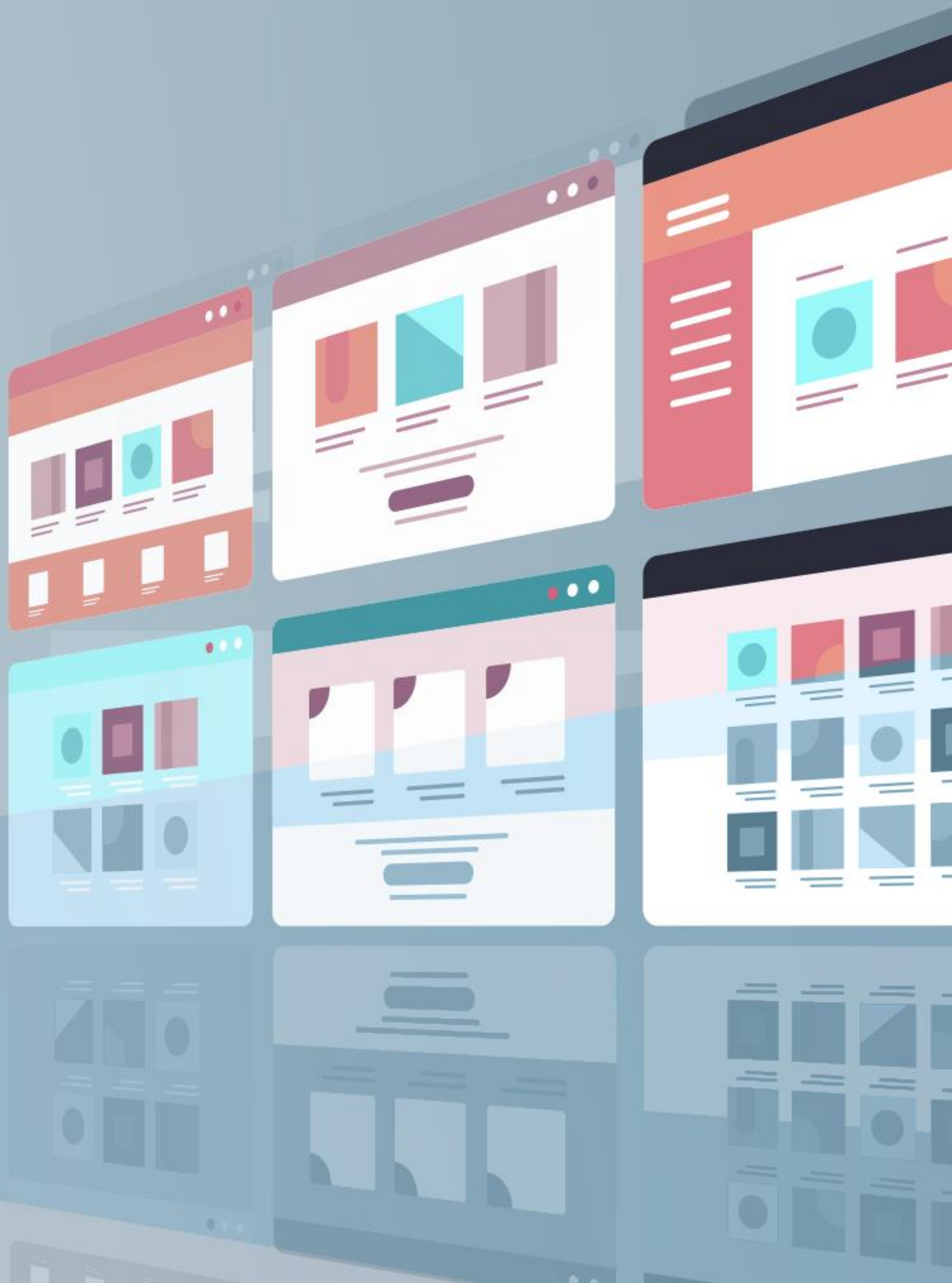
10

데이터분석방법론(1)

Qualitative Variables as Regressors

통계·데이터과학과 장영재 교수



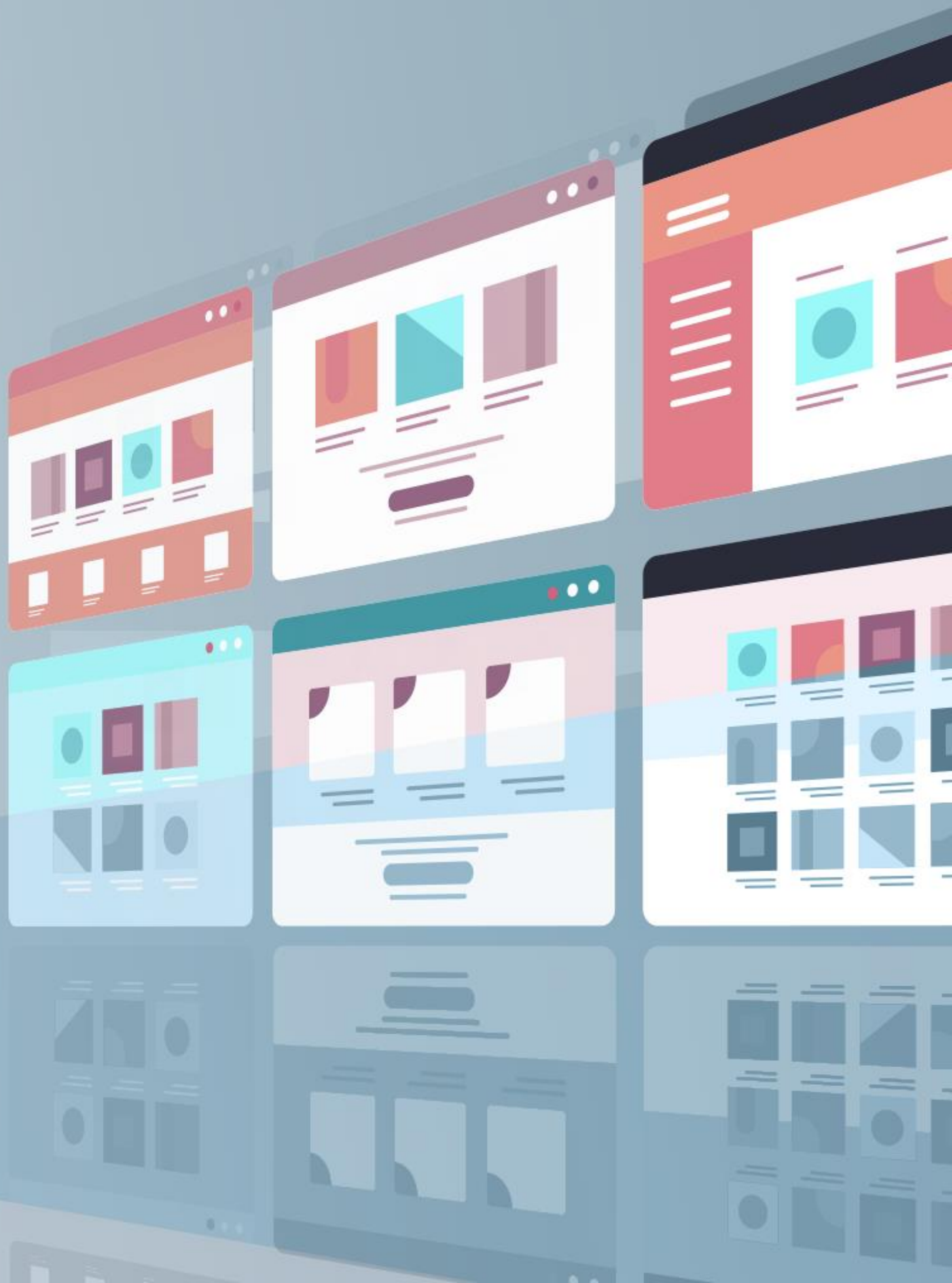
한국방송통신대학교 대학원

학습목차

1. Dummy Regressors (Dichotomous)
2. Dummy Regressors (Polytomous)
3. Modelling Interactions

01

Dummy Regressors (Dichotomous : 이항형)



1. 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 variables : constant).
- Other regression assumptions

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

1. A dichotomous Explanatory Variable

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.

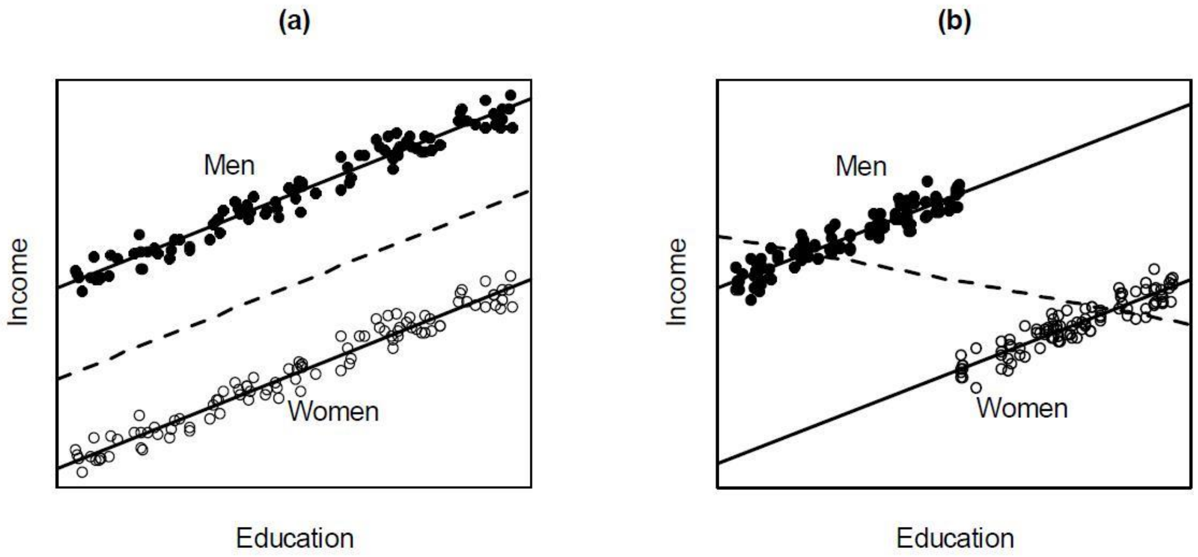
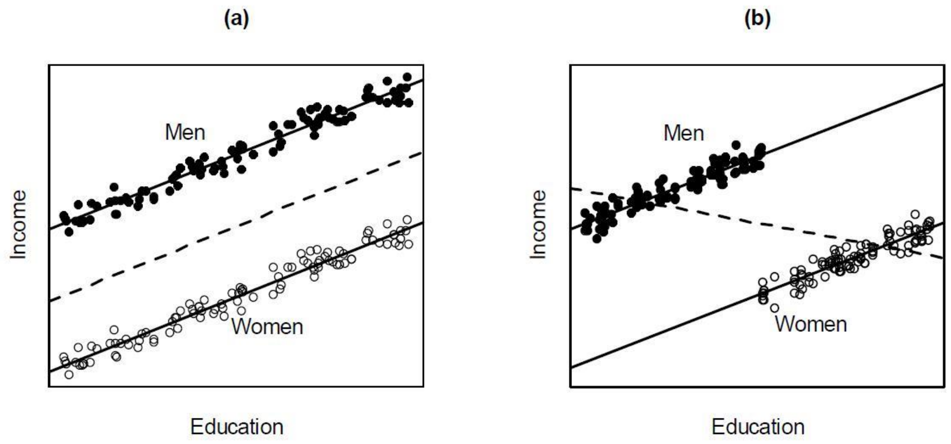


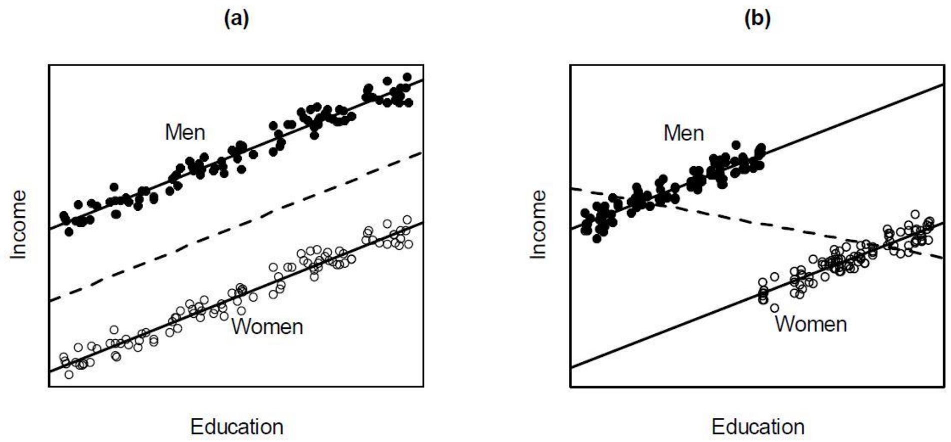
Figure 1 represents idealized examples, showing the relationship between education and income among women and men.

1. A dichotomous Explanatory Variable



- 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 (dashed line) as those of the separate within-gender regressions (ignoring gender inflates the size of the errors).
- 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.

1. A dichotomous Explanatory Variable



- We could perform separate regressions for women and men. This approach is reasonable, but it has its limitations:
  - Makes it difficult to estimate and test for gender differences in income.
  - If we can assume parallel regressions, how about pooling sample data from both groups?

we can more efficiently estimate the common education slope by pooling the data.

2. 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_i, called a dummy-variable regressor or an indicator variable, is coded 1 for men and 0 for women:

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

  - Thus, for women the model becomes

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

  - and for men

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

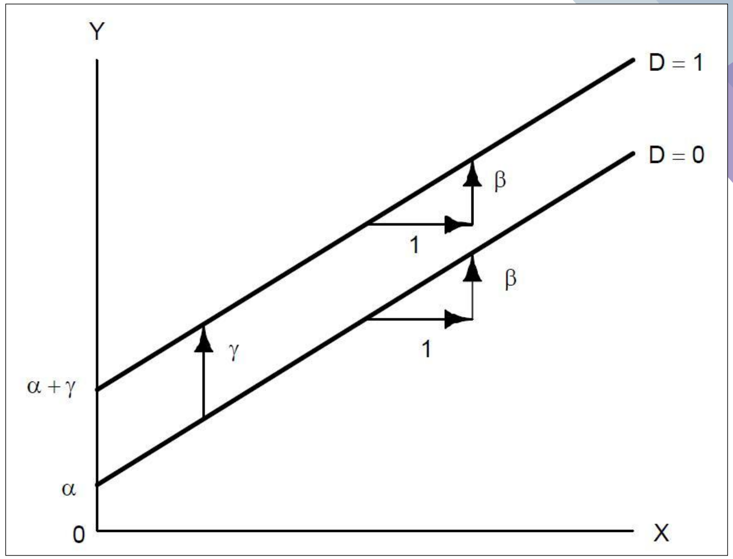


Figure 2.

- These regression equations are graphed in Figure 2.

3. 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.
  - Gender : a qualitative explanatory variable (male and female).
  - The dummy variable D : a regressor representing the categories
  - While, the quantitative explanatory variable income and the regressor X : the same.

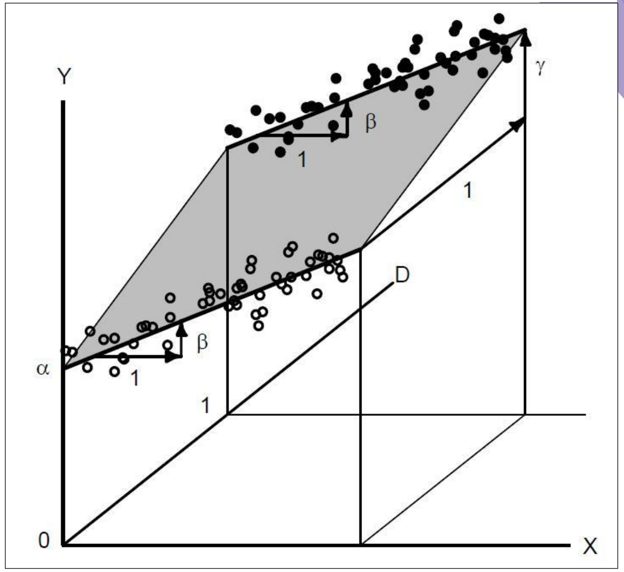
- An explanatory variable can give rise to several regressors.

- Some regressors are functions of more than one explanatory variable.

4. How and Why Dummy Regression Works

- Interpretation of parameters in the additive dummy-regression model:
  - γ gives the difference in intercepts for the two regression lines.
    - Regression lines : parallel, so γ : constant separation between the lines — the expected income gap in favor of men (when education is held constant)
    - If men were disadvantaged relative to women, then γ would be negative.
  - α gives the intercept for women, when D = 0.
  - β is the common within-gender education slope.
- Figure 3 shows the basic geometric 'trick' used when dealing with a dummy regressor.
  - What's happening is that we're creating a flat surface to match the data, like a floor. But this dummy regressor D only has two values: zero and one.

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



4. How and Why Dummy Regression Works

- If we code D the other way around (vice versa), D = 0 for men, 1 for women (Figure 4):
  - Reversed sign of γ, but same magnitude
  - The coefficient α is now the income intercept for men.

- 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):

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

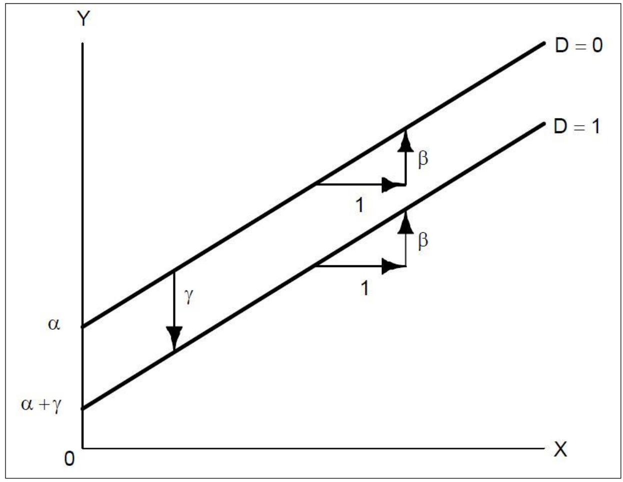
  - For D = 0 we have

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

  - and for D = 1

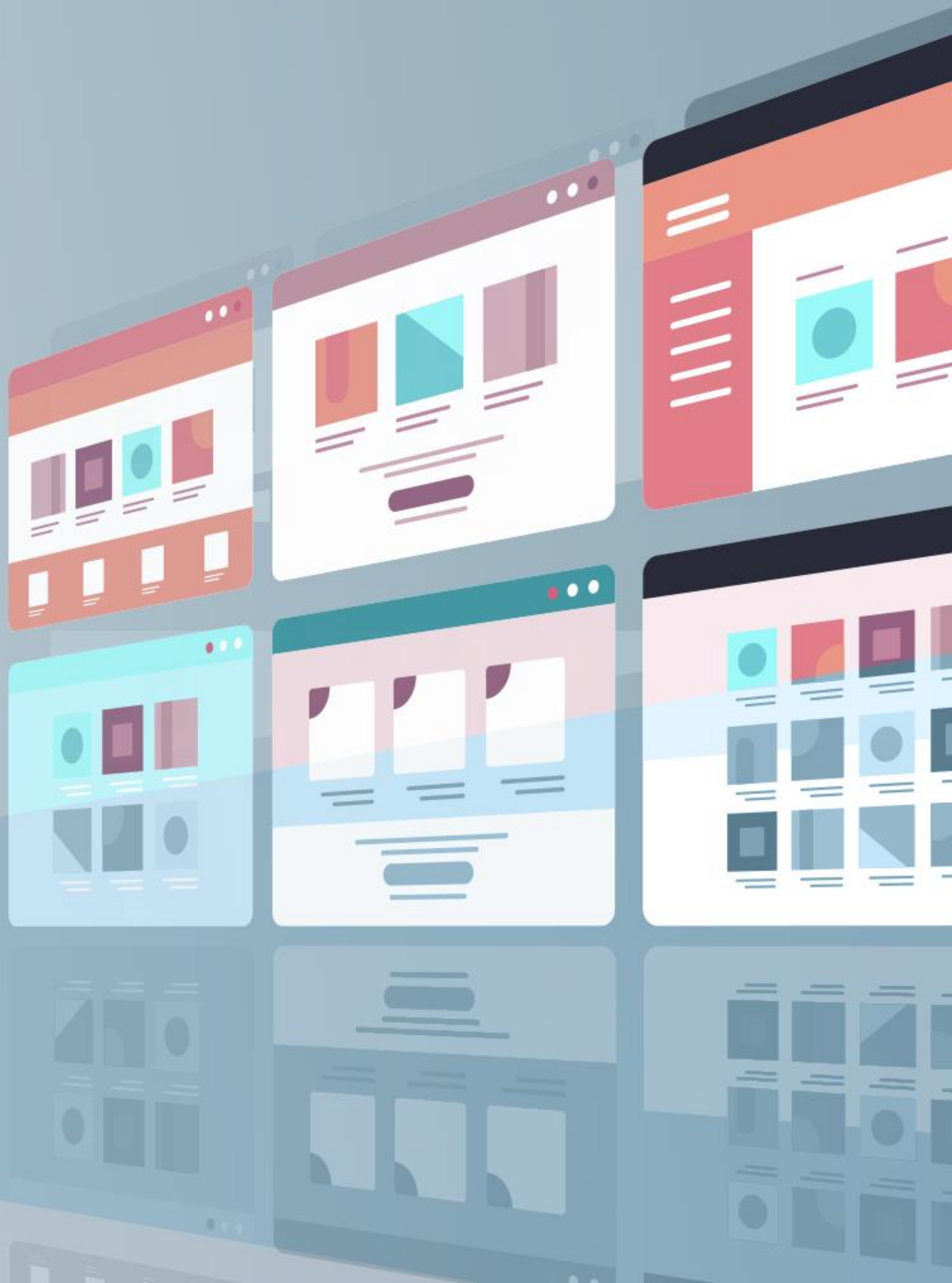
$$Y_i = (\alpha + \gamma) + \beta_1 X_{i1} {}_{+\cdots+} \beta_k X_{ik} {}_{+} \varepsilon_i$$

Figure 4. *D* = 0 for men and *D* = 1 for women.



02

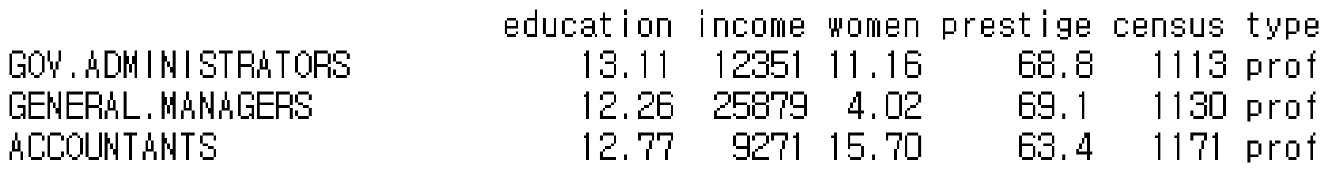
Dummy Regressors (Polytomous : 다항형)



1. Polytomous Explanatory Variables

- Example of the regression of the rated prestige of 102 Canadian occupations on their income and education levels.
  - Classified 98 of the occupations into three categories:
    (1) professional and managerial; (2) ‘white-collar’; and (3) ‘blue-collar’
  - The three-category classification : introducing two dummy regressors:

Category	D_1	D_2
Professional & Managerial	**1**	**0**
White Collar	**0**	**1**
Blue Collar	**0**	**0**



```
                       education income women prestige census type
GOV.ADMINISTRATORS         13.11  12351 11.16     68.8   1113 prof
GENERAL.MANAGERS           12.26  25879  4.02     69.1   1130 prof
ACCOUNTANTS                12.77   9271 15.70     63.4   1171 prof
```

http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/datasets/Prestige.txt

- 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_{i1} is income and X_{i2} is education for observation i.

1. Polytomous Explanatory Variables

- Three parallel regression planes with different intercepts (see Figure 5):

  Professional: $\boldsymbol{Y_i = (\alpha + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i}$
  White Collar: $\boldsymbol{Y_i = (\alpha + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i}$
  Blue Collar: $\boldsymbol{Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i}$

  - α : the intercept for blue-collar occupations
  - γ_1 : the constant vertical difference between the parallel regression planes for professional and blue-collar occupations (given that the values of education and income are constant)
  - γ_2 : the constant vertical distance between the parallel 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
- The choice of a baseline category is usually arbitrary

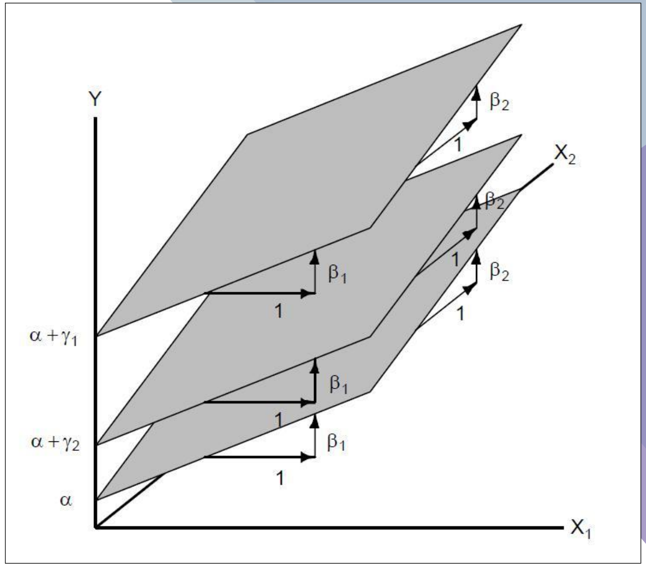


Figure 5.

1. Polytomous Explanatory Variables

- Since the selection of a baseline is arbitrary, we aim to examine the null hypothesis that there is no partial effect of occupational type,

$$H_0 : \gamma_1 = \gamma_2 = 0$$

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

- The hypothesis $H_0 : \gamma_1 = \gamma_2 = 0$ can be tested by the incremental-sum-of-squares approach (comparing the improvement in the model's fit when a particular variable (or group of variables) is added).

2. How many dummy regressors are needed?

- It may seem more natural to code three dummy regressors:

Category	D_1	D_2	D_3
Professional & Managerial	1	0	0
White Collar	0	1	0
Blue Collar	0	0	1

  - Then, for the jth occupational type, we would have

$$Y_i = (\alpha + \gamma_j) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

- It may seem more natural to code three dummy regressors:
  - Four parameters $(\alpha, \gamma_1, \gamma_2, \gamma_3)$ to represent only three group intercepts

    : can not find unique values for these four parameters.
  - The set of three dummy variables is perfectly collinear like $\boldsymbol{D3} = \boldsymbol{1} - \boldsymbol{D1} - \boldsymbol{D2}$

    : cannot calculate unique least-squares estimates for the model.

2. How many dummy regressors are needed?

- 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:
  - 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 0, delete its dummy regressors from the model and compute an incremental F-test.

Category	D_1	D_2	· · ·	D_{m-1}
1	1	0	· · ·	0
2	0	1	· · ·	0
.	.	.	· · ·	.
.	.	.	· · ·	.
.	.	.	· · ·	.
m-1	0	0	· · ·	1
m	0	0	· · ·	0

2. How many dummy regressors are needed?

- The regression of prestige on income and education

$$\hat{Y} = -7.621 + 0.001241X_1 + 4.292X_2 \quad R^2 = .81400$$
$$(3.116) \quad (0.000219) \quad (0.336)$$

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

$$\hat{Y} = -0.6229 + 0.001013X_1 + 3.673X_2 + 6.039D_1 - 2.737D_2$$
$$(5.2275) \quad (0.000221) \quad (0.641) \quad (3.867) \quad (2.514)$$
$$R^2 = .83486$$

- The three fitted regression equations are:

$$\begin{aligned} &\text{Professional:} && \hat{Y} = 5.416 + 0.001013X_1 + 3.673X_2 \\ &\text{Whitecollar:} && \hat{Y} = -3.360 + 0.001013X_1 + 3.673X_2 \\ &\text{Bluecollar:} && \hat{Y} = -0.623 + 0.001013X_1 + 3.673X_2 \end{aligned}$$

2. How many dummy regressors are needed?

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

$$\boldsymbol{H_0 : \gamma_1 = \gamma_2 = 0}$$

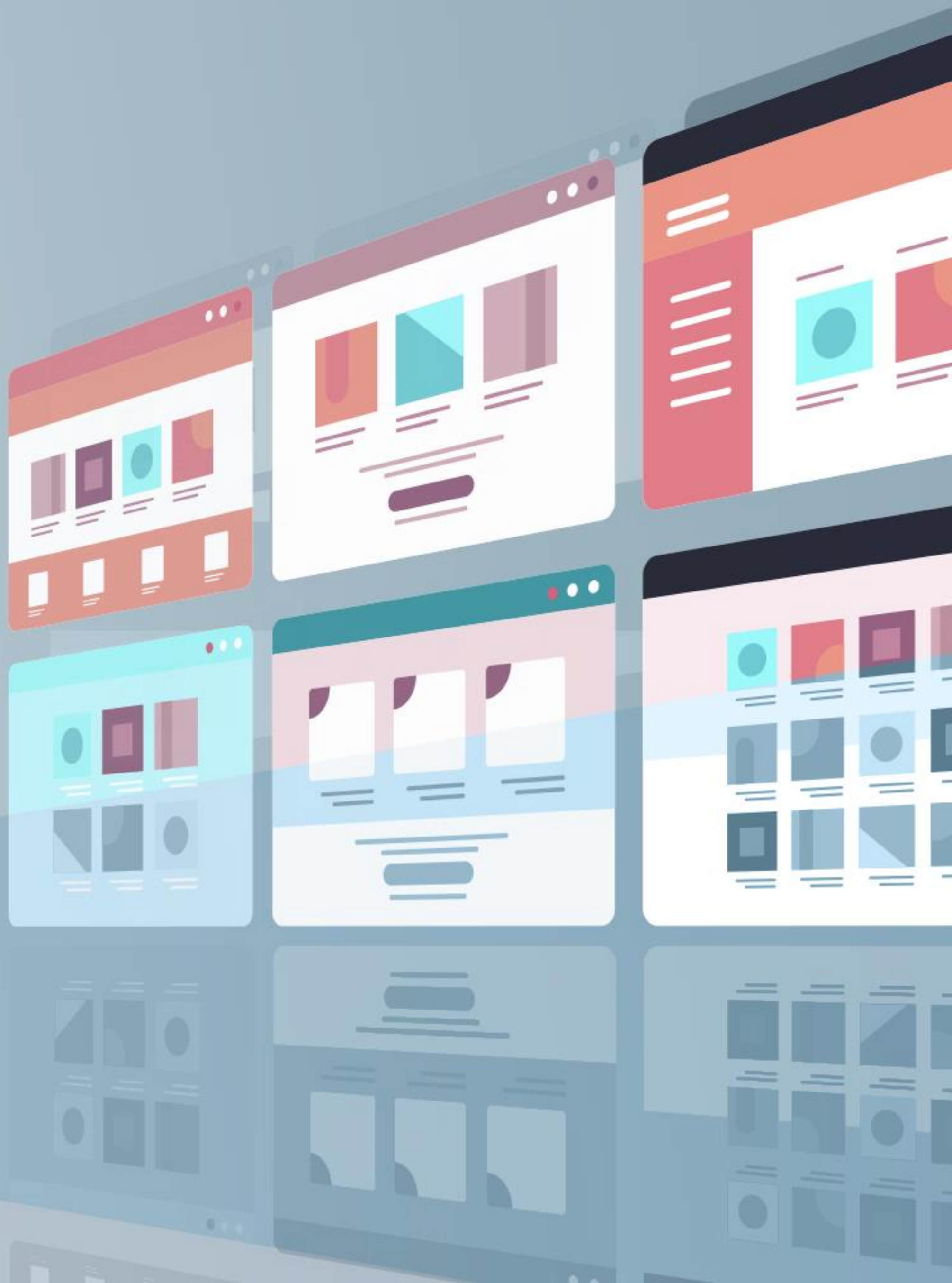
calculate the incremental F-statistic

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

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

03

Modelling Interactions



1. Interaction

- 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 regression lines 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 6

  (The effects of gender and education were additive in Figure 1)
  - In (a), gender and education are independent,

    since women and men have identical education distributions.
  - In (b), gender and education are related,

    since women have higher levels of education than men on average.

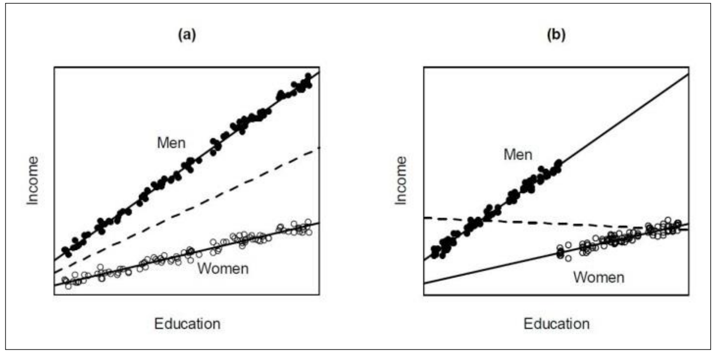


Figure 6

1. Interaction

- In both (a) and (b), the within-gender regression lines 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.

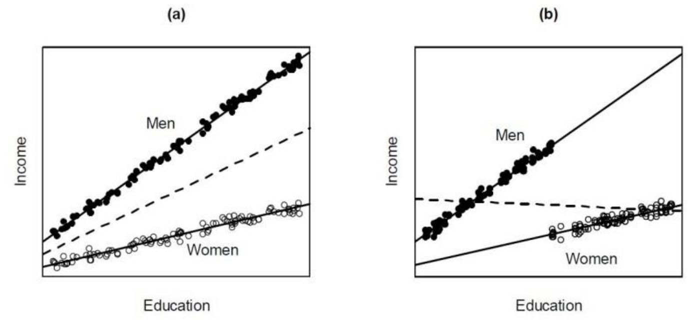


Figure 6

1. Interaction

- 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.

2. 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:

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

  - The dummy regressor D for gender
  - The quantitative regressor X for education,
  - Interaction regressor XD.

2. Constructing Interaction Regressors

- 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.
- For women,

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

- And for men,

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

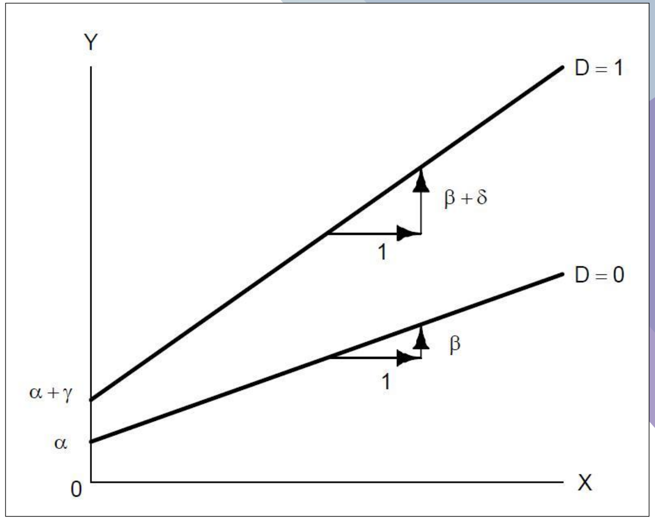


Figure 7

- These regression equations are graphed in Figure 7:
  - 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.
  - To test for interaction, we can test the hypothesis $H_0 : \delta = 0$.

3. The Principle of Marginality

- Consider the model

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

  - As shown in Figure 8 (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 8 (b), constrains the slope for women to 0, which is needlessly restrictive.

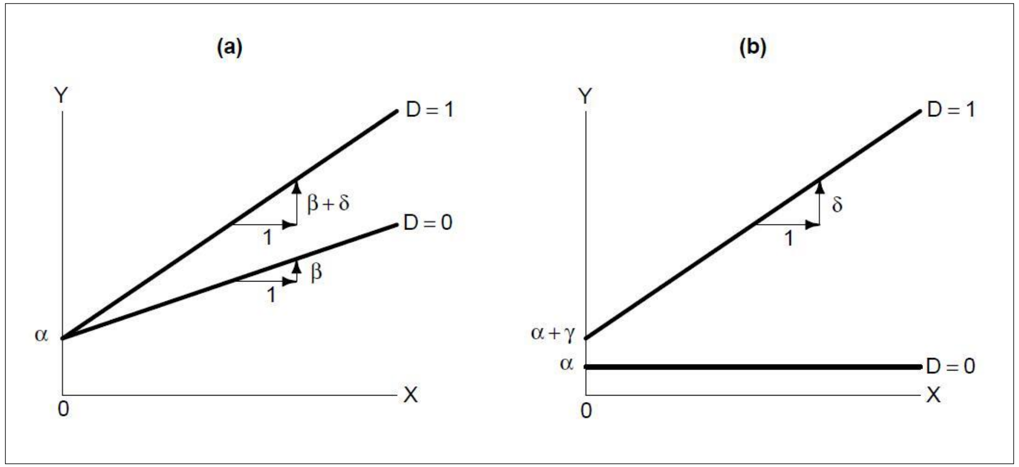


Figure 8

4. 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 $X_{i1}D_{i1}$ $X_{i1}D_{i2}$ capture the interaction between income(X_{i1}) and occupational type
  - $X_{i2}D_{i1}$ and $X_{i2}D_{i2}$ capture the interaction between education(X_{i2}) and occupational type.

4. Interactions With Polytomous Explanatory Variables

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

Professional: $\boldsymbol{Y_i = (\alpha + \gamma_1) + (\beta_1 + \delta_{11})Xi_1 + (\beta_1 + \delta_{21})Xi_1 + \varepsilon_i}$

Wihite Collar: $\boldsymbol{Y_i = (\alpha + \gamma_2) + (\beta_1 + \delta_{12})Xi_1 + (\beta_2 + \delta_{22})Xi_2 + \varepsilon_i}$

blue Collar: $\boldsymbol{Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i}$

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

4. Interactions With Polytomous Explanatory Variables

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

$$
\begin{aligned}
\hat{Y}_i = {} & 2.276 + 0.003522X_1 + 1.713X_2 \\
& (7.057) \quad (0.000556) \quad (0.927) \\
& + 15.35D_1 - 33.54D_2 \\
& \quad (13.72) \qquad (17.54) \\
& + 0.002903X_1D_1 - 0.002072X_1D_2 \\
& \quad (0.000599) \qquad (0.00894) \\
& + 1.388X_2D_1 - 4.291X_2D_2 \\
& \quad (1.289) \qquad (1.757) \\
& \qquad\qquad\qquad\qquad R^2 = .8747
\end{aligned}
$$

- The regression equation for each group:

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

5. Hypothesis Tests for Main Effects and Interactions

- To test the null hypothesis of no interaction between income and type,
  $H_0 : \delta_{11} = \delta_{12} = 0$,
  delete the interaction regressors X_1D_1 and X_1D_2 from the full model
  and calculate an incremental F-test.
  - To test the null hypothesis of no interaction between education and type,
    $H_0 : \delta_{21} = \delta_{22} = 0$,
    delete the interaction regressors X_2D_1 and X_2D_2 from the full model
    and calculate an incremental F-test.

5. Hypothesis Tests for Main Effects and Interactions

Model	Terms(Income, Education,Type)	Parameters	Regression 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	$I, E, T, I\times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}$	24,556	6
3	$I, E, T, E\times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{21}, \delta_{22}$	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	$I, T, I\times T$	$\alpha, \beta_1, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}$	23,488	5
7	$E, T, E\times T$	$\alpha, \beta_2, \gamma_1, \gamma_2, \delta_{21}, \delta_{22}$	22,710	5

5. Hypothesis Tests for Main Effects and Interactions

Source	Models Contrasted	Sum of 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	<.0001
Income × Type	1 – 3	952	2	11.92	<.0001
Education × Type	1 – 2	238	2	2.98	.56
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	$\delta_{21} = \delta_{22} = 0$

5. Hypothesis Tests for Main Effects and Interactions

- 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.
  - 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 ($\delta_{21} = \delta_{22} = 0$).

다음시간 안내

11

Analysis of Collinear Data