REVIEW OF METHODS FOR ASSESSING THE APPLICABILTY DOMAINS OF SARS AND QSARS

PAPER 3: Joint Applicability Domain and Predictive Uncertainty in QSAR Regression

Author:

Dr Tom Aldenberg (RIVM, Bilthoven, NL)
E-mail: tom.aldenberg@rivm.nl

Sponsor:

The European Commission - Joint Research Centre
Institute for Health & Consumer Protection - ECVAM
21020 Ispra (VA)
Italy

Contact: Dr Andrew Worth
E-mail: andrew.worth@jrc.it
http://ecb.jrc.it/QSAR

JRC Contract ECVA-CCR.496575-Z

VERSION OF 27 SEPTEMBER 2004

Joint Applicability Domain and Predictive Uncertainty in QSAR Regression

Tom Aldenberg
RIVM
Aug. 2004

Introduction

In this section we will discuss methods of evaluating the predictive uncertainty of QSAR models. Most QSAR models are developed on the basis of linear regression, so we will treat the assumptions underlying this approach and examine ways of assessing the quality of the fit, and analyzing their predictive value.

Most quality assessments focus on the quality of the fit on the training data used to develop the QSAR. Predictive uncertainty is certainly related to the fit and its uncertainty, but, the performance of the line on data, preferably validation data from other sources than the training data set is equally important, and may not result into satisfying predictions, even if the quality of the line was acceptable.

The prediction error can be evaluated with regard to:
(1) the training data, or subsets of it
(2) new data, the test or validation set,
(3) and on both sets combined.

We will mainly focus on Multiple Linear Regression (MLR), but other approaches are used in QSAR modeling, such as Partial Least Squares (PLS), Regression Trees, and Clustering.

In Statistics–and in MLR in particular–one has two 'philosophies' of calculation and interpreting results: Classical (Frequentist) and Bayesian Statistics. Roughly taken, the classical vision assumes the model as given, and the data are a particular sample of a theoretically infinite number of similar samples. Bayesians hold that the data are unique, and the model is uncertain, which is implemented by considering model parameters probabilistically. Actually, the latter mental picture is the one most applied scientists maintain, even when interpreting classical results, like confidence limits.

For MLR, the mathematics of both types of statistical analysis is *almost identical*, but the Bayesian interpretation yields more tools to evaluate (interpret) predictive checks (Chaloner and Brant, 1988, Gelman et al., 2004, p. 359). This is exactly what is needed in QSAR predictive uncertainty assessment. So, we will treat both classical and Bayesian model fitting, and examine how they can be combined (re-interpreted) in a clarifying manner. Our analysis will lead to some new plots to assess the predictive uncertainty of a (QSAR) regression model.

Multiple Linear Regression

In Sokal and Rohlf (1995, p. 455) a distinction is made between Model I Regression and Model II Regression.

Model I Regression is what most applied scientists are familiar with: the independent (predictor) variable(s), X, are measured *without error*. The expected value of the prediction variable, Y, is a linear function of the predictor(s). The residuals between prediction and observed Y-values follow a normal distribution, and the variance along the regression line, or surface is a constant, not depending on the predictors (*homoscedasticity*).

In Model II Regression (Sokal and Rohlf, 1995, p. 541), some or more of the Model I Regression provisos are not satisfied. Both X and Y follow some bivariate, or multivariate, distribution. A special case is, when X and Y are bivariate normal, which is easy to handle. In that case, the regression of Y on X has the same equation as that for the Model I regression. In other cases, Sokal and Rohlf express a generally felt uneasiness: "Research on and controversy over Model II regression continues".

Many applied scientists think that regression is not justified when the predictors are not chosen at experimentally designed values, or if not measured without error.

This is simply too restrictive a vision. The modern treatment in statistics holds a much more general view, which is based on Decision Theory. In Hastie et al. (2001, p. 18), the justification of regression runs as follows.

Let X be a random input vector of predictors, and Y a random output variable. Suppose, they have a joint distribution $\Pr(X,Y)$. We seek a function $f(X)$ for predicting Y at given values of the input X. In decision theory one may define a so-called loss function for judging prediction errors. A convenient loss function is *squared error loss*:

$$(Y - f(X))^2 .$$

When we want to make the best prediction minimizing the expected (squared) prediction error at given values of x of X, then the solution is

$$f(x) = \mathrm{E}(Y \mid X = x),$$

called the *conditional expectation*, a.k.a. the *regression function*. Hence, the *best* prediction (best in the sense of the loss function) of Y at any point $X = x$ is the conditional mean.

In this view both predictors and prediction variables are *random*, that is distributed, there is no assumption on being error-free; there is no assumption on the joint distribution, nor on the linearity of Y over X.

In linear regression $f(x)$ takes a special form. There may be confusion about the word *linearity*. The inputs X can come from different sources (cf. Hastie et al., 2001, p. 42):

- quantitative
- transformations of quantitative inputs (logs, square roots, etc.)
- basis expansions, i.e. power of inputs
- numeric codes of qualitative inputs, e.g. 0 and 1 for some grouping
- interactions between variables

No matter the source, or non-linearity, of the X_j predictors, the model is *linear* in the *parameters* β_j:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \cdot \beta_j \, .$$

In this framework, some predictors may enter the linear regression in a *non-linear* way, which yields a potentially unlimited number of possible regressions. To choose the predictors in MLR is a problem called *model selection*. This is a difficult subject that is still evolving (cf. Liang, 2002). Appendix 1 shortly discusses some aspects of model selection, touching on a few alternative modeling techniques, such as Ridge Regression and Partial Least Squares (PLS).

Whatever the final model, part of the variation of the response variable is left unexplained. This is the *residual variation*, or *un-modeled* part. The important point we are going to make is that predictive uncertainty is not only a matter of the quality of the best fitting model, and its uncertainty, but also on the residual variation that is to be expected when applying the model. Thus, we do not only focus on the confidence limits of the model (regression line), but also on the predictive limits of the occurrence of response values. We will approach this by working through an example.

Case Study

We will use Debnath et al. (1992, Table IIa), and Glende et al. (2001, Table 7) as a case study. In Table IIa, Debnath et al. (1992) present the mutagenicity of *S. typhimurium* TA98 with metabolic activation (S9) caused by aromatic and heteroaromatic amines. The linear regression (QSAR) is based on 4 physico-chemical parameters.

The table has 95 cases, from which records # 89 up to 95 are left out of the model. This leaves a data set of 88 points (cases).

The single response variable is: $\log TA98$ [log Revertants/nmol]. Potential (untransformed) predictor variables are: $x_1 = \log P$, $x_2 = e_{LUMO}$, $x_3 = e_{HOMO}$, and $x_4 = I_L$.

Only $x_4 = I_L$ is categorical, taking values 0 and 1 (meaning that the compound has three or more fused rings).

We will examine two single predictor models. The first does regression on log *P*:

(**M1A**) $$\log TA98 = -4.23(\pm 0.86) + 1.69(\pm 0.35) \log P ,$$

with reported statistics $n = 88$, $r = 0.723$, and $s = 1.329$ (Debnath et al., 1992, p. 41). This is a 1+1 predictor fit: intercept and 1 predictor.

The other single predictor model takes only I_L into account:

(**M1B**) $$\log TA98 = -1.31(\pm 0.32) + 3.09(\pm 0.55) I_L ,$$

with statistics $n = 88$, $r = 0.771$, and $s = 1.224$ (same reference). This is also a 1+1 predictor fit, but the single predictor variable is now categorical {0, 1}. We observe that this model fits somewhat better than the one against log *P*. Note that we may compare models of equal complexity in this way.

A 1+2 predictor model that combines the two previous predictors into one regression is:

(**M2**) $$\log TA98 = -3.57(\pm 0.62) + 1.09(\pm 0.27) \log P + 2.23(\pm 0.47) I_L$$

with statistics $n = 88$, $r = 0.875$, and $s = 0.938$ (same reference).

The full model, a 1+4 predictor fit, is reported as

(**M4**) $$\begin{aligned} \log TA98 = {} & 7.20(\pm 5.4) + 1.08(\pm 0.26) \log P - 0.73(\pm 0.41) e_{LUMO} \\ & + 1.28(\pm 0.64) e_{HOMO} + 1.46(\pm 0.56) I_L \end{aligned}$$

with $n = 88$, $r = 0.898$, and $s = 0.860$ (same reference).

For some later comparisons, we may add the *null model*, i.e. a model with no predictor variables at all. The 'intercept' is just the mean, and the residual standard error of the model is the sample standard deviation of all observations:

(**M0**) $$\log TA98 = -0.257(\pm 0.41) ,$$

with $n = 88$, $r = 0.0$, and $s = 1.912$ (calculated here).

Model 1A (Single Predictor Log *P*)

Best Fit

One measure of the quality of the fit is the coefficient of determination (COD), also called the squared multiple correlation. It is $R^2 = 0.5228$ for the **Model 1A**. The well-known interpretation is explained divided by total variation, but another useful way of looking at it is as the square of the *ordinary* correlation coefficient of *observed* against *fitted* model values (Cook and Weisberg, 1999, p. 165).

The observed over fitted plot should be routinely reported for any (QSAR) regression. The great advantage is that it can always be made, no matter how many predictor variables are included.

The next Fig. 1 shows a plot of observed mutagenicity against fitted values for **Model 1A**. The correlation of the scatter is 0.7231, i.e. the *r* reported by Debnath et al. (1992) for this model, which indeed yields $R^2 = 0.7231^2 = 0.5228$.

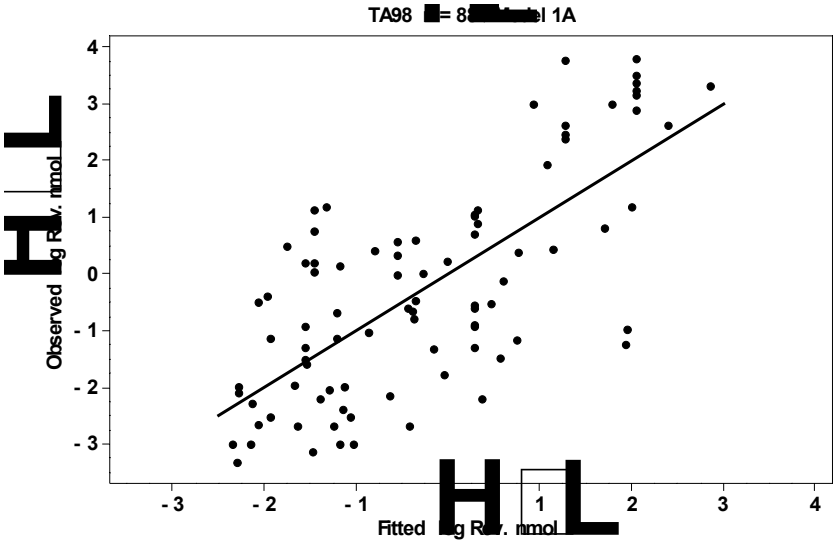

Fig. 1. **Model 1A.** Plot of observed values against fitted values for the 1+Log *P* predictor fit of Table IIa in Debnath et al. (1992).

From this plot, one derives the *residuals against fitted* plot by subtracting the *fitted* values from the *observed* ones (Fig. 2), which is also always possible even with multiple predictors. It gives an overview of how the model performs, how large the *un-modeled* part is, whether the variance of the residuals seem constant (*homoscedasticity*), etc.

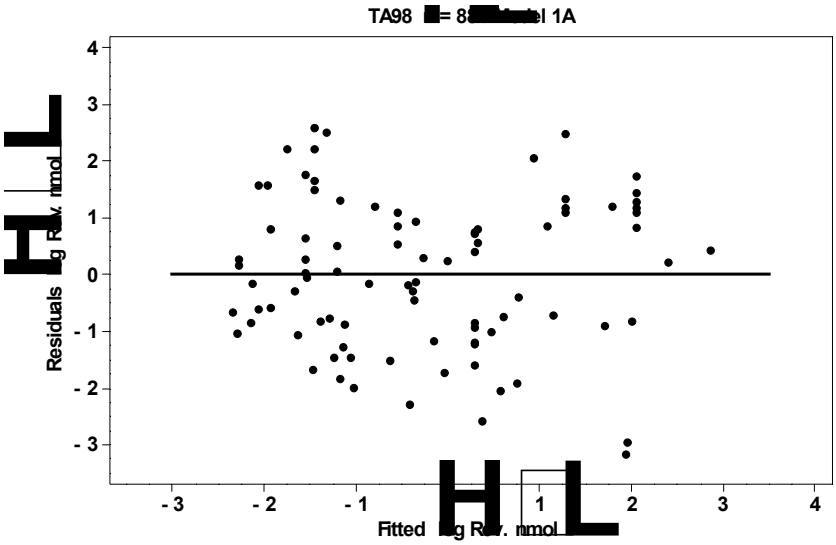

Fig. 2. **Model 1A.** Plot of model prediction residuals against fitted values for the 1+ Log *P* predictor fit of Table IIa in Debnath et al. (1992).

However, with only one predictor variable in Model 1A, it is also possible to plot the observed values and best fit model against the single predictor variable, which is shown in Fig. 3.

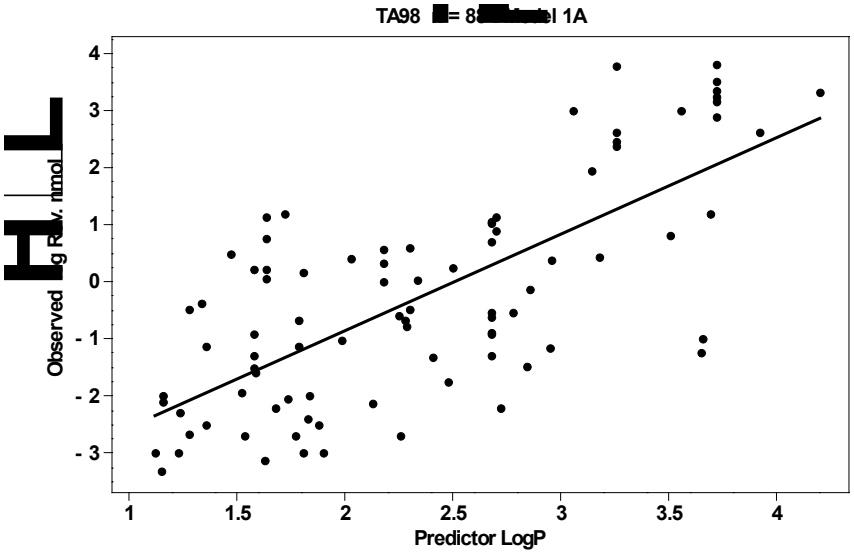

Fig. 3. **Model 1A.** Plot of observations and fitted model against the single predictor variable for the 1+ Log *P* predictor model.

Best Fit Uncertainty and Predictive Uncertainty

If we want to model predictions, we have to look beyond the best fit and its uncertainty, and hypothesize the distribution of the residuals. A quick way to do this is to add say 2

times the standard error of estimate to the best fit. This leads to straight line uncertainty bands parallel to the best fit.

Here the predictive uncertainty limits are calculated from a Bayesian interpretation of the ordinary linear regression estimates (cf. Appendix for technical details) and are only approximately straight lines. They are mathematically identical to the classical prediction limits (cf. Helsel and Hirsch, 1992, p. 242):

$$\hat{y} \pm t_{\alpha/2} \cdot \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{1/2} \cdot \hat{s} \,,$$

with x_0 a given input (predictor) value, and $\hat{y}$ the model fit at that input.

The Student-t value for $n - 2 = 86$ degrees of freedom with $\alpha = 0.02$ leads to 2.37 resulting in a coverage of 98% of the predictive distribution. The reduction of 2 for the degrees of freedom relates to the number of coefficients fitted (intercept and one predictor coefficient). The estimate $\hat{s}$ is the residual standard error also with $n - 2 = 86$ degrees of freedom. It is the same as reported by Debnath et al. for the model *s*.

Note the resemblance to the confidence limits for the fitted line:

$$\hat{y} \pm t_{\alpha/2} \cdot \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{1/2} \cdot \hat{s}$$

The leading 1 inside the parentheses causes the prediction limits to be less sensitive to input value x_0, than the confidence limits of the fitted line.

This can be shown, when we overlay the prediction limits (red) and credibility limits of the line (blue) in Fig. 3 to yield Fig. 4. Note that the red lines are only *approximately* linear.

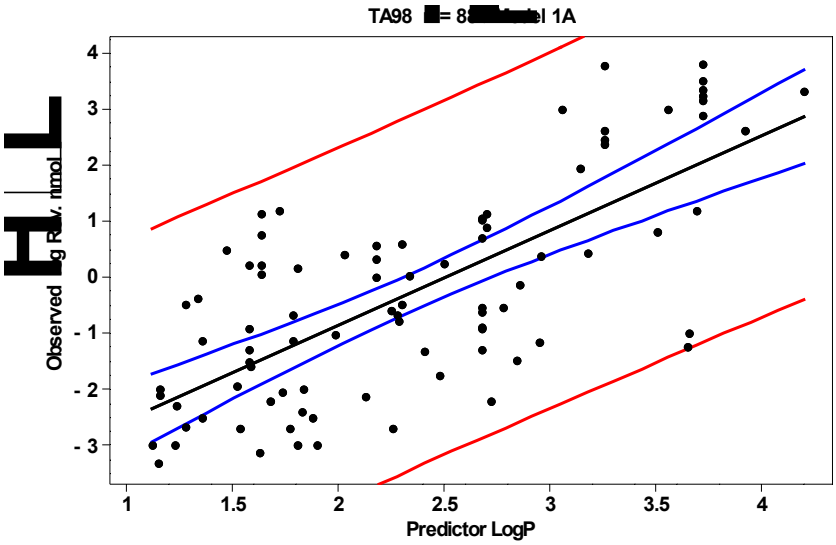

Fig. 4. **Model 1A.** Plot of observed values against the single descriptor for Model 1A. The inner limits represent the best fit uncertainty. The outer limits indicate the predictive uncertainty.

Model Test Run on Independent Data

For a one-predictor model, one can overlay the best fit limits and predictive limits with independently measured data for some other compounds. We overlay new data (n = 18) from Glende et al. (2001, Table 7) with the prediction uncertainty against predictor graph of Fig. 4, to yield Fig. 5. The case numbers of the new data are attached to the points.

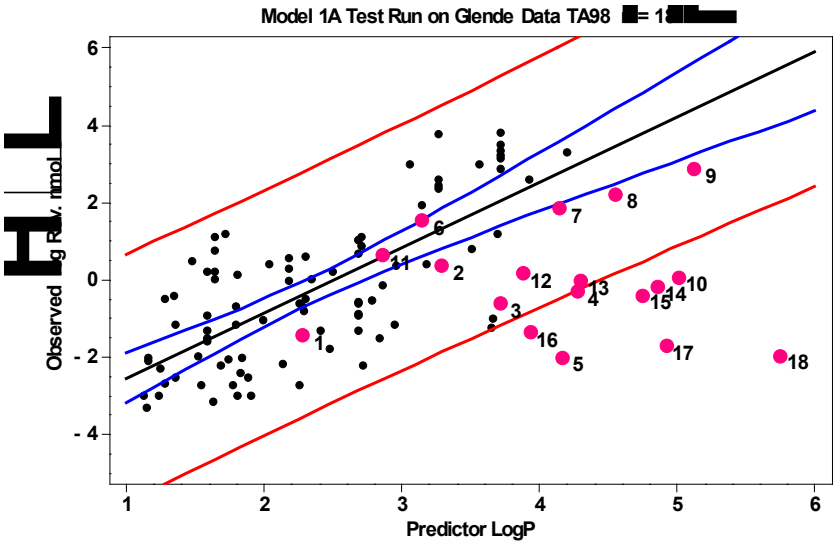

Fig. 5. **Model 1A.** Test run of Model 1A on 18 independent test results (Glende et al., 2001, Table 7) . Integers are the case numbers of the Glende et al. data.

From this graph, it becomes obvious that validation *must* employ predictive uncertainty, not only fit uncertainty. Because, even with these wider bands, the model uncertainty is

overly optimistic about independent test values (provided there are no chemical or biological reasons explaining the systematic deviations right away). One may reason that 7 tests (#8, 9, 10,14, 15, 17, and 18) are outside (above) the applicability domain, but clearly the model performs well for #8 and 9, which are also out of bounds, while it performs questionably for #5 and 16, which are definitely within the one-dimensional predictor space.

Joint Applicability Domain

In this section, we propose to assess a so-called joint applicability domain for both predictor(s) and response on the training set. As a first approach, we calculate probability contours for the joint distribution of *X* (predictor(s)) and *Y* on the basis of the bivariate/ multivariate distribution. This is for the time being sufficient to label data points as inside or outside the domain.

The question of what probability cuts-offs to take, can be provided by the well-known graphical device in exploratory data analysis (EDA) called box-and-whisker plots, or box-plots for short (Tukey, 1977). For univariate empirical data, outside values are characterized by being *beyond* 1.5 times the interquartile range (IQR) above the third quartile (or below the first quartile). If data are normally distributed this happens in roughly 1% of the cases. Extreme ("far") outside values are beyond 3.0 time the IQR above, or below the nearest quartile. This only happens one in a million times. So, when a data set displays several near or far outside values, they may contain erroneous data, and/or the data may come from another distribution, e.g. a more skewed, and perhaps need transformation.

This idea of outside values, we transfer to the multivariate normal distribution. Points within the (elliptical) contour modeling 99% of the probability, are inside points. They are in the 'code green' zone (OK). The points between this boundary and the outer ellipse covering 99.9999% of the cases are 'outside'. A model should be used for those values with *caution*. This is the 'code orange' zone. Data points outside the outer ellipse are in the 'code red' zone. To find data points there only happens once in a million cases, and are therefore extremely unlikely given the training data.

In Fig. 6, we draw these elliptical contours for the one-predictor model M1A. The ellipse inside the cloud is analogous to the 'box' of the box-plot: it contains 50% of the probability of finding data points there. The contour that circumscribes the data is the 99% probability contour. This ellipse marks the joint applicability domain, covering 99% of the possible variation. Points between this contour and the outer thin ellipse are 'outside' values. There the model may still apply, but warnings or additional investigation would be justified. Points that are outside the thin ellipse are 'far outside' values, to be signaled as outside the applicability domain. The model should not be routinely applied there.

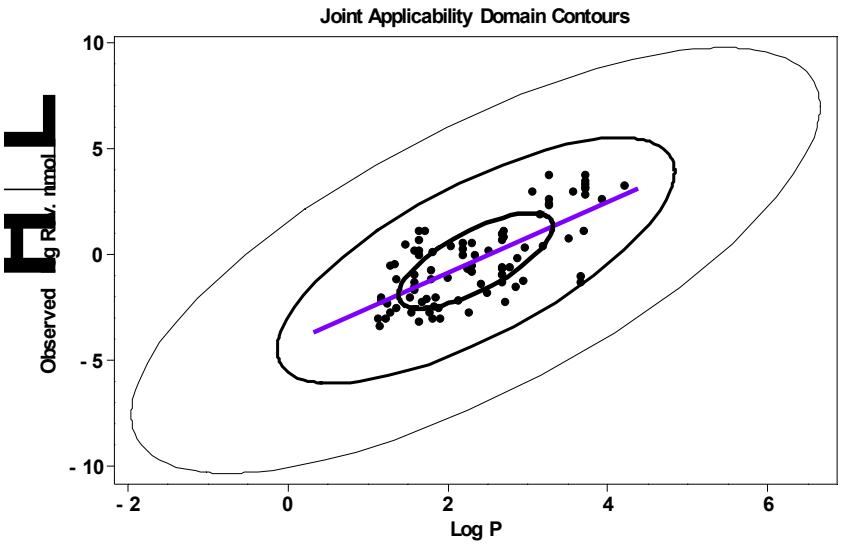

Fig. 6. **Model 1A.** Joint Applicability Domain of Predictor and Observed Response values for the single descriptor Model 1A. The straight line is the best fit regression line from Figs. 3 and 4. The elliptical contours contain a given fraction of the bivariate normal probability mass: 50%, 99% (used as the domain of applicability) and 99.9999%. The model should not be applied outside the outer ellipse, and only used with caution within the outer two ellipses. These values correspond to the usual *outside* and *far outside* values in box plots.

It is important that the best regression line is compatible with the multivariate normal model for the joint variation of predictor(s) and response. In the introduction, we have seen that for *any* joint distribution of X and Y, the regression line is the *best* prediction of Y given X one can make. For the bivariate and multivariate normal distribution, the regression line is a *linear equation*, and exactly equal to the ordinary least squares (OLS) fit, as is commonly fitted. Note that the regression line is less tilted than the principal axis of the elliptical contours. This is the reason why Y on X regression and X and Y regression is different.

In Fig. 7, we overlay Fig. 4, the predictive uncertainty with the applicability domain contours, to see how they correspond. We note that the conditional predictive uncertainty corresponds roughly with the fat ellipse containing the joint 99% of the domain.

It would be possible to assess the joint domain with more complex models, involving the x-dependent conditional distribution of Y, and a more sophisticated estimate for the predictor (descriptor) domain, but this needs further study.

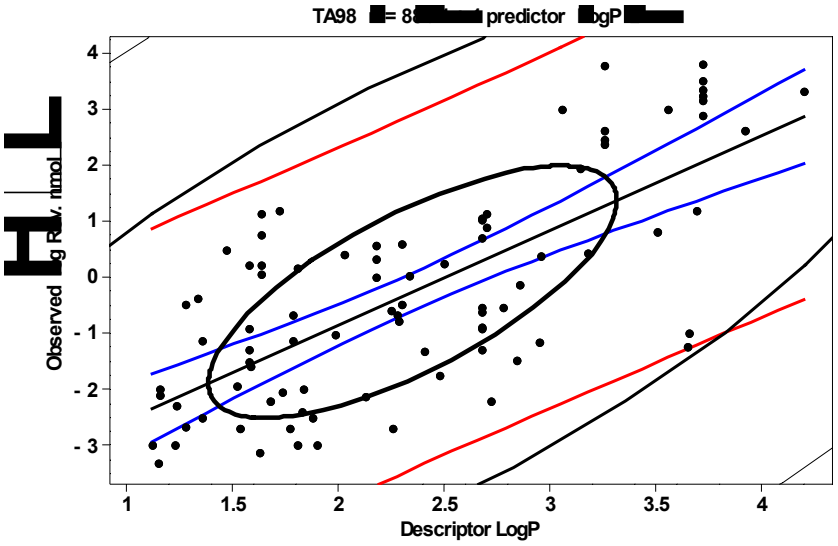

Fig. 7. **Model 1A.** Joint Applicability Domain of Predictor and Observed Response values for the single descriptor Model 1A, the best fit confidence limits (blue), and the conditional predictive limits (red). The straight line is the best fit regression line. The elliptical contours are as in Fig. 6. The thick outer ellipse models the joint applicability domain. The thin elliptical contours that are just visible in the corners mark the occurrence of extremely unlikely data, corresponding to far outside values in a boxplot.

To see how this works on a validation set, Fig. 8 puts the Glende data (18 cases) on the joint applicability contours.

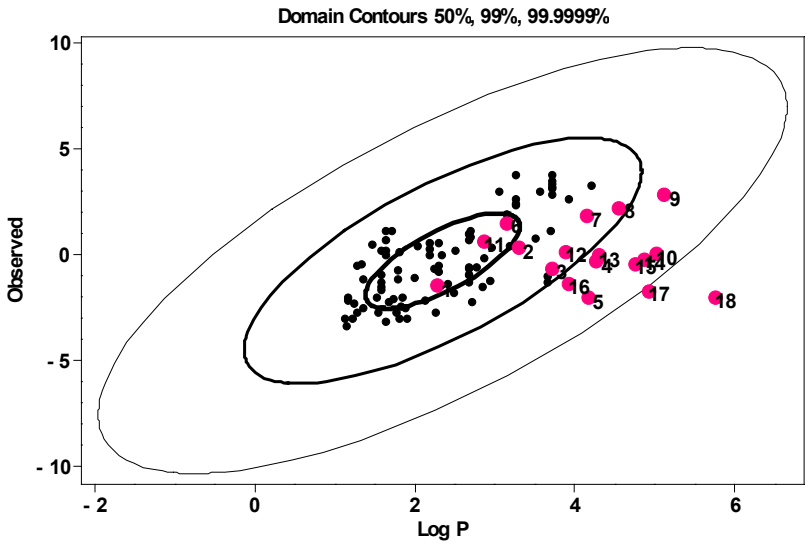

Fig. 8. **Model 1A.** Joint Applicability Domain of Predictor and Observed Response values for the single descriptor Model 1A. The labeled points are the 18 substances in the Glende et al. (2001, Table 7) test data. We note that cases #1, 2, 3, 6, 7, 8, 11, and 12 are within the joint applicability domain. Cases # 4, 5, 9, 10, 13, 14, 15, 16 are in the code orange zone (triggering warnings or further analysis), while 17 and 18 are in the red zone, not warranting the use of the model.

Model 4: Full Model with 4 Predictors

We start with repeating observed versus fitted and residuals over fitted for the full model. The advantage of this plot, as mentioned above, is that it can be made for *any* number of predictor variables (Cook, 1998, p. 5).

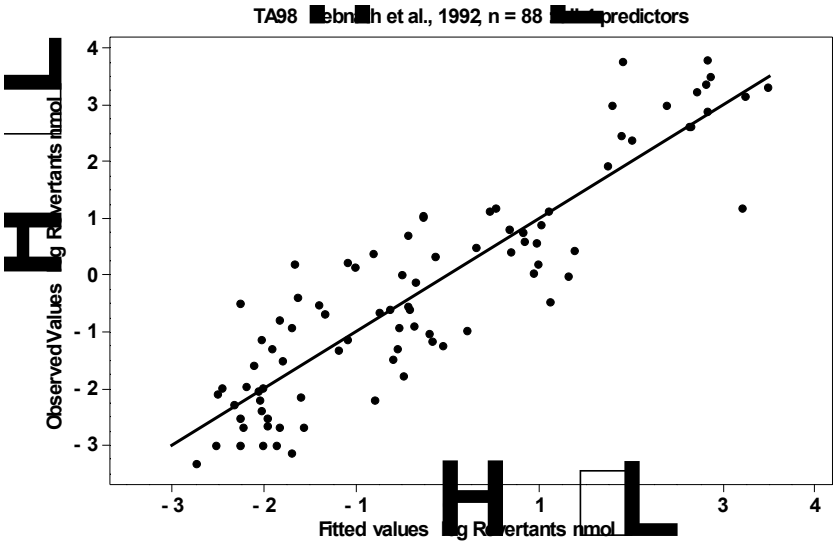

Fig. 9. **Model 4.** Plot of observed values against fitted values for the 1+4 predictor fit of Table IIa in Debnath et al. (1992). Correlation coefficient is 0.8982. The square of this equals the COD: R^2 = 0.8068. The straight line is the 1:1 reference line, it is not fitted.

From this plot, we construct a plot of the residuals of the fit against the fitted values, by subtracting the fitted values from the observed values.

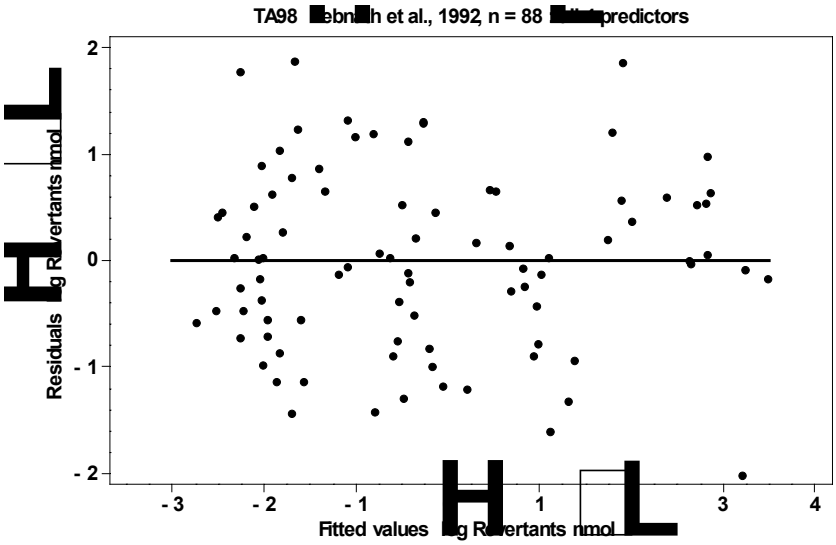

Fig. 10. **Model 4.** Plot of model prediction residuals against fitted values for the 1+4 predictor fit of Table IIa in Debnath et al. (1992).

We might now do some distribution checking of the residuals: histograms, kernel fitting, probability plots. The next Fig. 11 shows a histogram of the residuals of the 1+4 fit. The distribution seems well-behaved, and we could further examine this in normal probability plots.

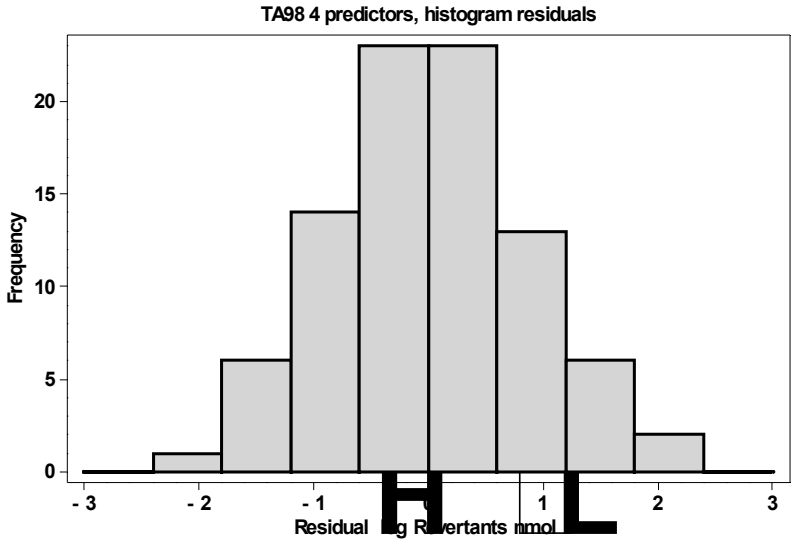

Fig. 11. **Model 4.** Histogram of model residuals for Model 4 on the training data from Debnath et al. (1992), n = 88.

But let us focus on the apparent uncertainty of the residuals. They span 4 orders of magnitude. This gives an idea on the remaining, i.e. *un-modeled* predictive uncertainty when using the full model. Even in the case of the full model, the uncertainty of the regression line severely underestimates the prediction error.

Predictive Uncertainty Plots

We redraw Fig. 9, the classical regression *observed over fitted* plot for the full model, now indicating the case numbers of the Debnath et al. (1992, Table IIa). This leads to Fig. 12. Here fitted means best fit (point) values. The observed values are the output data themselves.

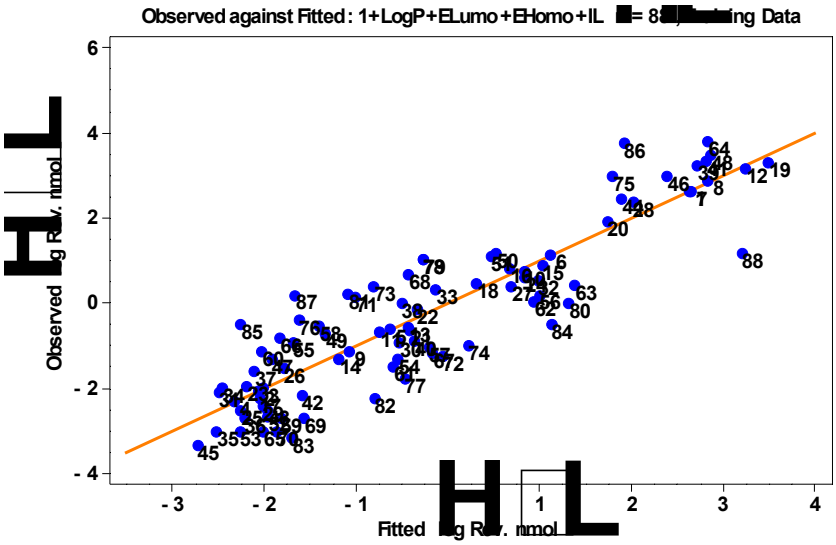

Fig. 12 Classical observed over fitted plot for the full (four-predictor) model best fit. Case numbers (n = 88) are those corresponding with Table IIa in Debnath et al. (1992).

Fig. 13 subtracts the model best fit from the observations to obtain the *residuals over fitted* plot. This looks fine, there seems no concern about heteroscedasticity, i.e. unequal variance. We note the extreme cases #45, 35 on the left, and #88, 12 and 19 on the right. Note also that #82, 83, 84, up to #88 constitute the residual boundary cases.

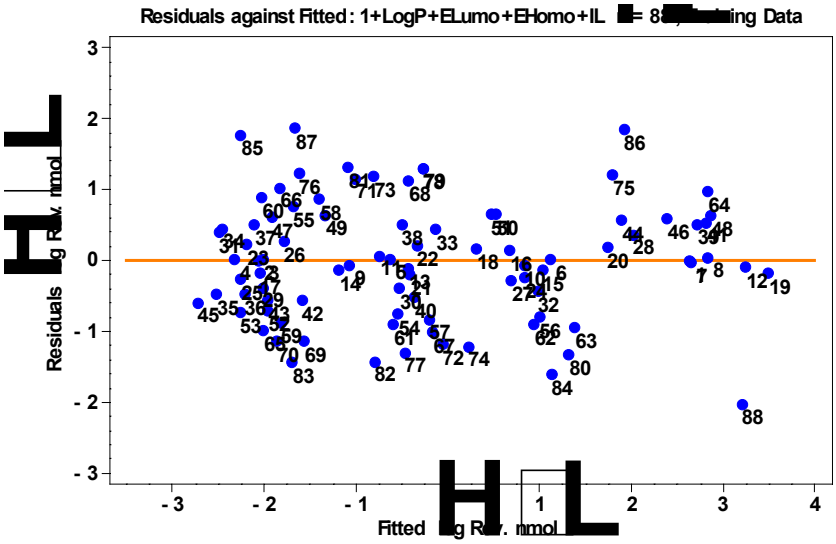

Fig. 13 Classical residuals over fitted plot for the full (four-predictor) model best fit. Case numbers (n = 88) are those corresponding with Table IIa in Debnath et al. (1992).

Now, we want to include the predictive uncertainty into these plots. Like in the *Y* against measured predictor plot, in the case of a single predictor (cf. Figs. 4 and 5), we put the predictive interval against observed (measured) response values (Fig. 14). The straight line is the 1:1 correspondence line to find the deviations.

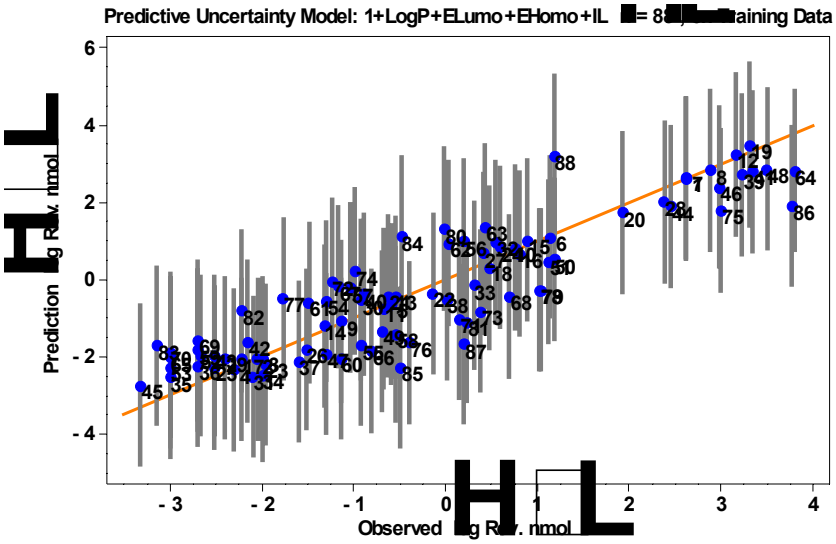

Fig. 14 **Model 4.** *Predictive uncertainty against observed* plot for the full (four-predictor) model. The grey prediction bands are the 98% credibility limits of the predictive distribution at each predictor input combination. The uncertainty bands weakly depend on the predictors. Case numbers (n = 88) are those corresponding with Table IIa in Debnath et al. (1992). Since the mean of the predictive distribution is the best fit value this graph is the transpose of Fig. 12.

One observes that the predictive uncertainty is approximately 4 orders of magnitude as it was in the classical residual plot. However, middle values are more probable than credibility limit values or values beyond. This follows from a Bayesian interpretation of the prediction intervals (see Appendix). Note also that all the ranges cover the observed data.

To test the performance of the model on the Glende et al. (2001, Table 7) validation data set, we develop the same prediction over observed plots at given predictor values for the full 4-predictor model (Fig. 15).

We note that cases #5, 10, 16, 17, and 18 of the Glende data seem to predict systematically too high. This corresponds well to the joint applicability assessment done for Model 1A in Fig. 8. Cases #5, 10, and 16 are in the 'code orange' zone of Model 1A, while 17 and 18 are in the 'code red' zone. The model should not be applied to these data without further detailed analysis. The model may be applied to 'code orange' cases only with caution.

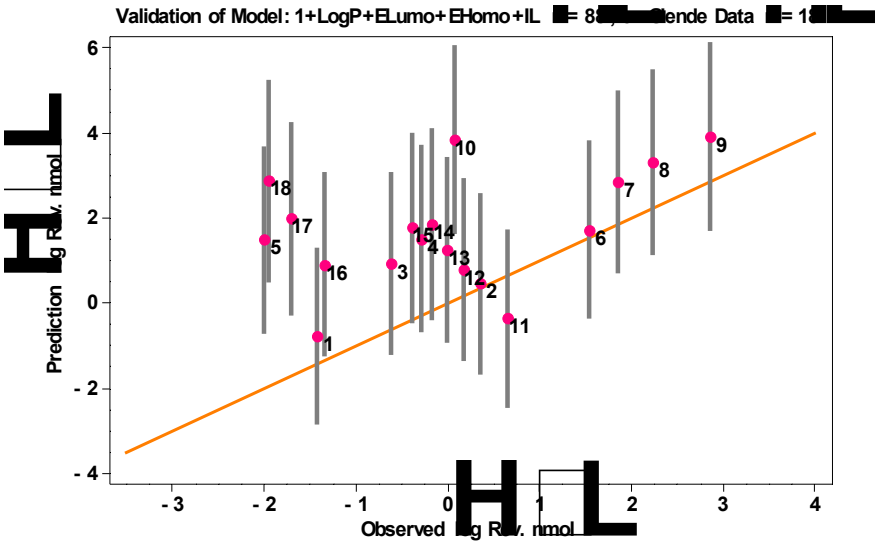

Fig. 15 **Model 4.** *Predictive uncertainty against observed response* plot for the full (four-predictor) model applied to the Glende et al. (2001, Table 7) test data. The grey prediction bands are the 98% credibility limits of the predictive distribution at each predictor input combination. The case numbers correspond to the Glende data. Note that cases #5, 10, 16, 17, and 18 seem to systematically overpredict the data. This corresponds to the joint applicability domain assessment for **Model 1A** in Fig. 8. Cases #5, 10, and 16 are in the code orange zone of Model 1A, while 17 and 18 are in the red zone.

Note that the joint applicability domain was assessed for Model 1A only–yet gives a good indication of where to expect trouble with that model, or more complicated models apparently. Incorporating more predictors when going from Model 1A to Model4 has not substantially improved the prediction in these cases.

The joint applicability domain assessment needs refinement in the case of a discrete predictor, which only takes integer values. For *each* value (0, and 1), the joint applicability domain has to be assessed *separately* for the response and remaining three continuous predictor variables, using a quadrivariate normal distribution.

Appendix 1. Model Selection

This section only touches on some aspects of model selection. A thorough review would require a much wider discussion and depth.

Some Alternatives to MLR

When reading statistical literature on validation, model testing and predictive assessment, a large part of these discussions addresses model *selection*. Model selection is important when developing a model for Y on a potential set X of descriptors. The question may be which descriptors, or functions of them (cf. the five bullets above), best model Y. To

simplify the discussion, we will assume Y to be univariate (a single variable). We will closely follow Hastie et al. (2001, Section 3.4: Subset selection and coefficient shrinkage)

If the set of descriptors exceeds the number of data combinations (cases), MLR will not be possible on the whole set. One has to rely on (predictor) subset selection. *Best subset regression* finds for a given limited number of predictors the subset yielding the smallest Residual Sum of Squares (RSS). If one allows powers and cross-products for interactions, one has to set a limit to the total number of predictors, or predictor combinations, and the resulting number of possibilities may be beyond calculating capacity. The criterion to set the limit cannot come from RSS, because this will increase, whenever more predictors or functions of them are added.

Often best subset is replaced by stepwise procedures (forward, backward or hybrid). These are very popular, but may not lead to the best subset obtainable.

Moreover, it is commonly known that putting too many predictors in a MLR may be dangerous. The problem of collinearity causes coefficient variance to be inflated, badly influencing prediction accuracy. Setting some of the coefficients manually to zero, is a crude way of discarding input variables.

A more sophisticated way to protect coefficients for blowing up the regression due to collinearity is *ridge regression*. This is also a least squares estimate, but subject to the constraint that the sum of the squared coefficients does not exceed a given size.

Ridge regression has two interesting interpretations. One is a Bayesian interpretation in which the coefficients β_j have a normal prior distribution with mean 0 and a given variance. So, one restricts the coefficients by supposing them to not to differ from 0 more than a certain a priori amount in the 'normal' sense. The best fit is the posterior mode (also the mean because of symmetry).

The other interpretation is very interesting compared to other MLR alternatives, like Principal Components Regression (PCR) and Partial Least Squares (PLS). The gradients of the response surface in the direction of the smallest principal components in correlated X space are protected against potentially high variance. These directions are exactly the causes of trouble in MLR. The shrinkage is greatest in the worst directions.

Principal Components Regression (PCR) is a *reduced* regression that uses derived inputs, based on principal components, of less than or equal dimension than the original inputs. So, while ridge regression shrinks the coefficients of all the principal components (without essentially reducing predictor dimensionality), PCR discards the smallest eigenvalue components. If none of the components are discarded, Ordinary Least Squares (OLS) results.

Partial Least Squares (PLS) also constructs linear combinations of inputs for regression, but it uses *both X and Y* for their determination. PLS is also used in QSAR modeling, cf. (Ericksson et al., 1997, Ericksson et al., 2003). The inputs are weighted by the strength of their univariate effect on Y. As in PCR, stopping at a lower dimension produces a reduced

regression, while continuing to the full input dimension leads to OLS, again. Hastie et al. (2001, p. 70) mention that PLS also tends to shrink the low-variance component directions, but can actually inflate some of the higher variance directions.

Quoting Frank and Friedman (1993), Hastie et al. (2001, p. 70) summarize strength and weaknesses of the different fit procedures as follows: "For minimizing prediction error, ridge regression is generally preferable to variable subset selection, principal components regression and partial least squares. However, the improvement over the latter two methods was only slight. (…) PLS, PCR and ridge regression tend to behave similarly. Ridge regression may be preferred because it shrinks smoothly rather than in discrete steps."

Clearly, if the number of potential inputs exceeds the number of data combinations (cases), full MLR or ridge regression won't work, and either one of the constructive methods adding (subsets of) components at the time are implied.

Prediction Error of a Fit

The criterion for model selection cannot be based solely on either of the fit methods mentioned. If we increase the complexity of the model, e.g. by adding components, the error as determined on the training data set will continually decrease and may lead to overfitting. Models that are overfitted tend to generalize poorly.

The training error rate, on the basis of squared error loss, is

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 ,$$

with y_i the measured ith response, and $\hat{f}(x_i)$ the model fit at predictor point (vector) x_i. In general the training error can be made arbitrarily small.

Prediction error approaches to model selection may be based on *analytical* (probabilistic) considerations or *sample re-use* (Hastie et al., 2001, p. 196). Some well-known analytical criteria are Mallows C_p statistic, the Akaike Information Criterion (AIC), and the Bayesian (or Schwartz) Information Criterion (BIC), among others. They relate to likelihood and Bayesian methods of model selection and dimension estimation. A recent, very thorough study of connections between most Bayesian and non-Bayesian model selection methods is presented by Liang (2002).

Sample re-use includes *cross-validation* and *bootstrap*, among other methods.

Cross-validation is extensively used in QSAR modeling (cf. Ericksson et al. 1997, Ericksson et al., 2003). In K-fold cross-validation, the data is divided into K approximately equal parts (Hastie et al., 2001, p. 214). Recommended values are 5 or 10 groups. Each observation is randomly allocated to belong to one of the K groups. Let

$\hat{f}^{-k}(x)$ be the fitted model computed with the *k*th part omitted. Then the cross-validation estimate of prediction error is

$$CV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{-\kappa(i)}(x_i))^2 ,$$

where $\kappa(i)$ is the indexing function associating group *K* with observation *i*. Taking the square root of *CV* retrieves the original unit of *Y*.

So, the fit at the *i*th point does not use that point, nor the other points belonging to its group. The procedure can be repeated by drawing new groups, to estimate the variation of *CV* or its square root.

If the number of groups equals the number of observations, *K* = *n*, this is called *leave-one-out* cross-validation. For the *i*th point the model is fitted to all data except the *i*th. Hence, $\kappa(i) = i$, i.e. each point is its own group, so there is no use in repeating the procedure.

Leave-one-out used to be popular in a non-averaged version, *n.CV*, known as *PRESS* (Prediction Error Sum of Squares) and is still considered a good model selection device (cf. Helsel and Hirsch, 1992, p. 248) and often suggested. Nowadays, 5– or 10–fold cross-validation is recommended over leave-one-out cross-validation (Hastie et al., 2001, p. 215).

PRESS or its grouped version is advocated by Ericksson et al. (1997, 2003) in QSAR modeling, to assess their predictive ability.

If the model depends on some *model complexity* or *tuning* parameter α, which may be the ridge regression parameter, number of predictors or predictor components in either of the regression methods, smoothness of a spline fit, and so on, then

$$CV = CV(\alpha)$$

is the prediction error test curve as a function of model complexity. The modeler may choose the tuning parameter $\hat{\alpha}$ that minimizes it, which is then fitted to all the data.

The one-standard error rule in model selection through cross-validation is to choose the most *parsimonious* (least parameterized) model, whose prediction error is no more than one standard error above the prediction error of the *best* (possibly over-parameterized) model (Hastie et al., 2001, p. 216).

The considerations so far refer to the prediction error of the (best) model fit, which results in a 'deterministic' line or response surface. To evaluate the model prediction uncertainty

including the residual error left un-modeled by the model fit, one has to assess the true variation of response data to be expected when the model is applied in new situations.

Appendix 2. Bayesian Multiple Linear Regression

The mathematics of Bayesian MLR is given in Box and Tiao (1973, p. 113), Tanner (1996, p. 17) and Gelman et al. (2004, p. 355).

In classical (frequentist) statistics an estimate of sigma, MSE, is $\hat{s} = 0.8605$ for the present 1+4 predictor fit.

From a Bayesian point of view the posterior uncertainty of sigma has a scaled inverse Chi distribution, with $df = n - k = 88 - 5 = 83$ the degrees of freedom and scale factor $\sqrt{n-k} \cdot \hat{s}$. Note that the standard deviation is calculated as in the classical case:

$$\hat{s}^2 = (y - \hat{y})^T (y - \hat{y}) / (n - k),$$

with $\hat{y} = X\hat{\beta}$ the best fit (estimate) at predictor matrix X. Here k is the number predictor variables ($k = 5$), including the intercept.

The next graph is a plot of the PDF of the posterior sigma uncertainty. The mode is at $\sqrt{df/(df+1)} \cdot \hat{s} = 0.8553$, a little smaller than the point estimate $\hat{s} = 0.8605$.

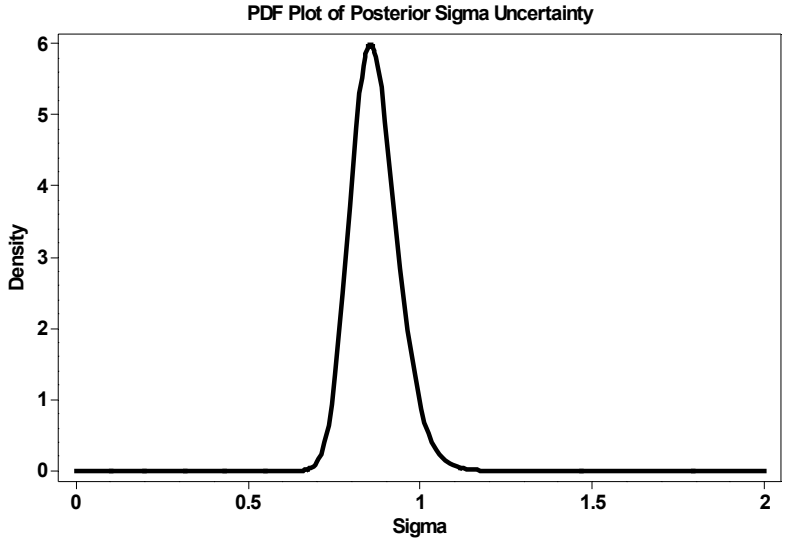

Fig. 15. Plot of the Bayesian posterior uncertainty of sigma estimated in the 1+4 predictor model.

In the present paper, we have summarized the sigma uncertainty to one value: $\sigma = \hat{s} = 0.8605$.

Given σ, the posterior distribution of the MLR coefficients β is

$$\beta \mid \sigma, y \sim N(\hat{\beta}, [(X^T X)^{-1}]^{1/2} \cdot \sigma)$$

in which $N(\mu, \sigma)$ denotes a Normal distribution with mean μ and standard deviation σ. The mean is calculated by the classical (best) fit equation

$$\hat{\beta} = (X^T X)^{-1} X^T y .$$

The predictive uncertainty of an observation y_0 at (new) input X_0, and *given* β and σ, is

$$y_0 \mid \beta, \sigma \sim N(X_0 \beta, \sigma) .$$

With the uncertainty of β incorporated, the predictive uncertainty *given* σ only is

$$y_0 \mid \sigma, y \sim N(X_0 \hat{\beta}, (1 + X_0 (X^T X)^{-1} X_0^T)^{1/2} \cdot \sigma) .$$

We see that the mean is the model best fit at the given input.

This is the classical prediction interval, Bayesian style, well-known in confidence interval form, cf. Helsel and Hirsch (1992, p. 242).
Note that $(1 + X_0 (X^T X)^{-1} X_0^T)^{1/2} \cdot \sigma$ reduces to

$$\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{1/2} \cdot \sigma$$

in simple linear regression (Draper and Smith, 1981, p. 78). If the new input is in the mean: $x_0 = \overline{x}$, then the standard deviation of the prediction is minimal: $\sqrt{(1+1/n)} \cdot \sigma$. The predictive uncertainty limits are relatively insensitive to the distance from the mean, much less than the uncertainty of the best fit.

If the uncertainty of σ is taken into account, the predictive distribution is Student-t (Gelman et al., 2004, p. 359). Drawing individual normal curves with varying β and σ leads to 'spaghetti plots' at given input values (Aldenberg and Jaworska, 2000). The PDF spaghetti plot there can be interpreted as a regression model with only an intercept (the mean of the data), and no predictor variable.

References

Aldenberg, T., and J.S. Jaworska (2000). Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions. *Ecotoxicology and Environmental Safety*, **46**, 1-18.

Box, G.E.P., and G.C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. New York: John Wiley.

Chaloner, K, and R. Brant (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, **75**, 651-659

Cook, R.D. (1998). *Regression Graphics*. New York: John Wiley.

Cook, R.D., and S. Weisberg (1999). *Applied Regression Including Computing and Graphics*. New York: John Wiley.

Debnath, A.K., G. Debnath, A.J. Shusterman, and C. Hansch (1992). A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in Salmonella typhimurium TA98 and TA100. *Env. Mol. Mut.*, **19**, 37-52.

Draper, N.R., and H. Smith (1981). *Applied Regression Analysis*. Second Edition. New York: John Wiley.

Ericksson, L., E. Johansson, and S. Wold (1997). Quantitative structure-activity relationship model validation. In: *Quantitative Structure-Activity Relationships in Environmental Sciences–VII* (F. Chen, and G. Schüürmann, eds.), Pensacola: SETAC Press, p. 381-397.

Ericksson, L., J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, and P. Gramatica (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environmental Health Perspectives*, **111**, 1361-1375.

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

Glende, C, H. Schmitt, L. Erdinger, G. Engelhardt, and G. Boche (2001). Transformation of mutagenic aromatic amines into non-mutagenic species by alkyl substituents. Part I. Alkylation *ortho* to the amino function. *Mutation Research*, **498**, 19-37.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis*. Second Edition. Boca Raton: Chapman & Hall.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. New York: Springer.

Helsel, D.R., and R.M. Hirsch (1992). *Statistical Methods in Water Resources*. Amsterdam: Elsevier.

Liang, F. (2002). Some connections between Bayesian and non-Bayesian methods for regression model selection. *Statistics & Probability Letters*, **57**, 53-63.

Sokal, R.R., and F.J. Rohlf (1995). *Biometry*. Third Edition, New York: W.H. Freeman.

Tanner, M.A. (1996). *Tools for Statistical Inference*. Third Edition. New York: Springer.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.