Shrinkage Estimation and Direct Marketing Scoring Models

Edward C. Malthouse[*]

July 17, 1998

[*] This paper is under review at *Journal of Interactive Marketing.* Thanks to Robert Kestnbaum, Pamela Ames, and the rest of the technical staff at Kestnbaum & Company for many helpful discussions and for use of their computing facilities.

Abstract

This paper shows how ridge regression and other shrinkage estimates can be used to improve the performance of direct marketing scoring models. It reviews the key property of shrinkage estimates — that they produce more stable estimates having smaller mean squared error than ordinary least squares models. It relates the idea of the effective number of parameters or degrees of freedom from the smoothing literature to ridge regression. The ridge estimates are shown to fit fewer degrees of freedom than the ostensible number of parameters in the model. This means that direct markets can include more variables in a scoring model without danger of overfitting the data. Reducing the degrees of freedom by shrinking the estimates is shown to be more stable than dropping variables from the model with, e.g., stepwise regression. These results are corroborated by comparing shrinkage estimates with stepwise regression on two data sets. Improved ways of drawing samples from a database for estimating and validating models are proposed.

keywords: scoring models, shrinkage estimation, ridge regression, principal components regression, weight decay

1 Introduction

One of the fundamental questions of direct marketing is deciding who should receive a particular offer. Whenever marketers have the ability to target an offer or communication to an individual — not a large group of people such as the mass market — they must judiciously decide who should receive the offer. If a company wants to have a good relationship with its customers, it should not make offers that are not relevant. This undermines the company's efforts to earn customers' loyalty, build trust, and strengthen the relationship. Also, there is often a high marginal cost for each communication, and sending offers to people who will not respond is unprofitable. For example, some catalog companies spend $5 to send a catalog to a single customer, making it very important that those who receive these catalogs are likely to respond, or at least are highly valued customers.

Scoring models can help make this decision. They use other information that a company has about a customer to predict whether or not the customer will be interested in the current offer. These models were originally developed by direct marketers, and an example from direct marketing should make the idea clear. Suppose that a catalog company must decide which of its customers should receive its spring fashion catalog. It has the entire purchase history of its customers, beginning with variables such as recency (how recently the customer has purchased), frequency (number of times the customer has placed an order during some fixed time period), and monetary value (amount of money spent during a fixed time period). The catalog company sent a spring fashion catalog last year, and recorded which customers made purchases from it and how much they spent. The catalog company could build a scoring model using data from the previous year's spring fashion catalog. It would develop a the model by using the purchase history it had prior to the mailing to predict who actually responded to the mailing. It would then apply the model to the current purchase history to estimate how likely each customer is to respond, and send the catalog to those who are most likely. The estimate of response likelihood or expected purchase amount is called a *score* and applying the model to the current purchase history is called *scoring the database*.

In general, let vector $\mathbf{x}(p \times 1)$ contain information that a company has on a single customer to whom it is considering sending an offer. In the catalog example, $\mathbf{x}$ would contain the previous purchase history. Sometimes scoring models are used to prospect for new customer; in this case the company would not have as much information about the prospect and $\mathbf{x}$ might contain only a set of overlayed demographics. The variables represented by $\mathbf{x}$ are called *predictor variables*. Let y be some measure of the customer's response. In the catalog example, y would usually be the amount the customer bought from the previous spring fashion catalog. In other cases, such as a theater company sending a mailing to sell subscriptions to its program, y is a dichotomous variable taking values "respond" (1) or "did not respond" (0). I shall refer to y as the *dependent variable*. A scoring model,

f, tells how the average of y is related to $\mathbf{x}$:

$$E(y|\mathbf{x}) = f(\mathbf{x}). \qquad (1)$$

For a fixed value of $\mathbf{x}$, the scoring model tells us the average of the response variable.

There has been a lot of research on modeling (1). The reason for this is that the financial gains from even small improvements can be great. Catalog companies routinely circulate millions of copies of their books. With a circulation of say two million, models yielding even a one-cent improvement in average order size are appreciated. The same is true in many other situations where direct marketing is used such as credit card offers, solicitations from internet service providers, and solicitations from charities.

Many different functional forms have been proposed for f (e.g., see Shepard (1995, chs. 12, 13, 16)). Perhaps the most commonly used functional form by practitioners is to model y as a linear function of $\mathbf{x}$, $f(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ (Banslaben, 1992). Hansotia and Wang (1997) use logistic regression. Peltier, Davis, and Schibrowsky (1996) give an example of how discriminant analysis can be used in direct marketing. It could also be a tree-based model fitted with, e.g., CHAID or CART (Magidson, 1988). Zahavi and Levin (1997) investigate neural networks. Bult (1993) investigates semi-parametric forms for f.

This paper investigates how scoring models can be improved by using an alternative method of estimation, *shrinkage estimation.* Shrinkage estimation has a long history in many fields, including engineering (e.g., Frank and Friedman, (1993)), statistics (e.g., Neter et al. (1996), economics (e.g., Vinod and Ullah, (1981), machine learning and artificial intelligence (called the *weight-decay method*, Rumelhart, et al. (1996, pp. 552–3)), and marketing (e.g., Mahajan, Jain, and Bergier, (1977); Saunders (1987, pp. 12–14); Sharma and James (1981)). But they are not well known in the direct marketing community. This paper summarizes the key properties of shrinkage estimates and shows how the performance of existing scoring models can be improved simply by reestimating them with a shrinkage method such as ridge regression (RR) and principal components regression (PCR). It also proposes a new interpretation of the RR estimates — they are shown to reduce the effective number of parameters in the model. The usual way to reduce the number of parameters is to drop variables from the model. When it is important do reduce the degrees of freedom (e.g., to prevent overfitting), this new interpretation implies that practitioners can reduce the effective number of parameters without having to drop seemingly important variables from a model. Shrinkage estimation is demonstrated on two direct marketing datasets.

2 Ridge regression

This section focuses on one particular type of shrinkage estimate, ridge regression (RR) for the linear regression model. Ridge estimates are also available for other models including logistic regression (Schaefer et al., 1984). A second approach to shrinkage estimation, principal components regression (PCR), is also discussed briefly in this section. The section defines RR and develops two key properties of its estimates. The first property, which is covered in most regression texts, is that ridge estimates are more reliable than ordinary least squares (OLS) estimates in that they have smaller mean squared error. This means that on average they will come closer to estimating the true model parameters than the OLS estimates. Because of this property, RR is often applied to problems where there is a large amount of multicollinearity among the predictor variables and the OLS estimates are unstable. The second property is that the effective number of parameters in a model estimated with RR is smaller than the number of variables in the model. This is important because it gives the modeler a way of reducing the risk of overfitting the data without dropping variables, the approach currently favored by many direct marketers. This is established by applying results from the smoothing literature to RR. For these two reasons, along with the fact that commercial software currently offer RR and PCR, these estimates could be attractive to direct marketing modelers.

Let $\mathbf{X}$ be an $n \times p$ matrix scaled so that the columns have mean zero and variance one. The n rows of $\mathbf{X}$ contain p measurements on a customer. Assume that

$p < n$ and that the rank of $\mathbf{X}$ is p. Let $\mathbf{y}$ be an $n \times 1$ vector also scaled to have mean zero and variance one, containing measurements of the dependent variable, usually demand or customer value for direct marketing models. The linear regression model quantifies the relationship between $\mathbf{X}$ and $\mathbf{y}$:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + e,$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters that must be estimated from the data and e is a noise term with mean zero and variance σ^2. The ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is the solution to the least-squares objective function

$$\mathbf{b} = \underset{\mathbf{b}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

2.1 Reducing mean squared error of b

The overall quality of an estimate is usually measured by its mean squared error (MSE)

$$\begin{aligned} \mathrm{MSE}(\mathbf{b}) &= E[(\mathbf{b} - \boldsymbol{\beta})'(\mathbf{b} - \boldsymbol{\beta})] \\ &= \mathrm{tr}[V(\mathbf{b})] + \mathrm{bias}(\mathbf{b})'\mathrm{bias}(\mathbf{b}). \end{aligned}$$

The MSE of $\mathbf{b}$ tells us, on average, how far the estimate $\mathbf{b}$ is from the true value of the parameter $\boldsymbol{\beta}$? It can be decomposed into a sum of two terms: the variance and the squared bias of the estimate. The OLS estimate $\mathbf{b}$ is an unbiased estimate of $\boldsymbol{\beta}$, and therefore MSE($\mathbf{b}$) is simply the variance of $\mathbf{b}$. The Gauss-Markov theorem tells us that $\mathbf{b}$ is the best linear unbiased estimate, meaning that it has the lowest variance and MSE among all linear *unbiased* estimates. But if we also consider biased estimates of $\boldsymbol{\beta}$, we can do better. This is exactly what RR and PCR do: they modify $\mathbf{b}$ by introducing a bias, which just so happens to decrease the variance by more than the squared bias, so that the RR/PCR estimates of $\boldsymbol{\beta}$ have lower MSE than $\mathbf{b}$.

RR biases the coefficients by shrinking them towards zero. This is the same idea as weight decay in fitting feedforward neural networks, where absolute magnitude of the weights in a neural networks are reduced slightly during each step of the estimation process. RR biases the coefficients by introducing a penalty term in the objective function that penalizes estimates of $\boldsymbol{\beta}$ that are large in absolute magnitude. The ordinary ridge estimate (ORE) with biasing constant $k \geq 0$ is

$$\begin{aligned} \mathbf{b}_k &= \underset{\mathbf{b}_k}{\operatorname{argmin}} \left[\|\mathbf{y} - \mathbf{X}\mathbf{b}_k\|^2 + nk\mathbf{b}_k'\mathbf{b}_k\right] \\ &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned}$$

When $k = 0$, the ORE is equal to the OLS estimate. The larger the value of k, the greater the penalty for having large coefficients. The ridge existence theorem shows that for $0 < k < 2\sigma^2/\boldsymbol{\beta}'\boldsymbol{\beta}$, MSE($\mathbf{b}_k$) < MSE($\mathbf{b}$) (Vinod and Ullah, 1981, §7.2.1). Thus for an appropriate choice of k, on average the ORE $\mathbf{b}_k$ will not equal $\boldsymbol{\beta}$ because it is biased, but on average it will be closer to $\boldsymbol{\beta}$ than $\mathbf{b}$. In this sense, the ORE is a more reliable estimate of $\boldsymbol{\beta}$ than the OLS estimate. By using a more reliable estimate, hopefully the performance of scoring models can be improved.

2.2 Effective number of parameters

Another very important property of the ORE is that it reduces the number of degrees of freedom used by the model. The OLS estimate has p parameters and therefore uses p degrees of freedom. The number of parameters in a model is very important to direct marketers. The main reason for keeping the degrees of freedom small is that a model with too many parameters in it is likely to *overfit* the data — it will begin to describe the stochastic variation in the data rather than the underlying relationship between $\mathbf{X}$ and $\mathbf{y}$. Overfitting can be detected with a validation sample. Often, direct marketers will split the available data into two groups: an estimation sample used to fit a model, and a validation sample use to evaluate how well the model performs. If the performance of the model, which is usually measured by a gains table, for the estimation and validation samples are approximately equal, the model does not overfit the data. If the performance on the estimation sample is much better than for the validation sample, the model overfits the data.

Hastie and Tibshirani (1990, §3.5) discuss the *ef-*

fective number of parameters or *degrees of freedom* for linear smoothers. A linear smoother can be written as

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y},$$

where the linear operator $\mathbf{S}(n \times n)$ is a *smoother* matrix and $\hat{\mathbf{y}}$ contains the predicted values of $\mathbf{y}$. The degrees of freedom of $\mathbf{S}$ are given by the trace of the smoother matrix, $\mathrm{tr}(\mathbf{S})$. The OLS and ORE are both linear smoothers, because the predicted values from either estimate are given by the product

$$\hat{\mathbf{y}} = \mathbf{H}_k\mathbf{y},$$

where $\mathbf{H}_k = \mathbf{X}\mathbf{b}_k$ is the "hat" matrix. Let $\mathbf{H} = \mathbf{H}_0$ be the hat matrix for the OLS estimate. It is easy to confirm that $\mathrm{tr}(\mathbf{H})$ gives the degrees of freedom for the OLS estimate. Recall that $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$ for $\mathbf{A}(n \times p)$ and $\mathbf{B}(p \times n)$. Then,

$$\mathrm{tr}(H) = \mathrm{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \mathrm{tr}(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) = p.$$

Therefore $\mathrm{tr}(\mathbf{H})$ gives the correct degrees of freedom for the OLS case.

To find the degrees of freedom for the ORE, let $\lambda_1 \geq \cdots \geq \lambda_p > 0$ be the eigenvalues of $\mathbf{X}'\mathbf{X}$. Let the SVD of $\mathbf{X}$ be $\mathbf{U}\Lambda^{1/2}\mathbf{V}'$, where $\mathbf{U}(n \times p)$ and $\mathbf{V}(p \times p)$ are column-orthogonal matrices.

$$\begin{aligned}
\mathrm{df}(k) &= \mathrm{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X} + kI)^{-1}\mathbf{X}') \\
&= \mathrm{tr}(\mathbf{U}\Lambda^{1/2}\mathbf{V}'(\mathbf{V}\Lambda\mathbf{V}' + k\mathbf{V}\mathbf{V}')^{-1}\mathbf{V}\Lambda^{1/2}\mathbf{U}') \\
&= \mathrm{tr}(\mathbf{U}\Lambda^{1/2}\mathbf{V}'(\mathbf{V}(\Lambda + k\mathbf{I})\mathbf{V}')^{-1}\mathbf{V}\Lambda^{1/2}\mathbf{U}') \\
&= \mathrm{tr}(\mathbf{U}\Lambda^{1/2}(\Lambda + k\mathbf{I})^{-1}\Lambda^{1/2}\mathbf{U}') \\
&= \mathrm{tr}(\Lambda(\Lambda + k\mathbf{I})^{-1}) \\
&= \sum_{j=1}^{p}\delta_j, \qquad (2)
\end{aligned}$$

where $\delta_j = \lambda_j/(\lambda_j + k)$ are called *shrinkage fractions.* When k is positive, the shrinkage fractions will all be strictly less than one, and $\mathrm{tr}(\mathbf{H}_k) < p$. Thus, the effective number of parameters in a RR model is less than the actual number of predictor variables.

This result could be very important to direct marketers because it implies that they can include more variables in a predictive model than they could with OLS without danger of overfitting. Direct marketers usually use a variable subset selection (VSS) procedure such as stepwise regression to reduce the number of predictors to a small number, say 10, and the final model is estimated with OLS. Suppose that $\mathbf{X}$ contains a modest number of relevant predictors, say 15 or 20, that is greater than the number of predictors after VSS. The ORE using the larger number of predictor variables and an appropriate k has the same number of effective parameters as the reduced OLS model. Because the ridge model uses information from a larger number of predictors, it is possible that the ORE will give better scoring model results. VSS and RR both reduce the effective number of parameters in a model. VSS does it explicitly by dropping variables from the model. RR does it by reducing the effect of any single variable, without dropping any entirely. Frank and Friedman (1993) also note this distinction. They give the results from a massive simulation study and show that RR outperforms VSS on chemical engineering applications. They use predictive mean squared error as the quantity to maximize in their simulation instead of lift at relevant quantiles, which is of greater interest to direct marketers. They also examine two other shrinkage estimates, PCR and partial least squares (PLS). Their are summarized "Although [OLS and VSS] are the oldest and most widely used techniques in the statistical community, the results presented here suggest that there might be much to be gained by considering one of the more modern methods (RR, PCR, or PLS) as well (p. 127)." They also find that the performance of RR dominates PCR and PLS, but not by much.

The shrinkage fractions δ_j come up elsewhere in the theory of RR. One of the other properties of δ_j will give us a deeper understanding of how RR shrinks the estimates and help us explain some of the experimental results in the next section. Let $\mathbf{T} = \mathbf{X}\mathbf{V}$ be the principal components of $\mathbf{X}$. We can equivalently regress $\mathbf{y}$ on $\mathbf{T}$:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = \mathbf{X}\mathbf{V}\mathbf{V}'\boldsymbol{\beta} + \mathbf{e} = \mathbf{T}\boldsymbol{\gamma} + \mathbf{e}, \qquad (3)$$

where $\boldsymbol{\gamma} = \mathbf{V}'\boldsymbol{\beta}$ are called the principal components regression (PCR) parameters. Let $\hat{\boldsymbol{\gamma}} = \Lambda^{-1/2}\mathbf{U}'\mathbf{y}$ be the OLS estimates of $\boldsymbol{\gamma}$. Now use RR to shrink the OLS estimates of $\boldsymbol{\gamma}$ (Vinod and Ullah, 1981, §7.1.2):

$$\begin{aligned} \hat{\gamma}_k &= (\mathbf{T}'\mathbf{T} + k\mathbf{I})^{-1}\mathbf{T}'\mathbf{y} \qquad (4) \\ &= (\Lambda + k\mathbf{I})^{-1}\boldsymbol{\Lambda}^{1/2}\mathbf{U}'\mathbf{y} \\ &= \boldsymbol{\Delta}\hat{\gamma}, \end{aligned}$$

where $\boldsymbol{\Delta} = \mathrm{diag}(\delta_j)$. Thus, the magnitude of the j^{th} estimated PCR parameter is shrunken by the fraction $\delta_j \leq 1$. Because $\delta_1 \geq \cdots \geq \delta_p$, the slope estimates of principal components corresponding to large eigenvalues are shrunken less than those corresponding to small eigenvalues. This may not be the optimal approach, because the principal components corresponding to large eigenvalues are not necessarily the best predictors of $\mathbf{y}$.

2.3 Finding biasing constant k

As indicated in § 2.1, the ridge existence theorem guarantees the existence of a k so that the ORE of $\boldsymbol{\beta}$ has smaller mean squared error than the OLS estimate. The catch is that the value of k depends on $\boldsymbol{\beta}$ and σ^2, which are the parameters being estimated. This is one of the criticisms of RR (Draper and Smith, 1998, p. 397). Most regression texts, which are usually concerned with finding more reliable estimates of $\boldsymbol{\beta}$ in the presence of multicollinearity, suggest computing OREs for a variety of k values. Next, they suggest constructing a ridge trace, which plots the values of the individual slope estimates against k, to assist in picking a value of k. I suggest using a similar strategy to find a k that produces a better scoring model. Estimate the model for a variety of k values. Instead of constructing a ridge trace, compare the gains tables resulting from evaluating the models on a validation sample. Choose the value of k giving the best lift. A figure similar to a ridge trace can be constructed. Figure 3 gives an example, and is discussed further in the next section.

One problem with this approach is that the values of k have no interpretation. It is also difficult to compare ridge models with VSS and PCR models. Equation (2) showed that every value of k is associated with an effective number of model parameters (degrees of freedom), which is a more interpretable quantity and allows the analyst to compare OREs to VSS and PCR estimates. Instead of picking a sequence of k values, we could also pick a sequence of df values and find their corresponding k values. To do so, we must address the question of finding a value of k to give a specified number of degrees of freedom, i.e., find k so that $\mathrm{df}(k) = c$ where $0 < c < p$. This is equivalent to finding the root $k > 0$ of function $g(k) = \mathrm{df}(k) - c$. I suggest finding this root numerically using the bisection method, which is guaranteed to converge to the root if we can find an interval $[k_l, k_u]$ such that $g(k_l)g(k_u) < 0$ and $g(k)$ is monotonic on $[k_l, k_u]$ (1988, §4.2.4). Pick $k_l = 0$ and $k_u = \lambda_1(p - c)/c$. Function $\mathrm{df}(k)$ and thus $g(k)$ decrease monotonically for $k \geq 0$ because $\mathrm{df}(k)$ is the sum of the δ_i, each of which decreases as k increases in this interval. Note that $g(0) = p - c > 0$ and

$$g(k_u) = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + k_u} - c \leq \sum_{i=1}^{p} \frac{\lambda_1}{\lambda_1 + k_u} - c = c - c = 0.$$

Therefore, there is a unique shrinkage constant k in the interval $[0, \lambda_1(p - c)/c]$ such that $\mathrm{df}(k) = c$.

The relationship between k and the degrees of freedom is illustrated with an example. Figure 2.3 shows the relationship for one of the data sets discussed in the next section. There are 34 variables in the model. If $k = 0$, then the degrees of freedom are also 34. As k increases, the degrees of freedom decrease, quickly at first, and then more slowly. The value 0 is an asymptote of the function.

2.4 Principal Components Regression

The PCR model was given in equation (3). It changes the basis of the column space of $\mathbf{X}$ to the principal component basis $\mathbf{T}$, which has orthogonal columns. Be reparameterizing the model in terms of the principal components, we could better understand how RR does its shrinking. Equation (4) showed that RR shrinks PCR parameter $\hat{\gamma}_j$ with the shrinkage fraction δ_j. An alternative approach to shrinkage is to drop columns of $\mathbf{T}$ from the regression, i.e., perform VSS on the principal components.

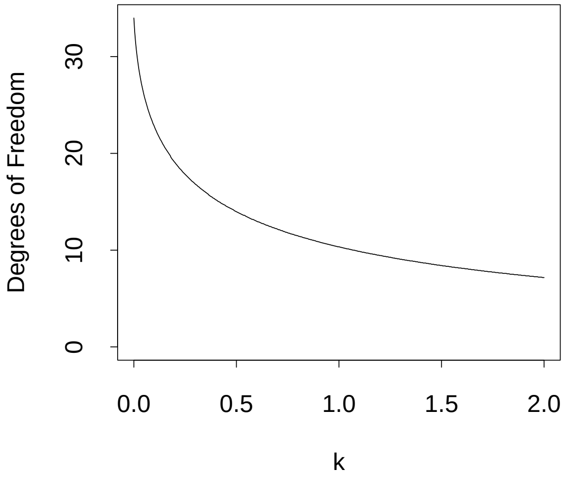

Figure 1: Plot of the effective number of parameter (degrees of freedom) against the ridge parameter k for a 34-variable model of DMEF2

3 Empirical Evaluation

The properties developed in the preceding section suggest that RR and PCR can provide better estimates of a scoring model. This section compares the performances of RR and PCR with VSS on some direct marketing data. The methods are compared using two of the Direct Marketing Educational Foundation's (DMEF) data sets. The DMEF makes these data sets available to educators for a nominal fee[1]. I shall use DMEF2 and DMEF4 in the comparison. DMEF2 contains data from a business with multiple divisions, each mailing different catalogs to a unified customer base. The DMEF4 data set contains data from an up-scale gift business that mails general and specialized catalogs to its customer base several times each year. Each data set contains approximately 100,000 observations. DMEF2 has 163 fields and DMEF4 has 93 fields. The data sets were used to fit scoring models that led to actual mailings. The `targdol` variable in each data set is a dependent variable and gives the number of dollars spent by each customer during a fixed time period. The distributions are summarized by the boxplots in table 3.2.

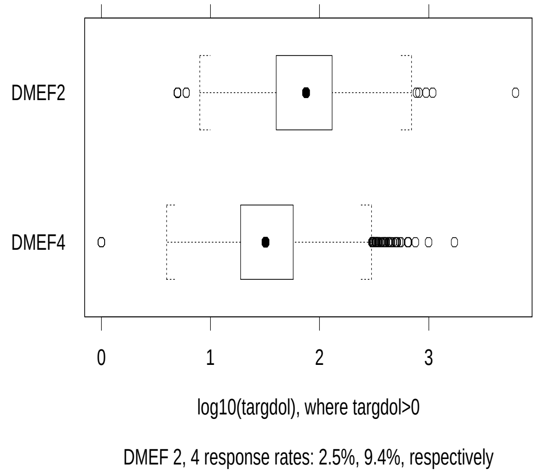

Figure 2: Boxplots for $\log_{10}$(`targdol`) in the DMEF data sets, responders only

The organization of this section follows the steps used to compare RR with VSS.

1. *Sampling.* The observations in both data sets were divided into three groups, an *estimation sample*, which was used to fit scoring models, a *validation sample*, was used to select a final model, and a *holdout* sample, which was used to evaluate the final model.

2. *Selecting a dependent variable.* I performed transformations on the `targdol` variable both data sets to remove outliers and make the distributions less pathological.

[1] Contact Laurie Spar at `lspar@the-dma.org`.

3. *Creating a set of candidate predictors.* I created a set of variables that could be used to predict the dependent variable. This set included the raw predictor variables in each data set, transformed versions of the raw predictors (e.g., $\log(x)$, $\sqrt{x}$, x^2, x^3), and combinations (interactions) of the raw and transformed predictors (e.g., dollars per buy, or log(dollars per buy)).

4. *Fitting models.* I used stepwise regression to reduce the number of predictor variables to p_{step}. Using the estimation data only, I fitted OLS models with $8, 9, \ldots, p_{\text{step}}$ predictor variables. I computed OREs using all p_{step} variables and a variety of k values.

5. *Comparing models.* I selected one stepwise model and one RR model by applying each model to the validation sample and comparing the gains tables. I then applied the two selected models to the holdout sample.

3.1 Sampling

The first step was to divide the available observations into three groups, an estimation sample used to fit the scoring models, a validation sample used to select a model, and a holdout sample used to evaluate the performance of the models. It is necessary to have separate validation and holdout samples rather than a single sample used to both choose the model and evaluate its performance. The data used to evaluate a model should be independent of data used to fit or select a model. One acceptable procedure would be to have two samples, fit and select a model using one, and evaluate the model using the other. The problem with this is that we may select a model that overfits the data. The way to avoid this problem is to have three samples, which is more similar to what would actually happen in practice where the analyst commonly splits the data into estimation and validation samples. The holdout sample plays the role of the actual mailing. Although the three samples are large, there is still considerable variation between the samples, as we shall see when we compare the models later in this section.

Each data set had a number of customers with extreme values of `targdol` (see figure 3). To ensure that the estimation, validation, and holdout samples were similar, I used the following stratified sampling procedure. I first sorted a dataset by the `targdol` variable. Next, I created $N/3$ strata of size three, where N is the number of observations in the data set. The first three customers in the sorted dataset constituted the first stratum, the fourth, fifth, and sixth customers the second stratum, the seventh, eighth, and ninth customers the third stratum, etc. There are six ways of assigning three observations to three different groups, i.e., estimation, validation, holdout; estimation, holdout, validation, etc. I drew a pseudorandom integer between 1 and 6 and assigned the observations based on this random integer.

By using this procedure the accuracy of the estimated average response is greatly improved. Consider the DMEF4 dataset, which has $N = 101,532$ observations. The mean of `targdol` for all customers is \$4.50 and the standard deviation is \$21.45. Suppose we wish to estimate the average of `targdol` using a sample of size $n = 33,844$. If we drew a simple random sample the standard deviation of the sample mean is (Cochran, 1977, Theorem 2.2)

$$\frac{S}{\sqrt{n}}\sqrt{(1-f)} = \frac{\$21.45}{\sqrt{33,844}}\sqrt{1-1/3} = 9.52 \times 10^{-2},$$

where S is the sample standard deviation of the population, n is the sample size, and f is the sampling fraction. Using the stratified sampling procedure outlined above, the standard deviation of the sample mean is (Cochran, 1977, Theorem 5.3)

$$\frac{\sqrt{253,613}}{33,844}\sqrt{(1-1/3)} = 1.21 \times 10^{-2},$$

where 253,613 is the sum of within-strata sample variances. Thus the accuracy is improved by nearly a factor of 8 by using this sampling procedure.

The accuracy can be improved further by Winsorizing the data before drawing the sample. In § 3.2 I suggest using a .1% Winsorized version `targdol`, i.e., replacing any value greater than the 99.9th percentile of \$236.95 (see Table 3.2) with the value

\$236.95. By doing this, the standard deviation becomes $S = \$18.92$ and the sum of within-strata sample variances becomes 29.34. The standard deviations of the means from SRS and stratified samples become 1.21×10^{-2} and 1.31×10^{-4}, respectively. The accuracy is improved by nearly a factor of 100.

A slightly simpler procedure that is closely related to the proposed stratified procedure would be to draw a systematic sample from the sorted database, e.g., assign all observations with $i \bmod 3 = 0$ to the estimation sample, all observations with $i \bmod 3 = 1$ to the validation sample, and the rest to the holdout sample, where i is the observation number. A problem with this approach is that the mean of observations with $i \bmod 3 = 2$ is guaranteed to be greater than (or equal to) the mean of the observations with $i \bmod 3 = 1$, which is greater than (or equal to) the mean of the observations with $i \bmod 3 = 0$, assuming the dataset is sorted in ascending order. For only marginally greater computational expense, one can draw the stratified sample as outlined above.

3.2 Selecting a dependent variable

The dependent variables in both data sets are nonnegative dollar amounts. In each case there is a large number of extreme values and the distributions are highly skewed. These outlying cases can be very problematic:

> Frequently in regression analysis applications, the data set contains some cases that are outlying or extreme ... These outlying cases may involve large residuals and often have dramatic effects on the fitted least squares regression function. It is therefore important to study the outlying cases carefully and decide whether they should be retained or eliminated, and if retained, whether their influence should be reduced in the fitting process and/or the regression model should be revised (Neter et al., 1996, p. 368).

One reasonable approach to reduce the influence of the extreme observations is to Winsorize[2] the extreme values and model the logarithm of the Winsorized `targdol` values. Table 3.2 gives the values of the 99.9th percentiles and and maximum values. Let `targdolw` be the .1% Winsorized version of the `targdol` variable. I model log(`targdolw`). Taking the log will further reduce the effect of the extreme values by bringing in the long tail.

Table 1: Values of the 99.9th and maximum value of `targdol` in the DMEF datasets

Data set	99.9 Percentile	Maximum
DMEF2	\$292.25	\$6248.75
DMEF4	\$236.95	\$1720.00

3.3 Creating a set of candidate predictors

The predictor variables provided in the DMEF datasets are most likely not the best predictors available for fitting a linear regression for several reasons. First, figure 3 showed that the distribution of `targdol` is problematic because it is highly skewed with many outliers. The same is true for many of the predictor variables. Just as extreme dependent-variable values can have a great influence on the estimated regression model, so can extreme predictor-variable values. Second, nonlinear combinations of the predictor variables are often better predictors than the original variables. Third, the relationships between predictor variables and log(`targdolw`) may not be linear. When this is the case, the fit can often be improved by using a transformed version of the predictor variable. Therefore I created additional candidate predictors in three steps:

[2] The $\alpha < .5$ Winsorized version of a set of number $x_1, \ldots, x_n$ is defined as follows. Let $x_{[\alpha]}$ be the value of the α quantile of the distribution. Let $w_i = x_i$ if $x_{[\alpha]} \leq x_i \leq x_{[1-\alpha]}$, $w_i = x_{[\alpha]}$ if $x_i < x_{[\alpha]}$, and $w_i = x_{[1-\alpha]}$ if $w_i > x_{[1-\alpha]}$. See, e.g., Dixon and Massey (1969, § 16-4).

1. I looked as the distributions of all the original variables. For variables with outliers and/or highly skewed distributions, I created an additional variable by Winsorizing the variable, taking logarithms, or both. This will reduce the effect of extreme observations on the estimation process.

2. I used judgment to create combination variables such as dollars per order, fraction of orders from division D, and number of divisions purchased from.

3. I created smooths of log(`targdolw`) against every original variable and transformed variable from steps 1 and 2. If the relationship appeared to be curvilinear, I created additional variables with log, square-root, square, and cube transformations. I also created dummy variables for categorical variables.

In total, I created 135 new candidate predictors for DMEF2 and 124 for DMEF4.

3.4 Comparing models

VSS and ridge models are compared based on the cumulative average order size at decile d (CAOS$_d$), where d is 1, 2, 3, or 4, on the validation and holdout samples. Cumulative average order size is defined as follows. Let $\hat{y}_i$ be the estimated dependent variable for individual $i = 1, \ldots n$ in the validation/holdout data, i.e., the estimate of log(`targdolw`). Order $\hat{y}_1, \ldots, \hat{y}_n$ from largest to smallest giving $\hat{y}_{(1)} \geq \ldots \hat{y}_{(n)}$. Let w_i be the Winsorized number of dollars spent by individual i, i.e., the value of `targdolw`, and let $w_{(i)}$ correspond to $\hat{y}_{(i)}$. Then the cumulative average order size at decile d is

$$\text{CAOS}_d = \frac{1}{\text{int}(nd/10)} \sum_{i=1}^{\text{int}(nd/10)} w_{(i)},$$

where int$(\cdot)$ is the greatest integer function. Note that CAOS$_d$ is not defined in terms of the deciles of $\hat{y}_1, \ldots, \hat{y}_n$. The problem with using deciles is that there can be many ties among the $\hat{y}_i$, particularly when there are very few predictor variables in the model. In these cases, decile d may not contain $10d\%$ of the observations, which could affect the comparisons of models.

To compare methods on DMEF2, I did the following. First, I applied stepwise regression to all original, combination, and transformed variables. In total 34 variables entered the regression, although five of them were also dropped. The final model had $34 - 5 = 29$ variables in it. Second, I scored the validation and holdout samples on the 8, 9, ..., 29 variable models from stepwise regression. The idea is to choose a model based on the validation sample and then evaluate it on the holdout sample. Third, I used all 34 variables to compute OREs giving models with $8, 9, \ldots, 34$ degrees of freedom. I scored the validation and holdout samples for the resulting models. I did the same for PCR.

CAOS is plotted against the degrees of freedom for the first four deciles in Figure 3. The following conclusions can be drawn.

1. RR dominates VSS and PCR. The RR curves are always higher than the corresponding stepwise curves, with a few minor exceptions for the 27- and 28-variable models at decile 4. The 10-variable PCR model also does very well.

2. The maximum CAOS values for RR are substantially higher than those for stepwise regression. For example, consider selecting a model to mail to a depth of decile 3. Using the validation sample to choose a model, we might choose the stepwise solution with 12 variables and the ORE with 14 degrees of freedom. The improvement on the holdout sample is $\$5.67 - \$5.52 = \$0.15$, which is a $.15/5.52 = 2.7\%$ improvement.

3. The VSS and PCR models are unstable, in that the CAOS values for models with different numbers of variables vary greatly. The CAOS values for the OREs follow a smooth curve. Breiman (1994) also notes that stepwise regression is very unstable.

4. The degrees of freedom giving a maximum value of CAOS for the ridge models on the validation sample are approximately the same as those for

the holdout sample. Thus, if we select a model on the validation sample, we feel confident that it is also a good model for the holdout sample. This is not always the case for the stepwise models. For example, if we were picking a model for decile 2, we would choose the 12-variable model based on the validation sample. CAOS on the holdout sample, however, improves by adding more variables. This makes it very difficult to choose a stepwise regression model.

I repeated the procedure with different seeds for the random number generator used to draw the stratified sample (which gave different sets of VSS predictors), and the conclusions were similar. RR gave smoother curves peeks around 10-15 degrees of freedom, depending on the decile. The curves for stepwise regression varied wildly.

The results from DMEF4 were similar. A total of 63 variables entered a stepwise regression using the original, combination, and transformed variables as candidate predictors. The marginal improvement after the 20th variable was very small, so I shall consider only the 20 variable model here. Other models gave similar conclusions. To arrive at the 20-variable model, 21 variables entered the regression, but the fifth variable to enter the model was dropped at step 14. I used all 21 variables to fit the ridge and PCR models. The results are summarized in figure 4.

1. The best ridge models perform better than the best VSS and PCR models. Consider selecting a model to mail to a depth of decile 3. One possibility is to choose the stepwise and ridge models with 20 degrees of freedom. This improves CAOS by \$24.16 − \$23.92 = \$0.24 or $0.24/23.92 \approx 1\%$.

2. The stepwise regression models do not improve much after the eighth variable entered the model.

3. The ORE perform worse than the stepwise estimates for smaller values of df. This was not the case with the DMEF2 models. One explanation for this is that the principal components corresponding to smaller eigenvalues are more highly correlated with $\mathbf{y}$ in this data set than for DMEF2. Equation (4) showed that the PCR parameter estimates of components corresponding to small eigenvalues are shrunken more than those corresponding to large eigenvalues. RR shrinks the wrong directions in this problem. The stepwise estimates do well here because the important components are strongly represented in all models. This is confirmed by the PCR models. PCR also does better than RR for a small number of degrees of freedom, because the components that best predict the dependent variable are in the model.

The comparisons were repeated with different stratified samples and sets of predictor variables. The conclusions were similar.

4 Discussion

The process of building a scoring model that doesn't overfit the data is comparable to getting a symphony orchestra to play with less volume. To reduce volume, the conductor can either remove players or have have all players continue playing, but more softly. Although both achieve the goal, there is a great difference in sound between an octet and a full orchestra playing softly, and both methods are appropriate in certain situations. This paper argues that modelers faced with the task of fitting scoring models that do not overfit the data have a similar two choices. Currently, modelers consider reducing the number of variables to be the only option. A second option is to use RR. A section of violins playing the same notes softly sounds different than a single violin playing the same music. Similarly, shrinkage estimates from highly correlated predictor variables will describe the variation in a dependent variable differently than an estimate using only one of the predictors.

This paper provides empirical evidence that models with many predictors estimated with RR have several advantages over VSS models. They can be substantially more profitable by increasing the lift at key deciles. The RR models seem to be more stable than VSS models in two different ways. The ridge model giving the best lift on the validation sample was also the model giving the best lift on the holdout sample; this is not always the case with the VSS models. The lift for the RR models was more consistent across models with different degrees of freedom than VSS models; the RR curves in figures 3 and 4 are very smooth whereas the points representing the VSS fits varied widely. The stability of the RR models makes selecting a final model much easier.

References

Banslaben, J. (1992). Predictive modelling. In Nash, E., editor, *The Direct Marketing Handbook*. McGraw Hill, New York.

Breiman, L. and Friedman, J. (1994). Predicting multivariate responses in multiple linear regression. Technical Report Technical Report Number 111, Laboratory for Computational Statistics, Department of Statistics, Stanford University. To be published in J. Roy Stat. Soc., Ser. B.

Bult, J. R. (1993). Semiparametric versus parametric classification models: An application to direct marketing. *Marketing Science*, XXX:380 – 90.

Cochran, W. G. (1977). *Sampling Techniques*. Wiley.

Dixon, W. J. and Massey, F. J. (1969). *Introduction to statistical analysis*. McGraw-Hill, third edition.

Draper, N. and Smith, H. (1998). *Applied Regression Analyisis*. Wiley, third edition.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–148. with discussion.

Hansotia, B. J. and Wang, P. (1997). Analytical challenges in customer acquisition. *Journal of Direct Marketing*, 11(2):7–19.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Magidson, J. (1988). Improved statistical techniques for response modeling. *Journal of Direct Marketing*, 2(4):6–18.

Mahajan, V., Jain, A., and Bergier, M. (1977). Parameter estimation in marketing models in the presense of multicollinearity. *Journal of Marketing Research*, 14(4):586–91.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistical Models*. Irwin, fourth edition.

Peltier, J. W., Schibrowsky, J. A., and Davis, J. (1996). Predicting payment and nonpayment of direct mail obligations: profiling good and bad credit risks. *Journal of Direct Marketing*, 10(2):36–43.

Rumelhart, D. E., Durbin, R., Golden, R., and Chauvin, Y. (1996). Backpropagation: the basic theory. In Smolensky, P., Mozer, M. C., and Rumelhart, D. E., editors, *Mathematical perspectives on neural networks*, pages 552–3. Lawrence Erlbaum Associates Publishers.

Saunders, J. (1987). The specification of aggregate market models. *European Journal of Marketing*, 21(2):5–47.

Schaefer, R., Roi, L., and Wolfe, R. (1984). A ridge logistic estimator. *Commun. Statistics-Theory Methods*, 13(1):99–113.

Sharma, S. and James, W. L. (1981). Latent root regression: an alternative procedure for estimating parameters in the presence of multicollinearity. *Journal of Marketing Research*, 18:154–161.

Shepard, D. (1995). *The New Direct Marketing*. Irwin, second edition.

Thisted, R. (1988). *Elements of Statistical Computing*. Chapman and Hall, New York.

Vinod, H. D. and Ullah, A. (1981). *Recent Advances in Regression Methods*. Marcel Dekker, Inc., New York.

Zahavi, J. and Levin, N. (1997). Applying neural computing to target marketing. *Journal of Direct Marketing*, 11(1):5–22.

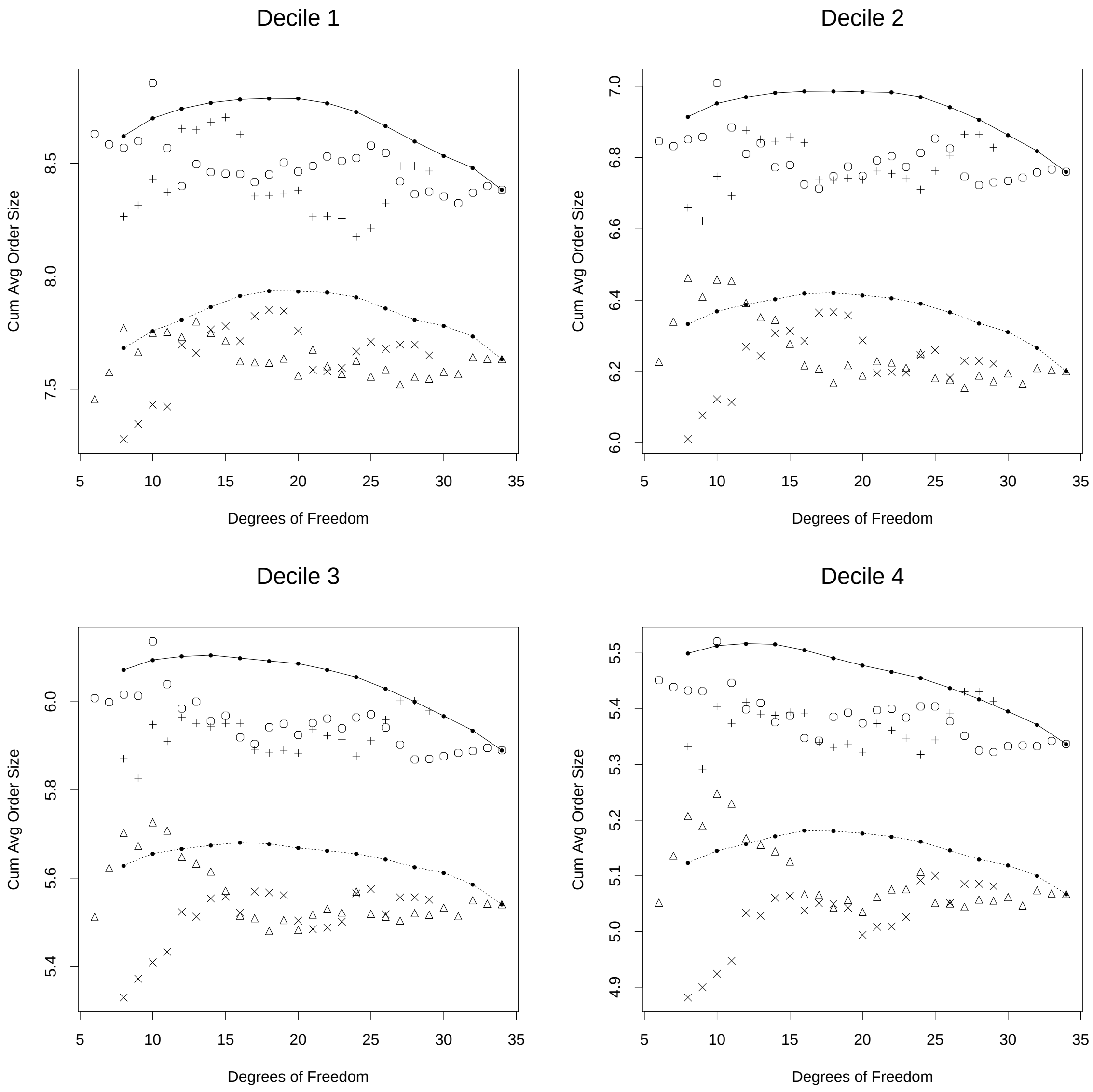

Figure 3: Lift traces and for a 34-variable model of DMEF2. The following plotting symbols are used for the *validation* sample: RR (solid line), PCR (∘), VSS (+). The symbols for the *holdout* sample are: RR (dotted line), PCR (△), VSS (×).

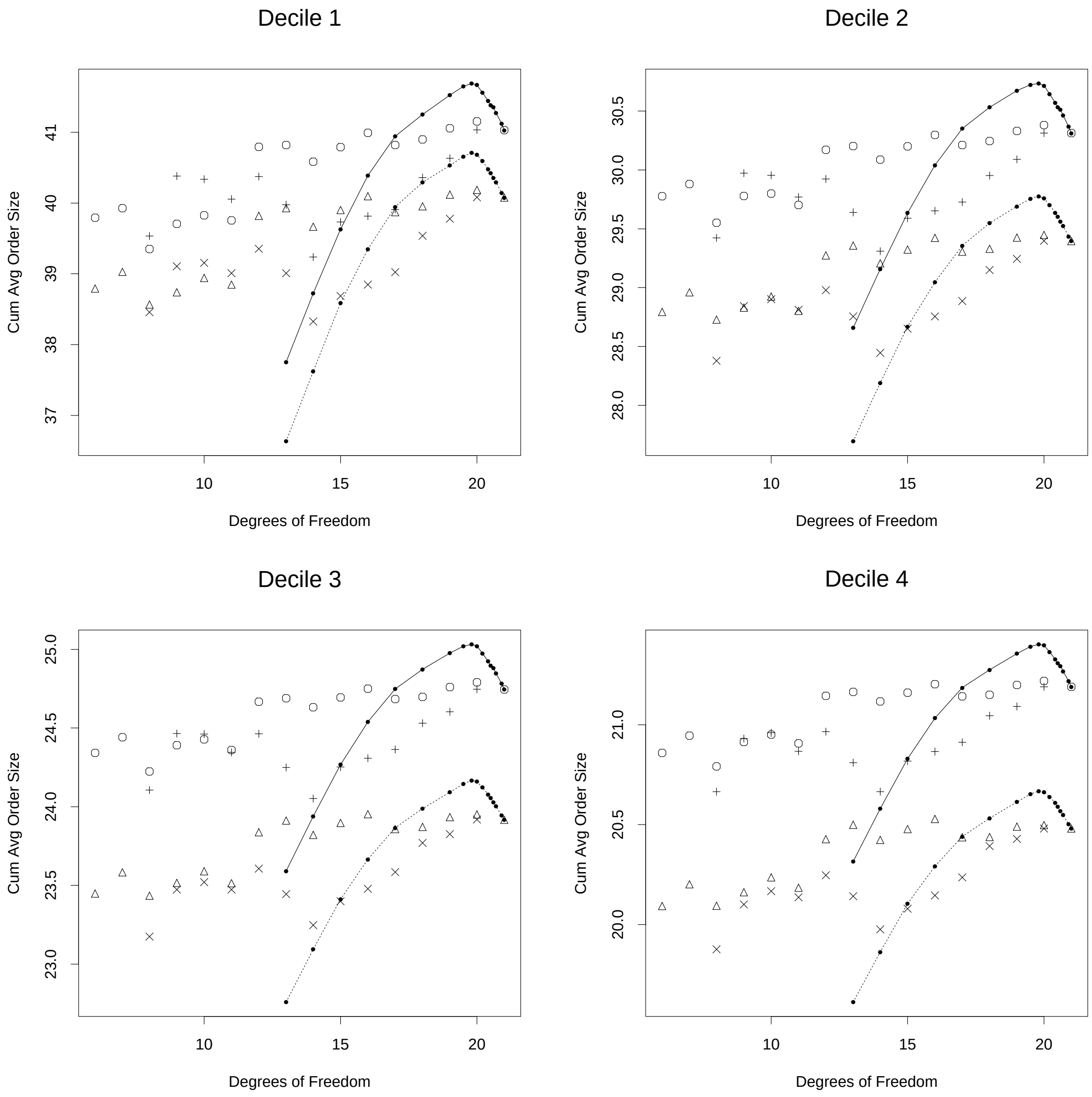

Figure 4: Lift traces and for a 21-variable model of DMEF2. The following plotting symbols are used for the *validation* sample: RR (solid line), PCR (∘), VSS (+). The symbols for the *holdout* sample are: RR (dotted line), PCR (△), VSS (×).