York SPIDA John Fox

Notes

Dummy-Variable Regression

1. Topics

- A Dichotomous explanatory variable
- Polytomous Explanatory Variables
- Modeling Interactions
- The Principle of Marginality

2. A Dichotomous Explanatory Variable

- The simplest case: one dichotomous and one quantitative explanatory variable.
- Assumptions:
  - Relationships are *additive* — the partial effect of each explanatory variable is the same regardless of the specific value at which the other explanatory variable is held constant.
  - The other assumptions of the regression model hold.

- The motivation for including a qualitative explanatory variable is the same as for including an additional quantitative explanatory variable:
  - to account more fully for the response variable, by making the errors smaller; and
  - to avoid a biased assessment of the impact of an explanatory variable, as a consequence of omitting another explanatory variables that is related to it.

- Figure 1 represents idealized examples, showing the relationship between education and income among women and men.
  - In both cases, the within-gender regressions of income on education are parallel. Parallel regressions imply additive effects of education and gender on income.
  - In (a), gender and education are unrelated to each other: If we ignore gender and regress income on education alone, we obtain the same slope as is produced by the separate within-gender regressions; ignoring gender inflates the size of the errors, however.
  - In (b) gender and education are related, and therefore if we regress income on education alone, we arrive at a biased assessment of the effect of education on income. The overall regression of income on education has a *negative* slope even though the within-gender regressions have positive slopes.

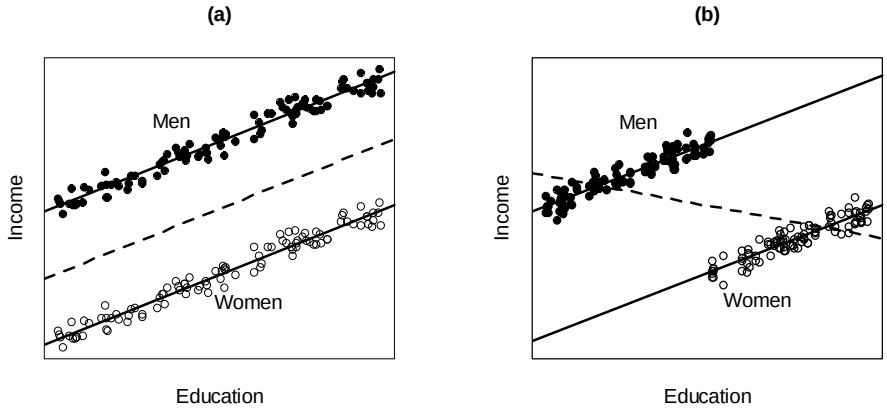

Figure 1. In both cases the within-gender regressions of income on education are parallel: in (a) gender and education are unrelated; in (b) women have higher average education than men.

- We could perform separate regressions for women and men. This approach is reasonable, but it has its limitations:
  - Fitting separate regressions makes it difficult to estimate and test for gender differences in income.
  - Furthermore, if we can assume parallel regressions, then we can more efficiently estimate the common education slope by pooling sample data from both groups.

2.0.1 Introducing a Dummy Regressor

- One way of formulating the common-slope model is

$$Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i$$

where D, called a *dummy-variable regressor* or an *indicator variable*, is coded 1 for men and 0 for women:

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}$$

  - Thus, for women the model becomes

$$Y_i = \alpha + \beta X_i + \gamma(0) + \varepsilon_i = \alpha + \beta X_i + \varepsilon_i$$

  - and for men

$$Y_i = \alpha + \beta X_i + \gamma(1) + \varepsilon_i = (\alpha + \gamma) + \beta X_i + \varepsilon_i$$

- These regression equations are graphed in Figure 2.

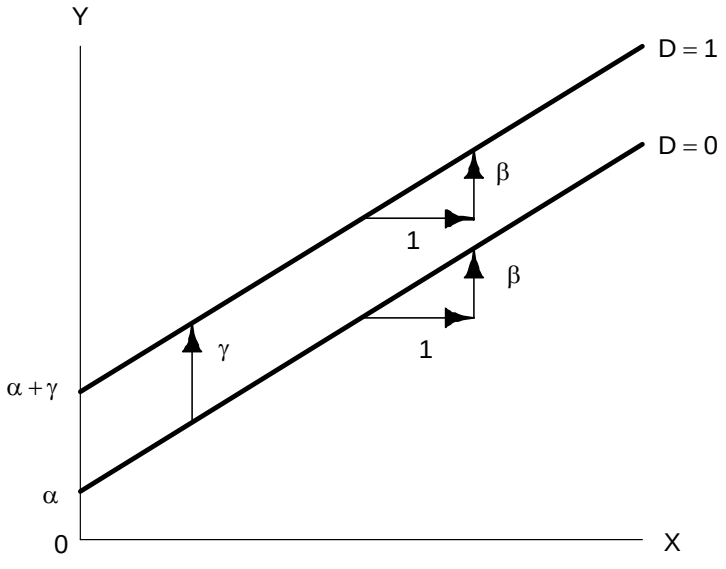

Figure 2. The parameters in the additive dummy-regression model.

2.1 Regressors vs. Explanatory Variables

- This is our initial encounter with an idea that is fundamental to many linear models: the distinction between *explanatory variables* and *regressors.*
  - Here, *gender* is a qualitative explanatory variable (or *factor*), with categories (also called *levels*) *male* and *female*.
  - The dummy variable D is a regressor, representing the explanatory variable gender.
  - In contrast, the quantitative explanatory variable (or *covariate*) *income* and the regressor X are one and the same.
- We will see later that an explanatory variable can give rise to several regressors, and that some regressors are functions of more than one explanatory variable.

2.2 How Dummy Regression Works

- Interpretation of parameters in the additive dummy-regression model:
  - γ gives the difference in intercepts for the two regression lines.
    - Because these regression lines are parallel, γ also represents the constant separation between the lines — the expected income advantage accruing to men when education is held constant.
    - If men were *dis*advantaged relative to women, then γ would be *negative*.
  - α gives the intercept for women, for whom $D = 0$.
  - β is the common within-gender education slope.

- Essentially similar results are obtained if we code D zero for men and one for women (Figure 3):
  - The sign of γ is reversed, but its magnitude remains the same.
  - The coefficient α now gives the income intercept for men.
  - It is therefore immaterial which group is coded one and which is coded zero.

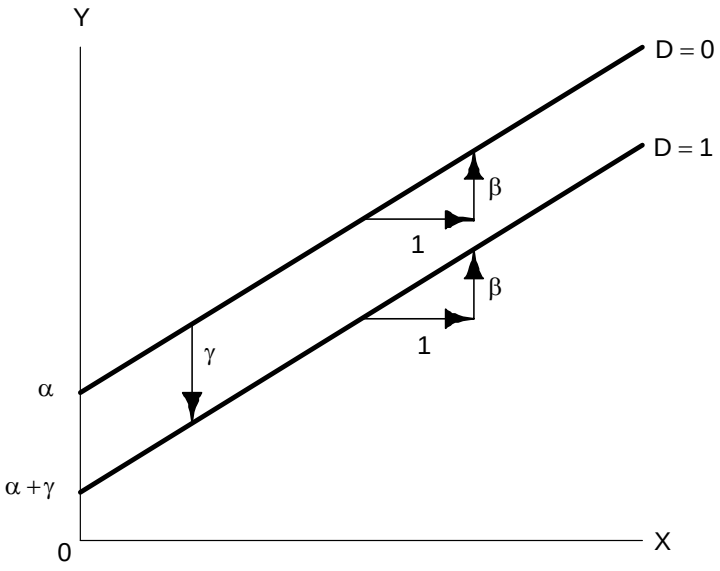

Figure 3. Parameters corresponding to the alternative coding $D = 0$ for men and $D = 1$ for women.

3. Polytomous Explanatory Variables

- Consider the regression of the rated prestige of occupations on their income and education levels.
  - Let us classify the occupations into three categories: (1) professional and managerial; (2) 'white-collar'; and (3) 'blue-collar'.
  - The *three*-category classification can be represented in the regression equation by introducing *two* dummy regressors:

Category	D_2	D_3
Blue Collar	0	0
White Collar	1	0
Professional & Managerial	0	1

  - The regression model is then

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_2 D_{i2} + \gamma_3 D_{i3} + \varepsilon_i$$

    where X_1 is income and X_2 is education.

  - This model describes three parallel regression planes, which can differ in their intercepts (see Figure 4):

$$\begin{aligned}
\text{Blue Collar: } Y_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\
\text{White Collar: } Y_i &= (\alpha + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\
\text{Professional: } Y_i &= (\alpha + \gamma_3) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i
\end{aligned}$$

    - α gives the intercept for blue-collar occupations.
    - γ_2 represents the constant vertical distance between the regression planes for white-collar and blue-collar occupations.
    - γ_3 represents the constant vertical difference between the parallel regression planes for professional and blue-collar occupations (fixing the values of education and income).
  - Blue-collar occupations are coded 0 for both dummy regressors, so 'blue collar' serves as a *baseline* category to which the other occupational categories are compared.

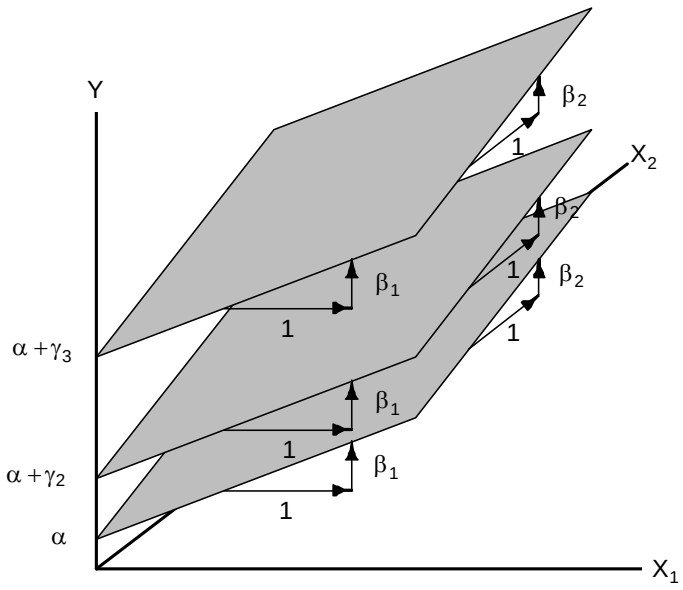

Figure 4. The additive dummy-regression model showing three parallel regression planes.

- The choice of a baseline category is usually arbitrary, for we would fit the same three regression planes regardless of which of the three categories is selected for this role.

▶ Because the choice of baseline is arbitrary, we want to test the null hypothesis of no partial effect of occupational type,

$$H_0\text{: } \gamma_2 = \gamma_3 = 0$$

but the individual hypotheses $H_0\text{:}\,\gamma_2 = 0$ and $H_0\text{:}\,\gamma_3 = 0$ are of less interest.

- The hypothesis $H_0\text{: } \gamma_2 = \gamma_3 = 0$ can be tested by the incremental-sum-of-squares approach, removing D_2 and D_3 from the model.

- For a polytomous explanatory variable with m categories, we code $m-1$ dummy regressors.
  - One simple scheme is to select the first category as the baseline, and to code $D_{ij} = 1$ when observation i falls in category j, and 0 otherwise, for $j = 2, \ldots, m$:

Category	D_2	D_3	$\cdots$	D_m
1	0	0	$\cdots$	0
2	1	0	$\cdots$	0
.	.	.		.
.	.	.		.
.	.	.		.
m	0	0	$\cdots$	1

  - To test the hypothesis that the effects of a qualitative explanatory variable are nil, delete its dummy regressors from the model and compute an incremental F-test.

4. Modeling Interactions

- Two explanatory variables *interact* in determining a response variable when the partial effect of one depends on the value of the other.
  - Additive models specify the absence of interactions.
  - If the regressions in different categories of a qualitative explanatory variable are not parallel, then the qualitative explanatory variable interacts with one or more of the quantitative explanatory variables.
  - The dummy-regression model can be modified to reflect interactions.
- Consider the hypothetical data in Figure 5 (and contrast these examples with those shown in Figure 1, where the effects of gender and education were additive):
  - In (a), gender and education are independent, since women and men have identical education distributions.
  - In (b), gender and education are related, since women, on average, have higher levels of education than men.

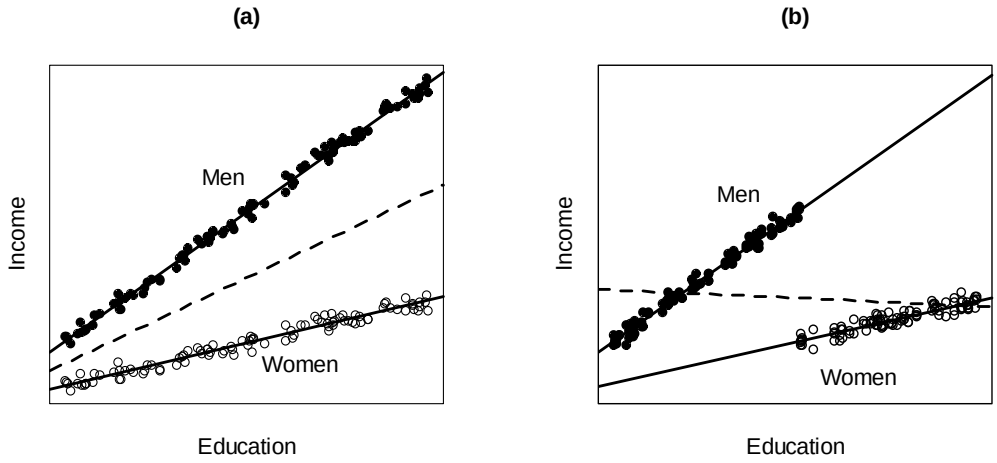

Figure 5. In both cases, gender and education interact in determining income. In (a) gender and education are independent; in (b) women on average have more education than men.

- In both (a) and (b), the within-gender regressions of income on education are not parallel — the slope for men is larger than the slope for women.
  - Because the effect of education varies by gender, education and gender interact in affecting income.
- It is also the case that the effect of gender varies by education. Because the regressions are not parallel, the relative income advantage of men changes with education.
  - *Interaction is a symmetric concept — the effect of education varies by gender, and the effect of gender varies by education.*

- These examples illustrate another important point: *Interaction* and *correlation* of explanatory variables are empirically and logically distinct phenomena.
  - Two explanatory variables can interact *whether or not* they are related to one-another statistically.
  - Interaction refers to the manner in which explanatory variables combine to affect a response variable, not to the relationship between the explanatory variables themselves.

4.1 Constructing Interaction Regressors

- We could model the data in the example by fitting separate regressions of income on education for women and men.
  - A combined model facilitates a test of the gender-by-education interaction, however.
  - A properly formulated unified model that permits different intercepts and slopes in the two groups produces the same fit to the data as separate regressions.
- The following model accommodates different intercepts and slopes for women and men:

$$Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$$

  - Along with the dummy regressor D for gender and the quantitative regressor X for education, I have introduced the *interaction regressor* XD.

- The interaction regressor is the *product* of the other two regressors: XD is a function of X and D, but it is not a *linear* function, avoiding perfect collinearity.
- For women,

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma(0) + \delta(X_i \cdot 0) + \varepsilon_i \\ &= \alpha + \beta X_i + \varepsilon_i \end{aligned}$$

- and for men,

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma(1) + \delta(X_i \cdot 1) + \varepsilon_i \\ &= (\alpha + \gamma) + (\beta + \delta) X_i + \varepsilon_i \end{aligned}$$

▶ These regression equations are graphed in Figure 6:

- α and β are the intercept and slope for the regression of income on education among women.
- γ gives the *difference* in intercepts between the male and female groups
- δ gives the *difference* in slopes between the two groups.

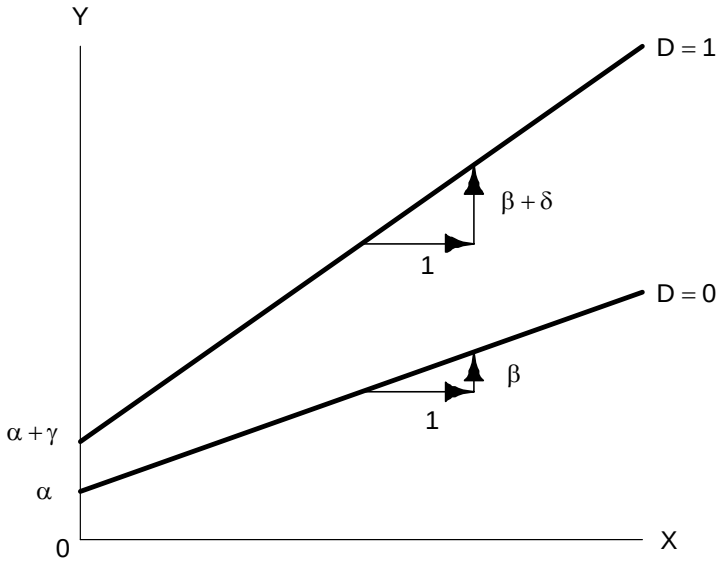

Figure 6. The parameters in the dummy-regression model with interaction.

  - To test for interaction, we can test the hypothesis H_0: $\delta = 0$.

- In the additive, no-interaction model, γ represented the unique partial effect of gender, while the slope β represented the unique partial effect of education.
  - In the interaction model, γ is no longer interpretable as the unqualified income difference between men and women of equal education — γ is now the income difference at $X = 0$.
  - Likewise, in the interaction model, β is not the unqualified partial effect of education, but rather the effect of education among women.
    - The effect of education among men ($\beta + \delta$) does not appear directly in the model.
- Extension to polytomous factors is straight-forward.

5. The Principle of Marginality

- The separate partial effects, or *main effects*, of education and gender are *marginal* to the education-by-gender interaction.
- In general, we neither test nor interpret main effects of explanatory variables that interact.
  - If we can rule out interaction either on theoretical or empirical grounds, then we can proceed to test, estimate, and interpret main effects.
- It does not generally make sense to specify and fit models that include interaction regressors but that delete main effects that are marginal to them.
  - Such models — which violate the *principle of marginality* — are interpretable, but they are not broadly applicable.

- Consider the model

$$Y_i = \alpha + \beta X_i + \delta(X_i D_i) + \varepsilon_i$$

  - As shown in Figure 7 (a), this model describes regression lines for women and men that have the same intercept but (potentially) different slopes, a specification that is peculiar and of no substantive interest.

- Similarly, the model

$$Y_i = \alpha + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$$

  graphed in Figure 7 (b), constrains the slope for women to 0, which is needlessly restrictive.

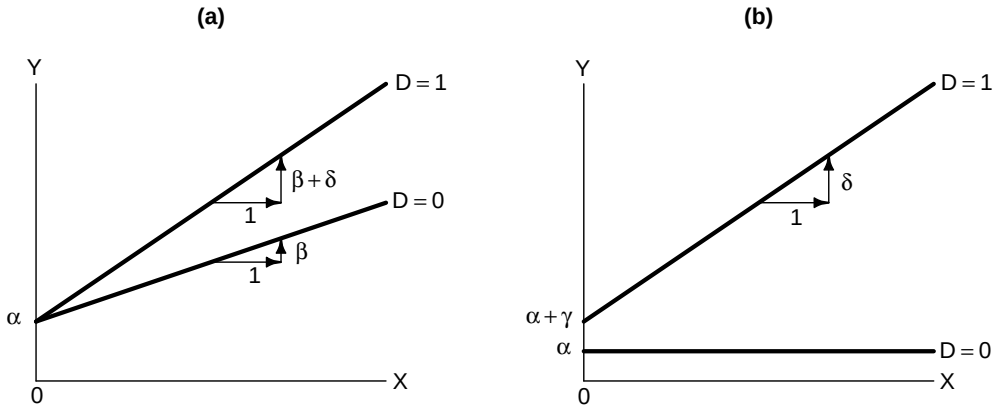

Figure 7. Two models that violate the principle of marginality, by including the interaction regressor XD but (a) omitting D or (b) omitting X.