Contents lists available at GrowingScience

Management Science Letters

homepage: www.GrowingScience.com/msl

A Semi parametric approach to dual modeling

Majid Navaee[a], **Mohammad Sadegh Mobin**[b], **Mohsen Haghverdi Vardani**[a] and **Nima Ahmadi**[c]

[a]*Department of Economics and Statistics, Allameh Tabataba'i University, Tehran, Iran*
[b]*Department of Management, Allameh Tabataba'i University, Tehran, Iran*
[c]*Department of Philosophy, Tarbiat Modares University, Iran*

ARTICLEINFO

Article history:
Received July 25, 2011
Received in Revised form
October, 19, 2011
Accepted 25 October 2011
Available online
6 November 2011

Keywords:
Robust parameter design
Nonparametric dual model
Parametric model
DMRR
RPD

ABSTRACT

Parameter design or robust parameter design (RPD) is a statistical methodology used mostly in engineering fields as a cost-effective approach for improving the quality of products and processes. The primary goal of parameter design is to choose the levels of the control variables, which optimizes a defined quality characteristic. Modeling both the mean and variance is commonly referred to as dual modeling. In parametric dual modeling, estimations of the mean and variance parameters are interrelated. When one or both of the models (the mean or variance model) are mis-specified, parametric dual modeling can lead to faulty inferences. An alternative to parametric dual modeling is nonparametric dual modeling. However, nonparametric techniques often result in estimates characterized by high variability, which leads us to ignore important knowledge. We develop a dual modeling approach called dual model robust regression (DMRR), which is robust against user misspecification of the mean and/or variance models. Numerical and asymptotic results illustrate the advantages of DMRR over several other dual model procedures. The proposed method will be illustrated with simulations.

1. Introduction

During the past three decades, there have been tremendous efforts on proposing new methods to design parameters. Taguchi proposed a cost-efficient method to quality improvement known as robust parameter design (RPD) (Aitkin, 1987; Einsporn & Birch, 1993; Mays et al., 2000). According to Taguchi method, there are two types of factors namely control factors and noise factors. Control factors are variables whose levels do not change in the process once they are set. However, the levels of the noise factors vary randomly within the process and may cause un-wanted changes in the response, *y*. The goal of robust parameter design is to find levels of the control factors, which cause the response to be robust against changes in the levels of the noise components.

There are several disadvantages on using Taguchi method, which leads other researchers to propose other sophisticated methods. In response surface modeling (RSM), we first determine some influencing factors and design some experiments to collect the behavior of response variable and

* Corresponding author. +989365540165
E-mail addresses: bigane2004@yahoo.com

doi: 10.5267/j.msl.2011.11.001

using simple linear or quadratic function, a surface model is estimated using different statistical methods such as ordinary least square, etc. There are different regression techniques of parametric, non-parametric and semi-parametric (Rahman et al., 1997; Ruppert et al., 2003). In most models specially for estimating variance, non-parametric models provides better estimation of curvature but they may not result promising results for the information with unusual skewness. Parametric models, on the other hand, may result unbiased results. The proposed model of this paper presents a hybrid of two models to incorporate the advantages of both methods by introducing a new convex parameter.

The proposed model of this paper first presents the parametric and non-parametric models in section 2 and the semi-parametric model is introduced in section 3. Section 4 compares the performance of the proposed model with parametric and non-parametric models. Finally, concluding remarks are given in the last to summarize the contribution of the paper.

2. Parametric and non-parametric models for estimating mean and variance

2.1. Parametric models

Given the data from a crossed array, there are different techniques to directly modeling the mean and variance as a function of various control factors. A general approach is to consider the underlying functional forms for the mean and variance models, which could be stated, parametrically. Let *d* be point design with n_i replicates at each location (*i* = 1,…, *d*), the point estimators of the process mean and variance, $\bar{y}_i$ and s_i^2, respectively. A popular form of response surface is defined in linear form as follows,

$$\bar{y}_i = h(x_i) + g^{1/2}(z_i\gamma)\varepsilon_i = x_i'\beta + g^{1/2}(x_i^{*\prime};\gamma)\varepsilon_i, \tag{1}$$

where $x_i' = (1, x_{i1}, x_{i2}, ..., x_{ik})$ and $x_i^{*\prime} = (1, x^*_{i1}, x^*_{i2}, ..., x^*_{il})$ are independent variables associated with mean and variances, respectively. In addition, $\beta_{1\times(k+1)}$ and $\gamma_{1\times(l+1)}$ are parameters for mean and variances models respectively, *g* is the function of variance and ε_i are error terms, which are assumed with mean of zero and variance of σ^2. Bartlett and Kendall (1946) presented a logarithmic form of regression function when Var(σ^2) is not uniform, which is as follows,

$$\ln(s_i^2) = g^*(x_i^*) + \eta_i = x_i^{*\prime}\gamma + \eta_i, \tag{2}$$

where η_i determines the behavior of error terms. We normally use ordinary least square or expected weighted least square techniques (EWLS) to estimates the parameters with the following steps,

Step 1. Using the ordinary least square technique, all parameters are estimated as follows,

$$\hat{\gamma}^{(ols)} = (x^{*\prime}x^*)^{-1}x^*y^*. \tag{3}$$

where $y^*_{d\times 1}$ is the logarithm of variance of sample size.

Step 2. $\hat{\sigma}_i^2 = \exp(x_i^{*\prime}\hat{\gamma}^{(OLS)}) = \exp(\hat{y}_I^{(OLS)})$ is calculated as the variance of the mean model and it is used to calculate the following,

$$\hat{V} = diag(\hat{\sigma}_1^2, \hat{\sigma}_2^2, ..., \hat{\sigma}_d^2) \;\; if \;\; n_i = n \;\; for(i = 1, 2, ..., d). \tag{4}$$

Step 3. Using the inverse of *V* the parameters estimated weighted least square is estimated as follows,

$$\hat{\beta}^{(EWLS)} = (x'\hat{V}^{-1}x)^{-1}x'\hat{V}^{-1}\bar{y}x = (x_1, x_2, ..., x_d)', \tag{5}$$

where $\bar{y}_{d\times 1}$ is the vector of mean of samples with $X = (X_1, X_2, ..., X_d)'$. Finally, mean and variance parameters are calculated as follows,

$$\hat{E}(y_i)^{EWLS} = x_i'\hat{\beta}, \hat{V}(y_i)^{(ols)} = \exp(x_i^{*\prime}\hat{\gamma}^{(ols)}) \tag{6}$$

2.2. Non-parametric models

There are many events where we do not have much information on curvatures of mean and variance model. In this case, using parametric models yields weak results. Anderson-Cook and Prewitt (2005) presented a non-parametric models as follows,

$$\bar{y}_i = h(x_i') + g^{1/2}(x_i'^{*})\varepsilon_i, \tag{7}$$

$$\ln(s_i^2) = g^{*}(x_i'^{*}) + \eta_i, \tag{8}$$

where Eq. (7) and Eq. (8) represent the mean and variance, respectively. h and g^* are functions of mean and variance with unknown curvature but known smooth curvatures. There are different non-parametric models and one of the most well known ones is called local polynomial regression.

2.1.1 Local polynomial regression

Pickle et al. (2008) first introduced local polynomial regression (LPR), where a Kernel function (Fan et al., 1995) at an arbitrary point $x_0 = (x_{01}, x_{02}, ..., x_{0k})$ is defined as follows,

$$K(x_0, x_i) = \frac{1}{b^k}\prod_{j=1}^{k} K(\frac{x_{0j} - \tilde{x}_{ij}}{b}), \tag{9}$$

where $\tilde{x}_i = (x_{i1}, x_{i2}, ..., x_{ik})$ and $K(x_{0j} - \tilde{x}_{ij})$ is a single variable Kernel function and *b* is the band wide. There are different versions of Kernel function and a popular one is represented as $K(z) = e^{-u^2}$, which is called Gaussian. One important factor is to find a fair value for band wide, *b* (Starnes, 1993; Mays et al., 2000). Anderson-Cook & Prewitt (2005) showed that when we choose a large value for *b* we get a relatively small value for variance but the estimation becomes unbiased. On the other hand, choosing a small value for *b* yields bigger values for variances but the estimate becomes less unbiased. Mays et al. (2001) presented different methods for estimating band wide and introduced the following as an estimators,

$$PRESS^{**} = \frac{PRESS}{d - trace(H^{(LLR)}) + (d - (k+1))\dfrac{SSE_{\max} - SSE_b}{SSE_{\max}}}, \tag{10}$$

where $SSE_{\max}$ is the sum of the square error of band wide for the whole estimation, SSE_b is the sum of the square error of a specific band wide, *b* and *k* is the number of estimated parameters.

2.1.2. Estimating mean and variance using Local polynomial and Estimated Weighted Local Linear Regression

Lin and Carroll, (2000) presented a method for non-parametric function estimation for clustered data when the predictor is measured without/with error, which is as follows,

$$\hat{E}(y_0)^{(EWLLR)} = X_0^{'}\beta^{(EWLLR)} = X_0{}'(X'W_0X)^{-1}X'W_0\overline{y} = h_0^{(EWLLR)'}\overline{y}\,, \tag{11}$$

$$\hat{V}ar[y_0]^{(LLR)} = \hat{\sigma}_0^2 = \exp[X_0{}^{*\prime}\hat{y}^{(LLR)}] = \exp[X_0{}^{*\prime}(X^{*\prime}W_0^{*}X^{*})^{-1}X^{*\prime}W_0^{*}y^{*}] = \exp[h_0^{(LLR)'}y^{*}]\,, \tag{12}$$

Note that in parameter estimation we look for minimizing the variance by using ordinary least square techniques and in the events the information of the function does not exist explicitly, we may switch to metaheuristics approaches such as genetic algorithm, simulated annealing, etc.

3. Semi-parametric method

In this section, we present the proposed semi-parametric model of this paper. There are different techniques to combine parametric and non-parametric models and the proposed model of this paper uses dual model robust regression model.

3.1. Dual model robust regression

Robinson and Birch (2000) presented a dual model robust regression model, where there is misrepresentation in data and provided good estimation for the results. The proposed regression model is stated as follows,

$$\overline{y}_i = h(x_i{}';\beta) + f\,(x_i) + g^{1/2}(x_i{}^{*\prime};\gamma)\varepsilon_i\,, \tag{13}$$

$$\ln(s_i{}^2) = g^{*}(x_i{}^{*\prime};\gamma) + l\,(x_i{}^{*}) + \eta_i\,, \tag{14}$$

where $h(x_i{}';\beta)$ and $g^{*}(x_i^{*'};\gamma)$ represent the mean and the variance, respectively. $f\,(x_i)$ and $l\,(x_i{}^{*})$ represent the unfitness of the functions, respectively. The method assumes there are two smooth functions for *f* and *l*. The algorithm has the following steps,

Step 1. Calculate the variance of model $\ln(s_i{}^2) = x_i{}^{*\prime}\gamma + l\,(x_i{}^{*}) + \eta_i$ as follows,

$$\begin{aligned} Var(\hat{y})^{(VMMR)} &= \exp[\lambda_\sigma\hat{y}^{*(LLR)} + (1-\lambda_\sigma)\hat{y}^{*(OLS)}] \\ &= \exp[\lambda_\sigma\hat{y}^{*(LLR)} + (1-\lambda_\sigma)H_\sigma^{(OLS)}y^{*}\;] = \exp[\lambda_\sigma H_\sigma^{(VMMR)}y^{*}], \end{aligned} \tag{15}$$

where $\lambda_\sigma \in [0,1]$ is the convex parameter and $H_\sigma^{(VMMR)}$ is a smooth matrix for variance of VMMR.

Step 2. Use $\hat{\sigma}_i{}^2 = \exp(h_{i,\sigma}^{(VMMR)'}y^{*})$ to calculate covariance of mean model, $\hat{V} = diag(\hat{\sigma}_1^2, \hat{\sigma}_2^2, ..., \hat{\sigma}_d^2)$ and $h_{i,\sigma}^{(VMMR)'}$ is the i^th^ row of matrix $H_\sigma^{(VMMR)}$,

Step 3. Use V^{-1} to estimate the parameters of EWLS which is as follows,

$$\hat{\beta}^{(`EWLS)} = (x'\hat{V}^{-1}x)^{-1}x'\hat{V}^{-1}\overline{y}\,, \hat{E}(y_i)^{EWLS} = x_i{}'\hat{\beta}^{(EWLS)} = H_\mu^{(EWLS)}\overline{y}\,. \tag{16}$$

Step 4. Calculate $r = \overline{Y} - \hat{E}(y)^{(EWLS)}$ and $\hat{r} = H_r^{(EWLS)}r$

Step 5. The robust mean model is calculated as follows,

$$\hat{E}[Y]^{(`MMRR)} = \hat{E}[Y]^{(`EWLS)} + \lambda_\mu\hat{r} = [H_\mu^{(EWLS)} + \lambda_\mu H_r^{(LLR)}(1-H_\mu^{(EWLS)})\overline{Y}] = H_\mu^{(MMRR)}\overline{Y}\,, \tag{17}$$

where $\lambda_\mu \in [0,1]$ is the parameter for the model. Both parameters b_μ and b_σ are calculated to minimize $PRESS^{**}$. One alternative to set the values is as follows,

$$\hat{\lambda}_\mu = \frac{\left\langle \hat{r}, y - \hat{y}^{(EWLS)} \right\rangle}{\left\| \hat{r} \right\|^2}, \; \hat{\lambda}_\sigma = \frac{\langle \hat{y}_{-1}{}^{(LLR)} - \hat{y}_{-1}{}^{(EWLS)}, y - \hat{y}_{-1}{}^{(EWLS)} \rangle}{\left\| \hat{y}^{(LLR)} - \hat{y}^{(EWLS)} \right\|}, \tag{18}$$

where $\hat{y}_{-1}{}^{(EWLS)}$ and $\hat{y}_{-1}{}^{(LLR)}$ are the same as $\hat{y}_{i,-i}{}^{(EWLS)}$ and $\hat{y}_{i,-i}{}^{(LLR)}$ for ith observation, $\langle \; \rangle$ indicates the inner product and $\| \; \|$ is the Euclidian norm.

4. Simulation

In this section, we compare three methods of parametric, non-parametric and semi-parametric using simulation technique. We compare the results for four different scenarios. In the scenario, we assume to have the precise shape of mean and variance functions. The second scenario considers to have precise form of mean but form of variance is not available. The third scenario studies the performance of three methods when we have the precise form of variance but the exact form of mean is not available. Finally, the last scenario assumes to have no information for either mean or variance. We use 500 set of data for dual problem using the following benchmark problem (Box & Draper, 1987; Burman & Chaudhuri, 1992),

$$\begin{aligned} y_i = \mu_i(x_i) &= 2 - 4x_{1i} - 8x_{2i} - 10x_{3i} + 2x_{1i}^2 + 4x_{2i}^2 + 2x_{3i}^2 \\ &+ \gamma_\mu [2Sin(\pi x_{1i}) + 2Sin(\pi x_{2i}) + 2Sin(\pi x_{3i}) + 4Cos(\pi x_{1i} x_{2i} x_{3i})] + g^{1/2}(X_i^*)\varepsilon_i, \end{aligned} \tag{19}$$

$$Ln(\sigma_i^2) = g^*(X_i) = 0.25 - 0.5x_{1i} - x_{2i} + 0.75x_{3i} + \gamma_\sigma [-2x_{1i}x_{2i}x_{3i} + x_{1i}^2 - 0.5x_{2i}^2 + 0.5x_i^3], \tag{20}$$

where $x_i = (x_{1i}, x_{2i}, x_{3i})'$ is the ith observation and all error terms are assumed to be normally distributed with the mean of zero and variance of 1. γ_μ and γ_σ are the parameters of mean and variance, respectively. All three variables of x_1, x_2 and x_3 receive three values of 0, -1 and 1. Therefore, there are 27 points with 81 experiments. We have also considered five values of 0, 0.25, 0.50, 0.75 and 1 for γ_μ and γ_σ. We have used simulated integrated mean square error of mean (SIMSEM) and simulated integrated mean square error of variance (SIMSEV). These formulas are calculated as follows,

$$SIMSEM = \frac{\sum asem}{500} \;,\; asem = \frac{\sum (E[y_i] - \hat{y}_i)^2}{8000}, \; SIMSEV = \frac{\sum asev}{500}, asev = \frac{\sum (\sigma_i^2 - \hat{\sigma}_i^2)^2}{8000}, \tag{22}$$

where *asem* and *asev* are the sum of squares of error terms for mean and variance, respectively. $E[y_i]$ and σ_i^2 are actual values of mean and variance at x_i, respectively. In addition, $\hat{y}_i$ and $\hat{\sigma}_i^2$ are estimated values of mean and variance at x_i, respectively.

4.1. The mean and variance have specified functions

In this case, we have $\gamma_\mu = \gamma_\sigma = 0$ and expect that parametric method perform better than other non-parametric and semi-parametric methods and it yields 0.742 and 21.272 for mean and variance of SIMSEM, respectively. These values are calculated as 0.849 and 21.012 for non-parametric and 2.376 and 32.016 for semi-parametric methods, respectively. As we expect, the first method outperforms other two methods.

4.2. The mean function is specified precisely but the variance function is not

Our experience indicate that parametric and semi-parametric methods performs reasonably better than non-parametric method.

4.3. The variance function is specified precisely but the mean function is not

In this case, the semi-parametric performs better than other methods for $\gamma_\mu \prec 0.5$ but non-parametric method performs better than other two methods when $\gamma_\mu \succ 0.5$. In this case, parametric method has the weakest performance compared with other methods.

4.4. Neither mean nor variance function is specified precisely

In this case, semi-parametric method performs better than other two methods for $\gamma_\mu \le 0.5 \; and \; \gamma_\sigma \succ 0.5$ but non-parametric method performs better than other methods with $\gamma_\mu \succ 0.5 \; and \; \gamma_\sigma \prec 0.5$. In this case, parametric method yields the weakest performance compared with other two methods. Table 1 and Table 2 show the results of Simulated integrated mean squared error values for the means model (**SIMSEM**) for 500 Monte Carlo runs.

Table 1
Simulated integrated mean squared error values for the means model (**SIMSEM**) for 500 Monte Carlo runs

γ_σ	γ_μ	Semi Parametric	Non-parametric	Parametric
0	0	0.849	2.376	**0.742**
0.25	0	0.859	2.556	**0.761**
0.5	0	0.862	2.665	**0.778**
0.75	0	0.868	2.733	**0.792**
1	0	0.871	2.814	0.803
0	0.25	**5.544**	8.704	6.926
0.25	0.25	**5.401**	8.772	7.012
0.5	0.25	**5.455**	8.920	7.228
0.75	0.25	**5.302**	8.988	7.332
1	0.25	**5.423**	8.089	7.448
0	0.50	**17.366**	18.889	19.489
0.25	0.50	**17.299**	18.922	19.667
0.50	0.50	**17.203**	18.988	19.778
0.75	0.50	**17.148**	19.1	19.882
1	0.50	**16.992**	19.221	19.976
0	0.75	52.718	**52.03**	55.987
0.25	0.75	52.322	**52.228**	56.102
0.5	0.75	**51.665**	53.114	56.204
0.75	0.75	**50.904**	53.922	56.322
1	0.75	**49.223**	54.966	56.387
0	1	81.021	**79.667**	92.880
0.25	1	80.884	**79.819**	92.923
0.5	1	80.803	**80.221**	93.107
.75	1	**80.664**	80.763	93.214
1	1	**80.120**	80.907	93.355

Table 2
Simulated integrated mean squared error values for the variance model (**SIMSEV**) for 500 Monte Carlo runs

γ_σ	Parametric	Non-parametric	Semi-parametric
0	**21.272**	24.449	22.012
0.25	**21.301**	26.665	21.338
0.50	24.334	28.334	**23.560**
0.75	26.365	29.806	**25.772**
1.00	29.409	33.912	**27.667**

5. Conclusion

In this paper, we have presented a dual modeling approach called dual model robust regression (DMRR), which is robust against user misspecification of the mean and/or variance models. The performance of the proposed model of this paper has been compared with two other parametric and non-parametric methods by implementing Monte Carlo simulation technique using a benchmark problem. To compare three approaches more generally, a simulation study was conducted. Variance model misspecification was observed to have little impact on the quality of the estimated mean. If the user correctly specifies the mean and variance models, the parametric approach came as the best strategy followed by semi-parametric method a the second best method. The nonparametric method, on the other hand, is vastly inferior in terms of SIMSEM. The nonparametric method, while best for large degrees of mean misspecification, is only slightly better than the proposed semi-parametric approach. When the mean is mis-specified, the parametric method is clearly worst one. For small to moderate mean misspecification, the semi parametric method is always superior. Since, in practice, one never knows if the forms of the underlying models are correctly specified, we specified a method that performs consistently well based on different degrees of potential misspecification.

Acknowledgment

The authors would like to thank the anonymous referees for constructive comments on earlier version of this work.

References

Aitkin, M. (1987). Modeling variance heterogeneity in Normal regression using GLIM. *Applied Statistics*, 36, 332-339.

Anderson-Cook, C.M., & Prewitt, K. (2005). Some guidelines for using nonparametric methods for modeling data from response surface designs. *Journal of Modern Applied Statistical Methods*, 4, 106-119.

Bartlett, M.S., & Kendall, D.G. (1946). The statistical analysis of variance heterogeneity and the logarithmic transformation. *Journal of Royal Statistical Society Series B*, 8, 128- 138.

Box, G., & Draper B. (1987). *Empirical Model Building and Response Surface*, Wiley, New York.

Box, G.E.P., & Meyer, R.D. (1986). Dispersion effects from fractional designs. *Technimetrics*, 28, 19-27.

Box, G.E.P., & Draper, N.R., (1987). *Empirical Model Building and Response Surface*. Wiley, New York.

Burman, P., & Chaudhuri, P. (1992). A Hybrid Approach to Parametric and Nonparametric Regression, *Technical Report No. 243*, Division of Statistics, University of California Davis, CA, USA.

Einsporn, R., & Birch, J.B. (1993). Model robust regression: using nonparametric regression to improve parametric regression analyses. Technical Report 93-5.Department of Statistics, Virginia Polytechnic Institute & State University, Blacksburg, VA.

Fan, J., & Gijbels, I. (1996). *Local Polynomial Modeling and its Applications*. Chapman & Hall, London.

Fan, J., Heckman, N.E., Wand, M.P., (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of American Statistical Association*, 90, 141-150.

Lin, X., & Carroll, R.J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association.*95, 520-534.

Mays, J. E., Birch, J. B., & R. L. Einsporn. (2000). An overview of model robust regression. *Journal of Statistical Computation and Simulation*. 66, 79-100.

Mays, J., Brich, J., & Starnes, B. (2001). Model robust regression: combining parametric, nonparametric and semi parametric methods. *Journal of Nonparametric Statistics*, 13, 245-270.

Pickle, S. M., Robinson, T.J., Birch, J.B., Anderson-Cook, C.M. (2008). A semi-parametric approach to robust parameter design. *Journal of Statistical Planning and inference*, 138, 114-131.

Robinson, T. J., Birch, J. and Alden Starnes, B. (2010). A semi-parametric approach to dual modeling when no replication exists. *Journal of Statistical Planning and Inference*, 140, 2860-2869.

Rahman, M., Gokhale, D.V., & Ullah, A. (1997). A Note on Combining Parametric and Nonparametric Regression. *Communications in Statistics-Simulation and Computation*, 26, 519-529.

Robinson T.J., & Birch, J.B. (2000). Model misspecification in parametric dual modeling. *Journal of Statistical Computation and Simulation*, 66, 113-126.

Ruppert, D., Wand, M.R., & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Starnes, B.A. (1999). Asymptotic results for model robust regression. Unpublished dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.