מודלים סטטיסטיים ויישומיהם 52518 תשע"ח – בוחן בית 3

להגשה עד 24.1.18 בשעה 23:55

1. The concept of two-way ANOVA can be generalized to multi-way ANOVA. For example, a three-way ANOVA model is given by
$$Y_{ijkl} = \mu_{ijk} + \epsilon_{ijkl}$$
with $\epsilon_{ijkl} \underset{iid}{\sim} N(0,1)$, and can be written as
$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \zeta_{ik} + \eta_{jk} + \theta_{ijk} + \epsilon_{ijkl}$$
in terms of weights $\{\pi_i\}, \{\tau_j\}, \{\nu_k\}$. Consider the null hypothesis H_0 that all the two-way and three-way interactions are 0.
    a. Prove that if H_0 holds with respect to one system of weights, it holds with respect to all systems of weights.
    b. Write a program in R to take a set of data and produce a p-value for a test of H_0. You may use vector/matrix operations, but you should not use any built-in statistical procedures. Write an explanation (1-2 pages long) of how your program carries out the computation. Run your program on the attached dataset 'popcorn.csv' (see explanation in 'popcorn.txt').

2. In the context of a four-variable loglinear model with the variables A, B, C, and D, consider the model that includes the following λ terms: A, B, C, AB, AC, AD, BC, BD, CD, ABC, and BCD. Define
$$\psi_{jk}^{AD}(i,i',l,l') = \frac{\pi_{ijkl}\pi_{i'jkl'}}{\pi_{ijkl'}\pi_{i'jkl}}$$
Prove that for fixed i, i', l, l', $\psi_{jk}^{AD}(i,i',l,l')$ does not depend on j or k.

3. The attached file *titanic.csv* contains the Titanic passengers' data, as explained in *titanic.txt*. The response variable is 'Survived' while the other variables are the predictors.
    a. Read the data file into R and remove columns that you won't use in your prediction model. Explain your decisions. After removing these columns, remove rows with missing values for one or more of the data items.

    Warning: in real life, deleting records with some data items missing is often not an appropriate approach, and there are methods for dealing with missing data, but they are not part of the course.

    b. Build a relevant logistic regression model. Which are the more significant predictors?
    c. Run the model checking tools discussed in class and consider the possibility of improving the model by adding nonlinear terms in age or doing a transformation on age. One additional term/transformation is enough, be sure to include all relevant plots.
    d. Add a new column named 'child', a binary variable indicating whether the passenger is younger than 15. Remove the 'age' column and any other variable you've added on (c).
    e. Build a logistic regression model using the updated dataset. In terms of predictors' significance, are there any apparent changes?

When discussing linear regression and ANOVA, a common method for comparing two models is F test. However, this test isn't applicable for logistic regression as $y_i - \hat{y}_i = e_i \in \{0,1\}$. Instead, we will use a tool called Analysis of Deviance (AoD), which is in fact a likelihood ratio test. The relevant R function compares two models using a given test method and you should call it in the following manner:

$$\boldsymbol{anova}(reduced, full, test\ = \text{"Chisq"})$$

where $reduced$ and $full$ are two glm objects representing the two models. The test statistic is the difference in 2 times the loglikelihood between the two models, and the (approximate) distribution, based on Wilks's theorem, is χ^2 with degrees of freedom equal to the difference in the number of parameters between the two models.

f. We would like to examine the popular hypothesis ‘women and children first’. Formulate H_0, H_1 properly.

g. Build two logistic regression models to represent your hypotheses, then examine the p-value for this hypothesis testing process using the aforementioned AoD function.