Workshop 5: solutions

1. Count data suggests using the Poisson distribution, i.e., $y_i \overset{\text{ind}}{\sim} \text{Poisson}(\mu_i)$, for $i = 1, \ldots, n$. The model is

$$\mathbb{E}(y_i) = \mu_i = (a + bx_i)^2,$$

and can be made linear in its parameters by employing a square root link function:

$$g(\mu_i) = \sqrt{\mu_i} = a + bx_i.$$

The linear predictor is hence of the form

$$\boldsymbol{\eta} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}.$$

2. Note that $Y_i \in \{0, 1, 2, 3, \ldots\}$ and as such

$$Z_i = \begin{cases} 1, \text{ if } Y_i > 0, \\ 0, \text{ if } Y_i = 0, \end{cases}$$

and hence it holds that

$$\begin{aligned} \pi_i &= \Pr(Z_i = 1) \\ &= 1 - \Pr(Z_i = 0) \\ &= 1 - \Pr(Y_i = 0). \end{aligned}$$

Because $Y_i \overset{\text{ind}}{\sim} \text{Poisson}(\mu_i)$ we know that

$$\Pr(Y_i = 0) = \frac{e^{-\mu_i}\mu_i^0}{0!} = e^{-\mu_i},$$

and hence

$$\pi_i = 1 - e^{-\mu_i}.$$

Solving for μ_i we get

$$1 - \pi_i = e^{-\mu_i} \Rightarrow \mu_i = -\log(1 - \pi_i).$$

Because $\log(\mu_i) = \mathbf{x}_i\boldsymbol{\beta}$, we have that

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

Note that $\mathbb{E}(Z_i) = 1 \times \Pr(Z_i = 1) + 0 \times \Pr(Z_i = 0) = \Pr(Z_i = 1) = \pi_i$. Out of curiosity, this link function is known as the complementary log-log link (see more in Exercise 6).

3. (a) As we have done before, we start by writing $f(y^*;\mu,k) = \exp\{\log f(y^*;\mu,k)\}$:

$$\begin{aligned} f(y^*;\mu,k) &= \exp\{\log f(y^*;\mu,k)\} \\ &= \exp\left\{\log\Gamma(ky^*+k) - \log\Gamma(k) - \log\Gamma(ky^*+1) + ky^*\log\left(\frac{\mu}{\mu+k}\right) + k\log\left(\frac{k}{\mu+k}\right)\right\} \\ &= \exp\left\{\frac{y^*\log\left(\frac{\mu}{\mu+k}\right) - \left[-\log\left(\frac{k}{\mu+k}\right)\right]}{1/k} + \left[\log\Gamma(ky^*+k) - \log\Gamma(k) - \log\Gamma(ky^*+1)\right]\right\} \end{aligned}$$

This is in the exponential family of distributions with:

- $\theta = \log\left(\frac{\mu}{\mu+k}\right) \Rightarrow \mu = \frac{e^\theta}{1-e^\theta}k$.
- $b(\theta) = -\log\left(\frac{k}{\mu+k}\right)$. Replacing μ by $\frac{e^\theta}{1-e^\theta}k$, we get that

$$\begin{aligned} \mu + k &= \frac{e^\theta}{1-e^\theta}k + k = k\left(\frac{e^\theta}{1-e^\theta} + 1\right) = k\left(\frac{1}{1-e^\theta}\right), \\ \frac{k}{\mu+k} &= \frac{k}{k\left(\frac{1}{1-e^\theta}\right)} = 1 - e^\theta, \end{aligned}$$

  and therefore

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

- $a(\phi) = \frac{1}{k}$.

(b) By the properties of the exponential family we have that

$$\mathbb{E}(Y^*) = b'(\theta) = \frac{e^\theta}{1-e^\theta} = \frac{\mu}{k}.$$

We further have that

$$\mathbb{E}(Y) = \mathbb{E}(kY^*) = k\mathbb{E}(Y^*) = k\frac{\mu}{k} = \mu.$$

We also know that $\text{var}(Y^*) = b''(\theta)a(\phi)$. Let's workout $b''(\theta)$:

$$b''(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{e^\theta}{1-e^\theta}\right) = \frac{e^\theta}{(1-e^\theta)^2} = \frac{e^\theta}{1-e^\theta} \times \frac{1}{1-e^\theta} = \frac{\mu}{k} \times \left(1 + \frac{\mu}{k}\right),$$

where we note that

$$1 + \frac{\mu}{k} = 1 + \frac{e^\theta}{1-e^\theta} = \frac{1}{1-e^\theta}.$$

Thus,

$$\text{var}(Y^*) = b''(\theta)a(\phi) = \frac{\mu}{k} \times \left(1 + \frac{\mu}{k}\right) \times \frac{1}{k} = \frac{\mu}{k^2} + \frac{\mu^2}{k^3}.$$

We finally have that

$$\text{var}(Y) = \text{var}(kY^*) = k^2\text{var}(Y^*) = k^2\left(\frac{\mu}{k^2} + \frac{\mu^2}{k^3}\right) = \mu + \frac{\mu^2}{k}.$$

(c) For the Poisson distribution we have that both the mean and variance are the same and therefore in data applications where the variance is (much) larger than the mean, the Poisson distribution is not appropriate but the negative binomial distribution may be, since as we have concluded in part (b), the variance is greater than the mean for this distribution. Note that if k is large relative to μ^2, the term $\frac{\mu^2}{k}$ is close to zero and the negative binomial converges to the Poisson distribution.

4. We have that

$$\exp\{\log f(y;\nu)\} = \exp\left\{\log\Gamma\left(\frac{\nu+1}{2}\right) - \frac{1}{2}\log(\nu\pi) - \log\Gamma\left(\frac{\nu}{2}\right) - \frac{\nu+1}{2}\log\left(1+\frac{y^2}{\nu}\right)\right\}$$

As we can observe this expression does not simplify into a form where there is a linear term in y that would correspond to $y\theta$ in the exponential family of distributions form. Thus, the t-distribution does not belong to the exponential family.

5. From Equation (29) in the lecture notes, we have that under the null hypothesis of no formulation effect then we would expect that

$$F = \frac{(D_0 - D_1)/3}{D_1/114} \sim F_{3,114},$$

where $D_0 = 201$ is the null model deviance and $D_1 = 193$ the alternative model deviance (the F ratio result is needed because the scale parameter is unknown). Note that the full model has 6 parameters: μ, α_1, α_2, and α_3 (let us assume, without loss of generality, that α_4 representing the placebo effect is set to zero for identifiability reasons), β_1, and β_2 (where, again, one of the three dose rates levels was set to zero due to identifiability reasons). And because we have $n = 120$ patients, $n - p_1 = 120 - 6 = 114$. Also, $p_1 - p_0$ is just the difference in the number of parameters of the two models and it is 3 due to the fact that we have three levels that can be identified for the drug formulation. So the p-value can be evaluated as follows

```
F <- ((201-193)/3)/(193/114)
pf(F, 3, 114, lower.tail = FALSE)

## [1] 0.1992741
```

so there is really no evidence against the null: formulation does not seem to matter.

6. We start by fitting the required binomial model with the three different link functions.

```
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)
alive <- n - dead

fit_logit <- glm(cbind(dead, alive) ~ logdose,
                 family = binomial(link = "logit"))
#alternative way
#fit_logit_alt <- glm(dead/n ~ logdose,
```

Observed	Logit	Probit	Comp.log.log
6	3.5	3.4	5.7
13	9.8	10.7	11.3
18	22.4	23.4	20.9
28	33.9	33.8	30.3
52	50.0	49.6	47.7
53	53.3	53.4	54.2
61	59.2	59.7	61.1
60	58.8	59.2	59.9

```
#                       family = binomial(link = "logit"), weights = n)
fit_probit <- glm(cbind(dead, alive) ~ logdose,
                  family = binomial(link = "probit"))
fit_cloglog <- glm(cbind(dead, alive) ~ logdose,
                   family = binomial(link = "cloglog"))
```

We can start by comparing the expected number of deaths, under the three models, against the observed number of deaths.

```
df <- data.frame("Observed" = dead,
                 "Logit" = n*fit_logit$fitted.values,
                 "Probit" = n*fit_probit$fitted.values,
                 "Comp log-log" = n*fit_cloglog$fitted.values)
require(kableExtra)
knitr::kable(df, escape = FALSE, digits = 1) %>%
  kable_styling(position = "center")
```

The fitted values obtained from the binomial model with the complementary log-log link function appear to more closely match the observed values, compared to those from the binomial model using either a logit or a probit link function (and note that the three models use the same number of parameters: 2, β_0 and β_1). The models with different link functions are not nested, so we cannot compare them with likelihood-ratio tests. However, we can compare them on the basis of the AIC criterion.

```
AIC(fit_logit, fit_probit, fit_cloglog)

##              df      AIC
## fit_logit     2 41.31361
## fit_probit    2 40.18499
## fit_cloglog   2 33.71237
```

The AIC is the lowest for the binomial model with complementary log log link function. Let us inspect further this model.

```
summary(fit_cloglog)

##
```

```
## Call:
## glm(formula = cbind(dead, alive) ~ logdose, family = binomial(link = "cloglog"))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -39.522      3.236  -12.21   <2e-16 ***
## logdose       22.015      1.797   12.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284.2024  on 7  degrees of freedom
## Residual deviance:   3.5143  on 6  degrees of freedom
## AIC: 33.712
##
## Number of Fisher Scoring iterations: 4
```

We have that the deviance is 3.51 on 6 degrees of freedom and therefore

```
pchisq(3.51, df = 6, lower.tail = FALSE)
```

```
## [1] 0.7426388
```

shows that there is a quite high probability of a χ^2_6 random variable being as large as 3.51.

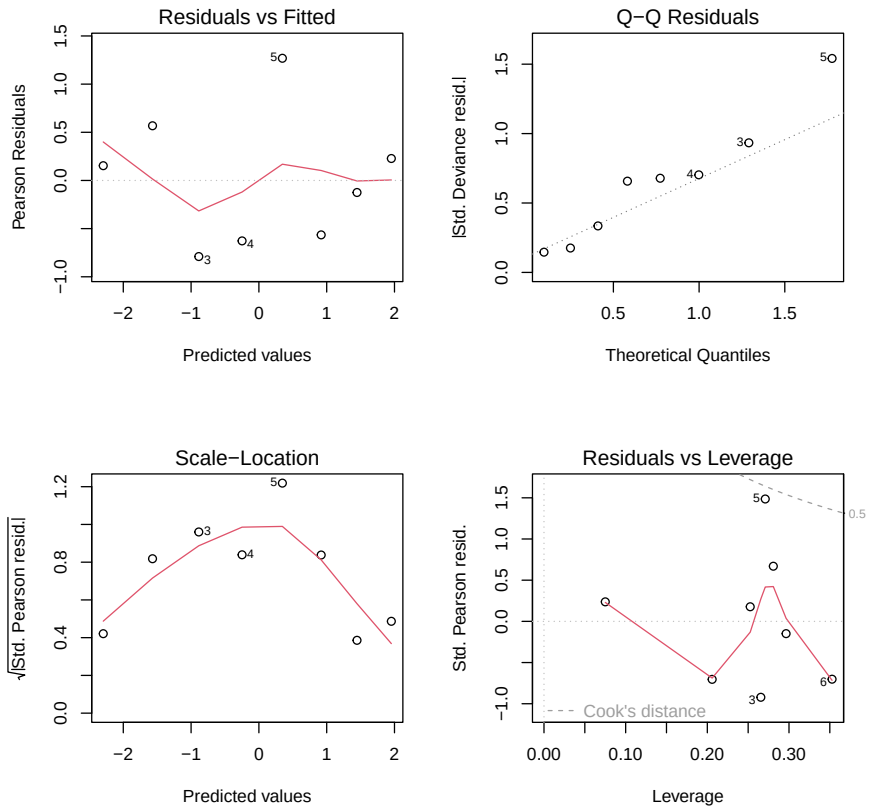
We have that

$$\log(-\log(1-\hat{p}_i)) = -39.522 + 22.015 x_i \Rightarrow \hat{p}_i = 1 - e^{-e^{39.522+22.015x_i}},$$

which allows us to compute the predicted probability of death at any dose of interest.

Finally, let us look at the residual plots (although we have very few observations).

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



If we had more data then such a departure from constant variance would be a cause for concern.

7. (a)
   - Writing m_i, o_i, p_i, v_i, n_i, a_i, e_i and r_i for variables in the order presented in the question (excluding `pollib`), then the initial model fitted is

     $$\log\{\mathbb{E}(m_i)\} = \alpha+\beta_1 o_i+\beta_2 p_i+\beta_3 v_i+\beta_4 n_i+\beta_5 a_i+\beta_6 e_i+\beta_7 r_i+\gamma_j, \quad \text{if country } i \text{ is liberalization type } j,$$

     where $m_i \sim$ Poi.
   - The residuals are ok, although Sudan and Liberia are very influential.
   - Backward selection is applied, using generalized likelihood ratio tests to compare models with and without individual terms. When comparing nested models 1 and 0 with deviances D_1 and D_0 and model degrees of freedom (number of identifiable parameters) p_1 and p_0, respectively, then

     $$D_0 - D_1 \sim \chi^2_{p_1-p_0}$$

     if model 0 is correct, and this is the basis for p-value calculations.
   - Terms drop out sequentially until we are left with the model

     $$\log\{\mathbb{E}(m_i)\} = \alpha + \beta_1 o_i + \gamma_j \quad \text{if country } i \text{ is liberalization type } j$$

     Actually, the p-value for testing $H_0 : \gamma_j = 0 \; \forall \; j$ is above the p-value for other terms that have previously been dropped, so there is an argument for selecting the model depending only on o_i, but perhaps the analyst is really interested in the γ_j's, and therefore leaves them in.

(b) The coefficients suggest that political liberalization reduces the number of coups, but the evidence that these effects are real is only marginal. There is a significant positive relationship between years of military rule and number of coups — this is a rather unsurprising effect: more coups is almost always going to lead to more years of military rule.

(c) The proportion of deviance explained for the selected model is

```
1 - (45.431/79.124)

## [1] 0.4258253
```

(d) Only in a rather limited sense. The main predictor is number of years under military rule: almost without exception, a country only gets to be under military rule because of a coup. So we're predicting number of coups using a variable that implicitly contains the information that at least one coup happened. There's no meaningful sense in which you could use the model to predict whether a country not in the list would be likely to suffer a coup.

(e) Using AIC model selection the model

$$\log\{\mathbb{E}(m_i)\} = \alpha + \beta_1 o_i + \beta_2 p_i + \gamma_j, \text{ if country } i \text{ is liberalization type } j,$$

would be selected. Notice the odd fact that the AIC for *the same model* sometimes changes between one call to `drop1` and the next (even though the scale parameter is known to be 1 here). This happens because there are missing values in the predictor variable data (see final model summary for evidence of this). Any country with missing values is simply omitted when model fitting, but as the number of predictors is reduced fewer and fewer countries need to be omitted. When comparing models by AIC (or hypothesis testing) it is important that they are fitted to the same response data, so `drop1` always fits all model alternatives to the response data useable by the largest model. However different calls to `drop1` will use different numbers of countries as the largest model under consideration changes. This explains the apparent discrepancy.