Mathematical Statistics
Stockholm University

Small sample bias and selection bias effects in multivariate calibration, exemplified for OLS and PLS regressions

Rolf Sundberg

Research Report 2005:16

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2005:16**,
http://www.math.su.se/matstat

Small sample bias and selection bias effects in multivariate calibration, exemplified for OLS and PLS regressions

Rolf Sundberg*

October 2005

Abstract

In multivariate calibration by for example ordinary least squares (OLS) multiple regression or by partial least squares regression (PLSR) the predictor $\hat{y}(x)$ is perfect for the calibration sample itself, in the sense that the regression of observed y on predicted $\hat{y}(x)$ is $y = \hat{y}(x)$. Plots of y against $\hat{y}(x)$ are much used to illustrate how good the calibration is and how well prediction works. Usually and rightly, this will be combined with cross-validation. In particular, cross-validation can show that for small samples the predictor $\hat{y}(x)$ will be biased, in the sense of making the regression coefficient of y on $\hat{y}(x)$ less than one, typically only slightly so for PLSR but substantially for OLSR, Another bias effect appears when y-values for the calibration are more or less selected. An increase in the spread of y might appear desirable because it increases the precision in the calibration. However, the resulting selection bias can affect both PLSR and OLSR substantially, and an additional problem with this bias is that it cannot be detected by cross-validation. These bias effects will here be illustrated by resampling from a large data-set, containing measurements on 344 pigs from slaughter pig grading.

Key words: Bias; Cross-validation; Multivariate calibration; OLS; PLSR; Prediction; Representativity

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: rolfs@math.su.se. Website: wwww.math.su.se/∼rolfs

1 Introduction

In multivariate calibration of, say, an instrument measuring absorbance at a number of wavelengths, a regression method is typically used to relate the concentration y of an analyte linearly to the instrument measurement vector x, and to form a predictor $\hat{y}(x)$ for the y-value of a future specimen. Regression methods used are for example ordinary least squares (OLS) multiple regression in simple cases, and partial least squares regression (PLSR) and principal components regression (PCR) for higher-dimensional or for other reasons near-collinear x.

To get an impression how well the predictor predicts, it is common and good practice to plot the actually observed y-values for the calibration sample or for a separate validation sample against the corresponding predicted values $\hat{y}(x)$. Provided a similar relation holds for new samples, this plot tells how certain we are about a future y when it is predicted to be $\hat{y}(x)$. For the calibration sample, the simple linear regression of y on $\hat{y}(x)$ is the perfect 45° line $y = \hat{y}(x)$, unless the predicted values are determined by cross-validation (e.g. CV leave-one-out), when this relation holds more or less approximately. Likewise, if a separate validation sample is used, the same relation is expected to hold approximately, at least if the validation sample is like the calibration sample. The latter condition is realistic in *natural calibration* which is when the calibration sample is regarded as randomly sampled from the population of interest.

The aim of this paper is to point out the restricted truth of these statements for the OLS and PLS regression methods (PCR is not included, but expected to behave similarly to PLSR). Firstly, we will demonstrate that if the calibration sample is small, the cross-validation-based regression of y on $\hat{y}(x)$ will be biased, slightly so for PLSR but substantially for OLS. For OLS, the existence of this bias was pointed out already 1983 by Copas [1]. Another bias appears when the y-values for the calibration are selected, so they do not represent the population for which prediction is desired. A typical aim of the selection is to yield a good spread in the calibration sample, because the more widely spread are the data points, the more precisely can a linear relationship be estimated. Therefore we will specifically study the effect of an overspread in the calibration sample y-values.

The results will be demonstrated in a simulation study by resampling from a large multivariate data-set from fat-grading in pig slaughteries.

For more general statistical aspects of multivariate calibration and of PLSR in this context, see for example [2–4].

2 Selection scenarios

The variable y is supposed to represent a concentration or some other intrinsic property of a specimen, that we want to be able to predict. The observed y can be the true property itself, but alternatively y is a measurement of the true property, with pure random error. In the latter case, the true property will here be denoted η. Note, however, that the predictors $\hat{y}(x)$ and $\hat{\eta}(x)$ of y and η, respectively, will be identical, since the pure random difference between y and η is best predicted by zero. For simplicity we use the notation $\hat{y}(x)$ for both of them, because any bias of the predictors will also be the same (but their precision will differ, because it is measured in different ways).

The kind of multivariate calibration data we have in mind are supposed to have a linear latent structure, as for example according to the general latent variable model of Burnham et al [5]. When the observed y is the true property, η takes the role of a y-related intrinsic property, a latent variable that is supposed to have a linear relationship with a latent vector t (on the x-side), $\eta = q't$ for some vector q. More precisely, η is a linear function of t, and the low-dimensional latent part of the instrument response vector x can be expressed as Pt for some matrix P. On the other hand, when η is the true property we could either think of η as being also a latent variable as above, or as being intermediate between y and another latent variable that is more primarily correlated with t.

All these cases basically fit in the natural calibration situation, in which x and y are correlated random variables, and the natural, ideal predictor would be the theoretical regression of y on x, if this were known. As a substitute we use the fitted linear regression of y on x, determined by a suitable regression method that is applied to the calibration sample, and of which OLSR and PLSR are important representatives.

Now suppose the calibration sample has been selected in order to yield a higher spread in y-values or in the underlying η, and in this way to achieve a more precise estimate of the relationship between y and x. Selection could be carried out in different ways.

Selection could be in terms of the true property itself, be it y or η. In practice, typical such situations would be when there are standards to choose

from, or when the calibration specimens can be prepared in the laboratory according to specifications as desired. Such selection would be deterministic. When the selection is carried out in terms of another intrinsic property, it is likely to have a more or less random or haphazard character. From a bias point of view, selection in y is likely to have more serious consequences than selection in η.

When η is the true property and y includes measurement error, a selection with respect to y will usually be less natural than selection with respect to η. However, when y is the true property of interest, or close to it, and it deviates substantially from the latent $\eta = q't$, it might be more adequate to think in terms of selection of y than selection of η. In any case, the properties to be studied here are not expected to differ much between selection in y and selection in η, because a substantial overspread in y by selection of y-values is hardly possible without a substantial overspread also in η, and vice versa.

Selection in η is equivalent to selection in the corresponding direction of the latent vector t, since η is a linear function of t. This is not equivalent with selection in a single x-variable from the instrument or in some other way in x. A recent paper [5] provides an example of selection in x. The authors study empirically some different strategies for division of the total set of specimens into a calibration set and a validation set, based on the x-values. When their model is linear, it is a multiple regression model, fitted by OLS. Analogously, approximate selection with respect to some aspect of the latent vector t could occur if a latent factor model is first fitted and the selection is made with respect to estimated latent factor values. The effects of such selection will not be included in this study, which will be restricted to selection in y.

3 A simulation study

3.1 Data for the simulation study

For demonstration of sampling effects, an excellent data-set will be used, representing 344 slaughter pigs. In the middle of the 1990es the Danish Meat Research Institute made a large study of some methods for grading slaughter pig carcasses. The quantity of interest, to be predicted, was the lean meat percentage. A reference value was obtained by dissection, $y =$ dissection lean meat %, This value is assumed to be without bias, but is certainly not without

measurement error. Various quick methods for x-measurements were tried. In one such method, called KC, the slaughter weight was supplemented by ten physical meaurements of fat and muscles thickness in specified positions. This made $\dim(x) = 11$. The KC method, together with the reference dissection method, was applied to the whole set of 344 pigs. These pigs were considered as reasonably well representing the population of Danish slaughter pigs, but this is not essential in our experiment, because we will consider the set of 344 pigs as the population in question. Also, we are not here interested in how good was the KC method, but we will use this data-set only to study sampling properties when taking small samples or samples with selected y-values.

The 11 x-variables explained $R^2 = 77.68\%$ of the total variation in y around the mean. A PLS regression (with centring but not scaling) showed that three latent variables was clearly the appropriate number of latent factors (and that was true for PCR, too). The three factors explained 77.27% and 90.56% of the total variation in y and x, respectively (around the means), Hence, in the experiments with PLSR the number of PLS-factors was fixed to be three, and this will be indicated by the notation PLS(3). Both the OLS regression and the PLS regression model fit the data well, and no peculiarities have been seen in the data. However, it may be worth noticing that spectroscopy data typically explain a higher percentage of the variation in y, at least when only the property represented by y has been varied in the calibration. This is likely to make the small-sample bias stand out relatively more clear for the present data.

3.2 Simulations

The large sample size of 344 pigs allows various kinds of further sampling to be carried out in the computer. The following resampling scheme was repeated 400 times on the KC dataset. First the set of 344 observations was randomly split in a basic calibration set of 172 observations and an equally large validation set. From the basic calibration set a subset of desirable size was taken, to form the actual calibration sample. Either this subset was just randomly sampled from the basic calibration set, or else the calibration sample was selected in a partially systematic way, as follows. A subset of the calibration set was first selected to represent overspread in the calibration. More specifically the k observations representing the $k/2$ largest and $k/2$ smallest y-values were selected. The study was repeated twice, with two

different k-values, $k = 86$ corresponding to overspread parameter $\theta = 1.30$, and $k = 56$ corresponding to $\theta = 1.50$, where θ is the ratio of standard deviations with and without overspread.

3.3 Simulation results

As mentioned above, simulations were run on the KC data, with varying calibration sample size and varying degree of selection in y, in order to see small-sample bias effects and selection bias effects. Each curve in the three subfigures constituting Figure 1 shows how sample size influences the slope of the linear regression of y on $\hat{y}(x)$. Each subfigure 1a - 1c shows four curves each, representing PLS(3) and OLS in combination with cross-validation-leave-one-out (CV) and test set validation. Three latent factors is the adequate number of factors for the PLS method on the KC data.

If the predictor $\hat{y}(x)$ had been applied on the same calibration sample as used to construct the predictor (i.e. without CV), the slope would have had the perfect, ideal value 1. Slopes less than 1, as shown in the diagrams, say that the predictor has a systematic tendency to exaggerate the influence of x by a factor that is the inverse of the slope, in other words, future observations tend to vary less in true y than the corresponding predicted values do.

Figure 1a shows that the slope gets smaller when the calibration sample is reduced. For PLS(3) the two upper curves of Figure 1a show that the influence of sample size is small or moderate. For the OLS regression method with small or moderate sample sizes, the bias in the CV-based slope is much higher. The CV-based slope is able to catch these reductions to a large extent, but both for PLS and OLS the CV-based slope is lagging about 0.1 units behind when the sample size is reduced. For OLS and sample size 16 (4 residual degrees of freedom) the slope is about 0.5, which means that the OLS predictor will exaggerate the influence of x by a factor 2.

Figures 1b and 1c show the influence of selection in calibration y-values. The diagrams are (perhaps remarkably) similar, which shows that already an overspread in standard deviation by a factor 1.3 is able to generate substantial bias effects, but these effects will not be much worse if the overspread is much higher. Remark: This is consistent with theoretical results indicating that there is a limit to the degree of bias as the overspread θ is increased.

The most important feature of Figures 1b and 1c is the difference between the CV-based upper curves for OLS and PLS(3) and the corresponding test-

set-validated curves (the two lower ones). The interpretation is not only that there is a substantial bias effect due to the selection, corresponding to a slope of magnitude 0.8 instead of 1 both for PLSR and OLSR, but also that this bias cannot be revealed by cross-validation. Of course a test set suffering from the same overspread would also be unable to reveal the bias.

The curves in Figure 1 show mean values of the slope taken over the 400 replicates of the resampling procedure. Median values were also computed, and they were always close to the mean values, thereby indicating essentially symmetric distributions of the regression coefficient. The standard deviation of the slope was of course smaller for large calibration samples than for small samples, and it was also larger when it was computed for the validation set than with cross-validation. Here are some examples: With a natural calibration sample of the small size 20 the four standard deviations were between 0.1 and 0.2. Selection with $\theta = 1.30$ reduced these standard deviations to fall between 0.05 and 0.12. Standard errors of the mean values are of course obtained from the standard deviations through division by the factor 20. Hence, even when the standard deviation was as large as 0.2, the mean value was sufficiently precise for the purpose of the study.

Illustrations of the features of Figure 1 are provided in Figure 2. Each of the four subfigures show y plotted against $\hat{y}(x)$ for the calibration data with cross-validation-leave-on-out (circles) and for the test set data (asterisks). Three regression lines are shown in each diagram, the ideal 45° line (solid), the line for the calibration data (dashed), and the line for the test set (also solid, but with slope less than 1). The data sets are chosen by slecting randomizations such that the regression lines agree reasonably well with the mean values shown in Figure 1. Both lower diagrams show the same data, while each of the upper diagrams show a separate data-set. The upper diagrams represent a small calibration sample without selection. We see that PLSR(3) works fine, whereas OLSR have much too low regression slopes. The lower diagrams illustrate that under selection in y, the cross-validated lines have a slope about 1, but the test set shows that this is systematically wrong, both for PLSR(3) and OLSR.

4 Discussion

In a discussion of confidence intervals for PLSR, Faber [7] commented that "in typical chemometrics applications bias is likely to be small". That is

probably true for most such applications, but a purpose of the present study is to warn against uncritical use of PLSR, or PCR or any similar method.

It is true that both the PLS and OLS predictors are essentially unbiased in the sense of equal mean values of the predictions $\hat{y}(x)$ and of the true values y. However, the present study shows that under certain circumstances the degree of dependence on x in the predictor can be seriously biased, exaggerating the true dependence. A small calibration sample is not very serious for PLSR, as long as it is much larger than the number of PLS factors needed, as seen from Figure 1a. OLS is much more sensitive, and this effect largely explains why OLS fails in much multivariate calibration. In comparison with sample size, selection can be more serious for PLS. The pig grading example showed that a moderate selection for overspread in y could have a substantial bias effect. It can be argued that selection in y is an odd idea in the pig grading situation, and that more natural selection procedures in this case would have less serious effects. However, in other calibration situations, where y is the property of interest and this y can be specified or determined beforehand, it is not unlikely that calibration samples are selected with the purpose of achieving large spread in y.

It should perhaps also be mentioned in this context that an often more serious cause for failure of the PLS predictor (and other predictors) is that the calibration sample does not reflect the full dimension of the natural population. If the latent dimension of the calibration sample is smaller than that of the natural population we risk serious prediction errors, because the variability in the missing dimensions may be correlated with y.

Note that even if the calibration sample has been randomly taken from the natural population, this does not guarantee that this sample shows a variation similar to that of the natural population, in particular if the sample is small. If there are reasons to expect lack of representativity of any kind in the calibration sample, this lack of representativity might also go over to the predictor.

It would be helpful to have at least approximate formulas for the bias effects, which could then be used in a practical situation to check that there is no serious bias to be expected, or possibly even to adjust the predictor itself. A formula for the small-sample bias of OLS under restrictive assumptions was given in [1], but more general such formulae for small-sample and selection effects have been derived and will be published.

Could the bias effects be corrected for? Copas [1] proposed this for the

small-sample effects with OLS. However, even if we have approximate formulas for a bias effect, the quantities in the formula must be estimated from data, and there is no guarantee that this does not introduce larger random errors than the bias that should be corrected for. Alternatively, it might appear from Figure 1 as if we could use a separate validation sample to find a correction factor for selection. However, this would neglect the randomness in the correction factor due to both the calibration sample and the validation sample. Also, this form of correction would imply that both the original calibration sample and the validation sample would be used in the calibration, with different roles, one part used only to find the direction of a predictor and the other part only to estimate a scalar factor along that direction. This would be inefficient use of the data.

The topic discussed in this paper could alternatively or additionally have been studied by simulating from an explicit statistical model for data, and by deriving approximate formulas for the bias effects. That will be done in a separate publication.

5 Acknowledgement

I am grateful to Eli Vibeke Olsen of the Danish Meat Research Institute for giving me access to the pigs grading data.

References

[1] J.B. Copas, Regression, prediction and shrinkage (with discussion), Journal of the Royal Statistical Society, Series B 45 (1983) 311–354.
[2] H. Martens, T. Næs, Multivariate Calibration, Wiley, Chichester, 1989.
[3] P.J. Brown, Measurement, Regression and Calibration. Oxford University Press, Oxford, 1993.
[4] R. Sundberg, Multivariate calibration—direct and indirect regression methodology (with discussion), Scandinavian Journal of Statististics 26 (1999) 161-207.
[5] K. Rajer-Kanduč, J. Zupan, N. Majcen, Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment. Chemometrics and Intelligent Laboratory Systems 65 (2003) 221–229.
[6] A.J. Burnham, J.F. MacGregor, R. Viveros, Latent variable multivariate regression modelling, Chemometrics and Intelligent Laboratory Systems, 48 (1999) 167–180.
[7] N.M. Faber, Response to 'Comments on construction of confidence intervals in connection with partial least squares', Journal of Chemometrics, 14 (2000) 363–369.

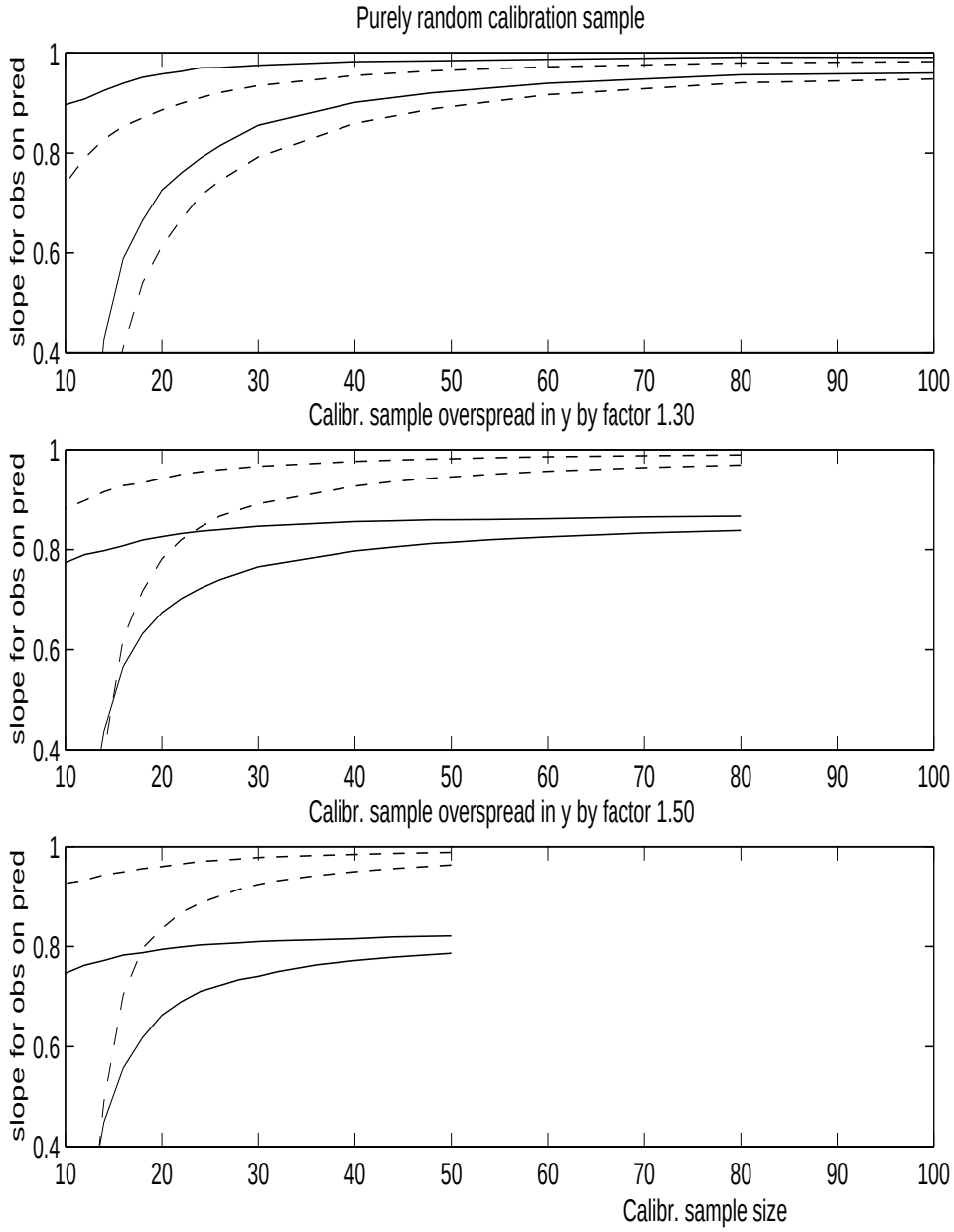

Figure 1: Small-sample bias and selection bias effects in regression coefficient of y on $\hat{y}(x)$ as functions of sample size
Top diagram: Natural calibration
Middle and bottom diagrams: Calibration sample overspread by factor $\theta = 1.30$ and $\theta = 1.50$.
In each diagram the dashed curves represent CV-leave-one-out and the solid curves represent test-set validation.
In each diagram the upper dashed and the upper solid curve represent PLSR, whereas the lower dashed and lower solid curve represent OLSR

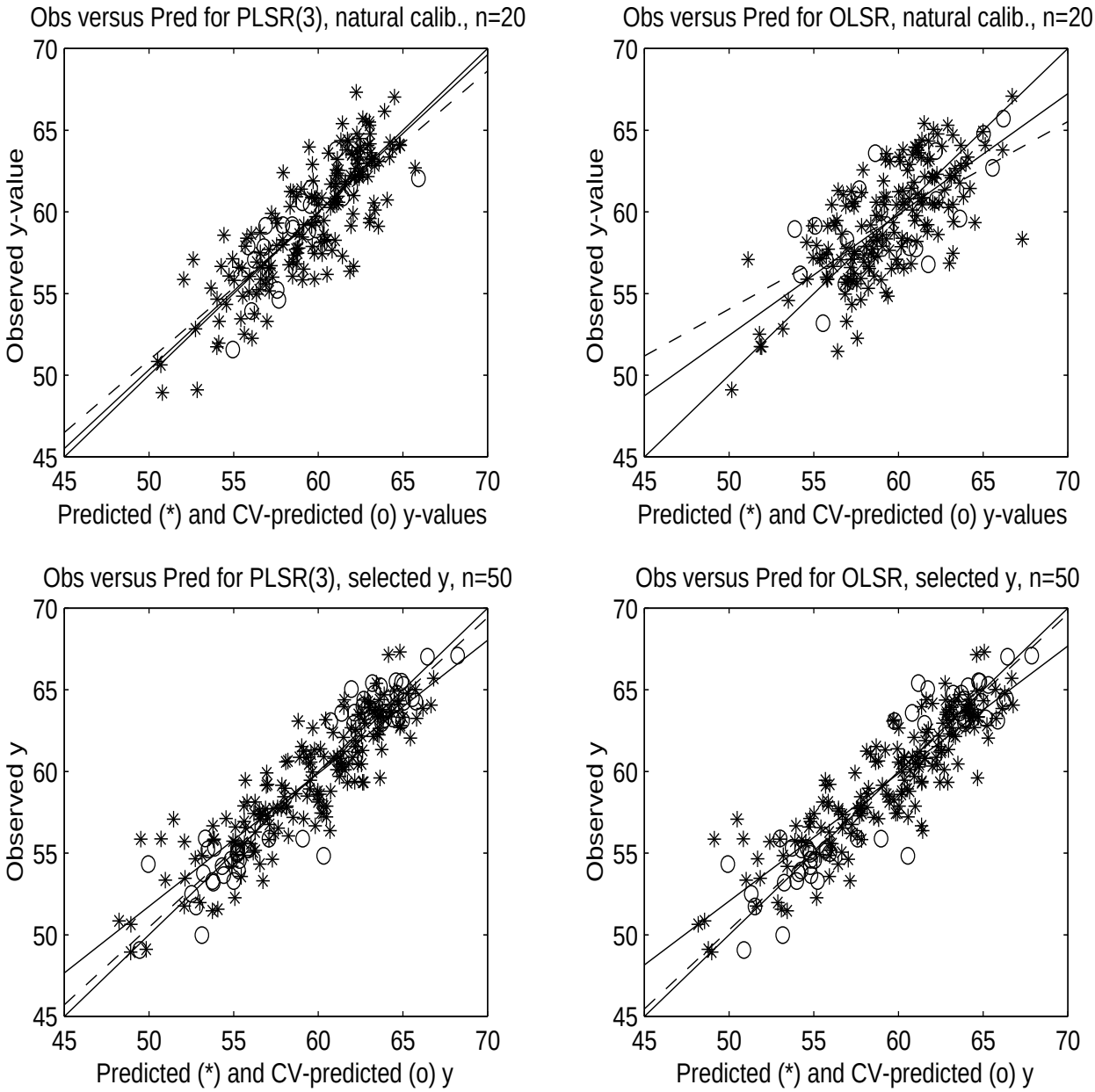

Figure 2: Illustration on KC-data of small-sample and selection biases for PLSR and OLSR.
Top: Natural calibration, small sample, size $n = 20$
Bottom: Calibration with selected y, sample size $n = 50$, $\theta = 1.50$
Left: PLSR(3); Right: OLSR
45° line between corners: Regression of y on $\hat{y}(x)$ in calibration sample
Dashed line: CV-regression of y on $\hat{y}(x)$ when y is left out of calibration (o)
Solid line of slope $< 45°$: Regression of y on $\hat{y}(x)$ for test set ($*$-data)