Nonlinear Regression Models

A nonlinear regression model can be written as

$$y_i = x_i(\boldsymbol{\beta}) + u_i, \quad u_i \sim \text{IID}(0, \sigma^2), \quad i = 1, \ldots, N. \tag{1}$$

Here $x_i(\boldsymbol{\beta})$ is a **nonlinear regression function**. It depends (implicitly) on explanatory variables and a k-vector $\boldsymbol{\beta}$ of parameters.

In vector terms,

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{2}$$

where the i^{th} element of $\boldsymbol{x}(\boldsymbol{\beta})$ is $x_i(\boldsymbol{\beta})$.

A simple example of a nonlinear regression function is

$$\beta_0 + \beta_1 X_{1i}^{\gamma} + \beta_2 X_{2i}^{\gamma}. \tag{3}$$

This is nonlinear in the regressors and in 1 of the 4 parameters.

If we combine the linear regression model $y_t = \boldsymbol{Z}_t \boldsymbol{\beta} + v_t$ with the **AR(1) error process**

$$v_t = \rho v_{t-1} + u_t, \quad u_t \sim \mathrm{IID}(0, \sigma_u^2), \quad |\rho| < 1, \tag{4}$$

we obtain the nonlinear regression model

$$\begin{aligned} y_t &= \rho y_{t-1} + \boldsymbol{Z}_t \boldsymbol{\beta} - \rho \boldsymbol{Z}_{t-1} \boldsymbol{\beta} + u_t &(5)\\ &= x_t(\boldsymbol{\beta}, \rho) + u_t, \quad u_t \sim \mathrm{IID}(0, \sigma_u^2). &(6)\end{aligned}$$

This is linear in the regressors but nonlinear in the parameters.

A linear regression model can include all sorts of nonlinear functions of the original regressors (squares, square roots, cross-products, logarithms, inverses, exponentials, etc.). We can still use OLS.

But we cannot use OLS if $x_i(\boldsymbol{\beta})$ is nonlinear in one or more parameters. The term “nonlinear regression” is usually reserved for such models.

Nonlinear Least Squares

ETM discusses method-of-moments (MM) estimation in some detail, but in practice nonlinear least squares, or NLS, is almost always used.

To obtain NLS estimates, we minimize

$$\mathrm{SSR}(\boldsymbol{\beta}) = \big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^{\top}\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big). \tag{7}$$

Let $X_{ij}(\boldsymbol{\beta})$ denote the derivative of $x_i(\boldsymbol{\beta})$ with respect to β_j.

Then we can let $\boldsymbol{X}_i(\boldsymbol{\beta})$ denote a $1 \times k$ vector and $\boldsymbol{X}(\boldsymbol{\beta})$ denote an $N \times k$ matrix, each having typical element $X_{ij}(\boldsymbol{\beta})$.

These are the analogs of the vector $\boldsymbol{X}_i$ and the matrix $\boldsymbol{X}$ for the linear regression model. In that case, the regression function is $\boldsymbol{X\beta}$, so that $\boldsymbol{X}_i(\boldsymbol{\beta}) = \boldsymbol{X}_i$ and $\boldsymbol{X}(\boldsymbol{\beta}) = \boldsymbol{X}$.

The first-order conditions to minimize (7) (omitting a factor of -2) are

$$\boldsymbol{X}^{\top}(\hat{\boldsymbol{\beta}})\big(\boldsymbol{y} - \boldsymbol{x}(\hat{\boldsymbol{\beta}})\big) = \boldsymbol{0}. \tag{8}$$

Multiplying by the appropriate power of N, these become

$$N^{-1/2}\boldsymbol{X}^{\top}(\hat{\boldsymbol{\beta}})\big(\boldsymbol{y}-\boldsymbol{x}(\hat{\boldsymbol{\beta}})\big)=\mathbf{0}. \tag{9}$$

The factor of $N^{-1/2}$ is here because we are going to derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ from (9).

Since $\boldsymbol{y}-\boldsymbol{x}(\boldsymbol{\beta}_0)=\boldsymbol{u}$, a first-order Taylor expansion around $\boldsymbol{\beta}_0$ yields

$$N^{-1/2}\boldsymbol{X}_0^{\top}\boldsymbol{u}-\left(\frac{1}{N}\boldsymbol{X}_0^{\top}\boldsymbol{X}_0+\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{A}_i(\boldsymbol{\beta}_0)\,u_i\right)\right)N^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)=\mathbf{0}, \tag{10}$$

where $\boldsymbol{x}_0\equiv\boldsymbol{x}(\boldsymbol{\beta}_0)$ and $\boldsymbol{X}_0\equiv\boldsymbol{X}(\boldsymbol{\beta}_0)$.

Equation (10) involves the matrix with typical element

$$\big[\boldsymbol{A}_i(\boldsymbol{\beta})\big]_{jl}=\frac{\partial^2 x_i(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_l}. \tag{11}$$

This is the second derivative of the regression function for the i^{th} observation with respect to the j^{th} and l^{th} parameters.

The leading-order terms in (10) are $O_p(1)$. They are

$$N^{-1/2}\boldsymbol{X}_0^\top \boldsymbol{u} \quad \text{and} \quad \frac{-1}{N}\,\boldsymbol{X}_0^\top \boldsymbol{X}_0\, N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0). \tag{12}$$

But the remaining term is $O_p(N^{-1/2})$ because

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{A}_i(\boldsymbol{\beta}_0)\, u_i \tag{13}$$

is $1/N$ times a weighted average of the u_i, which have mean 0.

Thus, asymptotically, we can ignore this third term. The moment conditions (10) are asymptotically equivalent to

$$N^{-1/2}\boldsymbol{X}_0^\top \boldsymbol{u} - \frac{1}{N}\,\boldsymbol{X}_0^\top \boldsymbol{X}_0 N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \overset{a}{=} \mathbf{0}. \tag{14}$$

These look very much like the moment conditions for OLS.

Solving these equations, we obtain

$$N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \overset{a}{=} \Big(\frac{1}{N}\boldsymbol{X}_0^\top\boldsymbol{X}_0\Big)^{-1} N^{1/2}\boldsymbol{X}_0^\top\boldsymbol{u}. \tag{15}$$

The asymptotic normality of $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ follows directly.

The asymptotic covariance matrix of (15) is

$$\Big(\operatorname*{plim}_{N\to\infty}\frac{1}{N}\boldsymbol{X}_0^\top\boldsymbol{X}_0\Big)^{-1}\Big(\operatorname*{plim}_{N\to\infty}\frac{1}{N}\boldsymbol{X}_0^\top\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{X}_0\Big)\Big(\operatorname*{plim}_{N\to\infty}\frac{1}{N}\boldsymbol{X}_0^\top\boldsymbol{X}_0\Big)^{-1}. \tag{16}$$

In the i.i.d. case,

$$\mathrm{Var}\left(N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right) \overset{a}{=} \sigma^2\Big(\operatorname*{plim}_{N\to\infty}\frac{1}{N}\boldsymbol{X}_0^\top\boldsymbol{X}_0\Big)^{-1}. \tag{17}$$

In practice, of course, we ignore the factors of N and replace σ^2 by s^2 and $\boldsymbol{X}_0$ by $\hat{\boldsymbol{X}} \equiv \boldsymbol{X}(\hat{\boldsymbol{\beta}})$:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = s^2(\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})^{-1}. \tag{18}$$

More generally, when $\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}$, the asymptotic variance is

$$\Big(\operatorname*{plim}_{N\to\infty} \frac{1}{N}\boldsymbol{X}_0^\top \boldsymbol{X}_0\Big)^{-1}\Big(\operatorname*{plim}_{N\to\infty} \frac{1}{N}\boldsymbol{X}_0^\top \boldsymbol{\Omega}\boldsymbol{X}_0\Big)\Big(\operatorname*{plim}_{N\to\infty} \frac{1}{N}\boldsymbol{X}_0^\top \boldsymbol{X}_0\Big)^{-1}. \tag{19}$$

and we actually use

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{X}}(\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})^{-1}. \tag{20}$$

The middle matrix in (20) can be chosen in various ways, depending on the assumptions we make.

- When the middle matrix is $\sum_{i=1}^N \hat{u}_i^2 \hat{\boldsymbol{X}}_i^\top \hat{\boldsymbol{X}}_i$, perhaps with $\hat{u}_i$ rescaled, then (20) is an HCCME.
- When the middle matrix is $\sum_{g=1}^G \hat{\boldsymbol{X}}_g^\top \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g^\top \hat{\boldsymbol{X}}_g$, perhaps with $\hat{\boldsymbol{u}}_g$ rescaled, then (20) is a CRVE.
- The middle matrix can also be specified so as to make (20) a HAC estimator. There are many such estimators.

Computing NLS Estimates

We wish to minimize a function $Q(\boldsymbol{\beta}) = \text{SSR}(\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a k-vector and $Q(\boldsymbol{\beta})$ is assumed to be twice continuously differentiable.

Given any initial value of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}_{(0)}$, we can perform a second-order Taylor expansion of $Q(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_{(0)}$.

This yields $Q^*(\boldsymbol{\beta})$, a quadratic approximation to $Q(\boldsymbol{\beta})$:

$$Q^*(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}_{(0)}) + \boldsymbol{g}_{(0)}^{\top}(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)})^{\top}\boldsymbol{H}_{(0)}(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}), \tag{21}$$

where $\boldsymbol{g}(\boldsymbol{\beta})$ is the **gradient** of $Q(\boldsymbol{\beta})$ and $\boldsymbol{H}(\boldsymbol{\beta})$ is the **Hessian**.

The gradient has typical element $\partial Q(\boldsymbol{\beta})/\partial \beta_j$, and the Hessian has typical element $\partial^2 Q(\boldsymbol{\beta})/\partial \beta_j \partial \beta_l$.

Let $\boldsymbol{g}_{(0)}$ and $\boldsymbol{H}_{(0)}$ denote $\boldsymbol{g}(\boldsymbol{\beta}_{(0)})$ and $\boldsymbol{H}(\boldsymbol{\beta}_{(0)})$, respectively.

The first-order conditions for a minimum of $Q^*(\boldsymbol{\beta})$ are

$$\boldsymbol{g}_{(0)} + \boldsymbol{H}_{(0)}(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}) = \mathbf{0}. \tag{22}$$

Solving these yields

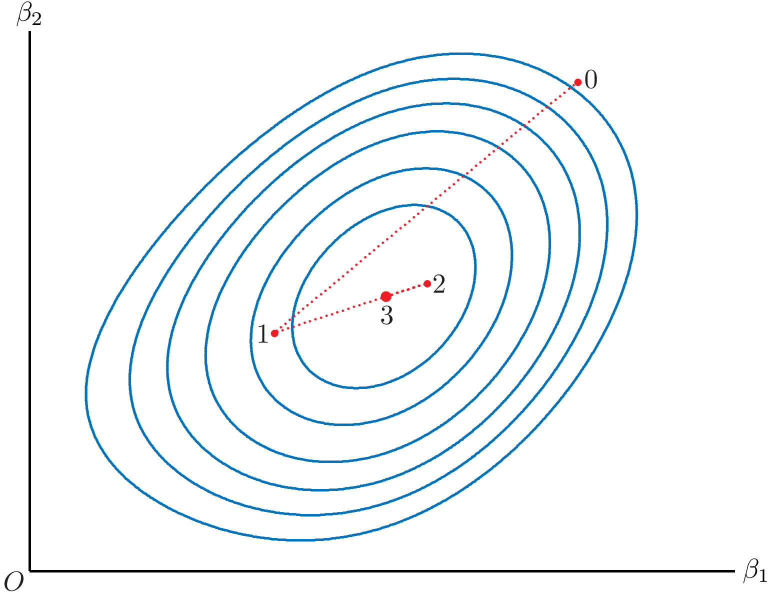
$$\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(0)} - \boldsymbol{H}_{(0)}^{-1}\boldsymbol{g}_{(0)}. \tag{23}$$

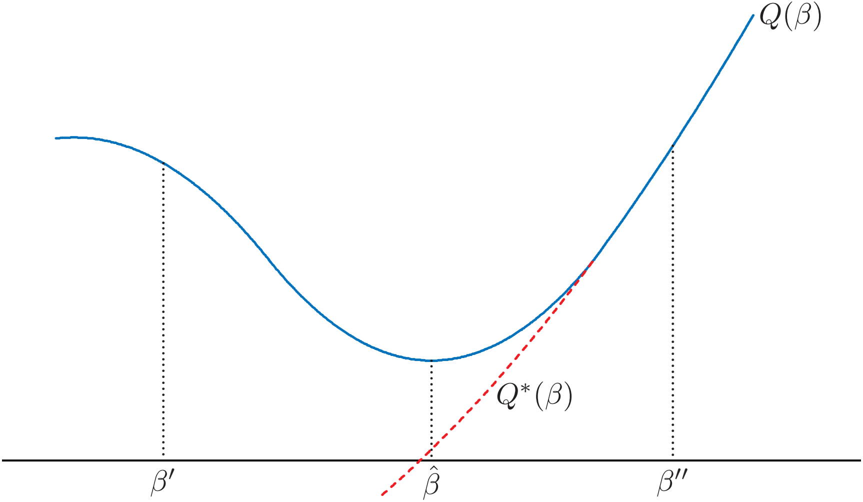
This is the key equation of **Newton's Method**. The idea is to apply it repeatedly so as to obtain $\boldsymbol{\beta}_{(2)}$, $\boldsymbol{\beta}_{(3)}$, and so on.

The figure shows an example for a two-dimensional function. Note that we will need a **stopping rule**.

Newton's Method can fail when the quadratic approximation to $Q(\boldsymbol{\beta}_{(0)})$ is not good enough. The second figure shows two examples of failure for a one-dimensional function.

In theory, the quadratic approximation will always be very good when $\boldsymbol{\beta}$ is close to the minimum.


β_2
0
2
3
1
O
β_1


$Q(\beta)$
$Q^*(\beta)$
β'
$\hat{\beta}$
β''

Quasi-Newton methods attempt to retain the good qualities of Newton's Method while surmounting problems of non-convexity and poor quadratic approximations.

They replace the Newton step by the more complicated formula

$$\boldsymbol{\beta}_{(s+1)} = \boldsymbol{\beta}_{(s)} - \alpha_{(s)} \boldsymbol{D}_{(s)}^{-1} \boldsymbol{g}_{(s)}. \tag{24}$$

Here $\alpha_{(s)}$ is a scalar which is determined at each step s, and $\boldsymbol{D}_{(s)} \equiv \boldsymbol{D}(\boldsymbol{\beta}_{(s)})$ is a matrix which approximates $\boldsymbol{H}_{(s)}$ near the minimum but is always positive definite.

Newton's Method itself sets $\boldsymbol{D}_{(s)} = \boldsymbol{H}_{(s)}$ and $\alpha_{(s)} = 1$.

Effective quasi-Newton methods often choose $\alpha_{(s)}$ to minimize

$$Q^{\dagger}(\alpha) \equiv Q\big(\boldsymbol{\beta}_{(s)} - \alpha \boldsymbol{D}_{(s)}^{-1} \boldsymbol{g}_{(s)}\big). \tag{25}$$

This is easy, because $Q^{\dagger}(\alpha)$ is one-dimensional.

Any nonlinear optimization algorithm needs a **stopping rule**. One that often works well is to stop when

$$\boldsymbol{g}_{(s)}^{\top}\boldsymbol{D}_{(s)}^{-1}\boldsymbol{g}_{(s)} < \epsilon. \tag{26}$$

Here the **tolerance** ϵ should be very small, say $10^{-16} < \epsilon < 10^{-8}$.

The recipe for a quasi-Newton optimization algorithm is:

1. Pick the starting point $\boldsymbol{\beta}_{(0)}$, and set $s = 0$.
2. Compute $\boldsymbol{g}_{(s)}$ and $\boldsymbol{D}_{(s)}$ and use them to determine the direction vector $\boldsymbol{D}_{(s)}^{-1}\boldsymbol{g}_{(s)}$.
3. Find $\alpha_{(s)}$, perhaps by solving a one-dimensional minimization problem like (25). Then use (24) to determine $\boldsymbol{\beta}_{(s+1)}$.
4. Decide whether $\boldsymbol{\beta}_{(s+1)}$ provides a sufficiently accurate approximation to $\hat{\boldsymbol{\beta}}$. If so, stop. If not, return to step 2 with $s = s + 1$.

Numerical optimization methods based on Newton's Method generally work well when $Q(\boldsymbol{\beta})$ is globally convex.

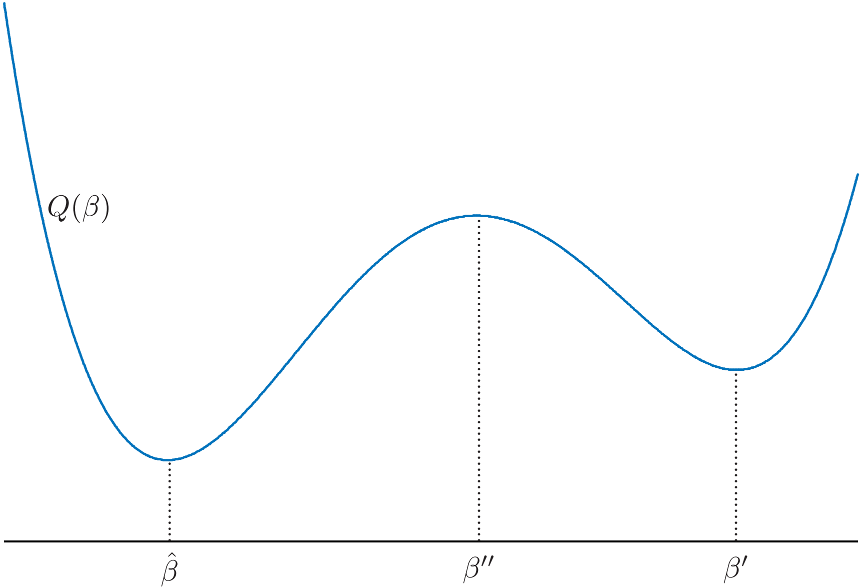
- For such a function, there can be at most one local minimum, which is also the global minimum.
- Optimization methods of this type can fail when there are multiple minima. They generally converge to a local minimum, but there is no guarantee that it is the global one.

The choice of the **starting values** $\boldsymbol{\beta}_{(0)}$ can be extremely important.

The usual way to guard against finding the wrong local minimum is to minimize $Q(\boldsymbol{\beta})$ several times, for several different starting values.

This is not feasible unless k is very small. For just 10 starting values for each of k parameters, the total number of starting values is 10^k.

Global optimization methods, such as **simulated annealing**, **particle swarm optimization**, and **ant colony optimization** are designed to find global optima when there are many local ones.


$Q(\beta)$
$\hat{\beta}$
β''
β'

The Gauss-Newton Regression

The **Gauss-Newton regression** or **GNR** is a linear approximation to a nonlinear regression model. It is an example of an **artificial regression**.

A first-order Taylor series approximation to $\boldsymbol{x}(\boldsymbol{\beta})$ at $\bar{\boldsymbol{\beta}}$ is

$$\boldsymbol{x}(\boldsymbol{\beta}) \cong \boldsymbol{x}(\bar{\boldsymbol{\beta}}) + \boldsymbol{X}(\bar{\boldsymbol{\beta}})(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}). \tag{27}$$

Substituting this into $\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}$, moving the first term to the left-hand side, and replacing $\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}$ by $\boldsymbol{b}$ yields the GNR:

$$\boldsymbol{y} - \boldsymbol{x}(\bar{\boldsymbol{\beta}}) = \boldsymbol{X}(\bar{\boldsymbol{\beta}})\boldsymbol{b} + \text{residuals}. \tag{28}$$

It is simply a regression of $\boldsymbol{y} - \boldsymbol{x}(\bar{\boldsymbol{\beta}})$, the residuals for $\boldsymbol{\beta} = \bar{\boldsymbol{\beta}}$, on $\boldsymbol{X}(\bar{\boldsymbol{\beta}})$, the matrix of derivatives of $\boldsymbol{x}(\boldsymbol{\beta})$.

In practice, we have to replace $\bar{\boldsymbol{\beta}}$ by some value of $\boldsymbol{\beta}$. Which one depends on what we want to use the GNR for.

1. Gauss-Newton Optimization

Since $\boldsymbol{H}(\boldsymbol{\beta}) \overset{a}{=} \boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})$, and the latter is positive definite, it makes sense to set $\boldsymbol{D}(\boldsymbol{\beta}) = \boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})$.

Furthermore, $\boldsymbol{g}(\boldsymbol{\beta}) = \boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{u}$.

We can use the GNR in a quasi-Newton procedure. If $\boldsymbol{\beta} = \boldsymbol{\beta}_{(s)}$,

$$\boldsymbol{b}_{(s)} = \big(\boldsymbol{X}_{(s)}^\top \boldsymbol{X}_{(s)}\big)^{-1}\boldsymbol{X}_{(s)}^\top(\boldsymbol{y} - \boldsymbol{x}_{(s)}). \tag{29}$$

But if $\boldsymbol{D}_{(s)} = \boldsymbol{X}_{(s)}^\top \boldsymbol{X}_{(s)}$ and $\boldsymbol{g}_{(s)} = \boldsymbol{X}_{(s)}^\top(\boldsymbol{y} - \boldsymbol{x}_{(s)})$, this is just

$$\boldsymbol{b}_{(s)} = \boldsymbol{D}_{(s)}^{-1}\boldsymbol{g}_{(s)}. \tag{30}$$

Thus $\boldsymbol{b}_{(s)}$ gives us a direction. Setting $\alpha_{(s)} = 1$ or choosing it by line search gives us the Gauss-Newton method.

2. NLS Covariance Matrices

Suppose we evaluate the GNR at $\hat{\boldsymbol{\beta}}$. It is

$$\boldsymbol{y} - \hat{\boldsymbol{x}} = \hat{\boldsymbol{X}}\boldsymbol{b} + \text{residuals.} \tag{31}$$

We should find that

$$\hat{\boldsymbol{b}} = (\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^\top(\boldsymbol{y} - \hat{\boldsymbol{x}}) = \mathbf{0}, \tag{32}$$

because $\hat{\boldsymbol{X}}^\top(\boldsymbol{y} - \hat{\boldsymbol{x}}) = \hat{\boldsymbol{X}}^\top\hat{\boldsymbol{u}} = \mathbf{0}$. If all the elements of $\hat{\boldsymbol{b}}$ are not very close to 0, the NLS routine is defective.

The standard covariance matrix of $\hat{\boldsymbol{b}}$ is

$$s^2(\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})^{-1}, \tag{33}$$

where s^2 is the same as for the nonlinear regression because $\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{u}} = \mathbf{0}$. Regressing $\hat{\boldsymbol{u}}$ on $\hat{\boldsymbol{X}}$ does not change it.

We can also use the GNR to obtain hetero-robust, cluster-robust, or HAC covariance matrices. For example, the cluster-robust one is

$$\left(\frac{G}{G-1}\right)\left(\frac{N-1}{N-k}\right)(\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})^{-1}\left(\sum_{g=1}^{G}\hat{\boldsymbol{X}}_g^\top\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g^\top\hat{\boldsymbol{X}}_g\right)(\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})^{-1}, \tag{34}$$

where $\hat{\boldsymbol{X}}_g$ contains the rows of $\hat{\boldsymbol{X}}$ for cluster g.

3. Testing Restrictions

We wish to test the restriction that $\boldsymbol{\beta}_2 = \boldsymbol{0}$ in the model

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u} = \boldsymbol{x}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) + \boldsymbol{u}. \tag{35}$$

It is easy to perform the test as an F test based on the GNR

$$\boldsymbol{y} - \tilde{\boldsymbol{x}} = \tilde{\boldsymbol{X}}\boldsymbol{b} + \text{residuals}, \tag{36}$$

where $\tilde{\boldsymbol{x}} \equiv \boldsymbol{x}(\tilde{\boldsymbol{\beta}})$ and $\tilde{\boldsymbol{X}} \equiv \boldsymbol{X}(\tilde{\boldsymbol{\beta}})$.

The GNR can also be written as

$$\boldsymbol{y} - \tilde{\boldsymbol{x}} = \tilde{\boldsymbol{X}}_1 \boldsymbol{b}_1 + \tilde{\boldsymbol{X}}_2 \boldsymbol{b}_2 + \text{residuals}. \tag{37}$$

The first-order conditions for $\tilde{\boldsymbol{\beta}}$ imply that

$$(\boldsymbol{y} - \tilde{\boldsymbol{x}})^\top \tilde{\boldsymbol{X}}_1 = \boldsymbol{0}. \tag{38}$$

This does *not* mean that we can omit $\tilde{\boldsymbol{X}}_1$ from (37). But it does mean that $\tilde{\boldsymbol{M}}_1(\boldsymbol{y} - \tilde{\boldsymbol{x}}) = \boldsymbol{y} - \tilde{\boldsymbol{x}}$.

The FWL regression that corresponds to equation (37) is

$$\tilde{\boldsymbol{M}}_1(\boldsymbol{y} - \tilde{\boldsymbol{x}}) = \boldsymbol{y} - \tilde{\boldsymbol{x}} = \tilde{\boldsymbol{M}}_1 \tilde{\boldsymbol{X}}_2 \boldsymbol{b}_2 + \text{residuals}. \tag{39}$$

Therefore, the numerator of the F statistic is just $1/k_2$ times

$$(\boldsymbol{y} - \tilde{\boldsymbol{x}})^\top \tilde{\boldsymbol{X}}_2 (\tilde{\boldsymbol{X}}_2^\top \tilde{\boldsymbol{M}}_1 \tilde{\boldsymbol{X}}_2)^{-1} \tilde{\boldsymbol{X}}_2^\top (\boldsymbol{y} - \tilde{\boldsymbol{x}}). \tag{40}$$

It can be shown that $1/s^2$ times this is asymptotically $\chi^2(k_2)$ when the disturbances are i.i.d.

We can also use the GNR (37) to compute a hetero-robust, or a cluster-robust, or perhaps a HAC Wald statistic for $\boldsymbol{b}_2 = \mathbf{0}$.

Of course, we can also compute Wald tests directly using $\hat{\boldsymbol{\beta}}$, without computing $\tilde{\boldsymbol{\beta}}$ or running a GNR.

For the i.i.d. case, we can use F tests based on $\mathrm{SSR}(\tilde{\boldsymbol{\beta}})$ and $\mathrm{SSR}(\hat{\boldsymbol{\beta}})$.

3a. Testing for Serial Correlation

Consider the linear regression model with autoregressive errors.

In this case, H_0 is the model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \mathrm{IID}(0, \sigma^2), \tag{41}$$

and H_1 is the model

$$y_t = \rho y_{t-1} + \boldsymbol{X}_t\boldsymbol{\beta} - \rho \boldsymbol{X}_{t-1}\boldsymbol{\beta} + u_t, \quad u_t \sim \mathrm{IID}(0, \sigma^2). \tag{42}$$

We wish to test the null hypothesis that $\rho = 0$.

The GNR that corresponds to (42) is

$$\begin{aligned} y_t - \rho y_{t-1} - \boldsymbol{X}_t\boldsymbol{\beta} + \rho\boldsymbol{X}_{t-1}\boldsymbol{\beta} \\ = (\boldsymbol{X}_t - \rho\boldsymbol{X}_{t-1})\boldsymbol{b} + b_\rho(y_{t-1} - \boldsymbol{X}_{t-1}\boldsymbol{\beta}) + \text{residual}, \end{aligned} \tag{43}$$

where $\boldsymbol{b}$ corresponds to $\boldsymbol{\beta}$ and b_ρ corresponds to ρ.

If (43) is evaluated at $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ and $\rho = 0$, it becomes

$$y_t - \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} = \boldsymbol{X}_t\boldsymbol{b} + b_\rho(y_{t-1} - \boldsymbol{X}_{t-1}\tilde{\boldsymbol{\beta}}) + \text{residual}. \tag{44}$$

If we denote the OLS residuals from (41) by $\tilde{u}_t$, the GNR (43) takes on the very simple form

$$\tilde{u}_t = \boldsymbol{X}_t\boldsymbol{b} + b_\rho\tilde{u}_{t-1} + \text{residual}. \tag{45}$$

This is just a linear regression of the OLS residuals $\tilde{\boldsymbol{u}}$ on $\boldsymbol{X}$ and the residuals lagged once.

A suitable test statistic is the t statistic for the artificial parameter b_ρ in (45) to equal 0.

We can replace the regressand in (45) by y_t. By the FWL Theorem, the t statistic is numerically identical.

This procedure, due to Durbin (1970), can easily be generalized to

- nonlinear regression models;
- tests for higher-order autoregressive errors;
- tests for moving-average errors.

For example, to test for AR(2) errors in a nonlinear regression model, we would run the test regression

$$\tilde{u}_t = \tilde{\boldsymbol{X}}_t \boldsymbol{b} + b_{\rho_1} \tilde{u}_{t-1} + b_{\rho_2} \tilde{u}_{t-2} + \text{residual}, \tag{46}$$

where $\tilde{\boldsymbol{X}}_t$ is the vector of derivatives of $x_t(\boldsymbol{\beta})$ evaluated at the NLS estimates $\tilde{\boldsymbol{\beta}}$.

Near the null hypothesis, an MA(p) process looks like an AR(p) process; see Godfrey (1978a,b). So tests against MA(p) and AR(p) errors are the same.

3b. Optimization Tricks

For purposes of numerical optimization, is often convenient to divide $\boldsymbol{\theta}$ into two subvectors.

Perhaps we can **concentrate** the objective function by writing

$$\min_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} \mathrm{SSR}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \min_{\boldsymbol{\theta}_1} \mathrm{SSR}\big(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(\boldsymbol{\theta}_1)\big), \tag{47}$$

where $\boldsymbol{\theta}_2(\boldsymbol{\theta}_1)$ minimizes $\mathrm{SSR}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with respect to $\boldsymbol{\theta}_2$ conditional on $\boldsymbol{\theta}_1$.

This is true for (42), where ρ and $\boldsymbol{\beta}$ play the roles of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. If we ignore the first observation, we have

$$\mathrm{SSR}(\rho, \boldsymbol{\beta}) = \sum_{t=2}^{T} (y_t - \rho y_{t-1} - \boldsymbol{X}_t \boldsymbol{\beta} + \rho \boldsymbol{X}_{t-1} \boldsymbol{\beta})^2. \tag{48}$$

Conditional on ρ, the least squares estimate of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}(\rho) = \big(\boldsymbol{X}^\top(\rho)\boldsymbol{X}(\rho)\big)^{-1}\boldsymbol{X}^\top(\rho)\,\boldsymbol{y}(\rho), \tag{49}$$

where the t^{th} rows of $\boldsymbol{X}(\rho)$ and $\boldsymbol{y}(\rho)$ are

$$\boldsymbol{X}_t - \rho\boldsymbol{X}_{t-1} \quad \text{and} \quad \boldsymbol{y}_t - \rho\,\boldsymbol{y}_{t-1} \quad \text{for } t = 2, \ldots, T. \tag{50}$$

We can use any sort of one-dimensional search routine to find $\hat{\rho}$ that minimizes $\text{SSR}(\rho) = \text{SSR}\big(\rho, \boldsymbol{\beta}(\rho)\big)$. This gives us $\hat{\boldsymbol{\beta}}$ as a byproduct.

But the covariance matrix of $\hat{\boldsymbol{\beta}}$ from the regression of $\boldsymbol{y}(\hat{\rho})$ on $\boldsymbol{X}(\hat{\rho})$ is, in general, wrong. It is correct only if everything in $\boldsymbol{X}$ is exogenous.

Instead, we need to run the GNR (43) evaluated at $\hat{\rho}$ and $\hat{\boldsymbol{\beta}}$:

$$\boldsymbol{y} - \hat{\rho}\boldsymbol{y}_{-1} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\rho}\boldsymbol{X}_{-1}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\hat{\rho})\boldsymbol{b} + b_\rho(\boldsymbol{y}_{-1} - \boldsymbol{X}_{-1}\hat{\boldsymbol{\beta}}) + \text{residuals}. \tag{51}$$

The covariance matrix of $\boldsymbol{b}$ and b_ρ is $\widehat{\text{Var}}(\hat{\rho}, \hat{\boldsymbol{\beta}})$.

Another possibility is to employ **sequential minimization**.

1. Choose $\boldsymbol{\theta}_2^{(0)}$ somehow, and set $s = 0$.
2. Minimize $\mathrm{SSR}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ w.r.t. $\boldsymbol{\theta}_1$ conditional on $\boldsymbol{\theta}_2^{(s)}$ to find $\boldsymbol{\theta}_1^{(s+1)}$.
3. Minimize $\mathrm{SSR}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ w.r.t. $\boldsymbol{\theta}_2$ conditional on $\boldsymbol{\theta}_1^{(s+1)}$ to find $\boldsymbol{\theta}_2^{(s+1)}$.
4. Check to see whether we are close enough to $\hat{\boldsymbol{\theta}}$.
5. If not, return to #2 and increment s by 1.

If the algorithm converges (it may not!), we have found $[\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2]$. Then use the GNR (or another artificial regression, or just matrix algebra) to obtain $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$. Why use an artificial regression?

Since economists very rarely estimate models that are nonlinear in every parameter, this sort of procedure is often very useful.

Minimization procedures that concentrate the objective function may be able to handle more severe nonlinearities than ones that use sequential minimization, at least when $\boldsymbol{\theta}_1$ is low-dimensional.

Bootstrap Inference in Nonlinear Regression Models

Finite-sample inference in nonlinear regression models, and other nonlinear models, can be problematic.

- In most cases, $\hat{\boldsymbol{\beta}}$ is biased, sometimes severely.
- $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ may not be close to being normally distributed with covariance matrix $(\boldsymbol{X}_0^\top \boldsymbol{X}_0)^{-1} \boldsymbol{X}_0^\top \boldsymbol{\Omega} \boldsymbol{X}_0 (\boldsymbol{X}_0^\top \boldsymbol{X}_0)^{-1}$.
- $\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}}$ may provide a poor approximation to $\boldsymbol{X}_0^\top \boldsymbol{X}_0$.
- $\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{\Omega}} \hat{\boldsymbol{X}}$ may provide a poor approximation to $\boldsymbol{X}_0^\top \boldsymbol{\Omega} \boldsymbol{X}_0$.
- There is likely to be correlation between $\hat{\boldsymbol{\beta}}$ and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$.

Thus hypothesis tests based on asymptotic theory may over-reject (or, much less commonly, under-reject) severely.

Similarly, confidence sets may under-cover (or over-cover).

Bootstrap methods usually provide more reliable inferences.

But bootstrapping can be expensive when a model is nonlinear.

For hypothesis tests, we can often get away with a small value of B by using a trick proposed in Davidson and MacKinnon (ER, 2000).

1. Start with $B = 99$ and compute $\hat{p}^*$.
2. Stop if $\hat{p}^*$ is *significantly* larger or smaller than α, or if $B \geq B_{\max}$.
3. If necessary, add more bootstrap samples, taking B to, say, $2B + 1$. Compute $\hat{p}^*$ again, and go to #2.

Eventually, we either stop with $B \geq B_{\max}$ and $\hat{p}^*$ very close to α, or with $\hat{p}^*$ clearly on one side of α.

When testing whether $\hat{p}^* = \alpha$ in step #2, use a very small level for the test (e.g., .0001).

It is not clear whether this trick could be adapted to work with bootstrap confidence intervals.

Another trick is to stop the nonlinear estimation early for the bootstrap samples; see Davidson and MacKinnon (IER, 1999).

Estimation of nonlinear models using bootstrap data has an advantage over estimation using real data: The model we are estimating actually generated the data, and we can start at the "true" value of $\boldsymbol{\beta}$.

But nonlinear estimation can fail for some bootstrap samples.

- Perhaps $\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})$ is numerically singular for some value of $\boldsymbol{\beta}$ that the Gauss-Newton algorithm encounters.
- If this happens, the algorithm can get stuck in a space of dimension lower than k and never converge.
- Well-written numerical optimization programs check for this sort of thing, but poorly-written ones can either die or iterate forever.
- We may have to throw out bootstrap samples where this happens.
- This is not a big deal if it happens in 2 out of 999 cases.
- But it is a big deal if it happens in 83 out of 999! If the problem cannot be corrected, it should certainly be reported.