Bayesian Robust Variable and Transformation Selection: A Unified Approach

Raphael Gottardo[†] and Adrian E. Raftery[‡]
Department of Statistics
[†]University of British Columbia and [‡]University of Washington

Technical Report no. 508
Department of Statistics
University of Washington

January 5, 2007

Abstract

We consider the problem of simultaneous variable and transformation selection for linear regression. We propose a fully Bayesian solution to the problem, which allows us to average over all possible models including transformations of the response and predictors. We use the Box-Cox family of transformations to transform the response and each predictor. To deal with the change of scale induced by the transformations, we propose to focus on new quantities rather than the estimated regression coefficients. These quantities, that we call generalized regression coefficients, have a similar interpretation to the usual regression coefficients on the original scale of the data, but do not depend on the transformations. This allows us to make probabilistic statements about the size of the effect associated with each variable, on the original scale of the data. Finally, in addition to variable and transformation selection, there is also uncertainty involved in the identification of outliers in regression. In this paper, we also propose a more robust model to account for such outliers based on a t-distribution with unknown degrees of freedom. Parameter estimation is carried out using an efficient Markov chain Monte Carlo algorithm, which allows us to move around the space of all possible models. Using three real data sets and a simulated one, we show that there is considerable uncertainty between model selection, transformation and outlier identification and that the three should be done simultaneously.

Contents

List of Tables

List of Figures

1 Introduction

Variable selection in linear regression is an important problem, whose purpose is to select a group of variables that best predict an outcome variable. Given a dependent variable $\boldsymbol{Y}$ and a set of potential independent variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, we wish to compare models of the form $\boldsymbol{Y} = \beta_0 + \boldsymbol{X}_{j_1}\beta_{j_1} + \cdots + \boldsymbol{X}_{j_q}\beta_{j_q}$, where $\boldsymbol{X}_{j_1}, \ldots, \boldsymbol{X}_{j_q}$ is a selected subset of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$. For this problem it is common to assume a standard linear model to describe the relationship between the response and independent variables, namely

$$y_i = \beta_0 + \sum_{k=1}^{q} X_{ij_k}\beta_{j_k} + \epsilon_i, \tag{1}$$

where the β_{j_q}'s are the unknown regression coefficients and ϵ_i follows a normal distribution with constant variance ψ^{-1}. In many cases, such assumptions (common variance, additive error structure, normal distribution) might be unrealistic and one solution is to look for transformations of the outcome variable and/or regressors so that (1) is appropriate after transformation.

Box and Cox (1964) discussed the power transformation family of models. In particular, they gave methods for estimating the parameters in the nonlinear model

$$h(\boldsymbol{Y}, \lambda) = \boldsymbol{A}\boldsymbol{\beta} + \psi^{-1}\boldsymbol{\epsilon}, \tag{2}$$

where ϵ follows a standard normal distribution, ψ is the precision, i.e. the reciprocal of the variance, $\boldsymbol{A}$ is a known design matrix, $\boldsymbol{\beta}$ is a vector of parameters, and

$$h(\boldsymbol{Y}, \lambda) = \begin{cases} (\boldsymbol{Y}^{\lambda} - 1)/\lambda, & \text{if } \lambda \neq 0 \\ \log(\boldsymbol{Y}) & \text{otherwise.} \end{cases}$$

Note that this transformation is valid only if $\boldsymbol{Y} > 0$. There exist several methods for estimating the unknown parameters such as maximum likelihood (Box and Cox 1964) and Bayesian approaches (Box and Cox 1964; Perrichi 1981; Sweeting 1984; Hinkley and Runger 1984).

There has been some discussion about the correct way to make inference about the regression parameter $\boldsymbol{\beta}$ when the transformation parameter λ is unknown (Bickel and Doksum 1981; Box and Cox 1982; Hinkley and Runger 1984). Bickel and Doksum (1981) showed that the variance of $\hat{\boldsymbol{\beta}}$ is greatly inflated when λ is estimated from the data compared to the case where it is known. Box and Cox (1982) and Hinkley and Runger (1984) argued that linear parameters have meanings only with reference to a particular scale and thus recommend taking a conditional approach. Chen and Lockhart (1997) obtained the Fisher information matrix and its inverse for all the unknown parameters, namely $\boldsymbol{\beta}$, ψ^{-1} and λ. They showed that the asymptotic distributions of $(\psi^{-1}, \boldsymbol{\beta})$ conditionally and unconditionally on λ were different and concluded that conditioning on λ was one step short of performing valid analyses. Carroll and Ruppert (1981) and Taylor (1986) proposed restricting attention to the predictive distribution of new observations, which can be defined independently of the scale. They studied the properties of the conditional median and conditional mean respectively. They showed that there is some cost in estimating λ but that the cost is not severe. However, their approach ignored the regression coefficients, which themselves are often of interest: they summarize the relationship between the dependent and independent variables.

To avoid scaling issues, Box and Cox (1964) and Smith and Kohn (1996) considered rescaling the transformation so that the scale of the transformed data is approximately the same for each λ, but Dagenais and Dufour (1994) argued against this as there is no clear interpretation of the parameters after such rescaling.

Variable and transformation selection are often done sequentially, and the resulting model depends on the order in which they are performed. It would seem more appropriate to do them simultaneously. Since both can be viewed as model selection procedures, we can unify them within a Bayesian framework, which would allow us to get more realistic measures of uncertainty by averaging over all possible models. Hoeting et al. (2001) proposed a simultaneous approach to variable selection and transformation based on a power transformation for the outcome variable and change-point transformations for the independent variable. The change-point transformation has the advantage of not inducing a change of scale. However, the power transformation does, and Hoeting et al. (2001) did not consider this scaling issue. Moreover, they used an approximate algorithm based on Bayes factors to estimate the cut-points of the change point transformation, and therefore did not average over all possible transformations. Finally, they used the Markov chain Monte Carlo model composition (MC^3) of Madigan and York (1995) to perform variable selection, which requires one to integrate out all model parameters. Geweke (1996) and George and McCulloch (1997) showed that such integration can be avoided.

Hoeting and Ibrahim (1998) proposed taking a predictive Bayesian viewpoint to perform variable and transformation selection. The authors used the Box-Cox family of transformations, but did not have to worry about scaling issues as they did not average over all possible transformation but instead selected the

best transformation. Finally, their variable selection approach requires one to compute all possible models and this is not practical even for moderate sized problems.

In this paper, we introduce a Bayesian model for variable and transformation selection. Our transformation selection approach is based on the Box-Cox family of transformations but could be generalized to other types of transformation. Parameter estimation is carried out using an efficient Markov chain Monte Carlo (MCMC) algorithm, which allows us to move around the space of all possible models (including transformations). To deal with the change of scale induced by the transformations, we focus on new quantities, which we call generalized regression coefficients. These have a similar interpretation to the usual regression coefficients on the original scale of the data, but do not depend on the transformations selected. Finally, in addition to variable and transformation selection, there is also uncertainty associated with the identification of outliers in regression. We also include the identification of outliers in our methodology; our approach is based on t-distributions with unknown degrees of freedom.

Section 2 starts by describing our basic model and prior specification for variable and transformation selection, and then extends it to deal with outliers. In Section 3, we introduce the MCMC algorithm used for parameter estimation and discuss our solution to the scaling problem. In Section 4, we illustrate our methodology on three real data sets and a simulated one. Finally, in Section 5 we discuss our findings and possible extensions.

2 A model for variable and transformation selection

Given a dependent variable $\boldsymbol{Y}$ and a set of potential regressors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, we wish to compare models of the form $g(\boldsymbol{Y}) = \beta_0 + g_{j_1}(\boldsymbol{X}_{j_1})\beta_{j_1} + \cdots + g_{j_q}(\boldsymbol{X}_{j_q})\beta_{j_q}$, where $\boldsymbol{X}_{j_1}, \ldots, \boldsymbol{X}_{j_q}$ is a selected subset of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, and g and the g_{j_q}'s are transformations from a predefined set of real functions, $\mathcal{T}$. In general, the set of possible transformation $\mathcal{T}$ would be a parametric family, $\mathcal{T} = \{g_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d\}$. When the response is transformed, a Jacobian term enters into the likelihood for the untransformed response, and so we require $g_{\boldsymbol{\theta}}$ to be a diffeomorphism for each $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$. Note that this assumption is not required for the independent variables; see for example Hoeting et al. (2001) where the authors used a change-point transformation for the independent variables. Here, for simplicity, we assume that the set of possible transformations is the same for the outcome variable and the independent variables.

We use the Box-Cox family of transformations, which we define as

$$\mathcal{T} = \{g_\lambda(x) \equiv (x^\lambda - 1)/\lambda : \lambda \in \Lambda \subset \mathbb{R}\},$$

where Λ is a given subset of $\mathbb{R}$. Note that the methodology presented here could easily be extended to other parametric transformations. In particular, if the data are not positive, then one could use the shifted power transformation of Box and

Cox (1964) or the extended power transformation family of Bickel and Doksum (1981).

2.1 A basic model

We now assume that a standard linear model can be used to describe the relationship between the transformed response and independent variables, namely

$$\begin{aligned} g_\lambda(y_i) &= \beta_0 + \sum_{j=1}^{p} g_{\lambda_j}(X_{ij})\beta_j + \epsilon_i, \\ (\epsilon_i|\psi) &\sim \mathrm{N}(0, \psi^{-1}), \end{aligned} \tag{3}$$

where the $\boldsymbol{\beta}_i$'s are the unknown regression coefficients, and λ and λ_j's are the transformation parameters.

In order to allow each variable to be either in or out of the model, we model each regression coefficient as a mixture of a Normal distribution and a point mass at zero, as follows:

$$(\beta_j|\lambda, \lambda_j, \sigma_\beta) \sim (1-w)\delta_0 + w\mathrm{N}\left(0, \frac{S^2_{g_\lambda(\boldsymbol{Y})}}{S^2_{g_{\boldsymbol{\lambda}_j}(\boldsymbol{X}_j)}}\sigma^2_\beta\right), \tag{4}$$

where w is the prior probability of being in the model for each variable, $S^2_{\boldsymbol{z}}$ denotes the empirical variance of $\boldsymbol{z}$, and σ^2_β is a common variance parameter. Note that the prior distribution for the regression coefficients is allowed to depend on the scales of the variables; this is to account for the change of scale induced by the transformations. We do not view this as being in contradiction with the Bayesian paradigm, but rather as an approximation to the prior information of an investigator who knows something, but not much, about the problem at hand. The prior distributions of the regression coefficients are as spread out as they can reasonably be given the marginal standard deviations of the variables. Note that data dependent priors have been used by other researchers in this context (Box and Cox 1964; Sweeting 1984).

The parameter w is assumed to have a Beta distribution, Beta $(\hat{w}\kappa, (1-\hat{w})\kappa)$, where $\hat{w}$ and κ are fixed hyperparameters. The mean of this distribution is $\hat{w}$ and the variance is $\hat{w}(1-\hat{w})/(1+\kappa)$ so that $\hat{w}$ can be seen as a "best guess" for w, whereas κ controls the spread around this prior guess. Putting a prior on w allows us to estimate the proportion of variables, which can be seen as a Bayesian solution to the multiple testing problem arising when the number of variables to be included is large (Scott and Berger 2006). Throughout this paper we use $\hat{w} = 0.5$ and $\kappa = 4$, i.e. a fairly noninformative Beta (2,2) prior. Note that this prior makes every model (marginally) equally likely *a priori*.

The prior for the variance parameter σ^2_β is taken to be uniform on the interval $[0, 1]$. The rationale for this is as follows. If all the variables are standardized by dividing by their standard deviation, then (4) implies that the prior variance of β_j is σ^2_β. On this scale, β_j rarely exceeds 1 in absolute value. It never does so when

$p = 1$ by the Cauchy-Schwarz inequality, and empirical evidence that it rarely does so when $p > 1$ was given by Raftery, Madigan, and Hoeting (1997). Thus $\sigma_\beta^2 = 1$ will almost always be more than large enough to cover the range of values of β_j, and much smaller values can easily be appropriate. Thus a $U[0, 1]$ prior for $\sigma_\beta^2 = 1$ covers the range of possibilities fairly well. The results are typically insensitive to reasonable changes in this prior.

The prior for the scaling parameter ψ is taken to be improper, $\pi(\psi) \propto \psi^{-1}$. The prior for the intercept β_0 is $(\beta_0|\lambda) \propto 1/S_{g_\lambda(\boldsymbol{Y})}$, to account for the change of scale induced by λ. Our rational for this prior follows from an idea used in Box and Cox (1964). Suppose that for a fixed value λ_1, the transformation over the range of observation is nearly linear, $g_\lambda(y) \approx \text{ const } + l_\lambda g_{\lambda_1}(y)$, where l_λ is a rescaling constant (depending on λ). Suppose furthermore that for λ_1, the regressors have little effect on the response, that is $E[\boldsymbol{g}_{\lambda_1}(Y)|\boldsymbol{X}] \approx \beta_0$, and using the approximate linear relationship $E[\boldsymbol{g}_\lambda(Y)|\boldsymbol{X}] \approx l_\lambda \beta_0$. Choosing $\lambda_1 = 1$, a simple estimate for l_λ is $S_{g_\lambda(\boldsymbol{Y})}$. The prior for λ only reflects the Jacobian term coming from the linear transformation. Again, this prior is data-dependent and is in the spirit of the one used in Box and Cox (1964). While simple, this prior accounts for the change of scale induced by the transformation of the response variable. We have found this prior to give reasonable results in practice. Finally we used a uniform prior on $[-1, 1]$ for the transformation parameters λ and λ_j's.

2.2 A robustified model

It has been shown that transformation selection can be heavily influenced by the presence of a few outliers (Carroll and Ruppert 1982; Cook and Wang 1983; Carroll and Ruppert 1985; Atkinson 1988; Hinkley and Wang 1988; Cheng 2005). As with transformation and variable selection, the order in which outlier detection and transformation is done usually leads to different answers. Thus, once again, we wish to do everything simultaneously, and our approach is based on t distributions with unknown degrees of freedom. We introduce a more robust version of (3), as follows:

$$\begin{aligned} g_\lambda(y_i) &= \beta_0 + \sum_{j=1}^{p} g_{\lambda_j}(X_{ij})\beta_j + \frac{\epsilon_i}{\sqrt{\varpi}_i}, \qquad (5) \\ (\epsilon_i|\psi) &\sim \mathrm{N}(0, \psi^{-1}), \\ (\varpi_i|\nu) &\sim \text{ Gamma } (\nu/2, \nu/2), \end{aligned}$$

where the β_i's are the unknown regression coefficients and the ϖ_i's are independent of the ϵ_i's. Since the ϖ's are independent of the ϵ's, we have $\epsilon_i/\sqrt{\varpi_i} \sim t_{(\nu,0,\psi^{-1})}$, i.e. the errors have a t distribution with ν degrees of freedom and scale parameter ψ^{-1}. The advantage of writing the model this way is that, conditionally on the ϖ_i, the sampling errors are again normal, but with different precisions.

The prior for the degrees of freedom ν is taken to be uniform on the set $\{1, 2, \ldots, 10, 20, \ldots, 100\}$. All other priors remain the same. As we will see in

the results section, the accommodation of outliers can have a substantial influence on both variable and transformation selection.

3 Parameter estimation and MCMC computation

3.1 MCMC sampler

It can be difficult to devise good MCMC algorithms in the context of transformation and variable selection due to the change of scale induced by the transformation (Smith and Kohn 1996). Here we introduce an efficient algorithm that accommodates such changes of scale. Our MCMC algorithm cycles through the following steps:

1. Block update $(\lambda, \beta_0, \boldsymbol{\beta}, \psi)$ by Metropolis-Hastings .
2. Update β_0 by Gibbs sampling.
3. For $j = 1$ to p
   (a) If $\beta_j > 0$ block update $(\lambda_j, \beta_0, \beta_j, \psi)$ by Metropolis-Hastings .
   (b) Update β_j by Gibbs sampling.
4. Update σ_β^2 by Metropolis-Hastings .
5. Update ψ by Gibbs sampling.
6. Block update $\boldsymbol{\varpi}$ and ν by Gibbs sampling (for the model with t distributed errors).

For move 1, we first select a candidate transformation λ^* using a symmetric proposal centered at the current value λ. We then compute candidate values $\beta_j^* = S_{g_{\lambda^*}(\boldsymbol{Y})}/S_{g_\lambda(\boldsymbol{Y})}\beta_j$; this basically rescales the regression coefficients. Then we find new candidate values β_0^* and ψ^* by maximizing the likelihood ratio over the set of possible transformations $\beta_0^* = \beta_0 + \delta$ and $\psi^* = \alpha\psi$ where δ and α are constants in $\mathbb{R}$ and $\mathbb{R}^+$ respectively. Straightforward calculations lead to the optimal values,

$$\delta = \frac{\sum_i \varpi_i \{g_{\lambda^*}(y_i) - g_\lambda(y_i) - \sum_j (\beta_j^* - \beta_j) g_{\lambda_j}(X_{ij})\}}{\sum_i \varpi_i}$$

and

$$\alpha = \frac{\sum_i \varpi_i \{g_\lambda(y_i) - \beta_0 - \sum_j \beta_j g_{\lambda_j}(X_{ij})\}^2}{\sum_i \varpi_i \{g_{\lambda^*}(y_i) - \beta_0^* - \sum_j \beta_j^* g_{\lambda_j}(X_{ij})\}^2}.$$

Note that these are given for the t-distributed model, and the corresponding estimates for the Gaussian model can be obtained by setting $\varpi_i \equiv 1$. We then use the Metropolis-Hastings algorithm to decide whether or not to accept these

new candidate values. This proposal induces a Jacobian term in the acceptance ratio equal to $\alpha \prod_{\{j:\beta_j \neq 0\}} [S_{g_{\lambda^*}(\boldsymbol{Y})}/S_{g_\lambda(\boldsymbol{Y})}]$, due to the change of scale in $\boldsymbol{\beta}$ and ψ.

Move 3a is similar to move 1. We first select a candidate transformation λ_j^* using a symmetric proposal centered at the current value λ_j and set $\beta_j^* = S_{g_j(\boldsymbol{X_j})}/S_{g_j^*(\boldsymbol{X_j})}\beta_j$. Then we find new candidate values β_0^* and ψ^* by maximizing the likelihood ratio over the set of possible transformations $\beta_0^* = \beta_0 + \delta$ and $\psi^* = \alpha\psi$ where δ and α are constants in $\mathbb{R}$ and $\mathbb{R}^+$ respectively. Straightforward calculations lead to the optimal values

$$\delta = \frac{\sum_i \varpi_i \{\beta_j g_{\lambda_j}(X_{ij}) - \beta_j^* g_{\lambda_j^*}(X_{ij})\}}{\sum_i \varpi_i}$$

and

$$\alpha = \frac{\sum_i \varpi_i \{g_\lambda(y_i) - \beta_0 - \sum_j \beta_j g_{\lambda_j}(X_{ij})\}^2}{\sum_i \varpi_i \{g_\lambda(y_i) - \beta_0^* - \sum_{k\neq j} \beta_k g_{\lambda_k}(X_{ik}) - \beta_j^* g_{\lambda_j^*}(X_{ij})\}^2}.$$

We then use the Metropolis-Hastings algorithm to decide whether or not to accept these new candidate values. Again this proposal induces a Jacobian term in the acceptance ratio equal to $\alpha S_{g_{\lambda_j}(\boldsymbol{X_j})}/S_{g_{\lambda_j^*}(\boldsymbol{X_j})}$ due to the change of variable.

The Gibbs sampler step 3b is performed using the full conditionals for the β_k's given by

$$(\beta_j|\dots) \sim (1-w_j^*)\delta_0 + w_j^* \mathrm{N}\left(\psi \sum_i r_{ij}/(\psi \sum_i \varpi_i g_{\lambda_j}(X_{ij})^2 + \psi_{\beta_j}), (\psi \sum_i \varpi_i g_{\lambda_j}(X_{ij})^2 + \psi_{\beta_j})^{-1}\right), \tag{6}$$

where

$$w_j^* = 1 - \frac{1-w}{1-w+w\sqrt{\psi_\beta/(\psi\sum_i \varpi_i g_{\lambda_j}(X_{ij})^2 + \psi_{\beta_j})}\exp(0.5(\psi\sum r_{ij})^2/(\psi\sum_i \varpi_i g_{\lambda_j}(X_{ij})^2 + \psi_{\beta_j}))},$$

the residual r_{ij} is defined by

$$r_{ij} = \varpi_i(g_\lambda(y_i) - \beta_0 - \sum_{k\neq j} \beta_k g_{\lambda_k}(X_{ik}))g_{\lambda_j}(X_{ij}),$$

and

$$\psi_{\beta_j} = \left(\frac{S^2_{g_\lambda(\boldsymbol{Y})}}{S^2_{g_{\lambda_j}(\boldsymbol{X}_j)}}\sigma_\beta^2\right)^{-1}.$$

All other updates are straightforward, involving Gibbs or random walk type proposals, and are not described here.

3.2 Generalized Regression Coefficients

In each case, the posterior model probability can be estimated from the MCMC output as the proportion of time spent in the corresponding model. The marginal posterior probability that the coefficient for each predictor does not equal zero,

namely $\Pr(\beta_j \neq 0|\boldsymbol{y})$, can be obtained by summing the posterior model probabilities across models for each predictor. However, note that in (6), w_j^* corresponds to the probability that $\beta_j \neq 0$ given everything else in the model. Thus, one can obtain more efficient estimates by Rao-Blackwellization when averaging the w_k^* values computed at each iteration at no extra cost.

In regression analysis, the β_j's themselves are also of interest as they summarize the relationship between the dependent variable and the independent variables. One difficulty with transformations is that the scale, and thus the interpretation, of each β_j depends on the transformations currently being applied. As a result, the usual MCMC posterior summaries for the β_j's such as means and standard deviations, are meaningless. As a solution we focus on a different quantity, which has a similar interpretation to the β_j on the original scale of the response and independent variables. On the original scale (i.e. when no transformation is applied), we have, for each observation (omitting the observation index i),

$$\text{med}(Y) = \beta_0 + \sum_j \beta_j X_j,$$

and as a result $d[\text{med}(Y)]/dX_j = \beta_j$. Similarly, after transformation of Y and X,

$$\text{med}((Y^\lambda - 1)/\lambda) = \beta_0 + \sum_j \beta_j (X_j^{\lambda_j} - 1)/\lambda_j.$$

Thus by invariance of the median we get

$$\text{med}(Y) = [1 + \lambda\{\beta_0 + \sum_j \beta_j (X_j^{\lambda_j} - 1)/\lambda_j\}]^{1/\lambda},$$

and so

$$d[\text{med}(Y)]/dX_j = \beta_j X_j^{\lambda_j - 1}[1 + \lambda\{\beta_0 + \sum_j \beta_j (X_j^{\lambda_j} - 1)/\lambda_j\}]^{1/\lambda - 1}.$$

The quantity $d[\text{med}(Y)]/dX_j$ does not depend on the transformations and has a similar interpretation to β_j on the original scale. Thus we use the sample average of values of this quantity, namely

$$\beta_j^G \equiv \frac{1}{n} \sum_i \beta_j X_{ij}^{\lambda_j - 1}[1 + \lambda\{\beta_0 + \sum_j \beta_j (X_{ij}^{\lambda_j} - 1)/\lambda_j\}]^{1/\lambda - 1}, \tag{7}$$

as a measure of the marginal change in Y when X_j is varied, all else equal. This has a similar interpretation across transformations, and it is equal to β_j on the original scale of the data. We call the quantity in (7) the generalized regression coefficient, as it generalizes the usual regression coefficient in the presence of transformations. The generalized regression coefficients can easily be estimated from the MCMC output.

Note that the quantity $1 + \lambda\{\beta_0 + \sum_j \beta_j (X_{ij}^{\lambda_j} - 1)/\lambda_j\}$ might be negative in which case (7) is not defined. However, this is unlikely to happen in practice

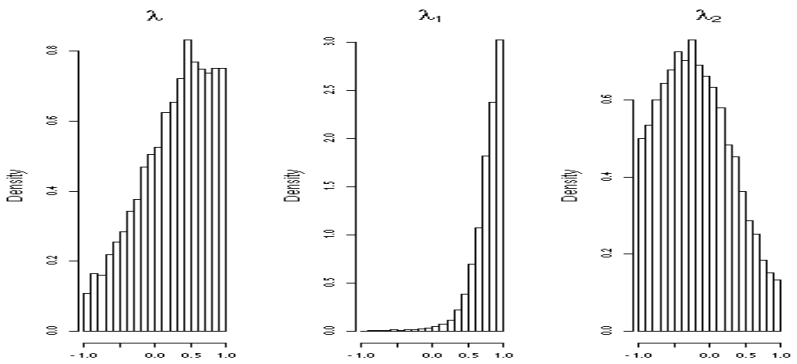

Figure 1: Histograms of MCMC samples from the posterior distributions of the transformation parameters, λ and λ_k's for the Hald data.

since the observations are required to be positive. In fact the probability of $1 + \lambda\{\beta_0 + \sum_j \beta_j (X_{ij}^{\lambda_j} - 1)/\lambda_j\}$ being negative will go to zero as the sample size increases. The generalized regression coefficients, given by (7), were all well defined for the examples used here. However, if β_j^G is not defined for a few possible values of the β_k parameters, one solution would be to find the posterior distribution of β_j^G conditional on its being well defined.

4 Examples

We now give results for three real data examples and a simulated example. All results from our method are obtained using an MCMC algorithm based on one million iterations thinned by 100 after a burn-in period of 1000 iterations.

4.1 The Hald data

The first data we used to illustrated our methodology is the Hald cement data. These data have been analyzed by many researchers; see for example Cook (1977), George and McCulloch (1993) and Hoeting and Ibrahim (1998). A full description of the data can be found in Draper and Smith (1981). There are four predictors, each one measuring the percentage composition of a particular ingredient in samples of cement concrete. The response is the heat evolved in calories per gram of cement.

Here, we use the model given by (3) as it has been noted that the Hald data were well behaved with no influential observations (Cook 1977). This was also confirmed by fitting the robust model as given by (5), and observing a large estimated number of degrees of freedom. Figure 1 shows histograms of the marginal posterior distribution of each transformation parameter, there is evidence that the second independent variable, X_2, needs to be transformed and some evidence that the outcome variable needs to be transformed as well. These results are in agreement with the estimated transformation parameters given in Hoeting and

Ibrahim (1998) even though these authors did not consider transformation of the outcome variable. However, there is substantial uncertainty about the transformations, and such uncertainty is naturally taken into account with our approach, which is not the case for the approach described in Hoeting and Ibrahim (1998).

Figure 2 shows that there are obvious changes of scale in the regression coefficients induced by the transformations. As a result, the usual ergodic averages from the MCMC output are meaningless, as explained in Section 3.2. The bottom graphs of Figure 2 show the trace plots of the generalized regression coefficients. The trace plot looks much better, with no obvious changes of scale, and the usual MCMC estimates can now be used.

The posterior model probabilities are given in Table 1 and are in broad agreement with previous results (George and McCulloch 1993; Hoeting and Ibrahim 1998). Note that the exact posterior model probabilities computed in George and McCulloch (1993) and Hoeting and Ibrahim (1998) differ from ours as the models used are slightly different and we consider transformations of both the outcome and the independent variables.

Table 1: Estimated posterior model probabilities for the Hald data. Only models with posterior probabilities greater than 0.01 are displayed. The marginal posterior probabilities of each variable being included in the model are shown in the last row.

X_1	X_2	X_3	X_4	Prob.
•	•			0.478
•	•		•	0.205
•	•	•		0.137
•	•	•	•	0.132
•			•	0.023
•		•	•	0.013
0.988	0.959	0.288	0.374	1

4.2 The US crime data

We now turn to the US crime data (Ehrlich 1973), a larger data set with 15 independent variables and so potentially $2^{15} = 32,768$ different models. Ehrlich's analysis concentrated on the relationship between crime rate and predictors 14 and 15 (probability of imprisonment and average time served in state prisons). In his original analysis, Ehrlich (1973) focused on two regression models, consisting of the predictors (9, 12, 13, 14, 15) and (1, 6, 9, 10, 12, 13, 14, 15), respectively, which were chosen in advance based on theoretical grounds.

After logarithmic transformations of all non-discrete variables, Raftery et al. (1997) stated that standard diagnostic checking (e.g. Draper and Smith (1981)) did not reveal any gross violations of the assumptions underlying normal linear

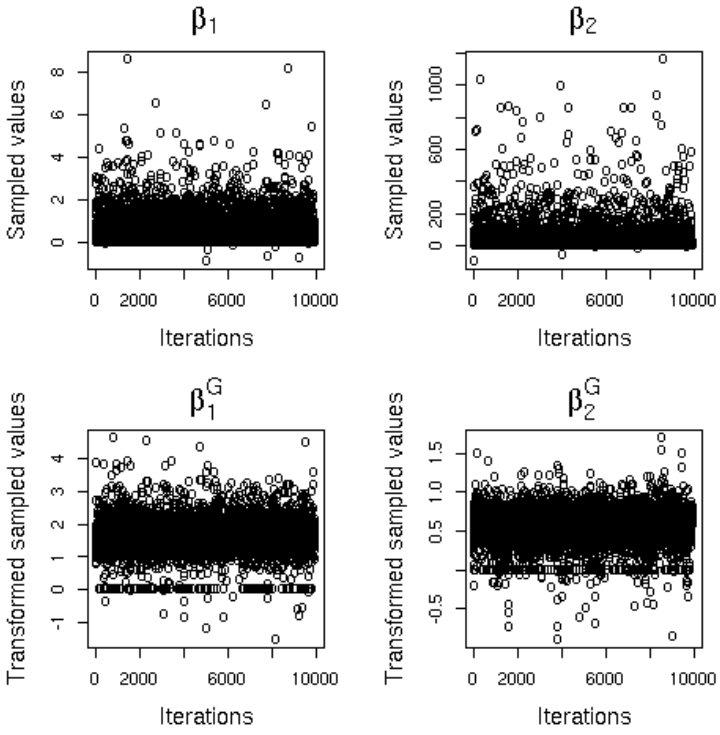

Figure 2: Trace plots of the regression coefficients, β_k's (top row), and the generalized regression coefficients (bottom row) for the Hald data. Note the y-axis scales: the regression coefficients show large changes of scale, while the generalized regression coefficients do not.

regression. Thus, we use the model with Gaussian errors given by (3). All variables were considered for transformation except $\boldsymbol{X}_2$, which is binary.

Table 2: Estimated posterior model probabilities for the crime data. Only models with posterior probabilities greater than 0.04 are displayed. The marginal posterior probabilities are given by the last row.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	Prob.
•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	0.0163
				•								•	•		0.0065
•	•	•	•	•		•	•	•	•	•	•	•	•		0.0052
•	•	•	•	•	•	•		•	•	•	•	•	•	•	0.0051
•	•	•	•	•	•	•	•	•		•	•	•	•	•	0.0049
•	•	•	•	•	•		•	•	•	•	•	•	•	•	0.0048
•	•	•	•	•	•	•	•	•	•	•	•	•	•		0.0038
•		•	•	•	•	•	•	•	•	•	•	•	•	•	0.0041
0.707	0.411	0.865	0.790	0.703	0.304	0.361	0.384	0.672	0.3122	0.513	0.514	0.984	0.828	0.428	1

Figure 3 shows histograms of the marginal posterior distribution of each transformation parameter. The histogram for the transformation parameter of the outcome variable, λ, supports the logarithmic transformation originally used by Ehrlich (1973). However, our analysis suggests that independent variables 13 and 14, $\boldsymbol{X}_{13}$ and $\boldsymbol{X}_{14}$, should be transformed. Again, there is a lot of uncertainty about the transformation; and such uncertainty is naturally taken into account. As with the Hald data, there are obvious changes of scale in the regression coefficients induced by the transformations; see the top graphs of Figure 4. The bottom graphs of Figure 4 show the trace plots of the generalized regression coefficients, which are much better with no obvious changes of scale, and the usual MCMC estimates can now be used.

The posterior model probabilities are given in Table 2. These are different from the ones given in Raftery et al. (1997) for two main reasons. Firstly, we consider variable transformation, which they did not, and more importantly we estimate the proportion of independent variables to be included in the model. Raftery et al. (1997) considered every model equally likely *a priori*, which implicitly fixes the proportion of independent variables to be included in the model at 0.5. In our case, the estimated posterior mean for the proportion parameter w is 0.57, which suggests that more variables should be included. Note however, that the marginal posterior probabilities are in broad agreement, with for example, the same three most significant variables, all with probability greater than 0.8 (shown in Figure 4).

4.3 The Scottish Hill Racing Data

Our third example involves data supplied by the Scottish Hill Runners Association. This example has been used by many researchers to illustrate the influence of outliers in linear regression (Atkinson 1986; Atkinson 1988; Hoeting et al. 1996). Here, we use it to illustrate the influence of outliers on both variable and transformation selection. The purpose of the study was to investigate the

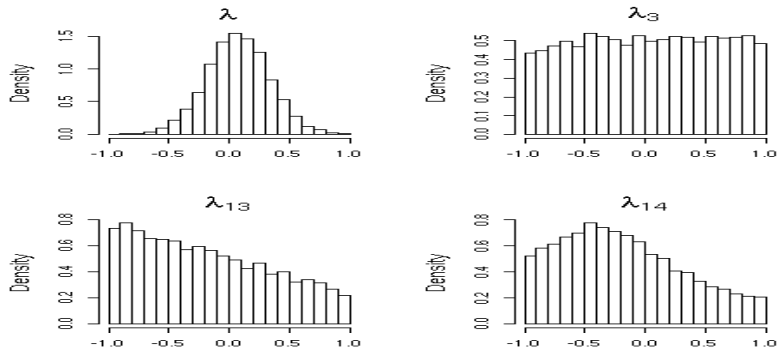

Figure 3: Histograms of the transformation parameters, λ and λ_k's, for the crime data. Only λ_k's for which the corresponding posterior probability for the regression coefficient of being non zero is greater than 0.8 are displayed.

relationship between record times of 35 hill races and two predictors: distance, defined as the total length of the race measured in miles, and climb, defined as the total elevation gained in the race, measured in feet; see Atkinson (1986) for further details.

In particular, Atkinson (1986) and Hadi (1990) concluded that races 7 and 18 are outliers. After they removed observations 7 and 18, their methods indicated that observation 33 is also an outlier. Thus, they conclude, observations 7 and 18 mask observation 33.

We start by fitting the Gaussian model (3). The top graphs of Figure 5 show histograms of the marginal posterior distribution of each transformation parameter for the model with Gaussian errors. All three histograms, except perhaps that of λ_1, clearly suggest transformation of the corresponding variables. The bottom graphs of Figure 5 show histograms of the marginal posterior distribution of each transformation parameter for the model with t errors. Now, there is not much evidence that the first independent variable $\boldsymbol{X}_1$ should be transformed and less evidence that the outcome variable should be transformed. From the Gaussian model, the posterior mean of λ is 0.12 and a 95% credible interval is $(-.34, .52)$, while from the t model the posterior mean of λ is 0.48 and 95% credible interval is $(.33, .64)$. This suggests that transformations for both $\boldsymbol{X}_1$ and $\boldsymbol{Y}$ are heavily influenced by the presence of outliers. Table 4 shows the estimated posterior means of the ϖ_i's, which can be interpreted as weights. Observation 18, 19 and 24 have small weights suggesting that they are outliers. On the other hand, observation 7 is slightly downweighted while observation 33 is not downweighted at all, suggesting that transformations of the response and independent variables are enough to accommodate such outliers. For comparison, we fitted the same model without transforming the independent variables, and observations 7, 18 and 33 had the three smallest weights, all smaller than .5 (results not shown). Note however that transformations of the independent variables were not considered by Atkinson (1986). Overall, this shows that transformation selection

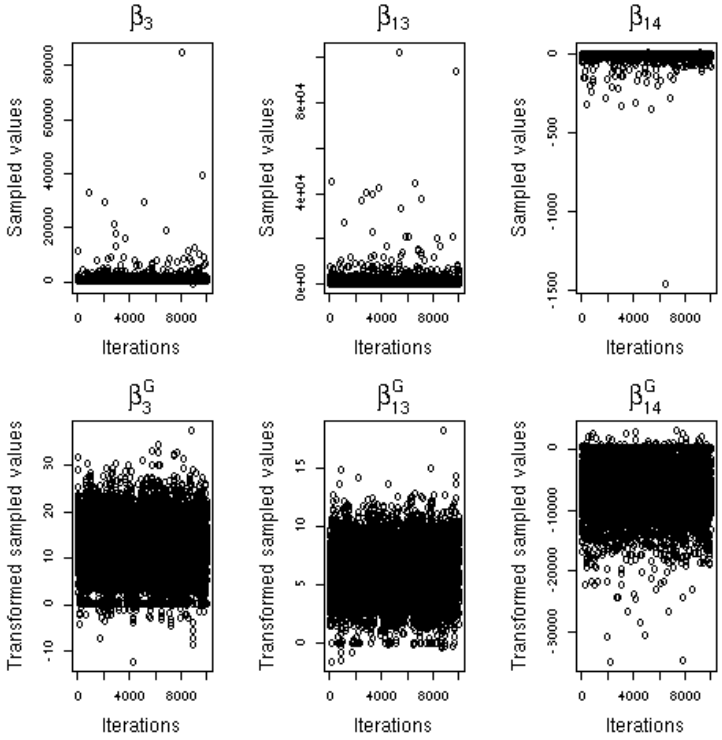

Figure 4: Trace plots of the regression coefficients, β_k's, and the generalized regression coefficients for the crime data. Only coefficients with posterior probability of being non zero greater than 0.8 are displayed.

and treatment of outliers affect one another, and that the two should be done simultaneously.

Finally, the posterior model probabilities are given in Table 3. There are substantial differences between the Gaussian and t models. After accounting for potential outliers the posterior probability of $\boldsymbol{X}_1$ being included in the model increased substantially.

Table 3: Estimated posterior model probabilities for the Scottish hill race Data. Only models with posterior probabilities greater than 0.01 are displayed. The marginal posterior probabilities are given by the last row.

Gaussian			t		
X_1	X_2	Prob.	X_1	X_2	Prob.
•	•	0.359	•	•	0.765
	•	0.641		•	0.235
0.359	1.000	1	0.767	1.000	1

Table 4: Posterior weights, i.e. posterior means of the ϖ's, associated with each observation of the Scottish hill race data. Observations with small weights are downweighted. Observations 18, 19 and 24 have small weights, suggesting that they might be outliers.

Obs.	1	2	3	4	5	6	7	8	9	10	11	12
Weight	0.97	1.17	1.07	1.27	1.29	0.90	0.72	1.33	1.34	1.15	0.94	0.80
Obs.	13	14	15	16	17	**18**	**19**	20	21	22	23	**24**
Weight	1.17	0.79	1.10	1.10	1.17	**0.06**	**0.42**	1.31	1.21	0.83	1.16	**0.42**
Obs.	25	26	27	28	29	30	31	32	33	34	35	
Weight	1.14	1.16	1.22	1.16	1.10	0.56	1.24	1.12	1.17	1.30	0.91	

4.4 Simulated Data

In the three previous examples considered, the true answer is unknown. We now turn to simulated data to evaluate the performance of our methodology when the truth is known. We follow the format of George and McCulloch (1997). We constructed n=100 observations on p=40 potential regressors as follows

$$y_i = \left[\lambda\{5 + \sum_k \beta_k (\lambda_k X_{ki} + 1)^{1/\lambda_k} + \epsilon_i\} + 1 \right]^{1/\lambda}$$

where ϵ_i follows a t-distribution with 2 degrees of freedom and the transformation parameters $\lambda, \lambda_1, \ldots, \lambda_p$ were randomly selected from the set $\{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\}$.

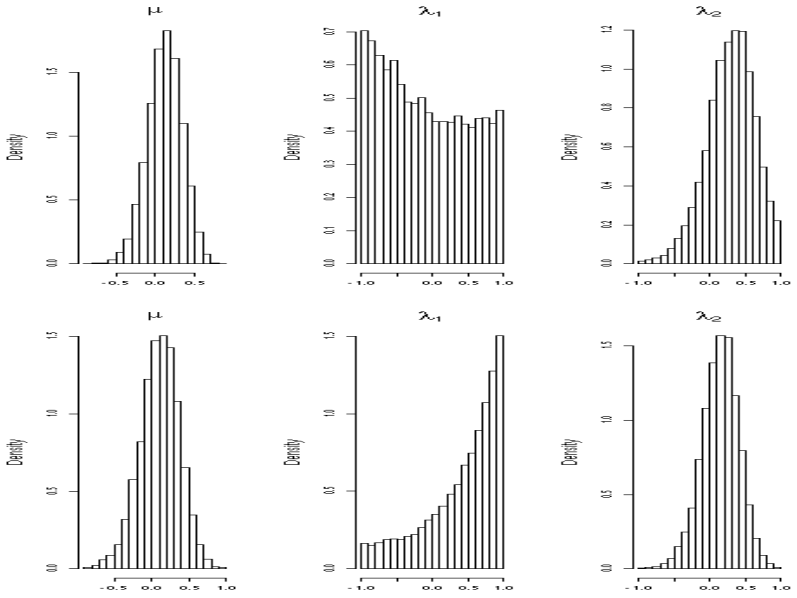

Figure 5: Histograms of the transformation parameters, λ and λ_k's for the Scottish hill race data for the Gaussian model (top) and t model (bottom).

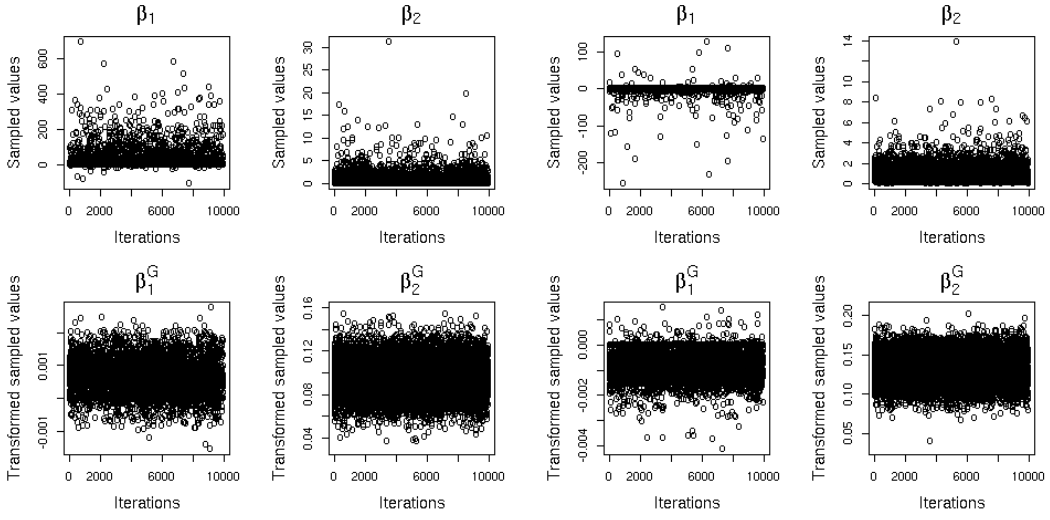

Figure 6: Trace plots of the regression coefficients, β_k's, and the generalized regression coefficients for the Scottish hill race data for the Gaussian model (left) and t model (right).

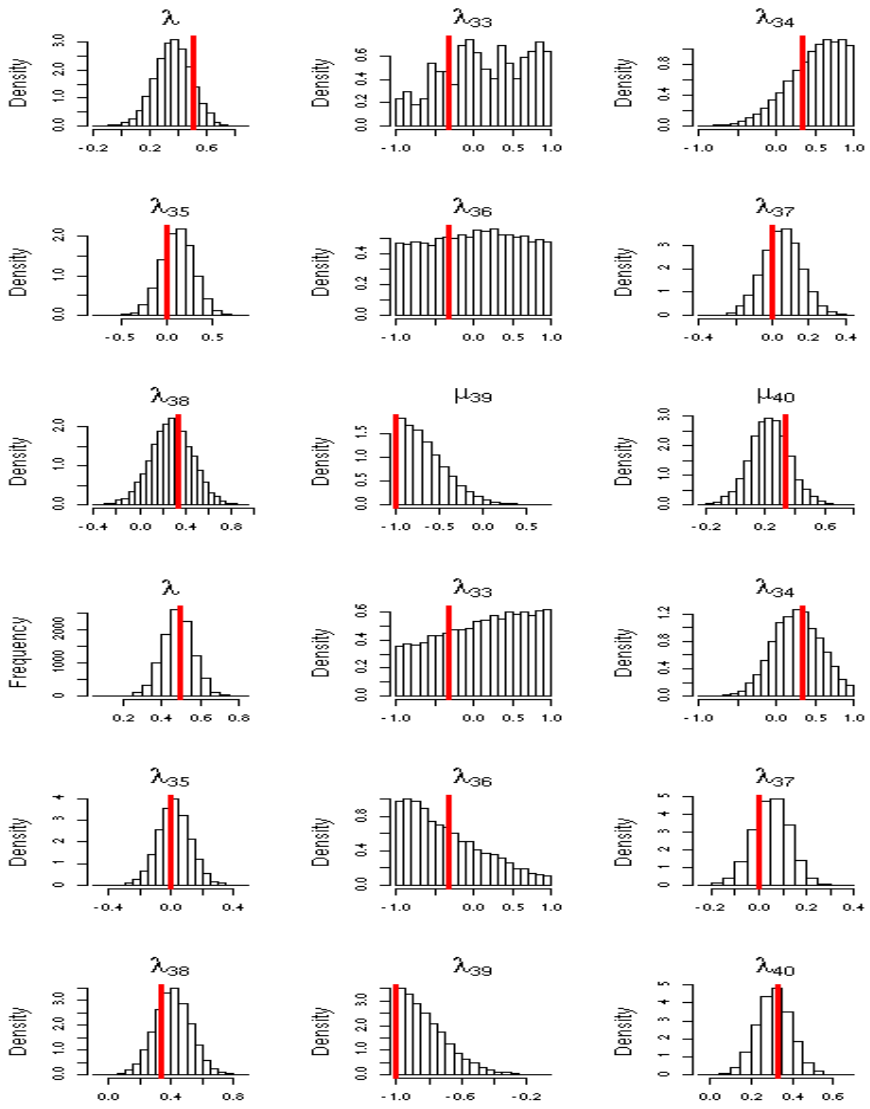

Figure 7: Histograms of the transformation parameters, λ and λ_k's for the simulated data for the Gaussian model (top) and t model (bottom). Only coefficients with posterior probability of being non zero greater than 0.1 are displayed.

The $\boldsymbol{X}_p$'s were obtained by generating independent observations from a $\mathrm{N}(2, 0.4)$ distribution if the corresponding transformation is greater than 0 and from a $\mathrm{N}(.5, 0.1)$ otherwise. This ensures that all regressors are positive. Only the last eight regressors were included in the model, i.e. $\beta_p = 0$ for $p = 1, .., 32$ and the remaining eight coefficients were set to 1, 1, 2, 2, 3, 3, 4 and 4, respectively.

The first three columns of Figure 7 show histograms of the marginal posterior distribution of the transformation parameter for the non zero regression coefficients estimated under the Gaussian model. Overall most histograms are well concentrated around the true values (red vertical bars), but again there is quite a bit of uncertainty. The last three columns of Figure 7 show histograms of the marginal posterior distribution of the transformation parameter for the non zero regression coefficients estimated under the t model. The transformations are better estimated than with the Gaussian model. This is particularly true of the outcome variable; the Gaussian model clearly underestimates the parameter λ to accommodate for the presence of outliers.

Finally, the posterior model probabilities are shown in Table 5 for both the Gaussian and t models. The t model gives better results than the Gaussian model. The posterior probability for the true model (including regressors 33 to 40) increases from 0.13 (Gaussian) to 0.70 (t).

Table 5: Estimated posterior model probabilities for the simulated data. Only models with posterior probabilities greater than 0.02 are displayed. The marginal posterior probabilities are given by the last row. Only variables included in one of the models with posterior probability greater than 0.02 are shown.

Gaussian										
X_{11}	X_{27}	X_{33}	X_{34}	X_{35}	X_{36}	X_{37}	X_{38}	X_{39}	X_{40}	Prob.
			•	•	•	•	•	•	•	0.456
		•	•	•	•	•	•	•	•	0.134
•		•	•	•	•	•	•	•	•	0.048
0.137	0.097	0.313	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t										
X_{11}	X_{27}	X_{33}	X_{34}	X_{35}	X_{36}	X_{37}	X_{38}	X_{39}	X_{40}	Prob.
		•	•	•	•	•	•	•	•	0.696
	•	•	•	•	•	•	•	•	•	0.052
0.021	0.073	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

5 Conclusion

We have introduced a unified approach to the problems of variable selection, transformation selection and outlier identification in regression. Using three real examples and a simulated one, we have shown that there can be considerable

uncertainty about each of these three modeling choices, and that all three should be done simultaneously. We have also shown how to deal with the change of scale induced by each transformation and make inference (including probabilistic statements) about the size of the effect associated with each predictor.

Similarly to Box and Cox (1964) and Hinkley and Runger (1984), our prior formulation for the regression coefficients depends on the transformation parameters to accommodate any change of scale. We have found a data dependent prior based on the scales of the variables to work well in practice, but other prior formulations have been proposed (Perrichi 1981; Sweeting 1984).

Here we chose a continuous prior for the Box-Cox transformation parameters. For scientific reasons, such as interpretation of the transformation parameters, an investigator might want to restrict each transformation to a finite number of values. Our model can easily be extended to the case where the parameters are restricted to a finite set. However, we have found our MCMC algorithm to mix poorly in the discrete case due to the possible large difference in likelihood between two very different transformation parameters. It would be possible to derive more efficient algorithms to overcome this problem, such as simulated tempering algorithms (Geyer 1991; Marinari and Parisi 1992). However, we have introduced new parameters that we call generalized regression coefficients, which have the same interpretation as the regression coefficients on the original scales of the data. Thus, using our approach one can make inference about the effect of each variable on the original scales, whose interpretation remains valid regardless of the transformation selected. Finally, as pointed out by Carroll (1982) in the context of maximum likelihood inference, restricted (discrete) and unrestricted inferences can lead to significantly different answers, and in such a case the unrestricted approach might be preferable. We used a univariate update Gibbs proposal for the regression coefficient, which we found to work well in the examples explored here. However, for very large problems with multicollinearity, an algorithm that performs block updates might be desirable. For example, this could be done by adding an extra step to our MCMC scheme such as the block update described in Nott and Green (2004).

References

Atkinson, A. C. (1986). Aspects of diagnostic regression analysis (discussion of "Influential observations, high leverage points and outliers in linear regression," by S. Chatterjee, A. S. Hadi). *Statistical Science 1*, 397–402.

Atkinson, A. C. (1988). Transformations unmasked. *Technometrics 30*, 311–318.

Bickel, P. J. and K. A. Doksum (1981). An analysis of transformations revisited. *Journal of the american statistical association 76*, 296–311.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B 26*, 211–252.

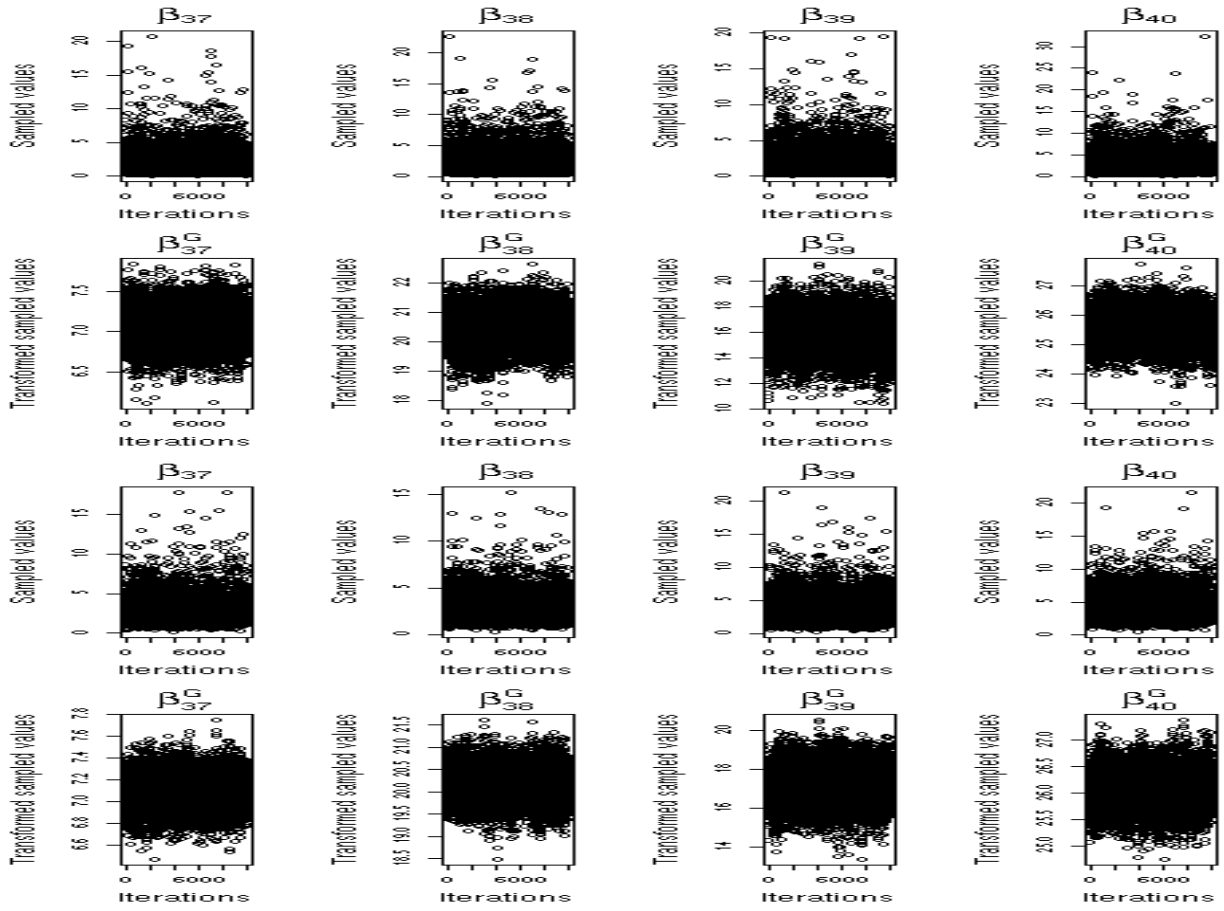

Figure 8: Trace plots of the regression coefficients, β_k's, and the generalized regression coefficients for the simulated data for the Gaussian model (top) and t model (bottom). Only the last four coefficients are displayed.

Box, G. E. P. and D. R. Cox (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association 77*, 209–210.

Carroll, R. J. (1982). Prediction and power transformation when the choice of power is restricted to a finite set. *Journal of the American Statistical Association 77*, 908–915.

Carroll, R. J. and D. Ruppert (1981). On prediction and the power transformation family. *Biometrika 68*, 609–615.

Carroll, R. J. and D. Ruppert (1982). Robust estimation in heteroscedastic linear models. *The annals of statistics 10*, 429–441.

Carroll, R. J. and D. Ruppert (1985). Transformations in regression: A robust analysis. *Technometrics 27*, 1–12.

Chen, G. and R. A. Lockhart (1997). Box-Cox transformed linear models: A parameter-based asymptotic approach. *Canadian Journal of Statistics 25*, 517–529.

Cheng, T. (2005). Robust regression diagnostics with data transformations. *Computational Statistics & Data Analysis 49*, 875–891.

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

Cook, R. D. and P. C. Wang (1983). Transformations and influencial cases in regression. *Technometrics 25*, 337–343.

Dagenais, M. G. and J. Dufour (1994). Pitfalls of rescaling regression models with box-cox transformations. *Review of Economics and Statistics 76*, 571–575.

Draper, N. R. and H. S. Smith (1981). *Applied regression analysis*. New York: Wiley.

Ehrlich, I. (1973). Participation in illegitimate activities: A theoritical and empirical investigation. *Journal of Political Economy 81*, 521–565.

George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*, 881–889.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica 7*, 339–374.

Geweke, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics 5 – Proceedings of the Fifth Valencia International Meeting*, pp. 609–620.

Geyer, C. (1991). Markov chain monte carlo maximum likelihood. In *Computing Science and Statistics: Proceeding of the 23rd Sumposium on the Interface*, pp. 156–163.

Hadi, A. S. (1990). A stepwise procedure for identifying multiple outliers in linear regression. *American Statistical Association Proceedings of the Statistical Computing Section*, 137–142.

Hinkley, D. V. and G. Runger (1984). The analysis of transformed data. *Journal of the American Statistical Association 79*, 302–309.

Hinkley, D. V. and S. Wang (1988). More about transformations and influencial cases in regression. *Technometrics 30*, 435–440.

Hoeting, J., A. E. Raftery, and D. Madigan (2001). Bayesian variable and transformation selection in linear regression. *Journal of Computational and Graphical Statistics 11*, 485–507.

Hoeting, J. A. and J. G. Ibrahim (1998). Bayesian predictive simultaneous variable and transformation selection in linear model. *Computational Statistics & Data Analysis 28*, 87–103.

Hoeting, J. A., A. E. Raftery, and D. Madigan (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics & Data Analysis 22*, 251–270.

Madigan, D. and J. York (1995). Bayesian graphical models for discrete data. *International Statistical Review 63*, 215–232.

Marinari, E. and G. Parisi (1992). Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett. 19*, 451–458.

Nott, D. and P. J. Green (2004). Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics 13*, 141–157.

Perrichi, L. R. (1981). A Bayesian approach to tranformations to normality. *Biometrika 68*(1), 35–43.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*, 179–191.

Scott, J. G. and J. O. Berger (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference 136*, 2144–2162.

Smith, M. and R. Kohn (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics 75*, 317–343.

Sweeting, T. J. (1984). On the choice of prior distribution for the box-cox transformed linear model. *Biometrika 71*, 127–134.

Taylor, J. (1986). The retransformed mean after a fitter power transformation. *Journal of the American Statistical Association 81*, 114–118.