Coding Categorical Variables in Regression: Indicator or Dummy Variables

Professor George S. Easton

DataScienceSource.com

This video is embedded on the following web page at DataScienceSource.com:

DataScienceSource.com/DummyVariables

On that page you will find links to a copy of these slides as well as links to the code file used in the video.

It is best to start from that page!

Introduction

- Indicator or “dummy” variables are regression X variables that take on only the values 0 and 1.
- Dummy variables are use to indicate the presence or absence of something. Examples:
  - Own a dog.
  - Is college educated.
  - Is a male.

Introduction (con’d)

- It is very natural to use a dummy variable to represent a binary x-variable in a regression. For example, we can code gender using Female = 0 and Male = 1.
- But we can also use dummy variables to code more than two categories.

Reminder - Terminology

- Quite a few different terms are used to describe variables that have categories (like defective or good, spam or ham, or grades A, B, C, D, F):
  - Categorical variable
  - Nominal variable
  - Attribute variable
  - Ordinal variable (ordered categorical)

Note:

- The development that follows is about using dummy variables to represent a categorical variable as an explanatory variable in regression.
- That is, it is about coding a categorical variable as an “x-variable” in a regression (LS or logistic).
- This is not about Y-variables.

Multiple Categories

- To represent a categorical variable with multiple categories, we will need multiple dummy variables – one fewer than the number of categories.
- Example: If we have four categories (“red,” “blue,” “green,” and “yellow”), we will need three dummy variables.

Multiple Categories

- We will need to select one of the categories as a “base case.” For example, we might select “yellow” as the base case.
- The three dummy variables will then be used to indicate the presence of the other colors.
- Absent other considerations, I usually select the most frequent case as the base case.

Multiple Categories

D1 = 1 if color is red, 0 otherwise

D2 = 1 if color is blue, 0 otherwise

D3 = 1 if color is green, 0 otherwise.

Multiple Categories

- The following table shows how the dummy variables will encode the four categories:

Case	D1	D2	D3
Red	1	0	0
Blue	0	1	0
Green	0	0	1
Yellow	0	0	0

Multiple Categories

- We will then use the variables D1, D2, and D3 as X-variables in our regression to encode the four color categories.
- Note that the base case (yellow) correspond to all of the dummy variables taking on a value of 0.
- Note that this approach requires that the regression be run with an intercept term.

Understanding Dummy Variables

- To really understand the effects of dummy variables, you need to understand their effects on the regression equations.

Indicator Variable – 2 Categories

- Model: $Y = \alpha + \beta_1 X_1 + \beta_2 I + \varepsilon$
- Reference Group: When Indicator Variable is 0:

$$Y = \alpha + \beta_1 X_1 + \varepsilon$$

- 2nd Category: When Indicator variable is 1

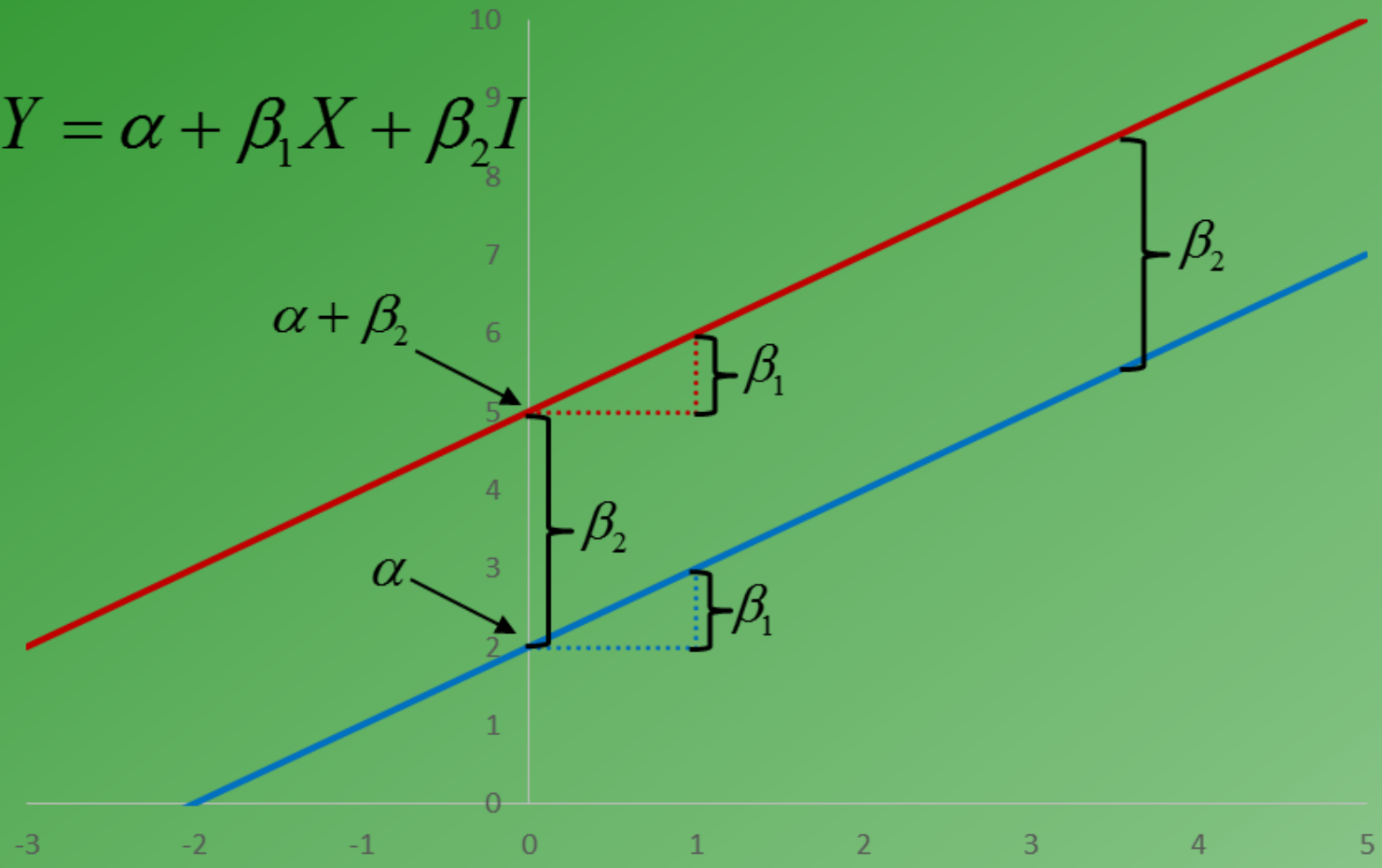
$$Y = \alpha + \beta_1 X_1 + \beta_2 + \varepsilon$$

$$= (\alpha + \beta_2) + \beta_1 X_1 + \varepsilon$$

- We have fit 2 parallel lines!

One Dummy Variable

$$Y = \alpha + \beta_1 X + \beta_2 I$$



Indicator Variable – 3 Categories

- Model: $Y = \alpha + \beta_1 X_1 + \beta_2 I_1 + \beta_3 I_2 + \varepsilon$
- Reference Group: When Both Indicator Variables are 0:

$$Y = \alpha + \beta_1 X_1 + \varepsilon$$

- 2nd Category: When first Indicator variable is 1 and 2nd is 0.

$$Y = \alpha + \beta_1 X_1 + \beta_2 + \varepsilon$$

$$= (\alpha + \beta_2) + \beta_1 X_1 + \varepsilon$$

3 Categories (continued)

- 3rd Category: When first Indicator variable is 0 and 2nd is 1.

$$Y = \alpha + \beta_1 X_1 + \beta_3 + \varepsilon$$

$$= (\alpha + \beta_3) + \beta_1 X_1 + \varepsilon$$

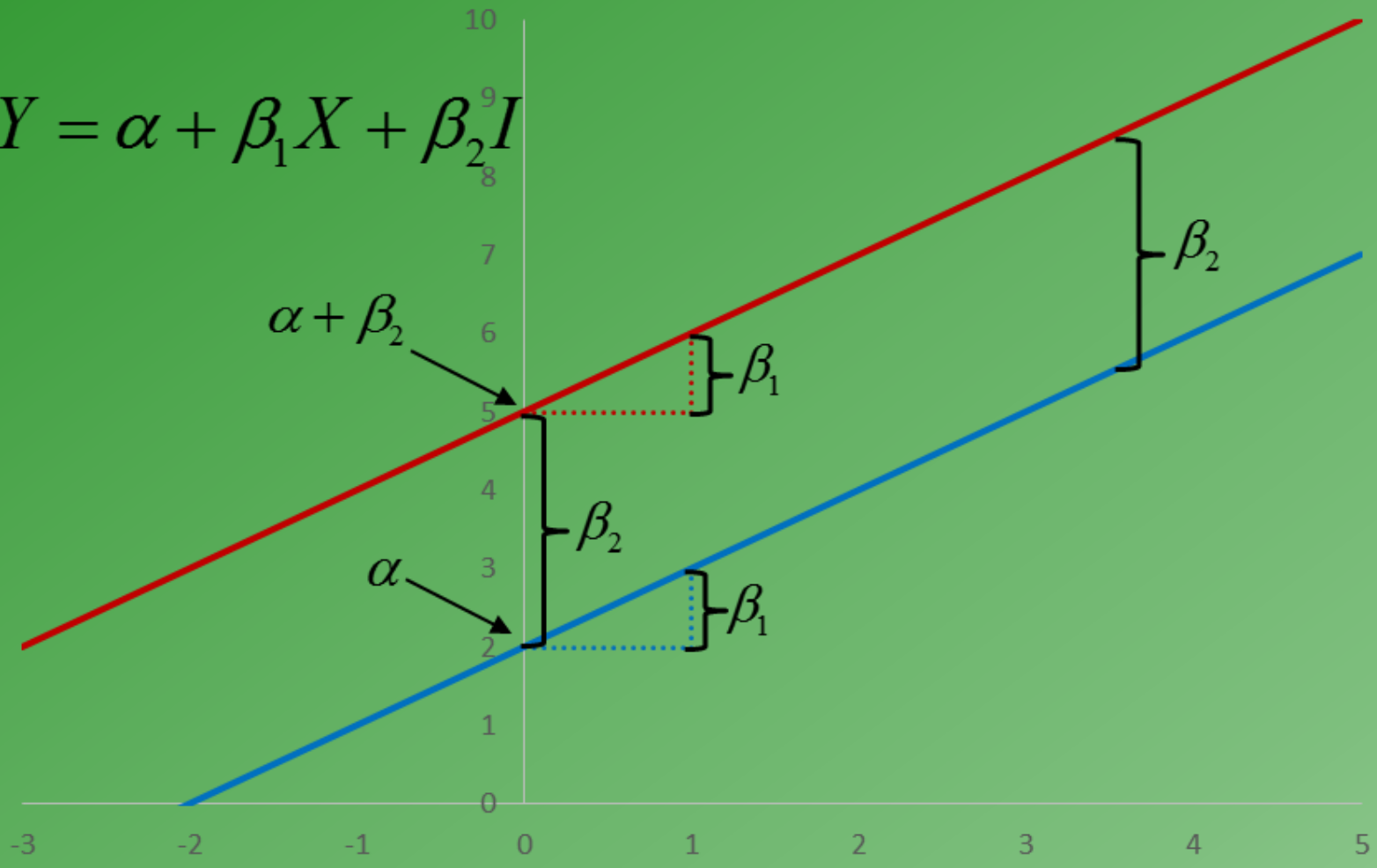
- β_2 (the coefficient on the 1st Indicator variable) is the change in the intercept between the first category (the base case) and the second.

3 Categories (continued)

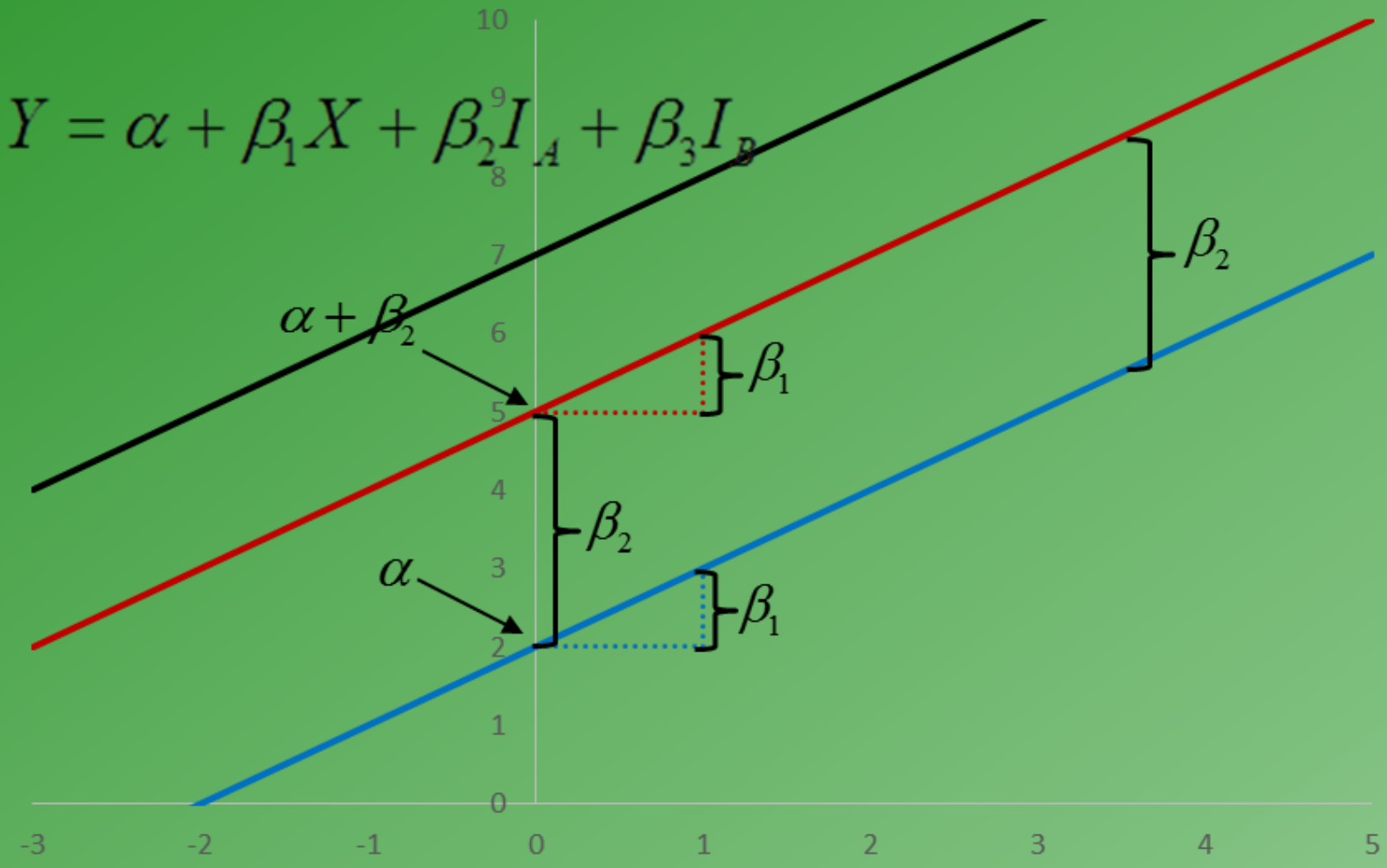
- β_3 (the coefficient on the 2nd Indicator variable) is the change in the intercept between the first category (the base case) and the 3rd category.
- We have fit 3 parallel lines!

One Dummy Variable

$$Y = \alpha + \beta_1 X + \beta_2 I$$

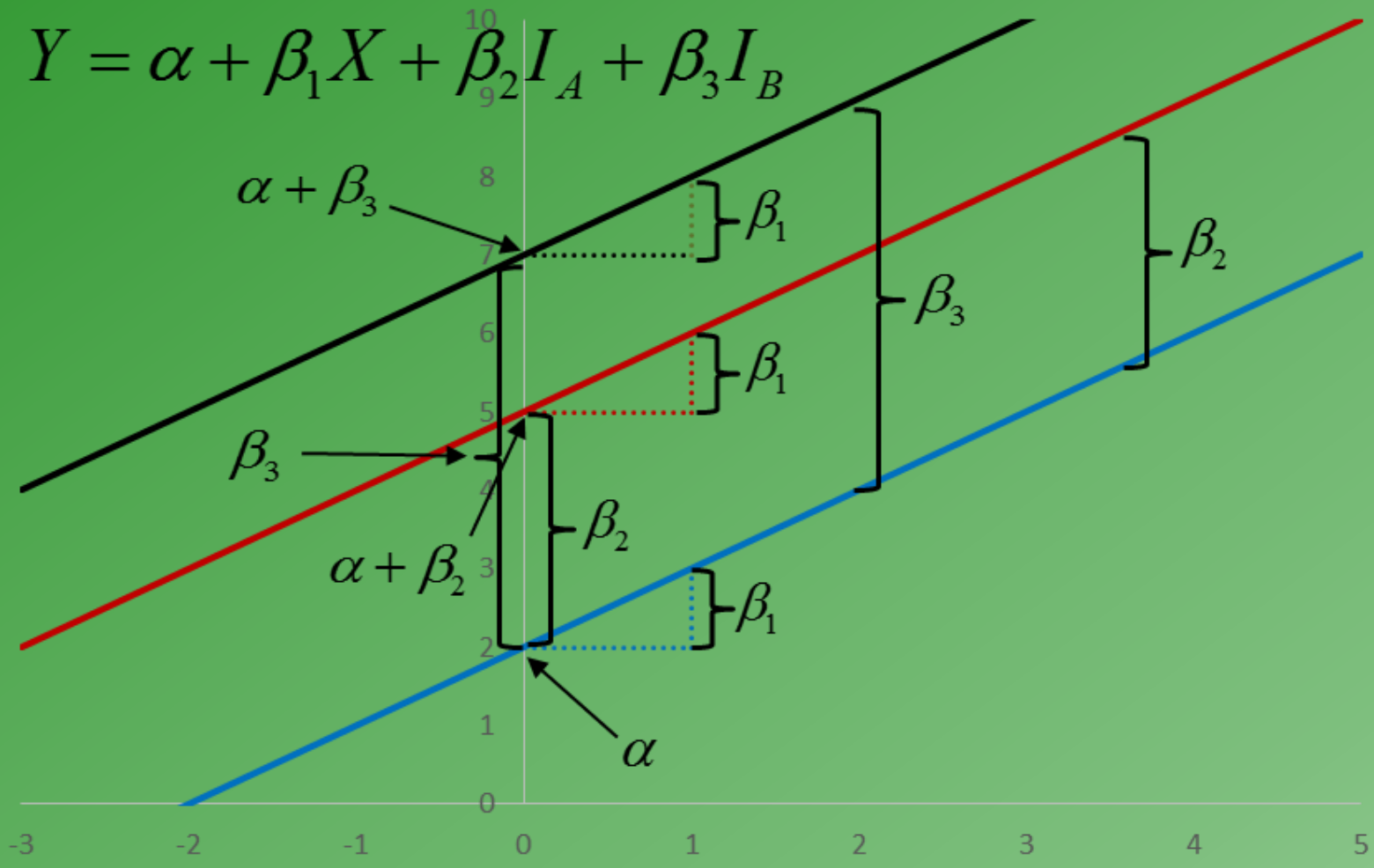


Two Dummy Variables



Two Dummy Variables

$$Y = \alpha + \beta_1 X + \beta_2 I_A + \beta_3 I_B$$



Two Dummy Variables

$$Y = \alpha + \beta_1 X + \beta_2 I_A + \beta_3 I_B$$