M1 INTERMEDIATE ECONOMETRICS

NONLINEAR MODELS

Koen Jochmans

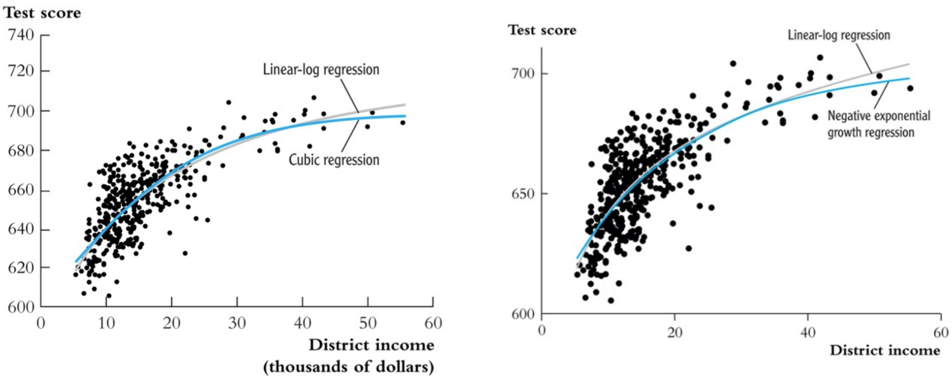
October 29, 2025

Nonlinear models arise naturally in a variety of situations. We discuss several pertinent examples here.

1. Negative exponential-growth model

Our first example goes back to the analysis of the Californian test score data. Recall that we had observed a concave pattern between (average) income and test performance. The left panel in Figure 1 contains the scatterplot from before along with fits based on a linear-log specification and on a cubic approximation in income. Both curves are nonlinear in income but remain linear in parameters; hence, they could be computed using (multiple) least squares.

Figure 1: Test score on income



Our main argument for preferring such specifications over a simple linear regression was that they are able to accommodate the apparent nonlinear relationship between income and test scores. The linear-log specification was especially attractive because of its simple interpretation and its parsimony. Both are, however, leading to predicted test scores that are unbounded as a function of income. However, tests are graded on a bounded scale; they cannot increase beyond the maximum. This naturally implies a nonlinear relationship.

A simple parsimonious specification for the mean of test scores (Y) as a function of income (X) would be the negative exponential-growth model, which is

$$Y = \alpha(1 - e^{-\beta X}) + e, \qquad \mathbb{E}(e|X) = 0.$$

for parameters $\theta = (\alpha, \beta)$. The predicted test score for income level x then is

$$\hat{y}(x) = \alpha(1 - e^{-x\beta}).$$

Clearly (with $\beta > 0$),

$$\lim_{x \downarrow 0} \hat{y}(x) = 0 \qquad \lim_{x \uparrow +\infty} \hat{y}(x) = \alpha,$$

and so lies between the minimum value of zero and maximum value of α. Furthermore, a marginal increase in income changes the predicted test score by

$$\frac{\partial \hat{y}(x)}{\partial x} = \alpha\beta e^{-x\beta}.$$

This reflects decreasing returns, as

$$\frac{\partial \hat{y}(x)/\partial x}{\hat{y}(x)} = \beta\,\Lambda(x\beta), \qquad \Lambda(a) = \frac{e^{-a}}{1 - e^{-a}},$$

is monotone decreasing. The right panel in Figure 1 shows the fit of the negative exponential-growth model (by nonlinear least squares) on the test score data.

2. Production functions

A second example, which we have also encountered before, is production functions. Suppose that we are interested in learning production function for output Y under CES technology with two (equally-weighted) inputs X_1, X_2. The production function is

$$Y = A\,(X_1^\gamma + X_2^\gamma)^{1/\gamma},$$

where A is random factor productivity, and γ is the substitution parameter. We do not observe A. Let $\alpha = \mathbb{E}(A|X_1, X_2)$. Then we can write $A = \alpha e$ for some e that satisfies $\mathbb{E}(e|X_1, X_2) = 1$. This implies that

$$\mathbb{E}(Y|X_1, X_2) = \alpha\,(X_1^\gamma + X_2^\gamma)^{1/\gamma}.$$

or, equivalently, $Y = \alpha\,(X_1^\gamma + X_2^\gamma)^{1/\gamma}\,e$, This is again a nonlinear problem, with parameters $\theta = (\alpha, \gamma)$.

A not uncommon approach to proceeding in an attempt to evade the nonlinearity is to linearize the problem by taking logs on the left and right of the above equation to arrive at

$$\log(Y) = \log(\alpha) + \frac{1}{\gamma}\log(X_1^\gamma + X_2^\gamma) + \log(e).$$

While this step is fine, it is important to realize that $\mathbb{E}(e|X_1, X_2) = 1$ does not imply that

$$\mathbb{E}(\log(e)|X_1, X_2) = 0.$$

Hence,

$$\mathbb{E}(\log(Y)|X_1, X_2) \neq \log(\alpha) + \frac{1}{\gamma}\log(X_1^\gamma + X_2^\gamma)$$

in general. (When e is fully independent of X_1, X_2, log-linearization will only affect the constant term, which becomes $\log(\alpha) + \mathbb{E}(\log(e)) = \mathbb{E}(\log A)$.) So, (X_1, X_2) are effectively to be treated as endogenous in the linearized equation under the conditions on the model in levels. Conversely, imposing instead the restriction that $\mathbb{E}(\log(e)|X_1, X_2) = 0$ does not imply that $\mathbb{E}(e|X_1, X_2) = 1$ holds.

3. Censoring

Another issue that can lead to nonlinearity are various data limitations. One empirically relevant example is the issue of bottom- or top-coding censoring. To proceed with a specific example, the variable Y^* is said to be left-censored at zero when we observe

$$Y = \max(Y^*, 0),$$

but not Y^* itself. Right censoring would be analogous and leads to top-coded variables instead.

To see how censoring creates dfficulties let us work with the classical linear regression model, where

$$Y^* = X'\beta + e, \qquad e|X \sim N(0, \sigma^2).$$

Then censoring below zero leads to the (observed) distribution $Y|X$ having a mass point at zero.

It is not obvious how to construct an estimator for β (or σ^2) here. A naive approach to 'dealing' with censoring is to retain only data that is not censored. That is, proceed with a least squares regression of Y on X given that $Y > 0$. However,

$$\mathbb{E}(Y^*|X, Y^* > 0) = X'\beta + \mathbb{E}(e|X, Y^* > 0) = X'\beta + \mathbb{E}(e|X, e > -X'\beta).$$

The last term can be worked out by virtue of normality of the errors. Indeed, for any (a, u) with $a < u$,

$$\mathbb{P}(e \leq u|e > a) = \frac{\mathbb{P}(e \leq u) - \mathbb{P}(e \leq a)}{1 - \mathbb{P}(e \leq a)} = \frac{\Phi(u/\sigma) - \Phi(a/\sigma)}{1 - \Phi)(a/\sigma)}.$$

The conditional density is found by taking the derivative of this expression with respect to u and equals

$$\frac{(1/\sigma)\,\phi(u/\sigma)}{1 - \Phi(a/\sigma)}.$$

Here, the denominator appears because of the censoring. Now, the mean of this conditional density function is

$$\mathbb{E}(e|e > a) = \frac{\int_a^\infty (u/\sigma)\phi(u/\sigma)\,du}{1 - \Phi(a/\sigma)} = \sigma \frac{\int_{a/\sigma}^\infty z\phi(z)\,dz}{1 - \Phi(a/\sigma)} = \sigma \frac{\phi(a/\sigma)}{1 - \Phi(a/\sigma)} =: \lambda_\sigma(a),$$

where we have used the change of variable $z = u/\sigma$ and the last step follows from fact that $\phi'(z) = -z\phi(z)$. Therefore, while the population regression line in the uncensored population is $x'\beta$, the one in the subpopulation of non-censored units is

$$x'\beta + \lambda_\sigma(-x'\beta),$$

which is again nonlinear.

The notion of censored versus non-censored population also relates to the interpretation of the parameters. Here, β captures average partial effects of a change in X on Y^*, as before in the classical linear regression model; that is,

$$\frac{\partial \mathbb{E}(Y^*|X)}{\partial X} = \beta.$$

The marginal effect on Y is nonlinear, however. We have

$$\frac{\partial \mathbb{E}(Y|X)}{\partial X} = \frac{\partial \mathbb{E}(Y|X, Y > 0)\, \mathbb{P}(Y > 0|X)}{\partial X} = \beta\, \Phi\left(\frac{X'\beta}{\sigma}\right),$$

which follows from an application of the chain rule, using the calculations from above. This is a function of X. The average marginal effect is the average of this quantity with respect to X,

$$\mathbb{E}\left(\beta\, \Phi\left(\frac{X'\beta}{\sigma}\right)\right).$$

Here, the nonlinear term accounts for the fact that a change in X also has an effect on whether or not censoring will occur.

3.1. Binary outcomes

Remaining within the classical normal model,

$$Y^* = X'\beta + e, \qquad e|X \sim N(0, \sigma^2),$$

an extreme form of censoring is to only observe the sign of Y^*. That is, we have

$$Y = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{if } Y^* \leq 0 \end{cases}.$$

In this case,

$$\mathbb{E}(Y|X) = \mathbb{P}(Y = 1|X) = \mathbb{P}(e \geq -X'\beta|X) = 1 - \Phi(-X'\theta) = \Phi(X'\theta)$$

for $\theta = \beta/\sigma$. Indeed, $Y|X$ is Bernoulli distributed with success probability $\Phi(X'\theta)$. This distribution depends on the parameters β and σ^2 only through the ratio θ. Statistically we will not be able to discriminate between different combinations of (β, σ^2) that lead to the same θ.

Here, the marginal effect is

$$\frac{\partial \mathbb{P}(Y = 1|X)}{\partial X} = \theta \, \phi(X'\theta).$$

This is larger (in magnitude) when $X'\theta$ is closer to zero. This is sensible as a small change in X will not greatly affect the probability of a sign switch when $|X'\theta|$ is already large.

Binary models are often used in the context of choice behavior based on a revealed-preference argument. Suppose that u_1 is the utility derived from an action (e.g., buying a good); u_0 is the utility from the outside option (not buying). Then

$$Y = \begin{cases} 1 & \text{if } u_1 > u_0 \\ 0 & \text{if } u_1 \leq u_0 \end{cases}.$$

Only differences in utility matter for the choice behavior. We can normalize u_0 to zero and interpret u_1 as net utility from taking the action. We can parametrize net utility as $u_1 = u_\theta(x, e)$, where x can be observed and e are other factors that are not observed to the econometrician. A simple example has $u_\theta = x'\theta + e$ but this is not essential to the argument. With e and X independent and u_θ strictly monotonic (normalized increasing) in e we have

$$\mathbb{P}(Y = 1|X) = \mathbb{P}(u_\theta(X, e) > 0|X) = \mathbb{P}(e > \varphi_\theta(X)|X) = 1 - F(\varphi_\theta(X)),$$

where F is the cumulative distribution function of e, $\varphi_\theta(x) = u_\theta^{-1}(x, 0)$, and u_θ^{-1} the inverse function of u_θ with respect to its second argument. With linear utility $u_\theta = x'\theta + e$, $\varphi_\theta(x) = -x'\theta$, and so the success probability is $1 - F(-X'\theta)$. When F is symmetric about zero this further simplifies to $F(X'\theta)$. This was the case for the normal distribution; this gives raise to the probit model,

$$\mathbb{P}(Y = 1|X) = \Phi(X'\theta).$$

It is also the case for the logistic distribution, where

$$\mathbb{P}(Y = 1|X) = \Lambda(X'\theta), \qquad \Lambda(a) = \frac{1}{1 + e^{-a}}.$$

This is the logit model.

3.2. Ordered choice

In a first possible generalization of the binary-choice problem the choice set consists of m alternatives, $\{1, 2, \ldots, m\}$, and choice behavior is modelled as the threshold-crossing problem

$$Y = c \text{ if } \alpha_{c-1} < Y^* \leq \alpha_c,$$

where $\alpha_0 = -\infty$ and $\alpha_m = +\infty$. Here the choices are ordered, and the value that Y takes is determined by the thresholds between the latent Y^* falls. Proceeding with

$$Y^* = u_\theta(X, e)$$

as before we have

$$\mathbb{P}(Y = c|X) = \mathbb{P}(\alpha_{c-1} < Y^* \leq \alpha_c|X) = \mathbb{P}(u_\theta^{-1}(X, \alpha_{c-1}) < e \leq u_\theta^{-1}(X, \alpha_c)|X).$$

Hence, when $e \sim F$ independent of X,

$$\mathbb{P}(Y = c|X) = F(u_\theta^{-1}(X, \alpha_c)) - F(u_\theta^{-1}(X, \alpha_{c-1})).$$

With linear utility $u_\theta(x, e) = x'\theta + e$ and so we arrive at the conventional result that

$$\mathbb{P}(Y = c|X) = F(\alpha_c - X'\theta)) - F(\alpha_c - X'\theta).$$

Note that, here, it is not sensible to include a constant term in the set of regressors, as its effect cannot be separated from the thresholds. With normal errors or logistic errors this amounts to what is called the ordered-probit and ordered-logit model, respectively.

3.3. Unordered discrete choice

An alternative situation arises when outcomes to not have a natural ordering. In this case we think about Y_c^* as the utility derived from choosing alternative c. We have $Y = c$ when

$$Y_c^* \geq Y_{c'}^*, \qquad Y_c^* = X_c'\theta_c + e_c$$

for all c'. The choice probability is

$$\mathbb{P}(Y = c|X_1, \ldots, X_m) = \mathbb{P}(Y_c^* \geq Y_1^*, Y_c^* \geq Y_2^*, \ldots, Y_c^* \geq Y_m^*|X_1, \ldots, X_m)$$

and does not, in general, admit a closed-form expression. One case that gives nice formulas arises when $(e_1, e_2, \ldots, e_m)$ follows a generalized extreme value distribution, i.e.,

$$F(e_1, e_2, \ldots, e_m) = e^{\left(-\sum_{c=1}^m e^{-e_c}\right)}.$$

Then

$$\mathbb{P}(Y = c|X_1, \ldots, X_m) = \frac{e^{X_c'\theta_c}}{\sum_{c'=1}^{m} e^{X_{c'}'\theta_{c'}}}.$$

This is the multinomial-logit model.

One limitation of this model is that it suffers from the independence of irrelevant alternatives problem. Moreover, the relative choice probabilities between options c_1 and c_2,

$$\frac{\mathbb{P}(Y = c_1|X_1, \ldots, X_m)}{\mathbb{P}(Y = c_2|X_1, \ldots, X_m)} = \frac{e^{X_{c_1}'\theta_{c_1}}}{e^{X_{c_2}'\theta_{c_2}}} = e^{X_{c_1}'\theta_{c_1} - X_{c_2}'\theta_{c_2}}$$

do not depend on the other possible choices. This can lead to unrealistic substitution patterns. A classic example related to travel options, where the first option is to go by car and the second is to go by taking a blue bus. Suppose that each option is taken with probability $1/2$. Then the ratio of probabilities is equal to one. Now say a third option arises: to travel by red bus. It seems logical to consider the blue and red bus as equivalent, so that they are taken with the sample probability. But in that case the model predicts that substitution away from the blue bus and car towards to red bus are the same; that is, the model states that the probability of taking each of the three different modes of transportation is $1/3$. In this example, one would expect substitution to be away from the blue bus only, and that the introduction of the red bus leaves the probability of taking the car unaltered. Several generalisations exist to deal with this limitation of the multinomial logit model, such as the nested logit model.