In this module we'll talk about how we build a linear regression model using R. Let us read a dataset inside R. Let us try to understand what this data is all about.

>>>

This is a direct marketing dataset and this dataset is from a direct marketer who sells his products only via direct mail. He sends catalogues with product characteristics to customers who then order directly from the catalogues. The marketer has developed customer records to learn what makes some customers spend more than others. The dataset has one thousand customer records with information on the age of customers categorised as old, middle and young customers. Information on the gender of customers as males and females, information about whether the customers own a house or not own a house, are they single or married, how far do the customers live away from the Brick and Mortar stores that sell the similar kind of products that this direct marketer also sells. The salary of the people or the customers, how many children they have ranging from zero to three, what has been their past history in terms of volume of purchase and how many catalogues this direct marketer has sent to these customers. We also have information about the amount spent by each of these customers. What we want to do is, we want to build a linear regression model which would predict the amount spent by each of these customers based on these attributes.

© Jigsaw Academy Education Pvt Ltd

>>

Before we proceed ahead with the task of building a linear regression model, the first thing we'll do is we'll explore our data and then prepare our data for the linear regression task. In order to explore the data better, we will be using packages such as "dplyr" which would help us in manipulating a data and also the visualisation library "ggplot2". We will use a special library called "car" which has relevant functions, which we will need when we perform the linear regression. Let us load these libraries.

>>

Let us start with the first variable in our dataset which is the age variable. Since age is a categorical variable here, what I'll do is, I'll try to understand how this continuous variable which is the amount spent, which I to predict, varies with each of the levels of this categorical variable.

>>

So let us produce a box plot for all the categories in this variable.

>>

So what I see here is the amount spent by each of the categories. Now I can see people of middle age and old age, they have similar behaviour in terms of amount spent and people who are young they spend a lot less compared to these two categories.

© Jigsaw Academy Education Pvt Ltd

Since these two categories behave similarly and the only difference that I will observe in my linear regression model would be with respect to if the person is young or not. What I want to do is I want to combine these two categories in my data and treat them as a single category.

>>>

So, let us do that. Here, I'm combining the middle and old age category people as one category. So now in my dataset I just have two categories, one category belonging to people who are middle aged and old aged, and one category belonging to people who are young. Let us now again take a look at the box plot.

>>>

And I can see that the means of these two groups are different. So if I'll run my linear regression model including this variable as one of the predictors, I have a high expectation that this would turn out to be significant in my model.

>>>

Let us analyse the other variable which is the gender.

>>>

I can see there is some difference between the female gender and male gender in terms of how much they spend and interestingly, the male

© Jigsaw Academy Education Pvt Ltd

category appears to be spending more as compared to the female category.

>>>

Let us analyse the next variable which is the variable which talks about if the person has his own house or not.

>>>

I see a higher expenditure for the group having their own house as compared to those who rent their houses, and it also makes sense.

>>>

Let us see, what is the behaviour of people in terms of their marital status?

>>>

I see people who are married, they spend on an average more as compared to people who are single.

>>>

Let us take a look at the variable location.

>>>

Now I can see that the people who live far away from Brick and Mortar stores that sell the similar kind of products as I do or this direct marketer

© Jigsaw Academy Education Pvt Ltd

does, they have higher expenditure as compared to people who live nearby the stores that sell similar products and it also makes sense.

>>

Let us take a look at the variable salary. Now since salary is a continuous variable so what I'll do is, I'll produce this scatter plot between salary and the amount spent.

>>

Now I can see a clear trend, as the salary increases the amount spent also increases. But you can notice that the variation in the amount spent is also increasing with the increase in levels of salary. So if I include this variable, there are high chances that my model would suffer from heteroscedasticity.

>>>

Let us take a look at the variable children. Now I can see that people who have two to three children, they behave similarly as compared to the other groups. So what I would like to do is to combine the group of people who have two to three children and treat them as one group and then take a look at the three groups that I would have, one group would be the people who have zero children, people who have one child and compare that with people who have two to three children and see if I find a pattern.

© Jigsaw Academy Education Pvt Ltd

>>>

So what I've done now is I have collapsed the total number of categories from four to three. So I'm calling the category of people who have two to three children as just a single category. Let us take look at the box plot with these new categories created.

>>>

I see that there are some difference between people who have no children and people who have children, and there is also some difference between or rather there is also a significant difference between people who have two to three children as compared to the other groups. So what I'll do is I'll just keep three groups as far as this variable is concerned and then run on the model on these three groups.

>>>

Let us talk about the variable which is the purchase history, and if I run a summary on this I see that I have some missing values in this particular variable. So I might as well impute these missing values first. So what I will do is, I will figure out what is the average amount spent for people who have high volume purchases, low volume purchases, medium volume purchases and I will also find out what is the average amount spent by people about whom I don't have any information in terms of their purchase history. So let us find the group means first and then find the

© Jigsaw Academy Education Pvt Ltd

mean amount spent by people about whom I don't have any purchase history. So, I can see that the mean amount spent by people belonging to the missing value category is twelve hundred and the mean amount spent by all the other categories is here. Now this mean amount matches with the medium group, but still 1239 is very different from 950, and I might as well not impute the missing values the same as the medium class.

So what I'll do is, I'll create a new category in my data and call that category as the missing value. So this is what I'm doing here. Let us take a look at the summary of this new variable that I have created, so now I have four categories in my data. So this category which was earlier labelled as NA, I have relabelled it as missing, so that I can run my linear regression model. Let us take a look at the variable catalogues. Now ideally the number of catalogues that I send to people should be a categorical variable. So, I should talk about the group of people to whom I have sent six catalogues, eight catalogues, ten catalogues, but for me to treat this variable as a categorical variable what is important, is the range of the number of catalogues that I've sent should not be too high. So let us take a look at the summary of this variable. I can see that the range is from six to twenty-four. So I have eighteen distinct groups for this variable, eighteen distinct possible groups for this variable which might become too much so I will not treat this variable as a categorical variable and I will treat this variable just as a continuous variable.

© Jigsaw Academy Education Pvt Ltd

So now what I'm doing is, I'm taking out the columns from my data from which I had derived other columns and creating a new dataset with fewer columns, and using this dataset to build a linear regression model. Now, the way we build a linear regression model in R is using this LM command; LM stands for linear model. The first argument is the dependent variable. Since I want to predict the amount spent, I'm writing that this is my dependent variable, that I put the Tilde sign and put a dot there. Now Tilde and dot is a short hand for saying that I want to build a linear regression model keeping amount spent as the dependent variable and all other variables in my dataset as the independent variables.

So I'm starting my model building process by taking into consideration all the other variable and I'm seeing how my amount spent is being predicted if I build a model taking all the independent variables into consideration. Let us take a look at the summary of the model object that I've just created. So this is how the summary for a linear regression model object looks like. This first statement is about the formula that I have used to build the model. So I'm saying that amount spent is a function of the entire predicted variable in my data. This here, talks about the five point summary of residuals. Now residuals are the error terms in my model. Then what I see is I see the coefficient estimate for each of the variables, their standard errors, their T values and the corresponding T values. Now I see that some of the variables are not significant.

© Jigsaw Academy Education Pvt Ltd

How do I know that? I can take a look at the magnitude of P value. The P value corresponding to the variable gender is .012, which is a lot larger than 5% or .05. So I can see that some variables are significant, some variables are not significant. If I come down, I see that my adjusted R^2 is 74.21% which means that this model is explaining 74.21% variation in my sample data. I can also see a P value here. Now this P value corresponds to the an nova for regression. So what it means is, that my current model with all the variables as predicted is performing significantly better than a model in which I include no predictors at all. It implies that whatever model building I have done, it can be generalised to the population. But keep in mind there are some variables which are not significant so I might need to get rid of them. So let us get rid of the variables that are not turning out to be significant and build the second model, and let us take a look at the summary. Now I can see still, the variable gender is not turning out to be significant and the variable history with the level missing is not turning out to be significant. Now one thing you would've noticed is the variable gender.

If I go back to my data, is a categorical variable with two levels and the variable history1 which I had created has four levels, low, medium, high and missing. Now if I take a look at the output of the linear regression model I just see one level for the variable gender and three levels for the variables history1. The reason for that is since these are categorical variables so they behave or they are included in the model as dummy

© Jigsaw Academy Education Pvt Ltd

variables, and whenever we include dummy variables we include one less level. So if I had two levels here, the model is automatically including one level. If I had four levels for the history variable, the model is automatically including one less level and is omitting the level high.

Now in order for me to remove this level missing, I would need to create dummy variables for all the levels and build my regression model only using the dummy variables. So I'm creating dummy variables for the variable gender and for each of the levels, male and female, and also for the variable history with all the levels missing, low, and high.

So, let us take a look at the dataset in which I have created these dummy variables and let us take a look at how these dummy variables appear in my dataset. Now I created dummy's for each of the levels of gender variable, male and female, so wherever I have observations corresponding to males I'm putting a one, wherever I have no observations corresponding to male I'm putting zero, similarly for females. Also for the variable history, I had four levels, missing, low, medium and high, and for each of these levels I'm doing the same procedure. Wherever I had an observation which is missing I'm putting a one and wherever I'm having an observation, if someone has had a high expenditure I'm putting a one and so on and so forth.

So let us now run a model by including the dummy variables. So I will now build a model in which I

© Jigsaw Academy Education Pvt Ltd

have a male dummy and I have a dummy for medium and low purchase histories. Let us take a look at the summary. Now I can see that the male dummy is not turning out to be significant, so I would now take this out. I can see that the medium dummy and the low dummy's are significant, so I will retain them. Let us take a look at the summary. Now I can see that all the variables are significant.

Now does that mean that we are done with the model building process? Definitely not, what we need to now do is look at the signs of the variables. Now I see that location far has a positive sign. Now it makes sense because the people who are living far away from a Brick and Mortar store that sells the same products as I do would make more purchases from me. Similarly I see a positive sign for the salary variable, makes sense. As the salary of people increases, the tendency to spend also increases and their tendency is also reflected in the fact that people with higher salaries purchase more from me and the people to whom I send more catalogues, they buy more from me because that is the reason that I'm sending them more catalogues.

Now I can see that for the variable children, for the level one I have a negative sign and for level three and two I have a negative sign. Now, if I think about it, children variable is a categorical variable and it has three levels in it. A level corresponding to the group with zero children, a level corresponding to the group with one child and a level corresponding to the group with three

© Jigsaw Academy Education Pvt Ltd

to two children. Now the model has automatically omitted the level zero, because this is a categorical variable meaning that it should be interpreted as a dummy variable, so one level should be left out. Now compared to the group of people who have zero children, I can see that group of people who have one child, they have lesser mean and people who have three to two children have a lesser mean. So these signs are in line.

Also, I noticed that the signs for the dummy variable talking about the medium purchase history and the low purchase history are also negative. What I need to do is, I need to figure out what is the mean of the group of people who do not belong to the medium category and the low category. So this is what I'm doing here. So I see that the mean of people who do not belong to the medium category and the low category is 1672. Let us take a look at the mean amount spent by people who belong to the medium purchase category and low purchase category. I can see that the medium purchase category mean is 950 and the low purchase category mean is 357, which is less than the mean of all the other groups. So, that is why the negative signs here make sense, because when I include dummies for the medium category and the low category, I'm excluding all the other groups and I'm making a comparison with respect to the groups that I've left, and the groups that I've left have a higher mean so that is why a negative sign here makes sense.

© Jigsaw Academy Education Pvt Ltd

So now I know that all my coefficients have a proper sign and also have significant P values. What I would do next is, I would check the linear regression model assumptions. Now the first assumption I'll check is the assumption of normality. So what I'll do is, I'll plot the histogram of residuals.

>>>

If I take a look at the histogram of residuals, I see that on the right tail I have some values which are not looking to be symmetric to what I have on the left tail. In the sense, what I see here are the signs of skewness. My right tail is elongated as compared to the left tail.

>>>

I can verify this by drawing a quantile, quantile plot.

>>>

So, if I take a look at this quantile, quantile plot, this 45 degree line here talks about the behaviour of data had it been normal and then I have these black dots which talk about the behaviour of data which I have, and the data that I am talking about is the residuals. So I can see that there is significant departure from normality at the higher quantiles for my residuals. So my data is not normal or my residuals are not normal.

© Jigsaw Academy Education Pvt Ltd

>>>

Next what I need to do is, I need to check if there is any form of multicollinearity. So I will use this function vif, and I can see that the vif is comfortably less than 10 for all the coefficients. So, I don't have any issue of multicollinearity but although I do have an issue of normality. Now next let us check the assumption of constant variation. So what I'll do is, I'll plot the fitted values with respect to the variables and I'll take a look at this scatter plot.

>>>

Now in this is scatter plot I see that there is a funnel pattern. As the magnitude of fitted values increases, the variation in the residual also increases. So my data suffers from heteroscedasticity.

>>>

So this model does not satisfy the assumptions of classical linear regression model. Now, in order to make the residuals normal and in order to get rid of heteroscedasticity, what I can do is I can apply transformations to my data. So the first transformation I'll apply is the log transformation and I'll take the logarithm of my dependent variable and keep all the other independent variables as it is. Let us build this model, take a look at the summary.

© Jigsaw Academy Education Pvt Ltd

Now I can see that the signs are in line and also all the coefficients are significant, so what I'll do now is I'll check the assumptions. Let us first of all check the normality assumptions using a qqplot.

>>

Now I can see there is some departure from normality but it looks okay. Earlier the departure was very much visible.

>>

Let us take a look at the fitted versus residuals graph to see if there is any heteroscedasticity in my data.

>>

Now, this looks okay but there is still a little funnelling pattern here but overall it looks okay. The variance looks fine across the values, although at the higher ends I see some funnelling pattern.

>>

What I'll do now is I'll try another transform on my data, so I will take the square root of my dependent variable and build a model. Let us take a look at the summary. Now I can again observe that the signs are in line as well as the coefficients have significant values. So I will take a look at the quantile, quantile plot for this particular model with square root transformation.

© Jigsaw Academy Education Pvt Ltd

>>

I see there is some departure from normality here and the departure of normality for this model and the previous model is almost similar. They depart a little bit from normality that should not be a lot of concern for me.

>>

Let us take a look at the fitted versus residuals plot.

>>

And here I see some funnelling pattern. There is definitely a funnelling pattern at the lower ends of my data. And as I move towards the higher magnitude of the fitted values, the funnel expands.

>>

So it seems like I can't get rid of heteroscedasticity. So, what I'll do is I'll finalise this model. Before finalising this model, I'll take a look at the multicollinearity and vif values. So, all the vif values are comfortably less than ten. Model seems to be okay in terms of multicollinearity, it seems to be okay in terms of the normality of residuals. There is some heteroscedasticity present in my model. I can try other transforms on my model, for example, I can try a cube root transformation or I can try a natural logarithm

© Jigsaw Academy Education Pvt Ltd

transformation and see if something changes in order to get rid of heteroscedasticity.

The last thing I will do is, I will find the predicted values and I'll find the actual values and I'll make a plot of predicted versus actual values and see if there is a good FIT. So this graph of predicted versus actual values suggests to me that the predicted values, they follow the actual values closely.

>>>

In this module we'll take a look at how we build a linear regression model in R. We will emphasise upon how we explore and prepare data for linear regression modelling task and will see how we build linear regression models and also how we do the model assumption checking.

>>>

© Jigsaw Academy Education Pvt Ltd