STAT3015/4030/7030 Generalised Linear Modelling

Tutorial 8

1. For the purposes of setting hull construction standards, the rate of damage by waves to certain types of cargo ships needs to be determined. Ships of three different types were examined and the cumulative data is given below, as well as in the file `Wave.txt` on Wattle:

Ship Type	Year of Construction	Period of Operation	Aggregate Months Service	Number of Damage Incidents
A	1960-64	1960-74	127	0
A	1960-64	1975-79	63	0
A	1965-69	1960-74	1095	3
A	1965-69	1975-79	1095	4
A	1975-79	1975-79	2244	11
B	1960-64	1960-74	44882	39
B	1960-64	1975-79	17176	29
B	1965-69	1960-74	28609	58
B	1965-69	1975-79	20370	53
B	1975-79	1975-79	7117	18
C	1960-64	1960-74	1179	1
C	1960-64	1975-79	552	1
C	1965-69	1960-74	781	0
C	1965-69	1975-79	676	1
C	1975-79	1975-79	274	1

Clearly, a good approximation to the distribution of the number of damage incidents would be the Poisson distribution. Without any other information, it seems reasonable to start by employing the canonical link function, which is the logarithm in this case. Of course, the number of damage incidents is not necessarily the best reponse variable in this case, since clearly the total amount of time in service is important. So, modelling the rates of damage incidents appears to be a more pertinent approach. Our model for the expected response then becomes:

$$\log\left(\frac{\texttt{dmge}}{\texttt{mnths}}\right) = \beta_0 + \beta_1\texttt{typb} + \beta_2\texttt{typc} + \beta_3\texttt{cons65} + \beta_4\texttt{cons75} + \beta_5\texttt{opr75},$$

where `typb` and `typc` are indicators of the second two ship types, respectively, and `cons65`, `cons75` and `opr75` are indicators of the obvious categories for year of construction and period of service.

(a) Use R to fit this model, recalling that the rates actually have a Poisson$(\lambda T)/T$ distribution, which means that we must take account of the different number months of observation for each data point by employing the weights option.

Solution: The required R commands are:

```
> wave <- read.table("Wave.txt",header=T)
> attach(wave)
> names(wave)

[1] "typ"   "cons"  "opr"   "mnths" "dmge"

> wave

   typ    cons     opr mnths dmge
1    A 1960-64 1960-74   127    0
2    A 1960-64 1975-79    63    0
3    A 1965-69 1960-74  1095    3
4    A 1965-69 1975-79  1095    4
5    A 1975-79 1975-79  2244   11
6    B 1960-64 1960-74 44882   39
7    B 1960-64 1975-79 17176   29
8    B 1965-69 1960-74 28609   58
9    B 1965-69 1975-79 20370   53
10   B 1975-79 1975-79  7117   18
11   C 1960-64 1960-74  1179    1
12   C 1960-64 1975-79   552    1
13   C 1965-69 1960-74   781    0
14   C 1965-69 1975-79   676    1
15   C 1975-79 1975-79   274    1

> typb <- ifelse(typ=="B", 1, 0)
> typc <- ifelse(typ=="C", 1, 0)
> cons65 <- ifelse(cons=="1965-69", 1, 0)
> cons75 <- ifelse(cons=="1975-79", 1, 0)
> opr75 <- ifelse(opr=="1975-79", 1, 0)
> rtes <- dmge/mnths
> wave.glm <- glm(rtes~typb+typc+cons65+cons75+opr75, family=poisson,
                  weights=mnths)
> summary(wave.glm)$coefficients

              Estimate Std. Error    z value      Pr(>|z|)
(Intercept) -6.4917786  0.2953465 -21.980211 4.453892e-107
```

```
typb        -0.4500336  0.2647756  -1.699680  8.919119e-02
typc        -0.8566385  0.5602355  -1.529069  1.262474e-01
cons65       0.6422717  0.1532117   4.192054  2.764402e-05
cons75       0.6449072  0.2513088   2.566195  1.028211e-02
opr75        0.4090553  0.1475958   2.771456  5.580614e-03
```

Notice that the effects due to ships are not quite significant. Perhaps this is due to a confounding interaction effect, which we are about to examine in the next part of the question.

Alternative code for model fit using `offset` term instead of `weights`

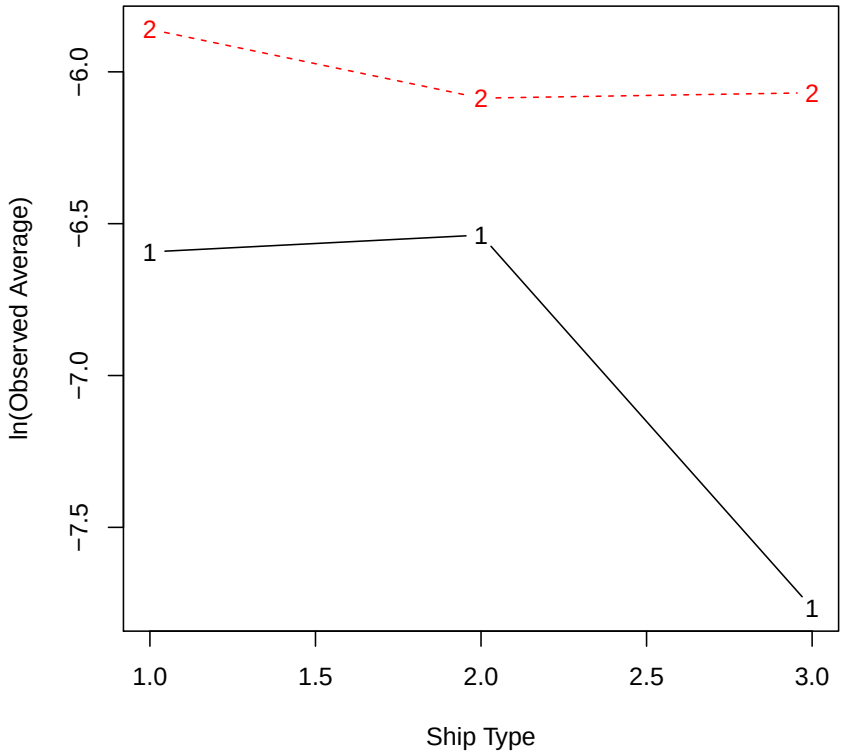
```
> wave.glm <- glm(dmge~offset(log(mnths))+typb+typc+cons65+cons75+opr75,
                  family=poisson)
> summary(wave.glm)$coefficients
```

```
              Estimate Std. Error    z value      Pr(>|z|)
(Intercept) -6.4917786  0.2953465 -21.980210 4.453939e-107
typb        -0.4500336  0.2647756  -1.699680  8.919119e-02
typc        -0.8566385  0.5602355  -1.529068  1.262475e-01
cons65       0.6422717  0.1532117   4.192054  2.764405e-05
cons75       0.6449072  0.2513088   2.566195  1.028211e-02
opr75        0.4090553  0.1475958   2.771456  5.580615e-03
```

(b) Examine the potential need for an interaction term between type of ship and period of operation in the model by plotting the logarithms of the observed averages within each ship type and period of service category combination (i.e., ignore interaction with year of construction by averaging the values for the three different levels of year of construction) against ship type and connecting the points associated with similar period of service category (note that this is the analog of the so-called "cell-means" plot for two-way ANOVA). What should this plot look like if there is no interaction between the two variables? Do you think an interaction term is necessary?

Solution: To create the desired plot:

```
> avg <- tapply(rtes, list(typ,opr), mean)
> matplot(c(1, 2, 3), log(avg), xlab="Ship Type",
+         ylab="ln(Observed Average)", type="b")
```



If there was no interaction the lines of the plot would be parallel. The plot does seem to indicate that some interaction may be present, though it is difficult to say conclusively.

(c) Fit the model with an interaction term between ship type and period of operation and examine whether the effect appears statistically significant. Do the results bear out your visual assessment in part (b)?

Solution: To test the interaction, we can look at the coefficients for the interaction indicators:

```
> int1 <- typb*opr75
> int2 <- typc*opr75
> wave.glm1 <- glm(rtes~typb+typc+cons65+cons75+opr75+int1+int2,
+                  family=poisson, weights=mnths)
> summary(wave.glm1)$coefficients

             Estimate Std. Error     z value     Pr(>|z|)
(Intercept) -6.6036013  0.5952118 -11.0945405 1.333435e-28
typb        -0.3284856  0.5893262  -0.5573918 5.772597e-01
typc        -1.2851065  1.1561472  -1.1115422 2.663350e-01
```

```
cons65       0.6446099  0.1535293   4.1986104 2.685579e-05
cons75       0.6373067  0.2534923   2.5141070 1.193342e-02
opr75        0.5486146  0.6503675   0.8435455 3.989234e-01
int1        -0.1596733  0.6583862  -0.2425222 8.083756e-01
int2         0.6730588  1.3210209   0.5094990 6.104025e-01
```

The last two lines of the above table show that the interaction coefficients do not appear significant. Of course, we really would like to do a formal test similar to the one for linear regression models. This can indeed be done through the so-called deviance table:

```
> int <- cbind(int1,int2)
> wave.glm2 <- glm(rtes~typb+typc+cons65+cons75+opr75+int,
+                  family=poisson, weights=mnths)
> anova(wave.glm2, test="Chisq")

Analysis of Deviance Table

Model: poisson, link: log

Response: rtes

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      14     50.493
typb    1   4.1702        13     46.322 0.0411416 *
typc    1   6.0436        12     40.279 0.0139565 *
cons65  1  13.5869        11     26.692 0.0002278 ***
cons75  1  13.4679        10     13.224 0.0002427 ***
opr75   1   7.5665         9      5.658 0.0059464 **
int     2   0.6240         7      5.034 0.7319802
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In this case, we have a deviance of 0.6240 which is clearly much smaller than $\chi_{(2)}(0.95) = 5.99$, and thus we cannot reject the null hypothesis. In other words, the interaction term seems unnecessary.

(d) The initial model form can clearly be rewritten as:

$$\log(\texttt{dmge}) = \beta_0 + \beta_1\texttt{typb} + \beta_2\texttt{typc} + \beta_3\texttt{cons65} + \beta_4\texttt{cons75} + \beta_5\texttt{opr75} + \log(\texttt{mnths}).$$

So, we could fit a Poisson generalised linear model to `dmge` using the predictors `typb`, `typc`, `cons65`, `cons75`, `opr75` and `log(mnths)`. What would we expect the

coefficient for the predictor log(`mnths`) to be? Fit the appropriate model and test whether this value is compatible with the actual observed data.

Solution: If we fit the stated model, we would certainly expect the coefficient associated with the predictor mnths to be equal to 1. Fitting the model, we have:

```
> wave.glm3 <- glm(dmge~log(mnths)+typb+typc+cons65+cons75+opr75,
                   family=poisson)
> summary(wave.glm3)$coefficients

              Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -6.7702933  1.8573035 -3.6452272 2.671557e-04
log(mnths)   1.0334744  0.2199790  4.6980592 2.626454e-06
typb        -0.5245450  0.5547128 -0.9456155 3.443448e-01
typc        -0.8142343  0.6274464 -1.2976955 1.943920e-01
cons65       0.6498637  0.1613113  4.0286307 5.610266e-05
cons75       0.6705229  0.3035486  2.2089475 2.717829e-02
opr75        0.4275957  0.1914953  2.2329305 2.555353e-02
```

So, a t-test for whether the coefficient of `log(mnths)` is equal to one would have a test statistic of:

$$T = \frac{1.0334744 - 1}{0.2199790} = 0.15217,$$

which yields a p-value for the test of:

```
> 2*(1-pt(0.152171,15-7))

[1] 0.8828198
```

Or

```
> 2*(1-pnorm(0.152171))

[1] 0.8790521
```

So, one seems a reasonable value for the coefficient of `log(mnths)`.

2. An experiment to determine the effects of temperature and storage time on the loss of ascorbicacid (vitamin C) in snap-beans was performed and the observed concentrations are shown below:

Temp ($^\circ F$)	Weeks of Storage			
	2	4	6	8
0	45	47	46	46
10	45	43	41	37
20	34	28	21	16

(a) Suppose that ascorbic acid concentration decays exponentially, and that the expected concentration after t weeks for the beans at temperature T is $\mu_T = \exp(\alpha + \beta_T t)$, where the initial concentration of acid, e^α, is assumed to be the same for each temperature group and the decay rate, β_T, is dependent on the temperature group. Fit a gamma generalised linear model with logarithmic link. Examine the dispersion parameter estimate. Does the data look consistent with the idea of an exponential distribution? [HINT: This model is a bit unusual in that it contains an interaction between time and temperature, but no main effect of temperature, and we can re-write the model as:

$$\log \mu = \alpha + \beta_0 t + \beta_{10} t z_1 + \beta_{20} t z_2,$$

where z_1 and z_2 are indicators for the temperature categories 10 $^\circ F$ and 20 $^\circ F$, repectively.]

Solution: The required *R* commands are:

```
> conc <- c(45, 47, 46, 46, 45, 43, 41, 37, 34, 28, 21, 16)
> temp <- c("0", "0", "0", "0", "10", "10", "10", "10", "20", "20", "20", "20")
> wks <- c(2, 4, 6, 8, 2, 4, 6, 8, 2, 4, 6, 8)
> tz1 <- ifelse(temp=="10", wks, 0)
> tz2 <- ifelse(temp=="20", wks, 0)
> bean.glm <- glm(conc~wks+tz1+tz2, family=Gamma(link=log))
> anova(bean.glm, test="Chisq")

Analysis of Deviance Table

Model: Gamma, link: log

Response: conc

Terms added sequentially (first to last)
```

```
     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    11     1.1665
wks   1  0.08688        10     1.0797 < 2.2e-16 ***
tz1   1  0.06691         9     1.0128 < 2.2e-16 ***
tz2   1  1.00665         8     0.0061 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(bean.glm)$dispersion

[1] 0.0007555935
```

$= \phi = \frac{1}{\alpha}$

Recall that an exponential distribution is just a special case of the Gamma distribution where $\alpha = 1$. The dispersion for the exponential distribution is $1/\alpha = 1$. Clearly, the dispersion parameter is not very close to 1, so the data do not appear to be exponential (though perhaps they are gamma distributed, however, we would need to do some diagnostic tests to assess this fact).

(b) Estimate the time taken for the concentration to be halved at each temperature. [HINT: Recall that the initial concentration is assumed to be e^{α}, and we want to find the time when the predicted concentration is $0.5e^{\alpha}$.]

Solution: Using the mean relationship $\mu = \exp(\alpha + \beta_T t)$, we see that we want to solve the equation:

$$0.5 \exp \alpha = \exp(\alpha + \beta_T t),$$

which means we want to estimate the value $t = \log(0.5)/\beta_T$. Now, for zero degrees, $\beta_T = \beta_0$, while for 10 degrees and 20 degrees, $\beta_T = \beta_0 + \beta_{10}$ and $\beta_T = \beta_0 + \beta_{20}$, respectively. So, the desired values are:

```
> btas <- coef(bean.glm)
> btas

  (Intercept)           wks           tz1           tz2
 3.8356719282 -0.0008096128 -0.0231865792 -0.1315290446

> betaT <- as.vector(c(btas[2], btas[2]+btas[3], btas[2]+btas[4]))
> hlflfe <- log(0.5)*betaT^-1
> hlflfe

[1] 856.146528  28.885716   5.237677
```

So, at zero degrees, the half-life is over 16 years, while at 10 degrees, the half-life is only about 7 months, and at 20 degrees the half-life is just over 5 weeks.

(c) Suppose that we do not assume that the initial concentrations were the same for each temperature. Create additional indicators and fit this model. Do you think

the assumption of equal initial concentrations is reasonable?

Solution: The necessary *R* commands are:

```
> z1 <- ifelse(temp=="10", 1, 0)
> z2 <- ifelse(temp=="20", 1, 0)
> bean.glm1 <- glm(conc~z1+z2+wks+tz1+tz2, family=Gamma(link=log))
> coef(bean.glm1)

 (Intercept)           z1           z2          wks          tz1          tz2
 3.817664284  0.064144486 -0.010907090  0.002193039 -0.033913189 -0.129707706
```

It appears that the assumption of a common initial concentration is reasonable, though we need to have some idea of standard errors for our parameter estimates before we can say for sure. In fact, we could refit the model with the predictors in an appropriate order and use the analysis of deviance table to determine whether the new indicators `z1` and `z2` were significant in the model:

```
> bean.glm2 <- glm(conc~wks+tz1+tz2+cbind(z1,z2), family=Gamma(link=log))
> anova(bean.glm2, test="Chisq")

Analysis of Deviance Table

Model: Gamma, link: log

Response: conc

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                           11    1.16655
wks           1  0.08688        10    1.07966   <2e-16 ***
tz1           1  0.06691         9    1.01275   <2e-16 ***
tz2           1  1.00665         8    0.00610   <2e-16 ***
cbind(z1, z2) 2  0.00218         6    0.00393   0.1918
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(bean.glm2)$dispersion

[1] 0.0006592066

> dstar <- anova(bean.glm2)$Deviance[5]/summary(bean.glm2)$dispersion
> dstar

[1] 3.30234
```

```
> 1-pchisq(dstar,2)

[1] 0.1918253
```

So, we can accept the null hypothesis that these new indicators are not necessary in the model. In other words, the assumption that the initial concentrations of ascorbic acid in the three temperature groups was the same is indeed plausible.