DSNP DJPb Kementerian Keuangan RI

Regression

Instructor: Muhammad Hilman, Ph.D

Slide by Fariz Darari, Ph.D.

Case Study: Restaurant Tipping

Let's assume that you are a **small restaurant owner** at a nice restaurant.

In the US, **tips** are a very important part of a waiter's pay. Most of the time **the dollar amount of the tip is related to the dollar amount of the total bill.**

Can you identify what can be a **data science problem** we can tackle here?

Case Study: Restaurant Tipping

Let's assume that you are a **small restaurant owner** at a nice restaurant.

In the US, **tips** are a very important part of a waiter's pay. Most of the time **the dollar amount of the tip is related to the dollar amount of the total bill.**

Can you identify what can be a **data science problem** we can tackle here?

As an owner who happens to be a data science geek, you would like to develop a model allowing you to make a prediction about:

Case Study: Restaurant Tipping

Let's assume that you are a **small restaurant owner** at a nice restaurant.

In the US, **tips** are a very important part of a waiter's pay. Most of the time **the dollar amount of the tip is related to the dollar amount of the total bill.**

Can you identify what can be a **data science problem** we can tackle here?

As an owner who happens to be a data science geek, you would like to develop a model allowing you to make a prediction about:

What amount of tip to expect for any given bill amount?

Case Study: Restaurant Tipping

Let's assume that you are a **small restaurant owner** at a nice restaurant.

In the US, **tips** are a very important part of a waiter's pay. Most of the tim

the dollar amount of the tip is related to the dollar am

Can you think of similar cases based on this tipping problem?

Can you identify wh ... e can tackle here?

ppens to be a data science geek, you would like to develop

a odel allowing you to make a prediction about:

What amount of tip to expect for any given bill amount?

Case Study: Student Grading

You are a lecturer at the best university in Indonesia. Grading is one important factor in student evaluation, capturing how students progress throughout your course. Student grading components can be related to each other.

Can you identify what can be a **data science problem** we can tackle here?

Case Study: Student Grading

You are a lecturer at the best university in Indonesia. Grading is one important factor in student evaluation, capturing how students progress throughout your course. Student grading components can be related to each other.

Can you identify what can be a **data science problem** we can tackle here?

As a lecturer who happens to be a data science geek, you would like to develop a model allowing you to make a prediction about?

Case Study: Household Food Expenditures

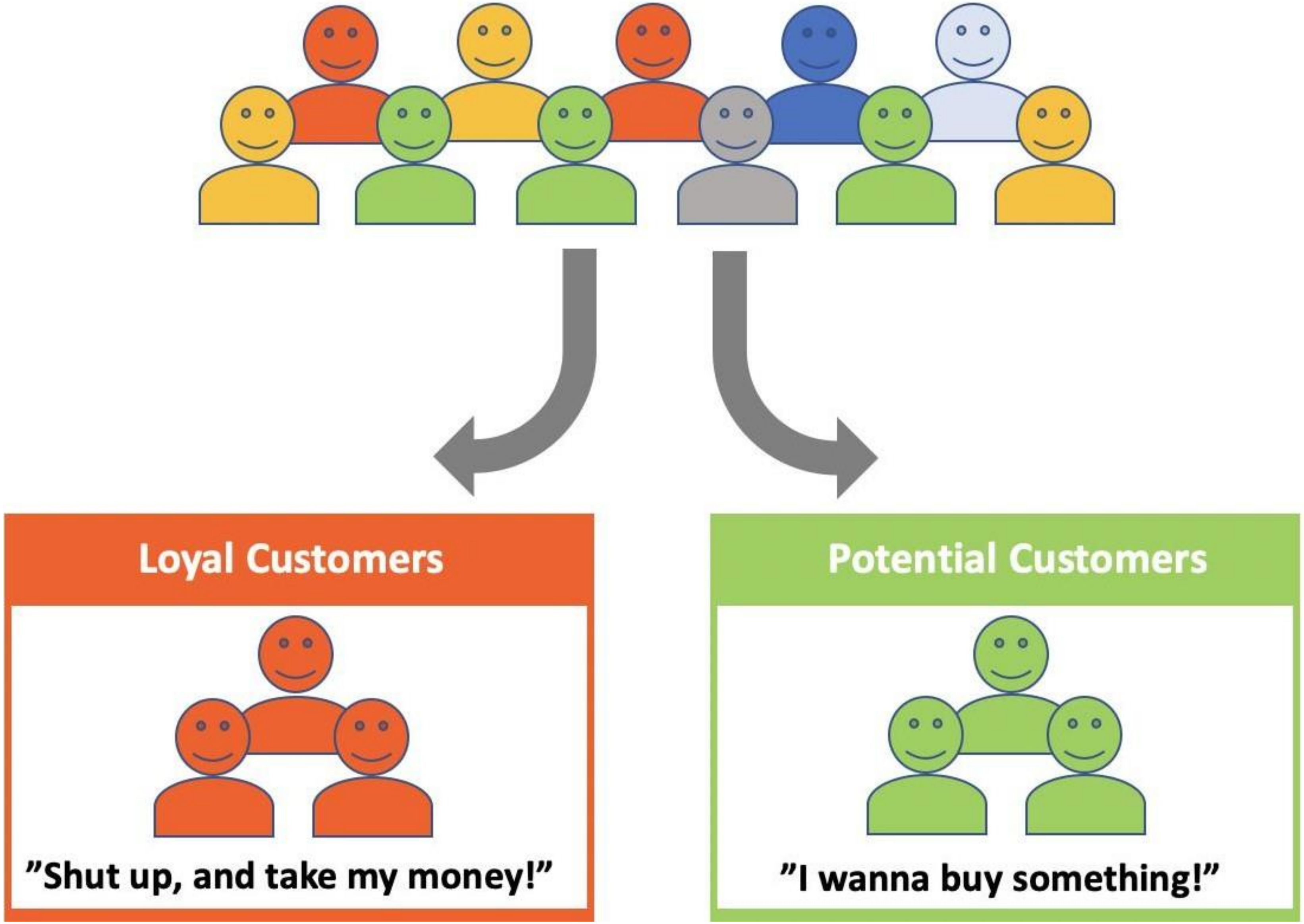
You have a family that you must support. As a good dad/mom, you would like to analyze food expenditures of your family.

What do you think would be the most important feature (variable) relating to food expenditure?

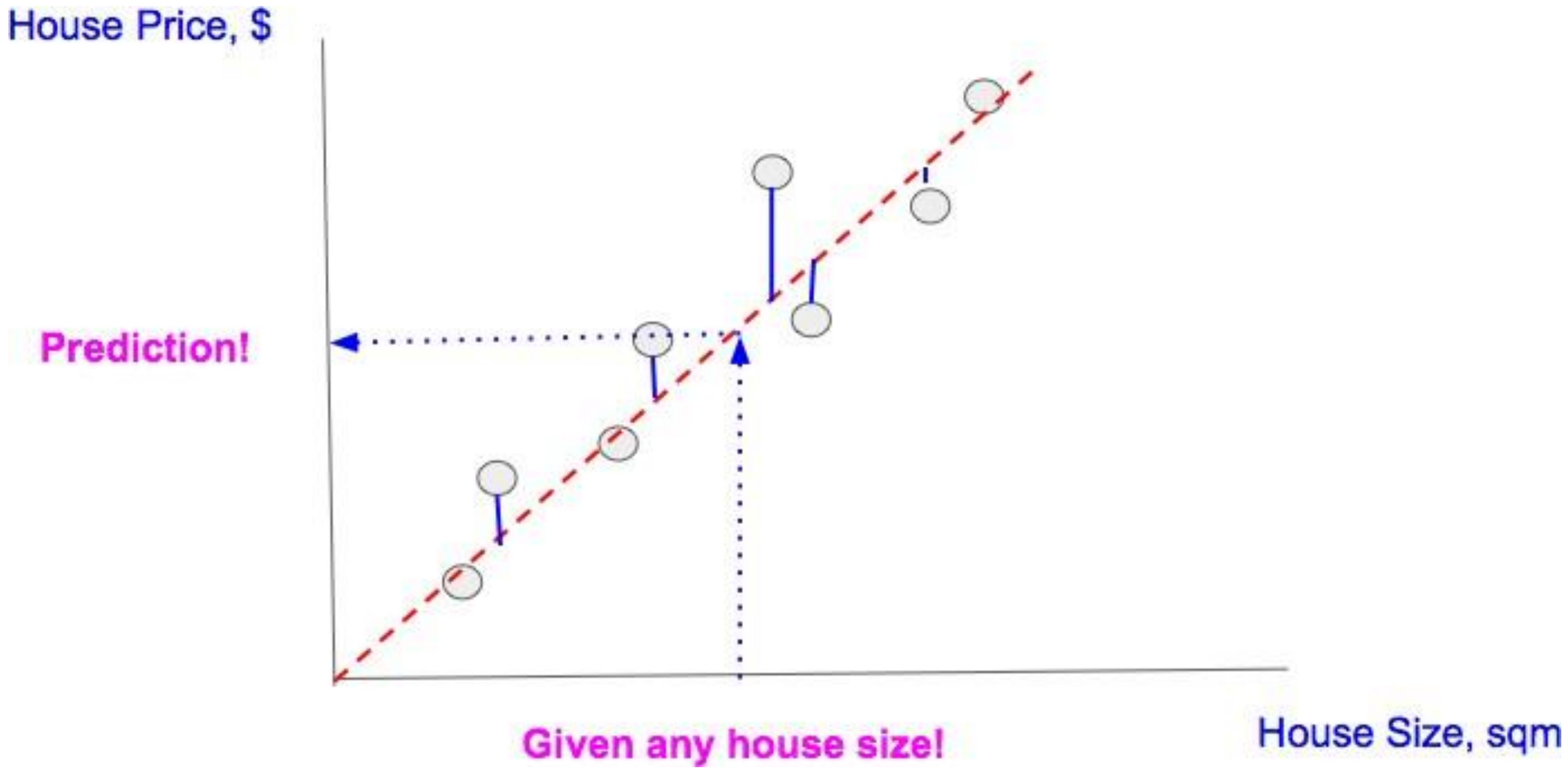
What can data science help here?


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

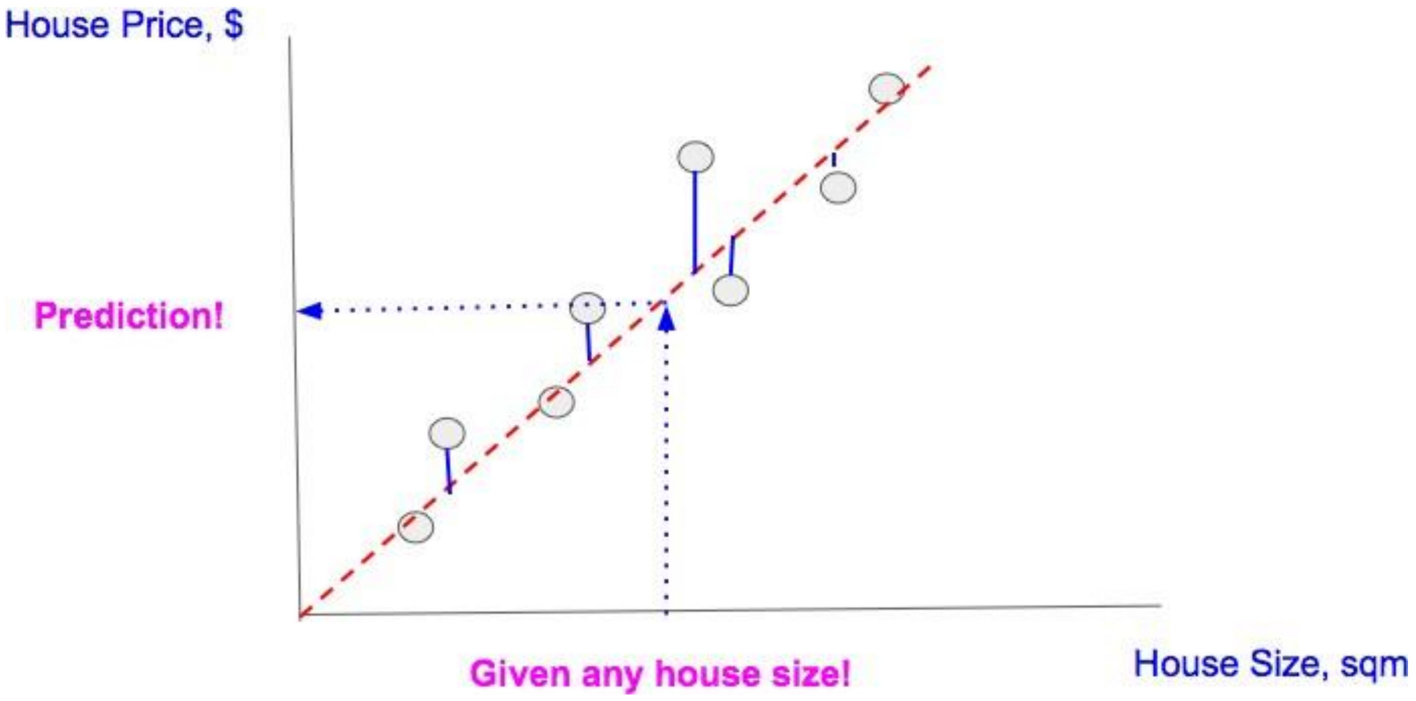
?? Learning	?? Learning
Data: x	Data: (x, y) where y is the label
Goal: Learn underlying structure/pattern	Goal: Learn function to map $X \to Y$
Example: Customer segmentation	Example: Price prediction



We have learned about clustering


House Price, $
Prediction!
Given any house size!
House Size, sqm

Simple Linear Regression



A simple linear regression model gives a straight-line relationship between two variables.

Case Study: Ice Cream Shop

The local ice cream shop keeps track of:

how much ice cream they sell

versus

the noon temperature on that day.

The figure on the right shows the records for the last 12 days.

Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

Case Study: Ice Cream Shop

The local ice cream shop keeps track of:

how much ice cream they sell

versus

the noon temperature on that day.

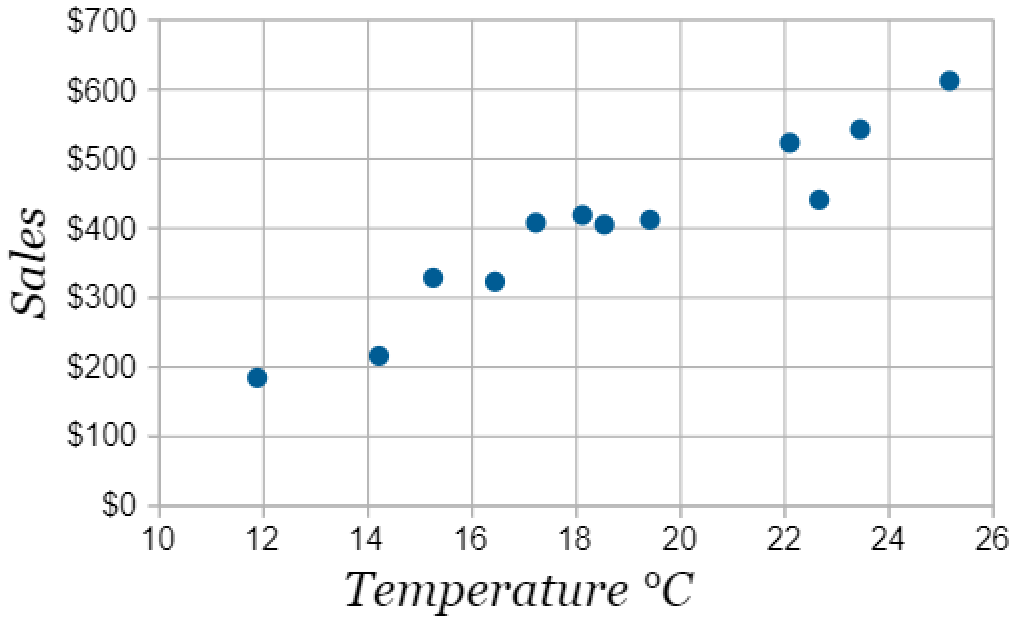
The figure on the right shows the records for the last 12 days.

What's next?

Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

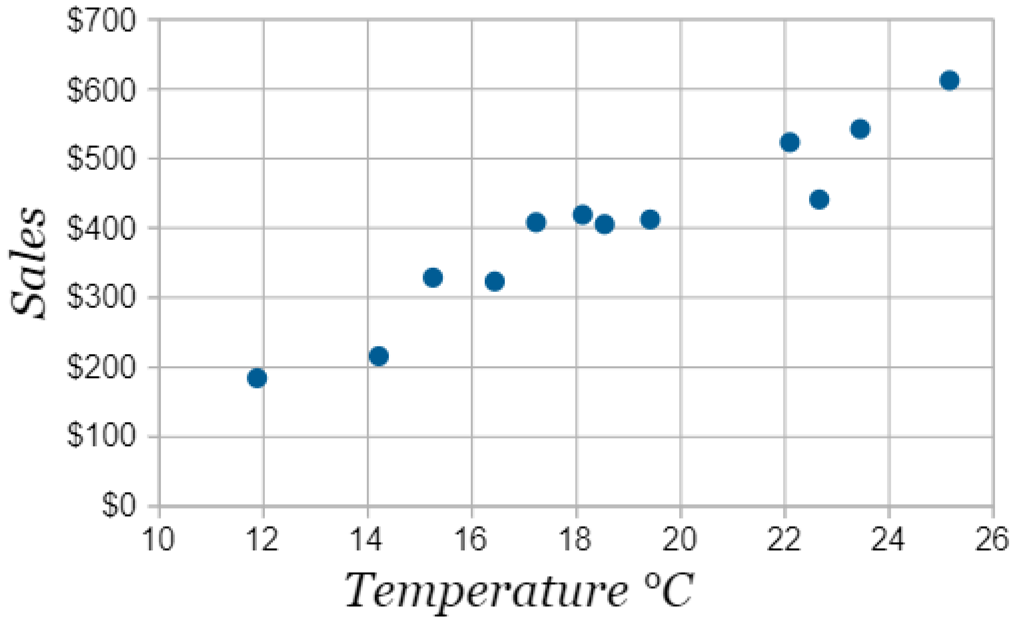
Case Study: Ice Cream Shop

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408



Case Study: Ice Cream Shop

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408



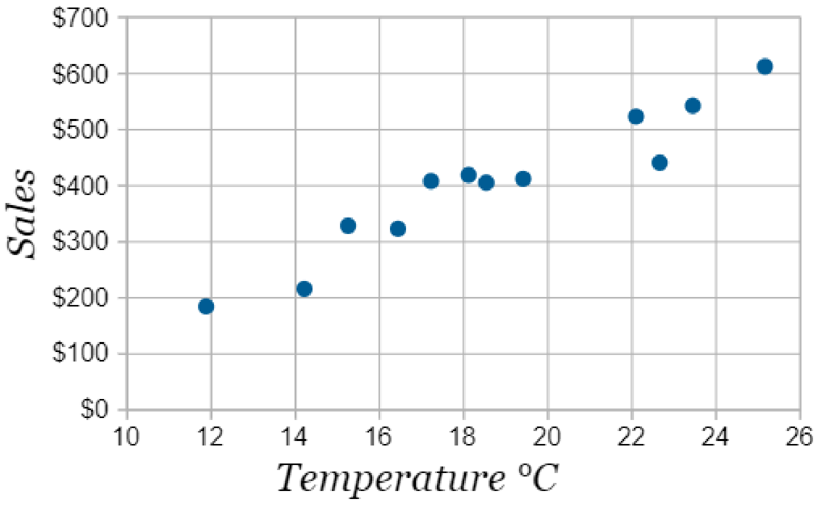
From our EDA using a scatterplot,

we can see that: **Warmer weather leads to more sales!**

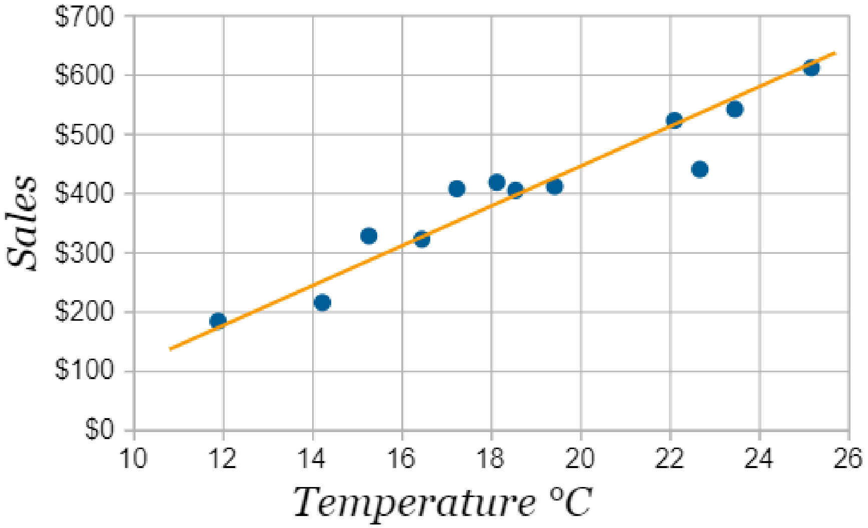
Case Study: Ice Cream Shop

Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

Raw Data



EDA (Scatterplot)

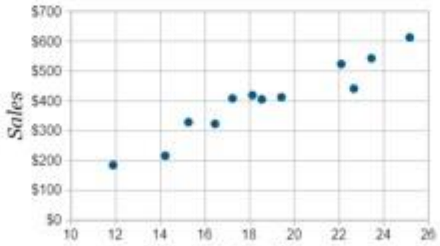


Regression Model

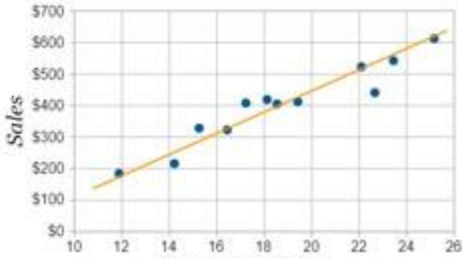
Case Study: Ice Cream Shop

Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

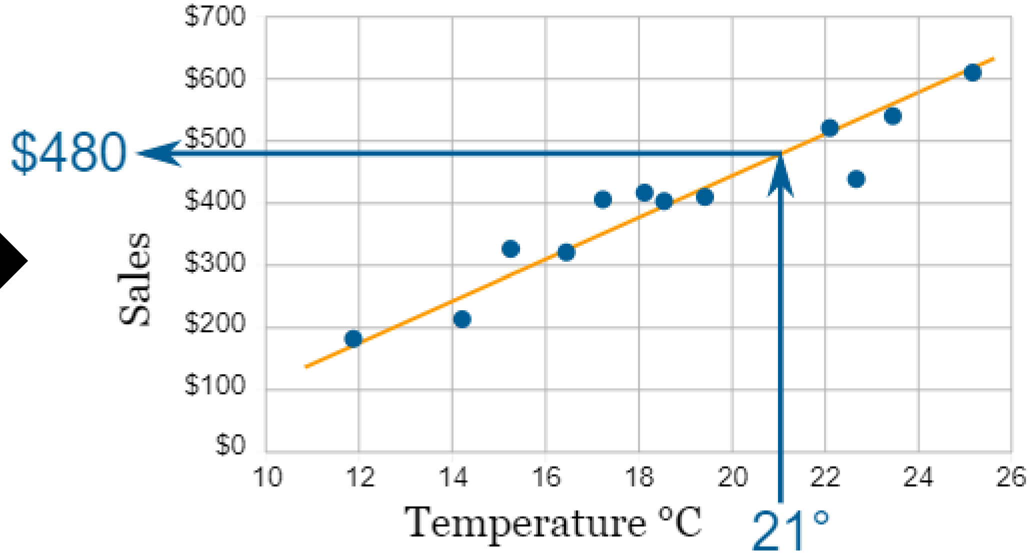
Raw Data



EDA (Scatterplot)



Regression Model

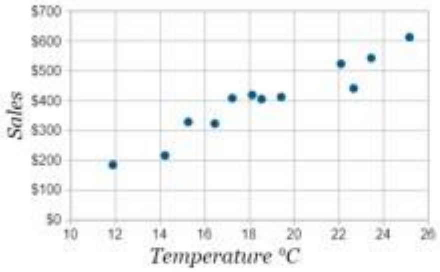


Prediction Using Regression Model

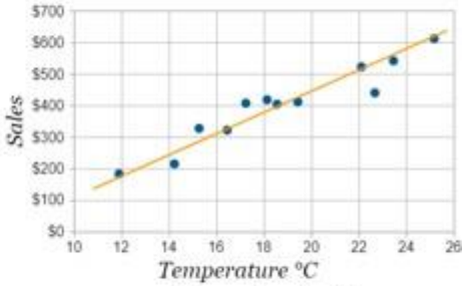
Case Study: Ice Cream Shop

Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

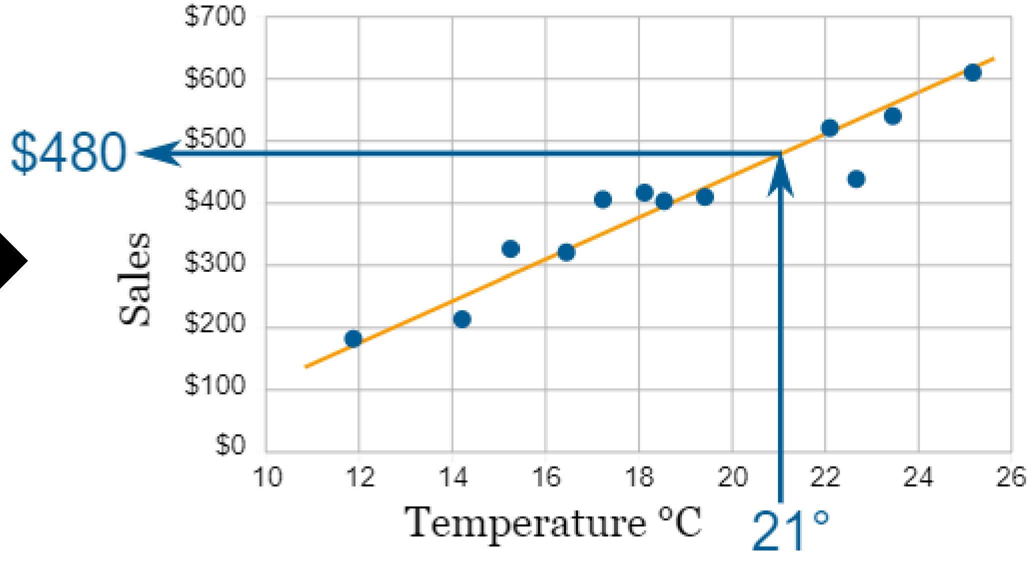
Raw Data



EDA (Scatterplot)



Regression Model



Prediction Using Regression Model

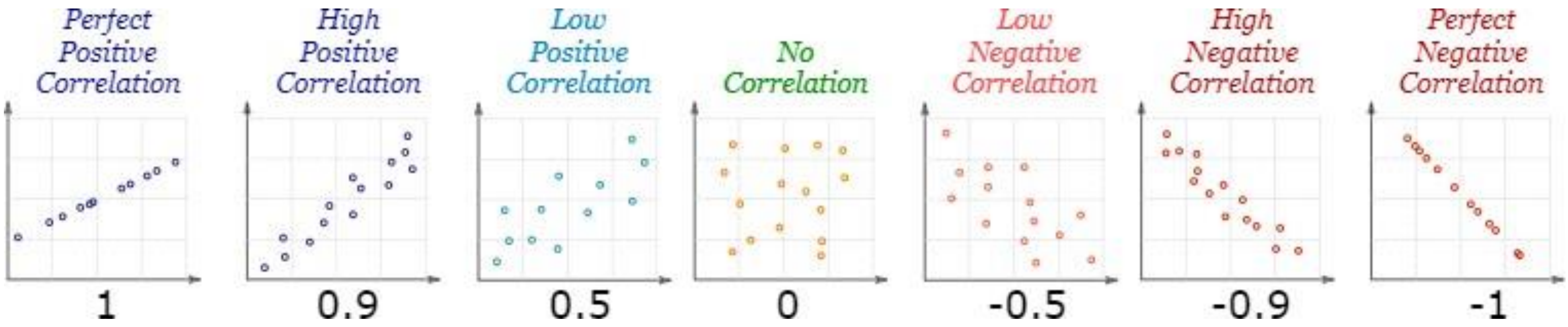
This is a **common scenario** when we perform regression analysis!

Correlation and regression

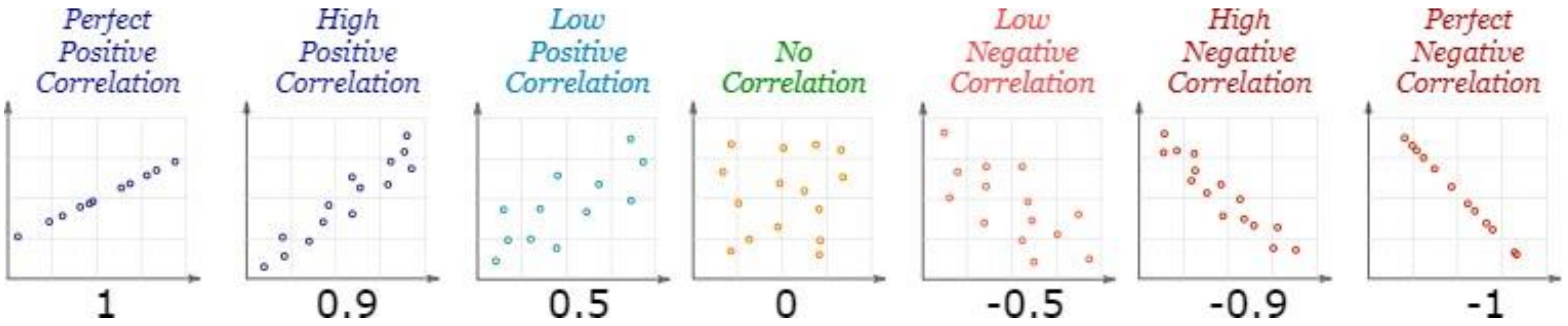
We use **correlation** to denote association between two quantitative variables.

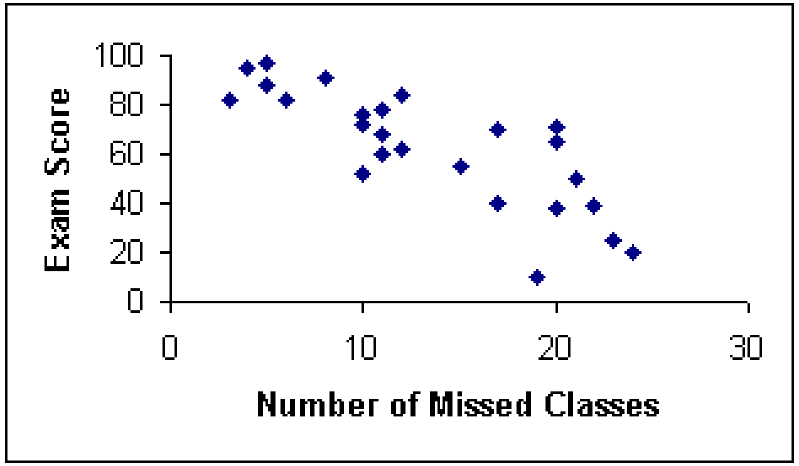
On the other hand, **regression** estimates the best straight line to summarize the association.

Which is the easiest one to model using regression?



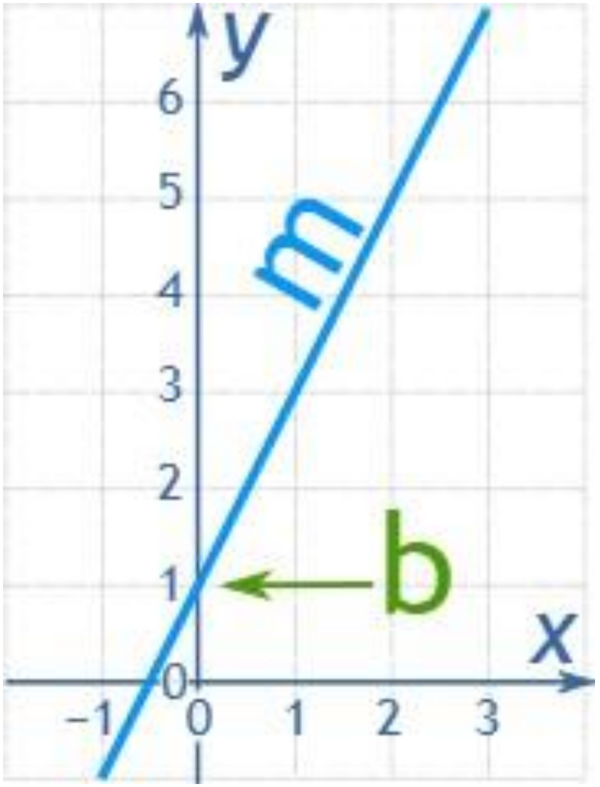
Which is the easiest one to model using regression?





What kind of correlation exists here?

Refresher: Linear Equation



A linear equation is an **equation for a straight line**

$y = 2x + 1$ is a linear equation as graphed

When x increases, y increases twice as fast (=**slope**)

When x is 0, y is already 1 (=**intercept**)

So, $y = 2x + 1$

Refresher: Linear Equation Exercise

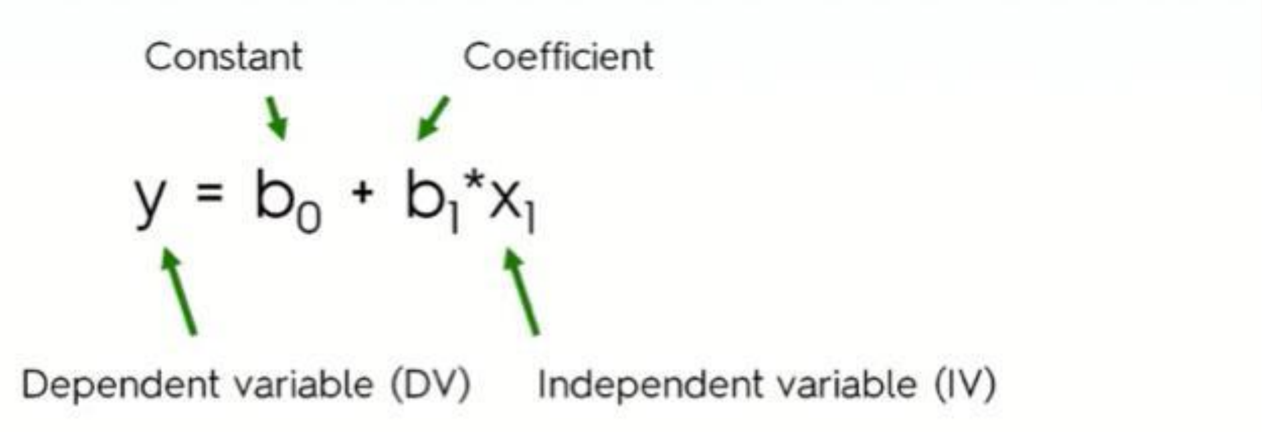
Try out and graph the following linearequations at
https://www.desmos.com/calculator

1) $y = 5$
2) $y = 2x$
3) $y = 2x + 1$
4) $y = x - 5$
5) $y = 0.5x + 2$

6) $y = 10000x + 30000$*

*You might need to adjust the scaling on X- and Y-axis

Linear Regression



- **Dependent variable (DV)**: the variable that you try to understand in terms of its dependence on another variable

- **Independent variable (IV)**: the variable that affects the dependent variable

- **Coefficient**: The independent variable's coefficient basically determines how a one-unit change in the IV can affect the DV

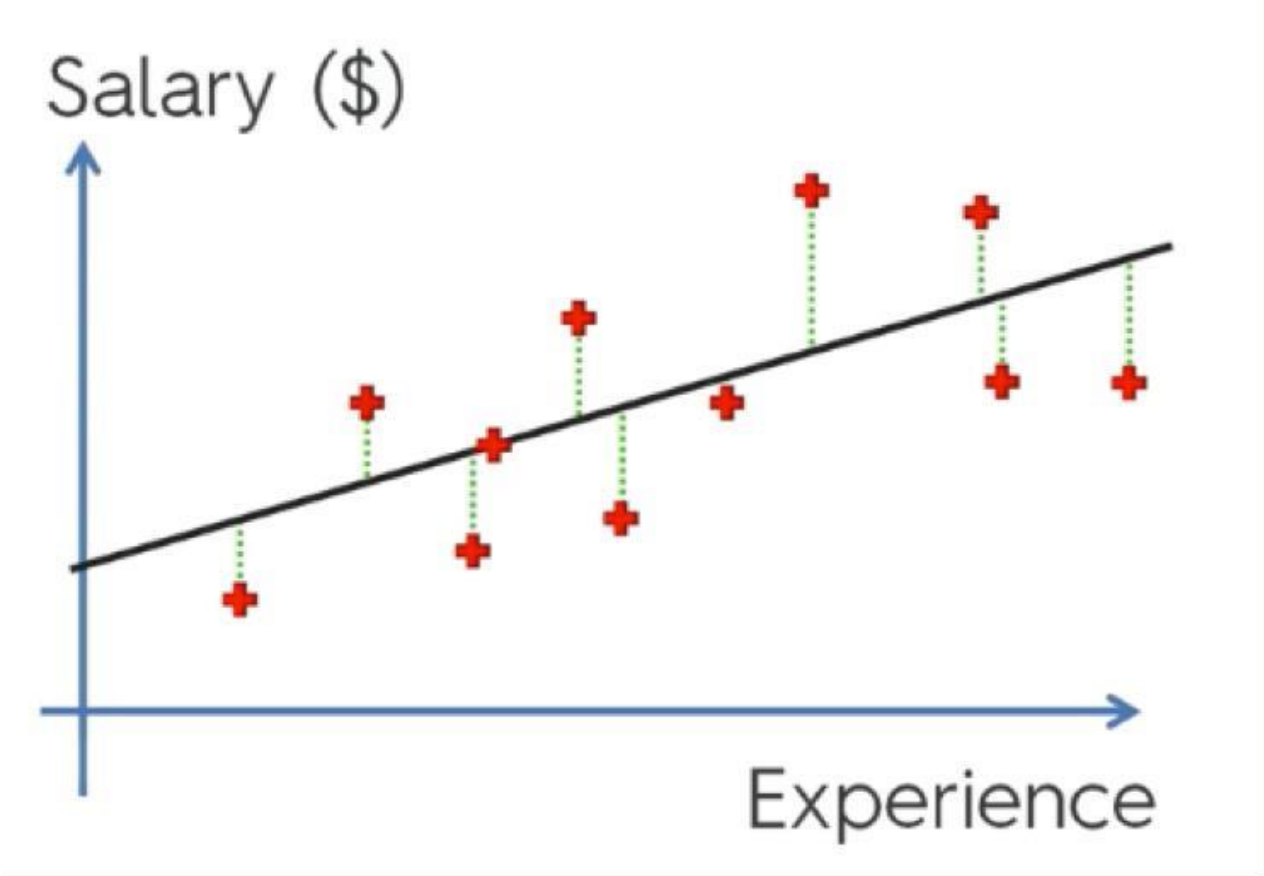
- **Constant**: The point where the straight line intersects with the Y-axis.

Linear Regression: Problem Examples

I want to know how number of hours jogging would affect the body fat level.
- Independent Variable (IV)?
- Dependent Variable (DV)?

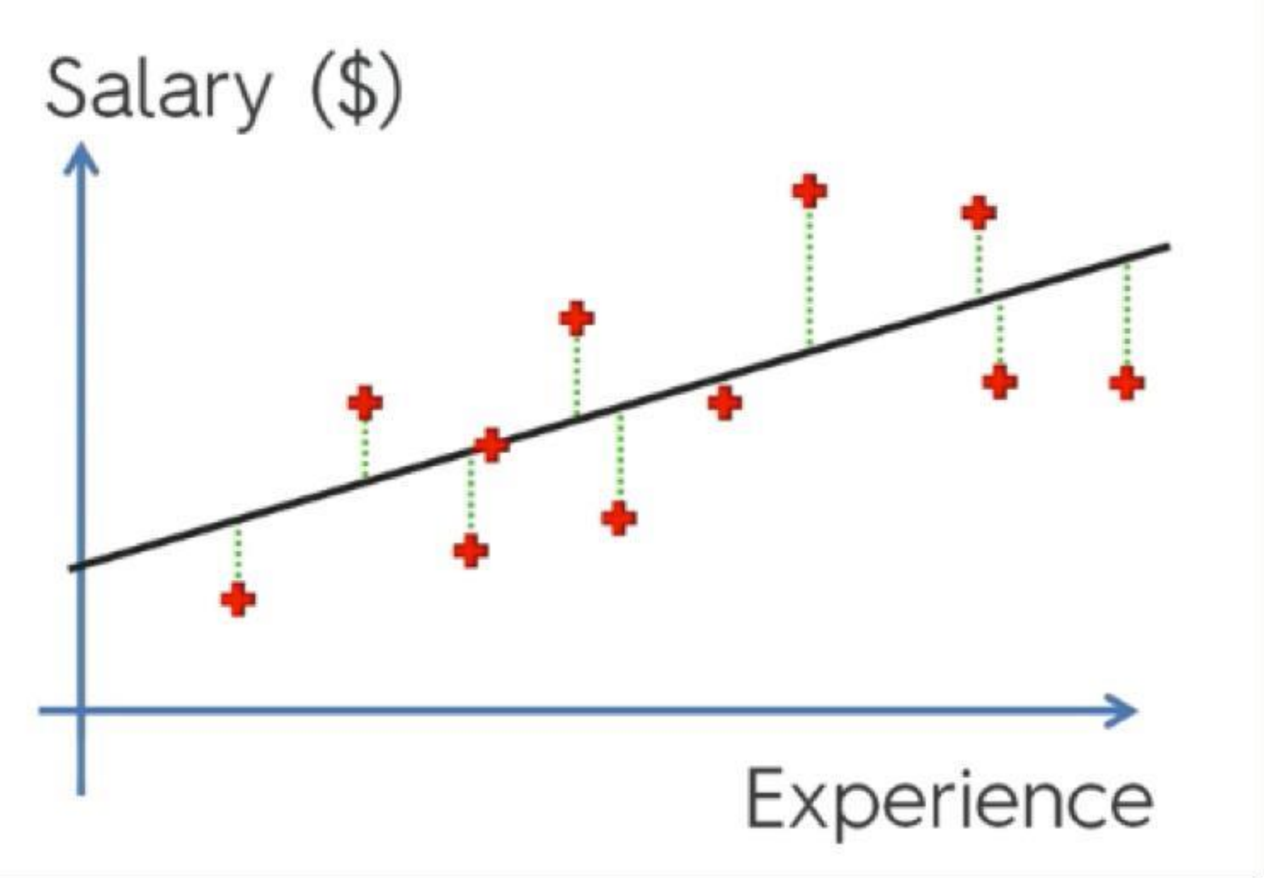
I want to know how studying hours would affect the GPA.
- Independent Variable (IV)?
- Dependent Variable (DV)?

Regression: Experience vs. Salary

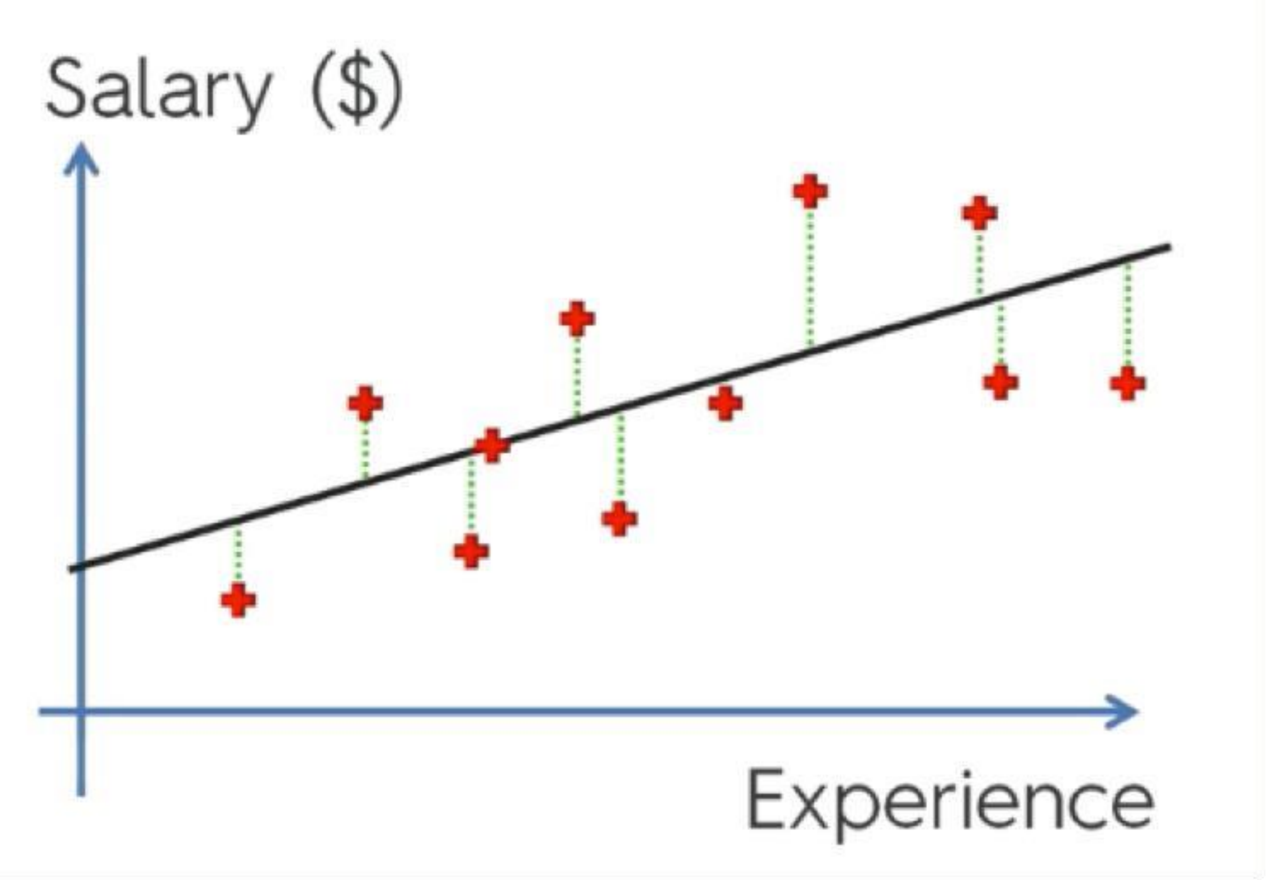


(Linear) Regression is the problem to find the **best fitting straight line** of data

Linear Regression Interpretation



Linear Regression Interpretation

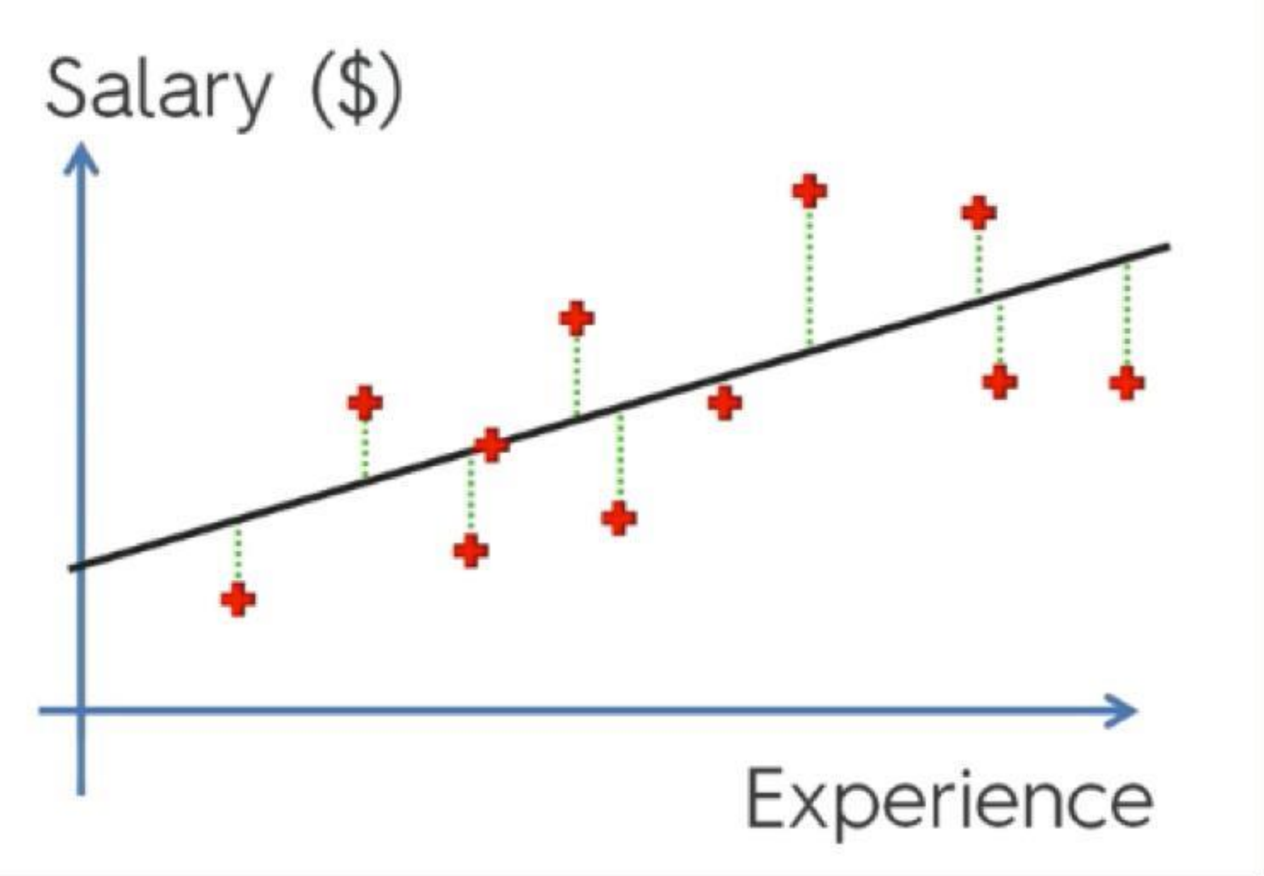


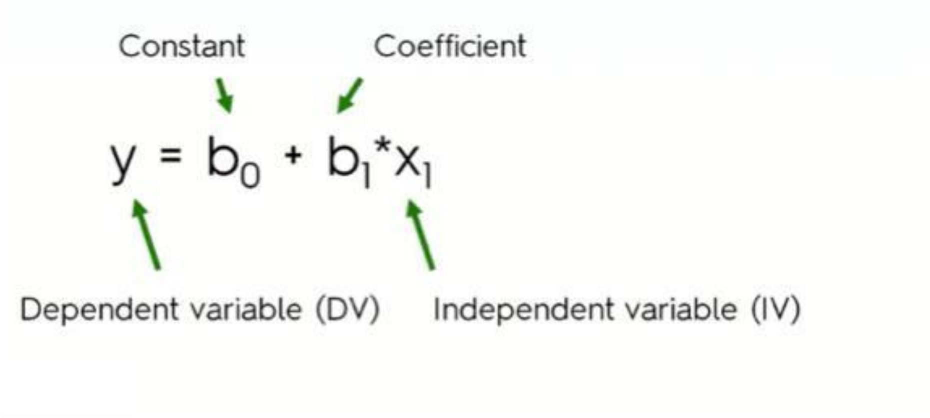
Constant　　Coefficient

$$y = b_0 + b_1 * x_1$$

Dependent variable (DV)　　Independent variable (IV)

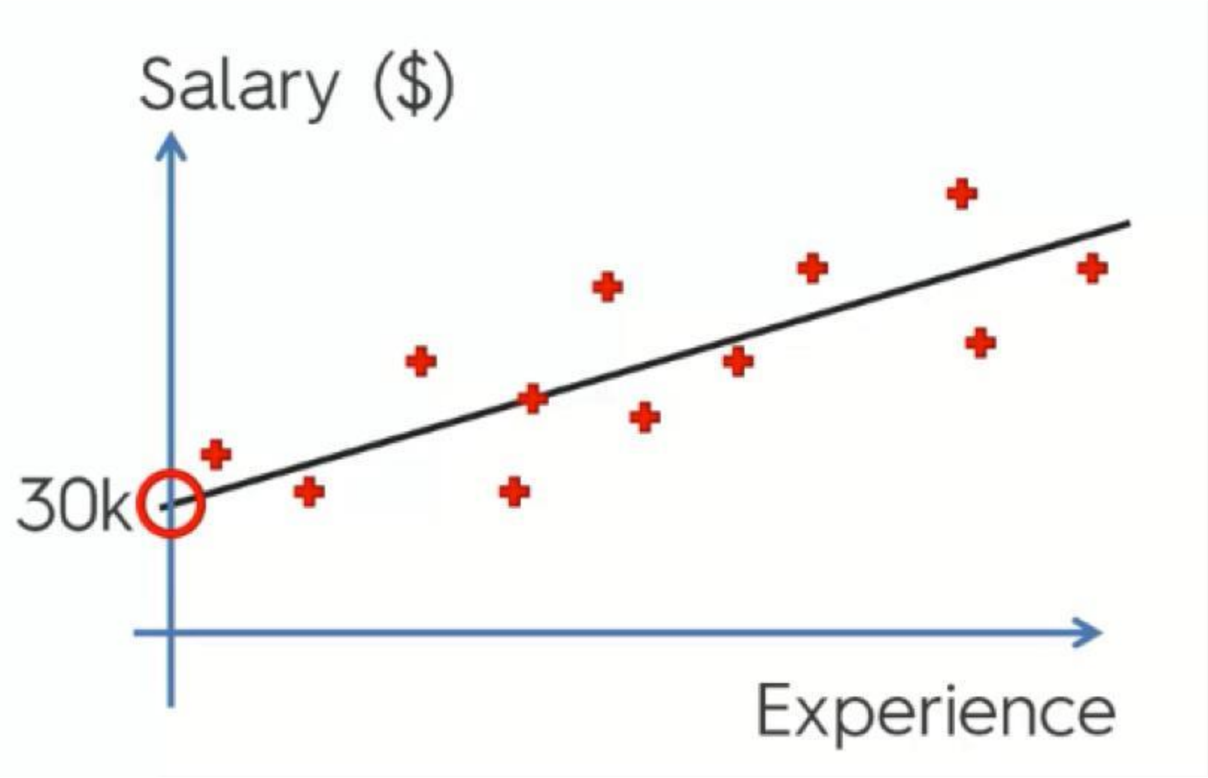
Linear Regression Interpretation

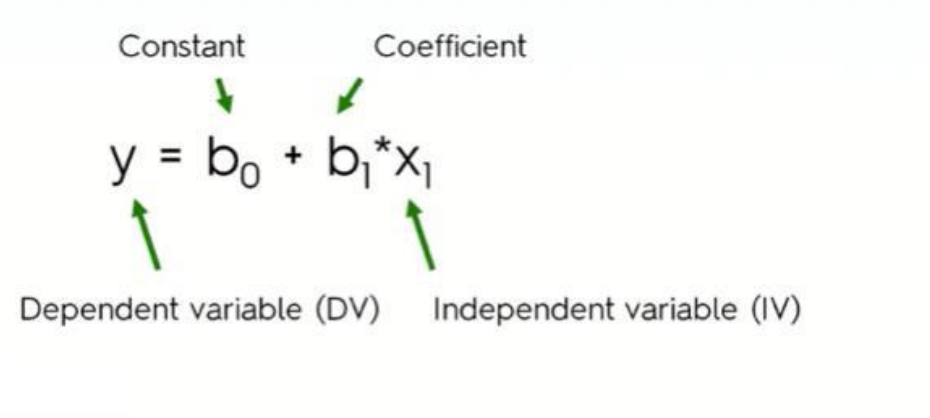




Salary = b_0 + b_1 * Experience

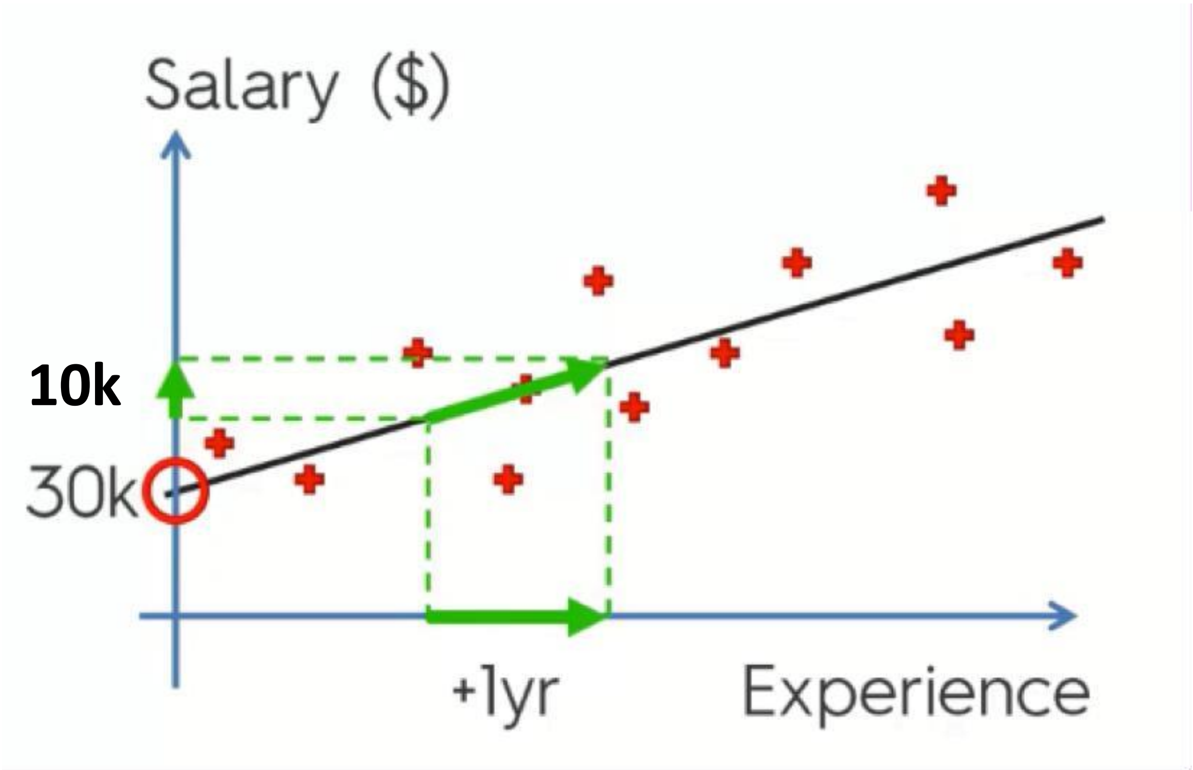
Linear Regression Interpretation





Salary = 30000 + b_1 * Experience

Linear Regression Interpretation



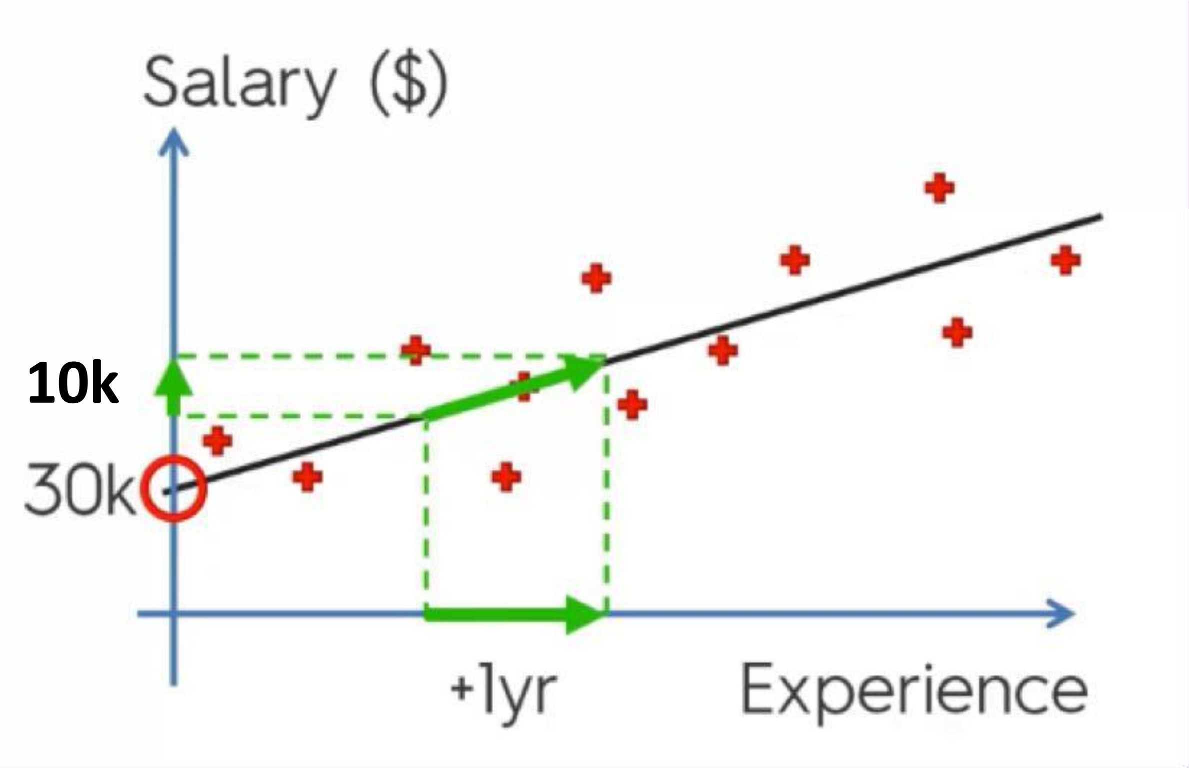
Constant Coefficient

$$y = b_0 + b_1{*}x_1$$

Dependent variable (DV) Independent variable (IV)

Salary = 30000 + b_1 * Experience

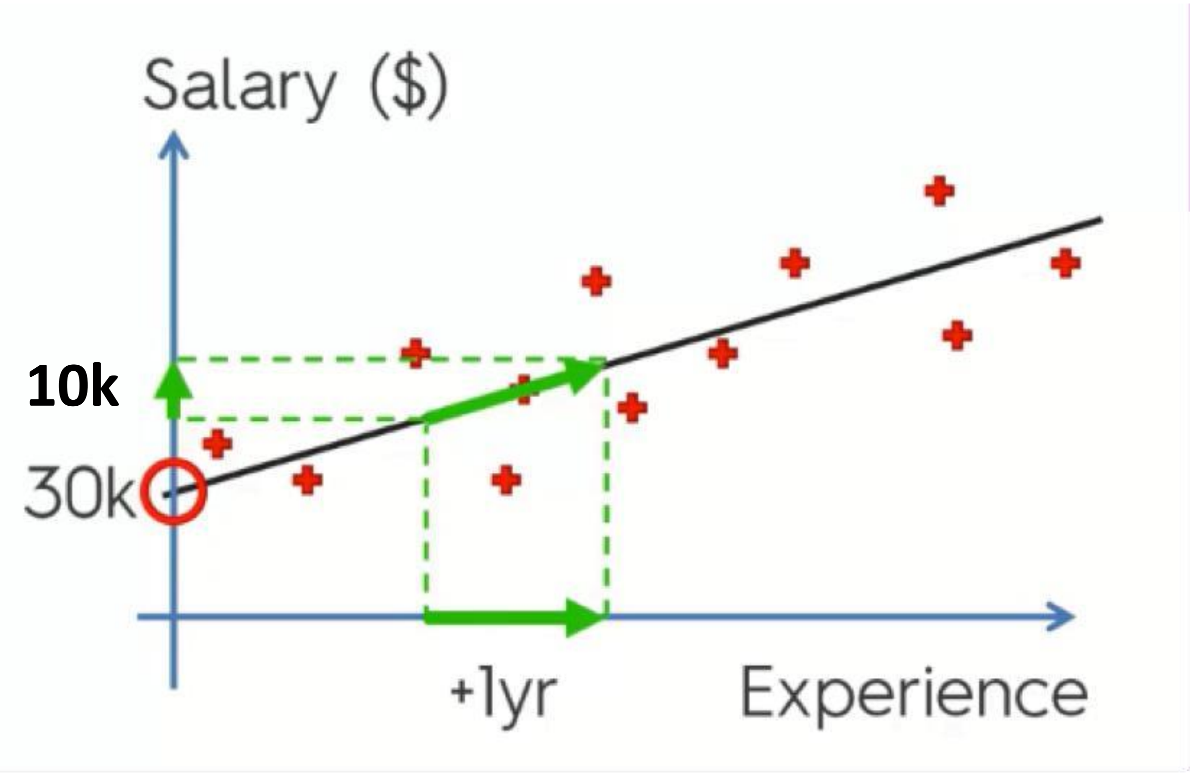
Linear Regression Interpretation



$$\underbrace{y}_{\text{Dependent variable (DV)}} = \overbrace{b_0}^{\text{Constant}} + \overbrace{b_1}^{\text{Coefficient}} * \underbrace{x_1}_{\text{Independent variable (IV)}}$$

Salary = 30000 + 10000 * Experience

Linear Regression Interpretation



Constant Coefficient

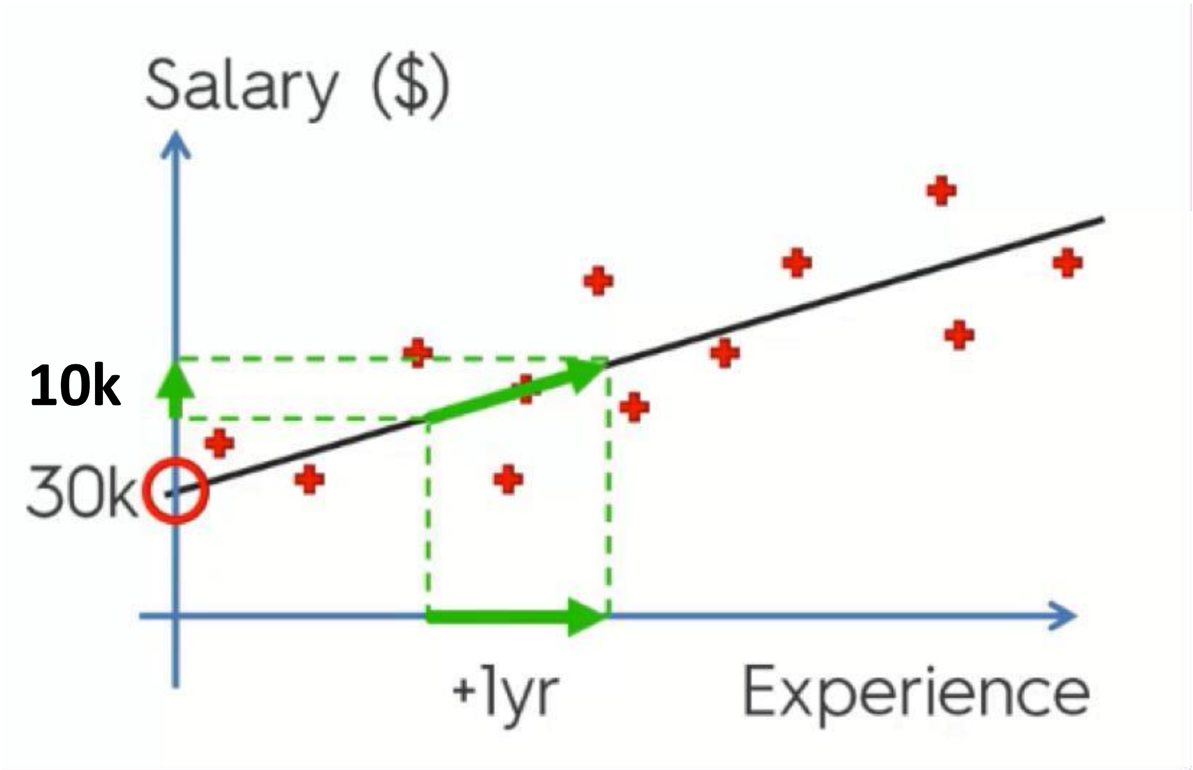
$$y = b_0 + b_1 {}^{*} x_1$$

Dependent variable (DV) Independent variable (IV)

Salary = 30000 + 10000 * Experience

Question: Salary after 5 years?

Linear Regression Interpretation



Constant   Coefficient

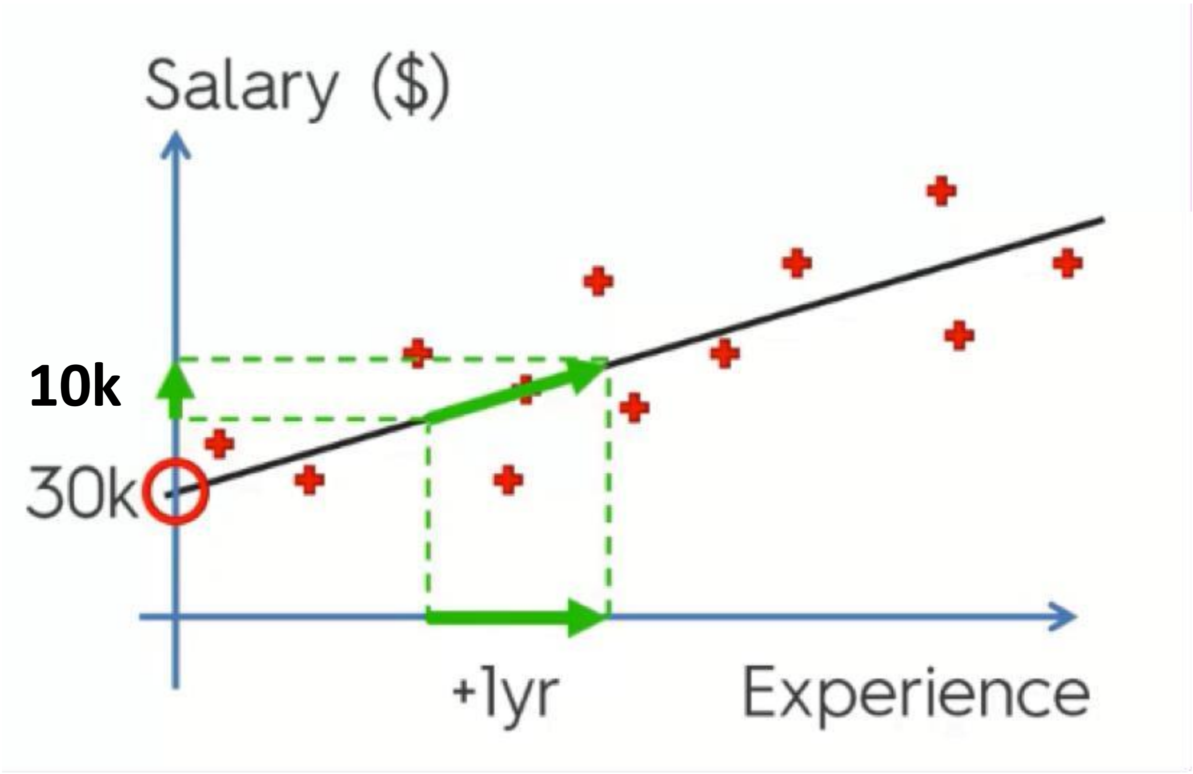
$$y = b_0 + b_1 * x_1$$

Dependent variable (DV)   Independent variable (IV)

Salary = 30000 + 10000 * Experience

Question: Salary after 10 years?

Linear Regression Interpretation



Constant Coefficient

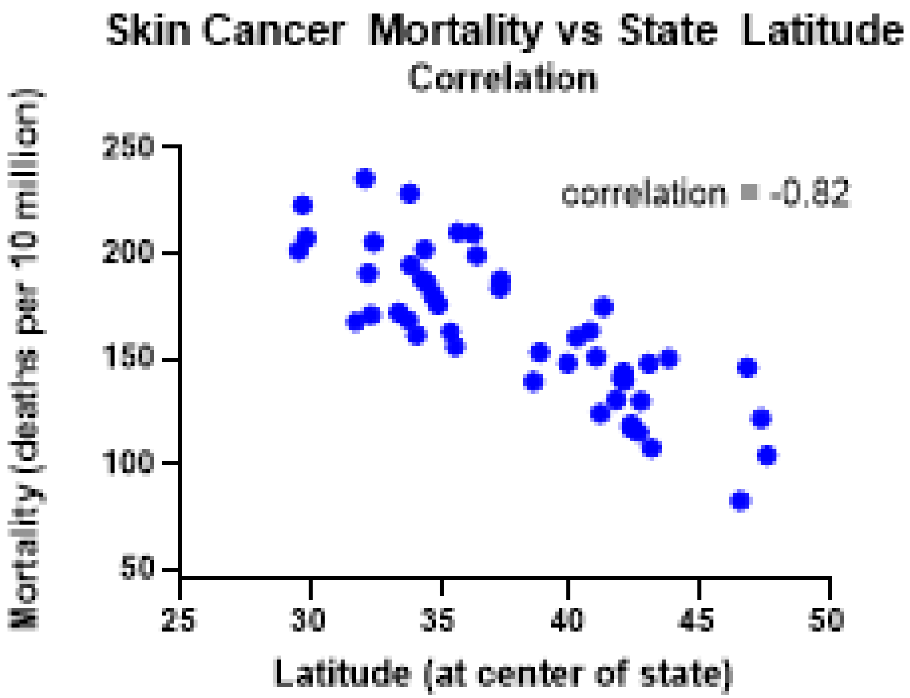
$$y = b_0 + b_1 * x_1$$

Dependent variable (DV) Independent variable (IV)

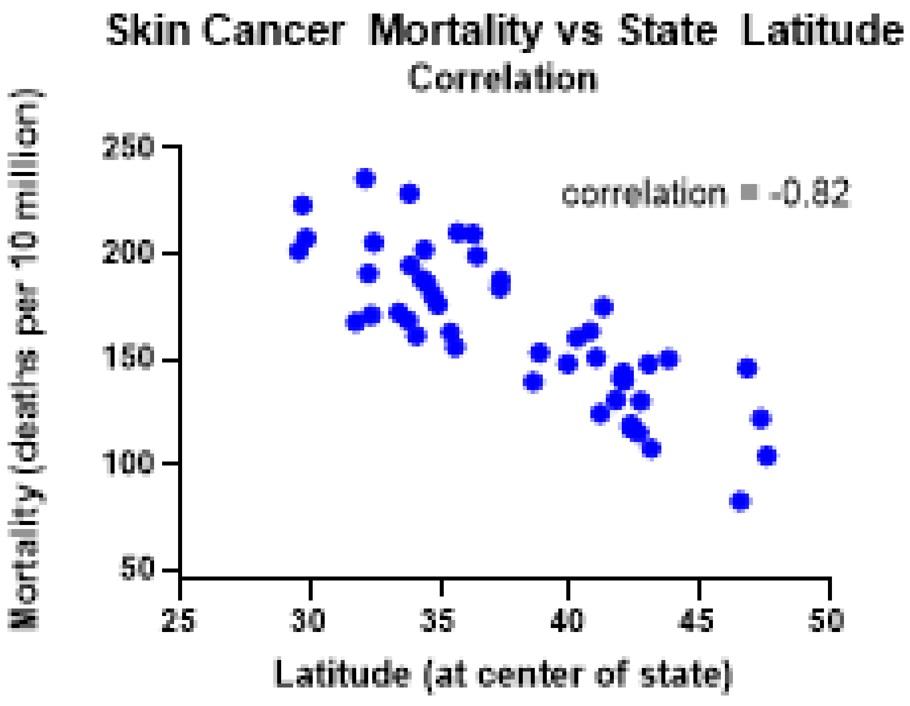
Salary = 30000 + 10000 * Experience

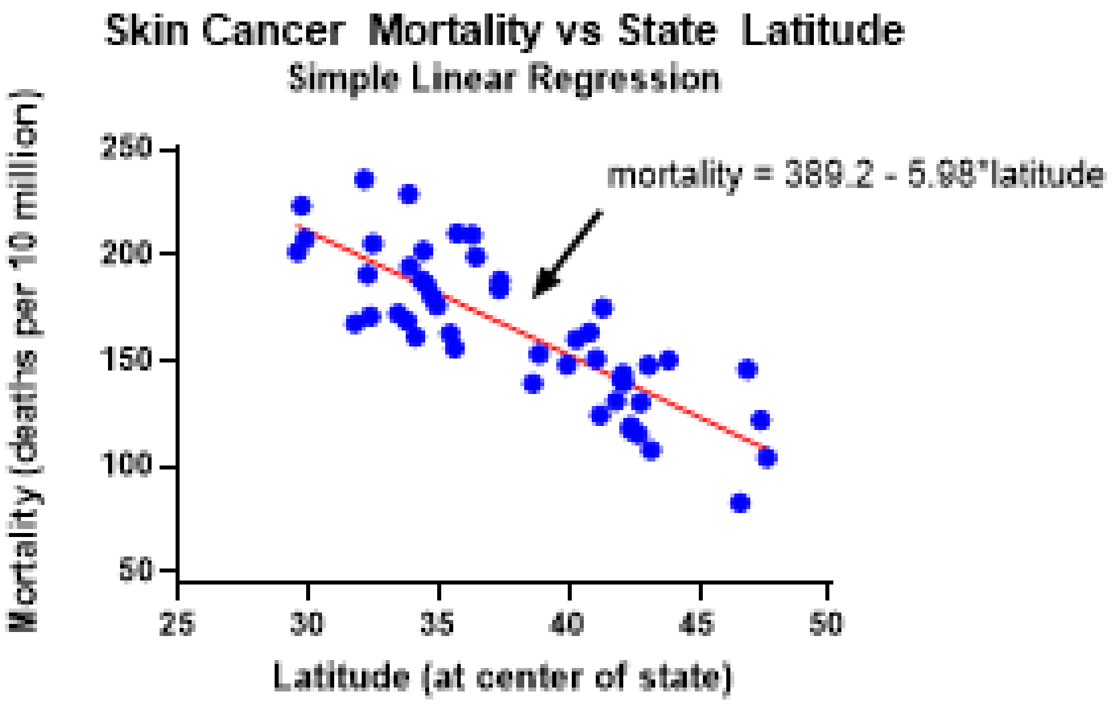
Question: Starting salary?

Linear Regression Interpretation

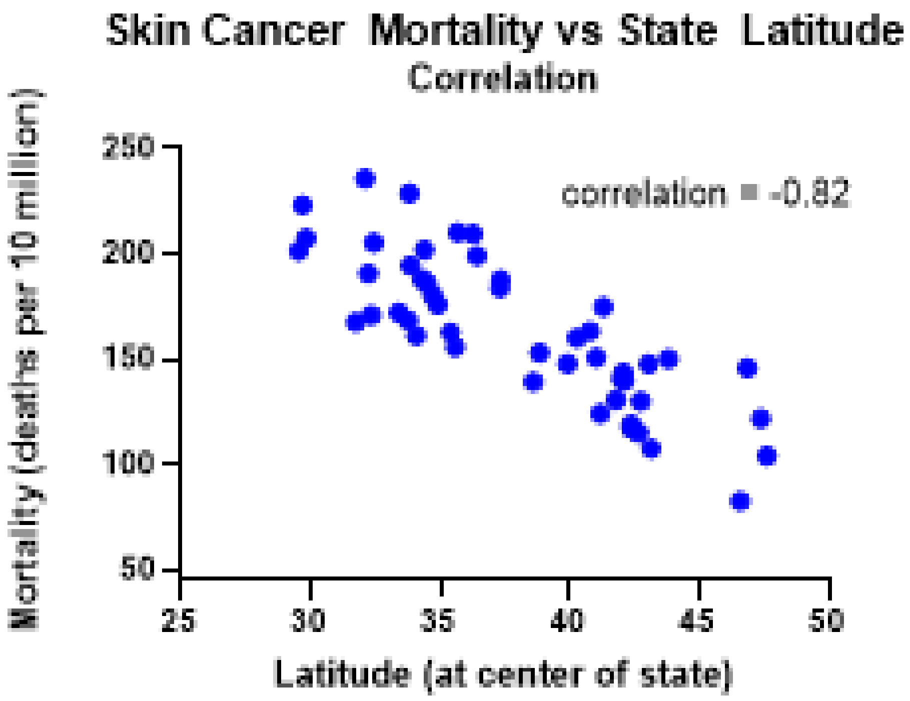


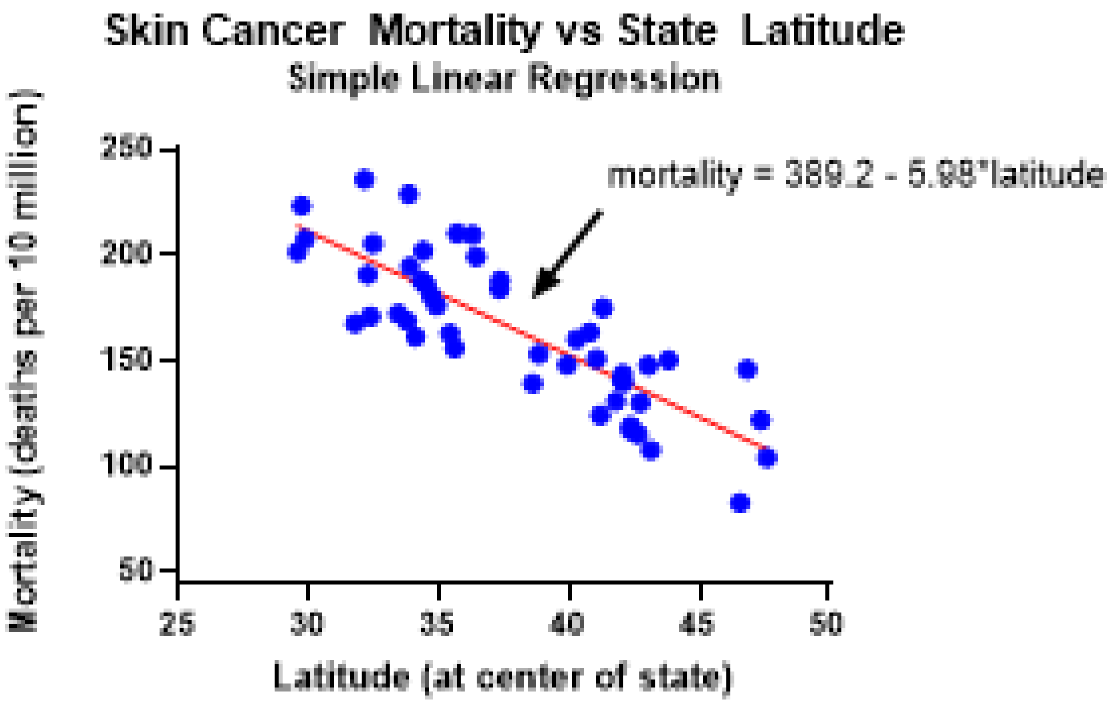
Linear Regression Interpretation





Linear Regression Interpretation

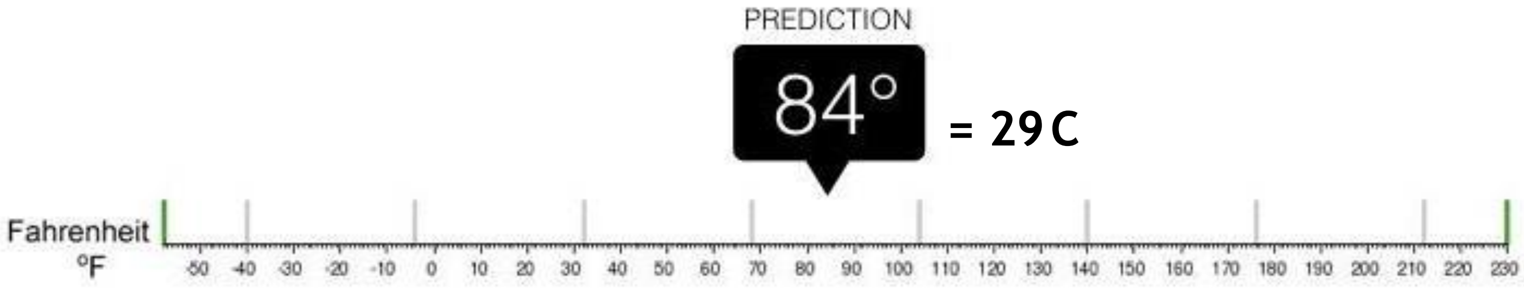




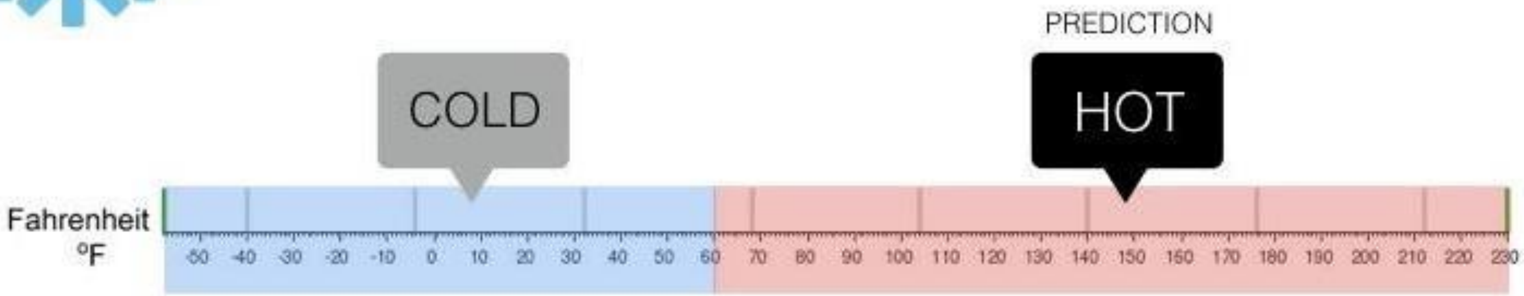
Question: A city at latitude 40 would be expected to have mortality rate of?

More on Regression

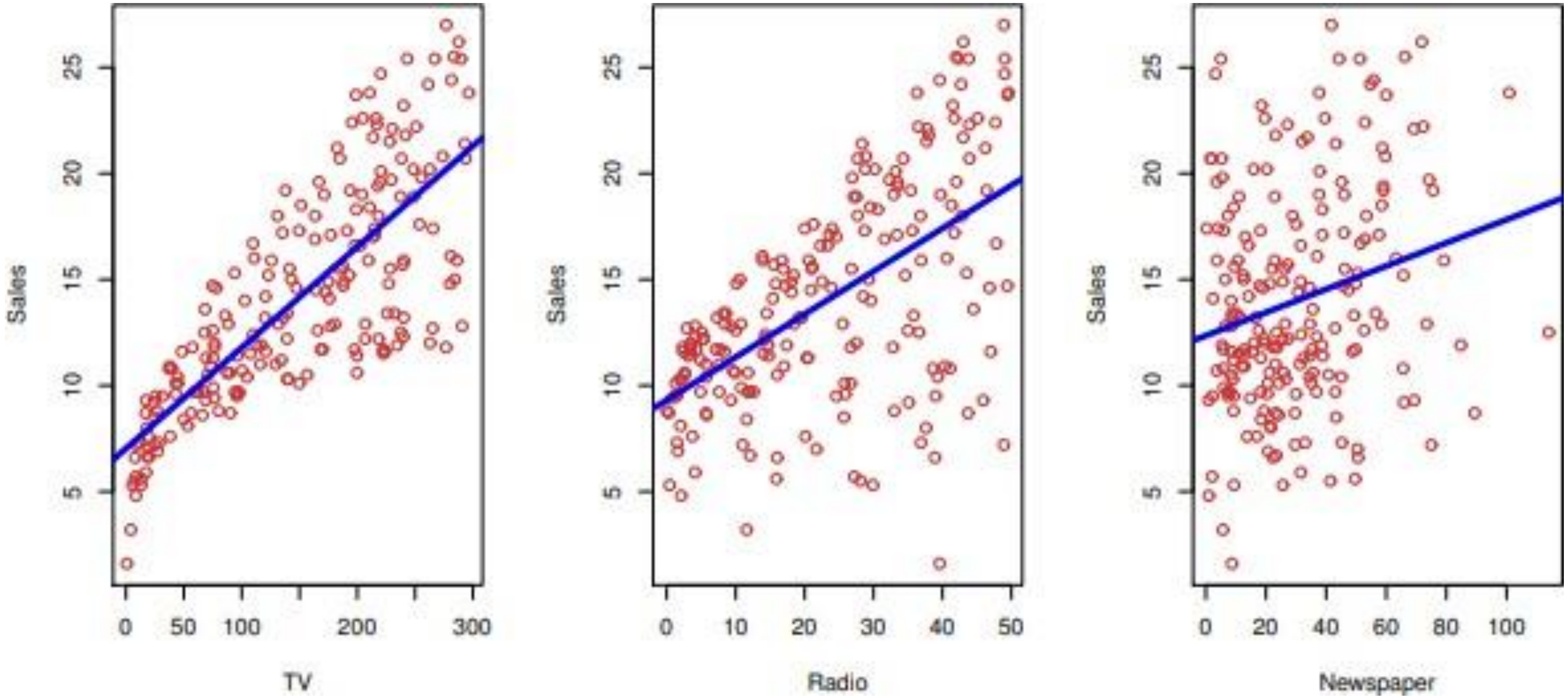
What is the temperature going to be tomorrow?



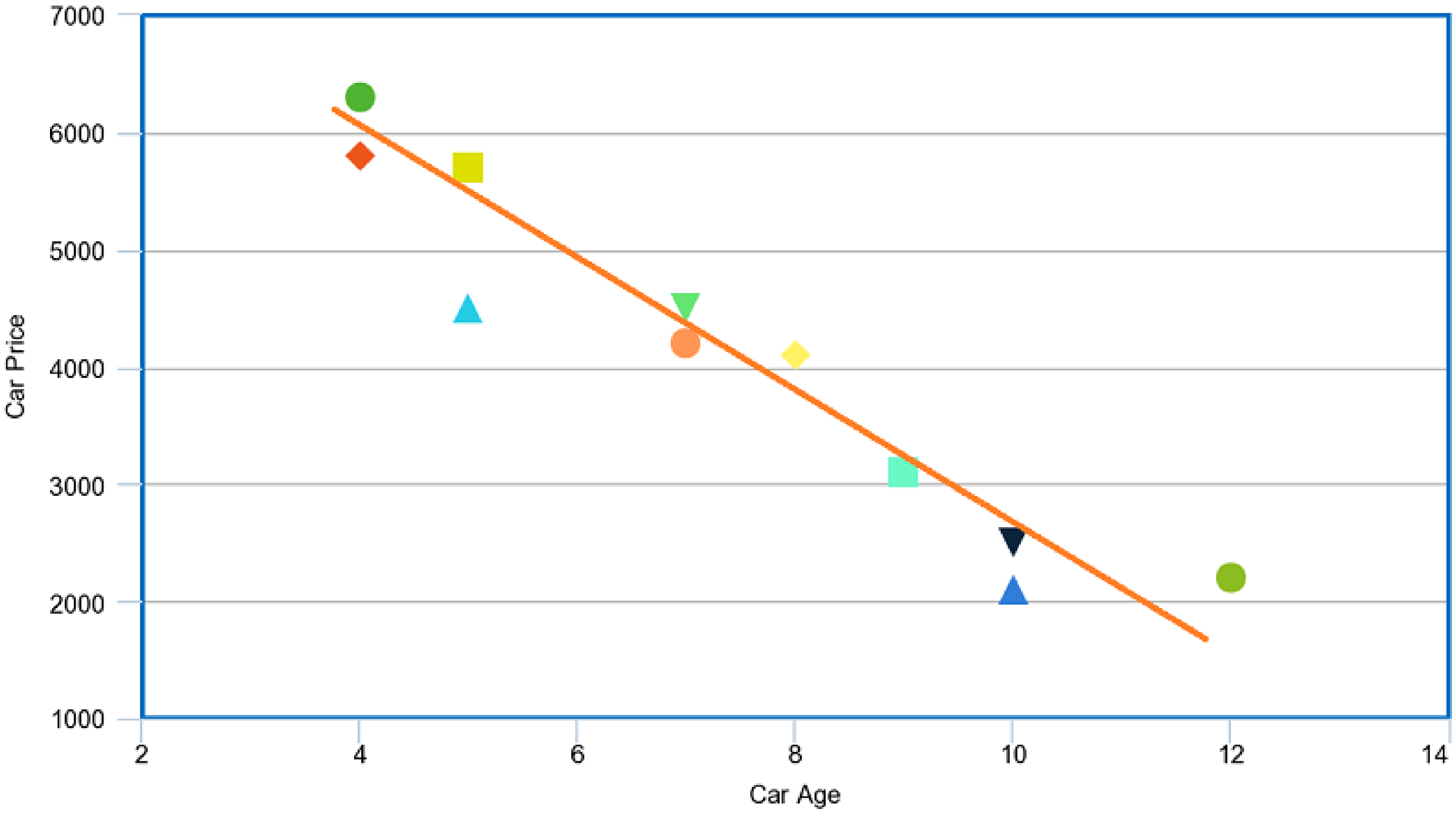
Will it be Cold or Hot tomorrow?

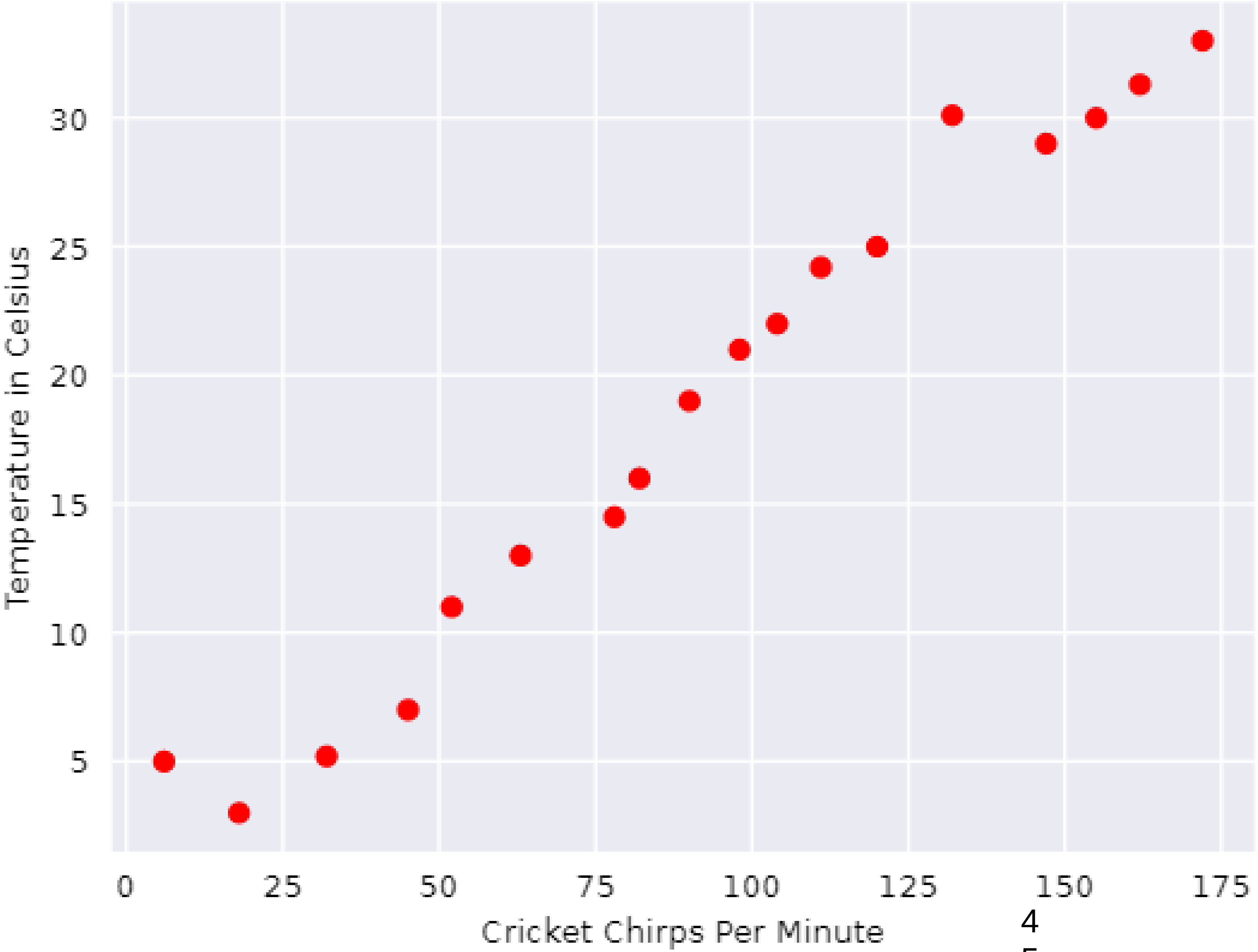


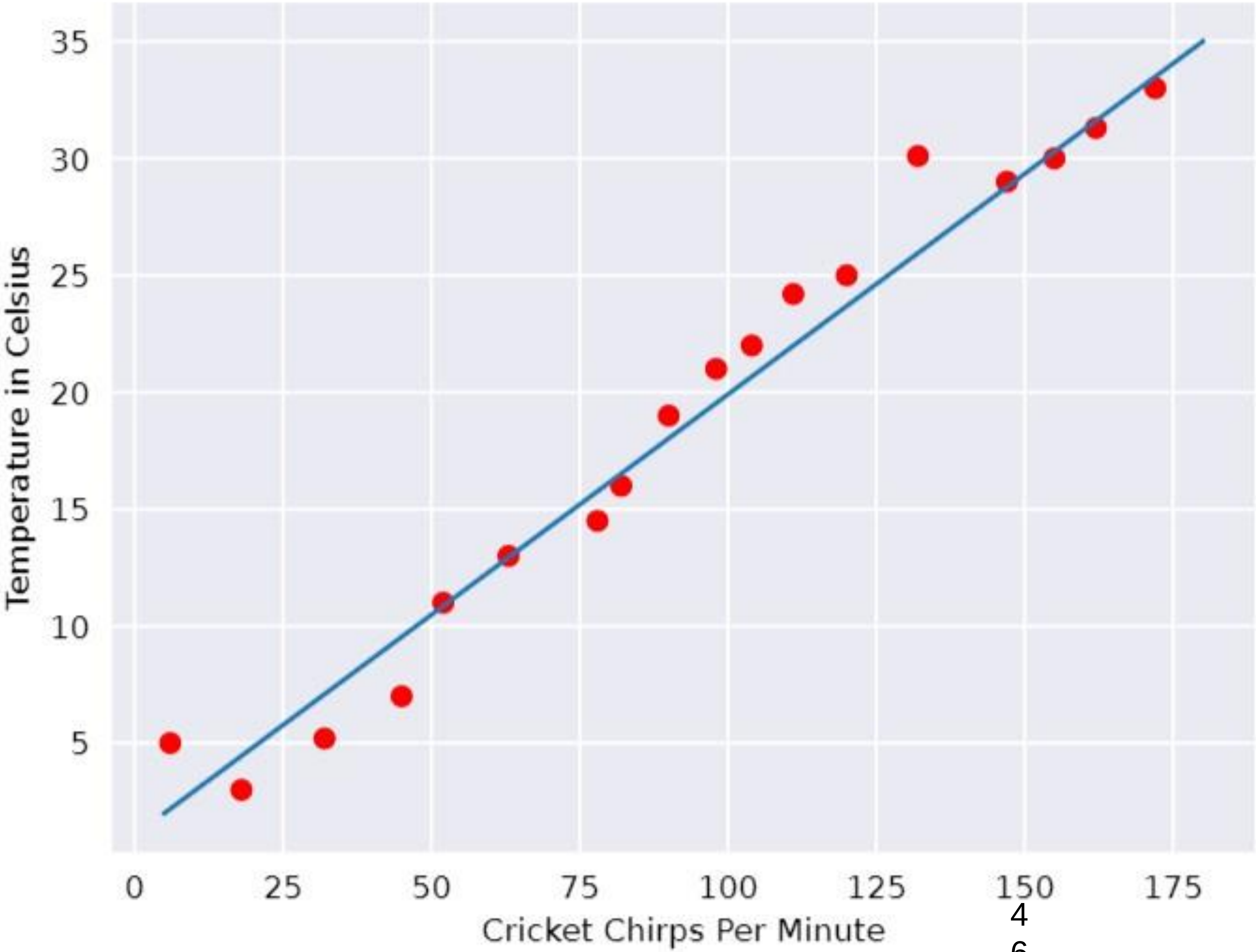
Advertising budgets on different mediums vssales



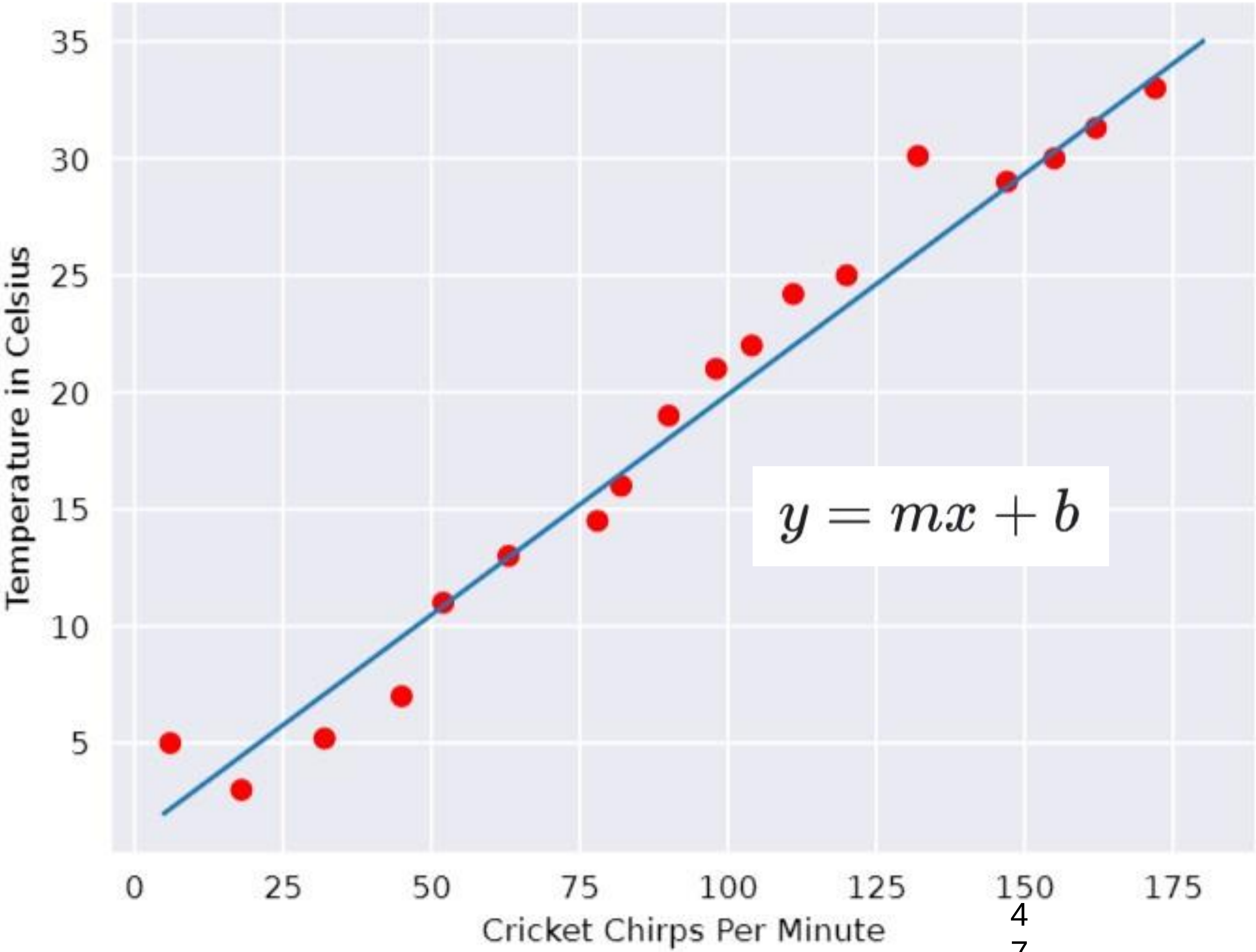
Car age vs car price




Temperature in Celsius
30
25
20
15
10
5
0
25
50
75
100
125
150
175
Cricket Chirps Per Minute

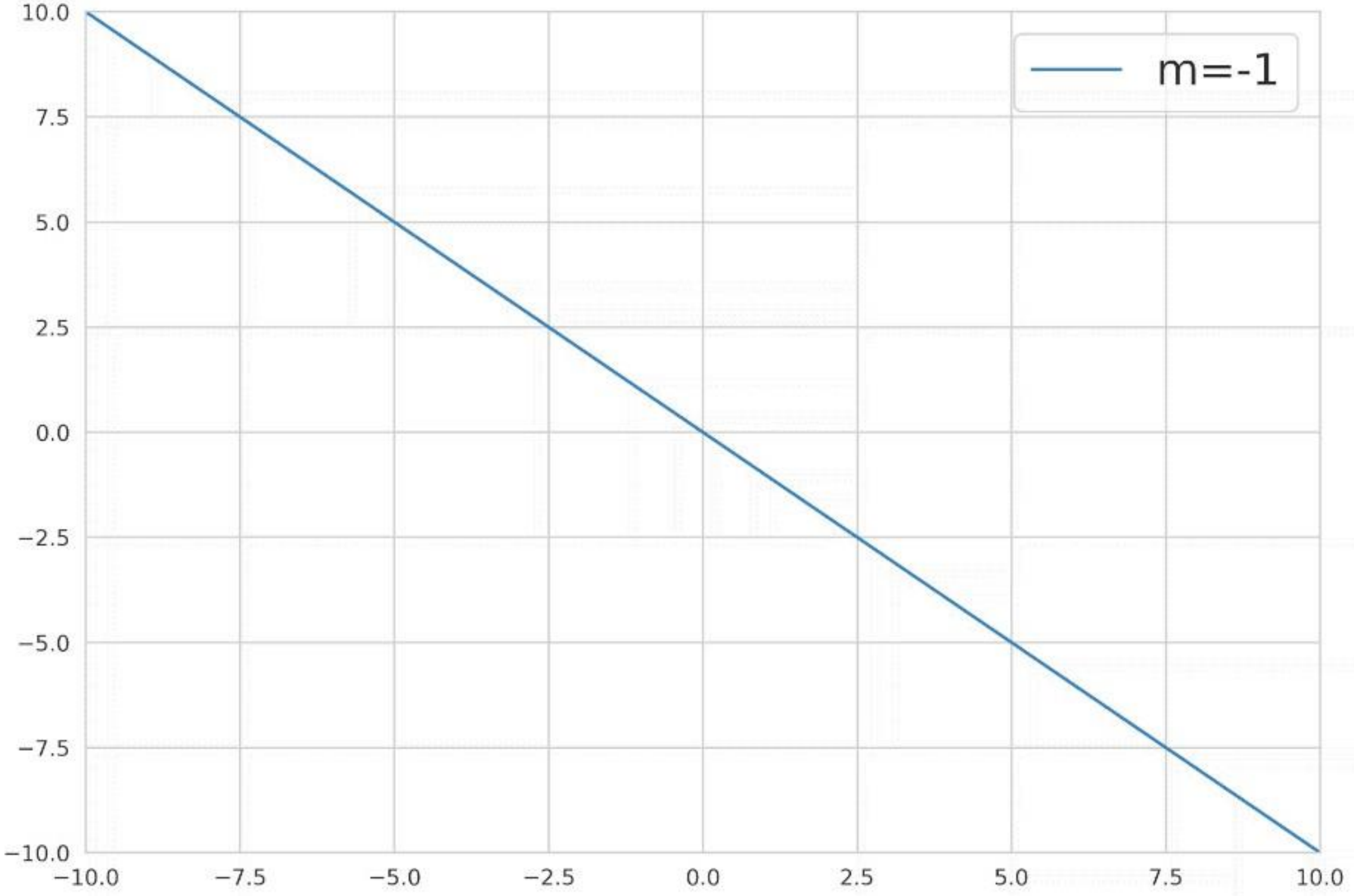


35
30
25
20
15
10
5
Temperature in Celsius
0
25
50
75
100
125
150
175
Cricket Chirps Per Minute

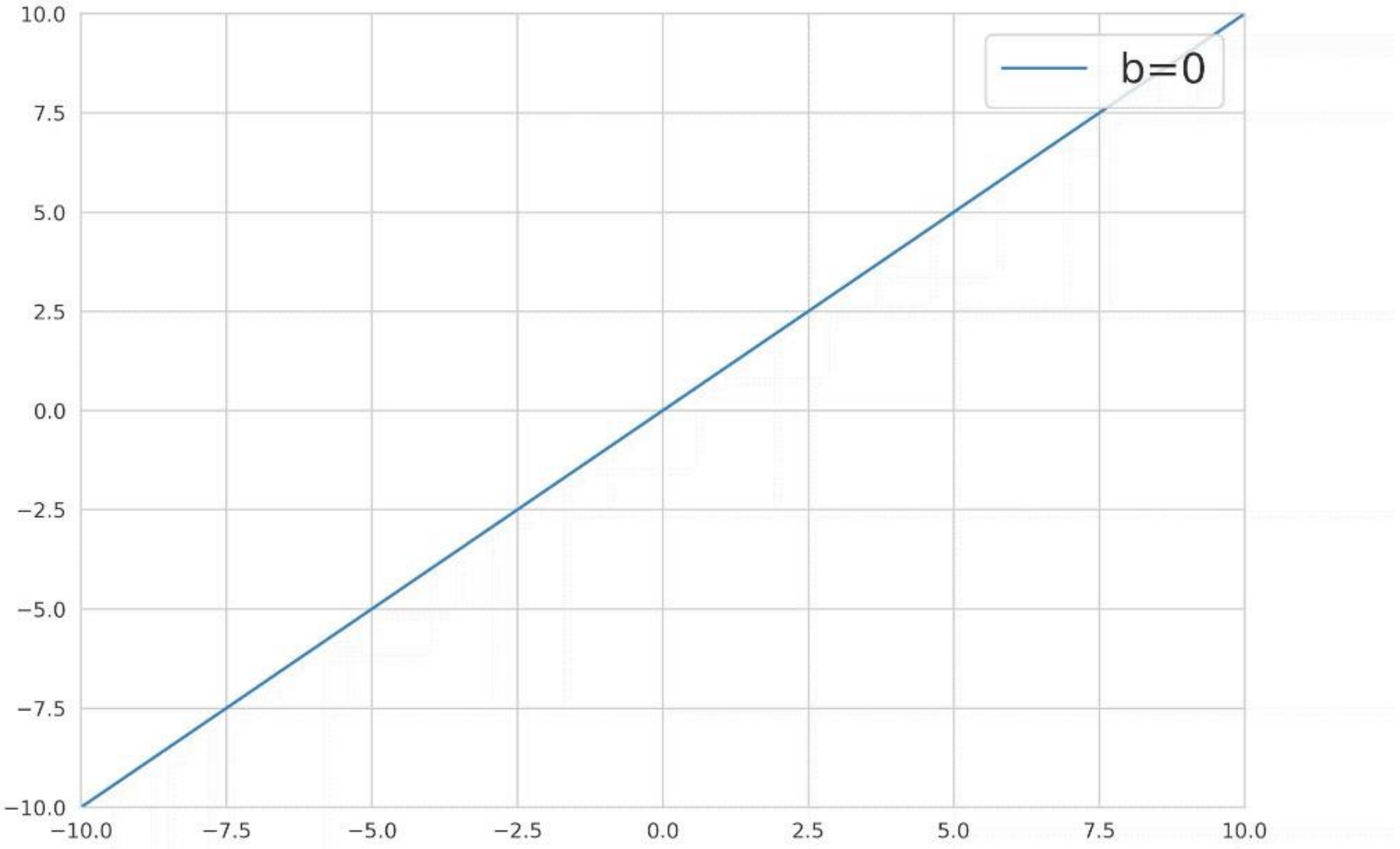


Temperature in Celsius
35
30
25
20
15
10
5
0
25
50
75
100
125
150
175
Cricket Chirps Per Minute
$y = mx + b$

Behavior of m on the line

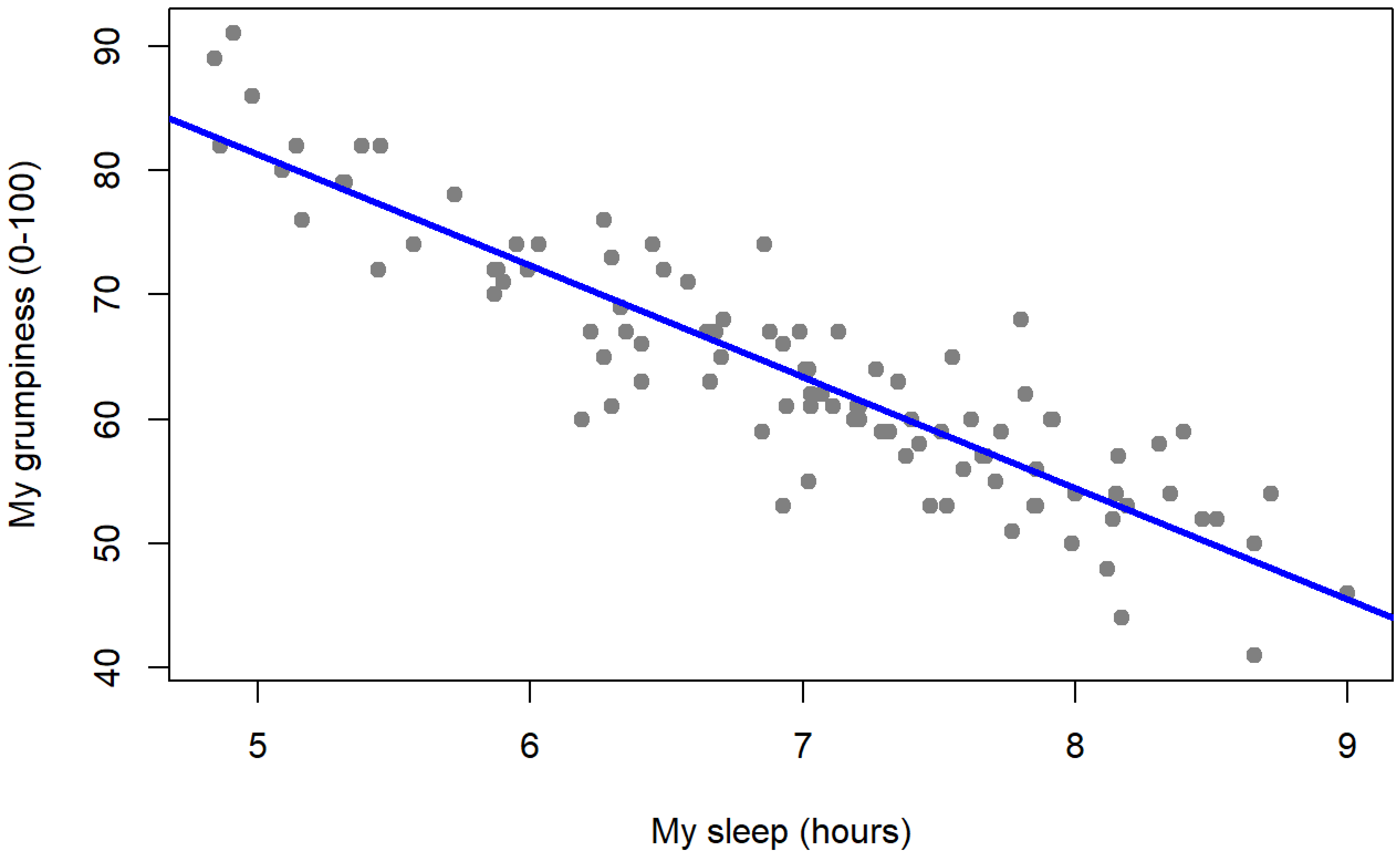


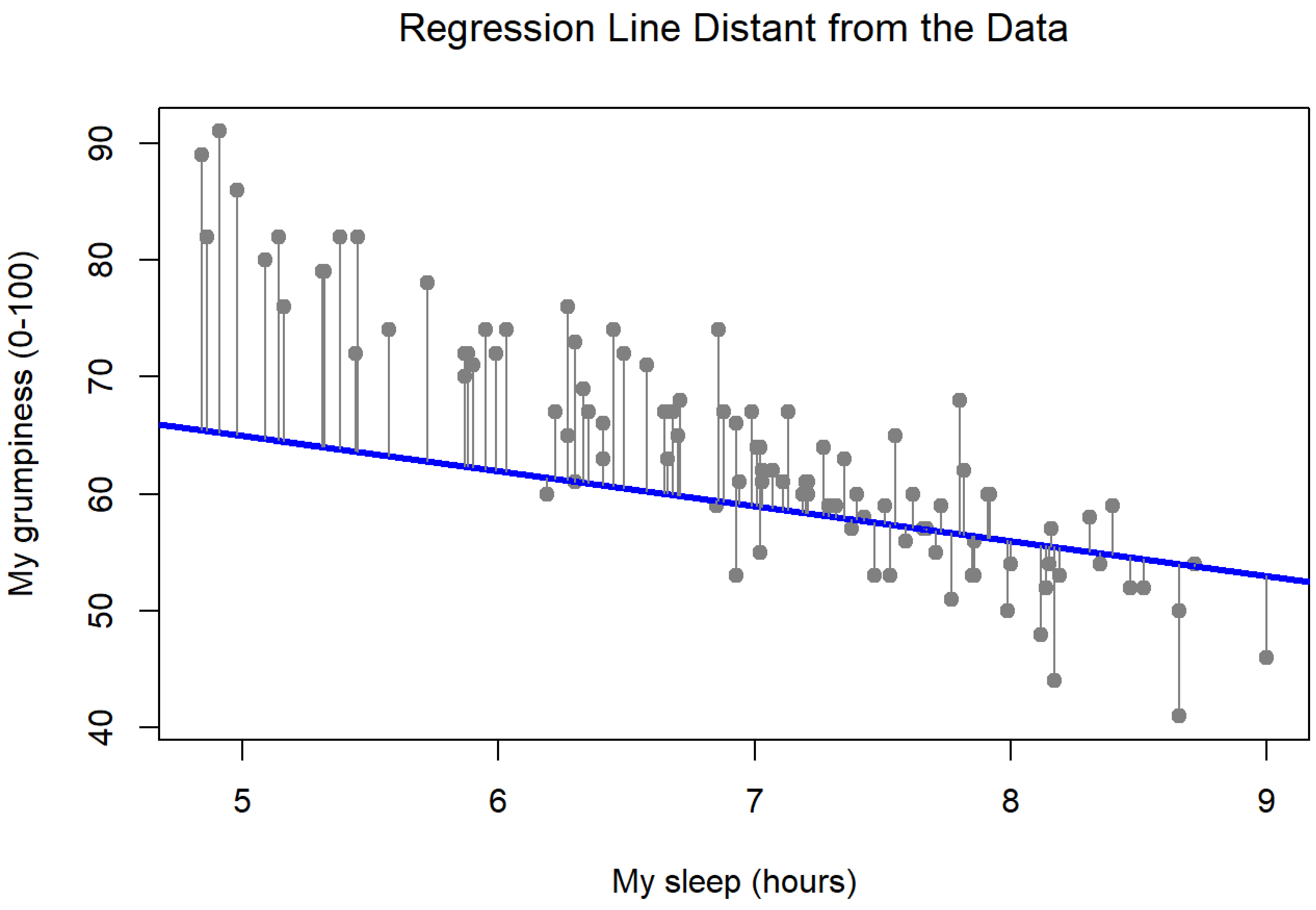
Behavior of b on the line




Not The Best Fitting Regression Line!
My grumpiness (0-100)
40
50
60
70
80
90
5
6
7
8
9
My sleep (hours)

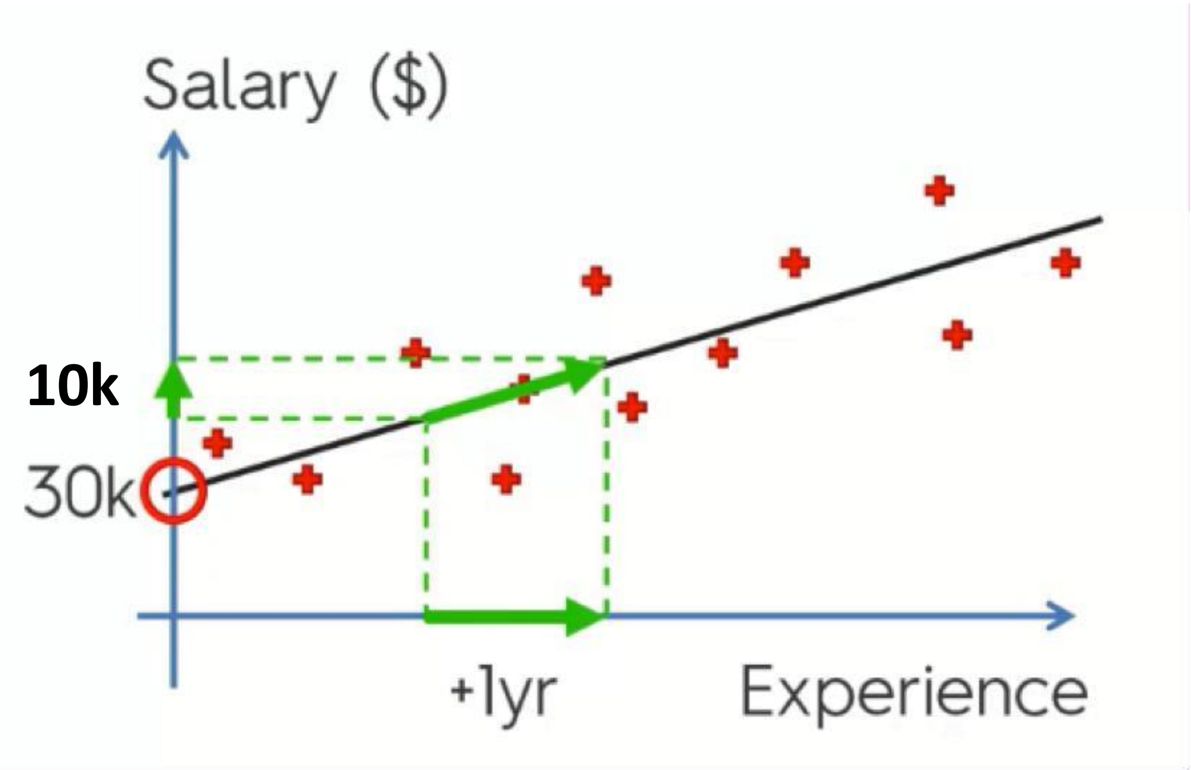
The Best Fitting Regression Line

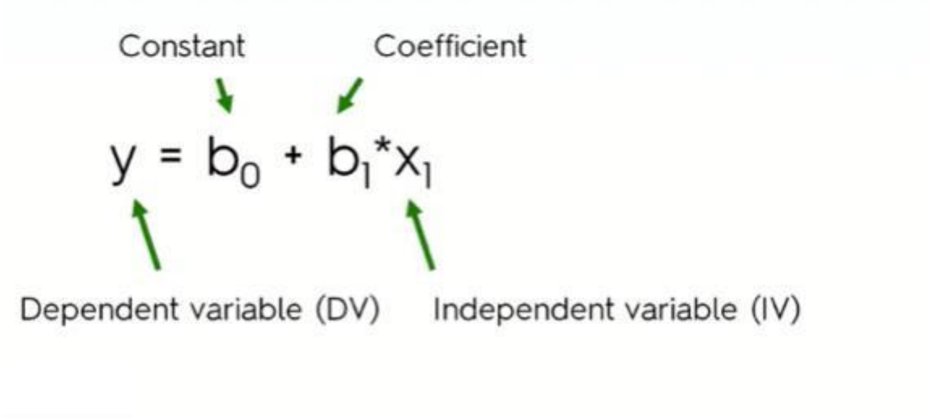



Regression Line Distant from the Data
My grumpiness (0-100)
40
50
60
70
80
90
5
6
7
8
9
My sleep (hours)


Regression Line Close to the Data
My grumpiness (0-100)
40
50
60
70
80
90
My sleep (hours)
5
6
7
8
9

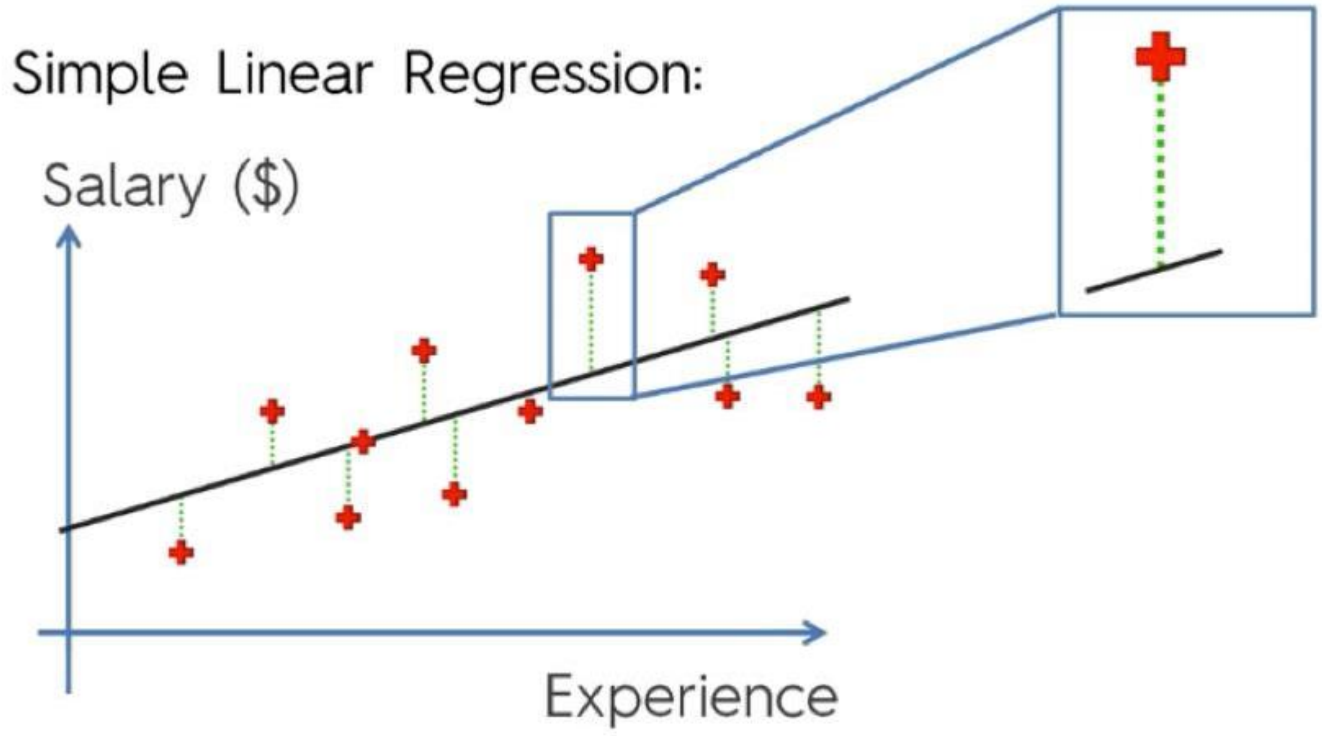
Linear Regression





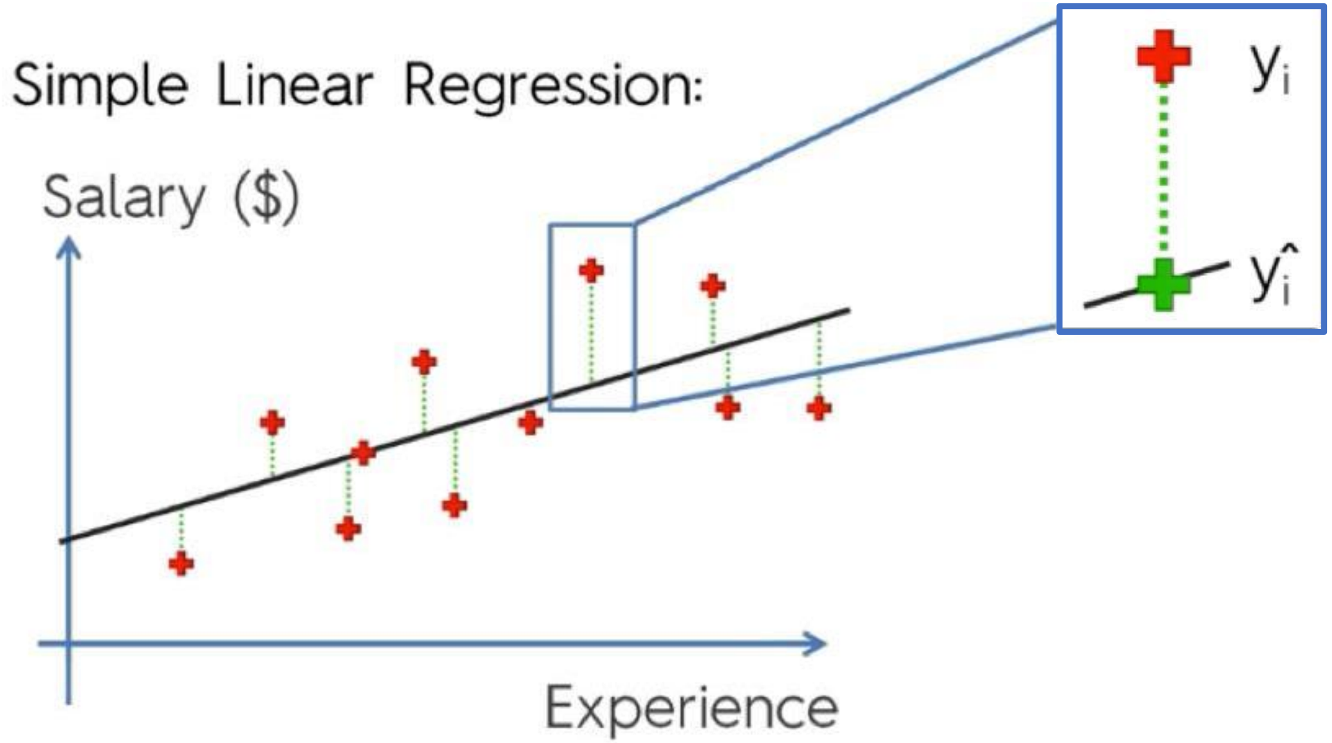
Salary = 30000 + 10000 * Experience

Linear Regression: Best fitting line



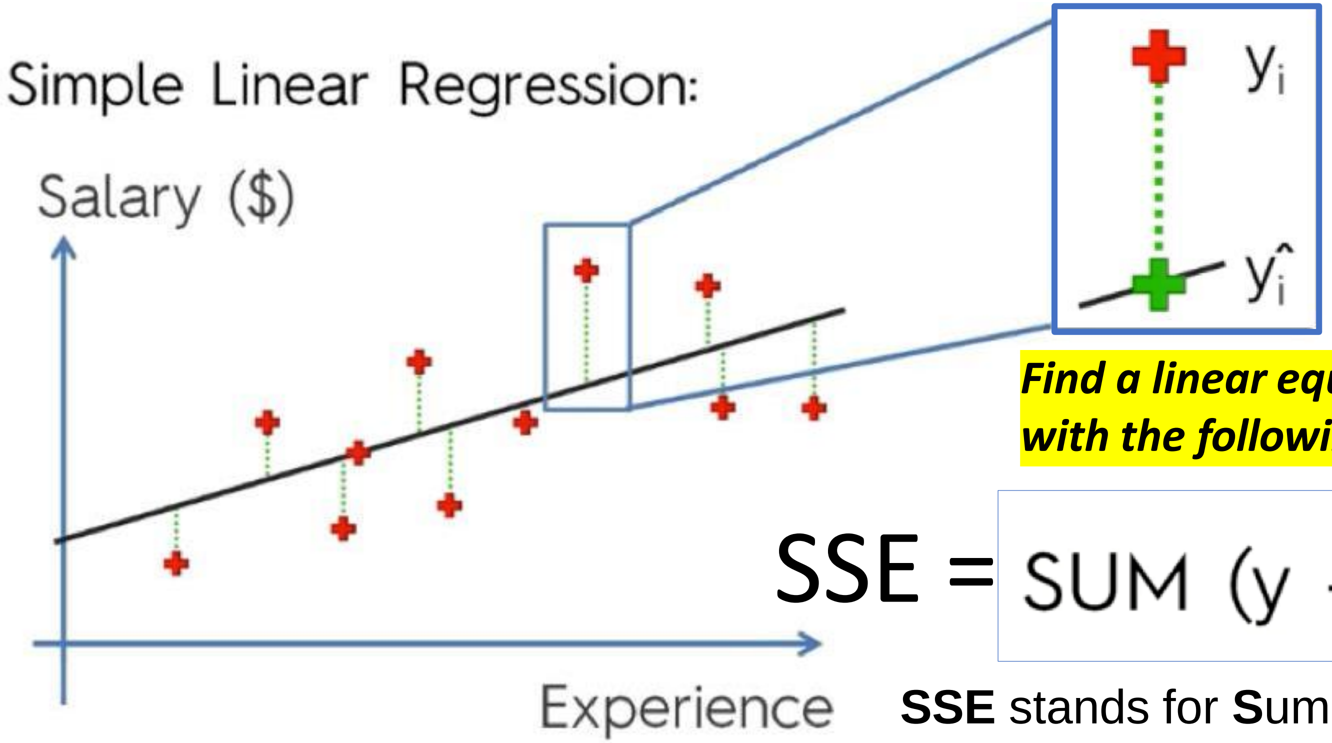
Here we have the salary of someone with x years of experience.
The straight line represents where that person's salary should be according to our linear regression model, whereas the red point is what that person is actually earning.

Linear Regression: Best fitting line



Here we have the salary of someone with x years of experience.
The straight line represents where that person's salary should be according to our linear regression model, whereas the red point is what that person is actually earning.

Linear Regression: Best fitting line



Find a linear equation with the following property:

$$SSE = \text{SUM } (y - \hat{y})^2 \text{ -> min}$$

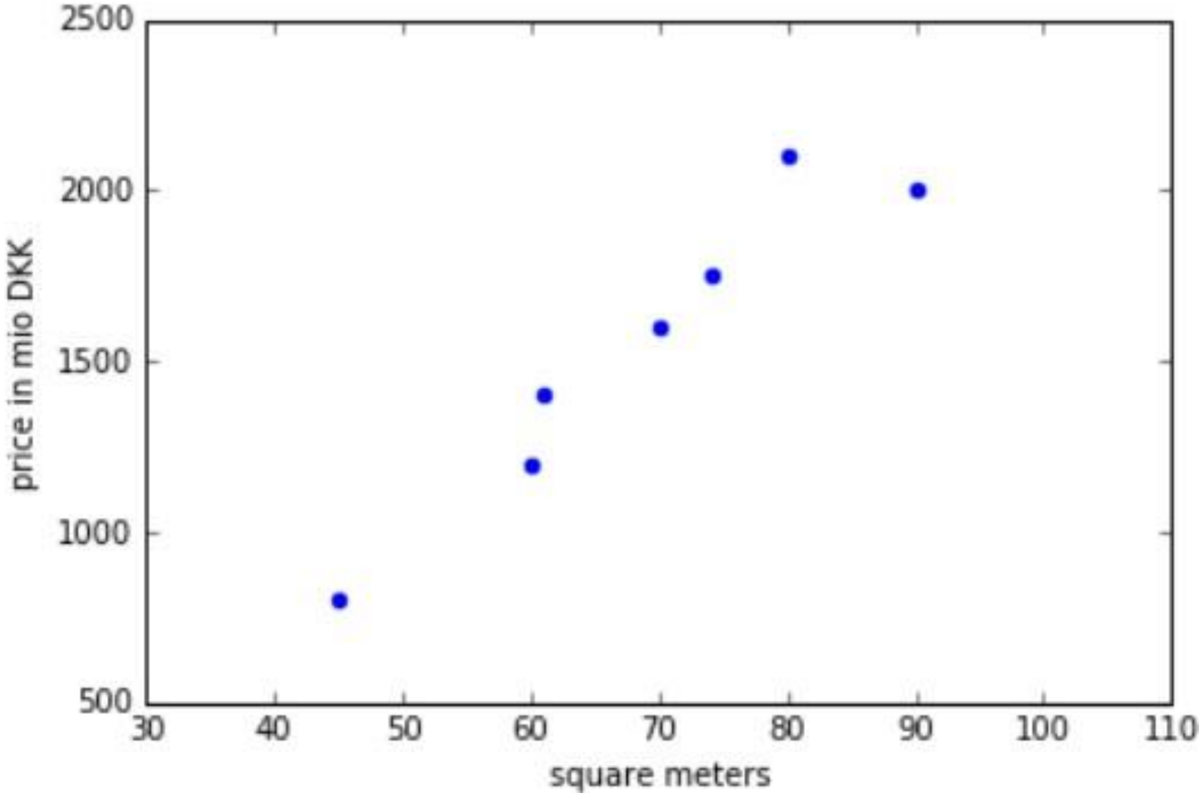
SSE stands for **S**um of **S**quared **E**rrors

Here we have the salary of someone with x years of experience.
The straight line represents where that person's salary should be according to our linear regression model, whereas the red point is what that person is actually earning.

Linear Regression: House price prediction

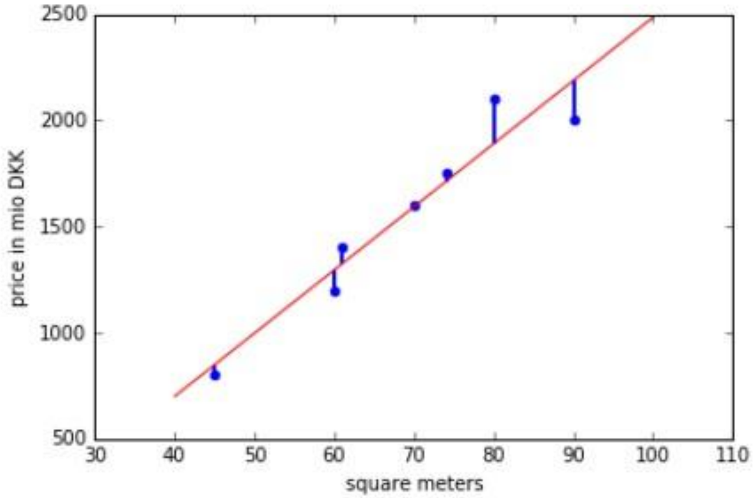
Example: House price prediction:

Size in m^2	Price in mio DKK
45	800
60	1200
61	1400
70	1600
74	1750
80	2100
90	2000



Linear Regression: House price prediction

Size in m^2	Price in mio DKK
45	800
60	1200
61	1400
70	1600
74	1750
80	2100
90	2000



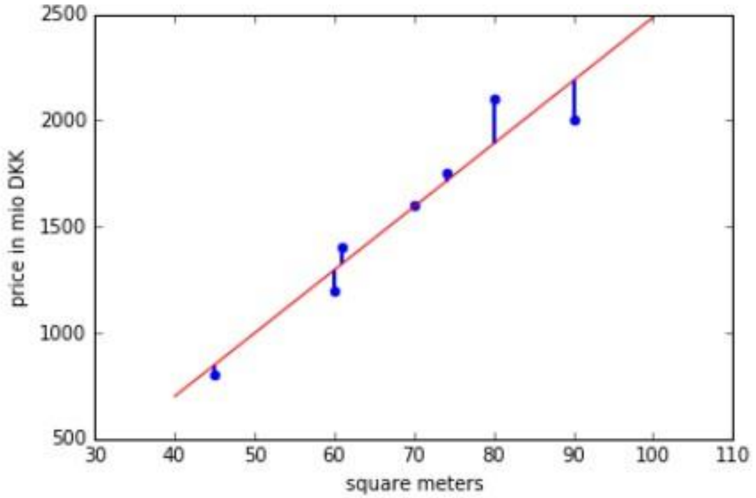
$$\begin{bmatrix}(x_1, y_1)\\(x_2, y_2)\\\vdots\\\vdots\\(x_m, y_m)\end{bmatrix} \rightsquigarrow \begin{bmatrix}(45, 800)\\(60, 1200)\\(61, 1400)\\(70, 1600)\\(74, 1750)\\(80, 2100)\\(90, 2000)\end{bmatrix}$$

$$f(x) = -489.76 + 29.75x$$

x	$\hat{y}$	y
45	848.83	800
60	1295.03	1200
61	1324.78	1400
70	1592.5	1600
74	1711.48	1750
80	1889.96	2100
90	2187.43	2000

Linear Regression: House price prediction

Size in m^2	Price in mio DKK
45	800
60	1200
61	1400
70	1600
74	1750
80	2100
90	2000



$$\begin{bmatrix} (x_1, y_1) \\ (x_2, y_2) \\ \vdots \\ \vdots \\ (x_m, y_m) \end{bmatrix} \rightsquigarrow \begin{bmatrix} (45, 800) \\ (60, 1200) \\ (61, 1400) \\ (70, 1600) \\ (74, 1750) \\ (80, 2100) \\ (90, 2000) \end{bmatrix}$$

$$f(x) = -489.76 + 29.75x$$

x	$\hat{y}$	y
45	848.83	800
60	1295.03	1200
61	1324.78	1400
70	1592.5	1600
74	1711.48	1750
80	1889.96	2100
90	2187.43	2000

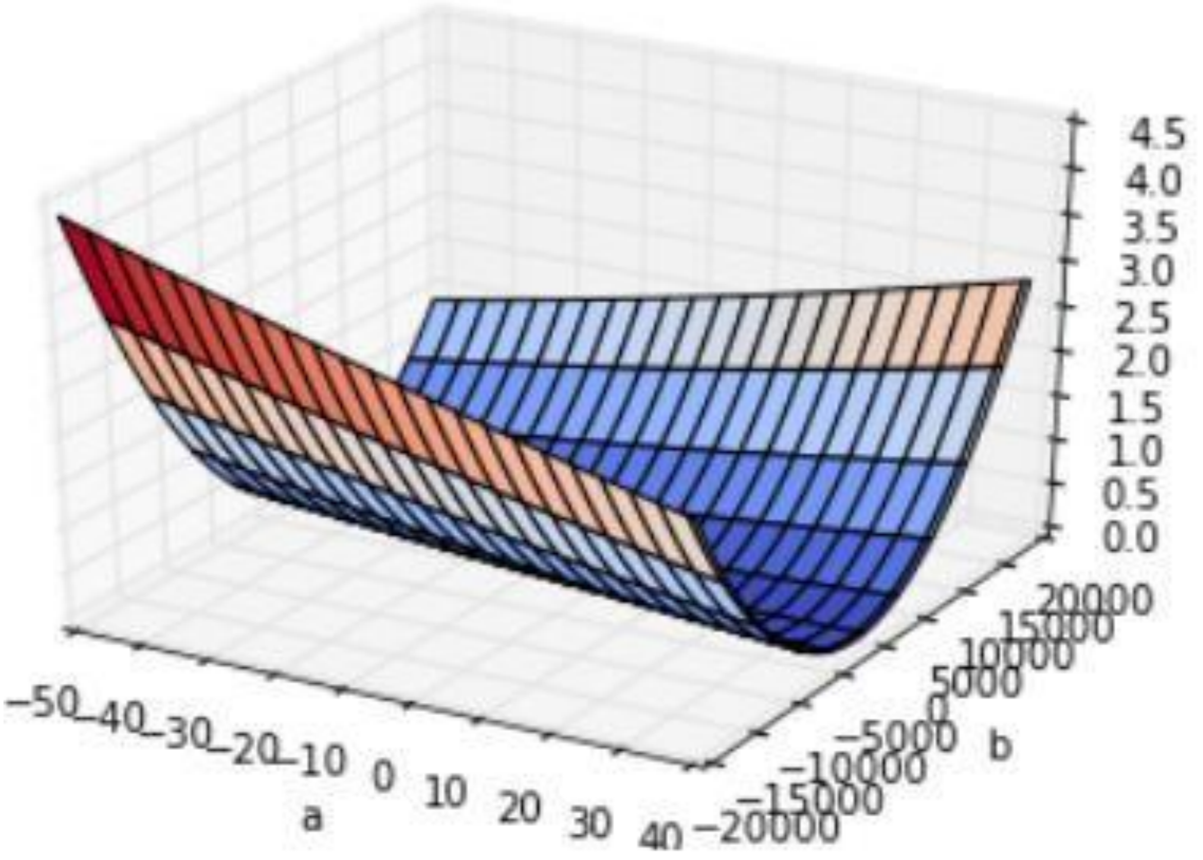
$$\begin{aligned} \text{SSE} &= \textstyle\sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \\ &= (800 - 848.83)^2 \\ &\quad +(1200 - 1295.03)^2 \\ &\quad +(1400 - 1324.78)^2 \\ &\quad +(1600 - 1592.5)^2 \\ &\quad +(1750 - 1711.48)^2 \\ &\quad +(2100 - 1889.96)^2 \\ &\quad +(2000 - 2187.43)^2 = 97858.86 \end{aligned}$$

Linear Regression: House price prediction

For

$$f(x) = b + ax$$

$$\begin{aligned}
\text{SSE} &= \textstyle\sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \\
&= (800 - b - 45 \cdot a)^2 \\
&\quad +(1200 - b - 60 \cdot a)^2 \\
&\quad +(1400 - b - 61 \cdot a)^2 \\
&\quad +(1600 - b - 70 \cdot a)^2 \\
&\quad +(1750 - b - 74 \cdot a)^2 \\
&\quad +(2100 - b - 80 \cdot a)^2 \\
&\quad +(2000 - b - 90 \cdot a)^2
\end{aligned}$$



Linear Regression: House price prediction

Theorem (Closed form solution)

The value of the coefficients of the line that minimizes the sum of squared errors for the given points can be expressed in closed form as a function of the input data:

$$a = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \qquad b = \bar{y} - a\bar{x}$$

where:

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

Linear Regression: House price prediction

Theorem (Closed form solution)

The value of the coefficients of the line that minimizes the sum of squared errors for the given points can be expressed in closed form as a function of the input data:

$$a = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \qquad b = \bar{y} - a\bar{x}$$

where:

Coding time: Code this in Python!

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

Warming Up

Theorem (Closed form solution)

The value of the coefficients of the line that minimizes the sum of squared errors for the given points can be expressed in closed form as a function of the input data:

$$a = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \qquad b = \bar{y} - a\bar{x}$$

where:

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

X	Y
10	80
30	40
15	70
55	-10

Warming up: What's the coefficient a and the constant b for the data above?

Linear Regression: Exercise

Theorem (Closed form solution)

The value of the coefficients of the line that minimizes the sum of squared errors for the given points can be expressed in closed form as a function of the input data:

$$a = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \qquad b = \bar{y} - a\bar{x}$$

where:

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

Example: House price prediction:

Size in m^2	Price in mio DKK
45	800
60	1200
61	1400
70	1600
74	1750
80	2100
90	2000

Exercise: For the given house price data, find the best fitting-line linear equation using the closed form solution!
(You may use your spreadsheet application)

Multiple linear regression

Simple Linear Regression

$$y = b_0 + b_1{*}x_1$$

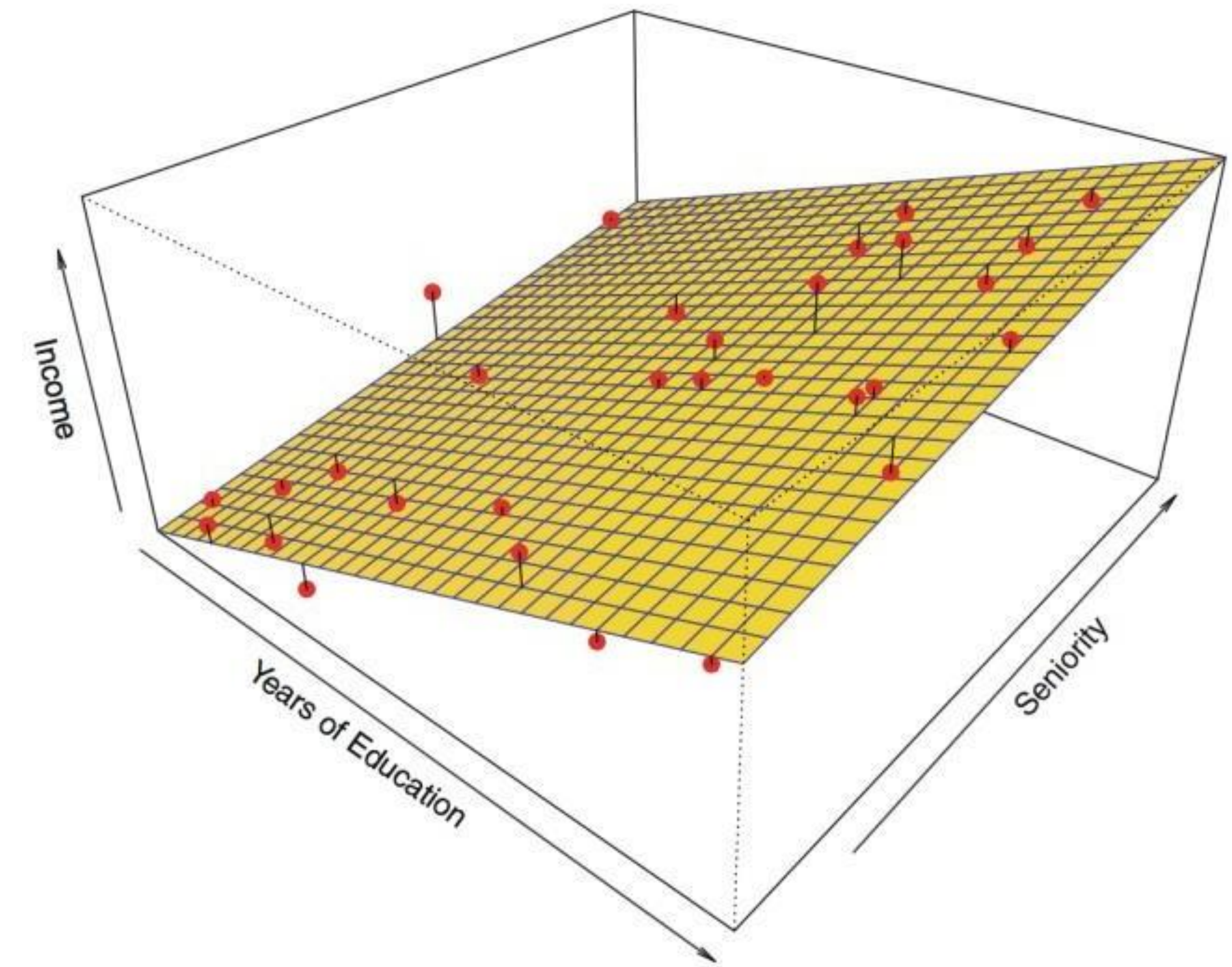
Multiple Linear Regression

$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + \dots + b_n{*}x_n$$

We use the multiple linear regression model when we're dealing with a dependent variable that is affected by more than one factor.

For example, a person's salary can be affected by their years of experience, years of education, daily working hours, etc.

Multiple linear regression



Multiple Linear Regression: DIY

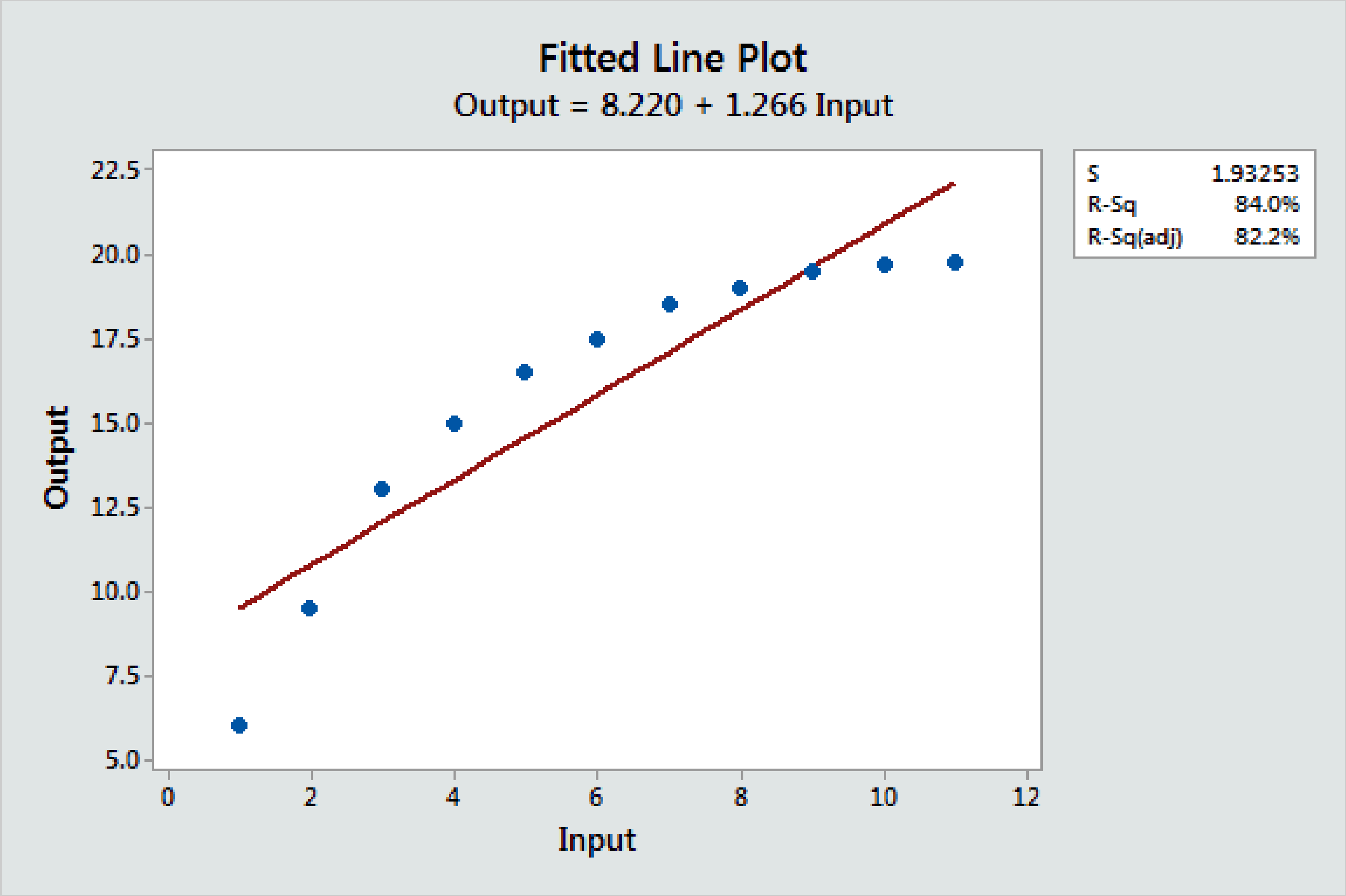
Play around with: http://al-roomi.org/3DPlot/index.html

How does the visualization look like with:

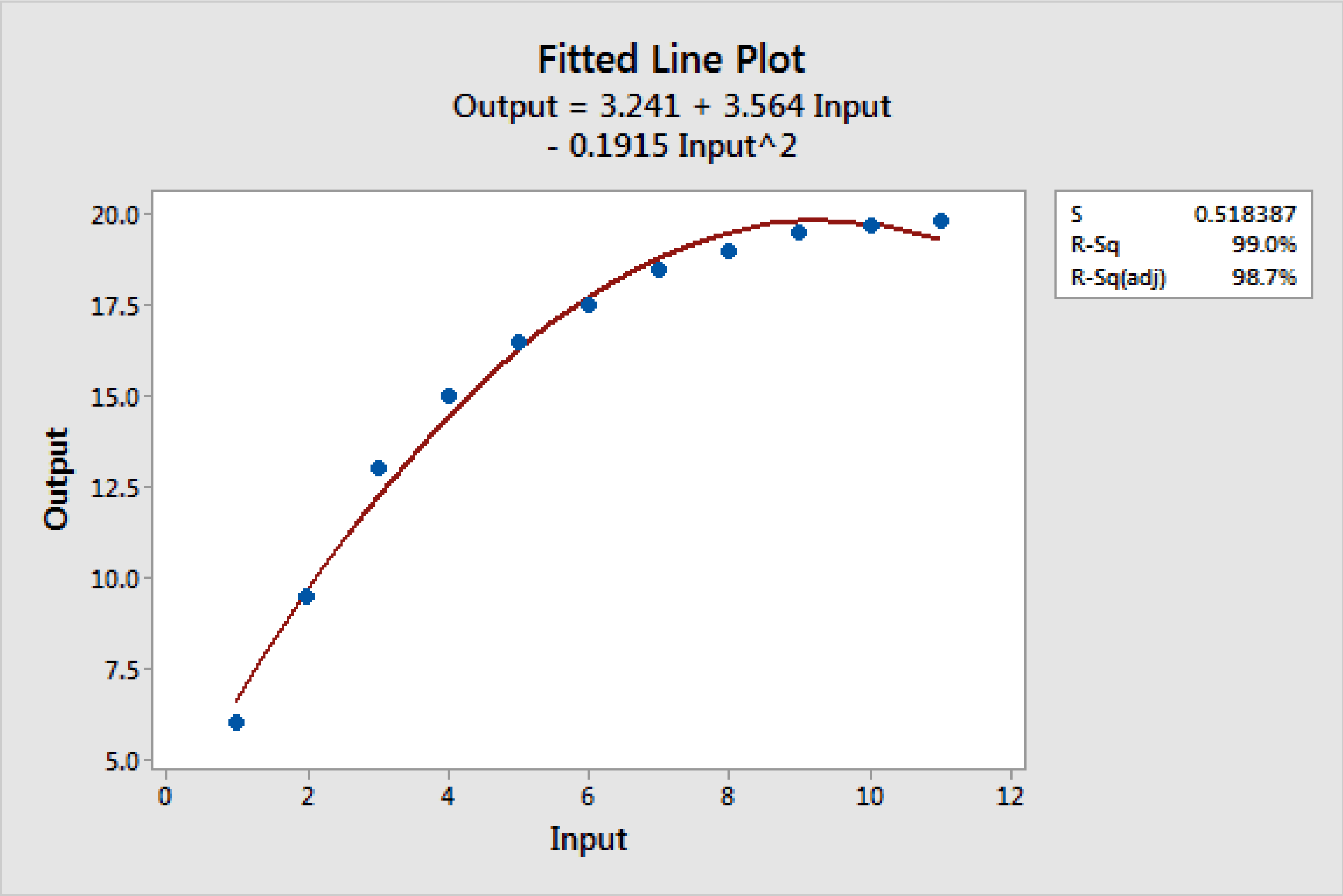
- x +y
- x +y +100
- 5*x +y - 10
- 20*x +(-100)*y +20
- x - 10

PS: You may rotate the generated plane and also click on specific points.

Nonlinear Regression

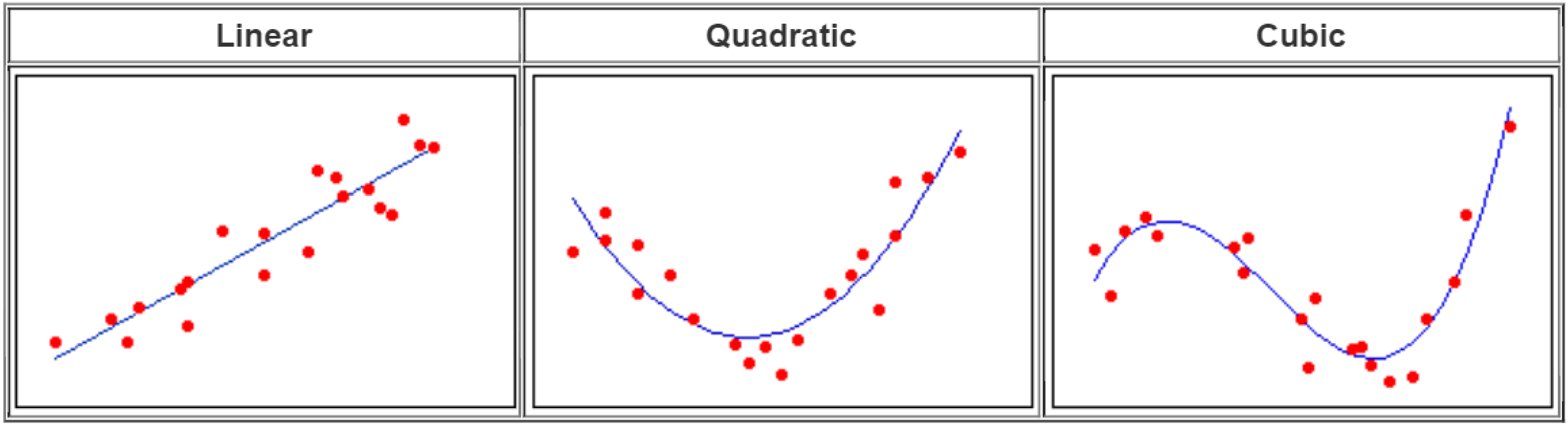


Nonlinear Regression

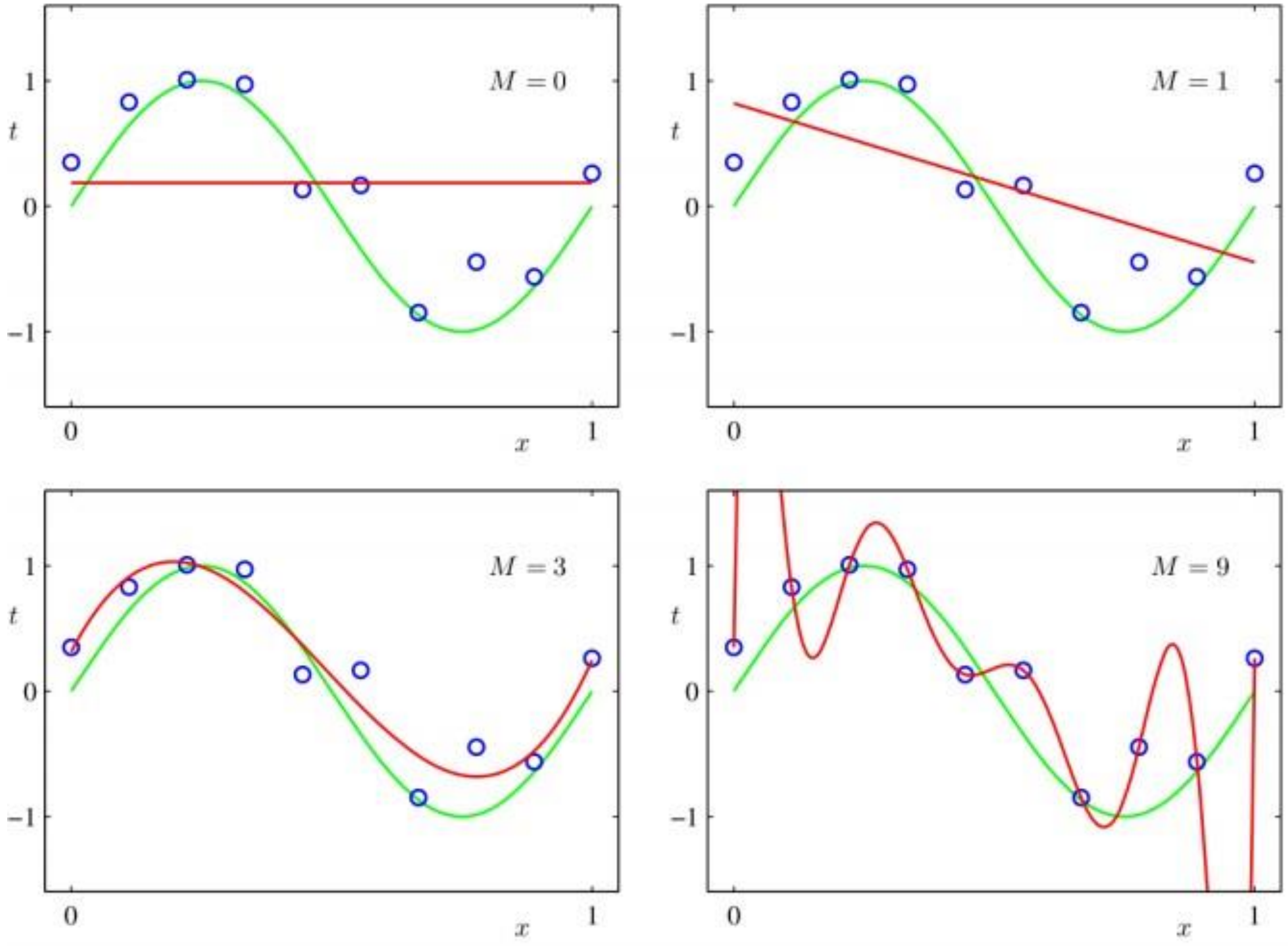


Nonlinear Regression

Model Order

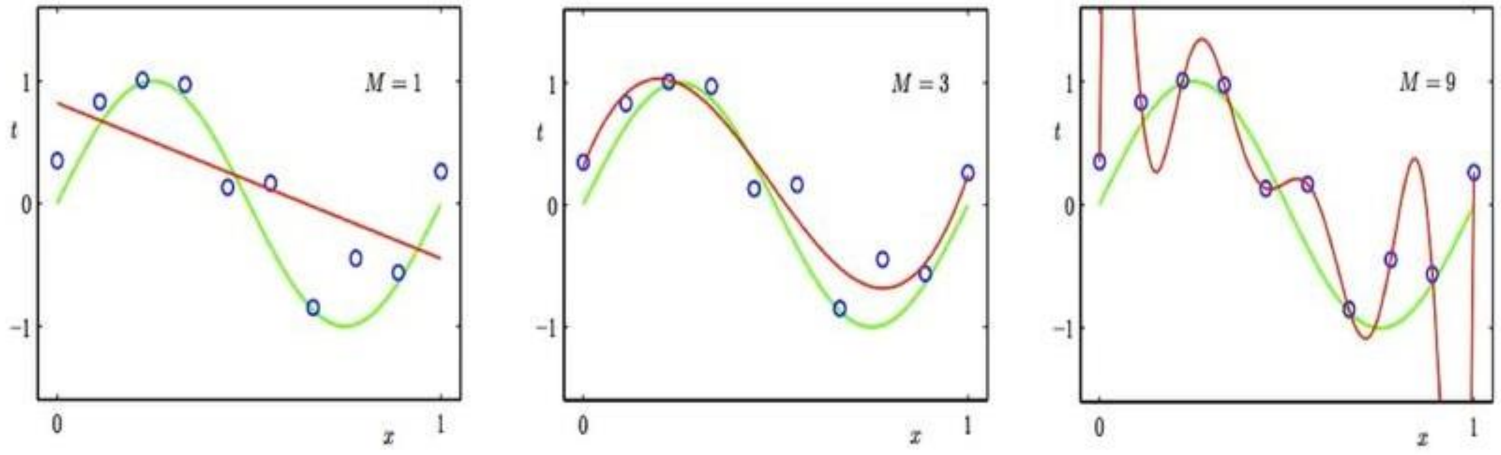


Polynomial regression



Underfitting and overfitting

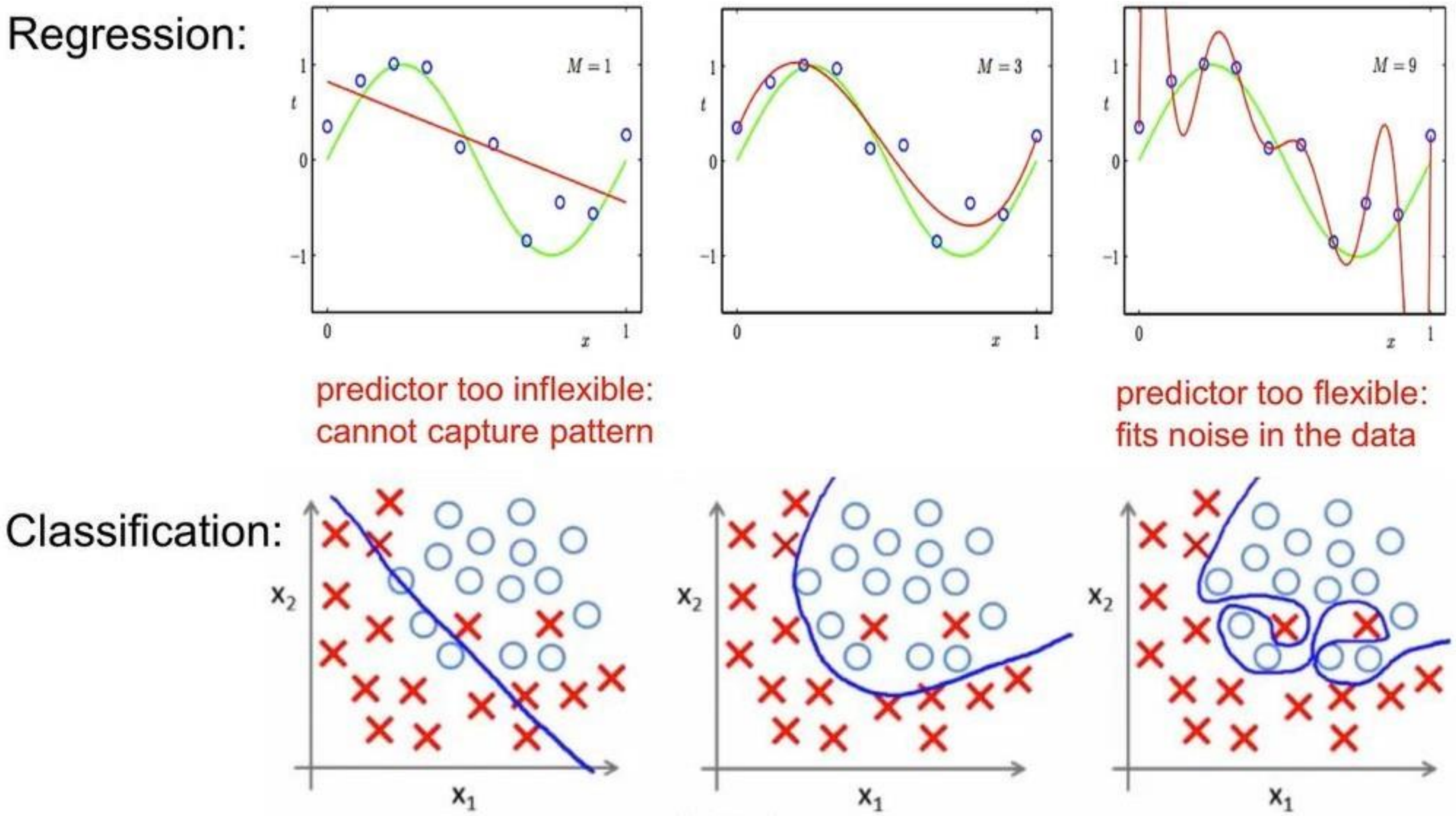
Regression:



predictor too inflexible:
cannot capture pattern

predictor too flexible:
fits noise in the data

Underfitting and overfitting



Nonlinear Regression: DIY

Play around with: https://www.wolframalpha.com/

How does the visualization look like with:

- y = x
- y = x^2
- y = x^2 – 5
- y = x^2 - 10x
- y = x^3 - x
- y = -x^4 + 2x^3

Bonus: x^2+(y-(x^2)^(1/3))^2 = 1

Evaluation: Error rates

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Similarities: Both MAE and RMSE express average model prediction error in units of the variable of interest. Both metrics can range from 0 to ∞. They are negatively-oriented scores, which means lower values are better.

Evaluation: Error rates

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Differences: Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable.

Evaluation: Error rates

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

CASE 1: Evenly distributed errors

ID	Error	|Error|	Error^2
1	2	2	4
2	2	2	4
3	2	2	4
4	2	2	4
5	2	2	4
6	2	2	4
7	2	2	4
8	2	2	4
9	2	2	4
10	2	2	4

MAE	RMSE
2.000	2.000

Evaluation: Error rates

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

CASE 1: Evenly distributed errors

ID	Error	|Error|	Error^2
1	2	2	4
2	2	2	4
3	2	2	4
4	2	2	4
5	2	2	4
6	2	2	4
7	2	2	4
8	2	2	4
9	2	2	4
10	2	2	4

MAE	RMSE
2.000	2.000

CASE 2: Small variance in errors

ID	Error	|Error|	Error^2
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	3	3	9
7	3	3	9
8	3	3	9
9	3	3	9
10	3	3	9

MAE	RMSE
2.000	2.236

Evaluation: Error rates

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

CASE 1: Evenly distributed errors

ID	Error	|Error|	Error^2
1	2	2	4
2	2	2	4
3	2	2	4
4	2	2	4
5	2	2	4
6	2	2	4
7	2	2	4
8	2	2	4
9	2	2	4
10	2	2	4

MAE	RMSE
2.000	2.000

CASE 2: Small variance in errors

ID	Error	|Error|	Error^2
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	3	3	9
7	3	3	9
8	3	3	9
9	3	3	9
10	3	3	9

MAE	RMSE
2.000	2.236

CASE 3: Large error outlier

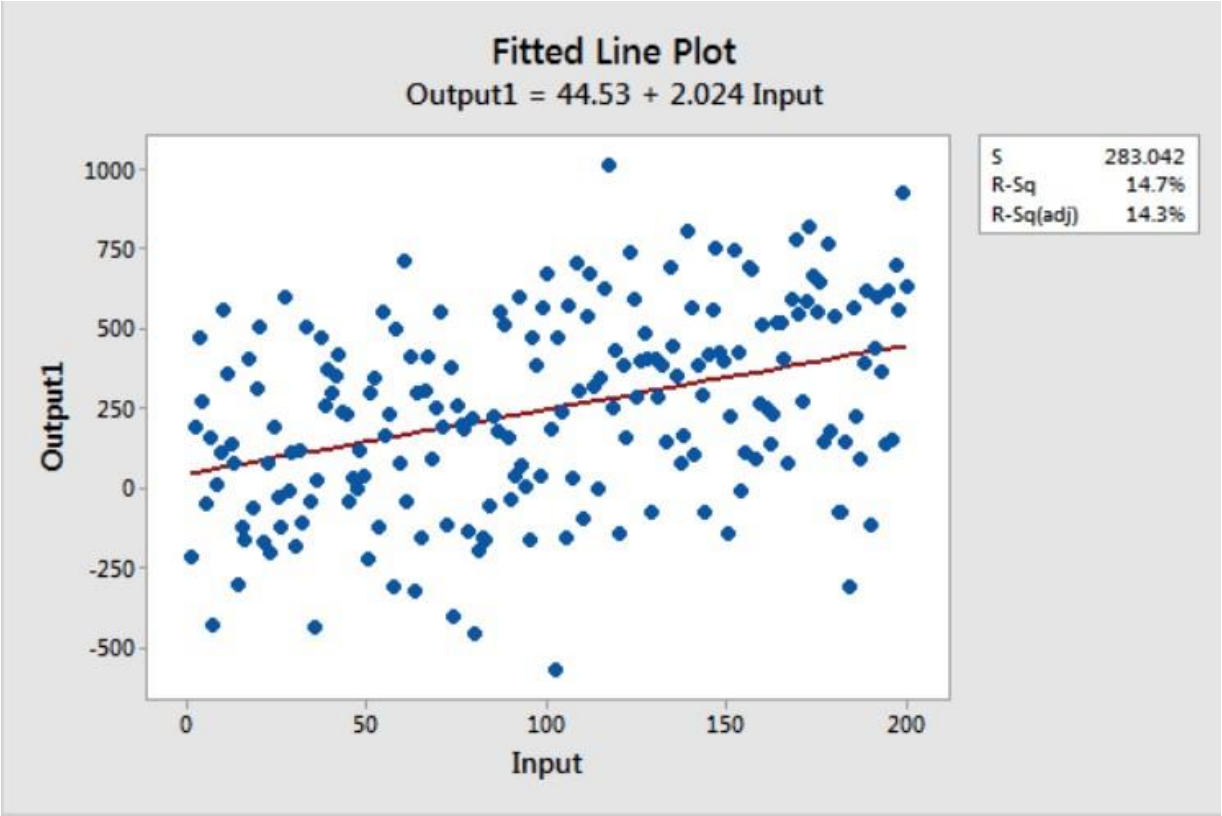
ID	Error	|Error|	Error^2
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	20	20	400

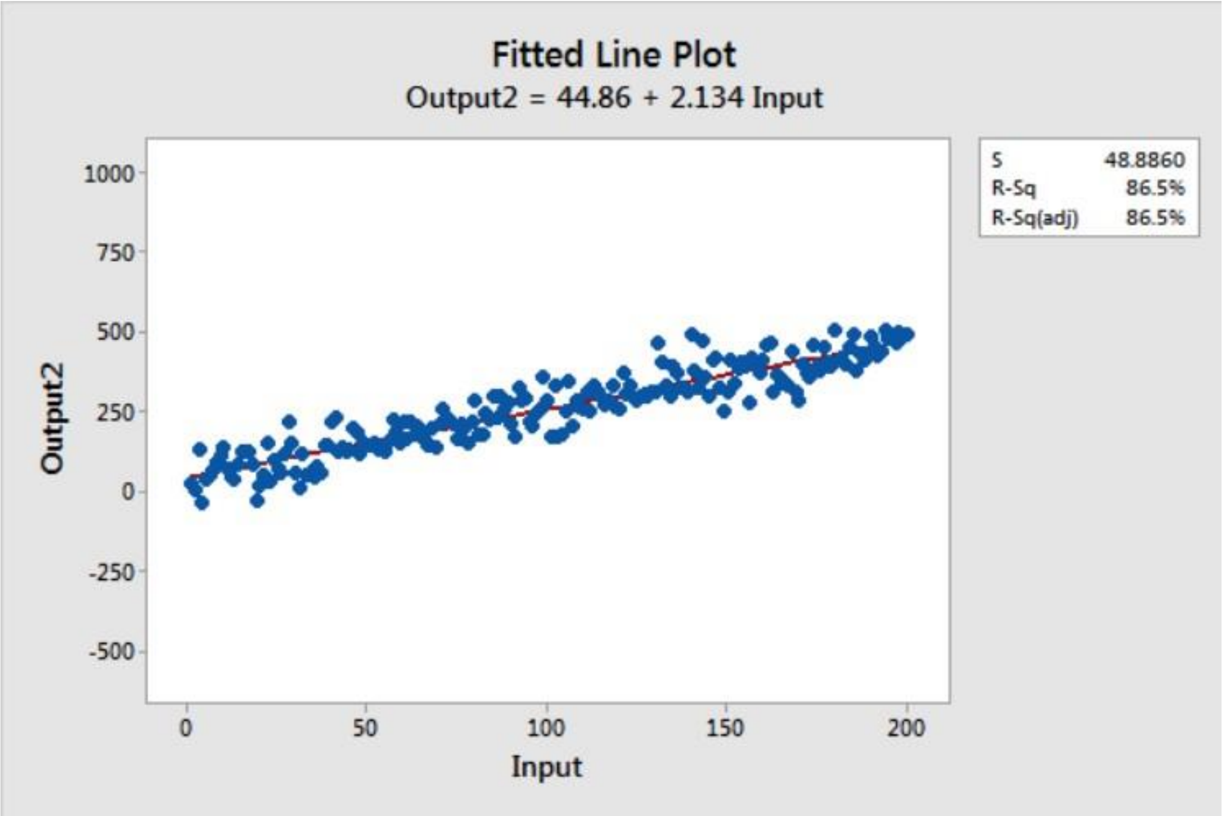
MAE	RMSE
2.000	6.325

Evaluation: R^2

- R^2 (called R-Squared) is a metric to assess regression performance.
- It is also known as coefficient of determination.
- Generally, the value ranges between 0 and 1.
- The closer R^2 is to 1, the better our model will be at predicting our dependent variable.

Evaluation: R^2






THANK YOU