APPLIED STATISTICS, MATH10096

Workshop 5

1. Suppose that y_i is the number of bacterial cells in a sample from a petri dish and x_i is the amount of nutrient in the dish. If it is expected that:

$$\mathbb{E}(y_i) = a^2 + 2bax_i + b^2x_i^2,$$

write down a suitable GLM for the (y_i, x_i) data.

2. Consider the GLM

$$Y_i \overset{\text{ind}}{\sim} \text{Poisson}(\mu_i), \quad \log(\mu_i) = \mathbf{x}_i\boldsymbol{\beta}, \quad i = 1, \ldots, n.$$

Let $Z_i = I(Y_i > 0)$ and define $\pi_i = \Pr(Z_i = 1)$. Then the Z_is follow a binomial GLM

$$Z_i \overset{\text{ind}}{\sim} \text{Bin}(1, \pi_i), \quad i = 1, \ldots, n.$$

Show that the link function for this binomial model is given by

$$\log(-\log(1 - \pi_i)) = \mathbf{x}_i\boldsymbol{\beta}.$$

3. Assume that Y follows a negative binomial distribution with probability mass function given by

$$f(y; \mu, k) = \frac{\Gamma(y + k)}{\Gamma(k)\Gamma(y + 1)} \left(\frac{\mu}{\mu + k}\right)^y \left(\frac{k}{\mu + k}\right)^k, \quad y = 0, 1, 2, \ldots$$

Further assume that $k > 0$ is a given constant and consider the random variable $Y^* = Y/k$. Then $\Pr(Y^* = y^*) = \Pr(Y = ky^*)$ such that Y^* has probability mass function given by

$$f(y^*; \mu, k) = \frac{\Gamma(ky^* + k)}{\Gamma(k)\Gamma(ky^* + 1)} \left(\frac{\mu}{\mu + k}\right)^{ky^*} \left(\frac{k}{\mu + k}\right)^k, \quad y^* = 0, \frac{1}{k}, \frac{2}{k}, \ldots \tag{1}$$

   (a) Show that (1) is a distribution in the exponential family with

$$\theta = \log\left(\frac{\mu}{\mu + k}\right), \quad b(\theta) = -\log(1 - e^\theta), \quad a(\phi) = \frac{1}{k}.$$

   (b) Find the mean and variance of Y^* using the expressions for $b(\theta)$ and $a(\phi)$. Use these results to show that $\mathbb{E}(Y) = \mu$ and $\text{var}(Y) = \mu + \frac{\mu^2}{k}$.

   (c) Compare the relationship between the mean and variance for the negative binomial distribution to the relationship between the mean and variance of the Poisson distribution, and comment on when the Poisson is a good model and when you need the negative-binomial.

4. Show that the t distribution, which is suitable for modelling heavy tail data and whose probability density function is given by

$$f(y; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad y \in \mathbb{R}, \quad \nu > 0,$$

is not in the exponential family.

5. A blood pressure drug trial looked at change in blood pressure, y_i, over 6 weeks of therapy for 40-45 year old hypertensive patients. Patients were given one of 3 drug formulations or a placebo, each at one of three dose rates. The data for 120 patients were modelled as

$$\mathbb{E}(y_i) = \mu + \alpha_j + \beta_k, \text{ if patient } i \text{ is formulation } j, \text{ dose group } k,$$

with the y_i assumed to be independent gamma random variables. Model checking plots suggested that the full model assumptions were met, and the model was re-fitted without the α_j terms. The deviance for the full model fit was 193, while the deviance for the reduced version was 201. What hypothesis can you test using this information, what is the associated p-value, and what can you conclude?

6. This exercise involves data from a dose-response study. Specifically, it pertains to data on the number of adult flour beetles that died after 5 hours of exposure to gaseous carbon disulfide at various dosages. The data can be entered in `R` as follows

```
logdose <- c(1.691, 1.724, 1.755, 1.784, 1.811, 1.837, 1.861, 1.884)
dead <- c(6, 13, 18, 28, 52, 53, 61, 60)
n <- c(59, 60, 62, 56, 63, 59, 62, 60)
```

The goal of the exercise is to fit a binomial regression model using three link functions and check which one provides a better fit to this data. Let p_i be the probability of death at (log) dose level x_i. The three link functions to be considered are:

- *Logit link function*:
$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_i = \beta_0 + \beta_1 x_i.$$
- *Probit link function*:
$$\Phi^{-1}(p_i) = \eta_i = \beta_0 + \beta_1 x_i,$$
where Φ stands for the standard normal cumulative distribution function. In `R`, please specify `family = binomial(link = "probit")`.
- *Complementary log-log link function*:
$$\log(-\log(1-p_i)) = \eta_i = \beta_0 + \beta_1 x_i.$$
In `R`, please specify `family = binomial(link = "cloglog")`.

7. The dataset `africa` from `R` package `faraway` gives information about the number of military coups in sub-Saharan Africa and various political and geographical information. Namely, the following variables are available:
   - `miltcoup`: number of successful military coups from independence to 1989.
   - `oligarchy`: number years country ruled by military oligarchy from independence to 1989.
   - `pollib`: Political liberalization - 0 = no civil rights for political expression, 1 = limited civil rights for expression but right to form political parties, 2 = full civil rights.
   - `parties`: Number of legal political parties in 1993.

- `pctvote`: Percent voting in last election.
- `popn`: Population in millions in 1989.
- `size`: Area in 1000 square km.
- `numelec`: Total number of legislative and presidential elections.
- `numregim`: Number of regime types.

The below `R` session attempts to build a simple model of the number of military coups, based on the other variables.

(a) Write short notes to accompany the `R` session, explaining what is being done, statistically, and what model you would finally choose for these data.

(b) Interpret the coefficients of the finally selected model. Do they make sense?

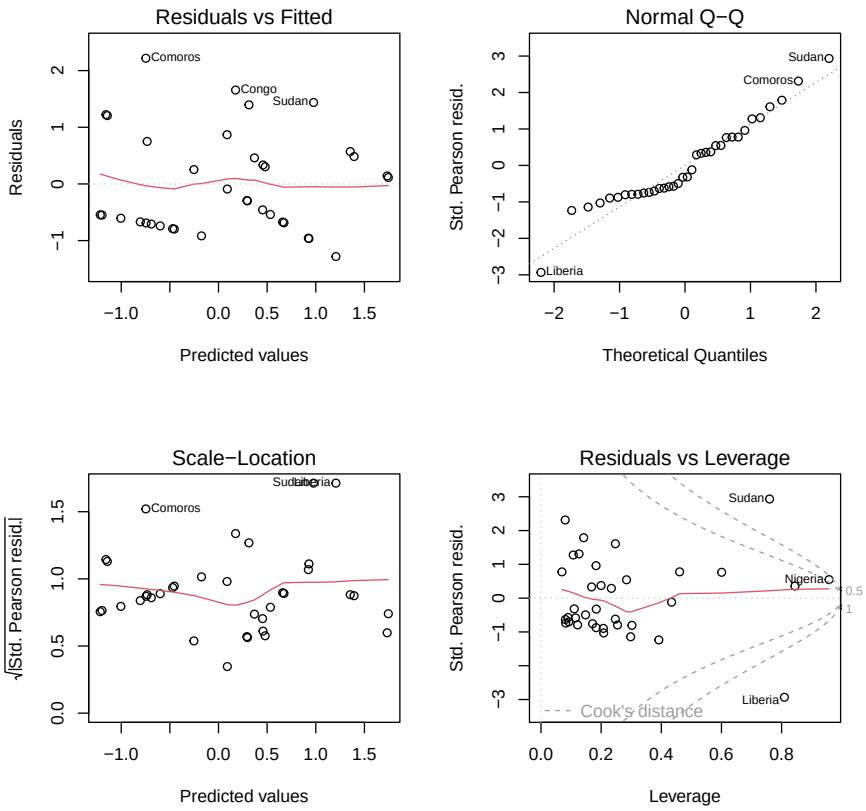
(c) What is the 'proportion deviance explained' for the selected model?

(d) Is the model a useful predictive model for number of coups?

(e) From the models fitted which model would you have selected if using AIC for model selection?

```
library(faraway)
data(africa)
ac1 <- glm(miltcoup ~ oligarchy + as.factor(pollib) +parties +
             pctvote + popn + size + numelec + numregim,
           family = poisson, data = africa)

par(mfrow = c(2,2))
plot(ac1)
```




```
drop1(ac1, test = "Chisq")

## Single term deletions
##
## Model:
## miltcoup ~ oligarchy + as.factor(pollib) + parties + pctvote +
##     popn + size + numelec + numregim
##                   Df Deviance    AIC    LRT Pr(>Chi)
## <none>                 28.249 113.06
## oligarchy          1   32.354 115.16 4.1045 0.042769 *
## as.factor(pollib)  2   35.581 116.39 7.3314 0.025587 *
## parties            1   35.311 118.12 7.0616 0.007875 **
## pctvote            1   30.572 113.38 2.3222 0.127537
## popn               1   30.601 113.41 2.3513 0.125178
## size               1   29.238 112.05 0.9881 0.320197
## numelec            1   28.430 111.24 0.1810 0.670499
## numregim           1   29.059 111.87 0.8099 0.368147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ac2 <- update(ac1, .~. - numelec)
drop1(ac2, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## miltcoup ~ oligarchy + as.factor(pollib) + parties + pctvote +
##     popn + size + numregim
##                   Df Deviance    AIC    LRT Pr(>Chi)
## <none>                 28.430 111.24
## oligarchy          1   36.595 117.40 8.1649 0.004271 **
## as.factor(pollib)  2   36.872 115.68 8.4417 0.014686 *
## parties            1   35.773 116.58 7.3428 0.006733 **
## pctvote            1   30.590 111.40 2.1600 0.141648
## popn               1   30.605 111.41 2.1746 0.140305
## size               1   29.452 110.26 1.0214 0.312196
## numregim           1   29.081 109.89 0.6508 0.419818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ac3 <- update(ac2, .~. - numregim)
drop1(ac3, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## miltcoup ~ oligarchy + as.factor(pollib) + parties + pctvote +
##     popn + size
##                   Df Deviance    AIC     LRT  Pr(>Chi)
## <none>                 29.081 109.89
## oligarchy          1   40.291 119.10 11.2094 0.0008139 ***
## as.factor(pollib)  2   37.830 114.64  8.7487 0.0125960 *
## parties            1   36.304 115.11  7.2224 0.0071998 **
## pctvote            1   31.599 110.41  2.5175 0.1125903
## popn               1   30.614 109.42  1.5324 0.2157552
## size               1   30.040 108.85  0.9582 0.3276300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ac4 <- update(ac3, .~. - size)
drop1(ac4, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## miltcoup ~ oligarchy + as.factor(pollib) + parties + pctvote +
##     popn
##                   Df Deviance    AIC     LRT Pr(>Chi)
## <none>                 30.040 108.85
## oligarchy          1   40.468 117.28 10.4283 0.001241 **
## as.factor(pollib)  2   38.022 112.83  7.9826 0.018476 *
```

```
## parties            1   37.547 114.36  7.5070 0.006146 **
## pctvote            1   32.241 109.05  2.2010 0.137920
## popn               1   31.069 107.88  1.0299 0.310189
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ac5 <- update(ac4, .~. - popn)
drop1(ac5,test="Chisq")
```

```
## Single term deletions
##
## Model:
## miltcoup ~ oligarchy + as.factor(pollib) + parties + pctvote
##                   Df Deviance    AIC     LRT  Pr(>Chi)
## <none>                 31.069 107.88
## oligarchy          1   48.196 123.00 17.1266 3.497e-05 ***
## as.factor(pollib)  2   39.762 112.57  8.6926   0.01295 *
## parties            1   37.547 112.36  6.4773   0.01093 *
## pctvote            1   32.822 107.63  1.7521   0.18561
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ac6 <- update(ac5,.~. - pctvote)
drop1(ac6,test="Chisq")
```

```
## Single term deletions
##
## Model:
## miltcoup ~ oligarchy + as.factor(pollib) + parties
##                   Df Deviance    AIC     LRT  Pr(>Chi)
## <none>                 42.235 125.92
## oligarchy          1   64.574 146.25 22.3389 2.285e-06 ***
## as.factor(pollib)  2   48.597 128.28  6.3615   0.04155 *
## parties            1   45.431 127.11  3.1964   0.07380 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ac7 <- update(ac6, .~. - parties)
drop1(ac7, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## miltcoup ~ oligarchy + as.factor(pollib)
##                   Df Deviance    AIC     LRT  Pr(>Chi)
## <none>                 45.431 127.11
## oligarchy          1   70.453 150.13 25.0219 5.668e-07 ***
```

```
## as.factor(pollib)  2   50.101 127.78  4.6694   0.09684 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ac7)
```

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + as.factor(pollib), family = poisson,
##     data = africa)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3725  -1.1365  -0.3315   0.4076   1.9161
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.4795     0.4274   1.122   0.2618
## oligarchy            0.1023     0.0207   4.942 7.73e-07 ***
## as.factor(pollib)1  -0.5393     0.4695  -1.149   0.2506
## as.factor(pollib)2  -0.9168     0.4426  -2.071   0.0383 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 79.124  on 41  degrees of freedom
## Residual deviance: 45.431  on 38  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 127.11
##
## Number of Fisher Scoring iterations: 5
```