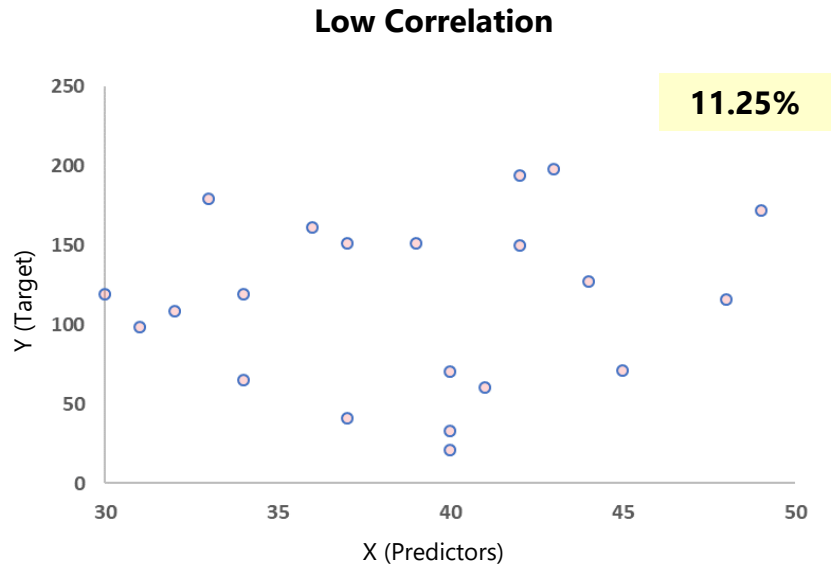
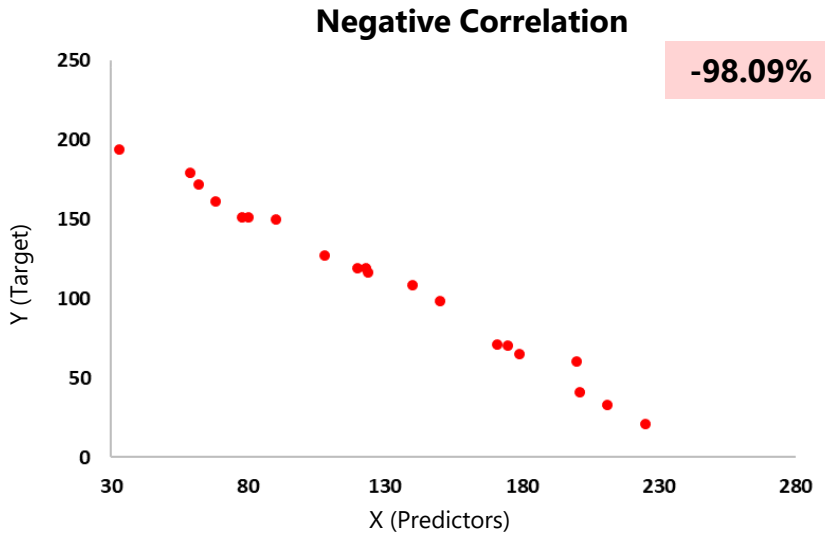
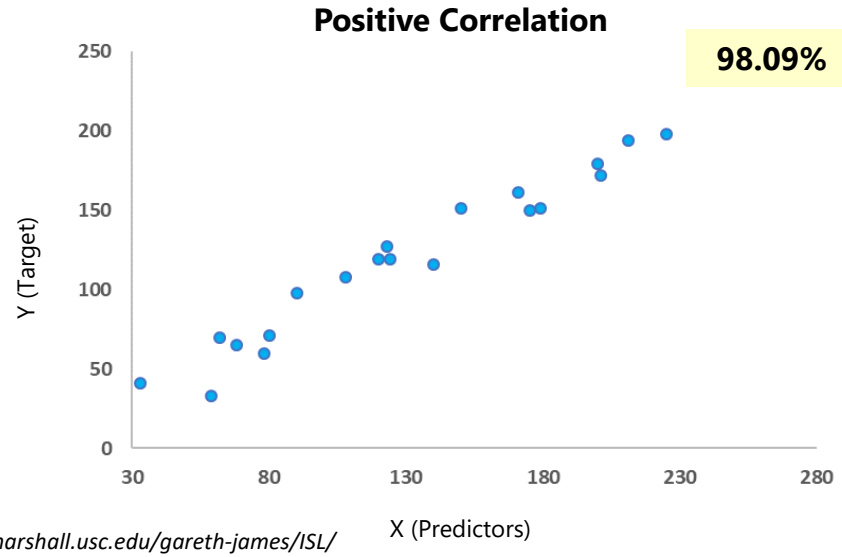
Intro to Regression

Correlation



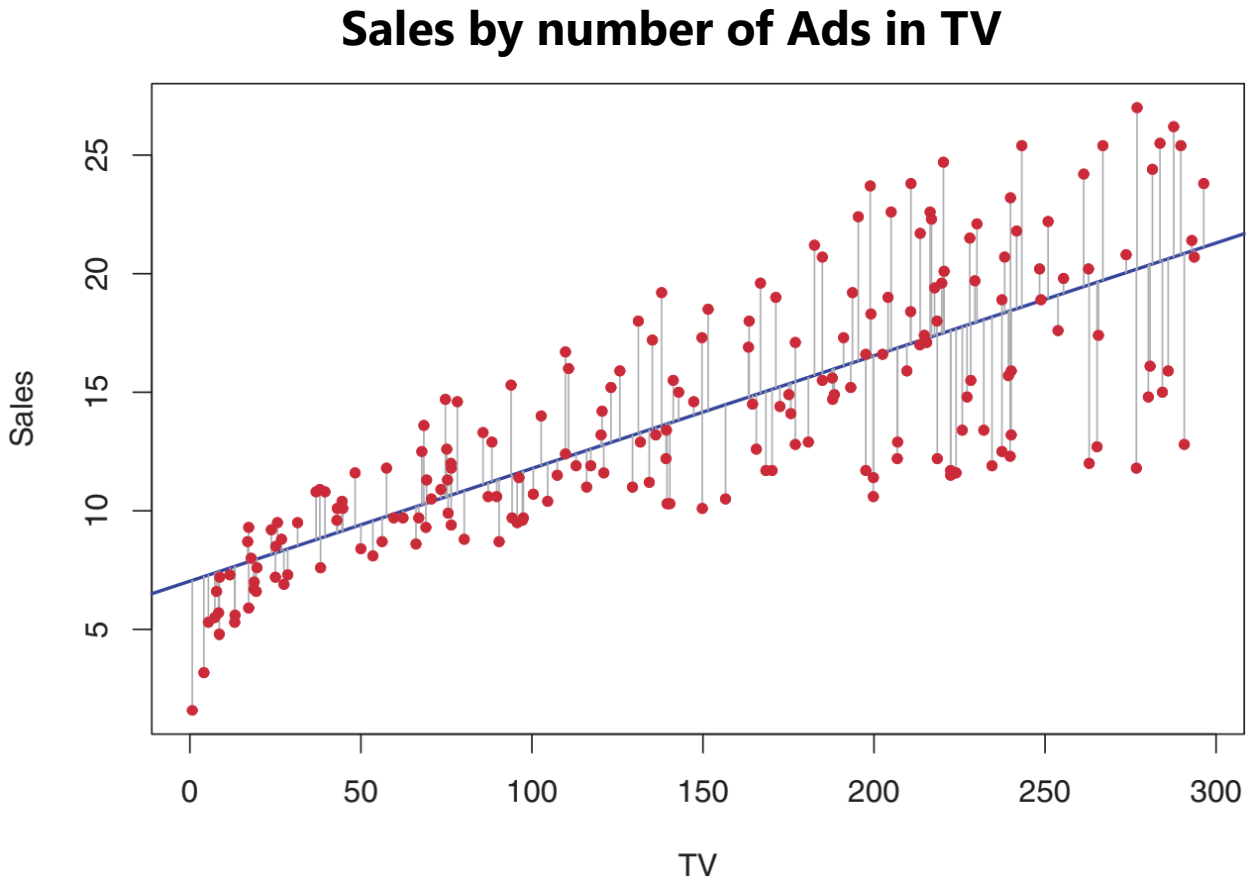




The correlation coefficient can be either Positive (>0) or Negative (<0)

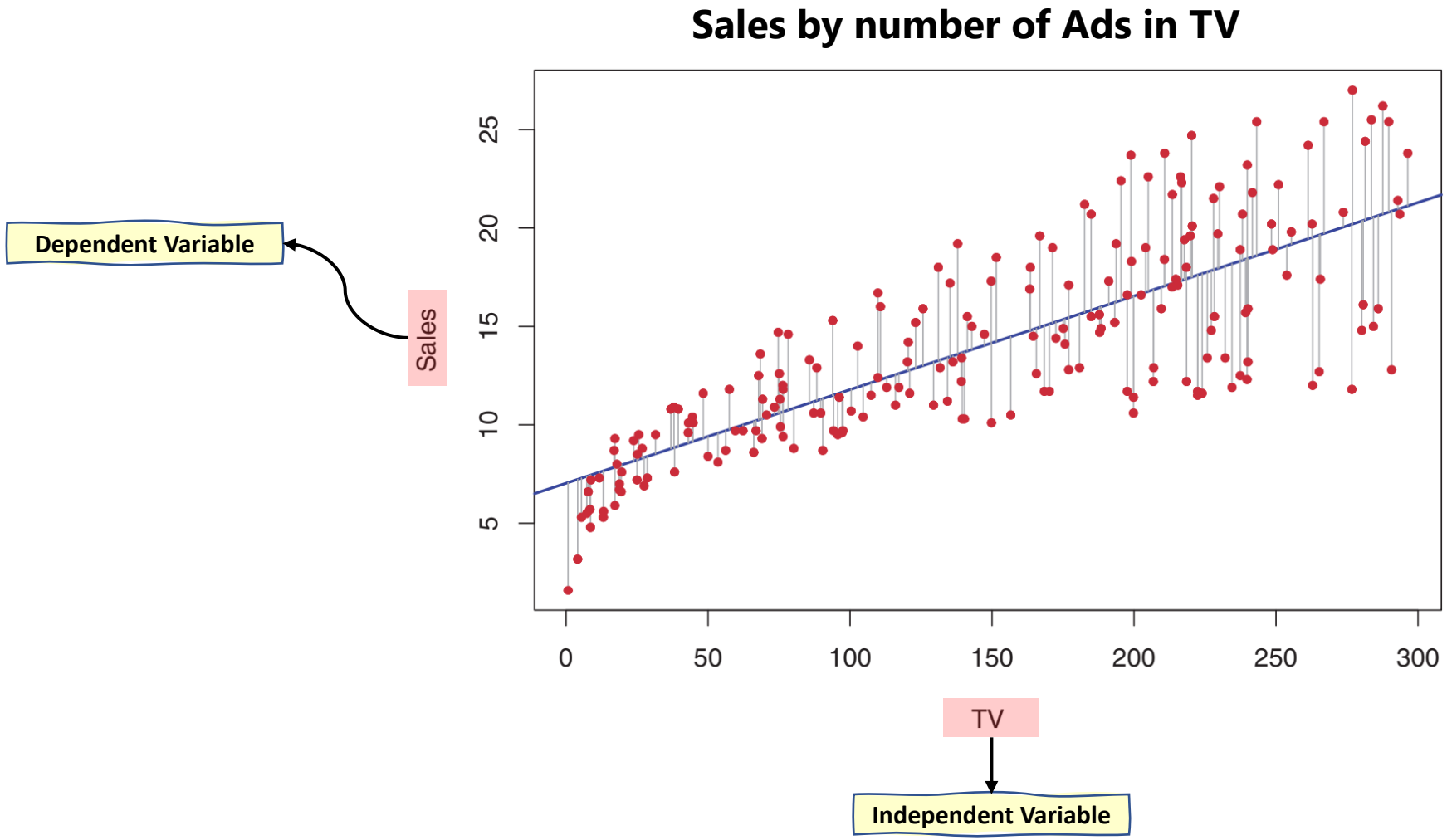
http://faculty.marshall.usc.edu/gareth-james/ISL/

Linear Regression



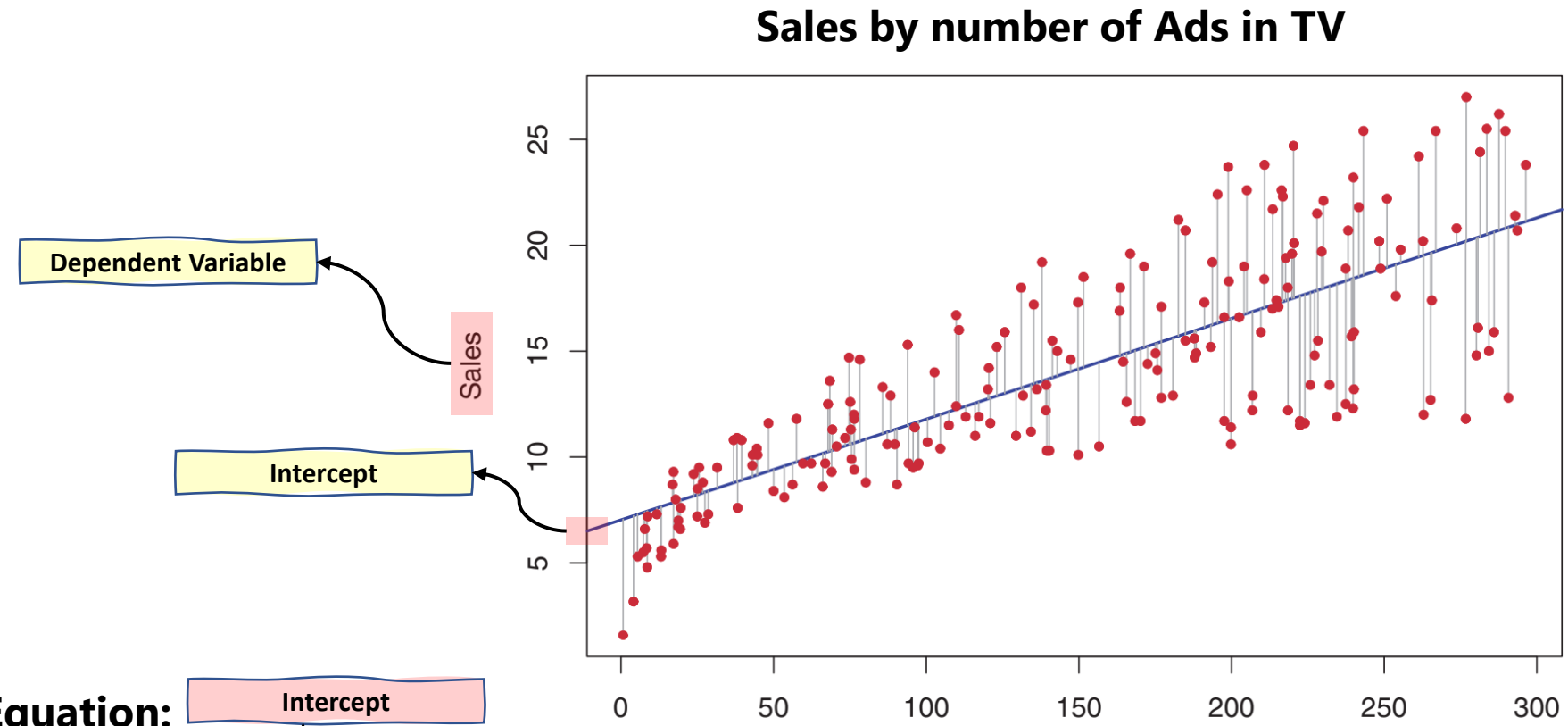
http://faculty.marshall.usc.edu/gareth-james/ISL/

Linear Regression



http://faculty.marshall.usc.edu/gareth-james/ISL/

Linear Regression



Dependent Variable

Intercept

Equation:

Intercept

$$Expected\ Sales \sim B_0 + B_1(Budget)$$

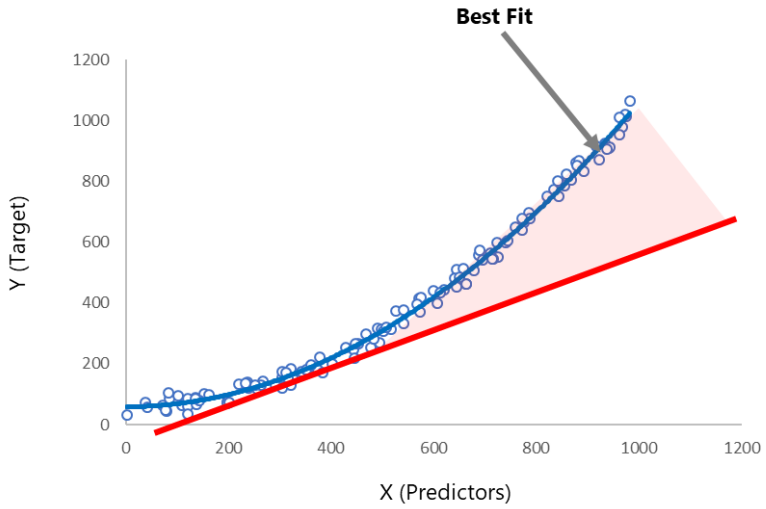
Dependent Variable

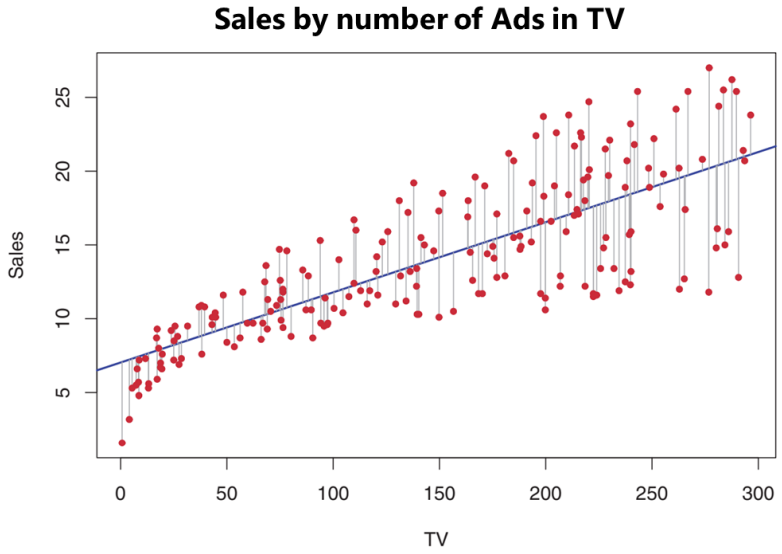
Independent Variable

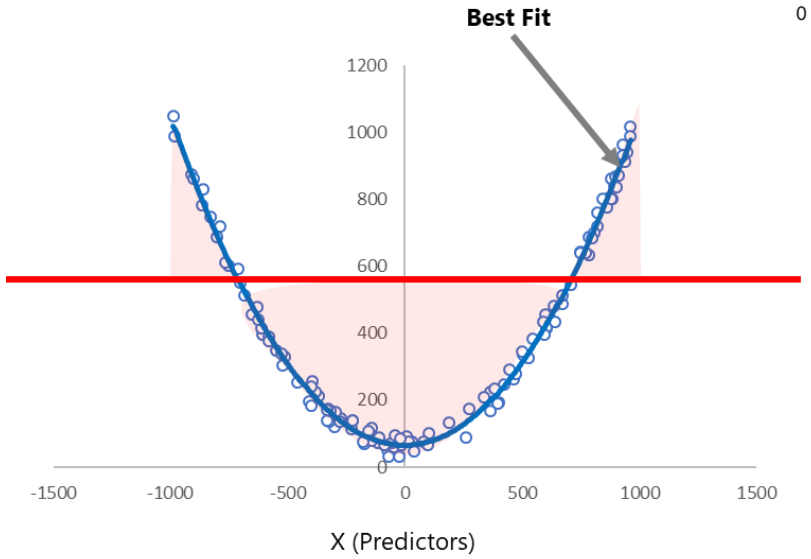
Independent Variable

http://faculty.marshall.usc.edu/gareth-james/ISL/

How do we measure the error on a Regression problem?







http://faculty.marshall.usc.edu/gareth-james/ISL/

How do we measure the error on a Regression problem?

Predicted	Actual Value	Difference	Squared Difference
218.4	212.4	-6.0	36.0
826.6	829.6	3.0	9.0
1,016.4	1,016.4	0.0	0.0
248.8	247.8	-1.0	1.0
173.2	175.2	2.0	4.0
449.8	441.8	-8.0	64.0
409.2	395.2	-14.0	196.0
203.6	196.6	-7.0	49.0
770.3	776.3	6.0	36.0
186.0	192.0	6.0	36.0
198.7	184.7	-14.0	196.0
863.4	862.4	-1.0	1.0
47.3	52.3	5.0	25.0
		RSS	653.0
		rMSE	7.7
		R^2	-99.95%

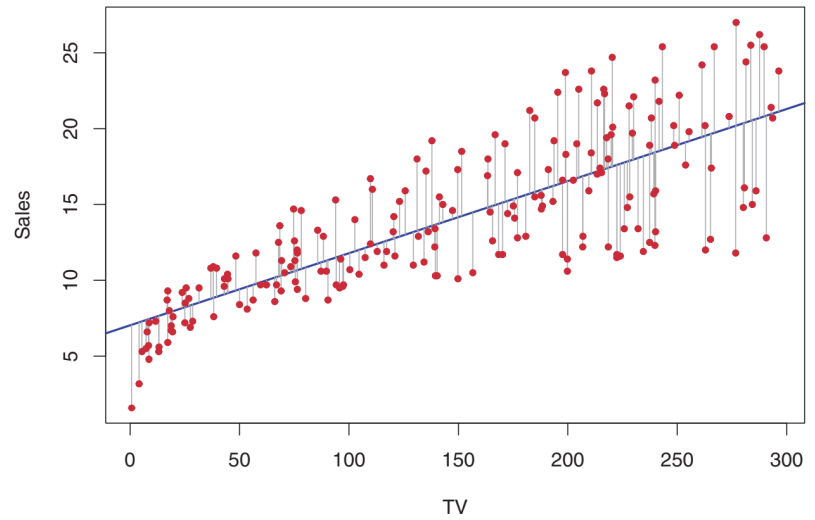
We have the (i) subscript because we need to calculate the sum of those errors by each observation.

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Observation — **Actual Value** — **Predicted Value**

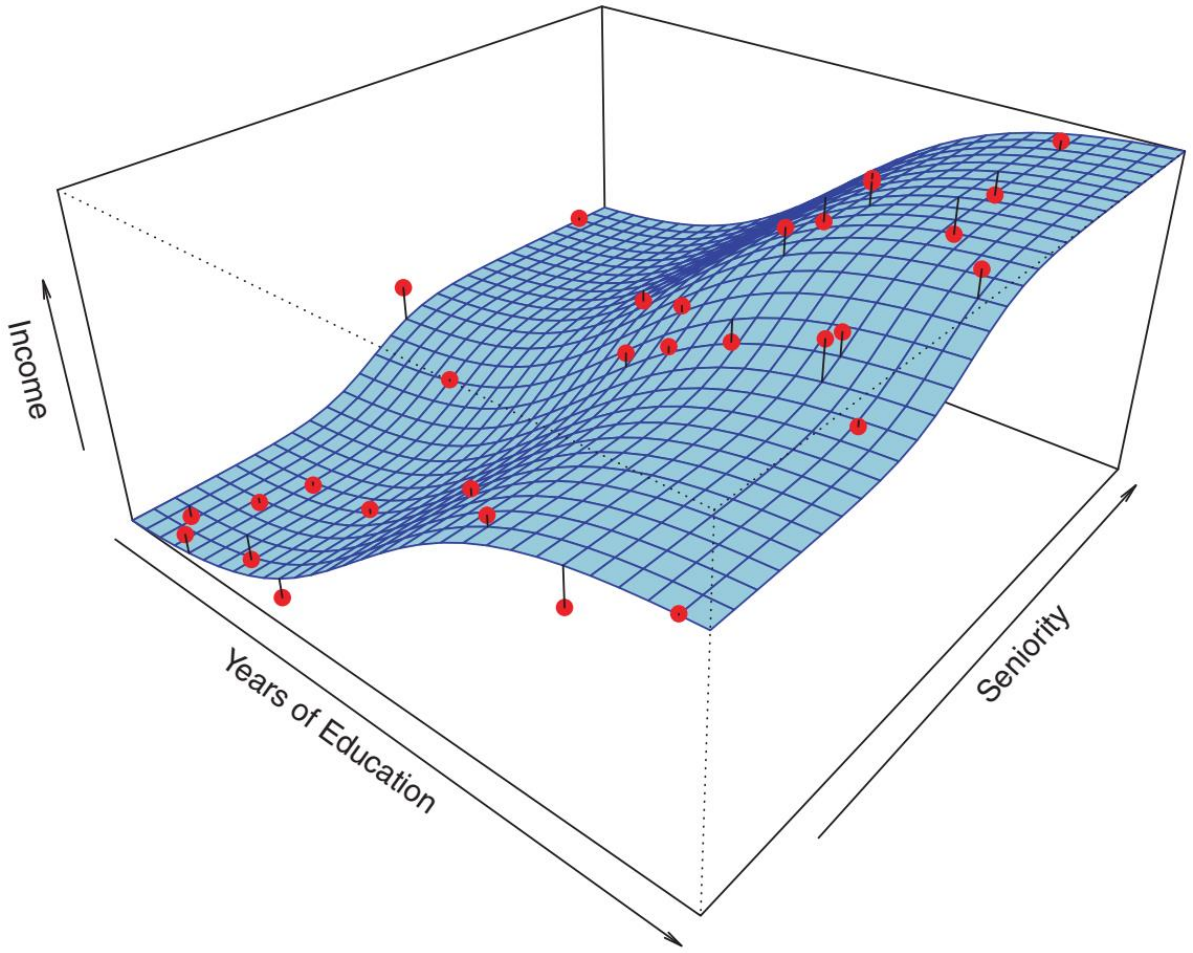
By using the RSS we calculate the **R-Squared,** that basically represents in percentage, how well our model describes our data, in percentage.

From RSS we can also calculate the **rMSE** (root Means Squared Error) which basically tells us the expected deviation of each data point.



http://faculty.marshall.usc.edu/gareth-james/ISL/

Multiple Linear Regression



http://faculty.marshall.usc.edu/gareth-james/ISL/

Multiple Linear Regression

p-value: probability that a variable does not have an important effect on the model. Ideally, we want P to be smaller than 0.05.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

Coefficients: impact (rate of change) that a certain variable has in the model.

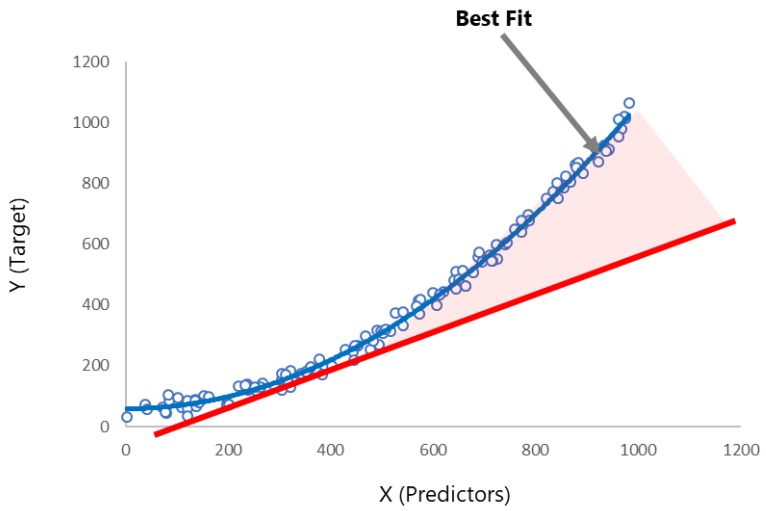
p-value > 0.05

http://faculty.marshall.usc.edu/gareth-james/ISL/

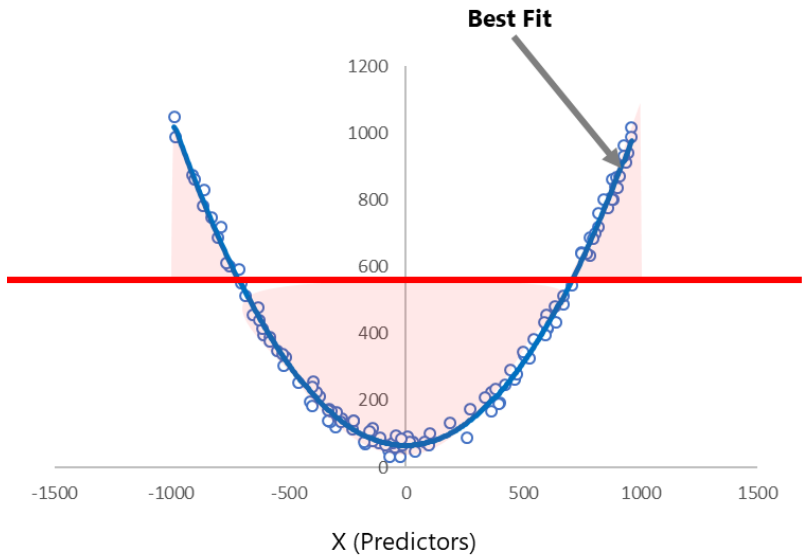
Polynomial Regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i,$$

Exponential Behavior



Quadratic Behavior



http://faculty.marshall.usc.edu/gareth-james/ISL/

Walk-through of concepts

Chapter 3 - Linear Regression

In [39]:
```
# %load ../standard_import.txt
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import axes3d
import seaborn as sns

from sklearn.preprocessing import scale
import sklearn.linear_model as skl_lm
from sklearn.metrics import mean_squared_error, r2_score
import statsmodels.api as sm
import statsmodels.formula.api as smf

%matplotlib inline
plt.style.use('seaborn-white')
```

Load Datasets

Datasets available on http://www-bcf.usc.edu/~gareth/ISL/data.html

In [2]:
```
advertising = pd.read_csv('Data/Advertising.csv', usecols=[1,2,3,4])
advertising.info()
```

https://github.com/JWarmenhoven/ISLR-python

https://github.com/JWarmenhoven/ISLR-python/blob/master/Notebooks/Chapter%203.ipynb

http://faculty.marshall.usc.edu/gareth-james/ISL/

Lab

This lab on Linear Regression is a python adaptation of p. 109-119 of "Introduction to Statistical Learning with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Written by R. Jordan Crouser at Smith College for SDS293: Machine Learning (Spring 2016).

3.6.1 Importing Libraries

In [2]:
```
# Tells matplotlib to display images inline instead of a new window
%matplotlib inline

import numpy as np
import pandas as pd
import statsmodels.api as sm
```

https://github.com/jcrouser/islr-python

https://github.com/jcrouser/islr-python/blob/master/Lab%202%20-%20Linear%20Regression%20in%20Python.ipynb

http://faculty.marshall.usc.edu/gareth-james/ISL/