Journal of Modern Applied Statistical Methods

Volume 5 | Issue 2 Article 3

11-1-2005

Inferences About the Components of a Generalized Additive Model

Rand R. Wilcox
University of Southern California, rwilcox@usc.edu

Journal of Modern Applied Statistical Methods
November, 2006, Vol. 5, No. 2, 309-316

1538 – 9472/06/$95.00

Inferences About the Components of a Generalized Additive Model

Rand R. Wilcox
Department of Psychology
University of Southern California

A method for making inferences about the components of a generalized additive model is described. It is found that a variation of the method, based on means, performs well in simulations. Unlike many other inferential methods, switching from a mean to a 20% trimmed mean was found to offer little or no advantage in terms of both power and controlling the probability of a Type I error.

Key words: Nonparametric regression, smoothers, backfitting algorithm, wild bootstrap

Introduction

When dealing with a regression problem, a standard approach is to assume

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, \qquad (1)$$

i=1,...,n, where ε_i is independent of $X_{i1},...,X_{ip}$, $E(\varepsilon)$=0, and then test hypotheses about the unknown parameters $\beta_o,...,\beta_p$. This approach seems appropriate when the assumed model, given by (1), is a reasonable approximation of the true regression surface. But experience with smoothers suggests that, at least in some situations, the assumption that Y is linearly related to the p regressors is unsatisfactory, and often it is unclear how to correct this problem when using a parametric approach to modeling the data, particularly when p>2. That is, simple transformations of the regressors might be used, such as taking logarithms, but situations arise where effective transformations are not evident and difficult to discern.

Rand R. Wilcox (rwilcox@usc.edu) is Professor of Psychology at the University of Southern California.

One problem is that smooths often suggest that over some region of the predictor space, the regression surface is, approximately, a horizontal plane, meaning there is virtually no association at all, but for other regions a curvilinear association appears that can be difficult to model.

Figure 1 provides an example where the goal is to predict reading ability (measured by a word identification score) based on two measures of phonological awareness. Shown is a smooth using the loess method derived by Cleveland and Devlin (1988). Note that for low measures associated with both predictors, the regression surface is nearly flat, but in other regions there appears to be a nonlinear association. (Switching to a robust smooth, namely the running interval smoother in Wilcox, 2003, with the span set to 1.2, results in a plot nearly identical to Figure 1.)

A more flexible approach, when modeling the data, is to use a generalized additive model (Hastie & Tibshirani, 1993). That is, assume that there exists functions $f_1,...,f_p$ such that

$$Y = \beta_0 + f_1(X_1) + \cdots + f_p(X_p) + \varepsilon. \qquad (2)$$

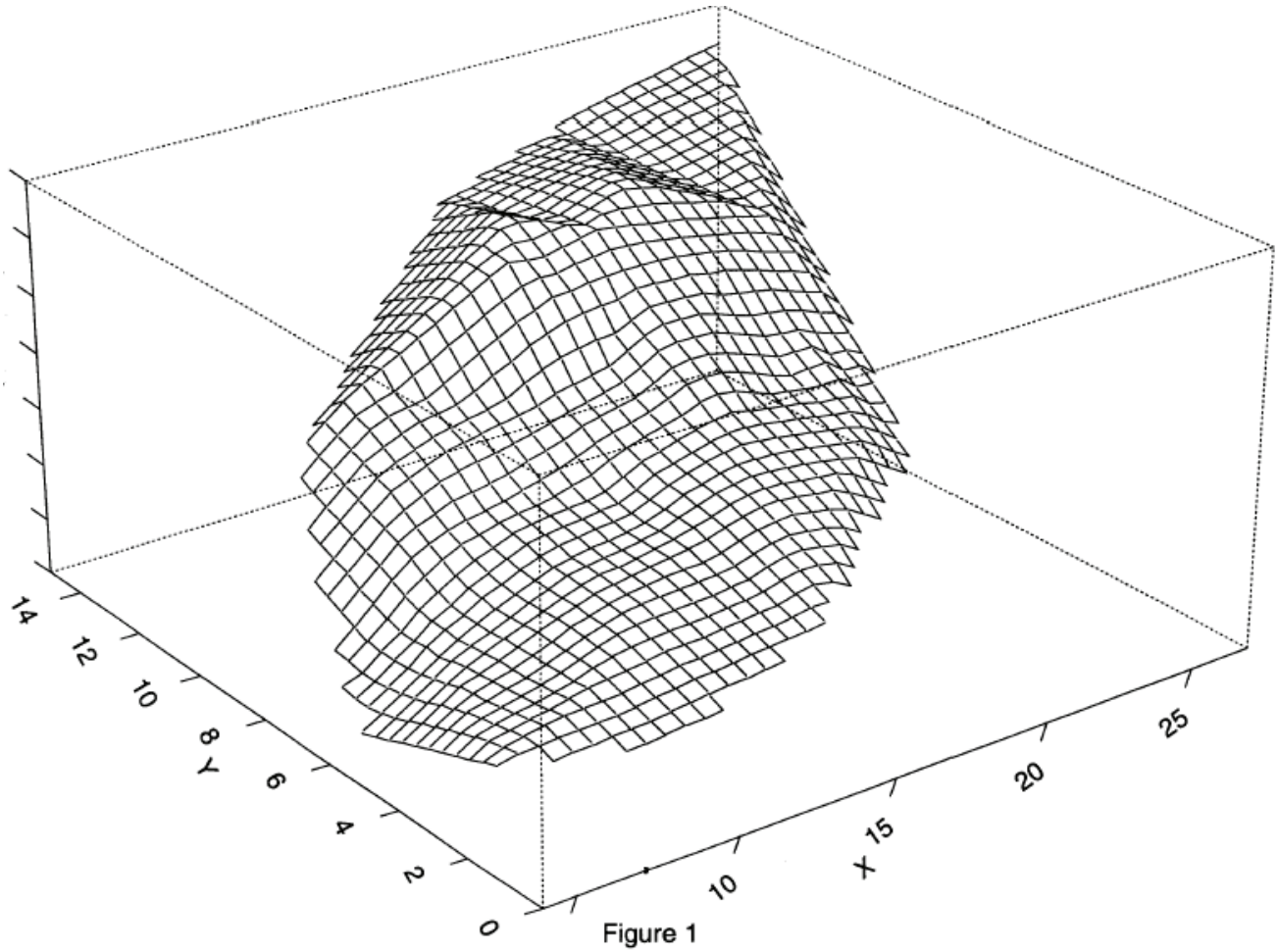

Figure 1

Of course, equation (2) contains the usual model, given by (1), as a special case. For any fixed *j*, the goal in this paper is to consider the problem of testing

$$H_0: f_j(X_j)=0. \qquad (3)$$

The general strategy used here is to fit a generalized additive model omitting the *j*th variable and then check what is essentially a simple extension of the partial residual plot (e.g. Berk & Booth, 1995) for an association. (The approach used here for detecting an association is closely connected to what Berk and Booth call the AMALL method for detecting curvature, which stems from Breiman and Friedman, 1985, p. 618). That is, test the hypothesis that the regression line between the resulting residuals and X_j is straight and horizontal; this is done with the wild bootstrap method derived by Stute, González-Manteiga and Presedo-Quindimil (1998). Details of the proposed method are given in the next two sections.

A Generalized Additive Fit

There are many ways of fitting the model given by (2) with most methods assuming that the goal is to estimate the mean of *Y* given $\left(X_{i1},...,X_{ip}\right)$. The method used here was chosen because it represents a particularly simple way of including virtually any robust measure of location. Robust measures of location are known to have many advantages, versus the mean, for a wide range of situations

(e.g., Hampel, Ronchetti, Rousseeuw and Stahel, 1986; Huber, 1981; Staudte & Sheather, 1990; Wilcox, 2005).

Two fundamental advantages are improved control over the probability of a Type I error in situations where methods based on means perform poorly, and substantial gains in power, even under small departures from normality. Here, however, when using means, good control over the probability of a Type I error is obtained in simulations and using means actually offers higher power. So the method used here provides an interesting example of a situation where a non-robust estimator performs better than a robust estimator in terms of power, even when sampling from a heavy-tailed distribution.

The robust location estimator used here is the 20% trimmed mean which is computed as follows. Let $X_1,...X_m$ be any m values and let $X_{(1)} \le ... \le X_{(m)}$ be the values written in ascending order. Let $g = [.2m]$, where $[x]$ is the greatest integer less than or equal to x. Then the 20% trimmed mean is

$$\frac{1}{m-2g}\sum_{i=g+1}^{m-g} X_{(i)} \quad .$$

The reason for choosing 20% trimming, over alternative amounts of trimming, stems from results on efficiency reported by Rosenberger and Gasko (1983).

As explained in Hampel, Ronchetti, Rousseeuw and Stahel (1986), Huber (1981), and Staudte and Sheather (1990), a reasonable alternative to the 20% trimmed mean is some robust M-estimator. The only reason for choosing a 20% trimmed over the better-known robust M-estimators is to avoid division by zero in certain situations to be described.

First consider the one-predictor case (p=1). There are many ways of estimating f_1 using so-called smoothers (e.g., Hastie & Tibshirani, 1990; Härdle, 1990). Here, a running interval smoother is used mainly because it is readily extended to robust measures of location such as the 20% trimmed mean. This is not to suggest that other smoothers have no value for the problem at hand. Rather, the goal is find at least one method that performs well in simulations, and the running interval smoother is relatively easy to implement.

A fairly well-known alternative is the smoother derived by Cleveland (1979) which includes a method of down weighting extreme Y values. One reason for choosing a running interval smoother is that when used with a 20% trimmed mean, it seems to be a bit better at handling moderately large or small outliers, versus Cleveland's method, and it seems to perform reasonably well compared to a variety of other smoothers that might be used (Wilcox, 2005). Again, this is not to suggest that all other smoothers be eliminated from consideration for the problem at hand, but the relative merits of using other smoothers is left for future investigations.

The running interval smoother is applied as follows. Let M be the median of the values $X_1,...,X_n$. The median absolute deviation (MAD), based on $X_1,...,X_n$, is the median of the n values $|X_1 - M|,...,|X_n - M|$. Let MADN=MAD/.6745; under normality, MADN estimates σ, the standard deviation. Let κ be some constant that is chosen in a manner to be described. Then the point X is said to be close to X_i if

$$|X_i - X| \le \kappa \times MADN.$$

The constant κ is called the span. Thus, for normal distributions, X is close to X_i if X is within κ standard deviations of X_i. Let

$$N(X_i) = \{j: |X_j - X_i| \le \kappa \times MADN\}.$$

That is, $N(x_i)$ indexes all X_j values that are close to X_i. Now consider the random sample $(X_1, Y_1),...,(X_n, Y_n)$ and let $\hat{\theta}_i$ be an estimate of some parameter of interest, based on the Y_j values such that $j \in N(X_i)$. That is, use all of the Y_j values for which X_j is close to X_i. Here, as

previously indicated, a mean or 20% trimmed mean is used. So, for example, $\hat{\theta}_i$ might be estimated with the 20% trimmed mean of the Y_j values such that $j \in N(X_i)$. In exploratory work, a good choice for the span is often κ=.8 or 1, but for the situation at hand an alternative choice is needed.

Virtually any smoother, including the one used here, can be extended to the generalized additive model given by (2) using the backfitting algorithm in Hastie and Tibshirani (1990). Set k=0 and let f_j^0 be some initial estimate of f_j (j=1,,...,p). Here, $f_j^0 = S_j\left(X \mid Y_j\right)$, where $S_j(Y|X_j)$ is the running interval smooth based on the jth predictor, ignoring the other predictors under investigation. Next, iterate as follows:

1. Increment k.
2. For each j, j=1,...,p, let
$$f_j^k = S_j(Y - \sum_{l \neq j} f_l \mid X_j).$$
3. Repeat steps 1 and 2 until convergence.

Finally, estimate β_0 with

$$b_0 = m(Y - \sum f_j^k),$$

where m indicates the measure of location used when computing the smooth, which here is taken to be a 20% trimmed mean or the usual mean.

Testing H_0

For convenience, momentarily assume the goal is to test

$$H_0 : f_1(X_1) = 0.$$

The proposed method begins by fitting the generalized additive model as described in the previous section using the $X_2,...,X_p$ values, ignoring X_1, yielding

$$\hat{Y}_i = b_0 + \hat{f}_2(X_{i2}) + \cdots + \hat{f}_p(X_{ip}),$$

where $\hat{f}_j(X_{ij})$ is the estimate of $f_j(X_{ij})$ based on the backfitting algorithm. Let $r_i = Y_i - \hat{Y}_i$, i=1,...,n. Then the strategy is to test the hypothesis that when predicting the residuals, given X_1, the regression is a straight horizontal line. This is done using the wild bootstrap method derived by Stute, González-Manteiga and Presedo-Quindimil (1998). As is evident, the method is readily modified to test (3) for any j.

To elaborate, let $\bar{r}_t$ be the mean or 20% trimmed mean based on the residuals $r_1,...,r_n$. Fix j and set I_i=1 if $\boldsymbol{X}_i \leq \boldsymbol{X}_j$, otherwise I_i=0. The notation $\boldsymbol{X}_i \leq \boldsymbol{X}_j$ means that for every k, k=1,...,p, $X_{ik} \leq X_{jk}$. Let

$$R_j = \frac{1}{\sqrt{n}} \sum I_i \left(r_i - \bar{r}_t\right) = \frac{1}{\sqrt{n}} \sum I_i v_i \quad , \qquad (4)$$

where

$$v_i = r_i - \bar{r}_t .$$

The test statistic is the maximum absolute value of all the R_j values. That is, the test statistic is

$$D = max \mid R_j \mid . \qquad (5)$$

An appropriate critical value is estimated with the *wild bootstrap method* as follows. Generate $U_1,...,U_n$ from a uniform distribution and set

$$V_i = \sqrt{12}(U_i - .5),$$

$$v_i^* = v_i V_i,$$

and

$$r_i^* = \bar{r}_t + v_i^*.$$

Then based on the n pairs of points $(\mathbf{X}_1, r_1^*),...,(\mathbf{X}_n, r_n^*)$, compute the test statistic as described in the previous paragraph and label it D^*. Repeat this process B times and label the resulting (bootstrap) test statistics $D_1^*,...,D_B^*$. Finally, put these B values in ascending order yielding $D_{(1)}^* \leq \cdots \leq D_{(B)}^*$. Then the critical value is $D_{(u)}^*$, where u=(1-α)B rounded to the nearest integer. That is, reject if

$$D \geq D_{(u)}^*.$$

Based on Theorem 1 in Stute et al. (1998), this method is valid under weak assumptions placed on $\boldsymbol{X}$.

For convenience, when using means, the technique just described will be called method V1. When using a 20% trimmed mean, it will be called method V2. Note that a smooth can be fit using a 20% trimmed mean, but when using the wild bootstrap in conjunction with the resulting residuals, one could use the mean of the residuals, rather than a 20% trimmed mean, when testing H_0. This will be called method V3.

Choosing the Span

There remains the issue of choosing the span, κ, when fitting the generalized additive model. Preliminary simulations indicated that if the span is too large, the actual probability of a Type I error can exceed the nominal level (cf. Härdle & Mammen, 1993). A proper choice for the span, given n and the amount of trimming, was found to correct this problem. That is, the choice of the span when using means differs from the choice when using a 20% trimmed mean instead. Here the span was determined by assuming that $X_1,...,X_p$ and ε have independent standard normal distributions, and then for a given sample size and depending on whether means or 20% trimmed means were used, κ was determined via simulations so that the actual probability of a Type I error is approximately equal to the nominal level when testing at the .05 level. Then given n, and depending on whether a trimmed mean was to be used, this value for κ was used in the simulations described in the next section. All indications are that the choice for the span does not depend on p for p=2, 3, 4 and 5. (Whether this remains true for p>5 has not been investigated.) The results are summarized in Table 1.

Table 1: Choices for the span, κ

n	20% trimming	mean
20	1.20	.80
40	1.0	.70
60	.85	.55
80	.75	.50
120	.65	.50
160	.65	.50

Simulation Results

Simulations were used to check the small-sample properties of the proposed method. Observations were generated where the marginal distributions have a g-and-h distribution (Hoaglin, 1985) which includes the normal distribution as a special case. More precisely, observations Z_{ij}, (i=1,...,n; j=1, 2) were initially generated from a multivariate normal distribution having correlation ρ, then the marginal distributions were transformed to

$$X_{ij} = \begin{cases} \dfrac{exp(gZ_{ij})-1}{g} exp(hZ_{ij}^2/2), & if \quad g > 0 \\ Zexp(hZ_{ij}^2/2), & if \quad g = 0 \end{cases}$$

where g and h are parameters that determine the third and fourth moments. The four (marginal) g-and-h distributions examined were the standard normal (g=h=0), a symmetric heavy-tailed distribution (g=0, h=.5), an asymmetric

distribution with relatively light tails (g=.5, h=0), and an asymmetric distribution with heavy tails (g=h=.5). Here, two choices for ρ were considered: 0 and .5. Table 1 shows the theoretical skewness (κ_1) and kurtosis (κ_2) for each distribution considered. When g>0 and h>1/k, $E(X^k)$ is not defined and the corresponding entry in Table 1 is left blank. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985). Some of these distributions might appear to represent extreme departures from normality, but the idea is that if a method performs reasonably well in these cases, this helps support the notion that they will perform well under conditions found in practice.

Table 2: Some properties of the g-and-h distribution.

g	h	κ_1	κ_2	$\hat{\kappa}_1$	$\hat{\kappa}_2$
0.0	0.0	0.00	3.0	0.00	3.0
0.0	0.5	0.00	—	0.00	11,896.2
0.5	0.0	1.75	8.9	1.81	9.7
0.5	0.5	—	—	120.10	18,393.6

A possible objection to Table 2 when performing simulations is that the distribution of observations generated on a computer does not always have the theoretical skewness and kurtosis values shown. The reason is that observations generated on a computer come from a bounded interval, so the skewness and kurtosis of the distribution will be finite, even when in theory it should be infinite. Accordingly, Table 2 also reports the estimated skewness ($\hat{\kappa}_1$) and kurtosis ($\hat{\kappa}_2$) values based on simulations with 10,000 replications.

Two sets of simulations were run. The first was for p=3 with the goal of testing H_0:$f_3(X_3)$=0. The correlation between X_1 and X_2 was taken to be either 0 or .5, and observations were generated according to one of three models: $Y=\varepsilon$, $Y = X_1 + X_2 + \varepsilon$ and $Y = X_1 + 2X_2^2 + \varepsilon$. Table 3 contains $\hat{\alpha}$, the estimated probability of making a Type I error when testing at the .05 level with n=20, when ρ=0. Increasing ρ to .5 had a negligible effect, so for brevity the results are not reported.

It is noted, however, that if n=20 and ρ=.7, then some effect on the probability of a Type I error results: it tends to decrease somewhat versus situations where ρ=.5 or 0. But for n=40, this was no longer the case. Introducing curvature had more of an effect, and so results for this case are reported. No situation was found where the estimated probability of a Type I error exceeded .065 when testing at the .05 level, and the lowest estimate was .030 except when ρ=.7, in which case, with n=20, $\hat{\alpha}$ goes as low as .017.

The second set of simulations was for p=5. Again observations were generated according to the models $Y=\varepsilon$, $Y = X_1 + X_2 + \varepsilon$ and $Y = X_1 + 2X_2^2 + \varepsilon$, only now the goal was to test H_0: $f_5(X_5)$=0. Obviously, with p=5, it is difficult to consider the many variations that might arise when the null hypothesis is true. Here, as a partial check on the method, some additional simulations were run assuming normality. If, for example, the models

$$Y = X_1 + X_2 + X_3 + X_4 + \varepsilon$$

and

$$Y = X_1 + X_2 + X_3^2 + X_4^3 + \varepsilon$$

are used to generate the data, the estimated probability of a Type I error when testing at the .05 level was .049 and .045, respectively.

Power

Now consider power. Various situations were considered and it was found that regardless of the distributions used, or the model used to generate the data, method V1 always had higher power than V2, and often the gain in power was substantial. For example, with n=40 and $Y = X_1 + .5X_3 + \varepsilon$, if X_1, X_2 and ε have independent standard normal random variables, power is .63 for method V1 and .40 for method V2. If instead ε has a symmetric, heavy-tailed distribution (g=0 and h=.5), now power is .19 and .06 for methods V1 and V2, respectively.

So, based purely on Type I error and power, all indications are that the approach based on means performs well, and there is no known reason for preferring the method based on a trimmed mean. This is not to suggest completely ruling out a trimmed mean, because using a trimmed can result in a better fit to the data, but in terms of detecting situations where a component of a generalized additive model differs from zero, using the mean appears to be preferable.

The reason method V2 has relatively low power is evidently related to using a trimmed mean applied to the residuals when using the wild bootstrap method. If method V3 is used instead, the problem of relatively low power, when using a general additive model based on a trimmed mean, is reduced substantially. Table 4 shows power estimates when using V1 versus V3. Often there is little separating the two methods, but even now, V1 has uniformly higher power. This continued to be the case when data were generated from non-linear models.

Conclusion

Of course, simulations cannot prove that a particular method always controls the probability of a Type I error, or that one particular method always has higher power than another. Nevertheless, all indications are that method V1, based on means, dominates in terms of power, and it performs well in terms of controlling the probability of a Type I error under what would seem like fairly extreme departures from normality.

References

Breiman, L., & Friedman, J. H. (1985). Rejoinder to comments on estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, *80*, 614–619.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*, 829–836.

Cleveland, W. S., & Devlin, S. J., (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, *83*, 596–610.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, *76*, 817–823.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. NY: Wiley.

Härdle, W. (1990). *Applied nonparametric regression*. Cambridge: Cambridge University Press.

Härdle, W., & Mammen, E. (1993). Comparing non-parametric versus parametric regression fits. *Annals of Statistics*, *21*, 1926–1947.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman and Hall.

Hoaglin, D. C. (1985) Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.) *Exploring data tables, trends, and shapes*. NY: Wiley, p. 461–515.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.) *Undeststanding robust and exploratory data analysis*. NY: Wiley, p. 297–336.

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. NY: Wiley.

Stute, W., Gonzalez-Manteiga, W. G., & Presedo-Quindimil, M. P. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association*, *93*, 141–149.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. (2nd ed.). San Diego, CA: Academic Press.

Table 3: Estimated probability of a Type I error, n=20

				p=3		p=5	
g	h	g	h	$Y = X_1 + X_2 + \varepsilon$	$Y = X_1 + X_2^2 + \varepsilon$	$Y = X_1 + X_2 + \varepsilon$	$Y = X_1 + X_2^2 + \varepsilon$
0.0	0.0	0.0	0.0	.051	.065	.049	.050
		0.0	0.5	.041	.048	.030	.038
		0.5	0.0	.044	.062	.040	.046
		0.5	0.5	.033	.037	.023	.038
0.0	0.5	0.0	0.0	.047	.045	.039	.051
		0.0	0.5	.043	.046	.036	.040
		0.5	0.0	.040	.043	.043	.050
		0.5	0.5	.038	.039	.036	.043
0.5	0.0	0.0	0.0	.053	.055	.047	.042
		0.0	0.5	.038	.053	.041	.042
		0.5	0.0	.050	.056	.042	.041
		0.5	0.5	.036	.043	.031	.034
0.5	0.5	0.0	0.0	.043	.040	.040	.049
		0.0	0.5	.039	.043	.041	.046
		0.5	0.0	.045	.043	.041	.046
		0.5	0.5	.036	.043	.031	.046

Table 4: Estimated power, n=40, p=3, $Y = X_1 + .5X_3 + \varepsilon$

g	h	g	h	Method V1	Method V3
0.0	0.0	0.0	0.0	.63	.63
		0.0	0.5	.19	.18
		0.5	0.0	.53	.52
		0.5	0.5	.15	.15
0.0	0.5	0.0	0.0	.64	.61
		0.0	0.5	.29	.27
		0.5	0.0	.59	.37
		0.5	0.5	.26	.24
0.5	0.0	0.0	0.0	.68	.45
		0.0	0.5	.21	.20
		0.5	0.0	.57	.42
		0.5	0.5	.19	.17
0.5	0.5	0.0	0.0	.56	.34
		0.0	0.5	.28	.24
		0.5	0.0	.52	.34
		0.5	0.5	.26	.24