Regression

Table of Content

© Copyright by Digital Skola 2022

Regression

© Copyright by Digital Skola 2022

Regression analysis is a way of mathematically sorting out which of those variables does indeed have an impact. It answers the questions: Which factors matter most? Which can we ignore? How do those factors interact with each other? And, perhaps most importantly, how certain are we about all of these factors?

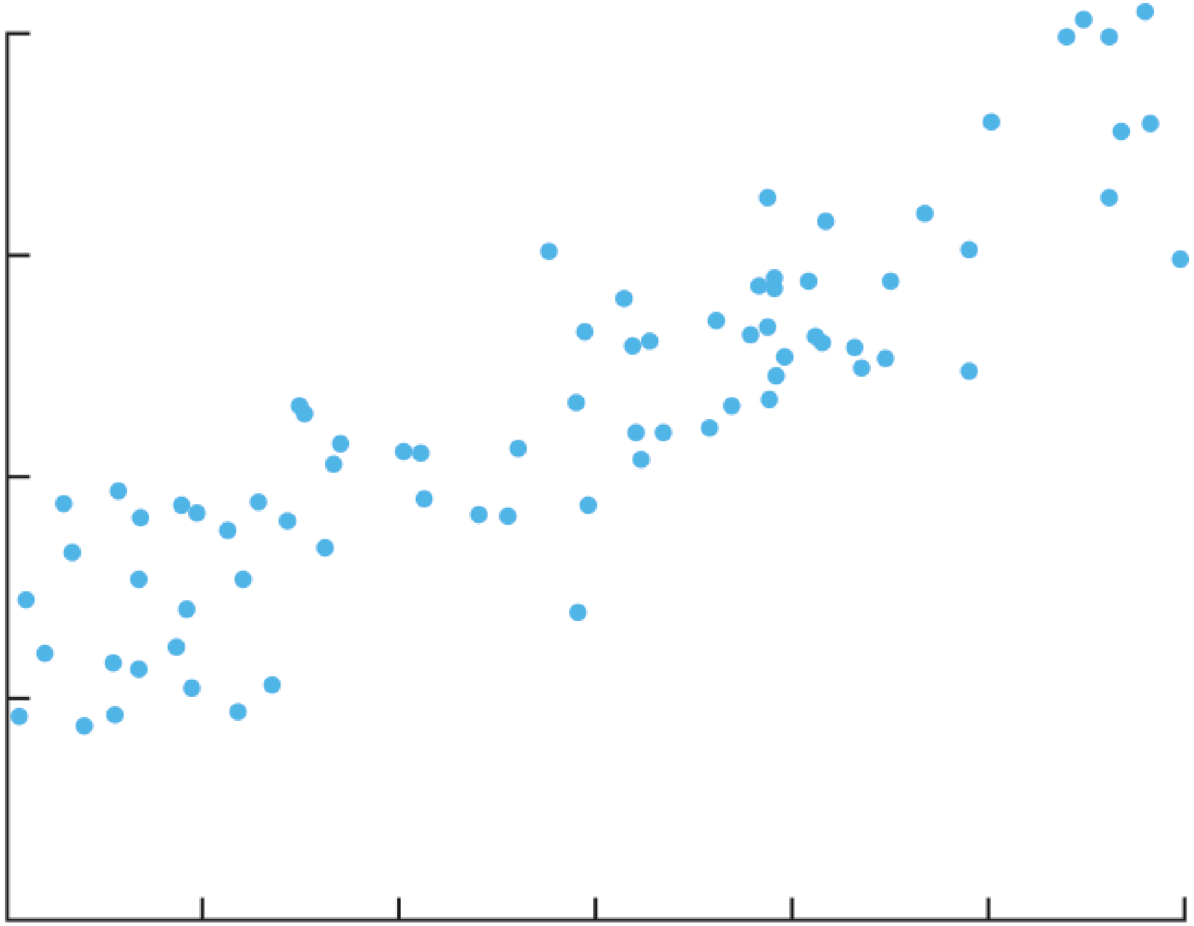
In regression analysis, those factors are called variables. You have your
dependent variable
— the main factor that you're trying to understand or predict. In Redman's example above, the dependent variable is monthly sales.

independent variables
— the factors you suspect have an impact on your dependent variable.

© Copyright by Digital Skola 2022

Is There a Relationship Between These Two Variables?

Plotting your data is the first step in figuring that out.



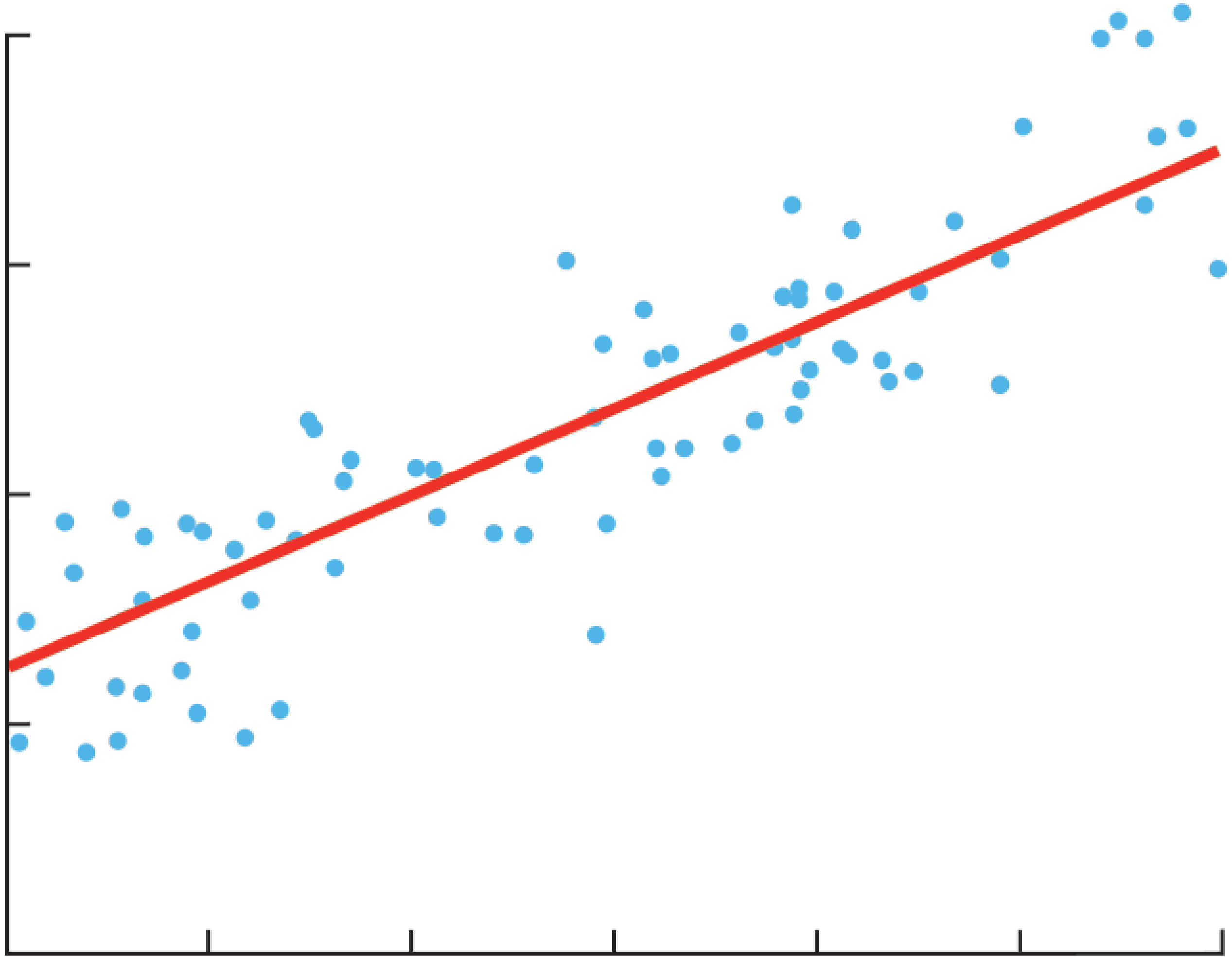
SOURCE HBR.ORG

© HBR.ORG

© Copyright by Digital Skola 2022

Building a Regression Model

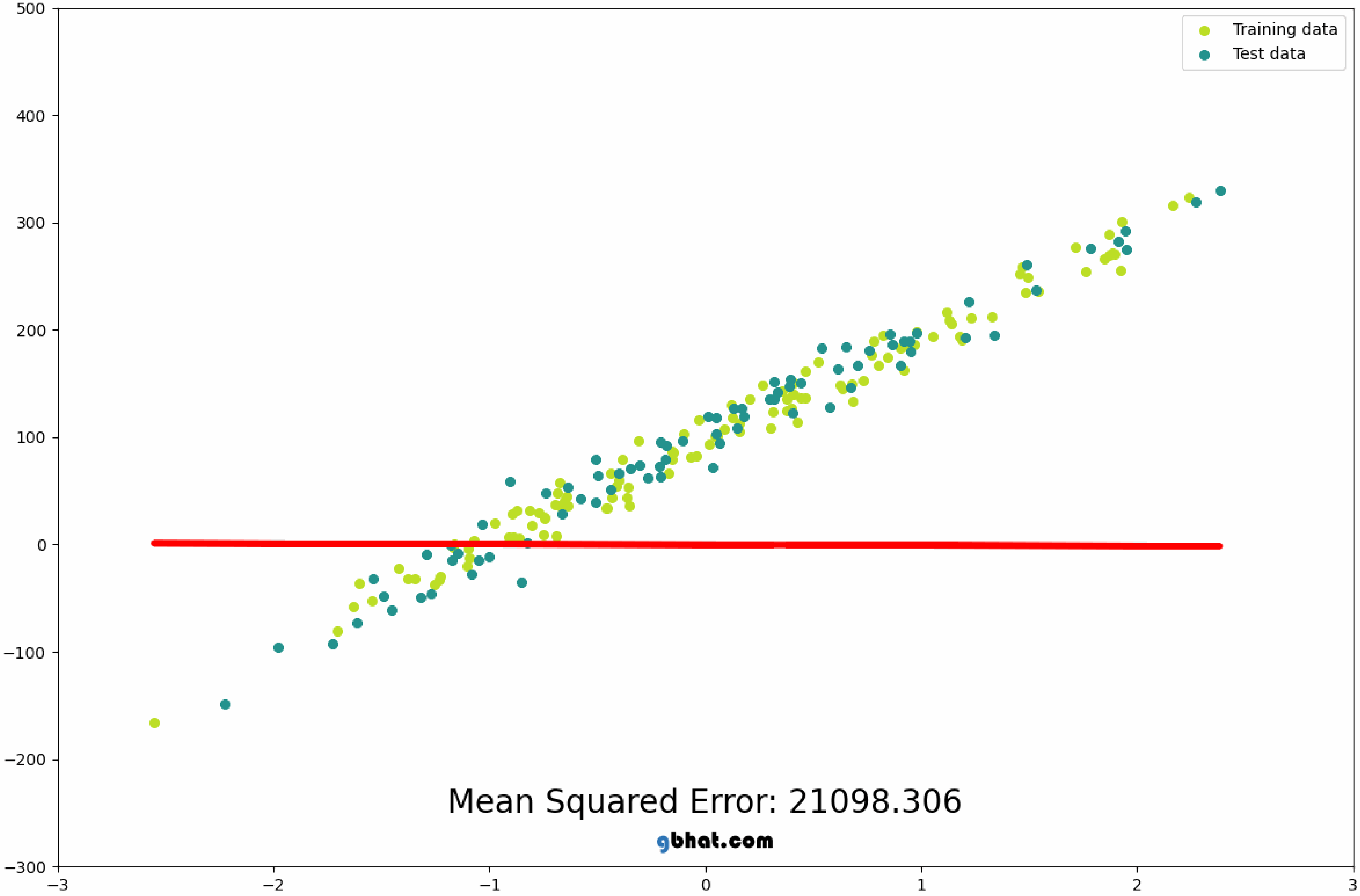
The line summarizes the relationship between x and y.



SOURCE HBR.ORG

© HBR.ORG

© Copyright by Digital Skola 2022



al Skola 2022

How do companies use it?

Regression analysis is the “go-to method in analytics,” says Redman. And smart companies use it to make decisions about all sorts of business issues. “As managers, we want to figure out how we can impact sales or employee retention or recruiting the best people. It helps us figure out what we can do.”

Most companies use regression analysis to explain a phenomenon they want to understand (e.g. why did customer service calls drop last month?); predict things about the future (e.g. what will sales look like over the next six months?); or to decide what to do (e.g. should we go with this promotion or a different one?).

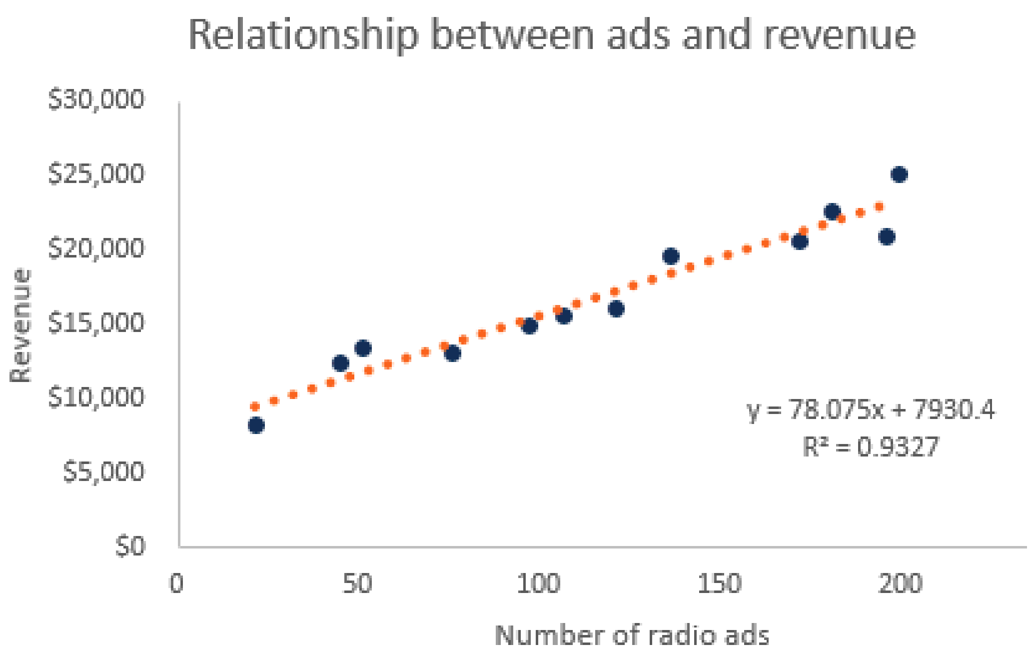
© Copyright by Digital Skola 2022

=FORECAST(B20,$C$4:$C$15,$B$4:$B$15)

Method #3: Simple Linear Regression

Data	Radio ads	Revenue
Jan	21	$8,350.0
Feb	180	$22,755.0
Mar	50	$13,455.0
Apr	195	$21,100.0
May	96	$15,000.0
Jun	44	$12,500.0
Jul	171	$20,700.0
Aug	135	$19,722.0
Sep	120	$16,115.0
Oct	75	$13,100.0
Nov	106	$15,670.0
Dec	198	$25,300.0
Totals	**1,391**	**$203,767.0**
Average	**116**	**$16,980.6**

Forecast function		
	100	=FORECAST(B20,$C$4:$C$15,$B$4:$B$15)
	150	FORECAST(x, known_ys, known_xs)
	200	$23,545.4



Regression Analysis – Simple Linear Regression

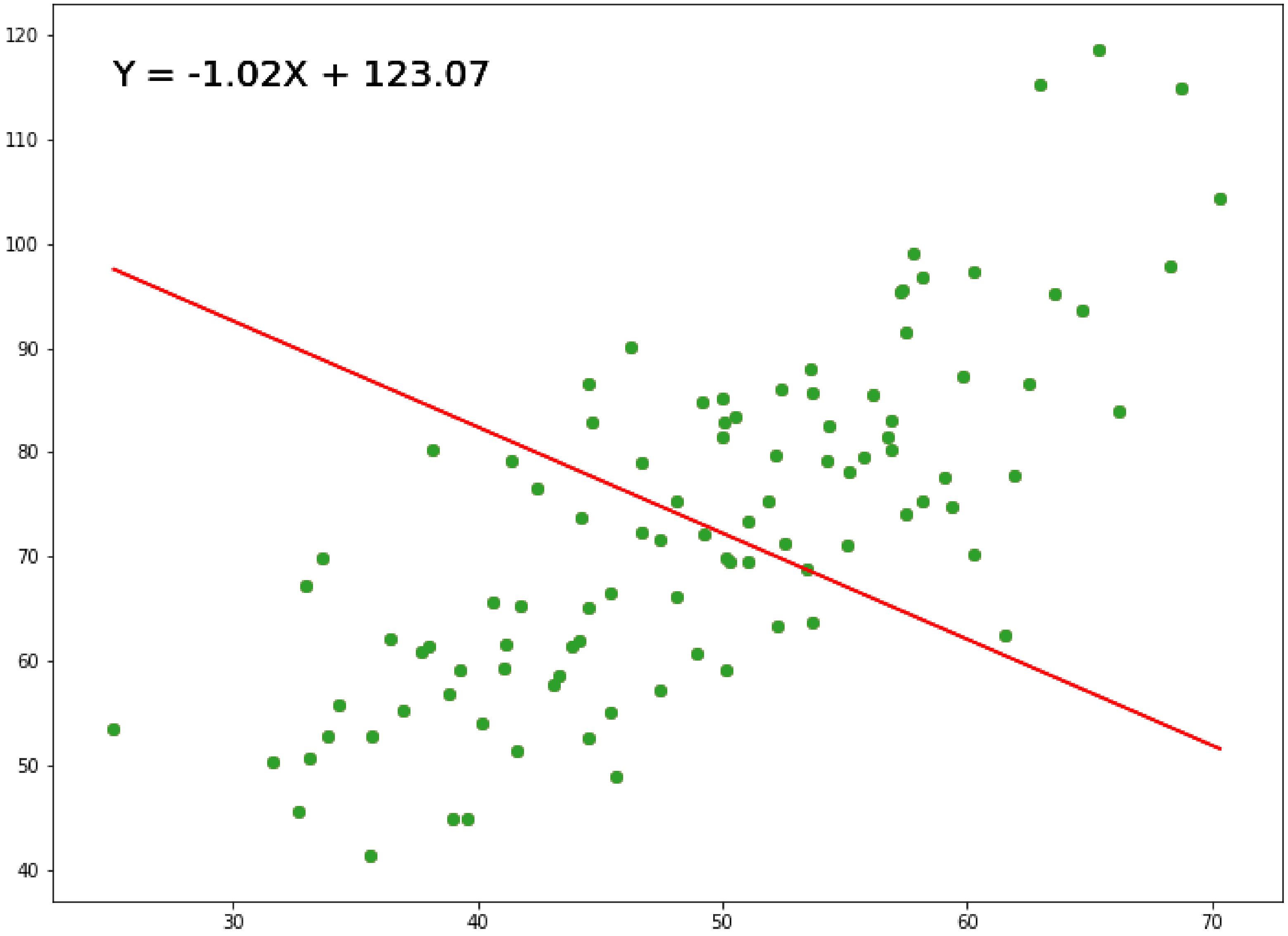
Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation:

$$Y = a + bX$$

Where:

- **Y** – Dependent variable
- **X** – Independent (explanatory) variable
- **a** – Intercept
- **b** – Slope

© Copyright by Digital Skola 2022



Y = -1.02X + 123.07
120
110
100
90
80
70
60
50
40
30
40
50
60
70

Regression Analysis – Multiple Linear Regression

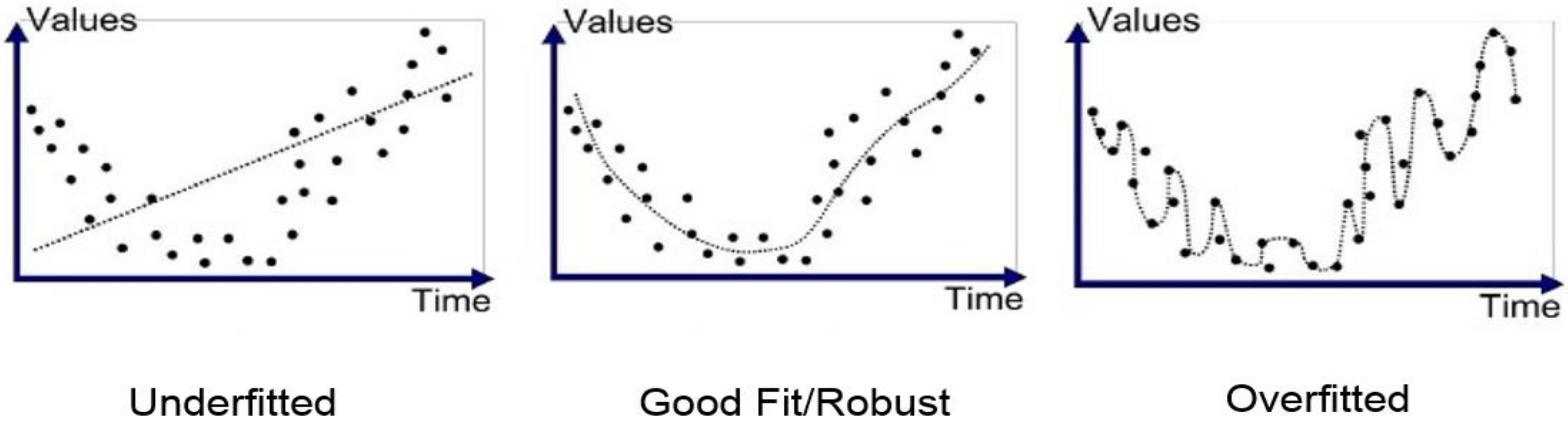
Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$$Y = a + bX_1 + cX_2 + dX_3 + \epsilon$$

Where:

- **Y** – Dependent variable
- **X_1, X_2, X_3** – Independent (explanatory) variables
- **a** – Intercept
- **b, c, d** – Slopes
- **ϵ** – Residual (error)

© Copyright by Digital Skola 2022



© Copyright by Digital Skola 2022

Lasso and Ridge Regression

© Copyright by Digital Skola 2022

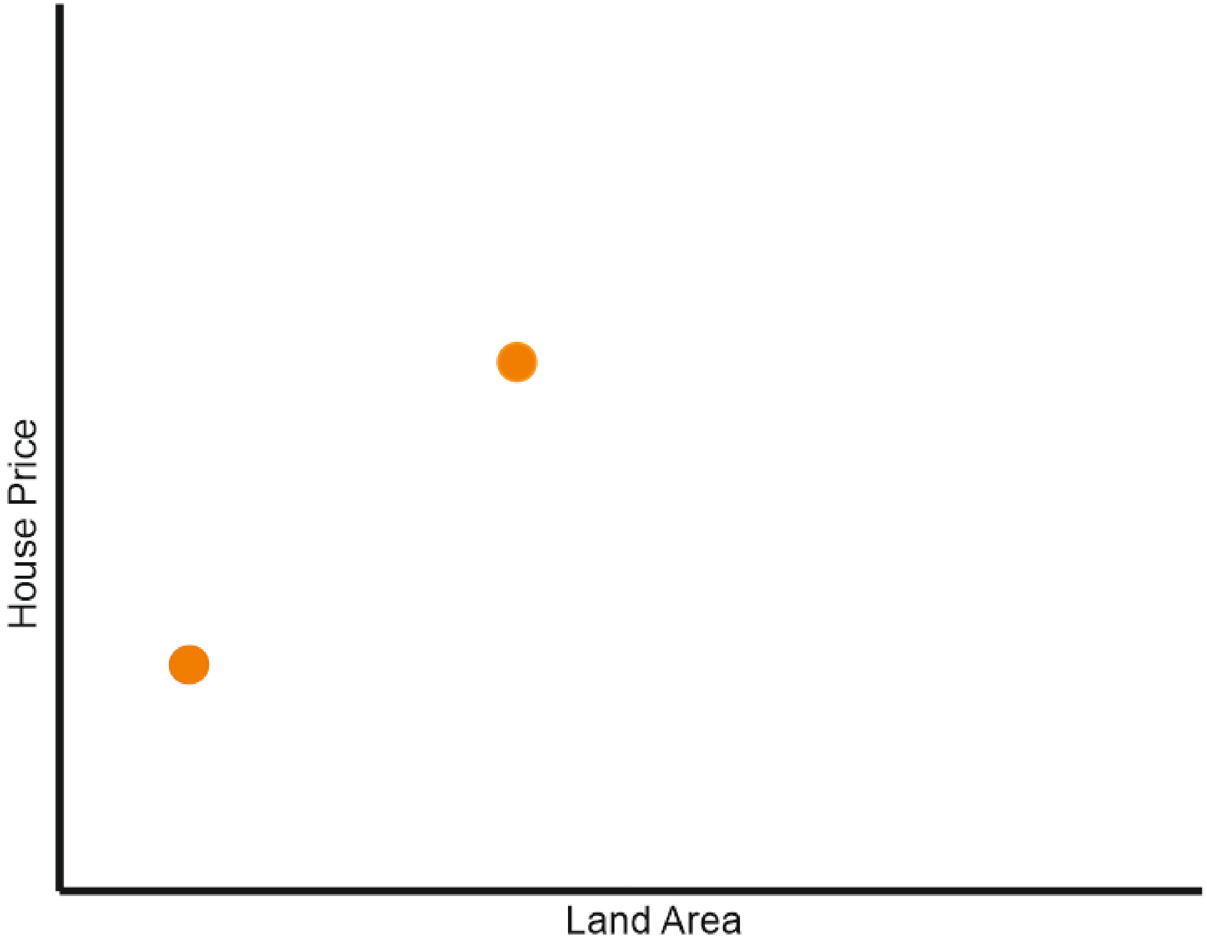
The procedure for selecting a regression line uses an error value, known as Sum Square Error (SSE). Regression lines are formed when minimizing SSE values.

Where the SSE formula is as follows:

$$SSE = \sum_{x=1}^{n} (y_i - \hat{y}_i)^2$$

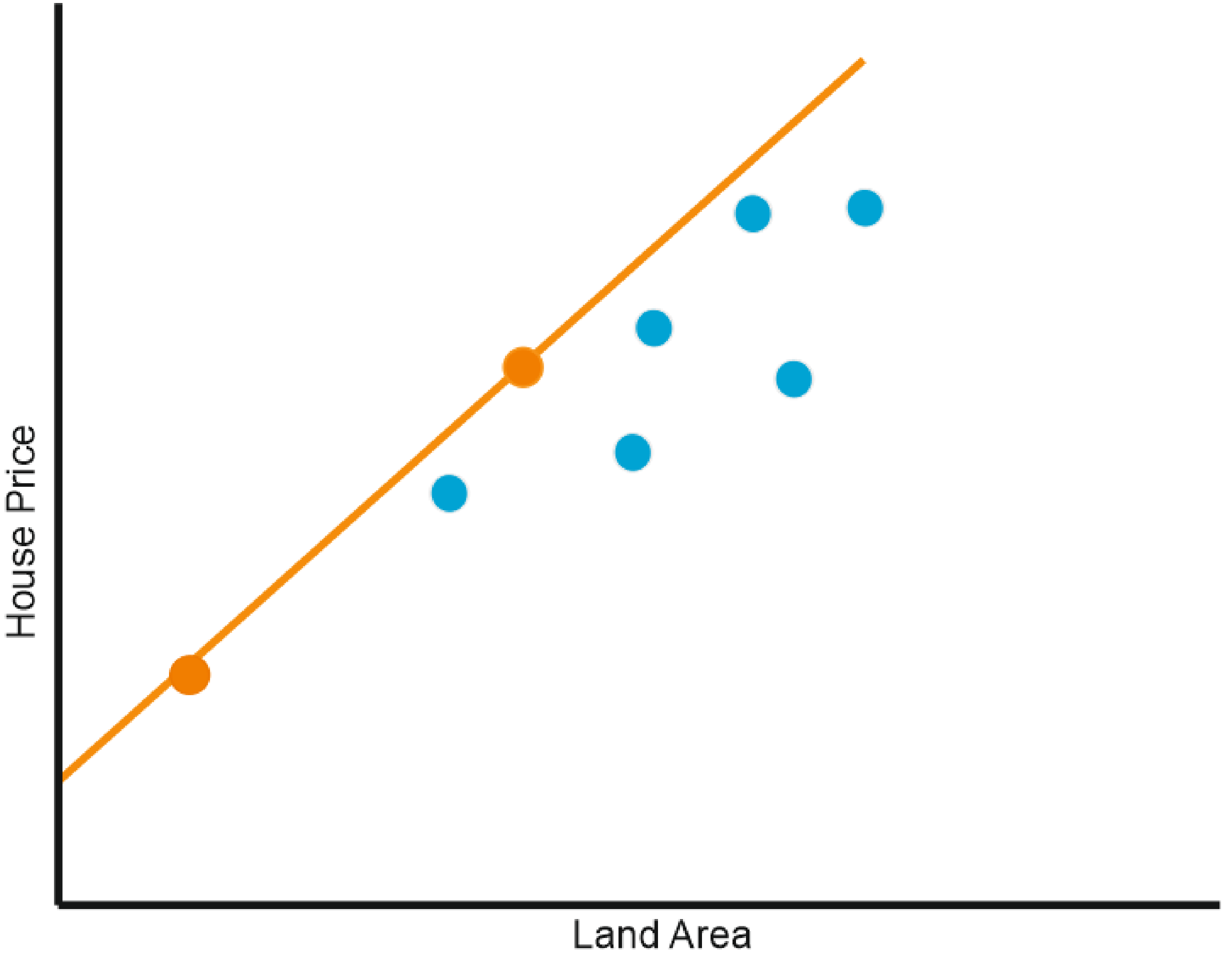
© Copyright by Digital Skola 2022

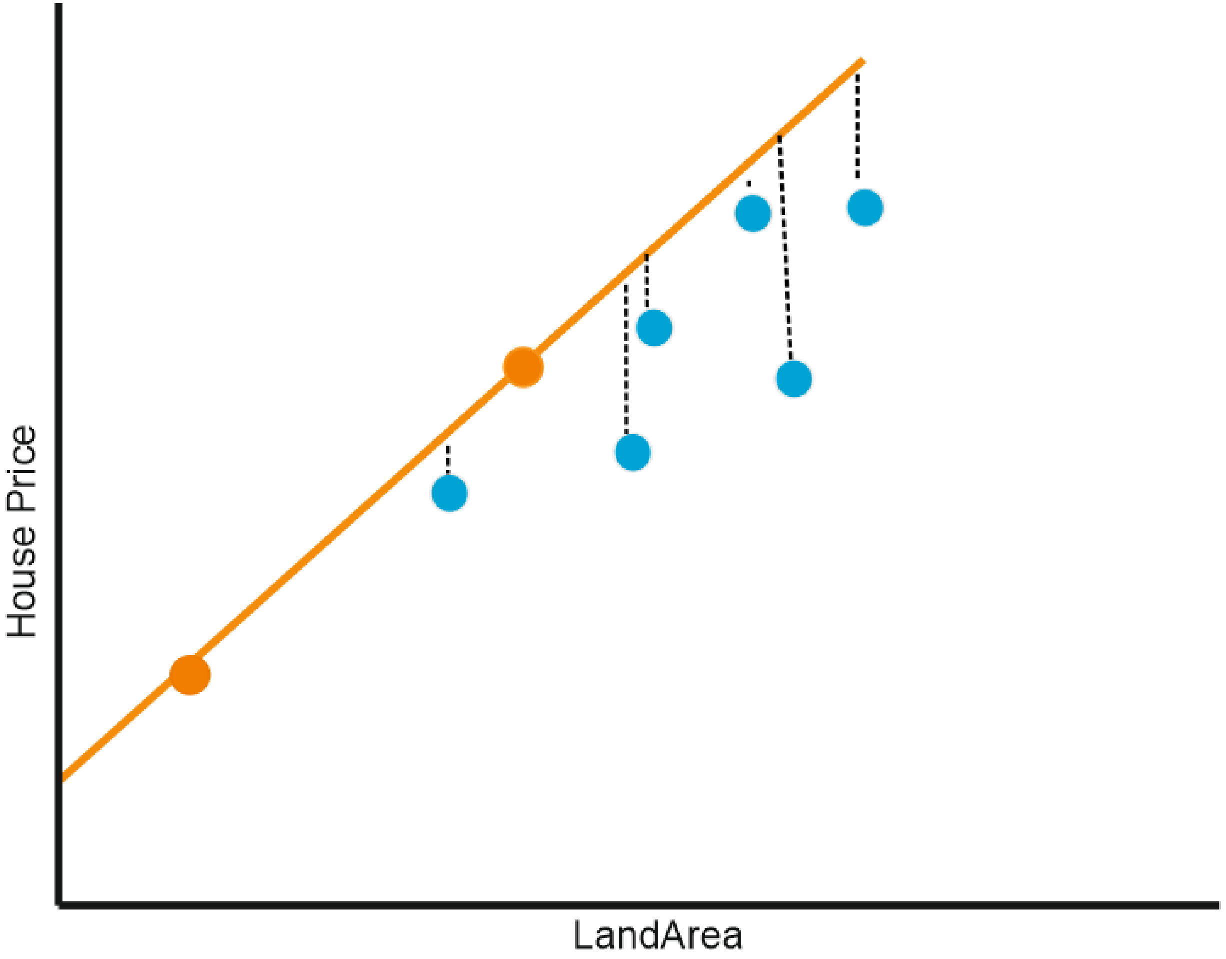
Fitting linear regression



© Copyright by Digital Skola 2022

But the model is overfit





© Copyright by Digital Skola 2022

To tackle overfit in liner regression
we can use 2 method:

1. Ridge Regression
2. Lasso Regression

© Copyright by Digital Skola 2022

RIDGE REGRESSION

Ridge Regression is a variation of linear regression. We use ridge regression to tackle the multicollinearity problem. Due to multicollinearity, we see a very large variance in the least square estimates of the model. So to reduce this variance a degree of bias is added to the regression estimates.
Ordinary Least Square (OLS) will create a model by minimizing the value of Sum Square Error (SSE), Whereas The Rigde regression will create a model by minimizing :

$$SSE + \lambda \sum_{i=1}^{n} (\beta_i)^2$$

© Copyright by Digital Skola 2022

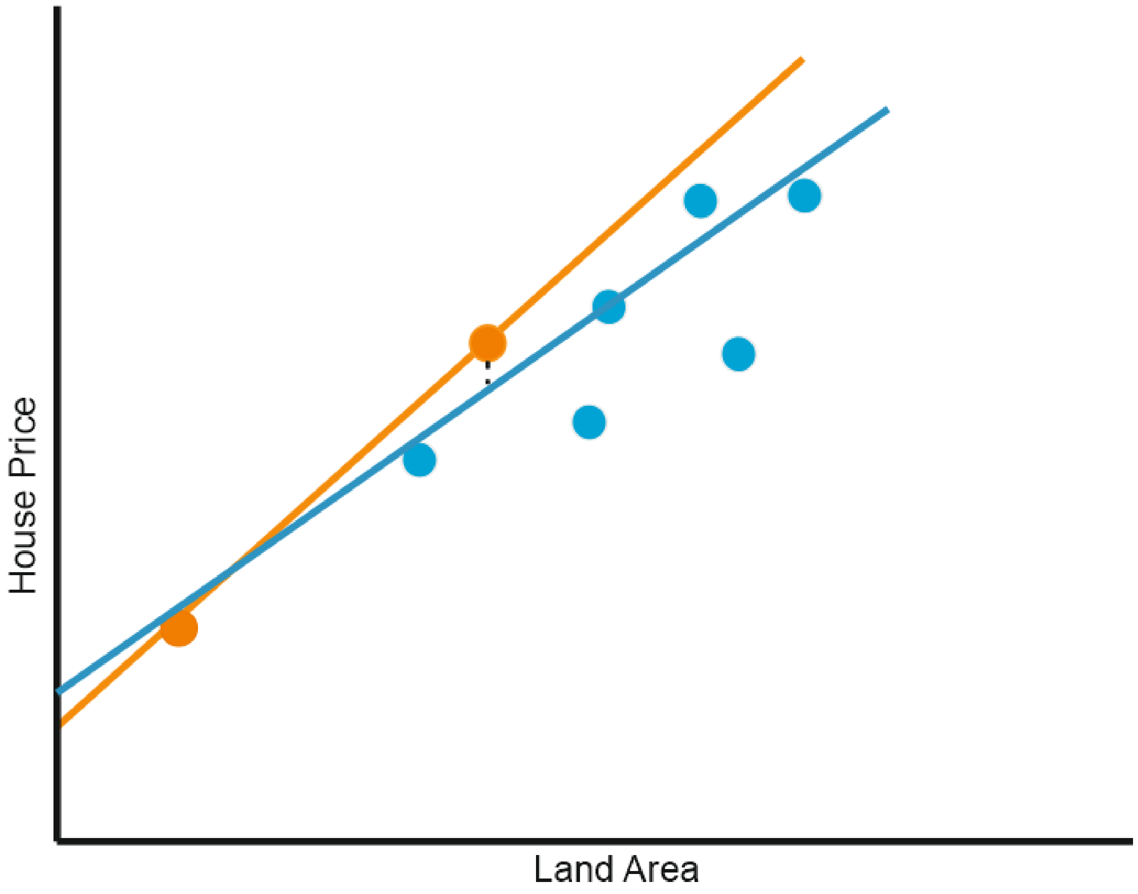
RIDGE REGRESSION

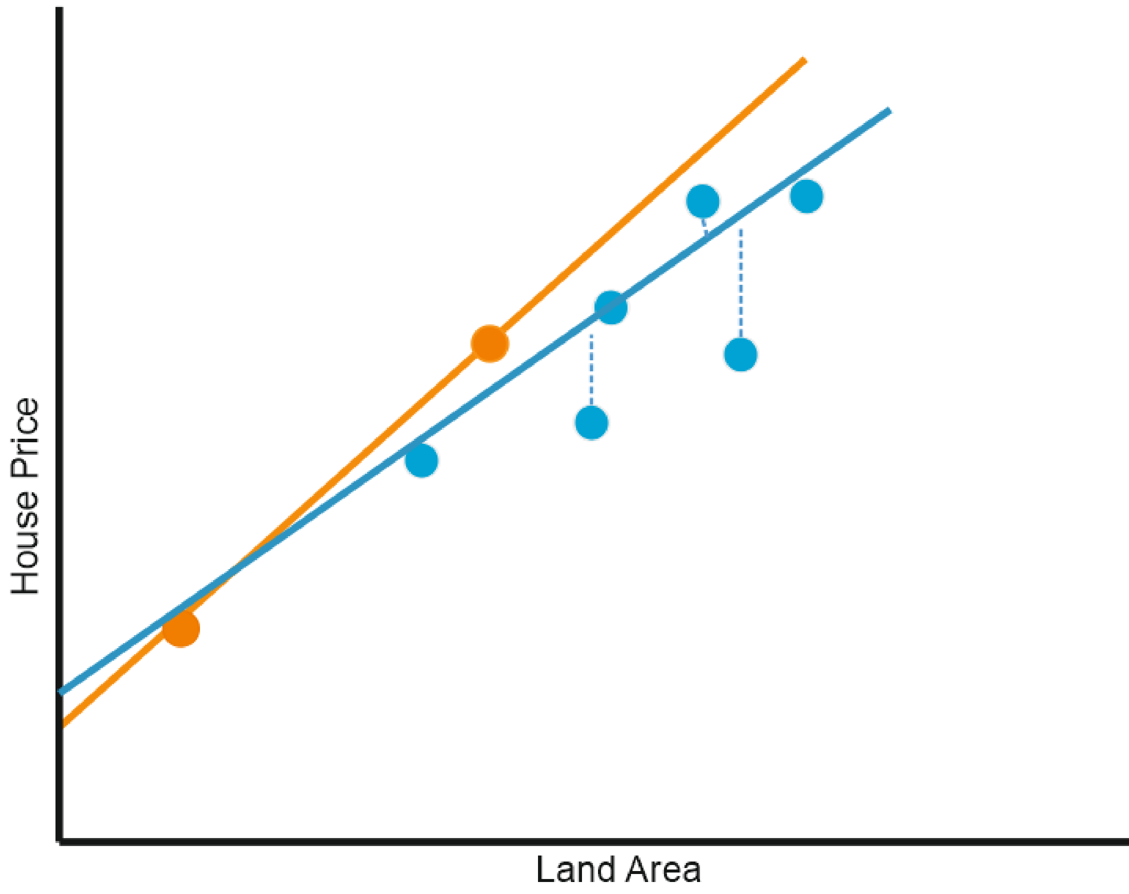
Ridge Regression is a variation of linear regression. We use ridge regression to tackle the multicollinearity problem. Due to multicollinearity, we see a very large variance in the least square estimates of the model. So to reduce this variance a degree of bias is added to the regression estimates.
Ordinary Least Square (OLS) will create a model by minimizing the value of Sum Square Error (SSE), Whereas The Rigde regression will create a model by minimizing :

$$SSE + \lambda \sum_{i=1}^{n} (\beta_i)^2$$

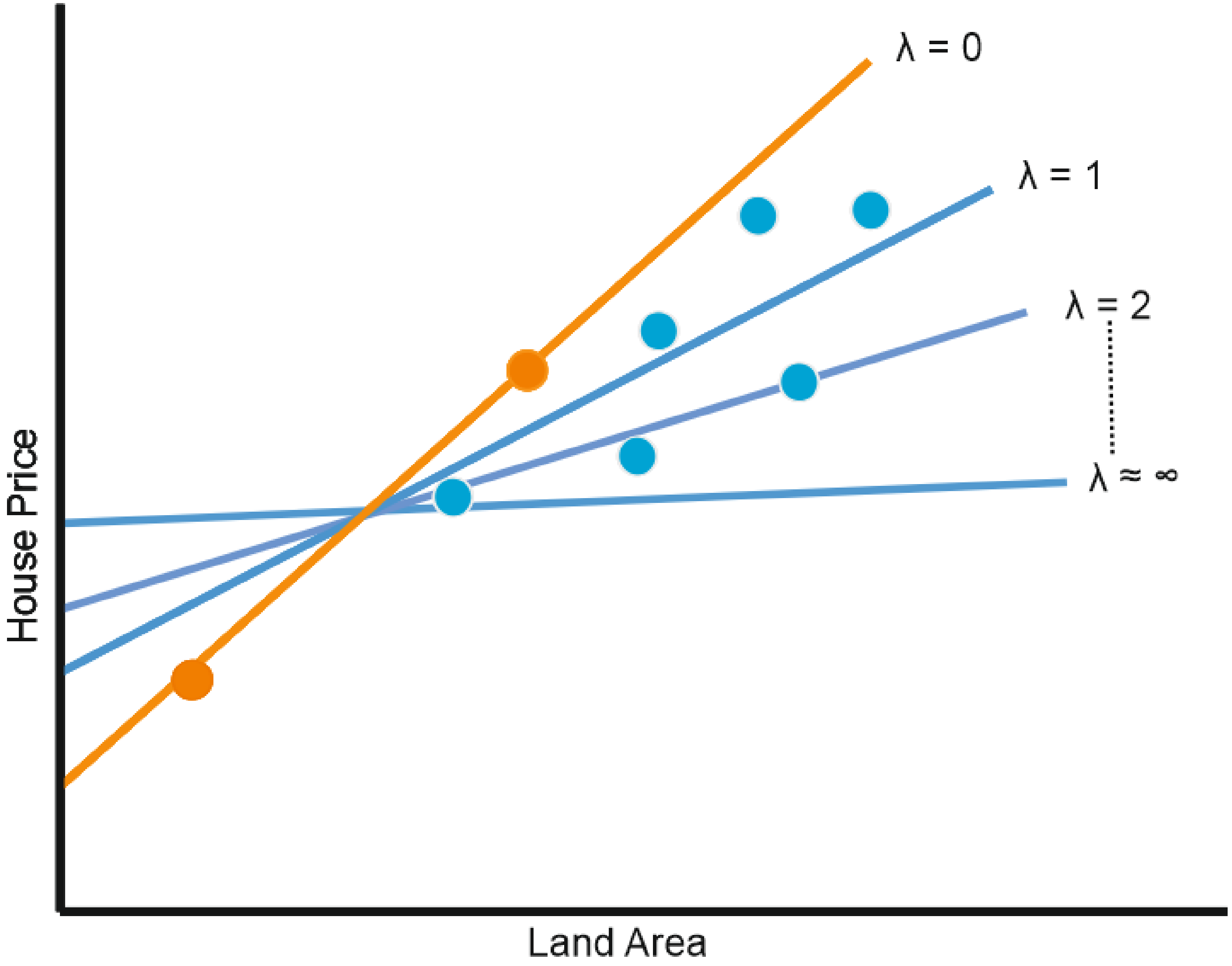
© Copyright by Digital Skola 2022

Linear (Orange) vs Ridge (Blue) Regression

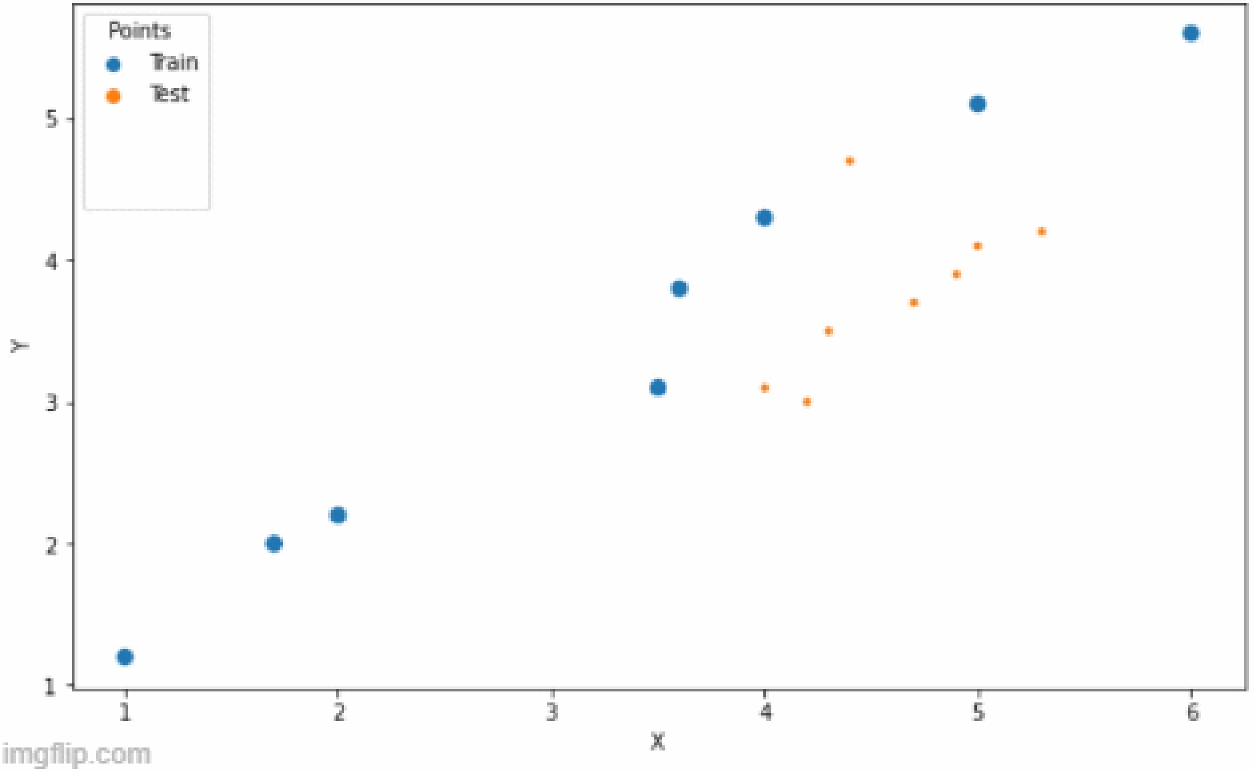




© Copyright by Digital Skola 2022



© Copyright by Digital Skola 2022



© Copyright by Digital Skola 2022

LASSO REGRESSION

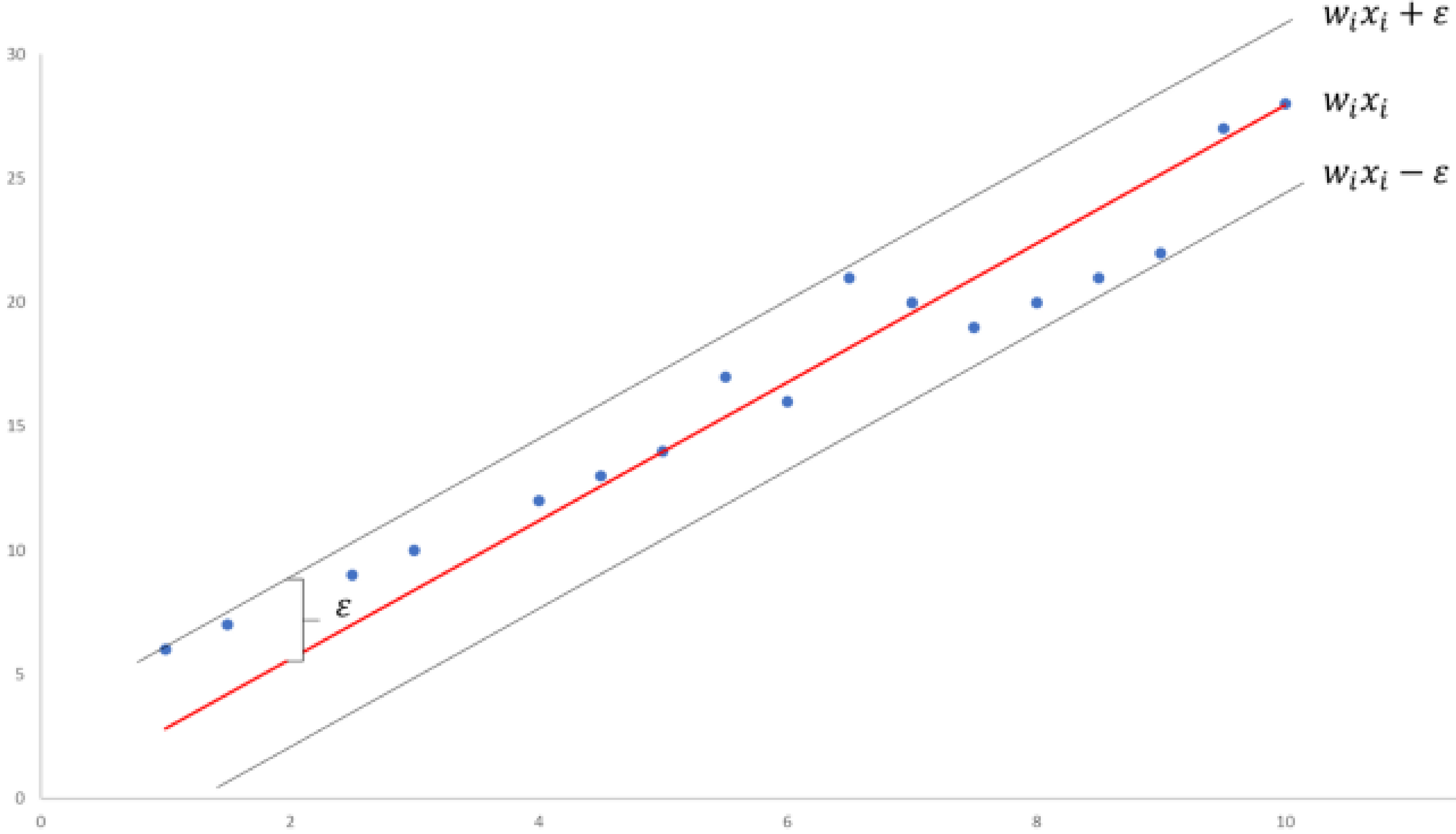
LASSO (Least Absolute Shrinkage Selector Operator), The algorithm is another variation of linear regression like ridge regression. We use lasso regression when we have large number of predictor variables. The equation of LASSO is similar to ridge regression and looks like as given below.

$$SSE + \lambda \sum_{i=1}^{n} |\beta_i|$$

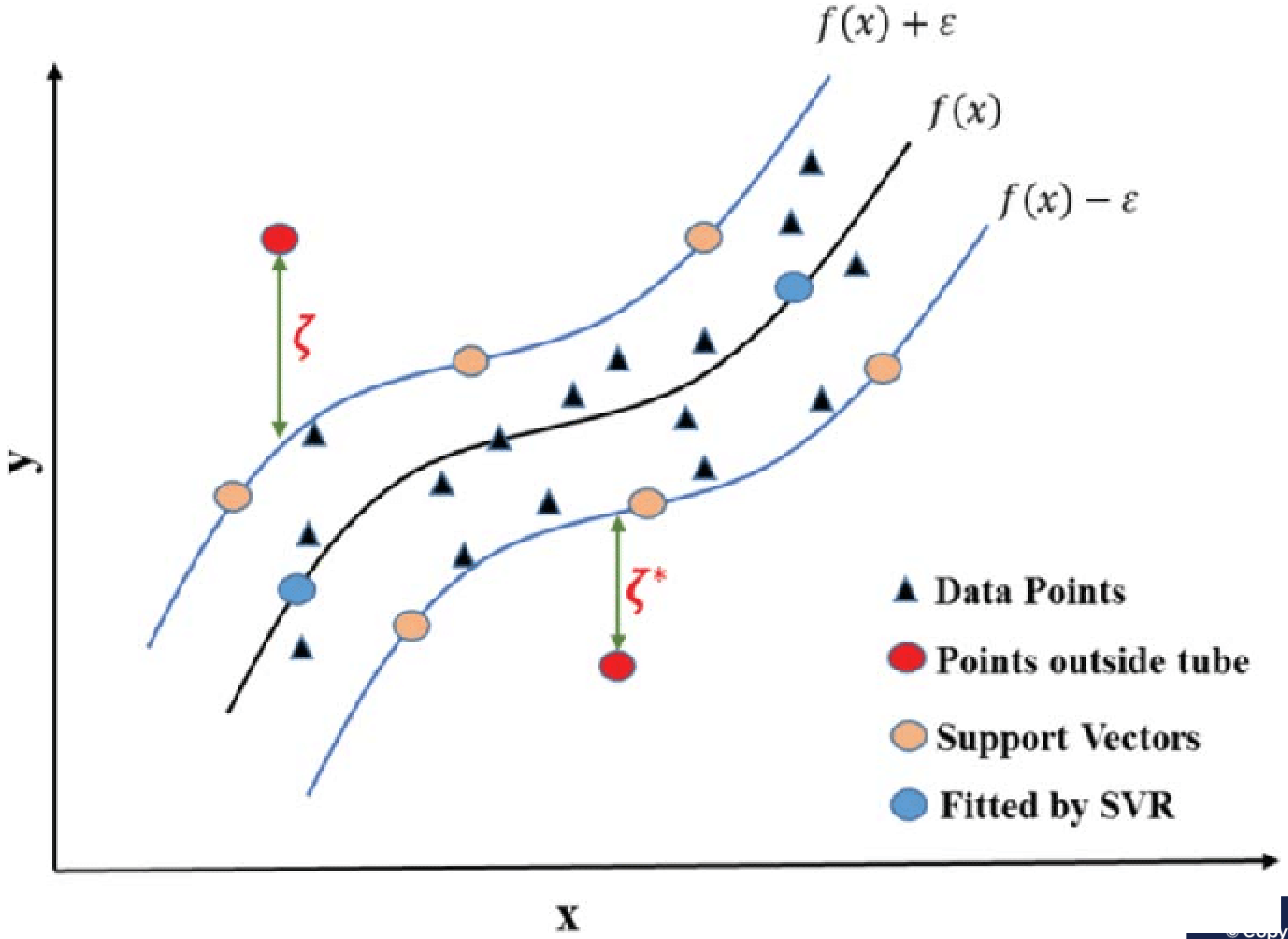
© Copyright by Digital Skola 2022

Support Vector Regression

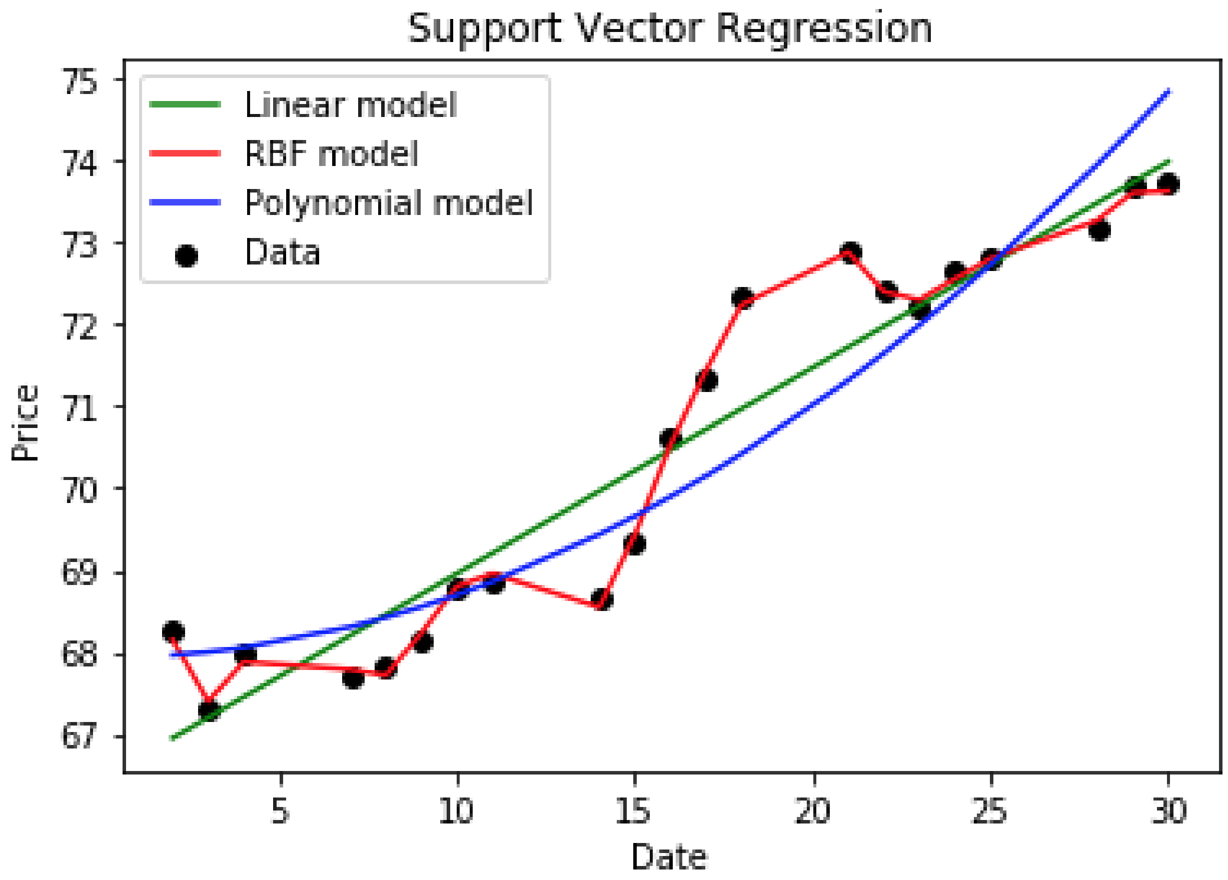
© Copyright by Digital Skola 2022



© Copyright by Digital Skola 2022



© Copyright by Digital Skola 2022



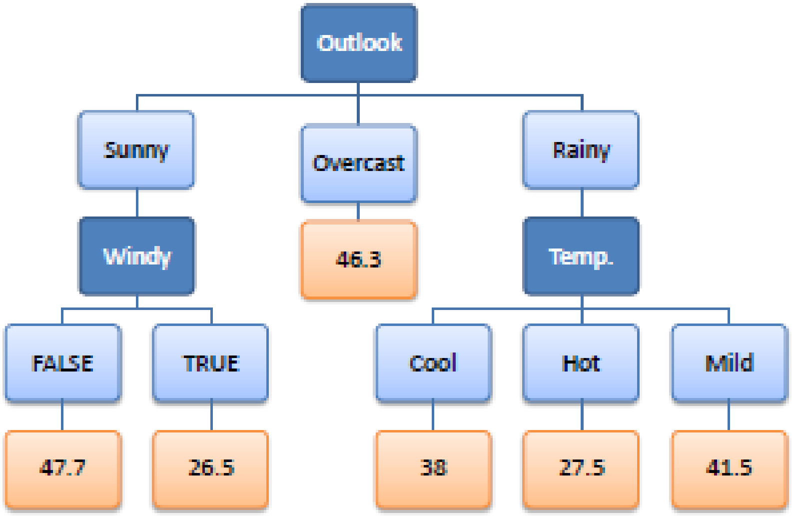
© Copyright by Digital Skola 2022

Decision Tree Regression

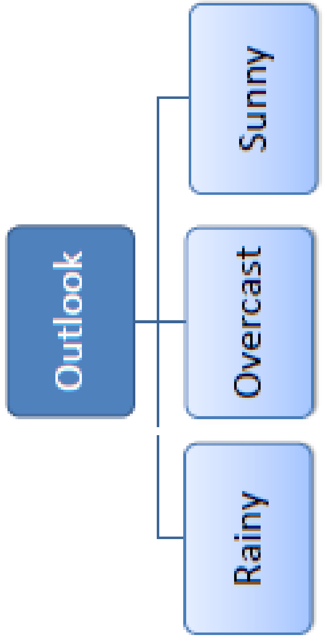
© Copyright by Digital Skola 2022

Predictors | Target

Outlook	Temp.	Humidity	Windy	Hours Played
Rainy	Hot	High	False	26
Rainy	Hot	High	True	30
Overcast	Hot	High	False	48
Sunny	Mild	High	False	46
Sunny	Cool	Normal	False	62
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	36
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	48
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	62
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30



© Copyright by Digital Skola 2022

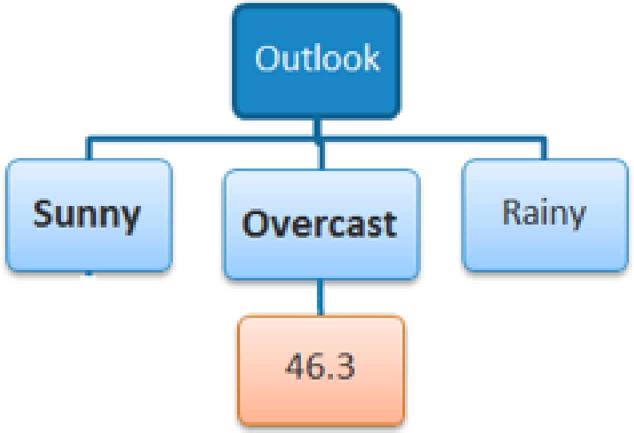


Outlook	Temp	Humidity	Windy	Hours Played
Sunny	Mild	High	FALSE	45
Sunny	Cool	Normal	FALSE	52
Sunny	Cool	Normal	TRUE	23
Sunny	Mild	Normal	FALSE	46
Sunny	Mild	High	TRUE	30
Overcast	Hot	High	FALSE	46
Overcast	Cool	Normal	TRUE	43
Overcast	Mild	High	TRUE	52
Overcast	Hot	Normal	FALSE	44
Rainy	Hot	High	FALSE	25
Rainy	Hot	High	TRUE	30
Rainy	Mild	High	FALSE	35
Rainy	Cool	Normal	FALSE	38
Rainy	Mild	Normal	TRUE	48

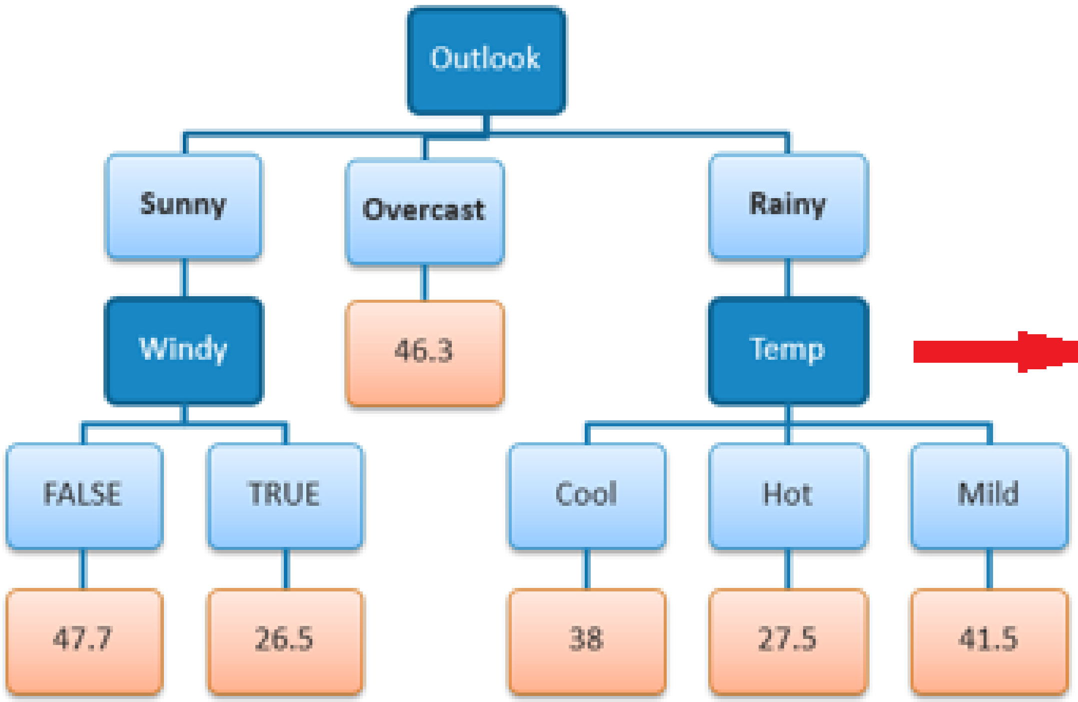
© Copyright by Digital Skola 2022

Outlook - Overcast

		Hours Played (StDev)	Hours Played (AVG)	Hours Played (CV)	Count
Outlook	Overcast	3.49	46.3	8%	4
	Rainy	7.78	35.2	22%	5
	Sunny	10.87	39.2	28%	5



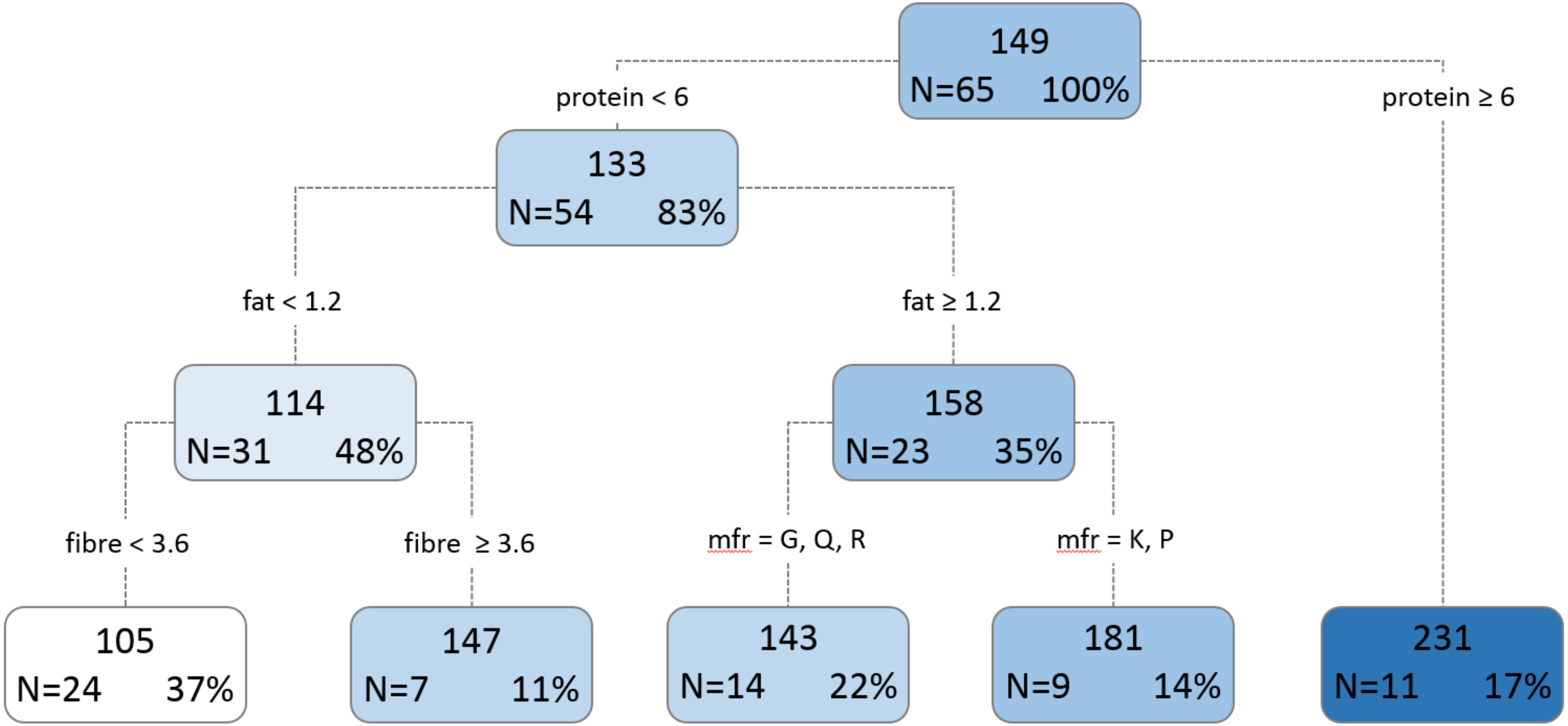
© Copyright by Digital Skola 2022



Temp	Hours Played
Cool	38
Hot	25
Hot	30
Mild	35
Mild	48

© Copyright by Digital Skola 2022

UScereal Calorie Prediction Decision Tree



© Copyright by Digital Skola 2022

Metric and Model Evaluation

© Copyright by Digital Skola 2022

R Square/Adjusted R Square

R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

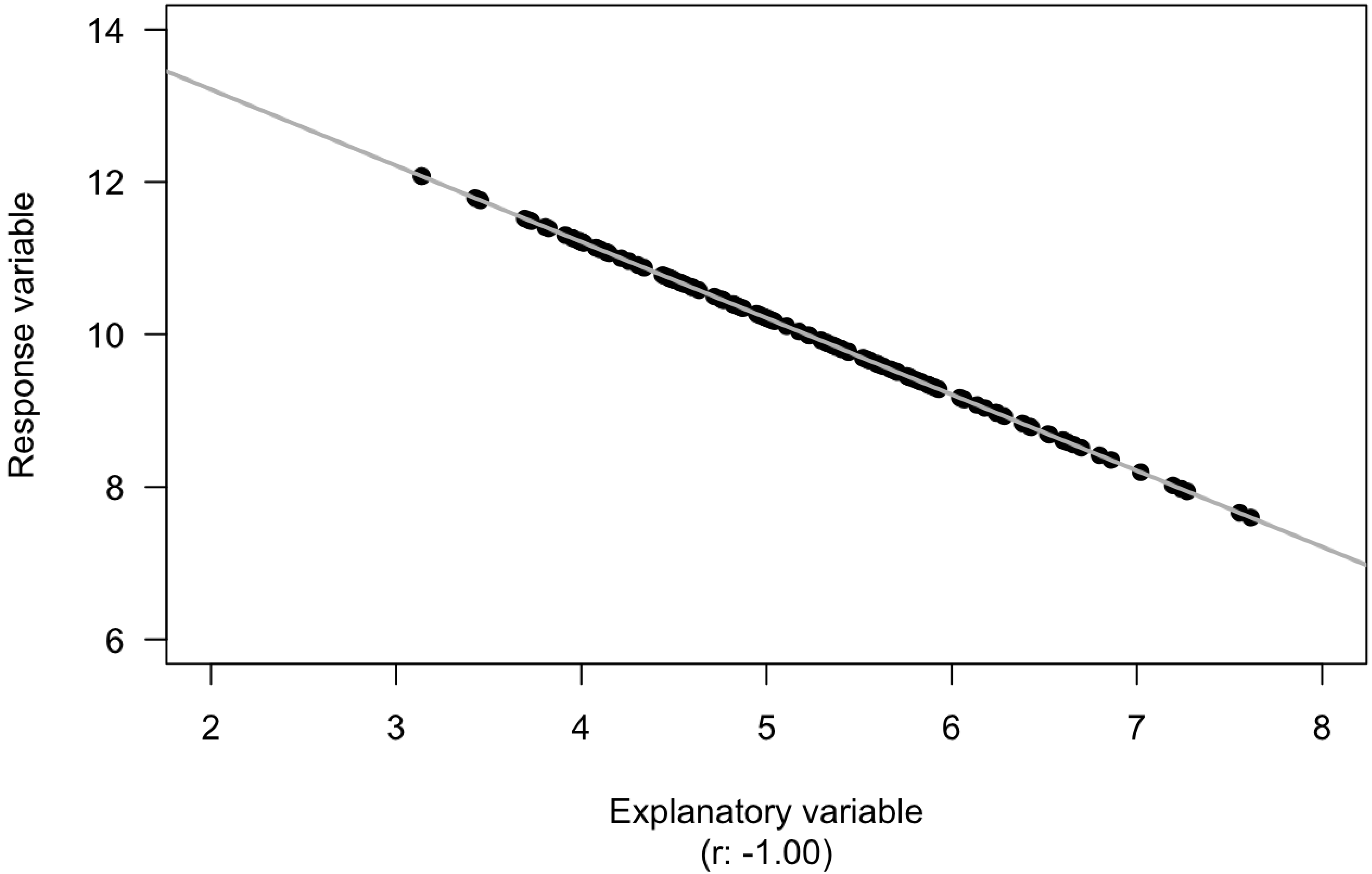
$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$R^2_{adjusted} = \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

R Square is a good measure to determine how well the model fits the dependent variables. **However, it does not take into consideration of overfitting problem**. If your regression model has many independent variables, because the model is too complicated, it may fit very well to the training data but performs badly for testing data. That is why Adjusted R Square is introduced because it will penalize additional independent variables added to the model and adjust the metric to prevent overfitting issues.

© Copyright by Digital Skola 2022

R-squared: 100%



Mean Square Error(MSE)/Root Mean Square Error(RMSE)

While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

© Copyright by Digital Skola 2022

MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. You cannot interpret many insights from one single result, but it gives you a real number to compare against other model results and help you select the best regression model.

Root Mean Square Error(RMSE) is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and makes it easier for interpretation.

© Copyright by Digital Skola 2022

Mean Absolute Error(MAE)

Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

© Copyright by Digital Skola 2022

Thank You!

© Copyright by Digital Skola 2022