536 Homework 5

Adrian Dobra adobra@uw.edu

Problem 1. A single-nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide (A, T, C, or G) in the genome differs between members of a species or paired chromosomes in an individual. We want to study whether the genotype associated with a particular SNP (which we generically call "SNP2") is related to the occurrence of Crohn's disease. Its three possible values represent the genotype of this biallelic SNP: 1 (BB), 2 (Bb) and 3 (bb). Here b is the minor allele (proportion of occurrence $\frac{649+2*527}{2*4970} = 0.171$) and B the wildtype allele (proportion of occurrence $\frac{2*3794+649}{2*4970} = 0.829$). The data is Table 1 is a cross-classification of 4970 individuals by SNP2 ("BB" vs. "Bb or bb") and Crohn's disease status ("yes" and "no"). The two binary variables are denoted by X_1 (occurrence of Crohn's disease) and X_2 (absence or presence of the minor allele b at location SNP 2).

		SNP 2 (X_2)		
		BB (1)	Bb or bb (2)	Row Totals
Disease (X_1)	No (1)	2037	958	2995
	Yes (2)	1757	218	1975
	Column Totals	3794	1176	4970

Table 1: An example of a 2×2 table for genotype-disease association.

Analyze the data in Table 1 as follows:

1. Fit the log-linear model of independence and determine if it fits the data well.
2. Derive the logistic regression of X_1 given X_2 from the log-linear model you chose. Provide an interpretation for this regression equation.
3. Use Fisher's exact test to test whether there is an association between the occurrence of disease and the presence/absence of the minor allele b in location SNP2. Make sure you consider all three alternative hypotheses.

Problem 2. Consider the 2×3 categorical data in Table 2. It cross-classifies the same 4970 individuals based on their Crohn's disease status and genotype at locus SNP2. Remark that Table 1 is obtained from Table 2 by collapsing categories Bb or bb of SNP2.

Analyze the data in Table 2 as follows:

		SNP 2 (X_2)			
		BB (1)	***Bb*** (2)	***bb*** (3)	Row Totals
Disease (X_1)	No (1)	2037	631	327	2995
	Yes (2)	1757	18	200	1975
	Column Totals	3794	649	527	4970

Table 2: An example of a 2×3 table for genotype-disease association.

1. Calculate the asymptotic p-value for testing independence vs. interaction.
2. Calculate the exact p-value for testing independence vs. interaction.
3. Based on your choice of log-linear model, derive the regression of disease given SNP 2.