An Introduction to Discrete Data Models

J.K. Lindsey

1. Simple Models
2. An Application: Models of Change
3. Overdispersion
4. Serial Dependence
5. Conclusions

1. Simple Models

For a brief introduction to logistic and log linear models, consider simple applications to modelling various forms of repeated observations.

Observations over Time

Suppose some response variable with two possible values, A and B, was recorded at two points in time.

A two-way table for change over time.

		Time 2	
		A	B
Time 1	A	45	13
	B	12	54

A first characteristic is a relative stability over time, indicated by the large frequencies on the diagonal.

Suppose that the responses at time 2 have a binomial distribution and that this distribution depends on what response was given at time 1.

We might have the simple linear regression model

$$\log\left(\frac{\pi_{1|j}}{\pi_{2|j}}\right) = \beta_0 + \beta_1 x_j$$

x_j is the response at time 1;

$\pi_{i|j}$ is the *conditional* probability of response i at time 2 given the observed value of x_j at time 1.

Then, if $\beta_1 = 0$, this indicates independence, that is, that the second response does not depend on the first.

This is a *logistic regression model*, with a logit link, the logarithm of the ratio of probabilities.

It is the direct analogue of classical (normal theory) linear regression.

On the other hand, if x_j is coded $(-1,1)$ or $(0,1)$, we may rewrite this as

$$\log\left(\frac{\pi_{1|j}}{\pi_{2|j}}\right) = \mu + \alpha_j$$

where $\mu = \beta_0$, the direct analogue of an analysis of variance model, with the appropriate constraints.

For our table, the parameter estimates are $\widehat{\beta}_0 = \widehat{\mu} = 1.242$ and $\widehat{\beta}_1 = \widehat{\alpha}_1 = -2.746$, when x_j is coded $(0,1)$.

That with $\alpha_1 = \beta_1 = 0$, that is, independence, fits the data much more poorly.

Clustered Observations

Suppose now that the same table are some data on the two eyes of people.

A two-way table of clustered data.

		Right eye	
		A	B
Left	A	45	13
eye	B	12	54

We again have repeated observations on the same individuals, but here they may be considered as being made simultaneously rather than sequentially.

Again, there will usually be a large number with similar responses, resulting from the dependence between the two similar eyes of each person.

Here, we would be more inclined to model the responses simultaneously.

Take a multinomial distribution over the four response combinations, with *joint* probability parameters, π_{ij}.

In that way, we can look at the association between them.

We might use a log link such that

$$\log(\pi_{ij}) = \phi + \mu_i + \nu_j + \alpha_{ij}$$

With the appropriate constraints, this is again an analogue of classical analysis of variance.

It is called a *log linear model*.

Here, the parameter estimates are $\widehat{\phi} = 2.565$, $\nu_1 = 1.424$, $\widehat{\mu}_1 = 1.242$, and $\widehat{\alpha}_{11} = -2.746$.

The conclusion is identical, that the independence model is much inferior to that with dependence.

Log Linear and Logistic Models

The two models just described have a special relationship to each other.

With the same constraints, the dependence parameter, $\boldsymbol{\alpha}$, is identical in the two cases because

$$\log\left(\frac{\pi_{1|1}\pi_{2|2}}{\pi_{1|2}\pi_{2|1}}\right) = \log\left(\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\right)$$

Inferences are also identical:

the normed profile likelihoods for $\boldsymbol{\alpha} = \mathbf{0}$ are also identical.

This is a general result:

in cases where both are applicable, logistic and log linear models yield the same conclusions.

The choice is a matter of convenience.

This is a very important property, because it means that such models can be used for *retrospective sampling*.

Common examples of this include, in medicine, case-control studies, and, in the social sciences, mobility studies.

These results extend directly to larger tables, including higher dimensional tables.

There, direct analogues of classical regression and ANOVA models are still applicable.

Thus, complex models of dependence among categorical variables can be built up by means of multiple regression.

Explanatory variables can be discrete or continuous (at least if the data are not aggregated in a contingency table).

2. An Application: Models of Change

One of the most important uses of log linear models has been in sample survey data.

A particularly interesting area of this field is *panel data*.

The same survey questions are administered at two or more points in time to the same people.

Let us restrict attention to the observation of responses at only two points in time.

Suppose that the response has I categories, called the *states*.

We have a $I \times I$ table and are studying changes in state over time.

The dependence parameter, $\boldsymbol{\alpha}$, will be a $I \times I$ matrix.

Because of the need for constraints, there will be only $(I-1) \times (I-1)$ independent values.

When $I > 2$, the idea is to reduce this number of parameters by structuring the values in some informative way.

The minimal model will be independence, that is, when $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$ or, equivalently, $\alpha_{ij} = 0\ \forall i, j$.

The maximal model is the saturated or "nonparametric" one.

Most interesting models are based on *Markov chains*:

the current response simply is made to depend on the previous one.

These are models describing the *transition probabilities* of changing from one state to another between two points in time.

Mover–Stayer Model

We have noticed that there is often a rather large number of individuals who will give the same response the two times.

Let us first see how to model this.

Suppose that we have a mixture of two subpopulations or latent groups.

One is susceptible to change while the other is not.

This is called a *mover–stayer* model.

We know that individuals recorded off the main diagonal will all belong to the first subpopulation, the movers, because they have changed.

The main diagonal frequencies are more complex:

they will contain both the stayers and any movers who did not happen to change within the observation period.

Let us assume that the locations of the movers at the two points in time are independent.

If we ignore the mixture on the diagonal, we can model the rest of the table by *quasi-independence*.

With this independence assumption, we can obtain estimates of the number of movers on the diagonal and, hence, of the number of stayers.

Example

Place of residence in Britain in 1966 and 1971.

	1971			
1966	CC	ULY	WM	GL
Central Clydesdale	118	12	7	23
Urban Lancs. & Yorks.	14	2127	86	130
West Midlands	8	69	2548	107
Greater London	12	110	88	7712

The deviance for independence is 19,884 with nine d.f., a strong indication of dependence.

That for the mover–stayer model (quasi-independence), fitted in the same way but to the table without the main diagonal, is 4.4 with 5 d.f.

The dependence arises almost entirely from stayers being in the same place at the two time points.

The numbers of movers on the diagonal are estimated to be only 1.6, 95.2, 60.3, and 154.6, respectively.

Most people in the table can have their 1971 place of residence exactly predicted by that of 1966:

they will be in the same place.

Symmetry

Because, in panel data, the same response variables are being recorded two (or more) times, we might expect some symmetry among them.

Complete Symmetry

Suppose that the probability of changing between any pair of categories is the same in both directions:

$$\pi_{i|j} = \pi_{j|i} \qquad \forall i, j$$

a model of *complete symmetry*. In terms of Markov chains, this is equivalent to the combination of two characteristics,

reversibility and *equilibrium*.

Equilibrium

Here, the marginal probabilities are the same at the two time points,

$$\pi_{i\bullet} = \pi_{\bullet i} \qquad \forall i$$

The marginal distribution of the states remains the same at the different time points.

In the analysis of contingency tables, this is called *marginal homogeneity*.

Reversibility

Reversibility implies (more or less) equal transition probabilities both ways between pairs of response categories, within the constraints of the marginal probabilities being those values observed.

In terms of log linear models, this is called *quasi-symmetry*.

Combining quasi-symmetry with marginal homogeneity yields complete *symmetry* (about the main diagonal) in the table.

Example

Sweden election votes in 1968 and 1970.

	1970				
1968	SD	C	P	Con	Total
SD	850	35	25	6	916
C	9	286	21	6	322
P	3	26	185	5	219
Con	3	26	27	138	194
Total	865	373	258	155	1651

SD - Social Democrat C - Centre
P - People's Con - Conservative

Besides the relatively large diagonal values, there also appears to be a "distance" effect:

a defecting voter seems more likely to switch to a nearby party on the left–right scale.

The equilibrium or marginal homogeneity model has a deviance of 65.2 with 3 d.f.

The reversibility or quasi-symmetry model has 2.5 with 3 d.f.

The overall election results changed, but, given this, the transfers between parties were equal in both directions.

They are highest between adjacent parties.

3. Overdispersion

Models based on the binomial, multinomial, and Poisson distributions involve strong assumptions.

The variance has a fixed relationship to the mean.

For example, for a Poisson distribution, the mean equals the variance.

In certain circumstances, such a relationship will be found not to hold.

Generally, this occurs when the events being counted are not independent.

Usually, the empirically calculated variance will be found to be too large as compared to the theoretical one.

This is called overdispersion.

The usual model for overdispersed binomial data is the beta-binomial distribution.

One way that this can be derived is by assuming that the binomial probability varies in a heterogeneous population according to a beta distribution.

This is then integrated to obtain the marginal beta-binomial distribution of the counts.

The negative binomial distribution can be obtained for Poisson count data in a similar way.

These distributions have an extra parameter measuring dispersion.

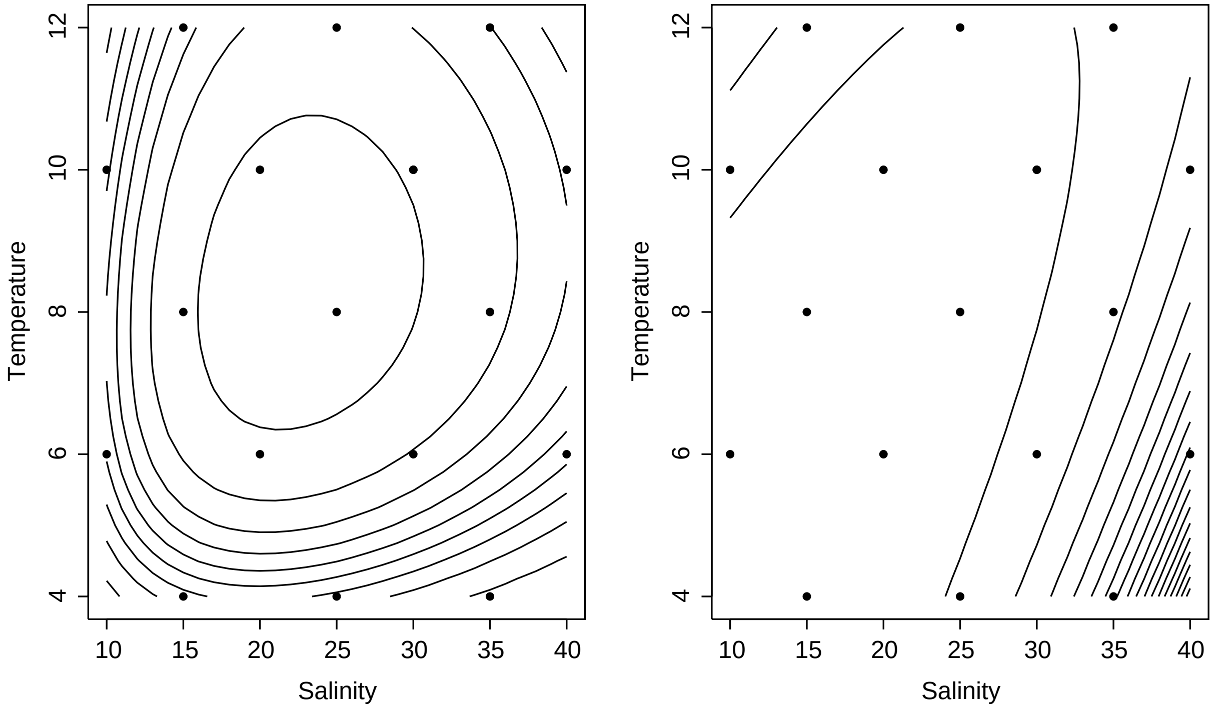
However, there is no reason that this remains constant under all conditions.

Example

Consider a study using a response surface design for fish eggs hatching under various conditions of temperature and salinity.

Four sets of eggs were kept in separate cells of each tank corresponding to a point of the design chosen.

There is more variability among among cells within a tank, all under the same controlled conditions, than would be expected under a binomial distribution.



Contours for the response surfaces for the probability of sole eggs hatching (left) and for the correlation among the eggs (right), along with the design points where observations were made. Probability contours range from 0.1 to 0.9 in steps of 0.1; correlation contours range from 0.04 to 0.32 in steps of 0.02.

4. Serial Dependence

Consider counts of events over time.

These will follow some profile of change.

For example, this might be a growth curve, having perhaps the logistic

$$\mu_t = \frac{N \exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

or Gompertz

$$\mu_t = N\{1 - \exp[-\exp(\beta_0 + \beta_1 x)]\}$$

form where N is the asymptotic maximum number of events.

Suppose that this common underlying profile exists for all individuals under the same conditions.

However, a given individual may deviate momentarily from the curve.

Obtain individual profiles by predicting the result at time $t + 1$ from the previously available information.

Use the common profile corrected by how far that individual (i) was from it at the previous time point:

$$\mu_{i,t+1} = \mu_{t+1} + \rho^{\Delta t}(n_{it} - \mu_t)$$

with $0 < \rho < 1$ and $n_{i0} = \mu_0$.

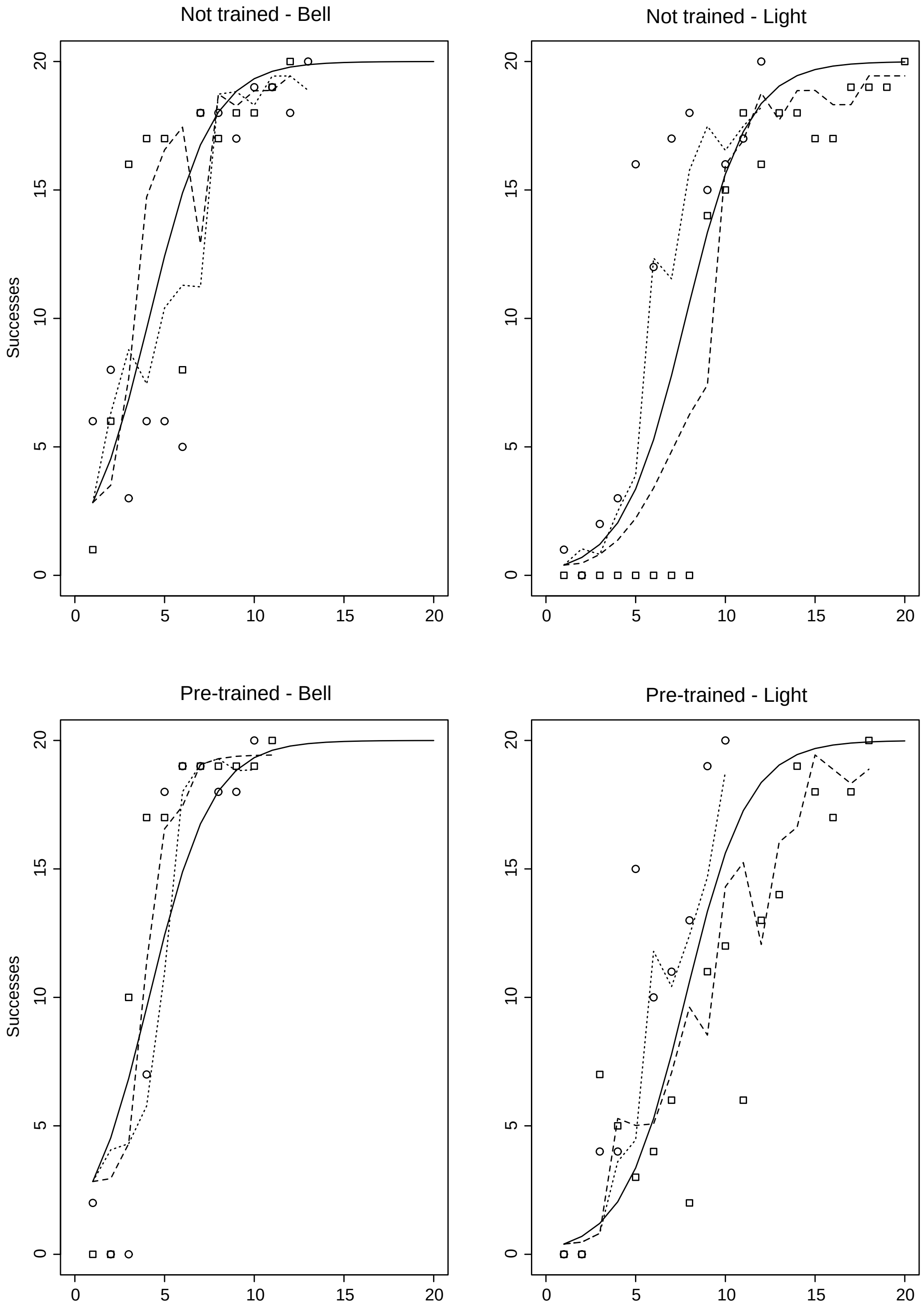
Example

16 laboratory animals were tested for learning in a 2×2 factorial experiment with training or not and light or bell stimulus.

Each animal was allowed 20 attempts to complete a task in each of a series of trials.

Trials for an animal stopped when a perfect score was reached.

The counts are overdispersed but there is also correlation of the numbers of successes over time.


Not trained - Bell
Not trained - Light
Pre-trained - Bell
Pre-trained - Light
Successes
Successes
0
5
10
15
20

5. Conclusions

Over the last 30 years, models for frequency and count data have become the most important area of applied statistics.

Many good textbooks are available.

The standard (logistic and log linear) models are relatively simple and easy to understand.

Close relationships exist to analysis of time to event (survival, failure time) data.

Many new and more complex models for realistic modelling of dependent events have appeared in the last 5 to 10 years.

However, many important areas still require further exciting research.