Discrete Choice & Cross-Validation

Ethan Ligon

April 26, 2021

Living in the RW (Separable Models)

Last time we thought about a RWDGP represented as a triple $(\mathcal{M}, \mathcal{R}, \mathcal{F})$, with data observed by an agent $X^* = \mathcal{R}(s^t)$ determining their actions $y^* = \mathcal{M}(X^*)$.

- An econometrician typically won't observe all of X^*; instead, partition $X^* = (X, U)$, where we observe X and don't observe U.
- If $\mathcal{M}$ is *separable* in (X, U) then there's some transformation T such that $\mathrm{T}\mathcal{M}((X, U)) = f(X) + g(U)$. In such cases we can then write $\mathrm{T}y^* = y = f(X) + u$.

Discrete Choice

We've mostly considered models in which y is continuous, yet the data we have on choices is often discrete (took a job, bought a house, went to college).

- In many ways better to think of choices y^* being continuous—it's often super-helpful to think of agents choosing *probabilities* of discrete outcomes, with nature rolling a die to map the continuous choice into a discrete outcome.
- Probabilities live in $[0, 1]$; useful to convert to *odds*, which live in $\mathbb{R}_+$, or log odds, which live in $\mathbb{R}$.

Binary Outcomes

Suppose we have data on a *binary* outcome. No loss of generality in coding $Y \in \{0,1\}$ (e.g., dummy variable for "bought a house"). Let $\Pr(y=1|X) = \pi(X) = \frac{e^{f(X)}}{1+e^{f(X)}}$; then likelihood

$$L_N(f|Y,X) = \prod_{j=1}^{N} \pi(X_j)_j^{y}(1-\pi(X_j))^{1-y_j}$$

While the log-likelihood is

$$logL_N(f|Y,X) = \sum_{j=1}^{N} y_j \log \pi(X_j) + (1-y_j)\log(1-\pi(X_j)).$$

If, e.g., $f(X) = X\beta$ then we have *logit* model, with score

$$\sum_{j=1}^{N} X_j(y_j - \pi(X_j)) = 0.$$

Too Many Parameters

Consider a parametric model.
Loosely speaking, the more parameters we introduce into a model the better we can fit a given *sample*. In the limit if we have a sample of size N then with N parameters we can fit it exactly; i.e.,

$$y_j = f(X_j, \beta) + u_j \qquad j = 1, \ldots, N.$$

Setting $u_j = 0$, this gives us N equations with N parameters.

Bias vs. Variance

Introducing a large number of parameters reduces *variance* in sample, but increases *bias*.

Out of Sample Prediction

Our focus now is on our ability to *predict* outcomes y given data X.

Estimation

Suppose we have an iid sample $d_N = ((y_1, X_1), \ldots, (y_N, X_N))$.

- We want to use these these data to estimate a model $\hat{y}(X)$ such that $\mathbb{E}(y - \hat{y}(X)|X) = 0$.
- The *prediction error* of the model is $e = y - \hat{y}(X)$ for any (y, X).
- We want to estimate the model such that, say, $\mathbb{E}(e^2|X)$ is minimized, even when (or especially when) (y, X) are realizations that aren't in the sample used to estimate $\hat{y}(X)$. Call this the *mean squared out-of-sample prediction error*.

Leave-one-out estimators

Whatever estimator we have of $\hat{y}(X)$, we presume that can be estimated using observations $j = 1, \ldots, N$.

Estimator

An old idea for improving out-of-sample predictive power is the "leave-one-out" estimator. In this case for each observation j we construct a prediction function using every observation *except* j; i.e.,

$$\hat{y}_{-j}(X) = f(y_{-j}, X_{-j}) \qquad j = 1, \ldots, N.$$

Then the usual leave-one-out estimator is

$$\hat{y}(X) = \frac{1}{N} \sum_{j=1}^{N} \hat{y}_{-j}(X).$$

Note that *every single* $\hat{y}_{-j}(X_j)$ is an out-of-sample prediction, since (y_j, X_j) weren't used to construct $\hat{y}_{-j}$.

Cross-validation criterion

Let $e_{-j} = y_j - \hat{y}_{-j}(X_j)$. Call this the "leave-one-out error".

Criterion

Define the *cross-validation* criterion as

$$\mathrm{CV} = \frac{1}{N}\sum_{j=1}^{N} e_{-j}^2.$$

Given the iid sampling assumption it's easy to see that this is an unbiased estimator of the expected squared out-of-sample prediction error.

General Approach

If we have multiple possible models or estimators of the relationship between y and X, say $(\hat{y}^1(X), \ldots, \hat{y}^q(X))$ we *select* a model by choosing the one that minimizes the CV criterion.

Parameter Estimation

Rather than selecting among a finite number of discrete "models", we can also use the CV criterion as the basis for selecting a vector of continuous parameters.

K-fold Cross-Validation

A practical problem is that as N grows large our proposal to choose models by minimizing CV become impractical; typically the computational cost of implementing the estimator becomes $O(N^2)$ or worse. Enter K-fold cross-validation.

1. Divide sample of N observations into K different "folds" $d_{(k)}$, each of roughly size N/K.
2. Estimate a "leave N/K out" estimator that uses data from $N-1$ folds to estimate $\hat{y}^{(-k)}(X)$.
3. Let
$$e_j^{(k)} = y_j - \hat{y}^{(-k)}(X_j) \qquad \text{for } j \in d_{(k)}.$$
4. The criteria
$$\mathrm{CV}^{(k)} = \sum_{k=1}^{K} \sum_{j \in d_{(k)}} (e_j^{(k)})^2$$
then approximates the CV criterion. (If $K = N$ the two are identical).