Maximum likelihood estimation
Logit regression (continued)

MAXIMUM LIKELIHOOD ESTIMATION

The philosophy behind maximum likelihood estimation is that we pick as our parameter estimates those that give us
the highest probability, under our model, of getting the observations we actually got.

Let's start with an example with which we are all familiar - coin tossing, where the results are

HEAD 1
or
TAIL 0

Suppose I told you we tossed a coin 10 times and we got 4 HEADs and 6 TAILs.

We know that, if P = probability of a head, the probability of this result is \( P(4\text{Heads}, 6\text{Tails}) = \binom{10}{4} p^4 (1-p)^6 \).

If we have 4 HEADS and 6 TAILS, what should we choose as our estimate of P?

The whole notion of maximum likelihood estimation is that we choose p to be the one that makes the probability of
getting our set of observations the largest possible: i.e. maximize \( P^4 (1-P)^6 \).

We refer to the expression for the probability of getting our set of observations as the *likelihood function*:

\[
\text{Like} = P^4 (1-P)^6.
\]

We then estimate P by choosing the value of P which makes Like as large as possible. Frequently, we work with the
log of the likelihood function because it is simpler mathematically.

There are two ways to find maximum likelihood estimates:

1. take the derivative of the likelihood function with respect to each parameter, set the resulting equations
equal to 0, and solve for the parameter estimates.

2. find the maximum using numeric search procedures.

We will illustrate the latter, because this is what most programs do.
Step 1. For $P = 0, .1, .2, \ldots, 1.0$ calculate Like and log Like

```stata
set obs 11
. gen p = (_n -1)/10
. gen like = (p^4) * (1-p)^6
. gen loglike = log(like)
. list

<table>
<thead>
<tr>
<th>p</th>
<th>like</th>
<th>loglike</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0000531</td>
<td>-9.842504</td>
</tr>
<tr>
<td>2</td>
<td>0.0004194</td>
<td>-7.776613</td>
</tr>
<tr>
<td>3</td>
<td>0.000953</td>
<td>-6.955941</td>
</tr>
<tr>
<td>4</td>
<td>0.0011944</td>
<td>-6.730117</td>
</tr>
<tr>
<td>5</td>
<td>0.0009766</td>
<td>-6.931472</td>
</tr>
<tr>
<td>6</td>
<td>0.0005308</td>
<td>-7.541047</td>
</tr>
<tr>
<td>7</td>
<td>0.000175</td>
<td>-8.650537</td>
</tr>
<tr>
<td>8</td>
<td>0.0000262</td>
<td>-10.5492</td>
</tr>
<tr>
<td>9</td>
<td>6.56e-07</td>
<td>-14.23695</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0.000000</td>
</tr>
</tbody>
</table>
```

What we can conclude from this is that it looks like the maximum is around $P = 0.4$. So I added lots more values close to 0.4 to see if the maximum is exactly 0.4 or something close to it.

Step 2: Drop like and loglike

```stata
input p = .35, .36, ..., .39, .41, .42 ..., .45
sort p
recompute like and loglike
```

We can see on the graph that the maximum is 0.4, but we can also see it if we list the values. We could also plot loglike and get the same form - if like is at its maximum, so is loglike. We then say $P = 0.4$ is the maximum likelihood estimate of $P$. 

![Graph of like vs p](image.png)
Let's start over again and see how we can get this likelihood function in a slightly different way.

For the first coin, the result is H or T. The probability of H is P; the probability of T is 1-P.

The probability of the 10 outcomes is

\[(\text{prob of outcome for case 1}) \times (\text{prob of outcome for case 2}) \times \ldots \times (\text{prob of outcome for case 10})\]

<table>
<thead>
<tr>
<th>Coin</th>
<th>Outcome</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T</td>
<td>1-P</td>
</tr>
<tr>
<td>2</td>
<td>H</td>
<td>P</td>
</tr>
<tr>
<td>3</td>
<td>T</td>
<td>1-P</td>
</tr>
<tr>
<td>4</td>
<td>T</td>
<td>1-P</td>
</tr>
<tr>
<td>5</td>
<td>T</td>
<td>1-P</td>
</tr>
<tr>
<td>6</td>
<td>H</td>
<td>P</td>
</tr>
<tr>
<td>7</td>
<td>H</td>
<td>P</td>
</tr>
<tr>
<td>8</td>
<td>T</td>
<td>1-P</td>
</tr>
<tr>
<td>9</td>
<td>H</td>
<td>P</td>
</tr>
<tr>
<td>10</td>
<td>T</td>
<td>1-P</td>
</tr>
</tbody>
</table>

Multiplying together, we get our likelihood function, as before.

**Application to Logistic Regression**

For logistic regression, we started with

\[L = \text{Logit } P = \ln \frac{P}{1-P} = \hat{\alpha}_0 + \hat{\alpha}_1 X_1 + \hat{\alpha}_2 X_2 + \hat{\alpha}_3 X_3 + \ldots + \hat{\alpha}_{K-1} X_{K-1}\]

where P is the probability that a given individual has outcome 1 - exactly equivalent to the coin example except now P can be different for every case - depending on the X values.

We also said last time that, in this case,

\[P = \frac{1}{1 + e^{-L}}.\]

We can also say that \[1-P = \frac{e^L}{1 + e^L}.\]

If we have the 10 individuals specified above and we know the outcomes for them and their X values, we can then find the likelihood function:
The likelihood function is the product of all of these. One more wrinkle:

\[ D^n_y \cdot P^y_i \cdot (1\cdot P)_i^{15y_i} \]

If the outcome, \( Y_2 = 1 \), then we get from this formulation the first term as the contribution of case 2, since \( 1-Y_2 = 0 \) and any number raised to the 0 power equals 1.

If we take logs, the log likelihood function becomes

\[
\log \text{like} = \sum (Y_i \log P_i + (1-Y_i) \log (1-P_i)) = \sum Y_i \log P_i + \sum (1-Y_i) \log (1-P_i).
\]

Once the likelihood or the log likelihood can be written down, there are methods for searching for the \( \alpha \)'s that maximize the likelihood. That's what we do in logit regression.

The search procedure is indicated by the iterations on the output.

. logit inter female black

Iteration 0:  Log Likelihood = -264.6267
Iteration 1:  Log Likelihood = -246.37474
Iteration 2:  Log Likelihood = -245.89753
Iteration 3:  Log Likelihood = -245.89738

Logit Estimates

| Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|-------|-----------|---|------|-----------------|
| female | -.6478314 | .2250376 | -2.879 | 0.004 | -1.090063, -0.2055996 |
| black  | 1.31346 | .2377786 | 5.524 | 0.000 | .8461905, 1.78073 |
| _cons  | -1.12112 | .1624118 | -6.903 | 0.000 | -1.440282, -0.8019567 |

Iteration 0 gives the log likelihood if \( P \) is the same for all individuals - i.e. if the model has intercept only. The program then does searches to find the \( \alpha \)'s that maximize the log likelihood. The table gives the best estimates.

It turns out that the -2 Log Likelihood (which is frequently called the likelihood statistics) has a chi-square distribution with degrees of freedom equal to \( n - K \), where, as before, \( n = \) sample size, \( K = \) number of parameters.
estimated.

Also, the difference in these likelihood statistics (large model statistic - small model statistic) follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters estimated.

In the STATA output, you find the chi-square value for the test of whether our model is better than the null model (with intercept only). In this case chi2(2) = 37.46, which is twice the difference between the log likelihood from iteration 1 (-264.627) and from iteration 3 (-245.897).

The probability of a chi-square value of 37.46 or larger with 2 degrees of freedom is also given and is small. We conclude that the model with Black and Female fits better than the null model.

GOODNESS OF FIT

We just concluded that our model fits better than the null model. There is no agreement on a good statistic that would be comparable to R^2 for multiple regression. One that is frequently used is the pseudo R^2 given by STATA, which is simply

\[ R^2 = \frac{\text{chi-sq stat}}{\text{chi-sq stat} + n} \]

where the chi-sq stat is the one we just discussed - that compares our model to the null model, and n is the sample size.

INFLUENCE STATISTICS

There are a whole series of influence statistics that can be used to see how much change dropping an observation would make in the estimates of the â’s and diagnostic graphs that may be useful. I refer you to Hamilton for these methods and will discuss them next time.
USE OF CASE-CONTROL STUDIES IN LOGIT REGRESSION

What follows is a comparison of the results of two logit regression analyses: 1) based on a random sample; 2) based on a case-control sample. The comparison illustrates that the case-control study estimates relative risks or relative odds well.

The following data set, which is from a study in Bangladesh. The data were based on a random sample. We can estimate a model using logit regression in which the probability of dying in a year depends on the child’s age at the start of the year (age 1 is the reference age), mother’s education, the presence and number of older female and male sibling, the sex of the child, and whether the child lived in an area where there was a health intervention program.

```
contains data from prob7-2.dta
Obs: 4434 (max= 4596)
Vars: 12 (max= 939)
Width: 49 (max= 968)
```

1. idchild str10 %10s Child ID
2. Progarea byte %8.0g in health plan area
3. Mothed byte %8.0g Mother has some education
4. FS1 byte %8.0g 1 older female sibling
5. FS2 byte %8.0g 2+ older female siblings
6. MS1 byte %8.0g 1 older male sibling
7. MS2 byte %8.0g 2+ older male siblings
8. Female byte %8.0g Index child female
9. Y2 byte %8.0g Age 2 at start of year
10. Y3 byte %8.0g Age 3 at start of year
11. Y4 byte %8.0g Age 4 at start of year
12. Died byte %8.0g Died during year

The first model considers only age, mother’s education, sex, and access to the health intervention as the predictors of child death.

```log
.logit Died Mothed Progarea Female Y2 Y3 Y4
```

Iteration 0:  Log Likelihood = -558.65643
Iteration 1:  Log Likelihood = -534.89626
Iteration 2:  Log Likelihood = -531.89359
Iteration 3:  Log Likelihood = -531.78721
Iteration 4:  Log Likelihood = -531.7867

Logit Estimates

|             | Coef.  | Std. Err. | t      | P>|t| | [95% Conf. Interval] |
|-------------|--------|-----------|--------|-----|---------------------|
| Died        |        |           |        |     |                     |
| Mothed      | -.4205 | .2164     | -1.94  | 0.052| -.8447 to .0037     |
| Progarea    | -.5532 | .1952     | -2.834 | 0.005| -.9359 to -.1705    |
| Female      | .6108  | .19115    | 3.195  | 0.001| .2360 to .9856      |
| Y2          | -.0847 | .21679    | -0.391| 0.696| -.5098 to .3402     |
| Y3          | -.5518 | .25357    | -2.176| 0.030| -1.049 to -.0547    |
Using the results of this logit regression, we can estimate the probability of dying for a child specific for the mother's education and residence and the child's age and gender.

But we can also treat this data set as if it were from a case-control study as follows:

Keep 90% of observations where the child died, 80% where the child survived. This strategy is equivalent to picking cases and controls, where the probability that a child gets into the study differs according to the outcome.

```
. set seed 32195
. gen rand=uniform()
. drop if rand>.9 & Died==1
   (14 observations deleted)
. drop if rand>.8 & Died==0
   (856 observations deleted)

. logit Died Mothed Progarea Female Y2 Y3 Y4
```

```
Iteration 0:  Log Likelihood =  -483.96967
Iteration 1:  Log Likelihood =  -463.16168
Iteration 2:  Log Likelihood =  -460.58692
Iteration 3:  Log Likelihood =  -460.49758
Iteration 4:  Log Likelihood =  -460.49717

Logit Estimates                                         Number of obs =   3564
                                                           chi2(6)       =  46.94
                                                           Prob > chi2   = 0.0000
                                                           Log Likelihood = -460.49717                             Pseudo R2     = 0.0485

```

| Died | Coef.  | Std. Err. |    t  | P>|t|  | [95% Conf. Interval] |
|------|--------|-----------|-------|------|-----------------------|
| Mothed | -0.3893636 | .2295023 | -1.697 | 0.090 | -.8393329 .0606058 |
| Progarea | -0.5178888 | .2069545 | -2.502 | 0.012 | -.9236502 -.1121274 |
| Female | 0.6114731 | .2034168 | 3.006 | 0.003 | .2126478 1.010298 |
| Y2    | -0.0856938 | .2289005 | -0.374 | 0.708 | -.5344832 .3630955 |
| Y3    | -0.9044317 | .274805 | -2.199 | 0.028 | -1.143223 -.0656404 |
| Y4    | -1.652398 | .4103384 | -4.027 | 0.000 | -2.45692 -.8478756 |
| _cons | -3.006877 | .2209988 | -14.058 | 0.000 | -3.340174 -.2788196 |

Nearly all the coefficients are the same as in the original model on the previous page, which was based on the full sample -- so the relative risks are estimated well. The constant is different --- therefore we cannot estimate the absolute risk for the reference group -- or for anyone.