Week 6: Maximum Likelihood Estimation

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2017
Outline

- An alternative way of estimating parameters
- Maximum likelihood estimation (MLE)
- Simple examples
- Adding explanatory variables
- Variance estimation
Bernoulli example

- Suppose that we know that the following numbers were simulated using a Bernoulli distribution:
  - 0 0 0 1 1 1 0 1 1 1
- We can denote by $y_1, y_2, \ldots, y_{10}$
- Recall that the pdf of a Bernoulli random variable is $f(x; p) = p^y(1 - p)^{1-y}$, where $y \in \{0, 1\}$. The probability of 1 is $p$ while the probability of 0 is $(1 - p)$
- We want to figure out what is the $\hat{p}$ we used to generate those numbers
- The probability of the first number will be given by $p^{y_1}(1 - p)^{1-y_1}$, the probability of the second by $p^{y_2}(1 - p)^{1-y_2}$ and so on...
- If we assume that the numbers above are independent, the joint probability of seeing all 10 numbers will be given by the multiplication of the individual probabilities
Bernoulli example

- We use the product symbol $\prod$. For example, $\prod_{i=1}^{2} x_i = x_1 \times x_2$
- So we can write the joint probability or the likelihood ($L$) of seeing those 10 numbers as:
  \[ L(p) = \prod_{i=1}^{10} p^{y_i} (1 - p)^{1-y_i} \]
- Remember that we are trying to find the $\hat{p}$ that was used to generate the 10 numbers
- In other words, we want to find the $p$ that maximizes the likelihood function $L(p)$ (we use $\hat{p}$ because that’s the optimal one)
- Said another way, we want to find the $\hat{p}$ that makes the joint likelihood of seeing those numbers as high as possible
- Sounds like calculus... We can take the derivative of $L(p)$ with respect to $p$ and set it to zero to find the optimal $\hat{p}$
Bernoulli example

Before, we take the log to simplify taking the derivative; the log function is a **monotonic transformation**, it won’t change the optimal \( \hat{p} \) value

We will use several properties of the log, in particular:

\[
\log(x^a y^b) = \log(x^a) + \log(y^b) = a \cdot \log(x) + b \cdot \log(y)
\]

The advantage of taking the log is that the multiplication becomes a summation. So now we have:

\[
\ln L(p) = \sum_{i=1}^{n} y_i \ln(p) + \sum_{i=1}^{n} (1 - y_i) \ln(1 - p)
\]

\[
\ln L(p) = n \bar{y} \ln(p) + (n - n \bar{y}) \ln(1 - p)
\]

This looks a lot easier; all we have to do is take \( \frac{d \ln(p)}{dp} \), set it to zero, and solve for \( p \) (I made it more general, \( n = 10 \))
Bernoulli example

\[
\frac{d \ln(p)}{dp} = \frac{n \bar{y}}{p} + \frac{(n - n \bar{y})}{(1 - p)} = 0
\]

- After solving, we'll find that \( \hat{p}(y_i) = \bar{y} = \sum_{i=1}^{n} \frac{y_i}{n} \)

- So that’s the MLE estimator. This is saying more or less the obvious: our best guess for the \( p \) that generated the data is the proportion of 1s, in this case \( p = 0.6 \)

- We will need to verify that our estimator satisfies the three basic properties of an estimator: bias, efficiency, and consistency (this will be in your exam)

- Note that we can plug in the optimal \( \hat{p} \) back into the \( \ln \) likelihood function:

\[
\ln L(\hat{p}) = n \bar{y} \ln(\hat{p}) + (n - n \bar{y}) \ln(1 - \hat{p}) = a, \text{ where } a \text{ will be a number that represents the highest likelihood we can achieve (we chose } \hat{p} \text{) that way}
\]
Example

Simulated 100 Bernoulli rvs with $p = 0.4$

```
set obs 100
gen bernie = uniform() < 0.4
sum bernie
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>bernie</td>
<td>100</td>
<td>0.46</td>
<td>0.5009083</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

* We just showed that $\hat{p}$ is 0.46

* Let’s get the highest value of the ln likelihood
* Plug in $\hat{p}$ and the other values

```
di 100*0.46*ln(0.46) + (100-100*0.46)*ln(1-0.46)
-68.994376
```

By the way, we just did a **logistic regression** with only a constant (no covariates). Convince yourself
Example

- We will use the logit command to model indicator variables, like whether a person died

logit bernie
Iteration 0:  log likelihood =  -68.994376
Iteration 1:  log likelihood =  -68.994376

Logistic regression
Number of obs =  100
LR chi2(0) =  -0.00
Prob > chi2 =  .
Log likelihood =  -68.994376  Pseudo R2 =  -0.0000

| bernie | Coef.  | Std. Err. | z   | P>|z|  | [95% Conf. Interval] |
|--------|--------|-----------|-----|------|---------------------|
| _cons  | -0.1603427 | .2006431 | -0.80 | 0.424 | -.5535959 - .2329106 |

di 1/(1+exp( .1603427 ))
.45999999

- Note that Log likelihood = -68.994376 matches what we found “by hand”; the coefficient is in the log-odds scale
- This is a model with no explanatory variables. We can easily make the parameter p be a linear function of predictors
A couple of things to note

- Statistical software maximizes the log likelihood numerically (also the \( \log \) likelihood because of numerical precision)
- The algorithm is given a starting value for some parameters (sometimes using the null model)
- Each iteration “improves” the maximization
- The second derivatives are also computed (we will see why in a sec)
- In many cases, we need to be mindful of the difference between the scale of estimation and the scale of interest
- Logit models report coefficients in the log odds scale
Let's plot the \(-\ln(L)\) function with respect to \(p\)

```plaintext
twoway function y = -(100*x*ln(x) + (100-100*x)*ln(1-x)), range(0 1) ///
xtitle("p") ytitle("-Ln P") saving(1100.gph, replace)
```

![Graph of the -\ln(L) function](image)
What about the precision of the estimate?

- There is some intuition in the plot above. The precision of the estimate \( \hat{p} \) can be measured by the curvature of the \( \ln L(\theta) \) function around its peak.
- A flatter curve has more uncertainty.
- The Fisher information function, \( I(\theta) \) formalizes that intuition:
  \[
  I(\theta) = -E\left[\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta}\right]
  \]
- It turns out that we can calculate \( \text{var}(\theta) \) using the inverse of \( I(\theta) \).
- For the Bernoulli, \( I(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})} \) (evaluated at \( \hat{p} \)).
- The variance is \( 1/I(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} \).
- Note something. Once we know \( \hat{p} \) we also know its variance.
Normal example

- What about if we do the same but now we have numbers like

90.46561  
105.1319  
117.5445  
102.7179  
102.7788  
107.6234  
94.87266  
95.48918  
75.63886  
87.40594

- I tell you that they were simulated from a normal distribution with parameters $\mu$ and $\sigma^2$. The numbers are independent. Your job is to come up with the best guess of the two parameters.

- Pretty much the same problem as with the Bernoulli example and we can solve it in exactly the same way.
Normal example

- As before, we know that formula for the pdf of a normal and because the observations are independent we multiply the densities:

\[ L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i-\mu)^2}{2\sigma^2}\right) \]

- Remember the rules of exponents, in particular \(e^a e^b = e^{a+b}\). We can write the likelihood as:

\[ L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2) \]

- As before, we can simplify the problem by taking the \(\ln\) to help us take the derivatives. But before:

- **Alert**: Perhaps you are wondering, why are we using the pdf of the normal if we know that the probability of one number is zero? Because we can think of the pdf as giving us the probability of \(y_i + d\) when \(d \to 0\)
Normal example

- After taking the ln, we have:
  \[ \ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \]
- All we have left is to take the derivative with respect to our two unknowns, \( \mu \) and \( \sigma^2 \) and set them to zero. Let’s start with \( \mu \):
  \[ \frac{\partial \ln(L(\mu, \sigma^2))}{\partial \mu} = 2 \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu) = 0 \]
- The above expression reduces to (I added the ^ to emphasize that’s the optimal):
  \[ \sum_{i=1}^{n} (y_i - \hat{\mu}) = 0 \]
- **Does it look familiar?** Replace \( \hat{\mu} \) by \( \hat{y}_i \). That’s exactly the same as the first order condition we saw when minimizing the sum of squares
- Solving, we find that \( \hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y} \). In other words, our best guess is just the mean of the numbers
Normal example

- We can also figure out the variance by taking the derivative with respect to $\sigma^2$
- We will find that $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{\mu})^2}{n}$
- If you remember the review lecture on probability and statistics, we know that this formula is biased. We need to divide by $(n - 1)$ instead.
- (What is the definition of bias?)
- This happens often in MLE but it is easy to correct for it
Normal example Stata

- We just figured out that the best guess is to calculate the sample mean and sample variance
- We can easily verify in Stata

```
clear
set seed 1234567
set obs 100
gen ynorm = rnormal(100, 10)
sum ynorm
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>ynorm</td>
<td>100</td>
<td>98.52294</td>
<td>10.03931</td>
<td>74.16368</td>
<td>123.5079</td>
</tr>
</tbody>
</table>

- The sum commands divides the sample variance by (n-1)
Linear regression

- What about if I told you that the number I generated is a linear function of one variable, say, $x_1$? In other words, I’m saying that the mean of the normal distribution is $\mu = \beta_0 + \beta_1 x_1$

- Now we want to find the parameters $\beta_0, \beta_1, \sigma^2$ that maximize the likelihood function. Once we know the optimal $\hat{\beta}_0, \hat{\beta}_1$ we find the optimal $\hat{\mu}$

- The likelihood function is now:
  $$L(\beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i})^2\right)$$

- The ln likelihood is:
  $$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i})^2$$
Linear regression

- If we take the derivatives with respect to $\beta_0$ and $\beta_1$ we will find exactly the **same first order conditions** as we did with OLS. For example, with respect to $\beta_1$:

$$\sum_{i=1}^{n} x_1(y_i - \beta_0 - \beta_1 x_1) = 0$$

- All the algebraic properties of OLS still hold true here.
- The MLE estimate of $\sigma^2$ will be biased but we divide by $(n-p-1)$ instead as we saw before.
- So what do we gain with MLE?
- **We do gain a lot in the understanding of linear regression**
What we get from MLE

■ 1) It is clear that we are modeling a conditional expectation function: \( E[Y|X] \)

■ 2) It is clear from the start of setting up the problem that we are assuming that \( Y \) distributes normal with mean given by \( \beta_0 + \beta_1 X \). This is equivalent to saying that the error is additive and distributes normal \( N(0, \sigma^2) \)

■ 3) It is clear that we assume that the observations are independent; otherwise, we can’t multiply the densities

■ 4) The value of the optimal log likelihood function gives us a measure of the goodness of fit, much like SSR (i.e. the explained part) did. By comparing the log likelihood of alternative models, we will test if the reduced model is adequate like we did with the F test

■ 5) The curvature of the log likelihood function provides information about the precision of the estimates
What we get from MLE

- 6) MLE is much more general than OLS. You will use MLE for logit, Probit, Poisson, mixture models...
- 7) Learning to model using likelihood ratio tests is more useful for more type of models
- 8) AIC and BIC to compare non-nested models are based on the log likelihood function
- Here is a more detailed proof:
Likelihood ratio test (LRT)

- The null $H_0$ is that the restricted (constrained) model is adequate
- The alternative $H_1$ is that the full (unconstrained) model is adequate
- The likelihood ratio test compares the log-likelihoods of both models and can be written as:
  \[
  LR = -2[L(RM) - L(FM)],
  \]
  where $L(RM)$ is the log-likelihood of the restricted model and $L(FM)$ that of the full model
- Under the null that the restricted model is adequate, the test statistics $LR$ distributes $\chi^2$ with degrees of freedom given by $df = df_{full} - df_{restricted}$; that is, the difference in degrees of freedom between the restricted and full models
Example

- Compare the likelihood and other criteria

```stata
qui reg colgpa
est sto m1
...
est table m1 m2 m3, star stat(r2 r2_a ll bic aic) b(%7.3f)
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsgpa</td>
<td>0.482***</td>
<td>0.459***</td>
<td></td>
</tr>
<tr>
<td>skipped</td>
<td></td>
<td>-0.077**</td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>3.057***</td>
<td>1.415***</td>
<td>1.579***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
</tr>
</thead>
<tbody>
<tr>
<td>r2</td>
<td>0.000</td>
<td>0.172</td>
<td>0.223</td>
</tr>
<tr>
<td>r2_a</td>
<td>0.000</td>
<td>0.166</td>
<td>0.211</td>
</tr>
<tr>
<td>ll</td>
<td>-60.257</td>
<td>-46.963</td>
<td>-42.493</td>
</tr>
<tr>
<td>bic</td>
<td>125.462</td>
<td>103.823</td>
<td>99.832</td>
</tr>
<tr>
<td>aic</td>
<td>122.513</td>
<td>97.925</td>
<td>90.985</td>
</tr>
</tbody>
</table>

legend: * p<0.05; ** p<0.01; *** p<0.001

- Note that the log likelihood (ll) gets larger for better fitting models; we will cover AIC and BIC later
Example

- LR tests

`lrtest m3 m2`
Likelihood-ratio test  LR chi2(1) = 8.94
(Assumption: m2 nested in m3)  Prob > chi2 = 0.0028

`. lrtest m3 m1`

Likelihood-ratio test  LR chi2(2) = 35.53
(Assumption: m1 nested in m3)  Prob > chi2 = 0.0000

- It seems logical that LRT and F-test comparing nested models should be equivalent (at least asymptotically)
LRT and F-tests

- Compare tests

```stata
qui reg colgpa
est sto m0
scalar l10 = e(ll)
reg colgpa male campus

Source | SS   df  MS
-------------+---------------------------------- F(2, 138) = 0.62
Model       | .171856209 2  .085928105 Prob > F = 0.5413
Residual    | 19.2342432 138 .139378574 R-squared = 0.0089
-------------+---------------------------------- Adj R-squared = -0.0055
Total       | 19.4060994 140 .138614996 Root MSE = .37333

est sto m1
scalar l11 = e(ll)

lrtest m0 m1
Likelihood-ratio test LR chi2(2) = 1.25
(Assumption: m0 nested in m1) Prob > chi2 = 0.5341

* By hand
di -2*[l10 - l11]
1.2542272```

- p-value of both 0.5341 (I chose bad predictors so p-values would be high)
Summary

- MLE is not more difficult than OLS
- The advantage of learning MLE is that it is by far the most general estimation method
- Learning the concept of log-likelihood and LRT will help us when modeling linear models, logistics, Probit, Poisson and many more
- AIC and BIC use the log-likelihood
- We are using the log-likelihood in a similar way we used SSR, although we did the F-test in terms of SSE but we know that \( \text{SST} = \text{SSE} + \text{SSR} \)
- Never forget the main lesson of MLE with a normal: **We are modeling the mean as a function of variables**
- Stata added a way to maximize likelihood without much programming. Type “help mlexp”
- See more examples under Code on my website: http://tinyurl.com/mcpberraillon