GLM Practical 4 Solutions

Part A: Investigation into low birthweights in Massachusetts, USA

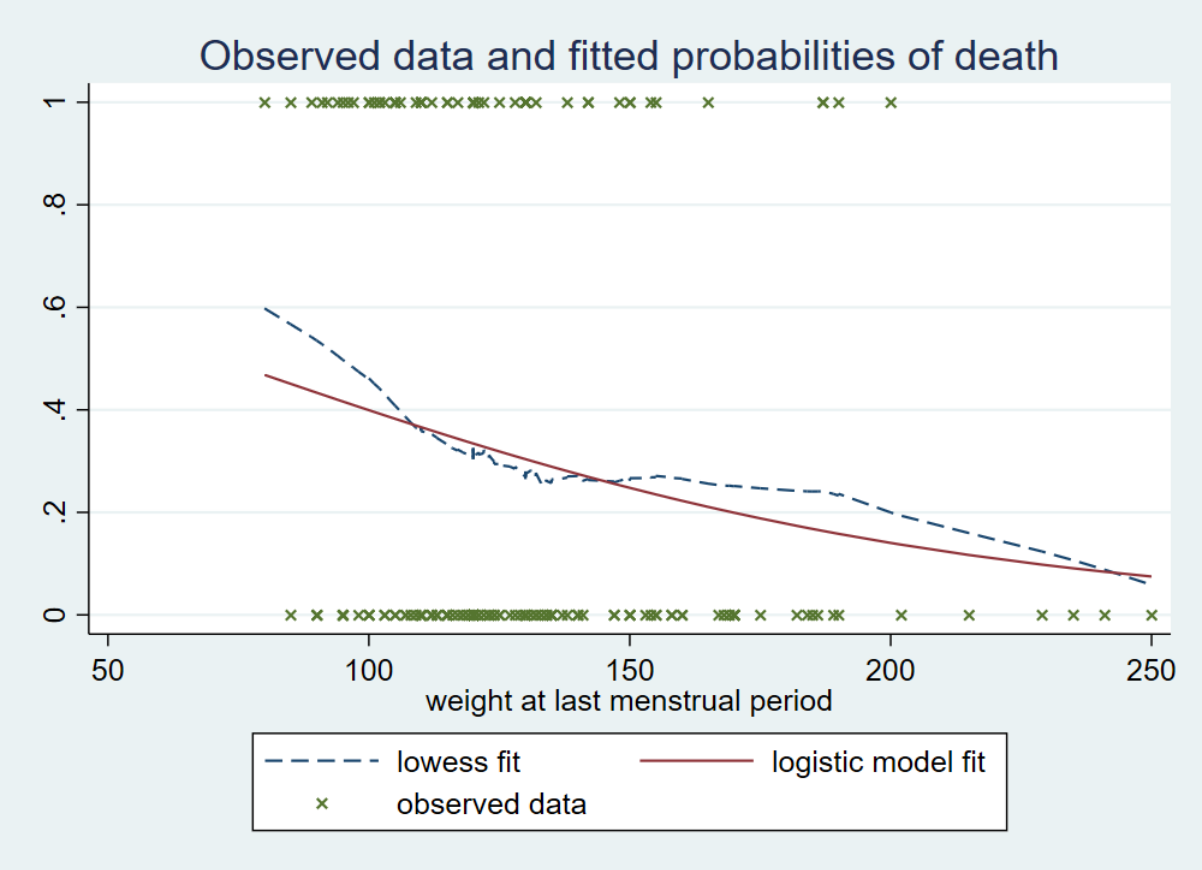
1.

```
glm low lwt, fam(bin) link(logit)
logistic low lwt
logit low lwt
```

Discussion: What are the main differences between the commands and the output they produce?

The glm command reports estimates of the log odds (at mother's weight zero, an extrapolation) and log odds ratios (for a 1 lb increase in mother's weight), and also deviance and Pearson statistics. The logistic command reports odds and odds ratios. The logit command, like glm, reports log odds and log odds ratios. Note that the commands differ somewhat with regards to what post estimation commands are available after fitting the model.

2.



Discussion: What do you conclude regarding the appropriateness of the linearity assumption here?

The lowess line shows some variation around the fitted values (there will always be some due to sampling variability), but it broadly follows the line of fitted values calculated from the model. From the plot we can conclude that assuming a linear (on the log odds scale) effect of mother's weight is not unreasonable here.

3. The estimated log odds of low birthweight for a mother who weighed 120 lbs is

$$(120 \times -0.014037) + 0.995763 = -0.6887 .$$

Hence, the probability is

$$\exp(-0.6887) / (1 + \exp(-0.6887)) = 0.334 .$$

4. Using lincom:

```
lincom 120*lwt + _cons, eform

( 1)  120*[low]lwt + [low]_cons = 0

------------------------------------------------------------------------------
         low |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .5022331   .0813097    -4.25   0.000      .365678    .6897819
------------------------------------------------------------------------------
```

This gives the estimated odds of low birthweight for a mother who weighed 120 lbs.

Rearranging the formula odds=probability/(1-probability) gives probability=; odds/(odds +1). So, the estimated odds of low birthweight for a mother who weighed 120 lbs is

$$0.5022331 / (1 + 0.5022331) = 0.334 \text{ (as calculated by hand above).}$$

The lower limit of the 95% confidence interval for the probability can be calculated as:

$$0.365678 / (1 + 0.365678) = 0.268 .$$

Similarly, the upper limit is given by

$$0.6897819 / (1 + 0.6897819) = 0.408 .$$

Alternatively, re-parameterise the model:

```
. logistic low lwt120

Logistic regression                                     Number of obs     =        189
                                                        LR chi2(1)        =       5.96
                                                        Prob > chi2       =     0.0146
Log likelihood = -114.35403                             Pseudo R2         =     0.0254

------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lwt120 |   .9860609   .0060825    -2.28   0.023     .9742112    .9980549
       _cons |   .5022331   .0813097    -4.25   0.000      .365678    .6897819
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Here, the constant term is the odds of low birth weight for a woman of 120 lbs; it is the same result as given by the lincom command above.

5.

```
. logit low i.race lwt

Logistic regression                                     Number of obs     =        189
                                                        LR chi2(3)        =      11.40
                                                        Prob > chi2       =     0.0098
Log likelihood = -111.63836                             Pseudo R2         =     0.0486

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |
       black |   1.080719   .4880135     2.21   0.027     .1242306    2.037208
       other |   .4810363   .3566464     1.35   0.177    -.2179778     1.18005
             |
         lwt |  -.0152009   .0064381    -2.36   0.018    -.0278193   -.0025825
       _cons |   .8029594   .8450592     0.95   0.342    -.8533262    2.459245
------------------------------------------------------------------------------

. est store A

. logit low lwt

Logistic regression                                     Number of obs     =        189
                                                        LR chi2(1)        =       5.96
                                                        Prob > chi2       =     0.0146
Log likelihood = -114.35403                             Pseudo R2         =     0.0254

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |  -.0140371   .0061685    -2.28   0.023    -.0261272    -.001947
       _cons |   .9957626   .7852434     1.27   0.205    -.5432862    2.534811
------------------------------------------------------------------------------

. est store B
```

```
. lrtest A B

Likelihood-ratio test                                 LR chi2(2)  =      5.43
(Assumption: B nested in A)                           Prob > chi2 =    0.0662
```

Discussion: What do you conclude?

Comparing the fits of these models tests the null hypothesis that race has no independent association with the probability of having a low birthweight baby, after adjusting for mother's weight. The log-likelihood ratio statistic for this comparison is $-2 \times (-114.35403 - -111.63836) = 5.43$ which is compared to a chi-square distribution with two degrees of freedom, giving p = 0.066. There is some weak evidence that race is independently associated with low birthweight, after adjusting for mother's weight. A Wald test of the same null hypothesis gives a very similar p-value here (see do file).

6. From the model with both mother's weight and race as covariates we have:

```
LR chi2(3) = 11.40
Prob > chi2 = 0.0098
```

The LR value and p-value match those from the profile likelihood ratio test comparing this model with the null model (no covariates). We can therefore conclude that the LR statistic and p-value displayed in the top right after fitting a logistic regression model corresponds to the likelihood ratio test of the null hypothesis that the coefficients corresponding to the covariates included in the non-null model are all zero. That is, it is testing whether the outcome/dependent variable is independent of all the included covariates. It is analogous to the overall F-test from a linear regression model.

Note that this test is often not very useful in practice. In this example all we can conclude from this test is that there is evidence that mother's weight and/or race are associated with the odds of a low birthweight baby. We cannot conclude that there is evidence that both factors are associated with the outcome, only that there is evidence that at least one of them is.

7.

```
. qui logit low i.race lwt

. estat gof, group(10)

Logistic model for low, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =       189
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =      7.61
                 Prob > chi2 =      0.4720
```

Discussion: What do you conclude about the goodness of fit of this model and about the Hosmer-Lemeshow test?

The Hosmer-Lemeshow test with 10 groups gives a p-value of 0.47, indicating no evidence that the model does not fit well. However, performing the test with 5 groups gives a rather different p-value of 0.18. Such differences when different choices of the number of groups are made are quite common, and have led some to conclude that the test is unreliable.

Part B: Investigation into effect of CS_2 at different doses on insect survival

8.

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

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
------------------------------------------------------------------------------

. dis chi2tail(6, 4.615)

.59405092
```

The deviance reported from the fitted model is 4.615. To test whether the model which assumes a linear (on the log odds scale) dose effect fits the data well (or rather if there is evidence of poor fit), we compare 4.615 to a chi-squared distribution on n - p degrees of freedom. Here n = 8 and p = 2, so we have six degrees of freedom and we obtain a p-value of 0.59. Thus, there is no evidence against the null hypothesis that the linear dose model is correctly specified. Note that this test is equivalent to the profile likelihood ratio test comparing this model to the saturated model which treats dose as a factor variable.

Discussion: Why can the deviance be used to test the fit of this model, but not that of any of the models in Part A?

The deviance can be used to test the fit of this model because the data are grouped. It could not be used for the models in Part A because there the data is for individuals.

9.

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

. gen dose2 = dose^2

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

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

------------------------------------------------------------------------------
             |                 OIM
           r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |  -.1500052   .3291798    -0.46   0.649    -.7951857    .4951753
       dose2 |    .003187   .0027273     1.17   0.243    -.0021584    .0085325
       _cons |  -2.488177   9.867752    -0.25   0.801    -21.82862    16.85226
------------------------------------------------------------------------------

. est store F
. lrtest D F

Likelihood-ratio test                             LR chi2(1)  =      1.43
(Assumption: D nested in F)                       Prob > chi2 =    0.2315
```

The likelihood ratio test comparing the two models gives p = 0.2315, indicating no evidence against the null hypothesis that the linear dose model is correctly specified. There is thus no evidence of a quadratic effect of dose. The Wald test p-value for the same null hypothesis is p = 0.243. They are very close, and typically they will be, since they are asymptotically equivalent.