BUSS975 Causal Inference in Financial Research

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

Chapter 6

Linear Regression 2

6.1 Homoskedasticity, Heteroskedasticity, and OLS Variance

Before turning to formal hypothesis testing, it is important to understand the variance of the OLS estimators under different assumptions about the error term. In particular, we distinguish between *homoskedastic* errors (constant variance) and *heteroskedastic* errors (variance depends on X). This will guide us in obtaining the correct standard errors for inference.

Definition 6.1. We say the error term u_i is **homoskedastic** if $Var(u_i \mid X_i = x) = \sigma^2$ is a constant, i.e. the disturbance variance does *not* depend on the level of the regressors. Conversely, if $Var(u_i \mid X_i = x)$ varies as a function of x, the errors are **heteroskedastic**. In other words, under heteroskedasticity the disturbance variance changes with X.

Consider a simple investment regression example: $Investment_i = \alpha + \beta\, Q_i + u_i$. Homoskedasticity would mean that the variability of the investment unexplained by Q is the same for firms of all sizes or characteristics. In practice, this is often unrealistic — one can easily imagine that firms with higher Q or different sizes have more volatile investment, violating the constant variance assumption. Thus, in many economic settings, heteroskedasticity is a safer and more realistic assumption.

Importantly, the presence of heteroskedasticity *does not introduce bias* into the OLS slope estimates, so long as our core assumption $E(u_i \mid X_i) = 0$ holds. Recall that the unbiasedness of OLS requires zero conditional mean of errors, which is a separate assumption from homoskedasticity. Heteroskedasticity affects the *variance* and efficiency of the estimators, but not their expectation. In particular, OLS remains consistent for the true coefficients even if the errors have non-constant variance. However, as we will see, when errors are heteroskedastic the usual formulas for standard errors (which assume homoskedasticity) are no longer valid, and OLS is no longer the most efficient linear estimator.

6.1.1 Variance of OLS Estimators under Homoskedasticity

Under the classical linear model assumptions, including homoskedasticity, we have a well-known formula for the covariance matrix of the OLS estimator. Suppose our linear model (with intercept) is

$$Y = X\beta + u,$$

where X is the $N \times (k+1)$ design matrix (first column all 1's for the intercept, and k regressors). If $Var(u \mid X) = \sigma^2 I_N$ (homoskedasticity and no serial correlation), then the variance of the OLS estimator is

$$Var(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1}. \tag{6.1}$$

In particular, the variance of each coefficient $\hat{\beta}_j$ is given by the jth diagonal element of $\sigma^2(X'X)^{-1}$. This can be unpacked to see how various factors influence the precision of $\hat{\beta}_j$.

Theorem 6.2 (Variance of a Coefficient Estimate)**.** *Consider the jth regressor X_j in a multiple regression with an intercept and define R_j^2 as the R-squared from regressing X_j on all the* other *independent variables (and an intercept). Then under assumptions including homoskedasticity,*

$$Var(\hat{\beta}_j) \; = \; \frac{\sigma^2}{(1 - R_j^2) \sum_{i=1}^{N} (X_{ij} - \bar{X}_j)^2},$$

where $\bar{X}_j = \frac{1}{N} \sum_{i=1}^{N} X_{ij}$ is the sample mean of X_j.

Proof. The idea of the proof is to apply the *Frisch–Waugh–Lovell* theorem to isolate the contribution of X_j. First, regress X_j on all other regressors to obtain the fitted values $\hat{X}_{ij}$ and residuals $v_{ij} = X_{ij} - \hat{X}_{ij}$ for each observation i. By construction, these residuals $\{v_{ij}\}$ are orthogonal to the other regressors and have mean zero. Let $S^2_{v_j} = \sum_{i=1}^N v_{ij}^2$ denote the total variation of the residual v_{ij}. It can be shown that

$$S^2_{v_j} \;=\; (1 - R_j^2) \sum_{i=1}^{N} (X_{ij} - \bar{X}_j)^2 \,, \tag{6.2}$$

since R_j^2 is the fraction of X_j's sample variance explained by the other regressors. Equation (6.2) follows directly from the definition of R_j^2:

$$R_j^2 \;=\; 1 - \frac{\sum_{i=1}^N v_{ij}^2}{\sum_{i=1}^N (X_{ij} - \bar{X}_j)^2},$$

so rearranging gives $\sum_i v_{ij}^2 = (1 - R_j^2) \sum_i (X_{ij} - \bar{X}_j)^2$.

Now, by Frisch–Waugh–Lovell, the OLS estimate $\hat{\beta}_j$ can be obtained in a two-step: (i) purge Y of the influence of the other regressors (regress Y on the other X_{-j} to get residuals $\tilde{Y}_i$), and (ii) regress $\tilde{Y}$ on v_j (the residuals of X_j) in a simple regression. In this equivalent regression, the "x" variable is v_{ij} and the error term is the same original u_i (since the part of Y explained by other X has been removed). The slope estimate from this simple regression is $\hat{\beta}_j = \frac{\sum_i v_{ij} \tilde{Y}_i}{\sum_i v_{ij}^2}$. Because v_{ij} is uncorrelated with the other regressors, one can show $Var(\tilde{Y}_i \mid v_{ij}) = Var(u_i \mid X) = \sigma^2$ (homoskedasticity of u_i). Thus, the variance formula for a simple regression slope applies here, giving

$$Var(\hat{\beta}_j \mid X) = \frac{\sigma^2}{\sum_{i=1}^N v_{ij}^2} \,.$$

Finally, substitute the identity (6.2) for $\sum v_{ij}^2$ to obtain the stated result. □

The formula in Theorem 6.2 provides intuition about what influences the estimator's variance:

- The factor $\sum_i (X_{ij} - \bar{X}_j)^2$ is the total sample variation in regressor X_j. A larger spread or variance in X_j **reduces** $Var(\hat{\beta}_j)$. Intuitively, more variation in the independent variable helps us pin down its effect on Y more precisely. This is one reason larger sample sizes (which typically offer more variability in X as well as more observations) lead to more precise estimates.

- The error variance $\sigma^2 = Var(u_i)$ appears in the numerator: more intrinsic noise in Y (due to unobserved factors) makes it harder to precisely estimate β_j, thus increasing the variance of $\hat{\beta}_j$. If a lot of the variability in Y comes from factors unrelated to X_j (i.e., a large σ^2), then our estimate of β_j will naturally be more noisy. Including additional relevant variables that absorb some of this variation in Y (thereby reducing σ^2) can improve the precision of all coefficient estimates.

- The term $(1 - R_j^2)$ in the denominator shows the impact of multicollinearity. Here R_j^2 is how well X_j can be predicted by the other regressors. If X_j is nearly a linear combination of other X's (i.e., R_j^2 is high, close to 1), then $(1 - R_j^2)$ is small, making $Var(\hat{\beta}_j)$ large. In other words, when X_j is highly *collinear* with other covariates, it is difficult to disentangle the separate effect of X_j on Y, so the variance of $\hat{\beta}_j$ is inflated. This phenomenon is known as **variance inflation due to multicollinearity**. In the extreme case where X_j is an exact linear combination of other regressors ($R_j^2 = 1$), $Var(\hat{\beta}_j)$ would be infinite and the regression cannot be estimated (perfect multicollinearity causes the $X'X$ matrix to be singular).

The term $\frac{1}{1-R_j^2}$ is often called the **variance inflation factor (VIF)** for the jth variable. It measures how much the variance of $\hat{\beta}_j$ is multiplied due to the presence of other correlated regressors. A VIF of 1 means X_j is uncorrelated with others (no inflation of variance), while a very large VIF indicates that multicollinearity is a serious concern for that coefficient.

To summarize, we prefer to have high independent variation in each regressor and not too much collinearity between them, in order to minimize standard errors. Adding an *irrelevant regressor* (one that truly has no effect on Y) will not bias our estimates, but it can increase the R_j^2 for the other variables and thus inflate their variances. Therefore:

- Including unnecessary regressors (especially if they are correlated with other X's) tends to make it harder to find statistically significant effects, because the standard errors of the estimates increase.

- On the other hand, omitting relevant regressors can bias our estimates (violating $E(u \mid X) = 0$). So in model selection there is a trade-off: include enough variables to satisfy the zero conditional mean assumption, but avoid including extraneous ones that only add noise.

Example 6.3 (Irrelevant Regressor)**.** Suppose the true model is $Y = \beta_0 + \beta_1 X_1 + u$, so X_2 is irrelevant (its true coefficient is 0). We nonetheless estimate an expanded model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$. OLS will still recover an unbiased estimate of β_1. In fact, under $E(u \mid X_1, X_2) = 0$, one can show $E[\hat{\beta}_1] = \beta_1$ and $E[\hat{\beta}_2] = 0$. Thus the inclusion of an irrelevant X_2 does not *bias* $\hat{\beta}_1$. However, $\hat{\beta}_1$ will typically have a larger variance than it would in a regression without X_2. This follows from Theorem 6.2: here X_2 is uncorrelated with X_1 in expectation (since X_2 is irrelevant, there is no reason for it to be correlated with X_1 in the data-generating process), but in any given sample it may introduce some sample correlation (increasing R_1^2 slightly above 0). Moreover, X_2 adds another degree of freedom used, slightly reducing the effective sample variation in X_1 around its mean. Thus the standard error of $\hat{\beta}_1$ will generally rise. In practice, if $\hat{\beta}_2$ turns out insignificantly different from 0 (as expected) and including it does not dramatically change $\hat{\beta}_1$, one might drop X_2 in the final model to regain some precision.

6.1.2 Heteroskedasticity and Robust Standard Errors

We have established the OLS variance formula under homoskedastic errors. When the error variance is not constant (heteroskedasticity), OLS is still unbiased but Equation (6.1) no longer holds with $\sigma^2(X'X)^{-1}$. In fact, under heteroskedasticity the true variance of $\hat{\beta}$ (conditional on X) is

$$Var(\hat{\beta} \mid X) = (X'X)^{-1} \left(\sum_{i=1}^{N} Var(u_i \mid X)\, x_i x_i' \right) (X'X)^{-1},$$

which generally cannot be simplified to a scalar times $(X'X)^{-1}$. The usual OLS standard error formulas (which assume homoskedasticity) will misestimate the true sampling variability. In particular, the conventional formula $\hat{\sigma}^2(X'X)^{-1}$ will be *biased* if the errors are heteroskedastic.

If we proceed with inference (e.g., t-statistics) using the wrong standard errors, our hypothesis tests and confidence intervals may be incorrect. For example, the conventional t-statistic might not actually follow a t-distribution under the null, or the calculated p-value could be misleading (maybe too small, claiming significance where there is none, or vice versa). Thus, detecting and addressing heteroskedasticity is crucial for valid inference.

The good news is that we can correct the standard errors for heteroskedasticity without changing our coefficient estimates. The most common solution is to use **heteroskedasticity-robust standard errors** (also known as Eicker–Huber–White standard errors). These are sometimes simply called "robust" standard errors. The idea is to use an estimator of $Var(\hat{\beta})$ that does not assume $Var(u_i)$ is constant. One such estimator (for large samples) is:

$$\widehat{Var}(\hat{\beta}) \;=\; (X'X)^{-1}\Big(\sum_{i=1}^{N} \hat{u}_i^2\, x_i x_i'\Big)(X'X)^{-1}, \tag{6.3}$$

where $\hat{u}_i$ are the OLS residuals. Intuitively, we weight the outer product $x_i x_i'$ by $\hat{u}_i^2$ for each observation, allowing each observation to contribute a different amount to the variance (since a larger residual suggests that observation had a higher error variance). The square roots of the diagonal entries of (6.3) give us the robust standard errors for each coefficient. Under mild assumptions, (6.3) is a consistent estimator for the true variance of $\hat{\beta}$ even when the errors are heteroskedastic.

Remark 6.4. In most regression software (e.g., `R`, `Stata`, `Python statsmodels`), the default reported standard errors assume homoskedasticity. It is typically up to the user to request robust standard errors (for instance, by using a command or option like `robust` or `HC3`). If one has any doubt about the homoskedasticity assumption, it is safer to use robust standard errors. In practice, robust SEs often turn out to be slightly larger than homoskedastic SEs (reflecting the fact that OLS was not fully efficient). In some cases, the robust SEs may be much larger, altering the conclusions of a hypothesis test. On rare occasions, a robust SE might even be slightly smaller than the conventional one; this can happen if the pattern of heteroskedasticity is such that the OLS formula *over*-estimates variability. In any case, a prudent rule is to use the larger of the two estimates for inference. Robust SEs ensure valid hypothesis tests even when the form of heteroskedasticity is unknown.

In summary, heteroskedasticity by itself does *not* bias the OLS coefficients, but it does invalidate the default standard error formulas and thereby the usual t-tests and p-values. By using heteroskedasticity-robust standard errors, we can perform correct inference without having to explicitly model the variance of the errors.

6.1.3 Weighted Least Squares (WLS)

Rather than just adjusting the standard errors for heteroskedasticity, one might attempt to directly account for the unequal error variances in estimation. **Weighted Least Squares** is an estimation technique that can yield more efficient (i.e., lower-variance) estimates than OLS in the presence of heteroskedasticity, provided we know or can accurately model the variance function. The idea is to give each observation a weight inversely proportional to the variance of its error term. Intuitively, observations that are measured with less noise (lower variance) get more weight in fitting the regression line, since they contain more information about the true relationship.

Theorem 6.5 (Efficiency of Weighted Least Squares)**.** *Suppose the true variance of the error term is $Var(u_i \mid X_i = x_i) = \sigma^2\,\omega(x_i)$ for some known function $\omega(x_i) > 0$. Define weights $w_i = \frac{1}{\sqrt{\omega(x_i)}}$. Consider the transformed model*

$$w_iY_i \;=\; \beta_0 w_i \;+\; \beta_1(w_iX_{i1}) \;+\; \cdots \;+\; \beta_k(w_iX_{ik}) \;+\; w_iu_i.$$

If we apply OLS to this transformed model (i.e., minimize the weighted sum of squared residuals $\sum_i w_i^2(Y_i - \beta_0 - \sum_j \beta_j X_{ij})^2$), then the errors w_iu_i are homoskedastic with variance σ^2. In this scenario, the WLS estimator is the Best Linear Unbiased Estimator (BLUE), achieving the smallest variance among linear unbiased estimators of β.

In less formal terms, if we know the variance of each observation's error, we can improve efficiency by using WLS: we divide each observation by the standard deviation of its error, thereby equalizing the error variance across observations, and then run OLS on the rescaled data. This yields more precise estimates than unweighted OLS (which implicitly gives equal weight to all observations).

Example 6.6 (WLS in action)**.** Suppose in the investment example that larger firms (with higher Q) have more volatile investment, say $Var(u_i \mid Q_i) = \sigma^2 Q_i^2$. Here $\omega(Q_i) = Q_i^2$. If we know this, we can perform WLS by weighting each observation by $w_i = 1/Q_i$. The transformed regression would be

$$\frac{Y_i}{Q_i} = \alpha \left(\frac{1}{Q_i} \right) + \beta \left(\frac{Q_i}{Q_i} \right) + \frac{u_i}{Q_i}.$$

This simplifies to $\dfrac{Y_i}{Q_i} = \alpha \cdot \dfrac{1}{Q_i} + \beta + \tilde{u}_i$, where $Var(\tilde{u}_i) = Var(u_i/Q_i) = \sigma^2$ is now constant. In effect, we have stabilized the variance and can estimate β more precisely.

Remark 6.7. In practice, the true variance function $\omega(x_i)$ is usually not known. A **feasible** WLS (FWLS) procedure can be used: first run an OLS, obtain residuals $\hat{u}_i$, then model $Var(u_i)$ as a function of x_i (for instance, regress $\log(\hat{u}_i^2)$ on X to estimate how the variance changes with X). From this, generate predicted variances $\hat{\sigma}_i^2$ for each observation and use weights $w_i = 1/\sqrt{\hat{\sigma}_i^2}$ to re-run a weighted regression. This two-step estimator is consistent if the variance model is correctly specified. However, if the variance model is misspecified, WLS can perform poorly (even yielding biased estimates in finite samples). By contrast, OLS with robust standard errors remains consistent and will give valid (if perhaps slightly less efficient) inference regardless of the form of heteroskedasticity.

In many cases, the efficiency gains from WLS over OLS are modest, especially if the heteroskedasticity is not severe or the variance model is approximate. Unless we have strong prior knowledge of the variance structure, it is often recommended to simply use OLS with robust standard errors for inference. This approach avoids the risk of misspecifying the weights and keeps the interpretation of coefficients straightforward (the OLS estimates still represent the best linear approximation of the conditional expectation function). In summary, WLS is a useful tool when the variance function is known or reliably estimated, but otherwise one “should not bother” with WLS as a default, and rely on OLS with robust inference instead.

6.2 Hypothesis Testing in Linear Regression

We often hear the phrase: "the estimate $\hat{\beta}$ is statistically significant." What does this mean in the context of regression analysis? It means that based on a hypothesis test, we have evidence that the true coefficient β is different from zero (or some other benchmark value) at a given significance level. Hypothesis testing allows us to quantify the uncertainty in our estimates and make probabilistic statements such as "we are 95% confident that β exceeds zero."

In this section, we focus on testing a single coefficient, typically the null hypothesis $H_0 : \beta_j = 0$ against the alternative $H_1 : \beta_j \neq 0$. This is the test underlying the claim of "statistical significance" for $\hat{\beta}_j$. We will discuss test statistics, p-values, and the distinction between statistical and economic significance.

6.2.1 The t-Test for a Single Coefficient

Because OLS estimates are random variables (functions of the random sample), we can construct test statistics to evaluate hypotheses about the true coefficients. Under the classical assumptions (including that either the errors u_i are normally distributed, or the sample size is large enough to apply asymptotic normality), the OLS estimator $\hat{\beta}_j$ is approximately normally distributed around β_j. We can standardize this estimator by subtracting the null hypothesis value and dividing by its standard error:

$$t_j = \frac{\hat{\beta}_j - \beta_{j,0}}{\widehat{SE}(\hat{\beta}_j)}.$$

Typically, for testing significance we take $\beta_{j,0} = 0$, so the t-statistic simplifies to $t_j = \hat{\beta}_j / \widehat{SE}(\hat{\beta}_j)$. This statistic measures how many standard deviations away from zero our estimate is.

If the null hypothesis is true ($\beta_j = 0$) and classical assumptions hold, this t_j statistic follows a t-distribution with $N-(k+1)$ degrees of freedom (in finite samples with normal errors) or approximately a standard normal distribution (for large N, by the Central Limit Theorem). We can then conduct a test by comparing $|t_j|$ to critical values of the t (or normal) distribution. For example, at the 5% significance level in a two-sided test, the critical value is about 1.96 (for large N), meaning if $|t_j| > 1.96$ we reject $H_0 : \beta_j = 0$. Equivalently, one can compute the ***p*-value** of the test, which is more informative.

Definition 6.8. The ***p*-value** for testing a given null hypothesis is the probability, under the assumption that the null is true, of obtaining a test statistic as extreme as (or more extreme than) the one actually computed from the sample. In a two-sided test for $H_0 : \beta_j = 0$, the p-value is $\Pr(|T| \geq |t_j|)$ where T denotes the test statistic under the null distribution.

If the p-value is below the chosen significance level α (say 0.05), we reject the null hypothesis. For instance, a p-value of 0.03 indicates that if β_j were truly zero, there is only a 3% chance we would observe an estimate as far from zero as $\hat{\beta}_j$ purely due to random sampling variation. Such a low probability leads us to conclude that β_j is likely not zero (i.e., the estimate is *statistically significant at the 5% level*).

On regression output tables, one often sees the t-statistic and p-value reported for each coefficient. The phrase "$\hat{\beta}_j$ is statistically significant" typically implies that $H_0 : \beta_j = 0$ was rejected at some conventional level (often 5%). Many tables also mark coefficients with stars to denote significance: for example, *** for $p < 0.01$, ** for $p < 0.05$, * for $p < 0.1$.

It is important to remember that the standard error $\widehat{SE}(\hat{\beta}_j)$ used in the t-statistic should be the appropriate one given our assumptions. If we have heteroskedasticity, we must use the robust standard error in computing t_j; otherwise, the test will not be valid. Most statistical software can report robust t-statistics when requested.

6.2.2 Statistical vs. Economic Significance

Rejecting $H_0 : \beta_j = 0$ tells us that we have evidence $\beta_j \neq 0$, but it does not by itself tell us whether the effect is *large* or important in a practical sense. We must distinguish between **statistical significance** and **economic (or substantive) significance**:

- A coefficient can be statistically significant (different from zero in a precise sense) but economically trivial in magnitude.

- Conversely, a coefficient can be economically large in magnitude but not statistically significant (often due to a small sample or high noise making the estimate imprecise).

Statistical significance relates to our confidence that an effect is non-zero. Economic significance relates to the size of the effect in real-world terms. Always examine the magnitude of $\hat{\beta}_j$ and consider its context. For example, with a very large sample, one might find that a policy intervention has an effect on income that is statistically different from zero but extremely small (say, \$1 increase in annual income on a base of \$50,000). Such an effect, while "statistically significant," is economically negligible. On the other hand, if we have an estimate that increasing education by one year raises wages by 10%, that is economically meaningful; but if our sample is tiny, the estimate might come with a huge standard error, and we cannot be statistically sure that the true effect isn't zero.

Example 6.9. Imagine a regression of a country's GDP growth on the number of new libraries built. Suppose we obtain $\hat{\beta} = 0.0005$ with a standard error of 0.0001 from a very large dataset, yielding a t-statistic of 5. This is statistically significant at 1% level ($p < 0.01$). However, the magnitude suggests that even building 100 new libraries would increase GDP growth by only 0.05%—an economically tiny effect. Now consider a different study where $\hat{\beta} = 0.5$ (meaning a very large effect of libraries on growth), but the standard error is 0.4 because the sample size is small. This gives a t-statistic of 1.25, which is not statistically significant at conventional levels ($p \approx 0.22$). We cannot rule out that the true effect is zero, even though the point estimate is large. In the first case, we have statistical but not economic significance; in the second, possibly economic significance but not statistical significance (due to imprecision).

Always report and discuss both the magnitude and the significance of key estimates. A good practice is to translate the coefficient into a more interpretable effect size. For instance, if $\hat{\beta}_1$ corresponds to the effect of an additional year of education on earnings, one could compute: "according to our estimate, an increase of one standard deviation in education (about 2.5 years) is associated with a \$5,000 increase in annual income." Then judge if \$5,000 is a large or small change relative to typical incomes, and discuss whether that is a meaningful impact economically. If an effect seems implausibly large or small, it may signal model misspecification or data issues.

In summary, statistical significance addresses the question “is the effect likely non-zero?” whereas economic significance asks “how large is the effect, and does it matter in context?”. Both questions are important for a complete analysis.

6.3 Miscellaneous Topics in OLS Regression

We now address a collection of additional issues in linear regression modeling: multicollinearity, use of binary (dummy) variables, interaction terms, and proper presentation of regression results.

6.3.1 Multicollinearity and Irrelevant Regressors

As discussed earlier, **multicollinearity** refers to the presence of high correlation or linear relationships among the regressors. Perfect multicollinearity (an exact linear dependence) prevents OLS estimation altogether, whereas imperfect multicollinearity (high but not perfect correlation among X's) can lead to large standard errors for the affected coefficients.

It is important to note that multicollinearity *does not bias* the OLS estimates; it only affects their variance. If two regressors X_2 and X_3 are highly collinear, we can still interpret the regression as long as an exact linear relationship does not exist. However, $\hat{\beta}_2$ and $\hat{\beta}_3$ will individually be very imprecise (large standard errors) because the data does not contain enough independent variation in X_2 versus X_3 to precisely attribute effects to each one. In contrast, the combination of them might be estimated more precisely (for example, $\beta_2 + \beta_3$ could be well-identified even if β_2 and β_3 separately are not).

Example 6.10 (Effect of Correlated Regressors)**.** Consider a model

$$Y \ = \ \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u,$$

and suppose X_2 and X_3 are highly correlated with each other (say, $X_3 \approx X_2$ for most observations). This might happen if, for instance, X_2 is years of education and X_3 is age at which a person started working—these tend to move together (more education implies entering the workforce later). In this case, $\hat{\beta}_2$ and $\hat{\beta}_3$ will each have a large variance (as reflected by big standard errors) because it is difficult for the regression to distinguish the separate impacts of X_2 and X_3. However, this multicollinearity does *not* affect $\hat{\beta}_1$'s variance directly, as long as X_1 is not collinear with X_2 or X_3. In fact, if X_1 is orthogonal to the other two ($R_1^2 = 0$ in regressing X_1 on X_2, X_3), then $Var(\hat{\beta}_1)$ is unchanged by whether X_2 and X_3 are in the model or not.

The key takeaways regarding multicollinearity are:

- It does not cause bias or inconsistency in $\hat{\beta}$. The OLS estimates are still centered on the true values (assuming exogeneity holds).
- It does inflate the standard errors of the affected coefficients, making it harder to find those coefficients statistically significant. You might have an important variable that shows up as insignificant simply because it moves in tandem with another variable.
- If two variables essentially measure the same thing, it's often unnecessary to include both. Avoid redundant controls that are highly correlated with your variable of interest unless they are needed to satisfy $E(u|X) = 0$. Include control variables that are necessary for validity, but be mindful that adding controls which are not essential can increase multicollinearity and thus reduce precision.
- The remedy for multicollinearity is usually *more data*. With a larger sample, even highly correlated variables can eventually be estimated precisely because you are more likely to find variation that separates them. In the short run, if multicollinearity is severe, you may consider dropping one of the collinear variables (especially if it is theoretically less central) to gain precision on the other.

6.3.2 Dummy Variables and Interaction Terms

Many of our regressors of interest are not continuous measurements but rather categorical or binary characteristics. A **dummy variable** (also called an indicator or binary variable) is a variable that takes the value 0 or 1 to indicate the absence or presence of a particular attribute. For example, we might have a dummy variable $Female_i$ which is 1 if individual i is female and 0 if male. Dummy variables allow us to model qualitative differences between groups in a regression framework.

Definition 6.11. A **dummy variable** (indicator) D is a variable that equals 1 if a certain condition is met or if an observation belongs to a certain category, and 0 otherwise. It is used in regression to shift intercepts or interact to allow different effects for different groups.

Models with a Single Dummy Variable

Including a dummy variable as a regressor allows for a shift in the intercept for the group indicated by the dummy. For instance:

$$\text{Wage}_i \;=\; \beta_0 \;+\; \delta_0\, Female_i \;+\; \beta_1\, Educ_i \;+\; u_i.$$

In this wage regression, $Female_i$ is the dummy (1 for women, 0 for men). The coefficient δ_0 measures the difference in wages between females and males, holding education constant. To see this, consider the expected wage for men versus women:

$$\begin{aligned} E[\text{Wage} \mid Female = 0,\, Educ] &= \beta_0 + \beta_1\, Educ, \\ E[\text{Wage} \mid Female = 1,\, Educ] &= \beta_0 + \delta_0 + \beta_1\, Educ. \end{aligned}$$

The difference is $E[\text{Wage} \mid Female = 1] - E[\text{Wage} \mid Female = 0] = \delta_0$ for the same education level. Thus δ_0 is the gender wage gap (female minus male) at a given education level. The intercept for males is β_0 (the baseline group when the dummy is 0), and the intercept for females is $\beta_0 + \delta_0$. In this specification, the only effect of gender is a parallel shift in the wage equation (an intercept shift); the slope on education β_1 is assumed to be the same for men and women.

Example 6.12 (Gender Wage Gap)**.** Suppose we estimate:

$$\widehat{\text{Wage}} \;=\; -1.57 \;-\; 1.80\, Female \;+\; 0.57\, Educ \;+\; 0.03\, Exp \;+\; 0.14\, Tenure.$$

All else equal, the coefficient on $Female$ is $\hat{\delta}_0 = -1.80$, indicating that women earn \$1.80 less *per hour* than men with the same education, experience, and tenure. Here the unit of the dependent variable is in dollars per hour (assuming Wage is measured that way), so the difference is \$1.80/hour. The intercept $\beta_0 = -1.57$ is the predicted wage for a male with zero education, zero experience, and zero tenure. That intercept is not economically meaningful in this context (a negative wage is impossible and nobody has zero years of education in the sample), reminding us that we should not over-interpret the intercept. The key result is the gap of \$1.80/hour attributable to gender after controlling for other factors.

If the dependent variable is in logarithms, dummy variable coefficients can be interpreted in percentage terms (approximately). For example:

$$\ln(\text{Price}) \;=\; \beta_0 + 0.054\,Colonial + 0.17\ln(\text{LotSize}) + 0.71\ln(\text{SqFt}) + \cdots,$$

where $Colonial$ is a dummy for a house being of colonial style. The coefficient 0.054 on $Colonial$ suggests a colonial house is associated with about 5.4% higher price compared to a non-colonial house with similar lot size, square footage, etc. (To be precise, one could compute $e^{0.054} - 1 \approx 5.55\%$ increase.)

Multiple Categories and the Dummy Variable Trap

If a categorical variable has more than two categories, we represent it with multiple dummy variables. For example, marital status could have categories {single, married, divorced, widowed}. We would introduce three dummy variables (since one category will serve as the baseline). Generally, for a categorical variable with M categories, we include $M-1$ dummy variables in the regression in addition to the intercept. Including all M would cause perfect multicollinearity (the sum of all dummies would equal 1 for each observation, duplicating the intercept). This is known as the **dummy variable trap**. Always omit one category as the reference group.

Which category to omit is arbitrary for model fit, but it changes the interpretation of coefficients. Each included dummy's coefficient measures the effect relative to the omitted baseline category.

Example 6.13 (Multiple Dummy Variables)**.** Suppose we want to estimate wage differences by gender and marital status. There are 4 groups: {single men, married men, single women, married women}. We can set single men as the baseline. Then include dummies for married men, single women, and married women:

$$\begin{aligned}\ln(\text{Wage}) \;=\; \beta_0 \;+\; \delta_1\,(\text{MarriedMale}) \;+\; \delta_2\,(\text{SingleFemale}) \\ +\; \delta_3\,(\text{MarriedFemale}) \;+\; \beta_1\,Educ \;+\; u.\end{aligned}$$

Here,

- β_0 = intercept for single men (baseline group).
- δ_1 = difference in log-wage between married men and single men (with same education).
- δ_2 = difference between single women and single men.
- δ_3 = difference between married women and single men.

If the estimation yields (standard errors omitted for brevity):

$$\begin{aligned}\ln(\text{Wage}) \;=\; 0.30 \;+\; 0.21\,\text{MarriedMale} \;-\; 0.11\,\text{SingleFemale} \\ -\; 0.20\,\text{MarriedFemale} \;+\; 0.08\,Educ,\end{aligned}$$

we interpret these as:

- Single men (baseline): intercept $\exp(0.30) \approx 1.35$, meaning baseline average wage \$1.35 (this number itself isn't meaningful, since no one in the sample likely has 0 education, but it anchors the other comparisons).
- Married men earn about 21% more than single men, ceteris paribus (since 0.21 in logs is approximately 0.21 in percentage).
- Single women earn about 11% less than single men, ceteris paribus.
- Married women earn about 20% less than single men, ceteris paribus.

We can also compare married women to married men: add δ_2 and δ_3 vs δ_1. For example, married women are about $0.21 - 0.20 - 0.11 = -0.10$ (10%) lower than married men (this kind of comparison can be done after estimation).

Whether we include separate dummy variables for each combination (as above) or use an interaction of simpler dummies (as we will see next) the fitted values and overall model fit will be identical. It's purely a matter of parameterization. The golden rule is to avoid including a full set of dummies for all categories along with an intercept; always omit or leave out one category as the reference.

If you accidentally include all categories (e.g., a dummy for each marital/gender combination plus an intercept), most software will automatically drop one for you to resolve the multicollinearity. It's then up to you to recognize which was dropped and interpret accordingly.

Interactions Between Dummy Variables

We can achieve the same model as the above example using fewer dummy variables and an interaction. For instance, we could define two main dummies: $Female$ (1 for women, 0 for men) and $Married$ (1 for married, 0 for single). Then include both and their interaction:

$$\ln(\text{Wage}) \;=\; \beta_0 + \beta_1\, Married + \beta_2\, Female + \beta_3\,(Married{\times}Female) + \beta_4\, Educ + u.$$

Let's interpret the coefficients:

- Baseline (when $Female = 0, Married = 0$) is single men: intercept β_0.
- β_1: effect of being married if male ($Female = 0$). So married men vs single men difference.
- β_2: effect of being female if single ($Married = 0$). So single women vs single men difference.
- β_3: the interaction term captures the *additional* effect of being female *and* married, beyond the sum of being female + being married. In other words, β_3 adjusts the wage for married women specifically.

For a married female, the log-wage would be:

$$\beta_0 + \beta_1(1) + \beta_2(1) + \beta_3(1 \cdot 1) + \beta_4\, Educ \;=\; \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4\, Educ.$$

For a married male: $\beta_0 + \beta_1$. For a single female: $\beta_0 + \beta_2$. For a single male: β_0. Thus:

- Difference married male vs single male = β_1.
- Difference single female vs single male = β_2.
- Difference married female vs single male = $\beta_1 + \beta_2 + \beta_3$.
- Difference married female vs married male = $(\beta_2 + \beta_3)$.

To connect with the previous example, if we estimate the interaction model we might get:

$$\begin{aligned}\ln(\text{Wage}) \;=\; 0.30 \;+\; 0.21\,Married \;-\; 0.11\,Female \;-\; 0.30\,(Married \times Female)\\ +\; 0.08\,Educ.\end{aligned}$$

For a married female, her log-wage relative to a single male baseline is $0.21 - 0.11 - 0.30 = -0.20$, which is a 20% deficit. Married female vs married male: $-0.11 - 0.30 = -0.41$. This matches the alternative parameterization above. Thus the interaction specification yields the same predictions, just distributed differently among the coefficients:

- $\beta_1 = +0.21$ (main effect of Married, effectively the married male premium).
- $\beta_2 = -0.11$ (main effect of Female, the single female gap).
- $\beta_3 = -0.30$ (interaction: additional penalty for being a married female on top of being female and being married).

The key point is that you can either create a dummy for each category or use main dummies and interactions; the model flexibly captures different intercepts for each group either way. The choice of parameterization does not affect the fitted values or predictions, only how the coefficients are presented.

Example 6.14 (Using Dummy Interactions: Computer Use and Wages)**.** In a famous study, Krueger (1993) examined the return to computer use on wages. One of his regressions included two dummies: *ComputerWork* (1 if the person uses a computer at work) and *ComputerHome* (1 if the person uses a computer at home), as well as their interaction. A simplified result was:

$$\begin{aligned}\ln(\text{Wage}) \;=\; \beta_0 \;&+\; 0.18\,ComputerWork \;+\; 0.07\,ComputerHome \\ &+\; 0.02\,(ComputerWork \times ComputerHome) \;+\; \cdots\end{aligned}$$

The omitted category here is people who use no computer at all. From these coefficients:

- Using a computer at work is associated with an $\approx$ 18% higher wage (for those who don't use a computer at home).
- Using a computer at home (but not at work) is associated with $\approx$ 7% higher wage.
- The interaction term 0.02 suggests that using a computer both at work and at home has an extra 2% premium on top of the individual effects. So a person who uses computers both at work and home would have roughly $0.18 + 0.07 + 0.02 = 0.27$ (27%) higher wages than someone using no computer.

If we convert to exact percentage: a coefficient of 0.27 in logs means about $e^{0.27} - 1 \approx 31\%$ higher wages. This reflects a slightly higher combined effect than the sum (due to the compounding in logs). The interaction being positive (0.02) means the two forms of computer use complement each other slightly in their association with wages.

Interactions between Continuous Variables

An **interaction term** need not involve only dummies; we can also include interactions between continuous variables, or between a continuous and a dummy. An interaction allows the effect of one regressor to depend on the level of another regressor.

For example, consider a model with an interaction between two continuous variables X_1 and X_2:

$$Y \;=\; \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + u.$$

This specification implies that the partial effect of X_1 on Y is not constant but varies with X_2. Specifically, holding X_2 constant,

$$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 X_2.$$

If $\beta_3 \neq 0$, the slope on X_1 depends on the value of X_2.

Interpretation:

- If $\beta_3 > 0$, then X_1's effect on Y is larger (steeper) when X_2 is larger.
- If $\beta_3 < 0$, then X_1's effect on Y is smaller (flatter) when X_2 is larger. In other words, X_2 "dampens" the influence of X_1 on Y.

For a concrete scenario, suppose X_1 is years of education and X_2 is years of job experience, and Y is log-wage. If we find $\beta_1 > 0$ (education raises wages) but $\beta_3 < 0$ (negative interaction), it could mean additional education yields a smaller percentage wage increase for those with more experience. Perhaps early in one's career, an extra year of education has a big payoff, but later in the career (with high experience), the marginal effect of another degree is less.

One must be careful in interpreting the coefficients in the presence of interactions:

- β_1 is the partial effect of X_1 when $X_2 = 0$ (since $\partial Y/\partial X_1 = \beta_1 + \beta_3 X_2$ and plugging $X_2 = 0$ gives β_1). Thus, β_1 is the slope on X_1 for the special case when $X_2 = 0$. If $X_2 = 0$ is outside the range or not meaningful, then β_1 alone may not be directly relevant. Similarly, β_2 is the effect of X_2 when $X_1 = 0$.
- β_3 itself indicates how much the slope on X_1 changes when X_2 increases by 1 (or vice versa, how X_2's slope changes with X_1).

Example 6.15 (Interactions Between Continuous Variables)**.** Consider firms where X_1 is capital investment and X_2 is market size, and Y is profit. Theory might suggest diminishing returns to investment in larger markets. An interaction model

$$Profit = \beta_0 + \beta_1 Investment + \beta_2 MarketSize + \beta_3 (Investment \times MarketSize) + u$$

with $\beta_1 > 0$ and $\beta_3 < 0$ would mean investment boosts profit, but the boost is less in big markets (X_2 large) than in small markets. For a given high $MarketSize$, the marginal profit from an extra unit of investment ($\partial Profit/\partial Investment$) could be quite small or even zero if β_3 is sufficiently negative. If the average market size $\bar{X}_2$ is such that $\beta_1 + \beta_3\bar{X}_2$ is still positive, then at the average market size the effect of Investment is positive. However, if one naively looked at β_1 alone, one might misstate the average effect if $\bar{X}_2$ is not 0.

A common mistake in interpretation is to quote β_1 as "the effect of X_1" without conditioning on X_2. If $X_2 = 0$ is not a typical scenario, β_1 is not very meaningful by itself. To get the effect of X_1 for a more relevant scenario (say, the average value of X_2), one can plug that in: effect at $X_2 = \bar{x}_2$ is $\beta_1 + \beta_3\bar{x}_2$. Alternatively, one can **reparameterize** the model by centering the variables.

Centering for Interpretation. If we subtract the mean from each variable (sometimes called mean-centering or demeaning), the interaction coefficient remains the same but the main effects become the effect at the mean of the other variable. For example, define $\tilde{X}_1 = X_1 - \mu_{X_1}$ and $\tilde{X}_2 = X_2 - \mu_{X_2}$, where μ_{X_1} and μ_{X_2} are sample means. Now run:

$$Y \;=\; \delta_0 + \delta_1\tilde{X}_1 + \delta_2\tilde{X}_2 + \delta_3(\tilde{X}_1 \times \tilde{X}_2) + u.$$

In this specification,

$$\frac{\partial Y}{\partial X_1} \;=\; \delta_1 + \delta_3\tilde{X}_2,$$

so at $\tilde{X}_2 = 0$ (which corresponds to $X_2 = \mu_{X_2}$, the average value of X_2),

$$\left.\frac{\partial Y}{\partial X_1}\right|_{X_2=\mu_{X_2}} \;=\; \delta_1.$$

Thus δ_1 now represents the effect of X_1 at the *average level of* X_2. Similarly, δ_2 would be the effect of X_2 at the average level of X_1. The interaction term's coefficient δ_3 will be identical to β_3 from before (centering does not affect the interaction slope, only the interpretation of main effects and the intercept).

Example 6.16 (Centering Example)**.** Returning to the education and experience on wages example:

$$\ln(\text{Wage}) \;=\; 0.39 - 0.23\,Female + 0.08\,Educ - 0.01\,(Female{\times}Educ) + u.$$

In this model, $\hat{\beta}_{\text{Female}} = -0.23$ indicates that, at $Educ = 0$, women earn 23% less than men. Of course, nobody has zero years of education in the data (assume all have at least, say, 8 years of schooling). If the average education in the sample is $\mu_{Educ} = 12$ years, a more relevant comparison is at 12 years of schooling: the female effect at 12 years is $-0.23+(-0.01)\times 12 = -0.35$. So at the mean education, women earn about 35% less than men. If we center education at 12, the coefficient on Female will directly give -0.35. Indeed, if we define $\widetilde{Educ} = Educ - 12$ and regress

$$\ln(\text{Wage}) \;=\; \tilde{\beta}_0 \;+\; \tilde{\delta}_0\,Female \;+\; \tilde{\beta}_1\,\widetilde{Educ} \;+\; \tilde{\delta}_1\,(Female \times \widetilde{Educ}) \;+\; u,$$

then $\tilde{\delta}_0$ will equal -0.35 (the gender gap at 12 years of education). The interaction coefficient $\tilde{\delta}_1$ remains -0.01. The intercept also changes accordingly. The slopes on $\widetilde{Educ}$ (for men) and the additional slope for women are unchanged by centering.

The main lesson: when using interactions, one must interpret coefficients carefully. Non-interacted coefficients (like β_1, β_2) usually represent effects at the zero value of the other variable, which might not be of interest. By centering variables, you can make those coefficients represent effects at more meaningful baseline levels (often the mean).

One must also be cautious about extrapolating interaction effects beyond the data range. In the earlier scenario with $\beta_1 > 0, \beta_3 > 0$ (women have lower starting wage but higher return to education), one might find a crossing point where women's predicted wages overtake men's at very high education levels. If that crossing point (solve $\beta_0 + \beta_1 X_1 + \beta_2(\text{female} = 1) + \beta_3 X_1(\text{female} = 1) = \beta_0 + \beta_1 X_1$ to find when female = male wage) occurs at, say, 25 years of education, but the maximum education in the data is 20 years, then within-sample women never actually catch up. You should not claim "with enough education, women earn more than men" unless that regime is supported by the data. Always check whether interaction-driven crossings occur within the support of your data. The crossing point for when two regression lines intersect (for example, men's and women's wage lines) can be found by setting the predicted outcomes equal and solving: in the example,

$$\beta_0 + \beta_2 + (\beta_1 + \beta_3)\,X \;=\; \beta_0 + \beta_1 X,$$

implying $X = \frac{-\beta_2}{\beta_3}$ (if $\beta_3 \neq 0$). If $\frac{-\beta_2}{\beta_3}$ is outside the sample range of X, the lines cross only out of sample.

Interactions Involving Dummy Variables and Continuous Variables

Combining the two ideas, we often interact dummy variables with continuous variables to allow different slopes for different groups. For example, to allow males and females to have different returns to education, we include an interaction $Female \times Educ$ in the wage model:

$$\ln(\text{Wage}) \;=\; \beta_0 + \delta_0\, Female + \beta_1\, Educ + \delta_1\,(Female \times Educ) + u.$$

Now we have:

- Intercept (for males): β_0.
- Intercept for females: $\beta_0 + \delta_0$ (when $Female = 1$).
- Slope on education for males: β_1.
- Slope on education for females: $\beta_1 + \delta_1$.

So δ_0 represents the gender gap when $Educ = 0$ (again, careful if 0 is not realistic), and δ_1 represents how much the female education slope differs from the male education slope. If $\delta_1 < 0$, women have a lower return to education than men.

We can visualize such a model as two lines on a wage vs. education plot: one line for men and one for women. If $\delta_0 < 0$ and $\delta_1 < 0$, the female line starts lower and has a flatter slope (so is always below the male line). If $\delta_0 < 0$ but $\delta_1 > 0$, the female line starts lower but is steeper—so it might eventually catch up or cross the male line. However, as cautioned, check whether the crossing happens beyond the data.

If we find a crossing within the data range, it means at some level of education, predicted female wages equal male wages, and beyond that, female wages would exceed male wages (according to the model). If this result is trusted, one could say "for education beyond X years, women are predicted to earn more than men, reversing the wage gap." If the crossing is out-of-sample, we should refrain from such a statement, as it would be extrapolation.

Again, to interpret δ_0 as the gender gap at a meaningful education level, one can center education at its mean. Often researchers will report the gender gap at the average education or at some relevant value, which can be obtained by centering or by manually plugging values into the equation.

The general advice is: **if you want the coefficient on a non-interacted dummy to be interpretable as the difference at the "average" value of the other covariates, then center those covariates around their means before interacting.** This will not change the fit or the interaction coefficient, but will yield more interpretable intercept and main effects.

Ordinal Independent Variables

A brief digression: sometimes we encounter an independent variable that is ordinal, meaning it represents a ranking or level, but differences between levels are not necessarily equal. An example is credit rating: AAA, AA, A, BBB, BB, B, etc. If we naively assign numbers (AAA=1, AA=2, A=3, etc.) and include this as a numeric regressor, we are imposing a linear structure: we assume the difference between AAA and AA is the same as between BBB and BB, etc. This might be a strong assumption if those rating notches have nonlinear effects on, say, interest rates.

A safer approach is to treat an ordinal variable as categorical and use dummy variables for each category (minus one). For the credit rating example, create dummies $D(\text{AAA}), D(\text{AA}), \ldots, D(\text{B})$. Choose one category (perhaps the lowest rating, or highest) as the baseline. The regression might look like:

$$IR = \beta_0+\gamma_{\text{AAA}}D(\text{AAA})+\gamma_{\text{AA}}D(\text{AA})+\gamma_{\text{A}}D(\text{A})+\cdots+\gamma_{\text{BB}}D(\text{BB})+\beta_1X_{\text{other}}+u,$$

omitting $D(\text{D})$ or whichever lowest category is chosen as the reference group. Each γ_{Rating} then measures how much higher (or lower) the interest rate is for that rating compared to the omitted category. This allows for a nonlinear pattern: maybe the jump from BBB to BB has a much bigger effect on interest rates than the jump from AA to A, etc., which a linear coding would not capture.

The trade-off is that using indicators for each category uses more degrees of freedom and yields many coefficients, but it avoids a potentially incorrect linearity assumption. If the ordinal variable truly has a linear effect, the dummy method will still capture it (the γ's might form a roughly linear pattern). If not, the dummy method is more flexible and thus preferred to avoid bias.

6.3.3 Presenting and Reporting Regression Results

Finally, a few words on effectively reporting regression results, especially in academic writing (like a paper or thesis).

Typically, regression results are presented in a table format. A well-constructed regression table should include:

- **Clear column labels** indicating the dependent variable for each regression (or each column can represent a different model specification with possibly the same dependent variable). If all columns share the same dependent variable, indicate it clearly in the table title or heading.

- **Row labels** for each independent variable included. Use descriptive names so the reader knows what each variable is (e.g., use "Female" rather than a generic "D2").

- **Coefficient estimates** with their standard errors (or t-statistics) indicated below in parentheses or ± notation. It is common to report something like: $\hat{\beta} = 0.105$ (0.032), where 0.032 is the standard error. Alternatively, some use symbols to denote significance and place standard errors in the table note.

- **Statistical significance indicators** (stars or bolding) for quick visual reference of which coefficients are significant at conventional levels.

- R^2 **(and possibly adjusted** R^2**)** for each regression, to indicate goodness-of-fit.

- **Number of observations** used in each regression (N).

- Optionally, other model diagnostics or fixed effects dummies, etc., can be noted (often at the bottom of the table, one might say "Year dummies: Yes" or "Estimation: OLS").

An example stub of a regression table could be:

	(1)	**(2)**	**(3)**
Dependent variable:	Wage (or ln(Wage) in col. (3))		
Education (years)	0.57	0.50	0.08
	(0.10)	(0.12)	(0.01)
Experience (years)	0.03^{*}	0.02	
	(0.01)	(0.01)	
Tenure (years)	0.14	0.10	
	(0.05)	(0.06)	
Female (dummy)	−1.80	−1.50	-0.23^{*}
	(0.60)	(0.80)	(0.08)
Female × Educ			−0.01
			(0.005)
Constant	−1.57	−0.50	0.39
	(1.00)	(1.10)	(0.20)
Observations	500	500	500
R^2	0.25	0.27	0.30

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors in parentheses.

(This is just a fabricated example for illustration.)

In the text of your paper or report, you should not simply restate all the numbers in the table. Instead, focus on the key results:

- Highlight the coefficients of interest (the ones related to your main hypotheses). Comment on their sign (positive/negative), magnitude, and statistical significance.
- Interpret the economic meaning: e.g., "We find that the coefficient on *Educ* is 0.08 (column 3), implying that each additional year of education is associated with an 8% higher wage, significant at the 1% level."
- For dummy variable effects: e.g., "The coefficient on Female in column 3 is -0.23, indicating that at the average education level, women earn

about 23% less than men, holding other factors constant. This gender gap is statistically significant at the 5% level."

- Discuss any surprises or notable points, such as significant interactions or unexpected signs.
- Do not waste text describing uninteresting coefficients unless needed for completeness. Control variables can usually be summarized briefly.
- Mention N and R^2 only if relevant. They are secondary to the coefficients but can be important for context or comparison across specifications.

Ensure that every regression table you include is discussed in the text. If a table presents multiple specifications, guide the reader through them: "Column 1 presents the baseline specification without interactions; column 2 adds additional controls; column 3 introduces the interaction term. We see that the coefficient on Female becomes more negative once we add the interaction, suggesting that part of the gender gap is associated with different returns to education." If a table or regression is not important enough to talk about, consider omitting it.

In summary, when reporting regression results: present them clearly in tables with all necessary information, and interpret them in writing focusing on what matters for your argument. Discuss both statistical significance and the practical significance of the findings. Avoid simply listing numbers; instead, tell the story of what the numbers mean.

With that, we conclude this chapter on linear regression extensions. We covered the implications of heteroskedasticity and how to handle it, the concept of hypothesis testing in regression, the difference between statistical and economic significance, issues of multicollinearity and irrelevant regressors, and how to use and interpret interaction terms and dummy variables. These tools and insights will be invaluable as we delve deeper into causal inference and more complex regression models.