19
Log-Linear Models

In this chapter we study **log-linear models** which are useful for modeling multivariate discrete data. There is a strong connection between log-linear models and undirected graphs.

19.1 The Log-Linear Model

Let $X = (X_1, \ldots, X_m)$ be a discrete random vector with probability function

$$f(x) = \mathbb{P}(X = x) = \mathbb{P}(X_1 = x_1, \ldots, X_m = x_m)$$

where $x = (x_1, \ldots, x_m)$. Let r_j be the number of values that X_j takes. Without loss of generality, we can assume that $X_j \in \{0, 1, \ldots, r_j - 1\}$. Suppose now that we have n such random vectors. We can think of the data as a sample from a Multinomial with $N = r_1 \times r_2 \times \cdots \times r_m$ categories. The data can be represented as counts in a $r_1 \times r_2 \times \cdots \times r_m$ table. Let $p = (p_1, \ldots, p_N)$ denote the multinomial parameter.

Let $S = \{1, \ldots, m\}$. Given a vector $x = (x_1, \ldots, x_m)$ and a subset $A \subset S$, let $x_A = (x_j : j \in A)$. For example, if $A = \{1, 3\}$ then $x_A = (x_1, x_3)$.

L. Wasserman, *All of Statistics*

19.1 Theorem. *The joint probability function $f(x)$ of a single random vector $X = (X_1, \ldots, X_m)$ can be written as*

$$\log f(x) = \sum_{A \subset S} \psi_A(x) \tag{19.1}$$

where the sum is over all subsets A of $S = \{1, \ldots, m\}$ and the ψ's satisfy the following conditions:

1. *$\psi_\emptyset(x)$ is a constant;*

2. *For every $A \subset S$, $\psi_A(x)$ is only a function of x_A and not the rest of the x_j's.*

3. *If $i \in A$ and $x_i = 0$, then $\psi_A(x) = 0$.*

The formula in equation (19.1) is called the **log-linear expansion** of f. Each $\psi_A(x)$ may depend on some unknown parameters β_A. Let $\beta = (\beta_A : A \subset S)$ be the set of all these parameters. We will write $f(x) = f(x; \beta)$ when we want to emphasize the dependence on the unknown parameters β.

In terms of the multinomial, the parameter space is

$$\mathcal{P} = \left\{ p = (p_1, \ldots, p_N) : \ p_j \geq 0, \ \sum_{j=1}^{N} p_j = 1 \right\}.$$

This is an $N - 1$ dimensional space. In the log-linear representation, the parameter space is

$$\Theta = \left\{ \beta = (\beta_1, \ldots, \beta_N) : \ \beta = \beta(p), p \in \mathcal{P} \right\}$$

where $\beta(p)$ is the set of β values associated with p. The set Θ is a $N - 1$ dimensional surface in $\mathbb{R}^N$. We can always go back and forth between the two parameterizations we can write $\beta = \beta(p)$ and $p = p(\beta)$.

19.2 Example. Let $X \sim$ Bernoulli(p) where $0 < p < 1$. We can write the probability mass function for X as

$$f(x) = p^x (1-p)^{1-x} = p_1^x p_2^{1-x}$$

for $x = 0, 1$, where $p_1 = p$ and $p_2 = 1 - p$. Hence,

$$\log f(x) = \psi_\emptyset(x) + \psi_1(x)$$

where

$$\begin{aligned}\psi_\emptyset(x) &= \log(p_2)\\ \psi_1(x) &= x\log\left(\frac{p_1}{p_2}\right).\end{aligned}$$

Notice that $\psi_\emptyset(x)$ is a constant (as a function of x) and $\psi_1(x) = 0$ when $x = 0$. Thus the three conditions of Theorem 19.1 hold. The log-linear parameters are

$$\beta_0 = \log(p_2), \quad \beta_1 = \log\left(\frac{p_1}{p_2}\right).$$

The original, multinomial parameter space is $\mathcal{P} = \{(p_1, p_2) : \ p_j \geq 0, p_1 + p_2 = 1\}$. The log-linear parameter space is

$$\Theta = \left\{(\beta_0, \beta_1) \in \mathbb{R}^2 : \ e^{\beta_0+\beta_1} + e^{\beta_0} = 1.\right\}$$

Given (p_1, p_2) we can solve for (β_0, β_1). Conversely, given (β_0, β_1) we can solve for (p_1, p_2). ■

19.3 Example. Let $X = (X_1, X_2)$ where $X_1 \in \{0, 1\}$ and $X_2 \in \{0, 1, 2\}$. The joint distribution of n such random vectors is a multinomial with 6 categories. The multinomial parameters can be written as a 2-by-3 table as follows:

multinomial		x_2	0	1	2
x_1	0		p_{00}	p_{01}	p_{02}
	1		p_{10}	p_{11}	p_{12}

The n data vectors can be summarized as counts:

data		x_2	0	1	2
x_1	0		C_{00}	C_{01}	C_{02}
	1		C_{10}	C_{11}	C_{12}

For $x = (x_1, x_2)$, the log-linear expansion takes the form

$$\log f(x) = \psi_\emptyset(x) + \psi_1(x) + \psi_2(x) + \psi_{12}(x)$$

where

$$\begin{aligned}\psi_\emptyset(x) &= \log p_{00}\\ \psi_1(x) &= x_1\log\left(\frac{p_{10}}{p_{00}}\right)\\ \psi_2(x) &= I(x_2 = 1)\log\left(\frac{p_{01}}{p_{00}}\right) + I(x_2 = 2)\log\left(\frac{p_{02}}{p_{00}}\right)\\ \psi_{12}(x) &= I(x_1 = 1, x_2 = 1)\log\left(\frac{p_{11}p_{00}}{p_{01}p_{10}}\right) + I(x_1 = 1, x_2 = 2)\log\left(\frac{p_{12}p_{00}}{p_{02}p_{10}}\right).\end{aligned}$$

Convince yourself that the three conditions on the ψ's of the theorem are satisfied. The six parameters of this model are:

$$\begin{array}{lll} \beta_1 = \log p_{00} & \beta_2 = \log\left(\frac{p_{10}}{p_{00}}\right) & \beta_3 = \log\left(\frac{p_{01}}{p_{00}}\right) \\ \beta_4 = \log\left(\frac{p_{02}}{p_{00}}\right) & \beta_5 = \log\left(\frac{p_{11}p_{00}}{p_{01}p_{10}}\right) & \beta_6 = \log\left(\frac{p_{12}p_{00}}{p_{02}p_{10}}\right). \end{array}$$

■

The next theorem gives an easy way to check for conditional independence in a log-linear model.

19.4 Theorem. *Let (X_a, X_b, X_c) be a partition of a vectors $(X_1, \ldots, X_m)$. Then $X_b \amalg X_c | X_a$ if and only if all the ψ-terms in the log-linear expansion that have at least one coordinate in b and one coordinate in c are 0.*

To prove this theorem, we will use the following lemma whose proof follows easily from the definition of conditional independence.

19.5 Lemma. *A partition (X_a, X_b, X_c) satisfies $X_b \amalg X_c | X_a$ if and only if $f(x_a, x_b, x_c) = g(x_a, x_b)h(x_a, x_c)$ for some functions g and h*

Proof. (Theorem 19.4.) Suppose that ψ_t is 0 whenever t has coordinates in b and c. Hence, ψ_t is 0 if $t \not\subset a \bigcup b$ or $t \not\subset a \bigcup c$. Therefore

$$\log f(x) = \sum_{t \subset a \bigcup b} \psi_t(x) + \sum_{t \subset a \bigcup c} \psi_t(x) - \sum_{t \subset a} \psi_t(x).$$

Exponentiating, we see that the joint density is of the form $g(x_a, x_b)h(x_a, x_c)$. By Lemma 19.5, $X_b \amalg X_c | X_a$. The converse follows by reversing the argument. ■

19.2 Graphical Log-Linear Models

A log-linear model is **graphical** if missing terms correspond only to conditional independence constraints.

19.6 Definition. *Let $\log f(x) = \sum_{A \subset S} \psi_A(x)$ be a log-linear model. Then f is* **graphical** *if all ψ-terms are nonzero except for any pair of coordinates not in the edge set for some graph $\mathcal{G}$. In other words, $\psi_A(x) = 0$ if and only if $\{i, j\} \subset A$ and (i, j) is not an edge.*

Here is a way to think about the definition above:

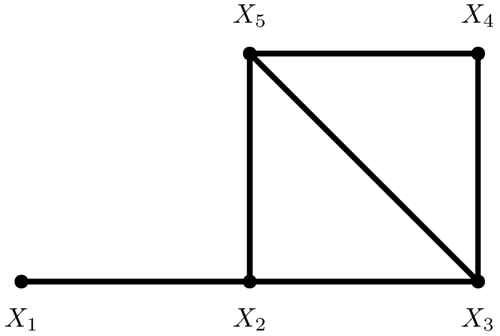

FIGURE 19.1. Graph for Example 19.7.

If you can add a term to the model and the graph does not change, then the model is not graphical.

19.7 Example. Consider the graph in Figure 19.1.

The graphical log-linear model that corresponds to this graph is

$$\begin{aligned}\log f(x) &= \psi_\emptyset + \psi_1(x) + \psi_2(x) + \psi_3(x) + \psi_4(x) + \psi_5(x) \\ &+ \psi_{12}(x) + \psi_{23}(x) + \psi_{25}(x) + \psi_{34}(x) + \psi_{35}(x) + \psi_{45}(x) + \psi_{235}(x) + \psi_{345}(x).\end{aligned}$$

Let's see why this model is graphical. The edge $(1,5)$ is missing in the graph. Hence any term containing that pair of indices is omitted from the model. For example,

$$\psi_{15},\ \psi_{125},\ \psi_{135},\ \psi_{145},\ \psi_{1235},\ \psi_{1245},\ \psi_{1345},\ \psi_{12345}$$

are all omitted. Similarly, the edge $(2,4)$ is missing and hence

$$\psi_{24},\ \psi_{124},\ \psi_{234},\ \psi_{245},\ \psi_{1234},\ \psi_{1245},\ \psi_{2345},\ \psi_{12345}$$

are all omitted. There are other missing edges as well. You can check that the model omits all the corresponding ψ terms. Now consider the model

$$\begin{aligned}\log f(x) &= \psi_\emptyset(x) + \psi_1(x) + \psi_2(x) + \psi_3(x) + \psi_4(x) + \psi_5(x) \\ &+ \psi_{12}(x) + \psi_{23}(x) + \psi_{25}(x) + \psi_{34}(x) + \psi_{35}(x) + \psi_{45}(x).\end{aligned}$$

This is the same model except that the three way interactions were removed. If we draw a graph for this model, we will get the same graph. For example, no ψ terms contain $(1,5)$ so we omit the edge between X_1 and X_5. But this is not graphical since it has extra terms omitted. The independencies and graphs

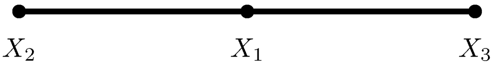

FIGURE 19.2. Graph for Example 19.10.

for the two models are the same but the latter model has other constraints besides conditional independence constraints. This is not a bad thing. It just means that if we are only concerned about presence or absence of conditional independences, then we need not consider such a model. The presence of the three-way interaction ψ_{235} means that the strength of association between X_2 and X_3 varies as a function of X_5. Its absence indicates that this is not so. ■

19.3 Hierarchical Log-Linear Models

There is a set of log-linear models that is larger than the set of graphical models and that are used quite a bit. These are the hierarchical log-linear models.

19.8 Definition. *A log-linear model is* **hierarchical** *if* $\psi_A = 0$ *and* $A \subset B$ *implies that* $\psi_B = 0$.

19.9 Lemma. *A graphical model is hierarchical but the reverse need not be true.*

19.10 Example. Let

$$\log f(x) = \psi_\emptyset(x) + \psi_1(x) + \psi_2(x) + \psi_3(x) + \psi_{12}(x) + \psi_{13}(x).$$

The model is hierarchical; its graph is given in Figure 19.2. The model is graphical because all terms involving (2,3) are omitted. It is also hierarchical. ■

19.11 Example. Let

$$\log f(x) = \psi_\emptyset(x) + \psi_1(x) + \psi_2(x) + \psi_3(x) + \psi_{12}(x) + \psi_{13}(x) + \psi_{23}(x).$$

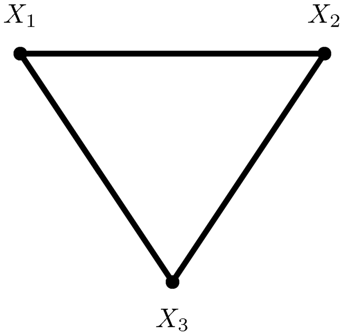

FIGURE 19.3. The graph is complete. The model is hierarchical but not graphical.

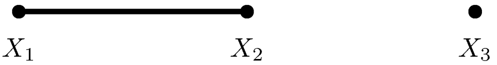

FIGURE 19.4. The model for this graph is not hierarchical.

The model is hierarchical. It is not graphical. The graph corresponding to this model is complete; see Figure 19.3. It is not graphical because $\psi_{123}(x) = 0$ which does not correspond to any pairwise conditional independence. ■

19.12 Example. Let

$$\log f(x) = \psi_\emptyset(x) + \psi_3(x) + \psi_{12}(x).$$

The graph corresponding is in Figure 19.4. This model is not hierarchical since $\psi_2 = 0$ but ψ_{12} is not. Since it is not hierarchical, it is not graphical either. ■

19.4 Model Generators

Hierarchical models can be written succinctly using **generators**. This is most easily explained by example. Suppose that $X = (X_1, X_2, X_3)$. Then, $M = 1.2 + 1.3$ stands for

$$\log f = \psi_\emptyset + \psi_1 + \psi_2 + \psi_3 + \psi_{12} + \psi_{13}.$$

The formula $M = 1.2{+}1.3$ says: "include ψ_{12} and ψ_{13}." We have to also include the lower order terms or it won't be hierarchical. The generator $M = 1.2.3$ is the **saturated** model

$$\log f = \psi_\emptyset + \psi_1 + \psi_2 + \psi_3 + \psi_{12} + \psi_{13} + \psi_{23} + \psi_{123}.$$

The saturated models corresponds to fitting an unconstrained multinomial. Consider $M = 1 + 2 + 3$ which means

$$\log f = \psi_\emptyset + \psi_1 + \psi_2 + \psi_3.$$

This is the mutual independence model. Finally, consider $M = 1.2$ which has log-linear expansion

$$\log f = \psi_\emptyset + \psi_1 + \psi_2 + \psi_{12}.$$

This model makes $X_3 | X_2 = x_2, X_1 = x_1$ a uniform distribution.

19.5 Fitting Log-Linear Models to Data

Let β denote all the parameters in a log-linear model M. The loglikelihood for β is

$$\ell(\beta) = \sum_{i=1}^{n} \log f(X_i; \beta)$$

where $f(X_i; \beta)$ is the probability function for the i^{th} random vector $X_i = (X_{i1}, \ldots, X_{im})$ as give by equation (19.1). The MLE $\widehat{\beta}$ generally has to be found numerically. The Fisher information matrix is also found numerically and we can then get the estimated standard errors from the inverse Fisher information matrix.

When fitting log-linear models, one has to address the following model selection problem: which ψ terms should we include in the model? This is essentially the same as the model selection problem in linear regression.

One approach is is to use AIC. Let M denote some log-linear model. Different models correspond to setting different ψ terms to 0. Now we choose the model M which maximizes

$$\text{AIC}(M) = \widehat{\ell}(M) - |M| \tag{19.2}$$

where $|M|$ is the number of parameters in model M and $\widehat{\ell}(M)$ is the value of the log-likelihood evaluated at the MLE for that model. Usually the model search is restricted to hierarchical models. This reduces the search space. Some also claim that we should only search through the hierarchical models because other models are less interpretable.

A different approach is based on hypothesis testing. The model that includes all possible ψ-terms is called the **saturated model** and we denote it by M_{sat}. Now for each M we test the hypothesis

$$H_0 : \text{the true model is } M \quad \text{versus} \quad H_1 : \text{the true model is } M_{sat}.$$

The likelihood ratio test for this hypothesis is called the deviance.

19.13 Definition. *For any submodel M, define the* **deviance** $\text{dev}(M)$ *by*

$$\text{dev}(M) = 2(\widehat{\ell}_{sat} - \widehat{\ell}_M)$$

where $\widehat{\ell}_{sat}$ is the log-likelihood of the saturated model evaluated at the MLE *and $\widehat{\ell}_M$ is the log-likelihood of the model M evaluated at its* MLE.

19.14 Theorem. *The deviance is the likelihood ratio test statistic for*

$$H_0 : \text{the model is } M \quad \text{versus} \quad H_1 : \text{the model is } M_{sat}.$$

Under H_0, $dev(M) \xrightarrow{d} \chi^2_\nu$ with ν degrees of freedom equal to the difference in the number of parameters between the saturated model and M.

One way to find a good model is to use the deviance to test every sub-model. Every model that is not rejected by this test is then considered a plausible model. However, this is not a good strategy for two reasons. First, we will end up doing many tests which means that there is ample opportunity for making Type I and Type II errors. Second, we will end up using models where we failed to reject H_0. But we might fail to reject H_0 due to low power. The result is that we end up with a bad model just due to low power.

After finding a "best model" this way we can draw the corresponding graph.

19.15 Example. The following breast cancer data are from Morrison et al. (1973). The data are on diagnostic center (X_1), nuclear grade (X_2), and survival (X_3):

	X_2	malignant	malignant	benign	benign
	X_3	died	survived	died	survived
X_1	Boston	35	59	47	112
	Glamorgan	42	77	26	76

The saturated log-linear model is:

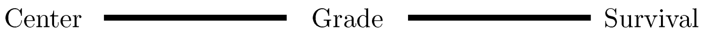

FIGURE 19.5. The graph for Example 19.15.

Variable	$\widehat{\beta}_j$	$\widehat{\mathsf{se}}$	W_j	p-value
(Intercept)	3.56	0.17	21.03	0.00 ***
center	0.18	0.22	0.79	0.42
grade	0.29	0.22	1.32	0.18
survival	0.52	0.21	2.44	0.01 *
center×grade	-0.77	0.33	-2.31	0.02 *
center×survival	0.08	0.28	0.29	0.76
grade×survival	0.34	0.27	1.25	0.20
center×grade×survival	0.12	0.40	0.29	0.76

The best sub-model, selected using AIC and backward searching is:

Variable	$\widehat{\beta}_j$	$\widehat{\mathsf{se}}$	W_j	p-value
(Intercept)	3.52	0.13	25.62	$<$ 0.00 ***
center	0.23	0.13	1.70	0.08
grade	0.26	0.18	1.43	0.15
survival	0.56	0.14	3.98	6.65e-05 ***
center×grade	-0.67	0.18	-3.62	0.00 ***
grade×survival	0.37	0.19	1.90	0.05

The graph for this model M is shown in Figure 19.5. To test the fit of this model, we compute the deviance of M which is 0.6. The appropriate χ^2 has $8 - 6 = 2$ degrees of freedom. The p-value is $\mathbb{P}(\chi^2_2 > .6) = .74$. So we have no evidence to suggest that the model is a poor fit. ■

19.6 Bibliographic Remarks

For this chapter, I drew heavily on Whittaker (1990) which is an excellent text on log-linear models and graphical models. Some of the exercises are from Whittaker. A classic reference on log-linear models is Bishop et al. (1975).

19.7 Exercises

1. Solve for the $p'_{ij}s$ in terms of the β's in Example 19.3.

2. Prove Lemma 19.5.

3. Prove Lemma 19.9.

4. Consider random variables (X_1, X_2, X_3, X_4). Suppose the log-density is

$$\log f(x) = \psi_\emptyset(x) + \psi_{12}(x) + \psi_{13}(x) + \psi_{24}(x) + \psi_{34}(x).$$

   (a) Draw the graph G for these variables.

   (b) Write down all independence and conditional independence relations implied by the graph.

   (c) Is this model graphical? Is it hierarchical?

5. Suppose that parameters $p(x_1, x_2, x_3)$ are proportional to the following values:

	x_2	0	0	1	1
	x_3	0	1	0	1
x_1	0	2	8	4	16
	1	16	128	32	256

   Find the ψ-terms for the log-linear expansion. Comment on the model.

6. Let $X_1, \ldots, X_4$ be binary. Draw the independence graphs corresponding to the following log-linear models. Also, identify whether each is graphical and/or hierarchical (or neither).

   (a) $\log f = 7 + 11x_1 + 2x_2 + 1.5x_3 + 17x_4$

   (b) $\log f = 7 + 11x_1 + 2x_2 + 1.5x_3 + 17x_4 + 12x_2x_3 + 78x_2x_4 + 3x_3x_4 + 32x_2x_3x_4$

   (c) $\log f = 7 + 11x_1 + 2x_2 + 1.5x_3 + 17x_4 + 12x_2x_3 + 3x_3x_4 + x_1x_4 + 2x_1x_2$

   (d) $\log f = 7 + 5055x_1x_2x_3x_4$