University of Pavia

Maximum Likelihood Statistical Inference

Eduardo Rossi

Parameter vector partitioned into

$$\boldsymbol{\theta} = [\boldsymbol{\theta}_1', \boldsymbol{\theta}_2']'$$

Given any $\boldsymbol{\theta}_2$ one can find the optimal value of $\boldsymbol{\theta}_1$ as a function of $\boldsymbol{\theta}_2$ by solving f.o.c.:

$$\left.\frac{\partial E_N[L(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}_1}\right|_{\boldsymbol{\theta}_1 = \widehat{\boldsymbol{\theta}}_1} \equiv E_N[L_1(\widehat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2)] = \mathbf{0} \Leftrightarrow \widehat{\boldsymbol{\theta}}_1 = \widehat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)$$

substituting this function into the original log-likelihood yields the concentrated average log-likelihood function

$$E_N[L^c(\boldsymbol{\theta}_2)] \equiv E_N[L(\widehat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2), \boldsymbol{\theta}_2)]$$

which yields the MLE for $\boldsymbol{\theta}_2$

$$\widehat{\boldsymbol{\theta}}_2 = \arg\max_{\boldsymbol{\theta}_2} E_N[L^c(\boldsymbol{\theta}_2)]$$

Consider the model for which the MLE for the variance can be expressed as a function of the MLE for $\boldsymbol{\beta}$:

$$\begin{aligned} 0 &= E_N[L_{\sigma^2}(\widehat{\boldsymbol{\theta}}; y_t|\mathbf{x}_t)] \\ &= -\frac{1}{2[\widehat{\sigma}^2(\boldsymbol{\beta})]^2}\left[\widehat{\sigma}^2(\boldsymbol{\beta}) - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{N}\right] \end{aligned}$$

so that

$$\widehat{\sigma}^2(\boldsymbol{\beta}) = \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{N}$$

substituting into the log-lik function we obtain the concentrated log-lik function

$$E_N[L(\boldsymbol{\theta})] = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/N}{2\sigma^2}$$

$$E_N[L^c(\boldsymbol{\beta}; y_t|\mathbf{x}_t)] = -\frac{1}{2}\left\{\log\left[2\pi\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{N}\right] + 1\right\}$$

In this way the dimension of maximization problem has been reduced by one and the maximization of the log-lik function has become much simpler. The concentrated log-lik is a monotonic transformation of a quadratic problem.

Asymptotically the concentrated log-lik has properties similar to an ordinary log-likelihood. We replace expectations with the probability limits of empirical moments.

Let

$$L^c(\boldsymbol{\theta}_2; u) \equiv L(\boldsymbol{\theta}_c; u)$$

where

$$\boldsymbol{\theta}_c \equiv [\widehat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)', \boldsymbol{\theta}_2']'$$

Under the assumptions for MLE asymptotic results

$$E\left[\sup_{\boldsymbol{\theta}\in\Theta} |L_{\boldsymbol{\theta}}(\boldsymbol{\theta}; U)|\right] \text{ exists}$$

then

$$E_N[L^c_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_{02})] \xrightarrow{\text{p}} \mathbf{0}$$

$$E_N[L^c_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_{02})L^c_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_{02})'] + E_N[L^c_{\boldsymbol{\theta}_2\boldsymbol{\theta}_2}(\boldsymbol{\theta}_{02})] \xrightarrow{\text{p}} \mathbf{0}$$

and

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{02}) \xrightarrow{\text{d}} N(\mathbf{0}, \mathfrak{I}_{2|1}(\boldsymbol{\theta}_0)^{-1})$$

where

$$E_N[L^c_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_{02})L^c_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_{02})'] \xrightarrow{\text{p}} \mathfrak{I}_{2|1}(\boldsymbol{\theta}_0) \equiv \mathfrak{I}_{22}(\boldsymbol{\theta}_0) - \mathfrak{I}_{21}(\boldsymbol{\theta}_0)\mathfrak{I}_{11}(\boldsymbol{\theta}_0)^{-1}\mathfrak{I}_{12}(\boldsymbol{\theta}_0).$$

For practical purposes we may treat the concentrated log-lik as though it were an ordinary log-likelihood for all calculations.
We may interpret $\mathfrak{I}_{2|1}(\theta_0)$ as the asymptotic conditional variance of $\sqrt{N}E_N[L_2(\boldsymbol{\theta}_0)]$ given $\sqrt{N}E_N[L_1(\boldsymbol{\theta}_0)]$. They have a joint multivariate normal distribution.

To factorize the joint density into conditional and marginal components, remember that for $\mathbf{x} \sim \mathbf{N}(\mu, \mathbf{\Sigma})$, where

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \begin{matrix} (n_1 \times 1) \\ (n_2 \times 1) \end{matrix}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \begin{matrix} (n_1 \times 1) \\ (n_2 \times 1) \end{matrix}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}$$

$$\mathbf{x}_1 | \mathbf{x}_2 \sim N[\boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}, \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}]$$

$$\mathbf{x}_2 \sim N[\boldsymbol{\mu}_2, \mathbf{\Sigma}_{22}]$$

Because

$$\left.\frac{\partial E_N[L(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}_1}\right|_{\boldsymbol{\theta}_1=\widehat{\boldsymbol{\theta}}_1} \equiv E_N[L_1(\widehat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2)] = \mathbf{0}$$

$$\begin{aligned}\frac{\partial L^c(\boldsymbol{\theta}_2; u)}{\partial \boldsymbol{\theta}_2} \equiv \frac{\partial L(\boldsymbol{\theta}_c; u)}{\partial \boldsymbol{\theta}_2} \\ &= L_2(\boldsymbol{\theta}_c; u) + \left[\frac{\partial}{\partial \boldsymbol{\theta}_2}\widehat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)'\right] L_1(\boldsymbol{\theta}_c; u) \\ &= L_2(\boldsymbol{\theta}_c; u)\end{aligned}$$

the score of the concentrated log-likelihood is L_2 evaluated at a point for which $L_1 \equiv \mathbf{0}$.

The asymptotic distribution theory reflects this conditional fact.

The score and the Hessian of the average concentrated log-lik function are

$$E_N[L^c_{\boldsymbol{\beta}}(\boldsymbol{\beta}; y_t|\mathbf{x}_t)] = \frac{2}{\widehat{\sigma}^2(\boldsymbol{\beta})} E_N[\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})]$$

$$\begin{aligned} E_N[L^c_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}; y_t|\mathbf{x}_t)] &= -\frac{2}{\widehat{\sigma}^2(\boldsymbol{\beta})} E_N[\mathbf{x}_t\mathbf{x}_t'] + \\ & \frac{1}{\widehat{\sigma}^2(\boldsymbol{\beta})} E_N[\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})] E_N[(y_t - \mathbf{x}_t'\boldsymbol{\beta})\mathbf{x}_t'] \end{aligned}$$

If we replace expectations with probability limits then we obtain an analogy to the score identity. Note that

$$\sigma^2(\boldsymbol{\beta}_0) \xrightarrow{p} \sigma_0^2$$

and

$$E_N[L^c_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0; y_t|\mathbf{x}_t)] = \frac{2}{\widehat{\sigma}^2(\boldsymbol{\beta}_0)} E_N[\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}_0)] \xrightarrow{p} \mathbf{0}$$

An analogy to the information matrix is

$$Var_N[L^c_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0; y_t|\mathbf{x}_t)] \xrightarrow{\text{p}} \frac{1}{\sigma_0^2} E[\mathbf{x}_t\mathbf{x}_t']$$

which is $\Im_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)$. Furthermore

$$E_N[L^c_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0; y_t|\mathbf{x}_t)] \xrightarrow{\text{p}} -\frac{1}{\sigma_0^2} E[\mathbf{x}_t\mathbf{x}_t']$$

As expected $\sqrt{N}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0)$ has an asymptotic variance equal to $\sigma_0^2 E[\mathbf{x}_t\mathbf{x}_t']^{-1}$.

Fewer parameters can be estimated more efficiently, when the restrictions imposed to reduce the number of parameters are correct. Compare the restricted MLE with the unrestricted MLE

$$\boldsymbol{\theta} = [\boldsymbol{\theta}_1', \boldsymbol{\theta}_2']'$$

$$\Theta = \Theta_1 \times \Theta_2$$

Restrictions $\boldsymbol{\theta}_2 = \mathbf{0}$.

The restricted MLE

$$\widetilde{\boldsymbol{\theta}} = \begin{bmatrix} \widetilde{\boldsymbol{\theta}}_1 \\ \widetilde{\boldsymbol{\theta}}_2 \end{bmatrix} = \begin{bmatrix} \arg\max_{\boldsymbol{\theta}\in\Theta:\boldsymbol{\theta}_2=\mathbf{0}} E_N[L(\boldsymbol{\theta})] \\ \mathbf{0} \end{bmatrix}$$

$$= \begin{bmatrix} \arg\max_{\boldsymbol{\theta}_1} E_N[L(\boldsymbol{\theta}_1, \mathbf{0})] \\ \mathbf{0} \end{bmatrix}$$

the asymptotic covariance matrix is

$$\mathbf{V}_R = \begin{bmatrix} \Im_{11}(\boldsymbol{\theta}_0)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Compared to the asymptotic variance of the unrestricted MLE, $\Im(\boldsymbol{\theta}_0)^{-1}$, the $\mathbf{V}_R$ is smaller in the p.s.d. sense. In fact

$$\Im(\boldsymbol{\theta}_0)[\Im(\boldsymbol{\theta}_0)^{-1} - \mathbf{V}_R]\Im(\boldsymbol{\theta}_0) = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Im_{22}(\boldsymbol{\theta}_0) - \Im_{21}(\boldsymbol{\theta}_0)\Im_{11}(\boldsymbol{\theta}_0)^{-1}\Im_{12}(\boldsymbol{\theta}_0) \end{bmatrix}$$

the matrix

$$\Im_{22}(\boldsymbol{\theta}_0) - \Im_{21}(\boldsymbol{\theta}_0)\Im_{11}(\boldsymbol{\theta}_0)^{-1}\Im_{12}(\boldsymbol{\theta}_0)$$

is p.s.d., i.e. is the $Var[\boldsymbol{\theta}_2|\boldsymbol{\theta}_1]$.

Three methods for computing hypothesis test statistics in the likelihood framework:

- Wald test
- Score test
- Likelihood Ratio (LR) test

$$H_0 : \mathbf{r}(\boldsymbol{\theta}_0) = \mathbf{0}$$

where $\mathbf{r} : \mathbb{R}^K \to \mathbb{R}^{K-M}$ is twice continuously differentiable function and its partial derivative

$$\mathbf{r}_{\boldsymbol{\theta}} : \quad (K - M) \times K$$

Estimation under both the null and alternative hypothesis is necessary. The LR test compares the goodness-of-fit of the unrestricted and restricted models using as a criterion of g-o-f the log-likelihood function.

1. Compute the restricted MLE

$$\widehat{\boldsymbol{\theta}}_R \equiv \arg \max_{\boldsymbol{\theta} \in \Theta : \mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}} E_N[L(\boldsymbol{\theta})]$$

2. Compute the log-likelihood function:

$$L(\widehat{\boldsymbol{\theta}}_R; U_1, \ldots, U_N) = N E_N[L(\widehat{\boldsymbol{\theta}}_R)]$$

3. compute the unrestricted MLE $\widehat{\boldsymbol{\theta}}$ and the value of the log-likelihood function

$$L(\widehat{\boldsymbol{\theta}}; U_1, \ldots, U_N) = N E_N[L(\widehat{\boldsymbol{\theta}})]$$

LR test statistic is

$$\begin{aligned}\mathcal{LR} &\equiv 2N\left\{\max_{\boldsymbol{\theta}\in\Theta} E_N[L(\boldsymbol{\theta})] - \max_{\boldsymbol{\theta}\in\Theta:\mathbf{r}(\boldsymbol{\theta})=\mathbf{0}} E_N[L(\boldsymbol{\theta})]\right\} \\ &= 2N\left\{E_N[L(\widehat{\boldsymbol{\theta}})] - E_N[L(\widehat{\boldsymbol{\theta}}_R)]\right\} \\ &= 2[L(\widehat{\boldsymbol{\theta}};U_1,\ldots,U_N) - L(\widehat{\boldsymbol{\theta}}_R;U_1,\ldots,U_N)] \xrightarrow{d} \chi^2_{K-M}\end{aligned}$$

Compare $\mathcal{LR}$ with the critical value of χ^2_{K-M}.

Linear restrictions:

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{c}$$

$$\mathbf{y}|\mathbf{X} \sim N(\mathbf{x}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N)$$

$$\text{rank}(\mathbf{R}) = r$$

Because

$$\widehat{\boldsymbol{\beta}}|\mathbf{X} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

under H_0

$$\mathbf{R}\widehat{\boldsymbol{\beta}} \sim N[\mathbf{c}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']$$

when σ^2 is known we compute the test statistic recalling that

$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})$$

$$
\begin{aligned}
(\mathbf{R}\widehat{\boldsymbol{\beta}}-\mathbf{c})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}}-\mathbf{c}) &= (\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}) \\
&= \|\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\|^2_{\mathbf{X}'\mathbf{X}} \\
&= \|\mathbf{y}-\widehat{\boldsymbol{\mu}}_R\|^2 - \|\mathbf{y}-\widehat{\boldsymbol{\mu}}\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\frac{(\mathbf{R}\widehat{\boldsymbol{\beta}}-\mathbf{c})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}}-\mathbf{c})}{\sigma^2} \\
&= \frac{1}{\sigma^2}\{(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) - (\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})\} \\
&= -\frac{1}{\sigma^2}\{(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}) - (\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})\} \\
&= 2[L(\widehat{\boldsymbol{\beta}},\sigma^2;\mathbf{y}|\mathbf{X}) - L(\widetilde{\boldsymbol{\beta}},\sigma^2;\mathbf{y}|\mathbf{X})] \\
&= \mathcal{LR}
\end{aligned}
$$

The Wald test compares the unrestricted estimator with the values specified by the null hypothesis in a quadratic form normalized with the inverse of the variance matrix of the estimator.

1. compute the unrestricted MLE

$$\widehat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} E_N[L(\boldsymbol{\theta})]$$

2. compute an estimator of the variance matrix of the asymptotic distribution of $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, for example the $\mathcal{I}(\widehat{\boldsymbol{\theta}})^{-1}$
3. compute the unrestricted estimator of $\mathbf{r}(\boldsymbol{\theta}_0)$, $\mathbf{r}(\widehat{\boldsymbol{\theta}})$, and of $\mathbf{r}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)$, $\mathbf{r}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})$
4. compute the quadratic form

$$\mathcal{W} = N\mathbf{r}(\widehat{\boldsymbol{\theta}})' \left[\mathbf{r}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})\mathfrak{I}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{r}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})'\right]^{-1} \mathbf{r}(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_{K-M}$$

5. compare $\mathcal{W}$ with the critical value of χ^2_{K-M}.

The null hypothesis as a restriction on a subset of the parameter vector $\boldsymbol{\theta} = [\boldsymbol{\theta}_1', \boldsymbol{\theta}_2']'$, $H_0 : \boldsymbol{\theta}_{02} = \mathbf{0}$. $\boldsymbol{\theta}_2 : (K - M \times K)$ so that under H_0 there are M unknown parameters.

1. compute the unrestricted MLE

$$\widehat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} E_N[L(\boldsymbol{\theta})]$$

2. compute an estimator of the variance matrix of the asymptotic distribution of $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, $(\mathfrak{I}(\widehat{\boldsymbol{\theta}})^{-1})$.

3. compute

$$\mathcal{W} = N\widehat{\boldsymbol{\theta}}_2'\widehat{\mathbf{V}}_W^{-1}\widehat{\boldsymbol{\theta}}_2$$

where $\widehat{\mathbf{V}}_W$ is the $(2,2)$ block of $[\mathfrak{I}(\widehat{\boldsymbol{\theta}})]^{-1}$ partitioned conformably with $\boldsymbol{\theta}$:

$$\widehat{\mathbf{V}}_W = \left\{\mathfrak{I}_{22}(\widehat{\boldsymbol{\theta}}) - \mathfrak{I}_{21}(\widehat{\boldsymbol{\theta}})\mathfrak{I}_{11}(\widehat{\boldsymbol{\theta}})^{-1}\mathfrak{I}_{12}(\widehat{\boldsymbol{\theta}})\right\}^{-1}$$

The limiting distribution of $\mathcal{W}$ is obtained from

$$\widehat{\mathbf{V}}_W^{-1/2}\sqrt{N}\widehat{\boldsymbol{\theta}}_2 \xrightarrow{\text{d}} N(\mathbf{0}, \mathbf{I}_{K-M})$$

under $H_0 : \boldsymbol{\theta}_{02} = \mathbf{0}$. Therefore

$$\mathcal{W} = \left(\widehat{\mathbf{V}}_W^{-1/2}\sqrt{N}\widehat{\boldsymbol{\theta}}_2\right)' \widehat{\mathbf{V}}_W^{-1/2}\sqrt{N}\widehat{\boldsymbol{\theta}}_2 = N\widehat{\boldsymbol{\theta}}_2'\widehat{\mathbf{V}}_W\widehat{\boldsymbol{\theta}}_2 \xrightarrow{\text{d}} \chi^2_{K-M}$$

by the Continuous Mapping Theorem.

This statistic is invariant to reparameterizations of $\boldsymbol{\theta}$, but not of $\mathbf{r}(\boldsymbol{\theta})$. Given any set equivalent of restrictions

$$\mathbf{g}[\mathbf{r}(\boldsymbol{\theta})] - \mathbf{g}[\mathbf{0}] \equiv \mathbf{s}(\boldsymbol{\theta}) = \mathbf{0}$$

the alternative Wald statistic is

$$\mathcal{W}_a = N\mathbf{s}(\widehat{\boldsymbol{\theta}})' \left[\mathbf{s}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})\mathfrak{I}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{s}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})'\right]^{-1} \mathbf{s}(\widehat{\boldsymbol{\theta}})$$

because

$$\mathbf{s}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathbf{g}_r[\mathbf{r}(\boldsymbol{\theta})]\mathbf{r}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

$$\begin{aligned}\mathcal{W}_a &= N\mathbf{s}(\widehat{\boldsymbol{\theta}})' \left[(\mathbf{g}_r[\mathbf{r}_{\boldsymbol{\theta}}(\boldsymbol{\theta})]\mathbf{r}_{\boldsymbol{\theta}}(\boldsymbol{\theta}))\mathfrak{I}(\widehat{\boldsymbol{\theta}})^{-1}(\mathbf{g}_r[\mathbf{r}_{\boldsymbol{\theta}}(\boldsymbol{\theta})]\mathbf{r}_{\boldsymbol{\theta}}(\boldsymbol{\theta}))'\right]^{-1} \mathbf{s}(\widehat{\boldsymbol{\theta}}) \\ &= N\{\mathbf{g}_r[\mathbf{r}_{\boldsymbol{\theta}}(\boldsymbol{\theta})]^{-1}\mathbf{s}(\widehat{\boldsymbol{\theta}})\}' \left[\mathbf{r}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})\mathfrak{I}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{r}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})'\right]^{-1} \{\mathbf{g}_r[\mathbf{r}_{\boldsymbol{\theta}}(\boldsymbol{\theta})]^{-1}\mathbf{s}(\widehat{\boldsymbol{\theta}})\}\end{aligned}$$

which equals $\mathcal{W}$ only if $\mathbf{g}$ is exactly linear.

Linear restrictions:

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{c}$$

$$\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_N)$$

$$\text{rank}(\mathbf{R}) = r$$

The Wald test statistic is

$$\mathcal{W} = N\mathbf{r}(\widehat{\boldsymbol{\theta}})' \left[\mathbf{r}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) \Im(\widehat{\boldsymbol{\theta}})^{-1} \mathbf{r}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})' \right]^{-1} \mathbf{r}(\widehat{\boldsymbol{\theta}})$$

$$\mathbf{r}(\widehat{\boldsymbol{\theta}}) = \mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c}$$

$$\mathbf{r}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathbf{R}$$

$$\Im(\widehat{\boldsymbol{\theta}})^{-1} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

$$\mathcal{W} = (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})'[\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})$$

The score test rests upon the restricted estimator alone:

1. compute the restricted MLE

$$\widehat{\boldsymbol{\theta}}_R \equiv \arg \max_{\boldsymbol{\theta} \in \Theta : \mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}} E_N[L(\boldsymbol{\theta})]$$

   and the score for the restricted parameters $E_N[L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_R)]$

2. compute a consistent estimator $\Im(\widehat{\boldsymbol{\theta}}_R)$

3. Finally compute the quadratic form

$$\mathcal{S} = N E_N[L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_R)]' \Im(\widehat{\boldsymbol{\theta}}_R)^{-1} E_N[L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_R)] \xrightarrow{d} \chi^2_{K-M}$$

4. Compare $\mathcal{S}$ with the critical value of a χ^2_{K-M}

The second method of computing the score statistic uses the outer-product estimator (OPG) for the variance matrix estimator. Let

$$\mathbf{G} = L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_R; U_t)' \qquad (N \times K)$$

be the matrix of derivatives, then

$$E_N[L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_R)] = \frac{1}{N}\widehat{\mathbf{G}}'\iota$$

$$Var_N[L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_R)] = \frac{1}{N}\widehat{\mathbf{G}}'\widehat{\mathbf{G}}$$

the score test statistic is

$$\begin{aligned}\mathcal{S}_{OLS} &= N E_N[L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_R)]'\{Var_N[L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_R)]\}^{-1}E_N[L_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_R)] \\ &= \iota'\widehat{\mathbf{G}}(\widehat{\mathbf{G}}'\widehat{\mathbf{G}})^{-1}\widehat{\mathbf{G}}'\iota\end{aligned}$$

This statistic is the RSS from the regression of ι on the columns of $\widehat{\mathbf{G}}$. This statistic is not identical to $\mathcal{S}$ in practice, but they are asymptotically equivalent under the null hypothesis, because they differ only in the estimation of the variance matrix.

Linear restrictions:

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{c}$$

$$\mathbf{y}|\mathbf{X} \sim N(\mathbf{R}\boldsymbol{\beta}, \sigma_0^2 \mathbf{I}_N)$$

$$\text{rank}(\mathbf{R}) = r$$

The Score evaluated at $\widetilde{\boldsymbol{\beta}}$:

$$\begin{aligned} L_{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\beta}}, \sigma_0^2; \mathbf{y}|\mathbf{X}) &= \frac{1}{\sigma_0^2}\mathbf{X}'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}) \\ &= \frac{1}{\sigma_0^2}\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) \\ &= \frac{1}{\sigma_0^2}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c}) \end{aligned}$$

Under H_0

$$L_{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\beta}}, \sigma_0^2; \mathbf{y}|\mathbf{X}) \sim N\left(\mathbf{0}, \frac{1}{\sigma_0^2}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}\right)$$

The generalized inverse $\mathbf{A}^{-}$ of matrix $\mathbf{A}$ has the defining property that

$$\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$$

If $\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$ then the standardized quadratic form

$$(\mathbf{z} - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-}(\mathbf{z} - \boldsymbol{\mu})$$

where $\boldsymbol{\Omega}^{-}$ denotes the generalized inverse of $\boldsymbol{\Omega}$ has a chi-square distribution with degrees of freedom equal to the rank of $\boldsymbol{\Omega}$.

The generalized inverse of

$$\mathbf{A} = \mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}$$

is $(\mathbf{X}'\mathbf{X})^{-1}$ because

$$\begin{aligned}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A} &= \mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R} \\ &= \mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R} \\ &= \mathbf{A}.\end{aligned}$$

The quadratic form in the score and a generalized inverse of its variance matrix

$$
\begin{aligned}
\mathcal{S} &= L_{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\beta}}, \sigma_0^2; \mathbf{y}|\mathbf{X})'[\sigma_0^2(\mathbf{X}'\mathbf{X})^{-1}]L_{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\beta}}, \sigma_0^2; \mathbf{y}|\mathbf{X}) \\
&= \frac{1}{\sigma_0^2}(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}[\sigma_0^2(\mathbf{X}'\mathbf{X})^{-1}] \\
&\quad \frac{1}{\sigma_0^2}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c}) \\
&= (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c})'[\sigma_0^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{c}) \\
&= \mathcal{W}
\end{aligned}
$$

In this model all three test statistics are equal.

The null hypothesis as a restriction on a subset of the parameter vector $\boldsymbol{\theta} = [\boldsymbol{\theta}_1', \boldsymbol{\theta}_2']'$, $H_0 : \boldsymbol{\theta}_{02} = \mathbf{0}$. The score test statistic examines how much $E_N[L_2(\widetilde{\boldsymbol{\theta}})]$ deviates from $\mathbf{0}$. Because

$$E[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] = \mathbf{0}$$

so that if $\widetilde{\boldsymbol{\theta}}$ is in the neighborhood of $\boldsymbol{\theta}_0$, as it should be under $H_0 : \boldsymbol{\theta}_2 = \mathbf{0}$ then $E_N[L_2(\widetilde{\boldsymbol{\theta}})]$ should not deviate significantly from zero.

The score test statistic is

$$\mathcal{S} = N E_N[L_2(\widetilde{\boldsymbol{\theta}})]' \widehat{\mathbf{V}}_s^{-1} E_N[L_2(\widetilde{\boldsymbol{\theta}})]$$

where $\widehat{\mathbf{V}}_s$ is a consistent estimator of the

$$Var\left[\sqrt{N} E_N[L_2(\boldsymbol{\theta}_0)] \middle| \sqrt{N} E_N[L_1(\boldsymbol{\theta}_0)]\right]$$

Consider

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

where $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \sigma^2\mathbf{I}_N)$, the variance σ^2 is unknown.

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$$

The restricted MLE is

$$\widetilde{\boldsymbol{\beta}} = \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} \\ \mathbf{0} \end{bmatrix}$$

$$\widetilde{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})}{N}$$

The score for $\boldsymbol{\beta}_2$ is

$$E_N[L_2(\widetilde{\boldsymbol{\theta}})] = \frac{1}{N\widetilde{\sigma}^2}\mathbf{X}_2'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\theta}}) = \frac{1}{N\widetilde{\sigma}^2}\mathbf{X}_2'(\mathbf{I}_N - \mathbf{P}_{\mathbf{X}_1})\mathbf{y}$$

$$E_N[L_2(\widetilde{\boldsymbol{\theta}})] = \frac{1}{N\widetilde{\sigma}^2}\mathbf{X}_2'(\mathbf{I}_N - \mathbf{P}_{\mathbf{X}_1})'(\mathbf{I}_N - \mathbf{P}_{\mathbf{X}_1})\mathbf{y} = \frac{1}{N\widetilde{\sigma}^2}\mathbf{X}_{2\perp 1}'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})$$

where $\mathbf{X}_{2\perp 1} \equiv (\mathbf{I}_N - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2$.

Given the block-diagonality of $\mathfrak{I}(\boldsymbol{\beta}, \sigma^2)$ in $\boldsymbol{\beta}$ and σ^2

$$\widehat{\mathbf{V}}_S = \frac{1}{N\widetilde{\sigma}^2}\mathbf{X}_2'(\mathbf{I}_N - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2$$

$$\begin{aligned}\mathcal{S} &= N E_N[L_2(\widetilde{\boldsymbol{\beta}})]'\widehat{\mathbf{V}}_S^{-1} E_N[L_2(\widetilde{\boldsymbol{\beta}})] \\ &= \frac{1}{\widetilde{\sigma}^2}(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})'\mathbf{X}_{2\perp 1}(\mathbf{X}_{2\perp 1}'\mathbf{X}_{2\perp 1})^{-1}\mathbf{X}_{2\perp 1}'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})\end{aligned}$$

$$\mathcal{S} = \frac{1}{\widetilde{\sigma}^2}(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})$$

because $(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}) \perp \mathbf{X}_1$.
This statistic could be calculated as the RSS from an OLS regression of $\frac{(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})}{\widetilde{\sigma}}$ on $\mathbf{X}$. The RSS is

$$\widehat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})^{-1}\widehat{\boldsymbol{\beta}} = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

if we replace $\mathbf{y}$ with $\frac{(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})}{\widetilde{\sigma}}$ we obtain the test statistic.