Chapter 4
Beyond Linear Regression: The Method of Maximum Likelihood

Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.
Box [2, p. 792]

4.1 Introduction and Overview

The linear regression model introduced in Chap. 2 assumes the variance is constant, possibly from a normal distribution. Many data types exist for which the randomness is not constant, and so other methods are necessary. This chapter demonstrates situations where the linear regression model fails. In these cases, least-squares estimation, as used in Chap. 2, is no longer appropriate. Instead, maximum likelihood estimation is appropriate. In Chap. 4, we discuss three specific situations in which linear regression models fail (Sect. 4.2) and then consider a general approach to modelling such data (Sect. 4.3). To fit these models, maximum likelihood estimation is needed and is reviewed in Sect. 4.4. We then examine maximum likelihood estimation in the case of one parameter (Sect. 4.5) and more than one parameter (Sect. 4.6), and then using matrix algebra (Sect. 4.7). Fitting models using maximum likelihood is discussed in Sect. 4.8, followed by a review of the properties of maximum likelihood estimators (Sect. 4.9). Results concerning hypothesis tests (Sect. 4.10) and confidence intervals (Sect. 4.11) are then presented, followed by a discussion of comparing non-nested models (Sect. 4.12).

4.2 The Need for Non-normal Regression Models

4.2.1 When Linear Models Are a Poor Choice

The random component of the regression models in Chap. 2 has constant variance, possibly from a normal distribution. Three common situations exist where the variation is not constant, and so linear regression models are a poor choice for modelling such data:

P. K. Dunn, G. K. Smyth, *Generalized Linear Models with Examples in R*,
Springer Texts in Statistics, https://doi.org/10.1007/978-1-4419-0118-7_4

1. The response is a *proportion*, ranging between 0 and 1 inclusive, of a total number of counts. As the modelled proportion approaches these boundaries of 0 and 1, the variance of the responses must approach zero. The variance must be smaller near 0 and 1 than the variation of proportions near 0.5 (where the observations can spread equally in both directions toward the boundaries). Thus, the variance is not, and cannot be, constant. Furthermore, because the response is between 0 and 1, the randomness cannot be normally distributed. For proportions of a total number of counts, the *binomial* distribution may be appropriate (Sect. 4.2.2; Chap. 9).
   A specific example of binomial data is *binary* data (Example 4.6) where the response takes one of two outcomes (such as 'success' and 'failure', or 'present' and 'absent').
2. The response is a *count*. As the modelled count approaches zero, the variance of the responses must approach zero. Furthermore, the normal distribution is a poor choice for modelling the randomness because counts are discrete and non-negative. For count data, the *Poisson* distribution may be appropriate (Example 1.5; Sect. 4.2.3; Chap. 10).
3. The response is *positive continuous*. As the modelled response approaches zero, the variance of the responses must approach zero. Furthermore, the normal distribution is a poor choice because positive continuous data are often right skewed, and because the normal distribution permits negative values. For positive continuous data, distributions such as the *gamma* and *inverse Gaussian* distributions may be appropriate (Sect. 4.2.4; Chap. 11).

In these circumstances, the relationship between y and the explanatory variables is usually non-linear also: the response has boundaries in all cases, so a linear relationship cannot apply for all values of the response.

4.2.2 Binary Outcomes and Binomial Counts

First consider binary regression. There are many applications in which the response is a binary variable, taking on only two possible states. In this situation, a transformation to normality is out of the question.

Example 4.1. (Data set: `gforces`) Military pilots sometimes black out when their brains are deprived of oxygen due to G-forces during violent manoeuvres. A study [7] produced similar symptoms by exposing volunteers' lower bodies to negative air pressure, likewise decreasing oxygen to the brain. The data record the ages of eight volunteers and whether they showed syncopal blackout-related signs (pallor, sweating, slow heartbeat, unconsciousness) during an 18 min period. Does resistance to blackout decrease with age?

```
> data(gforces); gforces
  Subject Age Signs
1      JW  39     0
2      JM  42     1
3      DT  20     0
4      LK  37     1
5      JK  20     1
6      MK  21     0
7      FP  41     1
8      DG  52     1
```

The explanatory variable is `Age`. The response variable is `Signs`, coded as `1` if the subject showed blackout-related signs and `0` otherwise. The response variable is binary, taking only two distinct values, and no transformation can change that. A regression approach that directly models the *probability* of a blackout response given the age of the subject is needed. □

The same principles apply to situations where a number of binary outcomes are tabulated to make a binomial random variable, as in the following example.

Example 4.2. (Data set: `shuttles`) After the explosion of the space shuttle *Challenger* on January 28, 1986, a study was conducted [3, 4] to determine if previously-collected data about the ambient air temperature at the time of launch could have been used to foresee potential problems with the launch (Table 4.1). In this example, the response variable is the number of damaged O-rings out of six for each of the previous 23 launches with data available, so only seven values are possible for the response. No transformation can change this.

A more sensible model would be to use a binomial distribution with mean proportion μ for modelling the proportion y of O-rings damaged out of m at various temperatures x. (Here, $m = 6$ for every launch.) Furthermore, a linear relationship between temperature and the proportion of damaged O-rings cannot be linear, as proportions are restricted to the range $(0, 1)$. Instead, a systematic relationship of the form

$$\log \frac{\mu}{1-\mu} = \beta_0 + \beta_1 x$$

may be more suitable, since $\log\{\mu/(1-\mu)\}$ has a range over the entire real line. □

Combining the systematic and random components, a possible model for the data is:

$$\begin{cases} ym \sim \text{Bin}(\mu, m) & \text{(random component)} \\ \log \dfrac{\mu}{1-\mu} = \beta_0 + \beta_1 x & \text{(systematic component).} \end{cases} \tag{4.1}$$

Table 4.1 The ambient temperature and the number of O-rings (out of six) damaged for 23 of the 24 space shuttle launches before the launch of *Challenger*; *Challenger* was the 25th shuttle. One engine was lost at sea and so its O-rings could not be examined (Example 4.2)

Temperature (in °F)	O-rings damaged	Temperature (in °F)	O-rings damaged	Temperature (in °F)	O-rings damaged
53	2	68	0	75	0
57	1	69	0	75	2
58	1	70	0	76	0
63	1	70	0	76	0
66	0	70	1	78	0
67	0	70	1	79	0
67	0	72	0	81	0
67	0	73	0		

4.2.3 Unrestricted Counts: Poisson or Negative Binomial

Count data is another situation where linear regression models are inadequate.

Example 4.3. (Data set: `nminer`) A study [9] of the habitats of the noisy miner (a small but aggressive native Australian bird) counted the number of noisy miners y and the number of eucalypt trees x in two-hectare buloke woodland transects (Table 1.2, p. 15). Buloke woodland patches with more eucalypts tend to have more noisy miners (Fig. 1.4, p. 15).

The number of noisy miners is more variable where more eucalypts are present. Between 0 and 10 eucalypts, the number of noisy miners is almost always zero; between 10 and 20 eucalypts, the number of noisy miners increases. This shows that the systematic relationship between the number of eucalypts and the number of noisy miners is not linear. A possible model for the systematic component is $\log \mu = \beta_0 + \beta_1 x$, where x is the number of eucalypt trees at a given site, and μ is the expected number of noisy miners. Using the logarithm ensures $\mu > 0$ even when β_0 and β_1 range between $-\infty$ and ∞, and also models the non-linear form of the relationship between μ and x.

Between 0 and 10 eucalypts, the number of noisy miners varies little. Between 10 and 20 eucalypts, a larger amount of variation exists in the number of noisy miners. This shows that the randomness does not have constant variance. Instead, the variation in the data may be modelled using a *Poisson distribution*, $y \sim \text{Pois}(\mu)$, where $y = 0, 1, 2, \ldots$, and $\mu > 0$.

Combining the systematic and random components, a possible model for the data is:

$$\begin{cases} y \sim \text{Pois}(\mu) & \text{(random component)} \\ \log \mu = \beta_0 + \beta_1 x & \text{(systematic component).} \end{cases} \tag{4.2}$$

□

Table 4.2 The time for delivery to soft drink vending machines (Example 4.4)

Time (in mins)	Cases	Distance (in feet)	Time (in mins)	Cases	Distance (in feet)	Time (in mins)	Cases	Distance (in feet)
16.68	7	560	79.24	30	1460	19.00	7	132
11.50	3	220	21.50	5	605	9.50	3	36
12.03	3	340	40.33	16	688	35.10	17	770
14.88	4	80	21.00	10	215	17.90	10	140
13.75	6	150	13.50	4	255	52.32	26	810
18.11	7	330	19.75	6	462	18.75	9	450
8.00	2	110	24.00	9	448	19.83	8	635
17.83	7	210	29.00	10	776	10.75	4	150
			15.35	6	200			

4.2.4 Continuous Positive Observations

A third common situation where linear regressions are unsuitable is for positive continuous data.

Example 4.4. (Data set: `sdrink`) A soft drink bottler is analyzing vending machine service routes in his distribution system [11, 13]. He is interested in predicting the amount of time y required by the route driver to service the vending machines in an outlet. This service activity includes stocking the machine with beverage products and minor maintenance or housekeeping. The industrial engineer responsible for the study has suggested that the two most important variables affecting the delivery time are the number of cases of product stocked x_1 and the distance walked by the route driver x_2. The engineer has collected 25 observations on delivery time, the number of cases and distance walked (Table 4.2).

In this case, the delivery times are strictly positive values. They are likely to show an increasing mean–variance relationship with standard deviation roughly proportional to the mean, so a log-transformation might be approximately variance stabilizing. However the dependence of time on the two covariates is likely to be directly linear, because time should increase linearly with the number of cases or the distance walked (Fig. 4.1); that is, a sensible systematic component is $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. No normal linear regression approach can achieve these conflicting aims, because any transformation to stabilize the variance would destroy linearity. A regression approach that directly models the delivery times using an appropriate probability distribution for positive numbers (such as a gamma distribution) is desirable. Combining the systematic and random components, a possible model for the data is:

$$\begin{cases} y \sim \text{Gamma}(\mu; \phi) & \text{(random component)} \\ \mu = \beta_0 + \beta_1 x & \text{(systematic component)} \end{cases} \tag{4.3}$$

where ϕ is related to the variance of the gamma distribution. □

Fig. 4.1 A plot of the soft drink data: time against the number of cases of product sold (left panel) and time against the distance walked by the route driver (right panel)

Table 4.3 The time to death (in weeks) and white blood cell count (WBC) for leukaemia patients, grouped according to AG type (Example 4.5)

AG positive patients				AG negative patients			
WBC	Time to death	WBC	Time to death	WBC	Time to death	WBC	Time to death
2300	65	7000	143	4400	56	28000	3
750	156	9400	56	3000	65	31000	8
4300	100	32000	26	4000	17	26000	4
2600	134	35000	22	1500	7	21000	3
6000	16	100000	1	9000	16	79000	30
10500	108	100000	1	5300	22	100000	4
10000	121	52000	5	10000	3	100000	43
17000	4	100000	65	19000	4	27000	2
5400	39						

Example 4.5. (Data set: `leukwbc`) The times to death (in weeks) of two groups of leukaemia patients (grouped according to a morphological variable called the AG factor) were recorded (Table 4.3) and their white blood cell counts were measured (Fig. 4.2). The authors originally fitted a model using the exponential distribution [5, 6].

We would like to model the survival times on a log-linear scale, building a linear predictor for $\log \mu_i$, where $\mu_i > 0$ is the expected survival time. However the log-survival times are not normally distributed, as the logarithm of an exponentially distributed random variable is markedly left-skewed. Hence normal linear regression with the log-survival times as response is less than desirable. Furthermore, linear regression would estimate the variance of the residuals, whereas the variance of an exponential random variable is known once the mean is specified. An analysis that uses the exponential distribution explicitly is needed. □

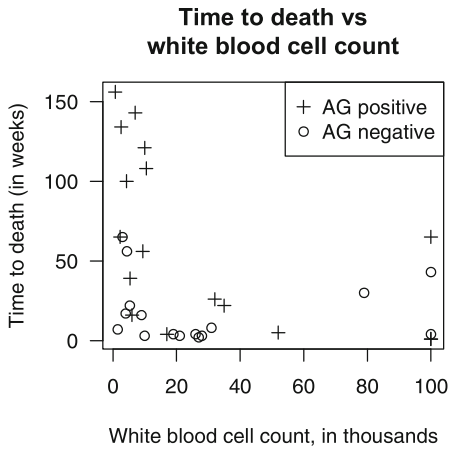

Fig. 4.2 A plot of the leukaemia data: time to death against the white blood cell count (Example 4.5)

Table 4.4 Different models discussed so far, all of which are generalized linear models. In all cases $\eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$ for the appropriate explanatory variables x_j (Sect. 4.3)

Data	Reference	Random component	Systematic component
FEV data	Example 1.1 (p. 1)	Normal	$\mu = \eta$
Challenger data	Example 4.2 (p. 167)	Binomial	$\log\{\mu/(1-\mu)\} = \eta$
Noisy miner data	Example 4.3 (p. 168)	Poisson	$\log \mu = \eta$
Soft drink data	Example 4.4 (p. 169)	Gamma	$\mu = \eta$
Leukaemia data	Example 4.5 (p. 170)	Exponential	$\log \mu = \eta$

4.3 Generalizing the Normal Linear Model

For the data in Sect. 4.2, different models are suggested (Table 4.4): a variety of random and systematic components appear. The theory in Chaps. 2 and 3, based on linearity and constant variance, no longer applies.

To use each of the models listed in Table 4.4 requires the development of separate theory: fitting algorithms, inference procedures, diagnostic tools, and so on. An alternative approach is to work more generally. For example, later we consider a *family* of distributions which has the normal, binomial, Poisson and gamma distributions as special cases. Using this general family of distributions, any estimation algorithms, inference procedures and diagnostic tools that are developed apply to *all* distributions in this family of distributions. Implementation for any one specific model would be a special case of the general theory. In addition, later we allow systematic components of the form $f(\mu) = \eta$ for certain functions $f()$.

This is the principle behind generalized linear models (GLMs). GLMs unify numerous models into one general theoretical framework, incorporating all the models in Table 4.4 (and others) under one structure. Common estimation algorithms (Chap. 6), inference methods (Chap. 7), and diagnostic tools (Chap. 8) are possible under one common framework. The family of distributions used for GLMs is called the *exponential dispersion model* (or EDM) family, which includes common distributions such as the normal, binomial, Poisson and gamma distributions, among others.

Why should the random component be restricted to distributions in the EDM family? For example, distributions such as the Weibull distribution and von Mises distribution are not EDMs, but may be useful for modelling certain types of data. GLMs are restricted to distributions in the EDM family because the general theory is developed by taking advantage of the structure of EDMs. Using the structure provided by the EDM family enables simple fitting algorithms and inference procedures, which share similarities with the normal linear regression models. The theory does not apply to distributions that are not EDMs. Naturally, if a non-EDM distribution really is appropriate it should be used (and the model will not be a GLM). However, EDMs are useful for most common types of data:

- Continuous data over the entire real line may be modelled by the normal distribution (Chaps. 2 and 3).
- Proportions of a total number of counts may be modelled by the binomial distribution (Example 4.2; Chap. 9).
- Discrete count data may be modelled by the Poisson or negative binomial distributions (Example 4.3; Chap. 10).
- Continuous data over the positive real line may be modelled by the gamma and inverse Gaussian distributions (Example 4.4; Chap. 11).
- Positive data with exact zeros may be modelled by a special case of the Tweedie distributions (Chap. 12).

The advantages of GLMs are two-fold. Firstly, the mean–variance relationship can be chosen separately from the appropriate scale for the linear predictor. Secondly, by choosing a response distribution that matches the natural support of the responses, we can expect to achieve a better approximation to the probability distribution.

4.4 The Idea of Likelihood Estimation

Chapter 2 developed the principle of least-squares as a criterion for estimating the parameters in the linear predictor of linear regression models. Least-squares is an appropriate criterion for fitting regression models to response data that are approximately normally distributed. In the remainder of this chapter, we develop a much more general estimation methodology called