Example 3: Case control study

A study was initiated to investigate the association between a woman’s having periodontal
disease and her chance of having a pre-term, low birth weight infant. The researchers asked
46 women who had delivered a low weight infant and 20 women who had delivered a normal
weight infant about their history of periodontal disease. It appears that women who had
low weight infants are more likely to have had periodontal disease.

<table>
<thead>
<tr>
<th>Periodontal Disease</th>
<th>Low Weight Infant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>41</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
</tr>
</tbody>
</table>

In this study, the number of women who had periodontal disease in each birth weight group
is the response variable. This is a retrospective study: a case control study where the total
in column \(j\) was fixed a priori at \(m_j\) (the number of women that the researchers chose to
study in the \(j^{th}\) birth weight group, \(j = 1, 2\)). Although birth weight in each disease group
might seem like a more natural response variable, we cannot use it as such since the numbers
in each group were fixed. This is a common design for studying the relationship between
exposure (e.g. to periodontal disease) and a rare outcome (e.g. low birth weight).

Let \(Y_{ij}\) be the count in the \((i, j)\)th cell, \(i, j = 1, 2\). Here, the counts \(Y_{1,j}, Y_{2,j}\) are observed
conditional on their sum (for each column \(j\)), so are not independent. Since the column
totals are fixed, again, the \(Y_{ij}\)’s cannot be Poisson distributed.

Instead, assuming that each woman’s exposure is independent, the counts in each column
have a multinomial distribution. For column \(j\), the probability that a count falls in cell \((i, j)\)
is denoted by \(\theta_{ij}\), where

\[
\sum_i \theta_{ij} = 1
\]

for each column \((j = 1, \ldots, J)\). In other words, the joint distribution of the counts in column
\(j\) is

\[
f \left( \mathbf{y} \left| \sum_i y_{ij} = m_j \right. \right) = m_j! \prod_i \frac{\theta_{ij}^{y_{ij}}}{y_{ij}!}.
\]

Since counts in different columns are independent, the joint distribution for all the cells in
the table is given by the product multinomial distribution:

\[
f \left( \mathbf{y} \left| \sum_i y_{i1} = m_1, \ldots, \sum_i y_{iJ} = m_J \right. \right) = \prod_j m_j! \prod_i \frac{\theta_{ij}^{y_{ij}}}{y_{ij}!}.
\]
Under this model,
\[ E[Y_{ij}] \equiv \mu_{ij} = m_j \theta_{ij}, \]
so
\[ \log \mu_{ij} = \log m_j + \log \theta_{ij}. \]

We can now model \( \theta_{ij} \) to address the question of whether a woman is more likely to have had periodontal disease if she gives birth to a low weight infant.

While the interpretation of a case control study is different than that of a randomized controlled experiment, the models are identical. (It doesn’t matter whether it’s the row totals or the column totals that are fixed; the choice of which variable to put on each side of the contingency table is arbitrary.)

**Analyzing Multinomial Data**

Suppose we have data in the following form:

<table>
<thead>
<tr>
<th>Covariate Class (i)</th>
<th>Response Category (j)</th>
<th>Row Totals</th>
<th>Explanatory Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>( Y_{11} ) \cdots ( Y_{1c} )</td>
<td>( m_1 )</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>( Y_{21} ) \cdots ( Y_{2c} )</td>
<td>( m_2 )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>( n )</td>
<td>( Y_{n1} ) \cdots ( Y_{nc} )</td>
<td>( m_n )</td>
<td>( x_n )</td>
</tr>
</tbody>
</table>

Here, a covariate class is a unique combination of covariate values shared by the observations in this class, and \( Y_{ij} \) is the count observed in cell \((i, j)\). Note that both the randomized controlled trial and case control study designs fit into this paradigm. We will look at the analysis of classification study data later.

We can model these data using a **multinomial logit model**. In particular, we let \( \pi_{ij} \) be the probability that an item of covariate class \( i \) will be in category \( j \). Note that \( \sum_{j=1}^{c} \pi_{ij} = 1 \) for all \( i \). Then, we choose one category as a reference category, say category 1. The logit for category \( j \) \((j = 2, \ldots, c)\) of the \( i^{th} \) covariate class is defined as
\[
\log \left( \frac{\pi_{ij}}{\pi_{i1}} \right) = \sum_{k=1}^{p} x_{ik} \beta_{jk}. 
\]

Note that here the regression coefficients (the \( \beta_{jk} \)'s) depend on \( j \). This formulation of the model ensures that the constraint \( \sum_{j=1}^{c} \pi_{ij} = 1 \) is satisfied.

If \( c = 2 \) then we have binomial data. By labelling category 1 as failure and category 2 as success, the multinomial model reduces to the binomial GLM with logistic link. In particular, the logit parameter for category 2 (success) is \( \log(\frac{\pi_{i2}}{\pi_{i1}}) = \log(\frac{\pi_{i2}}{1-\pi_{i2}}) \), which is the usual logistic
link function in terms of the probability of success. In this case, we can use S-PLUS in the usual way to fit the model.

However, when $c > 2$, the multinomial logit model does not fit into the GLM framework (e.g. because this distribution is multivariate). Fortunately, it turns out that we can estimate the parameters in the model using a Poisson GLM with log link. The key is in the specification of the linear predictor.

**Definition:** A model of the form

$$\log E[Y_i] \equiv \log \mu_i = \text{constant} + x_i^\prime \beta$$

is called a *log-linear model*.

**Important Fact:** For the multinomial models considered in this course, the MLEs can be obtained by fitting a Poisson GLM with log link as long as the parameters which correspond to the fixed marginal totals are included in the model in the appropriate way. In other words, we can use the Poisson GLM functionality in S-PLUS in the usual way for estimating the parameters associated with the contingency table. We can also use the usual goodness-of-fit assessment tools (e.g. deviance and $\chi^2$ tests, residual analysis, etc.).

Q: Why does this work?

We will consider an intuitive answer to this question.

**FACT:** Suppose that $Y_1, Y_2, \ldots, Y_m$ are independent Poisson random variables with means $\mu_1, \mu_2, \ldots, \mu_m$ respectively. Then the sum $Y = Y_1 + Y_2 + \ldots + Y_m$ also has a Poisson distribution with mean $\mu = \mu_1 + \mu_2 + \ldots + \mu_m$.

Assume (for simplicity) that we have a $2 \times 1$ table, with $Y_1$ representing the count in the first cell, and $Y_2$ representing the count in the second. Assume further that $Y_1$ and $Y_2$ are independent with distributions Poisson($\mu_1$) and Poisson($\mu_2$), respectively. We’re given that $Y_1 + Y_2 = n$. Then

$$P(Y_1 = y_1, Y_2 = y_2 \mid Y_1 + Y_2 = n) = \frac{P(Y_1 = y_1, Y_2 = n - y_1)}{P(Y_1 + Y_2 = n)} = \frac{e^{-\mu_1} \mu_1^{y_1}}{y_1!} \cdot \frac{e^{-\mu_2} \mu_2^{n-y_1}}{(n-y_1)!} \cdot \frac{1}{\frac{e^{-\mu_1-\mu_2}(\mu_1+\mu_2)^n}{n!}}$$

$$= \binom{n}{y_1} \left( \frac{\mu_1}{\mu_1 + \mu_2} \right)^{y_1} \left( \frac{\mu_2}{\mu_1 + \mu_2} \right)^{n-y_1}$$

... which is the binomial distribution with P(success) = $\frac{\mu_1}{\mu_1 + \mu_2}$. In general, if we have $m$ counts (and again condition on their sum), we will end up with the multinomial distribution,
i.e.
\[
P \left( Y_1 = y_1, \ldots, Y_m = y_m \ \middle| \sum_{i=1}^{m} Y_i = n \right) = \frac{n!}{y_1! \cdots y_m!} \left( \frac{\mu_1}{\sum_{i=1}^{m} \mu_i} \right)^{y_1} \cdots \left( \frac{\mu_m}{\sum_{i=1}^{m} \mu_i} \right)^{y_m}.
\]

This illustrates the connection between the Poisson and binomial distributions: the likelihood associated with multinomial observations is the same as the likelihood associated with Poisson observations which are constrained by a fixed total.

To apply this result, let \( \mu_{ij} \) be the expected count in the \((i, j)\)th cell, and let \( \mu_i = \sum_j \mu_{ij} \). The multinomial logit model for category \( j \) \((j = 2, \ldots, c)\) of the \(i\)th covariate class can be written as
\[
\log \left( \frac{\pi_{ij}}{\pi_{i1}} \right) = \log \left( \frac{\mu_{ij}/\mu_i}{\mu_{i1}/\mu_i} \right) = \log \left( \frac{\mu_{ij}}{\mu_{i1}} \right) = x_i' \beta_j,
\]
with \( \beta_1 \equiv 0 \). So, the effect of the covariates is allowed to depend on the response category.

Equivalently,
\[
\log \mu_{ij} = \log \mu_{i1} + x_i' \beta_j \equiv \mu + \psi_i + x_i' \beta_j,
\]
i, j \geq 1, where \( \log \mu_{i1} = \mu + \psi_i \) and \( \psi_i = \beta_1 \equiv 0 \). Using the “important fact” above, this model can be fit by taking each \( Y_{ij} \) as a Poisson(\( \mu_{ij} \)) random variable and then fitting a Poisson GLM with log link function and linear predictor \( \log \mu_{ij} = \mu + \psi_i + x_i' \beta_j \).

Note that there is a set of “nuisance parameters” (the \( \psi_i \)'s). There is one parameter for each row. It is these parameters in the model which must be included in the model in order to satisfy the condition of the “important fact”. So, a nuisance factor with \( n \) levels has to be created and included as well. This factor guarantees that the sum of observed values is the same as the sum of fitted values in each covariate class.

As mentioned earlier, the vector of regression coefficients \( \beta_j \) will in general be different for each \( j \). We can achieve this in S-PLUS by creating a categorical predictor called \texttt{Category} to denote the categories. We denote the effect of the \( j \)th category by \( \gamma_j \). We then fit the model with the interactions between category and the other predictor variables.

NOTES:

1. Remember that our goal is to estimate the parameters of the multinomial logit model \((\beta_j)\). We are \textit{not} interested in the log-linear model, except as a tool for estimating \( \beta_j \).

2. The log-linear model can be derived from the multinomial logit model. The only “trick” that we use is, when fitting the model, to treat the observed response as Poisson, rather than multinomially, distributed. The MLE of \( \beta_j \) that we obtain based on the log-linear model will be identical to that based on the multinomial model as long as we include the nuisance parameters (the \( \psi_i \)'s) in the proper way.