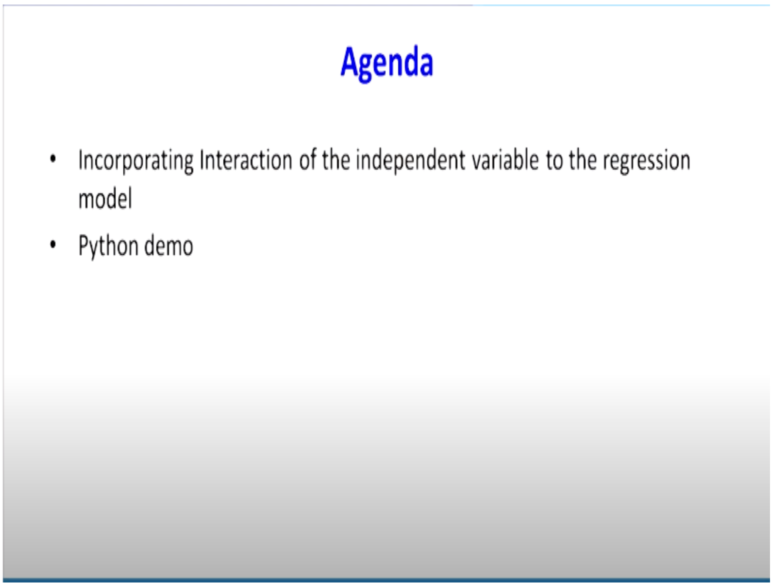
Data Analytics with Python
Prof. Ramesh Anbanandam.
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 45
Regression Analysis Model Building (Interaction) – II

In this lecture, we are going to see if there are two independent variable. If they have some interaction, how to incorporate this effect of interaction onto the dependent variable. Before that I will explain with an example what is interaction, then I will construct regression model for incorporating this regression. At this end, I will use the Python to run this interaction regression model.

(Refer Slide Time: 00:55)

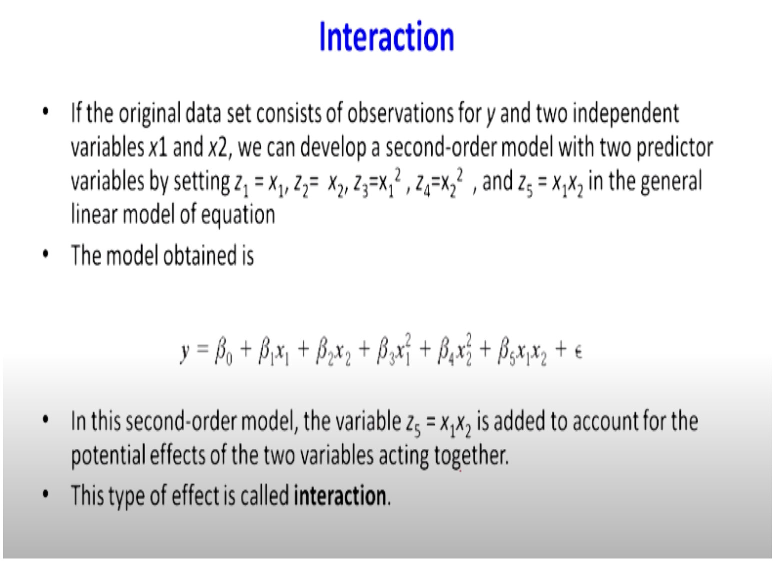


Agenda

- Incorporating Interaction of the independent variable to the regression model
- Python demo

The agenda for this lecture is incorporating interaction among independent variables to the regression model and Python demo.

(Refer Slide Time: 01:07)



Interaction

- If the original data set consists of observations for *y* and two independent variables *x*1 and *x*2, we can develop a second-order model with two predictor variables by setting $z_1 = x_1$, $z_2 = x_2$, $z_3 = x_1^2$, $z_4 = x_2^2$, and $z_5 = x_1 x_2$ in the general linear model of equation
- The model obtained is

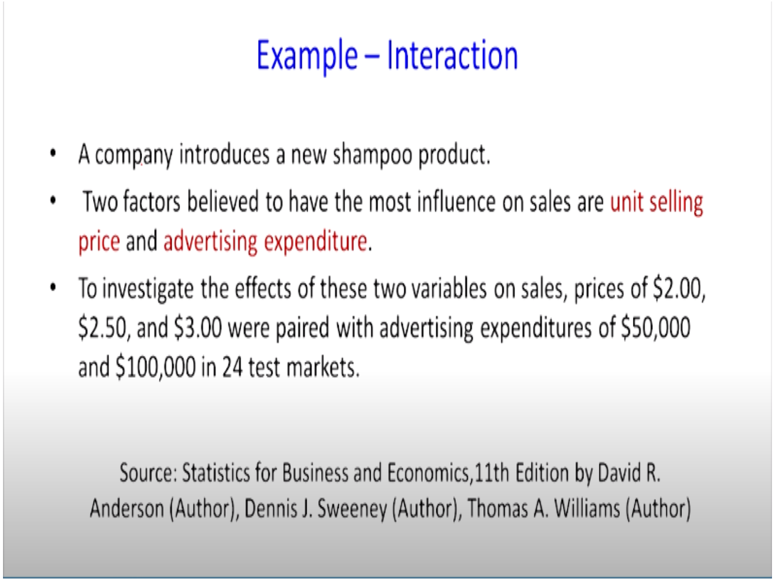
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

- In this second-order model, the variable $z_5 = x_1 x_2$ is added to account for the potential effects of the two variables acting together.
- This type of effect is called **interaction**.

First, we will see what is interaction. If the original dataset consist observation for y and two independent variable x1 and x2, we can develop a second-order model with two predictor variables setting z1 = x1, z2=x2, z3=x1 square, and z4 is x2 square and z5 is the fifth independent variable that is x1 and x2 in the general linear model equation. So, when you bring this interaction, our regression equation will become like this, y = beta 0 + beta 1x1 + beta 2x2 + square of the first independent variable x1 square plus the square of the second independent variable x2 square and interaction.

In this second-order model, this is called as second order regression model, the variable z5, that is x1 and x2 is added to the account for the potential effect of two variables acting together. This type of effect is called interaction. So this term we say as interaction.

(Refer Slide Time: 02:19)



Example – Interaction

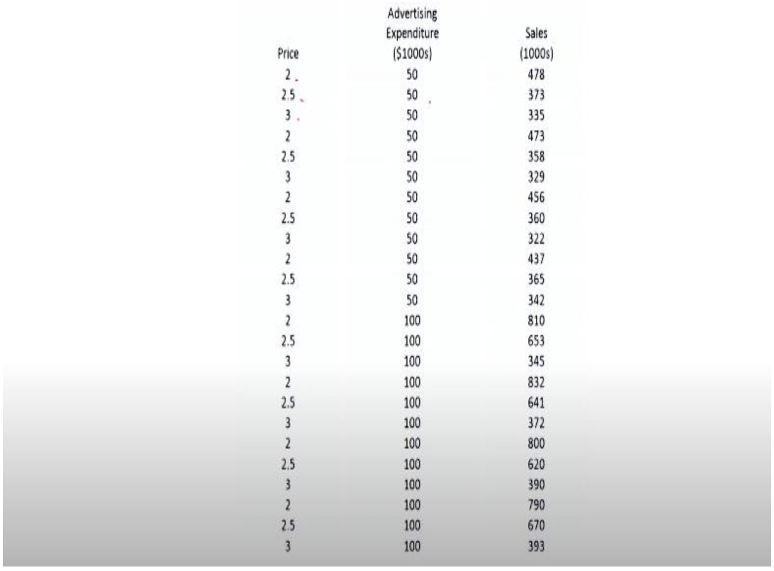
- A company introduces a new shampoo product.
- Two factors believed to have the most influence on sales are unit selling price and advertising expenditure.
- To investigate the effects of these two variables on sales, prices of $2.00, $2.50, and $3.00 were paired with advertising expenditures of $50,000 and $100,000 in 24 test markets.

Source: Statistics for Business and Economics,11th Edition by David R. Anderson (Author), Dennis J. Sweeney (Author), Thomas A. Williams (Author)

We take one example problem to understand how to do interaction into the regression model. This problem is taken from this book statistics for business and economics, 11th edition. A company produces a new shampoo product, two factors believed to have the most influence on sales are unit selling price and advertising expenditure. So, there are two variables that is going to affect our sales. One variable is unit selling price, another variable is advertising expenditure.

These two variables are independent variable. The sales is the dependent variable. So, in this problem setting, there is one dependent variable, two independent variables. To investigate, the effect of these two variables on the sales, the prize of 2.5 dollar and 3 dollar were paired with advertising expenditure of 50,000 dollars and 100,000 dollars in 24 test markets. I will show you this dataset.

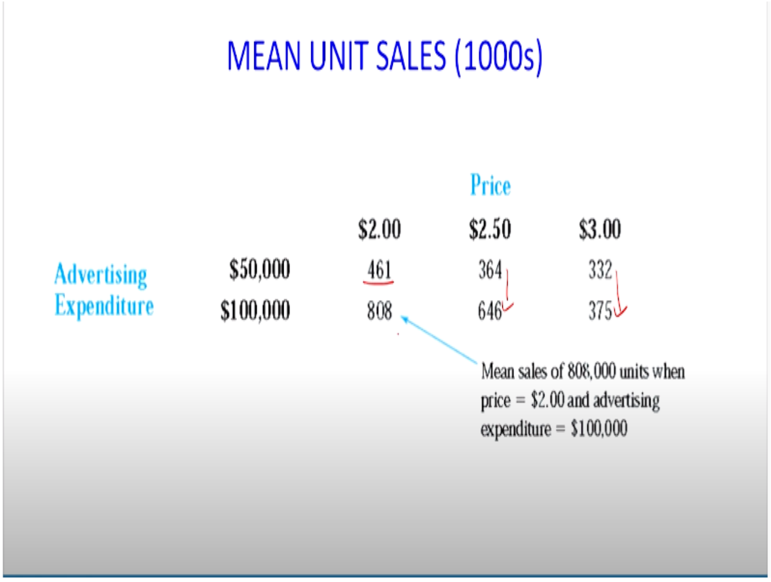
(Refer Slide Time: 03:24)



Price	Advertising Expenditure ($1000s)	Sales (1000s)
2	50	478
2.5	50	373
3	50	335
2	50	473
2.5	50	358
3	50	329
2	50	456
2.5	50	360
3	50	322
2	50	437
2.5	50	365
3	50	342
2	100	810
2.5	100	653
3	100	345
2	100	832
2.5	100	641
3	100	372
2	100	800
2.5	100	620
3	100	390
2	100	790
2.5	100	670
3	100	393

This dataset, you say that, there are 3 levels in prize, 2, 2.5, and 3. There are 2 level in the advertising expenditure. One is 50, another one is 100. So that there will be a 24 different alternatives. The last column is sales.

(Refer Slide Time: 03:44)



MEAN UNIT SALES (1000s)

		Price		
		$2.00	$2.50	$3.00
Advertising Expenditure	$50,000	461	364	332
	$100,000	808	646	375

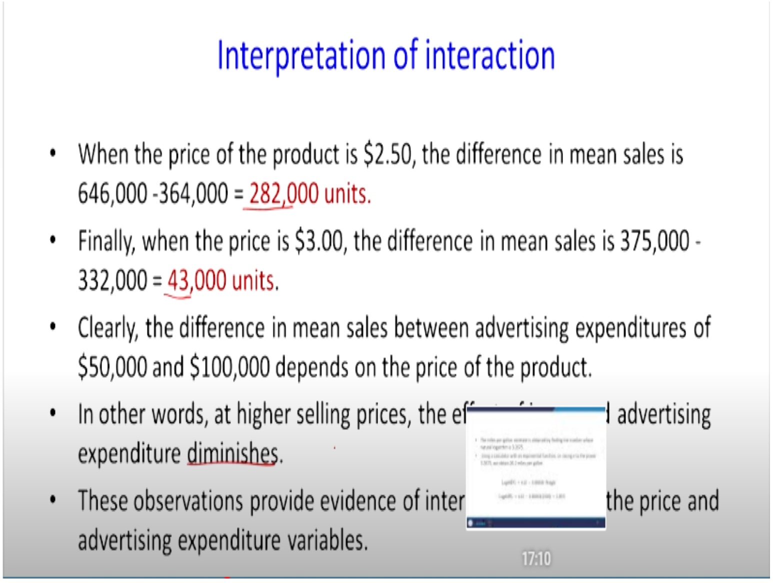
Mean sales of 808,000 units when price = $2.00 and advertising expenditure = $100,000

Now, we have made a summary of the previous table. What the summary says, when the price is 2 dollars, when the advertising expenditure is 50,000 dollars, this 461 says the mean sales. So, for example, in another sales, look at this one. When the price of the shampoo is 2 dollar, the expenditure is 100,000 dollars, this was the mean of all that combinations. So, the mean sales of 888,000 units when the price is 2 dollars and the advertising expenditure.

How it was done where there is a; 2 is there and you have to look at the corresponding sales value. The average of these four element is our (()) (04:36). Similarly, the cells is nothing but the mean of that level and the corresponding variable. Similarly, how we got the 461. When the price is 2, advertising expenditure is 50. Because the next one, it is going to the 100. So, the average of this value is 461. Now look at this table.

What it says that by keeping the selling prices to 2 dollars, when you increase the advertising expenditure, the mean value of the sales is increasing. Here it is increasing. The second case by keeping 2005 dollar as the prize, when you increase the expenditure 50,000 dollar to 100,000 dollar what is happening here, your sales is increasing. Here although the sales is increasing. This is one way. By looking at another way, when you find the difference between the 50,000 and 100,000 dollars that we will show you in the next slide, what will happen the difference, instead of increasing, it will start decreasing.

(Refer Slide Time: 06:18)



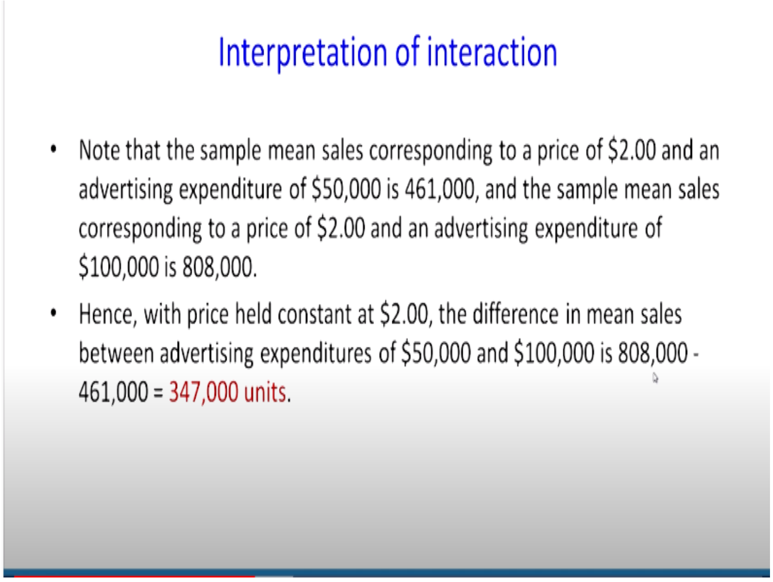
Interpretation of interaction

- When the price of the product is $2.50, the difference in mean sales is 646,000 -364,000 = 282,000 units.
- Finally, when the price is $3.00, the difference in mean sales is 375,000 - 332,000 = 43,000 units.
- Clearly, the difference in mean sales between advertising expenditures of $50,000 and $100,000 depends on the price of the product.
- In other words, at higher selling prices, the ef d advertising expenditure diminishes.
- These observations provide evidence of inter the price and advertising expenditure variables.

This is the explanation for our previous slide. When the price of the product is 2.5 dollars, the difference in mean sale is when it is 2.5 dollar, so the difference in mean sale is 646,000 dollar minus 364,000 dollars, this was your 282,000 dollars, 3 dollars, the difference in mean sale is 43. So what is happening, the difference is mean sale is decreasing. Clearly, the difference in mean sales between advertising expenditures of 50,000 dollars and 100,000 dollars depends on the price of the product.

In other words, at higher selling prices, the effect of increased advertising expenditure diminishes. Actually what it has to do, when the price of the product increases, then we go for increasing the advertising expenditure, the sales also has to increase, but it is not happening so. So, what is happening when the selling price is increasing, the effect of advertising expenditure on the sale diminishes. These observations provide evidence of interaction between the price and the advertising expenditure variables.

(Refer Slide Time: 07:15)



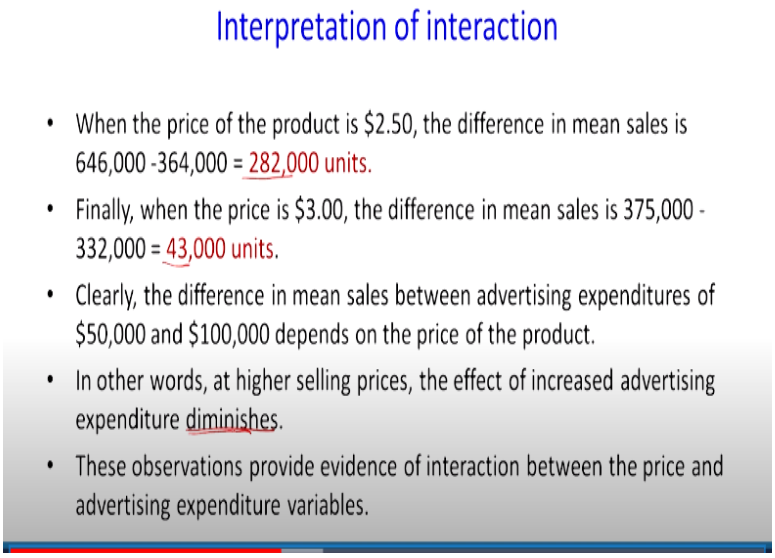
Interpretation of interaction

- Note that the sample mean sales corresponding to a price of $2.00 and an advertising expenditure of $50,000 is 461,000, and the sample mean sales corresponding to a price of $2.00 and an advertising expenditure of $100,000 is 808,000.
- Hence, with price held constant at $2.00, the difference in mean sales between advertising expenditures of $50,000 and $100,000 is 808,000 - 461,000 = 347,000 units.

I am going to interpret this mean unit sale against advertising expenditure. Note that the sample mean sales corresponding to the price of 2 dollars and an advertising expenditure of 50,000 dollars is 461,000, and the sample mean sales corresponding to the price of 2 dollars, and the advertising is 808 dollars. I am referring to this 461 and 808. Hence the prize held constant 2 dollars.

The difference in the mean sales between advertising expenditures 50,000 dollars and 100,000 dollars is 808,000 dollars minus 461,000 dollars, the difference is 347,000. We will go to the next column.

(Refer Slide Time: 08:03)



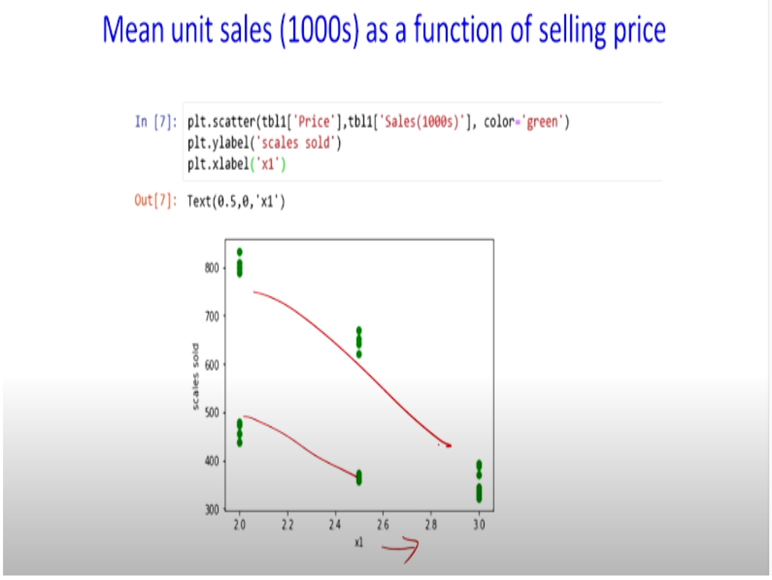
Interpretation of interaction

- When the price of the product is $2.50, the difference in mean sales is 646,000 -364,000 = 282,000 units.
- Finally, when the price is $3.00, the difference in mean sales is 375,000 - 332,000 = 43,000 units.
- Clearly, the difference in mean sales between advertising expenditures of $50,000 and $100,000 depends on the price of the product.
- In other words, at higher selling prices, the effect of increased advertising expenditure diminishes.
- These observations provide evidence of interaction between the price and advertising expenditure variables.

When the price of the product is kept 2.50 dollars, the difference in mean sale is 282,000 units. Finally, when the price is 3 dollars, the difference in mean sale is 43,000 units. Clearly, the difference in mean sales between the advertising expenditure of 50,000 dollars and 100,000 dollars depends on the price of the product. In other words, at higher selling prices, the effect of increased advertising expenditure diminishes.

What it happens, when the price increases, when the advertising expenditure also increases, the sales has to increase, but instead of increasing, it starts decreasing. So, the expenditure diminishes. These observations provide evidence of interaction between the price and advertising expenditure variables.

(Refer Slide Time: 08:57)

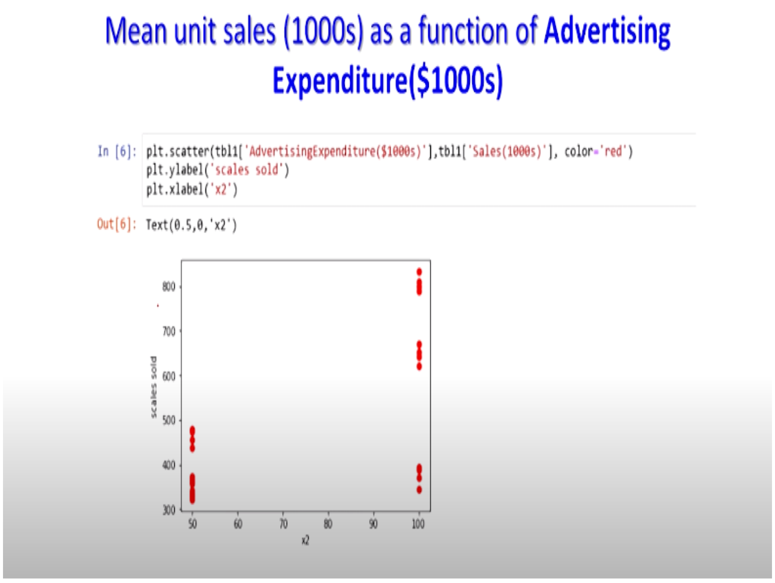


Mean unit sales (1000s) as a function of selling price

```
In [7]: plt.scatter(tbl1['Price'],tbl1['Sales(1000s)'], color='green')
        plt.ylabel('scales sold')
        plt.xlabel('x1')

Out[7]: Text(0.5,0,'x1')
```

First, we will do the Python code for that. I have imported the data, the prices 2, 2.5, 3 level, there is advertising expenditure is 50 and 100. The sale is in terms of unit, that is 478 and so on. When you plot this scatterplot, see that there are three different levels. What it says that, whenever the price of the product is increasing, the sales it is not increasing. The sales you see that there is a decreasing trend. It has to increase. Why it is decreasing, so there is no effect of amount spent on expenditure when x1 increases.

(Refer Slide Time: 09:47)

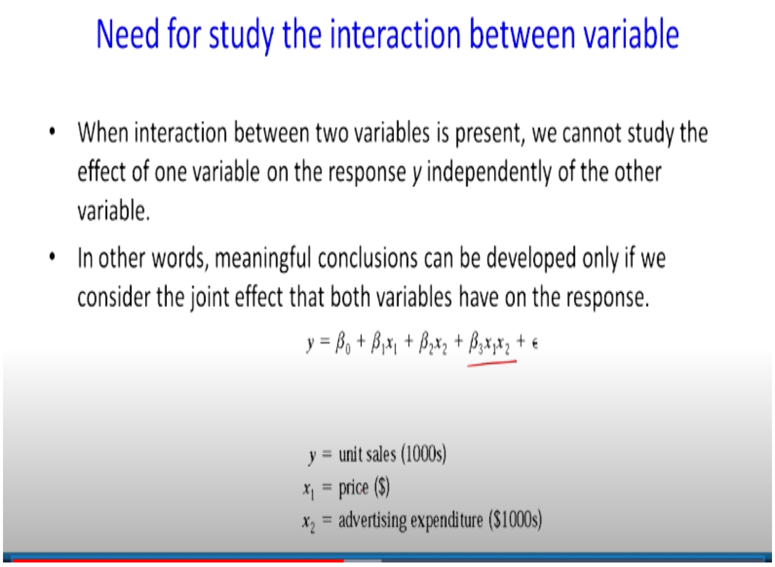


Mean unit sales (1000s) as a function of Advertising Expenditure($1000s)

```
In [6]: plt.scatter(tbl1['AdvertisingExpenditure($1000s)'],tbl1['Sales(1000s)'], color='red')
        plt.ylabel('scales sold')
        plt.xlabel('x2')

Out[6]: Text(0.5,0,'x2')
```

So, this graph shows that there is effect of interaction. So this scatterplot shows between the advertising expenditure, there are two level, one is 50,000 dollars, another one is 100,000 dollars. The y-axis is the number of scales sold.

(Refer Slide Time: 10:00)



Need for study the interaction between variable

- When interaction between two variables is present, we cannot study the effect of one variable on the response *y* independently of the other variable.
- In other words, meaningful conclusions can be developed only if we consider the joint effect that both variables have on the response.

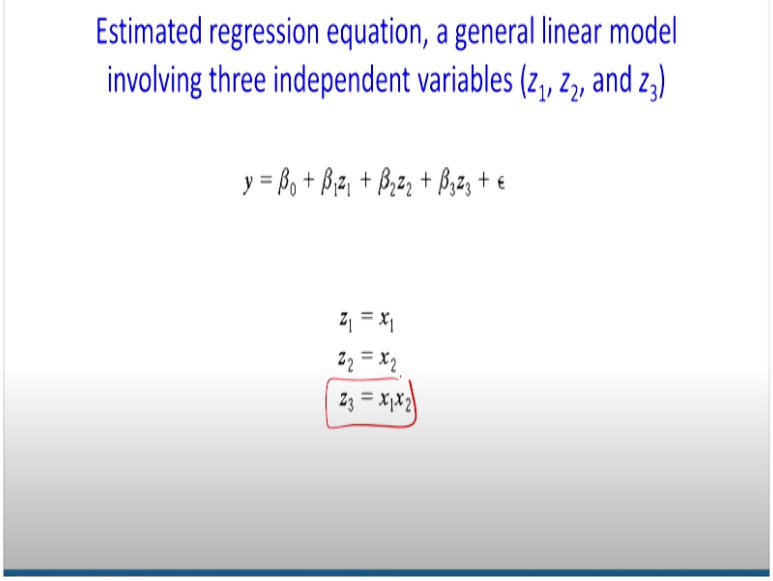
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \underline{\beta_3 x_1 x_2} + \epsilon$$

y = unit sales (1000s)
x_1 = price ($)
x_2 = advertising expenditure ($1000s)

In our summary table and our scatterplot, we have realized there is interaction between x1 and x2. When interaction between two variables are present, we cannot study the effect of one variable on the response variable y independently of each variable. In other words, a meaningful conclusion can be developed only if we consider the joint effect of both the variables having the response.

So, what is the joint effect is this x1 and x2. We have realized in that summary table, that there is a interaction between both the variable x1 and x2. Here y is the unit sales, in terms of units, x1 is the price, it has three level. x2 is advertising expenditure, it has two levels.

(Refer Slide Time: 10:54)



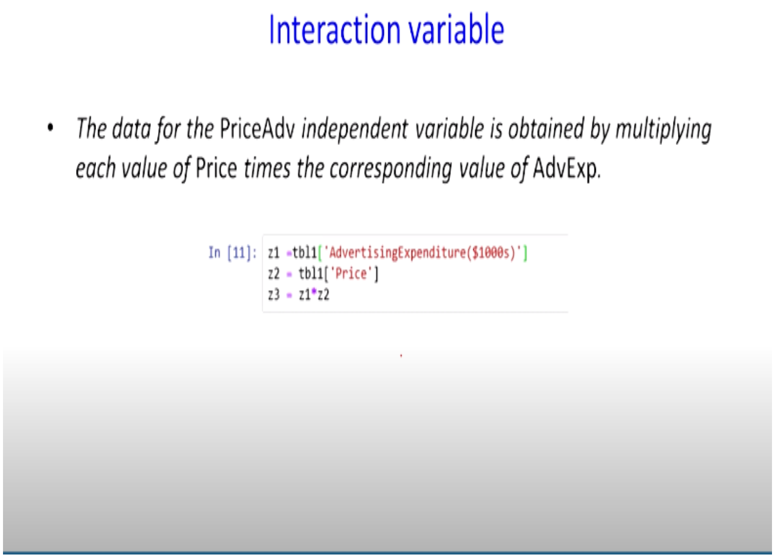
Estimated regression equation, a general linear model involving three independent variables (z_1, z_2, and z_3)

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon$$

$$z_1 = x_1$$
$$z_2 = x_2$$
$$\boxed{z_3 = x_1 x_2}$$

Now, the estimated regression equation, a general linear model involving 3 independent variables, that is z1, z2, and z3. Here, the z1 is x1, z2 is x2, and z3 is this interaction variable, that is x1 multiplied by x2. What we have to do, apart from x1 and x2, we have to introduce another variable, that is product of two variable x1 and x2.

(Refer Slide Time: 11:21)



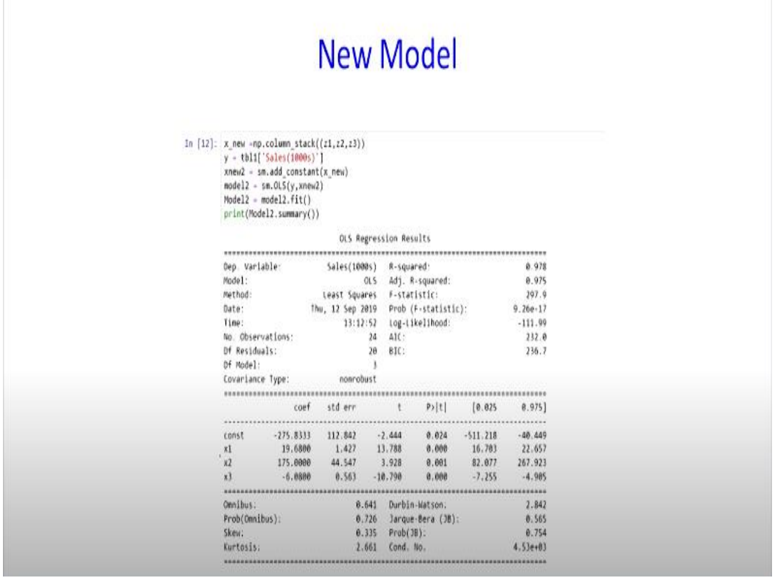
Interaction variable

- *The data for the* PriceAdv *independent variable is obtained by multiplying each value of* Price *times the corresponding value of* AdvExp.

```
In [11]: z1 =tbl1['AdvertisingExpenditure($1000s)']
         z2 = tbl1['Price']
         z3 = z1*z2
```

Now, we will create a new variable, that is z3, that is the product of z1 and z2. The data for price advertisement of independent variable is obtained by multiplying each value of the price times, the corresponding value of advertising expenditure. So, both variable z1 and z2 has to be multiplied, that will be our new variable.

(Refer Slide Time: 11:43)



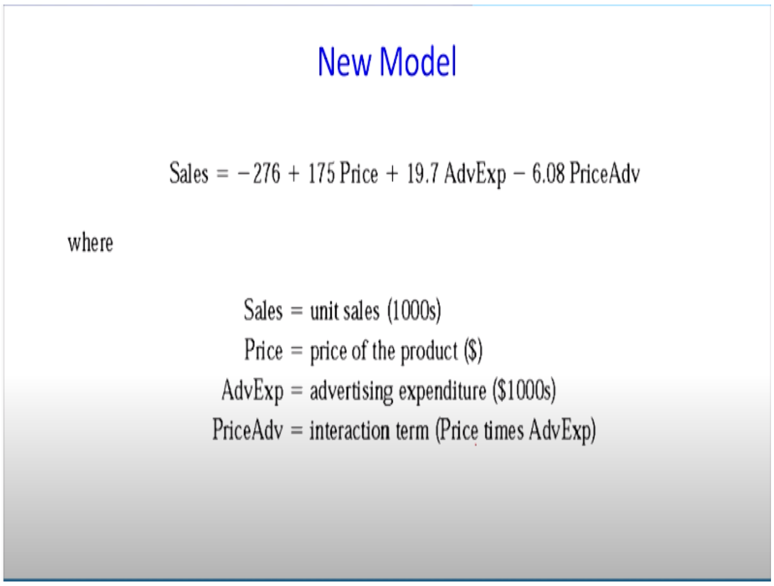
New Model

```
In [12]: x_new =np.column_stack((z1,z2,z3))
         y = tbl1['Sales(1000s)']
         xnew2 = sm.add_constant(x_new)
         model2 = sm.OLS(y,xnew2)
         Model2 = model2.fit()
         print(Model2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Sales(1000s)   R-squared:                       0.978
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                     297.9
Date:                Thu, 12 Sep 2019   Prob (F-statistic):           9.26e-17
Time:                        13:12:52   Log-Likelihood:                -111.99
No. Observations:                  24   AIC:                             232.0
Df Residuals:                      20   BIC:                             236.7
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -275.8333    112.842     -2.444      0.024    -511.218     -40.449
x1            19.6800      1.427     13.788      0.000      16.703      22.657
x2           175.0000     44.547      3.928      0.001      82.077     267.923
x3            -6.0800      0.563    -10.790      0.000      -7.255      -4.905
==============================================================================
Omnibus:                        0.641   Durbin-Watson:                   2.842
Prob(Omnibus):                  0.726   Jarque-Bera (JB):                0.565
Skew:                           0.335   Prob(JB):                        0.754
Kurtosis:                       2.661   Cond. No.                     4.53e+03
==============================================================================
```

After multiplying, now this is our output model for our interaction. So look at the R-square. R-square is 0.978, x1 is our one independent variable, x2 is another independent variable. This x3 is the interaction.

(Refer Slide Time: 12:53)



New Model

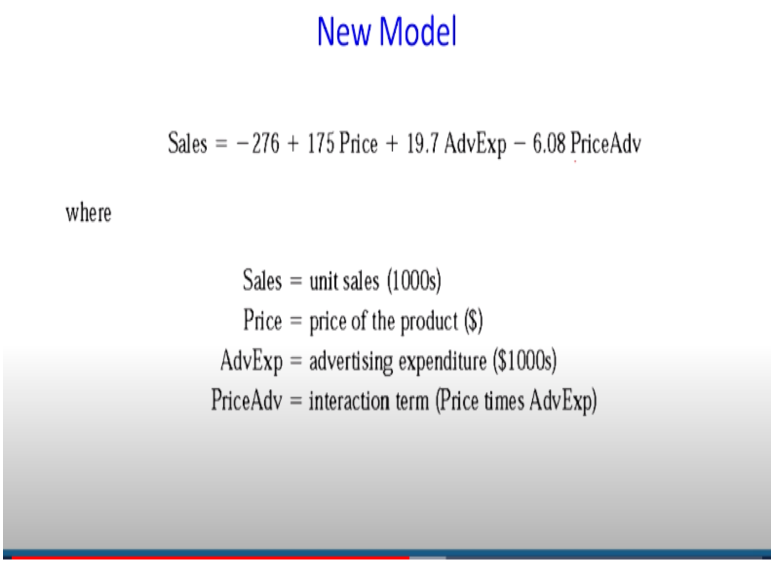
$$\text{Sales} = -276 + 175\,\text{Price} + 19.7\,\text{AdvExp} - 6.08\,\text{PriceAdv}$$

where

Sales = unit sales (1000s)
Price = price of the product ($)
AdvExp = advertising expenditure ($1000s)
PriceAdv = interaction term (Price times AdvExp)

So, for this how we can write the regression equation. -276+175 price, that is our x2, then 19.7 advertising expenditure, that is our x1. The third one is our interaction variable, that is x3, that is -6.08. Look at the p value of f statistics, that is very low, the overall model is significant. For all variables, x1, x2 and interaction variables, look at the p-value this one, all are less than 0.05, so each independent variable is significant variables.

(Refer Slide Time: 12:46)



New Model

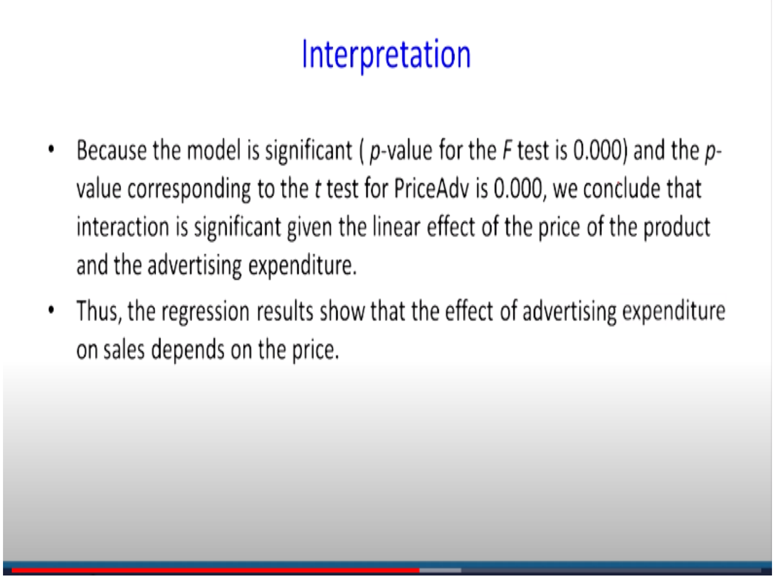
$$\text{Sales} = -276 + 175\,\text{Price} + 19.7\,\text{AdvExp} - 6.08\,\text{PriceAdv}$$

where

$$\begin{aligned} \text{Sales} &= \text{unit sales (1000s)} \\ \text{Price} &= \text{price of the product (\$)} \\ \text{AdvExp} &= \text{advertising expenditure (\$1000s)} \\ \text{PriceAdv} &= \text{interaction term (Price times AdvExp)} \end{aligned}$$

So what is the new model now. Sales equal to -276+175 price + 19.7 AdvExp – the price end advertisement. This is our interaction term. How to interpret this.

(Refer Slide Time: 13:05)

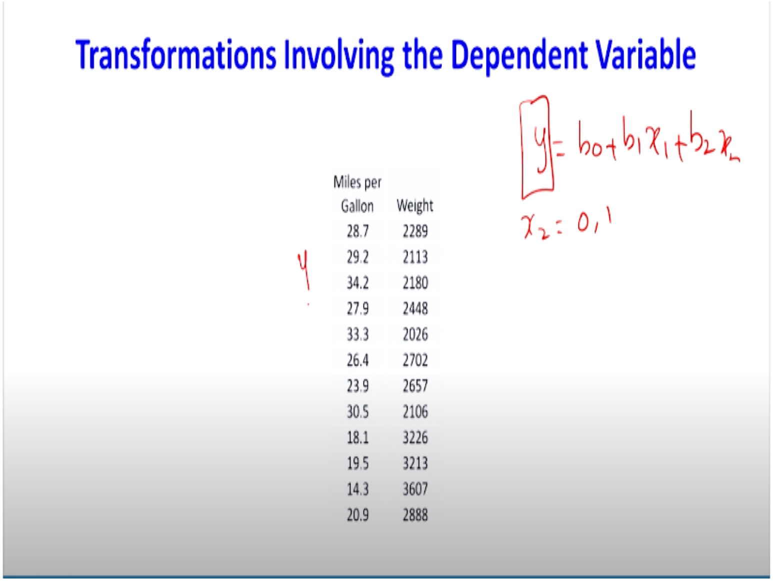


Interpretation

- Because the model is significant (*p*-value for the *F* test is 0.000) and the *p*-value corresponding to the *t* test for PriceAdv is 0.000, we conclude that interaction is significant given the linear effect of the price of the product and the advertising expenditure.
- Thus, the regression results show that the effect of advertising expenditure on sales depends on the price.

Because the model is significant, the p-value for the F test is 0.0000 and the p value corresponding to the t test PriceAdv is 0.00, we conclude that interaction is significant given the linear effect of the price of the product and the advertising expenditure. Thus, this regression results shows that the effect of advertising expenditure on sales depends on the price.

(Refer Slide Time: 13:18)



Transformations Involving the Dependent Variable

$$\boxed{y} = b_0 + b_1 x_1 + b_2 x_2$$

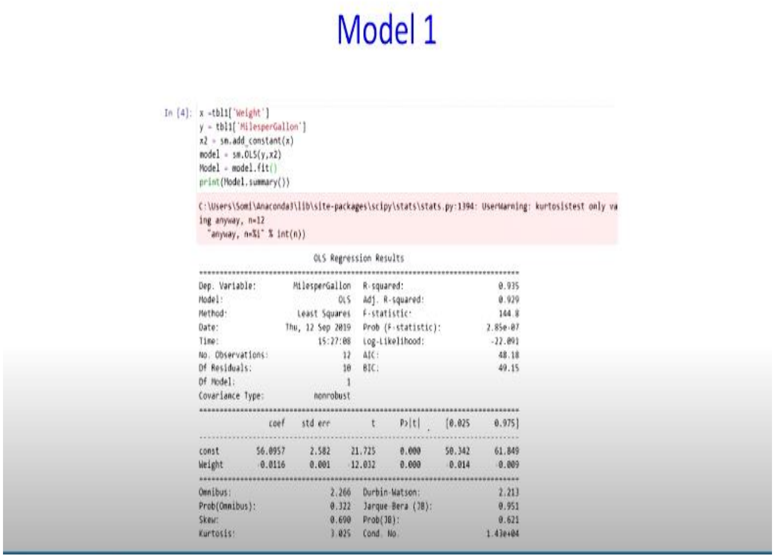
$x_2 = 0, 1$

Miles per Gallon	Weight
28.7	2289
29.2	2113
34.2	2180
27.9	2448
33.3	2026
26.4	2702
23.9	2657
30.5	2106
18.1	3226
19.5	3213
14.3	3607
20.9	2888

y

So far, we have done some transformations only on independent variable. For example, y = b0 +b1x1 + b2x2. Suppose, x2 is a categorical variable, assume that. What you have done, if x2 can have only two variables, say 0, 1 gender. So we have done a modification. We have introduced a dummy variable and we have done the model. Now, there may be a situation that your y variable also has to be transformed.

(Refer Slide Time: 14:23)



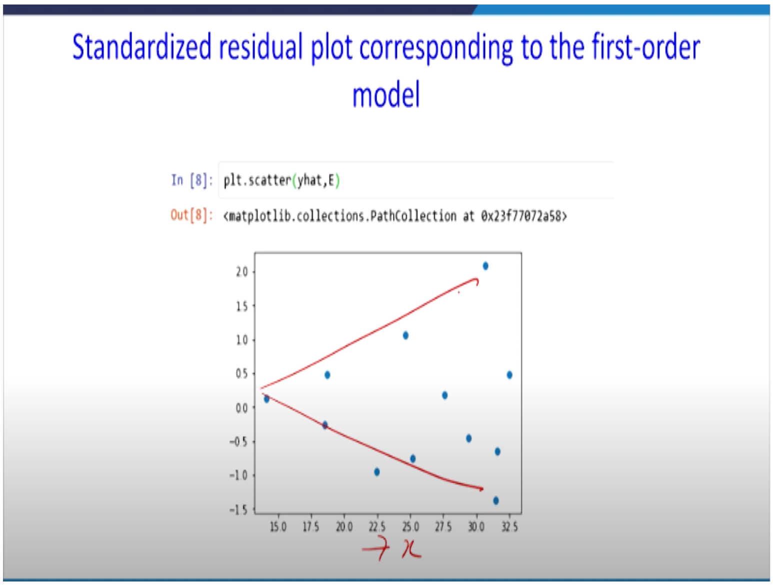
Model 1

```
In [4]: x =tbl1['Weight']
        y = tbl1['MilesperGallon']
        x2 = sm.add_constant(x)
        model = sm.OLS(y,x2)
        Model = model.fit()
        print(Model.summary())
```

```
C:\Users\Somi\Anaconda3\lib\site-packages\scipy\stats\stats.py:1394: UserWarning: kurtosistest only va
ing anyway, n=12
  "anyway, n=%i" % int(n))
                            OLS Regression Results
==============================================================================
Dep. Variable:          MilesperGallon   R-squared:                       0.935
Model:                             OLS   Adj. R-squared:                  0.929
Method:                  Least Squares   F-statistic:                     144.8
Date:                 Thu, 12 Sep 2019   Prob (F-statistic):           2.85e-07
Time:                         15:27:08   Log-Likelihood:                -22.091
No. Observations:                   12   AIC:                             48.18
Df Residuals:                       10   BIC:                             49.15
Df Model:                            1
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         56.0957      2.582     21.725      0.000      50.342      61.849
Weight        -0.0116      0.001    -12.032      0.000      -0.014      -0.009
==============================================================================
Omnibus:                         2.266   Durbin-Watson:                   2.213
Prob(Omnibus):                   0.322   Jarque-Bera (JB):                0.951
Skew:                            0.690   Prob(JB):                        0.621
Kurtosis:                        3.025   Cond. No.                     1.43e+04
```

Suppose, the miles per gallon, that is your y variable. This weight is your independent variable. Suppose, if you do a regression analysis for this dataset, see that there is a negative relationship. When the weight increases, the miles per gallon decreases. There is a negative relationship for scatterplot. Now, when you look at the Regression model, the Regression model, y is equal to 56.0957 – 0.0116. This is significant.

(Refer Slide Time: 14:40)



Standardized residual plot corresponding to the first-order model

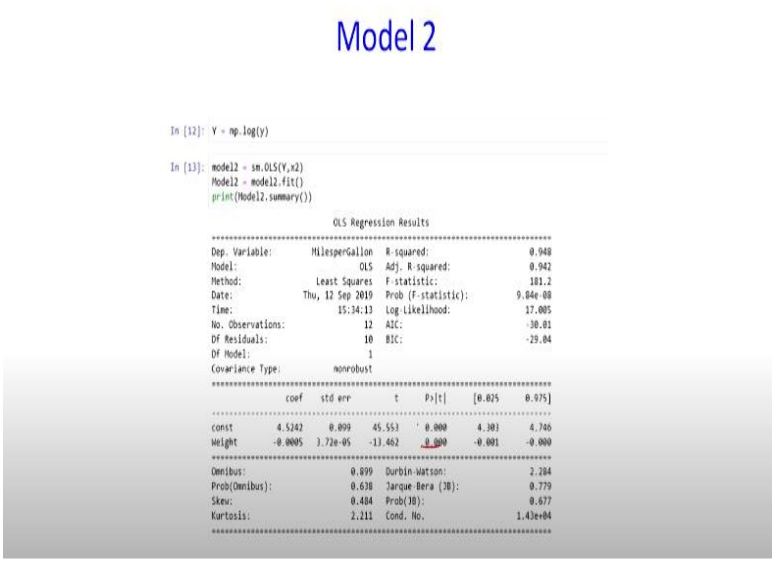
```
In [8]: plt.scatter(yhat,E)

Out[8]: <matplotlib.collections.PathCollection at 0x23f77072a58>
```

Now, look at the standard residual plot. First, we will predict the residuals, then we will standardize. Then, we will predict the y value. Now, we will draw a graph between predicted y hat and standardize residual. When we look at this, you see that there is a conical relationship. What is happening, whenever the value of x increases, the variance is not constant. This is violating our model.

What is the model? When the variance or the error term should be same for all value of x, but now what is happening, when the value of x is increasing, the variance also increases. So, it is not fitting to our assumption of regression equation. We are going to take log of y, so the y is there, so we are going to take log of y values, b0+b1x1. This is going to be the same. Our independent variable will not be disturbing.

But for the dependent variable, we are going to take the log of; the purpose of taking log is that the error term, instead of getting this conical shape, we may get a kind of a rectangular shape. So, that means the variance of the error terms is going to be same.

(Refer Slide Time: 16:02)

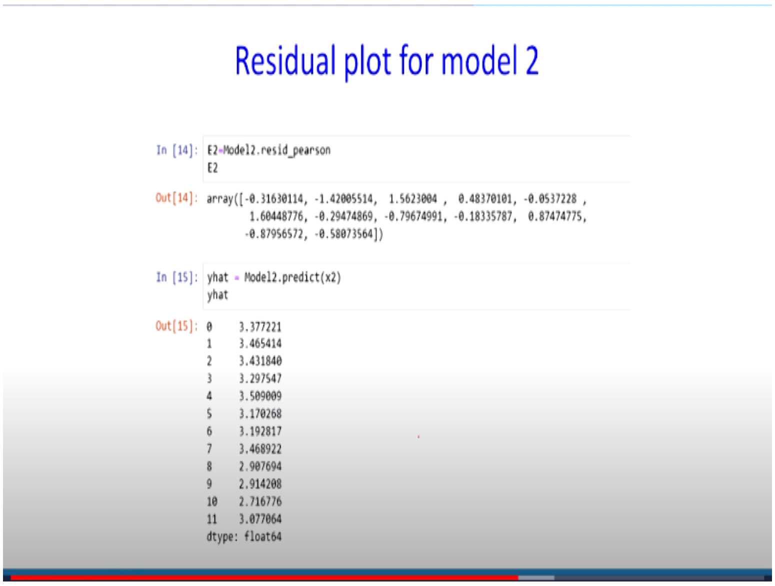


Model 2

```
In [12]: Y = np.log(y)

In [13]: model2 = sm.OLS(Y,x2)
         Model2 = model2.fit()
         print(Model2.summary())

                            OLS Regression Results
==============================================================================
Dep. Variable:         MilesperGallon   R-squared:                       0.948
Model:                            OLS   Adj. R-squared:                  0.942
Method:                 Least Squares   F-statistic:                     181.2
Date:                Thu, 12 Sep 2019   Prob (F-statistic):           9.84e-08
Time:                        15:34:13   Log-Likelihood:                 17.005
No. Observations:                  12   AIC:                            -30.01
Df Residuals:                      10   BIC:                            -29.04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.5242      0.099     45.553      0.000       4.303       4.746
Weight        -0.0005   3.72e-05    -13.462      0.000      -0.001      -0.000
==============================================================================
Omnibus:                        0.899   Durbin-Watson:                   2.284
Prob(Omnibus):                  0.638   Jarque-Bera (JB):                0.779
Skew:                           0.484   Prob(JB):                        0.677
Kurtosis:                       2.211   Cond. No.                     1.43e+04
==============================================================================
```

First, what you have done, I have taken log of all dependent variable, that is I call it Y. Now, this log of Y is taken as the new dependent variable. After substituting this, you look at the new variable, one is weight, the R square is increased, and F is good, the model is okay. Now, we will go for the residual plot for this.

(Refer Slide Time: 16:24)

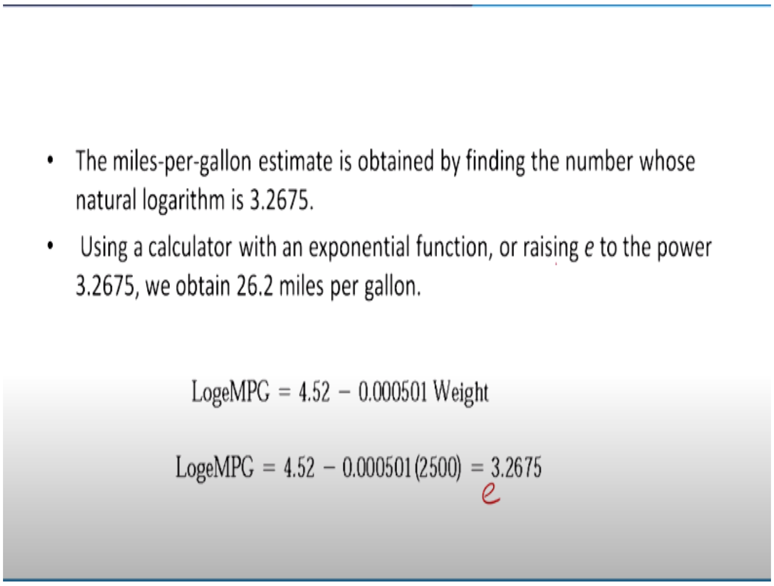


Residual plot for model 2

```
In [14]: E2=Model2.resid_pearson
         E2

Out[14]: array([-0.31630114, -1.42005514,  1.5623004 ,  0.48370101, -0.0537228 ,
                 1.60448776, -0.29474869, -0.79674991, -0.18335787,  0.87474775,
                -0.87956572, -0.58073564])

In [15]: yhat = Model2.predict(x2)
         yhat

Out[15]: 0     3.377221
         1     3.465414
         2     3.431840
         3     3.297547
         4     3.509009
         5     3.170268
         6     3.192817
         7     3.468922
         8     2.907694
         9     2.914208
         10    2.716776
         11    3.077064
         dtype: float64
```

When you go for residual plot against y hat, this is our standardize residual, so what is happening.

(Refer Slide Time: 16:34)



- The miles-per-gallon estimate is obtained by finding the number whose natural logarithm is 3.2675.
- Using a calculator with an exponential function, or raising *e* to the power 3.2675, we obtain 26.2 miles per gallon.

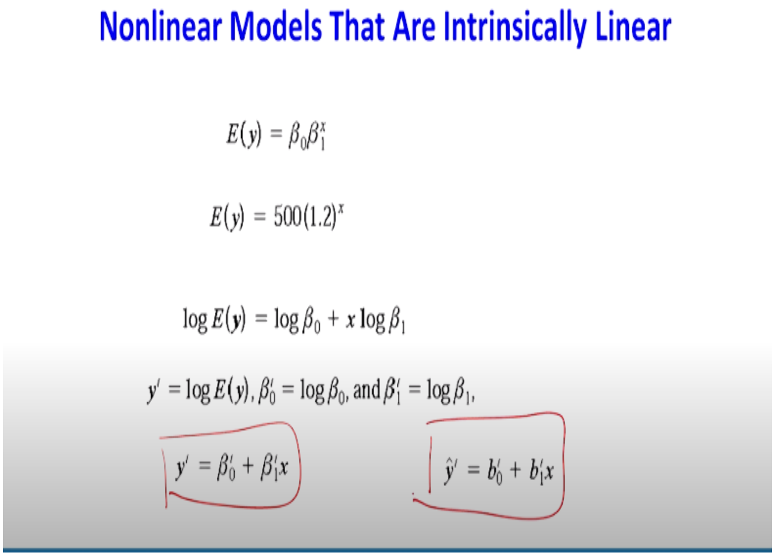
$$\text{LogeMPG} = 4.52 - 0.000501 \text{ Weight}$$

$$\text{LogeMPG} = 4.52 - 0.000501(2500) = \underset{e}{3.2675}$$

Now, there is no conical shape, there is rectangular shape is appearing, but you should be very careful while interpreting the answer because it is not actual y, it is log of y. So, when you substitute the values into this, the miles per gallon estimate is obtained by finding the number whose natural logarithm is 32.675. So what you have to do, suppose if you substitute weight is 2500, we are getting the log of y value, that is miles per gallon is 3.26.

If you want to know the actual value, you take e to the power 3.26, that is why to bring you to normal term, you have to take natural logarithm is 3.26 using a calculator or any exponential function using our Python, we have to rising e to the power 3.26, you will get 26.2 miles per gallon, that is your original y values.

(Refer Slide Time: 17:37)



Nonlinear Models That Are Intrinsically Linear

$$E(y) = \beta_0\beta_1^x$$

$$E(y) = 500(1.2)^x$$

$$\log E(y) = \log \beta_0 + x \log \beta_1$$

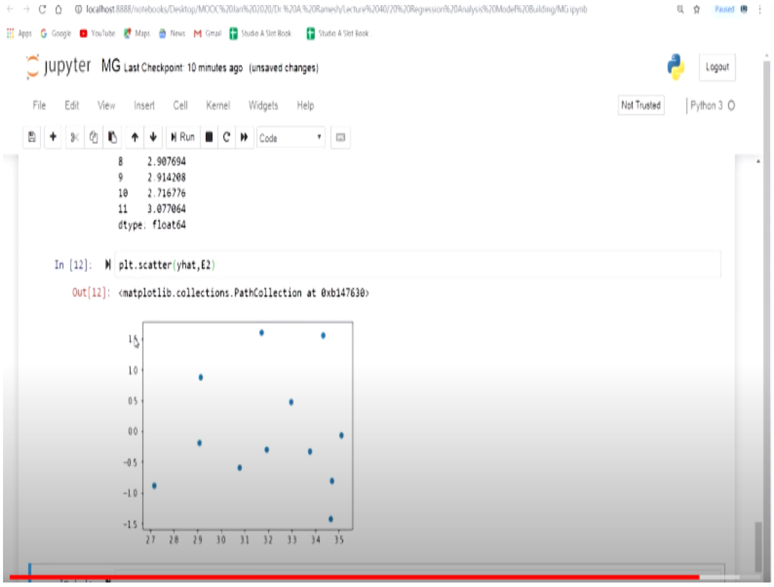
$$y' = \log E(y), \beta_0' = \log \beta_0, \text{ and } \beta_1' = \log \beta_1,$$

$$\boxed{y' = \beta_0' + \beta_1' x} \qquad \boxed{\hat{y}' = b_0' + b_1' x}$$

There are some more nonlinear model. How to do that one, I will explain. Suppose, there may be a nonlinear relationship that the power is there, beta0 + beta1x, so the expected value, suppose if you substitute beta0 is 500, it is 1.2 to the power x. So for this kind of model, you take log of both the sides. It will become log of expected value of y. So log of beta0 + log of x log beta 1. Here the constant term is, this can be written y dash equal to beta0 dash, beta1 dash x.

So, the y dash is the log of ey, beta 1 is log of b0 and beta1 dash, log of beta1. This equation can be estimated with the sample of this Regression equation, but we should be very careful while interpreting, you have to remember it has to be brought into the original term.

(Refer Slide Time: 18:37)



```
8     2.907694
9     2.914208
10    2.716776
11    3.077064
dtype: float64
```

```
In [12]:  plt.scatter(yhat,E2)
Out[12]: <matplotlib.collections.PathCollection at 0xb147630>
```

localhost:8888/notebooks/Desktop/MOOC%20Jan%202020/Dr.%20A.%20Ramesh/Lecture%2040/20%20Regression%20Analysis%20Model%20Building/MG.ipynb
Apps
Google
YouTube
Maps
News
Gmail
Studio A Slot Book
Studio A Slot Book
jupyter
MG Last Checkpoint: 10 minutes ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
27
28
29
30
31
32
33
34
35

Now, we are going to do the interaction among independent variable with y with the help of this Python code. So I have imported, the file name is Tyler. So, this was our portion of our file name. First, we should do the scatterplot. So, what is happening here, when the price of the product is increasing, see that the y variable, it is the number of scales sold, it is decreasing. So, this table is suggesting that there is a interaction effect between the prize and the dependent variable.

Look at this. These are another dependent variable, that is advertising expenditure. This also shows that whenever the advertising expenditure is increasing, the car sold is increasing, but it is not linearly increasing because there seems to be some other variable, which is affecting the advertising expenditure. That variable is nothing but the prize. From our scatterplot, plot number one and plot number 2, we realize that there is interaction effect.

So the two variable that is z1 and z2 are multiplied. What is our z1 variable, that is our advertising expenditure, our z2 variable is price, so new variable is z1 multiplied by z2. We will do this one. Now, the third variable, that is new variable taken as another dependent variable. Now, there are 3 variable, one is for advertising expenditure, another variable for prize, the third one is interaction among these two.

So, when you run this model, we are getting all these three variables, that is x1, x2, and x3 is our interaction variable. All are significant. So, we can say that there is interaction effect between x1 and x2. See, our R square is better, 0.978, our fp value also very less, and individual significance of each independent variable is also less than 0.05, also variables are significant.

Now, in our class I have explained one more problem, that is how to do transformation of our dependent variable. So, I have imported the necessary libraries with the data file is this one. So, here the weight is independent variable, but the miles per gallon is dependent variable. So, when you do the scatterplot between these two, there seems to be a negative relationship. When you do a simple linear regression by taking x’s weight independent variable, y is the miles per gallon, we are getting this one.

So even though the model is significant when you go for residual plot. What is happening between standardize residual and predicted value? there is a conical shape is there. So, what this implies that the value of x increases, the variance or the error term is not the same. It is getting increased. This is violation of Regression model. To compensate this, we are going to do the transformation, log transformation of our dependent variable. After log transformation when you do again, there is a regression equation.

So, you look at this the third one, now the new dependent variable is the log of y, so the independent variable. So, this one, we will go for a standardized residual plot. Now, what is happening when you go for that, now there is no conical relationship. Then we can say that the log of transformation of dependent variable is correct, you should go for log transformation of our dependent variable.

In this lecture, we have seen how to incorporate if there is interaction among variable, how to incorporate this interaction into our Regression model. We have taken one sample example, when we are plotting the summary table, we have realized that there is a interaction between two variables, then we have taken the product of the two variable that introduces a third variable, then we have done a multiple regression model, we realize that the interaction is significant.

In another problem, what we have seen in this class is, generally we do the transformation in the independent variable, but sometimes, we need to do the transformation for the dependent variable also. So what transformation we have done, we have done log of our y value. Before doing log of y value, we have realized that the variance of the error is not the same. After doing the log transformation, we have realized that the variance of the error term is same, then we have accepted that taking log of our dependent variable is correct. Thank you.