8 Constructing QSARs for Metal Ions

FROM:
Walker, J., M.C. Newman, and M. Enache. *Fundamental QSARs for Metal Ions.* Taylor & Francis, Boca Raton, FL., 213, p. 288

8.1 SELECTION AND TRANSFORMATIONS OF EXPLANATORY VARIABLES

Defining the mode of action to be modeled is the first step in QSAR generation (McKinney et al. 2000). Chapter 1 covers broadly the various modes of action. The class of relevant toxicants and candidate explanatory variables (descriptors) are identified once the mode of action is defined. Previous chapters describe the potential explanatory variables for metal ions. The final step in QSAR development is the generation of quantitative means for selecting and relating the explanatory variable(s) to the effect of interest. The purpose of this chapter is to provide essential details about this last step.

What are the issues that require consideration while generating a tool to relate metal ion qualities and bioactivity? First, a method must be selected for determining which, and perhaps how many, explanatory variables to use. Second, the best approach must be applied to fit the most appropriate model to these data. What is best will depend on the intended use of the model and the data set qualities. Third, some validation method is applied that specifies how useful the resulting model is relative to the intended predictions of bioactivity for a metal ion not used to generate the model. Again, usefulness will depend on the intended application of resulting predictions. Finally, unique issues must be addressed for building predictive models for metal ion mixtures. The objective of this chapter is to explore these four activities: selecting the best explanatory variables, fitting the appropriate model, assessing predictive value of a model, and modeling metal ion mixtures.

8.2 SELECTION AND ADJUSTMENT OF INDEPENDENT VARIABLES

How does one determine which and how many explanatory variables to include in a model? There is no substitute for a sound understanding of the subject: the statistical methods described below should augment, not supplant, a sound understanding of the chemical, physical, and biological processes that translate metal exposure to bioactivity. How many candidate explanatory variables to consider might depend on the group of cations for which predictions are being made. A few candidate models with a single explanatory variable each might be the focus for some sets of metal ions, such as a group of divalent class (b) and intermediate metal ions. Models incorporating more than one explanatory variable might be thoughtfully explored for more diverse sets, such as ones composed of class (a), intermediate,

and class (b) metal ions with differing charges. Several metal–ligand binding trends might be anticipated to influence metal ion bioactivity for such a heterogeneous set of cations.

8.3 QUANTITATIVE ION CHARACTERISTIC-ACTIVITY RELATIONSHIP (QICAR) MODELS

8.3.1 Intermetal Ion Trends

Models predicting bioactivity of single metal ions based on binding characteristics can be generated for particular subsets of metals of interest or for metal ions in general. No general rules are needed if intent alone determines the subset, that is, relative toxicity of trivalent lanthanides used by the computer display industry. Wolterbeck and Verburg (2001) recommend a periodic table corner calibration of elements if the intent is to produce a more general metal model for a particular effect to a particular biological entity. A series of metals with minimum and maximum values for relevant properties are selected from the periodic table "corners." For example, they select boron, cesium, germanium, lithium, selenium, and uranium for general model calibration. This theme of selecting metals ranging along relevant binding axes is also the natural choice for more narrowly focused studies such as that of the lanthanides just mentioned. In so doing, the distribution of metal ions along the final scale(s) used for the explanatory variable(s) should be as uniform as reasonable to facilitate later regression fitting.

Selecting the most useful model from among a group of candidates is not as simple as fitting a regression and picking the model with the highest coefficient of determination (r^2) value,

$$r^2 = \frac{b^2 \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2} \tag{8.1}$$

where, for a simple model with one explanatory variable, b = slope, X_i and Y_i = the i^{th} X and Y of n data pairs, and $\bar{X}$ and $\bar{Y}$ are averages of the X and Y observations. The coefficient of determination increases as additional explanatory parameters are added to a model, so it will not provide directly useful insight for identifying the model whose parameters contain the most information for making predictions per estimated explanatory variable.

One general scheme for assessing model adequacy and then selecting from among candidate models is provided here. The model adequacy assessment has three initial components: application of subject knowledge to identify candidate models, conventional regression methods, and residual analysis. Additional crucial components involved with gauging predictive adequacy will be discussed later. Approaches to selecting the best model from among candidates can take several forms, including the Minimum Akaike Information Criterion Estimation (MAICE) and Mallows's C_p approaches. Each of these approaches will be described and illustrated. During discussions, it is important that the statistical details not distract

the reader from the central theme that subject knowledge should be the touchstone for decisions at all steps.

Modeling begins by satisfying the basic conditions for regression analysis. It is assumed at the onset that the explanatory variable values are independent of each other. Also, Type I regression techniques formally require no error in the explanatory variable(s), although this requirement is relaxed often to the general premise that the explanatory variable has an immaterial amount of error relative to that of the response variable. Another approach, such as functional regression, might be required if this premise was unjustified. Next, the regression residuals (the difference between the observed response variable value and its predicted value) are assumed to be normally distributed. Finally, the sample variance around the regression line is assumed to be independent of the magnitude of the explanatory variable(s), that is, homoscedasticity is assumed. The last two requirements can be satisfied in some cases by transforming one or more of the variables prior to regression fitting. A common instance in this chapter is the logarithmic transformation of the response variable, such as the log of EC_{50}. This transformation can resolve nonnormality of residuals and heteroscedasticity issues while also conforming to the general toxicological paradigm that response is more often related linearly to the logarithm of dose than to the arithmetic dose (Finney 1942, 1947).

A combination of univariate statistics and plots allow exploration of a candidate model relative to these last two requirements. Bacterial bioluminescence 15-minute EC_{50} data for 20 metal ions (McCloskey et al. 1996, Appendix 8.1) and the recently developed softness index (Kinraide 2009) can be applied to illustrate this approach. The following statistical analysis system (SAS) code implements analyses with normality plots and tests of regression residuals (Figure 8.1 top). It also plots predicted and observed data (Figure 8.1 middle) and regression residuals versus the explanatory variable, σ_{con} (Figure 8.1 bottom).

```
PROC GLM;
  MODEL TOTLEC = SOFTCON;
  OUTPUT OUT = LINEAR2 PREDICTED = PRED2 RESIDUAL = RES2;
  RUN;
PROC UNIVARIATE NORMAL PLOT;
  VAR RES2;
  RUN;
SYMBOL1 V = dot COLOR = black; SYMBOL2 V = star COLOR = black;
SYMBOL3 V = dot COLOR = black;
PROC GPLOT;
  PLOT TOTLEC*SOFTCON PRED2*SOFTCON/OVERLAY HAXIS = -1.5 to
  1.5 by 0.5;
  PLOT RES2*SOFTCON/VREF = 0 HAXIS = -1.5 to 1.5 by 0.5;
RUN;
```

The middle plot in Figure 8.1 shows the values predicted (asterisks) with the model shown in the upper right corner and the original observations (solid dots). The observations are distributed uniformly along the axis of the explanatory variable with no obvious gaps. This minimizes the chance of a few extreme observations having more influence on the model fitting than others. The coefficient of determination, r^2,

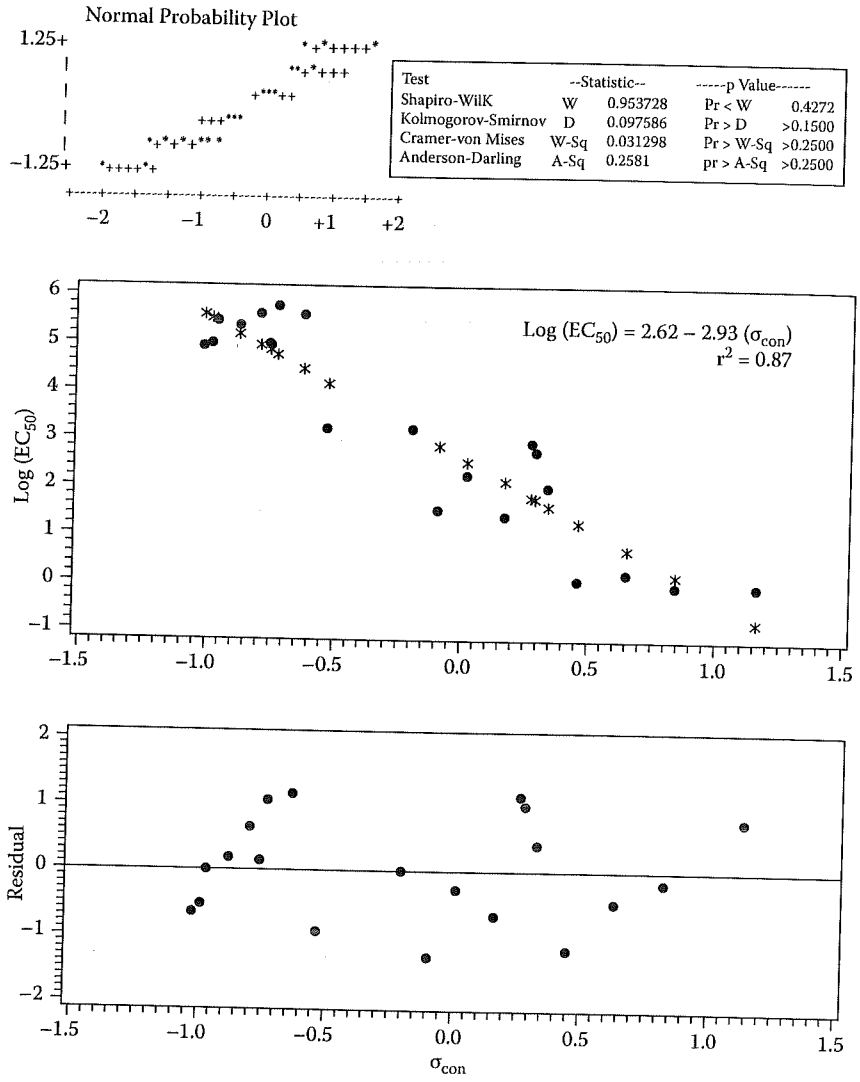
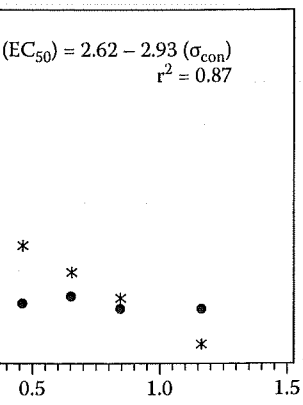
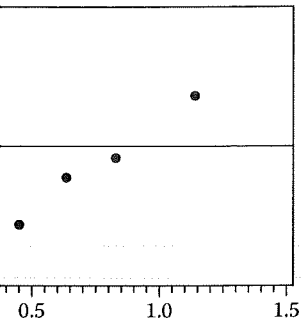


FIGURE 8.1 Bacterial bioluminescence inhibition 15-minute EC_{50} data for 20 metal ions (McCloskey et al. 1996) modeled with Kinraide's softness index (Kinraide 2009). The SAS code listed in Appendix 8.1 produced the normality plots and tests for the regression residuals (top) and also plots of predicted and observed data (middle) and plots of regression residuals versus the explanatory variable, σ_{con} (bottom). (Data from J.T. McCloskey, M.C. Newman, and S.B. Clark. "Predicting the Relative Toxicity of Metal Ions Using Ion Characteristics: Microtox® Bioluminescence Assay. *Environ. Toxicol. Chem.* 15 (1996):1730–1737; and T.B. Kinraide, Improved Scales for Metal Ion Softness and Toxicity. *Environ. Toxicol. Chem.* 28 (2009):525–533.)

--Statistic--		-----p Value------	
W	0.953728	Pr < W	0.4272
D	0.097586	Pr > D	>0.1500
-Sq	0.031298	Pr > W-Sq	>0.2500
-Sq	0.2581	pr > A-Sq	>0.2500





nute EC_{50} data for 20 metal ions
index (Kinraide 2009). The SAS
d tests for the regression residuals
and plots of regression residuals
J.T. McCloskey, M.C. Newman,
l Ions Using Ion Characteristics:
n. 15 (1996):1730–1737; and T.B.
icity. *Environ. Toxicol. Chem.* 28

of 0.87 indicates that the model accounts for approximately 87% of the variation in the response variable, with only 13% of the variability remaining unexplained. The residuals appear to be randomly distributed around the predicted values (bottom panel). The residuals show no pattern if plotted against the explanatory variable with a reference line indicating the state of perfect prediction at any point (residual = 0). There was a random distribution of residuals with no obvious trends left unexplained by the model along the range of values for the explanatory variable. This will be explored more closely later. Further, there was no trend in the amount of variation in the residuals along the abscissa, providing no reason to doubt the assumption of homoscedasticity. In a normal probability plot (top of Figure 8.1) with the distribution of points expected for a normal distribution (asterisks) and positions of the residuals (+ signs), the residuals conform to the assumed normal distribution. Four tests for normality of the regression residuals also provide no evidence of deviation from this assumption (top right of Figure 8.1).

The next task required in QSAR development is selection of the best model. Several approaches are used and range from statistically uninformed judgment of the researcher, MAICE, and Mallows's C_p method. Model selection guided solely by the researcher's informal judgment can produce the best model, but consistency of good judgment is enhanced by application of more formal methods. For example, model selection from among candidate models based only on the smallest χ^2 value will not always produce the best model,

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(Y_i - \hat{Y}_i\right)^2}{\hat{Y}_i}. \tag{8.2}$$

Use of coefficients of determination or χ^2 values *for two models of similar complexity* might be adequate if combined with subject knowledge and the residual plots just described. As an example, such use would be appropriate if the above SAS code fitting the [ln EC_{50i}, σ_{con}] data were also modified to assess the alternative model generated with the more conventional softness index, σ_p. The r^2 for σ_{con} was 0.87 and that for σ_p was 0.81, lending support to Kinraide's (2009) argument that σ_{con} will perform better than the conventional σ_p during model generation. But the model with the most information per fitted explanatory variable cannot be identified with these otherwise useful goodness-of-fit statistics. The r^2 will increase with each addition of an explanatory variable, but the incremental improvement in fit might carry the cost of increased variance in parameter estimates (Hocking 1976). A straightforward change can be made to Equation (8.1) to generate an adjusted r^2 that incorporates the number of explanatory parameters and model degrees of freedom (Hocking 1976; Walker et al. 2003),

$$r^2_{Adjusted} = 1 - \frac{(n-1)\left(1-r^2\right)}{n-p} \tag{8.3}$$

where n = the number of observations, r^2=coefficient of determination, and p=the number of estimated parameters.

Alternatively, criteria can be estimated for each model based on the principle of parsimony, that is, all else being equal, select the simplest model. The Akaike Information Criterion (AIC) is one of the most widely used information criterion that combines the model error sum of squares and the number of parameters in the model,

$$AIC = n\ ln\left(\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}\right) + 2p \tag{8.4}$$

The MAICE approach involves computing AIC values for each candidate model and then selecting the model with the lowest AIC value. The model with the lowest AIC value contains the most information per estimated parameter. Similar criteria, such as the Sawa or Schwarz Bayesian information criteria, can also be applied.

Another similar approach is described by Mallows (1973, 1995), Hocking (1976), Burman (1996), Der and Everitt (2006), and numerous others. Mallows's C_p statistic is estimated as the following,

$$C_p = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{s^2} + 2\mathrm{p} - \mathrm{n} \tag{8.5}$$

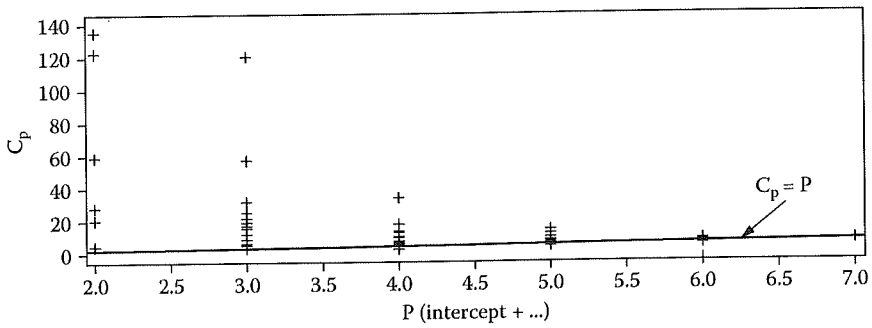
where s^2 = the residual (error) mean square for the model including all available explanatory variables and an intercept. According to Hocking (1976, p. 18), "Cp is an estimate of the standardized total mean squared error of estimation for the current data, X." A set of explanatory variables are selected for possible inclusion and models built with increasing numbers of these variables incorporated. Mallows's Cp statistics are computed for the 2p − 1 possible models and tabulated beginning with models with the highest r^2 and ending with those with the lowest r^2. The best or most parsimonious one, two, three, and more variable models are identified as those with the lowest Cp statistics. Commonly, Mallows's Cp statistics for all 2p − 1 models are plotted against the number of estimated parameters in each model. The line for Cp = p is included in this plot because Cp values close to this line are those of the most parsimonious models.

The following SAS code implements AIC and Mallows's C_p statistic-based model selection for the bacterial bioluminescence inhibition by 20 metal ions (Appendix 8.1) using 6 candidate explanatory variables. The explanatory variables as defined in Newman et al. (1998) are the following: SOFTCON (Kinraide's σ_{con} softness index), ION (ionic index or the square of the ion charge divided by the Pauling ionic radius, Z^2/r), COVAL (covalence index, $\chi^2 r$, the square of the electronegativity times the radius), HYD (|log K_{OH}| where K_{OH} is the first hydrolysis constant of the metal ion), DELE (ΔE_0, the difference in electrochemical potential of the ion and its first stable reduced state), and ANIP (atomic number divided by ΔIP, the difference in ionization potentials for the ion oxidation numbers OX and OX − 1). Both of the PROC GLMSELECT procedures in the SAS code use forward model selection, that is, they begin with a model only fitting an intercept and then progressively

add explanatory variables that most improve the model. Selection in this example is based on the AIC (CHOOSE = AIC) and C_p (SELECT = CP), although simply replacing these specifications at the end of the MODEL line with SELECTION = FORWARD(SELECT = ADJRSQ STOP = SL SLE = 0.2) would permit selection based on an $r^2_{adjusted}$ instead.

```
PROC GLMSELECT;
  MODEL TOTLEC = SOFTCON ION COVAL HYD DELE ANIP/SELECTION =
  FORWARD(SELECT = SL CHOOSE = AIC SLE = 0.2);
RUN;
PROC GLMSELECT;
  MODEL TOTLEC = SOFTCON ION COVAL HYD DELE ANIP/SELECTION =
  FORWARD(SELECT = CP);
RUN;
PROC REG;
  MODEL TOTLEC = SOFTCON ION COVAL HYD DELE ANIP/SELECTION = CP;
  PLOT CP.*NP./CMALLOWS = BLACK;
RUN;
```

The first PROC GLMSELECT in the code uses the AIC statistic to select the combinations of these six variables that produce the most parsimonious model. The model with the lowest AIC was that combining the softness, covalence, and ionic indices as shown in the inset table of Figure 8.2. Adding any of the other variables to the



Model	AIC	Cp
Intercept	53.1	
plus σ_{con}	14.9	4.4
plus $\chi_m{}^2 r$	13.1	2.9
plus Z^2/r	11.3	1.9

FIGURE 8.2 Results from the SAS code that implements AIC and Mallows's C_p statistic-based model selection for the bacterial bioluminescence inhibition by 20 metal ions using 6 candidate explanatory variables (see text for details). The model with the lowest AIC was that including the softness, covalence, and ionic indices (inset table). Application of Mallows's C_p statistic also results in selection of the model containing the softness, covalence, and ionic indices.

models did not reduce the AIC any lower than 11.3. Note, however, that the selection used here involved significance levels for variables (SELECT = SL) and an associated p-value ≤ 0.2 was required (SLE = 0.2) for an explanatory variable to be considered in a model. The second PROC GLMSELECT in the code does the same except it selects models based on Mallows's C_p statistic. Again, the model containing the softness, covalence, and ionic indices was selected. This final model had an r^2 of 0.91 and $r^2_{adjusted}$ of 0.89.

The PROC REG that specifies forward variable selection with the C_p statistic produces a C_p versus p plot (Figure 8.2), and generates a table of models and associated r^2 and C_p values. The best 1, 2, and 3 explanatory variable models are highlighted in Table 8.1. Note that similar $r^2_{adjusted}$-p plots could also have been produced but break points for C_p-p plots tend to be clearer than for $r^2_{adjusted}$-p plots (Hocking 1976).

8.3.1.1 Nonmonotonic Models

It is important to mention at this point in discussions that, based on subject knowledge, some metal ion data sets should not be expected to conform to a monotonic trend. Model development should involve attention to such exceptions. The deviation of K^+ in the metal-valinomycin stability constant versus $AN(\Delta IP)^{-1}$ relationship discussed in Chapter 1 (Figure 1.2) is one important example. A similar example involves acute Ba^{2+} toxicity to the nematode, *Caenorhabditis elegans* (Tatara et al. 1998). This divalent cation is much more toxic than predicted by the general model constructed with 18 mono-, di- and trivalent cations. This could have been anticipated based on an understanding of its interference with K^+ channels and Na^+/K^+-ATPase. The Ba^{2+} has a radius very similar to K^+ but a much higher Z^2r^{-1}. Its bonds with K^+ channel ligand sites are much more stable than those of the K^+. It outcompetes K^+ for binding at the K^+-channels, resulting in blockage of K^+ channels in excitable tissues.

8.3.1.2 Cross-Validation

The procedures described to this point have not assessed model usefulness for prediction. Several approaches allow estimation of predictive usefulness with differing degrees of effectiveness. The most prominent approaches will be described: validation, statistical rules of thumb, and two cross-validations approaches.

The best approach, validation, generates a completely new data set and uses the model generated with the earlier data set to make predictions for these new data. The model is validated if new predictions are close to their corresponding observed bioactivities. Another approach might involve using the new data set to estimate new model parameters and subsequent comparison of those estimates to those generated earlier by fitting the first data set. Comparable estimates suggest good prediction for the first model. Understandably, all data are often pooled into a larger data set after successful validation to generate a final model.

Other approaches can generate acceptable estimates of usefulness if the above approach is not feasible. At the other extreme from the above validation approach is application of a statistical rule of thumb such as the γ_m criterion estimated from

TABLE 8.1
Results of Mallows's Cp Analysis

Number of Parameters	C(p)	r^2	Variables in Model (plus an Intercept)
3	**1.9423**	**0.9083**	**SOFTCON ION COVAL**
3	2.3011	0.9060	SOFTCON COVAL HYD
3	2.4306	0.9051	SOFTCON COVAL DELE
2	**2.9164**	**0.8888**	**SOFTCON COVAL**
4	3.0258	0.9143	SOFTCON COVAL HYD DELE
2	3.3478	0.8859	SOFTCON DELE
4	3.3727	0.9121	SOFTCON ION COVAL DELE
4	3.6318	0.9104	SOFTCON ION COVAL ANIP
3	3.7654	0.8963	SOFTCON COVAL ANIP
4	3.7978	0.9093	SOFTCON COVAL HYD ANIP
4	3.9168	0.9085	SOFTCON ION COVAL HYD
4	4.3910	0.9054	SOFTCON COVAL DELE ANIP
1	**4.4219**	**0.8657**	**SOFTCON**
3	4.7552	0.8898	SOFTCON HYD DELE
2	4.8865	0.8758	SOFTCON HYD
3	4.9336	0.8886	SOFTCON DELE ANIP
5	5.0081	0.9145	SOFTCON ION COVAL HYD DELE
5	5.0134	0.9144	SOFTCON COVAL HYD DELE ANIP
3	5.3152	0.8861	SOFTCON ION DELE
5	5.3324	0.9123	SOFTCON ION COVAL DELE ANIP
4	5.3882	0.8988	SOFTCON ION HYD DELE
2	5.4163	0.8723	SOFTCON ION
5	5.5826	0.9107	SOFTCON ION COVAL HYD ANIP
4	5.6538	0.8971	ION COVAL HYD DELE
3	6.0443	0.8813	ION HYD DELE
3	6.0851	0.8811	COVAL HYD DELE
4	6.1357	0.8939	SOFTCON HYD DELE ANIP
2	6.3207	0.8664	SOFTCON ANIP
4	6.8304	0.8893	SOFTCON ION DELE ANIP
3	6.8336	0.8761	SOFTCON ION HYD
3	6.8852	0.8758	SOFTCON HYD ANIP
4	6.9399	0.8886	ION HYD DELE ANIP
5	6.9701	0.9016	SOFTCON ION HYD DELE ANIP
6	7.0000	0.9145	SOFTCON ION COVAL HYD DELE ANIP
3	7.4085	0.8724	SOFTCON ION ANIP

Note: Only the first 35 models of 2^P-1 or 63 possible models with 6 parameters (P) are tabulated.

F statistics (Draper and Smith 1988). The computed statistic is compared to some arbitrary threshold for model predictive usefulness. Alone, a statistically significant F statistic estimated for a regression model parameter provides only limited insight about how useful any predictions from a model might be; however, some F statistic value greater by a preestablished magnitude from an $F_{(df_m, df_r, 0.95)}$ has served as a tool for separating useful from nonuseful models. The general rule of thumb, (F statistic)/$\left(F_{(df_m, df_r, 0.95)}\right) \geq 4$ to 5 is often applied for this purpose as detailed in Draper and Smith (1988).* For example, the estimated F statistic for the above model predicting bacterial bioluminescence inhibition based on the metal ions' softness (σ_{con}), ionic, and covalence indices was 52.83 and had an associated critical $F_{(3,16,0.95)}$ of 3.24. The resulting (F statistic)/$(F_{(3,16,0.95)})$ = 52.83/3.24 = 16.31 is much greater than 5, suggesting that the model would be a useful one for prediction.

Two cross-validation methods provide a better approach than that just described but generally not as good as the validation method. The first involves splitting of the available data into two subsets and the second involves removal of one datum at a time from the data set prior to building models.

If large enough, a data set can be split into two subsets called the *training* and *validation* sets. This procedure simulates the validation technique by producing a data set not used to build the original model. The disadvantage of this approach is that all available data are not used to generate the model. As a general rule, the number of observations should be at least 6 to 10 times more than the number of explanatory variables in order to successfully apply this approach (Neter et al. 1990). Individual observations can be randomly split between the training and validation sets, but in some cases, a completely random assignment might not be the best approach. For example, it might be preferable to randomly pick observations from within regions along a gradient for some explanatory variable. This ensures that both the training and validation data sets will have observations representing all relevant regions along the gradient.

If the data set (n) is small, one observation can be removed at a time from the data set to produce a data set of size $n-1$, a model is generated with the $n-1$ observations, predictions made with the model for that one removed observation, and the difference between the observed value and predicted value calculated. The removed datum is then placed back into the data set and another datum removed and the above process repeated. This process is repeated for a data set to build n models. Each model has a different observation missing for which predictions are done. Analysis of the n differences between the observed and predicted values (prediction residuals) suggests how useful predictions will be from a model. Prediction residuals can be examined directly or some summary statistic might be generated from the prediction residuals. The following SAS code generates individual prediction residuals and also produces a summary statistic for the bacterial bioluminescence data set listed in Appendix 8.1. Figure 8.3 suggests good prediction (top panel) and no apparent trend in prediction residuals with predicted ln of the EC_{50} (bottom panel).

* The df_m = model degrees of freedom or the number of estimated parameters minus 1, df_r = the residuals degrees of freedom, and 0.95 = 1-α.

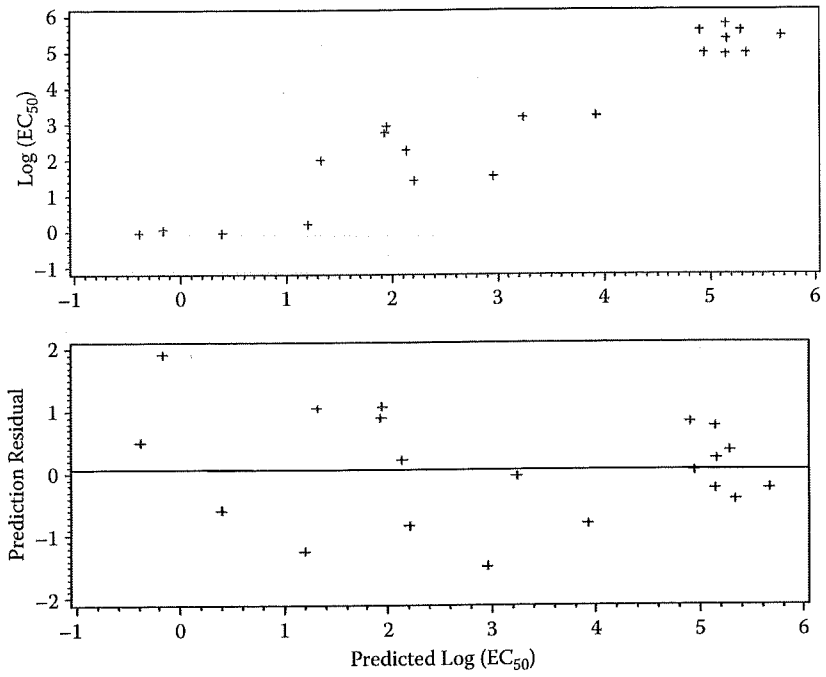


FIGURE 8.3 Good prediction is suggested in this plot of observed Ln (EC_{50}) versus the Ln (EC_{50}) predicted for each point that was omitted from the model (top panel). No trends in prediction quality over the range of predictions were evident in the bottom panel of prediction residuals versus the predicted ln of the EC_{50}.

```
PROC REG PRESS OUTEST = PSOFT;
  MODEL TOTLEC = SOFTCON/AIC BIC;
  OUTPUT OUT = LINEAR PREDICTED = PRED RESIDUAL = RES PRESS =
  PRES;
  RUN;
PROC PRINT;
  VAR TOTLEC PRED RES PRES;
RUN;
PROC PRINT DATA = PSOFT;
RUN;
```

The OUTPUT statement specifies that each prediction residual (PRESS = PRES) be listed in the output. The PROC REG specifies that the following prediction residual sum of squares statistic also be generated,

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 . \tag{8.6}$$

This PRESS statistic and the model total sum of squares (SS_T) can be combined to produce a statistic similar to the coefficient of determination (Equation [8.1]) except

now the variation in expected predictions is quantified for a model that was built without the observation of interest.

$$r^2_{prediction} = 1 - \frac{PRESS}{SS_T}. \quad (8.7)$$

Continuing with the bacterial bioluminescence inhibition example, the model including the softness, covalence, and ionic indices had a *PRESS* of 14.260 and a SS_T of 85.567. The $r^2_{prediction}$ is 1 – (14.269/85.567) or 0.83. This is slightly lower than the r^2 (0.91) or even the $r^2_{adjusted}$ (0.89) that were estimated earlier. Given the ions used to build the model are a representative sample of ions being modeled, the $r^2_{prediction}$ best reflects the amount of variation to be expected in bioactivity predictions for a metal ion not used to build the QICAR model.

8.3.1.3 Metal Interactions

Discussions to this point have focused on predicting bioactivity of a single metal ion in isolation from others; yet, metal ions are often present as mixtures. Models applicable to metal mixtures were described in Chapter 1 (Section 1.3.3) and included those based on the assumptions of either joint independent (Equations [1.3] and [1.4]) or joint similar (Equations [1.6] to [1.9]) action.

The joint similar action model is based on the assumption that probit models (bioactivity versus concentration) for a set of metal ions sharing the same mode of action (and toxicokinetics) will have a common slope (Equations [1.6] and [1.7]). So one simple metric gauging conformity to or deviation from the assumption of joint similar action is the absolute difference in the estimated slopes for two metal ions. To produce Figure 8.1 (bottom panel), maximum likelihood fitting of single metal ion concentration versus proportional inhibition of bacterial bioluminescence to a probit model was done using the SAS package procedure, PROC PROBIT. (Examples of such an application of PROC PROBIT are given for the combined influence of La^{3+} and Ce^{3+} concentration on bacterial bioluminescence in Appendix 8.2.) This was done for each metal separately and the absolute value of the difference in model slopes used to produce the figure. Note that a full factorial experimental design involving a matrix of two metal mixtures is not required. However, as the figure should also make clear, the more involved experimental design required for the approach based on independent action produced clearer metal interaction QICARs in this case.

An estimated interaction coefficient, ρ, was used in Chapter 1 to quantify potential interactions in binary mixtures of metal ions based on the independent joint action model (Figure 1.8, top panel). The associated data set was generated with a matrix of binary metal mixtures. The SAS code in Appendix 8.2 shows an example La^{3+} and Ce^{3+} data set with five La^{3+} concentrations (including 0) combined with five Ce^{3+} concentrations (including 0). The top row of data includes those for different concentrations of Ce^{3+} and no added La^{3+}. The leftmost column of data includes those for different concentrations of La^{3+} and no added Ce^{3+}. All other data reflect mixtures of the two metal ions at different concentrations. The first row and column of data can be used to generate the probit models for the bioactivity of each metal

alone and then the slopes of the two models to be compared as described above under the assumption of similar mode of action. The confidence intervals for slopes of the La^{3+} (2.50 with a 95% confidence interval of 1.26 to 3.73) and Ce^{3+} (3.08 with a 95% confidence interval of 2.23 to 3.94) models suggest no obvious deviation from the assumption of similar action. All of the other data could be used to estimate the interaction coefficient ρ, under the assumption of independent action of the paired metal ions. The resulting interaction coefficient estimate of 1.50 (95% confidence interval: 0.91 to 2.10) provides little evidence for or against this assumption. (The interaction coefficient would be 1 if the metal ions had completely independent action.) Fortunately, the evidence was much clearer in the data sets used to generate Figure 1.8.

8.4 CONCLUSION

The general steps of QSAR production were described at the beginning of this chapter: (1) define the mode of action, (2) define the relevant toxicants sharing that mode of action, (3) define the variables with the most potential for quantifying differences among toxicants, and (4) generate a tool for selecting and relating the explanatory variable(s) to the effect of interest. Most attention was given in this chapter to the last step, although relevant aspects of the other steps were discussed. Approaches to selecting metal ions were identified, including spreading choices along the ranges of candidate explanatory variables, and in the case of a general metal ion QICAR, the periodic table corner calibration method. The coefficient of determination combined with regression residual plots was described as a useful but insufficient means of selecting the best model for making predictions. For candidate models with differing numbers of explanatory variables, MAICE and Mallows's C_p approaches were advocated for picking the model with the most information for making predictions per estimated parameter. Several techniques were described for quantifying how good predictions will be for metal ions not included during model fitting, including the γ_m criterion based on F statistics, model validation with a completely new data set, and cross-validation. The γ_m criterion based on F statistics is useful but involves an arbitrary threshold. Model validation with a completely new data set is ideal but might require more resources than are available to the researcher. Cross-validation involving splitting of a data set into training and validation data subsets simulates validation with new data but requires a relatively large data set. Often, the potential influence of the resulting data set sizes on model generation is assessed by conducting cross-validation twice. Data subset A is the training data and subset B the validation data during the first cross-validation, and then subset A becomes the validation data and subset B becomes the training data during a second cross-validation. How similar the results are for the two cross-validations suggests the influence of data splitting on the cross-validation process. The final cross-validation approach can be applied with smaller data sets. A set of n models are generated by omitting one observation from each model and then comparing predicted values for the omitted observation to the observed values for the observation. Prediction residuals or summary statistics are then generated. Summary statistics might include the PRESS (prediction sum of squares) or $r^2_{prediction}$.

In addition to the methods described above, those useful for quantifying metal ion mixture bioactivities were also discussed. The general models described in Chapter 1 are implemented with specific computer code to illustrate the two potential approaches. The first is based on similar joint action, assumes identical slopes for similar acting metal ions, and measures deviations from identical slopes in order to quantify departure from the assumption of similar action. This involves only probit models for the individual metal ions alone. The second approach, which is based on joint independent action, requires an experimental design in which different concentrations of each metal ion are in mixture with those of the second metal ion. An interaction coefficient, ρ, quantifies deviations from the assumption of independent action.

REFERENCES

Burman, P. 1996. Model fitting via testing. *Stat. Sinica* 6:589–601.

Der, G., and B.S. Everitt. 2006. *Statistical Analysis of Medical Data Using SAS*. Boca Raton, FL: Chapman and Hall/CRC.

Draper, N.R., and H. Smith. 1998. *Applied Regression Analysis, 3rd Edition*. New York: John Wiley & Sons, Inc.

Finney, D.J. 1942. The analysis of toxicity tests on mixtures of poisons. *Ann. Appl. Biol.* 29:82–94.

Finney, D.J. 1947. *Probit Analysis*. Cambridge: Cambridge University Press.

Hocking, R.R. 1976. The analysis and selection of variables in linear regression. *Biometrics* 32:1–49.

Kinraide, T.B. 2009. Improved scales for metal ion softness and toxicity. *Environ. Toxicol. Chem.* 28:525–533.

Mallows, C.L. 1973. Some comments on C_p. *Technometrics* 15:661–675.

Mallows, C.L. 1995. More comments on C_p. *Technometrics* 37:362–372.

McCloskey, J.T., M.C. Newman, and S.B. Clark. 1996. Predicting the relative toxicity of metal ions using ion characteristics: Microtox® bioluminescence assay. *Environ. Toxicol. Chem.* 15:1730–1737.

McKinney, J.D., A. Richard, C. Waller, M.C. Newman and F. Gerberick. 2000. The practice of structure activity relationships (SAR) in toxicology. *Toxicol. Sci.* 56:8–17.

Neter, J., W. Wasserman, and M.H. Kutner. 1990. *Applied Linear Statistical Models*. Homewood, IL: Richard D. Irwin, Inc.

Newman, M.C. 1995. *Quantitative Methods in Aquatic Ecotoxicology*. Boca Raton, FL: Lewis Publishers/CRC Press.

Newman, M.C., J.T. McCloskey, and C.P. Tatara. 1998. Using metal–ligand binding characteristics to predict metal toxicity: Quantitative ion character-activity relationships (QICARs). *Environ. Health Persp.* 106:1419–1425.

Tatara, C.P., M.C. Newman, J.T. McCloskey, and P.L. Williams. 1998. Use of ion characteristics to predict relative toxicity of mono-, di-, and trivalent metal ions. *Caenorhabditis elegans* LC50. *Aquat. Toxicol.* 42:255–269.

Walker, J.D., J.C. Dearden, T.W. Schultz, J. Jaworska, and M.H.I. Comber. 2003. QSARs for new practitioners. In *QSARs for Pollution Prevention, Toxicity Screening, Risk Assessment, and Web Applications*, ed. J.D. Walker, 3–18, Pensacola, FL: SETAC Press.

Wolterbeck, H.T. and T.G. Verburg. 2001. Predicting metal toxicity revisited: General properties vs. specific effects. *Sci. Total Environ.* 279:87–115.

APPENDIX 8.1: SAS BACTERIAL BIOLUMINESCENCE EC_{50} QICAR DATA SET

```
/* This code models ion characteristics against all metals */
/* from Microtox toxicity data - Ba included - McCloskey   */
/* et al. 1996 COVAL is covalence index which reflects the */
/* tendency to form covalent bonds with soft ligands such  */
/* as sulfur. It is the eletronegativity squared times the */
/* radius. ION is Z squared/radius. It is the polarizing   */
/* power or the energy of the metal ion during             */
/* electrostatic interaction with a ligand, SOFT is sigma  */
/* sub p or the softness index. It reflects the tendency   */
/* for the outer electron shell to deform (polarizability) */
/* and the ion's tendency to share electrons with ligands. */
/* ANIP reflects the ionization potential (IP) and inertia */
/* or size (AN). LGANZIP is the log of ANIP that           */
/* Kaiser (1980) preferred to ANIP. DELE this the absolute */
/* difference between the electrochemical potential of the */
/* ion and its first stable reduced state which is a       */
/* measure of the ion's ability to change electronic state. */
/* HYD is the absolute value of the log of the first       */
/* hydrolysis constant which reflects the ion's affinity   */
/* to intermediate ligands such as oxygen donor atoms. The */
/* TOTLEC is the log (base 10) of the EC50 at 15 minutes   */
/* exposure and expressed as total dissolved metal, not the */
/* free ion. This also includes Hg for which the chloride  */
/* species are not considered. Note that SOFTCON was added */
/* as a potentially better softness index. It is the       */
/* computed softness index Sigma Con Comp from             */
/* Kinraide 2009 Env. Tox. Chem. 28:525-533, Table 2.      */

OPTIONS PS = 58;
DATA REVIEW;
    INPUT METAL $ COVAL ION SOFT ANIP LGANIP DELE HYD TOTLEC
    SOFTCON @@;
    CARDS;
HG1+  4.08  3.92  0.065  9.62  0.983  0.91  3.40  -0.037   1.16
CA2+  1.00  4.00  0.181  3.47  0.540  2.76  12.7   4.976  -0.99
CD2+  2.71  4.21  0.081  6.07  0.783  0.40  10.1   1.424   0.17
CU2+  2.64  5.48  0.104  2.31  0.364  0.16  8.00   0.208   0.65
MG2+  1.24  5.56  0.167  1.62  0.210  2.38  11.6   4.941  -1.02
MN2+  1.99  4.82  0.125  3.05  0.484  1.03  10.6   3.196  -0.20
NI2+  2.52  5.80  0.126  2.66  0.425  0.23  9.90   2.753   0.29
PB2+  6.41  3.39  0.131  10.8  1.033  0.13  7.70   0.061   0.46
ZN2+  2.04  5.33  0.115  3.50  0.544  0.76  9.00   1.547  -0.09
CO2+  2.65  5.33  0.130  2.94  0.468  0.28  9.70   2.942   0.27
CR3+  1.71  14.5  0.107  1.66  0.220  0.41  4.00   2.265   0.02
FE3+  2.18  13.9  0.103  1.80  0.255  0.77  2.20   2.009   0.34
CS1+  1.06  0.59  0.218  14.1  1.149  2.92  14.9   5.606  -0.63
K1+   0.93  0.72  0.232  4.38  0.641  2.92  14.5   5.796  -0.73
SR2+  1.02  3.54  0.174  7.12  0.852  2.89  13.2   5.372  -0.88
```

```
LI1+  0.71  1.35  0.247  0.56  -.252  3.05  13.6   5.469  -0.97
NA1+  0.88  0.98  0.211  2.14  0.330  2.71  14.2   5.603  -0.80
BA2+  1.08  2.94  0.183  11.7  1.068  2.90  13.4   4.980  -0.76
LA3+  1.27  8.57  0.171  7.36  0.867  2.37  8.50   3.229  -0.53
AG1+  4.28  0.87  0.074  6.21  0.793  0.80  12.0  -0.034   0.84
;
```

APPENDIX 8.2: SAS BACTERIAL BIOLUMINESCENCE—BINARY METAL MIXTURE EXAMPLE

```
/* LA AND CE ARE THE CONCENTRATIONS OF LANTHANUM AND CERUM. */
/* PAPB IS THE MEASURED BIOLUMINESCENCE AFTER 15 MINUTES OF
   EXPOSURE. */
DATA LACE;
INPUT LA CE PAPB @@;
PAPB = 100*((PAPB-.372)/(1-.372)); NORMZ = 100;
CARDS;
     0 0.372     0 3.125 .359     0 6.25 .385     0 12.50 .481     0 25.00 .662
 3.125 0.333 3.125 3.125 .370 3.125 6.25 .447 3.125 12.50 .533 3.125 25.00 .684
 6.250 0.368 6.250 3.125 .419 6.250 6.25 .449 6.250 12.50 .568 6.250 25.00 .747
 12.50 0.500 12.50 3.125 .548 12.50 6.25 .569 12.50 12.50 .629 12.50 25.00 .761
 25.00 0.667 25.00 3.125 .725 25.00 6.25 .708 25.00 12.50 .757 25.00 25.00 .821
;
DATA LAN; SET LACE; IF CE = 0; RUN;
PROC PROBIT LOG10 INVERSECL LACKFIT DATA = LAN; /* PROC PROBIT A */
  MODEL PAPB/NORMZ = LA/D = NORMAL ITPRINT;
  OUTPUT OUT = PLAN P = PPROB;
RUN;
DATA LAN2; SET PLAN; PAPB = PAPB/100; RUN;
DATA CEN; SET LACE; IF LA = 0; RUN;
PROC PROBIT LOG10 INVERSECL LACKFIT DATA = CEN; /* PROC PROBIT B */
  MODEL PAPB/NORMZ = CE/D = NORMAL ITPRINT;
  OUTPUT OUT = PCEN P = PPROB;
RUN;
DATA NEW;
  SET LACE;
  IF LA NE 0; IF CE NE 0;
  PAPB = PAPB/100;
  LCE = LOG10(CE); LLA = LOG10(LA);
  INTERLA = -3.5687+2.4985*LLA; /* Resulting Model from PROC
                                   PROBIT A */
  PLA = PROBNORM(INTERLA);
  INTERCE = -4.3893+3.0872*LCE; /* Resulting Model from PROC
                                   PROBIT B */
  PCE = PROBNORM(INTERCE);
  EXPECT = PLA+PCE;
  RUN;
PROC GLM DATA = NEW;
  MODEL PAPB = PLA PCE PLA*PCE/CLPARM;
RUN;
```