

CHAPTER 2

Linear Models and Regression

The purpose of models is not to fit the data, but to sharpen the questions
Samuel Karlin (1924), Evolutionary Geneticist

Abstract

A model is said to be linear if the partial derivatives with respect to any of the model parameters are independent of the other parameters. This chapter introduces linear models and regression, both simple linear and multiple regression, within the framework of ordinary least squares and maximum likelihood. Influence diagnostics, conditional models, error in variables, and smoothers and splines are discussed. How to appropriately handle missing data in both the dependent and independent variables is discussed.

Introduction

A model is said to be linear if the partial derivatives with respect to any of the model parameters are independent of the other parameters. All models of the form

$$Y = \theta_0 + \sum_{k=1}^{p-1} \theta_k x_k, \quad (1)$$

where Y is a $n \times 1$ vector of responses called the dependent variable, x is a $n \times 1$ matrix of predictor or independent variables, n is the total number of observations, θ is a $p \times 1$ vector of regression parameters, and p is the number of estimable parameters, are linear because

$$\frac{\partial Y}{\partial \theta_k} = x_k, \quad (2)$$

which does not depend on any other θ_j , $k \neq j$. Much has been written about linear regression models and little will be devoted towards its exposition herein, except for a few general properties of the linear model and a review of some of its salient features. The reader is referred to Neter et al. (1996) or Myers (1986) for further details. The goal is to develop the concepts necessary for the exposition of the nonlinear model, the most common model type seen in pharmacokinetics.

The purpose of a model is explain the behavior of a system and/or to predict the current or future observations. Let

$$Y = \hat{\theta}_0 + \sum_{k=1}^{p-1} \hat{\theta}_k x_k + e_i, \quad (3)$$

and let the predicted value (\hat{Y}_i) be defined as

$$\hat{Y} = \hat{\theta}_0 + \sum_{k=1}^{p-1} \hat{\theta}_k x_k, \quad (4)$$

where $\hat{\theta}$ is the estimator for θ and e are independent, normally distributed residuals with mean 0 and variance σ^2 . In general, the hat-notation, $\hat{\cdot}$, indicates that the value is estimated. By definition, the residuals are calculated as the difference between (3) and (4), i.e.,

$$e = Y - \hat{Y}. \quad (5)$$

It should be noted that for notation purposes, the symbol “ ε ” will be used interchangeably with the symbol “ e ,” although technically “ ε ” is an estimator of “ e .”

The goal is to find the “best” line through the data and consequently find the “best” estimators for θ . One method is to find the set of \hat{Y}_S that are closest to the observed Y based on some type of minimization criterion or objective function. Thus,

$$\hat{\theta} : \min[f(Y, \hat{Y})], \quad (6)$$

where $f(Y, \hat{Y})$ is a specific function based on the observed and predicted values. It should be noted that many different types of objective functions exist. If

$$f(Y, \hat{Y}) = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n e_i^2, \quad (7)$$

then the solution to the minimization problem is the method of ordinary least squares (OLS). The function defined in (7) is called the residual sum of squares or error sum of squares. The use of the word “ordinary” is used to differentiate it from weighted least squares, which will be discussed in the chapter on “Variance Models, Weighting, and Transformations.” For weighted least-squares the objective function is

$$\begin{aligned} f(Y, \hat{Y}) &= \sum_{i=1}^n w_i (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^n w_i e_i^2, \end{aligned} \quad (8)$$

where w_i is the weight associated with the i th data point. A robust procedure for curve fitting is the least-absolute value criterion,

$$f(Y, \hat{Y}) = \sum_{i=1}^n |Y_i - \hat{Y}_i|, \quad (9)$$

sometimes called the L_1 -norm criterion. Most often least squares is used as the minimization criterion because of its statistical properties. Since no pharmacokinetic software package provides alternative objective functions, like the L_1 -norm, only least squares and its modifications will be discussed.

The Method of Least Squares and Simple Linear Regression

The Concept of Ordinary Least Squares Applied to the Simple Linear Model

At the minimum of a function, the first derivative equals zero. In the case of the simple linear regression (SLR) model, $Y = \theta_0 + \theta_1 x + \varepsilon$, where the function being

minimized is the residual sum of squares (7), the following equalities must hold

$$\frac{\partial}{\partial \theta_0} \sum_{i=1}^n [Y_i - (\theta_0 + \theta_1 x)]^2 = 0 \quad (10)$$

$$\frac{\partial}{\partial \theta_1} \sum_{i=1}^n [Y_i - (\theta_0 + \theta_1 x)]^2 = 0.$$

Applying the derivatives, the following pair of equations are obtained

$$n\theta_0 + \theta_1 \sum_{i=1}^n x_i = \sum_{i=1}^n Y_i, \quad (11)$$

$$\theta_0 \sum_{i=1}^n x_i + \theta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i Y_i. \quad (12)$$

These equations are referred to as the least squares normal equations. Solving (11) and (12) simultaneously, θ_0 and θ_1 may be estimated by

$$\hat{\theta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (13)$$

$$\hat{\theta}_0 = \bar{Y} - \hat{\theta}_1 \bar{x}. \quad (14)$$

Intuitively, the concept of least squares makes sense since the predicted model attempts to minimize the squared deviations from the observed values (Fig. 1). Under OLS assumptions, every data point contributes equally to the estimate of the slope and intercept.

The variance of the parameter estimates may then be obtained using the linear expectation rule

$$\begin{aligned} \text{Var}(\hat{\theta}_0) &= \text{Var}(\bar{Y} - \hat{\theta}_1 \bar{x}) \\ &= \text{Var}(\bar{Y}) + \bar{x}^2 \text{Var}(\hat{\theta}_1) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right), \end{aligned} \quad (15)$$

$$\begin{aligned} \text{Var}(\hat{\theta}_1) &= \text{Var} \left(\frac{\sum_{i=1}^n Y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\ &= \frac{1}{S_{xx}^2} \text{Var} \left(\sum_{i=1}^n \sigma^2 (x_i - \bar{x})^2 \right) \\ &= \frac{\sigma^2}{S_{xx}}. \end{aligned} \quad (16)$$

The square roots of $\text{Var}(\hat{\theta}_0)$ and $\text{Var}(\hat{\theta}_1)$ are called the standard error of the parameter estimates denoted as $\text{SE}(\hat{\theta}_0)$ and $\text{SE}(\hat{\theta}_1)$, respectively. The residual variance estimator, $\hat{\sigma}^2$, is estimated by

$$\hat{\sigma}^2 = \text{MSE} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - p}, \quad (17)$$

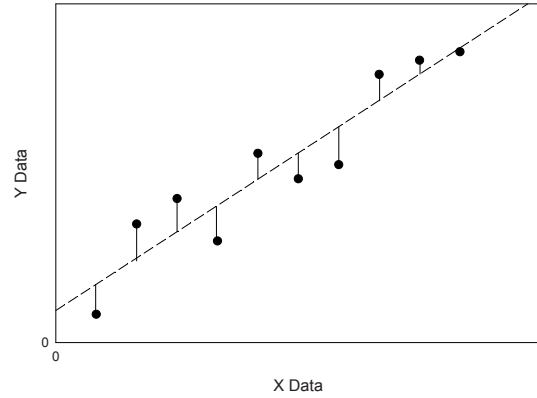


Fig. 1 Illustration of the concept of least squares linear regression. The *dashed line* minimizes the squared deviation (indicated by *solid lines*) between the observed data and the predicted value

where MSE is referred to as the mean square error or residual mean square error. The numerator in (17) is called the residual sum of squares or sum of squares error, while the denominator is called the residual degrees of freedom or simply degrees of freedom. Degrees of freedom is a term that estimates the amount of known information (n) less the amount of unknown information (p). It can be shown that $E(\hat{\sigma}^2) = \text{MSE}$, which means that MSE is an unbiased estimate for the residual variance under the assumption that the model is correct. Actual estimation of (15) and (16) is made using the MSE estimator for the residual variance.

The following assumptions are made with a linear model:

- The x s or independent variables are fixed and known with certainty.
- The residuals are independent with mean zero and constant variance.

When both X and Y are measured with error, this is called error-in-variables (EIV) regression, which will be dealt with in a later section. When x is not fixed, but random and X and Y have a joint random distribution, this is referred to as conditional regression, and will also be dealt with later in the chapter. When the residual's have nonconstant variance, this is referred to as heteroscedasticity, which will be dealt with in later chapters. Under OLS assumptions, the fitted regression line has the following properties:

1. The sum of the residuals equals zero.
2. The sum of the squared residuals is a minimum (hence least squares).
3. The sum of the observed Y values equals the sum of the predicted Y values.
4. The regression line always goes through the point (\bar{x}, \bar{Y}) .

Also under OLS assumptions, the regression parameter estimates have a number of optimal properties. First, $\hat{\theta}$ is an unbiased estimator for θ . Second, the standard error of

the estimates are at a minimum, i.e., the standard error of the estimates will be larger than the OLS estimates given any other assumptions. Third, assuming the errors to be normally distributed, the OLS estimates are also the maximum likelihood (ML) estimates for θ (see below). It is often stated that the OLS parameter estimates are best linear unbiased predictors (BLUE) in the sense that “best” means “minimum variance.” Fourth, OLS estimates are consistent, which in simple terms means that as the sample size increases the standard error of the estimate decreases and the bias of the parameter estimates themselves decreases.

Maximum Likelihood Estimation of Parameters in a Simple Linear Model

Let $\hat{Y}(\hat{\theta}, x)$ be the vector of predicted values for Y . When the errors are normally distributed the likelihood function is given by

$$L(Y | \theta, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[Y_i - \hat{Y}_i]^2}{2\sigma^2}\right). \quad (18)$$

The log-likelihood function is the logarithm of the likelihood and is given by

$$LL(Y | \theta, \sigma) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n [Y_i - \hat{Y}_i]^2. \quad (19)$$

To find the maximum likelihood estimates for θ and σ^2 the log-likelihood must be concentrated with respect to σ^2 . After concentrating the log-likelihood, differentiate with respect to σ^2 , set the derivative equal to zero, solve for σ^2 , and substitute the result back into (19). The concentrated log-likelihood is then maximized with respect to θ .

Differentiating with respect to σ^2 and setting the derivative equal to zero leads to

$$\frac{dLL(Y | \theta, \sigma)}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n [Y_i - \hat{Y}_i]^2 = 0. \quad (20)$$

Solving for σ^2 leads to

$$\sigma^2(\theta) = \frac{\sum_{i=1}^n [Y_i - \hat{Y}_i]^2}{n} \quad (21)$$

where $\sigma^2(\theta)$ denotes the dependence of σ^2 on θ . Substituting back into (19) leads to

$$LL(Y | \theta) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln\left[\frac{\sum_{i=1}^n [Y_i - \hat{Y}_i]^2}{n}\right] - \frac{n}{2}. \quad (22)$$

The middle term in (22) is a function of the residual sum of squares. The first and last terms are constants. Only the middle term in the equation matters for maximization. By minimizing the negative of (22) (which is equivalent to maximizing the log-likelihood function) the maximum likelihood estimate of θ , which is equivalent to the OLS solution, is found. Once θ is found, the maximum likelihood estimate of σ^2 can be found, although the estimate is biased since the choice of denominator (n for maximum likelihood

and $n - p$ for least squares) is different. The same result can be obtained if the likelihood is concentrated with respect to θ first.

The fact that the same result was obtained with the OLS estimates is dependent on the assumption of normality and that the residual variance does not depend on the model parameters. Different assumptions or a variance model that depends on the value of the observation would lead to different ML estimates. Least squares estimates focus completely on the structural model in finding the best parameter estimates. However, ML estimates are a compromise between finding a good fit to both the structural model and the variance model. ML estimates are desirable because they have the following properties (among others):

1. They are asymptotically unbiased.
2. Asymptotically they have minimum variance.
3. They are scale invariant.

For more on the properties and derivation of likelihood functions, the reader is referred to the Appendix given at the end of the book.

Precision and Inference of the Parameter Estimates for the Simple Linear Model

Under normal theory assumptions on the residuals, i.e., $\varepsilon \sim N(0, \sigma^2)$, a $(1 - \alpha)100\%$ confidence interval for $\hat{\theta}_j$ can be computed from

$$\hat{\theta}_j \pm t_{\alpha/2, n-p} \sqrt{\text{Var}(\hat{\theta}_j)}, \quad (23)$$

where t is Student's two-tailed t -distribution with $n - p$ degrees of freedom. A corresponding test for whether a model parameter equals zero (null hypothesis).

$$H_0 : \theta_j = 0$$

vs. the alternative hypothesis that the parameter does not equal zero

$$H_a : \theta_j \neq 0$$

can be made from the $(1 - \alpha)100\%$ confidence interval. If the $(1 - \alpha)100\%$ confidence interval does not contain zero, the null hypothesis is rejected at level α . Similarly, an equivalent T -test can be developed where

$$T = \frac{\text{ABS}(\hat{\theta})}{\text{SE}(\hat{\theta})}, \quad (24)$$

where $\text{ABS}(\cdot)$ is the absolute value function. If T is greater than Student's two-tailed t -distribution with $n - p$ degrees of freedom, then the null hypothesis is rejected. Both the confidence interval approach and the T -test approach produce equivalent results. This latter approach is sometimes referred to as a T -test. For larger sample sizes, the T -test is replaced by a Z -test based on a $N(0,1)$ distribution. For this book, the T -test and Z -test will be used interchangeably.

If θ_j is the slope and the null hypothesis is rejected, then there is evidence to suggest that x affects Y in a linear manner. However, it is unwise to read too much into the

rejection of the null hypothesis for the slope because rejection simply states that there is a trend in the data and speaks nothing to the quality of the fit. θ_j may be rejected but the quality of the regression line is poor, i.e., the model does a poor job at explaining the data. Also, rejection of the null hypothesis says nothing about the ability to predict future observations.

Regression Through the Origin

Sometimes the regression model is linear and is known to go through the origin at the point (0,0). An example may be the regression of dose against area under the curve (AUC). Obviously, when the dose of the administered drug is zero then the AUC should be zero as well. In this case, x becomes a $n \times 1$ matrix of predictor variables with the column of ones removed and for the SLR model, the model reduces to $Y = \theta_1 x + \varepsilon$. The solution to the SLR model is

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \quad (25)$$

with variance estimate

$$\text{Var}[\hat{\theta}_1] = \frac{\text{MSE}}{\sum_{i=1}^n x_i^2}. \quad (26)$$

Regression through the origin is presented here because of a number of peculiarities to the model, some of which may be unfamiliar to pharmacokineticists. First, the residuals may not necessarily sum to zero and a residual plot may not fall around the zero line. But $\sum_{i=1}^n x_i e_i = 0$ and, thus, a residual plot using $x_i e_i$, instead of e_i , may be of more use. Second, it may be possible for the coefficient of determination to be negative because sometimes the residual sum of squares may be greater than the total sum of squares, an event that may occur if the data are curvilinear. Hence, the coefficient of determination is a meaningless statistic under this model. Third, confidence intervals for predicted values will increase in range as x_0 , the value to be predicted, becomes removed from the origin, as opposed to the confidence intervals typically seen with SLR. Neter et al. (1996) suggest that using a regression through the origin model is not “safe practice,” that an intercept model always be used. They argue that if the regression line does go through the origin, then θ_0 will be very close to zero using an intercept model, differing only by a small sampling error, and unless the sample size is small there will be no deleterious effects in using an intercept model. But if the regression line does not go through the origin and a no-intercept model is used, the resulting model may be quite biased.

An example of a model that perhaps should have used a no-intercept model is the so-called Calvert formula used to dose carboplatin, a platinum-containing oncolytic agent used to treat a wide range of tumors. Carboplatin is primarily excreted through the kidney by filtration. Calvert et al. (1989)

developed a semimechanistic model from 18 adult patients to dose carboplatin. Since clearance (CL) is the ratio of dose to AUC then

$$\text{Dose} = \text{CL} \times \text{AUC}. \quad (27)$$

Using linear regression the authors estimated that

$$\text{CL in mL/min} = 1.21 \times \text{GFR in mL/min} + 23 \quad (28)$$

where GFR is the glomerular filtration rate for the patient estimated using $^{51}\text{Cr-EDTA}$ clearance. Hence, a suitable dosing equation (after rounding) was

$$\text{Dose in mg} = (1.2 \times \text{GFR in mL/min} + 20) \times \text{AUC in mg/(mL min)}, \quad (29)$$

where the target AUC was 3 mg min/mL. However, a quick examination of their parameter estimates shows that the standard error associated with the intercept was 16. A T -test for this parameter was 1.44 with a corresponding p -value of 0.39. Also, examination of Fig. 1 in the paper shows that the 95% confidence interval for the regression of GFR against carboplatin CL crosses the ordinate when GFR equals zero. The authors interpreted the intercept as the degree of nonrenal elimination, but the intercept in this example was not statistically different from zero and by all accounts should have been removed from the model and a no-intercept model used instead. Perhaps the authors are correct and the intercept does reflect nonrenal clearance and that with a larger sample size the standard error of the intercept will be reduced making its estimation more precise. Based on the data at hand, however, a no-intercept model appeared to be more appropriate in this case. This issue will be revisited in the chapter on “Case Studies in Linear and Nonlinear Regression,” when a similar equation will be developed in children.

Goodness of Fit Tests for the Simple Linear Model

As just mentioned, the T -test tests the significance of a particular parameter estimate. What is really needed is also a test of the overall significance of a model. To start, the total sum of squares of the observed data, SS_{total} , is partitioned into a component due to regression, $\text{SS}_{\text{regression}}$, and a component due to residual, unexplained error, SSE ,

$$\sum_{i=1}^n (Y - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y} - \bar{Y})^2 + \sum_{i=1}^n (Y - \hat{Y})^2. \quad (30)$$

$$\text{SS}_{\text{total}} = \text{SS}_{\text{regression}} + \text{SSE}.$$

Equation (30) can be seen conceptually as

$$\left(\begin{array}{c} \text{Total Variability} \\ \text{of the Observations} \end{array} \right) = \left(\begin{array}{c} \text{Variability explained} \\ \text{by the model} \end{array} \right) + \left(\begin{array}{c} \text{Unexplained} \\ \text{Variability} \end{array} \right) \quad (31)$$

Equally, terms on the right-hand side of (30) can be viewed as variability due to the regression line and variability around the regression line. Clearly, a good model is one where $\text{SS}_{\text{regression}} \gg \text{SSE}$. Assuming that the residuals are independent and normally distributed with mean 0 and variance σ^2 , a F -test can be computed to test the null hypothesis that $\theta = 0$,

$$F = \frac{[SS_{\text{regression}} / 1]}{[SSE / (n - p)]} = \frac{SS_{\text{regression}}}{MSE}. \quad (32)$$

Under the null hypothesis, F is distributed as an F -distribution with $p, n - p$ degrees of freedom. If $F > F_{p, n-p, \alpha}$ the null hypothesis is rejected. This is called the analysis of variance approach to regression. The power of this approach comes in when multiple covariates are available (see “Multiple Linear Regression” section later in this chapter). The F -test then becomes an overall test of the “significance” of the regression model.

One of the most commonly used yardsticks to evaluate the goodness of fit of the model, the coefficient of determination (R^2), develops from the analysis of variance of the regression model. If SS_{total} is the total sum of squares then

$$R^2 = \frac{SS_{\text{regression}}}{SS_{\text{total}}}. \quad (33)$$

The correlation coefficient is the square root of R^2 . These metrics have been discussed in greater detail in the previous chapter.

Prediction and Extrapolation in the Simple Linear Model

The goal of regression analysis is usually twofold. First, a model is needed to explain the data. Second, using the model, predictions about mean responses or future observations may be needed. The distinction between mean responses and future observations must be clarified. Mean responses are based on already observed data. Future observations are unobserved. The confidence interval for a future observation should be wider than that of a mean response because of the additional uncertainty in future observations compared to the known observation. Now, let $\hat{Y}(x_0)$ be the estimated response (or expected value) given x_0 is

$$\hat{Y}(x_0) = \hat{\theta}_0 + \hat{\theta}_1 x_0. \quad (34)$$

The standard error for $\hat{Y}(x_0)$ is interpreted as the standard error the mean response conditional on x_0 . Thus, the variance of $\hat{Y}(x_0)$ is

$$\text{Var}[\hat{Y}(x_0)] = \text{Var}[\hat{\theta}_0 + \hat{\theta}_1 x_0] \quad (35)$$

and using the estimate for σ^2 , the estimated standard error of prediction is

$$\text{SE}[\hat{Y}(x_0)] = \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]} \quad (36)$$

with a corresponding $(1 - \alpha)100\%$ confidence interval given by

$$\hat{Y}(x_0) \pm t_{\alpha/2, n-p} \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}. \quad (37)$$

Note that the standard error of prediction is not a constant for all values of x_0 but reflects where x_0 is collected in relation to the mean. Observations removed from the mean

of x will have larger standard errors of prediction than values close to the mean. Equation (37) is developed as the confidence interval for a single observation measured at x_0 . If more than one observation is observed at x_0 , the term $1/n$ in (36) and (37) is substituted with the term m/n , where m is the number of observations at x_0 . Note that m is contained within n . If the confidence interval is made for all points on the regression line, the result would be a confidence band.

The confidence interval for a future response, one not in the original data set, must be more variable due to the additional uncertainty in its measurement. Thus, (36) is modified to

$$\text{SE}[\hat{Y}(x_0)] = \sqrt{\text{MSE} \left[\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}, \quad (38)$$

where m is the number of future observations to be collected. The corresponding prediction interval is

$$\hat{Y}(x_0) \pm t_{\alpha/2, n-p} \sqrt{\text{MSE} \left[\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}. \quad (39)$$

Clearly prediction intervals are wider than corresponding confidence intervals.

To illustrate further the distinction between confidence intervals for mean responses and prediction intervals for future observations, consider allometric scaling. In allometric scaling, the systemic clearance or volume of distribution is calculated for many different species, usually mouse, rat, and dog. A regression line of the log-transformed pharmacokinetic parameter is regressed against the log-transformed weight. One may then ask “What is the 95% confidence interval for clearance in the rat?” This is an example of confidence interval using (37). Next, someone may ask “If 10 rats were taken from the population, what is the 95% confidence interval for clearance in the rat?” This is another example of a confidence interval using (37) with the term $1/n$ replaced by $1/10$. Then, someone may ask “What is the 95% prediction interval for clearance in a guinea pig?” This is an example of a prediction interval using (39) because guinea pigs were not in the original population. A similar question can be asked about humans – what is the clearance in humans given a dataset based entirely on animal data. This approach, called prospective allometric scaling, is often used in choosing the starting dose for a new drug in a first time in man study. Bonate and Howard (2000) argue that prospective allometric scaling can lead to unreasonably large confidence intervals because the extrapolation from animals to humans, based on body weight, is tremendous and that using this approach in practice should be done with great care. A further example of allometric scaling is presented in the chapter on “Case Studies in Linear and Nonlinear Modeling.”

An amusing report of extrapolation, and the pitfalls thereof, is presented by Tatem et al. in the journal *Nature*. The authors plotted the winning times of the men’s and women’s Olympic 100 m finals for the past 100 years. A linear model was able to adequately describe the relationship for both males and females. In both cases, males and females

are getting faster over time. In 1932, males and females had a finish time of 10.3 and 11.9 s, respectively. By 2000, the times had decreased to 9.85 and 10.75 s, respectively. Females, however, are improving at a faster rate than males, and the authors speculated that “should these trends continue, the projections will intersect at the 2156 Olympics, when – for the first time ever – the winning women’s 100 m sprint time of 8.079 s will be lower than the men’s winning time of 8.098 s” (Fig. 2). The authors themselves question whether the trend will indeed continue but it is nevertheless an amusing example of extrapolation.

An even more amusing example of extrapolation was reported by Mark Twain in 1874. He said

In the space of one hundred and seventy six years the Lower Mississippi has shortened itself two hundred and forty-two miles. That is an average of a trifle over a mile and a third per year. Therefore, any calm person, who is not blind or idiotic, can see that in the Old Oölitic Silurian Period, just a million years ago next November, the Lower Mississippi was upwards of one million three hundred thousand miles long, and stuck out over the Gulf of Mexico like a fishing-pole. And by the same token any person can see that seven hundred and forty-two years from now the Lower Mississippi will be only a mile and three-quarters long, and Cairo [Illinois] and New Orleans will have joined their streets together and be plodding comfortably along under a single mayor and a mutual board of aldermen. There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact.

Both of the Olympics example and Twain’s quote illustrate the risks one takes when extrapolating. In both cases, the results lead to an absurd result, although to be fair, the Olympics example may indeed come true. While making predictions outside the bounds of an observed dataset has its uses, blind extrapolation needs to be cautioned against.

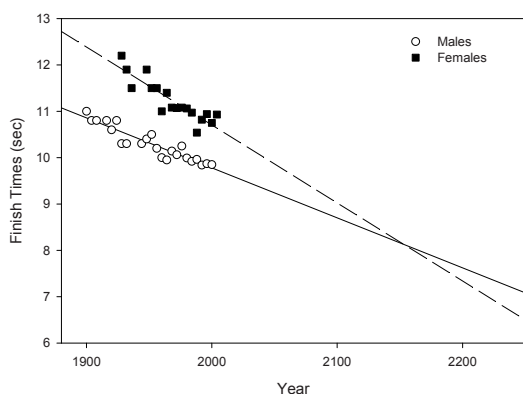


Fig. 2 Plot of winning 100 m sprint times in the Olympics for males (open circles) and females (closed squares), superimposed with the linear regression lines, for the twentieth century

Categorical Independent Variables

Up until now it has been assumed that x consists of continuous variables. OLS is not predicated on x being continuous, although this makes it convenient to explain the model. An extremely important data type is a categorical variable where the variable of interest takes on discrete values. These variables are also called factors or class variables. For instance, whether a person is considered a *smoker* can be coded as either “yes” or “no.” The variable *race* may take on the values: White, Black, Asian, or Hispanic. These variables must enter the model through what are called dummy variables or indicator variables which are themselves categorical variables that take on the value of either 0 or 1. If there are k levels in the categorical variable, then $k - 1$ dummy variables are needed to uniquely define that variable. For example, the variable *smoker* has two levels and thus needs a single dummy variable (0 or 1) to define that variable.

In general, there are three different types of coding dummy variables for nominal variables. One is reference cell coding, which is the most common, where one category serves as the reference cell (such as a placebo group) and all other categories are interpreted relative to the reference cell (such as active treatment groups). For example, suppose the categorical variable *race* has four levels: White, Black, Hispanic, and Other. Three dummy variables (D1 – D3) are needed to uniquely define that variable. In reference cell coding, using White as the reference cell, the categories can be defined as:

Variable: race	Dummy variables		
	D1	D2	D3
White	0	0	0
Black	1	0	0
Asian	0	1	0
Hispanic	0	0	1

Another type of coding is deviation from the means coding whereby the contrast compares the “group mean” from the “overall mean.” This coding is accomplished by setting all the design variables equal to -1 for one of the groups and then coding the other groups as 0 or 1. So, returning to the race example, the deviation from the mean coding schema is:

Variable: race	Dummy variables		
	D1	D2	D3
White	-1	-1	-1
Black	1	0	0
Asian	0	1	0
Hispanic	0	0	1

A modification of reference cell coding is incremental effects coding, where one group is the reference and all other categories are coded as increments from the prior group. So the design matrix in the race example would be

Variable: race	Dummy variables		
	D1	D2	D3
White	1	0	0
Black	1	0	1
Asian	1	1	0
Hispanic	1	1	1

Lastly, if the categorical variable is ordinal then orthogonal polynomials, which are typically used to assess trends in the analysis of variance models, could be used to code the design matrix. The advantage of orthogonal polynomials is that they provide a test for whether the logit has a significant linear, quadratic, cubic, etc. component. So, suppose that weight was categorized into four variables. The design matrix could be coded as:

Variable: weight	Dummy variables		
	D1	D2	D3
70 kg or lower	-0.67	0.5	-0.22
70 – 80 kg	-0.22	-0.5	0.67
80 – 90 kg	0.22	-0.5	-0.67
90 kg or higher	0.67	0.5	0.22

If the coefficient associated with D1 were statistically significant based on a T -test or Wald's test then this would be indicative of a linear trend. If D2 were significant, this would be indicative of a quadratic trend and so on.

The presence of a dummy variable results in a shift in the regression through its effect on the intercept (Fig. 3). The difference between the regression lines is an indication of the difference between the levels of the variable assuming that the regression coefficients for the continuous variables across classes remain constant among the factor levels. Also, note that the inferential statistics on the

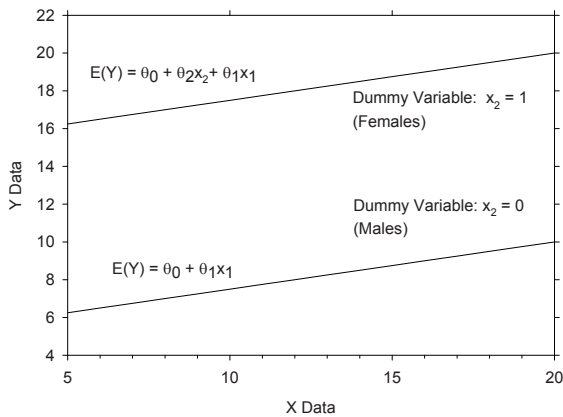


Fig. 3 Plot of regression line for a single categorical covariate (sex) with two levels (males and females). The effect of the categorical variable is to shift the model intercept

regression parameters, even the regression estimates themselves, are independent of how the factor levels are coded. For instance, with variable *sex* it makes no difference whether “males” are coded as 0 or 1 as long as “females” are coded 1 or 0, respectively.

Multiple Linear Regression

Rarely in a single experiment is one dependent variable and one independent variable collected. More often, many dependent variables and many independent variables are collected. Then, a scientist may wish to use the independent variables to explain a particular dependent variable. For example, suppose from a population pharmacokinetic analysis (which will be discussed in later chapters) total systemic clearance (CL) was estimated in a group of subjects. Also available were demographic information, such as age, weight, and smoking status. Of interest would be whether any of the demographic variables were related to clearance. It may be that smokers have higher clearance estimates than nonsmokers and require more drug to achieve the same therapeutic effect.

In this case, multiple linear regression may be used to determine the significance of the demographic variables, which are often called covariates. The model may then be formulated as

$$CL = \theta_0 + \theta_1 \text{Weight} + \theta_2 \text{Age} + \theta_3 \text{Smoker} + \varepsilon. \quad (40)$$

As in SLR, the same assumptions are made: ε_i is normally distributed, uncorrelated with each other and have mean zero with variance σ^2 . In addition, the covariates are measured without error. In matrix notation then, the general linear model can be written as

$$Y = x\theta + \varepsilon, \quad (41)$$

with solution

$$\hat{\theta} = (x^T x)^{-1} x^T Y. \quad (42)$$

In this case, x is a $n \times (k+1)$ matrix of independent variables where the first column of the matrix is a column of ones, which is necessary for inclusion of the intercept in the model, and k is the number of independent variables. An estimate of MSE is obtained by

$$MSE = \frac{(Y - x\hat{\theta})^T (Y - x\hat{\theta})}{n - p} = \frac{\sum_{i=1}^n (Y_i - x_i\hat{\theta})^2}{n - p}, \quad (43)$$

which is exactly the same as (17), but written in matrix notation. The standard error of the parameter estimates is calculated by

$$SE(\hat{\theta}) = \sqrt{\text{diag}(x^T x)^{-1} MSE} \quad (44)$$

where $\text{diag}(\cdot)$ is the diagonal elements of $x^T x$. Similarly, T -tests and confidence intervals for the parameter estimates can be calculated using (24) and (23), respectively. $(1 - \alpha)100\%$ confidence intervals for mean responses can be computed from

$$\hat{Y}(x_0) \pm t_{\alpha/2, n-p} \sqrt{MSE[x_0^T (x^T x)^{-1} x_0]}, \quad (45)$$

and $(1 - \alpha)100\%$ prediction intervals for future responses can be calculated from

$$\hat{Y}(x_0) \pm t_{\alpha/2, n-p} \sqrt{\text{MSE}[1 + x_0^T (x^T x)^{-1} x_0]}, \quad (46)$$

Similarly, a $(1 - \alpha)100\%$ confidence band for the response function at any x can be developed using

$$x\hat{\theta} \pm \sqrt{\text{MSE}[1 + x^T (x^T x)^{-1} x]} \sqrt{F_{p, n-p, \alpha} p}. \quad (47)$$

Confidence bands differ from confidence intervals in that they consider all the values of x simultaneously, as opposed to a single value x_0 . Confidence bands are larger than confidence intervals.

Model Selection and Sequential Variable Selection Procedures in Multiple Linear Regression

Even though many different covariates may be collected in an experiment, it may not be desirable to enter all these in a multiple regression model. First, not all covariates may be statistically significant – they have no predictive power. Second, a model with too many covariates produces models that have variances, e.g., standard errors, residual errors, etc. that are larger than simpler models. On the other hand, too few covariates lead to models with biased parameter estimates, mean square error, and predictive capabilities. As previously stated, model selection should follow Occam's Razor, which basically states "the simpler model is always chosen over more complex models."

To strike the proper balance between an over-parameterized model and an underparameterized model, one must strike a balance between a biased model and an overinflated variance model. Mallows (1973) proposed his C_p criterion which is defined as

$$C_p = \frac{\text{SSE}^*}{\text{MSE}} - (n - 2p^*), \quad (48)$$

where SSE^* is the sum of squares error from the model containing p^* parameters, where $p^* \leq p$. When $p^* = p$, then $C_p = p$. For example, if a model with four possible covariates is examined, the submodel with covariates $\{x_1, x_2\}$ becomes

$$C_p = \frac{\text{SSE}(x_1, x_2)}{\text{MSE}} - (n - 6). \quad (49)$$

When there is no bias in the model, the expected value of C_p is p^* , the number of parameters in the model. Thus, when C_p is plotted against p^* , models with little bias will fall near the line $C_p \cong p^*$. Models with substantial bias will have C_p values greater than the line. In using Mallows C_p as a model selection criterion one chooses a C_p that is small and near p^* .

One way to identify important predictor variables in a multiple regression setting is to do all possible regressions and choose the model based on some criteria, usually the coefficient of determination, adjusted coefficient of determination, or Mallows C_p . With this approach, a few candidate models are identified and then further explored for residual analysis, collinearity diagnostics, leverage analysis, etc. While useful, this method is rarely seen in the literature and cannot be advocated because the method is a

"dummy-ing down" of the modeling process – the method relies too much on blind usage of the computer to solve a problem that should be left up to the modeler to solve.

Related to all possible regressions, a variety of automated algorithms have been developed to screen a large number of covariates in a multiple regression setting and select the "best" model. Forward selection algorithms begin with no covariates in the model. Each covariate is then screened using SLR. F -tests are then calculated reflecting each covariate's contribution to the model when that covariate is included in the model. These F -tests are then compared to a significance level criteria (F_{in}) set by the user a priori and if the F -tests meets F_{in} the covariate is included in the model. At each step only one covariate is added to the model – that covariate having the highest contribution to the F -test. For example, suppose $\{x_1, x_2, x_3, x_4\}$ were possible covariates and using SLR x_3 was found to be the most significant covariate based on the F -test. The next step then compares the models $\{x_1, x_3\}$, $\{x_2, x_3\}$, and $\{x_3, x_4\}$. The contribution x_1, x_2 , and x_4 make to their respective models is then compared and the covariate having the highest contribution is compared to F_{in} . The new variable is then added to the model if that F -test meets the entry criteria. If in this case, that variable was x_1 , then the next models tested will be $\{x_1, x_3, x_2\}$ and $\{x_1, x_3, x_4\}$. This process repeats until no further variables are available or until the model with the highest contribution does not meet the entry criteria, at which point the algorithm stops.

Backward elimination is similar to forward selection, except that the initial model contains all the covariates and removal from the model starts with the covariate of the least significance. Removal from the model then proceeds one variable at a time until no covariates meet the criteria for removal (F_{out}). Stepwise regression is a blend of both forward and backward selection in that variables can be added or removed from the model at each stage. Thus, a variable may be added and a variable may be removed in the same step. The algorithm quits when no additional covariates can be added on the basis of F_{in} and no covariates can be removed on the basis of F_{out} .

The problem with using all possible regressions or sequential methods is that they lead to the "dumbing down" of statistical analysis. The user plugs in some data and the computer spits out a "best model." Simply because a software manufacturer includes an algorithm in a package does not mean it should be used. Scientific judgment must play a role in covariate selection and model selection. Explanatory covariates should be based on physiological or physical sense. As an example, suppose volume of distribution were screened against clinical chemistry laboratories and inorganic phosphate was identified as a significant covariate. How does one interpret this? It is better to use a priori covariates that make sense in the model and then build on that model. As a rule, sequential variable selection procedures and all possible regressions should be used with caution. Harrell presents some very valid criticisms of

stepwise regression and all possible subsets regression. They are:

1. The coefficient of determination is often biased high.
2. The F - and chi-squared distribution next to each variable do not have the prescribed theoretical distribution.
3. Confidence intervals for effects and predicted values are too narrow.
4. p -Values do not have the proper meaning anymore because of multiplicity.
5. The regression coefficients are biased.
6. The algorithm has problems with collinearity.
7. It is based on methods, i.e., F -tests for nested models, that were designed to test prespecified hypotheses.
8. Increasing the sample size does not improve things.
9. It is too easy to use and causes people to quit thinking about their problem.
10. It uses a lot of paper.

In summary, automated techniques should not be used blindly, even though they often are.

Collinearity and Ill-Conditioning

When multiple covariates are included in the regression model, the possibility for collinearity, which is sometimes called multicollinearity or ill-conditioning, among the predictors arises. The term collinear implies that there is correlation or linear dependencies among the independent variable. Entire books (Belsley et al. 1980) have been written on collinearity and all its nuances will not be discussed in its entirety here. Nevertheless, an analyst should at least understand what it is, how to detect it, and how to combat it.

Collinearity is actually simple to understand, although there are complex geometric reasons for its effect on parameter estimation. Consider two variables x_1 and x_2 that are regressed against Y . Now suppose x_1 and x_2 are correlated to the extent that they essentially are the same thing. Thus, x_2 does not provide any more information than x_1 and vice versa. As the correlation between x_1 and x_2 increases, it becomes more and more difficult to isolate the effect due to x_1 from the effect due to x_2 , such that the parameter estimates become unstable. The bottom line is when collinearity exists among a set of predictors, the parameter estimates become extremely sensitive to small changes in the values of the predictors and are very much dependent on the particular data set that generated them. A new data set may generate completely different parameter estimates. Although collinearity is often due to correlation between variables, collinearity may be due to a few influential observations and not necessarily to the whole vector of data. Careful examination of the scatter plots between possible correlated variables should be done to rule out this cause of collinearity.

Collinearity manifests itself during the inversion of the matrix $x^T x$ in (42), such that small changes in x lead to large changes in the parameter estimates and their standard errors. When the predictors are uncorrelated, the values of the parameter estimates remain unchanged regardless of any other predictor variables included in the model. When the predictors are correlated, the value of a regression parameter depends on which other parameters are entered into the model and which others are not, i.e., collinearity destroys the uniqueness of the parameter estimate. Thus, when collinearity is present a “regression coefficient does not reflect any inherent effect of the particular predictor variable on the response variable but only a marginal or partial effect, given whatever other correlated predictor variables are included in the model” (Neter et al. 1996). Correlation between predictor variables in and of itself does not mean that a good fit cannot be obtained nor that predictions of new observations are poorly inferred, provided the inferences are made within the sample space of the data set upon which the model was derived. What it means is that the estimated regression coefficients tend to widely vary from one data set to the next.

There are a variety of methods to detect collinearity (Belsley et al. 1980). First, examine the parameter estimates. A priori variables that are expected to be important which are not found to be statistically significant is a clue that collinearity may be present. If the values of the parameters change drastically if a row of x or column of x is deleted (such as a sign change), that is another clue. Second, examine the various collinearity diagnostics, of which there are many, some of which are better than others. Keep in mind, however, that there are no definitive cut-off values indicating whether collinearity is present.

The first simple diagnostic is to examine the correlation matrix of the covariates. High correlations, either positive or negative, are indicative of collinearity. However, the correlation matrix is sometimes unable to detect the situation where three or more covariates are collinear but no two correlations are high (Belsley et al. 1980). Related to the inverse of the correlation matrix are variance inflation factors (VIF), calculated as

$$VIF = \frac{1}{1 - R_i^2}, \quad (50)$$

where R_i^2 is the coefficient of determination of x_i regressed against all other x . The higher the coefficient of determination, the higher the VIF, and the greater the collinearity. Possible collinearity is present when the VIF is greater than 5 and multicollinearity is almost certainly occurring when the VIF is greater than 10.

Another useful tool is to examine the eigenvalues of the $x^T x$ matrix, l_i . The number of eigenvalues near zero indicate the number of collinear covariates among the regressors. One of the most commonly used yardsticks to measure the degree of collinearity is the condition number (K), which can be calculated using many different methods.

The first definition is simply the ratio of the largest to smallest eigenvalue

$$K = \frac{l_1}{l_p}, \quad (51)$$

where l_1 and l_p are the largest and smallest eigenvalues of the correlation matrix (Jackson 1991). The second way is to define K as

$$K = \sqrt{\frac{l_1}{l_p}}. \quad (52)$$

The latter method is often used simply because the condition numbers are smaller. The user should be aware how a software package computes a condition number. For instance, SAS uses (52). For this book (51) will be used as the definition of the condition number. Condition numbers range from 1, which indicates perfect stability, to infinity, which indicates perfect instability. As a rule of thumb, $\text{Log}_{10}(K)$ using (51) indicates the number of decimal places lost by a computer due to round-off errors due to matrix inversion. Most computers have about 16 decimal digits of accuracy and if the condition number is 10^4 , then the result will be accurate to at most 12 (calculated as $16 - 4$) decimal places of accuracy.

It is difficult to find useful yardsticks in the literature about what constitutes a large condition number because many books have drastically different cut-offs. For this book, the following guidelines will be used. For a linear model, when the condition number is less than 10^4 , no serious collinearity is present. When the condition number is between 10^4 and 10^6 , moderate collinearity is present, and when the condition number exceeds 10^6 , severe collinearity is present and the values of the parameter estimates are not to be trusted. The difficulty with the use of the condition number is that it fails to identify which columns are collinear and simply indicates that collinearity is present. If multicollinearity is present wherein a function of one or more columns is collinear with a function of one or more other columns, then the condition number will fail to identify that collinearity. See Belsley et al. (1980) for details on how to detect collinearity among sets of covariates.

Collinearity may also be caused by poor scaling and/or near singularity of the $x^T x$ matrix. If the collinearity is due to scaling, then one simple way to remove the collinearity is by centering. Centering creates a new variable x^* using

$$x_{ij}^* = x_{ij} - \bar{x}_i, \quad (53)$$

where x_{ij} is the value of the j th row of the i th variable and \bar{x}_i is the mean of the i th variable. An expansion of centering is standardizing the covariates which is done using

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_i}{s_i}, \quad (54)$$

where s_i is the standard deviation of the i th variable. After centering, x^* has zero mean with the same variance as the

original data. After standardization, x^* has zero mean and variance 1, which forces approximate orthogonality between the covariates. A third method is scaling where each observation is divided by a column-dependent constant, such as the mean, making each column approximately the same scale.

For example, suppose with the linear model

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \quad (55)$$

with corresponding $x^T x$ matrix

$$x^T x = \begin{bmatrix} 8 & 117 & 3607 \\ & 2251 & 58112 \\ & & 1861257 \end{bmatrix}. \quad (56)$$

The condition number of $x^T x$ is 1.92×10^5 , which is quite ill-conditioned. The model could be centered on the mean of 15 and 450, respectively,

$$Y = \theta_0^* + \theta_1^* (x_1 - 15) + \theta_2^* (x_2 - 450) \quad (57)$$

with corresponding $x^T x$ matrix

$$x^T x = \begin{bmatrix} 8 & -3 & 7 \\ & 541 & 5357 \\ & & 234957 \end{bmatrix} \quad (58)$$

and condition number 29,475, a 65-fold reduction over the original model. The “*” superscript in (57) denotes that the parameter estimates are not the same as those in (55). Or the model could be scaled to its mean

$$Y = \theta_0^* + \frac{\theta_1^* x_1}{15} + \frac{\theta_2^* x_2}{450}. \quad (59)$$

Then

$$x^T x = \begin{bmatrix} 8.0 & 7.8 & 8.0 \\ & 10.0 & 8.6 \\ & & 9.2 \end{bmatrix}. \quad (60)$$

and the condition number becomes 47, a 40,000-fold reduction from the original condition number. In the original domain, inverting $x^T x$ would lead to a loss of about six decimals of precision on a double-precision computer, but inversion after transformation would lead to only a 2 decimal loss in precision. Lastly, the model could be standardized

$$Y = \theta_0^* + \frac{\theta_1^* (x_1 - 15)}{8.78} + \frac{\theta_2^* (x_2 - 450)}{183.21} \quad (61)$$

with $x^T x$ matrix

$$x^T x = \begin{bmatrix} 8.00 & -0.34 & 0.038 \\ & 7.02 & 3.33 \\ & & 7.00 \end{bmatrix} \quad (62)$$

and corresponding condition number of 2.83, a 682,000-fold reduction over the original condition number. Less than 1 decimal loss of precision would occur after standardization. It makes little difference whether centering or standardizing with the mean or median, except that these estimates tend to be study specific. A more robust method of centering would be to use a consistent value across all

studies and all drugs (Holford 1996). For example, all BSA values would be centered by 1.7 m^2 , weight by 70 kg, age by 40 years (70 years for elderly studies), 7.5 L/h for creatinine clearance, etc. In this manner, parameter estimates can be compared across studies making them more relevant.

One advantage of centering over standardization or scaling is that the parameter estimates associated with x are the same as the original data. The only difference being the estimate of the intercept. However, since centering only transforms the data to have the same mean, the variance of the columns of x may still be of differing magnitudes. Even after centering, ill-conditioning may still be present. Scaling presents the opposite problem. After scaling, the variance of the columns of x may be of the same magnitude but the means may be vastly different. Hence, ill-conditioning may still be present after scaling. Only standardization transforms the data to the same mean and variance and from a purely numeric point of view is the method of choice. However, with standardization and scaling the parameter estimates obtained from the transformed data are not the same as the original data and must be transformed back to the original domain should one wish to interpret the parameter estimates. A disadvantage of transforming the predictor variables to the same scale is that the transformation does not always cure ill-conditioning. For example, centering will not prevent loss of numerical accuracy if any of the predictor variables are correlated with the model intercept (Simon and Lesage 1988).

A fourth method to remove the collinearity is by transforming the collinear variables into another variable and use that variable as a surrogate. For example, height and weight are often highly correlated and can be combined into a composite variable called body surface area (BSA), which is a measure of the overall surface area on an individual. There are a number of different measures to compute BSA, but a common one is based on the height and weight on an individual

$$\text{BSA} = 0.0235(\text{Weight})^{0.51456}(\text{Height})^{0.42246}, \quad (63)$$

where BSA is in m^2 , weight is in kg, and height is in cm (Gehan and George 1970). As an example, consider the data in Table 1. Apparent oral clearance was obtained from 65 individuals. Height and weight were collected on all subjects. Both height (Pearson's r : 0.2219, $p = 0.0757$) and weight (Pearson's r : 0.4684, $p < 0.0001$) were marginally correlated with clearance (see Fig. 4). Height and weight had a better correlation with each other (Pearson's r : 0.6038, $p < 0.0001$) than with clearance. The SAS output from the regression analysis is presented in Table 2.

When height and weight were included in the models alone, they were both positively related to clearance ($p < 0.10$). When both variables were included in the model, height showed a sign change and now has a negative relationship with clearance. This is the first warning sign that something is wrong. The eigenvalues of $x^T x$ were $\{2.99, 0.00854, 0.000939\}$. The condition number of the model

with both covariates was 3,185, which is not exceedingly large, but nevertheless indicated that the resulting inverted matrix lost three to four decimal places during large. But, there were two eigenvalues near zero indicating that two variables were collinear. When BSA was used as the sole covariate, the coefficient of determination was slightly smaller than using weight alone, but far better than height. A further refinement in the model might be one where the intercept is removed from the model since the 90% confidence interval for the intercept included zero. In summary, when the covariates were regressed alone they both were statistically significant as predictor variables for clearance. But when entered together, collinearity among predictors occurred and the effect of height became opposite what was expected.

Sometimes, even after rescaling, when the $x^T x$ matrix is still ill-conditioned, then either ridge regression or principal components regression may be necessary. Briefly, in ridge regression a small constant (k) is added to the $x^T x$ matrix prior to inversion so as to stabilize the matrix. Hence, the estimator for θ becomes

$$\hat{\theta} = x(x^T x + kI)^{-1} x^T Y, \quad (64)$$

where I is the identity matrix. The choice of the constant must be chosen with care because the resulting parameter estimates become biased to some degree. However, the reduction in the variance of the estimators may be greater than the resulting increase in bias such that the trade-off is of merit.

Principal components regression is another biased regression technique but when done successfully is superior to OLS in terms of prediction and estimation. Principal components (PC) are linear transformations of the original variables such that each PC is orthogonal or uncorrelated to the others (Jackson 1991). There will be k principal components if there are k variables. Of these k principal components, j ($j < k$) components may contain most of the "information" contained in k . Thus, regression of the j principal components, instead of the original k variables, may be used for regression. The predicted values can then be back-transformed to the original domain for prediction. The reader should see Neter et al. (1996) for further details of these algorithms.

Influence Diagnostics

Frequently, data contain samples that are different from the bulk of the remaining data, i.e., these observations may be outliers. Outliers may arise from improper recording of data, assay error (both random and systematic), choice of an invalid model, or may not be outliers at all, but are in fact legitimate data points. Residual analysis is a tool to assess the fit of a model. Although useful, it fails to provide information on how individual observations may affect the parameter estimates or their standard errors. As most modelers have seen, a single observation may have a dramatic influence on the estimation of the relationship between Y and x . Similarly,

Table 1

Clearance, weight, and height estimates from 65 subjects

Clearance (mL/min)	Weight (lb.)	Height (in.)	Clearance (mL/min)	Weight (lb.)	Height (in.)
62,612	124.5	67.7	51,530	117.2	66.4
54,951	136.5	65.1	55,333	142.4	65.1
54,897	140.7	68.6	48,292	115.0	66.5
55,823	148.8	65.2	51,453	143.9	69.5
68,916	185.1	70.8	56,779	122.5	70.2
74,333	185.7	70.5	56,346	145.6	71.1
62,203	143.4	71.9	58,239	168.9	72.6
40,359	126.7	67.5	64,677	182.0	67.9
51,205	134.5	66.8	67,045	167.8	71.1
57,108	151.8	67.2	51,764	140.0	71.7
51,574	131.2	60.2	69,917	165.1	74.6
49,579	127.6	63.4	38,738	107.4	63.7
62,450	152.5	75.6	59,912	132.2	66.3
49,879	144.6	68.6	53,475	134.4	67.6
53,818	161.5	73.6	51,197	154.2	72.4
53,417	155.8	71.9	55,603	149.6	72.4
65,510	171.0	72.6	53,013	123.0	70.7
45,320	114.5	65.5	63,697	155.0	76.4
53,174	128.4	67.0	71,911	137.8	65.8
56,905	131.1	65.9	52,606	138.2	71.1
67,193	145.6	68.6	45,523	153.3	73.9
48,135	146.9	71.4	54,643	157.6	72.6
53,952	104.8	65.1	55,699	135.7	65.9
51,145	147.0	67.3	51,787	132.1	73.6
58,154	173.1	74.5	59,247	140.9	69.8
51,574	141.0	71.4	56,044	141.9	68.7
59,407	144.5	70.6	47,898	134.8	72.9
69,394	145.4	71.4	45,694	152.0	70.2
60,276	167.0	72.3	41,664	116.2	66.3
50,626	126.8	67.2	53,827	130.6	70.2
37,266	128.1	72.5	57,166	141.7	74.2
52,343	120.6	65.5	50,248	147.1	70.5
43,509	149.9	70.4			

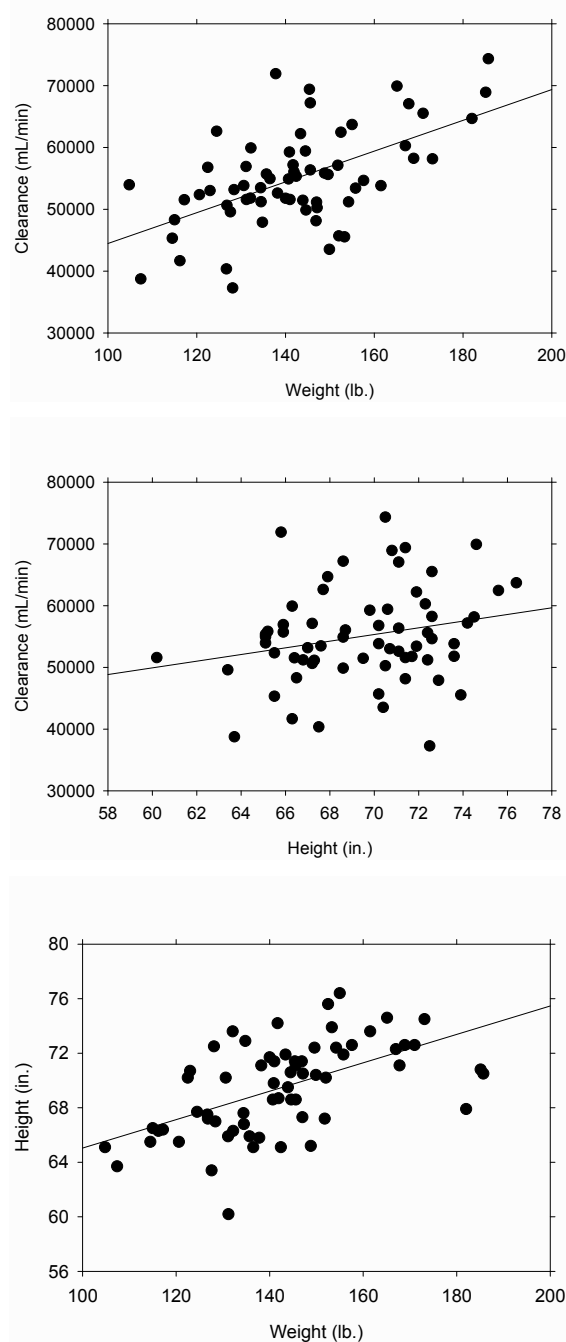


Fig. 4 Correlation plot of data in Table 1. Top plot is clearance against weight. Middle plot is height against clearance and bottom plot is weight against height. Solid line is least squares fit to the data. All three plots show evidence for a linear relationship between the respective variables

Table 2

SAS output from regression analysis of Table 1 using clearance as the dependent variable

Both Variables						
Variable	df	Parameter estimate	Standard error	T for H0: parameter=0	Prob > T	Variance inflation
Intercept	1	34810	17179.538954	2.026	0.0470	0.00000000
Height	1	-278.561305	290.86043976	-0.958	0.3419	1.44322425
Weight	1	277.806275	54.70976093	5.078	0.0001	1.44322425
Collinearity Diagnostics						
Number	Eigenvalue	Condition Index ^a	Var Prop Intercept	Var Prop Height	Var Prop Weight	
1	2.99052	1.00000	0.0002	0.0002	0.0012	
2	0.00854	18.70927	0.0690	0.0152	0.7982	
3	0.0009391	56.43213	0.9308	0.9846	0.2006	
Height only						
Variable	df	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T	
Intercept	1	17537	19877.615084	0.882	0.3810	
Height	1	539.916181	285.79575217	1.889	0.0635	
Weight only						
Variable	df	Parameter estimate	Standard error	T for H0: Parameter=0	Prob > T	
Intercept	1	19595	6,535.0993235	2.998	0.0039	
Weight	1	248.769673	45.51058177	5.466	0.0001	
BSA only						
Variable	df	Parameter estimate	Standard error	t Value	Pr > t	
Intercept	1	1,695.00446	10848	0.16	0.8763	
BSA	1	30090	6,100.33895	4.93	<.0001	

^aDenotes that the condition index reported by SAS is calculated using (52) and is the square root of the condition number otherwise used throughout this book

deleting a single observation in a nonlinear model may result in convergence, whereas inclusion of the data point may not. An observation which individually, or together with other observations, has a larger impact on a parameter estimate, such as the slope, its standard error, or associated T -test, than other observations is said to be influential. Influence diagnostics provide rational, objective measures to assess the impact individual data points have on the regression coefficients and their standard errors. Thus, by using influence diagnostics a modeler can have an impartial measure by which to either remove a data point from an analysis or weight that data point sufficiently so as to force it to have equal influence as other observations in the data set.

The purpose of this section is to provide a primer on influence diagnostics with the ultimate hope being that more rational decision making rules will be used before discarding data points from an analysis and greater use of influence diagnostics will result in their incorporation in pharmacokinetic software packages (something that is definitely lacking at this time). The reader is referred to Belsley et al. (1980) or Neter et al. (1996) for further in-depth discussion on using influence diagnostics.

Influence in the x -direction

Although most are familiar with the influence a discordant observation in the Y -direction has on parameter estimation, the independent variables themselves also influence the parameter estimates. Recall that ordinary least squares minimizes the quantity

$$\sum_{i=1}^n (Y - \hat{Y})^2 = \sum_{i=1}^n (Y - x\hat{\theta})^2, \quad (65)$$

which can be expanded to

$$\sum_{i=1}^n (Y - \hat{Y})^2 = \sum_{i=1}^n (Y - x(x^T x)^{-1} x^T Y)^2. \quad (66)$$

Let $h = x(x^T x)^{-1} x^T$ be called the HAT matrix. Then least squares minimizes

$$\sum_{i=1}^n (Y - hY)^2, \quad (67)$$

and an alternative method for determining the predicted values of the dependent variable is

$$\hat{Y} = x\hat{\theta} = hY. \quad (68)$$

The HAT matrix can be thought to map the observed values (Y) to the predicted values (\hat{Y}). One important aspect of the least squares model is that a better fit is observed at remote

observations than at observations near the middle of the data. By corollary, observations that have large HAT values will be better predicted because the method of least squares attempts to find parameter estimates that result in residuals near zero. Thus, it is said that observations with large HAT values have more influence than observations with small HAT values. Another term used to indicate influence in the x -direction is called *leverage*.¹ Observations with high leverage exert greater influence on parameter estimates than observations with low leverage.

Another way to look at the HAT matrix is as a distance measure – values with large HAT values are far from the mean of x . It can be shown that the HAT matrix has two useful properties: $0 \leq h_i \leq 1$ and $\sum h_i = p$ for $i = 1$ to n . The average size of h_i is then p/n . It is desirable to have all independent variables to have equal influence, i.e., each data point has $h_i \cong p/n$. As a rule of thumb, an independent variable has greater *leverage* than other observations when h_i is greater than $2p/n$. Figure 5 presents an example of noninfluential and influential x -values.

Consider the previous example where clearance was modeled as a function of BSA. There were 65 observations and two estimable parameters in the model. Hence, under the rule of thumb, observations with HAT values greater than 0.062 exerted greater leverage than other observations. Figure 6 presents the HAT values plotted against BSA. Four observations met the criteria for having high leverage. This plot illustrates that observations with large HAT values in a model including an intercept are at the extremes of x . In the single predictor case, this corresponds to observations at the tails of the distribution of x . In the two-dimensional case this would correspond to observations near the ends of the ellipse. In the case where no intercept is in the model, only observations far removed from zero can have high leverage. It must be kept in mind that a large HAT value is not necessarily a bad thing. An observation with a large HAT value that is concordant with the rest of the data probably will not change the parameter estimates much. However, a large HAT value coupled with a large DFBETAS (see below) is a combination that spells trouble.

Influence in the Y -direction

Most pharmacokineticists are familiar with this case, when a single observation(s) is discordant from the other observations in the Y -direction. Outliers in the Y -direction are often detected by visual examination or more formally by residual analysis. One common statistic is standardized residuals

$$e_s = \frac{e_i}{\sqrt{\text{MSE}}} \quad (69)$$

¹ More formally, leverage is defined as the partial derivative of the predicted value with respect to the corresponding dependent variable, i.e., $h_i = \partial \hat{Y}_i / \partial Y_i$, which reduces to the HAT matrix for linear models.

Under the assumption that the residuals are independent, normally distributed with mean 0 and constant variance, when the sample size is large, standardized residuals greater than ± 2 are often identified as suspect observations. Since asymptotically standardized residuals are normally distributed, one might think that they are bounded by $-\infty$ and $+\infty$, but in fact, a standardized residual can never exceed $\pm \sqrt{(n-p)(n-1)n^{-1}}$ (Gray and Woodall 1994). For a simple linear model with 19 observations, it is impossible for any standardized residual to exceed ± 4 . Standardized residuals suffer from the fact that they prone to “ballooning” in which extreme cases of x tend to have smaller residuals than cases of x near the centroid of the data. To account for this, a more commonly used statistic, called studentized or internally studentized residuals, was developed

$$e_{si} = \frac{e_i}{\sqrt{\text{MSE}(1-h_i)}} \quad (70)$$

Under the assumption that the residuals are independent, normally distributed with mean 0 and constant variance, when the sample size is large, studentized residuals greater than ± 2 are often identified as suspect observations. Like standardized residuals, studentized residuals are not bound by $-\infty$ and $+\infty$, but are bounded by $\pm \sqrt{(n-p)}$ (Gray and Woodall 1994). An alternative statistic, one that is often erroneously interchanged with standardized residuals, are studentized deleted residuals, which are sometimes called jackknifed residuals, externally studentized residuals, or R-student residuals

$$e_i^* = \frac{e_i}{\sqrt{\text{MSE}(i)(1-h_i)}} \quad (71)$$

where $\text{MSE}(i)$ is the square root of the mean square error with the i th data point removed. Fortunately, a simple relationship exists between MSE and $\text{MSE}(i)$ so that e_i^* can be recalculated without having to fit a new regression after each data point is removed

$$\text{MSE}(i) = \frac{\left[(n-p)\text{MSE} - \frac{e_i^2}{1-h_i} \right]}{n-p-1} \quad (72)$$

Upper bounds for externally studentized residuals have not been developed. Externally studentized residuals are distributed as a Student's t -distribution with $n-p-1$ degrees of freedom. Thus, in the case of a single outlier observation, a quick test would be to compare the value of the external studentized residual to the appropriate t -distribution value, although as Cook and Weisberg (1999) point out, because of issues with multiplicity a more appropriate comparison would be Student's t -distribution with α/n critical value and $n-p-1$ degrees of freedom. In general, however, a yardstick of ± 2 or ± 2.5 is usually used as a critical value to flag suspect observations.

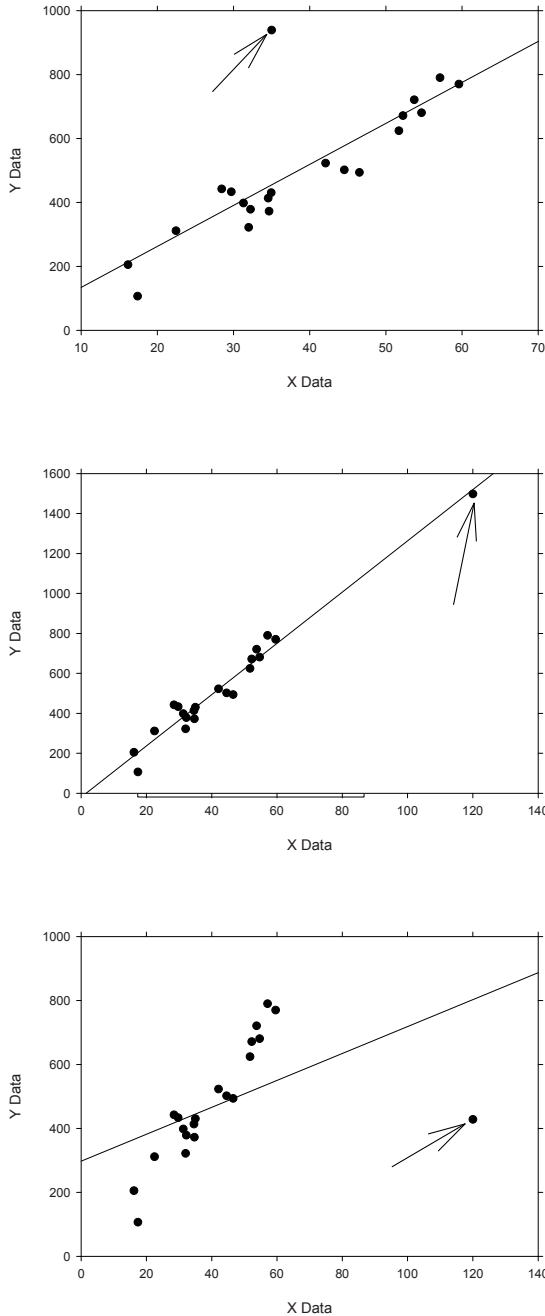


Fig. 5 Example of influential and noninfluential observations. *Top plot:* Y-value is discordant from bulk of data but does not influence the estimate of the regression line. *Middle plot:* x-value is discordant from bulk of data but does not influence the estimate of the regression line. *Bottom plot:* x-value and Y-value are discordant from bulk of data and have a profound influence on the estimate of the regression line. Not all outlier observations are influential and not all influential observations are outliers

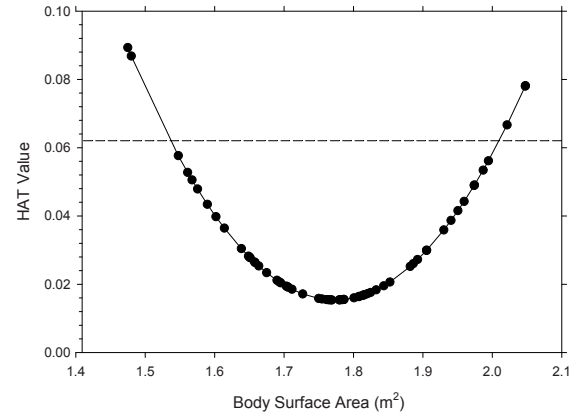


Fig. 6 Plot of HAT values against body surface area under a simple linear model using the data in Table 1. The plot illustrates that HAT values are a function of the x-matrix and that observations with high HAT values are at the extremes in x. The dashed line is the yardstick for observations with high leverage, $2p/n$

Identification of Influential Observations

Influential observations are ones that significantly affect the values of the parameter estimates, their standard errors, and the predicted values. One statistic used to detect influential observations has already been presented, the HAT matrix. An obvious way to detect these observations is to remove an observation one at a time and examine how the recalculated parameter estimates compare to their original values. This is the row deletion approach to influence diagnostics and on first glance it would appear that this process requires n -iterations – a numerically intensive procedure. Statisticians, however, have derived equations that directly reflect the influence of the i th observation without iteration. One useful diagnostic is DFFITS

$$\text{DFFITS} = \sqrt{\frac{h_i}{1-h_i}} \left[\frac{e_i}{\sqrt{\text{MSE}(i)(1-h_i)}} \right], \quad (73)$$

which measures the impact of deleting the i th data point on predicted values and is the number of standard errors that the i th predicted value changes if that observation is deleted from the data set. DFFITS are basically studentized deleted residuals scaled according to the leverage of the i th observation.

Another useful statistic that is used is called DFBETAS,

$$\begin{aligned} \text{DFBETAS} &= \frac{\beta - \beta(i)}{\sqrt{\text{MSE}(i)(x^T x)^{-1}}} \\ &= \frac{(x^T x)^{-1} x_i^T e_i}{(1-h_i)\sqrt{\text{MSE}(i)(x^T x)^{-1}}} \end{aligned} \quad (74)$$

where $\beta(i)$ denotes the least squares parameter estimates with the i th data point removed. DFBETAS measures the number of standard errors that a parameter estimate changes with the i th observation deleted from the data set.

A large change in DFBETAS is indicative that the i th observation has a significant impact on the value of a regression coefficient. As a yardstick for small to moderate sample sizes, DFFITS and DFBETAS greater than ± 1 are indicative of influential observations. For larger sample sizes a smaller absolute value may be needed as a yardstick: one rule of thumb is $2n^{-0.5}$ for DFBETAS and $2\sqrt{p/n}$ for DFFITS (Belsley et al. 1980).

One problem with DFBETAS is that there will be $n \times p$ DFBETAS for the analyst to examine, which can be tedious to examine. Cook's distance, D_i , is a composite score that assesses the influence an observation has on the set of regression parameters and is computed by

$$D_i = \left(\frac{e_i^2}{(1 - h_i)^2} \right) \left(\frac{h_i}{p \times \text{MSE}} \right). \quad (75)$$

As its name implies, Cook's distance is a distance measure that represents the standardized distance in p -dimensional space between β and $\beta(i)$. A large value of D_i indicates that the i th observation has undue influence on the set of regression parameters. Once an observation has been identified as exerting undue influence then DFBETAS can be examined to determine which regression parameters are affected. Interpreting Cook's distance and finding a yardstick is much more difficult than DFFITS or DFBETAS. Myers (1986) recommends interpreting a particular Cook's distance as follows: If Cook's D is about 50% of the F -value from an $F_{p, n-p}$ distribution then deletion of the i th observation moves the centroid of confidence region to the 50% confidence region.

Although DFFITS and DFBETAS provide a flag that the i th observation has an impact on the value of the j th regression coefficient, they do not give any indication of whether the influence that is exerted is positive or negative. Like the HAT matrix, a large DFFITS or DFBETAS is not necessarily a bad thing. It is the combination of a high leverage observation in the presence of large DFFITS or DFBETAS that results in erratic regression parameter estimates.

The variance-covariance of linear regression parameter estimates is given by $\sigma^2(x^T x)^{-1}$ and a statistic that summarizes the properties of the variance/covariance matrix is the generalized variance of the regression parameters

$$\text{GV} = |\text{Var}(\beta)| = |\text{MSE}(x^T x)^{-1}|, \quad (76)$$

where $|\cdot|$ is the determinant function. Precise estimation of the regression parameters results in small determinants or GV. COVRATIO measures the ratio of the variance/covariance without and with the i th observation and is calculated using

$$\text{COVRATIO} = \frac{|\text{MSE}(i)(x_{(i)}^T x_{(i)})^{-1}|}{|\text{MSE}(x^T x)^{-1}|}. \quad (77)$$

where $x_{(i)}$ denotes the x matrix without the i th observation. COVRATIOs greater than one are indicative that the i th observation improves the performance of the model over what would be seen without the observation in the data set. A combination of high leverage and a small residual results in an observation that improves the properties of the regression parameters. As a yardstick, observations with $\text{COVRATIO} > 1 + 3p/n$ or $\text{COVRATIO} < 1 - 3p/n$ (applies only when $n > 3p$) show undue influence on the generalized variance of the regression parameters.

Unless the number of observations is small, influence diagnostics are best examined graphically. Gray (1986) recommended for the linear model that a useful diagnostic plot is h_i against e_i^2 / SSE , the normalized residual for the i th subject. Such a plot is called an L-R triangle for leverage and residual. Regardless of the data set, the L-R triangle data should show low leverage and small residuals such that the majority of the data cluster near $(p/n, 0)$. Cases will that have undue influence will be discordant from the bulk of the data. Obviously, plots of h_i against any influence diagnostics will find utility. Lastly, bubble plots having one of the other influence diagnostics, such as COVRATIO, may be used to gain a trivariable influence plot.

Belsley et al. (1980) present many more diagnostics, including ones for multiple row deletion, but most of the ones that have been presented herein are easily obtained using most, if not all, linear regression software. One last point is that these diagnostics are not independent of each other, they are often correlated themselves and will show overlap in observations that are flagged.

So What Now?

Once an outlier or an influential observation is detected what can be done about it? Obviously an observation can be deleted, but clearly what is needed is a further examination of why that observation was flagged in the first place. If nothing of interest arises in re-examination of the data points, then there is no sound rationale for removal of the observation in question. One might then consider that the model itself is wrong. This is a very important concept because model misspecification is often discovered through outlier and influential observation analysis. Lastly, one might try a weighted linear regression model where the weights are proportional to the inverse of the HAT matrix. In other words, influential observations are given less weight in the model than uninfluential observations. Alternatively, all observations could have weights equal to "1," except the data point(s) in question which is given a much smaller weight. In this manner the observation is not removed from the data set, but is simply given less weight in the modeling process.

Given the level of research activity devoted to identification of influential observations, considerably less effort has been devoted to what to do about them. Under guidelines (E9: Statistical Principles for Clinical Trials) developed by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (1997), more commonly called ICH, several principles for dealing with outliers or influential observations are presented. First, data analysis should be defined prior to analyzing the data, preferable before data collection even begins. The data analysis plan should specify in detail how outliers or influential observations will be handled. Second, in the absence of a plan for handling outliers or influential observations, the analyst should do two analyses, one with and the other without the points in question, and the differences between the results should be presented in the discussion of the results. Lastly, identification of outliers should be based on statistical, as well as scientific rationale, and that the context of the data point should dictate how to deal with it.

Example

Port et al. (1991) administered 5-fluorouracil (5-FU) treatments to 26 patients with advanced carcinomas of various origin under a variety of doses and treatment schedules. Monotherapy was given as 5-day courses weekly for 3 weeks, once weekly for 3 weeks, or once every 3 weeks. Combination therapy with methotrexate (MTX) was given once every 2–3 weeks. Serial blood samples for pharmacokinetic analysis were collected on Day 1 and 5-FU clearance was determined by noncompartmental methods. Some patients had multiple cycles of therapy and for those subjects only data from the first cycle was included in this analysis. The following covariates were available for analysis: sex, age, BSA, 5-FU dose, and the presence or absence of MTX. Scatter plots and box and whisker plots are shown in Fig. 7 with the data presented in Table 3.

Of interest was to determine whether a useful model relating 5-FU clearance and patient demographics could be developed for possible use in future individualized dosing regimens. Nonparametric correlation analysis between the covariates revealed that sex and BSA were correlated ($r = -0.4689$, $p = 0.0157$), a not surprising result since both males and females were enrolled in the study and males (which were coded as “1”) would be expected to have higher BSA than females (which were coded as “0”). The sign of the correlation would change to positive had the coding been reversed. Also, 5-FU dose was correlated with the presence or absence of MTX ($r = 0.4382$, $p = 0.0251$). This too was not surprising given the study design in that patients who were treated with MTX were also the ones who were treated with relatively high-dose 5-FU. The

magnitude of the correlations indicated that mild collinearity may be a problem during the analysis.

Examination of the univariate distribution of 5-FU clearance revealed it to be skewed and not normally distributed suggesting that any regression analysis based on least squares will be plagued by non-normally distributed residuals. Hence, Ln-transformed 5-FU clearance was used as the dependent variable in the analyses. Prior to multiple regression analysis, age was standardized to 50 years old, BSA was standardized to 1.83 m^2 , and dose was standardized to 1,000 mg. A p -value less than 0.05 was considered to be statistically significant. The results from the SLRs of the data (Table 4) revealed that sex, 5-FU dose, and presence or absence of MTX were statistically significant.

Multiple regression of all covariates (Table 5) had a condition number of 1,389, indicating that the model had little collinearity. Notice that presence or absence of MTX as a variable in the model was not statistically significant, possibly a result of the collinearity between presence or absence of MTX and 5-FU dose. Since with the univariate models, 5-FU dose had a higher coefficient of determination than presence or absence of MTX, a second multivariate model was examined where presence or absence of MTX was removed from the model. Table 6 presents the results. Now, age was not statistically significant. This variable was removed from the model and the reduced model's results are shown in Table 7. Sex was almost significant and it was decided to remove this variable from the model. The resulting model and influence diagnostics are shown in Tables 8 and 9, respectively. Influence plots, including an L - R plot, are shown in Fig. 8. The condition number of this model was 451 indicating the new model had good parameter stability.

Examination of the collinearity diagnostics indicated that two of the observations had HAT values greater than the yardstick of $2 \times 3/26$ or 0.23. One studentized residual was greater than ± 3 (Subject 3). Subject 3 had a DFBETA of 1.023 for the intercept and -1.084 for the parameter associated with BSA, indicating that these parameters would change by more than one standard error should this subject be removed from the data set. This subject had a COVRATIO of 0.444, much lower than the critical value of 0.65, and the largest absolute DFFITs in the data set. Clearly, there was something unusual about this subject. At this point, one might then go back and examine what was unique about this subject. Although not the lowest clearance observed in the study, this subject did have the second lowest value. Why? Since this data set was taken from the literature this question cannot be answered. For purposes of this analysis, it was decided that Subject 3 would be removed from the data set. The resulting model after removal of Subject 3, as shown in Table 10 with

Table 3

Treatment groups, patient demographics, and 5-FU clearance values from Port et al. (1991)

Subject	Sex	Age (Years)	BSA (m ²)	Dose (mg)	MTX	5-FU CL (L/min)
1	1	43	1.65	1,500	1	0.58
2	1	48	1.63	750	0	0.56
3	1	50	2.14	1,500	1	0.47
4	0	68	2.14	1,800	1	0.85
5	1	50	1.91	1,500	1	0.73
6	1	48	1.66	1,500	1	0.71
7	1	45	1.6	1,500	1	0.61
8	0	53	2.05	1,600	1	0.86
9	0	44	1.94	850	0	1.36
10	0	58	1.7	1,500	1	0.53
11	1	61	1.83	1,600	1	0.91
12	0	49	1.67	1,500	1	0.81
13	0	70	1.89	1,600	1	0.64
14	0	47	1.64	1,500	1	0.56
15	0	63	1.88	600	0	0.98
16	1	46	1.67	1,500	1	0.79
17	0	45	2.01	1,000	0	1.92
18	0	46	1.82	1,000	0	1.65
19	0	57	1.68	1,400	1	0.83
20	1	52	1.76	750	0	1.19
21	1	64	1.27	1,200	1	0.57
22	0	65	1.67	750	0	1.12
23	1	75	1.67	1,500	0	0.5
24	1	64	1.57	1,500	0	0.44
25	0	60	2.02	1,800	0	0.67
26	0	54	2.13	1,800	0	0.93

Sex: 0 = males, 1 = females; MTX: 0 = no methotrexate given, 1 = methotrexate given; CL, clearance

influence diagnostics shown in Table 11, resulted in a model accounting for more than 59% of the total variance with all model parameters being statistically significant. The condition number of the final model was 481 indicating the model to be quite stable. Examination of the influence diagnostics showed that now possibly Subject 2 showed undue influence. Some modelers would indeed remove this subject from the model, but removal of Subject 2 is not advised given the sample size of the analysis. So, the final model was one where BSA positively affected 5-FU clearance and dose negatively affected 5-FU clearance, an indication of Michaelis-Menten elimination kinetics.

Conditional Models

Up to now it has been assumed that x is fixed and under control of the experimenter, e.g., the dose of drug given to subjects or sex of subjects in a study, and it is of interest to make prediction models for some dependent variable Y or make inferences on the regression parameters. There are times when x is not fixed, but is a random variable, denoted X . An example would be a regression analysis of weight vs. total clearance, or age vs. volume of distribution. In both cases, it is possible for the experimenter to control age or weight, but more than likely these are samples randomly drawn from subjects in the population.

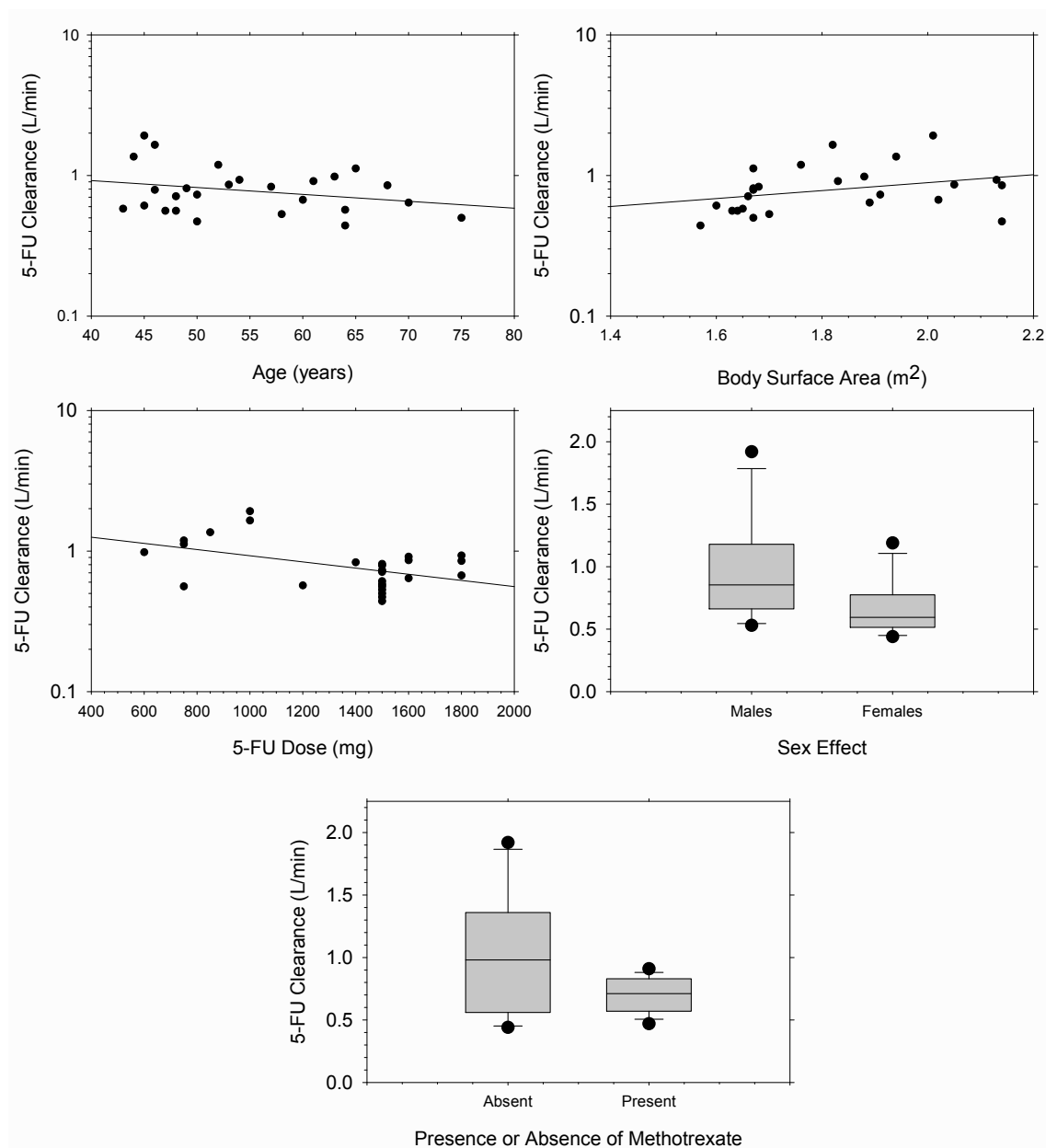


Fig. 7 Scatter plots and box and whisker plots of 5-fluorouracil (5-FU) clearance as a function of patient demographics. Data are presented in Table 3. Solid line is the least squares fit to the data. Note that some plots are shown on a log-scale

Table 4

Results of univariate regression analysis of the data in Table 3 using a simple linear model with Ln-transformed 5-FU clearance as the dependent variable

Variable	Intercept	SE(Intercept)	Slope	SE(Slope)	R^2
Sex	-0.0922	0.0916	-0.346	0.135	0.2158
Age	0.366	0.453	-0.564	0.408	0.0738
BSA	-1.416	0.620	1.188	0.628	0.1297
Dose	0.428	0.264	-0.505	0.190	0.2278
MTX	-0.0763	0.107	-0.305	0.140	0.1640

Note: Bold values were statistically significant at $p < 0.05$

Table 5

Results of multivariate linear regression of data in Table 3 using Ln-transformed 5-FU clearance as the dependent variable

Variable	Estimate	SE(Estimate)	<i>t</i> -value	<i>p</i> -value
Intercept	0.104	0.696	0.15	0.883
Sex	−0.247	0.123	−2.00	0.059
Age	−0.490	0.323	−1.51	0.146
BSA	0.995	0.589	1.69	0.106
Dose	−4.78	0.212	−2.26	0.035
MTX	−0.061	0.146	−0.42	0.681

Note: R^2 was 0.5750 with an adjusted coefficient of determination of 0.4688

Table 6

Results of multivariate linear regression of data in Table 3 using Ln-transformed 5-FU clearance as the dependent variable without MTX included in the model

Variable	Estimate	SE(Estimate)	<i>t</i> -value	<i>p</i> -value
Intercept	0.025	0.656	0.04	0.971
Sex	−0.246	0.121	−2.04	0.054
Age	−0.452	0.305	−1.48	0.153
BSA	1.076	0.545	1.97	0.062
Dose	−0.535	0.160	−3.35	0.003

Note: R^2 was 0.5713 with an adjusted coefficient of determination of 0.4897

Table 7

Results of multivariate linear regression of data in Table 3 using Ln-transformed 5-FU clearance as the dependent variable without MTX and age included in the model

Variable	Estimate	SE(Estimate)	<i>t</i> -value	<i>p</i> -value
Intercept	0.522	0.558	0.04	0.971
Sex	−0.219	0.122	−1.79	0.087
BSA	1.176	0.556	2.12	0.046
Dose	−0.580	0.161	−3.60	0.002

Note: R^2 was 0.5263 with an adjusted coefficient of determination of 0.4617

Table 8

Results of multivariate linear regression of data in Table 3 using Ln-transformed 5-FU clearance as the dependent variable without MTX, age, and sex included in the model

Variable	Estimate	SE(Estimate)	<i>t</i> -value	<i>p</i> -value
Intercept	−1.004	0.512	1.96	0.062
BSA	1.622	0.520	3.12	0.005
Dose	−0.621	0.167	−3.73	0.001

Note: The coefficient of determination was 0.4574 with an adjusted coefficient of determination of 0.4102. BSA and dose were standardized prior to analysis.

Table 9
Influence diagnostics for the regression model presented in Table 8

Subject	Residual	RStudent	HAT	COV Ratio	DFFITS	DFBETAs		
						Intercept	BSA	DOSE
1	-0.071	-0.247	0.071	1.220	-0.068	-0.033	0.041	-0.031
2	-0.555	-2.249	0.156	0.728	-0.967	-0.582	0.201	0.750
3	-0.716	-3.130	0.148	0.444	-1.305	1.023	-1.084	-0.041
4	0.064	0.237	0.182	1.386	0.112	-0.089	0.074	0.048
5	-0.071	-0.247	0.055	1.199	-0.059	0.025	-0.024	-0.016
6	0.123	0.428	0.068	1.196	0.116	0.053	-0.066	0.052
7	0.024	0.084	0.089	1.253	0.026	0.015	-0.018	0.012
8	0.031	0.109	0.107	1.277	0.038	-0.027	0.025	0.010
9	0.120	0.442	0.159	1.322	0.192	-0.022	0.100	-0.152
10	-0.205	-0.719	0.058	1.131	-0.178	-0.062	0.081	-0.080
11	0.282	0.999	0.059	1.063	0.249	-0.044	0.004	0.141
12	0.246	0.867	0.065	1.105	0.229	0.099	-0.125	0.103
13	-0.123	-0.428	0.063	1.189	-0.111	0.042	-0.028	-0.055
14	-0.097	-0.340	0.074	1.215	-0.096	-0.048	0.059	-0.043
15	-0.310	-1.239	0.245	1.236	-0.706	-0.054	-0.259	0.637
16	0.221	0.776	0.065	1.127	0.205	0.088	-0.112	0.092
17	0.496	1.948	0.142	0.826	0.792	-0.265	0.540	-0.517
18	0.513	1.946	0.081	0.772	0.578	0.082	0.145	-0.415
19	0.199	0.694	0.053	1.131	0.164	0.081	-0.083	0.040
20	0.084	0.307	0.151	1.329	0.130	0.041	0.015	-0.111
21	0.062	0.247	0.286	1.588	0.157	0.145	-0.144	0.009
22	0.103	0.377	0.151	1.321	0.159	0.083	-0.018	-0.129
23	-0.237	-0.836	0.065	1.113	-0.221	-0.095	0.120	-0.099
24	-0.276	-1.001	0.102	1.113	-0.337	-0.209	0.250	-0.144
25	-0.068	-0.245	0.130	1.302	-0.095	0.061	-0.043	-0.055
26	0.162	0.606	0.176	1.320	0.281	-0.220	0.181	0.124

Note: Bolded data indicate data that are questionable.

As subjects enroll in a study, the experimenter usually cannot control how old they are or what their weight is exactly. They are random. Still, in this case one may wish to either make inferences on the parameter estimates or predictions of future Y values. Begin by assuming that Y can be modeled using a simple linear model and that X and Y have a joint probability density function that is bivariate normal

$$f_{xy}(X, Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right) + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}, \quad (78)$$

where μ_X and μ_Y are the population means for X and Y , respectively, σ_X and σ_Y are the standard deviations for X and Y , respectively, and ρ is the correlation between X and Y which can be expressed as

$$\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}, \quad (79)$$

where σ_{XY} is the covariance between X and Y . Further details regarding joint probability densities and conditional inference is presented in Appendix given at the end of the book. What is of interest is to find the conditional density function of Y given X . The probability density function for the conditional distribution of Y given X is

$$f_{XY}(Y | X) = \frac{f_{XY}(X, Y)}{f_X(X)}, \quad (80)$$

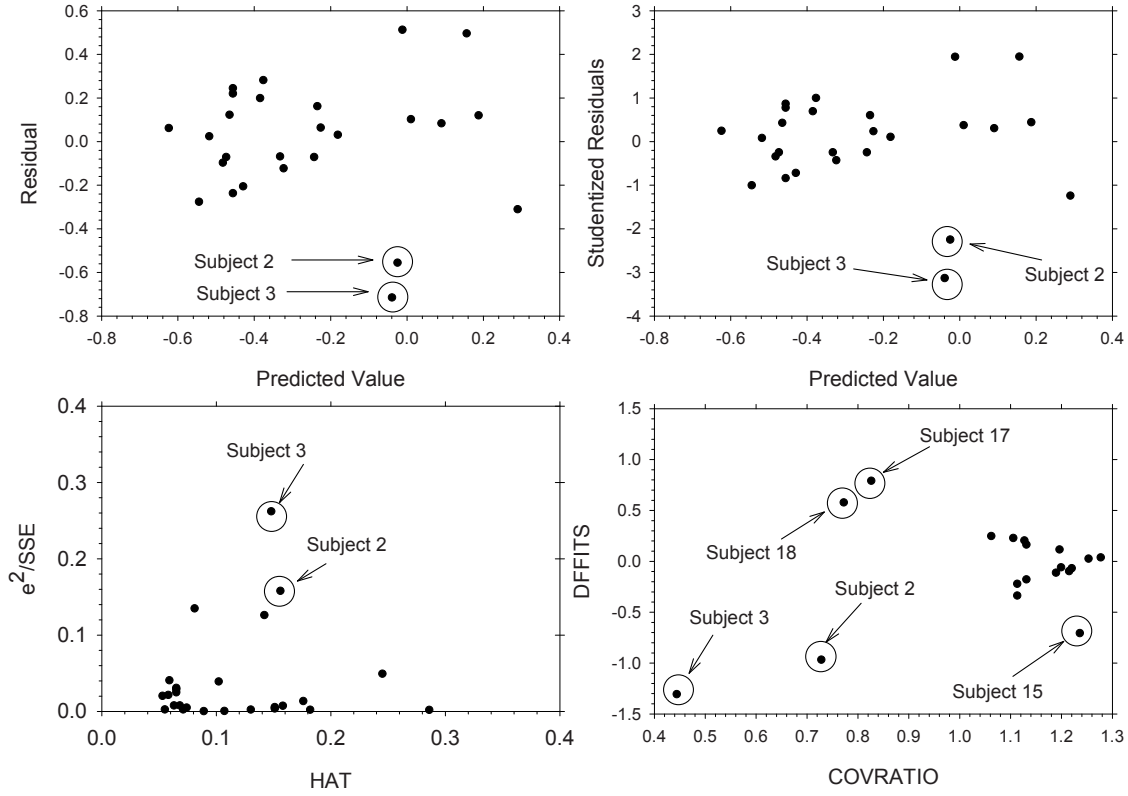


Fig. 8 Residual plots and influence plots for final linear model shown in Table 8 using data presented in Table 3. Suspect values are noted in the plots

Table 10

Results of multivariate linear regression of data in Table 3 using Ln-transformed 5-FU clearance as the dependent variable using only BSA and 5-FU dose with subject 3 removed from the analysis

Variable	Estimate	SE(Estimate)	<i>t</i> -value	<i>p</i> -value
Intercept	-1.445	0.458	-3.16	0.0045
BSA	2.102	0.468	4.49	0.0002
Dose	-0.616	0.142	-4.34	0.0003

Note: The coefficient of determination was 0.5950 with an adjusted coefficient of determination of 0.5581.

where $f_X(X)$ is the marginal density of X , which is assumed normal in distribution. Hence, the conditional distribution of Y given X is the ratio of the bivariate normal density function to a univariate normal distribution function. After a little algebra then

$$f_{Y|X}(Y|X) = \frac{1}{\sigma_{Y|X}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y - \theta_0 - \theta_1 X}{\sigma_{Y|X}}\right)^2\right], \quad (81)$$

where

$$\theta_0 = \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X}, \quad (82)$$

$$\theta_1 = \rho \frac{\sigma_Y}{\sigma_X}, \quad (83)$$

and

$$\sigma_{Y|X}^2 = \sigma_Y^2 (1 - \rho^2). \quad (84)$$

Notice that two assumptions have been made: normality of the responses and constant variance. The result is that the conditional distribution itself is normally distributed with mean $\hat{\theta}_0 + \hat{\theta}_1 X$ and variance $\sigma_{Y|X}^2$. Thus, the joint distribution function at any level of X can be “sliced” and still have a normal distribution. Also, any conditional probability distribution function of Y has the same standard deviation after scaling the resulting probability distribution function to have an area of 1.

If data are collected from a random population (X, Y) from a bivariate normal distribution and predictions about Y given X are desired, then from the previous paragraphs it may be apparent that the linear model assuming fixed x is applicable because the observations are independent, normally distributed, and have constant variance with mean $\theta_0 + \theta_1 X$. Similar arguments can be made if inferences are to be made on X given Y . Thus, if X and Y are random, all calculations and inferential methods remain the same as if X were fixed.

EIVs Regression

One assumption until now has been that the dependent and independent variables are measured without error. The impact of measurement error on the regression parameter estimates depends on whether the error affects the dependent or independent variable. When Y has measurement error, the effect on the regression model is not problematic if the measurement errors are uncorrelated and unbiased. In this case, the linear model becomes

$$Y = \theta_0 + \sum_{k=1}^{p-1} \theta_k x_k + \varepsilon + \kappa, \quad (85)$$

where κ is the measurement error in Y . This model can be rewritten as

$$Y = \theta_0 + \sum_{k=1}^{p-1} \theta_k x_k + \varepsilon^*, \quad (86)$$

where ε^* is the sum of the measurement error and model error. Equation (86) is functionally equivalent to (5). Thus, measurement error in Y is absorbed by the model error term and standard OLS techniques may be used.

Before proceeding, a distinction needs to be made between X being simply a random variable and X being random due to random measurement error. This distinction is important and the question is sometimes asked, what is the difference? If X is random but measured accurately, the experimenter has no control over its measurement, and its value may vary from study to study. An example of this might be the weight of subjects in a clinical study. If this random variable X is measured without error, then an exact, accurate measurement of X can be obtained only for *that* study. If, however, X is random due to measurement error, then repeated measurement of X within the same study will result in differing values of X each time X is measured and a misleading relationship between X and Y will be obtained.

One other distinction needs to be made between random X and X with random measurement error. Neither implies that X is biased. Bias implies a constant effect across all measurements. For example, if a weight scale is not calibrated properly and when no one is standing on it, the scale records a measure of 1 kg, then when any person is measured their weight will be biased high by 1 kg. This is not the type of measurement error that is being discussed here because any constant bias in a measuring instrument will be reflected in the estimate of the intercept. Random measurement error means that repeated measuring of a

variable will vary from measurement to measurement even though its value has not changed. An example of this might be when a patient goes to the doctor's office and their weight is measured at 180 lb. The nurse forgets to write down the value and so the patient is weighed again. This time their weight is 179 lb. That patient has not lost a pound in the few moments between measurements; they are still the same weight. But due to random measurement error, their weight changed from one reading to the next.

If both X and Y are random variables and X is measured without random error, then all the theory presented for the case of fixed x is still applicable if the following conditions are true:

1. The conditional distribution for each of the Y_i given X_i is independent and normally distributed with conditional mean $\theta_0 + \theta_1 X_i$ and conditional variance σ^2 .
2. The X_i are independent random variables whose distribution does not depend on the model parameters θ or σ^2 .

This was discussed in the previous section. In contrast, when the independent variable has measurement error, then the analyst observes

$$X_k = x_k + \delta_k, \quad (87)$$

where X_k is the observed value of x_k and δ_k is the vector of measurement errors for x_k . It is usual to assume that $\delta \sim N(0, \sigma_k^2)$ with independent measurement errors. The model is then

$$Y = \theta_0 + \sum_{k=1}^{p-1} \theta_k x_k + \varepsilon. \quad (88)$$

Since X_k is observed, not the true value of x_k , the true value must be replaced with the observed value. Then the linear model becomes

$$Y = \theta_0 + \sum_{k=1}^{p-1} \theta_k (X - \delta)_k + \varepsilon \quad (89)$$

which can be expanded to

$$Y = \theta_0 + \sum_{k=1}^{p-1} \theta_k X_k + (\varepsilon - \theta_k \delta_k). \quad (90)$$

Equation (90) looks like an ordinary regression model with predictor variable X and model error term $(\varepsilon - \theta_k \delta_k)$

$$Y = \theta_0 + \sum_{k=1}^{p-1} \theta_k X_k + \varepsilon^*. \quad (91)$$

However, the expected value of ε^* is zero with variance $\sigma^2 + \sum_{k=1}^{p-1} \theta_k^2 \sigma_k^2$. Thus the variance of the measurement errors are propagated to the error variance term, thereby inflating it. An increase in the residual variance is not the only effect on the OLS model. If X is a random variable due to measurement error such that when there is a linear relationship between x_k and Y , then X is negatively correlated with the model error term. If OLS estimation procedures are then used, the regression parameter estimates are both biased and inconsistent (Neter et al. 1996).

Table 11
Influence diagnostics for the model presented in Table 10

Subject	Residual	RStudent	HAT	COV Ratio	DFFITS	DFBETAs		
						Intercept	BSA	DOSE
1	-0.067	-0.274	0.071	1.225	-0.076	-0.035	0.043	-0.034
2	-0.541	-2.696	0.156	0.559	-1.161	-0.680	0.244	0.900
4	-0.062	-0.277	0.208	1.436	-0.142	0.116	-0.099	-0.058
5	-0.135	-0.555	0.062	1.173	-0.142	0.069	-0.067	-0.036
6	0.124	0.510	0.068	1.189	0.138	0.060	-0.075	0.062
7	0.041	0.170	0.089	1.257	0.053	0.030	-0.036	0.023
8	-0.071	-0.298	0.124	1.295	-0.112	0.084	-0.080	-0.029
9	0.052	0.225	0.166	1.369	0.100	-0.018	0.055	-0.078
10	-0.214	-0.888	0.058	1.093	-0.220	-0.069	0.091	-0.099
11	0.239	0.995	0.062	1.067	0.255	-0.060	0.023	0.142
12	0.244	1.022	0.065	1.063	0.270	0.110	-0.138	0.121
13	-0.182	-0.755	0.068	1.139	-0.205	0.089	-0.066	-0.098
14	-0.090	-0.372	0.074	1.218	-0.105	-0.051	0.063	-0.047
15	-0.361	-1.760	0.249	1.014	-1.015	-0.031	-0.392	0.906
16	0.219	0.913	0.065	1.095	0.241	0.098	-0.124	0.109
17	0.409	1.899	0.155	0.845	0.812	-0.319	0.578	-0.505
18	0.476	2.168	0.083	0.684	0.654	0.053	0.189	-0.461
19	0.196	0.805	0.053	1.108	0.190	0.088	-0.090	0.046
20	0.064	0.273	0.152	1.342	0.116	0.033	0.015	-0.099
21	0.168	0.805	0.305	1.510	0.533	0.497	-0.491	0.028
22	0.106	0.458	0.151	1.315	0.193	0.096	-0.021	-0.156
23	-0.238	-0.994	0.065	1.071	-0.263	-0.107	0.135	-0.118
24	-0.251	-1.075	0.103	1.092	-0.364	-0.224	0.266	-0.155
25	-0.163	-0.702	0.145	1.254	-0.288	0.197	-0.148	-0.161
26	0.039	0.172	0.202	1.434	0.087	-0.070	0.060	0.036

Note: Bold data indicate data that were questionable

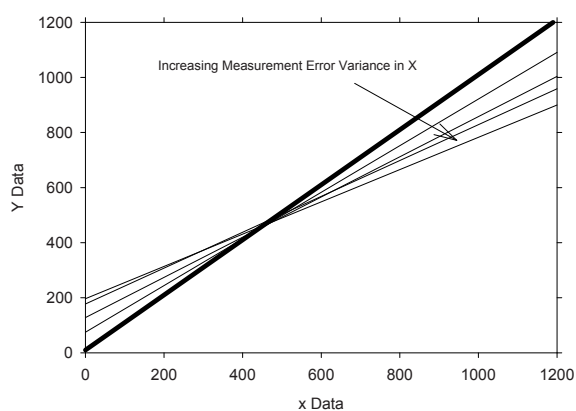


Fig. 9 Effect of increasing measurement error in X on least squares fit. Heavy line is the true least squares fit to model $Y = x + 10$. Y has no measurement error associated with it. x has increasing degrees of measurement error as indicated by the direction of the arrow, the result being the slope is attenuated and the intercept is inflated

Obtaining unbiased and consistent parameter estimates under these conditions using OLS is difficult. Measurement error in x is traditionally handled by two types of models:

- Classical error models and calibration models, where the relationship between X given x is modeled
- Regression calibration models, where the relationship between x given X is modeled

Alternative models may be developed to include additional covariates which are not measured with error, e.g., $X = f(R, Z)$. The classical model is used when an attempt to measure x is made but cannot be done so due to various measurement errors. An example of this is the measurement of blood pressure. There is only one true blood pressure reading for a subject at a particular point in time, but due to minor calibration errors in the instrument, transient increases in blood pressure due to diet, etc., possible recording errors and reading errors by the nurse, etc., blood pressure is a composite variable that can vary substantially both within and between days. In this case, it makes sense to try and model the observed blood pressure using (87). Under this model, the expected value of X is x . In

regression calibration problems, the focus is on the distribution of x given X . For purposes herein, the focus will be on the classical error model. The reader is referred to Fuller (1987) and Carroll et al. (1995) for a more complete exposition of the problem.

In the pharmacokinetic arena, there are many cases where the independent variable is measured with error and a classical measurement model is needed. Some examples include in vitro–in vivo correlations, such as the relationship between $\log P$ and volume of distribution (Kaul and Ritschel 1990), in vivo clearance estimates based on in vitro microsomal enzyme studies (Iwatsubo et al. 1996, 1997), or the estimation of drug clearance based on creatinine clearance (Bazunga et al. 1998; Lefevre et al. 1997). In these three examples, $\log P$, in vitro clearance, and creatinine clearance, all have some measurement error associated with them that may be large enough to produce significantly biased regression parameter estimates.

Before a solution to the problem is presented, it is necessary to examine what happens when the measurement error in x is ignored and the SLR model applies. When a classical error model applies, the effect of measurement error in x is attenuation of the slope and corresponding inflation of the intercept. To illustrate this, consider the linear model $Y = x + 10$ where x is a set of triplicate measurements at $\{50, 100, 250, 500, 750, 1,000\}$. Y is not measured with error, only x has error. Figure 9 plots the resulting least squares fit with increasing measurement error in x . As the measurement error variance increases, the slope of the line decreases with increasing intercept. Sometimes the attenuation is so severe that bias correction techniques must be used in place of OLS estimates.

Let us assume the SLR model applies, where x has mean μ_x and variance σ_x^2 and $\varepsilon \sim N(0, \sigma^2)$. The predictor x cannot be observed, but X can, where $X = x + \delta$ with δ being the difference between the observed and true values having mean 0 and variance σ_k^2 . Thus, the total variance of X is $\sigma_x^2 + \sigma_k^2$. Then the OLS estimate of the slope of Y on X is not $\hat{\theta}_1$, but

$$\hat{\theta}_1^* = \lambda \hat{\theta}_1, \quad (92)$$

where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_k^2} < 1. \quad (93)$$

The denominator in (93) represents the total variability of X , whereas the numerator is the variability in x , the true values. λ is sometimes called the attenuation factor or reliability factor and represents the proportion of variation in x found in X . The net effect of measurement variance of the predicted values is greater than when x has error in x is that $\hat{\theta}_1$ is attenuated toward zero and the no measurement error. Corresponding to this is that as the slope decreases, the intercept increases in response.

Measurement error causes double-trouble: attenuation of the slope and increased error about the regression line. However, when more complex error structures are assumed, such as when X is not an unbiased estimate of x or the variance of δ depends on x , then it is possible for the opposite effect to occur, e.g., $\hat{\theta}_1$ is inflated (Carroll et al. 1995). Rarely are these alternative measurement error models examined, however. The bottom line is that measurement error in the predictors leads to biased estimates of the regression parameters, an effect that is dependent on the degree of measurement error relative to the distribution of the predictors.

Hodges and Moore (1972) showed for the linear model, the maximum bias introduced by measurement error in the predictors, assuming an additive error model, can be estimated by

$$\text{bias} = \hat{\theta} - (n - p - 1)(x^T x)^{-1} U \hat{\theta}, \quad (94)$$

where

$$U = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ & & \dots \\ 0 & & & \sigma_k^2 \end{bmatrix} \quad (95)$$

with the diagonal elements of U being the measurement error variance for the k th predictor variable. Bias estimates can be transformed to relative bias estimates by

$$\text{relative bias} = \frac{\text{bias}}{\hat{\theta}} \times 100\%. \quad (96)$$

If (96) indicates that severe bias is present in the parameter estimates, then the parameter estimates need to be bias-corrected. It should be mentioned, however, that correcting for bias is not without its downside. There is a trade-off involved, the bias vs. variance trade-off, which states that by correcting for bias in measurement error models, the variance of the unbiased estimator increases relative to the biased estimator leading to larger confidence intervals. In general, for large sample sizes and for moderate attenuation correction, bias correction is beneficial. The reader is referred to Fuller (1987) for further details.

In the case of SLR when λ is known, an unbiased estimate of the slope can be obtained by rearrangement of (92), i.e.,

$$\hat{\theta}_1 = \frac{\hat{\theta}_1^*}{\lambda}. \quad (97)$$

Stefanski et al. (Carroll et al. 1995, 1996; Cook and Stefanski 1994; Stefanski and Cook 1995) present a “remeasurement method” called simulation-extrapolation (SIMEX), which is a Monte Carlo approach to estimating and reducing measurement error bias, in the same vein as the bootstrap is used to estimate sampling error. The advantage of the SIMEX algorithm is that it is valid for linear and nonlinear models and for complex measurement error structures, included heteroscedastic variance models. The method

assumes that σ_k^2 , the variance of the measurement error, is known to some degree of certainty. If σ_k^2 is not known, then it must be estimated. If no estimate of σ_k^2 can be obtained, then no method can be used to obtain unbiased parameter estimates.

The basic idea is to add random measurement error to the predictor variables using Monte Carlo and develop the relationship between measurement error and parameter estimates. Using this relationship, the parameter estimates for the case of no measurement error can then be extrapolated. When asked what does SIMEX offer over other methods in reducing the bias of parameter estimates in regression models, Stefanski (personal communication) responds by asking “does the bootstrap offer any advantage for computing the standard error of the sample mean?” Thus, SIMEX is analogous to bootstrap methods, i.e., it may be over-kill for simple problems or it may be the only solution but for complex problems.

SIMEX is easiest to understand in the linear regression case and its exposition will be as described by Carroll et al. (1995). Begin by assuming the simple linear model. Recall that σ_k^2 represents the variance in x with no error and σ_k^2 is the measurement variance of X . Now suppose that there are $m - 1$ additional data sets in addition to the original data with each of these additional data sets having successively larger measurement error variances, i.e., $(1 + \lambda_m)\sigma_k^2$ where $0 = \lambda_1 > \lambda_2 > \lambda_3 > \dots > \lambda_m$. Then for each of these datasets the slope of the m th data set, $\hat{\theta}_{1,m}^*$, does not consistently estimate $\theta_{1,m}$ but instead estimates

$$\hat{\theta}_m^* = \frac{\theta_k \sigma_x^2}{\sigma_x^2 + (1 + \lambda_m)\sigma_k^2}. \quad (98)$$

This problem can now be thought of as a nonlinear regression problem where $\hat{\theta}_m^*$ is regressed against λ_m . The regression parameters in the absence of measurement error can be obtained by extrapolating λ to -1 . However, modeling (98) is not practical since σ_k^2 and σ_k^2 may not be known. Carroll et al. (1995) suggest that in their experience it is much easier to regress λ_m against $\hat{\theta}_{1,m}^*$ using a quadratic polynomial

$$\hat{\theta}_{1,m}^* = \gamma_0 + \gamma_1 \lambda_m + \gamma_2 \lambda_m^2 \quad (99)$$

evaluated over the equally spaced interval $0 < \lambda_m \leq 2$. Estimation of the standard error of SIMEX parameter estimates can be calculated using the bootstrap or jackknife, a process which should not increase computing time to prohibitive levels given the current processor speed of most personal computers.

Therefore, the SIMEX algorithm is as follows. First a simulation step is performed:

1. Define $X(\lambda_m) = X_i + \sqrt{\lambda_m} \sigma_k Z$ where Z are independent, random variates with mean zero and variance 1.

Table 12

Desirudin clearance as a function of creatinine clearance

Creatinine CL (mL/min)	Desirudin CL (mL/min)
8.22	13.61
9.79	17.33
25.07	16.09
24.28	19.80
25.07	23.51
27.42	27.23
36.19	29.21
44.41	47.03
44.26	56.93
58.75	70.54
63.45	133.66
76.37	105.20
82.25	134.90
82.64	141.09
93.21	102.72
96.34	170.79
107.70	148.51
105.74	170.79
106.14	199.26
111.23	195.54
125.72	170.79

2. For each data set, regression is done and the parameter estimates saved.
3. Repeat steps 1 and 2 many times (>100).
4. Calculate the average parameter estimate.
5. Following the extrapolation step regress the average parameter estimates vs. λ using a quadratic polynomial.
6. Extrapolate to $\lambda = -1$.

If the error assumed for X is not normally distributed, a suitable transformation needs to be found prior to performing the algorithm. Carroll et al. (1995) stress that the “extrapolation step should be approached as any other modeling problem, with attention paid to the adequacy of the extrapolant based on theoretical considerations, residual analysis, and possible use of linearizing transformations and that extrapolation is risky in general even when model diagnostics fail to indicate problems.”

As an example, consider the data presented by Lefevre et al. (1997). In that study, eight healthy subjects with normal renal function and 15 patients with varying degrees of renal impairment were given an infusion of desirudin, the recombinant form of the naturally occurring anticoagulant hirudin, found in the European leech *Hirudo medicinalis*,

with doses ranging from 0.125 to 0.5 mg/kg infused over a 30-min period. Serial blood samples were collected and the clearance of hirudin was calculated using noncompartmental methods. The raw data values were not reported in the publication, but the data were presented as a plot. This plot was reanalyzed by taking the X - Y coordinates for each data point and determining the associated X - Y value.

The reanalyzed data are presented in Table 12 and plotted in Fig. 10. Lefevre et al. (1997) reported that plasma clearance (CL) of hirudin could be related to creatinine clearance (CrCL) by the equation: $CL = 1.73 \times CrCL - 17.5$. The reanalyzed model gave OLS estimates of 1.71 ± 0.13 for the parameter associated with CrCL and -15.1 ± 9.3 mL/min for the intercept ($R^2 = 0.9058$, $MSE = 442.9$). Ignore for the moment that a better model might be a no-intercept model. Residual analysis suggested that the residuals were normally distributed and that data weighting was unnecessary.

In order to use the SIMEX algorithm on this data, an estimate of the measurement variance for creatinine clearance must be obtained. In discussions with clinical chemists, the upper limit of measurement error associated with measuring serum or urine creatinine using the Jaffe reaction is 5%. Assuming mean 24 h values for urinary volume, urinary daily creatinine excretion, and serum creatinine of 1,000 mL, 1.5 g, and 1.1 mg/dL, respectively, an approximate measurement error variance for creatinine clearance was found to be 60 (mL/min)^2 .

With this as an estimate of the assay measurement variance, the SIMEX algorithm was applied. Figure 11 plots the mean regression parameter against varying values of λ using 1,000 iterations for each value of λ . Extrapolation of λ to -1 for both the slope and intercept leads to a SIMEX equation of

$$CL = -19.4 + 1.78 \times CrCL, \quad (100)$$

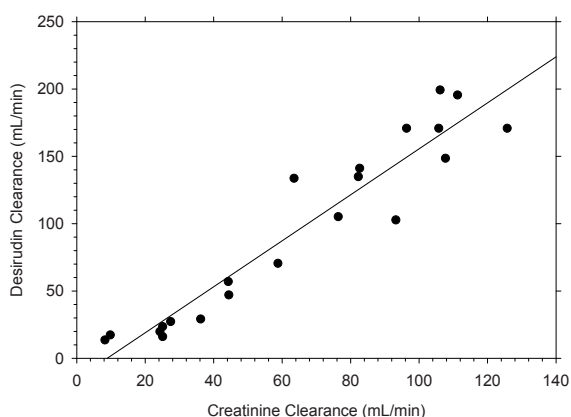


Fig. 10 Plot of desirudin clearance as a function of creatinine clearance. Data redrawn from Lefevre et al. (1997). Solid line is the ordinary least squares fit

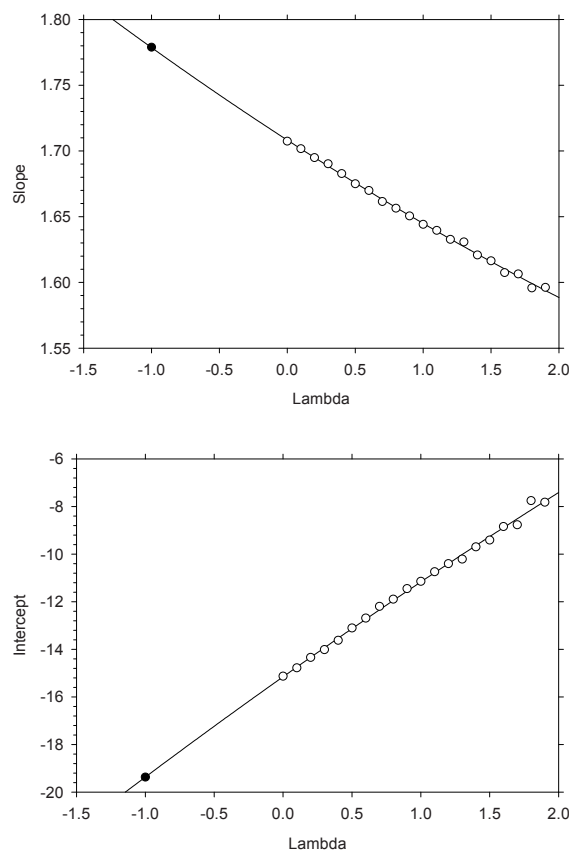


Fig. 11 Plot of SIMEX extrapolation to desirudin data show in Fig. 10. extrapolated value for slope (top) and intercept (bottom) at $\lambda = -1$; o, mean regression parameter using 100 iterations for each value of λ ; solid line is the second order polynomial fit

values not too different from the OLS estimates. The bias of the OLS estimates for slope and intercept using (94) was 0.06 and -4.1 mL/min, respectively, with a relative error of 23 and 4%, respectively. The jackknife SIMEX estimates for slope and intercept were 1.90 ± 0.43 (mean \pm standard error of mean) and -21.0 ± 4.9 mL/min, respectively. Hence, the OLS estimates in this case were relatively unbiased.

Surprisingly, even though the parameter estimates obtained from regression of independent variables with measurement errors are biased, one can still obtain unbiased prediction estimates and corresponding confidence intervals. The reason is that even though X has measurement error, the model still applies to the data set on hand. The problem arises when one wishes to make predictions in another population or data set. In this case, three options are available (Buonaccorsi 1995). First, carry out the regression of Y on X and calculate the predicted response ignoring the measurement error. Second, regress Y on X , recognizing that X is measured with error, but obtain a modified estimate of σ^2 , and calculate a modified prediction interval. Third, correct for the measurement error of X and

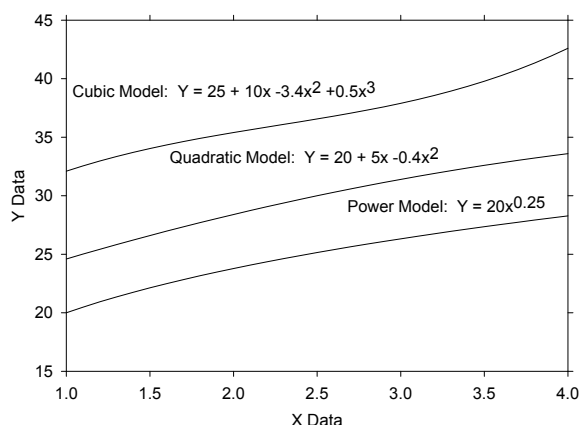


Fig. 12 Plot of a cubic, quadratic, and power function

regress Y against the corrected X . The prediction interval then uses the parameters obtained from the corrected regression. Options 1 and 2 are reasonable assuming that the value to be predicted has the same measurement error distribution as the current data.

Buonaccorsi (1995) present equations for using Option 2 or 3 for the SLR model. In summary, measurement error is not a problem if the goal of the model is prediction, but keep in mind the assumption that the predictor data set must have the same measurement error distribution as the modeling data set. The problem with using option 2 is that there are three variance terms to deal with: the residual variance of the model, σ^2 , the uncertainty in θ , and the measurement error in the sample to be predicted. For complex models, the estimation of a corrected σ^2 may be difficult to obtain.

Polynomial Regression

Sometimes one sees in the literature models of the form

$$Y = \theta_0 + \sum_{k=1}^m \theta_k x_k + \sum_{l=m+1}^{p-1} \theta_l x_l^q + \varepsilon, \quad (101)$$

where q is the power term, being described as “nonlinear models.” This is in error because this model is still linear in the parameters. Even for the terms of degree higher than 1,

$$\frac{\partial}{\partial \theta_l} = q x_l^{q-1}, \quad (102)$$

which means that the parameters are independent of other model parameters. What may confuse some people is that polynomial models allow for curvature in the model, which may be interpreted as nonlinearity. Since polynomial models are only special cases of the linear model, their fitting requires no special algorithms or presents no new problems.

Often a polynomial may be substituted as a function if the true model is unknown. For example, a quadratic model may be substituted for an E_{\max} model in a pharmacodynamic analysis or a quadratic or cubic function may be used in place of a power function

$$Y = \theta_1 X^{\theta_2} \quad (103)$$

as shown in Figure 12. It is almost impossible to distinguish the general *shape* of the quadratic model and power model. The change in intercept was added to differentiate the models graphically. Also note that an increase in the number of degrees of freedom in the model increases its flexibility in describing curvature as evidenced from the cubic model in Fig. 12.

Polynomial model development proceeds the same as model development when the degree of the equation is 1. However, model development generally proceeds first from simpler models and then terms of higher order are added later. Hence, if a quadratic term is added to a model, one should keep the linear term as well. The function of the linear term is to provide information about the basic shape of the curve, while the function of the quadratic term is to provide refinements to the model. The LRT or information criteria can be used to see if the additional terms improves the goodness of fit. Extreme caution should be made in extrapolating a polynomial function as the function may deviate significantly from the interval of data being studied. Also, higher order models (greater than 2) are usually avoided because, even though they often provide good fits to the data, it is difficult to interpret their coefficients and the predictions they make are often erratic. When two or more predictors are modeled using quadratic polynomials, response surfaces of the type shown in Fig. 13 can be generated. These are extremely useful in examining how two variables interact to generate the response variable. They are also very useful for detecting

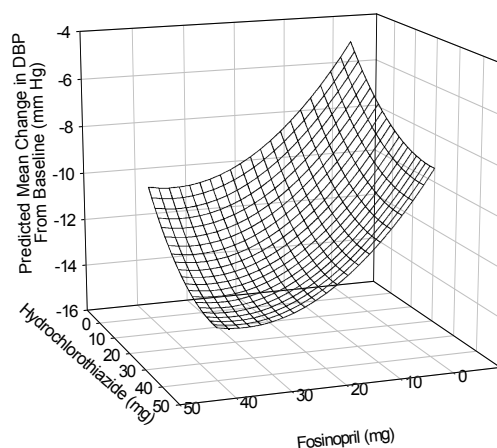


Fig. 13 Quadratic response surface of predicted mean change in diastolic blood pressure following 8 weeks of randomized therapy to fosinopril and/or hydrochlorothiazide. Data presented in Pool et al. (1997)

and characterizing antagonism or synergy between drug combinations. Although not used commonly clinically, response surfaces are useful both in vitro and in vivo models. See Greco et al. (1995) for details and Carter et al. (1985), Rockhold and Goldberg (1996), and Stewart (1996) for examples.

Smoothers and Splines

Linear regression models are of the form

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p + \varepsilon \quad (104)$$

ignoring higher order polynomial and interaction terms. Additive models are of the form

$$Y = f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \varepsilon, \quad (105)$$

where $f_1(\cdot), f_2(\cdot), \dots, f_p(\cdot)$ are now generic smooth functions that do not have to be of a linear or nonlinear form or even of the same functional form. Linear models and additive models can be combined to form semiparametric models of the form

$$Y = \theta_0 + \theta_1 x_1 + f(x_2) + \cdots + f(x_p) + \varepsilon, \quad (106)$$

where the model consists of both a parametric structural form and an additive structural form. Semiparametric models are sometimes also called partially linear models, partial linear models, partly linear models, or partial spline models. Additive and semiparametric models can be further extended through generalized additive models (GAMs) to allow for categorical dependent variables or survival data similar to how generalized linear models extend linear models. When the structural model contains some elements which are additive in nature, the exact nature of the function $f(\cdot)$ or the parameter estimates of $f(\cdot)$ may not be of interest, simply whether a “relationship” between x and Y exists. So unlike linear models, inference on the model parameters in a GAM or semiparametric model is usually not of interest. The purpose of this chapter will be to introduce semiparametric models and smoothing models and to illustrate how semiparametric or additive models can be easily developed using linear mixed effect methodology.

Smoothers and Locally Weighted Regression (LOESS)

A smoother describes the trend in Y as a function of some set of predictors and are generally nonparametric in nature. If only one predictor variable is available, these smoothers are called scatterplot smoothers. Of importance is that the smoother does not assume a rigid structural form like in (104). Smoothers work through the concept of local averaging or neighborhoods, i.e., the predicted value is based on observations near the reference value as opposed to linear or nonlinear models which base predictions on the totality of the data. By assuming a nonparametric form the smoother can become linear in parts of a curve and curvilinear in other parts.

Two main decisions must be made for any smoother. First, how big should the neighborhood be around the reference value and then, second, how should the predicted response at the reference value be calculated within each

neighborhood. How the predicted value is estimated within the neighborhood is what distinguishes the types of smoothers: running mean, running median, exponential, LOESS, etc. The size of the neighborhood can vary from very small to very large. As the size of the neighborhood increases, the curve becomes smoother but flatter (variance decreases but bias increases). As the size of the neighborhood decreases, the curve becomes more jittery and not very useful (variance increases but bias decreases). So clearly there will be some optimum neighborhood that minimizes the bias-variance trade-off. Many algorithms exist for finding the optimum neighborhood, such as cross-validation, but many times the choice of the neighborhood is simply based on graphical examination and the analyst's discretion.

The idea behind a smoother will begin through the exposition of running mean and running median smoothers. First, the x observations are ordered from smallest to largest. Starting at $x_{(1)}$, the smallest value of x , a neighborhood of k observations near $x_{(1)}$ is chosen. Few software packages require the neighborhood be defined in terms of k , the number of observations in the neighborhood. Rather it is more convenient to have each neighborhood consist of some proportion of the total number of observations

$$w = \frac{(2k+1)}{n} \quad (107)$$

a value referred to as the span. So, if the span was defined as the nearest 10% of the observations to $x_{(i)}$ and the total number of observations was 100, then the neighborhood around $x_{(i)}$ would consist of the ten nearest observations. For a running mean smoother, the average of Y in the neighborhood of $x_{(1)}$ is calculated and used as the first predicted value of the smoothed line. This process is repeated for $x_{(2)}, x_{(3)}, \dots, x_{(n)}$. The predicted values are then joined by a line segment and the entire line is called the running mean smoother. A running median smoother uses the same algorithm except the median Y -value is calculated within each neighborhood and used as the predicted value.

The neighborhood around $x_{(i)}$ can be based on either the nearest symmetric neighbors in which the $k/2$ observations to the left and $k/2$ observations to the right of $x_{(i)}$ are chosen as the neighborhood. In the case where $x_{(i)}$ is near the tail of x and it is not possible to take all the points both to the right and left of $x_{(i)}$, as many observations as possible are taken for the calculation. Alternatively, symmetry can be ignored and the nearest neighborhood may consist of the nearest neighbors to $x_{(i)}$ regardless of whether the observations are to the right or left of $x_{(i)}$. Hastie and Tibshirani (1990) suggest that nearest neighborhoods are preferable to symmetric neighborhoods because in a neighborhood with a fixed number of observations the average distance of the observations to the reference value is less with nearest neighborhoods, unless the observations are equally spaced, resulting in less bias in the predictions.

Running mean and median smoothers, while often seen in time series analysis, are not very useful because the smoothed line tends to be too jittery to be useful and tends

to flatten out at near the tails of x leading to large bias in the fit. However, a simple solution to the bias problem exists. Instead of computing the mean in the neighborhood, a running line smoother is computed where within each neighborhood ordinary least squares linear regression is used to compute the predicted value at $x_{(i)}$. Cleveland (1979) further improved on the algorithm by suggesting that within each neighborhood weighted least squares linear regression (which will be discussed in the chapter on “Variance Models and Transformations”) be used to predict the value at $x_{(i)}$ where the weights decrease smoothly away from $x_{(i)}$. This algorithm is more commonly referred to as the Locally Weighted Scatterplot Smoother (LOWESS) or LOESS algorithm.

The algorithm proceeds by starting at $x_{(1)}$ and calculating the distance from $x_{(1)}$ for each $x_{(i)}$

$$\Delta_{(i)} = |x_{(i)} - x_{(1)}|, \quad (108)$$

where $|\cdot|$ is the absolute value function. The k nearest neighbors having the smallest $\Delta_{(i)}$ are then identified, as is the observation having the largest $\Delta_{(i)}$, denoted $\max(\Delta)$. A scaled distance is then calculated as

$$u_{(i)} = \frac{\Delta_{(i)}}{\max(\Delta)}. \quad (109)$$

Once the scaled distances are calculated, the tricube weight function is formed for each $x_{(i)}$ in the neighborhood of $x_{(1)}$

$$W_{(i)} = \begin{cases} (1 - u_{(i)}^3)^3 & \text{for } u_{(i)} < 1 \\ 0 & \text{for } u_{(i)} \geq 1 \end{cases}. \quad (110)$$

Weighted linear regression using the above weights is then performed on the observations in the neighborhood and the Y value at $x_{(1)}$ (denoted $Y_{(1)}$) is predicted. This process is then repeated on $x_{(2)}$, $x_{(3)}$, ..., $x_{(n)}$ replacing $x_{(1)}$ in the above equations with $x_{(2)}$, $x_{(3)}$, etc. The LOESS smoother then joins each predicted $Y_{(i)}$ by a line segment. Figure 14 illustrates the concepts just described for a single observation in a data set. Each observation in the neighborhood does not contribute equally to the predicted $x_{(1)}$. Observations near $x_{(1)}$ have greater weight than observations within the neighborhood but further removed from $x_{(1)}$. An optional robustness factor can be built into the model by providing less weight to observations having large residuals.

Figure 15 presents a representative concentration-time profile with a LOESS smooth to the data. To create a LOESS smooth to the data, an analyst must first decide whether the weighted regression model within each neighborhood will be linear or quadratic in nature. Higher order polynomial models are of course possible, but are rarely used. Quadratic models are useful if the data exhibit a large degree of curvature or has many inflection points. Secondly, the proportion of observations in each neighborhood must be selected. The span may range from 0

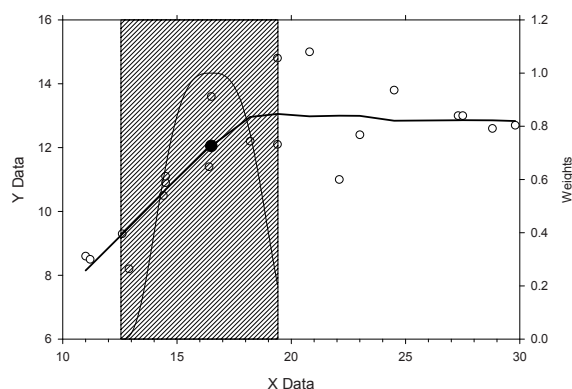


Fig. 14 Example of a LOESS smoother and nearest neighbor concept. Twenty observations were simulated. A LOESS smoother having span 0.5 and using a weighted linear regression model is shown as a heavy-set solid line. The hatched area is the k nearest neighbors (ten observations because $20 \times 0.5 = 10$) to a target value of 16.5. The solid line within the neighborhood is the tricube weight function. The solid dot within the window is the predicted value. Notice that the tricube weight function is not necessarily symmetric around the target value

to 1, but typically, a good default value is 0.3–0.5, i.e., each neighborhood consists of the nearest half to third of the data surrounding $x_{(i)}$. The number of observations in each neighborhood is then $k = w \times n$. If k is not a whole number, the value must be either truncated or rounded. The default span in S-Plus is 2/3. No default value is used in SAS; the smoothing parameter is an optimized one that is data-dependent. If x is sorted from smallest to largest then the physical width of each window may change in size as the smoother proceeds from $x_{(1)}$, ..., $x_{(n)}$, as will the degree of asymmetry. For example, the neighborhood will lie entirely to the right of $x_{(1)}$, but will lie entirely to the left of $x_{(n)}$.

Kernel Smoothers

The smoothers just presented use neighborhoods of constant span with the neighborhood being either the nearest k observations or the nearest symmetric neighbors.

In contrast, kernel smoothers use neighborhoods of constant width or bandwidth (denoted as b) as is more often used (Altman 1992). A kernel is a continuous bounded and symmetric function K that integrates to one. Given a reference point, say $x_{(1)}$, the difference between $x_{(1)}$ and all other $x_{(i)}$ s is scaled to the bandwidth

$$u_{(i)} = \frac{|x_{(i)} - x_{(1)}|}{b}. \quad (111)$$

The scaled difference $u_{(i)}$ is then passed to the kernel function. Popular kernel functions include the Epanechnikov kernel

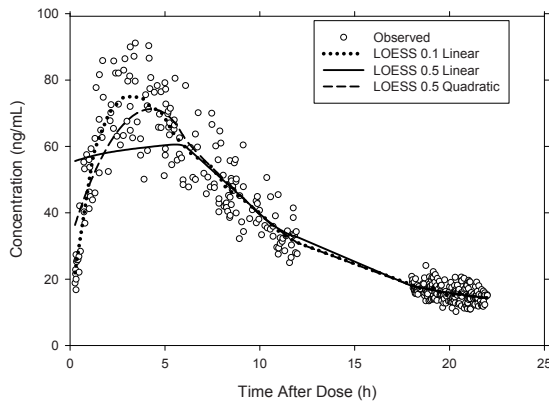


Fig. 15 Concentration data from 97 subjects having from one to four samples collected per patient at steady state. The *dotted*, *solid*, and *dashed* lines are the LOESS fits of varying span and either linear or quadratic regression

$$K(u_{(i)}) = \begin{cases} 0.75(1-u_{(i)}^2) & \text{for } u_{(i)} \leq 1 \\ 0 & \text{for } u_{(i)} > 1 \end{cases} \quad (112)$$

the standard Gaussian kernel

$$K(u_{(i)}) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{u_{(i)}^2}{2}\right] \text{ for all } u_{(i)}, \quad (113)$$

and the quartic kernel

$$K(u_{(i)}) = \begin{cases} \frac{15}{16}(1-u_{(i)}^2)^2 & \text{for } u_{(i)} \leq 1 \\ 0 & \text{for } u_{(i)} > 1 \end{cases} \quad (114)$$

Figure 16 plots the differences in the kernels and how they decrease as they move from their reference point.

In general, the Gaussian kernel weights less drastically than does the Epanechnikov kernel, while the quartic kernel weights most dramatically as the distance from the reference point increases. The weight given to the i th by

$$W_{(i)} = \frac{c}{b} K(u_{(i)}), \quad (115)$$

where c is a constant defined such that the weights sum to unity. The weight given an observation is only dependent on how close the observation is to $x_{(i)}$. The predicted value for $x_{(1)}$ is then the weighted average of all the Y values

$$\hat{Y}_{(1)} = \frac{\sum_{i=1}^n W_{(i)} Y}{\sum_{i=1}^n W_{(i)}}. \quad (116)$$

This process is repeated for all $x_{(i)}$.

Kernel smoothing is more like a weighted running mean smoother, even though it is sometimes referred to as kernel regression. With kernel smoothing, the analyst must choose the bandwidth and the choice of kernel. For constant bandwidth the number of data points in the neighborhood varies from neighborhood to neighborhood.

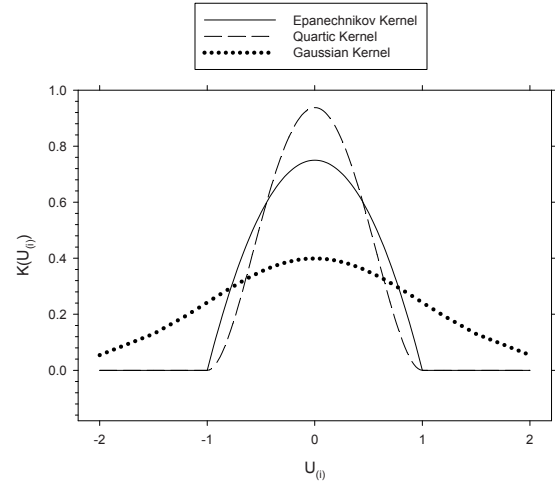


Fig. 16 The Epanechnikov (*solid line*), standard Gaussian (*dotted line*), and quartic kernel (*dashed line*) smoother

As the bandwidth increases the variance of the estimate decreases, but the bias of the predicted fit decreases. Conversely, the weights around the reference value increase as the bandwidth decreases. As for choice of the kernel function, in practice the choice of kernel is relatively unimportant compared to the choice of bandwidth as most kernel functions produce roughly equivalent smooths to the data (Hastie and Tibshirani 1990). The general opinion is that kernel smoothing is inferior to local weighted regression as kernel smoothing suffers from the same boundary bias as running mean smoothers and under performs when the “true” regression function is linear (Ryan 1997).

Spline Interpolation

Another type of smoother sometimes seen in the pharmacokinetic arena is a spline smoother, which comes in many different flavors. Splines have their history in drafting where draftsmen needed to draw a smooth curve through a set of points. To do this, the draftsman would place a piece of paper over a board, hammer in nails or push in pins where the points were, and then a thin piece of wood was interwoven between the points. The result was a smooth curve that passed through each point. This type of spline is referred to as an interpolating splines. The problem with interpolating splines is that the spline passes through every point. Interpolating splines applied to data with replicate values or noisy data are not very appealing and so penalized regression splines, which use a penalized least squares minimization approach to the problem, are often used instead.

The concept of an interpolating spline is best explained through linear splines and then expanded to higher order splines (Chapra and Canale 1998). Given a set of ordered data points $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ a first order spline can be defined through the linear functions

$$\begin{aligned} f(x) &= f(x_{(1)}) + m_1(x - x_{(1)}) & x_{(1)} \leq x \leq x_{(2)} \\ f(x) &= f(x_{(2)}) + m_2(x - x_{(2)}) & x_{(2)} \leq x \leq x_{(3)} \end{aligned} \quad (117)$$

...

$$f(x) = f(x_{(n-1)}) + m_{n-1}(x - x_{(n-1)}) \quad x_{(n-1)} \leq x \leq x_{(n)}$$

where m_i is the slope of the straight line connecting the points

$$m_i = \frac{f(x_{(i+1)}) - f(x_{(i)})}{x_{(i+1)} - x_{(i)}}. \quad (118)$$

The points where two splines meet are called the knots. A first order spline is a simple linear interpolating line to the function, but suffers from the problem that the change from one interpolating line to another is not smooth – the derivatives are discontinuous.

To overcome the problem of discontinuity, higher order splines may be developed. In order for the spline to be smooth at the knots, both the first and second derivative of the spline must exist and be continuous. In general, a spline of at least $m + 1$ must be used for m -derivatives to be continuous and exist. Hence, for both the first and second derivative to be continuous a cubic spline of the following form must be used

$$Y = \theta_0 + \theta_1 x + \theta_3 x^2 + \theta_4 x^3. \quad (119)$$

The following conditions must also be met:

1. The function values must be equal at the interior knots, i.e., the splines must join at the knots.
2. The first and last functions must pass through the first and last observed data points (the end points).
3. The first and second derivatives of the interior knots must be equal.

Under these constraints there are $n - 2$ equations but n unknowns. Thus the spline cannot be solved as is. Two unknowns can be eliminated and the problem solved by imposing some constraints on the end points. Under the constraint that the second derivatives at the end knots equal zero creates what is referred to as a natural spline. The result is that at the end points the end cubics approach linearity and have zero curvature. This type of spline is the mathematical equivalent to the draftsman's spline from earlier. If instead the slope of the first and last cubics at the end points is specifically defined the result is a clamped cubic spline. See Fig. 17 for an example of a natural cubic spline.

Unfortunately, as is apparent in the bottom plot of Fig. 17, interpolating cubic splines may not be valuable in some instances, such as trying to smooth concentration-time data pooled across individuals in a population or when the data are particularly noisy. In this case, regression splines, which do not force the spline curve to interpolate the observed data points, may be more useful. Like interpolating splines, regression splines use piecewise polynomials to interpolate between the knots, the most common polynomial being cubic. As the number of knots increases the flexibility of the spline increases. Thus, a regression spline may pass near the observed data but not be constrained to interpolate it.

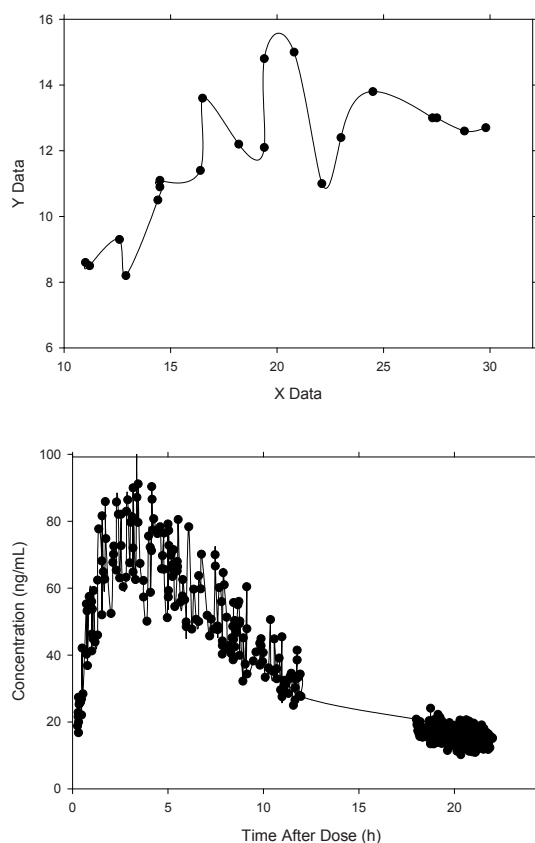


Fig. 17 Interpolating cubic spline fit to the data in Fig. 14 (top) and Fig. 15 (bottom)

Handling Missing Data

Anyone who does data analysis will eventually run into the problem of missing data, either the dependent variable is missing or one or more of the independent variables is missing. The problem of handling missing data is far too complex to cover it in its entirety within this book and many excellent books are available on the subject for readers who wish greater detail. These include books by Allison (2002), Little and Rubin (2002), and Schafer (1997).

It is worthwhile to consider the regulatory opinion on missing data, keeping in mind that these guidances were written with an eye toward formal statistical analysis, such as hypothesis testing, and not with an eye toward pharmacokinetic or pharmacodynamic modeling per se. Having said that, more and more modeling is done to support New Drug Applications and that in the future it is likely that increased scrutiny will be paid toward these issues. ICH E9 (1998) states that the missing data is a potential source of bias and as such every effort should be done to collect the data in the first place. The guidance also recognizes that despite best efforts, missing data is a fact of life in clinical studies. Also, trial results are valid “*provided the methods for dealing with missing data are sensible, ... particularly those pre-defined in the protocol.*” Unfortunately,

no recommendations are made in the guideline on what those “methods” are. The guideline does state that no universal method for handling missing data is available and that any analysis based on data containing missing values should also have a corresponding sensitivity analysis to see what effect the method of data handling has on the analysis results.

The Committee for Proprietary Medicinal Products (CPMP) (2001) has also issued a points to consider document related to missing data that expands on the ICH E9 guideline. The CPMP document is mainly concerned with the issue of bias and how missing data affects detecting and estimating treatment effects. The CPMP does not generally accept analyses where all missing data is deleted and only data with complete cases is analyzed. They recommend that all efforts be directed at avoiding missing data in the first place, something that seems intuitively obvious but needs to be restated for its importance. The CPMP also recommends that whatever method used to handle missing data be stated a priori, before seeing the data, in a data analysis plan or the statistical methods section of the study protocol. The final report should include documentation on any deviations from the analysis plan and defend the use of the prespecified method for handling missing data. Lastly, a sensitivity analysis should be included in the final report indicating the impact of the missing data handling procedure on treatment outcomes. This may be as simple as a complete case analysis vs. imputed data analysis (which will be discussed later).

Types of Missing Data and Definitions

Little and Rubin (2002) define three types of missing data mechanisms. The first and most restrictive is missing completely at random (MCAR) in which cases that are missing are indistinguishable from cases that have complete data. For example, if a sample for drug analysis was broken in the centrifuge after collection and could not be analyzed then this sample would be MCAR. If the data are MCAR then missing data techniques such as casewise deletion are valid. Unfortunately, data are rarely MCAR.

Missing at random (MAR), which is a weaker assumption than MCAR, is where cases of missing data differ from cases with complete data but the pattern of missingness is predictable from other variables in the dataset. For example, suppose in a Phase 3 study all patients at a particular site failed to have their weight collected at study entry. This data would be MAR because the missingness is conditional on whether the data were collected at a particular site or not. When data are MAR, the missing data mechanism is said to be “ignorable” because the missing data mechanism or model is independent of the parameters to be estimated in the model under consideration. Most data sets consist of a mixture of MCAR and MAR.

If data are missing because the value was not collected then that value is truly missing. In more statistical terms, if the data are missing independent of the actual value of the

missing data then the missing data mechanism is said to be *ignorable*. If, however, data are missing because their value is above or below some level at which obtaining quantifiable measurements is not possible then this type of missing data is an entirely different problem. In this case, the data are missing because of the actual value of the observation and the missing data mechanism is said to be nonignorable. These last type of data are extremely tricky to handle properly and will not be discussed in any great detail herein. The reader is referred to Little (1995) and Diggle and Kenward (1994) for details.

Last, is the pattern of missingness as it relates to missing covariates. Figure 18 presents a schematic of the general pattern of missingness. Some covariates have missing data, others do not. There may be gaps in the covariates. But if the covariates can be re-arranged and re-ordered x_1, x_2, \dots, x_p , such that the degree of missingness within each covariate is less than the preceding covariate then such a pattern of missingness is monotonic or nested. Monotonic missingness is useful because there are specific ways to impute monotonic missing data.

Methods for Handling Missing Data: Missing Dependent Variables

If the missing data are the dependent variable and the reason for missingness is not nonignorable then the missing data should be deleted from the analysis. If however, the missing dependent variable is missing because it is below or above some threshold value then more complicated methods to analyze the data are needed. For instance, the dependent variable may be missing because its value was below the lower limit of quantification (LLOQ) of the method used to measure it. For example, white blood cell count may be near zero after chemotherapy and may not be measurable using current technologies. In this case, the value may be reported as $<0.1 \times 10^9$ per liter. In such a case, the true value for white blood cell count lies between 0 and the LLOQ of the assay. Such data are said to be censored. Another instance might be when the value exceeds some threshold beyond which an accurate and quantifiable measurement cannot be made, such as determining the weight of a super-obese individual whose weight exceeds the limit of the scale in a doctor’s office. In this case only that the subject’s weight was larger than c , the upper limit of the scale, is known. A value is said to be censored from below if the value is less than some threshold or censored from above if the value exceeds some constant. With censored data, the usual likelihood function does not apply and parameter estimates obtained using maximum likelihood will be biased.

For data where the dependent variable is censored, the log-likelihood function is the sum of the log-likelihoods for observations not censored plus the sum of the log-likelihoods for censored observations. To obtain parameter

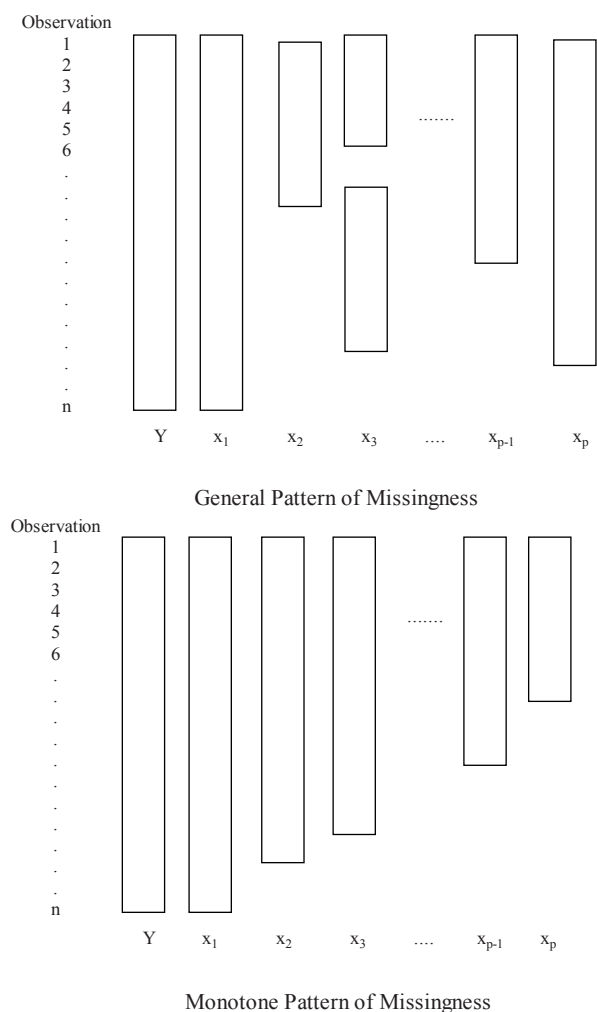


Fig. 18 General (*top*) and monotonic (*bottom*) pattern of missingness in the covariates

estimates one needs access to an optimization package that can fit a general likelihood function, like MATLAB (The MathWorks Inc., Natick, MA). A simpler more pragmatic (but certainly more biased) approach in the case where observations are censored from below, imputation is usually done by setting the value equal to zero, equal to some fraction of the constant, such as one-half the LLOQ of the assay, or randomly assigning the data point a value based on a probability distribution. For instance, a sample may be randomly drawn from the interval $[0, \text{LLOQ}]$ based on a uniform distribution. Observations censored from above are more problematic because there may be no theoretical upper limit and in such cases, imputation is usually done by setting the missing value equal to the upper threshold. Whatever the imputation method used, the usual caveats apply. The reader is referred to Breen (1996) for a good exposition to the problem. Unfortunately at this time, no major statistical package, such as SAS or S-Plus, or pharmacokinetic software package can handle the censored data case using the correct log-likelihood equations.

Methods for Handling Missing Data: Missing Independent Variables

There are many different ways for handling missing data including ignore the missing data (complete case analysis), mean or median substitution, hot deck methods, regression methods, and maximum likelihood and its variants. The simplest method, called listwise deletion or complete case analysis, is to ignore the missing data and model only the data that have no missing data. The advantages of this method are that it can be applied to any type of statistical model and is easy to do. Hence, casewise deletion is the method of choice for handling missing data in most statistical software packages. A disadvantage of this method is that it may lead to biased results, especially if the data are not MCAR, but are MAR, such as if the data were more likely to be missing because of assignment to a particular treatment arm. If the data are MCAR, then the model parameters will be unbiased but the standard errors will be larger due to a reduced sample size. Hence, power will be decreased at detecting significant treatment effects. The CPMP does not generally accept listwise deletion analysis because it violates the intent to treat principle.² The Points to Consider document does state, however, that listwise deletion may be useful in certain circumstances, such as in exploratory data analysis and confirmatory trials as a secondary endpoint, to illustrate the robustness of other conclusions.

Imputation, which is basically making up data, substitutes the made-up data into the missing data and treats the imputed data as if it were real. Imputation is generally recognized as the preferred approach to handling missing data and there are many different ways to impute missing data. The first approach is naïve substitution wherein the mean or median value is substituted for all missing values. For example, if a person's weight was missing from a data set then the mean weight, perhaps stratified by sex, would be substituted. While preserving the mean of the marginal distribution of the missing variable, it biases the distribution of the variable. The result is that if the variable is indeed related to the dependent variable and the proportion of missing data is large, then naïve substitution may distort the relationship between variables. It is generally recognized that this approach does more harm than good, unless the proportion of missing data is small (less than a few percent), where at best the substitution adds no information.

If the missing value is one of the independent variables then naïve substitution ignores any correlations that may be present among predictor variables. To account for any correlations between variables, conditional mean imputation may be used wherein for cases with complete data the variable with missing data is regressed against the other

² The intent to treat principle essentially states that all patients are analyzed according to the treatment they were randomized to, irrespective of the treatment they actually received. Hence, a patient is included in the analysis even if that patient never received the treatment.

predictor variables and then the predicted value is substituted for the missing value. In general, all variables are used in the analysis and no attempt is made to reduce the imputation model to its simplest form.

A variant of naïve substitution is to use random substitution wherein an observation is randomly sampled from the observed values and substituted for the missing value. This approach too tends to maintain the mean on-average but may obscure real relationships among the variables. Another variant of naïve substitution is hot-deck imputation, which requires pretty large data sets to be useful and has been used for many years by the U.S. Census Bureau. The basic idea is that each missing value is randomly replaced from other subjects having similar covariates. Suppose the weight of a 67-year-old male was missing, but weight was collected on three other 67-year-old males in the study, then the weight of the missing value is randomly drawn from one of the three observable weights. The advantage of the method is that it imputes realistic values since the imputed value is itself actual data and is conceptually simple. But what if there were no other 67-year-old males in the study. How would the imputation work? This is where hot deck is often criticized, in the choice of the “donor” cases since one then must set up “similarity” criterion to find matching donors. Also, besides SOLAS (Statistical Solutions, Saugus MA), no other software package has a built-in hot deck imputation algorithm. The user must program their own filters and similarity measures which makes the method data-specific and difficult to implement.

Regression-based methods impute the missing values using least-squares regression of the missing covariate against the observed covariates (Little 1992). In other words, the missing covariate becomes the dependent variable and the other covariates with no missing data become the independent variables. Ordinary least-squares, or sometimes weighted least-squares that downweights incomplete cases, is then used to obtain the regression model and the missing value is imputed based on the predicted value. A modification of this approach is to add random error to the predicted value based on the residual mean square error to account for unexplained variability. Little (1992) suggests that when the partial correlation between Y and the observed x s is high then a better imputation can be had by including Y , as well as the observed x s, in the imputation process. This may seem like cheating but if Y is not included in the imputation then biased parameter estimates may result using the filled-in data.

If the covariates show a monotone pattern of missingness (Fig. 18) then the imputation procedure can be done sequentially. For instance, suppose that x_1, x_2, x_3 , and x_4 are the covariates that exhibit monotone missingness and that x_1 and x_2 have no missing data. In the first step, x_3 would be imputed based on the regression of x_3 against Y, x_1 , and x_2 . Then given imputed values for x_3 , x_4 would be imputed using the regression of Y, x_1, x_2 , and x_3 against x_4 . In this manner all the covariates can be imputed. One problem that

may arise using regression-based methods is that the covariates may show collinearity. The covariate design matrix used in the imputation may be near singular with the resulting parameter estimates showing instability. A check of the correlation matrix prior to imputation may be useful to detect which covariates show collinearity. Collinearity could also arise if one or more of the covariates show excessive skewness. In which case, a transformation to normality may be useful prior to imputation.

A parametric method for handling missing data is maximum likelihood. Recall that in linear regression maximum likelihood maximizes the likelihood function $L(\cdot)$

$$L(Y | \theta, \sigma) = \prod_{i=1}^n f(Y_i | \theta), \quad (120)$$

where f is the probability density function. In the case where missing data are present the likelihood function becomes the entire sample

$$L(Y | \theta, \sigma) = \prod_{i=1}^m f(x_i, Y_i | \theta) \prod_{i=m+1}^n g(Y_i | \theta), \quad (121)$$

where g is the probability density function for the missing data and there are m cases with observed data and $n - m$ cases of missing data. The problem then becomes to find the set of θ that maximizes the likelihood. In order to maximize the likelihood certain distributional assumptions must be made, the most common being a multivariate normal distribution. Although direct maximization of the likelihood is possible, the software to do such maximization is not readily available.

Two alternatives to direct maximization of the likelihood are available: the EM algorithm, which is the default multiple imputation (MI) algorithm in SAS, and Markov chain data augmentation. The expectation-maximization (EM) is difficult to explain in lay terms, but in brief, the EM approach to missing data proceeds in two steps. In the first step, the expectation step, the algorithm essentially computes a regression-based imputation to the missing values using all available variables. After the expectation step, the maximization step computes new estimates of the likelihood as if the variable had no missing data. Then the E-step is repeated, etc., until stability of the estimates is obtained.

Data augmentation using Markov Chain Monte Carlo (MCMC), which has its basis in Bayesian statistics, is much like the EM algorithm except that two random draws are made during the process. Markov chains are a sequence of random variables where the current value depends on the value of the previous step. In the first step, starting values are made. For a multivariate normal model, the starting values are the means and covariance matrix or the means and covariances obtained using the EM algorithm. For each missing variable, given the estimates of the mean and covariance, estimates of the regression parameters relating the variable with missing data to the other variables are obtained. Using the regression estimates, the predicted values for all missing data are calculated. Then (and this is

the first random, stochastic step in the process) normally distributed random variability is added to the predicted values and substituted for the missing data. The means and covariances for the imputed data set are then computed. Based on these updated means and covariances (and this is the second random stochastic step in the process) a random draw from the posterior distribution of the means and covariances is made. Using the randomly drawn means and covariances, the entire process is repeated until convergence is achieved. The imputations obtained at the final step in the process are those that are used in the statistical analysis.

Related to MCMC are two fundamental issues: how many iterations are needed before convergence is achieved and what posterior distribution should be used. There is no satisfactory answer for whether or not convergence (or stationarity) has been achieved. The default in SAS is to use 50 burn-in iterations before the first imputation is available for use. Schaffer (1997) used anywhere from 50 to 1,000 iterations in examples used in his book. Of course the more iterations the better, but increasing the number of iterations also increases the computation time, which may become prohibitive. Allison (2002) suggests that as the proportion of missing data increases the number of iterations should increase. If only 5% of the data are missing then fewer iterations are needed, although typically 500–1,000 iterations is usually seen in the literature for most realistic data sets. The reader is referred to Gelman et al. (1995) for further details on MCMC and convergence. The second fundamental issue related to MCMC is the choice of the posterior distribution. In order to obtain the posterior distribution, one needs a prior distribution, which is a probability distribution associated with the prior beliefs of the data before actual collection of any data. An uninformative prior is often used in the absence of any prior knowledge, which is what SAS does as a default.

The problem with any imputation method wherein a single value is substituted for the missing data and then the data set is analyzed as if it were all complete cases is that the standard errors of the model parameters are underestimated because the sampling variability of the imputed values is not taken into account. For this reason multiple imputation arose. With multiple imputation many different datasets are generated, each with their own set of imputed values, and each imputed data set is analyzed as if complete. The parameter estimates across data sets are then combined to generate improved estimates of the standard errors. Multiple imputation, when done correctly, can provide consistent, asymptotically normally distributed, unbiased estimates of model parameters given the data are MAR. Problems with multiple imputation include generation of different parameter estimates every time it is used, is difficult to implement if not built into a statistical package, and is easy to do the wrong way (Allison 2002).

Rubin (1987) proposed that if m -imputed data sets are analyzed that have generated m -different sets of parameter estimates then these m -sets of parameter estimates need to be combined to generate a set of parameter estimates that

takes into account the added variability from the imputed values. He proposed that if θ_i and $SE(\theta_i)$ are the parameter estimates and standard errors of the parameter estimates, respectively, from the i th imputed data set, then the point estimate for the m -multiple imputation data sets is

$$\theta_{MI} = \frac{1}{m} \sum_{i=1}^m \theta_i. \quad (122)$$

So the multiple imputation parameter estimate is the mean across all m -imputed data sets. Let $\bar{U}(\theta_i)$ be the variance of θ_i , i.e., the standard error squared, averaged across all m -data sets

$$\bar{U}(\theta_i) = \frac{1}{m} \sum_{i=1}^m [SE(\theta_i)]^2 \quad (123)$$

and let $\bar{B}(\theta_i)$ be the variance of the point estimates across imputations

$$\bar{B}(\theta_i) = \frac{1}{m-1} \sum_{i=1}^m (\theta_i - \bar{\theta}_i)^2. \quad (124)$$

Then the variance associated with θ_i is

$$\text{Var}(\theta_i) = \bar{U}(\theta_i) + \left(1 + \frac{1}{m}\right) \bar{B}(\theta_i). \quad (125)$$

The multiple imputation standard error of the parameter estimate θ_i is then the square root of (125). Examination of (125) shows that the multiple imputation standard error is a weighted sum of the within- and between-data set standard errors. As m increases to infinity the variance of the parameter estimate becomes the average of the parameter estimate variances.

Rubin (1987) also showed that the relative increase in variability (RIV) due to missing data is a simple function

$$\text{RIV} = \frac{\left(1 + \frac{1}{m}\right) \bar{B}(\theta_i)}{\bar{U}(\theta_i)}, \quad (126)$$

and that the overall fraction of “missing data” can be calculated as

$$\zeta = \frac{\left[\frac{\text{RIV} + 2}{(m-1) \left(1 + \frac{1}{\text{RIV}}\right)^2} \right]}{\text{RIV} + 1}. \quad (127)$$

Given an estimate of RIV the relative efficiency of a parameter estimate based on m imputations to the estimate based on an infinite number of imputations can be calculated by $(1 + \zeta/m)^{-1}$. For example, with five imputations and 40% missing data the relative efficiency is 93%. With ten imputations the relative efficiency is only 96%. Thus, the difference between five and ten imputations is not that large and so the increase in relative efficiency with the larger number of imputation sets may not be worth the computational price. Typically, the gain in efficiency is not very large when more imputation data sets are used and it is for this reason that when multiple imputation is used and

reported in the literature the number of imputed data sets is usually five or less.

There are two additional twists related to multiple imputation using MCMC. Obviously multiple imputation creates multiple datasets. For a fixed amount of computing time, one can either increase the number of iterations in the Markov chain generating a fixed number of imputed data sets or one can increase the number of imputed data sets to be analyzed using a smaller number of iterations in the Markov chain. Allison (2002) suggests that more imputation data sets be generated instead of spending more time on increasing the number of iterations in the Markov chain. The second twist is that several different data sets need to be generated. To do this, independent Markov chains are generated, one for each data set, using perhaps different starting values; this is called the parallel approach. Care must be taken with this approach that convergence has been achieved with each individual Markov chain. Alternatively, one very long Markov chain can be generated and then the data sets generated every k iterations are chosen. For example, a Markov chain of 3,000 iterations could be generated with the first 500 iterations used for burn-in and then every 500th data set thereafter used for the imputed data sets. With this method the question of independence must be raised – are the imputed data sets truly independent if they are run from the same Markov chain? As k decreases the issue of correlated data sets becomes more and more important, but when k is very large, the correlation is negligible. For example, the issue of correlation would be valid if the imputed data every ten iterations were used, but becomes a nonissue when k is in the hundreds. In general, either method is acceptable, however.

To illustrate these concepts a modification of the simulation suggested by Allison (2000) will be analyzed. In this simulation 10,000 observations of three variables were simulated: Y , x_1 , and x_2 . Such a large sample size was used to insure that sampling variability was small. x_1 and x_2 were bivariate normally distributed random variables with mean 0 and variance 1 having a correlation of 0.5. Y was then generated using

$$Y = 1 + x_1 + x_2 + Z, \quad (128)$$

where Z was normally distributed random error having mean 0 and variance 1. Four missing data mechanisms were then examined:

1. Missing completely at random: x_2 was missing with probability 0.5 independent of Y or x_1
2. Missing at random, dependent on x_1 : x_2 was missing if $x_1 < 0$
3. Missing at random, dependent on Y : x_2 was missing if $Y < 0$
4. Nonignorable: x_2 was missing if $x_2 < 0$

The data were then fit using linear regression of (x_1, x_2) against Y . The results are presented in Table 13. Listwise deletion resulted in parameter estimates that were unbiased, except when the data were MAR and dependent on the value of Y in which case all three parameter estimates were severely biased. Surprisingly, even when the missing data

mechanism was nonignorable the parameter estimates were unbiased and precise. The standard errors for all models with missing data were about 25–200% larger than the data set with no missing data because of the smaller sample sizes. MI tended to decrease the estimates of the standard errors compared to their original values. When the data were MAR or MCAR, the parameter estimates remained unbiased using MI with MCMC, even when the data were MAR and dependent on Y . The bias that was observed was now removed. But when the missing data were non-ignorable, the parameter estimates obtained by MI became biased because MI assumes the data MAR.

So how is MI incorporated in the context of exploratory data analysis since obviously one would not wish to analyze m different data sets. A simple method would be to impute $m+1$ data sets, perform the exploratory analysis on one of the imputed data sets, and obtain the final model of interest. Then using the remaining m -data sets compute the imputed parameter estimates and standard errors of the final model. It should be kept in mind, however, that with the imputed data set being used to develop the model, the standard errors will be smaller than they actually are since this data set fails to take into account the sampling variability in the missing values. Hence, a more conservative test of statistical significance for either model entry or removal should be considered during model development.

A totally different situation arises when covariates are missing because of the value of the observation, not because the covariate was not measured. In such a case the value is censored, which means that the value is below or above some critical threshold for measurement. On the other hand, a covariate may be censored from above where the covariate reported as greater than upper limit of quantification (ULOQ) of the method used to measure it. In such a case the covariate is reported as $>ULOQ$, but its true value may lie theoretically between ULOQ and infinity. The issue of censored covariates has not received as much attention as the issue of censored dependent variables. Typical solutions include any of the substitution or imputation methods described for imputed missing covariates that are not censored.

In summary, case-deletion is easy but can lead to biased parameter estimates and is not generally recommended by regulatory authorities. In contrast, multiple imputation, although computationally more difficult, is generally recognized as the preferred method to handling missing data and like any statistical analysis requires certain assumptions be met for validity. The analyst is especially warned in the case of censored data and the effects of case-deletion or multiple imputation on parameter estimation. This section has presented a high-level overview of MI and handling missing data that is far from complete. The reader is strongly encouraged to read more specialized texts on the topic prior to actually implementing their use in practice. Before closing this section, the best advice for missing data is to have none – do everything possible to obtain the data in the first place!

Table 13
Parameter estimates and standard errors from simulated multiple imputation data set

Missing Data Mechanism	Parameter	Number of Observations Without Missing Data	Listwise Deletion Mean (Standard deviation)	MCMC Mean (Standard deviation)
No missing data	Intercept	10,000	1.005 (0.0101)	—
	x_1		0.987 (0.0116)	—
	x_2		0.998 (0.0117)	—
MCAR	Intercept	4,982	1.000 (0.0141)	0.993 (0.0101)
	x_1		0.986 (0.0162)	0.978 (0.0118)
	x_2		0.996 (0.0164)	1.012 (0.0133)
MAR on x_1	Intercept	5,023	0.998 (0.0234)	1.001 (0.0284)
	x_1		0.996 (0.0250)	0.994 (0.0126)
	x_2		0.992 (0.0166)	0.993 (0.0142)
MAR on Y	Intercept	6,942	1.419 (0.0123)	1.000 (0.0122)
	x_1		0.779 (0.0133)	0.988 (0.0162)
	x_2		0.789 (0.0136)	0.996 (0.0150)
Nonignorable on x_2	Intercept	5,016	1.017 (0.0234)	1.66 (0.0187)
	x_1		0.973 (0.0166)	1.13 (0.0148)
	x_2		1.016 (0.0254)	1.22 (0.0300)

Note: True values are 1.000 for intercept, 1.000 for x_1 , and 1.000 for x_2 . Results based on 10,000 simulated observations. MI was based on five imputed data sets having a burn-in of 500 iterations

Software

Every statistical package, and even spreadsheet programs like Microsoft Excel[®], has the capability to perform linear regression. SAS (SAS Institute, Cary, NC, <http://www.sas.com>) has the REG procedure, while S-Plus (Insightful Corp., Seattle, WA, <http://www.insightful.com>) has available its lm function. Statistical packages are far more powerful than the spreadsheet packages, but spreadsheet packages are more ubiquitous. The choice of either S-Plus or SAS is a difficult one. S-Plus has better graphics, but SAS is an industry standard – a workhorse of proven capability. S-Plus is largely seen in the pharmacokinetic community as a tool for exploratory data analysis, while SAS is viewed as the de facto standard for statistical analysis in pharmaceutical development. All the examples in this book were analyzed using SAS (version 8) for Windows.

Using statistical reference data sets (certified to 16 significant digits in the model parameters) available from the National Institute of Standards and Technology (NIST) Information Technology Department (<http://www.itl.nist.gov/div898/strd>), McCullough (1999) compared the accuracy of SAS (version 6.12) and S-Plus (version 4.5), both of which are older versions than are currently available, in fitting a variety of linear models with varying levels of difficulty. Data sets of low difficulty should be easily fit by most algorithms, whereas data sets of high

difficulty, which are highly collinear, may produce quite biased parameter estimates because different software may use different matrix inversion algorithms. McCullough found that SAS and S-Plus both demonstrate reliability in their linear regression results. However, for analysis of variance problems, which should use the same linear regression algorithms, the results were quite variable. Neither SAS or S-Plus passed average difficulty problems. When the linear regression data sets were analyzed using Microsoft's Excel 97 (Microsoft Corp., Seattle, WA, <http://www.microsoft.com>) built-in data analysis tools, most performed reasonably well, but failed on a problem that was ill-conditioned, leading the authors to conclude that Excel 97 is "inadequate" for linear regression problems (McCullough and Wilson 1999). Furthermore, when Excel 97 analyzed the analysis of variance data sets, the software delivered acceptable performance on only low-level difficulty problems and was deemed inadequate.

Unfortunately, there have been no studies of this type with more recent software versions or with other software, such as R or Matlab. It might be expected that more recent versions of software that previously performed adequately, such as SAS and S-plus, still perform adequately. It is not clear whether R, Matlab, or Excel perform adequately with reasonable accuracy. It seems likely that R and Matlab do in fact perform adequately, but Excel should still be considered questionable until shown otherwise.

Summary

Linear regression is one of the most important tools in a modelers toolbox, yet surprisingly its foundations and assumptions are often glossed over at the graduate level. Few books published on pharmacokinetics cover the principles of linear regression modeling. Most books start at nonlinear modeling and proceed from there. But, a thorough understanding of linear modeling is needed before one can understand nonlinear models. In this chapter, the basics of linear regression have been presented, although not every topic in linear regression has been presented – the topic is too vast to do that in one chapter of a book. What has been presented are the essentials relevant to pharmacokinetic and pharmacodynamic modeling. Later chapters will expand on these concepts and present new ones with an eye towards developing a unified exposition of pharmacostatistical modeling.

Recommended Reading

- Allison PD. *Missing Data*. Sage Publications, Inc., Thousand Oaks, CA, 2002.
- Fox J (2000) *Nonparametric Simple Regression*. Sage Publications, Newbury Park, CA.
- International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *Statistical Principles for Clinical Trials* (E9). 1998.
- Myers RH. *Classical and Modern Regression with Applications*. Duxbury Press, Boston, 1986.
- Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. *Applied Linear Statistical Models*. Irwin, Chicago, 1996.

References

- Allison PD. Multiple imputation for missing data: a cautionary tale. *Sociological Methods and Research* 2000; 28: 301-309.
- Allison PD. *Missing Data*. Sage Publications, Inc., Thousand Oaks, CA, 2002.
- Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician* 1992; 46: 175-185.
- Bazunga M, Tran HT, Kertland H, Chow MSS, and Massarella J. The effects of renal impairment on the pharmacokinetics of zalcitabine. *Journal of Clinical Pharmacology* 1998; 38: 28-33.
- Belsley DA, Kuh E, and Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley and Sons, Inc., New York, 1980.
- Bonate PL and Howard D. Prospective allometric scaling: does the emperor have clothes? *Journal of Clinical Pharmacology* 2000; 40: 335-340.
- Breen R. *Regression Models: Censored, Sample Selected, and Truncated Data*. Sage Publications, Thousand Oaks, CA, 1996.

- Buonaccorsi JP. Prediction in the presence of measurement error: general discussion and an example predicting defoliation. *Biometrics* 1995; 51: 1562-1569.
- Calvert AH, Newell DR, Gumbrell LA, O'Reilly S, Burnell M, Boxall FE, Siddik ZH, Judson IR, Gore ME, and Wiltshaw E. Carboplatin dosing: prospective evaluation of a simple formula based on renal function. *Journal of Clinical Oncology* 1989; 7: 1748-1756.
- Carroll RJ, Juchenhodd H, Lombard F, and Stefanki LA. Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association* 1996; 91: 242-250.
- Carroll RJ, Ruppert D, and Stefanski LA. *Measurement Error in Nonlinear Models*. Chapman & Hall, New York, 1995.
- Carter WH, Jones DE, and Carchman RA. Application of response surface methods for evaluating the interactions of soman, atropine, and pralidoxime chloride. *Fundamental and Applied Toxicology* 1985; 5: S232-S241.
- Chapra SC and Canale RP. *Numerical Methods for Engineers: With Programming and Software Applications*. McGraw-Hill, New York, 1998.
- Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 1979; 74: 829-836.
- Cook JR and Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* 1994; 89: 1314-1328.
- Diggle PJ and Kenward MG. Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics* 1994; 43: 49-93.
- European Agency for the Evaluation of Medicinal Products and Committee for Proprietary Medicinal Products. Points to Consider on Missing Data. 2001.
- Fuller WA. *Measurement Error Models*. John Wiley and Sons, Inc., New York, 1987.
- Gehan EA and George SL. Estimation of human body surface area from height and weight. *Cancer Chemotherapy Reports* 1970; 54: 225-235.
- Gelman A, Carlin JB, Stern HS, and Rubin DB. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.
- Gray JB. A simple graphic for assessing influence in regression. *Journal of Statistical Computing and Simulation* 1986; 24: 121-134.
- Gray JB and Woodall WH. The maximum size of standardized and internally studentized residuals in regression analysis. *American Statistician* 1994; 48: 111-113.
- Greco WR, Bravo G, and Parsons JC. The search for synergy: A critical review from a response surface perspective. *Pharmacological Reviews* 1995; 47: 331-385.

- Hastie TJ and Tibshirani RJ. *Generalized Additive Models*. Chapman and Hall, New York, 1990.
- Hodges SD and Moore PG. Data uncertainties and least squares regression. *Applied Statistics* 1972; 21: 185-195.
- Holford NHG. A size standard for pharmacokinetics. *Clinical Pharmacokinetics* 1996; 30: 329-332.
- Hollander M and Wolfe DA. *Nonparametric Statistical Methods*. John Wiley & Sons, New York, 1999.
- International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. General Considerations for Clinical Trials (E8). 1997.
- International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Statistical Principles for Clinical Trials (E9). 1998.
- Iwatsubo T, Hirota N, Ooie T, Suzuki H, Shimada N, Chiba K, Ishizaki T, Green CE, Tyson CA, and Sugiyama Y. Prediction of in vivo drug metabolism in the human liver from in vitro metabolism data. *Pharmacology Therapeutics* 1997; 73: 147-171.
- Iwatsubo T, Hirota N, Ooie T, Suzuki H, and Sugiyama Y. Prediction of in vivo drug disposition from in vitro data based on physiological pharmacokinetics. *Biopharmaceutics and Drug Disposition* 1996; 17: 273-310.
- Jackson JE. *A User's Guide to Principal Components*. John Wiley and Sons, Inc., New York, 1991.
- Kaul S and Ritschel WA. Quantitative structure-pharmacokinetic relationship of a series of sulfonamides in the rat. *European Journal of Drug Metabolism and Pharmacokinetics* 1990; 15: 211-217.
- Lefevre F, Duval M, Gauron S, Brookman LJ, Rolan PE, Morris TM, Piraino AJ, Morgan JM, Palmisano M, and Close P. Effect of renal impairment on the pharmacokinetics and pharmacodynamics of desirudin. *Clinical Pharmacology and Therapeutics* 1997; 62: 50-59.
- Little RJ. Regression with missing X's: A review. *Journal of the American Statistical Association* 1992; 87: 1227-1237.
- Little RJ. Modeling the drop-out mechanism in repeated measures studies. *Journal of the American Statistical Association* 1995; 90: 1112-1121.
- Little RJ and Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 2002.
- Malloves CL. Some comments on Cp. *Technometrics* 1973; 15: 661-675.
- McCullough BD. Assessing the reliability of statistical software: Part II. *American Statistician* 1999; 53: 149-159.
- McCullough BD and Wilson B. On the accuracy of statistical procedures in Excel 97. *Computational Statistics & Data Analysis* 1999; 31: 27-37.
- Myers RH. *Classical and Modern Regression with Applications*. Duxbury Press, Boston, 1986.
- Neter J, Kutner MH, Nachtsheim CJ, and Wasserman W. *Applied Linear Statistical Models*. Irwin, Chicago, 1996.
- Pool JL, Cushman WC, Saini RK, Nwachuku CE, and Battikha JP. Use of the factorial design and quadratic response models to evaluate the fosinopril and hydrochlorothiazide combination in hypertension. *American Journal of Hypertension* 1997; 10: 117-123.
- Port RE, Daniel B, Ding RW, and Hermann R. Relative importance of dose, body surface area, sex, and age in 5-fluorouracil clearance. *Oncology* 1991; 48: 277-281.
- Rockhold FW and Goldberg MR. An approach to the assessment of therapeutic drug interactions with fixed combination drug products. *Journal of Biopharmaceutical Statistics* 1996; 6: 231-240.
- Rubin DR. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.
- Ryan TP. *Modern Regression Methods*. John Wiley & Sons, Inc., New York, 1997.
- Schafer J. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London, 1997.
- Simon SD and Lesage JP. The impact of collinearity involving the intercept term on the numerical accuracy of regression. *Computer Science in Economics and Management* 1988; 1: 137-152.
- Stefanski LA and Cook JR. Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association* 1995; 90: 1247-1256.
- Stewart WH. Application of response surface methodology and factorial designs for drug combination development. *Journal of Biopharmaceutical Statistics* 1996; 6: 219-231.

Pharmacokinetic-Pharmacodynamic Modeling and
Simulation

Bonate, P.L.

2011, XIX, 618 p. 301 illus., 4 illus. in color., Hardcover

ISBN: 978-1-4419-9484-4