HR Metrics & Analytics
MSB21T2001

Index-Session 24

session Objectives- Topics to be discussed

1. Types of Regression
2. Regression Modelling

Recap of Session 26

- Regression

I am going to talk about different types of Regression Models. Regression can be classified broadly into two categories. One is Simple Regression, and the other is Multiple Regression. In Simple Regression, we'll have one independent variable, whereas in Multiple Regression, we'll have more than one independent variable. Simple Regression can be further classified as Linear Simple Regression or Non-Linear Regression. Similarly, Multiple Regression can be classified into either Linear, or Non-Linear. Let us try to understand this a bit more. In case of simple regression, I may have a function which is $Y = \beta 0 + \beta 1x1 + \varepsilon$. In this case, there is only one explanatory variable which is x1. So, that's why this is simple linear regression. Below that, we have multiple linear regression. So here, $Y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta kxk + \varepsilon$. So, there are K explanatory variables and the functional form is linear, so that's why this is multiple linear regression. Now, below that we have a non-linear function which says $Y = \beta 0 + 1/ (\beta 1 + \beta 2x1) + x2\beta 3 + \varepsilon$. So, here, we have many nonlinear term. Now, we have to understand that whenever we talk about non-linear regression, the function is non-linear with respect to the Beta coefficients or the regression coefficients, not with respect to the explanatory variables or independent variables.

what is a problem when we have a non-linear regression

The problem comes when we try to estimate the regression parameters. So, typically, what happens is, I'm going to talk about this later when we come to ordinary least squares method. So, when you have a non-linear regression, when we try to estimate, we'll have system of non-linear equation. And, they are difficult to solve. So, typically, we may not have any close form solution. So, in most cases, we may use some kind of numerical approximations to estimate the Beta coefficients. Okay, let us try to understand the difference between linear and non-linear regression. If you look at these two equations in this slide, the first one is $Y = \beta0 + \beta1x1 + \beta2x2 + ... + \beta kxk + \varepsilon$. In the second case, we have $\beta0 + \beta1x1 + \beta2x2 + \beta3x1x2 + \beta4x2\ 2$. Now, even though you have non-linear term, in terms of explanatory variables or independent variables, there is no non-linear term with respect to the Beta coefficients or regression coefficient. So, this also will be considered as linear regression, as far as regression models are concerned. So, we have to understand the difference when we call something as non-linear regression and when we call a model as a linear regression.

I'm going to now talk about regression model development.
the steps involved in regression model development. The first thing is, to basically, to explore the data. Because, the dataset may have quantitative variable as well as qualitative variable. So, we need to understand what kind of data we have. So, once we have done with that step, we go on do a descriptive statistics. As part of descriptive statistics, we may come up with some data visuals, things like scatter plot. The next step is pre-processing the data. Because there may be some missing information, so we may have to use some kind of data interpretation. Or, if we have qualitative data, Predective Analytics Prof. Dinesh Kumar Regression Types DK_PA_W02_C03 then we may have to use dummy variable because we cannot use qualitative data directly. So, once the pre-processing is done, then we go and define the functional form of relationship. The functional form can be like Y equal to β0 plus β1x1, or it can be log natural of Y equal to β1 plus β1x1.

The idea for the functional form may come from the scatter plot that we do in the second stage. So, once we have the functional form, we go and estimate the regression parameters. So, in case of linear regression, the regression parameters are estimated using ordinarily squares. So, once we have the regression parameters, then we perform diagnostic test because while developing a regression model, we use lot of assumptions. So, we have to actually check whether those assumptions have been met. So, if they are met, then we go and use the model, otherwise we have to go and preprocess the data because we may not have correctly pre-processed the data or the functional form we have used may not be appropriate. And, we repeat this process until we develop a model that satisfies all the diagnostic tests, or we can conclude that there is no relationship between Y and the set of explanatory variables that we have considered.

why can't diagnostic test be carried before parameter estimation?

The reason is that most of the diagnostic test uses the estimated value of Beta. So, since, they need the estimated value of Beta, we cannot use diagnostic test before we estimate the parameters. So, that's the reason why we do the diagnostic test once we have estimated the regression parameters.

In Simple Linear Regression, we have one explanatory variable. So, first thing is to check what is explanatory variable that we are using. For example, in case of die-another-day hospital, my explanatory variable could be the weight of the patient, right? And then, I had to specify the nature of the relationship between the explanatory variable and the response variable. So, in case of dieanother-day hospital, that is a total cost of treatment. So, my Y is total cost of treatment and my X is weight. So, I'll write the functional form Y equal to β0 plus β1X1. So, this is basically what we assume. We assume that there is a relationship in the population, and there is a population parameter, β0 is the intercept, and β1 is the slope of that function. And, this is what we assume in the population. So, what we are trying to do is we have a sample data, and we try to estimate β0 and β1 using that sample data. Now, before we go and do that, we make large number of assumptions.

And, I'm going to talk about only the most important assumptions here. The first assumption is that the error follows normal distribution or the residuals follow normal distribution. The second important assumption is that for different values of X, so, if I have to speak about the DID case, for different values of body weight, the variance of εi or the residuals is constant. So, this is called homoscedasticity. If the variance is not constant, then it is called heteroscedasticity. We also assume that there is no multi-collinearity. This is valid only when we have multiple linear regression. This basically assumes that when we have more than one explanatory variable, there is no high correlation between those explanatory variables. And, the last important assumption is that there is no autocorrelation between two ε values. So, this is valid when were are talking about a time series data.

what will happen if the error does not follow normal distribution?

Now, for parameter estimate itself we don't need that assumption. But, if it is not a normal distribution, some of those hypotheses tests which we will be doing as part of diagnostics may not be valid. For example, to check whether the Beta is significant or not, or the regression coefficient is significant or not, I'll do a T-test. So, if my error is not a normal distribution, then the statistic I derive may not be a T-distribution. So, my diagnostic test or hypotheses test is not valid.

what would be the impact of multi-collinearity?

Multi-collinearity is valid when we have more than one explanatory variable. Now, what it does is, it destabilizes the model. We cannot trust the Beta coefficients, and also we may remove one significant variable from the model itself. So, it has lot of impact, and we are going to discuss this in detail when we come to Multiple Linear Regression.

In linear regression, the estimation of parameter is done using a technique called Ordinarily Square. So, if you look at regression model building, we start with a population and belief about that population. So, we assume that there is relationship that exists. So, if I go back to DAD case, we assume that the treatment cost is related to the body weight. But, we don't have the entire population. But, we have a sample. So, now, using this sample I have to estimate the population parameters which are β0 and β1. Now, the question is, "What kind of logic I can use?" So, if you look at β0 and β1Xi, we can prove that this is nothing but the expected value of Y for a given X. So, the error is difference between the estimated line and the actual value. So, since E (Y|X) = ⍰ ^ 0 + ⍰ ^ 0X1, if you add the error which is deviation from the mean, they will all add up to zero. So, we cannot say that we'll just look all the errors and then try to minimize. So, to overcome this, what we do is we take square of those errors. So, here, I've written that ε1 2 , ε2 2 , and all that. So, for the entire observation I have the line, I calculate the error from that line, and square that and try to minimize that. So, that's called least squares method

what will happen if we use absolute error values?

Why should I use the square of error? Because, if the problem is that, if I add just the error then it'll add up to zero. Why can't I use the absolute value of error? Now, I'm going to talk about this later. The primary reason is that we can mathematically prove that if we use ordinarily square, we get unbiased estimate. So, I'm going to talk about that later. So, when we try to estimate, we are looking at sum of squared errors. So, which is the εi 2 for the entire sample data I have. So, that's given in the first equation. And, this is a classical optimization now. So, what we do is we take the partial derivative with respect to those two parameters. First with β0 and with β1. So, I have those equations below that. And, I equate that to zero and solve those equation. So, basically, what I have is system of two linear equations. So, when I solve that, I get the estimate of β1 as ⍰ ^ 1=∑(∑(Xi − X)(Yi − Y) ∑(Xi − X) 2) So, that’s the estimate of β1. And, now, I can substitute this in the first equation. I'll get β0 estimate. So, the β0 = Y _ - ⍰ ^ 1 X _ . So, I have, now, estimate of β0 and β1. And, if it satisfies all the diagnostic test, I can go and use this regression model.

REFERENCES

- Bhattacharya ,Deepak (2019). HR Metrics and Analytics, Pearson
- Predictive Analysis course from SWAYAM
- Batt, R. (2002). Managing customer services: Human resource practices, quit rates, and sales growth. Academy of Management Journal, 45(3), 587–597.
- Becker, B. E., & Gerhart, B. (1996). Human resources and organizational performance: Progress and prospects. Academy of Management Journal [Special Issue: Human Resources and Organizational Performance], 39(4): 779–801.
- Becker, Gary, S. (1962). Investment in human capital: A theoretical analysis. Journal of Political Economy, 70(2), 437–448.
- (1964). Human capital. New York: National Bureau for Economic Research.
- Beer, M., Spector, R., Lawrence, P., Quinn Mills, D., & Walton, R. (1984). Human resource management: A general managers perspective. Glencoe, IL: Free Press.
- Begin, J.P. (1993). Identifying patterns in HRM systems: Lessons learned from organizational theory. In J. Shaw, P. Kirkbride and K. Rowlands (Eds.), Research in Personnel and Human Resource Management, 3 (pp. 3–20), Greenwich, CT: JAI Press.
- Bhattacharyya, D. K. (2012). Human resource management (3rd ed.). New Delhi: Excel Books

In case, you find any difficulty in understanding the concepts of lecture, please feel free to contact.

Dr Mamta Gaur
9891121396
mamtagaur@galgotiasuniversity.edu.in

Thanks