11 Generalized linear models for nonnormal response

11.1 Introduction

So far, in our study of "regression-type" models for longitudinal data, we have focused on situations where

- The response is **continuous** and reasonably assumed to be **normally distributed**.
- The model relating mean response to **time** and possibly other covariates is **linear** in parameters that characterize the relationship. For example, regardless of how we modeled covariance (by direct modeling or by introducing random effects), we had models for the mean response of a data vector of the form

  $$E(\boldsymbol{Y}_i) = \boldsymbol{X}_i\boldsymbol{\beta};$$

  i.e. for the observation at time t_{ij} on unit i,

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

Under these conditions, we were led to methods that were based on the assumption that

$$\boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i);$$

the form of the matrix $\boldsymbol{\Sigma}_i$ is dictated by what one assumes about the nature of variation. To fit the model, we used the methods of **maximum likelihood** and **restricted maximum likelihood under the assumption** that the data vectors are distributed as **multivariate normal**. Thus, the fitting method was based on the normality assumption.

As we noted at the beginning of the course, the assumption of normality is not always relevant for some data. This issue is not confined to longitudinal data analysis – it is an issue even in ordinary regression modeling. If the response is in the form of small **counts**, or is in fact **binary** (yes/no), it is clear that the assumption of normality would be quite unreasonable. Thus, the modeling and methods we have discussed so far, including the classical techniques, would be inappropriate for these situations.

One possibility is to analyze the data on a **transformed** scale on which they appear to be more nearly normal; e.g. count data may be transformed via a square-root or other transformation, and then represented by linear models on this scale. This is somewhat unsatisfactory, however, as the model no longer pertains directly to the original scale of measurement, which is usually of greatest interest. Moreover, it tries to "force" a model framework and distributional assumption that may not be best for the data.

In the late 1970'/early 1980's, in the context of ordinary regression modeling, a new perspective emerged in the statistical literature that generated much interest and evolved into a new standard for analysis in these situations. For data like counts and binary outcomes, as well as for continuous data for which the normal distribution is not a good probability model, there are **alternative** probability models that might be better representations of the way in which the response takes on values. The idea was to use these more appropriate probability models as the basis for developing new regression models and methods, rather than to try and make things fit into the usual (and inappropriate) normal-based methods. Then, in the mid-1980's, these techniques were extended to allow application to longitudinal data; this topic still is a focus of current statistical research.

In this chapter, we will gain the necessary background for understanding longitudinal data methods for nonnormal response. To do this, we will step away from the longitudinal data problem in this chapter, and consider just the ordinary regression situation where responses are **scalar** and **independent**. Armed with an appreciation of regression methods for nonnormal response, we will then be able to see how these might be extended to the harder problem of **longitudinal data**. As we will see, this extension turns out to not be quite as straightforward as it was in the normal case.

Thus, in this chapter, we will consider the following problem as a prelude to our treatment of nonnormal longitudinal data:

- As in multiple regression, suppose we have responses $Y_1, \ldots, Y_n$ each taken at a setting of k covariates $x_{j1}, \ldots, x_{jk}$, $j = 1, \ldots, n$.
- The Y_j values are mutually **independent**.
- The goal is to develop a **statistical model** that represents the response as a function of the covariates, as in usual linear regression.
- However, the nature of the response is such that the **normal** probability model is **not** appropriate.

We might think of the data as arising either as

- n observations on a single unit in a longitudinal data situation, where we focus on this individual unit **only**, so that the only relevant variation is **within** the unit. If observations are taken far enough apart in time, they might be viewed as independent.
- n **scalar** observations, each taken on a different unit (thus, the independence assumption is natural). Here, j indexes observations and units (recall the oxygen intake example in section 3.4).
- Either way of thinking is valid – the important point is that we wish to fit a regression model to data that do not seem to be normally distributed. As we will see, the data type might impose **additional** considerations about the form of the regression model.
- We use the subscript j in this chapter to index the observations; we could have equally well used the subscript i.

The class of regression models we will consider for this situation is known in the literature as **generalized linear models** (not to be confused with the name of the SAS procedure `GLM` standing for `G`eneral `L`inear `M`odel). Our treatment here is not comprehensive; for everything you ever wanted to know and more about generalized linear models, see the book by McCullagh and Nelder (1989).

11.2 Probability models for nonnormal data

Before we discuss regression modeling of nonnormal data, we review a few probability models that are ideally suited to representation of these data. We will focus on three models in particular; a more extensive catalogue of models may be found in McCullagh and Nelder (1989):

- The **Poisson** probability distribution as a model for **count** data (discrete)
- The **Bernoulli** probability distribution as a model for **binary** data (discrete) (this may be extended to model data in the form of **proportions**
- The **gamma** probability distribution as a model for **continuous** but nonnormal data with **constant coefficient of variation**.

We will see that all of these probability models are members of a special class of probability models. This class also includes the **normal** distribution with constant variance (the basis for classical linear regression methods for normal data); thus, generalized linear models will be seen to be an **extension** of ordinary linear regression models.

COUNT DATA – THE POISSON DISTRIBUTION: Suppose we have a response Y that is in the form of a **count** – Y records the number of times an event of interest is observed. Recall the epileptic seizure data discussed at the beginning of the course; here, Y was the number of seizures suffered by a particular patient in a two-week period.

When the response is a count, it should be clear that the possible values of the response must be nonnegative integers; more precisely, Y may take on the values $0, 1, 2, 3, \ldots$. In principle, **any** nonnegative integer value is possible; there is no upper bound on how large a count may be. Realistically, if the thing being counted happens infrequently, large counts may be so unlikely as to almost never be seen.

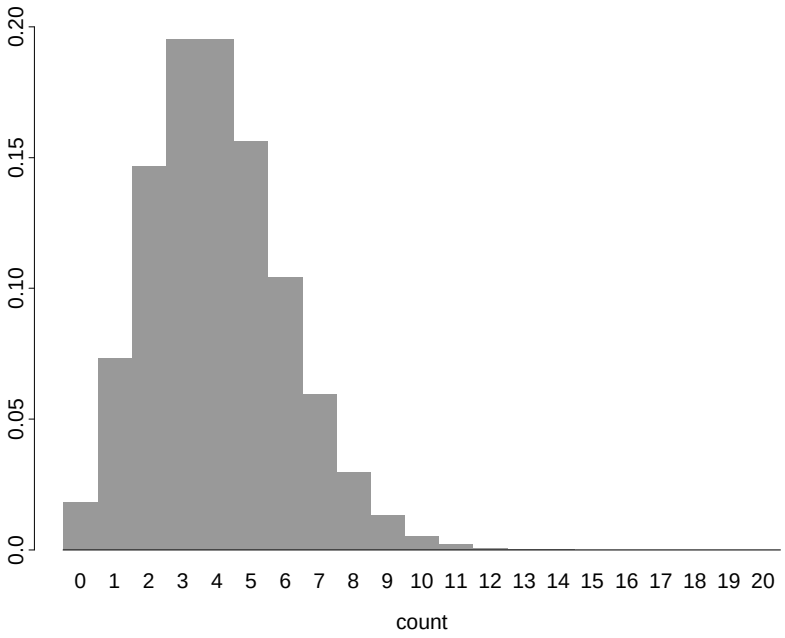
The **Poisson** probability distribution describes probabilities that a random variable Y that describes counts takes on values in the range $0, 1, 2, 3, \ldots$. More precisely, the probability density function describes the probability that Y takes on the value y:

$$f(y) = P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, 2, \ldots, \quad \mu > 0. \tag{11.1}$$

- It may be shown that the **mean** (**expectation**) of Y is μ; i.e. $E(Y) = \mu$. Note that μ is **positive**, which makes sense – the average across all possible values of counts should be positive.
- Furthermore, it may be shown that the **variance** of Y is also equal to μ; i.e. $\text{var}(Y) = \mu$. Thus, the variance of Y is **nonconstant**. Thus, if Y_1 and Y_2 are both Poisson random variables, the only way that they can have the **same variance** is if they have the **same mean**.
- This has implications for regression – if Y_1 and Y_2 correspond to counts taken at **different** settings of the covariates, so thus at possibly different mean values, it is inappropriate to assume that they have the same variance. Recall that a standard assumption of ordinary regression under normality is that of **constant** variance regardless of mean value; this assumption is clearly not sensible for count data.

Figure 1 shows the **probability histogram** for the case of a Poisson distribution with $\mu = 4$. Because the random variable in question is **discrete**, the histogram is not smooth; rather, the blocks represent the probabilities of each value on the horizontal axis by **area**.

Figure 1: *Poisson probabilities with mean = 4.*



Some features:

- Probabilities of seeing counts larger than 12 are virtually negligible, although, in principle, counts may take on **any** nonnegative value.
- Clearly, if μ were larger, the values for which probabilities would become negligible would get larger and larger.
- For "smallish" counts, where the mean is small (e.g. $\mu = 4$), the shape of the probability histogram is **asymmetric**. Thus, discreteness aside, the normal distribution would be a lousy approximation to this shape. For larger and larger μ, it may be seen that the shape gets more and more symmetric. Thus, when counts are very large, it is common to approximate the Poisson probability distribution by a normal distribution.

EXAMPLE – HORSE-KICK DATA: As an example of a situation where the response is a (small) count, we consider a world-famous data set. These data may be found on page 227 of Hand *et al.* (1994). Data were collected and maintained over the 20 years 1875 – 1894, inclusive, on the numbers of Prussian militiamen killed by being kicked by a horse in each of 10 separate corps of militiamen. For example, the data for the first 6 years are as follows:

Year	Corps									
	1	2	3	4	5	6	7	8	9	10
1875	0	0	0	0	1	1	0	0	1	0
1876	0	0	1	0	0	0	0	0	1	1
1877	0	0	0	0	1	0	0	1	2	0
1878	2	1	1	0	0	0	0	1	1	0
1879	0	1	1	2	0	1	0	0	1	0
1880	2	1	1	1	0	0	2	1	3	0

Thus, for example, in 1877, 2 militiamen were killed by kicks from a horse in the 9th corps. Note that, technically, counts may not be **any** number – there is an "upper bound" (the total number of men in the corps). But this number is so huge relative to the size of the counts that, for all practical purposes it is "infinite." Clearly, the numbers of men killed (counts) in each year/corps combination are small; thus, the normal distribution is a bad approximation to the true, Poisson distribution.

It was of interest to determine from these data whether differences in the numbers of men kicked could be attributed to systematic effects of year or corps. That is, were members of certain corps more susceptible to horse-kick deaths than others? Were certain years particularly bad for horse-kick deaths?

- If the data were normal, a natural approach to this question would be to postulate a **regression model** that allows mean response to depend on the particular corps and year.
- Specifically, if we were to define 19 **dummy** variables for year and 9 for corps, we might write a **linear model** for the mean of the jth observation in the data set ($n = 200$ total) as

$$\beta_0 + \beta_1 x_{j1} + \cdots + \beta_{19} x_{j,19} + \beta_{20} z_{j1} + \cdots + \beta_{28} z_{j9}, \tag{11.2}$$

$$\begin{aligned} x_{jk} &= 1 \text{ if observation } j \text{ is from year } k = 1875, \ldots, 1893 \\ &= 0 \text{ otherwise} \\ z_{jk} &= 1 \text{ if observation } j \text{ is from corps } k = 1, \ldots, 9 \\ &= 0 \text{ otherwise} \end{aligned}$$

With these definitions, note that β_0 corresponds to what happens for year 1894 with corps 10. The remaining parameters describe the change from this due to changing year or corps.

- Note that, aside from the normality issue, letting (11.2) represent the mean of observation Y_j, $E(Y_j)$ has a problem. Recall that counts **must** be nonnegative by definition. However with this model, it is possible to end up with an estimated value for $E(Y_j)$ that is **negative** – this restriction is not enforced. This seems quite possible – many of the observations are 0, so that it would not be surprising to end up estimating some means as negative. More on this later.

BINARY DATA – THE BERNOULLI DISTRIBUTION: Suppose we have a response y that takes on either the value 0 or 1 depending on whether an event of interest occurs or not. Recall the child respiratory data at the beginning of the course; here, y was 0 or 1 according to whether a child did not or did "wheeze."

Here, the response can take on only two possible values. Clearly, the normal distribution should not even be considered as a model.

The **Bernoulli** probability distribution describes probabilities that a random variable Y that characterizes whether an event occurs or not takes on its two possible values (0, 1). The probability density function is given by

$$f(1) = P(Y = 1) = \mu, \quad f(0) = P(Y = 0) = 1 - \mu$$

for $0 \leq \mu \leq 1$. The extremes $\mu = 0, 1$ are not particularly interesting, so we will consider $0 < \mu < 1$. This may be summarized succinctly as

$$f(y) = P(Y = y) = \mu^y (1 - \mu)^{(1-y)}, \quad 0 < \mu < 1, \quad y = 0, 1. \tag{11.3}$$

- It may be shown that the **mean** of Y is μ. Also, note that μ is also the probability of seeing the event of interest ($y = 1$). As a probability, it must be between 0 and 1, so that the mean of Y must be between 0 and 1 as well.

- Furthermore, it may be shown that the **variance** of Y is equal to $\mu(1-\mu)$; i.e. $\text{var}(Y) = \mu(1-\mu)$. As with the Poisson distribution, the variance of Y is **nonconstant**. Thus, if Y_1 and Y_2 are both Bernoulli random variables, the only way that they can have the **same variance** is if they have the **same mean**.

- This has implications for regression – if Y_1 and Y_2 correspond to binary responses taken at **different** settings of the covariates, so thus at possibly different mean values, it is inappropriate to assume that they have the same variance. Thus, again, the usual assumption of constant variance is clearly not sensible when modeling binary data.

EXAMPLE – MYOCARDIAL INFARCTION DATA: The response is often binary in medical studies. Here, we consider an example in which 200 women participated in a study to investigate risk factors associated with myocardial infarction (heart attack). On each woman, the following information was observed:

- Whether the woman used oral contraceptives in the past year (1 if yes, 0 if no)
- Age in years
- Whether the woman currently smokes more than 1 pack of cigarettes per day (1 if yes, 0 if no)
- Whether the woman has suffered a myocardial infarction – the response ($y = 0$ if no, $y = 1$ if yes).

The data for the first 10 women are given below:

Woman	Contracep.	Age	Smoke	MI
1	1	33	1	0
2	0	32	0	0
3	1	37	0	1
4	0	36	0	0
5	1	50	1	1
6	1	40	0	0
7	0	35	0	0
8	1	33	0	0
9	1	33	0	0
10	0	31	0	0

The objective of this study was to determine whether any of the covariates, or potential **risk factors** (oral contraceptive use, age, smoking), were associated with the chance of having a heart attack. For example, was there evidence to suggest that smoking more than one pack of cigarettes a day raises the probability of having a heart attack?

- If the data were normal, a natural approach to this question would be to postulate a **regression model** that allows mean response (which is equal to probability of having a heart attack as this is a binary response) to depend on age, smoking status, and contraceptive use.

- Define for the jth woman

$$\begin{aligned} x_{j1} &= 1 \text{ if oral contraceptive use} \\ &= 0 \text{ otherwise} \\ x_{j2} &= \text{ age in years} \\ x_{j3} &= 1 \text{ if smoke more then one pack/day} \\ &= 0 \text{ otherwise} \end{aligned}$$

  Then we would be tempted to model the mean (probability of heart attack) as a **linear model**, writing the mean for the j observation

$$\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \beta_3 x_{j3}.$$

- Using a linear function of the covariates like this to represent the mean (probability of heart attack) has an immediate problem. Because the mean is a probability, it must be between 0 and 1. There is **nothing** to guarantee that the estimates of means we would end up with after fitting this model in the usual way would honor this restriction. Thus, we could end up with **negative** estimates of probabilities, or estimated probabilities that were **greater** than one! More on this later.

CONTINUOUS DATA WITH CONSTANT COEFFICIENT OF VARIATION – THE GAMMA DISTRIBUTION: As we have already remarked, just because the response is continuous does not mean that the normal distribution is a sensible probability model.

- For example, most biological responses take on only **positive** values. The normal distribution in principle assigns positive probability to **all** values on the real line, negative and positive.

- Furthermore, the normal distribution says that values to the left and right of its mean are **equally likely** to be seen, by virtue of the **symmetry** inherent in the form of the probability density. This may not be realistic for biological and other kinds of data. A common phenomenon is to see "unusually large" values of the response with more frequency than "unusually small" values. For example, if the response is **annual income**, the distribution of incomes is mostly in a limited range; however, every so often, a "chairman of the board," athlete, or entertainer may command an enormous income. For this situation, a distribution that says small and large values of the response are equally likely is not suitable.

Other probability models are available for continuous response that better represent these features. Several such models are possible; we consider one of these.

The **gamma** probability distribution describes the probabilities with which a random variable Y takes on values, where Y can only be **positive**. More precisely, the probability density function for value y is given by

$$f(y) = \frac{1}{y\Gamma(1/\sigma^2)} \left(\frac{y}{\sigma^2 \mu} \right)^{1/\sigma^2} \exp\left(-\frac{y}{\sigma^2 \mu} \right), \quad \mu, \sigma^2 > 0, \quad y > 0. \tag{11.4}$$

- In (11.4), $\Gamma(\cdot)$ is the so-called "Gamma function." This function of a positive argument may only be evaluated on a computer. If the argument is a positive integer k, however, then it turns out that $\Gamma(k) = (k-1)! = (k-1)(k-2)\cdots(2)(1)$.

- It may be shown that the **mean** of Y is μ; i.e. $E(Y) = \mu$. Note that μ must be **positive**, which makes sense.

- It may also be shown that the **variance** of Y is $\text{var}(Y) = \sigma^2\mu^2$. That is, the variance of Y is **nonconstant**; it depends on the value of μ. Thus, if Y_1 and Y_2 are both gamma random variables, then the only way that they can have the same variance is if they have the same mean μ and the same value of the parameter σ^2.

- Thus, for regression, if Y_1 and Y_2 correspond to responses taken at different covariate settings, it is inappropriate to take them to have the same variance. Thus, as above, the assumption of constant variance is not appropriate for a response that is well-represented by the gamma probability model.

- In fact, note here that the symbol σ^2 is being used here in a different way from how we have used it in the past, to represent a **variance**. Here, it turns out that σ (not squared) has the interpretation as the **coefficient of variation** (CV), defined for any random variable Y as

$$CV = \frac{\{\text{var}(Y)\}^{1/2}}{E(Y)};$$

that is, CV is the ratio of standard deviation of the response to mean, or "**noise to signal**." This ratio may be expressed as a **proportion** or a **percentage**; in either case, CV characterizes the "quality" of the data by quantifying how large the "noise" is relative to the size of the thing being measured.

- "Small" CV ("high quality") is usually considered to be $CV \leq 0.30$. "Large" CV ("low quality") is larger.

- Note that for the gamma distribution,

$$CV = \frac{(\sigma^2\mu^2)^{1/2}}{\mu} = \sigma,$$

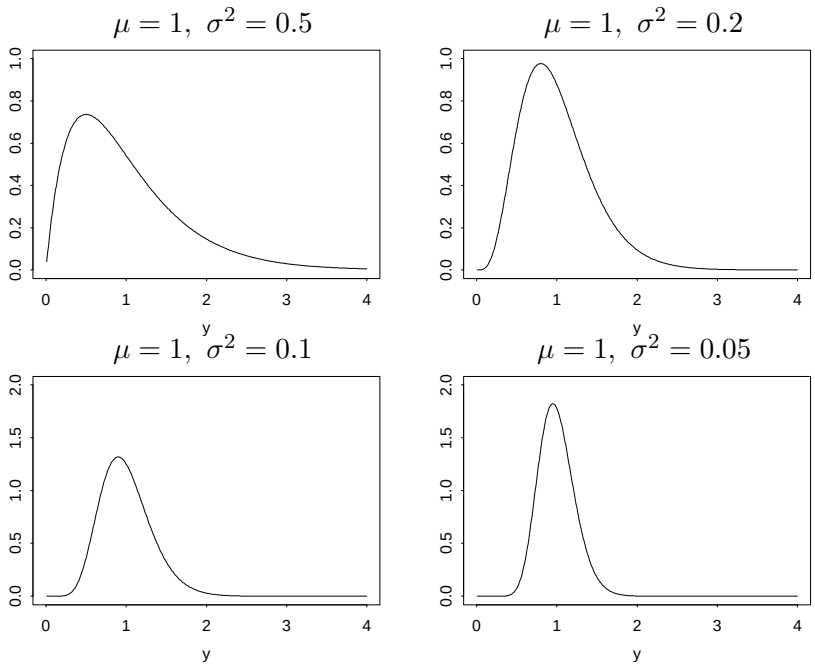
so that, **regardless** of the value of μ, the ratio of "noise" to "signal" is the same. Thus, rather than having **constant variance**, the gamma distribution imposes **constant coefficient of variation**. This is often a realistic model for biological, income, and other data taking on positive values.

Figure 2 shows gamma probability density functions for $\mu = 1$ and progressively smaller choices of σ^2, corresponding to progressively smaller CV.

- As σ^2 becomes smaller, the shape of the curve begins to look more **symmetric**. Thus, if CV is "small" ("high quality" data), gamma probability distribution looks very much like a normal distribution.

- On the other hand, when σ^2 is relatively large, so that CV is "large" ("low quality" data), the shape is **skewed**. For example, with $\sigma^2 = 0.5$, corresponding to $CV = 0.707$, so "noise" that is 70% the magnitude of the "signal" (upper left panel of Figure 2), the shape of the gamma density does not resemble that of the normal at all.

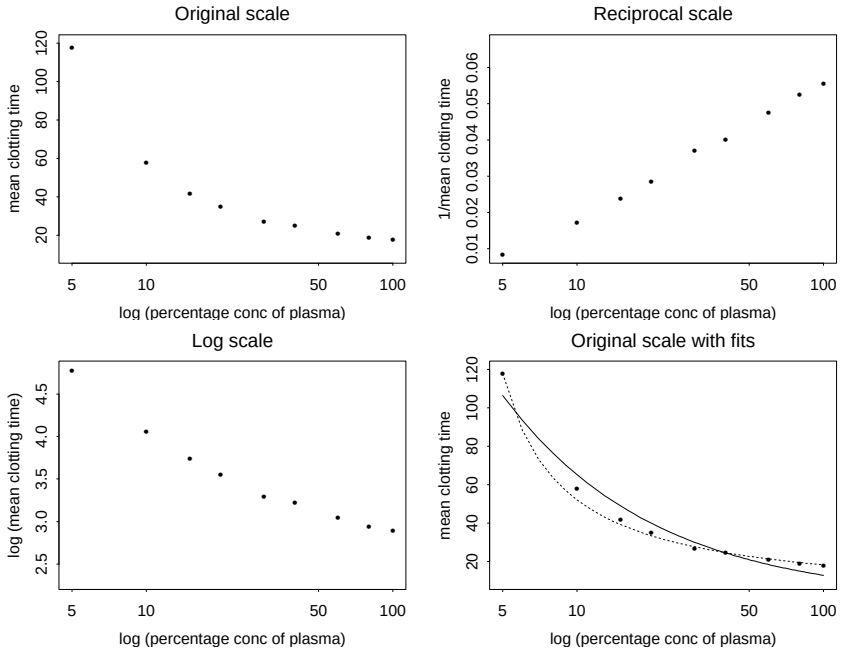
Figure 2: *Gamma probability density functions.*



EXAMPLE – CLOTTING TIME DATA: In the development of clotting agents, it is common to perform *in vitro* studies of time to clotting. The following data are reported in McCullagh and Nelder (1989, section 8.4.2), and are taken from such a study. Here, samples of normal human plasma were diluted to one of 9 different percentage concentrations with prothrombin-free plasma; the higher the dilution, the more the interference with the blood's ability to clot, because the blood's natural clotting capability has been weakened. For each sample, clotting was induced by introducing thromboplastin, a clotting agent, and the time until clotting occurred was recorded (in seconds). 5 samples were measured at each of the 9 percentage concentrations, and the mean clotting times were averaged; thus, the response is mean clotting time over the 5 samples. The response is plotted against percentage concentration (on the log scale) in the upper left panel of Figure 3. We will discuss the other panels of the figure shortly.

It is well-recognized that this type of response, which is by its nature always positive, does **not** exhibit the same variability at all levels. Rather, large responses tend to be more variable than small ones, and a constant coefficient of variation model is often a suitable model for this nonconstant variation.

Figure 3: *Clotting times (seconds) for normal plasma diluted to 9 different concentrations with prothrombin-free plasma. In the lower right panel, the solid line is the loglinear fit, the dashed line is the reciprocal (inverse) fit.*



From the plot, it is clear that a straight-line model for mean response as a function of log(percentage concentration) would be inappropriate. A quadratic model seems better, but, because such models eventually curve "back up," this might not be a good model, either. In the upper right and lower left panels, the **reciprocals** $(1/y)$ and **logarithms** $(\log y)$ of the response, respectively, are plotted against log(percentage concentration). These appear to be roughly like straight lines, the former more-so than the latter. We will return to the implications of these two plots for choosing a model for mean response shortly. Note, of course, that a sensible model for mean response would be one that honors the positivity restriction for the response.

Also noticeable from the plot is that the data are of "high quality" – the pattern of change in the response with log(percentage concentration) is very clear and smooth, with very little "noise." This would suggest that if the data really are well-represented by the gamma probability distribution, then the coefficient of variation is "small." From the plot, it is very difficult to see any evidence of that the variance really is nonconstant as the response changes – this is due to the fact that variation is just so small, so it is hard to pick up by eye.

We will return to these data shortly.

SUMMARY: The Poisson, Bernoulli, and gamma distributions are three different probability distributions that are well-suited to modeling data in the form of counts, binary response, and positive continuous response where constant coefficient of variation is more likely than constant variance, respectively. As mentioned above, still other probability distributions for other situations are available; discussion of these is beyond our scope here, but the implications are similar to the cases we have covered. We now turn to regression modeling in the context of problems where these probability distributions are appropriate.

11.3 Generalized linear models

THE CLASSICAL LINEAR REGRESSION MODEL: The classical linear regression model for scalar response Y_j and k covariates $x_{j1}, \ldots, x_{jk}$ is usually written as

$$Y_j = \beta_0 + \beta_1 x_{j1} + \cdots + \beta_k x_{jk} + \epsilon_j$$

or, defining $\boldsymbol{x}_j = (1, x_{j1}, \ldots, x_{jk})'$, where $\boldsymbol{x}_j$ is $(p \times 1)$, $p = k + 1$,

$$Y_j = \boldsymbol{x}_j'\boldsymbol{\beta} + \epsilon_j, \quad \boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'. \tag{11.5}$$

The Y_j are assumed to be independent across j. When the response is continuous, it is often assumed that the ϵ_j are independent $\mathcal{N}(0, \sigma^2)$, so that

$$Y_j \sim \mathcal{N}(\boldsymbol{x}_j'\boldsymbol{\beta}, \sigma^2).$$

That is, the classical, normal-based regression model may be summarized as:

(i) **Mean**: $E(Y_j) = \boldsymbol{x}_j'\boldsymbol{\beta}$.

(ii) **Probability distribution:** Y_j follow a normal distribution for all j and are independent.

(iii) **Variance**: $\text{var}(Y_j) = \sigma^2$ (constant regardless of the setting of $\boldsymbol{x}_j$).

As we have discussed through our examples, this approach has several deficiencies as a model for count, binary, or some positive continuous data:

- The normal distribution may not be a good probability model.
- Variance may not be constant across the range of the response.

- Because the response (and its mean) are restricted to be positive, a model that does not build this in may be inappropriate – in (11.5), there is nothing that says that estimates of the mean response **must** be positive everywhere – it could very well be that the estimated value of $\boldsymbol{\beta}$ could produce **negative** mean estimates for some covariate settings, even if ideally this is not possible for the problem at hand.

Models appropriate for the situations we have been discussing would have to address these issues.

GENERALIZATION: For responses that are not well represented by a normal distribution, it is not customary to write models in the form of (11.5) above, with an **additive** deviation.. This is because, for distributions like the Poisson, Bernoulli, or gamma, there is no analogue to the fact that if ϵ is normally distributed with mean 0, variance σ^2, then $Y = \mu + \epsilon$ is also normal with mean μ, variance σ^2.

It is thus standard to express regression models as we did in (i), (ii), and (iii) above – in terms of (i) an assumed model for the mean, (ii) an assumption about probability distribution, and (iii) an assumption about variance. As we have noted, for the Poisson, Bernoulli, and gamma distributions, the form of the distribution dictates the assumption about variance.

We now show how this modeling is done for the three situations on which we have focused. We will then highlight the common features. Because these models are more complex that usual linear regression models, special fitting techniques are required, and will be discussed in section 11.4.

COUNT DATA: For data in the form of counts, we have noted that a sensible probability model is the Poisson distribution. This model dictates that variance is equal to the mean; moreover, any sensible representation of the mean ought to be such that the mean is forced to be positive.

(i) **Mean**: For regression modeling, we wish to represent the mean for Y_j as a function of the covariates $\boldsymbol{x}_j$. However, this representation should ensure the mean can only be positive. A model that would accomplish this is

$$E(Y_j) = \exp(\beta_0 + \beta_1 x_{j1} + \cdots + \beta_k x_{jk}) = \exp(\boldsymbol{x}_j'\boldsymbol{\beta}). \tag{11.6}$$

In (11.6), the positivity requirement is enforced by writing the mean as the **exponential** of the **linear function** of $\boldsymbol{\beta}$ $\boldsymbol{x}_j'\boldsymbol{\beta}$. Note that the model implies

$$\log\{E(Y_j)\} = \beta_0 + \beta_1 x_{j1} + \cdots + \beta_k x_{jk} = \boldsymbol{x}_j'\boldsymbol{\beta};$$

i.e. the **logarithm** of the mean response is being modeled as a **linear function** of covariates and regression parameters. As a result, a model like (11.6) is often called a **loglinear model**.

Loglinear modeling is a standard technique for data in the form of counts, especially when the counts are **small**. When the counts are small, it is quite possible that using a **linear** model instead, $E(Y_j) = \boldsymbol{x}_j'\boldsymbol{\beta}$, would lead to an estimated value for $\boldsymbol{\beta}$ that would allow estimates of the mean to be **negative** for some covariate settings. This is less of a worry when the counts are very large. Consequently, loglinear modeling is most often employed for small count data.

It is important to note that a loglinear model for the mean response is not the **only** possibility for count data. However, it is the most common.

(ii) **Probability distribution**: The Y_j are assumed to arise at each setting $\boldsymbol{x}_j$ from a Poisson distribution with mean as in (11.6) and are assumed to be independent.

(iii) **Variance**: Under the Poisson assumption and the mean model (11.6), we have that the variance of Y_j is given by

$$\text{var}(Y_j) = E(Y_j) = \exp(\boldsymbol{x}_j'\boldsymbol{\beta}) \tag{11.7}$$

BINARY DATA: For binary data, the relevant probability model is the Bernoulli distribution. Here, the mean is also equal to the probability of seeing the event of interest; thus, the mean should be restricted to lie between 0 and 1. In addition, the model dictates that the variance of a response is a particular function of the mean.

(i) **Mean**: For regression modeling, we wish to represent the mean for Y_j as a function of the covariates $\boldsymbol{x}_j$ with the important restriction that this function always be between 0 and 1. A model that accomplishes this is

$$E(Y_j) = \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_j'\boldsymbol{\beta})}. \tag{11.8}$$

Note that, **regardless** of the value of the **linear combination** $\boldsymbol{x}_j'\boldsymbol{\beta}$, this function must **always** be less than 1. Similarly, the function must **always** be greater than 0. (Convince yourself).

It is an algebraic exercise to show that (try it!)

$$\log\left(\frac{E(Y_j)}{1 - E(Y_j)}\right) = \boldsymbol{x}_j'\boldsymbol{\beta}. \tag{11.9}$$

The function of $E(Y_j)$ on the left hand side of (11.9) is called the **logit** function. Recall that here $E(Y_j)$ is equal to the probability of seeing the event of interest. Thus, the function

$$\left(\frac{E(Y_j)}{1 - E(Y_j)}\right)$$

is the ratio of the probability of seeing the event of interest to the probability of **not** seeing it!

This ratio is often called the **odds** for this reason. Thus, the model (11.8) may be thought of as modeling the **log odds** as a **linear combination** of the covariates and regression parameters.

Model (11.8) is not the only model appropriate for representing the mean of a Bernoulli random variable; any function taking values only between 0 and 1 would do. Other such models are the **probit** and **complementary log-log** functions (see McCullagh and Nelder 1989, page 31). However, (11.8) is by far the most popular, and the model is usually referred to as the **logistic regression model** (for binary data).

(ii) **Probability distribution**: The Y_j are assumed to arise at each setting $\boldsymbol{x}_j$ from a Bernoulli distribution with mean as in (11.8) and are assumed to be independent.

(iii) **Variance**: For binary data, if the mean is represented by (11.8), then we must have that the variance of Y_j is given by

$$\text{var}(Y_j) = E(Y_j)\{1 - E(Y_j)\} = \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}\left(1 - \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}\right) \tag{11.10}$$

CONTINUOUS, POSITIVE DATA WITH CONSTANT COEFFICIENT OF VARIATION: For these data, there are a number of relevant probability models; we have discussed the **gamma** distribution. Here, the mean must be positive, and the variance must have the constant CV form.

(i) **Mean**: For regression modeling, we wish to represent the mean for Y_j as a function of the covariates $\boldsymbol{x}_j$ If the size of the responses is not too large, then using a linear model, $E(Y_j) = \boldsymbol{x}_j'\boldsymbol{\beta}$ could be dangerous; thus, it is preferred to use a model that enforces positivity. One common model is the **loglinear model** (11.6), which is also commonly used for count data. Both types of data share the requirement of positivity, so this is not surprising.

When the size of the response is larger, it is often the case that the positivity requirement is not a big concern – even if a **linear model** is used to represent the data, because the responses are all so big, estimated means will still all be positive for covariate settings like those of the original data. This opens up the possibility for other models for the mean.

With a **single covariate** ($k = 1$), linear models are seldom used – here, the linear model would be a **straight line**. This is because it is fairly typical that, for phenomena where constant coefficient of variation occurs, the relationship between response and covariate seldom looks like a straight line; rather it tends to look more like that in the upper left panel of Figure 3.

Note that in the lower left panel of Figure 3, once the response is placed on the **log** scale, the relationship looks much more like a straight line. This suggests that a model like

$$\log\{E(Y_j)\} = \beta_0 + \beta_1 x_j,$$

where x_j = log percent concentration, might be reasonable; that is, log of response is a straight line in x_j. This is exactly the loglinear model (11.6) in the special case $k = 1$, of course.

However, note that in the upper right panel, once the response is **inverted** by taking the **reciprocal** (so plotting $1/Y_j$ on the vertical axis), the relationship looks even more like a straight line. This observation indicates that a model like

$$\frac{1}{E(Y_j)} = \beta_0 + \beta_1 x_j$$

might be appropriate.

More generally, for k covariates, this suggests the model

$$E(Y_j) = \frac{1}{\boldsymbol{x}_j'\boldsymbol{\beta}}. \tag{11.11}$$

This model does **not** preserve the positivity requirement; however, for situations where this is not really a concern, the **inverse** or **reciprocal** model (11.11) often gives a better representation than does a plain linear model for $E(Y_j)$, as was the case for the clotting time data.

(ii) **Probability distribution**: The Y_j are assumed to arise at each setting $\boldsymbol{x}_j$ from a gamma distribution with mean as in (11.6), (11.11), or some other model deemed appropriate. The Y_j are also assumed to be independent.

(iii) **Variance**: Under the gamma assumption, the variance of Y_j is proportional to the square of the mean response; i.e. constant coefficient of variation. Thus, if the mean is represented by (11.6), then we must have that the variance of Y_j is given by

$$\text{var}(Y_j) = \sigma^2 E(Y_j)^2 = \sigma^2\{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})\}^2. \tag{11.12}$$

If the mean is represented by (11.11), then we must have that

$$\text{var}(Y_j) = \sigma^2 E(Y_j)^2 = \sigma^2 \left(\frac{1}{\boldsymbol{x}_j'\boldsymbol{\beta}}\right)^2. \tag{11.13}$$

IN GENERAL: All of the regression models we have discussed share the features that

- Appropriate models for **mean response** are of the form

$$E(Y_j) = f(\boldsymbol{x}_j'\boldsymbol{\beta}), \tag{11.14}$$

  where $f(\boldsymbol{x}_j'\boldsymbol{\beta})$ is a suitable function of a **linear combination** of the covariates $\boldsymbol{x}_j$ and regression parameter $\boldsymbol{\beta}$.

- The **variance** of Y_j may be represented as a function of the form

$$\text{var}(Y_j) = \phi V\{ E(Y_j) \} = \phi V\{ f(\boldsymbol{x}_j'\boldsymbol{\beta}) \}, \tag{11.15}$$

  where V is a function of the **mean response** and ϕ is a constant usually assumed to be the same for all j. For the Poisson and Bernoulli cases, $\phi = 1$; for the gamma case, $\phi = \sigma^2$.

SCALED EXPONENTIAL FAMILY: It turns out that these regression models share even **more**. It was long ago recognized that certain probability distributions all fall into a **general class**. For distributions in this class, if the mean is equal to μ, then the variance **must be** a specific function $\phi V(\mu)$ of μ. Distributions in this class include:

- The **normal** distribution with mean μ, variance σ^2 (not related to μ in any way, so a function of μ that is the same for all μ).
- The **Poisson** distribution with mean μ, variance μ.
- The **gamma** distribution with mean μ, variance $\sigma^2\mu^2$.
- The **Bernoulli** distribution with mean μ, variance $\mu(1-\mu)$.

The class includes other distributions we have not discussed as well. This class of distributions is known as the **scaled exponential family**. As we will discuss in section 11.4, because these distributions share so much, fitting regression models under them may be accomplished by the **same** method.

GENERALIZED LINEAR MODELS: We are now in a position to state all of this more formally. A **generalized linear model** is a regression model for response Y_j with the following features:

- The mean of Y_j is assumed to be of the form (11.14)

$$E(Y_j) = f(\boldsymbol{x}_j'\boldsymbol{\beta}).$$

  It is customary to express this a bit differently, however. The function f is almost always chosen to be **monotone**; that is, it is a **strictly increasing** or **decreasing** function of $\boldsymbol{x}_j'\boldsymbol{\beta}$. This means that there is a **unique** function g, say, called the **inverse** function of f, such that we may re-express (11.14) model in the form

$$g\{E(Y_j)\} = \boldsymbol{x}_j'\boldsymbol{\beta}.$$

  For example, for binary data, we considered the logistic function (11.8); i.e.

$$E(Y_j) = f(\boldsymbol{x}_j'\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_j'\boldsymbol{\beta})}.$$

  This may be rewritten in the form (11.9),

$$\log\left(\frac{E(Y_j)}{1-E(Y_j)}\right) = g\{E(Y_j)\} = \boldsymbol{x}_j'\boldsymbol{\beta}.$$

  The function g is called the **link function**, because it "links" the mean and the covariates. The linear combination of covariates and regression parameters $\boldsymbol{x}_j'\boldsymbol{\beta}$ is called the **linear predictor**. Certain choices of f, and hence of link function g, are popular for different kinds of data, as we have noted.

- The probability distribution governing Y_j is assumed to be one of those from the **scaled exponential family** class.

- The variance of Y_j is thus assumed to be of the form dictated by the distribution:

$$\text{var}(Y_j) = \phi V\{\, E(Y_j)\,\},$$

  where the function V depends on the distribution and ϕ might be equal to a known constant. The function V is referred to as the **variance function** for obvious reasons. The parameter ϕ is often called the **dispersion parameter** because it has to do with variance. It may be known, as for the Poisson or Bernoulli distributions, or unknown and estimated, which is the case for the gamma.

The models we have discussed for count, binary, and positive continuous data are thus all generalized linear models. In fact, the **classical** linear regression model assuming normality with constant variance is also a generalized linear model!

11.4 Maximum likelihood and iteratively reweighted least squares

The class of generalized linear models may be thought of as extending the usual classical linear model to handle special features of different kinds of data. The extension introduces some complications, however. In particular:

- The model for mean response need no longer be a **linear** model.
- The variance is allowed to **depend** on the mean; thus, the variance depends on the **regression parameter** $\boldsymbol{\beta}$.

The result of these more complex features is that it is no longer quite so straightforward to estimate $\boldsymbol{\beta}$ (and ϕ, if required). To appreciate this, we first review the method of least squares for the normal, linear, constant variance model.

LINEAR MODEL AND MAXIMUM LIKELIHOOD: For the **linear model** with constant variance σ^2 and normality, the usual method of **least squares** involves minimizing in $\boldsymbol{\beta}$ the **distance** criterion

$$\sum_{j=1}^{n}(y_j - \boldsymbol{x}_j'\boldsymbol{\beta})^2, \tag{11.16}$$

where $y_1, \ldots, y_n$ are observed data. This approach has another motivation – the estimator of $\boldsymbol{\beta}$ obtained in this way is the **maximum likelihood estimator**. In particular, write the observed data as $\boldsymbol{y} = (y_1, \ldots, y_n)'$. Because the Y_j are assumed independent, the **joint density** of all the data (that is, the joint density of $\boldsymbol{Y}$), is just the product of the n individual normal densities:

$$f(\boldsymbol{y}) = \prod_{j=1}^{n}(2\pi)^{-1/2}\sigma^{-1}\exp\{-(y_j - \boldsymbol{x}_j'\boldsymbol{\beta})^2/(2\sigma^2)\}.$$

It is easy to see that the only place that $\boldsymbol{\beta}$ appears is in the exponent; thus, if we wish to maximize the likelihood $f(\boldsymbol{y})$, we must maximize the exponent. Note that the **smaller** $(Y_j - \boldsymbol{x}_j'\boldsymbol{\beta})^2$ gets, the **larger** the exponent gets (because of the negative sign). Thus, to **maximize** the likelihood, we wish to **minimize** (11.16), which corresponds **exactly** to the method of **least squares**!

- Thus, obtaining the least squares estimator in a linear regression model under the normality and constant variance assumptions is the same as finding the maximum likelihood estimator.
- In this case, minimizing (11.16) may be done **analytically**; that is, we can write down an **explicit** expression for the estimator (as a function of the random vector $\boldsymbol{Y}$):

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

where $\boldsymbol{X}$ is the usual design matrix.

- This follows from calculus – the minimizing value of (11.16) is found by setting the first derivative of the equation to 0 and solving for $\boldsymbol{\beta}$. That is, the least squares (ML) estimator solves the set of p equations

$$\sum_{j=1}^{n}(Y_j - \boldsymbol{x}_j'\boldsymbol{\beta})\boldsymbol{x}_j = \mathbf{0}. \tag{11.17}$$

- Note that the the estimator and the equation it solves are **linear** functions of the data Y_j.

GENERALIZED LINEAR MODELS AND MAXIMUM LIKELIHOOD: A natural approach to estimating $\boldsymbol{\beta}$ in all generalized linear models is thus to appeal to the principle of maximum likelihood. It is beyond the scope of our discussion to give a detailed treatment of this. We simply remark that it turns out that, fortuitously, the form of the joint density of random variables $Y_1, \ldots, Y_n$ that arise from **any** of the distributions in the scaled exponential family class has the same general form. Thus, it turns out that the ML estimator for $\boldsymbol{\beta}$ in **any** generalized linear model solves a set of p equations of the **same** general form:

$$\sum_{j=1}^{n}\frac{1}{V\{f(\boldsymbol{x}_j'\boldsymbol{\beta})\}}\{Y_j - f(\boldsymbol{x}_j'\boldsymbol{\beta})\}f'(\boldsymbol{x}_j'\boldsymbol{\beta})\boldsymbol{x}_j = \mathbf{0}, \tag{11.18}$$

where $f'(u) = \dfrac{d}{du}f(u)$, the derivative of f with respect to its argument.

The equation (11.18) and the equation for the linear, normal, constant variance model (11.17) share the feature that they are both **linear** functions of the data Y_j and are equations we would like to solve in order to obtain the maximum likelihood estimator for $\boldsymbol{\beta}$. Thus, they are very similar in **spirit**. However, they differ in several ways:

- Each **deviation** $\{Y_j - f(\boldsymbol{x}_j'\boldsymbol{\beta})\}$ in (11.18) is **weighted** in accordance with its **variance** (the scale parameter ϕ is a constant). Of course, so is each deviation in (11.17); however, in that case, the variance is **constant** for all j. Recall that **weighting** in accordance with variance is a sensible principle, so it is satisfying to see that, despite the difference in probability distributions, this principle is still followed. Here, the variance function depends on $\boldsymbol{\beta}$, so now the weighting **depends** on $\boldsymbol{\beta}$! Thus, $\boldsymbol{\beta}$ appears in this equation in a very complicated way.

- Moreover, $\boldsymbol{\beta}$ also appears in the function f, which can be quite complicated – the function f is certainly not a **linear** function of $\boldsymbol{\beta}$!

The result of these differences is that, while it **is** possible to solve (11.17) **explicitly**, it **is not** possible to do the same for (11.18). Rather, the solution to (11.18) must be found using a numerical algorithm.

The numerical algorithm is straightforward and works well in practice, so this is not an enormous drawback.

ITERATIVELY REWEIGHTED LEAST SQUARES: It turns out that there is a standard algorithm that is applicable for solving equations of the form (11.18); discussion of the details is beyond our scope. The basic idea is (operating on the observed data)

- Given a **starting value**, or guess, for $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(0)}$, say, evaluate the **weights** at $\boldsymbol{\beta}^{(0)}$: $1/V\{f(\boldsymbol{x}_j, \boldsymbol{\beta}^{(0)})\}$.
- Pretending the weights are **fixed constants** not depending on $\boldsymbol{\beta}$, solve equation (11.18). This still requires a numerical technique, but may be accomplished by something that is **approximately** like solving (11.17). This gives a new guess for $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(1)}$, say.
- Evaluate the weights at $\boldsymbol{\beta}^{(1)}$. and repeat. Continue updating until two successive $\boldsymbol{\beta}$ values are the same.

The repeatedly updating of the weights along with the approximation to solve an equation like (11.17) gives this procedure its name: **iteratively reweighted least squares**, often abbreviated as IRWLS or IWLS.

Luckily, there are standard ways to find the **starting value** based on the data and knowledge of the assumed probability distribution. Thus, the user need not be concerned with this (usually); software typically generates this value automatically.

SAMPLING DISTRIBUTION: It should come as no surprise that the **sampling distribution** of the estimator $\widehat{\boldsymbol{\beta}}$ solving (11.18) **cannot** be derived in **closed form**. Rather, it is necessary to resort to **large sample theory** approximation. Here, "large sample" refers to the sample size, n (number of independent observations). This is sensible – each Y_j is typically from a different unit.

We now state the large sample result. For n "large," the IRWLS/ML estimator satisfies

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}\{\boldsymbol{\beta}, \phi(\boldsymbol{\Delta}'\boldsymbol{V}^{-1}\boldsymbol{\Delta})^{-1}\}. \tag{11.19}$$

Here,

- $\boldsymbol{\Delta}$ is a $(n \times p)$ matrix whose (j, s) element ($j = 1, \ldots, n$, $s = 1, \ldots, p$) is the derivative of $f(\boldsymbol{x}_j'\boldsymbol{\beta})$ with respect to the sth element of $\boldsymbol{\beta}$.
- $\boldsymbol{V}$ is the $(n \times n)$ **diagonal** matrix with diagonal elements $V\{f(\boldsymbol{x}_j'\boldsymbol{\beta})\}$.

A little thought about the form of $\boldsymbol{\Delta}$ and $\boldsymbol{V}$ reveals that both **depend on** $\boldsymbol{\beta}$. However, $\boldsymbol{\beta}$ is **unknown** and has been **estimated**. In addition, if ϕ is not dictated to be equal to a specific constant (e.g. $\phi = 1$ if Y_j are Poisson or Bernoulli but is unknown if Y_j is gamma), then it, too, must be estimated. In this situation, the standard estimator for ϕ is

$$\widehat{\phi} = (n-p)^{-1} \sum_{j=1}^{n} \frac{\{Y_j - f(\boldsymbol{x}_j'\widehat{\boldsymbol{\beta}})\}^2}{V\{f(\boldsymbol{x}_j'\widehat{\boldsymbol{\beta}})\}}.$$

In the context of fitting generalized linear models, this is often referred to as the **Pearson chi-square** (divided by its degrees of freedom). Other methods are also available; we use this method for illustration in the examples of section 11.6.

Thus, it is customary to approximate (11.19) by replacing $\boldsymbol{\beta}$ and ϕ by estimates wherever they appear. **Standard errors** for the elements of $\widehat{\boldsymbol{\beta}}$ are then found as the square roots of the diagonal elements of the matrix

$$\widehat{\boldsymbol{V}}_\beta = \widehat{\phi}(\widehat{\boldsymbol{\Delta}}'\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{\Delta}})^{-1},$$

where the "hats" mean that $\boldsymbol{\beta}$ and ϕ are replaced by estimates. We use the same notation, $\widehat{\boldsymbol{V}}_\beta$, as in previous chapters to denote the estimated covariance matrix; the definition of $\widehat{\boldsymbol{V}}_\beta$ should be clear from the context.

HYPOTHESIS TESTS: It is common to use **Wald** testing procedures to test hypotheses about $\boldsymbol{\beta}$. Specifically, for null hypotheses of the form

$$H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{h},$$

we may approximate the sampling distribution of the estimate $\boldsymbol{L}\widehat{\boldsymbol{\beta}}$ by

$$\boldsymbol{L}\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}(\boldsymbol{L}\boldsymbol{\beta}, \boldsymbol{L}\widehat{\boldsymbol{V}}_\beta\boldsymbol{L}').$$

Construction of test statistics and confidence intervals is then carried out in a fashion identical to that discussed in previous chapters. For example, if $\boldsymbol{L}$ is a row vector, then one may form the "z-statistic"

$$z = \frac{\boldsymbol{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{h}}{SE(\boldsymbol{L}\widehat{\boldsymbol{\beta}})}.$$

More generally, the Wald χ^2 test statistic would be

$$(\boldsymbol{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{h})'(\boldsymbol{L}\widehat{\boldsymbol{V}}_\beta\boldsymbol{L}')^{-1}(\boldsymbol{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{h})$$

(of course $= z^2$ in the case $\boldsymbol{L}$ has a single row).

REMARK: Note that all of this looks very similar to what is done in classical, linear regression under the assumption of constant variance and normality. The obvious difference is that the results are now just **large sample** approximations rather than exact, but the form and spirit are the same.

11.5 Discussion

Generalized linear models may be regarded as an extension of classical linear regression when the usual assumptions of normality and constant variance do not apply. Because of the additional considerations imposed by the nature of the data, sensible models for mean response may no longer be **linear functions** of covariates and regression parameters directly. Rather, the mean response is modeled as a **function** (**non**linear) of a linear combination of covariates and regression parameters (the **linear predictor**). Although the models and fitting methods become more complicated as a result, the spirit is the same.

11.6 Implementation with SAS

We illustrate how to carry out fitting of generalized linear models for the three examples discussed in this section:

1. The horsekick data
2. The myocardial infarction data
3. The clotting times data

As our main objective is to gain some familiarity with these models in order to appreciate their extension to the case of longitudinal data from m units, we do not perform detailed, comprehensive analyses involving many questions of scientific interest. Rather, we focus mainly on how to specify models using SAS `PROC GENMOD` and how to interpret the output. In the next chapter, we will use `PROC GENMOD` with the `REPEATED` statement to fit longitudinal data.

EXAMPLE 1 – HORSEKICK DATA: Recall that it was reasonable to model these data using the Poisson distribution assumption. Define Y_j to be the jth observations of number of horsekick deaths suffered corresponding to a particular corps and year denoted by dummy variables

$$\begin{aligned} x_{jk} &= 1 \text{ if observation } j \text{ is from year } k = 1875, \ldots, 1893 \\ &= 0 \text{ otherwise} \\ z_{jk} &= 1 \text{ if observation } j \text{ is from corps } k = 1, \ldots, 9 \\ &= 0 \text{ otherwise} \end{aligned}$$

We thus consider the loglinear model

$$E(Y_j) = \exp(\beta_0 + \beta_1 x_{j1} + \cdots + \beta_{19} x_{j,19} + \beta_{20} z_{j1} + \cdots + \beta_{28} z_{j9}) \tag{11.20}$$

for the mean response. This model represents the mean number of horse kicks as an exponential function; for example, for j corresponding to 1894 and corps 10,

$$E(Y_j) = \exp(\beta_0);$$

for j corresponding to 1875 and corps 1,

$$E(Y_j) = \exp(\beta_0 + \beta_1 + \beta_{20}).$$

An obvious question of interest would be to determine whether some of the regression parameters are different from 0, indicating that the particular year or corps to which they correspond does not differ from the final year and corps (1894, corps 10). This may be addressed by inspecting the Wald test statistics corresponding to each element of $\boldsymbol{\beta}$. To address the issue of how specific years compared, averaged across corps, one would be interested in whether the appropriate differences in elements of $\boldsymbol{\beta}$ were equal to zero. For example, if we were interested in whether 1875 and 1880 were different, we would be interested in the difference $\beta_1 - \beta_6$.

PROGRAM:

```
/*****************************************************************

  CHAPTER 11, EXAMPLE 1

  Fit a loglinear regression model to the horse-kick data.
  (Poisson assumption)

*****************************************************************/

options ls=80 ps=59 nodate; run;

/*****************************************************************

  The data look like (first 6 records)

1875   0   0   0   0   1   1   0   0   1   0
1876   0   0   1   0   0   0   0   0   1   1
1877   0   0   0   0   1   0   0   1   2   0
1878   2   1   1   0   0   0   0   1   1   0
1879   0   1   1   2   0   1   0   0   1   0
1880   2   1   1   1   0   0   2   1   3   0
          .
          .
          .

  column 1      year
  columns 2-11  number of fatal horsekicks suffered by corps 1-10.

*****************************************************************/

data kicks; infile 'kicks.dat';
  input year c1-c10;
run;

/*****************************************************************

  Reconfigure the data so that the a single number of kicks
  for a particular year/corps combination appears on a separate
  line.

*****************************************************************/

data kicks2; set kicks;
  array c{10} c1-c10;
  do corps=1 to 10;
   kicks = c{corps};
   output;
  end;
  drop c1-c10;
run;

proc print data=kicks2 ; run;

/*****************************************************************

  Fit the loglinear regression model using PROC GENMOD.  Here,
  the dispersion parameter phi=1, so is not estimated.  We let SAS
  form the dummy variables through use of the CLASS statement.
  This results in the model for mean response being parameterized
  as in equation (11.20).

  The DIST=POISSON option in the model statement specifies
  that the Poisson probability distribution assumption, with its
  requirement that mean = variance, be used.  The LINK=LOG option
  asks for the loglinear model.  Other LINK= choices are available.

  We also use a CONTRAST statement to investigate whether there is
  evidence to suggest that 1875 differed from 1880 in terms of numbers
  of horsekick deaths. The WALD option asks that the usual large sample
  chi-square test statistic be used as the basis for the test.

 *****************************************************************/

proc genmod data=kicks2;
  class year corps;
  model kicks = year corps /  dist = poisson link = log;
  contrast '1875-1880' year 1 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 / wald;
run;
```

OUTPUT: Following the output, we comment on a few aspects of the output.

The SAS System 1

Obs	year	corps	kicks
1	1875	1	0
2	1875	2	0
3	1875	3	0
4	1875	4	0
5	1875	5	1
6	1875	6	1
7	1875	7	0
8	1875	8	0
9	1875	9	1
10	1875	10	0
11	1876	1	0
12	1876	2	0
13	1876	3	1
14	1876	4	0
15	1876	5	0
16	1876	6	0
17	1876	7	0
18	1876	8	0
19	1876	9	1
20	1876	10	1
21	1877	1	0
22	1877	2	0
23	1877	3	0
24	1877	4	0
25	1877	5	1
26	1877	6	0
27	1877	7	0
28	1877	8	1
29	1877	9	2
30	1877	10	0
31	1878	1	2
32	1878	2	1
33	1878	3	1
34	1878	4	0
35	1878	5	0
36	1878	6	0
37	1878	7	0
38	1878	8	1
39	1878	9	1
40	1878	10	0
41	1879	1	0
42	1879	2	1
43	1879	3	1
44	1879	4	2
45	1879	5	0
46	1879	6	1
47	1879	7	0
48	1879	8	0
49	1879	9	1
50	1879	10	0
51	1880	1	2
52	1880	2	1
53	1880	3	1
54	1880	4	1
55	1880	5	0

The SAS System 2

Obs	year	corps	kicks
56	1880	6	0
57	1880	7	2
58	1880	8	1
59	1880	9	3
60	1880	10	0
61	1881	1	0
62	1881	2	2
63	1881	3	1
64	1881	4	0
65	1881	5	1
66	1881	6	0
67	1881	7	1
68	1881	8	0
69	1881	9	0
70	1881	10	0
71	1882	1	0
72	1882	2	0
73	1882	3	0
74	1882	4	0
75	1882	5	0
76	1882	6	1
77	1882	7	1
78	1882	8	2
79	1882	9	4
80	1882	10	1

81	1883	1	1
82	1883	2	2
83	1883	3	0
84	1883	4	1
85	1883	5	1
86	1883	6	0
87	1883	7	1
88	1883	8	0
89	1883	9	0
90	1883	10	0
91	1884	1	1
92	1884	2	0
93	1884	3	0
94	1884	4	0
95	1884	5	1
96	1884	6	0
97	1884	7	0
98	1884	8	2
99	1884	9	1
100	1884	10	1
101	1885	1	0
102	1885	2	0
103	1885	3	0
104	1885	4	0
105	1885	5	0
106	1885	6	0
107	1885	7	2
108	1885	8	0
109	1885	9	0
110	1885	10	1

Obs	year	corps	kicks
111	1886	1	0
112	1886	2	0
113	1886	3	1
114	1886	4	1
115	1886	5	0
116	1886	6	0
117	1886	7	1
118	1886	8	0
119	1886	9	3
120	1886	10	0
121	1887	1	2
122	1887	2	1
123	1887	3	0
124	1887	4	0
125	1887	5	2
126	1887	6	1
127	1887	7	1
128	1887	8	0
129	1887	9	2
130	1887	10	0
131	1888	1	1
132	1888	2	0
133	1888	3	0
134	1888	4	1
135	1888	5	0
136	1888	6	0
137	1888	7	0
138	1888	8	0
139	1888	9	1
140	1888	10	0
141	1889	1	1
142	1889	2	1
143	1889	3	0
144	1889	4	1
145	1889	5	0
146	1889	6	0
147	1889	7	1
148	1889	8	2
149	1889	9	0
150	1889	10	2
151	1890	1	0
152	1890	2	2
153	1890	3	0
154	1890	4	1
155	1890	5	2
156	1890	6	0
157	1890	7	2
158	1890	8	1
159	1890	9	2
160	1890	10	2
161	1891	1	0
162	1891	2	1
163	1891	3	1
164	1891	4	1
165	1891	5	1

The SAS System 4

Obs	year	corps	kicks
166	1891	6	1
167	1891	7	0
168	1891	8	3
169	1891	9	1
170	1891	10	0
171	1892	1	2
172	1892	2	0
173	1892	3	1
174	1892	4	1
175	1892	5	0
176	1892	6	1
177	1892	7	1
178	1892	8	0
179	1892	9	1
180	1892	10	0
181	1893	1	0
182	1893	2	0
183	1893	3	0
184	1893	4	1
185	1893	5	2
186	1893	6	0
187	1893	7	0
188	1893	8	1
189	1893	9	0
190	1893	10	0
191	1894	1	0
192	1894	2	0
193	1894	3	0
194	1894	4	0
195	1894	5	0
196	1894	6	1
197	1894	7	0
198	1894	8	1
199	1894	9	0
200	1894	10	0

The SAS System 5

The GENMOD Procedure

Model Information

Data Set	WORK.KICKS2
Distribution	Poisson
Link Function	Log
Dependent Variable	kicks

Number of Observations Read	200
Number of Observations Used	200

Class Level Information

Class	Levels	Values
year	20	1875 1876 1877 1878 1879 1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894
corps	10	1 2 3 4 5 6 7 8 9 10

Parameter Information

Parameter	Effect	year	corps
Prm1	Intercept		
Prm2	year	1875	
Prm3	year	1876	
Prm4	year	1877	
Prm5	year	1878	
Prm6	year	1879	
Prm7	year	1880	
Prm8	year	1881	
Prm9	year	1882	
Prm10	year	1883	
Prm11	year	1884	
Prm12	year	1885	
Prm13	year	1886	
Prm14	year	1887	
Prm15	year	1888	
Prm16	year	1889	
Prm17	year	1890	
Prm18	year	1891	
Prm19	year	1892	
Prm20	year	1893	
Prm21	year	1894	
Prm22	corps		1
Prm23	corps		2
Prm24	corps		3
Prm25	corps		4
Prm26	corps		5

Prm27	corps	6
Prm28	corps	7
Prm29	corps	8
Prm30	corps	9

The SAS System 6
The GENMOD Procedure

Parameter Information

Parameter	Effect	year	corps
Prm31	corps		10

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	171	171.6395	1.0037
Scaled Deviance	171	171.6395	1.0037
Pearson Chi-Square	171	160.6793	0.9396
Scaled Pearson X2	171	160.6793	0.9396
Log Likelihood		-161.8886	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-2.0314	0.7854	-3.5707	-0.4921	6.69	0.0097
year	1875	1	0.4055	0.9129	-1.3837	2.1947	0.20	0.6569
year	1876	1	0.4055	0.9129	-1.3837	2.1947	0.20	0.6569
year	1877	1	0.6931	0.8660	-1.0042	2.3905	0.64	0.4235
year	1878	1	1.0986	0.8165	-0.5017	2.6989	1.81	0.1785
year	1879	1	1.0986	0.8165	-0.5017	2.6989	1.81	0.1785
year	1880	1	1.7047	0.7687	0.1981	3.2114	4.92	0.0266
year	1881	1	0.9163	0.8367	-0.7235	2.5561	1.20	0.2734
year	1882	1	1.5041	0.7817	-0.0281	3.0363	3.70	0.0544
year	1883	1	1.0986	0.8165	-0.5017	2.6989	1.81	0.1785
year	1884	1	1.0986	0.8165	-0.5017	2.6989	1.81	0.1785
year	1885	1	0.4055	0.9129	-1.3837	2.1947	0.20	0.6569
year	1886	1	1.0986	0.8165	-0.5017	2.6989	1.81	0.1785
year	1887	1	1.5041	0.7817	-0.0281	3.0363	3.70	0.0544
year	1888	1	0.4055	0.9129	-1.3837	2.1947	0.20	0.6569
year	1889	1	1.3863	0.7906	-0.1632	2.9358	3.07	0.0795
year	1890	1	1.7918	0.7638	0.2948	3.2887	5.50	0.0190
year	1891	1	1.5041	0.7817	-0.0281	3.0363	3.70	0.0544
year	1892	1	1.2528	0.8018	-0.3187	2.8242	2.44	0.1182
year	1893	1	0.6931	0.8660	-1.0042	2.3905	0.64	0.4235
year	1894	0	0.0000	0.0000	0.0000	0.0000	.	.
corps	1	1	0.4055	0.4564	-0.4891	1.3001	0.79	0.3744
corps	2	1	0.4055	0.4564	-0.4891	1.3001	0.79	0.3744
corps	3	1	-0.0000	0.5000	-0.9800	0.9800	0.00	1.0000
corps	4	1	0.3185	0.4647	-0.5923	1.2292	0.47	0.4931
corps	5	1	0.4055	0.4564	-0.4891	1.3001	0.79	0.3744
corps	6	1	-0.1335	0.5175	-1.1479	0.8808	0.07	0.7964
corps	7	1	0.4855	0.4494	-0.3952	1.3662	1.17	0.2799
corps	8	1	0.6286	0.4378	-0.2295	1.4867	2.06	0.1510

The SAS System 7
The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
corps	9	1	1.0986	0.4082	0.2985	1.8988	7.24	0.0071
corps	10	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Contrast Results

Contrast	DF	Chi-Square	Pr > ChiSq	Type
1875-1880	1	3.98	0.0461	Wald

INTERPRETATION:

- Pages 1–4 of the output show the reconfigured data set.
- The results of running `PROC GENMOD` appear on pages 5–7 of the output. On page 6, the results of the fit by IRWLS/ML are displayed. The table `Analysis of Parameter Estimates` contains the estimates of the parameters β_0 – β_{28}, along with their estimated standard errors (square roots of the elements of $\widehat{\boldsymbol{V}}_\beta$). The column `Chi-Square` gives the value of the Wald test statistic for testing whether the parameter in that row is equal to zero.
- The row `SCALE` corresponds to ϕ; here, for the Poisson distribution, $\phi = 1$, so nothing is estimated. This is noted at the bottom of page 6 (`The scale parameter was held fixed.`).
- Page 7 shows the result of the `contrast` statement to address the null hypothesis that there was no difference in mean horsekick deaths in 1875 and 1880 (see the program). The Wald test statistic is 3.98 with an asociated p-value of 0.046, suggesting that there is some evidence to support a difference. Note that if β_1 and β_6 are different, then the mean responses for 1875 and 1880 must be different for any corps. However, note that the difference $\beta_1 - \beta_6$ does **not** correspond to the actual difference in mean response. Inspection of the estimates of β_1 and β_6 on page 6 shows $\widehat{\beta}_1 = 0.4055$ and $\widehat{\beta}_6 = 1.7047$. This suggests that the mean response for 1880, which depends on $\exp(\beta_6)$, is larger than that for 1875, which depends on $\exp(\beta_1)$.

EXAMPLE 2 – MYOCARDIAL INFARCTION DATA: Here, the response (whether or not a woman has suffered a myocardial infarction) is **binary**, so we wish to fit a generalized linear model assuming the Bernoulli distribution. The mean function must honor the restriction of being between 0 and 1; here, we fit the **logistic regression** model, using the **logit** link.

Recall that we defined

$$\begin{aligned} x_{j1} &= 1 \text{ if oral contraceptive use} \\ &= 0 \text{ otherwise} \\ x_{j2} &= \text{ age in years} \\ x_{j3} &= 1 \text{ if smoke more then one pack/day} \\ &= 0 \text{ otherwise} \end{aligned}$$

Thus, we model the mean response, equivalently, the probability of suffering a heart attack, as

$$E(Y_j) = \frac{\exp(\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \beta_3 x_{j3})}{1 + \exp(\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \beta_3 x_{j3})}. \tag{11.21}$$

Interest focuses on whether or not β_1, β_2, and β_3. corresponding to the association of oral contraceptive use, age, and smoking, respectively, with probability of myocardial infarction, are different from zero.

If β_1 is different from zero, for example, the interpretation is that oral contraceptive use does change the probability of suffering a heart attack. We say more about this shortly.

PROGRAM:

```
/*****************************************************************

  CHAPTER 11, EXAMPLE 2

  Fit a logistic regression model to the myocardial infarction
  data.

*****************************************************************/

options ls=80 ps=59 nodate; run;

/*****************************************************************

  The data look like (first 10 records)

     1 1 33 1 0
     2 0 32 0 0
     3 1 37 0 1
     4 0 36 0 0
     5 1 50 1 1
     6 1 40 0 0
     7 0 35 0 0
     8 1 33 0 0
     9 1 33 0 0
     10 0 31 0 0
          .
          .
          .

  column 1      subject id
  column 2      oral contraceptive indicator (0=no,1=yes)
  column 3      age (years)
  column 4      smoking indicator (0=no,1=yes)
  column 5      binary response -- whether MI has been suffered
                (0=no,1=yes)

*****************************************************************/

data mi; infile 'infarc.dat';
  input id oral age smoke mi;
run;

/*****************************************************************

  Fit the logistic regression model using PROC GENMOD.
  We do not use a CLASS statement here, as the covariates are
  either continuous (AGE) or already in "dummy" form (ORAL, SMOKE).
  The model statement with the LINK=LOGIT option results in the
  logistic regression model in equation (10.21).  The DIST=BINOMIAL
  specifies the Bernoulli distribution, which is the simplest case
  of a binomial distribution.

  In versions 7 and higher of SAS, PROC GENMOD will model by
  default the probability that the response y=0 rather than
  the conventional y=1!  To make PROC GENMOD model probability
  y=1, as is standard, one must include the DESCENDING option in
  the PROC GENMOD statement.  In earlier versions of SAS, the
  probability y=1 is modeled by default, as would be expected.

  If the user is unsure which probability is being modeled, one
  can check the .log file.  In later versions of SAS, an explicit
  statement about what is being modeled will appear. PROC GENMOD
  output should also contain a statement about what is being
  modeled.

*****************************************************************/

proc genmod data=mi descending;
  model mi = oral age smoke / dist = binomial link = logit;
run;
```

OUTPUT: Following the output, we comment on a few aspects of the output.

```
                              The SAS System                              1

                           The GENMOD Procedure

                            Model Information

                        Data Set               WORK.MI
                        Distribution          Binomial
                        Link Function            Logit
                        Dependent Variable          mi

                     Number of Observations Read         200
                     Number of Observations Used         200
                     Number of Events                     43
                     Number of Trials                    200

                          Response Profile

                       Ordered              Total
                         Value    mi    Frequency

                             1    1           43
                             2    0          157

PROC GENMOD is modeling the probability that mi='1'.

                          Parameter Information

                       Parameter       Effect

                       Prm1            Intercept
                       Prm2            oral
                       Prm3            age
                       Prm4            smoke

                  Criteria For Assessing Goodness Of Fit

          Criterion                 DF           Value        Value/DF

          Deviance                 196        150.3748          0.7672
          Scaled Deviance          196        150.3748          0.7672
          Pearson Chi-Square       196        177.5430          0.9058
          Scaled Pearson X2        196        177.5430          0.9058
          Log Likelihood                      -75.1874

  Algorithm converged.

                              The SAS System                              2

                           The GENMOD Procedure

                    Analysis Of Parameter Estimates

                             Standard       Wald 95%          Chi-
   Parameter  DF  Estimate     Error   Confidence Limits    Square  Pr > ChiSq

   Intercept   1   -9.1140    1.7571  -12.5579   -5.6702    26.90      <.0001
   oral        1    1.9799    0.4697    1.0593    2.9005    17.77      <.0001
   age         1    0.1626    0.0445    0.0753    0.2498    13.32      0.0003
   smoke       1    1.8122    0.4294    0.9706    2.6538    17.81      <.0001
   Scale       0    1.0000    0.0000    1.0000    1.0000

NOTE: The scale parameter was held fixed.

                          Contrast Estimate Results

                                   Standard                                 Chi-
Label                    Estimate     Error   Alpha   Confidence Limits  Square

smk log odds ratio         1.8122    0.4294    0.05    0.9706    2.6538   17.81
Exp(smk log odds ratio)    6.1241    2.6297    0.05    2.6396   14.2084

                          Contrast Estimate Results

                     Label                    Pr > ChiSq

                     smk log odds ratio           <.0001
                     Exp(smk log odds ratio)
```

INTERPRETATION:

- From the output, the Wald test statistics in the `Chi-Square` column of the table `Analysis Of`

`Parameter Estimates` of whether $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$ are all large, with very small p-values. This suggests that there is strong evidence that oral contraceptive use, age, and smoking affects the probability of having a heart attack.

- In each case, note that the estimate is **positive**. The logistic function

$$\frac{\exp(u)}{1+\exp(u)}$$

is an **increasing** function of u. Note that because the estimated values of β_1, β_2, and β_3 are positive, if x_{j1} changes from 0 (no contraceptives) to 1 (contraceptives), the **linear predictor**

$$\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \beta_3 x_{j3}$$

evaluated at the estimates increases, and the same is true if age x_{j2} increases or if x_{j3} changes from 0 (no smoking) to 1 (smoking). Thus, the fit indicates that the probability of having a heart attack **increases** if one uses oral contraceptives or smokes, and increases as women age.

- In fact, we can say more. According to this model, the **odds** of having a heart attack, given a woman has particular settings of contraceptive use, age, and smoking $(x_{j1},,x_{j2},x_{j3})$ is, from (11.9), which is the ratio of the probability of having a heart attack to not having one, is

$$\exp(\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \beta_3 x_{j3}).$$

A common quantity of interest is the so-called **odds ratio**. For example, we may be interested in comparing the odds of having a heart attack if a randomly chosen woman smokes ($x_{j3} = 1$) to those if she does not ($x_{j3} = 0$). The ratio of the odds under smoking to those under not smoking, for any settings of age or contraceptive use, is thus

$$\frac{\exp(\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \beta_3)}{\exp(\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2})} = \exp(\beta_3).$$

Thus, $\exp\beta_3$ is a multiplicative factor that measures by how much the odds of having a heart attack change if we move from not smoking to smoking. If $\beta_3 > 0$, this multiplicative factor is > 1, meaning that the odds go up; if β_3 is negative, the factor is < 1, and the odds go down. β_3 itself is referred to as the **log odds ratio** for obvious reasons.

Here, we estimate the log odds ratio for smoking as 1.81 and the odds ratios as $\exp(\widehat{\beta}_3) = \exp(1.81) = 6.12$; the odds increase by 6-fold if a woman smokes! Note that, ideally, we would like a **standard error** to attach to this estimated odd ratios.

One can actually get `PROC GENMOD` to print out a log odds ratio and odds ratio and associated standard errors in an `estimate` statement with the `exp` option by choosing $\boldsymbol{L}$ appropriately. Here, to get the log odds ratio, which is just β_3, we take $\boldsymbol{L} = (0, 0, 0, 1)$. The `estimate` tatement would be

```
estimate "smk log odds ratio" int 0 oral 0 age 0 smoke 1 / exp;
```

try adding this to the program and see what happens (see the program on the class web site for the results).

- An interesting aside: Logistic regression is a standard technique in public health studies. Chances are, when you read in the newspaper that a certain behavior increases the risk of developing a disease, the analysis that was performed to arrive at that conclusion was like this one.

EXAMPLE 3 – CLOTTING TIME DATA: These data are positive and continuous with possible constant coefficient of variation. Thus, we consider the gamma probability model. Letting Y_j be the clotting time at percentage concentration x_j, we consider two models for the mean response:

- Loglinear: $E(Y_j) = \exp(\beta_0 + \beta_1 x_j)$
- Reciprocal (inverse): $E(Y_j) = 1/(\beta_0 + \beta_1 x_j)$.

Note that although in both models β_1 has to do with how the changing percentage concentration affects the mean response, this happens in different ways in each model, so the parameters have different interpretations, so it is not interesting to compare their values for the different models.

Here, because of the gamma assumption, the dispersion parameter ϕ is not equal to a fixed, known constant. It is thus estimated from the data. Note that `PROC GENMOD` does not print out the estimate of ϕ; rather, it prints out $1/\phi$.

We also show how to obtain results of the fit in a table that may be output to a SAS data set using the `ods` statement, which is relevant in versions 7 and higher of SAS. Earlier versions use the `make` statement.

PROGRAM:

```
/*****************************************************************

  CHAPTER 11, EXAMPLE 3

  Fitting loglinear and reciprocal models to the clotting data.
  (Gamma assumption)

*****************************************************************/

options ls=80 ps=59 nodate; run;

/*****************************************************************

  The data look like

       5 118
      10  58
      15  42
      20  35
      30  27
      40  25
      60  21
      80  19
      100 18

  column 1      percentage concentration plasma
  column 2      clotting time (seconds)

*****************************************************************/

data clots; infile 'clot.dat';
  input u y;
  x=log(u);
run;

/*****************************************************************

  Fit the loglinear regression model using PROC GENMOD.  The
  DIST=GAMMA option specifies the gamma distribution assumption.
  We then fit two models: the loglinear model in the first
  call to PROC GENMOD, obtained with the LINK=LOG option,
  and the reciprocal (inverse) model, obtained with the
  LINK=POWER(-1) option -- this option asks that the linear
  predictor be raised to the power in parentheses as the model
  for the mean response.

  Here, the dispersion parameter phi is unknown so must be estimated.
  This may be done a number of ways -- here, we use the PSCALE
  option in MODEL statement to ask that phi be estimated
  by the Pearson chi-square divided by its degrees of freedom.
  Actually, for the gamma distribution, what is printed under
  SCALE parameter is the reciprocal of this quantity, so we must
  remember to invert the result from the output to obtain the estimate
  of phi.

  Also, use the OBSTATS option in the MODEL statement to output a
  table of statistics such as predicted values (estimates of the mean
  response) and residuals (response-estimated mean).   We show
  how to output these to a data set using the ODS statement for
  for the loglinear fit (although we don't do anything with them).
  The ODS statement works with version 7 and higher of SAS.
  Note that the obstats option causes the output of GENMOD to contain
  these statistics; printing the output data set simply repeats
  these values.

*****************************************************************/

proc genmod data=clots;
  model y = x /  dist = gamma link = log obstats pscale;
  ods output obstats=outlog;
run;

proc print data=outlog; run;

/*****************************************************************

  Fit the inverse reciprocal regression model using PROC GENMOD.
  Phi is again calculated by the Pearson chi-square/dof.

*****************************************************************/

proc genmod data=clots;
  model y = x /  dist = gamma link = power(-1) obstats pscale;
run;
```

OUTPUT: Following the output, we comment on a few aspects of the output.

```
                              The SAS System                                1

                            The GENMOD Procedure

                             Model Information

                          Data Set              WORK.CLOTS
                          Distribution               Gamma
                          Link Function                Log
                          Dependent Variable             y

                       Number of Observations Read           9
                       Number of Observations Used           9

                        Criteria For Assessing Goodness Of Fit

          Criterion                 DF           Value          Value/DF

          Deviance                   7          0.1626            0.0232
          Scaled Deviance            7          6.6768            0.9538
          Pearson Chi-Square         7          0.1705            0.0244
          Scaled Pearson X2          7          7.0000            1.0000
          Log Likelihood                      -26.4276

    Algorithm converged.

                          Analysis Of Parameter Estimates

                               Standard       Wald 95%          Chi-
     Parameter  DF  Estimate      Error   Confidence Limits    Square  Pr > ChiSq

     Intercept   1    5.5032     0.1799     5.1506    5.8559   935.63      <.0001
     x           1   -0.6019     0.0520    -0.7039   -0.4999   133.80      <.0001
     Scale       0   41.0604     0.0000    41.0604   41.0604

NOTE: The Gamma scale parameter was estimated by DOF/Pearson's Chi-Square

                          Lagrange Multiplier Statistics

                          Parameter     Chi-Square    Pr > ChiSq

                          Scale             0.3069        0.5796

                                  Observation Statistics

 Observation          y          x       Pred      Xbeta        Std    HessWgt
                                 Lower      Upper     Resraw     Reschi     Resdev
                              StResdev   StReschi     Reslik

           1        118  1.6094379  93.175154  4.5344811  0.1026374  52.000165
                         76.196496  113.93712  24.824846   0.266432  0.2458801
                         2.1728608  2.3544798  2.2608074

                              The SAS System                                2

                            The GENMOD Procedure

                                  Observation Statistics

 Observation          y          x       Pred      Xbeta        Std    HessWgt
                                 Lower      Upper     Resraw     Reschi     Resdev
                              StResdev   StReschi     Reslik

           2         58  2.3025851   61.39102  4.1172636  0.0738424  38.792341
                         53.119026  70.951174   -3.39102  -0.055236  -0.056288
                        -0.413325  -0.405606  -0.411497
           3         42  2.7080502  48.096382   3.873207  0.0607149  35.855825
                         42.700382  54.174268  -6.096382  -0.126753  -0.132544
                           -0.9248    -0.8844  -0.918591
           4         35  2.9957323  40.449166   3.700046  0.0545252  35.528863
                         36.349431  45.011297  -5.449166  -0.134716  -0.141291
                        -0.967048   -0.92205  -0.961605
           5         27  3.4011974  31.689627  3.4559894   0.052237     34.984
                         28.605721  35.106001  -4.689627  -0.147986  -0.155989
                        -1.060815  -1.006389  -1.054851
           6         25  3.6888795  26.651048  3.2828285  0.0556359  38.516653
                         23.897747  29.721562  -1.651048  -0.061951  -0.063278
                        -0.434509  -0.425393  -0.433342
           7         21  4.0943446  20.879585  3.0387719  0.0661298  41.297168
                         18.341382  23.769042  0.1204152  0.0057671  0.0057561
                         0.0409427  0.0410213  0.0409576
           8         19  4.3820266  17.559778   2.865611  0.0762872  44.428066
                         15.121094  20.391766  1.4402218  0.0820182  0.0798774
                         0.5932165  0.6091154  0.5973195
```

```
          9         18  4.6051702  15.352785  2.7312969  0.0851497  48.140231
                        12.992945  18.141231  2.6472147  0.1724257  0.1634065
                        1.2715487  1.3417313  1.2945556
```


Obs	Observation	y	x	Pred	Xbeta
1	1	118	1.6094379	93.175154	4.5344811
2	2	58	2.3025851	61.39102	4.1172636
3	3	42	2.7080502	48.096382	3.873207
4	4	35	2.9957323	40.449166	3.700046
5	5	27	3.4011974	31.689627	3.4559894
6	6	25	3.6888795	26.651048	3.2828285
7	7	21	4.0943446	20.879585	3.0387719
8	8	19	4.3820266	17.559778	2.865611
9	9	18	4.6051702	15.352785	2.7312969

Obs	Std	Hesswgt	Lower	Upper	Resraw
1	0.1026374	52.000165	76.196496	113.93712	24.824846
2	0.0738424	38.792341	53.119026	70.951174	-3.39102
3	0.0607149	35.855825	42.700382	54.174268	-6.096382
4	0.0545252	35.528863	36.349431	45.011297	-5.449166
5	0.052237	34.984	28.605721	35.106001	-4.689627
6	0.0556359	38.516653	23.897747	29.721562	-1.651048
7	0.0661298	41.297168	18.341382	23.769042	0.1204152
8	0.0762872	44.428066	15.121094	20.391766	1.4402218
9	0.0851497	48.140231	12.992945	18.141231	2.6472147

Obs	Reschi	Resdev	Stresdev	Streschi	Reslik
1	0.266432	0.2458801	2.1728608	2.3544798	2.2608074
2	-0.055236	-0.056288	-0.413325	-0.405606	-0.411497
3	-0.126753	-0.132544	-0.9248	-0.8844	-0.918591
4	-0.134716	-0.141291	-0.967048	-0.92205	-0.961605
5	-0.147986	-0.155989	-1.060815	-1.006389	-1.054851
6	-0.061951	-0.063278	-0.434509	-0.425393	-0.433342
7	0.0057671	0.0057561	0.0409427	0.0410213	0.0409576
8	0.0820182	0.0798774	0.5932165	0.6091154	0.5973195
9	0.1724257	0.1634065	1.2715487	1.3417313	1.2945556

The GENMOD Procedure

Model Information

Data Set	WORK.CLOTS
Distribution	Gamma
Link Function	Power(-1)
Dependent Variable	y

Number of Observations Read	9
Number of Observations Used	9

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	7	0.0167	0.0024
Scaled Deviance	7	6.8395	0.9771
Pearson Chi-Square	7	0.0171	0.0024
Scaled Pearson X2	7	7.0000	1.0000
Log Likelihood		-16.1504	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.0166	0.0009	-0.0184	-0.0147	318.53	<.0001
x	1	0.0153	0.0004	0.0145	0.0162	1367.15	<.0001
Scale	0	408.8247	0.0000	408.8247	408.8247		

NOTE: The Gamma scale parameter was estimated by DOF/Pearson's Chi-Square

Lagrange Multiplier Statistics

Parameter	Chi-Square	Pr > ChiSq
Scale	0.2600	0.6101

Observation Statistics

Observation	y	x	Pred	Xbeta	Std	HessWgt
		Lower	Upper	Resraw	Reschi	Resdev
		StResdev	StReschi	Reslik		

```
          1        118  1.6094379  122.85904  0.0081394  0.0003814  6170940.5
                        112.52367  135.28505  -4.859041   -0.03955  -0.040083
                        -2.535827  -2.502059   -2.50553

                                The SAS System                                5

                            The GENMOD Procedure

                            Observation Statistics

 Observation          y          x       Pred      Xbeta        Std    HessWgt
                             Lower      Upper     Resraw     Reschi     Resdev
                          StResdev   StReschi     Reslik

          2         58  2.3025851  53.263889  0.0187744  0.0003353  1159852.7
                        51.462321  55.196169  4.7361113  0.0889179  0.0864112
                        1.8736358  1.9279877  1.8808138
          3         42  2.7080502  40.007131  0.0249955  0.0004121  654352.76
                        38.754832  41.343065  1.9928686  0.0498128   0.049009
                        1.0510498  1.0682898  1.0529795
          4         35  2.9957323  34.002638  0.0294095  0.0004948  472674.68
                        32.917102  35.162214  0.9973619  0.0293319  0.0290499
                        0.6246313  0.6306943   0.625336
          5         27  3.4011974  28.065779  0.0356306  0.0006317  322026.28
                         27.12331  29.076102  -1.065779  -0.037974  -0.038466
                        -0.833125  -0.822477  -0.831765
          6         25  3.6888795  24.972206  0.0400445  0.0007367   254947.6
                        24.103101  25.906332  0.0277938   0.001113  0.0011126
                        0.0242347  0.0242437   0.024236
          7         21  4.0943446  21.614323  0.0462656  0.0008909  190994.29
                        20.828244  22.462064  -0.614323  -0.028422  -0.028696
                        -0.629919  -0.623908  -0.629011
          8         19  4.3820266  19.731822  0.0506796   0.001003  159173.77
                         18.99499  20.528126  -0.731822  -0.037088  -0.037557
                        -0.828624  -0.818283  -0.826977
          9         18  4.6051702   18.48317  0.0541033  0.0010911  139665.78
                        17.780391  19.243791   -0.48317  -0.026141  -0.026372
                        -0.583988  -0.578865  -0.583139
```

INTERPRETATION:

- Pages 1–2 of the output show the results of fitting the loglinear model. The estimates of β_0 and β_1 and their estimated standard errors are given in the table `Analysis of Parameter Estimates`. The `SCALE` parameter estimate corresponds to an estimate of $1/\phi$; thus, the estimate of ϕ itself is $1/41.0604 = 0.02435$. Recall that the coefficient of variation σ is defined as $\sigma^2 = \phi$; thus, the estimated coefficient of variation under the loglinear fit is 0.15606.

- The table `Observation Statistics` on pages 1 and 2 lists a number of results based on the fit. Of particular interest is the column `PRED`, which gives the estimates of the mean response at each x_j value (the column `Y` contains the actual data values for comparison). These numbers are repeated on page 3, which shows the result of the call to `proc print` to print the data set created by the `ods` statement. This illustrates how it is possible to output such results so that further manipulation may be undertaken.

- Pages 4–5 contain the same information for the reciprocal link fit. Here, the estimate of ϕ is $1/408.8247 = 0.002446$, so that the estimated coefficient of variation σ is 0.04946.

- Note that the estimates of CV do not agree well at all between the two fits. The reason can be appreciated when one inspects the lower right panel of Figure 3. Here, the estimated mean

response for each fit is superimposed on the actual data – the solid line represents the fit of the loglinear model, the dashed line is the fit of the reciprocal model. Note that this second model appears to provide a much better fit to the data. The calculation of ϕ, and hence of σ, is based on squared deviations $\{Y_j - f(\boldsymbol{x}_j'\widehat{\boldsymbol{\beta}})\}^2$. Because the loglinear model fits poorly, these deviations are large, leading to an estimate of CV that is misleading large. The reciprocal model, which fits the data very well, leads to a much smaller estimate because the deviations of the fit from the observed responses are much smaller. Based on the visual evidence, the fit of the reciprocal model is preferred for describing the percentage concentration of plasma-clotting time relationship.