Lecture 8: The Classical Linear Model

1. The Assumptions of the Classical Linear Regression Model

The phrase *classical model* refers to a set of assumptions under which ordinary least squares (OLS) can be said to be the best possible estimator. (“Best” is a term we will use loosely for now, but which has a generally agreed upon definition in terms of the variance of the estimator across different possible samples.) Therefore, the “quality” of a regression will depend in large part on whether or not these standard assumptions are satisfied. If there is evidence that one or more of them are violated, it will be necessary to apply an appropriate correction, if available, or to use a method other than OLS. We will discuss several such corrections later on, but other regression methods are beyond the scope of this course.

The *classical assumptions* are as follows:

1. The regression model is *linear* and *additive* in $\epsilon_i$.
2. The stochastic error term, $\epsilon_i$, has zero mean.
3. The independent variables, $X_{ik}$, are uncorrelated with $\epsilon_i$ (that is, they are *exogenous*).
4. The stochastic error terms, $\epsilon_i$, are uncorrelated.
5. The stochastic error term has constant variance that does not vary with $X_i$ (no *heteroskedasticity*).
6. No independent variable is a perfect linear combination of the other independent variables (no perfect *multicollinearity*).

Note that these are almost all assumptions involving the stochastic error term. We will discuss each of these assumptions in turn below.

1.1. Linear Model with Additive Errors

The first assumption requires that the *population* model, or the “true” model, be linear and that the stochastic error term enters additively (as opposed to, say, multiplicatively):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i.$$ 

If the actual model which generated the data you observe is not of this form, then the OLS coefficients you estimate will be invalid. This is an *assumption*—we can never really know the true model.

This assumption is also not as restrictive as it seems, since even models such as

$$Y_i = e^{\theta_0} X_i^{\theta_1} e^{\epsilon_i}$$
are linear under transformations. Taking the logarithm of both sides of the above equation gives

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + \epsilon_i.$$  

Then, if we let $Y_i^* \equiv \ln(Y_i)$ and $X_i^* \equiv \ln(X_i)$, we have a linear model with an additive error term:

$$Y_i^* = \beta_0 + \beta_1 X_i^* + \epsilon_i.$$  

Implicitly, we are also assuming that the model is correctly specified in the sense that all relevant independent variables (X's) are included in the regression. That is, even though the model is linear, we must be estimating the correct linear specification which includes all of the right variables.

1.2. Zero Mean Errors

The second classical assumption is that the population distribution of $\epsilon_i$ has zero mean:

$$E[\epsilon_i] = 0.$$  

We add this term to the regression to represent the additional variation in the dependent variable ($Y_i$) that cannot be explained by the included independent variables (the $X_i$'s).

Think about there being, for each observation, an associated draw $\epsilon_i$ from some unknown distribution. This variable reflects the unobservable determinants of $Y_i$. Although we don't observe this distribution, we assume it has a zero mean. This means that, aside from the variation in $Y_i$ that we capture via $X_i$, the remaining part has zero mean. For example, $\epsilon_i$ could be Normally distributed with mean $\mu = 0$ and variance $\sigma^2$.

The constant term in the regression, $\beta_0$, essentially captures the fixed part of $Y_i$ that cannot be explained by the regressors while the stochastic error term, $\epsilon_i$, captures the random portion of the unexplained variation. So, although we restrict $\epsilon_i$ to have mean zero, the intercept still allows the unexplained component of $Y_i$ to have nonzero mean.

1.3. Exogenous Explanatory Variables

The third classical assumption is that the explanatory variables are uncorrelated with the error term. If this assumption were violated—if $X_i$ was correlated with $\epsilon_i$—the regression coefficients would probably attribute to $X_i$ some of the variation in $Y_i$ that is actually due to the error term. Suppose that $X_i$ and $\epsilon_i$ are positively correlated. Then we would likely estimate a value of $\hat{\beta}_1$ that is too high, or biased upward.

This assumption can be violated when a researcher omits a relevant independent variable. We call the resulting bias in the OLS coefficients the omitted variable bias. When such a variable is omitted, it is effectively being lumped together with the error term. If the variable is correlated, positively or negatively, with the other independent variables, then the coefficient on the included variables will be biased.

To see this, suppose that $X_{1i}$ and $X_{2i}$ are positively correlated but we run the regression with only $X_{1i}$. The true model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$
but the model we estimate is

\[ Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\epsilon}_i \]

where \( \hat{\epsilon}_i \) captures the variation of both \( \epsilon_i \) and \( X_{2i} \). The estimated coefficients from the second equation will be different from those of the first, and will be biased.

1.4. Uncorrelated Errors (No Serial Correlation)

The fourth classical assumption requires realizations of the error term to be uncorrelated with each other. That is, there must be no systematic correlation across the error terms associated with the observations. In the case of a time series, this type of correlation is referred to as *serial correlation* and this assumption prohibits serial correlation in the error terms. In a cross section, there must be no correlation across individuals. For example, there may be neighborhood-level correlation in unobservables—individuals in the same neighborhood may have similar tastes or preferences which we don’t observe which are important but get picked up by the stochastic error term. In general though, this is more of a concern in time series applications where the effects of a particular shock, say an external event which increases oil prices for several quarters. The values of \( \epsilon_t \) (where \( t \) indexes observations in different time periods) may be correlated over time as a result.

1.5. No Heteroskedasticity

The fifth assumption requires the error term to have constant variance. The variance must not vary with values of \( X_{ki} \) for any \( k \). For example, if there is reason to believe that people in our dataset with high values of \( X_{1i} \) have values of \( \epsilon_i \) with higher variance, then this assumption would be violated.

Referring back to our height-weight example, where we regressed WEIGHT\(_i\) on HEIGHT\(_i\), we must be willing to assume that \( \epsilon_i \) has the same variance for both tall people and short people. The variance must not change with HEIGHT\(_i\). If for some reason the unobserved determinants of WEIGHT\(_i\) had higher variance for taller people, then this assumption would be violated. This is called *heteroskedasticity*.

**Example 1.** Suppose we are studying expenditure on education in the United States at the state level and that we are controlling for population in the regression. We might think that the unobservable determinants of this expenditure are more variable for larger states than for smaller states. Since population is one of our independent variables, the variance of \( \epsilon_i \) depends on population, and so there is heteroskedasticity.

1.6. No Perfect Multicollinearity

The sixth assumption is that no independent variable is a perfect linear combination of the other variables. This situation is called perfect multicollinearity.

**Example 2.** In our education expenditure example from before, suppose that in addition to total population (POP\(_i\)) we control for subsets of the population: children under age 18 (CHILDREN\(_{1i}\)),
adults aged 18 to 65 ($\text{ADULTS}_i$), and senior citizens, age 65 and older ($\text{SENIORS}_i$). These variables are collinear since

$$\text{POP}_i = \text{CHILDREN}_i + \text{ADULTS}_i + \text{SENIORS}_i.$$  

Similarly,

$$\text{CHILDREN}_i = \text{POP}_i - \text{ADULTS}_i - \text{SENIORS}_i$$

and so on.

In addition to the pure additive example above, collinearity can arise in more complex situations. Two variables $A$ and $B$ are collinear if we can write

$$B = \alpha_0 + \alpha_1 \cdot A$$

for any numbers $\alpha_0$ and $\alpha_1$. So, if $B = 2 - \frac{1}{3}A$, then $A$ and $B$ are collinear.

In the case of indicator variables, we have to be careful to exclude at least one categorical indicator variable. For instance, we cannot include both indicators $\text{MALE}_i$ and $\text{FEMALE}_i$ or $\text{USED}_i$ and $\text{NEW}_i$ because

$$\text{USED}_i = 1 - \text{NEW}_i.$$  

For indicator variables with multiple categories, let’s say we have data on sales of American cars produced by the three major manufacturers, with indicators named $\text{CHRYSLER}_i$, $\text{FORD}_i$, and $\text{GM}_i$. Since each car model in our dataset must be manufactured by either Chrysler, Ford, or General Motors, and since each car is manufactured by exactly one manufacturer, these indicator variables must sum to one:

$$\text{CHRYSLER}_i + \text{FORD}_i + \text{GM}_i = 1.$$  

Given this relationship, we know that these variables are multicollinear and we must omit one of them from the regression. In the case of categorical dummy variables like these, the omitted variable is the baseline category, and the coefficients on the included dummies are to be interpreted relative to the baseline.