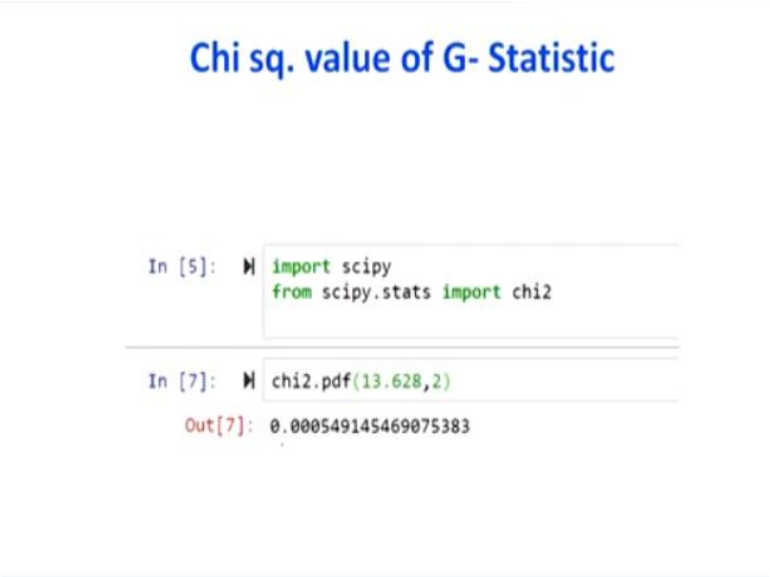
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 39
Logistic Regression - II

In the previous class, we have done the overall significance of the logistic regression model, in this class, we will go for testing the individual significance of all independent variable.

(Refer Slide Time: 00:41)



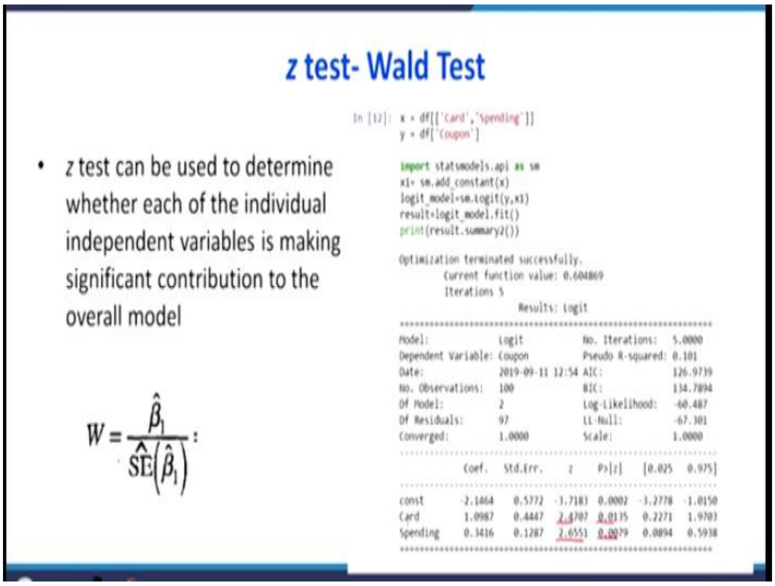
Chi sq. value of G- Statistic

```
In [5]:   import scipy
          from scipy.stats import chi2

In [7]:   chi2.pdf(13.628,2)

Out[7]:   0.000549145469075383
```

So, the agenda for this class is; testing the significance of logistic regression coefficient then we will do the Python demo on logistic regression. In the previous class I have stopped by saying the G statistics and corresponding p value, that p value has less than 0.05, then we have seen the overall model is significant. So, this is the code from Python to get the for different chi square G value and degrees of freedom.

So, chi square dot pdf 13.628, 2 has our g value in the previous lecture, 2 was our degrees of freedom because there was 2 independent variable, so this was the p value, so the p value is very low, we can say that the model is significant.

(Refer Slide Time: 01:30)



z test- Wald Test

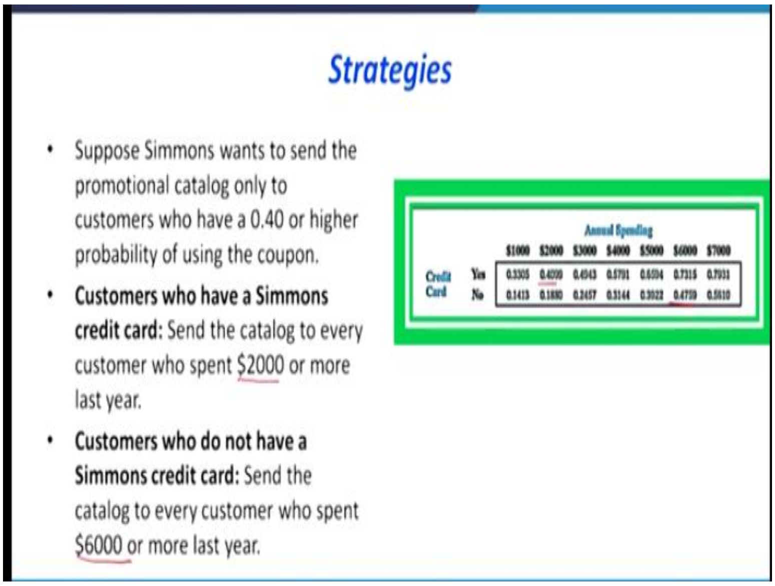
- z test can be used to determine whether each of the individual independent variables is making significant contribution to the overall model

$$W = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)};$$

```
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
         ==================================================================
```

Z test or Wald test; z test can be used to determine whether each of the individual independent variable is making significant contribution to the overall model or not. For example, how we got the z value; if you divide 1.0987 divided by 0.447, you will get the z value. Similarly, when you divide 0.3416 by 0.1287, you will get this z value, so corresponding probability you see this one, both the probabilities are less than 0.05. So, we can say both the independent variable is significant, as a whole model also significant, the independent variable in a logistic regression model also significant.

(Refer Slide Time: 02:15)



Strategies

- Suppose Simmons wants to send the promotional catalog only to customers who have a 0.40 or higher probability of using the coupon.
- **Customers who have a Simmons credit card:** Send the catalog to every customer who spent $2000 or more last year.
- **Customers who do not have a Simmons credit card:** Send the catalog to every customer who spent $6000 or more last year.

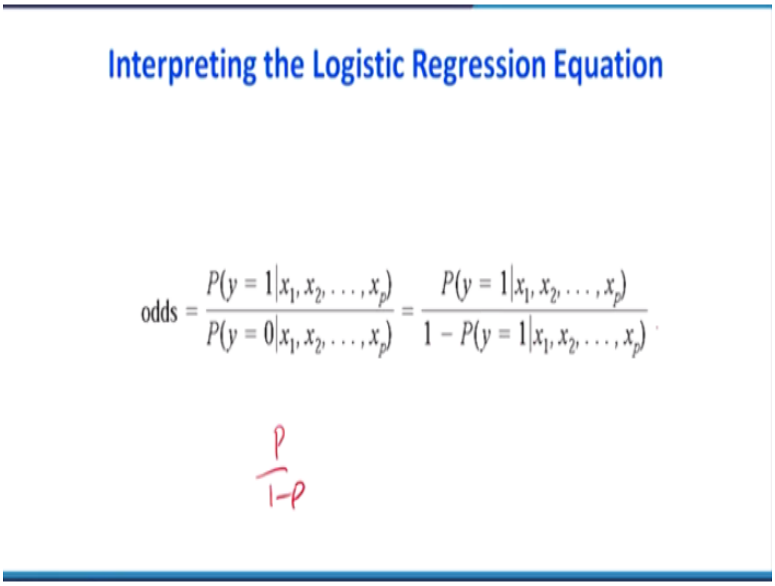
		Annual Spending						
		$1000	$2000	$3000	$4000	$5000	$6000	$7000
Credit Card	Yes	0.3305	0.4099	0.4943	0.5791	0.6594	0.7315	0.7931
	No	0.1413	0.1880	0.2457	0.3144	0.3922	0.4759	0.5610

Once we came to know both are significant, then we will go for interpretation of its output, so what kind of different strategies that company has to adopt, so that they can improve their revenue by selling more coupons. Suppose Simmons wants to send the promotional catalog only to customers who have a 0.40 or higher probability of using the coupon. So, what will happen, you look at where this 0.4, here this one 0.4, those who are having credit card.

Those who are not having credit card, 0.4 is here, so what interpretation from this table is; whoever having the credit card and whose spending is 2,000 dollar and above, for them you can send the coupon, they will use the coupon. Those who are not having the coupon but their spending is above 6,000 dollar, for them you can send the coupon, so that they will use that one.

So, this is the managerial interpretation, so customers who have a Simmons credit card, send the catalog to every customers who spend 2,000 dollar or more last year because the 0.4 is the cut-off, customers who do not have the Simmons credit card send the catalog to every customer who spent 6,000 dollar or more in the last year, so that is the strategy for the promotion.

(Refer Slide Time: 03:49)



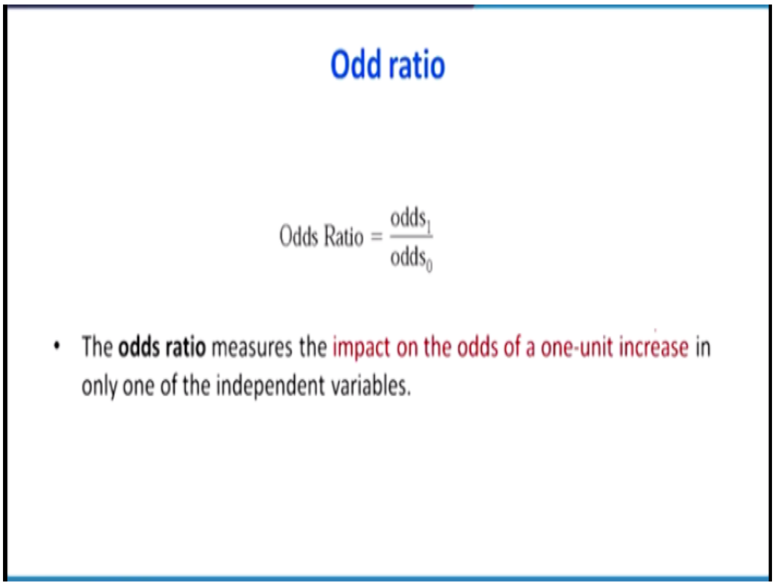
Interpreting the Logistic Regression Equation

$$\text{odds} = \frac{P(y=1|x_1, x_2, \ldots, x_p)}{P(y=0|x_1, x_2, \ldots, x_p)} = \frac{P(y=1|x_1, x_2, \ldots, x_p)}{1 - P(y=1|x_1, x_2, \ldots, x_p)}$$

$$\frac{P}{1-P}$$

Now, we will go for interpreting the logistic regression equation; for that there is a close connection between odd. What is odd is, probability of success divided by probability of not getting success, so generally the odd is P divided by 1 - P that is odd, so probability of y equal to 1 is the success, probability of y equal to 0 is not success. So, probability of y equal to 1 given that a different independent variable that is a numerator P, so 1 - P of y equal to 1 that is 1 – P, this function is called odds function.

(Refer Slide Time: 04:33)



Odd ratio

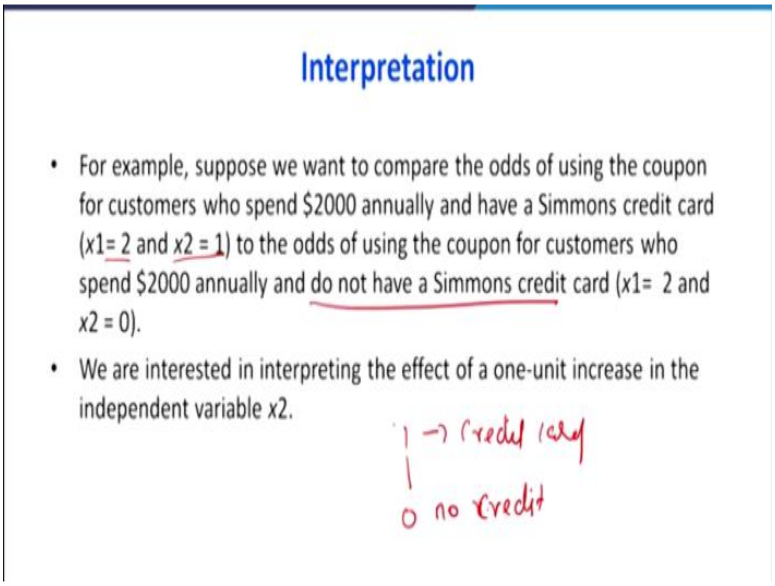
$$\text{Odds Ratio} = \frac{\text{odds}_1}{\text{odds}_0}$$

- The **odds ratio** measures the impact on the odds of a one-unit increase in only one of the independent variables.

Then from odds function, we are going to the next term odds ratio because this odds ratio is very, very helpful to explain the coefficient of logistic regression equation. So, the odds ratio is odds 1 divided by odds 0, that means when 0th level what is the odd, when the first level what is the odd, so 0th level means for example, those who are not having credit card that is odds is 0.

Those who are having the credit card suppose, we are jumping from one level to another level, odds 1 is those who are having the credit card, so the odd ratio measures the impact of odds of 1 unit increase in only one of the independent variable, so this odds ratio will help us to interpret to the logistic regression equation, if the independent variable is increased by 1 unit, what is the effect of that on the dependent variable will interpret this.

(Refer Slide Time: 05:33)



Interpretation

- For example, suppose we want to compare the odds of using the coupon for customers who spend $2000 annually and have a Simmons credit card ($x1= 2$ and $x2 = 1$) to the odds of using the coupon for customers who spend $2000 annually and do not have a Simmons credit card ($x1= 2$ and $x2 = 0$).
- We are interested in interpreting the effect of a one-unit increase in the independent variable $x2$.

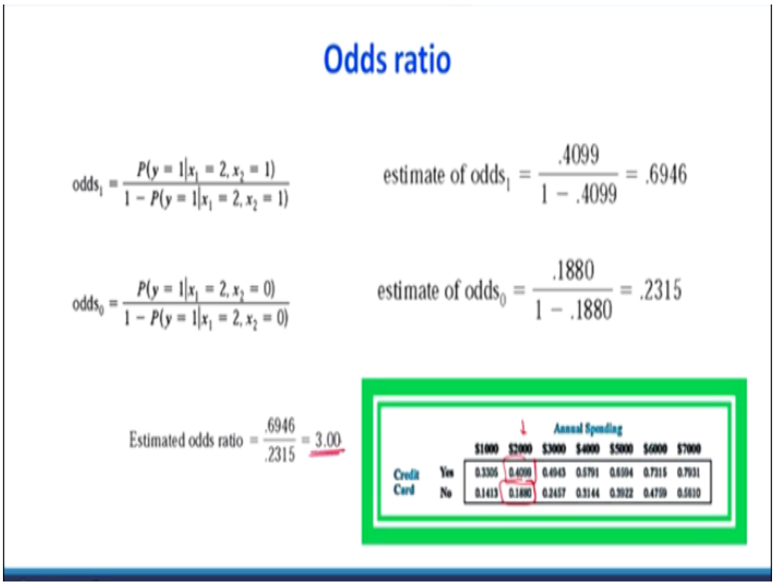
1 → Credit card

0 no Credit

What is interpretation? For example, suppose we want to compare the odds of using the coupon for customers who spent 2,000 dollar annually and have a Simmons credit card, so that is this category; x1 is a 2000 having the credit card to the odds of the coupon for customers who spent 2,000 dollar annually and do not have the Simmons credit card. So, what will happen here; x2 will becomes 0, so x1 equal to 2, x2 equal to 0, so that is the second category is odd 0, the first one is odds 1.

So, the ratio of that 2 is called odd ratio, so we are interested in interpreting the effect of 1 unit increase on the independent variable x2, so 1 unit increase means here person having not credit card to a person having the credit card, so the 0 level to 1 level. The 0 level is no credit card, the 1 level is credit card, okay. So, what is the impact of this one on the estimated value of y?

(Refer Slide Time: 06:48)



Odds ratio

$$\text{odds}_1 = \frac{P(y=1|x_1=2, x_2=1)}{1-P(y=1|x_1=2, x_2=1)} \qquad \text{estimate of odds}_1 = \frac{.4099}{1-.4099} = .6946$$

$$\text{odds}_0 = \frac{P(y=1|x_1=2, x_2=0)}{1-P(y=1|x_1=2, x_2=0)} \qquad \text{estimate of odds}_0 = \frac{.1880}{1-.1880} = .2315$$

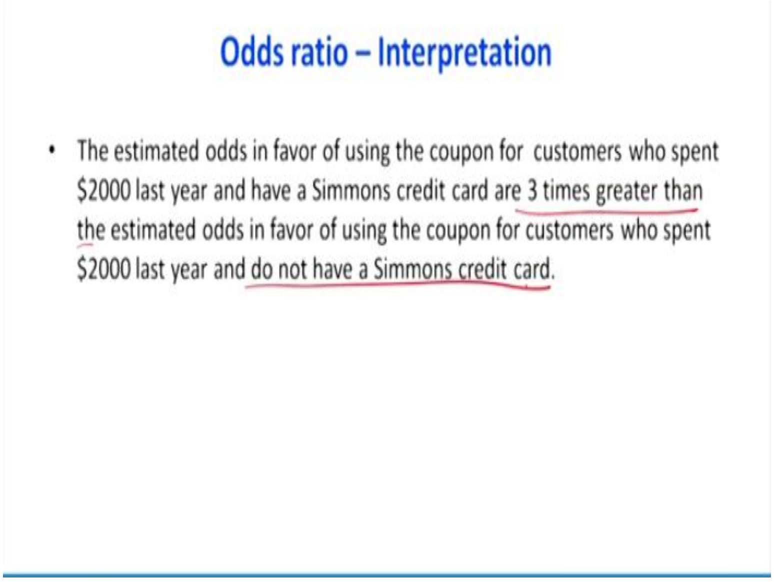
$$\text{Estimated odds ratio} = \frac{.6946}{.2315} = \underline{3.00}$$

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

First, we will see odds 1; odds 1 is a person having the correct card, so P of y equal to 1, x1 equal to 2, spending is 2,000 dollar, x2 equal to 1 having the credit card, this is the numerator, probability of success we can say probability of not success, 1 - P of y equal to 1, x1 equal to 2, x2 equal to 1. So, what will happen; you can substitute this, when x1 equal to 2, this value and having the credit card that is this 0.0499.

So when you substitute this 0.4099 divided by 1 – 0.4, we are getting 0.6946, we will go to the odds 0; 0th level. So, P of y equal to 1, x1 equal to 2, x2 equal to 0, this case similar to previous one spending amount is same but he is not having credit card, for that the probability is numerator is P, denominator is 1 – P. So, what is that category; this one, 0.1880, so 0.1880 divided by 1 – 0.1880 that is giving 0.2315, so when you divide this 0.6946 divided by 0.2315, this 3 so, the 3 is very useful for interpreting.

(Refer Slide Time: 08:21)

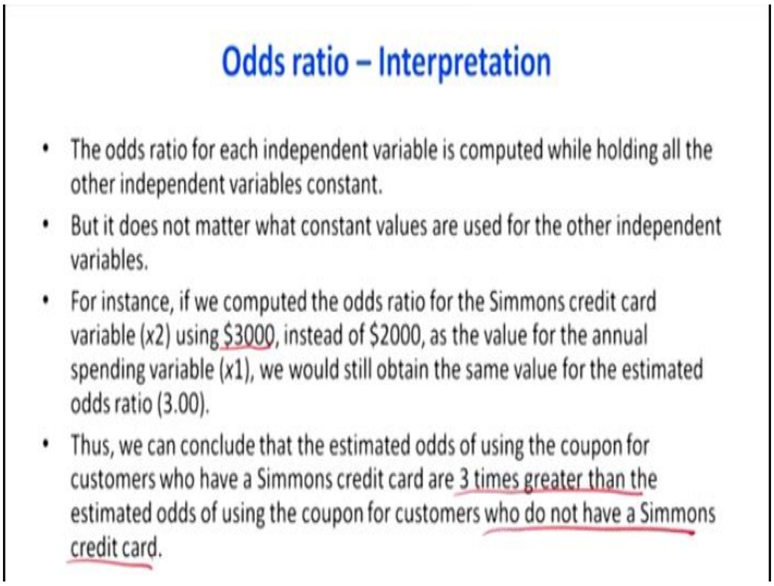


Odds ratio – Interpretation

- The estimated odds in favor of using the coupon for customers who spent $2000 last year and have a Simmons credit card are 3 times greater than the estimated odds in favor of using the coupon for customers who spent $2000 last year and do not have a Simmons credit card.

What is the meaning of 3 is; the estimated odds in favour of using the coupon for customers who spent 2,000 dollar last year and have a Simmons credit card are 3 times greater than the estimated odds in favour of using the coupon for customers who spent to 2,000 dollar last year and do not have the Simmons credit card. So that means, expenditure is, the spending amount is same. But when you sent this coupon to the person who is having the credit card, there is a 3 times more chance that person will use that coupon, okay that is the meaning of this odds ratio.

(Refer Slide Time: 09:03)



Odds ratio – Interpretation

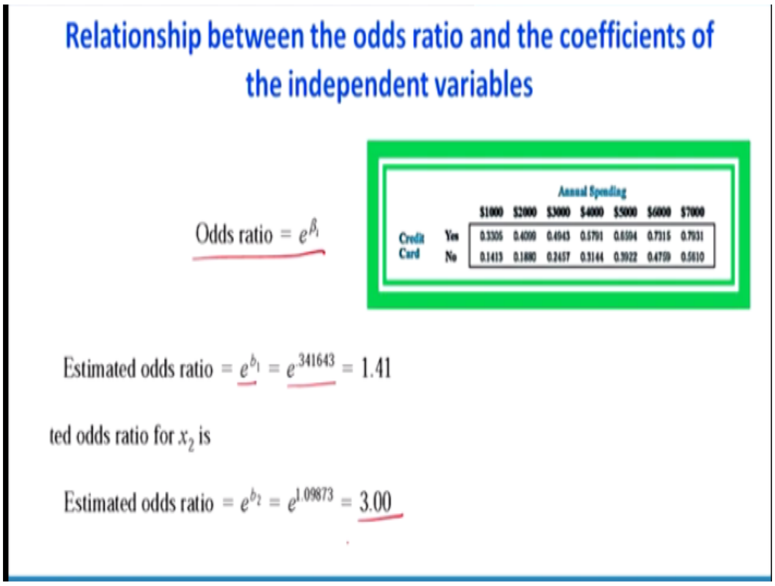
- The odds ratio for each independent variable is computed while holding all the other independent variables constant.
- But it does not matter what constant values are used for the other independent variables.
- For instance, if we computed the odds ratio for the Simmons credit card variable ($x2$) using $3000, instead of $2000, as the value for the annual spending variable ($x1$), we would still obtain the same value for the estimated odds ratio (3.00).
- Thus, we can conclude that the estimated odds of using the coupon for customers who have a Simmons credit card are 3 times greater than the estimated odds of using the coupon for customers who do not have a Simmons credit card.

The odds ratio for each independent variable is computed while holding all other independent variable is constant for example, in the previous case also where the expenditure; the amount spent that is expenditure is taken as the constant, we have interpret only for a person having the credit card or not having the credit card, it does not matter constant values are used for other independent variables.

So, we do not bother about the constant variables for instance, if we computed the odd ratio for Simmons credit card variable x2 instead of 2,000 you say, 3,000 dollar expenditure, instead of 2,000 as the value of the annual spending variable is x1, we would still obtain the same value of estimated odd ratio, so the constant does not matter.

Thus we can conclude that the estimated odds of using coupon for customers who have a Simmons credit card are 3 times greater that is important interpretation, 3 times greater than the estimated odds of using the coupon for customers who do not have the Simmons credit card. So, you have to target to a person who has the credit card when you target to them, there is a 3 times more chance that people will use when compared to those who are not having credit card they will use the coupon.

(Refer Slide Time: 10:27)



Relationship between the odds ratio and the coefficients of the independent variables

$$\text{Odds ratio} = e^{\beta_i}$$

		Annual Spending						
		$1000	$2000	$3000	$4000	$5000	$6000	$7000
Credit Card	Yes	0.3306	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

$$\text{Estimated odds ratio} = e^{b_1} = e^{.341643} = 1.41$$

ted odds ratio for x_2 is

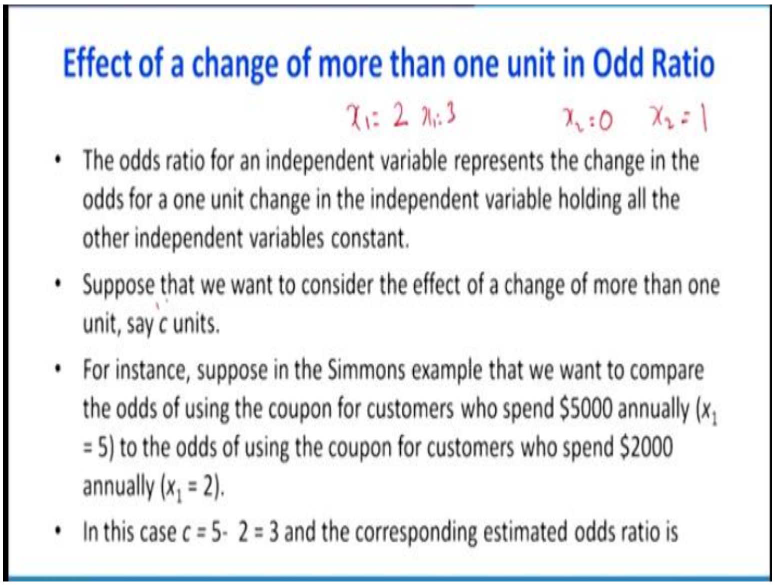
$$\text{Estimated odds ratio} = e^{b_2} = e^{1.09873} = 3.00$$

Now, another very useful relationship instead of finding odd ratio that way, there is a connection between the coefficient of logistic regression equation and the odds ratio, that is this one; e to the power beta i, what is the title says; relationship between odds ratio and the coefficient of independent variable, the beta i is called the coefficient of independent variable.

So, if you want to know the estimated odds ratio for x1 variable that is the amount spent, e to the power b1, in our equations b1 is 0.34, where we got this one b1, I am going back this one, so here we are taking x1 variable, this is x2 variable, so e to the power 0.3416 that will give you the odd ratio for this variable because in some software packages for example, Minitab they directly give the odd ratio for each independent variable.

But in Python we can calculate the odd ratio using that relationship e to the power beta 1 for example, if we want to know odd ratio for card, those who are having card or not, so e to the power 1.0987 that will become 3, I will show you that one yeah, see that b2; e to the power b2, e to the power 1.09873 is 3, so odd ratio we can directly get from the coefficient of logistic regression equation.

(Refer Slide Time: 12:05)



Effect of a change of more than one unit in Odd Ratio

$x_1: 2$ $x_1: 3$ $x_2: 0$ $x_2 = 1$

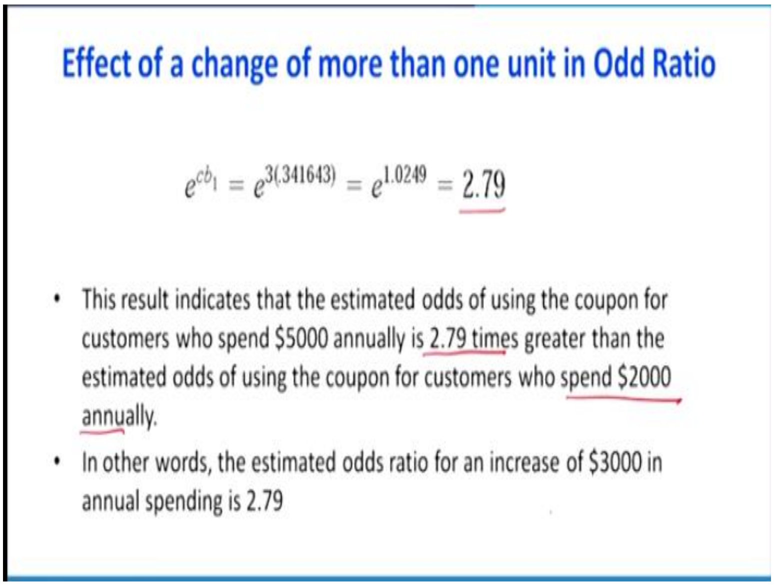
- The odds ratio for an independent variable represents the change in the odds for a one unit change in the independent variable holding all the other independent variables constant.
- Suppose that we want to consider the effect of a change of more than one unit, say c units.
- For instance, suppose in the Simmons example that we want to compare the odds of using the coupon for customers who spend \$5000 annually (x_1 = 5) to the odds of using the coupon for customers who spend \$2000 annually (x_1 = 2).
- In this case c = 5- 2 = 3 and the corresponding estimated odds ratio is

Now, so far we have seen there is 1 unit is change then we have seen, what is the corresponding effect on the dependent variable, sometime what will happen; what is the meaning of 1 unit change means, suppose we have taken for x2 equal to 0 and 1, we have seen 1 unit change, if there is 1 unit change, we have seen effect of that on the dependent variable.

For example, this is a discrete variable, if the independent variable is a continuous variable for example, say x1 is amount spent suppose, somebody is spending 2,000 dollar, what is the probability that people will use the coupon, so we can see 2,000 to 3,000 that is x1 equal to 2 to 3, if it instead of 1 unit jump, we can go for 6 unit or 5 unit jump at a time and corresponding interpretation we can find out that is the meaning of that change of more than 1 unit in the odd ratio.

The odd ratio for an independent variable represents the change in odds of 1 unit each change in the independent variable holding all other independent variables are constant. Suppose, we want to consider the effect of a change of more than 1 unit for example, c unit instead of 2 to 3, I want to say 2 to 5 for instance, suppose the Simmons example that we want to compare the odds of using the coupon for customers who spent 5,000 dollar annually to the odds of using the coupon for customers who spent 2,000, the increment is not 1. Because it is a 3 okay, in this case c equal to 5 - 2 is 3 and the corresponding estimated odd ratio is very, very useful (()) (14:03).

(Refer Slide Time: 14:05)



Effect of a change of more than one unit in Odd Ratio

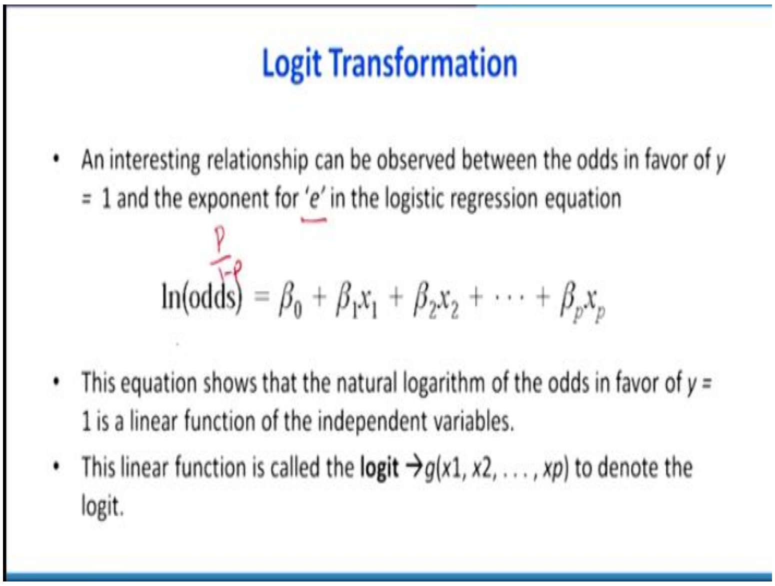
$$e^{cb_1} = e^{3(.341643)} = e^{1.0249} = \underline{2.79}$$

- This result indicates that the estimated odds of using the coupon for customers who spend $5000 annually is 2.79 times greater than the estimated odds of using the coupon for customers who spend $2000 annually.
- In other words, the estimated odds ratio for an increase of $3000 in annual spending is 2.79

So, e to the power c, this c is how much is we are increasing, so when you multiply by 3 of this one, we are getting the odd ratio is 2.79, this result indicates that the estimated odds of using the coupons for the customers who spend 5,000 dollar annually is 2.79 times greater than the estimated odds of using the coupon for customers who spend only 2,000 dollar annually.

You see that here the increment is not the unit increment, it is the 3 times increment in other words, the estimated odd ratio for increase of 3,000 in annual spending yeah, it is a 3 unit means, 3000 is 2.79.

(Refer Slide Time: 14:49)



Logit Transformation

- An interesting relationship can be observed between the odds in favor of *y* = 1 and the exponent for *'e'* in the logistic regression equation

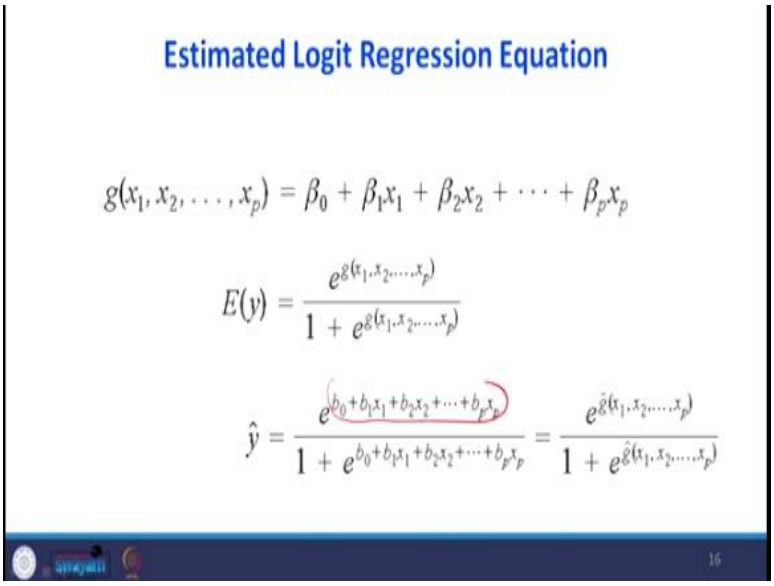
$$\ln(\overset{\frac{p}{1-p}}{\text{odds}}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- This equation shows that the natural logarithm of the odds in favor of *y* = 1 is a linear function of the independent variables.
- This linear function is called the **logit** →*g*(*x*1, *x*2, . . . , *xp*) to denote the logit.

Then we will come to some theory portions of this logistic regression equation, first we will see what is the logit transformation. An interesting relationship can be observed between odds in favour of y equal to 1 and the exponent of e in the logistic regression equation, so when you say, as I told you previously probability of success, probability of non-success when you take log of that, that is nothing but a logit function.

So, log of odds equal to beta 0 + beta 1 x1 + beta 2 x2 + beta p xp, this equation shows that the natural logarithm of the odds in favour of y equal to 1 is a linear function of independent variable. So, why we are taking log, so that will become a linear function, this linear function is called logit generally, in the custom is g of x1, x2 up to xp to denote the logit function. So, when you take logit function it will become linear, so interpretation is easy.

(Refer Slide Time: 16:04)



Estimated Logit Regression Equation

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

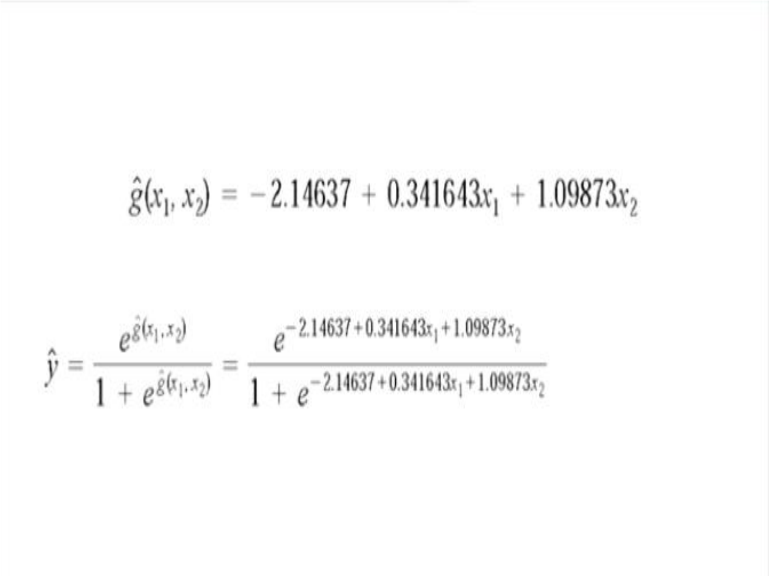
$$E(y) = \frac{e^{g(x_1, x_2, \ldots, x_p)}}{1 + e^{g(x_1, x_2, \ldots, x_p)}}$$

$$\hat{y} = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}} = \frac{e^{\hat{g}(x_1, x_2, \ldots, x_p)}}{1 + e^{\hat{g}(x_1, x_2, \ldots, x_p)}}$$

So, how to estimate the logistic regression equation, we know this a logit function, if you want to know the expected value of this logistic function is nothing but e to the power g of x1, x2 up to xp divided by 1 + e of g of x1, x2 up to xp, so you can expand this with the help of sample data; b0, b1, b2 up to bp, so this is the your sample value, with the help of the sample data we can predict the population parameter.

Sample; generally, the name parameter is used only for the population not for the sample, so e to the power; when I write the hat symbol it is the estimated value, y hat, g hat, so you can this can be written as because this is right, this can be is nothing but this value, so we can do logit function, so e to the power g hat x1, x2 up to xp divided by 1 + e to the power g hat this one.

(Refer Slide Time: 17:15)

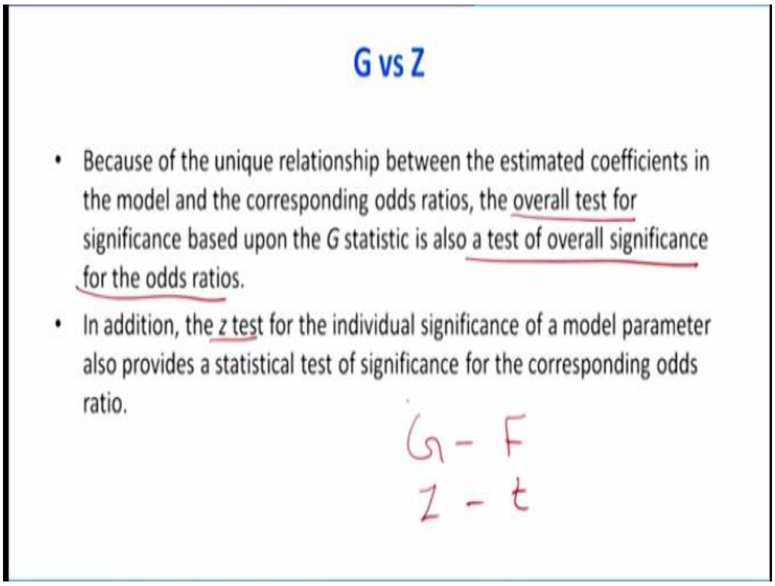


$$\hat{g}(x_1, x_2) = -2.14637 + 0.341643x_1 + 1.09873x_2$$

$$\hat{y} = \frac{e^{\hat{g}(x_1, x_2)}}{1 + e^{\hat{g}(x_1, x_2)}} = \frac{e^{-2.14637 + 0.341643x_1 + 1.09873x_2}}{1 + e^{-2.14637 + 0.341643x_1 + 1.09873x_2}}$$

Why we are taking e to the power and taking log; to make it linear, in our problem so far which we have discussed, so this was our logit equation is -2.14, 0.34164 x1 + 1.09873 x2, and if you want to predict y, here y where is the probability value, e to the power -2141, this we got this answer.

(Refer Slide Time: 17:40)



G vs Z

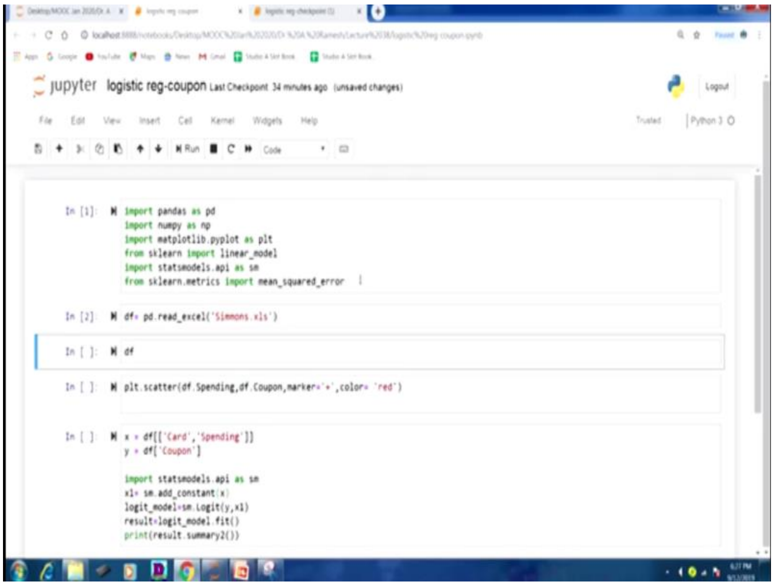
- Because of the unique relationship between the estimated coefficients in the model and the corresponding odds ratios, the overall test for significance based upon the *G* statistic is also a test of overall significance for the odds ratios.
- In addition, the *z* test for the individual significance of a model parameter also provides a statistical test of significance for the corresponding odds ratio.

G - F
Z - t

Very important things; we will compare what is the purpose of G statistics and Z statistic, as I told you because of the unique relationship between estimated coefficient in the model and the corresponding odds ratio, the overall test is very important; the overall test for the significance based upon G statistics also is test of a overall significance for the odd ratio but the z test or the Wald test for the individual significance of model parameters also provide a statistical test of significance for the corresponding odd ratio.

This is similar to G is similar to F test, z is similar to t test in the; this is for, the right side one is for linear regression, the left side one is for logistic regression. Now, we will go for Python the data which I have explained to you which I have brought you in the screenshot, I will run that model then I will show you how to do the logistic regression using Python.

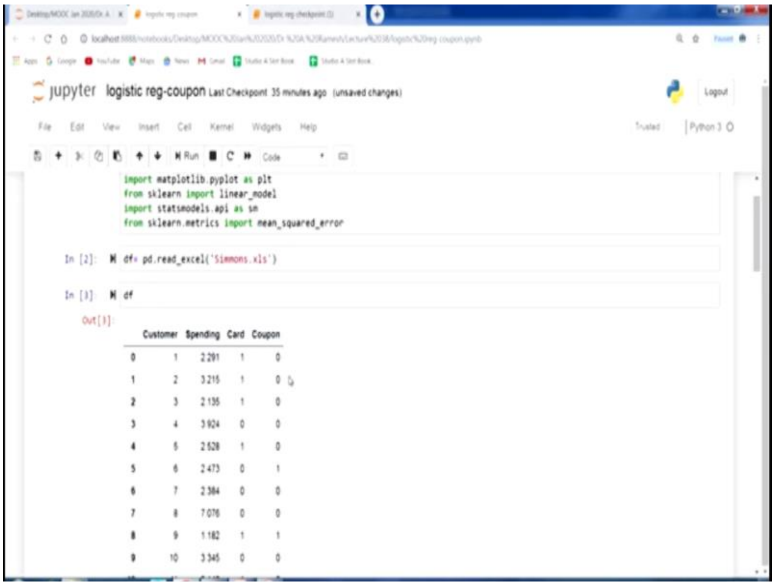
(Refer Slide Time: 18:55)



```
In [1]:  import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         from sklearn import linear_model
         import statsmodels.api as sm
         from sklearn.metrics import mean_squared_error

In [2]:  df= pd.read_excel('Simmons.xls')

In [ ]:  df

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

In [ ]:  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())
```

Now, we will go to our Python environment, then I will teach you how to do the logistic regression, so what are the libraries required? You need pandas, you need a numpy, you need a matplotlib dot pyplot, you need a sklearn for doing linear model, you can import statsmodel dot api, then sklearn metrics import mean squared error, the file name is Simmons dot xls, as I told you this was taken from Anderson, Sweeney and Williams book.

(Refer Slide Time: 19:26)


jupyter logistic reg-coupon Last Checkpoint 35 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code

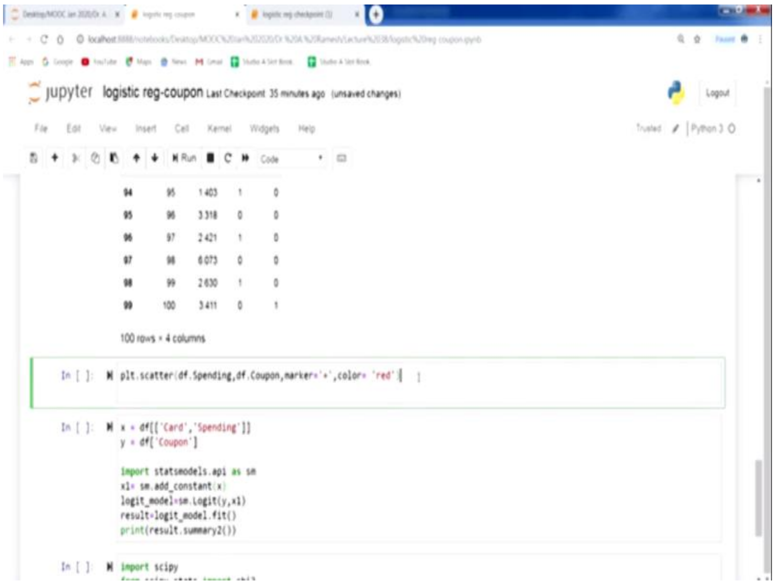
```
import matplotlib.pyplot as plt
from sklearn import linear_model
import statsmodels.api as sn
from sklearn.metrics import mean_squared_error
```

```
In [2]:  df= pd.read_excel('Simmons.xls')
```

```
In [3]:  df

Out[3]:
     Customer  Spending  Card  Coupon
  0         1     2.291     1       0
  1         2     3.215     1       0
  2         3     2.135     1       0
  3         4     3.924     0       0
  4         5     2.528     1       0
  5         6     2.473     0       1
  6         7     2.384     0       0
  7         8     7.076     0       0
  8         9     1.182     1       1
  9        10     3.345     0       0
```

So, this was the data, so what is happening I am scrolling, there is a 100 data set is there, so what are the variable is there; customer number is there, 1, 2, 3 up to 100, spending; how much they spend last time, then whether the possession of the card or not, if 1 means they are having the card, 0 means not having the card, then coupon; whether they have used the coupon, 0 means not use the coupon, 1 means uses the coupon.

(Refer Slide Time: 19:56)


jupyter logistic reg-coupon Last Checkpoint 35 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code

```
 94        95     1.403     1       0
 95        96     3.318     0       0
 96        97     2.421     1       0
 97        98     6.073     0       0
 98        99     2.630     1       0
 99       100     3.411     0       1

100 rows × 4 columns
```

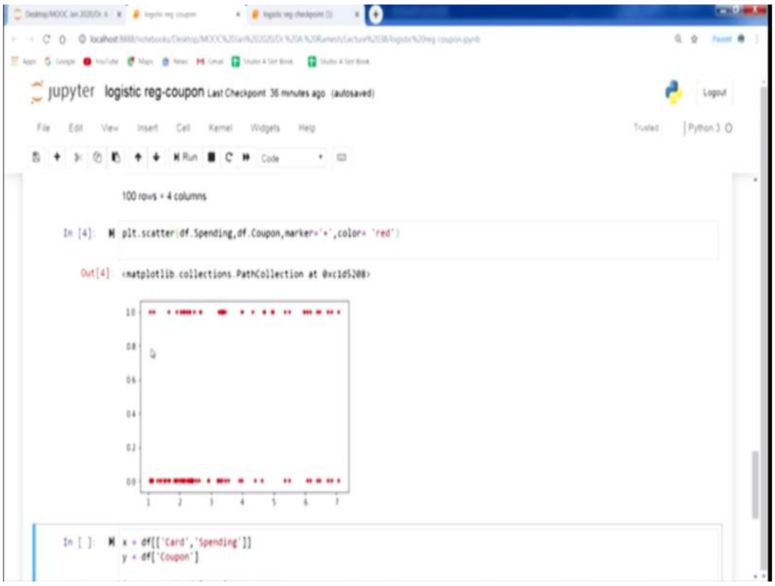
```
In [ ]:  plt.scatter(df.Spending,df.Coupon,marker='+',color= 'red')
```

```
In [ ]:  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())
```

```
In [ ]:  import scipy
```

First, we will do the scatterplot between spending and coupon, when you look at this data spending is a continuous variable, coupon is the categorical variable, coupon is our dependent variable.

(Refer Slide Time: 20:13)


jupyter logistic reg-coupon Last Checkpoint 36 minutes ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Logout
Run
Code
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7

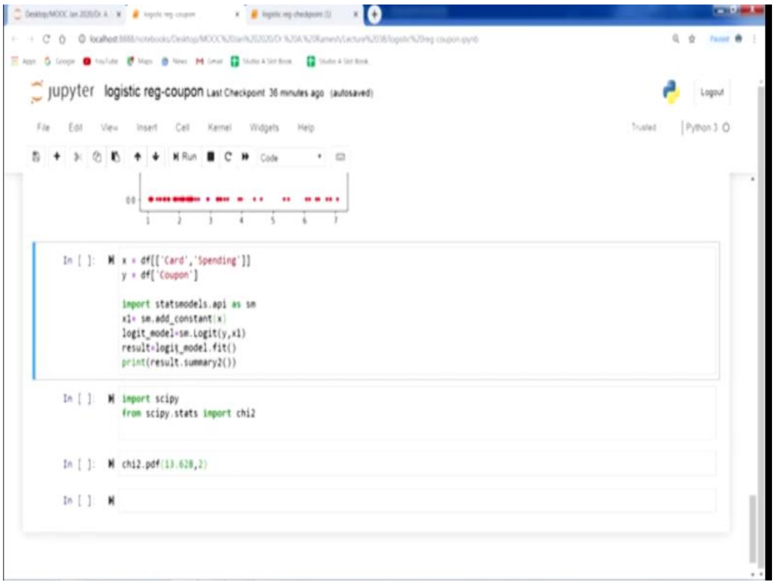
100 rows × 4 columns

```
In [4]:  plt.scatter(df.Spending,df.Coupon,marker='+',color= 'red')

Out[4]: <matplotlib.collections.PathCollection at 0xc1d5208>
```

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

So, when you run this you see that you are getting this way, so for this kind of model whether there is a 2 possibility, it is 1 or 0, in between there is no possibility, so the linear regression model is not valid here, we should go for logistic regression model that is one point. The another point is here the assumption is a linear regression model the error term will follow normal distribution but the logistic regression model error term will follow binomial distribution, so you cannot go for linear regression model.

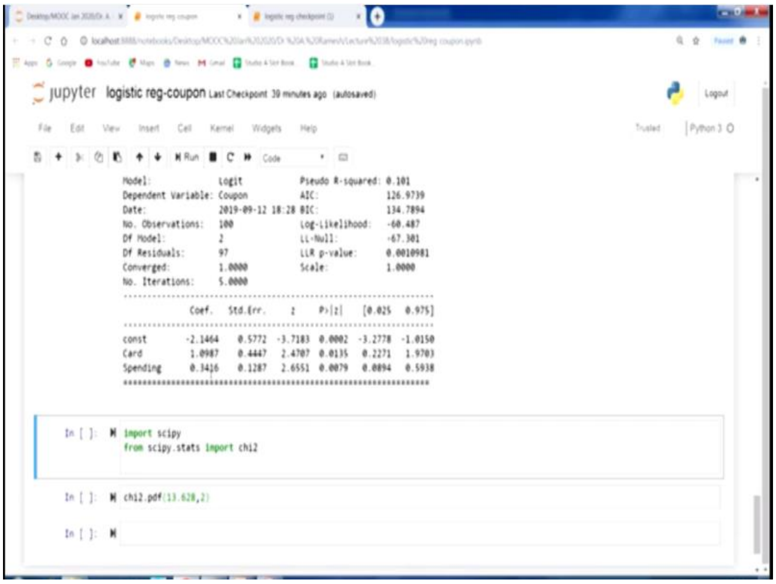
(Refer Slide Time: 20:47)


jupyter logistic reg-coupon Last Checkpoint 36 minutes ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Logout
Run
Code
0.0
1
2
3
4
5
6
7

```
In [ ]:  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())

In [ ]:  import scipy
         from scipy.stats import chi2

In [ ]:  chi2.pdf(13.628,2)

In [ ]:
```

In x, I am taken card and spending, the Y is a dependent variable is coupon, so x1 equal to sm dot add underscore constant x, logit model sm dot Logit y, x1, result logit underscore model fit, so I am going to summary of the logistic regression.

(Refer Slide Time: 21:10)



```
Model:              Logit            Pseudo R-squared: 0.101
Dependent Variable: Coupon           AIC:              126.9739
Date:               2019-09-12 18:28 BIC:              134.7894
No. Observations:   100              Log-Likelihood:   -60.487
Df Model:           2                LL-Null:          -67.301
Df Residuals:       97               LLR p-value:      0.0010981
Converged:          1.0000           Scale:            1.0000
No. Iterations:     5.0000
-----------------------------------------------------------------
              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
=================================================================
```

```
In [ ]: import scipy
        from scipy.stats import chi2
```

```
In [ ]: chi2.pdf(13.628,2)
```

```
In [ ]:
```

So, summary of logistic regression is see the constant value we need not bother about this, for card it is 1.0987, for spending it is 0.3416, so after getting this output what do you have to see; you have to check the overall significance with the help of G statistics that it is not here but you have to find out how to multiplied by - 60.487 minus, minus of – 67.301, so that value is your G value.

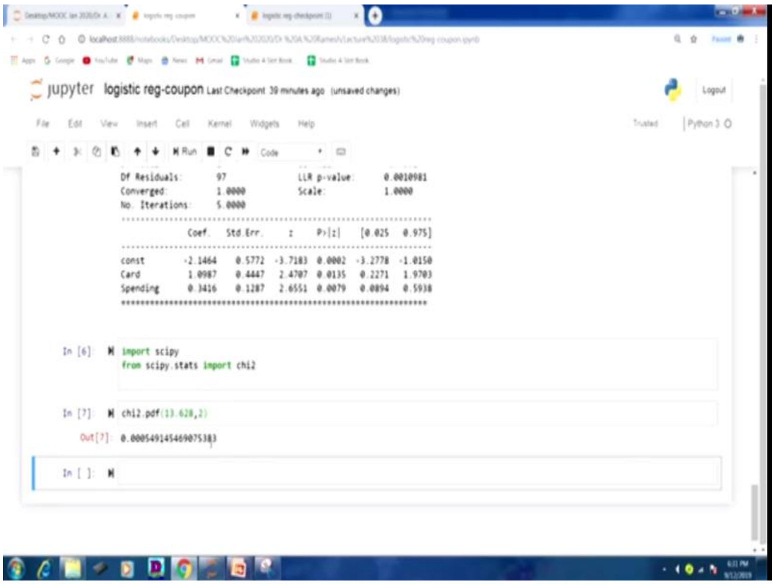
For that G value you have to find out by having 2 degrees of freedom, that G value is nothing but your chi square value, you have to find out what is the corresponding p value, with the p value is less than 0.05, we can say the overall model is significant. The next aspect is checking whether each independent variable is significant or not that is done with the help of Wald test.

So, you can look at here, it is the p value is 0.01 less than 0.05, for second variable also the p value is less than 0.05, then we can say both the variable is significant. Suppose, if you want to interpret this model with the help of odd ratio, what you have to do; when you take e to the power this beta 1 that is e to the power 1.0987, you will get a corresponding odd ratio that is used to explain 1 unit increase.

Suppose, the card person is not having card to having the card, so what was the corresponding effect on the dependent variable that can be found out. The spending is a continuous variable, here also we can find out e to the power – 0.3416 will give you the odd ratio, so that will help you to interpret, suppose a person is spending 2,000, another person is spending 3,000.

Suppose, there is 1 unit jump what is the corresponding chances because of that 1 unit jump that the person will use the coupon suppose, if there is a c unit jump simply you have to see; you have to find out e to the power c of that is a c of 0.3416, so you will get a c unit odd ratio that you can directly interpret it.

(Refer Slide Time: 23:33)



jupyter
logistic reg-coupon Last Checkpoint: 39 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code

```
Df Residuals:        97           LLR p-value:       0.0010981
Converged:           1.0000       Scale:             1.0000
No. Iterations:      5.0000
==================================================================
              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
==================================================================
```

```
In [6]:  import scipy
         from scipy.stats import chi2
```

```
In [7]:  chi2.pdf(13.628,2)
Out[7]:  0.0005491454690753 83
```

```
In [ ]:
```

This is the code to check the G value for example, as I told you previously the G value is 13.628, it is mentioned in my PPT also, so chi square value is 13.628 and the degrees of freedom is 2, why it is 2 because there are 2 independent variable, so corresponding p value is 0.0054 that is less than 0.05, overall model is significant. In this class I have explained how to test the significance of each independent variable in a logistic regression equation that I have done with the help of z statistics otherwise, Wald statistics.

Then I have explained what is the odds, then I have explained what is the odds ratio, then I explained how to use this odds ratio to interpret the coefficient of logistic regression equation, at the end I have used Python and I have shown how to use Python for running logistic regression equation. Thank you very much.