Lecture 6

Regression Specifications and Estimation

Ananya Iyengar

January 2026

St. Stephen's College

This corresponds to Sections 7.1-7.3 of Studenmund (2017) and 6.1-6.5 of Gujarati and Porter (2009).

Theoretical considerations guide the choice of the functional form of the regression equation. As long as the specified model is linear in parameters, we have briefly encountered non-linear relationships between the dependent and explanatory variables. In this lecture, we will be introduced to different functional forms that lend a variety of interpretations to econometric models and allow us to examine a wide range of relationships between measurable variables.

1 The Regression Intercept β_0

The *estimated* $\hat{\beta}$ may theoretical interpretation, such as fixed costs in estimating a cost function. Even when some direct theoretical interpretation does not lend itself to the intercept, it is important to include it in the regression.

We only must suppress the intercept if we believe that $E[Y/X = 0] = 0$ i.e. the conditional distribution of the dependent variable is centered at 0. If not, suppressing the *intercept forces us to fit a regression line through* the origin, which can lead to biased estimates in the case of a mis-specified model.

Of course, in some cases, a regression through the origin may be what is theoretically accurate.

Consider $Y = \beta X + \epsilon$. The sum of squared errors is given by $\sum(Y_i - \beta X_i)^2$. Differentiating w.r.t. β, we get $-2\sum(Y_i - \beta X_i)X_i = 0 \implies \hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2}$. This $\hat{\beta}$ is unbiased. However,

1. This estimated regression line may not pass through the point $(\bar{X}, \bar{Y})$.

2. The coefficient of determination is no longer allowed the usual interpretation.

3. The residuals may not sum up to 0.

2 Alternate functional forms that are linear in parameters

2.1 Double Log

Natural log is applied to both the X and Y variables. The regression equation takes the form $lnY_i = \beta_0 + \beta_1 lnX_{1i} + ... + \beta_k lnX_{ki} + \epsilon$. Then, $\beta_j = \frac{\delta lnY}{\delta lnX_j} = \frac{\delta Y}{\delta X_j}\frac{X_j}{Y}$, which is the *elasticity* of Y with respect to X_j

2.2 Log-Linear

Natural log is applied to the Y variable. The regression equation takes the form $lnY_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon$. Then, $\beta_j = \frac{\delta lnY}{\delta X_j} = \frac{\delta Y/Y}{\delta X_j}$. Here, for a one unit increase in X_j, keeping all other X_{-j}s constant, is associated with a $\beta_j \times 100$ percent change in the mean value of Y_i.

2.3 Linear-Log

Natural log is applied to some (not necessarily all) X variables. The regression equation takes the form $Y_i = \beta_0 + \beta_1 lnX_{1i} + ... + \beta_k lnX_{ki} + \epsilon$. Then, $\beta_j = \frac{\delta Y}{\delta lnX_j}$. Here, a one percent increase in X_j, keeping all other X_{-j}s constant, is associated with a $\frac{\beta_j}{100}$ units change in the mean value of Y_i.

2.4 Polynomial Forms

Economic theory dictates functional form. For example, the marginal cost function (relationship between cost and marginal production) is often considered to be U-shaped. Any n-th powered polynomial can be used to estimate such economic relationships. In case of the quadratic polynomial, we estimate $Y_i = \beta_0 + \beta_1 X_i^2 + \epsilon$. Here, a one unit change in X_i is associated with, on an average, a $2\beta_1$ unit change in Y_i.

2.5 Lags

Consider time series data (X_t, Y_t), where the slope coefficients from regression Y_t on X_t is the instantaneous rate of change. It might be of interest to also look at the impact of past values of the explanatory variable $X_{t-1}, X_{t-2}, ..$ on the dependent variable. Different lag structures are used on the basis of the context, data structure, and characteristics of the error term. A regression with a lag of one time period can take the form $Y_t = \beta_0 + \beta_1 X_{t-1} + \epsilon$.

3 Indicator/Dummy Variables

Also called qualitative variables, indicator/dummy variables (D_i) switch on when an observation fulfills a certain qualitative attribute and off when it does not. This qualitative attribute can be deterministic (for e.g., $D_i = 1$ if individual i lives in Delhi, else $D_i = 0$), self-reported (for e.g., $D_i = 1$ if individual i is a conservative, else $D_i = 0$) or constructed by the researcher (for e.g., $D_i = 1$ if a child has weight-to-age ratio less than 1.5 sd below the mean, else $D_i = 0$).

Consider a simple regression that only has dummy variables as explanatory variables. This is also called an *ANOVA (Analysis of Variance)* model.

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

Here, suppose Y_i are wages and $D_i \in \{0, 1\}$ is gender. The dummy variable takes value 1 for men. Then,

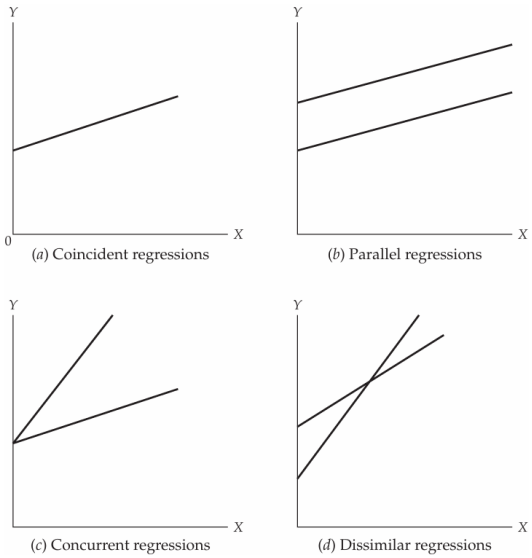
$$E[Y/D = 1] = \beta_0 + \beta_1 \text{ and } E[Y/D = 0] = \beta_0$$

Therefore, the coefficient on the dummy variable β_1 depicts the gap in the mean wages between men and women and provides a **differential intercept**. This model can be extended to include other measurable regressors as well (these are called *Analysis of Covariance* or ANCOVA models).

Note:

1. The qualitative category which is attributed to $D_i = 0$ is called the *base* or *reference* category.
2. The assignment of the reference category will impact the direction and magnitude of the coefficient, but will not change the overall fit or interpretation of the model.

Figure 1: Comparing regressions with dummy variables



3. If we have m categories, we must use $m-1$ dummy variables when running a regression with intercept.

4. Failure to do so leads to a situation of perfect multicollinearity (aka the dummy variable trap), which a violation of the CLRM assumptions and prevents estimating the regression equation.

Dummy variable regressions can be used to encapsulate a variety of relationships between groups, keeping other covariates constant. Some illustrative examples are as follows:

1. *Intercept-shift:* $Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \epsilon_i$.

2. *Multiple Categorical Variables:* $Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \epsilon_i$.

3. *Interaction of two Dummy Variables:* $Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + \epsilon_i$.

4. *Differential slope coefficients:* $Y_i = \beta_0 + \beta_1 X_i + (\beta_2 \times D_i) X_i + \epsilon_i$.

5. *Differential intercept and slope coefficients:* $Y_i = \beta_0 + \beta_1 X_i + (\beta_2 \times D_i) X_i + \beta_3 D_i + \epsilon_i$. The differential slope coefficient is also called the *slope drifter*.

Therefore, for every pair of categories being compared using dummy variables, we can have 4 types of comparisons as shown in Figure 1