

# 1 Introduction and Motivation

## 1.1 Scope and objectives

*OBJECTIVE:* The goal of this course is to provide a comprehensive treatment of modern regression models and associated inferential methods for univariate and multivariate response.

- By *univariate* response, we mean the case where the response is a single, scalar value.
- By *multivariate* response, we mean the case where several scalar responses may be thought of together as a group.

We will clarify these designations through a series of examples momentarily.

Our emphasis will be on methods that have become commonplace tools for the practicing statistician.

We will discuss the models and methods from two perspectives:

- We will derive relevant theoretical results and study their implications for practical use, and
- We will discuss practical implementation of the methods using available software, and demonstrate them on data from a variety of applications.

*THEME:* As we will see, statistical models and associated inferential methods for univariate and multivariate response share common features. Thus, we will first study the univariate case in some detail. Many of the concepts and results will then carry over readily to the case of multivariate response.

We begin by reviewing the form of the “classical” linear regression model, which provides a convenient point of departure for discussing the need for more sophisticated and broadly applicable methods.

## 1.2 “Classical” linear regression model

In the usual regression framework, we consider a (scalar) response  $Y$ , say, and an associated vector of covariates  $\mathbf{x}$ . Formally, let

$Y$  = response, dependent variable (scalar)

$\mathbf{x}$  =  $(p \times 1)$  covariate, predictor, independent variable  $(p \times 1)$  ( $\mathbf{x}$  may include the constant “1”)

Here, we have noted some of the common terminology used in the regression literature.

*ASSUME:* The values of  $\boldsymbol{x}$  may be set by an experimenter or may be observed. In either case, the values are *known without error*. For example:

- $x$  is a planned dose of a drug given to a rat by injection in a pharmacological experiment. The actual dose received is exactly equal to the planned value – no errors were committed in preparing the injection solution.
- $\boldsymbol{x}$  contains values of age, weight, height, and race of subjects sampled from a population of interest in an epidemiological study. None of the values of age or race of the subjects are recorded incorrectly. Errors of measurement committed in ascertaining weight and height are negligible.

In the first example,  $x$  is often thought of as “fixed,” as it is set by the experimenter. In the second, as the subjects are sampled from a population,  $\boldsymbol{x}$  may be thought of as a *random vector* taking its values according to some (multivariate) probability distribution. In this case, values sampled from this population may be ascertained perfectly, without error. Alternatively, in the first example if the dose scale is continuous and there are  $n$  different doses in the study, one could also view the doses as being drawn from a population of doses (along the continuum of possible doses).

*NOTATION:* Although  $\boldsymbol{x}$  may be a *random vector*, we use the lower case symbol instead of  $\boldsymbol{X}$ .

*USUAL PERSPECTIVE:* The usual way of thinking is that  $Y$  is a *random variable*. Given  $\boldsymbol{x}$ , the values of  $Y$  that might be observed *vary*, so that the different values of  $Y$  may be seen with the same value of  $\boldsymbol{x}$ . This may be because of one or more of the following:

- Error in measurement of  $Y$ . Here, ideally, there is a unique value of  $Y$  corresponding to each  $\boldsymbol{x}$ , so that, in truth, a deterministic relationship between  $Y$  and  $\boldsymbol{x}$  does exist. However,  $Y$  cannot be measured exactly. For example, in an (idealized) physics experiment, if  $Y$  is force and  $x$  is acceleration, then for an object with a particular mass  $m$ ,  $Y = mx$ . However, with available devices, it may only be possible to measure  $Y$  for a given  $x$  with some uncertainty, so that different values of  $Y$  may be recorded for the same  $x$ , reflecting the error in the device.
- “Sampling variation” or variation among individuals. This may happen in several ways. In the epidemiological study example above, suppose  $Y$  is total cholesterol. If we consider all subjects in the population having a particular age, weight, height, and race combination, total cholesterol will not be identical for all of them, reflecting natural biological variation across humans. Of course, total cholesterol is likely to be measured with error as well.

As another example, consider the pharmacological study above.  $x$  may be dose of drug given to a rat and  $Y$  concentration 5 minutes after injection, where  $Y$  is determined by drawing a blood sample from the rat. If the drug is not “well mixed” within the rat, then different values of  $Y$  might be seen from different samples. Of course, again,  $Y$  may be measured with error.

- Features of the biological or physical process under study. For example, in the epidemiological study, for a given subject, total cholesterol may fluctuate over time about some “typical” value. Thus, the value of  $Y$  recorded for a subject will reflect this fluctuation.

In many applications, it would indeed be likely for these features to occur simultaneously, although the effects of some might be negligible relative to those of others.

*FORMALLY:* We may conceptualize for each value of  $\mathbf{x}$  a *probability distribution* characterizing possible values of the response  $Y$  that might be observed.

- This is easy to think of in the case where the values of  $\mathbf{x}$  are fixed constants.
- When the values of  $\mathbf{x}$  vary, for example, in a population, it is natural to think of a *joint* distribution of values  $(Y, \mathbf{x})$  that might be observed. Under this perspective, the probability distribution characterizing possible values of  $Y$  for each value of  $\mathbf{x}$  would be the conditional distribution.

We will not concern ourselves with the distinction between fixed and random  $\mathbf{x}$ . The important point is that the basis for regression modeling is to consider a statistical model for the values of  $Y$  conditional on  $\mathbf{x}$  values. We will adopt this perspective throughout.

*DATA:* We observe pairs  $(Y_j, \mathbf{x}_j)$ ,  $j = 1, \dots, n$

- We may think of these pairs as draws from the joint probability distribution of  $(Y, \mathbf{x})$  in the event  $\mathbf{x}$  is observed rather than set by the experimenter.
- Alternatively, we may think of  $Y_j$  as being a draw from the distribution of  $Y$  corresponding to the fixed value  $\mathbf{x}_j$ .

We summarize the data as

$$\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix},$$

where  $\mathbf{Y}$  is a  $(n \times 1)$  vector and  $\mathbf{X}$  is a  $(n \times p)$  matrix.

*NOTE:* Our notation here blurs the distinction between random variables and observed values, as is often conventional (although not very precise).

*USUAL LINEAR REGRESSION MODEL:* This model is a statistical model for the pairs  $(Y_j, \mathbf{x}_j)$ , usually written in terms of *deviations*  $e_j = Y_j - \mathbf{x}_j^T \boldsymbol{\beta}$  as

$$Y_j = \mathbf{x}_j^T \boldsymbol{\beta} + e_j, \quad j = 1, \dots, n, \quad \text{or} \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} = (e_1, e_2, \dots, e_n)^T$$

in matrix notation. Here, the  $Y_j$  are assumed to be related to the  $\mathbf{x}_j$  approximately through a function *linear* in a parameter vector  $\boldsymbol{\beta}$  ( $p \times 1$ ). For example

- $\mathbf{x}_j^T \boldsymbol{\beta} = \beta_0 + \beta_1 c_j$ , a straight line, where  $\mathbf{x}_j = (1, c_j)^T$ ,  $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ ,  $p = 2$
- $\mathbf{x}_j^T \boldsymbol{\beta} = \beta_0 + \beta_1 c_j + \beta_2 c_j^2$ , a quadratic function in  $c_j$  with  $\mathbf{x}_j = (1, c_j, c_j^2)^T$ ,  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ ,  $p = 3$
- $\mathbf{x}_j^T \boldsymbol{\beta} = \beta_0 + \beta_1 c_{1j} + \beta_2 c_{2j} + \beta_3 c_{1j} c_{2j}$ , a simple response surface model with  $\mathbf{x}_j = (1, c_{1j}, c_{2j}, c_{1j} c_{2j})^T$ ,  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ ,  $p = 4$ .

The models may be a characterization of what is believed to be the truth or an approximation to a more complicated model (e.g., a *nonlinear* model).

The deviations represent the fact that the relationship between observed values for  $Y$  and  $\mathbf{x}$  does not exactly follow the smooth relationship dictated by  $\mathbf{x}_j^T \boldsymbol{\beta}$ . Rather, observed values deviate from this relationship, presumably due to one or more of the reasons given above (e.g., measurement error, sampling variation, etc.). As the  $Y_j$  are viewed as random variables, the  $e_j$  are also random variables.

*ASIDE:* The  $e_j$  are often referred to as “errors” in the regression literature. We prefer the term *deviations*, as use of the former term can be misleading, as it might suggest that the fact that observed values do not follow a smooth relationship with  $\mathbf{x}$  is due entirely to the effects of measurement error. In reality, the fact that the  $Y_j$  deviate from the relationship for one or more of the reasons given.

*“CLASSICAL” REGRESSION ASSUMPTIONS:* In a first course on regression analysis, inferential methods for this model are developed under a set of assumptions. For lack of a better term, we will call these the “classical” assumptions. A good part of this course will be devoted to relaxation of these assumptions and study of the consequences when they are incorrect.

Usually, the classical assumptions are stated assuming the  $\mathbf{x}_j$  are fixed constants; that is, the conditioning on  $\mathbf{x}_j$  is not made explicit. Here, we will be a little more careful.

- (0) The expected value of  $e_j$  is equal to zero. Such an assumption implies no systematic tendency, or *bias*, in the way the  $Y_j$  deviate from  $\mathbf{x}_j^T \boldsymbol{\beta}$ . Note that we could interpret this as  $E(e_j | \mathbf{x}_j) = 0$ , so that the expected value is conditional on the value of  $\mathbf{x}_j$ , or as  $E(e_j) = 0$ , so that this is an unconditional statement. Technically, these are two different statements; the first implies the second, of course, but not vice versa. Under the usual assumption that the  $\mathbf{x}_j$ 's are fixed, this distinction is usually not discussed. Given (2) below, the assumption  $E(e_j) = 0$  is made directly.
- (1) The model  $\mathbf{x}_j^T \boldsymbol{\beta}$  is *correct*. Along with (0), treating the  $\mathbf{x}_j$  as fixed constants, this implies that  $E(Y_j | \mathbf{x}_j) = \mathbf{x}_j^T \boldsymbol{\beta}$ .
- (2) The  $e_j$  are identically distributed for all  $j$ , independently of  $\mathbf{x}_j$ , and  $\text{var}(e_j) = \sigma^2$  ( $= \text{var}(e_j | \mathbf{x}_j)$  by independence).
- (3) The  $e_j$  are independent.
- (4) The  $e_j$  are normally distributed.

The assumption that  $e_j$  and  $\mathbf{x}_j$  are independent is usually not stated explicitly. Note that (4) implies that the response may be viewed, at least approximately, as continuous. Often, (3) and (4) are combined into a single statement.

Under *all* of these assumptions,  $e_j$  are independent, normal random variables. Usually, what is really meant is that the pairs  $(Y_j, \mathbf{x}_j)$  are independent. In any event, the assumptions taken all together imply

$$\begin{aligned} E(Y_j | \mathbf{x}_j) &= \mathbf{x}_j^T \boldsymbol{\beta}, & \text{var}(Y_j | \mathbf{x}_j) &= \sigma^2 & (1.1) \\ \Rightarrow & & \mathbf{Y} &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n). \end{aligned}$$

*REGRESSION PARAMETER  $\boldsymbol{\beta}$* : If  $\boldsymbol{\beta}$  were known, then the model (1.1) would provide a complete characterization of the response “on average.” Usually,  $\boldsymbol{\beta}$  is unknown and must be estimated.

The approach that is advocated under the assumptions above is that of *ordinary least squares* (OLS). The OLS estimator  $\hat{\boldsymbol{\beta}}_{OLS}$  is defined as

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \left( \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \sum_{j=1}^n \mathbf{x}_j Y_j.$$

- Under all assumptions (0) – (4),  $\hat{\boldsymbol{\beta}}_{OLS}$  is in fact the maximum likelihood estimator for  $\boldsymbol{\beta}$ ; i.e.,  $\hat{\boldsymbol{\beta}}_{OLS}$  maximizes the loglikelihood

$$\log L = -(n/2) \log 2\pi - (n/2) \log \sigma^2 - (1/2) \sum_{j=1}^n (Y_j - \mathbf{x}_j^T \boldsymbol{\beta})^2 / \sigma^2.$$

Thus, if the assumptions are all correct,  $\hat{\beta}_{OLS}$  is a natural estimator to consider. Moreover, for large samples,  $\hat{\beta}_{OLS}$  enjoys all the usual “optimality” properties associated with maximum likelihood estimation.

- Note that  $\hat{\beta}_{OLS}$  satisfies

$$\text{minimize } \sum_{j=1}^n (Y_j - \mathbf{x}_j^T \boldsymbol{\beta})^2 \quad \Rightarrow \quad \text{solve } \sum_{j=1}^n (Y_j - \mathbf{x}_j^T \boldsymbol{\beta}) \mathbf{x}_j = \mathbf{0}. \quad (1.2)$$

The expression to be minimized in (1.2) may be interpreted as a sensible “distance” criterion. Even if we do not adopt (4), normality,  $\hat{\beta}_{OLS}$  still has the interpretation as being the minimizer of “distance between data and model.” In fact, it is often motivated this way in elementary regression texts.

- Here, all observations receive *equal weight* in the distance measure, reflecting assumption (2), which implies that all data are of equal “quality” (precision). Note further that all observations are treated “separately” (in a sum), reflecting the independence assumption.
- In addition,  $\hat{\beta}_{OLS}$  is a *linear* function of the  $Y_j$  and has expectation  $\boldsymbol{\beta}$  under (1.1), whether assumption (4) holds or not, so is unbiased. In fact, it may be shown that, under (0)–(3),  $\hat{\beta}_{OLS}$  has the smallest (sampling) variance among all linear functions of the  $Y_j$  used to estimate  $\boldsymbol{\beta}$  – it is the Best Linear Unbiased Estimator (BLUE) of  $\boldsymbol{\beta}$ .
- More generally,  $\hat{\beta}_{OLS}$  solves the set of  $p$  *estimating equations* in (1.2) that are *linear* in the  $Y_j$ . If we adopt the sum of squared deviations above as the criterion to minimize, then this is regardless of whether or not (4) holds.

Thus, an implication to keep in mind is that, even without the assumption of normality,  $\hat{\beta}_{OLS}$  may be viewed as a “sensible” estimator for  $\boldsymbol{\beta}$  with a nice interpretation and nice properties.

### 1.3 Violation of the “classical” assumptions

All of the “classical” assumptions are violated routinely in practice. Moreover, several of them may be violated simultaneously. We illustrate by considering a series of examples.

*EXAMPLE 1.1 Pharmacokinetics of indomethacin.* A common objective is to investigate the *pharmacokinetics* of a drug. A known dose of the drug is given to a subject (human or animal), and then at several subsequent time points, blood samples are drawn and the concentration of drug in blood or plasma (the response) is determined. Interest focuses on characterizing the concentration-time profile.

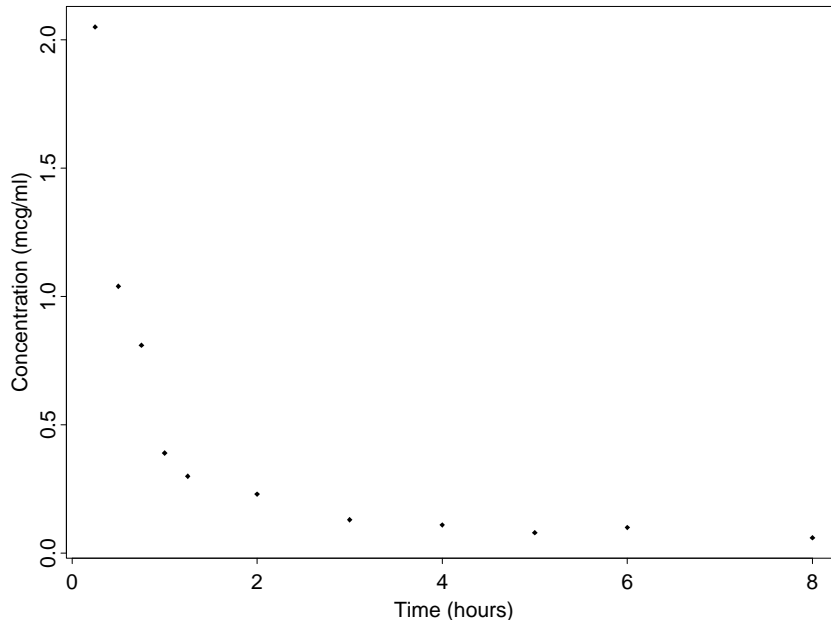
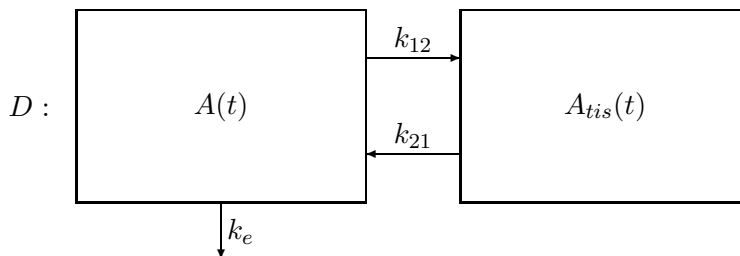
Figure 1.1: *Concentration-time data for a subject receiving intravenous indomethacin at time 0.*

Figure 1.1 shows plasma concentrations plotted against time for a human subject receiving an intravenous dose of the drug indomethacin at time 0 (the time scale is defined so that 0 represents dose time). Here,  $x_j$  is the  $j$ th time point (in hours) and  $Y_j$  is the measured concentration of indomethacin.

- From the plot, one might be tempted to approximate the relationship by a polynomial in time (which is, of course, a linear model). However, this is not a good approximation in general.
- In fact, theoretical considerations suggest a more scientifically relevant way to represent the relationship. Standard practice in pharmacology is to represent the body as a system of “compartments” corresponding to parts of the system such as “blood” and “deeper tissues.” For indomethacin, pharmacologists believe a two compartment model as follows provides a reasonable representation:



Here, dose  $D$  is given intravenously into the left blood compartment at time  $t = 0$ , and the amount of drug present is denoted by  $A(t)$ . Transfer is assumed to take place between the blood compartment and the “deeper tissues” compartment according to fractional rates of transfer  $k_{12}$  and  $k_{21}$ , and the amount in the deeper tissue compartment at  $t$  is denoted by  $A_{tis}(t)$ . The rate  $k_e$  corresponds to removal of drug from the blood by elimination from the system, e.g., via the kidneys.

From this model, one can write down a set of differential equations describing amounts of drug in each compartment:

$$\begin{aligned}\frac{dA(t)}{dt} &= k_{21}A_{tis}(t) - k_{12}A(t) - k_eA(t) \\ \frac{dA_{tis}(t)}{dt} &= k_{12}A(t) - k_{21}A_{tis}(t)\end{aligned}\tag{1.3}$$

subject to the initial conditions that  $A_{tis}(0) = 0$  (no drug in the deeper tissues when the dose is given) and  $A(0) = D$  (the dose fills the blood compartment instantaneously when it is given).

- Solution of the differential equations (1.3) yields an expression for  $A(t)$ . This can be divided by a parameter corresponding to the “volume” of the blood compartment to give concentration (amount per unit volume) at time  $t$ ,  $C(t)$ , say. The form of the expression for  $C(t)$  is

$$\beta_1 \exp(-\beta_2 t) + \beta_3 \exp(-\beta_4 t),$$

where the elements of  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T$  are functions of  $k_{12}$ ,  $k_{21}$ ,  $k_e$ , and the volume; that is, functions of *physically meaningful* quantities (with respect to the compartment model). In fact, pharmacokineticists are often more interested in the values of these parameters than they are in the actual profile itself! This is because the values  $k_{12}$ ,  $k_{21}$ ,  $k_e$ , and volume tell them something about the individual’s inherent biological features in terms of processing this drug.

Other meaningful quantities, for example the drug *terminal half-life*

$$\log 2/\beta_4,$$

may also be derived from these elements.

- The above suggests that an appropriate regression model to describe blood concentration  $Y_j$  at time  $x_j$  would be

$$\beta_1 \exp(-\beta_2 x_j) + \beta_3 \exp(-\beta_4 x_j).\tag{1.4}$$

This model is a function of a “covariate”  $x_j$  (time) and parameters  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T$ , but is *nonlinear* in some of the elements of  $\boldsymbol{\beta}$ . In particular, the dependence in (1.4) on the parameters  $\beta_2$  and  $\beta_4$  appearing in the exponential terms is nonlinear.

- The model has a *theoretical* interpretation. Rather than just being an empirical representation for the profile, the form of and parameters in the model have physical meaning, and the model itself is derived from theoretical considerations.

*CLASSICAL ASSUMPTIONS VIOLATED:* (1). A linear model does not provide a scientifically relevant (or very good) representation for these data.

*ASIDE:* We put the word “covariate” in quotes above, because “time” is really not a covariate in the strict sense in this setting. If we think rigorously about this problem, we may envision an entire *process* in *continuous time*. In particular, we can think of  $Y(t)$ , the concentration we would observe at time  $t$ . The *deterministic* compartmental model provides a description of what amount of drug  $A(t)$  and hence exact concentration  $C(t)$  of drug (if we divide by “volume”) would look like at any time  $t$  if we believed this model and if there were no measurement error or other source of variation such as “fluctuations” within a person operating. If we think of  $Y(t)$  as the concentration we would observe including these effects, we can think of  $Y(t)$  as a *stochastic process* in time, i.e.,

$$Y(t) = \beta_1 \exp(-\beta_2 t) + \beta_3 \exp(-\beta_4 t) + e(t),$$

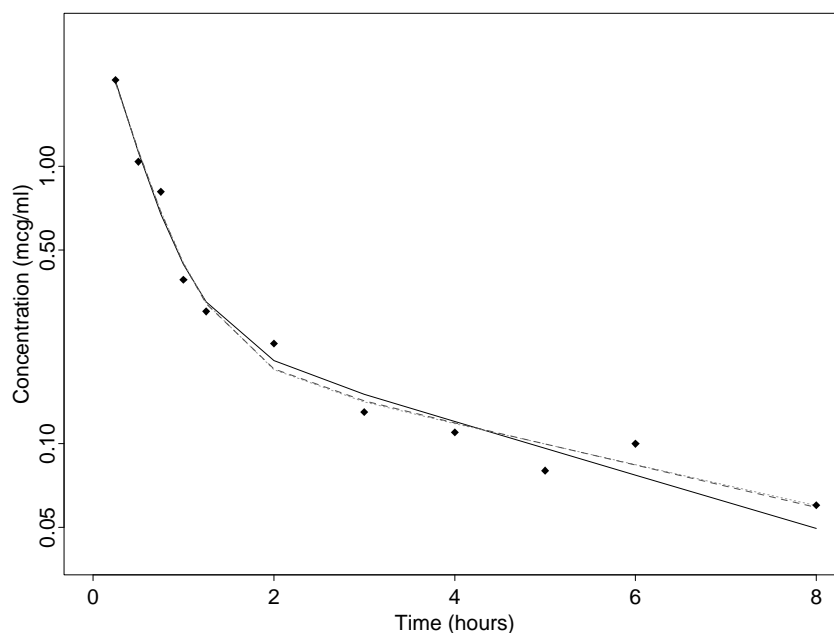
say, where  $e(t)$  is the “deviation” at time  $t$  that takes into account the effects of fluctuations and measurement error.

From this perspective, then, “time” is a fundamental part of the response, not a “covariate” in the traditional sense. When we take blood samples at specific times  $x_1, \dots, x_n$ , say, we are seeing *realizations* of the process  $Y(t)$  at these time points.

It is conventional in much of the literature on univariate regression to sweep this distinction under the rug, and to treat “time” as a “covariate” in the usual way. In the sequel, we will do the same, as for our purposes how we regard time in problems like this will not have implications for the concepts and results we will study. When we discuss multivariate response later in the course, this issue becomes somewhat more important; we defer further comment until then.

*EXAMPLE 1.2 Pharmacokinetics of indomethacin, continued.* As we will discuss, model (1.4) may be fitted to the indomethacin data using a *nonlinear* version of ordinary least squares. In fact, there are alternative fitting methods that we will also discuss. Figure 1.2 shows the data plotted on the log-concentration scale, with the model (1.4) fitted to the data using nonlinear ordinary least squares (solid line) and another method called *generalized least squares* (dashed line) that we will introduce in Chapter 2. Note that these fits are different, with the difference being most pronounced for the largest time points, where the response is smallest.

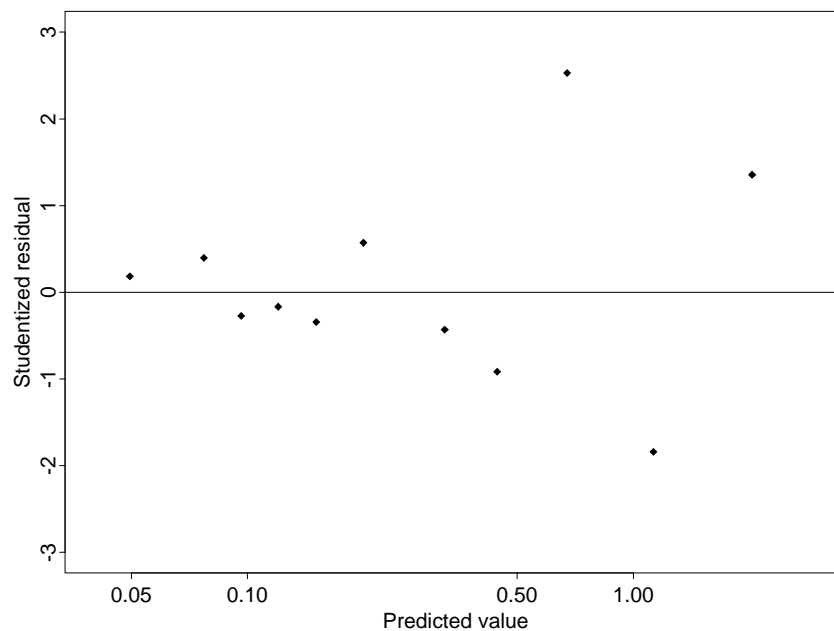
Figure 1.2: *Log Concentration-time data with fits of model (1.4) superimposed.*



From the nonlinear ordinary least squares fit, it is possible to construct a familiar residual plot. Figure 1.3 shows a plot of residuals, which have been “studentized” (so standardized in a certain way) in a manner we will discuss in Chapter 7, against predicted values, where predicted values are constructed in the usual way by evaluating the model (1.4) for each time point at the estimated value of  $\beta$ .

- The plot shows clear evidence of a “fan” shape, such that the magnitude of the (studentized) residuals is larger for larger predicted values (so for larger values of the response). Just as in linear regression, this pattern suggests that variability in the response  $Y$  increases with the level of  $Y$  – the larger  $Y$ , the larger the variation. The increase seems to take place in a “smooth” fashion, as suggested by the regularity of the fan shape.

Figure 1.3: Plot of (studentized) residuals versus predicted values for the nonlinear ordinary least squares fit of (1.4).



- In fact, this phenomenon is commonplace with pharmacokinetic data. In particular, it is widely noted that the variability in measured drug concentrations about a “smooth” concentration-time profile as dictated by a model like (1.4), increases with increasing concentration.

*CLASSICAL ASSUMPTIONS VIOLATED:* (2). The assumption of constant variance (2) for all  $j$  is obviously not appropriate.

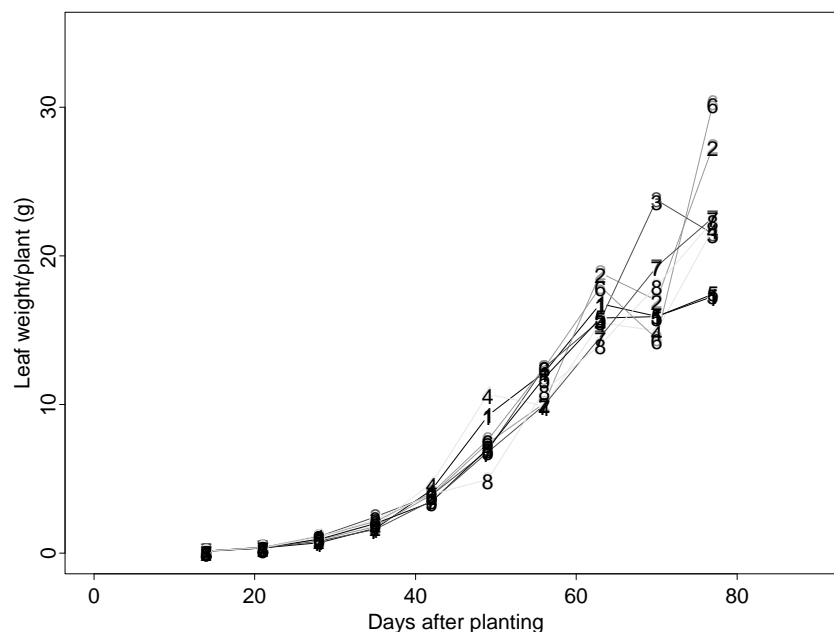
In fact, we can identify still another possible violation. Concentration measurements are taken over time on the *same* subject. As we will discuss in greater detail later in the course, there is a possibility that, if responses are ascertained sufficiently close together in time (separated by only a small time interval), they may be “more alike” than those that are farther apart in time. That is, they may tend to be “large” or small (larger or smaller than the fitted profile would suggest) together.

This tendency for *serial correlation* would clearly violate the independence assumption (3).

As it turns out, pharmacokinetic data are often reasonably assumed to be (approximately) normally distributed at any time  $x$ , so assumption (4) may be reasonable.

*EXAMPLE 1.3 Soybean growth study.* Figure 1.4 depicts data from an experiment carried out by researchers in the Department of Crop Science at North Carolina State University, reported in Davidian and Giltinan (1995, p. 7). In a series of field experiments over several years, plots in a field were planted with two genotypes of soybean, and the goal was to compare the growth patterns across the two genotypes. Each plot was sampled over the course of the growing season at approximately weekly intervals. At each sampling time, six plants were randomly selected from each plot, their leaves were removed and mixed together, and the mixture weighed. The weight divided by 6 was the average leaf weight per plant (g). The figure shows average leaf weight per plant versus time for the eight plots planted with one of the genotypes in one of the years.

Figure 1.4: *Soybean growth data for 8 plots.*



- From the figure, one might again be tempted to use a polynomial to represent the growth profile for a given plot. However, as with the pharmacokinetic example, this will not lead to a very good approximation. For one thing, intuitively, as time goes on, growth cannot continue to increase forever. Rather, it must “level off” as the plants in the plot reach maturity.

Moreover, as the plants are measured over the season where growth is taking place, it is unlikely that they will begin to “shrink,” so that the profile begins to decrease. A polynomial model such as a quadratic function of time would not be able capture the “asymptotic” behavior expected toward the end of the season; moreover, it allows the possibility of decrease.

- Several theoretical models have been postulated to describe growth processes. The most common such model is the *logistic growth model*, which says that growth rate relative to present size declines linearly with increasing size. Formally, letting  $Y$  be the growth value (average leaf weight per plant here) and  $x$  be time, this may be expressed by the deterministic relationship

$$\frac{dY}{dx} / Y = k \left( 1 - \frac{Y}{a} \right),$$

where the right hand side is a linear function of present size  $Y$  and  $k > 0$  and  $a > 0$ .

- Upon integration, this model leads to

$$Y = \frac{\beta_1}{1 + \beta_2 \exp(-\beta_3 x)},$$

where  $\beta_1 = a$ ,  $\beta_3 = k$ , and  $\beta_2$  is the value such that  $\beta_1/(1 + \beta_2)$  represents size at time  $x = 0$ . Note also that as time gets large, i.e.,  $x \rightarrow \infty$ , the function approaches  $\beta_1$ . Thus,  $\beta_1$  is a physically meaningful parameter characterizing the asymptotic behavior, and, together with  $\beta_2$ , it characterizes the physically meaningful feature of “starting growth” at  $x = 0$ . As  $\beta_3 > 0$ , this parameter describes the change of growth with time. A plot of this function over a range of  $x$  for particular choices of  $\beta = (\beta_1, \beta_2, \beta_3)^T$  reveals that it has an “S-shape.” Thus, this model appears to be a reasonable way to represent the growth profile for a given plot.

- A common phenomenon for growth data is that variability in the response about an S-shaped pattern increases with size. Furthermore, for a given plot, the measurements of growth are taken *over time*. These features are similar to those for the pharmacokinetic example.

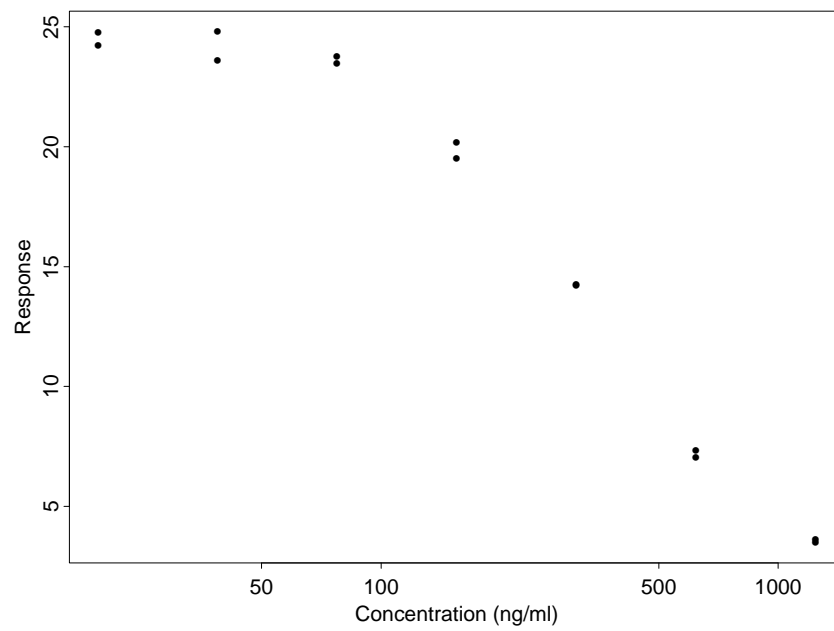
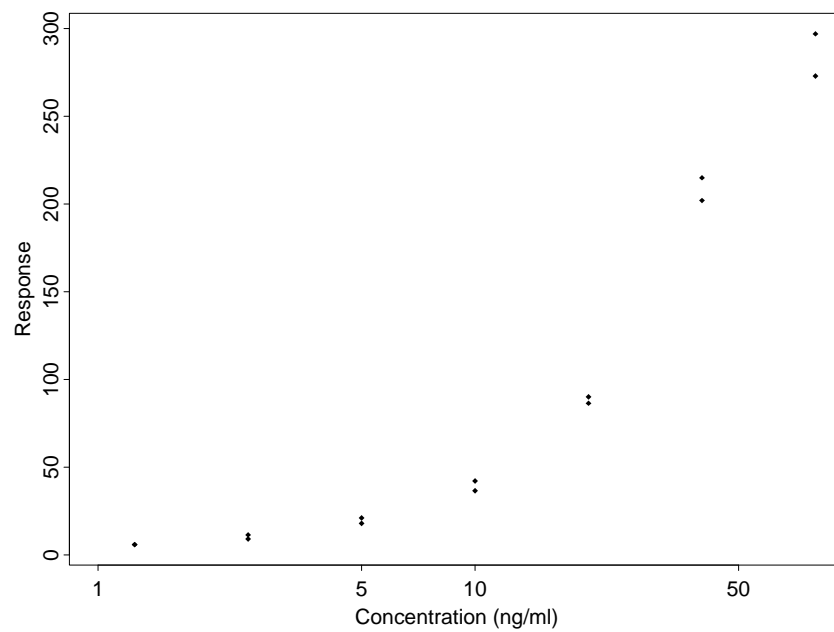
*CLASSICAL ASSUMPTIONS VIOLATED:* (1), (2), (3) are clearly violated: an appropriate model is not linear in parameters, and the assumptions of constant variance and independence are suspect. Whether or not the normality assumption (4) is reasonable is not immediately clear.

*EXAMPLE 1.4 Assay data.* An assay is a procedure used to determine the level of drug or other substance in a sample. For example, pharmacokinetic analysis requires that the concentration of drug be determined for the blood samples drawn from a particular subject at each time point.

Ordinarily, such levels cannot be determined directly. Under the assay paradigm, the level is inferred by observing a response that is related to level. An assay is conducted as follows:

- Samples with *known* levels of the substance  $x_j$  are prepared, the so-called *standards*. Typically, replicate samples with the same  $x_j$  value are prepared. For each standard, a response  $Y_j$  is measured. The resulting pairs  $(Y_j, x_j)$  may be used to establish a relationship between level (e.g., concentration of drug) and the response.
- Responses for samples of interest with *unknown* levels are also obtained. The relationship may be used to infer the unknown level. This procedure is known as *calibration*, and a popular way to do it is often referred to in regression texts as “inverse regression.”
- Different kinds of assays are performed, depending on the substance to be calibrated. *Radioimmunoassay* (RIA) is one type of assay procedure. Here, the response  $Y$  is generally a radioactive gamma count whose value decreases as level increases. Figure 1.5 shows standards data for a RIA developed at Becton Dickinson Research Center in Research Triangle Park, NC, to determine the concentration of a certain drug in porcine (pig) blood serum for an experiment involving the pharmacokinetic properties of the drug (in pigs); the data are reported in Belanger, Davidian, and Giltinan (1996).
- Alternatively, Figure 1.6 shows standards data from a type of assay known as an ELISA, or *enzyme-linked immunosorbent assay*. The response for an ELISA is generally something like an absorbance or optical density (color) reading or rate of change of this that increases with level. This assay was developed in the context of a large epidemiological study of childhood asthma. The study involved relating the levels of common allergens in samples of house dust to development of childhood asthma. To determine levels of allergen in house dust samples, an assay was required. The data in Figure 1.6 are standards data from a run of an assay developed to measure concentration of *Dermatophagoides pteronyssinus* (roach) allergen and are reported in Higgins et al. (1998).
- Both of these examples exhibit some common features of assay data: an “S-shaped” relationship between concentration and response and variability that tends to increase with the level of the response. This latter feature is easily deduced directly from a plot of the data because of the replication present at each standard concentration in each case. Thus, variance is obviously not constant in either case.
- A standard model to represent the relationship between response  $Y$  and concentration or level  $x$  is the so-called *four parameter logistic function*, one parameterization of which is given by

$$\beta_1 + \frac{\beta_2 - \beta_1}{1 + (x/\beta_3)^{\beta_4}}.$$

Figure 1.5: *Standards data for a RIA for drug in porcine serum.*Figure 1.6: *Standards data for a ELISA for roach allergen*

Here,  $\beta_1$  is the response at “ $x = \infty$ ,”  $\beta_2$  is the “background” response at  $x = 0$ ,  $\beta_3$  is the concentration giving response halfway between  $\beta_2$  and  $\beta_1$ , often called the “ED50,” and  $\beta_4$  is a “shape” parameter governing the steepness of the increasing or decreasing part of the curve. This model provides an excellent representation of most assay concentration-response relationships.

*CLASSICAL ASSUMPTIONS VIOLATED:* (1) and (2) are clearly violated, as before. Typically, the responses are determined separately for each sample, so are considered independent. In the case of RIA, the response is a count. Thus, one might question the relevance of the normality assumption (4). Usually, however, the counts are quite large. Recall for large mean, the Poisson distribution, often a model for count data, approaches the normal.

Table 1.1: *Data for 4 children in the Six Cities study.*

<i>Subject</i>	<i>City</i>	<i>Mother's smoking status</i>	<i>Wheezing status</i>
1	Portage	2	0
2	Kingston	0	0
3	Portage	0	0
4	Portage	1	1

*EXAMPLE 1.5 Six Cities study – Binary data.* Often, the response is not continuous, but rather is discrete. The most profound example is that where the response takes on only two possible values, generally denoted as  $Y = 0$  or  $1$ , corresponding to “absence” or “presence” of a condition or “failure” or “success.” An example is provide by data from a large public health study called the Six Cities Study, which was undertaken in six small, American cities to investigate a variety of public health issues. The full situation is reported in Lipsitz, Laird, and Harrington (1992). The data in Table 1.1 are those for the first 4 of 300 children in a study focused on the association between maternal smoking and child respiratory health at age 10. The response of interest was “wheezing status,” a measure of the child’s respiratory health, which was coded as either “no” (0) or “yes” (1), where “yes” corresponds to respiratory problems. Also recorded at each examination was a code to indicate the mother’s level of smoking: 0 = none, 1 = moderate, 2 = heavy.

Formally, for the  $j$ th child, the response is the binary variable  $Y_j$  taking on values 0 and 1. Mother’s smoking status is a categorical variable with three levels, so we represent it by two dummy variables defined as

$$\begin{aligned} s_{0j} &= 1 \text{ if none} & s_{1j} &= 1 \text{ if moderate} \\ &= 0 \text{ otherwise} & &= 0 \text{ otherwise} \end{aligned}$$

Let  $\mathbf{x}_j = (1, s_{0j}, s_{1j})^T$ . Following the same reasoning as in “classical” regression modeling, we would like to specify a model for  $E(Y_j|\mathbf{x}_j)$ . Now, here, as  $Y_j$  is binary,

$$E(Y_j|\mathbf{x}_j) = P(Y_j = 1|\mathbf{x}_j) = \pi(\mathbf{x}_j) = \pi_j,$$

say. So, from the usual perspective, we would like to write a model for the association between the probability of presence of wheezing and the covariate mother’s smoking status. Obviously, as probabilities must be between 0 and 1, a model that doesn’t respect this restriction is a poor candidate. A linear model  $\mathbf{x}_j^T \boldsymbol{\beta} = \beta_0 + \beta_1 s_{0j} + \beta_2 s_{1j}$ , say, has no property that forces its value to be between 0 and 1, so is not an appropriate choice.

Alternatively, a common approach is to postulate a model for  $\pi_j$  that does respect this restriction. The most popular such model is the so-called *logistic regression model*. This model is often motivated as follows. The *odds* is the ratio of the probability of seeing the event of interest to not seeing it, i.e.,  $\pi_j/(1-\pi_j)$ . The logistic regression model models the logarithm of the odds, or *logit*, as a linear function of  $\mathbf{x}_j$ . In particular, in the case of the Six Cities data, the model would be

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \mathbf{x}_j^T \boldsymbol{\beta} \quad \text{or} \quad \pi_j = \frac{\exp(\mathbf{x}_j^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_j^T \boldsymbol{\beta})}.$$

Note from the second expression that  $0 \leq \pi_j \leq 1$ .

- The interpretation is that an increase in one of the elements of  $\mathbf{x}_j$  by one unit changes the log odds by the value of the corresponding coefficient in  $\boldsymbol{\beta}$ ; for example, going from  $s_{0j} = 0$  to  $s_{0j} = 1$  changes the log odds by the additive amount  $\beta_1$ . Thus, the odds change by the multiplicative factor  $\exp(\beta_1)$ .

Note that this model is *nonlinear* in the elements of  $\boldsymbol{\beta}$ .

- Now with  $E(Y_j|\mathbf{x}_j) = \pi_j$ , we have immediately that

$$\text{var}(Y_j|\mathbf{x}_j) = \pi_j(1 - \pi_j),$$

so that, under the logistic regression model for  $E(Y_j|\mathbf{x}_j)$ ,

$$\text{var}(Y_j|\mathbf{x}_j) = \frac{\exp(\mathbf{x}_j^T \boldsymbol{\beta})}{\{1 + \exp(\mathbf{x}_j^T \boldsymbol{\beta})\}^2}.$$

Obviously, the variance is a function of the mean and hence  $\mathbf{x}_j$ , so varies with  $j$ . In fact, if we write

$$e_j = Y_j - \frac{\exp(\mathbf{x}_j^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_j^T \boldsymbol{\beta})},$$

because  $Y_j$  can only take on the values 0 and 1, for each  $j$ ,  $e_j$  can also only take on only two possible values for each  $\mathbf{x}_j$ .

CLASSICAL ASSUMPTIONS VIOLATED: (1), (2), and (4) are clearly violated. For (4), as  $Y_j$  is a Bernoulli random variable for each  $j$ , the normality assumption is not even a good approximation.

Table 1.2: *Data on damage to cargo ships – 6 ships.*

Ship type	Year constructed	Period of operation	Months service	# damage incidents
A	1960-64	1960-74	127	0
A	1970-74	1975-79	1512	6
B	1960-64	1960-74	44882	39
B	1960-64	1975-79	17176	29
C	1960-64	1960-74	1179	1
D	1970-74	1975-79	1208	11

*EXAMPLE 1.6 Wave damage to cargo ships.* The data in Table 1.2 are a subset of those reported in McCullagh and Nelder (1989, p. 205) concerning a type of damage caused by waves to the forward section of cargo ships. Each of a number of classes of ships, indexed by  $j$  is represented by the ship type, year of construction, period of operation, and months of aggregate service. The first three variables are categorical; defining appropriate dummy variables, the information on the  $j$ th combination of ship type constructed in a certain period with a certain period of operations, along with the aggregate months of service for that combination, may be summarized in a vector  $\mathbf{x}_j$ . The response  $Y_j$  is the number of damage incidents suffered by combination  $j$ . This response is in the form of a count; note from Table 1.2 that the value varies from very small, including no incidents (0) to rather large.

A natural distributional model for data in the form of counts is the Poisson distribution. In the regression context, we envision a Poisson distribution describing possible values of the number of damage incidents for each ship type represented by  $\mathbf{x}_j$ . That is,

$$P(Y_j = k|\mathbf{x}_j) = \frac{\exp(-\mu_j)\mu_j^k}{k!}, \quad k = 0, 1, 2, \dots,$$

where  $\mu_j = E(Y_j|\mathbf{x}_j)$  and, as dictated by the Poisson distribution,  $\text{var}(Y_j|\mathbf{x}_j) = \mu_j$ .

- Note that  $\mu_j$  appears to be quite small for some combinations, as indicated by the very small observed counts. As  $\mu_j > 0$  must hold, it might be dangerous to postulate a linear model such as  $E(Y_j|\mathbf{x}_j) = \mathbf{x}_j^T\boldsymbol{\beta}$ , as fitting such a model to data by, say, ordinary least squares could lead to negative predicted values in these cases. Rather, a model that respects the requirement of positive mean would be more sensible.
- A popular such model is the so-called *loglinear* model in which it is assumed that

$$\log E(Y_j|\mathbf{x}_j) = \mathbf{x}_j^T\boldsymbol{\beta} \quad \text{or} \quad E(Y_j|\mathbf{x}_j) = \exp(\mathbf{x}_j^T\boldsymbol{\beta}).$$

*CLASSICAL ASSUMPTIONS VIOLATED:* (1), (2), and (4) are clearly violated. Because of the presence of rather small counts, using the normal distribution as an approximation to the Poisson would not be reasonable, as such an approximation is only good when the mean is large.

#### 1.4 Further violation: multivariate response

So far in our examples, we have restricted attention to situations where the response may be viewed as *univariate* in the sense that each  $Y_j$  is a scalar quantity. In some of the cases, responses were observed on a single subject or plot, so that the scope of inference of necessity is restricted to observations on this individual or “unit.” Thus, even though the observations are repeated measurements on the same unit, we are only interested in modeling and inference for that unit. In other cases, like the Six Cities or cargo ship data, a single observation was available on each of several units, and interest focuses on inference about the population of units. Despite this distinction, in all of these cases, the data structure is the same: pairs consisting of scalar responses  $Y_j$  and associated covariates  $\boldsymbol{x}_j$ .

We now consider situations where there are several units, each with several responses. This puts us in the realm of what we will refer to as *multivariate response*. We describe some examples to illustrate.

*EXAMPLE 1.7 Developmental toxicology studies.* Developmental toxicology studies in rodents (rats, mice) are used in testing and regulation of potentially toxic substances that may pose danger to developing fetuses. An overview is given by Ryan (1992) and references therein. As described by Ryan, a typical study includes a control group of pregnant dams (exposed to 0 dose of the toxic agent) and several additional groups of pregnant dams exposed to different doses of the agent; usually, each group involves 20 to 30 dams. Exposure to non-zero doses typically leads to malformations of the fetuses, prenatal deaths, decreased birthweight, and so on, depending on the agent, which are considered responses to the exposure. The issues involved in risk assessment based on such observations are rather complicated; here, we do not attempt to discuss these issues but rather just use this situation to discuss the notion of *clustered data*.

For example, supposed that the response of interest is the continuous variable birthweight, and consider a study in which a total of  $m$  pregnant rats are exposed to different doses of a toxic agent. The objective is to characterize the effect on birthweight of different doses of the agent across the population of all exposed mothers and their pups.

Formally, index the pregnant rats by  $i = 1, \dots, m$ , and suppose that rat  $i$  is exposed to dose  $x_i$  and gives birth to  $n_i$  pups, where the  $j$ th pup,  $j = 1, \dots, n_i$ , has birthweight  $Y_{ij}$ . Several (e.g., 20) pregnant rats may be exposed to the same dose  $x_i$ . For each mother rat, then, we have observations  $Y_{i1}, \dots, Y_{in_i}$  on her  $n_i$  pups. These responses may be collected into a vector

$$\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T.$$

Thus, each mother yields a *multivariate response*, the vector of birthweights of each of her pups (with possibly different lengths  $n_i$ ).

- Suppose that we believe that, on average, birthweight is a simple linear function of dose. That is, suppose we believe that, for mother rats receiving dose  $x_i$ , the average birthweight across all such mothers and their pups is

$$E(Y_{ij}|x_i) = \beta_0 + \beta_1 x_i.$$

- Suppose further that it is reasonable to assume that, for all doses  $x_i$ , associated birthweights vary across all mothers and their pups in a similar way, so that

$$\text{var}(Y_{ij}|x_i) = \sigma^2,$$

say (constant for all  $i$  and  $j$ ). It may even be that (continuous) responses like birthweights for a given  $x_i$  may be reasonably thought to be normally distributed.

- An important issue, however, is that birthweights for pups born to the same mother may well be expected to be “more alike” than those compared across different mothers. In particular, some mothers may have a tendency to have “heavier” pups than other mothers, so all pups from a “heavy”-type mother given dose  $x_i$  will tend to deviate in a positive direction from the average birthweight for pups across all mothers given  $x_i$ . On the other hand, we would expect that the way in which birthweights for pups from mother  $i$  turn out would have nothing to do with how those for mother  $i'$  turn out, once we have accounted for the possibility that both mothers may have received the same dose.
- More formally, we may summarize these observations as follows. Clearly, the above reasoning indicates that we believe that birthweights for pups from the same mother may be *correlated*, simply because they may tend to be “heavy” or “light” together. However, birthweights for pups from different mothers would be reasonably assumed to be independent, as they may turn out “heavy” or “light” in a completely unrelated way. We may write this as

$$\begin{aligned} \text{corr}(Y_{ij}, Y_{i'j'}) &= \rho, \quad i = i' \text{ (same mother)} \\ &= 0 \quad i \neq i' \text{ (different mothers)} \end{aligned}$$

- Summarizing, we may write a model incorporating all of these considerations in terms of assumptions about the vectors  $\mathbf{Y}_i$ ,  $i = 1, \dots, m$ :

$$E(\mathbf{Y}_i|x_i) = (\beta_0 + \beta_1 x_i)\mathbf{1}_{n_i} = \boldsymbol{\mu}_i, \quad \text{var}(\mathbf{Y}_i|x_i) = \mathbf{V}_i,$$

where  $\mathbf{1}_{n_i}$  is a  $(n_i \times 1)$  vector of ones,  $\boldsymbol{\mu}_i = (\beta_0 + \beta_1 x_i, \dots, \beta_0 + \beta_1 x_i)^T$ ,  $(n_i \times 1)$ , and

$$\mathbf{V}_i = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \cdots & \rho & \rho & 1 \end{pmatrix} \quad (n_i \times n_i).$$

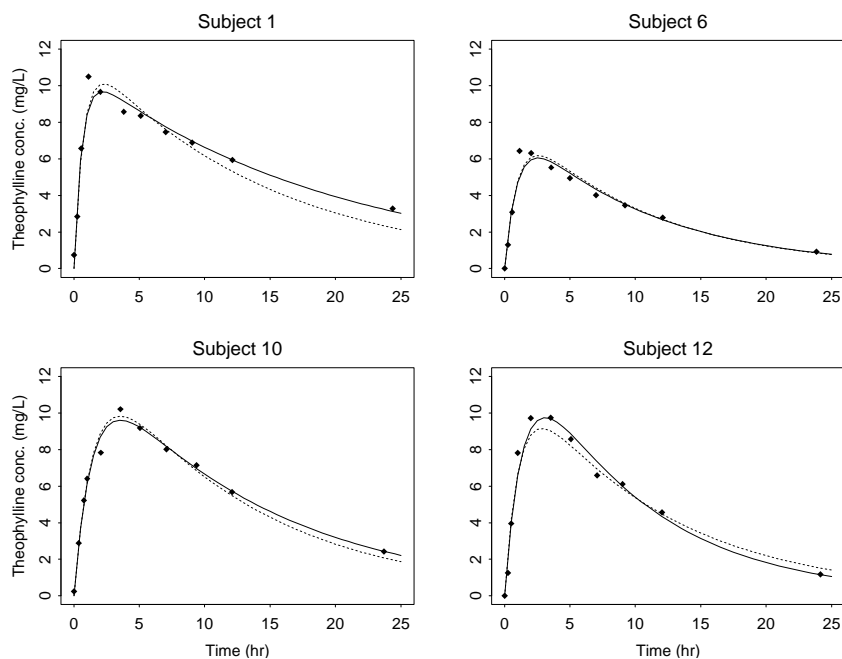
Note that, although the elements of the covariance matrix  $\mathbf{V}_i$  do not depend on  $i$ , the subscript  $i$  is necessary, as it denotes dependence of its dimension on  $n_i$ .

*Notational note:* Throughout, we will use  $\text{var}(\cdot)$  to denote both variance of a scalar random variable and covariance matrix of a random vector. The meaning should be clear from the context.

- This has the flavor of a “regression model,” but with *multivariate* response, so pertaining to a set of pairs  $(\mathbf{Y}_1, x_1), \dots, (\mathbf{Y}_m, x_m)$ , where the response is a random vector. The assumption on variance is replaced with an assumption about the covariance matrix of such random vectors.
- In fact, the response need not be continuous. For example, suppose that, instead, the pregnant dams were sacrificed prior to giving birth, and the fetuses of each examined. For dam  $i$ , the status of each of the  $n_i$  fetuses is either “normal” or “malformed,” which may be summarized as  $Y_{ij} = 0$  (normal) or  $Y_{ij} = 1$  (malformed) for the  $j$ th fetus from dam  $i$ . In this case, the vector  $\mathbf{Y}_i$  for dam  $i$  would be a vector whose elements are binary, taking on only the values 0 or 1. Nonetheless, the same issues arise: We may wish to model the probability of malformation as a function of dose, taking into account that the way in which fetuses turn out for a given dam  $i$  will tend to be “more alike” (correlated) than across dams.

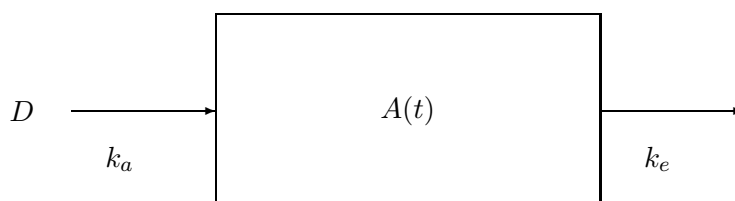
*EXAMPLE 1.8 Pharmacokinetics of theophylline.* As another example of multivariate response, we consider data from a study of the anti-asthmatic agent theophylline. Figure 1.7 shows concentration-time data for four of the 12 subjects in the study, each of whom received a single oral dose of theophylline (given in units of mg/kg, so scaled to each individual’s body weight in kg). For each subject, 10 blood samples were drawn following the dose and assayed for theophylline concentration.

Figure 1.7: *Theophylline concentration-time profiles for 4 subjects receiving an oral dose of theophylline at time 0, with fits of model (1.5) superimposed.*



In Examples 1.1 and 1.2 we considered pharmacokinetic data from a single subject, and the objective was to characterize the pharmacokinetic behavior (the concentration-time profile, underlying parameters) for that subject only. Although this is sometimes done to aid in the selection of an appropriate dosing regimen for a particular subject, it is far more common that data are obtained on several subjects with the broader objective of understanding pharmacokinetic behavior in the entire population of subjects. In fact, the data for the single subject in Examples 1.1 and 1.2 were taken from such a study.

Understanding pharmacokinetic behavior in the population means understanding how individual concentration-time profiles and the parameters that characterize them vary across the population of individuals. To appreciate this, recall the usual compartmental modeling strategy discussed in Example 1.1. For an oral dose given at time 0, a standard model that appears to represent well theophylline concentration-time profiles, is the one compartment open model with first order absorption, represented pictorially as



Here,  $A(t)$  is the amount of drug in the “blood compartment” at time  $t$ . The dose  $D$  given at time 0 is absorbed through the gut into the blood at a fractional absorption rate of  $k_a$ , and is eliminated (e.g., excreted by the kidneys and metabolized by the liver) at fractional elimination rate  $k_e$ . A system of differential equations, with appropriate initial conditions may be written down and solved for  $A(t)$ , see, for example, Gibaldi and Perrier (1982). Dividing by  $V$ , the volume of the blood compartment, yields the following expression for concentration of drug at time  $t$ :

$$C(t) = \frac{A(t)}{V} = \frac{k_a DF}{V(k_a - k_e)} \{\exp(-k_e t) - \exp(-k_a t)\}, \quad k_e = Cl/V, \quad (1.5)$$

Now, the compartment model and the ensuing model for concentration  $C(t)$  in (1.5) pertain to *individual* subject behavior; that is, the model is a theoretical description of biological processes taking place over time *within* a given subject, as that subject processes the drug. The parameter  $F$ , the bioavailability, is usually taken to be equal to 1. The parameter  $Cl$ , the clearance rate, is a measure of the volume of blood cleared of drug per unit time, and is of primary importance in understanding how the drug is eliminated from the system.

- If we were interested only in a particular subject’s behavior, then, if that subject received dose  $D$ , letting  $t_j$  denote the  $j$ th observation time for that subject where concentration  $Y_j$  was measured, and  $\mathbf{x}_j = (D, t_j)^T$ , we could postulate a regression model

$$E(Y_j | \mathbf{x}_j) = \frac{D\beta_3}{\beta_2(\beta_3 - \beta_1/\beta_2)} \{\exp(-\beta_1 t_j / \beta_2) - \exp(-\beta_3 t_j)\}, \quad (1.6)$$

where  $\beta_1 = Cl$ ,  $\beta_2 = V$ , and  $\beta_3 = k_a$ ,  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$ . As before, interest would focus on estimation of  $\boldsymbol{\beta}$ .

- However, we are interested in the population of subjects. Indexing subjects by  $i = 1, \dots, m$ , let  $Y_{ij}$  denote the  $j$ th concentration measurement for subject  $i$  receiving dose  $D_i$  at time 0, where the time points at which concentration is measured are  $t_{ij}$ ,  $j = 1, \dots, n_i$  (the times may possibly be different for different subjects). Letting  $\mathbf{x}_{ij} = (D_i, t_{ij})^T$ , we have pairs  $(Y_{i1}, \mathbf{x}_{i1}), \dots, (Y_{in_i}, \mathbf{x}_{in_i})$  for subject  $i$ . The responses on subject  $i$  may be collected in time order into the vector

$$\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T, \quad (n_i \times 1).$$

From our previous discussion, note that including “time” in the covariate vector  $\mathbf{x}_{ij}$  may not be entirely satisfactory, as what we have are realizations of a stochastic process for each subject  $i$  at times  $t_j$ , but we do not dwell on this for now.

- From the reasoning above, *each* subject could be thought of as having a regression relationship of the form (1.6); however, as pharmacokinetic behavior is a *within-individual* process, it would be expected that each subject would have his or her own pharmacokinetic parameters  $\beta$  governing his or her individual behavior. For example, it is biologically implausible that the process of elimination (excretion and metabolism) takes place in exactly the same way for all subjects. From the point of view of the model (1.6), different subjects would thus be expected to have different values of the clearance rate  $\beta_2$ .

We could thus think of a model for subject  $i$  depending on his or her *individual-specific* set of pharmacokinetic parameters  $\beta_i = (\beta_{1i}, \beta_{2i}, \beta_{3i})^T$ :

$$E(Y_{ij} | \mathbf{x}_{ij}, \beta_i) = \frac{D\beta_{3i}}{\beta_{2i}(\beta_{3i} - \beta_{1i}/\beta_{2i})} \{ \exp(-\beta_{1i}t_j/\beta_{2i}) - \exp(-\beta_{3i}t_j) \}.$$

- Of course, the elements of  $\mathbf{Y}_i$  would be expected to be correlated, as concentrations from the same individual might tend to be “more alike.” Moreover, as discussed in Example 1.2, the variance of observations on a given subject tends to increase with concentration level, and there is the potential for within-subject serial correlation (which is different from correlation due to differences *across* subjects). We defer discussion of these issues until later chapters.
- Returning to the objective of the study, understanding how pharmacokinetic behavior varies in the population of subjects would clearly involve understanding of how the parameters  $\beta_i$  vary across subjects; more precisely, understanding their *distribution*. Of course, these parameters cannot be observed, rather, information on them is only available through the multivariate response vectors  $\mathbf{Y}_i$  available from each of  $m$  subjects drawn from the population.

Clearly, the modeling and analysis associated with this objective is considerably more complicated than that involved in the usual, “classical” regression framework.

Note that the data structure in this example is similar to that in the soybean growth study Example 1.3. In the discussion of that example, we focused on modeling and analysis for a single plot and postulated the logistic growth model as a way to represent the individual-plot growth process. Of course, in this study, the real objective was to understand soybean growth patterns across the entire populations of plots planted with the two genotypes. Thus, it should be evident that, to address this objective, a similar type of model framework as that for Example 1.8 would be required.

In these last two examples, we see that the extension of regression-type modeling to multivariate response obviously requires a greater level of complexity. Moreover, it appears that different ways of thinking may be more natural for different problems.

- In Example 1.7 on developmental toxicity, we wrote down a model to describe dose-response in the population of rats directly. This made sense – each mother rat was seen at a single dose.
- In contrast, in Example 1.8, the theophylline pharmacokinetic study, a model for individual behavior was postulated, where the population came into the model through the idea of individual-specific parameters  $\beta_i$  governing individual processes.

We will formalize and discuss both types of modeling strategies in later chapters.

In any event, a general issue relevant in all multivariate response situations is the theme that observations from the same unit  $i$  (subject, rat, plot) may be *correlated*, while responses taken from different units may be reasonably assumed as independent.

- Thus, it is important to recognize that, although it may be tempting to simply think of all the  $N = \sum_{i=1}^m n_i$  observations altogether as a single vector  $\mathbf{Y}$  ( $N \times 1$ ), it would not be appropriate to assume that all the elements of  $\mathbf{Y}$  are *independent*, as would be the case if one adopted the “classical” regression assumptions without careful thought.
- That is, it is not appropriate to try to “force” situations like those in Examples 1.7 and 1.8 into the “classical” regression framework. In situations such as these, the assumption (3) of independence is surely violated. Indeed, the other assumptions of linearity, constant variance, and normality are also violated.

## 1.5 Summary and a look ahead

The preceding series of examples illustrates the issues associated with regression modeling when one moves away from the “classical” regression framework that is generally introduced in a first course on regression analysis. The examples emphasize that more complex and flexible models and associated inferential methods are required.

It is important to recognize that, although not emphasized in a typical first course in regression, a *regression model* is, in broad generality, nothing more than a postulated description of the conditional expectation  $E(Y|\mathbf{x})$ , where  $Y$  and  $\mathbf{x}$  are viewed as random variable/vectors. This, of course, is a function of  $\mathbf{x}$ , so part of the art of regression modeling is positing a plausible functional form for this conditional expectation. A regression model may also embody further assumptions, such as the “classical” assumption that  $\text{var}(Y|\mathbf{x})$  does not depend on  $\mathbf{x}$ , or even assumptions on the entire conditional distribution of  $Y$  given  $\mathbf{x}$ . As we discuss at the beginning of the next chapter, we will take this point of view throughout. The extent to which one is willing (or not) to make such assumptions will be an important focus of this course.

We will begin our study by confining attention to *univariate* response problems. We will begin in Chapter 2 with an introduction to a general (univariate) nonlinear regression model framework that will form the basis for the material in Chapters 3–12, all of which focus on the model, its implementation, and theoretical results that form the groundwork for commonly-used inferential procedures. In terms of the “classical” assumptions, the models and methods we will discuss accommodate violation of assumptions (1) (linearity), (2) (constant variance and independence of  $e_j$  and  $\mathbf{x}_j$ ), and (4) (normality). We will maintain assumption (3) (independence).

In Chapters 13–15, we will take up the multivariate response case. This will also involve violation of assumption (3). The issues are much more complex than those in univariate regression modeling, as we will see. In fact, a critical issue will be the fact that, in contrast to univariate modeling, there are different approaches, of which we will discuss two of the most popular. Which approach is most suitable for a given problem will depend on the context and the scientific objectives.