Up to this point in the course, we have focused on using the ordinary least squares model. OLS estimations minimize the squared differences between the actual points and a line that passes through them. To be valid, the five Gauss-Markov conditions must be met. When one or more of the conditions are not met, we have generally found “corrections” that “repair” the problem. However, there are some types of Gauss-Markov failures that cannot be corrected within the OLS framework. The most common failure that cannot be fixed in OLS is when the assumption of normally distributed error term cannot be maintained. Regressions on qualitative dependent variables are one prominent example of this problem. Another is when there is “truncation” or “censoring” of the error term. Truncation and censoring occur when by nature or by policy the dependent variable cannot assume any value from negative to positive infinity (at least in theory). In the face of these failures, we must seek a different form of estimation: maximum likelihood estimation (MLE). As we shall see, MLE is an extraordinarily flexible framework. It will also turn out that OLS is a form of maximum likelihood estimation.

**The Linear Model and OLS**

Recall first the standard linear model:

\[ y_i = X_i \beta + \varepsilon_i \]

where \( \varepsilon_i \) is a disturbance term and \( X_i \beta \) is shorthand for:

\[ X_i \beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k. \]

It is important to keep in mind that this model simply says that the expected value of \( Y \) depends on \( X \) so that:

\[ E(y_i|X_i) = X_i \beta \]

and that the variance of \( y_i \), given \( X_i \), is:

\[ \text{Var}(y_i|X_i) = \text{Var}(\varepsilon_i) = \sigma^2 \]

This is the variance of the disturbance term. In other words, the variance of \( y \) given \( X \) is simply the variance of the disturbance term given \( X \). The empirical counterpart is the residual variance, which is the variance around the estimated regression line. If we assume that the disturbance term is normally distributed with the same variance for each \( y \), we can describe this model by:

\[ (y_i|X_i) \sim N(X_i \beta, \sigma^2). \]
The solid line in the figure just below illustrates the expected value of \( y \) given \( X \), that is \( \text{E}(y|X) = X\beta \). The disturbance term is assumed to be normally distributed with a variance equal to 4 and the sample values (the dots) were randomly selected from this distribution. The variation of the dots around the expected value of \( y \) represents the variance of \( y \) given \( X \). The estimated regression line is the dashed line. The actual expected value of \( y \) given \( X \) is \( \text{E}(y|X) = 0 + 1X \). The estimated regression line is given by \( y = -0.4939 + 1.1050X \), which differs somewhat from the true “population” relationship between \( X \) and \( y \).

The ordinary least squares estimate of \( \beta \) is found by minimizing the sum of squared residuals and is given by:

\[
\text{Min } \sum_{i=1}^{N} (e_i)^2 = \sum_{i=1}^{N} (y_i - X_i \beta)^2
\]

**The Maximum Likelihood Method**

The linear model as set out above can be estimated by another method called maximum likelihood estimation (MLE). But before setting out this method for the general model above, we will consider a simpler example.

Recall that we are often concerned that estimators (e.g., the estimated mean, coefficients in a regression, etc.) have particular properties, in particular that they be unbiased and have small variance. (Here the reference is to the variance of an estimate, not the residual variance as mentioned above.) In this respect, maximum likelihood
estimates are very attractive. While often biased in small samples, they become unbiased as the number of data points used in their calculation increases. Furthermore, these estimators have the smallest possible variance. More formally, they attain what is known as the Cramer-Rao lower bound. No other possible estimator has a variance smaller than the Cramer-Rao lower bound, and thus no estimator has a variance smaller than that of the maximum likelihood estimator.

While maximum likelihood estimation is a powerful technique, it does require important a priori information about the problem. While maximum likelihood always gives us the "best" estimator, we can only use the technique if we know, or are willing to assume, the distribution followed by the observed data. This is a strong requirement, but one that we may be comfortable with in a wide variety of situations. And, in many cases involving non-linear models (such as qualitative choice specifications, which cannot be estimated using ordinary least squares linear models), we have no choice but to assume that the data follow a particular distribution.

Now, we will discuss the basic ideas behind maximum likelihood estimation and then turn to calculating maximum likelihood estimators in practice.

The basic principal behind maximum likelihood estimation is simple. Suppose that we observe a number of random draws from a known algebraic distribution with an unknown parameter. This parameter might be the mean $\mu$ of a normal distribution, or the parameter $\lambda$ of the Poisson distribution, or a $\beta$ from a multiple regression. We wish to use the data, and the knowledge of the underlying algebraic distribution, to estimate the unknown parameter. Maximum likelihood procedures estimate an unknown population parameter (i.e., $\mu$, $\lambda$, $\beta$, etc.) with the value of that parameter that is most likely to have generated the observed sample. Let’s elaborate on this last, somewhat confusing statement.

Consider a very simple example: standardized test scores. Let’s assume the scores are drawn from a normal distribution with unknown mean and standard deviation 200. We pick a single test score at random from the group of examinees and wish to estimate the mean of the underlying normal distribution from which we are drawing. Suppose the score chosen at random is 1200. What is our maximum likelihood estimate of the population parameter $\mu$? It is simply 1200 because, based on the rather limited sample at our disposal, the most likely distribution to have generated the observation at 1200 is a distribution with mean 1200. As shown in the figure below, a normal distribution with mean 1200 is more likely (in fact, eleven times more likely) to have generated the observation at 1200 than a normal distribution with mean 800 or mean 1600. And working with a sample of one, we can construct no estimator which is in any sense “better” than this maximum likelihood estimator.
Now consider the (slightly) more reasonable case of picking more than a single observation to use in calculating a maximum likelihood estimator. To do this, we will need to return to the probability distribution function (PDF) for a normal distribution. The PDF for a distribution with mean \( \mu \) and variance 200 is:

\[
f(y) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y - \mu)^2}{2\sigma^2}}
\]

\[
= \frac{1}{200 \sqrt{2\pi}} e^{-\frac{(y - \mu)^2}{2*200^2}}
\]

Suppose we draw three examinees from the group with scores \( y_1, y_2 \) and \( y_3 \). Since these are independent draws, the joint PDF for the three draws is simply the product of the PDF’s for each individual draw:
Notice what we have written is a function which gives the probability of observing the three observations we drew, $y_1$, $y_2$ and $y_3$, as a function of the unknown population parameter $\mu$. Such a joint probability function for a group of observed values is referred to as a "likelihood function" and is usually denoted $\mathcal{L}$ (in the equations, I will use "L" since Microsoft Equation 3.0 does not support "\mathcal{L}"). To find the maximum likelihood estimator, or MLE, we need only to maximize this expression over all possible values of $\mu$. In other words, we find the $\mu$ which maximizes the probability that we observe the particular sample we have in fact drawn. This is qualitatively no different than what we showed graphically while working with only a single observation.

To implement this maximization, we take advantage of the fact that the logarithm of a function is what is called a monotone transformation. This means simply that if A is greater than B, log(A) is greater than log(B). In this case, we will use the fact to argue that maximizing the logarithm of the likelihood function, or log likelihood, gives the same answer as maximizing the likelihood function itself. But the mathematical problem of maximizing a log likelihood is much easier to solve. This results directly from the property that the logarithm of a product is the sum of two factors:

$$\ln(L) = \ln(L_1 L_2) = \ln(L_1) + \ln(L_2)$$

So we wish to maximize:

$$L = \prod_{i=1}^{3} \left( \frac{1}{200\sqrt{2\pi}} e^{-\frac{(y_i-\mu)^2}{2*200^2}} \right)$$

We take the derivative of this rather awful looking expression with regard to the parameter $\mu$:

$$\frac{d\ln}{d\mu} = \frac{y_1-\mu}{200^2} + \frac{y_2-\mu}{200^2} + \frac{y_3-\mu}{200^2}$$

Setting this expression equal to zero gives the first order condition for a maximum:

$$y_1-\mu + y_2 - \mu + y_3 - \mu = 0$$
Solving this expression for the maximum likelihood estimator, which we denote here as \( \hat{\mu} \):

\[
3 \hat{\mu} = y_1 + y_2 + y_3
\]

\[
\hat{\mu} = \frac{y_1 + y_2 + y_3}{3} = \frac{\sum_{i} y_i}{3} = \bar{Y}
\]

So the maximum likelihood estimator for this problem is simply the sample mean. After crawling through all this notation, we still end up with our usual choice of estimator for the population mean, the sample mean. But we can now say that this is not only the obvious, but also the best estimator, at least as the sample size grows large.

In the above example, we assumed that the variance of the normal distribution was known. That assumption is not necessary, although it helps simplify the algebra. We could have written a likelihood function in terms of two unknown parameter, \( \mu \) and \( \sigma \), and maximized the likelihood function over both of these.

**An Exponential Example**

Consider another example of maximum likelihood estimation, this time involving the exponential distribution. This probability distribution is often used to model the waiting time between consecutive events such as automobile accidents, presidential vetoes, and patent applications. The PDF of an exponential random variable is given by:

\[
f(y) = \frac{1}{\theta} e^{-\frac{y}{\theta}}
\]

Suppose we wish to estimate the parameter \( \theta \) -- the mean of the distribution -- using maximum likelihood and three data points: \( y_1 = 2.0, y_2 = 4.0 \) and \( y_3 = 3.6 \). Obviously, in a real application we would want more than three points, but here an unrealistically small number makes it possible to show what is going on graphically, and leaves the basic story unchanged.

The figure below shows the P.D.F.’s for several exponential distributions with varying values for the parameter \( \theta \). The maximum likelihood procedure searches for the value of \( \theta \) which is most likely to generate the three observed data points. Notice that the two points to the right are more likely to be drawn if \( \theta \) is larger:
Now that we understand the basic problem graphically, we can proceed analytically. First, we write the likelihood function, or joint PDF for the three draws from the exponential with unknown parameter $\theta$:

$$ L = F(y_1, y_2, y_3) $$

$$ = \frac{1}{\theta} e^{\frac{-y_1}{\theta}} \cdot \frac{1}{\theta} e^{\frac{-y_2}{\theta}} \cdot \frac{1}{\theta} e^{\frac{-y_3}{\theta}} $$

$$ = \left( \frac{1}{\theta} \right)^3 \cdot e^{\frac{-y_1 - y_2 - y_3}{\theta}} $$

Next, we maximize this likelihood function with respect to $\theta$, the parameter we wish to estimate. But first, taking the log of the likelihood function makes the algebra easier:

$$ \ln L = -3 \ln \theta - \frac{y_1 + y_2 + y_3}{\theta} $$

Taking the derivative of the likelihood function with respect to $\theta$:

$$ \frac{d \ln L}{d \theta} = -3 \frac{\theta}{\theta^2} + \frac{y_1 + y_2 + y_3}{\theta^2} $$

Now set this expression equal to zero to find a first order condition which can be solved for the maximum likelihood estimator $\theta$: 
0 = -\frac{3}{\hat{\theta}} + \frac{y_1 + y_2 + y_3}{\hat{\theta}^2}

\frac{3}{\hat{\theta}} = \frac{y_1 + y_2 + y_3}{\hat{\theta}^2}

3\hat{\theta} = y_1 + y_2 + y_3

\hat{\theta} = \frac{\frac{y_1 + y_2 + y_3}{3}}{\bar{Y} = 3.2}

Once again, we find that the sample mean is the maximum likelihood estimator for \( \theta \).

**The General Linear Model**

As it turns out, the least squares estimator is in fact a maximum likelihood estimator. Remember that we presented regression as a way of estimating the expected value of a (random) variable \( Y \), when the expected value varies with another variable \( X \). We described the relationship as:

\[ E(y_i | X_i) = X_i \beta \]

Just above, we derived the MLE of the mean \( \mu \) of a normally distributed random variable \( Y \). We wrote the PDF of each realization \( y_i \) as:

\[ f(y_i) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - \mu)^2}{2\sigma^2}} \]

Maximum likelihood estimation of the general linear model can easily be understood if we replace \( \mu \) with \( X_i \beta \) and we assume that each \( y_i \) is distributed normally and conditional on \( X_i \) with mean \( X_i \beta \) and variance \( \sigma^2 \). As above, the model can be written in standard statistical notation as \( (y_i | X_i) \sim N(X_i \beta, \sigma^2) \). Then the PDF of \( y_i \) conditional upon \( X_i \) can be written:

\[ f(y_i | X_i) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - X_i \beta)^2}{2\sigma^2}} \]

Once we have the correct PDF, we can write the likelihood function for a sample of \( N \) independent realizations of \( X \) and \( Y \):
\[ L = f(y_1|X_1, y_2|X_2, \ldots, y_N|X_N) \]
\[ = f(y_1|X_1)^* f(y_2|X_2)^* \ldots f(y_N|X_N)^* \]
\[ = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^N \frac{\sum_{i=1}^{N} (y_i - X_i \hat{\beta})^2}{2\sigma^2} \]

Now, we take the logarithm:

\[ \ln L = -N \ln(\sqrt{2\pi}\sigma) - \frac{\sum_{i=1}^{N} (y_i - X_i \hat{\beta})^2}{2\sigma^2} \]

The likelihood function is maximized with respect to \( \beta \) and \( \sigma \). Notice that maximizing this expression over possible values of the maximum likelihood estimator for \( \hat{\beta} \) involves picking the \( \hat{\beta} \) that minimizes the expression \((y_i - X_i \hat{\beta})^2\) summed over all observations. This is, of course, just the familiar criteria we've already used in defining the least squares estimator above. So, least squares regression estimates and maximum likelihood estimates for \( \beta \) are the same, assuming that the distribution of \( Y \) given \( X \) is normal (or, equivalently, that the disturbance term \( \varepsilon \) is normally distributed).

In practice, both \( \beta \) and \( \sigma \) would be estimated. The values would be found by taking the derivative of \( \ln L \) with respect to \( \beta \) and \( \sigma \), setting the results equal to zero, and solving for the estimated values of \( \beta \) and \( \sigma \). To simplify, suppose that there is only one \( X \). Then

\[ \frac{d \ln L}{d\beta} = \frac{1}{\sigma^2} \sum_{i=1}^{N} [y_i X_i - \hat{\beta} X_i^2] \]

Setting this expression equal to zero and solving for \( \beta \) we find

\[ \hat{\beta} = \frac{\sum_{i=1}^{N} y_i X_i}{\sum_{i=1}^{N} X_i^2} \]

which is in fact the least squares estimate. Similarly, it can be shown that the MLE estimate of \( \sigma^2 \) is given by

\[ \hat{\sigma}^2 = \frac{\sum_{i=1}^{N} (y_i - X_i \hat{\beta})^2}{N} \]

which in large samples is the same as the least squares estimate. (Recall that the least squares estimate divides by \( N - k \), where \( k \) is the number of parameters.)