Under Assumption 5 we have specified the distribution of the error, so we can estimate the model parameters $\theta = (\beta, \sigma^2)$ with the principle of maximum likelihood. Under the assumption that the error is Gaussian, we will see that the OLS estimator $B$ is equivalent to the MLE and the OLS estimator of $\sigma^2$ differs only slightly from its ML counterpart. Further, $B$ achieves the Cramer-Rao lower bound.

**ML Principle**

The intuitive idea of the ML principle is to choose the value of the parameter that is most likely to have generated the data. Precisely, we assume that the probability distribution of a sample $(Y)$ is a member of a family of functions indexed by $\alpha$ (this is described as parameterizing the distribution). This function, viewed as a function of the parameter vector $\alpha$ is called the likelihood function.

In general, the likelihood function has the form of the joint density function

$$L(\alpha|Y_1 = y_1, \ldots, Y_n = y_n) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \alpha).$$

For an i.i.d. sample of a continuous random variable, we form the likelihood function as

$$L(\alpha|Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{t=1}^n f_Y(y_t; \alpha).$$

**Definition.** The maximum likelihood estimator (MLE) of $\alpha$, $\hat{\alpha}_{ML}$, is the value of $\alpha$ (in the parameter space) that maximizes $L(\alpha|Y_1 = y_1, \ldots, Y_n = y_n)$.

**Conditional versus Unconditional Likelihood**

For the regression model, we have a sample $(Y, X)$, whose joint density we parameterize. Because the joint density is the product of a marginal density and a conditional density, we can write the joint density of the data as

$$f(y, x; \alpha) = f(y|x; \theta) \cdot f(x; \psi).$$

The parameter vector of interest is $\theta$.

If we knew the parametric form of $f(x; \psi)$, then we could maximize the joint likelihood function. We cannot do this, as the classic model does not specify $f(x; \psi)$. However, if there is no functional relation between $\theta$ and $\psi$ (such as
the value of an element of \( \psi \) depending on an element of \( \theta \), then maximizing the joint likelihood is achieved by separately maximizing the conditional and marginal likelihoods. In such a case, the ML estimate of \( \theta \) is obtained by maximizing the conditional likelihood alone.

**Log-Likelihood for the Regression Model**

As we have already seen, Assumptions 1.2 (strict exogeneity), Assumption 1.4 (spherical error variance) and Assumption 1.5 (Gaussian) together imply \( U_i \sim N(0, \sigma^2 I_n) \). Because \( Y = X\beta + U \), we have

\[
Y|X \sim N(X\beta, \sigma^2 I_n).
\]

The log-likelihood function, which is simpler to maximize, is

\[
\ln L(\tilde{\beta}, \tilde{\sigma}^2 | (Y_1, X_1) = (y_1, x_1), \ldots, (Y_n, X_n) = (y_n, x_n)) = -\frac{n}{2} \ln (2\pi) - \frac{n}{2} \ln \tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2} (Y - X\tilde{\beta})' (Y - X\tilde{\beta}).
\]

(Because the likelihood function has the form of a joint density function, the likelihood function takes values on the unit interval. Because the likelihood function takes values on the unit interval, the log-likelihood function is negative.)

**ML via Concentrated Likelihood**

We could maximize the log likelihood in two stages. First, maximize over \( \tilde{\beta} \) for any given \( \tilde{\sigma}^2 \). The \( \tilde{\beta} \) that maximizes the objective function could (but in this case, does not) depend on \( \tilde{\sigma}^2 \). Second, maximize over \( \tilde{\sigma}^2 \), taking into account that the \( \tilde{\beta} \) from the first stage could depend on \( \tilde{\sigma}^2 \). The log likelihood function in which \( \tilde{\beta} \) is constrained to be the value from the first stage is called the concentrated log likelihood (concentrated with respect to \( \tilde{\beta} \)). Because the first stage for the Gaussian log-likelihood amounts to minimizing the sum of squares \( (Y - X\tilde{\beta})' (Y - X\tilde{\beta}) \), the value of \( \tilde{\beta} \) is simply the OLS estimator \( \beta \) (so \( B_{ML} \) and \( B_{OLS} \) are identical if the regression error is Gaussian).

In consequence, the minimized sum of squares is \( \hat{U}' \hat{U} \), so the concentrated log likelihood is

\[
\ln L_C(\tilde{\sigma}^2 | (Y_1, X_1) = (y_1, x_1), \ldots) = -\frac{n}{2} \ln (2\pi) - \frac{n}{2} \ln (\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2} \hat{U}' \hat{U}.
\]

This is a function of \( \tilde{\sigma}^2 \) alone and, because \( \hat{U}' \hat{U} \) is not a function of \( \tilde{\sigma}^2 \), one can simply take the derivative with respect to \( \tilde{\sigma}^2 \) (taking the derivative with respect...
to $\hat{\sigma}^2$, rather than $\hat{\sigma}$ can be tricky; replace $\hat{\sigma}^2$ with $\hat{\gamma}$). If we set this derivative equal to zero, we obtain

**Proposition (ML Estimator of $(\beta, \sigma^2)$):** Suppose Assumptions 1.1-1.5 hold. Then the ML estimator of $\beta$ is the OLS estimator and the ML estimator of $\sigma^2$ is

$$\frac{1}{n} \hat{U}' \hat{U} = \frac{n - K}{n} S^2.$$ 

As $S^2$ is an unbiased estimator of the variance, the ML estimator of $\sigma^2$ is biased, which indicates that a best estimator of the variance does not exist. The resultant maximized log likelihood is

$$-\frac{n}{2} \ln (2\pi) - \frac{n}{2} \ln \left( \frac{\hat{U}' \hat{U}}{n} \right) - \frac{1}{2} n = -\frac{n}{2} \ln \left( \frac{2\pi}{n} \right) - \frac{n}{2} - \frac{n}{2} \ln \left( \hat{U}' \hat{U} \right).$$

**Cramer-Rao Bound for the Classic Regression Model**

Recall from 241A, the Cramer-Rao inequality for the covariance matrix of any unbiased estimator. Let $S (\hat{\theta})$ be the score vector, which is the gradient (vector of partial derivatives) of the log likelihood

$$S (\hat{\theta}) = \frac{\partial \ln L (\hat{\theta})}{\partial \theta}.$$

**Cramer-Rao Inequality:**

1. Let $Z$ be a vector of random variables (not necessarily independent) with joint density $f(z; \theta)$.

2. Let $\theta$ be an $m$-dimensional vector of parameters, defined in a parameter space $\Theta$.

3. Let $L (\hat{\theta})$ be the likelihood and let $\hat{\theta} (z)$ be an unbiased estimator of $\theta$ with finite covariance matrix.

Under certain regularity conditions on $f(z; \theta)$,

$$\text{Var} \left[ \hat{\theta} (z) \right] \geq I (\theta)^{-1} \text{ (Cramer-Rao Lower Bound)},$$
where $I(\theta)$ is the information matrix defined by

$$I(\theta) = E \left[ S(\theta) S(\theta)' \right].$$

(Note that the score is evaluated at the true parameter value $\theta$.) Also under the regularity conditions, the information matrix equals the negative of the expected value of the Hessian (matrix of second partial derivatives) of the log likelihood:

$$I(\theta) = -E \left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right].$$

This is called the information matrix equality.

The regularity conditions guarantee that the operations of differentiation and taking expectations can be interchanged

$$E \left[ \frac{\partial L(\theta)}{\partial \theta} \right] = \partial E \left[ L(\theta) \right] / \partial \theta.$$

For the classic regression model, the Cramer-Rao bound is (derivation in Hayashi)

$$I(\theta)^{-1} = \begin{bmatrix} \sigma^2 \cdot (X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}.$$

Therefore the OLS estimator, which is equivalent to the MLE, achieves the Cramer-Rao bound and is the best unbiased estimator.

What about the estimator of $\sigma^2$? We have already seen that the MLE for $\sigma^2$ is biased, so the Cramer-Rao bound does not apply. But $S^2$ is unbiased, does it achieve the bound? It can be shown that

$$Var \left( S^2 | X \right) = \frac{2\sigma^4}{n - K},$$

so the estimator does not achieve the bound. However, it can also be shown that an unbiased estimator with lower variance does not exist, so the bound is not attainable.

Quasi-Maximum Likelihood

Of course, if the Gaussian assumption is incorrect, then the resultant estimator is not the MLE. Rather, as the likelihood is misspecified, the resultant estimator is the quasi-MLE. In many cases the Gaussian quasi-MLE performs well. Unfortunately, in general a quasi-MLE performs quite poorly.
OLSE as a Method of Moments Estimator

The OLS estimators are constructed so that the population moments hold in the sample and so are method of moments estimators. An assumption of the classic model is that each regressor is uncorrelated with the error term (captured in Assumption 2, where the regressors are assumed exogenous and measured without error). To understand the mathematical implications of the assumption, recall that two random variables are uncorrelated if they have zero covariance, which in turn implies

$$\text{Cov}(X_t, U_t) = E(X_tU_t) - EX_tEU_t = 0.$$ 

Under Assumption 3, $EU_t = 0$, so a zero covariance implies $E(X_tU_t) = 0$. The two population moments used to construct the estimators are

$$EU_t = 0,$$

which can be viewed as $E(X_{t,0}U_t) = 0$ where $X_{t,0} = 1$ is the intercept regressor, and

$$E(X_tU_t) = 0.$$

The method of moments sets sample moments equal to population moments. To construct sample analogs of these moments, we need a sample value of the unobserved error $U_t$. For a given estimator, the residual (prediction of $U_t$) is observed

$$U_t^P = Y_t - Y_t^P = Y_t - B_0 - B_1X_t.$$ 

Equality of sample and population moments yields

$$\frac{1}{n} \sum_{t=1}^{n} U_t^P = 0 \quad \text{and} \quad \frac{1}{n} \sum_{t=1}^{n} X_tU_t^P = 0.$$ 

From the definition of $U_t^P$, $\frac{1}{n} \sum_{t=1}^{n} U_t^P = 0$ implies

$$\frac{1}{n} \sum_{t=1}^{n} Y_t = \frac{1}{n} \sum_{t=1}^{n} Y_t^P.$$ 

One can readily verify that the OLS residuals do satisfy the population moments, as asserted above, by replacing the OLS estimators with their data formulae

$$\frac{1}{n} \sum_{t=1}^{n} U_t^P = \frac{1}{n} \sum_{t=1}^{n} (Y_t - B_0 - B_1X_t) = \bar{Y}_n - (\bar{Y}_n - B_1\bar{X}_n) - B_1\bar{X}_n = 0.$$
and

\[
\frac{1}{n} \sum_{t=1}^{n} X_t U_t^P = \frac{1}{n} \sum_{t=1}^{n} X_t (Y_t - B_0 - B_1 X_t)
\]

\[
= \frac{1}{n} \sum_{t=1}^{n} X_t Y_t - \frac{1}{n} \left( \bar{Y}_n \sum_{t=1}^{n} X_t - B_1 \bar{X}_n \sum_{t=1}^{n} X_t \right) - \frac{1}{n} B_1 \sum_{t=1}^{n} X_t^2
\]

\[
= \frac{1}{n} \sum_{t=1}^{n} X_t Y_t - \frac{1}{n} \bar{Y}_n \sum_{t=1}^{n} X_t + \frac{1}{n} B_1 \left( \bar{X}_n \sum_{t=1}^{n} X_t - \sum_{t=1}^{n} X_t^2 \right).
\]

Because \(\sum_{t=1}^{n} (X_t - \bar{X}_n) (Y_t - \bar{Y}_n) = \sum_{t=1}^{n} X_t Y_t - \bar{Y}_n \sum_{t=1}^{n} X_t \) and \(\sum_{t=1}^{n} (X_t - \bar{X}_n)^2 = \sum_{t=1}^{n} X_t^2 - \bar{X}_n \sum_{t=1}^{n} X_t\), the above displayed equation becomes

\[
\frac{1}{n} \sum_{t=1}^{n} X_t U_t^P = \frac{1}{n} \left[ \sum_{t=1}^{n} (X_t - \bar{X}_n) (Y_t - \bar{Y}_n) - B_1 \sum_{t=1}^{n} (X_t - \bar{X}_n)^2 \right].
\]

Because \(B_1 = \frac{\sum_{t=1}^{n} (X_t - \bar{X}_n) (Y_t - \bar{Y}_n)}{\sum_{t=1}^{n} (X_t - \bar{X}_n)^2}\), the above expression equals

\[
\frac{1}{n} \left[ \sum_{t=1}^{n} (X_t - \bar{X}_n) (Y_t - \bar{Y}_n) - \sum_{t=1}^{n} (X_t - \bar{X}_n) (Y_t - \bar{Y}_n) \right] = 0.
\]

Finally, note that orthogonality between \(U_t^P\) and the regressors implies orthogonality between \(U_t^P\) and \(Y_t^P\), which is a linear combination of the regressors. In detail:

The conditional expectations used to define the model contain important information. If we treat the regressor as a random variable, then we must distinguish between conditional and unconditional expectations. For example, the conditional expectation of \(Y_t^P\) is

\[
E \left( Y_t^P | X_t \right) = E \left( A_{OLS} + B_{OLS} X_t | X_t \right) = \alpha + \beta X_t = E \left( Y_t | X_t \right).
\]

The unconditional expectation of \(Y_t^P\) is

\[
E \left( Y_t^P \right) = E \left( A_{OLS} + B_{OLS} X_t \right) = \alpha + \beta E X_t,
\]

which is constant if the expectation of the regressor is constant across observations. While the conditional and unconditional expectations of \(Y_t^P\) differ, the conditional and unconditional expectations of \(U_t^P\) are the same

\[
E \left( U_t^P | X_t \right) = E \left( Y_t - Y_t^P | X_t \right) = \alpha + \beta X_t - E \left( Y_t^P | X_t \right) = 0.
\]
and
\[ E(U_t^P) = E(Y_t - Y_t^P) = \alpha + \beta EX_t - E(Y_t^P) = 0. \]
The (unconditional) covariance between \( Y_t^P \) and \( U_t^P \) is
\[
E \left[ (Y_t^P - EY_t^P) U_t^P \right] = E \left[ (A_{OLS} + B_{OLS}X_t - \alpha - \beta EX_t) U_t^P \right] \\
= E \left[ (A_{OLS} + B_{OLS}X_t) U_t^P \right] = 0,
\]
where the second line follows because \( \alpha - \beta EX_t \) is not random and \( EU_t^P = 0 \), and the third line follows because \( A_{OLS} + B_{OLS}X_t \) is uncorrelated with \( U_t^P \) by construction (recall, if \( X \) and \( Y \) are uncorrelated, then \( E(XY) = EX \cdot EY \)).

Clearly, if the predicted values of the dependent variable were correlated with the estimated residuals, then the predicted values could be improved, so we expect zero covariance.

To show that the sample estimate is always zero, the sample estimate of the covariance between \( Y_t^P \) and \( U_t^P \) is
\[
\frac{1}{n-2} \sum_{t=1}^{n} \left( y_t^P - \frac{1}{n} \sum_{t=1}^{n} y_t^P \right) u_t^P = \frac{1}{n-2} \sum_{t=1}^{n} \left( bx_t - b \frac{1}{n} \sum_{t=1}^{n} x_t \right) u_t^P \\
= \frac{b}{n-2} \left( \sum_{t=1}^{n} x_t u_t^P - \bar{x} \sum_{t=1}^{n} u_t^P \right) = 0,
\]
where the third line follows from the normal equations that state \( \sum_{t=1}^{n} u_t^P = \sum_{t=1}^{n} x_t u_t^P = 0 \). Of course the normal equations ensure that the sample analogs equal the population moments. The relevant population moments are \( E(U_t|X_t) = 0 \) (the residuals are mean zero) and \( E(X_tU_t|X_t) = 0 \) (the residuals are uncorrelated with the regressors).

Recall Assumption 2: Issues of identification are in play here. To make the issues clear, consider the model
\[ Y_t = \alpha_0 + X_t \gamma_0 + U_t, \]
in which \( x_t \) is the \( k-1 \times 1 \) vector that does not include the intercept. We now ask, under what conditions are the coefficients identified? If the covariance matrix of \( X_t \) is nonsingular and \( X_t \) is independent of \( U_t \), then \( \gamma_0 \) is identified. An additional assumption is needed to identify \( \alpha_0 \). Two alternative assumptions that identify
\( \alpha_0 \) are: \( EU_t = 0 \) and \( Med (U_t) = 0 \). The only difference is in interpretation of \( \alpha_0 + X_t \gamma_0 \), as discussed above.

Alternatively, we could assume that \( U_t \) is symmetrically distributed around 0, conditional on \( X_t \). Then \( \alpha_0 \) and \( \gamma_0 \) are identified and \( \alpha_0 + X_t \gamma_0 \) is both the conditional mean and the conditional median, as well as being equal to other location measures.

Both \( \alpha_0 \) and \( \gamma_0 \) are identified under a conditional location restriction that is weaker than either the assumption of independence (between the regressor and the error) or the assumption of conditional symmetry. Further, each conditional location restriction is associated with a conditional moment restriction \( E [f (U_t) | X_t] = 0 \) for some function \( f (U_t) \) from which an estimator is constructed. Consider the two location assumptions introduced earlier. If \( E (U_t | X_t) = 0 \), then \( f (U_t) = U_t \) and the resultant estimator is OLS (and, again, \( \alpha_0 + X_t \gamma_0 \) is the conditional mean of \( Y_t \)). If \( Med (U_t | X_t) = 0 \), the corresponding moment condition is \( E [sgn (U_t) | X_t] = 0 \) and the resulting estimator is least absolute deviations (and, again, \( \alpha_0 + X_t \gamma_0 \) is the conditional median of \( Y_t \)).\(^1\)

To derive the moment condition for OLS, note that \( E (U_t | X_t) = 0 \) is clearly a moment condition that can be used for estimation. The OLSE \( B \) thus satisfies

\[
\sum_{t=1}^{n} X_t U_t (B) = 0.
\]

While \( Med (U_t | X_t) = 0 \) is a moment condition, it may not be as clear how it can be used to form an estimator. Consider first the case in which \( U_t \) is continuous. The assumption \( Med (U_t | X_t) = 0 \) implies

\[
P (U_t < 0 | X_t) = P (U_t > 0 | X_t) = \frac{1}{2},
\]

which implies \( E [sgn (U_t) | X_t] = 0 \), which in turn implies

\[
E [X_t sgn (U_t)] = 0.
\]

\(^1\)The signum, or sign, function is defined as

\[
sgn (u) = \begin{cases} 
1 & \text{if } u > 0 \\
0 & \text{if } u = 0 \\
-1 & \text{if } u < 0
\end{cases}.
\]
The LAD estimator $B_L$ satisfies the sample analog

$$\sum_{t=1}^{n} X_t \text{sgn} (U_t (B_L)) = 0.$$ 

There are two problems here. First it may not be apparent that the sample analog with $B_L$ admits a unique solution. In fact, in Powell’s symmetrically trimmed LAD paper in Econometrica, his conditional moment equation has many solutions. Also, if $U_t$ is not distributed symmetrically, then the assumption $\text{Med}(U_t|X_t) = 0$ does not necessarily lead to a simple moment condition for estimation. The problem is, if $U_t$ does not have a continuous distribution, then it is possible that there is positive point mass at the median, so it is possible that $E[\text{sgn} (U_t) | X_t] \neq 0$. The alternative is to return to the loss function (also termed the objective function). The loss function approach solves both problems. First, there is clearly a unique solution (as Powell shows in the appendix to the above mentioned paper). Second, the loss function approach works well even if $U_t$ does not have a continuous distribution.