

LECTURE /DISCUSSION

Specification of the OLS Regression Model

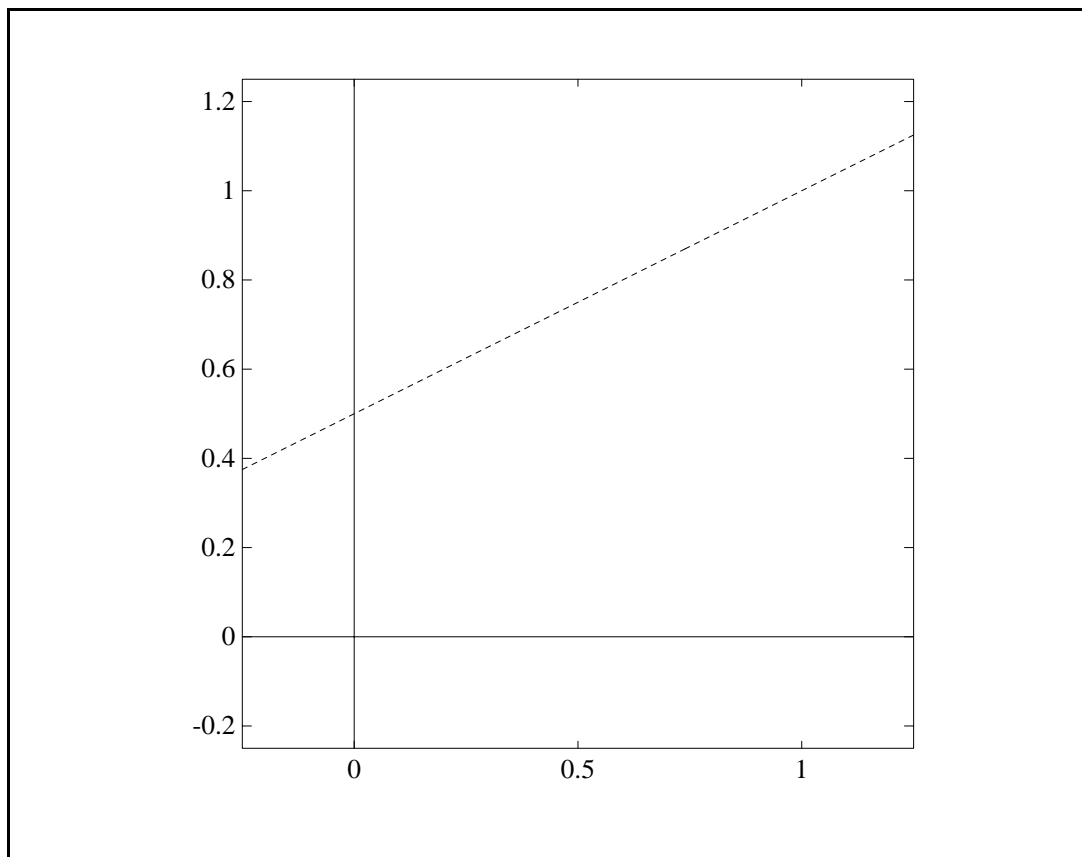
The Specification

The specification is the selection of explanatory variables and the transformations, if any, of the dependent variable and the explanatory variables.

The simplest linear model is

$$y = \alpha + \beta x + \varepsilon$$

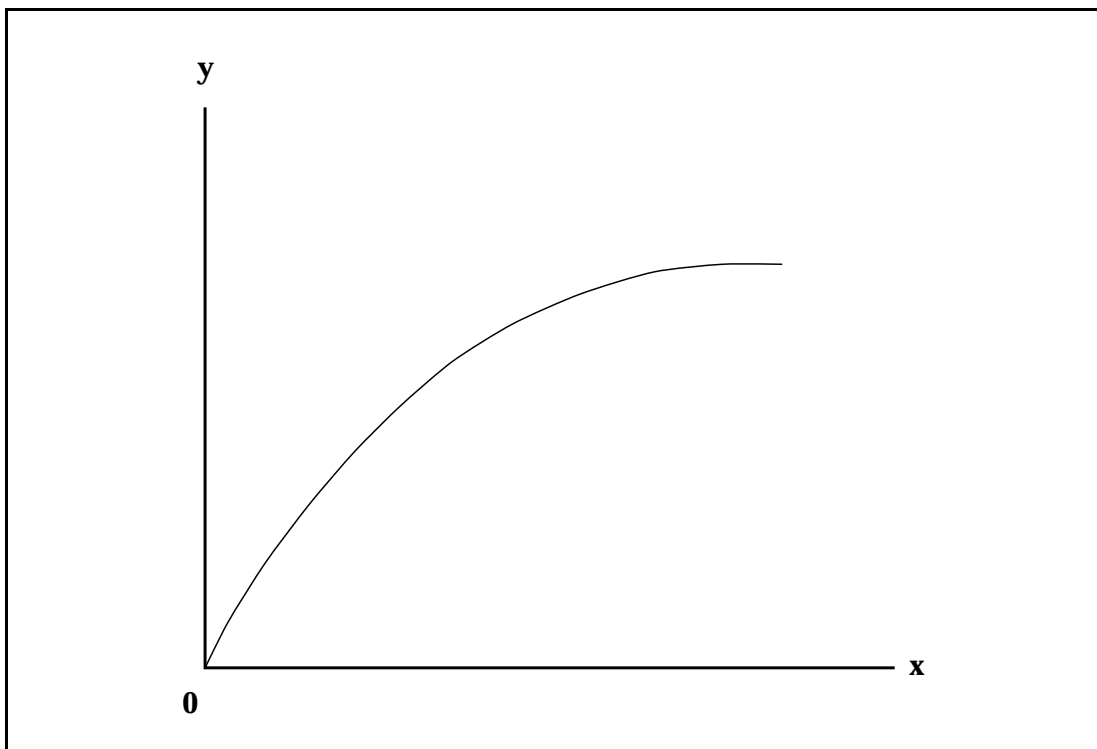
which specifies a **linear** relationship between the dependent variable y and the single explanatory variable x .



Nonlinear Relationships

By transforming the dependent variable or the explanatory variable, the linear model can handle **nonlinear** relationships. For example, one could alter the simple model above by replacing the explanatory variable x with the transformed variable $\sqrt{x} = z$:

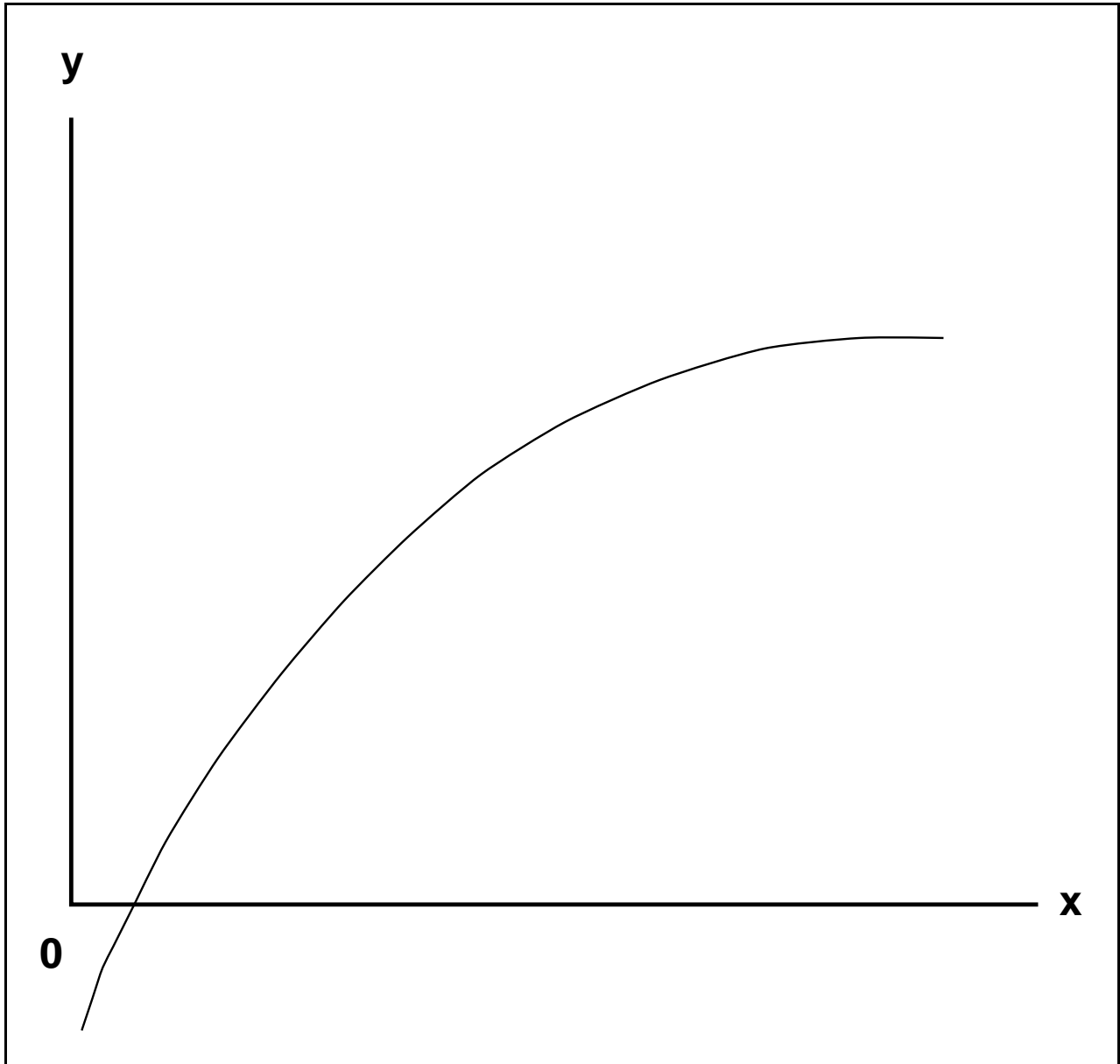
$$y = \gamma + \delta\sqrt{x} + \varepsilon = \gamma + \delta z + \varepsilon$$



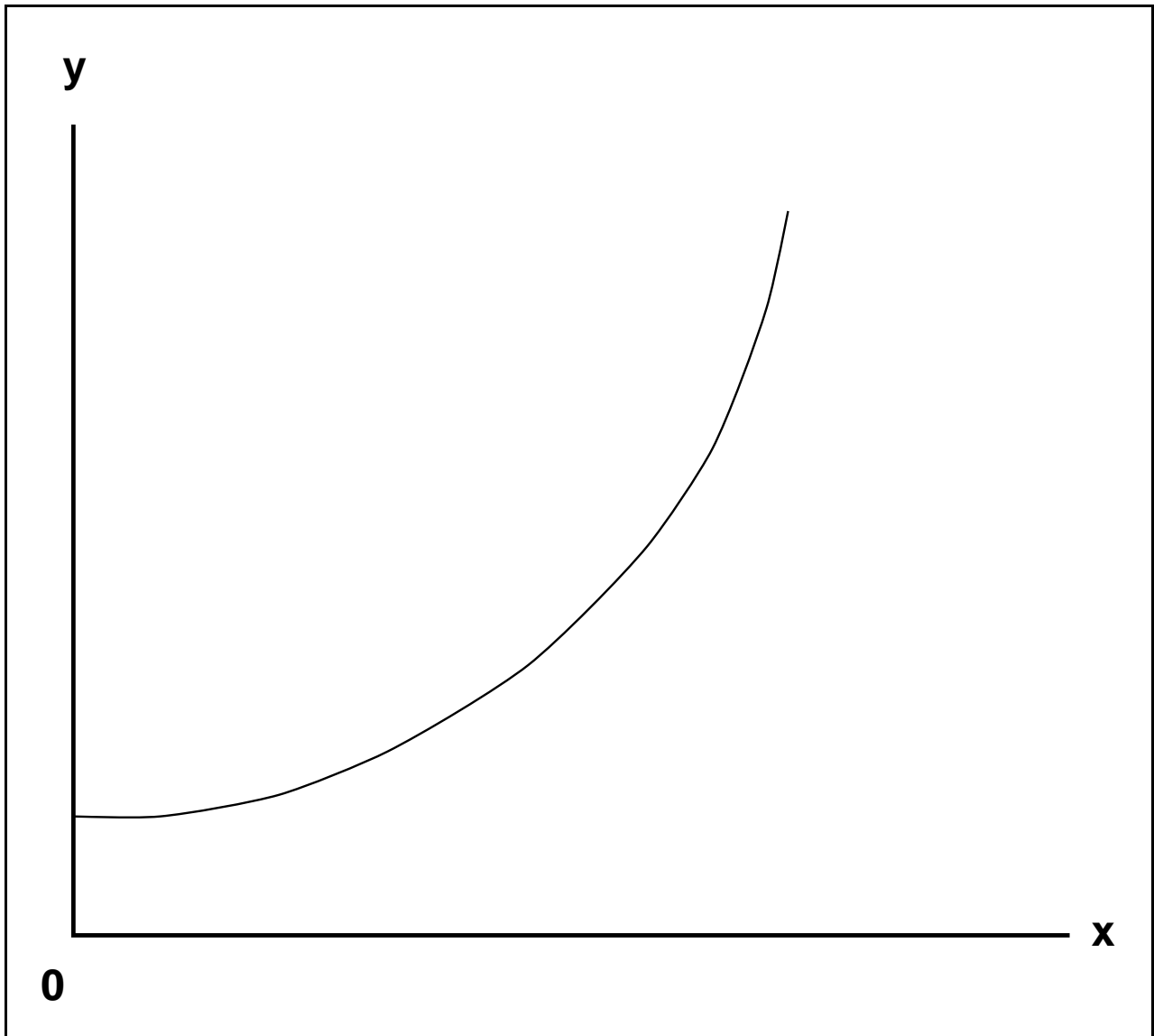
The transformation specifies a strongly diminishing influence of x on y such that for large values of x further increases in x leave y virtually unchanged.

Other transformations:

$$y = \alpha + \beta \ln(x) + \varepsilon$$



$$y = \alpha + \beta x^2$$

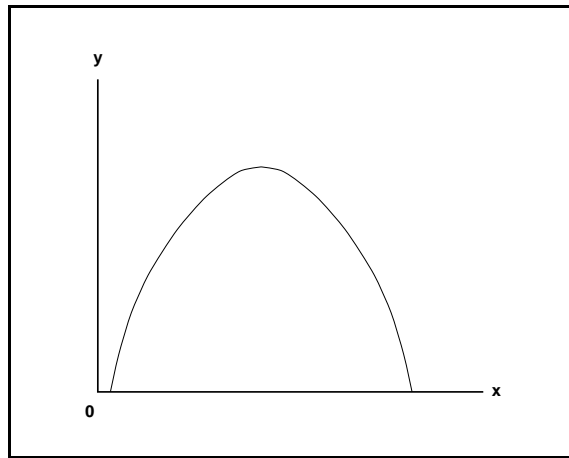


x^2 increases at an increasing rate.

Entering several transformations:

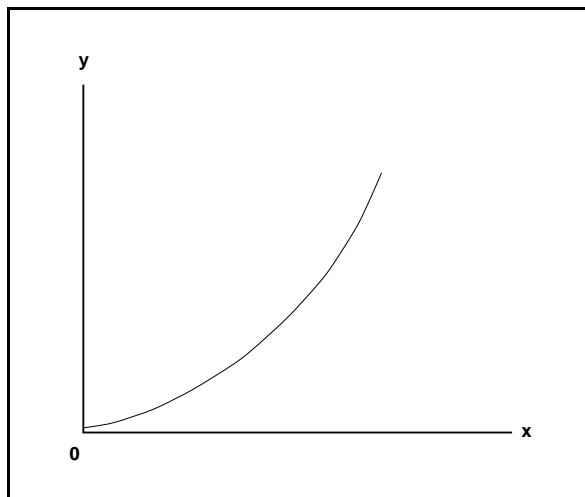
Example: $y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$

If $\beta_1 > 0$ and $\beta_2 < 0$,

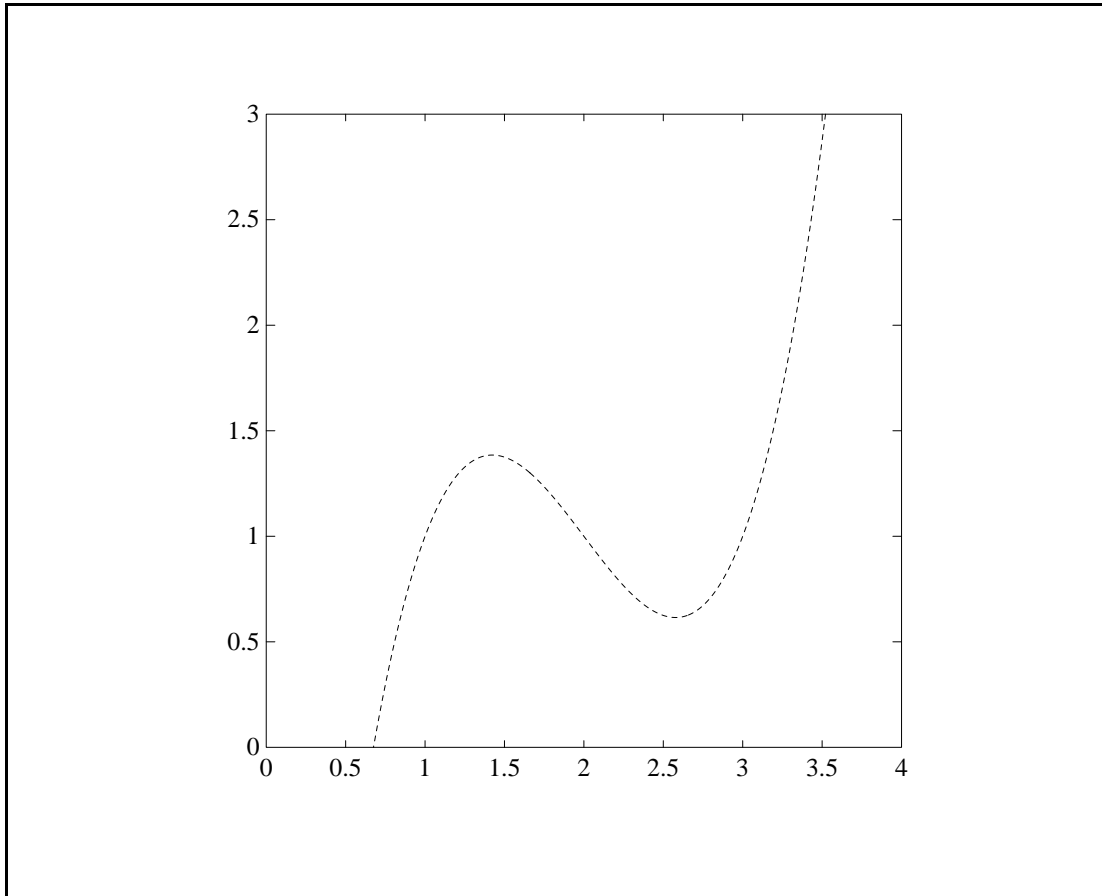


$$y = 2x - .1x^2$$

If $\beta_1 > 0$ and $\beta_2 > 0$,



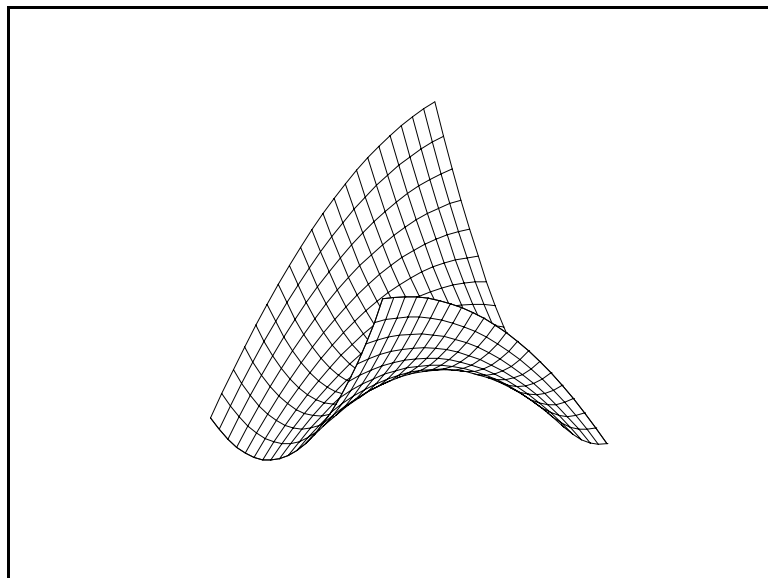
$$y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$
$$\beta_1 > 0, \beta_2 < 0, \beta_3 > 0$$



The figure depicts a relationship that is rarely found in applications, but it illustrates the flexibility that polynomial specifications can exhibit.

Transformations of more than one explanatory variable:

$$y = \beta_1 + \beta_2 w + \beta_3 x + \beta_4 w^2 + \beta_5 x^2 + \beta_6 wx + \varepsilon$$



These specifications reflect changes in the marginal effect of one explanatory variable given others: We can rewrite the equation above as

$$y = (\beta_1 + \beta_2 w + \beta_4 w^2) + (\beta_3 + \beta_5 x + \beta_6 w)x + \varepsilon$$

and interpret the intercept as a function of w and the slope of x as changing with w and x .

Note that the flexibility of polynomials is limited by an important feature: Such specifications always drive the predicted value of the dependent variable to extreme values at extreme values of the explanatory variables.

Transforming the Dependent Variable

One can transform the dependent variable, too:

$$\ln (y) = \alpha + \beta x + \varepsilon$$

The log transformation is often applied to all the explanatory variables as well. Such specifications are called ***log-linear*** or ***log-log*** models:

$$\ln (y) = \alpha + \beta \ln (x) + \varepsilon$$

Linear Model

$$y = \alpha + \beta x + \varepsilon$$

Impact of a unit change in x :

$$\frac{dy}{dx} = \beta$$

Elasticity = percent change in y for a percent change in x :

$$\text{Elasticity} = \frac{\frac{dy}{dx}}{\frac{y}{x}} = \frac{dy}{dx} \cdot \frac{x}{y} = \beta \frac{x}{y}$$

Log-log Model

$$\ln (y) = \alpha + \beta \ln (x) + \varepsilon$$

or

$$y = \exp (\alpha + \beta \ln (x) + \varepsilon)$$

Impact of a unit change in x :

$$\frac{dy}{dx} = \exp(\alpha + \beta \ln (x) + \varepsilon) \cdot \beta \cdot \frac{1}{x} = \beta \frac{y}{x}$$

Elasticity:

$$\text{Elasticity} = \frac{\frac{dy}{dx}}{\frac{y}{x}} = \frac{dy}{dx} \cdot \frac{x}{y} = \left(\beta \frac{y}{x}\right) \frac{x}{y} = \beta$$

Choosing Explanatory Variables

Specification also includes the selection of the explanatory variables that are entered, transformed or untransformed, on the right-hand side of the regression equation. This selection is limited often by the data sets available for analysis. When the researcher participates in data collection, perhaps through the design of a survey, the possibilities may be greater but budget constraints still restrict choices. We will show later that one cannot yield to the temptation to put all possible explanatory variables into a regression model. This approach does not solve this specification issue.

Given a data set, the selection of explanatory variables is governed by

- theory (economic, engineering)
- purpose (forecasting, experimental design, testing theory, summarizing data)
- experience, practice, and common sense

Generally, one must strike a compromise between these three sources of ideas for specification.

Theory is helpful for determining which variables must be included and how they should enter.

For example, the analysis of consumer demand for a commodity should include not only its price, but the prices of close substitutes and complementary goods. Household income determines household demand as well.

Furthermore, the price of a good should decrease demand as price increases, as should the prices of complementary goods, while the prices of substitute goods should increase demand as those prices rise.

Finally, the demand for a good is rarely a linear function of prices and income: At a low price, demand will fall rapidly as price increases while demand is inelastic at high prices.

Experience, industry practice, and common sense also govern specification. Explanatory variables that should appear on theoretical grounds are often omitted because their influence is already known to be slight (based on previous experience). A desire to compare results with previous studies may lead a researcher to specify a regression equation identical to previous practice. Common sense may be one of the most powerful motivations of all.

Common sense compels analysts to put a constant term in virtually all regressions. Ultimately, a regression model is an approximation to a more realistic statistical model and a constant term is the first term in any approximation.

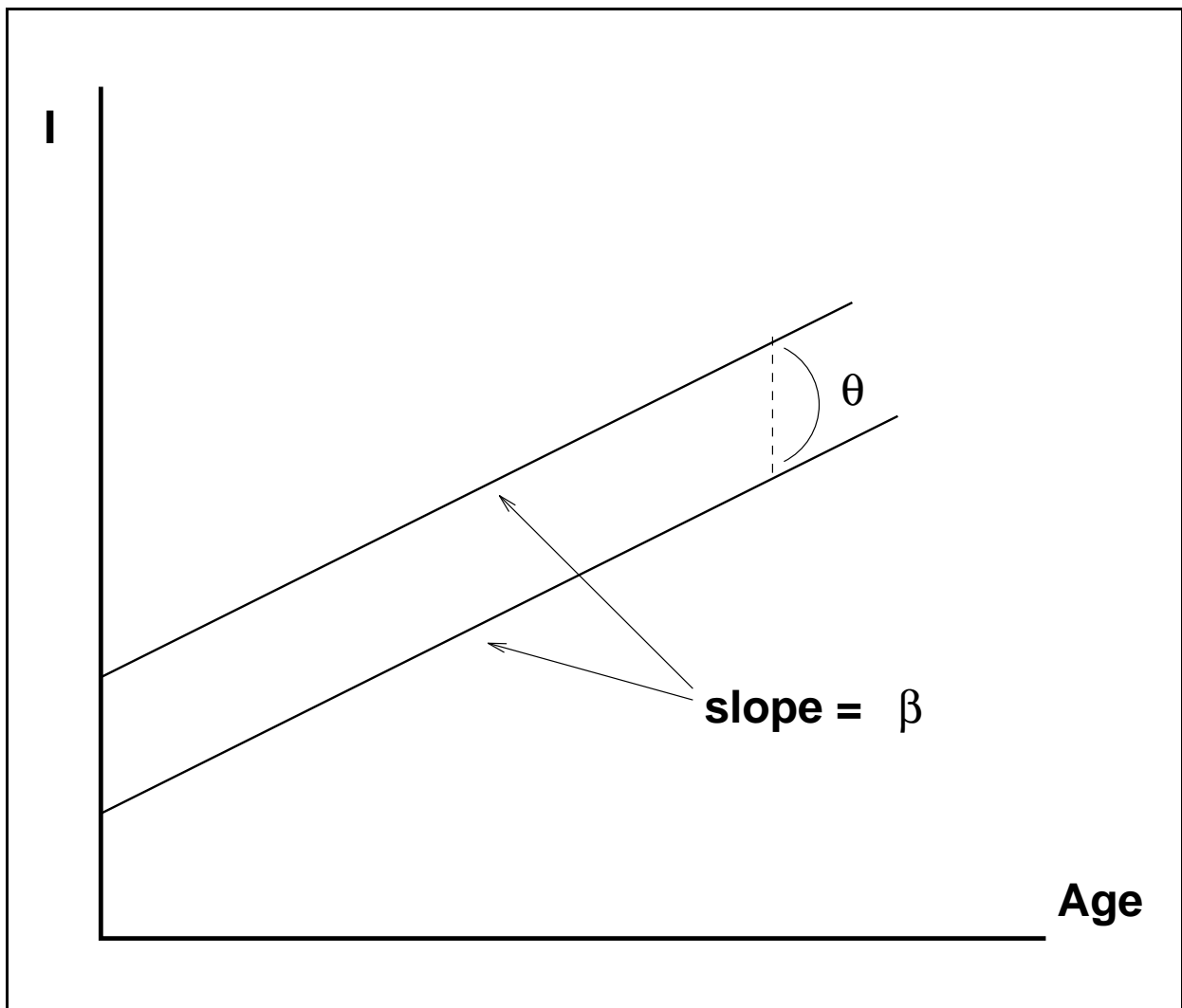
Dummy Variables

Common sense also leads to the introduction of ***dummy variables*** to regression models for improvements in the approximation. A dummy variable is a variable that can have only two possible values, zero and one. When a dummy variable is one, the variable indicates the occurrence of a particular state.

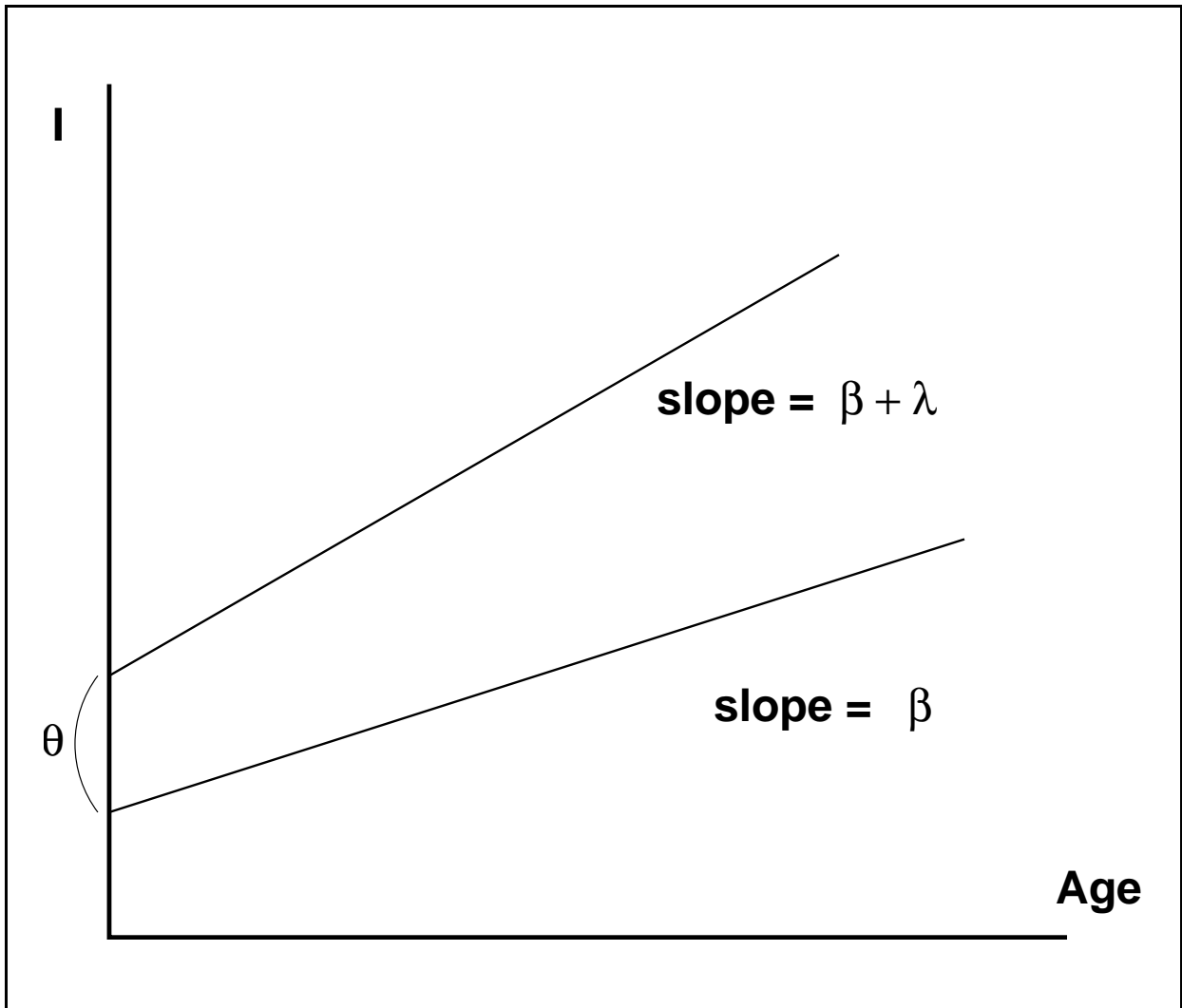
Example 1: Effect of Education on Income

$$\text{Income} = \alpha + \beta \text{Age} + \theta D + \varepsilon ,$$

where $D = 1$ if college degree, zero otherwise.



$$\text{Income} = \alpha + \beta \text{Age} + \theta D + \lambda (\text{Age} * D)$$



Example 2: Conditional Demand Models

CDMs estimate the Unit Energy Consumption (UEC) of each appliance:

$$E = \alpha + \beta_1 d_1 + \beta_2 d_2 + \dots + \varepsilon$$

E = annual kWh consumption

d_1 = 1 if household has an air conditioner, 0 otherwise

d_2 = 1 if household has electric heat, 0 otherwise

d_3 = 1 if household has electric water heater, 0 otherwise

etc.

β_1 is air conditioner UEC

β_2 is electric heater UEC

β_3 is water heater UEC

etc.

For the commercial sector, annual consumption is usually expressed as per square foot. CDMs estimate the Energy Use Intensity (EUI) of each end-use.

$$\text{EPS} = \alpha + \theta_1 d_1 + \theta_2 d_2 + \theta_{d3} + \dots + \varepsilon$$

EPS is annual kWh consumption of building divided by its square feet of space.

Incorporating Prior Information: Statistically Adjusted Engineering (SAE) Models

Let A_1 be an engineering estimate of the annual energy consumption of an air conditioner, and A_2 be an engineering estimate of the annual energy consumption of an electric heater.

Estimate a CDM with A_1, A_2, \dots , replacing the dummies d_1, d_2, \dots .

$$E = \alpha + \lambda_1 A_1 + \lambda_2 A_2 + \dots + \varepsilon$$

λ_1 is an adjustment factor for air conditioning load.

- $\lambda_1 = 1$ if A_1 is correct
- $\lambda_1 < 1$ if A_1 is too high
- $\lambda_1 > 1$ if A_1 is too low

Example

Adjustment coefficients from an SAE model of residential customers.

Central air	.48
Window air	.19
Water heater	.92
Range/oven	1.23
Dishwasher	.73
Washer/dryer	2.04
Refrigerator	1.06