

1 General objective

1.1 Understand the relationship.

Suppose you think that, for all observations, $i=1,\dots,N$, there is a relationship

$$E(y_i) = f(x_i)$$

In words, “the expected value of y depends on x .” In the “usual OLS” case, the function f is linear, and we write this:

$$E(y_i) = b_0 + b_1x_i$$

We never get to observe the expected value directly, we always just observe “realizations” from the random process, so the more usual thing is

$$y_i = b_0 + b_1x_i + e_i$$

We directly observe x_i and y_i , but not the b 's or e_i .

1.2 The importance of the residual.

The error term, e_i , is never known, but rather, once we have a set of estimates of the b 's, we can calculate a predicted value, \hat{y}_i , and the difference between the observed value and the predicted value, known as a residual can be calculated. That residual value is sometimes called \hat{e}_i .

1.2.1 What's OLS:

The Ordinary Least Squares approach is just one way to try to estimate the parameters. It does so by minimizing the sum of the \hat{e}_i^2 , arriving at \hat{b}_0 and \hat{b}_1 that have “known properties”.

You could say OLS is maximizing

$$-\text{Sum of squares}$$

Maximum likelihood is just maximizing a different objective function.

2 Maximum Likelihood as an Alternative

2.1 Adopt a different objective

Suppose you instead want to maximize the probability of observing this sample.

The key words: **sample & probability**.

The probability of observing a sample of y 's: y_1, y_2, \dots, y_N , is found by multiplying their individual probabilities:

$$\text{Probability } (y_1, y_2, \dots, y_N) = p(y_1) * p(y_2) * p(y_3) * \dots * p(y_N)$$

If you give me enough information about your model of y_i , then I can figure how likely each observation is.

2.2 What details about y_i are needed?

The “generalized linear model” was developed in the 1980s and 1990s to deal with a variety of possible distributions for y . We aren't going into detail about it, but it is one useful way to introduce maximum likelihood analysis.

If you think y_i is Normal with a mean that depends on some coefficients $b' = (b_0, b_1, \dots, b_m)$ and variables $x_i = (x_{0i}, x_{1i}, \dots, x_{mi})$, then it is pretty common to combine your b 's and x 's in “some formula” like

$$g(x; b) = b_0 + b_1 x_{1i} + \dots + b_m x_{mi}$$

Then you might write

$$y_i = N(g(x_i; b), \sigma^2)$$

That is, the value of y_i is drawn from a normal distribution, and the average of those y_i 's, given x_i and b , is equal to $g(x; b)$. Because the expected value of a normal distribution is equal to the first parameter in it, some people write:

$$E(y_i|x) = g(x; b)$$

or, if there is a linear relationship:

$$E(y_i|x) = b_0 + b_1 x_{1i} + \dots + b_m x_{mi}$$

or if they use matrix algebra, they might write

$$E(y_i|x) = x_i b$$

or

$$E(y|x) = Xb$$

where X is a matrix that collects up data for all observations and y is a column vector of all observations.

See?

Anyway, the “generalized linear model” (GLM) is an effort to allow you to use a variety of distributions within this framework. Off hand, I can think of examples with a Poisson distribution, Gamma distribution, Binomial distribution, negative Binomial distribution.

In POLS 707 we concentrate almost exclusively on the Normal and Binomial case, but touch briefly on the Poisson.

2.3 Suppose you have an “error term”.

You might be used to thinking of regressions with error terms. So suppose you have some model with an error term, e_i . Suppose you know/assume that

- e_i has a given statistical distribution, such as Normal $(0, \sigma^2)$.
- Any other statistical distribution can be used, some are too mathematically tedious for practical modeling. Some models with the GLM perspective are hard to think of if you insist on having this separate error term interpretation.
- If you assume that the variable x_i is “fixed”, and the coefficients b_0 and b_1 are fixed, then the distribution of y_i is pretty easy to see. In the formula

$$y_i = b_0 + b_1x_i + e_i$$

then any variation in y_i is caused solely by the error term, once we have taken the b 's and x_i into account.

It is not necessary to assume a linear relationship, though. So you could just write

$$y_i = g(x_i; b) + e_i$$

- The probability of observing a particular value of y_i is the same as the probability of observing the error term that “goes along” with that value of y_i . Given values for b_0 , b_1 and x_i and y_i , we know the residual, so it is obvious that

$$p(y_i) = p(e_i = y_i - g(x_i))$$

I suppose to keep things “tidy”, I should emphasize that $f(x_i)$ takes into account the b 's, so some books write:

$$p(y_i) = p(e_i = y_i - g(x_i; b))$$

where the term b means “any coefficients you have in mind”.

If you tell me the error is normal, that means

$$p(y_i) = N(y_i - g(x_i; b), \sigma^2)$$

2.4 Maximize something

Since you give me a probability distribution for y_i , either because you stipulate some distribution for y itself or for the error term, then we can get down to business. With the information you give me, I can say how likely each y_i is to be observed. If e_i is Normal, then

$$p(y_1) = p(e_1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-e_1^2/2\sigma^2} \tag{1}$$

We've got a potential confusion here because this formula has “Euler's constant” e , approximately 2.7, as well as the error term, e_i . So pay attention.

The reasoning proceeds as follows:

- The probability of observing the whole sample is found by repeating the previous statement for each observation, and multiplying them all together. That yields some big, ugly formula, something like

$$p(y_1, y_2, \dots, y_N) = p(e_1) * p(e_2) * p(e_3) * \dots * p(e_N)$$

- Since $p()$ depends on the x 's and the b 's, then this is usually rewritten with them instead of e 's or y 's.

The likelihood function is the thing you want to maximize. You have to rewrite this probability value so it depends only on the parameters being estimated.

$$L(b, \sigma) = p(y|x, b, \sigma)$$

- Now, consider the linear regression problem. You want to adjust the coefficients b and σ^2 to make this model fit as closely as possible to the data. As you vary those parameters, the probability of obtaining the sample goes up and down. This process eventually maximizes the likelihood.
- The likelihood function is written by “thinking backwards”, taking the x_i and y_i as givens and adjusting the parameter estimates. If you assume the linear model

$$y_i = b_0 + b_1 x_i + e_i$$

then the probability of observing y_i is the same as the probability of observing

$$e_i = y_i - b_0 - b_1 x_i$$

And if e_i is normal, that means

$$p(y_i) = p(e_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_i - b_0 - b_1 x_i)^2 / 2\sigma^2} \quad (2)$$

If you did this for each observation, and multiplied them all together, you would have the likelihood function,

$$L(b_0, b_1, \sigma^2; y) = \prod_{i=1}^N p(e_i)$$

If you plug in the results from equation (2) into this formula over and over, you arrive at something I'm too lazy to type. But if you did, and then adjusted the b 's and σ^2 , you would maximize the likelihood.

- Since the above formula is a big string of things multiplied together, it is often recommended to take the logarithm of that formula, thus converting it into a sum:

$$\ln L(b_0, b_1, \sigma^2; y) = \sum_{i=1}^N \ln p(e_i)$$

That's why people talk about “log likelihood” so often.

Here's where the calculus kicks in.

You find the slope of the log likelihood as it depends on each parameter. Then set that equal to 0.

$$\partial l / \partial b = 0$$

You have one of those “conditions” for each b or other parameter being estimated. Those equations are called the “score equations”

3 Maximum likelihood properties

3.1 It is not always unbiased.

Sometimes MLE estimates are unbiased, but it cannot be proven that they always are unbiased.

3.2 Consistent.

MLE's are always consistent. As the sample size grows, the expected difference between the estimate and the true value declines.

3.3 Asymptotic normality.

As the sample size grows, the sampling distribution of an MLE tends toward a Normal distribution.

One important aspect of this result is that MLE's allow t-tests. We can calculate estimates of the standard error of the coefficients (using means discussed in various books—look for terms like “inverse of information matrix”).

There are other ways of testing hypotheses, such as a Wald test or a likelihood ratio test. Details on these will be encountered in the future.

They often write complicated looking things. If we are estimating a vector of parameters θ by an MLE $\hat{\theta}$ and the variance matrix of the testimates is Σ_{θ} then the estimate converges in probability to a multivariate normal dist:

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, \Sigma_{\theta})$$

3.4 Asymptotic efficiency.

The MLE estimator has the smallest variance (in its sampling distribution) when compared when the sample size exceeds “some” arbitrary size. Sometimes I've seen them say it has variance as low as any consistent estimators. Sometimes they say it achieves the Cramer-Rao lower bound.

4 Computational challenges/difficulties.

Computer algorithms to maximize the likelihood sometimes fail, may take a long long time, and might not give confidence a maximum has been found. Sometimes they may arrive at a “local maxima”