Contents

1 Modelling Categorical Outcomes

If the outcome that you wish to model is a categorical variable with more than 2 categories, a more complex model than logistic regression is required. The exact model will depend on whether the categorical variable is nominal or ordinal.

1.1 Nominal Outcomes

1.1.1 Cross-Tabulation

We have seen previously that a 2 × 2 table could be used to examine an association between two dichotomous variables. In fact, the same approach can be used for any two categorical variables, irrespective of the number of categories. If the variable used to define the rows of the table has R categories, and the variable used to define the columns has C categories, you end up with an $R \times C$ table. The expected number of observations in each cell which lies at the intersection of row r and column c can be calculated as

$$E_{rc} = P_r \times P_c \times N = \frac{N_r \times N_c}{N} \tag{1.1}$$

where

P_r Proportion of the sample that is in row r.

P_c Proportion of the sample that is in column c.

N Total sample size.

N_r Number of observations that are in row r.

N_c Number of observations that are in column c.

Just as in the dichotomous case, we can calculated a test statistic by summing $\frac{(O_{rc}-E_{rc})^2}{E_{rc}}$ for each cell in the table. However, we will now have $R \times C$ terms to add, and this statistic will follow a χ^2 distribution on $(R-1) \times (C-1)$ degrees of freedom if the null hypothesis is true. In stata, this test can be performed in exactly the same way as the χ^2-test for 2 × 2 tables, the only difference being the number of categories in the two variables passed to the command.

For example, consider the table below, drawn up to see if males and females tend to have different preferences for their medical insurance (data taken from stata's built in Health Insurance dataset, and can be loaded into stata with the command `webuse sysdsn1`).

	Females		Males		Total	
Indemnity	234	(50.7%)	60	(39.0%)	294	(47.7%)
Prepaid	196	(42.4%)	81	(52.6%)	277	(45.0%)
No Insurance	32	(6.9%)	13	(8.4%)	45	(7.3%)
Total	462	(100%)	154	(100%)	616	(100%)

Table 1.1: A 3 by 2 table

A smaller proportion of men than women have indemnity insurance, whereas a larger proportion of men than women have prepaid or no insurance. We can quantify these differences with what stata calls the "Relative Risk Ratio". The relative risk of having prepaid rather than indemnity insurance in males is $\frac{0.526}{0.390} = 1.35$, whereas in females it is $\frac{0.42}{0.51} = 0.84$. The relative risk ratio is therefore $\frac{1.35}{0.84} = 1.61$. Similarly, the relative risk ratio for no insurance rather than indemnity is $\frac{0.08/0.39}{0.07/0.51} = 1.58$

1.1.2 Multinomial Logistic Regression

Whilst it is possible to use a tabulation to get some information about the magnitude of the association between a categorical predictor and a categorical outcome. However, there are occasions when we want to include multiple predictors at the same time, some of which may be quantitative. It is possible to extend the notion of logistic regression to the case where the outcome has more than two categories: this is known as multinomial logistic regression.

Multiple Logistic Regressions

It is easiest to think of multinomial logistic regression as a series of dichotomous logistic regressions. If our outcome variable has R possible categories, we choose one of the categories as our reference or baseline category, and perform $R - 1$ binary logistic regressions, with the outcome taking the value 0 for the reference category in each case, and the value 1 for one particular outcome category (and is considered missing for the other possible outcome categories.

So in the data presented above, we could do a logistic regression of prepaid vs indemnity, and a second logistic regression of no insurance vs indemnity. In this case, indemnity is our reference category. The output of performing these two logistic regressions

```
. logistic insure1 male

------------------------------------------------------------------------------
     insure1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.611735   .3157844     2.44   0.015      1.09779     2.36629
       _cons |   .8376068   .0811033    -1.83   0.067     .6928203    1.012651
------------------------------------------------------------------------------

. logistic insure2 male

------------------------------------------------------------------------------
     insure2 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.584375   .5693029     1.28   0.200     .7834322    3.204163
       _cons |   .1367521   .0257746   -10.56   0.000     .0945154    .1978636
------------------------------------------------------------------------------
```

```
. logistic insure1 male

------------------------------------------------------------------------------
     insure1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.611735   .3157844     2.44   0.015      1.09779     2.36629
       _cons |   .8376068   .0811033    -1.83   0.067     .6928203    1.012651
------------------------------------------------------------------------------

. logistic insure2 male

------------------------------------------------------------------------------
     insure2 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.584375   .5693029     1.28   0.200     .7834322    3.204163
       _cons |   .1367521   .0257746   -10.56   0.000     .0945154    .1978636
------------------------------------------------------------------------------
```

You can see that the odds ratios calculated by the two logistic regressions correspond to the relative risk ratios we calculated from Table 1.1. However, I suggest that you think of them as relative risk ratios rather than odds ratios, since they are only odds ratios if you ignore any possible outcome other than the reference outcome and the one being predicted.

In that case, $P(referenceoutcome) = 1 - P(regressionoutcome)$, so that the relative risk $\frac{P(regressionoutcome)}{P(referenceoutcome)} = \frac{P(regressionoutcome)}{1-P(regressionoutcome)}$, which is the odds of the regression outcome. The ratio of the relative risks is then a ratio of odds.

Combining Multiple Logistic Regressions

Rather then perform multiple logistic regressions in this way, it is possible to fit a single model covering all possible outcomes. Rather than a single linear predictor, as in binary logistic regression, there will need to be $R - 1$, corresponding to the separate logistic regressions that could be fitted. If we represent the linear predictor for the i^{th} observation when comparing the j^{th} outcome to the reference outcome (which we will take to be R) as LP_{ij}, then probability the the outcome for the i^{th} observation takes the value j can be calculated as

$$P(Y_i = j|\boldsymbol{X_i}) = \begin{cases} \frac{exp(LP_{ij})}{1+\sum_{m=2}^{R} exp(LP_{im})} & \text{if } j < R \\ \frac{1}{1+\sum_{m=2}^{R} exp(LP_{im})} & \text{if } j = R \end{cases} \qquad (1.2)$$

Multinomial Logistic Regression in Stata

The stata command for fitting this model is `mlogit`. To fit a multinomial logistic regression model for the above data, the command would be

```
mlogit insure male
```

The output from the above command would be

```
Multinomial logistic regression                   Number of obs     =        616
                                                  LR chi2(2)        =       6.38
                                                  Prob > chi2       =     0.0413
Log likelihood = -553.40712                       Pseudo R2         =     0.0057

------------------------------------------------------------------------------
      insure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Indemnity    |  (base outcome)
-------------+----------------------------------------------------------------
Prepaid      |
        male |    .477311   .1959283     2.44   0.015     .0932987    .8613234
       _cons |  -.1772065   .0968274    -1.83   0.067    -.3669847    .0125718
-------------+----------------------------------------------------------------
Uninsure     |
        male |     .46019   .3593233     1.28   0.200    -.2440708    1.164451
       _cons |  -1.989585   .1884768   -10.56   0.000    -2.358993   -1.620177
------------------------------------------------------------------------------
```

All of this output has exactly the same interpretation as the output we have seen previously, except that since we have three possible outcomes, so we have 2 linear predictors. Stata chose indemnity as the reference outcome (we will see how to change that shortly), and produced two linear predictors, one for comparing prepaid to indemnity, the other comparing no insurance to indemnity. In each case, the linear predictor takes the form

$$LP = \beta_0 + \beta_1 \times \texttt{male}$$

with β_0 taking the values -0.177 for the prepaid LP an -1.990 for the uninured LP. The coefficients β_1 take the values 0.477 and 0.460 respectively.

More often than not, these coefficients are not particularly meaningful, and we would prefer to see relative risk ratios. This can be achieved using the `rrr` option. We can also change the outcome used as the reference with the `baseoutcome()` option. For example, the command

```
mlogit insure male, rrr baseoutcome(3)
```

produces the following output:

```
. mlogit insure male, baseoutcome(3) rrr

Iteration 0:   log likelihood = -556.59502
Iteration 1:   log likelihood = -553.40794
Iteration 2:   log likelihood = -553.40712
Iteration 3:   log likelihood = -553.40712

Multinomial logistic regression                   Number of obs     =        616
                                                  LR chi2(2)        =       6.38
                                                  Prob > chi2       =     0.0413
Log likelihood = -553.40712                       Pseudo R2         =     0.0057

------------------------------------------------------------------------------
      insure |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Indemnity    |
        male |   .6311637   .2267918    -1.28   0.200      .312094    1.276435
       _cons |     7.3125   1.378237    10.56   0.000     5.053987    10.58029
-------------+----------------------------------------------------------------
Prepaid      |
        male |   1.017268   .3605507     0.05   0.961       .50786    2.037639
       _cons |      6.125   1.167805     9.51   0.000     4.215161    8.900163
-------------+----------------------------------------------------------------
Uninsure     |  (base outcome)
------------------------------------------------------------------------------
```

The likelihood ratio χ^2 in this case is 6.38 on 2 degrees of freedom, suggesting that there is a significant association between sex and insurance type. Howeve, the individual RRRs are not statistically significantly different from 1. That is because we have chosen to use Uninsured as our reference outcome, and this has the smallest numbers, and hence lowest power to detect differences. There is, in fact, a statistically significant difference between the RRRs for Indemnity and Prepaid, as we saw in the previous model, but our choice of reference category means that it is not automatically presented as part of the output. We can use `lincom` to recover it, as we will see in section 1.1.2.

`predict` Using `predict` after `mlogit` is slightly complicated by the fact that there are multiple linear predictors. You therefore need to either provide multiple variables for the predictions to be put into, or specify which one particular prediction you want (if you provide a single variable name, stata assumes you want outcome 1, which may or may not be useful). The most useful options to `predict` after `mlogit` are `p`, which gives the predicted probability for each outcome, and `xb`, which gives the linear predictor for each outcome (or 0 for the baseline outcome). For example, we the command

```
predict prob*, p
```

would create 3 new variables, called `prob1`, `prob2` and `prob3`, containing the predicted probabilities of belonging having each of the 3 types of insurance.

`lincom` The command `lincom` can also be uaed after the `mlogit` command, but again the multiple linear predictors make it more complicated. You cannot simple use a variable name (or `_cons`) to identify a coefficient, you also need to specify which linear predictor the coeffienct belongs to. We do this by putting the value of the relevant outcome in square brackets before the variable name. For example, suppose that we want to test whether there is a statistically significant difference between the coefficients for male in the prepaid and uninsured linear predictors. We can do this with the command

```
lincom [Prepaid]male - [Uninsure]male
```

and get the following output:

```
 ( 1)  [Prepaid]male - [Uninsure]male = 0

------------------------------------------------------------------------------
      insure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    .017121   .3544302     0.05   0.961    -.6775495    .7117916
------------------------------------------------------------------------------
```

As you might have expected, the difference between the relative risk ratios is not statistically significant.

1.2 Ordinal Variables

If the variable that you want to predict is ordinal, rather than nominal, there are a number of possible approaches.

1. Ignore the ordinal nature of the variable, treat it as nominal

2. Ignore the ordinal nature of the variable, treat it as interval

3. Respect the ordinal nature of the variable.

Options 1 and 2 may seem inappropriate if the data is ordinal, but both can be useful in certain circumstances.

Option 1 may be the best approach if different predictors are important at different levels of the outcome variable. For example, with an ordinal pain rating, rated as none, mild, moderate, severe, it may be that a particular variable reduces the probability of reporting “none”, but has no impact on the level of pain reported. In this case, the variable may be ordinal, but the association between predictors and outcome is not, and an ordinal regression model may not fit well.

Option 2 can be useful for ordinal data with lots of possible values, such as a visual analog scale. Technically, this is an ordinal variable, but it is usually appropriate to treat it as interval.

Option 2 is also commonly used when the ordinal variable is a *predictor*, rather than the outcome. It does make the assumption that each time the predictor goes up 1 category, the outcome goes up by an equal amount, which is quite a strong assumption. Ways of testing this assumption and including the predictor in the most appropriate way are outlined in section 1.2.2. Option 2 for predictor variables is also the basis of the “test for trend”, discussed in section 1.2.1.

1.2.1 Trend Test

We have seen that the χ^2-test can be used to test for an association betwen two categorical variables. However, it treats any deviation of the observed values from the expected values in the same way. It does not specifically test if observations in higher categories of one variable tend to be in higher (or lower) categories of the second variable. However, the χ^2 statistic can be broken down into two components, one of which measures the linear trend, and the other deviations around that trend.

As an example, consider the data in Table 1.2. This is looking to see if there is an association between the reading score (dichotomised as "High" or "Low") and the writing score (categorised into 4 ordinal levels, labelled 0, 1, 2, 3 for ease of interpretation when we get round to doing some regression).

Reading Score	Writing Score			
	Low		High	
0	18	(82%)	4	(28%)
1	42	(53%)	37	(47%)
2	10	(19%)	42	(81%)
3	4	(9%)	43	(91%)

Table 1.2: Association between reading and writing scores

The proportion of children with high writing scores increases as the reading score increases. The χ^2 statistic is 51.2 on 3 degrees of freedom, giving $p < 0.001$ and suggesting there is an association, but saying nothing about how the association works.

This χ^2 statistic can be decomposed into two parts, one testing the trend for the proportion to increase as the ordinal predictor increases and on testing for variation around this trend. The test is sometimes referred to as the Cochran-Armitage test, and it can be performed in stata with a user-written command `ptrend`:

```
  trend test]
. ptrendi 4 18 0 \ 37 42 1 \ 42 10 2 \ 43 4 3

     +------------------------+
     |  r   nr   _prop      x |
     |------------------------|
  1. |  4   18   0.182   0.00 |
  2. | 37   42   0.468   1.00 |
  3. | 42   10   0.808   2.00 |
  4. | 43    4   0.915   3.00 |
     +------------------------+

Trend analysis for proportions
-----------------------------

Regression of p = r/(r+nr) on x:

Slope =  .24784, std. error =  .03548, Z =   6.984

Overall chi2(3) =          51.222,  pr>chi2 = 0.0000
Chi2(1) for trend =        48.781,  pr>chi2 = 0.0000
Chi2(2) for departure =     2.441,  pr>chi2 = 0.2951
```

Here, the χ^2 test for trend is highly significant, the test for departures from a linear trend is non-significant.

This trend test appears regularly in the literature, and was developed by two highly respected statisticians. However, I would suggest that it is *never* the best analysis available. The reasoning behind it is perfectly sound, but if you look at the mathematics, it is exactly equivalent to performing a linear regression, and we have seen that linear regression with dichotomous outcomes is not a good idea. However, we can combine the idea of the Cochran-Armitage test with logistic regression as outlined in section 1.2.2.

1.2.2 Ordinal Predictors

We have seen that we can include a categorical predictor in a logistic regression model by putting an `i.` before the variable's name. If we create a variable called `oread` containing the ordinal reading score and a variable called `gwrite` containing the dichotomous writing score, we can perform a logistic regression.

```
. logistic gwrite i.oread

Logistic regression                               Number of obs     =        200
                                                  LR chi2(3)        =      55.25
                                                  Prob > chi2       =     0.0000
Log likelihood = -104.16821                       Pseudo R2         =     0.2096

------------------------------------------------------------------------------
      gwrite | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       oread |
          1  |   3.964286   2.366622     2.31   0.021     1.230296     12.7738
          2  |       18.9   12.38441     4.49   0.000     5.232435     68.2684
          3  |     48.375   36.80353     5.10   0.000     10.89005     214.888
             |
       _cons |   .2222222    .122838    -2.72   0.007     .0752087     .656609
------------------------------------------------------------------------------

. testparm i.oread

 ( 1)  [gwrite]1.oread = 0
 ( 2)  [gwrite]2.oread = 0
 ( 3)  [gwrite]3.oread = 0

           chi2(  3) =   40.22
         Prob > chi2 =    0.0000
```

Performing `testparm i.oread` will test whether there is any association between `oread` and `gwrite`. It is conceptually, but not mathematically, comparable to the overall χ^2-test.

To get a test for trend, we add `oread` as a continuous variable a to the model. We can still only have 3 coefficients for `oread` altogether, so one coefficient will need to be dropped. For this reason, it is important to add `oread` *before* `i.oread`, otherwise it will be dropped as the unidentifiable 4^{th} coefficient.

```
. logistic gwrite oread i.oread
note: 3.oread omitted because of collinearity

Logistic regression                             Number of obs     =        200
                                                LR chi2(3)        =      55.25
                                                Prob > chi2       =     0.0000
Log likelihood = -104.16821                     Pseudo R2         =     0.2096

------------------------------------------------------------------------------
      gwrite | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       oread |   3.643681   .9240332     5.10   0.000     2.216545    5.989686
             |
       oread |
          1  |   1.087989   .5068218     0.18   0.856     .4366229    2.711083
          2  |   1.423578   .7522192     0.67   0.504     .5053657    4.010112
          3  |          1  (omitted)
             |
       _cons |   .2222222    .122838    -2.72   0.007     .0752087     .656609
------------------------------------------------------------------------------

. testparm oread

 ( 1)  [gwrite]oread = 0

           chi2(  1) =   26.00
         Prob > chi2 =    0.0000

. testparm i.oread

 ( 1)  [gwrite]1.oread = 0
 ( 2)  [gwrite]2.oread = 0

           chi2(  2) =    0.49
         Prob > chi2 =    0.7844
```

Now, the `testparm oread` tests the linear trend effect of `oread` whilst the `testparm i.oread` tests the departures around the linear trend. In this case, the latter test is not significant, whilst the trend test is, so the best way to include `oread` in our model is as a continuous predictor.

It should be noted that the test for trend using logistic regression is testing linearity on the log-odds scale, where as `ptrend` tests for linearity on the probability scale. However, the important thing is whether the effect is linear *in the model you are using*, and for dichotomous outcomes, the appropriate model is logistic.

1.2.3 Ordinal Outcomes

There are a number of possible approaches when the outcome variable is ordinal. A simple crosstabulation can be used to calculate and ordinal odds ratio. This approach can then be extended to produce a model predicting the probaility of being in each outcome category, respecting the ordinal nature of the data in a number of different ways.

We will explore these methods by applying them to the data in Table 1.3. This compares two treatments, A and B, with the outcome being an ordinal variable with 4 levels: “Healed”, “Improved”, “No change” and “Worse”. On treatment A, most subjects are in the “Healed” and “Improved” categories, while on treatment B, most subjects are in the “No change” or “worse” categories, sugggesting that treatment A is better.

	Treatment A		Treatment B		Total	
Healed	12	(38%)	5	(16%)	17	(27%)
Improved	10	(31%)	8	(25%)	18	(28%)
No Change	4	(13%)	8	(25%)	12	(19%)
Worse	6	(19%)	11	(34%)	17	(27%)
Total	32	(100%)	32	(100%)	34	(100%)

Table 1.3: Ordinal Outcome Example

Crosstabulation with Ordinal outcomes

From this Table 1.3, we could calculate three odds ratios, by dichotomising the outcome in 3 ways: “Healed” vs “Improved”, “No Change” and “Worse”; “Healed” and “Improved” vs “No Change” and “Worse”; or “Healed” “Improved”, “No Change” vs “Worse”. These odds ratios can be calculated as in we saw for 2 by two tables:

$$OR_1 = \frac{(12)\times(8+8+11)}{5\times(10+4+6)} = 3.2 \tag{1.3}$$

$$OR_2 = \frac{(12+10)\times(8+11)}{(5+8)\times(4+6)} = 3.2 \tag{1.4}$$

$$OR_3 = \frac{(12+10+4)\times 11}{(5+8+8)\times 6} = 2.3 \tag{1.5}$$

This is as much as we can do with cross-tabulation. However, the 3 odds ratios above are not too dissimilar. Would it be reasonable to assume they are all estimates of the same population parameter ? And if so, what is our best estimate of the value of that parameter ?

Ordinal Regression

We can get an estimate of this parameter using ordinal logistic regression, sometimes referred to as ordered polytomous logistic regression. I’ll show how this works in Stata first, then explain what is happening in the background.

Suppose we have a variable `treatmentA` containing 1 for those on treatment A and 0 for those on treatment B, and a variable y containing 1 for “Worse”, 2 for “No Change”, 3 for “Improved” and 4 for “Healed”. We could fit an ordinal logistic regression with the command

```
ologit y treatment, or
```

(the `or` says that we want to see odds ratios rather then coefficients, since we don’t know what the coefficients might mean yet). The output we get is

```
Iteration 0:   log likelihood = -87.993692
Iteration 1:   log likelihood = -85.260015
Iteration 2:   log likelihood = -85.249205
Iteration 3:   log likelihood =   -85.2492

Ordered logistic regression                       Number of
                                                  LR chi2(1
                                                  Prob > ch
Log likelihood =   -85.2492                       Pseudo R2

---------------------------------------------------------------
           y | Odds Ratio   Std. Err.      z    P>|z|
-------------+-------------------------------------------------
  treatmentA |   2.932027   1.367426     2.31   0.021
-------------+-------------------------------------------------
       /cut1 |  -.5635603   .3435512                          -
       /cut2 |   .3179999   .3363157                          -
       /cut3 |   1.616945    .396272
---------------------------------------------------------------
```

`predict` By default, the command `predict` after `ologit` gives predicted probabilities for each outcome, exactly the same as it does after `mlogit`. There is now only a single linear predictor, so if you use the `xb` option, you only need to give a single new variable. However, if you want the probabilities of each possible outcome, you need to give as many variables as there are outcomes.

1.2.4 Alternatives

There are a number of alternatives to the ordered polytomous regression model for ordinal data[1, 2]. One approach is to assume that there is a normally distributed latent variable underlying the ordinal outcome, and that there are thresholds in this latent variable which define which category is manifest. This is the ordinal probit model, which can be fitted with `oprobit`.

Another approach to ordinal data is the Stereotype Regression model. This can be thought of as lying between the ordered polytomous model and the multinomial model, in that it allows variables to affect different transitions in different ways. If a variable has an effect on the transition from level 1 to level 2, but not on the transition from level 2 to level 3, a stereotype regression model is a useful way to model this. Stereotype regression models can be fitted with the command `slogit`.

2 Modelling Categorical Outcomes: Practical

2.1 Practical For Session 8: Categorical Outcomes

Datasets

The datasets that you will use in this practical can be accessed via http from within stata. However, the directory in which they are residing has a very long name, so you can save yourself some typing if you create a global macro for this directory. You can do this by entering

```
global basedir http://personalpages.manchester.ac.uk/staff/mark.lunt
global datadir $basedir/stats/8_categorical/data
```

(In theory, the global variable `datadir` could have been set with a single command, but fitting the necessary command on the page would have been tricky. Far easier to use two separate commands as shown above). If you wish to run the practical on a computer without internet access, you would need to:

1. Obtain copies of the necessary datasets

2. Place them in a directory on your computer

3. Define the global macro `$datadir` to point to this directory.

2.1.1 Binomial & Multinomial Logistic Regression

The data used for this section was collected as part of a survey of alligator food choices in 4 lakes in Florida. The largest contributor to the volume of the stomach contents was used as the outcome variable `food`, and the charactertics of the alligators are their length (dichotomised as $\leq$ 2.3m and $>$ 2.3m), their gender and which of the four lakes they were caught in.

1.1 Load the alligators data into stata with the command `use $datadir/alligators`, and familiarise yourself with the values used for each of the variables and their meanings with the command `label list`

1.2 Create a new variable `invertebrate` which takes the value 0 if the main food was fish, 1 if the main food was invertebrates and missing if the main food was anything else. This can be done with the command `gen invertebrate = food - 1 if food < 3`

1.3 Produce a cross-tabulation of food against length, with the command

```
tabulate invertebrate size, co
```

You should see that whilst fish and invertebrates are equally common in the smaller alligators, the larger ones are more likely to eat fish than invertebrates.

1.4 Obtain an odds ratio for the effect of size on the probability that the main food is either fish or invertebrates with

```
logistic invertebrate size
```

Is size a significant predictor of food choice ?

1.5 Now create another outcome variable which compares the probability that the main food is reptiles to the probability that the main food is fish with

```
gen reptile = (food == 3) if (food == 1) | (food == 3)
```

1.6 Obtain an odds ratio for the effect of size on the probability that the main food is either fish or reptiles with

```
logistic reptile size
```

Is size a significant predictor of this food choice ?

1.7 Now use `mlogit food size, rrr` to get the odds ratios for the effect of size on all food choices. Which food category is the comparison group ?

1.8 Check that the odds ratios for the invertebrate vs. fish and reptile vs. fish comparisons are the same as before.

1.9 Are larger alligators more likely to choose reptiles rather than invertebrates ? You can test this with

```
lincom [Reptile]size - [Invertebrate]size, eform
```

What is the odds ratio for size in this food choice ?

1.10 Generate a new variable to enable you to check this result using a single logistic regression model (`gen rep_inv = food == 3 if food == 3 | food == 2`). Perform the logistic regression with

```
logistic rep_inv size
```

Are the results the same as you got with `lincom` ?

1.11 Now we are going to look at the influence of the lakes on the food choices. Produce a table of main food choice against lake with

```
tabulate food lake, co chi2
```

Does the primary food differ between the 4 lakes ?

1.12 What proportion of alligators from Lake Hancock had invertebrates as their main food choice ?

1.13 How does this proportion compare to the other three lakes ?

1.14 Now fit a multinomial logistic regression model with

```
mlogit food i.lake, rrr
```

Look at the LR χ^2 statistic at the top: does this suggest that the primary food differs between the lakes ?

1.15 What is the odds ratio for preferring invertebrates to fish in lake Oklawaha compared to Lake Hancock ? Does this agree with what you saw in the table ?

1.16 Confirm your answer to the previous question by using the command `logistic invertebrate i.lake`

2.1.2 Using `mlogit`

This section uses the dataset `$datadir/politics`, which contains information on the effect of gender and race on political party identification.

1.17 Use `label list` to find out the meanings of the variables

1.18 Use `mlogit party race, rrr` to determine the effect of race on party affiliation. How does being black affect the odds of being a republican rather than a democrat ?

1.19 How does being black affect the odds of being an independent rather than a democrat ?

1.20 Use `tabulate party race, co` to confirm that your answers to the previous questions are sensible.

1.21 What is the odds ratio for being a republican rather than a democrat for women compared to men (use `mlogit party gender, rrr` to find out).

1.22 Fit a multinomial model in which party identification is predicted from both race and gender (`mlogit party race gender, rrr`).

1.23 Add the interaction between race and gender, to see if the race influence differs between men and women. Is this difference statistically significant ?

2.1.3 Ordinal Models

This section uses the data in `$datadir/housing`. This data concerns levels of satisfaction among tenants of different types of housing, according how much contact they have with other residents and how much influence they feel they have over the management of their housing.

1.24 Use `label list` to find out the meanings of the variables.

1.25 Does the degree of satisfaction depend on which type of housing the tenant lives in ? (Use `ologit satisfaction i.housing` to find out).

1.26 Of which type of housing are the tenants most satisfied ?

1.27 Test whether `influence` and `contact` are significant predictors of satisfaction

1.28 Create a multivariate model for predicting satisfaction from all of the variables that were significant univariately. Are these predictors all independently significant ? (You may need to use `testparm` for categorical predictors).

1.29 Does the effect of influence depend on which type of housing a subject lives in ? (Fit an interaction term and use `testparm` to test its significance).

Bibliography

[1] Ben G. Armstrong and Margaret Sloan. Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, 129(1):191–204, 1989.

[2] Sander Greenland. Alternative models for ordinal logistic regression. *Statistics in Medicine*, 13:1665–1677, 1994.