

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

Lecture - 06
Categorical Variable as Predictor Part -2




Welcome to the Predictive Analytics Regression and Classification. In this lecture we are going to continue on how to model Categorical Variable as Predictor in linear model setup.

(Refer Slide Time: 00:32)

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






◀ ▶ 🔍 🔄

In this experiment we are continuing from the previous video that we are assessing the relative effect of three toxin groups; toxin 1, toxin 2 and toxin 3 with a control group on the lever of a certain species of trout. The area of and we are measuring the how the liver is condition is deteriorating with the effect of toxin.

(Refer Slide Time: 01:31)

A model with intercept as baseline

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


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


So, our response vector y in this case are still the same. The change we see is a design matrix let us see the change design matrix. In the change design matrix for the intercept for each observation we will have a intercept as 1 and then toxin 1 has 1 1 and 0 rest of them are 0, toxin 2 000 and then 111 and then 00. So, for each level we will. So, whenever whatever observation belongs to that group we will have 1 in that level and rest of the things are 0, this is called typically one hot encoding or dummy variable.

In the previous video we found that if you create the level for each of the control design each of the dummy variable for each of the levels of your predictor variable in this case treatment, then it is going to give you a trouble in a sense your design matrix is going to be have a less than full rank. As a result your design matrix is will not be invertible and you will not have a solution unique solution.

(Refer Slide Time: 02:54)




A model with baseline

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.






So, in this case if you see carefully that actually the first column is essentially sum of the four columns, if you just column toxin 1, toxin 2, toxin 3 and control if you add all these columns you will get the intercept column and that is mainly the problem; that means, your first column is a is fully dependent on the 2nd, 3rd, 4th and 5th.

So, if you add directly some 2nd column plus 3rd column plus 4th column and plus 5th column what you will get? The 1st column back, this is not good thing because in this case you will get a $x^T x$ which is again the same problem. The first column of $x^T x$ is simply some of the first 2nd, 3rd, 4th and 5th column and it is not invertible. So, you will not have a unique solution.

(Refer Slide Time: 03:55)

A model with baseline

One solution is to drop one label from the X , i.e.,

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


So, solution to this is you sacrifice one of the level. So, in this case we are sacrificing control as a last column control from the design matrix. Now, this is our modified design matrix. In this case we have 4 columns and now you can see the first column is not anymore sum of the 2nd, 3rd and 4th column. Because in the last 4 rows if you add them up its not you are not getting back the values of the intercept first column.

So, this is that is how we are breaking the dependency between the columns. So, we are just sacrificing one level whatever level you can sacrifice toxin 1 you can sacrifice toxin 2 for this case we are sacrificing the last level the control group.

(Refer Slide Time: 04:51)



A model with baseline

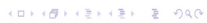
The $X^T X$ matrix is

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

The $(X^T X)^{-1}$ is

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$






This gives us the x transpose x matrix 16 4 4 4, 4 4 0 0, 4 0 4 0 and 4 0 0 4 and we have a nice x transpose x inverse in this case.

(Refer Slide Time: 05:05)

Try yourself

Find $\hat{\beta}_i = (X^T X)^{-1} X^T y$






Navigation icons: back, forward, search, etc.

And now my request is to try yourself solve this and find the beta hat by hand. So, pause the video try to solve the beta hat and see if the solution matches with the my solution. I hope now you are back and we can now look into the solution check if your solution matches with my solution.

(Refer Slide Time: 05:37)

Try yourself


$$\text{Find } \hat{\beta}_i = \begin{pmatrix} 13 \\ 10 \\ 20 \\ 7.25 \end{pmatrix} = \begin{pmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}$$


Navigation icons: back, forward, search, etc.




So, this is my solution 13 10 20 and 7.25, 13 is mu hat, 10 is beta 1 hat, 20 is beta 2 hat and 7.25 is beta three hat I hope your solution matches with my solution.

(Refer Slide Time: 05:56)

Try yourself

Find $\hat{\beta} = \begin{pmatrix} 13 \\ 10 \\ 20 \\ 7.25 \end{pmatrix} = \begin{pmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}$

- ▶ For Toxin 1:
 $\hat{\theta}_1 = \hat{\mu} + \hat{\beta}_1 = 23$
- ▶ For Toxin 2:
 $\hat{\theta}_2 = \hat{\mu} + \hat{\beta}_2 = 33$
- ▶ For Toxin 3:
 $\hat{\theta}_3 = \hat{\mu} + \hat{\beta}_3 = 20.5$
- ▶ For Control:
 $\hat{\theta}_4 = \hat{\mu} = 13$



So, for toxin 1 if theta 1 hat is mu hat plus beta 1 hat which is 23, for toxin 2 theta 2 hat equal to mu hat plus beta 2 hat is 33 for toxin 3 theta 3 hat the mu hat plus beta 3 hat is 20.5. So, you can see carefully this 23, 33 and 20.5 is actually the sample mean or group mean of that particular toxin group.

You can go and check it out and for theta 4 hat is mu hat is 13 which is the sample mean of the group of the control group. So, mu hat is essentially representing the control group if you sacrifice the. So, that why you do not need the mu the baseline. So, mu hat is creating as a baseline and then beta 1 hat is the difference between the baseline and the toxin 1 group, beta 2 hat is the difference between the baseline and the toxin 2 group and beta 3 hat is the difference between the baseline and the toxin 3 group.

