Chapter 4: Performing a multiple regression analysis

For an overview of key terms, refer to page 152-157

What is multiple regression analysis?
MRA is a statistical technique used to analyze the relationship between a single **dependent variable** and several **independent variables**. The objective is to use the independent variables to predict the single dependent value selected by the researcher. The set of weighted independent variables forms the regression variate, a linear combination of the independent variables that best predicts the dependent variable. The regression equation is the most widely known example of a variate among the multivariate techniques. It is required that both dependent and independent variables are metric. Sometimes it is possible to use nonmetric variables, by replacing them with dummy variables.
When there is a single independent variable, the statistical technique is called simple regression. When there are more than one independent variables involved, it is called multiple regressison.

Simple regression
For researchers, identifying the single independent variable that is the best prediction is the starting point. We can select this variable based on the correlation coefficient. The higher this coefficient, the stronger the relationship. In the regression equation, we represent the intercept as b0 and the regression coefficient as b1. With a method called least squares we can estimate the values of b0 and b1 such that the sum of squared errors of prediction is minimized. The prediction error is called the residual, e. Interpretation of the simple regression model:

- **Regression coefficient** = If it is found significant, the value of the regression coefficient indicates the extent to which the independent variable is associated with the dependent variable.
- **Intercept** = The intercept has explanatory value only within the range of values for the independent variable. Its interpretation is based on the characteristics of the independent variable. The intercept only has interpretive value if zero is a conceptually valid value for the independent variable . If the independent variable represents a measure that never can have a true zero value , it aids in improving the prediction process, but has no exploratory value.
- **Assessing prediction accuracy.** The most commonly used measure of predictive accuracy is the coefficient of determination (R2). It is calculated as the squared correlation between the actual and predicted values of the independent variable and represents the combined effects of the entire value in predicting the dependent variable. It ranges from 1 to 0. Another measure is the expected variation in the predicted values, termed the standard error of the estimate (SEe). It allows the researcher to understand the confidence interval that can be expected for any prediction from the regression model. Smaller confidence intervals denote greater predictive accuracy.

Multiple regression equation

The impact of multicollinearity. Collinearity is the association, measured as the correlation, between two independent variables. Multicollinearity refers to the correlation among three or more independent variables. Multicollinearity reduces the single independent variable's predictive power by the extent to which it is associated with the other independent variables. To maximize prediction from a given number of variables, the researcher should try to look for independent variables with low multicollinearity with the other independent variables. The task for the researcher is to expand upon the simple regression model by adding independent variables that have the greatest additional predictive power. The addition of extra independent variables is based on trade-offs between increased predictive power versus overly complex and even misleading regression models.

There is a six stage decision process for multiple regression analysis.

Stage 1- objectives of multiple regression.

The necessary starting point is the research problem. In selecting suitable applications for multiple regression, the researcher must consider three primary issues:

1. The appropriateness of the research problem
   The ever-widening applications of MRA fall into two broad categories: prediction and explanation.

- **Prediction** involves the extent to which the regression equation can predict the dependent variable. This type of multiple regression fulfils one of two objectives. The first objective is to maximize the overall predictive power of the independent variables as represented in the variate. Multiple regression can also achieve a second objective of comparing two or more sets of independent variables to ascertain the predictive power pf each variables. This use of MRA is concerned with the comparison of results across two or more alternative or competing models.
- **Explanation** examines the regression coefficients for each individual independent variable and attempts to develop a substantive or theoretical reason for the effects of the independent variables. Interpretation of the variate may rely on any of three perspectives. The most direct interpretation is a determination of the relative importance of each independent variable in the prediction of the dependent measure. MRA assesses simultaneously the relationships between each independent variable and the dependent measure. MRA can in addition also afford the researcher a means of assessing the nature of the relationships between independent variables and the dependent variable. Finally MRA provides insights into the relationships among independent variables in their prediction of the dependent measure. These relationships are important because correlation between these variables may make some variables redundant in their predictive efforts and .

2. Specification of a statistical relationship
   MRA is useful when the researcher wants to obtain a statistical relationship. A statistical relationship is characterized by two elements. First, when multiple observations are collected, more than one value of the independent variable will usually be observed for any of an independent variable. Second, based on the use of a random sample, the error in predicting the dependent variable is also assumed to be random. A statistical relationship estimated an average value, where a functional relationship estimates an exact value.
3. Selection of the dependent and independent variables
   The success of any multivariate technique starts with the selection of variables. The researcher must specify which variables are dependent and which are independent. The

researcher should always consider three issues:

- **Strong theory**, the selection of variables should always be based on theory;
- **Measurement error**, this refers to the degree to which the variable is an accurate and consistent measure of the concept being studied. Problematic measurement error may be addressed by using summated scales or structural equation modeling;
- **Specification error**, this is probably the most problematic issue and it is concerning the inclusion of irrelevant variables or the omission of relevant variables from the set of independent variables. Both types of specification can have substantial impacts on any regression analysis. The first type can reduce model parsimony, the additional variables may mask the effects of more useful variables and the additional variables may make the testing of statistical significance less precise and reduce the statistical significance. The second type can bias the results and negatively affect any interpretation of the variables.

Stage 2- research design of a Multiple Regression Analysis

MRA can represent a wide range of dependence relationships, in which the researcher incorporates three features.

1. Sample size: MRA maintains necessary levels of statistical power and practical/ statistical significance across a broad range of sample sizes.

- **Power levels** in various regression models. In multiple regression power refers to the probability of detecting as statistically significant a specific level of R2. Sample size plays a role in assessing power, but also in anticipating the statistical power of a proposed analysis. The researcher can also consider the role of sample size in significance testing before collecting the data.
- Sample size affects the **generalizability** of the results by the ratio of observations to independent variables. The ratio should never fall below 5:1. As the ratio is lower, there is a risk of overfitting.
  Degrees of freedom (df)= Sample size - number of estimated parameters, = Sample size – Number of independent variables +1
  The larger the degrees of freedom, the more generalizable the results. If the number of independent variables are reduced, the degrees of freedom increase.

2. Creating additional variables
   Unique elements of the dependence relationship; Independent variables are assumed to be metric and linear. However, these two assumptions can be relaxed by creating additional variables to represent these special aspects of the relationship. Metric variables are required. However, sometimes nonmetric data should be incorporated in the analysis. In this cases we can use **variable transformations**. Our purpose is here to provide the researcher with a means to modify the independent or the dependent variable for one of two reasons. Firstly, to improve or modify the relationship between independent and dependent variables. Secondly, to enable the use of nonmetric variables in the regression variate. Data transformations may be based on reasons that are either theoretical or data derived. In each case the researcher must proceed many times by trial and error. If we want to **incorporate nonmetric data** we can do this by using dummy variables. Each dummy variable represents one category of a nonmetric independent variable, and each nonmetric variable with K categories can be represented by K-1 dummy variables.

- One of the two forms of dummy variable coding is **indicator coding**. In this type of coding each category is represented by a) or an 1. The regression coefficients for the dummy variables represent differences on the dependent variable for each group of respondence

from the reference category. These group differences can be assessed directly, because the coefficients are in the same units as the dependent variable. This form of coding is most appropriate when a logical reference group is present, such as in an experiment.

- The second type of coding is **effects coding**. It is the same as indicator coding, except that the comparison or omitted group is now given a value of -1 instead of 0 for the dummy variables. Now the coefficients represent differences for any group from the mean of all groups rather than from the omitted group.

Several types of data transformations are appropriate for **linearizing a curvilinear relationship**. Direct approaches involve modifying the values through some arithmetic transformation. These methods have several limitations. Firstly, these are applicable only in a simple curvilinear relationship. Secondly, they do not provide any statistical means for assessing whether the curvilinear or linear model is more appropriate. Finally, they accommodate only univariate relationships and not the interaction between variables when more than one independent variable is involved.
Now a means of creating new variables to explicitly model the curvilinear components is discussed.
First, the curvilinear effect should be specified. Power transformations of an independent variable that add a nonlinear component for each additional power of the independent variable are known as polynomials. The power of 1 represents a linear relationship, the power of 2 represents a quadratic relationship and this is the first inflection point of a curvilinear relationship. The power of 3 (cubic relationship) adds a second inflection point. The cubic term is usually the highest power used. Multivariate polynomials are created when the regression equation contains two or more independent variables. We must here also add an interaction term (X1X2), which is needed for each variable combination to represent fully the multivariate effects. Now, the curvilinear effect should be interpreted. Multicollinearity can cause problems in assessing the statistical significance of the individual coefficients to the extent that the researcher should assess incremental effects as a measure of any polynomial terms in a three step process:

1. Estimate the original regression equation
2. Estimate the curvilinear relationship (original equation plus polynomial term)
3. Assess the change in R2. If this is a significant change, a curvilinear relationship exists.

Common practice is to start with the linear component and then sequentially add higher order polynomials until there is nonsignificance. Polynomials also have their limitations. The first problem is concerning degrees of freedom. An additional term requires a degree of freedom , which may be restrictive. Also, multicollinearity is introduced by adding extra terms.
Now we can represent **interaction or moderator effects**. The nonlinear relationships discussed before require the creation of an extra variable to represent the changing slope of the relationship over the range of the independent variable. This variables focuses on the relationship between a single independent variable and the dependent variable. If an independent-dependent relationship is affected by another independent variable this is called a moderator effect. It changes the form of a relationship between an independent and the dependent variable. It is known as the interaction effect. The moderator term is a compound variable formed by multiplying X1 by the moderator X2, which is entered in the regression equation.
$Y= b0+ b1X1 + b2X2+ b3X1X2$
Because of multicollinearity effects, an approach similar to testing the significance of polynomial effects is employed.
The researcher follows a three-step process:

1. Estimate the original equation
2. Estimate the moderated equation (original plus the moderator variable)
3. Assess the change in R2. If it significant, then a moderator effect is present.

Now, we can interpret the moderating effects. The interpretation of the regression coefficient changes slightly in the moderated relationship. The moderator effect(b3 coefficient) indicates the unit change in the effect of X1 as X2 changes. The B1 and B2 coefficients now represent the effects of X1 and X2, respectively, when the other independent variable is zero. In the unmoderated regression, the regression coefficient B1 and B2are averaged across all levels of the other independent variables, whereas in a moderated relationship, they are separate from the other independent variables. To determine the total effect of an independent variables, the separate and moderated variables have to be combined.
btotal=b1+b3X2

Nature of the independent variables. MRA accommodates metric independent variables that are assumed to be fixed in nature as well as those with a random component.

Stage 3- Assumptions in Multiple Regression Analysis
The assumptions to be examined are:
1. Linearity of the phenomenon measured
2. Constant variance of the error terms
3. Independence of the error terms
4. Normality of the error term distribution

These assumptions apply for both individual variables as for relationships as a whole. This section focusses on examining the variate and its relationship with the dependent variable for meeting the assumptions of multiple regression. We cannot just test the variate only, but we should also test the single variables, because two questions cannot be answered if we would only test the variate:
1. Have assumption violations for individual variables caused their relationships to be misrepresented?
2. What are the sources and remedies of any assumptions violations for the variate?

Methods of diagnosis
The principial measure of prediction error for the variate is the residual, the difference between the observed and predicted variables. Some standardization is recommended to make the residuals comparable. The studentized residual is the most widely used. The values correspond with T values. Plotting the residuals versus the predicted values is a basis method of identifying assumption violation. However, there are some considerations:
- The most common plot is residuals (ri) versus the predicted dependent variable (Yi). In MRA only the predicted dependent variables represent the total effect of the regression variate, so the dependent variable is used.
- Violations of each assumption can be identified by specific patterns of the residuals (see figure 4-5). An important plot is the null plot, the plot of residuals where all assumptions are met. The residuals are falling randomly, with relative equal dispersion about zero and no strong tendency to be either greater or less than zero.

Linearity of the phenomenon.
The linearity of the relationship represents the degree to which the change in the dependent variable is associated with the independent variable. The regression coefficient is constant across

the range of values for the independent variable. The concept of correlation is based on a linear relationship, so it is crucial. Corrective action can be taken by one of three options:

1. Transforming data values
2. Directly including the nonlinear relationships in the regression model, such as through the creation of polynomial terms
3. Using specialized methods such as nonlinear regression

Identifying the independent variables for action. How do we know which variables we have to select for corrective action? We use partial regression plots, which show the relationship of a single independent variable to the dependent variable controlling for the effects of all other independent variables. Now the line running through the points in the plots will slope upward or downward, instead of being horizontal.

Constant variance of the error term.

The presence of unequal variances is one of the most common assumption violations. Diagnosis is made with residual plots or simple statistical tests. The most common plot is triangle-shaped in either direction. If variation is expected in the midrange more than in the tails, a diamond-shape is expected. Spss provides the levene test for testing heteroscedacity.

If heteroscedacity is present, there are two remedies:

1. The procedure of weighted least squares can be employed when the violation can be attributed to a single independent variable through the analysis of residual plot
2. Easier are a number of variance-stabilizing transformations

Independence of the error terms.

We can best identify independentness by plotting residuals against any possible sequencing variable. The pattern should appear random and similar to the null plot od residuals. Violations are identified by consistent patterns in the plots. (Examples are shown in figure 4-5e and f). Normality of the error term distribution. Perhaps the most encountered violation is the assumption of normality. The simplest test is creating a histogram. This is more difficult in smaller samples. A better test is a normal probability plot. These differ from residual plotsin that the standardized residuals are compared with the normal distribution. The normal distribution makes a straight diagonal line.

Stage 4- Estimating the regression model and assessing the overall model fit

Having specified the objectives of the regression analysis, selected the independent and dependent variables, addressed the issues of research design and assessed the variables for meeting the assumptions of the regression, the researcher now is ready to estimate the regression model and assess the overall predictive accuracy of the independent variables. We must accomplish three basic tasks:

1. Select a method for specifying the regression model to be estimated
   There are three approaches to specifying the regression model:

- **Confirmatory specification**
  The researcher specifies the exact set of variables to be included. The researcher has total control over the variable selection. It is simple in concept, but the researcher is responsible for all trade-offs between more independent variables and greater predictive accuracy versus model parsimony and concise explanation.
- **Sequential search methods**

There is a general approach of estimating the regression equation by considering variables until some overall criterion measure is achieved. There are two types of sequential search methods: stepwise estimation and forward addition and backward elimination.

Stepwise estimation
This approach enables the researcher to examine the contribution of each developing equation. The independent variable with the greatest contribution is added first. Independent variables are then selected for inclusion based on their incremental contribution over the variables already in the equation.

The specific issues at each stage are as follows:

- Start with a simple regression model by selecting the on independent variable that is the most highly correlated with the dependent variable (equation Y=b0+b1X1)
- Examine the partial correlation coefficients to find an additional independent variable that explains the largest statistically significant portion of the unexplained variance
- Recomputed the regression equation using the two independent variables, and examin the partial F value for the original variable in the model to see whether it still makes a significant contribution. If it doesn't, eliminate the new variable
- Continue this procedure until none of the candidates would contribute significantly.

Forward addition and backward elimination
These are mainly trial and error processes for finding the best regression estimates. The forward addition model is similar to the stepwise method in that it builds the regression equation starting with a single independent variable, whereas the backward elimination method starts with a regression equation including all the independent variables and then deletes independent variables that do not contribute significantly. The primary distinction is that the stepwise method can delete or add variables in every stage. In the forward addition and backward elimination methods adding and deleting cannot be reversed.
To many researchers, sequential methods seem to be a perfect solution to the dilemma faced in the confirmatory approach by achieving the maximum predictive power with only those variables that contribute in a statistically significant way.

Three critical caveats markedly affect the resulting regression equation

1. The multicollinearity among independent variables has substantial impact on the final model specification
2. All sequential search methods create a loss of control on the part of the researcher
3. Especially in the step-wise procedure, the researcher should employ more conservative thresholds to ensure that the overall error rate across all significance tests is reasonable.

- **Combinatorial approach**
  This is a generalized search process across all possible combinations of independent variables. The best known procedure is the all-possible- subsets regression, which is exactly as the name suggests. All possible combinations of independent variables are examined, and the best-fitting set is identified. Usage of this approach has decreased because of criticism on the a-theoretical nature and the lack of consideration of such factors as multicollinearity, the identification of outliers and influential, and the interpretability of the results.

The most important criterion in selecting an approach, is the researcher's substantive knowledge of the research context and any theoretical foundation that allows for an objective and informed perspective as to the variables to be included as well as the expected signs and the magnitude of their coefficients. With the independent variables selected and the regression coefficient estimated, the researcher must now assess the estimated model for meeting the assumptions underlying multiple regression.

2. Assess the statistical significance of the overall model in predicting the independent variable. We always expect some variation if we would take different random samples to test the hypothesis that the amount of variation explained by the regression model is more than the baseline production, the F-ratio is calculated as:
   F ratio= Dfregression = number of estimated coefficients-1
   dfresidual = sample size- number of estimated coefficients

   Three important features of this ratio should be noted:

- Dividing each sum of squared by its degrees of freedom results in an estimate of the variance
- If the ratio of the explained variance to the unexplained is high, the regression variate must be of significant value in explaining the dependent variable. If the F value is statistically significant, we can state that the regression model is not just specific for this sample, but for multiple samples from the population.
- Although larger R2 values result in higher F values, practical significance must be assessed separately from statistical significance.

Adjusting the coefficient of determination. The addition of an extra variable will always lead to a higher R2 value. This creates concerns with generalizability. The impact is most noticeable if the sample size is close to the number of predictor variables. What is needed is a more accurate measuring relating the level of overfitting to the R2 achieved by the model. This measure involves an adjustment in the number of independent variables to the sample size. In this way, adding non-significant independent variables to increase R2 can be discounted. We also get the adjusted R2 in output of statistical programs. This adjusted R2 is useful in comparing across regression equations involving different numbers of independent variables or different sample sizes because it is making allowances for the degrees of freedom. Statistical significance testing for the estimated coefficients in regression analysis is appropriate when the analysis is based on a sample. Establishing the significance level denotes the chance the researcher is willing to take being wrong about whether the estimated coefficient is different from zero.

- Sampling error is the cause for variation in the estimated regression coefficients for each sample drawn from a population. For small sample sizes, sampling error is larger and the estimated coefficients will most likely vary widely from sample to sample. The standard error is the expected variation of the estimated coefficients due to sampling error. With the significance level and the standard error, we can compute the confidence interval for a regression coefficient. With the confidence interval in hand and the researcher must ask three questions about the statistical significance of any regression coefficient.
- Was statistical significance established?
- How does the sample size come into play?
- Does it provide practical significance in addition to statistical significance?

3. Determine whether any of the observations exert an undue influence on the results.
   We shift our attention here to individual observations that lie outside the general patterns of data and strongly influence the regression results.
   There are different **types of influential observations**, these are the three basic types:

- Outliers: observations that have large residual values and can be identified only with respect to a specific regression model.
- Leverage points: observations that are distinct from the remaining observations based on coefficients for one or more independent variables.
- Influential observations: all observations that have disproportional effect on the regression results.

Many times, influential observations are difficult to **identify** through the traditional analysis. Their patterns are not that different as outliers. General forms of influential observations(figure 4-8):

- Reinforcing the general pattern and lowering standard error
- Conflicting, contrary to the general pattern
- Multiple influential points may work towards the same results
- Shifting, influential observations may affect all of the results in a similar manner

Remedies for influentials. Influentials, outliers and leverage points are all based on one of four conditions, each of which has a specific course of corrective action:

- An error in observation or data entry: remedy by correcting the data or deleting the case
- A valid but exceptional observation that is explainable by an extraordinary situation: remedy by deletion of the case unless the variables reflecting the extraordinary situation are included in the regression equation
- An exceptional observation with no likely explanation: remedy by analyzing with and without the observation
- An ordinary observation in its individual characteristics but exceptional in its combination of characteristics: remedy by modifying the conceptual basis

Stage 5- interpreting the regression variate

The next task is to interpret the regression variate by evaluating the estimated regression coefficients for their explanation of the dependent variable. Not only the regression model should be evaluated, but also the potential independent variables that were omitted if a sequential search or combinatorial method was employed.

The estimated regression coefficients , termed the b coefficients, represent both the type of relationship and the strength of the relationship between independent and dependent variables.

Prediction is an integral element in the regression analysis, both in the estimation process and in the forecasting situations.

First, in the ordinary least square estimation procedure, used to derive the regression variate, a **prediction** of the dependent variable is made for each observation in the data set. A single predicted variable is computed. As such, the predicted value represents the total of all effects of the regression model and allows the residual to be used exensively as a diagnostic measure for the overall regression model.

The real benefit of prediction comes in forecasting applications.

Many times the researcher is interested in more than just prediction. It is important for a regression model to have accurate predictions to support its validity, but many research questions are more focused on assessing the nature and impact of each independent variable in making the prediction

of the dependent variable. For **explanatory** purposes, the regression coefficients become indicators of the relative impact and importance of the independent variables in their relationship with the dependent variable. In many instances, the coefficients do not give us thus information directly. We must ensure that all of the independent variables are on comparable scales. Differences in variability can also disturb the analysis. What can help us is the beta coefficient. Standardization converts variables to a common scale and variability. Multiple regression does not only give us the regression coefficients, but also coefficients resulting from the analysis of standardized data called βcoefficients. They eliminate the problem of dealing with different units of measurement. Now we can determine which variable has the most impact.
Two cautions must be taken into consideration when dealing with beta coefficients:

1. They should be used as a guide to the relative importance of individual independent variables only when collinearity is minimal.
2. The beta values can be interpreted only in the context of the other variables in the equation.

A key issue in interpreting the regression variate is the correlation among independent variables. Some degree of multicollinearity is unavoidable. The researcher's task is to i**dentify multicollinearity.**The easiest way is an examination of the correlation matrix for the independent variables. We need a measure expressing the degree to which each independent variable is explained by other independent variables. The 2 most common measures for assessing both pairwise and multiple variable collinearity are tolerance and the variance inflation factor.
Tolerance is defined as the amount of variability of the selected independent variable not explained by the other independent variables. Tolerance can be defined simply in two steps:

1. Take each independent variable and calculate R2* (the amount of the independent variable that is explained by other independent variables)
2. Tolerance is calculated as 1-R2*
   If tolerance is high, multicollinearity is low.

The **Variance inflation factor** is calculated as the inverse of the tolerance value. High VIF values mean high multicolinearity.

Determine its impact on the results
The impact of multicollinearity can be categorized in terms of estimation or explanation. It always creates shared variance between the variables and thus decreases the ability to predict the roles of each independent variable.
Firstly, in the extreme case of multicollinearity, singularity, the estimation of any coefficients is prevented. Secondly, as multicollinearity increases, the ability to demonstrate that the estimated regression coefficients are significantly different from zero become markedly impacted due to increases in the standard error as shown in the VIF value. This is the most problematic with small sample sizes. Finally, high degrees of multicollinearity can also result in regression coefficients being incorrectly estimated and even having the wrong signs. In some instances, this reversion of signs is desired, that is called a suppression effect. In other instances, signs are reversed because of multicollinearity. In these cases, the researcher may need to revert using bivariate correlations to describe the relationship rather than the estimated coefficients that are impacted by multicollinearity.

As multicollinearity occurs, identifying the effects of each independent variable on the dependent variable becomes increasingly difficult. Each researcher must determine what degree of multicollinearity is too much. Some suggested guidelines have been developed and can be found on page 200.

Apply remedies if needed

The researcher can apply a number of remedies:

1. Omit one or more highly correlated independent variables and identify other independent variables to help the prediction.
2. Use the model with the highly correlated independent variables for prediction only, while acknowledging the lowered level of overall predictive accuracy.
3. Use the simple correlations between each independent variable and the dependent variable to understand the independent-dependent variable relationship.
4. Use a more sophisticated method of analysis, such as Bayesian regression.

Stage 6- Validation of the results

The final step is to ensure that it represents the general population and is appropriate for the situations in which it will be used.

The most appropriate validation approach is to test the regression model on a new sample drawn from the general population. It will ensure representativeness and can be used in several ways. It can predict values in the new sample and the predictive fit can be calculated. Also, a separate model can be estimated with the new sample and then be compared with the original equation. Many times, however, the ability to collect new data is limited by factors as cost, time or availability of respondents. Then the researcher can use a split sample.

An alternative approach a researcher can use is calculating the PRESS statistic. It is a measure similar to R2 used to assess predictive accuracy. N-1 regression models are estimated. One observation is omitted in the estimation of the regression model and then predicts the omitted observation with the estimated model. The procedure is applied again and the residuals can be summed to provide an overall measure of predictive fit.

When comparing regression models, the most common standard used is overall predictive fit. R2 gives this information, but had one disadvantage: as more variables are added, R2 will always increase. Therefore, we use the adjusted R2, it will compensate for different sample sizes.

In forecasting, we must consider several factors that can have a serious impact on the quality of the new predictions:

1. When applying the model to anew sample, we must remember that the predictions now have not only the sampling variations from the original sample, but also those of the newly drawn sample. So we should calculate confidence intervals of our predictions in addition to the point.
2. We must make sure that the conditions have not changed after our original measurements.
3. We must not use the model beyond the range of independent variables found in the sample.

For an example of a regression analysis, refer to page 203-226.