ELSEVIER

Food Quality and Preference 11 (2000) 17–26

Food Quality and Preference

www.elsevier.com/locate/foodqual

Aspects of statistical regression in sensometrics☆

Rolf Sundberg*

Mathematical statistics, Stockholm University, S-106 91 Stockholm, Sweden

Accepted 9 May 1999

Abstract

A discussion is presented of roles of regression analysis in sensometric studies, distinguishing description, interpretation and prediction purposes. A brief review is given of linear regression methods for prediction in situations with near-collinear explanatory variables, including for example ridge regression and partial least squares, and latent variable models are discussed. Finally problems with statistical and causal inference from regression on covariates in designed experiments are discussed. Illustrations in the paper are based on a sensometric study of apple flavour under varied storage conditions [Brockhoff, P., Skovgaard, I., Poll, L., & Hansen, K. (1993). A comparison of methods for linear prediction of apple flavor from gas chromatographic measurements. *Food Quality and Preference*, *4*, 215–222], with sensory response data and with gas chromatography measurements as covariates.

Keywords: Causality; Covariates; Designed experiments; Errors in variables; Multiple regression; Near-collinearity; Prediction; PCR; PLS; Ridge regression

1. Introduction

In this paper aspects are discussed of sensometric studies in the form of more or less explicitly designed experiments involving auxiliary physical or chemical measurements, 'covariates', which might be used as explanatory variables in a regression or covariance analysis. A review is given of modern linear regression methodology, with discussion of how the information in the explanatory variables should be judged according to what is wanted from the study, and in particular the problems with covariates in the interpretation of designed experiments.

Typically a sensometric study involves several sensory response variables, concerning texture, taste, smell, etc., but different responses are nevertheless treated individually. Multivariate aproaches are possible, with the idea of borrowing strength from one response variable to another. However, the potential of a multivariate approach is unclear. Also, this aspect was the topic of a talk by Sijmen de Jong at the same meeting, so here the discussion will concentrate on aspects concerning individual responses. Also problems with missing data and nonlinearities will not be discussed here, even though they may be important concerns in a statistical analysis of sensory data.

2. Two sensometric experiments

A first simple example of the type of situations to have in mind in this paper is a study by Hough et al. (1996) of relationships between sensory descriptors for cheese and physical or chemical descriptors. Sensory texture descriptors were related to a set of four instrumental compression tests variables (plus moisture). Aroma and flavour descriptors were related to concentrations of a set of nine organic acids. A sample of 23 cheeses were analysed, cheeses taken at different stages of ripening. This was the implicit design of the study, not used for a study of ripening per se, but apparently primarily used as a guarantee for substantial variation. Nine trained assessors gave sensory profiles. Averages over assessors were used. Sensory and instrumental descriptors were "correlated" by the use of multivariate PLS.

The second situation is taken from Brockhoff, Skovgaard, Poll and Hansen (1993), where it is described in

☆ Paper presented at the 4th Sensometrics Meeting, Copenhagen 6–8 August 1998.

* *E-mail address:* rolfs@matematik.su.se (R. Sundberg).

0950-3293/99/$ - see front matter
PII: S0950-3293(99)00048-8

more detail, and where its data are analysed from the predictive point of view. These data will be used to illustrate features of the following discussion. It must be stressed that the present study does not invalidate the one previously published, but serves to supply some additional aspects. This example involves an imbedded factorial experiment. Apples were given four different treatments (storage conditions, specified by oxygen/nitrogen ratio), during a shorter or longer period (109–190 days). The apples were then subject to evaluation 6 times during a post-storage ripening period. This constitutes a complete factorial experiment with $2\times4\times6=48$ design points. The smell of the apples was evaluated by a trained sensory panel of 8–10 assessors, who gave scores on a 0–5 point scale (in half units) for preference, intensity, banana flavour, etc. Also, gas chromatographic (GC) measurements were made of a number of 15 volatiles (different acetates, propanoates etc), so that to each design point was attached a set of such GC-data. An explicit purpose of the study was to investigate the predictive ability of these volatiles.

3. Regression methodology, a review and discussion

3.1. Aims of regression studies

A standard ingredient in a statistical analysis of situations like those described above will be a regression analysis, where the original sensory response variables, or some functions of them (as for example the residuals from an ANOVA model fit in the apples example) are regressed on the instrumental data. The reason is not only that the primary aim is prediction of sensory properties from instrumental readings, but also because directly concentration-related instrument measurements are considered causal to sensory variables, at least more than the other way round, and because random errors in sensory variables are thought to be larger than in the chemical variables, cf. the discussion around Fig. 4(a)–(c) below. Assuming that the variation between assessors has nothing to do with variation in instrumental values, i.e. that the difference between any two assessors is uncorrelated with the instrumental variables, response values from now on will be taken to be the averages over assessors, to be denoted $y(n\times1)$. This was also done in the examples referred to above. A complication that will not be addressed is that all assessors had not evaluated all apple samples. The matrix of potential explanatory instrumental variables will be denoted $X(n\times p)$. For convenience, the x-variables are assumed centered. This means that the intercept need not be much discussed in the following. The apples data will be particularly considered, with preference as y-variable (48×1) and the GC data as X (48×15). Preference was one of the responses found by Brockhoff et al. to be influenced by the experimental factors. However, the reader should temporarily forget about the imbedded designed experiment, until it will be returned to in the next section.

In the apples example all the GC x-variables are substantially positively correlated with the preference response y. Fig. 1 shows the individual regressions of y on six of the x-variables, including those who show the highest variation per se (nos. 4, 5 and 10).

It is seen from Fig. 1 that some of these individual relationships deviate clearly from linearity. This could be expected from the fact that response data are sensory data. However, because the x-variables are correlated, individual x-variables may show a non-linear relation with y at the same time as a suitable linear combination of the same x-variables fit the response linearly. So, not only are non-linearities outside the scope of this review; linear modelling cannot be rejected just because of individual non-linearities. An extreme example with spectroscopic x-data is discussed in Sundberg (1999).

All 15 correlation coefficients are between 0.40 and 0.75. In fact, the variation in correlation is not even higher than expected from only sample variation in normal samples of size $n=48$, under the assumption of a common underlying value (of about 0.6). These correlations show that one cannot immediately reject any x-variables as irrelevant; rather one is led to use all of them jointly for inference about the relation between the x-variables and preference y. Then an important feature is that the x-variables are also (necessarily) mutually strongly correlated themselves. The empirical correlations in x fall between 0.23 and 0.99. A principal components analysis (PCA) on their covariance matrix shows that the first two principal components together explain nearly 96% of this variability, see Fig. 2 for a combined scores and loadings plot. This means that a multitude of other linear forms in x, with coefficients orthogonal to the first two PC directions, will be almost constant. There is near-collinearity in X. Whether this collinearity is a problem or not depends on what is wanted from the data:

- a descriptive relationship?
- an understanding or interpretation of a possible causal relationship?
- a relation to be used for prediction?

The best fit of a linear regression function $y=\alpha+\beta'x$ is obtained by classical ordinary least squares, OLS (in chemometrics sometimes misleadingly called multiple linear regression, MLR, which is rather the statistical model for data). In a descriptive sense OLS regression on the apple data will explain practically all variations in y by the 15 variables in X. However, when the explanatory variables are near-collinear, the OLS regression

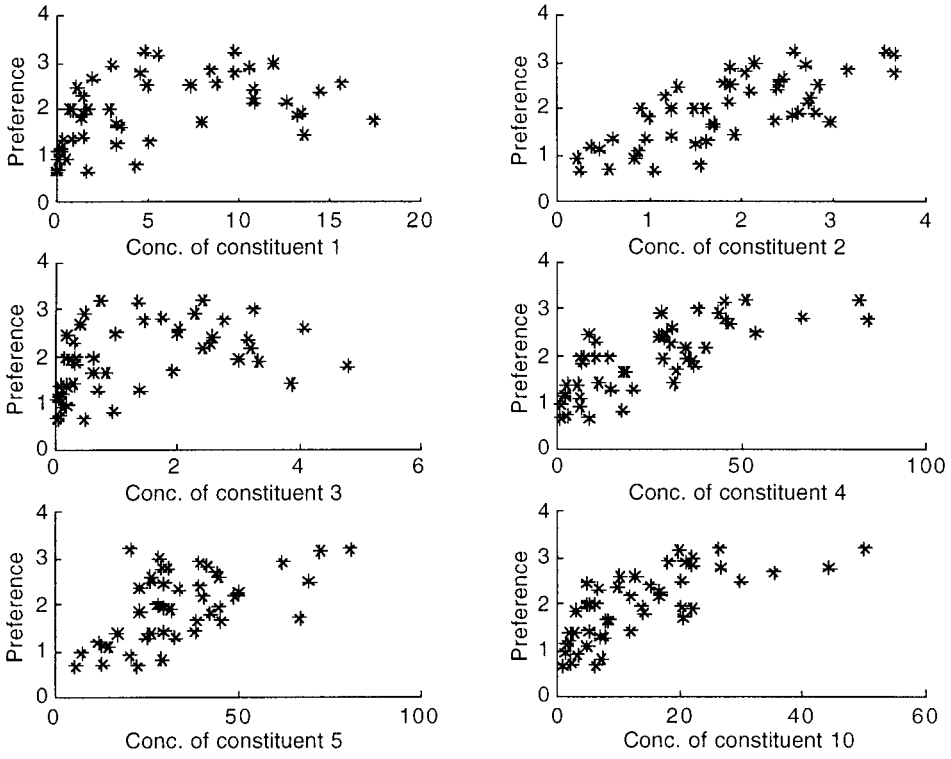

Fig. 1. Plots of preference against concentrations for 6 GC components. Note the different scales on the *x*-axis.

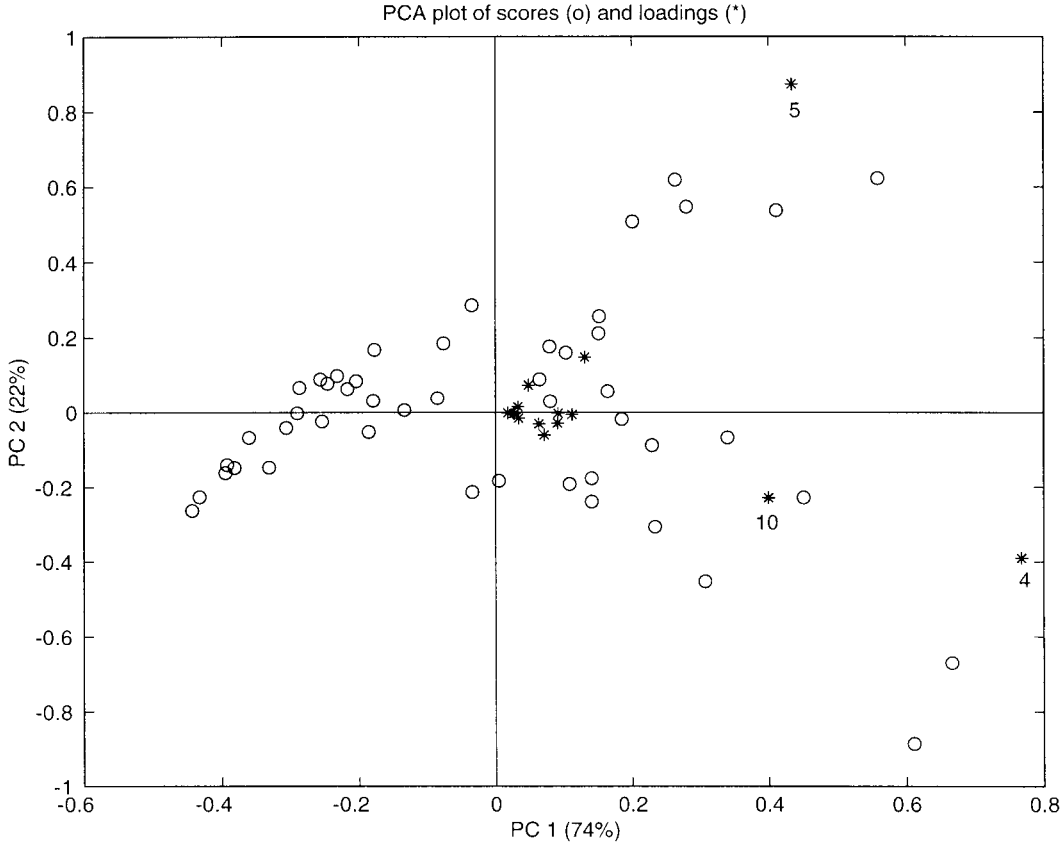

Fig. 2. First two principal components of GC covariance matrix. Asterisks are loadings and circles are proportional to scores. The three variables with high loadings are labelled by numbers (4, 5 and 10).

coefficient estimator $\hat{\beta}_{\mathrm{OLS}}$ is likely to be misleading. In a standard regression model its uncertainty, as given by its variance–covariance matrix, is $\sigma^2(X'X)^{-1}$. Near-collinearity in X implies that $X'X$ is near-singular, and its inverse will have numerically very large elements. This also means that $\hat{\beta}_{\mathrm{OLS}}$ will tend by randomness to have unrealistic, numerically very large values. Hence, the coefficients of the best description of y may be very far from the coefficients of the true relationship, assuming there is one. Also for prediction, OLS is likely to be inefficient due to this overfitting. The feature is seen in the apples data, where the 15 OLS coefficients cover a wide interval, from about -1 for x-variables 6 and 7 to $+1.5$ for variable 12. The high weight on several x-variables, which are of little variation per se, and the several high negative coefficients in spite of the clear positive regression of y on any single x (cf. Fig. 1) are indications that the OLS method overfits. From the predictive point of view this will also be clear from comparisons of model fit R^2 with cross-validatory prediction results. But how could one better estimate how y will change with x in such little informative directions? Perhaps one should not try?

OLS yields the best fit, but many other linear functions in x fit y almost equally well. For future predictions it is wise not to use the OLS estimates of the relationship in little informative directions, because they are likely to overreact. Instead one could shrink the OLS relationship in these directions (e.g. RR) or even assume that y does not depend on x in such directions (PCR, PLS). For brief descriptions of these relatively well-known methods, see formula box. For more details the reader is referred to the books by Brown (1993) or Martens and Næs (1989). These methods "trade bias for variance" by replacing the unbiased OLS estimator by a (slightly) biased estimator of higher precision.

An alternative would be to exclude x-variables so that the remaining ones are no longer near-collinear, for example, by a stepwise regression procedure. Then, in this example, almost all of them would have to be excluded, however, and that would probably imply a loss of precision. On the other hand it might be economical for future predictions not having to measure all x-variables.

Anyhow, for the second type of question one must live with some intrinsic uncertainty from the design of the study: The given 48 observations of the 15 x-variables are not able to answer in a unique way which x-variables influence y, and in what proportions. The situation would not even change essentially if many more x-variables were available than the number of observations. The OLS fit would be perfect, but there would be an intrinsic lack of uniqueness. However, one could still hope for good predictors through ridge regression or the latent factor methods PLS and PCR.

Formula box
Conventional regression procedures
Linear regression model:

$$y = \alpha + \beta' x + \text{noise} = \alpha + \Sigma\beta_j x_j + \varepsilon,$$
$$\bar{x} = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2.$$

Intercept estimate: $\hat{\alpha} = \bar{y}$.

Ordinary least squares (OLS):

$$\hat{\beta}_{\mathrm{OLS}} = (X'X)^{-1}X'y$$

Ridge regression (RR):

$$\hat{\beta}_{\mathrm{RR}}(\delta) = (X'X + \delta I)^{-1}X'y.$$

where δ is desired 'ridge constant', and I the identity matrix of size p.

Partial least squares (PLS):
form regressors ('factors') successively.
First factor $t_1 = c_1'x$, $|c_1| = 1$, to maximize $\mathrm{Cov}(t_1, y)$.
Regress y on t_1,

$$\hat{\beta}_{\mathrm{PLS}}(1) = (t_1't_1)^{-1}t_1'y.$$

Repeat on y-residuals to form t_2 and regress y on (t_1, t_2), to form $\hat{\beta}_{\mathrm{PLS}}(2)$, etc.

Principal components regression (PCR):
like PLS but replace $\mathrm{Cov}(t, y)$ by $\mathrm{Var}(t)$.
This is equivalent to:
regress y on major PCs t from a PCA.

3.2. Latent factor models

To some extent a latent factor model helps interpretation and understanding. If a PCA shows that the variation in x lies within a small-dimensional subspace, represented by a few PCs, one may argue that the relationship between y and x stays within these PCs. In chemometrics this latent dimension is often called the chemical rank of the system. The conceptual problem whether there is a well-defined such rank will be left aside, and the practical problem to decide how large it is. Anyhow, note that such an identification does not necessarily require many more observations than the chemical rank number itself, and in particular it may be sufficient with fewer observations than variables.

A latent factor model motivates PCR and PLS regression. PCR is regression on the first few principal components of x as regressors. By definition they maximize variance with respect to x. PLS is based on a principle of maximization of covariance between x and y, and will therefore tend to yield a lower rank by

down-weighting potential regressors which correlate little with y. In the apples example all x-variables correlate about equally with y. This implies that first factor PLS and first factor PCR should differ little here, cf. Fig. 3 below.

A latent factor model assumes that a latent variable or latent vector, t say, is randomly generated from some population, and that x and y are linearly dependent on t (+noise). For interpretation and understanding, this population for t is crucial, and not less so for prediction. Only if new observations can be regarded as generated from the same population, the predictions can be trusted. This restriction is obvious if the latent dimension is estimated by leave-one-out cross-validation, since the prediction tests are made within the same original set of observations, but the restriction is also crucial with an external test set. The quality of the predictions is doubtful as soon as one goes outside the conditions of the test set. It is important that the test set, in all essential ways (whatever they are) mimics the future. If the test set is not representative enough, the predictor will overfit. Note that this is a risk with any predictor construction method. Diagnostics for checking that a new x falls within the data of the calibration have an important role. These are easily formed in latent factor models.

Here a warning might be appropriate concerning predictor validation in factorial and similar designed experiments by latent factor models and by leave-one-out and similar forms of internal cross-validation. The observations in a designed experiment are not randomly generated from a natural population. Leave-one-out is particularly questionable if one of a set of replicates is left out, or if one factor has no real effect, which implies that this factor will serve to generate replicates. The remaining replicates are likely to take care of the fit such that the one left out will always be well predicted.

Suppose now that a latent factor model is generating the data. Some proponents of PLS seem to think this means the regression model is necessarily invalid, and that PLS, therefore, should yield better predictors than those motivated from the regression model (see e.g. contributions by S. Wold to the discussions of Frank & Friedman, 1993, and of Sundberg, 1999). However, say that the distribution for the latent factor t is (as usual) assumed normal, and the noise in x and y as well, given t. Then x and y are also normally distributed, jointly, and as a mathematical consequence the best unbiased predictor of y is the theoretical linear regression of y on x. Hence the linear regression model is adequate for the prediction problem even with an underlying latent structure, although this best predictor is unknown. In a comparison of different predictor constructions it is not obvious which will be the best one. For example, the regression model apparently involves fewer parameters in the linear structure than a typical latent factor model in its loadings matrices. In fact, the most remarkable feature is the close relationship between all these methods, and their approximate equivalence in typical

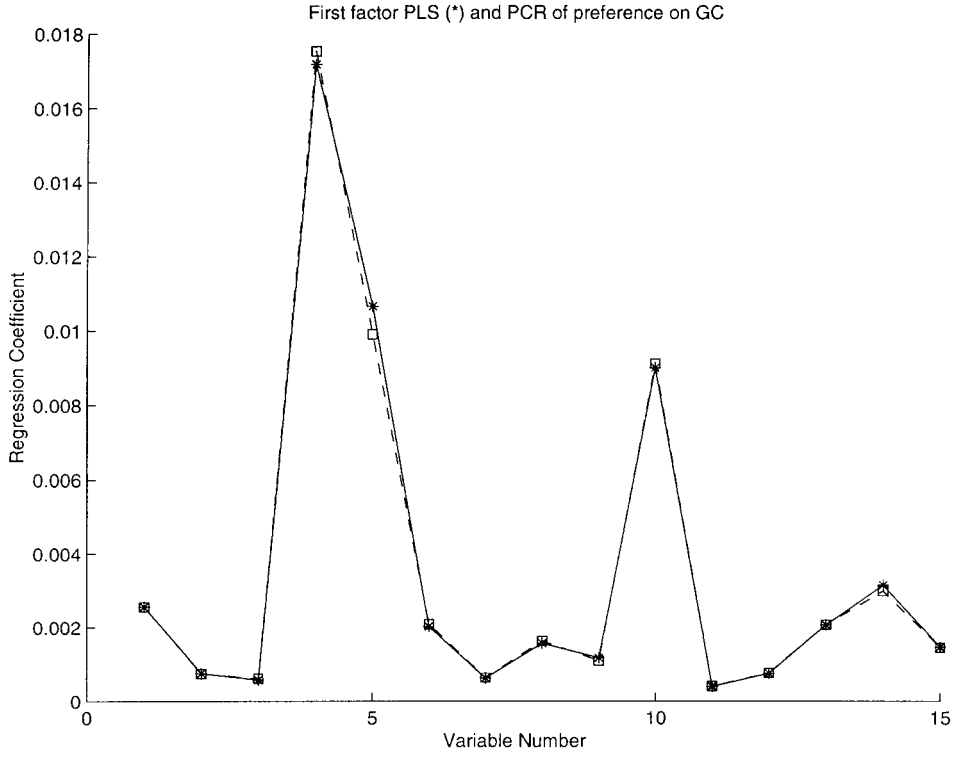

Fig. 3. Regression coefficients of preference on GC variables by first factor PLS (∗) and first factor PCR (square).

cases (cf. below). The choice whether or not to scale x to unit variances will usually be more important than the choice between PLS, PCR and ridge regression. For the apples data, as an example, variance standardization of the GC variables gave a less good fit for the same number of PLS or PCR factors.

3.3. Relationships between regression procedures

Before this discussion is continued, a brief review will be given of how the above mentioned methods for multiple linear regression are related, by introducing a more general framework for regression method construction.

Suppose one wants to select one or several regressors, in the form of linear combinations $t = c'x$ of the original x-variables. High correlation with y would of course be desirable. If the only criterion is to maximize the (squared) correlation $R^2(t, y)$ with y, one is led to the OLS regression; the regressor t will be proportional to the OLS fitted relationship. But it has also already been remarked above that the uncertainty about the regression is high in directions where x has little variation. Therefore it is also desirable that the regressor $t = c'x$ to be selected shows a large variation in the data, i.e. a criterion that $\mathrm{Var}(t) = c'\mathrm{Var}(x)c$ should be high.

Remark. *To impose this criterion, some length restriction on c must also be inserted, usually chosen as* $|c| = 1$ *in the Euclidean metric. It is unavoidable that this implies a lack of scaling invariance, if different x-components are rescaled differently. In other words, which regressors will show large or little variation, and what is near or less near collinearity, depends on the metric chosen.*

Now suppose a criterion is formulated, involving both $R^2(t, y)$ and $V = \mathrm{Var}(t)$, in the form "maximize a function $g(R^2, V)$ under the restriction $|c| = 1$". An example is the product of R^2 and V, which is equivalent with the covariance squared, and which is maximized by PLS, cf. the formula box. From the function g it is only required that it is increasing in each of its arguments separately, a very weak and natural demand. Then, as shown in Björkström and Sundberg (1999) it follows, remarkably, that the regressor t must be proportional to a ridge regression type estimator for some δ, more precisely $t = c'_{\mathrm{RR}}x$ with

$$C_{\mathrm{RR}} \propto (X'X + \delta I_p)^{-1}X'y. \tag{1}$$

Typical δ-values in traditional ridge regression (RR) are small positive, whereas when g is the product R^2V, (first factor) PLS is obtained in the limit as δ tends to $+\infty$ or $-\infty$. On the negative axis all δ-values may be allowed between $-\infty$ and minus the largest eigenvalue of $X'X$ (PCR via the first PC). This representation includes Stone and Brooks' (1990) continuum regression (CR), with explicit criterion function

$$g(R^2, V) = R^2V^{\gamma}, \tag{2}$$

where they let γ run through the continuum from 0 (OLS) via 1 (PLS) to $+\infty$ (PCR). As a rule, but a rule with its exceptions, there is a (non-explicit) one-to-one correspondence between δ in (1) and γ in (2).

Now, when a first regressor t has been constructed, y can be regressed on t, by simple least squares. This is so in OLS, PLS, PCR and CR, but not in conventional RR. Therefore Björkström and Sundberg (1999) advocate least squares ridge regression (LSRR) to replace conventional RR, that is to use c_{RR} in (1) above not as an estimate of β but to form a regressor $t = c'_{\mathrm{RR}}x$, as described above. The difference is only a numerical factor, typically close to 1 (for small δ).

Usually PLS and PCR will require more than this first regressor. After regression of y on the first t, the residuals from this regression can replace y for a second step, to construct a second regressor (second CR- or PLS-factor or PC) to be used jointly with the first one. Since LS residuals are orthogonal to the regressors, the new c-vector will be orthogonal to the previous ones. How many steps should be gone through, that is how many regressors (factors) should be constructed and used? If the purpose is description, one makes sure one is close enough to the maximum R^2 (attained for $\delta = 0$). If the purpose is prediction of new but similar observations, cross-validation leave-one-out or some similar validation procedure can be used. Else the choice is more delicate.

One might think there need not be any particular similarity between LSRR regression (with only one regressor but a selected δ) and PLS or PCR with several factors. However, experience indicates that typically all three methods tend to give quite similar regressions. More precisely, by the best choices of the parameter δ for LSRR and of the number of factors for PLS and for PCR, respectively, they will all yield about the same optimal predictors.

With the apples data it was seen that there is a single major source of variation common to both preference y and the GC-variables, namely one of the experimental factors (storage condition: oxygen/nitrogen ratio). Neither the PLS or PCR predictors will benefit from more than the first latent factor, and this factor will then be almost identical for PLS and PCR, and represent the experimental factor. The optimality of a single factor was established by cross-validation. The corresponding PLS and PCR regressions are almost indistinguishable, as illustrated in Fig. 3. More generally interpreted, the LSRR regression is quite insensitive to the choice of its parameter δ over a wide region of δ-values, including the PLS and PCR special cases.

4. Covariates in designed experiments

In this section a discussion will be found of the role of covariates (the GC-values) in a designed experiment, where the main purpose is to find out about the influence of the factors (storage conditions etc.) of the experiment. In a designed experiment with covariates, analysis of covariance (ANCOVA) is a statistical technique used to obtain increased precision in the estimated treatment effects. This is done by so called covariate adjustment of the treatment means, and it is based on a combination of an ANOVA type structure in the experimental factors and a linear regression on the covariates. For a situation where covariance analysis would be legitimate, imagine a modified apples example with the GC measurements made on the apples before storage. A regression on these GC values could increase the precision by eliminating variability between apples before storage, if this source of variability still affects the response values after storage.

In the apples experiment as it was really carried out, a covariance analysis would not be legitimate, however. For covariance analysis it is crucial that the covariates be unaffected by treatment, whereas in the apples experiment the covariate values are measured after treatment and intrinsically connected with the factor levels of the experiment. If storage conditions are systematically changed there is reason to expect changes not only in smell- and taste-related sensory variables (y) but also in the concentrations of the volatiles measured by GC (x). A detailed discussion of the postulates behind covariance analysis and the problems when these postulates fail were given more than 40 years ago (Cochran, 1957; Smith, 1957). Not that I have seen misuse in sensometrics, but a warning might nevertheless be motivated, because misuse can be found even in examples from statistics texts of relatively recent date.

For the GC data of the apples example, one can feel quite convinced that the covariation between x and y is more or less caused by the experimental factors. But what is the implication of the underlying experimental design for the interpretation of this covariation? By help of the x-data the variation in y can evidently be described more completely. How well will one be able to predict from x? Are the reasons for variation in y better understood? Does the variation in x completely describe the variation in y due to treatment, or is there variation in y caused by the experimental factors that does not go through x? Is there variation in y caused by variation in x that is not related to the experimental factors? Questions like these will be discussed, and whether they can be answered or not. Three different models for the variation will be discussed and compared:

(a) a correlation model for (x, y);
(b) a simple causal model for the influence of x on y;
(c) a more saturated model for causal influences.

The weakest (near absence of) structure just states that the experimental factors influence x and y, so that as a consequence x and y are seen to be mutually correlated. This correlation will imply a regression of y on x, but without any causal interpretations of this regression within the model. This is depicted in Fig. 4(a).

At the other end in causality is the structure stating that the experimental factors influence x, and x influences y, so that all influence on y of the experimental factors goes through x. A simple such structure is indicated in Fig. 4(b).

The structure of Fig. 4(b) can be represented by the following statistical model:

$$x = \tau_x + \gamma,$$
$$y = \alpha + \beta x + \varepsilon, \tag{3}$$

where τ_x denotes the direct treatment effect on x, the letter γ represents the effects on x from other causes of variation, and ε denotes experimental and measurement error variation in y (unrelated with treatments and with x). The impact of x on y is assumed linear with coefficient β. Since also in model (a), y has a regression on x, models (a) and (b) are not distinguishable without further information. The feature of model (b) that

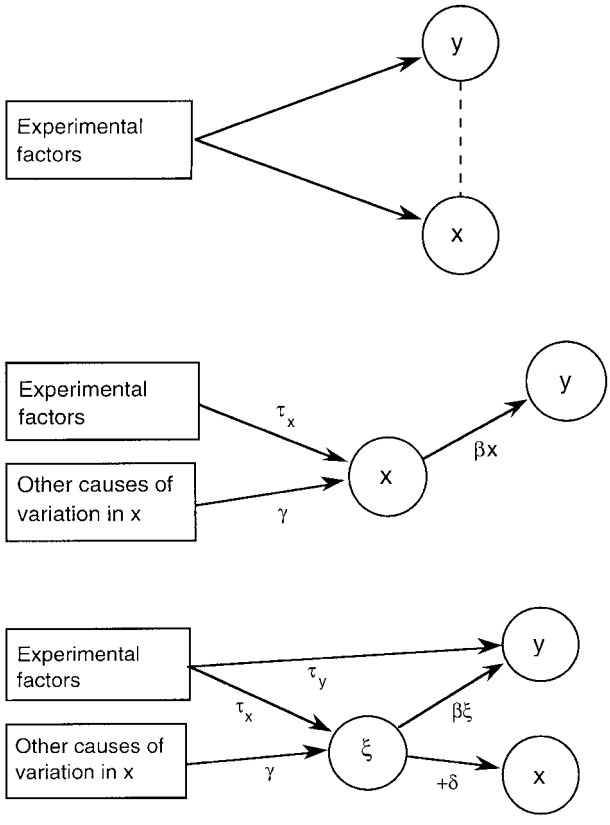

Fig. 4. (a). Correlation structure; (b). A simple causal structure; (c). An extended causal model.

makes it distinguishable from the correlation structure of Fig. 4(a) is that not only the experimental variation τ_x but also other variation in x influences y in the same way (by the same coefficient vector β). If there is enough variation in x not caused by the systematic experimental variation, one can hope to see this feature after elimination of the treatment effects. For example, by forming residuals in x and y, from an ANOVA treatment structure model, new data are obtained, which should be unaffected by treatments if the treatment structure was correctly modelled. Safer than residuals are differences between replicates in (x, y), since these do not rely on any treatment structure model.

Fig. 4(c) generalizes structure (b) in two ways and makes it more realistic, by allowing a direct part of the influence of the experimental factors on y, i.e. some treatment effect τ_y on y that is not explainable through x, and by allowing measurement errors in x, which should of course not have influence on y.

The statistical model representation extending (3) above is

$$\begin{aligned} \xi &= \tau_x + \gamma, \\ x &= \xi + \delta, \\ y &= \alpha + \beta'\xi + \tau_y + \varepsilon \end{aligned} \tag{4}$$

where the treatment effect term τ_y on y is added, and ξ is introduced as an unobserved idealized version of x, so that x is ξ perturbed by the measurement error δ. Note that the influence on y goes through the ξ part of x only. Can these structure elements be distinguished in data?

First inference concerning the possible presence of substantial measurement errors in x will be discussed. Even in the absence of a direct treatment effect τ_y on y, the model that relates y with x is no longer a regression model when measurement errors δ are allowed in x, but a so called errors in variables model. If a regression of y on x is then fitted, it will be more or less shrunk relative to the true, error-free β. If replicates of (x, y) are available there is a chance to see different degrees of such shrinkage, freed from treatment effects. For a difference in (x, y) between two replicates, all treatment effects τ_x and τ_y would vanish from (3) and from (4). According to model (3) there would still be a regression with the same regression coefficients as before differencing. However, with measurement errors δ allowed in x, according to (4), these differences in y and x would yield a shrunk regression, because the influence of the measurement errors in x would be much higher in the differences. Hence, in the presence of replicates this shrinkage effect might possibly be seen, even though the reduced variation in x by going over to differences between replicates will also imply considerably increased statistical uncertainty.

In the apples example there are no true replicates. However, the storage length factor on two levels seemed to have little effect, and might be tried as a substitute for a true replicate factor. Regressions of y differences on x differences, and y residuals on x residuals in an ANOVA model allowing two-factor interactions (that is residuals equal three-factor interactions), gave regression coefficient vectors as illustrated in Fig. 5. Conclusions are shaky, because the "true" β-vectors of the models above are not available, but only the first factor PLS estimates of β. Under this reservation, note that differences and residuals yield slightly lower β-estimates than the original data. This might be interpreted to say that the x-variables have measurement errors, whose influence is amplified in the differences and residuals when most of the underlying variation, that there is in ξ, has been eliminated.

Consider now how the possible presence could be investigated of a treatment effect τ_y directly on y, as allowed in model (4). Regressing y on x would leave residuals without remaining treatment effects according to model (b), whereas if such treatment effects are seen, this is an indication that model (c) is the right one, with a non-zero effect τ_y. In the apples example a 3-way ANOVA on the residuals from a first factor PLS regression (for regression coefficients, see Figs. 3 and 5, asterisks and solid lines) shows no indication of any treatment effects, as judged from a comparison of main effects and 2-factor interactions with 3-factor interaction, or main effects with all interactions. In other words, there is no indication of any treatment effect τ_y directly on y in the apples example.

The final discussion will concern the implications for prediction of y by its linear regression on the vector x. Prediction does not require causal relationships, so all models (a), (b) and (c) will allow this prediction, provided as usual that the prediction situation is the same as the calibration situation. The latter assumption can be crucial when the variation in the calibration data is generated in a designed experiment, like with the apples data, or by some other deliberate variation of circumstances, as with the cheese data. This is so, even in the causal model versions (b) and (c), because in the experiment performed one is not likely to see variation in more than a few dominating directions of the multidimensional x. By construction, the predictor in the apples example is good at predicting the effect of varied storage conditions, but that does not at all imply that the relation between y and x must be the same if one later wants to predict differences between apple sorts or between shiploads, for example. Analogously, in the cheese example the variation in data was generated by different ripening times, which does not imply that the relationship would well represent variation between different producers at the same ripening stage, if that should be desired. The coefficient vector β can be not only of wrong magnitude, but also of the wrong direction. If one wants to predict a special type of variation,

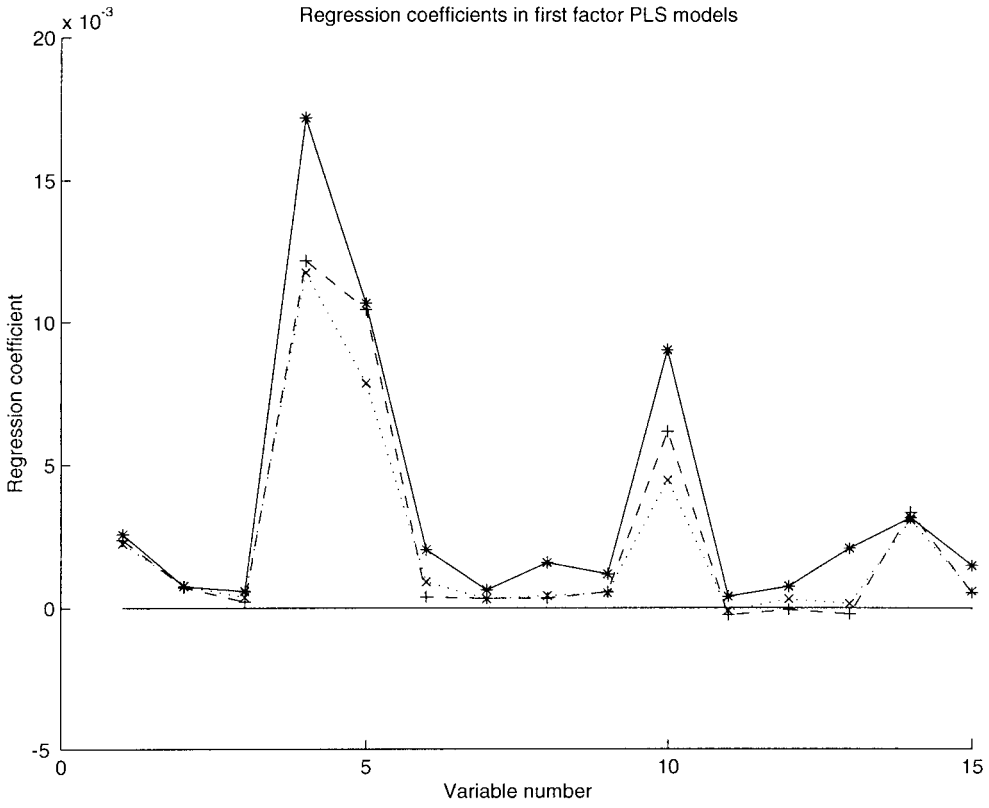

Fig. 5. Regression coefficients by first factor PLS for different sets of data: ——∗——: preference on GC; – – –+– – –: preference pairwise differences on GC pairwise differences (over two-level factor); ···×···: preference residuals on GC residuals according to ANOVA models allowing pairwise interactions.

this source of variation must be well represented in the calibration and it is risky to replace it by some other, more easily generated variation.

5. Conclusions

This paper has discussed several problems and features connected with explanatory variables in sensometrics, in particular as covariates in connection with designed experiments. Near-collinearities among the explanatory variables is a problem for understanding causal relationships, but often not for prediction purposes, when they rather support each other. Latent variable methods (PCR, PLS) are popular for prediction and the paper has discussed their theoretical basis and relationships with other prediction methods (ridge regression and least squares ridge regression). Causal models with covariates in designed experiments have also been discussed, and the possibilty to draw statistical inference about them. The final discussion concerned the dangers in forming predictors from designed experiments, when they are based on systematically generated variation.

Acknowledgements

The author is grateful to Per Brockhoff for making the apple data available, and to the Swedish Natural Science Research Council for financial support.

References

Björkström, A., & Sundberg, R. (1999). A generalized view on continuum regression. *Scandinavian Journal of Statistics*, *26*, 17–30.

Brockhoff, P., Skovgaard, I., Poll, L., & Hansen, K. (1993). A comparison of methods for linear prediction of apple flavour from gas chromatographic measurements. *Food Quality and Preference, 4*, 215–222.

Brown, P. J. (1993). *Measurement, regression, and calibration*. Oxford: Oxford University Press.

Cochran, W. G. (1957). Analysis of covariance: its nature and uses. *Biometrics, 13*, 261–281.

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

Hough, G., Califano, A. N., Bertola, N. C., Bevilacqua, A. E., Martinez, E., Vega, M. J., & Zaritzky, N. E. (1996). Partial least squares correlations between sensory and instrumental measurements of flavor and texture for reggianito grating cheese. *Food Quality and Preference, 7*, 47–53.

Martens, H., & Næs, T. (1989). *Multivariate calibration*. Chichester: John Wiley.

Smith, H. F. (1957). Interpretation of adjusted treatment means and regressions in analysis of covariance. *Biometrics, 13*, 282–307.

Stone, M., & Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *Journal of the Royal Statistical Society B*, *52*, 237–269; Corrigendum (1992) *54*, 906–907.

Sundberg, R. (1999). Multivariate calibration—direct and indirect regression methodology (with discussion). *Scandinavian Journal of Statistics, 26*, 161–207.