Class notes 5

Regression with nominal predictors

Introduction

Until now, we have focused entirely on variables measured on a metric scale. How can we incorporate qualitative data measured on a nominal scale into the model? Examples of qualitative data are

- Race
- Gender
- Customer segment
- Sector (the industry of a firm)
- Geographic region (i.e. country)

To aid the understanding on how to use of dummy variables in regression analysis, it will be useful to consider a specific example. Suppose we want study wages and how wages depend on various background variables. One such background variable is *gender*, which is a variable that can take the values, *Male* or *Female*. In order to incorporate gender into the model, we need to construct a binary variable, denoted D, that takes the value 0 or the value 1. We may attach the value of 0 to *Male* and the value of 1 to *Female*. In that case, we say that *Male* represents the *reference category*. Formally, gender is defined in the following way

$$D = \begin{cases} 1 & \text{if Female} \\ 0 & \text{if Male} \end{cases}$$

Intercept dummies

We now consider the two-variable regression model

$$Wage = B_0 + B_1 D + \varepsilon$$

Recall that a regression model is a model for the conditional mean. Let us consider the mean wage conditional on $D = 0$

$$\begin{aligned} E(Wage|D = 0) &= B_0 + B_1 \cdot 0 \\ &= B_0 \end{aligned}$$

The equation expresses the mean wage for men, which is B_0. Next, consider the mean wage when $D = 1$

$$\begin{aligned} E(Wage|D = 1) &= B_0 + B_1 \cdot 1 \\ &= B_0 + B_1 \end{aligned}$$

This equation expresses the mean wage for females, which is $B_0 + B_1$. Summarizing the analysis, we make the following observations

- The parameter B_1 expresses the difference in mean wage across gender. For instance, suppose that B_1 is smaller than zero. We then have $B_0 > B_0 + B_1$, and the conclusion is that

the mean wage for females is lower than the mean wage for males. On the other hand, if B_1 is larger than zero, we conclude that the mean wage for males is lower than the mean wage for females.

- To establish statistical difference across gender involves estimating the model, from which we can apply the (standard) t-test to evaluate the hypothesis

$$H_0 : B_1 = 0$$
$$H_A : B_1 \neq 0$$

- The model is also known as a one-way ANOVA model (analysis of variance), and the t-test of B_1 is simply a two-sample t-test for the difference in means.

Let us extend the simple ANOVA-model and include *education* as a predictor. The model is now

$$Wage = B_0 + B_1 D + B_2 Educ + \varepsilon$$

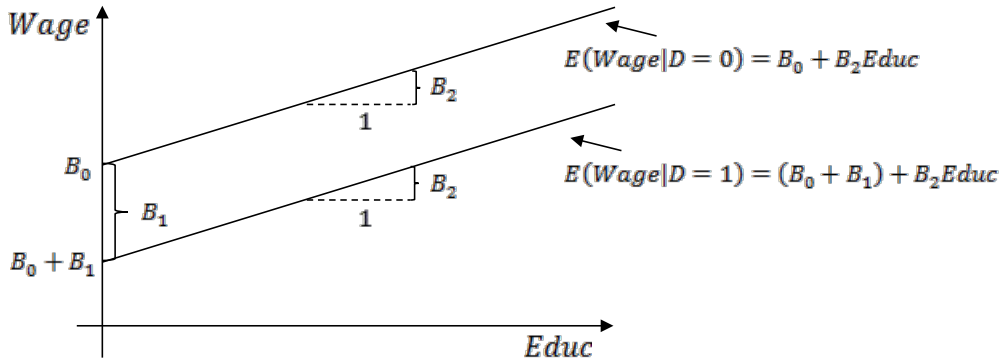
As before, we study the mean for difference values of D (the conditional mean). First, let $D = 0$, which gives

$$\begin{aligned} E(Wage|D = 0) &= B_0 + B_1 \cdot 0 + B_2 Educ \\ &= B_0 + B_2 Educ \end{aligned}$$

This equation is simply the mean wage among men (given some educational attainment). Next, we let $D = 1$, from which we have

$$\begin{aligned} E(Wage|D = 1) &= B_0 + B_1 \cdot 1 + B_2 Educ \\ &= (B_0 + B_1) + B_2 Educ \end{aligned}$$

The equation is the mean wage among woman (given some educational attainment). The two equations are illustrated below:



(note that in the illustration, $B_1 < 0$)

Summarizing the content, we make the following observations:

- The parameter B_1 expresses the difference in mean wage among gender (given some educational attainment).

- Regardless of the value of D (regardless of gender), the slope parameter B_2 stays the same in the two equations. Thus, the change in wage for a one-unit increase in educational attainment is the same for both men and woman. Simply put, the reward associated with one more year of education is the same across gender.
- As before, establishing statistical difference across gender involves estimating the model and evaluating the hypothesis

$$H_0 : B_1 = 0$$
$$H_A : B_1 \neq 0$$

- The model is known as an ANCOVA-model (analysis of covariance).

Slope dummies

Thus far, our focus has been on dummy variables that relates to the constant term in the model. Dummy variables that allows for variation in the constant term (or the intercept) are referred to as intercept dummies. We shall extend the framework and allow for different slopes as well. Such dummies are referred to as slope dummies or interaction dummies. The dummy variable itself is the same as before. The difference is how we use it.

Consider the following model

$$Wage = B_0 + B_1 D + B_2 Educ + B_3 D \cdot Educ + \varepsilon$$

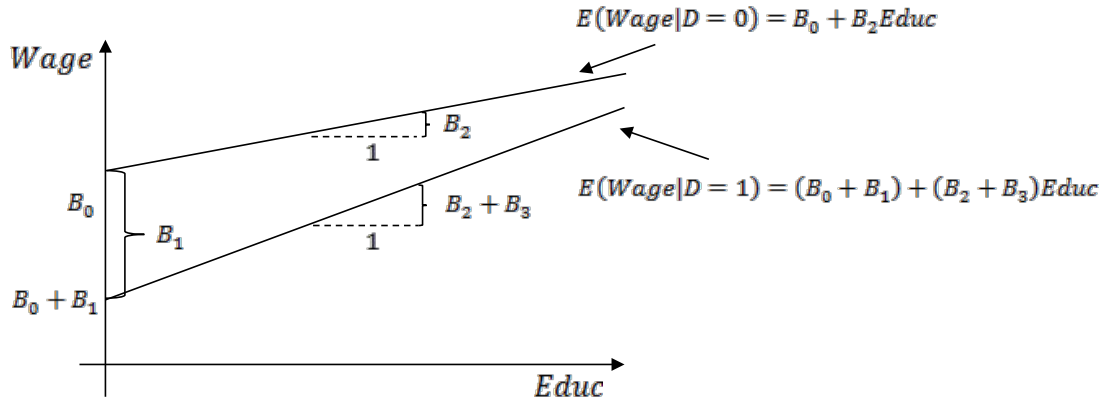
This model contains both an intercept dummy and a slope dummy. The term $D \cdot Educ$ is called an interaction term. Again, we consider the mean conditional on D. First, let $D = 0$

$$\begin{aligned} E(Wage|D = 0) &= B_0 + B_1 \cdot 0 + B_2 Educ + B_3 \cdot 0 \cdot Educ \\ &= B_0 + B_2 Educ \end{aligned}$$

As previously, this equation is simply the mean wage among men (given some educational attainment). Next, consider the case when $D = 1$

$$\begin{aligned} E(Wage|D = 1) &= B_0 + B_1 \cdot 1 + B_2 Educ + B_3 \cdot 1 \cdot Educ \\ &= (B_0 + B_1) + (B_2 + B_3) Educ \end{aligned}$$

The equation is the mean wage among woman (given some educational attainment). Below, we illustrate the difference between the two equations. For illustrative purposes, let $B_1 < 0$ and $B_3 > 0$.



Summing up, we have

- The parameters B_1 and B_3 expresses the differences in mean wage among gender (given some educational attainment).
- In this case, we allow for difference in slope across gender. Thus, the change in wage for a one-unit increase in educational attainment may be different for men and women. Simply put, the reward associated with one more year of education may be different across gender.
- Establishing statistical difference across gender involves estimating the model and evaluating the joint hypothesis

$$H_0 : B_1 = 0,\ B_3 = 0,,$$
$$H_A : \text{At least one } B_j \neq 0 \ \text{ for } j = 1, 3$$

More than Two Categories

We shall now extend the dummy variable approach even further. As we have seen, using dummy variables, it is straightforward to incorporate qualitative information such as gender into the regression model. What about variables that have more than two categories? This problem is addressed by creating several dummy variables. In general, if we have m categories, we need to create $m-1$ dummy variables. The last (the m-th) category is the reference category and is represented by the constant term. Again, consider the wage model presented before. In the example, it was shown how a single dummy variable is sufficient to handle the two-categorical variable *gender*.

Let us now consider an example with more than two categories. Let *timedrs* denote the frequency of visits to a health professional (for instance a doctor). In some countries, the demand for health care is satisfied by non-governmental entities such as private hospitals and health care firms. Thus, one would conjecture that $income$ is an important predictor of *timedrs*

$$timedrs = B_0 + B_1 income + \varepsilon$$

It is preferable to have the exact value of income for every individual in the data. However, it is often the case that individuals are divided into groups according to their income level. For instance, suppose that individuals in the data are divided into three groups (or subsamples) with the different income levels denoted *low-income*, *middle-income* and *high-income*. It is then sufficient to specify two dummy variables to handle the three income levels.

The dummy variables are specified in the following way:

$$D_1 = \begin{cases} 1 & \text{if low income group} \\ 0 & \text{Otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{if middle income group} \\ 0 & \text{Otherwise} \end{cases}$$

The reference category is then the high-income group. The model becomes

$$timedrs = B_0 + B_1 D_1 + B_2 D_2 + \varepsilon$$

Writing out the conditional means involves evaluating the expected number of visits to a health professional for different combinations of D_1 and D_2. First, consider the mean number of visits to a health professional for the group belonging to the high-income level (the reference category)

$$\begin{aligned} E(timedrs|D_1 = D_2 = 0) &= B_0 + B_1 \cdot 0 + B_1 \cdot 0 \\ &= B_0 \end{aligned}$$

The mean number of visits for the middle-income group is

$$\begin{aligned} E(timedrs|D_1 = 0, D_2 = 1) &= B_0 + B_1 \cdot 1 + B_2 \cdot 0 \\ &= B_0 + B_1 \end{aligned}$$

Finally, we have the mean number of visits for the low-income group

$$\begin{aligned} E(timedrs|D_1 = 1, D_2 = 0) &= B_0 + B_1 \cdot 0 + B_2 \cdot 1 \\ &= B_0 + B_2 \end{aligned}$$

In the special case when $B_1 = B_2 = 0$, the mean number of visits is the same for all income levels.

Suppose that we want to test if the mean number of visits to a health professional is the same across all income levels. Let μ_1, μ_2 and μ_3 denote the mean number of visits for the low-income group, the middle-income group and the high-income group, respectively. The test can be performed using the ANOVA-framework. Using ANOVA in this case involves evaluating the following hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_A : \text{At least one } \mu_i \neq \mu_j \text{ for } i \neq j$$

Here, the alternative hypothesis states that at least one of the means is different. Evaluating this hypothesis is equivalent to evaluating the following hypothesis

$$H_0 : B_1 = B_2 = 0$$
$$H_A : \text{At least one } B_j \neq 0 \text{ for } j = 1,2$$

Evaluating this hypothesis is a straightforward application of the *F*-test. Thus, ANOVA is just a special case of regression analysis where the predictors are dummy variables.

The model can be extended to an ANCOVA-model by including slope dummies. Suppose we are able to quantify the individual's state of physical health (here denoted *phyheal*). This variable is an obvious predictor of the number of visits to a health professional.

First, consider the model

$$timedrs = B_0 + B_1 D_1 + B_2 D_2 + B_3 phyheal + \varepsilon$$

Extending the model by including slope dummies, we obtain

$$timedrs = B_0 + B_1 D_1 + B_2 D_2 + B_3 phyheal +$$
$$B_4 D_1 \cdot phyheal + B_5 D_2 \cdot phyheal + \varepsilon$$

As previously, writing out the conditional means involves evaluating the expected number of visits to a health professional for different combinations of D_1 and D_2, for some level of physical health .