

Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 05
Categorical Variable as Predictor Part - 1

Welcome to the Predictive Analytics – Regression and Classification course. In this lecture, we are going to discuss how a Categorical Variable as Predictor is going to play in the linear model setup and typically these models sometimes called ANOVA. And, in a special case it is called ANOVA, but overall it is part of the linear model setups and regression also is part of the linear model setup.

(Refer Slide Time: 00:53)

Experiment



An experiment performed to assess the relative effect of three toxins and a control on the liver of a certain species of trout. The are about the amounts of deterioration (in standard units) of the liver in each sacrificed fish.

Toxin 1	Toxin 2	Toxin 3	Control
28	33	18	11
23	36	21	14
14	34	20	11
27	29	22	16

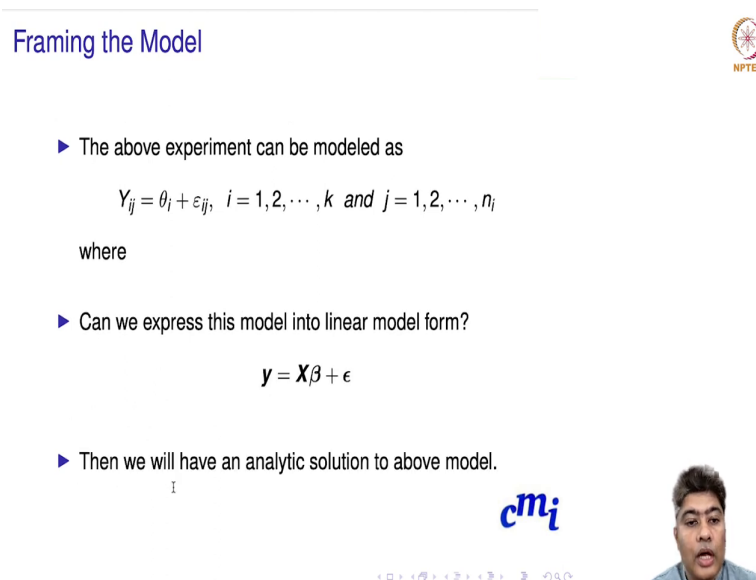


First, we will consider an experiment. So, suppose an experiment is performed to assess the relative effect of three toxins and the control on the liver of a certain species of trout. Trout is

So, little bit more clarification of the model. We can take sigma i square to be different we can take sigma i square to be different, but for a simplicity we are assuming sigma i square is equal to sigma square for all groups. So, all group has the similar variance. So, this assumption is called homoscedasticity assumption and then we assume that epsilon ij is following normal distribution with mean at 0 and variance sigma square.

So, this is the model setup that we are going to work out. So, this is the from the model we are framing from the experiment we are framing the model.

(Refer Slide Time: 04:33)



The slide is titled "Framing the Model" and contains the following content:

- ▶ The above experiment can be modeled as
$$Y_{ij} = \theta_i + \varepsilon_{ij}, \quad i = 1, 2, \dots, k \text{ and } j = 1, 2, \dots, n_i$$
where
- ▶ Can we express this model into linear model form?
$$y = X\beta + \epsilon$$
- ▶ Then we will have an analytic solution to above model.

The slide also features the NPTEL logo in the top right corner, the CMi logo in the bottom right corner, and a small video inset of a man speaking in the bottom right corner. Navigation icons are visible at the bottom of the slide.

Now, as we frame the model question is can we express this model into linear model form like y equal to x beta plus epsilon? Then we will have an analytic solution of the above model I mean this model will have a analytic solution.

(Refer Slide Time: 05:00)

Finding Solutions

We define the design matrix X as follows:

$$X = \begin{pmatrix} \text{Toxin 1} & \text{Toxin 2} & \text{Toxin 3} & \text{Control} \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$



cmj



So, first we have to define the X matrix or the design matrix. So, we have 16 observations for you can see there. So, we have 16 rows here each the first 4 rows of my data belongs to toxin group 1 and rest of the places called toxin 2 for the toxin 2 I created 0s, toxin 3 I created 0 and the control group 0s.

Similarly, the from row 5 to row 8, I have toxin group all 1s; toxin 2 goes to 1; then 3 is 0 and control group from row 5 to row 8 is 0. This is called dummy creation of dummy variable in statistics in machine learning often it is called one hot encoding. This is very useful this one hot encoding is very useful solving many problems and we will see soon how it is going to make our life very simple.

So, we have 16 rows for see each samples we have one representation and if it goes to toxin 1 we have 1, if it not if a particular row does not belong to that group then that row will get 0.

(Refer Slide Time: 06:42)

Finding Solutions

We have the response vector y as follows:

$$y = \begin{pmatrix} 28 \\ 23 \\ 14 \\ 27 \\ 33 \\ 36 \\ 34 \\ 29 \\ 18 \\ 21 \\ 20 \\ 22 \\ 11 \\ 14 \\ 11 \\ 16 \end{pmatrix}$$



cmj



All the observations are stacked over another. The first four row observations are belongs to toxin 1; from 5 to 8 observation number 5 to 8 belongs to toxin group 2 from 8 to 12 belongs to toxin group 3 and from 12 13 to 16 it group belongs to control group.

(Refer Slide Time: 07:14)

Finding Solutions



We can find the $\mathbf{X}^T \mathbf{X}$ as follows

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}$$

and $(\mathbf{X}^T \mathbf{X})^{-1}$ is

$$(\mathbf{X}^T \mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{pmatrix}$$

I

cmj



Then we calculated X transpose X turns out X transpose X in this case become very simple.

(Refer Slide Time: 07:24)

Finding Solutions

We define the design matrix X as follows:

$$X = \begin{pmatrix} \text{Toxin 1} & \text{Toxin 2} & \text{Toxin 3} & \text{Control} \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$



cmj



If you go back to the X what happens is $X^T X$ is essentially it $X^T X$ is kind of you know when it is first row first column will be just it is with itself. So, it will have created 4 and then 0 0 0 in this way the $X^T X$ creates this matrix. So, $X^T X$ inverse is very simple, just diagonals will be 1 out of 4 and the half diagonals are all 0. So, this is simple solution.

(Refer Slide Time: 08:01)

Finding Solutions



We can find the $\mathbf{X}^T \mathbf{X}$ as follows

$$\mathbf{X}^T \mathbf{y} = \begin{pmatrix} \sum_{j=1}^4 y_{1j} \\ \sum_{j=1}^4 y_{2j} \\ \sum_{j=1}^4 y_{3j} \\ \sum_{j=1}^4 y_{4j} \end{pmatrix}$$

Hence the solution is:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \begin{pmatrix} \frac{1}{4} \sum_{j=1}^4 y_{1j} \\ \frac{1}{4} \sum_{j=1}^4 y_{2j} \\ \frac{1}{4} \sum_{j=1}^4 y_{3j} \\ \frac{1}{4} \sum_{j=1}^4 y_{4j} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \\ \bar{y}_4 \end{pmatrix}$$

So the model yields group sample mean as the solution for $(\theta_1, \theta_2, \theta_3, \theta_4)$.

cmj



Now, we calculate X transpose y. X transpose y is if you look it carefully that sum of the first group of observations the second of element is the sum of the second group of observation and third is the third group of observation and fourth is the fourth group of observation fourth element.

Hence the solution is X transpose X inverse X transpose y. X transpose X inverse is all diagonal elements is 1 by 4 and off diagonals are 0. So, my beta hat is basically sample group mean y 1 bar on the of the first group. Similarly, second. So, y 2 bar is the sample group mean of the second group of the toxin group 2.

So, we also want to incorporate global mean for all data. In this case situation I would like you to take a pause, pause your video here for about 10 minutes think about it and try for yourself to develop the model with global mean.

(Refer Slide Time: 10:27)

A model with global mean



- ▶ We can try

$$Y_{ij} = \mu + \theta_i + \varepsilon_{ij}, \quad i = 1, 2, \dots, k \text{ and } j = 1, 2, \dots, n_i$$

where

- ▶ But we are going to face a problem. Can you identify the problem?
- ▶ Pause the video and think about it for 10 minutes.

cmj



◀ ▶ ⏪ ⏩ 🔍 🔄

I believe you have now got a possible solution. Let us see how can you solve the solutions get a solution. If one possible solution is you take Y_{ij} equal to μ μ is the sort of a global mean plus θ_i θ_i are the group means plus some ε_{ij} . Now, i equal to runs to 1 to k and j equal to 1s to n_i . This is a simple solution looks like, but we are going to face a problem.

Can you identify the problem? So, again I will request you to pause the video for a while and think about maybe for 10 minutes that what problem you are going to face this particular model.

(Refer Slide Time: 11:24)

A model with global mean

Our response vector y is still same. The change we see is in the design matrix X . Let us see the changed design matrix.

$$X = \begin{pmatrix} \text{Intercept} & \text{Toxin 1} & \text{Toxin 2} & \text{Toxin 3} & \text{Control} \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

cm_i



I believe now you have you have got the idea that what problem you are going to solve you are going to face and let us we know that in this case our response vector will still be same. The change we will see if you look into the model our response vector will going to be same. The only change that we are going to see is in the design matrix.

So, the changes that whatever design matrix we have had before that same design matrix we are still going to have, but now what is what we will have is we are going to have a fifth column called intercept, ok.

(Refer Slide Time: 12:36)

A model with global mean



The $X^T X$ matrix is

$$X^T X = \begin{pmatrix} 16 & 4 & 4 & 4 & 4 \\ 4 & 4 & 0 & 0 & 0 \\ 4 & 0 & 4 & 0 & 0 \\ 4 & 0 & 0 & 4 & 0 \\ 4 & 0 & 0 & 0 & 4 \end{pmatrix}$$

- ▶ Now if you look at carefully the first column of $X^T X$ is direct sum of 2nd, 3rd, 4th and 5th column. So $X^T X$ is not invertible.
- ▶ That means solution does not exists if we create dummy variable for each labels of categorical variables.

cmj



◀ ▶ ⏪ ⏩ 🔍 🔄

So, we are going to have a intercept column. Now, this is an interesting phenomena that if we have this design matrix then our X transpose X matrix is going to be 16 4 4 4 4 and then 4 4 0 0 0, then 4 0 4 0 0, 4 0 0 4 0, 4 0 0 0 4. In this matrix, if you look at very carefully the first column of this X transpose X is actually direct sum of second column, third column, fourth column and fifth column.

So, if you just add the second column, third column, fourth column and fifth column you will get the first column back. So, the first column is completely dependent on the second, third, fourth and fifth column. So, that means, X transpose X is not going to be invertible; that means, solution does not exist, if we create a dummy variable for each labels of categorical variable. I hope you understand the problem. Let me repeat.

We can let me go back to the model. In this model what we have? We have only effectively one predictor; predictor is treatment and there are four possible levels of the treatment toxin 1, toxin 2, toxin 3 and control. So, for each levels of the treatment so, there is only one predictor that is treatment in treatment level there are four possible levels. So, one categorical variable with four levels – toxin 1, toxin 2, toxin 3 and control group.

For each level if you create a if you create a for each level you create a dummy variable or one hot encoding, then what happens if and also you keep a intercept parameters then the in the X matrix the intercept parameter becomes completely dependent on all the columns of your predictor variables.

As a result one column will become completely dependent on some other columns and hence we will not be going to we are not going to have a invertible matrix or going to have a analytical solution at all. For this case we will solution does not exist. So, you have to be very careful about when you are going to handle categorical variables with different levels.

Let us stop here and we will continue on the next part.