# MATH3823 - Solutions to Chapter 6 Exercises

**Exercise 6.1**

a. If $Y \sim P(\lambda)$, then $E(Y) = \lambda,\ \text{var}(Y) = \lambda,\ E(Y^2) = \lambda + \lambda^2$. Hence for the mixture model

$$E(Y) = E(Y|X=1)P(X=1) + E(Y|X=2)P(X=2) = \frac{1}{2}(\lambda_1 + \lambda_2)$$
$$E(Y^2) = E(Y^2|X=1)P(X=1) + E(Y^2|X=2)P(X=2) = \frac{1}{2}(\lambda_1 + \lambda_1^2 + \lambda_2 + \lambda_2^2),$$

so that

$$\text{Var}(Y) = \frac{1}{2}(\lambda_1 + \lambda_1^2 + \lambda_2 + \lambda_2^2) - \left\{\frac{1}{2}(\lambda_1 + \lambda_2)\right\}^2 = \lambda + (\lambda_1 - \lambda_2)^2/4.$$

b. Writing $\overline{Y} = \frac{1}{2}(\overline{Y}_1 + \overline{Y}_2)$, where $\overline{Y}_j$ has mean $\lambda_j$ and variance $\lambda_j/30$, $j = 1, 2$, we see that

$$E[(\overline{Y})^2] = \frac{1}{4}E\left[(\overline{Y}_1)^2 + 2\overline{Y}_1\overline{Y}_2 + (\overline{Y}_2)^2\right] = \frac{1}{4}\{\lambda_1^2 + \lambda_1/30 + 2\lambda_1\lambda_2 + \lambda_2^2\ + \lambda_2/30\}.$$

Also

$$E[\sum_{i=1}^{60} Y_i^2] = E[\sum_{i=1}^{30} Y_i^2] + E[\sum_{i=31}^{60} Y_i^2] = 30(\lambda_1^2 + \lambda_1 + \lambda_2^2 + \lambda_2).$$

Hence

$$E(S^2) = \frac{1}{59}E\left[\sum_{i=1}^{60} Y_i^2 - 60(\overline{Y})^2\right] = \frac{1}{2}(\lambda_1 + \lambda_2) + \frac{60}{4 \cdot 59}(\lambda_1 - \lambda_2)^2,$$

which is larger than $\lambda$ when $\lambda_1 \neq \lambda_2$.

The $\chi^2$ goodness of fit statistic is just

$$X^2 = \sum_k (O_k - E_k)^2/E_k = (n-1)S^2/\overline{Y},$$

in terms of observed values $O_i = Y_i$ and expected values $E_i = \overline{Y}$. Here $n = 60$. Under the null Poisson model, the numerator and denominator have the same expectation and for large n, $X^2 \sim \chi^2_{n-1}$ approximately. Under the mixture model the numerator has a larger expectation and $X^2$ will be larger in distribution than $\chi^2_{n-1}$.

c. When a scale parameter $\phi$ is present, it can be used to represent the true scale variability in the model, plus the effects of overdispersion.

**Exercise 6.2 I Percentages adding to 100% across housing**

**Low Contact:**

| Satisfaction | Low | Medium | High |
|---|---|---|---|
| Tower blocks | 25 | 30 | 37 |
| Apartments | 50 | 43 | 41 |
| Houses | 26 | 27 | 23 |
| Total | 100 | 100 | 100 |

**High Contact:**

| Satisfaction | Low | Medium | High |
|---|---|---|---|
| Tower blocks | 11 | 48 | 25 |
| Apartments | 46 | 43 | 48 |
| Houses | 43 | 39 | 26 |
| Total | 100 | 100 | 100 |

**II Percentages adding to 100% across satisfaction**

**Low Contact:**

| Satisfaction | Low | Medium | High | Total |
|---|---|---|---|---|
| Tower blocks | 30 | 25 | 46 | 100 |
| Apartments | 41 | 24 | 35 | 100 |
| Houses | 38 | 27 | 35 | 100 |

**High Contact:**

| Satisfaction | Low | Medium | High | Total |
|---|---|---|---|---|
| Tower blocks | 19 | 26 | 55 | 100 |
| Apartments | 31 | 26 | 43 | 100 |
| Houses | 38 | 31 | 31 | 100 |

**III Percentages adding to 100% across contact**

**Low Satisfaction:**

| Contact | Low | High | Total |
|---|---|---|---|
| Tower blocks | 66 | 34 | 100 |
| Apartments | 48 | 52 | 100 |
| Houses | 34 | 66 | 100 |

**Medium Satisfaction:**

| Contact | Low | High | Total |
|---|---|---|---|
| Tower blocks | 53 | 47 | 100 |
| Apartments | 40 | 60 | 100 |
| Houses | 31 | 69 | 100 |

**High Satisfaction:**

| Contact | Low | High | Total |
|---|---|---|---|
| Tower blocks | 50 | 50 | 100 |
| Apartments | 37 | 63 | 100 |
| Houses | 37 | 63 | 100 |

The three tables give percentages adding to 100 over housing, satisfaction and contact, respectively.

The $R$ output shows the result of fitting a model with all 3 *first-order* interactions (also known as *two-way* interactions) but no *second-order* interaction (*three-way* interaction). The deviance $6.89 \sim \chi^2_4$ is not significant, so there is no need to consider the saturated model. The residuals all lie between $\pm 2$, again indicating the model is an adequate fit. There are strongly significant coefficients within each first-order interaction; hence, further simplification has not been attempted. For this dataset it does not seem reasonable to condition on any of the marginal totals so the counts are assumed to arise from a Poisson model, not a product-multinomial model.

```
count = c(65,54,100,34,47,100,130,76,111,141,116,191,67,48,62,130,105,104)

sat=rep(1:3,6)
housing = rep(1:3,rep(6,3))
contact=rep(rep(1:2,rep(3,2)),3)

# Convert into factors
sat = as.factor(sat)
housing = as.factor(housing)
contact = as.factor(contact)

## Modelling
glm1 = glm(count ~ sat+housing+contact+
             sat:housing+housing:contact+sat:contact,
           poisson)
summary(glm1)
```

```
##
## Call:
## glm(formula = count ~ sat + housing + contact + sat:housing +
##     housing:contact + sat:contact, family = poisson)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         4.0943     0.1127  36.338  < 2e-16 ***
## sat2               -0.1073     0.1524  -0.704 0.481589
## sat3                0.5608     0.1329   4.219 2.46e-05 ***
## housing2            0.7402     0.1302   5.687 1.30e-08 ***
## housing3            0.2395     0.1417   1.690 0.090995 .
## contact2           -0.4306     0.1293  -3.331 0.000867 ***
## sat2:housing2      -0.4068     0.1713  -2.375 0.017570 *
## sat3:housing2      -0.6416     0.1501  -4.275 1.91e-05 ***
## sat2:housing3      -0.3371     0.1804  -1.869 0.061627 .
## sat3:housing3      -0.9456     0.1645  -5.749 8.98e-09 ***
## housing2:contact2   0.5744     0.1256   4.575 4.76e-06 ***
## housing3:contact2   0.8906     0.1387   6.419 1.37e-10 ***
## sat2:contact2       0.2960     0.1301   2.275 0.022909 *
## sat3:contact2       0.3282     0.1182   2.777 0.005483 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 294.477  on 17  degrees of freedom
## Residual deviance:   6.893  on  4  degrees of freedom
## AIC: 148
##
## Number of Fisher Scoring iterations: 4
```

```
residuals(glm1, type="pearson")
```

```
##           1           2           3           4           5           6
##  0.64620407  0.01457774 -0.49864032 -0.80142840 -0.01559242  0.52481845
##           7           8           9          10          11          12
##  0.37705287  0.08967078 -0.46480966 -0.35088696 -0.07196879  0.36708512
##          13          14          15          16          17          18
## -1.05756656 -0.12654027  1.40479462  0.84025074  0.08670781 -0.94719781
```

Since there is no second-order interaction, the log odds ratios for any two variables are constant given the third. Here are some interpretations based on the tables of percentages and the coefficients in the $R$ output.

- For each level of contact, Tables I and II show that the ratio of high to low satisfaction is higher for tower block residents than for apartment dwellers, which is in turn higher than for house residents. This interpretation is confirmed by the coefficients `sat3.housing2` $= -0.64$ and `sat3.housing3` $= -0.95$.
- For each level of satisfaction, Tables I and III show that contact with neighbours increases as one moves from housing category tower block to apartment to house. This interpretation is confirmed by the coefficients `housing2.contact2` $= 0.57$ and `housing3.contact2` $= 0.89$.
- For each level of housing, Tables II and III show that high satisfaction goes with high contact. This interpretation is confirmed in particular by the coefficient `sat3.contact2` $= 0.33$.

For me the overall conclusions are partly expected and partly unexpected. I am surprised that tower blocks have a higher satisfaction than houses, and that house residents have greater contact than tower block residents. But it is natural that higher satisfaction is associated with higher contact.

**End of Solutions to Chapter 6 Exercises**