14.771 Development Economics: Microeconomic Issues and Policy Models
Fall 2008

14.771: Recitation Handout #1
Nonparametric Regression and Measurement Error

We will cover today two topics. First, I will do a brief overview of non-parametric regressions. Second, I will discuss measurement error, it's problems for empirical work, and what we can do about it.

Nonparametric Regression

You will see nonparametric estimation come up a lot in some of the health and nutrition papers. For example Deaton and Subramanian (1996) use nonparametric regression to estimate Engel curves and the elasticity of calorie consumption with respect to income (this is a great paper, by the way - you should read it). Why not just use OLS for such an exercise? There are many reasons why one may prefer a non-parametric model. First, it may be that we do not want to impose a particular form of relationship between two variables. It may be that we are more interested in describing the relationship than producing causal estimates of y on x. Finally, in some applications, like regression discontinuity (we will cover it next week), it is essential for the identification assumption to have as flexible of a relationship between the two variables as possible.

Assume one has a relationship between a variable Y and X that they would like to estimate non-parametrically, ie

$$E\left(Y|X\right)=g_0\left(X\right)$$

There are, in general, three different approaches to non-parametric regressions:

- series regressions
- kernel regressions
- local linear regressions

Series regressions

Series regressions are the easiest to implement in terms of computing and standard errors. They are simply OLS regressions where one uses flexible functions of x to approximate the nonlinear relationship. The most common types of approximating functions used are power series and splines. Power series are good for approximating smooth functions - all you do is add polynomials of x on the right hand side of an OLS regression. However, when functions jump or have kinks, the approximation can be pretty bad - in this case, a better option is to use a spline. Denote the vector of K approximating functions as $p^K(x) = [p_{1K}(x), ...p_{KK}(x)]'$. Then when x is a scalar, all we have is:

$$\begin{aligned} \text{Power series} &: \; p_{jK}(x) = x^{j-1},\; (j = 1, 2, ..., K) \\ \text{Spline} &: \; p_{jK}(x) = x^{j-1},\; (j = 1, 2, 3, 4) \\ &\quad\; p_{jK}(x) = 1(x > l_{j-4,K})(x - l_{j-4,K})^3,\; (j = 5, ...K) \end{aligned}$$

Where $l_1, ...l_{K-4}$ are "knots" in the spline. You have to decide where to place the knots. In general, people tend to evenly space them.

Series are biased because it would require an infinite number of polynomial functions to achieve unbiasedness. They also provide a "global" fit - remember that OLS just minimizes mean squared error - this is in contrast to kernel regression and local linear (Fan) regression, which provides a local fit. Given this, how do we choose the number of terms? In practice, many people just add terms until the fit "looks" good (think RD papers). However, we have theory that can guide us - it is asymptotially optimal to choose the number of terms to minimize the cross validation critera - $\widehat{CV}(K)$ - in the sense that

$$\frac{\min_K MSE(K)}{MSE\left(\hat{K}\right)} \xrightarrow{p} 1$$

So how do we compute this thing? Denote

$$\hat{g}_{-i,K}(X_i) = Y_i - \frac{Y_i - \hat{g}_K(X_i)}{1 - Q_{ii}}$$

where Q_{ii} is the iith element of the idempotent projection matrix: $Q = P(P'P)^{-1}P'$ and P is just the matrix of all the approximating terms of X : $P = \left[p^K(X_1), ..., p^K(X_n)\right]$. Note that $\hat{g}_{-i}(x_i)$ is just the predicted value for the i^{th} observation using all other observations. Using the formula above, we don't have to run n different regressions to get this, which saves a lot of time. Then

$$\widehat{CV}(K) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{g}_{-i,K}(X_i)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{g}_{-i}(x_i)}{1 - Q_{ii}}\right)^2$$

To pick an optimal number of terms, you'd compute this statistic for a bunch of different K and choose the K that gives the smallest cross validation statistic. Standard errors are easier can be simply taken from the OLS regressions. It is also very easy to get the derivative of the function at any given point once you have the coefficient on all of the approximating terms.

Kernel regressions

Sometimes series regressions are ill-suited to a purpose (for example, a function is very wiggly or you want a more local fit). The first (and most intuitive) approach one can take is to take the mean of the outcome variables for a series of points over the domain of x. It is easy to compute and available in all statistical packages including stata.

Formally, this implies:

$$\begin{aligned}\hat{g}_h(x) &= \arg\min_g \sum_{i=1}^{n} (Y_i - g)^2 K_h(x - x_i) \\ \hat{g}_h(x) &= \frac{\sum Y_i K\left(\frac{x-X_i}{h}\right)}{\sum K\left(\frac{x-X_i}{h}\right)}\end{aligned}$$

where K is a kernel function. Despite its ugly looking character, this is simply a weighted mean. Rewrite this as

$$\hat{g}_h(x) = \sum w(X_i) Y_i \;\; where \; w(X_i) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum K\left(\frac{x-X_i}{h}\right)}$$

K is simply a function that weights observations depending on how far they are of the value of x of interest, here little x. If the kernel is symmetric, it will give more weight to observations that are very closed to the value of interest. Can you think of weighting functions?

h is the bandwith, ie, how far away are you allowing observations to matter for estimating the mean at x. What happens when the bandwith is larger? smaller?

Kernel regressions are biased. This is because one can never restrict oneself to small enough of a region around x to arrive to unbiasedness. Furthermore, the smaller the bandwith, the larger the variance. How do we pick the "best" bandwith? Practically, many people just try a few and pick the one they like the most. However, there is a theory that provides the "optimal" bandwith. We have already said that kernel regressions are biased. We can compute the mean square error as

$$MSE = Bias^2 + Variance$$

The optimal bandwith is the one that minimizes this value. It is not that easy to compute the MSE however, so there are "easier" solutions. Silverman's rule of thumb provides a very crude approximation if the kernel is the standard normal and the real distribution is also normal. Then, the estimate is given by:

$$h = 1.06 n^{\frac{1}{5}} \sqrt{Var(x)}$$

Again, the way to get at optimal bandwith is cross-validation. The formula is the same as it is for series estimation, but we cannot make use of the projection matrix trick to get the leave one out estimator without rerunning a bunch of regressions:

$$\widehat{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \hat{g}_{-i,h}(X_i)\right]^2$$

One repeats the kernel regression with various bandwiths, compute the CV and finds the bandwith that minimizes it.

What about standard errors? One in general bootstraps the standard errors in this case.

Local linear regressions

Kernel regression can be very biased at boundaries. For this reason, it isn't used in practice very much anymore - but we teach it (and learn it) because it has a nice intuitive interpretation and it's closely related to a much better technique - local linear regression. In local linear regression we basically use linear regression to take out the "linear" part of the bias in kernel regression. Incidentally, we can also do local polynomial regression which takes out higher order terms of the bias, but the improvement is small in practice, so you'll mostly see local linear in applications (but STATA has a command "lpoly" that will let you do many different degrees). So, how do we do local linear regression? As its name suggests, instead of taking a weighted mean around each data point, we take the predicted value from a weighted least squares regression around each data point. And the weights? Kernel weights, of course! So this means that points closest to the point of interest (x) will be weighted most, and points very far away will receive little weight in the regression. Formally:

$$\hat{g}_h(x) = \arg\min_{g,\beta} \sum_{i=1}^{n} \left(Y_i - g - (x - X_i)' \beta\right)^2 K_h(x - x_i)$$

$$\hat{g}_h(x) = \frac{\sum (x - x_i) K\left(\frac{x-x_i}{h}\right) y_i}{\sum (x - x_i)^2 K\left(\frac{x-x_i}{h}\right)}$$

Less formally, this is what you would do: First, datasets often have lots of observations, so it's impractical to run a regression for each point. Instead, people will often "grid" up the domain - i.e. if x runs from 0 to 1000, run a regression for $x = 5j$, $j = 1 - 190$. For each x, create the kernel weight for all observations: $w_i(x) = K\left(\frac{x-x_i}{h}\right)$. Run weighted OLS of Y on X and take the predicted value at x. Save all the predicted values, and plot them - now you've done local linear regression.

This produces smaller bias than the kernel and reduces MSE. In addition, local linear also makes it easy to get derivatives of the nonlinear function (Deaton and Subramanian use this a lot) - all we need to do is look at the regression coefficient on x at each point!

$$\hat{\beta}(x) = \widehat{\frac{\partial g_0(x)}{\partial x}}$$

There is also a cross validation statistic to help you pick optimal bandwidth - it's given by:

$$CV(h) = \frac{\sum_{i=1}^{n} \left[Y_i - \hat{g}_{-i,h}(x_i)\right]^2}{n}$$

The Curse of Dimensionality

All of these methods are relatively easy to implement when x is a scalar - that is, you want to model a flexible two-dimensional function. What if you want two independent variables - or three, or four? The answer is that the convergence rate of your estimator plummets at a disturbingly high rate, and you'll need tons of data to estimate your function precisely. The curse of dimensionality affects all types of nonparametric estimation - if you want to allow for a fully flexible relationship between the dependent and independent variables, there is just no way around it. Can you think of ways to add some structure but still allow for a flexible functional form on some dimensions?

Measurement Error

In applied work, we generally use survey data (the alternative is administrative data - social security earnings records, health insurance claims, etc). There are many reasons why the variables we see in surveys are measured with error:

1. Respndents don't know the exact answer and estimate, or the misreport on purpose
2. Respondents don't understand the question
3. Gap between what people say and what they would do in the real world (contingent valuation surveys)
4. The survey does not measure the concept we're really interested in (i.e. we use schooling to proxy for human capital)

Let's do a simple univariate example. Suppose the "true" model is

$$Y_i^* = \alpha + X_i^* \beta + \varepsilon_i$$

but instead we observe these things measured with error:

$$\begin{aligned} Y_i &= Y_i^* + \eta_i \\ X_i &= X_i^* + \nu_i \end{aligned}$$

So when we run a regression, what regression coefficient to we get. Asymptotically:

$$\begin{aligned} p\lim \hat{\beta} &= \frac{Cov\left(X_i, Y_i\right)}{Var\left(X_i\right)} = \frac{Cov\left(X_i, Y_i^* + \eta_i\right)}{Var\left(X_i\right)} \\ &= \frac{Cov\left(X_i, \alpha + X_i^* \beta + \varepsilon_i + \eta_i\right)}{Var\left(X_i\right)} \\ &= \frac{Cov\left(X_i, X_i^*\right)}{Var\left(X_i\right)} \beta + \frac{Cov\left(X_i, \varepsilon_i + \eta_i\right)}{Var\left(X_i\right)} \end{aligned}$$

We can already tell from looking at this that there's no reason to think that this is going to be equal to β, which is what we'd get if there were no measurement error. Without imposing some more structure on the problem, this is as far as we can get - we cannot say if β is biased up or down, or how large the bias is. This is unsatisfying if you're an economist - are there cases where we can sign the bias?

Classical Measurement Error

When people refer to measurement error in papers and in seminars, this is what they usually mean. Classical measurement error imposes that:

$$\begin{aligned} Cov\left(X_i, \nu_i\right) &= 0 \\ Cov\left(Y_i, \eta_i\right) &= 0 \\ Cov\left(\nu_i, \eta_i\right) &= 0 \end{aligned}$$

So what happens to our measurement error formula with these additional assumptions?

$$\begin{aligned} p\lim\hat{\beta} &= \frac{Cov\left(X_i, X_i^*\right)}{Var\left(X_i\right)}\beta + \underbrace{\frac{Cov\left(X_i, \varepsilon_i + \eta_i\right)}{Var\left(X_i\right)}}_{=0} \\ &= \frac{Cov\left(X_i^* + \nu_i, X_i^*\right)}{Var\left(X_i\right)}\beta \\ &= \frac{Var\left(X_i^*\right)}{Var\left(X_i\right)}\beta \end{aligned}$$

We often call $\frac{Var\left(X_i^*\right)}{Var\left(X_i\right)} = \lambda$ the reliability ratio - think of it as the share of signal (vs. noise) in the data that we observe. What can we say about the bias on $\hat{\beta}$ now? What about when there are additional covariates in the regression (say $W_i = W_i^* + \mu_i$)? Adding controls will exacerbate measurement error as long as they are correlated with X_i^*. (You can do a bunch of math to convince yourself of this, but I'll omit it here).

Measurement Error in Panel Data

Now suppose we have panel data and we observe:

$$\begin{aligned} Y_{it} &= Y_{it}^* + \eta_{it} \\ X_{it} &= X_{it}^* + \nu_{it} \end{aligned}$$

When using panel data, we often use first differences of fixed effects. What are the implications of measurement error for these models? Consider the first differences case where measurement error remains classical:

$$p\lim\hat{\beta} = \frac{Var\left(\Delta X_{it}^*\right)}{Var\left(\Delta X_{it}\right)}\beta = \frac{Var\left(\Delta X_{it}^*\right)}{Var\left(\Delta X_{it}^*\right) + Var\left(\Delta \nu_{it}\right)}\beta$$

If measurement error is uncorrelated over time then $Var\left(\Delta\nu_i\right) = 2Var\left(\nu_i\right)$. Similarly, note that

$$Var\left(\Delta X_{it}^*\right) = Var\left(X_{it} - X_{it-1}\right) - 2Cov\left(X_{it} - X_{it-1}\right)$$

so if the Xs are positively correlated over time (which is often true in practice) $Var\left(\Delta X_{it}^*\right) < Var\left(X_{it}^*\right)$. When this is the case, first differencing exacerbates measurement error. On the other hand, if the errors are more correlated over time than the Xs, then measurement error is less of a problem with first differences. In most cases, first differencing (or fixed effects) will make measurement error problems worse.

Division Bias

This is a particularly pernicious form of measurement error - it often occurs when the right hand size variable is constructed by dividing one variables by another that is measured with error (for example, contructing the hourly wage by dividing total earnings by hours worked). Suppose the true model is:

$$\ln h_i^* = \alpha + \beta \ln w_i^* + \varepsilon_i$$

where h^* is hours worked and w_i^* is the wage. But suppose we observe $\ln h_i = \ln h_i^* + \eta_i$, and we need to construct the wage: $\ln w_i = \ln y_i - \ln h_{i.}$ Then

$$\begin{aligned} p\lim \hat{\beta} &= \frac{Cov\left(\ln h_i^* + \eta_i, \ln w_i\right)}{Var\left(\ln w_i\right)} = \frac{Cov\left(\ln h_i^* + \eta_i, \ln y_i - \ln h_{i.}\right)}{Var\left(\ln w_i\right)} \\ &= \frac{Cov\left(\ln h_i^* + \eta_i, \ln y_i - \ln h_{i.}^* - \eta_i\right)}{Var\left(\ln w_i\right)} = \frac{Cov\left(\ln h_i^* + \eta_i, \ln w_{i.}^* - \eta_i\right)}{Var\left(\ln w_i\right)} \\ &= \frac{Cov\left(\beta \ln w_i^* + \varepsilon_i + \eta_i, \ln w_{i.}^* - \eta_i\right)}{Var\left(\ln w_i\right)} \\ &= \frac{Var\left(\ln w_i^*\right)}{Var\left(\ln w_i\right)}\beta - \frac{Var\left(\eta_i\right)}{Var\left(\ln w_i\right)} \end{aligned}$$

The second term is necessarily positive, so we see that division bias exacerbates the downward bias of classical measurement error.

What Can We Do?

So measurement error is clearly a problem - if you think it's nonclassical, we have no idea what it's doing our estimates, and if it's classical, it biases them to zero - which will make it harder to detect significant effects. If you have the means to assess to reliability ratio and are willing to assume classical measurement erro, you can adjust you estimates. If you have panel data with $T > 2$ you can try the approach outlined in Grilliches and Hausman (1984). However, the most common solution to the problem is instrumental variables. Returning to our original notation, suppose you have a second, independent measure of your X variable, also measured with error:

$$Z_i = X_i^* + \xi_i$$

Where $Cov\left(\xi_i, \nu_i\right) = 0$ and $Cov\left(\xi_i, \eta_i\right) = 0$. Then note that

$$p\lim \hat{\beta}_{IV} = \frac{Cov\left(Z_i, Y_i\right)}{Cov\left(Z_i, X_i\right)} = \frac{Cov\left(X_i^*, Y_i\right)}{Cov\left(X_i^*, X_i\right)} = \frac{Var\left(X_i^*\right)}{Var\left(X_i^*\right)}\beta$$

The use of IV to solve the measurement error problem is probably the most common solution that you'll see in practice.