GLM Practical 8 Solutions

1. The `los` variable has a highly positively skewed (non-normal) distribution. It ranges from 1 to 116 days, with an observed mean of 9.85 and median of 8. The variance (78.0) is much larger than the mean.

```
. ci means los

 Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
----------+---------------------------------------------------------------
      los |      1,495    9.854181    .2284457        9.406072    10.30229
```

The 95% CI extends from 9.406 to 10.302. This CI is valid provided the variable is normally distributed. Even if the variable is not normally distributed the coverage of the CI will be close to 95% if the sample size is sufficiently large. This is a consequence of the central limit theorem. Here the coverage should be reasonable since the sample size is quite large.

2.

```
. glm los, fam(poisson) link(log) nolog

Generalized linear models                         Number of obs   =      1,495
Optimization     : ML                             Residual df     =      1,494
                                                  Scale parameter =          1
Deviance         =  8901.134077                   (1/df) Deviance =   5.957921
Pearson          =  11828.70662                   (1/df) Pearson  =   7.917474

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   9.778116
Log likelihood   = -7308.141824                   BIC             =  -2019.829

------------------------------------------------------------------------------
             |                 OIM
         los |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   2.287896   .0082389   277.69   0.000     2.271748    2.304044
------------------------------------------------------------------------------
```

The estimated constant coefficient in this model (2.288) represents the log of the mean length of stay. We can see this by calculating $e^{2.288} = 9.854$. Analogously, we see that the 95% CI for the mean is estimated as $e^{2.272} = 9.696$ to $e^{2.304} = 10.015$.

3.

```
. glm los, fam(poisson) link(log) nolog robust

Iteration 0:   log pseudolikelihood = -7354.5568
Iteration 1:   log pseudolikelihood = -7308.2078
Iteration 2:   log pseudolikelihood = -7308.1418
Iteration 3:   log pseudolikelihood = -7308.1418

Generalized linear models                         Number of obs   =      1,495
Optimization     : ML                             Residual df     =      1,494
                                                  Scale parameter =          1
Deviance         =  8901.134077                   (1/df) Deviance =   5.957921
Pearson          =  11828.70662                   (1/df) Pearson  =   7.917474

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   9.778116
Log pseudolikelihood = -7308.141824               BIC             =  -2019.829

------------------------------------------------------------------------------
             |               Robust
         los | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _cons |   2.287896   .0231826    98.69   0.000     2.242459    2.333333
------------------------------------------------------------------------------
```

As in question 2 the exponent of the estimated constant coefficient in this model (2.288) represents the mean length of stay ($e^{2.288} = 9.854$). Using a robust standard error, the 95% CI for the mean is estimated as $e^{2.242} = 9.416$ to $e^{2.333} = 10.312$.

Discussion: Compare and contrast the 95% confidence intervals in questions 1, 2 and 3? Which is most appropriate? Which is least appropriate?

The estimated mean is the same using all three approaches. The 95% CI is much narrower when making the Poisson assumption (9.696 to 10.015) than using either of the other two approaches (9.406 to 10.302 and 9.416 to 10.312), which are very similar to each other.

The difference between the 95% CIs in questions 2 and 3 can be attributed to overdispersion in `los`: both the deviance and Pearson statistics, divided by the residual degrees of freedom, are considerably larger than one. This makes sense given the context; the underlying severity of incoming patients' conditions will vary, leading to more variability than predicted by the Poisson distribution. So, the approach in question 2 is not appropriate.

The approach in question 1 takes no account of the skewness of the distribution of `los` (the 95% CI here is symmetric about the mean on an arithmetic scale), so the approach in question 3 is probably to be preferred (here the 95% CI is symmetric about the mean on a multiplicative scale), but (because of the large sample size and the central limit theorem) the approach in question 1 is also acceptable.

4.

```
. glm los i.type2 i.type3, fam(poisson) link(log) eform

Iteration 0:   log likelihood = -7029.5631
Iteration 1:   log likelihood = -6949.5862
Iteration 2:   log likelihood = -6949.3886
Iteration 3:   log likelihood = -6949.3886

Generalized linear models                         Number of obs   =      1,495
Optimization     : ML                             Residual df     =      1,492
                                                  Scale parameter =          1
Deviance         =  8183.627612                   (1/df) Deviance =   5.485005
Pearson          =   9396.93324                   (1/df) Pearson  =   6.298213

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   9.300854
Log likelihood   = -6949.388592                   BIC             =  -2722.716

------------------------------------------------------------------------------
             |                 OIM
         los |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.type2 |   1.267877   .0265014    11.35   0.000     1.216985    1.320897
     1.type3 |   2.065477   .0535019    28.00   0.000     1.963233    2.173046
       _cons |   8.830688   .0882451   217.98   0.000     8.659413     9.00535
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.

. est store A

. lrtest A B

Likelihood-ratio test                                 LR chi2(2)  =    717.51
(Assumption: A nested in B)                           Prob > chi2 =    0.0000
```

(Note that the use of the `eform` option avoids the need to exponentiate estimates).

The mean length of stay for elective admissions is 8.83 days (95% CI 8.66 to 9.01 days). Urgent admissions are associated with stays that are on average 26.8% (95% CI 21.7% to 32.1%) longer than those resulting from elective admissions. Emergency admissions result in stays that are on average 2.06 (95% CI 1.96 to 2.17) times as long as those from elective admissions.

However, note that all of these 95% CI rely on the Poisson assumption being correct and in fact this assumption is likely incorrect due to overdispersion as described above (note that both the deviance and Pearson statistics, divided by the residual degrees of freedom, are considerably larger than one).

5. To use a likelihood ratio test of whether age group has an independent effect, we add `i.age` to the model and compare log-likelihoods.

```
. glm los i.age i.type2 i.type3, fam(poisson) link(log) nolog eform

Generalized linear models                         Number of obs   =      1,495
Optimization     : ML                             Residual df     =      1,484
                                                  Scale parameter =          1
Deviance         =   8165.18541                   (1/df) Deviance =   5.502147
Pearson          =  9346.752373                   (1/df) Pearson  =   6.298351

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   9.299221
Log likelihood   = -6940.167491                   BIC             =  -2682.679

------------------------------------------------------------------------------
             |                 OIM
         los |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |
          2  |   .7755658   .0956944    -2.06   0.039     .6089643    .9877464
          3  |   .8176414   .0972513    -1.69   0.091     .6476194      1.0323
          4  |   .7910486   .0931576    -1.99   0.047     .6280029    .9964251
          5  |   .7659533   .0901755    -2.26   0.024     .6081218    .9647483
          6  |   .7951697   .0935329    -1.95   0.051     .6314458    1.001345
          7  |   .7593596   .0901184    -2.32   0.020     .6017686    .9582205
          8  |   .7239164   .0878831    -2.66   0.008     .5706278    .9183832
          9  |   .7316548     .09231    -2.48   0.013     .5713648    .9369124
             |
     1.type2 |   1.265918   .0265407    11.25   0.000     1.214954    1.319021
     1.type3 |   2.064878    .053619    27.92   0.000     1.962417     2.17269
       _cons |   11.32913   1.320198    20.83   0.000      9.01581      14.236
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.

. est store B

. lrtest A B

Likelihood-ratio test                             LR chi2(8)  =     18.44
(Assumption: A nested in B)                       Prob > chi2 =    0.0181
```

Provided that the assumptions of the Poisson model hold (which is in fact extremely unlikely, as explained above) there is evidence (p = 0.018) that age group has an effect on length of hospital stay, after adjusting for admission type.

6. To allow for overdispersion refit the model with the `robust` option.

```
. glm los i.age i.type2 i.type3, fam(poisson) link(log) eform robust nolog

Generalized linear models                         Number of obs   =      1,495
Optimization     : ML                             Residual df     =      1,484
                                                  Scale parameter =          1
Deviance         =   8165.18541                   (1/df) Deviance =   5.502147
Pearson          =  9346.752373                   (1/df) Pearson  =   6.298351

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   9.299221
Log pseudolikelihood = -6940.167491               BIC             =  -2682.679

------------------------------------------------------------------------------
             |               Robust
         los |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |
          2  |   .7755658   .1792692    -1.10   0.272     .4930223    1.220031
          3  |   .8176414    .173137    -0.95   0.342     .5399076    1.238244
          4  |   .7910486   .1623995    -1.14   0.254     .5289985     1.18291
          5  |   .7659533   .1574871    -1.30   0.195     .5119027    1.146086
          6  |   .7951697   .1646582    -1.11   0.268     .5299061    1.193221
          7  |   .7593596   .1589473    -1.32   0.188     .5038207    1.144508
          8  |   .7239164   .1535639    -1.52   0.128     .4776652    1.097118
          9  |   .7316548   .1652243    -1.38   0.166     .4699868    1.139008
             |
     1.type2 |   1.265918    .067136     4.45   0.000     1.140942    1.404585
     1.type3 |   2.064878   .2416633     6.20   0.000     1.641625    2.597257
       _cons |   11.32913   2.271633    12.11   0.000     7.647503    16.78314
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.


.  testparm i.age

 ( 1)  [los]2.age = 0
 ( 2)  [los]3.age = 0
 ( 3)  [los]4.age = 0
 ( 4)  [los]5.age = 0
 ( 5)  [los]6.age = 0
 ( 6)  [los]7.age = 0
 ( 7)  [los]8.age = 0
 ( 8)  [los]9.age = 0

           chi2(  8) =    4.09
         Prob > chi2 =    0.8493
```

Discussion: Compare and contrast the results of the tests in questions 4 and 5. Which is the appropriate test?

Using the `robust` option does not change the parameter estimates. However, the SEs have changed dramatically: they have all increased substantially. Further the test of differences in rate by age category is no longer statistically significant.

These differences can be attributed to the fact that with `robust`, the validity of the SEs does not rely on the Poisson assumption holding. Here, the deviance and Pearson statistics are both much larger than the residual degrees of freedom, suggesting that there is overdispersion, violating the Poisson assumption. The robust option relaxes this, and we see that in fact the test of differences in rate by age category is not statistically significant and that all our estimates are much less precise than implied by the model-based SEs.

Discussion: Working together with one or more colleagues (in your Breakout Room if online), write a short paragraph (suitable for a medical journal) to summarise your findings concerning the effects of the type of admission on length of stay in hospital for this model. If online, one of you should post your group's paragraph in the Zoom chat.

"After adjusting for age urgent admissions are associated with stays that are on average 26.6% (95% CI 14.1% to 40.5%) longer than those resulting from elective admissions, whilst emergency admissions are associated with stays that are on average 2.065 (95% CI 1.642 to 2.597) times as long as those from elective admissions."

Or (if you don't like mixing percentage increases and multiplicative effects):

"After adjusting for age urgent admissions are associated with stays that are on average 26.6% (95% CI 14.1% to 40.5%) longer than those resulting from elective admissions, whilst emergency admissions are associated with stays that are on average 106.5% (95% CI 64.2% to 159.7%) times longer than those from elective admissions."

7. The negative binomial model can be fitted as follows.

```
. nbreg los , nolog

Negative binomial regression                    Number of obs     =      1,495
                                                LR chi2(0)        =       0.00
Dispersion     = mean                           Prob > chi2       =          .
Log likelihood =  -4856.494                     Pseudo R2         =     0.0000

------------------------------------------------------------------------------
         los |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   2.287896   .0198946   115.00   0.000     2.248903    2.326888
-------------+----------------------------------------------------------------
    /lnalpha |  -.7128727   .0431289                     -.7974038   -.6283417
-------------+----------------------------------------------------------------
       alpha |   .4902339   .0211432                      .450497    .5334758
------------------------------------------------------------------------------
LR test of alpha=0: chibar2(01) = 4903.30                 Prob >= chibar2 = 0.000
```

The likelihood ratio test of the null that $\alpha = 0$ is highly statistically significant: there is very strong evidence against the null hypothesis that `los` (marginally) follows a Poisson distribution. There is more variability than would be expected from a Poisson variable. This can also be seen by noting (use `summ los, detail`) that the sample variance of `los` (78.02) is much larger than its sample mean (9.85). This also indicates why the inferences from the Poisson model in part 2. were anti-conservative.

8. Fit the negative binomial regression model relating los to type of admission as follows.

```
. nbreg los i.type2 i.type3 , nolog

Negative binomial regression                    Number of obs     =      1,495
                                                LR chi2(2)        =     112.55
Dispersion     = mean                           Prob > chi2       =     0.0000
Log likelihood = -4800.2189                     Pseudo R2         =     0.0116

------------------------------------------------------------------------------
         los |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.type2 |   .2373439   .0502166     4.73   0.000     .1389211    .3357666
      1.type3 |   .7253612   .0757014     9.58   0.000     .5769893    .8737332
       _cons |   2.178233   .0222433    97.93   0.000     2.134637    2.221829
-------------+----------------------------------------------------------------
    /lnalpha |  -.8033559   .0443828                     -.8903446   -.7163671
-------------+----------------------------------------------------------------
       alpha |   .4478236   .0198757                      .4105142    .4885238
------------------------------------------------------------------------------
LR test of alpha=0: chibar2(01) = 4298.34                 Prob >= chibar2 = 0.000
```

From the fitted model the predicted mean lengths of admission are $e^{2.1782}$ = 8.83 days, $e^{(2.1782+0.2373)}$ = 11.20 days and $e^{(2.1782+0.7253)}$ = 18.24 days for elective, urgent and emergency admissions respectively. These are identical to the observed means in the three groups.

To compute the predicted variances use the result that, for negative binomial regression, if the expectation of the outcome is μ then its variance is $\mu(1 + \alpha\mu)$.

Since the estimate of α is 0.4478 the estimated variances are

$8.83 \times (1 + 0.4478 \times 8.83) = 43.7$ days^2,
$11.20 \times (1 + 0.4478 \times 11.20) = 67.3$ days^2 and
$18.24 \times (1 + 0.4478 \times 18.24) = 167.2$ days^2

for elective, urgent and emergency admissions respectively.

These can be compared with the observed variances (use `summ los if type1==1, detail` etc.) which are 41.7, 77.9 and 424.9 days^2 respectively.

Discussion: What do you conclude about the estimates and variances predicted from the negative binomial model? What are your conclusions about the best way to model these data?

The sample variances are close to those predicted by the negative binomial model for the elective and urgent admissions. However, for the emergency admissions the sample variance is quite a bit larger than predicted by the model. The number of emergency admissions is relatively small, meaning that the variance will be imprecisely estimated but the lack of agreement between the fitted and observed variances could also indicate that the Gamma random-effects distribution is not appropriate for these data (or that a different α is needed for each type of admission).

Another issue that we have overlooked in our analyses is that the hospital days only take positive values, whereas Poisson distributions (and negative binomials) should also sometimes take the value zero. One approach which could be used to allow for this is to instead model "number of days of stay after day of admission", which would take the value zero for some patients.