Choice of order in Regression Strategy

Julian J. Faraway
Department of Statistics, University of Michigan,
Ann Arbor, Michigan 48109, USA

Abstract

We view the pursuit of an appropriate model by the statistician in the following way: The desired model (or models) should satisfy several requirements, unimportant variables should be excluded, outliers identified, etc. The methods of regression data analysis such as variable selection, transformation and outlier detection, that address these concerns are characterized as functions acting on regression models and returning regression models. A model that is unchanged by the application of any of these methods is considered acceptable. The ordering of the methods is studied. A method for the generation of acceptable models supported by all possible orderings of the choice of regression data analysis methods is described with a view to determining if two capable statisticians may reasonably hold differing views on the same data. The consideration of all possible orders of analysis generates a directed graph in which the vertices are regression models and the arcs are data-analytic methods. The structure of the graph is of statistical interest. The ideas are demonstrated using a LISP-based analysis package The methods described are not intended for the entirely automatic analysis of data, rather to assist the statistician in examining regression data at a strategic level.

1 INTRODUCTION

Textbooks on linear regression have several chapters, each devoted to one particular aspect of building a regression model and checking its adequacy. One chapter may study variable selection and another diagnostics for the detection of outliers. If these may be viewed as tactics in the pursuit of a regression model, what of the strategy? Very little is said about how these various techniques fit together, other than that it is a skill gained by experience. Daniel & Wood(1980) is a notable exception in this respect. There are outstanding questions concerning the interaction between these tactics and the order in which they should be carried out. In this paper, we will look at, not a particular specific of regression analysis, but the process as a whole.

Regression strategy is usually discussed within the context of expert systems for Statistics, but production of such systems for the automatic analysis of data has been stalled primarily by the difficulty of integrating the real-world context of the data. Nevertheless, this does not preclude worthwhile study of regression strategy. This article does not describe an expert system and the tools discussed are not intended for the completely automatic analysis of regression data. These methods should be regarded in the same way as the usual tools of the regression analyst, such as Box-Cox transformations. The difference is that they are designed for strategic, not tactical, application. Just as the tactical tools require the supervision of a statistician for appropriate use, these strategic tools are not meant to be blindly or automatically applied. Gale(1986) and Phelps(1987) contain several articles in which expert systems and statistical strategies are discussed.

We view the pursuit of an appropriate model by the statistician in the following way: The desired model (or models) should satisfy several requirements: unimportant variables should be excluded, outliers identified, variables appropriately transformed, etc. An initial model is proposed and these requirements are checked sequentially by numerical or graphical methods and if necessary the model is changed in a way suggested by the particular method. For example, variable selection methods can detect redundant variables and propose their elimination. When the current best model satisfies all the requirements the analysis ends. The choice of requirements and methods used to test these and make appropriate model changes is made by the statistician. We take particular interest in the order of application of the methods and the influence of individual data points on the final model chosen. We must automate the methods used due to the amount of repetitive analysis required, which leads to two main difficulties with the tools we propose. Graphically-based methods are difficult to automate because they rely on human perception and because the automated methods are context-free, the statistician must examine the results to determine the sense or lack of it. These tools are not intended to replace the standard analysis - they are an additional aid and the results should be interpreted with due care.

Regression analysis is influenced by the taste of the statistician. The choice and ordering of methods are not universally agreed upon and so it is possible that two experienced analysts will construct different valid models and come to different conclusions from the same data. If one was aware of this, then one would hesitate to seize upon one conclusion and discard the other, rather one might say that the data do not support any strong conclusion. However, it is quite possible that the first two statisticians may agree and a third disagree, or a fourth or a fifth. Given that the number of reasonable analyses is likely to be large, one is unlikely to have the resources to collect all these opinions. Often, one will be the sole analyst so that it will be difficult to know if the data support many or only one conclusion.

Lubinsky & Pregibon(1987) discuss the search for "good" regression models although their methodology and motivation differ from ours. See also Brownstone(1988) and Adams(1990).

We show a way, given a particular choice of methods, of generating all of the acceptable models arrived at by different orderings of the methods chosen. Thus the statistician may discover if there are several competing models for the data which support different conclusions, in which case suitable doubt may be expressed, or that only one model is indicated, whence the conclusion may be infused with greater confidence.

Regression analytic methods are sometimes quite inexact, perhaps depending on the statistician's interpretation of a plot, but since a large number of possible regression analyses need to be considered, these methods need to be exactly specified. This we do in section 2, so that they may be programmed. In section 3, we discuss the generation of acceptable models derived by the various orderings of the methods. We also address the question of which of the generated models is best in section 4.

2 CHARACTERIZING REGRESSION DATA ANALYSIS

The process of building a regression model may consist of several stages such as outlier detection, variable selection and transformation which we shall call regression analytic procedures (RAPs). At each stage there is a candidate model that may be supplanted by another model according to the result of the RAP. A regression model might be specified by the original data, the functions specifying the link between the predictors and the response and weights on the observations. Of course, a richer formulation of regression models is possible, if not desirable, but for simplicity we will proceed with this.

We wish to characterize RAP's as functions acting on regression models and returning regression models. For some procedures such as variable selection and transformation, this is relatively straightforward, but for other diagnostic based methods depending on graphics, it is not so easy. Also the physical context of the data can be included to a limited extent by restricting the RAP's from transforming or eliminating certain variables or points but it is difficult to include knowledge concerning which functional forms are more appropriate than others. Thus, the RAP's provide only an approximation to a regression data analysis by a human but the information provided by this approach is additional to that given by a standard analysis, not a replacement, so nothing is lost and much may be gained.

When constructing a regression model, we have certain requirements for what is an acceptable final choice - for example, that there be no redundant predictors, that the expected response should be linear in the predictors, that there be no outliers included in the model etc. The RAP's are a response to these requirements in that they examine a candidate model and if necessary change that model to make it acceptable with respect to that particular requirement. Thus, a minimal list of RAP's is determined by the requirements we wish our model of choice to satisfy. An acceptable final model would be one which is not changed by the application of any of the RAP in the list.

The following RAP's have been programmed. The names for the methods are given brackets for reference. Check and remove outliers (outlier-test), check and remove influential points (test-influence), check for non-constant variance and reweight if necessary (hetero-test), check for and apply a Box-Cox transform on the response(box-cox-test), check for transformations of the predictors (tran-predictors), perform variable selection using the backward elimination method (bw-elim), perform variable selection using the forward selection method (fw-sel) and restore points previously eliminated that are not now outliers, but may be influential (restore-points).

This list is obviously not exhaustive, but is representative of the sort of data-analytic actions that may occur in practice and is appropriate for some common requirements for acceptable regression models. I certainly do *not* claim that these are the best methods to use all the time only that the ideas that follow are not restricted by these particular choices.

These functions have been programmed to take any regression model as input, and output a (possibly changed) regression model. The flexibility of Lisp and the object-oriented programming system that comes with LISP-STAT makes it easier to program these functions in full generality to keep track of the numeric (the data and weights) and non-numeric (the transformations and variable names) components of the model.

3 GENERATION OF ACCEPTABLE MODELS

In this section, we consider the generation of acceptable models by changing the order in which RAP's are applied. A natural, although arbitrary, choice of initial model is the regression of all possible predictors on the response with unit weights on the observations. Clearly, there are situations when there will be several reasonable initial choices, but for now we will use only that one initial model. Our ideas are not affected by this restriction.

First we should select a list of RAP's considered appropriate for the particular problem. Starting with the initial model, we apply each of the RAP's individually. Some will cause no change, but others may return different models. To each newly found model, we apply each of the RAP's until no new models are generated. The total analysis may be viewed as a directed graph, where the vertices are models and the arcs linking the vertices are RAP's that resulted in a change. Loops indicating RAP's which had no effect on a particular model could be explicitly drawn, but can be implicitly assumed. The absorbing vertices (having out-degree zero) in the graph are acceptable models. This approach to generating the acceptable models is much more efficient and comprehensive than simply generating all combinations of the list of RAP's and then applying them sequentially because much redundant calculation is eliminated, particularly since an RAP may occur more than once in a given path.

The work of Lubinsky and Pregibon (1987) differs from ours in that their graph is strictly a tree where the nodes of the tree are *features* of regression models and the arcs indicate the result of tests for the presence of such features. The strategy (ordering of the tests) is determined by calibrating the order of analysis on known data, and is then fixed. The user then chooses a model (or models) from the tree trading the simplicity of the model against its accuracy. We produce all the acceptable models determined by a particular fixed choice of RAP's and leave the user to choose amongst them. We do not fix the order.

Now, certainly some of these analyses might consist of sequences of RAP's that no statistician would ever try in practice, but this of little consequence since the final model is the ultimate subject of interest. The statistician should examine these models to ensure that they are physically sensible and discard those that are not. The difficulty is not that unreasonable models will be included amongst those considered, since the statisticians can easily screen these out, rather that important models will not be discovered due to the inflexibility of the RAP's. Thus we do not claim that this method will produce all reasonable models, but it may well find some that otherwise might have been missed.

We shall use several datasets in the following discussion:

The Galapagos dataset: 29 cases being islands, 5 geographic predictors (area, elevation, distance to nearest island, distance from Santa Cruz island and area of adjacent island) and number of species as the response, described in detail in Andrews & Herzberg (1985).

The Chicago dataset: 47 cases being zip codes in Chicago, 5 socio-economic predictors (% minority composition, fire rate, theft rate, age of housing and income) and no. homeowner insurance policies is the response, described in detail in Andrews & Herzberg (1985).

The Swiss dataset: 47 cases being provinces in 1888 Switzerland, 5 socio-economic predictors and a standardised fertility measure is the response. Described in Mosteller & Tukey (1977).

Some data will generate several possible models, others only one. For example, using the RAP's outlier-test, test-influence, box-cox-test, tran-predictors, bw-elim, fw-sel and restore-points, a digraph depicting the analysis of the Chicago dataset is shown in Figure 1. The initial model chosen was the one with all cases and predictors included and all variables untransformed. The Chicago dataset analysis considers 28 models in all, of which 3 are acceptable , whereas the Galapagos dataset produces 31 of which 2 were acceptable, where the transformations used, the variables included and the points excluded can all differ. In contrast the Swiss dataset produces only one acceptable model, as the result of applying backward elimination to the initial model with no other RAP having any effect.

Notice how cycles may occur in the graphs, for example in the Chicago data, $10 \rightarrow 12 \rightarrow 21 \rightarrow 25 \rightarrow 10$, so that any search method that applies RAP's sequentially without memorizing past models would be vulnerable

to being caught in such a cycle. Different vertices in the graph may be reached by different paths indicating different orders in the analysis. It is simple to see the effect of eliminating an RAP from the analysis by removing the apropriate arcs and vertices from the graph. If information on the models represented by the vertices is stored, RAP's may be added to the graph, without redoing what has already been computed. It is also possible to identify crucial stages in the analysis that led to the choice of one model or another. The models and RAP's where the branch occurs could be more closely investigated by the statistician. The software will allow the tree to be displayed and any vertex may be selected to display that model and allow further analysis.

4 SELECTING A MODEL

Having generated a list of acceptable models, can we choose which one is best? Much has been written about model selection - it is likely that model selection techniques are included among the RAP's, but we have nothing new to say about these selection methods. We are providing a wider choice of plausible models to the analyst, models that might have been discovered by hand given limitless time and patience.

Expert knowledge of the particular area may allow one to choose one model with confidence or at least eliminate some of the competitors. Some model selection methods are criterion based, like the adjusted R^2 or the Akaike information criterion. Given that the response may be transformed in different ways in the competing models, the criterion may have to allow for this, so Mallow's C_p may be inappropriate. Lubinsky and Pregibon (1987) discuss some ways of choosing from models that differ in structure, allowing for the simplicity and accuracy of the model.

We could simply pick the model that maximises the chosen criterion, but this may be precipitous. Suppose, prediction is our goal then the predictions from the acceptable models may vary greatly even if the criterion does not. Given that the value of the criterion may be sensitive to small perturbations of the model, it would seem inadvisable to put too much weight on it. Also, one reason for constructing the list of acceptable models was the possibility that two capable analysts may differ in the ordering of their analysis and arrive at different final models. So it's also quite possible that using different criteria may result in different choices from the list.

One objective in regression analysis is to assess the dependence of the response on a particular predictor. This dependence is quantified by the appropriate regression parameter. If different transformations are used in the acceptable models, say a log transform on the response in one model and a square root in another, it will be difficult to directly compare the relevant parameter estimates. One possibility for a consistent method of comparison is to assess the change in the response as the relevant predictor is changed (both in the original scale) at a specific point in the range of X, X_0, that should be chosen with respect to the context of the data, i.e at points of particular interest in the predictor space. Since the effect may differ over this space, several X_0's might be considered. Another concern in interpreting regression coefficients is collinearity, which is not specifically addressed here.

Here is an example - suppose we are interested in the dependence of the first predictor, minority, on the response in the Chicago dataset. We perform the analysis as above but restrict the RAP's from eliminating (or adding polynomial terms to) the variable minority. Three acceptable models are found and are described in the following table (the response is square-rooted in all these models so there is no consistent interpretation problem). Model no. refers to Figure 1, the predictors are labeled, 1,2,3,4,5 in order as above and excluded refers to eliminated points.

Table 1 - Analysis of Chicago data

Model	Predictors	Excluded	Adj. R^2	$\hat{\beta}_1$	$\hat{se}(\hat{\beta}_1)$
17	1,3,4,5	7	85.1	-0.00805	0.00253
19	1,2,4,5	7,24	87.0	-0.00642	0.00240
27	1,2,4,$\sqrt{4}$	24	83.3	-0.01398	0.00201

There are reasons to pick any of these models, #17 because it is simplest, model #19 because it has the highest adjusted R^2 and model #27 because the $\hat{se}(\hat{\beta}_1)$ is smallest. Yet the models have quite different

forms and the $\hat{\beta}_1$ vary quite widely (and more than just a single $\hat{se}(\hat{\beta}_1)$ might indicate).

It might be better that the analyst be aware that there are several possible candidate models giving different interpretions and that to select one of them capriciously and discard the rest would be to ignore the real uncertainty in the estimate. It would be far better to report the full range of acceptable models and the predictions they make. If, however, there is only one acceptable model generated, then the analyst can be a lot more confident in the estimate. So we can see that it is advantageous to have the list of acceptable models available.

Another concern in simply selecting a "best" model is in the reliability of inference from that model. If one accepts that allowance for the data analysis that precedes model selection should be made in the inference that follows, then since the amount of data analysis here has been increased substantially over the usual amount, naive inference from a "best model" is likely to be even more optimistic with regard to estimated standard errors.

It should be emphasised here that it would imprudent to rely on the generated models alone. We advise that the statistician perform their usual analysis without using the RAP's and paying particular attention to graphical methods and physical context. A weakness of the RAP's is they lack the human perception of graphical displays and thus may miss important features. The generated models should be regarded as additional information not as a replacement for a standard analysis.

5 CONCLUSION

We have seen that the generation of all acceptable models is a useful tool in regression analysis. Practical difficulties include the complete specification of the methods of regression data analysis and the programming of RAP's in sufficient generality. Further development of these ideas requires a more comprehensive and versatile set of RAP's.

We recommend that the statistician do the analysis their usual manner and use the methods we have described to provide additional information. It would be unwise to rely solely on the models generated automatic procedures we have described because it is quite possible that important visible features will be missed by them and that physical context will be ignored.

We wish to emphasise that without the full incorporation of physical context into the RAP's, which is a quantum leap beyond what we have here, and without a much more comprehensive set of RAP's, the methods we have discussed here are only appropriate for careful use by statisticians and not for unguided application by the uninitiated.

REFERENCES

Adams J. (1990) *Proc. Stat. Comp. ASA*

Andrews D. & Herzberg A (1985) "Data : a collection of problems from many fields for the student and research worker" *New York, Springer-Verlag.*

Brownstone D. (1988) "Regression strategies" *Proceedings on the 20th Symposium on the Interface, Ed. Wegman E. et al.* 74-79

Daniel C. & Wood F. (1980) "Fitting Equations to Data, 2nd Ed." *New York, John Wiley.*

Gale W. (Editor) (1986) "Artificial intelligence and statistics" *Addison-Wesley, Reading Mass.*

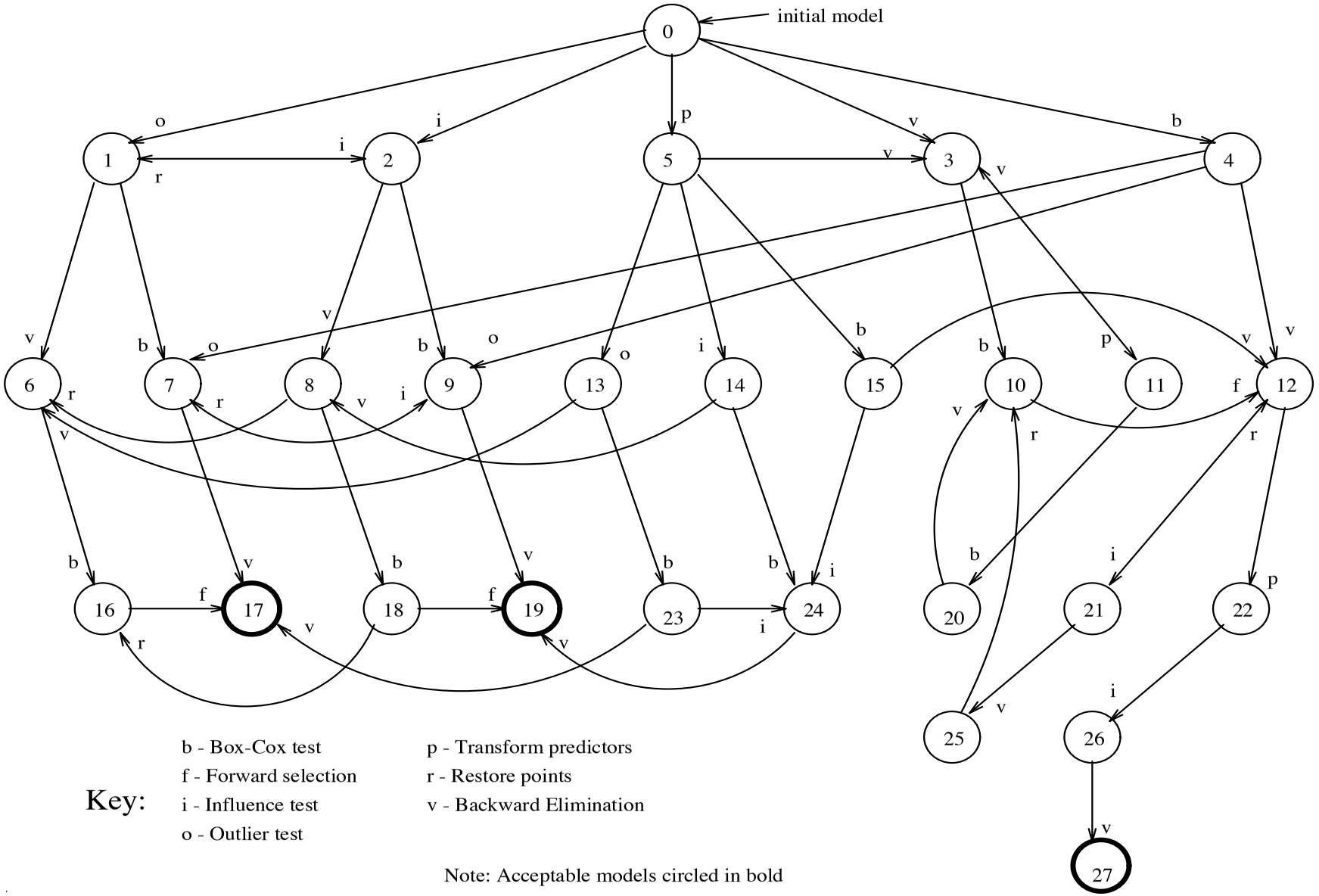
Lubinsky D. & Pregibon D. (1987) "Data Analysis as Search" in "Interactions in Artificial Intelligence and Statistical Methods" edited by Phelps B. *Gower Technical Press, Aldershot, Hants*

Mosteller F. & Tukey J. (1977) "Data Analysis and Regression" *Addison-Wesley, Reading Mass.*

Phelps B. (Editor) (1987) "Interactions in Artificial Intelligence and Statistical Methods" *Gower Technical Press, Aldershot, Hants*

Tierney L. (1990) "Lisp-Stat: An object-oriented environment for statistical computing and dynamic graphics." *Wiley, New York*

FIGURE 1 - Chicago Data



Key:

b - Box-Cox test
f - Forward selection
i - Influence test
o - Outlier test
p - Transform predictors
r - Restore points
v - Backward Elimination

Note: Acceptable models circled in bold