Economics 620, Lecture 18: Nonlinear Models

Nicholas M. Kiefer

Cornell University

The basic point is that smooth nonlinear models look like linear models locally.
Models linear in parameters are no problem even if they are nonlinear in variables. For example

$$\varphi(y) = \beta_0 + \beta_1\psi_1(x_1) + \beta_2\psi_2(x_2) + ...$$

with φ and ψ known functions of observable regressors, is still a linear regression model. However,

$$y = \theta_1 + \theta_2 e^{x\theta_3} + \varepsilon$$

is nonlinear (arising, for example, as a solution to a differential equation).

Notation:

$y_i = f(X_i, \theta) + \varepsilon_i = f_i(\theta) + \varepsilon_i \quad i = 1, 2, ..., N.$

Stacking up yields
$y = f(\theta) + \varepsilon$
where y is $N \times 1$, $f(\theta)$ is $N \times 1$, θ is $K \times 1$ and ε is $N \times 1$. For X_i $i = 1, 2, ..., N$ fixed, f: $\mathbb{R}^K \to \mathbb{R}^N$.

$$\frac{\partial f}{\partial \theta'} = F(\theta) = \begin{bmatrix} \frac{\partial f_1}{\partial \theta_1} & \cdots & \frac{\partial f_1}{\partial \theta_K} \\ \bullet & \cdots & \bullet \\ \frac{\partial f_N}{\partial \theta_1} & \cdots & \frac{\partial f_N}{\partial \theta_K} \end{bmatrix}.$$

Obviously $F(\theta)$ is $N \times K$. Assume $E\varepsilon = 0$ and $V\varepsilon = \sigma^2 I$.

The nonlinear least squares estimator minmizes
$S(\theta) = \sum_{i=1}^{N}(y_i - f_i(\theta))^2 = (y - f(\theta))'(y - f(\theta))$.
Differentiation yields

$$\begin{aligned}\frac{\partial}{\partial \theta'} S(\theta) &= \frac{\partial}{\partial \theta'}(y - f(\theta))'(y - f(\theta)) \\ &= -2(y - f(\theta))' F(\theta).\end{aligned}$$

Thus, the nonlinear least squares (NLS) estimator $\hat{\theta}$ satisfies

$$F(\hat{\theta})'(y - f(\hat{\theta})) = 0.$$

(*Are these equations familiar*?)

This is like the property $X'e = 0$ in the LS method.

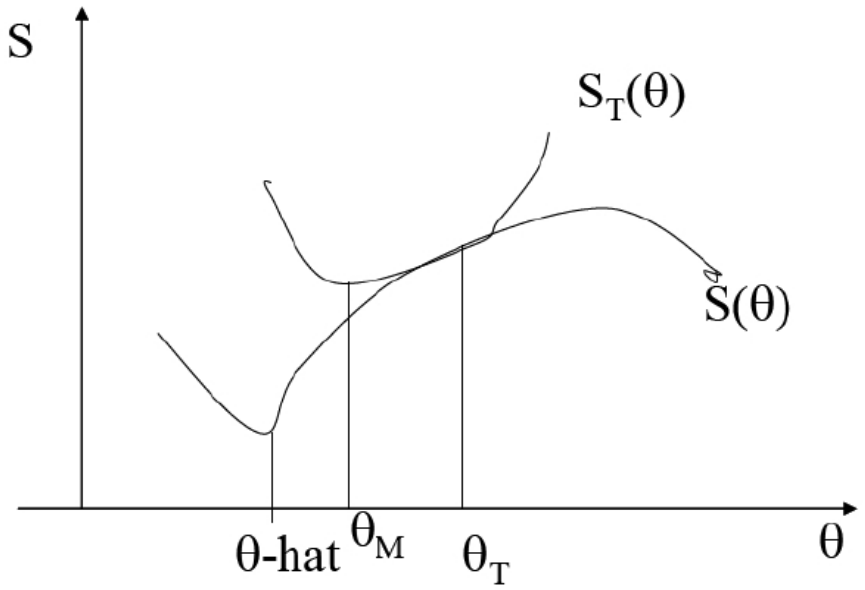
Computing $\hat{\theta}$:

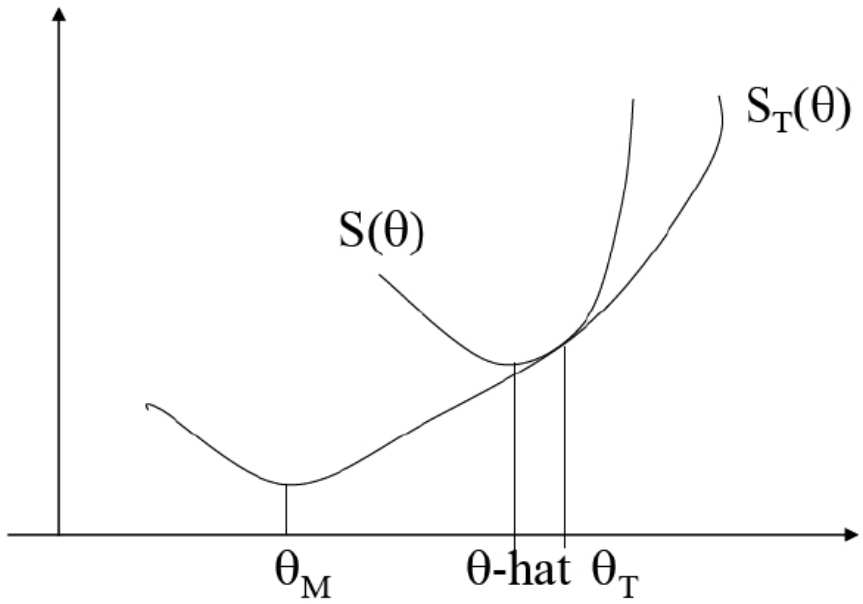
The *Guass-Newton method* is to use a first-order expansion of $f(\theta)$ and θ_T (a "trial" value) in $S(\theta)$, giving
$S_T(\theta) = (y - f(\theta_T) - F(\theta_T)(\theta - \theta_T))'(y - f(\theta_T) - F(\theta_T)(\theta - \theta_T))$.

Minimizing S_T in θ gives
$\theta_M = \theta_T + [F(\theta_T)'F(\theta_T)]^{-1}F(\theta_T)'(y - f(\theta_T))$.

(*Exercise*: Show θ_M is the minimizer.)

We know $S_T(\theta_M) \leq S_T(\theta_T)$.
Is it true that $S(\theta_M) \leq S(\theta_T)$?


S
$S_T(\theta)$
$S(\theta)$
θ-hat
θ_M
θ_T
θ


$S_T(\theta)$
$S(\theta)$
θ_M
θ-hat
θ_T

The method is to use θ_M for the new trial value, expand again, minimize and iterate. But there can be problems. For example, it is possible that $S(\theta_M) > S(\theta_T)$, but there is a θ^* between θ_T and θ_M,

$$\theta^* = \theta_T + \lambda(\theta_M - \theta_T)$$

for some $\lambda < 1$, with $S(\theta^*) \leq S(\theta_T)$. This suggests trying a decreasing sequence of λ's in $[0, 1]$, leading to the modified *Gauss-Newton method*.

(1) Start with θ_T and compute

$$D_T = [F(\theta_T)'F(\theta_T)]^{-1}F(\theta_T)'(y - f(\theta_T)).$$

(2) Find λ such that $S(\theta_T + \lambda D_T) < S(\theta_T)$.
(3) Set $\theta^* = \theta_T + \lambda D_T$ and go to (1) using θ^* as a new value for θ_T.

Stop when the changes in parameter values and S between iterations are small. Good practice is to try several different starting values.

Estimation of $\sigma^2 = V(\varepsilon)$:

$$
\begin{aligned}
s^2 &= \frac{(y - f(\hat{\theta}))'(y - f(\hat{\theta}))}{N - k} \\
\sigma^2 &= \frac{(y - f(\hat{\theta}))'(y - f(\hat{\theta}))}{N}
\end{aligned}
$$

Why are there two possibilities?

Note the simplification possible when the model is linear in some of the parameters (as in the example)

$$y = \theta_1 + \theta_2 \exp(\theta_3 x) + \varepsilon.$$

Here, given θ_3, the other parameters can be estimated by OLS. Thus the sum of squares function S can be concentrated, that is written as a function of one parameter alone. The nonlinear maximization problem is 1-dimensional, not 3. This is an important trick.

We used the same device to estimate models with autocorrelation the $\rho-$differenced model could be esimated by OLS conditional on ρ.

Inference:

We can show that

$$\hat{\theta} = \theta_0 + (F(\theta_0)'F(\theta_0))^{-1}F(\theta_0)'\varepsilon + r$$

where θ_0 is the true parameter value and plim$N^{1/2}r = 0$ (show). So r can be ignored in calculating the asymptotic distribution of $\hat{\theta}$. This is just like the expression in the linear model - decomposing $\hat{\beta}$ into the true value plus sampling error.
Thus the asymptotic distribution of $N^{1/2}(\hat{\theta} - \theta_0)$ is

$$N\left(0, \sigma^2\left(\frac{(F(\theta_0)'F(\theta_0)}{N}\right)^{-1}\right).$$

So the approximate distribution of $\hat{\theta}$ becomes

$$N(\theta_0, \sigma^2(F(\theta_0)'F(\theta_0))^{-1}).$$

In practice σ^2 is estimated by s^2 or $\hat{\sigma}^2$ and

$$F(\theta_0)'F(\theta_0)$$

is estimated by

$$F(\hat{\theta})'F(\hat{\theta}).$$

Check that this is OK.

Applications:

1. The overidentified SEM is a nonlinear regression model (linear in variables, nonlinear in parameters) - consider the reduced form equations in terms of structural parameters.
2. Many other models - more to come.

Efficiency: Consider another consistent estimator θ^* with "sampling error" in the form
$n^{1/2}(\theta^* - \theta) = (F(\theta_0)'F(\theta_0))^{-1}F(\theta_0)'\varepsilon + C'\varepsilon + r.$

It can be shown, in a proof like that of the Gauss-Markov Theorem, that the minimum variance estimator in this class (locally linear) has $C = 0$.

A better property holds when the errors are iid normal. Then, the NLS estimator is the MLE, and we have the Cramer-Rao efficiency result.

Nonlinear SEM

Let $y_i = f(Z_i, \theta) + \varepsilon_i = f_i(\theta) + \varepsilon_I$.

Where now Z is used to indicate included endogenous variables. NLS will not be consistent (why not?). The trick is to find instruments W and look at

$$W'y = W'f(\theta) + W'\varepsilon.$$

If the model is just-identified, we can just set $W'\varepsilon = 0$ (its expected value) and solve. Otherwise, we can do nonlinear GLS, minimizing the variance-weighted sum of squares

$$(y - f(\theta))'W(W'W)^{-1}W'(y - f(\theta)).$$

This $\hat{\theta}$ is called the nonlinear 2SLS estimator (by Amemiya, who studied its properties in 1974). Note that there are not two separate stages.

Specifically, it might be tempting to just obtain a first state $\hat{Z} = (I - M)Z = [\hat{Y}_2 - X_1]$ and do nonlinear regression of y on $f(\hat{Z}, \theta)$.

This does not necessarily work. $\hat{Z}$ is orthogonal to ε, but f may not be. In the linear regression case, we want $\hat{Z}'\varepsilon = 0$. The corresponding condition here is $\hat{F}'\varepsilon = 0$. Since $\hat{F}$ depends generally on θ, there is no real way to do this in stages.

Asymptotic Distribution Theory

$$n^{-1/2}(\theta - \theta_0)$$
$$\rightarrow \quad N\left(0, \sigma^2 \left(\frac{(F(\theta_0)'(W(W'W)^{-1}W')F(\theta_0)}{N}\right)^{-1}\right).$$

In practice σ^2 is estimated by

$$(y - f(Z, \hat{\theta}))'(y - f(Z, \hat{\theta}))/n$$

(or over $n - k$) and $F(\theta_0)$ is esimated by $F(\hat{\theta})$.
Don't forget to remove the factor N^{-1} in the variance when approximating the variance of your estimator.
The calculation is exactly like the usual calculation of the variance of the GLS estimator.

MISCELLANEOUS:

Proposition:
$\sum_{I=1}^{\infty}(f_i(\theta_0) - f_i(\theta'))^2 = \infty$, $\theta' \neq \theta_0$
is necessary for the existence of a consistent NLS estimator. (*Compare this with OLS.*)

Funny example: Consider $y_i = e^{-\alpha i} + \varepsilon_i$ where $\alpha \in (0, 2\pi)$ and $V(\varepsilon_i) = \sigma^2$. Is there a consistent estimator?

Consistency can be shown (generally, not just in this example) using regularity conditions, not requiring derivatives. Normality can be shown using first but not second derivatives.

Many models can be set up as nonlinear regressions - like qualitative dependent variable models. Sometimes it is hard to interpret parameters. We mgiht be interested in functions of parameters, for example elasticities.

Tests on functions of parameters:
Remember the delta-method. Let $\hat{\theta} \sim N(\theta_0, \sum)$ be a consistent estimator of θ whose true value is θ_0. Suppose $g(\theta)$ is the quantity of interest. Expanding $g(\hat{\theta})$ around θ_0 yields
$g(\hat{\theta}) = g(\theta_0) + \frac{\partial g}{\partial \theta}|_{\theta=\theta_0}(\hat{\theta} - \theta_0) +$ more.

This implies that

$$V(g(\hat{\theta})) = E(g(\hat{\theta}) - g(\theta_0))^2 \approx \left(\frac{\partial g}{\partial \theta_0}\right) \sum \left(\frac{\partial g}{\partial \theta_0}\right)'.$$

Asymptotically, $g(\hat{\theta}) \sim N(g(\theta_0), V(g(\hat{\theta}))$. This is a very useful method. It also shows that normal approximation may be poor. (*Why*?) How can the normal approximation be improved? This is a deep question. One trick is to choose a parametrization for which $\partial^2 \ell / \partial\theta \partial\theta'$ is constant or nearly so. Why does this work?

Consider the Taylor expansion of the score.....