Chapter 8

Use of "Dummy" Variables

Overview

- Categorical variable as predictor
- Dummy variable (Indicator variable)
  - A dummy variable takes two values, 0 or 1
- First order model with dummy variable
- Second order model with dummy variable
- Test for coincidence
- Testing for parallelism
- Categorical variable with more than 2 levels

8.1 Introduction

- How to handle predictor variables which are categorical?

- Examples of categorical variables
  1. Different types of fertilizers on the yield of rice.
  2. Different teaching methods on the exam score.

Introduction (Continued)

- Since the levels in these factors may just represent different categories and do not have any ordering, hence dummy variables are used to identify these categories.
- For example,

$$D = \begin{cases} 1, & \text{if the observation is from Machine A;} \\ 0, & \text{if the observation is from Machine B.} \end{cases}$$

- How about 3 machines?

Introduction (Continued)

- If there are c categories in a particular factor, then $c - 1$ dummy variables will be used to identify all these categories.

- For example

$$(D_1, D_2) = \begin{cases} (1,0), & \text{for Machine A;} \\ (0,1), & \text{for Machine B;} \\ (0,0), & \text{for Machine C.} \end{cases}$$

8.2 First Order Model with "Dummy" Variable

Example 1

- A lecturer in a university wishes to investigate the relationship between the achieved college grade point index (x) and the starting salary (y) of recent graduates majoring in business
- so that when advising students he may build a model to predict the starting salary of business majors based on the grade point index.

- A random sample of 30 recent graduates from the Faculty of Business Administration is drawn and the data are given as follows.

Example 1 (Continued)

x	2.7	3.1	3.0	3.3	3.1	2.4	2.9	2.1
y	17.0	17.7	18.6	20.5	19.1	16.4	19.3	14.5
Major	acc	acc	acc	acc	acc	acc	acc	acc

x	2.6	3.2	3.0	2.2	2.8	3.2	2.9
y	15.7	18.6	19.5	15.0	18.0	20.0	19.0
Major	acc	acc	acc	acc	acc	acc	acc

x	3.0	2.6	3.3	2.9	2.4	2.8	3.7	3.1
y	17.4	17.3	18.1	18.0	16.2	17.5	21.3	17.2
Major	m/m	m/m	m/m	m/m	m/m	m/m	m/m	m/m

x	2.8	3.5	2.7	2.6	3.2	2.9	3.0
y	17.0	19.6	16.6	15.0	18.4	17.3	18.5
Major	m/m	m/m	m/m	m/m	m/m	m/m	m/m

y: starting salary (1000); *x*: GPI

acc: accountancy; m/m: marketing/management

Example 1 (Continued)

- Let $D = \begin{cases} 0, & \text{for accounting major;} \\ 1, & \text{for marketing/management major.} \end{cases}$

  Model $y = \beta_0 + \beta_1 x + \beta_2 D + \varepsilon$ (1)

- β_2 can be interpreted as the expected additional amount in the starting salary for a marketing/management major over an accountancy major with the same GPI.
- The model in Equation (1) is equivalent to

$$y = \begin{cases} \beta_0 + \beta_1 x + \epsilon, & \text{for acc major;} \\ (\beta_0 + \beta_2) + \beta_1 x + \epsilon, & \text{for m/m major.} \end{cases}$$

Example 1 (Continued)

- Moreover, if there is an interaction effect between GPI and the major on the response (i.e. the rates of increase in the starting salary per 1 unit increase in GPI are different for different majors). Then the model should be

$$y = \beta_0 + \beta_1 x + \beta_2 D + \beta_3 xD + \varepsilon \quad (2)$$

- β_3 is the expected additional increase per unit increase in GPI for a marketing/management major.

Example 1 (Continued)

- The model in Equation (2) is equivalent to

$$y = \begin{cases} \beta_0 + \beta_1 x + \epsilon, & \text{for acc major} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x + \epsilon, & \text{for m/m major.} \end{cases}$$

- The above idea can be extended to the second order model.

8.3 Second Order Model with "Dummy" Variable

- Suppose that we have 2 sets of data on y and x

- We want to fit $y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \varepsilon$ for each set.

- We can handle both sets of data simultanenously by using the following model,

$$y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \alpha_0 D + \alpha_1 x D + \alpha_{11} x^2 D + \varepsilon \quad (3)$$

where $D = \begin{cases} 0, & \text{if the obs is from data set 1;} \\ 1, & \text{if the obs is from data set 2.} \end{cases}$

Second Order Model with "Dummy" Variable (Continued)

- The model in Equation (3) is equivalent to

$$y = \begin{cases} \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon, \\ \qquad\qquad \text{for obs from data set 1;} \\ (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1) x + (\beta_{11} + \alpha_{11}) x^2 + \epsilon, \\ \qquad\qquad \text{for obs from data set 2.} \end{cases}$$

Second Order Model with "Dummy" Variable (Continued)

- Partial F tests then enable us to check the various possibilities, as follows:

For examples

1. H_0: $\alpha_0 = \alpha_1 = \alpha_{11} = 0$ against H_1: H_0 is not true.

   If H_0 is rejected, we conclude that the models are not the same.

   If H_0 is not rejected, we take them to be the same.

   Model under H_0:

$$y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \varepsilon$$

Second Order Model with "Dummy" Variable (Continued)

2. If H_0 in (1) is rejected, we would look at the subsets of the α's.

   For example, we would like to test

   H_0: $\alpha_1 = \alpha_{11} = 0$ against H_1: H_0 is not true.

   If H_0 is not rejected, we conclude that the two sets of data exhibited only a difference in response levels, but with the same slope and curvature.

   Model under H_0:

$$y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \alpha_0 D + \varepsilon$$

Second Order Model with "Dummy" Variable (Continued)

3. If H_0 in (2) is rejected, we could test

   H_0: $\alpha_{11} = 0$ against H_1: $\alpha_{11} \neq 0$ to see if the models differ only in zero and first order terms, indicated by non-rejection of H_0.

   Model under H_0

   $$y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \alpha_0 D + \alpha_1 x D + \varepsilon$$

   or

   $y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \varepsilon$ for obs. from data set 1

   $y = (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1) x + \beta_{11} x^2 + \varepsilon$ for obs. from data set 2

8.4 Testing for Coincidence

- We want to test if the models from different groups are the same
- Refer to Example 1 on Slide 8.7
- Test H_0: $\beta_2 = \beta_3 = 0$ against

  H_1: $\beta_2 \neq 0$ or $\beta_2 \neq 0$ or both.
- That is to choose between
  - Reduced model under H_0: $y = \beta_0 + \beta_1 x + \varepsilon$ and
  - Model: $y = \beta_0 + \beta_1 x + \beta_2 D + \varepsilon$ or $y = \beta_0 + \beta_1 x + \beta_3 xD + \varepsilon$ or $y = \beta_0 + \beta_1 x + \beta_2 D + \beta_3 xD + \varepsilon$
- If H_0 is not rejected, then it implies that a dummy variable is not required.

Testing for Coincidence (Continued)

- Test statistics

$$F = \frac{\big(SSR(x, D, xD) - SSR(x)\big)/2}{SSE(x, D, xD)/26}$$

From the computer output, we have

$$F = \frac{(63.2451 - 58.0194)/2}{14.2819/26} = 4.76$$

- Since $F_{\text{obs}} = 4.76 > F_{0.05}(2, 26) = 3.37$ (or p-value = $0.0173 < 0.05$), we reject H_0 at the 5% significance level and hence the (dummy) variable pertaining to the business major is a significant contributor to a model predicting the starting salary.

Testing for Coincidence (Continued)

- In other words, the model for accounting major does not coincide with the model for marketing/management major.

Note:

- $SSR(x)$ can be obtained from the first row of the **Type I SS** if x is entered as the first variable in the model statement.
- Alternatively, $SSR(x)$ can be obtained by running separately for the model $y = \beta_0 + \beta_1 x + \varepsilon$
  - In SAS, use "proc reg;" with "model y=x;"
  - In R, use "lm(y~x)"

Testing for Coincidence (Continued)

- In SAS, we can use the following statements to perform the partial F test

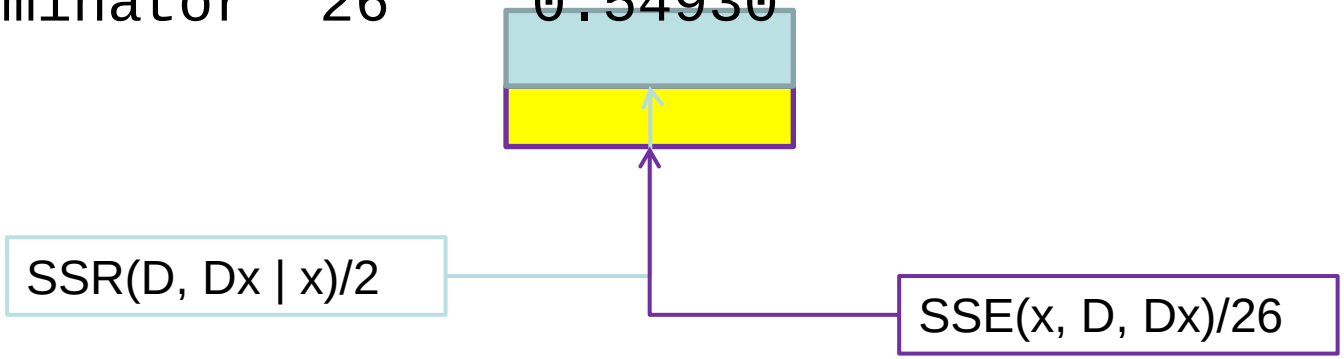
```
data ch8ex1;
  infile "d:\ST3131\ch8ex1.txt" firstobs=2;
  input x y d;
  xd=x*d;
proc reg data=ch8ex1;
  model y = x d xd;
  test d=0, xd=0;
run;
quit;
```

Testing for Coincidence (Continued)

- Partial output from running the above program

Test 1 Results for Dependent Variable y

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	2.61285	4.76	0.0174
Denominator	26	0.54930		



Testing for Coincidence (Continued)

- In R, we can use the following statements to perform the partial F test

```
> modelfull=lm(y~x+d+x*d)
> model1=lm(y~x)
> anova(model1,modelfull)
Analysis of Variance Table
Model 1: y ~ x
Model 2: y ~ x + d + x * d
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     28 19.5076
2     26 14.2819  2    5.2257 4.7567 0.01736 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
   ‘.’ 0.1 ‘ ’ 1
```

SSE(x)

SSE(x, d, x*d)

SSR(d, x*d | x)

8.5 Testing for Parallelism

- We want to test if the effect of the predictor is the same for different groups
- Refer to Example 1 on Slide 8.7
- Test H_0: $\beta_3 = 0$ against H_1: $\beta_3 \neq 0$.
- That is to choose between
  - Reduced model under H_0:

    $y = \beta_0 + \beta_1 x + \beta_2 D + \epsilon$ and
  - Full model: $y = \beta_0 + \beta_1 x + \beta_2 D + \beta_3 xD + \epsilon$
- It is equivalent to test if there is any interaction effect, xD.

Testing for Parallelism (Continued)

- Test statistic:

$$F = \frac{\big(SSR(x, D, xD) - SSR(x, D)\big)/1}{SSE(x, D, xD)/26}$$

- From the computer output, we have

$$F = \frac{(63.2451 - 62.4764)/1}{14.2819/26} = 1.40$$

- Since $F_{\text{obs}} = 1.40 < F_{0.05}(1, 26) = 4.23$ (or p-value = $0.2475 > 0.05$), we do not reject H_0 at the 5% significance level and we conclude that there is no interaction effect.

Testing for Parallelism (Continued)

- Hence the fitted model reduces to

$$\hat{y} = 6.2053 + 4.1370x - 0.7849D$$

- That is,

$$\hat{y} = 6.2053 + 4.1370x, \quad \text{for acc major,}$$

$$\hat{y} = 5.4204 + 4.1370x, \quad \text{for m/m major}$$

Note: $\hat{\beta}_2$ is a negative number means that there is a reduction in the starting salary for marketing management majors as compared to that for the accounting majors.

8.6 Factors with more than two levels

Example 2

- The following table shows turkey weights (y) in pounds, and age (x) in weeks, of thirteen thanksgiving turkeys.

- Four of these turkeys, were reared in Georgia (G), four in Virginia (V) and five in Wisconsin (W).

x	y	Origin	D_1	D_2
28	13.1	G	1	0
20	8.9	G	1	0
32	15.1	G	1	0
22	10.4	G	1	0
29	13.1	V	0	1
27	12.4	V	0	1
28	13.2	V	0	1
26	11.8	V	0	1
21	11.5	W	0	0
27	14.2	W	0	0
29	15.4	W	0	0
23	13.1	W	0	0
25	13.8	W	0	0

Example 2 (Continued)

- A simple regression model is fitted and the fitted equation is

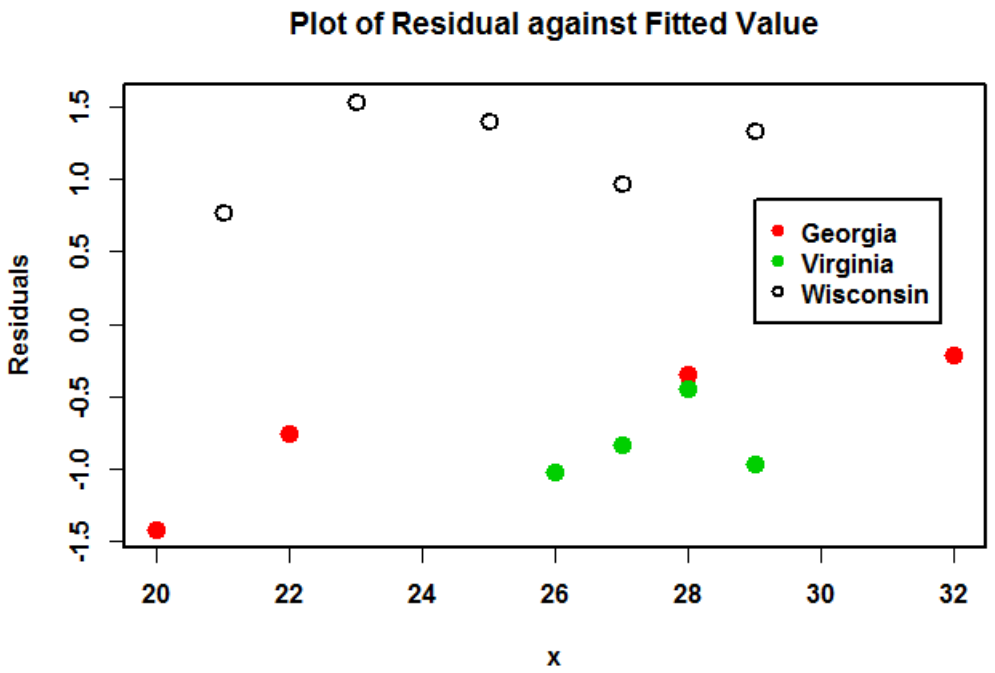
$$\hat{y} = 1.98 + 0.0417x$$

- The residuals from this fitted model are, in ascending order,

−0.35, −1.42, −0.22, −0.75, −0.97, −0.83, −0.45, −1.02, 0.77, 0.97 1.33, 1.53 and 1.40.

Residual Plot

- Residuals from G and V groups are negative and positive for W group.
- Hence different origins of the turkeys may have difference in weights
- Therefore dummy variables should be used.



Use of Dummy Variables for Factor with 3 Levels

- Consider the following model

$$y = \beta_0 + \beta_1 x + \beta_2 D_1 + \beta_3 D_2 + \varepsilon,$$

where $(D_1, D_2) = \begin{cases} (1,0), & \text{for G;} \\ (0,1), & \text{for V;} \\ (0,0), & \text{for W.} \end{cases}$

Note

- D_1 compares G with W while keeping age fixed
- D_2 compares V with W while keeping age fixed

Fitted Model

- The fitted equation is

$$\hat{y} = 1.4309 + 0.4868x - 1.9184D_1 - 2.1919D_2$$

- That is,

$$\hat{y} = -0.4875 + 0.4868x, \quad \text{for G;}$$
$$\hat{y} = -0.7610 + 0.4868x, \quad \text{for V;}$$

and

$$\hat{y} = \quad 1.4309 + 0.4868x, \quad \text{for W.}$$

Significance Test

- $R^2 = 0.9794$
- For testing H_0: $\beta_1 = \beta_2 = \beta_3 = 0$, we have

$$F_{\text{obs}} = (38.6058/3)/(0.8112/9) = 142.78$$

- Since the observed F-value $= 142.78 > F_{0.05}(3, 9) = 3.86$ (or p-value $= 6.6(10)^{-8} < 0.05$), we reject H_0 at the 5% significance level and conclude that there is a significant relationship between the weight of a turkey and the age and origin of the turkey.

- Are there any differences among the three origins of the turkeys?

Any Difference between G & W?

- Is there any difference between turkeys from Georgia (G) and Wisconsin (W)?

- Test H_0: $\beta_2 = 0$ against H_1: $\beta_2 \neq 0$

- Test Statistic

$$F = \frac{SSR(D_1|D_2, x)/1}{MSE(x, D_1, D_2)} = \frac{8.1449}{0.0901} = 90.04$$

Any Difference between G & W?

- Since $F_{\text{obs}} = 90.04 > F_{0.05}(1,9) = 5.12$ (or p-value = $5.44(10)^{-6} < 0.05$), H_0 is rejected at the 5% significance level.

- We conclude that there is significant difference in expected weights between turkeys from Georgia and Wisconsin of the same age.

Alternate Test

- An alternate test for testing H_0: $\beta_2 = 0$ against H_1: $\beta_2 \neq 0$ is as follows

$$t = \frac{\hat{\beta}_2}{\text{s.e.}(\hat{\beta}_2)} = -\frac{1.9184}{0.2018} = -9.51$$

- Sine $|t_{\text{obs}}| = 9.51 > t_{0.025}(9) = 2.228$ (or p-value = $5.44(10)^{-6} < 0.05$), H_0 is rejected at the 5% significance level.

Any Difference between V & W?

- Is there any difference between turkeys from Virginia (V) and Wisconsin (W)?

- Test H_0: $\beta_3 = 0$ against H_1: $\beta_3 \neq 0$

- Test Statistic:

$$F = \frac{SSR(D_2|D_1, x)/1}{MSE(x, D_1, D_2)} = \frac{9.6873}{0.0901} = 107.52$$

Any Difference between V & W?

- Since $F_{\text{obs}} = 107.52 > F_{0.05}(1,9) = 5.12$ (or p-value = $2.64(10)^{-6} < 0.05$), H_0 is rejected at the 5% significance level.

- We conclude that there is significant difference in expected weights between turkeys from Virginia and Wisconsin of the same age.

Any Difference between G & V?

- Is there any difference between turkeys from Georgia (G) and Virginia (V)?

- There is no β's in the original model that represents the difference between these 2 groups

Any Difference between G & V?

- We notice that

  β_2 measures G – W

  β_3 measures V – W

- Hence $\beta_2 - \beta_3$ can be considered measuring G – V

- Therefore we want to test H_0: $\beta_2 - \beta_3 = 0$ against H_1: $\beta_2 - \beta_3 \neq 0$

- We can use the testing general linear hypotheses approach for the above hypothesis testing problem

Any Difference between G & V? (Continued)

- The model under the hull hypothesis that $\beta_2 = \beta_3$ is given by

$$y = \beta_0 + \beta_1 x + \beta_2 (D_1 + D_2) + \epsilon,$$

- From the computer printout obtained by fitting the above reduced model, we have

$$SSE_{\mathrm{H}} = 0.9525 \text{ with } 10 \text{ d.f.}$$

Any Difference between G & V? (Continued)

- On the other hand, $SSE = 0.8117647$ with 9 d.f.

- Hence

$$F = [SSE_{\mathrm{H}} - SSE)/1]/[SSE/9] = 1.568.$$

- Since $F_{\mathrm{obs}} = 1.568 < F_{0.05}(1,9) = 5.12$ (or p-value = $0.2421 > 0.05$), we do not reject H_0 and conclude that there is no difference in the expected weights between groups G and V of the same age.

Interaction Effect

- Are there any interaction effects?
- Consider the model

$$y = \beta_0 + \beta_1 x + \beta_2 D_1 + \beta_3 D_2 + \beta_4 x D_1 + \beta_5 x D_2 + \epsilon$$

- Test H_0: $\beta_4 = \beta_5 = 0$ against H_1: H_0 is not true.
- Why is there no interaction term $D_1 D_2$ in the model?

Interaction Effect

$$F = \frac{SSR(xD_1, xD_2 | x, D_1, D_2)/2}{MSE(x, D_1, D_2, xD_1, xD_2)}$$

$$= \frac{\left(SSR(x, D_1, D_2, xD_1, xD_2) - SSR(x, D_1, D_2)\right)/2}{MSE(x, D_1, D_2, xD_1, xD_2)}$$

$$= \frac{(38.7107 - 38.6057)/2}{0.0101} = 0.52$$

- Since $F_{\text{obs}} = 0.52 < F_{0.05}(2,7) = 4.74$ (or p-value = $0.6158 > 0.05$), therefore H_0 is not rejected and hence $y = \beta_0 + \beta_1 x + \beta_2 D_1 + \beta_3 D_2 + \epsilon$ is an appropriate model.

8.7 Dummy Variables for More Than One Factor

- The above two examples have only one factor in the model.
  - The major with 2 levels (accounting, and marketing/management) in Example 1
  - The origin with 3 levels (Georgia, Virginia, and Wisconsin) in Example 2

- However dummy variables can be used for more than one factor.

Dummy Variables for More Than One Factor (Continued)

- Consider the following example.

  Response: salary (y)

  Predictors: experience (x_1) --- measured in years

  education (D_1, D_2):

  (1, 0) for O/A Levels

  (0, 1) for bachelor degree

  (0, 0) for advanced degree

  management responsibility (D_3):

  1 if the respondent has

  0 otherwise.

Model: $y = \beta_0 + \beta_1 x + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + \epsilon$