Nonlinear Regression Functions [1]

Jasmine(Yu) Hao

Faculty of Business and Economics
Hong Kong University

August 11, 2021

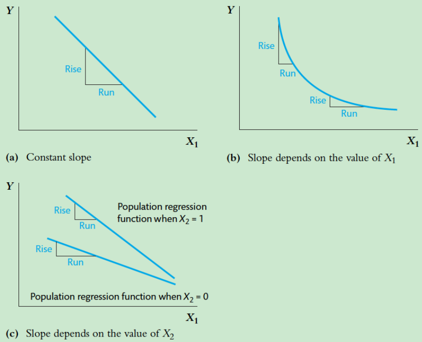
[1] This section is based on Stock and Watson (2020), Chapter 8.

Nonlinear model I

- This chapter develops two groups of methods for detecting and modeling nonlinear population regression functions.
- The methods in the first group are useful when the relationship between Y and an independent variable, X_1, depends on the value of X_1 itself.
- For example, reducing class sizes by one student per teacher might have a greater effect if class sizes are already manageably small than if they are so large that the teacher can do little more than keep the class under control.
- If so, the test score (Y) is a nonlinear function of the student–teacher ratio (X_1), where this function is steeper when X_1 is small.

Nonlinear model II

FIGURE 8.1 Population Regression Functions with Different Slopes



In Figure 8.1(a), the population regression function has a constant slope. In Figure 8.1(b), the slope of the population regression function depends on the value of X_1. In Figure 8.1(c), the slope of the population regression function depends on the value of X_2.

Nonlinear model III

- Whereas the linear population regression function in Figure 8.1(a) has a constant slope, the nonlinear population regression function in Figure 8.1(b) has a steeper slope when X_1 is small than when it is large.
- The methods in the second group are useful when the effect on Y of a change in X_1 depends on the value of another independent variable—say, X_2.
- For example, students still learning English might especially benefit from having more one-on-one attention; if so, the effect on test scores of reducing the student–teacher ratio will be greater in districts with many students still learning English than in districts with few English learners. In this example, the effect on test scores (Y) of a reduction in the student–teacher ratio (X_1) depends on the percentage of English learners in the district (X_2).

Nonlinear model IV

- As shown in Figure 8.1(c), the slope of this type of population regression function depends on the value of X_2. Although they are nonlinear in the X's, these models are linear functions of the unknown coefficients (or parameters) of the population regression model and thus are versions of the multiple regression model.
- Therefore, the unknown parameters of these nonlinear regression functions can be estimated and tested using OLS.
- In some applications, the regression function is a nonlinear function of the X's and of the parameters. If so, the parameters cannot be estimated by OLS, but they can be estimated using nonlinear least squares.

- the nonlinear models are extensions of the multiple regression model and therefore can be estimated and tested using the OLS.
- First, however, we return to the California test score data and consider the relationship between test scores and district income. (*Test Scores* and *District Income*)
- The economic background of the students is an important factor in explaining performance on standardized tests.
    - That analysis used two economic background variables (the percentage of students qualifying for a subsidized lunch and
    - the percentage of students whose families qualify for income assistance).
- A different, broader measure of economic background is the average annual per capita income in the school district ("district income"). The California data set includes district income measured in thousands of 1998 dollars. The sample contains a wide range of income levels: For the 420 districts in our sample, the median district income is 13.7 (that is, \$13,700 per person), and it ranges from 5.3 (\$5300 per person) to 55.3 (\$ 55,300 per person).

- Figure 8.2 shows a scatterplot of fifth-grade test scores against district income for the California data set, along with the OLS regression line relating these two variables.
- Test scores and district income are strongly positively correlated, with a correlation coefficient of 0.71; students from affluent districts do better on the tests than students from poor districts. But this scatterplot has a peculiarity: Most of the points are below the OLS line when income is very low (under $10,000) or very high (over $40,000), but they are above the line when income is between $15,000 and $30,000. There seems to be some curvature in the relationship between test scores and district income that is not captured by the linear regression.

A general formula for a nonlinear population regression function I

The nonlinear population regression models considered in this chapter are of the form

$$Y_i = f(X_{1i}, X_{1i}, \ldots, X_{ki}) + u_i, i = 1, \ldots, n, (8.3)$$

where $f(X_{1i}, X_{1i}, \ldots, X_{ki})$ is the population nonlinear regression function, a possibly nonlinear function of the independent variables $X_{1i}, X_{1i}, \ldots, X_{ki}$ and u_i is the error term.
For example, in the quadratic regression model in Equation (8.1), only one independent variable is present, so X1 is Income and the population regression function is

$$f(Income_i) = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 .(8.1)$$

- Because the population regression function is the conditional expectation of Y_i given $X_{1i}, X_{2i}, \ldots, X_{ki}$ in

A general formula for a nonlinear population regression function II

- The equation allow for the possibility that this conditional expectation is a nonlinear function of X_1; that is, $E[Y_i|X_{1i}, X_{2i}, \ldots, X_{ki}] = f(X_{1i}, X_{2i}, \ldots, X_{ki})$, where f can be a nonlinear function. If the population regression function is linear, then $f(X_{1i}, X_{2i}, \ldots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$, and Equation (8.3) becomes the linear regression model.
- This equation allows for nonlinear regression functions as well.

The effect on Y of a change in X_1 I

$$\widehat{TestScore} = \underset{(2.9)}{607.3} + \underset{(0.27)}{3.85}\,Income - \underset{(0.0048)}{0.0423}Income^2, (8.2)$$
$$R^2 = 0.554,$$

- Suppose an experiment is conducted on individuals with the same values of $X_1, X_2, \ldots, X_k$, and participants are randomly assigned treatment levels $X_1 = x_1$ or $X_1 + \Delta X_1 = x_1 + \Delta x_1$. Then the expected difference in outcomes is the causal effect of the treatment, holding constant $X_1, X_2, \ldots, X_k$.
- In the nonlinear regression model of Equation (8.3), this effect on Y is $Y = f(X_1 + \Delta X_1, \ldots, X_k) - f(X_1, X_2, \ldots, X_k)$. In the context of prediction $Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k)$ is the predicted difference in Y for two observations, both with the same values of $X_2, ..., X_k$, but with different values of X_1, specifically $X_1 + \Delta X_1$ and X_1.

The effect on Y of a change in X_1 II

- Because the regression function f is unknown, this population causal effect is also unknown. To estimate this effect, first estimate the regression function f. At a general level, denote this estimated function by $\hat{f}$; an example of such an estimated function is the estimated quadratic regression function in Equation (8.2). The estimated effect on Y (denoted Y_n) of the change in X_1 is the difference between the predicted value of Y when the independent variables take on the values $X_1 + \Delta X_1, X_2, \ldots, X_k$ and the predicted value of Y when they take on the values $X_1, X_2, \ldots, X_k$.
- The computational method in Key Concept 8.1 always works, whether X_1 is large or small and whether the regressors are continuous or discrete. Appendix 8.2 shows how to evaluate the slope using calculus for the special case of a single continuous regressor when X_1 small.

Application to test scores and district income I

- What is the predicted change in test cores associated with a change in district income of \$1000, based on the estimated quadratic regression function in Equation (8.2)? Because that regression function is quadratic, this effect depends on the initial district income. We therefore consider two cases: an increase in district income from 10 to 11 (i.e., from \$10,000 per capita to \$11,000 per capita) and an increase in district income from 40 to 41 (i.e., from \$40,000 per capita to \$41,000 per capita).

$$\Delta \hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k). (8.5)$$

- To compute Y_n associated with the change in income from 10 to 11, we can apply the general formula in Equation (8.5) to the quadratic regression model. Doing so yields

$$\Delta \hat{Y} = (\hat{\beta}_0 + \hat{\beta}_1 \times 11 + \hat{\beta}_2 \times 11^2) - (\hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 10^2). (8.6)$$

Application to test scores and district income II

- The term in the first set of parentheses in Equation (8.6) is the predicted value of Y when Income = 11, and the term in the second set of parentheses is the predicted value of Y when Income = 10. These predicted values are calculated using the OLS estimates of the coefficients in Equation (8.2). Accordingly, when Income = 10, the predicted value of test scores is 607.3 + 3.85 * 10 - 0.0423 * 102 = 641.57. When Income = 11, the predicted value is 607.3 + 3.85 * 11 - 0.0423 * 112 = 644.53.
- The difference in these two predicted values is $Y_n = 644.53 - 641.57 = 2.96$ points; that is, the predicted difference in test scores between a district with average income of $11,000 and one with average income of $10,000 is 2.96 points. In the second case, when income changes from $40,000 to $41,000, the difference in the predicted values in Equation (8.6) is $Y_n = (607.3 + 3.85 * 41 - 0.0423 * 41^2) - (607.3 + 3.85 * 40 - 0.0423 * 40^2) = 694.04 - 693.62 = 0.42$ points. Thus a change

Application to test scores and district income III

of income of \$1000 is associated with a larger change in predicted test scores if the initial income is \$10,000 than if it is \$40,000 (the predicted changes are 2.96 points versus 0.42 points). Said differently, the slope of the estimated quadratic regression function in Figure 8.3 is steeper at low values of income (like \$10,000) than at the higher values of income (like \$40,000).

Standard errors of estimated effects.

- The estimate of the effect on Y of changing X depends on the estimate of the population regression function, fn, which varies from one sample to the next. Therefore, the estimated effect contains a sampling error.

Application to test scores and district income IV

- One way to quantify the sampling uncertainty associated with the estimated effect is to compute a confidence interval for the true population effect. To do so, we need to compute the standard error of Y n in Equation (8.5).

$$SE(\Delta Y) = SE(\hat{\beta}_1 + 21\hat{\beta}_2)$$

- Thus, if we can compute the standard error of $\hat{\beta}_1 + 21\hat{\beta}_2$ then we have computed the standard error of Yn.
- Some regression software has a specialized command for computing the standard error in Equation (8.7) directly. If not, there are two other ways to compute it; these correspond to the two approaches in Section 7.3 for testing a single restriction on multiple coefficients.

1. The first method is to use approach 1 of Section 7.3, which is to compute the F-statistic testing the hypothesis that $\beta_1 + 21\beta_2 = 0$. The standard error of Y_n is then given by

Application to test scores and district income V

$$SE(\Delta\hat{Y}) = \frac{|\Delta\hat{Y}|}{\sqrt{F}}.(8.8)$$

the F-statistic testing the hypothesis that $\beta_1 + 21\hat{\beta}_2 = 0$ is $F = 299.94$. Because $\hat{Y} = 2.96$, applying Equation (8.8) gives $SE(\hat{Y}) = 2.96/\sqrt{299.94} = 0.17$.

1. The second approach is to compute the standard error using $var(\hat{\beta})$.

A comment on interpreting coefficients in nonlinear specifications I

- In the multiple regression model of Chapters 6 and 7, the regression coefficients had a natural interpretation.
- For example, β_1 is the expected change in Y associated with a change in X_1, holding the other regressors constant.
- But as we have seen, this is not generally the case in a nonlinear model. That is, it is not very helpful to think of β_1 in as being the effect of changing the district income, holding the square of the district income constant.
- In nonlinear models, the regression function is best interpreted by graphing it and by calculating the predicted effect on Y of changing one or more of the independent variables.

A General Approach to Modeling Nonlinearities Using Multiple Regression I

The general approach to modeling nonlinear regression functions taken in this chapter has five elements:

1. Identify a possible nonlinear relationship. The best thing to do is to use economic theory and what you know about the application to suggest a possible nonlinear relationship: ask yourself whether the slope of the regression function relating Y and X might reasonably depend on the value of X or on another independent variable.
    - For example, cutting class size from 18 students to 17 could have a greater effect than cutting it from 30 to 29.
2. Specify a nonlinear function, and estimate its parameters by OLS.(polynomial or logarithm?)
3. Determine whether the nonlinear model improves upon a linear model.
    - Most of the time you can use t-statistics and F-statistics to test the null hypothesis that the population regression function is linear against the alternative that it is nonlinear.

A General Approach to Modeling Nonlinearities Using Multiple Regression II

4. Plot the estimated nonlinear regression function. Does the estimated regression function describe the data well? Looking at Figures 8.2 and 8.3 suggests that the quadratic model fits the data better than the linear model.
5. Estimate the effect on Y of a change in X. The final step is to use the estimated regression to calculate the effect on Y of a change in one or more regressors X using the method in Key Concept 8.1.

Nonlinear Functions of a Single Independent Variable

- This section provides two methods for modeling a nonlinear regression function. To keep things simple, we develop these methods for a nonlinear regression function that involves only one independent variable, X.
- The first method discussed in this section is polynomial regression, an extension of the quadratic regression used in the last section to model the relationship between test scores and district income. The second method uses logarithms of X, of Y, or of both X and Y.
- Although these methods are presented separately, they can be used in combination.

Polynomials I

One way to specify a nonlinear regression function is to use a polynomial in X. In general, let r denote the highest power of X that is included in the regression.

The polynomial regression model of degree r is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots \beta_r X_i^r + u_r.$$

- When $r = 2$, the regression is the **quadratic regression model**.
- When $r = 3$, so that the highest power of X included is X_3, Equation (8.9) is called the **cubic regression model**.

Testing the null hypothesis that the population regression function is linear.

If the population regression function is linear, then the quadratic and higher-degree terms do not enter the population regression function. Accordingly, the null hypothesis (H_0) that the regression is linear and the alternative (H_1) that it is a polynomial of degree up to r correspond to

Polynomials II

$$
\begin{aligned}
H0 :& \beta_2 = 0, \beta_3 = 0, \ldots, \beta_r = 0 \text{ vs.} \\
H1 :& \text{ at least one } \beta_j \neq 0, j = 2, \ldots, r.
\end{aligned} \tag{8.10}
$$

- The null hypothesis that the population regression function is linear can be tested against the alternative that it is a polynomial of degree up to r by testing H_0 against
- H_1 in Equation (8.10). Because H_0 is a joint null hypothesis with $q = r - 1$ restrictions on the coefficients of the population polynomial regression model, it can be tested using the F-statistic.

Which degree polynomial should I use? I

How many powers of X should be included in a polynomial regression?

- The answer balances a trade-off between flexibility and statistical precision.
- Increasing the degree r introduces more flexibility into the regression function and allows it to match more shapes; a polynomial of degree r can have up to $r-1$ bends (that is, inflection points) in its graph.
- But increasing r means adding more regressors, which can reduce the precision of the estimated coefficients.

A practical way to determine the degree of the polynomial is to ask whether the coefficients in Equation (8.9) associated with largest values of r are 0. If so, then these terms can be dropped from the regression.

Sequential Hypothesis Testing I

This procedure, which is called sequential hypothesis testing because individual hypotheses are tested sequentially, is summarized in the following steps:

1. Pick a maximum value of r, and estimate the polynomial regression for that r
2. Use the t-statistic to test the hypothesis that the coefficient on X_r, is 0. If you reject this hypothesis, then X_r belongs in the regression, so use the polynomial of degree r.
3. If you do not reject $\beta_r = 0$ in step 2, eliminate Xr from the regression, and estimate a polynomial regression of degree $r - 1$. Test whether the coefficient on X_{r-1} is 0. If you reject, use the polynomial of degree $r - 1$.
4. If you do not reject $\beta_{r-1} = 0$ in step 3, continue this procedure until the coefficient on the highest power in your polynomial is statistically significant

Sequential Hypothesis Testing II

This recipe has one missing ingredient: the initial degree r of the polynomial. In many applications involving economic data, the nonlinear functions are smooth; that is, they do not have sharp jumps, or "spikes." If so, then it is appropriate to choose a small maximum degree for the polynomial, such as 2, 3, or 4—that is, to begin with $r = 2$ or 3 or 4 in step 1.

Application to district income and test scores. I

The estimated cubic regression function relating district income to test scores is

$$\widehat{TestScore} = \underset{(5.1)}{600.1} + \underset{(0.71)}{5.02}\,Income - \underset{(0.029)}{0.096}\,Income^2 + \underset{(0.00035)}{0.00069}\,Income^3,$$

$$\bar{R}^2 = 0.555.$$

The t-statistic on $Income^3$ is 1.97, so the null hypothesis that the regression function is a quadratic is rejected against the alternative that it is a cubic at the 5% level.
Moreover, the F-statistic testing the joint null hypothesis that the coefficients on $Income^2$ and $Income^3$ are both 0 is 37.7, with a p-value less than 0.01%, so the null hypothesis that the regression function is linear is rejected against the alternative that it is either a quadratic or a cubic.

Interpretation of coefficients in polynomial regression models

The coefficients in polynomial regressions do not have a simple interpretation. The best way to interpret polynomial regressions is to plot the estimated regression function and calculate the estimated effect on Y associated with a change in X for one or more values of X.

Logarithms I

Another way to specify a nonlinear regression function is to use the natural logarithm of Y and/or X. Logarithms convert changes in variables into percentage changes, and many relationships are naturally expressed in terms of percentages.
Here are some examples:

- box in SW Ch 3, "Social Class or Education? Childhood Circumstances and Adult Earnings Revisited," examined the household earnings gap by socioeconomic classification. In that discussion, the wage gap was measured in terms of pounds sterling. However, it is easier to compare wage gaps across professions and over time when they are expressed in percentage terms.
- In SW 8.1, we found that district income and test scores were nonlinearly related. Would this relationship be linear using percentage changes? That is, might it be that a change in district income of 1%—rather than \$1000—is associated with a change in test scores that is approximately constant for different values of income?

Logarithms II

- In the economic analysis of consumer demand, it is often assumed that a 1% increase in price leads to a certain percentage decrease in the quantity demanded. The percentage decrease in demand resulting from a 1% increase in price is called the price **elasticity**.

Logarithms III

The logarithm function has the following useful properties

$$
\begin{aligned}
\ln(1/x) &= -\ln x \\
\ln(ax) &= \ln a + \ln x \\
\ln(a/x) &= \ln a - \ln x \\
\ln(a^x) &= x \ln a.
\end{aligned}
$$

Logarithms and percentages I

The link between the logarithm and percentages relies on a key fact: When x is small, the difference between the logarithm of $x + \Delta x$ and the logarithm of x is approximately $\Delta x / x$, the percentage change in x divided by 100. That is,

$$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}.(\text{ when } \frac{\Delta x}{x} \text{ is small.})$$

Proof: using taylor expansion.

The three logarithmic regression models.

There are three different cases in which logarithms might be used: when X is transformed by taking its logarithm but Y is not; when Y is transformed to its logarithm but X is not; and when both Y and X are transformed to their logarithms. The interpretation of the regression coefficients is different in each case. We discuss these three cases in turn

1. linear-log model. $Y_i = \beta_0 + \beta_1 \ln(X_i)$. In the linear-log model, a 1% change in X is associated with a change in Y of $0.01\beta1$.

Logarithms and percentages II

2. log-linear model. $\ln Y_i = \beta_0 + \beta_1 X_i$. a one-unit change in $\Delta X_1 = 1$ is associated with a $(100 * \beta_1)\%$ change in Y.
3. log-log model. $\ln Y_i = \beta_0 + \beta_1 \ln X_i$. In the log-log model, a 1% change in X is associated with a β_1 change in Y.

A difficulty with comparing logarithmic specifications I

Which of the log regression models best fits the data? the R^2 can be used to compare the log-linear and log-log models; as it happened, the log-log model had the higher R^2. Similarly, the R^2 can be used to compare the linearlog regression in Equation (8.18) and the linear regression of Y against X.

- the linear-log regression has an R^2 of 0.561,
- while the linear regression has an R^2 of 0.508.
- so the linear-log model fits the data better.

How can we compare the linear-log model and the log-log model? Unfortunately, the R^2 **cannot** be used to compare these two regressions because their dependent variables are different: one is Y, the other is $\ln(Y)$.
R^2 measures the fraction of the variance of the dependent variable explained by the regressors. Because the dependent variables in the log-log and linear-log models are different, it does not make sense to compare their R^2's.

A difficulty with comparing logarithmic specifications II

- Because of this problem, the best thing to do in a particular application is to decide, using economic theory and either your or other experts' knowledge of the problem, whether it makes sense to specify Y in logarithms.
- For example, labor economists typically model earnings using logarithms because wage comparisons, contract wage increases, and so forth are often most naturally discussed in percentage terms.
- In modeling test scores, it seems natural (to us, anyway) to discuss test results in terms of points on the test rather than percentage increases in the test scores, so we focus on models in which the dependent variable is the test score rather than its logarithm.

Computing predicted values of Y when Y is in logarithms I

If the dependent variable Y has been transformed by taking logarithms, the estimated regression can be used to compute directly the predicted value of $\ln(Y)$. However, it is a bit trickier to compute the predicted value of Y itself.
To see this, consider the log-linear regression model and rewrite it so that it is specified in terms of Y rather than $\ln(Y)$.

$$Y_i = \exp(\beta_0 + \beta_1 X_i + u_i).$$

Polynomial and Logarithmic Models have of Test Scores and District Income

In practice, economic theory or expert judgment might suggest a functional form to use, but in the end, the true form of the population regression function is unknown.
In practice, fitting a nonlinear function therefore entails deciding which method or combination of methods works best.
As an illustration, we compare polynomial and logarithmic models of the relationship between district income and test scores.

Polynomial specifications

- We considered two polynomial specifications, quadratic and cubic.
- Because the coefficient on $Income^3$ was significant at the 5% level,
- the cubic specification provided an improvement over the quadratic, so we select the cubic model as the preferred polynomial specification.

Logarithmic specifications I

- The logarithmic specification in seemed to provide a good fit to these data, but we did not test this formally.
- One way to do so is to augment it with higher powers of the logarithm of income. If these additional terms are not statistically different from 0, then we can conclude that the specification in is adequate in the sense that it cannot be rejected against a polynomial function of the logarithm.

$$\widehat{TestScore} = \underset{(3.8)}{557.8} + \underset{(1.40)}{36.42}\ln(Income), \bar{R}^2 = 0.561. \tag{8.18}$$

- Accordingly, the estimated cubic regression (specified in powers of the logarithm of income) is

$$\begin{aligned}\widehat{TestScore} = &\underset{(79.4)}{486.1} + \underset{(87.9)}{113.4}\ln(Income) - \underset{(31.7)}{26.9}[\ln(Income)]^2 \\ &+ \underset{(3.74)}{3.06}[\ln(Income)]^3, \bar{R}^2 = 0.560\end{aligned} \tag{8.26}$$

Logarithmic specifications II

- The t-statistic on the coefficient on the cubic term is 0.818, so the null hypothesis that the true coefficient is 0 is not rejected at the 10% level. The F-statistic testing the joint hypothesis that the true coefficients on the quadratic and cubic term are both 0 is 0.44, with a p-value of 0.64, so this joint null hypothesis is not rejected at the 10% level.
- Thus the cubic logarithmic model in Equation (8.26) does not provide a statistically significant improvement over the model in Equation (8.18), which is linear in the logarithm of income.

Interactions Between Independent Variables I

- In the introduction to this chapter, we wondered whether reducing the student–teacher ratio might have a bigger effect on test scores in districts where many students are still learning English than in those with few still learning English.
- This could arise, for example, if students who are still learning English benefit differentially from one-on-one or small-group instruction.
  - If so, the presence of many English learners in a district would interact with the student–teacher ratio in such a way that the effect on test scores of a change in the student–teacher ratio would depend on the fraction of English learners.

Interactions Between Independent Variables II

- The possible interaction between the student–teacher ratio and the fraction of English learners is an example of the more general situation in which the effect on Y of a change in one independent variable depends on the value of another independent variable. We consider three cases:
  1. when both independent variables are binary,
  2. when one is binary and the other is continuous,
  3. and when both are continuous.

Interactions Between Two Binary Variables I

Consider the population regression of log earnings Y_i, where $Y_i = \ln(Earnings_i)$ against two binary variables:

1. whether a worker has a college degree(D_{1i}), where $D_{1i} = 1$ if the ith person graduated from college,
2. and the worker's sex (D_{2i}), where $D_{2i} = 1$ if the ith person is female).

In this regression model, β_1 is the effect on log earnings of having a college degree, holding sex constant, and β_2 is the mean difference between female and male earnings, holding schooling constant. The specification has an important limitation:

- The effect of having a college degree in this specification, holding constant sex, is the same for men and women.
- Phrased mathematically, the effect on Y_i of D_{1i}, holding D_{2i} constant, could depend on the value of D_{2i}.
- There could be an interaction between having a college degree and sex, so that the value in the job market of a degree is different for men and women.

Interactions Between Two Binary Variables II

It is easy to modify the specification so that it does by introducing another regressor to include interaction between having a college degree and sex, the product of the two binary variables, $D_{1i} * D_{2i}$. The resulting regression is

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

The new regressor, the product $D_{1i} * D_{2i}$, is called an interaction term or an **interacted regressor**, and the population regression model is called a binary variable interaction regression model. The interaction term in Equation (8.28) allows the population effect on log earnings Y_i of having a college degree (changing D_{1i} from $D_{1i} = 0$ to $D_{1i} = 1$) to depend on sex D_{2i}.

Interactions Between Two Binary Variables III

1. Compute the conditional expectation of Y_i for $D_{1i} = 0$ given a value of D_{2i}; this is $E[Y_i|D_{1i} = 0, D_{2i} = d_2] = \beta_0 + \beta_1 * 0 + \beta_2 * d2 + \beta_3 * 0 * d2$, note that under **conditional mean zero assumption**, $E[u_i|D_{1i}, D_{2i}] = 0$.
2. compute the conditional expectation of Y_i after the change:

$$E[Y_i|D_{1i} = 1, D_{2i} = d_2] - E[Y_i|D_{1i} = 0, D_{2i} = d_2] = \beta_1 + \beta_3 d_2.$$

- if the person is male $d_2 = 0$, the effect of acquiring a college degree is β_1, but if the person is female $(d_2 = 1)$, the effect s $\beta_1 + \beta_3$.
- the coefficient β_3 on the **interaction term** is the difference in the effect of acquiring a college degree for women versus that for men.
- The binary variable interaction regression allows the effect of changing one of the binary independent variables to depend on the value of the other binary variable.

Application to the student–teacher ratio and the percentage of English learners I

- Let $HiSTR_i$ be a binary variable that 1 if the student–teacher ratio is 20 or more,
- $HiEL_i$ be a binary variable that equals 1 if the percentage of English learners is 10% or more and that equals 0 otherwise.

$$\widehat{TestScore} = \underset{(1.4)}{664.1} - \underset{(1.9)}{1.9}\,HiSTR - \underset{(2.3)}{18.2}HiEL - \underset{(3.1)}{3.5}(HiSTR * HiEL),$$

$$\bar{R}^2 = 0.29$$

this effect thus is $-1.9 - 3.5HiEL$.

- if the fraction of English learners is low $(HiEL = 0)$, then the effect on test scores of moving from $(HiSTR = 0)$ to $(HiSTR = 1)$ is for test scores to decline by 1.9 points.
- If the fraction of English learners is high, then test scores are estimated to decline by $1.9 + 3.5 = 5.4$ points.

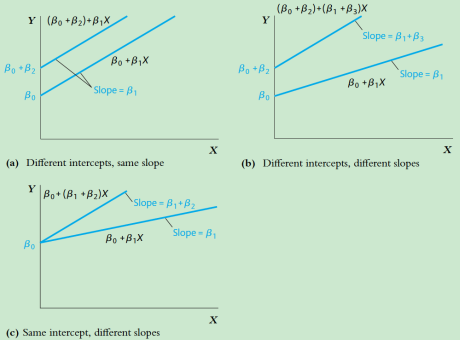
Application to the student–teacher ratio and the percentage of English learners II

We can compute the sample average using different combinations of the binary variables:

- when $HiSTR_i = 0$ and $HiEL_i = 0$, the sample average is 664.1.
- when $HiSTR_i = 0$ and $HiEL_i = 1$, the sample average is $645.9 (= 664.1 - 18.2)$
- when $HiSTR_i = 1$ and $HiEL_i = 0$, the sample average is $662.2 (= 664.1 - 1.9)$.
- when $HiSTR_i = 1$ and $HiEL_i = 1$, the sample average is $640.5 (= 664.1 - 1.9 - 18.2 - 3.5)$.

Interactions Between a Continuous and a Binary Variable I

FIGURE 8.8 Regression Functions Using Binary and Continuous Variables



Interactions of binary variables and continuous variables can produce three different population regression functions: (a) $\beta_0 + \beta_1X + \beta_2D$ allows for different intercepts but has the same slope, (b) $\beta_0 + \beta_1X + \beta_2D + \beta_3(X \times D)$ allows for different intercepts and different slopes, and (c) $\beta_0 + \beta_1X + \beta_2(X \times D)$ has the same intercept but allows for different slopes.

Interactions Between a Continuous and a Binary Variable II

- Different intercepts, same slope (Figure 8.8a):

$$Yi = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i;$$

- . Different intercepts and slopes (Figure 8.8b):

$$Yi = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i * D_i + u + i;$$

- Same intercept, different slopes (Figure 8.8c):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i * D_i + u_i.$$

All three specifications—Equations (8.31), (8.32), and (8.33)—are versions of the multiple regression model, and once the new variable $X_i * D_i$ is created, the coefficients of all three can be estimated by OLS.

Application to the student–teacher ratio and the percentage of English learners I

Does the effect on test scores of cutting the student–teacher ratio depend on whether the percentage of students still learning English is high or low? One way to answer this question is to use a specification that allows for two different regression lines, depending on whether there is a high or a low percentage of English learners. This is achieved using the different intercept/different slope specification where the binary variable HiELi equals 1 if the percentage of students still learning English in the district is greater than 10% and equals 0 otherwise

$$\widehat{TestScore} = \underset{(11.9)}{682.2} - \underset{(0.59)}{0.97}\, STR + \underset{(19.5)}{5.6}\, HiEL - \underset{(0.97)}{1.28}(STR * HiEL)$$

Application to the student–teacher ratio and the percentage of English learners II

- For districts with a low fraction of English learners $HiEL_i = 0$, the estimated regression line is $682.2 - 0.97STR_i$. For districts with a high fraction of Eng-lish learners $HiEL_i = 1$, the estimated regression line is $682.2 + 5.6 - 0.97STR_i - 1.28STR_i = 687.8 - 2.25STR_i$. According to these estimates, reducing the student–teacher ratio by 1 is predicted to increase test scores by 0.97 points in districts with low fractions of English learners but by 2.25 points in districts with high fractions of English learners.
- The difference between these two effects, 1.28 points, is the coefficient on the interaction term in Equation (8.34). The interaction regression model in Equation (8.34) allows us to estimate the effect of more nuanced policy interventions than the across-the-board class size reduction considered so far. For example, suppose the state considered a policy to reduce the student–teacher ratio by 2 in districts with a high fraction of

Application to the student–teacher ratio and the percentage of English learners III

English learners $HiEL_i = 1$ but to leave class size unchanged in other districts.

- The OLS regression in Equation (8.34) can be used to test several hypotheses about the population regression line. First, the hypothesis that the two lines are, in fact, the same can be tested by computing the F-statistic testing the joint hypothesis that the coefficient on $HiEL_i$ and the coefficient on the interaction term $STR_i * HiEL_i$ are both 0. This F-statistic is 89.9, which is significant at the 1% level.
- Second, the hypothesis that two lines have the same slope can be tested by testing whether the coefficient on the interaction term is 0. The t-statistic, $-1.28/0.97 = -1.32$, is less than 1.64 in absolute value, so the null hypothesis that the two lines have the same slope cannot be rejected using a two-sided test at the 10% significance level.

Application to the student–teacher ratio and the percentage of English learners IV

- Third, the hypothesis that the lines have the same intercept corresponds to the restriction that the population coefficient on *HiEL* is 0. The t-statistic testing this restriction is $t = 5.6/19.5 = 0.29$, so the hypothesis that the lines have the same intercept cannot be rejected at the 5% level.
- These three tests produce seemingly contradictory results: The joint test using the F-statistic rejects the joint hypothesis that the slope and the intercept are the same, but the tests of the individual hypotheses using the t-statistic fail to reject. The reason is that the regressors, *HiEL* and $STR * HiEL$, are highly correlated. This results in large standard errors on the individual coefficients. Even though it is impossible to tell which of the coefficients is nonzero, there is strong evidence against the hypothesis that both are 0.

Application to the student–teacher ratio and the percentage of English learners V

- Finally, the hypothesis that the student–teacher ratio does not enter this specification can be tested by computing the F-statistic for the joint hypothesis that the coefficients on STR and on the interaction term are both 0. This F-statistic is 5.64,
- which has a p-value of 0.004. Thus the coefficients on the student–teacher ratio are jointly statistically significant at the 1% significance level.

Application to the student–teacher ratio and the percentage of English learners.

$$\widehat{TestScore} = \underset{(11.8)}{686.3} - \underset{(0.59)}{1.12}\,STR - \underset{(0.37)}{0.67}\,PctEL + \underset{(0.019)}{0.0012}(STR * PctEL),$$

$$\bar{R}^2 = 0.422.(8.37)$$

Application to the student–teacher ratio and the percentage of English learners VI

- When the percentage of English learners is at the median ($PctEL = 8.85$), the slope of the line relating test scores and the student–teacher ratio is estimated to be $-1.11 (= -1.12 + 0.0012 * 8.85)$.
- When the percentage of English learners is at the 75th percentile ($PctEL = 23.0$), this line is estimated to be slightly flatter, with a slope of $-1.09 (= -1.12 + 0.0012 * 23.0)$.
- For a district with 8.85% English learners, the estimated effect of a one-unit reduction in the student–teacher ratio is to increase test scores by 1.11 points, but for a district with 23.0% English learners, reducing the student–teacher ratio by one unit is predicted to increase test scores by only 1.09 points.
- The difference between these estimated effects is not statistically significant, however: The t-statistic testing whether the coefficient on the interaction term is 0 is $t = 0.0012/0.019 = 0.06$, which is not significant at the 10% level.

References I

Stock, J. H. and Watson, M. W. (2020). *Introduction to econometrics*, volume 4. Pearson New York.