Introduction to Data Analytics

Prof. Nandan Sudarsanam and
Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

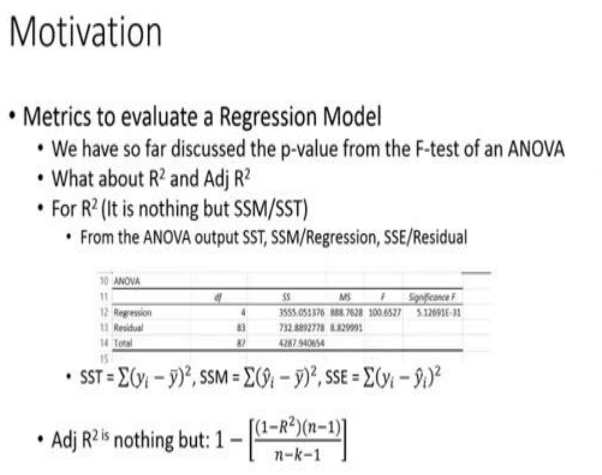
Module - 04
Lecture - 21
Adj R^2 and Regularization/Coefficient Shrinkage

Hello today we will continue our lecture series in the topic of regression and in our last class, we primarily focused on, mainly ended, fine with the ideas behind subset selection. So, essentially we were talking about this problem, where you have a lot of input variables in, that you could use in your multiple regression model, how do you go about picking, which one should stay in the model and which one should go out. And in discussing that we would discussing a more broader topic which is, how do we measure how good a model is and how do we do things in our regression practice, so that we get good predictive accuracy and it is also useful for interpretation.

In that regard, we will continue our lecture today and talk about two other small measures, which is the R^2 and the adjusted R^2. The R^2 has at least been introduced to you, you know you have an idea of where you can find that in your analysis.

We will talk a little bit about that and we will talk about them more as other matrix of measuring, how good our regression is and then, we will move on to this topic called regularization, which has to do with how do you fine tune the datas, the coefficients of your regression model.

(Refer Slide Time: 01:34)



Motivation

- Metrics to evaluate a Regression Model
  - We have so far discussed the p-value from the F-test of an ANOVA
  - What about R^2 and Adj R^2
  - For R^2 (It is nothing but SSM/SST)
    - From the ANOVA output SST, SSM/Regression, SSE/Residual

10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	4	3555.051376	888.7628	100.6527	5.12691E-31
13	Residual	83	732.8892778	8.829991		
14	Total	87	4287.940654			
15						

    - $\text{SST} = \sum(y_i - \bar{y})^2$, $\text{SSM} = \sum(\hat{y}_i - \bar{y})^2$, $\text{SSE} = \sum(y_i - \hat{y}_i)^2$
  - Adj R^2 is nothing but: $1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$

When we were discussing the topic of subset selection, which is, how do you go about choosing which variable should stay in the model. We said, we would go about making those choices based off of some metric, not of the individual variables, but of the model as a whole, correct and in that context, we primarily discuss the use of the p value from the f test that you do with your ANOVA. So, for instance I have copy pasted the same results from our examples that we took in our last class and so, this stable out here kind of…

It is just a copy paste from that previous excel sheet and we were focusing on this p value primarily saying, this ANOVA is essentially a measure of how good the overall regression is, not individual input variables, but can we use that, can we use this p value as a guide to see which variable stay in and which variable stay out and that is the context in which we discuss the topic of best subset regression, we spoke about backwards, forwards and hybrid stepwise regression.

But, today what we are going to talk about is, we talk about two other measures which are also available in which is, something that you could use and they primarily the R^2 and adjusted R^2. So, the first thing that we are going to start by saying is the R^2 is not a very good measure, in terms of how the overall regression model is, especially in a multiple regression context. I will explain that in a second, what we mean by…

Again just as a reminder, you know simple regression just means there is one input variable, multiple regression means there are multiple input variable. So, the R^2 essentially measures, is a measure of how much of your variation and here, you are talking about variations in terms of y, your output variable. So, how much if your output variables variation, the sum amount of variation and how much of that can I explain using my model and how much of that can I not explain using my model.

So, if I take the overall variation of y and I break it up into two chunks, the amount that can be explained by my model, which is my regression equation and the amount that cannot be explained by that. If I put those two together I should get the overall variation in y irrespective of x. So, R^2 is nothing, but this ratio how much variation is explained by the model divided by the total variation. So, depending on the software that you use and the text book that you use, people usually have sum of squares total and that is the total variation and sum of squares model that is the model variation.

But, I have also seen sum of squares regression as a proxy for sum of squares model and the sum of squares… If you take, if you say the total variation is sum of squares total and the model variation is either sum of squares model or the sum of squares regression, then the only the other quantity that is left is the sum of squares error or you know, as some people again call it the residual sum of squares. You might have seen it RSS or sum of squares error.

So, this, your R^2 value therefore, is nothing… You can just basically look at your ANOVA output, even if your R^2 for some reason disappears. It is nothing but, the ratio of this number to this number, in which case I think that looks like it is about 0.8, 0.9 invalid. To give you a little more mathematical intuition and also kind of giving you some kind of formulas that are used for calculating these three values. As I mentioned, the sum of squares total is just essentially the variability, in some sense the total squared distance of each data point from the grand mean, whereas sum of squares model is where you go to each, not each data point, but here y_i is each data point.

We look at the deviation of each data point, actual data point from the grand average. With sum of squares regression or sum of squares model, you look at each predicted value $\hat{y}_i$. It is nothing but, what value of y are you predicting for each x and look at the deviation of that from the grand mean and finally, y_i- $\hat{y}_i$ is how much is each data point deviating from each predicted value. So, that should hopefully give you some intuition us to how these values are calculated.

So, now, going back to a bigger topic, as a metric to see how well a regression model is especially in a multiple regression, where there many variables, many input variables. This R^2 is not great, because R^2 by definition will only increase as you keep on adding more and more variables. So, we will be discussing this in great detail in our lectures on the bias variance dichotomy, which is going to come up, but the core idea just to kind of give you some feel for what we are talking about is that, if you keep on adding variables and if you keep on adding complexity, at some point you should be able to explain away all the data that you have or in another words, take the example where you have 10 data points and you had only 10 data points.

You should technically be able to fit a line or essentially fit a function that will go through all these 10 data points. If you just choose, you know if you say I am willing to fit a 9th order polynomial, essentially if the fitted model can get more and more and more complex, then for a finite set of data point you should be able to explain the way everything, but that is not necessarily great. Because, when you go and try and predict with that model, you are not going to do too well, you just do kind of over fitted to the data by just constantly increasing the complexity of the model that you are using.

Now, given that is the base and given that we were talking about this idea loosely when we in our previous lectures spoke about, how it is really important to try and figure out which of the subset of the variables you want in the model and you want to throw the others away. This R^2 does not help, because R^2 will never decrease as you keep on adding more input variable. So, let say I had a model with 5 input variables, I suddenly come up and say, maybe I should have the 6th variable and add it, it is almost it is definite that R^2 will increase and so, R^2 itself is not a great measure of how good a model is.

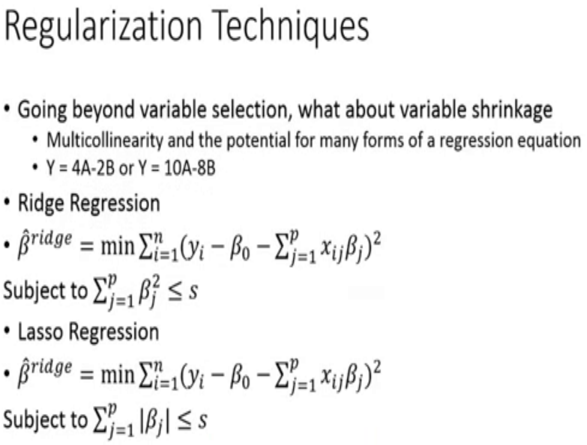
Now, as we discussed you could definitely use the p value from the f test of the NOVA, but another metric that is also popular is called the adjusted R^2 and that is essentially nothing but, a modified version of R^2 that is shown in the formula here. So, it uses essentially to compute it, you can use your R^2 values and the idea is that n is nothing but, the number of data points and k is the number of independent variables that are there.

So, for instance in our example, I believe we had 88 data points in the excel example that we were discussing yesterday. It is 88, I can infer that also from the total degrees of freedom and k essentially is 4, because we had four input variables. See, you can use k, the k = 4, so 88 and 4. So, the idea out here is that, when you use the adjusted R^2, it kind of penalizes modules, where the number of variables you are using to explain it is too high.

So, you can compare the adjusted R^2 of one model versus the other, where you used different numbers of input variables and yes, the higher your R^2 the better it is. So, if you can get a very higher R^2, you are going to get a, you know that is one way of getting a higher adjusted R^2, but not at the cost of, you know having too many variables. You want to keep k as small as possible and have a very high R^2.

Notice that, there is a - in front of the k, but this whole thing also has a 1 -, so just be careful when you are trying to get an intuition of, how increasing or decreasing these values is going to effect the R^2, adjusted R^2. So, that is more just, again introducing you to the concept that there is, it is not the p value of the f statistic as we discussed in the best subset selection and stepwise regression, but this is the concept of adjusted R^2 as well and that is just one of the them, there are other metrics also that can be used to measure regressions, a regression model.

(Refer Slide Time: 11:29)



Regularization Techniques

- Going beyond variable selection, what about variable shrinkage
  - Multicollinearity and the potential for many forms of a regression equation
  - Y = 4A-2B or Y = 10A-8B
- Ridge Regression
- $\hat{\beta}^{ridge} = \min \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$

Subject to $\sum_{j=1}^{p} \beta_j^2 \leq s$

- Lasso Regression
- $\hat{\beta}^{ridge} = \min \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$

Subject to $\sum_{j=1}^{p} |\beta_j| \leq s$

We now go on to the topic of regularization and you can also use the term coefficient shrinkage to express the same concept. What regularization does is also, you know as a huge over lap what you try to do with sub set selection, which is the core idea being how can I simplify my model in some sense. Because, I care about predictive accuracy, I care about interpretation, I do not care about just trying to get this function to go though all my data points.

So, how can I simplify the model in some way and one way of simplifying the model is to use fewer variables, that is what we saw in subset selection. But, once you fix the numbers of variables, once you say I have fixed these are the variables that need to go in the module, is there some way to more smoothly determine the coefficients. We saw how the coefficients were being determined through an optimization process, but somewhere can I go directly in there and put in my constraint of saying, I do want you to optimize something, but at the same time I want you to trying keep it as simple as possible without kind of over fitting the data.

So, in some sense the problem of regularization is a problem of saying I have an algorithm and I have a fixed number of input variables. So, I do not, I am no longer bargaining to throw variables out and keep them in, but is there some way of in my fitting methodology itself preventing this problem of over fitting or preventing this problem of trying to get the functional form to be so hardwired to the data, that it does not do such a good job of predicting and kind of, can I impose penalties upon myself to kind of over fitting.

One way in which over fitting happens is, in problems where the problems of what we call as multicollinearity. Multicollinearity is the idea that many of my input variables are correlated with each other and when you have that problem, the coefficients themselves through a regular regression process can get, the determining their βs can be kind of poor. Meaning that, there can be a lot of variance between one sample to the next.

So, it is like, ideally if I am doing a certain regression process and I take a sample of data and I then do this regression fit. If I take other sample of data and I get completely different set of βs, then that is not a very stable process and that tends to happen with least squares regression, when you have multicollinearity, meaning you have lots of highly correlated input variables. I mean take this example that we have on the slide, you have this idea, where you get this equation which says $y = 4A + 2B$.

Now, imagine a word in which A and B was so highly correlated to the point, where they were practically the same variable. Then, a fitted model that says $y = 10A - 8B$ should also a kind of give you the same results, in the sense that A and B are practically the same data points. They correlated to 1 and now assume that they are also equal in magnitude, the 4 A and 2 B, you can substitute B with A and you net getting 2 A, which is the same net you are getting in the second equations.

So, our both equations the same, now the idea with something like ridge regression is, you in cases like this, you want to have the simplest possible equation and you want to have the lowest possible magnitudes for your variables. So, you would be very happy with the $y = 2A$, I am just keeping it as simple as that or you can take it as $2A - 0B$. So, in addition to choosing which variables, you want to keep their magnitudes, the magnitude of the coefficients, you want to keep them as low as possible.

So, if you have off setting coefficients, you are not making a regression equation that says y is equal to 10,000 and 2 A and you know, you know really large say 1000 and 2 A - 1000 and B. Your kind of blowing things out of proposition, it will not generalize very well and so on and so forth. So, how can we achieve this? How can we achieve this goal of and really simple terms, not just allowing highly correlated variables to take up opposing sides and you know, have really large coefficients.

And the way we will do that is, by taking a standard least squares minimization problem and that is essentially what have written as the objective function here. So, the objective function here is no different for the ridge regression. At least the way have written it out here is no different than for least squares minimization, because all I am saying is let us minimize for each data point, the actual y - the predicted y and the predicted $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots$, goes on.

So, essentially j goes from 1 to p, meaning p is the total number of independent variables, n is the total number of data points, so you do it for each data point. You look at the deviation of the fitted value, the actual value to the estimate or the predicted value and the goal is, this deviation is what needs to be minimized. The square of this deviation is to be minimized in an ordinary least square regression, but in addition to that and what we do with the ridge regression is, we do this optimization with a constraint and the constraint can be written like this.

The constraint said it is the subject to some β, you take each β which is each coefficients, square it and the sum of those square should be less than sum value s and that value s is essentially, it is not like you have particular number in mind. What winds up happening is that you can rewrite this optimization by just not having this constraint, but essentially out here, just adding $– \lambda + \sum \beta_j^2$.

So, this is kind of like, you might have come across this with like Lagrange multipliers, when you have a single constraint on the coefficients. You can just rewrite the objective function to have that constraint integrated in to the thing and essentially, it is like solving just the minimization without the constraint. But, essentially the solution to that is what will ensure that your data's themselves are stable and you know, not taking really large values.

Another way to achieve the same thing is, again even with the lasso regression, this should not be ridge, this should be lasso. So, even with the lasso regression what you see is the same thing, you have the same objective function, but now it subjective to the constraint that the sum of the βs is less than some value s and again, out here you can just wind up rewriting your objective function. But, out here rewriting it does not do you too much good, because you cannot do the same trick as with standard calculus with Lagrange multipliers, where you have a beta square.

When you have a |β|, it just becomes computationally harder, but you know just like we saw in the simple regression case. If you got an excel sheet or a MATLAB, you can just do the optimization and put the sign as a constraint and as long as your optimization technique is good enough, you should be fine, you should be able to redo it. I hope that gives you some idea of regularization techniques and look forward to see you in the next class.

Thank you.