Marketing Research and Analysis -II (Application Oriented)

Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 50
Discriminant Analysis in SPSS

Welcome everyone to the class of marketing research and analysis, so as we have discussed in the last lecture about a new concept which comes under the you know, ambit of regression, which was called as logistic regression, so we learned that logistic regression is a technique which is used when your dependent variable is categorical nature although, the general assumption of regression is at both dependent and the predictor variable should both be continuous.

But we; this is a special case where we understood that sometimes we need to understand in life, whether I need to give permission to somebody to you know join my course or not or whether it will flood or it will not flood, whether it will rain or it will not rain, whether somebody will be a defaulter or not defaulter, so there are several you know occasions in life, should I give you know, should I marry or I should not marry.

So, in such conditions where the person is coming with a such a situation or such a condition, he needs to understand, these decisions must be based on certain statistical measures, so the independent variables, by using the independent variables, he tries to come to an outcome and predict whether he should do it or he should not do it or it will happen or it will not happen, so that was a case of logistic regression, right where we calculated the odds ratio and we measured.

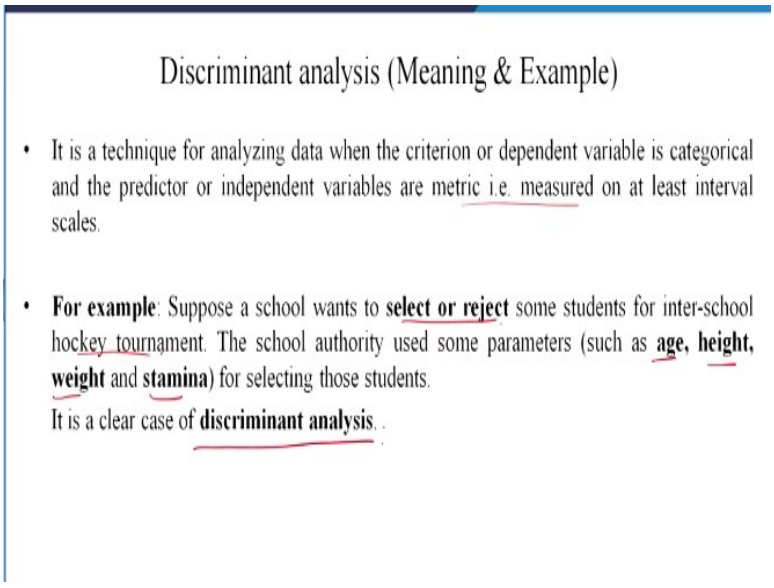
And then we find out, how this variables were impacting, the outcome variable, so we say and we also understood during that while in the last illustration that the amount of strength or you can say, the strength of the model, right through a classification results, we understood how strong are model was, how much correct early it was explaining the entire study, right, similarly today we will talk about another method which is very close to that logistic regression or has a similar purpose you can understand called the discriminant analysis.

As the name you understand from discriminant, so, right what it says the discriminant, to discriminate, right, to you know separate the one from the other, the good from the rest or something like that right, so discriminant analysis is the similar technique like logistic regression but the difference in the between the two is that the logistic regression was performing better when your data is not following the normal assumptions; the normality assumptions, right.

But if your data is following a normality assumption, right, all the independent variables are following the normality assumption, then in such a condition, then you have 2 options, right; the logistic and the discriminant, so the question is; should I use the logistic regression or should I not use; should I use the discriminant analysis, so please remember, if your data is following a normality right or the normal distribution, it is always better to use the discriminant analysis instead of the logistic regression.

But if your data is not following the normal distribution, then it is better to follow a use a logistic regression instead of the discriminant analysis, this is the basic understanding, right, so, let us start the lecture.

(Refer Slide Time: 03:43)



So, discriminant analysis is the technique for analysing data when the criterion or dependent variable is categorical and the predictor or independent variables are metric, right that is measured on at least interval scales. For example, a school wants to select or reject some students

for interschool hockey tournament, okay the school authority used parameters such as age, height, weight and stamina for selecting those students.

So, in this case selecting the students is a clear case where we are using discriminant analysis, right.

(Refer Slide Time: 04:20)

Now, this is what the meaning of discriminant is, you see now, the men have been separated from the woman or the girls and from the boys right, through some measures and this is what, this line is the mark of difference or the discriminant right, it is separating the 2, right.

(Refer Slide Time: 04:37)

Classification of Discriminant Analysis

- Discriminant analysis techniques are described by the number of categories possessed by the criterion variable. They are:
- **Two group discriminant analysis**: When the criterion variable has two categories , the technique is known as two group discriminant analysis.
- *Example*: Do user and non-user of social networking sites differ with respect to their age.
- **Multiple group discriminant analysis**: When the criterion variable has three or more categories, the technique is referred to as multiple discriminant analysis.
- *Example*: Do heavy, medium and light users of hard drink differ in terms of their weight.

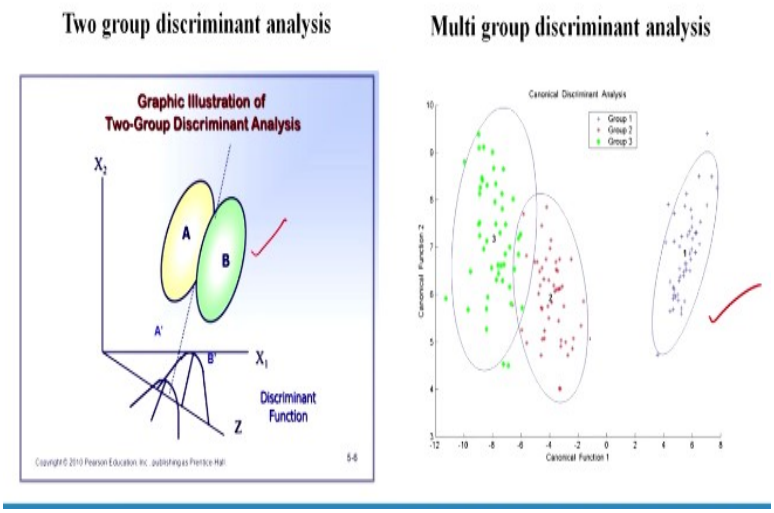
So, classification of discriminant analysis; discriminant analysis techniques are described by the number of categories possessed by the criterion variable, they are 2 group or multi group. So, 2 group discriminant analyses are when the criterion variable has only 2 categories, so this technique is known as 2 group's discriminant analysis. So, do user and non-user of social networking sites differ with respect to their age I want to know?

So, this is like a binomial logistic regression but the condition I have already said is that here the data the variables; independent variables need to follow a normal distribution, right, if they not do not follow a normal distribution, then in the same case do user and non- user of the social networking sites differ with respect to their age can be done through a logistic regression, okay. Multiple group discriminant analysis or comparatively, you had the multinomial logistic regression.

When the criterion variable has 3 or more categories those techniques is referred to as multiple discriminant analysis, right. Example; do heavy, medium and light users of hard drink, right differ in terms of their weight or maybe in terms of their heart you know, condition, right, so we want to check, so there are several such studies in real life that you want to know, right, somebody who is a vegetarian or a non-vegetarian, does it impact their stamina, does it impact their weight, right, you want to see does.

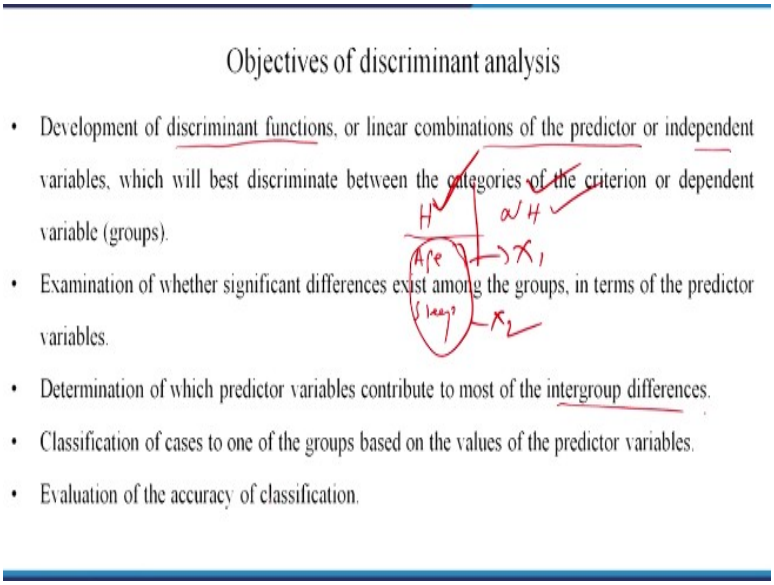
Or you can say is the weight of a person affected by the vegetarian or the type of food but here the case is slightly different, it is the reversal, right, here are independent variable is rather a you know categorical and the dependent variable weight is a continues but just assume the inverse of it where the dependent variable is categorical and the independent variable is continuous, so here heavy medium and light users of hard drink; do they differ in terms of their weight, right?

(Refer Slide Time: 06:42)



So, can we find out some relationship that is where we use a discriminant analysis, so this is the class of a two group discriminant analysis and this is a three group discriminant analysis, multi group or this is the; here this case is the three group discriminant analysis, right.

(Refer Slide Time: 06:51)



So, let us see what we exactly do in this case; the development of discriminant functions, the objective of a discriminant analysis is to develop discriminant functions, right or linear combinations of the predictor or independent variables which will best discriminate the categories or the criterion/dependent variables. Now, what you understand; it says that a linear combination of the independent variables is done in such a way that it best discriminates between the groups in the criterion variable, right.

So, in the criterion variable, you had let us say, heavy and non-heavy, let us say non-heavy, so if I take the you know independent variables like age, let us say you know, sleeping habit, right etc., then how are they helping; how are these helping to differentiate between heavy and non-heavy, so can I create a equation; a discriminant function is like a equation which will help me by putting on the values of X_1 and X_2, right, to predict whether he will be a heavy person or a non-heavy person, right in weight.

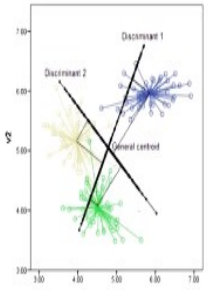
So, examination of whether discriminant significant differences exist among the groups in terms of the predictor variables, right, determination of which predictor variables; for example, in logistic regression, you have seen exponential beta right, the higher the exponential beta, we could say, higher is a contribution of that particular variable to the dependent variable, so if you remember in the last case where we have taken monthly salary and I think it was gender, right.

So, we had seen that monthly salary at a higher exponential beta, so that explained more in comparison to the gender, okay to explain the effect on the dependent variable, so determination of which a predictor variable contributes to most of the inter group differences, classification of cases to one of the groups based on the values of the predictor variables and the another objective is to evaluate the accuracy of classification.

(Refer Slide Time: 08:57)

Discriminant function

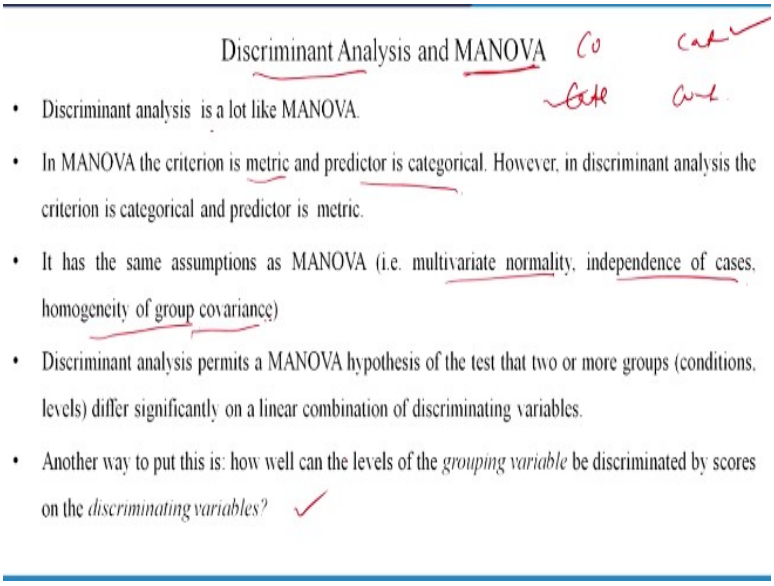
- Discriminant function is the linear combination of independent variables developed by discriminant analysis that will best discriminate between the categories of the dependent variable.
- In the two-group case it is possible to derive only one discriminant function, but in multiple discriminant analysis more than one function may be computed.



Geometric interpretation of two group discriminant analysis

This was same like in the logistic regression also, right, so discriminant function is the linear combination of independent variables developed by discriminant analysis that will best discriminates between the categories; the 2 or 3 categories or 4 categories whatever you have but you should not have too many categories also, then it will create confusion. In the 2 group case, it is possible to derive only one discriminant function.

(Refer Slide Time: 09:21)



But in multiple discriminant analysis, more than one function may be computed okay, very interestingly as I had given you an example and I said it is very, it looks very similar to ANOVA, right, actually there is a relationship, discriminant analysis and MANOVA which has got multiple independent variables, right and each independent variables has got several factors, right.

So, discriminant analysis actually is a reverse, is this called as the mirror image of the MANOVA that is very interesting, so discriminant analysis is a mirror image of the MANOVA, right, why? Because in the MANOVA, you had the dependent variable which was continuous and the independent variables were categorical, right, so you had more than one independent variables.

But here interestingly, you have a categorical; one categorical dependent variable and then you have some independent variables which are continuous, so that is why it is a mirror image, right, so you see the MANOVA, the criterion is metric and the predictor is categorical right, however in discriminant, the criterion is categorical and the predictor is metric, right. It is the same

assumptions as MANOVA, so multivariate normality, independence of observations or cases, homogeneity of group variance or co-variance, so that this case co-variance.

So, right, all these are similar right, discriminant analysis permits a MANOVA hypothesis or the test that 2 or more groups differ significantly on a linear combination of the discriminant variables, right. What we say is that 2 or more groups differ significantly on a linear combination of the discriminant or the independent variables that you have taken, right, another way to understand is how well can the levels of the grouping variable be discriminated by scores on the discriminating variables?

So, how are my predictor variables, how good will they explain the differences in the criterion variable of the dependent variable, okay, so look at this interesting table.

(Refer Slide Time: 11:25)

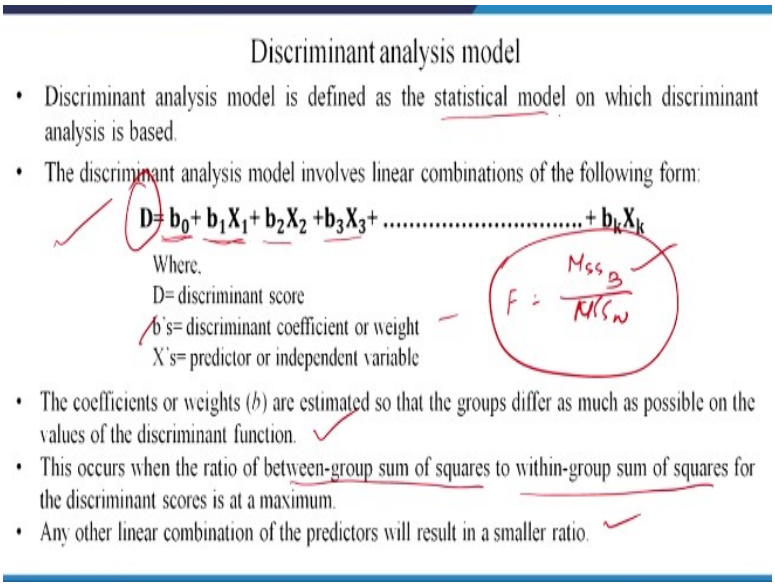
Relationships between ANOVA, Regression and Discriminant analysis

Relationships	ANOVA	Regression	Discriminant
Similarities			
Number of dependent variable	One	One	One
Number of independent variable	Multiple	Multiple	Multiple
Differences			
Nature of the dependent variable	Metric	Metric	Categorical
Nature of the independent variable	Categorical	Metric	Metric

So, the similarities between ANOVA, regression and discriminant; number of dependent variable in ANOVA, you had 1, right, in regression, you had; you also have 1 and discriminant also, you have 1, number of independent variable; multiple, multiple, multiple. Differences; nature of the dependent variable; metric in ANOVA, metric; continuous or metric right, in regression but here categorical.

Nature of the independent variable; categorical in ANOVA, metric in regression and in discriminant metric, right, so this is the basically the relationship between and the difference, right,

(Refer Slide Time: 12:03)



Now, how does it look like; the discriminant analysis model is defined as the statistical model on which discriminant analysis is based, right, so this is how the discriminant function looks like, the discriminant score where $D = b_0 + b_1 X_1 + b_2 X_{2+.....+} b_k X_k$

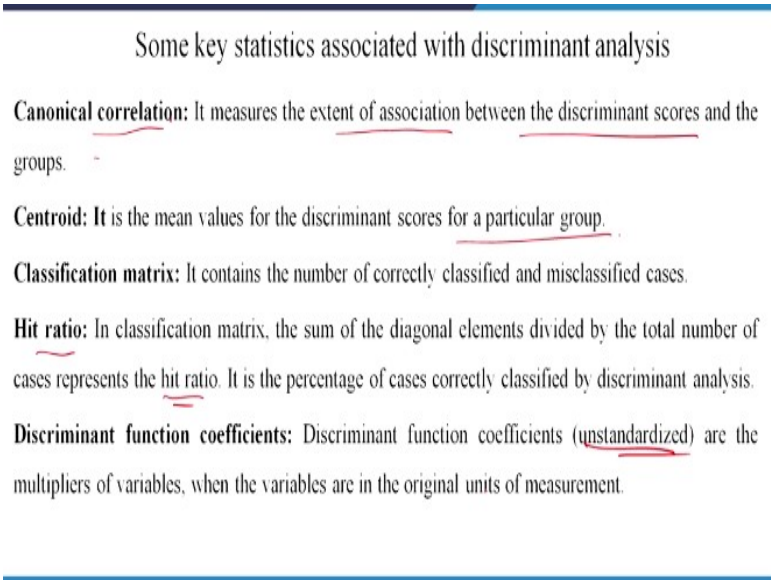
So, what are these b's let us say, this b's; b1, b2, b2, they all my coefficients, so this is my intercepts, these are my coefficients; b1, b2, b3 and X is my predictor variables or the independent variables, the coefficient or weights are estimated, so the coefficients will only help me to as in the regression we were doing, this coefficients only will explain me or have an impact on the value of D, correct, higher the value of b, the larger the impact on D.

So that the group differences differs; groups differ as much as possible on the values of the discriminant function, right, this occurs when the ratio of between group, sum of squares to within groups sum of squares for the discriminant scores at a maximum now, what does it mean?

That if you remember, in ANOVA, what was an F ratio = $MSSb/MSSw$

So, it says similarly, the ratio of between groups sum of squares to within group sum of squares for the discriminant scores should be maximum, right, so any other linear combination of the predictors will result in a smaller ratio, so this is; if this value should be maximum then only it is the best combination, okay.

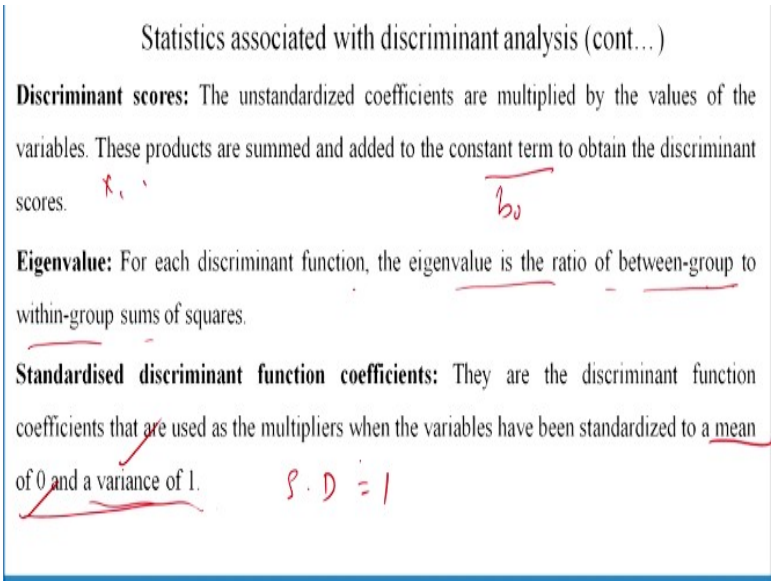
(Refer Slide Time: 13:55)



Some key statistics associated with discriminant analysis; first is canonical correlation; you will see when I will use the spaces also, this I will show you, it measures the extent of association between the discriminant score and the groups, right. Centroid; it is the mean values for the discriminant scores for a particular group, right. Classification matrix; it contains a number of correctly classified and misclassified cases. Hit ratio; in the classification matrix, when I will show you the sum of the diagonal elements divided by the total number of cases represents the hit ratio.

And it is a percentage of cases correctly classified, I will show you when I will show you the table but understand this terms are very important, the discriminant function coefficients are the; these are the unstandardized coefficients; are the multiplier of variables when the variables are in the original units of measurement, right. So, these are some of the terms that are associated.

(Refer Slide Time: 14:52)

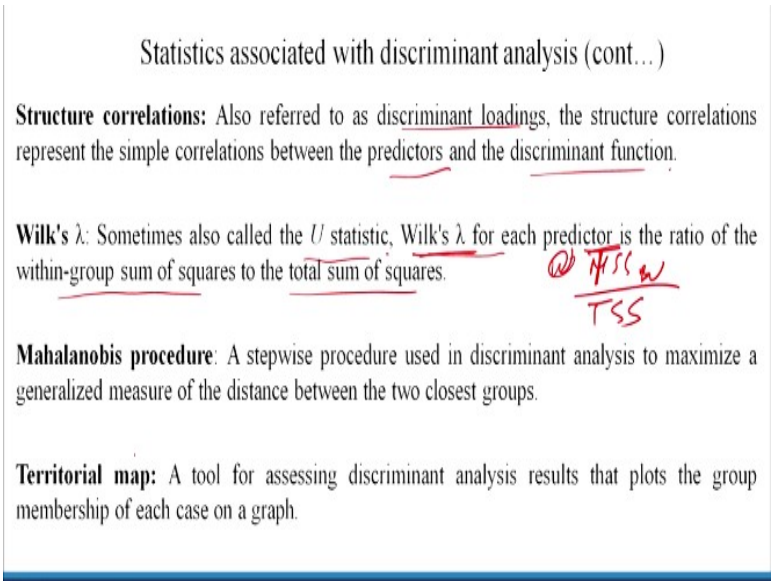


So, discriminant score; unstandardized coefficients are multiplied by the values of the variables, variables means they are X_1, X_2 whatever, this products are summed and added to the constant term which is the b_0 or the intercept right, a to obtain the discriminant score, so it is exactly like a you calculate the regression equation and the score, right, eigenvalue. For each discriminant function, the eigenvalue is the ratio of the between group to the within groups sum of squares, right.

So, higher the eigenvalue more is explanation, right, better it is. Standardised discriminant function coefficient; they are the discriminant function coefficients that are measured or that are used as the multipliers when the variables have been standardised to a mean of 0 and a variance of 1, so this means what; you have standardised it, so whenever you standardised a variable, so then only you can compare 2 different variables at the same level why?

Because when you have standardised all the value lies between 0 and 1 and the mean has a 0 and a standard deviation is actually 1, SD = 1, so SD is 1 means variance is 1, right but the mean is 0, so this is how you know the standardised value is taken and it helps for comparison, right.

(Refer Slide Time: 16:02)

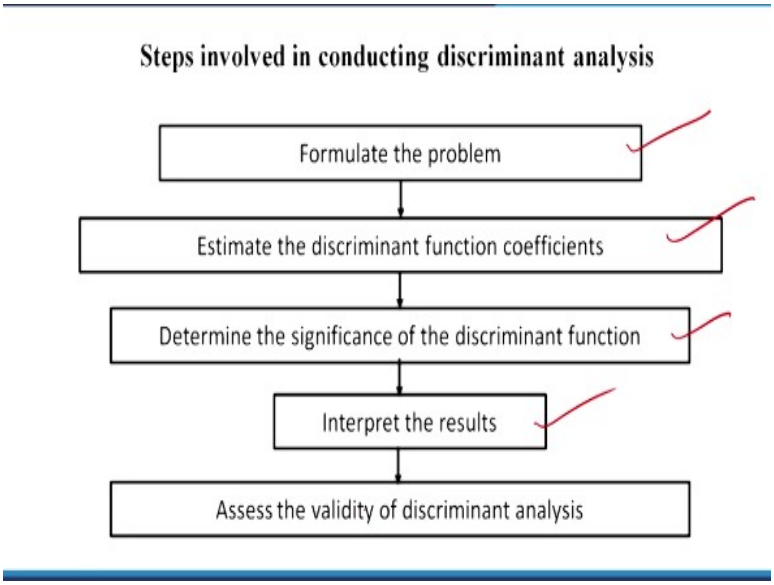


Some other terms are like for example, structure correlation is referred to as the discriminant loadings, so this term loadings you will understand later on also, the this is the correlation between the; between 2 you know variables, the structure correlation represent the simple correlation between the predictor and the discriminant function, right, so and then another term called Wilk's lambda (λ), sometimes called as the U statistic.

Wilk's lambda (λ) for each predictor is the ratio of the within group sum of squares to the total sum of squares not between, what it is saying; the within sum of square, so mean sum of square within or total sum of square within let us say total sum of square within divided by the total sum of square, so this Wilk's lambda (λ) is actually helping to measure the statistical significance of the model, okay.

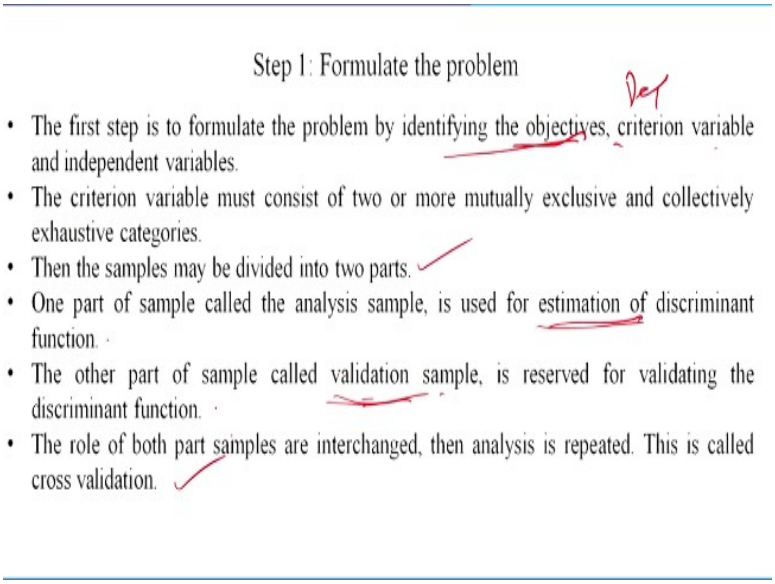
Mahalanobis procedure; a stepwise procedure used in discriminant analysis to maximise the distance between the 2 closest groups, Mahalanobis distance, you were also using in regression also to find outliers, if you remember right, so it helps to measure the distance between the 2 closest groups, right. Territorial map; a tool for assessing discriminant analysis results that plots the group membership of each case on a graph, which I generally do not use it but you can if you want to plot in a graph and see so that is through the territorial map, okay.

(Refer Slide Time: 17:31)



So, what is the step involved? First you formulate the problem, estimate the different discriminant function coefficients, determine the significance, interpret the results, okay and assess the validity.

(Refer Slide Time: 17:45)

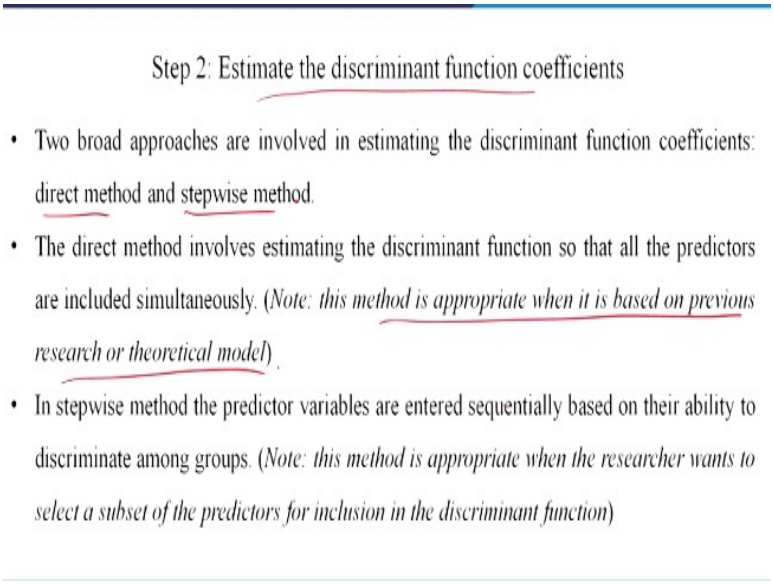


Now, step 1; how do you do that so, the first step is to formulate the problem by identifying the objectives, criterion variable and independent variables, so what is your dependent, what is your variable, what is your independent variable and what is your objective of the study, right, first you need to formulate. So, the criterion variable must consists of 2 or more mutually exclusive and collectively exhaustive categories.

So, there should be 2 categories which are mutually exclusive, very different and collectively exhaustive means, they all the respondents are lying in this two categories only, right, the sample may be divided into 2 parts, right, one part called the analysis sample is used for estimation of the discriminant function, the other part called the validation sample is used for validating the discriminant function.

So, one part is used for estimation, the other part is used for this validation, so if you have let us say 200, so you can divided into 2 parts of 100, 100 each, one to estimate the discriminant function and the other is to validate the discriminant function, right. The role of both part samples are interchanged, then analysis is repeated, this is called cross validation which you will see later on, right.

(Refer Slide Time: 18:52)

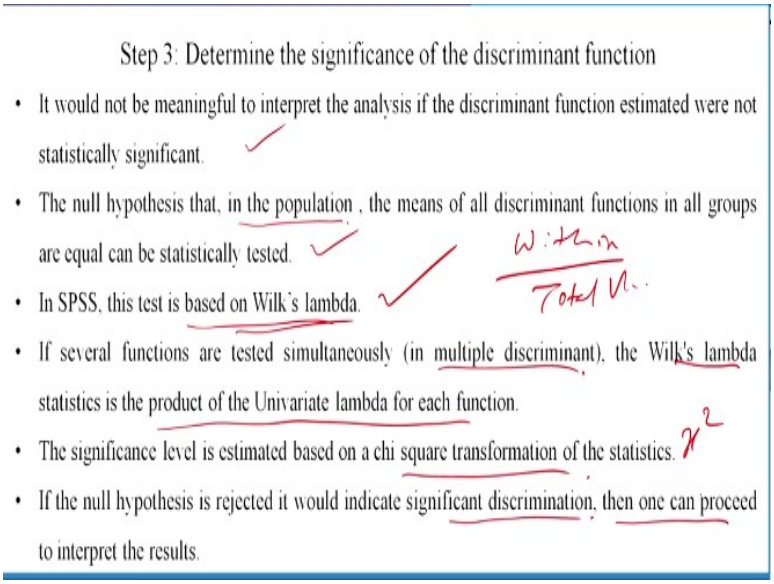


Second; how do I estimate the discriminant function coefficients; two broad approaches are there, right, one is the direct method and the other is the stepwise method, by now I think you have understood what is the stepwise method, in regression also, I will explained. The direct method involves estimating the discriminant function so that all the predictors are included simultaneously enter and in regression, you were using the method enter, okay.

This method is appropriate when it is based on previous research or a theoretical model, okay. In stepwise method, the predictor variables are entered sequentially based on the ability to

discriminate among the group, so you may use a forward method or something like in stepwise in the regression we are using step forward regression, backward regression, similar right, then determine the significance of the discriminant function.

(Refer Slide Time: 19:40)

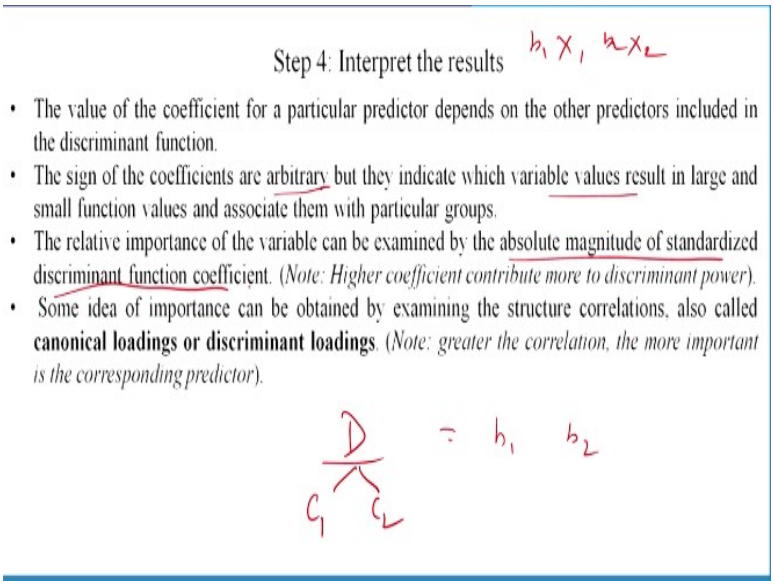


It would not be meaningful to interpret the analysis of the discriminant function estimated were if they were not statistically significant, so null hypothesis in the population, the means of all discriminant functions in all groups are equal and can be statistically tested, this test is based on the Wilk's lambda which I already said, what it said the within the group variance divided by the total variance, right.

If several functions are tested simultaneously like in multiple discriminant, the Wilk's λ is a product of the univariate lambda for each function, so that you need not worry, if you are doing it by hand it is a separate thing but you will be do not do it by hand, we do it by SPSS, so it gives us or any software it helps us, the significance level is estimated based on a chi-square transformation, so this is something which is which will be provided the chi-square value which helps in predicting the overall model significance, right.

If the null hypothesis is rejected, it would indicate significant discrimination then one can proceed to interpret the results.

(Refer Slide Time: 20:41)



Let us see this at the fourth step is to interpret the results, right, so the value of the coefficient right for a particular predictor variable, so whatever the coefficients you have X_1, X_2, so this b_1, b_2, right, we are talking about this, right, so the value of the coefficient for a particular predictor variable predicted depends on the other predictors included in the discriminant function, so they are somewhere related, okay.

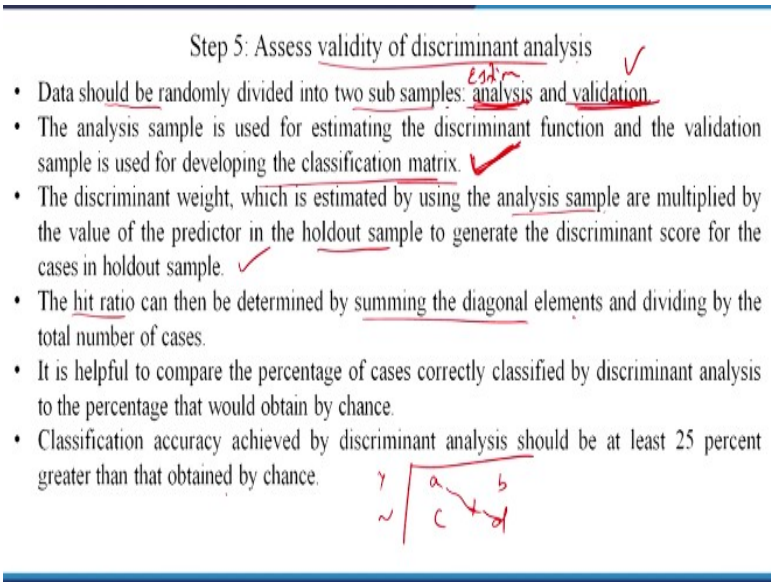
The sign of the coefficients are arbitrary but they indicate which variable results values result in large and small function values and associated them with the particular groups, now what does it mean; that means in simple terms, you need to understand that the slope values, right with the you know, the variables together we help in discriminating, we will give you the discriminant score right.

So, this discriminant score which will come, this will finally help you to discriminate whether this will come into the category 1 or category 2, right, this is what it means. The relative importance of the variable can be examined by the absolute magnitude of a standardised discriminant function coefficient, higher coefficient contribute more to the discriminating power; discriminant power.

That means a higher coefficient is better in explaining that these 2 are clearly different right, if the coefficient values is less, smaller coefficient is there, then you cannot exactly say that there is

a clear-cut difference between the two groups, right, some idea of importance can be obtained by examining the structure correlations also called as canonical loadings or discriminant loadings.

(Refer Slide Time: 22:24)



Now, remember greater the correlation, the more important is the corresponding predictor, right, so let us see and finally we will come to the validity of the discriminant analysis. So, to check whether the study is valid or not what you should do is; the data at; divide the data into 2 sub samples; one, analysis and the other is for validation, right, so this will help you in estimation, right and this one for validation.

So, the analysis is used for estimating the discriminant function and the validation sample is used for developing the classification matrix, now I will show you in regression logistic also, you had seen classification matrix which was helping you to calculate the hit ratio which was telling what percentage of the values were predicted correctly and thus how strong the model was. For example, in the last case, if I remember a predictor or model, they had classified 91.2 percent of the cases correctly in logistic regression if you remember.

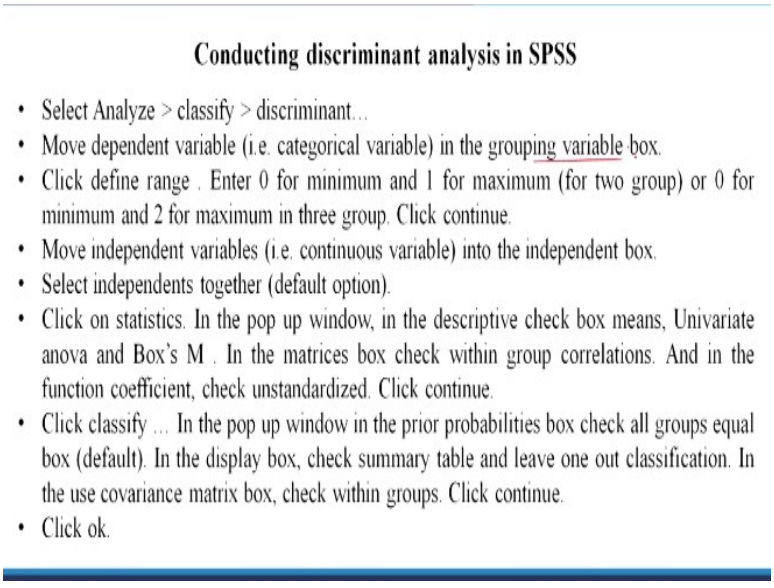
So, here that is what it helps you to find out the classification matrix which tells you whether the model is sufficiently explaining the whole process or not, the discriminant weight which is estimated by using the analysis sample right, so analysis sample are multiplied by the value of

the predictor in the holdout samples, so this is the analysis, this is the validation or the holdout sample to generate the discriminant score for the cases in the holdout sample, right.

So, one you are using for validation, the other you are using for estimation, if you do not do this also, it is not a very great deal, you can still you know, get to know but the point is if you validate it, it becomes more rigour, right. What is this hit ratio? Hit ratio is determined by summing the diagonal elements and then dividing by the total number of cases.

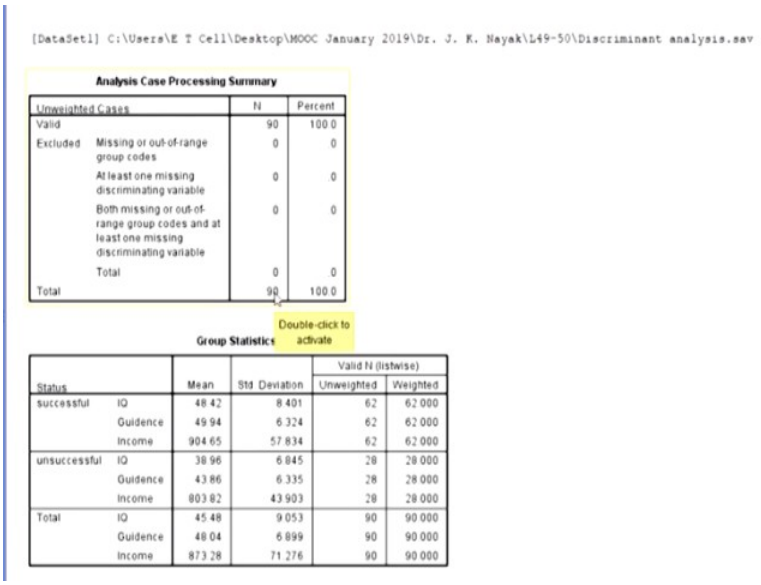
For example, so let us say a, b, c, d right, so this was a case of yes, no, right and whatever was, so this plus this, right, this plus this divided by the total, right, the total number of cases, it is helpful to compare the percentage of cases correctly classified by discriminant analysis to the percentage that would obtained by chance. So, what it says is classification accuracy achieved by discriminant analysis should be at least 25% greater than that obtained by chance, it should be at least 25% greater.

(Refer Slide Time: 24:49)



Now, how do you conduct the discriminant analysis? So, this example; move the dependent variable in the grouping variable, so let us go to the slide straight away.

(Refer Slide Time: 25:01)



[DataSet1] C:\Users\E T Cell\Desktop\MOOC January 2019\Dr. J. K. Nayak\L49-50\Discriminant analysis.sav

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		90	100.0
Excluded	Missing or out-of-range group codes	0	0
	At least one missing discriminating variable	0	0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	0
	Total	0	0
Total		90	100.0

Double-click to activate

Group Statistics

Status		Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
successful	IQ	48.42	8.401	62	62.000
	Guidence	49.94	6.324	62	62.000
	Income	904.65	57.834	62	62.000
unsuccessful	IQ	38.96	6.845	28	28.000
	Guidence	43.86	6.335	28	28.000
	Income	803.82	43.903	28	28.000
Total	IQ	45.48	9.053	90	90.000
	Guidence	48.04	6.899	90	90.000
	Income	873.28	71.276	90	90.000

So, this is the model were I will show you, so there are two cases; variable view, if you go to the variable view, so status of people were successful and unsuccessful, right and what is this; the level of IQ they have, the level of guidance given to them and the income of this people, right, the family income or something. Now, we want to seek whether the IQ, guidance and income, does it predict the status of a person, whether he will be successful or not successful, right?

So, we are interested there are 3 independent variables and one dependent variable which you can see, right, so there are only 2 cases; 0 and 1, so 0 is your case of successfu, 1 is unsuccessful, okay. Now, I want to run a discriminant equation, discriminant you know analysis. So, first you can check for normality of the data and if it is following normality, then you will automatically go for discriminant analysis.

So, how do I go; go to classify because it is the classification technique, so discriminant, right, what is my grouping variable; status, so what is the range; minimum is 0, so minimum is 0 and the maximum is 1, okay continue. Now, what are independent variables; I am taking all the 3 guidance, sorry, okay now statistics. What do I need; I need the means, I think if you go by if you go the; you know ppt also it is everything is mentioned here.

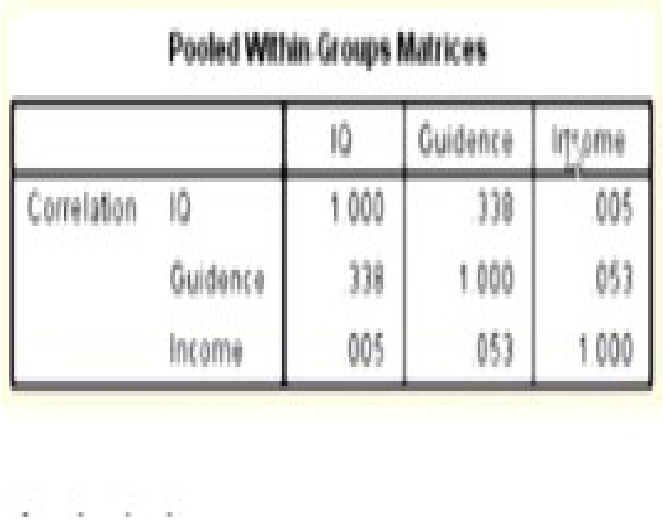
So, for example you see, move the independent variables, select the independent variables, click on statistics, go to the box M, right, so everything is given here, right, so let me show again, let

me go back, okay, so I need the box M, right and I need the unstandardized weights, right and the within groups correlation, so I need this, if you are interested to save, you can predict the group membership which group does it come to; the successful group or unsuccessful group, right.

And the probabilities, right, so this is all we need and I go to okay, so look at it if I; if what I am getting is; I have got N is 90 and all there is no missing case, right. Now, if you look at this group statistics, so successful people and unsuccessful people, their IQ level, the mean and standard deviation for IQ, guidance and income for successful people and unsuccessful is given to you, right and as you can see the IQ for successful people is more than the unsuccessful.

The guidance is more again and income is also more, right, so can we now say which one out of this may be all are impacting positively but how much or which one is the most important one right, let us say.

(Refer Slide Time: 27:55)

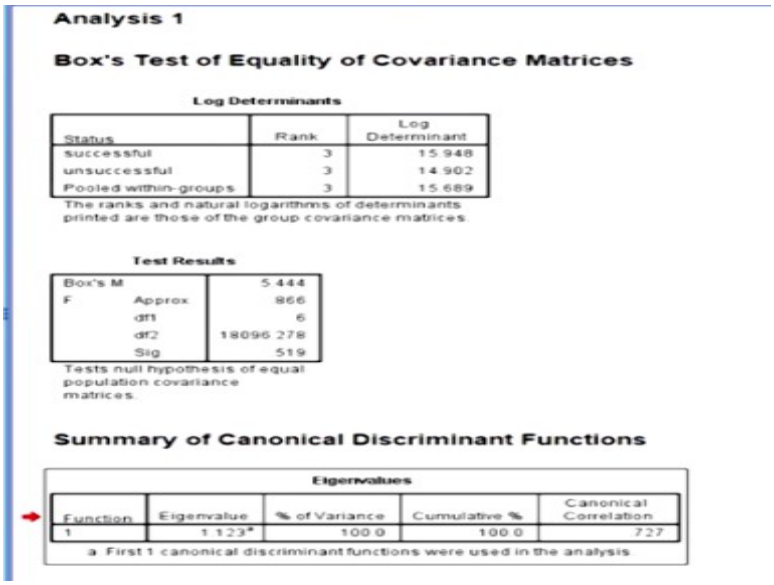


Pooled Within Groups Matrices

		IQ	Guidence	Income
Correlation	IQ	1.000	.338	.005
	Guidence	.338	1.000	.053
	Income	.005	.053	1.000

Now, if you look at the correlation, right, so correlation between IQ and guidance is .338, IQ and income .005, similarly between guidance and income is .053, so the strongest correlation we find is between IQ and guidance, okay.

(Refer Slide Time: 28:15)



Analysis 1

Box's Test of Equality of Covariance Matrices

Log Determinants

Status	Rank	Log Determinant
successful	3	15.948
unsuccessful	3	14.902
Pooled within-groups	3	15.689

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		5.444
F	Approx.	.866
	df1	6
	df2	18096.278
	Sig.	.519

Tests null hypothesis of equal population covariance matrices.

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.123[a]	100.0	100.0	.727

a. First 1 canonical discriminant functions were used in the analysis.

Now, let us go down, so if you see this box M test, now this is, this test the null hypothesis of equal population covariance, so the box M test should be not significant right, that means the value of the box M test, the significant value should be above .05 that means what; the null hypothesis says there is no difference in the group variances and the population or the covariance matrices; population covariance matrices.

And this is coming true, you cannot reject this null hypothesis and this is what is required, right, so this should always come above .05, right now, look at this canonical correlation value, so this value if you see this is exactly similar to or very much similar to the R square value, so if I take this value, this is similar to the R and if I take the R square, so that means what; .727 square is equal to something around 50 something will come, right.

So, this is my R square of the explained variance in the model, right when I am having, the 1 criterion and 3 predictor variables which we had in this model, right.

(Refer Slide Time: 29:29)

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig
1	.471	65.125	3	.000

Standardized Canonical Discriminant Function Coefficients

	Function
	1
IQ	.442
Guidence	.231

Now, the Wilk's lambda which is a test of significance is significance is significant this means what; it tells that the overall model is significant okay, now let us go to the file, so this values I have put it there, I will show you.

(Refer Slide Time: 29:44)

Problem

- Suppose a researcher wants to see the influence of **IQ** , **guidance** and **income** on the **result status (success or failure)** of a students sitting for a competitive exam.
- IV's are IQ , guidance and income.
- DV is result status i.e. success or failure.
- It is a clear case of **discriminant analysis**.

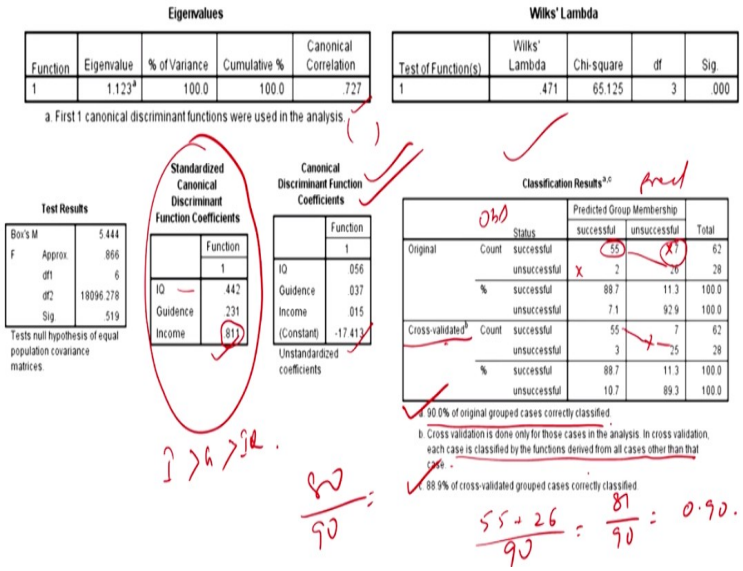
So, suppose let us go back to the problem, a researcher wants to see the influence of IQ, guidance and income on the results status, success or failure, right of students. IV's are IQ given; DV is given, success or failure, okay.

(Refer Slide Time: 29:58)

Reporting the SPSS result

- "Discriminant analysis was used to conduct a multivariate analysis of variance test of the hypothesis that successful and unsuccessful student would differ significantly on a linear combination of three variables i.e. IQ, guidance and income.
- The overall Chi-square test was significant (Wilk's λ = .471, Chi-square = 65.125, df = 3, Canonical correlation = .727, p < .001). The overall model was significant {see. Table: Wilk's λ and Eigenvalues}
- The discriminant functions extracted accounted for nearly (**52.85 %**) of the variance in student result status, confirming the hypothesis. {see. Table: Eigenvalues}
- Among the independent variable, the income is the best predictor. {see. Table: standardized canonical discriminant function coefficient}
- D= (-17.413) + 0.056 IQ + 0.037 (Guidance) + 0.015 (Income) {see. Table: canonical discriminant function coefficient}

(Refer Slide Time: 30:02)



Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.123[a]	100.0	100.0	.727

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.471	65.125	3	.000

Test Results

Box's M		5.444
F	Approx.	.866
	df1	6
	df2	18096.278
	Sig.	.519

Tests null hypothesis of equal population covariance matrices.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
IQ	.442
Guidence	.231
Income	.813

Canonical Discriminant Function Coefficients

	Function
	1
IQ	.056
Guidence	.037
Income	.015
(Constant)	-17.413

Unstandardized coefficients

Classification Results[a,c]

		Status	Predicted Group Membership: successful	Predicted Group Membership: unsuccessful	Total
Original	Count	successful	55	7	62
		unsuccessful	2	26	28
	%	successful	88.7	11.3	100.0
		unsuccessful	7.1	92.9	100.0
Cross-validated[b]	Count	successful	55	7	62
		unsuccessful	3	25	28
	%	successful	88.7	11.3	100.0
		unsuccessful	10.7	89.3	100.0

a. 90.0% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 88.9% of cross-validated grouped cases correctly classified.

Now, how do I report? So, discriminant analysis was used to conduct; let me show you the first this value, right, so eigenvalues is .727, so if I take this and square it, I am getting this value, right, 52.85, so this discriminant functions extracted accounted for nearly 53% of the variance in the student result status confirming the hypothesis, right. Now, if you look at this Wilk's lambda, it is saying it is significant, right.

So, here you see the overall chi square test was significant, Wilk's lambda is .471, chi-square is 65.125, degree of freedom is 3 and this is what the canonical correlation and significant .001, so the overall model was significant, right, so the overall model was significant, this value you have

to mention. Now, the third table was this box M test whether I mention it here or not about the box M, let me see?

If it is not mention also, you need to mention, right, so from the box M test, we found that the null hypothesis that the difference among the groups, right, there is no difference among the groups exist was found to be true that is also you can report, okay. Now, coming to this standardised you know, canonical discriminant function coefficients now, if you look at this, right, this 2 table, this tells you the importance of each predictor variables to the criterion variable.

Now, this has a different role, I will explain this, so what it is saying; income has .811, so income effects the success or a failure rate highest, the highest impact is of income, right, so next is IQ, IQ has a .442, right and third followed by guidance, so it is income, then guidance and then, right, income, guidance and IQ, okay. Now, what is this table saying, you will get this table, if you see, if you back to the output file.

(Refer Slide Time: 32:12)

Structure Matrix

	Function
	1
Income	826
IQ	525
Guidence	424

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function

Canonical Discriminant Function Coefficients

	Function
	1
IQ	056
Guidence	037
Income	015

So, if you see this, so you have seen this table right now, right and we are talking about now this table, so this table is the table which I am requiring again here, so this table says, now I want to find a discriminant score, right so, my equation will be made on this, so you see, so D =

-17.413+.056 IQ +.037 guidance +.015 income, so this is what the discriminant score you can this is how you calculate the discriminant score, right.

So, by putting on the actual values of IQ, guidance and income now you can predict the discriminant score and accordingly you can see the cut of value is always .5, so .5 it is let us say it is above .5, then it is move towards 1, if it is below .5, it moves towards 0 because there can be only 2 value, right; 0 and 1. Now, look at this classification result, so this is the one which tells you the hit ratio.

Now, what you do here; now you see 55 cases originally they have measured, the status is successful unsuccessful, this is observed, right, this is observed and this is predicted, so observed successful was let us say which was successful and they were actually predicted also were 55 cases, 7 people who are observed to be successful but you are predicted that there will not be successful is 7, right, so to this an error.

You had predicted, you had seen that; you have seen that there are unsuccessful but you have predicted that there would be successful, so again 2 cases which are wrong, it is like a type 2 error, okay and this is where you have unsuccessful and this is actually unsuccessful, so it is again correct, right, so now you see $\frac{55+26}{90} = \frac{81}{90}$ how much; 90%, right, so that is what it says here, 90 percent of original cases were correctly classified, right, the cross validation is done only for those cases in the analysis, right only for those cases in analysis. In cross validation, each case is classified by the functions derived from all cases other than that case now, let us see in this case what is happening; 55 cross validation, $\frac{55+25}{90} = \frac{80}{90}$, so 88.9% of the cross validated group cases correctly classified, so what basically it does; you need not bother about it, when you do this you know discriminant analysis on the software, it will help you to get this tables and once you understand this tables, you can even calculate the discriminant score and you can predict whether the model is a sufficiently explaining or it is not explaining, right.

And thus after that you can report it, right, so once you do this, it helps you in very clearly discriminating between 2 or more than 2 groups, right, so and accordingly draw an inference in the study, right, so this technique has a very high you know utility and it can be utilised, so you can understood by now that logistic regression and discriminate analysis both are similar kind of techniques.

The only difference lying in the data's behaviour whether it is a normal data or not, if it is normal, then discriminant analysis, if it is not normal not following a normality condition, then it is a logistic regression case, right. Well, this is all for the day, thank you very much.