Probability and Statistics for Business and Data

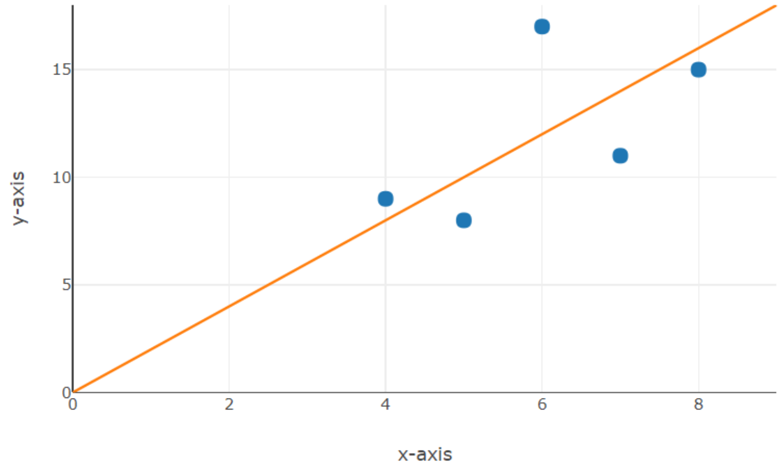
PART 6 - REGRESSION

PIERIAN DATA

Linear Regression

Linear Regression

- The goal of **regression** is to develop an equation or formula that **best describes** the relationship between variables.

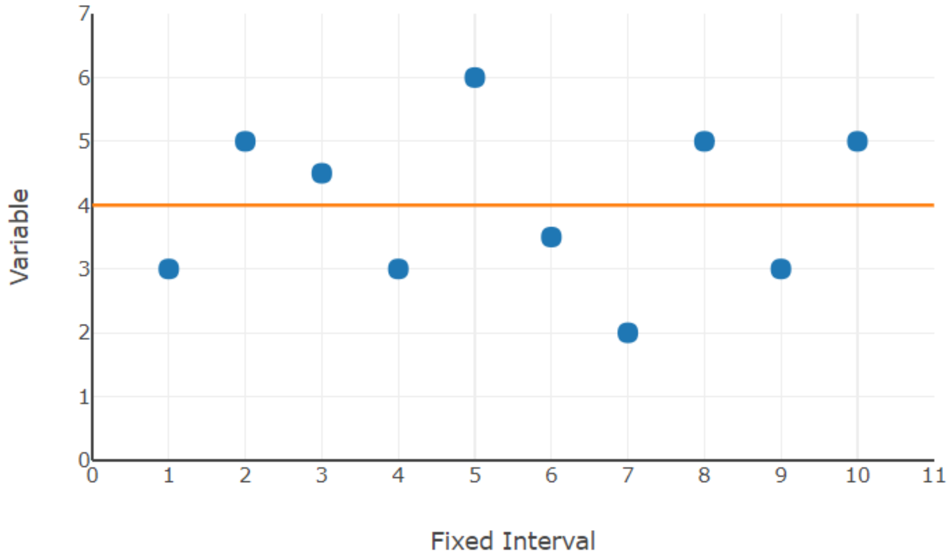


$y = 2x$

PIERIAN DATA

Linear Regression

- How do we find a best-fit line?
- Consider a dataset with only one variable
- The best-fit line is just the mean value of the data points

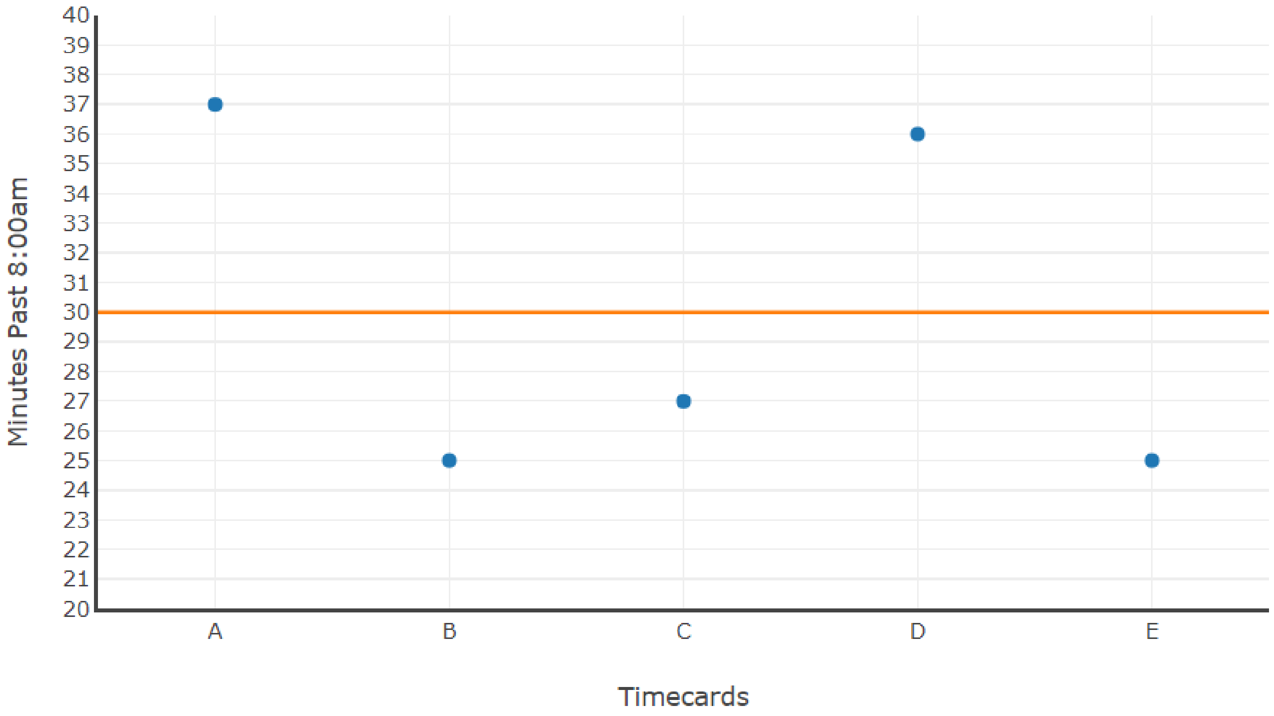


Understanding Best Fit

- A plant manager wants to know when employees arrive at work
- The shift starts a 8:30am
- She takes five random timecards and plots the minutes of arrival on a chart

Understanding Best Fit

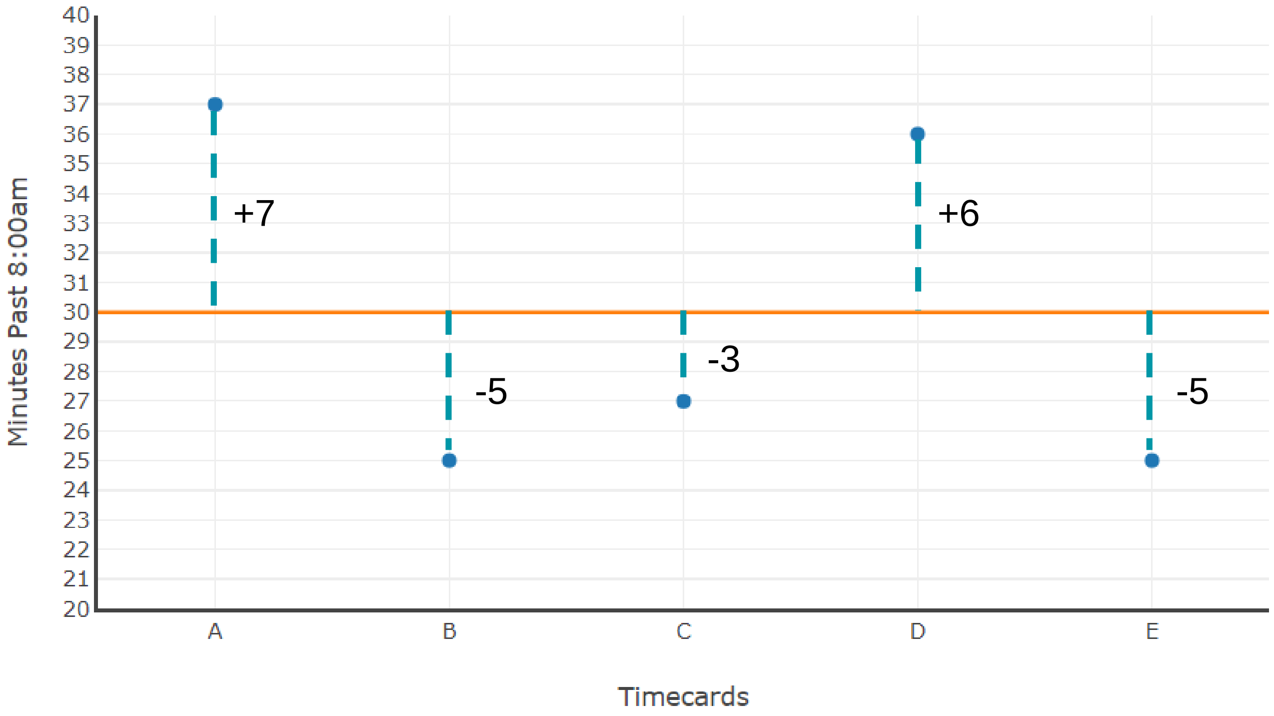
Timecard	Minutes past 8:00am
A	37
B	25
C	27
D	36
E	25
Total:	**150**
Mean	**30**



Understanding Best Fit

What makes $y = 30$ a best-fit line?

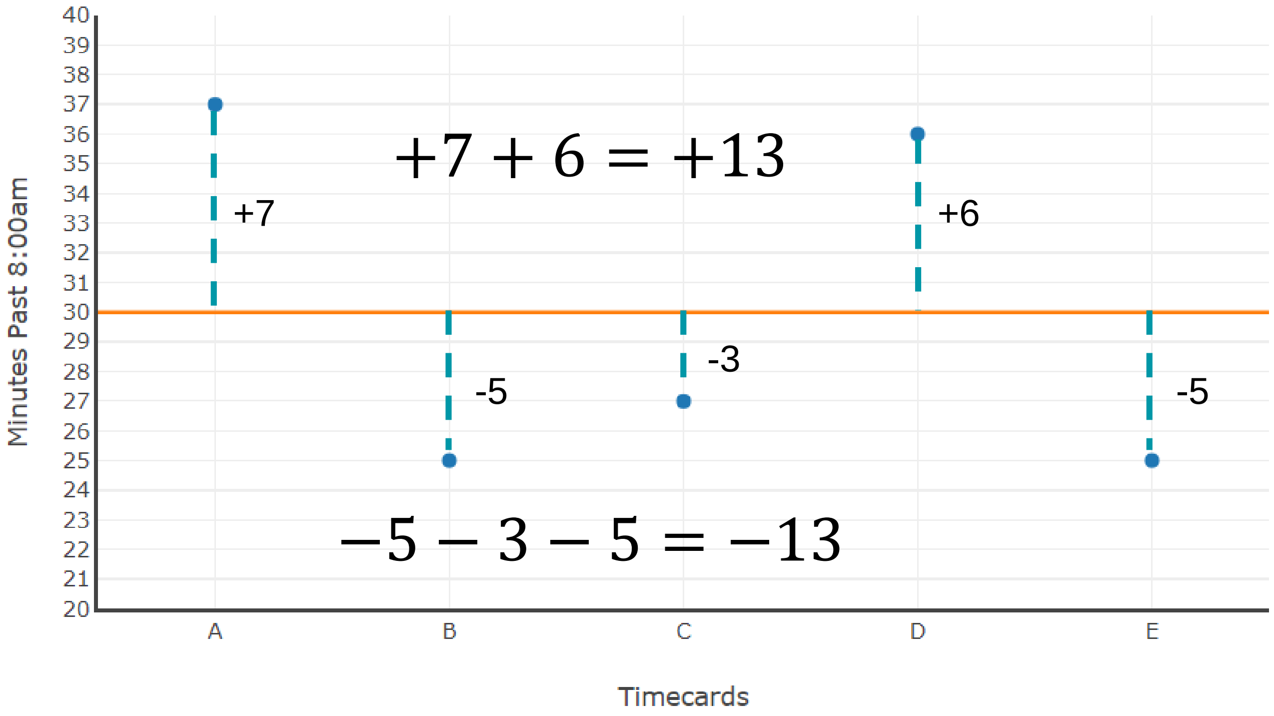
Consider the error



PIERIAN DATA

Understanding Best Fit

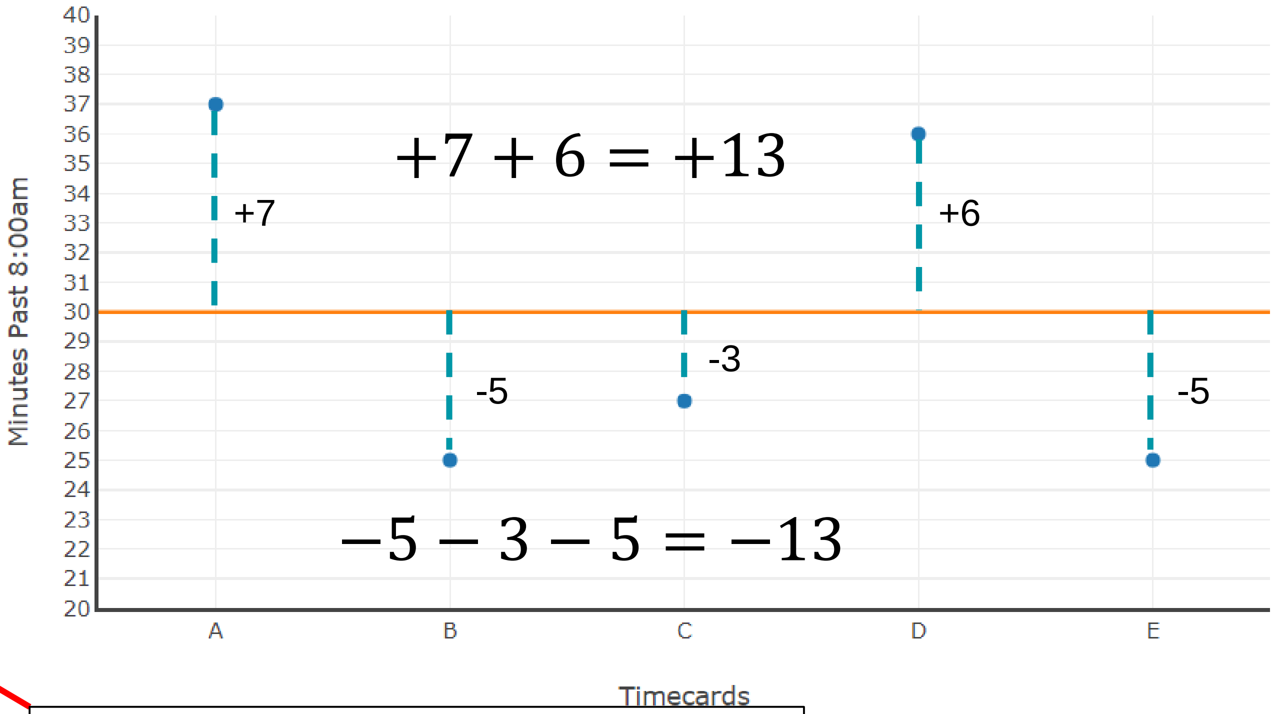
See that the sum of the distances above the line balances the sum of those below the line



PIERIAN DATA

Understanding Best Fit

Error (E)	Square Error (SE)
+7	49
-5	25
-3	9
+6	36
-5	25
Sum of Squares Error (SSE)	**144**



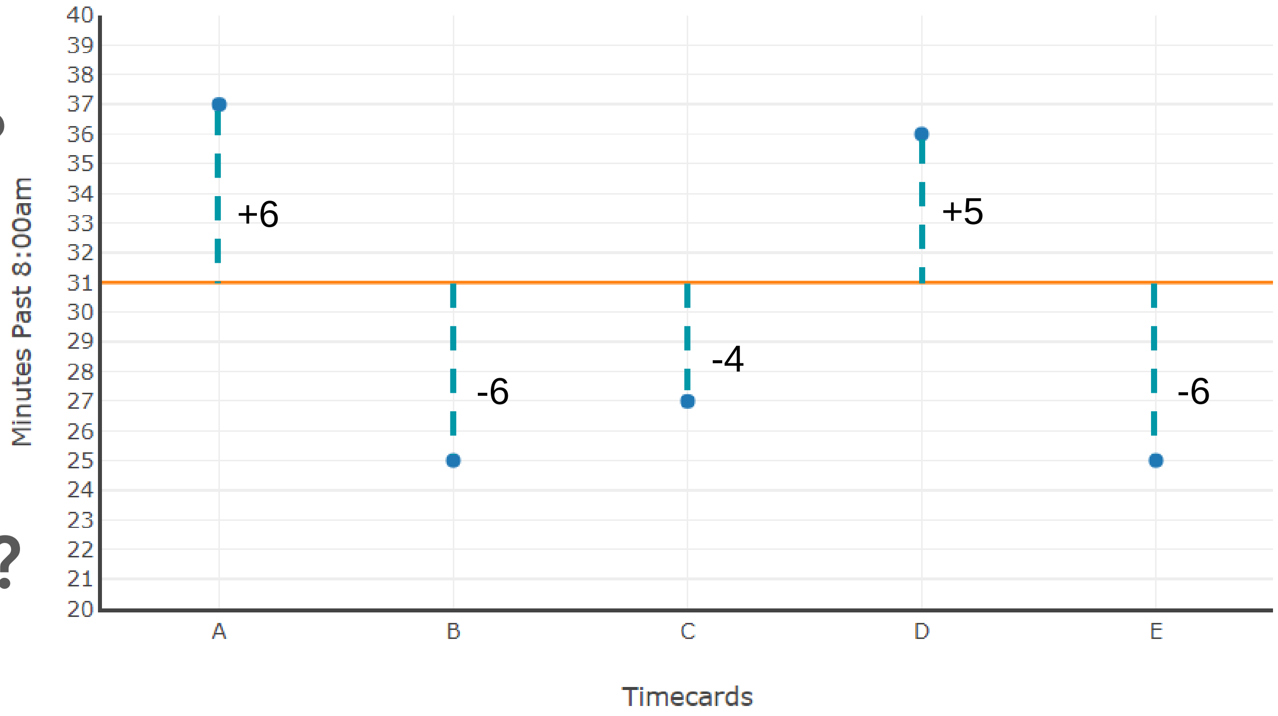
we want to MINIMIZE the SSE

Understanding Best Fit

What if we move the line?

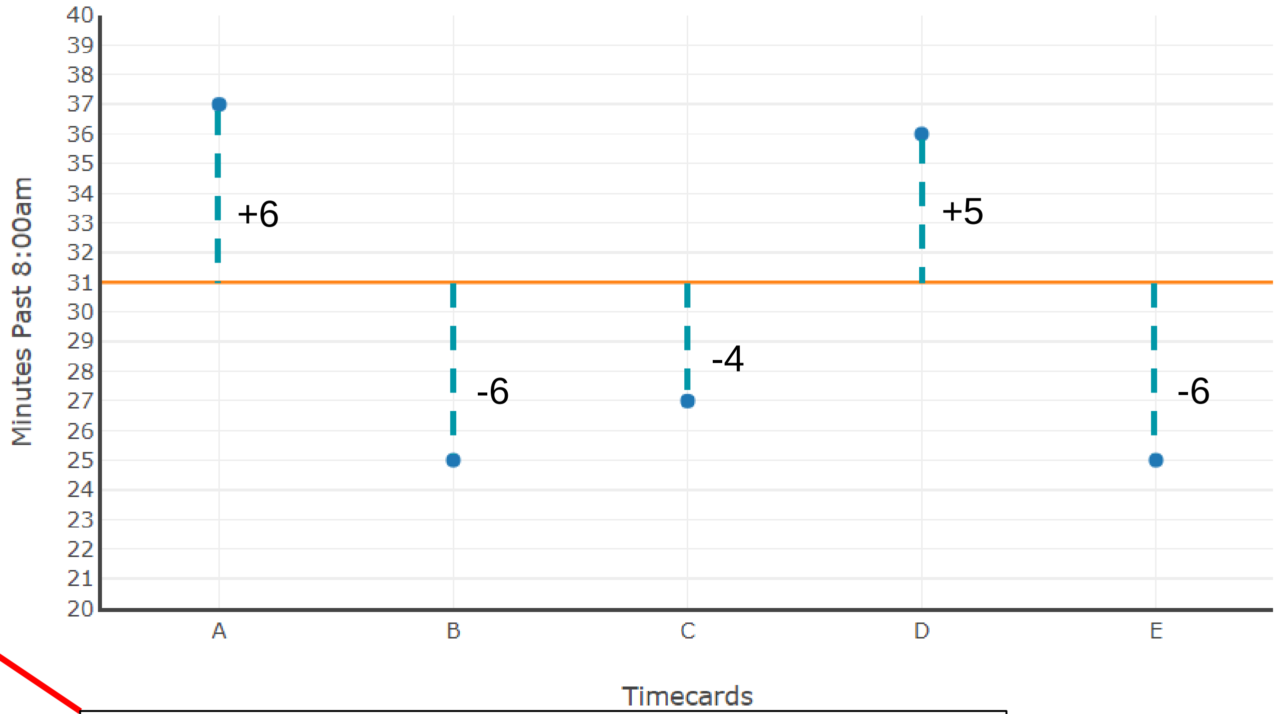
Let's set it to $y = 31$ instead

How does it affect the SSE?



Understanding Best Fit

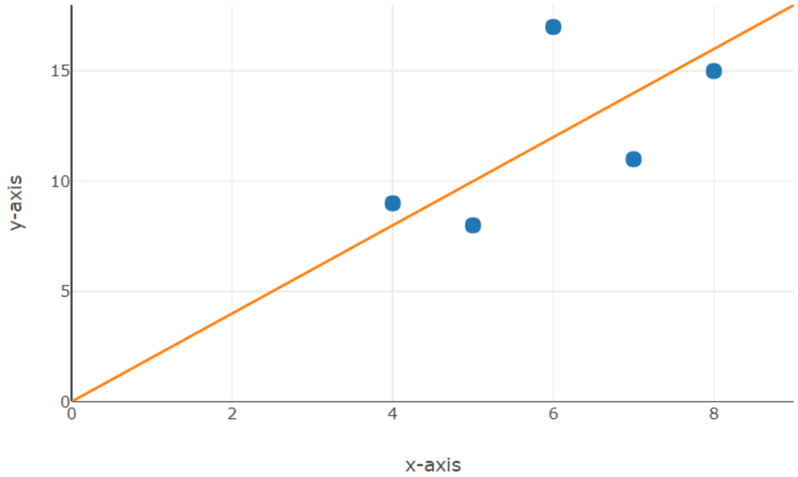
Error (E)		Square Error (SE)	
+7	+6	49	36
-5	-6	25	36
-3	-4	9	16
+6	+5	36	25
-5	-6	25	36
Sum of Squares Error (SSE)		144	149



moving the line INCREASED the SSE

Linear Regression

- That's it! The goal of regression is to find the line that best describes our data.
- Fortunately, we don't have to rely on trial-and-error.
- We have algebra!

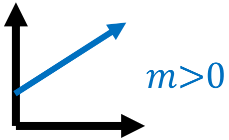


Linear Regression

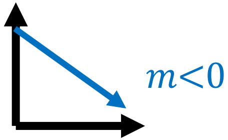
- Recall that the equation of a line follows the form $y = mx + b$ where

  m is the **slope** of the line, and

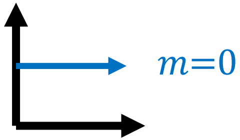
  b is where the line crosses the y-axis when x=0 (b is the **y-intercept**)



positive slope



negative slope



zero slope

PIERIAN DATA

Linear Regression

- In a linear regression, where we try to formulate the relationship between variables, $y = mx + b$ becomes

$$\hat{y} = b_0 + b_1 x$$

- Our goal is to predict the value of a **dependent variable** (y) based on that of an **independent variable** (x).

Linear Regression

$$\hat{y} = b_0 + b_1 x$$

- How to derive b_1 and b_0:

$$b_1 = \rho_{x,y} \frac{\sigma_y}{\sigma_x}$$

$\rho_{x,y} = Pearson\ Correlation\ Coefficient$
$\sigma_x, \sigma_y = Standard\ Deviations$

$$= \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} \cdot \frac{\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$

PIERIAN DATA

Linear Regression

$$\hat{y} = b_0 + b_1 x$$

- How to derive b_1 and b_0:

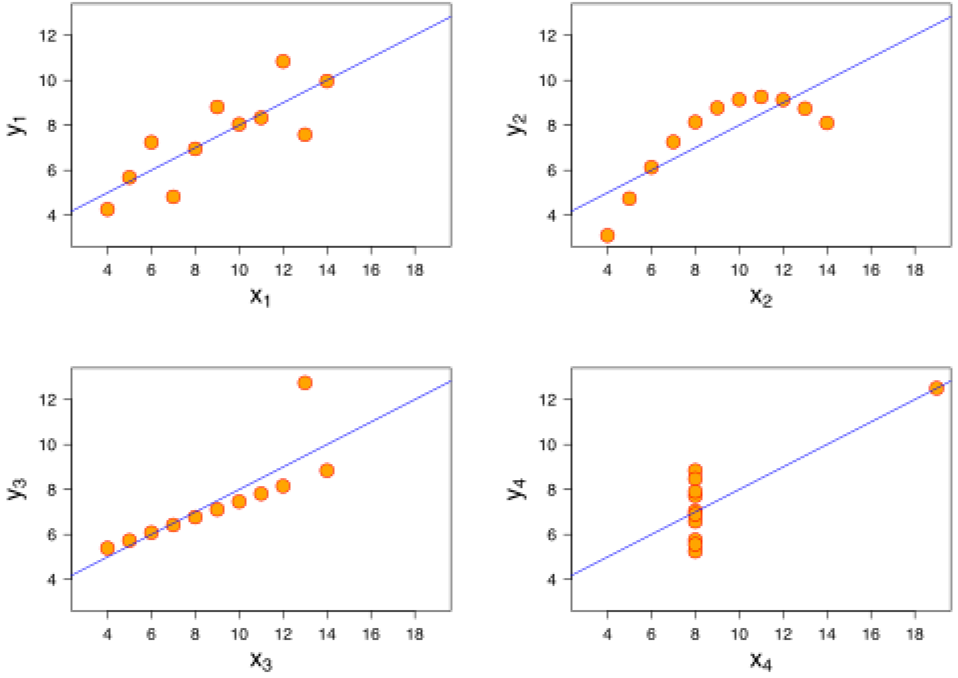
$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

Limitations of Linear Regression

Anscombe's Quartet illustrates the pitfalls of relying on pure calculation.

Each graph results in the same calculated regression line.



Regression Exercise #1

- A manager wants to find the relationship between the number of hours that a plant is operational in a week and weekly production.

Regression Exercise #1

- Here the **independent variable** x is hours of operation, and the **dependent variable** y is production volume.

PIERIAN DATA

Regression Exercise #1

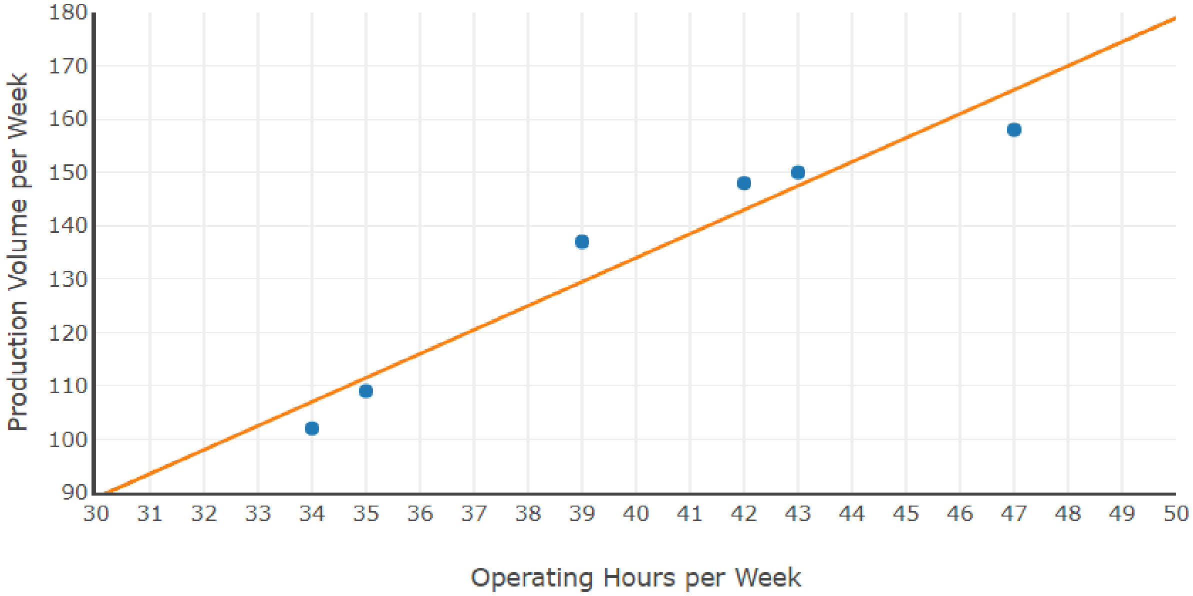
- The manager develops the following table:

Production Hours (x)	Production Volume (y)
34	102
35	109
39	137
42	148
43	150
47	158

Regression Exercise #1

- First, plot the data   Is there a linear pattern?

Production Hours (x)	Production Volume (y)
34	102
35	109
39	137
42	148
43	150
47	158



Regression Exercise #1

$$\hat{y} = b_0 + b_1 x$$

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

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

- Run calculations:

	Production Hours (x)	Production Volume (y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
	34	102	-6	-32	192	36
	35	109	-5	-25	125	25
	39	137	-1	3	-3	1
	42	148	2	14	28	4
	43	150	3	16	48	9
	47	158	7	24	168	49
$\bar{x}, \bar{y}$	**40**	**134**		**Sum:**	**558**	**124**
					$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma(x-\bar{x})^2$

Regression Exercise #1

$$\hat{y} = b_0 + b_1 x$$

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

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

- Run calculations:

	Production Hours (x)	Production Volume (y)
	34	102
	35	109
	39	137
	42	148
	43	150
	47	158
$\bar{x}, \bar{y}$	**40**	**134**

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{558}{124} = \mathbf{4.5}$$

$$b_0 = \bar{y} - b_1\bar{x} = 134 - (4.5 \times 40) = \mathbf{-46}$$

$$\boldsymbol{\hat{y} = -46 + 4.5x}$$

Sum:	**558**	**124**
	$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma(x-\bar{x})^2$

PIERIAN DATA

Regression Exercise #1

Based on the formula, if the manager wants to produce 125 units per week, the plant should run for:

Production Hours (x)	Production Volume (y)
34	102
35	109
39	137
42	148
43	150
47	158

$$\hat{y} = b_0 + b_1 x$$

$$125 = -46 + 4.5x$$

$$x = \frac{171}{4.5} = \mathbf{38}\ hours\ per\ week$$

REGRESSION with Excel Data Analysis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.966875047
R Square	0.934847357
Adjusted R Square	0.918559196
Standard Error	6.614378278
Observations	6

ANOVA

	df	SS	MS	F	Significance F
Regression	1	2511	2511	57.39428571	0.00162772
Residual	4	175	43.75		
Total	5	2686			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-46	23.91250292	-1.923679849	0.126733563	-112.3917517	20.39175167	-112.3917517	20.39175167
X Variable 1	4.5	0.593988704	7.575901644	0.00162772	2.85082297	6.14917703	2.85082297	6.14917703

Data Analysis

Analysis Tools

- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression

OK

Cancel

Help

LINEAR REGRESSION with Python

```
>>> from scipy.stats import linregress
>>> x = [34, 35, 39, 42, 43, 47]
>>> y = [102, 109, 137, 148, 150, 158]
>>> slope = round(linregress(x,y).slope,1)
>>> intercept = round(linregress(x,y).intercept,1)
>>> print(f'y = {intercept} + {slope}x')
y = -46.0 + 4.5x
```

Multiple Regression

Linear vs Multiple Regression

- In linear regression we have one independent variable that may relate to a dependent variable with the formula

$$\hat{y} = b_0 + b_1 x$$

Linear vs Multiple Regression

- Multiple regression lets us compare several independent variables to one dependent variable at the same time.
- Each independent variable is assigned a subscript: x_1, x_2, x_3 etc.

Linear vs Multiple Regression

- The general formula is expanded:

linear regression

$$\hat{y} = b_0 + b_1 x$$

multiple regression

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots$$

- b_1 is the coefficient on x_1
- b_1 reflects the change in $\hat{y}$ for a given change in x_1, all else remaining constant

Linear vs Multiple Regression

- The formulas for coefficients also expand:

multiple regression

$$b_1 = \frac{\sum(x_2-\overline{x_2})^2 \sum(x_1-\overline{x_1})(y-\bar{y}) - \sum(x_1-\overline{x_1})(x_2-\overline{x_2}) \sum(x_2-\overline{x_2})(y-\bar{y})}{\sum(x_1-\overline{x_1})^2 \sum(x_2-\overline{x_2})^2 - (\sum(x_1-\overline{x_1})(x_2-\overline{x_2}))^2}$$

$$b_2 = \frac{\sum(x_1-\overline{x_1})^2 \sum(x_2-\overline{x_2})(y-\bar{y}) - \sum(x_1-\overline{x_1})(x_2-\overline{x_2}) \sum(x_1-\overline{x_1})(y-\bar{y})}{\sum(x_1-\overline{x_1})^2 \sum(x_2-\overline{x_2})^2 - (\sum(x_1-\overline{x_1})(x_2-\overline{x_2}))^2}$$

$$b_0 = \bar{y} - b_1\overline{x_1} - b_2\overline{x_2}$$

Multiple Regression

- For example, a used car lot may want to know what variables affect net profits
- They would create a list of predictors that might correlate with profit:

age brand
price
color style

Multiple Regression

- They would want to measure the correlation of each variable to net profit
- However, some predictors might correlate with **each other**:

price age brand color style

Multiple Regression

- The age of a car would have a direct impact on its sales price
- You can't adjust one without affecting the other
- This is called **multicollinearity**

price age

brand

color

style

Regression Exercise #2

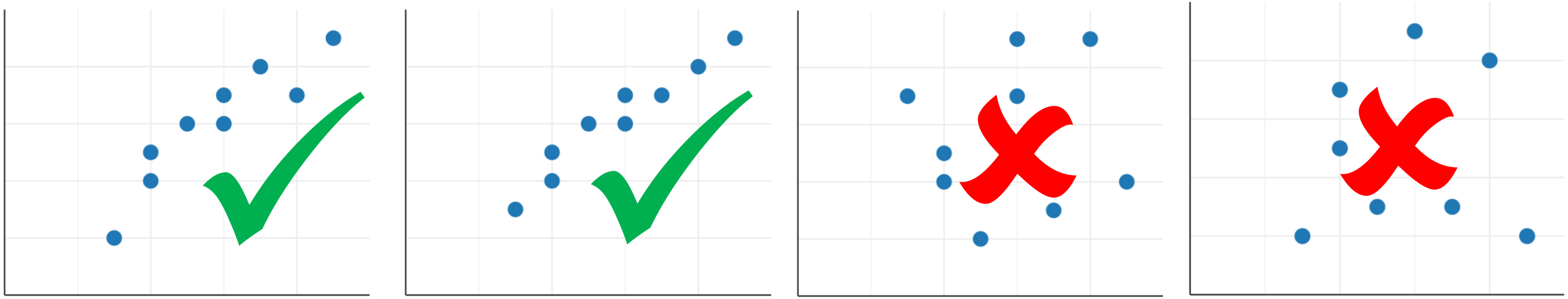
- A pharmacy delivers medications to the surrounding community.
- Drivers can make several stops per delivery.
- The owner would like to predict the **length of time** a delivery will take based on one or two related variables.

Regression Exercise #2

- First, consider what variables may have an effect on delivery time:
  - number of stops
  - driving distance
  - outside temperature
  - gasoline prices

Regression Exercise #2

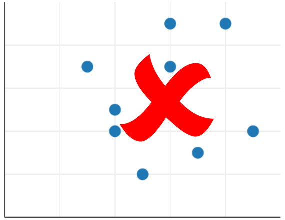
- Next, plot each variable against delivery time to see if there may be a relationship



Time vs Distance | Time vs Stops | Time vs Temperature | Time vs Gas Price

Regression Exercise #2

- Once we've chosen our variables x_1 and x_2, we'll usually test for **multicollinearity**
- We want to know if our two independent variables are closely related to each other
- If they are, it makes sense to discard one!



Stops vs Distance

A delivery might go to one customer that lives far away, or to a group of stops close by

PIERIAN DATA

Regression Exercise #2

$$y = Delivery\ Time\ (minutes)$$
$$x_1 = Number\ of\ Stops$$
$$x_2 = Distance\ (miles)$$

y	x_1	x_2	$(y-\bar{y})$	$(x_1-\bar{x}_1)$	$(x_1-\bar{x}_1)^2$	$(x_2-\bar{x}_2)$	$(x_2-\bar{x}_2)^2$	$(x_1-\bar{x}_1)(y-\bar{y})$	$(x_2-\bar{x}_2)(y-\bar{y})$	$(x_1-\bar{x}_1)(x_2-\bar{x}_2)$
29	**1**	**8**	-1	-1	1	2	4	1	-2	-2
31	**3**	**4**	1	1	1	-2	4	1	-2	-2
36	**2**	**9**	6	0	0	3	9	0	18	0
35	**3**	**6**	5	1	1	0	0	5	0	0
19	**1**	**3**	-11	-1	1	-3	9	11	33	3
$\bar{y}$	$\bar{x}_1$	$\bar{x}_2$			$\Sigma(x_1-\bar{x}_1)^2$		$\Sigma(x_2-\bar{x}_2)^2$	$\Sigma(x_1-\bar{x}_1)(y-\bar{y})$	$\Sigma(x_2-\bar{x}_2)(y-\bar{y})$	$\Sigma(x_1-\bar{x}_1)(x_2-\bar{x}_2)$
30	**2**	**6**			**4**		**26**	**18**	**47**	**-1**

Regression Exercise #2

y = *Delivery Time (minutes)*
x_1 = *Number of Stops*
x_2 = *Distance (miles)*

$$b_1 = \frac{\sum(x_2-\overline{x_2})^2 \sum(x_1-\overline{x_1})(y-\bar{y}) - \sum(x_1-\overline{x_1})(x_2-\overline{x_2}) \sum(x_2-\overline{x_2})(y-\bar{y})}{\sum(x_1-\overline{x_1})^2 \sum(x_2-\overline{x_2})^2 - \left(\sum(x_1-\overline{x_1})(x_2-\overline{x_2})\right)^2}$$

$$b_2 = \frac{\sum(x_1-\overline{x_1})^2 \sum(x_2-\overline{x_2})(y-\bar{y}) - \sum(x_1-\overline{x_1})(x_2-\overline{x_2}) \sum(x_1-\overline{x_1})(y-\bar{y})}{\sum(x_1-\overline{x_1})^2 \sum(x_2-\overline{x_2})^2 - \left(\sum(x_1-\overline{x_1})(x_2-\overline{x_2})\right)^2}$$

$\bar{y}$	$\overline{x_1}$	$\overline{x_2}$
30	2	6

$\Sigma(x_1-\overline{x_1})^2$	$\Sigma(x_2-\overline{x_2})^2$	$\Sigma(x_1-\overline{x_1})(y-\bar{y})$	$\Sigma(x_2-\overline{x_2})(y-\bar{y})$	$\Sigma(x_1-\overline{x_1})(x_2-\overline{x_2})$
4	26	18	47	-1

PIERIAN DATA

Regression Exercise #2

y = Delivery Time (minutes)
x_1 = Number of Stops
x_2 = Distance (miles)

$$b_1 = \frac{(26)(18) - (-1)(47)}{(4)(26) - ((-1))^2} = \frac{515}{103} = \mathbf{5}$$

$$b_2 = \frac{(4)\,(47) - (-1)(18)}{(4)(26) - ((-1))^2} = \frac{206}{103} = \mathbf{2}$$

$\bar{y}$	$\overline{x_1}$	$\overline{x_2}$
30	2	6

$\Sigma(x_1-\overline{x_1})^2$	$\Sigma(x_2-\overline{x_2})^2$	$\Sigma(x_1-\overline{x_1})(y-\bar{y})$	$\Sigma(x_2-\overline{x_2})(y-\bar{y})$	$\Sigma(x_1-\overline{x_1})(x_2-\overline{x_2})$
4	26	18	47	-1

PIERIAN DATA

Regression Exercise #2

y = *Delivery Time (minutes)*
x_1 = *Number of Stops*
x_2 = *Distance (miles)*

$$\hat{y} = 8 + 5x_1 + 2x_2$$

$$b_1 = \frac{(26)(18) - (-1)(47)}{(4)(26) - ((-1))^2} = \frac{515}{103} = \mathbf{5}$$

$$b_2 = \frac{(4)\ (47) - (-1)(18)}{(4)(26) - ((-1))^2} = \frac{206}{103} = \mathbf{2}$$

$$b_0 = \bar{y} - b_1\overline{x_1} - b_2\overline{x_2}$$
$$= 30 - (5)(2) - (2)(6)$$
$$= 30 - 10 - 12 = \mathbf{8}$$

$\bar{y}$	$\overline{x_1}$	$\overline{x_2}$
30	2	6

$\Sigma(x_1-\overline{x_1})^2$	$\Sigma(x_2-\overline{x_2})^2$	$\Sigma(x_1-\overline{x_1})(y-\bar{y})$	$\Sigma(x_2-\overline{x_2})(y-\bar{y})$	$\Sigma(x_1-\overline{x_1})(x_2-\overline{x_2})$
4	26	18	47	-1

Regression Exercise #2

y = *Delivery Time (minutes)*
x_1 = *Number of Stops*
x_2 = *Distance (miles)*

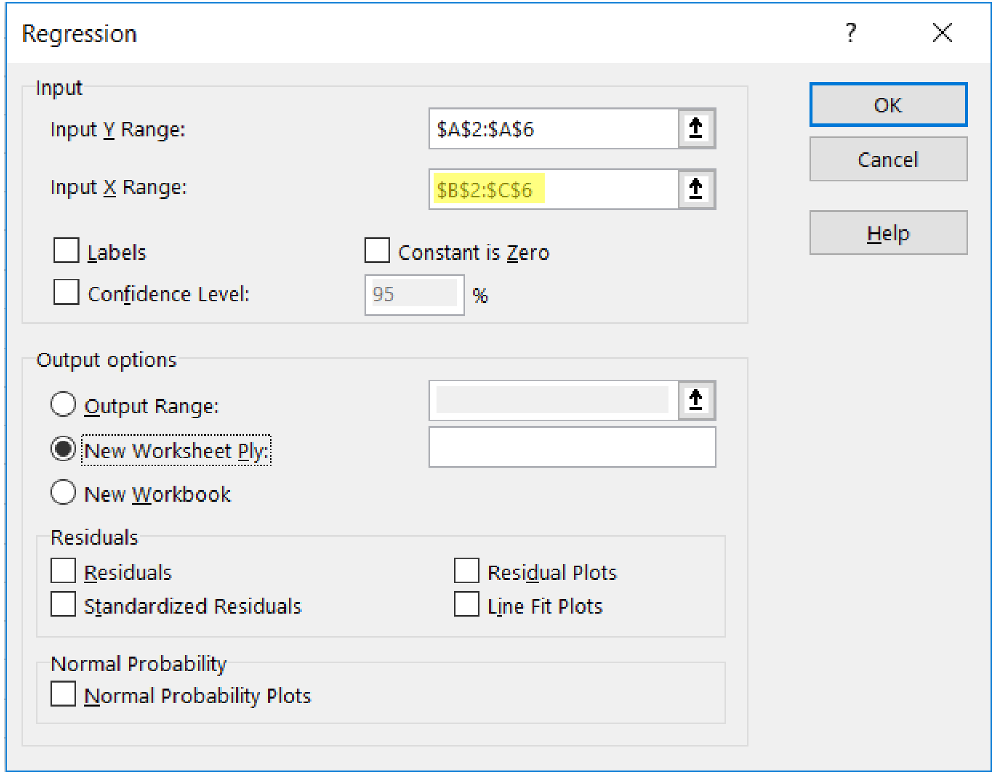
$$\hat{y} = 8 + 5x_1 + 2x_2$$

y	x_1	x_2
29	1	8
31	3	4
36	2	9
35	3	6
19	1	3

Based on our analysis, pharmacy deliveries have a fixed time of 8 minutes, plus 5 minutes for each stop, and 2 minutes for each mile traveled

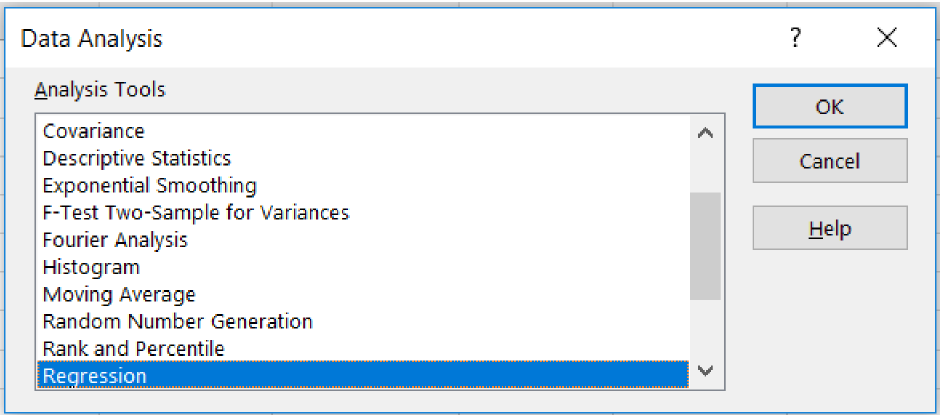
Multiple Regression in Excel

Steps are the same as linear regression, except you select a wider x-axis range



Multiple Regression in Excel

	A	B	C						
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	1							
5	R Square	1							
6	Adjusted R Square	1							
7	Standard Error	1.25607E-15							
8	Observations	5							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	184	92	5.83119E+31	1.71492E-32			
13	Residual	2	3.15544E-30	1.57772E-30					
14	Total	4	184						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	8	2.11706E-15	3.77882E+15	7.00306E-32	8	8	8	8
18	X Variable 1	5	6.31078E-16	7.92295E+15	1.59304E-32	5	5	5	5
19	X Variable 2	2	2.47529E-16	8.07985E+15	1.53177E-32	2	2	2	2
20									



MULTIPLE REGRESSION with Python

```
>>> from sklearn.linear_model import LinearRegression
>>> x1,x2 = [1,3,2,3,1], [8,4,9,6,3]
>>> y = [29,31,36,35,19]
>>> reg = LinearRegression()
>>> reg.fit(list(zip(x1,x2)), y)
>>> b1,b2 = reg.coef_[0], reg.coef_[1]
>>> b0 = reg.intercept_
>>> print(f'y = {b0:.{3}} + {b1:.{3}}x1 + {b2:.{3}}x2’)
y = 8.0 + 5.0x1 + 2.0x2
```

Next Up: CHI-SQUARE ANALYSIS