# Chapter 4

# Summary and Conclusions

This thesis has focused on using statistical methods to model over-dispersed outcomes with application to disease incidence rates in the south of Thailand, and sediment densities of macrobenthic fauna in Middle Songkhla Lake. This chapter summarizes the most important conclusions from the thesis, and suggests directions for further research by statisticians.

## 4.1 Pneumonia among children in Surat Thani province

The objective for this study was to investigate regional and temporal patterns of pneumonia incidence for young children in Surat Thani province to determine the districts and age-gender groups with high disease risk and thus offer a means to prevent epidemic outbreaks by using suitable timely measures.

We chose this research topic because after preliminary examination of data from the National Notifiable Disease Surveillance (Report 506) we found that in the seven provinces of the upper southern zone, pneumonia accounted for 6% of all disease cases over the nine-year period 1999-2004, and was the fourth most common disease reported, after diarrhea with 51.2% of cases, pyrexia of unknown origin with 10.4%, and conjunctivitis with 6.4%. Among these diseases reported, pneumonia was by far the most lethal, accounting for 47.7% of all deaths from hospital-diagnosed cases of infectious diseases in the region during the same period. It should be noted, however,

that while 59% of these pneumonia cases occurred among children less than 5 years of age, 89% of the deaths occurred among older peoples.

Of the seven provinces in the zone, Surat Thani province recorded the highest average incidence rate of pneumonia cases (8.3%) during the six years. Moreover, this province is the largest in area and the second largest in population. It is divided into 19 districts.

Following current statistical practice, we initially used a Poisson generalized linear model containing additive effects associated with the season of the year, gender-age group, and district. However, a diagnostic plot of residuals from this model indicated substantial over-dispersion, as Figure 4.1 shows.

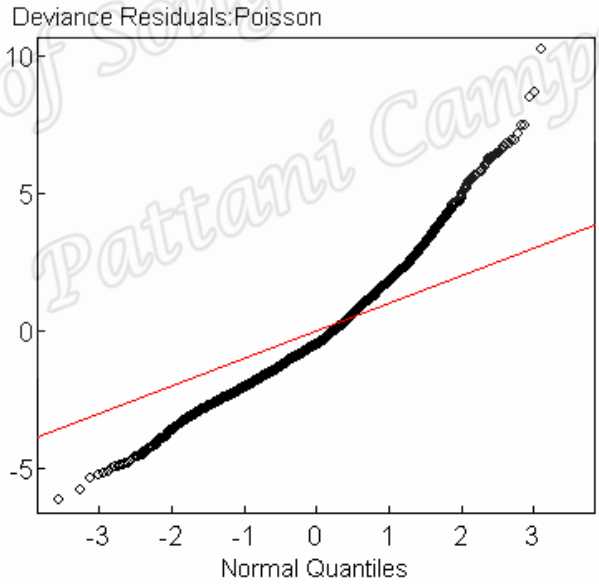


Figure 4.1: Deviance residuals from Poisson regression generalized linear model

Again following current statistical practice, we fitted a generalized linear model with a negative binomial distribution. This model can accommodate over-dispersion by means of an additional parameter $\theta$, with lower values indicating higher over-

dispersion, and the Poisson model arising in the limit as $\theta$ tends to infinity. For the pneumonia incidence rates the estimated value of $\theta$ was found to be 1.54 with standard error 0.061.

The Poisson model gives fitted values (also called *expected* values) for the cell counts. These expected cell counts are informative to public health authorities because the ratio of the observed number of cases to the expected number is a useful measure of the *observed* relative risk of disease associated with a cell. For example, 60 hospital-diagnosed pneumonia cases were reported in Surat Thani City among male infants less than 12 months of age in the September quarter of 2001, whereas the corresponding expected number according to the Poisson regression model was 11.2. This gives an estimate of 60/11.2 = 5.4 for the observed relative risk of disease in this particular cell. Using a formula for the standard error of a relative risk (see, for example, McNeil 1996, Chapter 4), an approximate 95% confidence interval is (2.8, 10.1).

The Poisson model has the property that the sum of the expected values based on the model is equal to the total number of observed cases. For the Surat Thani pneumonia data, the total number of disease cases was 16,253. But this property does hold for the negative binomial model, for which the sum of the expected values was found to be 23,000.3. This means that if the fitted values given by the negative binomial model are used to estimate relative risks of disease outcomes associated with specified cells, these estimates will be biased, and some adjustment is needed to correct for this bias. A reasonable method of adjustment is to rescale the expected cell counts based on the

model so that they sum to the total number of observed cases, and this is the method we used to create the graphs in Figure 3 of the first article.

However, this straightforward method of scaling fitted cell counts based on the negative binomial model can give rise to a further bias, as is seen in Figure 3 of the first article, where all the points corresponding to the Surat Thani City (denoted by blue circles) are located substantially below the line indicating a perfect fit of the model. For this reason, even though the reasonably close fit of the points to the straight line in the diagnostic plot of deviance residuals versus normal quantiles suggests that the negative binomial model is plausible, that model is unsatisfactory. It does not provide accurate estimates of the relative risks that health authorities need for effective planning of disease prevention policies.

The same theory applies to the alternative linear model based on the log-transformed incidence rates. In this case the fitted values given by the model in a cell must be exponentiated to reverse the log-transformation and then multiplied by the population associated with the cell to obtain a measure of the expected number of disease cases in the cell. But as for the negative binomial distribution, the sum of these expected cell counts will generally exceed the total observed disease count, so a scale factor is again needed to compensate for the non-linearity of the transformation*.

---

* For a given non-linear transformation $f(Z)$ of a normally distributed random probability theory as the ratio $E[f(Z)] / f(E[Z])$ where $E[X]$ is the expected value of X. If $f(x) = \log(X)$ this ratio is $\exp(\sigma^2/2)$ where $\sigma$ is the standard deviation.

For the log-transformation with the zero counts replaced by 1 (right panel of Figure 2 in the first article) the scale factor was found to be 1.31, and the corresponding expected cell count for the cell corresponding to male infants less than 12 months of age in Surat Thani City in the July-September quarter of 2001 is 40.0. This gives an estimate of 60/40.0 = 1.5 with 95% confidence interval (1.0, 2.2) for the observed relative risk of disease in this cell based on the log-transformed linear model. This is quite different from the 95% confidence interval (2.8, 10.1) based on the Poisson model and also quite different from the 95% confidence interval (2.0, 5.6) based on the negative binomial model using a similar calculation.

To summarize, the choice of distribution for the fitted model can be quite important, and choosing a model that appears to provide a plausible fit to the data can still give misleading estimates of relative risks associated with particular cells.

**4.2 Tuberculosis in southern Thailand**

The objective for this study was very similar to that for the first study. In this case the disease of interest was tuberculosis, and the scope of the study was more extensive, involving all fourteen provinces in southern Thailand and all ages in the population. However, due to the unavailability of more recent data for some provinces the study spanned only the six years from 1999 to 2004.

The conclusions from this study were remarkably similar to those for the study of pneumonia among young children in Surat Thani Province, in the sense that again the log-transformed linear model provided a superior method for data analysis to the alternative negative binomial model. Although tuberculosis incidence rates in Thailand are much lower than those in less developed countries such as Nepal

(Kakchapati et al 2010) it is a serious health problem, and the results obtained from such studies are important for reducing the health burden. We found substantially higher rates in five locations in the region, namely, all of Pattani province, all of the rural districts in Yala province, the districts in Naratiwat province bordering Malaysia, and in all districts of Phuket province. The reasons for these differences are yet to be explained, and call for further study.

Although other complex statistical models are now the preferred methods used by biostatisticians for analyzing incidence rates, their advantages are offset by (a) the difficulties of understanding and correctly applying the methods experienced by scientists who lack an adequate knowledge of statistical theory, (b) the lack of availability of software packages and the associated difficulties in using these packages where they exist, and (c) the possible higher risk of bias associated with the use of complex models rather than simpler stratified analyses (Greenland 1989).

**4.3 Densities of macrobenthic fauna families in middle Songkhla Lake**

The objective for the third study was to see if the statistical literature could provide more appropriate methods than those preferred by ecologists for investigating associations between abundances of taxa and their environmental determinants. The preferred method used by ecologists is canonical correspondence analysis, but this method is not widely used in other scientific disciplines, where regression analysis is the most common method of choice. Our study focused on the sediment densities of 24 commonly occurring families of macrobenthic fauna in Middle Songkhla Lake, and their dependence on 12 characteristics of the water and the sediment measured on the same occasions at the same locations.

We thus compared the results obtained from the canonical correspondence analysis with those obtained using multivariate multiple linear regression, where the multivariate outcomes comprised the 24 families, thus allowing for correlations between the observed densities with respect to these families. We concluded that canonical correspondence analysis is useful for informatively displaying the associations in a two-dimensional biplot, but this method is inferior to multivariate multiple linear regression in two important ways: Firstly, the fact that the biplot is restricted to two dimensions can fail to show some associations that are statistically significant but require a further dimension that cannot be seen in such a plot. For example, in our study we found that the *Spionidae* was clearly associated with a salinity factor in the regression analysis, but the biplot failed to show this association.

Secondly, multivariate multiple regression models give fitted values for their outcomes but canonical correspondence analysis does not.

These results are consistent with earlier findings by Warton and Hudson (2004).

**4.4 Suggestions for further study**

Given that the log-transformed linear regression model was found to be superior to the method preferred by biostatisticians in each of our investigations of disease incidence, it would be useful to know whether this finding is supported more generally in such epidemiological studies.

Similarly, it would be useful to apply the multivariate multiple regression method to additional examples of ecological studies.

There are also generalizations of each method that could prove useful. In particular, given that the transformed (univariate) linear multiple regression model is easily extended to multivariate linear multiple regression models as used for the ecological study whereas generalized linear regression models are not so easily extended, it would be useful to investigate this possibility for epidemiological studies.
These models can be extended even further to non-linear regression models such as those with multiplicative components (see, for example, McNeil and Tukey 1973, Lee and Carter 1992, Booth et al 2002, and Sriwattanapongse and Kuning 2009).

Seasonal effect is not only an important factor in common infectious disease such as influenza, chickenpox, and measles but also in non-infectious diseases. So, instead of analyzing the infectious diseases by combining triple monthly according to calendar year, it would be more significant to create a season or period following by monsoon. There are two seasons in southern Thailand: June- August; (south west monsoon season; light rain), December- February (north-east monsoon season; heavy rain). Hot, first rain, slightly rain and heavy rain seasons used in this study are defined as March to May, July to August (after the rain has begun from south-west monsoon), Sepember to November and December to February (heavy rain from North-east monsoon), respectively. However climate change may be indirectly related to other factors which have an important bearing on pneumonia and TB.