MATH3012 – Statistical Methods II
S-Plus Worksheet 9 – Log-linear models for the `lymphoma` data set.

1 Finding a suitable model

1. The `lymphoma` dataset represents classification of 30 lymphoma patients by sex, cell type of lymphoma and response to treatment and it is an example of a three-way contingency table. Use log-linear models to determine how these three variables are associated.

Cell Type	Sex	Remission No	Remission Yes
Nodular	Male	1	4
	Female	2	6
Diffuse	Male	12	1
	Female	3	1

2. First, get the `lymphoma` dataset from the usual place. Now get the source file containing S-Plus commands from the course website, usual place.

3. As usual we first set the treatment contrasts. `options(contrasts=c("contr.treatment", "contr.poly"))`

4. Here the saturated model is the three factor interaction model `Cell * Remis * Sex`. We issue:

   ```
   ly.sat <- glm(y ~ Cell * Remis * Sex, data=lymphoma, family=poisson)
   ```

5. The saturated model fits exactly. `summary(ly.sat)` confirms that we have zero deviance.

6. Now we drop the three factor interaction term. `ly.glm1 <- update(ly.sat, . ~ . - Cell:Remis:Sex)`

   Note the colon instead of the `*` in `Cell:Remis:Sex`. What happens if you use the `*`?

7. Issue `summary(ly.glm1)` followed by `anova(ly.glm1, test="Chisq")`. From the output, we can see that we can remove the `Remis:Sex` term and no more. We cannot remove any of the remaining two interaction terms, `Cell:Remis` and `Cell:Sex` because of low p-values; we cannot remove any lower order terms if they appear in higher order terms. Remember the *principle of marginality*!

8. Issue the following two commands.

   ```
   ly.glm2 <- update(ly.glm1, . ~ . - Remis:Sex)
   summary(ly.glm2)
   ```

9. We issue the following commands to see the quality of fit. Compare the observed and fitted counts!

   ```
   pcount <- predict(ly.glm2, type="response")
   lymphoma$pcount <- pcount # See the lymphoma dataset
   data.frame(observed=lymphoma$y, fitted=pcount)
   ```

2 Investigating the dependence structure

1. Absence of the interaction term `Remis:Sex` from `ly.glm2` does not imply the independence of remission and sex. It merely implies that remission is independent of sex *conditional on* cell type, that is

$$P(R, S|C) = P(R|C)P(S|C).$$

Another way of expressing this is

$$P(R|S, C) = P(R|C),$$

that is, the probability of each level of R given a particular combination of S and C, does not depend on which level C takes. [Equivalently, we can write $P(S|R, C) = P(S|C)$]. This can be observed by calculating the estimated odds in favour of $R =$ yes over $R =$ no for the `lymphoma` dataset.

2. We now illustrate the above theory. We first find the 8 fitted probabilities which are simply the fitted counts divided by 30 (which is the total number of patients classified).

```
fit.prob <- pcount/(sum(pcount))

lymphoma$fitprob <- fit.prob # See the lymphoma dataset

fit.prob
```

Using the above commands (and then by hand) we obtain the following table of fitted probabilities.

Cell Type	Sex	Remission No	Remission Yes
Nodular	Male	0.0385	0.1282
	Female	0.0615	0.2051
Diffuse	Male	0.3824	0.0510
	Female	0.1176	0.0157

Subsequently, we form the odds ratios by dividing the probabilities, e.g. $\frac{0.1282}{0.0385} = 3.33$.

Cell Type	Sex	Remission No	Remission Yes	Odds
Nodular	Male	0.0385	0.1282	3.33
	Female	0.0615	0.2051	3.33
Diffuse	Male	0.3824	0.0510	0.13
	Female	0.1176	0.0157	0.13

Therefore, the odds depend only on a patient's Cell type, and not on their Sex.

	Remission		
Sex	No	Yes	Total
Male	0.4208	0.1792	0.6
Female	0.1792	0.2208	0.4
Total	0.6	0.4	1

	Remission		
Sex	No	Yes	Odds
Male	0.4208	0.1792	0.43
Female	0.1792	0.2208	1.23

3. The above establishes that remission and sex are conditionally independent given cell type. It is easy to see that they are not marginally independent, as the following table (left) demonstrates; the cell probabilities are not the product of the marginal totals.

   From the table on the right hand side, we see that male patients have a much lower probability of remission. The reason for this is that, although R and S are not directly associated, they are both associated with C. Observing the estimated values (last column of the very first table of fitted probabilities) it is clear that patients with C = nodular have a greater probability of remission, and furthermore, that female patients are more likely to have this cell type than males. Hence females are more likely to have R = yes than males. However, the conditional independence of R and S given C implies that two patients with the same cell type are equally likely to have R = yes, even if one is male and the other female.

3 Demonstrating the equivalence of logistic and log-linear models

1. First fit the Poisson GLM.

   ```
   ly.pois <- glm(y ~ Cell+ Sex+ Remis+ Cell:Remis + Cell:Sex +
   Sex:Remis, data=lymphoma, family=poisson)
   summary(ly.pois)
   ```

2. Prepare the data for logistic regression.

   ```
   makepropdata  <- function() {
        Cell <- c("nodular", "nodular", "diffuse", "diffuse")
        Sex <- c("male", "female", "male", "female")
        y <- c(4, 6, 1, 1)
        n <- c(5, 8, 13, 4)
        data.frame(Cell=Cell, Sex=Sex, y=y, n=n)
   }
   newlymphoma <- makepropdata() # Bring up this data set
   ly.bino <- glm(y/n ~ Sex+Cell, data=newlymphoma, family=binomial, weights=n)
   summary(ly.bino)
   ```

3. Compare the coefficients. Issue the commands `coef(ly.pois)` and `coef(ly.bino)`. Using the equivalence interpret the parameter estimates.

4. The binomial proportion model is equivalent to models from equivalent Bernoulli observations.

   ```
   convbindata <- function()
   ```

```
{
Cell <- rep(lymphoma[, 2], lymphoma[, 1])
Sex <- rep(lymphoma[, 3], lymphoma[, 1])
Remis <- rep(lymphoma[, 4], lymphoma[, 1])
data.frame(Cell, Sex, Remis )
}
nlmph <- convbindata() # Bring up this data set
ly.bino2 <- glm(Remis~Sex+Cell, data=nlmph, family=binomial)
summary(ly.bino2)

coef(ly.pois)
coef(ly.bino)
coef(ly.bino2)
```

Exercises: Demonstrate this equivalence for the `heartattack` dataset. *Hint: See the source file for this exercise sheet.*