Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 06
Lecture - 30

Lecture 30: Regression methods and conclusions

Hello everyone, welcome back. Today is the last lecture in this module. This module is content based recommender system models and in this regard today we are going to talk about regression models and we will also conclude this particular module. So this is the content, so we will be talking about regression models followed by conclusions on recommender system. So now, coming to recommendation system models, so far we have covered four models decision tree, naive based algorithm, then rule based approach, then one more approach we have covered. So these four approaches we have seen that they work well with the response variable which is basically some kind of binary or in ordinal scale.

Regression Model	Nature of Rating (Target Variable)
Linear Regression	Real
Polynomial Regression	Real
Kernel Regression	Real
Binary Logistic Regression	Unary, Binary
Multiway Logistic regression	Categorical, Ordinal
Probit	Unary, Binary
Multiway Probit	Categorical, Ordinal
Ordered Probit	Ordinal, Interval-based

In fact, for all these systems that we have studied, all the methods that we have studied, the example that running example that we used was with binary variable. However, there can be situations in which we have the response variable which is a real number. So, in such setting linear regression is a very good choice. In fact, for large scale problems linear regression models is a very good choice.

So if you look at this table, we can see regression model and its variations fit with different kinds of rating variable. This linear regression, simple linear regression, multiple linear regression and kernel regression, they fit well with real type the target variable which is a real number. Then binary logistics regression is for unary and binary rating. And in case of unary rating, it is actually considered as a binary rating where the rating 0 is considered as something where the user has not given any input. Similarly, it can be extended to multi way regression and we have something called probit models for

unary and binary, then multi way probit for categorical and ordinal variable and order probits for ordinal and interval based variables.

The Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

The estimate

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

The error (Residual Sum Square)

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

So out of this, we will be choosing the first one because we have already seen various classification models which are fitting well with categorical variables. So therefore, we will be focusing only on linear regression. In fact, we have talked about linear regression on several occasions. While talking about the supervised learning algorithm, we saw one glimpse of linear regression model. Then we also saw about regression models during our discussion on user user and item item based collaborative filtering where we were trying to learn the weights using some kind of regression model.

So in the current setting, when we are discussing it in the context of we are discussing this in the content of a content based in the context of content based recommender system, we still will be able to use this regression model. This is a regression model of multiple linear regression. Now what is your data? Your data specifically in a content based recommender system where you have text documents. This text documents are represented in terms of TF or IDF rating. Now this TF IDF rating will be giving you certain numerical value.

So this numerical value represents the feature. So as per this model, this is your feature 1, this is feature 2, and this is your feature p. So you have p such features. So these features may be the TF IDF rating and if you are thinking of combining it, then it may be some kind of explicit features that we have seen in our last example. And we are supposed to determine this value of y which is our response variable.

Now in this particular model, we assume this response variable is numeric. So far all our variables were categorical. Now this is our model. What the model says that given the values of feature 1, feature 2, and feature k and these model parameters beta 0 to beta p, if we determine the value, we will be getting the value of y, but what happens? We may

not be able to get the exact value of y and there will be some error involved. So this epsilon basically tells you the residual error, the error term.

Ridge regression estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

shrinkage penalty
($l_2$penalty)

The lasso (least absolute shrinkage and selection operator)

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

($l_1$penalty)

So now when we solve this model and determine the estimated values of beta 0 to beta p, then based on this estimate, we find the estimate of y. And in our data set, we have total n number of observations. So if we have n number of observations, so how do we get these beta values? We get these beta values by solving this least square problem where sum of all these differences between the actual observation and the estimated value, we square this and take the sum. So this error is called residual sum square. So taking this one, we try fitting this function to some fitting to fit this to a linear function so that we determine these values.

And we know that how to determine these values. We know the ah gradient descent, stochastic gradient descent, etcetera, but anyway those things we have already discussed. So besides minimizing this difference, we also try adding certain penalty term and because of this penalty term, we avoid overfitting. So this we avoid overfitting, ok. So ah besides this, sometimes to understand which factors are relevant, which factors are not, we set up this with L1 penalty and as we have discussed in one of the earlier lectures, if we put this L1 penalty, what happens? Sum it enforces.

It not only helps avoiding overfitting, it also enforces some of the betas if they are not very relevant to be exactly 0. So which means this by this process, you will be able to understand in a much better manner which variables are actually irrelevant in your system. Now moving ahead, these things whatever I told so far in some setting like how to use regularizer, what is L2 regularizer, what is L1 regularizer, those things we had a bit elaborate discussion elsewhere in some of the earlier lecture. Now the point that we are going to make is given that we are fitting our model to this regression setting where we try fitting a linear function. How do we draw various inferences? So first inference we would like to draw is to understand how well does the model fit the data.

Second is is at least one of the predictors that is features is useful in predicting the response. So we would like to find out the contribution of this in determining the response variable. Then we identify whether all the predictors explain x, explain y that is the response or only a subset of the predictor. Predictor here means that feature. Now given a set of predictor values what response value should we predict and how accurate is our prediction.

So these four points we are going to discuss now. So first point is about assessing the accuracy of the regression model. Then we assess the accuracy of the regression model as the as we have already seen what we were trying to minimize. We were trying to minimize this residual sum of squares that was the value of y minus estimated value of y. So, our aim was to minimize this and while minimizing this to just the quality of this we can quality of the regression model we can use this arises in to in determining two related quantities.

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

So what are they? One is residual standard error and R square statistics. Now the regression model associates an error term with each observation. Is not it? Error term with each observation. This was the model; this was the error term. So it associates this epsilon which was the error term.

So which means with how many observations we had? We had total n number of observations. So with respect to n number of observations with each one we associate one epsilon value. Now the regression now what happens due to the presence of this error terms even if we know the true regression line we would not be able to perfectly predict predict x. Why so? Because this line is just the average line. Is not it? Suppose in case of a simple linear regression this is your x and this is your y.

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

So your points will be and you will be fitting the line which will be minimizing the error. But whatever you may say when you actually get a new value of x for which you are going to determine y that new value of x probably will be somewhere here and there will always be an associated error term. So as a result we would not be able to perfectly predict the value of y from x. Even if when we know this beta values because this beta

values basically define this line or a plane in case of a two variable regression or one hyper plane in case of a multiple variable regression. Now this RSC is an estimate of the standard deviations of the error term roughly speaking it is the average amount that the response will be deviate from the true.

Now here also we were finding the deviations. So what is the point in determining this? Now this term makes sense because it is a sample. So therefore, it is multiplied with 1 by n minus 2 where n is the total number of records. Where n is the total number of records. Now this RSC is considered as a measure of lack of field of a model.

So if the predictions obtained using the model are very close to the true outcome value of RSC will be small and we can conclude that the model fits the data very well. Now there is one more statistic using which we can also assess the accuracy of the regression model. So here RSC was supposed to be small when we have a good fit. In case of R square statistic, it has to be large in case of a good fit. Now RSC provides an absolute measure of lack of fit of the model to the data.

But since it is measured in terms of units of y because it is basically y minus y cap square root over. So it has the unit of y. So therefore, we need some kind of statistics which is not having any unit. Now this R square statistics provides one such alternative measure. What it does? It takes the form of a proportion so that that unit you miss.

The proportion of variance explained and so it always takes the value between 0 and 1 and is independent of the scale of y. So this is defined like this R square is TSS minus RSS divided by TSS. So where TSS is the total sum of squares. Now see when you found out this value of RSS what was RSS? RSS was sum over all n yi minus yi hat. Whereas when we talk about TSS it is y bar.

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

The total sum of squares $\text{TSS} = \sum (y_i - \bar{y})^2$

So what is this y bar? Remember your data set. In your data set you had the features like f1, f2 up to fp and your response variable was y and you had many values. Now when we consider this y of the training pattern this y values have some average and there will be some deviation of each of these each of these y's with respect to this y bar. So what does it indicate? It indicates that originally there is a lot of variation in the data.

So that is what it says. TSS measures the total variance in the response y and can be square and can be so total variance in response y and can be thought of as the amount of variability the amount of variability inherent in the response before the regression is

performed. So this variability was in the originally in the training data. So this variability was in the original original variability in the training data. Whereas RSS residual sum of squares measures the amount of variability that is left unexplained after performing the regression. So what was it? This was y hat and we had this RSS which is basically y minus y hat.

So RSS measures the amount of variability that is left unexplained after performing the regression task. Now what is remains unexplained? Because let us say this is our regression line and after the regression depending on the values of beta that you get this is the value which got predicted but this was the actual value. So this was the variability that was left unexplained. Now TSS minus RSS measures the amount of variability. TSS minus RSS measures the amount of variability in the response that is explained by performing the regression.

So this is TSS minus RSS this explains amount of variability in the response that is explained. Now what is R square? R square is what is R square? It is TSS minus RSS divided by TSS. So what was the original variability? Original variability minus what is getting explained. So R square measures the proportion of the variability in y that can be explained using x. Now an R square statistic that is close to 1 indicates that large proportion of variability in the response is explained by the regression.

So which means it is a good fit. Large proportion of the variability is now taken care of this regression line. Now the number near 0 indicates that a regression does not explain much of the variability in the response. So this might occur because the model fit is wrong or the error variance is high and many such reasons. For example, suppose there is a some kind of out layer in the data and because of that out layer the line regression line is getting distorted. So that may be a reason and because of the error variance high error variance also it may be the case.

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}$$

Now the relationship between response and predictor. How do we find whether the regression coefficients are 0 or not? So how many regression coefficients were there? Beta 0, beta 1 up to beta p. Now to understand the relationship between the response and the predictor. Predictor was y and what was the responses? Sorry the predictors were x 1 or f 1, f 2 up to let us say x p for the features f 1 to f p the values are x 1 to x p and y was the predicted value. So now under what condition you say that y has a relationship with x with sorry x has a relationship y.

This is beta into x 1, beta into beta 2 into x 2 and so on and beta p into x p. You take the sum of that so which means if any one of the beta is 0 correspond it indicates corresponding x value does not have any kind of contribution in y that is why beta 2 is being forced to be 0 or near 0. So therefore, in case of a multiple variable setting we need to check for all this or in general we have to test some hypothesis that all these values beta 0 equal to beta 1 up to equal to beta p all of them are 0. So we will be talking about hypothesis testing when we talk about the performance and all, but for the time being let us say this is our null hypothesis and there is typically there will be two hypothesis one is null hypothesis and the other one is alternate hypothesis.

So then our null hypothesis is all these are 0. Now if any one of this is not 0 how do we know? Do we have to individually check? No there is a another statistic called F-statistic. Using this F-statistic we can say when there is no relationship between response and predictors. So now this F-statistic is defined in this way this TSS, RSS etcetera you already know n is the number of records on which you are performing regression p is the number of attributes. So this way you have n number of records you have p number of attributes p attributes.

So you have p number of attributes. So when n is large so when there is no relationship between response and predictor one would expect F-statistic to take a value close to 1 otherwise F is always greater than 1. Now how large it depends on the data set when n is large and F-statistic that is just little larger than 1 might provide the evidence of for the evidence of the relationship to exist whereas, when the size of n is small it can be very large quantity as well. So next question we asked how to decide the important variables. So, we can decide on important variables by feature selection and many feature selection methods we have already studied, but with respect to regression particularly we have talked about stepwise regression. So stepwise regression we studied in a bit detail at the time of feature selection.

So this stepwise regression you can have forward selection and backward elimination where in case of forward selection you start with one feature and keep on adding otherwise if you have total p attributes and take with taking one attributes p taking two attributes p c 2 and so on total you may get 2 to the power p number of combinations. So therefore, to avoid this forward and backwards forward selection and backward elimination were adopted where you start with one feature and keep on adding so that some your performance improves. Now the performance the statistics various statistic which were used to just the quality of the model or the performance of the model we saw last time besides of course, RSS and RAC we have this Mallows CP Akai Kei information criterion, Bayesian information criterion and adjusted R square these things also we have already discussed. So all these values they actually give some kind of penalty in terms of number of attributes a new number of attributes to be determined on

which it is to be determined. So the next thing is we have to judge the uncertainty associated with the prediction.

Now when we know that we are actually fitting a line or a plane or an hyper plane. So which means if we determine the corresponding beta values we are and we determine that it is a good model is it sufficient no because the original model that we considered was having some error y was beta 0 plus beta 1 x 1 up to beta p x p plus some epsilon. So that epsilon is some error which will always be there even if you correctly determine with right optimization you determine the values of all these betas. So this is called irreducible error. So whenever this error is again going to come when you estimate when you get the estimated values of beta 0 beta 1 and beta p.

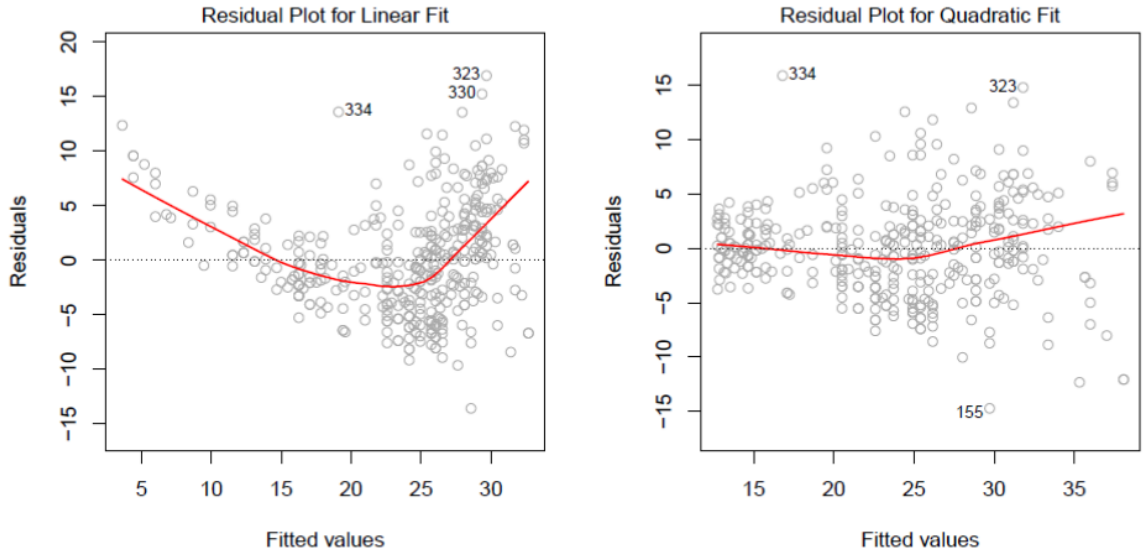
So when you get the estimates how do you get the estimates by solving that least square problem. So even if when you get the estimates and you get a good fit then also you cannot avoid this term this term is going to be there ok. So this part is called irreducible error or the model bias. Now when we use a linear model we are estimating the best linear approximation to the true relationship. Here we ignore this discrepancy and operate as if the linear model was correct.

So we assume that the linear model which we create using this estimated parameter is actually correct. So to understand how well we are taking care of this irreducible error or how well we are taking care of this estimation we ask some questions related to related to the prediction interval. This prediction interval typically will be little larger than the confidence interval. So these are few potential problems in regression modeling which we would like to see one after the other.

So first problem is non-linearity of the data. Remember through regression what are we trying to do? We are trying to fit a linear model, but it may so happen that the model the data is actually not giving a linear relationship. There are of course other measures like you can have non-linear version of the regression, you can have little more complicated network kind of thing like neural network and all, but right now focusing on only linear regression if we find that it is not a good fit how do we know? To know this, we have to look for the residual plot and find out if there is any pattern. For example, here what is the basic assumption of a linear regression model? These are the random errors. So this because this epsilon is a random value it has to be distributed without any pattern. However here on this if you try plotting you are able to get some pattern.

So which shows there is a lack of fit with the linear model and it is a some kind of non-linear model you must try. Still you can in a way you can think of reducing this effect by transforming y in certain manner. For example, in this second one this y is transformed if you have a log transformed y this is supposed to be no no no sorry sorry sorry. So in this

you have actually tried a quadratic fit. So if you try a quadratic fit then these errors are more or less becoming pattern free and more or less they are random.

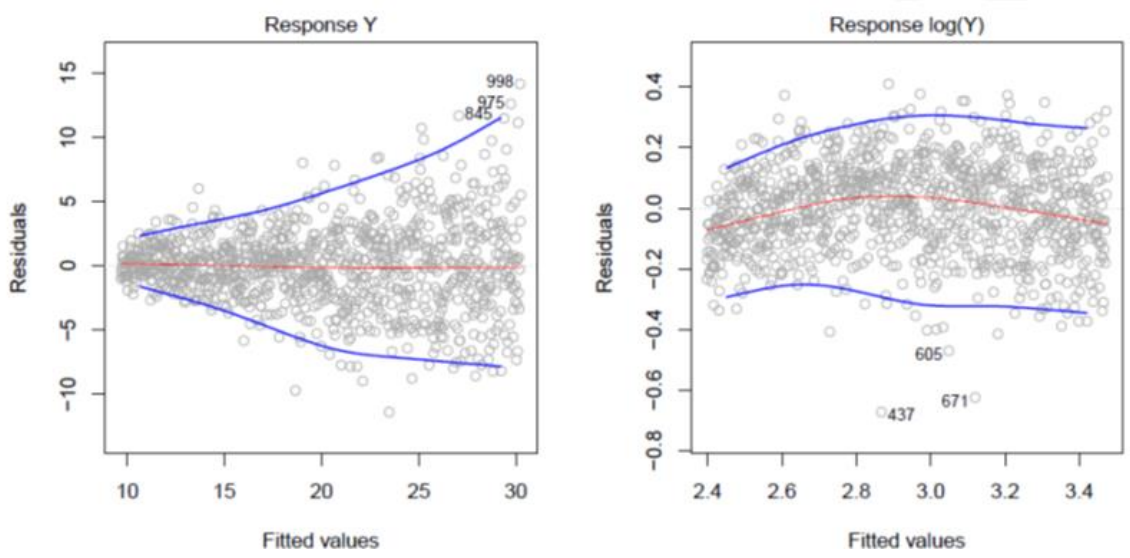


When you try fitting a data it is almost a line. There can be correlation of the error terms. Now what is the implication of the correlation? If the error terms are correlated this is this kind of thing often happen when you have the data when you have the time series data. So if the error terms are correlated we may have some kind of you know see we measure our confidence on the model. There are many ways we saw we can say whether it is a good fit or not. But with this correlation because two variables are contributing in the same manner unnecessarily this confidence will go up.

So because our confidence will go up we will be erroneously conclude that some of the parameters are statistically significant which may not be the case. So these correlations frequently occur in the case of time series data which consists of observations for which measurements are obtained at discrete point in time. In many cases these observations are actually these observations over the time are actually related so that is how you end up in such kind of pattern. So what do you do in this context? You can actually see the how the error is getting changed over the time.

So if you plot that you will be able to get some idea about this. With respect to time you plot your error term. Now there is one more assumption in linear regression model that error terms have constant variance they are random and their variance is constant. Now the standard error that is RAC that we determine as well as confidence interval and various hypothesis tests associated with the linear models generally rely on this assumption. Because every time you are taking y minus y bar or y minus y hat square so all these terms relates to the variance. So every time so if you have such kind of non-constant variance it is going to create problem in such calculations.

So as a result you will end up getting wrong inferences based on these values like standard error confidence etc. you will be getting wrong inferences. So this non-constant variance in error term if exist this problem is called the heteroscedasticity problem. So we can identify the heteroscedasticity problem in our data if we get some kind of funnel like pattern in the data set. Wherever however if we try making some transformation for example here we log transformed it the response variable and this is somewhat reduced.

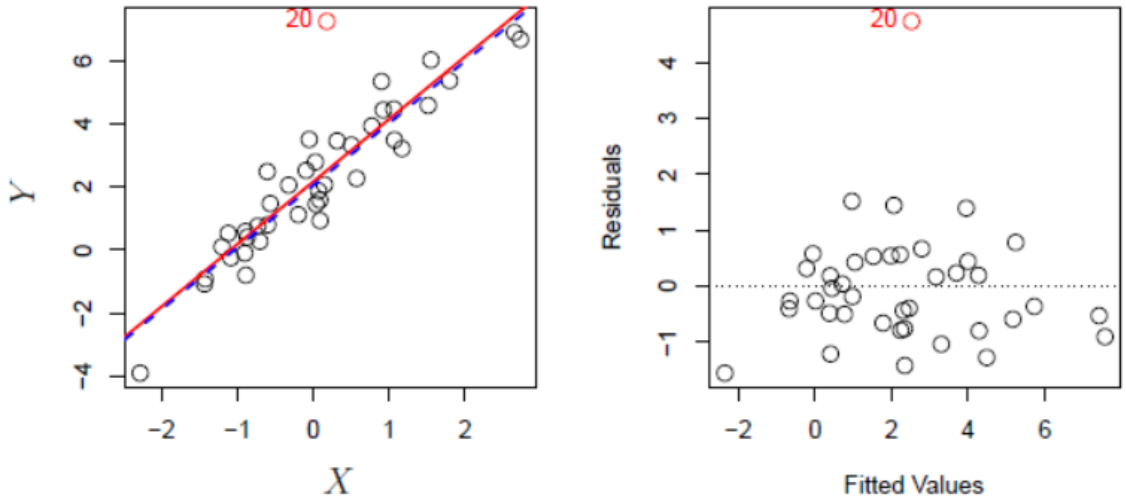


The next problem is that of outliers. Now existence of outlier may have little effect on least square feet. For example here in this diagram we can see this red line is with the outlier blue line is without the outlier, but there is hardly any difference, but this may not be the case all the time. You may get even a completely different feet which is not that of same as that of when you remove the outlier it is also possible. However, this outlier can also otherwise impact the standard error confidence interval etc. Now typically this outlier should be removed, but just like that blindly removing them is also not good.

So if you remove them how do you identify? So while discussing pre-processing steps we know that we can outrightly find out the outliers before even applying this regression we can find out the outlier. How do you do? There are you will look at box plots then in case of multivariate we look at some kind of you know Mahalanobis distance measures and so on. However, we can also detect it from the residual plots if we are trying with regression. This is the 20th point which is actual in outlier and you can also get it observed from the residual. So now the problem here the issue is here you have only one variable so you can visualize the outlier.

But in case of a multivariate setting residuals always are going to be one thing through which you will be able to determine the outlier. Next problem is your co-linearity. Now co-linearity refers to the situation in which two or more predictor variables are closely related to one another. The presence of co-linearity can pose problems in the regression

context since it can be difficult to separate out the individual effects of co-linear variables because they are basically indicate the same thing. So now the co-linearity reduces the accuracy of estimates of the regression coefficients it calls standard error to grow.



Now next question is how do you detect this co-linearity? This co-linearity first thing is that you can detect the co-linearity by looking at the correlation plots. In the correlation plots or the correlation matrix if some value is close to 1 or in a correlation plot if the you get a straight line then you know that there is co-linearity between two predict two feature feature vectors the predictor variables. Whereas this may not be the case always. So it may so happen that this may not be visible from this because this correlation matrix and the between two predictors that the scatter plot that you make.

So both these are between two variables. It may so happen that even if it the between two pair of variables there is not much correlation. Correlation as such with multiple variable can be high. So such and also give misleading results. So such problems are called multi co-linearity problems. So this multi co-linearity problems can be determined by computing something called variance in place and factor.

So for smallest possible value of v i f that is 1 it indicates complete absence of co-linearity. But higher values indicate how much co-linearity you have in your among your features. So with this we complete our topic on various types of models that can be used in recommender content based recommender system. But we may have to note that this content based recommender systems are actually some kind of supervised learning systems where you have a number of features and one dependent variable. So once your data set is ready by combining the item features and user preferences you can use any kind of of the self-classifiers.

So these such kind of methods particularly use very less engineering effort. And the kind of models that we have used starting from net base then we did rule based and so on

your decision trees and even regression. Regression this weights on the coefficients can tell you explain you which variable is important which is not. Looking at the regression coefficient you can decide. So all these are quite explainable methods and because of which you can provide some kind of explanation in a content based system. Look in case of in case of collaborative filtering also you can provide certain explanation for example if you are looking watching comedy movies then on that category only you are getting recommendation that can be written.

But in the feature level explanation will be missing because you are not considering feature level details in case of your collaborative filtering. Collaborative filtering relies only on the rating data. So therefore item level features for example in the running example that we are using what were the item level features. Item level features were director, the genre, then language and so on. So you may start getting suppose you keep on watching Telugu movies because it is one of the feature of the item it is not based on the similarity it is a feature of the item.

So you may keep getting recommendations on Telugu movies. So besides these two another advantage is it takes care of the cold start problem at the item level. In case of collaborative filtering we have cold start problem for both new user as well as new item in both cases sufficient ratings will not be available and we completely rely on the rating. But in case of content based system we are having item features. So when a new item comes you can at least compare it with other items, consider its similarity and push them into recommendations. And once they appear start appearing what is cold start problem? Cold start problem is you cannot make arrangement automatically pushing some value to the item recommendation process, pushing some item to the recommendation process.

But you will be do so by simply comparing the item level similarity. So item level cold start problem is taken care of. So there are many issues with content based recommender system as well. So they suffer from over specialization. So which means if you are watching comedy movies you will keep on watching comedy movies. So novelty which is about getting recommendations on newer items and serendipity which is about getting very surprising ones which is beyond your attributes preference of the attributes that cannot happen in case of content based system.

So some of the ways in which you can avoid it is by filtering out some of the very highly recommended item and going for the lower one. And you can also introduce this novelty and serendipity in an operational way at the time of programming. So by including certain random information generator then implementing something it relates to user profile, implementing poor similarity measures that will capture anomalies and exceptions and by reasoning you can introduce this. But anyway as we have already told it cannot really take care of the cold start problem for the new user.

So these are our references. So this regression model basically I have talked from this second book and details of which recommendation algorithm will go for which kind of data we got from here. And this is one reference which you can read to find out what are the recent trends in content based recommended system. So as we course becomes old probably this number you have to this year beyond this year you have to look and look for better references. So these are our conclusions. You can regression is a very good approach in the sense it can fit to different kinds of rating scenarios.

To judge the quality of fit we use R square statistics and standard error. We saw about various potential problems and how they are going to affect the results and we completed this module with a discussion on content based system and understood that there are certain advantages and disadvantages. With this we finish. Thank you.