Introduction to

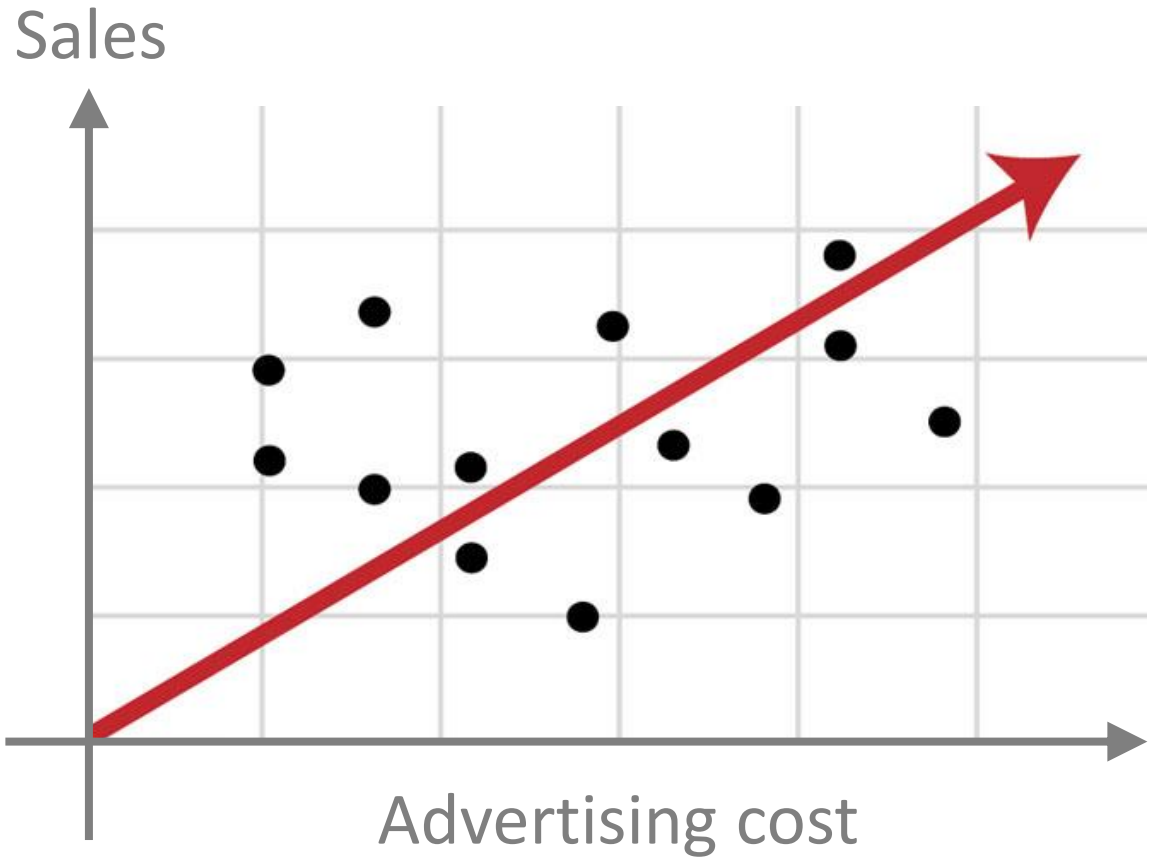
Big Data

in Business

Dr.Pongrapee Kaewsaiha

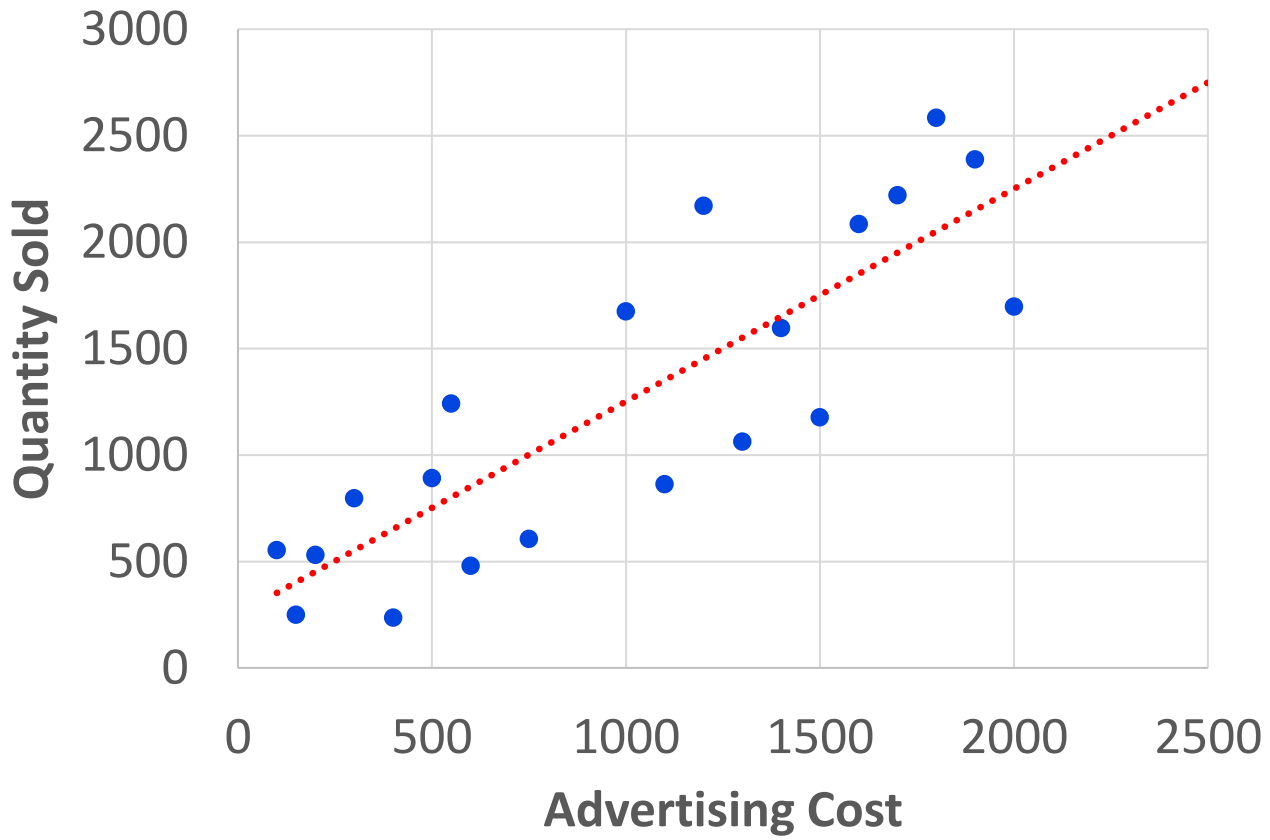
College of Hospitality Industry Management
Suan Sunandha Rajabhat University

Regression Analysis

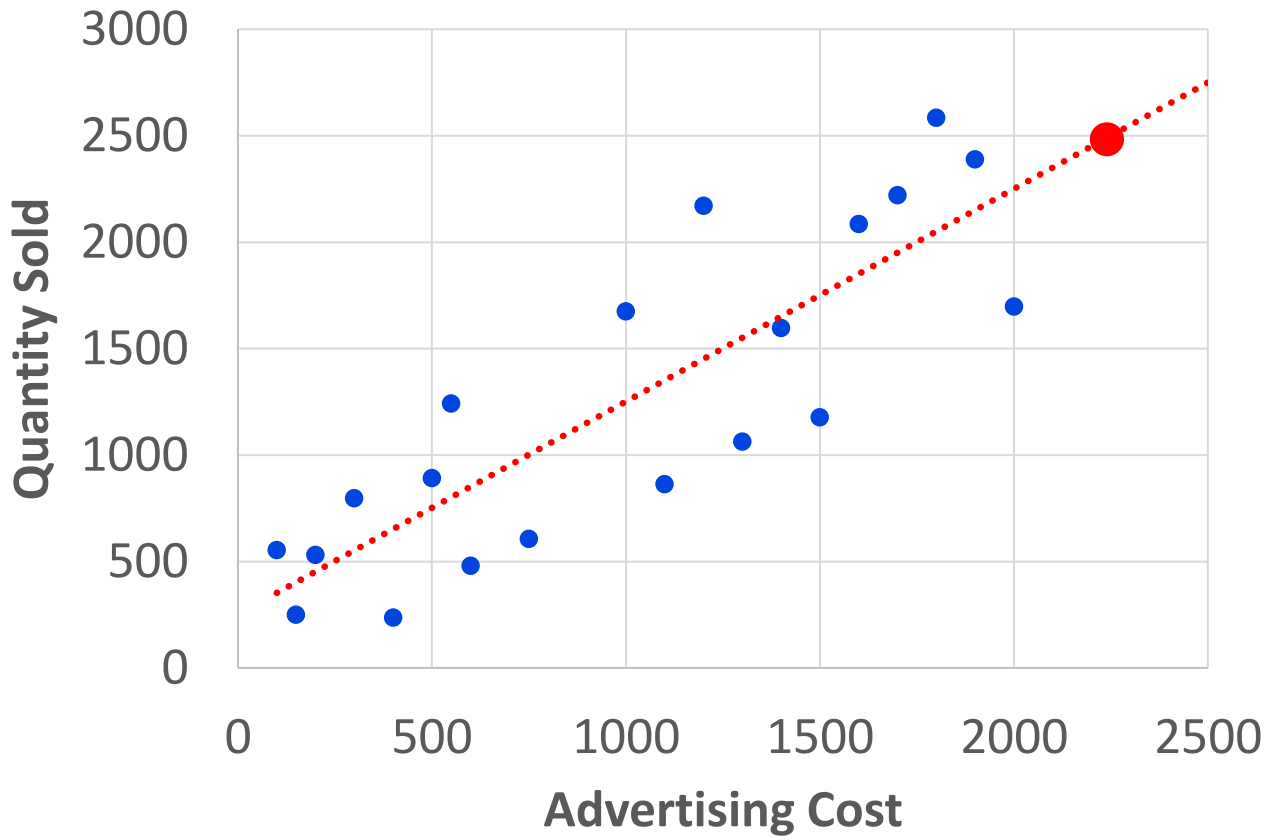


Example 1

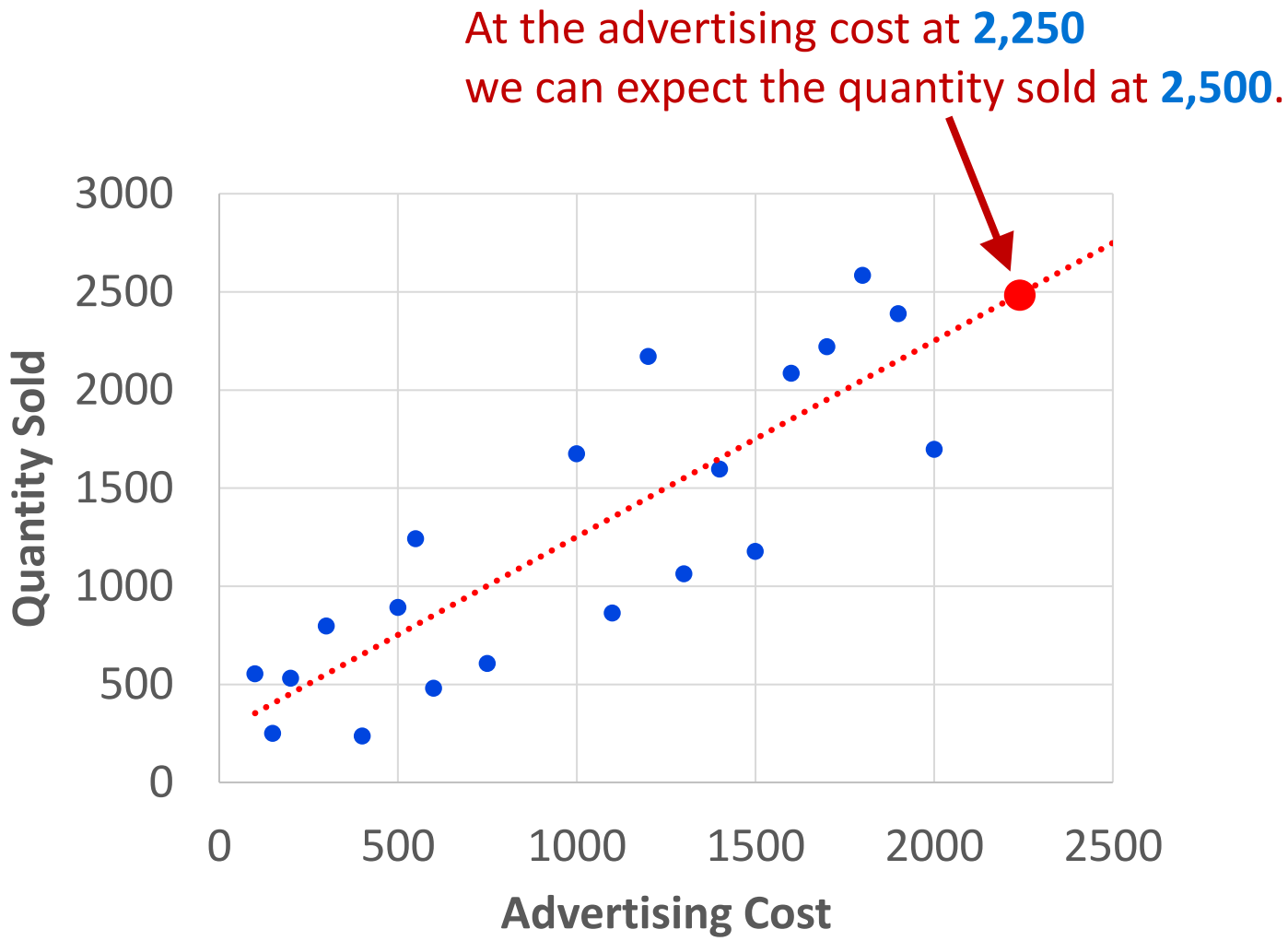
Advertising Cost	Quantity Sold
100	553
150	250
200	531
300	796
400	236
500	891
550	1241
600	480
750	606
1000	1674
1100	863
1200	2170
1300	1063
1400	1596
1500	1177
1600	2084
1700	2220
1800	2583
1900	2388
2000	1696



Advertising Cost	Quantity Sold
100	553
150	250
200	531
300	796
400	236
500	891
550	1241
600	480
750	606
1000	1674
1100	863
1200	2170
1300	1063
1400	1596
1500	1177
1600	2084
1700	2220
1800	2583
1900	2388
2000	1696



Advertising Cost	Quantity Sold
100	553
150	250
200	531
300	796
400	236
500	891
550	1241
600	480
750	606
1000	1674
1100	863
1200	2170
1300	1063
1400	1596
1500	1177
1600	2084
1700	2220
1800	2583
1900	2388
2000	1696

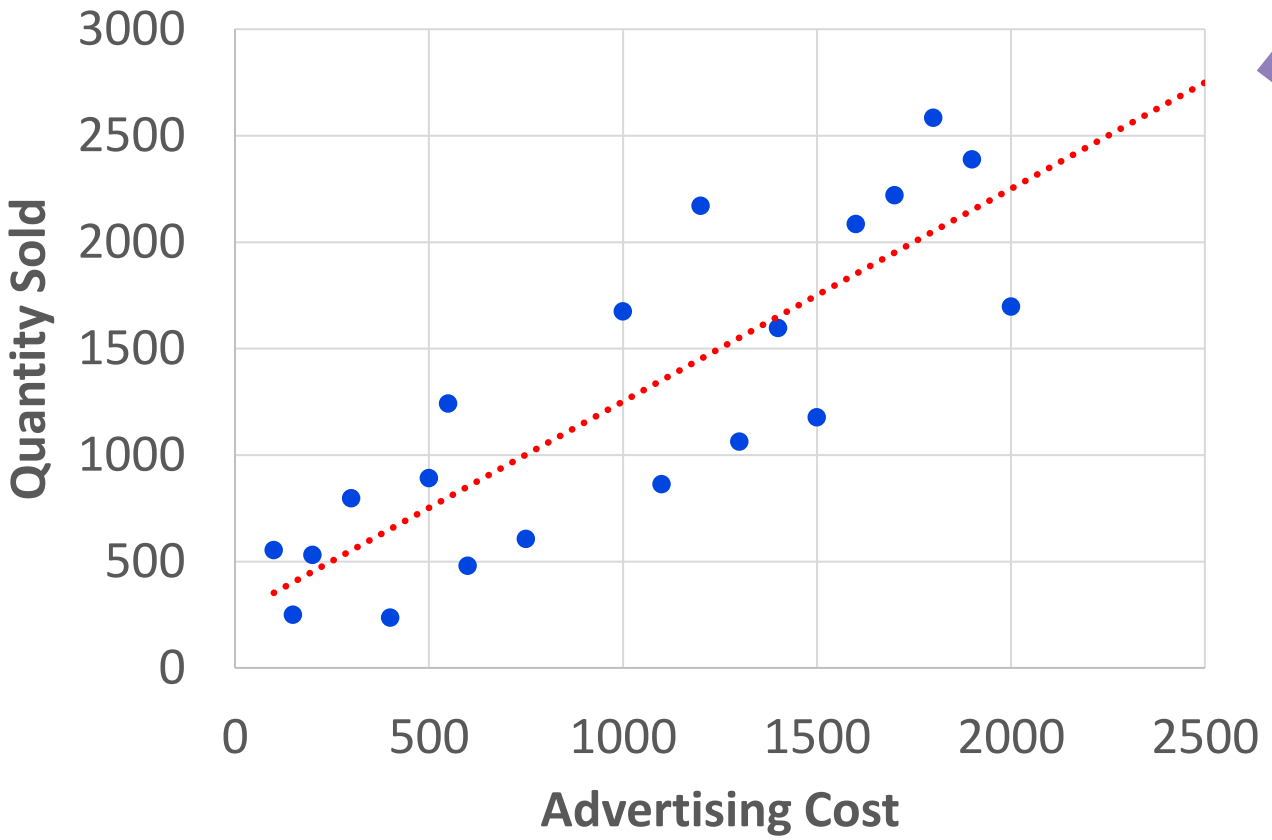


Advertising Cost	Quantity Sold
100	553
150	250
200	531
300	796
400	236
500	891
550	1241
600	480
750	606
1000	1674
1100	863
1200	2170
1300	1063
1400	1596
1500	1177
1600	2084
1700	2220
1800	2583
1900	2388
2000	1696

A common linear equation is $Y = a + bX$

Quantity sold

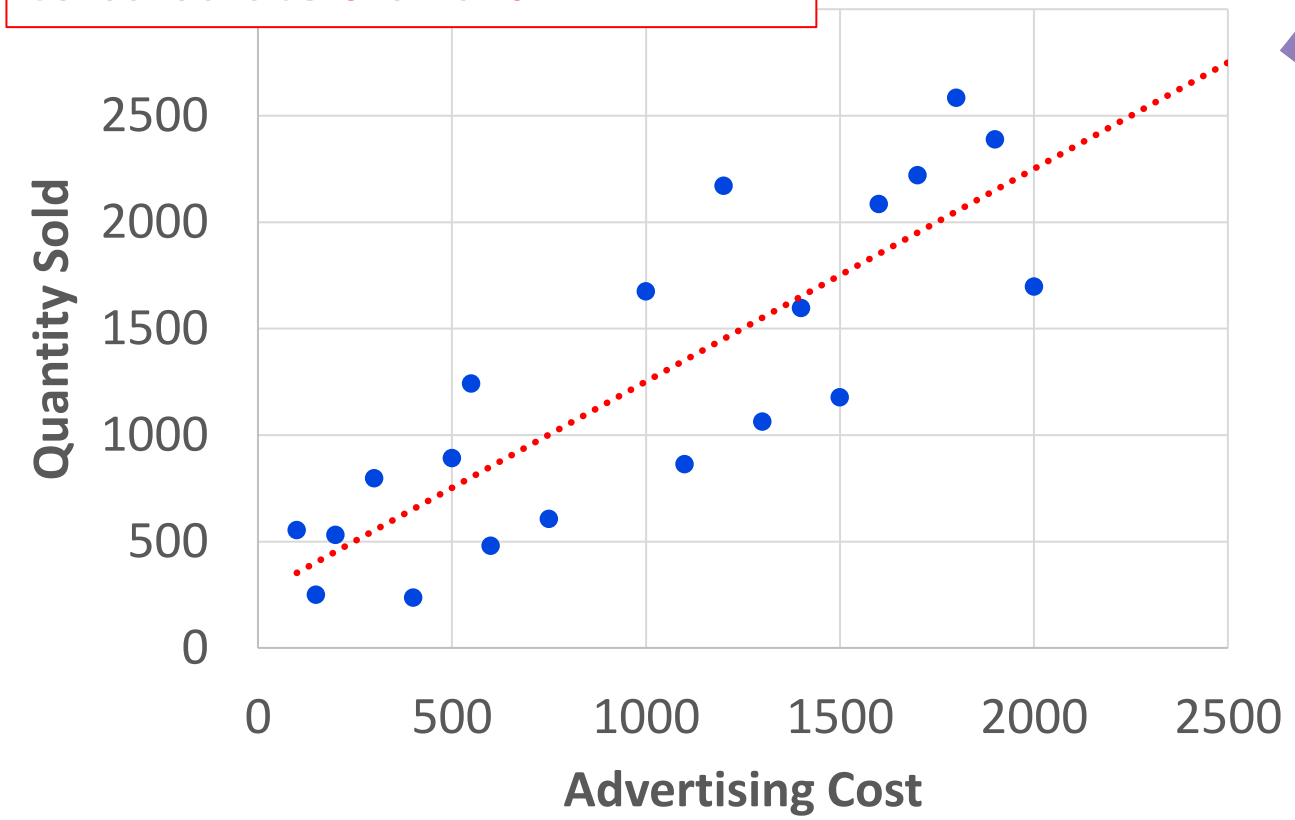
Ad cost



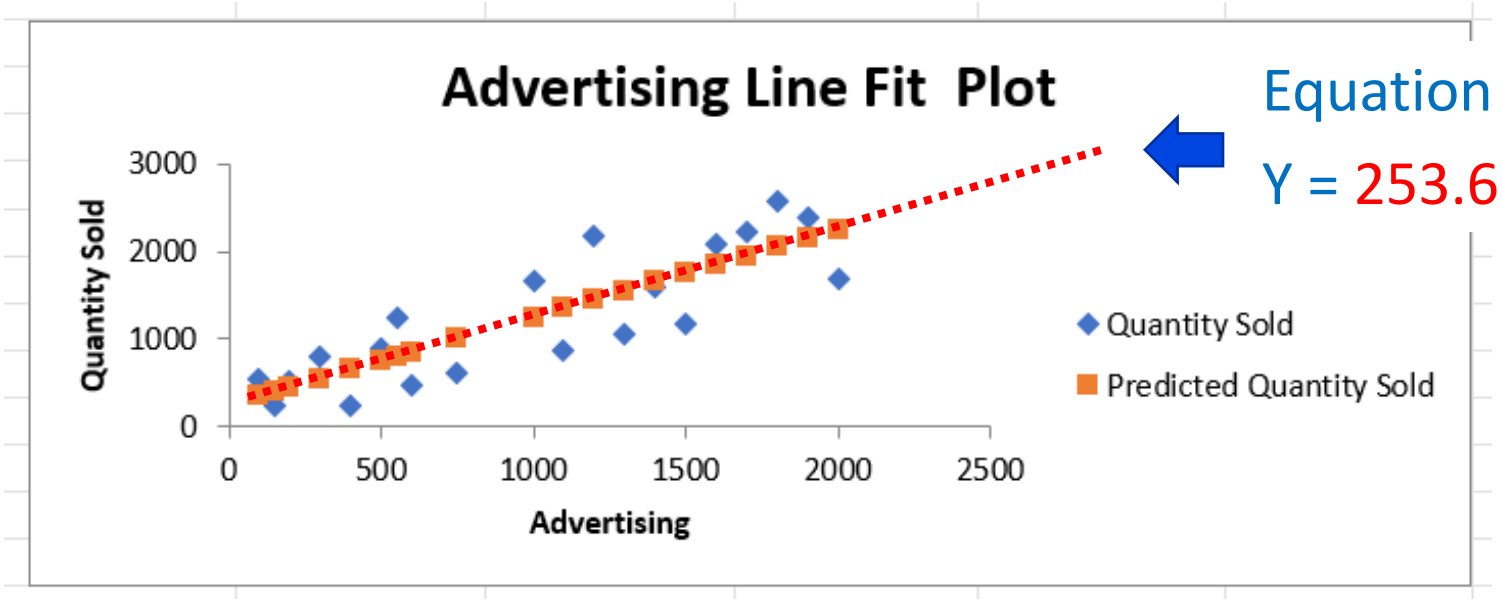
Advertising Cost	Quantity Sold
100	553
150	250
200	531
300	796
400	236
500	891
550	1241
600	480
750	606
1000	1674
1100	863
1200	2170
1300	1063
1400	1596
1500	1177
1600	2084
1700	2220
1800	2583
1900	2388
2000	1696

A common linear equation is $Y = a + bX$

The goal is to use this dataset to calculate **a** and **b**.



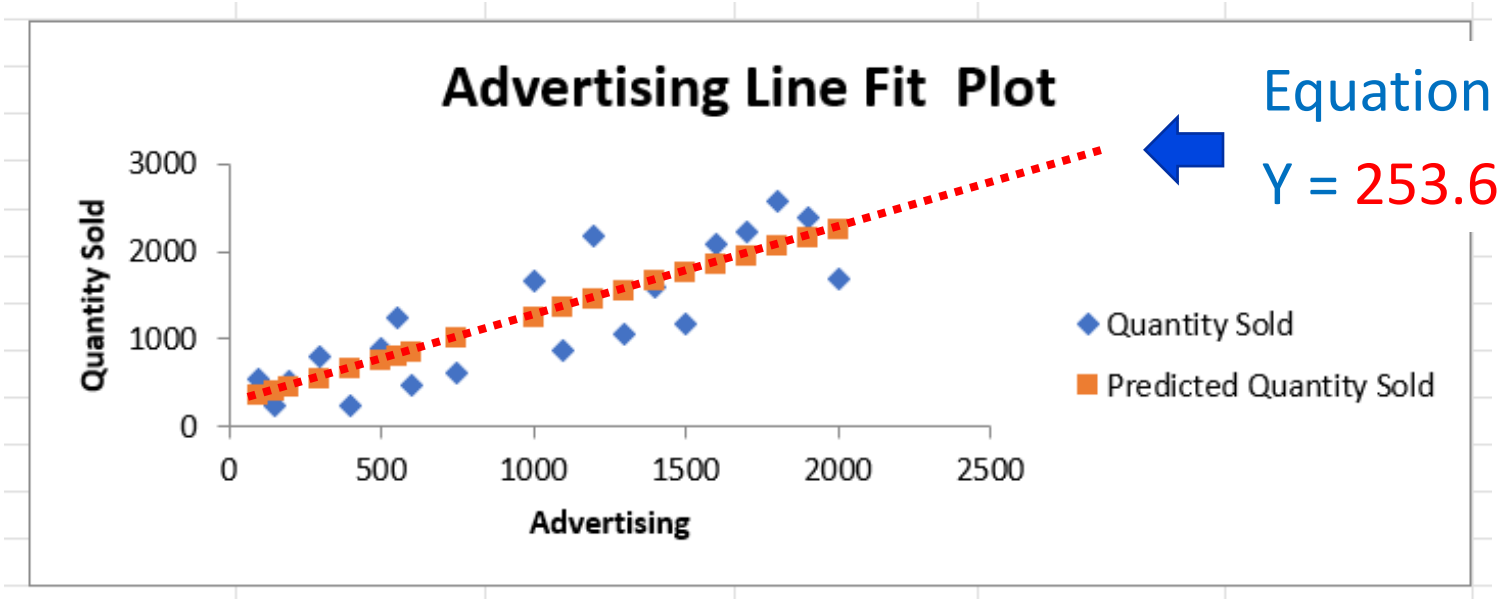
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	253.6313122	177.2517205	1.430910298	0.169588531
Advertising	0.998771758	0.150757013	6.625043433	3.21488E-06



Equation of this trend line:

Y = 253.6313 + 0.9988X

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	253.6313122	177.2517205	1.430910298	0.169588531
Advertising	0.998771758	0.150757013	6.625043433	3.21488E-06



Equation of this trend line:

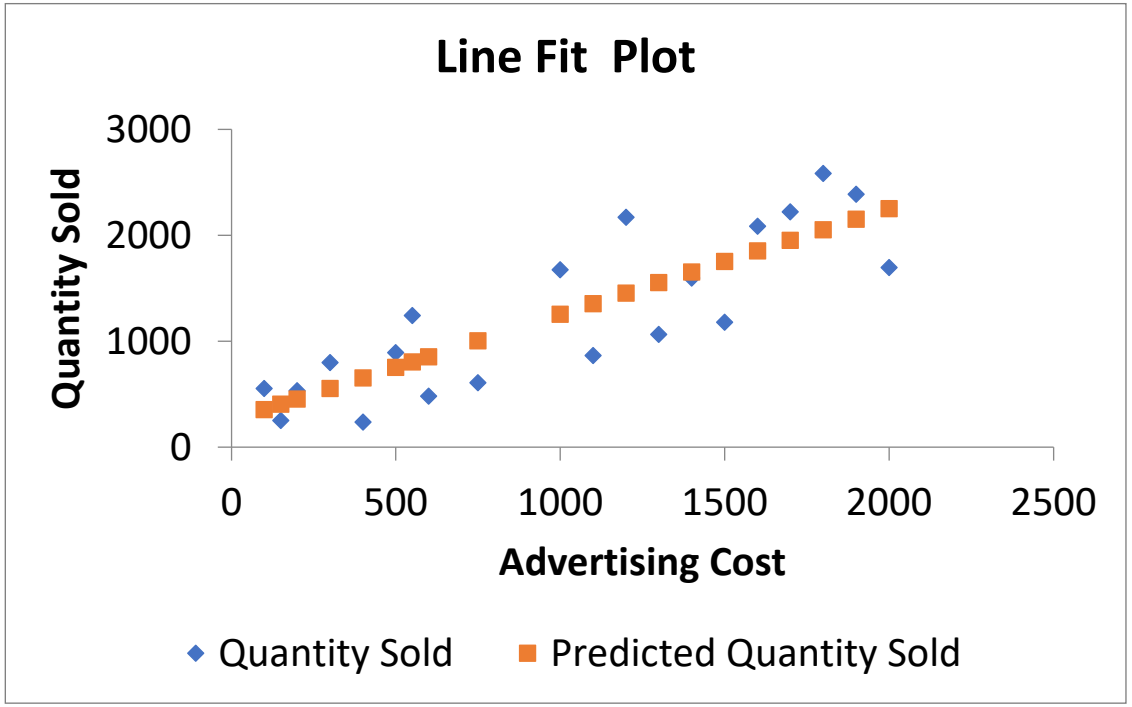
Y = 253.6313 + 0.9988X

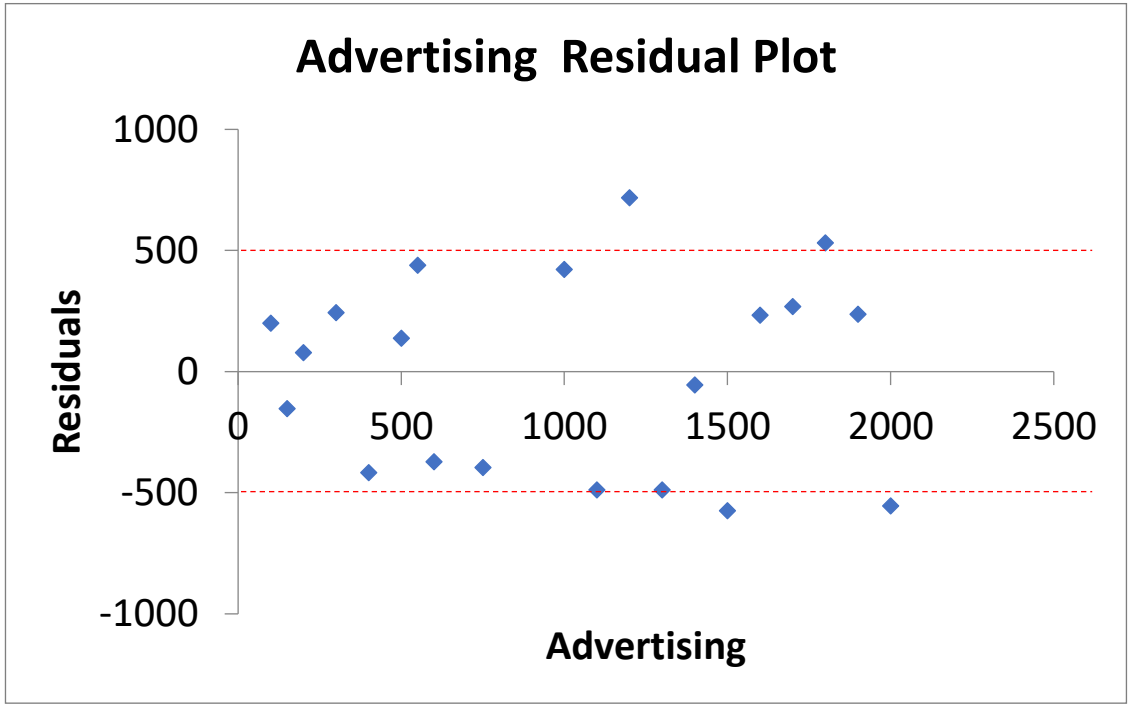
Try to predict quantity sold at ad cost equals 3,500.

Y = 253.6313 + 0.9988(3500)

Y = 3731.43

At ad cost equals 3,500. we could expect 3,731 items sold.





Using this equation, we can expect approx. 500 more or less than the predicted quantity sold.

Regression line equation

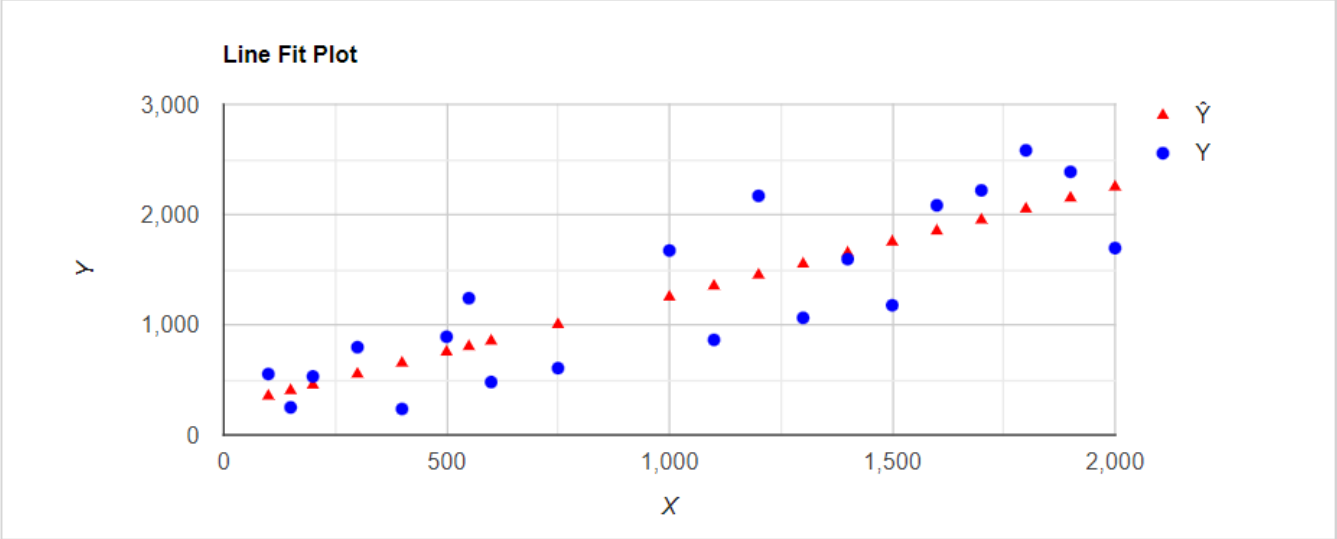
$$\hat{Y} = 253.6313 + 0.9988X$$

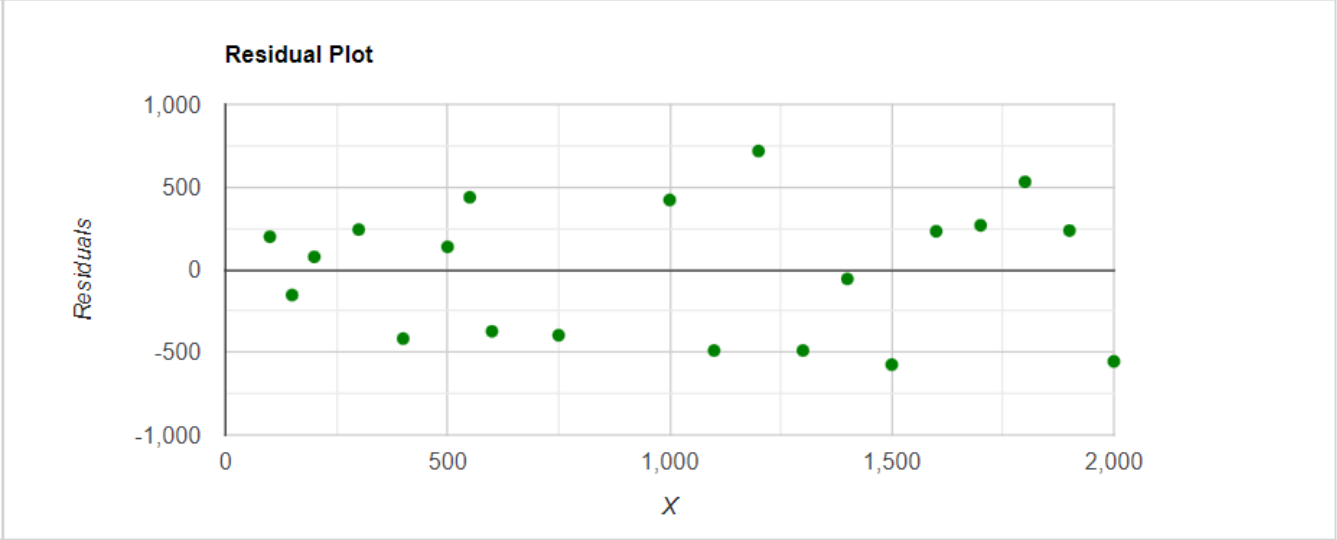
Reporting linear regression in APA style

X predicted Y, R^2 = .71, $F(1,18)$ = 43.89, p < .001.
β = , p < .001.

Results from a web application
(statskingdom.com)

Regression line





Excel output

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.842120571
R Square	0.709167057
Adjusted R Square	0.693009671
Standard Error	414.1669963
Observations	20

Web app output

X predicted Y, R^2 = .71, $F(1,18)$ = 43.89, p < .001.

β = , p < .001.

Excel output

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.842120571
R Square	0.709167057
Adjusted R Square	0.693009671
Standard Error	414.1669963
Observations	20

0.71 or **71%** of the output can be predicted by the input.

Advertising Cost	Quantity Sold
100	553
150	250
200	531
300	796
400	236
500	891
550	1241
600	480
750	606
1000	1674
1100	863
1200	2170
1300	1063
1400	1596
1500	1177
1600	2084
1700	2220
1800	2583
1900	2388
2000	1696

Web app output

X predicted Y, $R^2 = .71$, $F(1,18) = 43.89$, $p < .001$.

$\beta =$, $p < .001$.

Excel output

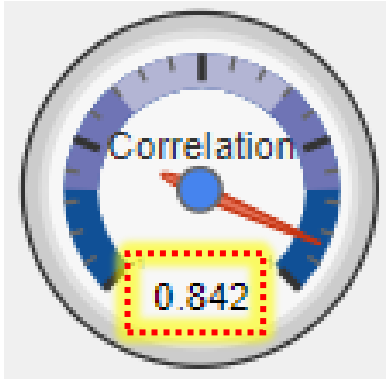
SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.842120571
R Square	0.709167057
Adjusted R Square	0.693009671
Standard Error	414.1669963
Observations	20

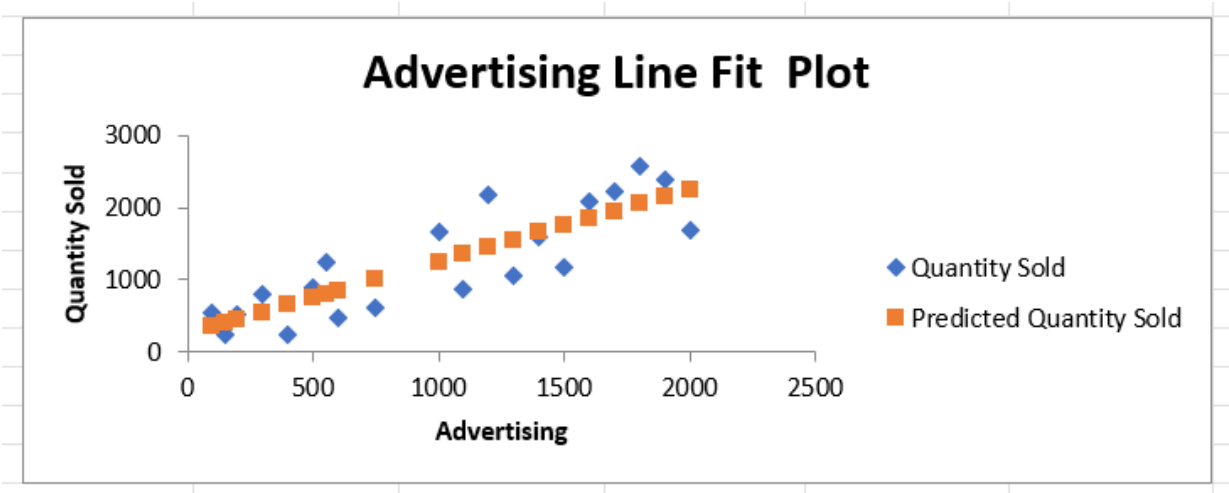
Multiple R approaches 1.0 means that input and output are highly correlated.

When we increase the ad cost, sales tend to increase as well, and vice versa.

Advertising Cost	Quantity Sold
100	553
150	250
200	531
300	796
400	236
500	891
550	1241
600	480
750	606
1000	1674
1100	863
1200	2170
1300	1063
1400	1596
1500	1177
1600	2084
1700	2220
1800	2583
1900	2388
2000	1696

Web app output





Excel output

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	253.6313122	177.2517205	1.430910298	0.169588531
Advertising	0.998771758	0.150757013	6.625043433	3.21488E-06

3.21×10^{-6}
$= 0.00000321$

Web app output

Source	DF	Sum of Square	Mean Square	F Statistic (df_1,df_2)	P-value
Regression (between $\hat{y}_i$ and $\bar{y}$)	1	7528846.386	7528846.386	43.8912 (1,18)	0.000003215
Residual (between y_i and $\hat{y}_i$)	18	3087617.414	171534.3008		
Total (between y_i and $\bar{y}$)	19	10616463.8	558761.2526		

Small P-value (P < 0.05):

Using ad cost as a factor in predicting sales is more accurate than without applying this variable.

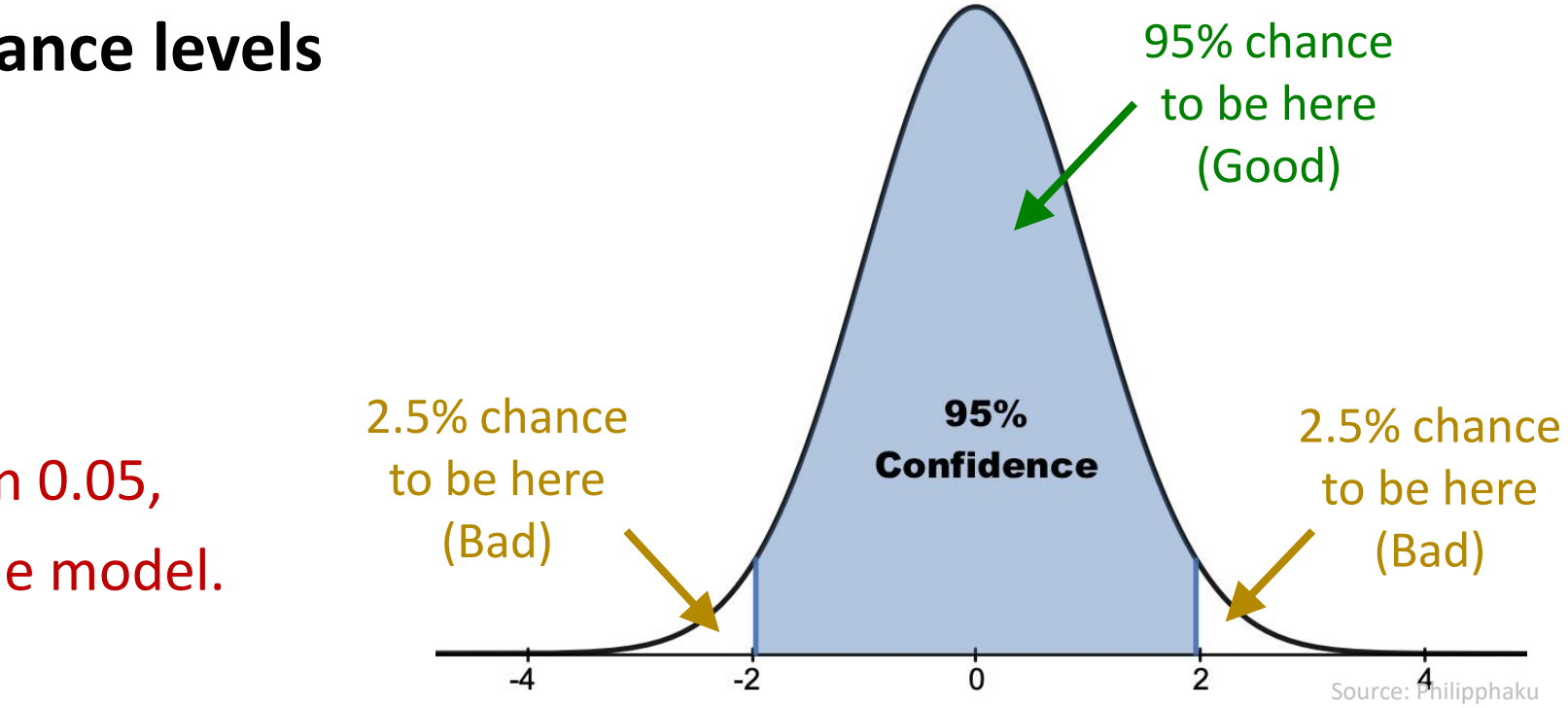
Three different significance levels

P < 0.001 (Best fit)

P < 0.01 (Good fit)

P < 0.05 (Acceptable fit)

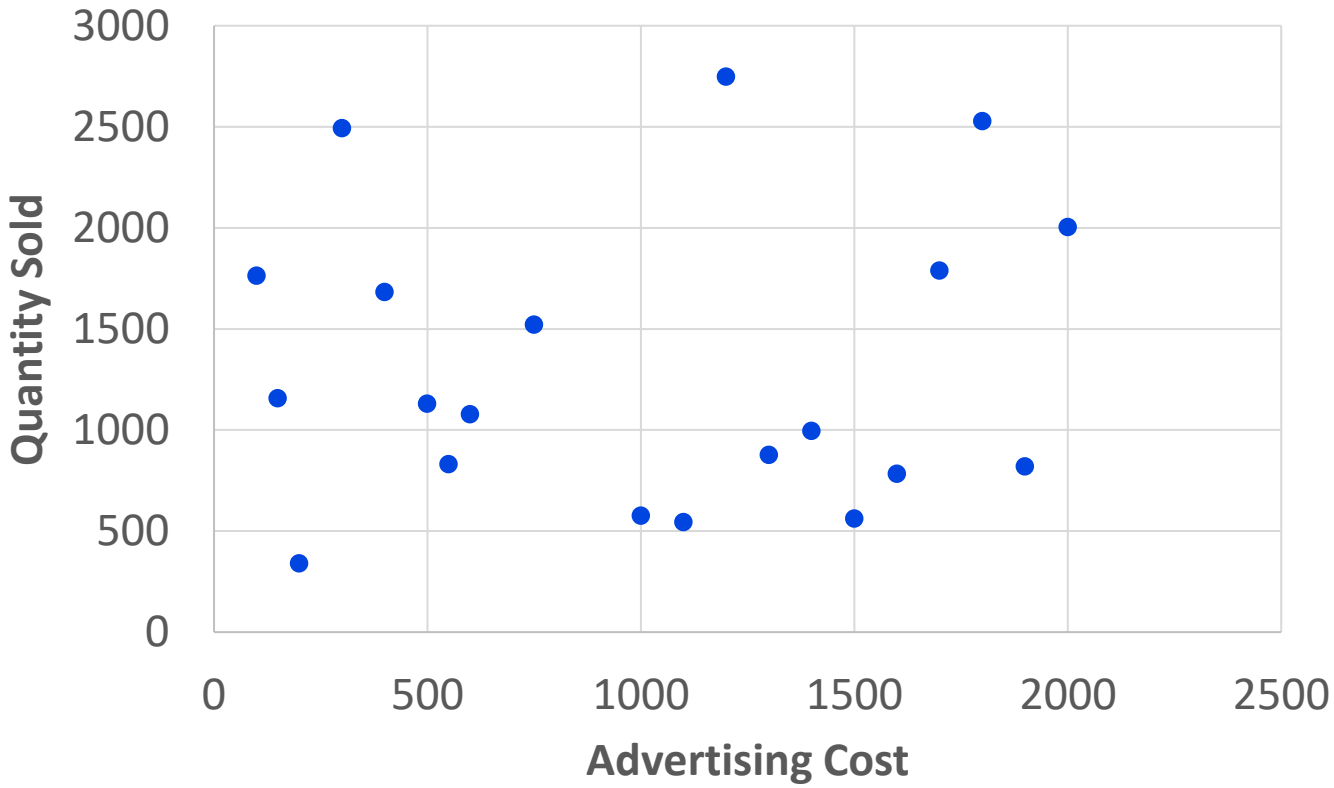
If the P value is greater than 0.05, that variable does not fit the model.



P < 0.05 means there is less than 5% chance that the prediction will be out of the confidence zone.

Example 2

Advertising Cost	Quantity Sold
100	1763
150	1156
200	339
300	2493
400	1682
500	1129
550	830
600	1077
750	1520
1000	576
1100	544
1200	2748
1300	876
1400	995
1500	561
1600	783
1700	1789
1800	2528
1900	819
2000	2003



Regression line equation

$$\hat{Y} = 1228.0943 + 0.08225X$$

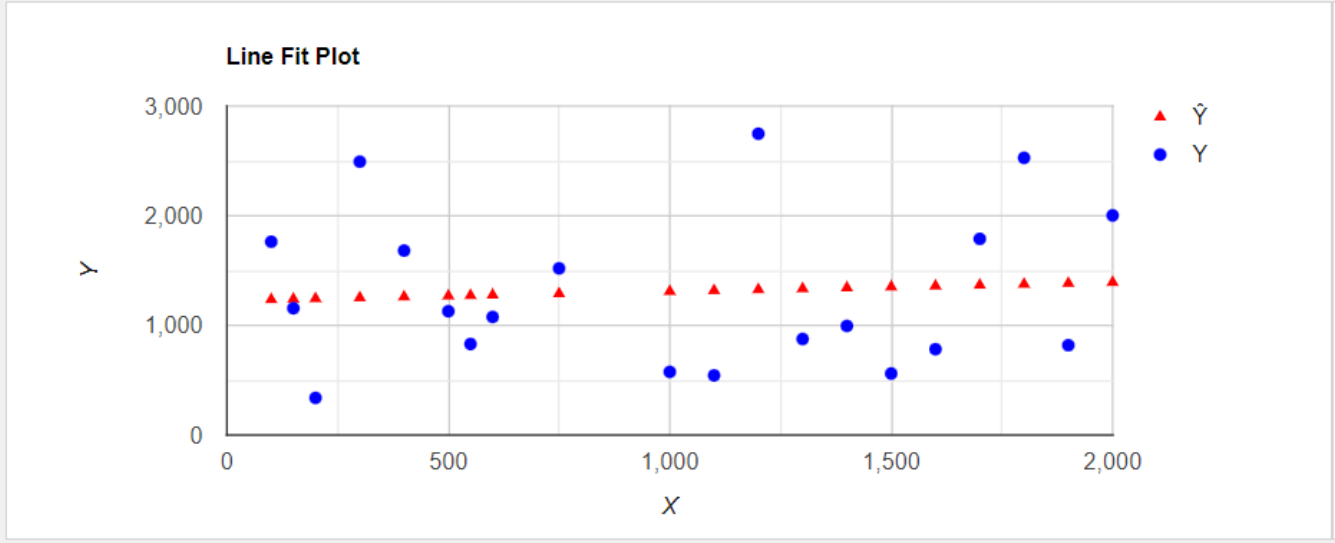
Reporting linear regression in APA style

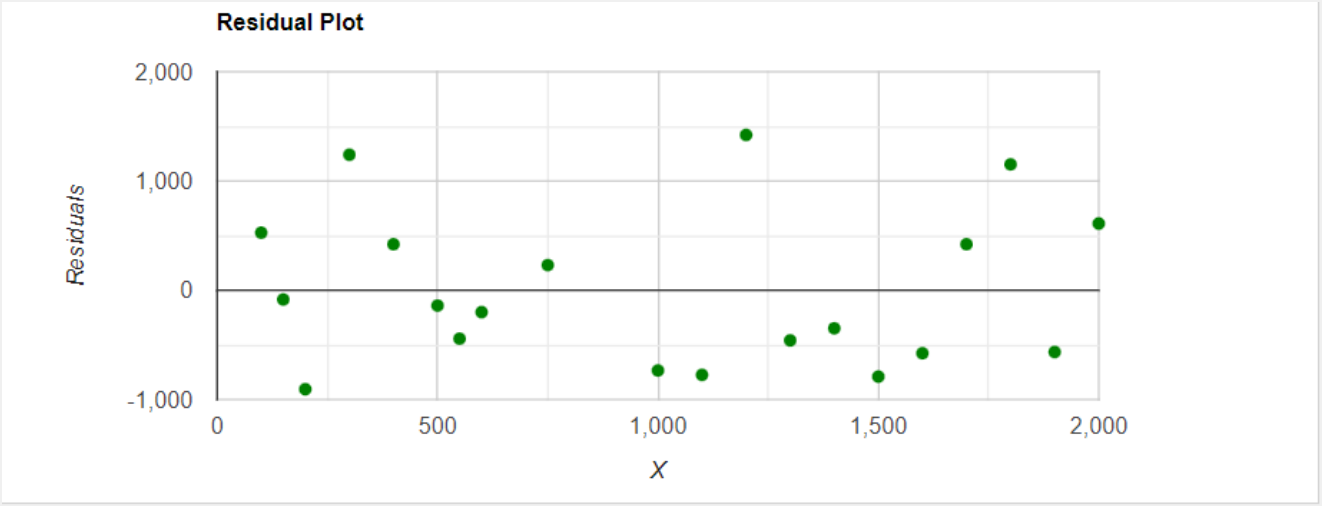
$R^2 = .0052$, $F(1,18) = 0.094$, $p = .763$.

$\beta = .082$, $p = .763$.

Results from a web application (statskingdom.com)

Regression line





0.0052 or **0.5%** of the output can be predicted by the input.

** Should be 0.7 (70%)or above to be good.*

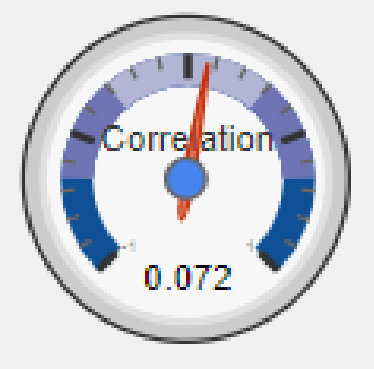
Regression line equation

$$\hat{Y} = 1228.0943 + 0.08225X$$

Reporting linear regression in APA style

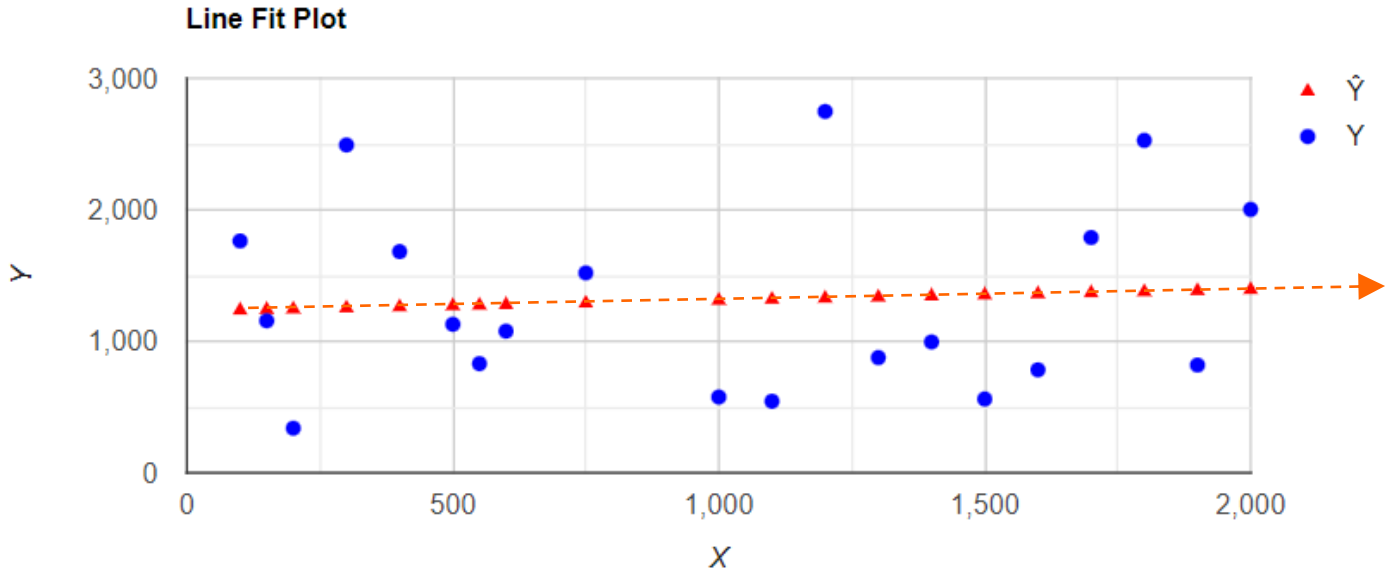
$R^2 = .0052$ $F(1,18) = 0.094$, $p = .763$.

$\beta = .082$, $p = .763$.



Too low correlation (Multiple R) indicates weak relationship between input and output.

- *+0.8 or above for positive correlation.*
- *-0.8 or below for negative correlation.*



Increasing the ad cost, **might not** increase sales.

Regression line equation

$$\hat{Y} = 1228.0943 + 0.08225X$$

Reporting linear regression in APA style

$R^2 = .0052$, $F(1,18) = 0.094$ $p = .763$.

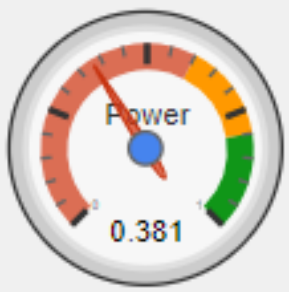
$\beta = .082$, $p = .763$.

P > 0.05 means that the variable (ad cost) **does not** fit the model.

Source	DF	Sum of Square	Mean Square	F Statistic (df_1,df_2)	P-value
Regression (between $\hat{y}_i$ and $\bar{y}$)	1	51058.6127	51058.6127	0.09403 (1,18)	0.7626
Residual (between y_i and $\hat{y}_i$)	18	9774086.337	543004.7965		
Total (between y_i and $\bar{y}$)	19	9825144.95	517112.8921		

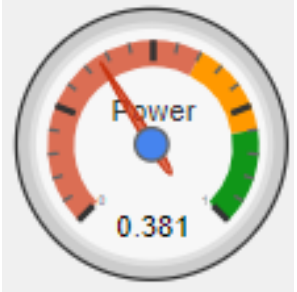
Example 1

Advertising Cost	Quantity Sold
100	553
150	250
200	531
300	796
400	236
500	891
550	1241
600	480
750	606
1000	1674
1100	863
1200	2170
1300	1063
1400	1596
1500	1177
1600	2084
1700	2220
1800	2583
1900	2388
2000	1696



Example 2

Advertising Cost	Quantity Sold
100	1763
150	1156
200	339
300	2493
400	1682
500	1129
550	830
600	1077
750	1520
1000	576
1100	544
1200	2748
1300	876
1400	995
1500	561
1600	783
1700	1789
1800	2528
1900	819
2000	2003

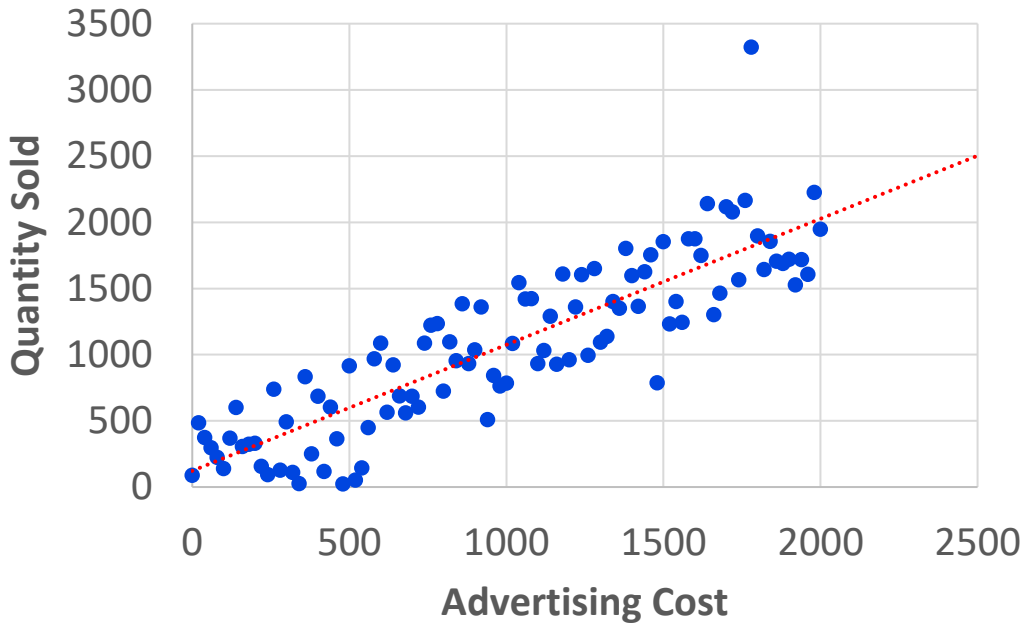


Low predictive power due to a small sample size (20 records).

** Should be 0.8 or above to be good.*

Example 3

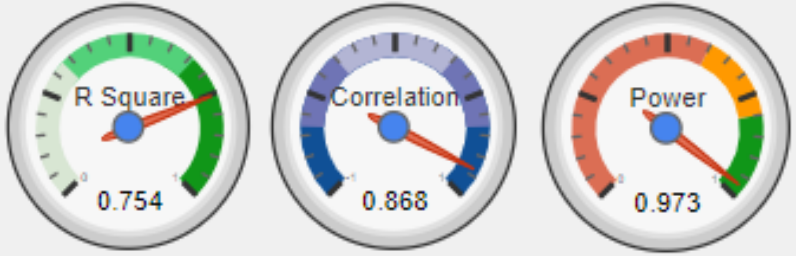
	A	B
1	Advertising Cost	Quantity Sold
2	0	90
3	20	486
4	40	374
5	60	297
6	80	224
7	100	140
8	120	369
9	140	602
10	160	307
11	180	325
12	200	331
13	220	158
14	240	95
15	260	740
16	280	127
17	300	494
18	320	112
19	340	26
20	360	833
21	380	251
22	400	687
23	420	118
24	440	605
25	460	366
26	480	24
27	500	917
28	520	54
29	540	146
100	1960	1608
101	1980	2226
102	2000	1948



Regression line equation

$$\hat{Y} = 122.0978 + 0.9521X$$

X predicted Y, $R^2 = .75$, $F(1,99) = 302.94$, $p < .001$.
$\beta = .95$, $p < .001$.

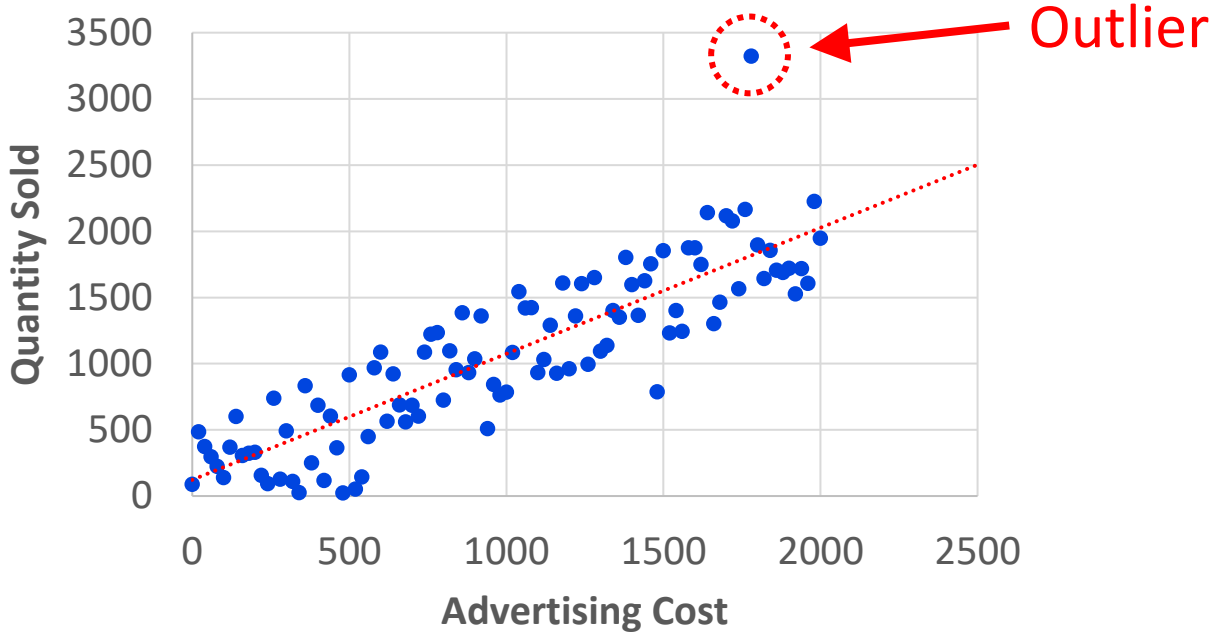


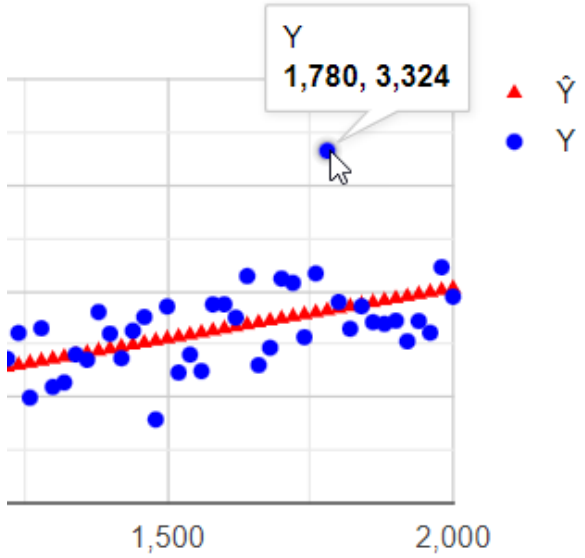
Good fit and high predictive power.

Running experiments too many times can waste a lot of resource.

	A	B
1	Advertising Cost	Quantity Sold
2	0	90
3	20	486
4	40	374
5	60	297
6	80	224
7	100	140
8	120	369
9	140	602
10	160	307
11	180	325
12	200	331
13	220	158
14	240	95
15	260	740
16	280	127
17	300	494

89	1740	1566
90	1760	2167
91	1780	3324
92	1800	1897
93	1820	1645
94	1840	1857
95	1860	1708
96	1880	1691
97	1900	1721
98	1920	1527
99	1940	1718
100	1960	1608
101	1980	2226
102	2000	1948

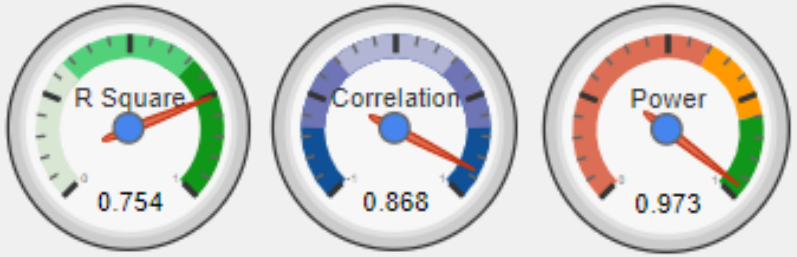




Before removing an outlier (N = 101)

Regression line equation

$$\hat{Y} = 122.0978 + 0.9521X$$



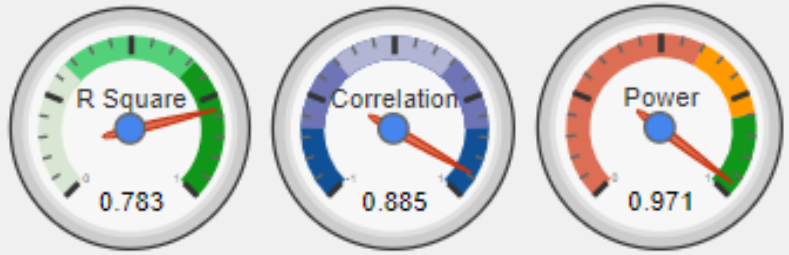
X predicted Y, $R^2 = .75$, $F(1,99) = 302.94$, $p < .001$.
$\beta = .95$, $p < .001$.

Good fit and high predictive power.

After removing an outlier (N = 100)

Regression line equation

$$\hat{Y} = 141.9575 + 0.9169X$$



X predicted Y, $R^2 = .78$, $F(1,98) = 354.55$, $p < .001$.
$\beta = .92$, $p < .001$.

Slightly increase the correlation
but slightly decrease the power.