the dependent variable y is related to the independent variables $x_1, x_2, \ldots, x_p$ and the error term is called a multiple regression model. We start by assuming that the multiple regression model has the following form.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

In a multiple regression model $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the parameters and the error term (ϵ) is a random variable. A close examination of this model reveals that y is a linear function of the variables $x_1, x_2, \ldots, x_p$ plus the error term ϵ epsilon. The error term takes into account the variability $y$which cannot be explained by the linear effect of p independent variables.

In section 5.10 we discuss the assumptions for the multiple regression model and epsilon. One of the assumptions is that the mean or expected value (ϵ) is zero. The implication of this assumption is that the mean or expected value of y, is denoted by $E(y)$, is equal to $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. The equation describing how the mean value is y is related to $x_1, x_2, \ldots, x_p$ is called a multiple regression equation.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

5.10 Estimated multiple regression equation

If the values are $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are known, equation (5.9) can be used to calculate the average value of y at given values of $x_1, x_2, \ldots, x_p$. Unfortunately, these parameter values will generally not be known and must be estimated from the sample data. A simple random sample is used to calculate the sample statistic $b_0, b_1, b_2, \ldots, b_p$to be used as point estimators of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. These sample statistics provide the following multiple regression equation estimation:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

Where they are

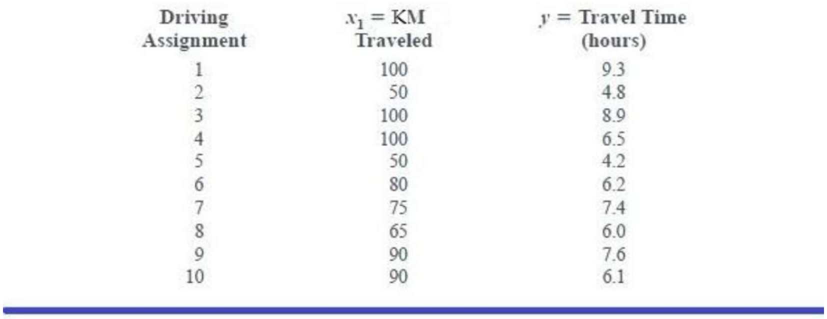
$b_0, b_1, b_2, \ldots, b_p$ are estimates $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$

$\hat{y}$ = the predicted value of the dependent variable

Example: Frigo Transport Company

As an illustration of multiple regression analysis, we will consider the problem faced by Frigo Trucking Company, an independent trucking company in Southern Italy. The largest part of Frigo's business involves deliveries throughout the local area. For better development work schedules, managers want to plan a common daily travel time for their drivers.

Initially, managers believed that the total daily journey time would be closely linked to the number of km travelled in daily deliveries. A simple random sample of 10 driver assignments provided the data shown in Figure 5.9 and a scatter plot. After reviewing this scatter diagram, the Managers assumed that a simple linear regression model could be used to describe the relationship $y = \beta_0 + \beta_1 x_1 + \epsilon$ between total journey time (y) and the number of km travelled (x_1).



Driving Assignment	x_1 = KM Traveled	y = Travel Time (hours)
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

Figure 5.9 Data for Frigo Transport Company example.

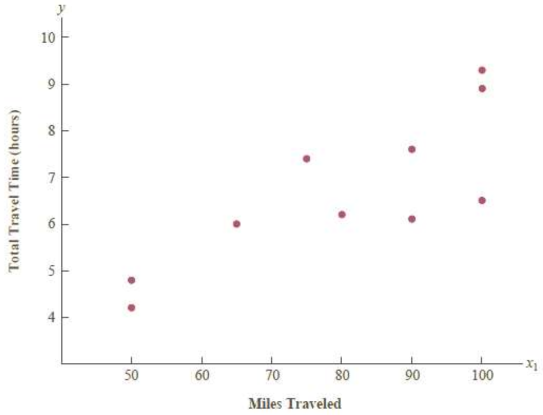


Figure 5.10 Scatter plot for Frigo Transport Company example.

To estimate the parameters β_0 and β_1 the least squares equation method was used to develop the estimated regression.

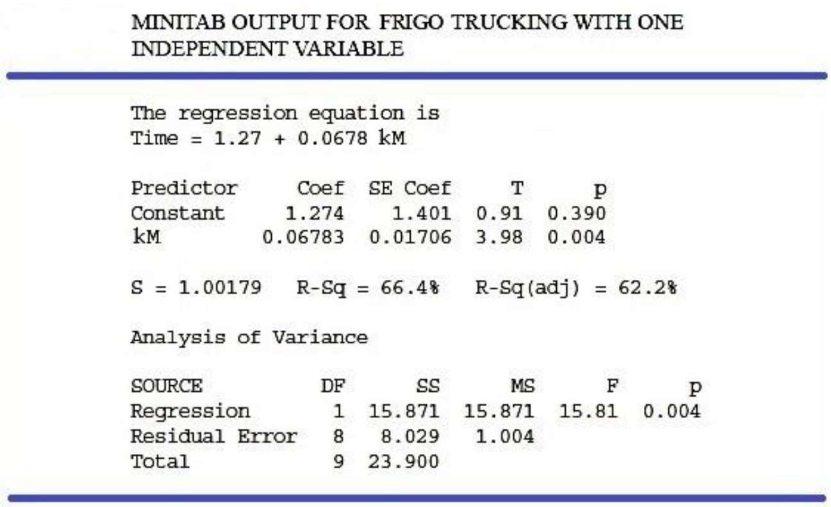
$$\hat{y} = b_0 + b_1x_1$$

Figure above shows the output of Minitab using simple linear regression to the data in Table above. The estimated regression equation is

$$\hat{y} = 1.27 + 0.0678x_1$$

At the 0.05 significance level, an F-value of 15.81 and a corresponding p-value of 0.004 indicate that the relationship is significant. This means that we can reject $H_0: \beta_1 = 0$ because the p-value is less than $\alpha = 0{,}05$. Note that the same conclusion follows from the value of $t = 3{,}98$ and the associated p-value of 0.004. Thus, we can conclude that the relationship between total journey time and number of miles travelled is significant. Longer journey times are associated with more kilometres travelled. With the coefficient of determination (expressed as a percentage) $R - Sq = 66{,}4\ \%$, we see that 66.4% of the variability in travel time can be explained by a linear effect of the number of miles travelled.

This finding is quite good, but managers may want to consider adding a second independent variable to explain some of the remaining variables in the dependent variable.



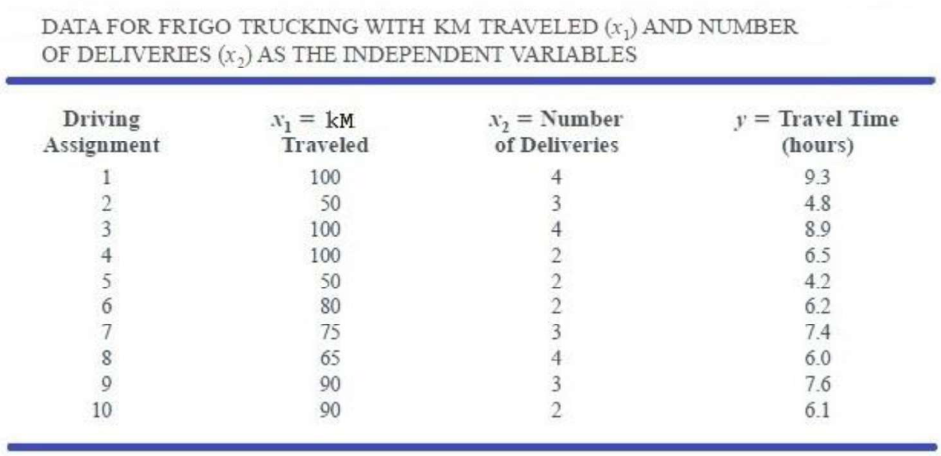
MINITAB OUTPUT FOR FRIGO TRUCKING WITH ONE INDEPENDENT VARIABLE

```
The regression equation is
Time = 1.27 + 0.0678 kM

Predictor       Coef   SE Coef     T      p
Constant       1.274     1.401  0.91  0.390
kM           0.06783   0.01706  3.98  0.004

S = 1.00179   R-Sq = 66.4%   R-Sq(adj) = 62.2%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       1  15.871  15.871  15.81  0.004
Residual Error   8   8.029   1.004
Total            9  23.900
```

Figure 5.11 Results with one independent variable.

When trying to identify the second independent variable, managers considered that the number of deliveries may also contribute to the total journey time. The Frigo Trucking data, with the number of deliveries added, is shown in Figure below. (x_1) and the number of deliveries (x_2), as independent variables, is shown in Figure 5.12. The estimated regression equation is

$$\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$$



DATA FOR FRIGO TRUCKING WITH KM TRAVELED (x_1) AND NUMBER OF DELIVERIES (x_2) AS THE INDEPENDENT VARIABLES

Driving Assignment	x_1 = kM Traveled	x_2 = Number of Deliveries	y = Travel Time (hours)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

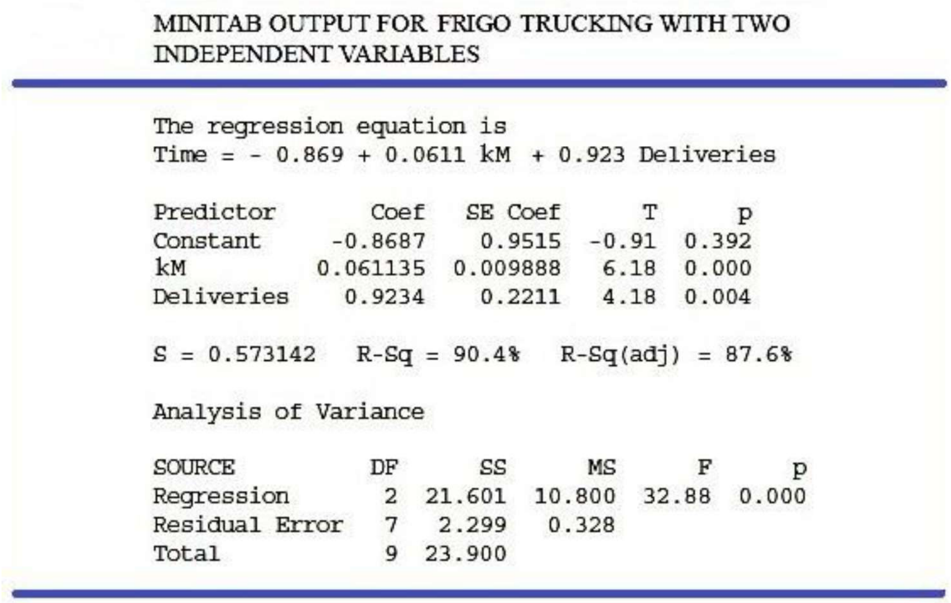
Figure 5.12 Frigo Trucing data and independent variables.

Let's take a closer look at the values b_1= 0.0611 and b_2= 0.923 in last equation.

Note on the interpretation of the coefficients

At this point we can make one comment on the relationship between the estimated regression equation with only miles travelled as the independent variable and the equation including the number of deliveries as the other independent variable. Value b_1 is not the same in both cases. In a simple linear regression we interpret b_1 as an estimate of the change y for a one-unit change in the independent variable. In multiple regression analysis this interpretation needs to be modified slightly. That is, in multiple regression analysis, each regression coefficient is interpreted as follows: would represent an estimate of the change in the y corresponding to the change in x_i by one unit when all other independent variables are held constant.

In the case of Frigo Trucking, that involves two independent variables, b_1=0.0611 and b_2= 0.923.



MINITAB OUTPUT FOR FRIGO TRUCKING WITH TWO INDEPENDENT VARIABLES

```
The regression equation is
Time = - 0.869 + 0.0611 kM  + 0.923 Deliveries

Predictor        Coef   SE Coef      T      p
Constant      -0.8687    0.9515  -0.91  0.392
kM           0.061135  0.009888   6.18  0.000
Deliveries     0.9234    0.2211   4.18  0.004

S = 0.573142   R-Sq = 90.4%   R-Sq(adj) = 87.6%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       2  21.601  10.800  32.88  0.000
Residual Error   7   2.299   0.328
Total            9  23.900
```

Figure 5.13 Results for Frigo Trucking with two independent variables.

Thus, 0.0611 hours is an estimate of the expected increase in travel time corresponding to an increase in one mile per distance travelled when the number of deliveries is constant. Similarly, since b_2 =0.923, the estimate of the expected increase in journey time corresponding to an increase of one delivery when the number of miles travelled is constant is 0.923 hours.

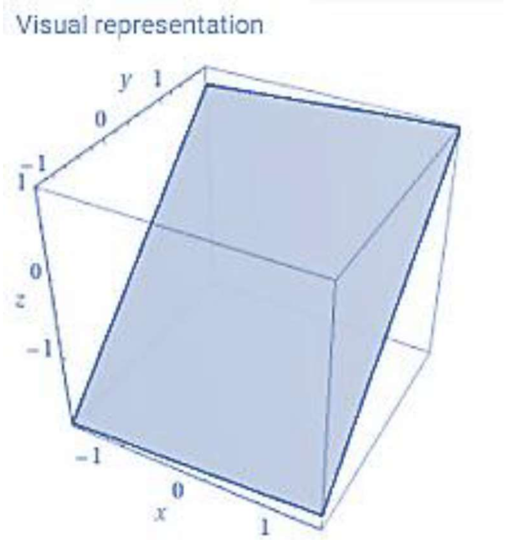


Figure 5.14 Visual representation of results for Frigo Trucking case.

References Chapter 5

- *Introductory Statistics.* Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/97898151231351230101

- Introductory Statistics 2e, Openstax, Rice University, Houston, Texas 77005, Jun 23, senior contributing authors: Barbara Illowsky and Susan dean, De anza college, Publish Date: Dec 13, 2023, (https://openstax.org/details/books/introductory-statistics-2e);

- Introductory Statistics 4th Edition, Susan Dean and Barbara Illowsky, Adapted by Riyanti Boyd & Natalia Casper (Published 2013 by OpenStax College) July 2021, (http://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf);

- Journal of the Royal Statistical Society 2024**,** A reputable journal publishing cutting-edge research and articles on various aspects of statistics, including theoretical advancements and practical applications. Recent issues have featured studies on sampling and hypothesis testing.

- Introductory Statistics 7th Edition, Prem S. Mann, eastern Connecticut state university with the help of Christopher Jay Lacke, Rowan university, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, 2011

- Introduction to statistics, made easy second edition, Prof. Dr. Hamid Al-Oklah Dr. Said Titi Mr. Tareq Alodat, March 2014

- Statistics for Business and Economics, Thirteenth Edition, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, 2017, 2015 Cengage Learning®

- Statistics for Business, First edition, Derek L Waller, 2008 Copyright © 2008, Derek L Waller, Published by Elsevier Inc. All rights reserved