Hypothesis Testing, Model Selection, and Prediction in Least Squares and Maximum Likelihood Estimation

Greene Ch.5, 14; Kennedy Ch. 4
R script `mod3s2a, mod3s2b, mod3s2c`

Testing for Restrictions in a LS Model

Linear Restrictions

Assume you have specified a CLRM of the usual form $\mathbf{y}=\mathbf{X\beta}+\boldsymbol{\varepsilon}$. Let's call this the "general" or "unrestricted" model. You may hypothesize that some elements of $\boldsymbol{\beta}$ are linearly related or constrained to take a specific value. Such constraints are called "linear constraints" or "linear restrictions". Here are a few examples:

Single value restriction:
$\beta_3=0$

Multiple value restrictions:
$\beta_3=0$
$\beta_5=-1$

Single restriction on linear relationships:
$\beta_2-\beta_3=0$

Multiple restrictions on linear relationships:
$\beta_2-\beta_3=0$
$\beta_1+\beta_3+4\beta_7=5$

Mixed linear restrictions:
$\beta_1-\beta_4=0$
$\beta_5=1$
$\beta_3+\beta_2=3$

In general, such hypotheses on restrictions arise from competing underlying theoretical models. For example, for our wage regression (see ***R*** script `mod3s2a`), you might hypothesize that a male worker earns \$3 more than a female worker, ceteris paribus. This would translate into the following hypothesis:

$$H_0:\beta_2=-3 \qquad (1)$$

As a general rule, you should always express a linear constraint such that there is only a numerical value left on the right hand side of the equality sign.

In general we can write a set of J linear constraints as

$$\begin{aligned}
&r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1k}\beta_k = q_1 \\
&r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2k}\beta_k = q_2 \\
&\vdots \\
&r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{Jk}\beta_k = q_J
\end{aligned} \tag{2}$$

or, more compactly as

$\mathbf{R\beta} = \mathbf{q}$ where

$$\underset{Jxk}{\mathbf{R}} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r_{21} & r_{22} & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{J1} & r_{J2} & \cdots & r_{Jk} \end{bmatrix}, \qquad \underset{Jx1}{\mathbf{q}} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_J \end{bmatrix} \tag{3}$$

In principle you can test any number of linear restrictions, but you must always obey the following 2 rules:

1. $J \le k$, i.e. you can't impose more restrictions than there are coefficients.
2. The rows of **R** must be linearly independent, i.e. you can't have any redundant or conflicting restrictions.

Here are a few examples for **R** and **q** using the wage data. Recall the element of X:

```
% contents of X
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%1  constant term
%2  female          1= worker = female
%3  non-white       1= worker = non-white
%4  union           1 = worker = unionized
%5  education       years of education
%6  experience      years of work experience
```

So *k*=6.

<u>Example 1:</u>

H$_0$: "Female workers earn $3 less than male workers, ceteris paribus"

$H_0: \beta_2 = -3 \qquad \rightarrow J = 1$

$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{q} = -3$

<u>Example 2:</u>

H$_0$:"1 additional year of education is worth 8 additional years of experience"

$H_0: \beta_5 - 8\beta_6 = 0 \qquad \rightarrow J = 1$

$\mathbf{R} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & -8 \end{bmatrix} \qquad \mathbf{q} = 0$

<u>Example 3:</u>

H$_0$: "Female workers earn $3 less than male workers, ceteris paribus" AND
"1 additional year of education is worth 8 additional years of experience"

$$H_0: \quad \beta_2 = -3$$
$$\beta_5 - 8\beta_6 = 0 \qquad \rightarrow J = 2$$
$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -8 \end{bmatrix} \qquad \mathbf{q} = \begin{bmatrix} -3 \\ 0 \end{bmatrix}$$

Example 4:
H_0: “Female workers earn $3 less than male workers, ceteris paribus” AND
”1 additional year of education is worth 8 additional years of experience” AND
“a unionized worker earns $1 more than a non-unionized worker, cet. par”.

$$H_0: \quad \beta_2 = -3$$
$$\beta_5 - 8\beta_6 = 0$$
$$\beta_4 = 1 \qquad \rightarrow J = 3$$
$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -8 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{q} = \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix}$$

Once **R** and **q** are explicitly defined, you can compute an *F-statistic* to implement the test as follows:

$$\begin{aligned} F_{(J,n-k)} &= \frac{1}{J}\left(\mathbf{Rb}-\mathbf{q}\right)'\left[\mathbf{R}\,\hat{V}\left(\mathbf{b}\right)\mathbf{R}'\right]^{-1}\left(\mathbf{Rb}-\mathbf{q}\right) = \\ & \frac{1}{J}\left(\mathbf{Rb}-\mathbf{q}\right)'\left[\mathbf{R}\,s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\left(\mathbf{Rb}-\mathbf{q}\right) \end{aligned} \tag{4}$$

A random variable that follows the F-distribution has 2 parameters, “Degrees of Freedom for the numerator (DoFn)” and “Degrees of Freedom for the denominator (DoFd)”. The DoFn are always equal to the number of restrictions (*J*), and the DoFd are equal to the sample size minus the number of parameters (*n*-*k*).

F-Tables with critical values for different DoFs and levels of significance can be readily obtained from the internet (see also p. 1096 in Greene's 6th edition). We’ll use the one for $\alpha = 0.05$. In that Table, n_1 = *J*, and n_2 = *n*-*k*. The table shows “critical values”. For example, if *J* = 4 and (*n*-*k*) >100, the critical F value is 2.37. If your computed *F* exceeds this value, you reject H_0. Otherwise don’t reject. More conveniently, you can directly compute the p-value for your test statistic in *R*. Then compare the p-value to your level of significance and draw your conclusion.

F-test vs. t-test

For a single restriction $t = \sqrt{F}$, so you can either do an F-test (using the F-table) or a t-test (using the t-table). Both will always arrive at the same test decision. Recall that both t- and F-statistics require the normality assumption for $\boldsymbol{\varepsilon}$ in the CLRM.

Testing for the inequality of the entire $\boldsymbol{\beta}$-vector over two groups: The Chow Test

Often times your data set can be conceptually split into two groups, such as “male / female”, “white / nonwhite” for the wage data, or “spin casters / fly fishers” for the angling data, etc. You may then hypothesize that the marginal effect of *all or some* of the remaining regressors differs over the two groups. Let's use the wage data and "white / nonwhite" as an example.

Using the full set of remaining regressors as example, your null and alternative hypothesis could be stated as:

$$\begin{aligned} H_0: \quad & \beta_{\text{constant,w}} - \beta_{\text{constant,nw}} = 0 \\ & \beta_{\text{gender,w}} - \beta_{\text{gender,nw}} = 0 \\ & \beta_{\text{union,w}} - \beta_{\text{union,nw}} = 0 \\ & \beta_{\text{education,w}} - \beta_{\text{education,nw}} = 0 \\ & \beta_{\text{experience,w}} - \beta_{\text{experience,nw}} = 0 \end{aligned} \tag{5}$$

H_a: *At least one* of these equalities doesn't hold.

Where "*w*" stands for "white" and "*nw*" stands for "nonwhite. In principle, you thus have *J*=5 restrictions and could use an F-test. The tricky part for a Chow test is in designing the unconstrained model, i.e. a model that estimates separate coefficients for these 5 regressors (incl. the constant) for the two race groups. You need to re-construct your **y** and **X** as follows:

1. Pick the variables of interest from the original **X** and group them into a new **X** (let's call it $\mathbf{X}_{\mathbf{int}}$ for now). Usually these will be all variables in the original **X** minus the column of ones and the variable that originally defined the two groups of interest (here "race").
2. Create indicator (dummy variables) for each group.
3. Create interaction terms between this dummy and $\mathbf{X}_{\mathbf{int}}$**.**
4. Estimate a CLRM of **y** against the first indicator dummy, its interactions, the opposite indicator dummy, and its interactions. Then test for the stated inequalities using an F-test.

Conceptually, the unconstrained model takes the following form (assume that observations have been sorted by group – this is not actually needed for estimation):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{where}$$

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_{\mathbf{w}} \\ \mathbf{y}_{\mathbf{nw}} \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} \mathbf{X}_{\mathbf{w}} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{\mathbf{nw}} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_{\mathbf{w}} \\ \boldsymbol{\beta}_{\mathbf{nw}} \end{bmatrix} \tag{6}$$

Clearly, if your H_a holds *for the entire set of restrictions*, this would be equivalent to running two separate regression models, one using only data for nonwhite workers, and one using only the sub-sample of white workers.

See *mod3s2a* for the implementation of this example.

Nonlinear restrictions

To test for nonlinear restrictions on $\boldsymbol{\beta}$ in the CLRM we need to invoke asymptotic results. Specifically, we need to apply the Delta Method to derive the estimated asymptotic variance of the restricted coefficient or set of coefficients.

Assume we impose e set of *J* nonlinear restrictions, each of which involves one or more elements of $\boldsymbol{\beta}$. Using Greene's (p. 114) notation, we can compactly express the set of restrictions as

$$c(\boldsymbol{\beta}) = \mathbf{q}, \quad \text{where} \quad \underset{Jx1}{c(\boldsymbol{\beta})} = \left[c_1(\boldsymbol{\beta}) \quad c_2(\boldsymbol{\beta}) \quad \cdots \quad c_J(\boldsymbol{\beta})\right]' \tag{7}$$

and **q** is a *J* by 1 vector of numerical values, (as for the linear restriction case). As a first step, we need to estimate the asymptotic variance of $c(\boldsymbol{\beta})$. By the Delta method, we have

$$\hat{V}\left(c(\mathbf{b})\right) = \mathbf{C}\hat{V}(\mathbf{b})\mathbf{C}' = \mathbf{C}\left(s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)\mathbf{C}' \text{ where } \underset{Jxk}{\mathbf{C}} = \frac{\partial c(\mathbf{b})}{\partial \mathbf{b}'}. \tag{8}$$

Next, we compute a *Wald test statistic* – essentially the asymptotic version of an F-test. As a general rule, we don’t use degrees of freedom based on sample size in asymptotic test statistics. Thus:

$$\begin{aligned} &W = \left(c(\mathbf{b}) - \mathbf{q}\right)'\left(\mathbf{C}\left(\hat{V}_a(\mathbf{b})\right)\mathbf{C}'\right)^{-1}\left(c(\mathbf{b}) - \mathbf{q}\right) \sim \chi^2(J) \qquad \text{where} \\ &\hat{V}_a(\mathbf{b}) = s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} \end{aligned} \tag{9}$$

The last term indicates that W follows a chi-squared distribution with J degrees of freedom. The table of critical values for this distribution is given on p. 1095 of Greene's 6th edition, or you can easily find it online. In Greene, the column labeled “0.95” corresponds to the 5% level of significance (i.e. $\alpha = 0.05$). For example, you can see that the critical value for *J*=1 equals 3.84.

If $c(\boldsymbol{\beta}) = \mathbf{q}$ contains only a single restriction, you can alternatively use a *z*-test as shown on p. 98. The *z*-*W* relationship is analogous to the t-F case, i.e. for a single restriction we have $z = \sqrt{W}$. An example of a Wald test involving nonlinear restrictions is given in script *mod3s2a*.

In theory, the Wald test could also be used for linear restrictions, in which case we have $c(\boldsymbol{\beta}) - \mathbf{q} = \mathbf{R}\boldsymbol{\beta} - \mathbf{q}$ as before. However, in the LS case where the finite properties of **b** are known the *F* test usually yields more accurate results, especially under small sample sizes.

Testing for Restrictions in an MLE Model

There exist three asymptotically equivalent test statistics for MLE estimation. They differ in the number and type of models you need to estimate before you can run a specific test. Here is a summary:

Test	Abbreviation	Required estimated models	Distribution of test statistic	Strengths / Limitations
Likelihood-Ratio	LR	Constrained & unconstrained	$\chi^2_{(J)}$	Constrained model often difficult to estimate; probably most reliable test of the three
Lagrange Multiplier	LM	Constrained	$\chi^2_{(J)}$	Constrained model often difficult to estimate
Wald	W	Unconstrained	$\chi^2_{(J)}$	Probably most straightforward to implement

LR test

The LR test requires the estimation of both the unconstrained and constrained models. The test statistic is then derived as

$$LR = 2\left(\ln L\left(\hat{\boldsymbol{\theta}}_{\mathbf{u}}\right) - \ln L\left(\hat{\boldsymbol{\theta}}_{\mathbf{c}}\right)\right) \sim \chi^2_{(J)} \qquad (10)$$

where $\ln L\left(\hat{\boldsymbol{\theta}}_{\mathbf{u}}\right)$ is the value of the log-likelihood function at the solution for the unconstrained model, and $\ln L\left(\hat{\boldsymbol{\theta}}_{\mathbf{c}}\right)$ is the analogous value for the constrained model. As before, J indicates the total number of joint restrictions to be tested.

The *intuition* for this test is that if the imposed constraint is correct, the value of *lnL* for the constrained model should be close to that for the unconstrained model (it can never be better, i.e. less negative than the *lnL* for the unconstrained model). A pronounced difference between the two values would suggest that the constraint (or set of constraints) does not hold. As before, the test decision is based on a comparison of the computed *LR* to the applicable critical value in a chi-squared table.

An example of an LR test is given in *mod3s2b.*

Wald Test

The Wald test requires estimation results from the unconstrained model. It takes the same form as given in (9) with the OLS estimator **b** replaced by the MLE estimator (say $\hat{\boldsymbol{\beta}}$), and the estimated asymptotic variance of $\hat{\boldsymbol{\beta}}$ given by any of the usual approximations, i.e. either the inverted negative Hessian or the outer product of gradients. Thus:

$$W=\left(c\left(\hat{\boldsymbol\beta}\right)-\mathbf{q}\right)'\left(\mathbf{C}\left(\hat{V}_a\left(\hat{\boldsymbol\beta}\right)\right)\mathbf{C}'\right)^{-1}\left(c\left(\hat{\boldsymbol\beta}\right)-\mathbf{q}\right)\sim\chi^2\left(J\right)\qquad\text{where}$$
$$\hat{V}_a\left(\hat{\boldsymbol\beta}\right)=\left(-H\left(\hat{\boldsymbol\beta}\right)\right)^{-1}\quad\text{or}\quad\hat{V}_a\left(\hat{\boldsymbol\beta}\right)=\left(\mathbf{G'G}\right)^{-1}\tag{11}$$

If the restrictions are linear $c\left(\hat{\boldsymbol\beta}\right)-\mathbf{q}$ takes the form of $\mathbf{R}\hat{\boldsymbol\beta}-\mathbf{q}$.

The intuition for the Wald test is that if the hypothesized restrictions are valid $c\left(\hat{\boldsymbol\beta}\right)-\mathbf{q}$ should be close to zero and "*W*" will be small. A large value for the Wald statistic would raise doubts as to the validity of the constraint.

The Lagrange Multiplier (LM) test

The LM test (also known as "score" test) is based on the constrained model and the intuition that if the imposed restrictions are correct, the sample gradient (or "score" function) should be close to zero at convergence. Since, by the information matrix identity, the variance of the gradient is equal to the information matrix, the test statistic is computed as follows:

$$LM=g\left(\hat{\boldsymbol\beta}\right)'\left(I\left(\hat{\boldsymbol\beta}\right)\right)^{-1}g\left(\hat{\boldsymbol\beta}\right)\tag{12}$$

where $I\left(\hat{\boldsymbol\beta}\right)$ is again estimated by the negative Hessian or the outer product of gradients.

An implementation of the three tests is given in script *mod3s2b*. As you can see, the results for the three tests can vary substantially under small to moderate sample sizes. This fact and the sensitivity of test statistics to the approximation used for $\hat{V}_a\left(\hat{\boldsymbol\beta}\right)$ is also illustrated in Greene's example 14.6.4 (p. 531).

So which test should be used in practice? In many cases the constrained model is difficult to specify & estimate, which makes the *LM* and *LR* tests somewhat less popular than the Wald.

The Wald test is also convenient when you have estimated the unrestricted model, and you want to test a whole series of hypotheses on different constraints. Using the LR, you would have to specify & estimate a separate restricted model for every set of restrictions.

In practice, if both unconstrained and constrained models are straightforward to estimate, the *LR* test is probably the most trusted approach as the value of the log-LH function at convergence is somewhat less sensitive to the choice of $\hat{V}_a\left(\hat{\boldsymbol\beta}\right)$ than the other two test statistics.

Ideally, you would perform all three tests & hope that they all point to the same decision rule. Good luck!

Model Selection based on Prediction

As mentioned in Poirier's "Intermediate Statistics and Econometrics", p. 405, the topic of prediction is somewhat neglected in most econometric textbooks, which focus more on estimation and tests on parameters. However, the prediction of yet unobserved outcomes is of central importance when econometric analysis is supposed to produce policy recommendations, for example in the context of benefit-cost analysis.

We will use the term "prediction" in the general sense of "combining some vector of explanatory values, **x**, with estimated parameters, $\hat{\boldsymbol{\theta}}$, to generate a point estimate $\hat{y} = f\left(\mathbf{x}, \hat{\boldsymbol{\theta}}\right)$ where *f* may be a linear or nonlinear function. (In time series analysis, "prediction" is usually referred to as "forecasting").

Within-Sample Prediction and Predictive Accuracy

The strategy of computing a predicted outcome, $\hat{y}_i$, based on combining an *observed* vector $\mathbf{x_i}$ with estimated parameters is called "*within sample prediction*". In contrast, combining estimated parameters with a new set of values for the regressors, say $\mathbf{x_p}$, is called "*out-of-sample prediction*".

Hypothesis tests are one approach to select between competing models. Another criterion relates to the *predictive accuracy* of competing models, i.e. how well the predicted dependent observations ("fitted values" in the CLRM context) agree with the actually observed y_i's. One such measure for the CLRM is R^2 and its adjusted counterpart.

Another, more generally applicable, criterion is the predictive *Mean Squared Error*, defined as

$$MSE = \tfrac{1}{\mathrm{n}} \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \tag{13}$$

Since this number can grow quite large, its square root is generally used instead, leading to the *Root Mean Squared Error(RMSE)*. An alternative criterion is the *Mean Absolute Error*, given as

$$MAE = \tfrac{1}{\mathrm{n}} \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right| \tag{14}$$

A smaller value for these measures indicates a better fit with the actual data. A (minor) problem with the RMSE and MAE is that they are not invariant to scaling of **y** – if **y** is multiplied by a factor, say α, then the RMSE and MAE will also be scaled by α. If a scale-free measure of predictive accuracy is desired, the **Theil *U*-statistic** can be used:

$$U = \sqrt{\frac{\tfrac{1}{\mathrm{n}} \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{\tfrac{1}{\mathrm{n}} \sum_{i=1}^{n} y_i^2}} \tag{15}$$

Note that a good fit based on these within-sample predictive statistics does not necessarily imply that the chosen model will generate accurate "out of sample" predictions as well. However, a poor fit with actual data would certainly raise serious doubts as to the model's out-of-sample predictive abilities, especially if the values in $\mathbf{x_p}$ are "not too different" from other values observed in the sample (e.g. predicting the price of a home with 5 bedrooms when the sample includes only homes with 2,3,4, and 6 bedrooms).

Out-of-Sample Prediction and Predictive Efficiency

When we predict out-of-sample we can no longer compare our predicted outcome to an actually observed value. However, we can still compute a standard error and confidence interval for the predicted value, and compare models based on the width of this interval. If two or more competing models produce similar point predictions, we would choose the one that generates the tightest confidence interval around its point estimate.

Linear predictions

Assume you are interested in the out-of-sample-predicted value $y_p = \mathbf{x}_\mathbf{p}'\hat{\boldsymbol{\beta}}$. Thus, y_p is a linear combination of regressor values and estimated coefficients. A 95% confidence interval can be quickly constructed as

$$C.I._{95\%} = \left\{\mathbf{x}_\mathbf{p}'\hat{\boldsymbol{\beta}} \pm 1.96 \cdot se\left(\mathbf{x}_\mathbf{p}'\hat{\boldsymbol{\beta}}\right)\right\}, \text{ where } \quad se\left(\mathbf{x}_\mathbf{p}'\hat{\boldsymbol{\beta}}\right) = \sqrt{\hat{V}\left(\mathbf{x}_\mathbf{p}'\hat{\boldsymbol{\beta}}\right)} \quad \text{and}$$
$$\hat{V}\left(\mathbf{x}_\mathbf{p}'\hat{\boldsymbol{\beta}}\right) = \mathbf{x}_\mathbf{p}\hat{V}_a\left(\hat{\boldsymbol{\beta}}\right)\mathbf{x}_\mathbf{p}' \tag{16}$$

Nonlinear predictions

Now assume your predictive construct of interest is a nonlinear function of regressor values and estimated coefficients, i.e. $y_p = f\left(\mathbf{x_p}, \hat{\boldsymbol{\beta}}\right)$. The derivation of a 95% confidence interval is as in (16), with $\mathbf{x}_\mathbf{p}'\hat{\boldsymbol{\beta}}$ replaced by $f\left(\mathbf{x_p}, \hat{\boldsymbol{\beta}}\right)$, and $\hat{V}_a\left(f\left(\mathbf{x_p}, \hat{\boldsymbol{\beta}}\right)\right)$ derived via the Delta Method. Alternatively, the entire confidence interval can be derived via simulation by drawing multiple sets of $\hat{\boldsymbol{\beta}}$ from its asymptotic distribution, computing $f\left(\mathbf{x_p}, \hat{\boldsymbol{\beta}}\right)$ for each draw, and using the 2.5th and 97th percentiles as confidence bounds. See script `mod3s2c` for an example.