MRU Mastering Econometrics
Spring 2020

M. Joshway

Regression Theory

1 Regression and the CEF

1.1 Defining Regression

Regression is a many-splendored thing. I like to define the regression of Y_i on a vector of covariates, X_i, as the *best linear predictor* (BLP) of Y_i given X_i. This regression origin story gives no quarter to the question of why we're running it.

To appreciate the power of the BLP framework, it's enough to consider a single regressor, X_{1i}. The regression slope and intercept can be defined to be the values of a and b that minimize mean (expected) squared prediction error of Y_i as a linear function of X_{1i}. This minimization problem can be written:

$$MSE_{Y|X_1}(a, b) = E(Y_i - a - bX_{1i})^2$$

The solution is the regression slope and intercept, β and α:

$$\begin{aligned} b = \beta &\equiv \frac{E[(Y_i - E(Y_i))X_{1i}]}{E[(X_{1i} - E(X_{1i}))X_{1i}]} = \frac{C(X_{1i}, Y_i)}{V(X_i)} \\ a = \alpha &\equiv E[Y_i] - E[X_{1i}]\beta \end{aligned}$$

Regression is the BLP for *any* Y_i given any X_{1i}. It's up to you to make your prediction problems interesting and relevant.

1.2 Regression and Causality

Regression is causal when the corresponding conditional expectation function (CEF) is causal. If, for example Y_i is fall grades and D_i is a treatment dummy indicating students receiving randomized GPA incentives, then $E[Y_i|D_i, W_i]$ has a causal interpretation, revealing differences in average potential GPAs indexed by D_i, conditional on control variables, W_i.

- The *regression* of Y_i on D_i and W_i inherits this CEFs causal interpretation
- In these notes, we're unconcerned with causality. Here, briefly, we're merely regression mechanics

1.3 Reasons to Love

The BLP property one of regression's key features. My love for regression is nourished by the following theorems as well:

The Linear CEF Theorem. Suppose that

$$E[Y_i \mid X_{1i}] = a + bX_{1i}, \tag{1}$$

for some constants, a and b. Then:

$$b = \beta \equiv \frac{C(X_{1i}, Y_i)}{V(X_i)} \tag{2}$$

$$a = \alpha \equiv E[Y_i] - E[X_{1i}]\beta \tag{3}$$

Proof. By properties of the CEF

$$E(Y_i - E[Y_i|X_{1i}]) = 0 \tag{4}$$
$$E((Y_i - E[Y_i|X_{1i}])X_{1i}) = 0 \tag{5}$$

Assuming $E[Y_i \mid X_i] = a + bX_i$ we have

$$E(Y_i - a - bX_{1i}) = 0 \qquad or \qquad a = \alpha = E[Y_i] - E[X_{1i}]b \tag{6}$$
$$E((Y_i - a - bX_{1i})X_{1i}) = 0 \qquad or \qquad E[(Y_i - E(Y_i))X_{1i} - b(X_{1i} - E(X_{1i}))X_{1i}] = 0 \tag{7}$$

So, b=$\beta \equiv \frac{E[(Y_i - E(Y_i))X_{1i}]}{E[(X_i - E(X_i))X_{1i}]} = \frac{C(X_{1i}, Y_i)}{V(X_{1i})}$

My favorite regression feature is this:

The Linear Approximation Theorem. α and β are the values of a and b that minimize:

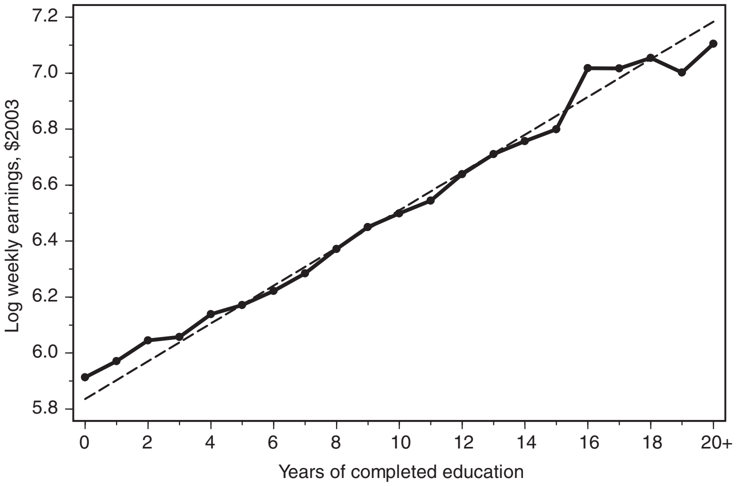
$$MSE_{E[Y|X]}(a, b) = E\left(E[Y_i \mid X_{1i}] - a - bX_{1i}\right)^2$$

Proof. The first order conditions for this minimization problem are

$$\partial MSE/\partial a = E(E[Y_i \mid X_{1i}] - a - bX_{1i}) = 0 \tag{8}$$
$$\partial MSE/\partial b = E((E[Y_i \mid X_{1i}] - a - bX_{1i})X_{1i}) = 0 \tag{9}$$

Iterating expectations (6) and (7) over X_i shows them to imply (8) and (9).

- The pic below shows how regression threads the CEF



- The regression slope and intercept provide the least square fit to $E[Y_i|X_i]$ as well as to Y_i

2 Regression for Dummies

When the conditioning variable is a dummy, say D_i, then $E[Y_i \mid D_i]$ is indeed linear:

$$E[Y_i \mid D_i] = E[Y_i \mid D_i = 0] + (E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0])D_i$$

By the Linear CEF theorem, therefore, the regression slope and intercept satisfy

$$\begin{aligned}\alpha =& E[Y_i|D_i = 0] = E[Y_i] - E[D_i]\beta \\ \beta =& E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = C(D_i, Y_i)/V(D_i)\end{aligned}$$

(Show this at home)

- *Regressions estimate differences in means.* We saw this use of regression in analysis of data from experiments like ALO (2009) and classroom electronics. Sometimes we add control variables to this regression (like high school GPA), making it a *multivariate regression.*
- When regressors are discrete and our model includes a dummy for all possible values they might take, the model is said to be *saturated.* The Linear CEF theorem applies to all saturated models.
- *Multivariate regression* is also an *automatic matchmaker,* that is, a simple strategy to make *ceteris paribus* comparisons involving the focal regressor, with control regressors held fixed. Consider coefficient δ in the regression of Y_i on D_i and a vector of saturated dummy controls, W_i:

  $$Y_i = W_i'\gamma + \delta D_i + \varepsilon_i,$$

  where ε_i is the regression error. Then

  $$\delta = E[\delta(W_i)\sigma_D^2(W_i)],$$

  where $\delta(W_i) = E[Y_i|D_i = 1, W_i] - E[Y_i|D_i = 0, W_i]$ and $\sigma_D^2(W_i)$ is conditional variance function for D_i given W_i. The proof of this uses the theorems given in Section 1, above. The ideas behind this important result are sketched in MM and detailed in MHE.

3 Ordinary Least Squares

We *estimate* α and β with sample analogs:

$$\begin{aligned}\hat{\alpha} &= \bar{Y} - \bar{X}_1\hat{\beta} \\ \hat{\beta} &= s_{X_1Y}/s_{X_1}^2\end{aligned}$$

These *Ordinary Least Squares* (OLS) estimators of α and β seem natural and indeed have good statistical properties.

- Traditional 'metrics texts derive OLS as the solution to a sample least squares problem:
  - Given observations on a pair of random variables: $\{(Y_i,\, X_{1i});\, i = 1, \ldots, n\}$, you'd like to "model" Y_i as a linear function of X_{1i}
  - How should you pick the slope and intercept? Minimizing the sample sum of squared errors,

    $$\widehat{MSE}_{Y|X}(a,\, b) = \sum_i (Y_i - a - bX_{1i})^2,$$

    generates $\hat{\alpha}$ and $\hat{\beta}$, above
  - Prove this at home

4 Make Mine Multivariate

- We're rarely interested in models with a single regressor. Rather, most 'metrics masters seek the regression of Y_i on a vector of k explanatory variables, X_i. The multivariate regression slope vector is then defined as the minimizer of

  $$MSE_{Y|X}(b) = E(Y_i - b'X_i)^2,$$

  where b is now a $k \times 1$ vector of coefficients

- The solution for this is

  $$b = \beta \equiv E[X_i X_i']^{-1} E[X_i Y_i] \tag{10}$$

  (the constant in this model is the coefficient on a regressor that equals 1 for every i)

- As many of you will know, the multivariate OLS estimator is the sample analog of this β, an estimator that can be written:

  $$\hat{\beta} = [X'X]^{-1} X'Y, \tag{11}$$

  where, in a sample of size n, Y is the $n \times 1$ column vector formed by stacking the Y_i and X is the $n \times k$ matrix of regressors with rows X_i'

- In my view, formulas (10) and (11) add little to our *understanding* of regression, though they're surely of use to computer programmers tasked with computing regression estimates
  - Regression anatomy and the variance-weighting interpretation of OLS estimates discussed in MHE tell us how, exactly, regression generates controlled comparisons
  - That's it for the theory

Fear no more!