Decision Modelling

Edited by T. Toronjadze

Tbilisi, Georgia, 2020

The book has been written by members of GAU Business Research Center

Teimuraz Toronjadze, Professor
Tamaz Uzunashvili, Professor
Michael Mania, Professor
Besik Chikvinidze, Professor
Revaz Tevzadze, Professor
Tsotne Kutalia, Professor
Irakli Chelidze, Professor
Levan Gachechiladze, PhD student
Beka Gogichashvili, PhD student

Technical Editor: *Maia Kvinikadze*

Contents

PREFACE

The purpose of this book is to introduce the computational solutions of decision making techniques to students, academicians and other interested parties. The book can be used to review and refresh knowledge in business forecasting, time series analysis and business modelling as well as for integration of academic achievements in the study process.

The book has been written based on Business Research Center's extended seminars.

Chapter 1. Simple Regression

1.1. Introduction

Establishment of correspondence between various types of phenomena is often crucial to make an applicable analysis of business process. Business and economics applications make extensive use of relationships between variables.

- Real estate agent may be interested in predicting the property price based on its area.
- Supermarket manager may wish to forecast the demand for a certain product given its selling price.
- Medical doctor may need to know the concentration of a certain drug in the bloodstream based on time passed after injection.

These types of relationships can mathematically be expressed as

$$Y = f(X)$$

where the function f can take linear and nonlinear forms.

In many applications, the form of the relationship is not precisely known. In some situations we are interested in the limited portion of the nonlinear relationship that can be approximated by linear relationship to some extent. Here, the primary goal is to present linear models based on least squares regression analyses. Once the linear relationship between variables is established, the next task is to measure the reliability of the model. Lastly, some coefficients measuring the strength of predictive power of the model are presented.

1.2. Correlation Analyses

The main goal of this section is to measure a linear relationship between two variables. First, the existence of linear dependence needs to be tested. As we begin our analyses, we conclude that if the pair of linearly related random variables X and Y is being considered, a scatter plot of the joint observations on this pair will tend to be clustered around a straight line. Conversely, if they are not linearly related, then the scatter plot will not follow a straight line.

Correlation coefficient has a wide range of applications in business and economics. In many applied business and economics related problems, there is an independent variable X, and a dependent variable Y, whose value depends on the value of X. In order to check the existence of linear relationship between X and Y, we test the following hypothesis

$$\begin{aligned} H_0&: \rho = 0 \\ H_1&: \rho \neq 0 \end{aligned} \qquad (1.2.1)$$

where ρ is the population correlation coefficient. The null hypothesis implies that there is no linear relationship between two random variables and the alternative hypothesis implies the opposite. As long as our interest is to test whether there is any kind of linear relationship (negative or positive), we do not concern ourselves with the sign of the coefficient. It can be shown that for a sample of n observations and in case of jointly normal distribution of the random variables X and Y, the random variable

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \tag{1.2.2}$$

follows a Student's t distribution with $n-2$ degrees of freedom. In (1.2.2), r is the sample correlation coefficient defined as

$$r = \frac{s_{xy}}{s_x s_y} \tag{1.2.3}$$

where the numerator is the sample covariance coefficient defined as

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

and the denominator of (1.2.3) is the product of sample standard deviations of X and Y given by

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}, \qquad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}.$$

where the sample means for X and Y are given respectively as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The decision rule for the hypothesis (1.2.1) is to

$$\text{reject } H_0 \text{ if } t < -t_{n-2,\alpha/2} \text{ or } t > t_{n-2,\alpha/2}. \tag{1.2.4}$$

Here, $t_{n-2,\alpha}$ is the number for which the random variable t_{n-2} satisfies

$$P\left(t_{n-2} > t_{n-2,\alpha}\right) = \alpha,$$

α is named as significance level of the test.

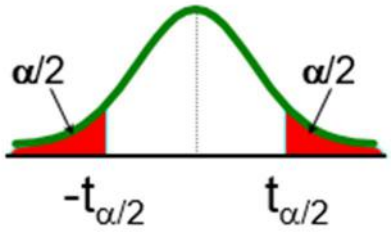


Figure 1.2.1

Figure 1.2.1 illustrates that if the t statistics computed by (1.2.2) falls within any of the shaded area called the rejection region (i.e. the condition of the decision rule (1.2.4) is met), the hypothesis (1.2.1) is rejected. The conclusion is that X and Y are linearly related. $-t_{\alpha/2}$ and $t_{\alpha/2}$ on the graph are the same as $t_{n-2,\alpha/2}$.

Hypothesis tests for positive and negative correlations

Similarly, the following hypothesis can be tested

$$H_0: \rho \geq 0$$
$$H_1: \rho < 0$$

with the decision rule

$$\text{reject } H_0 \text{ if } t < -t_{n-2,\alpha}.$$

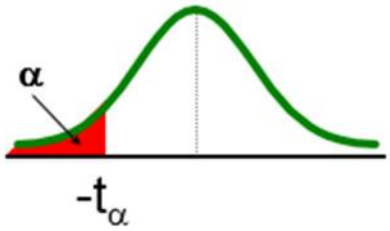


Figure 1.2.2

Or the hypothesis

$$H_0: \rho \leq 0$$
$$H_1: \rho > 0$$

with the decision rule

$$\text{reject } H_0 \text{ if } t > -t_{n-2,\alpha}.$$

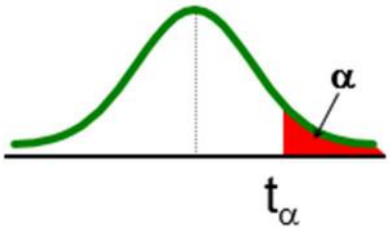


Figure 1.2.3

The difference between the hypothesis (1.2.1) and the rest of two is that in the former there is half a significance level $\alpha/2$ while in the latter there is α. Since the construction of a linear regression model makes sense when there is a linear dependence of any kind (either positive or negative), we concentrate only on the hypothesis (1.2.1).

Example 1.2

A real estate agent, Lisa Miller is concerned about the estimation of house prices. She needs a model to predict the price for a given house. She thinks that the most significant determinant of a house price is its area. So, she collects the data of houses sold. The following data in Figure 1.2.4 represents the sample of 10 observations on the independent variable - house area, X, measured in square meters, and the corresponding dependent variable – house price, Y, measured in \$1000s. (e.g. the 4th record in the data implies that the house with 1700 square meters was sold for \$302 000). Lisa decides to construct a linear model, but she realizes that the model will only be applicable if there actually is a linear relation between the house price and its area. So, the first task for her is to test the presence or absence of linear relationship between the house price and its area using the hypothesis (1.2.1). The following figure illustrates the computations

	A	B	C	D	E	F	G	H	I	J	K
1		X	Y		r	0.865866	<--"=CORREL(B2:B11,C2:C11)"				
2	1	1300	248		t	4.895369	<--"=F1*SQRT(COUNT(B2:B11)-2)/SQRT(1-F1^2)"				
3	2	2110	308		$t_{8,0.025}$	2.306004	<--"=T.INV.2T(0.05,COUNT(B2:B11)-2)"				
4	3	1935	239								
5	4	1700	302								
6	5	1050	169								
7	6	1455	223								
8	7	2250	385								
9	8	2550	367								
10	9	1765	232								
11	10	1600	245								

Figure 1.2.4

The cells F1, F2 and F3 compute the correlation coefficient r, t statistics from (1.2.2) and $t_{n-2,\alpha/2}$ respectively. The hypothesis is tested based on $\alpha = 0.05$ significance level (95% confidence level). Note that the T.INV.2T function accepts the significance level of 0.05 as an argument rather than 0.025 which is $a/2$. The reason for this is that the function itself makes the appropriate division. According to the decision rule (1.2.4), since $t = 4.8954 > t_{8,0.025} = 2.306$, it can be concluded that the null hypothesis in (1.2.1) is rejected.

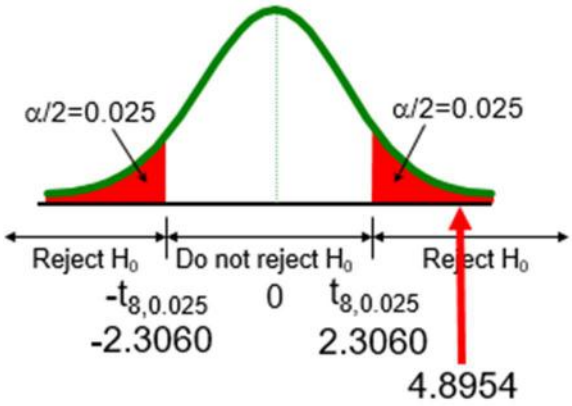


Figure 1.2.5

So, Lisa concludes that the house price is linearly related to its area. This result was quite expected as long as the sample correlation coefficient is 0.87. Therefore, it makes sense to construct a linear regression model.

1.3. Simple Regression

In a simple linear regression, we model the effect of all factors other than X (in our example, the house area) as part of random error term labeled as ε. This random error term is a random variable distributed normally with mean 0 and standard deviation σ. The linear model is

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

We assume that, for predetermined values of X, there are corresponding mean values Y plus a random term.

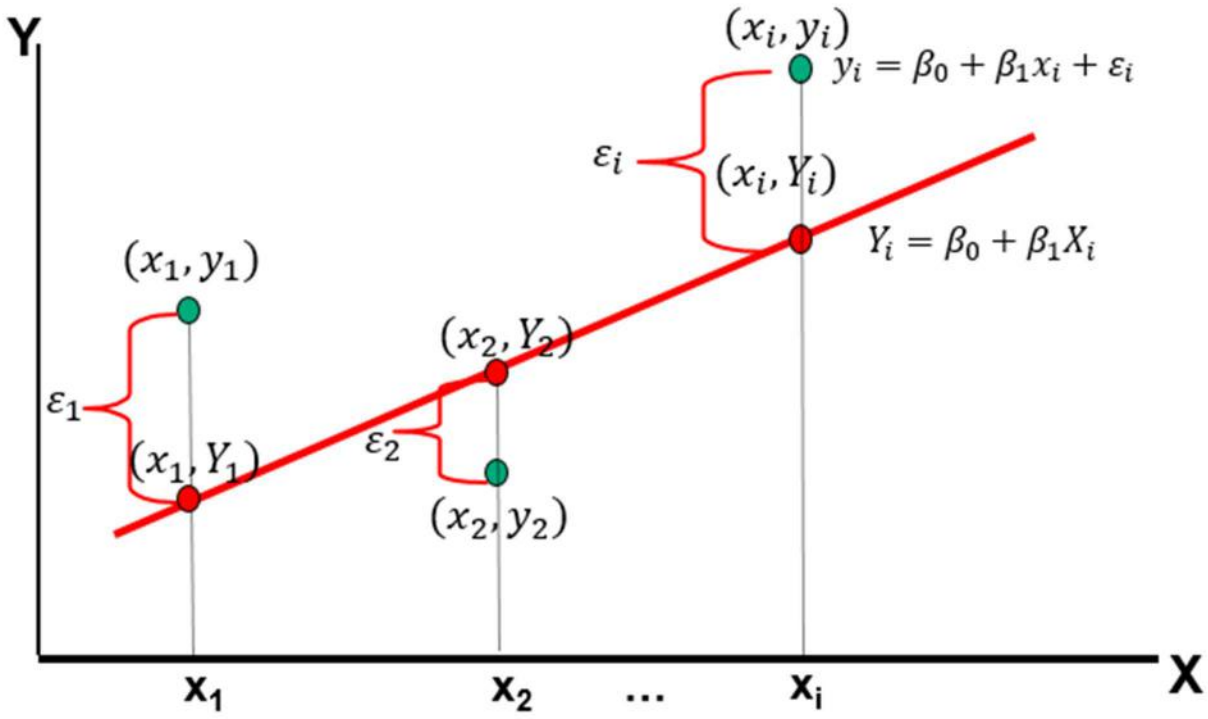


Figure 1.3.1

Figure 1.3.1 represents an example of the set of observations on the pairs of (X, Y) variables. The mean level Y for every X is represented by the population equation

$$Y = \beta_0 + \beta_1 X. \tag{1.3.1}$$

The simple linear regression model provides the mathematical expectation of the value of Y for a given value of X. Since (1.3.1) is a linear equation, the expected value of Y for a specific value of $X = x$ can be written as

$$E[Y|X = x] = \beta_0 + \beta_1 x$$

where β_0 is the Y intercept and β_1 is the slope of the line. These parameter values are unknown and must be estimated from the sample observations in a least squares sense (examined later in the text). The actual observed value of Y given the value of X is modeled as the computed value of Y plus an error term ε mentioned above. So, the actual observed value of Y can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

The estimated regression model as illustrated in Figure 1.3.2 is given by the equation

$$y_i = b_0 + b_1 x_i + e_i$$

where b_0 and b_1 are the estimated values of the coefficients and e_i is the difference between the predicted value of Y on the regression line, defined as

$$\hat{y}_i = b_0 + b_1 x_i \tag{1.3.2}$$

and the observed value y_i.

Generally, the fitted regression equation is

$$\hat{y} = b_0 + b_1 x \tag{1.3.2'}$$

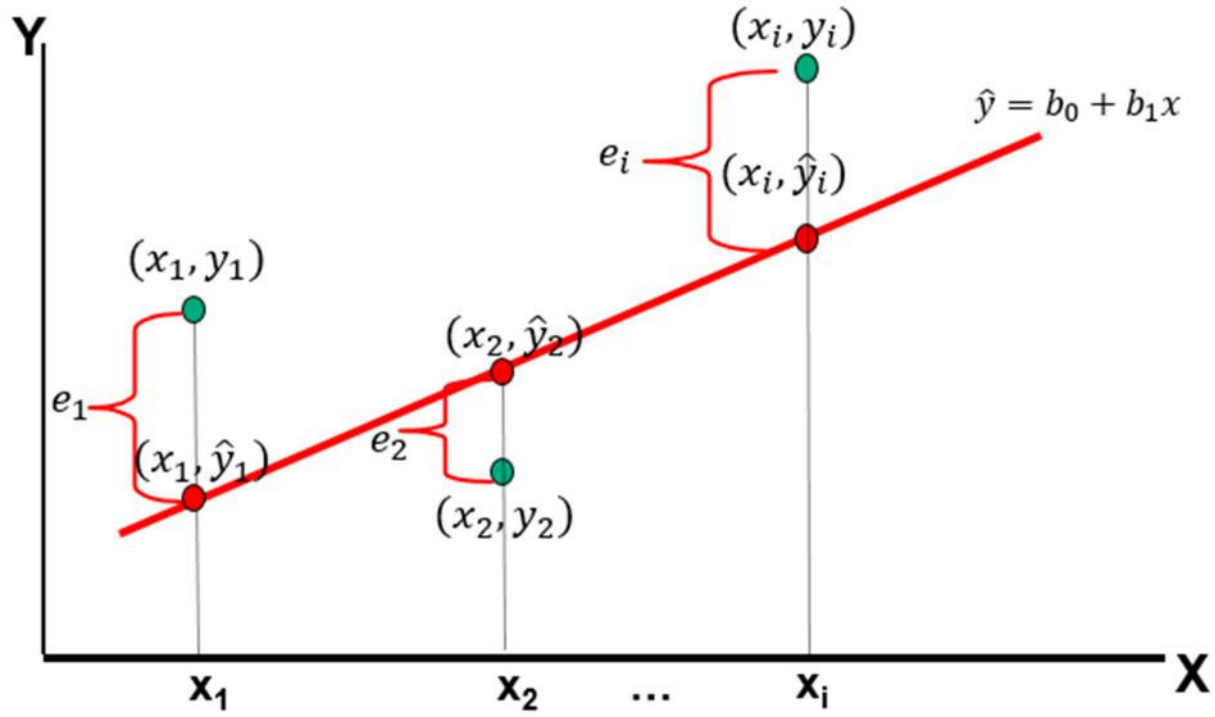


Figure 1.3.2

The difference between y_i and $\hat{y}_i$ for each value of X is defined as the residual

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

So, there is a predicted value of Y for each observed value of X. The difference between the observed value of Y and its predicted value is defined as the residual e.

The population regression line is just a theoretical construct. The model has to be estimated by the available sample data. Suppose that there are n pairs of observations, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We need to find estimators of the unknown coefficients β_0 and β_1 of the population regression line.

In order to obtain the coefficient estimators b_0 and b_1 for (1.3.2') according to the least squares procedure, the sum of squared residuals (errors) must be minimized. Let us define the sum of squared errors as the following mathematical function involving the b_0 and b_1 coefficients

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The idea behind the least squares regression is to obtain b_0 and b_1 such that SSE is minimized. Thus, the minimization procedure (which is beyond the scope of this book) yields

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \tag{1.3.3}$$

and

$$b_0 = \bar{y} - b_1 \bar{x} \tag{1.3.4}$$

Any other values of b_0 and b_1 increase the SSE. As a conclusion, the line given by the equation (1.3.2') can be interpreted as the one passing through the sample points in a "best" possible way. The "best" in the sense that the total (squared) deviation from actual observations is at a possible minimum. No other line achieves the same.

Assumptions of the population regression model

There are various assumptions regarding the population regression model that are stated below for a convenient reference.

1) The random variables Y are linear functions of X plus the random error term ε:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

2) The realizations of the random variable X are the fixed numbers $x_i (i = 1, \dots, n)$, which are independent of the error terms $\varepsilon_i (i = 1, \dots, n)$.
3) The error terms $\varepsilon_i (i = 1, \dots, n)$ are the random variables with the mean of 0 and standard deviation σ. This property is called homoscedasticity, or uniform variance:

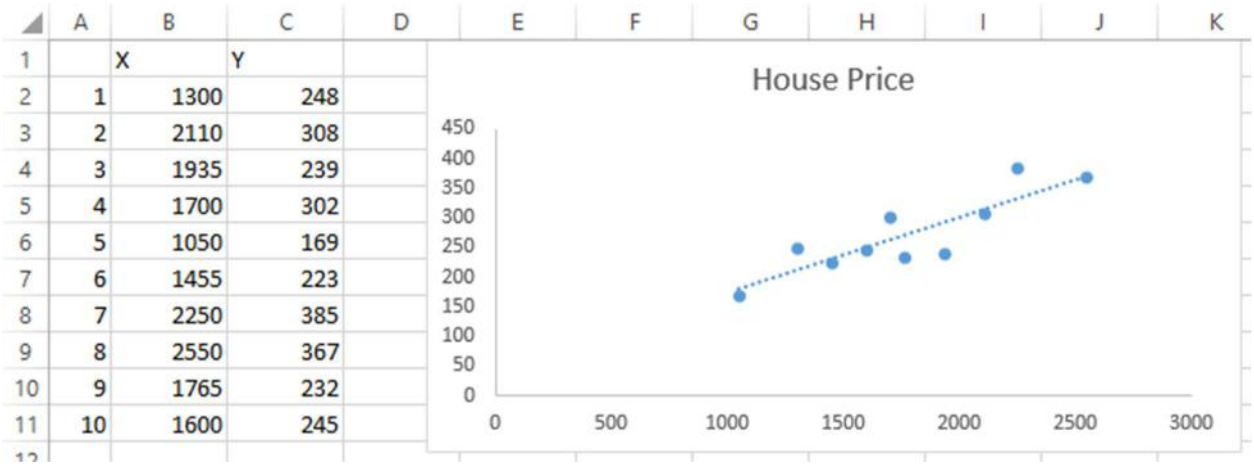
$$E\varepsilon_i = 0, \quad E\varepsilon_i^2 = \sigma^2, \quad \text{for} \quad i = 1, \dots, n$$

4) The random error terms ε_i are linearly independent of one another, so the correlation between them is 0:

$$E[\varepsilon_i \varepsilon_j] = 0, \qquad \text{for all } i \neq j$$

Example 1.3

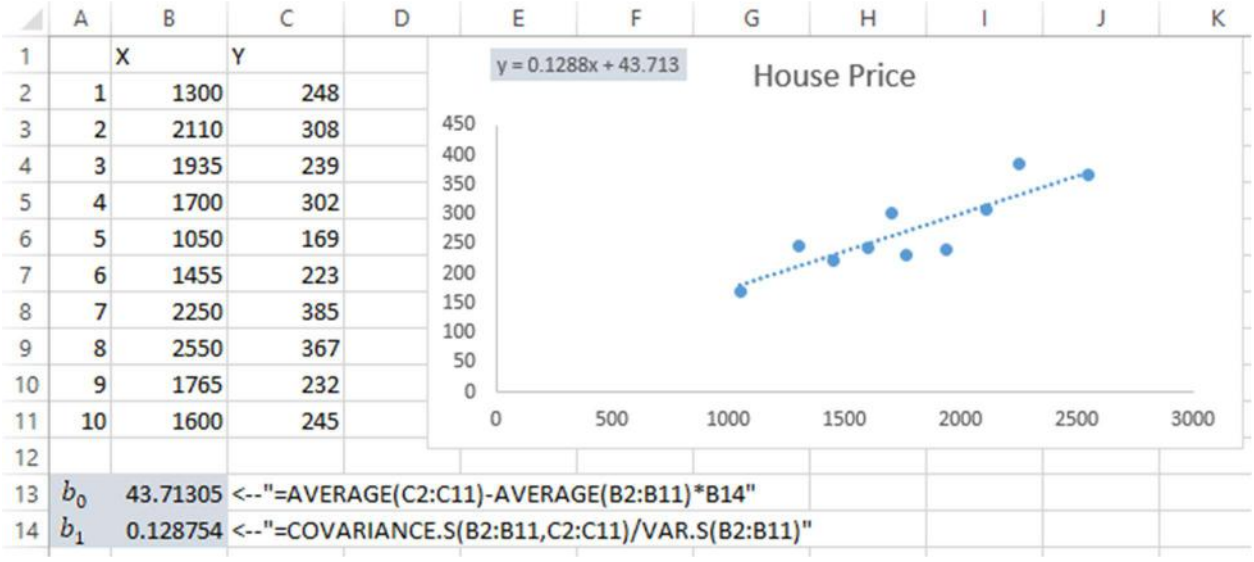
The scatter plot for the data in Figure 1.2.4.



	A	B	C
1		X	Y
2	1	1300	248
3	2	2110	308
4	3	1935	239
5	4	1700	302
6	5	1050	169
7	6	1455	223
8	7	2250	385
9	8	2550	367
10	9	1765	232
11	10	1600	245

Figure 1.3.3

For Lisa it is obvious that the scatter plot shows a positive linear pattern. Next, she computes b_0 and $b_1$coefficients based on the formulas (1.3.4) and (1.3.3) respectively. The results are shown in the cells B13 and B14 in Figure 1.3.4. Alternatively, these values could have been obtained directly from the scatter chart by right clicking on any of the points on the chart, selecting "Add Trendline" option and checking the "Display equation on chart" checkbox.



	A	B	C
1		X	Y
2	1	1300	248
3	2	2110	308
4	3	1935	239
5	4	1700	302
6	5	1050	169
7	6	1455	223
8	7	2250	385
9	8	2550	367
10	9	1765	232
11	10	1600	245
12			
13	b_0	43.71305	<--"=AVERAGE(C2:C11)-AVERAGE(B2:B11)*B14"
14	b_1	0.128754	<--"=COVARIANCE.S(B2:B11,C2:C11)/VAR.S(B2:B11)"

Figure 1.3.4

So the intercept coefficient $b_0 = 0.1288$ and the slope $b_1 = 43.713$. Thus, Lisa obtains the ultimate linear equation (1.3.2') to be

$$\hat{y} = 43.7131 + 0.1288x \qquad (1.3.5)$$

Both of these values carry important interpretations. In the context of an example of the house price depended on its price, Lisa has $b_0 = 43.7131$ which seemingly indicates that a house with 0 square feet area costs \$43.7131, which makes no sense. However, the value of b_0 just indicates that, for houses within the range of sizes observed, \$43.7131 is the portion of the house price not explained by square meter. The value of $b_1 = 0.1288$ on the other hand, tells us that the average value of a house increases by 0.1288(\$1000) = \$128.8 on average, for each additional square meter of size. b_1 can otherwise be regarded as the sensitivity coefficient. It indicates how sensitive a house price is with respect to its area. So, having the equation (1.3.5), Lisa can estimate the house price in terms of a given value of its area.

1.4. Measures of Variability

The estimated regression model developed so far explains the changes in the dependent variable Y that arise from changes in independent variable X. If there was an only random variable Y and its sample observations, then the central tendency of Y would be measured by the average value of Y being $\bar{y}$, and the total variability of the observed values of Y about $\bar{y}$ would be measured by $\sum_{i=1}^{n}(y_i - \bar{y})^2$. However, as long as there is an independent variable X whose linear function is Y, it is expected that the linear equation would be closer to the individual values of Y and therefore, the variability of the individual values of Y about the linear equation would be smaller than about the average value $\bar{y}$.

At this point, we are ready to introduce measures of variability. The analyses of variance, ANOVA, for least squares regression is developed by splitting the total variability of Y into explained and unexplained (or error) portions. The figure 1.4.1 illustrates a single observed point. The deviation of the point value from the average value $\bar{y}$ is the total variation $y_i - \bar{y}$. This variation consists of two components, the explained component by the linear equation $\hat{y}_i - \bar{y}$ and an unexplained or random component which we call the residual $e_i = y_i - \hat{y}_i$.

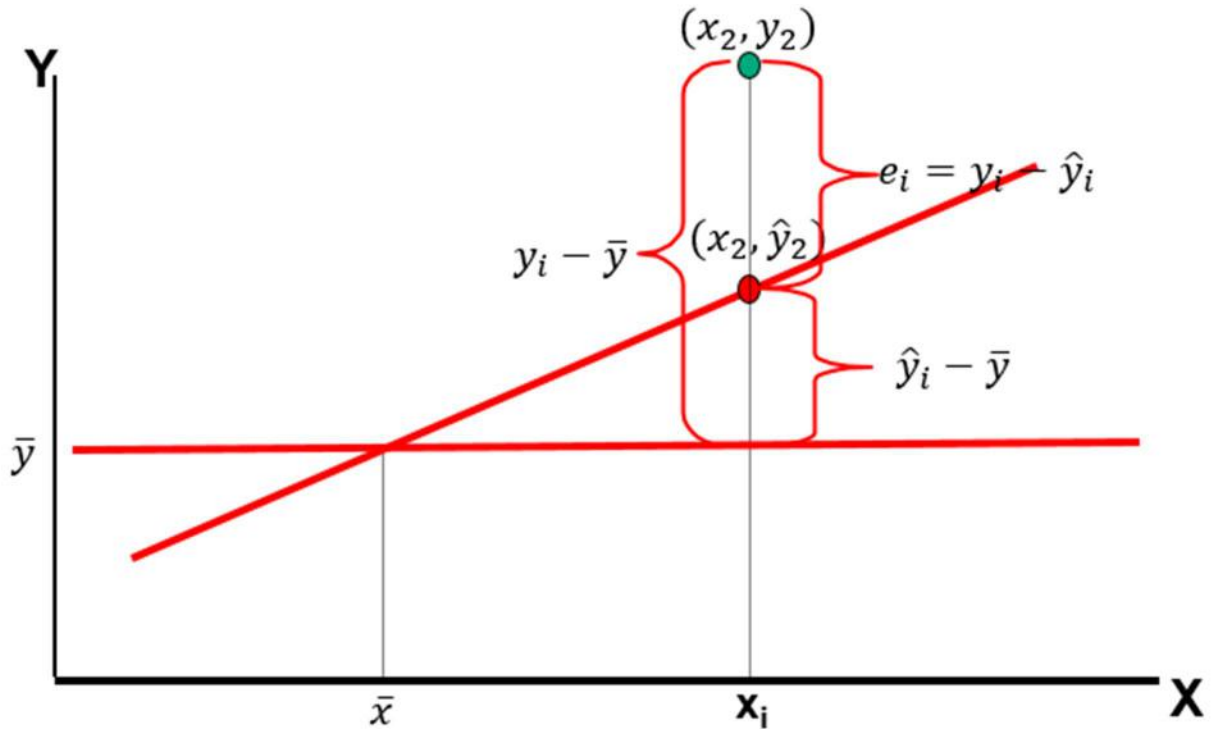


Figure 1.4.1

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

We defined these deviations just for a single point illustrated in the figure. If we sum such deviations for all observed points from the sample, we obtain

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1.4.1}$$

This equation can be expressed as

$$SST = SSR + SSE$$

Here, we see that the total variability - SST of the sample observed points about the mean can be split into an explained portion – SSR, representing the variability explained by the slope coefficient b_1, and an unexplained portion - SSE. The source of the latter is the uncertainty that arises from factors other than the explanatory variable X. The left side of the equation is the *sum of squares total*

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{1.4.2}$$

The portion of variability explained by the regression model is the *sum of squares regression* and is given by

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{1.4.3}$$

From here it is clear that the portion of variability explained by the regression depends solely on the value of b_1 coefficient and squared deviation in X. The deviations about the regression line, or the residual value, which computes the unexplained portion of variability or the *error sum of squares* can be defined as

$$SSE = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \tag{1.4.4}$$

From Figure 1.4.1, it is clear that the regression line is closer to the data point as the value of $\hat{y}_i - \bar{y}$ increases and hence, makes the value of $y_i - \hat{y}_i$ shrink. Similarly, the fit of the regression equation to the observed sample data improves if the value of SSR increases and correspondingly, the value of SSE decreases.

Example 1.4

In the previous section, Lisa Miller got the model predicting a house price for a given value of its area. So, she can use the model for the estimation of house prices. However, her colleague, Mary Wilson suggests that even though she has the model constructed, it would be better to check the reliability of the predictions made by the model. So, she now wants to determine how accurate her predictions are. She knows that changes in house area should cause changes in house price. She decides to measure the variability in house price by SSE, SSR and SST. Once having the values of b_0 and b_1 in place, she proceeds to compute the values of SSR, SSE and SST.

	A	B	C	D	E	F	G	H
1		X	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$			
2	1	1300	248	211.0927	3685.377			
3	2	2110	308	315.3831	1899.485			
4	3	1935	239	292.8512	443.1534			
5	4	1700	302	262.5941	84.74823			
6	5	1050	169	178.9043	8629.611			
7	6	1455	223	231.0495	1660.604			
8	7	2250	385	333.4086	3795.617			
9	8	2550	367	372.0347	10046.99			
10	9	1765	232	270.9631	0.700399			
11	10	1600	245	249.7188	487.581			
12								
13	b_0	43.71305		SSR	30733.86	<--"=SUM(E2:E11)"		
14	b_1	0.128754		SSE	10259.74	<--"=SUMX2MY2(C2:C11,D2:D11)"		
15				SST	40993.6	<--"=E13+E14"		

Figure 1.4.2

First, the estimated values of Y for each observed value of X are computed in the D column (by the function "=$B$13+$B$14*B2" in the cell D2). The column E contains elements of the SSR (computed by the function "=(D2-AVERAGE($C$2:$C$11))^2" in the cell E2) which when summed, gives us the value of SSR computed in the cell E13. On the other hand, in order to compute the value of SSE, as long as there are the values of Y and $\hat{y}$ listed in the columns C and D, the function $SUMXMY2$ can be used which performs the summation for the squared differences of the elements of the first and second arrays passed as arguments. The values of these measures are not self - explanatory at this point. So, Lisa's concern about the accuracy of predictions made by the model still remains unaddressed. However, having the values of SSR, SSE and SST computed, the next step is to measure the explanatory power of the regression equation using these measures of variability which helps determine the precision of the model.

1.5. The Explanatory Power of a Linear Regression Equation

In the previous section, based on the Figure 1.4.1, we claimed that greater the value of SSR, better the fit of the regression equation to the sample data. As long as $SST = SSR + SSE$, greater value of SSR implies less value of SSE in SST and thus, the explained portion of variation occupies part of unexplained portion of variation. The proportion of explained part of variation into the total variation in the dependent variable is captured by the coefficient of determination, R^2 defined as follows

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{1.5.1}$$

Closer this value is to one, greater the explained portion of variation in the dependent variable and thus, better the estimation accuracy. It can be shown that the coefficient of determination coincides the square of the correlation coefficient.

$$R^2 = r^2 \tag{1.5.2}$$

This equation has geometric interpretation. R^2 is close to one when the absolute value of the correlation coefficient is close to one. This happens when there is a strong linear relationship between two random variables and therefore, the regression line has a high explanatory power. The following Figures illustrate several cases.

Case 1: $R^2 = 1$

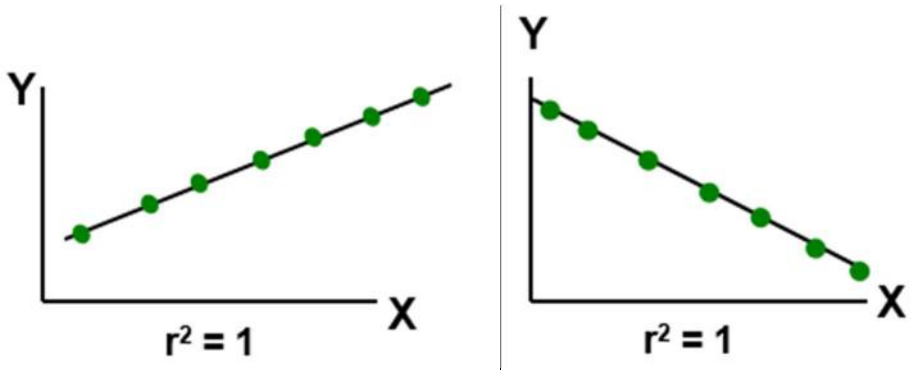


Figure 1.5.1

The coefficient of determination equals one if and only if $SSR = SST$, leaving no space for SSE. In order for SSE to be zero, there must not be a deviation from the regression line to any of the observed point. Thus, all points lie on the regression line. This implies perfect correlation.

Case 2: $0 < R^2 < 1$

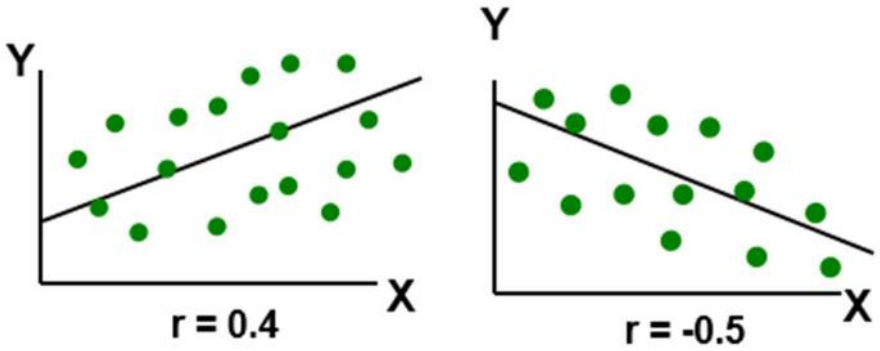


Figure 1.5.2

Figure 1.5.2 shows positive and negative correlation coefficients. In this case the R^2 is not equal to one, meaning there are some random components introducing unexplained portion of variation in the value of Y and therefore, there are some deviations from the regression line to the sample points.

Case 3: $R^2 = 0$

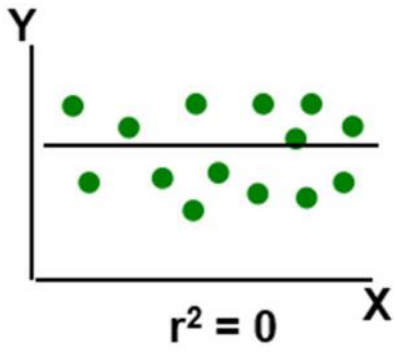


Figure 1.5.3

Correlation coefficient being zero implies the absence of linear dependence between two random variables. Similarly, the coefficient of determination being zero implies the absence of explained portion of variation in the value of the dependent variable. So, $SSR = 0$ and $SSE = SST$, meaning the random, unexplained component occupies the sum of squares total completely.

Example 1.5

In the previous example, Lisa computed the values of SSR, SSE and SST. However, she could not draw any conclusion solely based on these numbers. Now she computes the value of R^2

	A	B	C	D	E	F	G	H
1		X	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$			
2	1	1300	248	211.0927	3685.377			
3	2	2110	308	315.3831	1899.485			
4	3	1935	239	292.8512	443.1534			
5	4	1700	302	262.5941	84.74823			
6	5	1050	169	178.9043	8629.611			
7	6	1455	223	231.0495	1660.604			
8	7	2250	385	333.4086	3795.617			
9	8	2550	367	372.0347	10046.99			
10	9	1765	232	270.9631	0.700399			
11	10	1600	245	249.7188	487.581			
12								
13	b_0	43.71305		SSR	30733.86	<--"=SUM(E2:E11)"		
14	b_1	0.128754		SSE	10259.74	<--"=SUMX2MY2(C2:C11,D2:D11)"		
15				SST	40993.6	<--"=E13+E14"		
16				R^2	0.749723	<--"=E13/E15"		

Figure 1.5.4

The cell D16 contains the value of R^2 coefficient which is nearly 75%. In the context of the house prices, Lisa interprets this coefficient value as the percent of total variability in house price that is explained by the house area. So, she concludes that 75% of variation in the house price has been explained by the house area making it a significant factor determining the house price. There is still 25% unexplained. So, Lisa thinks that this is not enough to make an accurate prediction of a house price given its area. Concerned with precision of the prediction, she starts to think of other factors that might affect the house price and increase the coefficient of determination and hence, the prediction accuracy. These additional factors will be examined in the section of multiple regression.

1.6. Standard Error of the Estimate and Variance Estimators of the Regression Coefficients

In section 1.4, the *sum of squared errors* was interpreted as the unexplained portion of variation in the dependent variable Y around the mean. It can also be used to measure the variation of observed Y values from the regression line.

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2} \tag{1.6.1}$$

The square root from (1.6.1) is the standard error of the estimate $s_e = \sqrt{s_e^2}$ and measures the average dispersion of observed values of Y about the regression line. This effect is illustrated in the following figure

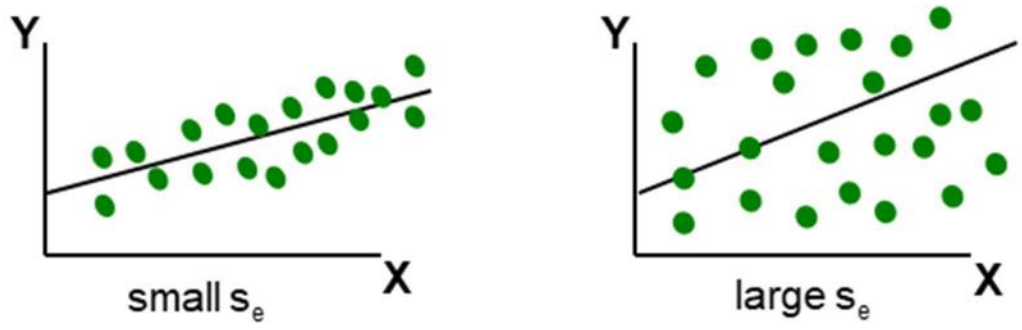


Figure 1.6.1

Small s_e implies the values of Y from the sample observations closely scattered around the regression line, while the large value of s_e implies the opposite. Note that the value of s_e is not self - explanatory by itself. Its value cannot tell us whether it is small or large. The magnitude of s_e should always be judged relative to the size of the Y values in the sample.

In section 1.3, b_1 was defined to be an unbiased estimator for β_1. The variance for this slope coefficient of the regression line is estimated by

$$s_{b_1}^2 = \frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_e^2}{(n-1)s_x^2} \tag{1.6.2}$$

while the population variance is

$$\sigma_{b_1}^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)s_x^2}.$$

The square root of (1.6.2), $s_{b_1} = \sqrt{s_{b_1}^2}$ is a measure of variation in the slope of regression lines from different possible samples. Smaller value of s_{b_1} implies a more precise estimate of the β_1 coefficient by b_1 and vice versa. This fact is illustrated in the following figure

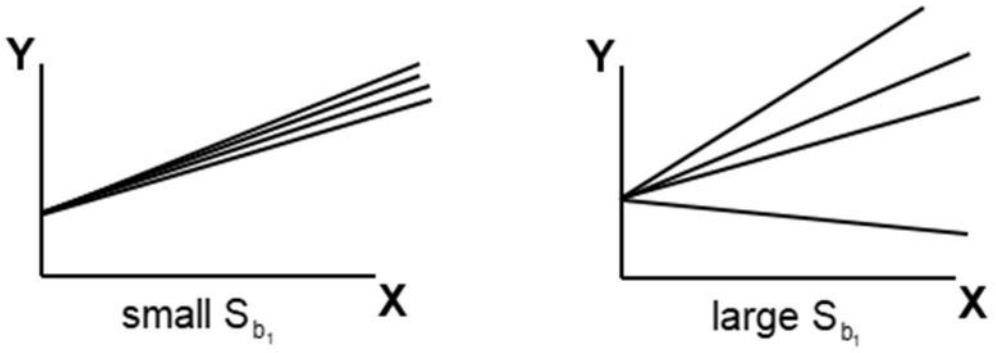


Figure 1.6.2

Similarly, the variance estimator for the regression intercept coefficient b_0 can be derived as

$$s^2_{b_0} = \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right) s_e^2 \tag{1.6.3}$$

Example 1.6

Lisa Miller computes the standard error of the estimate, the sample variance estimator and the variance estimator for the slope and intercept coefficients as follows

	A	B	C	D	E	F	G	H
1		X	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$			
2	1	1300	248	211.0927	3685.377			
3	2	2110	308	315.3831	1899.485			
4	3	1935	239	292.8512	443.1534			
5	4	1700	302	262.5941	84.74823			
6	5	1050	169	178.9043	8629.611			
7	6	1455	223	231.0495	1660.604			
8	7	2250	385	333.4086	3795.617			
9	8	2550	367	372.0347	10046.99			
10	9	1765	232	270.9631	0.700399			
11	10	1600	245	249.7188	487.581			
12								
13	b_0	43.71305		SSR	30733.86	<--"=SUM(E2:E11)"		
14	b_1	0.128754		SSE	10259.74	<--"=SUMX2MY2(C2:C11,D2:D11)"		
15				SST	40993.6	<--"=E13+E14"		
16				R^2	0.749723	<--"=E13/E15"		
17								
18	s_{b_0}	47.9489		s_e^2	1282.467	<--=E14/(COUNT(C2:C11)-2)"		
19	s_{b_1}	0.026301		s_e	35.81155	<--"=SQRT(E18)"		

Figure 1.6.3

The cells E18 and E19 contain the values of s_e^2 and s_e respectively. As noted above, the value of s_e itself cannot tell whether it is large or small. Lisa interprets the value of s_e (which is 35.81) as the average dispersion from the regression line of the sample observations and concludes that it is relatively small compared to the values of Y. So, the observed values are scattered close around the regression line. The cell B18 computes the value of s_{b_0} according to (1.6.3) contains the formula "=SQRT((1/A11+AVERAGE(B2:B11)^2/(A10*VAR.S(B2:B11)))*E18)" and the cell B19 computes the value of s_{b_1} based on (1.6.2) by the formula "SQRT(E18/((COUNT(C2:C11)-1)*VAR.S(B2:B11)))".

1.7. Hypothesis for the Regression Slope Coefficient

In section 1.3, the hypothesis for the population correlation coefficient was tested. The null hypothesis was formulated as $H_0: \rho = 0$ against the alternative $H_1: \rho \neq 0$. Rejection of the null hypothesis implies that the two random variables are linearly related. Existence of the linear relationship between two random variables can similarly be tested by the following hypothesis

$$\begin{aligned} &H_0: \beta_1 = 0 \ (no\ linear\ relationship) \\ &H_1: \beta_1 \neq 0 \quad\ \ (linear\ relationship) \end{aligned} \tag{1.7.1}$$

Under fairly general conditions, it can be concluded that the random variable

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1 - 0}{s_{b_1}} = \frac{b_1}{s_{b_1}} \tag{1.7.2}$$

follows the Student's t distribution with $n - 2$ degrees of freedom. So, the decision rule for the hypothesis is to

$$\text{reject } H_0 \text{ if } t \geq t_{n-2,\alpha/2} \quad \text{or} \quad t \leq -t_{n-2,\alpha/2} \tag{1.7.3}$$

In (1.7.2), β_1 is taken to be 0 since in the hypothesis (1.7.1), its value is tested against 0. Similarly, the hypothesis can be tested for any other value of β_1. In addition, there can be the hypothesis involving inequalities. Taking the specific value of the slope coefficient β_1^* and the significance level α, all cases are summarized below for a convenient reference:

Case 1: To test the null hypothesis

$$H_0: \beta_1 = \beta_1^* \ or\ H_0: \beta_1 \leq \beta_1^*$$

against the alternative

$$H_1: \beta_1 > \beta_1^*$$

the decision rule is to

$$reject\ H_0\ if\ t = \frac{b_1 - \beta_1^*}{s_{b_1}} \geq t_{n-2,\alpha}$$

Case 2: To test the null hypothesis

$$H_0: \beta_1 = \beta_1^*\ or\ H_0: \beta_1 \geq \beta_1^*$$

against the alternative

$$H_1: \beta_1 < \beta_1^*$$

the decision rule is to

$$reject\ H_0\ if\ t = \frac{b_1 - \beta_1^*}{s_{b_1}} \leq -t_{n-2,\alpha}$$

Case 3: To test the null hypothesis

$$H_0: \beta_1 = \beta_1^*$$

against the alternative

$$H_1: \beta_1 \neq \beta_1^*$$

the decision rule is to

$$reject\ H_0\ if\ t = \frac{b_1 - \beta_1^*}{s_{b_1}} > t_{n-2,\alpha/2}\ or\ t = \frac{b_1 - \beta_1^*}{s_{b_1}} \leq -t_{n-2,\alpha/2}$$

The last one is the generalized version of (1.7.1) with β_1^* being any number (not necessarily 0).

Remark: At this point, there are two hypotheses at our disposal testing the linear dependence of two random variables. So, there is no difference in which one is used when it comes to simple regression with a single explanatory variable. However, as we will see later, in multiple regression, the hypothesis (1.7.1) can be tested for any independent variable separately. Hence, it answers the question whether the dependent variable Y is in linear relation with that specific independent variable.

In simple regression, there is another way of testing the linear dependence using the F test. The hypothesis is similar to (1.7.1)

$$H_0: \beta_1 = 0\ (no\ linear\ relationship)$$
$$H_1: \beta_1 \neq 0\quad (linear\ relationship)$$

It can be shown that

$$F = \frac{MSR}{MSE} \tag{1.7.4}$$

follows and an F distribution with the numerator degrees of freedom of 1 and the denominator degrees of freedom $n-2$. (Note that the F distribution is characterized by two degrees of freedom). In (1.7.4) *mean squared regression* is defined to be

$$MSR = \frac{SSR}{k} \tag{1.7.5}$$

where k is the number of explanatory variables in the regression model. Since the simple regression has only one explanatory variable ($k = 1$), the value of $MSR = SSR$. On the other hand, *mean squared error* is defined to be

$$MSE = \frac{SSE}{n-k-1} = \frac{SSE}{n-2} = s_e^2 \tag{1.7.6}$$

So, in simple regression (1.7.4) can be rewritten as

$$F = \frac{SSR}{s_e^2}$$

The decision rule for the hypothesis is to

$$\text{reject } H_0 \text{ if } F \geq F_{1,n-2,\alpha} \tag{1.7.7}$$

where $F_{1,n-2,\alpha}$ is the critical value corresponding to α significance level satisfying

$$P\left(F_{1,n-2} > F_{1,n-2,\alpha}\right) = \alpha$$

Example 1.7

Similarly to the Example 1.2, where the hypothesis was tested for the population correlation coefficient, here the hypothesis (1.7.1) is tested. As long as both of the hypothesis carry the same interpretation, in particular, rejection of the null hypothesis means the existence of linear relation between the random variables X and Y, it is quite expected that both hypotheses give the same answer.

	A	B	C	D	E	F	G	H	I	J
1		X	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$		F	23.96464	<--"=E13/E18"	
2	1	1300	248	211.0927	3685.377		$F_{1,8,0.05}$	5.317655	<--"=F.INV(0.95,A2,A9)"	
3	2	2110	308	315.3831	1899.485					
4	3	1935	239	292.8512	443.1534					
5	4	1700	302	262.5941	84.74823					
6	5	1050	169	178.9043	8629.611					
7	6	1455	223	231.0495	1660.604					
8	7	2250	385	333.4086	3795.617					
9	8	2550	367	372.0347	10046.99					
10	9	1765	232	270.9631	0.700399					
11	10	1600	245	249.7188	487.581					
12										
13	b_0	43.71305		SSR	30733.86	<--"=SUM(E2:E11)"				
14	b_1	0.128754		SSE	10259.74	<--"=SUMX2MY2(C2:C11,D2:D11)"				
15				SST	40993.6	<--"=E13+E14"				
16				R^2	0.749723	<--"=E13/E15"				
17										
18	s_{b_0}	47.9489		s_e^2	1282.467	<--=E14/(COUNT(C2:C11)-2)"				
19	s_{b_1}	0.026301		s_e	35.81155	<--"=SQRT(E18)"				
20										
21				t	4.895369	<--"=B14/B17"				
22				$t_{8,0.025}$	2.306004	<--"=T.INV.2T(0.05,A9)"				

Figure 1.7.1

The hypothesis (1.7.1) is tested by the significance level of $\alpha = 0.05$. The cells E21 and E22 contain the values of the t statistics from (1.7.2) and $t_{8,0.025}$. Note that the value of t statistics exactly matches the one computed for the hypothesis (1.2.1). It can be shown that the value of t statistics computed by (1.2.2) and (1.7.2) are equal. This computation was illustrated in Figure 1.2.4, cell F2.

The cell H1 contains the value of the F statistics and the cell H2 contains the critical value $F_{1,8,0.05}$. Note that unlike the T.INV.2T function in the cell E22, the function F.INV in the cell H2 receives $1 - \alpha = 0.95$ as an argument. The rest of the two arguments are the numerator degrees of freedom (equal to 1) and the denominator degrees of freedom (equal to 8).

1.8. Hypothesis for the Regression Slope Coefficient Tested by p-value

As an alternative to the techniques examined in Section 1.7, all of the hypotheses can be tested by comparing probabilities rather than comparing the corresponding statistics (t or F) to the critical values ($t_{n-2,\alpha/2}$ or $F_{1,n-2,\alpha}$).

Consider the hypothesis (1.7.1) which is restated below

$$H_0: \beta_1 = 0 \ (no\ linear\ relationship)$$
$$H_1: \beta_1 \neq 0 \quad (linear\ relationship)$$

The t statistics defined by (1.7.2) is

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1 - 0}{s_{b_1}} = \frac{b_1}{s_{b_1}}$$

According to the rejection rule (1.7.3), this value must be compared to the critical value $t_{n-2,\alpha/2}$ and decide to reject H_0 if $|t| > |t_{n-2}, \alpha/2|$. This fact can be illustrated as follows

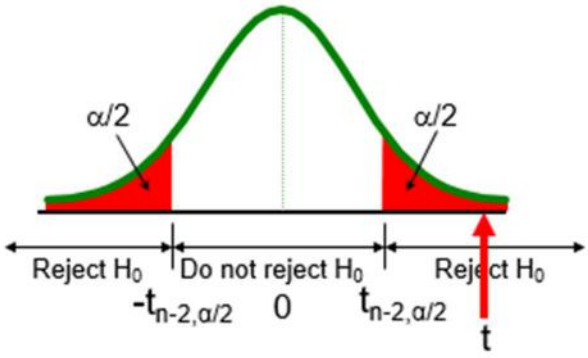


Figure 1.8.1

In the figure above, $t > t_{n-2,\alpha/2}$. Thus, the t value falls within one of the rejection regions and the null hypothesis H_0 must be rejected. In this case we have $P(t_{n-2} > t_{n-2,\alpha/2}) = \alpha/2$ and $P(t_{n-2} > t) < \alpha/2$. So, the area under the density function from t to the right is smaller than the shaded area which is $\alpha/2$. So, instead of comparing the values of t and $t_{n-2,\alpha/2}$, we can compare the probabilities (areas under the curve of the probability density function). Therefore, the decision rule (1.7.3) can be translated as

$$\text{reject } H_0 \text{ if } P(t_{n-2} > t) < \alpha/2.$$

More formally, (for positive t) we can define *probability value (p-value)* as

$$p - value = 2P(t_{n-2} > t) \tag{1.8.1}$$

and this value needs to be compared with the significance level α and we have the final version of the decision rule:

$$\text{reject } H_0 \text{ if } p - value < \alpha \tag{1.8.2}$$

and this is equivalent to (1.7.3). Note that in the discussion above, the assumption above is the positivity of value of t.

On the other hand, if t is positive and $t < t_{n-2,\alpha/2}$, then the area to the right of t would be greater than the area to the right of $t_{n-2,\alpha/2}$ and according to (1.7.3) the hypothesis H_0 would not be rejected. This fact is illustrated on the Figure 1.8.2 below

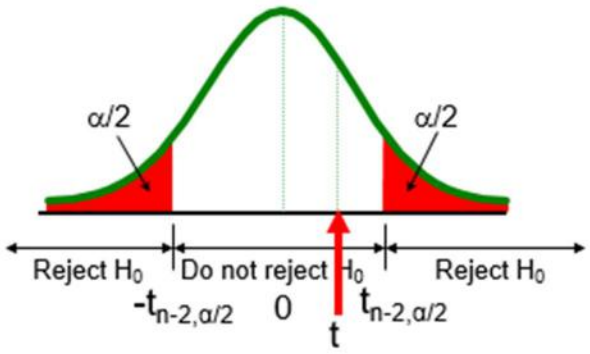


Figure 1.8.2

In this case (1.8.2) does not hold and H_0 is not rejected. The convenience of using p-values rather than comparison of t values is clear when using a statistical computer package (as shown in Section 1.10) where the p-value for a given hypothesis is generated by a computer. What remains is to simply compare this value to the significance level α.

In section 1.7, the same hypothesis was tested with the F test. The coefficient F computed by (1.7.4) is compared with the critical value $F_{1,n-2,\alpha}$ and the decision rule (1.7.7) states to

$$\text{reject } H_0 \text{ if } F \geq F_{1,n-2,\alpha}$$

which is equivalent to

$$\text{reject } H_0 \text{ if } p - value = P\left(F_{1,n-2} > F\right) < P\left(F_{1,n-2} > F_{1,n-2,\alpha}\right) = \alpha \tag{1.8.3}$$

The following figure illustrates the condition of (1.8.3) met

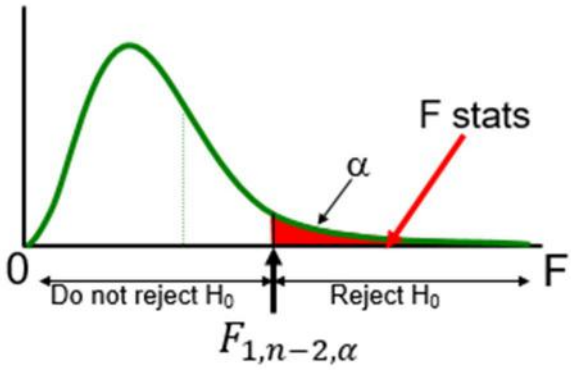


Figure 1.8.3

On the other hand, if the F statistics falls within the non-rejection region (i.e. $F < F_{1,n-2,\alpha}$), this implies that $p - value = P(F_{1,n-2} > F) > P(F_{1,n-2} > F_{1,n-2,\alpha}) = \alpha$ and is illustrated below

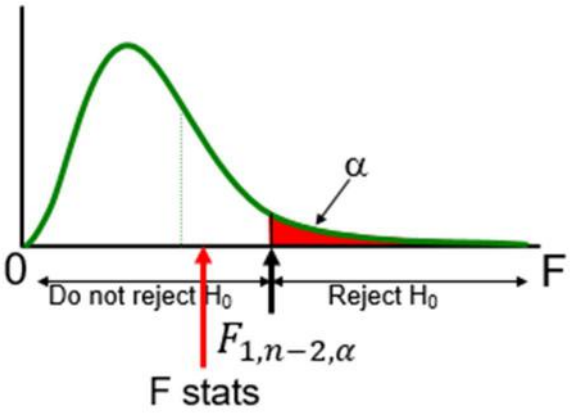


Figure 1.8.4

Similarly to the hypothesis test related to Student's t distribution, the p-value can directly be compared to the significance level α.

Example 1.8

The figure below illustrates the computations of p-values in both ways.

	A	B	C	D	E	F	G	H	I	J
1		X	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$		F	23.96464	<--"=E13/E18"	
2	1	1300	248	211.0927	3685.377		$F_{1,8,0.05}$	5.317655	<--"=F.INV(0.95,A2,A9)"	
3	2	2110	308	315.3831	1899.485		p-value	0.001201	<--"=F.DIST.RT(S16,1,8)"	
4	3	1935	239	292.8512	443.1534					
5	4	1700	302	262.5941	84.74823					
6	5	1050	169	178.9043	8629.611					
7	6	1455	223	231.0495	1660.604					
8	7	2250	385	333.4086	3795.617					
9	8	2550	367	372.0347	10046.99					
10	9	1765	232	270.9631	0.700399					
11	10	1600	245	249.7188	487.581					
12										
13	b_0	43.71305		SSR	30733.86	<--"=SUM(E2:E11)"				
14	b_1	0.128754		SSE	10259.74	<--"=SUMX2MY2(C2:C11,D2:D11)"				
15				SST	40993.6	<--"=E13+E14"				
16				R^2	0.749723	<--"=E13/E15"				
17										
18	s_{b_0}	47.9489		s_e^2	1282.467	<--=E14/(COUNT(C2:C11)-2)"				
19	s_{b_1}	0.026301		s_e	35.81155	<--"=SQRT(E18)"				
20										
21				t	4.895369	<--"=B14/B17"				
22				$t_{8,0.025}$	2.306004	<--"=T.INV.2T(0.05,A9)"				
23				p-value	0.001201	<--"=T.DIST.2T(E21,A9)"				

Figure 1.8.5

The p-value (1.8.1) is computed in the cell E23 by the function T.DIST.2T(E21,A9). This function receives the value of t and the degrees of freedom as arguments and returns the probability in (1.8.1). Similarly, the p-value (1.8.3) is computed in the cell H3 by the function F.DIST.RT(S16,1,8). This function receives the F value, first degrees of freedom and second degrees of freedom as arguments and returns the probability in (1.8.3). As long as there is only one explanatory variable in the regression model, these two values are the same.

1.9. Confidence Intervals

In Section 1.3, we defined b_0 and b_1 to be unbiased estimators of β_0 and β_1 respectively. So, the most likely values of the population intercept and slope coefficients are their sample estimates. However, it would be useful to know by high probability the interval within which the actual population intercept and slope coefficients will fall. The confidence intervals for β_0 and β_1 corresponding to a given confidence level $1-\alpha$ are

$$b_0 - t_{n-2,\alpha/2} s_{b_0} < \beta_0 < b_0 + t_{n-2,\alpha/2} s_{b_0} \tag{1.9.1}$$

$$b_1 - t_{n-2,\alpha/2}s_{b_1} < \beta_1 < b_1 + t_{n-2,\alpha/2}s_{b_1} \tag{1.9.2}$$

Once having this information, one can make predictions or forecasts for the dependent variable for a given value of the independent variable. For a specific value of the independent variable x_{n+1}, the corresponding forecast value of the dependent variable is

$$y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}$$

which has an expectation of

$$E[y_{n+1}|x_{n+1}] = \beta_0 + \beta_1 x_{n+1}$$

Suppose that the population regression model and the standard assumptions examined in Section 1.3 hold. Let b_0 and b_1 be the least squares estimates of β_0 and β_1, based on the sample observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Then it can be shown that the following are confidence intervals corresponding to $100(1-\alpha)\%$ confidence level.

1. For the forecast of the actual value resulting for Y_{n+1}, the prediction interval is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2}s_e\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{1.9.3}$$

So, the most likely value of the dependent variable for a specific value of the independent variable x_{n+1} is $\hat{y}_{n+1}$. However, the actual value of the dependent variable will fall within the interval of (1.9.3) by $100(1-\alpha)\%$ confidence level.

2. For the forecast of the conditional expectation $E[Y_{n+1}|X_{n+1}]$, the confidence interval is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2}s_e\sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{1.9.4}$$

Here the confidence interval is constructed for the average value of the dependent variable for a fixed value of the independent variable x_{n+1}. The average value of Y will fall within the interval given by (1.9.4) by $100(1-\alpha)\%$ confidence level.

All of the variables appearing in formulas (1.9.3) and (1.9.4) are examined above in the text.

Example 1.9

Lisa Miller continues gathering information about the model and constructs the confidence intervals for β_0 and β_1. The results are shown in Figure 1.9.1

	A	B	C	D	E	F	G	H	I	J
1		X	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$		F	23.96464	<--"=E13/E18"	
2	1	1300	248	211.0927	3685.377		$F_{1,8,0.05}$	5.317655	<--"=F.INV(0.95,A2,A9)"	
3	2	2110	308	315.3831	1899.485		p-value	0.001201	<--"=F.DIST.RT(S16,1,8)"	
4	3	1935	239	292.8512	443.1534					
5	4	1700	302	262.5941	84.74823		LCL		UCL	
6	5	1050	169	178.9043	8629.611		-66.8573	β_0	154.2834	
7	6	1455	223	231.0495	1660.604		0.068103	β_1	0.189404	
8	7	2250	385	333.4086	3795.617					
9	8	2550	367	372.0347	10046.99					
10	9	1765	232	270.9631	0.700399					
11	10	1600	245	249.7188	487.581					
12										
13	b_0	43.71305		SSR	30733.86	<--"=SUM(E2:E11)"				
14	b_1	0.128754		SSE	10259.74	<--"=SUMX2MY2(C2:C11,D2:D11)"				
15				SST	40993.6	<--"=E13+E14"				
16				R^2	0.749723	<--"=E13/E15"				
17										
18	s_{b_0}	47.9489		s_e^2	1282.467	<--=E14/(COUNT(C2:C11)-2)"				
19	s_{b_1}	0.026301		s_e	35.81155	<--"=SQRT(E18)"				
20										
21				t	4.895369	<--"=B14/B17"				
22				$t_{8,0.025}$	2.306004	<--"=T.INV.2T(0.05,A9)"				
23				p-value	0.001201	<--"=T.DIST.2T(E21,A9)"				

Figure 1.9.1

The cell G6 and G7 contain the lower confidence levels (LCL) and the cells I6 and I7 contain the upper confidence levels (UCL) for β_0 and β_1 respectively. Lisa interprets these levels as the minimum and maximum values β_0 and β_1 can obtain in $100(1-\alpha)\%$ of times. When it comes to the house example where the house price is explained by the area as an independent variable, the lower confidence level of -66.8573 (a negative value) makes no sense. In Section 1.3, b_0 was defined to be the portion of the value of Y unexplained by a given X variable. Similarly, the negative value for the house price (which makes no sense) can be regarded as the value that arises from insufficiency of explanatory power of the X variable.

The confidence interval for these coefficients provide useful information about the reliability of the model. Now Lisa is interested to predict the price of a house with 1550 square meters of area.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		X	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$	$(x-\bar{x})^2$				$\sum_{i=1}^{10}(x_i-\bar{x})^2$	1853953	<--"=SUM(F2:F11)"	
2	1	1300	248	211.0927	3685.377	222312.3							
3	2	2110	308	315.3831	1899.485	114582.3							
4	3	1935	239	292.8512	443.1534	26732.25							
5	4	1700	302	262.5941	84.74823	5112.25				$\hat{x}_{11}$	$\hat{y}_{11}$		
6	5	1050	169	178.9043	8629.611	520562.3				1550	243.2811	<--"=B13+B14*J6"	
7	6	1455	223	231.0495	1660.604	100172.3							
8	7	2250	385	333.4086	3795.617	228962.3				LCL		UCL	
9	8	2550	367	372.0347	10046.99	606062.3				155.6331	Y_{11}	330.929	
10	9	1765	232	270.9631	0.700399	42.25				213.9137	$E[Y_{11}\|X_{11}]$	272.6485	
11	10	1600	245	249.7188	487.581	29412.25							
12													
13	b_0	43.71305		SSR	30733.86	<--"=SUM(E2:E11)"							
14	b_1	0.128754		SSE	10259.74	<--"=SUMX2MY2(C2:C11,D2:D11)"							
15				SST	40993.6	<--"=E13+E14"							
16				R^2	0.749723	<--"=E13/E15"							
17													
18	s_{b_0}	47.9489		s_e^2	1282.467	<--=E14/(COUNT(C2:C11)-2)"							
19	s_{b_1}	0.026301		s_e	35.81155	<--"=SQRT(E18)"							
20													
21				t	4.895369	<--"=B14/B17"							
22				$t_{8,0.025}$	2.306004	<--"=T.INV.2T(0.05,A9)"							
23				p-value	0.001201	<--"=T.DIST.2T(E21,A9)"							

Figure 1.9.2

So, for the given area of $x_{11} = 1550$ square meters, Lisa computes the predicted value of price to be \$243.28 in the cell K6 containing the formula "=B13+B14*J6".

Since the formulas (1.9.3) and (1.9.4) contain $\sum_{i=1}^{n}(x_i - \bar{x})^2$, it is more convenient to have it computed separately. Individual terms $(x_i - \bar{x})^2$ are computed for $i = 1, \ldots 10$ from E2 through E11 by the formula "=(B2-AVERAGE($B$2:$B$11))^2" contained in E2. Since this cell formula is extended through E11, we need to freeze the values of the B column, thus there is $B$2:$B$11 in the function AVERAGE in E2. The sum of the values in E column is the desired quantity contained in K2. Since the desired value for which we are computing the confidence intervals is $x_{n+1} = x_{11} = 1550$, it is given in J6. K6 contains the predicted value $\hat{y}_{n+1}$ according to (1.3.2). The cells J9 and L9 compute the lower and upper confidence limits (1.8.3) and the cells J10 and L10 compute the same for (1.8.4). Based on the results, Lisa concludes that the house price with an area of 1550 square meters, will fall within the interval of \$155 633 to \$330 929 (by 95% confidence level) while the average house price of houses with 1550 square meters of area will fall within the interval of \$213 914 to \$272 649.

1.10. Regression Table

Most of the computations in the preceding sections can be summarized by the regression table. This section aims to analyze the table and make references to the computations above. The regression table can be obtained in Excel using the Data tab, Data Analysis, Regression. The resulting window is

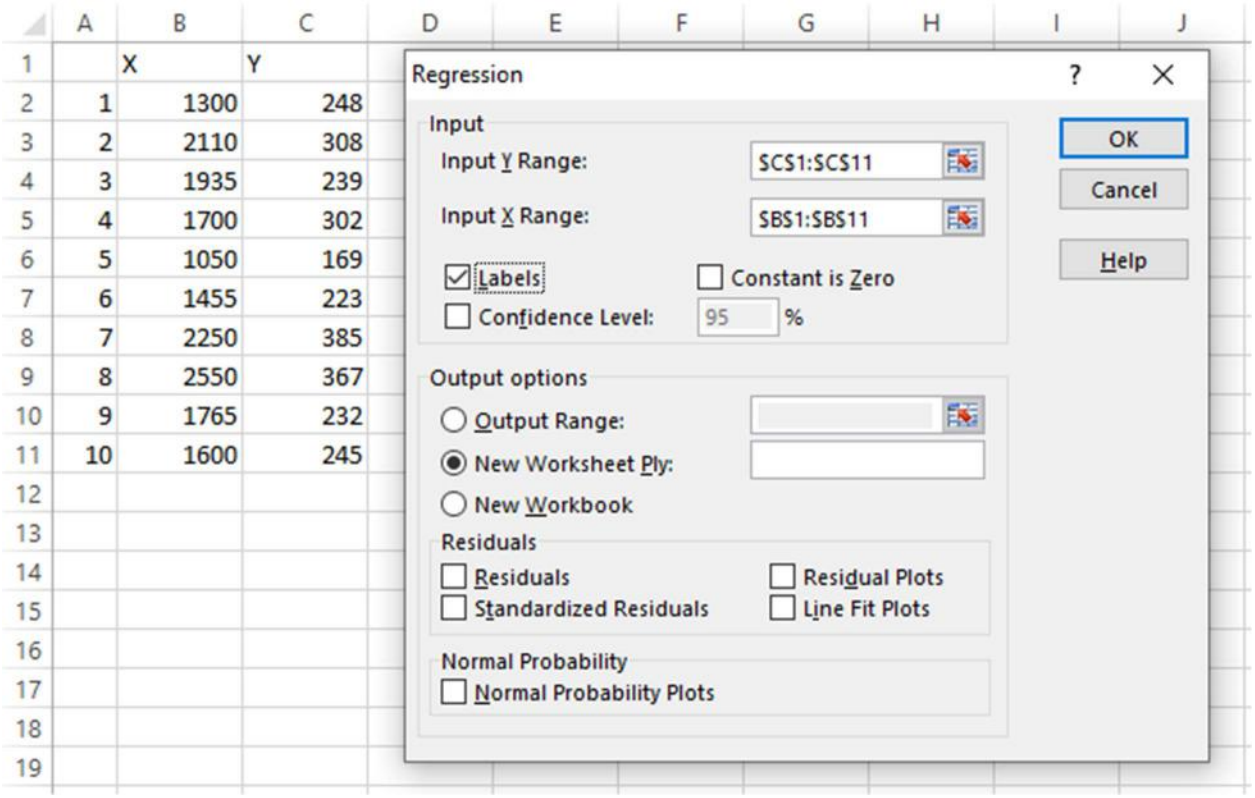


Figure 1.10.1

After filling the inputs as shown above, the resulting regression table is

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.86586572							
5	R Square	0.749723445							
6	Adjusted R Square	0.718438876							
7	Standard Error	35.8115501							
8	Observations	10							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	30733.86304	30733.9	23.96464	0.001200778			
13	Residual	8	10259.73696	1282.47					
14	Total	9	40993.6						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	43.7130536	47.94890387	0.91166	0.388593	-66.85731701	154.28342	-66.857317	154.283424
18	X	0.128753568	0.026301094	4.89537	0.001201	0.068103136	0.189404	0.06810314	0.189404

Figure 1.10.2

The regression table is divided into three sub tables. The topmost table, *Regression Statistics* provides general information about the regression. The middle table, *ANOVA (Analysis of Variance)* contains the information about the measures of variability of the fitted regression model. The last table contains information about the regression coefficients.

The table below summarizes the computations from the preceding sections

#	Cell	Description	Equation
1.	B5	R^2	(1.5.1)
2.	B7	s_e	square root from (1.6.1)
3.	C12	SSR	(1.4.3)
4.	C13	SSE	(1.4.4)
5.	C14	SST	(1.4.2)
6.	D12	MSR	(1.7.5)
7.	D13	MSE	(1.7.6)
8.	E12	F	(1.7.4)
9.	F12	$p-value\ (significance\ F)$	(1.8.3)
10.	B17	b_0	(1.3.4)
11.	B18	b_1	(1.3.3)
12.	C17	s_{b_0}	square root from (1.6.3)

13.	C18	s_{b_1}	square root from (1.6.2)
14.	D18	t	(1.7.2)
15.	E18	$p-value$	(1.8.1)
16.	F17	$LCL\ (\beta_0)$	(1.9.1)
17.	G17	$UCL\ (\beta_0)$	(1.9.1)
18.	F18	$LCL\ (\beta_1)$	(1.9.2)
19.	G18	$UCL\ (\beta_1)$	(1.9.2)

Table 1.10

Chapter 2. Multiple Regression Analyses

2.1. Introduction

In Chapter 1, we developed a simple regression model as a vehicle for estimating dependent variable Y (e.g. house price) in terms of independent variable X (e.g. house area). In this chapter we generalize the topic to multiple regression model. In many situations, more than one variable jointly affects the dependent variable.

- House price is determined by its area, location, availability of a parking space around and possibly the age of a house.
- Demand for a certain product is determined by its price and brand name.
- Concentration of a certain drug in the bloodstream may depend on time passed after injection and age of a patient.

Multiple regression analyzes helps predict the value of Y in terms of several independent variables X. Most of the topics included in this chapter are the generalizations of what has already been examined in Chapter 1. As an additional topic, we include the dummy variables enabling us to incorporate qualitative variables into the model.

2.2. Multiple Regression

The population regression model is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \qquad (2.2.1)$$

where Y is the dependent variable and X_j-s for $j = 1, \dots, k$ are the independent explanatory variables. Similar to the simple regression model, β_0 is the intercept coefficient, β_j-s are the slope coefficients for their corresponding X_j-s and ε is a normal random variable with mean 0 and variance σ^2. Y and X_j-s are random variables whose realizations are given by (2.2.2) below

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \qquad (2.2.2)$$

which states that the observed value y_i is a function of fixed values $\{x_{1i}, x_{2i}, \dots, x_{ki}\}$. As seen in Chapter 1, there are several standard assumptions about this model:

1) The terms x_{ji} are (fixed numbers) the realizations of random variables X_j and are independent of the error terms ε_i.
2) Mathematical expectation of the random variable Y is a linear function of the independent variables X_j.
3) The error terms are normally distributed with mean 0 and variance σ^2:

$$E\varepsilon_i = 0, \qquad E\varepsilon_i^2 = \sigma^2 \qquad for\ i = 1, \dots, n$$

4) The random terms ε_i are independent of one another (uncorrelated)

$$E[\varepsilon_i \varepsilon_j] = 0 \quad for\ all\ i \neq j$$

The corresponding sample estimated model for (2.2.2) is

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i \tag{2.2.3}$$

where e_i is the residual measuring the difference between the actually observed value of Y and its estimated value. The coefficients b_j are to be found using the least squares procedure. The least squares estimates of the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ in (2.2.1) are the values $b_0, b_1, \ldots, b_k$ for which the *sum of squared errors* defined as

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2$$

is minimized. The resulting equation is

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} \tag{2.2.4}$$

For simplicity we consider a model with only two predictor variables. Then (2.2.4) is reduced to

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} \tag{2.2.5}$$

This can visually be illustrated as follows

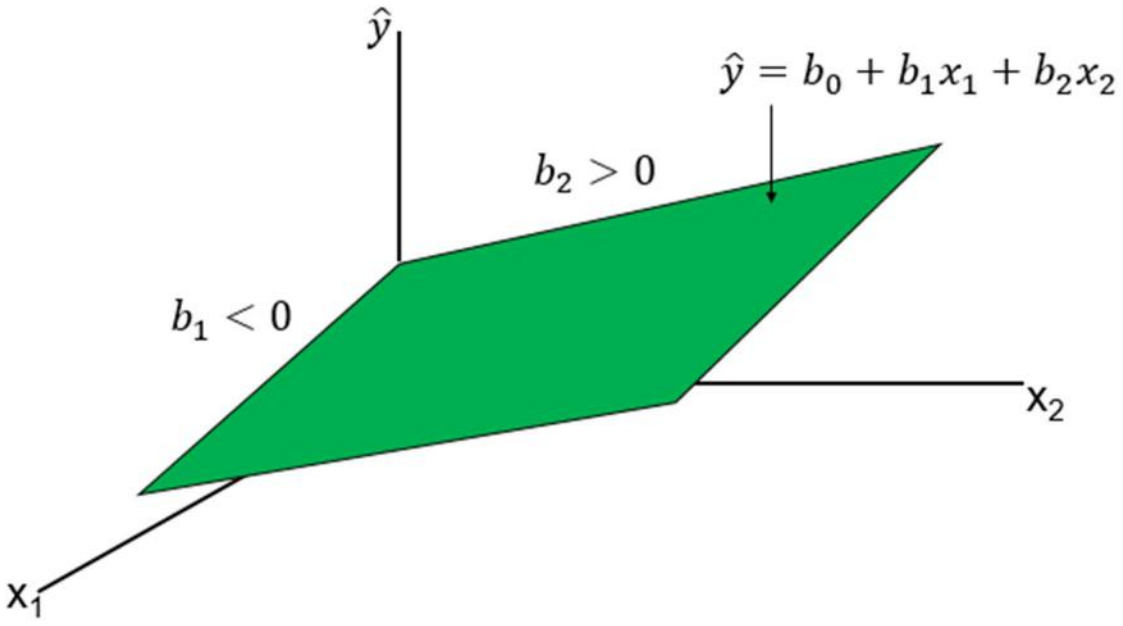


Figure 2.2.1

Unlike the Figure 1.3.2, where $\hat{y}$ was a function of a single variable whose graph was a line, here $\hat{y}$ is a function of two variables whose graph is a surface in three dimensions. It can be referred as the prediction surface. Note that depending on the slope coefficients b_1 and b_2, the surface is falling or rising with respect to the corresponding variable. In the Figure 2.2.1, $b_1 < 0$ and $\hat{y}$ is a decreasing function of x_1. On the other hand, $b_2 > 0$ making $\hat{y}$ an increasing function of x_2.

Minimization procedure for SSE (which is not examined in this book) yields the following results for the coefficients

$$b_1 = \frac{s_y\left(r_{x_1y} - r_{x_1x_2}r_{x_2y}\right)}{s_{x_1}(1 - r_{x_1x_2}^2)} \tag{2.2.6}$$

$$b_2 = \frac{s_y\left(r_{x_2y} - r_{x_1x_2}r_{x_1y}\right)}{s_{x_2}(1 - r_{x_1x_2}^2)} \tag{2.2.7}$$

$$b_0 = \bar{y} - b_1\,\bar{x}_1 - b_2\,\bar{x}_2 \tag{2.2.8}$$

where
s_y is the sample standard deviation of Y,
s_{x_1} is the sample standard deviation of X_1,
s_{x_2} is the sample standard deviation of X_2,
r_{x_1y} is the sample correlation coefficient between X_1 and Y,
r_{x_2y} is the sample correlation coefficient between X_2 and Y,
$r_{x_1x_2}$ is the sample correlation coefficient between X_1 and X_2,
$\bar{y}$ is the sample mean value of Y,
$\bar{x}_1$ is the sample mean value of X_1,
$\bar{x}_2$ is the sample mean value of X_2

Example 2.2

In Example 1.5, the real estate agent, Lisa Miller found that the R^2 coefficient defined by (1.5.1) was 75%. At that time, she had the simple regression model with an only predictor variable of the house price – its area. She thought she could increase the explanatory power of the model by addition more variables. She came up with the house age as an additional factor affecting the house price. So, now she has observations on the triple – house area measured by square meters and denoted by X_1, house age in years denoted by X_2 and the corresponding price – the dependent variable Y measured in \$1000s. (e.g. the 4th record in Figure 2.2.1 indicates that the

8 year old house with 1700 square meters was sold for $302 000). In order to base her predictions of house prices on a given pair of house area and age, she has to construct the equation (2.2.5). She uses the equations (2.4.6), (2.4.7) and (2.4.8) to compute the values of b_1, b_2 and b_0 respectively. The results are shown in the figure below. As expected, the $b_1 = 0.0474 > 0$ and $b_2 = -9.5446 < 0$. Lisa's interpretation of b_1 is the same as it was before, increasing the area of a house causes the house price to also increase. However, since the b_2 coefficient is negative, Lisa concludes that older the house, less its price. The equation (2.2.5) thus is

$$\hat{y} = 301.4223 + 0.0474x_1 - 9.5446x_2$$

At this point, Lisa is able to predict the house price for a given set of its area and age. However, she accepts her colleague's – Mary's advice to measure various other characteristics of the multiple regression model to have an idea about its prediction accuracy before she starts to use the model.

	A	B	C	D	E	F	G	H	I
1		X_1	X_2	Y					
2	1	1300	15	248					
3	2	2110	7	308					
4	3	1935	17	239					
5	4	1700	8	302					
6	5	1050	18	169					
7	6	1455	16	223					
8	7	2250	5	385					
9	8	2550	6	367					
10	9	1765	14	232					
11	10	1600	13	245					
12									
13	s_{x_1}	453.8664	<--"=STDEV.S(B2:B11)"						
14	s_{x_2}	4.909175	<--"=STDEV.S(C2:C11)"						
15	s_y	67.48959	<--"=STDEV.S(D2:D11)"						
16	r_{x_1y}	0.865866	<--"=CORREL(B2:B11,D2:D11)"						
17	r_{x_2y}	-0.94545	<--"=CORREL(C2:C11,D2:D11)"						
18	$r_{x_1x_2}$	-0.78809	<--"=CORREL(B2:B11,C2:C11)"						
19									
20	b_0	301.4223	<--"=AVERAGE(D2:D11)-B21*AVERAGE(B2:B11)-B22*AVERAGE(C2:C11)"						
21	b_1	0.047394	<--"=B15*(B16-B18*B17)/(B13*(1-B18^2))"						
22	b_2	-9.54456	<--"=B15*(B17-B18*B16)/(B14*(1-B18^2))"						

Figure 2.2.1

As shown in Figure 1.3.3, the house price positively depends on its area. Lisa can visualize the negative dependence of the house price by its age by constructing a similar scatter plot for the sample observations on X_2 and Y. The result is shown in the following figure below

	A	B	C	D
1		X_1	X_2	Y
2	1	1300	15	248
3	2	2110	7	308
4	3	1935	17	239
5	4	1700	8	302
6	5	1050	18	169
7	6	1455	16	223
8	7	2250	5	385
9	8	2550	6	367
10	9	1765	14	232
11	10	1600	13	245
12				
13				
14				
15				

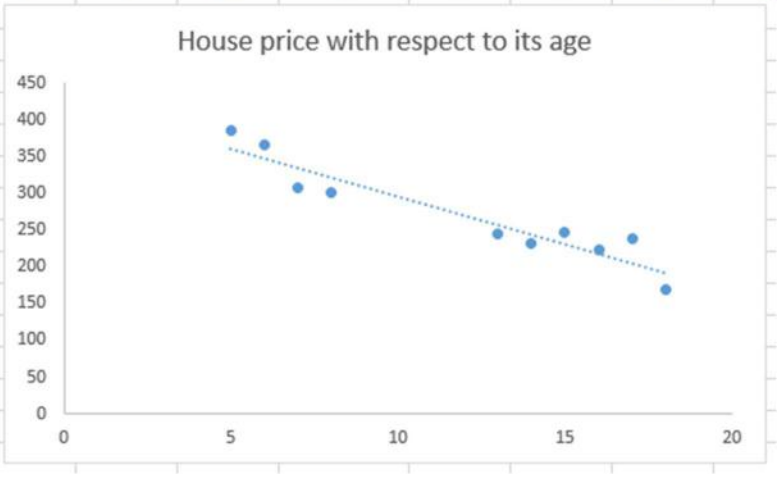


Figure 2.2.2

Note that the equation is deliberately NOT displayed using the "Display equation on chart" checkbox. The reason for this is that it would bring up an equation expressing the dependence of Y on X_2 alone. When modelling the dependence of Y on X_1 and X_2, co-movements of X_1 and X_2 also play a role as shown in (2.2.6) and (2.2.7), so b_1 and b_2 values are different from what they would be if Y was regressed only on any of these variables. So, the value of b_1 is also different from what Lisa had in Example 1.3.

2.3. Measures of Variability

The estimated model from the sample is given by (2.2.3) which is restated below

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i \tag{2.3.1}$$

This can be rewritten as

$$y_i = \hat{y} + e_i$$

as long as the sample fitted regression equation is defined by (2.2.4) as

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} \tag{2.3.2}$$

The difference between the sample mean and the dependent variable can be expressed as

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Squaring both sides yields the following equality

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y}+y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}e_i^2$$

which is the *sum of squares total* decomposed into *sum of squares regression* and *sum of squares error.* So the equality above can be rewritten as

$$SST=SSR+SSE$$

out of which

$$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{2.3.3}$$

$$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}e_i^2 \tag{2.3.4}$$

and

$$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \tag{2.3.5}$$

The figure below shows the differences for a single observed point

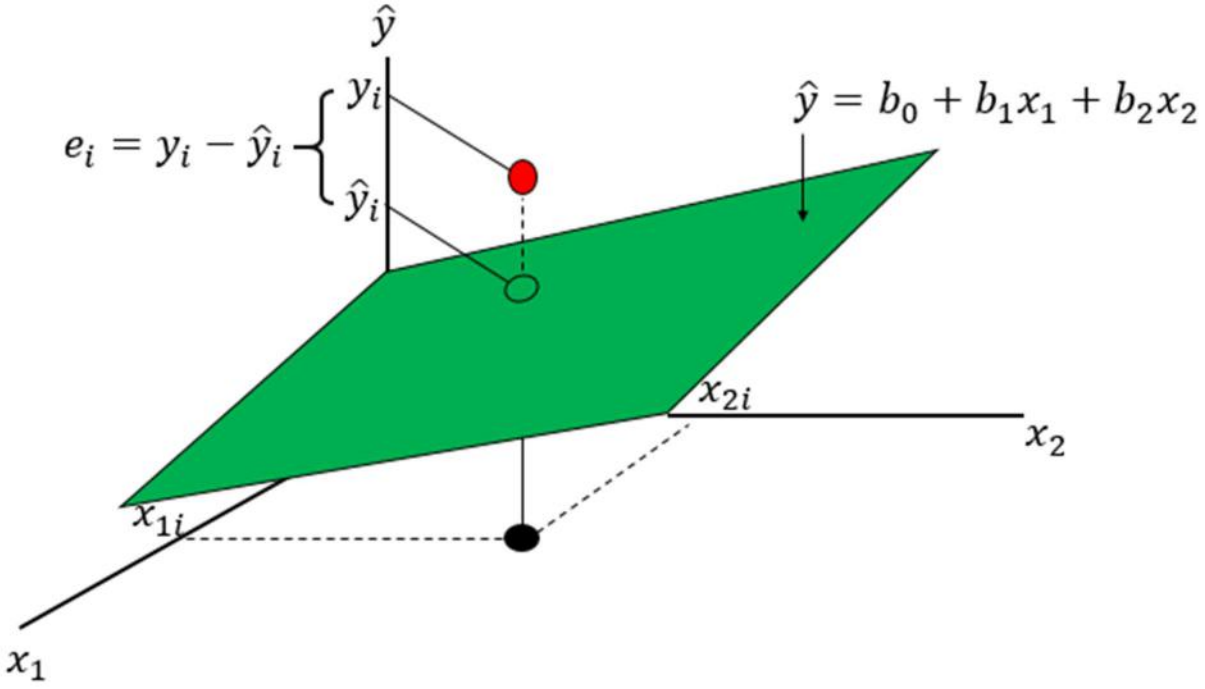


Figure 2.3.1

Example 2.3

Computing the values of (2.3.3), (2.3.4) and (2.3.5) is a right step towards determining whether the addition of the independent variable X_2 improved the existing simple regression model or not. Lisa knows that these values alone are not sufficient and she needs to proceed with measuring the explanatory power of the 2 variable model. The computations of SSE, SSR and SST are shown below.

	A	B	C	D	E	F	G	H
1		X_1	X_2	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$		
2	1	1300	15	248	219.8658	2697.165		
3	2	2110	7	308	334.6111	3945.232		
4	3	1935	17	239	230.8716	1675.131		
5	4	1700	8	302	305.6351	1144.815		
6	5	1050	18	169	179.3837	8540.777		
7	6	1455	16	223	217.6672	2930.357		
8	7	2250	5	385	360.3353	7838.5		
9	8	2550	6	367	365.0089	8687.89		
10	9	1765	14	232	251.4484	414.1887		
11	10	1600	13	245	253.173	346.9661		
12								
13	s_{x_1}	453.8664		SSE	2772.577	<--"=SUMXMY2(D2:D11,E2:E11)"		
14	s_{x_2}	4.909175		SSR	38221.02	<--"=SUM(F2:F11)"		
15	s_y	67.48959		SST	40993.6	<--"=SUM(E13:E14)"		
16	r_{x_1y}	0.865866						
17	r_{x_2y}	-0.94545						
18	$r_{x_1x_2}$	-0.78809						
19								
20	b_0	301.4223						
21	b_1	0.047394						
22	b_2	-9.54456						

Figure 2.3.2

The column E contains the computations of $\hat{y}_i$ for $i = 1, \dots, 10$ by the function "=$B$20+$B$21*B2+$B$22*C2" in the cell E2. The SSE is then computed in the cell E13 by the formula $SUMXMY2$ which receives two vectors as arguments and makes the summation of the differences squared. On the other hand, in order to compute the SSR value, at first, the squared differences have to be computed for the mean observed value of Y and its estimated value $\hat{y}$. This computation is carried out in column F with the formula "=(E2-AVERAGE($D$2:$D$11))^2" in the cell F2. The sum of these values is the SSR computed in the cell E14. Finally, the sum of SSE and SSR is SST in E15 which could have alternatively

been computed by taking the sum of the differences $y_i - \bar{y}$ for $i = 1, \dots, 10$. For the sake of illustration, this computation is carried out in the following figure

	A	B	C	D	E	F	G	H
1		X_1	X_2	Y	$\hat{y}$	$(\hat{y} - \bar{y})^2$	$(y_i - \bar{y})^2$	
2	1	1300	15	248	219.8658	2697.165	566.44	
3	2	2110	7	308	334.6111	3945.232	1310.44	
4	3	1935	17	239	230.8716	1675.131	1075.84	
5	4	1700	8	302	305.6351	1144.815	912.04	
6	5	1050	18	169	179.3837	8540.777	10567.84	
7	6	1455	16	223	217.6672	2930.357	2381.44	
8	7	2250	5	385	360.3353	7838.5	12814.24	
9	8	2550	6	367	365.0089	8687.89	9063.04	
10	9	1765	14	232	251.4484	414.1887	1584.04	
11	10	1600	13	245	253.173	346.9661	718.24	
12								
13	s_{x_1}	453.8664		SSE	2772.577	<--"=SUMXMY2(D2:D11,E2:E11)"		
14	s_{x_2}	4.909175		SSR	38221.02	<--"=SUM(F2:F11)"		
15	s_y	67.48959		SST	40993.6	<--"=SUM(E13:E14)"		
16	r_{x_1y}	0.865866		SST	40993.6	<--"=SUM(G2:G11)"		
17	r_{x_2y}	-0.94545						
18	$r_{x_1x_2}$	-0.78809						
19								
20	b_0	301.4223						
21	b_1	0.047394						
22	b_2	-9.54456						

Figure 2.3.3

The value in the cell E16 is the sum of the values in column G which are the squared differences between the observed value of Y and the sample mean value computed by "=(D2-AVERAGE($D$2:$D$11))^2" in the cell G2.

2.4. The Explanatory Power of a Multiple Regression Equation

In Section 1.5, the coefficient of determination of the fitted regression equation was defined to be the proportion of total variability explained by the regression. So, it is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{2.4.1}$$

It follows that

$$0 \leq R^2 \leq 1$$

Closer this value to one, better the capability of the fitted regression equation to make accurate predictions. In Section 1.5, we saw that the value of the coefficient of determination coincides with the square of the value of the sample correlation coefficient between X and Y variables (there was only a single independent variable X in Section 1.5). In particular, we had $R^2 = r_{xy}^2$. This equality implies that the explanatory power of the model, R^2, is the same as the measure of linear dependence between the dependent and independent variables. So, the goodness of the model can be determined directly by the sample correlation coefficient between Y and X. On the other hand, there are more than one independent variables in the multiple regression model. So, how can all $X_j, j = 1, \dots, k$ variables taken together explain the variation in the values of Y? It turns out that the value of R^2 is related to the correlation between the sample observations of Y and the estimated values of Y. In particular, we have

$$R^2 = r_{y\,\hat{y}}^2 \tag{2.4.2}$$

where $r_{y\,\hat{y}}$ is the correlation between the observed values of Y and its estimated values by the fitted regression equation (2.3.2). We can conclude that R^2 is still a correlation coefficient squared just like in (1.5.2). However, unlike (1.5.2), here the correlation is taken between Y and all the independent variables taken together and combined in $\hat{y}$. It is reasonable to assign $r_{y\hat{y}}$ a separate name emphasizing its role. So, the *coefficient of multiple correlation* is defined as

$$R = \sqrt{R^2} = r_{y\hat{y}} \tag{2.4.3}$$

Now consider what happens when we add new predictor variables to the regression model. It can be shown that addition of a new variable reduces SSE. It can be thought of this way: every new independent variable brings new information and reduces SSE. Therefore, the portion of SSR in SST increases and so does R^2. So, addition of a new variable evidently improves the model by increasing R^2. However, the new variable may not contribute much to the explanatory power of the existing model and yet decrease the degrees of freedom of the model (Regression Degrees of Freedom is $n - k - 1$ where k is the number of predictor variables). More degrees of freedom mean more accuracy of analyzes based on the sample. So, the dilemma whether or not to add a new variable depends on the measurement whether the new variable brings enough explanatory power to offset the loss of one degree of freedom is considered below.

In order to correct the fact that addition of a non-relevant explanatory variable will still increase SSR and correspondingly R^2, the *adjusted* R^2 denoted by $\bar{R}^2$ is used. It is defined as

$$\bar{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} \tag{2.4.4}$$

where n is the sample size and k is the number of explanatory variables. It can be seen in (2.4.4) that R^2 penalizes the excessive use insignificant independent variables. $\bar{R}^2 < R^2$ and provides a better way of comparing multiple regression models with different number of explanatory variables.

Example 2.4

In Example 2.3, it was unclear if addition of the new variable X_2, which is the age of a house, improved the model. The computations of the values of R^2, R and $\bar{R}^2$ are shown in the following figure.

	A	B	C	D	E	F	G	H
1		X_1	X_2	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$		
2	1	1300	15	248	219.8658	2697.165		
3	2	2110	7	308	334.6111	3945.232		
4	3	1935	17	239	230.8716	1675.131		
5	4	1700	8	302	305.6351	1144.815		
6	5	1050	18	169	179.3837	8540.777		
7	6	1455	16	223	217.6672	2930.357		
8	7	2250	5	385	360.3353	7838.5		
9	8	2550	6	367	365.0089	8687.89		
10	9	1765	14	232	251.4484	414.1887		
11	10	1600	13	245	253.173	346.9661		
12								
13	s_{x_1}	453.8664		SSE	2772.577	<--"=SUMXMY2(D2:D11,E2:E11)"		
14	s_{x_2}	4.909175		SSR	38221.02	<--"=SUM(F2:F11)"		
15	s_y	67.48959		SST	40993.6	<--"=SUM(E13:E14)"		
16	r_{x_1y}	0.865866						
17	r_{x_2y}	-0.94545		R^2	0.932366	<--"=E14/E15"		
18	$r_{x_1x_2}$	-0.78809		R	0.965591	<--"=CORREL(D2:D11,E2:E11)"		
19				$\bar{R}^2$	0.913041	<--"=1-(E13/(A11-2-1))/(E15/A10)"		
20	b_0	301.4223						
21	b_1	0.047394						
22	b_2	-9.54456						

Figure 2.4.1

The real estate agent now has meaningful numbers at her disposal. R^2 is 93%. However, in order to account for the possible unnecessary loss of degrees of freedom, Lisa relies on the adjusted coefficient of determination $\bar{R}^2$ which is also quite high, 91%. The root from R^2 is computed as $r_{y\hat{y}}$ whose value is 96.56% implying a very strong linear relation between the predicted and observed values of Y.

Lisa can now conclude that addition of the new variable (X_2, the age of a house) significantly increased the explanatory power of the existing simple regression model (with only X_1, the area of a house). Now she can make better predictions taking these two factors into account.

Lisa now wants to assess the variability for the individual components of the regression model. In particular, she determines to estimate the standard error of the estimate and the standard errors of the slope coefficients similarly to Section 1.6.

2.5. Standard Error of the Estimate and Variance Estimators of the Regression Coefficients

In Section 1.6, the unbiased estimate of the error variance was defined by the (1.6.1) and the root from this value was defined to be the standard error of the estimate. The geometric effects of this value was also described. In the context of the multiple regression model (2.2.2), the unbiased estimate of error variance is

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1} = \frac{SSE}{n-k-1} \tag{2.5.1}$$

Note that the equation (1.6.1) is just the special case of (2.5.1) when $k = 1$. In order to measure the effects of correlations between the independent variables, consider the generalized versions of (1.6.2). The sample estimators of the regression coefficient variance are

$$s_{b_1}^2 = \frac{s_e^2}{(n-1)s_{x_1}^2(1-r_{x_1x_2}^2)} \tag{2.5.2}$$

$$s_{b_2}^2 = \frac{s_e^2}{(n-1)s_{x_2}^2(1-r_{x_1x_2}^2)} \tag{2.5.3}$$

The square roots from each are the standard errors of the slope coefficients.

Example 2.5

The computations of (2.5.1), (2.5.2) and (2.5.3) are carried out below

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		X_1	X_2	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$			s_e^2	396.0825	<--"=E13/(A11-2-1)"		
2	1	1300	15	248	219.8658	2697.165			s_e	19.90182	<--"=SQRT(J1)"		
3	2	2110	7	308	334.6111	3945.232			$s_{b_1}^2$	0.000564	<--"=J1/(A10*B13^2*(1-B18^2))"		
4	3	1935	17	239	230.8716	1675.131			s_{b_1}	0.023745	<--"=SQRT(J3)"		
5	4	1700	8	302	305.6351	1144.815			$s_{b_2}^2$	4.819255	<--"=J1/(A10*B14^2*(1-B18^2))"		
6	5	1050	18	169	179.3837	8540.777			s_{b_2}	2.19528	<--"=SQRT(J5)"		
7	6	1455	16	223	217.6672	2930.357							
8	7	2250	5	385	360.3353	7838.5							
9	8	2550	6	367	365.0089	8687.89							
10	9	1765	14	232	251.4484	414.1887							
11	10	1600	13	245	253.173	346.9661							
12													
13	s_{x_1}	453.8664		SSE	2772.577	<--"=SUMXMY2(D2:D11,E2:E11)"							
14	s_{x_2}	4.909175		SSR	38221.02	<--"=SUM(F2:F11)"							
15	s_y	67.48959		SST	40993.6	<--"=SUM(E13:E14)"							
16	r_{x_1y}	0.865866											
17	r_{x_2y}	-0.94545		R^2	0.932366	<--"=E14/E15"							
18	$r_{x_1x_2}$	-0.78809		R	0.965591	<--"=CORREL(D2:D11,E2:E11)"							
19				$\bar{R}^2$	0.913041	<--"=1-(E13/(A11-2-1))/(E15/A10)"							
20	b_0	301.4223											
21	b_1	0.047394											
22	b_2	-9.54456											

Figure 2.5

The cells J1, J3 and J5 contain computations based on the above-mentioned formulas. The square roots from them are computed in the cells J2, J4 and J6.

2.6. Hypothesis for the Multiple Regression Slope Coefficients

When testing the existence of linear relationship between the dependent and independent variables in simple regression model, we conducted the correlation analyses. This was sufficient when dealing with a single explanatory variable. However, in the multiple regression model, there are more than one explanatory independent variables. In Section 1.7, we defined the hypothesis based on the Student's t test. Here, we generalize the hypothesis (1.7.1).

For the regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

consider the following hypothesis

$$\begin{aligned} &H_0: \beta_j = 0 \ (no\ linear\ relationship) \\ &H_1: \beta_j \neq 0 \quad (linear\ relationship) \end{aligned} \tag{2.6.1}$$

Here the existence of linear relationship between Y and X_j is tested. This test can be regarded as a filter to decide whether a given variable should be included in the regression model or

not. For the least squares coefficients, $b_0, b_1, \dots, b_k$ and the estimated standard deviations $s_{b_0}, s_{b_1}, \dots, s_{b_k}$ of the least squares estimators, it can be shown that

$$t = \frac{b_j - \beta_j}{s_{b_j}} = \frac{b_j - 0}{s_{b_j}} = \frac{b_j}{s_{b_j}}, \quad j = 1, \dots, k \tag{2.6.2}$$

follows the Student's t distribution with $n - k - 1$ degrees of freedom. So, the rejection rule for the hypothesis is to

$$\text{reject } H_0 \text{ if } t \geq t_{n-k-1,\alpha/2} \quad \text{or} \quad t \leq -t_{n-k-1,\alpha/2} \tag{2.6.3}$$

by the significance level α. Note that when $k = 1$, (2.6.3) reduces to its simple regression version with the degrees of freedom $n - 2$.

Similarly, the hypothesis can be tested for any other value of $\beta_j, j = 1, \dots, k$. Taking the specific value of the slope coefficient β_1^* and the significance level α, several possible formulations of the hypothesis are summarized below

Case 1: To test the null hypothesis

$$H_0: \beta_j = \beta_j^* \text{ or } H_0: \beta_j \leq \beta_j^*$$

against the alternative

$$H_1: \beta_1 > \beta_1^*$$

the decision rule is to

$$reject\ H_0\ if\ t = \frac{b_j - \beta_j^*}{s_{b_j}} \geq t_{n-k-1,\alpha}$$

Case 2: To test the null hypothesis

$$H_0: \beta_j = \beta_j^* \text{ or } H_0: \beta_j \geq \beta_j^*$$

against the alternative

$$H_1: \beta_j < \beta_j^*$$

the decision rule is to

$$reject\ H_0\ if\ t = \frac{b_j - \beta_j}{s_{b_j}} \leq -t_{n-k-1,\alpha}$$

Case 3: To test the null hypothesis

$$H_0: \beta_j = \beta_j^*$$

against the alternative

$$H_1: \beta_j \neq \beta_j^*$$

the decision rule is to

$$reject\ H_0\ if\ t = \frac{b_j - \beta_j^*}{s_{b_j}} > t_{n-k-2,\alpha/2}\ or\ t = \frac{b_j - \beta_j^*}{s_{b_j}} \leq -t_{n-k-1,\alpha/2}$$

The last one is the generalized version of (2.6.1). In all of the above cases (with one tailed test), the critical value $t_{n-k-1,\alpha}$ satisfies

$$P\left(t_{n-k-1} > t_{n-k-1,\alpha}\right) = \alpha$$

The for positive t, p-value is defined as

$$p - value = P(t_{n-k-1} > t) \tag{2.6.4}$$

So, the rejection rules above which compares the t statistics to the critical value can be replaced by comparison of the p-value to the significance level.

Hypothesis test for overall significance

Even though the hypothesis described above allow us to individually test the regression coefficient values β_j against a given value β_j^*. There is a way to test all of the coefficients against zero. Consider again the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

and the hypothesis formulated as

$$\begin{gathered} H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1: at\ least\ one\ \beta_j \neq 0 \end{gathered} \tag{2.6.5}$$

Accepting the null hypothesis would lead to a conclusion that none of the explanatory variables is statistically significant and they should not to be included in the regression model. Rather, the new set of the independent variables has to be proposed. On the other hand, if the null hypothesis is rejected (as it usually happens), then we can conclude that this set of variables contains at least one which is statistically significant for explaining the value of the dependent variable. However, accepting the alternative hypothesis does not provide any

information for identifying that variable beyond the fact that it is in a given set. In order to identify the relevant variables, the individual tests need to be carried out using the Student's t test.

Before defining the rejection rule for (2.6.5), let us define *mean square regression* as

$$MSR = \frac{SSR}{k} \tag{2.6.6}$$

and *mean square for error* as

$$MSE = \frac{SSE}{n-k-1} = s_e^2 \tag{2.6.7}$$

The ratio

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{SSR/k}{s_e^2} \tag{2.6.8}$$

follows the F distribution with k degrees of freedom for the numerator and $n-k-1$ degrees of freedom for the denominator. For a given significance level α, the decision rule for (2.6.5) is to

$$reject\ H_0\ if\ F = F_{k,n-k-1} > F_{k,n-k-1,\alpha} \tag{2.6.9}$$

where $F_{k,n-k-1,\alpha}$ is the critical value for which

$$P\left(F_{k,n-k-1} > F_{k,n-k-1,\alpha}\right) = \alpha \tag{2.6.10}$$

is defined to be the p-value for $F_{k,n-k-1,\alpha}$.

Hypothesis test for a subset of regression explanatory variables

In addition to testing all of the coefficients against zero as in (2.6.5), there is a way to test the hypothesis for a subset of variables. Consider the following multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_r z_r + \varepsilon \tag{2.6.11}$$

There are $k + r$ explanatory variables in (2.6.11) out of which the first k variables are assumed to be affecting the dependent variable. The variables $z_1, \ldots, z_r$ are to be tested by the following hypothesis

$$\begin{gathered} H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0 \\ H_1: at\ least\ one\ \alpha_j \neq 0 \end{gathered} \tag{2.6.12}$$

Accepting the null condition would imply that all of the α_j coefficients are simultaneously zero and none of the variables is affecting the dependent variable. On the other hand, rejecting the null implies that there is at least one variable relevant in the estimation of the dependent variable value and the t test described above needs to be conducted to identify such variables individually. It can be shown that under the standard regression assumptions

$$F = \frac{[SSE(r) - SSE\ ]/r}{s_e^2} \tag{2.6.13}$$

follows the F distribution with the numerator degrees of freedom r, and the denominator degrees of freedom $n - k - r - 1$. $SSE(r)$ is the *sum of squared errors* computed for the regression involving only the first k variables $x_1, x_2, \dots, x_k$ which is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

This is referred to as *restricted sum of squared errors.* SSE is the *sum of squared errors* for the entire model (2.6.11). Once having computed the value of F, the decision rule for the hypothesis is to

$$reject\ H_0\ if\ F > F_{r,n-k-r-1,\alpha} \tag{2.6.14}$$

Example 2.6.1

In Example 2.4, the real estate agent, Lisa Miller had to make a choice of the new variable to be added to the existing regression model. Ultimately, she saw that addition of X_2 improved the model by increasing the adjusted coefficient of determination significantly. Generally, in order to decide on a given explanatory variable, the hypothesis (2.6.1) is tested. Lisa decided to now test the following hypotheses by $\alpha = 0.1$.

$$\begin{aligned} &H_0: \beta_1 = 0 \\ &H_1: \beta_1 \neq 0 \end{aligned} \tag{2.6.15}$$

and

$$\begin{aligned} &H_0: \beta_2 = 0 \\ &H_1: \beta_2 \neq 0 \end{aligned} \tag{2.6.16}$$

The following figure illustrates the computations

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		X_1	X_2	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$			s_e^2	396.0825	<--"=E13/(A11-2-1)"		
2	1	1300	15	248	219.8658	2697.165			s_e	19.90182	<--"=SQRT(J1)"		
3	2	2110	7	308	334.6111	3945.232			$s_{b_1}^2$	0.000564	<--"=J1/(A10*B13^2*(1-B18^2))"		
4	3	1935	17	239	230.8716	1675.131			s_{b_1}	0.023745	<--"=SQRT(J3)"		
5	4	1700	8	302	305.6351	1144.815			$s_{b_2}^2$	4.819255	<--"=J1/(A10*B14^2*(1-B18^2))"		
6	5	1050	18	169	179.3837	8540.777			s_{b_2}	2.19528	<--"=SQRT(J5)"		
7	6	1455	16	223	217.6672	2930.357							
8	7	2250	5	385	360.3353	7838.5			t_1	1.995951	<--"=B21/J4"		
9	8	2550	6	367	365.0089	8687.89			t_2	-4.34776	<--"=B22/J6"		
10	9	1765	14	232	251.4484	414.1887			$t_{7,0.05}$	1.894579	<--"=T.INV.2T(0.1,7)"		
11	10	1600	13	245	253.173	346.9661							
12													
13	s_{x_1}	453.8664		SSE	2772.577	<--"=SUMXMY2(D2:D11,E2:E11)"							
14	s_{x_2}	4.909175		SSR	38221.02	<--"=SUM(F2:F11)"							
15	s_y	67.48959		SST	40993.6	<--"=SUM(E13:E14)"							
16	r_{x_1y}	0.865866											
17	r_{x_2y}	-0.94545		R^2	0.932366	<--"=E14/E15"							
18	$r_{x_1x_2}$	-0.78809		R	0.965591	<--"=CORREL(D2:D11,E2:E11)"							
19				$\bar{R}^2$	0.913041	<--"=1-(E13/(A11-2-1))/(E15/A10)"							
20	b_0	301.4223											
21	b_1	0.047394											
22	b_2	-9.54456											

Figure 2.6.1

Note that $t_{n-k-1,\alpha/2} = t_{7,0.05}$ is computed in the cell J10. Since the hypothesis is tested based on $\alpha = 0.1$, the T.INV.2T function receives the significance level as an argument and makes the division by 2. Lisa obtains the critical value to be $t_{7,0.05} = 1.8946$. As long as the values of both t_1 and t_2 statistics in absolute value are greater than the 1.8946, Lisa concludes that by the given significance level of $\alpha = 0.05$, both variables, the house area X_1 and the house age X_2 are linearly related to the house price Y and retains them in the model.

Example 2.6.2

For the sake of illustration, the following hypothesis could have been tested before the hypothesis above

$$H_0: \beta_1 = \beta_2 = 0 \qquad (2.6.17)$$
$$H_1: at\ least\ one\ \beta_j \neq 0$$

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		X_1	X_2	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$			s_e^2	396.0825	<--"=E13/(A11-2-1)"		
2	1	1300	15	248	219.8658	2697.165			s_e	19.90182	<--"=SQRT(J1)"		
3	2	2110	7	308	334.6111	3945.232			$s_{b_1}^2$	0.000564	<--"=J1/(A10*B13^2*(1-B18^2))"		
4	3	1935	17	239	230.8716	1675.131			s_{b_1}	0.023745	<--"=SQRT(J3)"		
5	4	1700	8	302	305.6351	1144.815			$s_{b_2}^2$	4.819255	<--"=J1/(A10*B14^2*(1-B18^2))"		
6	5	1050	18	169	179.3837	8540.777			s_{b_2}	2.19528	<--"=SQRT(J5)"		
7	6	1455	16	223	217.6672	2930.357							
8	7	2250	5	385	360.3353	7838.5			t_1	1.995951	<--"=B21/J4"		
9	8	2550	6	367	365.0089	8687.89			t_2	-4.34776	<--"=B22/J6"		
10	9	1765	14	232	251.4484	414.1887			$t_{7,0.05}$	1.894579	<--"=T.INV.2T(0.1,7)"		
11	10	1600	13	245	253.173	346.9661							
12													
13	s_{x_1}	453.8664		SSE	2772.577	<--"=SUMXMY2(D2:D11,E2:E11)"							
14	s_{x_2}	4.909175		SSR	38221.02	<--"=SUM(F2:F11)"			MSR	19110.51	<--"=E14/2"		
15	s_y	67.48959		SST	40993.6	<--"=SUM(E13:E14)"			MSE	396.0825	<--"=J1"		
16	r_{x_1y}	0.865866							F	48.24882	<--"=J13/J14"		
17	r_{x_2y}	-0.94545		R^2	0.932366	<--"=E14/E15"			$F_{2,7,0.1}$	3.257442	<--"=F.INV(0.9,2,7)"		
18	$r_{x_1x_2}$	-0.78809		R	0.965591	<--"=CORREL(D2:D11,E2:E11)"							
19				$\bar{R}^2$	0.913041	<--"=1-(E13/(A11-2-1))/(E15/A10)"							
20	b_0	301.4223											
21	b_1	0.047394											
22	b_2	-9.54456											

Figure 2.6.2

Since $F = 48.25 > F_{2,7,0.1} = 3.26$, the null hypothesis is rejected.

As an alternative, the p-values can be used to test the hypothesis.

Example 2.6.3

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		X_1	X_2	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$			s_e^2	396.0824666	<--"=E13/(A11-2-1)"		
2	1	1300	15	248	219.8658	2697.165			s_e	19.90182069	<--"=SQRT(J1)"		
3	2	2110	7	308	334.6111	3945.232			$s_{b_1}^2$	0.000563821	<--"=J1/(A10*B13^2*(1-B18^2))"		
4	3	1935	17	239	230.8716	1675.131			s_{b_1}	0.023744906	<--"=SQRT(J3)"		
5	4	1700	8	302	305.6351	1144.815			$s_{b_2}^2$	4.819255488	<--"=J1/(A10*B14^2*(1-B18^2))"		
6	5	1050	18	169	179.3837	8540.777			s_{b_2}	2.195280276	<--"=SQRT(J5)"		
7	6	1455	16	223	217.6672	2930.357							
8	7	2250	5	385	360.3353	7838.5			t_1	1.995951109	<--"=B21/J4"		
9	8	2550	6	367	365.0089	8687.89			t_2	-4.34776177	<--"=B22/J6"		
10	9	1765	14	232	251.4484	414.1887			$t_{7,0.05}$	1.894578605	<--"=T.INV.2T(0.1,7)"		
11	10	1600	13	245	253.173	346.9661			p-value (t_1)	0.086132093	<--"=T.DIST.2T(J8,7)"		
12									p-value (t_2)	0.003364031	<--"=T.DIST.2T(ABS(J9),7)"		
13	s_{x_1}	453.8664		SSE	2772.577	<--"=SUMXMY2(D2:D11,E2:E11)"							
14	s_{x_2}	4.909175		SSR	38221.02	<--"=SUM(F2:F11)"			MSR	19110.51137	<--"=E14/2"		
15	s_y	67.48959		SST	40993.6	<--"=SUM(E13:E14)"			MSE	396.0824666	<--"=J1"		
16	r_{x_1y}	0.865866							F	48.24881932	<--"=J13/J14"		
17	r_{x_2y}	-0.94545		R^2	0.932366	<--"=E14/E15"			$F_{2,7,0.1}$	3.257442051	<--"=F.INV(0.9,2,7)"		
18	$r_{x_1x_2}$	-0.78809		R	0.965591	<--"=CORREL(D2:D11,E2:E11)"			p-value (F)	8.04612E-05	<--"=F.DIST.RT(J16,2,7)"		
19				$\bar{R}^2$	0.913041	<--"=1-(E13/(A11-2-1))/(E15/A10)"							
20	b_0	301.4223											
21	b_1	0.047394											
22	b_2	-9.54456											

Figure 2.6.3

The p-values corresponding to t_1 and t_2 with the given degrees of freedom (which is 7) are computed in the cells J11 and J12. As long as the hypothesis is tested by $\alpha = 0.1$, the null hypotheses (2.6.15) and (2.6.16) are rejected since the values in the cells J11 and J12 are

$$P(t_{n-k-1} > t_1) = 0.0861 < \alpha = 0.1$$

and

$$P(t_{n-k-1} > |t_2|) = 0.0034 < \alpha = 0.1$$

Note that since the density function of the Student's t distribution is symmetric and $t_2 < 0$, the absolute value of t_2 is used as an argument of the T.DIST.2T function.

Similarly, the p-value for F statistics, also known as *Significance F,* computed in the cell J18 is less than the critical value

$$P\left(F_{k,n-k-1} > F\right) = 0 < \alpha = 0.1 \qquad (2.6.18)$$

and the hypothesis (2.6.17) is rejected.

Example 2.6.4

Suppose the real estate agent in the previous examples, Lisa Miller is considering to have three explanatory variables for the dependent variable house price - X_1 being the area measured in square meters, X_2 as the house age and X_3 as the number of bedrooms. Thus, the regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

She does not doubt that the house area affects the house price. So she does not need to test any hypothesis for X_1. She only wants to know if any of the rest of the variables affects the price. The regression model above can be rewritten as

$$y = \beta_0 + \beta_1 x_1 + \alpha_2 z_2 + \alpha_3 z_3$$

So, she attempts to test the following hypothesis

$$H_0: \alpha_2 = \alpha_3 = 0$$

$$H_1: at\ least\ one\ a_j \neq 0$$

The following figure illustrates the sample observations on three independent variables

	A	B	C	D	E
1		X_1	X_2	X_3	Y
2	1	1300	15	4	248
3	2	2110	7	6	308
4	3	1935	17	3	239
5	4	1700	8	3	302
6	5	1050	18	2	169
7	6	1455	16	3	223
8	7	2250	5	7	385
9	8	2550	6	6	367
10	9	1765	14	3	232
11	10	1600	13	2	245

Figure 2.6.4

In order to compute the F statistics from (2.6.13), Lisa has to split the regression model into two parts, the estimated regression equation for the first one is

$$\hat{y} = b_0 + b_1 x_1$$

for which she computes $SSE(r)$, and the second equation is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

for which she computes SSE. In order to follow the procedure of the hypothesis, we take a shortcut in computing these values. The following figure illustrates construction of the regression table only for X_1 using the Data tab in Excel, Data Analysis, Regression option.

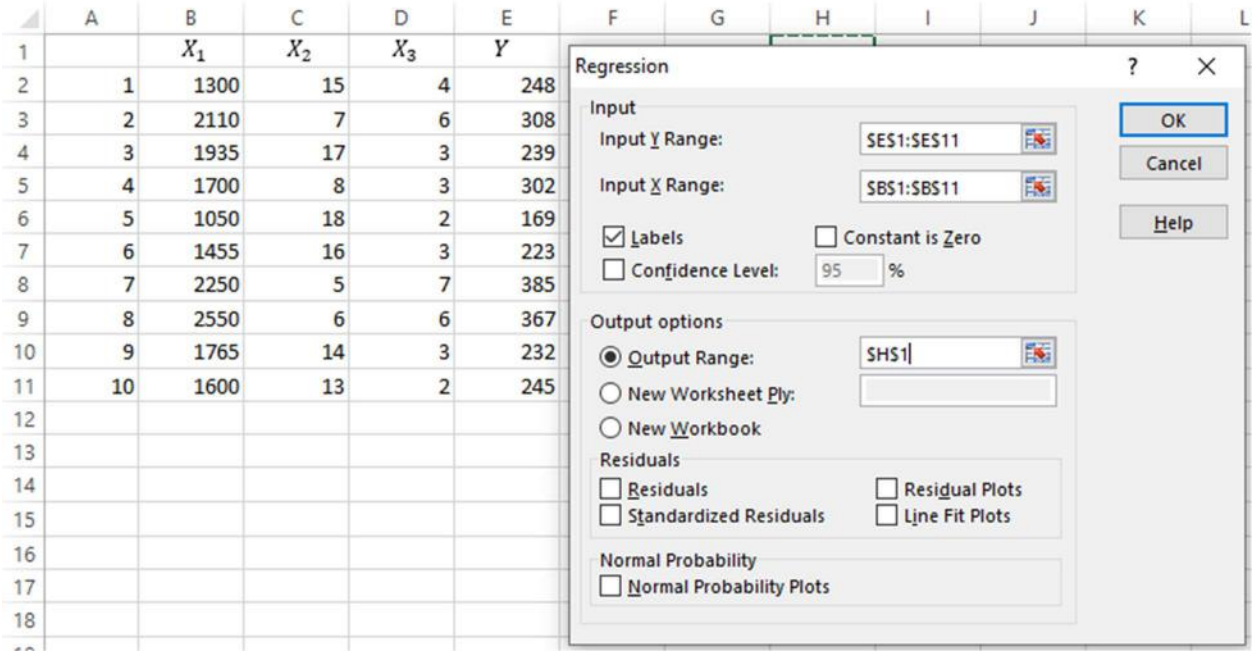


Figure 2.6.5

The resulting output table is

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.865866							
R Square	0.749723							
Adjusted R Square	0.718439							
Standard Error	35.81155							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	30733.86304	30733.86	23.96464	0.001200778			
Residual	8	10259.73696	1282.467					
Total	9	40993.6						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept b_0	43.71305	47.94890387	0.911659	0.388593	-66.85731701	154.2834	-66.8573	154.2834
b_1	0.128754	0.026301094	4.895369	0.001201	0.068103136	0.189404	0.068103	0.189404

Figure 2.6.6

From which we only need the highlighted $SSE(r)$. Similarly, constructing the regression table for X_1, X_2, X_3 as follows

	A	B	C	D	E
1		X_1	X_2	X_3	Y
2	1	1300	15	4	248
3	2	2110	7	6	308
4	3	1935	17	3	239
5	4	1700	8	3	302
6	5	1050	18	2	169
7	6	1455	16	3	223
8	7	2250	5	7	385
9	8	2550	6	6	367
10	9	1765	14	3	232
11	10	1600	13	2	245

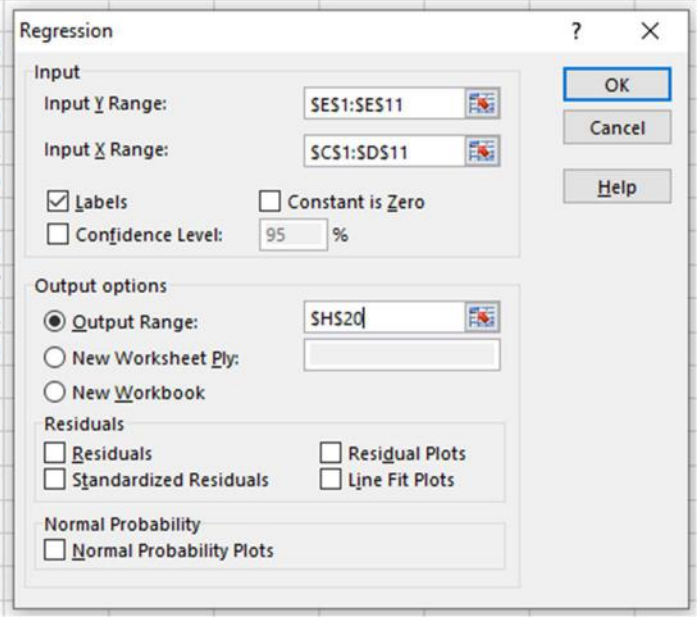


Figure 2.6.7

results in the following output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.973168
R Square	0.947056
Adjusted R Square	0.920584
Standard Error	19.01919
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	38823.22224	12941.07	35.77554	0.000318128
Residual	6	2170.377764	361.7296		
Total	9	40993.6			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept b_0	271.0952	66.40503699	4.08245	0.006484	108.6079456	433.5825	108.6079	433.5825
b_1	0.035209	0.024578322	1.43254	0.201964	-0.024931552	0.09535	-0.02493	0.09535
b_2	-7.94291	2.437658925	-3.25842	0.017283	-13.90764679	-1.97817	-13.9076	-1.97817
b_3	8.423564	6.528564697	1.290263	0.24445	-7.551258563	24.39839	-7.55126	24.39839

Figure 2.6.8

The highlighted numbers are s_e^2 and SSE. Computation of F from (2.6.13) and the critical value $F_{r,n-k-r-1,\alpha} = F_{2,7,0.05}$ is shown below

	A	B	C	D	E
1		X_1	X_2	X_3	Y
2	1	1300	15	4	248
3	2	2110	7	6	308
4	3	1935	17	3	239
5	4	1700	8	3	302
6	5	1050	18	2	169
7	6	1455	16	3	223
8	7	2250	5	7	385
9	8	2550	6	6	367
10	9	1765	14	3	232
11	10	1600	13	2	245
12					
13	r	2			
14	SSE(r)	10259.74			
15	SSE	2170.378			
16	s_e^2	361.7296			
17	F	11.1815			
18	$F_{2,7,0.05}$	0.051671			

Figure 2.6.9

According to the rejection rule (2.6.14), since $F = 11.1815 > F_{2,7,0.05} = 0.0517$, the null hypothesis is rejected. The conclusion Lisa can draw from here is that at least one of the variables X_2, X_3 is affecting Y. For a detailed identification, Lisa should conduct the t-based tests discussed above.

2.7. Confidence Intervals for the Regression Coefficients

For the multiple regression model, we defined $b_0, b_1, \dots b_k$ to be the unbiased estimators for the population coefficients $\beta_0, \beta_1, \dots, \beta_k$. If the standard regression assumptions hold, the confidence interval for the β_j coefficient corresponding to $100(1-\alpha)\%$ confidence level is

$$b_j \pm t_{n-k-1,\alpha/2} s_{b_j} \tag{2.7.1}$$

where $t_{n-k-1,\alpha/2}$ is the number for which

$$P\left(t_{n-k-1} > t_{n-k-1,\alpha/2}\right) = \alpha/2$$

and the random variable t_{n-k-1} follows the Student's t distribution with $n-k-1$ degrees of freedom.

Example 2.7

Once having computed the values of b_1 and b_2, Lisa Miller gets more information about the estimates of the corresponding population coefficients β_1 and β_2 by constructing the confidence intervals (2.7.1) for each. This way, she estimates the minimum and maximum values the given coefficients obtain by the confidence level $100(1-\alpha)\%$. The following figure illustrates the computations

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		X_1	X_2	Y	$\hat{y}$	$(\hat{y}-\bar{y})^2$			s_e^2	396.0824666	<--"=E13/(A11-2-1)"		
2	1	1300	15	248	219.8658	2697.165			s_e	19.90182069	<--"=SQRT(J1)"		
3	2	2110	7	308	334.6111	3945.232			$s_{b_1}^2$	0.000563821	<--"=J1/(A10*B13^2*(1-B18^2))"		
4	3	1935	17	239	230.8716	1675.131			s_{b_1}	0.023744906	<--"=SQRT(J3)"		
5	4	1700	8	302	305.6351	1144.815			$s_{b_2}^2$	4.819255488	<--"=J1/(A10*B14^2*(1-B18^2))"		
6	5	1050	18	169	179.3837	8540.777			s_{b_2}	2.195280276	<--"=SQRT(J5)"		
7	6	1455	16	223	217.6672	2930.357							
8	7	2250	5	385	360.3353	7838.5			t_1	1.995951109	<--"=B21/J4"		
9	8	2550	6	367	365.0089	8687.89			t_2	-4.34776177	<--"=B22/J6"		
10	9	1765	14	232	251.4484	414.1887			$t_{7,0.05}$	1.894578605	<--"=T.INV.2T(0.1,7)"		
11	10	1600	13	245	253.173	346.9661			p-value (t_1)	0.086132093	<--"=T.DIST.2T(J8,7)"		
12									p-value (t_2)	0.003364031	<--"=T.DIST.2T(ABS(J9),7)"		
13	s_{x_1}	453.8664		SSE	2772.577	<--"=SUMXMY2(D2:D11,E2:E11)"							
14	s_{x_2}	4.909175		SSR	38221.02	<--"=SUM(F2:F11)"			MSR	19110.51137	<--"=E14/2"		
15	s_y	67.48959		SST	40993.6	<--"=SUM(E13:E14)"			MSE	396.0824666	<--"=J1"		
16	r_{x_1y}	0.865866							F	48.24881932	<--"=J13/J14"		
17	r_{x_2y}	-0.94545		R^2	0.932366	<--"=E14/E15"			$F_{2,7,0.1}$	3.257442051	<--"=F.INV(0.9,2,7)"		
18	$r_{x_1x_2}$	-0.78809		R	0.965591	<--"=CORREL(D2:D11,E2:E11)"			p-value (F)	8.04612E-05	<--"=F.DIST.RT(J16,2,7)"		
19				$\bar{R}^2$	0.913041	<--"=1-(E13/(A11-2-1))/(E15/A10)"							
20	b_0	301.4223		LCL		UCL							
21	b_1	0.047394		-0.00875	β_1	0.103541							
22	b_2	-9.54456		-14.7356	β_2	-4.35354							

Figure 2.7.1

The cells D21 and D22 compute the lower confidence limits for β_1 and β_2 coefficients and the cells F21 and F22 compute the upper confidence limits respectively.

The formulas used are

D21: =B21-T.INV.2T(0.05,7)*J4

F21: =B21+T.INV.2T(0.05,7)*J4

D22: =B22-T.INV.2T(0.05,7)*J6

F22: =B22+T.INV.2T(0.05,7)*J6

2.8. Regression Table

Computations from the preceding sections can be summarized by the regression table. That is a multiple regression analogue of the table shown in Section 1.10. By applying the regression package in Excel using the Data tab, Data Analysis, Regression option as

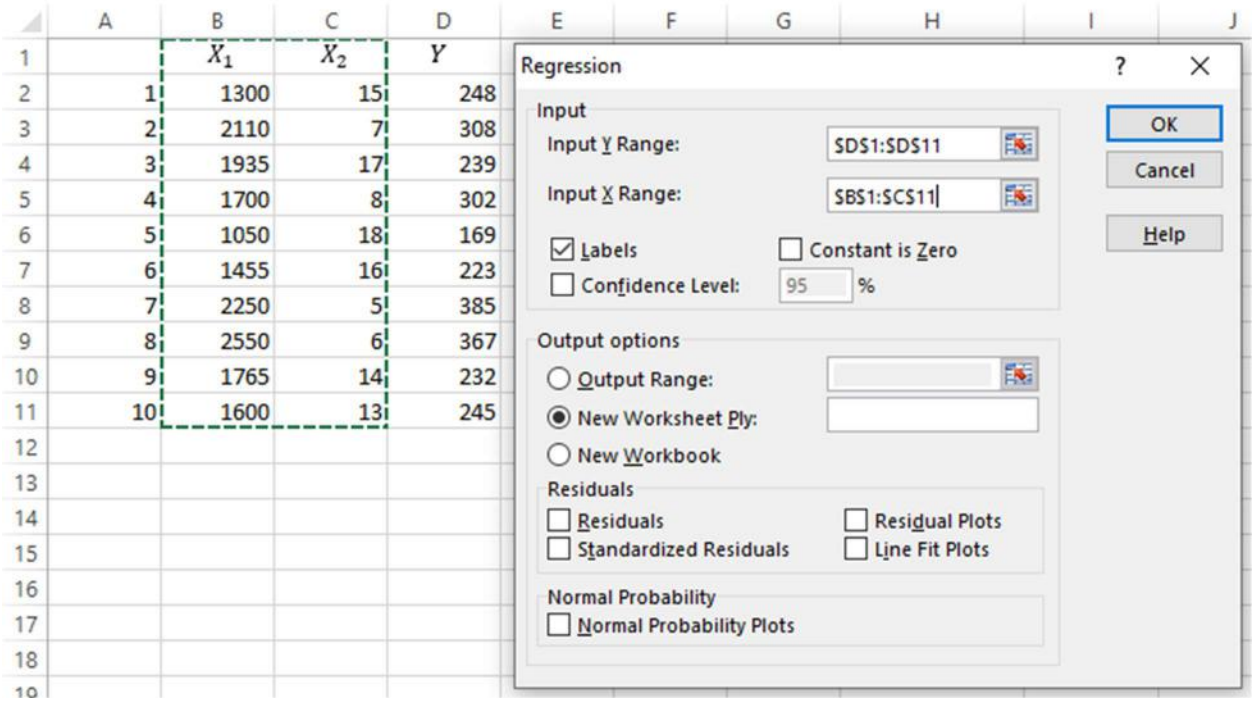


Figure 2.8.1

the summary table shown below is obtained

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.965591							
5	R Square	0.932366							
6	Adjusted R Square	0.913041							
7	Standard Error	19.90182							
8	Observations	10							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	38221.02273	19110.51	48.24882	8.04612E-05			
13	Residual	7	2772.577266	396.0825					
14	Total	9	40993.6						
15									
16		*Coefficients*	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept b_0	301.4223	64.98823685	4.638106	0.002375	147.7495633	455.0951	147.7496	455.0951
18	b_1	0.047394	0.023744906	1.995951	0.086132	-0.008754109	0.103541	-0.00875	0.103541
19	b_2	-9.54456	2.195280276	-4.34776	0.003364	-14.73556863	-4.35354	-14.7356	-4.35354

Figure 2.8.2

Almost all of the coefficients shown in Figure 2.8.2 have been computed in the preceding sections. The following table summarizes the references of each number from Figure 2.8.2 to the corresponding equations in the text.

#	Cell	Description	Equation
1.	B4	R	(2.4.3)
2.	B5	R^2	(2.4.1)
3.	B6	$\bar{R}^2$	(2.4.4)
4.	B7	s_e	(2.5.1)
5.	C12	SSR	(2.3.5)
6.	C13	SSE	(2.3.4)
7.	C14	SST	(2.3.3)
8.	D12	MSR	(2.6.6)
9.	D13	MSE	(2.6.7)
10.	E12	F	(2.6.8)
11.	F12	$P(F_{k,n-k-1} > F)$	mentioned in (2.6.19)
12.	B17	b_0	(2.2.8)
13.	B18	b_1	(2.2.6)
14.	B19	b_2	(2.2.7)

15.	C18	s_{b_1}	square root from (2.5.2)
16.	C19	s_{b_2}	square root from (2.5.3)
17.	D18	t_1	(2.6.2)
18.	D19	t_2	(2.6.2)
19.	E18	$P(t_{n-k-1} > t_1)$	(2.6.4)
20.	E19	$P(t_{n-k-1} > \lvert t_2 \rvert)$	(2.6.4)
21.	F18	$b_1 - t_{n-k-1,\alpha/2} s_{b_1}$	(2.7.1)
22.	G18	$b_1 + t_{n-k-1,\alpha/2} s_{b_1}$	(2.7.1)
23.	F19	$b_2 - t_{n-k-1,\alpha/2} s_{b_2}$	(2.7.1)
24.	G19	$b_2 + t_{n-k-1,\alpha/2} s_{b_2}$	(2.7.1)

Table 2.8

2.9. Dummy Variables

Up until this point in the text, all independent variables considered were numerical in nature. The real estate agent in the examples decided to predict the house price (measured in $1000s) by two explanatory variables - X_1 being the house area measured in square meters, and X_2 being the house age measured in years. What if she now wants to include another variable – location of the house as an additional factor determining the price? Location is not a numerical variable, but rather a categorical variable capturing a certain characteristic of a house.

Let us introduce a dummy variable as categorical independent variable obtaining only two values: yes or no, on or off, red or black (white also works), male or female, etc. The values are recorded as 0 (in case of a certain criteria not satisfied) or 1 (in case of a certain criteria satisfied).

Consider a two-variable model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 \tag{2.9.1}$$

where x_1 is a normal numerical variable and x_2 is dummy – it can either be 0 or 1. Let us consider both cases separately. When $x_2 = 0$ we have (2.9.1) rewritten as

$$\hat{y} = b_0 + b_1 x_1 \tag{2.9.2}$$

and when $x_2 = 1$ we have

$$\hat{y} = b_0 + b_1 x_1 + b_2 = (b_0 + b_2) + b_1 x_1 \tag{2.9.3}$$

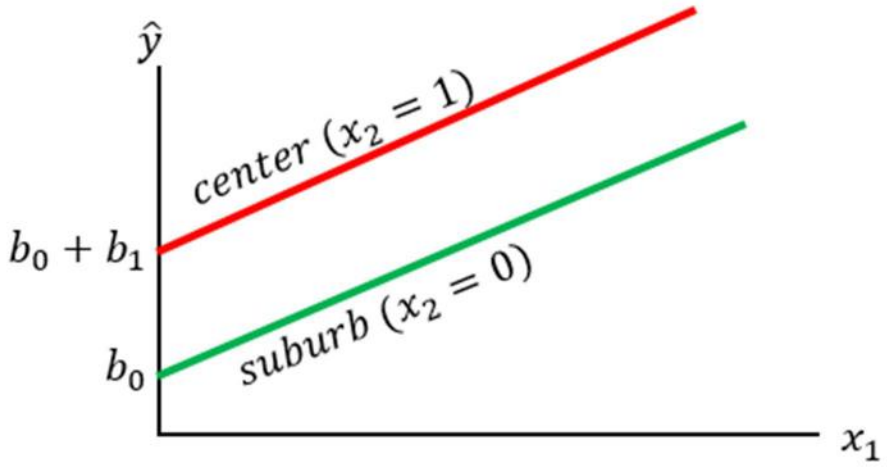


Figure 2.9.1

In any case, (2.9.1) is reduced to a (linear) function of a single variable x_1. We see that the slopes of these two equations are both b_1. Thus, the lines corresponding to these equations are going to be parallel. However, the intercept coefficients differ. b_0 is the intercept in the first line and $b_0 + b_1$ is the slope of another. So, it can be concluded that adding a dummy variable to the existing model only affects the intercept coefficient. In particular, it splits the model into two separate equations – one predicting the value of Y in case of $x_2 = 0$ and another for $x_2 = 1$. The goal is to compute the coefficients b_0, b_1 and b_2 in the least squares sense.

Multiple regression with more than one dummy variables

Every additional dummy variable splits the existing scenarios in two. So, if in case of one dummy variable, there are two of them, that will result in four different scenarios for all possible combinations of the dummy variable values. To make it clear, consider the multiple regression model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \tag{2.9.4}$$

where x_1 is a numerical variable while x_2 and x_3 are dummies. Then, there will be four versions of (2.9.4)

Case 1: $x_2 = 0, x_3 = 0$

$$\hat{y} = b_0 + b_1 x_1 \tag{2.9.5}$$

Case 2: $x_2 = 0, x_3 = 1$

$$\hat{y} = (b_0 + b_3) + b_1 x_1 \tag{2.9.6}$$

Case 3: $x_2 = 1, x_3 = 0$

$$\hat{y} = (b_0 + b_2) + b_1 x_1 \tag{2.9.7}$$

Case 4: $x_2 = 1. x_3 = 1$

$$\hat{y} = (b_0 + b_2 + b_3) + b_1 x_1 \quad (2.9.8)$$

Again, the slopes of these lines is the same b_1 and intercepts differ. The Figure 2.9.2 illustrates a possible scenario.

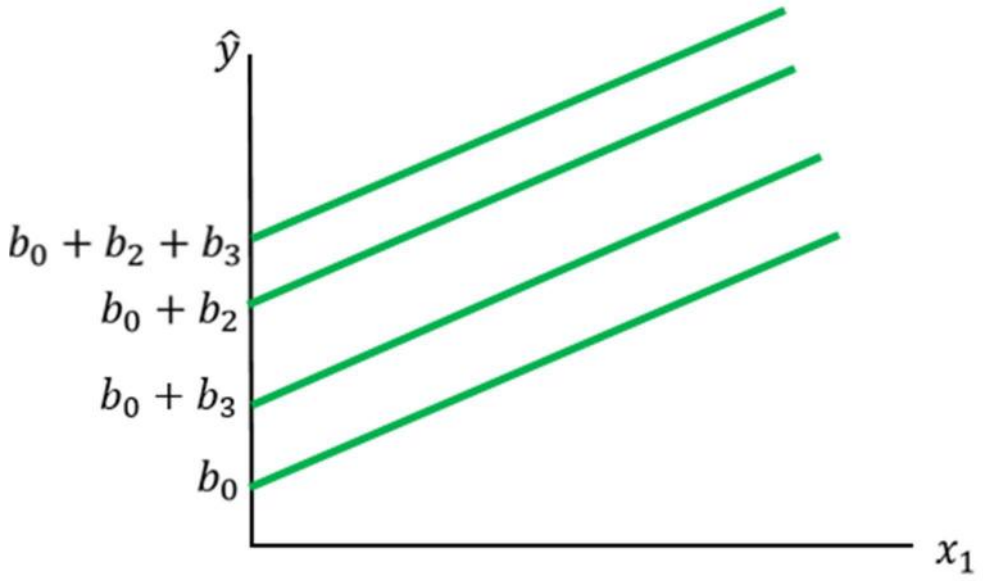


Figure 2.9.2

Example 2.9.1

Consider again the example of the real estate agent, Lisa Miller. She needs to predict the house price with 2000 square meters of area located in the city center and another house located outside of the city center with 1850 square meters of area. Initially she had the house price Y as a dependent variable with X_1 – house area measured in square meters as an explanatory variable. She thinks that in addition to area, location is a significant determinant of the house price in a given city. So, she decides to express house location as a dummy variable X_2 whose realizations are either 1 if the house is located in the city center, or 0 if it is located elsewhere. Thus, she is going to obtain the fitted regression equation (2.9.1) split into two separate equations (2.9.2) and (2.9.3).

She collects the data of observations on Y, X_1 and X_2 as shown below

	A	B	C	D
1		X_1	X_2	Y
2	1	1300	1	248
3	2	2110	1	308
4	3	1935	0	239
5	4	1700	1	302
6	5	1050	0	169
7	6	1455	0	223
8	7	2250	1	385
9	8	2550	1	367
10	9	1765	0	232
11	10	1600	1	245

Figure 2.9.2

e.g. the 4th record indicates that a house with 1700 square meters located in the city center was sold for $302 000 and the 5th record shows a house with 1050 square meters of area located outside of the city center was sold for $169 000.

Lisa computes b_0, b_1 and b_2 according to formulas (2.2.8), (2.2.6) and (2.2.7)

	A	B	C	D	E	F	G
1		X_1	X_2	Y		s_{x_1}	453.8664
2	1	1300	1	248		s_{x_2}	0.516398
3	2	2110	1	308		s_y	67.48959
4	3	1935	0	239		r_{x_1y}	0.865866
5	4	1700	1	302		r_{x_2y}	0.714779
6	5	1050	0	169		$r_{x_1x_2}$	0.417658
7	6	1455	0	223			
8	7	2250	1	385		b_0	57.23202
9	8	2550	1	367		b_1	0.102187
10	9	1765	0	232		b_2	55.90543
11	10	1600	1	245			

Figure 2.9.3

The procedure used in Figure 2.9.3 above is examined in detail in Section 2.2. As a shortcut, she could have alternatively gotten the same coefficients using the regression package of Excel in the Data tab, Data Analysis, Regression option. The resulting output table would be

I	J	K	L	M	N	O	P	Q
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.949097							
R Square	0.900785							
Adjusted R Sq	0.872437							
Standard Erro	24.10453							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*gnificance F*			
Regression	2	36926.4	18463.2	31.77676	0.000308			
Residual	7	4067.199	581.0284					
Total	9	40993.6						
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept b_0	57.23202	32.53869	1.758892	0.122	-19.7097	134.1738	-19.7097	134.1738
b_1	0.102187	0.019484	5.244718	0.001193	0.056115	0.148259	0.056115	0.148259
b_2	55.90543	17.12452	3.264643	0.013776	15.41238	96.39847	15.41238	96.39847

Figure 2.9.4

Note that addition of the house location as an additional (dummy) variable contributed to the explanatory power of the regression and the existing R^2 rose from 75% to 90% and adjusted coefficient of determination $\bar{R}^2$ rose to 87%.

Having the values of b_0, b_1 and b_2 computed, Lisa constructs the fitted regression equation (2.9.1) to be

$$\hat{y} = 57.2320 + 0.1022x_1 + 55.9054x_2$$

By inserting the desired values of x_1 and x_2 in the above equation, Lisa can predict the corresponding house price. The figure below shows the expected prices of houses Lisa needed to compute.

	A	B	C	D	E	F	G
1		X_1	X_2	Y		s_{x_1}	453.8664
2	1	1300	1	248		s_{x_2}	0.516398
3	2	2110	1	308		s_y	67.48959
4	3	1935	0	239		r_{x_1y}	0.865866
5	4	1700	1	302		r_{x_2y}	0.714779
6	5	1050	0	169		$r_{x_1x_2}$	0.417658
7	6	1455	0	223			
8	7	2250	1	385		b_0	57.23202
9	8	2550	1	367		b_1	0.102187
10	9	1765	0	232		b_2	55.90543
11	10	1600	1	245			
12							
13	x_1	x_2	$\hat{y}$				
14	2000	1	317.512				
15	1850	0	246.2784				

Figure 2.9.5

As illustrated above, the house with 2000 square meters of area located in the city center is expected to be sold for $317 512 and the house with 1850 square meters of area located outside of the city center is expected to be sold for $246 278.

Generally, Lisa can construct the prediction graphs (lines) corresponding to (2.9.2) and (2.9.3) as functions of area only as illustrated below

	A	B	C	D	E	F	G	H
1		X_1	X_2	Y		x_1	$\hat{y}\,(x_2=0)$	$\hat{y}\,(x_2=1)$
2	1	1300	1	248		1000	159.4193	215.3247
3	2	2110	1	308		1200	179.8567	235.7622
4	3	1935	0	239		1400	200.2942	256.1996
5	4	1700	1	302		1600	220.7316	276.6371
6	5	1050	0	169		1800	241.1691	297.0745
7	6	1455	0	223		2000	261.6065	317.5120
8	7	2250	1	385		2200	282.0440	337.9494
9	8	2550	1	367		2400	302.4814	358.3869
10	9	1765	0	232		2600	322.9189	378.8243
11	10	1600	1	245		2800	343.3563	399.2618
12								
13	b_0	57.232						
14	b_1	0.10219						
15	b_2	55.9054						

Figure 2.9.6

The cell G2 contains the formula "=$B$13+$B$14*F2" and the cell H2 – "=$B$13+$B$14*F2+$B$15".

Example 2.9.2

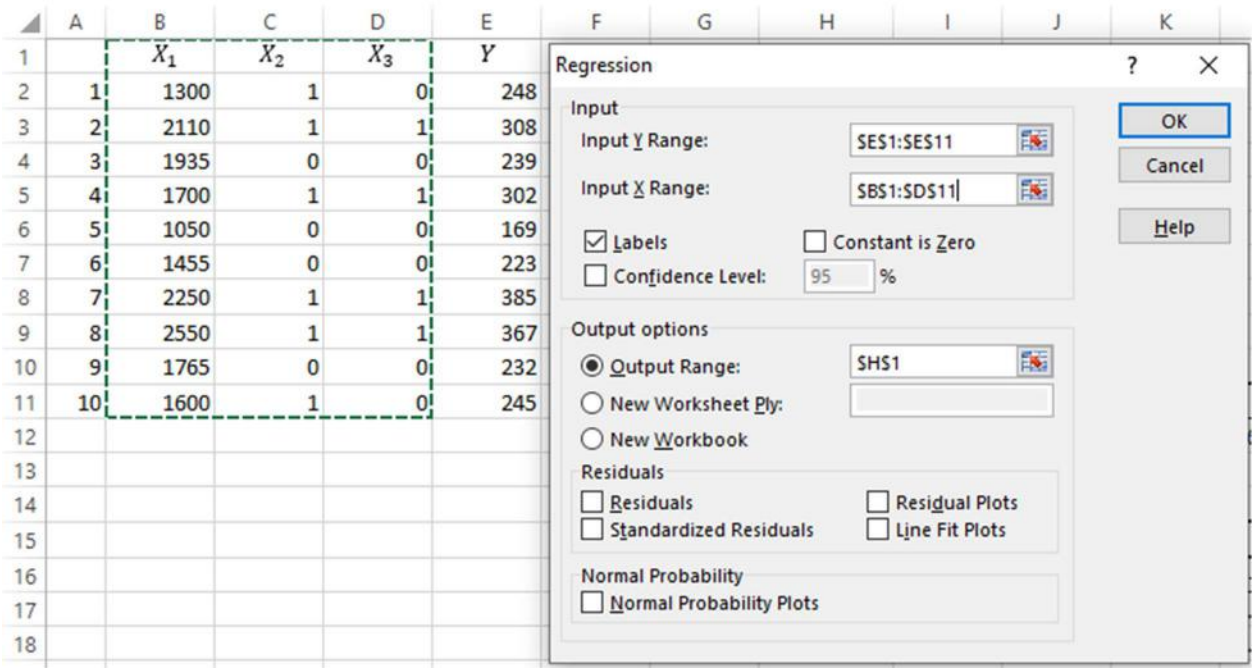
Suppose Lisa Miller wants to predict the prices of two houses. The first with 1880 square meters of area located in the city center with the available parking space and another with 2150 square meters of a house located outside of the city center without an available parking space.

She considers adding another dummy variable to the existing model in the previous Example 2.9.1. The new variable X_3 equals 1 if there is an available parking space around a house and 0 otherwise. She collects the sample data of 10 records shown below

	A	B	C	D	E
1		X_1	X_2	X_3	Y
2	1	1300	1	0	248
3	2	2110	1	1	308
4	3	1935	0	0	239
5	4	1700	1	1	302
6	5	1050	0	0	169
7	6	1455	0	0	223
8	7	2250	1	1	385
9	8	2550	1	1	367
10	9	1765	0	0	232
11	10	1600	1	0	245

Figure 2.9.7

e.g. the 4th record now indicates that a house with 1700 square meters of area, located in the city center with an available parking space was sold for $302 000 and the 5th record indicates that a house with 1050 square meters of area located outside of the city center without an available parking space was sold for $169 000. The fitted equation whose coefficients Lisa has to compute is (2.9.4). She proceeds with the regression package described above



	A	B	C	D	E
1		X_1	X_2	X_3	Y
2	1	1300	1	0	248
3	2	2110	1	1	308
4	3	1935	0	0	239
5	4	1700	1	1	302
6	5	1050	0	0	169
7	6	1455	0	0	223
8	7	2250	1	1	385
9	8	2550	1	1	367
10	9	1765	0	0	232
11	10	1600	1	0	245

Figure 2.9.8

which yields the regression tables

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.963814
R Square	0.928937
Adjusted R Squ	0.893405
Standard Error	22.03455
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	3	38080.47	12693.49	26.14403	0.000764
Residual	6	2913.13	485.5216		
Total	9	40993.6			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept b_0	94.22299	38.2151	2.465596	0.048748	0.714012	187.732	0.714012	187.732
b_1	0.078341	0.023589	3.321088	0.015983	0.020621	0.136062	0.020621	0.136062
b_2	38.68206	19.23137	2.011404	0.090977	-8.37541	85.73953	-8.37541	85.73953
b_3	38.96521	25.2735	1.541742	0.174074	-22.8768	100.8072	-22.8768	100.8072

Figure 2.9.9

With these coefficients at hand, (2.9.4) becomes

$$\hat{y} = 94.2230 + 0.0783x_1 + 38.6801x_2 + 38.9652x_3$$

Next, she computes the expected price of houses of her interest. The results are shown below

	A	B	C	D	E	F	G	H	I	J
1		X_1	X_2	X_3	Y		x_1	x_2	x_3	$\hat{y}$
2	1	1300	1	0	248		2150	1	1	340.3041
3	2	2110	1	1	308		1880	0	0	241.5047
4	3	1935	0	0	239					
5	4	1700	1	1	302					
6	5	1050	0	0	169					
7	6	1455	0	0	223					
8	7	2250	1	1	385					
9	8	2550	1	1	367					
10	9	1765	0	0	232					
11	10	1600	1	0	245					
12										
13	Intercept b_0	94.223								
14	b_1	0.07834								
15	b_2	38.6821								
16	b_3	38.9652								

Figure 2.9.10

In particular, with $x_1 = 1880, x_2 = 1, x_3 = 1$, she obtains $\hat{y} = \$340\,30$, and for $x_1 = 2150, x_2 = 0, x_3 = 0$, $\hat{y} = \$241\,505$.

The value in J2 is computed by the formula "=$B$13+$B$14*G2+$B$15*H2+$B$16*I2" and the value in J3 by "=$B$13+$B$14*G3+$B$15*H3+$B$16*I3".

Remark: Note that the p-value for the β_3 coefficient is $0.1741 > \alpha = 0.1$*. So, based on the significance level of 10%, X_3 does not affect the house price and needs not be included in the regression model. We are ignoring this fact for the current example.*

In addition, Lisa can specify equations (2.9.5), (2.9.6), (2.9.7) and (2.9.8) for four different scenarios.

$$\hat{y} = 94.223 + 0.0783x_1 \qquad (2.9.5')$$

$$\hat{y} = 133.188 + 0.0783x_1 \qquad (2.9.6')$$

$$\hat{y} = 132.905 + 0.0783x_1 \qquad (2.9.7')$$

$$\hat{y} = 171.87 + 0.0783x_1 \qquad (2.9.8')$$

In the following figure, different scenarios are listed below. e.g. the scenario 2 implies a house located outside of the city center with an available parking space and the scenario 4 implies a house located in the city center located in the city center. The columns H to K carry out the computations of the equations (2.9.5)-(2.9.8). For the listed values of x_1 in the L column, the formula in the cell H2 extended throughout the array is

"=$B$13+$B$14*$L2+$B$15*M$14+$B$16*M$15".

	A	B	C	D	E	F	G	H	I	J	K
1		X_1	X_2	X_3	Y		x_1	$\hat{y}(1)$	$\hat{y}(2)$	$\hat{y}(3)$	$\hat{y}(4)$
2	1	1300	1	0	248		1000	172.5643	211.5295	211.2464	250.2116
3	2	2110	1	1	308		1200	188.2326	227.1978	226.9147	265.8799
4	3	1935	0	0	239		1400	203.9009	242.8661	242.5829	281.5481
5	4	1700	1	1	302		1600	219.5691	258.5343	258.2512	297.2164
6	5	1050	0	0	169		1800	235.2374	274.2026	273.9195	312.8847
7	6	1455	0	0	223		2000	250.9057	289.8709	289.5877	328.5529
8	7	2250	1	1	385		2200	266.5739	305.5392	305.256	344.2212
9	8	2550	1	1	367		2400	282.2422	321.2074	320.9243	359.8895
10	9	1765	0	0	232		2600	297.9105	336.8757	336.5925	375.5578
11	10	1600	1	0	245		2800	313.5788	352.544	352.2608	391.226
12											
13	Intercept b_0	94.223						1	2	3	4
14	b_1	0.07834					x_2	0	0	1	1
15	b_2	38.6821					x_3	0	1	0	1
16	b_3	38.9652									

Figure 2.9.10

Finally, the Figure 2.9.2 can be drawn in terms of these numbers as

L	M	N	O	P
x_1	$\hat{y}(1)$	$\hat{y}(2)$	$\hat{y}(3)$	$\hat{y}(4)$
1000	172.5643	211.5295	211.2464	250.2116
1200	188.2326	227.1978	226.9147	265.8799
1400	203.9009	242.8661	242.5829	281.5481
1600	219.5691	258.5343	258.2512	297.2164
1800	235.2374	274.2026	273.9195	312.8847
2000	250.9057	289.8709	289.5877	328.5529
2200	266.5739	305.5392	305.256	344.2212
2400	282.2422	321.2074	320.9243	359.8895
2600	297.9105	336.8757	336.5925	375.5578
2800	313.5788	352.544	352.2608	391.226
	1	2	3	4
x_2	0	0	1	1
x_3	0	1	0	1

Figure 2.9.11

Note that there are 4 separate lines on the graph. The reason why only three appear is that the differences between the predicted house values for the scenarios 2 and 3 are very small making the distance between the corresponding lines invisible.

We can conclude that Lisa Miller has four equations at her disposal. Depending on the desired scenario, she can make predictions using the corresponding equation.

Chapter 3. Nonlinear Regression

3.1. Quadratic Regression

The simple regression model examined in Chapter 1 was a linear model implying linear relationship between the dependent and independent variables. However, in many applied problems, the dependence may not be linear. There are various types of dependences out of which we only investigate a quadratic model.

In Figure 1.3.3, the scatter plot of X and Y was illustrated. Since the scattered points are clustered around the straight line, it is expected that the relationship between these variables will be linear. This fact is shown in Figure 2.10.1 below

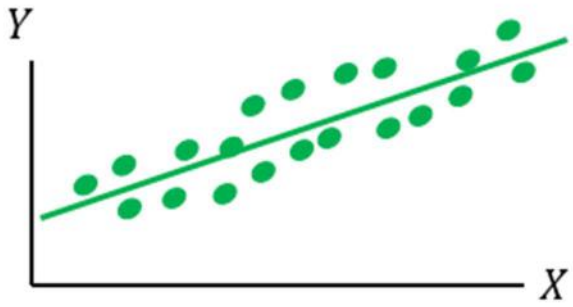


Figure 3.1.1

Note that the linear pattern is preserved for large and small values of X. Therefore, the residuals (differences between points on the line and the observed points for given values of X) is not showing any unusual change as we move through the extreme values of X on either side. When detecting the linear pattern of this kind, one would follow the procedures described in Chapter 1 to build a simple regression model. What if the scatter plot looks as follows?

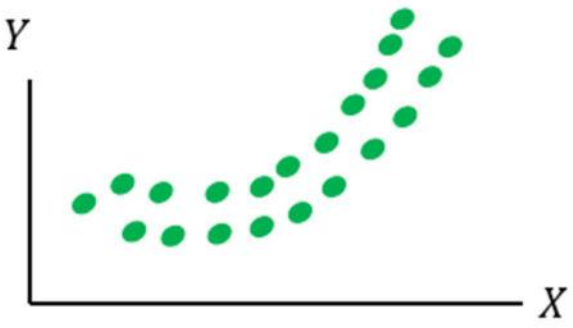


Figure 3.1.2

It is obvious that the scatter plot does not follow a linear pattern. Fitting a straight line into the points would be unreasonable. It would introduce higher residuals at the extreme values of *X* as the line seems to be diverging from the observed points. The following figure illustrates this effect

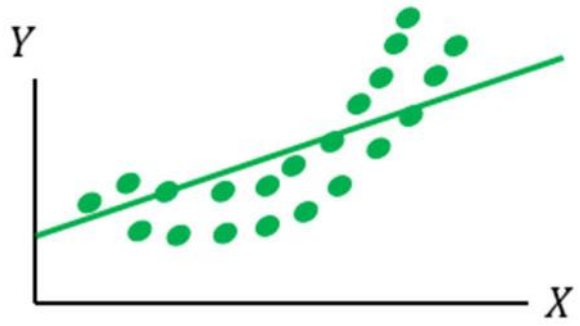


Figure 3.1.3

In an ideal case, there must be a more flexible fitting curve capable of following the observed pattern. The quadratic regression aims to capture the pattern following the shape of parabola. Figure 3.1.4 shows the fitting curve corresponding to the quadratic equation

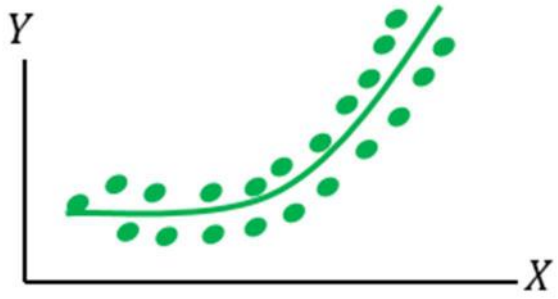


Figure 3.1.4

The quadratic regression model may be considered when scatter plot takes one of the following shapes

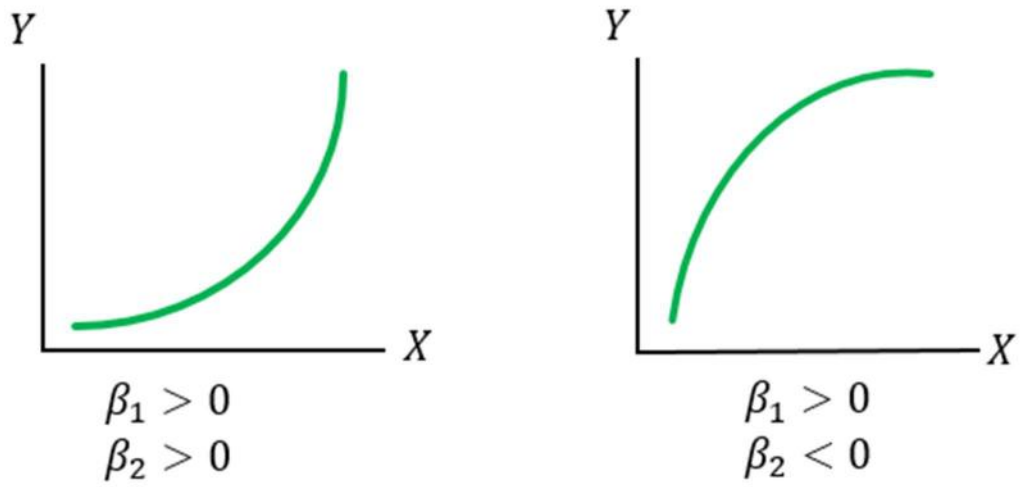


Figure 3.1.5

or

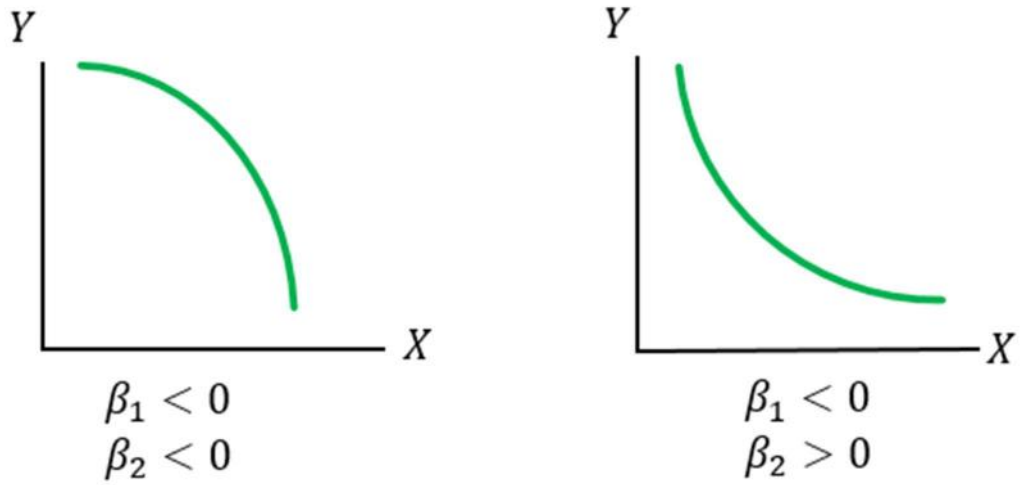


Figure 3.1.6

Recall that the simple linear regression model was defined as

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3.1.1}$$

which was estimated by

$$\hat{y} = b_0 + b_1 x \tag{3.1.2}$$

The quadratic regression model extends the simple linear regression model (3.1.1) by adding the quadratic component $\beta_2 X^2$ as follows

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \tag{3.1.3}$$

Note that there is still an only explanatory variable X. The second term $\beta_2 X^2$ involves the same variable and introduces concavity to the model. Corresponding fitting quadratic equation is

$$\hat{y} = b_0 + b_1 x + b_2 x^2 \tag{3.1.4}$$

Before computing the coefficients b_0, b_1 and b_2, the importance of quadratic term $\beta_2 X^2$ has to be measured. This is done by the following hypothesis

$$\begin{aligned} H_0&: \beta_2 = 0 \\ H_1&: \beta_2 \neq 0 \end{aligned} \tag{3.1.5}$$

Accepting the null hypothesis suggests that the quadratic term improves the model while rejecting it implies the opposite. It can be shown that under general conditions,

$$t = \frac{b_2 - \beta_2}{s_{b_2}} \tag{3.1.6}$$

follows a Student's t distribution with $n - 3$ degrees of freedom. ($n - 3$ comes from $n - k - 1$ where k is the number of explanatory coefficients - β_1 and β_2 in quadratic model). So, the rejection rule for the hypothesis (3.1.5) is to

$$reject\ H_0\ if\ |t| = \frac{b_2 - \beta_2}{s_{b_2}} > |t_{n-3,\alpha/2}| \tag{3.1.7}$$

where $t_{n-3,\alpha}$ is the number for which

$$P(|t_{n-3}| > |t_{n-3,\alpha\backslash 2}|) = \alpha$$

and the corresponding p-value is

$$p - value = P(|t_{n-3}| > |t|) \tag{3.1.8}$$

It is useful to ultimately based the decision of whether the model (3.1.2) must be extended by the quadratic component in (3.1.3) by comparing the coefficients of determination R^2 from the linear model and the adjusted coefficient of determination $\bar{R}^2$ from the quadratic model. The following example illustrates the comparison

Example 3.1

Tom Davis produces and exports wine. There are different regulations in each importing country. According to the regulations, the wine purity must be at a predetermined minimum

level in order to be allowed for selling on the market. Here is the list of purity levels for each importing country

Country	Purity of Wine in %
A	83
B	95
C	86
D	80
E	99
F	93
G	85

Table 3.1.1

As we can see, Tom is only allowed to export wines if its purity is at least 80%. Some countries have quite tough requirements. For example, the country E will only allow a wine with at least 99% purity. Tom uses a unique technology and method for wine production. So he needs a model to estimate (or predict) the wine purity based on time. He has observed 15 production outputs and recorded the purity level as a dependent variable Y and time in weeks it takes to get to this level as an independent explanatory variable X. The figure below shows the sample observations and the scatter diagram

	A	B	C
1		X	Y
2	1	1	2
3	2	3	4
4	3	4	6
5	4	5	8
6	5	6	12
7	6	7	15
8	7	8	19
9	8	9	23
10	9	11	33
11	10	13	44
12	11	15	58
13	12	16	65
14	13	17	75
15	14	19	91
16	15	20	99

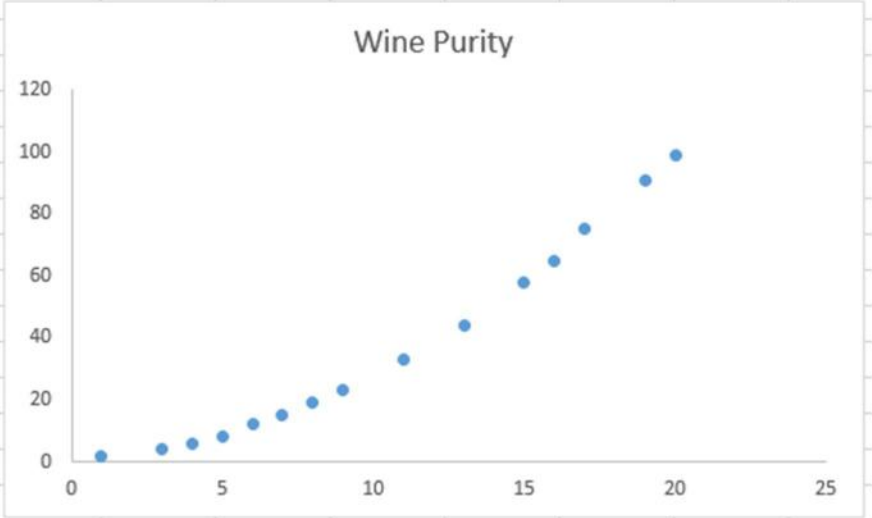


Figure 3.1.7

e.g. the 10th record indicates that wine having had been filtered for 13 weeks reached 44% of purity. Tom has two options to predict the wine purity. He can either use the linear simple

regression model (3.1.2) or the quadratic model (3.1.4). Even though the scatter diagram indicates a clear non-linear pattern and Tom is planning to use the quadratic model, linear regression is also expected to be a suitable model predicting the wine purity with a reasonably high precision. So, Tom starts with what he expects to be a better model – quadratic regression.

First, he computes the coefficient estimates for β_0, β_1 and β_2 from (3.1.3), which are b_0, b_1 and b_2. This is done by the scatter plot diagram directly by selecting the polynomial model of order 2.

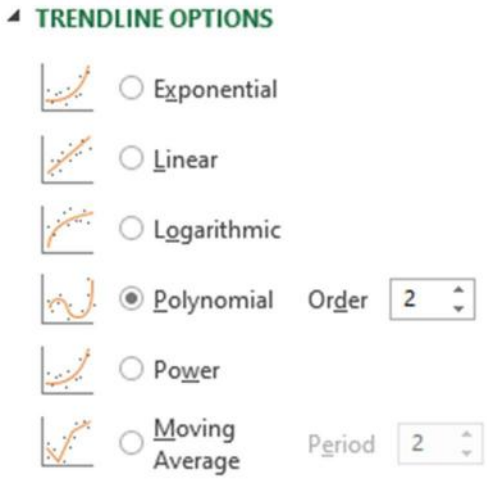


Figure 3.1.8

Selecting "Display Equation on chart" checkbox on the same window yields the quadratic regression equation shown in Figure 3.1.9 below. The values of the coefficients are copied in the cells B18-B20.

	A	B	C
1		X	Y
2	1	1	2
3	2	3	4
4	3	4	6
5	4	5	8
6	5	6	12
7	6	7	15
8	7	8	19
9	8	9	23
10	9	11	33
11	10	13	44
12	11	15	58
13	12	16	65
14	13	17	75
15	14	19	91
16	15	20	99
17			
18	b_0	0.9526	
19	b_1	0.4077	
20	b_2	0.2269	

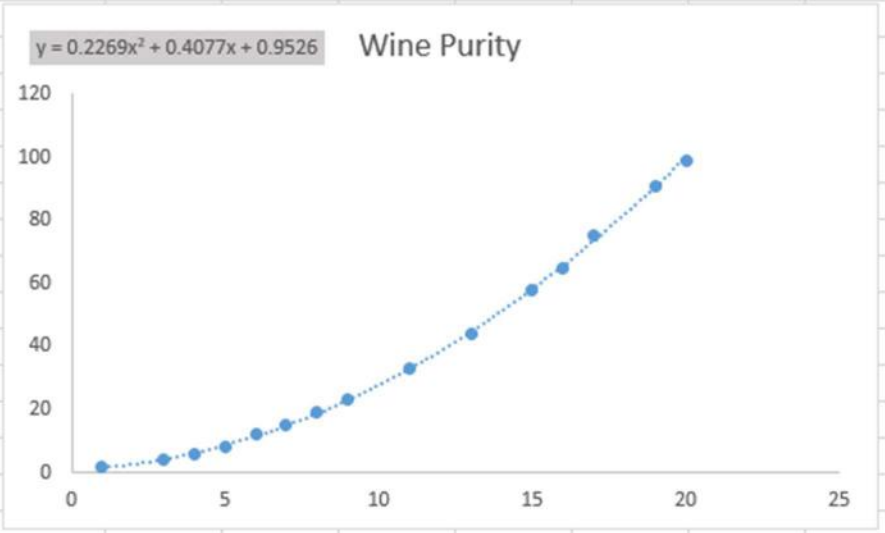


Figure 3.1.9

The equation (3.1.4) with these values is

$$\hat{y} = 0.9526 + 0.4077x + 0.9526x^2$$

Tom can now estimate Y values by $\hat{y}$ for all observations on X. The following figure illustrates relevant computations.

	A	B	C	D	E	F	G	H
1		X	Y	$\hat{y}$	$(\bar{y}-\hat{y})^2$			
2	1	1	2	1.5872	1249.349			
3	2	3	4	4.2178	1070.306			
4	3	4	6	6.2138	943.6897			
5	4	5	8	8.6636	799.1778			
6	5	6	12	11.5672	643.4407			
7	6	7	15	14.9246	484.3843			
8	7	8	19	18.7358	331.1502			
9	8	9	23	23.0008	194.1155			
10	9	11	33	32.8922	16.33076			
11	10	13	44	44.5988	58.75938			
12	11	15	58	58.1206	448.9003			
13	12	16	65	65.5622	819.612			
14	13	17	75	73.4576	1334.022			
15	14	19	91	90.6098	2881.163			
16	15	20	99	99.8666	3960.596			
17								
18	b_0	0.9526		SSR	15235	<--"=SUM(E2:E16)"		
19	b_1	0.4077		SSE	4.949764	<--"=SUMXMY2(C2:C16,D2:D16)"		
20	b_2	0.2269		SST	15239.95	<--"=E18+E19"		
21				R^2	0.999675	<--"=E18/E20"		
22				$\bar{R}^2$	0.999621	<--"=1-(E19/A13)/(E20/A15)"		

Figure 3.1.10

In order to compute the *sum of squares errors* and *sum of squares regression,* $\hat{y}$ values are first computed in the column D by "=$B$18+$B$19*B2+$B$20*B2^2" in D2 and the squared differences between the mean value of observed Y and the estimated value of Y are computed in the column E by "=(AVERAGE($C$2:$C$16)-D2)^2". Note that the R^2 coefficient and adjusted coefficient of determination $\bar{R}^2$ are close to 100% meaning that the model is extremely accurate. So, Tom can almost precisely predict the wine purity for a given filtering time.

Tom builds the simple linear regression model for comparison. The figure below shows the coefficients computed

	A	B	C
1		X	Y
2	1	1	2
3	2	3	4
4	3	4	6
5	4	5	8
6	5	6	12
7	6	7	15
8	7	8	19
9	8	9	23
10	9	11	33
11	10	13	44
12	11	15	58
13	12	16	65
14	13	17	75
15	14	19	91
16	15	20	99
17			
18	b_0	-17.3681	
19	b_1	5.289096	

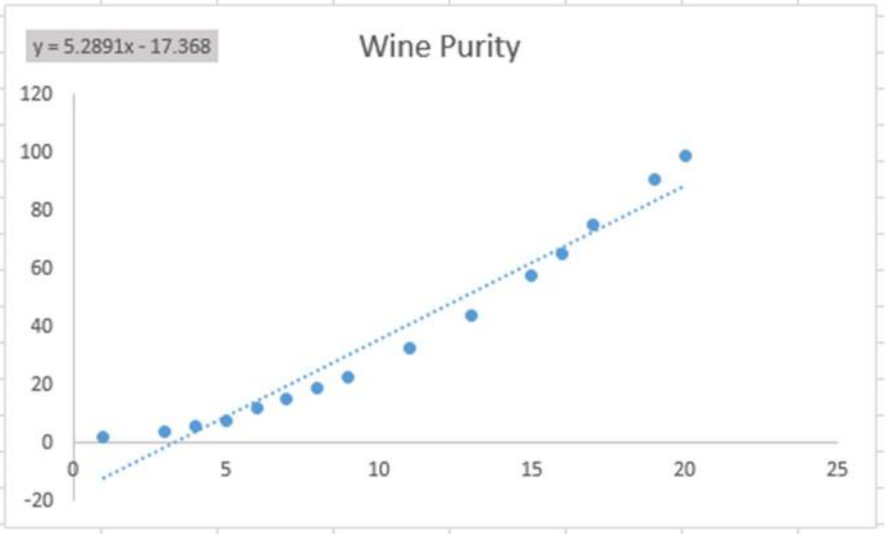


Figure 3.1.11

The equation (3.1.2) is

$$\hat{y} = -17.3681 + 5.2891x$$

Similar computations as in quadratic model above yields the coefficient of determination equal to 0.9563.

	A	B	C	D	E	F	G	H
1		X	Y	$\hat{y}$	$(\bar{y}-\hat{y})^2$			
2	1	1	2	-12.079	2402.205			
3	2	3	4	-1.50077	1477.18			
4	3	4	6	3.788329	1098.591			
5	4	5	8	9.077425	775.9516			
6	5	6	12	14.36652	509.261			
7	6	7	15	19.65562	298.5194			
8	7	8	19	24.94471	143.727			
9	8	9	23	30.23381	44.8836			
10	9	11	33	40.812	15.04409			
11	10	13	44	51.3902	209.0009			
12	11	15	58	61.96839	626.7541			
13	12	16	65	67.25749	919.5543			
14	13	17	75	72.54658	1268.304			
15	14	19	91	83.12478	2133.649			
16	15	20	99	88.41387	2650.246			
17								
18	b_0	-17.3681		SSR	14572.87	<--"=SUM(E2:E16)"		
19	b_1	5.289096		SSE	666.0621	<--"=SUMXMY2(C2:C16,D2:D16)"		
20				SST	15238.93	<--"=E18+E19"		
21				R^2	0.956292	<--"=E18/E20"		

Figure 3.1.12

As long as $R^2 = 0.9563$, it can be concluded that the linear model also provides a reasonably good explanatory power. However, comparing $\bar{R}^2$ from the quadratic model with R^2 of the linear model suggests that the quadratic model outperforms its linear counterpart.

As a conclusion, Tom can now predict times it takes to be able to export wine to various countries according to Table 3.1.1 with a great accuracy.

Chapter 4. Time Series Models

4.1. Introduction

All models examined so far in the text involved dependent random variable explained by one or more predictor (independent) variables. A random variable is a variable that can obtain various values at a given point in time. Sample observations on these variables helped construct the models later used for making predictions.

The objective of this chapter is to construct forecasting models based on time series. *Time Series* is a data recorded over successive increments of time.

Before selecting the prediction model for a given time series data, it is important to identify the data pattern. There can typically be four basic data pattern for the time series: horizontal, trend, cyclical and seasonal.

When time series data fluctuates around a constant level, horizontal or stationary pattern exists.

On the other hand, when the values of time series data grow or decline over a certain time period, the trend pattern emerges. Trend is a long term component of a time series data that that represents growth or decline over an extended period of time.

A wavelike fluctuations of observations from a time series data around the trend indicates a cyclical pattern.

Fluctuations repeating themselves year after year imply seasonal pattern. These fluctuations are influence by seasonal factors and usually have the effects of the same magnitude every year.

In this chapter, we uncover these patterns and examine different models used for predictions.

4.2. Autocorrelation Coefficient

When a measurements of variables are made over time, observations in different time periods frequently tend to be related. This relation is measured by autocorrelation coefficient that is formally defined as a correlation between a variable and itself lagged by one or more time periods. The following equation provides a formula for computing the lag k autocorrelation coefficient between observations y_t and y_{t-k}

$$r_k = \frac{\sum_{t=k+1}^{n}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}, k = 0,1,2 \dots \quad (4.2.1)$$

where

r_k is the autocorrelation coefficient for a lag of k periods

$\bar{y}$ is the mean value of the series

y_t is the observation in time period t

y_{t-k} is the observation at time period $t-k$ which is k time periods earlier than t

The autocorrelation coefficient obtains values within the interval of [-1,1] and implies a certain degree of similarity between the original data and its lagged version. $r_k = 1$ implies a perfect positive correlation between the original data and the lagged data while $r_k = -1$ implies a perfect negative correlation. The following example illustrates the computation.

Example

Leo Johnson, a small book store owner is interested to know how much autocorrelation between the book sales data and its lagged version is. He computes the autocorrelation coefficients for $k = 1$ as shown in the following figure

	A	B	C	D	E	F	G	H
1	*Time t*	*Month*	y_t	y_{t-1}	$(y_t - \bar{y})(y_{t-1} - \bar{y})$	$y_t - \bar{y}$		
2	1	January	125			370.5625		
3	2	February	133	125	216.5625	126.5625		
4	3	March	128	133	182.8125	264.0625		
5	4	April	136	128	134.0625	68.0625		
6	5	May	147	136	-22.6875	7.5625		
7	6	June	143	147	-3.4375	1.5625		
8	7	July	138	143	7.8125	39.0625		
9	8	August	151	138	-42.1875	45.5625		
10	9	September	149	151	32.0625	22.5625		
11	10	October	161	149	79.5625	280.5625		
12	11	November	157	161	213.5625	162.5625		
13	12	December	163	157	239.0625	351.5625		
14								
15				r_1	0.5960	<--"=SUM(E3:E13)/SUM(F2:F13)"		

Figure 4.2.1

$r_1 \approx 0.6$ implies that the correlation between the originally observed data and its lagged data by one period of time are correlated by 60%. So, the successive book sales are somewhat correlated with each other. The values in E column are computed by the formula "=(C3-AVERAGE($C$2:$C$13))*(D3-AVERAGE($C$2:$C$13))" in E3. Note that computations start

from the E3 because of the index of summation in the numerator of (4.2.1). The formula in F2 is "=(C2-AVERAGE($C$2:$C$13))^2". The denominator in (4.2.1) is the summation of all squared differences between the observed values of Y and its mean value. Ultimately, the sum of the values in column E represents the numerator of (4.2.1) and the sum of all values in column F represents the denominator. Their quotient is r_1 computed in E15.

If on the other hand Leo computed the autocorrelation coefficient for lag $k = 2$, it would be computed as follows

	A	B	C	D	E	F	G	H
1	*Time t*	*Month*	y_t	y_{t-2}	$(y_t - \bar{y})(y_{t-2} - \bar{y})$	$y_t - \bar{y}$		
2	1	January	125			370.5625		
3	2	February	133			126.5625		
4	3	March	128	125	312.8125	264.0625		
5	4	April	136	133	92.8125	68.0625		
6	5	May	147	128	-44.6875	7.5625		
7	6	June	143	136	10.3125	1.5625		
8	7	July	138	147	-17.1875	39.0625		
9	8	August	151	143	-8.4375	45.5625		
10	9	September	149	138	-29.6875	22.5625		
11	10	October	161	151	113.0625	280.5625		
12	11	November	157	149	60.5625	162.5625		
13	12	December	163	161	314.0625	351.5625		
14								
15				r_2	0.4618	<--"=SUM(E4:E13)/SUM(F2:F13)"		

Figure 4.2.2

The column D in the above figure contains the lagged data by 2 periods of time. Computations of the numerator of (4.2.1) starts from E4 now with the formula "=(C4-AVERAGE($C$2:$C$13))*(D4-AVERAGE($C$2:$C$13))" . F column remains the same and the r_2 coefficient in E15 turns out to be 0.4618. So, Leo can conclude that the original data is correlated to its lagged version of one period of time than two. In the context of the book sales, this means that successive book sales are more correlated with each other than the book sales of every second month.

4.3. Error Estimators

Residual in the text above was defined as

$$e_t = y_t - \hat{y}_t$$

the difference between the actually observed value y_t and its forecast value $\hat{y}_t$. In the following sections we examine various forecasting models to obtain $\hat{y}_t$. Regardless of the model, we can measure the forecast error in different ways. Measures of forecast errors help evaluate a forecasting technique and estimate parameters of a given model by its optimization.

One method for measuring a forecasting technique is the *mean absolute deviation (MAD).* It measures forecasting accuracy by averaging absolute deviations of the forecast value from the actually observed value

$$MAD = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t| \tag{4.3.1}$$

MAD is a useful measure on its own. It is expressed in the same units as the original time series and provides an average deviation regardless of direction.

Another way of evaluating a forecasting technique is *mean squared error (MSE).*

$$MSE = \frac{1}{n}\sum_{t=1}^{n} (y_t - \hat{y}_t)^2 \tag{4.3.2}$$

It first sums the squared deviations from time series and its forecast and then divides by number of observations. Computation of the MSE frequently leads to extremely large values. Its usefulness becomes clear when we need to provide an analytical derivations of parameters of a given forecasting method. In particular, since MSE is the average of squared deviations, it is an easily differentiable function and hence, minimization problem given the model $\hat{y}$ with some parameters are analytically solvable.

The square root from MSE is the *root mean squared error (RMSE)* which brings MSE back to units of original time series and thus, the magnitude of RMSE is interpreted in the same units.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (y_t - \hat{y}_t)^2} \tag{4.3.3}$$

Sometimes it is useful to compute the forecasting errors in percentages. *Mean absolute percentage error (MAPE)* measures the average percentage deviations in absolute values. So its value shows the magnitude of deviation in percentage and is always positive

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{|y_t|} \tag{4.3.4}$$

Note that the value of MAPE is not defined if $y_t = 0$ for any of $t = 1, \dots, n$.

In order to determine whether a forecasting method is biased, showing consistently low or high values, *mean percentage error (MPE)* is used. It is computed by taking the deviation between observed and forecast value in each period and dividing by actual value for that period.

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{y_t - \hat{y}_t}{y_t} \tag{4.3.5}$$

If the resulting value is close to zero, then the model is unbiased. Large positive value implies that the model is consistently underestimating and large negative value implies that the model is consistently overestimating.

Choice of the forecast error estimators depends on the situations described above. But the decision whether the selected forecasting model is producing reasonably accurate results depends on the judgment of values of selected error estimators. The following sections examine forecasting methods.

4.4. Simple Moving Average

Moving average method is based on smoothing historical data. The objective is to use past observations to forecast future values of the time series. The moving average method is appropriate to be used for forecasting when factors affecting the time series have stabilized and the environment in which the time series data is generated is generally unchanging. A *simple moving average (SMA)* method averages all past observations in the time series to obtain the forecast value for the next period

$$\hat{y}_{t+1} = \frac{1}{t}\sum_{i=1}^{t} y_i \tag{4.4.1}$$

where t is the number of observations and y_i is the actual observed value of the time series corresponding to time period i.

Whenever a new observation becomes available, the forecast for the next period is

$$\hat{y}_{t+2} = \frac{1}{t+1}\sum_{i=1}^{t+1} y_i \tag{4.4.2}$$

So, the forecast value evolves with the appearance of new observation in then time series data. The equation (4.4.1) uses all past observations starting from the first data point to the last available one. However, when dealing with many time series simultaneously, the data storage may be an issue. The following formula provides a way to compute the forecast value based only on the most recent observed data point and the most recent forecast

$$\hat{y}_{t+2} = \frac{t\hat{y}_{t+1} + y_{t+1}}{t+1} \tag{4.4.3}$$

As long as we keep storing the most recent data, there is no longer a need to average all past observations to forecast the next period value.

Example 4.4

ABC Transit Petroleum Inc. imports and distributes gasoline. John Meyer, a sales manager is responsible for weekly reporting and forecasting the gasoline sales (measured in thousands of liters). In order to forecast the sales for the next week, he observes sales data from the past month

	A	B	C	D	E
1	t	y_t			
2	1	177			
3	2	201			
4	3	305			
5	4	155			
6	5	381			
7	6	137			
8	7	122			
9	8	365			
10	9	122			
11	10	395			
12	11	152			
13	12	394			
14	13	456			
15	14	163			
16	15	221			
17	16	394			
18					
19	$\hat{y}_{15}$	251.7857	<--"=AVERAGE(B2:B15)"		
20	e_{15}	-30.7857	<--"=B16-B19"		
21	$\hat{y}_{16}$	249.7333	<--"=(A15*B19+B16)/A16"		
22	e_{16}	144.2667	<--"=B17-B21"		

Figure 4.4.1

Based on the observations on 14 weeks in the past, John forecasts the sales for the next (15th) month to be $\hat{y}_{15} = 251\,786$ liters. When the new data $y_{15} = 221\,000$ becomes available at the end of 15th week, John computes the residual between the observed and forecasted value to be -30 786 liters. So the company sold 30 786 less litters than predicted. In order to forecast the sales for the 16th week, John proceeds with (4.4.2) using only the most recent observed value y_{15} and the most recent forecast value $\hat{y}_{15}$. The forecast value for 16th week is 249 733 which is 144 267 litters less than 394 000 that was actually sold.

	A	B	C	D
1	t	y_t	$\hat{y}_t$	e_t
2	1	177		
3	2	201	177	24
4	3	305	189	116
5	4	155	227.6667	-72.6667
6	5	381	209.5	171.5
7	6	137	243.8	-106.8
8	7	122	226	-104
9	8	365	211.1429	153.8571
10	9	122	230.375	-108.375
11	10	395	218.3333	176.6667
12	11	152	236	-84
13	12	394	228.3636	165.6364
14	13	456	242.1667	213.8333
15	14	163	258.6154	-95.6154
16	15	221	251.7857	-30.7857
17	16	394	249.7333	144.2667
18				
19	$\hat{y}_{15}$	251.7857	<--"=AVERAGE(B2:B15)"	
20	e_{15}	-30.7857	<--"=B16-B19"	
21	$\hat{y}_{16}$	249.7333	<--"=(A15*B19+B16)/A16"	
22	e_{16}	144.2667	<--"=B17-B21"	

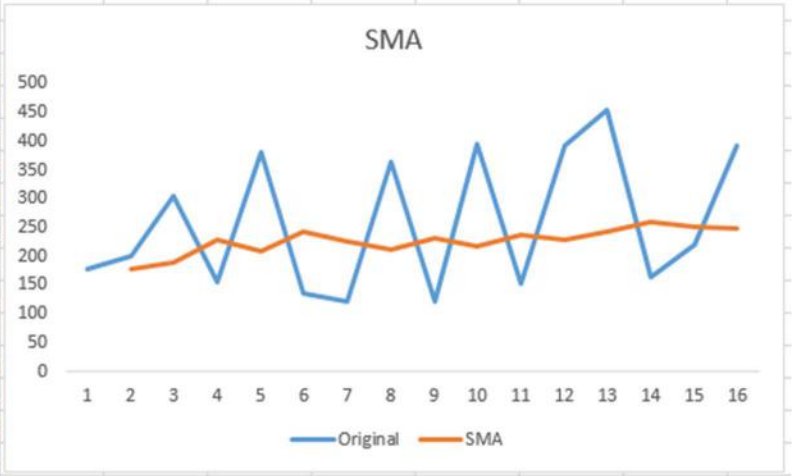


Figure 4.4.2

Computations of the forecast values for the *simple moving average* are illustrated in the Figure 4.4.2. Note that computations in the column C begin from the cell C3 which contains the formula "=AVERAGE($B$2:B2)". The starting cell of the array in the argument of the AVERAGE function is frozen. The results are the predicted values for the next time period. e.g. the row corresponding to $t = 5$ implies that the company sold 381 000 litters of gasoline in the 5th week while the forecast value for this time period would have been 209 500 litters which was found by averaging the previous 4 observations. In column D, the residuals are computed for each time period.

4.5. Moving Average

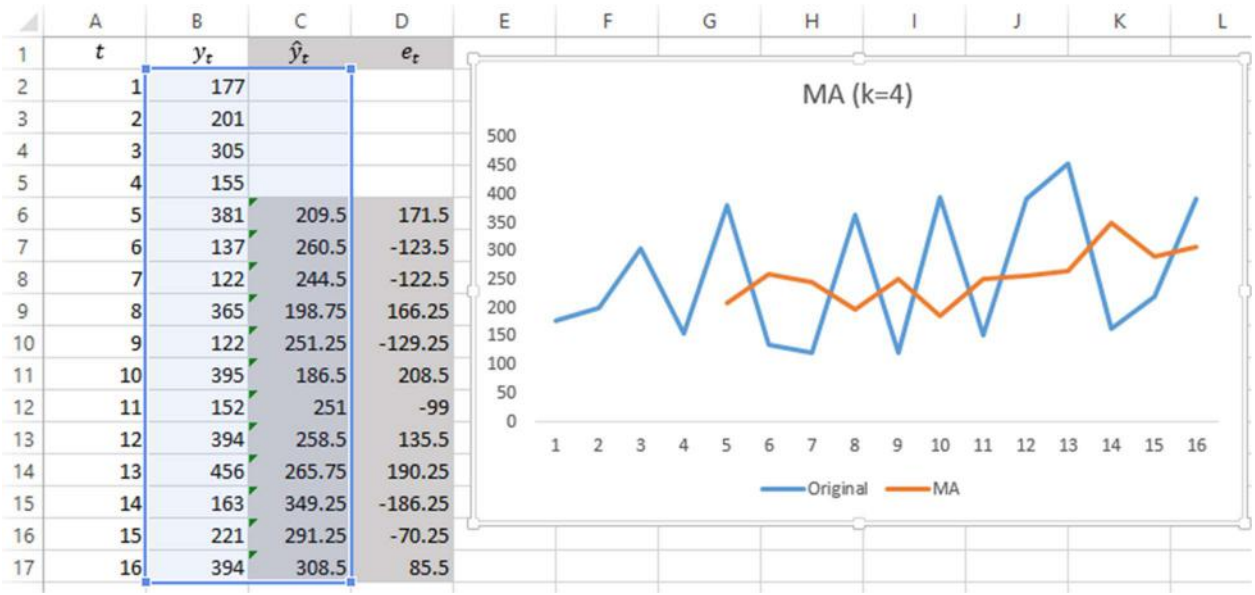
Simple Moving Average computes the average value of all past observations. The assumption here is that all the observed values have equal weight in computation of the forecast value. What if the most recent observations are more relevant than all the available data points? The term *moving average (MA)* is the generalized version of *SMA* which enables us to average the only the most recent observations

$$\hat{y}_{t+1} = \frac{y_t + y_{t-1} + \cdots + y_{t-k+1}}{k} = \frac{1}{k}\sum_{i=t-k+1}^{t} y_i \tag{4.5.1}$$

where k is the number of terms in the moving average. So, only the most recent consecutive k data points are averaged in order to forecast the value of the next period.

Example 4.5

John Meyer from Example 4.4 forecasts the values of the time series for the time periods 15 and 16 based on (4.5.1) now using $k = 4$. The following figure illustrates the computations of forecasts for the time periods from 5 to 16.



	A	B	C	D
1	t	y_t	$\hat{y}_t$	e_t
2	1	177		
3	2	201		
4	3	305		
5	4	155		
6	5	381	209.5	171.5
7	6	137	260.5	-123.5
8	7	122	244.5	-122.5
9	8	365	198.75	166.25
10	9	122	251.25	-129.25
11	10	395	186.5	208.5
12	11	152	251	-99
13	12	394	258.5	135.5
14	13	456	265.75	190.25
15	14	163	349.25	-186.25
16	15	221	291.25	-70.25
17	16	394	308.5	85.5

Figure 4.5.1

The formula in the cell C6 is "=AVERAGE(B2:B5)". Note that the cell references in the AVERAGE function are not frozen. The column D computes the residuals with a simple formula subtracting $\hat{y}_i$ from y_i in each cell.

John Meyer has two methods in his hands now – the simple moving average and moving average. He can base his choice of which method to use on one of the error estimators from section 4.3. He selects the model with smaller MSE. The following figure illustrates the computation of MSE

	A	B	C	D	E	F	G	H	I	J
1	t	y_t	$\hat{y}_t$	e_t	e_t^2		MSE	21564.86	<--"=AVERAGE(E6:E17)"	
2	1	177								
3	2	201								
4	3	305								
5	4	155								
6	5	381	209.5	171.5	29412.25					
7	6	137	260.5	-123.5	15252.25					
8	7	122	244.5	-122.5	15006.25					
9	8	365	198.75	166.25	27639.06					
10	9	122	251.25	-129.25	16705.56					
11	10	395	186.5	208.5	43472.25					
12	11	152	251	-99	9801					
13	12	394	258.5	135.5	18360.25					
14	13	456	265.75	190.25	36195.06					
15	14	163	349.25	-186.25	34689.06					
16	15	221	291.25	-70.25	4935.063					
17	16	394	308.5	85.5	7310.25					

Figure 4.5.2

The column E values are squares of corresponding values in the column D. According to (4.3.2), the average of these values is the *mean squared error* computed in H1. Similar computation of MSE for the *simple moving average* would yield MSE = 16 579.62.

	A	B	C	D	E	F	G	H	I	J
1	t	y_t	$\hat{y}_t$	e_t	e_t^2		MSE	16579.62	<--"=AVERAGE(E3:E17)"	
2	1	177								
3	2	201	177	24	576					
4	3	305	189	116	13456					
5	4	155	227.6667	-72.6667	5280.444					
6	5	381	209.5	171.5	29412.25					
7	6	137	243.8	-106.8	11406.24					
8	7	122	226	-104	10816					
9	8	365	211.1429	153.8571	23672.02					
10	9	122	230.375	-108.375	11745.14					
11	10	395	218.3333	176.6667	31211.11					
12	11	152	236	-84	7056					
13	12	394	228.3636	165.6364	27435.4					
14	13	456	242.1667	213.8333	45724.69					
15	14	163	258.6154	-95.6154	9142.302					
16	15	221	251.7857	-30.7857	947.7602					
17	16	394	249.7333	144.2667	20812.87					

Figure 4.5.3

Note that as long as the *simple moving average* computation includes more data, it has more components in the computation of residuals (D2, D3) and thus, its MSE is expected to be higher. However, increasing the number of observations offsets this effect and comparison by MSE is gives reasonable result.

4.6. Double Moving Average

One way of forecasting time series with a linear trend is a *double moving average (DMA)* method. It has a double smoothing effect of the original data. The predicted value by the DMA method is based on two sets of average values. First, average is computed from k number of the original data. Next this set is averaged once again. The resulting forecast value is a function of both, the average from the first set and the average from the second set.

The procedure of building a DMA method is as follows. First, equation (4.5.1) is used to compute the moving average of k order

$$M_t = \frac{y_t + y_{t-1} + \cdots + y_{t-k+1}}{k} = \frac{1}{k} \sum_{i=t-k+1}^{t} y_i \tag{4.6.1}$$

next, the following equation is used to compute the second moving average

$$M_t' = \frac{M_t + M_{t-1} + \cdots + M_{t-k+1}}{k} = \frac{1}{k}\sum_{i=t-k+1}^{t} M_i \tag{4.6.2}$$

Considering the forecast is to be made p periods ahead, the ultimate equation used to make the forecast is

$$\hat{y}_{t+p} = a_t + b_t p \tag{4.6.3}$$

where

$$a_t = M_t + (M_t - M_t') = 2M_t - M_t' \tag{4.6.4}$$

and

$$b_t = \frac{2}{k-1}(M_t - M_t') \tag{4.6.5}$$

Example 4.6

Ben Fischer is a local video game store manager. His job is to prepare financial reports and make forecasts for weekly rentals. He has observations for 15 weeks and plans to forecast the video game rentals for the next week (16th). The following figure illustrates the original time series and two moving averages, first by (4.6.1) and second by (4.6.2).

	A	B	C	D
1	t	y_t	M_t	M_t'
2	1			
3	2	436		
4	3	443	439.5	
5	4	448	442.3333	
6	5	451	447.3333	443.0556
7	6	447	448.6667	446.1111
8	7	468	455.3333	450.4444
9	8	468	461	455
10	9	476	470.6667	462.3333
11	10	483	475.6667	469.1111
12	11	477	478.6667	475
13	12	488	482.6667	479
14	13	485	483.3333	481.5556
15	14	486	486.3333	484.1111
16	15	503	491.3333	487
17	16			

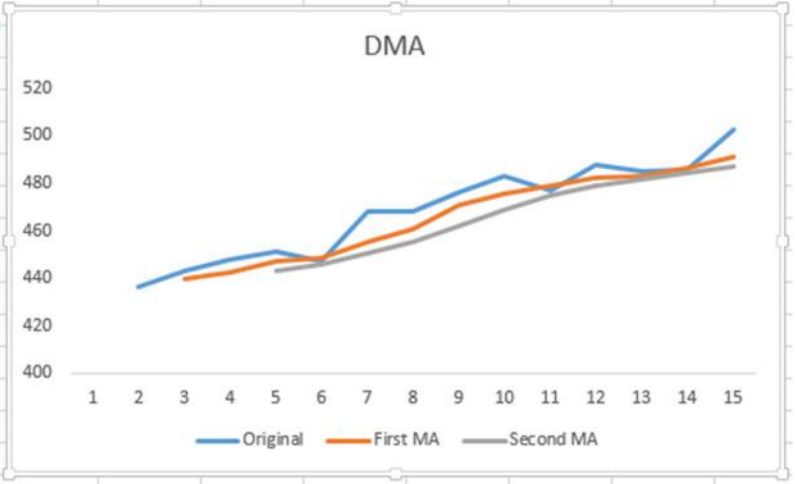


Figure 4.6.1

The column B contains the original time series. The values of M_t in the C column is computed according to (4.6.1) by "=AVERAGE(B2:B4)" in the cell C4. Similarly, the M_t' values in the D column correspond to equation (4.6.2) by "=AVERAGE(C4:C6)" in D6.

In order to make forecast for 16th week, Ben needs to use (4.6.3)-(4.6.4).

	A	B	C	D	E	F	G	H	I
1	t	y_t	M_t	M_t'	a_t	b_t	$\hat{y}_t = a_t + b_t p\ (p=1)$	e_t^2	
2	1								
3	2	436							
4	3	443	439.5						
5	4	448	442.3333						
6	5	451	447.3333	443.0556	451.6111	4.277778			
7	6	447	448.6667	446.1111	451.2222	2.555556	455.8888889	79.01235	
8	7	468	455.3333	450.4444	460.2222	4.888889	453.7777778	202.2716	
9	8	468	461	455	467	6	465.1111111	8.345679	
10	9	476	470.6667	462.3333	479	8.333333	473	9	
11	10	483	475.6667	469.1111	482.2222	6.555556	487.3333333	18.77778	
12	11	477	478.6667	475	482.3333	3.666667	488.7777778	138.716	
13	12	488	482.6667	479	486.3333	3.666667	486	4	
14	13	485	483.3333	481.5556	485.1111	1.777778	490	25	
15	14	486	486.3333	484.1111	488.5556	2.222222	486.8888889	0.790123	
16	15	503	491.3333	487	495.6667	4.333333	490.7777778	149.3827	
17	16						500		
18									
19									
20						MSE	63.52962963	<--"=AVERAGE(H7:H16)"	

Figure 4.6.2

In Figure 4.6.2, a_t values are computed in column E by "=2*C6-D6" in E6. b_t values are computed in column F by "=C6-D6" in F6. The forecast values in column G are computed for $p = 1$. Ben forecasts the rentals for the week 16 to be 500 units of video games. In the last column, residual values squared are computed whose average is the MSE computed in G20.

4.7. Simple Exponential Smoothing

Simple Exponential Smoothing is a moving average, exponentially weighted for all previously observed values. The objective of the model is to first estimate the current level and then use this estimated level to forecast future values of the time series. The simple exponential smoothing model is often appropriate for modelling data without predictable upward or downward trend. The model is based on averaging past values of a given time series in an exponentially decreasing order. The idea behind it is that the most recent observation receives the largest weight α $(0 < \alpha < 1)$ and is assumed to be the most significant determinant of the forecast. The next most recent observation receives less weight $\alpha(1-\alpha)$ which is less than α.

The observation one time period older receives the weight $\alpha(1-\alpha)^2$ which is even less and so forth.

The forecast for time $t+1$ is the weighted sum of the most recent observation y_t and the forecast for time t, which is $\hat{y}_t$. The most recent value y_t is assigned a weight α and the weight for the forecast value $\hat{y}_t$ is $(1-\alpha)$. So, the ultimate forecast equation is

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t \tag{4.7.1}$$

which can be rewritten as

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t = \hat{y}_t + \alpha(y_t - \hat{y}_t) \tag{4.7.2}$$

In this form, the forecast value for the next period of time is the predicted value for the previous period plus the weighted residual. According to (4.7.2), $\hat{y}_t$ turns out to be

$$\hat{y}_t = \alpha\hat{y}_{t-1} + (1-\alpha)\hat{y}_{t-1}$$

Substituting this equation in (4.7.2) yields

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t = \alpha y_t + (1-\alpha)[\alpha y_{t-1} + (1-\alpha)\hat{y}_{t-1}] \tag{4.7.3}$$

$$\hat{y}_{t+1} = \alpha y_t + \alpha(1-\alpha)y_{t-1} + (1-\alpha)^2\hat{y}_{t-1}$$

Continued substitution of $\hat{y}_t, \hat{y}_{t-1}, \hat{y}_{t-2}$, ... results in

$$\hat{y}_{t+1} = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \alpha(1-\alpha)^3 y_{t-3} + \cdots \tag{4.7.4}$$

So the predicted value $\hat{y}_{t+1}$ is an exponentially smoothed value. Past observations become less and less relevant as more weights are given to more recent values. The speed with which the "relevance" of past observations in forecasting the next value decreases, depends on the weight α. The value of α can be optimized by minimizing one of the error estimator from Section 4.3.

Example 4.7

A local cable TV provider was established in 2014. The number of additional subscriptions sold is a time series data observed every quarter from 2015. The Figure 4.7.1 below shows the quarterly observations for the number of additional subscribers from 2015 till 2020. In total there are 21 observations in column C. Assuming $\alpha = 0.1$ (which will later be optimized), the forecast values are computed in column D according to (4.7.2). In D1 we have the same value as in C1. The reason for this is that by convention $y_1 = \hat{y}_1$ is to be taken. All the rest of the computations of $\hat{y}_t$ values depend on the previously observed and forecast values based on

(4.7.2). So, the cell D2 contains the formula "=$I$1*C2+(1-$I$1)*D2" extended throughout the column below. The column E computes residuals for each period of time and the column F is the squares of the values in E.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Year	t	y_t	$\hat{y}_t$	e_t	e_t^2		α	0.1					
2	2015	1	520	520	0	0		MSE	23558.86					
3		2	370	520	-150	22500								
4		3	240	505	-265	70225								
5		4	390	478.5	-88.5	7832.25								
6	2016	5	590	469.65	120.35	14484.12								
7		6	450	481.685	-31.685	1003.939								
8		7	310	478.5165	-168.517	28397.81								
9		8	470	461.6649	8.33515	69.47473								
10	2017	9	650	462.4984	187.5016	35156.86								
11		10	540	481.2485	58.75147	3451.735								
12		11	370	487.1237	-117.124	13717.96								
13		12	580	475.4113	104.5887	10938.79								
14	2018	13	720	485.8702	234.1298	54816.77								
15		14	610	509.2832	100.7168	10143.88								
16		15	440	519.3548	-79.3548	6297.191								
17		16	650	511.4194	138.5806	19204.59								
18	2019	17	790	525.2774	264.7226	70078.04								
19		18	680	551.7497	128.2503	16448.14								
20		19	510	564.5747	-54.5747	2978.399								
21		20	720	559.1172	160.8828	25883.26								
22	2020	21	860	575.2055	284.7945	81107.9								
23		22		603.685										

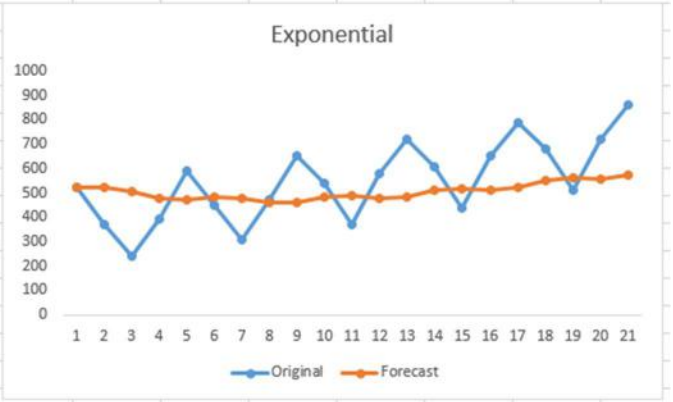


Figure 4.7.1

The very last value in the cell D23, which is $\hat{y}_{22} = 604$ new subscribers, is the forecast for the time period 22 (second quarter of 2020), meaning that the number of subscribers will increase by 458 at $t = 22$. Note that $\alpha = 0.1$ was assumed above. This alpha produces MSE = 23 558.86 in I2 by averaging the values in the column F. What if there can be found another value of α which produces lower MSE? This would mean that the exponential model with lower α better fits the observed sample of time series. The best possible value of α can be found by minimizing MSE. For this purpose, Solver package from Data tab is used as illustrated below

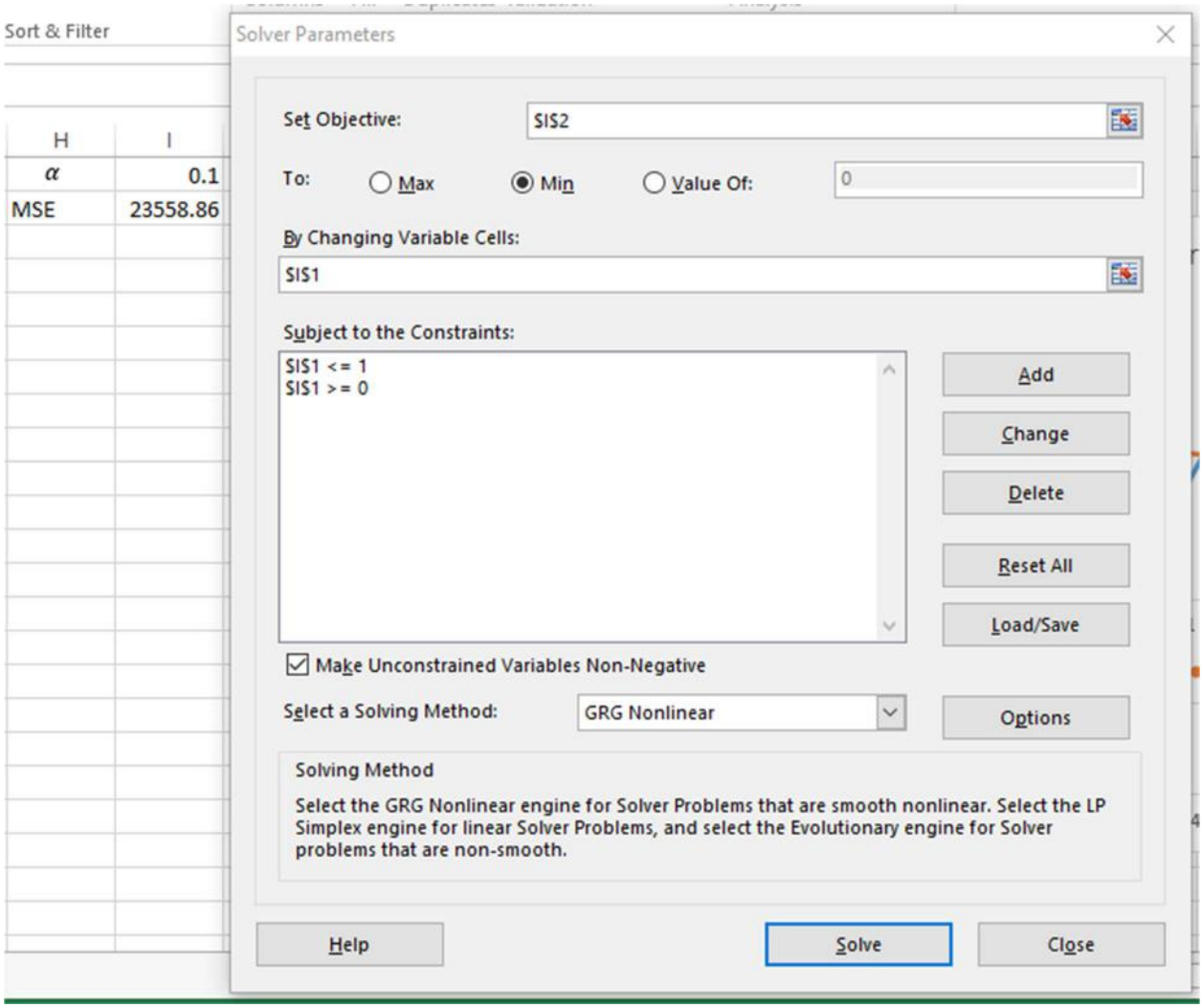


Figure 4.7.2

In the Set Objective text box, the cell containing the value to be optimized (minimized) is selected. That is I2 with the MSE value. In By Changing Variable Cells text box, the variable to be optimized is selected. For our model this is α in the cell I1. Next, since we have $0 < \alpha < 1$, the value of α is subject to constraints

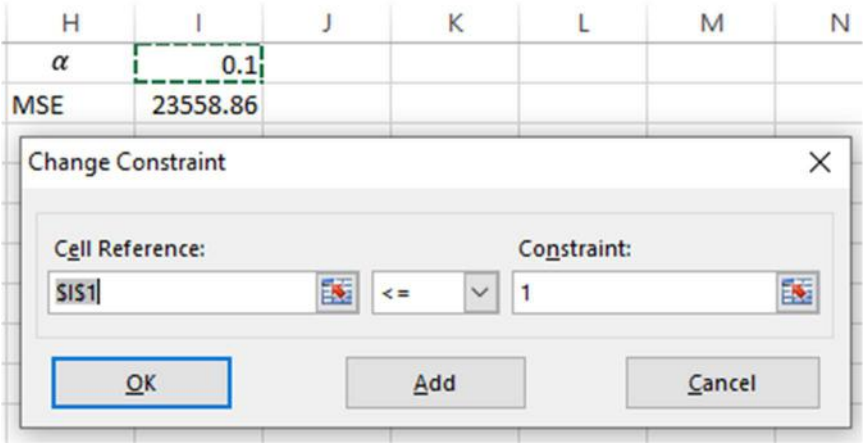


Figure 4.7.3

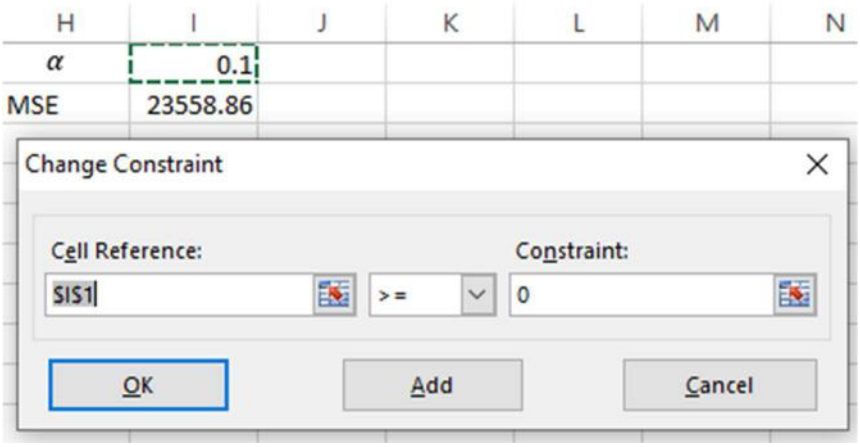


Figure 4.7.4

Once both constraints are in place, clicking the Solve button in Figure 4.7.2 yields different value of α minimizing MSE and therefore, changing the forecast values in column D. The result is shown in Figure 4.7.5. The forecast value for the time period 22 now is 708 which is obtained by $\alpha = 0.3036$ producing the minimum MSE of 20 604.15. No other value of α can produce lower MSE.

	A	B	C	D	E	F	G	H	I
1	Year	t	y_t	$\hat{y}_t$	e_t	e_t^2		α	0.303589
2	2015	1	520	520	0	0		MSE	20604.15
3		2	370	520	-150	22500			
4		3	240	474.4617	-234.462	54972.27			
5		4	390	403.2817	-13.2817	176.4037			
6	2016	5	590	399.2495	190.7505	36385.74			
7		6	450	457.1593	-7.15925	51.25491			
8		7	310	454.9858	-144.986	21020.88			
9		8	470	410.9697	59.03029	3484.575			
10	2017	9	650	428.8906	221.1094	48889.34			
11		10	540	496.017	43.983	1934.505			
12		11	370	509.3697	-139.37	19423.93			
13		12	580	467.0586	112.9414	12755.75			
14	2018	13	720	501.3464	218.6536	47809.4			
15		14	610	567.7272	42.2728	1786.99			
16		15	440	580.5607	-140.561	19757.32			
17		16	650	537.8881	112.1119	12569.09			
18	2019	17	790	571.924	218.076	47557.14			
19		18	680	638.1295	41.87054	1753.142			
20		19	510	650.8409	-140.841	19836.16			
21		20	720	608.0832	111.9168	12525.38			
22	2020	21	860	642.0599	217.9401	47497.9			
23		22		708.2241					

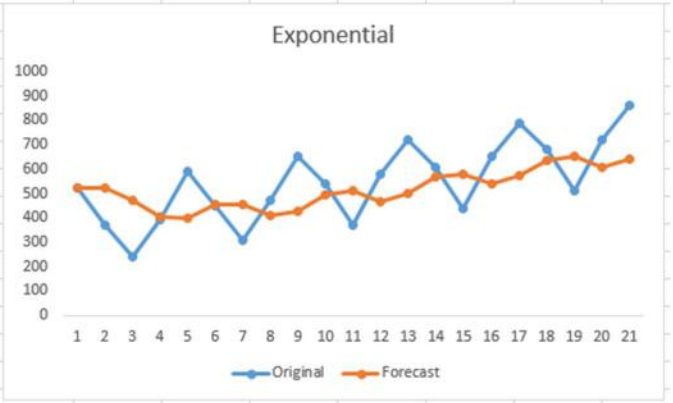


Figure 4.7.5

As a conclusion, the cable TV provider expects the number of subscribers to increase by 610 in the second quarter of 2020.

4.8. Holt's Exponential Smoothing Model Adjusted for Trend

In simple exponential smoothing, the level around which the time series fluctuates, is assumed to be changing over time and an estimate of the current level is required. What if in addition to estimating the current level, the observed time series data is trending? Then the necessity of anticipating upward or downward movements arises. So, in addition to the level estimate, the linear function estimating the trend is required. Holt's exponential smoothing method adjusted for trend intends to estimate linear trend in a time series and can be used to generate more accurate results for trending data.

The trend estimation on the other hand, requires estimate of the current slope and the current level. Holt's model weights the level and slope by different weights for each. These estimates themselves evolve over time as new observations show up and can be regarded as time series.

In particular, the level estimate in the Holt's model for a given period of time is

$$L_t = \alpha y_t + (1-\alpha)(L_{t-1} + T_{t-1}) \qquad (4.8.1)$$

where L_{t-1} and T_{t-1} are the previous estimates of the level and trend respectively and α is the smoothing constant satisfying $0 < \alpha < 1$. The trend estimate is given by

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \tag{4.8.2}$$

where $0 < \beta < 1$. Ultimately, the forecast for p periods ahead into the future is given by

$$\hat{y}_{t+p} = L_t + pT_t \tag{4.8.3}$$

So the forecast value at time t, p periods into the future is the linear function of p given the level and trend estimates fixed for the time period t.

Example 4.8

The objective of this example is to solve the forecasting problem given in *Example 4.7* using the Holt's method. Figure 4.8 shows the computations with $\alpha = 0.3$ and $\beta = 1$. The forecasts are computed for $p = 1$. Current levels at each time t are computed by the formula "=$K$1*C3+(1-$K$1)*(D2+E2)" in D2 and the trends are computed by "=$K$2*(D3-D2)+(1-$K$2)*E2" in E2. Note that by convention $L_1 = y_1$ and $T_1 = 0$ are taken. The forecast for the second quarter of 2020 is additional 749. MSE, which is the average of the sum of squared residuals is computed in the cell K4 and equals 20 437.63.

	A	B	C	D	E	F	G	H	I	J	K
1	Year	t	y_t	L_t	T_t	$\hat{y}_t$	e_t	e_t^2		α	0.3
2	2015	1	520	520	0	520	0	0		β	0.1
3		2	370	475	-4.5	520	-150	22500		p	1
4		3	240	401.35	-11.415	470.5	-230.5	53130.25		MSE	20437.63
5		4	390	389.9545	-11.4131	389.935	0.065	0.004225			
6	2016	5	590	441.979	-5.06929	378.5415	211.4586	44714.72			
7		6	450	440.8368	-4.67659	436.9097	13.09028	171.3554			
8		7	310	398.3122	-8.46139	436.1602	-126.16	15916.4			
9		8	470	413.8955	-6.05691	389.8508	80.14924	6423.9			
10	2017	9	650	480.487	1.207927	407.8386	242.1614	58642.13			
11		10	540	499.1865	2.957078	481.695	58.30504	3399.478			
12		11	370	462.5005	-1.00723	502.1436	-132.144	17461.92			
13		12	580	497.0453	2.547974	461.4933	118.5067	14043.85			
14	2018	13	720	565.7153	9.160176	499.5933	220.4067	48579.13			
15		14	610	585.4128	10.21391	574.8755	35.12455	1233.734			
16		15	440	548.9387	5.545111	595.6267	-155.627	24219.68			
17		16	650	583.1387	8.410596	554.4838	95.51618	9123.34			
18	2019	17	790	651.0845	14.36412	591.5493	198.4507	39382.69			
19		18	680	669.814	14.80066	665.4486	14.55139	211.743			
20		19	510	632.2303	9.562219	684.6147	-174.615	30490.29			
21		20	720	665.2547	11.90844	641.7925	78.2075	6116.413			
22	2020	21	860	732.0142	17.39355	677.1632	182.8368	33429.3			
23						749.4078					

Figure 4.8.1

On the other hand we should optimize the model by minimizing MSE as shown in the following figure.

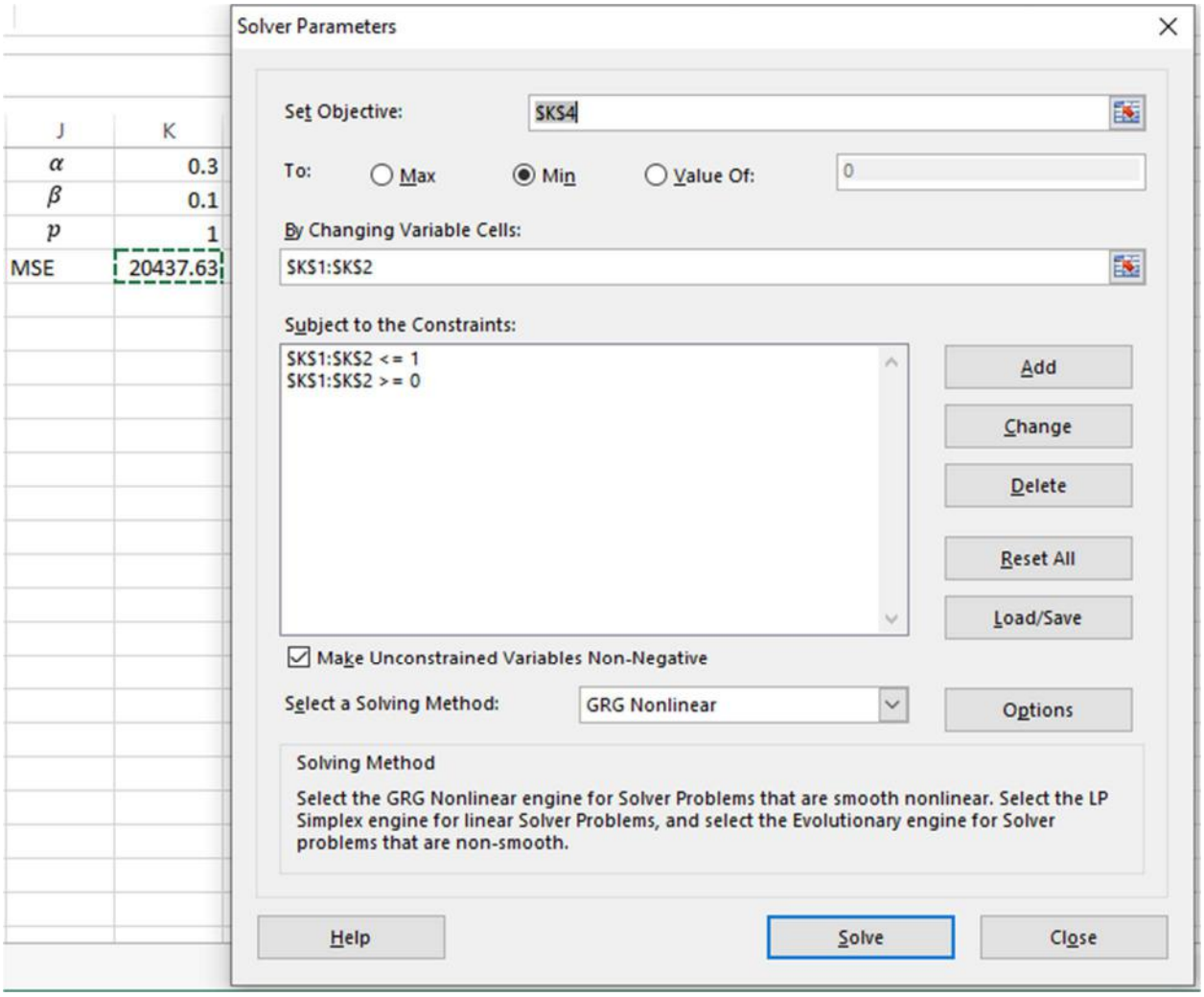


Figure 4.8.2

The constraints for the values of α and β were added as follows

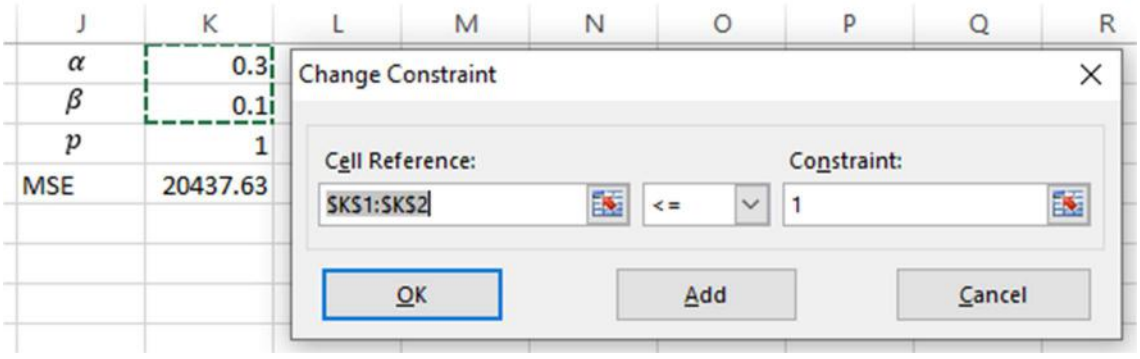


Figure 4.8.3

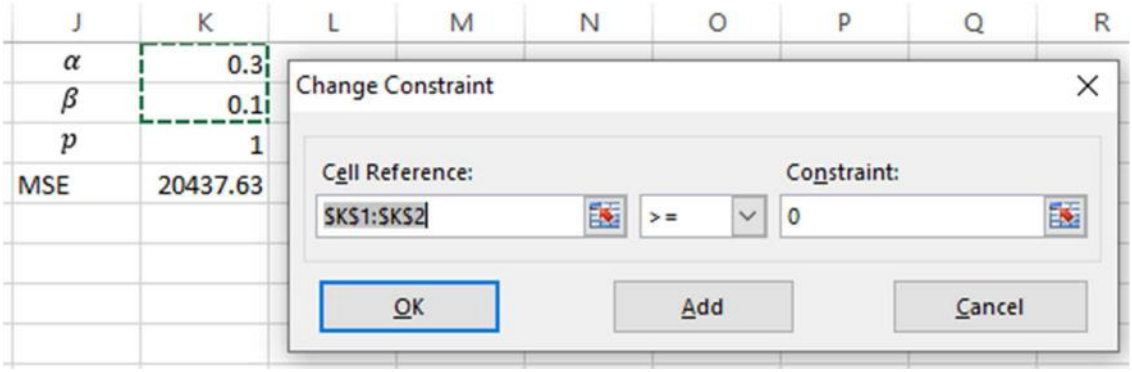


Figure 4.8.4

Solving the MSE minimization problem in Figure 4.8.2 gives the following results

	A	B	C	D	E	F	G	H	I	J	K
1	Year	t	y_t	L_t	T_t	$\hat{y}_t$	e_t	e_t^2		α	0.22323
2	2015	1	520	520	0	520	0	0		β	0.227133
3		2	370	486.5155	-7.60545	520	-150	22500		p	1
4		3	240	425.5781	-19.7189	478.91	-238.91	57077.99		MSE	20180.66
5		4	390	402.3189	-20.523	405.8592	-15.8592	251.5127			
6	2016	5	590	428.2734	-9.96644	381.7959	208.2041	43348.95			
7		6	450	425.3818	-8.35951	418.3069	31.69309	1004.452			
8		7	310	393.1317	-13.7859	417.0223	-107.022	11453.76			
9		8	470	399.5826	-9.18942	379.3458	90.6542	8218.185			
10	2017	9	650	448.3453	3.973435	390.3931	259.6069	67395.72			
11		10	540	471.8918	8.419142	452.3187	87.68131	7688.013			
12		11	370	455.6862	2.826044	480.311	-110.311	12168.51			
13		12	580	485.632	8.98584	458.5123	121.4877	14759.27			
14	2018	13	720	544.93	20.4134	494.6178	225.3822	50797.12			
15		14	610	575.3121	22.67762	565.3434	44.65665	1994.216			
16		15	440	562.7216	14.66707	597.9897	-157.99	24960.74			
17		16	650	593.5977	18.34868	577.3887	72.61132	5272.404			
18	2019	17	790	651.6934	27.37654	611.9464	178.0536	31703.08			
19		18	680	679.2775	27.4237	679.0699	0.930109	0.865103			
20		19	510	662.7915	17.45035	706.7012	-196.701	38691.37			
21		20	720	689.1171	19.46621	680.2419	39.7581	1580.706			
22	2020	21	860	742.3841	27.14349	708.5833	151.4167	22927.01			
23						769.5276					

Figure 4.8.5

So we obtain a more accurate model with $\alpha = 0.2232$ and $\beta = 0.2271$ producing minimum possible MSE which is 20 180.66 which is smaller than MSE computed by $\alpha = 0.3$ and $\beta = 0.1$. Based on this formula, taking the trend into consideration, the expected value for the new subscribers in the period 22 is 770. At this point, the local cable TV provider has a more accurate forecast compared to the forecast made by the simple exponential smoothing method.

Figure 4.8.6 shows the original time series data and the forecasts based on the optimized α and β on the same chart:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Year	t	y_t	L_t	T_t	$\hat{y}_t$	e_t	e_t^2		α	0.22323	
2	2015	1	520	520	0	520	0	0		β	0.227133	
3		2	370	486.5155	-7.60545	520	-150	22500		p	1	
4		3	240	425.5781	-19.7189	478.91	-238.91	57077.99		MSE	20180.66	
5		4	390	402.3189	-20.523	405.8592	-15.8592	251.5127				
6	2016	5	590	428.2734	-							
7		6	450	425.3818	-							
8		7	310	393.1317	-							
9		8	470	399.5826	-							
10	2017	9	650	448.3453	3							
11		10	540	471.8918	8							
12		11	370	455.6862	2							
13		12	580	485.632								
14	2018	13	720	544.93								
15		14	610	575.3121	2							
16		15	440	562.7216	1							
17		16	650	593.5977	1							
18	2019	17	790	651.6934	2							
19		18	680	679.2775								
20		19	510	662.7915	17.45035	706.7012	-196.701	38691.37				
21		20	720	689.1171	19.46621	680.2419	39.7581	1580.706				
22	2020	21	860	742.3841	27.14349	708.5833	151.4167	22927.01				
23						769.5276						

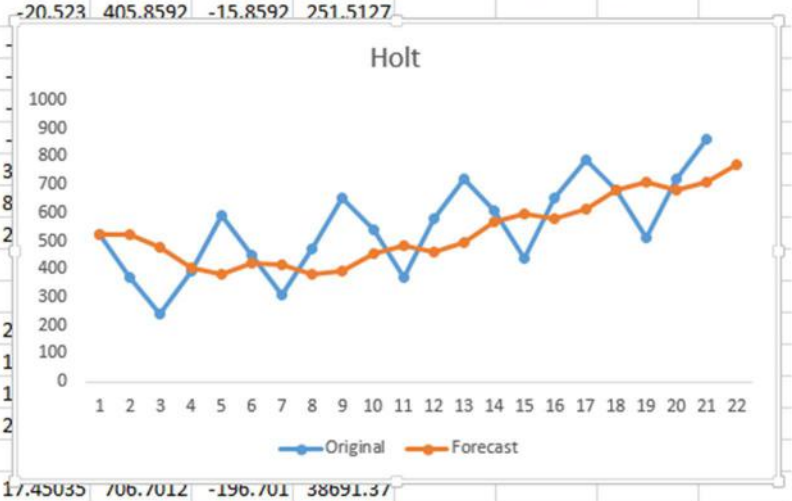


Figure 4.8.6

4.9. Holt-Winters' Exponential Smoothing Model Adjusted for Trend and Seasonal Variation

In Holt's model, the simple exponential smoothing method was extended by taking trend into account. If we observe the data in *Examples 4.8* and *4.7*, we notice that the observations for the first quarter each year are consistently higher than the observations of the third quarter. So, the seasonal pattern emerges that needs to be addressed in order to obtain a better model, taking one more component into consideration and therefore, producing more accurate results. To address this problem, a seasonal index is added to the existing Holt's model and we obtain Holt-Winters's Exponential Smoothing model estimating current level, trend and seasonal index for each period of time as follows

The level estimate is given by

$$L_t = \alpha \frac{y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + T_{t-1}) \tag{4.9.1}$$

where S_{t-s} is the seasonality estimate defined by (4.9.3) below. The trend estimate is

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \tag{4.9.2}$$

and the seasonality estimate defined as

$$S_t = \gamma\frac{y_t}{L_t} + (1-\gamma)S_{t-s} \tag{4.9.3}$$

where γ is the smoothing constant for the seasonality estimate satisfying $0 < \gamma < 1$. s in the index of the seasonality coefficient is the constant, time periods in which the seasonality pattern persists. Ultimately, the forecast for p periods into the future is

$$\hat{y}_{t+p} = (L_t + pT_t)S_{t-s+p} \tag{4.9.4}$$

Example 4.9

By adding the seasonality component to the model, the *Example 4.8* is further extended. The following figure demonstrates computations for the Hold-Winter's method

	A	B	C	D	E	F	G	H	I	J	K	L
1	Year	t	y_t	L_t	T_t	S_t	$\hat{y}_t$	e_t	e_t^2		α	0.4
2	2015	1	520	500	0	1		520	270400		β	0.1
3		2	370	500	0	1		370	136900		γ	0.3
4		3	240	500	0	1		240	57600		p	1
5		4	390	500	0	1		390	152100		MSE	54017.56
6	2016	5	590	536	3.6	1.030224		590	348100			
7		6	450	503.76	0.016	0.967985	539.6	-89.6	8028.16			
8		7	310	426.2656	-7.73504	0.918174	503.776	-193.776	37549.14			
9		8	470	439.1183	-5.67626	1.021098	418.5306	51.46944	2649.103			
10	2017	9	650	512.4376	2.223288	1.101691	446.5424	203.4576	41395.01			
11		10	540	531.9405	3.951254	0.982135	498.1839	41.81613	1748.589			
12		11	370	482.7246	-1.36547	0.872666	492.0418	-122.042	14894.2			
13		12	580	516.0219	2.100811	1.051964	491.5148	88.48518	7829.628			
14	2018	13	720	572.29	7.517538	1.148615	570.811	149.189	22257.34			
15		14	610	596.3229	9.169082	0.994375	569.449	40.55096	1644.381			
16		15	440	564.9759	5.117473	0.844505	528.3926	-88.3926	7813.248			
17		16	650	589.2129	7.029421	1.067324	599.7175	50.28251	2528.331			
18	2019	17	790	632.8594	10.69113	1.178521	684.8527	105.1473	11055.96			
19		18	680	659.669	12.30298	1.005308	639.9305	40.06946	1605.562			
20		19	510	644.7449	9.580266	0.828457	567.4836	-57.4836	3304.359			
21		20	720	662.4287	10.39062	1.0732	698.3772	21.62277	467.544			
22	2020	21	860	695.5828	12.66698	1.195877	792.9317	67.06833	4498.16			
23		22					712.0095					

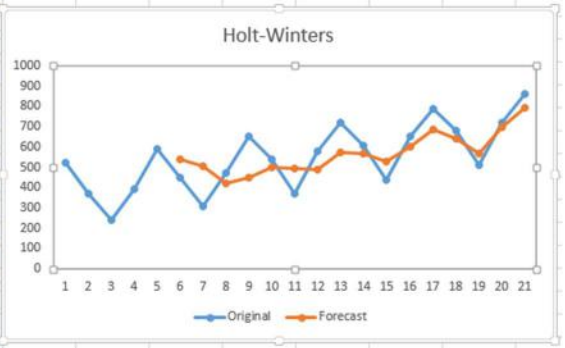


Figure 4.9.1

The initializing values of the level, trend and seasonality are by convention taken as $L_1 = y_1, T_1 = 0$ and $S_1 = 1$. The index of the seasonality coefficient in (4.9.3) is $s = 4$ in this example since we concluded that the seasonality pattern keeps repeating every quarter. So, the

computations begin from $t = 5$. The column D computes the levels according to (4.9.1) by "=$L$1*C6/F2+(1-$L$1)*(D5+E5)" in D6. The trend component values are computed in column E according to (4.9.2) by "=$L$2*(D6-D5)+(1-$L$2)*E5". The seasonality component is computed in column F based on (4.9.3) by "=$L$3*C6/D6+(1-$L$3)*F2" in F6. The forecasts taking all these three components into consideration are computed in the column G based on (4.9.4) by "=(D6+E6*$L$4)*F3". The highlighted value, 512 is the forecast value for 22nd month. The computations described above were based on $\alpha = 0.4, \beta = 0.1$ and $\gamma = 0.3$. The *mean squared error*, computed in L4 is 54 017.56. The result is $\hat{y}_{22} = 712$ new subscribers for the month 22. On the other hand, optimization gives the different and more accurate results. The following figure demonstrates the optimization by Solver

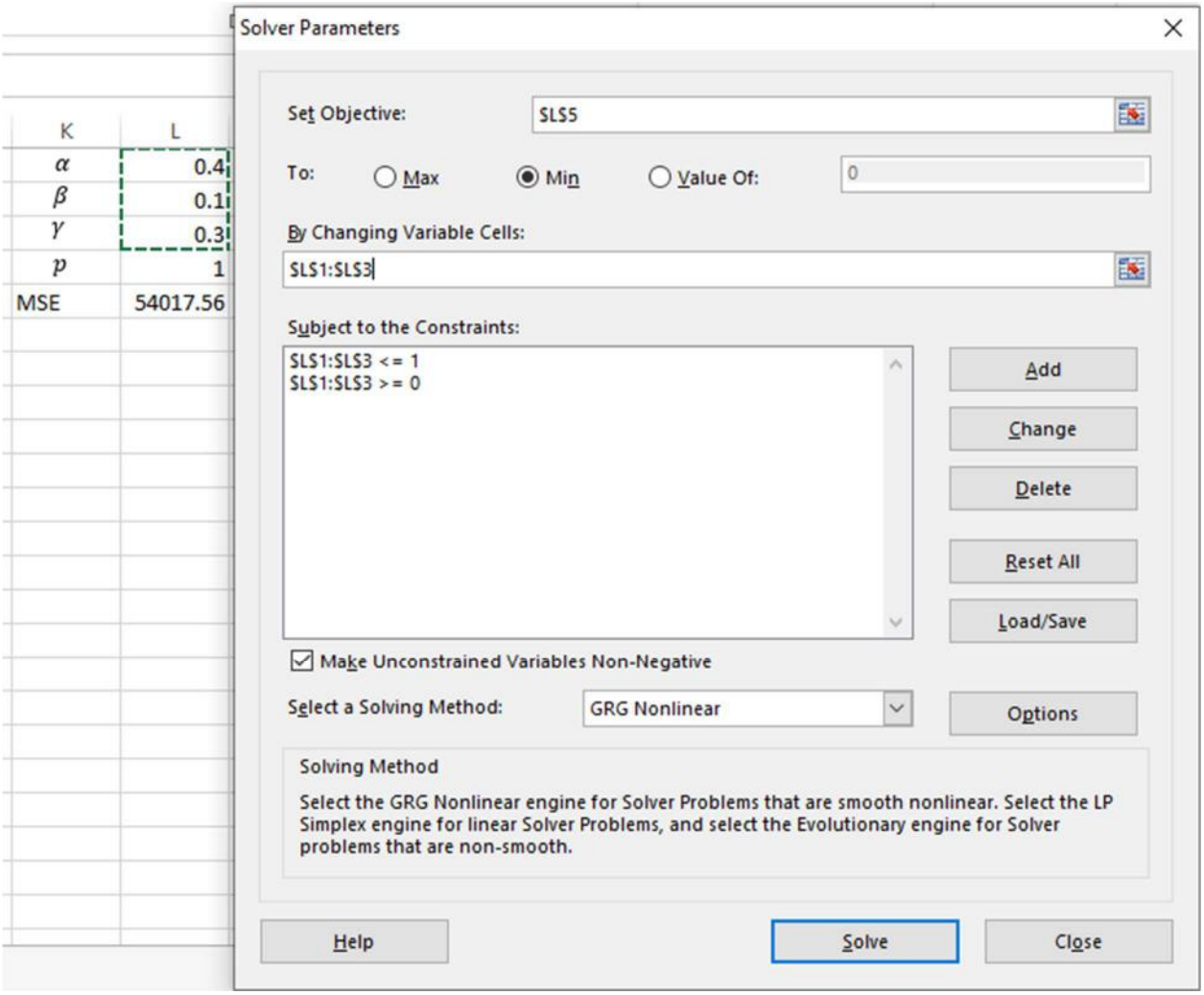


Figure 4.9.2

The constraints now involve all 3 constants

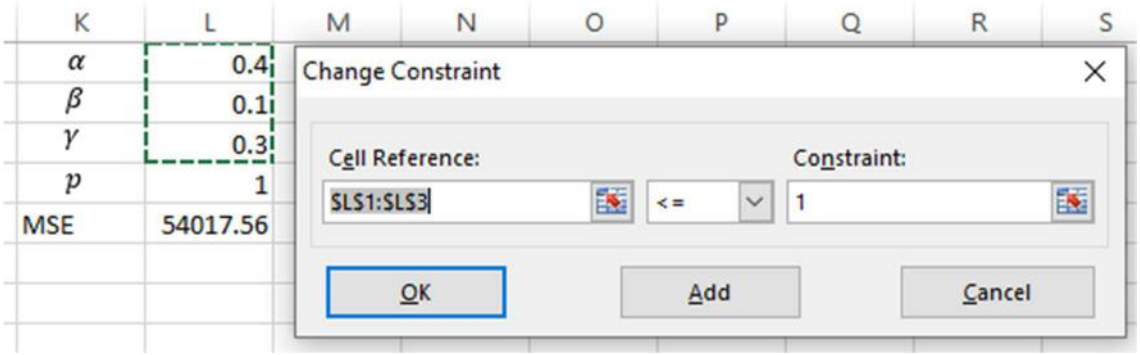


Figure 4.9.3

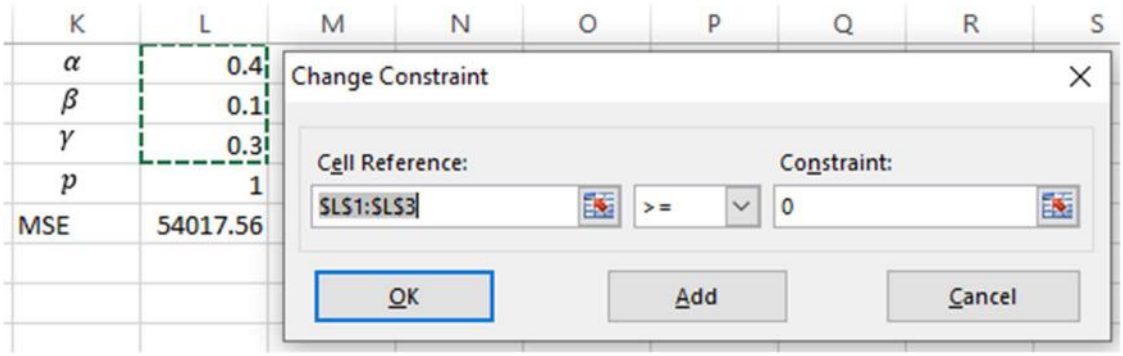


Figure 4.9.4

The result of optimization is given below

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Year	t	y_t	L_t	T_t	S_t	$\hat{y}_t$	e_t	e_t^2		α	0.101873	
2	2015	1	520	500	0	1		520	270400		β	0.34643	
3		2	370	500	0	1		370	136900		γ	1	
4		3	240	500	0	1		240	57600		p	1	
5		4	390	500	0	1		390	152100		MSE	49790.27	
6	2016	5	590	509.1686	3.17628	1.158752		590	348100				
7		6	450	505.9936	0.976004	0.889329	512.2449	62.2449	3896.885				
8		7	310	486.9036	-5.97545	0.636							
9		8	470	479.8149	-6.36112	0.979							
10	2017	9	650	482.3672	-3.27326	1.347							
11		10	540	492.1437	1.247587	1.09							
12		11	370	502.3309	4.344522	0.736							
13		12	580	515.3791	7.359751	1.125							
14	2018	13	720	523.9181	7.768274	1.374							
15		14	610	534.1572	8.624237	1.141							
16		15	440	548.3422	10.55065	0.802							
17		16	650	560.7966	11.21016	1.159							
18	2019	17	790	572.2969	11.31067	1.380							
19		18	680	584.8144	11.72875	1.162							
20		19	510	600.5198	13.10639	0.849							
21		20	720	614.3967	13.37333	1.171							
22	2020	21	860	627.2848	13.20522	1.370988	866.5753	-6.57533	43.23496				
23		22					744.7375						

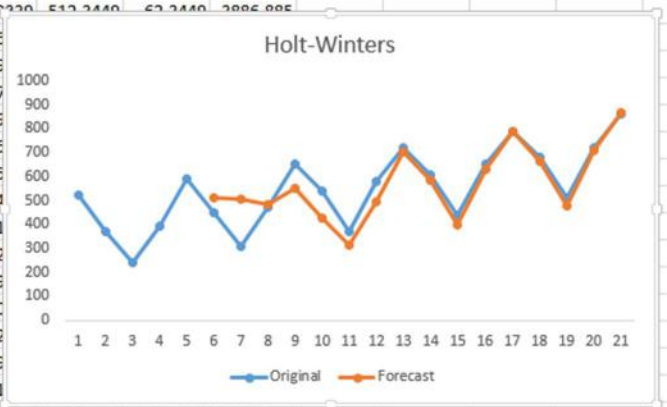


Figure 4.9.10

which sets $\alpha = 0.1019, \beta = 0.3464$ and $\gamma = 1$. The constant of seasonality was obtained to be 1. We see an explicit seasonal pattern in the time series. As a result, MSE = 49 790.27 which is lowest among all models described above. As a conclusion, the cable TV provider has a more accurate forecasts based on the Holt-Winters' method which predicts $\hat{y}_{22} = 745$.

Chapter 5. Inventory Management

5.1. Introduction

Inventory is defined as a stock of items maintained by an organization to meet demands of customers. Every type of organization keeps some form of inventory. Most frequently, demand is uncertain and can be considered as random. An organization is concerned to maintain a safe level of inventory to meet uncertain demand. Stocks of inventories may be built up to meet the demands of cyclical or seasonal nature. Similarly, a company may purchase a large amount of inventory to take advantage of price discounts and meet the surge in anticipated demand. On the other hand, maintaining an excessive amount of inventory may result in unreasonable cost.

There are several types of cost associated with inventory management. The most basic cost is the *cost of carrying.* This is simply the cost of holding items in storage. Carrying cost depends on the amount of inventory being stored and the length of storage time.

Another type of cost is the *ordering cost*. Ordering cost is the cost associated with replenishing the stock of inventory being held. Ordering cost usually reacts inversely to carrying costs. When the volume of inventory ordered is high, fewer orders are required, so the ordering cost reduces. However, as long as the volume to be maintained increases, the carrying cost follows accordingly.

The last type of cost covered in this chapter is the *shortage cost.* Shortage cost is associated with a loss caused from inability to meeting a customer demand. Part of shortage cost may not be measurable in dollar amounts (like loss of goodwill and reputation resulting in the loss of customers). In this chapter, we consider shortage costs measured in dollar amounts. Shortage costs may occur when the storage cost is unaffordable. So the shortage cost acts inversely to the carrying costs. As the amount of inventory maintained in storage increases, so does the carrying costs while the shortage costs decrease.

The goal of this chapter is to find an optimal balance (in the sense of total cost minimization) between the amount of inventory to order and carry, number of orderings and timing of ordering.

5.2. Basic Economic Order Quantity Model (EOQ)

The most basic form of the inventory model is the *basic economic order quantity model*. The objective in this model is to obtain an optimal order size that minimizes total cost. The total cost consists of the carrying cost and the ordering cost. The time span between ordering time and replenishment time is called the *lead time.*

There are various assumptions in this model

- The model does not take shortages into account
- Lead time is assumed to be constant for every order
- Demand is not random, it is assumed to be constant over time
- Quantity ordered is received all at once

These assumptions are shown in the following figure

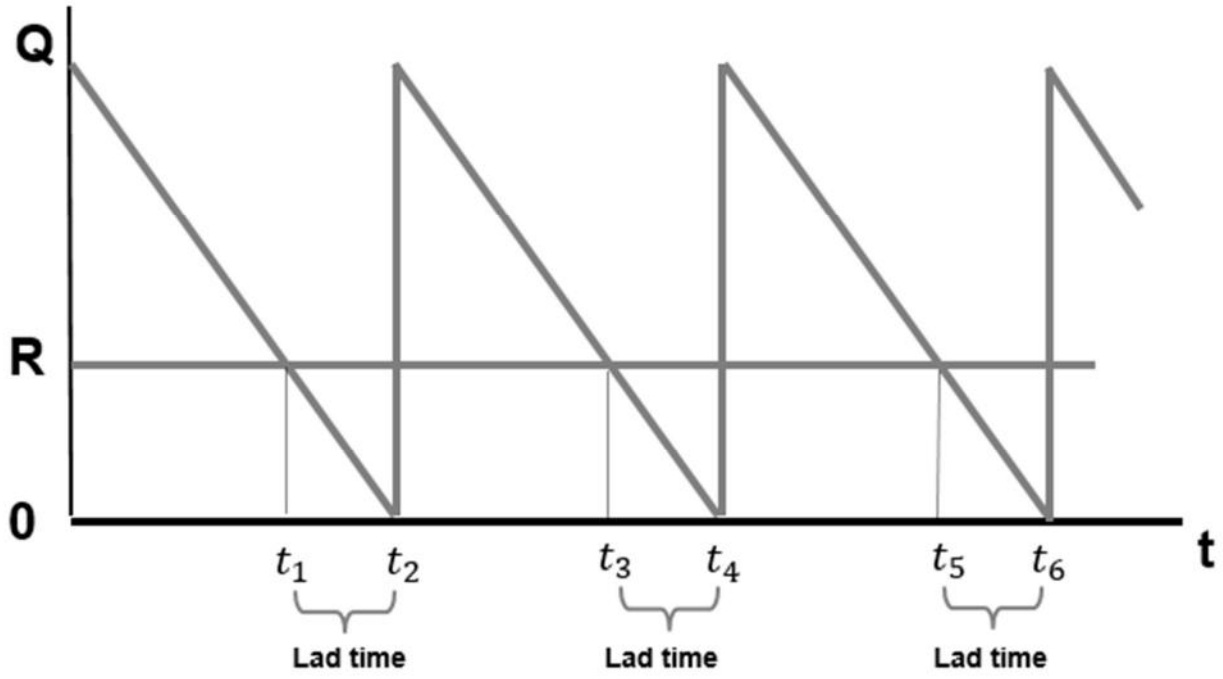


Figure 5.2.1

In the figure, Q denotes the order size which is constant and same for all orders. R is the level at which a new order is placed. This occurs at times $t_1, t_3, t_5, \ldots$ At times $t_2, t_4, t_6, \ldots$ the quantities reach zero and the orders are received immediately all at once. Lead times (or delivery times) illustrated in the figure are constant and equal. Since demands are assumed to

be constant every time, the quantity depletes at a constant rate. The average inventory on an annual basis maintained is equal to

$$average\ inventory = \frac{Q}{2} \tag{5.2.1}$$

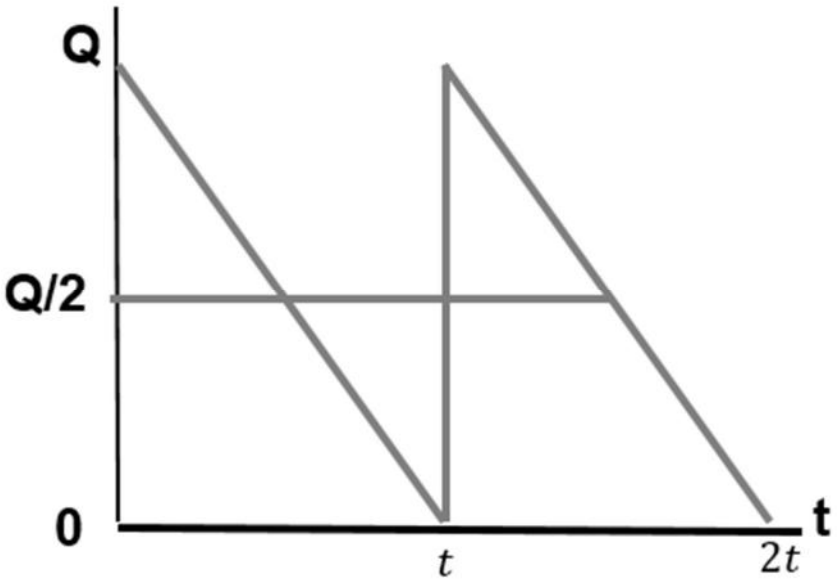


Figure 5.2.2

Since (5.2.1) holds, meaning the available inventory on an annual basis is $Q\backslash 2$, the annual carrying cost is given by

$$annual\ carrying\ cost = C_c \frac{Q}{2} \tag{5.2.2}$$

where C_c is the unit carrying cost per year. On the other hand, another type of cost in this model is the ordering cost. Since the demand was assumed to be known and constant, the number of orders per year is D/Q. As long as the cost per order, C_o is known, the annual cost of ordering is

$$annual\ ordering\ cost = C_o \frac{D}{Q} \tag{5.2.3}$$

Since in this simplest setting we assumed absence of the shortage cost (when quantity depletes to zero, replenishment occurs immediately, setting the inventory size back to Q), the total inventory cost consists of these two costs (5.2.2) and (5.2.3) only

$$TC = C_c \frac{Q}{2} + C_o \frac{D}{Q} \tag{5.2.4}$$

All quantities in (5.2.4) are constants and known except Q. Total cost as a function of an independent variable Q is illustrated in the following figure

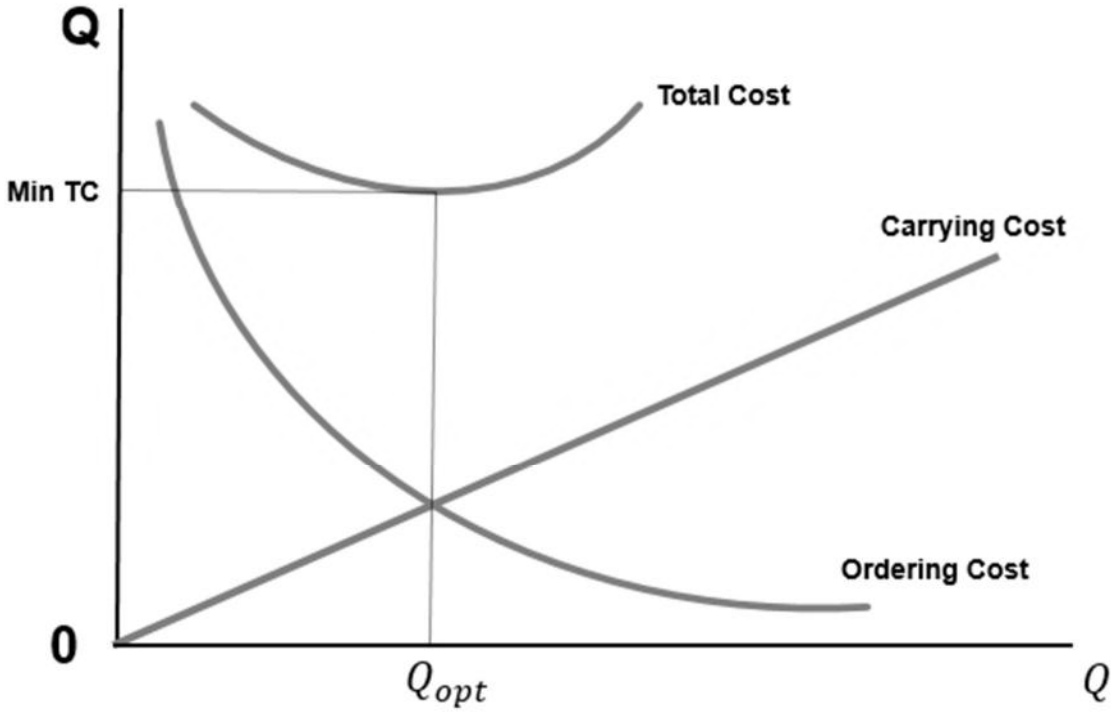


Figure 5.2.3

The objective is to minimize the total cost function. Q_{opt} in Figure 5.2.3 is the total cost minimizing quantity. The total cost is minimized for the value of Q where the carrying cost coincides with the ordering cost. Total cost minimization problem is solved by differentiating the function with respect to Q, equating the derivative function to zero and solving for Q. The solution is Q_{opt}. It has a geometric meaning shown in the figure above. In particular, at $Q = Q_{opt}$, the tangent line of the total cost function is horizontal. This is the point where minimum value is achieved. So, differentiating (5.2.4) and equating to zero yields

$$\frac{dTC}{dQ} = -\frac{C_0 D}{Q^2} + \frac{C_c}{2} = 0$$

$$Q_{opt} = \sqrt{\frac{2C_0 D}{C_c}} \tag{5.2.5}$$

As a result

$$TC_{min} = C_0 \frac{D}{Q_{opt}} + C_0 \frac{Q_{opt}}{2}$$

Example 5.2

Fast Vehicles Inc. is the largest retailer of tires for sport cars. The company has several large retail stores each getting supplies from the same warehouse. Inventory is kept in the central warehouse and distributed to the retail stores as demanded on a daily basis. The company has the carrying unit cost of $0.7 per tire and an ordering cost is $140. Demand is constant each ear and estimated to be 8 000 tires per year. The company would like to know the optimal order sizes minimizing the total cost. In addition, the company is interested in the number of orders it will need to make annually and the time between orders.

The following figure demonstrates the computations

	A	B	C	D	E
1	C_c	0.7			
2	C_o	140			
3	D	8000			
4	Q_{opt}	1788.854	<--"=SQRT(2*B2*B3/B1)"		
5	TC_{min}	1252.198	<--"=B2*B3/B4+B1*B4/2"		
6	# of orders	4.472136	<--"=B3/B4"		
7	cycle time	81.61648	<--"=365/B6"		

Figure 5.2.4

The quantities in B1, B2 and B3 are given. B4 computes the optimal quantity according to (5.2.5) that minimizes the total inventory cost computed in B5 by (5.2.4). So, 1789 tires must be ordered to minimize the total cost to $1 252.2. In total, 4 orders will be made per year and time between consecutive orders is 82 days. Note that the assumption in the above example is that there are 365 working days per year. If we only consider working days excluding the weekends and holidays, the remaining number of days would be divided by the number of orders per year in order to obtain the time between orders (which is called the cycle time).

5.3. EOQ Model with Non-Instantaneous Receipt

The EUQ model with non-instantaneous receipt assumes that the orders are not received all at once whenever placed. So the replenishment continues over some time which is constant. The rate at which the order is received over time is known as the production rate. The rate at which inventory is demanded is also assumed to be constant. However, the assumption that

the shortage is not possible persists here as well. Thus the production rate exceeds the demand rate. Note that in the model the ordering cost remains the same. The fact that the stocks are being replenished over time does not affect the ordering cost. However, the carrying cost changes since the average inventory level (5.2.1) is different. In addition, the maximum inventory level is no longer Q, but a quantity lower than Q. The reason for this is that the order quantity is depleted during the order receipt period. The following figure demonstrates the specifics of the model

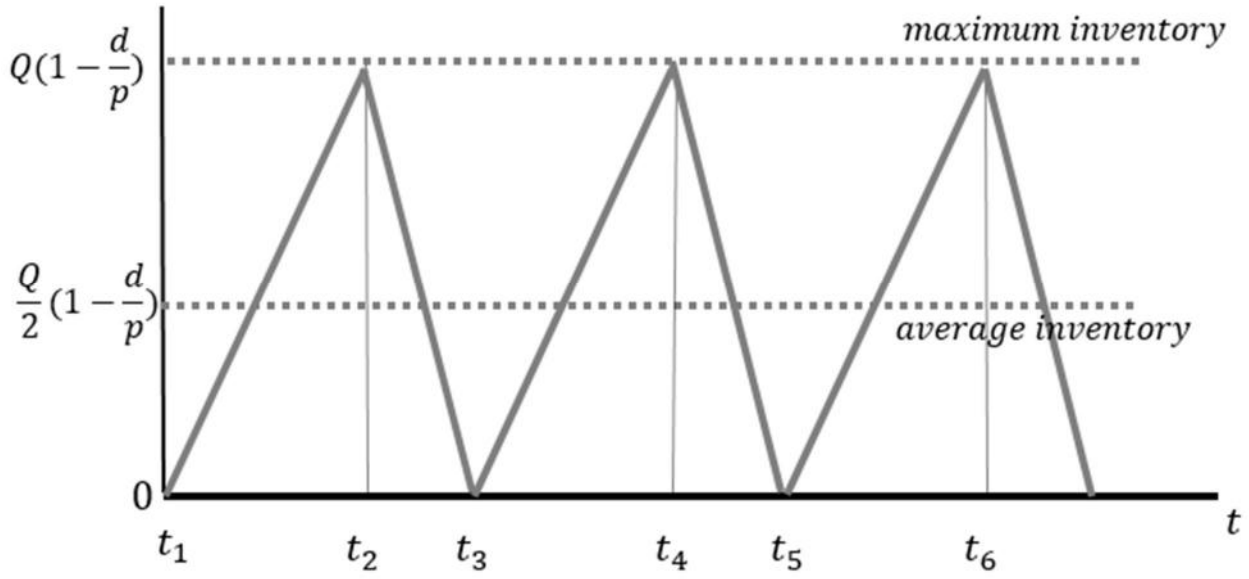


Figure 5.3.1

At times t_1, t_3, t_5, the orders are placed and the stocks start to fill at some rate. At times t_2, t_4, t_6, replenishment of stocks end. The maximum level the inventory can reach is

$$maximum\ inventory\ level = Q - \frac{Q}{p}d = Q(1-\frac{p}{d}) \tag{5.3.1}$$

Correspondingly, the average inventory is determined as

$$average\ inventory\ level = \frac{Q}{2}(1-\frac{d}{p}) \tag{5.3.2}$$

The carrying cost is affected by the level of inventory and is defined as

$$total\ carrying\ cost = C_c\frac{Q}{2}(1-\frac{d}{p}) \tag{5.3.3}$$

where C_c is the carrying cost per unit of inventory. Total annual inventory cost is the sum of the carrying cost and the ordering cost. The latter remains unchanged

$$TC = C_o\frac{D}{Q} + C_c\frac{Q}{2}(1-\frac{d}{p}) \tag{5.3.4}$$

In order to minimize the total cost by the quantity, proceeding by differentiating the function with respect to quantity variable and equating to zero yields

$$\frac{dTC}{dQ} = -C_o\frac{D}{Q^2} + \frac{C_c}{2}\left(1-\frac{d}{p}\right) = 0$$

from which

$$Q_{opt} = \sqrt{\frac{2C_oD}{C_c(1-\frac{d}{p})}} \tag{5.3.5}$$

Example 5.3

In the previous example, Fast Vehicles Inc. had an ordering cost of \$140 and a carrying cost of \$0.7 per tire annually. In addition, the annual demand was $D = 8\,000$. Since we assume 365 working days per year, it turns out that the daily demand is $d = 8000/365 = 21.92$. On the other hand, the production rate, (which must satisfy $p > d$) is $p = 160$ units per day.

Optimal value of Q and the corresponding minimum total cost is computed below

	A	B	C	D	E	F
1	C_c	0.7				
2	C_o	140				
3	D	8000				
4	p	160				
5	d	21.91781	<--"=B3/365"			
6	Q_{opt}	1994.64	<--"=SQRT(2*B2*B3/(B1*1-B5/B4))"			
7	TC_{min}	1163.995	<--"=B2*B3/B6+B1*B6/2*(1-B5/B4)"			
8	run length	12.4665	<--"=B6/B4"			
9	# of runs	4.010749	<--"=B3/B6"			
10	max Inv.	1721.402	<--"=B6*(1-B5/B4)"			

Figure 5.3.2

In Figure 5.3.2 above, Q_{opt} is computed by (5.3.5). The production run length is interpreted as time it takes to replenish the stock to Q_{opt} and is computed as follows

$$production\ run\ length = \frac{Q}{p}$$

Number of production runs is computed as

$$\#\ of\ production\ runs = \frac{D}{Q}$$

Ultimately, the maximum level of inventory is computed according to (5.3.1). As a result, the company has 4 production runs in total during a year, each lasting for 12 days receiving 1995 tires per order. Maximum inventory level that can be reached is 1721 units of tires. The total cost resulting from these numbers is $1 164.

5.4. EOQ Model with Possibility of Shortages

In the previous two models, the shortages were not allowed. At any point in time, the inventory stocked were enough to meet the external demand. In this section, we allow the possibility of shortage. When shortage occurs, the company is not able to meet the demand immediately. However, another assumption in this section is that the demand not met because of the shortage can be back ordered. Thus the customer gets the order later and all demand is eventually met. The following figure illustrates this scenario

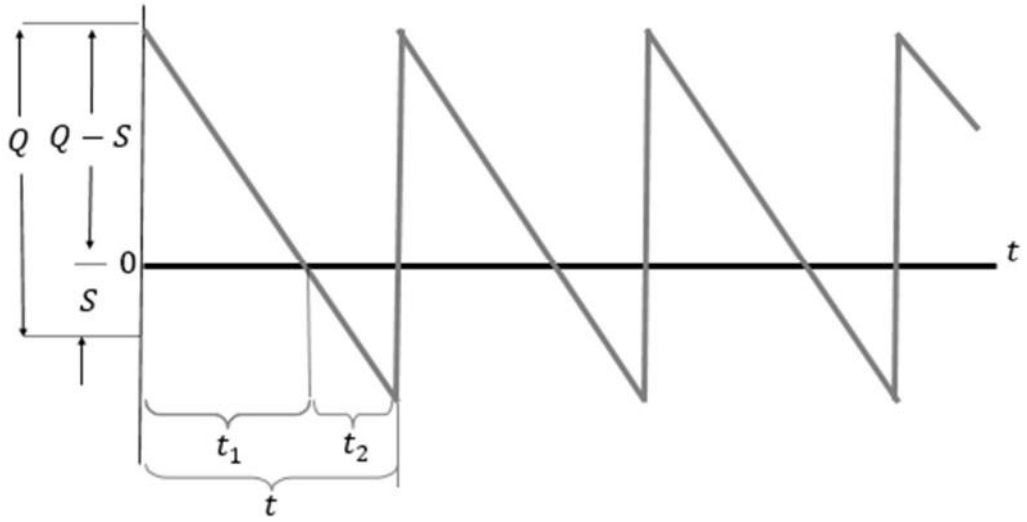


Figure 5.4.1

The maximum inventory never reaches Q because the demand during shortage must be compensated by late delivery. So, the maximum inventory level reached is $Q - S$. Q and S are inversely related. Greater the value of Q, less the shortage but the carrying cost increases accordingly. t_1 denotes the time it takes for the inventory level (starting from $Q - S$) to deplete completely. So, from the complete replenishment, it takes t_1 to start the shortage period. The

shortage period lasts for t_2 after which the stocks are replenished and set back to the maximum level $Q - S$ again and cycle starts over again. The total cost now consists of three components – the total ordering cost for which the ordering cost per unit remains unaffected, the total carrying cost which is diminished as shortage increases, and the total shortage cost which increases as Q decreases.

All of these costs are computed by the following formulas

$$total\ ordering\ cost = C_o \frac{D}{Q} \tag{5.4.1}$$

$$total\ carrying\ cost = C_c \frac{(Q-S)^2}{2Q} \tag{5.4.2}$$

$$total\ shortage\ cost = C_s \frac{S^2}{2Q} \tag{5.4.3}$$

In the last formula C_s is the unit cost of shortage.

Combining these components results in the total inventory cost function

$$TC = C_s \frac{S^2}{2Q} + C_c \frac{(Q-S)^2}{2Q} + C_o \frac{D}{Q} \tag{5.4.4}$$

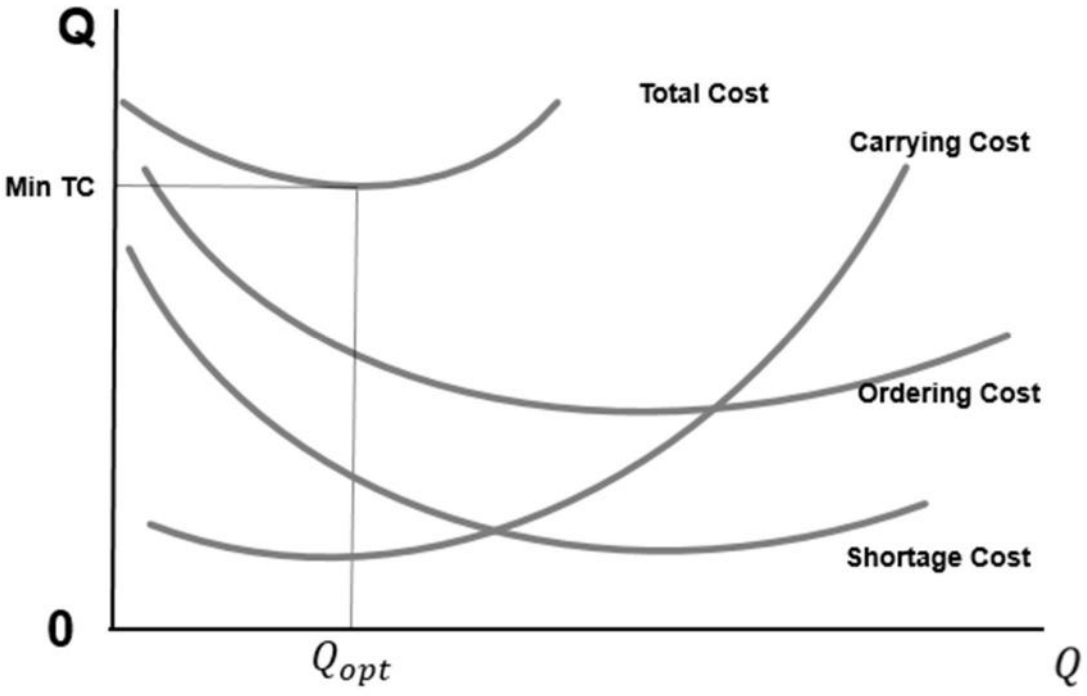


Figure 5.4.2

The slope of the total cost function is zero wherever it has a horizontal tangent line. In this model we do not only optimize Q, but also the shortage level S. Skipping the differentiation steps as shown in previous sections, it can be shown that the optimal quantity and the shortage levels are given by

$$Q_{opt} = \sqrt{\frac{2C_o D}{C_c}\left(\frac{C_s + C_c}{C_s}\right)} \tag{5.4.5}$$

$$S_{opt} = Q_{opt}\left(\frac{C_c}{C_c + C_s}\right) \tag{5.4.6}$$

The time during which inventory is on hand t_1 shown in Figure 5.4.1, is

$$t_1 = \frac{Q - S}{D} \tag{5.4.7}$$

and the time during which there is a shortage is

$$t_2 = \frac{S}{D} \tag{5.4.8}$$

Example 5.4

The examples in the previous sections are extended here by allowing the shortages. All quantities remain the same, the only additional constant here is the shortage cost per unit of tire, which is $C_s = 1.85$. In addition, note that in this model the replenishment occurs immediately. So there are no longer the need for p (production run) and d (demand per day) quantities.

	A	B	C	D	E	F	G	H
1	C_c	0.7						
2	C_o	140						
3	C_s	1.85						
4	D	8000						
5	Q_{opt}	2100.193	<--"=SQRT(2*B2*B4/B1*(B3+B1)/B3)"					
6	S_{opt}	576.5236	<--"=B5*B1/(B1+B3)"					
7	TC_{min}	1066.569	<--"=B3*B6^2/(2*B5)+B1*(B5-B6)^2/(2*B5)+B2*B4/B5"					

Figure 5.4.3

Here the optimal quantity to order each time is 2 100 units of tire and the optimal shortage is 577 tires. These quantities minimize the total cost and set it to $1 067. The total ordering,

carrying and shortage costs individually are computed according to (5.4.1), (5.4.2) and (5.4.3) shown below

	A	B	C	D	E	F	G
1	C_c	0.7					
2	C_o	140					
3	C_s	1.85					
4	D	8000					
5	Q_{opt}	2100.193	<--"=SQRT(2*B2*B4/B1*(B3+B1)/B3)"				
6	S_{opt}	576.5236	<--"=B5*B1/(B1+B3)"				
7	TC_{min}	1066.569	<--"=B3*B6^2/(2*B5)+B1*(B5-B6)^2/(2*B5)+B2*B4/B5"				
8	$C.Ordering$	533.2843	<--"=B2*B4/B5"				
9	$C.Carrying$	386.8925	<--"=B1*(B5-B6)^2/(2*B5)"				
10	$C.Shortage$	146.3918	<--"=B3*B6^2/(2*B5)"				

Figure 5.4.4

Ultimately, the times of inventory on hand and times of shortage computed in the cells B11 and B12 below.

	A	B	C	D	E	F	G
1	C_c	0.7					
2	C_o	140					
3	C_s	1.85					
4	D	8000					
5	Q_{opt}	2100.193	<--"=SQRT(2*B2*B4/B1*(B3+B1)/B3)"				
6	S_{opt}	576.5236	<--"=B5*B1/(B1+B3)"				
7	TC_{min}	1066.569	<--"=B3*B6^2/(2*B5)+B1*(B5-B6)^2/(2*B5)+B2*B4/B5"				
8	$C.Ordering$	533.2843	<--"=B2*B4/B5"				
9	$C.Carrying$	386.8925	<--"=B1*(B5-B6)^2/(2*B5)"				
10	$C.Shortage$	146.3918	<--"=B3*B6^2/(2*B5)"				
11	t_1	0.190459	<--"=(B5-B6)/B4"				
12	t_2	0.072065	<--"=B6/B4"				

Figure 5.4.5

The results are as follows. The time during which inventory is on hand is 0.19 year which is 69.5 days. And the time during which there is a shortage is 0.07 year, or 26.3 days.

5.5. Discounts on Ordered Quantity

The basic economic order quantity model examined in Section 5.1 only considers the carrying cost and the cost of order. Adding an additional component – price of purchase transforms the model by taking the possible discounts into account. Depending on the quantity ordered, there may be two scenarios – in case of a quantity discount, carrying costs may be fixed and constant or it may be computed as a percentage of the purchase price. This section covers the first case implying that the carrying costs are constant.

Purchase prices differ depending on the order size.

Quantity	Price
$q_1 \leq Q < q_2$	p_1
$q_2 \leq Q < q_3$	p_2
$q_3 \leq Q < q_4$	p_3
$q_4 \leq Q < q_5$	p_4
…	…

Table 5.5.1

$q_i < q_j, p_i > p_j, i < j$. So, the prices are set in a descending order in the table above. Greater quantity implies less purchase price. Given the values of cost of order, C_o, cost of carrying, C_c and demand, D in the basic EOQ model, the total cost minimizing optimal quantity (ignoring the price discounts) is obtained by (5.2.5) to be

$$Q_{opt} = \sqrt{\frac{2C_o D}{C_c}} \tag{5.5.1}$$

Total cost function now includes two elements from Section 5.2 – the carrying cost and the ordering cost, and an additional component, the purchase price to be paid

$$TC = \frac{C_o D}{Q} + C_c \frac{Q}{2} + PD \tag{5.5.2}$$

As long as the total cost minimizing quantity Q_{opt} is given by (5.5.1), depending on its value, the purchase price, PD depends on this value and is taken from Table 5.5.1.

Example 5.5

Fast Vehicles Inc. in Example 5.2 had the carrying cost per tire $C_c = \$0.7$, ordering cost $C_o = \$140$ and an annual demand of 8 000 tires. In addition, the following price discounts are offered by the supplier

Quantity	Price
$1 \leq Q < 500$	\$150
$500 \leq Q < 1000$	\$130
$1000 \leq Q < 5000$	\$120
$5000 \leq Q$	\$115

Table 5.5.2

e.g. the purchase price for an ordered quantity within 500 and 1000 is \$130 while the price for a quantity within the range of 1000 and 5000 is \$120.

Figure 5.5.1 shows the computations of the optimal quantity and total cost corresponding to the price discount

	A	B	C	D	E	F	G	H	I	J	K
1	C_c	0.7				Range		Price			
2	C_o	140				1	500	150			
3	D	8000				500	1000	130			
4	Q_{opt}	1788.854	<--"=SQRT(2*B2*B3/B1)"			1000	2000	120			
5	$C.Ordering$	626.099	<--"=B2*B3/B4"			2000	+	115			
6	$C.Carrying$	626.099	<--"=B1*B4/2"								
7	Purchase Price	120	<--"=IF(AND(B4>=F2,B4<G2),H2,IF(AND(B4>=F3,B4<G3),H3,IF(AND(B4>=F4,B4<G4),H4,H5)))"								
8	PD	960000	<--"=B7*B3"								
9	TC_{min}	961252.2	<--"=B5+B6+B8"								

Figure 5.5.1

The cell B4 computes Q_{opt} according to (5.5.1) to be 1 789 units. This number falls within the range of 1000 and 2000. So, the purchase price is \$120 computed in the cell B7. Total cost is computed in the column B9 which is \$961 252.2. This cost is minimum attainable in terms of carrying and order costs. However, when the purchase price is added to it, further investigation is required to check if the quantity ordered should be increased to take advantage of larger discount. If instead of purchasing 1 789 units of tires, what if 2 000 is purchased? Figure 5.5.2 shows the total cost computed for $Q = 2000$ (this is no longer Q_{opt} according to its definition from 5.5.1)

	A	B	C	D	E	F	G	H	I	J	K
1	C_c	0.7				Range		Price			
2	C_o	140				1	500	150			
3	D	8000				500	1000	130			
4	Q	2000	<--"=SQRT(2*B2*B3/B1)"			1000	2000	120			
5	$C.Ordering$	560	<--"=B2*B3/B4"			2000	+	115			
6	$C.Carrying$	700	<--"=B1*B4/2"								
7	Purchase Price	115	<--"=IF(AND(B4>=F2,B4<G2),H2,IF(AND(B4>=F3,B4<G3),H3,IF(AND(B4>=F4,B4<G4),H4,H5)))"								
8	PD	920000	<--"=B7*B3"								
9	TC	921260	<--"=B5+B6+B8"								

Figure 5.5.2

In Figure 5.5.2, the quantity is set to 2 000. Corresponding purchase price is \$115. Cost of ordering and carrying is recalculated accordingly. Most importantly, the total cost corresponding to this new purchase price now becomes \$921 260 which is less than \$961 252.2 computed by $Q_{opt} = 1\,789$. So, as long as \$921 260 < \$961 252.2, the company should take the maximum discount price and order 2 000 units of tire.

5.6. Discounts on Ordered Quantity with Constant Carrying Costs as a Percentage of Price

In addition to the previous model where the discounts were applied for increasing amount of inventory ordered, the discount can be applied to the carrying costs. This section covers the case where the annual carrying cost is a fixed percentage p (expressed in decimals) of the purchase price. Given the fixed percentage of the purchase price, Table 5.5.2 is

Quantity	Price	Carrying Cost
$q_1 \leq Q < q_2$	p_1	$p_1 p$
$q_2 \leq Q < q_3$	p_2	$p_2 p$
$q_3 \leq Q < q_4$	p_3	$p_3 p$
$q_4 \leq Q < q_5$	p_4	$p_4 p$
…	…	…

Table 5.6.1

Given the values of carrying and ordering costs, the optimal quantity without a price discount is again given by $Q_{opt} = \sqrt{\frac{2C_o D}{C_c}}$ which is the same as (5.5.1). The total cost function is also identical to (5.5.2) and is defined as $TC = \frac{C_o D}{Q} + C_c \frac{Q}{2} + PD$.The only difference is that the purchase price P and the carrying cost C_c now both depend on Q.

Example 5.6

In Example 5.5, consider that in addition to the information given, there is also a discount in carrying cost depending on ordered quantity. Assuming he percentage of price is $p = 0.1$. The following table illustrates the discounts for various ranges of quantities

Quantity	Price	Carrying Cost
$1 \leq Q < 500$	\$150	\$150(0.1)=\$15
$500 \leq Q < 1000$	\$130	\$130(0.1)=\$13
$1000 \leq Q < 5000$	\$120	\$120(0.1)=\$12
$5000 \leq Q$	\$115	\$115(0.1)=\$11.5

Table 5.6.1

The value of carrying cost without discount would be $C_c = \$0.7$. The given values of $C_o = \$140$ and $D = 8\,000$ per year result in the optimal quantity (without discount) shown in the following figure

	A	B	C	D	E	F	G	H	I	J
1	C_c	0.7				Range		Price	Carrying Cost	
2	C_o	140				1	500	150	15	
3	D	8000				500	1000	130	13	
4	Q_{opt}	1788.854	<--"=SQRT(2*B2*B3/B1)"			1000	2000	120	12	
5	$C.Ordering$	626.099	<--"=B2*B3/B4"			2000	+	115	11.5	
6	$C.Carrying$	626.099	<--"=B1*B4/2"							
7	Purchase Price	120	<--"=IF(AND(B4>=F2,B4<G2),H2,IF(AND(B4>=F3,B4<G3),H3,IF(AND(B4>=F4,B4<G4),H4,H5)))"							
8	Carrying Cost	12	<--"=IF(AND(B4>=F2,B4<G2),I2,IF(AND(B4>=F3,B4<G3),I3,IF(AND(B4>=F4,B4<G4),I4,I5)))"							
9	PD	960000	<--"=B7*B3"							
10	TC_{min}	961252.2	<--"=B5+B6+B9"							

Figure 5.6.1

Since $Q_{opt} = 1788.85$, then purchase price is \$120 and the carrying cost is \$13. Carrying costs in the column I are computed by taking 10% of the prices in the corresponding row. Ordering and carrying costs are computed as before and the total cost minimized is \$961 252.2. In order to determine which order quantity is more beneficial, we compute the total cost for 2 000 units of tire. The computations are demonstrated below

	A	B	C	D	E	F	G	H	I	J
1	C_c	0.7				Range		Price	Carrying Cost	
2	C_o	140				1	500	150	15	
3	D	8000				500	1000	130	13	
4	Q	2000	<--"=SQRT(2*B2*B3/B1)"			1000	2000	120	12	
5	$C.Ordering$	560	<--"=B2*B3/B4"			2000	+	115	11.5	
6	$C.Carrying$	700	<--"=B1*B4/2"							
7	Purchase Price	115	<--"=IF(AND(B4>=F2,B4<G2),H2,IF(AND(B4>=F3,B4<G3),H3,IF(AND(B4>=F4,B4<G4),H4,H5)))"							
8	Carrying Cost	11.5	<--"=IF(AND(B4>=F2,B4<G2),I2,IF(AND(B4>=F3,B4<G3),I3,IF(AND(B4>=F4,B4<G4),I4,I5)))"							
9	PD	920000	<--"=B7*B3"							
10	TC	921260	<--"=B5+B6+B9"							

Figure 5.6.2

In case of $Q = 2\,000$, the purchase price is set to \$115 and the carrying costs reduces to 11.5. Ultimately, the total cost is \$921 260 which is less than \$961 252.2 obtained in the context of the basic EOQ model. So, the conclusion here is that the Fast Vehicles Inc. should take advantage of the discounts and order 2 000 instead of 1789 units.

Chapter 6. Queuing Analysis

6.1. Introduction

Queues are one of the most common occurrences in everyone's daily life. Anyone who goes shopping or to a movie, frequently experiences the inconvenience of waiting in line. Queues are not only related to inconvenience for customers and companies, but might also be related to significant expenses. The expense arises from the fact that customers seeing the queue prefer to avoid it and make a choice for the competitor's service or buy an alternative product. Thus, reduction of queue is important for companies, especially with service related operations.

Queues form when customers or things arrive at a rate faster than then can be served. Most of the organizations have sufficient service capacity to handle queues in the long run.

Queuing analysis is a probabilistic form of analysis. So the managers have some operating characteristics to influence on - such as the average time a customer spends in the queuing line or the rate of arrival of new customers. There are several ways companies can influence these characteristics and speed up procedures of providing quality services.

The objective of this chapter is to cover two of the most common types of queuing systems – a single server system and a multiple server system. The simple server system is the simplest form of queuing system. It is covered in Section 6.2 and demonstrates the fundamentals of a queuing system. The multiple server system involves more complex analysis and assumes that the single waiting line is being served by several servers. As a result, decisions aimed to reduce the queues and speed up the service are discussed.

6.2. Single Server Model

The simplest form of the queuing system is the Single Server system. The assumption is that the customers arrive at a Poisson arrival rate and are served at exponential rates. Denoting the number of customer arrival rate as λ (average number of customer arrivals per time period) and the customer service rate as μ (average number of customers served per time period), the model assumes that $\lambda < \mu$ holds. If the inequality did not hold (i.e. $\lambda > \mu$), the model would result in paradoxical situation when at some point the queue gets unreasonably long and the waiting line length approaches infinity in the long run.

Customer in the queuing system is either in the waiting line or being served. Given the inequality $\lambda < \mu$, the probability that no customer is in the queuing system is given by

$$P_0 = 1 - \frac{\lambda}{\mu} \tag{6.2.1}$$

This is equivalent to probability that the server is being idle. The probability that n customers are in the queuing system is

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0 = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \tag{6.2.2}$$

The average number of customers in the queuing system is

$$L = \frac{\lambda}{\mu - \lambda} \tag{6.2.3}$$

and the average number of customers in then waiting line is

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} \tag{6.2.4}$$

The average time a customer spends in the queuing system is given by

$$W = \frac{1}{\mu - \lambda} = \frac{L}{\lambda} \tag{6.2.5}$$

and the average time a customer spends in the waiting line is

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} \tag{6.2.6}$$

The probability that the server is busy, or equivalently, the probability that the customer has to wait, known as the *utilization factor* is

$$U = \frac{\lambda}{\mu} \tag{6.2.7}$$

Note that since the event – server is busy, is mutually exclusive to the event – server is idle, the utilization factor coincides with $1 - P_0$.

Manager of the company has several mechanisms to influence the queue length. First option is to add a new employee. In this case the customers in the same queue line are served faster. In other words, the service rate λ increases. However, the company has to make additional expense of hiring an additional employee. Another option is to add a new checkout counter. Construction of the new checkout counter is significantly costly and it includes the costs of each additional cashiers. This results in splitting the queue line into two lines. It is assumed that the customers divide themselves equally between the two lines making the arrival rates half of the prior arrival rate for a single checkout counter. Comparing the total costs of both options, the manager of a company decides which option to choose. Total cost is the sum of the cost of service and the cost of waiting in the queue.

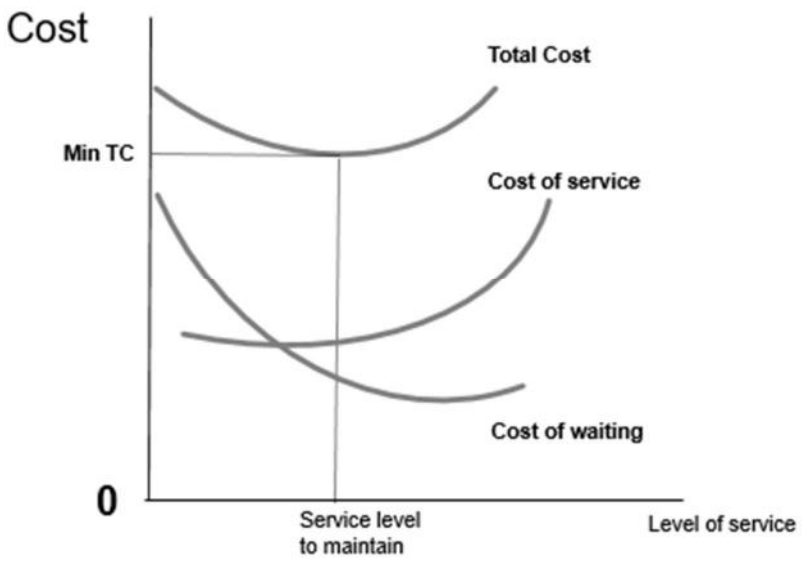


Figure 6.2.1

Example 6.2

Toy Gift store sells toys for kids. The store manager has to frequently handle queues, especially on Christmas. The customer arrival rate is computed to be $\lambda = 25$ customers per hour while the customer service rate is $\mu = 30$ per hour. By applying these values into the formulas above, the manager computes the following quantities

	A	B	C	D	E
1	λ	25			
2	μ	30			
3	P_0	0.166667	<--"=1-B1/B2"		
4	L	5	<--"=B1/(B2-B1)"		
5	L_q	4.166667	<--"=B1^2/(B2*(B2-B1))"		
6	W	0.2	<--"=1/(B2-B1)"		
7	W_q	0.166667	<--"=B1/(B2*(B2-B1))"		
8	U	0.833333	<--"=B1/B2"		

Figure 6.2.2

According to Figure 6.2.2, the probability that the server is idle is $P_0 = 0.1667$, which means that the server is serving customers by $U = 0.8333$ probability (i.e. 83% of times the server is busy). On average there are $L = 5$ customers waiting in the total queuing system (i.e. either waiting in the queuing line or being served). The average number of customers in the waiting line is $L_q = 4.17$. The average time a customer spends in the waiting line is $W = 0.2$ hours

which is 12 minutes and the average time a customer spends in the waiting line is W_q =0.1667 hours.

Considering these results, the manager now has two options. She can either add a new employee or add a new checkout counter. Let us first consider the case of adding a new employee that results in additional weekly cost.

In particular, addition of an extra employee will cost the Toy Gift store $140 per week. By analyzing statistical data, the manager concluded that by adding a new employee, for each reduced minute that customer spends in the waiting system, the store avoids a loss of $70 per week. This loss would have arisen from the situation when a customer simple walks away to avoid waiting in the queue. The effect of adding a new employee is the increase in service rate. If the previous service rate was $\mu = 30$ customers served per hour, now it is $\mu = 40$ per hour. Assuming the arrival rate remains the same $\lambda = 25$, we have the same quantities computed below

	A	B	C	D	E
1	λ	25			
2	μ	40			
3	P_0	0.375	<--"=1-B1/B2"		
4	L	1.666667	<--"=B1/(B2-B1)"		
5	L_q	1.041667	<--"=B1^2/(B2*(B2-B1))"		
6	W	0.066667	<--"=1/(B2-B1)"		
7	W_q	0.041667	<--"=B1/(B2*(B2-B1))"		
8	U	0.625	<--"=B1/B2"		

Figure 6.2.3

Note that by adding a new employee, the waiting time in the queuing system is reduced to $W = 0.0667$ which is 4 minutes. Initially it was 12 minutes. So there is 8 minute reduction in total waiting time. Since for each reduced minute the store saves the loss of $70, the total saving is $70 x 8 mins = $560 per week. Deducting the extra employee cost yields the profit of $560 - $140 = $520.

The manager has another option. Instead of adding a new employee to the existing checkout counter, she can add a new checkout counter. This would split the waiting line in two separate lines with equal number of customers waiting in each. The effect of adding the new checkout counter would be the reduction in customer arrival rate. However, the cost of constructing it is an initial $5 000 plus an extra $180 per week for an additional cashier. The service rate would remain the same $\mu = 30$, but the arrival rate per counter is now $\lambda = 12.5$ per hour.

	A	B	C	D	E
1	λ	12.5			
2	μ	30			
3	P_0	0.583333	<--"=1-B1/B2"		
4	L	0.714286	<--"=B1/(B2-B1)"		
5	L_q	0.297619	<--"=B1^2/(B2*(B2-B1))"		
6	W	0.057143	<--"=1/(B2-B1)"		
7	W_q	0.02381	<--"=B1/(B2*(B2-B1))"		
8	U	0.416667	<--"=B1/B2"		

Figure 6.2.4

Since the arrival rate is reduced, the probability that the server is idle is increased to $P_0 = 58\%$ and the utilization factor (otherwise interpreted as the probability that the server is busy) is reduced to $U = 42\%$. The waiting time in the total queuing system is now $W = 0.0571$ hours which is 3.43 minutes. Recall that initially this quantity was 12 minutes, so there is 8.57 minutes reduction. This would save the store manager 8.57 x \$70 = \$600 per week. Subtracting the additional cost per week, which is \$200 results in the final profit for the store which is \$600 - \$200 = \$400 per week.

Because the initial payment of this project was \$5 000, it would take the store \$5000/400=12.5 weeks to break even from this project. After 12.5 weeks, the store starts to make profits of \$400 per week.

6.3. Finite Queue Length

In (6.2.2), P_n is defined to be the probability of n customers waiting in the waiting line. If M is defined as the maximum number of customers allowed in the system, then P_M is the probability that a customer will not be allowed in the system. This situation may arise because of some natural restrictions like the space for the waiting line may be limited or the model can be applied to cars in the drive queue of a fast food restaurant. Correspondingly, the equations (6.2.4), (6.2.5) and (6.2.6) can be rewritten as

$$L_q = L - \frac{\lambda(1 - P_M)}{\mu} \tag{6.3.1}$$

which is the average number of customers waiting in the queuing line. The average time a customer spends in the entire queuing system is

$$W_q = \frac{L}{\lambda(1 - P_M)} \tag{6.3.2}$$

and the average time a customer spends in the entire queuing system is

$$W_q = W - \frac{1}{\mu} \tag{6.3.3}$$

The probability that the system is empty corresponds to the probability P_0 (i.e. no one is in the system) defined as

$$P_0 = \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{M+1}} \tag{6.3.4}$$

and the probability that the system is full and no customer will be allowed in is

$$P_M = P_0 \left(\frac{\lambda}{\mu}\right)^M \tag{6.3.5}$$

The average number of customers, L, used in the equation (6.3.1) is defined as

$$L = \frac{\lambda/\mu}{1 - \frac{\lambda}{\mu}} - \frac{(M+1)\left(\frac{\lambda}{\mu}\right)^{M+1}}{1 - \left(\frac{\lambda}{\mu}\right)^{M+1}} \tag{6.3.6}$$

Example 6.3

Given the values of $\lambda = 25$, $\mu = 30$ and the maximum number of customers allowed is $M = 5$. The quantities in equations (6.3.1) – (6.3.6) are computed below

	A	B	C	D	E	F	G	H	I
1	λ	25							
2	μ	30							
3	M	10							
4	P_0	0.192586	<--"=(1-(B1/B2))/(1-(B1/B2)^(B3+1))"						
5	P_M	0.031104	<--"=B4*(B1/B2)^B3"						
6	L	3.289291	<--"=(B1/B2)/(1-B1/B2)-((B3+1)*(B1/B2)^(B3+1))/(1-(B1/B2)^(B3+1))"						
7	L_q	2.481878	<--"=B6-B1*(1-B5)/B2"						
8	W	0.135795	<--"=B6/(B1*(1-B5))"						
9	W_q	0.102462	<--"=B8-1/B2"						
10	U	0.807414	<--"=1-B4"						

Figure 6.3.1

So in the example above where the average number of arrivals per hour is 25 customers and the average service rate is 30 customers per hour with an additional restriction that maximum 10 customers are allowed to wait in the system, the probability that no one is in the system is $P_0 = 0.1926$. and nearly $U = 80\%$ of times the server is busy. The probability that the maximum number of customers has been reached and no additional customer will be allowed is $P_{10} = 0.0311$. The average number of customers waiting in the queuing system is $L = 3.29$ and on average they spend $W = 0.1358$ hours in the system.

If the queues are associated with costs as described in Example 6.2, the manager may proceed with one of the options described there – either consider adding new employees which will result in additional cost per period, or consider adding a new checkout counter with additional cashiers that will result in significant down payment and additional cost per period.

6.4. Finite Calling Population

For some waiting systems, there might be a limited number of potential customers that can arrive at a service facility. This situation is referred to as a *finite calling population.* Given the limited number of potential customers N (we call it the population size), in the single server model with a Poisson arrival and exponential service times, the equations (6.2.1) – (6.2.7) can be redefined as

$$P_0 = \frac{1}{\sum_{n=0}^{N} \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n} \tag{6.4.1}$$

which is the probability that no customer is in the waiting system. The probability that n customers are in the waiting system is defined as

$$P_n = \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n P_0, n = 1,2,\dots,N \tag{6.4.2}$$

In addition, the average number of customers in the waiting line is

$$L_q = N - \frac{\lambda + \mu}{\lambda}(1 - P_0) \tag{6.4.3}$$

and the average number of customers in the entire system (being served or waiting in a line) is given by

$$L = L_q + (1 - P_0) \tag{6.4.4}$$

The time a customer spends on average in the queuing line is

$$W_q = \frac{L_q}{(N-L)\lambda} \tag{6.4.5}$$

while the average time a customer spends in the queuing system is

$$W = W_q + \frac{1}{\mu} \tag{6.4.6}$$

Example 6.4

Consider a Memory Devices Inc. manufacturing plant which produces memory cards of various sizes. Due to a large demand for production, the plant operates 7 days per week and has 10 automated manufacturing machines in total. Continuous operations causes the machines to break down frequently and they end up in the repairing queue. There is a repair person assigned to this task. In Memory Devices Inc. each machine operates on average of 180 hours before it breaks down and a repair person is called. The average repair time of a single machine is 4 hours. The breakdown rate follows a Poisson distribution and the service time follows an exponential distribution. The company needs to analyze the machine idle time due to breakdowns and determine if the repair stuff is sufficient or needs to hire assistants to the senior repair person.

Since P_0 in the equation (6.4.1) contains the denominator with the sum of numbers, it is convenient to have it computed separately. This is done in the following figure

	A	B	C	D	E	F	G	H	I	J	K
1	λ	0.005556	<--"=1/180"			$\frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n$					
2	μ	0.25	<--"=1/4"		n						
3	N	10			0	1	<--"=FACT($B$3)/FACT($B$3-F3)*($B$1/$B$2)^F3"				
4					1	0.222222					
5					2	0.044444					
6					3	0.007901					
7					4	0.001229					
8					5	0.000164					
9					6	1.82E-05					
10					7	1.62E-06					
11					8	1.08E-07					
12					9	4.8E-09					
13					10	1.07E-10					

Figure 6.4.1

The column F in the figure contain each of the elements in the sum of the denominator in (6.4.1). The sum of these numbers is the denominator as a single number. The following figure shows the computations of equations (6.4.1) – (6.4.6). Note that in Figure 6.4.2, the sum of the

values in the column F is not explicitly computed but is included directly in the cell B4 which computes P_0. As a conclusion, $U = 21.63\%$ is the probability that the repair person is busy. Out of 10 operating machines, an average of almost $L = 0.27$ which is 0.27/10=0.027=2.7% are broken down and waiting in the queue of being repaired. Each machine that is broken down is either waiting in the line or being repaired for $W = 4.94$ hours. This is the time the machine is idle for. As long as the repair person is busy only for 21.63% of times, it can be concluded that he is adequately handling his job and no assistant is needed to be hired.

	A	B	C	D	E	F	G	H	I	J	K
1	λ	0.005556	<--"=1/180"			$\frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n$					
2	μ	0.25	<--"=1/4"		n						
3	N	10			0	1	<--"=FACT($B$3)/FACT($B$3-F3)*($B$1/$B$2)^F3"				
4	P_0	0.783711	<--"=1/SUM(I3:I13)"		1	0.222222					
5	L	0.266988	<--"=B6+(1-B4)"		2	0.044444					
6	L_q	0.050699	<--"=B3-(B1+B2)/B1*		3	0.007901					
7	W	4.937619	<--"=B8+1/B2"		4	0.001229					
8	W_q	0.937619	<--"=B6/((B3-B5)*B1		5	0.000164					
9	U	0.216289	<--"=1-B4"		6	1.82E-05					
10					7	1.62E-06					
11					8	1.08E-07					
12					9	4.8E-09					
13					10	1.07E-10					

Figure 6.4.2

6.5. The Multiple Server Model

Up to this point, all models described the single server model implying that the customers waiting in the line would all end up with a single server. A little more complex scenario is the multiple server model where customers lined up in a single waiting line end up with different servers. The examples of this model would be the airport check in counter or a bank office. The assumption of the arrival rate being Poisson distributed and the service time being exponentially distributed still persists. In addition, there is an infinite calling population meaning the population size is not restricted (unlike in Section 6.4).

The characteristics of this model is the arrival rate λ, which is the average number of arrivals per time period. The service rate μ which is the average number of customers served per time period per server. c is the number of servers and $c\mu$ is the mean effective service rate for the model for which $c\mu > \lambda$ must hold to avoid infinitely long queue.

In this model, the probability that no customer is waiting in the queuing system is given by the following equation

$$P_0 = \frac{1}{\left[\sum_{n=0}^{c-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right] + \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c \frac{c\mu}{c\mu - \lambda}} \tag{6.5.1}$$

The probability that n customers are waiting in the queuing system for $n > c$

$$P_n = \frac{1}{c!\, c^{n-c}}\left(\frac{\lambda}{\mu}\right)^n P_0 \tag{6.5.2}$$

and for $n \leq c$ we have

$$P_n = \frac{1}{n}\left(\frac{\lambda}{\mu}\right)^n P_0 \tag{6.5.3}$$

The average number of customers in the entire queuing system is given by

$$L = \frac{\lambda\mu\left(\frac{\lambda}{\mu}\right)^c}{(c-1)!\,(c\mu - \lambda)^2} P_0 + \frac{\lambda}{\mu} \tag{6.5.4}$$

while the average number of customers in the waiting line is

$$L_q = L - \frac{\lambda}{\mu} \tag{6.5.5}$$

The time a customer spends on average in the waiting system is given by

$$W = \frac{L}{\lambda} \tag{6.5.6}$$

and the average time a customer spends in the queuing line is

$$W_q = W - \frac{1}{\mu} = \frac{L_q}{\lambda} \tag{6.5.7}$$

Finally, the probability that a customer arriving in the system must wait because all servers are busy at once is

$$P_w = \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c \frac{c\mu}{c\mu - \lambda} P_0 \tag{6.5.8}$$

Example 6.5

Consider again the Toy Gift store with $\lambda = 25$ per hour. In that example, the average service rate was 30 customers per hour. Suppose that instead of 30, now the average service rate is 12 customers per hour per service representative. There are 3 store representatives (servers) in total. Adding a new representative leads to an additional weekly expense of $200 and for each

reduced minute of waiting in the system, the store saves \$70. Using the formulas (6.5.1) – (6.5.8), the following figure illustrates the computations

	A	B	C	D	E	F	G	H	I
1	λ	25	<--"=1/180"						
2	μ	12	<--"=1/4"						
3	c	3							
4	P_0	0.098178	<--"=1/(SUM(F3:F5)+1/FACT(B3)*(B1/B2)^B3*B3*B2/(B3*B2-B1))"						
5	L	3.183847	<--"=B1*B2*(B1/B2)^B3/((B3-1)*(B3*B2-B1)^2)*B4+B1/B2"						
6	L_q	1.100513	<--"=B5-B1/B2"						
7	W	0.127354	<--"=B5/B1"						
8	W_q	0.044021	<--"=B6/B1"						
9	P_w	0.484226	<--"=1/FACT(B3)*(B1/B2)^B3*B3*B2/(B3*B2-B1)*B4"						
10									
11									
12		$\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n$							
13	n								
14	0	1	<--"=1/FACT(E3)*($B$1/$B$2)^E3"						
15	1	2.083333							
16	2	2.170139							

Figure 6.5.1

The probability that there is no customer waiting in the queuing system is $P_0 = 0.0982$ and the probability that all servers are busy at the same time and the customer arriving in the system must wait for the service is $P_w = 0.4842$. Since the denominator of P_0 includes the sum, it is separately computed in the cells B14 - B16.

The average number of customers waiting in the system is $L = 3.1838$ and they spend on average $W = 0.1274$ hours. The average number of customers in the waiting line (waiting to be served) is $L_q = 1.1005$ and they spend on average $W_q = 0.044$ hours.

If the manager decides to add a new server, that will result in the following computations

	A	B	C	D	E	F	G	H	I
1	λ	25	<--"=1/180"						
2	μ	12	<--"=1/4"						
3	c	4							
4	P_0	0.119067	<--"=1/(SUM(F3:F5)+1/FACT(B3)*(B1/B2)^B3*B3*B2/(B3*B2-B1))"						
5	L	2.50734	<--"=B1*B2*(B1/B2)^B3/((B3-1)*(B3*B2-B1)^2)*B4+B1/B2"						
6	L_q	0.424006	<--"=B5-B1/B2"						
7	W	0.100294	<--"=B5/B1"						
8	W_q	0.01696	<--"=B6/B1"						
9	P_w	0.195043	<--"=1/FACT(B3)*(B1/B2)^B3*B3*B2/(B3*B2-B1)*B4"						
10									
11									
12		$\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n$							
13	n								
14	0	1	<--"=1/FACT(E3)*($B$1/$B$2)^E3"						
15	1	2.083333							
16	2	2.170139							
17	3	1.507041							

Figure 6.5.2

Addition of the new server caused the probability that a customer has to wait to reduce to $P_w = 0.195$. The probability that the no customer has to wait is not increased significantly, only to $P_0 = 0.1191$. The average number of customers in the system with $c = 4$ servers now is $L = 2.5073$ and the average waiting time is $W = 0.1$ hours which is 6 minutes.

If the number of employees is raised to $c = 4$ (i.e. increased by 1), the difference in waiting time in the system is $60(0.1274 - 0.1003) = 1.6204$ minutes. So, the saving from this reduction is \$70 x 1.6204 = \$113.43. However, the weekly expense increased by \$200. So, there is a negative profit. The manager decides that adding a new employee does not contribute to a positive profit and unlike the situation described in Example 6.2, no further action is taken in this sense.