

Bayesian Semiparametric Regression: An Exposition and Application to Print Advertising Data

Michael Smith¹, Sharat K. Mathur² and Robert Kohn¹

¹ Australian Graduate School of Management, University of New South Wales

² Booz, Allen and Hamilton, Chicago

First Version June 1996. This Version November 1998

Summary

A new regression based approach is proposed for modeling marketing databases. The approach is Bayesian and provides a number of significant improvements over current methods. Independent variables can enter into the model in either a parametric or nonparametric manner, significant variables can be identified from a large number of potential regressors and an appropriate transformation of the dependent variable can be automatically selected from a discrete set of pre-specified candidate transformations. All these features are estimated simultaneously and automatically using a Bayesian hierarchical model coupled with a Gibbs sampling scheme. Being Bayesian, it is straightforward to introduce subjective information about the relative importance of each variable, or with regard to a suitable data

transformation. The methodology is applied to print advertising Starch data collected from thirteen issues of an Australian women's monthly magazine. The empirical results highlight the complex and detailed relationships that can be uncovered using the methodology.^{1 2}

Key words: Bayesian analysis; Gibbs sampler; Nonparametric regression; Starch print advertising data; Bayesian Variable Selection; Subset Selection.

¹The authors would like to thank Professor John Rossiter for helping them with the classification in section 2.4 and for permission to use the Starch data. The work of Michael Smith was undertaken with the support of an Australian Postgraduate Research Award and a Monash Faculty grant. The work of Sharat Mathur and Robert Kohn was partially supported by an Australian Research Council Grant. We are grateful Professors Gary Lilien, David Midgley, John Roberts and John Rossiter for comments on a previous version of the paper, as well as three anonymous referees whose comments helped improve the paper.

²An Splus compatible package called 'br' that implements the semiparametric approach in this paper was written by Mike Smith and is currently available at the Statlib repository on the world wide web at <http://www.stat.cmu.edu/S/>

1 Introduction.

A large proportion of research in marketing attempts to understand the factors affecting consumer behavior. The number of factors is frequently large and often complex and interrelated. With the growing sophistication of computing tools there is a shift from separated to integrative data bases (Curry 1993) which provide the capability of modeling these complex inter-relationships. Researchers modeling such data as a regression are faced with decisions on key issues, such as which variables and interactions to include in the analysis, the functional form of the effect of each independent variable, and the appropriate transformation for the dependent variable.

This paper introduces a highly flexible Bayesian regression approach to the marketing literature which addresses such issues and has the following advantages over existing approaches to regression.

- (a) Independent variables can enter the model either parametrically or nonparametrically. By parametric we mean that the functional form of the independent variable in the regression is known (for example, an independent variable can enter linearly or as a quadratic), while by nonparametric we mean that the functional form is not prescribed, but is estimated from the data. This is particularly important in modeling marketing data because the form of the relationship between the dependent and any particular independent variable is often unknown, and is usually chosen subjectively.
- (b) The procedure identifies regressors that have a high probability of being significant determinants of the dependent variable. It does so by estimating the posterior probability that each of the regression coefficients are non-zero, conditioned only on the information in the dataset. Traditional approaches to identify significant regressors include all-subsets regression and stepwise regression. How-

ever, all-subsets regression is computationally impractical in datasets with a large number of variables. Stepwise regression is also often unsatisfactory because at each stage only the probability that each regressor is non-zero, conditional on full knowledge of the significance (or otherwise) of other regressors, can be identified. This results in a procedure which considers only a restricted number of subsets and often identifies completely incorrect subsets (Kass and Raftery 1995). Our approach uses a Bayesian hierarchical regression model that is estimated using a Gibbs sampling scheme, an estimation methodology that involves a stochastic search which traverses most of the likely subsets of variables. The result is a variable selection procedure that does not suffer from the unreliability of stepwise regression, nor the computational problems of all-subsets regression. This is particularly useful with the large and complex datasets that often arise in marketing.

- (c) The Bayesian hierarchical model we use enables consideration of datasets where there are large numbers of independent variables, including cases where these are collinear. This can even include cases where there are more regressors than observations, but many of which are redundant.
- (d) An appropriate transformation of the dependent variable is selected from a discrete number of candidate transformations. Traditional regression approaches usually assume that an appropriate transformation of the dependent variable is known before model estimation takes place.
- (e) The functional forms of those variables that are modeled nonparametrically, the effect of the other independent variables, their probability of significance and the choice of the transformation of the dependent variable are estimated simultaneously. It is crucial that these aspects of the regression model are not estimated conditionally upon each other. For example, if an inappropriate transformation is chosen then the wrong variables may be identified as significant and their

effects incorrectly estimated. Conversely, if the wrong variables are identified, then there will be so much variability in the regression estimates that it may be difficult to choose the appropriate data transformation. Current approaches to regression either select the variables, given the data transformation, or choose a data transformation from a family of transformations, but do not select variables at the same time. To the best of our knowledge there are currently no other approaches that simultaneously identify significant variables, fit any required regressors nonparametrically, and select a suitable transformation for the data.

- (f) As this approach is Bayesian, it allows the researcher (and the manager) to introduce into the analysis prior subjective information on the relative importance of each independent variable. This is in addition to the information in the data, and enables the modeler to incorporate ‘soft’ data. It also permits the user to determine the sensitivity of the estimates to various if-then scenarios.

The Bayesian analysis is carried out by using a Bayesian hierarchical regression model that is estimated using a Gibbs sampling scheme. The Gibbs sampler is an estimation procedure similar in scope to maximum likelihood, with introductions given in Gelfand and Smith (1990) and Casella and George (1992). The hierarchical model in this paper was proposed in Smith and Kohn (1996) and improves and extends the work on such models by Mitchell and Beauchamp (1988) and George and McCulloch (1993). It makes it feasible to identify significant regressors from a large number of such variables, whereas this was not practical with previous approaches. Furthermore, previous approaches did not consider data transformations.

The features outlined above make our methodology particularly useful in exploratory analysis of large marketing datasets. We illustrate this by applying it to Starch print advertising data from an Australian women’s monthly magazine. Such data contains

measurements of advertisement (ad) readership, along with a large number of content and product category variables. One feature of this data is that there is likely to be substantial interaction between these two groups of variables, though exactly which variables is unknown. This gives rise to a dataset with a large number of main effects and interaction terms (in our case over two hundred) from which those of key importance require identification. In addition to such interactions, there are regressors (such as position of the ad in the magazine issue) that are unlikely to be related to the readership scores in a linear manner, but where the true functional form is unknown *a priori*. Finally, the readership scores themselves are far from normally distributed, being bounded between zero and one and also skewed, and require an appropriate data transformation. To estimate such a regression model using any existing alternative methodology would require the researcher to subjectively make the decisions about data transformations and functional form before carrying out variable selection using techniques such as forward or backward selection.

The paper is organized as follows. Section 2 describes the data and the proposed model and relates them to the current literature on print advertising. Section 3 describes the Bayesian approach to semiparametric regression. It sets out the Bayesian hierarchical model, discusses the prior assumptions and briefly discusses estimation using the Gibbs sampler. However, for details on the implementation and further discussion of the sampling scheme the reader is referred to Smith and Kohn (1996; 1997). Section 4 reports the results of the application to the Australian print advertising data, while section 5 summarizes the implications of the methodology for marketing researchers. The appendix lists the variables used in the study.

2 Modeling the Starch print advertising data

2.1 Introduction

Creating ads that attract consumer attention to their products is one of the major tasks facing companies. It is particularly important for advertisers to get an insight into the factors affecting readership of print ads (the largest category of ads) because the advertiser has much less control over the consumer's response than in a highly emotive medium like television (Rossiter and Percy 1987). This knowledge will help advertisers in their creative and media strategies.

Researchers have tried to determine what aspects of print advertising affect readership for over three decades. Previous research investigating print advertising can be classified into two groups. The first investigated the influence of specific aspects of print ads, such as the presence or absence of pictures (Edell and Staelin 1983), position in the magazine (Frankel and Solov 1962), type of headline (Soley and Reid 1983), the presence and use of color (Gardner and Cohen 1966) and size (Starch 1966; Trodahl and Jones 1975). Most of this research used controlled experiments, with a few variables investigated at any one time.

The second group of research was more integrative, and examined the collective impact of a number of ad characteristics on measures of ad effectiveness (Twedt 1952; Diamond 1968; Hanssens and Weitz 1980). These studies typically used some measure of recall or recognition as the dependent variable with a number of ad characteristics (both mechanical and content related) as the independent variables.

While previous research added to our understanding of the key drivers of readership, it had two important methodological limitations. First, all the models in the literature assumed that the functional relationship relating the dependent variable and the independent variables was known (except for a few parameters that were estimated from

the data). Most of the research used linear or log-linear regression models (Twedt 1952 and Diamond 1968) or a fixed nonlinear relationship (Hanssens and Weitz 1980). In practice, the form of the relationship for some of the regressors (such as position of the advertisement in an issue) is likely to be nonlinear, but difficult to pre-determine and needs to be estimated from the data. Second, few of the papers explicitly modeled the interaction between the ad content and product category variables, which would result in a large number of interaction terms, but used a multiplicative model to implicitly capture these interactions (Hanssens and Weitz 1980).

Using Bayesian semiparametric regression, such limitations no longer apply, with unknown functional relationships being modeled nonparametrically and the large number of interaction and main effects (even if collinear) entered into the regression and the key determinants identified. As such, the Starch data forms an interesting demonstration of the Bayesian methodology. To give our results face validity, we compare the procedure's predictive ability with some common alternatives.

The data investigated in this paper are collected for all advertisements in each issue of a leading Australian women's monthly magazine. They consist of three attention-relevant scores and about fifty ad content and product characteristics (the variables used in the study are listed in the appendix). The three attention-relevant scores are 'noted' scores (indicating the proportion of respondents who claim to recognize the ad as having been seen by them in that issue), 'associated' scores (indicating the proportion of the respondents who claim to have noticed the advertiser's brand or company name or logo) and 'read-most' scores (indicating the proportion of respondents who claim to have read half or more of the copy.) These have the structure where a noted score is at least as large as the corresponding associated score for an ad, which in turn is at least as large as the read-most score. Therefore, we normalized these so that $y^{(1)} = (\text{raw noted})$, $y^{(2)} = (\text{raw associated}/\text{raw noted})$ and $y^{(3)} = (\text{raw read-most}/\text{raw associated})$ and we use these measures as our dependent

variables.

The data used in the study are for the thirteen months between March 1992 and March 1993. During this period 1030 ads appeared in the magazine, but 25 were dropped because of missing observations on some of the independent variables. The data from the first twelve issues are used to fit the model ($n = 943$ observations), while the data from the thirteenth issue (62 observations) is used as a hold-out sample. The independent variables in the data describe the following four aspects of the ad: (i) the position of the ad in the magazine, including inserts, (ii) the major features of the ad, (iii) the product category, (iv) the message in the ad. How each of these were incorporated into the model is discussed below.

2.2 Variables relating to the position of the ad

Previous research has found the page number of the ad in an issue to be an important factor influencing readership (Diamond 1968). The Starch data include the page number of the ad and the total number of pages in all our issues. This was used to construct the continuous variable

$$P = \frac{\text{page number}}{\text{number of pages in issue}}$$

which takes values between 0 and 1, and is defined as the relative position of the ad in an issue. The front and inside-front cover are coded as the first two pages, while the inside-back and the back cover as the last two pages. Figure 1 plots the raw noted score against the position variable, indicating a distinctly nonlinear relationship.

The usual approach to modeling such a nonlinear response is to assume a specific functional form, for example a linear or a quadratic function. Such an approach is called parametric, as the functional form is specified *a priori*. This paper takes a non-parametric approach instead, in which the shape of a response function f is estimated

from the data. We do so by using a cubic regression spline, which is a piecewise cubic polynomial between m so-called ‘knots’ which partition the domain of the independent variable (in this case $P \in [0, 1]$) into $m + 1$ subintervals. If the location of the m knots is given by the points p_1, p_2, \dots, p_m , then the regression spline approximation for the function f is written as

$$f(P) = b_0 + b_1P + b_2P^2 + b_3P^3 + \sum_{i=1}^m b_{3+i}(P - p_i)_+^3 \quad \text{where } (x)_+^3 = \max(0, x)^3. \quad (2.1)$$

It can be readily shown that such an approximation for the unknown response function f is not only continuous, but has continuous first and second derivatives. Such regression splines are used frequently in practice because they form a linear model where the b_i ’s are regression parameters and the terms $\{P, P^2, P^3, (P - p_1)_+^3, \dots, (P - p_m)_+^3\}$ are terms in a regression.

An important issue with regression splines concerns m , the number of knots. If too few knots are used then f will have ‘local bias’ (that is, where important features of f are missed) as knots will not be located in all the appropriate positions to capture variations in the nonlinear relationship. However, if too many knots are used and the parameters are estimated using least squares, then f will have high ‘local variance’ (that is, the estimate will not be smooth). Our solution is to introduce a lot of knots ($m = 20$), but instead of ordinary least squares we use a hierarchical Bayesian model (as discussed in section 3) that explicitly accounts for the possibility that many of them are redundant. This results in an estimation procedure that provides smooth, flexible and reliable estimates for f (Smith and Kohn 1996).

Alternative nonparametric regression methodologies exist including the popular local regression and smoothing splines, for which general references are Hardle (1990) and Eubank (1988), respectively. However, the methodology presented here has several distinct advantages that lead us to prefer it in the empirical analysis of our Australian Starch data. First, it is locally adaptive in the sense that it allows the curvature of the regression function to change across its domain. Both local regression and spline

smoothing work best when the curvature of the regression function does not vary greatly across the whole domain.

This property is important in identifying the effect of the position variable on readership scores. To demonstrate this, we estimated the univariate nonparametric regression model

$$y^{(1)} = f_{\text{uni}}(P) + \epsilon \quad \text{where } \epsilon \sim \text{iid } N(0, \sigma^2).$$

Figure 1 plots the results of the estimation, with the bold line being the estimate \hat{f}_{uni} from the Bayesian approach discussed above.³ This captures the profile of the effect of the position variable on noted scores, while remaining smooth. The dotted line denotes the estimate from a local regression with a smoothing parameter set so that the peak in the noted scores for the last tenth of the issue could be caught. However, an unavoidable by-product of this is that the rest of the curve is non-smooth. The dashed line corresponds to a local regression where the smoothing parameter is set larger so that the estimate is smooth, but key features of the curve (such as the exposure of the middle fifth and last tenth of magazine) are bleached out.⁴ A further discussion of the comparison of local regression, Bayesian regression and other approaches can be found in Smith and Kohn (1996).

The second advantage of our approach is that there are no alternative nonparametric regression approaches that can extend to the large integrative regression models of the type developed here.

—Figure 1 About Here.—

³This was produced using the Splus compatible package ‘br’ freely available from Michael Smith.

⁴The local regression was undertaken using the default local regression estimator ‘loess’ in Splus with the smoothing parameter set to 0.18 for the first fit and 0.75 for the second. No single smoothing parameter for such an estimator would enable an estimate which is both smooth and captures the underlying profile.

In addition to ads within the magazine, each issue had a few ads that were inserts and thus had no page number. To deal with these a dummy variable I was created for the presence ($I = 1$) or absence ($I = 0$) of inserts. In our integrative regression model, the combined effect of position in the magazine and inserts is:

$$\begin{aligned}
aI + (1 - I)f(P) = \\
aI + b_0(I - 1) + b_1(I - 1)P + b_2(I - 1)P^2 + b_3(I - 1)P^3 \\
+b_4(I - 1)(P - p_1)_+^3 + \cdots + b_{23}(I - 1)(P - p_{20})_+^3, \tag{2.2}
\end{aligned}$$

which is obtained by substituting in a regression spline of the type outlined at (2.1) with $m = 20$ knots. When an ad is an insert, equation (2.2) is equal to an intercept a ; when an ad is within a magazine it is equal to $f(P)$.

2.3 Variables describing the major features of the ad

Previous research investigating the impact of major features of the ad on magazine readership found significant effects for size of the ad (Trodel and Jones 1975; Diamond 1968), presence of pictures (Edell and Staelin 1983) and type of headline (Soley and Reid 1983). There are thirty three independent variables in the data that describe the major features of the ad such as color, size of ad, presence or absence of bleed, headline type, etc. The appendix lists these, including variables X_1, \dots, X_{31} that are mostly binary and therefore enter the regression linearly as $\sum_{i=1}^{31} \beta_i X_i$. The other two variables are the size of the ad (S) and its brand prominence (B) which we model nonparametrically. Each assumes a discrete, but small, number of levels (0-3 and 0-4, respectively) and it is inappropriate for such variables to be modeled using a cubic regression spline (Smith, Sheather and Kohn 1996). Instead, we use a dummy variable basis for each, so that

$$g(S) = \sum_{i=1}^3 c_i \mathcal{I}_i(S) \quad \text{and} \quad h(B) = \sum_{i=1}^4 d_i \mathcal{I}_i(B) \quad \text{where} \quad \mathcal{I}_i(x) = \begin{cases} 1 & \text{if } x = i \\ 0 & \text{otherwise} \end{cases}$$

Here, g and h measure deviations from the base exposure attributable to an ad with $S = 0$ (size less than a full page) and $B = 0$ (brand not present). The terms c_i and d_i are regression coefficients that we estimate using the Bayesian hierarchical model. This ensures g and h measure *significant* deviations from the base cases of $S = 0$ and $B = 0$, as opposed to those that would be obtained using least squares. Therefore, the model for the main effects of the major features of the ad is:

$$\sum_{i=1}^{31} \beta_i X_i + g(S) + h(B) = \sum_{i=1}^{31} \beta_i X_i + \sum_{i=1}^3 c_i \mathcal{I}_i(S) + \sum_{i=1}^4 d_i \mathcal{I}_i(B). \quad (2.3)$$

2.4 Variables describing the product category

It is likely that consumers have different ad readership scores for different product categories. Hanssens and Weitz (1980) addressed this issue by classifying products into three groups— routine, unique, and important products— and modeled each group separately. In contrast, our model allows the simultaneous estimation of product effects by introducing specific interactions between the product type variables and ad content variables. This allows the investigation of product-specific ad characteristics that affect the readership of the ad.

There are twenty-three product types, T_1, \dots, T_{23} , recorded in our Starch data, which are described in the appendix. These variables are binary and enter the model as main effects, but are thought to be mainly of significance as linear interactions with budgetary and design variables. In order to explicitly model these interactions, we grouped the product types into four categories based on the criteria in Rossiter and Percy (1987, pp.166-174)— the type of decision and the type of motivation. Rossiter and Percy (1987) categorize decisions into two broad types— low involvement decisions which have a low level of perceived risk (both economic and psychosocial) for the consumer and high involvement decisions which have a high level of perceived risk. Similarly, there are two types of motivations for product purchases— informational

(which reduce the negative motivation associated with the purchase) and transformational (which increase the positive motivation associated with a purchase). We call the four categories G_1, \dots, G_4 and define them as follows. G_1 : low involvement and informational; G_2 : low involvement and transformational; G_3 : high involvement and informational; G_4 : high involvement and transformational. Each product type T_i is classified as belonging to one of the four categories, with the classification given in the appendix. For example, T_1 is women's apparel and accessories and is classified as G_4 (high involvement and transformational). There are 136 linear interactions of G_1, \dots, G_4 with the ad attributes X_1, \dots, X_{31} , S , B and I ; which we model as simple multiplicative interactions in (2.4) below. We use the Rossiter and Percy (1987) categories to illustrate the methodology, but there are other ways to classify products and the results may vary according to the classification.

The variable R for the presence or absence of recipes was treated as a special case and only its interaction with G_2 (the group that is low involvement and transformational, and includes food and drink products) is included. The following terms were included in the full model:

$$\underbrace{\sum_{i=1}^{23} \beta_i^T T_i}_{\text{product type main effects}} + \underbrace{\beta^R G_2 R}_{\text{recipe interaction}} + \underbrace{\sum_{j=1}^4 G_j (\beta_{Bj} B + \beta_{Sj} S + \beta_{Ij} I + \sum_{i=1}^{31} \beta_{ij} X_i)}_{\text{ad attribute/category interactions}} \quad (2.4)$$

2.5 Variables describing the message in the ad

Twenty eight ad message binary variables $X_i^{[j]}$ were created from four original Starch categorical variables for the purposes of inclusion in the model. These variables are described in the appendix and include predominant feature ($i = 1$), headline appeal ($i = 2$), predominant appeal ($i = 3$) and predominant color ($i = 4$) of the ad. They

enter into the model as main effects and the following terms included:

$$\sum_{i=1}^8 \beta_i^1 X_i^{[1]} + \sum_{i=1}^8 \beta_i^2 X_i^{[2]} + \sum_{i=1}^6 \beta_i^3 X_i^{[3]} + \sum_{i=1}^6 \beta_i^4 X_i^{[4]} \quad (2.5)$$

2.6 Issue number

The data consists of ads published in the thirteen consecutive monthly issues from March 1992 to March 1993. The first twelve issues ($M = 1, \dots, 12$) were used as a calibration sample and the March 1993 issue was used as a hold-out sample for model validation. It seems unreasonable to attempt to estimate any seasonal or trend components because the data used to fit the model spans only a single year. Nevertheless, it is important to control for issue specific effects. Consequently, we modeled the issue effect nonparametrically with a dummy variable basis, so that

$$l(M) = \sum_{i=2}^{12} r_i \mathcal{I}_i(M) \quad (2.6)$$

The function l measures any significant deviations in readership scores for the April 1992- February 1993 issues from that of March 1992 and controls for any potential monthly effects on the training sample. Here, r_i denote regression coefficients and $\mathcal{I}_i(\cdot)$ is as defined in section 2.3.

2.7 Transforming the dependent variables

There are three measures of ad readership in the Starch data: noted, associated and read-most scores, each of which is a proportion. Simple histograms and density estimates demonstrate that the raw scores are far from normally distributed and may require transformation to satisfy both the additivity assumptions in a regression model and the assumptions of a Gaussian error distribution. Previous researchers used transformations such as arcsine (Finn 1988) or logarithmic (Hanssens and Weitz 1980).

These transformations are typically decided in advance and imposed on the data. We considered nine different normalized transformations for the dependent variables of the type

$$T_\lambda(y) = a_\lambda + b_\lambda t_\lambda(y) .$$

In the above, $\lambda = 1, 2, \dots, 9$ indexes our nine transformations and $t_\lambda(y)$ is what we call the ‘base transformation’. This can be any monotonically increasing transformation, which we normalize using two constants a_λ and b_λ to get the actual data transformation $T_\lambda(y)$. These constants can be calculated from the data for each transformation so that the dependent variable has the same median and inter-quartile range pre and post-transformation; see Smith and Kohn (1996) for details. In effect, these constants normalize the base transformations considered to make them scale and location invariant— a property that eases the qualitative interpretation of the regression estimates.

Table 1 presents the base transformations that we consider, which include those of Finn (1988) and Hanssens and Weitz (1980). The other seven are of the form $\Phi^{-1}(y^\theta)$, which are probit transformations with various values for the skewing parameter θ . It is important to note that the methodology allows a data-driven selection of the most appropriate normalized transformation from those listed in table 1. A user could well select any finite number of transformations considered likely and use the procedure to select between these. As implemented, our methodology may lead to different transformations for the noted, associated and read-most scores.

—Table 1 About Here.—

3 Estimation Methodology

3.1 Introduction

All the terms constructed in the previous section listed at equations (2.2)-(2.6) can be collected together to form a large design matrix X with $p = 262$ terms for each of the three regressions

$$T_\lambda(\mathbf{y}^{(j)}) = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{for } j = 1, 2, 3. \quad (3.1)$$

Here, we consider a regression for each of the three transformed readership scores; so that $T_\lambda(\mathbf{y}^{(1)})$, $T_\lambda(\mathbf{y}^{(2)})$ and $T_\lambda(\mathbf{y}^{(3)})$ are n -vectors of the observations of the noted, associated and read-most scores, respectively, transformed according to the transformation T_λ . The vector $\boldsymbol{\beta}$ is made up of the regression coefficients introduced earlier (the a , b_i 's, c_i 's, d_i 's, r_i 's and various β 's) while the vector $\boldsymbol{\epsilon}$ is made up of errors distributed iid $N(0, \sigma^2)$.

Estimating such a regression using least squares presents a number of problems. First, there is often collinearity between the regressors. Second, even when collinearities are eliminated, variability in the regression coefficient estimates would be large due to a lack of degrees of freedom. Third, the estimates for the nonparametric function components f , g , h and l would be non-smooth.

Commonly used approaches for tackling this problem are stepwise regression and all-subset regression, both of which are severely limited. All-subset regression requires 2^p regressions to be undertaken—something that is computationally infeasible. The stepwise procedure, while feasible, traverses only a small number of subsets, which often leads to the situation that forward and backward stepwise procedures result in substantially different estimates (Kass and Raftery 1995).

All these problems lead us to use a Bayesian hierarchical model, coupled with a Gibbs sampler, to estimate this large regression model. This approach can handle

a large number of variables ($p = 262$ in this case), even when collinearity exists, without being hindered by the problem that there are only about four times as many observations as coefficients to be estimated. It can do so because it reduces the effective dimension of the problem by explicitly modelling the possibility that many of the terms in the design matrix may be redundant. Using the Gibbs sampler to undertake the computations results in the redundant variables being identified reliably. This is because the search over the distribution of important subsets is stochastic, rather than deterministic as in stepwise regression, and it traverses many more likely combinations of regressors than the stepwise regression approach. Furthermore, all other approaches require pre-specification of the data transformation, whereas the Bayesian methodology can simultaneously select the appropriate data transformation and identify significant variables.

3.2 The Bayesian hierarchical model

This section briefly describes the Bayesian hierarchical model, with a full exposition being found in Smith and Kohn (1996). The key idea behind such a model is that we explicitly model the possibility that terms in the regression at (3.1) may be superfluous and can be omitted from the regression. To do this, we use the vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_p)'$ of indicator variables, where each element is defined as

$$\gamma_i = \begin{cases} 0 & \text{if } \beta_i = 0 \\ 1 & \text{if } \beta_i \neq 0 \end{cases}$$

That is, γ_i is a binary variable that indicates whether, or not, β_i is zero and therefore whether, or not, the corresponding independent variable enters the regression. Therefore, the vector $\boldsymbol{\gamma}$ explicitly parameterises the subsets of regressors in the linear regression at (3.1). The regression can be rewritten, conditional on this ‘subset’ parameter, as

$$T_\lambda(\mathbf{y}^{(j)}) = X\boldsymbol{\gamma}\boldsymbol{\beta}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \quad \text{for } j = 1, 2, 3.$$

Here, β_γ are the regressors that are non-zero, given γ , and X_γ is a design matrix made up of the corresponding columns of X .

Because this is a Bayesian methodology it is necessary to place prior distributions on the unknown parameters γ , λ , β and σ^2 . Ideally, such priors are ‘uninformative’, unless there is some real information on a parameter or group of parameters (for example, from previous research). The priors we use here are listed below.

- (i) The prior distribution of β_γ conditional on γ , σ^2 and λ is $N(0, n\sigma^2(X'_\gamma X_\gamma)^{-1})$. The variance matrix in this prior is simply that of the least squares estimate of β_γ blown up by a factor of n and is therefore relatively uninformative. A fully uninformative prior for β_γ cannot be used as it results in what is commonly called Lindley’s paradox (Mitchell and Beauchamp 1988).
- (ii) The prior density for σ^2 given γ and λ is proportional to $1/\sigma^2$. This is a commonly used prior for σ^2 and means that $\log \sigma^2$ has a flat prior or uninformative prior (Box and Tiao 1973).
- (iii) We assume that given λ , the γ_i are *a priori* independent with $Pr(\gamma_i = 0) = \pi_i$, $i = 1, \dots, p$. The probabilities π_i are specified by the user. In this paper we take $\pi_i = 1/2$ for all i which can be considered uninformative. However, our methodology allows for a user to impose a prior assessment of the relative importance of individual terms in the form of a non-uniform prior on γ . For example, if X_7 (photo used) is considered more likely to have an impact on ad recognition than other variables, a higher value of π_i could be imposed on that variable.
- (iv) We assume that each of the nine candidate transformations is equally likely, that is, $Pr(\lambda = i) = 1/9$ for $i = 1, 2, \dots, 9$. This is a uninformative prior for λ , although if a researcher wants to incorporate prior beliefs, some transformations can be given a higher prior probability than others.

3.3 Estimation

Smith and Kohn (1996) develop a Gibbs sampling scheme to estimate this Bayesian hierarchical model and show how it can be implemented. The sampling scheme produces a Monte Carlo sample of the subset parameter γ from its distribution conditioned only on the data; that is, from its posterior distribution. As this is by far the most problematical of the parameters in the model, the sample enables Monte Carlo estimation of the posterior probability $\Pr(\lambda = i|\text{data})$.

From this distribution we can pick the single most likely transformation and label it the ‘mode’ transformation $\hat{\lambda}_M$. Following Box and Cox (1982) we estimate the remaining features of the model conditional on this best transformation. The sampling scheme is run again and another Monte Carlo sample for the subset parameter γ obtained. Using this, estimates of the expected value of β given the data (called the posterior mean), smoothing over the distribution of γ , can be obtained; that is, we estimate $E(\beta|\lambda = \hat{\lambda}_M, \text{data})$. Estimates of the probability each term in the regression is non-zero (and therefore significant), given the data, can also be calculated; that is, we estimate $\Pr(\gamma_i = 1|\lambda = \hat{\lambda}_M, \text{data})$. Our implementation followed that outlined in Smith and Kohn (1996), but where our sampling scheme was run for 10,000 iterations for convergence and 20,000 iterations to obtain the Monte Carlo sample of γ . This is a conservative sampling length and took approximately four hours on a standard DEC UNIX workstation. Similar results were obtained with sampling runs of one tenth the length and took under one hour to run.

This produces estimates of the transformation and the regression parameters that are robust to the fact that it is unknown *a priori* which terms actually are in the regression. This is in contrast to least squares regression which estimates the regression coefficients given some particular known subset, or value for, γ , as well as a pre-specified data transformation.

4 Empirical Estimates and Methodological Comparison

4.1 Introduction

We applied our estimation methodology to the integrative regression models for the Starch scores noted ($y^{(1)}$), associated ($y^{(2)}$) and read-most ($y^{(3)}$). The first 943 observations, consisting of the March 1992 till February 1993 issues, are used for estimation. The March issue, consisting of 62 ads, is used for model validation. Sections 4.2-4.5 below discuss various aspects of the model estimates. Section 4.6 discusses the predictive performance of the model on the hold-out sample compared to the application of standard regression approaches to such a dataset, while Section 4.7 discusses some of the managerial implications of the empirical results.

4.2 Data transformation estimates

The most likely transformations, given the data, were (once normalized) $\sin^{-1}(y^{0.5})$, $\Phi^{-1}(y^{1.5})$ and $\sin^{-1}(y^{0.5})$ for the noted, associated and read-most scores, respectively. The posterior probabilities of all nine candidate transformations for each of the three scores are given in table 2.

—Table 2 and Figure 2 About Here.—

The selected transformations normalize the errors effectively, as shown in figures 2(a)-(c) by the normal probability plots of the residuals from the three fits. The results confirm that the asymmetric arcsine transformation discussed by Finn (1988) is the most likely (in that it has the highest posterior probability) of those considered here for the noted and read-most score regressions. However, in the associated score regression

a skewed probit transformation is the most likely, given the data. The logarithmic transformation proposed by Hanssens and Weitz (1980) is the most unlikely, with a posterior probability equal to zero (to three decimal places) for all three scores in our dataset.

The posterior probabilities of the transformations seem to be fairly insensitive to choice of the prior distribution. For example, we used an informative prior that ascribed double the probability to the seven skewed probit transformations as prescribed for the arcsine and logarithmic. That is, the prior $\Pr(\lambda = i) = 1/8$ for $i = 1, \dots, 7$ and $\Pr(\lambda = 8) = \Pr(\lambda = 9) = 1/16$. The data transformations selected using this informative prior were the same as those using the uninformative prior; although the posterior probabilities change slightly (see table 2). These results are reassuring because they demonstrate that the transformation selection is based on information from the data and is not ‘prior driven’.

Although these three different transformations make it difficult to compare parameter magnitudes across scores, it is possible to examine the variation of ad characteristics that determine the noted, associated and read-most scores.

4.3 Impact of variables related to the position of the ad

Figures 3(a.1), (a.2) and (a.3) outline the impact of P , the position of the ad in the issue, on ad recognition scores. It is evident from these figures that the position of the ad is a determinant of noted and associated scores, but is not related to the read-most score.

—Figure 3 About Here.—

Figure 3(a.1) is a plot of the regression spline estimate of the response f for the noted score $y^{(1)}$. This estimate suggests that advertising in the front end of the maga-

zine (the first 15%) is immensely beneficial from the point of view of attracting casual attention to the ad leading to high noted scores. However, there is a dramatic decrease in exposure as ads are placed further into the magazine, up to about 20% in from the front page. This is consistent with previous literature (Diamond 1968; Hanssens and Weitz 1980) which showed a significant negative coefficient for the page number variable. However, the nonparametric regression is able to capture the more sophisticated relationship between noted scores and position than the fixed linear relationships imposed in prior research. After the rapid drop that occurs in the front of the magazine, there is a resurgence in ad exposure for the middle 40% of the magazine. After this, positioning an ad 80%-90% into the magazine appears to be the worst choice, while the last 5% seems to be an improvement. Notice that the estimate for the effect is slightly different than that uncovered in the univariate nonparametric regression found in figure 1 because the current regression controls for other effects that are correlated with the position variable, as well as having a transformed dependent variable.

Figure 3(b.1) provides the estimate of f for the associated scores $y^{(2)}$. The profile of the estimate differs substantially from that of the noted score and indicates that for high associated scores the worst place to advertise is 70%-90% into the magazine, while the last 5% is advantageous.

The insert dummy variable I was a significant contributor to all three scores, with table 3 providing the insert main effects and interactions with the product categories that had a greater than 35% chance of being non-zero (ie: $\Pr(\gamma_i = 1 | \text{data}, \lambda = \hat{\lambda}_M) > 0.35$). Also included in the table is the estimated intercept of $f(P)$ from (2.2), namely the coefficient \hat{b}_0 of $(1 - I)$. Notice that inserts have a similar effect as advertising on the front cover of the magazine, though for the associated and read-most scores they achieve slightly higher exposure.

—Tables 3 and 4 About Here—

4.4 Impact of the major features of the ad

The size of the ad has a strong effect on the noted scores, but the marginal gain decreases after a full page ad ($S = 1$). This is illustrated in figure 3(a.2) which plots the estimate of the main effect, g , of ad size as the solid curve. Interestingly, for products that are classified as ‘high involvement and transformational’ (for example, fashion apparel and accessories, household furnishings, and jewellery), the size of the ad has a much more pronounced effect. When combined with the main effect, it produces the relationship represented by the long dashed line. This is caused by a value of 0.045 for the slope of the interaction G_4S which is identified as having a posterior probability of being non-zero of 0.946.

For the associated scores, ad size has a marginal main effect, although some effect may occur on products that are high involvement and transformational as there is a posterior probability of 0.411 that SG_4 is non-zero and positive; something that is reflected in the interaction plots in figure 3(b.2). Figure 3(c.2) shows that for read-most scores there is a very flat ad size main effect. However, in a similar manner to associated scores, the figure indicates transformational products react positively to increases in ad size. This is caused by positive slopes for SG_2 and SG_4 which are non-zero with a probability of 0.649 and 0.765, respectively.

The relationships between the three scores and brand prominence are shown in figures 3(a.3),(b.3) and (c.3). For both the noted and associated regressions there are strong nonlinear main effects, while no significant interaction effects with any of the four product categories. In particular, for associated scores it seems that brand names have little effect unless they are highly prominent. For read-most scores, brand prominence is negatively related for high involvement and informational products, with the slope of BG_3 being -0.039 with a posterior probability of being non-zero of 0.863.

Table 4 provides an interpretation of the empirical results of our analysis. The

table lists all the slope coefficients for terms in the three regressions (excluding those involving P, I, B, S and M) which have a posterior probability greater than 0.35 of being non-zero. This indicates that a number of ad characteristic variables significantly impact on ad recognition scores for this magazine, irrespective of product type. For example, variables such as right hand position of the ad and headline color are likely to be related to noted scores for all products.

4.5 Product specific effects

It is apparent from table 4 that product specific effects are very important for the three models. First, there are products types which are significant main effects. For example, T_{12} (motor vehicles) has a negative coefficient in all three regressions and a posterior probability (to three decimal places) of being non-zero of 1.000, 0.530 and 1.000 for noted, associated and read-most scores. These, and the other product main effects found in table 4, appear consistent with the positioning of the magazine as primarily for women with a focus on domestic and entertainment issues. Second, the product specific effects manifest themselves as interactions between ad characteristics and the product categories G_1, G_2, G_3 and G_4 defined in section 2.4. This has already been seen, to some extent, with their interaction effects with B (brand prominence) and S (size of ad). However, there are many additional interaction terms found in table 4. For example, for G_2 (low involvement and transformational) products more than fifty words in the copy (X_{24}) appears to provide a decline in read-most scores, with a posterior probability of 0.996 of having a non-zero slope coefficient. Table 4 provides a large number of other such interaction effects between product categories and both ad attribute and ad message variables.

4.6 Model validation

To demonstrate the effectiveness of our estimation methodology we compare its predictive performance to some other alternative estimation procedures. The prediction is for the 62 ads in the hold-out sample of the subsequent March 1993 issue of the magazine and is undertaken for all three readership scores. Throughout, we use the transformed dependent variables (as selected by our procedure) to enable a fair comparison of the approaches.

The first alternative is a ordinary least squares regression based on the 260 linearly independent regressors in our sample, which is included simply as a benchmark. The second is a forward selection stepwise regression procedure coupled with Akaike's information criteria, or AIC, (Akaike 1978). Here, the model uncovered during the forward selection procedure with the maximum AIC value was selected as our regression model and least squares used to estimate the regression coefficients. The last alternative uses factor analysis on the 260 linearly independent regressors in our sample to reduce the dimensionality of the problem and then uses least squares to estimate the coefficients for these factors. We used maximum likelihood factor analysis (Mardia, Kent and Bibby 1979) with 30, 50 and 100 factors, which explained 54.62%, 66.71% and 81.85%, respectively, of the variation in the design.

Using these procedures we forecast the (transformed) readership scores and calculated (i) the sample correlation between these and the (transformed) actual readership scores, (ii) the mean absolute error and (iii) the mean squared error. These are found in table 5 and, by all measures, the Bayesian approach results in the most accurate forecasts. While these results demonstrate the advantage of using the Bayesian approach, this is, of course, specific to this single dataset. For comprehensive simulations on the reliability of such a semiparametric regression approach we refer the reader to Smith and Kohn (1996; 1997).

4.7 Managerial Implications

The methodology provides a way in which many of the design and budgetary issues facing prospective advertisers of this Australian women's magazine can be addressed. For example, the estimated effect of the position of the ad in an issue (P) on its noted score, found in figure 3(a.1), suggests that an advertiser who is interested in attaining high levels of casual exposure may be prepared to pay a premium to place the ad in the front 10% of the magazine. However, the nonlinearity of the relationship also indicates that locating the ad 40% into the magazine is also an advantageous position. If a more in-depth exposure is required, the estimated effect of the position variable on read-most scores found in figure 3(c.1) indicates that it is not worth any extra cost to position the ad at any particular location in the magazine.

Identifying such nonlinearities is important in identifying changing marginal relationships between ad characteristics and readership scores. For example, consider an advertiser who is faced with a decision on how prominent to make the brand of a product. Figures 3(a.3) and (b.3) suggest that, on average, making it 'impossible to miss' ($B = 4$) results in higher noted and associated scores than 'not present' ($B = 0$). However, the nonlinear relationship suggests that placing the brand somewhat in the middle of these extremes, as 'easy to miss' ($B = 2$), is no different than omitting it completely. It does not result in half the level of noted and associated exposure as would be suggested by a model that imposed a linear relationship.

The ability of the methodology to incorporate a large number of product type interaction variables into a regression framework with a relatively low sample size also proves important. It enables a more detailed modeling of the complex interaction effects and provides a measurement of the relationships between product type, ad

characteristics and exposure. For example, while figure 3(a.2) confirms that larger ads obtain higher casual attention, it is particularly important for products that are categorized as high involvement and transformational. Moreover, figure 3(c.2) identifies that higher in-depth exposure (as measured by the read-most score) can be obtained by larger ads when the product is transformational, but not when it is informational.

Advertising managers can also obtain predictive exposure ratings for proposed ads, given decisions on design and message factors. This is viable because the methodology allows a reliable estimation of the collective impact of ad characteristics. Such reliability is reflected in the relative improvement in predictive ability found for the ads in the March 1993 issue. Ad design can then be tailored to maximize this predicted exposure, given a product type and budgetary constraints.

5 Summary and conclusion.

This paper provides a new modeling approach to regression which allows the data to determine the functional form of the independent variables and to identify the significant independent variables, as well as an appropriate transformation of the dependent variable. It can handle regressors that lead to a regression matrix that is collinear (such as in the Starch print advertising dataset examined here) because it explicitly accounts for the possibility that many of the regressors are redundant. It reduces the subjective requirements of traditional techniques in pre-specifying functional forms and transformations of the dependent variable. While an uninformative prior is used in this analysis, if a researcher has strong prior information on which regressors are important, or which data transformation is most likely, then it can be easily incorporated into the analysis.

These features of the methodology prove useful in our analysis of a dataset from an

Australian women's monthly magazine. They enable the development of an integrative model which can capture the complex inter-relations between ad characteristics and exposure measures that exist in such data. All features of the model are estimated simultaneously, the reliability of which is reflected in its improved predictive performance, relative to standard regression techniques. The results provide an insight into the form of the relationship between variables such as position of the ad in an issue, size of the ad and brand prominence and the readership scores. They also identify product specific effects through the use of interaction terms based on a categorization suggested by Rossiter and Percy (1987). As such, they have implications for the design of ads by potential advertisers in this magazine. More generally, with the growth of complex databases in marketing, we feel that our technique provides the researcher with a powerful modeling tool which is practical, easy to implement, and can be applied to complex datasets containing a large number of variables, such as Starch print advertising data.

Appendix: Description of the data

There were 1005 observations. The first 943, March'92–Feb'93, formed the calibration sample. The remaining advertisements, numbers 944-1005 from March'93, were used used for model validation. The following is a list of the independent variables.

Advertisement attribute variables.

X_1 color (0-2), 0=monotone, 1=2-color, 2=4-color

X_2 left-hand position, (0,1)

X_3 right-hand position, (0,1)

X_4 size of Issue, (0-2) 0=296 pages or less, 1=300 to 320 pages, 2=more than 320 pages

X_5 square finish in main illustration (0,1)

X_6 silhouette or other shape in main illustration (0,1)

X_7 photo used (0,1)

X_8 size of photo (0,1), 0=<1/2 of space

X_9 ad content relates to 'end result of using product' (0,1)
 X_{10} ad content relates to 'finished result of using or eating the product' (0,1)
 X_{11} ad content relates to 'actual product or package' (0,1)
 X_{12} number of illustrations (0,1), 0=multi-illustrations, 1=single main illustration
 X_{13} bleed (unbordered) (0,1)
 X_{14} headline length (0,1), 0=<8 words
 X_{15} large headline (0,1), 1=>1.3cm high
 X_{16} small headline (0,1), 1=<1.3cm high
 X_{17} headline above main illustration (0,1)
 X_{18} headline beside main illustration (0,1)
 X_{19} headline below main illustration? (0,1)
 X_{20} headline color (0-2), 0=no color, 1=partial color, 2=all color
 X_{21} headline type (0-2), 0=all reverse, 1=partial reverse, 2=all straight
 X_{22} headline case (0-2), 0=all lower, 1=partial upper, 2=all upper
 X_{23} copy area (0,1), 1=>1/3 area
 X_{24} , number of words in copy (0,1), 0=<50 words
 X_{25} opposite page both not part of the advert and color (0,1)
 X_{26} opposite page is mainly other advert (0,1)
 X_{27} opposite page is mainly related editorial (0,1)
 X_{28} opposite page is mainly unrelated editorial (0,1)
 X_{29} opposite page is mainly part of the same multi-page ad (0,1)
 X_{30} advertorial (0,1)
 X_{31} coupon (0,1)
 P position in issue, continuous on $[0, 1]$, 0=front cover, 1=back cover
 I insert, (0,1)
 S size of ad (0-3), 0=Less than full page, 1=full page, 2=double page, 3=multiple pages
 B brand prominence (0-4), 0=not present, 4=impossible to miss
 R recipe (0,1)
 M issue number (1-13), 1=March 1992, 13=March 1993

Product Type Independent variables.

All the variables take the values 0 and 1 only, where 1=true and 0=false. The category to which each product belongs is provided in brackets. The product categories are G_1 (low involvement & infor-

mational), G_2 (low involvement & transformational), G_3 (high involvement & informational) and G_4 (high involvement & transformational).

$T_1(G_4)$ women's apparel/accessories	$T_2(G_2)$ food	$T_3(G_3)$ pharmaceutical
$T_4(G_1)$ toiletries and health	$T_5(G_3)$ household appliances	$T_6(G_4)$ household furnishings
$T_7(G_1)$ household goods	$T_8(G_4)$ building and decorating	$T_9(G_4)$ travel and holidays
$T_{10}(G_4)$ mens apparel/accessories	$T_{11}(G_4)$ children's wear	$T_{12}(G_4)$ motor vehicles
$T_{13}(G_2)$ books and magazines	$T_{14}(G_2)$ drink	$T_{15}(G_2)$ cosmetics/beauty
$T_{16}(G_1)$ hair	$T_{17}(G_4)$ jewellery	$T_{18}(G_2)$ records
$T_{19}(G_3)$ government	$T_{20}(G_2)$ craft	$T_{21}(G_1)$ baby
$T_{22}(G_1)$ pets	$T_{23}(G_3)$ finance	

Ad message dummy variables

1. *Predominant Feature.*

Each ad can have only one predominant feature, but this may have several of the following characteristics: $X_1^{[1]}$ person(s) 18+; $X_2^{[1]}$ baby(ies); $X_3^{[1]}$ child(ren); $X_4^{[1]}$ animal(s); $X_5^{[1]}$ women's fashions; $X_6^{[1]}$ furnishings; $X_7^{[1]}$ a premium offer; $X_8^{[1]}$ the product(s).

2. *Headline Appeal.*

Again, each ad can have only one headline, but may have any of the following appeals: $X_1^{[2]}$ a promise; $X_2^{[2]}$ new feature of the product; $X_3^{[2]}$ a question; $X_4^{[2]}$ an exclamation; $X_5^{[2]}$ news item; $X_6^{[2]}$ prize(s) in a contest; $X_7^{[2]}$ price; $X_8^{[2]}$ new product.

3. *Predominant appeal.*

This can have any of the following characteristics: $X_1^{[3]}$ social status; $X_2^{[3]}$ increased pleasure; $X_3^{[3]}$ personal security; $X_4^{[3]}$ health or hygiene; $X_5^{[3]}$ knowledge or quality; $X_6^{[3]}$ increased income/capital gain/savings.

4. *Predominant color of advertisement.*

The registered colors are: $X_1^{[4]}$ red; $X_2^{[4]}$ blue; $X_3^{[4]}$ green; $X_4^{[4]}$ grey, black or monotone; $X_5^{[4]}$ pastel shades, no predominant color; $X_6^{[4]}$ yellow, orange or brown.

References

- Akaike, H., (1978), "A Bayesian analysis of the minimum AIC procedure," *Annals Institute Statistical Mathematics*, 30, 9-14
- Box, George and Cox, David, (1982), "An analysis of transformations revisited, rebutted", *Journal of the American Statistical Association*, 77, 209-252.
- Box, George, and Tiao, George, (1973), *Bayesian inference in statistical analysis*, New York: John Wiley & Sons.
- Casella, George and George, Edward (1992), "Explaining the Gibbs sampler," *The American Statistician*, 46, 167-174.
- Curry, David J. (1993), "The New Marketing Research Systems: How To Use Strategic Database Information for Making Better Marketing Decisions," New York: John Wiley & Sons.
- Diamond, Daniel S. (1968), "A Quantitative Approach to Magazine Advertisement Format Selection," *Journal of Marketing Research*, 5, 376-86
- Edell, Julie A. and Staelin, Richard (1983), "The Information Processing of Pictures in Print Advertising," *Journal of Consumer Research*, 10(June) 45-61.
- Eubank, Randolph L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.
- Frankel, Lester R. and Solov, Bernard M. (1963), "Does Recall of an Advertisement Depend Upon its Position in the Magazine ?", *Journal of Advertising Research*, 2 (December), 28-32.

Finn, Adam (1988), "Print Ad Recognition Readership Scores: An Information Processing Perspective," *Journal of Marketing Research*, XXV(May), 168-177.

Gardner, Burleigh B. and Cohen, Yehudi A. (1964), "ROP Color and Its Effect on Newspaper Advertising," *Journal of Marketing Research*, 1 (May), 68-70.

Gelfand, Alan E. and Smith, Adrian F. M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398-409.

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

Hanssens, Dominique M. and Weitz, Berton A. (1980), "The effectiveness of Industrial Print Advertisements Across Product Categories," *Journal of Marketing Research*, 17, 294-306

Hardle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press
Econometric Society Monographs

Kass, Robert E. and Raftery, Adrian E. (1995), "Bayes Factors", *Journal of the American Statistical Association*, 90, 773-795.

Mardia, Kantilal, Kent. J. and Bibby, John (1979), *Multivariate Analysis*, London: Academic Press

Mitchell, T. J. and Beauchamp, J. J. (1988) "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1023-1036

Rossiter, John R. and Percy, Larry (1987), *Advertising and Promotion Management*, New York: McGraw-Hill.

Smith, Michael (1996), *Nonparametric Regression: A Markov chain Monte Carlo Approach*, Unpublished PhD Thesis, University of New South Wales.

Smith, Michael and Kohn, Robert (1996), "Nonparametric regression using Bayesian variable selection," *Journal of Econometrics*, 75, 317-344.

Smith, Michael and Kohn, Robert (1997), "A Bayesian Approach to Nonparametric Bivariate Regression", *Journal of the American Statistical Association*, 92, 1522-1535.

Smith, Michael, Sheather, Simon and Kohn, Robert (1996), "Finite sample performance of robust Bayesian regression," *Journal of Computational Statistics*, 11, 3, 317-343.

Soley, Lawrence C. and Reid, Leonard N. (1983), "Industrial Ad Readership as a Function of Headline Type," *Journal of Advertising*, 12 (1), 34-38.

Starch, Daniel (1966), "Measuring Advertising Readership and Results," New York: McGraw-Hill.

Trodahl, Verling C. and Jones, Robert L. (1965), "Predictors of Newspaper Advertising Readership," *Journal of Advertising Research*, 5 (March), 23-7.

Twedt, Dik W. (1952), "A Multiple Factor Analysis of Advertising Readership," *Journal of Applied Psychology*, 36 (June) 207-15.

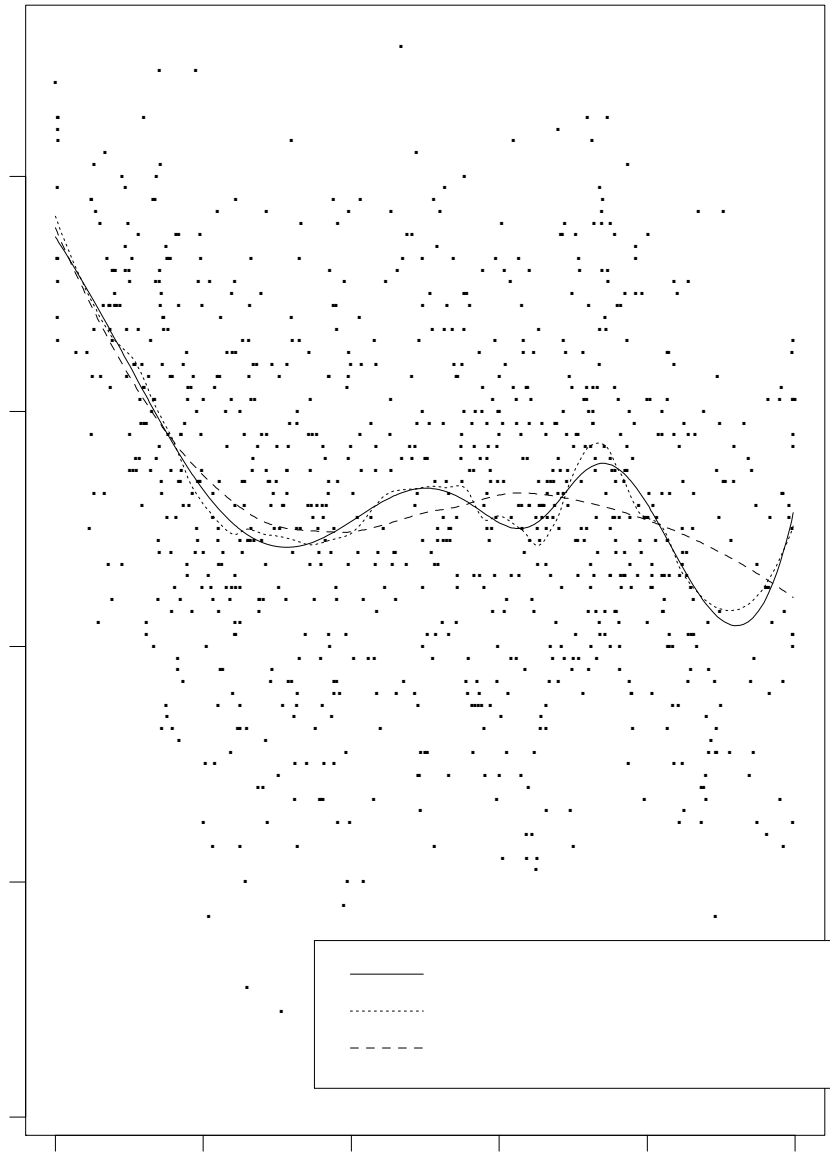


Figure 1: The scatter plot is of the noted scores $y^{(1)}$ against position in issue P . The bold line is the Bayesian regression fit to the data, the dotted is from a local regression with a low smoothing parameter value and the dashed is from a local regression with a higher value for the smoothing parameter. No smoothing parameter value would enable a local regression to simultaneously capture the curvature of the relationship, while remain smooth.

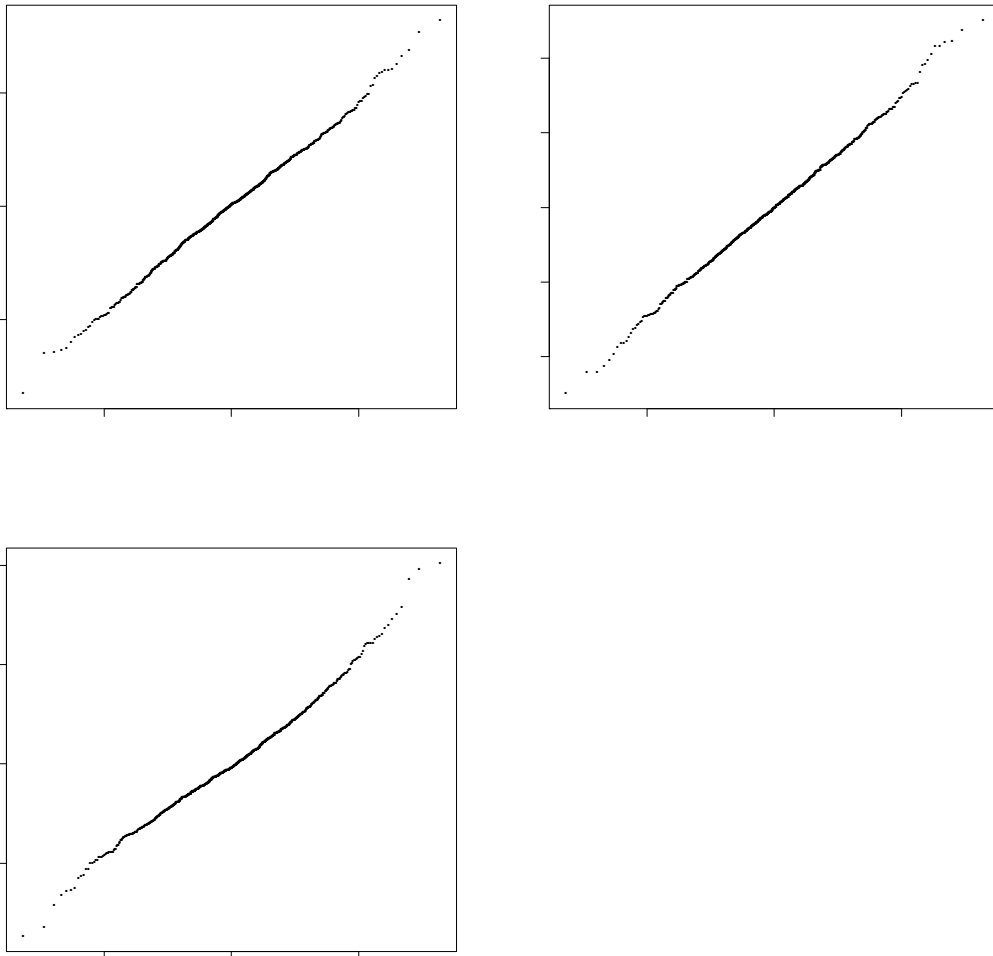


Figure 2: (a)-(c) Normal probability plots based on the residuals obtained from each of the three regressions. If the quantiles of a normal and the observed residuals are linearly related with a slope of 45 degrees, then this indicates that the residuals follow a normal distribution.

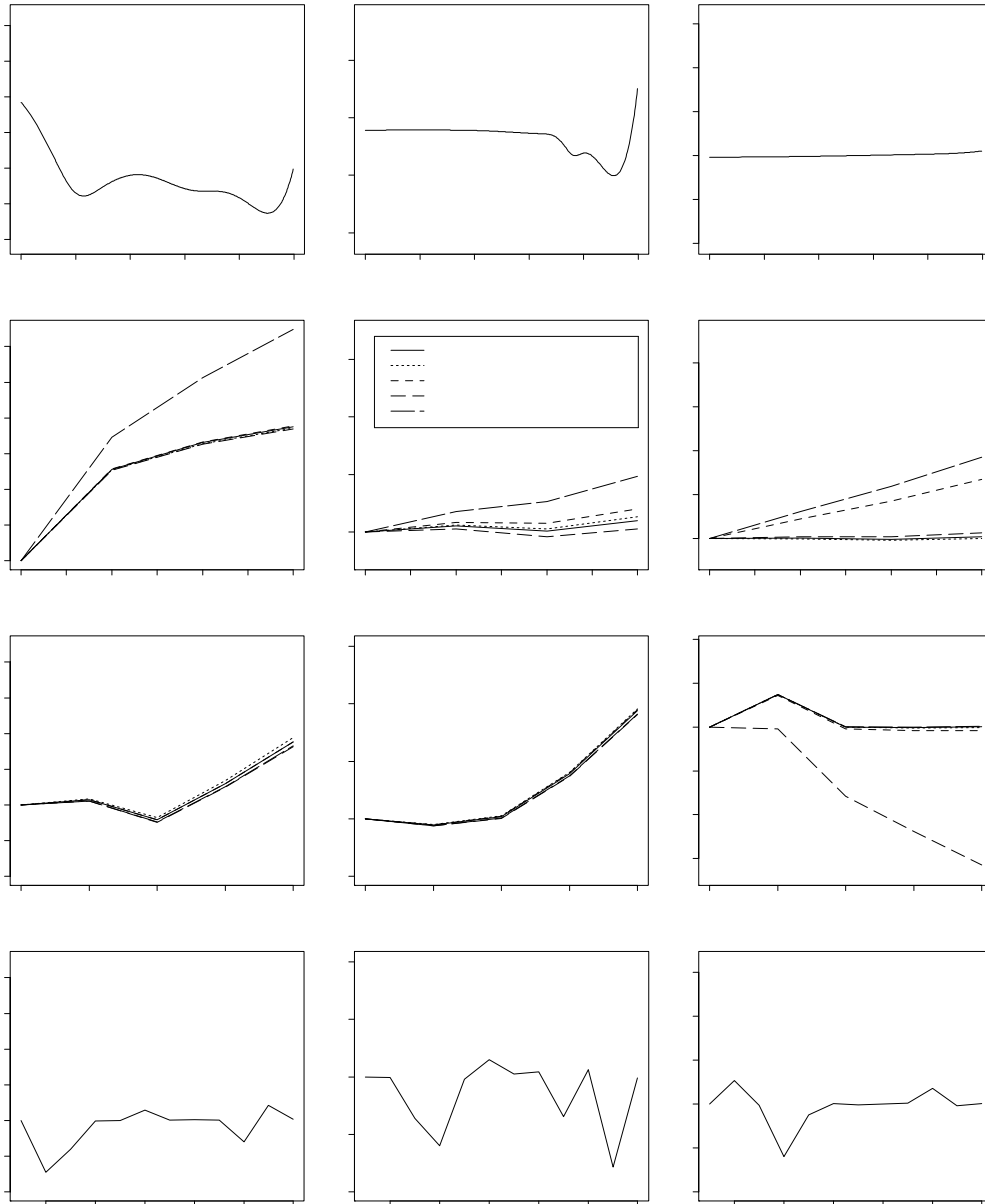


Figure 3: Estimates of the functions f, g, h and l for each of the three regressions. The rows correspond to the four functions, while the three columns correspond to the three regressions arising from the dependent variables $y^{(1)}, y^{(2)}$ and $y^{(3)}$, respectively. The figures in each column are plotted with the same range on the vertical axis, so that an idea of the comparative strength of the effects within each regression can be obtained. This range is set equal to the distance between the 85/100 quantile and 15/100 quantile of the transformed dependent variable of each regression (that is, $T_9(y^{(1)}), T_6(y^{(2)})$ and $T_9(y^{(3)})$). The main effects are in bold, while interactions are also plotted for the effects of B and S .

λ	$t_\lambda(y)$	For Noted Score		For Associated Score		For Read-Most Score	
		a_λ	b_λ	a_λ	b_λ	a_λ	b_λ
1	$\Phi^{-1}(y^{0.1})$	-0.516	0.678	-0.124	0.459	-0.537	0.692
2	$\Phi^{-1}(y^{0.25})$	-0.063	0.564	0.165	0.396	-0.072	0.573
3	$\Phi^{-1}(y^{0.5})$	0.240	0.478	0.368	0.346	0.237	0.483
4	$\Phi^{-1}(y^{0.75})$	0.398	0.428	0.478	0.316	0.397	0.431
5	$\Phi^{-1}(y)$	0.500	0.394	0.552	0.295	0.500	0.395
6	$\Phi^{-1}(y^{1.5})$	0.631	0.347	0.649	0.266	0.630	0.347
7	$\Phi^{-1}(y^2)$	0.713	0.315	0.712	0.246	0.712	0.315
8	$\log(1+y)$	0.819	0.625	0.892	0.873	0.798	0.584
9	$\sin^{-1}(y^{0.5})$	-0.278	0.991	-0.117	0.820	-0.281	0.994

Table 1: The first column gives the transformation index, while the second the base transformations. The remaining columns gives the constants required to make the normalized transformation scale and location invariant. These constants are provided for each of the three dependent variables considered in our dataset. The last two transformations were proposed by Finn (1998) and Hanssens and Weitz (1980)

Base Transformation	Uninformative Prior			Informative Prior		
	For $y^{(1)}$	For $y^{(2)}$	For $y^{(3)}$	For $y^{(1)}$	For $y^{(2)}$	For $y^{(3)}$
$\Phi^{-1}(y^{0.1})$	0.004	0.009	0.000	0.004	0.009	0.000
$\Phi^{-1}(y^{0.25})$	0.027	0.043	0.000	0.030	0.043	0.000
$\Phi^{-1}(y^{0.5})$	0.084	0.116	0.000	0.095	0.115	0.000
$\Phi^{-1}(y^{0.75})$	0.129	0.176	0.002	0.148	0.176	0.003
$\Phi^{-1}(y)$	0.153	0.213	0.006	0.178	0.213	0.011
$\Phi^{-1}(y^{1.5})$	0.158	0.231	0.034	0.186	0.232	0.059
$\Phi^{-1}(y^2)$	0.139	0.211	0.102	0.168	0.212	0.176
$\log(y + 1)$	0.000	0.000	0.000	0.000	0.000	0.000
$\sin^{-1}(y^{0.5})$	0.305	0.001	0.856	0.190	0.000	0.751

Table 2: Posterior posterior estimates for each of the nine candidate transformations. The first column gives the respective base transformation (though we consider the normalized version of this, as discussed in section 2.7). The first three columns correspond to the posterior probabilities using the default uninformative prior given at item (iv) in section 3.2, while the last three columns are for the informative prior discussed in section 4.2 that places more prior weight on the seven probit transformations. The posterior probabilities of the most likely transformations for each score and prior are highlighted in bold and do not change with respect to either prior.

Term	Probability	Coefficient
For Noted		
I	0.992	0.402
$(1 - I)$	0.992	0.442
For Associated		
I	1.000	0.743
$(1 - I)$	1.000	0.739
For Read-Most		
I	1.000	0.555
IG_1	0.522	0.094
$(1 - I)$	1.000	0.548

Table 3: The posterior mean estimates the regression coefficients, $E(\beta_i|\lambda = \hat{\lambda}_M, \text{data})$, are in the ‘Coefficient’ column. The ‘Probability’ column contains the posterior probabilities of these coefficients being non-zero, $\Pr(\gamma_i = 1|\lambda = \hat{\lambda}_M, \text{data})$. The results are for the terms involving the insert dummy variable, including the non-insert intercept $(1 - I)$. Only those coefficients that have a posterior probability greater than 0.35 of being non-zero are reported.

For Noted Score Regression							
Description	Term	Prob.	Coeft.	Description	Term	Prob.	Coeft.
RH position	X_3	0.777	0.033	cntnt is fin	X_{10}	0.724	0.02
hdln color	X_{20}	0.838	-0.015	copy area	X_{23}	0.428	-0.011
(cntnt is act) G_2	$X_{11}G_2$	0.398	0.012	(recipe) G_2	RG_2	1.000	0.107
(hdln below ill) G_3	$X_{19}G_3$	0.512	-0.036	(hdln color) G_3	$X_{20}G_3$	0.369	-0.02
(copy area) G_3	$X_{20}G_3$	0.422	0.029	(num words) G_3	$X_{24}G_3$	0.507	-0.042
pred feat person(s)	$X_1^{[1]}$	0.433	-0.01	pred feat animal(s)	$X_4^{[1]}$	0.881	0.05
pred feat product	$X_8^{[1]}$	0.521	0.014	hdln apl prize(s)	$X_6^{[2]}$	0.465	-0.023
pred apl security	$X_3^{[3]}$	0.785	-0.063	pred apl knowledge	$X_5^{[3]}$	0.571	-0.013
furnishings	T_6	0.431	-0.021	building	T_8	0.998	-0.211
travel	T_9	0.909	-0.096	motor vehicles	T_{12}	1.000	-0.157
cosmetics	T_{15}	0.813	0.045	jewellery	T_{17}	0.656	0.071

For Associated Score Regression							
Description	Term	Prob.	Coeft.	Description	Term	Prob.	Coeft.
hdln color	X_1	0.992	-0.018	advertorial	X_{30}	0.472	-0.023
coupon	X_{31}	0.506	-0.012	(cntnt is fin) G_1	$X_{10}G_1$	0.584	0.021
(hdln case) G_2	$X_{22}G_2$	0.922	-0.015	(recipe) G_2	RG_2	1.000	0.075
(opp pg same muli) G_2	$X_{29}G_2$	0.816	0.04	(small hdln) G_3	$X_{16}G_3$	0.692	-0.039
(opp pg unrel ed) G_3	$X_{28}G_3$	0.818	0.049	(advertorial) G_3	$X_{30}G_3$	0.815	0.104
(opp pg unrel ed) G_4	$X_{28}G_4$	0.456	-0.017	pred feat product	$X_8^{[1]}$	0.927	0.024
pred col blue	$X_2^{[4]}$	0.522	-0.01	pred col yel/org/brwn	$X_6^{[4]}$	0.582	0.009
pred apl income	$X_6^{[3]}$	0.57	0.014	building	T_8	0.459	-0.041
motor vehicles	T_{12}	0.53	-0.026	cosmetics	T_{15}	0.565	0.019
pets	T_{22}	0.777	0.047				

For Read-Most Score Regression							
Description	Term	Prob.	Coeft.	Description	Term	Prob.	Coeft.
hdln color	X_{20}	0.65	0.011	hdln type	X_{21}	0.467	-0.006
(num words) G_1	$X_{24}G_1$	0.972	-0.072	(square ill) G_2	X_5G_2	0.65	0.028
(hdln beside ill) G_2	$X_{18}G_2$	0.508	-0.019	(recipe) G_2	RG_2	1.000	0.104
(num words) G_2	$X_{24}G_2$	0.996	-0.095	(advertorial) G_2	$X_{30}G_2$	0.531	0.053
(coupon) G_2	$X_{31}G_2$	0.521	0.025	(photo) G_3	X_7G_3	0.467	0.038
(cntnt is end) G_3	X_9G_3	0.517	0.032	(hdln lnth) G_4	$X_{14}G_4$	0.956	-0.066
pred feat person(s)	$X_1^{[1]}$	0.993	-0.035	pred col blue	$X_2^{[4]}$	0.381	-0.009
pred apl exclamation	$X_4^{[2]}$	0.769	-0.022	pred apl news	$X_5^{[2]}$	0.684	-0.016
motor vehicles	T_{12}	1.000	-0.198	jewellery	T_{17}	0.633	0.071
government	T_{19}	0.923	0.184				

Table 4: This table provides the posterior means for the slope coefficients, $E(\beta_i|\lambda = \hat{\lambda}_M, \text{data})$, in the ‘Coeft.’ column. Their respective posterior probability of being non-zero, $\Pr(\gamma_i = 1|\lambda = \hat{\lambda}_M, \text{data})$, is provided in the ‘Prob.’ column. These are provided for all three regressions, but only for coefficients that had a posterior probability greater than 0.35 of being non-zero. Terms that involve the variables P, I, S, B and M are not included because they are dealt with elsewhere.

Regression	Procedure					
	Bayesian	OLS	Step & AIC	30-Factor	50-Factor	100-Factor
	(i) Correlation Between Predicted and Actual Values					
Noted	0.764	0.706	0.743	0.727	0.695	0.620
Associated	0.511	0.507	0.488	0.464	0.331	0.347
Read-Most	0.504	0.368	0.357	0.273	0.220	0.245
	(ii) Mean Absolute Error of Predicted Values					
Noted	0.075	0.086	0.092	0.109	0.102	0.100
Associated	0.070	0.074	0.151	0.111	0.114	0.086
Read-Most	0.095	0.115	0.120	0.146	0.145	0.121
	(iii) Mean Squared Error of Predicted Values					
Noted	0.009	0.012	0.011	0.019	0.017	0.015
Associated	0.008	0.009	0.043	0.027	0.023	0.011
Read-Most	0.015	0.021	0.023	0.036	0.032	0.022

Table 5: This table contains diagnostics for the predicted readership scores in the hold-out sample (March 1993 issue) using a variety of approaches. The six methods used are arrayed across the top of the table and are described in section 4.6, while the the results for all three readership scores are presented. The diagnostics include (i) the sample correlations between the actual readership scores and predicted scores, (ii) mean absolute prediction error and (iii) mean squared prediction error.