M3S12 BIOSTATISTICS: ASSESSED COURSEWORK 2

Deadline: Friday 30th April

The data below correspond to counts in a small cohort study of randomly selected patients who have been cross-categorized into a three-way, $2 \times 2 \times 2 = 2^3$ table, according to three factors

- presence of coronary heart disease (CHD) YES/NO
- hypertension, measured via diastolic blood pressure (BP) YES/NO
- genotype (genetic configuration) at a specific marker locus (G) - genotype 1, genotype 2

		CHD		
Genotype	BP	Yes	No	Total
1	Yes	15	5	20
	No	40	60	100
	Total	55	65	120
2	Yes	20	10	30
	No	10	40	50
	Total	30	50	80

In the proposed analysis, a log-linear model analysis is to be considered, that is, it is assumed that if N_{ijk} is the count for the $(i, j, k)^{th}$ cell entry in the $2 \times 2 \times 2$, then

$$N_{ijk} \sim Poisson\left(\lambda_{0ijk}\right)$$

where Poisson parameter λ_{0ijk} is to be represented using models of the form

$$\mu_{ijk} = \log \lambda_{0ijk} = x^T \beta$$

where the form of vector parameter β depends on the models being fitted.

(i) Carry out an analysis of deviance using SPLUS, for example the code below, to identify whether there is any difference in model adequacy when different predictors are included/excluded from the model. The following code may be useful.

```
G<-rep(c(1:2),each=4)
BP<-rep(rep(c(“YES”,“NO”),each=2),2)
CHD<-rep(c(“YES”,“NO”),4)
N<-c(15,5,40,60,20,10,10,40)
GBPCHD<-data.frame(factor(G),factor(BP),factor(CHD),N)
names(GBPCHD)<-c(“G”,“BP”,“CHD”,“N”)
summary(glm(N∼G+BP+CHD+G:BP,family=poisson,data=GBPCHD))
```

The SPLUS code above fits the model

$$G + BP + CHD + G\ .\ BP$$

that is, the main effects plus the G/BP interaction. The SPLUS symbol : specifies the interaction term. The symbol $*$ indicates a “main effects plus interaction” term, for example

$$G * BP \equiv G + BP + G\ .\ BP$$

There are eight possible models that could be fitted, including main effects and interactions. Note that the SPLUS command **anova** can be used as follows

anova(glm(****FILL-IN MODEL FORMULA HERE****,family=poisson,data=GBPCHD))

to assist in the analysis of deviance.

This analysis implements a standard "contrast" parameterization; for example, the model above, we have

$$\mu_{ijk} = \log \lambda_{0ijk} = \mu + \alpha_i + \beta_j + \xi_k + \gamma_{ij} \qquad i, j, k = 1, 2$$

where μ is the log expected response in the "baseline" category, which is the $G = 1, BP =$ **"No"** and $CHD =$ **"No"** category. Factor levels are labelled in alphabetical order, and thus

$$G = 1 \Longrightarrow \alpha_1 = 0 \qquad BP = \textbf{"No"} \Longrightarrow \beta_1 = 0 \qquad CHD = \textbf{"No"} \Longrightarrow \xi_1 = 0$$

For the model that you decide is the best fitting model, provide a table of parameter estimates, standard errors and t-statistics, and then describe carefully the implication of this model for the relationship between the three factors. Briefly discuss the adequacy of the fit of the model by inspecting the estimate of dispersion, $\widehat{\phi}_D$

[12 MARKS]

(ii) Show that the log-linear model log-odds ratio for any 2×2 subtable, say (without loss of generality)

$$\log\left(\frac{\mu_{i11}\mu_{i22}}{\mu_{i12}\mu_{i21}}\right) \qquad i = 1, 2$$

can be expressed in terms of parameters that appear in the linear predictor.

[4 MARKS]

(iii) Describe how a GLM analysis would proceed if CHD status was regarded as the binary response variable for each individual, and G and BP were regarded as factor predictors.

[4 MARKS]