Class 4: Multiple Regression and Categorical Variables

Table of Contents / Agenda

1 Categorical Variables

1.1 Use of Dummy Variables

- To capture the effect of categorical variables
  - Brands, In-store displays, Gender
- Dummy variable has a value of 0 or 1
  - 1 indicates presence of characteristic
  - 0 indicates absence of characteristic

1.2 Example

Sales	Store Type
10	A
4	B
8	A
6	B
7	A
6	B
7	B
8	A

- Categorical variables require recoding
- Use indicator variables / dummy variables

Sales	Store Type	Dummy
10	A	1
4	B	0
8	A	1
6	B	0
7	A	1
6	B	0
7	B	0
8	A	1

- Sales Estimate = 5.75 + 2.5 × (if store type is A).
- Note that this gives a **relative** measure.
- Store type A sales are estimated to be 2.5 units **more than** store type B.

1.3 Coding Dummy Variables

- If a category can either be present or absent, then code:
  - Presence as 1
  - Absence as 0
  - Example: Presence of "In Store Display"
- If a category can be of two types:
  - Code one of the category as 1
  - Code the other as 0
  - Example: Male/ Female; Cash/ Credit

1.4 Coding Dummy Variables: An Example

- Do male teachers get more wage in general?
- Are Texas drivers more likely to buy a pickup truck compared to drivers in other states?

1.4.1 Model:

- Let D_i be the dummy variable. Then, when it is true ($D_i = 1$), the model is:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 D_i \\ &= \underbrace{(\beta_0 + \beta_2)}_{\text{intercept}} + \beta_1 x_{1i} \end{aligned}$$

- When it is not true ($D_i = 0$), the model is:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 D_i \\ &= \beta_0 + \beta_1 x_{1i} \end{aligned}$$

- So β_2 represents the relative difference between the two groups in terms of their intercepts
- What does it mean when β_2 is not significant?

1.5 Dummy coding with more than 2 categories (L levels)

- At the most, $L - 1$ variables are needed
- Choose a base (comparison) variable
- Code each variable as being the category or not ...

Sales	REGION	R2ornot	R3ornot
10	1	0	0
4	2	1	0
8	1	0	0
6	2	1	0
7	3	0	1
6	3	0	1
7	3	0	1
8	1	0	0

1.6 Dummy Coding for Multi-Category

what if we have more than one category?
e.g., color = { **red**, **green**, **blue**} is independent variable (x) and preference is dependent variable (y)
use a separate dummy variable for each category, except one (e.g., the last)

$$\begin{aligned} &\text{color is } \textbf{red}: && D_{i1} = 1, D_{i2} = 0 \\ &\text{color is } \textbf{green}: && D_{i1} = 0, D_{i2} = 1 \\ &\text{color is } \textbf{blue}: && D_{i1} = 0, D_{i2} = 0 \end{aligned}$$

$$y_i = \beta_0 + \beta_1 D_{i1} + \beta_2 D_{i2} = \begin{cases} \beta_0 + \beta_1 & \text{if red} \\ \beta_0 + \beta_2 & \text{if green} \\ \beta_0 & \text{if blue} \end{cases}$$

1.6.1 Interpretation

$$y_i = \beta_0 + \beta_1 D_{i1} + \beta_2 D_{i2} = \begin{cases} \beta_0 + \beta_1 & \text{if red} \\ \beta_0 + \beta_2 & \text{if green} \\ \beta_0 & \text{if blue} \end{cases}$$

- β_0 preference of product if **blue** (blue is called the `baseline level`)
- β_1 preference of product if **red** as **compared to blue** product: "how much better (worse) is red product liked over blue"
- β_2 preference of product if **green** as **compared to blue** product: "how much better (worse) is green product liked over blue"

1.7 Another Example

- Brands `{ =Sony`, `Samsung`, `Bose`}
- Use a separate dummy variable for each brand, except one (e.g. the last one)
  - D_{Sony}, $D_{Samsung}$
- Dummy Coded Variables

Brand	Brand Code	D_{Sony}	$D_{Samsung}$
Sony	1	1	0
Samsung	2	0	1
Bose	3	0	0

- What is the baseline in this example?
- Let's say we have the following model to predict sales:

$$Sales = \beta_0 + \beta_1 \times Price + \beta_2 \times Ad + \beta_3 \times D_{Sony} + \beta_4 \times D_{Samsung}$$

- Then, sales for each brand is:
- $Sales_{Sony} = \beta_0 + \beta_1 \times Price_{Sony} + \beta_2 \times Ad_{Sony} + \beta_3$
- $Sales_{Samsung} = \beta_0 + \beta_1 \times Price_{Samsung} + \beta_2 \times Ad_{Samsung} + \beta_4$
- $Sales_{Bose} = \beta_0 + \beta_1 \times Price_{Bose} + \beta_2 \times Ad_{Bose}$

2 Multicollinearity

- Why do we use $L - 1$ variables instead of L in dummy coding?
- If you do, you will get **perfect multicollinearity**
- What is multicollinearity?

2.1 Multicollinearity

- Source: Two or more independent (x_k) variables in a multiple regression model are highly correlated
- Since two x_k's are moving together, it is hard to identify which one is causing the changes in y

2.2 Consequences of Multicollinearity

- Estimates of the effect (coefficients) are less precise
- Small `t-stat` (= large `p-value`)
- `Type 2 Error`: you do not reject the null ($H_0 : \beta = 0$) when you should
- But does **not** actually bias results

2.3 Fixes

- This is a data problem. If you have sufficient number of observations, high correlation between explanatory (predictor) variables is okay
  - Standard Errors for estimates become smaller as you increase number of sample

2.4 Perfect multicollinearity

- You have complete dependency among variables (predict one with others)
- Inversion in OLS estimate formula does not work and you cannot estimate the model
- Just like $1/0$ does not work
- Not a big problem - you will see the error right away

2.4.1 Dummy Variable Trap

- If you have L dummies for L number of categories, including a constant term in the regression together guarantee perfect multicollinearity
- Analogous to this is that when you know the mean first $n-1$ observations then you can infer n'th observation

3 Making Predictions in Regression Models

Once you have regression results (estimated coefficients, $\widehat{\beta}_k$'s), it is easy to make predictions given values of x_k's.

- Remember we are using the linear model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \varepsilon_i$$

- For example, estimation results can be:

$$y_i = \underbrace{10}_{\widehat{\beta}_0} + \underbrace{3}_{\widehat{\beta}_1} x_{i1} + \underbrace{3}_{\widehat{\beta}_2} x_{i2}$$

- Once we have $\widehat{\beta}_k$'s, given x_k values, we can calculate the **predicted** value of y, $\widehat{y}$ by plugging in those estimates:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + 0$$

  (Because $\hat{\varepsilon}_i = E[\varepsilon_i] = 0$)

- For example, if your estimation results are:

$$y_i = 10 + 3x_{i1} + 3x_{i2}$$

- The estimate of y for values of $x_1 = 5, x_2 = 4$ is:

$$\widehat{y} = 10 + 3 \times \underbrace{5}_{x_1} + 3 \times \underbrace{4}_{x_2} = 37$$