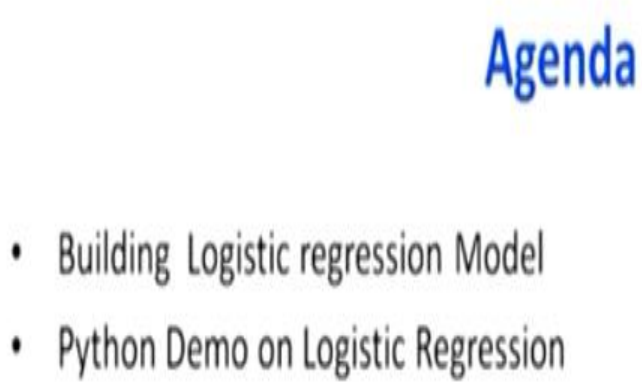
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture 38
Logistic Regression - I

In this class, we are going to new topic that is the logistic regression. I am going to explain, when you go for logistic regression over linear regression and see generally when we are doing linear regression, both independent variable and dependent variable are continuous. When an independent variable there is a categorical data, we have used dummy variable regression. There may be a chance even in the dependent variable, there may be some categorical variable.

In that case, we should go for logistic regression. I will explain this logistic regression with the help of example, then I will interpret our Python output, at the end I will explain the theory behind this logistic regression.

(Refer Slide Time: 01:14)



The class agenda is that we will build a logistic regression model, then I will do the demo for logistic regression model. We will see the application of logistic regression.

(Refer Slide Time: 01:24)

Application

$y = a + b_0 x_1 + b_2 x_2$

- In many regression applications the dependent variable may only assume two discrete values.
- For instance, a bank might like to develop an estimated regression equation for predicting whether a person will be approved for a credit card or not
- The dependent variable can be coded as y =1 if the bank approves the request for a credit card and y = 0 if the bank rejects the request for a credit card.
- Using logistic regression we can estimate the probability that the bank will approve the request for a credit card given a particular set of values for the chosen independent variables.

In many regression applications, the dependent variable may only assume 2 discrete variables. For example, in the linear regression also the dependent variable was a continuous variable, independent variable also continuous, but some time what may happen, the dependent variable may be discrete values. For example, gender. For example, good or bad or success or failure. So the y value, so what is happening? See in a linear regression, y = a + b1x1 +b2x2.

If x1, x2 are independent variable, y is a dependent variable, in the x1 if there is any categorical data, we should go for a dummy variable regression. Sometime in the dependent variable, there may be some categorical data. For example, 0, 1, it may be gender. It may be quality of product, good or bad, whether a person will buy the product or not buy the product. Whenever there is two options, it may be categorical. That time, we should go for a logistic regression.

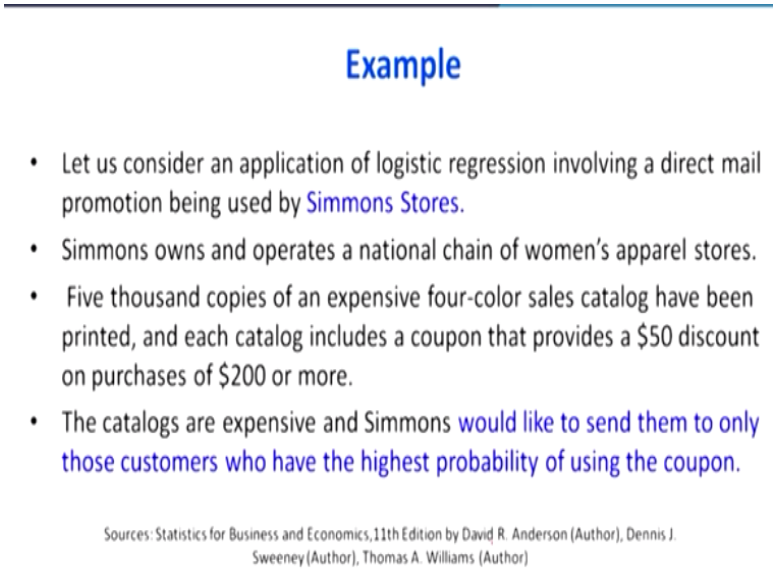
For instance, a bank might like to develop your estimated regression equation for predicting whether a person will be approved for a credit card or not. Here the y, the dependent variable is a person will be approved for getting credit card or not. We have two possibility, then you should go for logistic regression. The dependent variable can be coded as y = 1, if the bank approves the request for a credit card and y = 0 if the bank rejects request for a credit card.

Using logistic regression, we can estimate the probability that the bank will approve the request of a credit card given a particular set of values for the chosen independent variable. This may be

applicable, when you go for a plain loan. Whether this person will repay the loan or not; because there are two possibilities, so that also we can use logistic regression. Somebody applying for some jobs whether he will get the job or he will not get the job.

For that purpose, you can go for logistic regression. In your context, we can say whether you will get the placement or not. Here it is only two possibilities, a person may get or may not get the placement. So that case, we can have different independent variables. So what kind of independent variables will help you to get the placement, that kind of problem can be solved with the help of this logistic regression model.

(Refer Slide Time: 04:06)

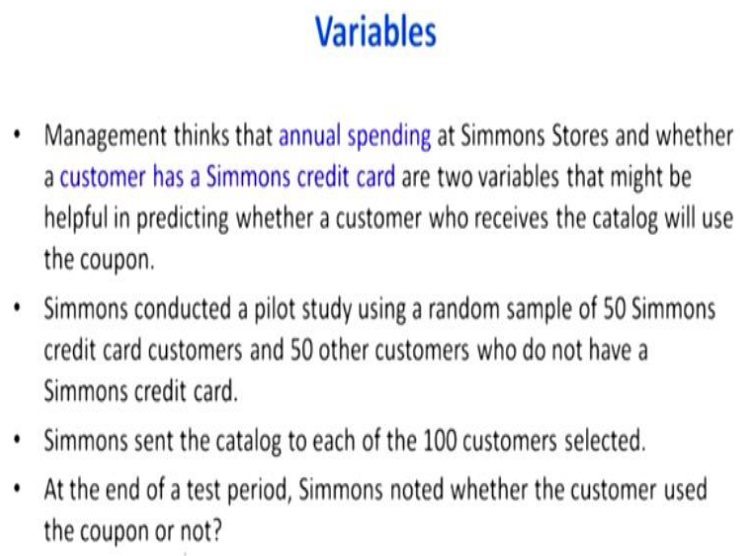


Let us take one example. This example is taken from the book Statistics for Business and Economics, 11th Edition by David Anderson, Dennis Sweeney and Thomas Williams. I will suggest you, this book is excellent book to understand the concepts. This problem also, which I have taken in this lecture, is from this book. Let us consider an application of logistic regression involving direct mail promotion being used by Simmons Stores. The store name is Simmon.

So they are going for a promotion. Simmon owns and operates a national chain of woman's apparel stores. 5000 copies of expensive four colour sales catalog have been printed and each catalog includes a coupon that provides a $50 discount on purchase of $200 or more. The catalogs are expensive and Simmons would like to send them to only those customers, who have

the highest probability of using the coupon. Now we have to identify, for what kind of customers we have to target, so that they will use the coupon.

(Refer Slide Time: 05:22)



What are the variables in these problems? The management thinks that the annual spending is one of the variable at Simmon stores and whether a customer has see a Simmon credit card are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon. So there are two independent variables. One is annual spending. Second one whether the person having some Simmon's credit card or not.

Simmons conducted a pilot study using a random sample of 50 Simmons' credit card customers and 50 other customers who do not have the Simmon credit card. Simmons sent to the catalog to each of the hundred customers selected. At the end of the test period, Simmons noted whether the customers used the coupon or not. By using this data set, they are going to construct a regression equation model, so that they can target to whom this catalog can be sent, so that they will use this coupon, so that the sales will increase.

(Refer Slide Time: 06:29)

Data (10 customer out of 100)

Customer	Spending (x_1)	Card (x_2)	Coupon (y)
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

This is a data set. I have shown for 10 customers, but there are 100 dataset; 50 dataset people are those who are not having the credit card, the remaining 50 are those who are having the credit card. So spending is one of the independent variable. Having or not having, for example 1 means having the credit card, 0 means not having the credit card. Here the coupon also 0 means they have not used the coupon, 1 means they have used the coupon.

So the coupon this variable, this is going to be our dependent variable. This is one independent variable x1, this is another independent variable x2. You see that the x2 variable is a categorical variable right. Actually in case, there are different levels. We have to convert into a dummy variable, then you have to run the analysis, but in this problem directly it is given, whether the person is having the credit card or not having the credit card.

(Refer Slide Time: 07:30)

Explanation of Variables

- The amount each customer spent last year at Simmons is shown in thousands of dollars and the credit card information has been coded as 1 if the customer has a Simmons credit card and 0 if not.
- In the Coupon column, a 1 is recorded if the sampled customer used the coupon and 0 if not.

Now we will go for what is the explanation of variables. The amount of each customer spent last year at Simmons is shown in 1000s of dollars and the credit card information has been coded as 1 if the customer has the Simmon credit card, 0 if not. So two variables, one is how much spent the last year, whether the person having the credit card or not. If the person is having credit card 1, otherwise it is 0. In the coupon column which is dependent variable 1 is recorded if the sampled customer used the coupon, 0 means if not.

(Refer Slide Time: 08:10)

Logistic Regression Equation

- If the two values of the dependent variable y are coded as 0 or 1, the value of $E(y)$ in equation given below provides the *probability* that $y = 1$ given a particular set of values for the independent variables $x_1, x_2, \ldots, x_p$.

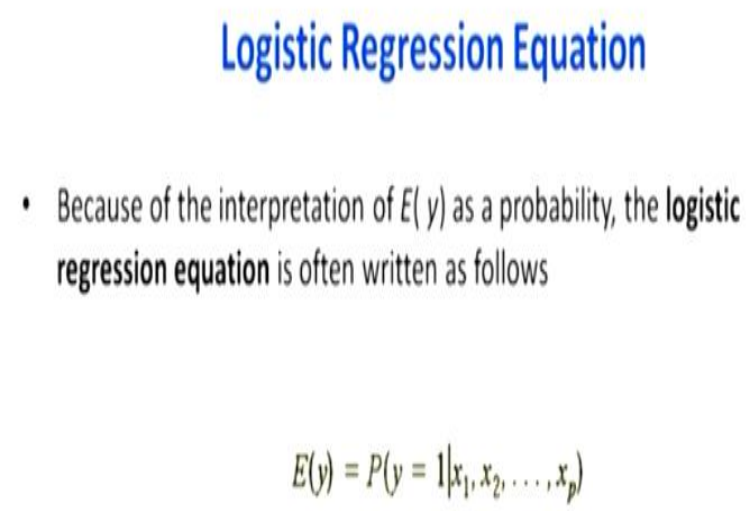
LOGISTIC REGRESSION EQUATION

$$E(y) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1 + e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}$$

First we will go for what is the logistic regression equation. If the two values of the dependent variable y are coded as 0 or 1, the value of expected y in equation given below provides the probability that y = 1 given you a particular set of values for the independent variable x1, x2, xp.

So logistic regression equation is expected value of y = e to the power b0 + beta 1x1 + beta 2x2 up to beta p. There are p independent variables + 1 + e to the power beta 0 + beta 1x1 + beta 2x2 and so on up to beta pxp. So this y is we are going to predict y. It is going to be 0 or 1.

(Refer Slide Time: 09:04)

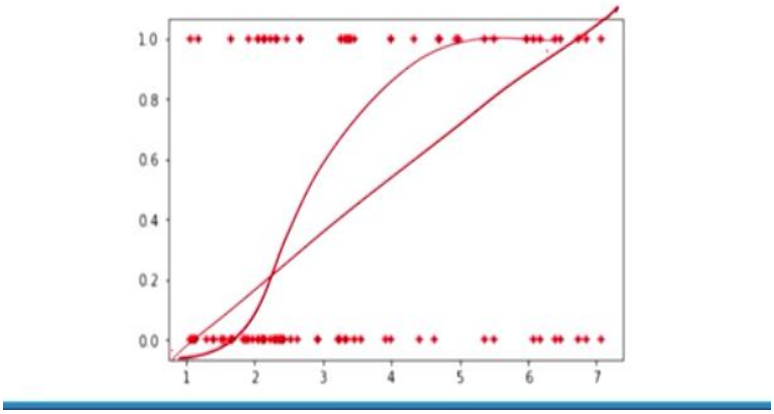


You may ask the question, why not we use simple linear regression equation here? Because the simple linear regression equation cannot be used for this problem, because there are two possibilities. We have assumed that when you plot this data set, I will tell you the next slide that is there. One assumption is that the error term in a simple linear regression should follow a normal distribution, but here the Y variable is only two possibilities. So that will follow binomial distribution and the error term of a logistic regression will follow binomial distribution.

So you cannot use your simple linear regression, whenever there is a y-value is categorical variable. Because of the interpretation of expected y is a probability of logistic regression equation is often written as expected value of y = p of y = 1 given x1, x2 up to xp. So we are going to find out the expected value of y.

(Refer Slide Time: 10:08)

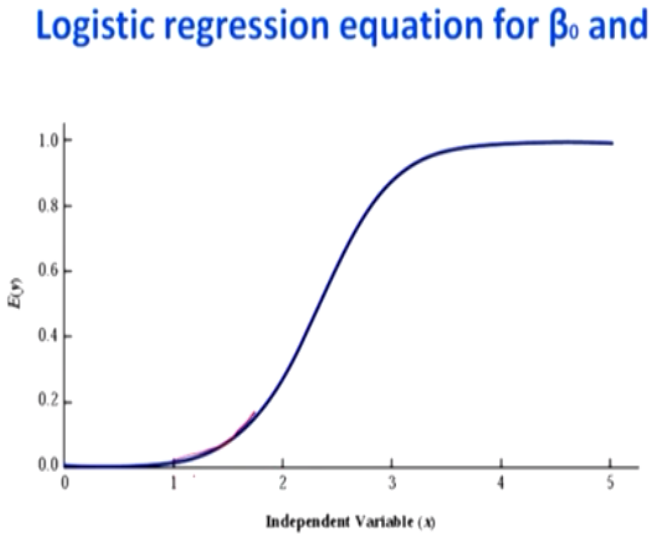
```
In [15]: plt.scatter(df.Spending,df.Coupon,marker='+',color= 'red')

Out[15]: <matplotlib.collections.PathCollection at 0x2a1b5b73c50>
```



This was the example that why we cannot use linear regression. So what will happen, here in no way you cannot construct any linearity, because here the x variable is spending. Spending is a continuous variable. The y variable is a person has used coupon or not. So what is happening? Whenever their income is low, they also used the coupon. When the incomes are more, that time also they have used the coupon.

So you cannot construct, for this data set, when you fit to this kind of linear regression, there is no meaning for that. So one way to fit the line for this kind of data is the S-shaped curve that I will show you in the next slide. It will become this way, that I will show in the next slide.

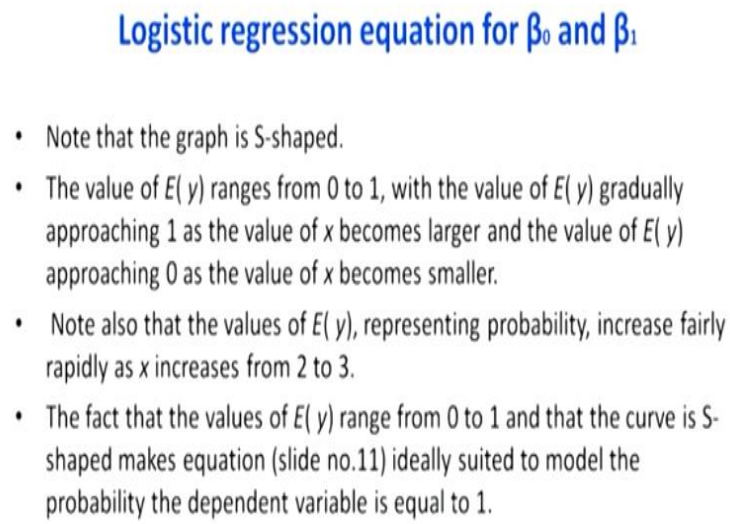
(Refer Slide Time: 11:01)



Yeah, this one assume that there are independent variable is there. The range of independent variable up to say 0 to 5, the expected value is nothing but the probability. You see that when this x = 3, it is getting the maximum value between 3 and 4. Whenever the value of x is between 3 and 4, there is a higher chances that expected value of y will be 1. When it is below 2, when the x value is below 2, there is a higher chance that the expected value will become 0.

So there are two possibilities and the rate of change also you see here, the rate of change also very high between 1 and 2, but between 2 & 3 the rate of change is low, but between 3 and 4 the rate of change is more. So this is an S-shaped curve for a logistic equation. So what we are understanding here, when x = 3 whenever the value of x is more than 3, there is a more chance the value of expected value of y will be 1. When it goes below 1 or below 2, there is more chance that the expected value of y will be 0. When you go right hand side, there is a more chance that the expected value of y becomes 1.

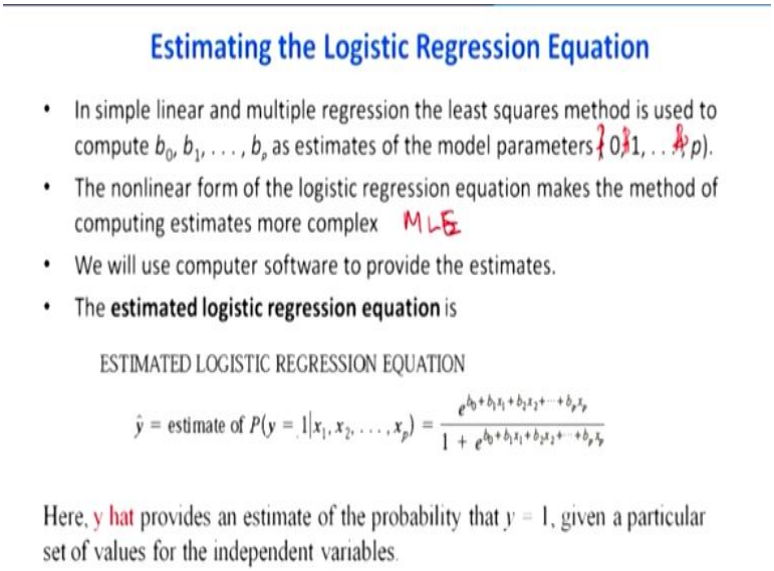
(Refer Slide Time: 12:24)



Now I will explain what is that previous curve? Note that the graph is S-shaped. The value of expected y range from 0 to 1, that is in x axis, with the value of expected value of y gradually approaching 1 as the value of x becomes larger and the value of expected value of y approaching 0 as the value of x become smaller. Note also that the value of expected y representing probability increase fairly rapidly as x increases from 2 to 3 after that it becoming constant.

The fact that the value of expected value of y range from 0 to 1 and that the curve S-shaped makes the equation, the previous slide this shape ideally suited to model the probability that dependent variable is equal to 1.

(Refer Slide Time: 13:31)



Now we will go for estimation of logistic regression equation. In a simple linear and multiple regression, the least square method is used to compute b0, b1 up to bp as estimates of the model parameter. What is the model parameter? Zero, this is beta 0, beta 0, beta 1 up to beta p. So what you have done? With the help of this sample parameter, we have estimated the population parameter, that is beta 0, beta 1 and beta p.

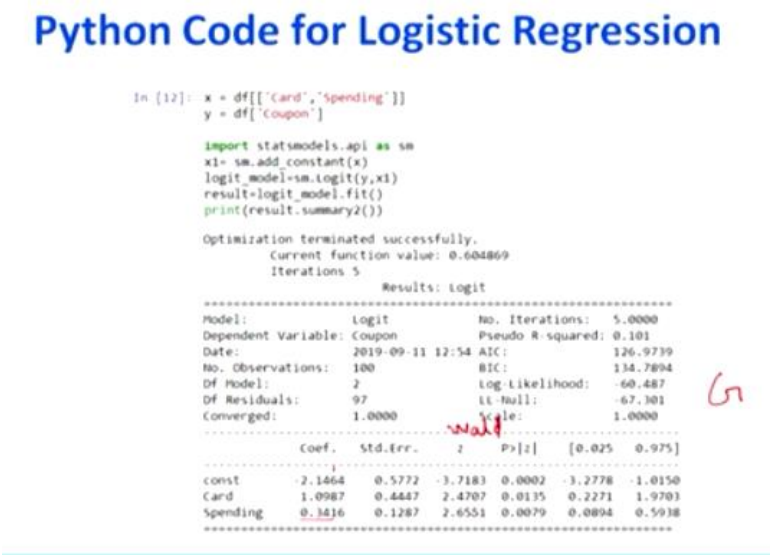
But the previous equation that is the logistic equation is non linear. So the non linear form of the logistic regression equation makes the method of computing estimates more complex. So what we are going to do? That is why in the previous class as I explained, whenever there is a non linear form of equation, instead of using that OLS method, you have to use your maximum likelihood estimation method, MLE to predict the population parameter.

So all software packages follow the concept of maximum likelihood estimation, I have explained the previous class to get, to predict the population parameter with the help of sample parameter. We will use computer software, the Python to provide the estimate. At the end of the class, I will

show you that. The estimated logistic regression equation is y hat is nothing but p of y = 1 for different x1, x2, xp equal to e to the power b0, b, b1. This b0 b1 is these sample parameter.

Divided by 1 + eb0 + b1x1 + b2x2 up to bpxp. Here why hat provides an estimate of the probability that y = 1 given a particular set of values of the independent variables. The y hat is the probability, that probability will tell us how much chance the p of y = 1. If it is a higher probability that p of y = 1. If y hat is low, you will get a lower probability.

(Refer Slide Time: 15:57)



I have brought the screenshot of the logistic regression. There are two independent variables. One is card and spending. There y is a dependent variable. So I am going to use a constant x1 = sm.add underscore constant. Here you see that we are going to use Logit model. Logit underscore model equal to sm dot Logit y, x1. Result equal to Logit underscore model dot fit. Print the result dot summary 2, then you will get this output. So look at this.

This is the constant is -2.14. This coefficient of card is 1.0987. The coefficient of spending is 0.3416. See there are 100 observations. I have shown only in my previous slides, only 10 observation only for understanding purpose. The model is Logit model and there are pseudo R square is 0.101. There is AIC. There is a log likelihood and log likelihood when the variable is not there. That is a log likelihood underscore null.

This we will use to find out the G statistics. I will tell you later. Then this is standard error of this regression coefficient. The z value, it is called wald statistic, we can say WALD statistics. That is nothing but the coefficient 1.0987 divided by 0.4447, you will get this one. This was the p-value. There are two things you have to understand before interpreting the answer. One is we have to look at the G statistics. In the coming slides I will explain what is the G statistic.

That G statistics is equivalent to F statistics of our linear regression. What we have done? The F statistics in the linear regression is helping to test the overall model and the T statistics in the linear regression is used to check the significance of an individual independent variable. The same way here the G statistics is to test the significance of overall logistic regression model. Here the z that is the WALD statistics is used to test the significance of individual the corresponding p value, is used to test the significance of individual independent variable. That is meant each independent variable. I will go further, then I will explain what is the meaning of that.

(Refer Slide Time: 18:50)

Variables

$$y = \begin{cases} 0 \text{ if the customer did not use the coupon} \\ 1 \text{ if the customer used the coupon} \end{cases}$$

x_1 = annual spending at Simmons Stores ($1000s)

$$x_2 = \begin{cases} 0 \text{ if the customer does not have a Simmons credit card} \\ 1 \text{ if the customer has a Simmons credit card} \end{cases}$$

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$$

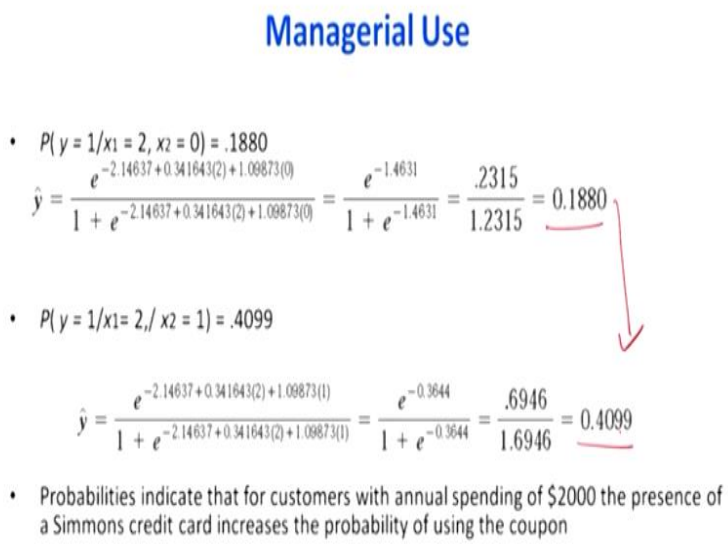
$$\hat{y} = \frac{e^{b_0+b_1 x_1+b_2 x_2}}{1 + e^{b_0+b_1 x_1+b_2 x_2}} = \frac{e^{-2.14637+0.341643 x_1+1.09873 x_2}}{1 + e^{-2.14637+0.341643 x_1+1.09873 x_2}}$$

So what are the variables? We have taken y, y can have two possibilities 0 if the customer did not use the coupon, 1 if the customer used the coupon. x1 is the annual spending at Simmon stores that is in terms of 1000, then x2 is a categorical variable. Categorical variable is 0 if the customer does not have the Simmon credit card, 1 if the customer has the Simmon credit card. So we know that the expected value of y = e to the power beta 0 + beta 1x1.

This was for the population, but this can be done with the help of y hat b0 + beta 1x1 + beta 2x2 divided by 1 + e to the power b0 beta 1x1 + beta 2x2. So this was the sample statistic. So from the previous output, what is the b0 here? See b0 is -2.1464. We got -2.1464. Now in our problem, the x1 spending, how much the customer has spent in the last time last year, see that is taken our x1 variable. Here x2, is the person is processing the card or not.

So the constant is -2.1464. So here x1 is, the coefficient of x1 is 0.34164. We got this one, 0.3416 and the coefficient of x2 is 1.0987. So we are getting 1.0987. So this was in the numerator, then 1 + e to the power the same value in the denominator.

(Refer Slide Time: 20:43)



We have got the output of logistic regression equation model. We will look at interpret and we will see the managerial use of that. How to interpret this one? For example, when y = 1, x1 = 2, x2 = 0. What is the meaning? Suppose the person's income is $2,000, is not having the credit card. When you substitute in our estimated regression equation, substitute x1 = 2, x2 = 0, both the numerator and denominator. When you simplify, we are getting 0.1880.

What is the meaning is that a person is having or not having credit card and having the expenditure of $2,000 that probability of that fellow to use that coupon is 0.1880. The same case instead of x2 = 0, I am going to see the interpretation x2 = 1, that means what? A person having

the credit card, spending $2,000, what is the probability that that person will use the coupon? So you substitute x1 = 2. Previously, we substituted x2 = 0, now substituted x2 = 1.

So when you simplify this, we are getting 0.4099. So what has happened? A person having the credit card is having the more possibility of that is the probability of y = 1, becomes higher. That means, there is a more chance that a person having the credit card will use the coupon. So probabilities indicates that the customers with the annual spending of $2,000, the presence of a Simmon credit card increases the probability of using the coupon.

How it is increasing? You see that this much. This is the probability of same income, but not having the credit card. This is the probability of having credit card. So what is happening? The probability is increased if the person is processing the credit card.

(Refer Slide Time: 22:45)

Managerial Use

- It appears that the probability of using the coupon is much higher for customers with a Simmons credit card.

		Annual Spending						
		$1000	$2000	$3000	$4000	$5000	$6000	$7000
Credit Card	Yes	0.3305	0.4099	0.4943	0.5791	0.6594	0.7315	0.7931
	No	0.1413	0.1880	0.2457	0.3144	0.3922	0.4759	0.5610

Like that, for different possibilities, so previously we have explained only this portion. Now x1 = 1, that means $1,000 x2 = 1, you will get this probability, then x1 = 1, x2 = 0, you will get this probability. Like that we have extended for up to $7,000. When you look at this figure, you see that when you compare the probability, the person having the credit card there is a more chances are that that fellow will use the coupon. If the person is not having credit card, there is a lesser chance to use the credit card. That is one interpretation.

(Refer Slide Time: 23:34)

Testing for Significance

t, F

$H_0: \beta_1 = \beta_2 = 0$

H_a: One or both of the parameters is not equal to zero

This before interpreting, we have to test whether the coefficient are significant or not, because the equation which we have can constructed is only for the population. The same thing also we have done our linear regression equation. The linear regression equation we have used a t-test and F test to predict the significance of the independent variable. So what is the null hypothesis? Beta 1 = beta 2 = 0, so one or both of the parameter is not equal to 0.

(Refer Slide Time: 24:07)

G Statistics

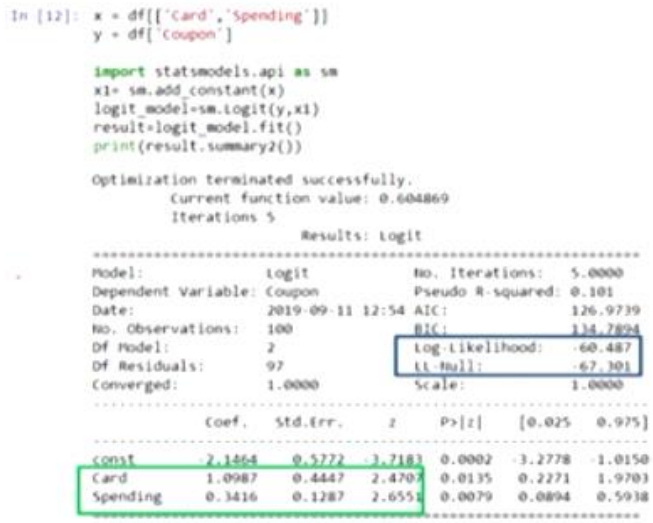
F

- The test for overall significance is based upon the value of a *G* test statistic.
- If the null hypothesis is true, the sampling distribution of *G* follows a chi-square distribution with degrees of freedom equal to the number of independent variables in the model.

Now we will go for G statistics. The test for overall significance is based upon the value of G test statistics. This is equivalent to F test statistics in our linear regression model. If the null hypothesis is true the sampling distribution of G follows a chi-square distribution with the degrees of freedom equal to the number of independent variable in the model. In our problem,

the number of independent variable is 2. So the degrees of freedom is 2. If there is only one independent variable, the degrees of freedom for G statistics is 1.

(Refer Slide Time: 24:43)



```
In [12]: x = df[['Card','Spending']]
         y = df['Coupon']

         import statsmodels.api as sm
         x1= sm.add_constant(x)
         logit_model=sm.Logit(y,x1)
         result=logit_model.fit()
         print(result.summary2())

         Optimization terminated successfully.
                 Current function value: 0.604869
                 Iterations 5
                              Results: Logit
         ================================================================
         Model:              Logit            No. Iterations:   5.0000
         Dependent Variable: Coupon           Pseudo R-squared: 0.101
         Date:               2019-09-11 12:54 AIC:              126.9739
         No. Observations:   100              BIC:              134.7894
         Df Model:           2                Log-Likelihood:   -60.487
         Df Residuals:       97               LL-Null:          -67.301
         Converged:          1.0000           Scale:            1.0000
         ----------------------------------------------------------------
                        Coef.   Std.Err.     z     P>|z|    [0.025   0.975]
         ----------------------------------------------------------------
         const         -2.1464   0.5772  -3.7183  0.0002  -3.2778  -1.0150
         Card           1.0987   0.4447   2.4707  0.0135   0.2271   1.9703
         Spending       0.3416   0.1287   2.6551  0.0079   0.0894   0.5938
         ================================================================
```

So this was the output. I am going to explain how we got this G statistics. So look at this value, which I have coloured in the blue colour log likelihood is - 60.487, log likelihood when the variable is not there. That is a log likelihood underscore null is – 67.307.

(Refer Slide Time: 25:05)

G Statistics

$$G = -2\ln\left[\frac{\text{(likelihood without the variable)}}{\text{(likelihood with the variable)}}\right].$$

$$G = 2(-60.487 - (-67.301) = 13.628$$

- The value of G is 13.628, its degrees of freedom are 2,and its p-value is 0.001.
- Thus, at any level of significance α >= .001, we would reject the null hypothesis and conclude that the overall model is significant.

Formula to find out the G statistics is G = - 2log n. There is a log likelihood with variable. First one is without variable numerator, the denominator is with variable. So G = 2, we got this – 60.487 you see that this value, - 60.487. So when it is null, that value is – 67. 307. So when you

find the difference and multiply by 2, this value is your G value. 13.628. So the value of G is we are getting the same answer. G is 13.628.

It is degrees of freedom for 2, because 2 independent variables and corresponding p-value is, it is 0.001. Thus at any level of significance, since alpha is greater than 0.001 is very low, we would reject null hypothesis and conclude that the overall model is significant. In this class, I have explained when to go for logistic regression equation. When should we go for logistic regression equation, whenever the dependent variable is the categorical variable, we should go for logistic regression equation.

Then I have taken a sample problem. With the help of sample problem, I have used Python to predict the different values that is the various parameters of the logistic regression equation. Then, I have explained what is the G statistics. The G statistics is equivalent to F statistics in our linear regression model. In the linear regression model, F statistics is used to test to the overall significance of the model.

The same way in the logistic regression equation or model, the G statistics is used to test the overall significance, but we have to check the individual significance of each independent variable. That we will continue in the next class. It is equivalent to looking at the t value of our linear regression model. The linear regression model and t statistics is used to test the significance of an individual variable. The same way in our logistic regression equation, the z statistics or the Wald statistics method is used to find out the significance of each independent variable. That we will continue in the next class.