Machine Learning with Python

Linear Regression

Cyrus Lentin

Linear Regression

Linear Regression Analysis Is Used To:

- Predict The Value Of A Dependent Variable Based On The Value Of At Least One Independent Variable
- Explain The Impact Of Changes In An Independent Variable On The Dependent Variable
- Dependent Variable
  The Variable We Wish To Predict Or Explain
- Independent Variable
  The Variable Used To Predict Or Explain The Dependent Variable
- It Is Used To Understand A Phenomena, Make Predictions, And/Or Test Hypotheses
- It Is One Of The Most Commonly Used Tools For Business Analysis
- It Is Easy To Use And Applies To Many Situations

Important

- Independent Variables Are All Continuous Or Categoric Numeric Values
- Dependent Variables Are Also Continuous Numeric Values

Examples

- House Price & Area Of The House
- Sales Of Product & Rate Per Unit

Linear Regression – Applications

- Human Resource - Salary Estimate
  Predicting or estimating salary of a person based on set of attributes such as years of experience, level of education, industry of work, previous job salary etc
- Human Resources - Churn
  Considering high level of employee churn, multiple regression based model to estimate months of stickiness (or job with a new employer) at the time of recruitment based on candidate attributes
- Human Resources - Resource Demand
  Forecasting or Demand Estimation for each of the technology skills; levels of bench in most of the big IT services provider is important level to get project & deliver but also add to the cost; an accurate estimation of demand by skills could be important measures to manage requirements at right cost
- Real Estate - House - Price Prediction
  Predicting House Prices considering house, locality and builder characteristics
- Real Estate - House Demand Forecast
  Developing a forecasting model to find volume of houses on sales in a month given economic factors, seasonality and other dimensions
- Retailer - Sales Volume & Return On Investment
  Finding out drivers of retail product sales as a function of spend across media channels, economic factors and competitor actions

Linear Regression – Applications

- Banking/Financial Services - Customer Value Estimation
  Considering customer level attributes, estimating short value of the customers
- Banking/Financial Services - Spend Value at a Customer
  Spend on Credit Card is a strong indicator of customer engagement on the card and whether the card is a front of wallet card; Predicting Spend value of card holder could help the product and marketing teams in engaging the customers with an appropriate treatment strategy
- Banking/Financial Services - Balance In Flow into Transaction or Saving Account
  Predicting amount of balance expected to be deposited into customers' transaction and saving account using customer level characteristics
- Banking/Financial Services - Drivers of New Account Volume
  Building Marketing or Media Mix Model to find economic, advertisement spend (across media or channels), competitor and offer related variables impacting new account open volume in a week
- Insurance/Financial Services - Claim Amount Estimation
  Insurance providers charge premium based on estimated claim amount for the target group of the customers; Claim could be against Motor, Home or Pet Policy; also, the estimated claim amount could be used for operational cash reserve calculations
- Banking/Financial Services - Revenue Regression Model
  Predicting revenue of customers and identifying parameters which are linked to increased revenue of the customers; this helps business bankers in realigning the priority and focus

Correlation

- The correlation coefficient is always a value between -1 and +1.
- It tells you how strongly two variables are related to each other.
- We use the corr() function to find the correlation coefficient between two variables.
- A correlation coefficient of +1 indicates a perfect positive correlation.
  As variable X increases, variable Y increases. As variable X decreases, variable Y decreases.
- A correlation coefficient of -1 indicates a perfect negative correlation.
  As variable X increases, variable Z decreases. As variable X decreases, variable Z increases.
- A correlation coefficient near 0 indicates no correlation.
  As variable X increases, variable W may increase or decrease.
  As variable X decreases, variable W may increase or decrease.
- A general way to interpret the calculated correlation coefficient value is as follows:
  - 0.0 to 0.2 / 0.0 to -0.2 – Very weak to negligible correlation
  - 0.2 to 0.4 / -0.2 to -0.4 – Weak correlation
  - 0.4 to 0.6 / -0.4 to -0.6 – Moderate correlation
  - 0.6 to 0.8 / -0.6 to -0.8 – Strong correlation
  - 0.8 to 1.0 / -0.8 to -1.0 – Very strong correlation

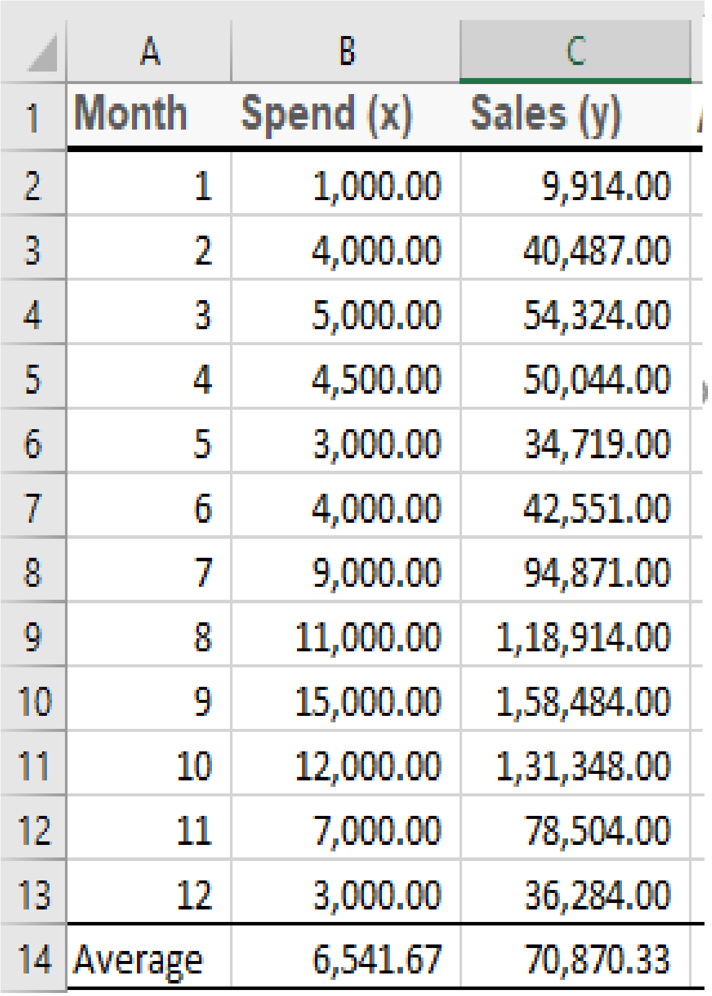
Simple Linear Regression

- Single Explanatory / Dependent Variable, Y
- Only One Independent Variable, X
- Correlation Must Exists Between X & Y
- Changes In Y Are Assumed To Be Directly Related To Changes In X
- Relationship Between X And Y Is Described By A Linear Function

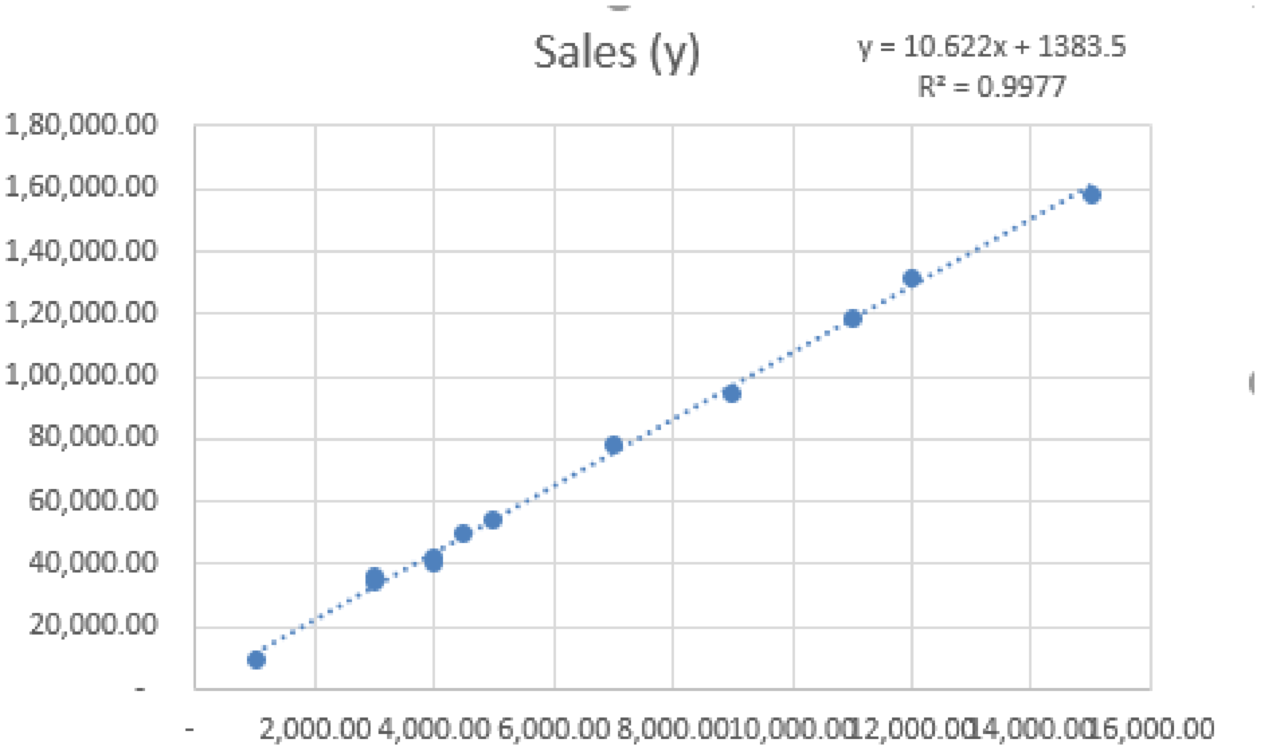
IMPORTANT

- Independent Variables must continuous numeric or categoric numeric values
- Dependent Variables are also continuous numeric values

Simple Linear Regression



	A	B	C
1	Month	Spend (x)	Sales (y)
2	1	1,000.00	9,914.00
3	2	4,000.00	40,487.00
4	3	5,000.00	54,324.00
5	4	4,500.00	50,044.00
6	5	3,000.00	34,719.00
7	6	4,000.00	42,551.00
8	7	9,000.00	94,871.00
9	8	11,000.00	1,18,914.00
10	9	15,000.00	1,58,484.00
11	10	12,000.00	1,31,348.00
12	11	7,000.00	78,504.00
13	12	3,000.00	36,284.00
14	Average	6,541.67	70,870.33



Linear Regression – Assumptions

- Correlation
  There exists a correlation between the dependent and independent variables
- No Internal Correlation
  The independent variables do not have any linear relationships between each other
- Linearity
  There exists a linear relationship between the dependent and independent variables
- Normality of Errors
  The error term is normally distributed
  The expected value of the error term, conditional on the independent variables is zero

Equation & R-Square

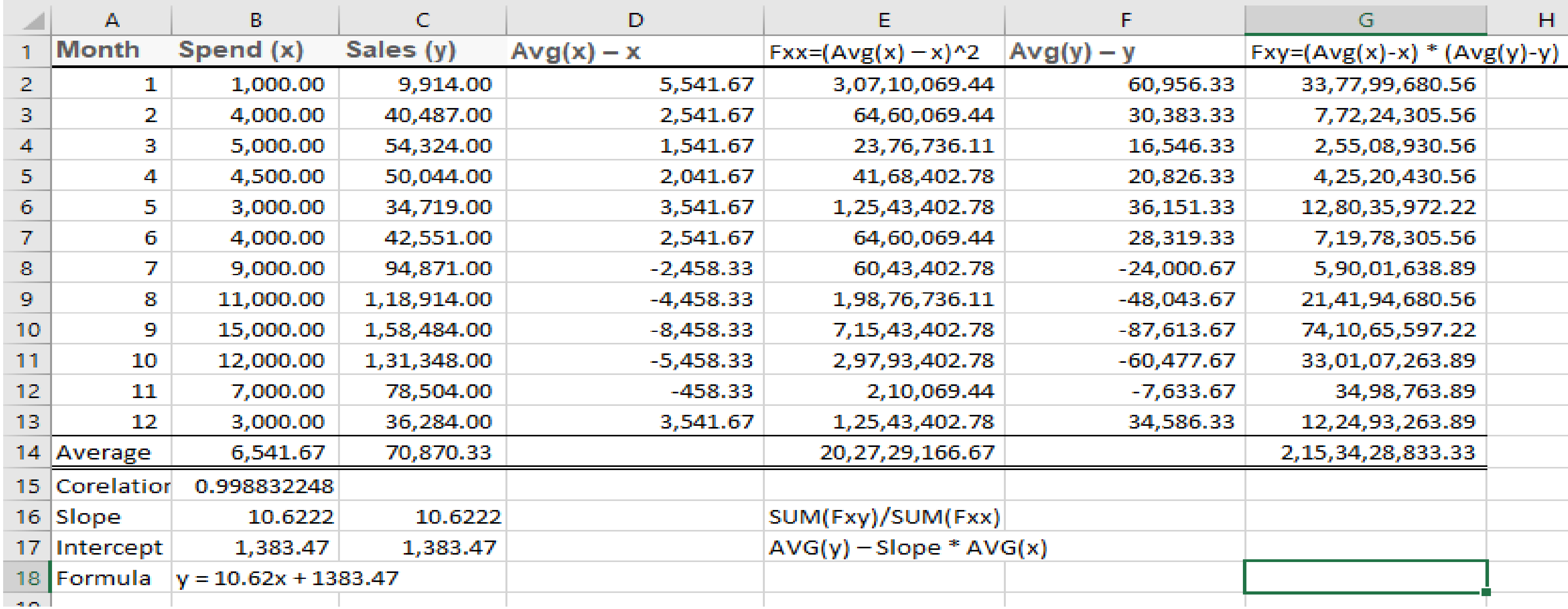
Equation

- Linear regression calculates an equation that minimizes the distance between the fitted line and all of the data points.
- A model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.
- Residual = Observed value - Fitted value
- Technically, ordinary least squares (OLS) regression minimizes the sum of the squared residuals.

R-Square

- R-squared is a statistical measure of how close the data are to the fitted regression line.
- It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.
- 0% indicates that the model explains none of the variability of the response data around its mean.
- If R-Square less than 0.65 (65%), dataset is not to be used for prediction via regression

Simple Linear Regression



	A	B	C	D	E	F	G	H
1	Month	Spend (x)	Sales (y)	Avg(x) – x	Fxx=(Avg(x) – x)^2	Avg(y) – y	Fxy=(Avg(x)-x) * (Avg(y)-y)	
2	1	1,000.00	9,914.00	5,541.67	3,07,10,069.44	60,956.33	33,77,99,680.56	
3	2	4,000.00	40,487.00	2,541.67	64,60,069.44	30,383.33	7,72,24,305.56	
4	3	5,000.00	54,324.00	1,541.67	23,76,736.11	16,546.33	2,55,08,930.56	
5	4	4,500.00	50,044.00	2,041.67	41,68,402.78	20,826.33	4,25,20,430.56	
6	5	3,000.00	34,719.00	3,541.67	1,25,43,402.78	36,151.33	12,80,35,972.22	
7	6	4,000.00	42,551.00	2,541.67	64,60,069.44	28,319.33	7,19,78,305.56	
8	7	9,000.00	94,871.00	-2,458.33	60,43,402.78	-24,000.67	5,90,01,638.89	
9	8	11,000.00	1,18,914.00	-4,458.33	1,98,76,736.11	-48,043.67	21,41,94,680.56	
10	9	15,000.00	1,58,484.00	-8,458.33	7,15,43,402.78	-87,613.67	74,10,65,597.22	
11	10	12,000.00	1,31,348.00	-5,458.33	2,97,93,402.78	-60,477.67	33,01,07,263.89	
12	11	7,000.00	78,504.00	-458.33	2,10,069.44	-7,633.67	34,98,763.89	
13	12	3,000.00	36,284.00	3,541.67	1,25,43,402.78	34,586.33	12,24,93,263.89	
14	Average	6,541.67	70,870.33		20,27,29,166.67		2,15,34,28,833.33	
15	Corelatior	0.998832248						
16	Slope	10.6222	10.6222		SUM(Fxy)/SUM(Fxx)			
17	Intercept	1,383.47	1,383.47		AVG(y) – Slope * AVG(x)			
18	Formula	y = 10.62x + 1383.47						

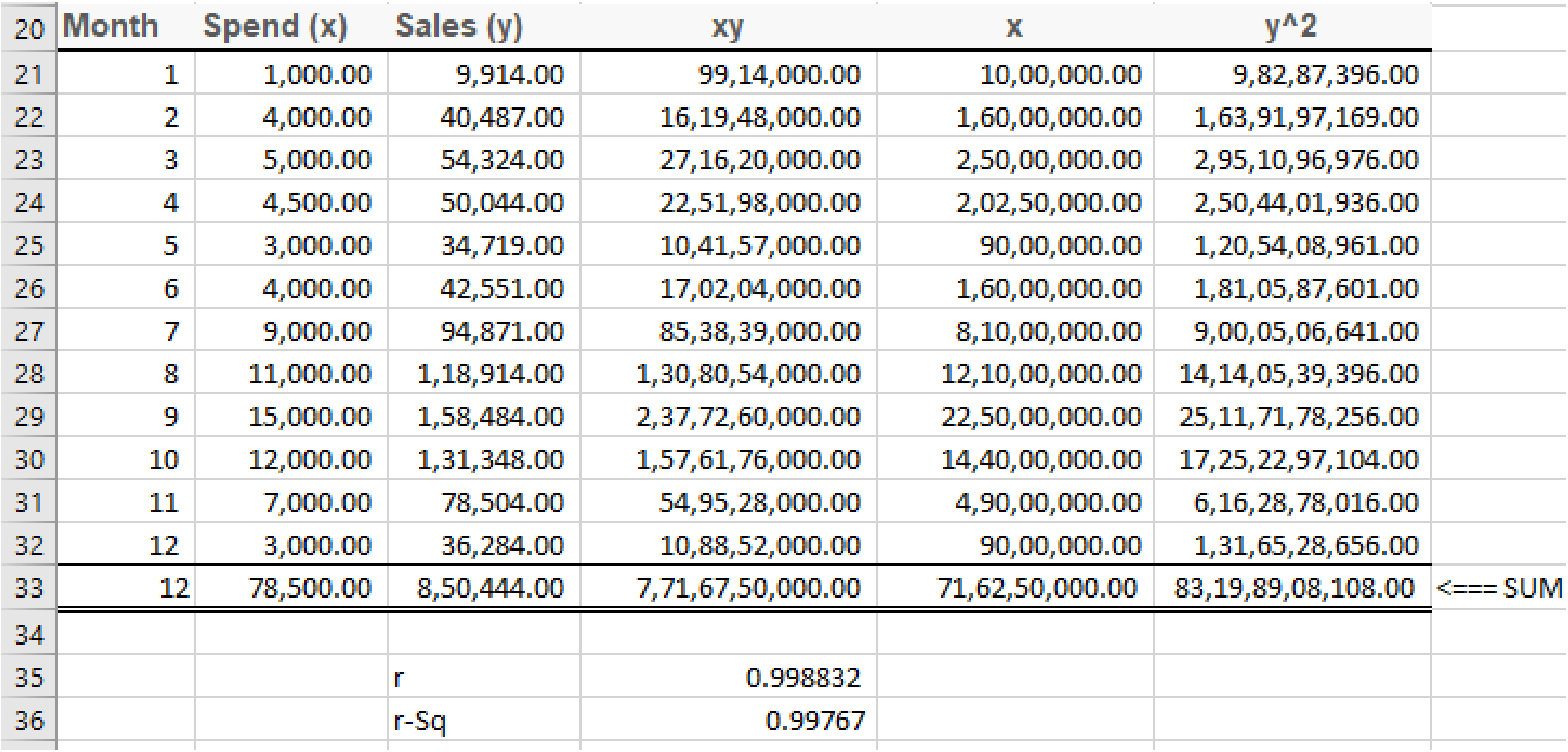
Slope

$$\sum((avg(x) - x) * (avg(y) - y)) / \sum((avg(x) - x)^2)$$

Intercept

$$avg(y) - (slope * avg(x))$$

Simple Linear Regression



20	Month	Spend (x)	Sales (y)	xy	x	y^2	
21	1	1,000.00	9,914.00	99,14,000.00	10,00,000.00	9,82,87,396.00	
22	2	4,000.00	40,487.00	16,19,48,000.00	1,60,00,000.00	1,63,91,97,169.00	
23	3	5,000.00	54,324.00	27,16,20,000.00	2,50,00,000.00	2,95,10,96,976.00	
24	4	4,500.00	50,044.00	22,51,98,000.00	2,02,50,000.00	2,50,44,01,936.00	
25	5	3,000.00	34,719.00	10,41,57,000.00	90,00,000.00	1,20,54,08,961.00	
26	6	4,000.00	42,551.00	17,02,04,000.00	1,60,00,000.00	1,81,05,87,601.00	
27	7	9,000.00	94,871.00	85,38,39,000.00	8,10,00,000.00	9,00,05,06,641.00	
28	8	11,000.00	1,18,914.00	1,30,80,54,000.00	12,10,00,000.00	14,14,05,39,396.00	
29	9	15,000.00	1,58,484.00	2,37,72,60,000.00	22,50,00,000.00	25,11,71,78,256.00	
30	10	12,000.00	1,31,348.00	1,57,61,76,000.00	14,40,00,000.00	17,25,22,97,104.00	
31	11	7,000.00	78,504.00	54,95,28,000.00	4,90,00,000.00	6,16,28,78,016.00	
32	12	3,000.00	36,284.00	10,88,52,000.00	90,00,000.00	1,31,65,28,656.00	
33	12	78,500.00	8,50,444.00	7,71,67,50,000.00	71,62,50,000.00	83,19,89,08,108.00	<=== SUM
34							
35			r	0.998832			
36			r-Sq	0.99767			

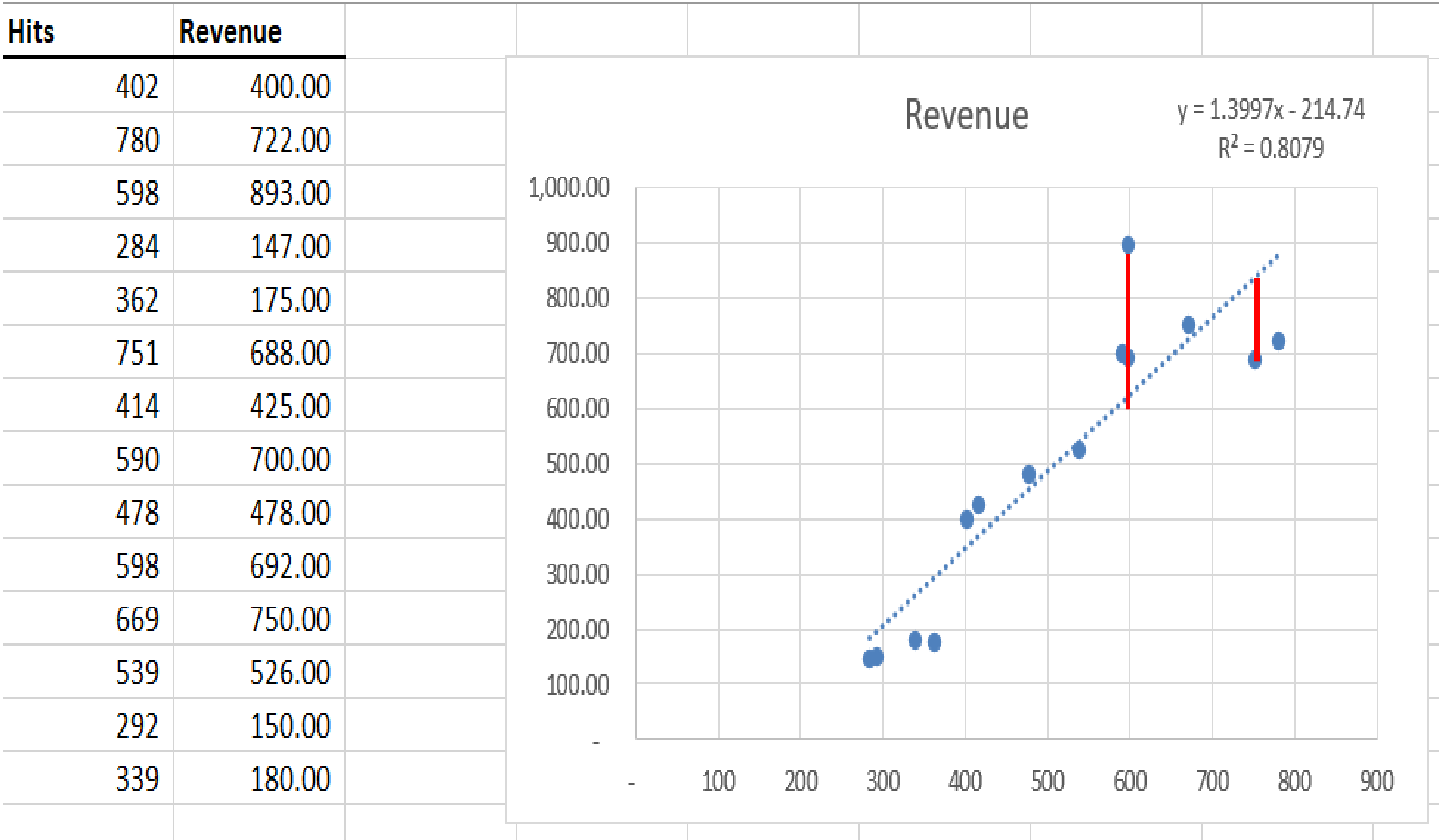
R Squared Formula = r²

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Simple Linear Regression – Steps

- Read Data
- Exploratory Data Analysis
- Data Transformation
- Data Imputation
- Visual Data Analysis
- **Generate OLS Summary**
- **Check Adj R-square**
- **Check P-value For Each Col**
- **Drop Column As Required**
- **Generate OLS Summary**
- **Create Model**
- **Train The Model Using Train Data**
- **Predict Using Test Data**
- **Check RMSE / Scatter Index**
- **Predict New Data**
- **Check RMSE / Scatter Index**

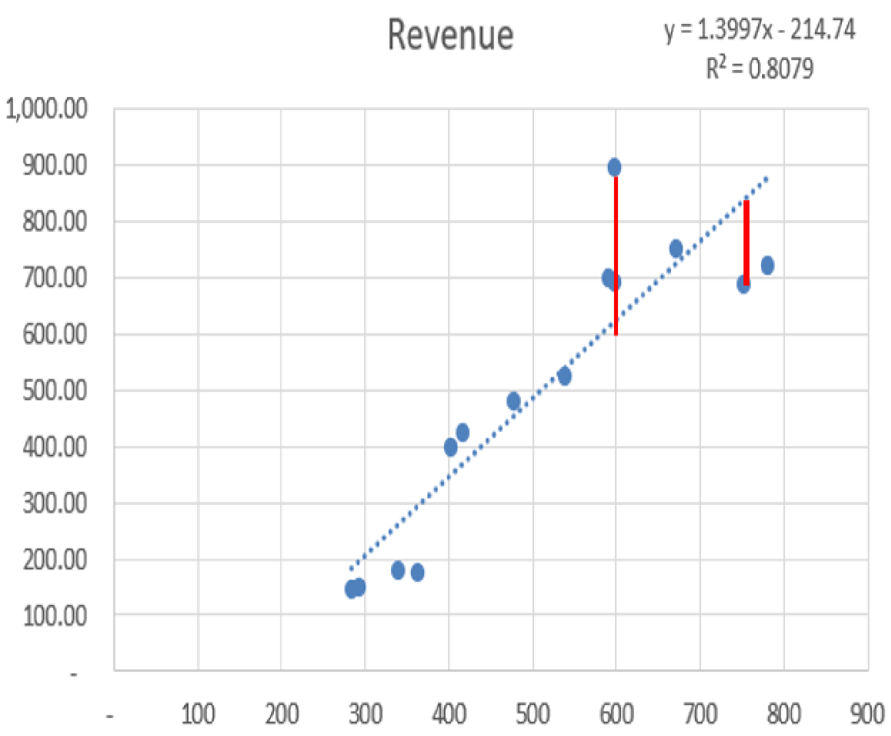
Evaluation – Linear Regression – Root Mean Square Error



Hits	Revenue
402	400.00
780	722.00
598	893.00
284	147.00
362	175.00
751	688.00
414	425.00
590	700.00
478	478.00
598	692.00
669	750.00
539	526.00
292	150.00
339	180.00

Evaluation - Linear Regression – Root Mean Square Error

- In Our Model, The Blue Dots Are The Actual Values And The Blue Line Is The Set Of Predicted Values.
- The Distance Between The Actual Value And The Predicted Line Represents The Error.
- Similarly, We Can Draw Straight Lines From Each Blue Dot To The Blue Line.
- Taking Mean Of All Those Distances After Squaring Them And Then Taking The Root Will Give Us RMSE.
- RMSE is a measure of how spread out these residuals / errors are.
- In other words, it tells you how concentrated the data is around the line of best fit.
- The Smaller The RMSE, The Better (Less Errors). Low RMSE of < 100 indicates a good fit. But is data dependent
- There Is No Fixed RMSE Benchmark. It Depends On The Distribution Of Data.



- For A Dataset Where The Error Which Ranges From 0 To 1000, An RMSE Of Around 100 Is Small
- In Case You Have A Higher RMSE Value, This Would Mean That You Probably Need To Change Your Evaluation To Scatter Index.

Evaluation – Linear Regression – Scatter Index

- Since There Is No Fixed RMSE Benchmark, We use Scatter Index (SI)
- Scatter Index is a metric which uses a combination of RMSE and Average Of Actual Values

$$ScatterIndex = \frac{RMSE}{mean(actual values)}$$

Interpretation

- If SI Is Less Than 1, Predictions Are Of Acceptable Quality
- SI Closer To 0 Is Better

Linear Regression – Multiple

- Single Explanatory / Dependent Variable, Y
- Many Independent Variable, X1 … Xn
- Correlation Must Exists Between X1 … Xn & Y
- Changes In Y Are Assumed To Be Directly Related To Changes In X1 … Xn
- Multi Collinearity Should Not Be Present In X1 … Xn
- Relationship between Xs and Y is described by a linear function

IMPORTANT

- Independent Variables must continuous numeric or categoric numeric values
- Dependent Variables are also continuous numeric values

Linear Regression – Multiple

- Regression Equation

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k\,.$$

- Slope for b1 & b2

For the two variable case:

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

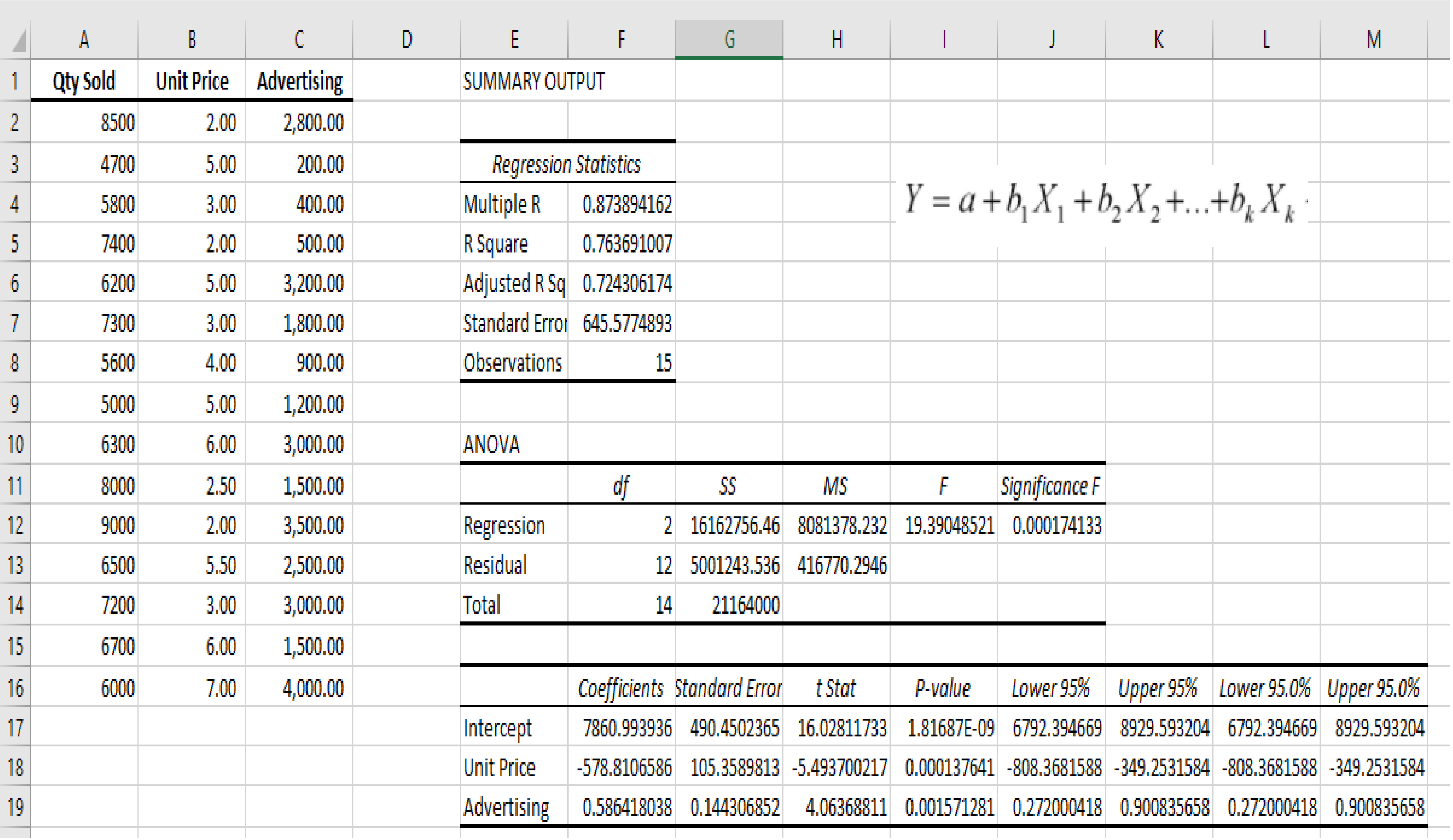
- Intercept

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

- R-Square

Link: http://faculty.cas.usf.edu/mbrannick/regression/Part3/Reg2.html

Linear Regression – Multiple



Qty Sold	Unit Price	Advertising
8500	2.00	2,800.00
4700	5.00	200.00
5800	3.00	400.00
7400	2.00	500.00
6200	5.00	3,200.00
7300	3.00	1,800.00
5600	4.00	900.00
5000	5.00	1,200.00
6300	6.00	3,000.00
8000	2.50	1,500.00
9000	2.00	3,500.00
6500	5.50	2,500.00
7200	3.00	3,000.00
6700	6.00	1,500.00
6000	7.00	4,000.00

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.873894162
R Square	0.763691007
Adjusted R Sq	0.724306174
Standard Error	645.5774893
Observations	15

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	16162756.46	8081378.232	19.39048521	0.000174133
Residual	12	5001243.536	416770.2946		
Total	14	21164000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	7860.993936	490.4502365	16.02811733	1.81687E-09	6792.394669	8929.593204	6792.394669	8929.593204
Unit Price	-578.8106586	105.3589813	-5.493700217	0.000137641	-808.3681588	-349.2531584	-808.3681588	-349.2531584
Advertising	0.586418038	0.144306852	4.06368811	0.001571281	0.272000418	0.900835658	0.272000418	0.900835658

Equation & R-Square & P-Value of F-Statistics

Equation

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k \cdot$$

R-Square / Adjusted R-Squre

- If R-Square / Adjusted R-Square should be more an 0.65 (65%). This indicate a good fit and the dataset may used for prediction via regression

P-Value of F-Statistics / F-statistics Probability / Significance-F

- If P-Value of F-Statistics should be less than 0.05. This indicate a good fit and the dataset may used for prediction via regression

Note

- R-Square / Adjusted R-Square & If P-Value of F-Statistics generally go together. If the two diverge something is wrong with our calculations.

Linear Regression – Steps

- Read Data
- Exploratory Data Analysis
- Data Transformation
- Data Imputation
- Visual Data Analysis
- **Split Train-test**
- **Generate OLS Summary**
- **Check Adj R-square**
- **Check P-value For Each Col**
- **Drop Column As Required**
- **Generate OLS Summary**
- **Create Model**
- **Train The Model Using Train Split**
- **Predict Train Split**
- **Check RMSE / Scatter Index**

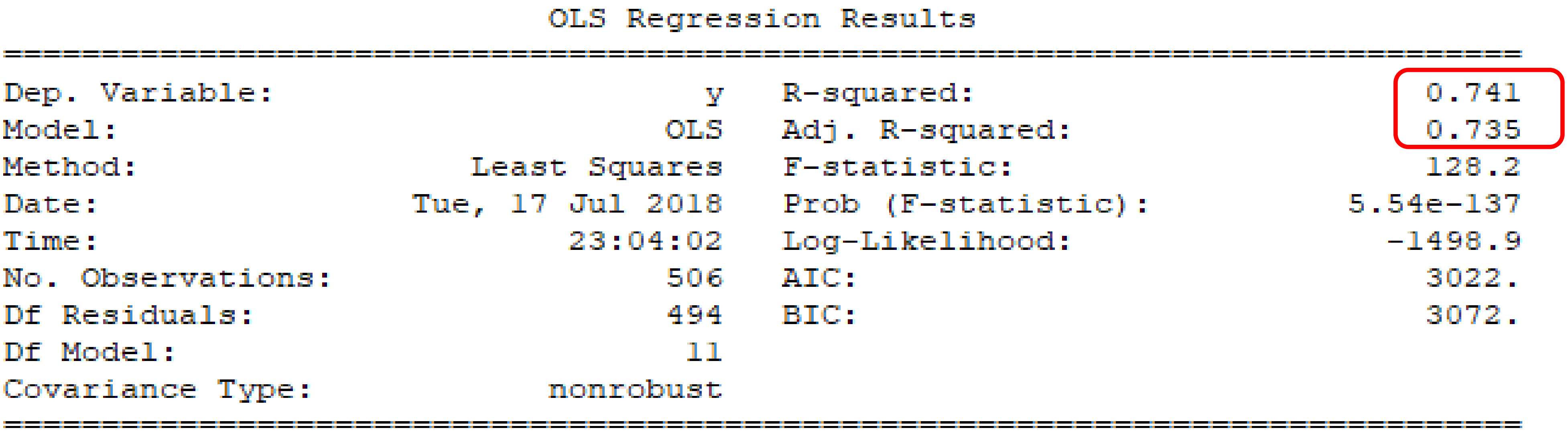
Linear Regression – Steps

- Predict Test Split
- Check RMSE / Scatter Index
- Create Model From Full Data
- Train The Model
- Read New Data
- Data Transformation
- Data Imputation
- Predict New Data
- Check RMSE / Scatter Index

Regression Output - Interpretation

R Square / Adj R Square (Regression Statistics)

- R Square signifies if regression is a good fit or not.
- The closer to 1, the better the regression line (read on) fits the data
- If R-Square less than 0.65 (65%), dataset is not to be used for prediction via regression

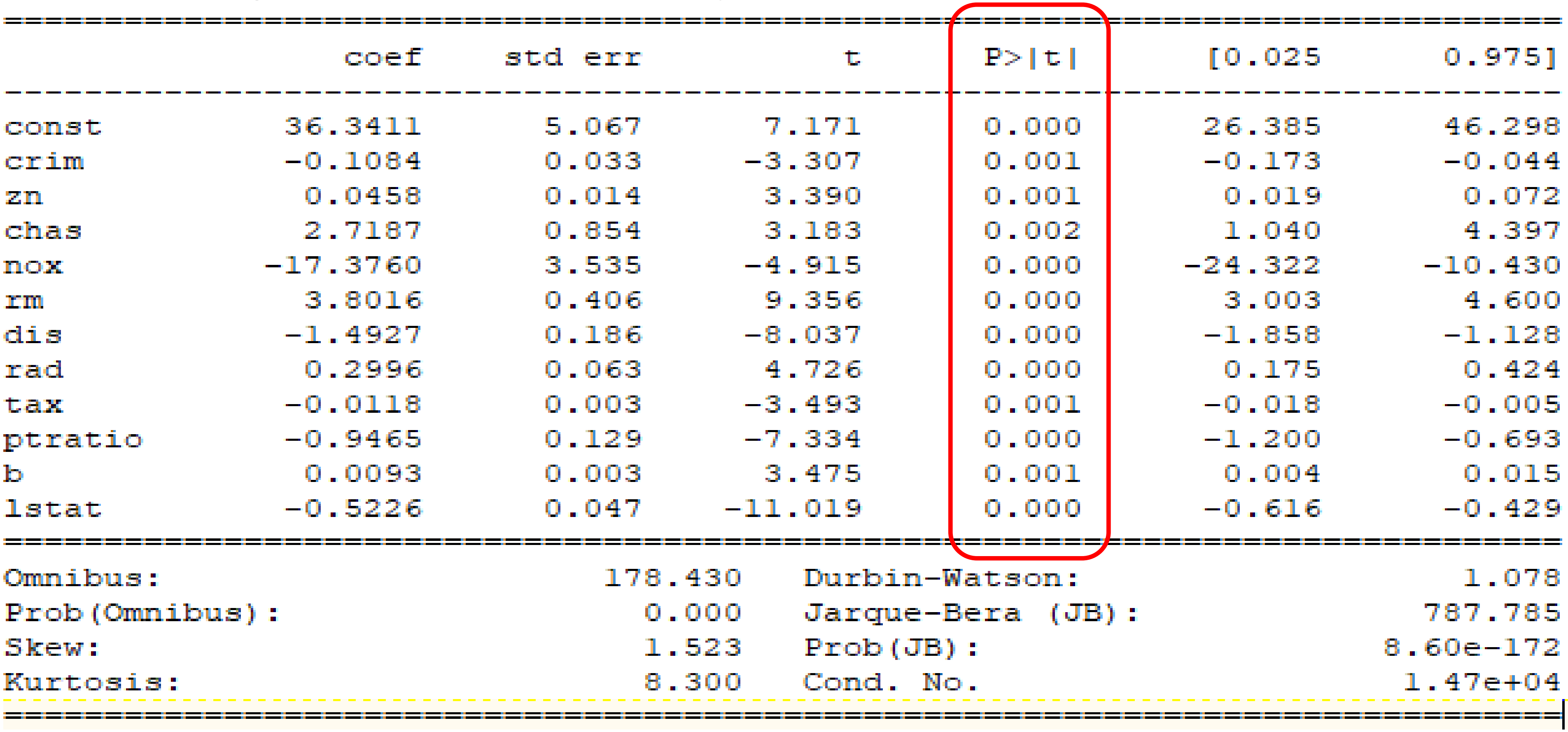


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.741
Model:                            OLS   Adj. R-squared:                  0.735
Method:                 Least Squares   F-statistic:                     128.2
Date:                Tue, 17 Jul 2018   Prob (F-statistic):          5.54e-137
Time:                        23:04:02   Log-Likelihood:                -1498.9
No. Observations:                 506   AIC:                             3022.
Df Residuals:                     494   BIC:                             3072.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
```

Regression Output - Interpretation

P-values (Probability Values)

- All P-values should be below 0.05
- Delete a variable with a high P-value (greater than 0.05)
- Re-run the regression until all P-values drops below 0.05

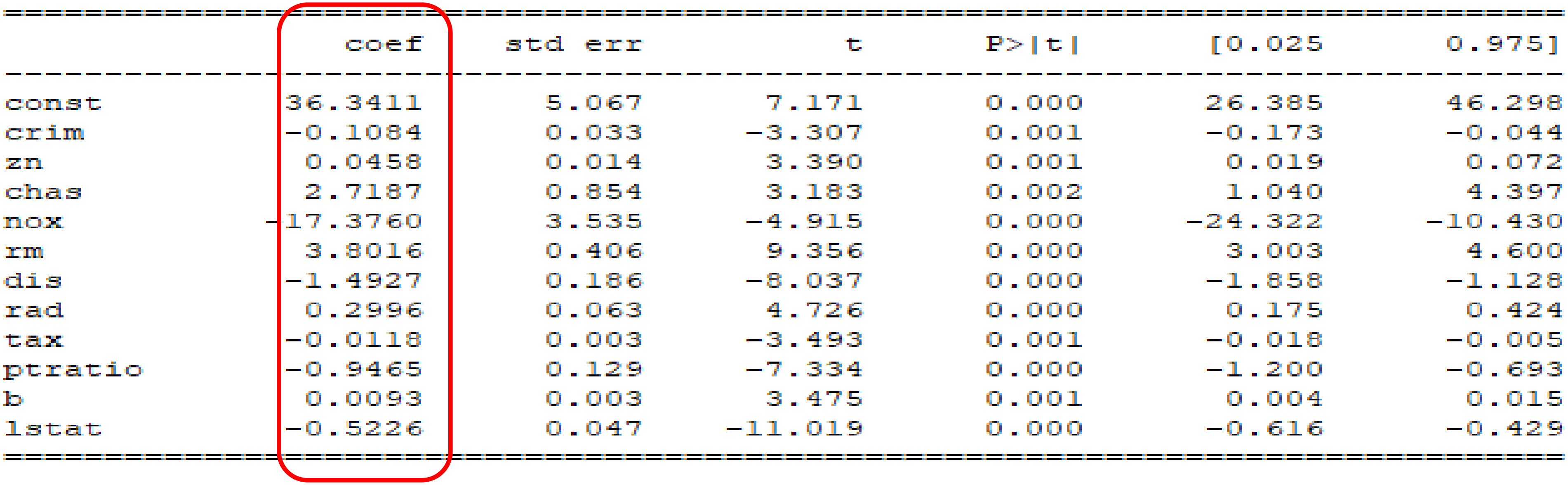


```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         36.3411      5.067      7.171      0.000      26.385      46.298
crim          -0.1084      0.033     -3.307      0.001      -0.173      -0.044
zn             0.0458      0.014      3.390      0.001       0.019       0.072
chas           2.7187      0.854      3.183      0.002       1.040       4.397
nox          -17.3760      3.535     -4.915      0.000     -24.322     -10.430
rm             3.8016      0.406      9.356      0.000       3.003       4.600
dis           -1.4927      0.186     -8.037      0.000      -1.858      -1.128
rad            0.2996      0.063      4.726      0.000       0.175       0.424
tax           -0.0118      0.003     -3.493      0.001      -0.018      -0.005
ptratio       -0.9465      0.129     -7.334      0.000      -1.200      -0.693
b              0.0093      0.003      3.475      0.001       0.004       0.015
lstat         -0.5226      0.047    -11.019      0.000      -0.616      -0.429
==============================================================================
Omnibus:                      178.430   Durbin-Watson:                   1.078
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              787.785
Skew:                           1.523   Prob(JB):                    8.60e-172
Kurtosis:                       8.300   Cond. No.                     1.47e+04
==============================================================================
```

Regression Output - Interpretation

Coefficients

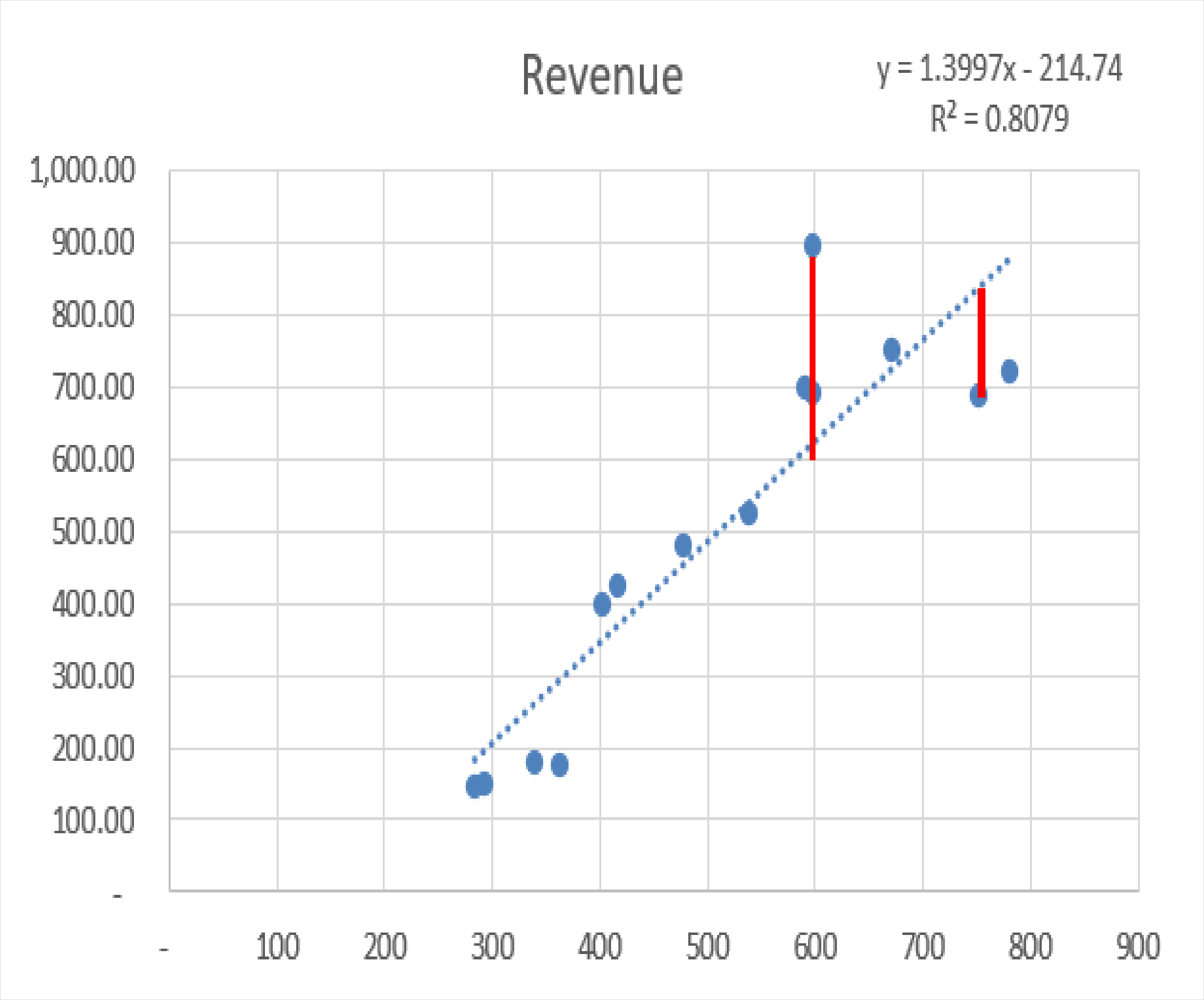
- You can use these coefficients to do a forecast.
- Coefficients help you put a formula to the regression line.
- y (dependent variable) = Coefficient-of-Intercept + Variable-1 * Coefficient(Variable-1)
  + Variable-n * Coefficient(Variable-n)
- Price Y = 8536.214 - (835.722 * rm) + (0.592 * lstat)+ …. + 36.3411
- This is automatically done by sklearn.LinearRegression



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         36.3411      5.067      7.171      0.000      26.385      46.298
crim          -0.1084      0.033     -3.307      0.001      -0.173      -0.044
zn             0.0458      0.014      3.390      0.001       0.019       0.072
chas           2.7187      0.854      3.183      0.002       1.040       4.397
nox          -17.3760      3.535     -4.915      0.000     -24.322     -10.430
rm             3.8016      0.406      9.356      0.000       3.003       4.600
dis           -1.4927      0.186     -8.037      0.000      -1.858      -1.128
rad            0.2996      0.063      4.726      0.000       0.175       0.424
tax           -0.0118      0.003     -3.493      0.001      -0.018      -0.005
ptratio       -0.9465      0.129     -7.334      0.000      -1.200      -0.693
b              0.0093      0.003      3.475      0.001       0.004       0.015
lstat         -0.5226      0.047    -11.019      0.000      -0.616      -0.429
==============================================================================
```

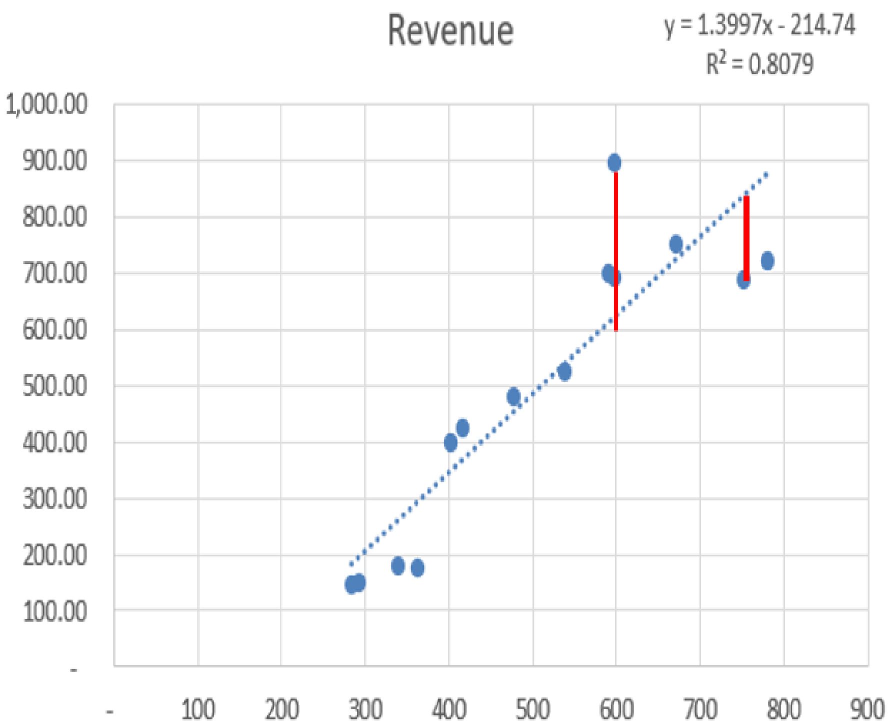
Evaluation – Linear Regression – Root Mean Square Error

Hits	Revenue
402	400.00
780	722.00
598	893.00
284	147.00
362	175.00
751	688.00
414	425.00
590	700.00
478	478.00
598	692.00
669	750.00
539	526.00
292	150.00
339	180.00



Evaluation - Linear Regression – Root Mean Square Error

- In Our Model, The Blue Dots Are The Actual Values And The Blue Line Is The Set Of Predicted Values.
- The Distance Between The Actual Value And The Predicted Line Represents The Error.
- Similarly, We Can Draw Straight Lines From Each Blue Dot To The Blue Line.
- Taking Mean Of All Those Distances After Squaring Them And Then Taking The Root Will Give Us RMSE.
- RMSE is a measure of how spread out these residuals / errors are.
- In other words, it tells you how concentrated the data is around the line of best fit.
- The Smaller The RMSE, The Better (Less Errors). Low RMSE of < 100 indicates a good fit. But is data dependent
- There Is No Fixed RMSE Benchmark. It Depends On The Distribution Of Data.



- For A Dataset Where The Error Which Ranges From 0 To 1000, An RMSE Of Around 100 Is Small
- In Case You Have A Higher RMSE Value, This Would Mean That You Probably Need To Change Your Evaluation To Scatter Index.

Evaluation – Linear Regression – Scatter Index

- Since There Is No Fixed RMSE Benchmark, We use Scatter Index (SI)
- Scatter Index is a metric which uses a combination of RMSE and Average Of Actual Values

$$ScatterIndex = \frac{RMSE}{mean(actualvalues)}$$

Interpretation

- If SI Is Less Than 1, Predictions Are Of Acceptable Quality
- SI Closer To 0 Is Better

Thank you!

Contact:

Cyrus Lentin
cyrus@lentins.co.in
+91-98200-94236