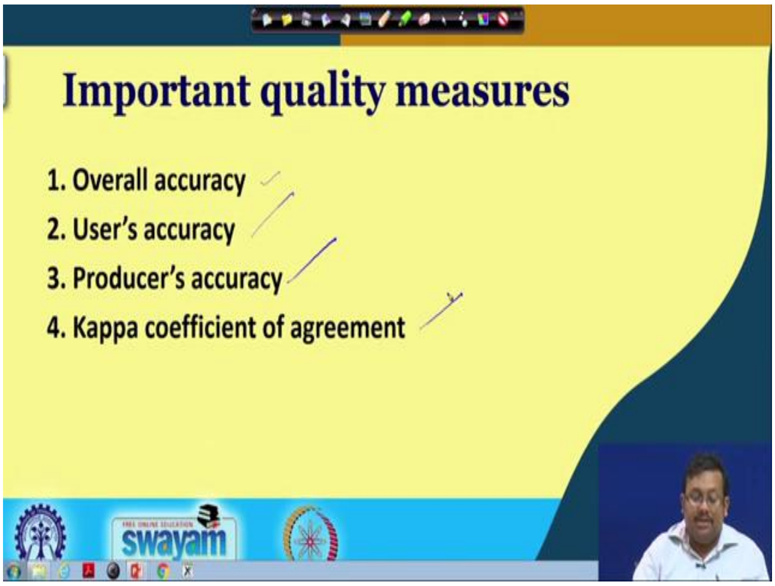
Soil Science and Technology
Prof. Somsubhra Chakraborty
Department of Agricultural and Food Engineering
Indian Institute of Technology, Kharagpur

Lecture – 59
Modeling Categorical Variables

Welcome friends to this forth lecture of week 12 of Soil Science and Technology and in this lecture we will be discussing about how to model and different categorical variables in digital soil mapping. And we will try to complete this and then we will discuss about the pedo transfer functions.

So, we will start from this slide where we left in the last lecture; obviously, we started this categorical modeling and let us see in details about how to measure different types of, how to measure different types of important quality measures in case of categorical modeling.

(Refer Slide Time: 01:00)

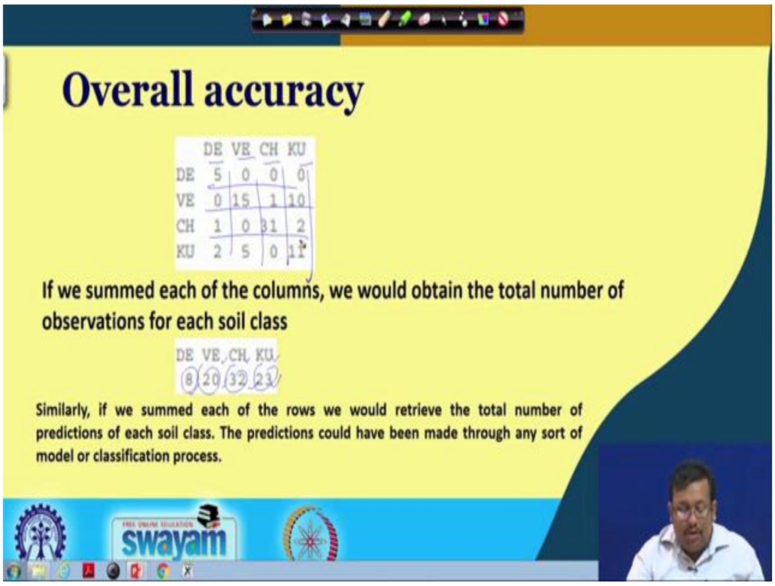


Important quality measures

1. Overall accuracy
2. User's accuracy
3. Producer's accuracy
4. Kappa coefficient of agreement

So, there are four different types of quality measures in case of categorical modeling one is called the overall accuracy, then users accuracy, then producers accuracy and kappa coefficient of agreement. So, let us see them one by one.

(Refer Slide Time: 01:19)



Overall accuracy

	DE	VE	CH	KU
DE	5	0	0	0
VE	0	15	1	10
CH	1	0	31	2
KU	2	5	0	11

If we summed each of the columns, we would obtain the total number of observations for each soil class

DE	VE	CH	KU
8	20	32	23

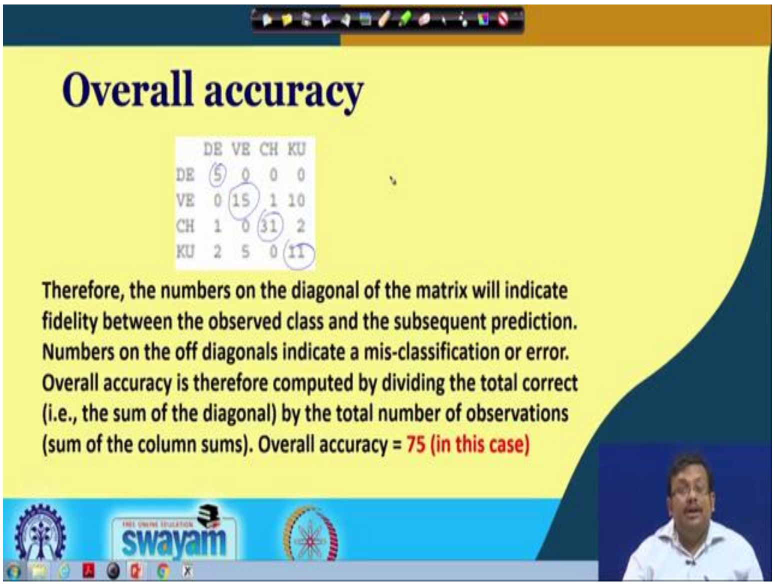
Similarly, if we summed each of the rows we would retrieve the total number of predictions of each soil class. The predictions could have been made through any sort of model or classification process.

So, overall accuracy let us consider this example. Let us see let us consider there are four different soil classes DE, VE, CH and KU they are the names of four different soil classes. So, this matrix is called confusion matrix which basically shows how many of the samples of using our categorical model have been correctly classified. Classification is basically assigning a particular observation correctly to a particular to their respective classes.

So, classification is a term used when we use categorical variables and regression is term when we use continuous variables. So, here you can see we have classified several soil classes into their respective classes and these are four different classes DE, VE, CH and KU their names. And, if we summed each of the column; so, this is a column, this another column, this another column, this another column. So, if you sum all each of the column we would obtain a total number of observation for each soil class. So, in case of DE we will get 8, VE 20, CH 32, KU 23.

So, these are the total number of observation, you know, for each soil class. So, similarly if we summed up each of the rows; each of the rows just like this row this row this row this row, we retrieve the total number prediction for each of the classes. So, if we sum again all the columns will get the number total number of observation for each soil class. And if we sum all the rows we retrieve the total number of prediction for each soil class.

(Refer Slide Time: 03:11)



Overall accuracy

	DE	VE	CH	KU
DE	5	0	0	0
VE	0	15	1	10
CH	1	0	31	2
KU	2	5	0	11

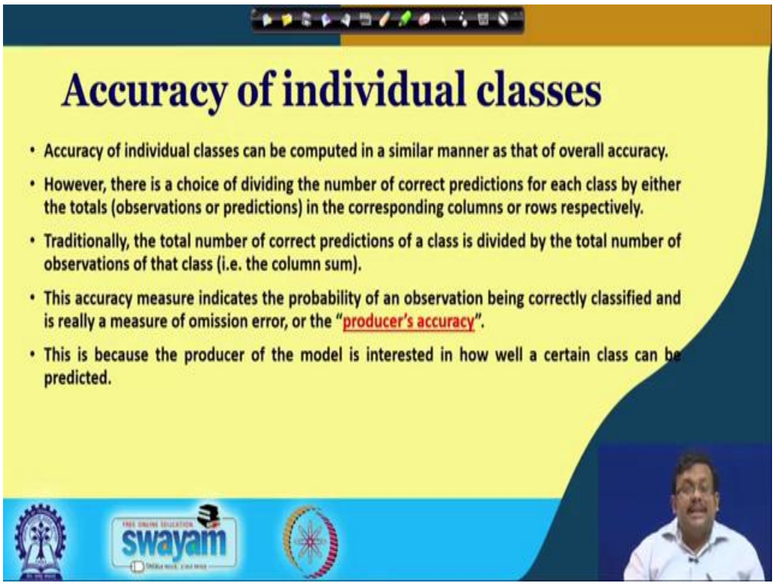
Therefore, the numbers on the diagonal of the matrix will indicate fidelity between the observed class and the subsequent prediction. Numbers on the off diagonals indicate a mis-classification or error. Overall accuracy is therefore computed by dividing the total correct (i.e., the sum of the diagonal) by the total number of observations (sum of the column sums). Overall accuracy = 75 (in this case)

So, what is the next step? The next step says the overall accuracy now another important thing I should mention that these diagonal values basically represents in the matrix basically indicate the fidelity between the observed class and the subsequent prediction.

That means the 5 samples have been correctly classified as DE, 15 samples are correctly classified as VE, 31 samples are correctly classified as CH and 11 samples are correctly classified as the KU samples and all other off diagonal samples are misclassification; that means, incorrect classification. So, numbers on the off-diagonals indicates a misclassification error and overall accuracy is therefore, computed by dividing the total correct that is a sum of the diagonal by the total number of observation sum of the column sums.

So, if we divide the; if we divide the total correct classification that is sum of the diagonal by the total number of observation sum of the column sums then we will get the overall accuracy. In our case this is 75. So, this is the one of the measures of classification accuracy.

(Refer Slide Time: 04:33)



Accuracy of individual classes

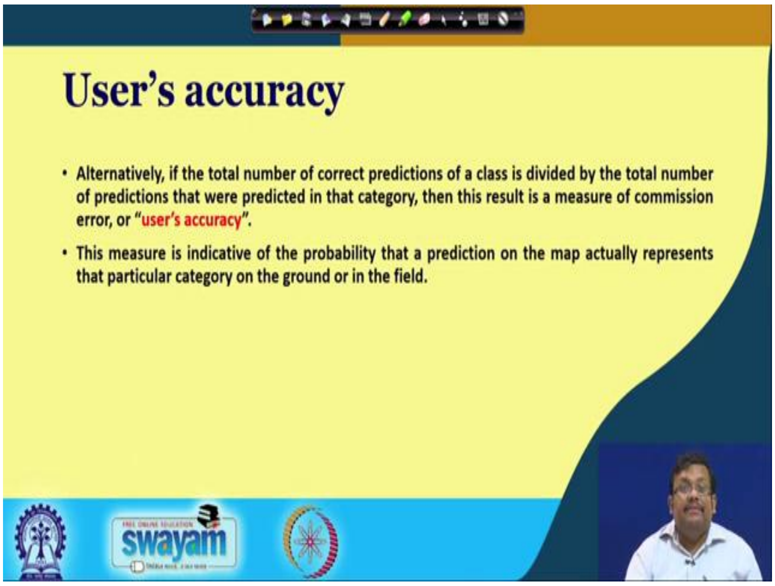
- Accuracy of individual classes can be computed in a similar manner as that of overall accuracy.
- However, there is a choice of dividing the number of correct predictions for each class by either the totals (observations or predictions) in the corresponding columns or rows respectively.
- Traditionally, the total number of correct predictions of a class is divided by the total number of observations of that class (i.e. the column sum).
- This accuracy measure indicates the probability of an observation being correctly classified and is really a measure of omission error, or the "producer's accuracy".
- This is because the producer of the model is interested in how well a certain class can be predicted.

The next measure of, you know, is called producers accuracy, but before the producers accuracy let us see what are the other important aspects. Like accuracy of individual classes can be computed in a similar manner as that of overall accuracy. However, there is a choice of dividing the number of correct prediction for each class by either the total that is observation or predictions in the corresponding columns or rows respectively. Traditionally, the total number of correct prediction of a class is divided by the total number of observation of the class that is column sum. We are already got the column sums for individual classes.

So, this accuracy measures the indicate the probability of an observation being correctly classified and it is really a measure of omission error or the producers accuracy. This is because producers of the model is interested in how well a certain class can be predicted. So, this is called producers accuracy. Again, the producers accuracy is computed when the total number of correct prediction of a class is divided by the total number of observation of that class. I have already told you the correct predictions by indicating this diagonal values or diagonal in this in this confusion matrix.

However, when we take the ratio of this correct predictions of a class with the total number of observation of that class, will get a producers accuracy.

(Refer Slide Time: 06:13)

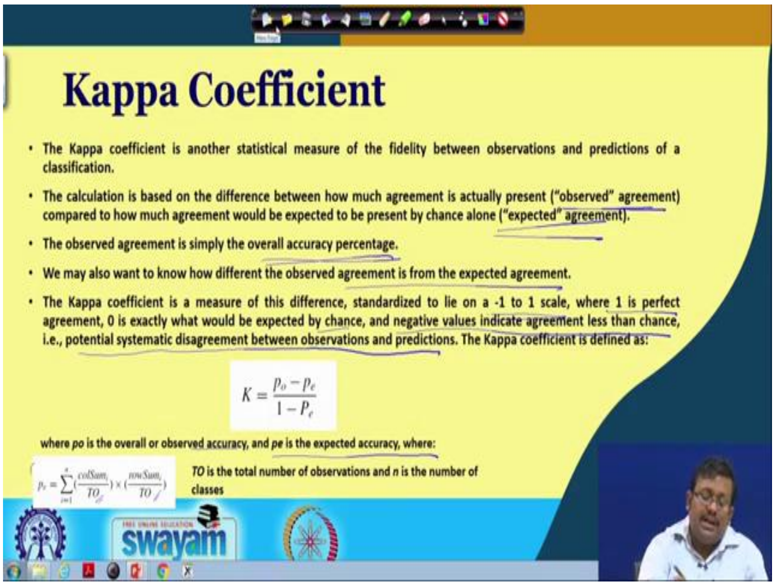


User's accuracy

- Alternatively, if the total number of correct predictions of a class is divided by the total number of predictions that were predicted in that category, then this result is a measure of commission error, or "user's accuracy".
- This measure is indicative of the probability that a prediction on the map actually represents that particular category on the ground or in the field.

Another important term is users accuracy. Now, alternatively if the total number of correct prediction of a class is divided by the total number of prediction that were predicted in that category, then it is result; then this result is a measure of commission error you know commission error or users accuracy. And, this measures is indicative at the probability that the prediction of the map actually represents that particular category on the ground or in the field.

(Refer Slide Time: 06:44)



Kappa Coefficient

- The Kappa coefficient is another statistical measure of the fidelity between observations and predictions of a classification.
- The calculation is based on the difference between how much agreement is actually present ("observed" agreement) compared to how much agreement would be expected to be present by chance alone ("expected" agreement).
- The observed agreement is simply the overall accuracy percentage.
- We may also want to know how different the observed agreement is from the expected agreement.
- The Kappa coefficient is a measure of this difference, standardized to lie on a -1 to 1 scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than chance, i.e., potential systematic disagreement between observations and predictions. The Kappa coefficient is defined as:

$$K = \frac{p_o - p_e}{1 - P_e}$$

where *po* is the overall or observed accuracy, and *pe* is the expected accuracy, where:

$$p_e = \sum_{i=1}^{n} \left(\frac{colSum_i}{TO}\right) \times \left(\frac{rowSum_i}{TO}\right)$$

TO is the total number of observations and *n* is the number of classes

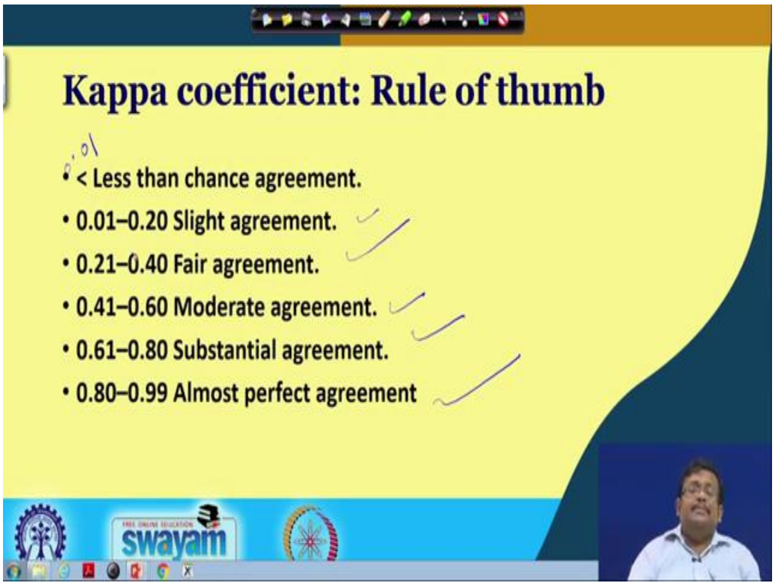
Another, so, we have covered the overall accuracy, then users accuracy, then producers accuracy, the fourth important point is Kappa coefficient. Now, the Kappa coefficient is another statistical measure of the fidelity between observations and predictions of a classification. And, the calculation is based on the difference between how much agreement is actually present that is observed agreement, compared to how much agreement would be expected to be present by chance alone, that is, expected agreement.

So, we calculate this Kappa coefficient by the difference between the observed agreement and the expected agreement again the observed agreement is how much agreement is actually present and expected agreement is, you know, how much agreement would be expected to the present by chance alone. So, the observed agreement is simply the overall accuracy percentage. In our case we have calculated already that overall accuracy is 75 percentage.

So, we may also want to know how different the observed agreement from the expected agreement. Now, the Kappa coefficient is a measure of this difference standardized to a line between minus 1 to plus 1 scale. So, again this kappa coefficient takes a value from minus 1 to plus 1; where 1 is the perfect agreement, 0 is exactly what would be expected by chance and negative values indicate agreements less than a chance. Again, plus 1 is a perfect agreement, 0 is exactly what would be expected by chance and negative values indicate the agreement less than chance.

That is potential systematic disagreement between observation and predictions. When there is minus 1 that again suggest that is a potential systematic disagreement between observation prediction. The Kappa coefficient is defined by this small k at this p 0 minus p e by 1 minus P e; where p 0 is the overall or observed accuracy and p e is a expected accuracy, where p e can be calculated by using this formula where TO is basically the total number of observation and n is the number of classes. In our case the number of classes is 4, total observations we already know and then column sum individual and the row sum for individual and then we will get the Kappa coefficient.

(Refer Slide Time: 09:21)



Kappa coefficient: Rule of thumb

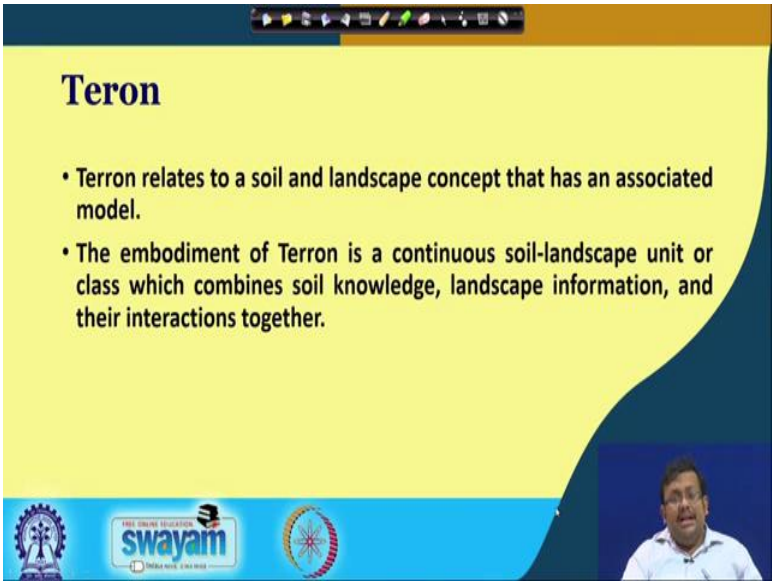
0.01

- < Less than chance agreement.
- 0.01–0.20 Slight agreement.
- 0.21–0.40 Fair agreement.
- 0.41–0.60 Moderate agreement.
- 0.61–0.80 Substantial agreement.
- 0.80–0.99 Almost perfect agreement

Now, once we calculate the Kappa coefficient what is the next step what are the different cut off values of kappa coefficient? The Kappa coefficient has you know as a rule of thumb we you know the scientist have device certain cutoff values for Kappa coefficient. First of all when I mean when there a less than a chance of agreement and you can see when the value is 0.01 to 0.20 there is a slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement, 0.80 to 0.99 almost perfect agreement.

However, when it is less than 0.01 there is less than chance agreement. So, these are some cut off values for Kappa coefficient. So, once we collect the, once we calculate the kappa coefficient you can predict whether there is how much the strength of the agreement you can calculate based on this cut off values.

(Refer Slide Time: 10:23)

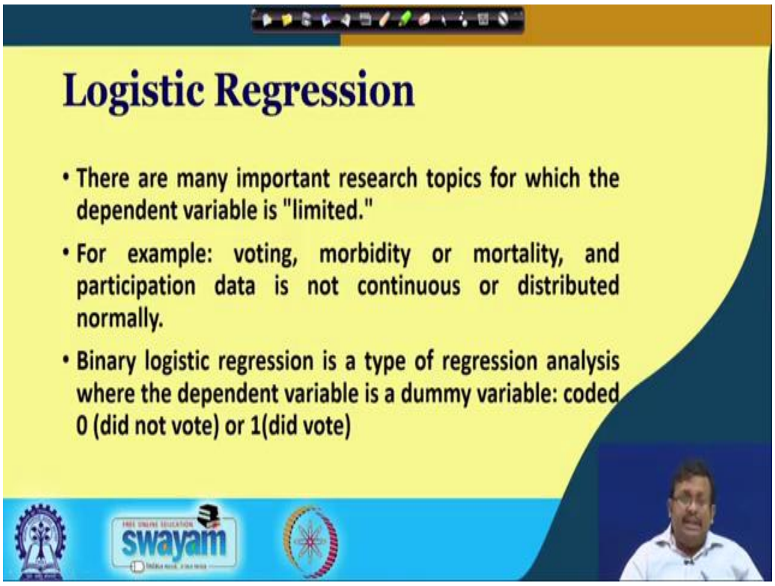


Teron

- Terron relates to a soil and landscape concept that has an associated model.
- The embodiment of Terron is a continuous soil-landscape unit or class which combines soil knowledge, landscape information, and their interactions together.

Now, you know for the classification of soil or you know or the categorical classification of the soil it is very important to know a particular term called Terron. Now, Terron relates to a soil and landscape concept that has an associated model. So, the embodiment of Terron is a continuous soil landscape unit or class which combines soil knowledge landscape information and their interaction together. So, this is very important term.

(Refer Slide Time: 10:49)



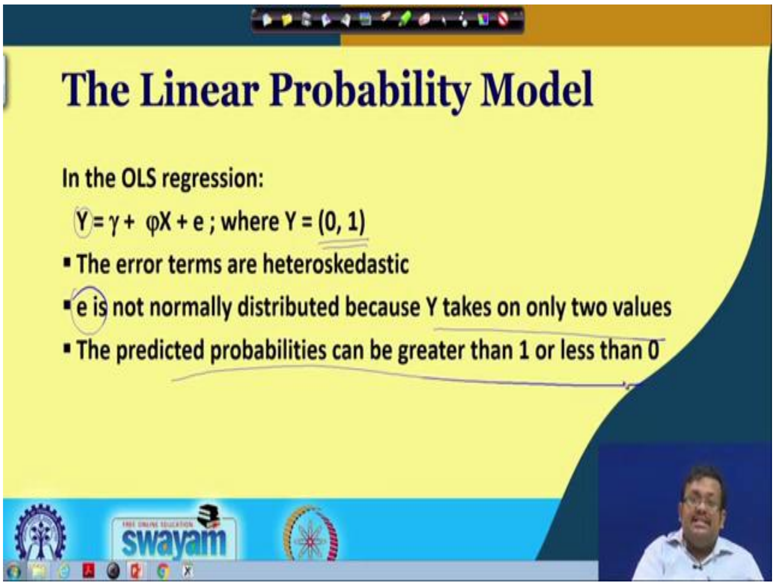
Logistic Regression

- There are many important research topics for which the dependent variable is "limited."
- For example: voting, morbidity or mortality, and participation data is not continuous or distributed normally.
- Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 (did not vote) or 1(did vote)

Now, before going to the applications of, you know, categorical model let us consider a couple of categorical models. So, the first important categorical model we will talk about is logistic regression. So, remember that there are many important research topics for which the dependent variable is limited. For example, voting, morbidity or mortality and participation data is not continuous or distributed normally.

So, binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable coded 0 or 1. For example, if you want to know whether there is a two outcome either they have voted or not voted for a particular candidate you have to, there are binary outcome either voted or not voted. So, you have to code them either 0 or 1; 0 means either did not vote and 1 means vote. So, these are binary coding we call it binary coding. So, you can see the output is very much limited we have binary outputs.

(Refer Slide Time: 11:52)



The Linear Probability Model

In the OLS regression:

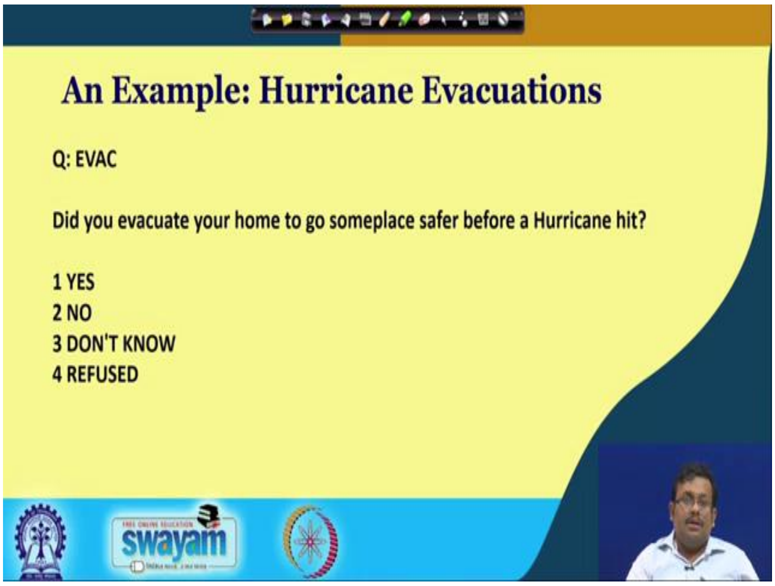
$Y = \gamma + \varphi X + e$; where $Y = (0, 1)$

- The error terms are heteroskedastic
- e is not normally distributed because Y takes on only two values
- The predicted probabilities can be greater than 1 or less than 0

So, if we use the linear probability model ordinary least squares regression, we will get this gamma plus phi X plus e; where Y in our case is only 0 to 1 and the error term is heteroskedastic. So, you can see here Y is only 2, I mean the value the y can take is only 2. So, e in this case you will see if we use a linear prediction model, e or the error will not be normally distributed because Y takes only two values, but these normal distribution of error is very much essential for, essential assumption for linear regression. However, in this case we will not get it because we are getting two binary output either 0 or 1.

So, the predicted probabilities can be greater than 1 or less than 0, but we know that probabilities always lie between 0 to 1; it cannot be less than 0 or greater than 1. However, if we have only this, you know, binary outputs or limited outputs we will get this, you know, some absurd probability.

(Refer Slide Time: 13:10)



An Example: Hurricane Evacuations

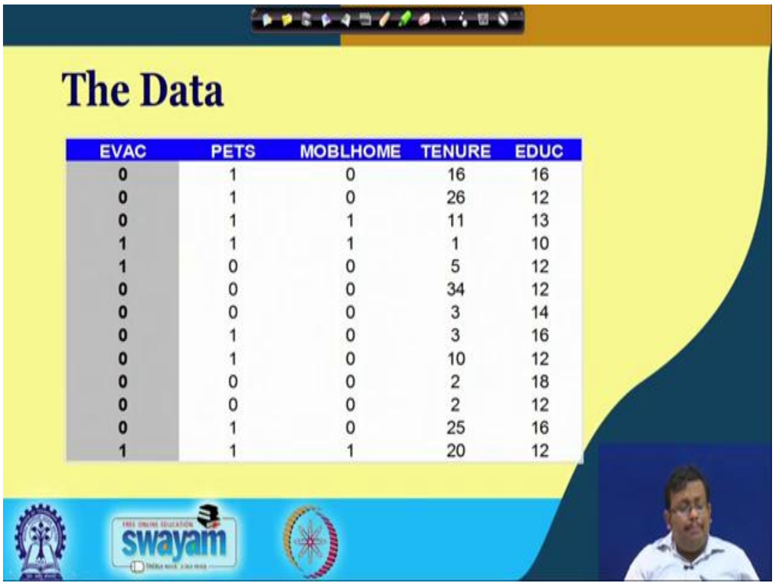
Q: EVAC

Did you evacuate your home to go someplace safer before a Hurricane hit?

1 YES
2 NO
3 DON'T KNOW
4 REFUSED

For example, let us see some example. I have seen this example recently. So, I am using this for your better understanding. There is a question here where ask, you know, they have asked people whether they will evacuate to some place safe, you know, some safer place if there is a hurricane. So, there are four options: yes, no, do not know or refused.

(Refer Slide Time: 13:33)

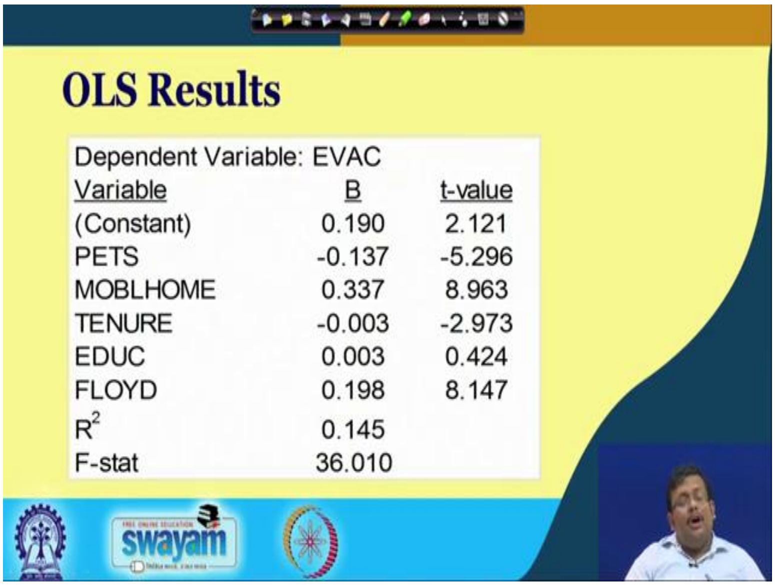


The Data

EVAC	PETS	MOBLHOME	TENURE	EDUC
0	1	0	16	16
0	1	0	26	12
0	1	1	11	13
1	1	1	1	10
1	0	0	5	12
0	0	0	34	12
0	0	0	3	14
0	1	0	3	16
0	1	0	10	12
0	0	0	2	18
0	0	0	2	12
0	1	0	25	16
1	1	1	20	12

So, the data looks like this. So, here you can see the evacuation either they will be, no they will not evacuate, they are, you know, they are denoted by 0 and evacuate they are denoted by 1. So, these are binary coding you can see. So, this is our target class either evacuate or not evacuate. What are the predicted variable? Number of pets. So, you can see hear the number of pets; mobile home, if they have mobile home; then they have tenure of this home and their education. So, based on this particular, you know, four different variables we have to classify either they will evacuate or they will not evacuate.

(Refer Slide Time: 14:20)



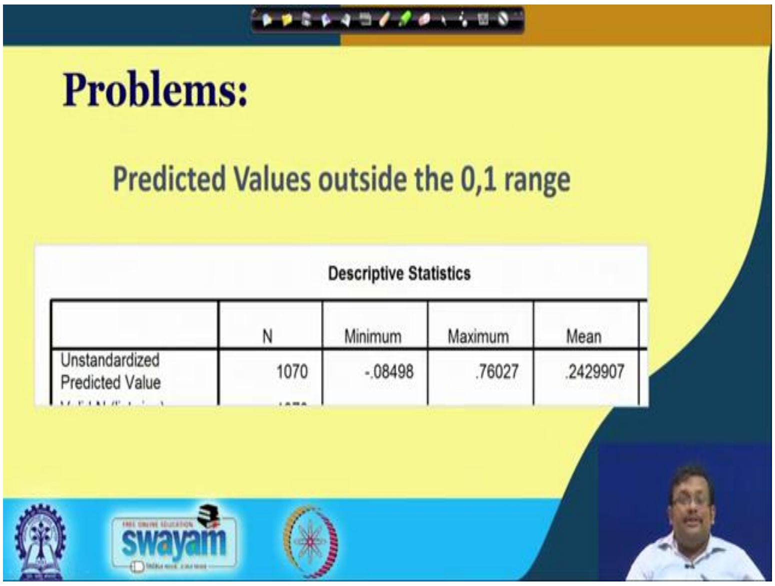
OLS Results

Dependent Variable: EVAC

Variable	B	t-value
(Constant)	0.190	2.121
PETS	-0.137	-5.296
MOBLHOME	0.337	8.963
TENURE	-0.003	-2.973
EDUC	0.003	0.424
FLOYD	0.198	8.147
R^2	0.145	
F-stat	36.010	

So, if we use the ordinary least squares regression or linear regression ultimately you will get this coefficient values.

(Refer Slide Time: 14:31)



Problems:

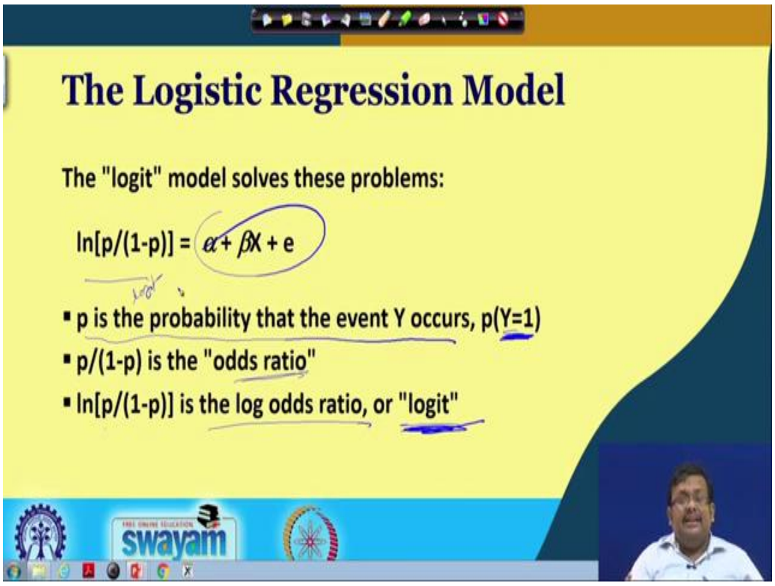
Predicted Values outside the 0,1 range

Descriptive Statistics

	N	Minimum	Maximum	Mean
Unstandardized Predicted Value	1070	-.08498	.76027	.2429907

And, if we use this coefficient values ultimately you will see for some and we created regression relationship we will get for certain number of, you know, numbers of subjects on certain number of N we will get minimum maximum values which is you can see minimum values here less than 0, which is not possible. So, this is the predicted values outside the 0 to 1 range. So, this is the problem in case of ordinary least squares when we are having limited numbers of outcomes.

(Refer Slide Time: 15:12)



The Logistic Regression Model

The "logit" model solves these problems:

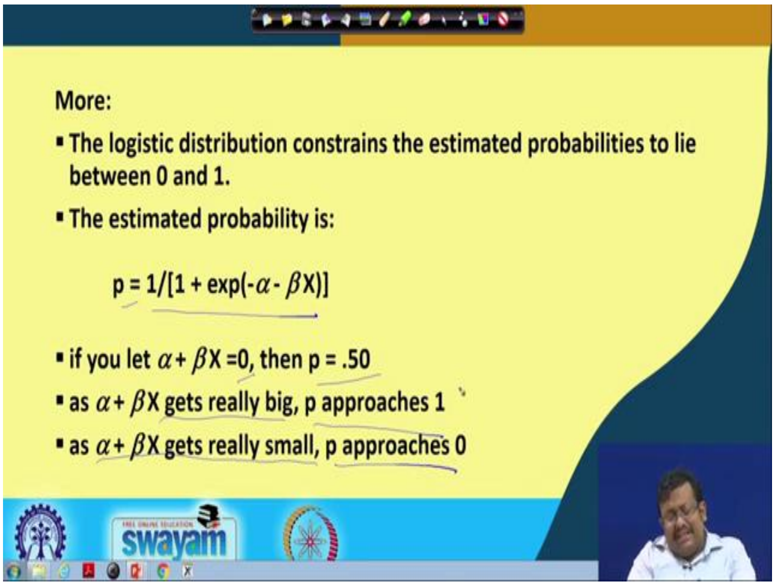
$\ln[p/(1-p)] = \alpha + \beta X + e$

- p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"
- ln[p/(1-p)] is the log odds ratio, or "logit"

So, what is the solution? The solution is to use logistic regression model. So, the logit model solves this problem where we use this form of logarithm natural logarithm of p by 1 minus p and we use this alpha plus beta X plus e. So, it is a linear regression formula and only in case of Y we are using this logarithm of p by 1 minus p, where p is the probability that event Y occurs. So, Y equal to 1, this probability is p and p by 1 minus p is the odds ratio or simply we call odd. And, natural logarithm of p by 1 minus p is the log odds ratio logs of odd ratio or logit. So, it is basically logit and this logit is basically modeled by linear regression form.

So, instead of targeting a particular you know target variable we are basically converting them into logit and we are predicting this logit using the simple linear regression or multiple linear regression. So, this is called the logistic regression model.

(Refer Slide Time: 16:35)



More:

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is:

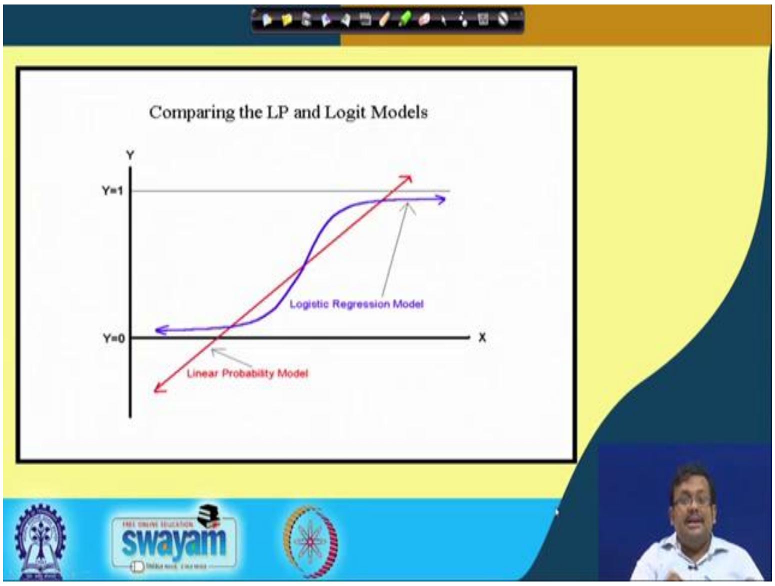
$$p = 1/[1 + \exp(-\alpha - \beta X)]$$

- if you let $\alpha + \beta X = 0$, then $p = .50$
- as $\alpha + \beta X$ gets really big, p approaches 1
- as $\alpha + \beta X$ gets really small, p approaches 0

Now, the logistic regression constrained the estimated probability to lie between 0 to 1. This is the solution because we see that in case of ordinary least squares the probability will go beyond 0 to 1. So, to cap them within 0 to 1, we need to apply this logistic regression. The estimated probability is, obviously, p equal to if we simplify it will be p equal to this. So, if you let alpha plus beta X equal to 0, then p equal to 0.50.

So, as alpha plus beta X get really big, p approaches to 1; as alpha plus beta X get really small p approaches to 0. So, you can see you are capping the lower limit and upper limit of the probability between 0 to 1. So, this is the solution of that problem where this probability goes beyond 0 to 1.

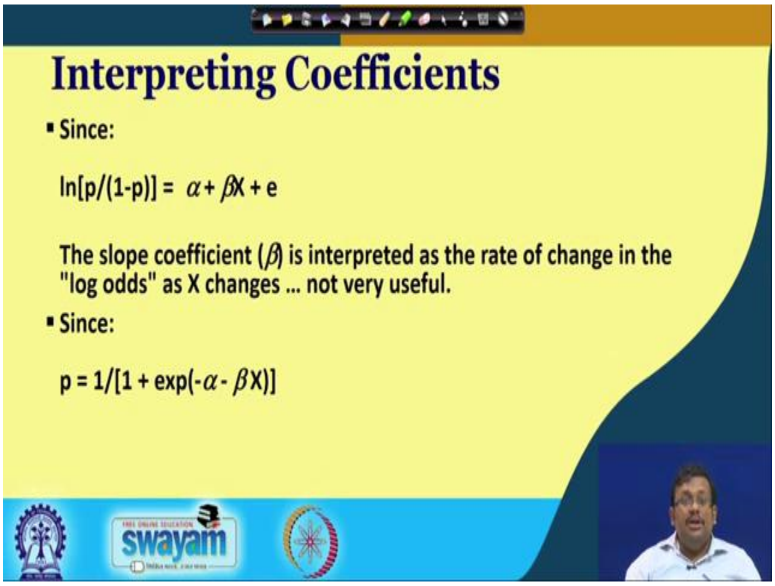
(Refer Slide Time: 17:34)



Comparing the LP and Logit Models

So, this can be shown very effectively by this next slide where you can see comparing the linear prediction and logit models. So, in case of linear prediction you can see there are two outcomes Y equal to 0 and Y equal to 1, these are two levels. So, in case of linear prediction model you can see there is some it will go beyond 0 to 1. However, when we are using logistic regression model by converting into logit and predicting that logit by simple linear regression ultimately you will see that logistic regression model will cap the probability between 1 to 0. So, that is why ultimately it is maintaining the assumption of linear regression.

(Refer Slide Time: 18:15)



Interpreting Coefficients

- Since:

ln[p/(1-p)] = $\alpha + \beta X + e$

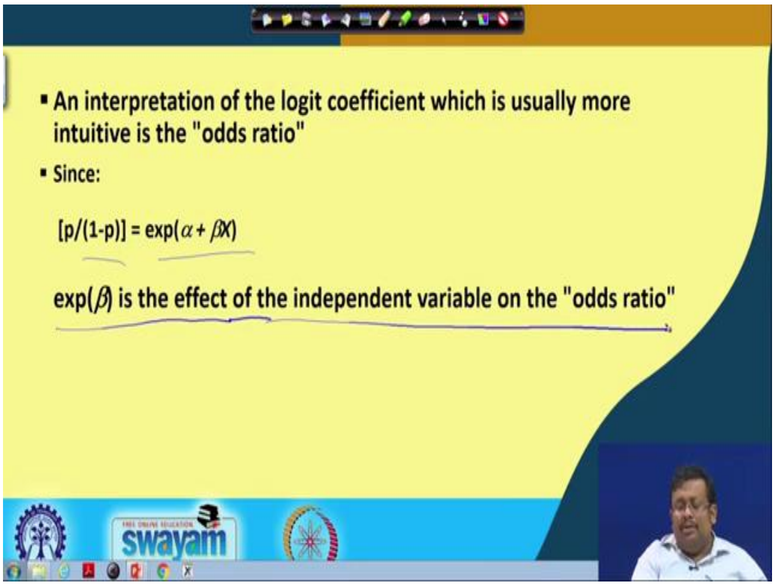
The slope coefficient (β) is interpreted as the rate of change in the "log odds" as X changes ... not very useful.

- Since:

p = 1/[1 + exp($-\alpha - \beta$X)]

So, so, interpreting the coefficient; obviously, log of p by 1 minus p alpha plus beta plus X e plus eta is error. So, the slope coefficient beta is interpreted as the rate of change in the logs odds as X changes which is not very useful. Since p equal to, you know, exponential p basically assume this form.

(Refer Slide Time: 18:36)



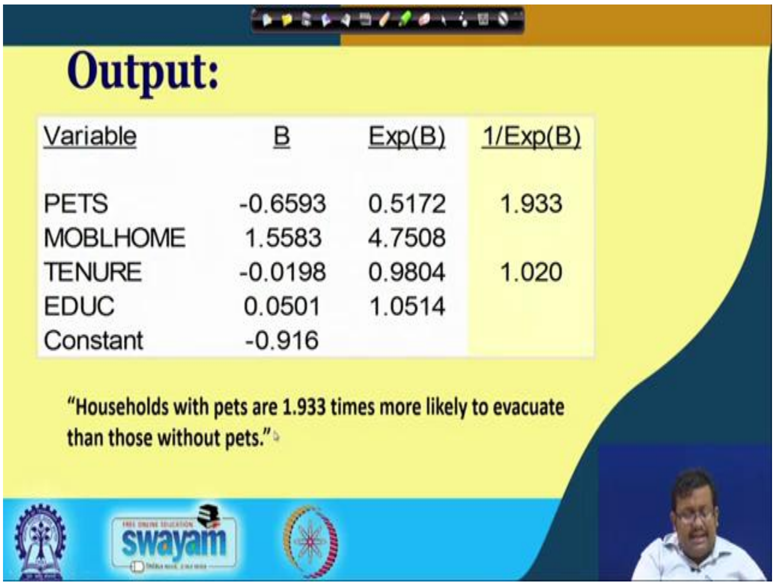
- An interpretation of the logit coefficient which is usually more intuitive is the "odds ratio"
- Since:

$[p/(1-p)] = \exp(\alpha + \beta X)$

$\exp(\beta)$ is the effect of the independent variable on the "odds ratio"

So, that is why an interpretation of the logit coefficient which is usually more intuitive is the odds ratio. So, since p takes the form of p by 1 minus p is equal to exponential of alpha plus beta X. So, exponential beta effect is the effect of the independent variable on the odds ratio.

(Refer Slide Time: 18:58)



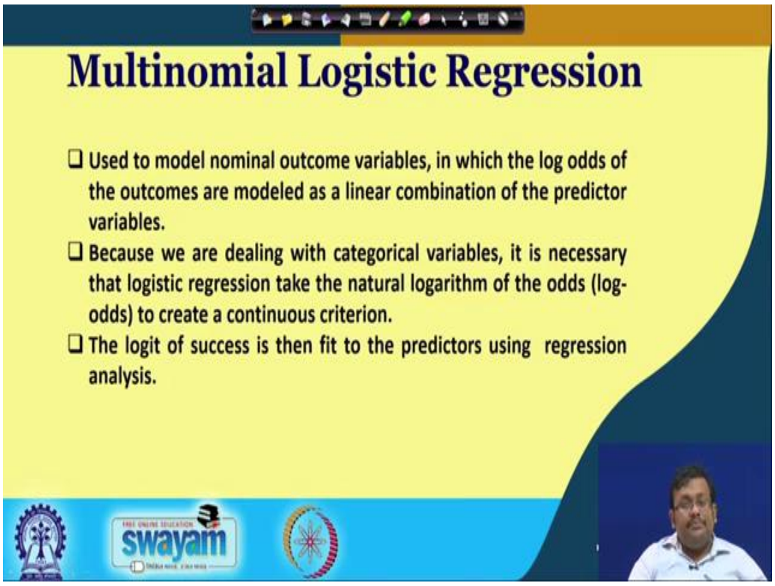
Output:

Variable	B	Exp(B)	1/Exp(B)
PETS	-0.6593	0.5172	1.933
MOBLHOME	1.5583	4.7508	
TENURE	-0.0198	0.9804	1.020
EDUC	0.0501	1.0514	
Constant	-0.916		

"Households with pets are 1.933 times more likely to evacuate than those without pets."

So, here you can see outputs. So, households with pets at you can see here 1.993 is ultimately the expected. So, the outputs with pets are 1.933 times more likely to evacuate than those without pets. So, this is ordinary logistic regression.

(Refer Slide Time: 19:17)

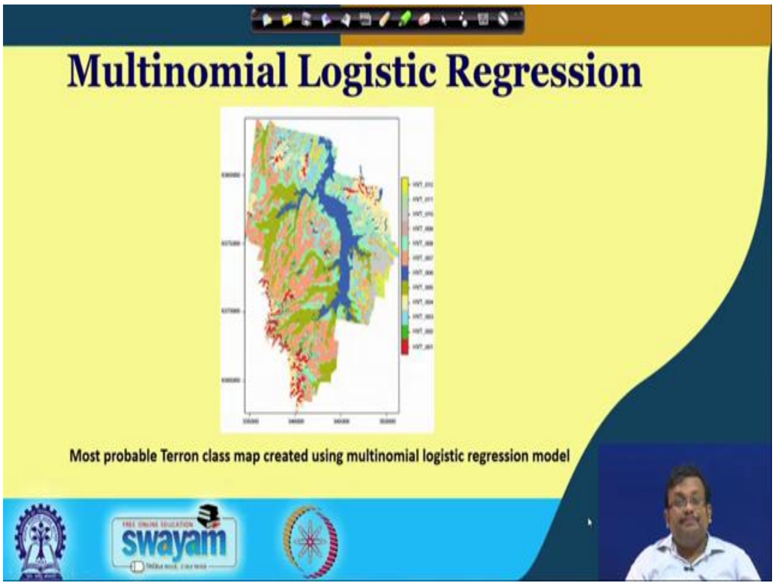


Multinomial Logistic Regression

- □ Used to model nominal outcome variables, in which the log odds of the outcomes are modeled as a linear combination of the predictor variables.
- □ Because we are dealing with categorical variables, it is necessary that logistic regression take the natural logarithm of the odds (log-odds) to create a continuous criterion.
- □ The logit of success is then fit to the predictors using regression analysis.

Now, multinomial logistic regression is basically, you know, is the same as logistic it is it is an extension of logistic regression. So, basically it is used to model nominal outcome variable in which the log odds of the outcomes are modeled as a linear combination of the predictor variables, in our case these are covariates. So, because we are dealing with categorical variables it is necessary that logistic regression takes the natural logarithm of the odds that is log odds to create a continuous criterion, we have already seen. And, the logit of success is then fit to the predictors using the regression analysis.

(Refer Slide Time: 19:56)

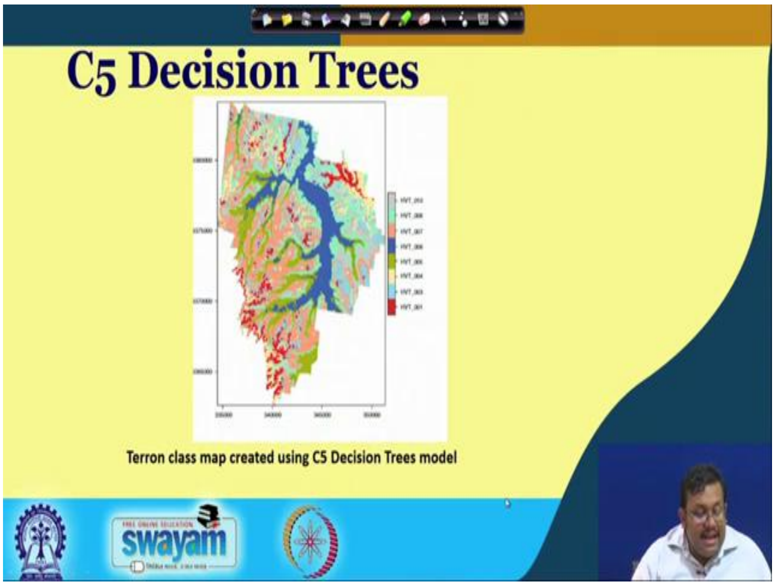


Multinomial Logistic Regression

Most probable Terron class map created using multinomial logistic regression model

So, these are example of multinomial logistic regression as you can see this is a most probable Terron class. Obviously, it is an area in Australia and they have you know they have classified the total area this area is called The Hunter Valley area and so, they have created this Terrons, you know, HVT 012, HVT 011. So, most probable Terron class they have classified based on this multiple logistic regression and this an example how multinomial logistic regression is used for predicting Terrons in digital soil mapping.

(Refer Slide Time: 20:38)



C5 Decision Trees

Terron class map created using C5 Decision Trees model

Another is C 5 decision tree. The C 5 decision tree is you can another type of, you know, categorical model and in this category it is basically based on the C 5 algorithms of Queenland and this also you can see Terron class map created using C 5 decision trees model. And, again random forest: random forest can be used for both regression and classification and here it is an example of classification. So, in this example, you know, you can see the Terron maps created using the random forest model.

So, guys we have completed this random forest, you know the categorical models and creation of categorical models using, using in DSM technique. Again, to wrap up, you know, categorical models are those where we are trying to predict some, you know, some categorical, you know, categorical model, you know, prediction models are those where we are trying to classify some categorical variables.

In case of digital soil mapping the categorical variables will be Terrons. Terrons are basically some classification which has some connotation with related landscape models and soil forming factors. And, this Terron can be classified based on several soil covariates as we know in case of continuous model we use some covariates to predict certain soil property and here also we can use those covariates or auxiliary variables to predict certain classes. And, these classes are, you know, these classes can be predicted by using several types of categorical models. And, this categorical models could be multinomial logistic model or C 5 decision tree model or random forest model.

And, again this multinomial logistic model basically assume that you convert the predictor variables into logit or log off odd ratios which will help, which will basically help to give you, to resolve the problem of finite outcomes in the target variables. And ultimately give it a linear regression forms and ultimately, you know, using this multinomial logistic regression you can predict several classes.

So, guys I hope that you have learned something new in this last couple of lectures of categorical models and continuous model. In case of continuous model we have discussed several important continuous model we started with multiple linear regression, then we discussed about then classification regression tree which is a non-linear model and then we talked about cubist, random forest and now, we have discuss this categorical models. And, all these are extensively used not only these there are some advanced other advanced model also which we do not have time to discuss right now; for example, artificial neural network and then ant colony optimization.

So, these all this advanced models which nowadays scientists are using for predicting certain soil properties through DSM and these are basically we use for this SCORPAN plus e for, you know, formula if you remember this function or soil special inference model. In case of the special inference model, we are using this mathematical models to predict certain soil property at a particular space.

So, guys let us wrap up here. And, in our last lecture we will be discussing about pedo transfer functions as well as some certain uncertainty measurements in association with digital soil mapping. And, thank you very much let us meet in our next and final lecture of Soil Science and Technology.

Thank you.