Sociology 740 John Fox

Lecture Notes

5. Dummy-Variable Regression and Analysis of Variance

Copyright © 2014 by John Fox

1. Introduction

- One of the limitations of multiple-regression analysis is that it accommodates only quantitative explanatory variables.
- *Dummy-variable regressors* can be used to incorporate qualitative explanatory variables into a linear model, substantially expanding the range of application of regression analysis.

2. Goals:

- To show how dummy regessors can be used to represent the categories of a qualitative explanatory variable in a regression model.
- To introduce the concept of interaction between explanatory variables, and to show how interactions can be incorporated into a regression model by forming interaction regressors.
- 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.
- To show how analysis-of-variance models can be handled using dummy variables.

3. 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.

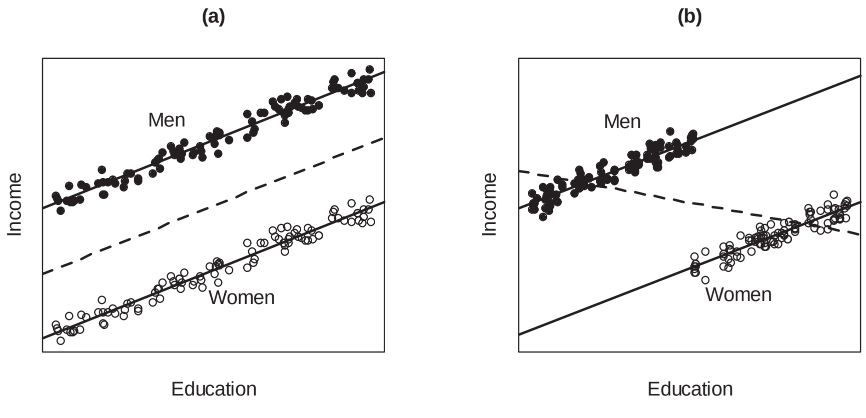

Figure 1. In both cases the within-gender regressions of income on education are parallel: in (a) gender and education are unrelated; in (b) women have higher average education than men.

- 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.

3.1 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.

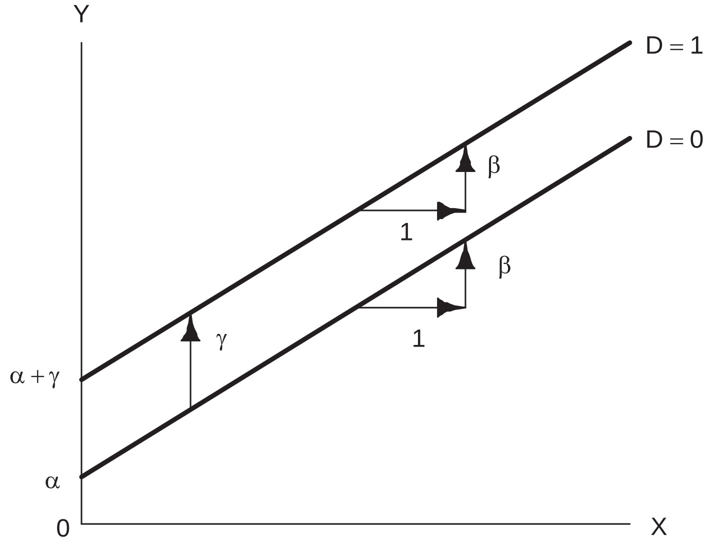

Figure 2. The parameters in the additive dummy-regression model.

© 2014 by John Fox

3.2 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, with categories *male* and *female*.
- The dummy variable D is a regressor, representing the explanatory variable gender.
- In contrast, the quantitative explanatory variable *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.

© 2014 by John Fox

- 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.
- 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):

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma D_i + \varepsilon_i$$

  - For $D = 0$ we have

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

  - and for $D = 1$

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

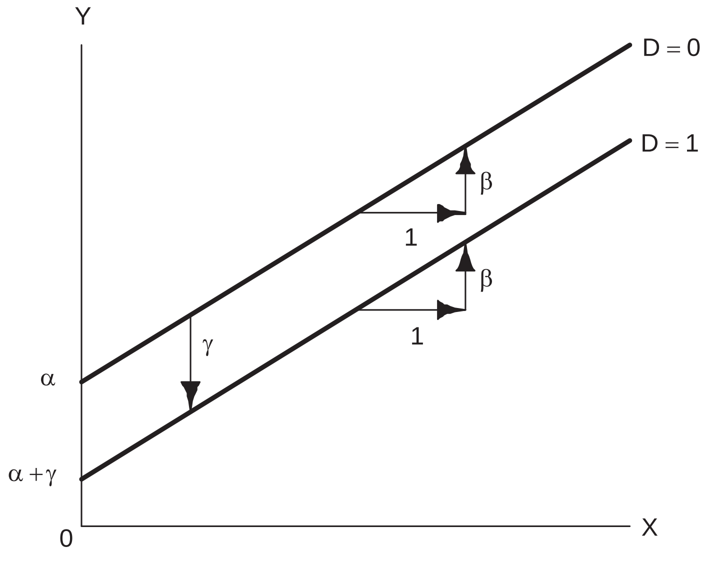

Figure 3. Parameters corresponding to alternative coding $D = 0$ for men and $D = 1$ for women.

4. Polytomous Explanatory Variables

▶ Recall the regression of the rated prestige of 102 Canadian occupations on their income and education levels.

- I have classified 98 of 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_1	D_2
Professional & Managerial	1	0
White Collar	0	1
Blue Collar	0	0

- The regression model is then

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

  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):

$$\begin{array}{rrl}
\text{Professional:} & Y_i = & (\alpha + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\
\text{White Collar:} & Y_i = & (\alpha + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\
\text{Blue Collar:} & Y_i = & \qquad\quad \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i
\end{array}$$

  – α gives the intercept for blue-collar occupations.

  – γ_1 represents the constant vertical difference between the parallel regression planes for professional and blue-collar occupations (fixing the values of education and income).

  – γ_2 represents the constant vertical distance between the regression planes for white-collar and blue-collar occupations.

- Blue-collar occupations are coded 0 for both dummy regressors, so 'blue collar' serves as a *baseline* category with which the other occupational categories are compared.

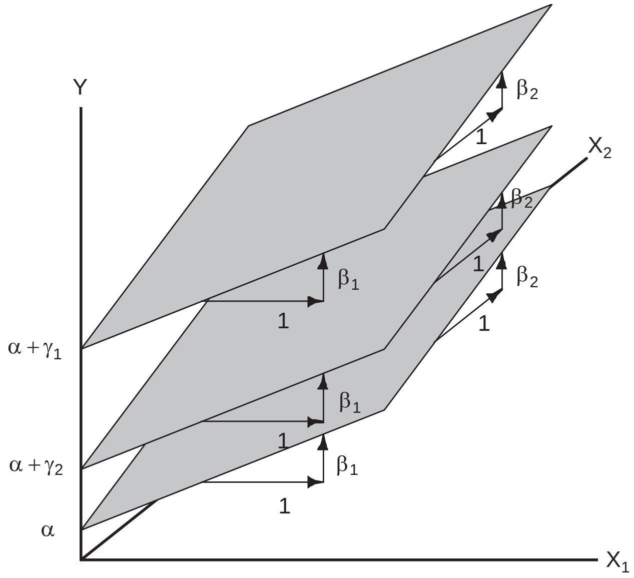

Figure 4. The additive dummy-regression model showing three parallel regression planes.

- 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,

$$H_0\colon \gamma_1 = \gamma_2 = 0$$

but the individual hypotheses $H_0\colon \gamma_1 = 0$ and $H_0\colon \gamma_2 = 0$ are of less interest.

- The hypothesis $H_0\colon \gamma_1 = \gamma_2 = 0$ can be tested by the incremental-sum-of-squares approach.

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

Category	D_1	D_2	$\cdots$	D_{m-1}
1	1	0	$\cdots$	0
2	0	1	$\cdots$	0
.	.	.		.
.	.	.		.
.	.	.		.
$m-1$	0	0	$\cdots$	1
m	0	0	$\cdots$	0

  - When there is more than one qualitative explanatory variable with additive effects, we can code a set of dummy regressors for each.
  - 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.

- The regression of prestige on income and education

$$\begin{array}{cccccccc} \widehat{Y} = & -7.621 & + & 0.001241X_1 & + & 4.292X_2 & \qquad & R^2 = .81400 \\ & (3.116) & & (0.000219) & & (0.336) & & \end{array}$$

- Inserting dummy variables for type of occupation into the regression equation produces the following results:

$$\begin{array}{l}\widehat{Y} = \underset{(5.2275)}{-0.6229} + \underset{(0.000221)}{0.001013X_1} + \underset{(0.641)}{3.673X_2} + \underset{(3.867)}{6.039D_1} - \underset{(2.514)}{2.737D_2} \\ R^2 = .83486\end{array}$$

- The three fitted regression equations are:

Professional:	$\widehat{Y} = 5.416 + 0.001013X_1 + 3.673X_2$
White collar:	$\widehat{Y} = -3.360 + 0.001013X_1 + 3.673X_2$
Blue collar:	$\widehat{Y} = -0.623 + 0.001013X_1 + 3.673X_2$

- To test the null hypothesis of no partial effect of type of occupation,

$$H_0\colon \gamma_1 = \gamma_2 = 0$$

calculate the incremental F-statistic

$$\begin{aligned} F_0 &= \frac{n-k-1}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2} \\ &= \frac{98-4-1}{2} \times \frac{.83486 - .81400}{1 - .83486} = 5.874 \end{aligned}$$

with 2 and 93 degrees of freedom, for which $p = .0040$.

5. 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.

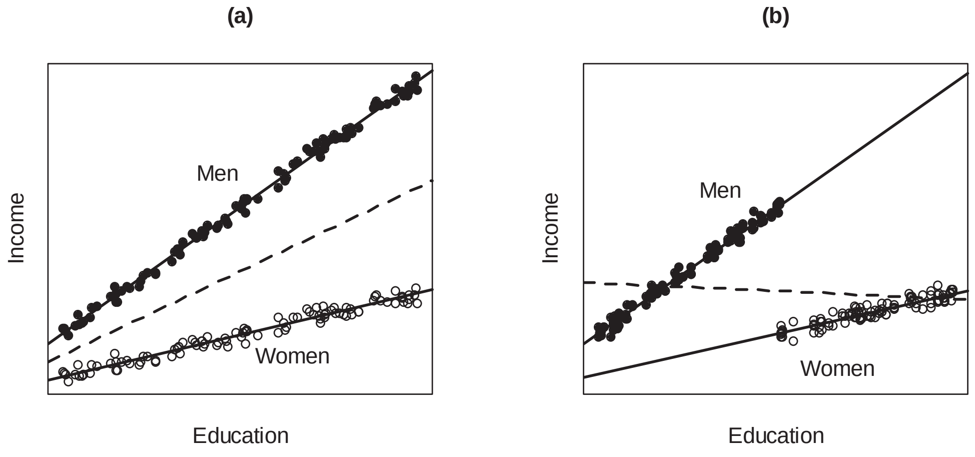

Figure 5. In both cases, gender and education interact in determining income. In (a) gender and education are independent; in (b) women on average have more education than men.

- 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.

5.1 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 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.

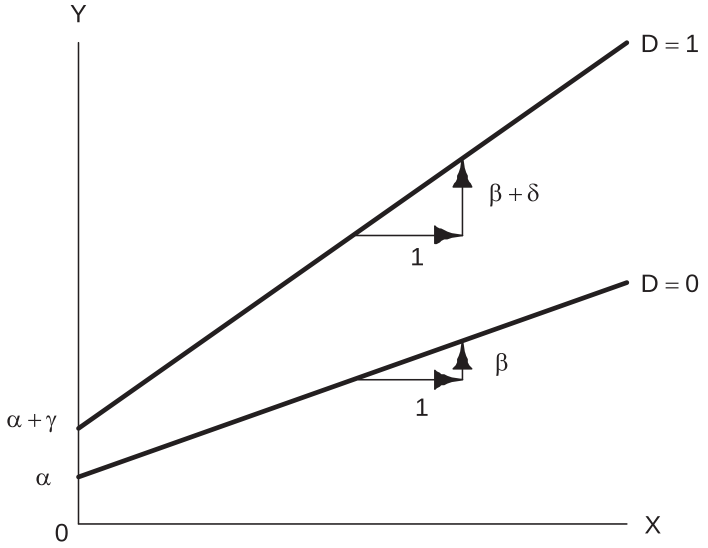

Figure 6. The parameters in the dummy-regression model with interaction.

© 2014 by John Fox

   - 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.

© 2014 by John Fox

5.2 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.

5.3 Interactions With Polytomous Explanatory Variables

- The method of modeling interactions by forming product regressors is easily extended to polytomous explanatory variables, to several qualitative explanatory variables, and to several quantitative explanatory variables.
- For example, for the Canadian occupational prestige regression:

$$\begin{aligned} Y_i = \; & \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} \\ & + \delta_{11} X_{i1} D_{i1} + \delta_{12} X_{i1} D_{i2} \\ & + \delta_{21} X_{i2} D_{i1} + \delta_{22} X_{i2} D_{i2} + \varepsilon_i \end{aligned}$$

  - We require one interaction regressor for each product of a dummy regressor with a quantitative explanatory variable.

- – The regressors X_1D_1 and X_1D_2 capture the interaction between income and occupational type;
- – X_2D_1 and X_2D_2 capture the interaction between education and occupational type.

- The model permits different intercepts and slopes for the three types of occupations:

$$\begin{array}{rrlll}
\text{Professional:} & Y_i = & (\alpha+\gamma_1) & + & (\beta_1+\delta_{11})X_{i1} \\
 & + & (\beta_2+\delta_{21})X_{i2} & + & \varepsilon_i \\
\text{White Collar:} & Y_i = & (\alpha+\gamma_2) & + & (\beta_1+\delta_{12})X_{i1} \\
 & + & (\beta_2+\delta_{22})X_{i2} & + & \varepsilon_i \\
\text{Blue Collar:} & Y_i = & \alpha & + & \beta_1 X_{i1} \\
 & + & \beta_2 X_{i2} & + & \varepsilon_i
\end{array}$$

- Blue-collar occupations, coded 0 for both dummy regressors, serve as the baseline for the intercepts and slopes of the other occupational types.

- Fitting this model to the Canadian occupational prestige data produces the following results:

$$\begin{array}{rl}
\widehat{Y}_i = & \underset{(7.057)}{2.276} + \underset{(0.000556)}{0.003522X_1} + \underset{(0.927)}{1.713X_2} \\
 & + \underset{(13.72)}{15.35D_1} - \underset{(17.54)}{33.54D_2} \\
 & - \underset{(0.000599)}{0.002903X_1D_1} - \underset{(0.000894)}{0.002072X_1D_2} \\
 & + \underset{(1.289)}{1.388X_2D_1} + \underset{(1.757)}{4.291X_2D_2}
\end{array}$$

$$R^2 = .8747$$

- The regression equation for each group:

$$\begin{array}{lrl}
\text{Professional:} & \widehat{\text{Prestige}} = & 17.63 + 0.000619 \times \text{Income} + 3.101 \times \text{Education} \\
\text{White-Collar:} & \widehat{\text{Prestige}} = & -31.26 + 0.001450 \times \text{Income} + 6.004 \times \text{Education} \\
\text{Blue-Collar:} & \widehat{\text{Prestige}} = & 2.276 + 0.003522 \times \text{Income} + 1.713 \times \text{Education}
\end{array}$$

© 2014 by John Fox

5.4 Hypothesis Tests for Main Effects and Interactions

- To test the null hypothesis of no interaction between income and type, H_0: $\delta_{11} = \delta_{12} = 0$, we need to delete the interaction regressors $X_1 D_1$ and $X_1 D_2$ from the full model and calculate an incremental F-test.
  - Likewise, to test the null hypothesis of no interaction between education and type, H_0: $\delta_{21} = \delta_{22} = 0$, we delete the interaction regressors $X_2 D_1$ and $X_2 D_2$ from the full model.
  - These tests, and tests for the main effects of occupational type, income, and education, are detailed in the following tables:

© 2014 by John Fox

			Regression	
Model	Terms	Parameters	Sum of Squares	df
1	$I, E, T, I \times T, E \times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2,$ $\delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}$	24,794.	8
2	$I, E, T, I \times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2,$ δ_{11}, δ_{12}	24,556.	6
3	$I, E, T, E \times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2,$ δ_{21}, δ_{22}	23,842.	6
4	I, E, T	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2$	23,666.	4
5	I, E	α, β_1, β_2	23,074.	2
6	$I, T, I \times T$	$\alpha, \beta_1, \gamma_1, \gamma_2,$ δ_{11}, δ_{12}	23,488.	5
7	$E, T, E \times T$	$\alpha, \beta_2, \gamma_1, \gamma_2,$ δ_{21}, δ_{22}	22,710.	5

	Models	Sum of			
Source	Contrasted	Squares	df	F	p
Income	$3-7$	1132.	1	28.35	$<$.0001
Education	$2-6$	1068.	1	26.75	$<$.0001
Type	$4-5$	592.	2	7.41	$<$.0011
Income × Type	$1-3$	952.	2	11.92	$<$.0001
Education × Type	$1-2$	238.	2	2.98	.056
Residuals		3553.	89		
Total		28,347.	97		

Source	Models	H_0
Income	$3-7$	$\beta_1 = 0 \mid \delta_{11} = \delta_{12} = 0$
Education	$2-6$	$\beta_2 = 0 \mid \delta_{21} = \delta_{22} = 0$
Type	$4-5$	$\gamma_1 = \gamma_2 = 0 \mid \delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0$
Income×Type	$1-3$	$\delta_{11} = \delta_{12} = 0$
Education×Type	$1-2$	$\delta_{21} = \delta_{22} = 0$

- Although the analysis-of-variance table shows the tests for the main effects of education, income, and type before the education-by-type and income-by-type interactions, the logic of interpretation is to examine the interactions first:
  - Conforming to the principle of marginality, the test for each main effect is computed assuming that the interactions that are higher-order relatives of that main effect are 0.
  - Thus, for example, the test for the income main effect assumes that the income-by-type interaction is absent (i.e., that $\delta_{11} = \delta_{12} = 0$), but not that the education-by-type interaction is absent ($\delta_{21} = \delta_{22} = 0$).
  - Tests formulated according to the principle of marginality are sometimes called *Type-II tests*.

© 2014 by John Fox

- The degrees of freedom for the several sources of variation add to the total degrees of freedom, but — because the regressors in different sets are correlated — the sums of squares do not add to the total sum of squares.
  - What is important is that sensible hypotheses are tested, not that the sums of squares add to the total sum of squares.

© 2014 by John Fox

6. Analysis-of-Variance Models

- *Analysis of variance (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 explanatory variables are categorical.
  - A single categorical explanatory variable (*factor* or *classification*) corresponds to *one-way* analysis of variance;
  - two factors to *two-way* analysis of variance;
  - three factors to *three-way* analysis of variance;
  - and so on.
- We will briefly consider one- and two-way ANOVA.

6.1 One-Way ANOVA

- Dummy regressors can be employed to code a one-way ANOVA model.
- For example, for a three-category classification:

$$Y_i = \alpha + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i$$

with

Group	D_1	D_2
1	1	0
2	0	1
3	0	0

- The response variable expectation (population mean) in group j is μ_j.

- Because the error ε has a mean of 0 under the usual linear-model assumptions, taking the expectation of both sides of the model produces the following relationships between group means and model parameters:

$$\begin{array}{ll}\text{Group 1:} & \mu_1 = \alpha + \gamma_1 \times 1 + \gamma_2 \times 0 = \alpha + \gamma_1 \\ \text{Group 2:} & \mu_2 = \alpha + \gamma_1 \times 0 + \gamma_2 \times 1 = \alpha + \gamma_2 \\ \text{Group 3:} & \mu_3 = \alpha + \gamma_1 \times 0 + \gamma_2 \times 0 = \alpha\end{array}$$

  - There are three parameters ($\alpha, \gamma_1,$ and γ_2) and three group means, so we can solve uniquely for the parameters in terms of the group means:

$$\begin{aligned}\alpha &= \mu_3 \\ \gamma_1 &= \mu_1 - \mu_3 \\ \gamma_2 &= \mu_2 - \mu_3\end{aligned}$$

  - Thus α represents the mean of the baseline category (group 3), and γ_1 and γ_2 capture differences between the other group means and the mean of the baseline category.

- One-way analysis of variance focuses on testing for differences among group means.
  - The omnibus F-statistic for the model tests H_0: $\gamma_1 = \gamma_2 = 0$, which corresponds to H_0: $\mu_1 = \mu_2 = \mu_3$, the null hypothesis of no differences among the population group means.

6.1.1 Example

▶ I will use Duncan's occupational-prestige data to illustrate one-way analysis of variance.

- Parallel boxplots for prestige in three types of occupations appear in Figure 7 (a).
  - Prestige, recall, is a percentage, and the data push both the lower and upper boundaries of 0 and 100 percent, suggesting the logit transformation in Figure 7 (b).
  - The data are better-behaved on the logit scale, which eliminates the skew in the blue-collar and professional groups and pulls in all of the outlying observations, with the exception of store clerks in the white-collar category.

© 2014 by John Fox

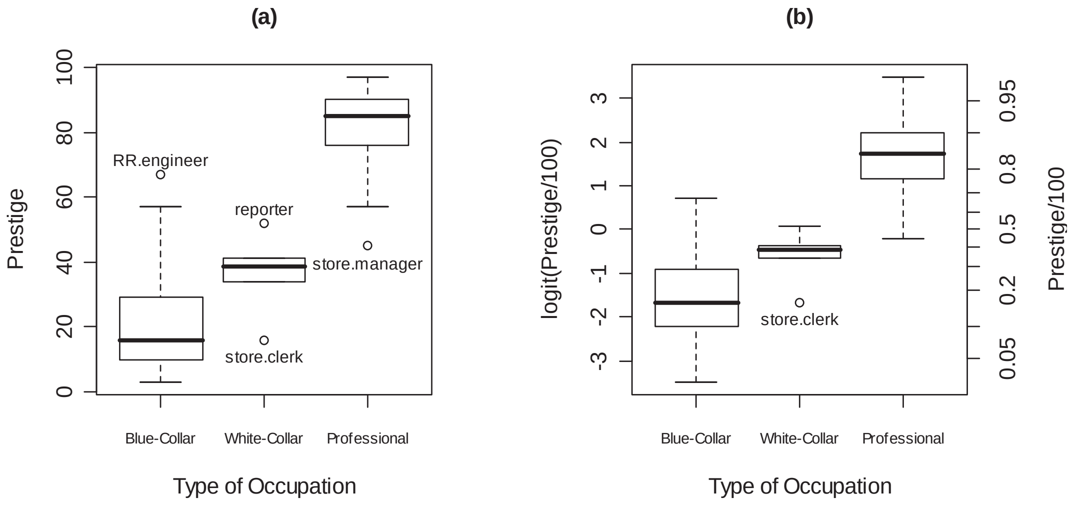

Figure 7. Parallel boxplots for (a) prestige and (b) the logit of prestige by type of occupation.

© 2014 by John Fox

- Means, standard deviations, and frequencies for prestige within occupational types are as follows:

	Prestige		
Type of Occupation	*Mean*	*Standard Deviation*	*Frequency*
Professional and managerial	80.44	14.11	18
White collar	36.67	11.79	6
Blue collar	22.76	18.05	21

  – Professional occupations therefore have the highest average level of prestige, followed by white-collar and blue-collar occupations.

- The order of the group means is the same on the logit scale:

	logit(Prestige/100)	
Type of Occupation	*Mean*	*Standard Deviation*
Professional and managerial	1.6321	0.9089
White collar	-0.5791	0.5791
Blue collar	-1.4821	1.0696

- On both scales, the standard deviation is greatest among the blue-collar occupations and smallest among the white-collar occupations, but the differences are not very large.

- Using the logit of prestige as the response variable, the one-way ANOVA for the Duncan data is

	Sum of		*Mean*		
Source	*Squares*	*df*	*Square*	F	p
Groups	95.550	2	47.775	51.98	$\ll .0001$
Residuals	38.604	42	0.919		
Total	134.154	44			

  – Occupational types account for nearly three-quarters of the variation in the logit of prestige among these occupations ($R^2 = 95.550/134.154 = 0.712$).

6.2 Two-Way ANOVA

6.2.1 Patterns of Means in the Two-Way Classification

▶ Several patterns of relationship in the two-way classification, all showing no interaction, are graphed in Figure 8:

- in (a) there are both row and column main effects;
- in (b) only column main effects;
- in (c) only row main effects;
- in (d) neither row nor column main effects.

▶ Figure 9 shows two different patterns of interactions:

- In (a), the interaction is dramatic: The order of row effects changes across columns and vice-versa. Interaction of this sort is sometimes called *disordinal*.
- In (b), the interaction is less dramatic.

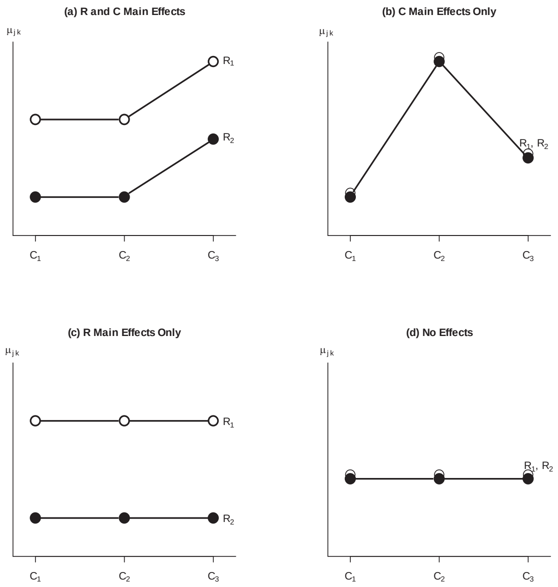

Figure 8. Patterns of association: (a) Row and Column main effects; (b) Column main effects only; (c) Row main effects only; (d) no effects.

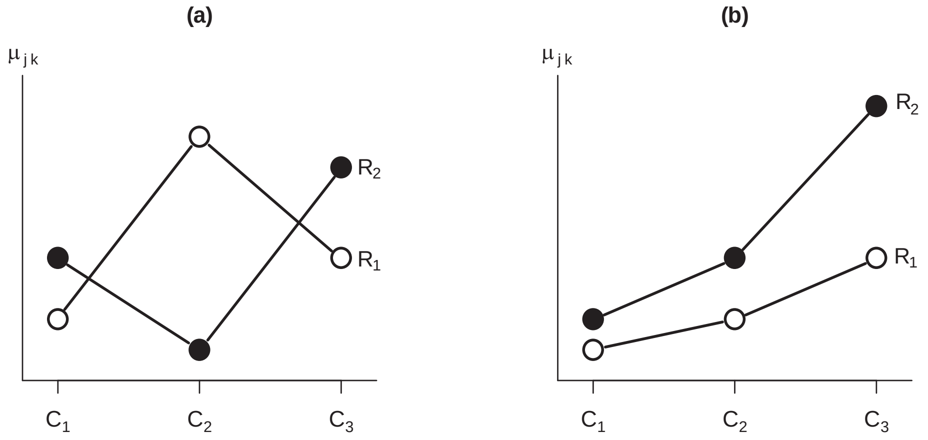

Figure 9. Two patterns of interaction in the two-way classification.

6.2.2 Example

▶ Even when interactions are absent in the population, we cannot expect perfectly parallel profiles of *sample* means: There is sampling error in sampled data.

- We have to determine whether departures from parallelism observed in a sample are sufficiently large to be statistically significant, and, if significant, are sufficiently large to be of interest.

- In general, if interactions are non-negligible, then we do not interpret the main effects of the factors — consistent with the principle of marginality.

▶ The following table shows means ($\overline{Y}_{jk}$), standard deviations (S_{jk}), and cell frequencies (n_{jk}) for data from a social-psychological experiment, reported by Moore and Krupat (1971), designed to determine how the relationship between conformity and social status is influenced by 'authoritarianism.'

		Authoritarianism		
Partner's Status		*Low*	*Medium*	*High*
Low	$\overline{Y}_{jk}$	8.900	7.250	12.63
	S_{jk}	2.644	3.948	7.347
	n_{jk}	10	4	8
High	$\overline{Y}_{jk}$	17.40	14.27	11.86
	S_{jk}	4.506	3.952	3.934
	n_{jk}	5	11	7

- Because of the conceptual-rigidity component of authoritarianism, Moore and Krupat expected that low-authoritarian subjects would be *more* responsive than high-authoritarian subjects to the social status of their partner.

- The cell means are graphed along with the data in Figure 10, and appear to confirm the experimenters' expectations.
  – There are two outlying observations in the low-status partner, high-authoritarianism condition.

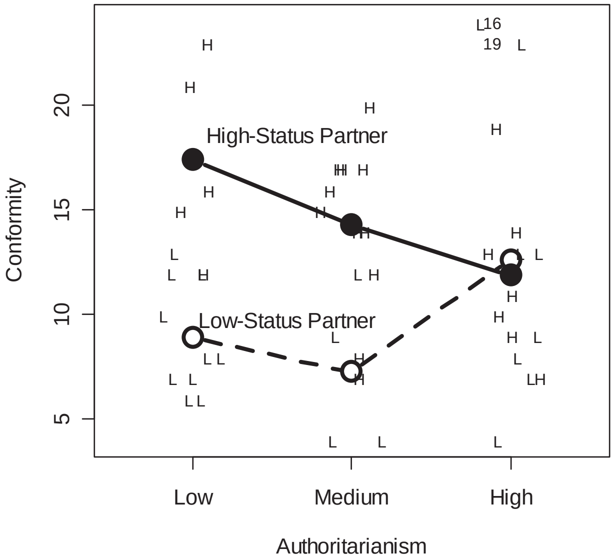

Figure 10. Mean conformity by authoritarianism and partner's status, for Moore and Krupat's data. The observations are jittered horizontally.

6.2.3 Two-Way ANOVA Model

- ▶ To model data in a two-way classification, we can code dummy regressors for each factor and take products between them for the interactions.
  - The resulting model will have as many parameters as there are cell means and consequently can represent any pattern of cell means.
- ▶ For the Moore and Krupat data, we can use the model

$$Y_i = \alpha + \gamma P_i + \delta_1 A_{i1} + \delta_2 A_{i2} + \lambda_1 (P_i A_{i1}) + \lambda_2 (P_i A_{i2}) + \varepsilon_i$$

  - Notice that there are 6 regression coefficients and $2 \times 3 = 6$ cell means.

- To test for main effects and interactions, we fit the following models to the data:

Model	*Terms*	*Parameters*	*RegSS*	*df*
1	$P, A, P \times A$	$\alpha, \gamma, \delta_1, \delta_2, \lambda_1, \lambda_2$	391.436	5
2	P, A	$\alpha, \gamma, \delta_1, \delta_2$	215.947	4
3	A	$\alpha, \delta_1, \delta_2$	3.733	3
4	P	α, γ	204.332	2

© 2014 by John Fox

- Analysis of variance table for Type-II tests:

Source	*Models Contrasted*	*SS*	*df*	*MS*	F	p
Partner's Status (P)	$2-3$	212.214	1	212.214	10.121	$.003$
Authoritarianism (A)	$2-4$	11.615	2	5.807	0.277	$.76$
P × A	$1-2$	175.489	2	87.745	4.185	$.023$
Error		817.764	39	20.968		

  - The sum of squares for 'error' is the residual sum of squares from the full model (model 1), and the degrees of freedom for error are $n-k-1 = 45-5-1 = 39$.
  - Thus, the interaction is statistically significant and we would not interpret the tests for the main effects.

© 2014 by John Fox

7. Summary

- ▶ A dichotomous explanatory variable can be entered into a regression equation by formulating a dummy regressor, coded 1 for one category of the variable and 0 for the other category.
- ▶ A polytomous explanatory variable can be entered into a regression by coding a set of 0/1 dummy regressors, one fewer than the number of categories of the variable.
  - The 'omitted' category, coded 0 for all dummy regressors in the set, serves as a baseline.
- ▶ Interactions can be incorporated by coding interaction regressors, taking products of dummy regressors with quantitative explanatory variables.
  - The model permits "different slopes for different folks" — that is, regression surfaces that are not parallel.

- ▶ The principle of marginality specifies that a model including a high-order term (such as an interaction) should normally also include the lower-order relatives of that term (the main effects that 'compose' the interaction).
  - The principle of marginality also serves as a guide to constructing incremental F-tests for the terms in a model that includes interactions.
- ▶ Dummy variables can be used to model the main effects and interactions of factors in analysis-of-variance models.