CHAPTER 4
Contingency Tables

A contingency table is used to show cross-classified categorical data on two or more variables. The variables can be *nominal* or *ordinal*. A nominal variable has categories with no natural ordering; for example, consider the automotive companies Ford, General Motors and Toyota. An ordering could be imposed using some criterion like sales, but there is nothing inherent in the categories that makes any particular ordering obvious. An ordinal variable does have a natural default ordering. For example, a disease might be recorded as absent, mild or severe. The five-point Likert scale ranging through strongly disagree, disagree, neutral, agree and strongly agree is another example.

An *interval scale* is an ordinal variable that has categories with a distance measure. This is often the result of continuous data that has been discretized into intervals. For example, age groups 0–18, 18–34, 34–55 and 55+ might be used to record age information. If the intervals are relatively wide, then methods for ordinal data can be used where the additional information about the intervals may be useful in the modeling. If the intervals are quite narrow, then we could replace interval response with the midpoint of the interval and then use continuous data methods. One could argue that all so-called continuous data is of this form, because such data cannot be measured with arbitrary precision. Height might be given to the nearest centimeter, for example.

4.1 Two-by-Two Tables

The data shown in Table 4.1 were collected as part of a quality improvement study at a semiconductor factory. A sample of wafers was drawn and cross-classified according to whether a particle was found on the die that produced the wafer and whether the wafer was good or bad. More details on the study may be found in Hall (1994). The data might have arisen under several possible sampling schemes:

Quality	No Particles	Particles	Total
Good	320	14	334
Bad	80	36	116
Total	400	50	450

Table 4.1 *Study of the relationship between wafer quality and the presence of particles on the wafer.*

1. We observed the manufacturing process for a certain period of time and observed 450 wafers. The data were then cross-classified. We could use a Poisson model.

2. We decided to sample 450 wafers. The data were then cross-classified. We could use a multinomial model.
3. We selected 400 wafers without particles and 50 wafers with particles and then recorded the good or bad outcome. We could use a binomial model.
4. We selected 400 wafers without particles and 50 wafers with particles that also included, by design, 334 good wafers and 116 bad ones. We could use hypergeometric model.

The first three sampling schemes are all plausible. The fourth scheme seems less likely in this example, but we include it for completeness. Such a scheme is more attractive when one level of each variable is relatively rare and we choose to oversample both levels to ensure some representation.

The main question of interest concerning these data is whether the presence of particles on the wafer affects the quality outcome. We shall see that all four sampling schemes lead to exactly the same conclusion. First, let's set up the data in a convenient form for analysis:

```
> y <- c(320,14,80,36)
> particle <- gl (2,1,4,labels=c("no","yes"))
> quality <- gl(2,2,labels=c("good","bad"))
> wafer <- data.frame(y,particle,quality)
> wafer
    y particle quality
1 320       no    good
2  14      yes    good
3  80       no     bad
4  36      yes     bad
```

We will need the data in this form with one observation per line for our model fitting, but usually we prefer to observe it table form:

```
> (ov <- xtabs(y ~ quality+particle))
       particle
quality no  yes
   good 320  14
   bad   80  36
```

Poisson Model

Suppose we assume that the process is observed for some period of time and we count the number of occurrences of the possible outcomes. It would be natural to view these outcomes occurring at different rates and that we could form Poisson model for these rates. Suppose we fit an additive model:

```
> modl <- glm(y ~ particle+quality, poisson)
> summary(modl)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.6934     0.0572   99.54   <2e-16
particleyes  -2.0794     0.1500  -13.86   <2e-16
```

```
qualitybad   -1.0576     0.1078   -9.81   <2e-16
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 474.10 on 3 degrees of freedom
Residual deviance:  54.03 on 1 degrees of freedom
```

The null model, which suggests all four outcomes occur at the same rate, does not fit because the deviance of 474.1 is very large for three degrees of freedom. The additive model, with a deviance of 54.03 is clearly an improvement over this. We might also want to test the significance of the individual predictors. We could use the *z*-values, but it is better to use the likelihood ratio test based on the differences in the deviance (not that it matters much for this particular dataset):

```
> dropl(modl,test="Chi")
Single term deletions
Model:
y ~ particle + quality
         Df Deviance AIC LRT Pr(Chi)
<none>            54  84
particle  1      364 392 310  <2e-16
quality   1      164 192 110  <2e-16
```

We see that both predictors are significant relative to the full model. By examining the coefficients, we see that wafers with particles occur at a significantly higher rate than wafers without particles. Similarly, we see that bad-quality wafers occur at a significantly higher rate than good-quality wafers.

The model coefficients are closely related to the marginal totals in the table. The maximum likelihood estimates satisfy:

$$X^T y = X^T \hat{\mu}$$

where the $X^T y$ is, in this example:

```
> (t(model.matrix(modl)) %*%y)[,]
(Intercept) particleyes qualitybad
        450          50        116
```

So we see that the fitted values, $\hat{\mu}$, are a function of marginal totals. This fact is exploited in an alternative fitting method known as *iterative proportional fitting*. The glm function in R, however, uses Fisher scoring, described in Section 6.2. In any case, the log-likelihood (ignoring any terms not involving μ) is:

$$\log L = \sum_i y_i \log \mu_i$$

which is maximized to obtain the fit.

The analysis so far has told us nothing about the relationship between the presence of particles and the quality of the wafer. The additive model posits:

$\log \mu = \gamma + \alpha_i + \beta_j$

where α represents the particle effect and β represents the quality outcome and i, j= 1,2. γ is the intercept term. Due to the log link, the predicted rate for the response in any cell in the table is formed from the product of the rates for the corresponding levels of the two predictors. There is no interaction term and so good- or bad-quality outcomes occur independently of whether a particle was found on the wafer. This model has a deviance of 54.03 on one degree of freedom and so does not fit the data.

The addition of an interaction term would saturate the model and so would have zero deviance and degrees of freedom. So an hypothesis comparing the models with and without interaction would use a test statistic of 54.03 on one degree of freedom. The hypothesis of no interaction would be rejected.

Multinomial Model

Suppose we assume that the total sample size was fixed at 450 and that the frequency of the four possible outcomes was recorded. In these circumstances, it is natural to use a multinomial distribution to model the response. Let y_{ij} be the observed response in cell *(i, j)* and let p_{ij} be the probability that an observation falls in that cell and let n be the sample size. The probability of the observed response under the multinomial is then:

$$\frac{n!}{\prod_i \prod_j y_{ij}} \prod_i \prod_j p_{ij}^{y_{ij}}$$

Now the p_{ij} will be linked to the predictor information according to the model we choose. To estimate the parameters, we would maximize the log-likelihood:

$$\log L = \sum_i \sum_j y_{ij} \log p_{ij}$$

where terms not involving p_{ij} are ignored. Notice that this takes essentially the same form as for the Poisson model above.

The main hypothesis of interest is whether the quality and presence of a particle on the wafer are independent. Let p_i for i=1,2 be the probabilities of the two quality outcomes and p_j for j=1,2 be the probability of the two particle categories. Let P_{ij} be the probability of a particular joint outcome. Under independence, $p_{ij}=p_i p_j$. Using the fact that probabilities must sum to one, the maximum likelihood estimates are:

$$\hat{p}_i = \sum_j y_{ij}/n \qquad \text{and} \qquad \hat{p}_j = \sum_i y_{ij}/n$$

We can compute these for the wafer data as, respectively:

```
> (pp <- prop.table( xtabs(y ~ particle)))
particle
     no     yes
0.88889 0.11111
> (qp <- prop.table( xtabs(y ~ quality)))
quality
   good     bad
0.74222 0.25778
```

The fitted values are then $\hat{\mu}_{ij} = np_i p_j = \sum_i y_{ij} \sum_j y_{ij}/n$ or:

```
> (fv <- outer(qp,pp)*450)
         particle
quality     no    yes
good    296.89 37.111
bad 103.11 12.889
```

To test the fit, we compare this model against the saturated model, for which $\hat{\mu}_{ij} = y_{ij}$. So the deviance is:

$$2\sum_i \sum_j y_{ij} \log(y_{ij}/\mu_{ij})$$

which computes to:

```
> 2*sum(ov*log(ov/fv))
[1] 54.03
```

which is the same deviance we observed in the Poisson model. So we see that the test for independence in the multinomial model coincides with the test for no interaction in the Poisson model. The latter test is easier to execute in R, so we shall usually take that approach.

This connection between the Poisson and multinomial is no surprise due to the following result. Let $Y_1,\ldots,Y_k$ be independent Poisson random variables with means $\lambda_1,\ldots,\lambda_k$, then the joint distribution of $Y_1,\ldots, Y_k|\Sigma_i Y_i=n$ is multinomial with probabilities $p_j=\lambda_j/\Sigma_i\lambda_i$.

One alternative to the deviance is the Pearson X^2 statistic:

$$X^2 = \sum_{i,j} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

which takes the value:

```
> sum( (ov-fv)^2/fv)
[1] 62.812
```

Yates' continuity correction subtracts 0.5 from $y_{ij} - \hat{\mu}_{ij}$ when this value is positive and adds 0.5 when it is negative. This gives superior results for small samples. This correction is implemented in:

```
> prop.test(ov)
        2-sample test for equality of proportions with
        continuity correction
data: ov
X-squared = 60.124, df = 1, p-value = 8.907e-15
```

The deviance-based test is preferred to the Pearson's X^2.

Binomial

It would also be natural to view the presence of the particle as affecting the quality of wafer. We would view the quality as the response and the particle status as a predictor. We might fix the number of wafers with no particles at 400 and the number with particles as 50 and then observe the outcome. We could then use a binomial model for the response for both groups. Let's see what happens:

```
> (m <- matrix(y,nrow=2))
     [,1] [,2]
[1,]  320   80
[2,]   14   36
> modb <- glm(m ~ 1, family=binomial)
> deviance(modb)
[1] 54.03
```

We fit the null model which suggests that the probability of the response is the same in both the particle and no particle group. This hypothesis of *homogeneity* corresponds exactly to the test of independence and the deviance is exactly the same.

For larger contingency tables, where there are more than two rows (or columns), we can use a multinomial model for each row. This model is more accurately called a *product* multinomial model to distinguish it from the unrestricted multinomial model introduced above.

Hypergeometric

The remaining case is where both marginal totals are fixed. This situation is rather less common in practice, but does suggest a more accurate test for independence. This sampling scheme can arise when classifying objects into one of two types when the true proportions of each type are known in advance. For example, suppose you are given 10 true or false statements and told that 5 are true and 5 are false. You are asked to sort the statements into true and false. We can generate a two-by-two table of the correct classification against the observed classification generated. Under the hypergeometric distribution and the assumption of independence, the probability of the observed table is:

$$\frac{(y_{11}+y_{12})!(y_{11}+y_{21})!(y_{12}+y_{22})!(y_{21}+y_{22})!}{y_{11}!y_{12}!y_{21}!y_{22}!n!}$$

If we fix any number in the table, say y_{11}, the remaining three numbers are completely determined because the row and column totals are known. There is a limited number of values which y_{11} can possibly take and we can compute the probability of all these outcomes. Specifically, we can compute the total probability of all outcomes more extreme than the one observed. This method is called *Fisher's exact test*. We may execute it as follows:

```
> fisher.test(ov)
        Fisher's Exact Test for Count Data
data:  ov
```

```
p-value = 2.955e-13
alternative hypothesis: true odds ratio is not equal to
1
95 percent confidence interval:
  5.0906 21.5441
sample estimates:
odds ratio
    10.213
```

Notice that the odds ratio, which is $\log(y_{11}y_{22})/(y_{12}y_{21})$, takes the value:

```
> (320*36)/(14*80)
[1] 10.286
```

and is a measure of the association for which an exact confidence interval may be calculated as we see in the output.

Fisher's test is attractive because the null distribution for the deviance and Pearson's χ^2 test statistics is only approximately χ^2 distributed. This approximation is particularly suspect for tables with small counts making an exact method valuable. The Fisher test becomes more difficult to compute for larger tables and some approximations may be necessary. However, for larger tables, the χ^2 approximation will tend to be very accurate.

4.2 Larger Two-Way Tables

Snee (1974) presents data on 592 students cross-classified by hair and eye color.

```
> data(haireye)
> haireye
     y   eye  hair
1    5 green BLACK
2   29 green BROWN
..etc..
16   7 brown BLOND
```

The data is more conveniently displayed using:

```
> (ct <- xtabs(y ~ hair + eye, haireye))
       eye
hair    green hazel blue brown
  BLACK   5    15    20   68
  BROWN  29    54    84  119
  RED    14    14    17   26
  BLOND  16    10    94    7
```

We can execute the usual Pearson's χ^2 test for independence as:

```
> summary(ct)
```

```
Call: xtabs(formula = y ~ hair + eye, data = haireye)
Number of cases in table: 592
Number of factors: 2
Test for independence of all factors:
        Chisq = 138, df = 9, p-value = 2.3e-25
```

where we see that hair and eye color are clearly not independent.

One option for displaying contingency table data is the *dotchart*:

```
> dotchart (ct)
```

which may be seen in the first panel of Figure 4.1. The mosaic plot, described in Hartigan and Kleiner (1981), divides the plot region according to the frequency of each level in a recursive manner:

```
> mosaicplot(ct,color=TRUE,main=NULL,las=1)
```

In the plot shown in the second panel of Figure 4.1, the area is first divided according to the frequency of hair color. Within each hair color, the area is then divided according to the frequency of eye color. A different plot could be constructed by reversing the order of hair and eye in the xtabs command above. We can now readily see the frequency of various outcomes. We see, for example, that brown hair and brown

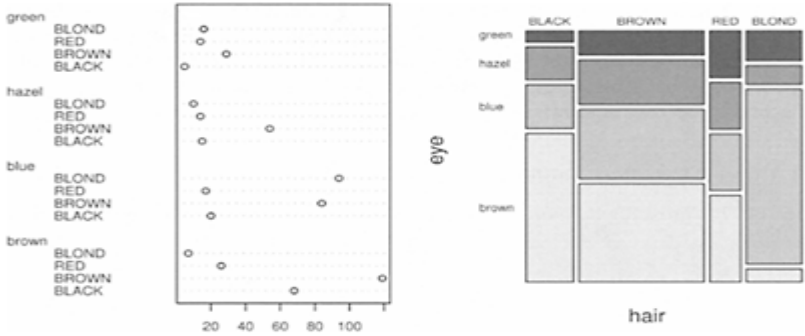

Figure 4.1 *Dotchart and Mosaic Plot*

eyes is the most common combination while green eyes and black hair is the least common.

Now we fit the Poisson GLM:

```
> modc <- glm(y ~ hair+eye,family=poisson,haireye)
> summary(modc)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.458      0.152   16.14   <2e-16
hair BROWN     0.974      0.113    8.62   <2e-16
```

```
hairRED       -0.419      0.153   -2.75    0.006
hairBLOND      0.162      0.131    1.24    0.216
eyehazel       0.374      0.162    2.30    0.021
eyeblue        1.212      0.142    8.51   <2e-16
eyebrown       1.235      0.142    8.69   <2e-16
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 453.31 on 15 degrees of freedom
Residual deviance: 146.44 on 9 degrees of freedom
AIC: 241.0
```

We see that most of the levels of hair and eye color show up as significantly different from the reference levels of black hair and green eyes. But this merely indicates that there are higher numbers of people with some hair colors than others and some eye colors than others. We already know this. We are more interested in the relationship between hair and eye color. The deviance of 146.44 on nine degrees freedom shows that they are clearly dependent. This does not tell us how they are dependent. To study this, we can use a kind of residual analysis for contingency tables called *correspondence analysis*.

Compute the Pearson residuals *rp* and write them in the matrix form R_{ij}, where i= 1,…, r and j=1,…, c, according to the structure of the data. Perform the singular value decomposition:

$$R_{r\times c} = U_{r\times w} D_{w\times w} V^T_{w\times c}$$

where r is the number of rows, c is the number of columns and w=min(r, c). U and V are called the right and left singular vectors, respectively. D is a diagonal matrix with sorted elements d_i, called *singular values*. Another way of writing this is:

$$R_{ij} = \sum_{k=1}^{w} U_{ik} d_k V_{jk}$$

As with eigendecompositions, it is not uncommon for the first few singular values to be much larger than the rest. Suppose that the first two dominate so that:

$R_{ij} \approx U_{i1} d_1 V_{j1} + U_{i2} d_2 V_{j2}$

We usually absorb the *ds* into *U* and *V* for plotting purposes so that we can assess the relative contribution of the components. Thus:

$$\begin{aligned} R_{ij} &\approx (U_{i1}\sqrt{d_1}) \times (V_{j1}\sqrt{d_1}) + (U_{i2}\sqrt{d_2}) \times (V_{j2}\sqrt{d_2}) \\ &\equiv U_{i1}V_{j1} + U_{i2}V_{j2} \end{aligned}$$

where in the latter expression we have redefined the *Us* and *Vs* to include the $\sqrt{d}$.

The two-dimensional correspondence plot displays U_{i2} against U_{i1} and V_{j2} against V_{j1} on the same graph. So the points on the plot will either represent a row level *(U)* or a column level *(V)*. We compute the plot for the hair and eye color data:

```
> z <- xtabs(residuals(mode, type="pearson")~hair+eye,
haireye)
> svdz <- svd(z,2,2)
```

```
> leftsv <- svdz$u %*% diag(sqrt(svdz$d[1:2]))
> rightsv <- svdz$v %*% diag(sqrt(svdz$d[1:2]))
> 11 <- 1.1*max(abs(rightsv),abs(leftsv))
> plot(rbind(leftsv,rightsv),asp=1,xlim=c(-
11,11),ylim=c(-11,11),
  xlab="SVl",ylab="SV2",type="n")
> abline(h=0,v=0)
> text(leftsv,dimnames(z)[[1]])
> text(rightsv,dimnames(z)[[2]])
```

The plot is shown in Figure 4.2. The correspondence analysis plot can be interpreted in light of the following observations:

- $\sum d_i^2 =$ Pearson's X^2 is called the inertia. When $r = c$, d_i^2 are the eigenvalues of R.
- Look for large values of |*Ui*| indicating that the row *i* profile is different. For example, the point for blonds in Figure 4.2 is far from the origin indicating that the distribution of eye colors within this group of people is not typical. In contrast, we see that the point for people with brown hair is close to the origin, indicating an eye color distribution that is close to the overall average. The same type of observation is true for the columns, |*Vj*|. Points distant from the origin mean that the level associated with the column *j* profile is different in some way.

    - If row and column levels appear close together on the plot and far from the origin, we can see that there will be a large positive residual associated with this particular combination indicating a strong positive association. For example, we see that blue eyes and blond hair occur close together on the plot and far from the origin indicating a strong association. On the other hand, if the two points are situated diametrically apart on either side of the origin, we may expect a large negative residual indicating a strong negative association. For example, there are relatively fewer people with blond hair and brown eyes than would be expected under independence.
- If points representing two rows or two column levels are close together, this indicates that the two levels will have a similar pattern of association. In some cases, one might consider combining the two levels. For example, people with hazel or green eyes have similar hair color distributions and we might choose to combine these two categories.
- Because the distance between points is of interest, it is important that the plot is scaled so that the visual distance is proportionately correct. This does not happen automatically, because the default behavior of plots is to fill the plot region out to the specified aspect ratio.

There are several competing ways to construct contingency tables. See Venables and Ripley (2002) who provide the function corresp in the MASS package. See also Blasius and Greenacre (1998) for a survey of methods for visualizing categorical data.

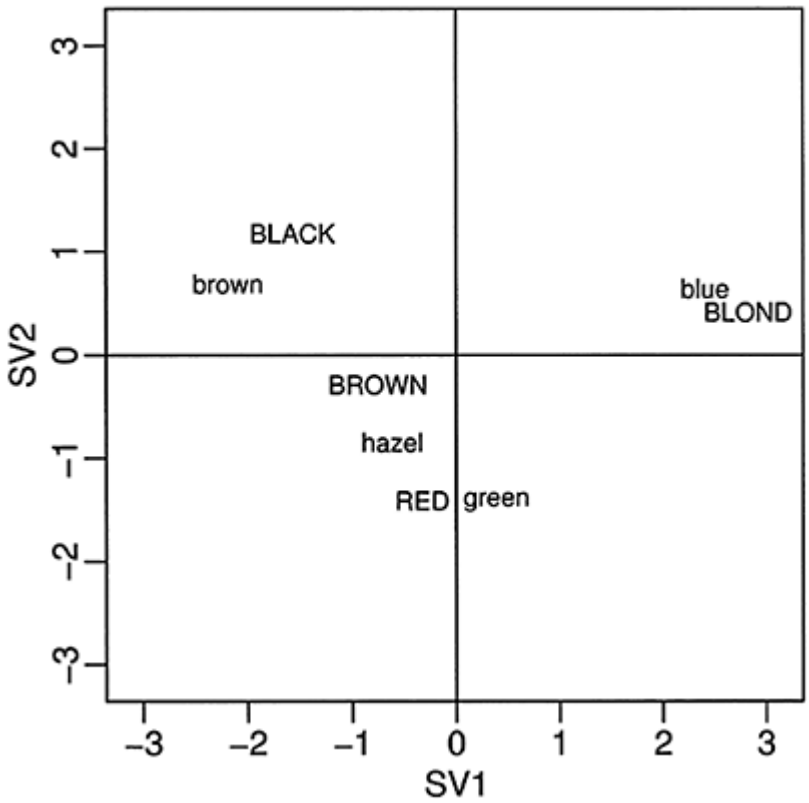

Figure 4.2 *Correspondence analysis for hair-eye combinations. Hair colors are given in upper-case letters and eye colors are given in lower-case letters.*

4.3 Matched Pairs

In the typical two-way contingency tables, we display accumulated information about two categorical measures on the same object. In matched pairs, we observe one measure on two matched objects.

In Stuart (1955), data on the vision of a sample of women is presented. The left and right eye performance is graded into four categories:

```
> data(eyegrade)
> (ct <- xtabs(y ~ right+left, eyegrade))
        left
```

```
right    best second third worst
  best   1520  266    124    66
  second  234 1512    432    78
  third   117  362   1772   205
  worst    36   82    179   492
```

If we check for independence:

```
> summary(ct)
Call: xtabs(formula = y ~ right + left, data =
eyegrade)
Number of cases in table: 7477
Number of factors: 2
Test for independence of all factors:
        Chisq = 8097, df = 9, p-value = 0
```

We are not surprised to find strong evidence of dependence. A more interesting hypothesis for such matched pair data is symmetry. Is $p_{ij}=p_{ij}$? We can fit such a model by defining a factor where the levels represent the symmetric pairs for the off-diagonal elements. There is only one observation for each level down the diagonal:

```
> (symfac <- factor (apply(eyegrade[,2:3],1,
  function(x) paste (sort(x), collapse="-"))))
[1] best-best      best-second    best-third    best-
worst
[5] best-second    second-second  second-third  second-
worst
[9] best-third     second-third   third-third   third-
worst
[13] best-worst     second-worst   third-worst   worst-
worst
10 Levels: best-best  best-second  best-third ...
worst-worst
```

We now fit this model:

```
> mods <- glm(y ~ symfac, eyegrade, family=poisson)
> c(deviance(mods),df.residual(mods))
[1] 19.249   6.000
> pchisq (deviance(mods),df.residual(mods),lower=F)
[1] 0.0037629
```

Here, we see evidence of a lack of symmetry. It is worth checking the residuals:

```
> round(xtabs(residuals(mods) ~ right+left,
eyegrade),3)
        left
right    best   second  third  worst
  best    0.000  1.001   0.317  2.008
  second -1.023  0.000   1.732 -0.225
```

```
third  -0.320 -1.783   0.000  0.928
worst  -2.219  0.223  -0.949  0.000
```

We see that the residuals above the diagonal are mostly positive, while they are mostly negative below the diagonal. So there are generally more poor left, good right eye combinations than the reverse. Furthermore, we can compute the marginals:

```
> margin.table(ct,1)
right
  best second  third  worst
  1976   2256   2456    789
> margin.table(ct,2)
left
  best second  third  worst
  1907   2222   2507    841
```

We see that there are somewhat more poor left eyes and good right eyes, so perhaps marginal homogeneity does not hold here. The assumption of symmetry implies marginal homogeneity (the reverse is not necessarily true). We may observe data where there is a difference in the frequencies of the levels of the rows and columns, but still be interested in symmetry. Suppose we set:

$P_{ij}=\alpha_i\beta_j\gamma_{ij}$

where $\gamma_{ij}=\gamma_{ji}$. This will allow for some symmetry while allowing for different marginals. This is the *quasi-symmetry* model. Now:

$\log EY_{ij}=\log np_{ij}=\log n+\log \alpha_i+\log \beta_j+\log \gamma_{ij}$

So we can fit this model using:

```
> modq <- glm(y ~ right+left+symfac, eyegrade,
family=poisson)
> pchisq(deviance(modq),df.residual(modq),lower=F)
[1] 0.06375
```

We see that this model does fit. It can be shown that marginal homogeneity together with quasi-symmetry implies symmetry. One can test for marginal homogeneity by comparing the symmetry and quasi-symmetry models:

```
> anova(mods,modq,test="Chi")
Analysis of Deviance Table
Model 1: y ~ symfac
Model 2: y ~ right + left + symfac
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1   6   19.25
2   3    7.27   3   11.98   0.01
```

So we find evidence of a lack of marginal homogeneity. This test is only appropriate if quasi-symmetry already holds.

When we examine the data here, we do see that many people do have symmetric vision. These entries lie down the diagonal. We might ask whether there is independence between left and right eyes among those people whose vision is not symmetric. This is the *quasi-independence* hypothesis and we can test it by omitting the data from the diagonal:

```
> modqi <- glm(y ~ right+left, eyegrade,
family=poisson,
  subset=-c(1,6,11,16))
> pchisq(deviance(modqi),df.residual(modqi),lower=F)
[1] 4.4118e-41
```

This model does not fit. This is not surprising since we can see that the entries adjacent to the diagonal are larger than those further away. The difference in vision between the two eyes is likely to be smaller than expected under independence.

4.4 Three-Way Contingency Tables

In Appleton, French, and Vanderpump (1996), a 20-year follow-up study on the effects of smoking is presented. In the period 1972-74, a larger study, which also considered other issues, categorized women into smokers and nonsmokers and according to their age group. In the follow-up, the researchers recorded whether the subjects were dead or still alive. Only smokers or women who had never smoked are presented here. Relatively few smokers quit and these women have been excluded from the data. The cause of death is not reported here. Here is the data:

```
> data(femsmoke)
> femsmoke
     y smoker dead   age
1    2    yes  yes 18-24
2    1     no  yes 18-24
3    3    yes  yes 25-34
...
28   0     no   no   75+
```

We can combine the data over age groups to produce:

```
> (ct <- xtabs(y ~ smoker+dead, femsmoke))
      dead
smoker yes no
   yes 139 443
    no 230 502
```

We can compute the proportions of dead and alive for smokers and nonsmokers:

```
> prop.table(ct,1)
      dead
```

```
smoker yes     no
   yes 0.23883 0.76117
    no 0.31421 0.68579
```

We see that 76% of smokers have survived for 20 years while only 69% of nonsmokers have survived. Thus smoking appears to have a beneficial effect on longevity. We can check the significance of this difference:

```
> summary(ct)
Call: xtabs(formula = y ~ smoker + dead, data =
femsmoke)
Number of cases in table: 1314
Number of factors: 2
Test for independence of all factors:
        Chisq = 9.1, df = 1, p-value = 0.0025
```

So the difference cannot be reasonably ascribed to chance variation. However, if we consider the relationship within a given age group, say 55–64:

```
> (cta <- xtabs(y ~ smoker+dead, femsmoke,
subset=(age=="55-64")))
      dead
smoker yes no
   yes 51  64
   no  40  81
> prop.table(cta,1)
      dead
smoker yes     no
   yes 0.44348 0.55652
   no  0.33058 0.66942
```

We see that 56% of the smokers have survived compared to 67% of the nonsmokers. This advantage to nonsmokers holds throughout all the age groups. Thus the *marginal* association where we add over the age groups is different from the *conditional* association observed within age groups. Data where this effect is observed are an example of *Simpson's paradox*. The paradox is named after Simpson (1951), but dates back to Yule (1903).

Let's see why the effect occurs here:

```
> prop.table(xtabs(y ~ smoker+age, femsmoke),2)
      age
smoker 18-24   25-34   35-44   45-54   55-64   65-
74   75+
   yes 0.47009 0.44128 0.47391 0.62500 0.48729 0.21818
0.16883
   no  0.52991 0.55872 0.52609 0.37500 0.51271 0.78182
0.83117
```

We see that smokers are more concentrated in the younger age groups and younger people are more likely to live for another 20 years. This explains why the marginal table gave an apparent advantage to smokers which is, in fact, illusory because once we control for age, we see that smoking has a negative effect on longevity.

It is interesting to note that the dependence in the 55–64 age group is not statistically significant:

```
> fisher.test(cta)
        Fisher's Exact Test for Count Data
data:  cta
p-value = 0.08304
alternative hypothesis: true odds ratio is not equal to
1
95 percent confidence interval:
0.92031 2.83340
sample estimates:
odds ratio
    1.6103
```

However, this is just a subset of the data. Suppose we compute the odds ratios in all the age groups:

```
> ct3 <- xtabs(y ~ smoker+dead+age,femsmoke)
> apply(ct3, 3, function(x)
(x[1,1]*x[2,2])/(x[1,2]*x[2,1]))
  18-24   25-34   35-44   45-54   55-64   65-74     75+
2.30189 0.75372 2.40000 1.44175 1.61367 1.14851     NaN
```

We see that there is some variation in the odds ratio, but they are all greater than one with the exception of the 25–34 age group. We could test for independence in each 2×2 table, but it is better to use a combined test. The Mantel-Haenszel test is designed to test independence in 2×2 tables across K strata. It only makes sense to use this test if the relationship is similar in each stratum. For this data, the observed odds ratios do not vary greatly, so the use of the test is justified.

Let the entries in the 2×2×K table be y_{ijk}. If we assume a hypergeometric distribution in each 2×2 table, then y_{11k} is sufficient for each table given that we assume that the marginal totals for each table carry no information. The Mantel-Haenszel statistic is:

$$\frac{(|\sum_k y_{11k} - \sum_k E y_{11k}| - 1/2)^2}{\sum_k \text{var}\, y_{11k}}$$

where the expectation and variance are computed under the null hypothesis of independence in each stratum. The statistic is approximately χ^2_1 distributed under the null, although it is possible to make an exact calculation for smaller datasets. The statistic as stated above is due to Mantel and Haenszel (1959), but a version without the half continuity correction was published by Cochran (1954). For this reason, it is sometimes known as the Cochran-Mantel-Haenszel statistic.

We compute the statistic for the data here:

```
> mantelhaen.test(ct3,exact=TRUE)
        Exact conditional test of independence in 2 × 2
× k
        tables
data:  ct3
S = 139, p-value = 0.01591
alternative hypothesis: true common odds ratio is not
equal to 1
95 percent confidence interval:
1.0689 2.2034
sample estimates:
common odds ratio
           1.5303
```

We used the exact method in preference to the approximation. We see that a statistically significant association is revealed once we combine the information across strata.

Now let's consider a linear models approach to investigating how three factors interact. Let p_{ijk} be the probability that an observation falls into the *(i, j, k)* cell. Let p_i be the marginal probability that the observation falls into the i^{th} cell of the first variable, p_j be the marginal probability that the observation falls into the j^{th} cell of the second variable and p_k be the marginal probability that the observation falls into the k^{th} cell of the third variable.

Mutual Independence: If all three variables are independent, then:

$P_{ijk}=P_iP_jP_k$

Now $EY_{ijk}=np_{ijk}$ so:

$\log EY_{ijk}=\log n+\log p_i+\log p_j+\log p_k$

So the main effects-only model corresponds to mutual independence. The coding we use will determine exactly how the parameters relate to the margin totals of the table although typically we will not be especially interested in these. Since independence is the simplest possibility, this model is the null model in an investigation of this type. The model $\log EY_{ijk}=\mu$ would suggest that all the cells have equal probability. It is rare that such a model would have any interest so the model above makes for a more appropriate null.

We can test for independence using the Pearson's χ^2 test:

```
> summary(ct3)
Call: xtabs(formula = y ~ smoker + dead + age, data =
femsmoke)
Number of cases in table: 1314
Number of factors: 3
Test for independence of all factors:
        Chisq = 791, df = 19, p-value = 2.1e-155
```

We can also fit the appropriate linear model:

```
> modi <- glm(y ~ smoker + dead + age, femsmoke,
family=poisson)
```

```
> c(deviance(modi),df.residual(modi))
[1] 735 19
```

Although the statistics for the two tests are somewhat different, in either case, we see a very large value for the degrees of freedom. We conclude that this model does not fit the data.

We can show that the coefficients of this model correspond to the marginal proportions. For example, consider the smoker factor:

```
> (coefsmoke <- exp(c(0,coef(modi)[2])))
         smokerno
1.0000     1.2577
> coefsmoke/sum(coefsmoke)
         smokerno
0.44292   0.55708
```

We see that these are just the marginal proportions for the smokers and nonsmokers in the data:

```
> prop.table(xtabs(y ~ smoker, femsmoke))
smoker
    yes      no
0.44292 0.55708
```

This just serves to emphasize the point that the main effects of the model just convey information that we already know and is not the main interest of the study.

Joint Independence: Let p_{ij} be the (marginal) probability that the observation falls into a *(i, j, ·)* cell where any value of the third variable is acceptable. Now suppose that the first and second variable are dependent, but jointly independent of the third. Then:

$P_{ijk}=P_{ij}P_k$

We can represent this as:

$logEY_{ijk}=\log n+\log p_{ij}+\log pk$

Using the hierarchy principle, we would also include the main effects corresponding to the interaction term log $p_{ij.}$ So the log-linear model with just one interaction term corresponds to joint independence. The specific interaction term tells us which pair of variables is dependent. For example, we fit a model that says age is jointly independent smoking and life status:

```
> modj <- glm(y ~ smoker*dead + age, femsmoke,
family=poisson)
> c(deviance(modj), df.residual(modj))
[1] 725.8  18.0
```

Although this represents a small improvement over the mutual independence model, the deviance is still very high for the degrees of freedom and it is clear that this model does

not fit the data. There are two other joint independence models that have the two other interaction terms. These models also fit badly.

Conditional Independence: Let $p_{ij|k}$ be the probability that an observation falls in cell (i,j,.) given that we know the third variable takes the value *k*. Now suppose we assert that the first and second variables are independent given the value of the third variable, then:

$P_{ij}|_k=P_i|_kP_j|_k$

which leads to:

$P_{ijk}=P_{ik}P_{jk}|P_k$

This results in the model:

$\log EY_{ijk}=\log n+\log pik+ \log p_{jk}-\log p_k$

Again, using the hierarchy principle, we would also include the main effects corresponding to the interaction terms and we would have model with main effects and two interaction terms. The minus for the log p_k term is irrelevant. The nature of the conditional independence can be determined by observing which of one of the three possible two-way interactions does not appear in the model.

The most plausible conditional independence model for our data is:

```
> modc <- glm(y ~ smoker*age + age*dead, femsmoke,
family=poisson)
> c(deviance(mode),df.residual(mode))
[1] 8.327 7.000
```

We see that the deviance is only slightly larger than the degrees of freedom indicating a fairly good fit. This indicates that smoking is independent of life status given age. However, bear in mind that we do have some zeroes and other small numbers in the table and so there is some doubt as to the accuracy of the χ^2 approximation here. It is generally better to compare models rather than assess the goodness of fit.

Uniform Association: We might consider a model with all three-way interactions:

$\log EY_{ijk}=\log n+\log p_i+\log p_j+\log p_k+\log p_{ij}+\log p_{ik}+\log P_{jk}$

The model has no three-way interaction and so it is not saturated. There is no simple interpretation in terms of independence. Consider our example:

```
> modu <- glm(y ~ (smoker+age+dead)^2, femsmoke,
family=poisson)
```

Now we compute the fitted values and determine the odds ratios for each age group based on these fitted values:

```
> ctf <- xtabs(fitted(modu) ~ smoker+dead+age,
femsmoke)
> apply(ctf, 3, function(x)
(x[1,1]*x[2,2])/(x[1,2]*x[2,1]) )
```

```
18-24  25-34  35-44  45-54  55-64  65-74    75+
1.5333 1.5333 1.5333 1.5333 1.5333 1.5333 1.5333
```

We see that the odds ratio is the same for every age group. Thus the uniform association model asserts that for every level of one variable, we have the same association for the other two variables.

The information may also be extracted from the coefficients of the fit. Consider the log-odds ratio for smoking and life status for a given age group:

$\log(EY_{11k}EY_{22k})/(EY_{12k}EY_{21k})$

This will be precisely the coefficient for the smoking and life-status term. We extract this:

```
> exp(coef(modu)['smokernordeadno'])
smokerno:deadno
         1.5333
```

We see that this is exactly the log-odds ratio we found above. The other interaction terms may be interpreted similarly.

Model Selection: Log-linear models are hierarchical, so it makes sense to start with the most complex model and see how far it can be reduced. We can use analysis of deviance to compare models. We start with the saturated model:

```
> modsat <- glm(y ~ smoker*age*dead, femsmoke,
family=poisson)
> dropl(modsat,test="Chi")
Single term deletions
Model:
y ~ smoker * age * dead
                Df Deviance   AIC  LRT Pr(Chi)
<none>              3.0e-10 190.2
smoker:age:dead  6      2.4 180.6  2.4    0.88
```

We see that the three-way interaction term may be dropped. Now we consider dropping the two-way terms:

```
> dropl(modu,test="Chi")
Single term deletions
Model:
y ~ (smoker + age + dead)^2
           Df Deviance AIC LRT Pr (Chi)
<none>                2 181
smoker:age   6       93 259  90   <2e-16
smoker:dead  1        8 185   6    0.015
age:dead     6      632 798 630   <2e-16
```

Two of the interaction terms are strongly significant, but the smoker: dead term is only just statistically significant. This term corresponds to the test for conditional

independence of smoking and life status given age group. We see that the conditional independence does not hold. This tests the same hypothesis as the Mantel-Haenszel test above. In this case the *p*-values for the two tests are very similar.

Binomial Model: For some three-way tables, it may be reasonable to regard one variable as the response and the other two as predictors. In this example, we could view life status as the response. Since this variable has only two levels, we can model it using a binomial GLM. For more than two levels, a multinomial model would be required.

We construct a binomial response model:

```
> ybin <- matrix(femsmoke$y,ncol=2)
> modbin <- glm(ybin ~ smoker*age, femsmoke[1:14,],
family=binomial)
```

This model is saturated, so we investigate a simplification:

```
> dropl(modbin,test="Chi")
Single term deletions
Model:
ybin ~ smoker * age
           Df Deviance  AIC  LRT Pr(Chi)
<none>         5.3e-10 75.0
smoker:age  6      2.4 65.4  2.4    0.88
```

We see that the interaction term may be dropped, but now we check if we may drop further terms:

```
> modbinr <- glm(ybin ~ smoker+age, femsmoke[1:14,],
family=binomial)
> dropl(modbinr,test="Chi")
Single term deletions
Model:
ybin ~ smoker + age
       Df Deviance AIC LRT Pr(Chi)
<none>            2  65
smoker 1         8  69   6   0.015
age    6       632 683 630  <2e-16
```

We see that both main effect terms are significant, so no further simplification is possible. This model is effectively equivalent to the uniform association model above. Check the deviances:

```
> deviance(modu)
[1] 2.3809
> deviance(modbinr)
[1] 2.3809
```

We see that they are identical. We can extract the same odds ratio from the parameter estimates as above:

```
> exp(-coef(modbinr)[2])
smokerno
  1.5333
```

The change in sign is simply due to which outcome is considered a success in the binomial GLM. So we can identify the binomial GLM with a corresponding Poisson GLM and the numbers we will obtain will be identical. We would likely prefer the binomial analysis where one factor can clearly be identified as the response and we would prefer the Poisson GLM approach when the relationship between the variables is more symmetric. However, there is one important difference between the two approaches. The null model for the binomial GLM:

```
> modbinull <- glm(ybin ~ 1, femsmoke[1:14,],
family=binomial)
> deviance(modbinull)
[1] 641.5
```

is associated with this two-way interaction model for the Poisson GLM:

```
> modj <- glm(y ~ smoker*age + dead, femsmoke,
family=poisson)
> deviance(modj)
[1] 641.5
```

So the binomial model implicitly assumes an association between smoker and age. In this particular dataset, there are more younger smokers than older ones, so the association is present. However, what if there was no association? One could argue that the Poisson GLM approach would be superior because it would allow us to drop this term and achieve a simpler model. On the other hand, one could argue that if the relationship between the response and the two predictors is the main subject of interest, then we lose little by conditioning out the marginal combined effect of age and smoking status, whether it is significant or not.

Correspondence Analysis: We cannot directly apply the correspondence analysis method described above for two-way tables. However, we could combine two of the factors into a single factor by considering all possible combinations of the two level. To make the choice of which two levels to combine, we would pick the pair whose association is least interesting to us. We could apply this to the smoking dataset here, but because there are only two levels of smoking and life status, the plot is not very interesting.

4.5 Ordinal Variables

Some variables have a natural order. One can use the methods for nominal variables described earlier in this chapter, but more information can extracted by taking advantage of the structure of the data. Sometimes one might identify a particular ordinal variable as the response. In such cases, the methods of Section 5.3 can be used. However, sometimes

one is simply interested in modeling the association between ordinal variables. Here the use of *scores* can be helpful.

Consider a two-way table where both variables are ordinal. We may assign scores u_i and v_j to the rows and columns such that $u_1 \leq u_2 \leq \ldots \leq u_I$ and $v_1 \leq v_2 \leq \ldots \leq v_j$. The assignment of scores requires some judgment. If you have no particular preference, even spacing allows for the simplest interpretation. If you have an interval scale, for example, 0–10 years old, 10–20 years old, 20–40 years old and so on, midpoints are often used. It is a good idea to check that the inference is robust to the assignment of scores by trying some alternative choices. If your qualitative conclusions are changed, this is an indication that you cannot make any strong finding.

Now fit the *linear-by-linear association* model:

$\log EY_{ij} = \log \mu_{ij} = \log np_{ij} = \log n + \alpha_i + \beta_j + \gamma_{ui} v_j$

So $\gamma=0$ means independence while γ represents the amount of association and can be positive or negative. γ is rather like an (unscaled) correlation coefficient. Consider underlying (latent) continuous variables which are discretized by the cutpoints u_i and v_j. We can then identify γ with the correlation coefficient of the latent variables

Consider an example drawn from a subset of the 1996 American National Election Study (Rosenstone, Kinder, and Miller (1997)). Considering just the data on party affiliation and level of education, we can construct a two-way table:

```
> data(nes96)
> xtabs( ~ PID + educ, nes96)
         educ
PID       MS HSdrop HS Coll CCdeg BAdeg MAdeg
  strDem   5 19     59 38   17    40    22
  weakDem  4 10     49 36   17    41    23
  indDem   1 4      28 15   13    27    20
  indind   0 3      12  9    3     6     4
  indRep   2 7      23 16    8    22    16
  weakRep  0 5      35 40   15    38    17
  strRep   1 4      42 33   17    53    25
```

Both variables are ordinal in this example. We need to convert this to a dataframe with one count per line to enable model fitting.

```
> (partyed <- as.data.frame.table(xtabs( ~ PID + educ,
nes96)))
       PID   educ Freq
1   strDem     MS    5
2  weakDem     MS    4
3   indDem     MS    1
...etc....
```

If we fit a nominal-by-nominal model, we find no evidence against independence:

```
> nomod <- glm(Freq ~ PID + educ, partyed, family=
poisson)
```

```
> pchisq(deviance(nomod),df.residual(nomod),lower=F)
[1] 0.26961
```

However, we can take advantage of the ordinal structure of both variables and define some scores. As there seems to be no strong reason to the contrary, we assign evenly spaced scores: one to seven for both PID and educ:

```
> partyed$oPID <- unclass(partyed$PID)
> partyed$oeduc <- unclass(partyed$educ)
```

Now fit the linear-by-linear association model and compare to the independence model:

```
> ormod <- glm(Freq ~ PID + educ + I (oPID*oeduc),
partyed,
  family= poisson)
> anova(nomod,ormod,test="Chi")
Analysis of Deviance Table
Model 1: Freq ~ PID + educ
Model 2: Freq ~ PID + educ + I (oPID * oeduc)
  Resid. Df Resid. Dev Df Deviance  P(>|Chi|)
1        36       40.7
2        35       30.6  1     10.2    0.0014
```

We see that there is some evidence of an association. So we see that using the ordinal information gives us more power to detect an association. We can examine $\hat{\gamma}$:

```
> summary(ormod)$coef['I(oPID * oeduc)',]
  Estimate Std. Error    z value   Pr(>|z|)
0.0287446  0.0090617  3.1720850  0.0015135
```

We see that $\hat{\gamma}$ is 0.0287. The p-value here can also be used to test the significance of the association although, as a Wald test, it is less reliable than the likelihood ratio test we used first. We see that $\hat{\gamma}$ is positive, which, given the way that we have assigned the scores, mean that a higher level of education is associated with a greater probability of tending to the Republican end of the spectrum.

Just to check the robustness of the assignment of the scores, it is worth trying some different choices. For example, suppose we choose scores so that there is more of a distinction between Democrats and Independents as well as Independents and Republicans. Our assignment of scores for apid below achieves this. Another idea might be that people who complete high school or less are not different; that those who go to college, but do not get a BA degree are not different and that those who get a BA or higher are not different. My assignment of scores in aedu achieves this:

```
> apid <- c(1,2,5,6,7,10,11)
> aedu <- c(1,1,1,2,2,3,3)
> ormoda <- glm(Freq ~ PID + educ +
I(apid[oPID]*aedu[oeduc]),
```

```
  partyed, family= poisson)
> anova(nomod,ormoda,test="Chi")
Analysis of Deviance Table
Model 1: Freq ~ PID + educ
Model 2: Freq ~ PID + educ + I(apid[oPID] *
aedu[oeduc])
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        36       40.7
2        35       30.9  1      9.8    0.0017
```

The numerical outcome is slightly different, but the result is still significant. Some experimentation with other plausible choices indicates that we can be fairly confident about the association here.

The association parameter may be interpreted in terms of log-odds. For example, consider the log-odds ratio for adjacent entries in both rows and columns:

$$\log \frac{\mu_{ij}\mu_{i+1,j+1}}{\mu_{i,j+1}\mu_{i+1,j}} = \gamma(u_{i+1} - u_i)(v_{j+1} - v_j)$$

For evenly spaced scores, these log-odds ratios will all be equal. For our example, where the scores are spaced one apart, the log-odds ratio is γ. To illustrate this point, consider the fitted values under the linear-by-linear association model:

```
> round(xtabs(predict(ormod,type="response") ~ PID +
educ, partyed),2)
         educ
PID       MS    HSdrop HS    Coll  CCdeg BAdeg MAdeg
  strDem   3.58 13.36  59.22 41.34 18.34 42.46 21.71
  weakDem  2.92 11.22  51.20 36.78 16.80 40.02 21.06
  indDem   1.59  6.27  29.45 21.78 10.23 25.09 13.59
  indind   0.49  2.00   9.65  7.34  3.55  8.96  5.00
  indRep   1.12  4.71  23.41 18.33  9.13 23.70 13.60
  weakRep  1.61  6.95  35.59 28.68 14.69 39.28 23.19
  strRep   1.69  7.49  39.48 32.74 17.26 47.49 28.85
```

Now compute log-odds ratio for, say, the lower two-by-two table:

```
> log(39.28*28.85/(47.49*23.19))
[1] 0.028585
```

We see this is, but for rounding, equal to $\hat{\gamma}$.

It is always worth examining the residuals to check if there is more structure than the model suggests. We use the raw response residuals (the unscaled difference between observed and expected) because we would like to see effects which are large in an absolute sense.

```
> round(xtabs(residuals(ormod,type="response") ~ PID +
educ, partyed),2)
         educ
```

```
PID       MS    HSdrop HS    Coll  CCdeg BAdeg MAdeg
  strDem   1.42  5.64  -0.22 -3.34 -1.34 -2.46  0.29
  weakDem  1.08 -1.22  -2.20 -0.78  0.20  0.98  1.94
  indDem  -0.59 -2.27  -1.45 -6.78  2.77  1.91  6.41
  indind  -0.49  1.00   2.35  1.66 -0.55 -2.96 -1.00
  indRep   0.88  2.29  -0.41 -2.33 -1.13 -1.70  2.40
  weakRep -1.61 -1.95  -0.59 11.32  0.31 -1.28 -6.19
  strRep  -0.69 -3.49   2.52  0.26 -0.26  5.51 -3.85
```

We do see some indications of remaining structure. For example, we see many more weak Republicans with some college than expected while fewer Republicans with master's degrees or higher. There may not be a monotone relationship between party affiliation and educational level.

To investigate this effect, we might consider an ordinal-by-nominal model where we now treat education as a nominal variable. This is called a *column effects* model because the columns (which are the education levels here) are not assigned scores and we will estimate their effect instead. A *row effects* model is effectively the same model except with the roles of the variables reversed. The model takes the form:

$$\log EY_{ij} = \log \mu_{ij} = \log np_{ij} = \log n + a_i + \beta_j + u_i\gamma_j$$

where the γ_j are called the column effects. Equality of the γ_js corresponds to the hypothesis of independence. We fit this model for our data:

```
> cmod <- glm(Freq ~ PID + educ + educ:oPID, partyed,
family= poisson)
```

We can compare this to the independence model:

```
> anova(nomod,cmod,test="Chi")
Analysis of Deviance Table
Model 1: Freq ~ PID + educ
Model 2: Freq ~ PID + educ + educ:oPID
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        36       40.7
2        30       22.8  6     18.0    0.0063
```

We find that the column-effects model is preferred. Now examine the fitted coefficients, looking at just the interaction terms as the main effects have no particular interest:

```
> summary(cmod)$coef[14:19,]
                  Estimate Std. Error   z value
Pr(>|z|)
educMS:oPID      -0.3122169   0.154051 -2.026710
0.042692
educHSdrop:oPID -0.1944513   0.077228 -2.517891
0.011806
educHS:oPID     -0.0553470   0.048196 -1.148384
0.250810
```

```
educColl:oPID    0.0044605   0.050603  0.088147
0.929760
educCCdeg:oPID  -0.0086994   0.060667 -0.143395
0.885978
educBAdeg:oPID   0.0345539   0.048782  0.708330
0.478740
```

The last coefficient, educMAdeg: oPID, is not identifiable and so this may be taken as zero. If there was really a monotone trend in the effect of educational level on party affiliation, we would expect these coefficients to be monotone. However, we can see that they are not. However, if we compare this to the linear-by-linear association model:

```
> anova(ormod,cmod,test="Chi")
Analysis of Deviance Table
Model 1: Freq ~ PID + educ + I(oPID * oeduc)
Model 2: Freq ~ PID + educ + educ:oPID
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        35      30.57
2        30      22.76  5     7.81      0.17
```

We see that the simpler linear-by-linear association is preferred to the more complex column-effects model. Nevertheless, if the linear-by-linear association were a good fit, we would expect the observed column-effect coefficients to be roughly evenly spaced. Looking at these coefficients, we observe that for high school and above, the coefficients are not significantly different from zero while for the lowest two categories, there is some difference. This suggests an alternate assignment of scores for education:

```
> aedu <- 0(1,1,2,2,2,2,2)
> ormodb <- glm(Freq ~ PID + educ + I
(oPID*aedu[oeduc]),
  partyed, family= poisson)
> deviance(ormodb)
[1] 28.451
> deviance(ormod)
[1] 30.568
```

We see that the deviance of this model is even lower than our original model. This gives credence to the view that whether a person finishes high school or not is the determining factor in party affiliation. However, since we used the data itself to assign the scores and come up with this hypothesis, we would be tempting fate to then use the data again to test this hypothesis.

The use of scores can be helpful in reducing the complexity of models for categorical data with ordinal variables. It is especially useful in higher dimensional tables where a reduction in the number of parameters is particularly welcome. The use of scores can also sharpen our ability to detect associations.

Further Reading: See books by Agresti (2002), Bishop, Fienberg, and Holland (1975), Haberman (1977), Le (1998), Leonard (2000), Powers and Xie (2000), Santner and Duffy (1989) and Simonoff (2003).

Exercises

1. The dataset parstum contains cross-classified data on marijuana usage by college students as it relates to the alcohol and drug usage of the parents. Analyze the data as if both factors were nominal. Redo the analysis treating both factors as ordinal. Contrast the results.
2. The dataset melanoma gives data on a sample of patients suffering from melanoma (skin cancer) cross-classified by the type of cancer and the location on the body. Determine whether the type and location are independent. Examine the residuals to determine whether any dependence can be ascribed to particular type/location combinations.
3. Data on social mobility of men in the UK may be found in cmob. A sample of men aged 45–64 was drawn from the 1971 census and 1981 census and the social class of the man was recorded at each timepoint. The classes are I=professional, II=semiprofessional, IIIN=skilled nonmanual, HIM=skilled manual, IV= semiskilled, V=unskilled.

   (a) Check for symmetry, quasi-symmetry, marginal homogeneity and quasi-independence.
   (b) Develop a score-based model. Find some good-fitting scores.

4. The dataset death contains data on murder cases in Florida in 1977. The data is cross-classified by the race (black or white) of the victim, of the defendant and whether the death penalty was given.

   (a) Consider the frequency with which the death penalty is applied to black and white defendants, both marginally and conditionally, with respect to the race of the victim. Is this an example of Simpson's paradox? Are the observed differences in the frequency of application of the death penalty statistically significant?
   (b) Determine the most appropriate dependence model between the variables.
   (c) Fit a binomial regression with death penalty as the response and show the relationship to your model in the previous question.

5. The dataset sex fun comes from a questionnaire from 91 couples in the Tucson, Arizona, area. Subjects answered the question "Sex is fun for me and my partner". The possible answers were "never or occasionally", "fairly often", "very often" and "almost always".

   (a) Check for symmetry, quasi-symmetry, marginal homogeneity and quasi-independence.
   (b) Develop a score-based model. Find some good-fitting scores.

6. The dataset suicide contains one year of suicide data from the United Kingdom cross-classified by sex, age and method.

   (a) Determine the most appropriate dependence model between the variables.
   (b) Collapse the sex and age of the subject into a single six-level factor containing all combinations of sex and age. Conduct a correspondence analysis and give an interpretation of the plot.
   (c) Repeat the correspondence analysis separately for males and females. Does this analysis reveal anything new compared to the combined analysis in the previous question?

7. A student newspaper conducted a survey of student opinions about the Vietnam War in May 1967. Responses were classified by sex, year in the program and one of four opinions. The survey was voluntary. The data may be found in the dataset uncviet.

   (a) Conduct an analysis of the patterns of dependence in the data assuming that all variables are nominal.
   (b) Assign scores to the year and opinion and fit an appropriate model. Interpret the trends in opinion over the years. Check the sensitivity of your conclusions to the assignment of the scores.

8. The dataset HairEyeColor contains the same data analyzed in this chapter as haireye. Repeat the analysis in the text for each sex and make a comparison of the conclusions.
9. A sample of psychiatry patients were cross-classified by their diagnosis and whether a drug treatment was prescribed. The data may be found in drugpsy. Is the chance that drugs will be prescribed constant across diagnoses?
10. The UCBadmissions dataset presents data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.

(a) Show that this provides an example of Simpson's paradox.
(b) Determine the most appropriate dependence model between the variables.
(c) Fit a binomial regression with admissions status as the response and show the relationship to your model in the previous question.