Chapter 2

Methodology

This chapter describes the methodology, including an overview of the statistical methods for data analysis aligned to the statistical models. It presents statistical methods adopted in the three studies contained in Chapter 3.

These methods include generalized linear models with Poisson and negative binomial distributional assumptions, and log-transformed linear regression modeling. For comparing various levels of a categorical determinant we prefer the method based on *sum* contrasts to compare each level with the over mean, rather than the more conventional *treatment* contrasts that compare each level with a specified referent level, thus facilitating the construction of thematic maps for visually displaying region effects. We also include methods for multivariate multiple linear regression (including factor analysis for reducing correlations between determinants that can mask their associations with outcomes), and canonical correspondence analysis. Mathematical derivations are not given here but are referred to in a number of textbooks and papers that describe the statistical theory.

2.1 Generalized linear models

Generalized linear models (GLMs) were formally introduced by Nelder and Wederburn (1972). The GLM method generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. Special cases used for our studies give rise to Poisson and negative binomial distributions for the count outcome.

Poisson regression model

Poisson regression is appropriate for fitting models where the response is a count (non-negative integer value). A random variable Y is said to have a Poisson distribution with parameter $\lambda > 0$ if it takes integer values $y = 0, 1, 2, \ldots$, with probabilities expressed as:

$$\text{Prob}(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}. \tag{2.1}$$

The mean and variance of this distribution can be shown to be

$$E(Y) = \text{var}(Y) = \lambda. \tag{2.2}$$

Since the mean is equal to the variance, any factor that affects one will also affects the other. The Poisson regression model can be fitted by using the generalized linear models with the log link function (McCullagh and Nelder 1989).

Poisson regression for rates

Poisson regression is commonly used for modeling the number of cases of disease in a specific population within a certain time period by an *offset* term (Greene 2003). Suppose that n_{ijqt} is the number of observed cases in cells defined by demographic (age-gender) group i, region j, season q and year t, and P_{ij} is the corresponding population at risk. Denoting the corresponding mean incidence rate by λ_{ijqt}, we consider additive models of the form:

$$\ln\left(\lambda_{ijqt}\right) = \mu + \alpha_i + \beta_j + \eta_q + \gamma_t + \ln(P_{ij}). \tag{2.3}$$

The terms α_i, β_j, η_q and γ_t represent demographic group, region, season and year effects, respectively, and are centered at 0, so that μ is a constant encapsulating the overall incidence rate when $\ln(P_{ij})$ is an *offset* term.

Adjusted incidence rates

After fitting the model, adjusted incidence rates for each factor of interest are obtained by suppressing the subscripts in Equation (2.3) corresponding to the other factors and replacing these terms with a constant satisfying the condition that the sum of the counts based on the adjusted incidence rates matches the total (Swennen et al 2009). For Poisson regression, this is achieved simply by multiplying the incidence rates for the specified factor of interest by a scale constant specific to the factor.

Overdispersion

After fitting a generalized linear model to the data, to check the adequacy of the respective model, one usually computes a residual deviance for each cell. Thus, the deviance statistic for an observation reflects its contribution to the overall goodness of fit of the model. Plotting these residual deviances against corresponding quantiles for the normal distribution gives an indication of the adequacy of the fit of the model to the data. If the plot is approximately linear with unit slope, the fit is satisfactory. Details are given in Chapter 7 of Venables and Ripley (2002).

Overdispersion in Poisson models occurs when the response variance is greater than mean. If the observed variance is greater than the mean the data are *over-dispersed* and the residual deviance plot will indicate that the model is not appropriate. Further details are given in Chapter 4 of Hilbe (2007).

Negative binomial regression model

Negative binomial regression is a common alternative to Poisson regression models when dealing with over-dispersed data.

The negative binomial distribution for a random variable Y is the number of trials with fixed probability π needed for a specified number of successes θ to occur. It thus has two parameters, and if the probability of success is re-expressed as $\lambda = (1\text{-}\pi)/\pi$ its density function is

$$\text{Prob}(Y = y) = \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+\lambda}\right)^{\theta}\left(\frac{\lambda}{\theta+\lambda}\right)^{y}. \tag{2.4}$$

The Poisson model arises in the limit as the parameter $\theta \to \infty$, so the Overdispersion parameter is actually the reciprocal of θ. The expected value of Y is λ and its variance is $\lambda + \lambda^2/\theta$.

The conventional negative binomial model, as defined above, gives rise to a generalized linear model, which is described in Chapter 7 of Venables and Ripley (2002). This model is also called the NB-2 model in contrast to other models including the NB-1 model used by Jansakul and Hinde (2004) for which the variance is $\lambda(1+\alpha)$ where $\alpha > 0$. However, these other models are not generalized linear models and require special software to fit them to data.

2.2 Log-transformed linear regression model

The conventional model for handling data where the outcome is continuous is linear regression, assuming independent error terms, each following an identical Gaussian distribution.

Let *Y* be a log-normally distributed random variable, that is, a random variable whose (natural) logarithm is Normal with mean μ and variance σ^2. This implies that the probability density function of ln(*Y*) is the density function of the normal distribution, namely

$$f(y) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(z-\mu)^2}{2\sigma^2} \right]. \tag{2.5}$$

In our studies, the incidence rates generally have positively skewed distributions so it is conventional to transform them by taking logarithms to obtain the outcome as

$$y_{ijqt} = \ln\left(1000 \times \frac{n_{ijqt}}{P_{ij}} \right). \tag{2.6}$$

Thus an additive linear model is fitted to the logarithms of the log-transformed incidence rates, namely

$$y_{ijqt} = \mu + \alpha_i + \beta_j + \eta_q + \gamma_t . \tag{2.7}$$

As in Equation (2.3), μ is a constant and α_i, β_j , η_q and γ_t represent demographic group, region, season and year effects, respectively, with zero means.

Handling zeroes

If any count n_{ijqt} is zero, Equation (2.6) needs to be modified to give a finite result, so that n_{ijqt} is replaced by a positive value n^*_{ijqt} .

Various methods may be considered for this data modification. Zero counts simply could be omitted, and the fitted model then used to impute counts for these cases before refitting the model (Ardkeaw and Tongkumchum 2009). This method has advantages in situations where under-reporting is known or suspected. Another method involves adding a constant *c* to all counts so that $n^{*}_{ijqt} = n_{ijqt} + c$. A third method involves replacing the zeroes by a suitably chosen constant *d* without changing any values of n_{ijqt} greater than 0.

Sum contrasts

Sum contrasts (Venables and Ripley 2002, Tongkumchum and McNeil 2009) are used to obtain confidence intervals for comparing adjusted incidence rates within each factor with the overall incidence rate. An advantage of these confidence intervals is that they provide a simple criterion for classifying levels of a factor into three groups according to whether each corresponding confidence interval exceeds, crosses, or is below the overall mean.

Methods for creating geographical maps

A thematic map is a type of map that uses different colours or shades to graphically display information about the underlying data representing estimated values of a variable at different locations on the map. The thematic map using data in regions might show one region in dark red to indicate that the region has high values, while showing another region in very pale red to indicate that the region has low values. A range map is a type of thematic map that displays data according to ranges set by the users. The ranges are shaded using colors or patterns. These types of maps are used to show the geographical distribution of the adverse outcome and to identify areas of high risk. Appropriate graphs are used for exploratory data analysis, visualizing the pattern of the data and highlighting possible errors in the data that could cause problems in further analysis.

Since the confidence intervals for factor-specific incidence rates obtained from a model divide naturally into three groups according to their location entirely above the mean, around the mean, or entirely below the mean, we used this trichotomy to create thematic maps of districts according to their estimated incidence rates.

2.3 Multivariate multiple regression model

Multivariate multiple regression is a logical extension of the multiple regression concept to allow for multiple response (dependent) variables. Multivariate regression estimates the same coefficients as one would obtain using separate univariate regression models. In addition, multivariate regression, being a joint estimator, also estimates the between-equation covariance. This means that it is possible to test coefficients across equations.

Suppose that data are available for n observations, and the response variables are arranged into a *matrix* whose columns are p outcome variables and rows correspond to the n observations. The model (Mardia et al 1979) is defined in matrix form as

$$\mathbf{Y}_{(n \times p)} = \mathbf{X}_{(n \times q)}\,\mathbf{B}_{(q \times p)} + \mathbf{E}_{(n \times p)}\ . \tag{2.8}$$

In this formulation $\mathbf{Y}_{(n \times p)}$ is an observed matrix of p response variables on each of the n observations, $\mathbf{X}_{(n \times q)}$ is the matrix of q predictors (including a vector of 1s) in columns and n observations in rows, $\mathbf{B}_{(q \times p)}$ contains the regression coefficients (including the intercept terms), and $\mathbf{E}_{(n \times p)}$ is a matrix of unobserved random errors with mean zero and common covariance matrix $\mathbf{\Sigma}$. Ordinary (univariate) multiple regression arises as the special case when $p = 1$. If $q-1$ determinants $f_i^{(k)}$ ($k = 1, 2,\ldots, q-1$) are available, the model for outcome j for observation i may be expressed as

$$y_{ij} = \mu_j + \sum_{k=1}^{q-1} \beta_j^{(k)} f_i^{(k)}. \tag{2.9}$$

In this equation, y_{ij} is the observed outcome *j* at observation *i*, μ_j is the overall mean associated with outcome *j*, and $\beta_j^{(l)}$ is the effect of factor *i* on outcome *j*.

For each outcome, the model fitting may be assessed by plotting the residuals against normal quantiles (Venables and Ripley 2002), and also by using the set of r-squared values for the response variables to see how much of the variation in each is accounted for by the model.

There are different multivariate test criteria such as Hotelling's trace, *Wilks' lambda*, Pilai's trace. And Poys'lastest root. A likelihood ratio, *Wilks' lambda*, for testing the significance of an extra parameter is used as an equivalent form of the *F* statistic in the univariate regression model (Wilks 1932). The *Wilks' lambda* can be expressed in terms of generalized variances in the matrices of sums of squares and cross products due to the factors and the residuals. We reject the null hypothesis if the likelihood ratio of generalized variations is too small. The *F* statistics are then used to test the significance of the factors on each response variable.

2.4 Canonical correspondence analysis

Assuming that the data structure comprises the **Y** and **X** matrices with rows corresponding to measurements of outcomes and predictors taken on the same occasions, canonical correspondence analysis (Ter Braak 1986) produces a two-dimensional *biplot* comprising arrows of variable lengths and directions (*gradients*) emanating from a common origin representing the predictor variables, together with superimposed points denoting the outcome variables. The relative lengths of the arrows and the angles between them are based on the correlation matrix of the predictor variables, and the coordinates of the points are planar projections of the density outcomes, computed in such a way that their positions relative to the arrows portray their associations with the environmental predictors. The method also produces coordinate scores and p-values for the overall associations based on Monte Carlo permutation tests.

2.5 Factor analysis

The objective of factor analysis is to identify underlying factors that describe the correlations among the variables (Mardia et al 1979, Johnson and Wichern 1998). Suppose that X is the n by p matrix with X_j denoting the j^{th} variable and element x_{ij} denoting the i^{th} observation of the j^{th} variable. Factor analysis is based on the correlation matrix R of X using the maximum likelihood estimation method. The appropriate number of common factors, m, can be determined using a chi-squared test on the residual correlations. Let L be the p by m matrix of factor loadings. The loading of variable j on common factor q is denoted by elements $l_j^{(q)}$. The total variance of X_j can be decomposed into two parts corresponding to communality and uniqueness. Communality is the proportion of the variance of X_j explained by the m common factors. It is the sum of squares of the loadings of the j^{th} variable contributed by all the factors, that is, $(l_j^{(1)})^2 + (l_j^{(2)})^2 + ... + (l_j^{(m)})^2$. The unexplained variance of X_j due to the random errors is often called the uniqueness, ψ_j, indicating how distinctive the measure of X_j is from the remaining variables. Therefore, any variables with high uniqueness should be dropped from the factor analysis. In this case, the correlation matrix is reproduced based on a lesser number of variables, followed by the steps described above.

The factor loadings obtained may not be readily interpretable and thus may need to be rotated to simplify their interpretation. That is, each variable contributes highly on a single factor and loads a smaller amount on the other factors. Each factor summarizes a general characteristic of a group of variables. Two basic types of factor rotation are *orthogonal* and *oblique*. Orthogonal rotation (for example, *varimax*) yields independent factors (Kaiser 1959) whereas oblique rotation (for example, *promax*) allows factors to be correlated (Hendrickson and White 1964). In practice it is desirable to try several methods to achieve a simple structure of factors.

Denoting the factor q by $F^{(q)}$, each $F^{(q)}$ is a weighted linear combination of the scaled variables and the estimated values of $F^{(q)}$ are called factor scores, $f_i^{(q)}$ (i = 1, 2, ..., n). The scores can be determined by the weighted least squares method. Each score i for factor q is the sum of the rotated loadings of X_j on $F^{(q)}$ multiplied by the corresponding standardized values of X_j. The formula is

$$f_i^{(q)} = \sum_{j=1}^{p} \tilde{l}_j^{(q)} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right) \quad \text{for } i = 1, 2, ..., n \text{ and } q = 1, 2, ..., m \qquad (2.10)$$

where $\tilde{l}_j^{(q)}$ is the rotated loading of variable j on common factor q, $\bar{x}_j$ and s_j are the mean and standard deviation of variable j, respectively. These common factors are often used as inputs to subsequent applications such as regression analysis and cluster analysis. Some important studies using this method include those reported by Kuppusamy and Giridhar (2006), Ouyang et al (2006), Senthamarai and Nagarajan (2008), Amiri and Nakane (2009), Boyacioglu et al (2009)and Sangun et al (2009).

2.6 Model assumptions

Data transformation

Many statistical procedures assume that the variables are normally distributed. A significant violation of the assumption of normality can seriously increase the chances of the researcher committing either a Type I or II error (depending on the nature of the analysis and the non-normality). There are a great variety of possible data transformations, from adding constants to multiplying, squaring or rising to a power, converting to logarithmic scales, inverting and reflecting, taking the square root of the values, and even applying trigonometric transformations such as sine wave transformations.

Logarithmic transformations are actually a class of transformations, rather than a single transformation. In brief, a logarithm is the power (exponent) a base number must be raised to in order to get the original number.Any given number can be expressed as y to the x power in an infinite number of ways. It was used when (a) the variances are not equal (heterogeneity of variances) ,(b) standard deviations are proportional to the means (CV's are equal) or (c)when the data is positively skewed. There are good reasons to consider a range of bases (Cleveland 1984) argues that base 10, 2, and *e* should always be considered at a minimum).For example, in cases where there are extremes of range base 10 is desirable, but when there are ranges that are less extreme, using base 10 will result in a loss of resolution, and using a lower base (*e* or 2) will serve (higher bases tend to pull extreme values in more drastically than lower bases).

Osborne (2002) concluded when using count variables researchers should use the square root of the counts in the analyses, which takes care of count data issues in most cases. Proportions require an arcsine-root transformation. In order to apply this transformation, values must be between 0 and 1.A square root of the values is taken, and the inverse sine (arcsine) of that number is the resulting value. However, in order to use this variable in an analysis, each observation must be weighted by the number in the denominator of the proportion.

Moreover making inferences about the results from the statistical models, it is necessary to check the model assumptions of normality and homoskedasticity. If these assumptions are not plausible, the results may not be valid and may lead to incorrect conclusions. Graphical methods are often used to check these assumptions.

Graphical Normality

A scatter plot of standardized residuals versus their normal scores is commonly used to assess normality of errors in a linear regression model. If these errors are normally distributed, then the standardized residuals should match the normal scores. The plot should then resemble a nearly straight line with intercept zero and slope one.

Homoskedasticity

A scatter plot of the standardized residuals against the fitted response variable can be used to check the homoskedasticity assumption. If the dots on the plot are randomly scattered evenly within a horizontal band around zero, then the homoskedasticity assumption is plausible.