Categorical Predictors

Goals

- To show how dummy indicator variables can be used to represent the categories of a categorical predictor variable in a regression model.
- To introduce the concept of interaction between predictor variables, and to show how interactions can be incorporated into a regression model by forming interaction terms.
- To introduce the principle of marginality, which serves as a guide to constructing and testing terms in complex linear models.
- To show how incremental F-tests are employed to test terms in dummy regression models.

Binary Predictor Variables

The simplest case: one binary and one quantitative (i.e., continuous) predictor 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 categorical predictor variable is the same as for including an additional quantitative predictor variable:

- to account more fully for the outcome variable, by making the errors smaller; and
- to avoid a biased assessment of the impact of an explanatory variable, as a consequence of omitting other explanatory variables that are related to it.

In (a), sex and education are unrelated to each other: If we ignore sex and regress income on education alone, we obtain the same slope as is produced by the separate within-sex regressions; ignoring sex inflates the size of the errors, however.

In (b) sex 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-sex regressions have positive slopes.

In both cases, the within-sex regressions of income on education are parallel. Parallel regressions imply additive effects of education and sex on income.

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 sex 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.

One way of formulating the common slope model is

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

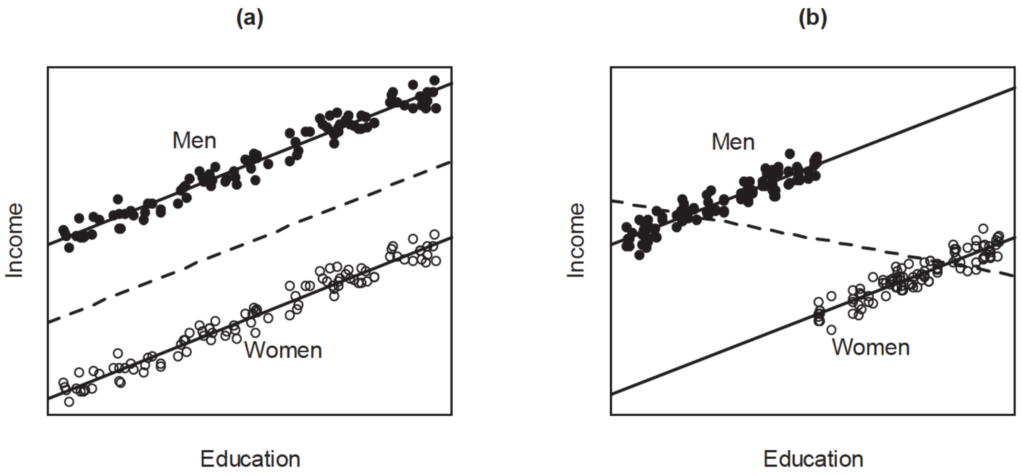

Figure 1: Suppose that we are interested in relationship between education and income among women and men.

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

$$Y_i = \alpha + \beta X_i + \gamma \times 0 + \epsilon_i = \alpha + \beta X_i + \epsilon_i$$

and for men

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

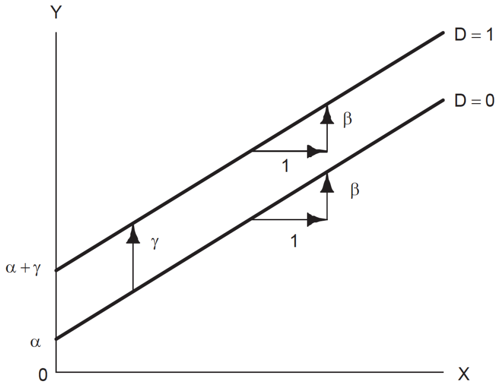

Figure 2: .

Essentially similar results are obtained if we code D zero for men and one for women:

- 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.

This method can be applied to any number of quantitative variables, as long as we are willing to assume that the slopes are the same in the two categories of the dichotomous explanatory variable (i.e., parallel regression surfaces):

$$Y_i = \alpha + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \gamma D_i + \epsilon_i$$

For $D = 0$

$$Y_i = \alpha + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \epsilon_i$$

For $D = 1$

$$Y_i = (\alpha + \gamma) + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \epsilon_i$$

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, sex is a qualitative explanatory variable, with categories male and female.

The dummy variable, D is a regressor, representing the explanatory variable sex.

In contrast, the quantitative explanatory variable income and the regressor X are one and the same.

We will see later that some regressors are functions of more than one explanatory variable when we discuss interactions.

More than two categories

An explanatory variable can give rise to several regressors - for example, if there more than two categories. A three-category variable can be represented with *two* dummy variables.

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i$$

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.

In general, for a categorical variable with C categories, $C - 1$ indicator variables are needed. Categorical variables may be *ordinal*, such as physical activity represented on a 5 point Likert scale (e.g., 1 = "much less active", 2 = "somewhat less active", 3="about as active", 4 = "somewhat more active", and 5 = "much more active"), or they may be *nominal*, meaning that there is no intrinsic ordering of the categories (e.g., race). Although ordinal categorical variables are meaningfully ordered, the increments between the categories may not be the same as the numbers used to represent them. Categories are usually set up to be mutually exclusive and exhaustive so that each individual belongs to one and only category, thereby defining subgroups of individuals.

If there are no other predictors in the model, a regression of an outcome on a binary categorical variable is equivalent to a t-test and a regression of an outcome on a categorical variable with more than two categories is equivalent to a one-way analysis of variance (ANOVA).

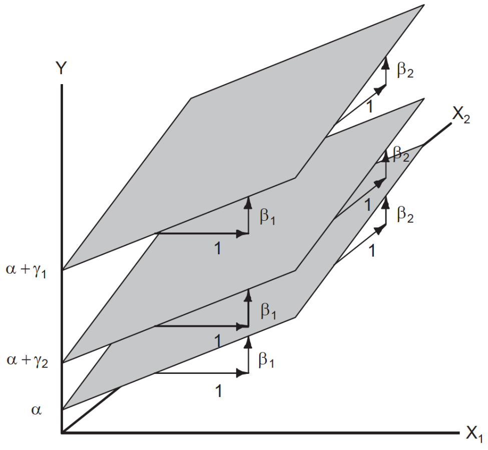

Figure 3: Three-categories

ANOVA

ANOVA describes the partition of the response variable sum of squares in a linear model into ‘explained’ and ‘unexplained’ components.

The term also refers to procedures for fitting and testing linear models in which the predictor variables are categorical.

- A single categorical predictor variable corresponds to one-way ANOVA;
- two categorical predictor variables to two-way ANOVA;
- three categorical predictor variables to three-way ANOVA;
- and so on.

When any of these models also contains a continuous predictor variable, it is called analysis of covariance (ANCOVA).

Interactions

Two variables interact in predicting an outcome variable when the partial effect of one depends on the value of the other (i.e., differs according to the levels of the other). Interaction is also referred to as effect modification and moderation.

Additive models specify the absence of interactions.

If the regression lines in different categories of a categorical predictor variable are **not** parallel, then the categorical predictor variable interacts with one or more of the quantitative predictor variables.

The regression model can be modified to reflect interactions.

Consider the hypothetical data above (and contrast these examples with those shown previously where the effects of sex and education were additive):

In (a), sex and education are independent, since women and men have identical education distributions.

In (b), sex and education are related, since women, on average, have higher levels of education than men.

In both (a) and (b) in the first figure, the within-sex 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 sex, education and sex interact in affecting income.

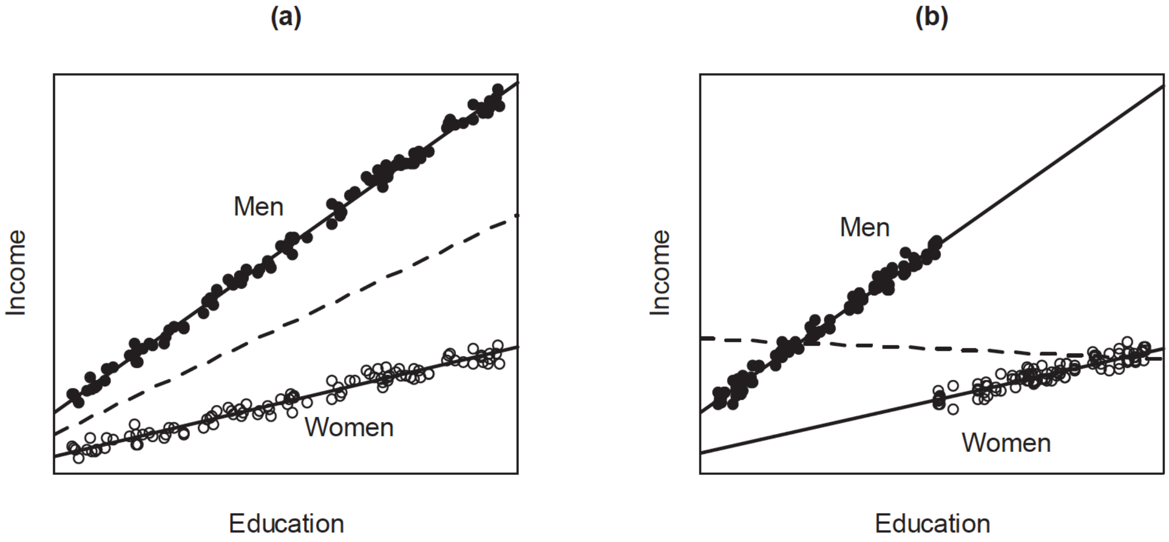

Figure 4: Interaction

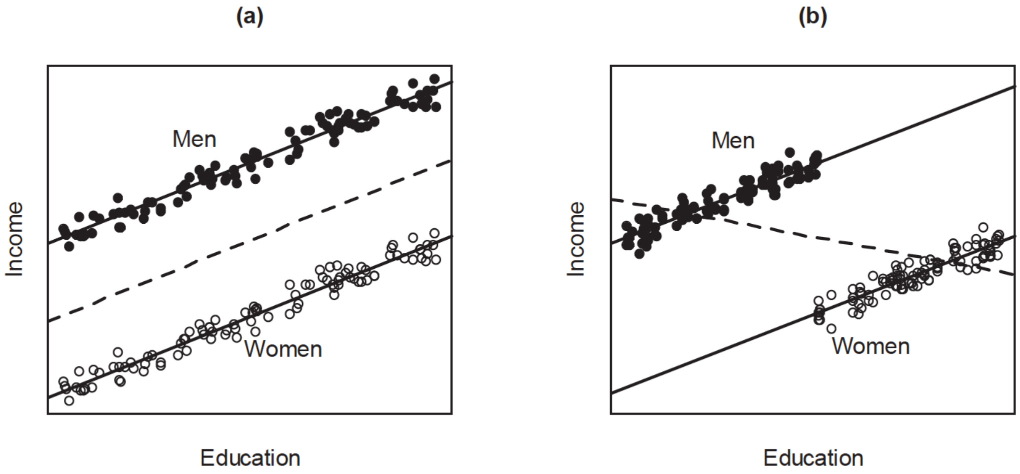

Figure 5: Relationship between education and income among women and men.

It is also the case that the effect of sex 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 sex, and the effect of sex varies by education.

These examples illustrate another important point: Interaction and correlation of predictor variables are empirically and logically distinct phenomena.

Two predictor variables can interact whether or not they are related to one another statistically. However, interaction can be difficult to detect if the interacting variables are highly correlated.

Interaction refers to the manner in which predictor variables combine to affect an outcome variable, not to the relationship between the predictor variables themselves.

Constructing Interaction Regressors

In modeling the HELP data, we could fit separate regressions of depression on age for women and men. But a combined model facilitates a test of the sex-by-age interaction.

A properly formulated unified model that permits different intercepts and slopes in the two groups produces the same fit as separate regressions.

The following model accommodates different intercepts and slopes for women and men:

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

Along with the dummy variable for sex, D, and the quantitative age predictor, X, there is an interaction term, XD, that 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 men, $D = 0$, thus the regression equation for predicting Y_i is:

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

For women, $D = 1$, thus the regression equation for predicting Y_i is:

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

Thus, an interaction term allows the *slope* in the two groups to differ.

The method of modeling interactions by forming product terms is easily extended to variables with more than two categories, to several categorical predictor variables, and to several quantitative predictor variables.

In interpreting interactions, it is helpful to write out the regression equation for one predictor variable for various values of the other predictor variable(s) that make up the interaction. If one of the variables is dichotomous and the other is continuous, then it is easiest to write the regression equation for the continuous variable for each value of the dichotomous variable. If both are continuous, then representative values for one of the variables can be chosen, such as the mean, maximum, minimum, +/- 1 std. deviation or the median, 25th, and 75th percentiles, and the regression model can be written for the other continuous variable at these chosen values. In addition to writing out the regression equation, it is helpful to plot the interaction.

Marginal Effects

The separate partial effects, or so-called main effects, of age and sex are marginal to the age-by-sex interaction.

It does not generally make sense to specify and fit models that include interaction terms but that delete main effects that are marginal to them.

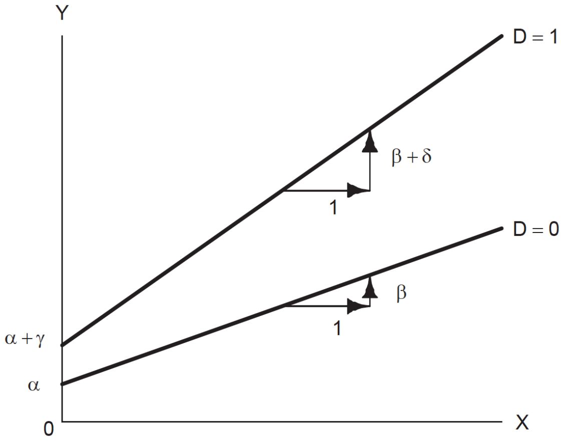

Figure 6: Parameters in a dummy variable regression with an interaction

Such models — which violate the so-called principle of marginality — are interpretable, but they are not broadly applicable.

Centering

It is useful to apply a linear transformation to the predictor variables so that zero is a legitimate, observable value. Otherwise, the intercept estimate is not interpretable. This transformation is called centering.

The mean is subtracted from all values of a variable. That is, compute the sample mean for the variable and subtract it from each observation.

The intercept is then interpretable as the expected value of the outcome variable when all predictor variables have their mean value.

Standardization includes not only mean centering but also divides by the standard deviation of the predictor variable to achieve a standard deviation of one. However, standardization will affect the interpretation of the regression coefficients and residual variances as well as the interpretation of the intercept whereas centering only affects the interpretation of the intercept.

The regression coefficients for variables that make up an interaction have a different meaning in the model with the interaction term than in a model without it. They are interpreted as the expected value of the regression slope for the case that the other variable is zero and vice versa. If one of the variables that make up the interaction does not have a meaningful zero in the data, then the regression coefficient for the other variable has no substantive interpretation. Thus, quantitative predictor variables that make up an interaction should be mean centered. In this case, the regression coefficient of one of the variables in an interaction can be interpreted as the expected value of the regression slope for individuals with an 'average' score on the other variable.

Also, note that when the variables that make up the interaction are mean centered, the regression coefficients for the individual variables do not change when an interaction term is added to the model.

Another reason that it can be important to mean center the variables that make up an interaction is computational stability. If the variables that make up the interaction have very different scales, then when multiplied together, the variance of the interaction term can become very large and result in convergence problems.

Confounding

A third variable, that is not the outcome variable or the primary predictor variable of interest, distorts the observed relationship between the predictor and outcome. Confounding complicates analyses owing to the presence of a third variable that is associated with both the primary predictor variable of interest and the outcome.

Criteria for a confounder:

1. A confounder must be a risk factor (or protective factor) for the outcome of interest. That is, it must be associated with the outcome.
2. A confounder must be associated with the main predictor variable of interest.
3. A confounder must not be an intermediate step in the causal pathway between the predictor and outcome.

The coefficient for the effect of the primary predictor on the outcome changes when a potential confounder is added to the model. **However, for logistic, Cox, and some other regression models, analogous changes are seen when nonconfounders associated with the outcome and not the predictor of interest are added to the model.**

Confounders often explain some of the association between a predictor of interest and an outcome, so the adjusted effect is often weaker than the unadjusted effect. But this is not always the case - the adjusted effect might be stronger than the unadjusted effect. We may find little or no association in the unadjusted analysis because it is masked or negatively confounded by another predictor.

Negative confounding can occur under the following circumstances:

- The predictors are inversely correlated, but have regression coefficients with the same sign.
- The predictors are positively correlated, but have regression coefficients with the opposite sign.

For the multivariable linear regression model to effectively control confounding and estimate unbiased causal effects, all potential confounders must have been:

- recognized and measured without error in the study.
- accurately represented in the systematic part of the model.

Of course, it is impossible to know if all confounders have been measured.

Measurement error can arise if the study has only measured proxies for the confounders.

Confounding by indication occurs when prognostic factors lead physicians to prescribe a particular drug or treatment.