STAT3015/4030/7030 Generalised Linear Modelling
Tutorial Week 12

1. A weekly lottery is conducted in which numbered and colored balls are chosen by a physical randomising device to determine the winners. The randomising machine contains 54 balls (consisting of six different colored sets of balls numbered 1 through 9) which are mixed by a draught of air in a closed, transparent container, and six balls are allowed to escape, one at a time. The randomisation of the machine is to be examined, and the number of times that each ball appears in the weekly winning six is tabulated over a one year period (52 weeks). The total number of times that each ball is in the winning six is tabulated in the data file `Lot.txt` on Wattle.

   (a) Clearly, if the machine is truly producing random balls, then the number and color of the winning balls should be independent of each other. Test this fact by directly calculating expected ball counts and constructing the Pearson chi-squared statistic.

   Solution: The necessary *R* commands are:

```
> lot <- read.table("Lot.txt", header=TRUE)
> attach(lot)
> lot

       no.1 no.2 no.3 no.4 no.5 no.6 no.7 no.8 no.9
red       5    8    5    6    9    4    7    7    6
blue      9    5    6    3    5    4    5    5    3
yellow   10    6    3    7   10    9    2    9    7
green     1    7   11    6    2    9   10    4    8
purple    4    9    6    5   10    3    7    8    3
orange    4    4    1    4    3    5    4    4    5

> rtot <- apply(lot, 1, sum)
> ctot <- apply(lot, 2, sum)
> eij <- rtot%*%t(ctot)/sum(rtot)
> pres <- (lot-eij)/sqrt(eij)
> c(sum(pres^2), 1-pchisq(sum(pres^2), (9-1)*(6-1)))

[1] 43.2834741  0.3330006
```

   So, the colors and the numbers of the balls do appear to be independent, as we would expect. (NOTE: 14/54=26% of the $\mathbb{E}_{ij}$'s are less than 5. However, only 9/54=17% of the $\mathbb{E}_{ij}$'s are less than 4.5, so the test is probably reasonably reliable.)

There can only be checked by examining E_{ij}'s (not shown here)

(b) Test the independence again, this time using a Poisson GLM approach. In other words, fit a Poisson model with link structure:

$$\ln(\mathbb{E}\{Y_{ij}\}) = \beta_0 + \beta_1 r_2 + \cdots + \beta_5 r_6 + \beta_6 c_2 + \cdots + \beta_{13} c_9 + \beta_{14} r_2 c_2 + \cdots + \beta_{53} r_6 c_9,$$

where r_i is the indicator for the i^{th} row and c_j is the indicator for the j^{th} column, and test whether $\beta_{14} = \cdots = \beta_{53} = 0$.

Solution: The necessary *R* commands are:

```
> oij <- as.matrix(lot)
> yij <- as.vector(oij)
> rfact <- rep(1:6, 9)
> cfact <- rep(1:9, rep(6, 9))
> rind2 <- ifelse(rfact==2, 1, 0)
> rind3 <- ifelse(rfact==3, 1, 0)
> rind4 <- ifelse(rfact==4, 1, 0)
> rind5 <- ifelse(rfact==5, 1, 0)
> rind6 <- ifelse(rfact==6, 1, 0)
> cind2 <- ifelse(cfact==2, 1, 0)
> cind3 <- ifelse(cfact==3, 1, 0)
> cind4 <- ifelse(cfact==4, 1, 0)
> cind5 <- ifelse(cfact==5, 1, 0)
> cind6 <- ifelse(cfact==6, 1, 0)
> cind7 <- ifelse(cfact==7, 1, 0)
> cind8 <- ifelse(cfact==8, 1, 0)
> cind9 <- ifelse(cfact==9, 1, 0)
> rinds <- cbind(rind2, rind3, rind4, rind5, rind6)
> cinds <- cbind(cind2, cind3, cind4, cind5, cind6, cind7, cind8, cind9)
> lot.glm <- glm(yij~rinds*cinds, family=poisson)
> anova(lot.glm, test="Chisq")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: yij

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                           53     61.787
rinds        5   11.577        48     50.210  0.04107 *
cinds        8    2.115        40     48.095  0.97727
```

```
rinds:cinds 40   48.095         0      0.000  0.17781
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> 1-pchisq(48.09458,40)
```

```
[1] 0.1778144
```

So, as before, we can see that the interaction terms are not significant in the model and the two variables can be reasonably assumed to be independent.

(c) If the winning balls are truly random, there should also be no row or column effects (i.e., the marginal distributions should be uniform, so that the chance of being in any row or column is the same for all the rows and columns, meaning that $\beta_1 = \cdots = \beta_5 = 0$ and $\beta_6 = \cdots = \beta_{13} = 0$). Use your Poisson GLM from part b to test these hypotheses. Refit the model without the interaction terms and examine the estimated row and column effects. Do there appear to be any rows or columns which are out of line with the hypothesis of true randomness?

Solution: Using the analysis of deviance table from part b, we see that:

```
> 1-pchisq(2.11547,8)
```

```
[1] 0.9772657
```

```
> 1-pchisq(11.57692,5)
```

```
[1] 0.04106808
```

So, it appears that the column effects are not significant, but there is a significant row effect. In other words, it appears that some of the colored balls appear more frequently than others. Refitting the model and examining the estimated effect values we have:

```
> lot.glm1 <- glm(yij~rinds+cinds, family=poisson)
> summary(lot.glm1)$coef
```

```
               Estimate Std. Error     z value     Pr(>|z|)
(Intercept)  1.79655564  0.2112860  8.50295758 1.848205e-17
rindsrind2  -0.23638878  0.1994143 -1.18541510 2.358534e-01
rindsrind3   0.10008346  0.1828028  0.54749403 5.840394e-01
rindsrind4   0.01739174  0.1865080  0.09324931 9.257055e-01
rindsrind5  -0.03571808  0.1890124 -0.18897219 8.501146e-01
rindsrind6  -0.51669074  0.2166925 -2.38444260 1.710502e-02
cindscind2   0.16705408  0.2365250  0.70628523 4.800108e-01
cindscind3  -0.03077166  0.2480988 -0.12402984 9.012916e-01
cindscind4  -0.06252036  0.2501222 -0.24995929 8.026188e-01
```

```
cindscind5   0.16705408  0.2365250  0.70628523 4.800108e-01
cindscind6   0.02985296  0.2443661  0.12216491 9.027684e-01
cindscind7   0.05884050  0.2426406  0.24250064 8.083923e-01
cindscind8   0.11441035  0.2394370  0.47783079 6.327706e-01
cindscind9  -0.03077166  0.2480988 -0.12402984 9.012916e-01
```

So, it appears that the last row is the only one with a significant t-statistic. Thus, it seems that the orange balls are out of line with the rest, and since the t-statistic is negative this means that orange balls are appearing significantly less frequently than the other balls.

(d) Suppose that we are now told that the orange balls were only added after 24 weeks of the lottery had been run. Thus, clearly there will be a row effect for this color ball. However, if the data is truly random, then the other rows should have no effect. Fit a Poisson GLM to test whether the other rows are uniform (i.e., test whether the model $\ln(\mathbb{E}\{Y_{ij}\}) = \beta_0 + \beta_5 r_6$ is adequate for this data). Moreover, a bit of algebra shows that if the orange balls were added after week twenty-four and the balls were truly random than it should be the case that $\beta_5 = -0.7073$. Test whether the observed data is consistent with this value.

Solution: Fitting the appropriate model in R shows:

```
> lot.glm2 <- glm(yij~rind6+cbind(rind2, rind3, rind4, rind5), family=poisson)
> anova(lot.glm2, test="Chisq")

Analysis of Deviance Table

Model: poisson, link: log

Response: yij

Terms added sequentially (first to last)


                                  Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                                 53     61.787
rind6                              1   8.3263        52     53.461 0.003908 **
cbind(rind2, rind3, rind4, rind5)  4   3.2506        48     50.210 0.516793
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> 1-pchisq(3.250624, 4)

[1] 0.5167933

> summary(lot.glm2)$coef[2, ]
```

```
   Estimate  Std. Error     z value    Pr(>|z|)
-0.51669074  0.21669239 -2.38444336  0.01710498

> (-0.5166907-(-0.7073))/0.2166924

[1] 0.8796308

> lot.glm3 <- glm(yij~rind6, family=poisson)
> summary(lot.glm3)$coef[2, ]

    Estimate   Std. Error       z value     Pr(>|z|)
-0.491822677  0.181683335 -2.707032407  0.006788763
```

So, it appears that the other columns are uniformly distributed. In addition, the observed value of the effect for the last row is within the critical t-test value, indicating that the theory is plausible (also, from R we can find out that $t_{54-2}(0.975) = 2.006647$ and $t_{54-6}(0.975) = 2.010635$). Thus, it appears that the randomness of the mixing machine is genuine.

We now give the algebra required to show how the value -0.7073 was derived, for those who are interested. We note that if the balls were indeed truly randomly generated, then during the first 24 weeks there were 6 × 24 = 144 balls drawn and we would expect them to be evenly spread among the first five rows (i.e., after 24 weeks the first 5 rows would be expected to have cell counts of 144/45=3.2). Similarly, during the remaining 28 weeks, there were 168 balls drawn and we would expect them to be evenly distributed among all six rows (i.e., during the remaining time, there should be an addition of 168/54=3.1111 balls to each cell). Therefore, under the model which contains only an intercept and an indicator of the last row, we would have cell expectations:

$$\begin{aligned} \mathbb{E}(Y_{ij}) &= 3.2 + 3.1111 = 6.3111, \quad 1 \le i \le 5, 1 \le j \le 9; \\ \mathbb{E}(Y_{6j}) &= 3.1111, \quad 1 \le j \le 9. \end{aligned}$$

Therefore, the theoretical value for the β's in the model $\ln \mathbb{E}Y_{ij} = \beta_0 + \beta_5 r_6$ would be:

$$\begin{aligned} \beta_0 &= \ln(\mathbb{E}\{Y_{ij}\}) = \ln 6.3111 = 1.8423, \quad 1 \le i \le 5, 1 \le j \le 9; \\ \beta_5 &= \ln(\mathbb{E}\{Y_{6j}\}) - \beta_0 = \ln 3.1111 - 1.8423 = -0.7073, \quad 1 \le j \le 9. \end{aligned}$$

2. The data file `HWCon.txt` is located on Wattle contains counts concerning the performance on homework assignments of schoolchildren learning from five different teaching styles and philosophies. The different teaching regimes were labeled A through E and the quality of the submitted homeworks was rated as either high, moderate or low.

   (a) Test whether the two categorical variables (i.e., teaching regime and homework quality) are independent using a Pearson chi-squared test and treating both variables as nominal.

Solution: The necessary *R* commands are:

```
> hw <- read.table("HWCon.txt", header=TRUE)
> attach(hw)
> hw

       A  B   C  D  E
high 114 35 147 73 65
mod  141 37 130 79 68
low   40 18  31 27 14

> rtot <- apply(hw, 1, sum)
> ctot <- apply(hw, 2, sum)
> eij <- rtot%*%t(ctot)/sum(rtot)
> pres <- (hw-eij)/sqrt(eij)
> c(sum(pres^2), 1-pchisq(sum(pres^2), (3-1)*(5-1)))

[1] 12.4826693  0.1309326
```

So, it appears that the variables are independent.

(b) Clearly, the quality of the submitted homework is an ordinal variable. Using this variable as the response, fit the proportional hazards model (i.e., the weighted complementary log-log binomial GLM to the appropriate transformed proportions) and test whether there is a significant column effect (i.e., that there is a difference in the quality of submitted homeworks for the different teaching regimes). Also, examine the suitability of the proportional hazards model for these data using the residual deviance.

Solution: The necessary *R* commands are:

```
> nij <- sweep(-rbind(0, apply(hw, 2, cumsum)[1, ]), 2, ctot, FUN="+")
> zij <- hw[1:2, ]/nij
> rfct <- as.vector(matrix(rep(1:2, 5), ncol=5))
> cfct <- as.vector(matrix(rep(1:5, 2), ncol=5, byrow=TRUE))
> prp <- c(zij$A, zij$B, zij$C, zij$D, zij$E)
> wgt <- as.vector(nij)
> hw.glm <- glm(prp~factor(rfct)+factor(cfct), family=binomial(link=cloglog),
+                weights=wgt)
> anova(hw.glm, test="Chisq")

Analysis of Deviance Table

Model: binomial, link: cloglog
```

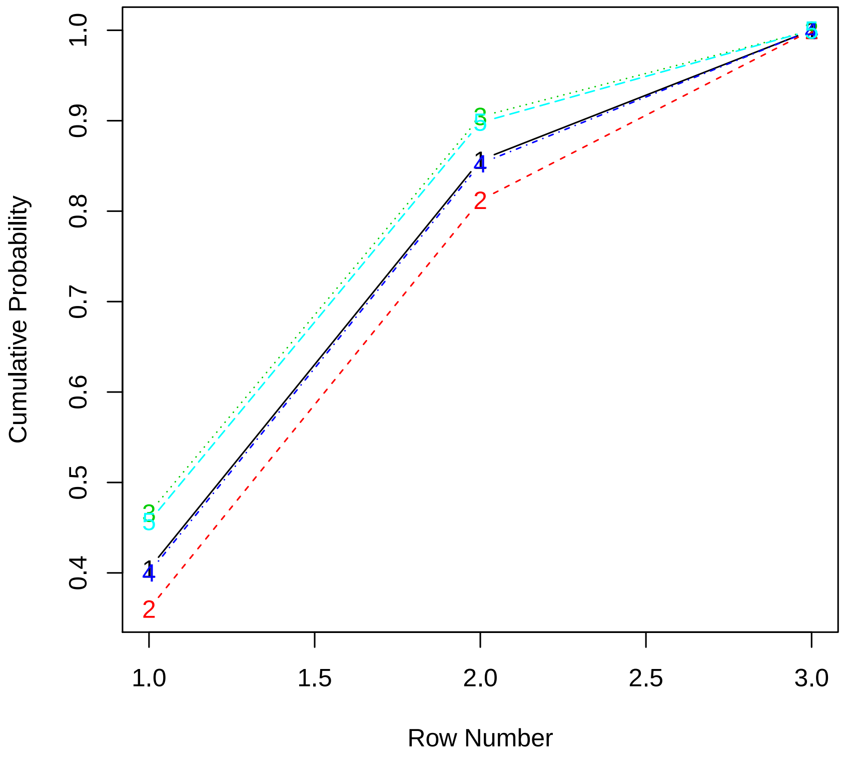
```
Response: prp

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                             9    206.885
factor(rfct)  1  194.772         8     12.113  < 2e-16 ***
factor(cfct)  4    9.963         4      2.150  0.04106 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> 1-pchisq(9.9629,4)

[1] 0.04105729

> 1-pchisq(2.1498,4)

[1] 0.7082295
```

So, it now appears that there is indeed a difference in the columns, so that some of the teaching regimes perform better than others at producing student homeworks of higher quality. Also, it appears that the proportional hazards model is reasonable for these data, at least in terms of the residual deviance. (NOTE: Technically, we should look and see if the dispersion estimate is too small as well as too large. To do this, we have to look at the other tail of the chi-squared distribution. In other words, we want to look at `pchisq(2.1498, 4)`, which represents the chance that we would have seen a dispersion estimate as low as we did if the model were truly adequate. In this case, this value is 0.2917705, which is reasonably large, so our model does seem adequate.)

(c) Plot the fitted cumulative probabilities for each teaching regime. Which regimes are the best and which the worst (at least in terms of the assessed quality of the submitted homework)?

Solution: The necessary *R* commands are:

```
> hatzij <- matrix(fitted(hw.glm), ncol=5)
> cumprbs <- matrix(0, 2, 5)
> cumprbs[1,] <- hatzij[1, ]
> for(i in 2:2) {
+   cumprbs[i,] <- hatzij[i, ]+(cumprbs[i-1, ]*(1-hatzij[i, ]))
+ }
> matplot(1:3,rbind(cumprbs, 1), xlab="Row Number", type="b",
+         ylab="Cumulative Probability")
```



So, from this plot we see that regimes C and E appear to be the best, while regime B appears to be the worst. Recall that the curves which are higher on the plot correspond to those columns where the initial rows are more prevelant, and in this case that means these columns have a larger proportion of high quality homeworks.