MSc: Statistical Analysis (Spring Term)

Generalized Linear Models: Exercises 2 Solutions

1. The suggested model is fitted in R as follows:

```
> leuk <- c(13, 5, 5, 3, 4, 18)
> other <- c(378, 200, 151, 47, 31, 33)
> tots <- leuk + other
> prop <- leuk/tots

> leuk.df <- data.frame(leuk, other, tots, prop)
> leuk.df
  leuk other tots    prop
  leuk other tots       prop
1   13   378  391 0.03324808
2    5   200  205 0.02439024
3    5   151  156 0.03205128
4    3    47   50 0.06000000
5    4    31   35 0.11428571
6   18    33   51 0.35294118

> dose <- c(0, 1, 10, 50, 100, 200)

> leuk.glm <- glm(prop ~ dose, family = binomial, weights = tots)
> summary(leuk.glm)

Call:
glm(formula = prop ~ dose, family = binomial, weights = tots)

Deviance Residuals:
       1         2         3         4         5         6
 0.41428  -0.48994  -0.13991   0.02835   0.00048   0.00269

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.488973   0.204062 -17.098  < 2e-16 ***
dose         0.014410   0.001817   7.932 2.15e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.35089  on 5  degrees of freedom
Residual deviance:  0.43206  on 4  degrees of freedom
AIC: 26.097

Number of Fisher Scoring iterations: 4
```

The fitted model appears adequate. The residual deviance, 0.4321, is clearly not significant when compared with a $\chi^2(4)$ distribution,

```
> dev <- deviance(leuk.glm)
> pchisq(dev, 4, lower.tail=F)
[1] 0.9797692
```

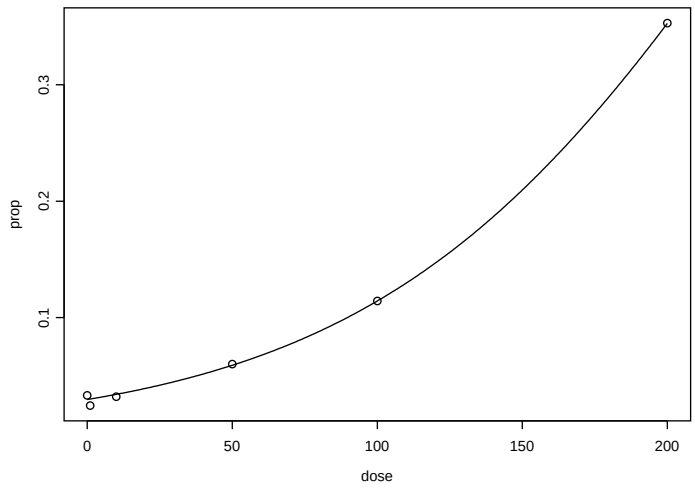
so that there is no need for any additional terms to be included in the model. Looking at the fitted values (for the proportion of cancer deaths that were due to leukemia),

```
> fits <- fitted(leuk.glm)
> diff <- prop - fits
> data.frame(prop, fits, diff)
        prop       fits          diff
1 0.03324808 0.02962762  3.620467e-03
2 0.02439024 0.03004473 -5.654483e-03
3 0.03205128 0.03406353 -2.012252e-03
4 0.06000000 0.05905247  9.475338e-04
5 0.11428571 0.11425978  2.593924e-05
6 0.35294118 0.35276092  1.802608e-04
```

we see that they are very close to the observed proportions. The model suggests that the dose response relationship is

$$\log\left(\frac{\pi}{1-\pi}\right) = -3.4890 + 0.0144\,\text{dose} \quad \text{or, equivalently} \quad \pi = \frac{\exp(-3.4890 + 0.0144)}{1 + \exp(-3.4890 + 0.0144)}$$

The fitted response curve can be shown on a plot of the observed proportions.



The commands used to produce this plot are

```
> plot(dose, prop)
> xx <- seq(0, 200, 0.5)
> lines(xx, predict(leuk.glm, data.frame(dose = xx), type = "response"))
```

2. Again,the suggested model is fitted in R as follows:

(a)
```
> dead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
> number <- rep(20, 12)
> sex <- c(rep("M", 6), rep("F", 6))
> prop <- dead/number
> budworm.df <- data.frame(dead, number, prop, sex)
> budworm.df
   dead number prop sex
1     1     20 0.05   M
2     4     20 0.20   M
3     9     20 0.45   M
4    13     20 0.65   M
5    18     20 0.90   M
6    20     20 1.00   M
7     0     20 0.00   F
8     2     20 0.10   F
9     6     20 0.30   F
10   10     20 0.50   F
11   12     20 0.60   F
12   16     20 0.80   F

> sex <- factor(sex)

> dose <- rep(c(1,2,4,8,16,32),2)
> ldose <- logb(dose,base=2)
> # Alternatively
> # ldose <- rep(0:5,2)

> budworm.glm <- glm(prop ~ sex * ldose, family = binomial, weights = number)
> summary(budworm.glm)

Call:
glm(formula = prop ~ sex * ldose, family = binomial, weights = number)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.39849  -0.32094  -0.07592   0.38220   1.10375

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9935     0.5527  -5.416 6.09e-08 ***
sexM          0.1750     0.7783   0.225    0.822
ldose         0.9060     0.1671   5.422 5.89e-08 ***
sexM:ldose    0.3529     0.2700   1.307    0.191
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   4.9937  on  8  degrees of freedom
AIC: 43.104

Number of Fisher Scoring iterations: 4
```

Recall, the command `factor(sex)` tells R that the 'character' vector `sex` is to be treated as a factor.

```
> anova(budworm.glm, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: prop

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                         11    124.876
sex        1    6.077        10    118.799   0.0137 *
ldose      1  112.042         9      6.757   <2e-16 ***
sex:ldose  1    1.763         8      4.994   0.1842
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The analysis of deviance table above, shows that the interaction term `sex:ldose` is not significant (change in deviance 1.8, compared with a $\chi^2(1)$ distribution), when added to a model which contains both `sex` and `ldose`.

Refitting the model without the interaction term can be done using the `update` command.

```
> budworm2.glm <- update(budworm.glm,  ~ . - sex:ldose)
> summary(budworm2.glm)
```


```
> budworm2.glm <- update(budworm.glm,  ~ . - sex:ldose)
> summary(budworm2.glm)
```


```
Call:
glm(formula = prop ~ sex + ldose, family = binomial, weights = number)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.10540  -0.65343  -0.02225   0.48471   1.42944

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.4732     0.4685  -7.413 1.23e-13 ***
sexM          1.1007     0.3558   3.093  0.00198 **
```

```
ldose         1.0642     0.1311   8.119 4.70e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   6.7571  on  9  degrees of freedom
AIC: 42.867

Number of Fisher Scoring iterations: 4
```

Note, the model with the main effects `sex` and `ldose` alone provides an adequate fit to the data.

```
> dev <- deviance(budworm2.glm)
> pchisq(dev, 9, lower.tail=F)
[1] 0.6623957
```

Also, we should check that the `sex` effect is needed

```
> anova(budworm2.glm, update(budworm2.glm,  ~ . - sex), test = "Chisq")
Analysis of Deviance Table

Model 1: prop ~ sex + ldose
Model 2: prop ~ ldose
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         9     6.7571
2        10    16.9840 -1  -10.227 0.001384 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The analysis of deviance table shows a significant `sex` effect. Hence `budworm2.glm` containing both `sex` and `ldose` main effects is our chosen model.

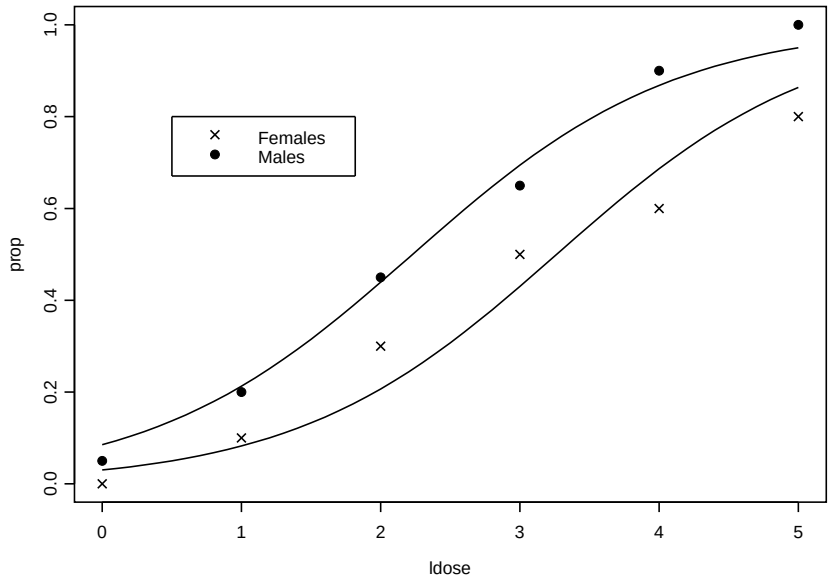
(b) Our analysis suggests a model with parallel lines (on the *logit* scale) for each sex. That is, a model with separate intercepts for each sex and a common slope. (Had the interaction term been present, we would have had different slopes also. C/f multiple linear regression). Since female, "F", is the first level of `sex` (they are in alphabetical order), the parameter `sex` represents the increase in intercept for males, "M". We can check this by reparametrizing the model to give separate intercepts.

```
> budworm0.glm <- update(budworm2.glm,  ~ . - 1)
> coef(budworm0.glm)
     sexF      sexM     ldose
-3.473155 -2.372412  1.064214
```

The fitted response curves are shown in the plot below.

For the females: $\log\left(\frac{p}{1-p}\right) = -3.473 + 1.064\text{ldose}$

For the males: $\log\left(\frac{p}{1-p}\right) = -2.3724 + 1.064\text{ldose}$



The commands used to produce this plot are

```
> plot(ldose, prop, type = "n")
> points(ldose[sex == "F"], prop[sex == "F"], pch = 4)
> points(ldose[sex == "M"], prop[sex == "M"], pch = 16)
> legend(0.5, 0.8, c("Females", "Males"), pch = c(4, 16))
> ld <- seq(0, 5, 0.1)
> lines(ld, predict(budworm2.glm, data.frame(ldose = ld, sex = factor(rep("F",length(ld)), levels
= levels(sex))), type = "response"))
> lines(ld, predict(budworm2.glm, data.frame(ldose = ld, sex = factor(rep("M",length(ld)), levels
= levels(sex))), type = "response"))
```