Mostly Harmless Econometrics: Notes

Matt Cocci

August 5, 2013

1 Regression Fundamentals

Throughout these notes, I will make use of the fact that Y_i is the dependent variable of interest, and X_i is a $K \times 1$ vector of covariates with entries x_{ki}.

Everything will be indexed by i, which isn't crucial when we discuss population values. In fact, we won't even use the index, but including it early will facilitate discussion of sample values and distributions later on. So we will start by discussing only X_i and Y_i, each of which has a population distribution. But when we start discussing sample values, we want to sum up many such variables, where the indexing will come in handy.

1.1 The Conditional Expectation Function (CEF)

Conditional Expectation We define the expectation of Y_i, conditional on the covariates $X_i = x$ as

$$E[Y_i \mid X_i = x] = \int y f_y(y \mid X_i = x)\, dy$$

$$\text{where} \quad f_y(y \mid X_i = x) = \frac{f_{x,y}(x, y)}{f_x(x)}$$

in the continuous case. Or, alternatively, in the discrete case,

$$\sum_t P(Y_i = t \mid X_i = x)$$

Viewing the CEF as a function of the random variable X_i, we see that the CEF itself is a random variable.

Law of Iterated Expectations $E[Y_i] = E\left\{E[Y_i \mid X_i]\right\}$

CEF Decomposition Property (3.1.1) Where the "CEF" stands for the conditional expectation function:

$$Y_i = E[Y_i \mid X_i] + \varepsilon_i$$
$$E[\varepsilon_i \mid X_i] = 0, \qquad E[\varepsilon_i X_i] = 0$$

In words, this says that the random variable Y_i can be broken down into it's conditional expectation along with some random component—the error term—that is mean independent of X_i (with expectation 0) and is uncorrelated with any function of X_i.

CEF Prediction Property (3.1.2) Where $m(X_i)$ is some function of X_i, the CEF solves the following expression

$$E[Y_i \mid X_i] = \arg\min_{m(X_i)} E\left[(Y_i - m(X_i))^2\right]$$

So we see that the CEF is the *minimum mean square error* predictor of Y_i conditional on X_i, among the class of all functions of X_i.

The ANOVA Theorem (3.1.3) Where we take $V(\cdot)$ to be the variance operator, we have

$$V(Y_i) = V\left(E[Y_i \mid X_i]\right) + E\left[V(Y_i \mid X_i)\right]$$

1.2 Regression and the CEF

Definition Where β represents the $K \times 1$ **Regression Coefficient Vector**, β solves

$$\beta = \arg\min_b E\left[(Y_i - X_i'b)^2\right]$$

which, upon using the first order condition—which is $E[X_i(Y_i - X_i'b)] = 0$—reduces to

$$\beta = E\left[X_i X_i'\right]^{-1} E\left[X_i Y_i\right].$$

Note by the construction of β the first, first order condition ensures that X_i will be uncorrelated with $(Y_i - X_i'\beta)$, which is just the definition of ε_i.

Bivariate Case, Single Regressor With the single regressor, x_i, the regression coefficient vector contains one term:

$$\beta_1 = \frac{Cov(Y_i, x_i)}{V(x_i)}$$

Multivariate Case, k Regressors Where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all the other covariates,

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})},$$

so that the kth element of the $K \times 1$ vector $E\left[X_i X_i'\right]^{-1} E\left[X_i Y_i\right]$ will be β_k as just defined.

This means that each coefficient in a multivariate regression is simply the coefficient in a *bivariate* regression of Y_i on the regressor of interest, after partialing out all the other covariates. (Alternatively, you can substitute $\tilde{Y}_{ki}$ for Y_i, where $\tilde{Y}_{ki}$ is the residual from a regression of Y_i on every covariate except x_{ki}.)

Linear CEF Theorem (3.1.4) (Regression Justification I) If the CEF is linear, then the population regression function is it.

- Note that the CEF will be linear if the vector $(Y_i, X_i')'$ has a multivariate normal distribution, which makes this justification pretty useful, since many variables are multivariate normal.

Best Linear Predictor Theorem (3.1.5) (Regression Justification II) The function $X_i'\beta$ is the best linear predictor of Y_i given X_i in a MMSE sense.

Regression CEF Theorem (3.1.6) (Regression Justification III) The function $X_i'\beta$ provides the best MMSE linear approximation to $E[Y_i|X_i]$. That is,

$$\beta = \arg\min_b \; E\left\{(E[Y_i|X_i] - X_i'b)^2\right\}$$

- Note that one important implication of this theorem is that β can be obtained by using the $E[Y_i|X_i]$ as the dependent variable instead of Y_i. Specifically, suppose X_i is discrete with mass function $g_x(u)$; then apply the theorem to get

$$E\left\{(E[Y_i|X_i] - X_i'b)^2\right\} = \sum_u (E[Y_i|X_i = u] - u'b)^2 \; g_x(u).$$

  This implies that β can be constructed from a *weighted least squares* (WLS) regression of $E[Y_i|X_i = u]$ on u, where the weights applied to the coefficients are $g_x(u)$ and u runs over the values of X_i.

  This approach, the **grouped data** version of the regression formula, turns out to be very useful when microdata (all the Y_i's and X_i's) are not availabe, and when you only have access to conditional averages, the $E[Y_i|X_i = u]$, such as wages conditional on race, sex, test scores, etc. Then you can estimate β by the equation above, adding dummy variables and such (although the standard errors will be larger.)

1.3 Elaboration and Motivation

Putting together everything that we've see, we cna assert that the CEF is the best *unrestricted* predictor of the dependent variable, Y_i. That we saw in 3.1.2.

Next, if we can't find the CEF, we can use regression, which furnishes the best *linear* predictor of the dependent variable, Y_i, by 3.1.5.

Finally, if we want to loosen up further, we can avoid trying to predict Y_i altogether. Instead, we can think about estimating the CEF itself—in effect, taking $E[X_i|Y_i]$ as the dependent variable and fitting a line to it. In that case, 3.1.6 tells us that regression offers the best linear predictor of the CEF (even if that CEF is non-linear). Now if it just so happens that we're in the special case where the CEF is linear, then regression works as desired, following Theorem 3.1.4.

1.4 Traditional Econometric Inference

This discussion has been somewhat less restrictive than the traditional econometric setup—sometimes called a *classical normal regression model* which postulates

- Fixed, non-stocahstic regressors.
- A linear CEF.
- Normally distributed errors.
- Homoskedasticity.

These stronger assumptions give us

1. Unbiasedness of the OLS estimator: $E[\hat{\beta}] = \beta$.
2. A formula for the sampling variance of the OLS estimator that's valid in small samples as well as large.

2 Asymptotic OLS Inference

In this section, we'll divorce ourselves from the question of what regression coefficients *mean*, and we'll instead consider their distributions and other purely statistical properties that are easily described and meaning-agnostic.

2.1 General Asymptotic Distribution Theory

Law of Large Numbers Sample moments converge in probability to population moments. That is, the sample moment can be made as close to a population moment as desired by taking a large enough sample.

Central Limit Theorem Sample moments are asymptotically normally distributed.

Slutsky's Theorem

1. Consider two random variables, A and B. Suppose that A converges in distribution—the statistic a_N has an asymptotic distribution—and that B converges in probability—the statistic b_N has a probability limit b.

   Then the the sum $a_N + b_N$ has the same asymptotic distribution as $a_N + b$.

2. Take A, B, a_N, b_N, and b to be the same as above.

   Then the product $a_N b_N$ has the same asymptotic distribution of $a_N b$.

Continuous Mapping Theorem Probability limits pass through continuous functions. Formally, plim of $h(b_N)$ equals $h(b)$ provided that plim of b_N equals b, where $h(\cdot)$ is continuous at b.

Delta Method Page 44.

- A **Quadratic Form** is a matrix-weighted sum of squares. To be more specific, take v as an $\text{N}\times 1$ vector and M as a diagnonal $\text{N}\times\text{N}$ matrix. Then a quadratic form in v is $v'Mv$.

  In the special case that M is diagonal with elements m_i, $i = 1, \ldots, \text{N}$, then

$$v'Mv = \sum_i m_i v^2.$$

2.2 Ordinary Least Squares Estimator

OLS Estimator First, we use summation notation, where X_i and Y_i are i.i.d with sample size N, to define the *OLS Estimator*:

$$\hat{\beta} = \left[\sum_{i=1}^{N} X_i X_i'\right]^{-1} \sum_{i=1}^{N} X_i Y_i$$

Or, in matrix notation, where X is a matrix with row entries X_i' and Y is a vector with elements Y_i, both with $i = 1, \ldots, N$:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Covariance Matrix As a consequence, $\hat{\beta}$ has an asymptotic normal distribution with proability limit β and *covariance matrix*

$$E[X_iX_i']^{-1}E[X_iX_i'e_i^2]E[X_iX_i']^{-1}. \tag{1}$$

Standard Errors When we want to construct t-tests and confidence intervals, we'll need to consider standard errors. Here we have a choice, depending on how many assumptions we want to impose on our data:

1. *Heteroskedasticity-Consistent, White, or "Robust" Standard Errors*: These standard errors are the square roots of the diagonal elements of Expression 1 above. In pracice they are computed by calculating the

$$\sum[X_iX_i\hat{e}_i^2]/N, \qquad \hat{e}_i = Y_i - X_i'\hat{\beta}.$$

2. *"Default" Standard Errors*: These are derived under the homoskedasticity assumption—that $E[e_i^2|X_i] = \sigma^2$. With this assumption, Expression 1 simplifies to

$$\sigma^2 E[X_iX_i']^{-1}.$$

   This convention is reported by most statistical software unless one explicity requests otherwise.

3 Saturated Models

Saturated regression models are regression models with discrete explanatory variables, where the model includes a separate parameter for all possible values taken on by the explanatory variables. In effect, we add many dummy variabls that can be switched "on" or "off" depending on whether an explanatory variable takes on a specific value.

Fit Saturated models fit the CEF *perfectly*—regardless of the distribution of Y_i—becuase the CEF is a linear function of the dummy regressors used to saturate. For this reason, it is standard to begin with a saturated model.

Interaction Terms In a saturated model with dummy regressors, a saturated model will include *interaction terms* which are defined as the product of the regressors.

4 Regression and Causality