On the Cost of Data Analysis

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

*

Abstract

A regression analysis usually consists of several stages such as variable selection, transformation and residual diagnosis. Inference is often made from the selected model without regard to the model selection methods that preceeded it. This can result in overoptimistic and biased inferences. We first characterize data analytic actions as functions acting on regression models. We investigate the extent of the problem and test bootstrap, jackknife and sample splitting methods for ameliorating it. We also demonstrate an interactive LISP-STAT system for assessing the cost of the data analysis while it is taking place.

*Keywords:** Regression analyis, bootstrap, jackknife, data splitting, model selection.

1 INTRODUCTION

A statistical model is usually defined by a systematic and a random component. For example, a linear regression model is specified by a particular linear relationship between the predictors and the expected response and by the distribution of the errors, often i.i.d normal. Inference might then proceed based on this knowledge. However, in practice, since the true model is rarely known, inference is usually preceded by several iterative stages of data analysis to search for a satisfactory model. At this point, inferences may be made on the basis of this model without conditioning on the data analytic methods that preceded it. These inferences are inaccurate, tending to err on the side of overstating the significance of predictors and making predictions with over-optimistic confidence. There may also be a problem with bias. Of course, this problem is well known to many statisticians and is mentioned in most good textbooks on regression; for example: "When the model is chosen through data analysis, the (usual) formula for the standard error of a prediction is likely to underestimate the prediction errors" in Weisberg(1985, p229). It is a failing that is easy to note but difficult to rectify. The difficulty applies to any problem which is not trivial and where the model assumptions are checked in any way. Since few would make inferences from data, even when the true model is supposedly known, without some kind of model verification, then the problem is virtually omnipresent. This paper presents general methods for adjusting regression inference for prior data analysis.

Many previous investigations of this problem have tackled the linear regression model. Bickel & Doksum(1981) consider the effect of the Box-Cox transformation applied to the response. They conclude that the cost (in terms of inflating the variance of the slope parameter) of estimating the index of the transformation can be substantial. There is some controversy regarding the interpretation of the parameters of the model in this situation which was brought up by Box & Cox(1982) and Hinkley & Runger(1984) with ensuing discussion. See also Doksum & Wong(1983) and Taylor(1989). Carroll & Ruppert(1984) look at predictions from non-linear regression models. Here there is no parameter interpretation problem and the conclusion was that there is little cost in estimating the correct transformation.

Other papers have focused on the effect of variable selection on inference. Miller (1984) lists some possible approaches to this problem. The simplest method is to split the data. One part may be used for finding the regression model and the other part for inference thus avoiding the conditionality problem. There are a number of concerns with this approach. For example, how should the data be split and what is the loss in efficiency? Picard & Cook(1984) discuss this. Another approach is based on resampling methods. Freedman et al.(1988) describe a bootstrap and a jackknife method for adjusting for the effects of variable selection alone. They find that the bootstrap works reasonably well for problems where the ratio of predictor variables to cases is small but begins to break down where this ratio is large. Hurvich & Tsai(1990) discuss some simulations which indicate the extent of the problem and prescribe sample splitting. Brownstone (1988) uses a bootstrap-based approach and Kipnis (1991) describes a pseudosample approach to the problem. A Monte-carlo approach may also be feasible, although this is very much related to the bootstrap approach. Miller also mentions a maximum-likelihood and a shrunken estimator method which only apply to the variable selection problem and do not address the general case.

Carroll, Ruppert & Wu (1988) discuss the effect of estimating the weights when using weighted least squares. They describe a bootstrap method for adjusting for this effect. Carroll & Rup-

pert(1991) expand on this to include allowance for transformation. Freedman & Peters(1984) give some empirical results in a generalized least squares situation.

Gong (1986) and Taylor(1988) looked at the effect of variable selection and estimating additional shape parameters respectively in a logistic regression setting. Dabrowska & Doksum (1988) look at similar problems in survival analysis. Hill(1985-86) gives a Bayesian viewpoint.

The aforementioned papers have a common characteristic in that they consider only one or two data analytic procedures at a time and assume that otherwise the model is correct. However, there are many stages in a regression analysis. Regression diagnostics may suggest various transformations or the exclusion or down-weighting of outlying or influential points. Variable selection may be used to produce a more parsimonious model. Many regression analyses consist of several passes so that some procedures may be repeated many times in any one analysis. In short, the whole analysis is rather complex and may take several paths, some of which may depend on the subjective assessment of the statistician. This makes any formal (in a mathematical sense) assessment and adjustment for the effect of the model selection virtually impossible.

In section 2, we describe a tool for the investigation of these problems. In section 3, we investigate possible solutions to these difficulties - the bootstrap, jackknife and sample splitting. In section 4, we use simulations to demonstrate the extent of the problem and the efficacy of the solutions proposed. In section 5, we demonstrate the use of the method on some real data and also propose a method for continuously evaluating the effects of the data analysis in progress. Section 6 is the conclusion.

2 REGRESSION ANALYSIS TOOL

Regression analysis is an inexact procedure: competing methods are available for the same task, some methods depend on the visual perception of the analyst and the choice of methods may depend on the physical nature of the problem. However, any empirical investigation to compare and assess the proposed methods for adjusting the inference for the data analysis will require a device for repetitive analysis of regression data. Given the large number of analyses necessary, particularly when resampling methods are used, and the inconsistency of human data analysts, a computer program to aid regression data analysis has been developed.

A regression analysis might be viewed as the application of a succession of procedures which suggest if or how the current "best" model may be changed. We will characterize regression analytic procedures as functions acting on regression models and returning regression models. For some procedures such as variable selection and transformation, this is relatively straightforward, because these methods are usually completely specified and need not require human judgment. However, for other diagnostic based methods, it is not so easy. We may have some method for detecting outliers but we also must specify what we will do with these outliers - for example, exclude or down-weight them. For the detection of heteroscedascity, the analyst may plot the residuals against the fitted values and then take action based on the visual assessment of this plot. We, however, require an exact specification and so our methods cannot be graphical and hence are less flexible. We must test for some *specific* form of deviation from non-constant variance and then say what we will do (transform or reweight) in response to this. So we may miss some gross deviation visible in a plot, although graphical methods are not perfect in that we are forced

to assess the significance of a graphical feature which may be problematic. Also the physical context of the data can be included to some extent by restricting the procedures 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 proposed data analytic functions can only approximate the behavior of a human data analyst.

Although this program could be one part of an expert system to do regression analysis, it lacks an interface to translate the real-world problem to and from the appropriate format. Furthermore, the analyst still needs to choose which methods are appropriate, what order they should be performed in, and when to stop the analysis. In short, the statistician should perform the analysis in the usual way except that the procedures are restricted to those that can be completely and exactly specified. To assist in this, a prototype regression analysis tool(RAT) has been developed using the LISP-STAT language - a statistical computing environment based on the Lisp language. (Tierney (1990)). Source code is available from Statlib.

Suppose we have data (X_{ij}, Y_i) with $i = 1, ..., n$ and $j = 1, ..., p$. We aim to find a model of the form

$$g(Y_i) = \beta_0 + \sum_{j=1}^{p'} \beta_j f_j(X'_{ij}) + w_i \varepsilon_i$$

where the primes indicate that some initial predictors may be excluded or additional predictors like squared or interaction terms may be introduced. So p' may be greater than, less than, or equal to p. The w_i are weights, which could be set to zero if we wish to exclude a point from the model. The g and f_j are transformations. Thus a regression model, in our sense, is specified by the original data, the transformations, and the weights.

The following data analytic functions have been built into RAT:

• Check for skewness of the variables and transform if necessary. The rule is if X is strictly positive and $\max(x)/\min(x) > c$ (where c is some specified critical value, defaulting to 100) then $X \mapsto \log(X)$.

• Check and remove outliers. The rule is to compute the jackknife residuals and check if any residuals exceed a critical value computed using the Bonferroni inequality. Points so identified have weights set to zero.

• Check and remove influential points. The rule is to compute the Cook statistics and check for those exceeding 1. Points so identified have weights set to zero.

• Check for non-constant variance and reweight if necessary. The squared residuals are regressed on the fitted values and the fitted values squared. If the regression is significant, then weights are computed by iteration using a method explained in Davidian & Carroll (1987).

• Check for a Box-Cox transform on the response. The profile log-likelihood is maximized and the index rounded to the nearest half between -2 and 2.

• Check for transformations of the predictors by testing the significance of adding X_i^{α} to the model, using the method described in Weisberg p153, add new predictor if appropriate.

• Perform variable selection using the backward elimination method.

• Restore previously excluded points: Check for outliers using the method described above and reinclude all points that are not outliers but were previously excluded.

A P-value of 0.05 is used in all tests but this may be changed by the user. This list is obviously not exhaustive but is representative of the sort of data-analytic actions that may occur in practice.

Furthermore, I do not wish to imply that these are the best methods to use all the time. Note that I have used least-squares estimation but robust estimates could also be used although some methods like backward elimination would have to be modified appropriately.

These functions have been programmed to take any regression model as input and output a (possibly changed) regression model. The flexibility of 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 regression model. Additional data-analytic functions may be added easily without disturbing the operation of RAT.

If we write the regression model, as specified by the data, weights and transformations, as m, and the data analytic functions as $a_1, a_2, ...$, then a typical data analysis for sequence of data analytic actions $a_i, a_j, ..., a_k$ (where a_k is done first) in our sense, could be written as

$$m_{final} = a_i(a_j(....(a_k(m_{initial})))) = (a_i \circ a_j \circ ... \circ a_k)(m_{initial})$$

We consider two particular problems in adjusting the inference for the effect of data analysis. The first is prediction and the second is assessing the dependence of the response on a predictor.

3 METHOD

3.1 Bootstrap

To estimate the distribution of complex estimates, the bootstrap is immediately appealing, but the choice of resampling algorithm is problematic. For regression data there are two methods - resample the residuals, which is conditional on the model or resample from the rows of the data, which is independent of the model. If the residual method is chosen, a model must be specified. We could perform the full data analysis to come up with a model and resample the residuals conditionally from this model to generate the bootstrap samples. However, the final model is almost certain to be closely fit to the data and so the bootstrap samples will not capture the full variation of the data. Freedman et al(1988) in a variable selection setting, and our simulations, reveal that this method seriously underestimates the variance of the quantities of interest often performing little better than the naive estimates. Freedman et al(1988) chose to resample residuals from the full (all predictors, all untransformed and no weights) model prior to variable selection which implicitly assumes that the full model is structurally correct. We are not willing to assume the correctness of any initially proposed model since any conditional resampling scheme based on that model would be suspect. Therefore, we use the unconditional resampling scheme, resampling from the rows (X_i, Y_i). This has the advantage of simplicity in that no model need be specified prior to the analysis and intermediate estimates of the distribution of the estimates of interest may be obtained, but the disadvantage that the bootstrap estimates of variance are biased, although it is asymptotically equivalent to the conditional method, see Freedman (1981).

Of course, this method will estimate the unconditional variance of the estimates - not the variance conditional on X. However, except for small or very skew samples these differ hardly at all and because our analysis is likely to remove leveraged outliers, the difference would reduced. The unconditional resampling would have to be modified in situations where X truly is fixed or is a biased sample from the population.

3.2 Prediction

Suppose we want a distribution for the mean response of Y given some X_0. For model m, write this prediction as $\hat{Y}(m_{X_0})$. Note that this is computed in the original scale of the response. The proposed method is

a) Perform the data analysis using only completely specified data analytic functions like those described in Section 2. The initial model, order and number of times each function is used will be at the discretion of the data analyst. Record the order in which these functions were performed. Thus, for sequence of actions $a_i, a_j, ..., a_k$, the prediction is

$$\hat{Y} \equiv \hat{Y}((a_i \circ a_j \circ ... \circ a_k(m_{initial}))_{X_0})$$

b) Generate bootstrap samples (X_i^*, Y_i^*) by sampling unconditionally from the original data and form $m_{initial}^*$. Perform the data analysis in the same order as for the original data. It is quite possible that the bootstrapped final models may have different weights, transformations and predictors from the model for the original data, nevertheless a prediction at the point X_0 may still be computed:

$$\hat{Y}^* \equiv \hat{Y}((a_i \circ a_j \circ ... \circ a_k(m_{initial}^*))_{X_0})$$

c) Form the predictive distribution as the empirical distribution of the bootstrap predictions, $\hat{Y}^*$.

A predictive distribution for a new observation at X_0 could be obtained by adding resampled residuals to the $\hat{Y}^*$'s assuming no heteroscedascity was detected, otherwise more complex modifications would be necessary. From now on we consider only predictive distributions for the mean response.

Note that the prediction is a function of the sequence of actions as well as X_0, so that a valid bootstrap predictive distribution may only be obtained by applying the same sequence of actions to the resampled datasets. Hence, at any given stage of the analysis, we cannot just start resampling from the current data to construct a valid predictive distribution as this will not reflect the effect of the prior data analysis. Therefore, it is advisable that the analyst resolve upon our procedure at the outset. In this way, a current estimate of the predictive distribution may be obtained that takes account of the whole analysis up to that point.

Since analytic adjustments are available for some actions, you might wonder if it is really necessary to apply such an action to all the resampled datasets as some computational expense might be saved by using the analytic adjustment. Unfortunately, this is not possible because the available analytic adjustments can only supply the marginal variance inflation of a particular action. Thus there is no way to coherently combine such information with that obtained using our resampling method. To clarify this point, consider a regression analysis consisting of just two actions for which analytic adjustments exist. Marginal variance inflation could be computed for each action but there is no obvious way to combine the two to assess the variance inflation due to the combination of both actions. Furthermore, bias and other distributional quantities may also be of interest.

3.3 Parameter Estimates

Estimating the distribution of a parameter estimate is more difficult than prediction. One major problem is that in some bootstrap samples, variables may be transformed in different ways making meaningful comparison of estimates from the different samples difficult. It is overly restrictive

to use only data-analytic functions that avoid transforming the relevant variables. One approach would be to determine some model-free measure of the physical meaning of the relevant parameter so that meaningful comparisons would be possible. There are a number of ways of doing this. See, for example, Hinkley & Runger(1984). Another possibility 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 is

$$\text{Effect}(\beta_j) = \frac{d}{dx_j} g^{-1}[\beta_0 + \sum_{i=1}^{p} \beta_i f_i(x_i)] \mid_{X=X_0}$$

If g, f_j are identity functions then the Effect(β_j) is just the usual β_j. Effect(β_j) can be estimated by using the sample $\hat{\beta}$'s.

It might be argued that this turns parameter interpretation into something like prediction. However, if the parameters have some intrinsic meaning, it is likely that the "true" model is known and so little if any data analysis will be needed and the usual estimates of error will apply. Certainly transformation will not be considered without abandoning the model that gave the parameters meaning in the first place. Otherwise, when the parameters have no intrinsic meaning, it is no loss to translate them into effects, as described above, which do have a definite physical interpretation. Note that when a particular parameter effect is of particular interest, the analyst needs to restrict the variable selection methods from eliminating the relevant predictor. RAT can do this.

Collinearity is a problem in parameter interpretation and the methods we propose do not avoid it. Nevertheless, the bootstrap distribution will tend to be wider to reflect the additional uncertainty. Of course, it is advantageous to try to remove the collinearity using the usual methods, provided they are completely specified.

Of course, this method has the disadvantage of dependency on the choice of X_0. However, in practice, we might be interested in the effect of a predictor on the response at a number of points in the predictor space and we would not be surprised to find that the effect differed from point to point, so a universal interpretation of an effect would be unwise in any case. Therefore, the analyst should choose a point or points at which the effects can be calculated according to the physical context of the problem.

If the qualitative effect of a predictor on the response is the sole subject of interest, then since the sign of β_i has the same interpretation no matter what the scale, no special adjustments are necessary.

3.4 Jackknife

The jackknife is a general purpose method that has been used to estimate the bias and variance of estimators and can be applied here in the same way that the bootstrap estimates are calculated. One disadvantage that the usual Jackknife method has relative to the bootstrap is that it cannot estimate the distribution of the quantity of interest. However, there are some variations on the usual leave-out-one method discussed in Wu(1986). Wu's leave out many method allows for the estimation of the distribution. For the leave-out-one method the bias and variance may be calculated in the usual way. We also try the leave-out-(n+p-1)/2 as described by Wu. Since it is infeasible to consider all possible sub-samples of size (n+p-1)/2, we merely take a random sample of these. These sub-samples are then analyzed similarly to the bootstrap resample.

3.5 Sample Splitting

Another possible approach to the conditionality problem is to split the sample into two (not necessarily equal) parts. One part is used to find a model employing whatever data analytic procedures are appropriate and the other part to make inferences from the chosen model. Miller (1984) and Hurvich C. & Tsai C-L. (1990) recommend this approach. The main drawback is that because only part of the data is used for model selection, the choice of final model is not likely to be as good as if all the data were used. Furthermore, only part of the data is used for estimation introducing further inaccuracy. There is also the technical problem of exactly how the data should be split. Snee(1977) and Picard & Cook(1984) describe ways in which the data may be split into two suitable parts. The relative size of the two parts depends on how accurate we want the estimate to be and how well we wish to assess the error in that estimate, but it is difficult to quantify these competing concerns. We have simply chosen to split the data randomly into two equal parts in our simulations.

A technical difficulty may arise using the sample splitting approach if some of the data is negative but the part chosen for the data analysis is all positive. It is possible that a model may be chosen for this half data that cannot be used for the other half. For example, a log transform might be chosen for the response but this could not be applied to the second half of the data. In situations such as these we shall avoid the use of inappropriate transformations.

4 SIMULATION

First we shall do some simulations to demonstrate the effectiveness of this method. We generate data from the model

$$g(Y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j f_j(X_{ij}) + w_i \varepsilon_i$$

where X_{ij} is distributed F_x, and the ε_i , F_ε with $\beta_0 = 0$ and $\beta_i = 1,\ i \neq 0$.

We consider the prediction of Y at a point $(0.2, 0.2, ..., 0.2)$ and the estimation of the effect of β_1. We fix the variable selection method so that X_1 cannot be removed from the model. Parameters were interpreted at the point of the medians (X_0).

The models considered are given in the table below. The default values are n=50, p=5, g & f_j are identity functions, F_x & F_ε are standard normal and $w_i = 1$.

Model Label	Description
Vanilla	Default values
Outlier	$F_\varepsilon \sim \frac{2}{3}N(0,1) + \frac{1}{3}N(0,3^2)$
Nonlinear	$F_x \sim U(0,0.2)$, $g^{-1}(x) = e^{x/5}$
Hetero	$F_x \sim U(0,0.2)$, $w_i = \sum_{j=1}^{p} \beta_j f_j(X_{ij})$
Collinear	$X_1 \sim$ standard normal, X_2 and $X_3 \sim N(0,.01) - X_1$
Saturation	n=25, p=15

Table 1: Models used in the simulation

400 replications were made for each model with 100 resamples being used for both the bootstrap and the jackknife. We used the the leave-out-(n+p-1)/2 method for the jackknife except for the "saturation" model where the usual leave-out-one method was employed. Starting with an initial model of all variables included, no transformations and unit weights, data-analytic actions were performed in the order indicated in Section 2, once only, to produce a final model. For each replication, the predicted value or estimated parameter effect using regular least squares methods $\hat{\theta}$ and resampled values $\theta_1^*, \theta_2^*, ..., \theta_{100}^*$ for both bootstrap and jackknife and the split sample estimate $\tilde{\theta}$ were calculated. We now focus on how well the R.M.S.E of an estimator may be estimated. The R.M.S.E of the regular estimator $rmse(\hat{\theta})$ is estimated from the 400 replications as

$$\widehat{rmse}(\hat{\theta}) = \sqrt{\frac{1}{400}\sum_{i=1}^{400}(\theta_i - \theta)^2} \quad (\dagger)$$

and similarly for the split sample estimator. $rmse(\hat{\theta})$ may be estimated from a given replication in 3 ways: The bootstrap and jackknife estimates of the rmse are given by

$$\widehat{rmse}^*(\hat{\theta}) = \sqrt{\frac{1}{100}\sum_{i=1}^{100}(\theta_i^* - \hat{\theta})^2}$$

The naive method which takes no account of the prior data analysis assumes that the final model is correct and thus that the bias is zero, so the rmse is estimated using the usual least-squares estimate of the standard error. The rmse of the split sample estimator is also estimated using the least-squares estimate of the standard error.

Note that when the model was transformed, the naive estimates of error were also appropriately transformed, that is if $\hat{Y}_T$ is the prediction and $\hat{se}(\hat{Y}_T)$ is the estimated standard error in the transformed scale, then the estimated standard error in the original scale is computed as

$$\lim_{\delta \to 0} \frac{g^{-1}(\hat{Y}_T + \delta \hat{se}(\hat{Y}_T)) - g^{-1}(\hat{Y}_T)}{\delta} = (g^{-1})'(\hat{Y}_T)\hat{se}(\hat{Y}_T).$$

A similar calculation may be made for the parameter effect using a numerical approximation for the derivative (let δ be small).

The estimated densities of estimated rmse's are shown in Figure 1, using a kernel density estimator with a manually chosen bandwidth. The naive estimate of the rmse of the regular estimator is marked "naive" and the split sample estimate of the rmse of the split sample estimator is marked "split". Each has been normalized by subtracting and then dividing the estimate by the estimated true rmse (from $(\dagger)$). We should note that the estimated densities are only as good as might be expected from a sample size of 400, but the estimate of the true rmse used to normalize the scale is sufficiently good so as not to be noticeable within the resolution of the plot. Four outlying points were not plotted to avoid compacting the range of interest. Thus a good estimate of the rmse should have a density tightly centered around zero. Note that the naive estimates of the rmse of the regular estimate tend to be too small, sometimes substantially so. The bootstrap and jackknife estimates are superior to the naive estimate, although these also tend to be on the low side. The estimate of rmse of the split sample estimator falls short in several cases. This is disappointing since half

the sample has been sacrificed, thereby increasing the true rmse, which is then not even estimated adequately.

Now some specific comments on the models: In the "Vanilla" model , most of the data-analytic actions do not change the original model and so the naive estimates work reasonably well. In this example, where we happen to have the correct model from the outset, splitting the sample is quite a loss. In the "Outlier" model, the estimators are now non-normal (and in subsequent models) and the bias corrections of the bootstrap and jackknife are noticeable. The naive estimate of rmse is clearly too low whereas the resampling methods are better centered even if the variation is quite high. The split sample method has an estimator with higher variation although this variation is quite successfully estimated. The results for the "Nonlinear" model and the "Hetero" model are quite similar - the bias corrections and error estimation of the resampling methods are quite successful. In the "Collinear" model, the bimodality is due to variable selection and the resampling methods again perform well. In the "Saturated" model, there is a high ratio of variables to cases. The split sample method is not applicable here since if the sample is halved there would be more variables than cases. We must use the leave-out-one jackknife estimate here so the results may be displayed more succinctly in a table:

	Estimator		RMSE Estimation			
	Bias	SD	True	Naive	Boot	Jack
Effect	0.019	0.582	0.582	0.288	13.5	1.18
Prediction	-0.309	0.822	0.877	0.338	15.9	1.50

Table 2: Results for the saturation model

We give the sample bias and SD of the regular estimator, it's sample RMSE and the average naive, bootstrap and jackknife estimates of that RMSE. The bootstrap fails, the naive estimate is too small by about a factor of 2 and the jackknife too large by about a factor of 2.

To summarize, these simulations show that the naive estimates of error can seriously underestimate the RMSE of the quantities of interest. The jackknife or bootstrap can provide a more realistic estimate of the error. These estimates are not perfect but are certainly superior to the naive ones. Splitting the sample introduces substantially more variation into the estimates without the certain reward of eliminating the bias and of being able to estimate the variation successfully. The regular estimates are often biased but the resampling methods can be used to correct this somewhat and the split sample estimator also has less bias but at the expense of additional variance. The jackknife method is cheaper computationally and possibly more robust than the bootstrap although the delete-many jackknife method we used in the first five models would not apply when the number of predictors became relatively large.

Of course, it would be rash to base these recommendations on 6 examples. I have carried out simulations under different conditions and achieved qualitatively similar results. The source code for the simulations is provided and the user need only write the function that generates the data (examples provided) to study his or her own models of interest. These simulations were quite time-consuming (about 2 days on a DECstation 5000/200 each) even though each complete regression

analysis takes about 2 seconds.

5 EXAMPLE

During a data analysis it would be helpful to have a concurrent assessment of the distribution of each quantity of interest, be it a parameter effect or a prediction. This would be of assistance in assessing the effect of an action and, if we have a maximum acceptable amount of variation in an estimate, in knowing when to stop. Using the bootstrap method (or the jackknife) we can do this. We take B resamples unconditionally from the original data and from these construct B regression models. We then analyze the original data as we would normally, except that any action we take on that original data, even if it results in no change in the current model, is carried out on all the resampled regression models. So, for example, although the outlier test may indicate that point #7 be excluded from the model for the original data, different points, if any at all, might be excluded from the resampled models. At each stage in the analysis we can access the quantities of interest, be it a parameter effect estimate or prediction (or both) in the each of the resampled regression models and use these to assess the progress of the analysis.

A tool for carrying out a data analysis in this manner has been developed and provided for the user. The program prompts the user for the choice of resampling method, the number of resamples, etc. After each data analytic function has been executed the user may view density estimates of the resampled quantities of interest and other numerical summaries. The state of the current regression model may be studied to determine the next step. A data analysis in progress is shown in figure 2. Other features include the ability to view the history of the data analysis, in terms of a description of the effect of each action and as a change in the estimated density displayed as a succession of boxplots or as a "density slice", a 3-D surface plot showing the density changing with the actions. Although the user is initially prompted for a point of interpretation (X_0) and/or points at which predictions will be made, it is possible to see how the estimated densities of these quantities change over a range of newly chosen values using boxplots or the density slice.

We will demonstrate this idea on the Chicago Insurance dataset given in Andrews & Herzberg(1985). We will take volact($\approx$ insurance policies issued) as the response and percentage minority composition, fire rate, theft rate, age of housing and median family income as the predictors. A particular concern with this data is to detect the practice of "redlining", that is denial of insurance on the basis of race, so we will assess the dependence of volact on minority composition taking X_0 as the point 10,6.2,29,60.4,11744 (where the order of the predictors is as above) and the prediction of the mean response at that same point. I do not wish to imply that what follows is a complete or appropriate analysis of this data, since clearly other considerations regarding the context and method of analysis apply. I merely want to show how taking account of the effect of the data analysis changes the conclusions that otherwise might be made.

We use the bootstrap method with 400 resamples, although we could just as well have used the jackknife method. The progress of the analysis may be seen in figure 3 and 4 where the percentiles of the resampled estimates for the estimated parameter effect and the same for the prediction, respectively, may be seen to change as we progress through the data analytic actions. The estimates from the original data are marked with a solid line. No initial skewness was found in the original data and hence none in any of the resamples. The Box-Cox test indicated a square root transform

on the response and at this point the parameter estimation problem enters. The resamples gave different transformations (7% log, 89% square root and 4% identity) which shows up as increased variation for the parameter effect. Attempting to transform the predictors did not alter the model for the original data and changed the distribution only slightly and variable selection eliminated income causing some increase again in the variation. No outliers were found but point 24 was influential and was thus excluded. No heteroscedascity was found in the original data but it must have been detected in some of the resamples causing a change in the estimated densities. Further variable selection eliminated theft and point restoration reincluded point 24.

The data analysis was stopped at that point. As for the effect of minority composition on the response, the RAT analysis indicates that 3.75% of the resampled estimated parameter effects are greater than zero, whereas the t-statistic from the final model for the test of $\beta_1 = 0$ is -6.71 with 43 degrees of freedom, giving a miniscule P-value. Thus, the standard analysis indicates a very significant predictor whereas the RAT analysis points leaves the issue in doubt depending on one's opinions about P-values. Lest it be thought that this is purely due to an ill-fit model, diagnostic checks indicated no outstanding problem and bootstrapping the residuals from the final model gave an estimated se for $\hat{\beta}_1$ of 0.0023, very similar to the naive estimate of 0.0022. There might be some concern about our choice of X_0. In figure 5, we show the percentiles of the resampled estimated effects with the solid line again marking the estimated effect from the original data at the end of the analysis where we vary the percentage minority composition from 10 up to 100 with values of the other predictors held constant as given previously. Notice how the size of the effect decreases as does the variation as the minority composition increases. Turning to the prediction problem, the RAT estimate of the rmse for prediction is 0.532 larger than the naive estimate of 0.441, with little bias indicated by the final plot.

Parameter estimation is much more sensitive than prediction to the choice of data analysis. Changing the order of the data analysis gave different models and hence t-statistics for β_1. However, the bootstrap analyses were much more consistent giving P-values for the test of $\beta_1 = 0$ not dissimilar to the one above. Thus, there is a suggestion that using this method might result in inference less dependent on the order of the data analysis.

In conclusion, we can see that making inference from the final model taking the data analysis into account may result in quite different qualitative and quantitative conclusions.

6 DISCUSSION

One major obstacle to widespread use of this procedure is the necessity of characterizing data-analytic actions as functions where the response to every observed condition must be completely specified. Realistic incorporation of flexible graphically based procedures within this framework is challenging to say the least. Nevertheless, given that over-optimistic inferences are frequently made from regression data, RAT points a way towards at least partial solution to the problem.

There remain a number of other outstanding questions. For instance, it might be argued that, during the course of the data analysis, we might examine a plot and, consciously or otherwise, decide on some action or lack of action. This informal activity would not be taken into account by the program and thus not take account of the variation introduced by this activity. So we might argue that the analysis should be conducted blind or in a completely automated manner to avoid

this problem but this would seem to impose a crushing inflexibility. The opportunity for abuse also exists in that the analyst, not obtaining a desirable result in his first analysis, might restart RAT and change the order of the actions, thus invalidating the whole procedure. The parameter interpretation problem is not completely resolved and other schemes may be appropriate.

ACKNOWLEDGEMENTS

The author thanks the referees for suggesting substantial improvements to this paper.

REFERENCES

Andrews D. & Herzberg A (1985) "Data : a collection of problems from many fields for the student and research worker" *New York, Springer-Verlag.*
Bickel P. & Doksum K. (1981) "An analysis of transformations revisited" *JASA* **76** 296-311
Box G. & Cox D. (1982) "An analysis of transformations revisited, rebutted" *JASA* **77** 209-210
Brownstone D. (1988) "Regression strategies" *Proceedings on the 20th Symposium on the Interface, Ed. Wegman E. et al.* 74-79
Carroll R. & Ruppert D. (1984) "Power transformations when fitting theoretical models to data" *JASA* **79** 321-329
Carroll R. & Ruppert D. (1991) "Prediction and Tolerance Intervals with Transformation and/or Weighting" *Technometrics* **33** 197-210
Carroll R., Wu J. & Ruppert D. (1988) "The effect of estimating weights in weighted least squares" *JASA* **83** 1045-1054
Dabrowska D. & Doksum K. (1988) "Partial Likelihood in Transformation models with censored data" *Scan J. Statist.* **15** 1-23
Davidian M. & Carroll R. (1987) "Variance function estimation" *JASA* **82** 1079-1091
Doksum K. & Wong C. (1983) "Statistical tests based on transformed data" *JASA* **78** 411-417
Freedman D. (1981) "Bootstrapping regression models" *Annals of Statistics* **9** 1218-1228
Freedman D. & Peters S. (1984) "Bootstrapping a regression equation: Some empirical results" *JASA* **79** 97-106
Freedman, D., Navidi W., and Peters, S. (1988). "On the Impact of Variable Selection in Fitting Regression Equations." *Lecture Notes in Economics and Mathematical Systems, Theo K. Dijkstra (ed.) Springer-Verlag* 1-16.
Gong G, (1986) "Cross-validation, the jackknife and the bootstrap: excess error estimation in forward logistic regression" *JASA* **81** 108-113
Hill B. (1985-86) "Some subjective Bayesian considerations in the selection of models (with discussion)" *Econometric Reviews* **4** 191-288
Hinkley D. & Runger G. (1984) "The analysis of transformed data (with discussion)" *JASA* **79** 302-319
Hurvich C. & Tsai C-L. (1990) "The impact of Model Selection on Inference in Linear Regression" *American Statistician* **44** 214-217
Kipnis V. (1991) "Evaluating the impact of exploratory procedures in regression prediction" *Comp. Stat. Data. Anal.* **12** 39-55

Miller J. (1984) “Selection of Subsets of Regression Variables (with discussion)” *JRSS A* **147** 389-425
Picard R. & Cook R. (1984) “Cross-Validation of Regression Models” *JASA* **79** 575-583
Snee R. (1977) “Validation of regression models. Methods and examples” *Technometrics* **19** 415-428
Taylor J. (1988) “The cost of generalizing logistic regression” *JASA* **83** 1078-1083
Taylor J. (1989) “A note on the cost of estimating the ratio of regression parameters after fitting a power transformation” *J.Stat.Plan.Inf* **21** 223-230
Tierney L. (1990) ”Lisp-Stat: An object-oriented environment for statistical computing and dynamic graphics.” *Wiley, New York*
Weisberg S. (1985) “Applied Linear Regression (2nd Ed.)” *Wiley, New York*
Wu J. (1986) “Jackknife, bootstrap and other resampling methods in regression analysis (with discussion)” *Annals of Statistics* **14** 1261-1350

FIGURES

Figures 1 and Figures 3,4 and 5 appear on the next two pages. Figure 2 is large screen shot of the software and is not included here.

Figure 1

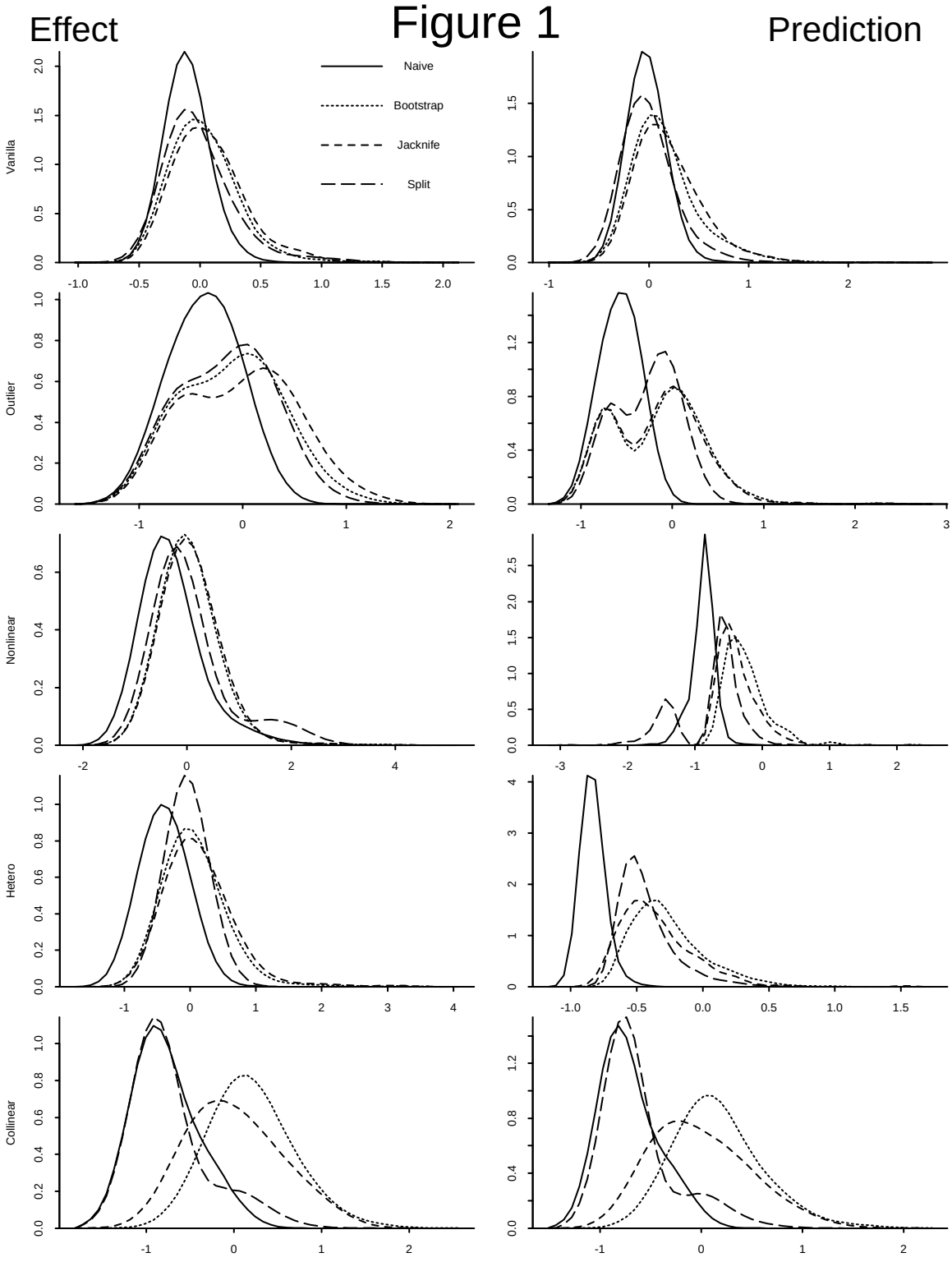
Effect
Prediction
Naive
Bootstrap
Jacknife
Split
Vanilla
Outlier
Nonlinear
Hetero
Collinear

Figure 3 - Analysis history for Minority effect

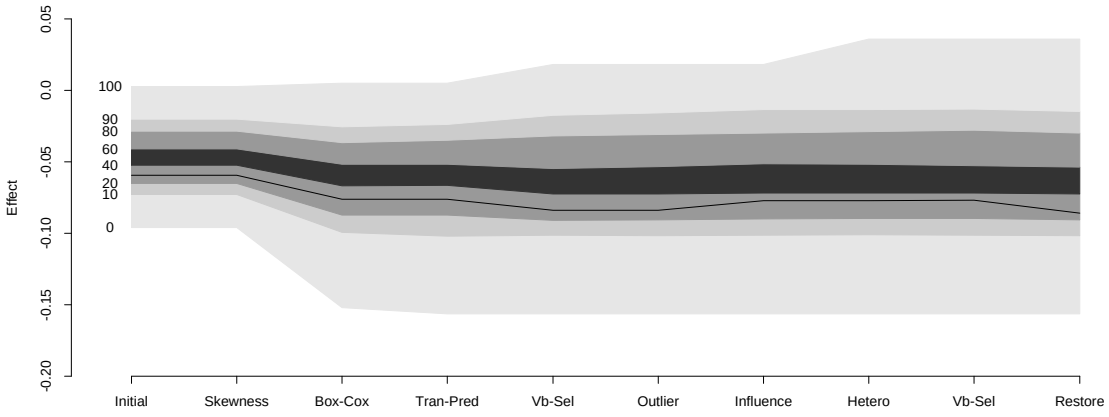

Figure 4 - Analysis history for the prediction

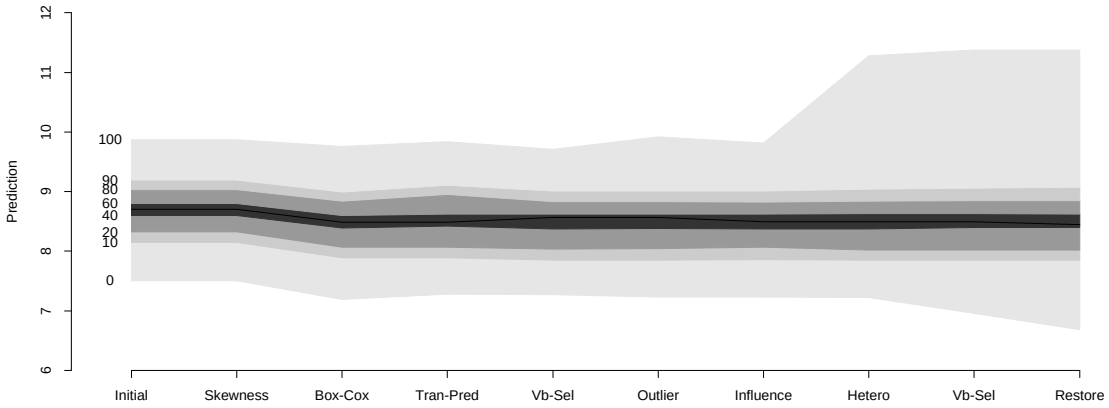

Figure 5 - Effect distribution as point of interpretation varies