This lecture consists of three sections:

- Three commonly used tests
- Composite hypothesis
- Application to linear regression model

\[
\ln \mathcal{L} = \sum_{t=1}^{T} \ln f(Y_t | \theta),
\]

where the $Y_t$ are i.i.d.
1 Three commonly used tests

In this section we examine 3 commonly used (asymptotically equivalent) tests:

1. Wald Test
2. Rao, Score or Lagrange Multiplier (LM) Test
3. LR Test

We motivate the derivation of the test statistic in each case and show asymptotic equivalence of the three tests.
1.1 Wald Test

Consider a simple null hypothesis:

\[ H_0 : \theta = \theta_0, \quad H_A : \theta \neq \theta_0. \]

Natural test uses the fact\(^1\) that:

\[ \sqrt{T}(\hat{\theta} - \theta_0) \sim N(0, I_{\theta_0}^{-1}) \]

so that \[ \sqrt{T}(\hat{\theta} - \theta_0)'I_{\theta_0}\sqrt{T}(\hat{\theta} - \theta_0) \sim \chi^2(k) \] \((k = \text{number of parameters})\), where:

\[ I_{\theta_0} = -E \left[ \frac{1}{T} \frac{\partial^2 \ln \mathcal{L}}{\partial \theta \partial \theta'} \right] = -E \left[ \frac{\partial^2 \ln f(Y_t | \theta)}{\partial \theta \partial \theta'} \right], \]

for \( Y_t \) i.i.d.

\(^1\)See Lecture III on Asymptotic theory.
This is the Wald Test and is similar to the $F$ test used in OLS hypothesis testing. We use the uniform convergence of $\hat{\theta}$ to $\theta_0$ to obtain $\text{plim } I_{\hat{\theta}} = I_{\theta_0}$, so that the test statistic for the Wald test:

$$W = \sqrt{T}(\hat{\theta} - \theta_0)' I_{\hat{\theta}} \sqrt{T}(\hat{\theta} - \theta_0) \overset{d}{\to} \chi^2(k)$$

in large samples.

Note that the Wald test is based on the *unrestricted* MLE model.
Recall at true parameter vector $\theta = \theta_0$,

$$E \left[ \frac{\partial \ln \mathcal{L}(y \mid \theta)}{\partial \theta} \right] = \int \frac{\partial \ln \mathcal{L}(y \mid \theta)}{\partial \theta} \bigg|_{\theta=\theta_0} \mathcal{L}(y \mid \theta_0) \, dy = 0$$

because

$$\int \mathcal{L}(y \mid \theta_0) \, dy = 1 \implies \int \frac{\partial \mathcal{L}(y \mid \theta)}{\partial \theta} \, dy = 0,$$

but

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \ln \mathcal{L}}{\partial \theta} \mathcal{L}$$

$$\implies \int \frac{\partial \ln \mathcal{L}(y \mid \theta)}{\partial \theta} \bigg|_{\theta=\theta_0} \mathcal{L}(y \mid \theta_0) \, dy = 0.$$
Differentiating again,

\[
\int \frac{\partial^2 \ln \mathcal{L}(y | \theta)}{\partial \theta \partial \theta'} \bigg|_{\theta=\theta_0} \mathcal{L}(y | \theta_0) \, dy \\
+ \int \frac{\partial \ln \mathcal{L}(y | \theta)}{\partial \theta} \bigg|_{\theta=\theta_0} \frac{\partial \ln \mathcal{L}(y | \theta)}{\partial \theta'} \bigg|_{\theta=\theta_0} \mathcal{L}(y | \theta_0) \, dy
\]

\[= 0.\]

Therefore,

\[I_{\theta_0} = E \left[ \frac{1}{T} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta} \bigg|_{\theta=\theta_0} \frac{\partial \ln \mathcal{L}}{\partial \theta'} \bigg|_{\theta=\theta_0} \right) \right].\]
Cramer-Rao Lower Bound (Scalar Case)

Consider an estimator \( t(y) \) of \( \theta \):

\[
E(t(y)) = \int t(y) \mathcal{L}(y; \theta) \, dy.
\]

Under regularity,

\[
\frac{\partial E(t(y))}{\partial \theta} = \int t(y) \frac{\partial \ln \mathcal{L}(y; \theta)}{\partial \theta} \, dy = Cov \left(t(y), \frac{\partial \ln \mathcal{L}}{\partial \theta} \right).
\]
\[
\frac{\partial \mathbb{E}(t(y))}{\partial \theta} = \int t(y) \frac{\partial \mathcal{L}}{\partial \theta} \, dy
\]
\[
= \int t(y) \left( \frac{\partial \ln \mathcal{L}}{\partial \theta} \right) \mathcal{L} \, dy
\]
\[
= \text{Cov} \left( t(x), \frac{\partial \ln \mathcal{L}}{\partial \theta} \right),
\]

because
\[
\mathbb{E}\left( \frac{\partial \ln \mathcal{L}}{\partial \theta} \right) = 0.
\]
From the Cauchy-Schwartz inequality,

\[
\left( \frac{\partial E(t(y))}{\partial \theta} \right)^2 \leq Var(t(y)) \cdot Var\left( \frac{\partial \ln \mathcal{L}}{\partial \theta} \right). \]

For full rank information matrix \( I_{\theta_0} \),

\[
Var(t(y)) \geq \frac{\left( \frac{\partial E(t(y))}{\partial \theta} \right)^2}{I_{\theta_0}}.
\]
If unbiased,

\[ E(t(y)) = \theta, \quad \frac{\partial \ln E(t(y))}{\partial \theta} = 1 \]

\[ \text{Var}(t(y)) \geq \frac{1}{I_{\theta_0}} \]

There is a vector version:

\[ \text{Var}(t(y)) \geq I_{\theta_0}^{-1}. \]

Cannot do better than MLE in terms of efficiency.
1.2 Rao Test (also LM or Score Test)

The second Test – Rao Test (LM) is based on the restricted model. Observes that in a large enough sample $\theta_0$ (true parameter value) should be a root of the likelihood equation:

$$
\frac{1}{T} \frac{\partial \ln \mathcal{L}}{\partial \theta} \bigg|_{\theta_0} \equiv \frac{1}{T} \sum_{t=1}^{T} \frac{\partial \ln f(y_t, \theta)}{\partial \theta} \bigg|_{\theta_0} = 0,
$$

i.e., it imposes the null onto the model for all sample sizes.

In contrast, the Wald Test gets its statistics from estimates of the unrestricted model (i.e., a model where the null is not imposed on the estimates). The Likelihood Ratio Test compares the likelihood restricted with the likelihood unrestricted.
\[
\frac{\partial \ln \mathcal{L}}{\partial \theta} \bigg|_{\hat{\theta}_{ML}} = 0 \quad \text{and} \quad \mathbb{E} \left( \frac{\partial \ln f(Y_T \mid \theta)}{\partial \theta} \right) = 0,
\]

but

\[
\frac{1}{T} \frac{\partial \ln \mathcal{L}}{\partial \theta} \xrightarrow{\text{P}} \mathbb{E} \left( \frac{\partial \ln f(Y_T \mid \theta)}{\partial \theta} \right) = 0
\]

\[
\Rightarrow \quad \frac{1}{T} \frac{\partial \ln \mathcal{L}}{\partial \theta} \bigg|_{\theta_0} \xrightarrow{\text{P}} 0.
\]
Now

\[
\frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta} \bigg|_{\theta_0} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\partial \ln f(y_t, \theta)}{\partial \theta} \bigg|_{\theta_0} \sim N(0, I_{\theta_0})
\]

implies that the hypothesis could be tested by testing if score
\[
\frac{\partial \ln \mathcal{L}}{\partial \theta} \bigg|_{\theta_0} = 0,
\]
at the restricted parameter values.\(^2\)

\(^2\)For the distributional results, refer to earlier lectures on Asymptotic theory.
Thus this test uses the statistic:

\[
LM = \left( \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta} \right)'_{\theta_0} \quad I_{\hat{\theta}}^{-1} \left( \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta} \right)_{\theta_0}
\]

\[\xrightarrow{d} \chi^2(k).\]

\(LM \sim \chi^2(k)\) in large samples can be shown using plim \(I_{\hat{\theta}} = I_{\theta_0}\).
The Rao test is also called the Score Test or the Lagrange Multiplier test. This is because this test can be motivated from the solution to a constrained maximization of the log-likelihood subject to constraint $\theta = \theta_0$. We get:

Lagrangian: \[ \ln \mathcal{L} - \lambda'(\theta - \theta_0) \]

FOC: \[ \frac{\partial \ln \mathcal{L}}{\partial \theta} - \lambda = 0 \]

$\Rightarrow$ At null: \[ \lambda = \frac{\partial \ln \mathcal{L}}{\partial \theta} = 0 \]

Thus one can test on $\lambda$ (the Lagrange Multiplier) or on the Score $\left( \frac{\partial \ln \mathcal{L}}{\partial \theta} \right)$. 
Asymptotic Equivalence of Wald and LM tests To establish the asymptotic relationships between the two tests, we use Taylor’s theorem to write:

\[
\frac{1}{T} \left( \sum \frac{\partial \ln f}{\partial \theta} \right)_{\hat{\theta}} = \frac{1}{T} \sum \frac{\partial \ln f}{\partial \theta} \bigg|_{\theta_0} = 0 \text{ by construction}
\]

\[
+ \frac{1}{T} \left( \sum \frac{\partial^2 \ln f}{\partial \theta \partial \theta'} \right)_{\theta^*} (\hat{\theta} - \theta_0),
\]

where \(\theta^*\) is an intermediate value with

\[
\|\theta_0\| \leq \|\theta^*\| \leq \|\hat{\theta}\|.
\]
From above, we get the duality relationships between score and parameter vectors:

$$(\hat{\theta} - \theta_0) = - \left( \sum \frac{\partial^2 \ln f}{\partial \theta \partial \theta'} \right)^{-1} \sum \frac{\partial \ln f}{\partial \theta} \bigg|_{\theta_0}$$

$$\Rightarrow \sqrt{T}(\hat{\theta} - \theta_0) = \left( -\frac{1}{T} \left( \sum \frac{\partial^2 \ln f}{\partial \theta \partial \theta'} \right)_{\theta^*} \right)^{-1} \frac{1}{\sqrt{T}} \sum \frac{\partial \ln f}{\partial \theta} \bigg|_{\theta_0}.$$ 

Noting that $\text{plim} \theta^* = \theta_0$, substituting into the Wald statistic we get:

$$\text{plim} W = \left( \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta} \right)' \bigg|_{\theta_0} \left( \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta} \right)_{\theta_0} = \text{plim} LM$$

Thus, asymptotically the Wald and Rao tests are equivalent.
1.3 Likelihood Ratio Test

The third commonly used test is the Likelihood Ratio Test and it uses both the restricted and the unrestricted models. Taylor expanding the Likelihood function around the point $\hat{\theta}$ we get:

$$
\ln \mathcal{L}(\theta_0) = \ln \mathcal{L}(\hat{\theta}) + \left. \frac{\partial \ln \mathcal{L}}{\partial \theta} \right|_{\hat{\theta}} (\theta_0 - \hat{\theta}) + \frac{1}{2} (\theta_0 - \hat{\theta})' \left( \frac{\partial^2 \ln \mathcal{L}}{\partial \theta \partial \theta'} \right)_{\theta^*} (\theta_0 - \hat{\theta}),
$$

where $\theta^*$ is an intermediate value with $\|\theta_0\| < \|\theta^*\| < \|\hat{\theta}\|$. 
\[
\ln \mathcal{L}_0 \equiv \ln \mathcal{L} \left( \hat{\theta}_0 \right) > \ln \mathcal{L} \left( \hat{\theta}_R \right) \equiv \ln \mathcal{L}_R
\]

\[
\ln \mathcal{L}_R = \ln \mathcal{L}_0 + \frac{\partial \ln \mathcal{L}_0 (\theta_R - \theta_0)}{\partial \theta} = 0
\]

\[
+ \frac{\theta_R - \theta_0}{2} \frac{\partial^2 \mathcal{L}_0 (\theta_R - \theta_0)}{\partial \theta \partial \theta'}
\]

\[
2 (\ln \mathcal{L}_R - \ln \mathcal{L}_0) = (\theta_R - \theta_0)' \frac{\partial^2 \mathcal{L}_0}{\partial \theta \partial \theta'} (\theta_R - \theta_0)
\]

\[
= -2 (\ln \mathcal{L}_R - \ln \mathcal{L}_0)
\]

\[
\oplus
\]
\[
= \sqrt{T} (\theta_R - \theta_0)' \left( \frac{-2^2 \mathcal{L}_0}{(T) \partial \theta \partial \theta'} \right) (\theta_R - \theta_0) \sqrt{T}
\]

where

\[
\left( \frac{-2^2 \mathcal{L}_0}{(T) \partial \theta \partial \theta'} \right) \xrightarrow{p} I_{\theta_0}
\]

and

\[
(\theta_R - \theta_0) \sqrt{T} \xrightarrow{d} N \left( 0, I_{\theta_0}^{-1} \right).
\]
\[ 2[\ln \mathcal{L}(\hat{\theta}) - \ln \mathcal{L}(\theta_0)] = \sqrt{T}(\hat{\theta} - \theta_0)' \left( \frac{-1}{T} \frac{\partial^2 \ln \mathcal{L}}{\partial \theta \partial \theta'} \right)_{\theta^*} \sqrt{T}(\hat{\theta} - \theta_0) \]

\[ \Rightarrow 2 \ln \left[ \frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\theta_0)} \right] \xrightarrow{d} \sqrt{T}(\hat{\theta} - \theta_0)' I_{\theta_0} \sqrt{T}(\hat{\theta} - \theta_0). \]

We may also write

\[ 2 \ln \left[ \frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\theta_0)} \right] \xrightarrow{d} \xi'I_{\theta_0}\xi \sim \chi^2(k); \]

\[ \xi \sim N(0, I_{\theta_0}^{-1}), \text{ and } \sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} \xi. \]
Based on the above derivation, we get the statistic for the likelihood ratio test:

\[ LR = -2 \ln \left( \frac{\hat{L}_R}{\hat{L}_U} \right), \]

where \( \hat{L}_R \) is the restricted maximum likelihood function and \( \hat{L}_U \) is the unrestricted maximum likelihood function, and as shown above:

\[ LR \xrightarrow{d} \chi^2(k). \]

Note that \( \hat{L}_R \leq \hat{L}_U \) (as an unrestricted maximized function would be greater than or equal to the restricted maximized function), so that always \( LR \geq 0 \).
Asymptotic Equivalence of LR and Wald tests

From the derivation above, it is clear that asymptotically, the LR test statistic converges to the Wald test statistic, i.e.,

$$\text{plim} LR = \sqrt{T} (\hat{\theta} - \theta_0) I_{\theta_0} \sqrt{T} (\hat{\theta} - \theta_0) = \text{plim} W$$

where $\xi \sim N(0, I_{\theta_0})$.

We saw earlier that the LM and Wald tests were asymptotically equivalent. Thus along with the above asymptotic equivalence of LR and Wald we get that asymptotically all three commonly used tests are equivalent, i.e., asymptotically:

$$LR \Leftrightarrow Wald \Leftrightarrow LM.$$
2 Composite Hypothesis

Consider a parameter vector \( \theta = (\theta_1, \theta_2) \) (where \( \theta_1 \) and \( \theta_2 \) are possibly vectors). In many situations, we may want to consider estimating only a subset of the parameters \( \theta_2 \). This could be because we omit \( \theta_1 \) or we hypothesize that \( \theta_1 = 0 \). In this context, 2 questions are relevant:

- When do we get an unbiased estimate for \( \theta_2 \) if we do the MLE estimation omitting \( \theta_1 \)?

- How do the results obtained in the previous section apply to test the composite hypothesis:

\[
H_0 : \theta_1 = 0 \text{ and } \theta_2 \text{ unrestricted} \\
H_A : \theta_1 \text{ and } \theta_2 \text{ both unrestricted}
\]

We explore these two questions in the next few subsections.
2.1 Definitions

We use the following defined expressions in the rest of the section.

- Define true parameter vector $\theta_0 \equiv (\bar{\theta}_1, \bar{\theta}_2)$;
- Define estimated parameter vector (unconstrained) as $\hat{\theta} \equiv (\hat{\theta}_1, \hat{\theta}_2)$;
- Define estimated parameter vector (constrained, i.e. under the null hypothesis setting $\theta_1 = 0$) as $\tilde{\theta} \equiv (0, \tilde{\theta}_2)$;
- In Taylor expansions hereafter, we use $\theta_0$ and $\theta^*$ interchangeably (neglecting $o_p(1)$ terms);
• Define the various information matrix related terms as follows:

\[ I_{\theta_i \theta_j} \equiv - \frac{1}{T} \left( \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta'_j} \right)_{\theta_0}, \quad I_{\theta_0} = \begin{pmatrix} I_{\theta_1 \theta_1} & I_{\theta_1 \theta_2} \\ I_{\theta_2 \theta_1} & I_{\theta_2 \theta_2} \end{pmatrix} \]
2.2 When is $\hat{\theta}_2$ unbiased if $\theta_1$ is omitted?

In this subsection, we show that the condition for $\hat{\theta}_2$ to be an unbiased estimator of $\bar{\theta}_2$ when the model is misspecified and estimated omitting $\theta_1$ is that the score vectors $\frac{\partial \ln L}{\partial \theta_1}$ and $\frac{\partial \ln L}{\partial \theta_2}$ be orthogonal to each other.
First, Taylor expanding the score vectors $\theta_1$ and $\theta_2$, we get:

\[
\frac{1}{\sqrt{T}} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \right)_{\hat{\theta}} = \frac{1}{\sqrt{T}} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \right)_{\theta_0} - I_{\theta_1 \theta_1} \sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1) - I_{\theta_1 \theta_2} \sqrt{T}(\hat{\theta}_2 - \bar{\theta}_2) + o_P(1)
\]

(1)

\[
\frac{1}{\sqrt{T}} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_2} \right)_{\hat{\theta}} = \frac{1}{\sqrt{T}} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_2} \right)_{\theta_0} - I_{\theta_2 \theta_2} \sqrt{T}(\hat{\theta}_2 - \bar{\theta}_2) - I_{\theta_1 \theta_2} \sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1) + o_P(1)
\]

(2)

(1) $\Rightarrow$ $0 = \frac{1}{\sqrt{T}} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \right)_{\theta_0} - I_{\theta_1 \theta_1} \sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1) - I_{\theta_1 \theta_2} \sqrt{T}(\hat{\theta}_2 - \bar{\theta}_2) + o_P(1)

(2) $\Rightarrow$ $0 = \frac{1}{\sqrt{T}} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_2} \right)_{\theta_0} - I_{\theta_2 \theta_2} \sqrt{T}(\hat{\theta}_2 - \bar{\theta}_2) - I_{\theta_2 \theta_1} \sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1) + o_P(1)
Collecting terms we get:

\[
\begin{pmatrix}
\sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1) \\
\sqrt{T}(\hat{\theta}_2 - \bar{\theta}_2)
\end{pmatrix} = [I_{\theta\theta}]^{-1}
\begin{pmatrix}
\frac{1}{\sqrt{T}} \left( \frac{\partial lnL}{\partial \theta_1} \right)_{\theta_0} \\
\frac{1}{\sqrt{T}} \left( \frac{\partial lnL}{\partial \theta_2} \right)_{\theta_0}
\end{pmatrix}
\]

Next consider \(\tilde{\theta}_2\) (MLE of \(\theta_2\) given \(\bar{\theta}_1 = 0\)). Expanding around root of likelihood \(\bar{\theta}_1 = 0\), we get:

\[
\frac{1}{\sqrt{T}} \left( \frac{\partial lnL}{\partial \theta_2} \right)_{\tilde{\theta}} = \frac{1}{\sqrt{T}} \left( \frac{\partial lnL}{\partial \theta_2} \right)_{\theta_0} - I_{\theta_2\theta_2} \sqrt{T}(\tilde{\theta}_2 - \bar{\theta}_2) + o_P(1)
\]

(3)

Now the first term on the right hand side is in common with the term on \(rhs\) of (2) above. Therefore we have that:

\[
I_{\theta_2\theta_2} \sqrt{T}(\tilde{\theta}_2 - \bar{\theta}_2) = I_{\theta_2\theta_2} \sqrt{T}(\hat{\theta}_2 - \bar{\theta}_2) + I_{\theta_2\theta_1} \sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1) + o_P(1)
\]
or:

\[
\sqrt{T}(\tilde{\theta}_2 - \bar{\theta}_2) = \sqrt{T}(\hat{\theta}_2 - \bar{\theta}_2) + I_{\theta_2\theta_2}^{-1} I_{\theta_2\theta_1} \sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1) + o_P(1)
\]  

Thus, we get that for \(\sqrt{T}(\tilde{\theta}_2 - \bar{\theta}_2) = \sqrt{T}(\hat{\theta}_2 - \bar{\theta}_2)\) we need:

\[
I_{\theta_2\theta_1} = E \left[ \frac{\partial \ln L}{\partial \theta_2} \left( \frac{\partial \ln L}{\partial \theta_1} \right)' \right] = 0,
\]

i.e., for \(\tilde{\theta}\) (the constrained estimator) to have the same properties as \(\hat{\theta}\) (the unconstrained estimator) we need that the score vectors \(\frac{\partial \ln L}{\partial \theta_1}\) and \(\frac{\partial \ln L}{\partial \theta_2}\) be uncorrelated.
Similarity with OLS result: Note that this result is analogous to a result for the OLS framework. In the OLS model we have:

\[ Y = X_1 \beta_1 + X_2 \beta_2 + U \]

If we run the regression omitting \( X_1 \), we have:

\[ Y = X_2 \beta_2 + \{U + X_1 \beta_1\} \]

Then we get, under standard OLS assumptions:

\[
\text{plim} \hat{\beta}_{2,OLS} = \beta_2 + \text{plim} \left( \frac{X'_2 X_2}{T} \right)^{-1} \text{plim} \left( \frac{X'_2 X_1}{T} \right) \beta_1,
\]

so that we get consistent estimate \( \hat{\beta}_{2,OLS} \) if the \( X_2 \) and \( X_1 \) are orthogonal to each other.\(^3\)

---

\(^3\)Note \( \text{plim} \frac{X'_2 X_1}{T} = E[X'_2 X_1] \) by appropriate LLN.
Note that the score vectors in MLE are analogous to the data vectors in OLS; we shall revisit this analogy again in Section 3 below.
2.3 Hypothesis testing results for the composite hypothesis

In this subsection, we show the following results:

- Asymptotic equivalence of LR test and Wald test;
- Asymptotic equivalence of Rao test (Score/LM test) and the LR test.
2.3.1 Equivalence between Likelihood Ratio and Wald Tests

From the definition of the LR test statistic in Section 1.3, we have the analogous definition for the composite hypothesis case as:

\[ LR = 2 \ln \frac{\mathcal{L}(\hat{\theta}_1, \hat{\theta}_2)}{\mathcal{L}(\theta_1 = 0, \tilde{\theta}_2)} = -2 \ln \frac{\mathcal{L}(\theta_1 = 0, \tilde{\theta}_2)}{\mathcal{L}(\hat{\theta}_1, \hat{\theta}_2)} \]

or

\[ LR = -2[\ln \mathcal{L}(\theta_1 = 0, \tilde{\theta}_2) - \ln \mathcal{L}(\tilde{\theta}_1, \tilde{\theta}_2)] + 2[\ln \mathcal{L}(\hat{\theta}_1, \hat{\theta}_2) - \ln \mathcal{L}(\tilde{\theta}_1, \tilde{\theta}_2)]. \]
Then following same steps as in derivation in Section 1.3, we get:

\[
LR = \sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1, \hat{\theta}_2 - \bar{\theta}_2)'I_{\theta_0}\sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1, \hat{\theta}_2 - \bar{\theta}_2) - \sqrt{T}(0, \tilde{\theta}_2 - \bar{\theta}_2)'I_{\theta_0}\sqrt{T}(0, \tilde{\theta}_2 - \bar{\theta}_2) + o_P(1)
\]

or

\[
LR = \sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1, \hat{\theta}_2 - \bar{\theta}_2)'I_{\theta_0}\sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1, \hat{\theta}_2 - \bar{\theta}_2) - \sqrt{T}(\tilde{\theta}_2 - \bar{\theta}_2)'I_{\theta_2\theta_2}\sqrt{T}(\tilde{\theta}_2 - \bar{\theta}_2) + o_P(1)
\]
Substituting for \((\tilde{\theta}_2 - \bar{\theta}_2)\) from equation (4) in Section 2.2, we get:

\[
\begin{align*}
\text{LR} &= \sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1, \hat{\theta}_2 - \bar{\theta}_2)'I_{\bar{\theta}_1, \bar{\theta}_2}\sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1, \hat{\theta}_2 - \bar{\theta}_2) \\
&- \left[\sqrt{T}(\hat{\theta}_2 - \bar{\theta}_2) + (I_{\theta_2\theta_2})^{-1}I_{\theta_2\theta_1}\sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1)\right]'I_{\theta_2\theta_2} \\
&\left[\sqrt{T}(\hat{\theta}_2 - \bar{\theta}_2) + (I_{\theta_2\theta_2})^{-1}I_{\theta_2\theta_1}\sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1)\right] + o_P(1)
\end{align*}
\]

Call \(\theta_1 - \bar{\theta}_1 \equiv a\) and \(\theta_2 - \bar{\theta}_2 \equiv b\).
Then we have:

\[ LR = \sqrt{T}a' I_{\theta_1\theta_1} \sqrt{T}a + \sqrt{T}b' I_{\theta_2\theta_1} \sqrt{T}a \]
\[ + \sqrt{T}a' I_{\theta_1\theta_2} \sqrt{T}b + \sqrt{T}b' I_{\theta_2\theta_2} \sqrt{T}b \]
\[ - \sqrt{T}a' I_{\theta_1\theta_2} (I_{\theta_2\theta_2})^{-1} I_{\theta_2\theta_2} (I_{\theta_2\theta_2})^{-1} I_{\theta_2\theta_1} \sqrt{T}a \]
\[ - \sqrt{T}a' I_{\theta_1\theta_2} (I_{\theta_2\theta_2})^{-1} I_{\theta_2\theta_2} \sqrt{T}b \]
\[ - \sqrt{T}b' I_{\theta_2\theta_2} \sqrt{T}b - \sqrt{T}b' I_{\theta_2\theta_2} I_{\theta_2\theta_2}^{-1} I_{\theta_2\theta_1} \sqrt{T}a + o_P(1) \]
\[ = \sqrt{T}a' [I_{\theta_1\theta_1} - I_{\theta_1\theta_2} (I_{\theta_2\theta_2})^{-1} I_{\theta_2\theta_2}] \sqrt{T}a + o_P(1) \]
\[ = \sqrt{T}a' (I^{\theta_1\theta_1})^{-1} \sqrt{T}a + o_P(1), \quad (5) \]

where \( I^{\theta_1\theta_1} \) is the upper left diagonal block of:

\[ I_{\theta_0}^{-1} = \left( \begin{array}{cc} I_{\theta_1\theta_1} & I_{\theta_1\theta_2} \\ I_{\theta_2\theta_1} & I_{\theta_2\theta_2} \end{array} \right)^{-1} \]
(See partitioned inverse result established in section 3.)

In the composite hypothesis case, the Wald test exploits the fact that: \( \sqrt{T} (\hat{\theta}_1 - \bar{\theta}_1) \sim N(0, I^{\theta_1\theta_1}) \) so that the Wald test statistic here is:

\[
W = \sqrt{T} (\hat{\theta}_1 - \bar{\theta}_1)' (I^{\theta_1\theta_1})^{-1} \sqrt{T} (\hat{\theta}_1 - \bar{\theta}_1) + o_P(1). \quad (6)
\]

From (5) and (6) above, we get directly the asymptotic equivalence of the Wald and LR tests, i.e., \( LR \iff W \), in large samples.
2.3.2 Equivalence between Rao (Score/LM) test and the other tests

To derive the statistic for the Rao (Score/LM) test, first compute the derivative with respect to $\theta_1$ when $\theta_1$ constraint is imposed):

$$\frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \bigg|_{\tilde{\theta}} = \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \bigg|_{\theta_0} - I_{\theta_1\theta_2} \sqrt{T} \left( \tilde{\theta}_2 - \bar{\theta}_2 \right) + o_P(1). \quad (7)$$
From Eqn (3) in Section 2.2 we have:

$$\sqrt{T} \left( \tilde{\theta}_2 - \bar{\theta}_2 \right) = I_{\theta_2 \theta_2}^{-1} \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_2} \bigg|_{\theta_0} + o_P(1)$$

Substituting above result in (7), we get:

$$\frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \bigg|_{\tilde{\theta}} = \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \bigg|_{\theta_0} - I_{\theta_1 \theta_2} I_{\theta_2 \theta_2}^{-1} \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_2} \bigg|_{\theta_0} + o_P(1)$$

$$\equiv A - I_{\theta_1 \theta_2} I_{\theta_2 \theta_2}^{-1} B + o_P(1),$$

defining $A \equiv \frac{1}{\sqrt{T}} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \right) \bigg|_{\theta_0}$ and $B \equiv \frac{1}{\sqrt{T}} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_2} \right) \bigg|_{\theta_0}$. 
Then we obtain variance of the key score parameter as:

\[
Var \left( \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \bigg| \hat{\theta} \right) \equiv E[(A - I_{\theta_1 \theta_2} I_{\theta_2 \theta_2}^{-1} B)(A - I_{\theta_1 \theta_2} I_{\theta_2 \theta_2}^{-1} B)'] \\
= I_{\theta_1 \theta_1} + I_{\theta_1 \theta_2} I_{\theta_2 \theta_2}^{-1} I_{\theta_2 \theta_2} I_{\theta_2 \theta_2}^{-1} I_{\theta_2 \theta_1} \\
- 2 I_{\theta_1 \theta_2} I_{\theta_2 \theta_2}^{-1} I_{\theta_2 \theta_1} \\
= (I_{\theta_1 \theta_1})^{-1}
\]
\[ \text{LM} = \left( \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \Bigg|_{\tilde{\theta}} \right)' I^{\theta_1 \theta_1} \left( \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \Bigg|_{\tilde{\theta}} \right) \]

\[ = [A - I_{\theta_1 \theta_2} I_{\theta_2 \theta_2}^{-1} B]' I^{\theta_1 \theta_1} [A - I_{\theta_1 \theta_2} I_{\theta_2 \theta_2}^{-1} B]. \]

From Section 2.2 and results for partitioned inverses (refer Section 2.3.1), we have:

\[ \sqrt{T}(\hat{\theta}_1 - \bar{\theta}_1) = I^{\theta_1 \theta_1} \left( \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \Bigg|_{\theta_0} \right) + I^{\theta_1 \theta_2} \left( \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_2} \Bigg|_{\theta_0} \right) + o_P(1) \]

\[ = I^{\theta_1 \theta_1} \left( \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \Bigg|_{\theta_0} \right) - I_{\theta_1 \theta_2} I_{\theta_2 \theta_2}^{-1} \left( \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta_2} \Bigg|_{\theta_0} \right) + o_P(1) \]

\[ = I^{\theta_1 \theta_1} [A - I_{\theta_1 \theta_2} I_{\theta_2 \theta_2}^{-1} B] + o_P(1). \]
Substituting this into expressions for the Wald statistic, we get:

\[
W = \sqrt{T} (\hat{\theta}_1 - \bar{\theta}_1)' I^{\theta_1 \theta_1} \sqrt{T} (\hat{\theta}_1 - \bar{\theta}_1) \\
= [A - I_{\theta_1 \theta_2} I_{\theta_2 \theta_2}^{-1} B]' I^{\theta_1 \theta_1} [A - I_{\theta_1 \theta_2} I_{\theta_2 \theta_2}^{-1} B] + o_P(1)
\]

\[\iff\] LM (Rao/Score)

Thus, along with the result in Section 2.3.1, we have even for the composite hypothesis case, asymptotically:

\[LR \iff W \iff LM \text{ (Rao)}\]
3 Application to linear regression model

In this section, we look at some analogies between standard MLE results derived in the earlier sections and OLS regression results. (Recall that we already look at some analogies between OLS and MLE in Section 2.2 above.)

First, we show analogy between OLS and MLE for estimation of parameter sub-vectors. Then we derive expressions for the three common test statistics and establish certain results for these.
3.1 Estimation of parameter subsets

Key result in regression is Theorem of the Partitioned Inverse:

\[
A = \begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix}
\]

assume nonsingular,

Assume \(A_{11}\) is nonsingular so \(A_{22} - A_{21}A_{11}^{-1}A_{12} = D\).

Then

\[
A^{-1} = \begin{bmatrix}
A_{11}^{-1}(I + A_{12}D^{-1}A_{21}A_{11}) & -A_{11}^{-1}A_{12}D^{-1} \\
-D^{-1}A_{21}A_{11}^{-1} & D^{-1}
\end{bmatrix}
\]

Proof: Multiply it out.
In OLS case, it produces a useful result (Frisch-Waugh-Goldberger):

\[
X'X = \begin{bmatrix}
X'_1X_1 & X'_1X_2 \\
X'_2X_1 & X'_2X_2
\end{bmatrix}
\]

In the OLS case we partition the matrix of independent variables \(X\) as

\[
X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}
\]

and also

\[
X'X = \begin{pmatrix} X'_1X_1 & X'_1X_2 \\
X'_1X_2 & X'_2X_2 \end{pmatrix}.
\]
We define:

\[
M_1 \equiv I - X_1(X'_1X_1)^{-1}X'_1 \\
D \equiv X'_2X_2 - X'_2X_1(X'_1X_1)^{-1}X'_1X_2 \\
= X'_2M_1X_2
\]

Now we have result for OLS regression:

\[
\hat{\beta}_{ols} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \\
= (X'X)^{-1}(X'Y) \\
= \begin{bmatrix} X'_1X_1 & X'_1X_2 \\ X_1X'_2 & X'_2X_2 \end{bmatrix}^{-1} \begin{bmatrix} X'_1Y \\ X'_2Y \end{bmatrix}
\] (8)
Then using the results from theorem of partitioned inverse above and simplifying, we can write:

\[
\hat{\beta}_1 = (X'_1X_1)^{-1}(X'_1Y) + (X'_1X_1)^{-1}X'_1X_2D^{-1}X'_2X_1(X'_1X_1)^{-1}X'_1Y - (X'_1X_1)^{-1}X'_1X_2D^{-1}X'_2Y
= (X'_1X_1)^{-1}X'_1Y - (X'_1X_1)^{-1}(X'_1X_2)D^{-1}X'_2M_1Y
\]

\[
\hat{\beta}_2 = -D^{-1}X'_2X_1(X'_1X_1)^{-1}X'_1Y + D^{-1}X'_2Y
= D^{-1}X'_2(I - X_1(X'_1X_1)^{-1}X_1)Y
= D^{-1}X'_2M_1Y
\]  

(9)

This leads to a Double Residual Regression Result. \(\hat{\beta}_2\): Regress \(Y\) on \(X_1\), \(X_2\) on \(X_1\); form residuals. Regress one set of residuals on another.
Result: $\beta_2$ is the regression of “cleaned out $Y$” on “cleaned out $X_2$”

This result follows directly from the results (8) and (9) above. Define:

“Cleaned out $Y$” $\equiv Y^* = [I - X_1(X_1'X_1)^{-1}X_1']Y = M_1Y$
(Residual from the regression of $Y$ on $X_1$)

“Cleaned out $X_2$” $\equiv X_2^* = [I - X_1(X_1'X_1)^{-1}X_1']X_2 = M_1X_2$
(Residual from the regression of $X_2$ on $X_1$)
Then from (9) above we have:

\[ \hat{\beta}_2 = D^{-1}X'_2M_1Y \]
\[ = [X'_2M_1X_2]^{-1}(X'_2M_1Y) \]
\[ = [X'_2X'_2]^{-1}(X'_2Y'). \]

Note that we really don’t have to clean out \( Y \), just \( X \) since \( M_1 \) is idempotent. Further, if \( X_1 \) and \( X_2 \) are orthogonal (uncorrelated), then \( X'_2 = X_2 \), and hence unbiased/consistent estimate of \( \hat{\beta}_2 \) can be obtained by directly regressing \( X_2 \) on \( Y \). (Recall that we derived the same result for MLE in Section 2.2.)
Also observe (derived directly from $\hat{Y} = X_1\hat{\beta}_1 + X_2\hat{\beta}_2$) that:

$$\hat{Y}'\hat{Y} = Y'X_1(X'_1X_1)^{-1}X'_1Y_1 + Y'M_1X_2D^{-1}X'_2M_1Y$$

Part due to $X_1$  \hspace{4cm} Part due to orthogonalized $X_2$

Thus, unless regressors are orthogonal, there are no unique contributions.

**Proof.**

1. Observe that:

$$M_1X_1 = 0 \hspace{1cm} M_2X_2 = 0$$

$$M_1M_X = M_X \hspace{1cm} M_2M_X = M_X$$

$$Y \equiv P_XY + M_XY = \left(X_1\hat{\beta}_1 + X_2\hat{\beta}_2\right) + M_XY$$
2. Observe that:

\[ M_1 X_2 = (I - P_{X_1}) X_2 = \text{linear combination of } X_2 \]

\[ \implies M_X \cdot M_1 X_2 = 0. \]

3. 

\[
X'_2 M_1 Y = X_2 M_1 X_1 \beta_1 + X'_2 M_1 X_2 \hat{\beta}_2 + X'_2 M_2 M_X Y \\
= X_2 M_1 X_1 \beta_1 + X'_2 M_1 X_2 \hat{\beta}_2 + Y' (M_X M_1 X_2),
\]

where \( X_2 M_1 X_1 \beta_1 \to 0 \) and \( M_X M_1 X_2 \to 0 \). Thus,

\[
\hat{\beta}_2 = (X'_2 M_1 X_2)^{-1} (X'_2 M_1 Y).
\]
4. Observe that $\hat{\beta}_2$ is the result of the regression

$$M_1Y = M_1X_2\hat{\beta}_2 + \text{error}_2,$$

but

$$Y = X_1\beta_1 + X_2\beta_2 + M_XY$$

$$\implies M_1Y = M_1X_1\beta_1 + M_1X_2\beta_2 + M_1M_XY$$

$$\implies M_1Y = M_1X_2\beta_2 + M_XY \therefore M_1X_1\beta_1 \to 0$$

$$\implies \text{error}_2 = M_XY.$$

(from Davidson & MacKinon)
Analogously, for MLE in neighborhood of optimum we have that:

$$\sqrt{T} (\hat{\theta} - \theta_0) \approx \left( \frac{\partial \ln \mathcal{L}}{\partial \theta} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta} \right)' \right)^{-1} \frac{1}{\sqrt{T}} \frac{\partial \ln \mathcal{L}}{\partial \theta} \nu_T + o_P(1)$$
\[ T^{-1} \left( \begin{array}{cc}
\frac{\partial \ln \mathcal{L}}{\partial \theta_1} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \right)', & \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_2} \right)'
\end{array} \right) \times \frac{1}{\sqrt{T}} \left( \begin{array}{c}
\frac{\partial \ln \mathcal{L}}{\partial \theta_1} \\
\frac{\partial \ln \mathcal{L}}{\partial \theta_2}
\end{array} \right) \nu'_T + o_P(1), \]

where \( \nu_T \) is a \( 1 \times T \) vector of ones, i.e. \( \nu_T = (1 1 1 1 \ldots 1) \).
Comparing above equation to the result for the standard OLS regression (see eqn (8) above), we note that the MLE result is analogous to regressing on the score vector $\nu_T$ (a vector of ones); i.e., we have that:

$$[X_1' \ X_2] \longleftrightarrow \begin{bmatrix} \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \right) & \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_2} \right)' \end{bmatrix} \quad \text{and} \quad Y \leftrightarrow \nu'_T.$$
Further, we also get:

\[
\sqrt{T} \left( \hat{\theta}_2 - \bar{\theta}_2 \right) = \left[ \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_2} \right)' \left[ I - \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \left( \frac{\partial \ln \mathcal{L}'}{\partial \theta_1} \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \right)^{-1} \frac{\partial \ln \mathcal{L}'}{\partial \theta_1} \right] \frac{\partial \ln \mathcal{L}}{\partial \theta_2} \right]^{-1} \\
\times \left( I - \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \left[ \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \right)' \frac{\partial \ln \mathcal{L}}{\partial \theta_1} \right]^{-1} \frac{\partial \ln \mathcal{L}'}{\partial \theta_1} \right) \nu'_T + o_P(1),
\]

which is analogous to regression result (9) above.
3.2 Hypothesis testing results for the linear regression model

Consider $MLE$ in classical linear regression model.

$$Y_t = X_t \beta + U_t$$

$$U_t \sim N \left( 0, \sigma^2_U \right) \text{ i.i.d.},$$

i.e.,

$$f(U_t) = \frac{1}{\sqrt{2\pi\sigma^2_U}} \exp \left( -\frac{1}{2} \frac{U_t^2}{\sigma^2_U} \right)$$

$$\implies$$

$$f \left( Y_t \mid X_t, \beta \right) = f \left( Y_t - X_t \beta \mid X_t \beta \right)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2_U}} \exp \left( -\frac{1}{2\sigma^2_U} (Y_t - X_t \beta)^2 \right).$$
With i.i.d. sampling, we have the log likelihood function:

$$\ln \mathcal{L} = -\frac{T}{2} \ln 2\pi - \frac{T}{2} \ln \sigma^2_U - \frac{1}{2\sigma^2_U} \sum_{t=1}^{T} (Y_t - X_t\beta)^2.$$
3.2.1 MLE Solution

\[
\frac{\partial \ln \mathcal{L}}{\partial \beta} = \sum U_t X_t = 0
\]

at optimum yields:

\[
\hat{\beta} = \left( \sum X_t' X_t \right)^{-1} \sum X_t' Y_t
\]

\[
\hat{\sigma}_U^2 = \frac{1}{T} \sum_{t=1}^{T} (Y_t - X_t \hat{\beta})^2
\]
\[
\frac{1}{T} \frac{\partial^2 \ln \mathcal{L}}{\partial \beta \partial \beta'} = -\frac{(X'X)}{T} \frac{1}{\sigma^2} \\
\frac{1}{T} \frac{\partial^2 \ln \mathcal{L}}{\partial (\sigma^2_U) \partial \beta} = \frac{\partial}{\partial \beta} \left[ -\frac{T}{2} \frac{1}{\sigma^2_U} + \frac{1}{2(\sigma^2_U)^2} \sum (Y_t - X_t\hat{\beta})^2 \right] \\
= \frac{1}{(2\sigma^2_U)^2} \sum (Y_t - X_t\hat{\beta})'X_t = 0
\]

(This holds unless we have that the model is such that there’s a functional relationship between \(\sigma^2_U\) and \(\beta\) – this we exclude by assumption.)
\[
\frac{1}{T} \frac{\partial^2 \ln \mathcal{L}}{\partial (\sigma_U^2)^2} = -\frac{T}{2(\hat{\sigma}_U^2)^2}
\]

This yields the information matrix:

\[
I_{\theta_0} = \begin{pmatrix}
\frac{1}{\sigma_U^2} X' X & 0 \\
0 & \frac{1}{2\sigma_U^4}
\end{pmatrix},
\]

which is $\beta$ and $\sigma_U$ block diagonal.

(By previous result we may ignore parameter estimation error between $\widehat{\beta}$ and $\widehat{\sigma}^2$).
3.2.2 Expressions for the various test statistics

Let: \( \beta = (\beta_1, \beta_2) \) and \( H_0: \beta_2 = 0 \).

Then the expression for the Likelihood Ratio test statistic is given by:

\[
LR = 2 \ln \left[ \frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\tilde{\theta})} \right],
\]

where \( \mathcal{L}(\hat{\theta}) \) is the \textit{unrestricted} maximized likelihood function and \( \mathcal{L}(\tilde{\theta}) \) is the \textit{restricted} maximized likelihood function.
We have:

\[
\ln \mathcal{L}(\tilde{\theta}) = -\frac{T}{2} \ln \tilde{\sigma}^2_U - \frac{1}{2\tilde{\sigma}^2_U} \sum_{t=1}^{T} (Y_t - X_{1t}\tilde{\beta}_1)^2
\]

\[
= -\frac{T}{2} \ln \left( \frac{\sum (Y_t - X_{1t}\tilde{\beta}_1)^2}{T} \right),
\]

using fact \(\tilde{\sigma}^2_U = \frac{1}{T} \sum_{t=1}^{T} (Y_t - X_{1t}\tilde{\beta}_1)^2\).
Similarly,

\[
\ln \mathcal{L}(\hat{\theta}) = -\frac{T}{2} \ln \left( \frac{\sum (Y_t - X_{1t}\hat{\beta}_1 - X_{2t}\hat{\beta}_2)^2}{T} \right)
\]

\[
\implies
\]

\[
LR = 2 \ln \left[ \frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\hat{\theta}_0)} \right]
\]

\[
= \frac{T}{2} \ln \left( \frac{Y' [I - X_1(X_1'X_1)^{-1}] Y}{Y' [I - X(X'X)^{-1}X'] Y} \right)
\]

(10)
The expression for the Wald Test statistic is obtained using partitioned inverse theorem on information matrix above and standard asymptotic normality result for MLE:

$$\sqrt{T}(\hat{\beta}_2 - \bar{\beta}_2) \sim N(0, \sigma^2 (X_2'[I - X_1(X_1'X_1)^{-1}X_1']X_2)^{-1})$$

$$\therefore \quad W = \sqrt{T} \left( \hat{\beta}_2 - \beta_2^H \right)' \left[ \frac{X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2}{\sigma^2 U T} \right]$$

$$\times \sqrt{T} \left( \hat{\beta}_2 - \beta_2^H \right),$$

where $\beta_2^H$ is the hypothesized value of $\beta_2$. 
The test statistic for the Rao (LM/Score) is obtained by computing the derivative of the log likelihood function at the point where we have that $\beta_2 = 0$ is imposed. We get:

$$
\ln \mathcal{L}(\theta_0) = -\frac{T}{2} \ln \sigma_U^2 - \sum \frac{(Y_{1t} - X_{1t}\beta_1 - X_{2t}\beta_2)^2}{2\sigma_U^2}
$$

$$
\Rightarrow 
\frac{1}{T} \frac{\partial \ln \mathcal{L}(\theta_0)}{\partial \beta_2} \bigg|_{\beta_2 = 0} = \frac{1}{T} \sum \frac{(Y_{1t} - X_{1t}\tilde{\beta}_1) X_{2t}}{2\tilde{\sigma}_U^2}
$$
As discussed in earlier sections, the Rao test statistic can now be formed by examining the asymptotic distribution of the score vector. In this linear regression context, the Rao test can be motivated in another intuitive way.

Basically, the Rao test considers the restrictions to be valid if the score $\approx 0$ (i.e., the log likelihood function is maximized) when the restrictions are imposed. Looking at the rhs of the expression above, in the linear regression case, this is equivalent to checking if the residual from the restricted MLE

$$\tilde{\varepsilon}_t = (Y_{1t} - X_{1t}\tilde{\beta}_1)$$

and $X_{2t}$ are orthogonal!
Accordingly, we have the Rao test statistic:

\[
LM = \left( \frac{1}{T} \sum \tilde{\varepsilon}_t X_{2t} \right)' \left( I_{\beta_2 \beta_2} \right) \left( \frac{1}{T} \sum \tilde{\varepsilon}_t X_{2t} \right) \\
= \left( \frac{1}{T} \sum \tilde{\varepsilon}_t X_{2t} \right)' \tilde{\sigma}_U^2 \left( X_2' (I - X_1 (X_1' X_1)^{-1} X_1') X_2 \right)^{-1} \\
\times \left( \frac{1}{T} \sum \tilde{\varepsilon}_t X_{2t} \right)
\]
Observe a feature of Rao test: Is it equivalent to regressing $M_1Y_1$ on $X_2$ and testing for statistical significance? The answer is: not in general, as shown below. Regressing $M_1Y_1$ on $X_2$ gives:

\[
M_1Y_1 = X_2\Pi + V \\
\hat{\Pi} = (X_2'X_2)^{-1}X_2'M_1Y
\]
OLS test of $\Pi = 0$ yields statistic:

$$T \hat{\Pi}'(X'_2 X_2) \hat{\Pi}$$

$$= T \frac{Y' M_1 X_2 (X'_2 X_2)^{-1} (X'_2 X_2) (X'_2 X_2)^{-1} X'_2 M_1 Y}{(Y - X_2 (X'_2 X_2)^{-1} X'_2 M_1 Y)'(Y - X_2 (X'_2 X_2)^{-1} X'_2 M_1 Y)}$$

$$= T \frac{Y' M_1 X_2 (X'_2 X_2)^{-1} X'_2 M_1 Y}{Y' [I - M_1 X_2 (X'_2 X_2)^{-1} X'_2] [I - X_2 (X'_2 X_2)^{-1} X'_2 M_1] Y}$$

$$= T \frac{Y' M_1 X_2 (X'_2 X_2)^{-1} X'_2 M_1 Y}{Y' I + M_1 X_2 (X'_2 X_2)^{-1} X'_2 M_1 - M_1 X_2 (X'_2 X_2)^{-1} X'_2 - X_2 (X'_2 X_2)^{-1} X'_2 M_1} Y$$
Whereas the Rao statistic as derived earlier is:

\[
LM = \frac{TY' M_1 X_2 (X'_2X_2)^{-1} X'_2 M_1 Y}{k Y' M_1 M_2 M_1 Y}.
\]
So only if \( M_1 X_2 = X_2 \) (i.e., \( X_1 \perp X_2 \)) are these equal, not in general.

Note that if we regress \( MY_1 \) on \( X_1 \) and \( X_2 \), then testing if \( X_2 \) in this regression is significant is equivalent to the Rao test. That is, if we set the model as:

\[
M_1 Y_1 = X_1 \Pi_1 + X_2 \Pi_2 + \hat{V} \\
\hat{\Pi}_2 = (X_2' M_1 X_2)^{-1} X_2' M_1 Y.
\]

Then the OLS test of \( \Pi_2 = 0 \) is equivalent to the Rao test (left as exercise for the reader).
3.2.3 Relationship between the various tests

While we demonstrated in earlier sections that, asymptotically, the three tests are equivalent, in this special linear regression case we can show that $W > LR > LM$ so that the Wald test is the most conservative (most likely to reject the null) and the Rao test is the least conservative (least likely to reject the null) in small/finite samples.

Define:

$$ESS(U) \equiv \text{Expected Sum of Squared Residuals (Unrestricted)}$$
$$= (Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2)'(Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2)$$

$$ESS(R) \equiv \text{Expected Sum of Squared Residuals (Restricted)}$$
$$= (Y - \tilde{\beta}_1 X_1)'(Y - \tilde{\beta}_1 X_1)$$
Using the results in the earlier subsections of Section 3, and from the expressions for Wald, LR and LM test statistics derived above it can be shown (left as exercise for the reader) that:

\[ W = T \left[ \frac{ESS(R) - ESS(U)}{ESS(U)} \right] \]

\[ LR = T \ln \left[ \frac{ESS(R)}{ESS(U)} \right] \]

\[ LM \ (Rao) = T \left[ \frac{ESS(R) - ESS(U)}{ESS(R)} \right] \]
1. Proof that $W \geq LR$:

Define $b \equiv \frac{ESS(R)}{ESS(U)}$

$$\implies LR = T \ln b, \quad W = T(b - 1)$$

Now $b - 1 \geq \ln b$ for $b \geq 1 \implies W \geq LR$

2. Proof that $LR \geq \text{Rao}$:

Define $c \equiv \frac{ESS(U)}{ESS(R)} \equiv \frac{1}{b}$.

$$\implies LM (\text{Rao}) = T[1 - c], \quad LR = -T \ln c$$

Now $-\ln c \geq 1 - c$ (since $c \leq 1$ always) $\implies LR \geq \text{Rao}$. 