Ch 8: Dummy variable regression models

Outline

1. Overview: the nature of dummy-variable regression
2. ANOVA models
3. ANOVA models with two qualitative variables
4. ANOVA models with interactions
5. ANCOVA: regression with qualitative **AND** quantitative dependent variables
6. Summary

1.1 Explaining the terminology

- The subjects of Analysis of Variance (ANOVA) and Analysis of Covariance (ANCOVA) present early examples of when the terminology used can be confusing and make things look harder than they really are. Quite simply

  1. ANOVA refers to situations where regression models contain purely qualitative X variables
  2. ANCOVA refers to situations where regression models contain a combination of qualitative and quantitative X variables

1.2 Introduction

- It is often convenient to discuss regression models where some of the X variables are qualitative in nature
- Among the wealth of examples discussed in Gujarati and Porter (2009) are categories such as male/female, black/white, Catholic/non-Catholic etc.
- In this case these variables essentially codify whether the effect is absent ($X = 0$) or present ($X = 1$)
- However, there are numerous examples where the categories involved consist of more than two levels. Examples include seasons of the year or regions in the UK

1.3 Example of dummy variables

- Dummy variables often (but not always) associated with the time of the year
- Suppose we have data on quarterly fridge sales
- A portion of this data is shown below

Fridge Sales	Durable Goods Sales	Q1	Q2	Q3	Q4	Quarter
1317	252.6	1	0	0	0	1
1615	272.4	0		1	0	2
1662	270.9	0	0	1	0	3
1295	273.9	0	0	0	1	4

Table: Portion of a dataset on quarterly fridge sales

1.4 Setting up dummy variable regression problems

- Data on the previous slide hints at two possible approaches
  1. Include 4-1=3 dummy variables in addition to the constant term
  2. Include one variable with all four categories listed
- Differences between the two approaches are due to the following
  1. This is how the subject is often taught from first principles
  2. This is the most efficient way of organising this data in R using the command `factor`

1.5 The dummy variable trap

- Suppose you follow the first approach
- If you have four different levels you would need $4 - 1 = 3$ dummy variables in the regression model in addition to the constant term
- In general terms if you have m different levels you would need $m - 1$ dummy variables in the regression model in addition to the constant term
- **But where do these numbers come from?**

1.6 Explanation of the dummy variable trap

• Suppose that you have a qualitative X variable that takes m different levels (e.g. the previous example has $m = 4$ quarters corresponding to the time of the year)
• Need at least $m - 1$ dummy variables plus the intercept term in order to represent each of the m categories that can arise
• Now suppose that you include m dummy variables together with the constant term
• The regression model now becomes

$$y = \beta_0(1) + \beta_1 D_1 + ... + \beta_m D_m + u,$$

where $D_1, D_2, ..., D_m$ denote dummy variables

1.7 Where the dummy variable trap comes from

- Starting from the regression model

$$y = \beta_0(1) + \beta_1 D_1 + ... + \beta_m D_m + u. \qquad (1)$$

- Now

$$\frac{1}{m} D_1 + ... + \frac{1}{m} D_m = 1$$

- This means there is an exact linear relationship between the X-variables on the right hand side of the regression model in equation (1)
- This contradicts the assumptions of the classical linear regression model

2.1 Fitting analysis of variance (ANOVA) models in R

- The data on Slide 1.3 links the sales of fridges and the sales of durable goods to the time of year
- Ignoring, for the moment, data on the sales of durable goods suppose you want to fit regression and analysis of variance models to link fridge sales to the time of the year
- There are two basic ways this can be achieved

1. A regression approach using the command `lm`
2. An analysis of variance (ANOVA) approach using the command `aov`

2.2 Reading the data into R

- Data is in the file `ancova.txt`

```
ancova<-read.table(‘‘E:ancova.txt")
fridge<-ancova[,1]
durables<-ancova[,2]
q1<-ancova[,3]
q2<-ancova[,4]
q3<-ancova[,5]
q4<-ancova[,6]
quarter<-ancova[,7]
```

2.3 Processing dummy variables in R

- Using the commands on the previous slide the variables `q1, q2, q3` and `q4` take the values 0 and 1
- No further data processing is needed for these though only three of these dummy variables can be included into the regression model if you also include an intercept term
- In order to follow the second, more efficient, approach you need to tell R that the variable `quarter` is a qualitative variable
- In R the command to do this is `factor`

```
quarter<-factor(quarter)
```

2.4 Regression and ANOVA

- There are two ways of fitting regression and ANOVA models in R
  1. A regression approach using the command `lm`
  2. An analysis of variance approach using the command `aov`
- We show that both approaches lead to the same numerical answers in our example

2.5 Regression approach

- Using a dummy variable approach (and arbitrarily excluding q1)

```
dummy.lm<-lm(fridge~q2+q3+q4)
summary(dummy.lm)
```

- Using the more advanced `factor` command

```
factor.lm<-lm(fridge~quarter)
summary(factor.lm)
```

- **Get the same numerical answers using both approaches**

2.6 Regression output for the dummy variable approach

```
Coefficients:
Estimate Std.  Error t value Pr(>|t|)
(Intercept) 1222.12 59.99 20.372 < 2e-16 ***
q2 245.38 84.84 2.892 0.007320 **
q3 347.63 84.84 4.097 0.000323 ***
q4 -62.12 84.84 -0.732 0.470091
```

2.7 Regression output for the `factor` approach

```
Coefficients:
Estimate Std.  Error t value Pr(>|t|)
(Intercept) 1222.12 59.99 20.372 < 2e-16 ***
quarter2 245.38 84.84 2.892 0.007320 **
quarter3 347.63 84.84 4.097 0.000323 ***
quarter4 -62.12 84.84 -0.732 0.470091
```

2.8 ANOVA for the same problem

- The analysis using ANOVA needs you to make use of the command `factor`
- Using the R syntax in exactly the same way as before leads to the following analysis of variance table

```
factor.aov<-aov(fridge~quarter)
summary(factor.aov)
Df Sum Sq Mean Sq F value Pr(>F)
quarter 3 915636 305212 10.6 7.91e-05 ***
Residuals 28 806142 28791
```

- Reconstructing the above F-statistic by hand then shows you that regression and analysis of variance lead to the same answers despite cosmetic differences in how you interpret model results

2.9 Reconstructing the *F*-statistic from scratch

- We are testing the null hypothesis that including the four levels of the variable `quarter` does not improve upon the simple model with just a constant term
- In R define the regression model with just a constant term

```
null.lm<-lm(fridge~1)
```

- Then use the command `anova` to obtain the same values as above

```
anova(factor.lm, null.lm)
```

2.10 Hand calculation of the F-statistic

- Alternatively we can show the two models give the same answers by repeating the earlier calculations by hand
- Using `summary(factor.lm)` we can see that the R^2 value is 0.5318 and the residual degrees of freedom is given by $n - p = 28$
- The change in the degrees of freedom is 4-1=3 (equivalent to the number of levels minus 1)
- The F-statistic can thus be constructed as

$$F = \frac{\frac{\Delta \ R^2}{\Delta \text{ d.f.}}}{\frac{1-R^2}{n-p}} = \frac{(0.5318/3)}{(1-0.5318)/28} = \frac{0.1772667}{0.01672143} = 10.60117,$$

giving the same answer as above subject to minor rounding error

3.1 ANOVA models with two-qualitative variables

- In a similar way it is also possible to define ANOVA models with two qualitative X variables
- The classical term used is two-way Analysis of Variance
- Really, these models get most interesting with the introduction of interaction terms (as we shall see below)
- It is important to note that you are not constrained to have a limit on the number of X variables
- Can have 3-way ANOVA, 4-way ANOVA etc. (though the 2-way ANOVA is sufficient to illustrate the general principles)
- The only real constraint would be that high-order interaction terms can prove rather difficult to interpret and so are not usually included into regression models

3.2 Two-way ANOVA example

• Adapting an example discussed in Gujarati and Porter (2009) suppose you have the following ANOVA model to investigate average hourly earnings in terms of gender and race:

$$Y_i = \beta_1 + \beta_2 D_{2,i} + \beta_3 D_{3,i} + u_i, \tag{2}$$

• $D_{2,i}$ is 1 if the respondent if female and 0 otherwise

• $D_{3,i}$ is 1 if the respondent is non-white and non-hispanic and 0 otherwise

• **How would you interpret the results of this model?**

3.3 Model interpretation I

$$Y_i = \beta_1 + \beta_2 D_{2,i} + \beta_3 D_{3,i} + u_i,$$

- Male white/hispanic. Average hourly wage given by

$$\beta_1 + \beta_2(0) + \beta_3(0) = \beta_1$$

- Female white/hispanic. Average hourly wage given by

$$\beta_1 + \beta_2(1) + \beta_3(0) = \beta_1 + \beta_2$$

- If $\beta_2 < 0$ the suggestion would be that females are generally paid less

3.4 Model interpretation II

$$Y_i = \beta_1 + \beta_2 D_{2,i} + \beta_3 D_{3,i} + u_i,$$

• Male non-white and non-hispanic. Average hourly wage given by

$$\beta_1 + \beta_2(0) + \beta_3(1) = \beta_1 + \beta_3$$

• Female non-white and non-hispanic. Average hourly wage given by

$$\beta_1 + \beta_2(1) + \beta_3(1) = \beta_1 + \beta_2 + \beta_3$$

• Compare, for example, with the average wage of β_1 for a male who is white or hispanic

• Suggestion is that if $\beta_3 < 0$ then non-white and non-hispanic workers are generally paid less

3.5 Two-way ANOVA in R

• There are at least two ways of fitting Two-way ANOVA models in R

- `lm` for linear model
- `aov` for analysis of variance

• This reflects that ANOVA models can be seen as a special case of linear regression models (Bingham and Fry, 2010)

• In both cases you need to define the X variables used as qualitative variables or factors. For example for the above wages example you would use

```
gender<-factor(gender)
race<-factor(race)
```

3.6 Two-way ANOVA via the `lm` and `aov` commands

- Both approaches use the familiar two-step approach
  1. A computational modelling step that generates no output
  2. Use the command `summary` to explicitly show you the results
- **For a regression approach using**

```
lm
a.lm<-lm(wage∼gender+race)
summary(a.lm)
```

- **For an ANOVA approach using**

```
aov
a.aov<-aov(wage∼gender+race)
summary(a.aov)
```

4.1 ANOVA models with interactions

• Higher order models are possible but as before a second-order model is sufficient to illustrate the general principles

• In higher order models three-way interaction terms and higher are possible

• However, in practice, you may be unlikely to see such models as they can quickly become hard to interpret

• Recurring themes

- Models remain part of the general class of linear regression models
- The R commands are essentially the same
- Models can, as before, be fitted either using the `lm` or the `aov` commands

4.2 Interaction means multiply

• Interaction terms allow us to account for multiplicative as opposed to purely additive effects within the standard class of general linear regression models

• It is easy to see how you can just multiply two quantitative variables X_2 and X_3 to form an additional regressor $X_4 := X_2X_3$

• For qualitative variables the effect of the interaction term is to allow for more subtle effects.

• For example in the context of the gender and race example earlier. Is the gender bias encountered more extreme for non-white and non-hispanic women?

• **Because interaction means multiply this has a direct effect on the R commands used**

1. To fit the full second-order model with interactions use $X_2{*}X_3$
2. To add a specific interaction term use $+X_2 : X_3$

4.3 A two-way ANOVA with interaction model

• Recall our previous example exploring how the hourly wage Y depends on gender and race

• From first principles a two-way ANOVA with interactions model can be constructed as

$$Y_i = \beta_1 + \beta_2 D_{2,i} + \beta_3 D_{3,i} + \beta_4 (D_{2,i} D_{3,i}) + u_i, \tag{3}$$

• $D_{2,i}$ is 1 if the respondent if female and 0 otherwise

• $D_{3,i}$ is 1 if the respondent is non-white and non-hispanic and 0 otherwise

• If $\beta_4 \neq 0$ then there is evidence of an interaction between gender and race

• If $\beta_4 = 0$ then equation (3) reduces to the simple two-way ANOVA model shown in equation (2)

• Apart from the above how would you interpret the results of this model?

4.4 Model interpretation I

$$Y_i = \beta_1 + \beta_2 D_{2,i} + \beta_3 D_{3,i} + \beta_4 (D_{2,i} D_{3,i}) + u_i,$$

• Male white/hispanic. Average hourly wage given by

$$\beta_1 + \beta_2(0) + \beta_3(0) + \beta_4(0) = \beta_1$$

• Female white/hispanic. Average hourly wage given by

$$\beta_1 + \beta_2(1) + \beta_3(0) + \beta_4(0) = \beta_1 + \beta_2$$

• If $\beta_2 < 0$ the suggestion would be that females are generally paid less

• Get exactly the same results as last time. The only difference is if both gender **AND** race effects are present.

4.5 Model interpretation II

$$Y_i = \beta_1 + \beta_2 D_{2,i} + \beta_3 D_{3,i} + \beta_4 (D_{2,i} D_{3,i}) + u_i,$$

- Female white/hispanic. Average hourly wage given by

$$\beta_1 + \beta_2(1) + \beta_3(0) + \beta_4(0) = \beta_1 + \beta_4$$

- Female non-white and non-hispanic. Average hourly wage given by

$$\beta_1 + \beta_2(1) + \beta_3(1) + \beta_4(1) = \beta_1 + \beta_2 + \beta_3 + \beta_4$$

- If $\beta_4 < 0$ the suggestion is that there is an additional racial effect that women workers are subjected to

4.6 Two-way ANOVA in R

- There are at least two ways of fitting Two-way ANOVA with interaction models in R
  - `lm` for linear model
  - `aov` for analysis of variance
- This reflects that ANOVA models can be seen as a special case of linear regression models (Bingham and Fry, 2010)
- In both cases you need to define the X variables used as qualitative variables or factors. For example for the above wages example you would use

```
gender<-factor(gender)
race<-factor(race)
```

4.7 Two-way ANOVA via the `lm` and `aov` commands

- **For a regression approach using**

`lm`

```
a1.lm<-lm(wage~gender*race)
summary(a1.lm)
```

Or

```
a2.lm<-lm(wage~gender+race+gender:race)
summary(a2.lm)
```

- **For an ANOVA approach using** `aov`

```
a1.aov<-aov(wage~gender*race)
summary(a1.aov)
```

Or

```
a2.aov<-aov(wage~gender+race+gender:race)
summary(a2.aov)
```

5.1 Analysis of Covariance (ANCOVA)

- Classical linear regression models include
  - Regular regression. All X variables quantitative
  - ANOVA. All X variables qualitative
- ANCOVA simply combines both of these cases
- ANCOVA refers to situations where regression models combine BOTH qualitative **AND** quantitative X variables

5.2 Proper ANCOVA

- Easy to envisage ANCOVA in terms of combining the mathematical and programming elements of previously defined regression models
- The main point of ANCOVA from a teaching perspective is it enables you to envisage regression models with different slopes and different intercept terms to be fitted to different parts of the data set
- These are sometimes referred to as segmented regression models as the effect is to potentially fit different regression lines to each segment of the data
- Consider an example with a qualitative variable Q that takes 2 levels (0 and 1) and a quantitative variable X
- Consider the regression model $Y = Q * X$ where $*$ means the interaction term and all main effects terms are present

5.3 ANCOVA interaction model

- The model $Y = Q * X$ leads to the regression model

$$Y = \beta_1 + \beta_2 D_Q + \beta_3 X + \beta_4 D_Q X + u$$

- If $Q = 0$

$$Y = \beta_1 + \beta_3 X + u$$

- If $Q = 1$ the result is a model with different intercepts and different slopes

$$Y = (\beta_1 + \beta_2) + (\beta_3 X + \beta_4) X + u$$

- If $Q = 1$ and $\beta_4 = 0$ the result is a model with a different intercept but the same slope term

$$Y = (\beta_1 + \beta_2) + \beta_4 X + u$$

5.4 An ANCOVA example

- Consider again the data shown in Slide 1.3 linking fridge sales (F) to durable goods expenditure (D)
- It is natural to fit a regression model of the form

$$F = \beta_1 + \beta_2 D + u. \tag{4}$$

- However, suppose we want to fit a regression model of the form shown in equation (4) in such a way as the values of β_1 and β_2 can potentially change depending on the time of the year
- The effect can be achieved by fitting a regression model of the form $Y = \text{Quarter} * D$ where Quarter is a categorical variable describing the time of the year

5.5 Entering the data into R

```
ancova<-read.table(‘‘E:ancova.txt")
fridge<-ancova[,1]
durables<-ancova[,2]
q1<-ancova[,3]
q2<-ancova[,4]
q3<-ancova[,5]
q4<-ancova[,6]
quarter<-ancova[,7]
```

- **Then need to tell R that quarter is a factor variable**

```
quarter<-factor(quarter)
```

5.6 Analysis of ANCOVA example I

- First fit the model with all interaction terms and compare with the simpler model with just the main effects terms present
- The non-signficant result shows that the simpler model without the interaction terms should suffice

```
ancova.lm<-lm(fridge∼quarter*durables)
main.lm<-lm(fridge∼durables+quarter)
anova(ancova.lm, main.lm)
Analysis of Variance Table
Res.Df RSS Df Sum of Sq F Pr(>F)
1 24 430992
2 27 465085 -3 -34093 0.6328 0.601
```

5.7 Analysis of ANCOVA example II

- Next fit the simpler model still with no quarterly term in it
- The significant results shows that the more complex model with the quarterly terms in it is required

```
simple.lm<-lm(fridge∼durables)
anova(main.lm, simple.lm)
Analysis of Variance Table
Res.Df RSS Df Sum of Sq F Pr(>F)
1 27 465085
2 30 1377145 -3 -912060 17.65 1.523e-06 ***
```

5.8 Regression output for the final model

```
Coefficients:
Estimate Std.  Error t value Pr(>|t|)
(Intercept) 456.2440 178.2652 2.559 0.016404 *
durables 2.7734 0.6233 4.450 0.000134 ***
quarter2 242.4976 65.6259 3.695 0.000986 ***
quarter3 325.2643 65.8148 4.942 3.56e-05 ***
quarter4 -86.0804 65.8432 -1.307 0.202116
```

5.9 A first segmented linear regression model

- Results on the previous slide suggest a model with different intercepts but the same slope term is appropriate
- In Quarter 1 the appropriate regression line is

$$F = 456.2440 + 2.7734D$$

- In Quarter 2 the appropriate regression line is

$$F = 456.2440 + 242.4976 + 2.7734D = 698.7416 + 2.7734D$$

- In Quarter 3 the appropriate regression line is

$$F = 456.2440 + 325.2643 + 2.7734D = 781.5083 + 2.7734D$$

- In Quarter 4 the appropriate regression line is

$$F = 456.2440 - 86.0804 + 2.7734D = 370.1636 + 2.7734D$$

6.1 Summary

• Dummy variable regression models occur when the dependent Y remains a quantitative measurement but some of the X variables are qualitative and denote membership categories rather than numerical measurements

• Some of the terminology can be confusing but essentially

- ANOVA all the X variables are qualitative
- ANCOVA all the X variables are a mixture of qualitative and quantitative
- Essentially the same regression theory and programming applies in each case

6.2 Additional overview: the `factor` command

- Often econometric textbooks give explicit examples of dummy-variable construction (see e.g. Gujarati and Porter, 2009)
- However, this approach is best suited for teaching purposes and not really best suited for analysing real datasets
- Care has to be taken to avoid the dummy variable trap. A variable with m categories needs $m-1$ dummy variables associated to it if an intercept term is also fitted
- In R qualitative variables can automatically be incorporated using the command `factor`
- In this case interpretation of the regression output etc. is the same but without the hassle of defining dummy variables
- Use of the R command `factor` is also useful to demonstrate the importance of abstract thinking in cases when the size of the dataset may make it inconvenient to look a spreadsheet of the entire data