Andrew Noymer, UC Irvine

Handout for 221: Logistic regression and maximum likelihood

For the purposes of this exposition, we need a small dataset, because we will be performing some calculations by hand.

I created a data set that is a random sample ($N = 20$) of the `hire771.dta` data set [**`sample 20, count`** gives a random sample of size 20]. I re-coded the "`homeoffice`" variable to $\{0, 1\}$ (from $\{1, 2\}$); the new variable is called "HO".

Look at a cross-tab of sex and home-office:

```
           |          HO
       sex |         0          1 |     Total
-----------+----------------------+----------
         0 |         1          2 |         3
         1 |         2         15 |        17
-----------+----------------------+----------
     Total |         3         17 |        20
```

We know the odds ratio will be: $(1 \times 15) \div (2 \times 2) = 15/4 = 3.75$.

And, sure enough, the odds ratio is 3.75:

```
. logistic HO sex

Logistic regression                               Number of obs   =         20
                                                  LR chi2(1)      =       0.77
                                                  Prob > chi2     =     0.3789
Log likelihood =  -8.067122                       Pseudo R2       =     0.0458

------------------------------------------------------------------------------
          HO | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |       3.75   5.390937     0.92   0.358     .2240568    62.76309
------------------------------------------------------------------------------
```

We can see also using the coefficients:

[please see overleaf]

```
. logit HO sex

Iteration 0:   log likelihood = -8.4541818
Iteration 1:   log likelihood = -8.1221907
Iteration 2:   log likelihood = -8.0672083
Iteration 3:   log likelihood =  -8.067122

Logistic regression                                     Number of obs   =         20
                                                        LR chi2(1)      =       0.77
                                                        Prob > chi2     =     0.3789
Log likelihood =  -8.067122                             Pseudo R2       =     0.0458

------------------------------------------------------------------------------
          HO |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   1.321756   1.437583     0.92   0.358    -1.495855    4.139367
       _cons |   .6931472   1.224736     0.57   0.571    -1.707292    3.093586
------------------------------------------------------------------------------
```

The coefficient is 1.32. We can see that that works out to the correct odds ratio:

```
. disp exp(1.321756 )
3.7500006
```

How does the software get the α, β values? What is the meaning of the "likelihood" numbers? Why does it go through four iterations?

In OLS (ordinary least squares), or "linear regression", the data determine unique parameter estimates according to a formula. For example, for the simple case of one X variable and N data points, the formula is:

$$\hat{\beta} = \frac{\sum_{i=0}^{N} x_i y_i}{\sum_{i=0}^{N} x_i^2}$$

where $x = X - \overline{X}$ and $y = Y - \overline{Y}$. And that's that. If we know all the data points, the above formula gives us the least-squares $\hat{\beta}$ (and, of course, $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$ [why?]). There is a generalization for more than one X variable; the textbooks cover this. In any case, what you need to know is that Stata can produce the OLS estimates simply by running the data through a big formula.

For logistic regression, there is no formula that produces spot-on estimates. The estimates need to be refined. That is the answer to why it performs several iterations — it's refining estimates. In the absence of a formula like the one above, for OLS, we need some way to get α, β from the data. Maximum likelihood is that way.

Instead of the logic of OLS, which is: $\{\text{data}\} \implies \{\alpha, \beta\}$, maximum likelihood takes a different approach. It assumes the data are fixed and asks:

$$\text{which } \alpha, \beta \implies \text{data?}$$

Some black box — a probability model — is assumed to account for the data. Maximum likelihood asks, which $\{\alpha, \beta\}$ are most likely, given the data and the model? Starting with a candidate $\{\alpha, \beta\}$, the program searches for $\{\alpha, \beta\}$ combinations that are more likely to have produced the observed data. The computer stops when it finds the $\{\alpha, \beta\}$ that are most likely to have produced the data.

To put it another way, given that the data come from a model, what are the parameters of that model?

Maximum likelihood gives us, literally, the parameters that are most likely to have produced the data, *assuming the data are distributed according to the model*.

Consider a single data point, {sex=1,HO=1}, say. What is the likelihood of observing this point? What is likelihood, anyway? It is probability.

As we saw in class, in logistic regression:

$$p = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}$$

So for this particular point, {sex=1,HO=1}:

$$p(\text{HO} = 1) = \frac{\exp(\alpha + \beta)}{1 + \exp(\alpha + \beta)}.$$

So if we know $\{\alpha, \beta\}$, we can know the probability of observing this point. The probability of observing the whole data set is the product of all these individual probabilities. (For two *independent* events, A and B, the probability of observing them both, $p(A \cap B)$, is $p(A) \times p(B)$.)

Consider the following table, which represents the whole data set ($N = 20$):

(1) homeoff	(2) sex	(3) p(HO)	(4) p(not HO)	(5) p(observed)	(6) log(p(observed))
0	0		0.33333	0.33333	−1.09861
0	1		0.11765	0.11765	−2.14007
0	1		0.11765	0.11765	−2.14007
1	0	0.66667		0.66667	−0.40547
1	0	0.66667		0.66667	−0.40547
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
1	1	0.88235		0.88235	−0.12516
					$\sum = -8.06712$
α: 0.69					
β : 1.32					

Columns (1) and (2) are simply the data, sorted by HO and then by sex. There are 3 "failures" (2 female, one male) for home-office and 17 "successes" (15 female, two male). You may verify that this matches the 2×2 table at the start of this document.

Column (3) is the probability of "success". I.e. the probability of being in the home-office. I.e., $p(\text{HO} = 1) = \exp(\alpha + \beta)/(1 + \exp(\alpha + \beta))$. Verify. It is blank for failure. Column (4) is the probability of "failure", i.e. $p(\text{HO} = 0) = 1 - p(\text{HO} = 1) = 1 - [\exp(\alpha + \beta)/(1 + \exp(\alpha + \beta))]$. It is blank for success.

Column (5) is the probability of observing the data point in question, i.e. it is the concatenation of columns (3) and (4). Column (6) is the logarithm of column (5). The sum of a bunch of logs is the same as the log of the product: $\log(A \times B \times C) = \log(A) + \log(B) + \log(C)$.

So the sum of column (6) (at the bottom, -8.06712) is the log of the product of all the probabilities in column (5). Assuming independence, it is the probability of observing exactly these data in 20 trials, if α and β are as stated.

Verify that this equals the log likelihood stated on p. 1 or p. 2. A likelihood really is just a probability. I hope this removes some of the mystery behind maximum likelihood. Those numbers in the regression table are just the probability of observing the data, for a given choice of parameter estimates.

Why use log likelihood? Adding a lot of numbers (logs) is easier than multiplying them, and tends to less propagation of roundoff error in the computer, and the logged numbers are easier to read — as data sets get larger, the likelihoods get very small, and log scale is better.

If the likelihood is so low, isn't there a better estimate of α and β? No. Changing either coefficient would result in an even smaller likelihood. This brings is back to the very first question, how does the software estimate α and β? It starts with a candidate value (estimated usually from linear regression, but this level of detail is beyond our scope), and bounces around that value until it finds the maximum likelihood estimate. The "likelihood function" is the sum of column (6).

So what?

It never hurts to know from where your estimates come. Moreover — for one thing — this really underscores the importance of understanding the independence assumption. The probability rule $P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}) \cdot P(\mathrm{B})$ used in the likelihood calculations assumes that A and B (et seq.) are independent. Are your data?

It also gives us some insight as to why there is colinearity at the level of the single-observation (so to say) for logistic but not for OLS. And it never hurts to know why. If success or failure is predicted perfectly then the likelihood is 0 or 1. Well, $\log(0)$ is undefined... we have a problem. And $\log(1) = 0$, which implies that the observation adds nothing to the log-likelihood, so it doesn't matter. This is why you will from time to time get error messages from Stata such as:

```
note: _Isc0695_848 != 0 predicts failure perfectly
      _Isc0695_848 dropped and 5 obs not used

note: _Isc0695_864 != 0 predicts success perfectly
      _Isc0695_864 dropped and 2 obs not used
```