海外优秀数学类教材系列丛书

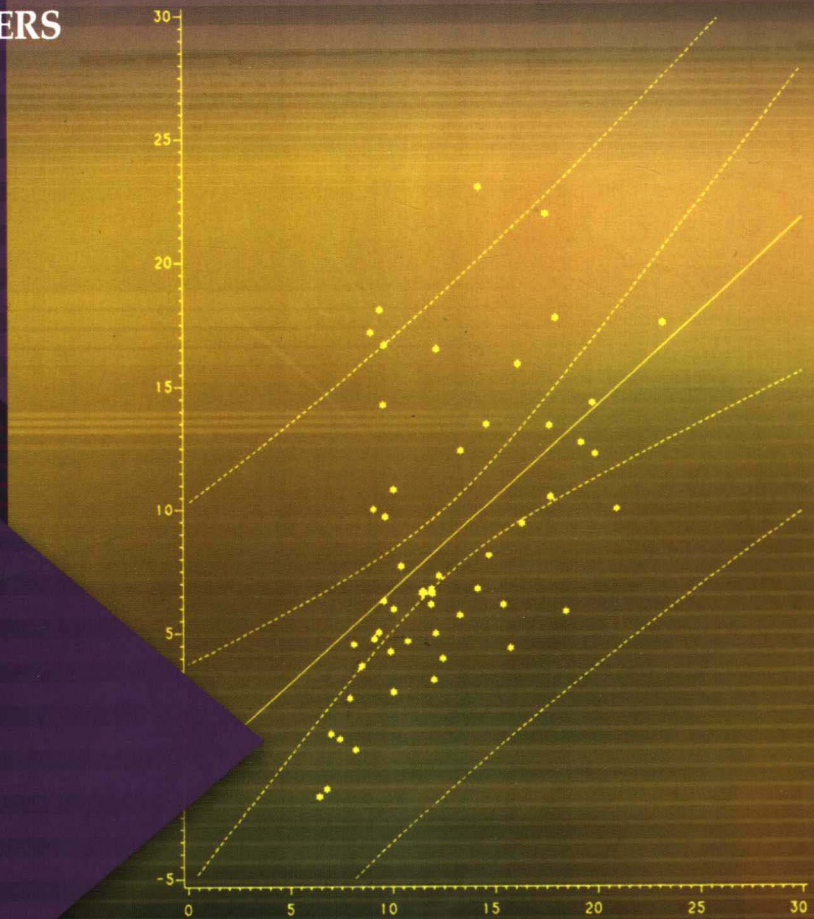
THOMSON

影印版

Classical and Modern Regression with Applications (Second Edition)

经典和现代回归分析及其应用（第2版）

RAYMOND H. MYERS



高等教育出版社
Higher Education Press

海外优秀数学类教材系列丛书

影印版

Classical and Modern Regression with Applications (Second Edition)

经典和现代回归分析及其应用（第2版）

Raymond H. Myers
Virginia Polytechnic Institute and State University

高等教育出版社
Higher Education Press

图字：01－2004－1625号

Raymond H. Myers

Classical and Modern Regression with Applications, Second Edition

ISBN:0－534－38016－6

981－265－457－7

图书在版编目（CIP）数据

经典和现代回归分析及其应用=Classical and Modern Regression with Applications：第2版/（美）麦尔斯（Myers, R.H.）著．—影印本．—北京：高等教育出版社，2005.5

（海外优秀数学类教材系列丛书）

ISBN 7-04-016323-3

Ⅰ.经… Ⅱ.麦… Ⅲ.回归分析－高等学校－教材－英文 Ⅳ.0212.1

中国版本图书馆CIP数据核字（2005）第028640号

出版发行	高等教育出版社	**购书热线**	010－58581118
社　　址	北京市西城区德外大街4号	**免费咨询**	800－810－0598
邮政编码	100011	**网　　址**	http://www.hep.edu.cn
总　　机	010－58581000		http://www.hep.com.cn
经　　销	北京蓝色畅想图书发行有限公司	**网上订购**	http://www.landraco.com
印　　刷	保定市印刷厂		http://www.landraco.com.cn
开　　本	787×1092　1/16	**版　　次**	2005年5月第1版
印　　张	31.25	**印　　次**	2005年5月第1次印刷
字　　数	700 000	**定　　价**	35.50元

物料号　16323－00

出版者的话

在我国已经加入 WTO、经济全球化的今天，为适应当前我国高校各类创新人才培养的需要，大力推进教育部倡导的双语教学，配合教育部实施的“高等学校教学质量与教学改革工程”和“精品课程”建设的需要，高等教育出版社有计划、大规模地开展了海外优秀数学类系列教材的引进工作。

高等教育出版社和 Pearson Education，John Wiley & Sons，McGraw - Hill，Thomson Learning 等国外出版公司进行了广泛接触，经国外出版公司的推荐并在国内专家的协助下，提交引进版权总数 100 余种。收到样书后，我们聘请了国内高校一线教师、专家、学者参与这些原版教材的评介工作，并参考国内相关专业的课程设置和教学实际情况，从中遴选出了这套优秀教材组织出版。

这批教材普遍具有以下特点：(1) 基本上是近 3 年出版的，在国际上被广泛使用，在同类教材中具有相当的权威性；(2) 高版次，历经多年教学实践检验，内容翔实准确、反映时代要求；(3) 各种教学资源配套整齐，为师生提供了极大的便利；(4) 插图精美、丰富，图文并茂，与正文相辅相成；(5) 语言简练、流畅、可读性强，比较适合非英语国家的学生阅读。

本系列丛书中，有 Finney、Weir 等编的《托马斯微积分》(第 10 版，Pearson)，其特色可用“呈传统特色、富革新精神”概括，本书自 20 世纪 50 年代第 1 版以来，平均每四五年就有一个新版面世，长达 50 余年始终盛行于西方教坛，作者既有相当高的学术水平，又热爱教学，长期工作在教学第一线，其中，年近 90 的 G. B. Thomas 教授长年在 MIT 工作，具有丰富的教学经验；Finney 教授也在 MIT 工作达 10 年；Weir 是美国数学建模竞赛委员会主任。Stewart 编的《微积分》(第 5 版，Thomson Learning) 配备了丰富的教学资源，是国际上最畅销的微积分原版教材，2003 年全球销量约 40 余万册，在美国，占据了约 50% ~60% 的微积分教材市场，其用户包括耶鲁等名牌院校及众多一般院校。本系列丛书还包括 Anton 编的经典教材《线性代数及其应用》(第 8 版，Wiley)；JayL. Devore 编的优秀教材《概率论与数理统计》(第 5 版，Thomson Learning) 等。在努力降低引进教材售价方面，高等教育出版社做了大量和细致的工作，这套引进的教材体现了一定的权威性、系统性、先进性和经济性等特点。

通过影印、翻译、编译这批优秀教材，我们一方面要不断地分析、学习、消化吸收国外优秀教材的长处，吸取国外出版公司的制作经验，提升我们自编教材的教学资源配套标准，使我国高校教材建设水平上一个新的台阶；与此同时，我们还将尝试组织海外作者和国内作者合编外文版基础课数学教材，并约请国内专家改编部分国外优秀教材，以适应我国实际教学环境。

这套教材出版后，我们将结合各高校的双语教学计划，开展大规模的宣传、培训工作，及时地将本套丛书推荐给高校使用。在使用过程中，我们衷心希望广大高校教师和同学提出宝贵的意见和建议。

高等教育出版社高等理科分社联系电话：010－58581384，E－mail：xuke@hep.com.cn。

高等教育出版社

2005年4月20日

PREFACE

No single statistical tool has received the attention given to regression analysis in the past 25 years. Both practical data analysts and statistical theorists have contributed to an unprecedented advancement in this important and dynamic topic. Many volumes have been written by statisticians and scientists with the result being that the arsenal of effective regression methods has increased manyfold.

My intent for this second edition is to provide a rather substantial increase in material related to classical regression while continuing to introduce relevant new and modern techniques. I have included major supplements in simple linear regression that deal with simultaneous influence, maximum likelihood estimation of parameters, and the plotting of residuals. In multiple regression, new and substantial sections on the use of the general linear hypothesis, indicator variables, the geometry of least squares, and relationship to ANOVA models are added. In addition, all new topics are illustrated with the use of real-life data sets and annotated computer printout. In the area of useful modern techniques, additional types of diagnostic residual plots are developed and illustrated, including component plus residual plots and augmented partial plots. These plots are designed to provide a two-dimensional picture of the role of each regressor in the multiple regression and graphically highlight the need for nonlinearities in the regression model.

The chapter on nonstandard conditions has received considerable enhancement in order to better highlight modern techniques. The Box–Cox transformation has been expanded considerably with illustrations. Regression with a categorical

response, and otherwise nonnormal error situations, are now given the consideration they deserve. Logistic regression is now expanded and is accompanied by Poisson regression. These two important techniques provide a foundation for a generalization to *generalized linear models* from which one can build models under error structures that have distributions such as gamma, negative binomial, exponential, and others. Other nonstandard conditions such as autocorrelated errors are discussed and illustrated. The text continues to emphasize applications that are selected from a wide variety of fields. The text puts emphasis on concepts and provides a blend between illustrations with real data sets and conceptual development.

Exercises in the text have been enhanced substantially. The number of exercises have been approximately doubled, and many new ones involve new data sets and problems to provide the student experience in using new techniques developed in the text. A data disk is available containing data sets in the exercises.

It is necessary that readers have been exposed to a course in basic statistical methods and thus are familiar with the use of the normal, t, χ^2, and F distributions. A basic course in calculus is required, and familiarity with matrix algebra is desirable. The text emphasizes applications with examples that illustrate nearly all techniques discussed. I have selected examples from a wide variety of fields including the physical sciences, engineering, biology, management science, and economics.

The book is designed for seniors and graduate students majoring in statistics or user subject-matter fields. There are nine chapters and three appendices. Chapter 1 introduces the general notion of regression models, and Chapter 2 deals specifically in simple linear regression with traditional material based on ordinary least squares estimation. Chapter 3 extends least squares estimation to multiple linear regression and introduces multicollinearity with illustrations, though modern methods of diagnosis and combating collinearity are relegated to Chapter 8. Both Chapters 2 and 3 have been expanded in areas previously discussed. Chapter 4 represents a blend between classical and contemporary methodology designed to choose the optimal subset model. Sequential analytic or *stepwise* model-building methods are discussed and illustrated. In addition, Mallow's C_p-statistic is developed and motivated as a criterion that represents a compromise between model overfitting and underfitting. The obvious connection is made in Chapter 4 between *cross-validation* criteria and model selection. To this end, the PRESS statistic and data splitting are presented and illustrated as criteria for discriminating among competing models. Illustrations are given of how modern software allows PRESS, C_p, and other model selection criteria to be used in an *all possible regression* type of procedure.

Chapter 5 deals solely in analysis of residuals when the user's intent is to do model criticism and detection of violation of assumptions. Modern diagnostic methods are discussed for outlier detection and plotting of residuals. The material in Chapter 6 revolves around modern methods for detecting data points that exert disproportionate influence on the regression results. Influence diagnostics,

which are rapidly becoming a standard part of the analyst's arsenal, are discussed. Chapter 7 follows logically by presenting procedures that are used as alternatives to standard methodology when assumptions are violated. Transformations are presented with illustrations. Weighted regression is used as an alternative to ordinary least squares when the homogeneous variance assumption fails. Robust regression is described for the case of outliers or non-Gaussian error assumptions. As I mentioned previously, Chapter 7 has been supplemented to include auto-correlated errors, additional development in transformations, and regression with categorical responses and other non-Gaussian conditions. Chapter 8 is dedicated completely to diagnostic and analytical methods in cases of data sets that contain multicollinearity. Ridge regression and principal components regression are developed and illustrated with real data sets. Chapter 9 covers nonlinear regression. Standard nonlinear regression methods for finding least squares estimators are developed and presented.

I have already mentioned the substantial number of real data sets that are a part of an expanded number of exercises. Other exercises are more conceptual or theoretical in nature and are certainly challenging to the graduate student or high-level undergraduate who has an appreciation of linear algebra.

The three appendices are each designed for a different purpose. Appendix A supplements the reader's background in linear algebra with a treatment of items such as eigenvalues, eigenvectors, and quadratic forms. Appendix B is designed to strengthen the reader's understanding of certain statistical concepts that are not likely to be covered in prerequisite courses. For example, treatment of notions such as maximum likelihood, the generalized variance, and expected values of quadratic forms are given. This will aid the student who is mathematically more sophisticated to achieve a higher level of understanding of some of the regression concepts in the text. In addition, Appendix B contains derivations of results dealing with quadratic forms, deletion formulae associated with influence diagnostics, and several other theoretical results that were deleted from the mainstream of the text so that the reader not be deflected from the practical concepts and consequences of the regression development. Appendix C contains statistical tables.

I would like to acknowledge those who made contributions to the project. First and foremost, I would like to thank my wife, Sharon, who proofread the manuscript at each stage and whose efforts resulted in many improvements. Thanks also go to Roy Welsh and Jeffrey Birch who made helpful suggestions on how to improve the first edition. In addition, I would like to thank the reviewers for this edition:

Fred Andrews, University of Oregon; Indra Chakavarti, University of North Carolina at Chapel Hill; Don Edwards, University of South Carolina at Columbia; Yu Sheng Hsu, Georgia State University; Jon Matta, University of Missouri; J. Michael Steel, Princeton University; R. Kirk Steinhorst, University of Idaho, and Wayne L. Winston, Indiana University.

Raymond H. Myers

CONTENTS

CHAPTER 1

INTRODUCTION: REGRESSION ANALYSIS

The term regression analysis describes a collection of statistical techniques that serve as a basis for drawing inferences about relationships among quantities in a scientific system. In the field of applied statistics, there is a plethora of modern data analysis techniques, which are branded with eye-catching names. The motivation of a particular method stems from the need to solve a specific problem. But regression analysis has been burdened with the responsibility—under one heading—of solving a variety of problems. For this reason, volumes are written about the topic, and its use continues to expand. New subtopics and subareas are hurled under the ever-growing umbrella that we continue to call *regression analysis*. The resulting analytical methodology, a product of the imagination of both professional statisticians and subject matter scientists, has become readily accessible today due to the rapid work of computer software personnel. Though it was not the intent or grand design, regression analysis is now perhaps the most used of all data analysis methods.

In 1885 Sir Francis Galton (1885) first introduced the word "regression" in a study that demonstrated that offspring do not tend toward the size of parents, but rather toward the average as compared to the parents. The upshot of the term's

application was the "regression towards mediocrity" of offspring. Credit for discovery of the method of least squares generally is given to Carl Friedrich Gauss, who used the procedure in the early part of the nineteenth century. There is some controversy concerning this discovery. Apparently Adrien Marie Legendre published the first work on its use in 1805. Regression analysis and the method of least squares nearly always were linked in practice and seemed to be compatible bedfellows until the latter part of the 1960s. It became very apparent that, in many nonideal situations, ordinary least squares (OLS) is not appropriate and, indeed, can be improved upon. Much controversy has evolved and will likely continue to cloud the notion of alternatives to OLS. However, biased estimation for combating multicollinearity (Chapter 8) and robust regression (Chapter 7) have been accepted by many data analysts and will attain a niche in the heritage of regression analysis.

Often a matter of no small question to the analyst is what effect quantities (variables) have on one another. The system that generates the data may be a chemical or biological process, the nation's economy, a group of patients in a medical experiment, or perhaps a group of metal specimen in a tensile strength study. In some cases the pertinent variables are random variables and are related in a probability sense through a joint probability distribution. In other cases, the variables are mathematical quantities, and the assumption is that there exists a functional relationship linking them. From a set of data involving measurements on the variables, regression analysis is designed to shed light on certain aspects of the mechanism that relate them. An example illustrates what kind of information regression analysis can acquire from the data. Suppose we select a group of men at random to take part in an experiment. Each individual must run a specified distance. We then make the following measures on each individual:

x_1: age
x_2: weight
x_3: pulse rate at rest (rest pulse)
x_4: pulse rate immediately following run (run pulse)
x_5: time that it takes to run the distance (run time)
y: oxygen consumption rate (oxygen rate)

For specific data on this type of application, the reader is referred to the *SAS User's Guide* (1985). The y variable represents the efficiency with which the individual utilizes oxygen. We can argue that perhaps y should play a somewhat different role than x_1, x_2, x_3, x_4, and x_5. It may be of interest to visualize the x variables as quantities that actually *determine* y or rather *predict* y. Thus the x's are called *independent variables* or *regressor variables*. Given the data, the analyst certainly would hope to derive information regarding the role of each regressor variable in terms of its influence on oxygen rate. If one or more of the variables has a negligible influence on oxygen, this is valuable information. In addition, we might expect that the data should allow for estimation of the relationship that exists among the variables (assuming a relationship exists). Before any specific techniques can be developed or even discussed, the notion of a statistical model must be introduced.

1.1

REGRESSION MODELS

In the following chapters much development revolves around the use and development of *statistical models*. Simply put, all procedures used and conclusions drawn in a regression analysis depend at least indirectly on the assumption of a regression model. A model is what the analyst perceives as the mechanism that generates the data on which the regression analysis is conducted. Regression models are usually found in an algebraic form. For example, in the oxygen consumption illustration, if the experimenter is willing to assume that the relationship is well represented by a structure that is linear in the regressor variables, then a suitable model may be given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \tag{1.1}$$

In Eq. (1.1), $\beta_0, \beta_1, \ldots, \beta_5$ are unknown constants called *regression coefficients*. Procedures embraced by regression analysis concern themselves with drawing conclusions about these coefficients. The investigator may be involved in a painstaking attempt to determine if an increase in x_4 (run pulse) truly does decrease or increase the efficiency of oxygen consumption. Perhaps the sign and magnitude of the coefficient is important. The ε term in the model is added to account for the fact that the model is not exact. It essentially describes the random disturbance or *model error*. When we apply (1.1) to a set of data, we can view the ε term as an aid in accounting for any variation due to the individual, that is *apart from the terms supplied by the model*. Chapter 2 gives a description that outlines the very important assumptions that often must be made on the ε's, assumptions on which the theory underlying regression analysis depends.

The model in Eq. (1.1) falls into the class of *linear models* (linear in the parameters, the β's). Any regression procedure involves *fitting the model* to a set of data, the latter defined as readings on the variables for the various experimental units being sampled (for example, the individuals sampled in the oxygen consumption situation). The term, *fit to a set of data*, actually involves estimation of the regression coefficients and the corresponding formulation of a *fitted regression model*, an empirical device that is the basis of any statistical inference made. Measures of quality of fit are important statistics that form the foundation of a statistical analysis. Clearly if the postulated model does not describe the data satisfactorily, any fundamental conclusions recovered from the fitted model are suspect.

This text is not confined to the treatment of linear models. Nonlinear models are commonplace in many of the natural sciences or engineering applications. A biochemist may postulate a growth model of the type

$$y = \frac{\alpha}{1 + e^{\beta t}} + \varepsilon \tag{1.2}$$

to represent the growth y of a particular organism as a function of time t. Here the parameters α and β are to be estimated from the data. Many of the same problems one attempts to solve by linear regression can often be handled by *nonlinear regression*. However, the computational aspects of building nonlinear models are less straightforward and thus require a special treatment (see Chapter 9).

The fitted regression model is produced by the model builder as an estimate of the functional relationship that describes the data. The type of model postulated quite often depends on the range of the regressor variables encountered in the data. For example, a chemical engineer may have knowledge of his system that necessitates the use of terms that impart curvature on the model. Suppose x_1 and x_2 represent temperature and a reactant concentration, and the response y is a simple yield of a chemical reaction. The engineer's goal is possibly two-fold:

1. To use the fitted or estimated regression (*prediction equation*) for yield estimation at locations x_1 and x_2 other than data conditions;
2. To study the relationship in the region of the data, perhaps with a view toward finding temperature and concentration conditions that provide satisfactory yield.

To reflect model curvature, a structure of the type

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \tag{1.3}$$

is used. Thus the ranges of the regressor variables may well dictate the model type. With narrow ranges in x_1 and x_2, the engineer may be successful at using a model that does not involve quadratic terms. Though Eq. (1.3) includes powers and products of order two in the regressor variables, it is, nevertheless, another example of a linear model since the coefficients enter linearly.

In nearly all regression applications in which linear models describe a set of data, the model formulation is an oversimplification of what occurs in the data observational process. Areas in the social and behavioral sciences represent examples where the systems from which the data are taken are far too complicated to model with an absolutely correct structure. The linear models used are approximations that hopefully *work well in the range of the data* used to build them. There is no intention here to cast stones on those who build linear models as approximations. When the sophistication of the subject matter field is not sufficient to provide a working theory, a linear and "common sense" empirical model approach can be very informative, particularly when used in conjunction with a set of data of reasonable quality.

1.2 FORMAL USES OF REGRESSION ANALYSIS

In this text much coverage is given to categories representing specific kinds of inferences, and distinctions are necessarily drawn among them. This is crucial since

either the estimation procedure or even the model that is adopted may well depend on what the intended goal of the study is. This often seems counter to the ideology of certain users of the methodology. To the inexperienced analyst, it may seem that the model that apparently best describes the data should be adopted for every purpose. However, a model that gives a satisfactory solution to one problem will not necessarily provide success in solving another. The uses of regression analysis fall into three or perhaps four categories, though there is some overlap. These categories are as follows:

1. Prediction
2. Variable screening
3. Model specification (system explanation)
4. Parameter estimation

The analyst is advised to know and bear in mind what goal is vital in his or her particular endeavor. Let us initially consider goal (1). Here, parameter estimates are not sought for their own sake. We do not search for the true model specification apart from how the functional form influences prediction. It is not important that we capture the role of each regressor in the model with strict preciseness. Certainly our example with the chemical process is a prediction problem. Reaction yield is important, and the engineer needs to predict it adequately.

Goal (2) above is relevant in a greater number of real-life applications than one might suspect. The model formulation is secondary; it is used merely as an instrument to detect the degree of importance of each variable in explaining the variation in response. Variables that are found to explain a reasonable amount of variation in the response are perhaps kept for further study. Regressors that seemingly play a minor role are eliminated. This practice often precedes a more extensive study or model-building process. A tobacco chemist, for example, conducts a taste-testing experiment with a panel in which a taste-type rating is given to each of several tobacco formulations. The regressor variables are the ingredient concentrations of the many additives put into cigarette tobacco. A model is fit to the test data with the sole purpose of determining, from a fitted regression model, which ingredients influence taste. Those that appear to play a role are retained as experimental variables for future studies.

Model specification explains itself. The analyst must take a great deal of care in postulating the model. Any analyst will acknowledge, either directly or obliquely, that various candidate models are often in competition, in different functional forms. Each functional form defines a different role for the regressor. When the model is linear, this type of exercise can be frustrating unless the complexion of each regressor variable is well defined in the data.

Parameter estimation is often the sole purpose of conducting a regression analysis in certain scientific fields. In Chapter 8, a data set is shown in which an agriculture production function is fit to a set of input regressors that represent expenditures. Six expenditure type regressors and a rainfall variable are used. The sampling unit is a year, and the data have been collected for the State of Virginia

for each of 25 years. These 25 data points are fit to a linear model, and prediction and variable screening are totally unimportant. However, specific ranges of the regression coefficients support (or refute) a particular economic theory. The *signs and magnitudes* of the coefficients are crucial.

1.3 THE DATA BASE

Much used (and perhaps overused) clichés in data analysis "Garbage In—Garbage Out" and "The results are only as good as the data that produced them" apply in the building of regression models. If the data do not reflect a trend involving the variables, there will be no success in model development or in drawing inferences regarding the system. Even if some type of association does exist, this does not imply that the data will reveal it in a clearly detectable fashion. Data may be produced from a designed experiment, a well-developed survey, a collection and tabulation over time, a computer simulation, or from one of many other sources. It should be clear that sample size is important. When the sample size is too small, the analyst cannot compute adequate measures of *error* in the regression results, and there can be no basis for checking model assumptions. However, the size of the data set is far from the only consideration. Many of the difficulties with data sets are obvious. For example, we cannot develop a model to produce a broad-based relationship if the sample of experimental units does not represent the population we are attempting to model. Broad generalizations regarding any of the goals of regression analysis cannot be made if the data are too specific. At times, even the ranges of the regressor variables in the data are such that the model built and conclusions drawn are *data specific*. For the oxygen consumption example, suppose all the individuals used in the experiment were well-conditioned athletes. Both the ranges of the variables and other characteristics of the inferences made would likely apply only to that population.

In many situations, difficulties with the regression analysis are a result of a failure of one or more assumptions. In particular, the multiple linear regression model of Eq. (1.1) is analyzed under the assumption that the regressor variables are measured without error. If excessive measurement error in the regressors exists, estimates of the regression coefficients can be severely affected and other inferences such as prediction, variable screening, etc. can be clouded with uncertainty. See Seber (1977).

Perhaps the most serious limitation in a regression data set is the failure to collect data on all potentially important regressors. This inadequacy may arise because the analyst isn't aware of all of the relevant regressors. Even if all or most of the quantities are identified, limitations in the data gathering process may prevent