Multiple linear regression

Multiple linear regression aims at studying the dependence of a quantitative variable Y on a set of m quantitative regressors, X_1, …, X_m, using a linear model:

$$Y = f(X_1, \ldots, X_m) + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon$$

Once again, linearity is referred to parameters.

As for simple linear regression, X_1, …X_m are deterministic variables, i.e., considered without error; the following additional hypotheses can be formulated:

$$E(\varepsilon) = 0 \quad \Rightarrow \quad E(Y|X_1, \ldots, X_m) = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m$$

$$V(\varepsilon) = \sigma^2 \quad \Rightarrow \quad V(Y|X_1, \ldots, X_m) = \sigma^2$$ (homoscedasticity)

$$\varepsilon \sim N(0, \sigma^2)$$

$$\mathrm{Cov}\,(\varepsilon_i, \varepsilon_j) = 0 \qquad \forall\, i \neq j$$ (uncorrelation)

Given n observations Y_i, corresponding to n sets of m values for the regressors:

$$(Y_i, X_{i1}, \ldots, X_{im}) \qquad i = 1, \ldots, n$$

The following equation can be written for the i-th observation:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_m X_{im} + \varepsilon_i$$

The following system of equations is thus obtained when all the n observations are considered:

$$\begin{cases} Y_1 = \beta_0 + \beta_1 X_{11} + \ldots + \beta_m X_{1m} + \varepsilon_1 \\ \ldots \\ Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_m X_{im} + \varepsilon_i \\ \ldots \\ Y_n = \beta_0 + \beta_1 X_{n1} + \ldots + \beta_m X_{nm} + \varepsilon_n \end{cases}$$

If the following vectors/matrices are introduced:

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ \ldots. \\ Y_i \\ \ldots. \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \ldots & x_{1m} \\ \ldots & & & \\ 1 & x_{i1} & x_{i2} \ldots & x_{im} \\ \ldots & & & \\ 1 & x_{n1} & x_{n2} \ldots & x_{nm} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \ldots. \\ \beta_i \\ \ldots. \\ \beta_m \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \ldots. \\ \varepsilon_i \\ \ldots. \\ \varepsilon_n \end{bmatrix}$$

The system can be written as a matricial equation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Previous assumptions can thus be expressed in matricial notation:

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \Rightarrow \quad \mathrm{E}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

$$\mathrm{V}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)$$

In this case the $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T$ notation indicates the outer product between vector $\boldsymbol{\varepsilon}$ and its transpose, i.e., a product based on the following rule:

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^{\mathrm{T}} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} = \begin{bmatrix} u_1v_1 & u_1v_2 & u_1v_3 \\ u_2v_1 & u_2v_2 & u_2v_3 \\ u_3v_1 & u_3v_2 & u_3v_3 \\ u_4v_1 & u_4v_2 & u_4v_3 \end{bmatrix}$$

Consequently:

$$\boldsymbol{\varepsilon\varepsilon}^{\mathsf{T}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{bmatrix} = \begin{bmatrix} \varepsilon_1\,\varepsilon_1 & \varepsilon_1\,\varepsilon_2 & \cdots & \varepsilon_1\,\varepsilon_n \\ \varepsilon_2\,\varepsilon_1 & \varepsilon_2\,\varepsilon_2 & \cdots & \varepsilon_2\,\varepsilon_n \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_n\,\varepsilon_1 & \varepsilon_n\,\varepsilon_2 & \cdots & \varepsilon_n\varepsilon_n \end{bmatrix}$$

Notably, the expectation of diagonal terms in the matrix correspond to variances of ε_i, that are all equal to σ^2, whereas all the other terms correspond to the covariances of ε_i with ε_j, that are all equal to 0, thus:

$$\mathrm{V}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon\varepsilon}^T) = \sigma^2 \mathbf{I}_n$$

where $\mathbf{I}_n$ is the identity matrix of n-th order: $I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$

$$\boldsymbol{\varepsilon} \sim \mathrm{NMV}(\mathbf{0}, \sigma^2\mathbf{I}_n) \quad \Rightarrow \quad \mathbf{y}|\mathbf{X} \sim \mathrm{NMV}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$$

where NMV stands for Normal MultiVariate distribution, whose variance is:

$$\mathrm{V}(\mathbf{y}|\mathbf{X}) = \mathrm{E}[(\mathbf{y} - \mathrm{E}(\mathbf{y}|\mathbf{X}))(\mathbf{y} - \mathrm{E}(\mathbf{y}|\mathbf{X}))^T] = \mathrm{E}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T] = \mathrm{E}(\boldsymbol{\varepsilon}\ \boldsymbol{\varepsilon}^T) = \sigma^2\mathbf{I}_n$$

The Ordinary Least Squares (OLS) method can be employed to find vector **b**, i.e., the estimator of the vector of unknown parameters **β**, thus the vector of response values **ŷ** predicted by the model can be expressed as:

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

or, equivalently:

$$\hat{\mathrm{y}}_{\mathrm{i}} = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots\ldots + b_m x_{im}$$ for *i = 1,2,..., n*

The vector corresponding to the difference between experimental and predicted values, i.e., the sampling residual, can be expressed as:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b}$$

or, equivalently:

$$e_i = \left(y - \hat{y}_i\right) = (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots\ldots - b_m x_{im})$$ for *i = 1,2,..., n*

According to the OLS method, vector **b** can be obtained by minimizing the sum of squared residuals:

$$\mathbf{S}(\mathbf{b}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^{\mathrm{T}}\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}) =$$

$$= \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{b} + \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}$$

The vector can be found by solving the following equation:

$$\frac{\partial \mathbf{S}(\mathbf{b})}{\partial \mathbf{b}^{\mathrm{T}}} = -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} = 0$$

$$\Downarrow$$

$$\mathbf{X}^{T}\mathbf{X}\,\mathbf{b} = \mathbf{X}^{T}\mathbf{y}$$

This matricial equation, expressed explicitly as:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{im} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^{2} & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{im} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} x_{im} & \sum_{i=1}^{n} x_{im}x_{i1} & \sum_{i=1}^{n} x_{im}x_{i2} & \cdots & \sum_{i=1}^{n} x_{im}^{2} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \cdots \\ b_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \cdots \\ \sum_{i=1}^{n} x_{im}y_i \end{bmatrix}$$

can be expressed as a system of equations, the first one being:

$$nb_0 + b_1 \sum_{i=1}^{n} x_{i1} + \ldots\ldots + b_m \sum_{i=1}^{n} x_{im} = \sum_{i=1}^{n} y_i$$

$$\Downarrow$$

$$n\frac{1}{n}b_0 + b_1 \frac{1}{n}\sum_{i=1}^{n} x_{i1} + \ldots\ldots + b_m \frac{1}{n}\sum_{i=1}^{n} x_{im} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\Downarrow$$

$$b_0 + b_1\bar{x}_1 + b_2\bar{x}_2 + \ldots\ldots + b_m\bar{x}_m = \bar{y}$$

If m = 1 the equation relating b_0 and b_1 for simple linear regression is easily obtained.

It can be easily demonstrated that the sum of residuals resulting from the OLS procedure is zero:

$$\hat{y}_i = b_0 + b_1 x_{i1} + \ldots\ldots + b_m x_{im}$$

⬇

$$\sum_{i=1}^{n} \hat{y}_i = n b_0 + b_1 \sum_{i=1}^{n} x_{i1} + \ldots\ldots + b_m \sum_{i=1}^{n} x_{im}$$

Since, as shown in the previous slide: $n b_0 + b_1 \sum_{i=1}^{n} x_{i1} + \ldots\ldots + b_m \sum_{i=1}^{n} x_{im} = \sum_{i=1}^{n} y_i$

the previous equation can be written as: $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$

thus: $\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$

Turning back to the matricial notation, the equation: $\mathbf{X}^T \mathbf{X}\, \mathbf{b} = \mathbf{X}^T \mathbf{y}$

can be used to calculate vector **b**: $\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y}$

It is worth noting that matrix **X^{T}X** has the following characteristics:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{im} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{im} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} x_{im} & \sum_{i=1}^{n} x_{im}x_{i1} & \sum_{i=1}^{n} x_{im}x_{i2} & \cdots & \sum_{i=1}^{n} x_{im}^2 \end{bmatrix}$$

1) is a square matrix (p × p, with p = m + 1)

2) is symmetric

3) terms on its main diagonal correspond to the sum of squares of values located in matrix **X** columns

4) terms outside the main diagonal are scalar products between couples of columns of matrix X.

Since: $\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

the equation $\hat{\mathbf{y}} = \mathbf{Xb}$ can be transformed into the following equation:

$$\hat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{Hy}$$

Matrix $\mathbf{X(X^{T}X)^{-1}X^{T}}$ is called *hat matrix*, since it is able to transform observed values of response y into calculated values ŷ, that are indicated as y with a symbol (^) similar to a hat.

The matricial equation $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$

can thus be written in the following form:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{y} = \mathbf{My}$$

where **I** is the identity matrix.

Both **H** and **M** are symmetric (n × n) matrices,

moreover: $\mathbf{HX} = \mathbf{X(X^{T}X)^{-1}X^{T}\,X} = \mathbf{X}$, thus $\mathbf{MX} = \mathbf{(I\text{-}H)\,X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$

Estimator of σ^2 and variance of vector b in multiple linear regression

By analogy with mth order univariate regression, the following equations can be written:

$$E(\mathbf{e}^{\mathrm{T}}\mathbf{e}) = \sigma^2(n\text{-}p) = \sigma^2(n\text{-}m\text{-}1)$$

thus the estimator of variance σ^2 can be calculated as:

$$\hat{\sigma}^2 = s_e^2 = \frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{n-m-1} = \frac{\sum_i e_i^2}{n-p}$$

This estimate can be exploited to calculate the variance of vector **b**:

$$\mathrm{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$$

As a consequence, the terms located on the main diagonal of matrix (X$^{\mathrm{T}}$X)$^{-1}$ determine variances of the model parameters, whereas those located outside the main diagonal determine their covariances.

Confidence interval for β_j in multiple linear regression

Since the vector of parameter is distributed as follows: $\mathbf{b} \sim \mathrm{NMV}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$

where NMV is the Normal MultiVariate distribution,

each parameter of the multiple linear regression model is also distributed according to the following normal distribution:

$$\mathrm{b}_j \sim \mathrm{N}(\beta_j, \sigma^2 \mathrm{c}_{jj})$$

where c_{jj} is the j-th element of the main diagonal of matrix **$(X^TX)^{-1}$**, thus the following relation is also true:

$$\mathrm{z} = \frac{b_j - \beta_j}{\sqrt{\sigma^2 c_{jj}}} \sim N(0,\ 1).$$

If σ^2 is estimated using s_e^2, the confidence interval at a 1-α level for β_j can be calculated as follows:

$$b_j \pm t_{n-m-1,1-\alpha/2} \sqrt{s_e^2 c_{jj}}$$

Multicollinearity

It is very important to point out that although regressors in a multiple regression model are indicated as «independent» variables, this does not imply necessarily that there is no correlation between them.

Multicollinearity is the term adopted to describe the presence of a significant correlation between regressors.

In particular, if columns of matrix **X** are considered, exact multicollinearity occurs when a linear combination of them corresponds to a null vector:

$$c_1\boldsymbol{X_1} + c_2\boldsymbol{X_2} + \ldots + c_m\boldsymbol{X_m} = \mathbf{0}$$

In a larger sense, a certain degree of multicollinearity exists if

$$c_1\boldsymbol{X_1} + c_2\boldsymbol{X_2} + \ldots + c_m\boldsymbol{X_m} = \mathbf{d}$$

where the following equation for the norm of vector **d** is valid: $\|\mathbf{d}\| < q\|\mathbf{c}\|$ with

$$\|\mathbf{c}\| = \sqrt{c_1^2 + c_2^2 + \ldots + c_m^2}$$

The number q represents the degree of multicollinearity.

The presence of multicollinearity may have some negative effects:

1. the variance of regression coefficients can be increased in such a way that single coefficients become non statistically significant, although the predictive capacity of the regression equation remains good

2. relative values and even the signs of coefficients can make their interpretation unreliable

3. values of single regression coefficients can radically change upon introduction or removal of one regressor (even their signs can change).

The presence of multicollinearity can be suspected when:

1. there is a high correlation between couples of predictors

2. values or signs of coefficients have no physical sense

3. regression coefficients are not statistically significant for predictors whose importance is well known

A quantity known as Variance Inflation Factor, VIF, can be used to evaluate the extent of multicollinearity between regressors. In particular, if the VIF is close or equal to 1 the corresponding regressor can be considered independent from other regressors.

Coefficient of determination in multiple linear regression

By analogy with simple linear regression, the following steps can be followed to calculate the coefficient of determination in multiple linear regression:

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = 1 - \frac{SS_{RES}}{SS_{TOT}}$$

In the case of simple linear regression it was shown that:

$$SS_{TOT} = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n\bar{y}^2$$

Thus, in the case of multiple linear regression, the following equation can be written:

$$SS_{TOT} = \mathbf{y}^T\mathbf{y} - n\bar{y}^2$$

Moreover:

$$SS_{RES} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^T\mathbf{e}$$

Starting from the previous equations, the coefficient of determination for multiple linear regression can be expressed with one of the following equations:

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SS}_{\mathrm{RES}}}{\mathrm{SS}_{\mathrm{TOT}}} \quad \Rightarrow \quad \mathrm{R}^2 = 1 - \frac{\mathbf{e}^T\mathbf{e}}{\mathbf{y}^T\mathbf{y} - n\bar{y}^2}$$

Coefficient R^2, that can change between 0 and 1, as usual, measures the fraction of variability of response Y due to its linear dependence on regressors.

Some possible problems for R^2 are:

1. it can be high even when the relation between response and regressors is not linear
2. it is increased at the increase of the number of regressors, thus it cannot be employed to compare models based on a different number of regressors.

As shown previously for other types of regressions, mean squares related to regression, MS$_{REG}$, and to residuals, MS$_{RES}$, can be exploited to evaluate the significance of the model.

In fact, the following statistic can be used for an F test:

$$F = \frac{MS_{REG}}{MS_{RES}} = \frac{SS_{REG}/m}{SS_{RES}/(n-m-1)}$$

with F distributed according to $F_{m,(n-m-1)}$.

If the realization of F is greater than the critical value for this distribution, at a fixed significance level:

1. the variability on Y explained by the model is significantly higher than the residual variability.

2. at least one of the m chosen regressors has a regression coefficient significantly different from 0.

When the realization of F is lower than the critical value, the model can be considered not adequate, thus no significant linear dependence exists between Y and the m regressors.

Criteria for the choice of regressors in multiple linear regression

The choice of regressors to be included in the model is a fundamental step in multiple linear regression.

Significant errors can be made by:

1) omitting relevant regressors

2) including irrelevant regressors.

Three methods can be exploited to choose regressors for multiple linear regression:

1) Forward selection

2) Backward elimination

3) Stepwise procedure

Forward selection

In this approach regressors are added sequentially to the model, starting from the simplest possible model:

$$Y = \beta_0 + \varepsilon$$

Supposing that X_1 is the first regressor added to the model:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

regression calculations are made, thus the estimator b_1 for model parameter β_1 is obtained, and the coefficient of determination R^2 is also calculated.

The same procedure can be repeated for other possible regressors that can be considered to be candidates as the X_1 regressor.

The regressor finally selected as X_1 is the one leading to the highest R^2 value.

Since: $$V(b_1) = \sigma^2 \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sigma^2}{S_{XX}}$$

A t-test to verify if b_1 is significantly different from zero can be performed, using this statistic:

$$t = \frac{b_1}{\sqrt{\frac{s_e^2}{dev(X_1)}}} = \frac{b_1}{\sqrt{\frac{\sum_{1=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)dev(X_1)}}} \sim t_{(n-2)}$$

with: $dev(X_1) = \sum_{i=1}^{n}(X_{1i} - \overline{X_1})^2$

If the realization of this statistic is lower than the critical value of the Student's t at a specific significance value the selected regressor is discarded and, since all other possible candidates lead to worse values for R^2, none of them deserves to be included in the model. An entirely new set of regressors has thus to be considered.

If the realization is greater than the critical value the regressor is confirmed as X_1 and the procedure goes on with the same approach, considering a set of regressors to be candidated as X_2, thus considering the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

The entire sequence of steps can be repeated for several other possible regressors, until a not significant t-test is obtained.

Backward elimination

In this approach single regressors are sequentially removed from an initial, complex model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon$$

Each time calculations are performed and the decrease in R^2 is evaluated.

The regressor whose removal leads to the smallest decrease in R^2 is subjected to the t-test shown before; if the test is not significant, the regressor is eliminated from the model and the procedure is repeated.

The procedure is stopped when a significant t-test is obtained.

Stepwise procedure

In this approach the first two steps are the same adopted for the forward selection, thus the following model is obtained:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Before proceeding to the consideration of a further regressor, the relevance of regressor X_1 is tested by the backward procedure explained before.

Generally speaking, in the stepwise procedure the elimination of each of the previously selected regressors is evaluated, after a new regressor is selected.

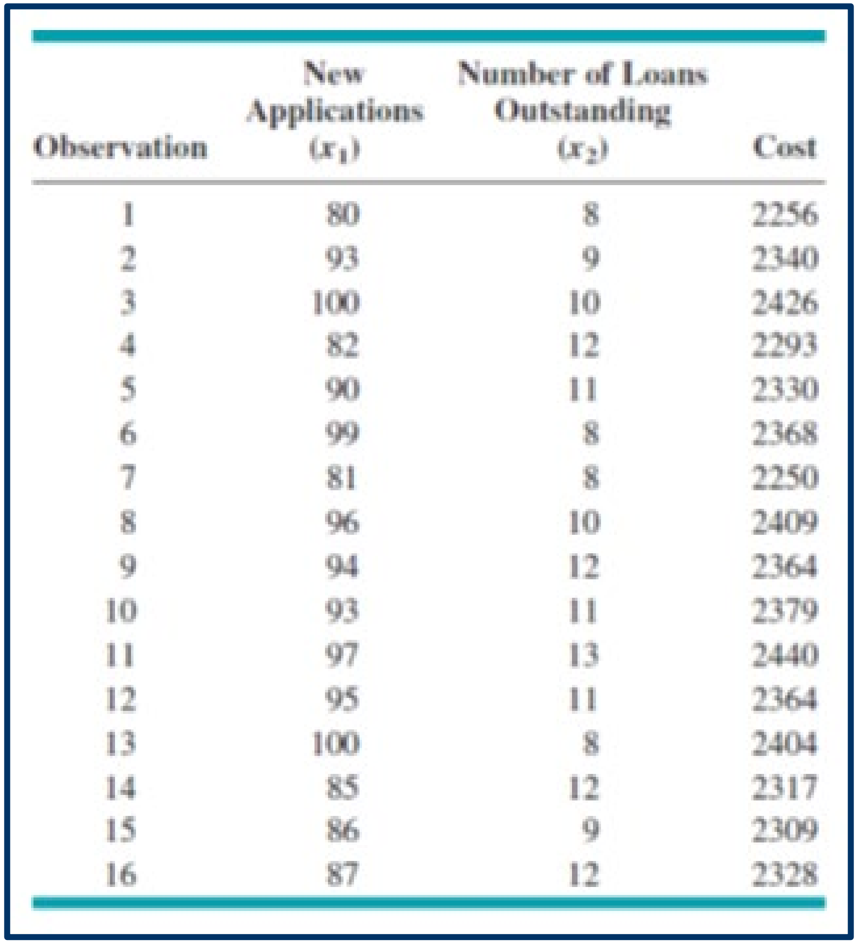
Based on this criterion, the decision to include a regressor is not irreversible (as it is in the forward selection approach): a previously selected regressor can be removed if the inclusion of further regressors makes its contribution to the explanation of response Y not significant.

It is worth noting that multiple linear regression models obtained according to the methods described so far do not represent the best models in an absolute sense.

They are the best models that can be obtained using an automatic approach.

A numerical example of multiple linear regression (two predictors)

The operating cost of a branch office of a finance company (response Y) was evaluated in 16 cases as a function of the numbers of new loan applications (x_1) and of existing (outstanding) applications (x_2):



Observation	New Applications (x_1)	Number of Loans Outstanding (x_2)	Cost
1	80	8	2256
2	93	9	2340
3	100	10	2426
4	82	12	2293
5	90	11	2330
6	99	8	2368
7	81	8	2250
8	96	10	2409
9	94	12	2364
10	93	11	2379
11	97	13	2440
12	95	11	2364
13	100	8	2404
14	85	12	2317
15	86	9	2309
16	87	12	2328

The following multiple linear regression model can be hypothesized to predict costs:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

In this case the **X** matrix and the **y** vector are:

$$\mathbf{X} = \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ 1 & 100 & 10 \\ 1 & 82 & 12 \\ 1 & 90 & 11 \\ 1 & 99 & 8 \\ 1 & 81 & 8 \\ 1 & 96 & 10 \\ 1 & 94 & 12 \\ 1 & 93 & 11 \\ 1 & 97 & 13 \\ 1 & 95 & 11 \\ 1 & 100 & 8 \\ 1 & 85 & 12 \\ 1 & 86 & 9 \\ 1 & 87 & 12 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 2256 \\ 2340 \\ 2426 \\ 2293 \\ 2330 \\ 2368 \\ 2250 \\ 2409 \\ 2364 \\ 2379 \\ 2440 \\ 2364 \\ 2404 \\ 2317 \\ 2309 \\ 2328 \end{bmatrix}$$

The vector of regression parameters **b** is given by the equation:

$$\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Matrices required for the calculation are:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 80 & 93 & \dots & 87 \\ 8 & 9 & \dots & 12 \end{bmatrix} \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ \vdots & \vdots & \vdots \\ 1 & 87 & 12 \end{bmatrix} = \begin{bmatrix} \boxed{16} & \boxed{1458} & \boxed{164} \\ 1458 & 133{,}560 & 14{,}946 \\ 164 & 14{,}946 & 1{,}726 \end{bmatrix}$$

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} = \boxed{\begin{bmatrix} 14.176004 & -0.129746 & -0.223453 \\ -0.129746 & 1.429184 \times 10^{-3} & -4.763947 \times 10^{-5} \\ -0.223453 & -4{,}763947 \times 10^{-5} & 2.222381 \times 10^{-2} \end{bmatrix}}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 80 & 93 & \dots & 87 \\ 8 & 9 & \dots & 12 \end{bmatrix} \begin{bmatrix} 2256 \\ 2340 \\ \vdots \\ 2328 \end{bmatrix} = \boxed{\begin{bmatrix} 37{,}577 \\ 3{,}429{,}550 \\ 385{,}562 \end{bmatrix}}$$

As a result, the following calculation is obtained for vector **b**:

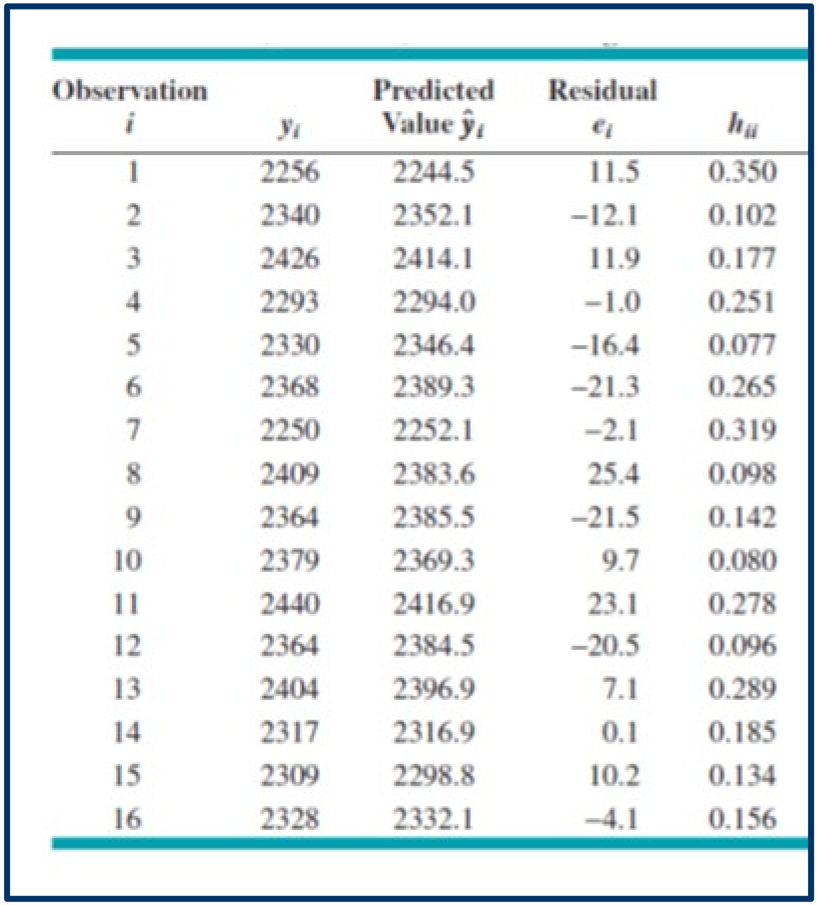
$$\mathbf{b} = \boxed{\begin{bmatrix} 14.176004 & -0.129746 & -0.223453 \\ -0.129746 & 1.429184 \times 10^{-3} & -4.763947 \times 10^{-5} \\ -0.223453 & -4{,}763947 \times 10^{-5} & 2.222381 \times 10^{-2} \end{bmatrix}} \boxed{\begin{bmatrix} 37{,}577 \\ 3{,}429{,}550 \\ 385{,}562 \end{bmatrix}} = \begin{bmatrix} 1566.07777 \\ 7.62129 \\ 8.58485 \end{bmatrix}$$

After approximating estimates of parameters to two decimal places, the following equation is obtained for the model:

$$\hat{y} = 1566.08 + 7.62\, x_1 + 8.58\, x_2$$

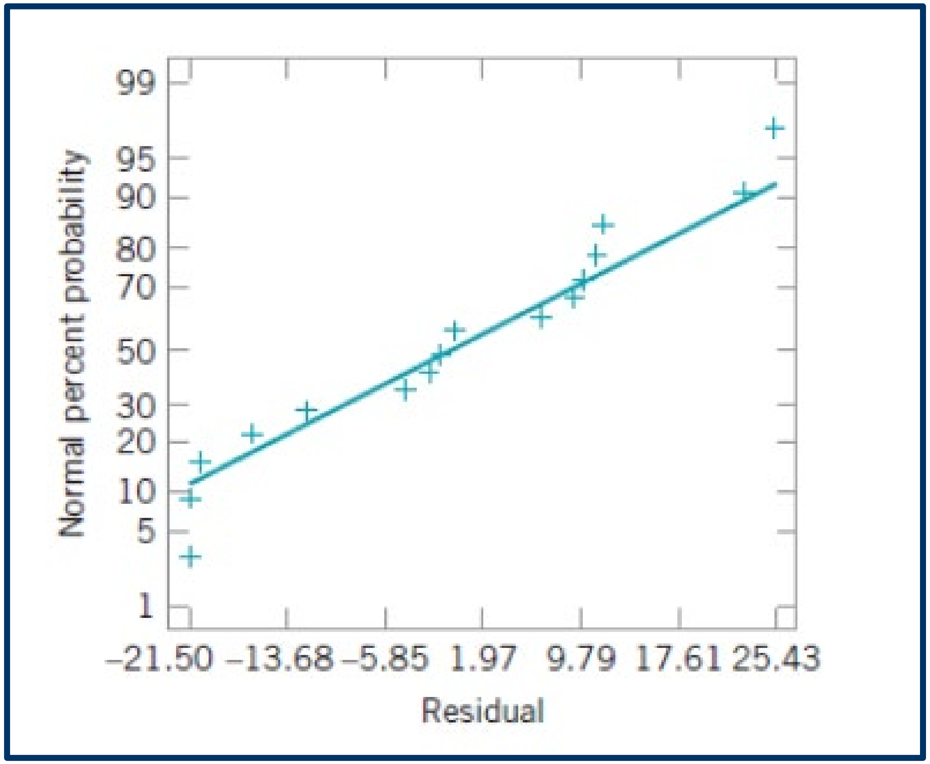
Notably, the contribution due to the two regressors is not very different.

A summary of experimental and predicted responses, residuals and diagonal elements of matrix H is reported in the table on the right:

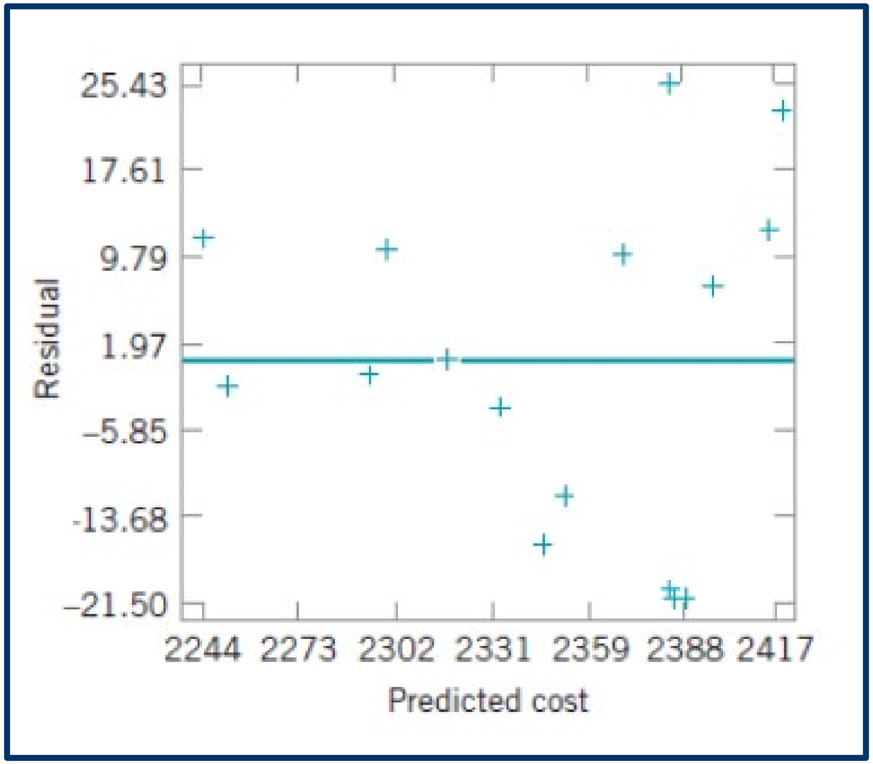


Observation i	y_i	Predicted Value $\hat{y}_i$	Residual e_i	h_{ii}
1	2256	2244.5	11.5	0.350
2	2340	2352.1	–12.1	0.102
3	2426	2414.1	11.9	0.177
4	2293	2294.0	–1.0	0.251
5	2330	2346.4	–16.4	0.077
6	2368	2389.3	–21.3	0.265
7	2250	2252.1	–2.1	0.319
8	2409	2383.6	25.4	0.098
9	2364	2385.5	–21.5	0.142
10	2379	2369.3	9.7	0.080
11	2440	2416.9	23.1	0.278
12	2364	2384.5	–20.5	0.096
13	2404	2396.9	7.1	0.289
14	2317	2316.9	0.1	0.185
15	2309	2298.8	10.2	0.134
16	2328	2332.1	–4.1	0.156

Different plots can be used to represent relevant information graphically:

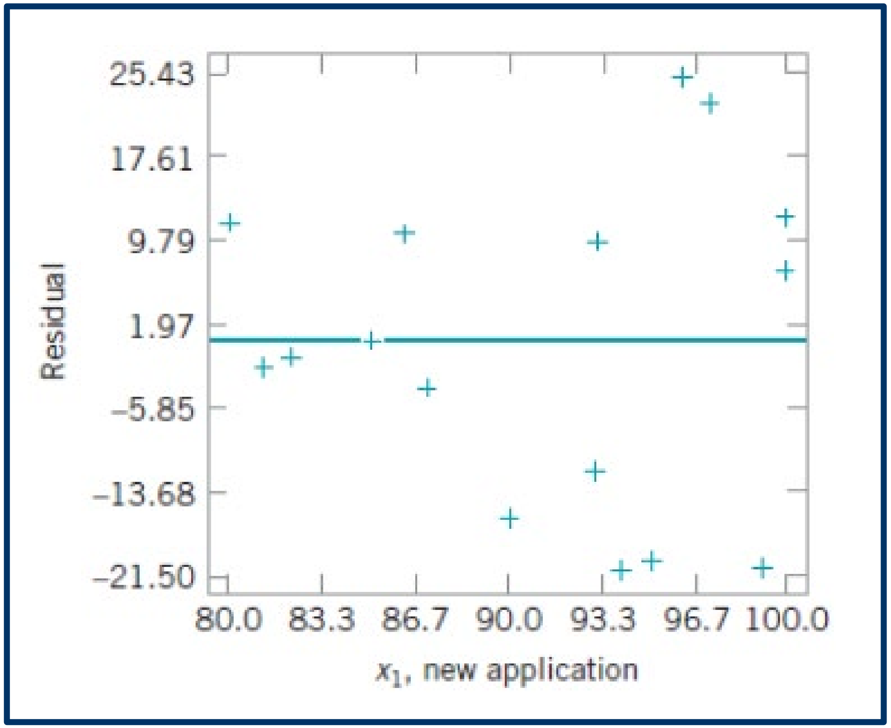


Normal probability plot of residuals

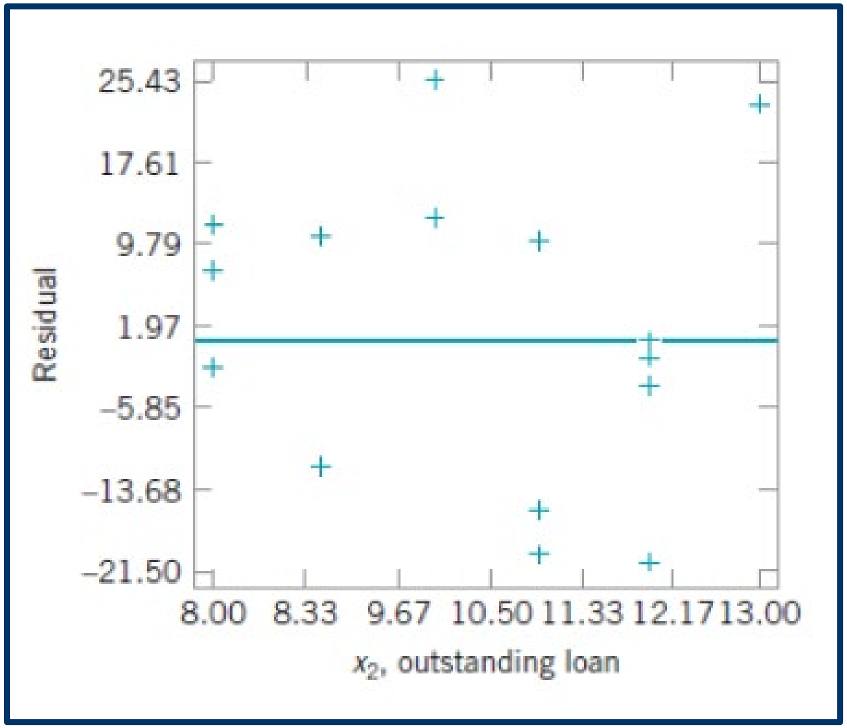


Plot of residuals vs predicted cost

The plot on the right indicates that the variance of the observed cost tends to increase with the magnitude of the cost.



Plot of residuals vs new loan applications

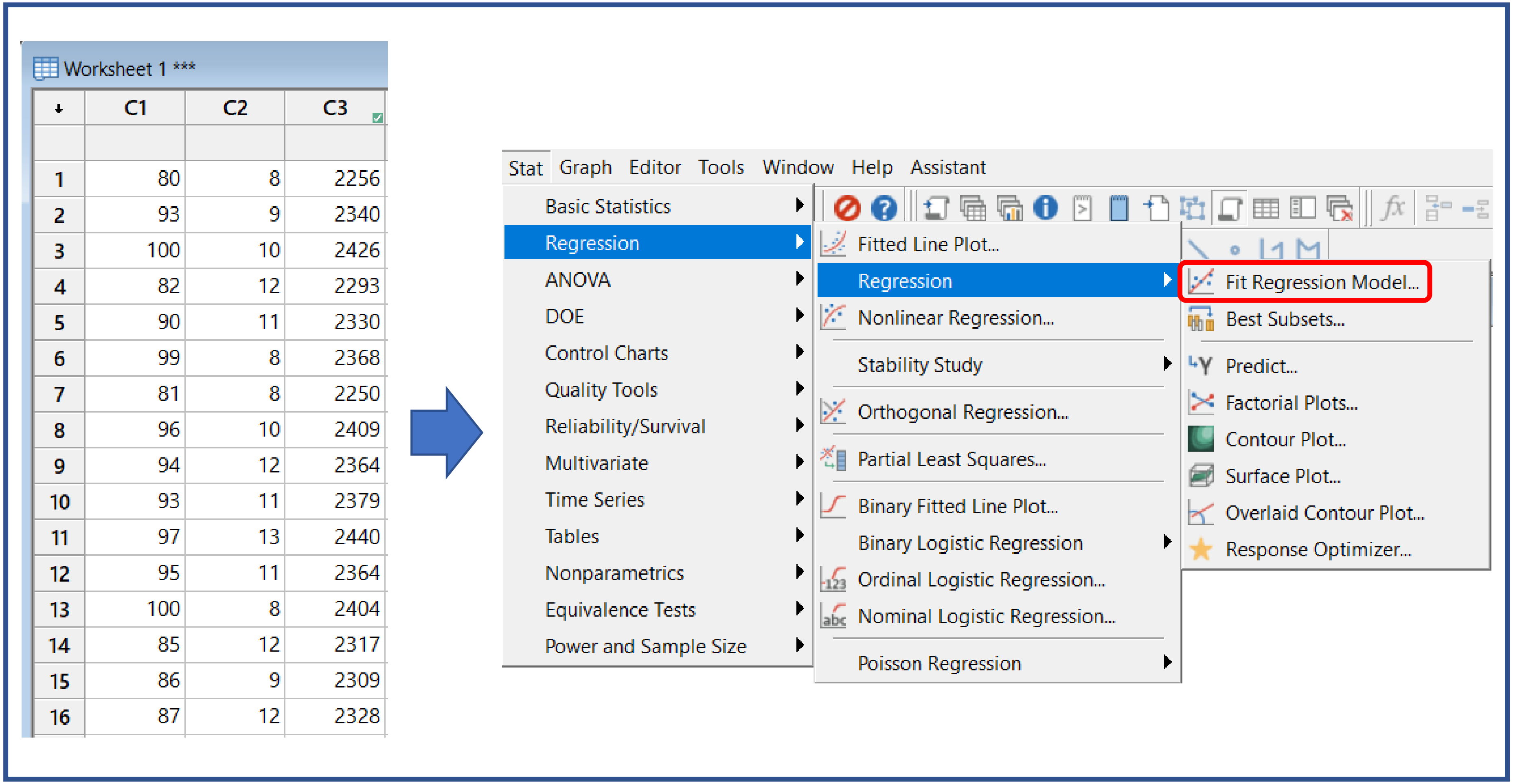


Plot of residuals vs outstanding loans

The plot on the left suggests that the variability in cost may be increased as the number of new applications increases, whereas the effect is not observed at the increase of outstanding loans.

Use of Minitab 18 to perform multiple linear regression (two predictors)

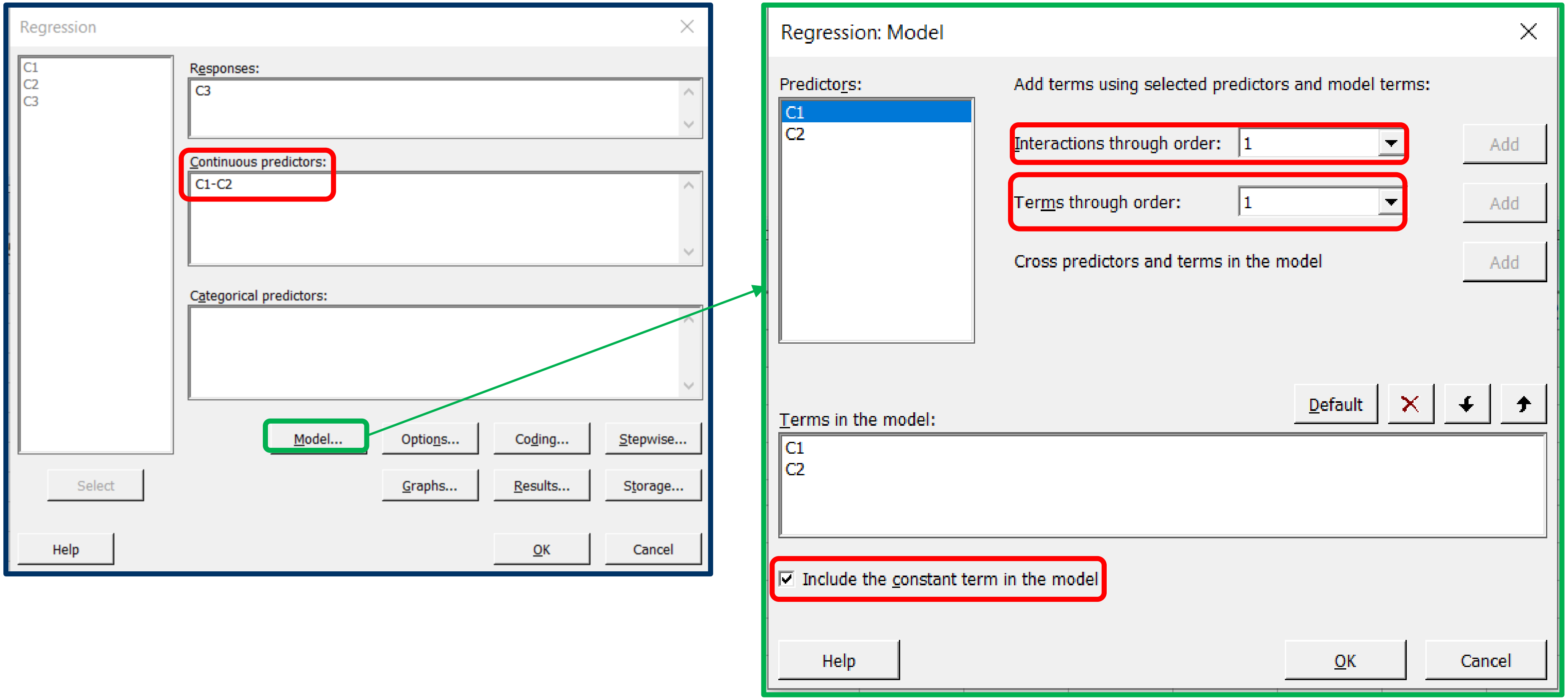
Minitab 18 can be used to perform multiple linear regression by introducing data of predictors and responses in appropriate columns of the Worksheet and then using the Stat > Regression > Regression > Fit Regression Model… option already considered for other types of regressions. The same dataset considered before for multiple linear regression with two predictors will be exploited as an example.



Worksheet 1 ***

↓	C1	C2	C3
1	80	8	2256
2	93	9	2340
3	100	10	2426
4	82	12	2293
5	90	11	2330
6	99	8	2368
7	81	8	2250
8	96	10	2409
9	94	12	2364
10	93	11	2379
11	97	13	2440
12	95	11	2364
13	100	8	2404
14	85	12	2317
15	86	9	2309
16	87	12	2328

In this case all columns of predictors have to be indicated in the box related to continuous predictors, then the order of regression (1 in the specific case) has to be indicated, for each predictor, in the Model… window (*Terms through order* box).



Regression

C1
C2
C3

Responses: C3

Continuous predictors: C1-C2

Categorical predictors:

Regression: Model

Predictors:
C1
C2

Add terms using selected predictors and model terms:

Interactions through order: 1

Terms through order: 1

Cross predictors and terms in the model

Terms in the model:
C1
C2

☑ Include the constant term in the model

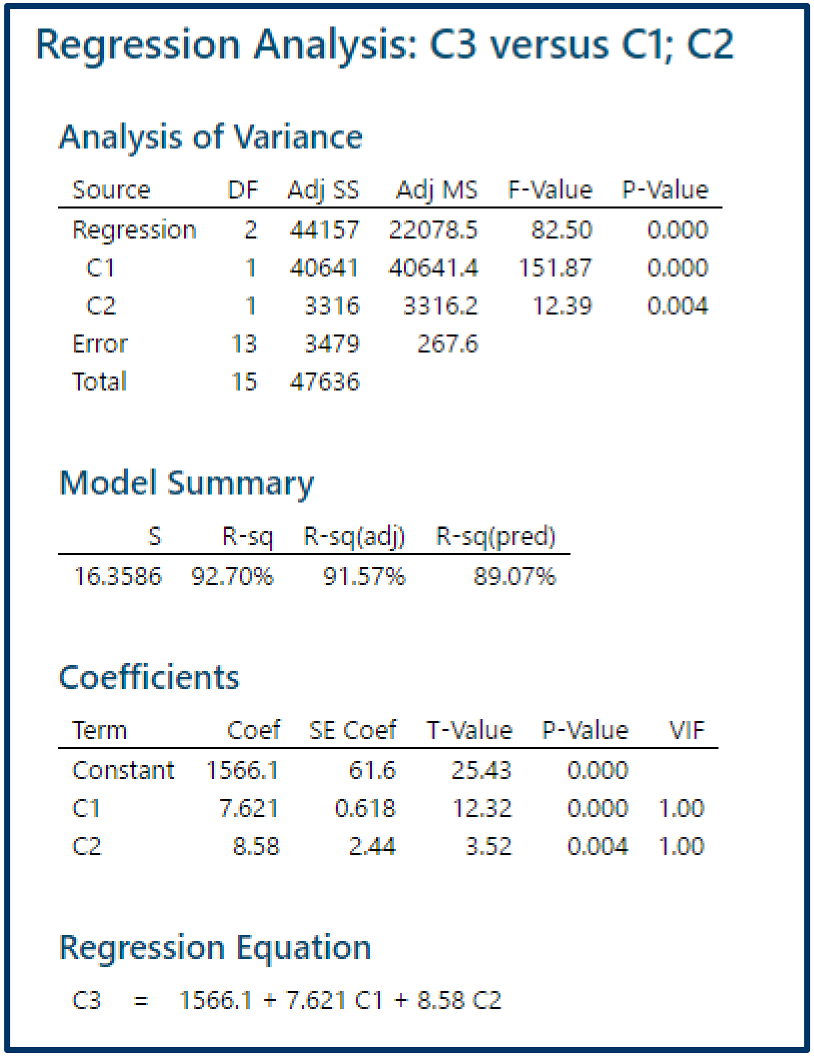
As shown for other regression methods, the inclusion of a constant term in the model can be selected in the same window.

Moreover, the value to be introduced in the *Interactions through order* box has to be put equal to 1 if no cross-terms (like X1 * X2 in the present case) have to be used in the model.

As for other methods of regression, a complete summary of information is displayed in the Session window once calculations are completed:

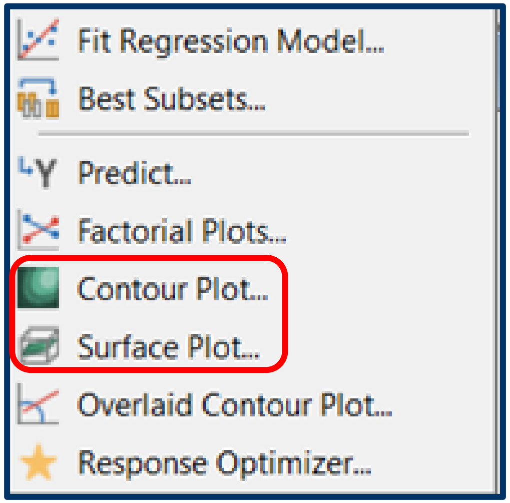
As apparent from the ANOVA table, both predictors are significant in the model, since their P-values are much lower even than 0.01.

Moreover, Variance Inflation Factors (VIF) are equal to 1 for both coefficients, thus indicating that collinearity between the two predictors is not present.



```
Regression Analysis: C3 versus C1; C2

Analysis of Variance

Source      DF  Adj SS   Adj MS  F-Value  P-Value
Regression   2   44157  22078.5    82.50    0.000
  C1         1   40641  40641.4   151.87    0.000
  C2         1    3316   3316.2    12.39    0.004
Error       13    3479    267.6
Total       15   47636

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
16.3586  92.70%     91.57%      89.07%

Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  1566.1     61.6    25.43    0.000
C1         7.621    0.618    12.32    0.000  1.00
C2          8.58     2.44     3.52    0.004  1.00

Regression Equation

C3  =  1566.1 + 7.621 C1 + 8.58 C2
```

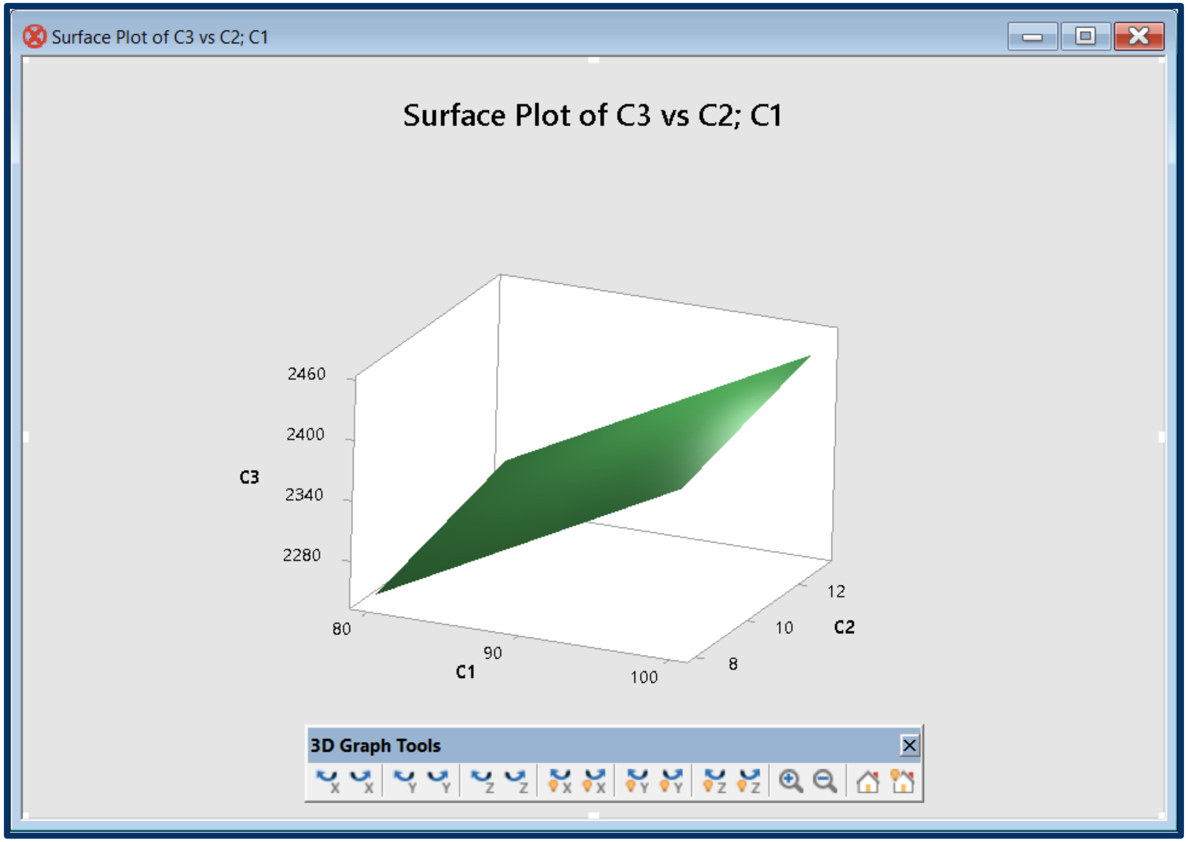
When multiple linear regression is performed using Minitab 18, useful graphs can be generated by using the Surface Plot… and the Contour Plot… options in the Stat > Regression > Regression menu.



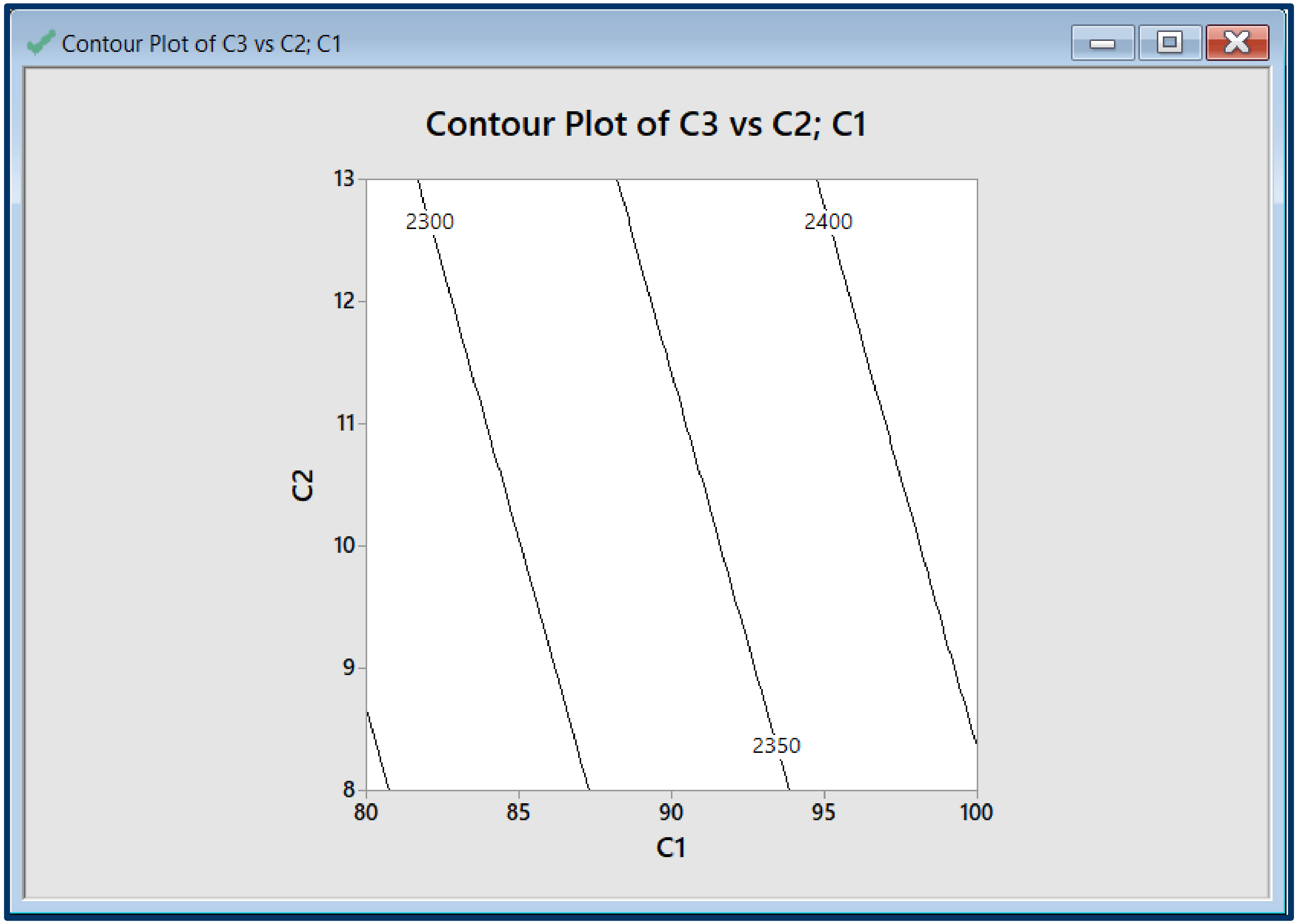
In particular, a 3D plot in which responses modeled by regression are shown as a function of the two predictors values is obtained using the Surface Plot… option.

A plane is actually obtained when no interaction term for predictors is present.

The plot can be conveniently rotated using 3D Graph Tools options.



In the Contour Plot lines including (C1, C2) coordinates leading to the same response are evidenced in a bi-dimensional graph:



Application of multiple linear regression to multicomponent analysis

Multicomponent analysis consists in the simultaneous determination of several components in low selectivity analytical systems, e.g., UV or IR spectroscopies.

Components can be elements, compounds, or chemical/physical properties. As an example, constituents of pharmaceutical formulations can be determined in the UV range, the water and protein content of cereals can be estimated from NIR spectra, and technological parameters of coal can be predicted using infrared (IR) spectra.

Let us start from Beer's law expressed in a simplified version, i.e., based on absorbances normalized to a constant cell thickness:

$$A_\lambda = k_\lambda c$$

where k_λ is the normalized absorption coefficient.

If absorbance additivity is valid, absorbances A_i obtained at a specific wavelength i for a multicomponent system can be represented by the sum of absorbances of the m individual components according to the following equation:

$$A_i = k_{i1}c_1 + k_{i2}c_2 + \cdots + k_{im}c_m = \sum_{j=1}^{m} k_{ij}c_j$$

where:

k_{ij} is the normalized absorption coefficient of the j-th component at wavelength i

c_j is the concentration of the j-th component

In multiwavelength spectroscopy, spectra are acquired at p wavelengths, so that either exactly determined linear equation systems (p = m) or over-determined systems (p > m) are obtained:

$$\begin{cases} A_1 = k_{11}c_1 + k_{12}c_2 + \cdots + k_{1m}c_m \\ A_2 = k_{21}c_1 + k_{22}c_2 + \cdots + k_{2m}c_m \\ \vdots \\ A_p = k_{p1}c_1 + k_{p2}c_2 + \cdots + k_{pm}c_m \end{cases}$$

The system of equations can be written in matricial notation as:

$$\mathbf{a} = \mathbf{Kc}$$

where:

a is a p × 1 vector representing the spectrum

K is a p × m matrix containing normalized molar absorptivities

c is a m × 1 vector representing concentrations

The system of equations can then be assimilated to the problem of multiple linear regression, in which the basic equation, in matricial notation, is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \Rightarrow \quad \mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

By analogy, the solution of the spectrophotometric equation for an unknown mixture of components is:

$$\mathbf{c}_0 = \left(\mathbf{K}^{\mathrm{T}}\mathbf{K}\right)^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{a}_{\mathbf{0}}$$

where $\mathbf{a}_0$ is the spectrum for the unknown mixture and $\mathbf{c}_0$ is the vector of predicted concentrations.

The described solution obviously requires the knowledge of normalized absorptivities for all components at all the selected wavelengths.

This information can be obtained and applied if:

1) pure component spectra can be measured, or obtained from external sources (scientific literature, databases, etc.)

2) there is no interaction between the different components in the sample or between constituents and the solvent, that could influence the spectrophotometric response

3) unknown matrix constituents do not interfere with the determination

This approach is known as the Direct Calibration Method.

An example of direct calibration: mixture of two components

Let us consider the following matrices:

$$\mathbf{K} = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 3 & 4 \end{bmatrix} \qquad \mathbf{a}_0 = \begin{bmatrix} 0{,}71 \\ 1{,}09 \end{bmatrix}$$

The solution of the system of spectrophotometric equations is:

$$\mathbf{c}_0 = \left(\mathbf{K}^{\mathrm{T}}\mathbf{K}\right)^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{a}_0 = \left(\begin{bmatrix} 3 & 3 \\ 2 & 4 \end{bmatrix} . \begin{bmatrix} 3 & 2 \\ 3 & 4 \end{bmatrix}\right)^{-1} \begin{bmatrix} 3 & 3 \\ 2 & 4 \end{bmatrix} . \begin{bmatrix} 0{,}71 \\ 1{,}09 \end{bmatrix} = \left(\begin{bmatrix} 18 & 18 \\ 18 & 20 \end{bmatrix}\right)^{-1} \begin{bmatrix} 3 & 3 \\ 2 & 4 \end{bmatrix} . \begin{bmatrix} 0{,}71 \\ 1{,}09 \end{bmatrix}$$

$$= \begin{bmatrix} 0.55556 & -0.5 \\ -0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 3 & 3 \\ 2 & 4 \end{bmatrix} . \begin{bmatrix} 0{,}71 \\ 1{,}09 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.55556 & -0.5 \\ -0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 5.4 \\ 5.78 \end{bmatrix} = \begin{bmatrix} 0{,}110 \\ 0{,}190 \end{bmatrix}$$

If the expected values for the component concentrations are 0.100 and 0.200, respectively, relative deviations of 10 and 5% are observed, respectively.

If a further couple of absorbance measurements is considered, the solution of the equation system becomes:

$$\mathbf{c}_0 = \left(\mathbf{K}^{\mathrm{T}}\mathbf{K}\right)^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{a_0} = \left(\begin{bmatrix} 3 & 3 & \boxed{2} \\ 2 & 4 & \boxed{6} \end{bmatrix} . \begin{bmatrix} 3 & 2 \\ 3 & 4 \\ 2 & 6 \end{bmatrix}\right)^{-1} \begin{bmatrix} 3 & 3 & 2 \\ 2 & 4 & 6 \end{bmatrix} . \begin{bmatrix} 0{,}71 \\ 1{,}09 \\ 1{,}41 \end{bmatrix} = \begin{bmatrix} 0{,}0995 \\ 0{,}2002 \end{bmatrix}$$

The deviation between calculated and expected concentrations has thus been reduced to 0.5 and 0.1 %, respectively.

As a general rule, it can be expected that the precision of the procedure increases at the increase of the number of measurements.

To some extent, the effect of using an over-determined system in multicomponent analysis is conceptually comparable to the effect of repeated measurements on the precision.

A further example for a mixture of two components

Let us consider the determination of Cl_2 and Br_2 in mixture in chloroform based on spectroscopic determination at six wavenumbers.

The following system of spectrophotometric equations was obtained:

$$\left\{\begin{array}{l} y_1 = 4.5c_1 + 168c_2 = 34.10 \\ y_2 = 8.4c_1 + 211c_2 = 42.95 \\ y_3 = 20c_1 + 158c_2 = 33.55 \\ y_4 = 56c_1 + 30c_2 = 11.70 \\ y_5 = 100c_1 + 4.7c_2 = 11.00 \\ y_6 = 71c_1 + 5.3c_2 = 7.98 \end{array}\right.$$

The system solution is $\mathbf{c} = \left(\mathbf{K}^{\mathrm{T}}\mathbf{K}\right)^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{y}$, thus:

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \left[\begin{bmatrix} 4.5 & 8.4 & 20 & 56 & 100 & 71 \\ 168 & 211 & 158 & 30 & 4.7 & 5.3 \end{bmatrix} \begin{bmatrix} 4.5 & 168 \\ 8.4 & 211 \\ 20 & 158 \\ 56 & 30 \\ 100 & 4.7 \\ 71 & 5.3 \end{bmatrix} \right]^{-1} \times \begin{bmatrix} 4.5 & 8.4 & 20 & 56 & 100 & 71 \\ 168 & 211 & 158 & 30 & 4.7 & 5.3 \end{bmatrix} \begin{bmatrix} 34.10 \\ 42.95 \\ 33.55 \\ 11.70 \\ 11.00 \\ 7.98 \end{bmatrix}$$

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 18667.81 & 8214.7 \\ 8211.7 & 98659.18 \end{bmatrix}^{-1} \begin{bmatrix} 4.5 & 8.4 & 20 & 56 & 100 & 71 \\ 168 & 211 & 158 & 30 & 4.7 & 5.3 \end{bmatrix} \begin{bmatrix} 34.10 \\ 42.95 \\ 33.55 \\ 11.70 \\ 11.00 \\ 7.98 \end{bmatrix} =$$

$$= \begin{bmatrix} 5.56\times10^{-5} & -4.63\times10^{-6} \\ -4.63\times10^{-6} & 1.052\times10^{-5} \end{bmatrix} \begin{bmatrix} 4.5 & 8.4 & 20 & 56 & 100 & 71 \\ 168 & 211 & 158 & 30 & 4.7 & 5.3 \end{bmatrix} \begin{bmatrix} 34.10 \\ 42.95 \\ 33.55 \\ 11.70 \\ 11.00 \\ 7.98 \end{bmatrix} =$$

$$= \begin{bmatrix} -0.00053 & -0.00051 & 0.00038 & 0.00298 & 0.00554 & 0.00392 \\ 0.00174 & 0.00218 & 0.00157 & 0.00006 & -0.00041 & -0.00027 \end{bmatrix} \begin{bmatrix} 34.10 \\ 42.95 \\ 33.55 \\ 11.70 \\ 11.00 \\ 7.98 \end{bmatrix}$$

The final result is:

c_1 = 0.099241 M, c_2 = 0.199843 M

These values are very close to the respective expected values, i.e., 0.100 and 0.200.

The variance on calculated concentrations can be obtained through formulas similar to those used to obtain the variance on p model parameters in multiple linear regression based on n data:

$$V(\mathbf{b}) = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1} = s_e^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \frac{\sum_i e_i^2}{n-p} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

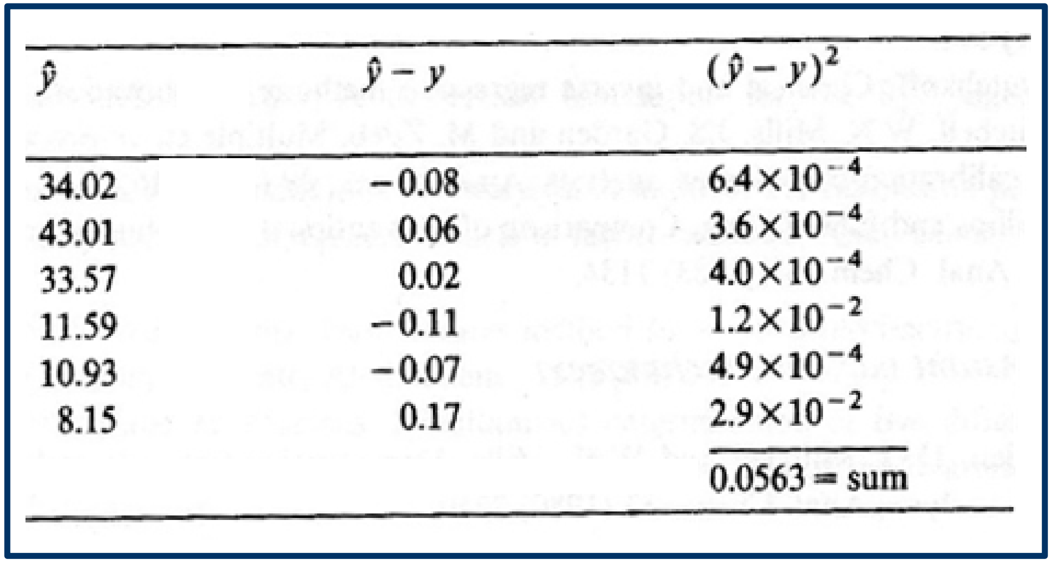
Consequently, the variance of concentrations for m components determined from p equations can be calculated as:

$$V(\mathbf{c}) = s_e^2 \left(\mathbf{K}^T\mathbf{K}\right)^{-1} = \frac{\sum_i e_i^2}{p-m} \left(\mathbf{K}^T\mathbf{K}\right)^{-1}$$

It is thus worth noting that in multicomponent analysis the propagation of error on the concentrations of components depends on the choice of normalized absorptivities and, consequently, on the choice of wavelenghts.

In the example of Cl_2 and Br_2 determination, $p = 6$ and $m = 2$.

Predicted values of absorbances, residuals and their squares are reported in the table on the right:



$\hat{y}$	$\hat{y} - y$	$(\hat{y} - y)^2$
34.02	−0.08	6.4×10^{-4}
43.01	0.06	3.6×10^{-4}
33.57	0.02	4.0×10^{-4}
11.59	−0.11	1.2×10^{-2}
10.93	−0.07	4.9×10^{-4}
8.15	0.17	2.9×10^{-2}
		0.0563 = sum

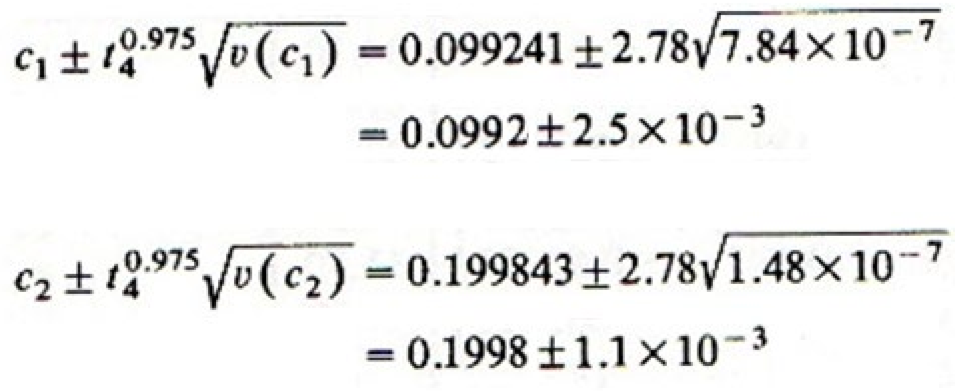
The s_e^2 quantity can be thus calculated:

$$s_e^2 = \frac{0.0563}{6-2} = 0.0141$$

Variances on single concentrations can be obtained by multiplying s_e^2 for the diagonal terms of the **$(K^TK)^{-1}$** matrix:

$$v(c_1) = (5.56\times10^{-5})(1.41\times10^{-2}) = 7.84\times10^{-7}$$
$$v(c_2) = (1.05\times10^{-5})(1.41\times10^{-2}) = 1.48\times10^{-7}$$

The 95% confidence limits for the concentrations can finally be obtained:



$$c_1 \pm t_4^{0.975}\sqrt{v(c_1)} = 0.099241 \pm 2.78\sqrt{7.84\times 10^{-7}}$$

$$= 0.0992 \pm 2.5\times 10^{-3}$$

$$c_2 \pm t_4^{0.975}\sqrt{v(c_2)} = 0.199843 \pm 2.78\sqrt{1.48\times 10^{-7}}$$

$$= 0.1998 \pm 1.1\times 10^{-3}$$