Prediction

Business Analytics

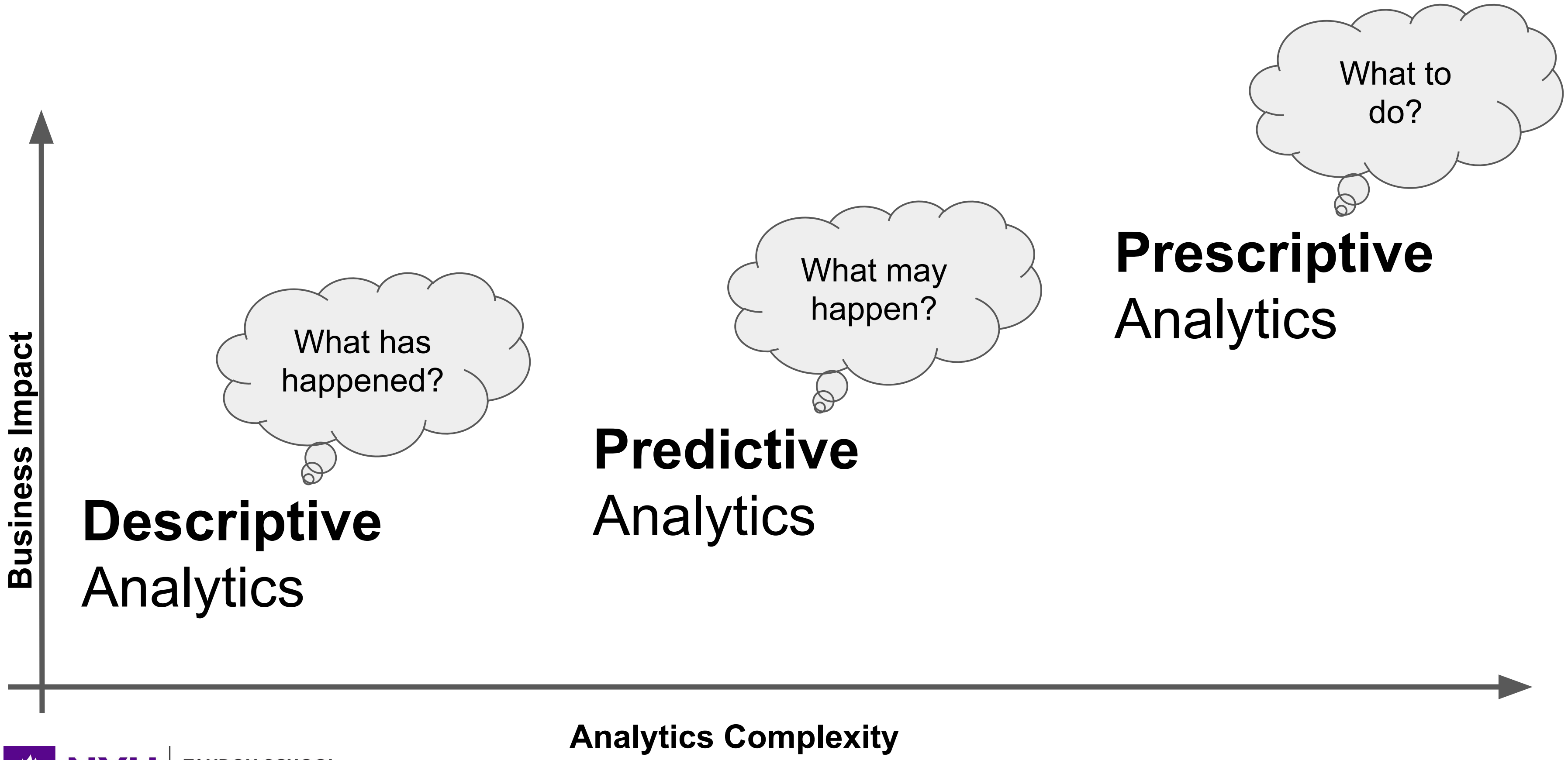
Business Analytics Definition

Business analytics refers to the application of data analysis and modeling techniques for understanding business situations and improving business decisions.

IMPLICATIONS:

- data → past business performance
- methods → statistics + mathematics + computational methods
- business decisions → actionable insight

Types of Analytics



Understanding Your Data

Exploratory analysis of data is useful for:

- understanding data properties
- detecting errors, ensuring data quality
- finding patterns in data
- determining relationships among variables
- checking assumptions
- mapping business problems into data mining tasks and suggesting modeling strategies

Lesson Objectives

1. Regression - Theory
   a. Linear Models
   b. Ordinary Least Squares
   c. Simple Linear Regression
2. Regression Applied
   a. Model strength
   b. Model interpretation
   c. Dummy variables
   d. Non-linear transformations
3. Classification
   a. Statistical classification
   b. Decision Trees

Linear Models

Regression models estimate the relationships among variables to predict outcomes.

Example: How does bike trip duration change as we introduce a new customer type, a new pricing scheme, or with different weather conditions.

In this week you will learn the basics of regression analysis and the specifics about linear regression models that example the relationship between numerical variables.

Business Case:

What is the influence of a variable (price, advertising, and etc.) on an outcome (market shares, sales, overall satisfaction)?

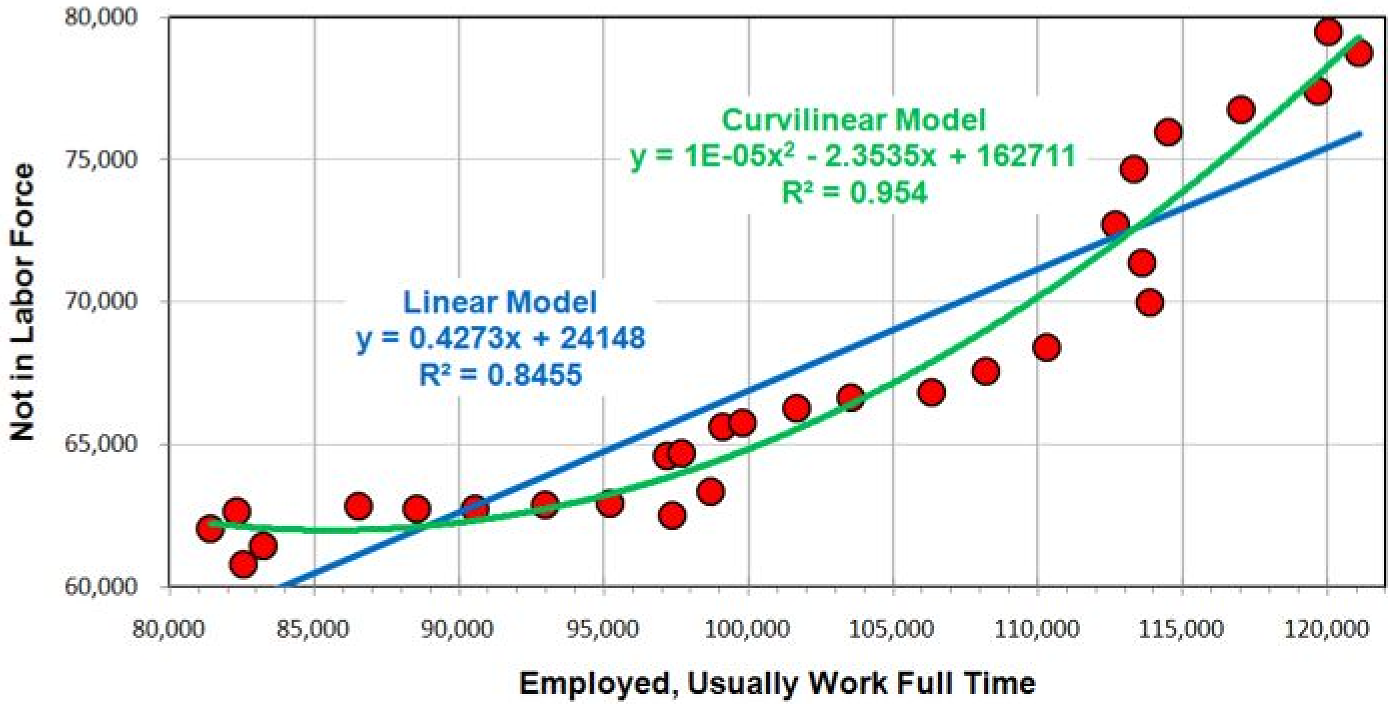
$$X \rightarrow Y$$

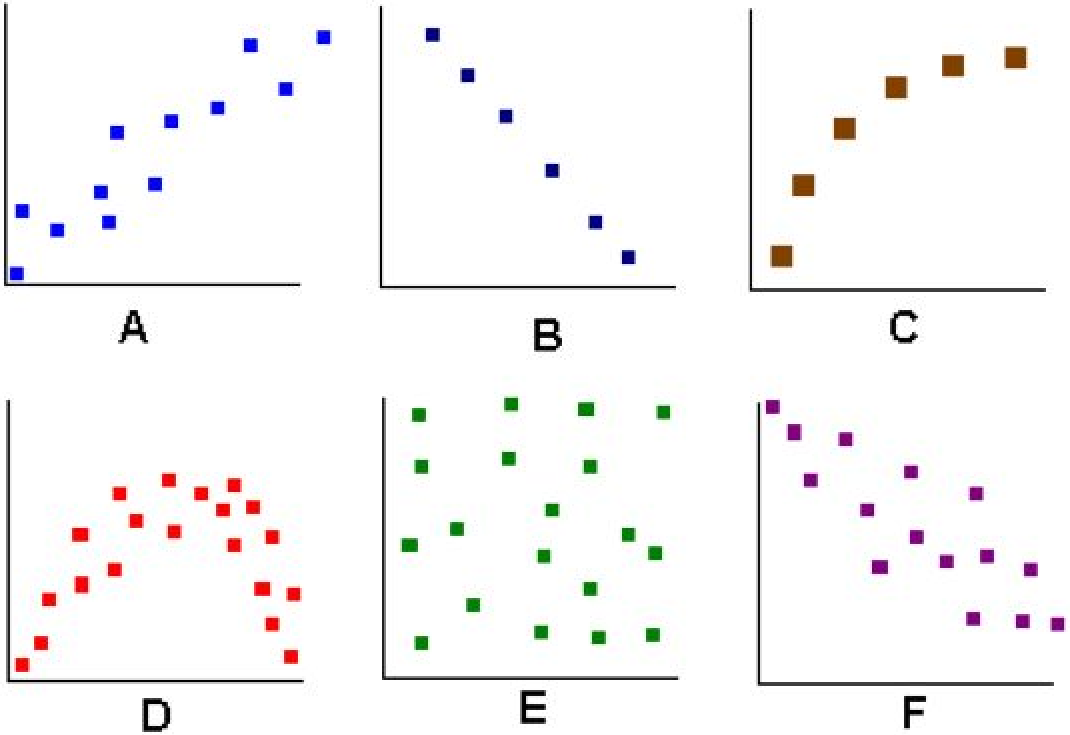
Independent variable (X) → Dependant variable (Y)

$$y=mx+b$$

NOTE IN TERMINOLOGY

- *Y is know as the dependent variable the variable that regression model seek to predict or response variable*
- *X is the independent variable, predictor or explanatory variable.*


Curvilinear Model
y = 1E-05x² - 2.3535x + 162711
R² = 0.954
Linear Model
y = 0.4273x + 24148
R² = 0.8455
Not in Labor Force
80,000
75,000
70,000
65,000
60,000
80,000
85,000
90,000
95,000
100,000
105,000
110,000
115,000
120,000
Employed, Usually Work Full Time


A
B
C
D
E
F

Simple Linear Regression

- The "workhorse" of statistical analysis is the simple linear regression.
- Used to determine the relationship between two variables.
  - Given one variable, a regression will provide the expected value of the other variable.
- The outcome of the regression → Y: response.
- The input variable → X: predictor.

$$Y_i = b_0 + b_1 X_i + \varepsilon_i, \; i = 1, \ldots, n$$

where:

Y_i = ith observation of the dependent variable, Y

X_i = ith observation of the independent variable, X

b_0 = regression intercept term

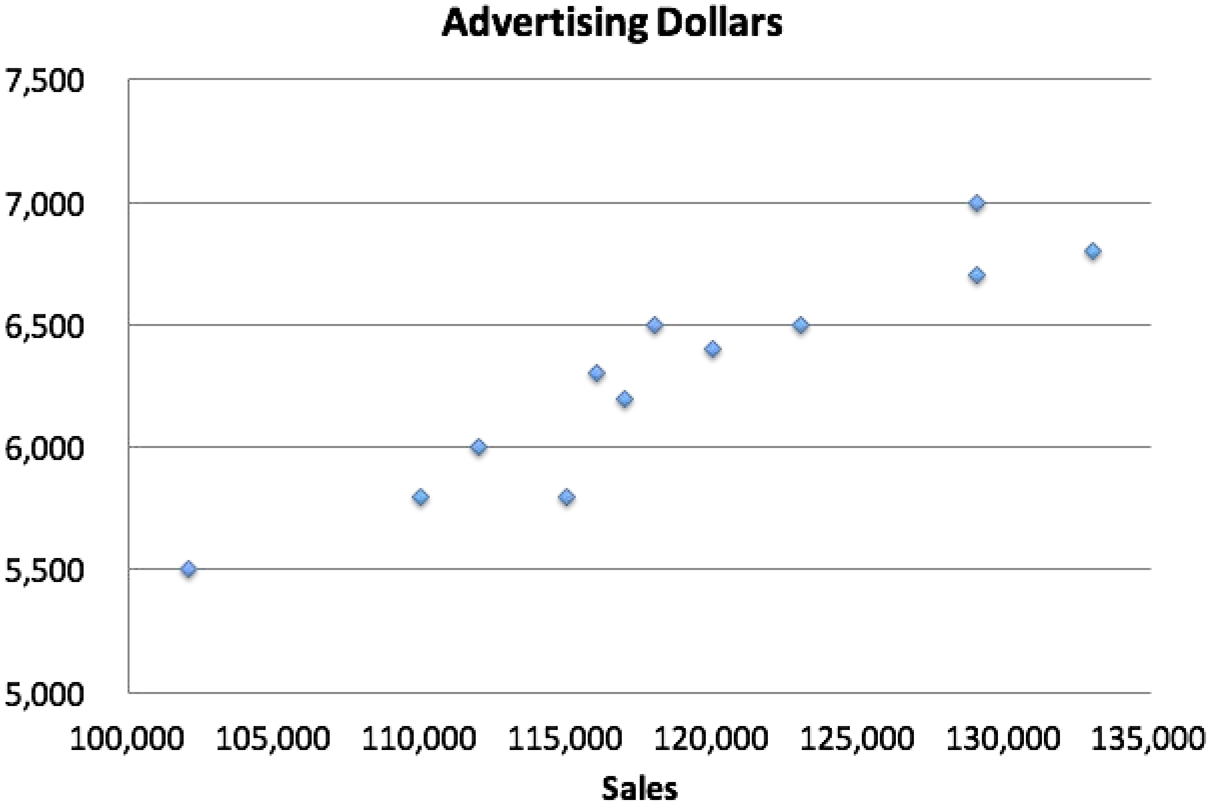
b_1 = regression slope coefficient

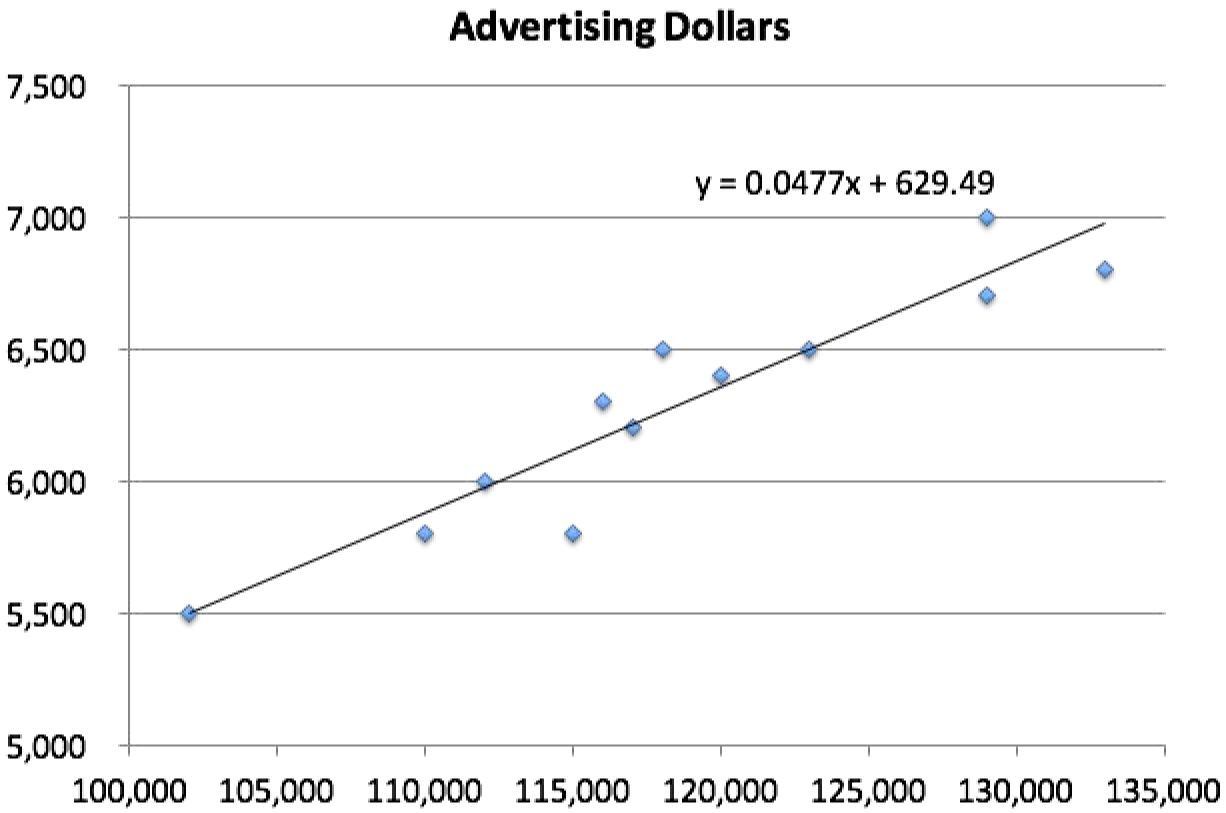
ε_i = residual for the ith observation (also referred to as the disturbance term or error term)

Sales vs Advertising

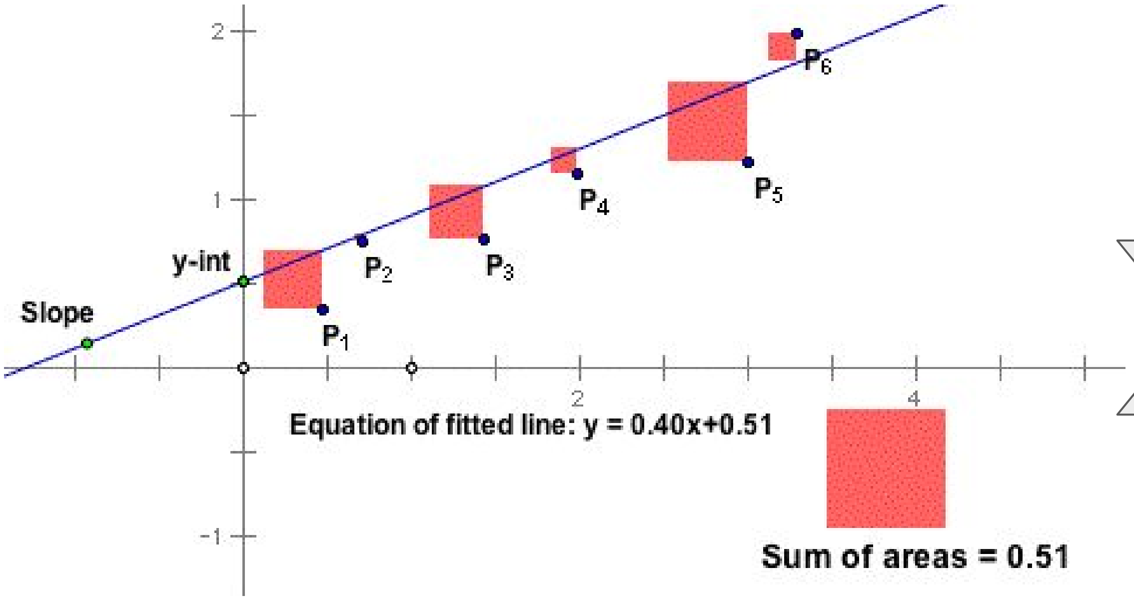
Month	Sales	Advertising Dollars
Jan	102,000	5,500
Feb	110,000	5,800
Mar	112,000	6,000
Apr	115,000	5,800
May	117,000	6,200
Jun	116,000	6,300
Jul	118,000	6,500
Aug	129,000	7,000
Sep	123,000	6,500
Oct	120,000	6,400
Nov	129,000	6,700
Dec	133,000	6,800

Scatter


Advertising Dollars
7,500
7,000
6,500
6,000
5,500
5,000
100,000
105,000
110,000
115,000
120,000
125,000
130,000
135,000
Sales


Advertising Dollars
7,500
7,000
6,500
6,000
5,500
5,000
y = 0.0477x + 629.49
100,000
105,000
110,000
115,000
120,000
125,000
130,000
135,000

Ordinary Least Squares



R^2

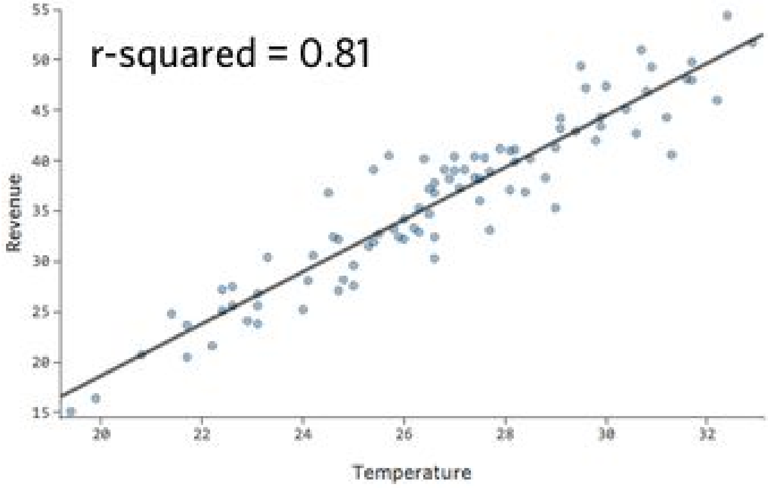
R^2 is an statistical measure of how close the data are to the fitted regression line.

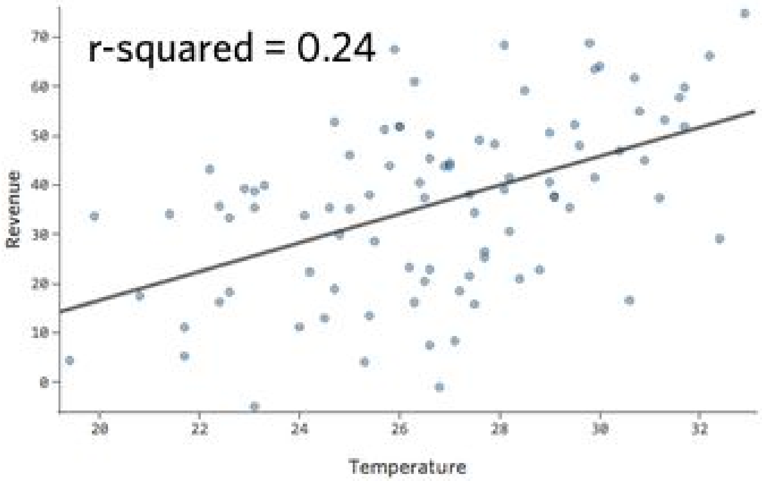
It indicates the goodness of fit of the model.

R^2 definition: Explained variation / Total variation

R^2 is always between 0 and 100%:

- 0% → model explains none of the variability
- 100% → model explains all the variability


r-squared = 0.81
Revenue
Temperature
55
50
45
40
35
30
25
20
15
20
22
24
26
28
30
32


r-squared = 0.24
Revenue
Temperature
70
60
50
40
30
20
10
0
20
22
24
26
28
30
32

Calculating b1: 0.0477

x-mean(x)	y-mean(y)	(x-mean(x))* (y-mean(y))	(x-mean(x))^2
-16,666.67	-791.67	13,194,444.44	277,777,777.78
-8,666.67	-491.67	4,261,111.11	75,111,111.11
-6,666.67	-291.67	1,944,444.44	44,444,444.44
-3,666.67	-491.67	1,802,777.78	13,444,444.44
-1,666.67	-91.67	152,777.78	2,777,777.78
-2,666.67	8.33	-22,222.22	7,111,111.11
-666.67	208.33	-138,888.89	444,444.44
10,333.33	708.33	7,319,444.44	106,777,777.78
4,333.33	208.33	902,777.78	18,777,777.78
1,333.33	108.33	144,444.44	1,777,777.78
10,333.33	408.33	4,219,444.44	106,777,777.78
14,333.33	508.33	7,286,111.11	205,444,444.44

Calculating b0: 629.4926

Intercept

$$\hat{y}_i = b_0 + b_1 x_i$$

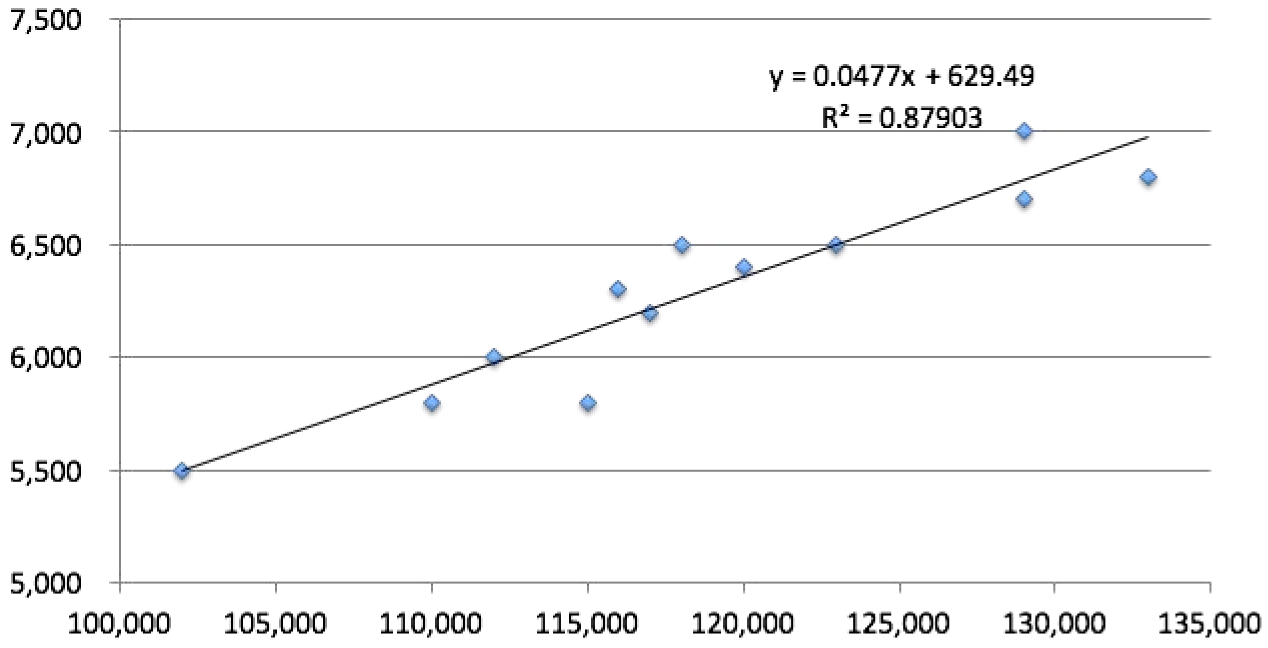
Slope

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$R^2 = \left(\frac{\sum(xy) - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x^2)}{n}\right]\left[\sum y^2 - \frac{(\sum y^2)}{n}\right]}} \right)^2$$

Advertising Dollars

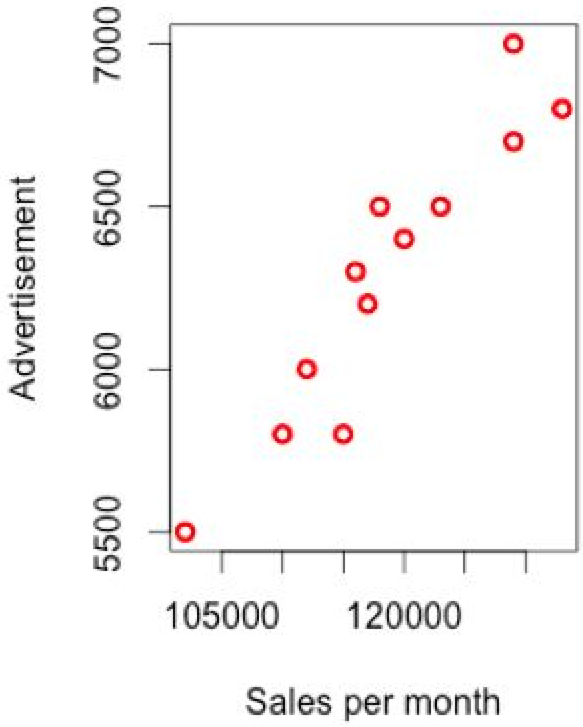


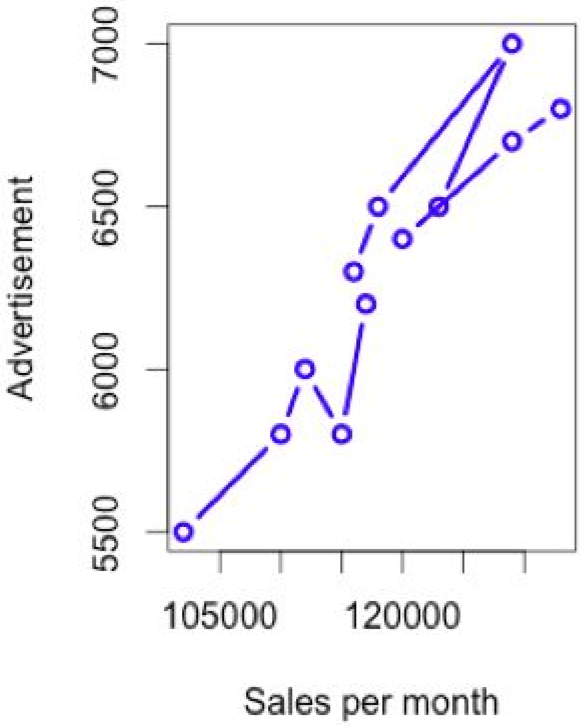
R Code for linear models

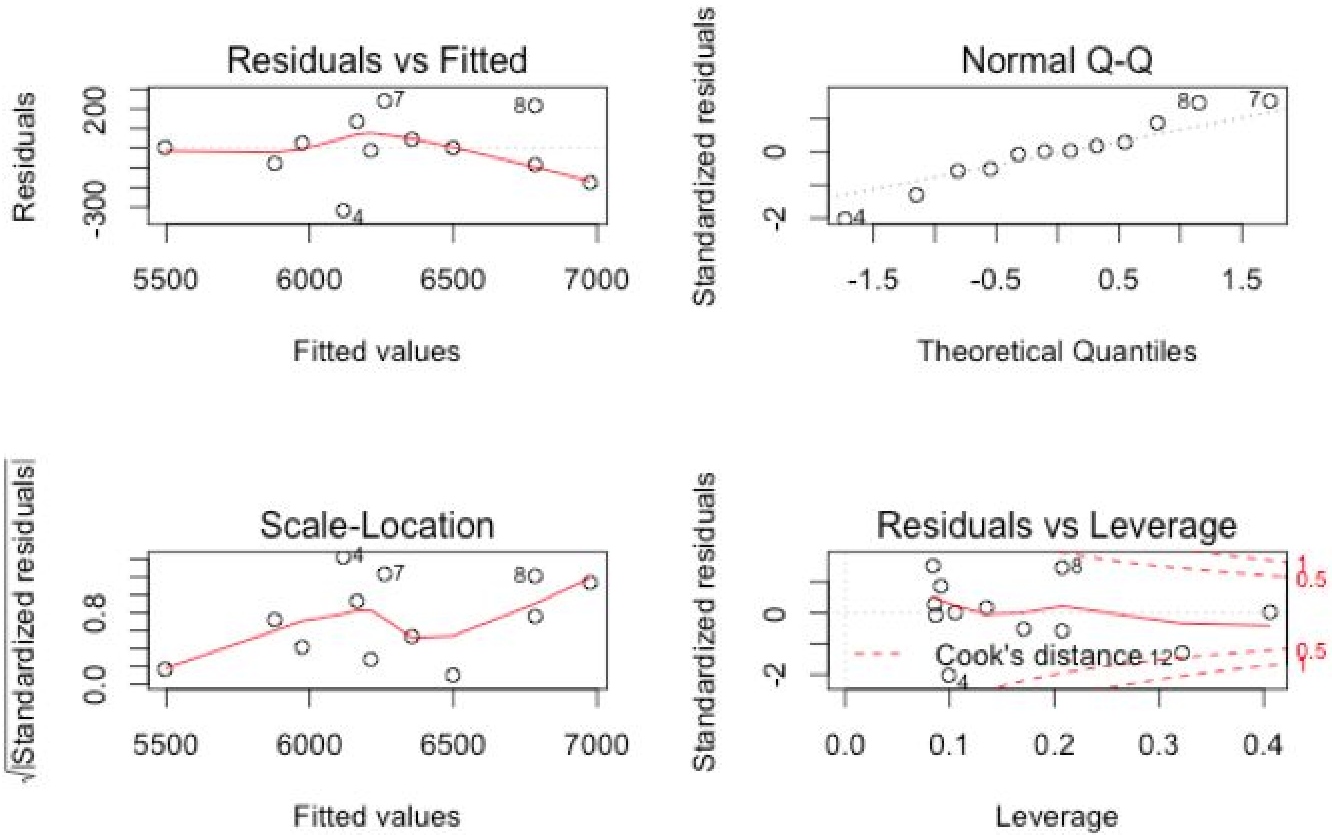
```
data <- read.csv("~/Google Drive/Business Analytics/Data/Sales vs Advertisement.csv",
     header=TRUE, stringsAsFactors=TRUE)
dim(data)
names(data)
x=data$Sales
y=data$Advertising.Dollars
par(mfrow=c(1,2))
 plot(x,y , col="red", lwd=3,
     ylab="Advertisement", xlab="Sales per month")
plot(x,y, type="b", col="blue", lwd=3,
     ylab="Advertisement", xlab="Sales per month")

model<-lm(y ~ x)
model
summary(model)

par(mfrow=c(2,2))
plot(model)
```


Advertisement
5500
6000
6500
7000
105000
120000
Sales per month


Advertisement
5500
6000
6500
7000
105000
120000
Sales per month


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

R Session

Build a linear model for Zagat using “Food” as and predictors and “Price” as a response.

Build a linear model for Zagat using “Food” and “Decor” as and predictors and “Price” as a response. Hint use *lm(y ~ x1+x2)*

Build a linear model for Zagat using “Food”, “Decor”, and “Service” as and predictors and “Price” as a response. Hint use *lm(y ~ x1+x2+x3)*