BUSS975 Causal Inference in Financial Research

Ji-Woong Chung
chung_jiwoong@korea.ac.kr
Korea University Business School

Chapter 6

Linear Regression 1

6.1 The CEF and Causality

6.1.1 The Conditional Expectation Function (CEF) and Decomposition of Y

In regression analysis and causal inference, *conditional expectations* $E[Y_i \mid D_i]$ play a pivotal role. A fundamental identity from probability theory is that any random variable Y can be decomposed into a part explained by some other variable(s) X and a residual part uncorrelated with X. Formally, for any two random variables Y and X, we can write:

$$Y = E(Y \mid X) + u,$$

where u is the "unexplained" part of Y given X, satisfying $E(u \mid X) = 0$. This equation is often referred to as the **CEF decomposition** of Y with respect to X. It says that $E(Y \mid X)$ is the best prediction of Y given X, and u is the deviation of Y from this prediction. By construction, u has mean zero conditional on X. In other words, $E(u \mid X = x) = 0$ for every value x of X.[1] An important implication is that u is mean-independent of X; in fact, u is uncorrelated with *any function* of X. For any function $h(X)$, we have

[1] This follows since $E(u \mid X) = E(Y - E(Y \mid X) \mid X) = E(Y \mid X) - E[E(Y \mid X) \mid X] = 0$.

$E[h(X)\,u] = E[\,E(h(X)u \mid X)\,] = E[h(X)E(u \mid X)] = 0$. Thus, u cannot be linearly predicted by X (or any known transformation of X).

In this decomposition:

- $E(Y \mid X)$ is called the **conditional expectation of Y given X**. It represents the part of Y that is systematically "explained" by X.
- u (often called the *error term*) captures everything about Y that is not explained by X. By construction, u has zero mean given X and is uncorrelated with X. Intuitively, u is the variation in Y that is unrelated (in a mean sense) to X.

This decomposition provides a natural way to think about the relationship between X and Y. The function $m(X) = E(Y \mid X)$ is often called the **conditional expectation function (CEF)** of Y given X. It tells us the expected value of Y for each value of X. If we know X, the best guess for Y (in terms of least mean squared error) is $E(Y \mid X)$. In fact, the CEF has an important optimality property: it is the *best predictor of Y given X in the mean-squared error sense*. Formally, $E(Y \mid X)$ minimizes the mean squared prediction error

$$E\big[(Y - g(X))^2\big]$$

over all measurable functions $g(X)$. That is, for any other function $g(X)$,

$$E\Big[(Y - E(Y \mid X))^2\Big] \;\le\; E\Big[(Y - g(X))^2\Big].$$

In other words,

$$E(Y \mid X) = \arg\min_{g(\cdot)} E\big[(Y - g(X))^2\big].$$

This can be shown by a simple argument: take any candidate function $g(X)$ and write

$$(Y - g(X))^2 = \big(Y - E(Y \mid X) + E(Y \mid X) - g(X)\big)^2.$$

Expanding this and taking expectations (conditioning on X), one finds

$$E\big[(Y - g(X))^2\big] = E\Big[(Y - E(Y \mid X))^2\Big] + E\Big[(E(Y \mid X) - g(X))^2\Big],$$

since $E(Y \mid X) - g(X)$ is a function of X and $E[Y - E(Y \mid X) \mid X] = 0$. The second term $E[(E(Y \mid X) - g(X))^2]$ is nonnegative (and is zero if and only if $g(X) = E(Y \mid X)$ almost surely). This establishes that $E(Y \mid X)$ yields the smallest possible mean squared error. In summary, the CEF $E(Y \mid X)$ is:

- A natural way to summarize the relationship between X and Y (it tells us the average Y for each X).
- The unique function of X that best predicts Y in terms of minimum mean squared error.

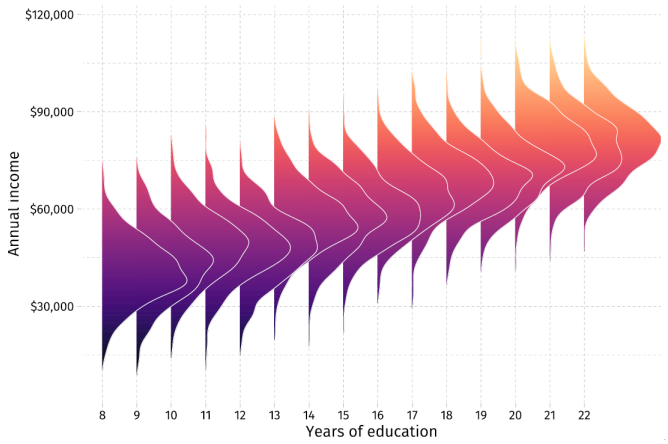

Figure 6.1: For any given value of X, the distribution of Y is centered at $E(Y \mid X)$. The dots indicate the mean of Y for each given X.

Visualizing the CEF

It may help to visualize what the CEF represents. Imagine we have many observations of (X, Y). For any fixed value of $X = x$, consider the distribution of the corresponding Y values. The conditional expectation $E(Y \mid X = x)$ is the mean of this conditional distribution of Y given $X = x$. Figure 6.1 illustrates this idea. For several different values of X along the horizontal axis, the figure depicts the distribution of Y (for example, as a cloud of points or a histogram) at that X value. Each of these conditional distributions is centered at its mean $E(Y \mid X = x)$, indicated by a dot. If we plot the point

$(x, E(Y \mid X = x))$ for every possible x, we trace out a curve: that curve is the conditional expectation function $E(Y \mid X)$ itself. Figure 6.2 shows the CEF connecting the means of the conditional distributions. This curve is fixed (for a given data-generating process) but generally unknown to us — it is fundamentally a property of the population. Our goal as analysts is often to *learn* or estimate this CEF from data. Figure 6.3 illustrates this goal: we observe sample data (points) and wish to infer the underlying conditional expectation function that relates Y to X.

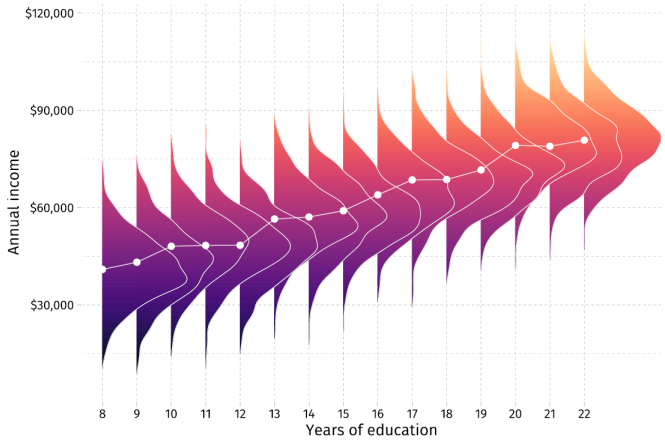

Figure 6.2: The conditional expectation function $E(Y \mid X)$ (solid line) connects the means of the conditional distributions of Y for each X.

6.1.2 Linear Regression as the Best Linear Predictor of the CEF

One of the most popular modeling approaches in empirical work is **linear regression**. Linear regression provides a simple, transparent, and intuitive way to summarize relationships in data. Even if our ultimate interest is not causal inference, linear regression is useful as a descriptive tool. Here we discuss how linear regression relates to the CEF.

In general, linear regression does *not* recover the full CEF unless the true CEF happens to be linear. Instead, what linear regression provides is the **best linear approximation** to the CEF. In other words, linear regression

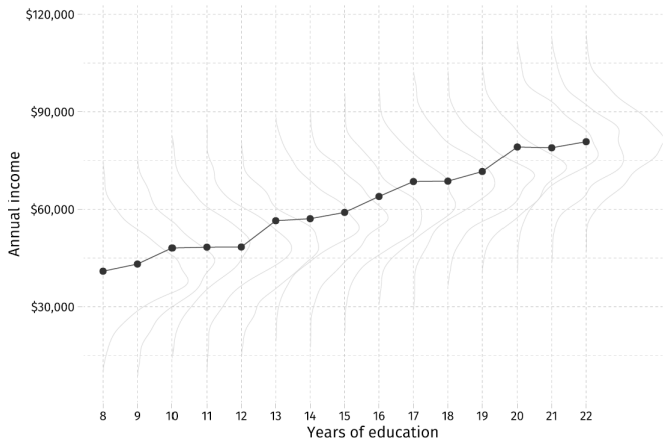

Figure 6.3: Our aim in regression analysis is often to learn or estimate the CEF from sample data (points).

finds the linear function of X that comes closest (in a mean-squared sense) to the actual $E(Y \mid X)$.

To be concrete, consider a simple linear regression model with one regressor X:

$$Y = \alpha + \beta X + u.$$

Here (Y, X) are observable random variables, and α and β are fixed (unknown) parameters. The term u represents everything that affects Y other than X – it is the part of Y not captured by the linear function $\alpha + \beta X$. Importantly, u may encompass a wide range of factors (other variables, randomness, etc.). We will usually assume $E(u) = 0$ (if not, the constant term α can absorb any non-zero mean in u). Our goal in running a regression is typically to estimate the slope coefficient β, which measures the association between X and Y in this linear specification.

Now, how is β related to the CEF $E(Y \mid X)$? If the true conditional expectation $E(Y \mid X)$ is exactly a linear function of X, say

$$E(Y \mid X) = \alpha^* + \beta^* X,$$

then clearly $\alpha^* + \beta^* X$ is the best possible linear predictor of Y given X. In fact, in that case $\alpha^* = \alpha$ and $\beta^* = \beta$ in our regression model (the linear regression would recover the true CEF parameters). But if $E(Y \mid X)$ is

nonlinear, no line can perfectly capture it. Nevertheless, there is a linear function of X that is "best" in terms of approximating $E(Y \mid X)$ as closely as possible. This is known as the **best linear predictor (BLP)** of Y given X (or equivalently, the best linear approximation to the CEF).

Formally, the BLP $(\alpha^{BLP}, \beta^{BLP})$ is defined as the solution to the following minimization problem:

$$(\alpha^{BLP}, \beta^{BLP}) \;=\; \arg\min_{\alpha,\beta} \; E\big[(Y - \alpha - \beta X)^2\big] \,.$$

That is, out of all lines $\alpha + \beta X$, we choose the one that minimizes the mean squared error when predicting Y. This optimization yields normal equations (first-order conditions):

$$E[\, Y - \alpha^{BLP} - \beta^{BLP} X \,] = 0,$$

$$E[\, X(Y - \alpha^{BLP} - \beta^{BLP} X) \,] = 0.$$

These conditions can be solved for α^{BLP} and β^{BLP}. The first implies $E(Y) - \alpha^{BLP} - \beta^{BLP} E(X) = 0$, so

$$\alpha^{BLP} = E(Y) - \beta^{BLP}\, E(X).$$

Substituting this into the second condition gives

$$E[X\,Y] - \big(E(Y) - \beta^{BLP} E(X)\big) E[X] - \beta^{BLP} E[X^2] = 0.$$

Simplifying,

$$E[X\,Y] - E(X)E(Y) - \beta^{BLP}\big(E[X^2] - [E(X)]^2\big) = 0.$$

Recognizing $E[X\,Y] - E(X)E(Y)$ as $\mathrm{Cov}(X, Y)$ and $E[X^2] - [E(X)]^2$ as $\mathrm{Var}(X)$, we find:

$$\beta^{BLP} = \frac{Cov(X, Y)}{Var(X)} \,,$$

provided $Var(X) > 0$. In other words, β^{BLP} is the slope of the linear regression of Y on X in the population. The intercept then is

$$\alpha^{BLP} = E(Y) - \beta^{BLP} E(X).$$

So the BLP line can be written as

$$\alpha^{BLP} + \beta^{BLP} X = E(Y) + \beta^{BLP}(X - E(X)).$$

By construction, this BLP has some important properties:

1. If the true conditional expectation is linear in X, then the BLP is the true CEF. In fact, if $E(Y \mid X) = \alpha^* + \beta^* X$ for some constants α^*, β^*, it must be that $\alpha^{BLP} = \alpha^*$ and $\beta^{BLP} = \beta^*$. [2]

2. $\alpha^{BLP} + \beta^{BLP} X$ is the *minimum-MSE linear predictor* of Y given X. By definition of β^{BLP}, no other linear combination of X yields a lower mean squared error in predicting Y. In fact, one can show

$$\beta^{BLP}(X - E(X)) = Proj(Y \mid X),$$

the projection of Y on the space of X (a one-dimensional space in this simple case).

3. $\alpha^{BLP} + \beta^{BLP} X$ is the *best linear approximation* to the actual $E(Y \mid X)$. That is, it is the linear function of X that comes closest on average to $E(Y \mid X)$. More formally, $(\alpha^{BLP}, \beta^{BLP}) = \arg\min_{\alpha,\beta} E\Big[\big(E(Y \mid X) - (\alpha + \beta X)\big)^2\Big]$. This holds because

$$E\big[(Y - (\alpha + \beta X))^2\big] = E\Big[\big(E(Y \mid X) - (\alpha + \beta X)\big)^2\Big] + E\big[Var(Y \mid X)\big],$$

and the second term $E[Var(Y \mid X)]$ does not depend on α, β. Thus minimizing $E[(Y - (\alpha + \beta X))^2]$ is equivalent to minimizing $E[(E(Y \mid X) - (\alpha + \beta X))^2]$. In other words, the BLP line is also the closest linear fit to the CEF itself.

The phrase "best" above always means in the least-squares sense (minimum mean squared error). Figure 6.4 provides a visual illustration. If the true CEF $E(Y \mid X)$ is non-linear (the curved solid line), the BLP is the straight dashed line that best approximates that curve. Even though it is not exact, it captures the overall increasing trend in $E(Y \mid X)$ in this example.

An important consequence of the normal equations is that the **population regression residual** $u = Y - (\alpha^{BLP} + \beta^{BLP} X)$ is uncorrelated with

[2] *Proof:* In this case, the linear function $\alpha^* + \beta^* X$ achieves zero mean squared error (since $Y = \alpha^* + \beta^* X + u$ with $E(u \mid X) = 0$), so it must solve the minimization defining the BLP. More directly, plugging $E(Y \mid X) = \alpha^* + \beta^* X$ into the normal equations above, we get $E[X(Y - \alpha^* - \beta^* X)] = E[X(u)] = 0$ and $E[Y - \alpha^* - \beta^* X] = E(u) = 0$, which means α^* and β^* satisfy the same equations as $\alpha^{BLP}, \beta^{BLP}$. Hence they must be equal.

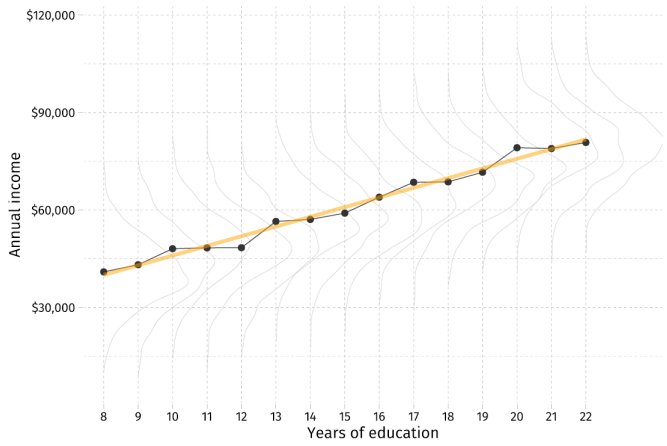

Figure 6.4: The population regression line (dashed) provides the best linear approximation to a possibly nonlinear CEF (solid curve).

X. Indeed, $E[Xu] = 0$ by construction of β^{BLP}. Also $E[u] = 0$ since $E(Y) = \alpha^{BLP} + \beta^{BLP}E(X)$. However, one should note that uncorrelatedness does *not* imply full independence; if the true relationship is nonlinear, u will generally be correlated with nonlinear functions of X. We can decompose the regression residual into two parts:

$$u = Y-(\alpha^{BLP}+\beta^{BLP}X) = \underbrace{Y - E(Y \mid X)}_{\text{pure prediction error}} + \underbrace{E(Y \mid X) - (\alpha^{BLP} + \beta^{BLP}X)}_{\text{linear approximation error}}.$$

The first part $Y - E(Y \mid X)$ is the irreducible uncertainty in Y given X (it has zero conditional mean by definition). The second part is the error from approximating the true CEF with a line; if $E(Y \mid X)$ is not exactly linear, this term is generally not zero and, importantly, will generally be correlated with X in a nonlinear way. Thus, the overall u is uncorrelated with X (in fact, orthogonal to the space spanned by 1 and X), but not necessarily independent of X.

Linear regression, therefore, gives us a very useful summary: even if $E(Y \mid X)$ is complicated, $\alpha^{BLP} + \beta^{BLP}X$ tells us the best linear approximation to that relationship. This is one reason linear regression is so ubiquitous: it provides a simple linear summary of possibly complex relationships, and it often serves as a starting point for more complex models.

6.1.3 Association versus Causality

Thus far, we have discussed linear regression as a descriptive tool for summarizing associations between X and Y. We must be careful, however, not to automatically interpret such associations as **causal effects**. The linear regression line $\hat{Y} = \alpha^{BLP} + \beta^{BLP}X$ captures how Y varies with X on average, but this need not reflect the outcome change we would see if we actively intervened on X.

In general, the coefficient β^{BLP} measures the change in the *predicted average of* Y associated with a one-unit difference in X. Specifically, $\beta^{BLP} = \frac{\partial}{\partial X}\big(\alpha^{BLP}+\beta^{BLP}X\big)$, so it is the slope of the best-fit line through the $E(Y \mid X)$ curve. If the CEF is exactly linear, then $\beta^{BLP} = \frac{\partial E(Y|X)}{\partial X}$ everywhere, and a one-unit increase in X is associated with an increase of β^{BLP} in Y on average. In such a case, if certain conditions hold (to be discussed below), one could interpret β^{BLP} as the *causal effect* of X on Y. However, if $E(Y \mid X)$ is nonlinear, β^{BLP} is better thought of as an average rate of change or a linear summary of the effect of X on Y. Even then, association is not causation without further assumptions.

To draw causal conclusions from a regression, we typically need to assume that X is **exogenous** or **unconfounded** with respect to Y—roughly speaking, X must not be related to other unobserved factors that also affect Y. The linear regression by itself cannot guarantee causality; it only describes associations. In a regression like $Y = \alpha+\beta X+u$, β captures the *approximate expected difference in* Y *associated with a one-unit difference in* X, but this difference is not necessarily due to X *causing* Y. To make a causal interpretation, we require additional assumptions about u and its relationship with X.

Key Assumptions for Causal Interpretation of OLS

The crucial assumption needed to interpret β causally is that X is **unrelated to the error term** u, which contains all other determinants of Y. In our simple model $Y = \alpha + \beta X + u$, the conditions are:

Assumption 1: $E(u) = 0$**.** Without loss of generality we can assume the error has zero mean. If it does not, a constant term in the regression

can absorb any constant mean in u. This assumption is easily satisfied by including an intercept (α) in the regression, so it is not restrictive.

Assumption 2: $E(u \mid X) = E(u)$**.** This means that the average of the error term does not depend on X. Given Assumption 1 (and an intercept in the model), $E(u) = 0$, so Assumption 2 is equivalent to $E(u \mid X) = 0$ for all values of X. This is the **conditional mean independence (CMI)** assumption. In words, it says: once we account for X, the remaining factors u have the same average effect on Y regardless of the value of X. Another way to put it is that X contains all the systematic information about Y, and whatever u represents is essentially "noise" that is unrelated to X on average.

Assumption 2 (CMI) is the formal statement that X is **exogenous** in this model. It implies $E(u \mid X = x) = 0$ for every x. If this holds, then indeed we have

$$E(Y \mid X = x) = \alpha + \beta x + E(u \mid X = x) = \alpha + \beta x,$$

meaning the linear regression function $\alpha + \beta X$ coincides with the true CEF. In that case, a change in X is associated with a change in Y only through this function, and we can interpret β as the causal effect of X on Y (under the linearity assumption). If X changes from a to b, the expected change in Y is:

$$E(Y \mid X = b) - E(Y \mid X = a) = [\alpha + \beta b + E(u \mid X = b)] - [\alpha + \beta a + E(u \mid X = a)].$$

Under CMI, $E(u \mid X = b) = E(u \mid X = a)$ (both equal $E(u)$, which is zero if we centered u). Thus the above difference simplifies to $\beta(b - a)$. This is precisely what we would think of as the causal effect of changing X by $b - a$ (assuming linearity). Without CMI, however, the difference in conditional means includes an extra term $E(u \mid X = b) - E(u \mid X = a)$ reflecting how the unobservables shift when X changes. For example,

$$E(Y \mid X = b) - E(Y \mid X = a) = \beta(b - a) + [E(u \mid X = b) - E(u \mid X = a)].$$

Unless $E(u \mid X)$ is constant, the second term is nonzero, meaning the difference in Y is not purely $\beta(b - a)$. In short, **the coefficient β represents the causal effect of X on Y only if $E(u \mid X)$ is the same for all X (zero, without loss of generality)**. This conditional mean independence assumption is the key to making a causal inference from an OLS regression.

Correlation vs. Causation: When we run OLS on observational data, Assumption 2 may or may not hold. If it fails, β will be biased as an estimator of the causal effect. It is worth noting that:

- CMI (Exogeneity) implies that X is not only uncorrelated with u, but in fact u has mean zero in every slice of the data defined by X. This is a strong condition (no dependence of the error on X at all).
- A weaker condition is that X is merely *uncorrelated* with u: $E[X\,u] = 0$. This is enough to ensure the *consistency* of OLS estimates (as we will discuss later), though not necessarily unbiasedness in finite samples. Intuitively, if $E[Xu] = 0$, then as the sample size grows, the sample covariance between X and the residual approaches zero, so the OLS slope converges to the correct value. However, in small samples there could still be correlation unless the stronger condition $E(u \mid X) = 0$ holds exactly. We often talk about regressors being *exogenous* or *uncorrelated with the error*, referring to this moment condition.

In practice, we usually focus on whether an estimator is **consistent** for the causal effect (i.e., whether it converges to the right answer as $n \to \infty$). For consistency we require $E[Xu] = 0$ (no overall correlation between X and u in the population). For unbiasedness in finite samples (a stronger criterion), one typically needs the full $E(u \mid X) = 0$. In most cases, it is hard to guarantee unbiasedness in finite samples because we only have one realization of X in our data. So econometric analysis emphasizes asymptotic properties like consistency, which rely on moment conditions such as $E[Xu] = 0$. In summary, we assume (or arrange, via research design) that X is as good as randomly assigned with respect to the unobservables in u. When that assumption is plausible, we can interpret β as the causal effect of X on Y.

How plausible is exogeneity? In many observational studies, one must be cautious because there are plenty of reasons why X might fail to satisfy CMI. The error term u captures "everything else" affecting Y, and it is hard to believe that X is completely unrelated to all those other factors. In fact, any of the following situations can violate the $E(u \mid X) = 0$ condition:

- **Omitted variables:** There may be some relevant variables that influence Y which we have not included in the model, and those omitted

factors are correlated with X. For example, suppose Y is a student's test score and X is the number of hours the student studied. If we omit the student's innate ability or prior knowledge (which certainly affects the score) and if that is correlated with study time, then X will pick up not just the effect of studying but also the effect of ability differences, leading to bias.

- **Reverse causality (Simultaneity):** X might be influenced by Y (or by factors related to Y). In this case, X is not a one-way causal input but part of a simultaneous relationship. For instance, in supply-and-demand settings, price and quantity are determined together; if we regress quantity on price, price is not exogenous because it responds to demand shocks (which are in u).

- **Measurement error:** If X is measured with error, then the observed X is $X_{\text{obs}} = X_{\text{true}} + (\text{error})$. The measurement error will end up in the residual u (since the regression uses X_{obs} in place of true X), and typically this error is correlated with X_{obs}, violating exogeneity. Measurement error in the independent variable generally biases OLS estimates toward zero (this is called attenuation bias).

(In econometrics jargon, any violation of $E(u \mid X) = 0$ is often referred to as an **"endogeneity problem"**—meaning some endogenous determination or correlation exists between X and the error term. It is always important to be specific about what the source of endogeneity is in a given context, rather than just stating that a model "has endogeneity." The three broad categories above cover most common sources of endogeneity: omitted confounders, simultaneity, and errors-in-variables.)

Examples illustrating potential violations of CMI: *Example 1: CEO Compensation and Firm Performance.* Suppose we regress CEO salary (Y) on the firm's return on equity (ROE, denoted X) across companies:

$$\text{Salary}_i = \alpha + \beta \text{ROE}_i + u_i,$$

where Salary is annual CEO pay (say, in thousands of dollars) and ROE is a percentage measure of firm profitability. We might find some association (e.g., $\beta > 0$ perhaps). However, is β capturing a causal effect of profitability

on salary? The error term u_i contains all other factors that affect CEO salary besides ROE. There are many such factors: firm size, industry, CEO experience, company growth opportunities, etc. If those factors are correlated with ROE, then $E(u \mid \text{ROE}) \neq 0$. For instance, more profitable firms might also be larger firms, and larger firms tend to pay their CEOs more (even at equal ROE). If we don't control for firm size, then ROE_i will partially proxy for size in explaining salary, biasing the estimate. Or consider risk: unprofitable firms might be facing higher bankruptcy risk, which (by trade-off theory in corporate finance) would lead to more conservative capital structures and potentially lower executive pay growth, etc. Conversely, firms with low profits might have less internal cash and thus operate with more debt (pecking order theory), which could also constrain or affect pay in different ways. These stories indicate likely correlations between ROE and the unobserved factors in u. Thus the simple regression would not isolate a clean causal effect of ROE on salary; β would be biased due to omitted variables like firm size or risk.

Example 2: Capital Structure and Profitability. Consider a regression examining whether less profitable firms use more debt (higher leverage):

$$\text{Leverage}_i = \alpha + \beta \text{Profitability}_i + u_i.$$

Here Leverage might be a debt-to-asset ratio, and Profitability could be return on assets, for firm i. Economic theories suggest contradictory influences: one theory (the trade-off theory) posits that firms with higher bankruptcy risk (often those with low profits) should borrow less to avoid financial distress costs, which would suggest a positive relation between profit and leverage (low profits $\rightarrow$ low leverage on average). Another theory (the pecking order theory) suggests that less profitable firms have less internal funds and therefore have to borrow more, implying a negative relation (low profits $\rightarrow$ high leverage). If we run the regression, u_i contains factors like bankruptcy risk, asset tangibility, growth opportunities, etc. Profitability might be correlated with those: unprofitable firms might indeed have higher risk or different asset types, so $E(u \mid \text{Profitability})$ is not constant. Without controlling for those factors, the coefficient β will not reliably measure a causal effect; it will reflect a mix of those underlying forces.

These examples show why the conditional mean independence assumption can fail: X often moves together with other relevant variables. In the salary

example, ROE alone cannot fully isolate performance effect because it's entangled with firm characteristics; in the leverage example, profitability correlates with other drivers of leverage decisions.

Can we test the exogeneity assumption? A common question is whether we can use the regression residuals to diagnose endogeneity. Suppose after running the OLS regression we obtain fitted values $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ and residuals $\hat{u}_i = Y_i - \hat{Y}_i$. One might consider checking if these residuals are correlated with X. Unfortunately, this check is uninformative for exogeneity, because by construction the OLS residuals in sample are uncorrelated with X. In fact, the OLS estimation ensures that $\frac{1}{n}\sum_i \hat{u}_i = 0$ and $\frac{1}{n}\sum_i X_i \hat{u}_i = 0$. Thus one will always find that the sample correlation between X and $\hat{u}$ is zero, regardless of whether $E(Xu) = 0$ in the population or not. The OLS residuals also have mean zero in the sample even if $E(u) \neq 0$ in reality, because the intercept takes up any mean. Therefore, one cannot "prove" the CMI assumption by looking at regression residual diagnostics alone. Exogeneity is fundamentally an identifying assumption that must be justified by the study design or substantive knowledge, or tested by external means (for example, testing whether adding controls or using instrumental variables changes the estimate).

To summarize this section: linear regression is a powerful tool for summarizing the relationship between X and Y. It gives us the best linear approximation to the true conditional expectation function. However, interpreting regression coefficients as causal effects requires strong assumptions. The critical assumption is that the regressor(s) X are not systematically associated with the unobserved determinants of Y (the error term). Violations of this assumption (endogeneity) can arise from omitted variables, measurement errors, simultaneity, etc. Much of econometric analysis (and many advanced techniques) is devoted to mitigating these issues and achieving credible causal inference.

In the remainder of this chapter, we delve into the mechanics of OLS estimation, interpretation of coefficients, and various practical considerations like rescaling, functional form (log transformations, polynomial terms), and extension to multiple regressors, all under the lens of the basic linear model.

6.2 The Linear OLS Model

6.2.1 Ordinary Least Squares Estimation and Interpretation

When we move from the population concepts of the previous section to actually using data, we rely on the **ordinary least squares (OLS)** method to estimate the coefficients α and β. Suppose we have a random sample of data $\{(X_i, Y_i)\}_{i=1}^n$. The OLS estimates $(\hat{\alpha}, \hat{\beta})$ are defined as the values that minimize the sum of squared residuals:

$$(\hat{\alpha}, \hat{\beta}) \;=\; \arg\min_{\alpha,\beta} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \alpha - \beta X_i\big)^2.$$

This is simply the sample analogue of the BLP problem discussed earlier. Solving this minimization yields the well-known formulas:

$$\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \;=\; \frac{\frac{1}{n}\sum_i X_i Y_i - \bar{X}\bar{Y}}{\frac{1}{n}\sum_i X_i^2 - \bar{X}^2},$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\,\bar{X},$$

where $\bar{X} = \frac{1}{n}\sum_i X_i$ and $\bar{Y} = \frac{1}{n}\sum_i Y_i$ are the sample means. The formula for $\hat{\beta}$ can be recognized as the ratio of the sample covariance of X and Y to the sample variance of X. It is the slope of the sample regression line. The intercept $\hat{\alpha}$ ensures that the regression line goes through the point of means $(\bar{X}, \bar{Y})$.

These OLS estimates $\hat{\alpha}, \hat{\beta}$ are our estimates of the population BLP coefficients $\alpha^{BLP}, \beta^{BLP}$. Under appropriate assumptions (particularly that X is uncorrelated with u), $\hat{\beta}$ will converge in probability to β^{BLP} as the sample size grows. Under the stronger assumption of exogeneity ($E(u \mid X) = 0$), $\hat{\beta}$ is also an unbiased estimator of the true β.

Interpreting the Coefficients: In the simple regression $Y = \alpha + \beta X + u$, the estimated slope $\hat{\beta}$ measures the average change in Y associated with a one-unit increase in X. More precisely, it is the difference in predicted Y when X increases by one (holding all else equal, though in a simple regression

there's no "else"). The intercept $\hat{\alpha}$ is the predicted value of Y when $X = 0$. Depending on the context, $X = 0$ may be outside the range of the data or may not make sense, so the intercept is not always meaningful on its own. However, it ensures the regression line is correctly positioned vertically.

Let's illustrate interpretation with a concrete example.

Example (CEO Pay and Performance). Suppose we regress CEO salary on firm performance (ROE) using data on a set of companies. The model is:

$$\text{Salary}_i = \alpha + \beta\,\text{ROE}_i + u_i,$$

where Salary_i is the annual salary of CEO i (in thousands of dollars) and ROE_i is the return on equity (as a percentage) of firm i. After running OLS, we obtain:

$$\widehat{\text{Salary}}_i = 963.2 + 18.5\ \text{ROE}_i,$$

with $\hat{\alpha} = 963.2$ and $\hat{\beta} = 18.5$. How do we interpret these numbers?

- *Slope ($\hat{\beta} = 18.5$):* This means that an increase in ROE by 1 percentage point is associated with an *increase* in CEO salary of about 18.5 (in the units of the dependent variable, which is thousands of dollars). In other words, if one firm has an ROE that is 1 point higher than another similar firm, this model predicts its CEO's salary would be about \$18,500 higher, on average. This is an *association*, not necessarily a causal effect (for causality we'd need to believe ROE is exogenous or control for other factors).
- *Intercept ($\hat{\alpha} = 963.2$):* This suggests that a firm with ROE = 0 (zero percent return, i.e., breaking even) would pay its CEO about \$963.2 (thousand) = \$963,200 per year, according to the regression line. This is an extrapolation because in practice firms with exactly 0% ROE might be rare, but it gives the baseline level of CEO pay when performance is at zero. Often intercepts are not of primary interest unless $X = 0$ is a meaningful baseline.

It is worth noting that in such a regression, the exogeneity assumption is questionable: factors like firm size, industry, CEO tenure, etc., are omitted and likely correlate with ROE and with salary. Thus, while $\hat{\beta} = 18.5$ tells us about the pay-performance relationship in the data, we should be cautious in

calling it a "return to performance" in a causal sense. We would likely need to control for other variables or use an instrumental variable to get closer to a causal effect.

Rescaling and Units of Measurement

Regression coefficients depend on the units of measurement of the variables. Changing the scale (units) of X or Y will change the numerical values of $\hat{\alpha}$ and $\hat{\beta}$, but it will *not change the underlying relationship or the statistical significance* of the coefficient. We often rescale variables for convenience or interpretability. Let's explore how rescaling works:

- **Scaling the dependent variable Y:** If we multiply Y by a constant c (for example, converting thousands of dollars to actual dollars by taking $c = 1000$), then all terms in the regression equation get multiplied by c. The new model can be written as

  $$cY = (c\alpha) + (c\beta)X + (c\,u).$$

  This implies the new OLS estimates will be $\hat{\alpha}_{new} = c\,\hat{\alpha}$ and $\hat{\beta}_{new} = c\,\hat{\beta}$. In other words, the intercept and slope are scaled by the same factor c. The residuals also scale by c. For example, in the CEO salary regression, if instead of measuring salary in thousands of dollars we measure it in actual dollars ($c = 1000$), the regression would become:

  $$\widehat{\text{Salary}}_i(\$) = 963,200 + 18,500 \times \text{ROE}_i.$$

  The coefficient 18.5 (thousands per %) has become 18,500 (dollars per %). The interpretation is unchanged: a 1 percentage point increase in ROE corresponds to \$18,500 higher salary. We just express it in different units. *Note:* While coefficients and standard errors scale, the t-statistics (and p-values) for significance do *not* change, because standard errors scale by the same factor, leaving $\hat{\beta}/SE(\hat{\beta})$ invariant. Thus rescaling Y is essentially a cosmetic change for presentation or interpretability; it does not affect inference or goodness-of-fit in a substantive way.

- **Scaling the independent variable X:** If we change the units of X, the slope will adjust inversely. Suppose we replace X by kX (for some

constant k). Then the model becomes

$$Y = \alpha + \beta(kX) + u = \alpha + (\beta k)X + u.$$

The new slope coefficient is $\beta_{new} = \beta k$. For example, if X was originally measured in percent and we decide to measure it in proportion (where 1 unit = 100%), then $k = 0.01$ (since $X_{\text{new}} = 0.01 \times X_{\text{old}}$). The new slope would be $\beta_{\text{new}} = \beta \times 0.01$. However, be careful: if we explicitly multiply X by 0.01 in the equation, that's equivalent to dividing the coefficient by 0.01 (or multiplying it by 100). It's often easier to think: a one-unit change in the new X is a 100-unit change in the old X. Thus the slope must become 100 times larger to reflect that difference. In our example, originally $\hat{\beta} = 18.5$ (per 1% ROE). If we measure ROE in decimals, then a 0.01 change in the new X is the same as a 1% change in the old units. The slope in the new regression would be $\hat{\beta}_{new} \approx 1850$ (per 1.0 change in decimal, i.e., per 100%). In effect, $\hat{\beta}_{new} = \hat{\beta}_{old} \times 100$. Again, the interpretation remains consistent: now $\hat{\beta}_{new} = 1850$ (in thousands) means if ROE goes up by 1.00 (i.e., 100 percentage points, a full change from 0 to 100%), salary increases by 1850 (thousand \$). Equivalently, a 0.01 (1%) increase leads to 18.5 (thousand \$) increase, same as before.

- **Scaling both X and Y:** If we multiply Y by c and X by k simultaneously, the net effect is:

$$cY = c\alpha + \frac{c}{k}\beta \cdot (kX) + cu.$$

So α gets multiplied by c, and β gets multiplied by c/k. In other words,

$$\hat{\alpha}_{new} = c\,\hat{\alpha} - \hat{\beta}\,c\,(\text{units shift of } X),$$

$$\hat{\beta}_{new} = \frac{c}{k}\,\hat{\beta}.$$

(This formula also encompasses the previous cases: setting $k = 1$ or $c = 1$ as needed.)

Why do we bother with rescaling variables? There are a few practical reasons:

1. **Readability of coefficients:** If a coefficient is extremely small or extremely large in absolute value due to units, it can be hard to read or interpret. For example, suppose Y is measured in dollars and X in billions of dollars. The slope might come out as something like 0.000000456, which is awkward to discuss. By measuring X in millions instead, the coefficient becomes 0.456, which is easier to interpret ("0.456 dollars increase per million dollars of X"). Conversely, if a coefficient is huge (e.g., 1234567890), perhaps the units are too small and should be scaled up. Rescaling by powers of 10 can make coefficients more modest in magnitude. Importantly, such rescaling will change the standard error by the same factor, so the t-statistic remains the same. Thus all hypothesis test conclusions remain identical.

2. **Comparing effect sizes (standardization):** Often we want to gauge the relative economic importance of different variables. If X and Y have very different variances or units, the raw coefficients don't directly tell which has a "bigger effect" in a standardized sense. One way to get a unit-free comparison is to standardize variables by their standard deviations. For instance, define $X^{sd} = X/\sigma_X$ and $Y^{sd} = Y/\sigma_Y$, where σ_X and σ_Y are the sample standard deviations. Regressing Y^{sd} on X^{sd} (with an intercept) will produce a slope $\hat{\beta}^*$ that indicates how many standard deviations Y changes, on average, for a one standard deviation increase in X. This $\hat{\beta}^*$ is essentially the correlation between X and Y (if both are standardized and a constant is included, the slope in simple regression equals the correlation r_{XY}). For example, if we find $\hat{\beta}^* = 0.25$, we interpret that as: a one standard deviation increase in X is associated with a 0.25 standard deviation increase in Y. This provides a sense of whether the effect is large or small in practical terms. A coefficient of 0.25 suggests a moderate effect, whereas 0.01 would be tiny. Standardized coefficients are often used to compare the relative importance of different predictors in a multiple regression.

Shifting (Adding Constants) to Variables: Another transformation is adding or subtracting a constant from variables. If we add a constant k to the independent variable X, the slope coefficient does not change, because the difference $X_i - X_j$ between any two observations remains the same. However,

the intercept will adjust. Consider $X_{\text{new}} = X + k$. Then

$$Y = \alpha + \beta(X_{\text{new}} - k) + u = (\alpha - \beta k) + \beta X_{\text{new}} + u.$$

So the new intercept $\alpha_{new} = \alpha - \beta k$, while $\beta_{new} = \beta$. Similarly, if we add a constant c to Y,

$$(Y + c) = (\alpha + c) + \beta X + u,$$

so the intercept increases by c (to $\alpha + c$) and the slope stays β. In short, shifting the data left/right (in X) or up/down (in Y) moves the regression line vertically but does not tilt it.

A common and useful practice is to **mean-center** certain variables for interpretability. For example, if X has sample mean $\bar{X}$, you might define $X^c = X - \bar{X}$ and regress Y on X^c (including an intercept). The slope will be the same as regressing on X itself, but the intercept now equals the expected value of Y when $X^c = 0$, i.e. when $X = \bar{X}$. In other words, $\hat{\alpha}$ from this regression gives the predicted Y at the mean of X. This can be much more meaningful than the predicted Y at $X = 0$ if 0 is outside the range or not of interest. For instance, if X is years of education (ranging from 8 to 20 in the sample), $X = 0$ is not relevant, but $\bar{X} = 14$ might correspond to a high school graduate. Centering would make the intercept the predicted outcome for an average person, which is interpretable.

We will see later that centering variables is also helpful in models with interaction terms or in differences-in-differences, to clarify the interpretation of coefficients.

6.2.2 Incorporating Nonlinear Relationships in a Linear Model

The term "linear regression" refers to linearity in parameters, not necessarily linearity in the raw variables. We can include transformations of variables (like squares, logs, etc.) as regressors to capture nonlinear relationships, while still using OLS to estimate the parameters linearly. Two very common ways to introduce nonlinearity are:

1. **Logarithmic transformations** of Y or X (or both).

2. **Polynomial terms** (e.g., X^2, X^3, etc.) or other nonlinear functions (like interactions).

These transformations allow a linear regression to fit a much wider variety of shapes, while maintaining ease of estimation and interpretation (with some care).

Why logs? Taking logarithms of variables is a particularly useful transformation in econometrics:

- **Elasticities and percent changes:** Logarithms convert multiplicative relationships into additive ones. If a 1% change in X consistently leads to a b% change in Y, then $\ln(Y)$ and $\ln(X)$ will have a linear relationship: specifically b will be the elasticity. Even if the relationship is not strictly constant elasticity, log transforms often stabilize variance and linearize growth relationships.
- **Reducing skewness and outlier impact:** Taking logs of a positive variable compresses the scale, so very large values are brought closer to the bulk of the data. This can reduce heteroskedasticity and the influence of extreme observations.
- **Interpretation in percentage terms:** Many times we care about proportional changes rather than absolute changes (e.g., a \$1 increase means different things if base is 10 vs 100, but a 10% increase is comparable). Logs allow coefficients to be interpreted in percentage terms, which are often more intuitive.

When we say "log" in economics or statistics, we typically mean the natural logarithm (base e). We use notation $\ln(Y)$.

Let's consider four typical cases of how logs can enter a regression and how to interpret coefficients in each case:

1. **Level-Level:** Neither X nor Y is logged. This is the standard linear model $Y = \alpha + \beta X + u$. Interpretation: β is the change in Y for a one-unit change in X. ("A one unit increase in X is associated with β units change in Y.")

2. **Log-Level:** The dependent variable is log-transformed, but the independent variable is in levels: $\ln(Y) = \alpha + \beta X + u$. Here $100 \cdot \beta$ is approximately the percentage change in Y for a one-unit increase in X. This is because

$$\Delta \ln(Y) = \beta \, \Delta X,$$

so for a small change ΔX, $\Delta \ln(Y) \approx \ln(Y + \Delta Y) - \ln(Y) = \ln(1 + \Delta Y/Y) \approx \Delta Y/Y$ (for small ΔY). Thus $\frac{\Delta Y}{Y} \approx \beta \Delta X$. Multiplying by 100,

$$\%\Delta Y \approx 100 \, \beta \, \Delta X,$$

where ΔX is in the original units. So if $\beta = 0.083$ and $\Delta X = 1$, then $\Delta \ln(Y) = 0.083$ implies roughly an 8.3% increase in Y. This interpretation is exact for infinitesimal changes and a good approximation for small discrete changes. For larger changes, the approximation error grows, and one should convert the log difference back to a percentage (we will address this shortly).

3. **Log-Log:** Both Y and X are in logs: $\ln(Y) = \alpha + \beta \ln(X) + u$. In this case, β itself is the elasticity of Y with respect to X. Specifically,

$$\beta = \frac{\partial \ln(Y)}{\partial \ln(X)} = \frac{\partial Y/Y}{\partial X/X},$$

so β is the percentage change in Y for a 1% change in X. For example, if $\beta = 0.5$, then a 10% increase in X is associated with a 5% increase in Y. The relationship assumes constant elasticity: the proportional effect does not depend on the level of X.

4. **Level-Log:** Y is in levels and X is in logs: $Y = \alpha + \beta \ln(X) + u$. Here, $\beta/100$ represents the absolute change in Y for a 1% change in X. Why? Because

$$\Delta Y = \beta \, \Delta \ln(X),$$

and $100 \, \Delta \ln(X)$ is approximately $\%\Delta X$. So,

$$\Delta Y \approx \beta \frac{\Delta X}{X},$$

thus for a 1% increase in X ($\frac{\Delta X}{X} = 0.01$),

$$\Delta Y \approx \beta \cdot 0.01.$$

Multiply both sides by 100:

$$100\frac{\Delta Y}{1} \approx \beta.$$

This suggests that increasing X by 1% changes Y by $\frac{\beta}{100}$ in the same units Y is measured. For example, if we have

$$\text{Salary} = \alpha + 1812.5 \ln(\text{Sales}) + u,$$

where Salary is in \$000s and Sales is in \$, then $\beta = 1812.5$ means that a 1% increase in Sales is associated with an increase in Salary of $1812.5/100 = 18.125$ (in \$000s), i.e. \$18,125.

These cases are summarized in the following table for clarity:

Model	**Dep. Var.**	**Ind. Var.**	**Interpretation of β**
Level-Level	Y	X	$dY = \beta\, dX$ (units of Y per unit of X)
Log-Level	$\ln(Y)$	X	$d(\%Y) = (100\,\beta)\, dX$ (percent change in Y per unit of X
Log-Log	$\ln(Y)$	$\ln(X)$	$d(\%Y) = \beta\, d(\%X)$ (percent change in Y per percent change i
Level-Log	Y	$\ln(X)$	$dY = (\beta/100)\, d(\%X)$ (change in Y per 1% change in X)

Table 6.1: Interpretation of β in different log/level specifications. Here dX denotes a small change in X, and $d(\%X)$ denotes a small percentage change in X.

As a concrete example, consider wages and education. If we believe each additional year of education yields a constant *percentage* increase in wages (rather than a constant dollar increase), a log-linear model is appropriate:

$$\ln(\text{wage}) = \alpha + \beta \times \text{education} + u.$$

If $\beta = 0.083$, we interpret that as: each additional year of schooling is associated with approximately an 8.3% increase in hourly wage on average (holding other factors constant, if it's multivariate). If a person has 1 more year than another, we expect their wage to be 8.3% higher. The intercept α in this model would be ln(wage) for someone with zero years of education (which is extrapolated and not meaningful except as part of the regression formula).

Another example: regress CEO salary on firm sales, both in log form:

$$\ln(\text{Salary}) = \alpha + 0.257\, \ln(\text{Sales}) + u.$$

The slope 0.257 is an elasticity: it says a 1% increase in firm sales is associated with a 0.257% increase in CEO salary. Or a doubling of sales (100% increase) is associated with roughly a 25.7% increase in salary. This model assumes the percentage increase in salary from a given percentage increase in sales is constant regardless of the firm's size (constant returns-to-scale in that relationship, so to speak).

And for a level-log case: if

$$\text{Salary}_{000} = \alpha + 1.8125 \ln(\text{Sales}),$$

(where Salary is in thousands), $\beta = 1.8125$ indicates that a 1% increase in Sales is associated with an increase of \$1.8125 (in \$000, which is \$1,812.5) in the CEO's salary. A 10% increase in sales would correspond to about \$18,125 higher salary.

Rescaling with Logs: Interestingly, if you change the units of a *logged* variable, the coefficient is unaffected (except the intercept will absorb a constant shift). For example, if Y is in dollars and you switch to thousands of dollars, $\ln(Y_{000}) = \ln(Y) - \ln(1000) = \ln(Y) - 6.9078$. The regression

$$\ln(Y_{000}) = \alpha' + \beta X + u$$

is equivalent to

$$\ln(Y) = (\alpha' + \ln(1000)) + \beta X + u.$$

So the slope β remains the same; only the intercept changes by $\ln(1000)$. Similarly, measuring X in different units inside a log has no effect on the slope. E.g., if X is population, using ln(population) whether population is counted in thousands or single units just shifts $\ln(X)$ by a constant $\ln(1000)$, which again shifts the intercept but not β. This is convenient, because when dealing with log variables, we need not worry about unit conversions for interpretation of slopes.

Interpreting Large Changes in Log Models: The approximation 100β as "percent change" works well for small β or small changes. But if β is large or the change in X is large, one should interpret carefully. The exact relationship in a log-linear model $\ln(Y) = \alpha + \beta X$ is:

$$\ln(Y_{new}) - \ln(Y_{old}) = \beta(X_{new} - X_{old}).$$

Exponentiating both sides:

$$\frac{Y_{new}}{Y_{old}} = \exp[\beta(X_{new} - X_{old})].$$

So the exact percentage change in Y when X changes from a to b is:

$$\%\Delta Y = 100\left(\frac{Y_{new} - Y_{old}}{Y_{old}}\right) = 100\left(\exp[\beta(b-a)] - 1\right).$$

If $\beta(b-a)$ is small (say 0.05), $\exp(0.05) \approx 1.051$, and $100(\exp(0.05) - 1) \approx 5.1\%$, close to $100\beta(b-a) = 5\%$. But if $\beta(b-a)$ is large, the difference is significant. For example, if $\beta = 0.56$ and $b - a = 1$ (a one-unit change in X), then the approximation suggests 56% increase in Y, but the exact change is $100(e^{0.56} - 1) \approx 75\%$. If β is negative, say -1.39 corresponding to a 75% decrease, $100\beta = -139\%$ is nonsensical, whereas $100(e^{-1.39} - 1) = -75\%$ is the correct interpretation (a 75% decrease). So, if a log-model implies a very large percentage change (say more than 10-20%), it is better to compute the exact effect: for a coefficient β and a change ΔX, the exact predicted percentage change in Y is $100(e^{\beta\Delta X} - 1)\%$.

In practice, one can take the coefficient and apply this formula to be precise. For instance, if $\beta = -0.2$ and $\Delta X = 5$, then $\beta\Delta X = -1$. The exact effect is $100(e^{-1} - 1) \approx -63\%$. The approximation would have given -100%, which overstates the effect (implying an impossible scenario of negative values if taken literally).

When to Use Log Variables: As a rule of thumb: - Logs are typically used for variables that are positive and reasonably skewed (e.g., income, sales, size measures, prices). They are great for modeling proportionate effects or growth rates. Many economic theories suggest constant elasticity or multiplicative effects, which logs handle naturally. - Do *not* log variables that are already in percentage form (like an unemployment rate of 5% should not be logged; it's already a percentage. Logging it would answer a weird question of percent changes in the percentage). Instead, you can use the percentage (or a proportion 0.05) directly or consider logistic transforms if bounded. - Do not log variables that take zero or negative values. $\ln(0)$ is $-\infty$ and undefined in regression. If a variable can be zero (e.g., number of patents, or debt issuance which might be zero for some firms), one cannot

straightforwardly log it. Some analysts use $\ln(1+Y)$ to handle zeros (adding a small constant before logging), but be cautious: $\ln(1+Y)$ does not have a clean percentage-change interpretation, especially if Y can be large or if many $Y = 0$. For example, $\ln(1+Y)$ for Y moving from 0 to 1 is a big jump (0 to $\ln(2)$), which does not correspond to a 100% increase or anything intuitive. - For variables measured in years (like age, education in years) or other inherently linear scales, logs are usually not used. An additional year of education is easier to interpret than a percentage change in education (what is a 10% increase in years of education? 1.2 years? Not so straightforward meaning). - If you have a lot of zeros (e.g., many people have zero income from a particular source), an alternative to $\ln(1+Y)$ is to use models designed for nonnegative data (like a Poisson regression or a tobit model if censoring is an issue). There is recent research (e.g., *Cohn, Liu, and Wardlaw (2022) in JFE*) explaining problems with $\ln(1+Y)$ and recommending alternatives such as Poisson pseudo-maximum-likelihood regression, which effectively models $E(Y \mid X)$ in a multiplicative form and can handle zeros properly. In short: avoid $\ln(1+Y)$ if possible; consider an appropriate model if Y has many zeros.

Example: Percentage Point vs. Percent Change. As a tangential clarification: if unemployment falls from 10% to 9%, it has decreased by 1 percentage point, which is a 10% *relative* decrease. It's important to distinguish percentage points (absolute difference in a rate) from percent change (relative difference). A drop from 10% to 8% is a 2 percentage point drop, which is a 20% decrease relative to the initial level. In discussing regression results, be precise: if X is a percentage (say, interest rate) and $\beta = -0.5$ (with Y not logged), one might say "a 1 percentage point increase in interest rate is associated with a 0.5 unit decrease in Y." Only use "percent increase/decrease" when dealing with log specifications or when explicitly talking about relative change.

Polynomial Terms (Quadratics): Another way to model nonlinearity is to include X^2 (and higher powers) as regressors. For example:

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + u.$$

This quadratic model allows the effect of X on Y to change with the level of X. The **marginal effect** of X is now:

$$\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X.$$

This means the impact of a one-unit increase in X depends on the current value of X. If β_2 is positive, the effect grows with X (convex relationship); if β_2 is negative, the effect diminishes with X (concave relationship).

When interpreting such a model: - If $\beta_2 \neq 0$, the effect of X is not constant. We often evaluate the effect at some meaningful value of X (e.g., at the mean of X, or at specific percentile values). For instance, "At the mean value of X (which is 5), a one-unit increase in X is associated with a $\beta_1 + 2\beta_2(5)$ increase in Y." - Quadratic models can exhibit a turning point. If $\beta_2 < 0$ (an inverted-U shape), the turning point (maximum) occurs at $X = -\beta_1/(2\beta_2)$. If $\beta_2 > 0$ (U-shape), the turning point (minimum) is at the same formula. It's important to check if this turning point lies within the range of your data; if it doesn't, the model is suggesting a curvature that goes beyond your observed data (which might not be relevant or might indicate the quadratic term is just capturing a gentle bend rather than a full turn within the sample). - Example: Suppose $\hat{\beta}_1 = 10$ and $\hat{\beta}_2 = -1$ in a model of Y on X and X^2. This implies a concave relationship. The turning point would be at $X = -10/(2 \cdot -1) = 5$. That's where Y is maximized. If your X data ranges from, say, 0 to 10, then $X = 5$ is squarely in the range, and it indicates that Y increases with X up to 5 and then decreases. If your data's range was 6 to 10, then the estimated parabola has a peak at 5 which is outside the range—within the observed range (6 to 10) the relationship would actually be monotonically decreasing. In such a case, including the quadratic might still help fit curvature, but one should be cautious in interpreting a turning point that lies outside the data. - Always graph or think about the shape implied by a polynomial. Sometimes a very large or very small turning point indicates the quadratic term is just adding a slight curve. If the turning point is extreme (e.g., negative or a huge number not in data range), it could mean the quadratic term is not really needed (the relationship might be effectively linear in the observed range) or it could mean one should consider a different functional form if theory suggests a saturating or asymptotic behavior.

In summary, polynomial terms allow flexible curvature, but interpretability becomes more complex since effects depend on levels. One strategy is to

report predicted effects at a few values of X or to compute where the effect is zero (if ever). We will later see that polynomials of higher order can be used for even more flexibility (though one must avoid extrapolating polynomials too far beyond data as they can behave wildly).

6.2.3 Multiple Regression: Adding More Regressors

So far we have focused on one independent variable X. Rarely in practice do we have a situation where only one factor matters. Most analyses involve multiple regressors. The multiple linear regression model can be written as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u,$$

for k regressors (not counting the intercept). We can also use vector/matrix notation: $Y = \alpha + \mathbf{X}'\boldsymbol{\beta} + u$, where $\mathbf{X} = (X_1, \ldots, X_k)'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$.

The interpretation of coefficients in multiple regression is **ceteris paribus**—holding all other included variables constant: - β_j is the partial effect of X_j on Y, i.e., the change in Y associated with a one-unit increase in X_j, keeping $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$ fixed. - The intercept α is the expected value of Y when all $X_1, \ldots, X_k = 0$ (if such a scenario is reasonable). If zero is not in the support for some variables or is not meaningful, the intercept is just a baseline that ensures the line passes through the means or adjusts for scale shifts. - The key assumption for causal interpretation generalizes: we need $E(u \mid X_1, X_2, \ldots, X_k) = 0$. That is, after controlling for all k regressors, any remaining factors in u are not systematically related to the included X's. In practice, including more variables in X can reduce omitted variable bias by absorbing some of the variation that could confound the relationship of interest. However, we must also be careful about *multicollinearity* and *overfitting* with many regressors.

Estimation: OLS extends naturally. The coefficients are chosen to minimize $\sum_i (Y_i - \alpha - \beta_1 X_{1i} - \cdots - \beta_k X_{ki})^2$. The normal equations lead to matrix formulas: $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$, where X is the $n \times (k+1)$ design matrix including a column of ones for the intercept and Y is the $n \times 1$ vector of outcomes. The solution exists if the X matrix has full column rank (no exact collinearity among regressors).

Interpreting Coefficients: Let's use an example.

Example: Predicting College GPA. Suppose we regress college GPA (Y) on high school GPA (X_1) and ACT score (X_2) for a sample of 141 students:

$$\widehat{\text{College GPA}} = 1.29 + 0.453 \text{ HS GPA} + 0.0094 \text{ ACT}.$$

Here, high school GPA is on a 4-point scale, ACT is on (roughly) a 0-36 scale, and college GPA is also on a 4-point scale. - $\hat{\beta}_1 = 0.453$ is the coefficient on high school GPA. This means that, holding ACT constant, a one-point higher high school GPA is associated with a 0.453 higher college GPA on average. In other words, if Student A had a high school GPA one point higher than Student B (e.g., 3.7 vs 2.7), and they had the same ACT score, we predict Student A's college GPA to be about 0.453 points higher than B's. This is a partial effect of high school GPA, net of ACT. - $\hat{\beta}_2 = 0.0094$ is the coefficient on ACT. It suggests that, holding HS GPA constant, each additional point on the ACT is associated with only a 0.0094 increase in college GPA. That is a very small effect: even a 10-point increase in ACT (which is a big difference in scores) corresponds to 0.094 higher college GPA. This might indicate that after accounting for high school GPA, ACT doesn't have much predictive power for college grades (or that the scaling is such that 1 ACT point is minor). - The intercept 1.29 would be the predicted college GPA for a student with HS GPA = 0 and ACT = 0. Of course, that's outside the plausible range (nobody has a 0 HS GPA if they made it to college, and ACT 0 is meaningless because min ACT is typically around 11 or so for someone who took it). So the intercept here is not something we interpret literally; it's just anchoring the plane. We might consider centering the predictors to give the intercept meaning (e.g., mean HS GPA and mean ACT yields intercept = mean college GPA, likely around 2.9 or so).

We can also consider combined changes: If a student increased her HS GPA by 1 point and her ACT by 1 point simultaneously, the predicted increase in college GPA would be $0.453(1) + 0.0094(1) = 0.4624$. If HS GPA rose by 2 and ACT by 10 (say comparing a student with 2.0 GPA/20 ACT to one with 4.0 GPA/30 ACT), the difference in predicted college GPA would be $0.453(2) + 0.0094(10) \approx 0.906 + 0.094 = 1.0$ point. That is, the second student would be predicted to have a college GPA one point higher than the first, which is a substantial difference on a 4-point scale. Notice how we just sum the contributions of each variable's change, reflecting the linear additivity of effects in this model.

Fitted Values and Residuals in Multiple Regression: After estimating a multiple regression, each observation has a fitted value $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$ and a residual $\hat{u}_i = Y_i - \hat{Y}_i$. Several useful properties hold (assuming an intercept is included):

- The sample average of the residuals is zero: $\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0$. Consequently, the average of the fitted values equals the average of the actual Y: $\frac{1}{n}\sum_i \hat{Y}_i = \bar{Y}$. The regression line (or hyperplane) thus passes through the point $(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k,\ \bar{Y})$.
- Each independent variable is uncorrelated with the residuals: $\frac{1}{n}\sum_i X_{ji}\,\hat{u}_i = 0$ for each $j = 1, \ldots, k$. This is essentially the sample version of the normal equations $X'\hat{u} = 0$. It means in the sample, the residual has zero sample covariance with each regressor.

Because of these properties, one should not try to "test" the model by regressing residuals on included X's: you will find no linear relationship by construction. (However, plotting residuals against X can still reveal if a nonlinear pattern remains or if there is heteroskedasticity, etc.)

Partial Regression (Frisch-Waugh-Lovell) Interpretation: Multiple regression coefficients can be understood as reflecting the unique contribution of each variable after "netting out" the influence of the others. A fundamental result, the **Frisch-Waugh-Lovell (FWL) theorem**, states that one can compute $\hat{\beta}_1$ (say) in a multiple regression Y on $X_1, X_2, \ldots, X_k$ by a three-step procedure:

1. Regress Y on all the other regressors $X_2, \ldots, X_k$ (excluding X_1). Obtain the residuals $\tilde{Y}_i$ from this regression. These residuals represent the part of Y that is orthogonal to $X_2, \ldots, X_k$ – effectively, Y with the linear effects of $X_2, \ldots, X_k$ removed.
2. Regress X_1 on $X_2, \ldots, X_k$ as well, and get residuals $\tilde{X}_{1i}$. These residuals are the portion of X_1 not explained by $X_2, \ldots, X_k$ – the variation in X_1 that is "left over" after accounting for those other factors.
3. Regress $\tilde{Y}$ on $\tilde{X}_1$ (simple linear regression with no additional controls). The slope coefficient in this regression will equal $\hat{\beta}_1$ from the full multiple regression of Y on all X's. Essentially, this final regression isolates

the relationship between Y and X_1 using only the variation that is unique to X_1 (unrelated to $X_2, \ldots, X_k$).

This result highlights an important insight: $\hat{\beta}_1$ measures the association between X_1 and Y *after removing any linear association of* X_1 *(and* Y*) with the other covariates.* That's why we call β_1 a **partial effect** or partial regression coefficient.

To understand this concretely, suppose a researcher wanted the effect of X on Y controlling for Z. A wrong approach some might take is: "First remove the effect of Z on Y by regressing Y on Z and taking residuals, then regress those residuals on X." This by itself is incomplete. According to FWL, one must also remove the effect of Z from X (get residuals of X on Z) and then do the second regression. If you fail to partial out Z from X as well, the coefficient you get will not generally equal the multiple regression coefficient from including Z. The omitted step means some of the variation in X that is correlated with Z could be erroneously attributed.

In practical terms, what FWL assures us is that the OLS coefficient on X_1 in a multivariate regression is capturing exactly the relationship between Y and the part of X_1 that is uncorrelated with the other covariates. So multiple regression correctly accounts for the overlapping influences of correlated regressors.

Implication: If you ever difference out or residualize data to control for something (like "industry-adjusted" performance = performance - industry average), be careful if you later relate that to another variable. If the other variable also has industry patterns, you should similarly adjust it. Otherwise, you have not truly controlled for industry effects in the relationship between the two. This is a common mistake: for example, subtracting out industry means from Y but not from X and then regressing the adjusted Y on X will generally *not* recover the coefficient you'd get if you included industry dummies in a full regression of Y on X. You must adjust both or, equivalently, just include the controls in a single regression.

6.2.4 Goodness-of-Fit: R^2 and Adjusted R^2

After estimating a regression, one might ask how well the model explains the data. The standard measure of goodness-of-fit in OLS is the R^2, or coefficient of determination. It is defined based on the **sum of squares** decomposition:

$$\text{SST} = \text{SSE} + \text{SSR}.$$

Here: - SST = Total Sum of Squares = $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. This measures the total variation in Y around its mean (how spread out the Y values are). - SSE = Explained Sum of Squares = $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$. This is the variation in the fitted values — how much of the variance in Y is captured by the model's linear prediction. - SSR = Residual (or Error) Sum of Squares = $\sum_{i=1}^{n}\hat{u}_i^2 = \sum(Y_i - \hat{Y}_i)^2$. This is the variation in Y that remains unexplained by the model.

The OLS normal equations ensure that SSE+SSR = SST (when an intercept is included). This is analogous to the formula in ANOVA: Total variability = Explained variability + Unexplained variability.

We then define:

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}.$$

R^2 is the fraction of the total variance in Y that is explained by the regression model. Its value ranges from 0 to 1 (for a model with intercept): - $R^2 = 0$ means the model explains none of the variability in Y (the best you can do is just use $\bar{Y}$; indeed $\hat{\beta} = 0$ would yield that). - $R^2 = 1$ means a perfect fit: all points lie exactly on the regression hyperplane ($\hat{u}_i = 0$ for all i). - Typically, R^2 lies somewhere in between. It's also true that R^2 is the square of the sample correlation between Y and $\hat{Y}$. In simple regression ($k = 1$), R^2 is just the square of the correlation between X and Y. In multiple regression, one can't reduce it to a simple pairwise correlation, but it still gives overall fit.

One caution: R^2 always weakly increases as you add more regressors. If you include an additional variable (even irrelevant ones), SSE cannot decrease (it may stay the same if the variable adds no explanatory power, but usually sample R^2 will go up at least a tiny bit due to sample peculiarities). Therefore, a high R^2 doesn't necessarily mean a good model in terms of causal insight or parsimony; it could mean you've thrown many predictors at it, some possibly spurious.

Because of this, sometimes we look at the **Adjusted** R^2:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1},$$

where k is the number of regressors (excluding the intercept) and n is sample size. Adjusted R^2 imposes a penalty for each additional regressor. If a new variable improves R^2 only slightly (less than what would be expected by chance given one degree of freedom), $\bar{R}^2$ might actually drop. It is possible for $\bar{R}^2$ to go down when adding a variable, whereas R^2 cannot.

Adjusted R^2 is often used to compare models with different numbers of predictors. However, it's not a strict test; it's just heuristic. (More formal model comparison can use F-tests or information criteria like AIC/BIC.)

Interpreting the magnitude of R^2: R^2 tells us how much variance is explained, but the "importance" of this depends on context. For example: - In a physics experiment, one might expect an R^2 of 0.9 or above if the model is correct (since physical laws often yield tight relationships). - In cross-sectional microeconomic data, an R^2 of 0.1 (10% explained variance) is not uncommon or necessarily troubling, because human behavior has a lot of idiosyncratic variation. A seemingly low R^2 does not preclude the key coefficient β from being statistically significant or economically important. - If your goal is accurate prediction, you might want a high R^2. If your goal is estimating a causal effect reliably, R^2 is secondary; you can have a low R^2 but still estimate a coefficient precisely if you have enough data and low noise on that dimension. Conversely, you could have a high R^2 but if it's mostly due to some other variables and your key X is barely varying independently, you might estimate β poorly.

So, a regression with $R^2 = 0.014$ (1.4%) means 98.6% of the variance in Y is left unexplained by the model. Is that "bad"? Not necessarily. It depends on what we're trying to do. If X is a policy variable that we suspect only has a modest effect on Y but is important to identify, an R^2 of 0.014 could be expected, and if we have enough data, we might still get a significant estimate of β. The low R^2 just indicates that Y has a lot of other stuff going on (which could be unobserved noise or many small factors). A classic example: in individual-level studies, variables like education, experience, etc., might only explain a fraction of the variation in earnings because there are

many unmeasured factors (skill, motivation, luck), but that doesn't mean those variables aren't important or that the regression is useless.

On the flip side, a very high R^2 could be suspicious if achieved too easily (like overfitting with many polynomial terms or including outcomes in predictors, etc.). One should use domain knowledge to judge if an R^2 is reasonable.

6.2.5 Unbiasedness and Consistency of OLS

When can we trust OLS estimates as reflecting true relationships? Two key theoretical properties are **unbiasedness** and **consistency**.

Unbiasedness: An estimator $\hat{\theta}$ is unbiased for θ if $E[\hat{\theta}] = \theta$. In our context, $\hat{\beta}_j$ is unbiased for the true coefficient β_j if the expectation of the sampling distribution of $\hat{\beta}_j$ equals β_j.

For OLS in the multiple regression model, under the following assumptions, the OLS estimates are unbiased:

1. *Linear in parameters:* The model indeed is $Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + u$ for some true parameters $(\alpha, \beta_1, \ldots, \beta_k)$, and we have an additive u.

2. *Random sampling:* We have a random sample of $(Y, X_1, \ldots, X_k)$ from the population (so that the data are representative and the error term properties carry over to sample moments).

3. *No perfect multicollinearity:* The independent variables are not exact linear combinations of each other (in the population and thus in the sample).

4. *Zero conditional mean (Exogeneity):* $E(u \mid X_1, \ldots, X_k) = 0$. This is essentially the same CMI assumption generalized to the vector of regressors. It implies each X_j is uncorrelated with u (and actually any function of X is uncorrelated with u).

If these hold, then $E(\hat{\beta}_j) = \beta_j$ for each j. The proof, in brief, is that $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\boldsymbol{\beta} + u) = \boldsymbol{\beta} + (X'X)^{-1}X'u$. Taking

expectations conditional on X (which is treated as non-random in deriving expectation, or using law of iterated expectations):

$$E[\hat{\boldsymbol{\beta}} \mid X] = \boldsymbol{\beta} + (X'X)^{-1}X'E[u \mid X] = \boldsymbol{\beta} + (X'X)^{-1}X' \cdot 0 = \boldsymbol{\beta}.$$

Thus $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. So OLS hits the mark on average.

However, unbiasedness is a finite-sample concept and depends on strict conditions. In many cases, especially with observational data, we worry that $E(u \mid X) \neq 0$. Then OLS is biased. But even if OLS is biased in small samples, it might still become accurate as n grows large under weaker conditions. That's consistency.

Consistency: An estimator $\hat{\theta}_n$ for θ is consistent if $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$. That means in large samples, the estimator will be arbitrarily close to the true value with high probability.

For OLS, a sufficient condition for consistency is:

$$plim_{n\to\infty}\frac{1}{n}X'u = 0,$$

in addition to technical regularity conditions (like the law of large numbers applying, which requires e.g. finite second moments and independent observations, etc.) and the regressors having full rank. The condition $\frac{1}{n}X'u \to 0$ essentially means $E[X_j u] = 0$ for each regressor j. In other words, we need **no correlation between X and u in the long run**. This is weaker than $E(u \mid X) = 0$. It allows for some forms of mild dependence as long as the covariance is zero.

For example, $E(u \mid X)$ might not equal a constant, but it could vary in a way that still yields zero overall covariance. A pathological scenario: if $u = X\epsilon$ where ϵ is a mean-zero shock uncorrelated with X, then $E(u \mid X) = XE(\epsilon) = 0$ actually in that case, but something more tricky: if u has some distribution that's symmetric around 0 for each X but maybe not zero at each X, $E(u \mid X)$ might be zero anyway. Actually, if $E(u \mid X) = 0$ is needed for unbiasedness, for consistency we just need $E(Xu) = 0$. If $Cov(X, u) = 0$, as sample size increases $\hat{\beta}$ will converge to the true β because the bias term involves $E(Xu)$. The formula we derived: $\hat{\beta} = \beta + (X'X)^{-1}X'u$. When n is large, $(X'X)/n \approx E(X_i X_i')$ and $(X'u)/n \approx E(X_i u_i)$. If the latter is zero, then $\hat{\beta} \to \beta$ in probability.

Thus the key for consistency is **zero correlation between each regressor and the error term**. This condition is often referred to as the regressors being exogenous (in a moment sense).

In practice, we emphasize consistency more than unbiasedness because: - Unbiasedness is nice but can be fragile (one small violation and it's gone). - We often rely on large samples; as long as the estimator is consistent, we can get close to the truth with enough data even if it's slightly biased in small samples. - Some estimators are biased in small samples but consistent (e.g., certain nonlinear estimators or ridge regression etc. can be biased but aim to reduce variance).

To reiterate: if X and u are uncorrelated (zero covariance) but $E(u \mid X)$ is not zero, OLS can be consistent but biased in finite samples. One classic scenario is measurement error in Y: $Y^{obs} = Y^{true} + \text{noise}$. The noise might have mean 0 and be independent of X (so $E(u \mid X) = 0$ in fact, giving unbiasedness in that case). But if we had some slight dependence that averages out, that's more unusual. More common is the reverse: $E(u \mid X) = 0$ might fail but $E(Xu) = 0$ holds? Possibly if the distribution of u changes with X but in a symmetric way around 0 such that overall correlation cancels out. It's a technical distinction.

Finally, note that even if OLS is unbiased/consistent for the *best linear approximation* of $E(Y \mid X)$, that still doesn't guarantee β is the true causal effect unless the exogeneity assumption holds. Bias and inconsistency as defined here mean with respect to the "true parameter" of the regression model. If the regression model itself is misspecified (due to omitted variables, etc.), β might consistently estimate a wrong quantity (like the association including bias).

Summary

Let's recap the major points covered:

- **CEF and Best Prediction:** The conditional expectation function $m(x) = E(Y \mid X = x)$ is central to thinking about Y vs X. It gives the true average relationship. It also minimizes mean squared error of prediction. Linear regression does not directly give $E(Y \mid X)$ unless the latter is already linear. Instead, OLS yields the best linear predictor (BLP) of Y given X, i.e., the

closest linear approximation to the CEF.

- **Causality vs Association:** For the regression slope to have a causal interpretation, we require that the error term u is (on average) unrelated to X (exogeneity). If $E(u \mid X) = 0$, then the regression line coincides with the causal CEF. Otherwise, regression may pick up spurious associations (endogeneity issues). We discussed common sources of endogeneity: omitted confounders, simultaneity, and measurement error. In practice, a lot of econometric work is about finding ways to ensure or approximate the condition $E(u \mid X) = 0$ (through study design, instruments, fixed effects, etc.).

- **Scaling and shifting variables:** We can change units or origin of variables for convenience without affecting the substantive relationship. Scaling Y by c scales all coefficients by c. Scaling X_j by k scales β_j by $1/k$. Shifting X (adding a constant) leaves slopes unchanged, only intercept shifts. A useful trick is centering variables (especially when interaction terms or nonlinearities are present) to make the intercept or main effects meaningful (like the effect at average levels). Also, standardizing variables (dividing by std. dev.) can help compare effect sizes in standard deviation terms.

- **Log transformations:** Taking logs of Y and/or X allows interpretation in terms of percentage changes and elasticities. We have to be careful to interpret correctly (especially for large changes, use the exact formula). Logs can handle wide distributions and often linearize growth relationships. But avoid logs when data can be zero/negative; consider alternative approaches for zeros.

- **Nonlinear relationships:** We can accommodate them by including transformed regressors (squares, cubes, interactions, piecewise linear terms, etc.). The model remains linear in parameters, so OLS still applies. For example, a quadratic term allows diminishing or increasing effects. The interpretation becomes: $\beta_1 + 2\beta_2 X$ is the marginal effect at a given X. Always check whether the implied curve makes sense in the observed range.

- **Multiple regression:** Introduce more variables $X_2, \ldots, X_k$ to control for other factors. Coefficients become partial derivatives (ceteris paribus effects). The hope is that controlling for enough confounders moves us closer to an unbiased estimate of the variable of interest. We saw that OLS will ensure residuals have zero correlation with each included regressor (in sample). We explained how partial regression works: the coefficient on X_1 is really

measuring the correlation between Y and X_1 after both have had the linear influence of other regressors removed.

- R^2 **and fit:** R^2 tells us the proportion of variance explained by the model. While an important descriptive statistic, a low R^2 does not invalidate a model if our purpose is estimation rather than prediction. Adjusted R^2 is a variant that accounts for model complexity. But neither R^2 nor $\bar{R}^2$ speaks to causality or correctness of specification; they only gauge in-sample fit.

- **Unbiasedness and consistency:** Under assumptions including exogeneity, OLS is an unbiased estimator of the true coefficients. Even if strict exogeneity fails, as long as regressors are uncorrelated with the error in expectation, OLS is consistent (with enough data, it converges to the best linear approximation parameters). However, if X is correlated with u, OLS will be biased and generally inconsistent for the causal effect. In such cases, other methods (instrumental variables, etc.) are needed, which are topics for subsequent chapters.

In conclusion, linear regression is a powerful and flexible tool that serves as a foundation for more advanced econometric techniques. It provides an easily interpretable summary of relationships and, under the right conditions, allows us to make causal inferences. The challenge in applied work is ensuring those right conditions (or approximations to them) hold, and being aware of the limitations of linear models when relationships are complex or data distributions violate assumptions. We will build on these concepts as we move into methods for dealing with endogeneity and enriching the model structure.