

Machine Learning for Soil and Crop Management
Professor. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology, Kharagpur
Lecture 09
Basics of Multivariate Data Analytics (Contd.)

Welcome friends to this lecture 9 of NPTEL online certification courses. Course of Machine Learning For Soil and Crop Management. And this is week 2. And in this week we are discussing about basics of multivariable data, Multivariate Data Analytics. In my previous three lectures of this week, we have discussed different types of association between variables. We have discussed correlation. We have discussed regression.

We have defined, what is covariance. Then we have seen the simple linear regression. We have seen the assumption of simple linear regression. We have also seen what is multiple linear regression and what are the assumptions of multiple linear regression. We have also seen the diagnostic plots for linear regression. And also we have seen different types of data transformation, box-cox transformation, centering and scaling, we have seen.

So, in this lecture we are going to continue. And we are going to discuss some of the very important concepts as far as the multiple linear regression is concerned. Also, we are going to discuss some of the pitfalls of multiple linear regression, and how to detect those pitfalls. And how to, what are the remedies for those type of situations. So, let us start.

(Refer Slide Time: 02:02)



KEYWORDS

- Multicollinearity
- VIF
- Overfitting
- Bias
- Variance

So, these are the broad concepts which we are going to cover in this week. First of all, as I have told you, we are going to cover first the multiple linear regression pitfalls. Then we are going to talk about multicollinearity, which is one of the major pitfall of multiple linear regression.

And then, we are going to talk about the variance inflation factor or VIF and then we will be talking about the over fitting, and then finally we will be talking about the bias variance tradeoff, for reducing the overfitting. So, these are the keywords for this lecture, we are going to talk about multicollinearity VIF over fitting bias and variance.

(Refer Slide Time: 02:53)

MLR: RECALL

- More than one predictor...

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

- y_i = dependent variable
- x_i = explanatory variables
- β_0 = y-intercept (constant term)
- β_p = slope coefficients for each explanatory variable
- ϵ = the model's error term (also known as the residuals)

So, let us recall that multiple linear regression is a type of regression where we have more than one predictor. Here you can see our target is y_i , whereas our predictors are $X_{i1}, X_{i2},$

then up to $x_i p$. So, we have more than one predictor. Since, we have more than one predictor, we call it multiple linear regression.

Here, this y_i is the dependent variable and we can see that x_i is the explanatory variable, and then β_0 is the Y-intercept, which is a constant term, and β_p is the slope coefficient for each explanatory variable, starting from β_1 β_2 up to β_p and then ϵ is the model error term, which we also known as residuals.

(Refer Slide Time: 3:59)

MULTICOLLINEARITY

- Occurs when two variables that measure the same thing or similar things (e.g., weight and BMI) are both included in a multiple regression model; they will, in effect, cancel each other out and generally destroy the MLR model.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

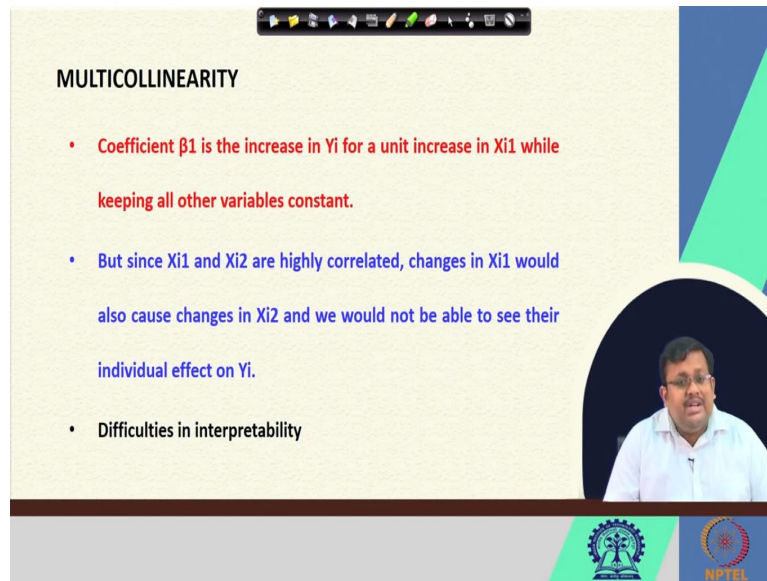
- y_i = dependent variable
- x_i = explanatory variables
- β_0 = y-intercept (constant term)
- β_p = slope coefficients for each explanatory variable
- ϵ = the model's error term (also known as the residuals)

So, what is multicollinearity? Multicollinearity is one of the major pitfall of multiple linear regression, and it occurs when two variables that measures the same thing or similar things, such as weight and basal metabolic rate or BMI are both included in a multiple regression model.

So, in any regression model, multiple regression model, if we include more than one predictor, which are correlated among themselves that creates the multi collinearity problem. Or in other words, when the multi collinearity happens that cancels each other out, and generally destroys the MLR model.

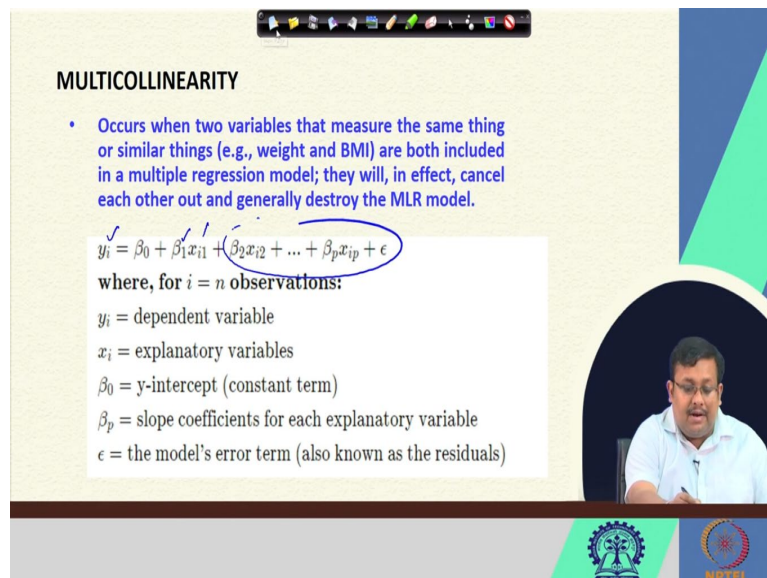


So, this is a one of the major problem for multiple linear regression. We have to select the variables very very carefully to be included in the multiple linear regression model.

(Refer Slide Time: 05:08)



MULTICOLLINEARITY

- Coefficient β_1 is the increase in Y_i for a unit increase in X_{i1} while keeping all other variables constant.
- But since X_{i1} and X_{i2} are highly correlated, changes in X_{i1} would also cause changes in X_{i2} and we would not be able to see their individual effect on Y_i .
- Difficulties in interpretability





MULTICOLLINEARITY

- Occurs when two variables that measure the same thing or similar things (e.g., weight and BMI) are both included in a multiple regression model; they will, in effect, cancel each other out and generally destroy the MLR model.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

- y_i = dependent variable
- x_i = explanatory variables
- β_0 = y-intercept (constant term)
- β_p = slope coefficients for each explanatory variable
- ϵ = the model's error term (also known as the residuals)

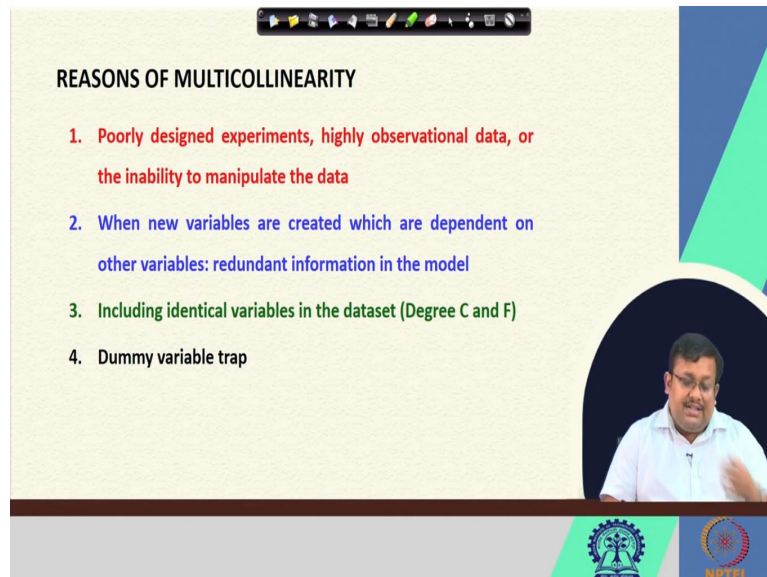


Now, what happens in multiple linear regression model? Multicollinearity. So, you know that coefficient beta 1 is the increase in y_i for a unit increase in X_{i1} ; we know that, while keeping all other variables constant. So, if we go back to this equation, we can see that this coefficient beta 1 denotes the increase in y_i for unit increase in X_{i1} , when all other are kept constant. However, when there is a multicollinearity, there are more than two variables, who are correlated among themselves.

So, suppose let us consider that since X_{i1} and X_{i2} are highly correlated. So, what happens when there is a change in X_{i1} that would also cause change in X_{i2} and we would not be able to see their individual effect on y_1 . So, that is the ill effect of multicollinearity, because, when we increase one variable that will impact the increase or decrease of another variable.

So, we will not be able to see their individual effect on the values of target parameter. So, what happens as a result of multicollinearity, there are always difficulties in interpretation. So, these are some of the important points of multicollinearity.

(Refer Slide Time: 06:48)



REASONS OF MULTICOLLINEARITY

1. Poorly designed experiments, highly observational data, or the inability to manipulate the data
2. When new variables are created which are dependent on other variables: redundant information in the model
3. Including identical variables in the dataset (Degree C and F)
4. Dummy variable trap

So, what are the reasons for multicollinearity? The most important reason for multicollinearity is the poorly designed experiment, with the highly observational data or the inability to manipulate the data. Sometime we design the experiment so poorly that we incorporate, whatever we can from different observations, without understanding whether they are correlated or not. And also sometime, we do not manipulate the data. We keep all the variables as it is, so that creates multicollinearity.

Also, when new variables are created, which are dependent on the other variables that can also include some of the redundant information in the model creating multicollinearity. And also, including identical variables in the data set can also create multi quantity. One example is, if we include two variables; one is measured in degree centigrade another measure in Fahrenheit, then obviously both of them will be highly correlated.

So, in that case that will incorporate the multicollinearity that will induce the multicollinearity. And also another reason is dummy variable trap. So, these are some of the major reasons for multicollinearity and multicollinearity, when it happens in any multiple linear regression that destroys the interpretation of the model.

(Refer Slide Time: 8:21)

DETECTION OF MULTICOLLINEARITY

- **VIF (Variance Inflation Factor)**
- VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable
- VIF score of an independent variable represents how well the variable is explained by other independent variables

$$\sqrt{VIF} = \frac{1}{\sqrt{1-R^2}}$$

- So, the closer the R^2 value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable

The slide also features a video inset of a man in a white shirt speaking, and logos for IIT Bombay and NPTEL at the bottom.

Now, how we can detect the multicollinearity. Multicollinearity can be detected by using the VIF or Variation Influence, Variance Inflation Factor. Now VIF, what is VIF? VIF determines the strength of the correlation between the independent variable and it is predicted by taking a variable and regressing it against every other variable. So, the VIF which is the variance inflation factor gives you the indication of whether certain variable are highly correlated or the certain variables are highly correlated to each other or not.

So, generally how we define VIF? VIF is calculated by taking a variable and regressing it against the every other variable. So, how we calculate the VIF score. So, VIF score of an individual variable represent, how well the variable is explained by other independent variables. So, if we make a regression of an independent variable with other independent variable then we calculate the R square and then we calculate their VIF score using this formula.

Now, so, VIF is equal to $1 / (1 - R^2)$. So, this is the formula VIF. So, closer the values, closer the R square value to 1, the higher the value of the VIF, and the higher the multi coordinate with the particular independent variable. So, that is why, this is how we calculate with the, we detect the multi coordinate using the VIF score.

(Refer Slide Time: 10:18)

VIF FEATURES

- VIF starts at 1 and has no upper limit
- VIF = 1, no correlation between the independent variable and the other variables
- VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others

No.	Variable	Collinearity statistics	
		Tolerance	VIF
1	X13	0.009	105.554
2	X1	0.017	60.497
3	X15	0.035	28.328
4	X19	0.056	17.775
5	X8	0.750	13.269
6	X6	0.137	7.295
7	X3	0.513	1.949
8	X4	0.721	1.387
9	X10	0.385	2.595
10	X11	0.260	3.851
11	X18	0.217	4.608
12	X21	0.517	1.935

Udomsin, W., Srungboonmee, K., Tantimongkolwaj, T., Naenna, T., Sritarin, A., Prachayasittikul, V., & Nantasenamat, C. (2014). Prediction of agricultural gross domestic product in Thailand using data mining. *Biomedical and Applied Technology Journal*, 2, 35-50.

So, let us see some example. So, VIF starts at 1 and it has no upper limit. And, when the VIF score is 1, that shows that there is no correlation between the independent variable and the other variables. So, when the variable VIF is 1 or near 1, then we are satisfied that there is no multicollinearity. However, when VIF exceeds 5 or 10 that indicates high multi quantity between the independent variables and the others.

So, here you can see one example is given. Here, the collinearity statistics are given; here you know number of variables are given X13, X1, X15, X19, X8, X6, 3, 4, 10, 11, 18 and 21. And, their VIF is given here. So, if we see their VIF, we can see that these X21, X18, X11, X10, X4, X3, they do not show any substantial multi coordinate.

However, when you go to X8, then X19, then X15, then X1, then X13, they are showing very high VIF score. So, that means they are highly correlated with each other. So, that implies that either we have to remove those samples to you know in the subsequent step for removing the multicollinearity effect in your model.

(Refer Slide Time: 11:58)

AVOIDING MULTICOLLINEARITY

- Dropping one of the correlated features will help in bringing down the multicollinearity between correlated features:

variables	VIF	variables	VIF
0 Gender	2.207155	0 Gender	1.863482
1 Age	13.706320	1 Years of service	2.478640
2 Years of service	10.299486	2 Education level	2.196539
3 Education level	2.409263		

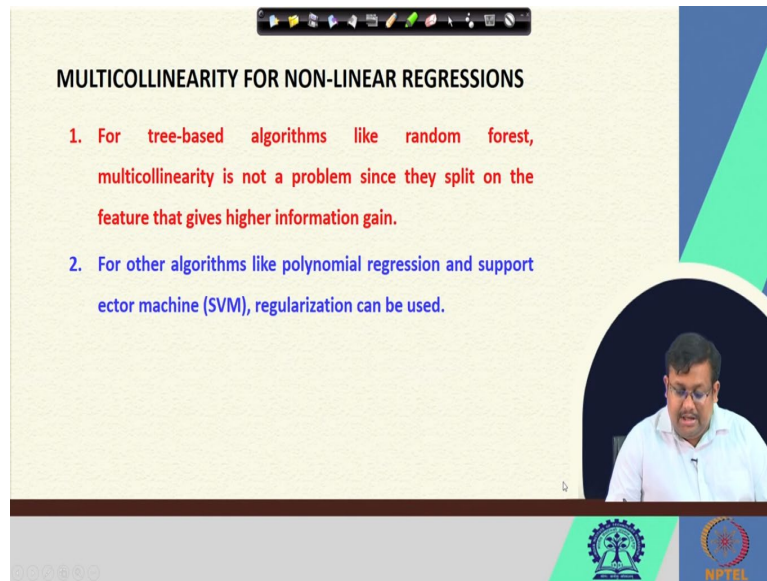
Source: <https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/>

So, what is the other way? So, as I have told you, once you identify the variable which shows the higher VIF score, to resolve the multi coordinate you have to drop that correlated variables. So, dropping one of the correlated variables will help in bringing down the multi coordinate between the correlated features.

For example, here one example is given. You see here, there are four variables gender, age, years of service and educational level and you can see that VIF scores are given 2.20, 13.7, 10.2 and 2.40. So, if we remove the age variable because it is showing the higher VIF. So, if we remove the ege variable then you can see, we can reduce the VIF score for all other variables. So, you can see that in after removing or dropping the age factor, we can see that the VIF of years of service also goes down from 10 to 2.

So, that shows that the avoiding multi coordinate means, you drop one variable at a time and see then the effect on the VIF score and then you repeat this step until you are satisfied that, ok, we have remove all the highly correlated variables and our VIF score for rest of the variables are within our tolerance limit. So, this is how you avoid the multicollinearity effect.

(Refer Slide Time: 13:50)



MULTICOLLINEARITY FOR NON-LINEAR REGRESSIONS

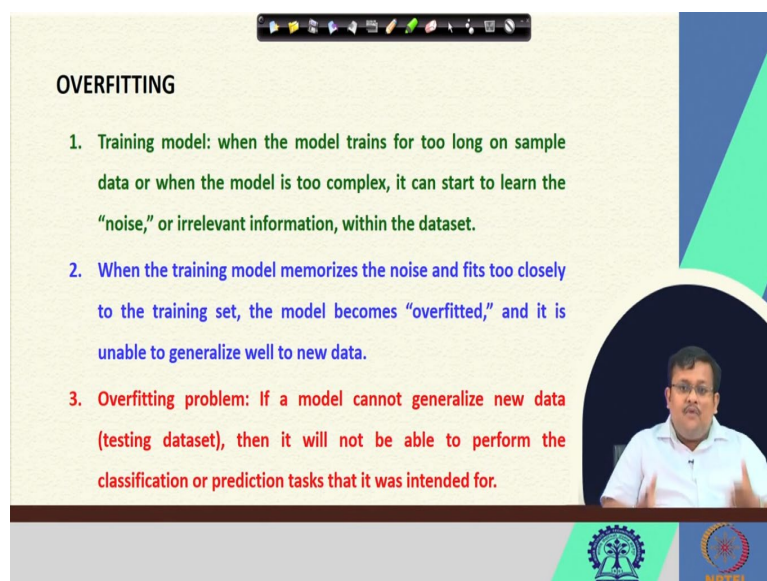
1. For tree-based algorithms like random forest, multicollinearity is not a problem since they split on the feature that gives higher information gain.
2. For other algorithms like polynomial regression and support vector machine (SVM), regularization can be used.

The slide includes a video inset of a speaker in the bottom right corner and logos for IIT Bombay and NPTEL at the bottom.

Now, so, the question arise, whether there is a multicollinearity problem in non-linear regressions or not? So, non-linear regressions like the tree base algorithms, like random forest, their multi collinearity is not a problem, because in case of tree base algorithm, they split on the feature that gives the higher information gain.

So, that is why the multicollinearity effect is not prevalent in case of non-linear regression like tree-based regression, like random forest. Now, for other algorithms, like polynomial regression and support vector machine, regularization can be used. Please, change it, it will be support vector. So, v is missing. So, it will be support vector machine.

(Refer Slide Time: 14:56)



OVERFITTING

1. Training model: when the model trains for too long on sample data or when the model is too complex, it can start to learn the "noise," or irrelevant information, within the dataset.
2. When the training model memorizes the noise and fits too closely to the training set, the model becomes "overfitted," and it is unable to generalize well to new data.
3. Overfitting problem: If a model cannot generalize new data (testing dataset), then it will not be able to perform the classification or prediction tasks that it was intended for.

The slide includes a video inset of a speaker in the bottom right corner and logos for IIT Bombay and NPTEL at the bottom.

Now, the next important thing pitfall of any regression problem is overfitting. Now, over fitting, you know, what is the training model? training model means, when the model is trained for too long on sample data or when the model is too complex, it can start to learn the noise in the data or irrelevant information with the, within the data set.

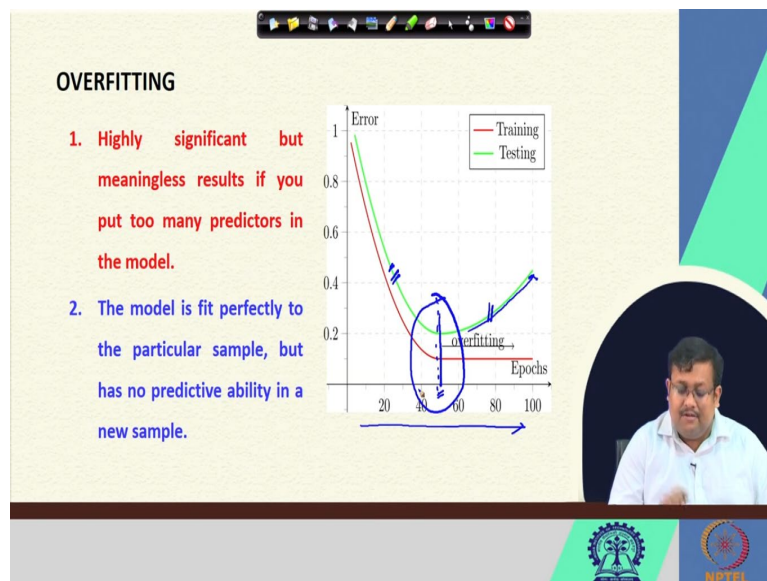
So, it depends on the, always with, the complexity of the data. So, if we make our model much complex summed or we train the model for too long then what will happen? It will learn not only the important feature of the dataset but also it will learn based on the noise, which is also present in the dataset. So, when the training model memorizes the noise and fits too closely to the training set, the model becomes overfitted.

Why we call it over fitting? Because, it is unable to generalize well the new data. So, that means, suppose, we have a training dataset and we have trained our model so well that it is giving us very high accuracy but when we try to predict the unknown sample based on the training model, it is failing miserably. The R square values is almost negligible. So, that is called overfitting.

So, that means, we have trained our training model so well or it is becoming too complex that it has memorized not only the important features but also the noise to perfectly fit to make a perfect fit. But, this perfect fit may not be useful for generating the useful prediction for unknown dataset. Because, unknown dataset are independent to the training dataset.

So, if a model cannot generalize new data or testing dataset or we sometime call it validation dataset, then it will not be able to perform the classification or prediction task, it was intended for. Sometime we create a very high over optimistic model, which is over fitted model, but in real life application that model is useless, because it is over fitted. So, that is why we should be very very careful about over fitting problem in any type of multivariate regression.

(Refer Slide Time: 17:27)



Now, this figure shows the overfitting problem. So, this overfitting is highly significant but meaning but produce meaningless results if you put too many predictors in the model. So, when you incorporate too many predictors in the model sometimes the model get overfitted. That is one of the reason. But, you may get high R square, where you may get very high model accuracy but it is overfitted. The model is fit perfectly to the particular sample but it has no predictive ability in a new dataset.

So, this graph shows a good example of overfitting. So, you can see with the number of epochs, when the number of epochs is increasing in any dataset and then you can see the error for the training data are continuously decreasing and also you will see that the error rate of the testing dataset is also increasing with each iteration, but after a certain point, it will start increasing. So, this feature is showing the overfitting that means our training model is not able anymore to generalize the testing dataset. So, this is called overfitting.

Now, it is our as a modeler, it is always desirable to find the sweet spot. This is called the switch for the optimum position where we should stop our modeling. So, here you can see this is where the error of the testing dataset produces the minimum error. So, we should select these epochs, as the suitable, we should use this split of the data as an optimum training and testing dataset.

So, this is called overfitting and here it is showing under fitting. So, here it is showing over fitting and here it is showing under fitting. So, as a modeler we should always try to find this 'sweet spot' to find out the point where the testing dataset is producing the lowest error and we should select that.

(Refer Slide Time: 19:59)

Estimating unknown function

- ▶ Suppose we observe Y_i and $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})$ for $i = 1, \dots, n$.
- ▶ We believe that there is a relationship between Y and at least one of the X 's. So we model the relationship as

$$Y_i = f(\mathbf{X}_i) + \epsilon_i \quad \text{with} \quad E\{\epsilon_i\} = 0,$$

where f is an unknown function and ϵ is a random error.

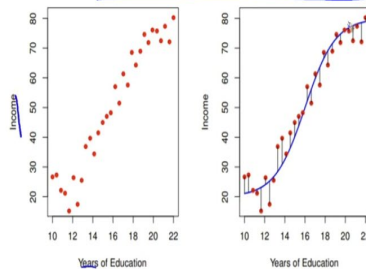


Figure from ISLR 2013

Now, how to estimate an unknown function? Suppose, we have observe y_1 and x_1 , you know there are different X_{i1}, X_{i2} , that for i values varies from 1 to n , we have already seen the example. So, we believe that there is a relationship between y and at least one of the x 's. So, we model the relationship as y equal to function of x_i plus ϵ_i , where expected values of ϵ_i equal to 0, we know that, where f are mean values of residual is 0, we know that, where f is unknown function and ϵ is the random error or residual.

So, you can see if we try to model these years of education with the income, we can fit a function like this. So, you know, when we do the modeling, we always want to develop this type of function.

(Refer Slide Time: 21:01)

Income vs. education and seniority

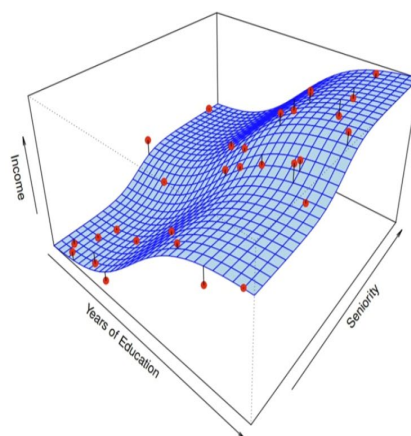
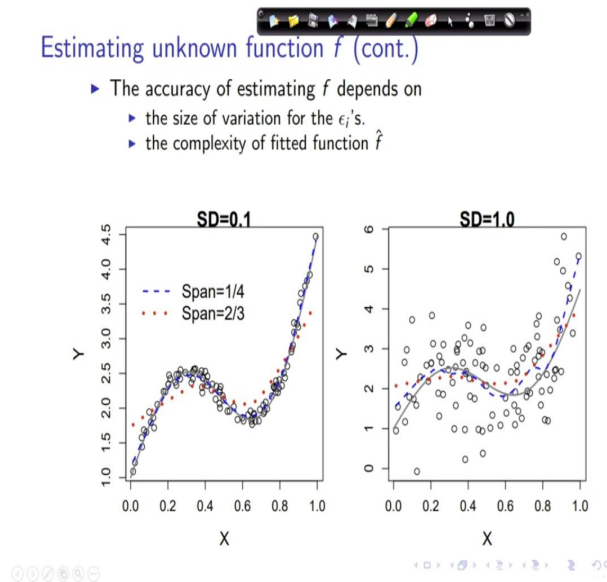


Figure from ISLR 2013

So, if we fit a model between income, years of education and seniority, we can fit a model like this and this is the three-dimensional plot for this type of regression.

(Refer Slide Time: 21:18)



Now, when we estimate the unknown function, the accuracy of the estimation of f depends on the size of the variation of the ϵ 's and the complexity of the fitted function. You can see here, where the standard deviation in this case, both in both these cases we are regressing y against x , but in both these cases one is showing, you know, relatively low standard deviation and here we are getting relatively higher standard deviation. So, it depends on the complexity size of the variation of the error term.

(Refer Slide Time: 21:54)

- How Do We Estimate f ?
- ▶ Use the training data $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ and a statistical method to estimate f .
 - ▶ Two groups of statistical learning methods:
 - ▶ Parametric methods:
 - ▶ Make some assumption about the functional form of f (e.g. MLR).
 - ▶ Pros: estimating $f \Rightarrow$ estimating a set of parameters (relatively easy task). Easy to interpret the model.
 - ▶ Cons: The form of model is too rigid. Low prediction accuracy when f is complicated.
 - ▶ Non-parametric methods:
 - ▶ Do not make explicit assumption about the functional form of f (e.g. neural network, tree).
 - ▶ Pros: accurately fit a wider range of possible shapes of f .
 - ▶ Cons: Large number of observations is required to obtain an accurate estimate of f .

Now, how do we estimate f ? So, if you use the training data that is $x_1, y_1, x_2, y_2, \dots, x_n, y_n$, and statistical method to estimate f there are two types of method. One is parametric method another is non-parametric method. So, what are the parametric methods? Parametric methods make some assumption about the functional forms of f , for example multiple linear regression.

Multiple linear regression is a parametric model. Because it makes assumption about the functional form of f . What are the advantage of parametric model? Because, estimation of f , you know estimating a set of parameter. Relatively, it is an easy task and easy to interpret the model. So, model interpretation is easy in case of parametric model.

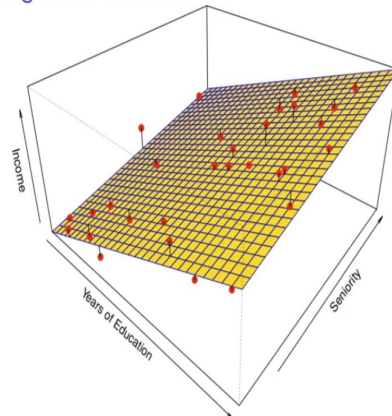
However, there is a disadvantage also. So, the form of model is too rigid. So, low prediction accuracy when f is complicated. So, what happens, when there is a parametric model? The model is too rigid, so, it will always give you the low prediction accuracy, when the f is too complicated.

The other type of model is non-parametric model, which do not make any explicit assumption about the functional form of f . For example, neural network, tree based model, random forest, these are all non-parametric methods. What are the advantages? Advantages is it can accurately fit wider range of possible shapes of function. What are the disadvantage? Disadvantage means large number of observation is required to obtain an accurate estimation of f .

So, the sample size is always a matter of concern in case of neural network. You need to have huge number of samples to get meaningful results from training a neural network model. So, now we know what are the parametric model and what are the non-parametric model.

(Refer Slide Time: 23:49)

A linear regression estimate



Even if the standard deviation is low, we will still get a bad answer if we use the wrong model.

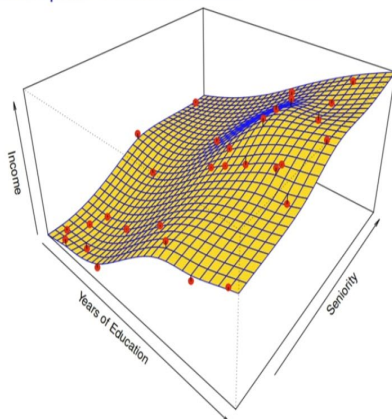
Figure from ISLR 2013

Navigation icons and page number 13

Now, if we do a linear regression estimate. So, we can see this type of plane, we can see this type of shape of the regression function. So, even if the standard deviation. So, in case of linear regression, even if the standard deviation is low, we will still get a bad answer, if we use the wrong model. So, in case of linear regression model, the standard deviation is low, but there is high bias.

(Refer Slide Time: 24:30)

A thin-plate spline



Non-linear regression methods are more flexible and can potentially provide more accurate estimates.

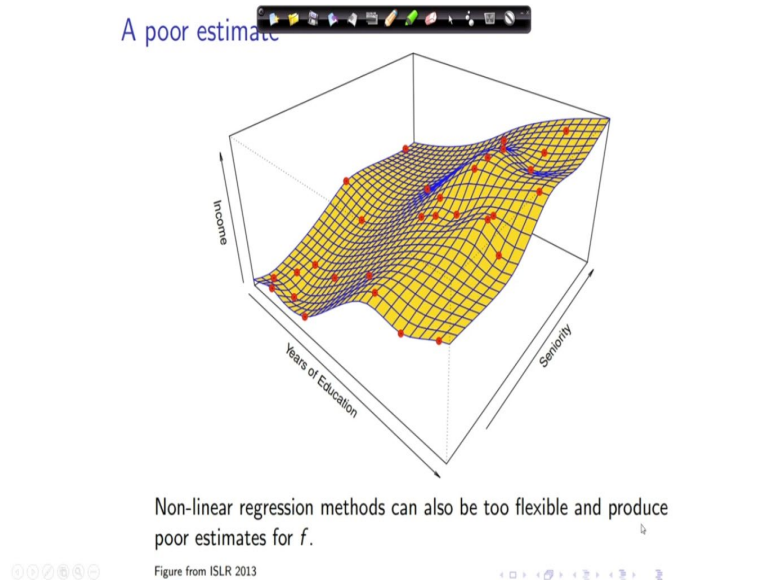
Figure from ISLR 2013

Navigation icons and page number 14

So, let us move to non-linear model. So, we have seen linear model, let us see a thin plate spline fitting the non-linear model. So, you can see here the non-linear regression methods are more flexible and can potentially provide more accurate estimate. But at the same time,

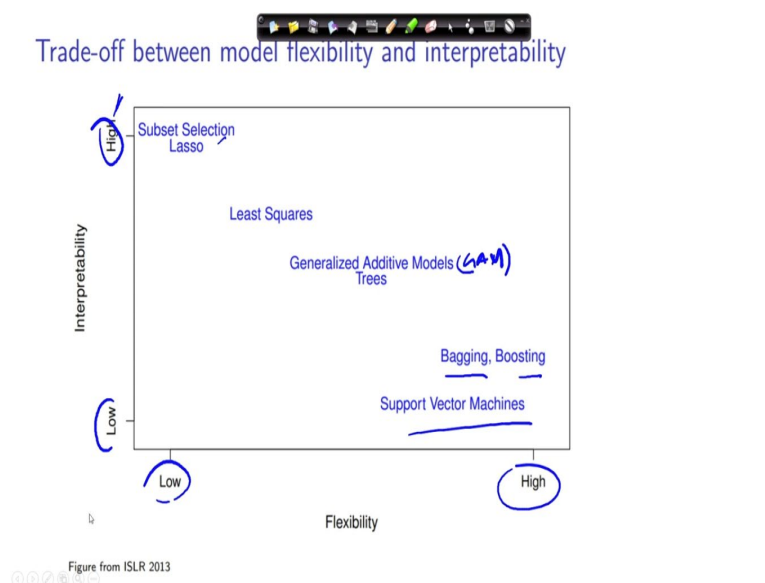
they have high variance but low bias. They have low bias but high variance. So, this is the problem of the non-linear regression model also.

(Refer Slide Time: 24:56)



So, what is the poor estimate? The poor estimate is non-linear regression methods can also be too flexible and produce poor estimate of f . So, when the non-linear methods are becoming too flexible, it can perfectly fit any data and it becomes too accustomed or too learned based on the noise in the data. So, it can show overfitting. So, when there is an overfitting, of course, there will be higher variance. Although there will be low bias but there will be higher variance. So, that is the problem.

(Refer Slide Time: 25:40)



So, what is the trade-off of the different model flexibility and interpretability? So, you can see here based on the flexibility and interpretability, we have, the models can be classified. So, here in the x-axis you can see that these are low flexible model and these are highly flexible model. Here also, they are low interpretable model and they are highly interpretable model.

So, you can see that although the Subset Selection, Lasso these are the low flexible model, they are highly interpretable however. So, Least Squares model are somewhat intermediate. So, then we go to Generalized Additive Models (GAM) or Trees. So, when we go from subset selection to least square, then GAM, we call it model and then Least Squares and then Tree base models the flexibility increases.

Then we go to Bagging and Boosting. Then Support Vector Machine. We are getting high flexibility but at the same time with the cost of the lower interpretability, because, this Subset Selection Lasso, Least Squares have higher interpretability. However, when you go to Support Vector Machine they have very low interpretability. So, we need to find a trade-off between this type of flexibility and interpretability of the model.

(Refer Slide Time: 27:08)

Bias variance tradeoff

- ▶ Two competing forces govern the choice of learning method, i.e. **bias** and **variance**.
- ▶ *Bias* refers to the error that is introduced by modeling a real life problem (that is usually extremely complicated) by a much simpler problem.
 - ▶ For example, linear regression assumes that there is a linear relationship between Y and X , which is unlikely in real life.
 - ▶ In general, the more flexible/complex a method is the less bias it will have.
- ▶ *Variance* refers to how much your estimate for f would change by if you had a different training data set.
 - ▶ Generally, the more flexible a method is the more variance.
 - ▶ In general, the more flexible/complex a method is the less bias it has.
- ▶ It can be shown the expected MSE for a new Y at x^{new} is:

$$E[MSE(x^{new})] = \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}$$

So, what is that is called Bias Variance Tradeoff. So, two competing forces that governs the choice of learning method are called bias and variance. What is bias? Bias refers to the error, that is introduced by modeling a real life problem. So, that is, usually, extremely complicated by a much simpler problem.

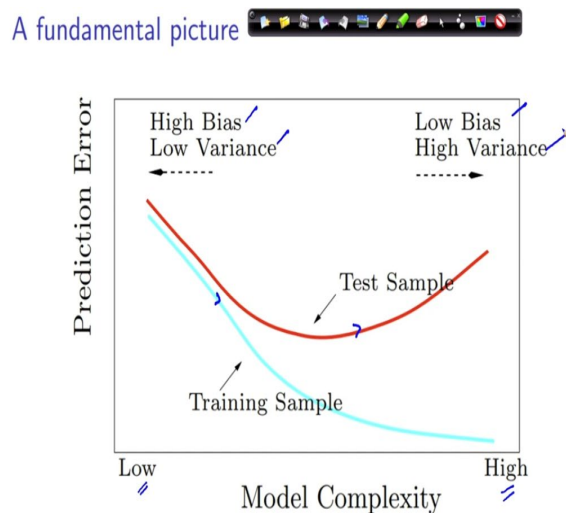
So, for example, let us consider there is a very complex relationship between y and x and we want to generalize that relationship by a simple linear regression or a linear regression. So, that will be creating bias. So, linear regression, where linear regression assume that there is a linear relationship between y and x , which is unlikely in real life that creates bias.

Now, in general, the more flexible or complex a method is, less bias it will have, because, the complex model will have the ability to perfect fit itself to the entire data set. So, in that way that will lower the bias. It will perfectly fit to the complexity of the model. So, that is, why it is called low bias model. So, higher complex model are showing the low bias.

Now, let us discuss, what is variance? Variance refers to, how much your estimate for function f would change by, if you had a different training data set. So, generally the more flexible the method, the more the variance. So, that is why complex model is used to have more variance than the simple model or linear model. So, in general, more flexibility or complexity a method is having the less bias.

So, we know that, basically the idea is, when the model is less complex, then the bias is high, but the variance is low. But when the model becomes too complex, the bias becomes low, but the variance becomes high. So, we have to find a sweet spot. So, we know that the expected MSE is, mean of MSE for a new y at a x new is, can be calculated by using this formula, where, the expected values of MSE will be irreducible error plus bias square plus variance. So, using this equation, there are different methods which have evolved to balance the bias and variance in a model. So, that is called bias variance tradeoff.

(Refer Slide Time: 30:03)



So, we can see here, a fundamental picture, we call it for the bias variance tradeoff, here you can see there is a model complexity, less complex model and this is a highly complex model. So, as the complexity of the model increases, you can see, the prediction error for the training set always going down, but the prediction error for the test sample goes up after a certain point. So, this type of thing, you can see, when there is a trade-off, this type of thing we can see, when you go from the lower complex model to higher complex model.

And you can see that when the model is less complex that will have high bias but low variance, but when the model is complex that will have low bias but they will have high variance. So, this is the picture of bias variance tradeoff and we should be very very careful for selecting the optimum model for modeling while taking care of this type of bias and variance and their interplay.

(Refer Slide Time: 31:21)

A cautionary note

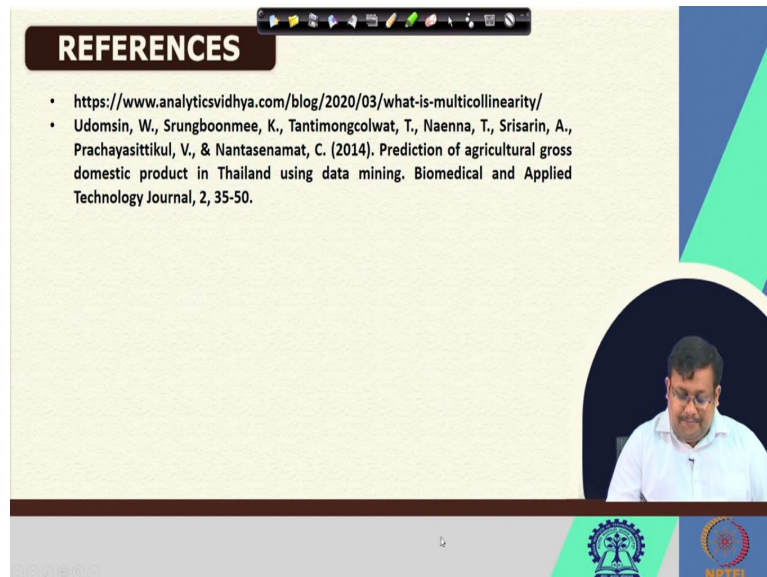
- ▶ George Box, a famous statistician and son-in-law of R.A. Fisher, once said:
"All models are wrong, but some are useful."
- ▶ In practice, there is really NO *true* model but a *good* model.
- ▶ A good model should achieve at least one of the following:
 - ▶ an interpretable model that can be explained by some known facts or knowledge;
 - ▶ reveals some unknown truth or relationship among the variables or observations;
 - ▶ a model with accurate prediction on new samples.
- ▶ The optimal model depends on:
 - ▶ the purpose of the study;
 - ▶ the complexity of the underlying mechanism;
 - ▶ the quality of the data and signal-noise-ratio;
 - ▶ the sample size.

So, a cautionary note we should mention that George Box a famous statistician and you know son-in-law of R. A. Fisher, you know, R. A. Fisher, once said that all models are wrong but some are useful. So, in practice there is really no true model, but a good model is there. There is no true model, but there is good model.

So, a good model should achieve at least one of the following. First of all, it should achieve an interpretable model that can be explained by some known facts or knowledge. Then it should reveal some unknown truth or relationship among the variables or observation. And thirdly, the model with accurate prediction on new samples.

So, these are the characteristics of a very good model. And the optimum model always depends on the purpose of the study, then the complexity of the underlying mechanism and the quality of the data and signal-to-noise ratio and finally the sample size.

(Refer Slide Time: 32:31)



REFERENCES

- <https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/>
- Udomsin, W., Srungboonmee, K., Tantimongcolwat, T., Naenna, T., Srisarin, A., Prachayasittikul, V., & Nantasenamat, C. (2014). Prediction of agricultural gross domestic product in Thailand using data mining. *Biomedical and Applied Technology Journal*, 2, 35-50.

So, now, we know these are the references. So, we have now discussed, we have now seen the pitfalls of MLRs, Multiple Linear Regression model, we have seen what is multicollinearity, we have seen the overfitting problem, we have seen how to detect the multicollinearity using VIF, and we have also discussed the bias variance trade-off.

Let us wrap up our lecture here and in the next lecture, we will start from here and we will discuss some other aspects of multivariate data analytics. So, thank you guys. Let us meet in our next lecture.