

Mgmt 469

Maximum Likelihood Estimation

In regression analysis, you estimate the parameters of an equation linking predictor variables X to a dependent variable Y . Your goal is to find the parameters – the regression coefficients – that *best* predict Y . This raises a question: how does one define *best*?

If you run ordinary least squares regression, you are implicitly assuming that the best predictions are those that minimize the sum of the squared residuals. There are other ways to define best, however, and statisticians have proposed many alternatives. One alternative is to estimate a “median regression.” This technique minimizes the sum of the absolute values of the residuals, rather than the squared values. This is a very useful approach if you are concerned about outlier values of Y . You can estimate median regression in Stata by typing **qreg** instead of **regress**

Another alternative that has many attractive properties involves “maximizing the likelihood function.”¹ Among these properties is the ability to estimate the coefficients of models that involve complex functional relationships, including nonlinear specifications of the LHS variable. Because many useful models involve such relationships, it is necessary to learn about maximum likelihood estimation.

Maximum Likelihood – An Intuitive Approach

Suppose that you are interested in predicting the scoring of Los Angeles Lakers’ guard Kobe Bryant. You believe that the statistical model generating Bryant’s scoring is:

$$(1) \quad S_i = M + \varepsilon_i,$$

where S_i is his score for any given night i , M is his predicted score, and ε_i is a random component that varies from night to night.

Equation (1) represents a simple statistical process -- it combines a *deterministic* part (M) and *random* part (ε_i) that together generate an *outcome* (S_i). You want to figure out the value of M that will give you the best predictions, recognizing that your predictions will rarely be exactly correct, due to random chance.

¹If the OLS assumptions (homoscedasticity, independent errors, normally distributed errors) are satisfied, then the OLS coefficients of a linear regression model will be identical to those that maximize the likelihood function.

To keep things simple, suppose you will base your predictions on Bryant's performance in his previous five games: 33, 22, 25, 40, 30. What should you predict for the next game?

- It makes sense to compute his average score – in this case 30 points – and use this for the prediction.
- We already learned that this prediction will minimize the SSE.

Here is how MLE generates the same prediction:

- Suppose that you believe Bryant's scoring follows the process described above; that is, $S_i = M + \varepsilon_i$.
- Suppose that ε_i is distributed normally with mean 0.
- Thus, if you believe that $M = 32$, you would predict $S = 32 + 0 = 32$
- Suppose you also believe that Bryant does not score the same amount every night. In particular suppose that the standard deviation of $\varepsilon_i = \sigma$.² (You can use the data to estimate the value of σ .)

To summarize, you have the following information:

- You expect Bryant to score M
- His actual score = $M + \varepsilon_i$
- ε_i is a random variable with mean 0 and standard deviation of σ .

You can now calculate *the probability of Bryant scoring the amounts he actually scored*. Let $f(\varepsilon)$ denote the density function for ε . (Recall that the density function is like a probability function, and that the density for a normal variable is a bell curve with its maximum at $\varepsilon=0$.)

- Given the prediction M and the density function, you can compute the probability of Bryant scoring any particular point total Y . This is given by the formula $f(Y-M) = f(\varepsilon)$.
- For example, if you believe that $M=32$, then the probability that Bryant scores 35 is given by $f(35-32) = f(3)$.
- If $\sigma=6$, for example, then examination of the normal table reveals $f(3) = 0.8$.

²The computer would determine the actual standard deviation from the data.

Assume that Bryant's scoring in one game is independent of what he scored in the prior game.

- Recall that the probability of two independent events occurring is just the product of the probability that each occurs.

- *It follows that the probability, or **likelihood**, of Bryant scoring exactly 33, 22, 25, 40, and 30 points is just the product of the probabilities of his getting each of these scores.*

Given any prediction M , you can write the *likelihood score* as:

$$\text{Likelihood score} = L = f(33-M) \cdot f(22-M) \cdot f(25-M) \cdot f(40-M) \cdot f(30-M).$$

You want to find "maximum likelihood estimator" (MLE) of M

- *This is the value of M that maximizes L .*

- Intuitively, you know that the MLE of M would not be 15 or 50 or some number far from his typical scoring output. It is almost impossible that a player who is predicted to score 15 points per game would actually score 33, 22, 25, 40, and 30.

- In fact, if $M = 15$ and $\sigma = 6$, then $L = f(33-15) \cdot f(22-15) \cdot f(25-15) \cdot f(40-15) \cdot f(30-15)$
 $= f(18) \cdot f(7) \cdot f(10) \cdot f(25) \cdot f(15) < .0000001$

- But 32 might be a good candidate to be the MLE. Someone predicted to score 32 points per game has a reasonable chance of scoring 33, 22, 25, 40, and 30.

- In this case, $L = f(1) \cdot f(-10) \cdot f(-7) \cdot f(8) \cdot f(-2) \approx .00005$

- It turns out that MLE estimate of M is given by the mean of the realized values of Y . That is, $M = 30$ and $L = .00014$

A Formal Definition of the MLE

Instead of minimizing SSE, maximum likelihood estimation maximizes the likelihood score. Here is how it works. (Don't worry; the computer will do the grunt work for you.)

- You specify a model you wish to estimate. Let's work with a simple model:

$$Y = \underline{B} \cdot \underline{X} + \varepsilon$$

- You make an assumption about the distribution of the errors. For example, you might suppose that $f(\varepsilon)$ is normal with mean = 0 and standard deviation = σ .
- You compute $f(\varepsilon) = f(Y - \underline{B} \cdot \underline{X})$ for each observation. This is the probability of observing Y , given \underline{X} and estimated coefficients \underline{B} .
- If you have done a good job of estimating the B 's, then ε should be small and $f(\varepsilon)$ should be large. (The probability of observing Y should be high.)

Assuming that the observations (and ε 's) are independent, you can compute the probability of observing all the values of Y in the data, given the X 's and the estimated B 's.

- For example, the probability of obtaining Y_1 and Y_2 given \underline{X}_1 , \underline{X}_2 and the estimated B 's is given by: $L = f(Y_1 - \underline{B} \cdot \underline{X}_1) \cdot f(Y_2 - \underline{B} \cdot \underline{X}_2) = f(\varepsilon_1) \cdot f(\varepsilon_2)$.
- In a regression with N observations, the likelihood score is:

$$L = \text{Likelihood score} = f(\varepsilon_1) \cdot f(\varepsilon_2) \cdot \dots \cdot f(\varepsilon_N).$$

The goal is to find values of \underline{B} that maximize L and thereby give the "best" overall predictions.

Note: the value of L gets very small very quickly, and depends on the number of observations.

- For example, if you have 10 observations and your model is so accurate that $f(\varepsilon) = .75$ for each observation, then $L = .75^{10} = .0563$.
- The same model applied to 20 observations with the same accuracy generates $L = .75^{20} = .0032$. The model is just as good, but the likelihood score is much smaller. This is only because you have more observations.
- For this reason, it is never appropriate to compare L across models unless you have the same number of observations.

The Log Likelihood Score

Likelihood scores can get very small.

- If $f(\epsilon) = .75$ and $n = 200$, then $L = 1.03e^{-25}$.
- To avoid dealing with such small numbers, researchers tend to report $\log(L)$. N
- In the example $\log(.0032) = -5.75$ and $\log(1.03e^{-25}) = -57.54$.
- Note that maximizing $\log(L)$ is the same as maximizing L . In general, statistical software will maximize and report $\log(L)$, also called the *log likelihood function*. The log likelihood function is a much more manageable number.
- Be careful: the log-likelihood is always a negative number (do you know why?). You maximize the log-likelihood by finding the value that is closest to zero.

Finding the MLE estimates of \underline{B}

For processes where MLE is required, the computer solves for \underline{B} by smart “trial and error”!

- Sophisticated algorithms allow computers to quickly converge on the values of \underline{B} that maximize L . Stata is preprogrammed with a very efficient MLE algorithm.

Here is how your computer finds the MLE estimates of the \underline{B} 's:

- 1) You specify a statistical process that relates a dependent variable Y to some predictors \underline{X} through some formula such as $Y = f(\underline{B}\underline{X} + \epsilon)$. Examples that we will use in this class include the Logit and Poisson processes. In medicine and engineering, hazard processes are commonly estimated.
- 2) The computer picks some "starting values" for \underline{B} and computes $\underline{B}\underline{X}$.
- 3) The computer then computes ϵ and $f(\epsilon)$ for each observation and computes the likelihood score L .
- 4) The computer examines how L would change as each coefficient B changes, and tries new values of \underline{B} that will improve the likelihood score.
- 5) It recomputes the likelihood score as in (3). (As you repeat this exercise the value of L will increase by smaller and smaller amounts.)
- 6) The computer repeats (4) and (5) until there are no material improvements in L .

The resulting values of \underline{B} are the MLE estimators of these parameters. They will give you the best estimates of the model that relates \underline{X} to Y . Your regression software will report the MLE estimates of \underline{B} as well as estimated standard errors and the log likelihood score.