STAT3015/4030/7030 Generalised Linear Modelling

Tutorial 11

1. The data file `Sprngs.txt` on Wattle contains data regarding the uncompressed height in inches of truck springs manufactured under various conditions. Five covariates define the manufacturing conditions, and each are two-level factor variables:

   - `ftmp` - furnace temperature (0 = low, 1 = high)
   - `htme` - heating time (0 = short, 1 = long)
   - `ttmp` - transfer time (0 = short, 1 = long)
   - `hdtmp` - hold-down time (0 = short, 1 = long)
   - `qot` - quench oil temperature (0 = low, 1 = high)

   (a) Fit an analysis of variance model to this data and determine the significant factors and interactions. The experiment was intended to identify conditions yielding springs having heights of exactly 8 inches. Which factor settings seem the best ones to achieve this goal?

   (b) The preceding analysis is only valid under the assumption of homoscedasticity. Examine this assumption by comparing the "within group" variances determined by the model you chose in part (a). In other words, calculate the variance of each group of data points for which the predictor variables included in your part (a) model are the same.

   (c) More formally, we can test our homoscedasticity assumption by fitting a GLM to the "within group" variances and determining if any of the predictors are significant. Generally, we would use a gamma GLM with logarithmic link and the same model structure as was chosen in part (a). Is there evidence of heteroscedasticity based on this approach?

   (d) In fact, it can be shown that sample variances tend to have gamma distributions with $\alpha = 0.5$. Does the analysis here seem consistent with this fact?

   (e) For the GLM fit in part (c), calculate the Cook's distances. Do any data points seem problematic? Rerun the analysis of part (c) without these data points. Does this new analysis seem consistent with the fact that sample variances have gamma distributions with $\alpha = 0.5$?

   (f) Use the predicted values for the "within group" variances from the model you fit in part (e) to perform an appropriate weighted ANOVA for the spring heights, thereby

accounting for any potential heteroscedasticity. Does this change your results for the appropriate settings you selected in part (a)?

2. The data file `Eyes.txt` is located on Wattle and contains a contingency table tabulating the eye test information on 7477 female employees aged between 30 and 39 years at the Royal Ordinance Factories. The data tabulates the visual acuity of each of the eyes of the subjects, rated on a four point scale: High, Moderate, Low and Poor. The columns are associated with the visual acuity of the left eye while rows are associated with the right eye.

   (a) Test whether there is independence/homogeneity in this contingency table. Briefly comment on the result of your test and the structure of the observed counts.

   (b) Another hypothesis of interest might be whether the table is "symmetric". In other words, if a woman has eyes which are of two different visual acuities, is the better one equally likely to be the right as the left? To answer this question, we can use a Pearson chi-squared test. The only difficulty is that we need to prescribe the appropriate E_{ij} values under the hypothesis of symmetry and determine the appropriate number of degrees of freedom. Now, the actual hypothesis of symmetry is $H_0 : \pi_{ij} = \pi_{ji}$. For this hypothesis, it can be shown that:

$$E_{ij} = \frac{Y_{ij} + Y_{ji}}{2}.$$

   In addition, the appropriate degrees of freedom is, as usual, the number of parameters which have been removed from the saturated model in the designation of the model under study. Using these facts, do you think that the table is symmetric? If not, where is the asymmetry primarily located in the table?

   (c) A slightly more flexible model to fit to these data is:

$$H_0 : \pi_{ij} = \theta\pi_{ji} \quad i < j,$$

   for some unknown parameter θ. This model implies that if a woman's eyes are of different visual acuities then the odds that it is the right eye which is better is equal to θ. Under this model, it can be shown that appropriate fitted values are:

$$E_{ij} = \begin{cases} \frac{Y_{ij}+Y_{ji}}{1+\widehat{\theta}} & i > j; \\ Y_{ii} & i = j; \\ \frac{\widehat{\theta}(Y_{ij}+Y_{ji})}{1+\widehat{\theta}} & i < j. \end{cases}$$

   where $\widehat{\theta}$ is the MLE of θ. Recall the definition of what θ actually measures and construct a common sense estimate $\widehat{\theta}$. Use your estimated value of θ to test whether the hypothesis H_0 fits the data adequately. Suppose we were asked to construct a 95% confidence interval for θ. Would you expect the value $\theta = 1$ to be in the interval?

3. The following table contains information regarding the one-year survival rates of heart transplant patients. Thirty-nine patients who received heart transplants during the period 1968 to 1971 were tracked to determine whether they survived for at least one year after their operation:

	Year of transplant:			
Survived 1 year	1968	1969	1970	1971
Yes	2	6	2	6
No	7	5	4	7

Use a contingency table analysis (treating both categorical variables as nominal) to test whether the two variables (i.e. one-year survival and transplant year) are independent of one another. Do you think such an analysis is very reliable in this instance?