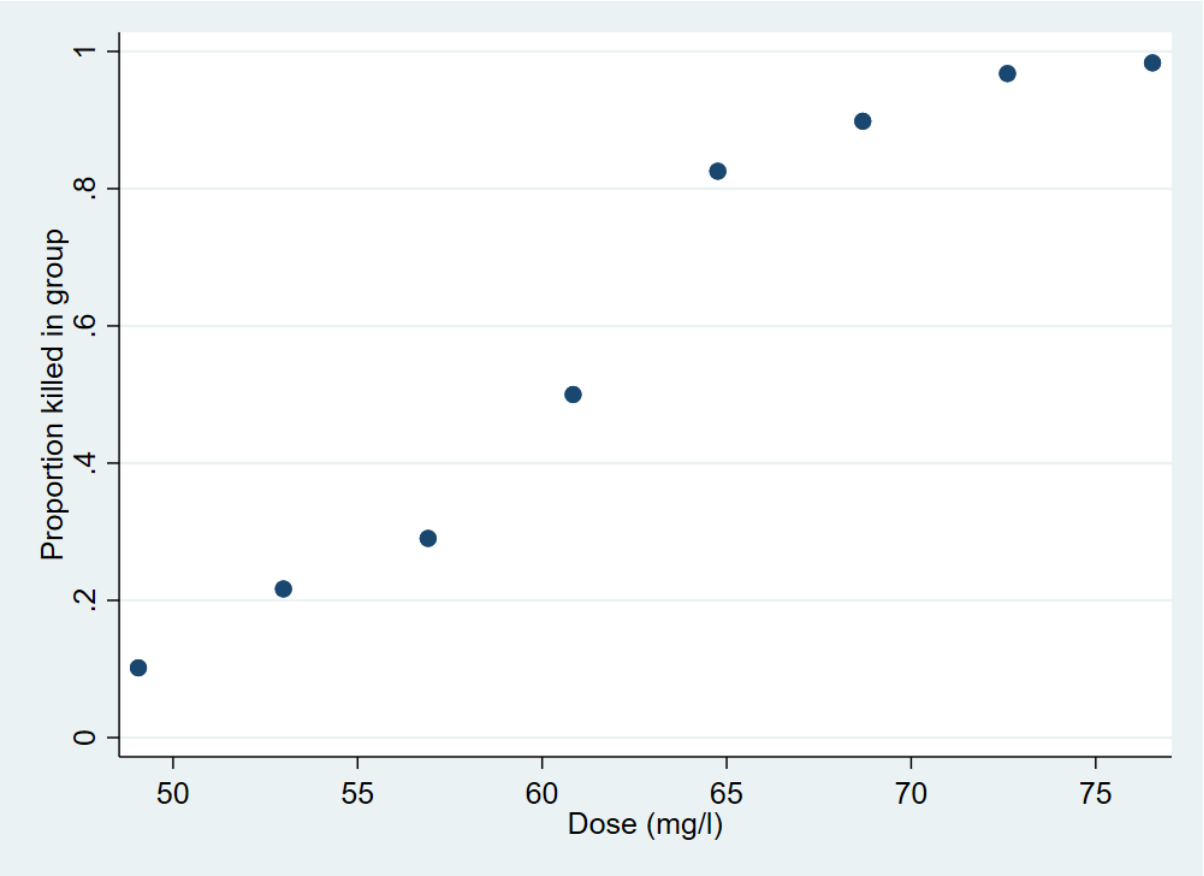
GLM Practical 3 Solutions

1. The dataset contains information on 481 insects. There are no obvious missing data, although, in general, when the data are grouped this may not always clear.

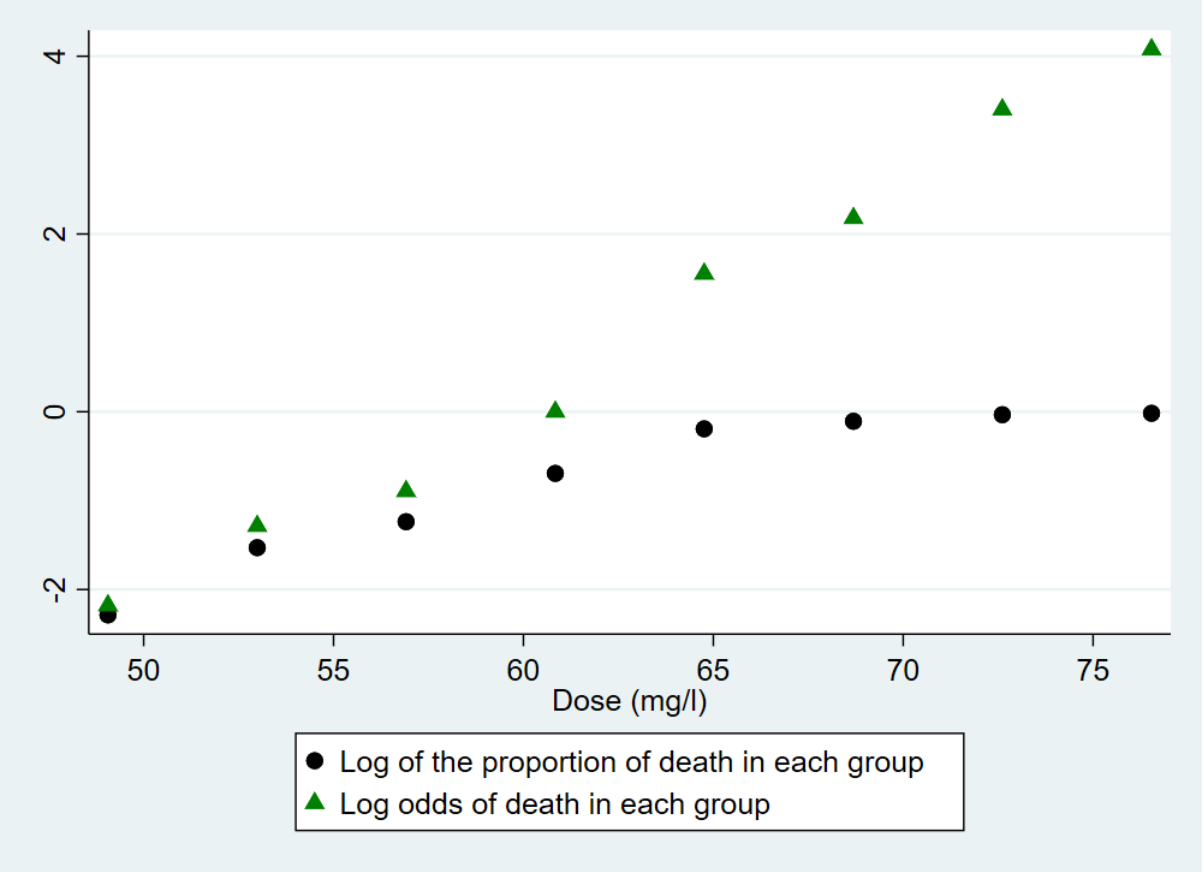
2. Linear regression is not appropriate because
   i. the plot shows a nonlinear sigmoid shape,
   ii. fitted values are not constrained to lie in [0,1] and
   iii. different observations have different variability because the variance of an estimated proportion depends both on the sample size and the proportion itself.



3. The plot (next page) for the log of the proportions is not linear. The plot for the log odds shows an approximately straight-line relationship.

Discussion: Looking at these plots, which link function is likely to best fit these data?

It is likely that the logit link function will give the best model fit for these data.



4. The random variable is the number of deaths in each group. Denote this by Y_i for the ith group and let d_i be the CS_2 dose (mg/l) in the i^{th} group. As for all GLMs, the model is defined by

i. the response distribution,
ii. the link function and
iii. the linear predictor

The response distribution is

$$Y_i \sim Bin(n_i, \pi_i) \quad for\ i = 1, 2, \dots, 8\ .$$

The expected value of Y_i is μ_i and the link function is

$$\eta_i = \log\left(\frac{\mu_i}{n_i - \mu_i}\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = logit(\pi_i)\ .$$

The linear predictor is

$$\eta_i = \beta_0 + \beta_1 d_i\ .$$

5.

```
. glm r dose , fam(bin n) link(logit)

Iteration 2:   log likelihood = -16.696989

Generalized linear models                         Number of obs   =          8
Optimization     : ML                             Residual df     =          6
                                                  Scale parameter =          1
Deviance         =  4.615478477                   (1/df) Deviance =   .7692464
Pearson          =  4.609224387                   (1/df) Pearson  =   .7682041

Variance function: V(u) = u*(1-u/n)               [Binomial]
Link function    : g(u) = ln(u/(n-u))             [Logit]

                                                  AIC             =   4.674247
Log likelihood   = -16.69698926                   BIC             =  -7.861171

------------------------------------------------------------------------------
             |                 OIM
           r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |   .2365929   .0203032    11.65   0.000     .1967995    .2763864
       _cons |   -14.0864   1.228393   -11.47   0.000    -16.49401    -11.6788
------------------------------------------------------------------------------
```

The Wald test of the linear term has a z-score of 11.65, and p<0.001.

a) The estimated probability of death at a dose of 55 mg/l is given by the following.

$$logit(\hat{\pi}_i) = -14.09 + (0.2366 \times 55) = -1.077$$

$$\text{So } \hat{\pi}_i = \frac{e^{-1.077}}{1+ e^{-1.077}} = 0.254$$

b) The dose that, on the basis of this model, would lead to a 50% death rate (LD50) is the dose that is associated with π = 0.5 and hence logit (π) = 0. This can be found by solving the following equation.

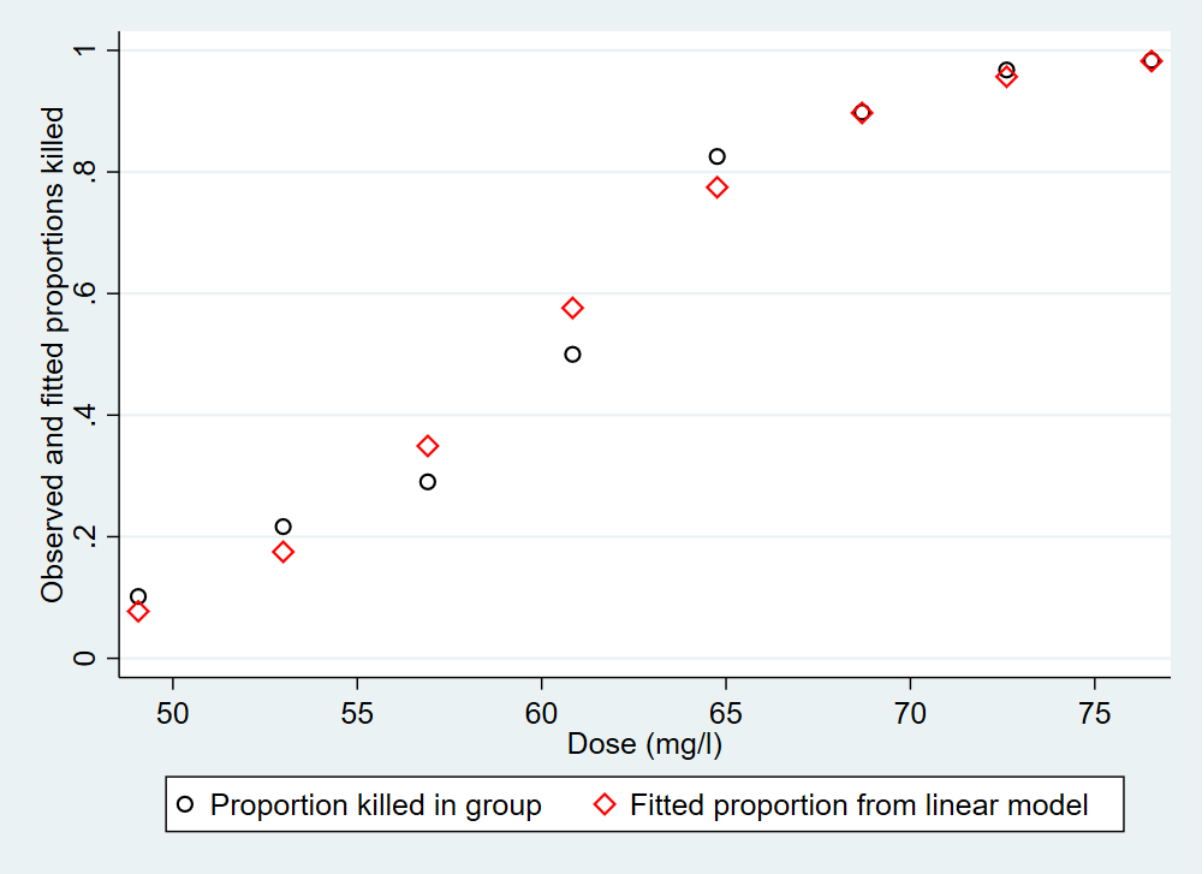
$$0 = -14.09 + (0.2366 \text{ x LD50})$$

Hence LD50 = 59.5 mg/l.

c) There is evidence of an increasing risk of death with increasing dose (z = 11.65, p<0.001, Wald test).

d) The odds ratio per unit mg/l increase in dose is $e^{0.2366}$ = 1.27. The 95% confidence interval extends from $e^{0.1968}$ (= 1.22) to $e^{0.2764}$ (= 1.32).

6. The predict command gives the fitted number of deaths in each group.



Discussion: Compare the fitted and observed proportions. What do you conclude?

Observed and fitted proportions are generally similar, with differences greatest in the third, fourth, and fifth dose groups.

7.

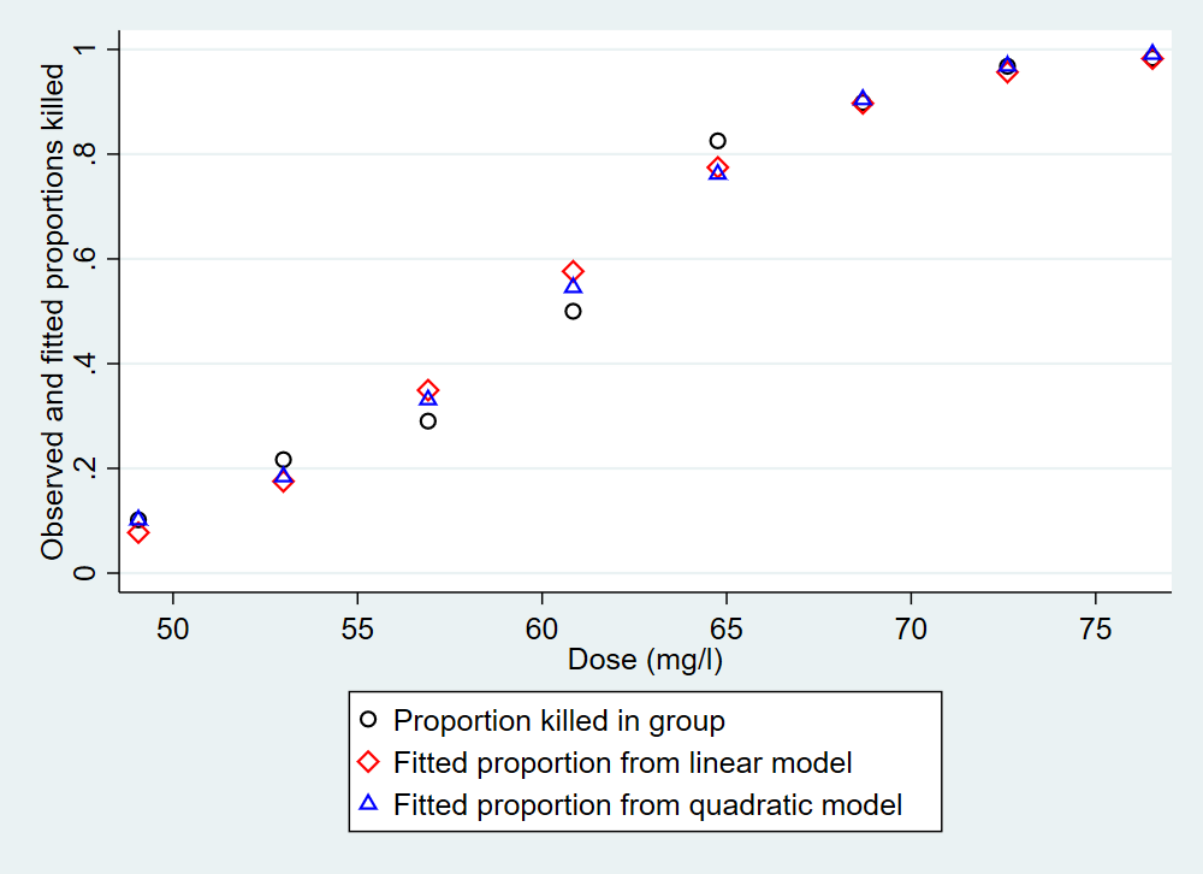
```
. glm r dose dose2, fam(bin n) link(logit) eform

Generalized linear models                         Number of obs   =          8
Optimization     : ML                             Residual df     =          5
                                                  Scale parameter =          1
Deviance         =  3.183602634                   (1/df) Deviance =   .6367205
Pearson          =  3.171441623                   (1/df) Pearson  =   .6342883

Variance function: V(u) = u*(1-u/n)               [Binomial]
Link function    : g(u) = ln(u/(n-u))             [Logit]

                                                  AIC             =   4.745263
Log likelihood   = -15.98105134                   BIC             =  -7.213605
```

r	Odds Ratio	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
dose	.8607035	.2833262	-0.46	0.649	.4514974	1.640786
dose2	1.003192	.002736	**1.17**	**0.243**	.9978439	1.008569
_cons	.0830612	.8196275	-0.25	0.801	3.31e-10	2.08e+07

a) The Wald test gives z=1.17 and p=0.243, so there is no evidence that the quadratic term improves the fit of the model.

b) The figure below shows that the quadratic model gives predictions that are somewhat closer to the observed proportions for the third and fourth dose groups, but the difference is marginal (and not statistically significant as tested above).

Because there is no compelling evidence of a lack of fit with the linear dose model we can use this linear dose model to describe the relationship between dose and probability of death.



8.

a)

Logit link

```
.  glm r i.dosecat, family(binomial n) link(logit)

Generalized linear models                         Number of obs   =          8
Optimization     : ML                             Residual df     =          0
                                                  Scale parameter =          1
Deviance         =  3.24185e-14                   (1/df) Deviance =          .
Pearson          =  4.74493e-24                   (1/df) Pearson  =          .

Variance function: V(u) = u*(1-u/n)               [Binomial]
Link function    : g(u) = ln(u/(n-u))             [Logit]

                                                  AIC             =   5.597313
Log likelihood   = -14.38925003                   BIC             =   3.24e-14

------------------------------------------------------------------------------
             |                 OIM
           r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dosecat |
           2 |   .8933342   .5326671     1.68   0.094    -.1506742    1.937343
           3 |   1.284715   .5136316     2.50   0.012     .2780151    2.291414
           4 |   2.178532   .5069153     4.30   0.000     1.184997    3.172068
           5 |   3.731881   .5437596     6.86   0.000     2.666132     4.79763
           6 |   4.357065   .6091545     7.15   0.000     3.163144    5.550986
           7 |    5.57973   .8379745     6.66   0.000      3.93733     7.22213
           8 |    6.25607   1.096578     5.71   0.000     4.106816    8.405324
             |
       _cons |  -2.178532   .4307373    -5.06   0.000    -3.022762   -1.334303
------------------------------------------------------------------------------
```

The constant is -2.179. This is the log odds of death in the baseline category (Group 1). Hence the probability is $e^{-2.179}/(1 + e^{-2.179}) = 0.10$.

The coefficient for the second dose category is 0.893. This is the log odds ratio comparing dose category 2 with category 1. Hence the log odds of death in category 2 is -2.179 + 0.893 = -1.285. It follows that the probability of death in group 2 is 0.22.

Log link

The constant is -2.286. This is the log probability of death in the baseline category (Group 1). Hence the probability is $e^{-2.286} = 0.10$.

The coefficient for the second dose category is 0.756. This is the log probability ratio comparing dose category 2 with category 1. Hence the log probability of death in category 2 is -2.286 + 0.756 = -1.529, and the probability of death is $e^{-1.529} = 0.22$.

Identity link

The constant is 0.102. This is simply the probability of death in the baseline category.

The coefficient for the second dose category is 0.115. This is the probability difference comparing categories 2 and 1. The probability of death in category 2 is 0.102 + 0.115 = 0.22.

Similar principles apply for the remaining coefficients.

Note that in all cases the predicted probabilities are the same as the observed probabilities. This is because the model contains the same number of parameters as there are groups, so it is possible for the model to exactly match the observed values. This is known as a saturated model (details in the next session).

b)

```
. glm r dose i.dosecat, family(binomial n) link(logit)
note: 8.dosecat omitted because of collinearity

Generalized linear models                         Number of obs   =          8
Optimization     : ML                             Residual df     =          0
                                                  Scale parameter =          1
Deviance         =  1.06581e-14                   (1/df) Deviance =          .
Pearson          =  4.68278e-24                   (1/df) Pearson  =          .

Variance function: V(u) = u*(1-u/n)               [Binomial]
Link function    : g(u) = ln(u/(n-u))             [Logit]

                                                  AIC             =   5.597313
Log likelihood   = -14.38925003                   BIC             =   1.07e-14

------------------------------------------------------------------------------
             |                 OIM
           r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |    .227659   .0399046     5.71   0.000     .1494475    .3058706
             |
     dosecat |
          2  |  -.0013658   .5052332    -0.00   0.998    -.9916047    .9888731
          3  |  -.5024084   .5059082    -0.99   0.321     -1.49397    .4891535
          4  |  -.5032906   .5646826    -0.89   0.373    -1.610048     .603467
          5  |    .157634   .6900568     0.23   0.819    -1.194852    1.510121
          6  |   -.111882   .8482936    -0.13   0.895    -1.774507    1.550743
          7  |     .21836    1.12576     0.19   0.846    -1.988089    2.424809
          8  |          0  (omitted)
             |
       _cons |  -13.34748   2.163485    -6.17   0.000    -17.58784   -9.107131
------------------------------------------------------------------------------

. test 2.dosecat 3.dosecat 4.dosecat 5.dosecat 6.dosecat 7.dosecat 8.dosecat

 ( 1)  [r]2.dosecat = 0
 ( 2)  [r]3.dosecat = 0
 ( 3)  [r]4.dosecat = 0
 ( 4)  [r]5.dosecat = 0
 ( 5)  [r]6.dosecat = 0
 ( 6)  [r]7.dosecat = 0

           chi2(  6) =    4.56
         Prob > chi2 =    0.6017
```

Discussion: Why are there estimates for only six (of eight) dose categories in this model? Discuss the interpretation of each parameter estimate and the test statistic computed by the final command above.

Since there are eight distinct dose categories models can include a maximum eight parameters. If a linear effect of dose and an intercept are included, only six indicator variables can be included. Stata has dropped the first and last categories.

The line defined by the intercept (_cons) and slope (dose) parameters will pass through the observed log odds for the highest and lowest dose categories (to check this use the following Stata commands).

```
gen extreme = (dosecat==1) + (dosecat==8)
glm r dose if extreme, family(binomial n) link(logit)
```

The indicator parameters therefore represent departures from this line for the other dose categories. The test command given tests their joint effects and hence can be regarded as (another) test of non-linearity.

Discussion: Working together with one or more colleagues (in your Breakout Room if online), write a paragraph to answer the aims of the analysis. If online, one of you should post your group's paragraph in the Zoom chat.

Overall conclusions from the study are that there is evidence that the probability of death increases with increasing CS_2 dose ($p<0.001$). Odds of death are estimated to increase by 27% (95% confidence interval 22% to 32%) per unit mg/l increase in dose and there is no evidence of non-linearity on a multiplicative odds (log odds) scale.