Regression and contrast estimates based on adaptive regressograms depending on qualitative explanatory variables

Olaf Bunke
Humboldt University, Berlin

Ernestina Castell
Havana University, Ciudad de la Habana

Summary: This methodological paper discusses the application of "adaptive" nonparametric procedures for estimating regression functions or contrasts in situations with quantitative regressands and qualitative regressors. We propose to apply an adaptive regressogram, that is the selection of a regressogram estimate among the class of regressograms corresponding to all possible partitions of the regressor range. Our selection criterion is an analog to Mallows' C_p and this allows to state some small sample and asymptotic properties of the adaptive estimator. We also comment on stepwise selection procedures. The details of the procedure are presented in several interesting special cases, e.g. the two-or three-sample problem and the twoway classification. We illustrate there possible improvements over the usual least squares (ANOVA-)estimates.

just the es**AMS 1991 subject classifications.** Primary 62G07, 62J05, 62J10; Secondary 62G05, 62C99, 62F12.

Key words and phrases. Adaptive least squares estimation, minimax regret, ANOVA, discrete explanatory variable, twoway classification.

---

* The research on this paper was carried out within the Sonderforschungsbereich 373 "Quantifikation und Simulation ökonomischer Prozesse" at Humboldt University Berlin and was printed using funds made available by the Deutsche Forschungsgemeinschaft.

1 Introduction

The classical treatment of a regression problem with a vector X of qualitative explanatory variables and a continuous (or quantitative) dependent variable Y is by least squares estimates (LSE) in ANOVA-models, that is, in additive regression models with (or without) interactions. The LSE are used for the estimation of the observation means, that is, of the values of the regression function at the regressor values or design points, as well as for the estimation of contrasts, that is, of linear functions of the means. Under a normal distribution of the observations there may be better estimators than the LSE in the sense of a small mean square error (MSE), as given by Stein related estimators (see e.g. Humak (1977)section 4.3.3.), but this not possible for one- or twodimensional contrasts and the Stein estimators may be notably worse than the LSE, if the distribution happens to be far from normal. Without normality assumption even a stronger improvement is possible, if some contrasts are known to be small. For such cases works the penalized LSE investigated by Green et al. (1985), a.o. (see also Bunke (1992)). Special cases of such situations occur, when some of the observations means are near together, that is, if the regression function has some "smoothness". Then we have conditions under which nonparametric regression estimators work, e.g. kernel or near-neighbour-estimators (see Härdle (1990)). But for their definition it is necessary to have a sensible notion of distance between values of the explanatory variable and this in general not possible for a qualitative (or nominal) variables.

Our paper is oriented toward methodology and its purpose is to propose and discuss nonparametric estimators, which behave especially well in the above mentioned situations, even if there is no prior information neither on the observation means nor on some contrasts. Our approach is introduced in section 2. consists in the interpretation of regressograms as LSE in linear models and in the application of model selection procedures to select a regressogram in the class of all regressograms. This is performed in section 3. and leads to an adaptive nonparametric estimator, which formally is also parametric. The properties of such intuitively well defined adaptive estimators, e.g. its MSE behaviour, are very difficult to clarify, but at least we discuss some asymptotic properties in section 4. Some small sample properties are only known (up to now) in the special case of a two-sample problem treated in section 5. This section illustrates our adaption procedure in three special cases, which often are of interest in applications: the two- and three-sample problem and the twoway classification. A case study indicates the possible improvement over the LSE. In section 6. we discuss stepwise selection procedures (e.g. a up- or downwards procedure), which may be useful if the number of different regressograms is too large for a feasable search of the best. The procedure is applied in section 7 in a case study to obtain a (suboptimal) regressogram.

2 On parametric and nonparametric estimators in regression models with qualitative explanatory variables

We consider a regression problem with qualitative or nominal explanatory variables $X_{(1)}, \ldots X_{(p)}$ and a continuous dependent variable Y. The observations of the vector $X = (X_{(1)}, \ldots, X_{(p)})$ of explanatory variables are assumed to be the elements of a finite set $\mathcal{X} = \{x_1, \ldots, x_m\}$. If $X_{(j)}$ takes values in $\mathcal{X}_j \quad (j = 1, \ldots, p)$, then $\mathcal{X}$ is a subset of the product set $\bigtimes_{j=1}^{r} \mathcal{X}_j$.

Let (x_i, y_{ij}) $(i = 1, \ldots, m; j = 1, \ldots, n_i)$ be $n = \sum_{i=1}^{m} n_i$ observations of the pair (X, Y). We assume the values $x_1, \ldots, x_m$ to be fixed, while $y_{11}, \ldots, y_{mn_m}$ are realizations of independent random variables $Y_{11}, \ldots, Y_{mn_m}$ with expectations and variances

$$EY_{ij} = f(x_i) =: \mu_i \ , \ DY_{ij} = \sigma^2. \quad (i = 1, \ldots, m; \ j = 1, \ldots, n_i). \tag{2.1}$$

There are statistical problems, that occur very frequently in applications and which have lead to a rich literature in statistics. Two of them are the following:

1. Estimation of vector $\mu = (\mu_1, \ldots, \mu_m)'$ of values of the "regression function" f or the related problem of prediction of the values of Y for fixed values of X.

2. Estimation of a linear parameter (or "contrast") $\gamma = C\mu$, where C is a $r \times m$ matrix and $\mu := (\mu_1, \ldots, \mu_m)'$. Problem 1 is a special case.

The classical treatment of the problems 1. and 2. is to use an additive regression model $g_\beta(x)$ with (or without) interactions and to estimate its parameter β by least squares. (see section 5). This leads to estimates $\hat{\mu}$ and $\hat{\gamma} = C\hat{\mu}$ of μ and γ. These estimators are BLUE (best linear unbiased estimators) if the additive model is adequate, that is, if $f(x) = g_{\beta_o}(x)$ for some value β_0, the "true parameter". In general, e.g. if the model g_β is not adequate, there is a bias in estimating μ or γ

$$E\hat{\mu} = \mu_* = Q\mu \quad , \quad E\hat{\gamma} = \gamma_* := C\mu_*, \tag{2.2}$$

where Q is the projection matrix corresponding to the linear subspace of R^m generated by the vectors

$$\underline{g}_\beta := (g_\beta(x_1), \ldots, g_\beta(x_m))' \quad , \quad \beta \in R^k. \tag{2.3}$$

If there is no completely reliable prior information on the dependence of Y on X leading to an adequate linear or nonlinear model, no restriction on the vector μ is allowed and the corresponding LSE or BLUE are given by the observation means.

$$\breve{\mu} = \overline{y} := (\overline{y}_1, \ldots, \overline{y}_m) \quad , \quad \overline{y}_i = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}. \tag{2.4}$$

As discussed in our introduction, for purely nominal explanatory variables we have (in general) to renounce to distances or to the concept of near neighbours. But it is at least possible to define abstract sets $N(x) \subset \mathcal{X}$ of "neighbours" of x. Then the "neighbour estimate" of μ_i ("running mean with window L_i") would be defined as

$$\hat{\mu}_i := \sum_{h \in L_i} n_h \overline{y}_h \,/\, \sum_{h \in L_i} n_h, \tag{2.5}$$

$$L_i := \{h \,|\, x_h \in N(x_i)\}. \tag{2.6}$$

Obviously this estimator will lead to a smaller MSE

$$MSE(\hat{\mu}) := E\|\hat{\mu} - \mu\|_W^2, \tag{2.7}$$

$$\|\mu\|_W^2 := \mu' W \mu \quad , \quad W : \text{ positive semidefinite matrix}, \tag{2.8}$$

than the mean (2.4), if in each subset $N(x_i)$ the values $f(x)$ of the regression function have a relatively small variability. As the regression function f is unknown, it is not known a priori, in which subsets $N(x_i)$ of $\mathcal{X}$ there is such a convenient behaviour of f. Therefore there is in general an overwhelming number of possible choices of such subsets. This number would be essentially restricted if the choice is limited to subsets given by a partition π

$$\mathcal{X} = \sum_{j=1}^{q} \mathcal{X}_j^{\pi} \tag{2.9}$$

of the set of values of X. The corresponding estimator $\hat{\mu}^{\pi}$ of μ is defined by (2.5) and

$$L_i = L_i^{\pi} := \{h \,|\, x_h,\ x_i \in \mathcal{X}_j^{\pi} \quad \text{for some } j\}. \tag{2.10}$$

It is just the estimator used on the CART procedure of Breiman et al. (1994). It may also be interpreted as a regressogram (see Härdle (1990)).

The assumption of a regression function f, which has a constant value η_j on each subset $\mathcal{X}_j^{\pi}$, determines a linear model $\mathcal{M}^{\pi}$ given by (2.1) and

$$\eta_j := \mu_i = \mu_h \quad \text{if} \quad i, h \in \mathcal{X}_j^{\pi} \ (j = 1, \ldots, q), \tag{2.11}$$

Under $\mathcal{M}^{\pi}$ the expectation vector μ may be expressed as a linear function $\mu = A^{\pi}\eta$ of the parameter $\eta \in R^q$. $\hat{\gamma}^{\pi} := C\hat{\mu}^{\pi}$ is just the BLUE of $\gamma = C\mu$ under the model $\mathcal{M}^{\pi}$.
In general the regression function has not exactly the above property and the model $\mathcal{M}^{\pi}$ is then inadequate, so that we have a bias in estimating μ which may be described in the form (2.2).

The choice of π in the set $\prod$ of all partitions of $\mathcal{X}$ (or equivalently of the estimator $\hat{\mu}^{\pi}$) corresponds to model selection in the class $\mathcal{M} := \{\mathcal{M}^{\pi} \,|\pi \in \prod\}$ of linear models. There is a rich literature on such procedures and we propose to use convenient modifications of the model selection criterrion introduced by Mallows (1973) (see also Bunke, Droge and Polzehl (1996)) which are directed to an efficient solution of the problems 1. and 2.
If m is relatively small it is even possible to use the possibly more accurate adaptive nonparametric "neighbour" estimator, selecting the system of m "neighbourhoods" $N(x_i)$ in (2.5), (2.6) among all such systems by the criterion.

3 Adaptive estimators of linear parameters based on model selection

The true regression function f^0 or the corresponding true expectation vector μ_0 uniquely determines a partition $\pi = \pi^0$ with (2.11) and different values $\eta_1, \ldots, \eta_{q_0}$. The model $\mathcal{M}^{\pi^0}$ may be called the true model. The estimators $\hat{\mu}^{\pi^0}$ of μ and $\hat{\gamma}^0 = C\hat{\mu}^{\pi_0}$ of γ given by the (unknown) true model are in general not the best in the sense of the MSE

$$MSE(\hat{\gamma}) := E\|\hat{\gamma} - \gamma\|_H^2 = E\|\hat{\mu} - \mu\|_W^2, \tag{3.1}$$

where $\hat{\gamma} = C\hat{\mu}$, H is a positive semidefinite matrix (weighting the errors in estimating the components of γ) and $W := C'HC$.
An "optimal" $\pi^* \in \prod$ defined by

$$MSE(\hat{\gamma}^{\pi^*}) = \min_{\pi \in \prod} MSE(\hat{\gamma}^{\pi}) \tag{3.2}$$

obviously may lead to an estimator with smaller MSE than $\hat{\mu}^{\pi^0}$ and also than the LSE $\breve{\gamma} := C\breve{\mu}$ given by the observation means, which corresponds to the trivial partition $\breve{\pi}$ of $\mathcal{X}$ in m singletons ($\breve{\mu} = \hat{\mu}^{\breve{\pi}}$).

As optimal estimators $\hat{\eta}$ depend on the unknown μ and σ^2, the determination of a data dependent partition $\hat{\pi} \in \prod$ aiming at least at a minimization of a good estimate $\widehat{M}(\pi)$ of MSE$(\hat{\mu}^{\pi})$ would be of interest:

$$\widehat{M}(\hat{\pi}) = \min_{\pi \in \prod} \widehat{M}(\pi). \tag{3.3}$$

A sensible unbiased estimator $\widehat{M}(\pi)$ is

$$\widehat{M}(\pi) := \|\hat{\gamma}^{\pi} - \overline{y}\|_W^2 + a_{\pi}\hat{\sigma}^2 \tag{3.4}$$

where

$$a_\pi := tr\{W(N[P^\pi]' + P^\pi N) - WN\}. \tag{3.5}$$

with

$$N := Diag[n_1^{-1}, \ldots, n_m^{-1}], \tag{3.6}$$

$$P^\pi := A^\pi([A^\pi]'N^{-1}A^\pi)^{-1}[A^\pi]'N^{-1}, \tag{3.7}$$

$$\hat{\sigma}^2 := (n-m)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_m}|y_{ij} - \overline{y}_i|^2, \tag{3.8}$$

assuming that there are at least two replicated observations for some $i \quad (n_i > 1)$. Under a normal distribution of the variables Y_{ij} the estimator $\widehat{M}(\pi)$ is obviously best unbiased (see Bunke and Droge (1984)). An alternative would be the use of a cross-validation estimate for the choice of π, but this seems to lead to less reliable estimators than "adaptive" estimators $\hat{\hat{\mu}} := \hat{\mu}^{\hat{\pi}}$, $\hat{\hat{\gamma}} := \hat{\gamma}^{\hat{\pi}}$ given by (3.4) (see Bunke and Droge (1984) and Bunke and Ilouga (1997)). In the special case $\gamma = \mu, W = I$ the estimate (3.5) is just the C_p-criterion for the model $\mathcal{M}^\pi$ (see Mallows (1973)) up to a term which does not depend on the model.

Remark on heteroscedastic variances

When the assumption of a homogeneous variance σ^2 is not realistic, the variances $\sigma_i^2 := DY_{ij}$ may depend on the values x_i ("design points") of the explanatory variables. If there are adequate estimates $\hat{\sigma}_i^2$ of the "heteroscedastic" variances σ_i^2 it is possible to use adaptive regressograms, provided the partition $\hat{\pi}$ (or π^*) is selected in view a small values of a sensible estimate

$$\hat{M}(\pi) := \|\hat{\gamma}^\pi - \overline{y}\|_W^2 + \hat{a}_\pi \tag{3.9}$$

of the MSE (2.7). The term $\hat{a}_\pi$ is defined by (3.4) replacing N by

$$\hat{N} := Diag[\hat{\sigma}_1^2 n_1^{-1}, \ldots, \hat{\sigma}_m^2 n_m^{-1}]. \tag{3.10}$$

In the case of replications at each design point ($n_i \geq 2$ for all i) we have the variance estimates

$$\hat{\sigma}_i^2 := (n_i - 1)^{-1}\sum_{j=1}^{n_i} |\, y_{ij} - \overline{y}_i \,|^2. \tag{3.11}$$

When there are no replications ($n_i = 1$) for some i then the estimate (4.3) is not defined. The application of usual nonparametric variance estimates (see e.g. Bunke, Droge and Polzehl 1995) will be possible only if there is a sensible notion of distance between design points x_i, which may not exist for nominal explanatory variables.

4 Asymptotic behaviour

From the theoretical point of view but also as a confirmation of an acceptable behaviour of the adaptive procedure $\hat{\hat{\gamma}}$ it is of interest to know large sample properties, e.g. under

(4.1) $\qquad n_i = n_i(n) \quad , \quad n \to \infty, \qquad n_i(n)/n \to c_i > 0.$

For simplicity we consider only the special case $\gamma = \mu, W = I_m$. We have then

(4.2) $\qquad MSE(\hat{\mu}^\pi) = \|\mu^\pi - \mu\|^2 + o(1) \quad$ a.s.,

where

(4.3) $\qquad \mu_i^\pi := \sum_{h \in L_i^\pi} n_h \mu_h \,/ \sum_{h \in L_i^\pi} n_h .$

It is obvious that a.s. for sufficiently large n a partition π_* will be optimal in the sense (3.2), if

(4.4) $\qquad \pi_* \in \prod_0 = \{\pi \mid \mu^\pi = \mu\}.$

For large n the unknown optimal partition is unique and just the true one ($\pi_* = \pi_0$). By the law of large numbers a.s. first term in (3.4) converges to $\|\mu^\pi - \mu\|^2$, while the second converges to zero.

If the components of the true μ_0 are different (as intuitively it should frequently be), then a.s. for large n we have $\hat{\pi} = \pi_0 = \pi_* = \breve{\pi}$ and $\hat{\hat{\mu}} = \breve{\mu}$. From this follows the strong consistency of $\hat{\hat{\mu}}$ and $\hat{\hat{\mu}}$ has the same asymtptotic normal distribution as the LSE $\breve{\mu}$.

If some components of μ_0 are identical, we have $\pi_0 \neq \breve{\pi}$. Then, even for large n, there will be a positive probability for $\hat{\pi} \neq \pi_0$, that is, we have $\lim_{n\to\infty} P(\hat{\pi} = \pi) > 0$, for all $\pi \in \prod_0$ but at least

(4.5) $\qquad \lim_{n\to\infty} P(\hat{\pi} \in \prod_0) = \lim_{n\to\infty} P(\mathcal{M}^{\hat{\pi}} \text{ is an adequate model }) = 1 \ ,$

see Nishii (1984) and Müller (1994). Therefore, in such "singular" cases (where $\pi_o \neq \breve{\pi}$), for large n the estimate $\hat{\hat{\mu}}$ will differ from the unknown optimal estimate $\hat{\mu}^{\pi_*}$ with positive probability. Nevertheless then the adaptive estimator $\hat{\hat{\mu}}$ will be strongly consistent and its mean square error $E\|\hat{\hat{\mu}} - \mu\|^2$ may be smaller than for the LSE $\overline{\mu}$ (see subsection 5.1.).

5 Special cases

For an illustration of the problem and the concepts introduced in section 2. and for the adaptive estimators presented in section 3. we treat some especially simple special cases. On the other hand these cases correspond to classical statistical applications, so that our approach suggests a different view at their treatment and possibly more reliable statistical inferences. Moreover in the most simple case our adaptive estimate coincides with an estimate already investigated by some authors.

5.1 Two-sample problem

The two-sample problem given by $m = 2$ is the most simple non-trivial special case of (2.1). Most frequently the difference (contrast) $\gamma = \mu_2 - \mu_1$ between the two sample expectations is of interest and sometimes $\mu = (\mu_1, \mu_2)'$, that is, the value of both means.
Here we have only two trivial partitions of $\mathcal{X}$:

$$\tilde{\pi} : \mathcal{X} = \{1,2\} \quad , \quad \breve{\pi} : \mathcal{X} = \{1\} + \{2\}. \tag{5.1}$$

The corresponding estimates $\tilde{\mu} = \hat{\mu}^{\tilde{\pi}}$, $\tilde{\gamma} := \hat{\gamma}^{\tilde{\pi}}$ and $\overline{\gamma}$ are given by

$$\tilde{\mu}_i = \overline{y} = n^{-1}(n_1\overline{y}_1 + n_2\overline{y}_2) \quad (i = 1,2) \tag{5.2}$$

$$\tilde{\gamma} = \tilde{\mu}_2 - \tilde{\mu}_1 = 0 \ , \ \breve{\gamma} = \overline{y}_2 - \overline{y}_1. \tag{5.3}$$

Taking $H = 1$ the MSE estimator (3.4) has the values

$$\widehat{M}(\tilde{\pi}) = \breve{\gamma}^2 - t\hat{\sigma}^2, \tag{5.4}$$

$$\widehat{M}(\breve{\pi}) = t\hat{\sigma}^2, \tag{5.5}$$

where $t := n/(n_1 n_2)$.
Consequently the adaptive estimator $\hat{\hat{\gamma}}$ has the form

$$\hat{\hat{\gamma}} = \hat{\hat{\gamma}}_c := \begin{cases} 0 & \breve{\gamma}^2 \le c\hat{\sigma}^2 \\ \breve{\gamma} & \breve{\gamma}^2 > c\hat{\sigma}^2 \end{cases} , \tag{5.6}$$

where $c := 2t$.

This estimator is a "testimator" or "pre-test-estimator" already well investigated in the literature. The corresponding MSE is under a normal distribution of the observations (see D/G: Droge and Georg (1995))

$$\begin{aligned} MSE(\hat{\hat{\gamma}}) \;=\; & t\sigma^2\{1 - \delta^2 F(2/5 \,|\, 5, n-2; \delta^2) - \\ & -F(2/3 \,|\, 3, n-2; \delta^2) + 2\delta^2 F(2/3 \,|\, 3, n-2; \delta^2)\} \end{aligned} \tag{5.7}$$

where $F(\cdot\,|\,r,s;\delta^2)$ denotes the distribution function of the noncentral F-distribution with r and s degrees of freedom and noncentrality δ^2 and

$$\delta^2 := \gamma^2/(t\sigma^2). \tag{5.8}$$

(i) They are especially good in comparison with standard estimators (here: $\breve{\gamma}$) in a certain region of the parameter space (here: "small" magnitude of the contrast γ),

(ii) in an intermediate region they are worse than the standard estimator but the MSE does not surpass some acceptable bound and

(iii) their MSE is neglegibly larger than that of the standard estimator in the complement of the above regions.

While the MSE of $\breve{\gamma}$ is the constant $MSE(\breve{\gamma}) = t\sigma^2$, in the most favourable case $\gamma = 0$ we have the smallest value

$$\min MSE(\hat{\hat{\gamma}}) = t\sigma^2\{1 - F(2/3\,|\,3, n-2; 0)\} \tag{5.9}$$

and in the most unfavourable case we have the largest value

$$\max_{\mu,\sigma} MSE(\hat{\hat{\gamma}}) = t\sigma^2(\tilde{R} + 1), \tag{5.10}$$

where $\tilde{R}$ is the maximum regret value corresponding to the adaptive estimator in the standard model selection problem with orthonormal explanatory variables treated by Droge and Georg (1995) (see our table 1 for some numerical values).

If we take $W = I$ in the MSE for estimating the expectation vector μ then from its estimate (3.5) we obtain similarily

$$\hat{\hat{\mu}}_i = \begin{cases} \overline{y} & \breve{\gamma}^2 \le 2t \quad (i = 1, 2). \\ \overline{y}_i & > 2t \end{cases} \tag{5.11}$$

and an analogous MSE behaviour.

An interesting question is, if the value of the constant c in an estimator $\hat{\hat{\gamma}}_c$ of the form (4.6) is sufficiently good although we have used a sensible model selection procedure. While the (unknown) value of c minimizing the $MSE(\hat{\hat{\gamma}}_c)$ would be $c = \infty$ if $\gamma^2 < t\sigma^2$ and $c = 0$ otherwise, there are also other applicable "optimal" values:

(i) The minimax value $c = 0$ and corresponds to the estimator $\breve{\gamma}$ (see D/G).

(ii) The minimax regret value

$$c = c_* := tc^*, \tag{5.12}$$

where c^* is the value corresponding to the minimax regret estimator in the standard problem treated in D/G. This follows because the form of the MSE (4.7) is identical to that investigated there. This value c_* is relatively near to the value $c = 2t$ in the adaptive estimator (5.6).
For the sake of comparison we state the minimal and maximal MSE values for $\hat{\hat{\gamma}}_{c_*}$:

$$\min_{\mu,\sigma} MSE(\hat{\hat{\gamma}}_{c_*}) = t\sigma^2\{1 - F(c^*/3, n-2; 0)\}, \tag{5.13}$$

$$\max_{\mu,\sigma} MSE(\hat{\hat{\gamma}}_{c_*}) = t\sigma^2\{R^* + 1\}, \tag{5.14}$$

where R^* is the maximum regret value corresponding to the minimax regret procedure in the standard problem of D/G (some numerical values are given in table 1).

Table 1 Values $c^*, R^*, \tilde{R}$ calculated in Droge and Georg (1995)

n-2	c^*	R^*	$\tilde{R}$
1	2.0739	0.6847	0.6919
2	1.9722	0.6501	0.6606
3	1.9387	0.6357	0.6594
4	1.9223	0.6277	0.6582
5	1.9126	0.6226	0.6572
7	1.9016	0.6164	0.6558
10	1.8935	0.6114	0.6545
15	1.8872	0.6074	0.6533
20	1.8841	0.6053	0.6527
30	1.8810	0.6032	0.6520
50	1.8785	0.6014	0.6515
100	1.8766	0.6001	0.6510
200	1.8757	0.5995	0.6508

5.2 Three-sample problem

The special case $m = 3$ of (2.1) is the next example in the order of simplicity already illustrating more realistically the working of our approach. There are five partitions of $\mathcal{X} = \{1, 2, 3\}$:

(5.15) $\tilde{\pi} : \mathcal{X},\ \pi^1 : \mathcal{X} = \{1\} + \{2,3\},\ \pi^2 : \mathcal{X} = \{2\} + \{1,3\},$
$\pi^3 : \mathcal{X} = \{3\} + \{1,2\} \quad , \quad \breve{\pi} : \mathcal{X} = \{1\} + \{2\} + \{3\}.$

If components in the vector $\gamma = (\mu_1 - \mu_2, \mu_2 - \mu_3, \mu_3 - \mu_1)$ are of equal interest the use of a unit weight matrix $H = I$ would be adequate. A comparison of the value of the MSE estimate (3.5) for the different partitions in (4.15) shows, that the adaptive estimator of γ is given with the notation $\hat{\tilde{\gamma}} = (\hat{\tilde{\mu}}_1 - \hat{\tilde{\mu}}_2, \hat{\tilde{\mu}}_2 - \hat{\tilde{\mu}}_3, \hat{\tilde{\mu}}_3 - \hat{\tilde{\mu}}_1)$, where
$\hat{\tilde{\mu}} = \overline{y}\mathbf{I}$ for $\breve{\nu} := \hat{\sigma}^{-1}\breve{\mu} \in \tilde{N}, \hat{\tilde{\mu}} = \breve{\mu}$ for $\breve{\nu} \in \breve{N}$
and

(5.16) $\hat{\tilde{\mu}}_j = \overline{y}_j, \qquad \hat{\tilde{\mu}}_i = (n_i + n_k)^{-1}(n_i\overline{y}_i + n_k\overline{y}_k), \qquad i \neq k \neq j),$

for $\breve{\nu} \in N_j$. The regions $\tilde{N}, \breve{N}, N_j$ are defined by:

(5.17) $N_{ik|j} = \overline{Q}_j \cap T_j \cap P_{ji} \cap P_{jk},$

(5.18) $\tilde{N} = Q \cap Q_j, \qquad \breve{N} = \overline{Q} \cap \overline{T}_j.$

(we assume here and in the following $i \neq k, j \neq k, i \neq j$),

(5.19) $Q := \{\breve{\nu} | \sum_{i<j} \Delta_{ij} < 2d\}$

(5.20) $Q_j := \{\breve{\nu} \mid \sum_{l \neq j} \Delta_{lj} < d + d_{ik}\}$

(5.21) $T_j = \{\breve{\nu} \mid \Delta_{ik} < d - d_{ik}\}$

(5.22) $P_{ji} := \{\breve{\nu} \mid \Delta_{ik} - \Delta_{ij} < d_{ij} - d_{ik}\}$

(5.23) $\Delta_{ij} := |\overline{y}_i - \overline{y}_j|^2/\hat{\sigma}^2,$

(5.24) $d := 2\sum_{i<j} n_i n_j / n_1 n_2 n_3$

(5.25) $d_{ij} := [n_1 n_2 n_3(n^2 - n_i^2 + n_j^2 + n_i n_j)]^{-1} \cdot$
$\cdot[n_i n_j(n_i + n_j) - n_k(n_i^2 + n_j^2)](n_i + n_j),$

The behaviour of $MSE(\hat{\tilde{\gamma}})$ will in principle follow the same pattern as that in the example 1, although now its calculation seems to be very cumbersome.

5.3 Twoway classification

We consider a four-sample-problem appearing in the case of two explanatory variables ($q = 2$ in section 2.), each variable having the two values 1,2. Then $\mathcal{X} = \{(i,j) \,|\, i,j = 1,2\}$ consists of four vectors and the corresponding n_{ij} replicated observations are denoted by y_{ijk} $(k = 1, \ldots, n_{ij})$. A twoway ANOVA model with interactions would be of the form

$$\mu_{ij} := EY_{ijk} = \overline{\mu}_{..} + \alpha_i + \beta_j + \gamma_{ij}, \quad DY_{ijk} = \sigma^2\,, \tag{5.26}$$

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0. \tag{5.27}$$

Often there is an interest in the estimation of the main effects of the explanatory variables, say of $\alpha = (\alpha_1, \alpha_2)'$. The classical estimate for α_i would be
$\breve{\alpha}_i = \overline{y}_{i..} - \overline{y}_{...}$, where

$$\overline{y}_{i..} := n_{i.}^{-1} y_{i..}, \quad \overline{y}_{...} := n_{..}^{-1} y_{...} \tag{5.28}$$

and where the dot . indicates sumation over the corresponding index.
Our approach is based on the observation means $\hat{\mu}_{ij}^{\pi}$ defined in section 2 corresponding to partitions π of the set $\mathcal{X}$. The corresponding estimate of $\alpha_i = \overline{\mu}_{i.} - \overline{\mu}_{..}$ will be

$$\hat{\alpha}_i^{\pi} = \hat{\mu}_{i.}^{\pi} - \hat{\mu}_{...}^{\pi} \tag{5.29}$$

The data dependent partition $\hat{\pi}$ is a partition under the 15 different partitions of $\mathcal{X}$ with minimal value of $\hat{M}(\pi)$, the estimator (3.5) of the MSE

$$MSE(\hat{\alpha}^{\pi}) = E|\hat{\alpha}_1^{\pi} - \alpha_1|^2 + E|\hat{\alpha}_2^{\pi} - \alpha_2|^2 = 2E|\hat{\alpha}_1^{\pi} - \alpha_1|^2. \tag{5.30}$$

It is again intuitively clear, that the adaptive estimator $\hat{\hat{\alpha}} = \hat{\alpha}^{\hat{\pi}}$ will be more accurate than $\breve{\alpha}$, if the magnitude of the contrast α_1 is small and will not differ much from that of $\breve{\alpha}$, if this magnitude is sufficiently large, while for all other intermediate magnitudes the MSE of $\hat{\hat{\alpha}}$ may be larger than that of $\breve{\alpha}$, but not surpass some bound $\overline{M}$.

Sometimes an estimate of the interaction terms $\gamma_{ij} = \mu_{ij} - \overline{\mu}_{i.} - \overline{\mu}_{.j} + \overline{\mu}_{..}$ may of be interest and our approach would yield estimates $\hat{\hat{\gamma}}_{ij} = \hat{\hat{\mu}}_{ij} - \hat{\hat{\mu}}_{.j} - \hat{\hat{\mu}}_{i.} + \hat{\hat{\mu}}_{..}$, where $\hat{\hat{\mu}} = \hat{\mu}^{\hat{\pi}}$ and $\hat{\pi}$ minimizes the unbiased estimate (3.5) corresponding to the vector γ with the components γ_{ij}.

5.4 A case study (gasoline mileage)

The objective of an experimental study was the comparison of the mileage reached by Fire-Hawks driven using gasoline types A, B and C (see Bowerman and O'Connell (1996)). The observed miles are given in the following table

type	A	B	C
	24	25,3	23,3
	25	26,5	24
	24,3	26,4	24,7
	25,5	27	
		27,6	
	$n_1 = 4$	$n_2 = 5$	$n_3 = 3$

We have a three-sample problem with replication sizes $n_1 = 4, n_2 = 5, n_3 = 3$, where the vector

$$\gamma' = (\gamma_1, \gamma_2, \gamma_3) = (\mu_A - \mu_B, \mu_A - \mu_C, \mu_B - \mu_C) \tag{5.31}$$

of differences of the mileage means under the gasoline types A, B, C is of interest. The calculation of the usual LSE yields

$$\breve{\gamma} = (-1.86, 0.7, 2.56), \tag{5.32}$$

while our adaptive procedure (described in 4.2) yields the estimate

$$\hat{\breve{\gamma}} = (-2.16, 0, 2.16). \tag{5.33}$$

A comparison of the MSE estimates (3.5) for the corresponding partitions $\breve{\pi}$ and $\hat{\pi}$ shows, that the partitions $\hat{\pi}$ leads to a reduction of 54.6% in the MSE(-estimates (3.5)):

$$\widehat{M}(\breve{\pi}) = 0.91422\ ,\ \widehat{M}(\hat{\pi}) \approx 0.41456. \tag{5.34}$$

6 Stepwise search procedure for a suboptimal estimator

The examples in section 4. have in common the very small number of elements in $\mathcal{X}$ and consequently of its partitions π. When the explanatory variables $X_{(1)}, \ldots, X_{(p)}$ have many possible values and (or) their number p is not small, the number $B(m)$ of possible partitions π of $\mathcal{X}$ may be so large (see (3.1)), that a comparison of the

MSE estimates $\widehat{M}(\pi)$ becomes computationally unfeasible. We remark , that the number $B(m)$ of different models in $\mathcal{M}$ (or of partitions in $\prod$) is

$$B(m) := \sum_{q=1}^{m} S(m,q) \quad , \quad S(m,q) := \sum_{j=1}^{q} (-1)^{q-j} j^m / j!(q-j)!, \tag{6.1}$$

where $S(m,q)$ is the number of q-th order partitions of $\mathcal{X}$ (see Stanley (1997)).
In such a case a stepwise procedure sequentially comparing partitions π with partitions in some sensible small "neighbourhood" $\sqcup(\pi)$ of π would replace the search for an optimal $\hat{\pi}$ by a search for hopefully suboptimal partitions.

Let for each $\pi \in \prod$ given a subset $\sqcup(\pi)$ of $\prod$, its elements being called "neighbours" of π. An example for a sensible definition leading to relatively few "neighbours" would be to call a partition π': $\mathcal{X} = \sum_{j \in I'} \mathcal{X}_i^{\pi'}$ a neighbour of $\pi : \mathcal{X} = \sum_{j \in I} \mathcal{X}_j^{\pi}$ if it is generated from π by the union of exactly two of the subsets $\mathcal{X}_j^{\pi}$, that is, if there are j_1, j_2 with

$$I = I' \cup \{j_2\}, \mathcal{X}_{j_1}^{\pi'} = \mathcal{X}_j^{\pi'} \ (j \in I' - \{j_1\}), \ \mathcal{X}_{j_1}^{\pi'} = \mathcal{X}_{j_1}^{\pi} \cup \mathcal{X}_{j_2}^{\pi}. \tag{6.2}$$

This would be the analog to downwards or backwards selection of regression models, while defining neighbours π' of π by partition of exactly one of the subsets $\mathcal{X}_j^{\pi}$ into two subsets $\mathcal{X}_{j_1}^{\pi'}, \mathcal{X}_{j_2}^{\pi'}$ would correspond to upwards or forwards regression. The parameter dimension in the linear models corresponding to neighbouring partitions π, π' of the above type obviously differ by one as in stepwise regression.

A special form of such a stepwise procedure is also given by the CART of Breiman et al. (1994) for the construction of regression trees which is oriented towards a large number p of explanatory variables. For this we may apply the corresponding software in S-plus (see Venables and Ripley (1997) or in XploRE (see Härdle et al. (1995)).

The search procedure starts with a fixed $\pi_0 \in \prod$, say $\pi_0 = \breve{\pi}$, if a downwards procedure based on (5.1) is chosen. The first step is to determine a π_1 with smallest $\widehat{M}(\pi)$ under the neighbours of π_0, the second step determines a π_2 with smallest $\widehat{M}(\pi)$ under the neighours of $\pi_1, \ldots$:

$$\widehat{M}(\pi_r) = \min\{\widehat{M}(\pi) | \pi \in \sqcup(\pi_{r-1})\}, \ r = 0, 1, \ldots \tag{6.3}$$

This minimization is simplified by the obvious fact, that the second term in (3.5) is the same for all neighbours of a fixed π. The procedure stops at the r-th step, when $\sqcup(\pi_r)$ constists only of π_r or when $\widehat{M}(\pi)$ has already been calculated for all partitions of $\prod$. In the downwards variant the procedure would stop at $\pi_{m-1} = \tilde{\pi}$,

while in the upwards variant it would start at $\pi_0 = \check{\pi}$ and stop at $\pi_{n-1} = \hat{\pi}$. The estimator γ^* corresponding to the partition $\pi^* = \pi_{r^*}$ with

$$\widehat{M}(\pi_{r^*}) = \min_r \widehat{\mathcal{M}}(\pi_r) \tag{6.4}$$

would qualify as an (adaptive) sensible alternative to the LSE $\check{\gamma}$ as an estimator of γ, although other estimators $\hat{\gamma}^{\pi_r}$ with $M(\pi_r)$ near to $M(\pi^*)$ could also be of interest, if the corresponding linear models $\mathcal{M}^{\pi_r}$ allow a sensible interpretation in the field of application. The stepwise procedure may also be repeated with alternative definitions of the neighbourhoods $\sqcup$, e.g. a downwards and an upwards procedure may both be performed.

7 A case study (steel quality)

The objective of an experimental study was the comparison of the quality of steel produced by 3 different types of rolling machines (see Hocking (1996)). It was also felt that there may be differences in the feedstock obtained from three different suppliers. Nine samples of feedstock were selected from each supplier and three samples were randomly assigned to each machine. The responses were ductibility as a measure of the quality of the product, given in table 3.

Table 3.

	machine 1	machine 2	machine 3
supplier 1	**1** 8,03 7,55 8,50	**4** 7,76 6,36 7,12	**7** 8,17 8,52 7,91
supplier 2	**2** 7,26 6,09 7.97	**5** 7,90 7,79 8,13	**8** 7,26 7,18 8,58
supplier 3	**3** 8,65 8,29 8,55	**6** 8,21 7,39 8,01	**9** 9,64 8,78 9,04

There may be interest in estimating the main effects of the machine type and of the

supplier as well as the interactions. In this case a good estimate of the vector μ of ductibility means μ_{ij} $((i,j) \in \mathcal{X} = \{(i,j)\} \,|\, i,j \in \{1,2,3\}\})$ corresponding to the nine different pairs of machine types and suppliers would be of interest, all other estimates could be derived from the estimate of μ (see 4.3). The use of a weight matrix $W = I$ seems to be sensible. The calculation of the means gives the usual LSE.

$$\breve{\mu} := ((\breve{\mu}_{ij})) = \begin{pmatrix} \begin{pmatrix} 8.02666 \\ 7.09333 \\ 8.49666 \\ 7.08000 \\ 7.94000 \\ 7.87000 \\ 8.20000 \\ 7.67333 \\ 9.15333 \end{pmatrix} \end{pmatrix}. \tag{7.1}$$

The number of partitions of $\mathcal{X}$ is already relatively large, so that in order to obtain a quick and easy improvement of the LSE $\breve{\mu}$ we apply the upwards stepwise procedure described in section 5. The adaptive estimate $\hat{\hat{\mu}}^*$ is given by

$$\hat{\hat{\mu}}^* = \begin{pmatrix} \begin{pmatrix} 8.02666 \\ 7.08666 \\ 8.34833 \\ 7.08666 \\ 7.82777 \\ 7.82777 \\ 8.34833 \\ 7.82777 \\ 9.15333 \end{pmatrix} \end{pmatrix} \tag{7.2}$$

and it corresponds to the partition

$$\pi^* = \{1\} + \{9\} + \{2,4\} + \{3,7\} + \{5,6,8\} \tag{7.3}$$

where we use the numbering of the pairs (i,j) given by the numbers presented in the upper right corner of the cells in table 3. The reduction in the MSE estimate (3.5) reached by the estimator $\hat{\mu}^{\hat{\pi}}$ in comparison with the LSE $\breve{\mu} = \hat{\mu}^{\breve{\pi}}$ is of $80,1\%$:

$$\widehat{M}(\breve{\pi}) = 9\,,\ \widehat{M}(\hat{\pi}) = 1,79. \tag{7.4}$$

But in this case study it happens by chance, that the values of the adaptive estimate $\hat{\hat{\mu}}$ and the mean $\breve{\mu}$ are very near. This is due to our data, because in each subset of the partition (6.3) the observation means that correspond to the pairs belonging to the subset are almost identical. The values of $\hat{\hat{\mu}}$ and $\breve{\mu}$ would be essentially different

(as in our case study 4.4), if there would be larger differences between the means corresponding to a subset of the partition.
The overall MSE $E\|\hat{\hat{\mu}} - \mu\|^2$ of the adaptive estimator $\hat{\hat{\mu}}$, which considers also the data dependence of the selected partition $\hat{\pi}$, may in principle be calculated by a Monte-Carlo approximation (e.g. under $\mu = 0$ or $\mu = \breve{\mu}$ and other tentative parameter values and under $\sigma^2 = \hat{\sigma}^2$). As easely seen, it must be larger than $\widehat{M}(\hat{\pi}) = 1,79$.

Acknowledgment

The authors are grateful to B. Droge for his valuable remarks leading to an improvement of the paper.

References:

Bowerman, B.L. and O'Connell, R.T. (1990). *Linear Statistical Models. An Applied Approach.* PWS-Kent, Publishing Co.

Breiman, L., Friedman, J.H., Olshen, R. and Stone, C.J. (1994). *Classification and Regression Trees*, Wadsworth, Belmont.

Bunke, O. (1992). Semiparametric modelling and prediction for a variable depending on time and explanatory variables. In: *Statistical Modelling* (eds. P.G.M. van der Heijden et al.), Elsevier, Amsterdam, 11 - 25.

Bunke, O. and Droge, B. (1984). Bootstrap and cross-validation estimates of the prediction error for linear regression models. *Ann. Statist.* **12**, 1400 - 1424.

Bunke, O., Droge, B. and Polzehl, J. (1995). Model selection, transformations and variance estimation in nonlinear regression. Discussion Paper No. 52, SFB 373, Humboldt University, Berlin.

Bunke, O. and Ilouga, P. (1998). Higher order asymptotic comparison of adaptive linear smoothers in regression models. (Discussion paper, SFB 373, in preparation).

Casella, G. and Berger, R.L. (1990). *Statistical Inference.* Wadsworth, Belmont, Calif.

Droge, B. and Georg, T. (1995). On selecting the smoothing parameter of least squares regression estimates using the minimax regret approach. *Statistics & Decisions* **13**, 1 - 20.

Green, P., Jennisen C. and Scheult, A. (1985). Analysis of field experiments by least squares smoothing. *J. Roy. Statist. Soc. Ser. B.* **47**, 299 - 315.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Cambridge

Härdle, W., Klinke, S. and Turlach, R.A. (1995). *XploRE: An Interactive Statistical Computing Environment*, Springer, New York 1995.

Hocking, R.R. (1990). *Methods and Applications of Linear Models.* Wiley, New York.

Humak, K.M.S. (1977), *Statistische Methoden der Modellbildung*, Band I. Akademie-Verlag, Berlin.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661 - 675.

Müller, M. (1993). Asymptotische Eigenschaften von Modellwahlverfahren in der Regressionsanalyse. Doctoral Thesis. Humboldt University, Berlin (in german).

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* 12, 758 - 765.

Stanley, R.P. (1997). *Enumerative Combinatorics.* Cambridge University Press, Cambridge.

Venables, W.N. and Ripley, B. (1997). *Modern Applied Statistics with S-Plus.* Springer, New York.