

ORIE 671: Intermediate Applied Statistics

Homework Assignment 3: Fall 2003 Due Dec. 1, 2003

NOTE: Do computations in R.

1. For the seed germination data discussed in class:
 - (a) Find the variance-covariance matrix of the estimated germination rates for the four treatment combinations under the “overdispersed” binomial counts model.
 - (b) Conduct a Wald test that the germination rates are all equal. Compare your answer with the corresponding test assuming binomial variation. Hand in a summary of your R code and annotated relevant output.
2. Suppose that Y_1 is $B(m, \pi_1)$. Show that

$$\frac{(Y_1 - m\pi_1)^2}{m\pi_1(1 - \pi_1)} = \frac{(Y_1 - m\pi_1)^2}{m\pi_1} + \frac{(Y_2 - m\pi_2)^2}{m\pi_2}$$

where $Y_2 = m - Y_1$ and $\pi_2 = 1 - \pi_1$.

3. (M&N: Problem 5.15) *Logistic discrimination*: Suppose that a population of individuals is partitioned into k sub-populations or groups, G_1, \dots, G_k , with relative frequencies, π_1, \dots, π_k . Multivariate measurements made on individuals in group j are distributed as $N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, for $j = 1, \dots, k$. Let \mathbf{z} be an observation made on an individual drawn at random from the combined population. The prior odds that the individual belongs to G_j rather than G_1 is π_j/π_1 , $j = 2, \dots, k$.
 - (a) Show that the corresponding posterior odds are of the form
$$\frac{\pi_j}{\pi_1} \exp(\alpha_j + \boldsymbol{\beta}_j^t \mathbf{z}),$$
and find expressions for α_j and $\boldsymbol{\beta}_j$ in terms of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}$.
 - (b) What simplifications can be made if the k normal means $\boldsymbol{\mu}_j$ lie in a straight line in R^p ?

- (c) Are the estimates of α_j and β_j obtained by multivariate logistic regression the same as the maximum likelihood estimates based on the normality assumptions? Explain.
4. Fisher's iris data involves measurements taken on samples (of size 50) of three different species of iris: Setosa, Versicolor and Virginica. The data can be included into R via the command "data(iris)". The measurements taken are lengths and widths of the sepals and petals of each of the 150 iris's.
- (a) Assuming that the three species are equally abundant, use the data to find a logistic discrimination rule for identifying the species of a new plant based on the sepal length and width measurements.
- (b) How well are the 150 iris's classified by the discriminant rule? Construct the 3×3 table in which the iris's are cross-classified by their true species and their "predicted" species.
- (c) Plot the data (using different colors and/or symbols for the three species). Include species classification boundaries on your plot.
5. Consider the Berkeley Graduate Admission data discussed in class. Treat the data as a six independent multinomial vectors (one for each department) with categories determined by the four admission \times sex combinations.
- (a) Find a loglinear model that is equivalent (in terms of inferences about the odds of admission) to the binomial GLM implying a common (sex \times admission) odds-ratio for each department. Explain why the models are equivalent.
- (b) Fit the loglinear model to the data from departments 2-5. Use the "anova" command to construct an analysis of deviance table for the model. Tabulate test statistics and p-values for testing goodness-of-fit of the common odds ratio model, and for testing conditional independence (of sex and admission) given department versus a common odds-ratio. Hand in a summary of your code and relevant output.