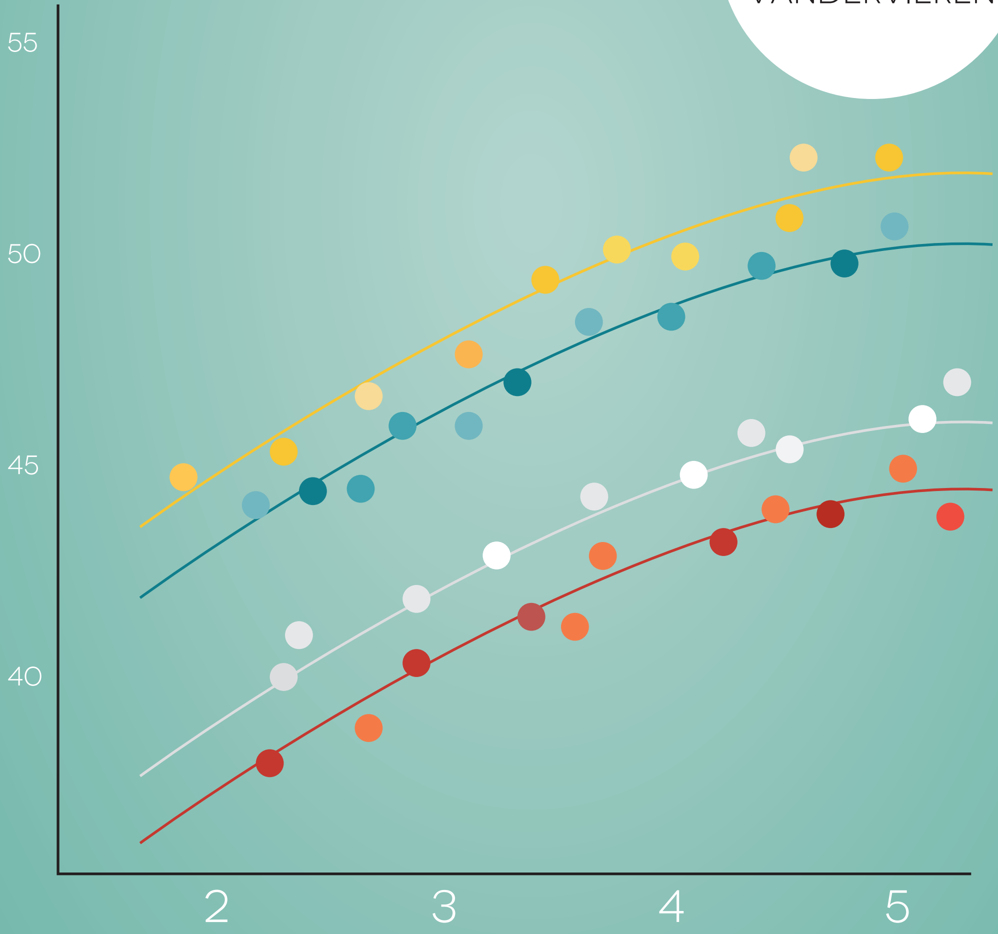
+ DIGITAL STUDY MATERIAL

REGRESSION AND ANALYSIS OF VARIANCE

PETER GOOS
& ELLEN
VANDERVIEREN



acco
learn

Contents

3

THE MULTIPLE LINEAR REGRESSION MODEL

The previous chapter has shown that the simple linear regression model is very useful. In some cases, however, the simple linear model can be improved upon by considering additional explanatory variables. In other words, it is sometimes possible to explain more of the variation in the response variable by including more than one explanatory variable in the regression model. In this chapter, we show how to generalize the simple linear regression model, involving just one explanatory variable, to a multiple linear regression model, involving more than one explanatory variable.

The focus in this chapter is on the use of multiple quantitative explanatory variables. It is possible to include qualitative explanatory variables in multiple regression models as well. We deal with qualitative (both nominal and ordinal) explanatory variables in Chapter 5.

Learning objectives of this chapter

Knowledge

- ✓ You understand the goal of multiple linear regression.
- ✓ You know how to interpret the parameters in a multiple linear regression model.
- ✓ You are familiar with the concepts of main effects, interaction effects and quadratic effects.
- ✓ You understand the principle of least squares regression and you can derive the least squares estimator for multiple linear regression.
- ✓ You are familiar with the assumptions behind the statistical inference in the context of multiple linear regression.
- ✓ You can derive the properties of the least squares estimator which form the basis for statistical inference.
- ✓ You can interpret the elements of a variance-covariance matrix.
- ✓ You know how to test the significance of the parameters in a multiple linear regression model.
- ✓ You know how to test hypotheses concerning one or more linear combinations of model parameters.
- ✓ You know how to make predictions using a multiple linear regression model.
- ✓ You can evaluate the quality of a multiple linear regression model using the (adjusted) coefficient of determination, various information criteria and the global F-test.
- ✓ You are familiar with the techniques to verify whether the assumptions behind the multiple linear regression model hold.
- ✓ You understand the mathematical derivations in the technical appendices.

Skills

- ✓ You are able to calculate the least squares regression model manually for small, well-structured data sets.
- ✓ You are able to compute the least squares regression model with the JMP software for any given data set.

- ✓ You can interpret the multiple linear regression output produced by JMP, including graphs such as the prediction profiler and the surface profiler.
- ✓ You can reconstruct most pieces of the JMP output for multiple linear regression.
- ✓ You know how to conduct tailor-made hypothesis tests for one or more linear combinations of model parameters in JMP.
- ✓ You can build an analysis of variance (ANOVA) table corresponding to a given multiple linear regression model.
- ✓ You can interpret residual plots made to check the model assumptions.

3.1 The model

In this chapter, we assume that the available data involves more than one explanatory variable. We denote the number of explanatory variables by k, and name the k explanatory variables $x_1, x_2, \dots, x_k$. If we denote the ith response by Y_i and the values of the k explanatory variables at the ith observation by $x_{i1}, x_{i2}, \dots, x_{ik}$, then the multiple linear regression model can be written as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + U_i. \tag{3.1}$$

Note that the expression on the right-hand side of this equation is a linear combination of the parameters $\beta_0, \beta_1, \dots, \beta_k$. As a result, the model belongs to the family of linear regression models. In total, the model involves $k+1$ unknown parameters or regression coefficients. These have to be estimated from the available data. As in the previous chapter, we denote the number of observations available by n.

The intercept β_0 represents the expected response when all explanatory variables take the value 0. The parameter β_i represents the expected change in the response when the ith explanatory variable increases by one unit, while keeping the other explanatory variables constant.

Obviously, the multiple linear regression model involves more variables than the simple linear regression model. Unavoidably, the mathematics are therefore more involved for the multiple linear regression model. The most elegant mathematical derivations and equations for multiple linear regression are obtained by making use of matrix algebra. In order to be able to use matrix algebra, we first need to express the multiple linear regression model in matrix form. To this end, we group all the responses $Y_1, Y_2, \dots, Y_n$ in an $(n \times 1)$-dimensional vector named $\boldsymbol{Y}$. Likewise, we group all the error terms $U_1, U_2, \dots, U_n$ in an $(n \times 1)$-dimensional vector named $\boldsymbol{U}$. In addition, we group all $k+1$ model parameters $\beta_0, \beta_1, \dots, \beta_k$ in a $((k+1) \times 1)$-dimensional vector named $\boldsymbol{\beta}$. Finally, we collect all n values for each of the k explanatory variables in a matrix.

To see how this works, we first write down the individual model for each of the n observations:

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_k x_{1k} + U_1. \\
Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_k x_{2k} + U_2. \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_k x_{nk} + U_n.
\end{aligned}$$

This system of n equations can be replaced by the following single matrix equation:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{bmatrix}. \tag{3.2}$$

The matrix involving all n values of the k explanatory variables has dimension $n \times (k+1)$ and is named $\mathbf{X}$. Using that name, the matrix equation can be written as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{U}. \tag{3.3}$$

The matrix $\mathbf{X}$ has as many rows as there are observations in the data set, and as many columns as there are parameters in the regression model. The first of these columns is a column of ones. This is because the first column of $\mathbf{X}$ corresponds to the intercept. The second column of $\mathbf{X}$ contains all values of the first explanatory variable, the third column of $\mathbf{X}$ contains all values of the second explanatory variable, and so on. The final column of $\mathbf{X}$ contains the n values of the kth explanatory variable. A commonly used name for the matrix $\mathbf{X}$ is **model matrix**.

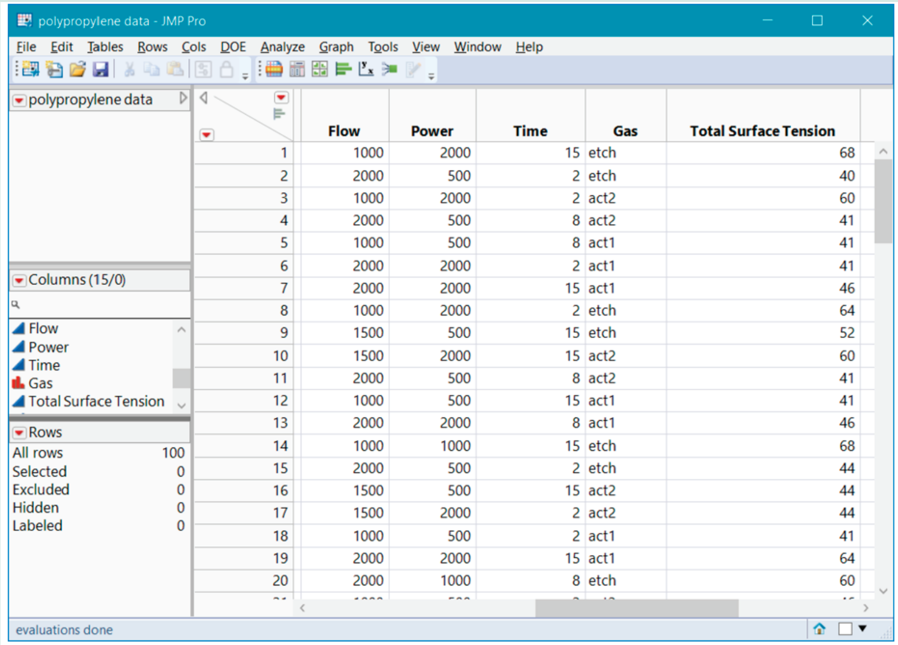
The multiple simple linear regression model in Equation (3.3) involves several symbols printed in bold. This is because, throughout this book, all vectors and matrices are printed in bold. The vectors $\boldsymbol{Y}$ and $\boldsymbol{U}$ are not only printed in bold, but also in italic. This is because they are random vectors, i.e., vectors whose elements are random variables. The model matrix $\mathbf{X}$ is not printed in italic, because its values are assumed to be constants rather than random variables. For the vector of unknown model parameters, the bold-printed Greek letter $\boldsymbol{\beta}$ is used. Bold print is used because it is a vector, while the Greek letter is used because it is a vector involving unknown parameters. The use of Greek letters for unknown parameters is a tradition in statistics. Finally, all vectors in this book ($\boldsymbol{Y}$, $\boldsymbol{U}$, $\boldsymbol{\beta}$, ...) are column vectors. We obtain row vectors by transposing column vectors, and we denote these transposes by $\boldsymbol{Y}'$, $\boldsymbol{U}'$ and $\boldsymbol{\beta}'$, for instance.

Example 3.1.1. Polypropylene is light and can be recycled easily. For these reasons, it is often used in the car industry for producing dashboards, door panels and bumpers. One problem, however, with polypropylene is that coatings do not adhere well to polypropylene. To improve the adhesive properties of the material, the polypropylene is subjected to a gas plasma treatment. Essentially, this means that a polypropylene to be coated is put in a sort of oven for a certain time. In the oven, the polypropylene undergoes a chemical treatment involving a well-chosen gas, a certain electrical power and a certain flow rate for the gas. Researchers from various Belgian companies performed a large experiment, involving $n = 100$ tests of various combinations of gas type, reaction time (expressed in minutes), electrical power (expressed in Watts) and gas flow rate (expressed in sccm). Part of the data is shown in Figure 3.1.

The researchers were interested in finding out how the total surface tension of the polypropylene depends on the reaction time, the power and the flow rate, because they believed that a large value for the total surface tension is required for a good adhesion of coatings to polypropylene. Therefore, one of the regression models they had in mind was

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + U_i, \tag{3.4}$$

where Y_i represents the total surface tension response at the ith observation (i.e., the *i*th experimental test), x_{i1} represents the flow rate, x_{i2} represents the power, and x_{i3} is the reaction time at the *i*th observation.



All rows	100
Selected	0
Excluded	0
Hidden	0
Labeled	0

	Flow	Power	Time	Gas	Total Surface Tension
1	1000	2000	15	etch	68
2	2000	500	2	etch	40
3	1000	2000	2	act2	60
4	2000	500	8	act2	41
5	1000	500	8	act1	41
6	2000	2000	2	act1	41
7	2000	2000	15	act1	46
8	1000	2000	2	etch	64
9	1500	500	15	etch	52
10	1500	2000	15	act2	60
11	2000	500	8	act2	41
12	1000	500	15	act1	41
13	2000	2000	8	act1	46
14	1000	1000	15	etch	68
15	2000	500	2	etch	44
16	1500	500	15	act2	44
17	1500	2000	2	act2	44
18	1000	500	2	act1	41
19	2000	2000	15	act1	64
20	2000	1000	8	etch	60

Figure 3.1. JMP data table showing part of the polypropylene data described in Example 3.1.1.

The intercept β_0 indicates what the expected total surface tension is for polypropylene that does not undergo any gas plasma treatment, in which case the flow rate, the power and the reaction time are all zero. The parameter or regression coefficient β_1 indicates what the expected change in total surface tension is when the flow rate is increased by 1 sccm. The parameter β_2 indicates what the expected change in total surface tension is when the power is increased by 1 Watt, and β_3 indicates what the expected change in total surface tension is when the reaction time is increased by one minute.

The three quantitative explanatory variables in the multiple linear regression model in Equation (3.4) will not explain the variation in total surface tension perfectly well. This is why the model involves an error term U_i. That error term captures the random variation in the total surface tension measured, as well as any influence of relevant explanatory variables that we overlook when using the model. For example, Figure 3.1 shows that three different gases were used when collecting the data. The possible impact of the gas type on the total surface tension is ignored in the model in Equation (3.4). Therefore, if the gas type indeed does have an influence on the surface tension, the error term will capture that influence.

Example 3.1.2. In Example 2.8.8, we considered the data set shown in Figure 2.43, concerning the monthly ice cream consumption in $n = 30$ successive months in the period 1951–1953. In that example, we explained that the variation in ice cream consumption, the response variable, could not be explained well by a single explanatory variable, the price index. A multiple linear regression model involving a second and third explanatory variable, the average monthly temperature (expressed in Fahrenheit) and the average income, does a much better job at explaining the ice cream consumption. That model can be written as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + U_i, \tag{3.5}$$

where Y_i represents the ice cream consumption in month i, x_{i1} represents the price level, x_{i2} represents the average temperature, and x_{i3} is the average income in month i.

The intercept β_0 indicates what the expected ice cream consumption is when the price level is zero, the average monthly temperature is zero and the average income is zero. This interpretation of the intercept is technically correct, but it lacks realism: a scenario in which the price level, the temperature and the income are all zero is not very likely to occur. This does not invalidate the regression model in this example, but it implies that we should perhaps seek for alternative ways to write the model. By centering the explanatory variables, for example, it is possible to obtain a model in which the intercept has a sensible interpretation.

The coefficient β_1 in Equation (3.5) indicates what the expected change in consumption is when the price level is increased by one unit. The parameter β_2 indicates what the expected change in consumption is when the temperature goes up by one degree, and β_3 indicates what the expected change in consumption is when the income goes up by one unit. Economic theory suggests that β_1 will be negative and that β_3 will be positive. As a matter of fact, higher prices generally lead to smaller sales numbers, and a larger income generally results in a larger consumption. Our common sense suggests that β_2 will be positive, since larger temperatures often stimulate us to seek refreshment in the form of ice cream. In this example, we know what the signs of the model parameters will be like. However, we have no idea about the absolute magnitude of the three parameters, β_1, β_2 and β_3. In other words, we do not know how large the impact of the three explanatory variables on the ice cream consumption is. That is why we need to estimate the multiple linear regression model in Equation (3.5).

3.2 Estimating the multiple linear regression model

As for the simple linear regression model discussed in the previous chapter, various approaches exist to estimate the multiple linear regression model. The most commonly used approach, however, is to seek those estimates for the $k + 1$ parameters $\beta_0, \beta_1, \dots, \beta_k$ that minimize the sum of the squared residuals. The resulting estimates are called the least squares estimates, while the corresponding estimator is called the least squares estimator.

In this section, we first introduce some of the notation we use and then derive an analytical expression for the least squares estimates and the least squares estimator for a general multiple linear regression model. Next, as a proof of concept, we apply that expression to a simple linear regression model and observe that it produces the same estimates as the formulas we derived in the previous chapter.

3.2.1 Notation

Once we have collected all data concerning the response variable and the k explanatory variables to be used in a multiple linear regression model, we can proceed to estimating the model. We denote all the individual observed response values by $y_1, y_2, \dots, y_n$ and the matrix containing all n observed response values by $\mathbf{y}$. The values $y_1, y_2, \dots, y_n$ thus represent the realizations of the random variables $Y_1, Y_2, \dots, Y_n$, while the vector $\mathbf{y}$ is the realization of the random vector $\boldsymbol{Y}$. We call the estimates we obtain for $\beta_0, \beta_1, \dots, \beta_k$ from the observed data $b_0, b_1, \dots, b_k$. For this reason, we call the vector containing all the estimates $\mathbf{b}$.

The estimated multiple linear regression model is generally written as

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik},$$

where $\hat{y}_i$ represents the response value predicted by the estimated model for the values $x_{i1}, x_{i2}, \dots, x_{ik}$ of the k explanatory variables at the ith observation. An alternative expression for the estimated model is

$$\hat{y}_i = \mathbf{x}_i'\mathbf{b},$$

where

$$\mathbf{x}_i' = \begin{bmatrix} 1 & x_{i1} & x_{i2} & \dots & x_{ik} \end{bmatrix}$$

is the ith row of the model matrix $\mathbf{X}$. A final way of writing the estimated regression model involves matrix notation:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b},$$

where $\hat{\mathbf{y}}$ is the vector containing the n individual predicted response values $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$.

3.2.2 The least squares estimates

In general, an estimated regression model will not fit the data perfectly well. As a result, the predicted responses $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ will deviate from the actually observed values $y_1, y_2, \dots, y_n$. The deviations between the predicted responses and the observed ones are named the residuals. The ith residual is calculated as

$$u_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik}).$$

The vector containing all n residuals can be calculated as

$$\mathbf{u} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b}. \tag{3.6}$$

Ideally, all residuals are close to zero, because this means that the estimated regression model fits the data very well. In that case, we say that we have a good model fit. In Section 2.2, we explained that minimizing the sum of all squared residuals,

$$\sum_{i=1}^{n} u_i^2,$$

is by far the most popular approach to ensure that all residuals are close to zero. This approach, which is called the least squares estimation approach, is not only mathematically convenient, but it also results in closed-form formulas for the parameter estimates $b_0, b_1, \dots, b_k$. The main weakness of the approach is that it is sensitive to outliers in the data. It is therefore important to verify that there are no outliers in the data, as they may severely affect the quality of the estimated model.

The least squares estimation approach requires determining the values for $b_0, b_1, \dots, b_k$ that minimize

$$\begin{aligned}
S(b_0, b_1, \dots, b_k) &= \sum_{i=1}^{n} u_i^2, \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \\
&= \sum_{i=1}^{n} \{y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik})\}^2, \\
&= \sum_{i=1}^{n} \{y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik}\}^2,
\end{aligned}$$

where n represents the number of observations.

To minimize a continuous differentiable function such as $S(b_0, b_1, \dots, b_k)$, we calculate the function's first derivatives with respect to the unknown parameters and set them to zero. This yields a set of equations. Solving that set of equations results in the values for the unknowns that minimize the function[1]. This is exactly the route we take here. The sum of squared residuals, $S(b_0, b_1, \dots, b_k)$, is a function of $k+1$ unknowns, $b_0, b_1, \dots, b_k$. Therefore, we differentiate $S(b_0, b_1, \dots, b_k)$ with respect to $b_0, b_1, \dots, b_k$, and set the $k+1$ resulting derivatives to zero.

[1] Strictly speaking, it is necessary to verify that the matrix of second derivatives is positive definite to be sure that the stationary point, detected by setting the first derivatives to zero, is indeed a minimum.

The derivatives of $S(b_0, b_1, \dots, b_k)$ with respect to the unknowns $b_0, b_1, \dots, b_k$ are as follows:

$$\begin{aligned}
\frac{\partial S(b_0, b_1, \dots, b_k)}{\partial b_0} &= \sum_{i=1}^{n} 2(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_kx_{ik})(-1), \\
&= -2\sum_{i=1}^{n}(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_kx_{ik}), \\
\frac{\partial S(b_0, b_1, \dots, b_k)}{\partial b_1} &= \sum_{i=1}^{n} 2(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_kx_{ik})(-x_{i1}), \\
&= -2\sum_{i=1}^{n} x_{i1}(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_kx_{ik}), \\
&\vdots \\
\frac{\partial S(b_0, b_1, \dots, b_k)}{\partial b_k} &= \sum_{i=1}^{n} 2(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_kx_{ik})(-x_{ik}), \\
&= -2\sum_{i=1}^{n} x_{ik}(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_kx_{ik}).
\end{aligned}$$

Setting these derivatives to zero and dividing the left- and right-hand sides of the resulting equations by -2 produces the following system of $k + 1$ equations with $k + 1$ unknown parameters:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_kx_{ik}) &= 0, \\
\sum_{i=1}^{n} x_{i1}(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_kx_{ik}) &= 0, \\
&\vdots \\
\sum_{i=1}^{n} x_{ik}(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_kx_{ik}) &= 0.
\end{aligned}$$

This can be rewritten as

$$\sum_{i=1}^{n} y_i - nb_0 - b_1\sum_{i=1}^{n} x_{i1} - b_2\sum_{i=1}^{n} x_{i2} - \cdots - b_k\sum_{i=1}^{n} x_{ik} = 0,$$

$$\sum_{i=1}^{n} x_{i1}y_i - b_0\sum_{i=1}^{n} x_{i1} - b_1\sum_{i=1}^{n} x_{i1}^2 - b_2\sum_{i=1}^{n} x_{i1}x_{i2} - \cdots - b_k\sum_{i=1}^{n} x_{i1}x_{ik} = 0,$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ik}y_i - b_0 \sum_{i=1}^{n} x_{ik} - b_1 \sum_{i=1}^{n} x_{i1}x_{ik} - b_2 \sum_{i=1}^{n} x_{i2}x_{ik} - \cdots - b_k \sum_{i=1}^{n} x_{ik}^2 = 0,$$

and as

$$\begin{aligned}
nb_0 + b_1 \sum_{i=1}^{n} x_{i1} + b_2 \sum_{i=1}^{n} x_{i2} + \cdots + b_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i, \\
b_0 \sum_{i=1}^{n} x_{i1} + b_1 \sum_{i=1}^{n} x_{i1}^2 + b_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + b_k \sum_{i=1}^{n} x_{i1}x_{ik} &= \sum_{i=1}^{n} x_{i1}y_i, \\
&\vdots \\
b_0 \sum_{i=1}^{n} x_{ik} + b_1 \sum_{i=1}^{n} x_{i1}x_{ik} + b_2 \sum_{i=1}^{n} x_{i2}x_{ik} + \cdots + b_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik}y_i.
\end{aligned}$$

Using matrix algebra, this system of equations can be written as a single matrix equation:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{i1}x_{ik} & \sum_{i=1}^{n} x_{i2}x_{ik} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}.$$

Now, observe that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{i1}x_{ik} & \sum_{i=1}^{n} x_{i2}x_{ik} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix}$$

and that

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}.$$

Consequently, our system of $k+1$ equations with $k+1$ unknown parameters can be written as

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}. \tag{3.7}$$

This expression contains the so-called normal equations. Premultiplying both sides of the expression by $(\mathbf{X}'\mathbf{X})^{-1}$, we first obtain

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Since

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}_{k+1},$$

where $\mathbf{I}_{k+1}$ denotes the $(k+1)$-dimensional identity matrix, this can be simplified to

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3.8}$$

This matrix expression produces the least squares estimate $\mathbf{b}$ of the parameter vector $\boldsymbol{\beta}$. An alternative way of writing this result is as follows:

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{bmatrix}.$$

In the above derivation, we made one implicit assumption, namely that the matrix product $\mathbf{X}'\mathbf{X}$ is not singular, and, as a consequence, that the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ exists. This technical condition is satisfied whenever the columns of the model matrix $\mathbf{X}$ are linearly independent, i.e., whenever no column of $\mathbf{X}$ is a linear combination of the other columns.

The derivation of the expression for the least squares estimates $b_0, b_1, \dots, b_k$ in the context of multiple linear regression, as given above, applies the same logic as the derivation for simple linear regression in the previous chapter. Therefore, it should be quite understandable to most readers. There also exists a short, mathematically more elegant derivation. That derivation, which involves matrix algebra right from the start and which requires differentiating with respect to a vector of unknowns rather than individual unknowns, is given in Section 3.7.1.

Example 3.2.1. To demonstrate how to use the matrix expression in Equation (3.8), we apply it to the data from Example 2.2.1 in Table 2.1, which we used for introducing simple linear regression. The data involves five observations for the response variable and the explanatory variable, so that n equals 5. The (5×1)-dimensional vector with observed response values, the (5×2)-dimensional model matrix and the (2×1)-dimensional parameter vector for the example are given by

$$\mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 4 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}, \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \tag{3.9}$$

respectively. Note that the model matrix involves two columns. The first column, involving five ones, corresponds to the intercept β_0, whereas the second column, involving the five values of the explanatory variable, corresponds to the slope β_1. Both of these parameters are included in the parameter vector $\boldsymbol{\beta}$.

To calculate the least squares estimates b_0 and b_1 of β_0 and β_1, we first need to calculate $\mathbf{X}'\mathbf{X}$, $(\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{X}'\mathbf{y}$. First, we have that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}.$$

The determinant of that matrix is $5 \times 55 - 15 \times 15 = 50$, so that

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{50} \begin{bmatrix} 55 & -15 \\ -15 & 5 \end{bmatrix} = \frac{1}{10} \begin{bmatrix} 11 & -3 \\ -3 & 1 \end{bmatrix}.$$

Finally,

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 10 \\ 37 \end{bmatrix},$$

so that

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{10} \begin{bmatrix} 11 & -3 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} 10 \\ 37 \end{bmatrix} = \frac{1}{10} \begin{bmatrix} -1 \\ 7 \end{bmatrix} = \begin{bmatrix} -0.1 \\ 0.7 \end{bmatrix}.$$

As a result, we obtain the same estimates for the intercept β_0 and the slope β_1 as in the previous chapter, in Example 2.2.3.

It is also instructive to apply the matrix expression for the least squares estimate **b** in Equation (3.8) to the general simple linear regression model in Equation (2.1). We conduct this exercise in Section 3.7.2.

3.2.3 Illustrations

Example 3.2.2. In Example 3.1.1, we sketched the context of an experiment involving $n = 100$ gas plasma treatment tests for improving the adhesive properties of polypropylene. Part of the data is shown in Figure 3.1. The researchers were interested in modeling the dependence of the total surface tension of the polypropylene on the flow rate, the power and the reaction time, by estimating the multiple regression model

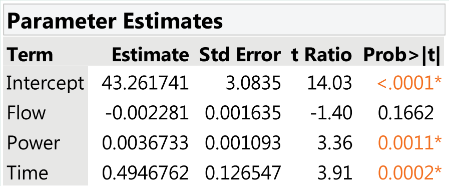
$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + U_i, \tag{3.10}$$

where Y_i represents the total surface tension response at the ith observation (i.e., the ith experimental test), x_{i1} represents the flow rate, x_{i2} represents the power, and x_{i3} is the reaction time at the ith observation.

Estimating the model using JMP results in the following estimated model:

$$\hat{y}_i = 43.2617 - 0.0023x_{i1} + 0.0037x_{i2} + 0.4947x_{i3}. \tag{3.11}$$

Consequently, the least squares estimates b_0, b_1, b_2 and b_3 of β_0, β_1, β_2 and β_3 are equal to 43.2617, −0.0023, 0.0037 and 0.4947. These point estimates are shown in the table labeled “`Parameter Estimates`” in the JMP output, more specifically in the column named “`Estimate`”. This is shown in Figure 3.2.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	43.261741	3.0835	14.03	<.0001*
Flow	-0.002281	0.001635	-1.40	0.1662
Power	0.0036733	0.001093	3.36	0.0011*
Time	0.4946762	0.126547	3.91	0.0002*

Figure 3.2. JMP output showing the parameter estimates b_0, b_1, b_2 and b_3 in Example 3.2.2, in which the electrical power is expressed in Watt.

To perform a multiple linear regression model in JMP, we can use the “`Fit Model`” platform in the “`Analyze`” menu. The dialog window of the “`Fit Model`” platform is shown in Figure 3.3. That window shows the names of all variables in the data set on the extreme left. To perform a regression analysis, we first need to select the response variable, the total surface tension, and click on the “`Y`” button. Next, we need to select the explanatory variables flow rate, power and time, and click on the “`Add`” button. This produces the window displayed in Figure 3.4. As soon as the response variable has been indicated, JMP recognizes its quantitative nature and it mentions the personality “`Standard Least Squares`” at the top right of the dialog window. In doing so, JMP assumes that we want to perform a multiple linear regression analysis. This may not always be appropriate. For instance, if we believe our response variable is Poisson distributed or gamma distributed, we may wish to switch the personality from “`Standard Least Squares`” to “`Generalized Linear Model`”. For the polypropylene data, however, a multiple linear regression analysis is what we want, so we can stick to “`Standard Least Squares`”. All we need to do to let JMP finally estimate the model is press the “`Run`” button. One of the pieces of output produced is shown in Figure 3.2.

One interesting piece of output that JMP provides is named the “`Prediction Profiler`”, which visualizes the estimated impacts of the explanatory variables. The “`Prediction Profiler`” for the estimated multiple linear regression model in Equation (3.11) is shown in Figure 3.5. The figure consists of three panels, one for each explanatory variable in the model. Each panel contains a solid black line, which shows the effect of the corresponding explanatory variable on the response, the total surface tension. The flow rate has a negative effect on the total surface tension, while the power and the reaction time have a positive effect.

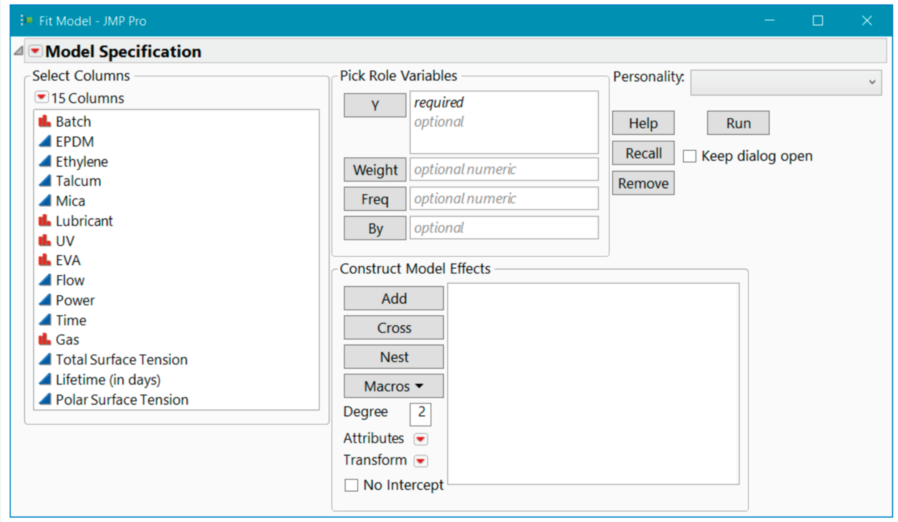
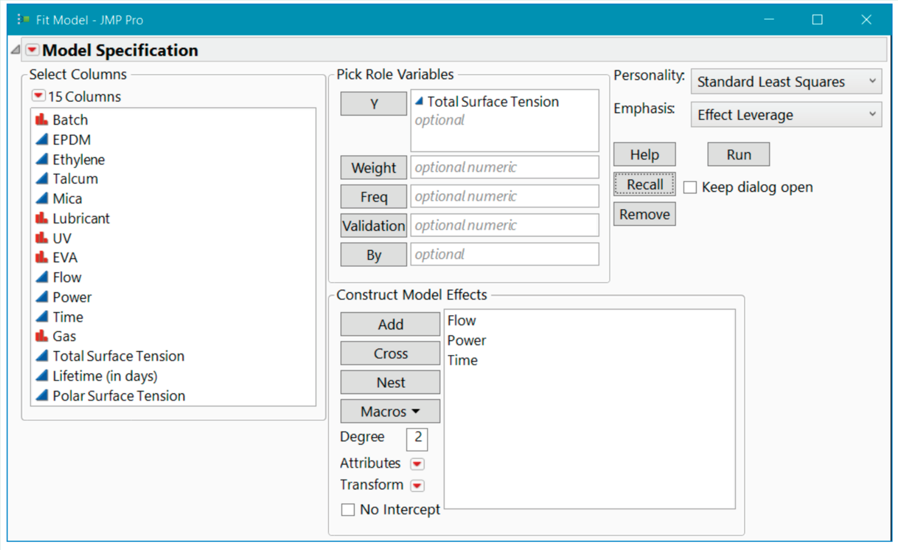


Figure 3.3. Initial dialog window of the "`Fit Model`" platform.



Y: Total Surface Tension

Personality: Standard Least Squares

Emphasis: Effect Leverage

Degree: 2

Flow
Power
Time

Figure 3.4. Dialog window of the "`Fit Model`" platform with input required for estimating the model in Equation (3.10) to the polypropylene data.

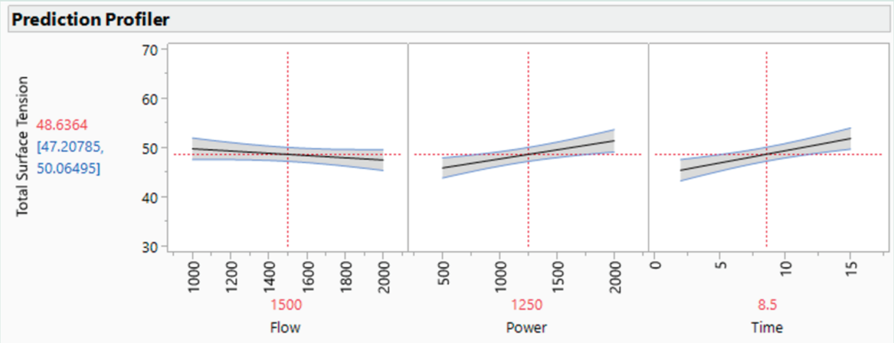


Figure 3.5. "`Prediction Profiler`" visualizing the estimated multiple linear regression model in Equation (3.11).

The dashed vertical lines in the three panels indicate the values of the explanatory variables under consideration, namely 1500 sccm for the flow rate, 1250 Watt for the power and 8.5 minutes for the reaction time. Substituting these three values in the estimated model in Equation (3.11) yields a point prediction for the total surface tension:

$$\hat{y}_i = 43.2617 - 0.0023 \times 1500 + 0.0037 \times 1250 + 0.4947 \times 8.5 = 48.6364.$$

This predicted response value is shown on the vertical axis. Increasing the reaction time by one minute results in an increase of 0.4947 of the total surface tension, from 48.6364 to 49.1311. This is shown by means of the "`Prediction Profiler`" in Figure 3.6. Increasing the reaction time by one minute in the "`Prediction Profiler`" can be done in two ways. One way is to click on the original reaction time on the horizontal axis, 8.5, and then enter the new value, 9.5. Another way is to grab the dashed vertical line in the profiler's panel for reaction time, and pull it to the right, so that the reaction time increases from 8.5 to 9.5. This shows that the "`Prediction Profiler`" is an interactive tool.

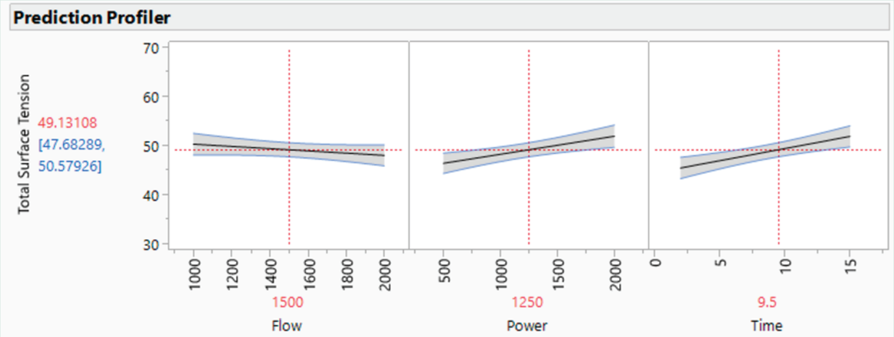
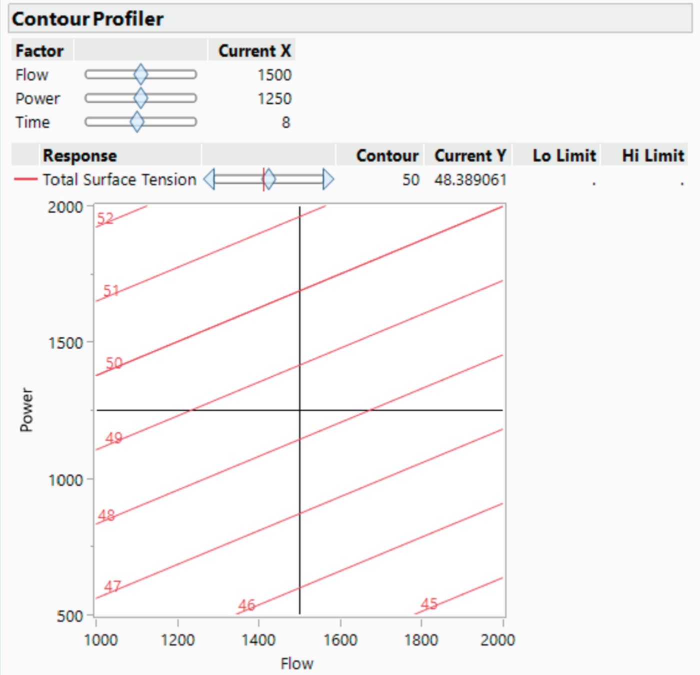


Figure 3.6. "`Prediction Profiler`" visualizing the impact on the total surface tension of an increase in reaction time of one minute.

The “Prediction Profiler” is one way to visualize the estimated regression model. It is possible to activate a two-dimensional “Contour Profiler” and a three-dimensional “Surface Profiler” as well. The “Contour Profiler” is shown in Figures 3.7 and 3.8. The “Contour Profiler” shows how the total surface tension depends on the flow (horizontal axis) and the power (vertical axis), for a given value of the reaction time. That given value is indicated at the top of the picture. It equals 8 minutes. Each of the parallel red lines in the figure is a contour. In this example, a contour shows all combinations of flow and power which result in a specified predicted value for the surface tension, provided the reaction time is 8 minutes. For instance, a flow of 2000 sccm in combination with a power of 1455 Watt produces a predicted total surface tension of 48, when the reaction time is 8 minutes. The same goes for a flow of 1250 sccm and a power of 985 Watt.



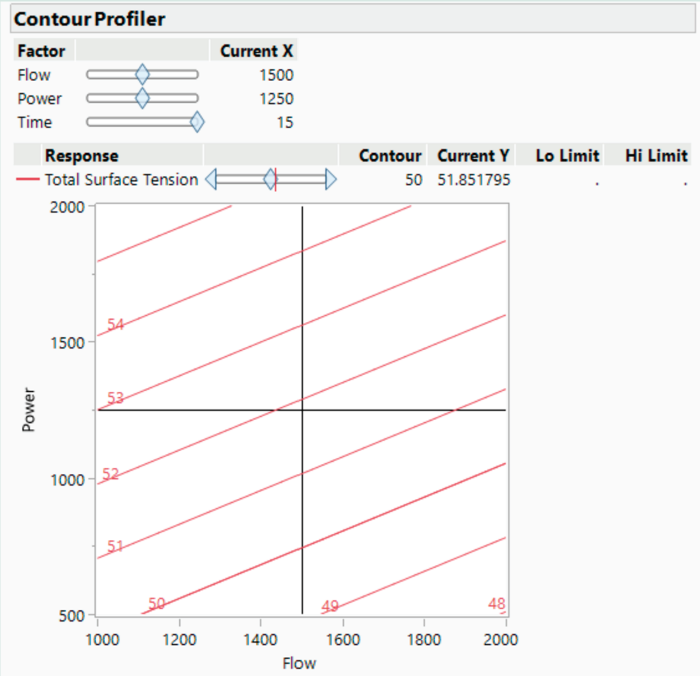
Contour Profiler

Factor		Current X
Flow		1500
Power		1250
Time		8

Response		Contour	Current Y	Lo Limit	Hi Limit
Total Surface Tension		50	48.389061	.	.

Figure 3.7. “Contour Profiler” visualizing the impact of flow and power on the total surface tension, given a reaction time of 8 minutes.

The “Contour Profiler” is also an interactive tool. We can use the sliders for flow, power and time to study the impact of these three explanatory variables on the response. We can also enter any specific value for the three explanatory variables directly, rather than by manipulating the sliders. Changing the flow and the power will affect the location of the black vertical and horizontal lines, as these indicate the current values of the explanatory variables flow rate and power, and the predicted value for the response. Changing the flow and the power does not affect the position of the contour lines. However, changing the reaction time will have an impact on the position of the contour lines. This is illustrated in Figure 3.8, where the reaction time has been raised from 8 to 15 minutes. The impact of that change is not only that the predicted response for the current flow-power combination went up from 48.39 to 51.85, but also that different contour lines appear in the picture. More specifically, the new figure contains contours for surface tensions ranging from 48 to 55, whereas the initial figure contained contours for surface tensions ranging from 45 to 52.



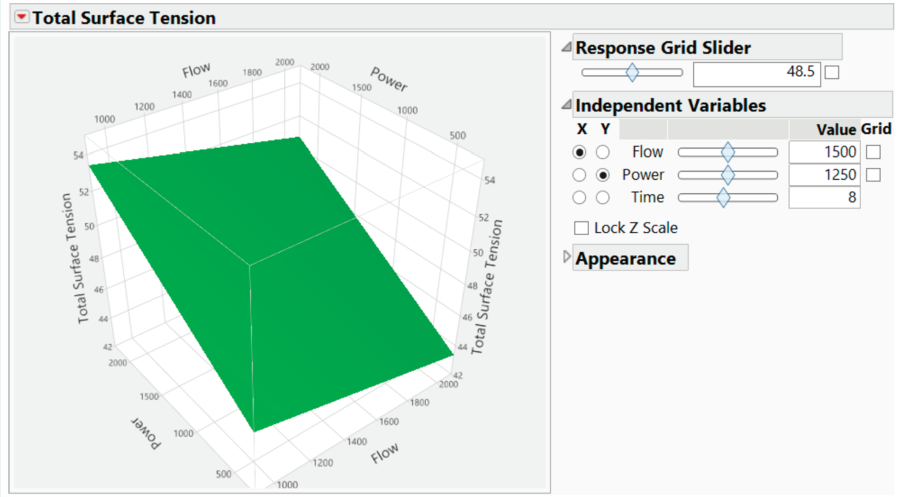
Contour Profiler

Factor	Current X
Flow	1500
Power	1250
Time	15

Response	Contour	Current Y	Lo Limit	Hi Limit
Total Surface Tension	50	51.851795	.	.

Figure 3.8. “Contour Profiler” visualizing the impact of flow and power on the total surface tension, given a reaction time of 15 minutes.

The “Surface Profiler” is shown in Figure 3.9. The “Surface Profiler” also shows how the total surface tension depends on the flow and the power, for a given value of the reaction time. The “Surface Profiler” is, however, a three-dimensional picture, in which the response is shown on the vertical axis, and the flow and the power appear on the two horizontal x- and y-axes. The given value for reaction time, 8 minutes, is indicated at the right-hand side of the picture. An interesting feature of the surface profiler is that it can be rotated, in order to study the dependence of the response on the two explanatory variables displayed from various angles.



Total Surface Tension

Response Grid Slider

48.5 □

Independent Variables

X	Y		Value	Grid
●	○	Flow	1500	□
○	●	Power	1250	□
○	○	Time	8	

□ Lock Z Scale

Appearance

Figure 3.9. “Surface Profiler” visualizing the impact of flow and power on the total surface tension, given a reaction time of 8 minutes.

3.2.4 A few notes concerning coefficient estimates

In Example 3.2.2, the estimated regression coefficient of the third explanatory variable, the reaction time, is substantially larger in absolute value than the estimated regression coefficient of the other two explanatory variables, the flow rate and the power. For this reason, it is tempting to interpret this as evidence that the reaction time is the most important explanatory variable. Drawing this conclusion would, however, not be correct. To understand why this is the case, consider a modified data set in which the electrical powers used in the polypropylene experiment are expressed in kiloWatt rather than in Watt.

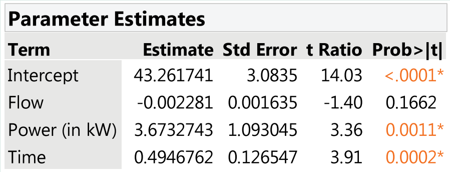
Example 3.2.3. In this example, we investigate to what extent the estimated multiple linear regression model for the polypropylene data changes when we switch the measurement unit for the electrical power from Watt to kiloWatt. The change in measurement unit for the power essentially means that we no longer work with the variable x_{2i} in the model, but with $x_{2i}^* = x_{2i}/1000$. The multiple regression model under consideration can then be written as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2^* x_{i2}^* + \beta_3 x_{i3} + U_i. \tag{3.12}$$

Estimating that model using JMP produces the following fitted model:

$$\hat{y}_i = 43.2617 - 0.0023 x_{i1} + 3.6733 x_{i2}^* + 0.4947 x_{i3}. \tag{3.13}$$

Comparing this model to the one in Equation (3.11), we can see that the least squares estimates b_0, b_1 and b_3 of β_0, β_1 and β_3 are not affected, but the estimated regression coefficient for the new explanatory variable, the power expressed in kiloWatt (x^*_{2i}), is 1000 times larger than the estimated regression coefficient for the original explanatory variable, the power expressed in Watt (x_{2i}). Figure 3.10 shows the "`Parameter Estimates`" table for the model in Equation (3.12). It is interesting to compare that table to the one in Figure 3.2 and to observe that the parameter estimate for the power increased by a factor of 1000, just like the standard error. As a result, the t-test statistic or t ratio and the p-value are not affected by the change in measurement unit. For details on the standard errors and t-tests for individual model parameters, we refer to Sections 3.4 and 3.5.4.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	43.261741	3.0835	14.03	<.0001*
Flow	-0.002281	0.001635	-1.40	0.1662
Power (in kW)	3.6732743	1.093045	3.36	0.0011*
Time	0.4946762	0.126547	3.91	0.0002*

Figure 3.10. JMP output showing the parameter estimates b_0, b_1, b^*_2 and b_3 in Example 3.2.3, in which the electrical power is expressed in kiloWatt.

Because the estimated regression coefficient for the power is increased by a factor of 1000 and the explanatory variable's value is decreased by a factor of 1000, the predicted response values are not affected by the change in measurement unit. To verify this, we can use the newly estimated model to make a prediction of the total surface tension when a flow rate of 1500 sccm, a power of 1250 Watt or 1.25 kiloWatt, and a reaction time of 8.5 minutes are used:

$$\hat{y}_i = 43.2617 - 0.0023 \times 1500 + 3.6733 \times 1.25 + 0.4947 \times 8.5 = 48.6364.$$

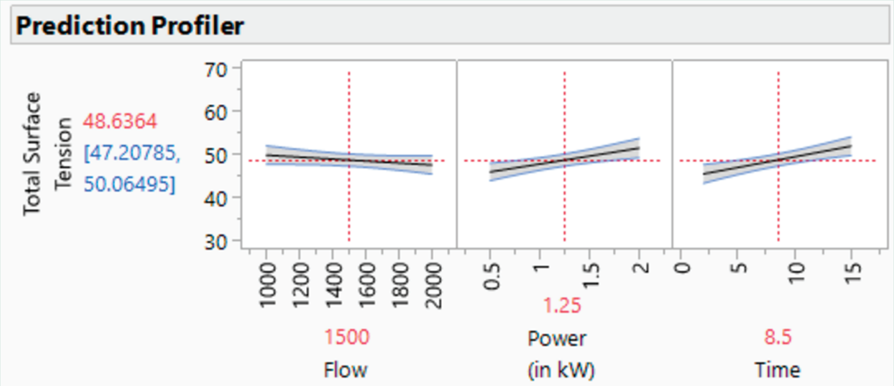


Figure 3.11. "`Prediction Profiler`" visualizing the estimated multiple linear regression model in Equation (3.13).

This prediction is identical to the one obtained in Example 3.2.2. This is confirmed by the "`Prediction Profiler`" in Figure 3.11, which visualizes the estimated impacts of the explanatory variables. Comparing Figures 3.5 and 3.11, we can see that the only difference is in the horizontal axis for the second explanatory variable, the electrical power. In Figure 3.11, the values on this axis range from 0.5 to 2 because they are expressed in kiloWatt, while they range from 500 to 2000 in Figure 3.5 because they are expressed in Watt.

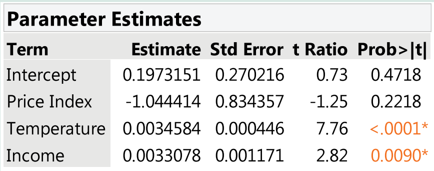
A key lesson from this example is that the absolute magnitude of an estimated regression coefficient in itself does not provide any information concerning the importance of the corresponding explanatory variable. This is because that magnitude is to a large extent determined by the measurement unit used, or, more generally, the scale used, for the explanatory variable. Judging how important each explanatory variable is for the model therefore requires a more subtle statistical analysis, for instance, involving coded explanatory variables and p-values resulting from significance tests. Significance tests are discussed in Section 3.5.4. The next example discusses the fact that, sometimes, we have a priori expectations concerning the signs of the regression coefficients, and that this offers a way to perform an initial quality check for a regression model.

Example 3.2.4. In Example 3.1.2, we discussed a sensible multiple linear regression model for the monthly ice cream consumption data shown in Figure 2.43. Estimating the model in JMP leads to the following expression:

$$\hat{y}_i = 0.1973 - 1.0444x_{i1} + 0.0035x_{i2} + 0.0033x_{i3}, \tag{3.14}$$

where $\hat{y}_i$ represents the predicted ice cream consumption in month i, x_{i1} represents the price index, x_{i2} represents the average temperature, and x_{i3} is the average income in month i. The JMP table with parameter estimates and the corresponding prediction profiler are shown in Figures 3.12 and 3.13, respectively.

An important quality check for an estimated model is to verify that the parameter estimates possess the expected signs. In this example, we expected a negative impact of the price index and a positive impact of the temperature and the income on the ice cream consumption. The point estimates -1.0444, 0.0035 and 0.0033 of the regression coefficients corresponding to these three explanatory variables indeed possess the expected signs.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.1973151	0.270216	0.73	0.4718
Price Index	-1.044414	0.834357	-1.25	0.2218
Temperature	0.0034584	0.000446	7.76	<.0001*
Income	0.0033078	0.001171	2.82	0.0090*

Figure 3.12. JMP output showing the parameter estimates b_0, b_1, b_2 and b_3 in Example 3.2.4.

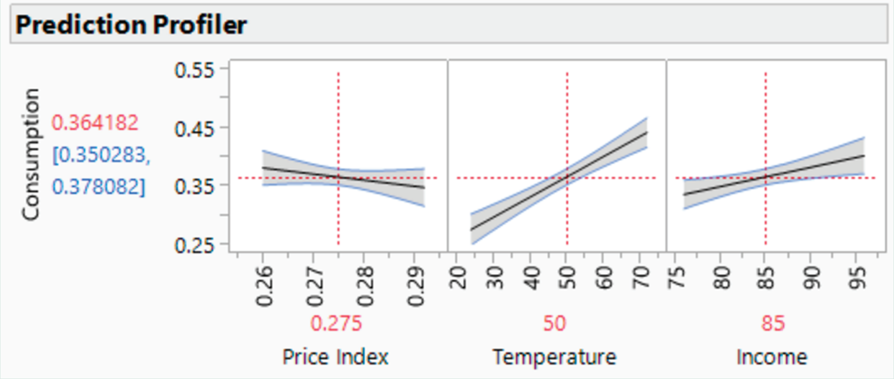


Figure 3.13. "`Prediction Profiler`" visualizing the estimated multiple linear regression model in Equation (3.14).

In order to be able to perform the quality check, we need to have expectations concerning the signs of the model parameters. In the above example, we built expectations based on economic theory. In some chemical or engineering applications, chemical insights or engineering knowledge may provide us with an idea about the nature of the relationships between explanatory variables and responses. In such scenarios, the regression analysis can be viewed as confirmatory: the signs of the estimated regression coefficients provide empirical support for the theory. However, it is quite common to not have any a priori idea about the signs of regression coefficients. In such cases, the regression analysis is exploratory in nature.

3.2.5 The least squares estimator

In the above derivations, we assumed that all the data had been collected. For this reason, we denoted the individual responses by $y_1, y_2, \ldots, y_n$ and the response vector by $\mathbf{y}$. We denoted the predicted response values by $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ and the predicted response vector by $\hat{\mathbf{y}}$. For the individual residuals, we used the symbols $u_1, u_2, \ldots, u_n$, and, for the vector of residuals, we used the symbol $\mathbf{u}$. In each case, we used a lowercase letter, to stress that, after the data has been collected, the responses and any quantities derived from the observed data are no longer random variables. The estimates we calculate from the observed data are also not random variables, which is why we denote them by $b_0, b_1, \ldots, b_k$ or by $\mathbf{b}$.

As long as we did not record the response values and the data set is not complete, the actual values are still unknown. The responses should then still be considered as random variables. We emphasize this by using uppercase letters for the responses: we denote individual responses by $Y_1, Y_2, \ldots, Y_n$ then, and the vector with all n responses by $\boldsymbol{Y}$. Every quantity derived from the responses then is a function of random variables, and is therefore a random variable itself. We stress this by using different symbols for these quantities whenever the data has not yet been collected, and we adjust our terminology as well. For instance, as long as the response values have not be acquired, we use the term least squares estimator instead of least squares estimate, and we use the symbols $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ instead of $b_0, b_1, \dots, b_k$, or the random vector $\hat{\boldsymbol{\beta}}$ instead of **b**. The least squares estimator is calculated as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y}.$$

The predicted responses should also be considered as random variables as long as the data has not been recorded. We denote an individual predicted response by $\hat{Y}_i$:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik},$$

or

$$\hat{Y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}},$$

where, as before,

$$\mathbf{x}_i = \begin{bmatrix} 1 & x_{i1} & x_{i2} & \dots & x_{ik} \end{bmatrix}'$$

and

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 & \hat{\beta}_2 & \dots & \hat{\beta}_k \end{bmatrix}'.$$

The vector containing all individual predicted responses $\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n$ is denoted by $\hat{\boldsymbol{Y}}$:

$$\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

This expression is often rewritten as

$$\hat{\boldsymbol{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y} = \mathbf{H}\boldsymbol{Y},$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

is called the **hat matrix**. This is because it is as if that matrix puts a hat on the vector $\boldsymbol{Y}$. The hat matrix is a symmetric $n \times n$ matrix. The hat matrix plays an important role in regression analysis, as we will see later.

Finally, as long as we did not yet record the responses, the ith residual is also a random variable, which we denote by $\hat{U}_i$:

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

The vector containing all n residuals can be calculated as

$$\hat{\boldsymbol{U}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \boldsymbol{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{Y}. \qquad (3.15)$$

In all of the above expressions, the values of the k explanatory variables at each of the n responses are regarded as constants, and, therefore, the model matrix $\mathbf{X}$ is considered as a matrix of constant values. Consequently, also the matrix expressions $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ are viewed as constant matrices, just like the hat matrix $\mathbf{H}$ and $\mathbf{I}_n - \mathbf{H}$. As a result, all the estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ contained within $\hat{\boldsymbol{\beta}}$, all the individual predicted responses $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$ contained within $\hat{\boldsymbol{Y}}$ and all residuals $\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_n$ contained within $\hat{\boldsymbol{U}}$ are linear combinations of the responses $Y_1, Y_2, \ldots, Y_n$ contained within $\boldsymbol{Y}$. This facilitates the derivation of the statistical properties of the least squares estimators, the predictions obtained from them and the corresponding residuals substantially.

3.3 Properties of the least squares estimator

In order to build confidence intervals, make interval predictions and perform statistical tests, it is important to know the statistical distribution of the least squares estimator $\hat{\boldsymbol{\beta}}$. We start by deriving the expected value of the least squares estimator and its variance-covariance matrix. The derivations of the expected value, the variance-covariance matrix and the statistical distribution, however, require four assumptions concerning the estimated multiple linear regression model.

3.3.1 Assumptions in the multiple linear regression model

The statistical analysis of the multiple linear regression model requires the same kinds of assumptions as in the case of a simple linear regression model. The four required assumptions are the following:

Assumption 1. *We assume that the true relationship between the response variable* Y_i *and the explanatory variables* $x_{i1}, x_{i2}, \ldots, x_{ik}$ *is described by* $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + U_i$ *for each observation* i*, where* $\beta_0, \beta_1, \ldots, \beta_k$ *are unknown constants, and that* $\mathrm{E}(U_i) = 0$. *Equivalently, we assume that* $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.

If the first assumption holds, we say that the regression model is correctly specified. This means, for example, that we do not overlook any other relevant explanatory variable(s), and that neither the response variable nor the explanatory variables have to be transformed (for instance, using a logarithmic or an inverse transformation) for the linear model to be true.

Assumption 2. *We assume that the random errors* $U_1, U_2, \ldots, U_n$ *are independent, and, hence, that* $\mathrm{cov}(U_i, U_j) = 0$ *whenever* $i \neq j$.

Another way of phrasing this assumption is by saying that the observations are made independently, so that the responses $Y_1, Y_2, \ldots, Y_n$ are independent. In that case, $\mathrm{cov}\left(Y_i, Y_j\right) = 0$ whenever $i \neq j$, and any pair of responses is uncorrelated.

Assumption 3. *We assume that* $\mathrm{var}(U_i) = \sigma^2(U_i) = \sigma^2$*: the variance of the error term is unknown, but it is the same for each observation.*

If this assumption holds, then also

$$\mathrm{var}(Y_i) = \mathrm{var}\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + U_i\right) = \mathrm{var}(U_i) = \sigma^2,$$

because $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ and $x_{i1}, x_{i2}, \ldots, x_{ik}$ are all considered as constants.

This situation, where the variance of all responses $Y_1, Y_2, \ldots, Y_n$ is the same, is referred to as **homoscedasticity**. The opposite situation, in which not all observations have the same variance, is called heteroscedasticity.

Assumption 4. *We assume that the error terms $U_1, U_2, \ldots, U_n$ are normally distributed.*

The assumption of normality is not only mathematically convenient, but it is also justifiable. Each error term U_i describes the fact that the response Y_i deviates to some extent from its expected value, $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$. These deviations occur because there exist many factors that influence the response to some extent, but which are not incorporated in the model. So, the error term can be thought of as the sum of the influences of all kinds of variables that impact the response and that differ from the explanatory variable. According to the central limit theorem, the sum or the linear combination of a large number of (independent) random variables is approximately normally distributed. Therefore, assuming that the error terms are normally distributed is justifiable.

The four assumptions made can also be expressed in matrix notation:

1. The first assumption states that $\boldsymbol{Y} = \mathbf{X}\beta + \boldsymbol{U}$, with $\mathrm{E}(\boldsymbol{U}) = \mathbf{0}_n$. This is equivalent to stating that $\mathrm{E}(\boldsymbol{Y}) = \mathbf{X}\beta$. Note that, by $\mathbf{0}_n$, we denote an n-dimensional vector of zeros.
2. Because of the normality assumption, the assumptions of homoscedasticity and independence of the error terms are equivalent to

$$\mathbf{var}(\boldsymbol{U}) = \mathbf{var}(\boldsymbol{Y}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_n, \tag{3.16}$$

   where $\mathbf{I}_n$ represents the n-dimensional identity matrix. In some of the later chapters, we denote the $n \times n$ variance-covariance matrix of the error terms and the responses by $\mathbf{V}$. Here, $\mathbf{V}$ is as simple as $\sigma^2\mathbf{I}_n$.
3. The random vector $\boldsymbol{U}$ has a multivariate normal distribution whose expected value is the n-dimensional null vector $\mathbf{0}_n$ and whose variance-covariance matrix is $\sigma^2\mathbf{I}_n$. The random vector $\boldsymbol{Y}$ then also has a multivariate normal distribution with variance-covariance matrix $\sigma^2\mathbf{I}_n$, but with expectation $\mathbf{X}\beta$. Mathematically, we write these distributional properties as

$$\boldsymbol{U} \sim N(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$$

   and

$$\boldsymbol{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}_n).$$

3.3.2 Unbiasedness

If the estimated model is correctly specified, the expectation of the least squares estimator $\hat{\boldsymbol{\beta}}$ equals the vector of unknown model parameters, $\boldsymbol{\beta}$. In other words, if Assumption 1 holds, then the least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. This can be shown as follows[2]:

$$\begin{aligned}\mathbf{E}(\hat{\beta}) &= \mathbf{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}),\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}\,(\mathbf{Y})\,.\end{aligned}$$

When Assumption 1 is true, then $\mathbf{E}(\mathbf{Y}) = \mathbf{X}\beta$, and we obtain

$$\begin{aligned}\mathbf{E}(\hat{\beta}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta,\\ &= \boldsymbol{\beta},\end{aligned}$$

since $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}_{k+1}$.

3.3.3 Variance-covariance matrix

The variance-covariance matrix of the least squares estimator $\hat{\boldsymbol{\beta}}$ contains all the variances of the estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ of the $k + 1$ model parameters $\beta_0, \beta_1, \dots, \beta_k$, as well as the covariances between all pairs of estimators, $(\hat{\beta}_0, \hat{\beta}_1), (\hat{\beta}_0, \hat{\beta}_2), \dots, (\hat{\beta}_{k-1}, \hat{\beta}_k)$. Because $\hat{\boldsymbol{\beta}}$ is a $(k + 1)$-dimensional random vector, its variance-covariance matrix is a symmetric $(k+1) \times (k+1)$ matrix. We denote that variance-covariance matrix by $\mathbf{var}(\hat{\beta})$ and it is equal to

$$\mathbf{var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

This expression for the least squares estimator's variance-covariance matrix is derived as follows[3]:

$$\begin{aligned}\mathbf{var}(\hat{\beta}) &= \mathbf{var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right),\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,(\mathbf{var}\,(\mathbf{Y}))\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)',\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,(\mathbf{var}\,(\mathbf{Y}))\,\mathbf{X}\left((\mathbf{X}'\mathbf{X})^{-1}\right)',\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,(\mathbf{var}\,(\mathbf{Y}))\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

[2] The first step in the proof of the unbiasedness makes use of the fact that the expectation of the product of a constant matrix, say $\mathbf{C}$, and a random vector, say $\boldsymbol{Y}$, equals the product of the constant matrix $\mathbf{C}$ and the expectation of the random vector $\boldsymbol{Y}$: $\mathbf{E}(\mathbf{C}\boldsymbol{Y}) = \mathbf{CE}(\boldsymbol{Y})$. In the proof, the matrix expression $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ plays the role of the constant matrix $\mathbf{C}$.

[3] The first step in the derivation of the variance-covariance matrix makes use of the fact that the variance-covariance matrix of the product of a constant matrix, say $\mathbf{C}$, and a random vector, say $\boldsymbol{Y}$, is equal to the product of the constant matrix $\mathbf{C}$, the variance-covariance matrix of the random vector $\boldsymbol{Y}$, and the transpose of the constant matrix: $\mathbf{var}(\mathbf{C}\boldsymbol{Y}) = \mathbf{C}(\mathbf{var}(\boldsymbol{Y}))\mathbf{C}'$. In the derivation, the matrix expression $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ plays the role of the constant matrix $\mathbf{C}$.

When Assumptions 2 and 3 hold, then $\mathbf{var(Y)} = \sigma^2\mathbf{I}_n$ and the covariance matrix matrix can be written as

$$\begin{aligned}\mathbf{var}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X'X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X'X})^{-1}, \\ &= \sigma^2(\mathbf{X'X})^{-1}\mathbf{X}'\mathbf{I}_n\mathbf{X}(\mathbf{X'X})^{-1}, \\ &= \sigma^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1}, \\ &= \sigma^2(\mathbf{X'X})^{-1}\mathbf{I}_{k+1}, \\ &= \sigma^2(\mathbf{X'X})^{-1}.\end{aligned} \tag{3.17}$$

The diagonal elements of the covariance matrix are the variances of the estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$. The off-diagonal elements are the covariances between these estimators. The meaning of each element of the variance-covariance matrix can be seen from the following expression:

$$\mathbf{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X'X})^{-1} = \begin{bmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{cov}(\hat{\beta}_0, \hat{\beta}_2) & \cdots & \text{cov}(\hat{\beta}_0, \hat{\beta}_k) \\ \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \cdots & \text{cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \text{cov}(\hat{\beta}_0, \hat{\beta}_2) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{var}(\hat{\beta}_2) & \cdots & \text{cov}(\hat{\beta}_2, \hat{\beta}_k) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\hat{\beta}_0, \hat{\beta}_k) & \text{cov}(\hat{\beta}_1, \hat{\beta}_k) & \text{cov}(\hat{\beta}_2, \hat{\beta}_k) & \cdots & \text{var}(\hat{\beta}_k) \end{bmatrix}.$$

In the remainder of this chapter, we will use the symbols $\sigma^2_{\hat{\beta}_0}, \sigma^2_{\hat{\beta}_1}, \dots, \sigma^2_{\hat{\beta}_k}$ to refer to the variances $\text{var}(\hat{\beta}_0), \text{var}(\hat{\beta}_1), \dots, \text{var}(\hat{\beta}_k)$ of the estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$. The symbols $\sigma_{\hat{\beta}_0}, \sigma_{\hat{\beta}_1}, \dots, \sigma_{\hat{\beta}_k}$ will therefore represent the standard deviations of the least squares estimators. These standard deviations are the (positive) square roots of the diagonal elements of $\mathbf{var}(\hat{\boldsymbol{\beta}})$. For the covariances $\text{cov}(\hat{\beta}_0, \hat{\beta}_1), \text{cov}(\hat{\beta}_0, \hat{\beta}_2), \dots, \text{cov}(\hat{\beta}_{k-1}, \hat{\beta}_k)$, we use a similar alternative notation: $\sigma_{\hat{\beta}_0,\hat{\beta}_1}, \sigma_{\hat{\beta}_0,\hat{\beta}_2}, \dots,$ $\sigma_{\hat{\beta}_{k-1},\hat{\beta}_k}$. Consequently, we can write the variance-covariance matrix of the least squares estimator $\hat{\boldsymbol{\beta}}$ as follows as well:

$$\mathbf{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X'X})^{-1} = \begin{bmatrix} \sigma^2_{\hat{\beta}_0} & \sigma_{\hat{\beta}_0,\hat{\beta}_1} & \sigma_{\hat{\beta}_0,\hat{\beta}_2} & \cdots & \sigma_{\hat{\beta}_0,\hat{\beta}_k} \\ \sigma_{\hat{\beta}_0,\hat{\beta}_1} & \sigma^2_{\hat{\beta}_1} & \sigma_{\hat{\beta}_1,\hat{\beta}_2} & \cdots & \sigma_{\hat{\beta}_1,\hat{\beta}_k} \\ \sigma_{\hat{\beta}_0,\hat{\beta}_2} & \sigma_{\hat{\beta}_1,\hat{\beta}_2} & \sigma^2_{\hat{\beta}_2} & \cdots & \sigma_{\hat{\beta}_2,\hat{\beta}_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{\hat{\beta}_0,\hat{\beta}_k} & \sigma_{\hat{\beta}_1,\hat{\beta}_k} & \sigma_{\hat{\beta}_2,\hat{\beta}_k} & \cdots & \sigma^2_{\hat{\beta}_k} \end{bmatrix}.$$

Because the variances $\sigma^2_{\hat{\beta}_0}, \sigma^2_{\hat{\beta}_1}, \dots, \sigma^2_{\hat{\beta}_k}$ as well as the standard deviations $\sigma_{\hat{\beta}_0}, \sigma_{\hat{\beta}_1}, \dots, \sigma_{\hat{\beta}_k}$ depend on the unknown variance or standard deviation of the error terms, σ^2 or σ, they are unknown too, and they will have to be estimated in order to perform hypothesis tests and construct confidence intervals. This will require an estimate for σ^2 or σ.

Ideally, the variance-covariance matrix $\mathbf{var}(\hat{\boldsymbol{\beta}})$ is a diagonal matrix with small diagonal elements. Small diagonal elements are desirable because they mean that the regression analysis produces precise estimates for the model parameters. A diagonal variance-covariance matrix is desirable because it means that all covariances $\text{cov}(\hat{\beta}_0, \hat{\beta}_1), \text{cov}(\hat{\beta}_0, \hat{\beta}_2), \ldots, \text{cov}(\hat{\beta}_{k-1}, \hat{\beta}_k)$ or $\sigma_{\hat{\beta}_0,\hat{\beta}_1}, \sigma_{\hat{\beta}_0,\hat{\beta}_2}, \ldots, \sigma_{\hat{\beta}_{k-1},\hat{\beta}_k}$ are zero. In that case, the estimates, hypothesis tests and confidence intervals for the individual model parameters are independent. A statistical statement concerning one model parameter, say β_i, then is independent of any statistical statement about another model parameter, say β_j. For the variance-covariance matrix $\mathbf{var}(\hat{\beta})$ to be diagonal, the columns of the model matrix $\mathbf{X}$ need to be orthogonal. When all columns of $\mathbf{X}$ are pairwise orthogonal, then $\mathbf{X}'\mathbf{X}$ is diagonal, and, as a result, $(\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{var}(\hat{\beta})$ as well.

When collecting observational data, we can only record the values of the explanatory variables, without intervening. In that case, we cannot influence the precision of the least squares estimator. However, when collecting experimental data, we often have the possibility to pick the values of the explanatory variables ourselves. In that case, it is often wise to ensure that the columns of the model matrix are orthogonal. The orthogonality or near-orthogonality of the columns of $\mathbf{X}$ is therefore of major importance in the design of experiments.

Example 3.3.1. In Example 3.2.2, we estimated a multiple regression model for the total surface tension of polypropylene based on $n = 100$ data points. The explanatory variables involved were the flow rate, the power and the reaction time. The model matrix for the polypropylene experiment is a 100×4 matrix, namely

$$\mathbf{X} = \begin{bmatrix} 1 & 1000 & 2000 & 15 \\ 1 & 2000 & 500 & 2 \\ 1 & 1000 & 2000 & 2 \\ & & \vdots & \\ 1 & 2000 & 500 & 2 \\ 1 & 2000 & 500 & 15 \\ 1 & 2000 & 2000 & 15 \end{bmatrix}.$$

Since this matrix contains large values for the explanatory variables, so does the matrix product $\mathbf{X}'\mathbf{X}$:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 100 & 151500 & 118000 & 851 \\ 151500 & 248750000 & 177000000 & 1295000 \\ 118000 & 177000000 & 182500000 & 1033000 \\ 851 & 1295000 & 1033000 & 10459 \end{bmatrix}.$$

Consequently,

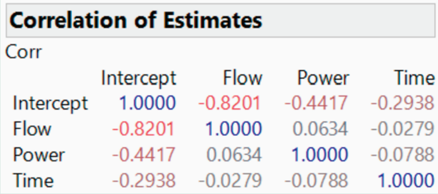
$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.185811157 & -8.0803 \times 10^{-5} & -2.90903 \times 10^{-5} & -0.00224067 \\ -8.08030 \times 10^{-5} & 5.22462 \times 10^{-8} & 2.21293 \times 10^{-9} & -1.12966 \times 10^{-7} \\ -2.90903 \times 10^{-5} & 2.21293 \times 10^{-9} & 2.33486 \times 10^{-8} & -2.13117 \times 10^{-7} \\ -0.00224067 & -1.12966 \times 10^{-7} & -2.13117 \times 10^{-7} & 0.00031296 \end{bmatrix}.$$

Many of the values in this matrix are small, due to the fact that $\mathbf{X}'\mathbf{X}$ contains large values. Note that both $\mathbf{X}'\mathbf{X}$ and $(\mathbf{X}'\mathbf{X})^{-1}$ are symmetric 4×4 matrices. These matrices have four rows and four columns because the estimated model involves four parameters. The first element of $\mathbf{X}'\mathbf{X}$ equals 100, the number of observations in the data set. This is not a coincidence. This happens whenever the model contains an intercept and the intercept is the first parameter in the model. The diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$ are all positive. This is logical because the diagonal elements of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ are variances, and, therefore, they should be positive. Since σ^2 is positive (as it is a variance), this is only possible when the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$ are positive too. Some of the off-diagonal elements of that matrix are positive, while others are negative. As a result of that, some pairs of parameter estimates will have a positive covariance, while other pairs will have a negative covariance.

At present, we do not know yet how to estimate σ^2 in the context of a multiple linear regression analysis. Therefore, we cannot yet estimate the variances $\text{var}(\hat{\beta}_0)$, $\text{var}(\hat{\beta}_1), \dots, \text{var}(\hat{\beta}_k)$ of the estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$. For the same reason, we cannot yet estimate the covariances between pairs of estimators. However, we can calculate the correlations between pairs of parameter estimators. This is because σ^2 is an irrelevant constant for computing the correlations. For instance, the correlation between the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ is

$$\begin{aligned}\rho(\hat{\beta}_1, \hat{\beta}_2) &= \frac{\text{cov}(\hat{\beta}_1, \hat{\beta}_2)}{\sqrt{\text{var}(\hat{\beta}_1)}\sqrt{\text{var}(\hat{\beta}_2)}} = \frac{2.21293 \times 10^{-9}}{\sqrt{5.22462 \times 10^{-8}}\sqrt{2.33486 \times 10^{-8}}} \\ &= 0.0634.\end{aligned}$$

This slightly positive correlation implies that, with this data set, β_1 and β_2 can be estimated almost independently. This is typical for experimental data sets that have been collected using a structured plan, called an experimental design, or a design for the experiment. An extensive discussion of experimental design can be found in the book “Optimal Design of Experiments: A Case Study Approach” by Peter Goos and Bradley Jones.



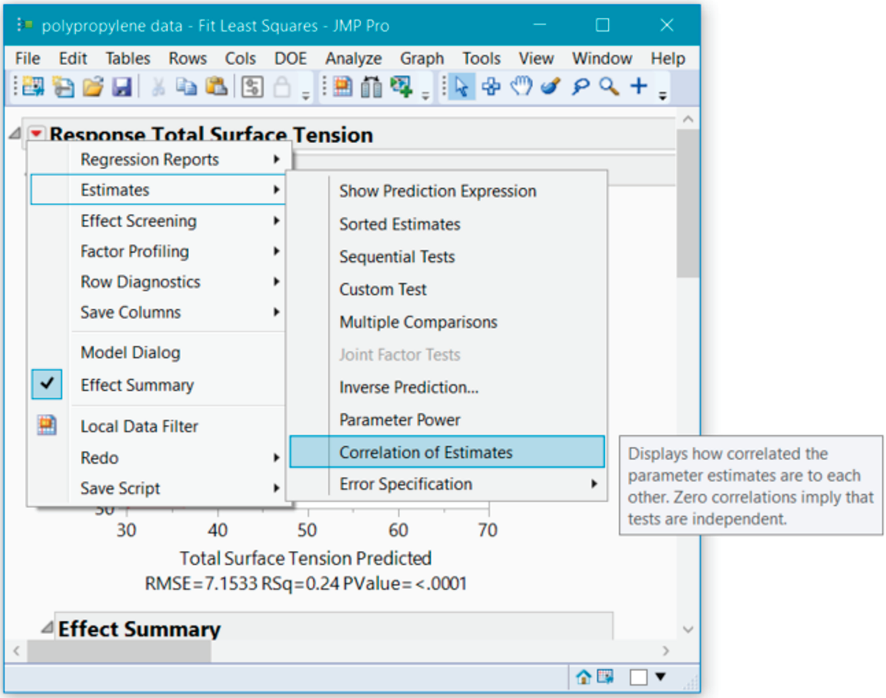
Correlation of Estimates

Corr

	Intercept	Flow	Power	Time
Intercept	1.0000	-0.8201	-0.4417	-0.2938
Flow	-0.8201	1.0000	0.0634	-0.0279
Power	-0.4417	0.0634	1.0000	-0.0788
Time	-0.2938	-0.0279	-0.0788	1.0000

Figure 3.14. Correlation matrix for the parameter estimators $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ in Example 3.3.1.

JMP does not allow you to calculate $\mathbf{X}'\mathbf{X}$ and $(\mathbf{X}'\mathbf{X})^{-1}$ in a few simple mouse clicks. It does, however, offer an option to obtain a correlation matrix for the parameter estimators. That matrix is shown in Figure 3.14 for the model considered here for the polypropylene data. It contains the value 0.0634 twice, because the correlation matrix is also symmetric. To produce the correlation matrix of the parameter estimators in JMP, we need to open the hotspot (red triangle) menu at the top of the regression output, next to the output's title "`Response Total Surface Tension`". One of the options in that menu is "`Estimates`". When selecting that option, we can indicate that we want the "`Correlation of Estimates`" to appear in the output. This is shown in Figure 3.15.



Displays how correlated the parameter estimates are to each other. Zero correlations imply that tests are independent.

RMSE=7.1533 RSq=0.24 PValue=<.0001

Figure 3.15. "`Estimates`" option offering the possibility to obtain the correlation matrix for the parameter estimators $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ in JMP.

In Example 3.3.1, the matrix product $\mathbf{X}'\mathbf{X}$ is characterized by large values. It is also far from being a diagonal matrix. For this reason, inverting the matrix requires appropriate software. Several decades ago, such software was unavailable and computations had to be done by hand, or using rudimentary software that lacked numerical precision. Inverting non-diagonal matrices involving large values was then cumbersome, and often imprecise. For this reason, it was common to rescale the explanatory variables, so as to avoid large values in $\mathbf{X}$ and $\mathbf{X}'\mathbf{X}$. For some experimental data, rescaling the experimental variables results in a diagonal $\mathbf{X}'\mathbf{X}$ matix, which is easy to invert. The possibility to rescale explanatory variables was already discussed in Example 3.2.3.

Example 3.3.2. In Example 3.2.4, we estimated a multiple regression model for the ice cream consumption in $n = 30$ successive months. The explanatory variables involved were the price index, the monthly average temperature and the monthly average income. For the estimated model in that example, the model matrix is a 30×4 matrix, namely

$$\mathbf{X} = \begin{bmatrix} 1 & 0.270 & 41 & 78 \\ 1 & 0.282 & 56 & 79 \\ 1 & 0.277 & 63 & 81 \\ & & \vdots & \\ 1 & 0.265 & 52 & 96 \\ 1 & 0.268 & 64 & 91 \\ 1 & 0.260 & 71 & 90 \end{bmatrix}.$$

Therefore,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 30 & 8.259 & 1473 & 2538 \\ 8.259 & 2.275721 & 405.087 & 698.549 \\ 1473 & 405.087 & 80145 & 123650 \\ 2538 & 698.549 & 123650 & 215846 \end{bmatrix},$$

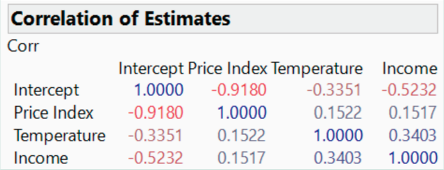
and

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 53.821472108 & -152.5598806 & -0.029740601 & -0.122082035 \\ -152.5598806 & 513.14146486 & 0.0417015490 & 0.1092729225 \\ -0.029740601 & 0.0417015490 & 0.0001463254 & 0.0001309171 \\ -0.122082035 & 0.1092729225 & 0.0001309171 & 0.0010114795 \end{bmatrix}.$$

Based on this result, we can again calculate the correlations between pairs of parameter estimators. For instance, the correlation between the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ is

$$\rho(\hat{\beta}_1, \hat{\beta}_2) = \frac{\text{cov}(\hat{\beta}_1, \hat{\beta}_2)}{\sqrt{\text{var}(\hat{\beta}_1)}\sqrt{\text{var}(\hat{\beta}_2)}} = \frac{0.0417015490}{\sqrt{513.14146486}\sqrt{0.0001463254}} = 0.1522.$$

The full correlation matrix produced by JMP for the parameter estimators $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ is shown in Figure 3.16. It contains the value 0.1522 twice. Overall, the correlations in Figure 3.16 are larger than those in Figure 3.14. This is due to the fact that the ice cream data is observational, while the polypropylene data is experimental and was collected using a structured plan for data collection, an experimental design.



Correlation of Estimates

Corr

	Intercept	Price Index	Temperature	Income
Intercept	1.0000	-0.9180	-0.3351	-0.5232
Price Index	-0.9180	1.0000	0.1522	0.1517
Temperature	-0.3351	0.1522	1.0000	0.3403
Income	-0.5232	0.1517	0.3403	1.0000

Figure 3.16. Correlation matrix for the parameter estimators $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ in Example 3.3.2.

3.3.4 Gauss-Markov theorem

Ideally, estimators are unbiased and efficient or precise. We already know that, if the multiple linear regression model is correctly specified, the least squares estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ are unbiased estimators of the model parameters $\beta_0, \beta_1, \dots, \beta_k$. Therefore, it is natural to wonder whether the estimators are also efficient.

In Section 2.3.5, we demonstrated that the least squares estimators of the intercept and the slope in simple linear regression were indeed efficient. More specifically, the least squares estimators we derived in the context of simple linear regression were best linear unbiased estimators. This followed from the Gauss-Markov theorem. It turns out that the Gauss-Markov theorem is valid for multiple linear regression as well, so that the least squares estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ are the best linear unbiased estimators of $\beta_0, \beta_1, \dots, \beta_k$. In other words, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$. We prove the Gauss-Markov theorem for multiple linear regression in Section 3.7.3.

Theorem 3.3.1 (Gauss-Markov theorem). *The least squares estimator $\hat{\boldsymbol{\beta}}$ has the smallest variance-covariance matrix of all unbiased estimators for $\boldsymbol{\beta}$ that are linear functions of the vector of responses $\boldsymbol{Y}$.*

Note that there exist linear estimators with a smaller variance than the least squares estimator, but these are biased estimators. So-called ridge regression produces such an estimator. Note also that Assumption 4, concerning the normality of the errror terms, is not required for the Gauss-Markov theorem to be valid. This means that the least squares estimators are the best linear unbiased estimators even when the error terms are not normally distributed.

3.3.5 Normal distribution

If Assumption 4 holds and the error terms $U_1, U_2, \dots, U_n$ are all normally distributed, then the responses $Y_1, Y_2, \dots, Y_n$ are also normally distributed, and the response vector $\boldsymbol{Y}$ is a multivariate normal random vector. Now, the least squares estimator $\hat{\boldsymbol{\beta}}$ involves $k+1$ linear combinations of the normally distributed responses $Y_1, Y_2, \dots, Y_n$, one for each individual estimator $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$. Each of these individual parameter estimators is therefore also normally distributed, and the vector containing all these individual estimators, $\hat{\boldsymbol{\beta}}$, then follows a multivariate normal distribution. Mathematically, we state this as follows for an individual estimator $\hat{\beta}_i$ and for the vector $\hat{\boldsymbol{\beta}}$:

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2_{\hat{\beta}_i}),$$

and

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right).$$

When deriving test statistics in the context of multiple linear regression, we will often rely on the normal distributions of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ and $\hat{\boldsymbol{\beta}}$.

In some cases, it will be unrealistic to assume that the errors terms are normally distributed. This is, however, not necessarily problematic. As a matter of fact, for large sample sizes, we can rely on the central limit theorem to justify our use of the confidence intervals and the hypothesis tests derived under the assumption of normal error terms. The central limit theorem states that any linear combination of a sufficiently large numer of independent random variables is approximately normally distributed. We should, however, be cautious using these intervals and tests when we encounter a clear deviation from normality in a small data sample.

3.4 Estimating σ^2

In order to perform significance tests, construct confidence intervals for estimated parameters and assess the quality of the model, we first need to estimate σ^2, the common variance of the error terms $U_1, U_2, \ldots, U_n$. This is because significance tests and confidence intervals are all based on one or more elements of the variance-covariance matrix of the least squares estimator in Equation (3.17), namely $\mathbf{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Because that variance-covariance matrix is proportional to σ^2, the variance of all parameter estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ and all their pairwise covariances depend on σ^2.

It can be shown that

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{U}_i^2}{n-(k+1)} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-(k+1)}$$

is an unbiased estimator of σ^2, if the regression model is correctly specified. This estimator generalizes the one for simple linear regression in Section 2.4 and is similar in structure to the estimator for a sample variance (see the book "Statistics with JMP: Graphs, Descriptive Statistics and Probability"). The numerator of the estimator is the sum of the squared residuals, which is named the **residual sum of squares** or **sum of squared errors**. We denote it by SSE. Therefore, the estimator of the variance σ^2 in multiple linear regression is often written as

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-(k+1)} = \text{MSE},$$

where MSE is short for **mean squared error**.

The division by $n-(k+1)$ in this estimator can be explained by the fact that we start from n observations (n units of information) and, in order to be able to calculate the predicted responses $\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n$, we need to estimate $k+1$ parameters, $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ (which costs us $k+1$ units of information). In the technical jargon, $n-(k+1)$ is the number of degrees of freedom associated with the sum of squared errors. That number is often called the **residual degrees of freedom**.

In the event that our model contains as many as $n-1$ explanatory variables, then $n-(k+1) = 0$ and there are no residual degrees of freedom. We then say that there are no degrees of freedom left to estimate σ^2. The MSE then does not exist. A model involving $n-1$ explanatory variables is called **saturated**. We discuss some of the charcteristics of saturated models in Section 3.6.2.

After we have collected the responses and estimated the multiple linear regression model, we can calculate the estimate s^2 of σ^2:

$$s^2 = \frac{\sum_{i=1}^{n} u_i^2}{n-(k+1)} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-(k+1)} = \frac{\text{sse}}{n-(k+1)} = \text{mse}\,.$$

Here, we again use lowercase letters for the sum of squared errors and the mean squared error because their values can be calculated as soon as we have collected the response data $y_1, y_2, \dots, y_n$.

When using matrix algebra, it is not very difficult to prove that the mean squared error is an unbiased estimator of σ^2. We provide the proof in Section 3.7.4.

Now that we are able to estimate σ^2, we can also estimate the entire variance-covariance matrix $\mathbf{var}(\hat{\boldsymbol{\beta}})$. The estimated variance-covariance matrix is

$$s^2(\mathbf{X}'\mathbf{X})^{-1}.$$

The diagonal elements of that matrix are the estimated variances of the least squares estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, , \hat{\beta}_k$. We denote these estimated variances by $s^2_{\hat{\beta}_0}, s^2_{\hat{\beta}_1}, s^2_{\hat{\beta}_2}, \dots, s^2_{\hat{\beta}_k}$, respectively. The off-diagonal elements are the estimated covariances between the parameter estimators. We denote them by $s_{\hat{\beta}_0,\hat{\beta}_1}, s_{\hat{\beta}_0,\hat{\beta}_2}, \dots, s_{\hat{\beta}_{k-1},\hat{\beta}_k}$. The full estimated variance-covariance matrix of the least squares estimator $\hat{\boldsymbol{\beta}}$ is therefore written as

$$s^2(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} s^2_{\hat{\beta}_0} & s_{\hat{\beta}_0,\hat{\beta}_1} & s_{\hat{\beta}_0,\hat{\beta}_2} & \cdots & s_{\hat{\beta}_0,\hat{\beta}_k} \\ s_{\hat{\beta}_0,\hat{\beta}_1} & s^2_{\hat{\beta}_1} & s_{\hat{\beta}_1,\hat{\beta}_2} & \cdots & s_{\hat{\beta}_1,\hat{\beta}_k} \\ s_{\hat{\beta}_0,\hat{\beta}_2} & s_{\hat{\beta}_1,\hat{\beta}_2} & s^2_{\hat{\beta}_2} & \cdots & s_{\hat{\beta}_2,\hat{\beta}_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{\hat{\beta}_0,\hat{\beta}_k} & s_{\hat{\beta}_1,\hat{\beta}_k} & s_{\hat{\beta}_2,\hat{\beta}_k} & \cdots & s^2_{\hat{\beta}_k} \end{bmatrix}.$$

The square roots of the estimated variances, $s_{\hat{\beta}_0}, s_{\hat{\beta}_1}, s_{\hat{\beta}_2}, \dots, s_{\hat{\beta}_k}$, are the estimated standard deviations or **standard errors** of the least squares estimators. They are reported in every output produced by software packages when carrying out multiple linear regression analyses.

Example 3.4.1. In Example 3.3.1, we found that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.185811157 & -8.08030 \times 10^{-5} & -2.90903 \times 10^{-5} & -0.00224067 \\ -8.08030 \times 10^{-5} & 5.22462 \times 10^{-8} & 2.21293 \times 10^{-9} & -1.12966 \times 10^{-7} \\ -2.90903 \times 10^{-5} & 2.21293 \times 10^{-9} & 2.33486 \times 10^{-8} & -2.13117 \times 10^{-7} \\ -0.00224067 & -1.12966 \times 10^{-7} & -2.13117 \times 10^{-7} & 0.00031296 \end{bmatrix}$$

for the multiple linear regression model with the total surface tension of polypropylene as a response and flow rate, power and reaction time as explanatory variables. The sum of the squared residuals is

$$\begin{aligned} \text{sse} &= \sum_{i=1}^{n} u_i^2 = (12.252688575)^2 + (-1.525487109)^2 + \cdots + (-7.466189455)^2 \\ &= 4912.33, \end{aligned}$$

so that the mean squared error equals

$$s^2 = \text{mse} = \frac{\text{sse}}{n-(k+1)} = \frac{4912.33}{100-(3+1)} = \frac{4912.33}{96} = 51.1701,$$

because $n = 100$ and $k = 3$. The estimated variance-covariance matrix thus is

$$s^2(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 9.507969529 & -0.004134695 & -0.001488551 & -0.114655239 \\ -0.004134695 & 2.67344 \times 10^{-6} & 1.13236 \times 10^{-7} & -5.78045 \times 10^{-6} \\ -0.001488551 & 1.13236 \times 10^{-7} & 1.19475 \times 10^{-6} & -1.09052 \times 10^{-5} \\ -0.114655239 & -5.78045 \times 10^{-6} & -1.09052 \times 10^{-5} & 0.016014196 \end{bmatrix}.$$

The diagonal elements of this matrix are the estimated variances $s^2_{\hat{\beta}_0}$, $s^2_{\hat{\beta}_1}$, $s^2_{\hat{\beta}_2}$ and $s^2_{\hat{\beta}_3}$ of the four least squares estimators $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$. Their square roots are the estimated standard deviations or standard errors of the parameter estimators:

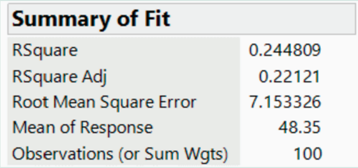
$$\begin{aligned} s_{\hat{\beta}_0} &= \sqrt{9.507969529} = 3.0835, \\ s_{\hat{\beta}_1} &= \sqrt{2.67344 \times 10^{-6}} = 0.001635, \\ s_{\hat{\beta}_2} &= \sqrt{1.19475 \times 10^{-6}} = 0.001093, \\ s_{\hat{\beta}_3} &= \sqrt{0.016014196} = 0.126547. \end{aligned}$$

Figure 3.2 shows that these values can be retrieved in the column labeled "`Std Error`" in the "`Parameter Estimates`" table in JMP.

The square root of the mean squared error is called the root mean squared error and abbreviated as rmse. It is a point estimate s of the standard deviation of the error terms, σ:

$$s = \text{rmse} = \sqrt{\text{mse}} = \sqrt{51.1701} = 7.1533.$$

As shown in Figure 3.17, it is one of the first statistics reported by JMP for any multiple linear regression analysis, in the table entitled "Summary of Fit".



Summary of Fit	
RSquare	0.244809
RSquare Adj	0.22121
Root Mean Square Error	7.153326
Mean of Response	48.35
Observations (or Sum Wgts)	100

Figure 3.17. "Summary of Fit" statistics for the multiple linear regression model for the polypropylene data in Example 3.4.1.

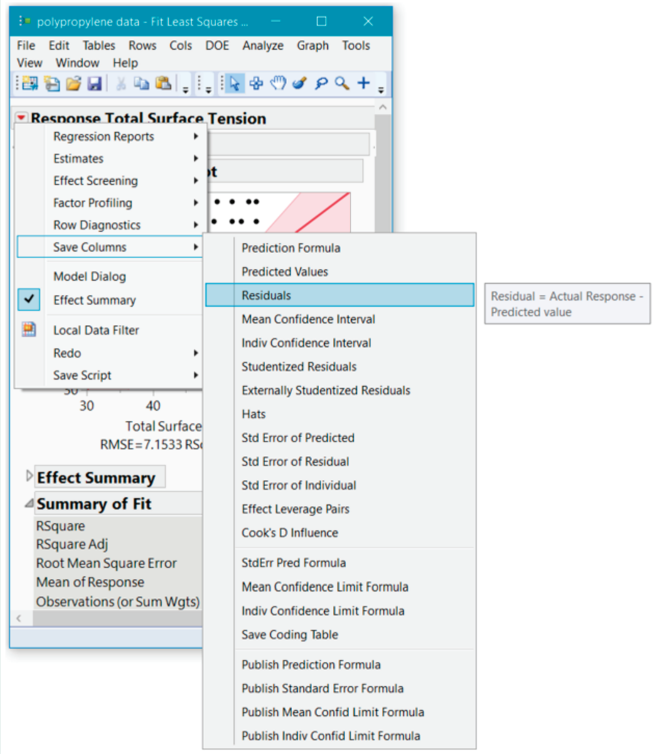


Figure 3.18. The options to add columns with the residuals u_i and the predicted response values $\hat{y}_i$ to the data table.