neue fische

School and Pool for Digital Talent

neue fische
School and Pool for Digital Talent

Linear Regression

neue fische
School and Pool for Digital Talent

Linear Regression

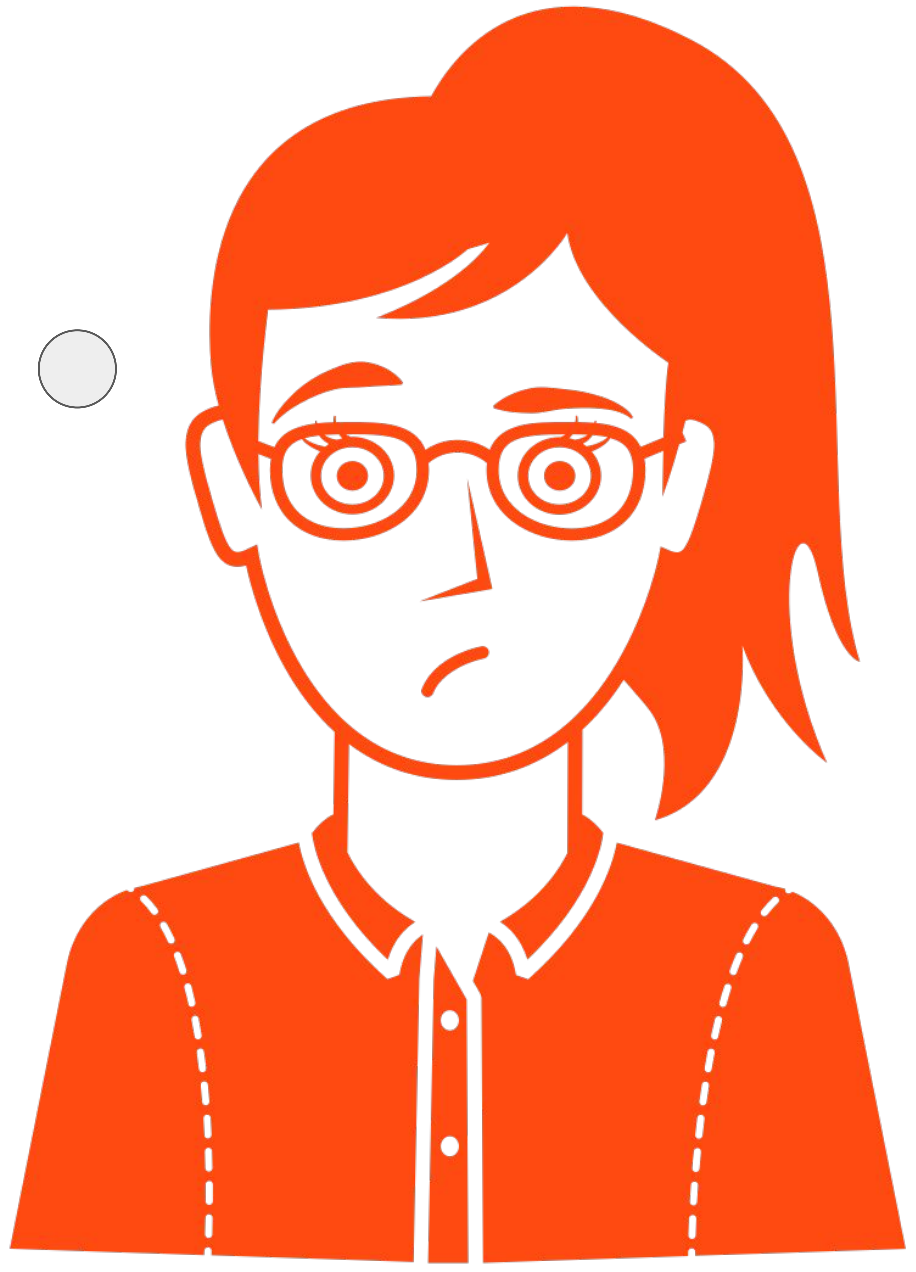
Part 1
Motivation

Goals of Linear regression

I own a house in King County!

It has 3 Bedrooms, 2 Bathroom, a nice 10.000 sqft lot and is only 10km away from Bill Gates mansion!

If only I had a way of estimating what it is worth...

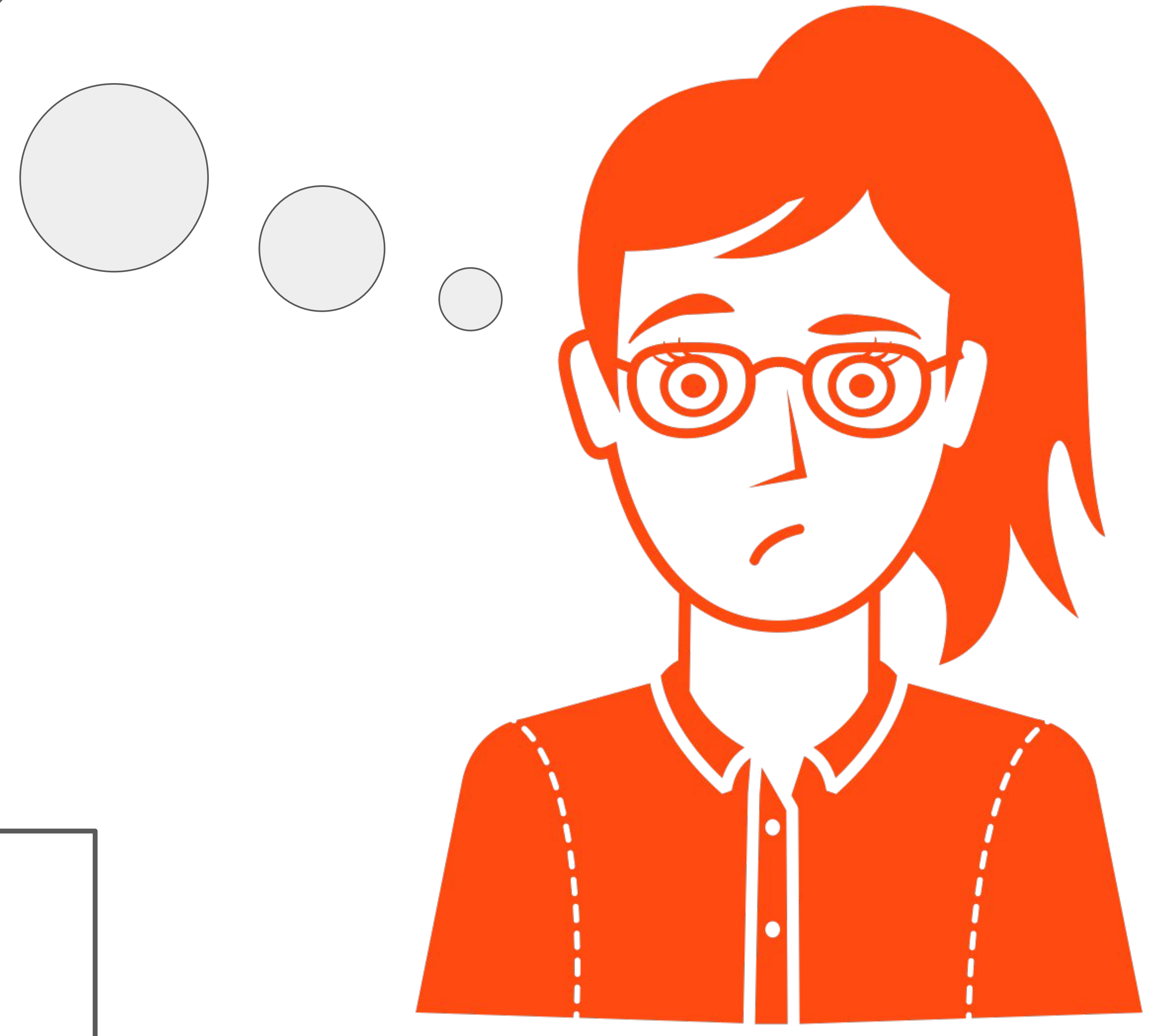


Goals of Linear regression

I own a house in King County!

It has 3 Bedrooms, 2 Bathroom, a nice 10.000 sqft lot and is only 10km away from Bill Gates mansion!

If only I had a way of estimating what it is worth...



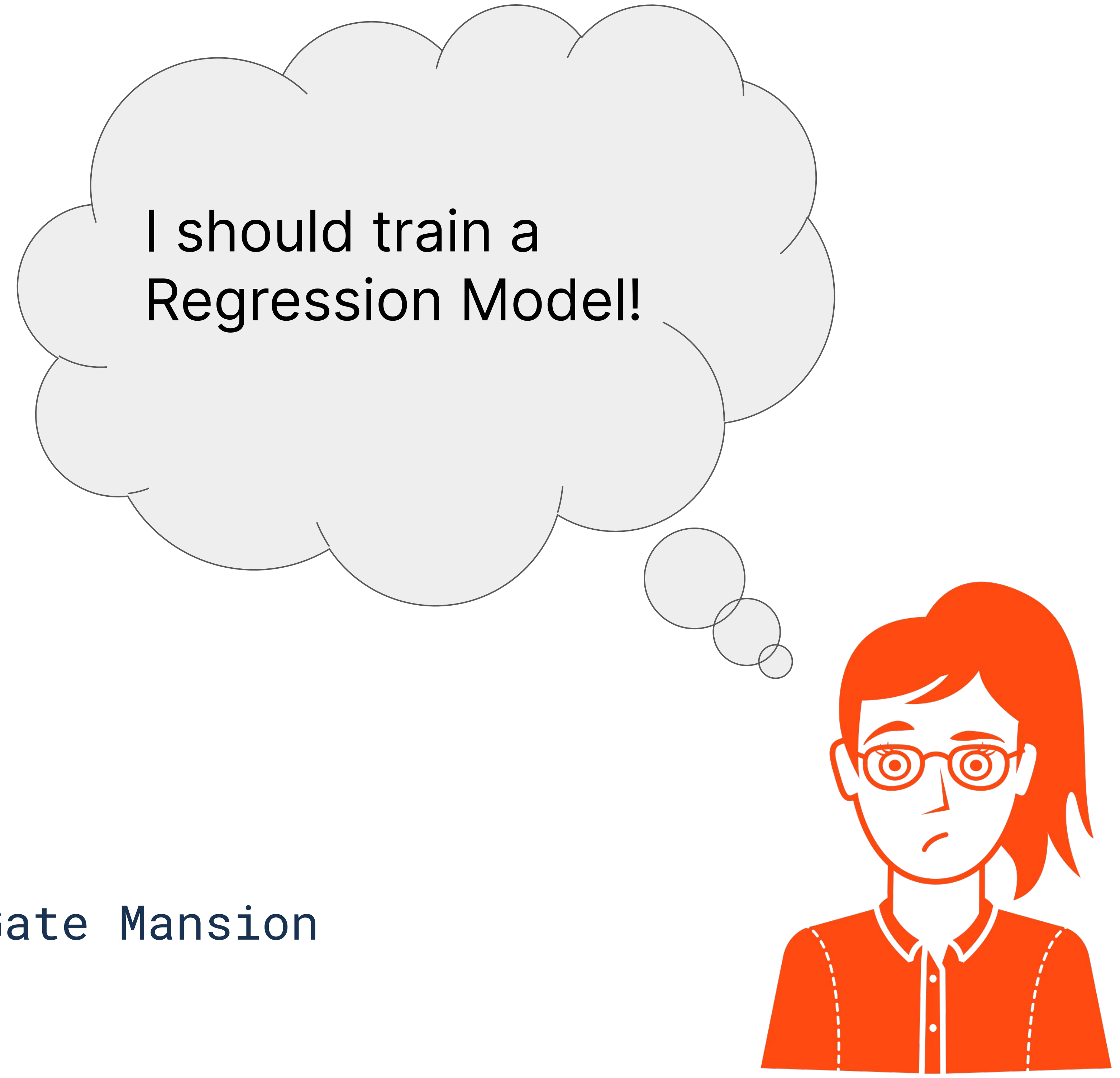
Use training data to...

find a similar data point

Use training data to...

find rules that can be used for generalization

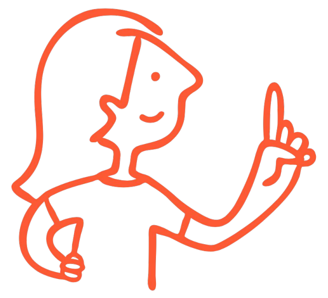
Goals of Linear regression



Building a model

```
    216.645 $ Basis-price
+    20.033 $ for each bedrooms
+   234.314 $ for each bathrooms
+         1 $ for each sqft lot
-    14.745 $ for each km distance from Bill Gate Mansion

=       xyz $ estimated house price
```



The term regression (e.g. regression analysis) usually refers to linear regression.

Goals of Linear regression

Building a model

```
   216.645 $ Basis-price
+   20.033 $ for each bedrooms
+  234.314 $ for each bathrooms
+        1 $ for each sqft lot
-   14.745 $ for each km distance from Bill Gate Mansion

=     xyz $ estimated house price
```

Descriptive statistics

Using LR for explanation (profiling)

→ Why is my house worth xyz?
→ How can I increase the price?

Goals of Linear regression

Building a model

```
   216.645 $ Basis-price
+   20.033 $ for each bedrooms,
+  234.314 $ for each bathrooms
+        1 $ for each sqft lot
-   14.745 $ for each km distance from Bill Gate Mansion

=      xyz $ estimated house price
```

Descriptive statistics
Using LR for explanation (profiling)

→ Why is my house worth xyz?
→ How can I increase the price?

Inferential statistics:
Using LR to make predictions

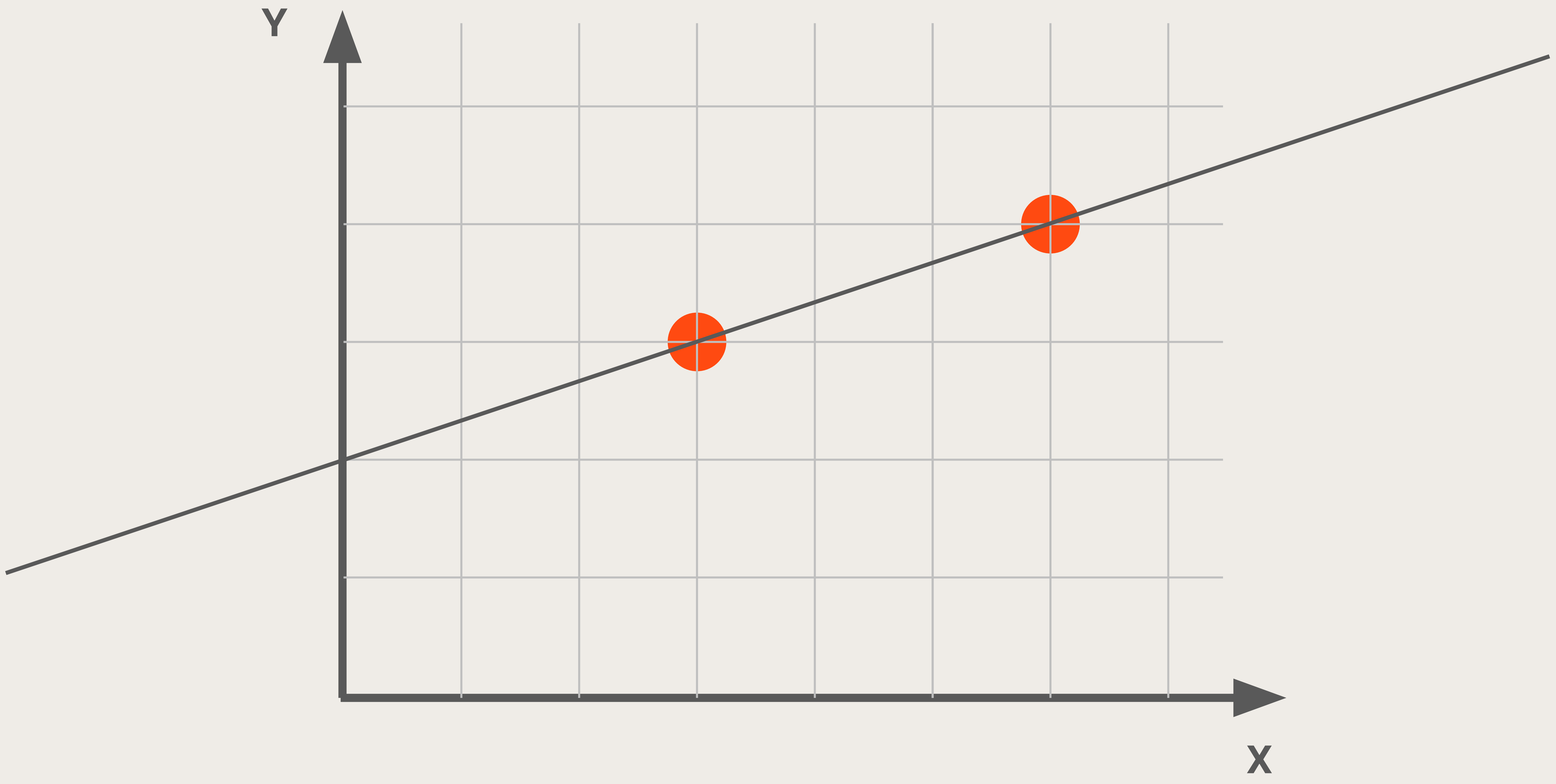
→ How much is my house worth?

neue fische
School and Pool for Digital Talent

Linear Regression

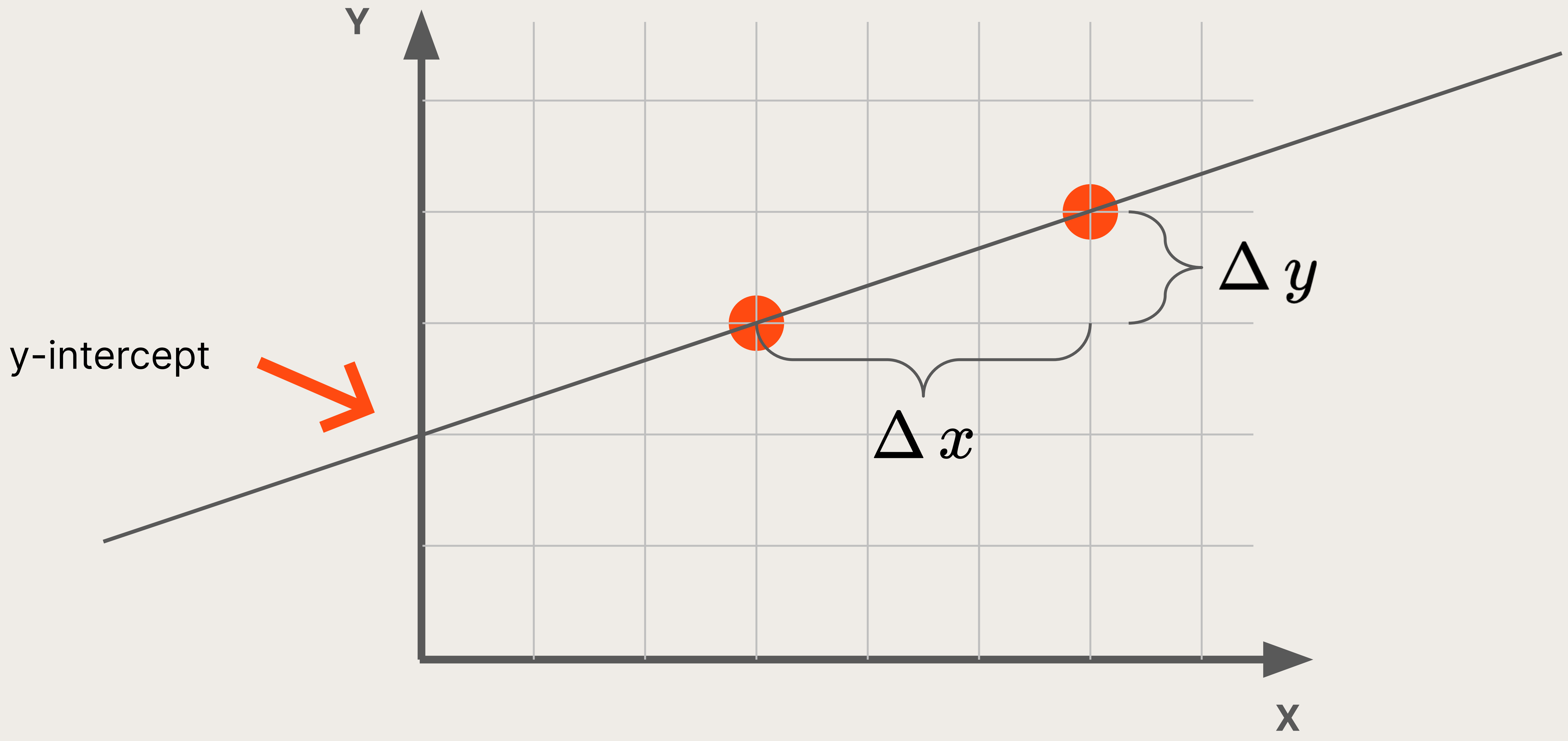
Part 2
Linear Equation

Linear Equation



Q: What is the equation of the line?

Linear Equation



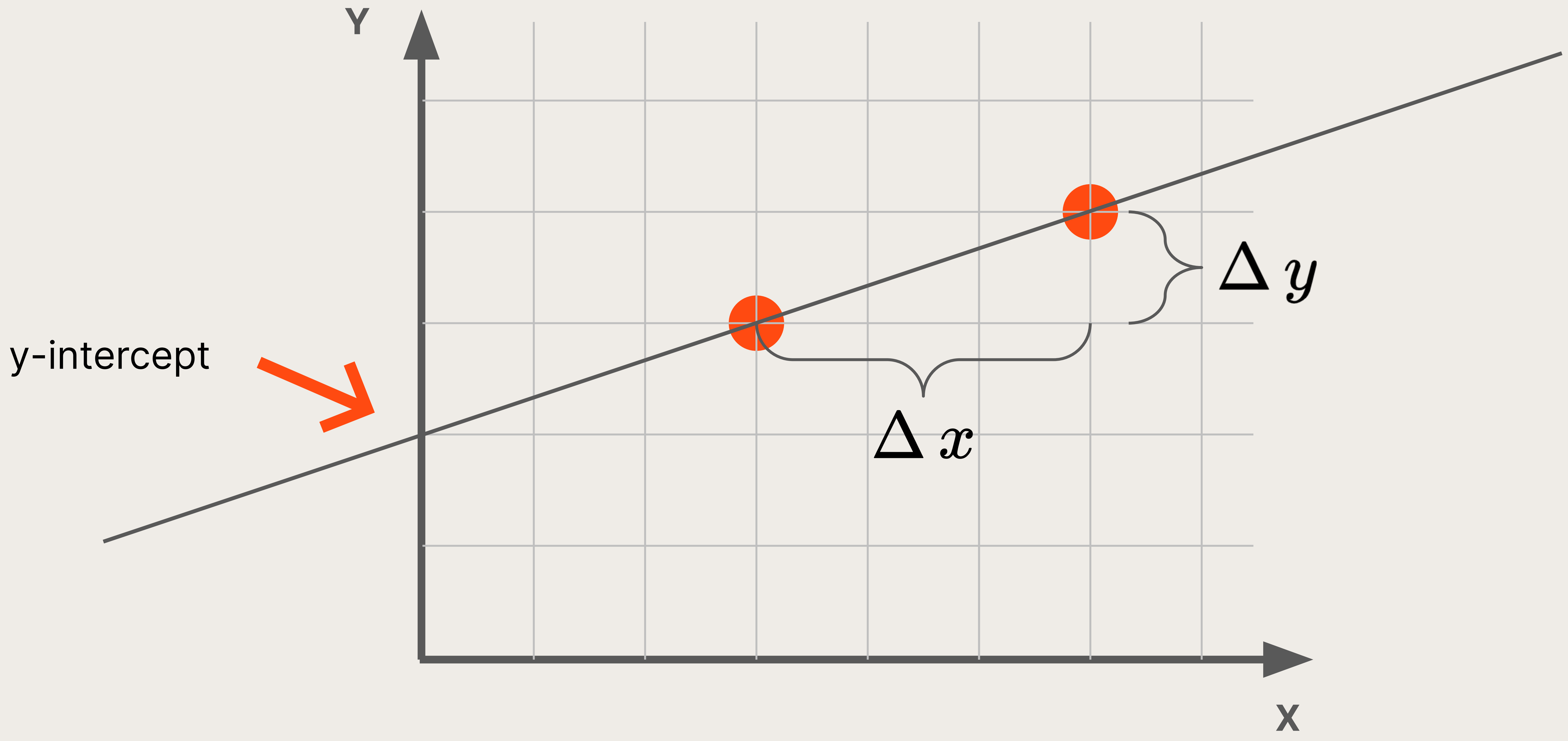
$$y = b_0 + b_1 \cdot x$$

$$b_1 = \frac{\Delta y}{\Delta x}$$

Q: What is the equation of the line?

$$y = 2 + \frac{1}{3} \cdot x$$

Linear Equation



$$y = b_0 + b_1 \cdot x$$

Key terms

- Intercept (b0 , value of y when x = 0)
- Slope (regression coefficient, weights, b1)

Linear Regression

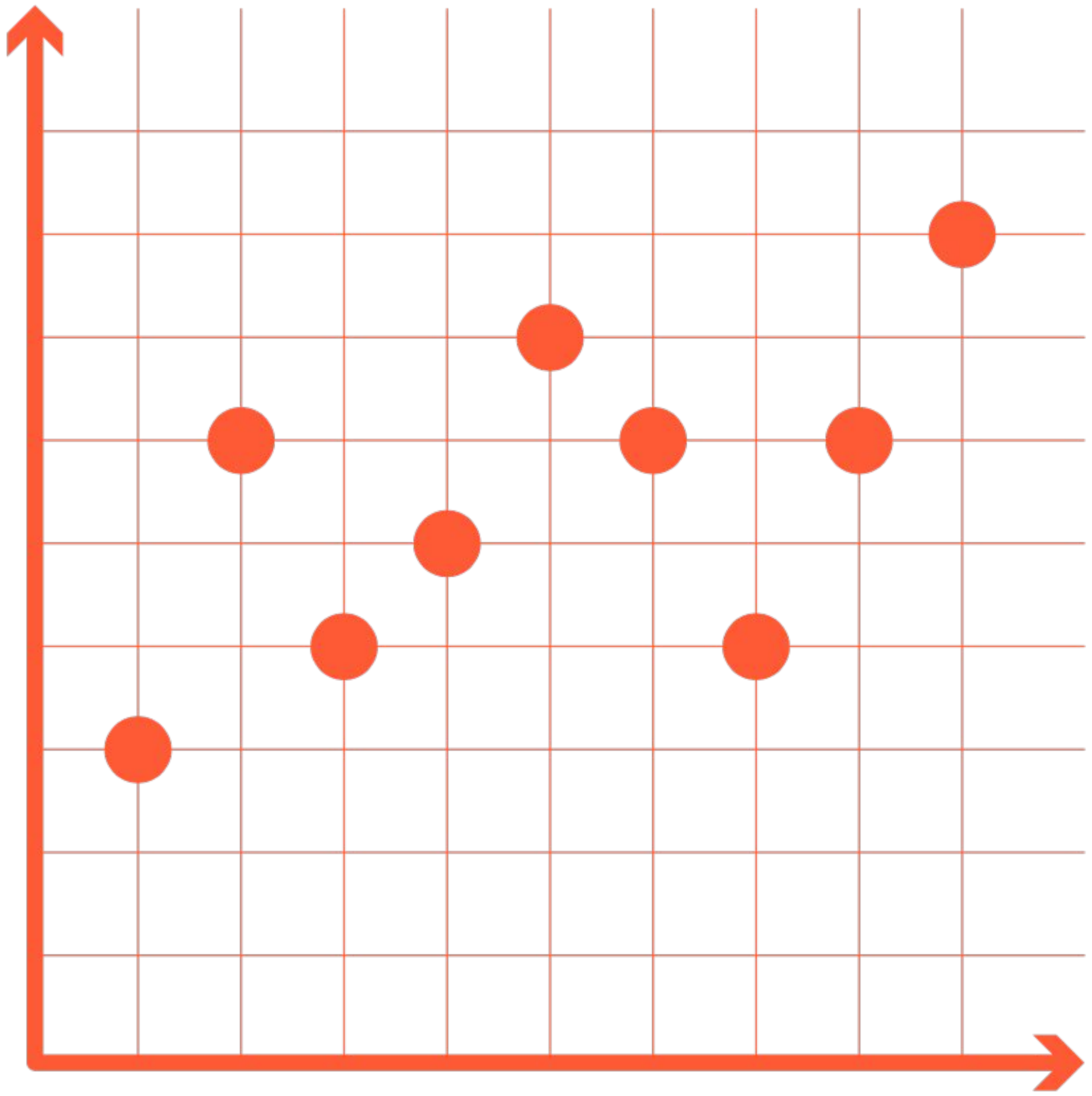
Part 3
Linear Regression

Linear Regression

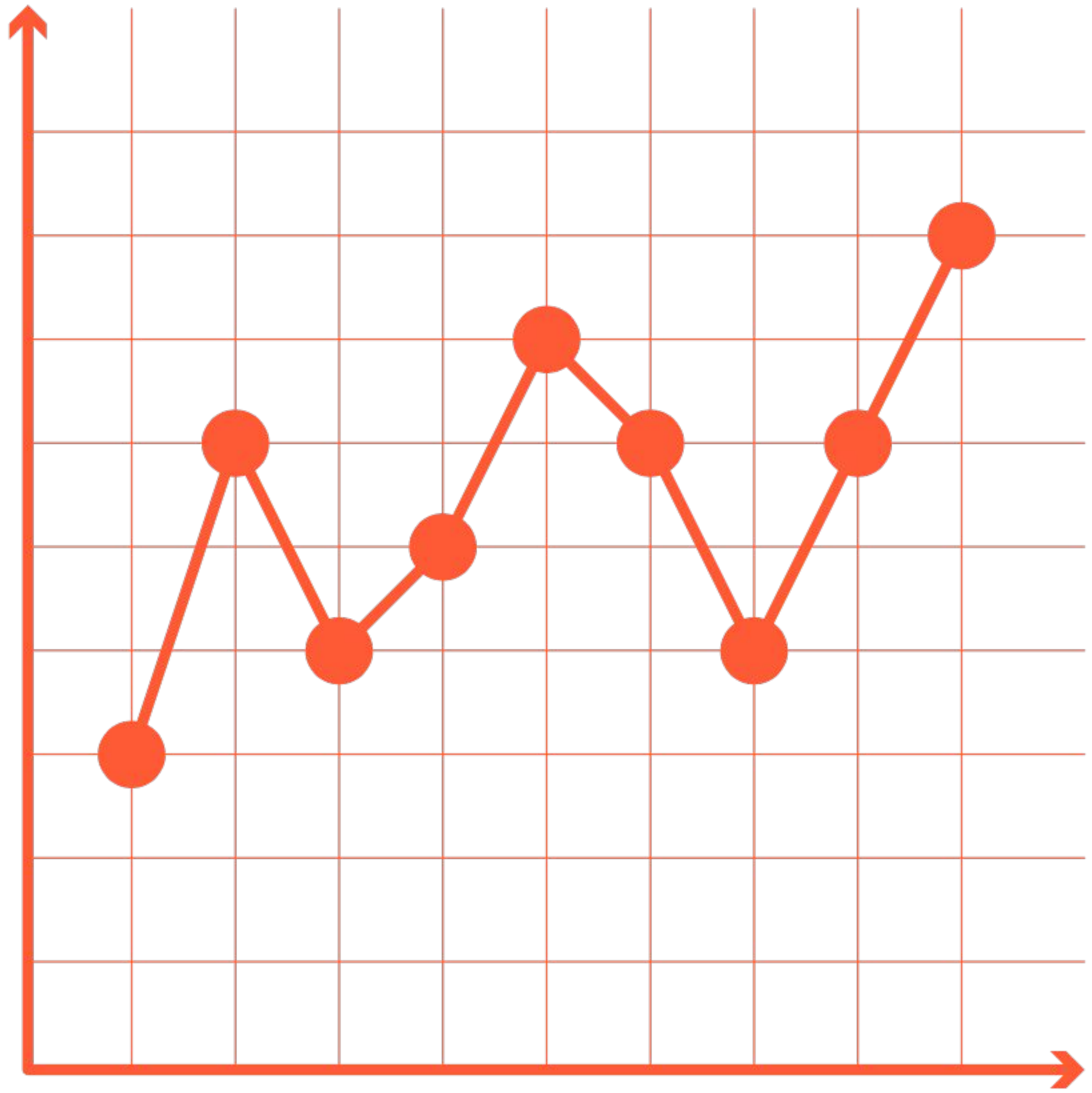
Is the variable X associated with a variable Y, and if so,
what is the relationship and can we use it to predict Y?

Correlation - measures the strength of the relationship
Regression - quantifies the nature of the relationship

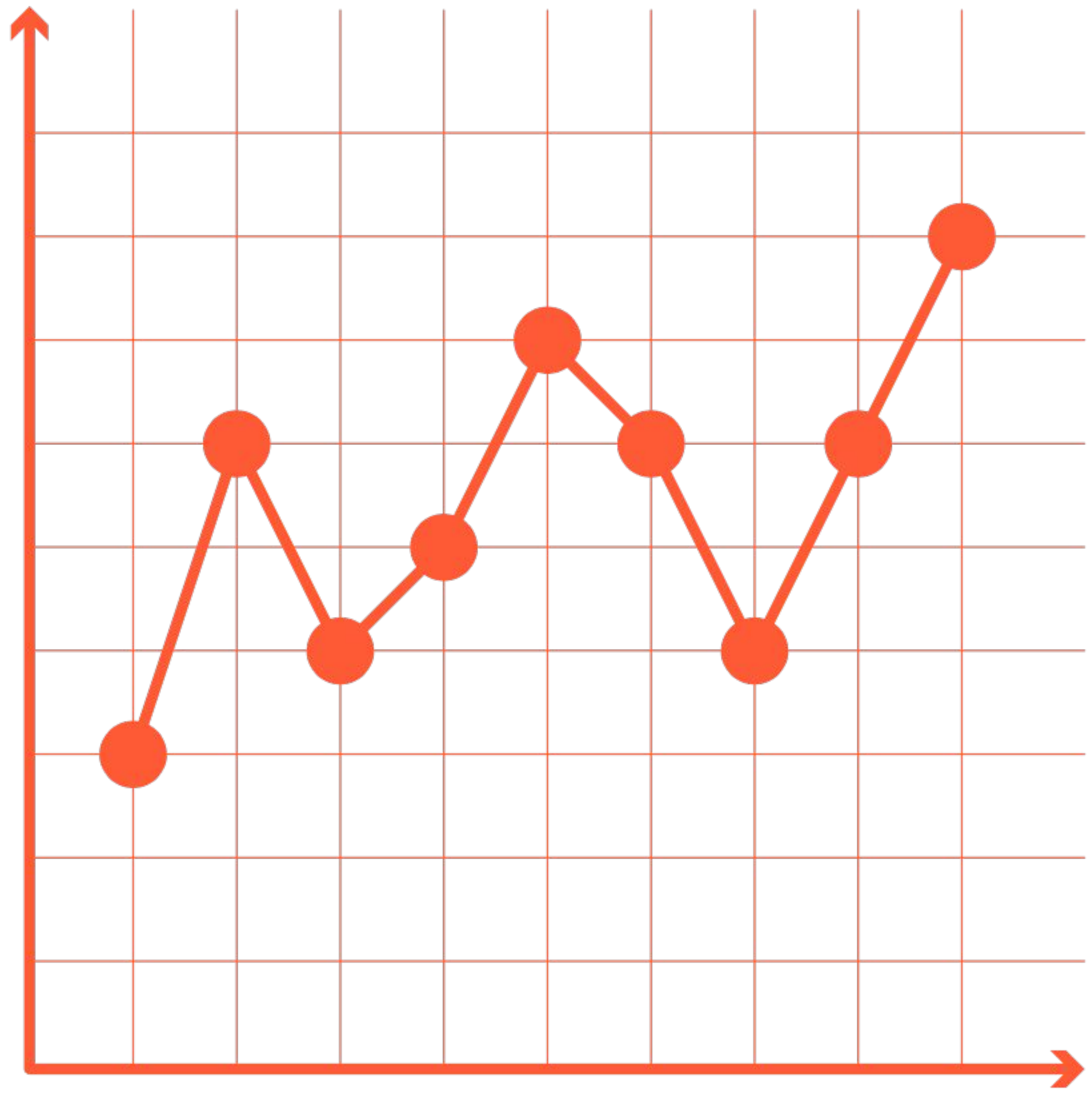
What about more than 2 points?

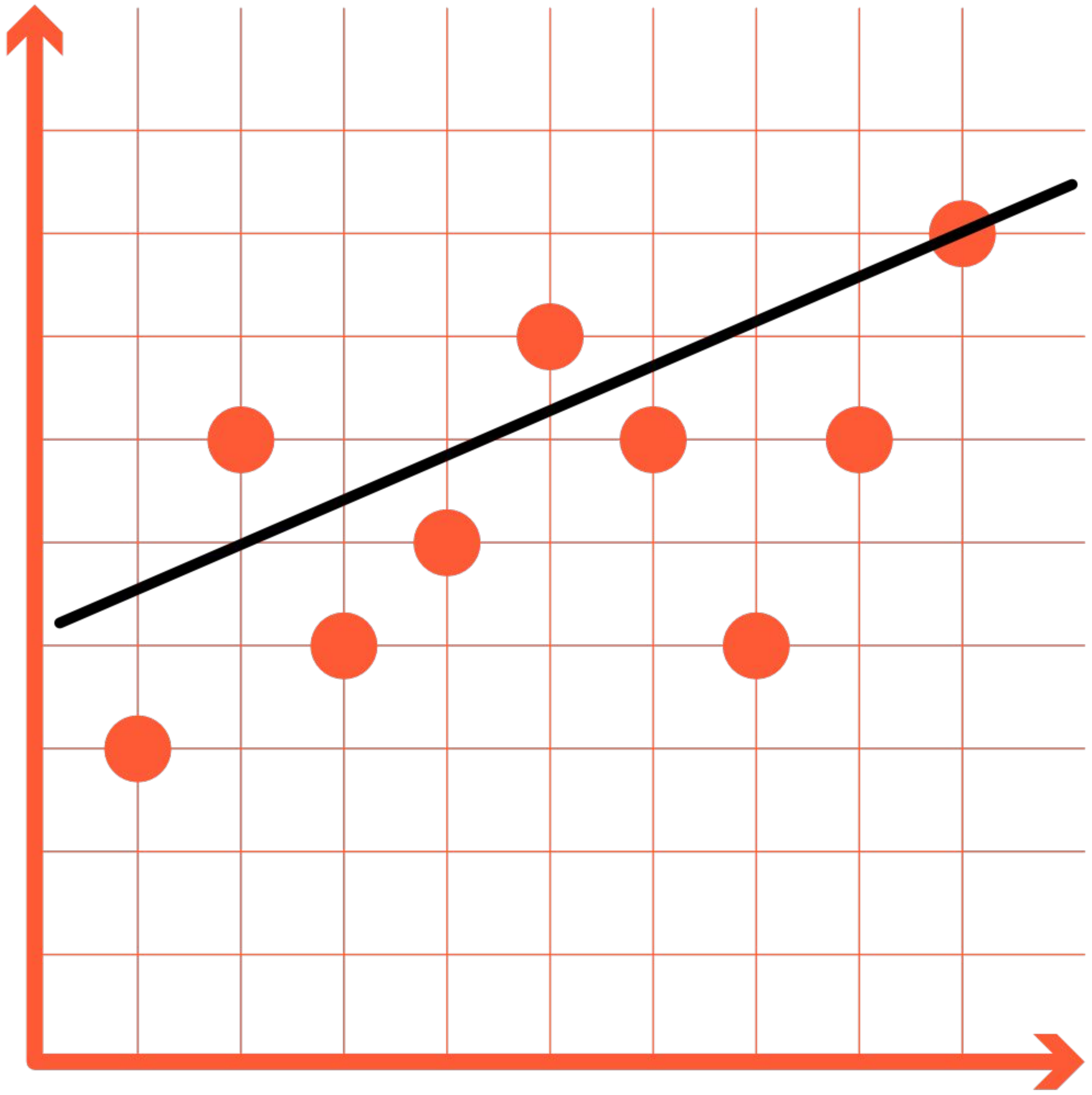


What about more than 2 points?

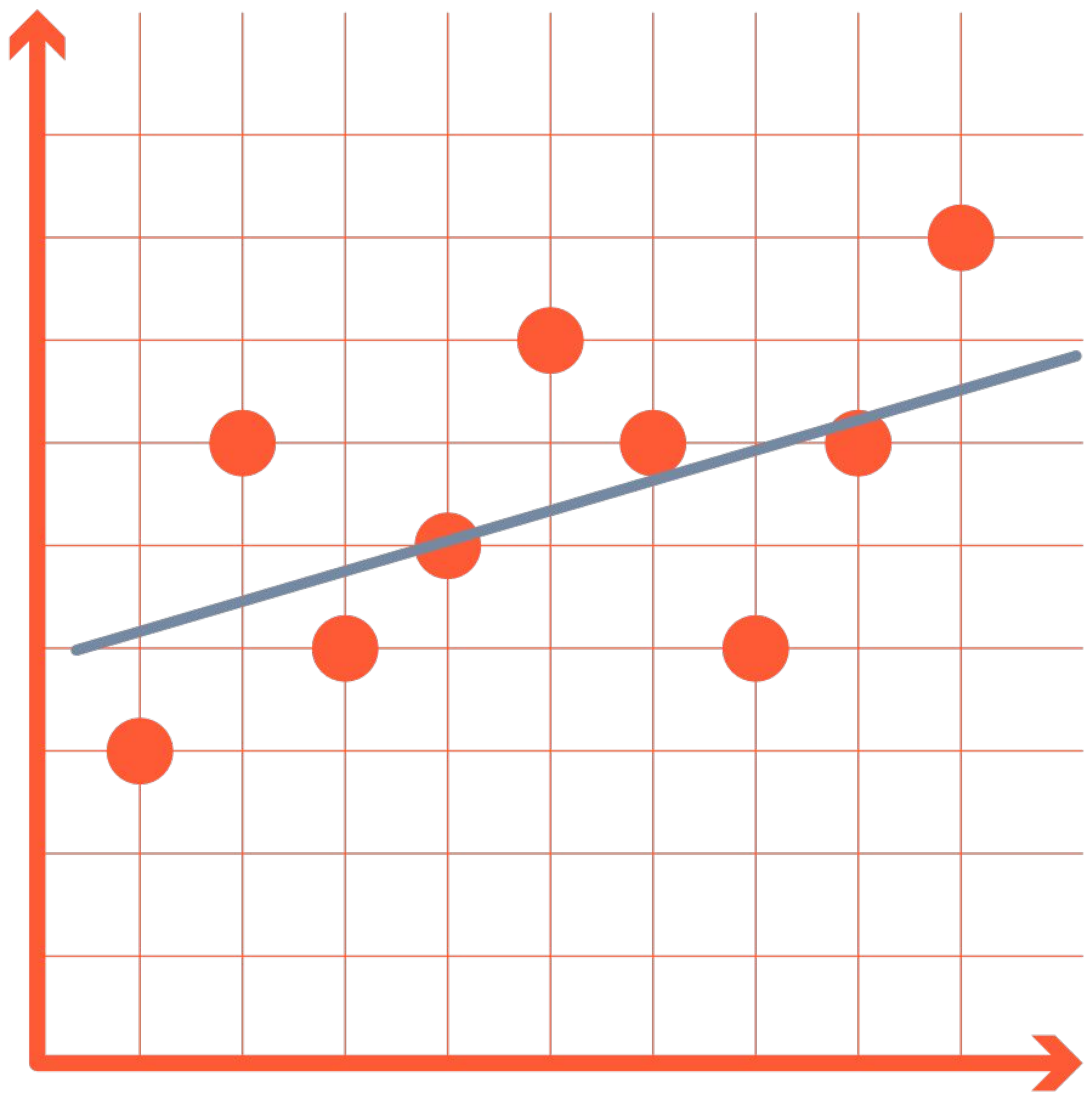


What about more than 2 points?

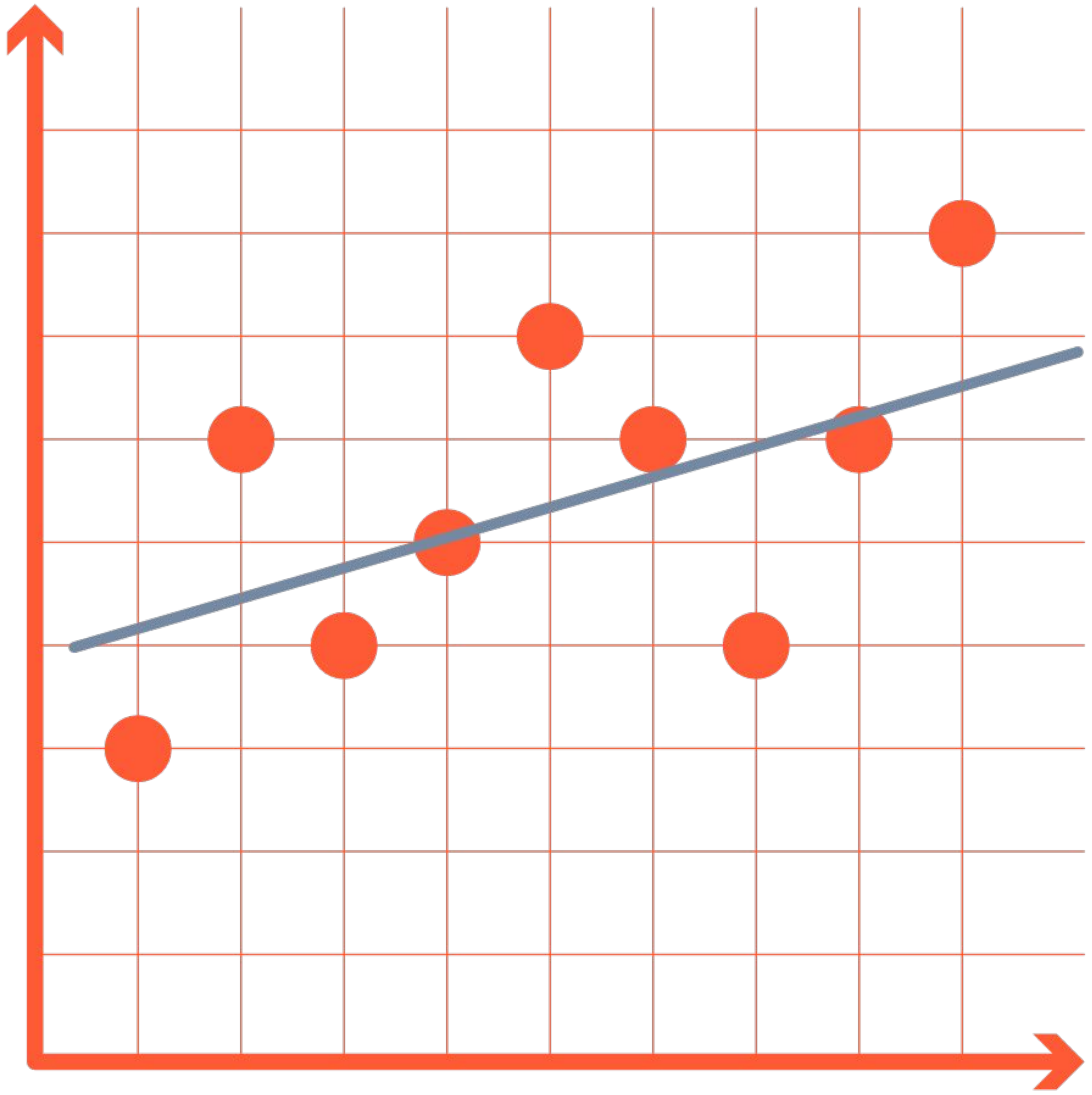


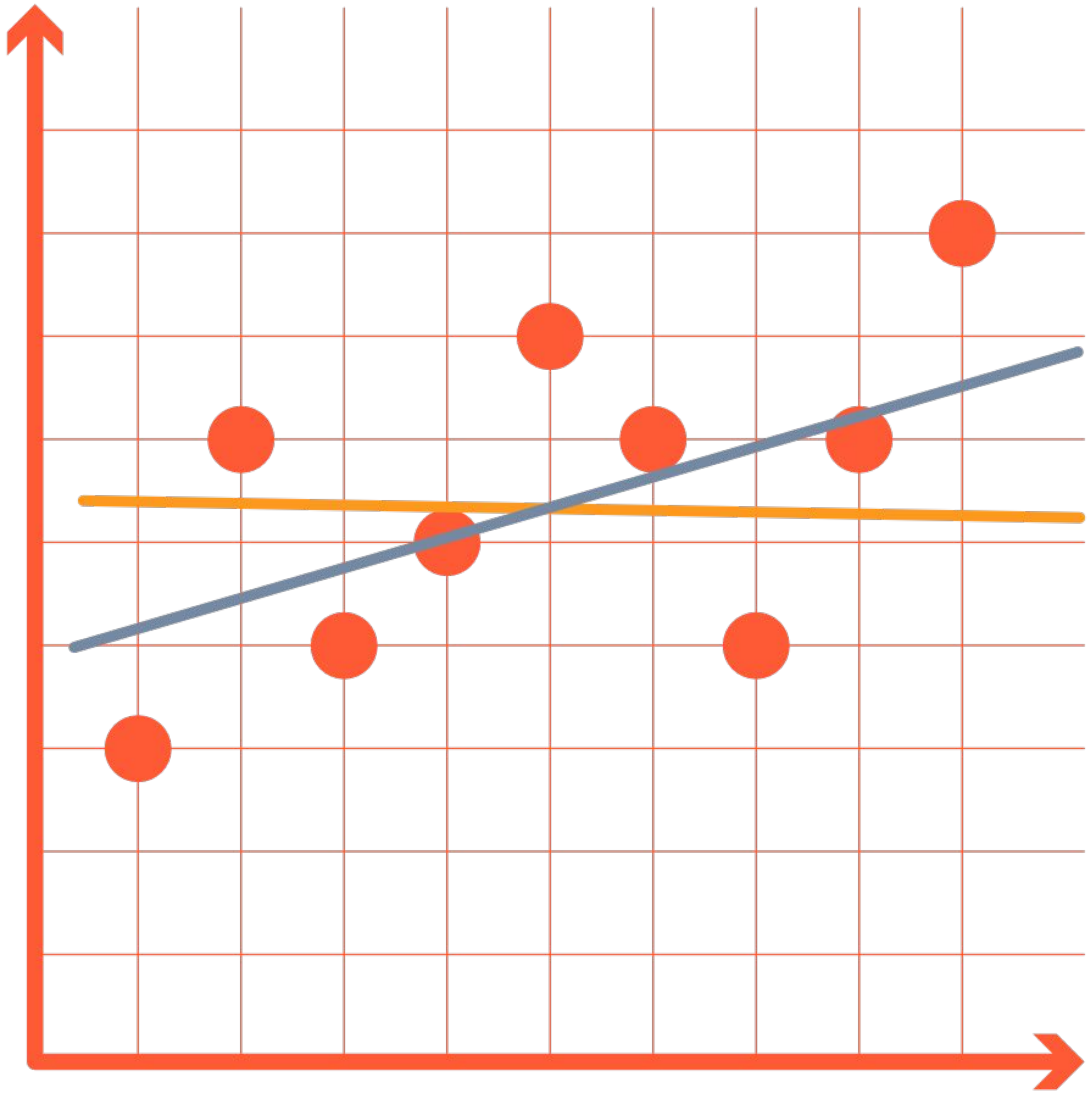


What about more than 2 points?

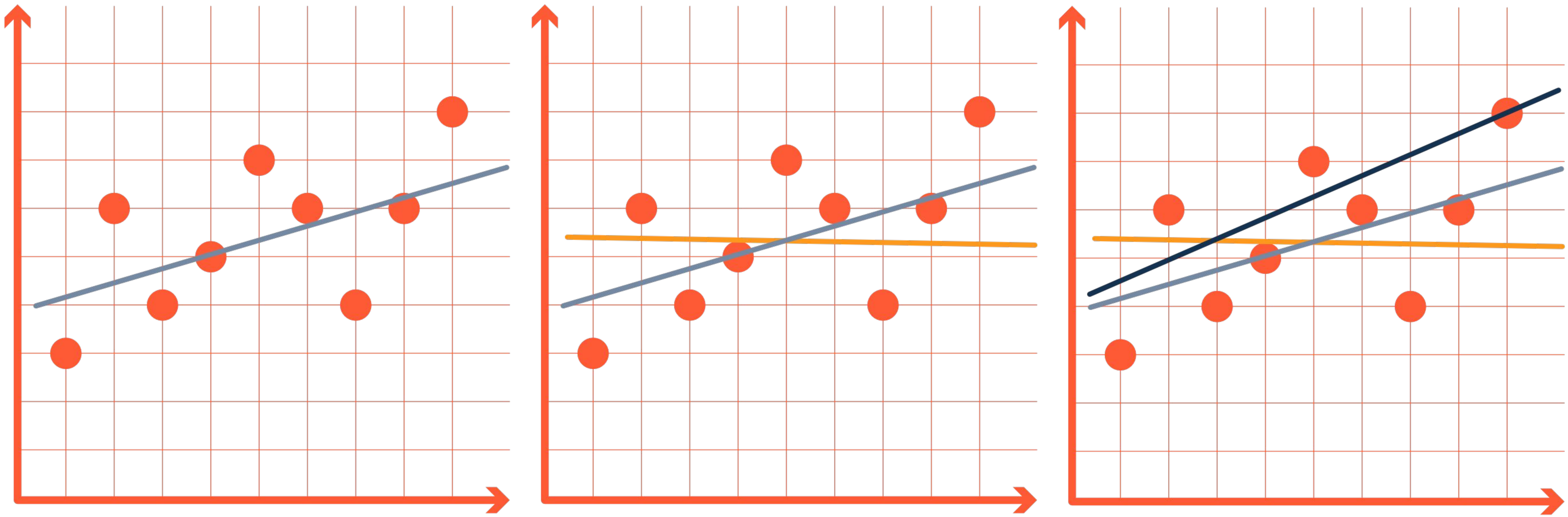


What about more than 2 points?





What about more than 2 points?

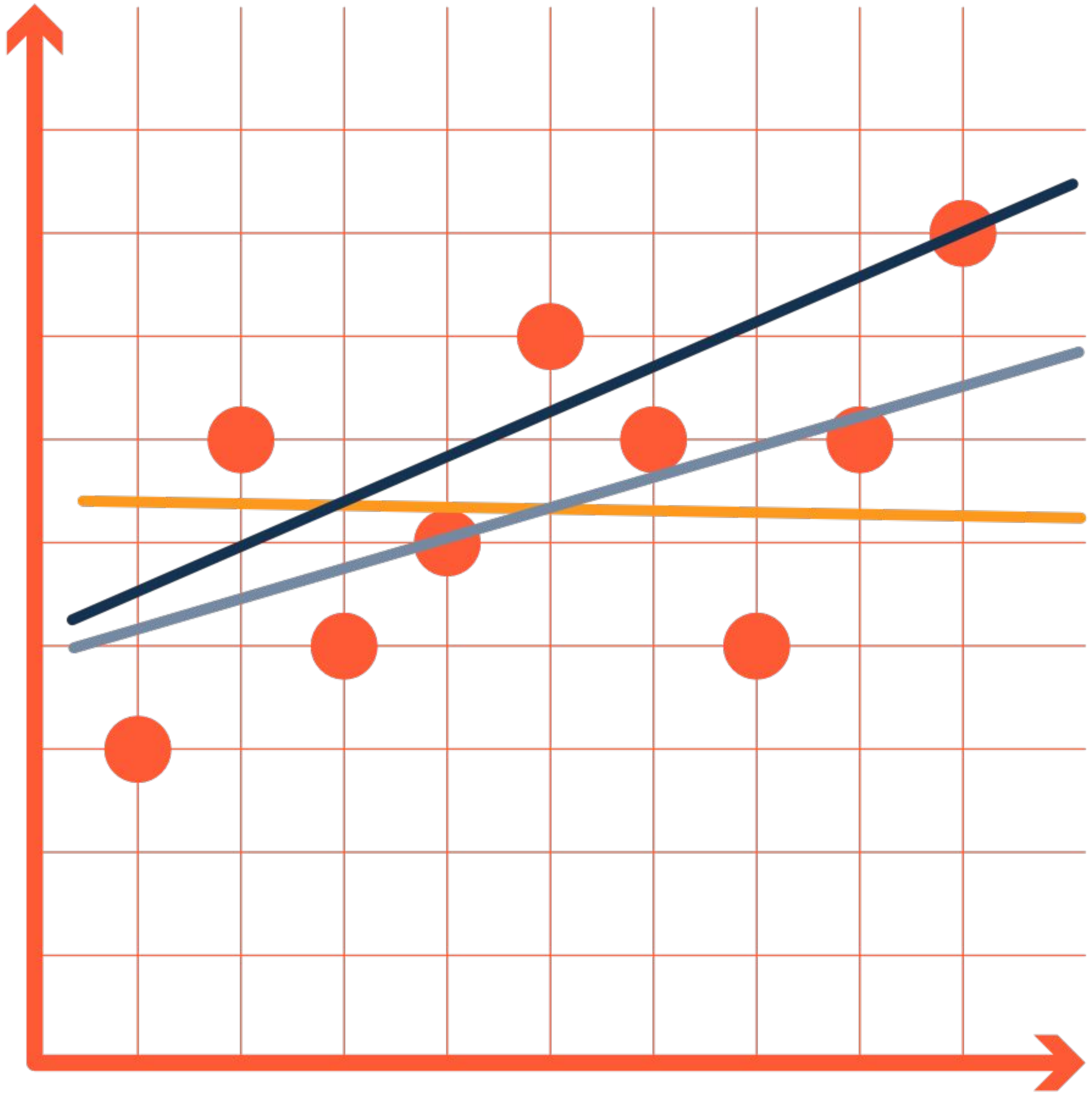


What about more than 2 points?

Which line best **fits** the data?

How do we get the best fitting line?

How do we know the line is best fitting?



neue fische
School and Pool for Digital Talent

Linear Regression

Part 3
Linear Regression

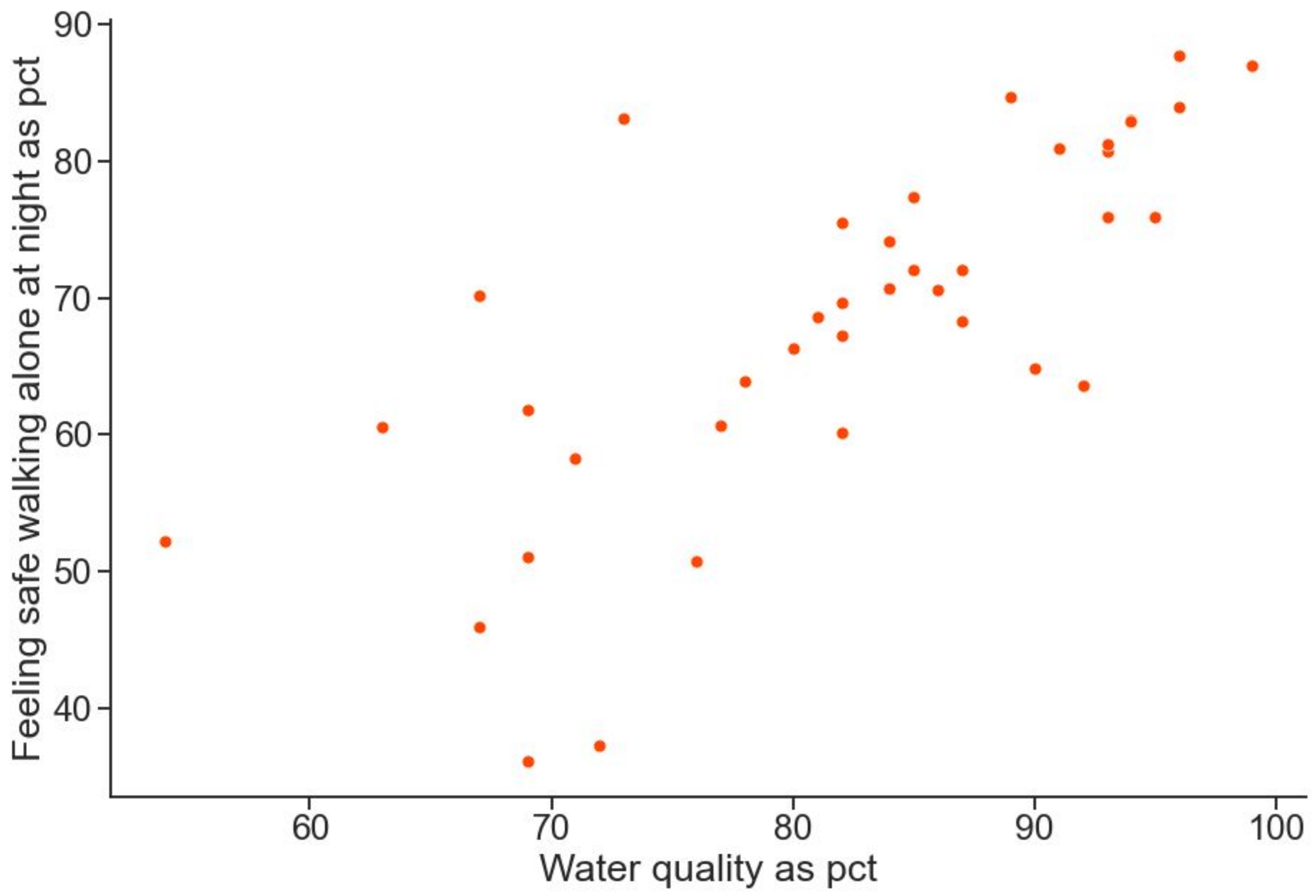
Let's look at the world happiness dataset (kaggle)

Two correlated variables

- water quality
- feeling safe walking alone at night
- r = 0.742054

$$y = b_0 + b_1 \cdot x + e$$

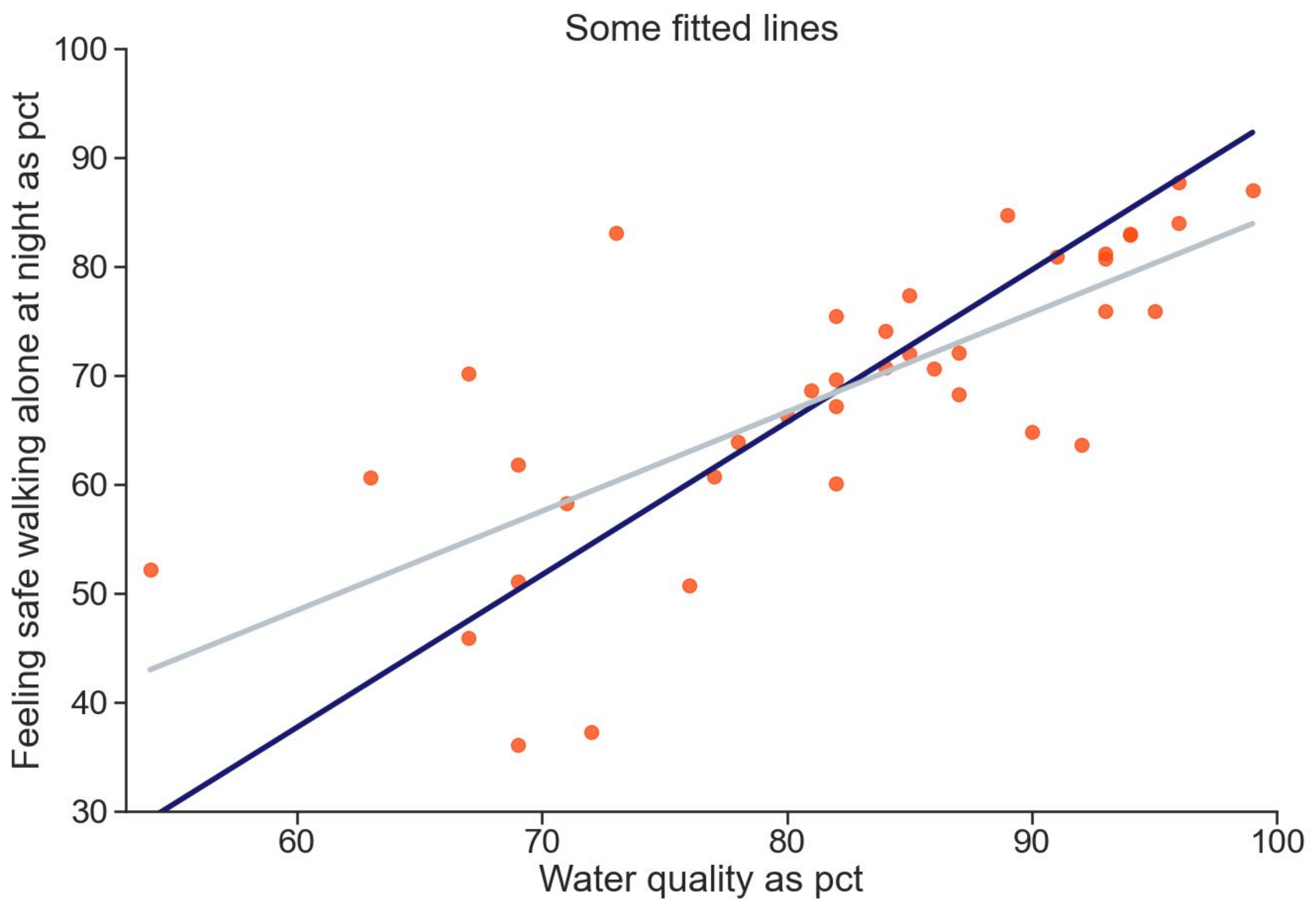
→ Find b_0 and b_1!



Trying out some lines.. which one is better?

Grey: $\hat{y} = -5.18 + 0.9 \cdot x$

Blue: $\hat{y} = -46.28 + 1.4 \cdot x$



***^ - the "hat" notation means the value is estimated** as opposed to a known value*
the estimate has uncertainty whereas the true value is fixed

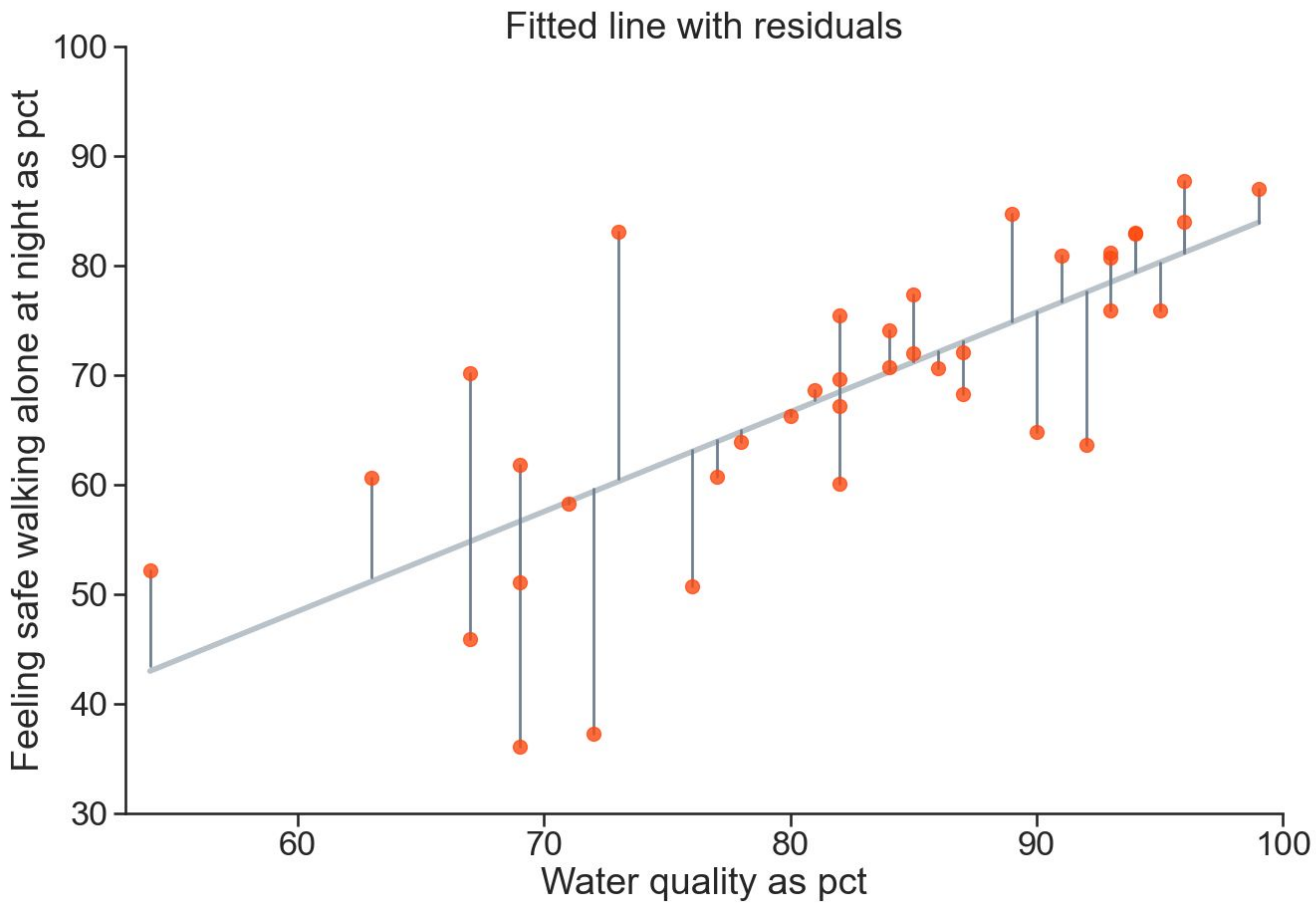
HOW DO WE KNOW WHICH LINE IS BETTER?

Residuals

$$e_i = y_i - \hat{y}_i$$

which means:

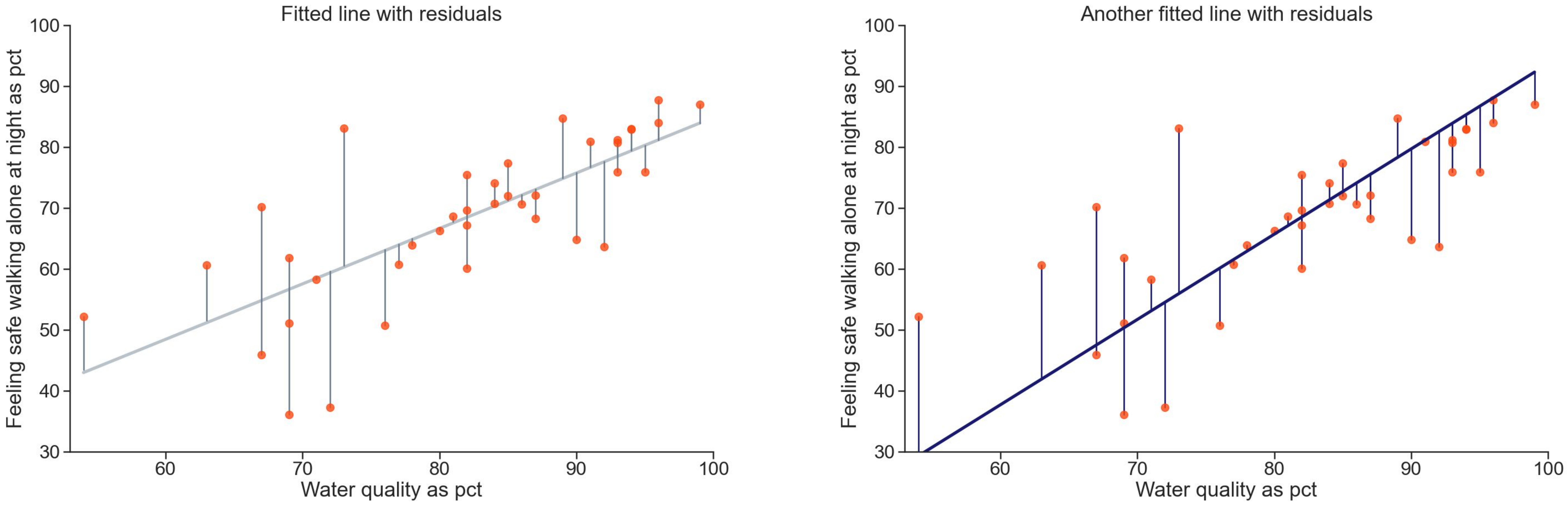
$$y_i = b_0 + b_1 \cdot x_i + e_i$$



Least squares criterion

By comparing the sum of squared residuals (SSR) we can find out which one is better:

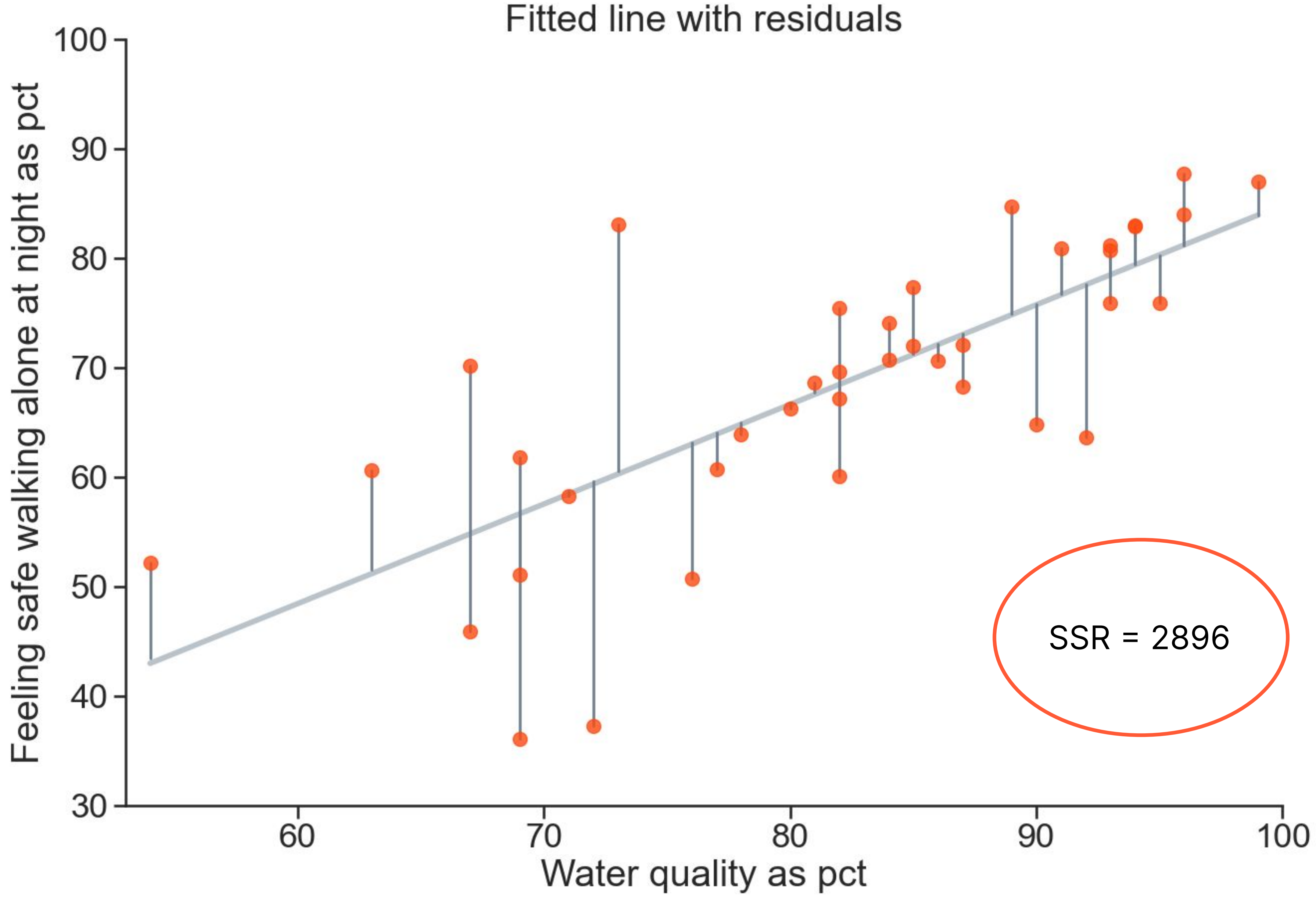
$$J(b_0, b_1) = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

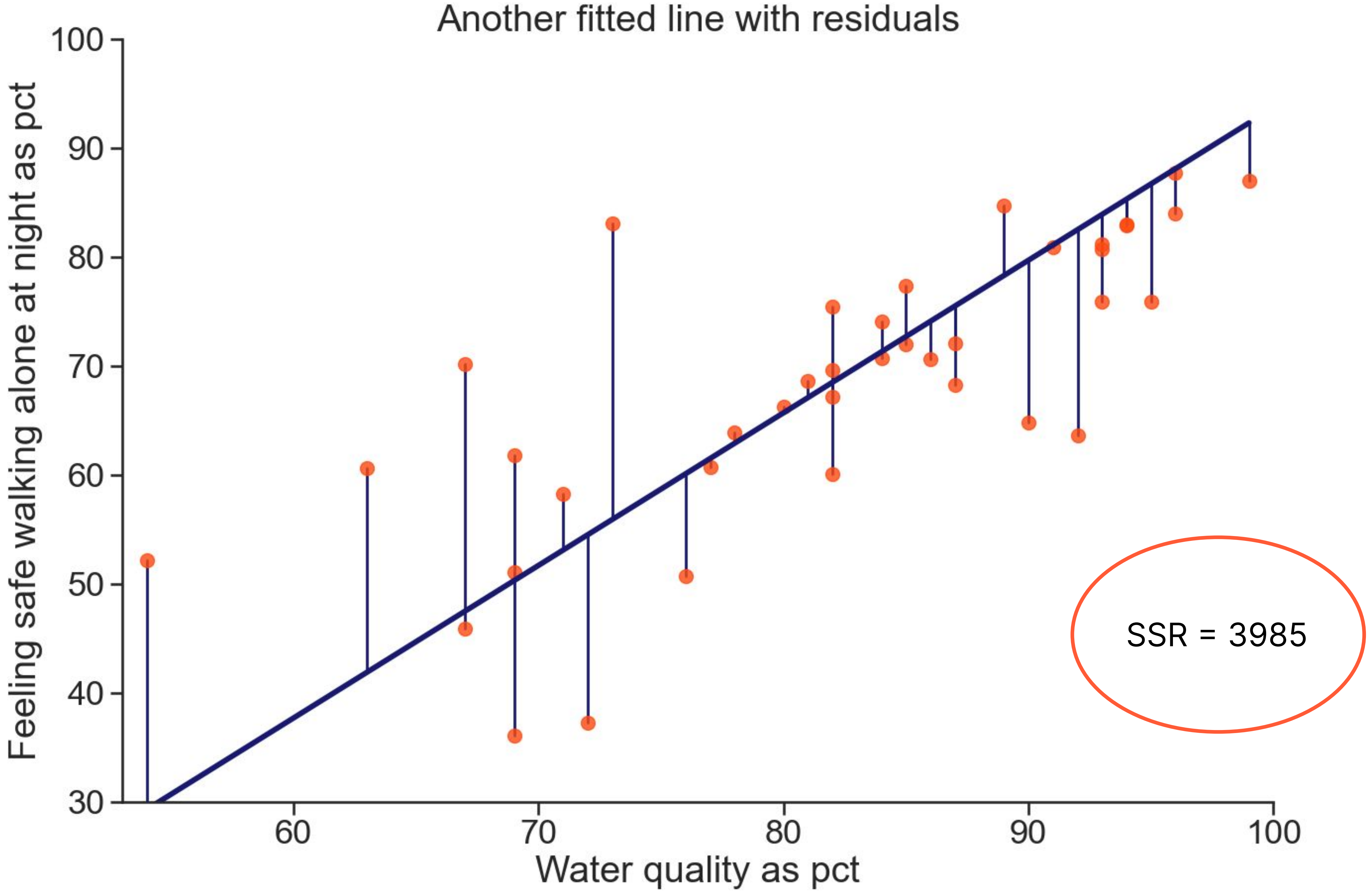


Least squares criterion

By comparing the sum of squared residuals (SSR) we can find out which one is better:

$$J(b_0, b_1) = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

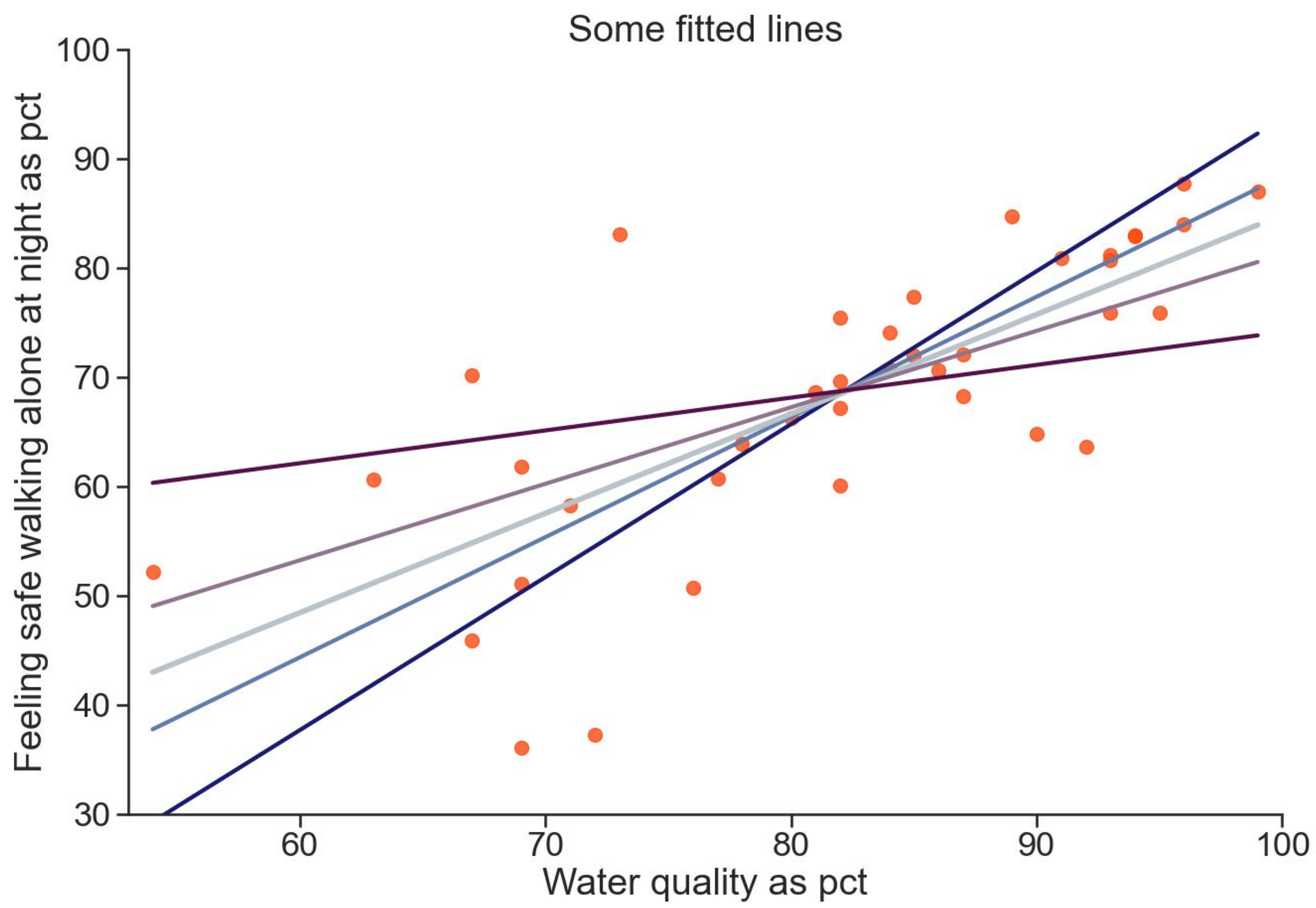


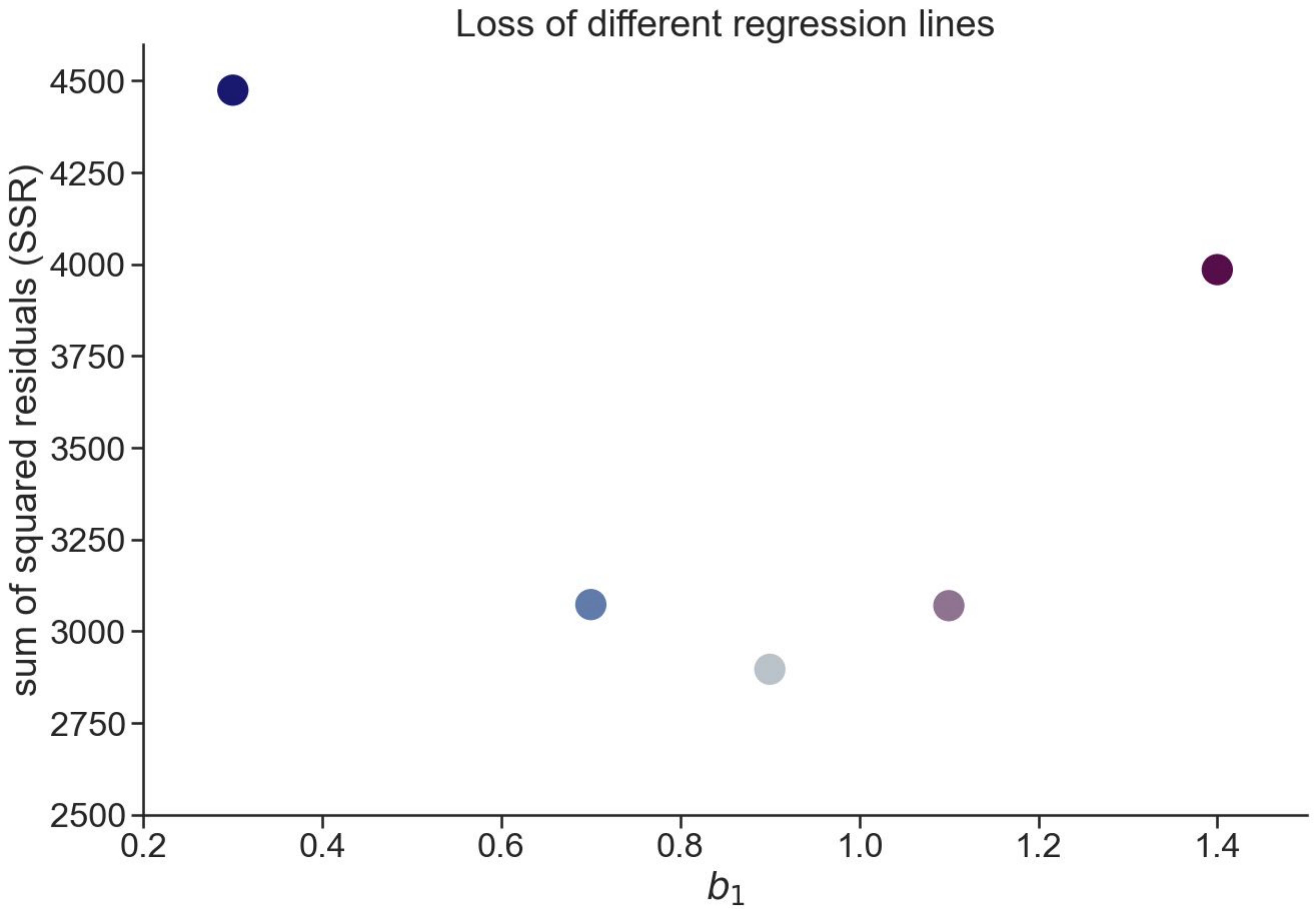


Trying out several fitted lines

By comparing the sum of squared residuals (SSR) we can find out which one is better:

$$J(b_0, b_1) = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

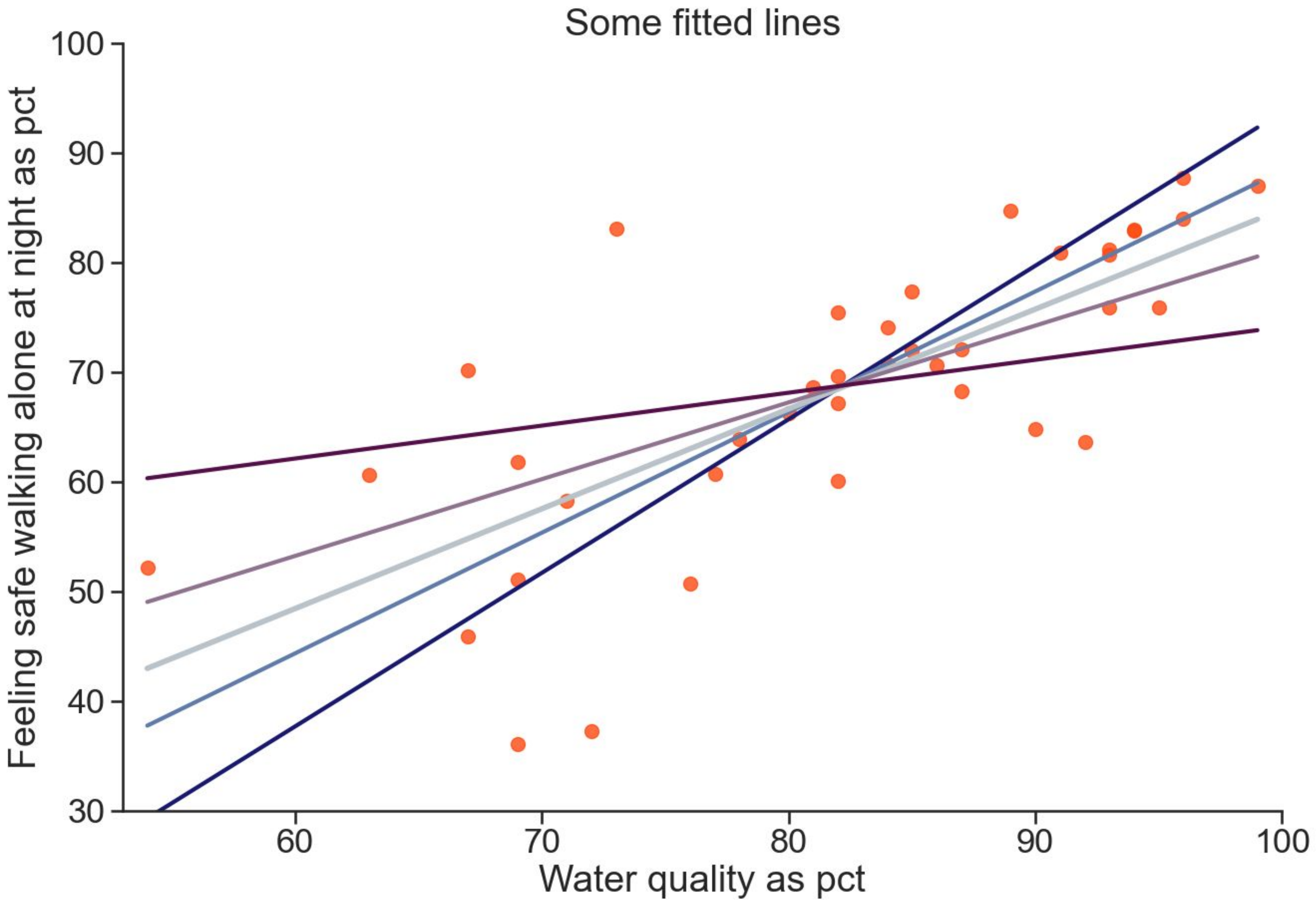


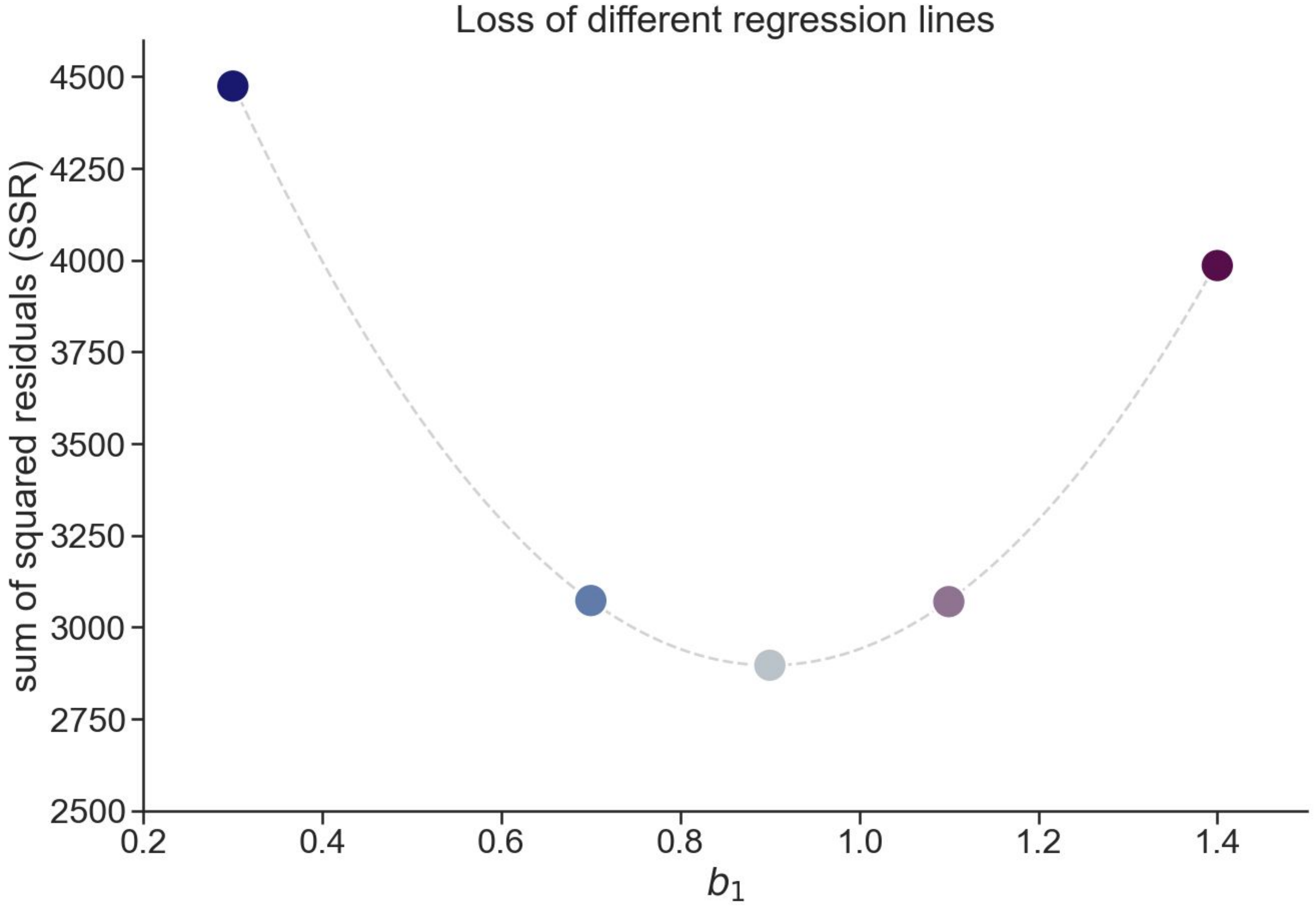


Trying out several fitted lines

By comparing the sum of squared residuals (SSR) we can find out which one is better:

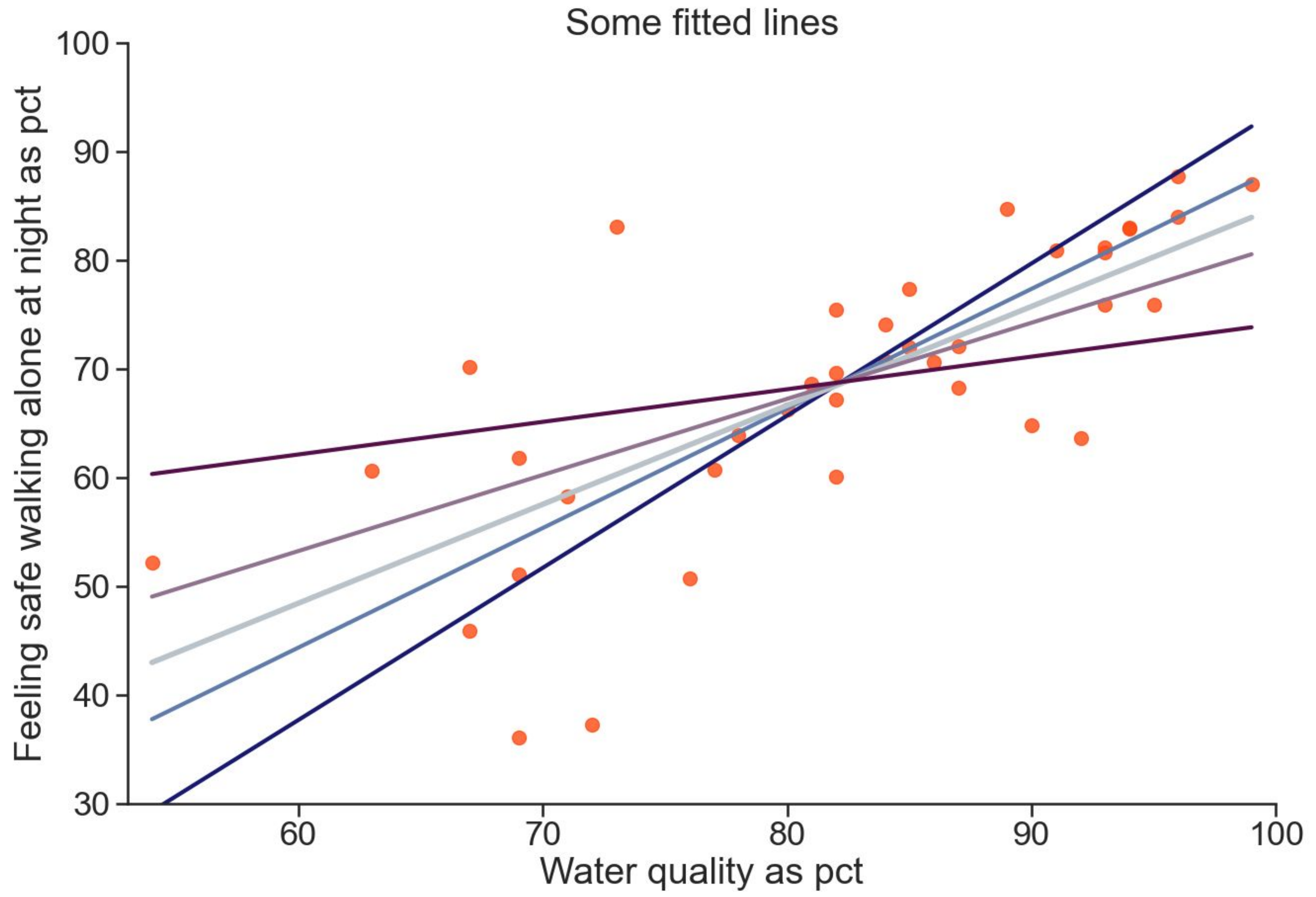
$$J(b_0, b_1) = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

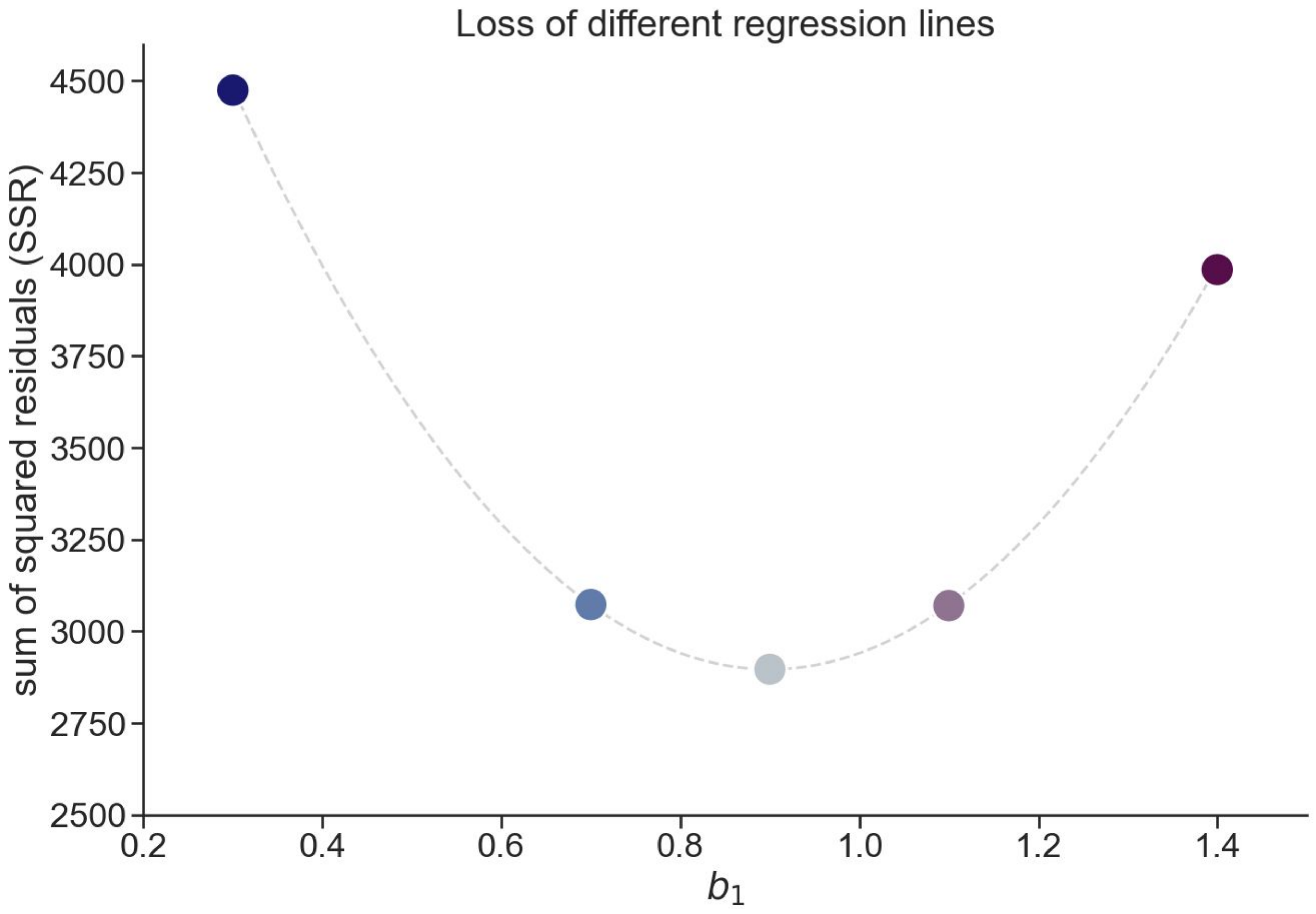




BUT THERE CAN BE AN INFINITE NUMBER OF LINES!

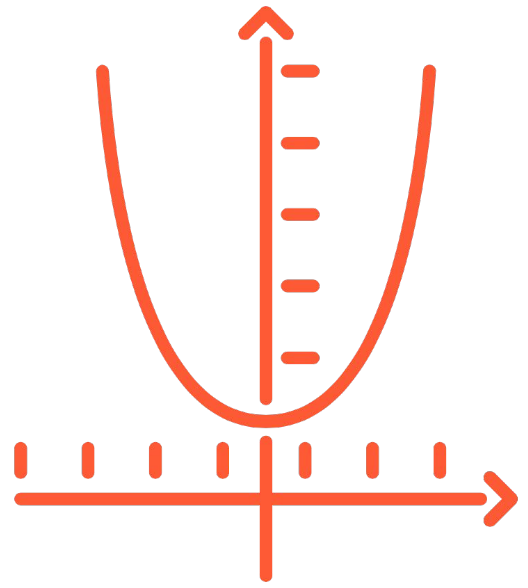
$$J(b_0, b_1) = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$





So how do we do this?

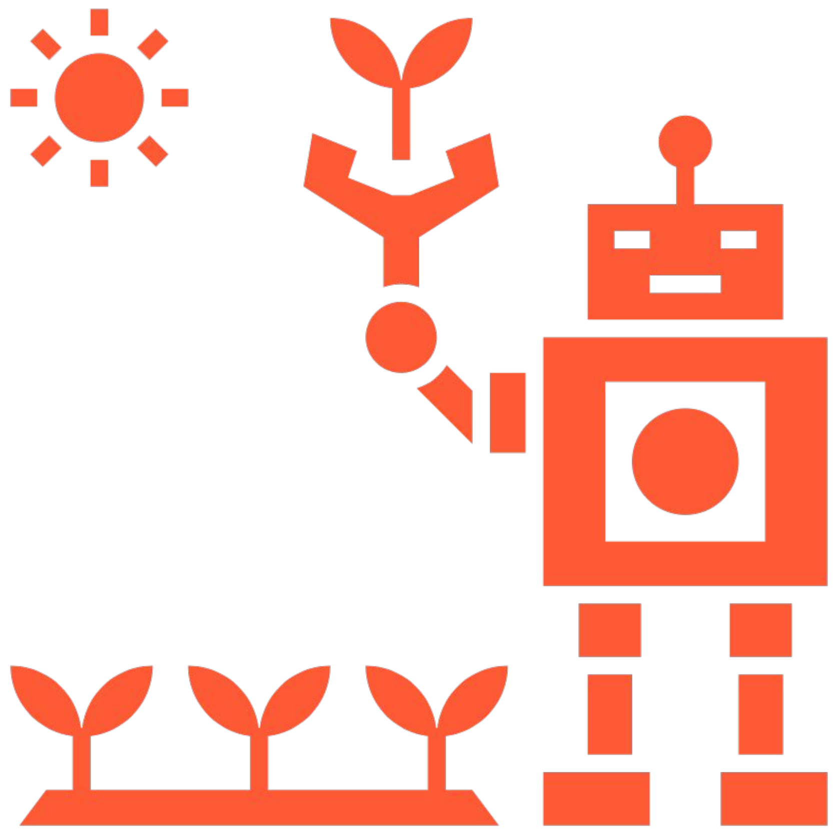
Obviously doing it manually is not really scalable



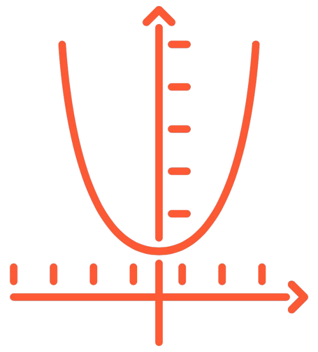
We minimize the OLS-function $J(b_0, b_1)$ with respect to b_0 and b_1 !

OLS - Ordinary Least Squares

$$J(b_0, b_1) = \sum (y_i - b_0 - b_1 x_i)^2$$



Ordinary least squares regression



$$\min\ J(b_0, b_1) = \sum (y_i - b_0 - b_1 x_i)^2$$

$$\frac{\partial J}{\partial b_0} = -2\,\Sigma(y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial J}{\partial b_1} = -2\ \Sigma x_i\,(y_i - b_0 - b_1 x_i) = 0$$

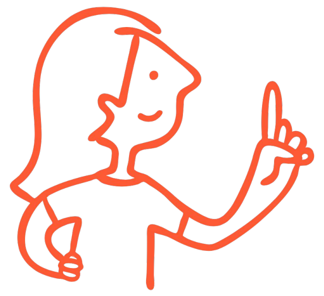
we divide the first equation by 2n:

$$-(\bar{y} - b_0 - b_1\bar{x}) = 0$$

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

... more math leads to:

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



the delta (or d) stands for first order derivative

Fun facts about residuals

$$y_i = b_0 + b_1 x_i + e_i$$

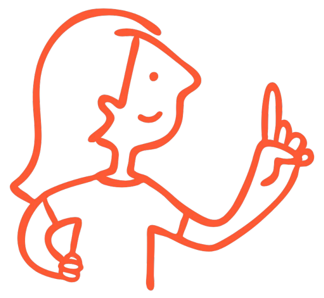
$$e_i = y_i - b_0 - b_1 x_i$$

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

Which leads to the following conclusions:

$$\Sigma e_i = 0$$

$$\Sigma (x_i - \bar{x})\, e_i = 0$$

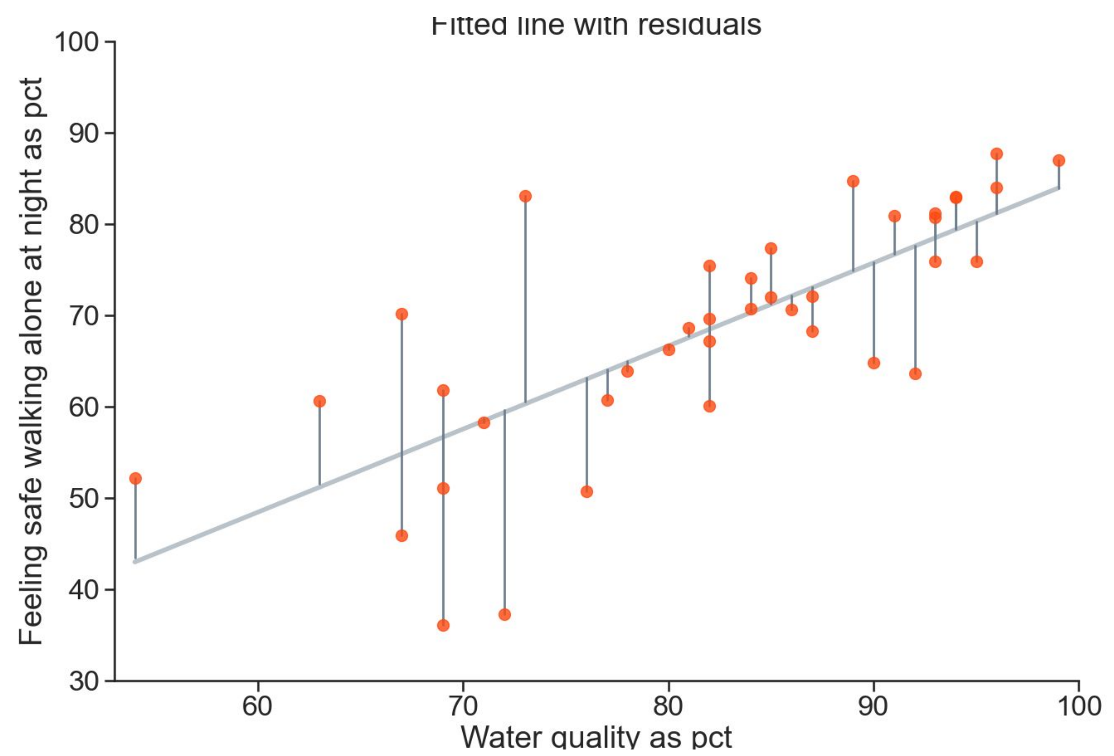


the second equation means the error/residual is uncorrelated with the explanatory variable

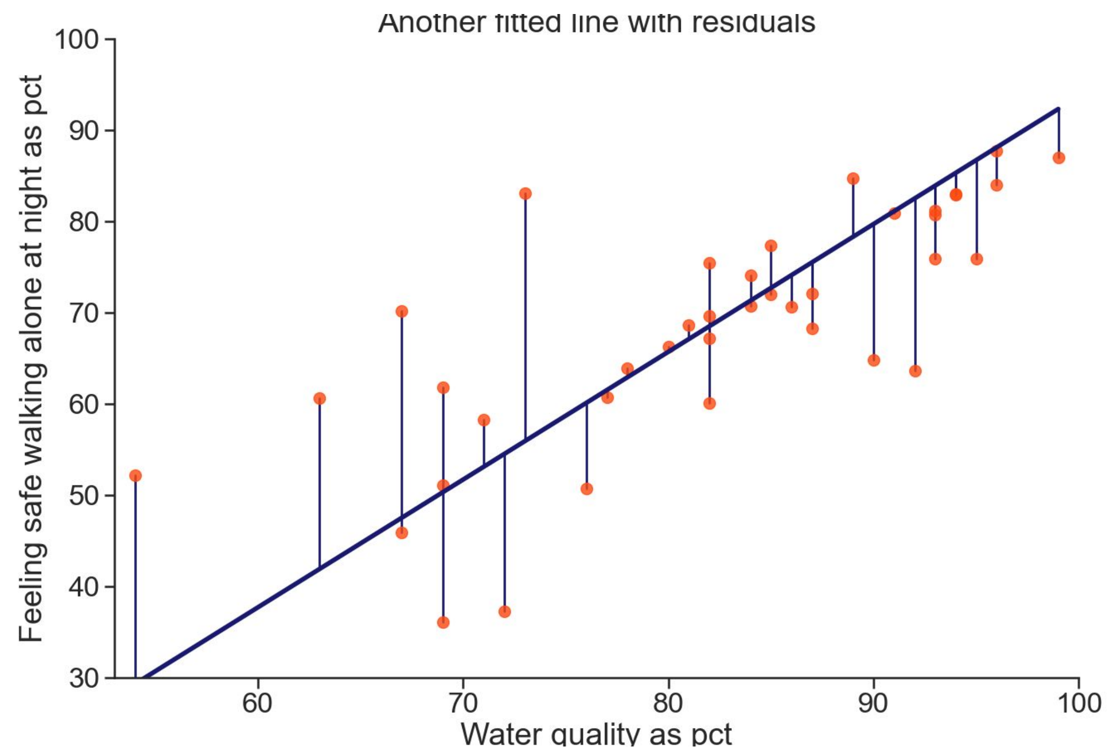
feel free to try this out for your models

Fun facts about residuals

$$y_i = b_0 + b_1 x_i + e_i$$
$$e_i = y_i - b_0 - b_1 x_i$$
$$b_0 = \bar{y} - b_1 \bar{x}$$



Σe = 0.0002

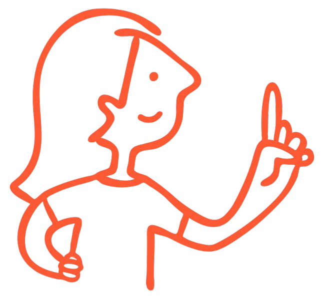


Σe = -8.2

Which leads to the following conclusions:

$$\Sigma e_i = 0$$

$$\Sigma(x_i - \bar{x})\, e_i = 0$$



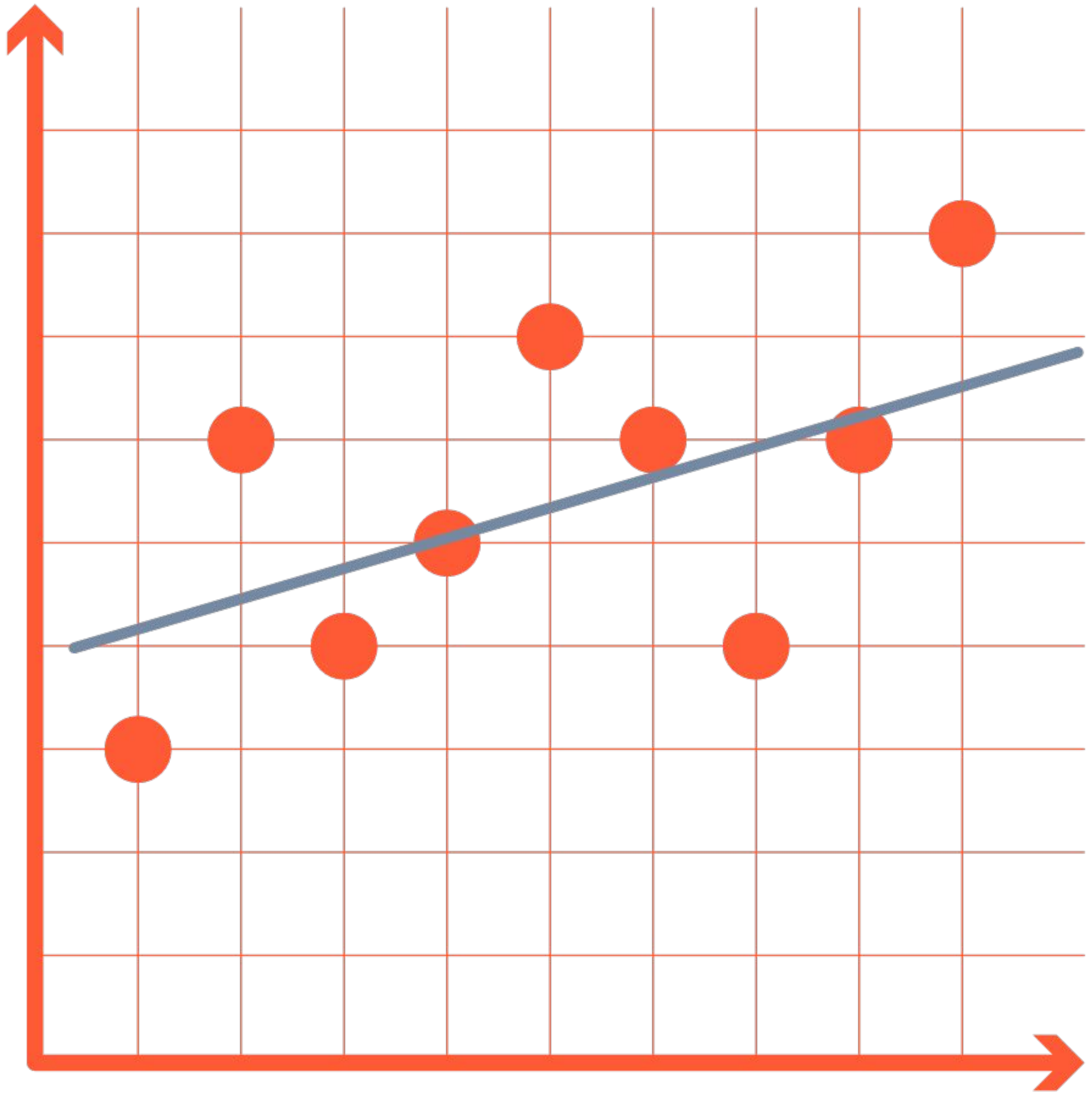
the second equation means the error/residual is uncorrelated with the explanatory variable

feel free to try this out for your models

Linear Regression

Part 4
Performance Metrics

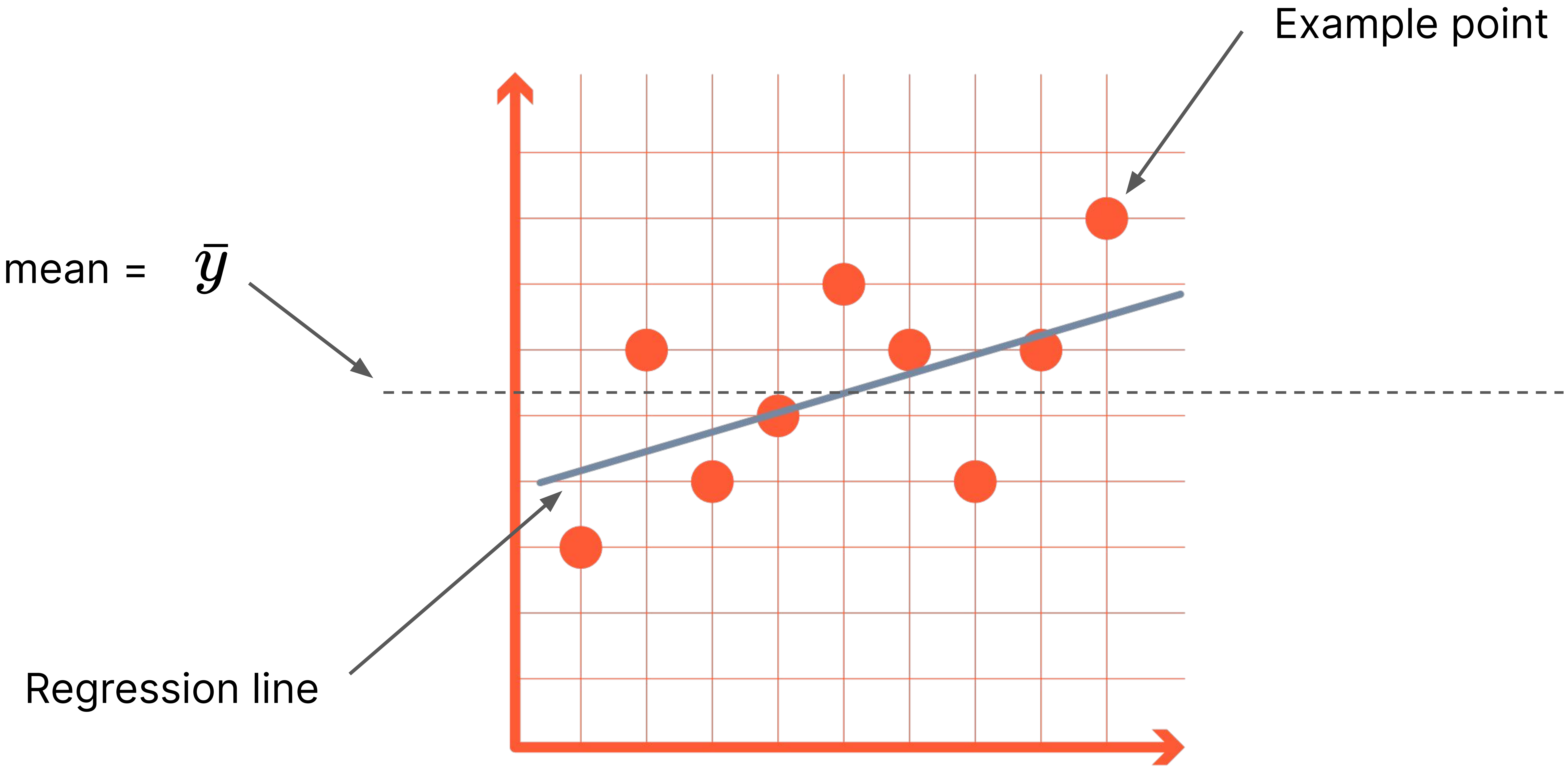
Sum of various squares (variance analysis)

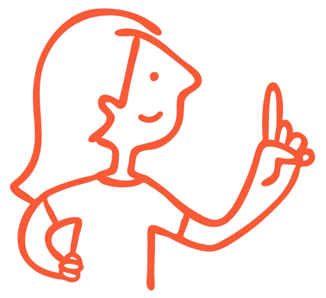


Fun Fact: the names are ridiculously stupid

SST = the total sum of squares
SSE = the explained sum of squares
SSR = the remaining sum of squares

Sum of various squares (variance analysis)

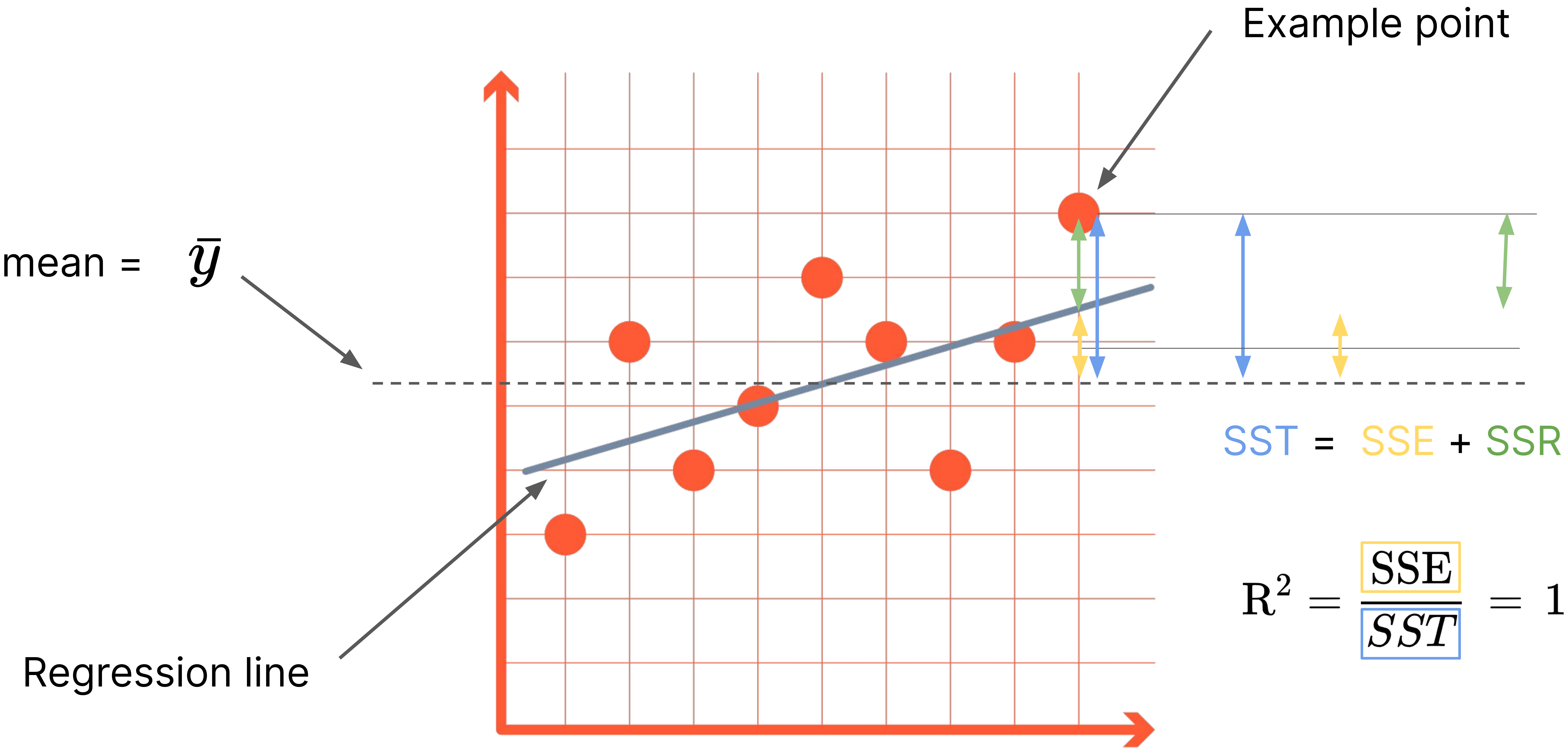




Fun Fact: the names are ridiculously stupid

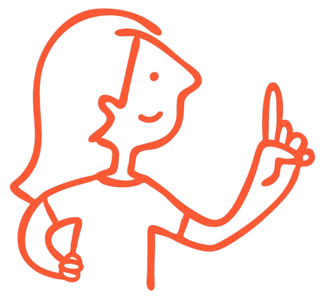
SST = the total sum of squares
SSE = the explained sum of squares
SSR = the remaining sum of squares

Sum of various squares (variance analysis)



SST = SSE + SSR

$$R^2 = \frac{\boxed{\mathrm{SSE}}}{\boxed{SST}} = 1 - \frac{\boxed{SSR}}{\boxed{SST}}$$



Fun Fact: the names are ridiculously stupid

SST = the total sum of squares
SSE = the explained sum of squares
SSR = the remaining sum of squares

Coefficient of determination: R^2

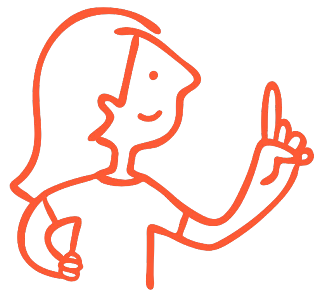
All the square sums depend on the scale of measurement of y.
We need a performance measure that is independent of scale
... enters the *coefficient of determination*

$$R^2 = \frac{SSE}{SST} = \frac{b_1^2 \Sigma (x_i - \bar{x})^2}{\Sigma (y_i - \bar{y})^2}$$

or:

$$R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma (y_i - \bar{y})^2}$$

being scale dependent means: that they could be cents, kms, meters, lots of meters, lots of money, depending on your problem.. thus you would always need to talk about the scale of y to put things into perspective.



$0 \leq R^2 \leq 1$

least squares criterion ~ maximizing R2

$R^2 = r^2$ (you know.. the Pearson correlation coefficient)

Sum of various squares

A traditional way to measure performance is to compare the **SSR** to the

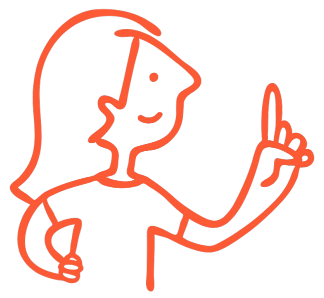
sum of squares of deviation of y:

$$y_i = b_0 + b_1 x_i + e_i$$
$$b_0 = \bar{y} - b_1 \bar{x}$$

this leads to the following conclusions:

$$y_i - \bar{y} = b_1(x_i - \bar{x}) + e_i$$
$$\Sigma(y_i - \bar{y})^2 = b_1^2 \Sigma(x_i - \bar{x})^2 + \Sigma e_i^2$$

SST = SSE + SSR



SST = the total sum of squares
SSE = the explained sum of squares
SSR = the remaining sum of squares

Coefficient of determination: R^2

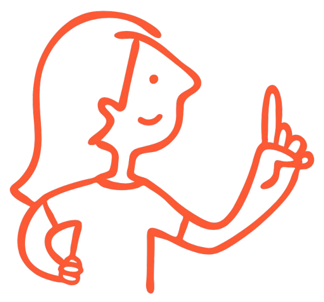
All the square sums depend on the scale of measurement of y.
We need a performance measure that is independent of scale
... enters the *coefficient of determination*

$$R^2 = \frac{SSE}{SST} = \frac{b_1^2 \Sigma(x_i - \bar{x})^2}{\Sigma(y_i - \bar{y})^2}$$

or:

$$R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma(y_i - \bar{y})^2}$$

being scale dependent means: that they could be cents, kms, meters, lots of meters, lots of money, depending on your problem.. thus you would always need to talk about the scale of y to put things into perspective.



$0 \leq R^2 \leq 1$

least squares criterion ~ maximizing R2

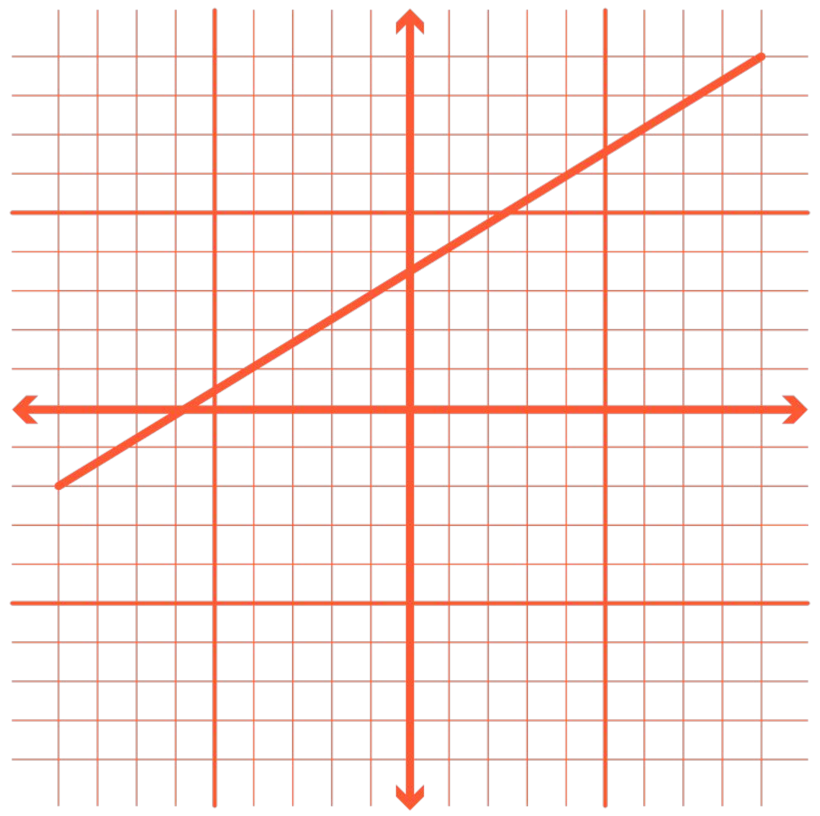
$R^2 = r^2$ *(you know.. the Pearson correlation coefficient)*

neue fische
School and Pool for Digital Talent

Linear Regression

Part 5
Key Terms

Key terms: Machine learning



Variables:

- Target (dependent variable, response, y)
- Feature (independent variable, explanatory variable, attribute, X)
- Observation (row, instance, example)

Model:

- Fitted values (predicted values) - denoted with the hat notation $\hat{y}$
- Residuals (errors, e)
- Least squares (method for fitting a regression)
- Coefficients (parameters)

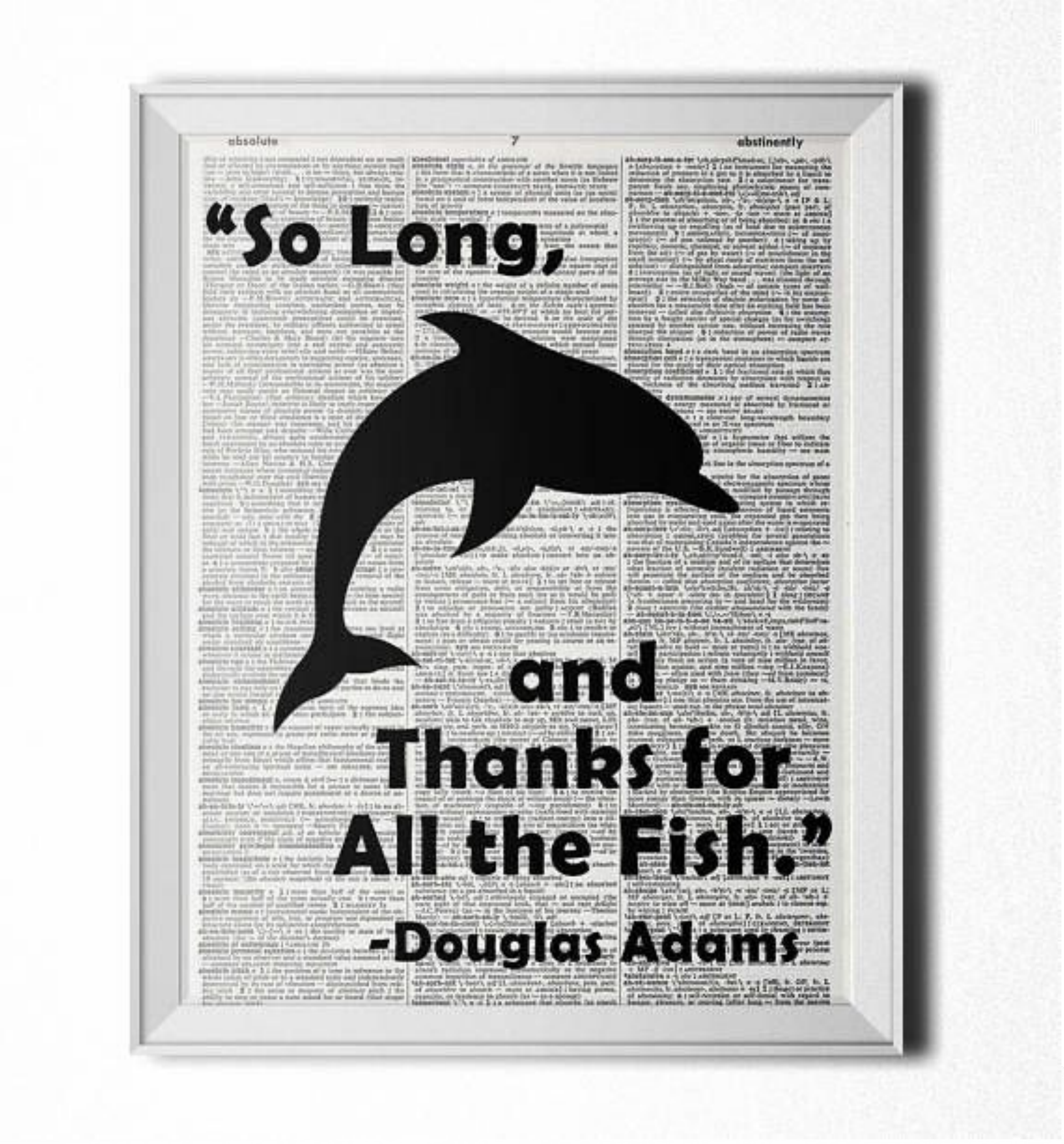
References

[*Practical Statistics for Data Science*](https://) - Peter Bruce & Andrew Bruce

Econometric Methods with Applications in Business and Economics - Christiaan Heij, Paul de Boer, Philip Hans Franses, Teun Kloek, Herman K. van Dijk

https://learningstatisticswithr.com/book/regression.html

https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp



calculate b1:

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

$$-2\,\Sigma\, x_i(y_i - b_0 - b_1 x_i) = 0$$

$$-2\,\Sigma\, x_i(y_i - \bar{y} + b_1\bar{x} - b_1 x_i) = 0$$

$$\Sigma\,(x_i y_i - x_i\bar{y} + b_1(x_i\bar{x} - x_i x_i)) = 0$$

$$\Sigma\,(x_i y_i - 2x_i\bar{y} + \bar{x}\bar{y} + b_1(-\bar{x}\bar{x} + 2x_i\bar{x} - x_i x_i)) = 0 \quad \Big| \quad \Sigma x_i = \Sigma\bar{x}$$

$$\Sigma\,(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y} + b_1(-\bar{x}\bar{x} + 2x_i\bar{x} - x_i x_i)) = 0 \quad \Big| \quad \Sigma x_i\bar{y} = \Sigma\bar{x}y_i = n\bar{x}\bar{y}$$

$$\Sigma(y_i - \bar{y})(x_i - \bar{x}) - b_1\Sigma(x_i - \bar{x})^2 = 0$$

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