Lecture Notes on

Regression

Prof. Kyle Butts

Conditional Expectation Function

The **conditional expectation function (CEF)** is one of the most fundamental concepts in applied econometrics. At its core, it asks a simple question: given that we know some characteristics about a unit, what do we expect their outcome to be?

Formally, we define the CEF as:

$$m(\boldsymbol{x}) \equiv \mathbb{E}[y_i|\boldsymbol{X}_i = \boldsymbol{x}]$$

This reads as "$m(\boldsymbol{x})$ is the expected value of y_i conditional on the unit having characteristics $\boldsymbol{X}_i = \boldsymbol{x}$." The beauty of this definition is that it connects directly to how we naturally think about prediction and relationships in data.

The most straightforward way to estimate the CEF for any given value $\boldsymbol{x}$ is remarkably simple: find all units in your data with $\boldsymbol{X}_i = \boldsymbol{x}$ and take their average outcome. This is the simplest estimation procedure for the CEF. For each possible value of $\boldsymbol{x}$,

1. Subset your data to units with $\boldsymbol{X}_i = \boldsymbol{x}$
2. Take average of y_i for those units. Call $m(\boldsymbol{x})$ this subsample average.

Let's start with the simplest case: a single discrete variable. Consider wages (w_i) and college attendance (D_i), where $D_i = 1$ indicates college attendance and $D_i = 0$ indicates no college. The CEF becomes:

$$m(1) = \mathbb{E}[w_i|D_i = 1] \quad \text{and} \quad m(0) = \mathbb{E}[w_i|D_i = 0]$$

This is exactly what we'd compute: the average wage for college attendees and non-attendees.

Now suppose we have two discrete variables: college attendance (D_i) and gender, where $F_i = 1$ indicates female. We now have four distinct combinations: $(D_i, F_i) \in \{(0,0), (1,0), (0,1), (1,1)\}$. Estimating the CEF requires computing four separate averages—one for each group. This highlights an important point: to fully characterize the relationship, we need to account for all possible interactions between our variables.

Moving to continuous variables introduces new challenges. If X_i is continuous, we can no longer find units with exactly $X_i = x$. Instead, we must find units with $X_i \approx x$. But how close is "close enough"? The standard approach is to create many small **bins** across the range of X and average outcomes within each bin. This non-parametric estimation approach works well when we have plenty of data and few variables.

However, as we add more variables, we quickly encounter the **curse of dimensionality**. With multiple continuous variables, the number of distinct combinations of $\boldsymbol{x}$ values grows exponentially. In practice, most combinations will have few or no observations, making reliable estimation impossible. This fundamental limitation pushes us toward modeling the CEF rather than estimating it directly.

Modelling the CEF

The curse of dimensionality forces us to make assumptions about the functional form of $m(\boldsymbol{x})$. Rather than estimating the CEF separately for every possible combination of characteristics, we must specify a model that can leverage all our data efficiently.

The most common approach is the **linear**

model:

$$m(\boldsymbol{x}) = \boldsymbol{x}'\beta = \sum_{k=1}^{p} x_k \beta_k$$

We assume that one of the $x_k = 1$ for all units, so that we have an intercept. This assumes that the expected outcome changes linearly with each variable x_k.

But linearity often fails to capture important features of real relationships. Consider a few examples where constant marginal effects seem implausible:

- Returns to experience in the labor market typically exhibit diminishing returns. The 20th year of experience adds less to wages than the second year.
- In housing markets, an additional square foot adds less value to a mansion than to a studio apartment.
- The return to experience may depend critically on whether someone already has a college degree. Here, there is complementarity between different types of human capital.

These examples suggest we need more flexibility while maintaining the tractability of linear models. The key insight is to distinguish between the variables we observe, $\boldsymbol{X}_i$, and the **terms** we include in our model. In general, we write

$$\boldsymbol{W}_i = \boldsymbol{g}(\boldsymbol{X}_i) = (g_1(\boldsymbol{X}_i), \ldots, g_K(\boldsymbol{X}_i))'.$$

For diminishing returns, we might include polynomial terms: if X_i is years of experience, we could include both X_i and X_i^2 in $\boldsymbol{W}_i$. For complementarity, we include interaction terms: the return to experience might be $\beta_1 X_{1i} + \beta_2 X_{1i} \cdot D_i$ where D_i indicates college education. For highly flexible relationships, we can use bins: divide a continuous variable into ranges and include indicator variables for each range.

The selection of terms to include in a linear regression model is a bit of an art. Including too many terms makes the model at risk of overfitting and makes the output more difficult to interpret. Too few terms risks making strong restrictions on how $\boldsymbol{X}_i$ and y_i relate.

One very common setting is wanting to consider a highly flexible relationship of one key variable ($X_{1,i}$) while maintaining a simple linear control for other variables ($\boldsymbol{Z}_i$). This model,

$$y_i = \mu(X_{1,i}) + \boldsymbol{Z}_i'\beta + u_i,$$

is often called the **partially linear model** because the function we let $\mu(X_{1,i})$ be very flexible while having a simple linear model for the other covariates.

There are many ways to estimate this model, but one popular method is called Binscatter Regression and is implementable via the `binsreg` package.

Properties of OLS

Understanding what ordinary least squares actually estimates requires us to think carefully about what's included in our model versus what's left in the error term. This leads us to two fundamental insights that shape how we interpret regression coefficients in practice.

Omitted Variable Bias

Every regression model makes a choice about which variables to include. There are a bunch of variables we do not include, perhaps because we think they're not important or we do not have them in our dataset. The error term in our regression model contains all the other factors that affect our outcome. When these omitted variables are correlated with our variables of interest, we get biased coefficient estimates.

Consider the simple case where the true relation-

ship is:

$$y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + \varepsilon_i$$

But we only estimate:

$$y_i = \delta_0 + X_{1i}\delta_1 + \text{error}_i$$

The omitted variable bias formula tells us exactly what happens:

$$\hat{\delta}_1 = \beta_1 + \beta_2 \frac{\text{Cov}\left(X_1, X_2\right)}{\text{Var}\left(X_1\right)}$$

Our estimate equals the true causal effect β_1 plus a bias term. This bias has two components: the effect of the omitted variable on the outcome (β_2) and how well our included variable predicts the omitted variable ($\frac{\text{Cov}\left(X_1, X_2\right)}{\text{Var}\left(X_1\right)}$).

The intuition is straightforward: regression is fundamentally about prediction. If X_1 and X_2 are correlated, then knowing X_1 gives us information about X_2. Since we haven't included X_2 directly, the regression "uses" the correlation between X_1 and X_2 to capture some of X_2's effect on y. Our coefficient on X_1, therefore, picks up both its direct effect and its indirect effect through predicting X_2.

This framework helps us sign the bias in practical applications. If the omitted variable has a positive effect on the outcome ($\beta_2 > 0$) and is positively correlated with our variable of interest, the bias is positive. Our estimate will be too large, overstating the true causal effect.

Frisch-Waugh-Lovell Theorem

The Frisch-Waugh-Lovell (FWL) theorem provides a deeper understanding of what "controlling for" variables actually means. It shows that any multivariate regression coefficient can be obtained through a simple two-step procedure.

Consider the regression:

$$y_i = \tau D_i + \boldsymbol{W}_i'\boldsymbol{\beta} + u_i.$$

We are interested in the coefficient τ but want to 'control' for other factors $\boldsymbol{W}_i$.

The FWL theorem tells us that our estimate $\hat{\tau}$ can be obtained by:

1. Regress D_i on $\boldsymbol{W}_i$ and obtain the residuals $\tilde{D}_i$
2. Regress y_i on $\boldsymbol{W}_i$ and obtain the residuals $\tilde{y}_i$
3. Regress $\tilde{y}_i$ on $\tilde{D}_i$ to get $\hat{\tau}$

The first step removes the variation in D_i that can be predicted by the control variables. What remains is the "unpredictable" variation in treatment. The second step removes the variation in y_i that can be predicted by the controls. The final step asks: is this unpredictable variation in treatment correlated with the unpredictable variation in the outcome?

This interpretation is powerful for thinking about causal inference. The regression coefficient captures the relationship between the outcome and the part of the treatment variable that cannot be explained by the included controls. If we believe this residual variation is "as good as random" – uncorrelated with other omitted factors – then we can interpret the coefficient causally.

This is why the choice of control variables is so important. Good controls are variables that predict both treatment assignment and the outcome, helping to isolate truly random variation in treatment. The FWL theorem shows that regression is fundamentally about identifying relationships in the variation that remains after accounting for observable differences. Understanding this helps us think more clearly about when regression estimates are likely to be credible and when we should worry about omitted variable bias.