On Strategies in Regression Analysis

by

Werner A. Stahel
Swiss Federal Institute of Technology, Zurich

Research Report No. 100
October

Seminar für Statistik
Eidgenössische Technische Hochschule (ETH)
CH-8092 Zürich
Switzerland

On Strategies in Regression Analysis

Werner A. Stahel
Swiss Federal Institute of Technology, Zurich
Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

October

Abstract. Data analysis remains a subject of practise and informal teaching. In the 1980's, there were several attempts to formulate ingredients of a theory of data analysis, in which John Tukey took an active part.

A branch of this field of statistical science consists of formalizing strategies for specified tasks. Such strategies can be used – in addition to training – for writing smart statistical procedures and programs for guided analysis. Furthermore, different strategies might be compared as to their success, as soon as there are evaluation criteria for the results – another aspect of the theory of data analysis.

This report focusses on sketching a strategy for multiple linear regression as a starting point. It is dedicated to John Tukey who gave his comments to a manuscript on it 17 years ago.

This version is based on a longer manuscript that dates back 17 years. I made a second version in 1989 and in March 2001, some editing in order to get ready for a workshop in Buenos Aires. The workshop was dedicated to the memory of John W. Tukey – and this was the reason to pick up this topic again. John had seen the original manuscript and contributed very valuable comments. Some of them are tacitly incorporated, whereas others will be marked explicitly.

The topic has not obtained too much attention in the meantime, and I did not have a chance to do much new research either except for some practical experience with the proposed strategy.

In the workshop in Minneapolis (IMA) in 1989, the participants discussed the need to start talking about the art of analyzing data in a more explicit and systematic way. We envisioned a meeting on the subject and found that one or a few papers would be needed to start discussions and determine the subject of such a meeting.

You may find that this work serves that purpose, or feel stimulated to bringing other approaches into the discussion.

As far as I am aware, there has not been much development of the topic since then. Any help to find newer material is welcome.

1 The Art of Applied Statistics

Traditional teaching of statistics and most of the statistical literature is centered on probabilistic models and the corresponding statistical procedures or on techniques directly. Applications usually appear as examples on which the methods are demonstrated.

In practice, however, a problem that is to be treated statistically will first be formulated in nonstatistical terms. To relate such a problem to an adequate statistical methodology and analysis is often a non-trivial task relegated to the "art" of applied statistics. The intention of this paper is to stimulate discussion of this relation.

Such a discussion should lead to new concepts and to "theories" on how to go from a non-statistical formulation of a problem to the selection of particular statistical techniques. In fact, the problem alone is found to be insufficient to guide the analysis. A number of other factors characterizing the "circumstances", from knowledge related to the problem to the human side of statistical consulting, bear their influence as well.

Potential Gains.

A set of concepts, relating the original problem to appropriate statistical methods, helps

- teaching the application of statistics;
- guiding non-professional statisticians to select adequate statistical methods;
- improving the quality and time efficiency of statistical analyses – also for experts;
- designing useful computing tools, including expert systems.

These aspects belong to the engineering side of statistics. Since professional statisticians have long neglected them, software engineers have devised their own tools for "knowledge discovery in data" and "data mining", which promise to work as push-button devices to solve all kinds of problems. Meanwhile, traditional statistical ways of dealing with problems, enriched by the many newer, more flexible statistical models and more explicit guidelines and recipes for their use bear a large potential to provide more appropriate answers to many problems. These methods are particularly more promising to the extent that they bring subject matter knowlegde about the problem into play.

Even though its large potential impact is clear, there have been few attempts to address the topic systematically. Some textbooks give an informal discussion. More formal contributions are those by Cox and Snell (1981) and Mallows and Walley (1981). Boen and Zahn (1982) address "The Human Side of Statistical Consulting". The textbook of Chatfield (1995) discusses many aspects of the "art" of statistical practise. (We have not followed up on this field of literature.)

At present, it seems hard to get beyond the stage of describing concepts, classifying, drawing plausible conclusions, and illustrating thoughts by examples – rather than defining concepts, proving relations, stating laws or theories and getting empirical evidence for their relevance or accuracy. In other words: We are at the stage of applying "soft science methods." This implies that most statisticians with some experience have their own opinions about the issues, even though they may never have attempted to conceptualize their perception and to make it accessible to discussion.

To this end, some notions or "vague concepts" are needed. The following descriptions of "*notions*" may illustrate this point.

2 Notions

A *statistical problem* consists of

- a *basic question*, described in terms of the "subject matter field" (that is, in non-statistical terms), or a list of such questions, preferably with priorities, and
- a potential source of data or a dataset.

It is commonplace that a statistical study should start by designing the collection of the data, and yet, too often decisions about the design cannot be, or are not, based on an overall view of the statistical problem. Note that a *data analysis project* may encompass several statistical problems – different basic questions, usually related to the same dataset.

The way in which the data is analyzed is also determined by the context and the circumstances. The *context* consists of any kind of qualitative or quantitative information related to the basic question or the origin of the data. The *circumstances* include the resources which are available for solving the statistical problem, the skills and characters of the researchers and statisticians involved, as well as any constraints to which they may be subject.

The statistical problem, the context, and the circumstances form the basis from which a data analysis starts. The course of the analysis is determined by the choice of the methods to be applied. The rules driving these can be grouped into three levels:

(A) The *basic approach* includes general decisions such as the selection of one or more fields of statistical methodology – for example, multiple regression.

(B) A detailed sequence of techniques is needed once the basic approach is clear. Good analysis often depends on characteristics of the actual data. A complete *strategy* includes rules for all kinds of data that are reasonably likely. We feel that good strategies should leave many decisions to the analyst, while providing sufficient information and guidance. They should also indentify points where the context and the circumstances may drive the decisions.

In order to facilitate the selection of a strategy, the questions for which a strategy is intended should be formulated in a generic, non-technical way. We call such a description the *purpose* of the strategy.

(C) The lowest level, called *recipes*, concerns details on carrying out the techniques and presenting their results in order to stimulate an appropriate reaction by the analyst.

In this paper, we concentrate on the second level and assume that the basic approach chosen consists of regression as the methodology, a classical non-Bayesian style, and a fairly high level of sophistication, and that the dataset is small enough to be handled in a single stage. Ideally, "regression" would include non-linear and generalized linear models as well as models for censored data and more. In our discussion, however, we have ordinary simple and multiple linear regression in mind, since this is the most extensively researched methodology in statistcs. Similar practical restrictions of the methodology covered may be needed when designing computer programs, even though they are undesirable from the point of view of the present theme.

3 Purposes

The purpose of an analysis, as seen by the potential users of the results, will usually require a sentence or a paragraph as an adequate description. Classes of potential purposes are determined below and are then related to statistical strategies. A more extensive text would describe these classes in terms that are readily understood by non-statisticians, since the aim of the classification is to guide less trained users of statistics.

Some relevant terminology here includes the distinction between explanatory variables and carriers. The multiple linear regression model reads

$$Y_i = \sum\nolimits_j \beta_j X_i^{(j)} + \varepsilon_i,$$

where the *carriers* $X^{(j)}$ depend on (may be identical with) the observed *explanatory variables* $U^{(j)}$,

$$X_i^{(j)} = g_j(U_i^{(1)},\ U_i^{(2)}, \ldots),$$

and the *response* Y may be a monotone transformation of an originally observed *target variable*, $Y_i = h(\widetilde{Y}_i)$ (if the purpose permits). The functions h and g_j determine the *functional form* of the model.

The following classes of purposes are subdivided and described in more detail in Appendix A.1.

(a) Modelling.

One of the basic tasks of the sciences is to find "*laws*" and mechanisms that explain the phenomena we observe. For a specific target variable $\widetilde{Y}$, we want to find "explanatory" variables $U^{(j)}$ that determine its outcome, and the functional form of this relationship. The error term ε_i may be assumed to represent an intrinsic randomness or the effects of relevant variables we failed to detect or include, even if the intention is to find all such variables.

Finding a law can be presumptuous and surely requires more than just a thorough analysis of a single dataset. If little is known about dependencies of the target variable on any potential explanatory variables, it may be enough to find some of the explanatory variables that influence the target variable and to get a rough idea of their relative importance.

(b) Prediction.

The model may be used to *predict* the value of the target variable from known values of the explanatory variables with or without pretending that the model represents an underlying mechanism.

(c) Optimization.

In industry, regression methods are frequently used to find the conditions for producing more or better products.

(d) Estimation and testing.

In these applications, it is clear which explanatory variables are of interest, and we want to draw some inference about the coefficients related to them. The equational form may be known from relevant theory in the subject matter field or from earlier statistical studies, which had model building as a purpose. A typical testing problem is the assessment of a treatment effect.

Apart from the carriers corresponding to the coefficients of interest, the model may not be completely specified. It is important to note that a coefficient β_j changes its value and its interpretation if carriers correlated with $X^{(j)}$ are included in or excluded from the model.

(e) Regression as auxiliary method.

This heading contradicts the intention of classifying the main purpose of the data analysis. It arises because we chose regression – which is a methodological category – as the scope of our discussion. But similar practical restrictions arise when designing courses or computer packages, so these lower level purposes are worthwhile mentioning here. We distinguish "*data reduction*" and "*filtering*" and discuss these terms in Appendix A.1.

What should be called a "purpose"?

The classification of purposes just described and the finer subdivision in Appendix A.1 are two attempts to cope with the tradeoff between too many classes and too

little homogeneity within the class to be useful to guide the strategy. A "purpose" should

- be described in non-statistical terms (which needs more space than we have used here),
- have some specific consequences for the statistical strategy to be applied, which should be distinct – at least partly – from those of all other "purposes", and
- be frequent enough to be worth separating.

It is not required that the purpose define the strategy by itself. Some other determinants have been and will be mentioned.

The fact that there are borderline cases should not stop us from formulating and classifying purposes. Hopefully, in such cases the differences in selected strategies would not affect the results seriously. Moreover, a basic question may give rise to more than one purpose suitable for guiding strategy. The main purpose is relevant, but the strategy need not be tuned to that purpose in a narrow-minded sense.

4 A Taxonomy of Exhibits

A large number of diagnostics for regression models have been proposed over the last thirty years. They include graphical displays as well as tables and other kinds of information that the analyst will examine during the course of an analysis. The information can be divided into blocks, which we will call *exhibits*. An exhibit is determined by the method used to *generate* the information contained in it and the way it is *presented*.

Remark. Specifically, most exhibits proposed for least squares regression can be adapted to robust fitting methods, thus changing the generation part and maintaining the presentation (although research is needed in this area). While some authors suggest that the use of diagnostics is an alternative to robust fitting, we conjecture that when there are multiple deviations from the classical assumptions, diagnostics must be based on robust estimates to be effective. Such robust diagnostic displays differ from their classical versions in the method used to generate the information, but can be identical in the way it is presented.

Lists of exhibits which are useful in multiple linear regression are given in Appendix A.2. They can be classified and summarized as follows. **(a) Preliminaries.**

Documentation of the data set; impossible values; "first aid" transformations; outliers. **(b) Inference for a fixed set of carriers.**

Coefficients; test statistics; confidence regions; region of validity; residuals. **(c) Diagnostics for stochastic part.**

Error distribution; heteroscedasticity; correlation; influential observations. **(d) Diagnostics for structural part.**

Non-linearity; transformations; lack of fit; collinearity. **(e) Model selection.** Characteristics for many possible equations. **(f) Selection of estimation method.** See Appendix A.2.

5 Strategies for two Purposes

Different purposes call for different emphases on the exhibits just discussed and on the associated steps of analysis. This contention is exemplified here; general considerations follow in later sections.

5.1 Finding the variables that matter

To fix ideas, assume that there are 100 to 500 observations on 10 to 30 explanatory variables and a response, and we want to find the variables that matter. The following difficulties arise:

- Any tentative model may have "multiple deficiencies" that obscure each other: An incomplete model gives rise to a distorted picture of the residuals. A wrong expression (transformation) of the target variable can mask the effect of an explanatory variable or make it necessary to include interaction or other second order terms.
- There are cases where two or more additional explanatory variables improve the fit remarkably, whereas none of them shows up individually as useful. Thus, backward elimination is safer than forward selection (if there are more observations than carriers).
- Since there are so many possible models, part of the search through them should be automated if possible.

Such considerations characterize the paradigm of John Tukey's Exploratory Data Analysis. It shaped the thinking of his students and coworkers in the 1970s and 80s.

A strategy taking these points into account consists of the following steps:

1. Data screening. Check for impossible values and univariate outliers.
2. "First aid" transformations (Mosteller and Tukey, 1977). Based on subject matter and general statistical knowledge and on marginal distributions, transform the variables in a plausible way.
3. A long equation. Fit an equation containing probably too many variables. This may be done by
   - assembling all variables that seem likely to influence the response according to prior judgement;

- taking all variables, if the number of observations is at least somewhat larger than the number of variables, thus leaving a reasonable number of degrees of freedom for the residuals;
- applying a stepwise forward procedure with low entry criterion.

4. Diagnostics for the stochastic part. Check for outliers in the residuals, non-normal error distribution, heteroscedasticity, correlations of errors. It may be necessary to transform the target variable, to introduce weights, to use robust estimators (which we recommend in any case), or to introduce blocking variables or refined fitting methods if the errors are correlated. Note, however, that the assumptions are not very crucial for the given purpose.

5. Nonlinearities. Examine residual plots (Appendix A.2, d1) to decide on transformations of the explanatory variables or inclusion of squared terms or model the transformation g_j by regression splines and the like.

6. Automated model selection with "all"-subsets techniques. They will usually produce several acceptable equations, reflecting partly problems with collinearities. (It is a shame that such techniques have not yet been extended to treat nominal explanatory variables adequately!)

7. Add variables. Plot residuals against explanatory variables not in the equation, including those that have been eliminated in Step 6.

8. Influential observations. Find multivariate outliers in the space of the carriers included in (one of) the equation(s). Use the methods described in c4-5 of Appendix A.2 or robust multivariate techniques.

9. Subject matter criticism. If the equation contains terms which are implausible or coefficients with the "wrong" sign according to subject matter knowledge, try to discover variables for which they may be surrogates, or suppress them, unless the latter deteriorates the statistical performance untolerably.

10. Goodness of fit. Compare the residual mean square with an external or internal estimate of "intrinsic" error variance (see App.A.2, d3). If satisfactory, it may be granted to go to Step 13.

    Comment by JWT: The result may help to modulate some conclusions, but should rarely have a major influence on the course of the analysis.

11. Interaction terms. For the carriers which have high influence on the response, check if there are interactions – among themselves or with others. Interactions among non-influential variables, though perfectly possible, are not common and undesirable according to Cox and Snell (1981, p.126). Therefore, they are of secondary priority to be explored. Interactions suggested by subject matter knowledge should be included in Step 3. The interactions may be introduced as new potential carriers and studied by means of residual plots. More time-efficient methods would be useful.

12. If the equation(s) has (have) been noticeably amended since Step 5, go back to that step, or even to Step 4.

13. Test for inclusion of any variables that have been suppressed in Step 9 and examine respective residual plots.

14. Collect the models that have been found to adequately describe the data. (See e3 of App.A.2.)

This is a rough draft of a strategy. A complete description would include a sequence of exhibits to be examined and rules for the necessary decisions. A different strategy for the same purpose has been sketched in a working paper by Tukey and Velleman in 1983.

5.2 Testing a hypothesis

Assume that the coefficient of a prespecified carrier – $X^{(j)}$, say – is to be tested against a pre-spaecified value (usually zero). For this task to be meaningful, the explanatory variables must be categorized into three groups: those that

(a) must be included in the model,

(b) must not be included in the model,

(c) may optionally be included in the model.

The analysis may then proceed along these steps:

1. Data screening, as above.

2. If applicable, develop model. Follow Steps 3 through 9 of the strategy for "finding the variables that matter," avoiding variables of group (b), caring only about those variables of group (c) that contribute substantially to the multiple correlation or that are involved in a collinearity problem with the carrier to be tested. Whether it is possible to transform the target variable, the carrier to be tested, or the variables of group (a), depends on the basic question.

3. Final examination of stochastic part: check assumptions (step 4 of 5.1) and influential observations (step 8 of 5.1) carefully. Amend the testing procedure if necessary.

4. Determine carriers of group (c) in the final model which suffer from a collinearity with the carrier to be tested ("critical" carriers).

5. Calculate the test result or results for models with and without critical carriers, if applicable. If necessary, discuss the different possible models and their implications for the basic question.

6 Classification of Purposes and Strategies

Not all the purposes of analysis listed in Appendix 1 lead to noticeably different strategies. We found that four different strategies correspond to four categories of purposes defined by two distinctions:

i. Model is given vs. to be developed.

ii. Emphasis is on the variables and their parameters vs. on the fit (estimated $E(Y|X)$).

Table 1 shows how the purposes of Appendix A.1 fit into these four classes. It becomes clear that the purposes "optimization" and "prediction" may lead to quite different strategies depending on the subject matter knowledge available, i.e., depending on how much is known about the model. Also, "prediction" had to be split up into "interpolation", that is, prediction for future observations with explanatory variables within the range covered by the dataset, and "extrapolation", when future observations are outside this range.

The two distinctions should be two continuous "dimensions". Along dimension "Model" of the table one might mark the "milestones" (based on a personal comment by J.W. Tukey)

– everything open;

– variables given, but functional forms open;

– equation given, but coefficients open;

– equation and coefficients given.

In between these situations, the open choices may be subject to slight to strong preferences. Also, parts of the model may be fixed and other parts open, as sketched in Subsection 5.2.

The emphasis the strategy should reflect in each of the four cases is on

(A) model building techniques: transformations, search for several equations describing the data adequately;

(B) collinearity analysis, diagnostics for stochastic part;

(C) same as in (A), but one reasonably good equation is sufficient; influential points (robustness), range of interpolation;

(D) analysis of residuals, influential points (robustness), range of interpolation (for the first 3 purposes mentioned).

		(i) Model	
		to be developed	known
(ii) emphasis on	equation	"*Case A*" Variables that matter Mechanism Optimization I Extrapolation I	"*Case B*" Test of specified hypothesis about coefficients Fitting Extrapolation II
	fit	"*Case C*" Interpolation I Outlier Identification	"*Case D*" Optimization II Interpolation II Calibration Reconciliation Data Reduction Filtering

Table 1: Classification of Purposes

7 Criteria for Strategies

The two strategies described in Section 5 have been derived from folklore, limited experience, and intuition. It is clear that alternative strategies, like the one proposed by Tukey and Velleman in the unpublished paper mentioned above for the first case, may be superior. In order to judge this, criteria are needed. Here are some points to be considered.

1. Reliability in controlled situations: When the right answer is known from external sources and the data is adequate, the strategy should find it.

2. Flexibility or range of application: The strategy should be adjustable to a reasonably wide range of datasets, "amounts" of relevant context, circumstances, and maybe purposes.

3. Economy: The effort to get a good answer should be low. The most economical strategy may depend considerably on the computing environment.

4. Adequacy for the purpose: The strategy should give an answer to the basic question unless either logic or the data suggests that the question is ill-posed.

5. Adequacy for the audience: The answer should be given in terms that are understandable to the users. The appropriate interpretation of the results should be facilitated.

The first three relate to a strategy as such. The last two are criteria for a successful data analysis for a specific statistical problem (Section 2). They may give support

to or discredit the strategy used, and are therefore more relevant, but less precise measures of the quality of strategies.

8 Conclusions

We started out to relate purposes to methods, or even to formulate purposes in a language that can be understood by non-statisticians, and to give rules that help them to select an adequate strategy for analyzing their data. It is clear that we have merely scratched the surface of such a task. The preceding considerations nevertheless allow us to draw some tentative conclusions.

The purpose alone is not enough to determine a strategy. Context and circumstances are also relevant, on the "second level" presently discussed as well as on the "basic approach" level. This is clear from the last two sections.

Context is crucial, for example, when a mechanism or law is wanted. Candidate equations that are compatible with the data will be discriminated on the bases of compatibility with earlier studies, plausibility, and reasonable behavior for extreme conditions, all of which can be subsumed under the term "generalizability".

A relevant characteristic of the dataset is the number of variables and its relation to the number of observations. A high number of variables (without a substantial relevant context) leaves only a few realistic purposes, and the respective strategies are similar. On the other extreme, in simple regression, an exhaustive analysis that gives the answers to most potential purposes is feasible. In both cases, formulation of a purpose may not be necessary, but it is not harmful either.

As to circumstances, the time budget is clearly important. In the academic environment, there is a temptation to do everything always (or rather: to apply one's preferred methods always). With the vast number of methods and diagnostics available for regression, however, it is unavoidable to make choices for virtually any time budget. This may be necessary at all levels – exploring methods, programming and including them in packages, and applying them to data.

In addition to economical constraints, some more circumstances can have a crucial impact on the choice of a strategy, including

(a) the computing environment;

(b) the audience (Sec. 7);

(c) the quality of the design and the data;

(d) the importance of the basic question;

(e) the availabilty of subject matter knowledge;

(f) the statistical tradition of the subject matter;

(g) the level of understanding of all the people involved in the analysis and of the ultimate audience, as well as the school of statistical thinking in which they have been educated; and

(h) the personalities of statistician(s) and researcher(s) and their ease of communication.

The classes of purposes to be described (in the sense of Section 3) are determined by two ends: They should be problem oriented, but they should also be helpful in determining the strategy. (Remember that "prediction" had to be split into "interpolation" and "extrapolation" for methodological reasons.) The analogy to a user-machine interface in computer science suggests that the methodological side should be subordinate to the problem side.

Clearly, there is no single best strategy for analyzing a given dataset. But the quality of analyses of statistical data can certainly be improved by making concious decisions about the strategy, keeping in mind purpose, context, and circumstances.

A.1 Detailed Description of Purposes

This Appendix lists the purposes on a more detailed level. For the sake of reference, we will refer to examples that can be found in the books by Daniel and Wood (1980), abbreviated by "DW", Draper and Smith (1981), by "DS", and Weisberg (1980), by "W."

(a) Modelling

(a1) Variables that matter, descriptive model. Find the explanatory variables that might be needed for a model that would satisfy either one of the three following purposes (b-d). If there are few explanatory variables or abundant and precise enough data, one may want to find even the functional form of the relationship.
Examples: Traffic accidents (DS Exercise 3K); Fuel consumption (W Examples 2.1, 3.4).

(a2) Mechanism, law. Find a functional relationship describing a plausible mechanism which explains an observed value of the target variable from the corresponding value(s) of the explanatory variables, or a relationship that generalizes to a number of situations or datasets.
Examples: Pressure and boiling pointg (W Problem 2.1); Development of ice crystals (DS Exercises 1R, 5V).

(b) Prediction

(b1) Interpolation. Find a rule to predict the target variable for future cases on the basis of the values of the explanatory variables. Future X values are within the range of the X values of the present dataset.
Example: Old Faithful (W Example 9.1).

(b2) Extrapolation. Find a rule to predict the target variable from X values that may be outside the range of the present dataset's X values. Note that predictions in time series should not be considered here – even if regression methods may be used in such contexts – since time series methods represent a different field of statistical methodology.

(b3) Calibration. A measurement of Y which is difficult or unfeasible under circumstances is to be replaced by a surrogate measurement X. This may or may not involve an inversion of the relationship given by the regression model.

Calibration may be seen as a special case of prediction. However, we know the variables involved (usually a single X) and there is a pronounced interest in the "prediction" error.
Examples: Cup and bottle loss (DS Exercise 1H); Pressure and boiling point

(W Problem 2.1; cf. 32b); Subjective estimation of the number of snow geese (W Problem 4.6).

(b4) Reconciliation. "Solve" an overspecified system of equations (more equations than unknowns). In the presence of calibration difficulties, for example, one may measure differences between pairs of objects. It is then desired to make all the differences consistent. Land surveys represent another instance of this purpose.

(c) Optimization

Find the X values that maximize or minimize the target variable (within a range of possible values of the explanatory variables), or at least improve its value, e.g., find best fertilizer for a crop, find better settings of a chemical reactor to get higher yield of the desired product. Tukey adds that the target variable may need to account for cost associated with different values of the X settings.

(c1) Optimization I. Find the variables that have a relevant effect on the target variable – similar to (a1).
Examples: stack loss (DW p. 71).

(c2) Optimization II. Find the optimum on a "response surface" determined by clearly identified explanatory variables.

(d) Estimation and Testing

(d1) Test of a specified hypothesis. Is there a treatment effect of known sign? Is there an effect of a certain explanatory variable (or several variables) in the presence of other specified variables? Test a parameter (or several) in a given model. The parts which do not involve the parameter(s) of interest may not be completely specified beforehand in practise – a fact that can have severe consequences on the interpretation of the results.
Examples: clinical trials (with covariables); effect of cloud seeding (W Example 7.1); sex differences in salaries (DS Exercise 6H); "Strong interaction" (W Example 4.1, Problem 4.2); Control charts (W Problem 4.4); Twin data, (W Examples 7.2); Rat data (W Example 5.3).

(d2) Fitting a model, estimating a coefficient. For a known model, estimate the coefficients. Again, a part of the model not involving the parameters of interest may be unknown in practise.

(e) Regression as an Auxiliary Method

(e1) Data reduction. If there are groups of data that can be summarized by fitting a model to each group, the parameter estimates may serve as a basis for further

data analysis.

(e2) Filtering. We may try to "filter out" the obvious effects of regression variables X and Y in order to find less obvious structure in the residuals from such a fit. As a special case, we may be interested in identifying any outliers, either for their special interest or for data screening.

A.2 Exhibits

Note. Generally, scatter plots should be fortified by a smooth of the variable shown vertically if more than about 30 points are plotted.

(a) Preliminaries

This group of exhibits are not specific to regression. They are useful in virtually all data analyses except for those of very small data sets. They include

(a1) Documentation: dataset or file name, its history; date; amount of data; printout of data for all observations or at least for the first and last one(s).

(a2) Other information that allows to check if the right data has been entered: format; range of variables.

(a3) Diagnostics for transformations: Skewness, tests for power transformation (BoxGC64Box and Cox 1964, Atkinson 1973, 1982 [with comment by J. W. Tukey]); normal plots; histograms;

(a4) Diagnostics for univariate outliers: Outlier test statistics.

(a5) Diagnostics for multivariate outliers: (Robust) Mahalanobis distance from (robust) center, numerical and graphical displays; scatterplot matrix of (original or transformed) variables or (robust) principal components.

(b) Inference for a fixed set of carriers

For a fixed set of carriers, the results of a regression analysis are captured in the following exhibits:

(b1) Equation, estimated coefficients: Identification of variables or terms (carriers); estimated coefficients; confidence intervals (they are more user friendly than the standard errrors, t values, and p values that are usually shown).

(b2) estimated standard deviation of error with degrees of freedom; multiple correlation; (F-) test for global hypothesis (all coefficients besides intercept are zero). (The usual ANOVA-table is not very helpful.)

(b3) Test statistics: t- or F-values (or analogous statistics) and corresponding p values for the deletion of individual terms or of prespecified blocks of terms.

(b4) Test for contrasts or other prespecified linear hypotheses.

(b5) Correlation or covariance matrix of estimated coefficients (see also d4).

(b6) Region of validity: Means and standard errors of carriers and their correlation or covariance matrix (or other characterization of their marginal and joint distribution).

(b7) Residuals and/or fitted values; standard errors of fitted values.

(b8) Prediction regions for specified values of explanatory variables.

(c) Diagnostics for stochastic part

There are numerous diagnostics that point to failures of the assumptions for the error term. We include several new suggestions.

(c1) Error distribution: table of ordinary, standardized, or studentized Weisberg, 1980, p. 105 residuals; normal or half normal (Atkinson 1982) or detrended normal plot; test statistics for normality, like the Durbin-Watson statistic; diagnostics for transformation of Y (see (a3)); outlier test statistics.
Comment by JWT: These diagnostics actually serve a variety of purposes. In particluar, they may suggest – or demand – improvements of the structural part of the model (see (c3), end).

(c2) Heteroscedasticity: Plot of ordinary, standardized, or studentized residuals against fitted values (Tukey-Anscombe plot); plot of residuals against actual or potential weight variable(s); scatter plot of the rank or the square root of the absolute residuals on the explanatory variables (or actual carriers); Cook and Weisberg (1982) give references and a more general method based on a score test.
Comment by JWT: All these may be clarified by inclusion of a formal fit, or by smoothing before plotting.

(c3) Correlation of errors: Plot of residuals against sequence number or time variable; tests for serial correlation; plots of residuals on grouping variables (this may show variance components in the error term); geographical plot of residuals (if applicable).
Comment by JWT: A more effective approach in some circumstances involves identifying the 3 nearest neighbors (in x-space [schould be time or geographical

space in this context, WSt]) of each data point and a random sample of 3 non-nearest neighbors. For each pair (selected point, neighbor or non-neighbor) calculate the absolute difference of the corresponding residuals, and display, comparatively, the ("letter values" of the) resulting distributions.

(c4) Influential observations for a single carrier ($X^{(j)}$): Plot of residuals against residuals from a regression of $X^{(j)}$ on all the other carriers ("partial regression leverage plot", Belsley, Kuh, and Welsch 1980, p.30); table of DFBETA$^{(j)}$ (Belsley, Kuh and Welsch 1980, p.??).

(c5) Globally influential observations: Plot of residuals against diagonal (h_u) of projection matrix (Williams in the Discussion of Atkinson 1982); Tables of the H_u's, Cook's distance Cook, 1977, $\| \hat{\beta}(i) - \hat{\beta} \|_x T_x$ Weisberg, 1980, p. 107, several functions of these values Belsley, Kuh and Welsch, 1980, Chap. 2.1, Daniel and Woods' WSSD Daniel and Wood, 1980, p. 59 and 129 (for an overview, see Hocking, 1983); half normal plot of Cook's distance Atkinson, 1982.

(d) Diagnostics for structural part

A number of diagnostics are used to check if a tentative equation should be modified by transforming, adding, or deleting carriers.

(d1) Non-linearity: Plots of (unstandardized) residuals against those independent variables or carriers that are in the tentative equation. Some authors prefer to add the "component effects," $\hat{\beta}_j \cdot (X^{(j)} - \bar{X}^{(j)})$, to the residuals and plot the sum against the carrier $X^{(j)}$ ("partial residual plot," see Larson and McCleary 1972, Landwehr 1983). Note that residual plots may be ineffective because other carriers in the equation may mask the non-linearity. Research is still needed to overcome this defect.

(d2) Test for power transformations of independent variables: cf. Box and Tidwell (1962).

(d3) Lack of fit information: Test for comparing the estimated error variance with an external value; estimation of the error variance from "near neighbors" in some sense Daniel and Wood, 1980, Ch. 7.5. Compare also comment of JWT to (c3).

(d4) Collinearity: Multiple correlation coefficient between each carrier and all the other carriers; singular value decomposition of the X matrix and condition numbers Belsley et al., 1980, Ch. 3.

(e) Model selection

There are two classes of automated model selection techniques. The relevant exhibits include the following.

(e1) Stepwise procedures: Tables of partial correlations with respective test statistics, coefficients and other items mentioned in b1 and b2 – they may be given at each step or organized into "summary tables"; plot of the (adjusted) multiple correlation coefficient (or other criteria, like C_p) or Akaike's information criterion on the number of carriers in the equation. Tukey suggests "mean square divided by degrees of freedom".

(e2) "All" subsets procedures (in fact, the procedure may not examine all the possible choices of sets of variables explicitly): table of equations with respective values of the criterion (e.g., C_p); Plot of the criterion on the number of carriers ("C_p plot," Daniel and Wood 1980, Ch. 6.2).

(e3) Comparative summary of several models.

(f) Selection of estimation method

While the ordinary least squares procedures are essentially unique, robust and ridge regression methods require the choice of one or several constants. These exhibits are listed for completeness even if their value seems questionable to the present author.

(f1) robustness constants: Plot of estimated coefficients against robustness constant Denby and Mallows, 1977.

(f2) ridge constant: Plot of "ridge trace" against ridge regression constant Hoerl and Kennard, 1970. Tukey suggests plotting the logarithm of the ridge trace in order to avoid spurious "elbows".

References

Atkinson, A. C. (1973). Testing for transformations to normality, *J. Royal Statist. Soc. B* **35**: 473–479.

Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables, *J. Royal Statist. Soc. B* **44**: 1–36.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics*, Wiley, N. Y.

Boen, J. R. and Zahn, D. A. (1982). *The Human Side of Statistical Consulting*, Lifetime Learning Publications, London.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *J. Royal Statist. Soc. B* **26**: 211–52.

Box, G. E. P. and Tidwell, P. W. (1962). Transformations of the independent variables, *Technometrics* **4**: 531–550.

Chatfield, C. (1995). *Problem Solving: A Statistician's Guide*, 3rd edn, Chapman & Hall, London.

Cook, R. D. (1977). Detection of influential observations in linear regression, *Technometrics* **19**: 15–18.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall.

Cox, D. R. and Snell, E. J. (1981). *Applied Statistics: Principles and Examples*, Chapman and Hall, London.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, 2nd edn, Wiley, N. Y.

Denby, L. and Mallows, C. L. (1977). Two diagnostic displays for robust regression, *Technometrics* **19**: 1–13.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, 2nd edn, Wiley, N. Y.

Hocking, R. R. (1983). Developments in linear regression methodology 1959-1982, *Technometrics* **12**: 55–67.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems, *Technometrics* **12**: 55–67.

Landwehr, J. (1983). Using partial residual plots to detect nonlinearity in multiple regression.

Larson, W. A. and McCleary, S. J. (1972). The use of partial residual plots in regression analysis, *Technometrics* **1**: 781–790.

Mallows, C. L. and Walley, P. (1981). A theory of data analysis?, *Proc. Bus. Econ. Statist. Sect., Amer. Statist. Assoc.* pp. 8–14.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: A Second Course in Statistics*, A.W. Series in Behavioral Science:Quantitative Methods, Addison-Wesley, Reading.

Weisberg, S. (1980). *Applied Linear Regression*, Wiley, N. Y.