Implementing Predictive Models for Continuous Data

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Regression to predict continuous variables

Simple and multiple regression

Multicollinearity and risks in regression

R-square and adjusted R-square

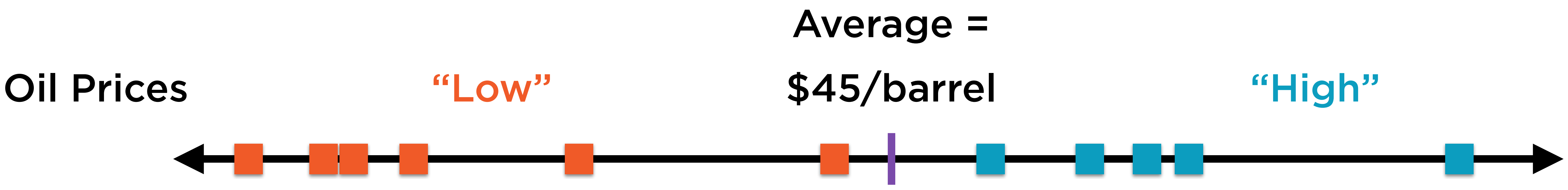
Selecting features for regression using statistical techniques

Linear Regression

Data in One Dimension

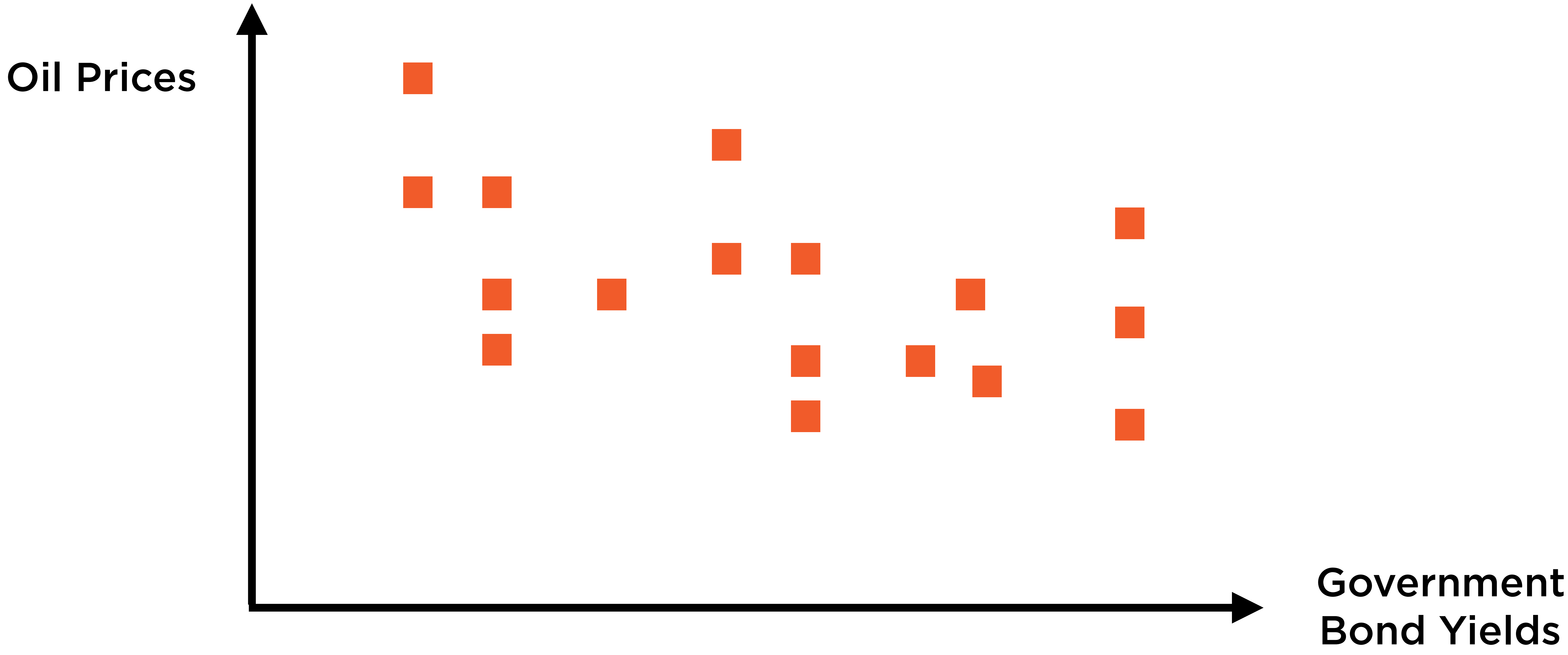
Unidimensional data points can be represented using a line, such as a number line

Data in One Dimension



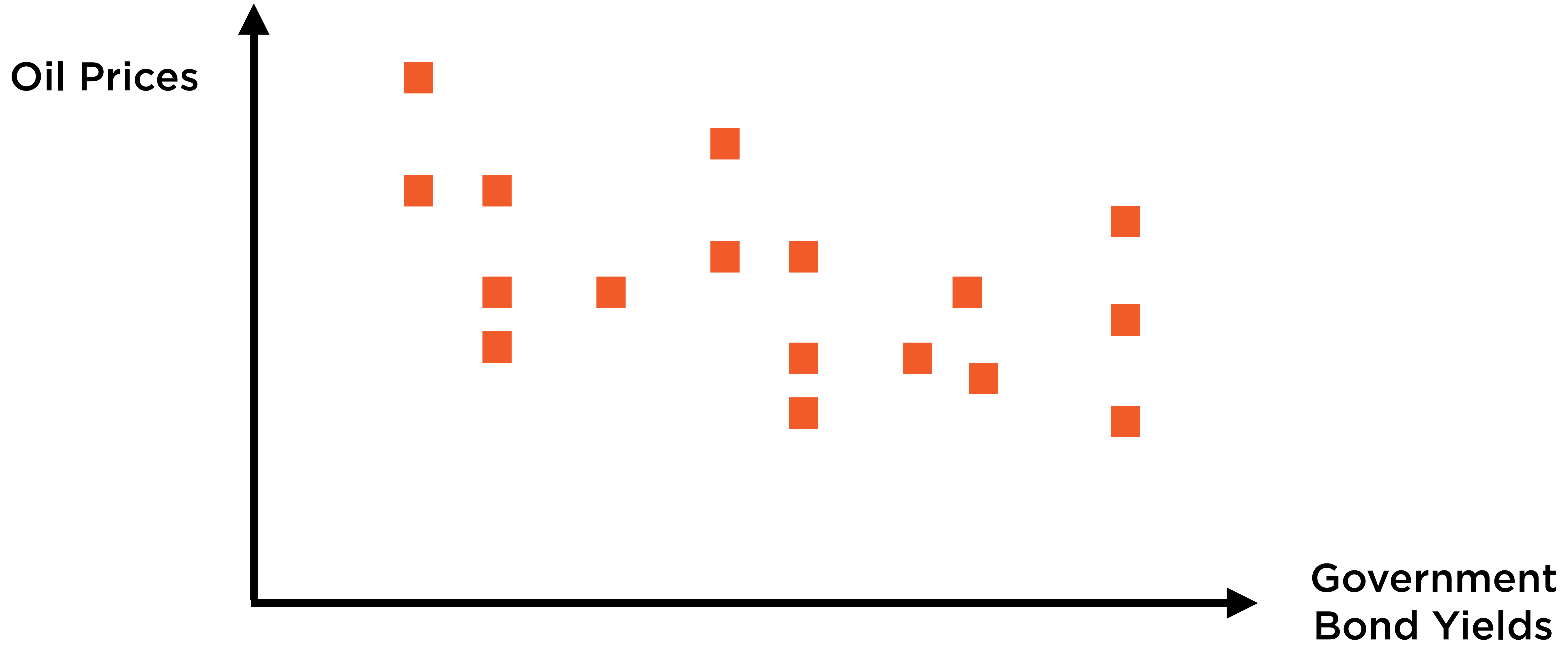
Unidimensional data is analysed using statistics such as mean, median, standard deviation

Data in Two Dimensions



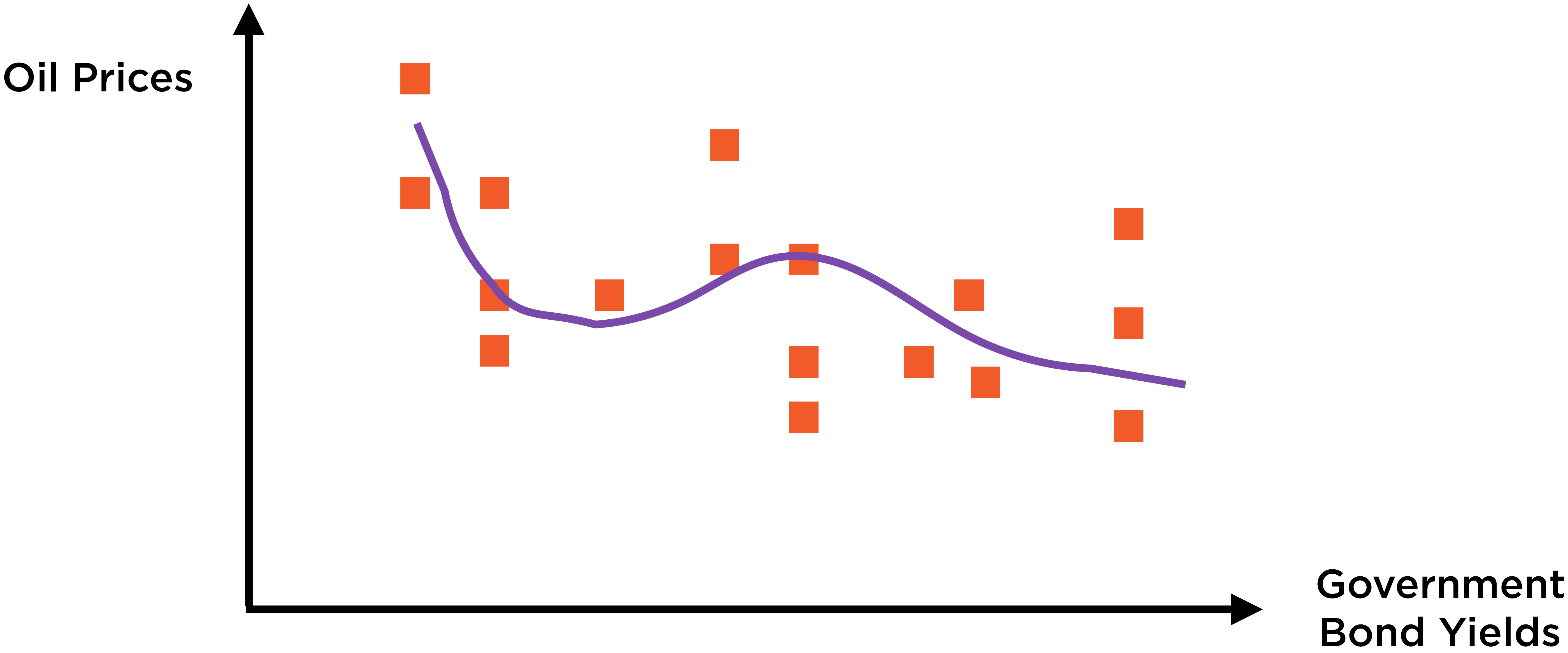
It's often more insightful to view data in relation to some other, related data

Data in Two Dimensions



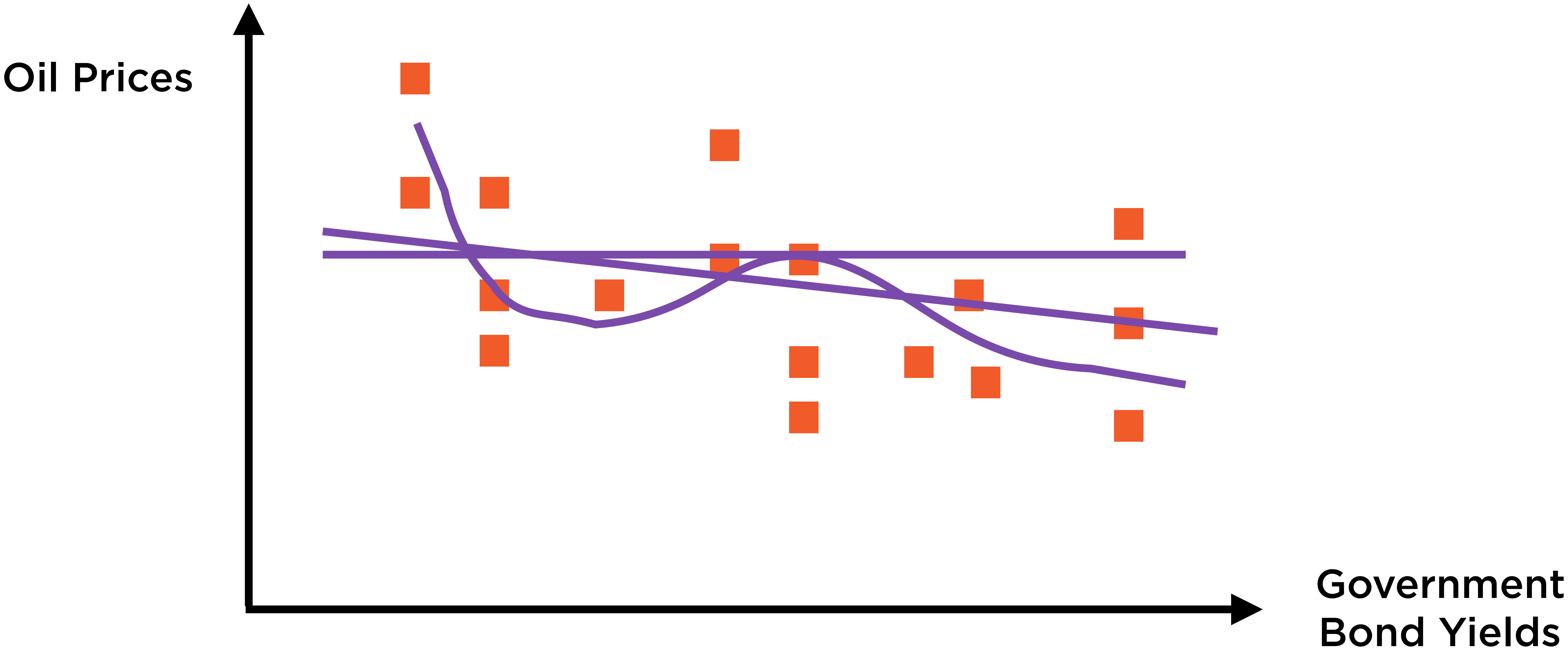
Bidimensional data can be represented in a plane

Data in Two Dimensions



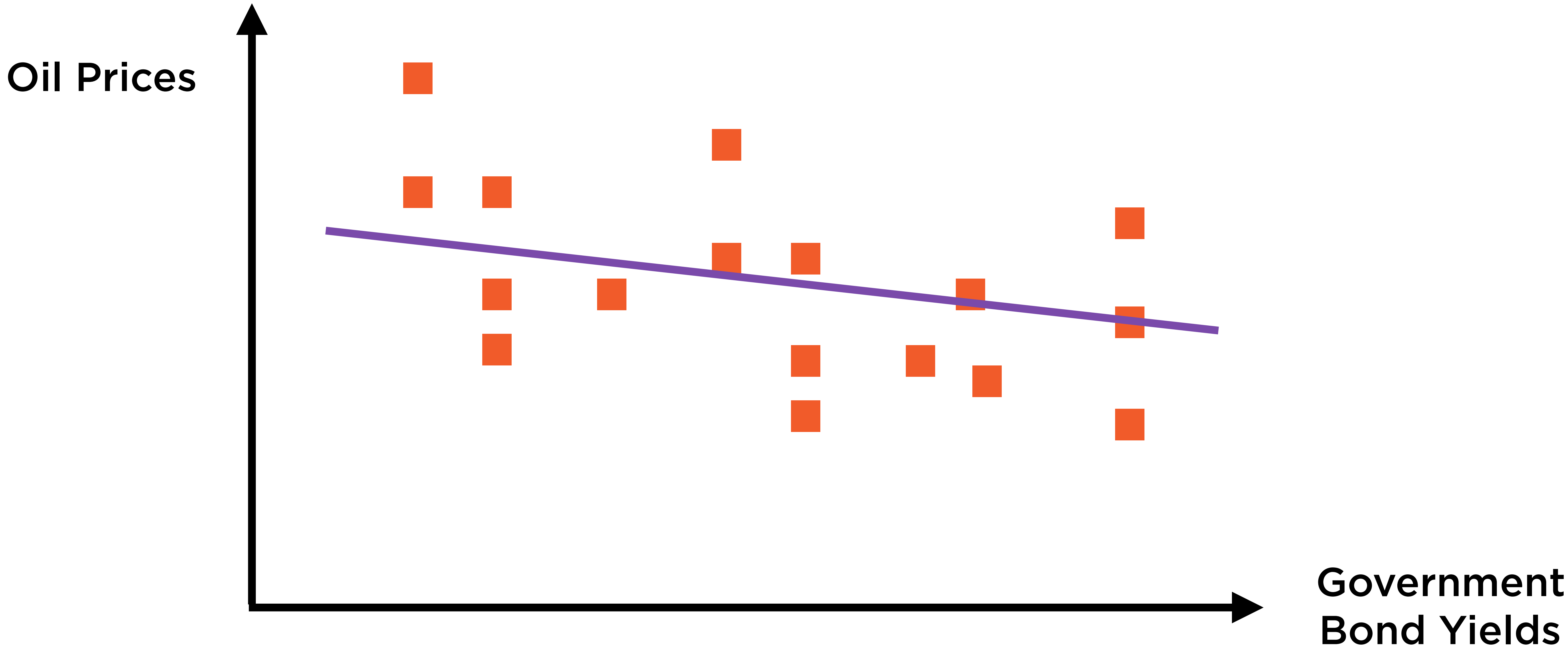
We can draw any number of curves to fit such data

Data in Two Dimensions



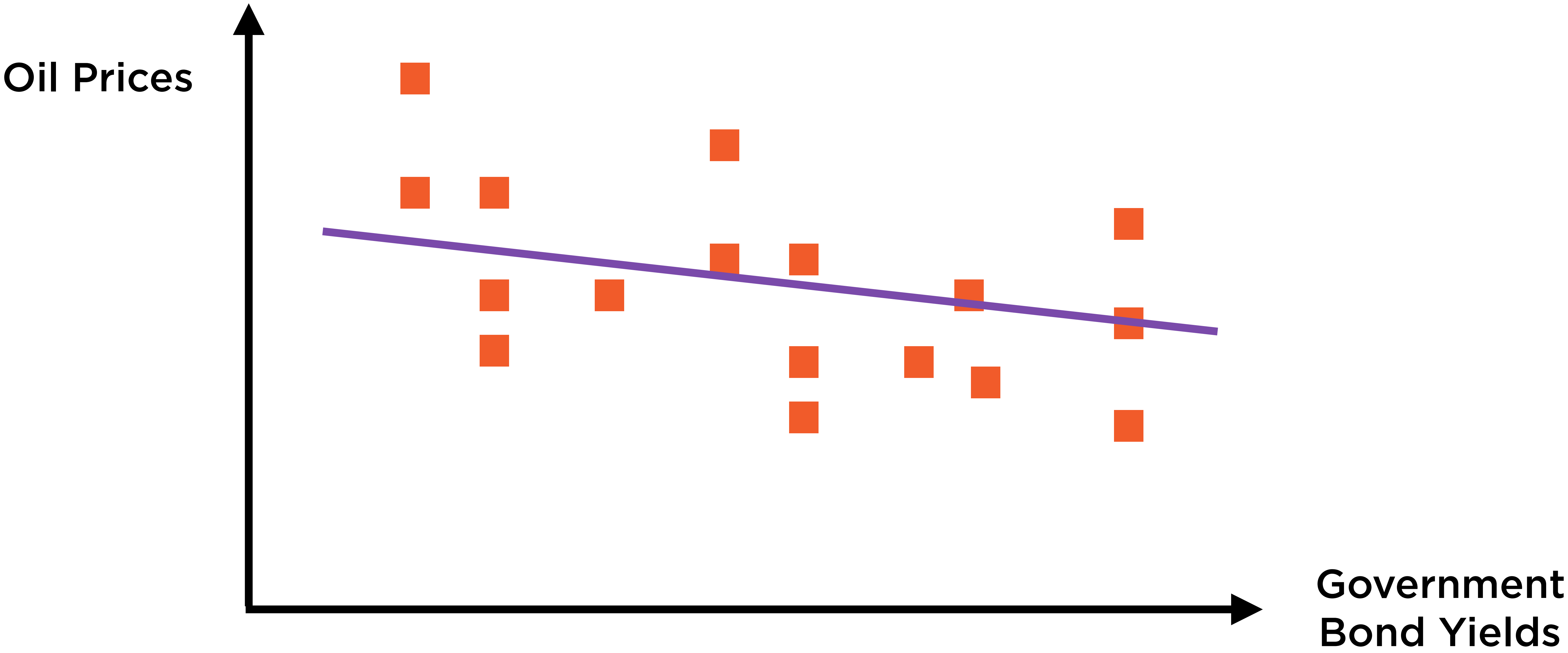
We can draw any number of curves to fit such data

Data in Two Dimensions



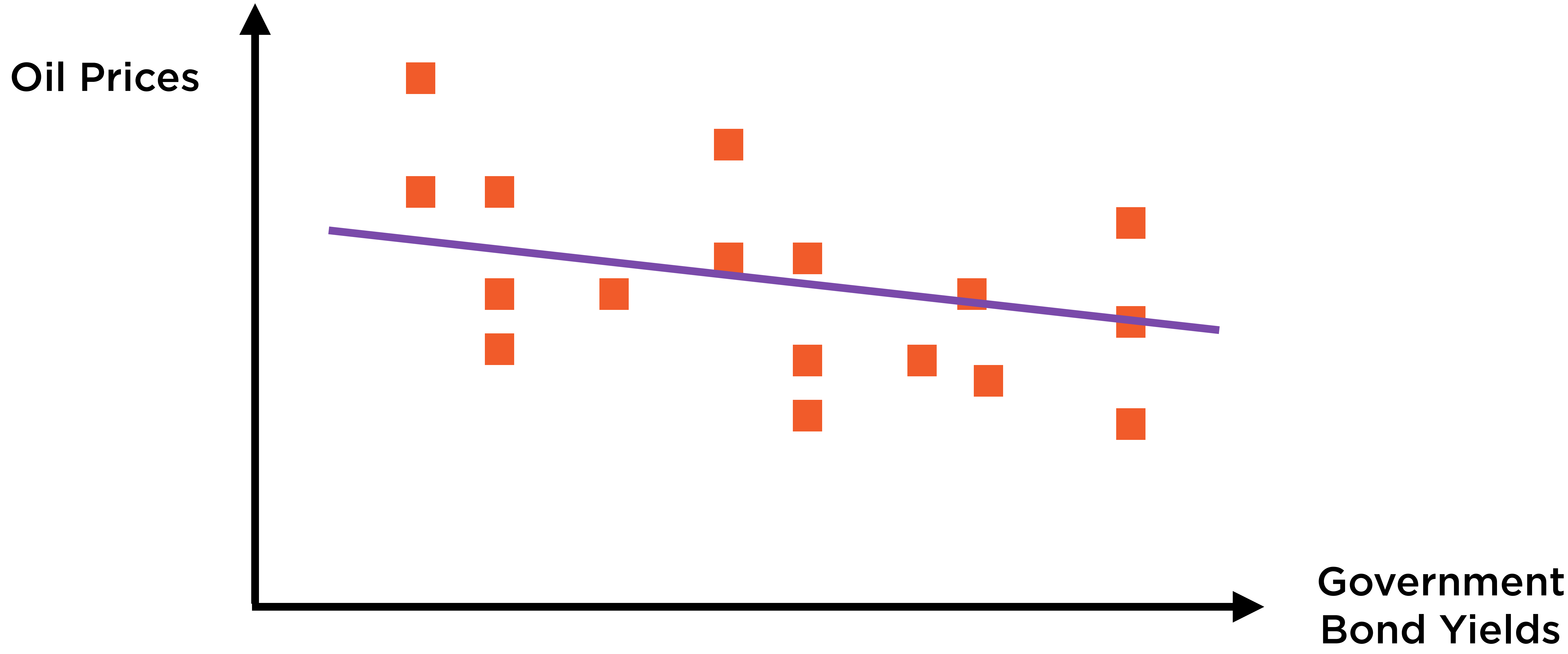
A straight line represents a linear relationship

Data in Two Dimensions



Finding the "best" such straight line is called Linear Regression

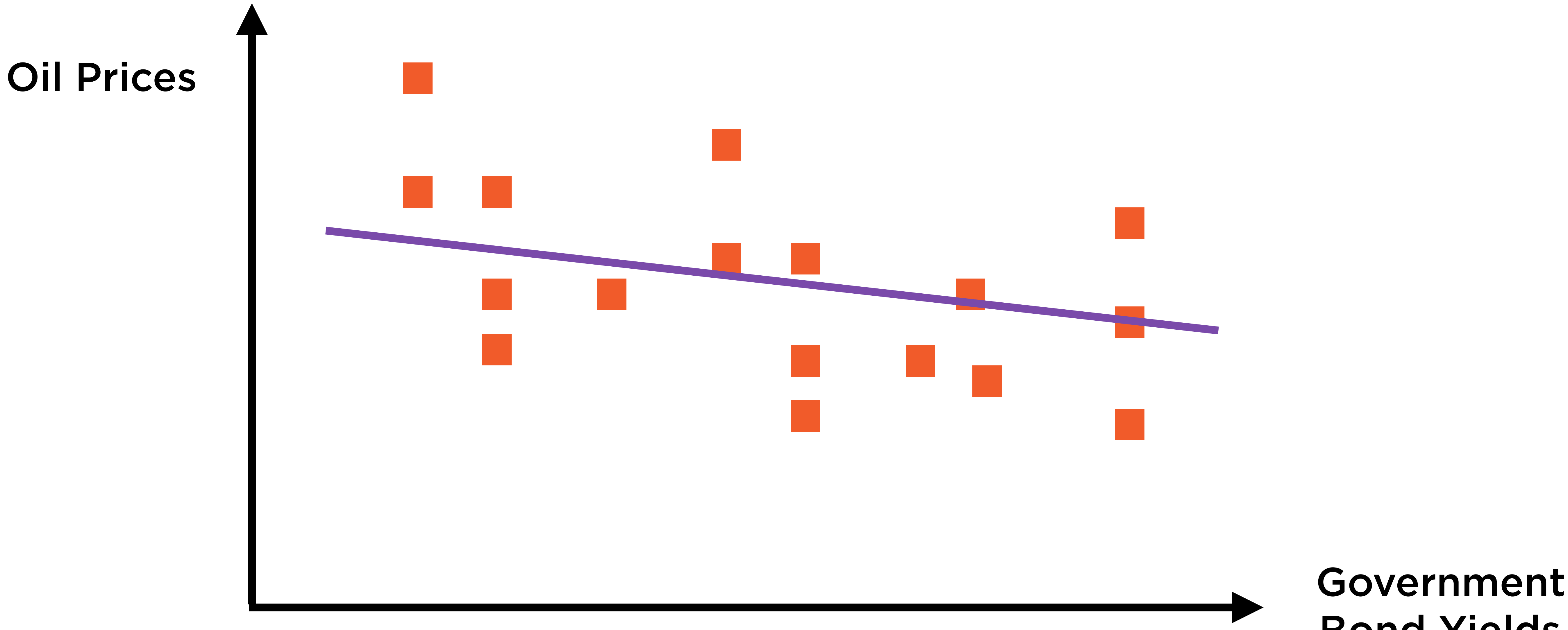
Linear Regression



The linear regression relationship can be expressed as

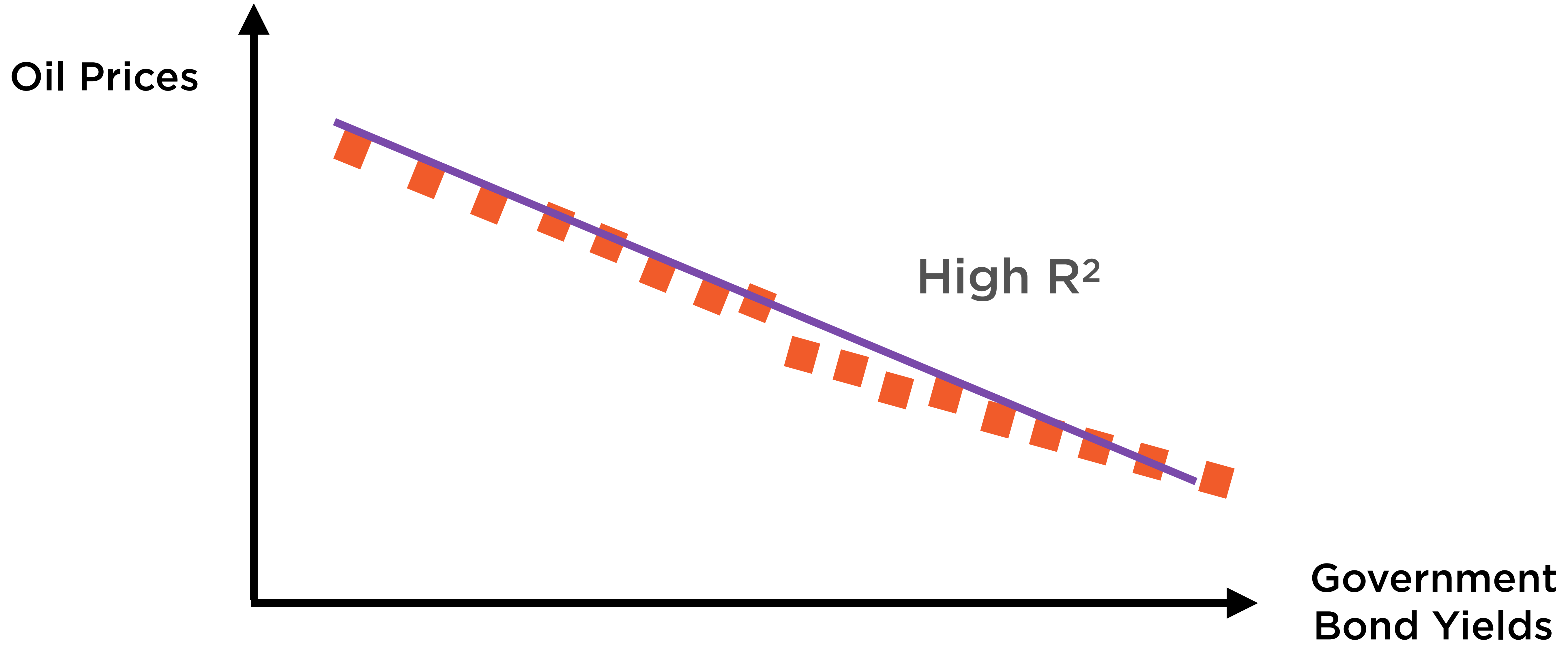
$y = A + Bx$

Linear Regression



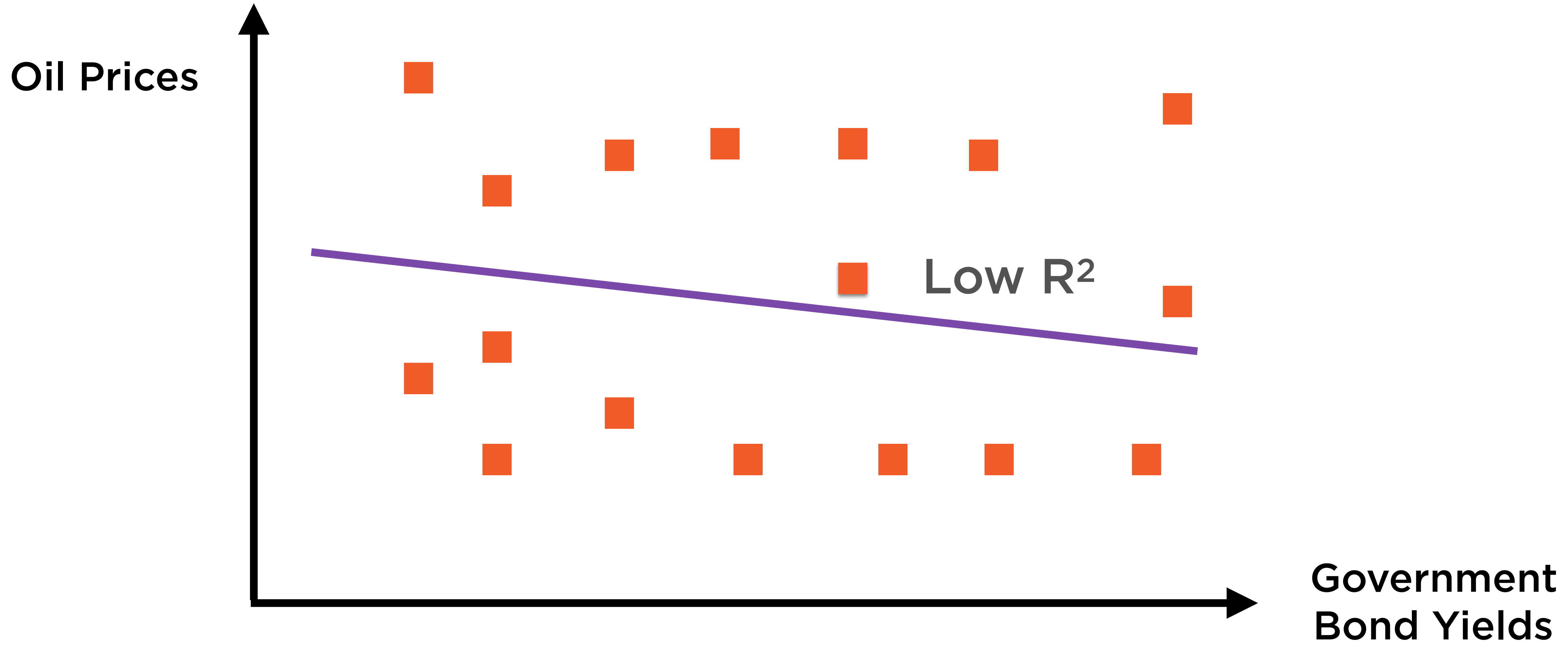
Regression not only gives us the equation of this line, it also signals how reliable the line is

Linear Regression



High quality of fit

Linear Regression

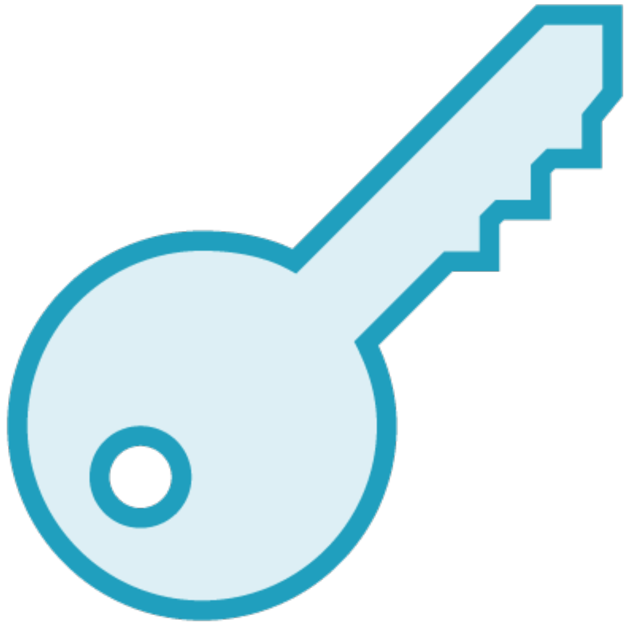


Low quality of fit

R^2 is a measure of how well the linear regression fits the underlying data

Setting Up The Regression Problem

X Causes Y



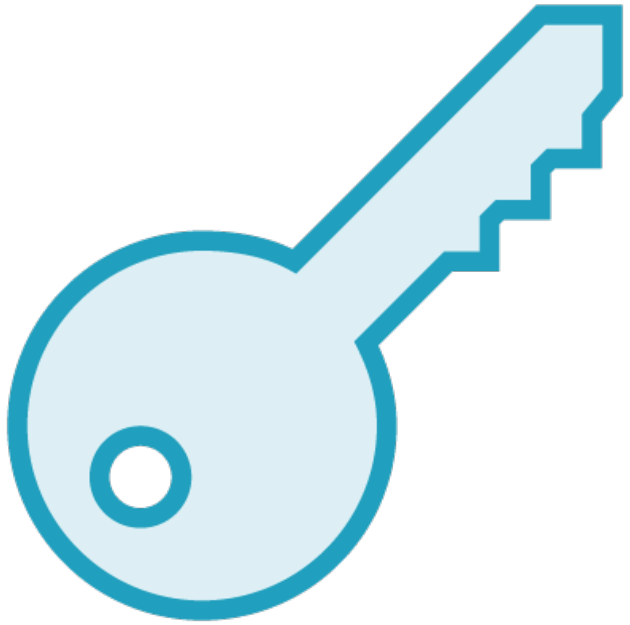
Cause

Independent variable

Effect

Dependent variable

X Causes Y



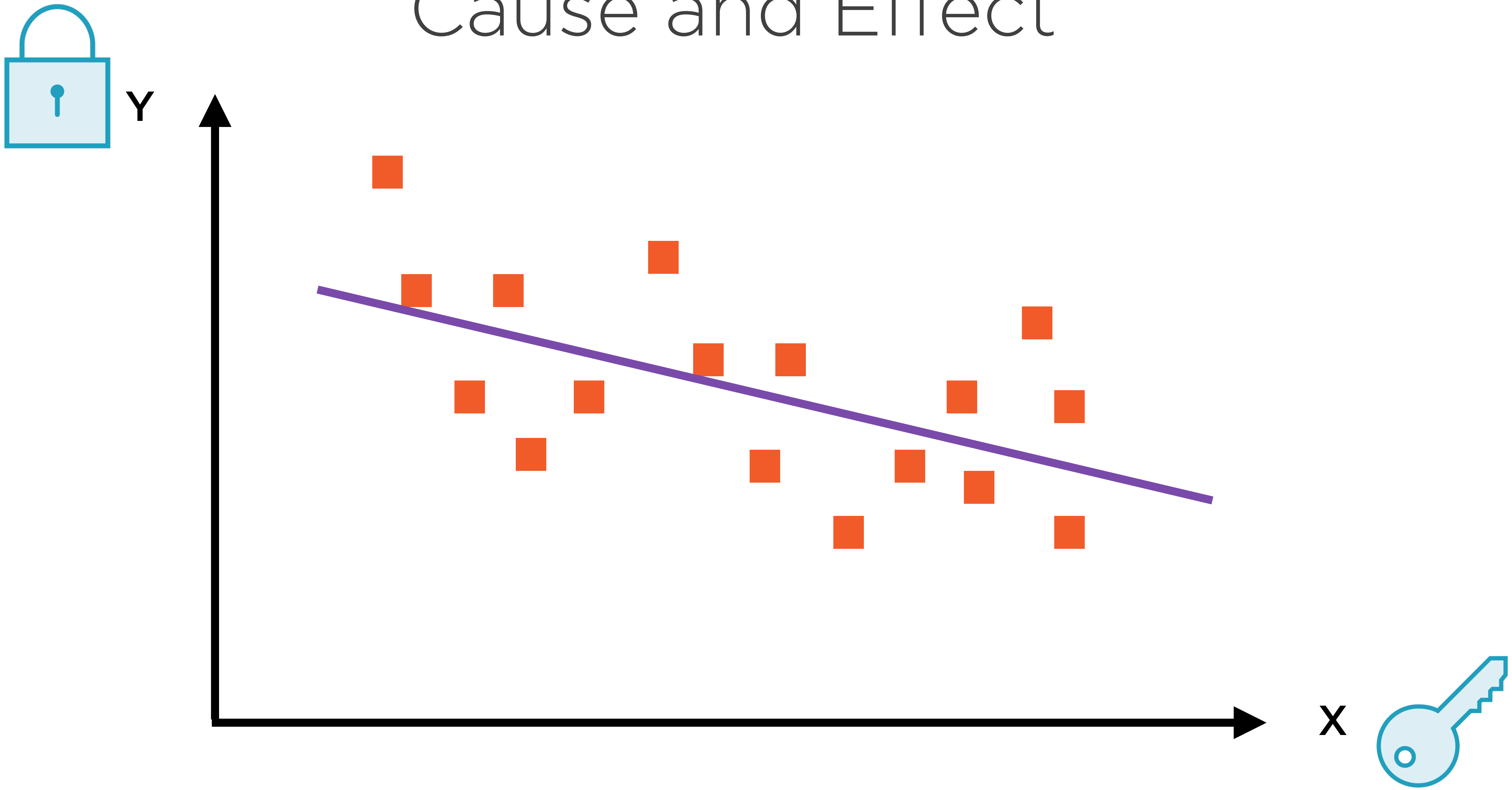
Cause

Explanatory variable

Effect

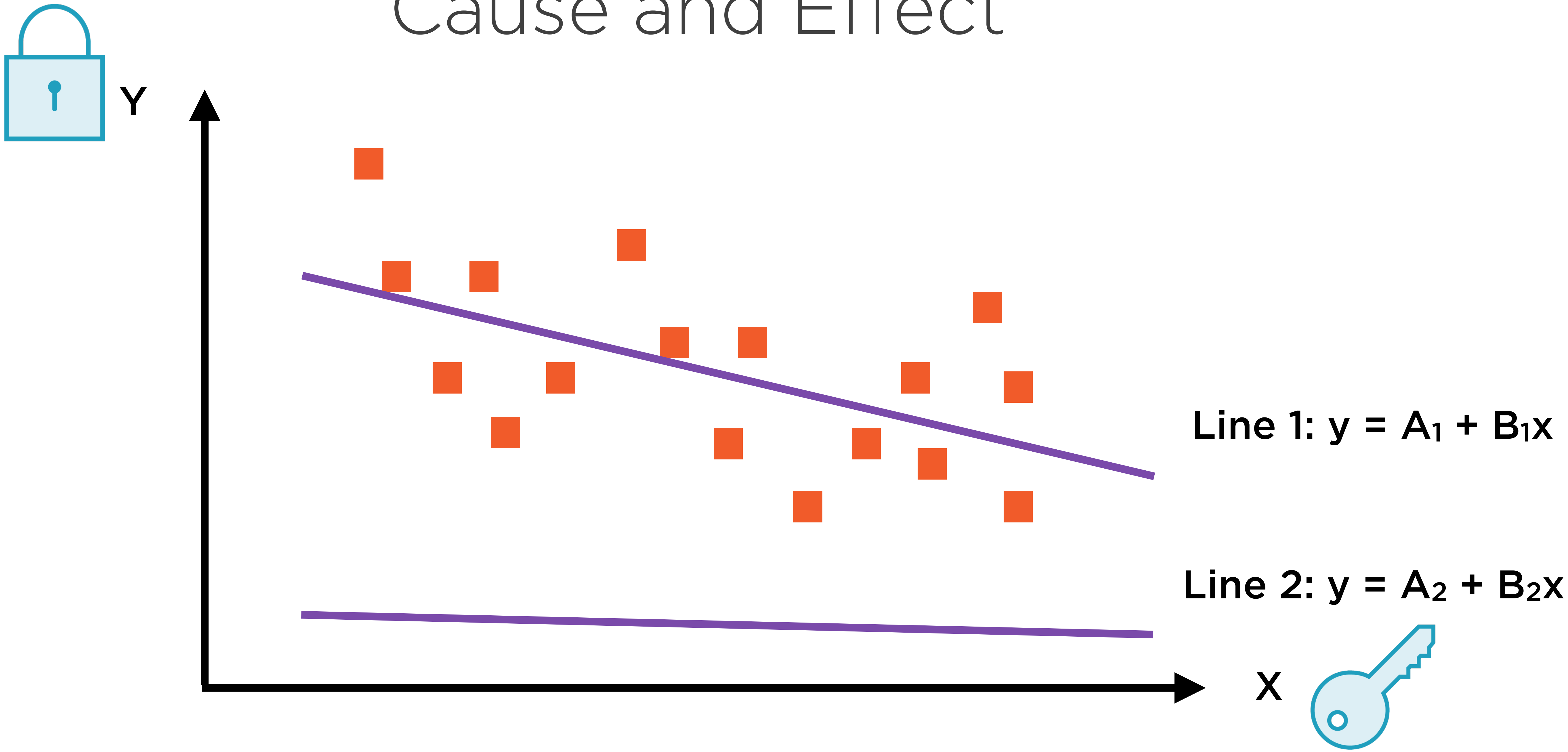
Dependent variable

Cause and Effect



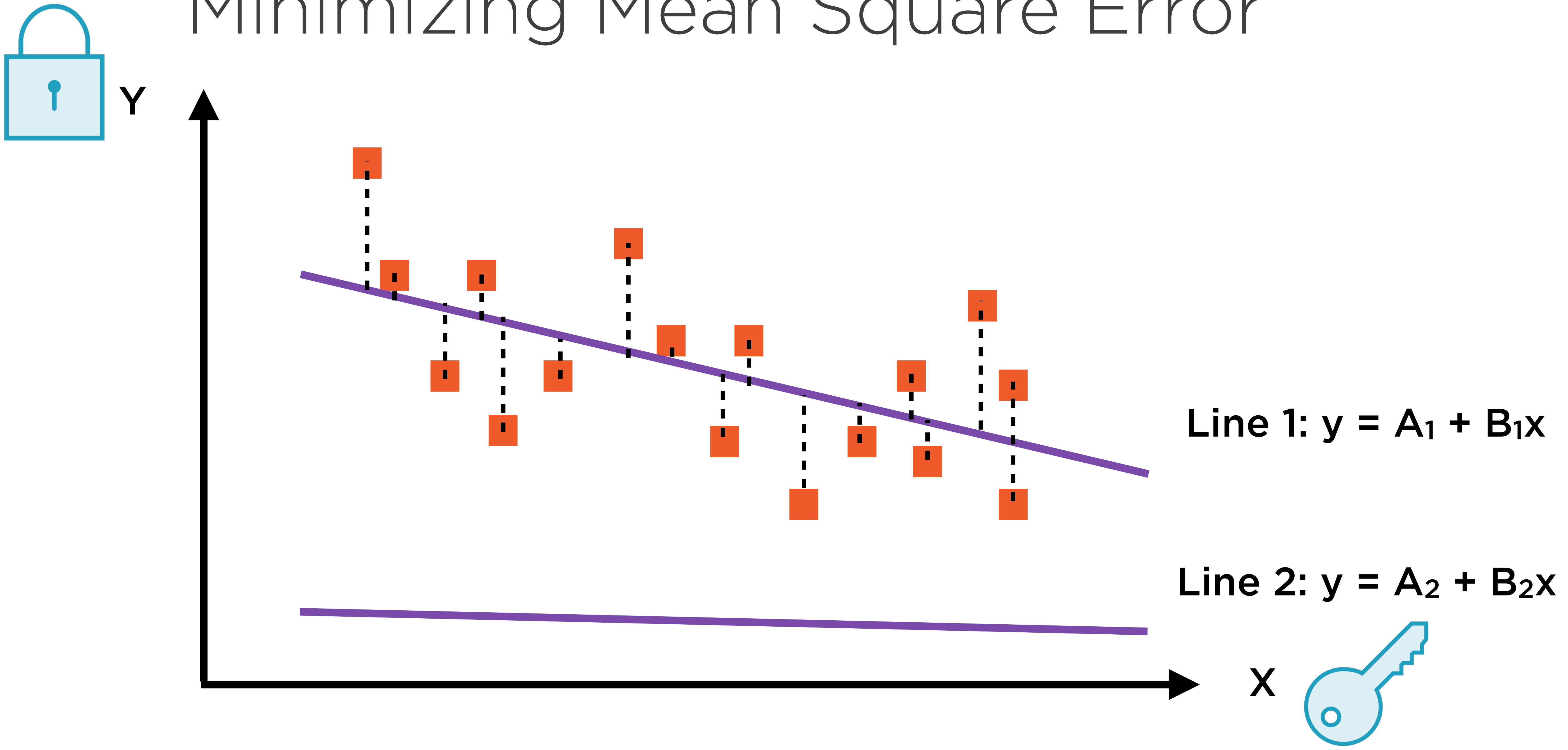
Linear Regression involves finding the “best fit” line

Cause and Effect

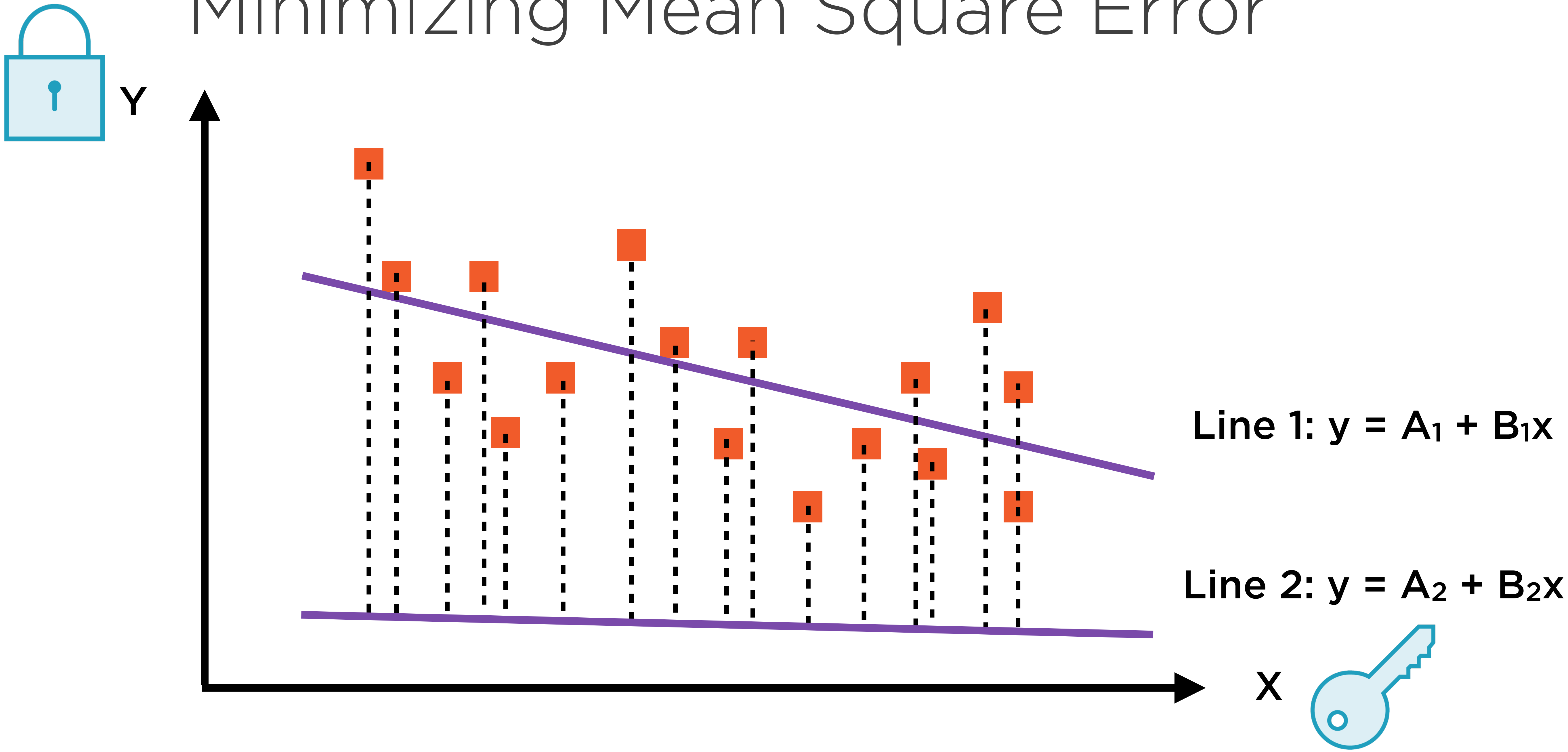


Which of these lines is a better fit?

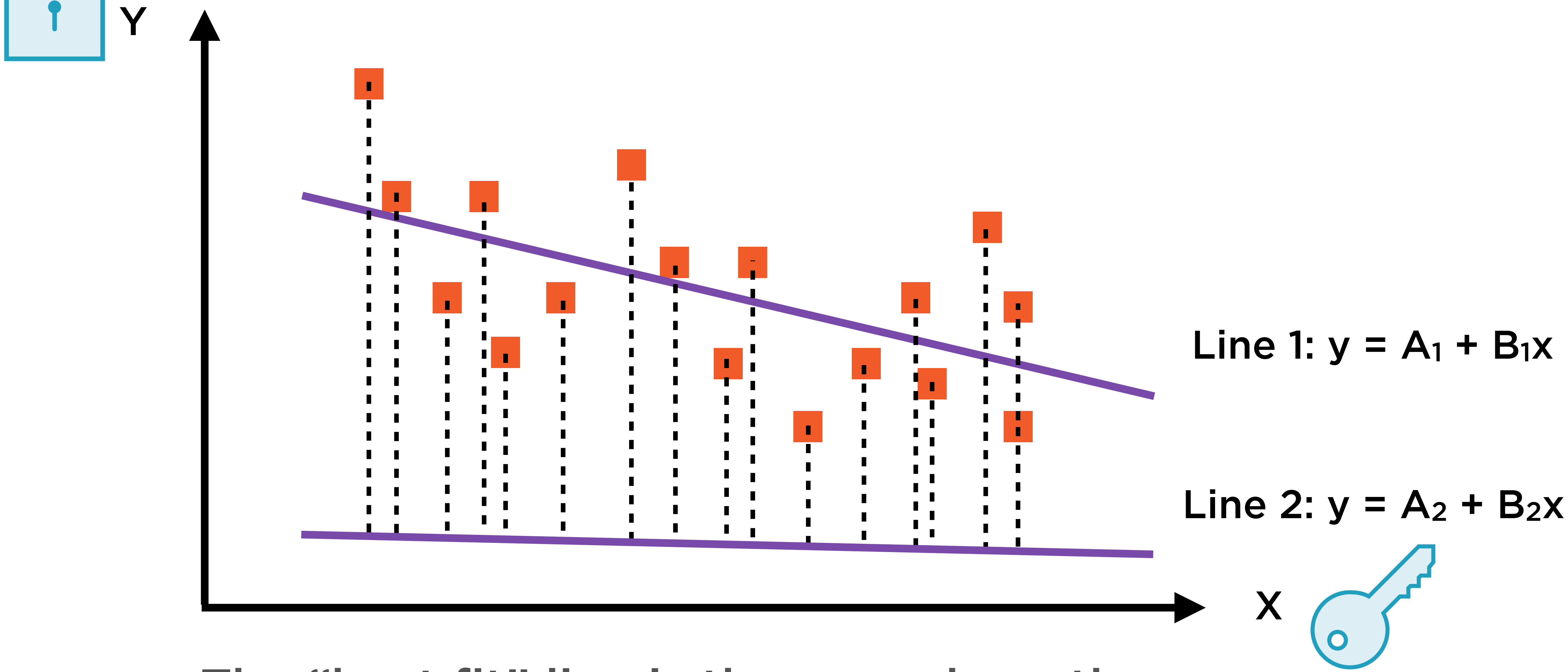
Minimizing Mean Square Error



Minimizing Mean Square Error



Minimizing Mean Square Error



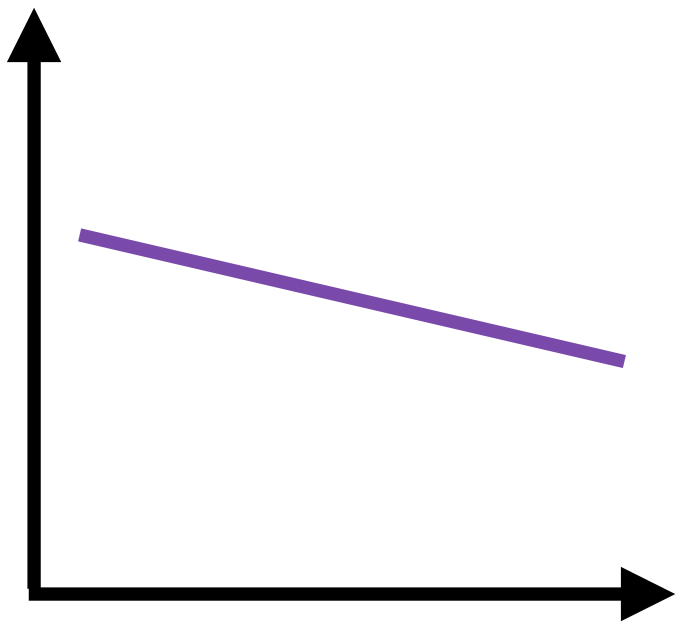
The "best fit" line is the one where the sum of the squares of the lengths of these dotted lines is minimum

The “best fit” line is the one where the sum of the squares of the lengths of the errors is minimized

Finding this line is the objective of the regression problem

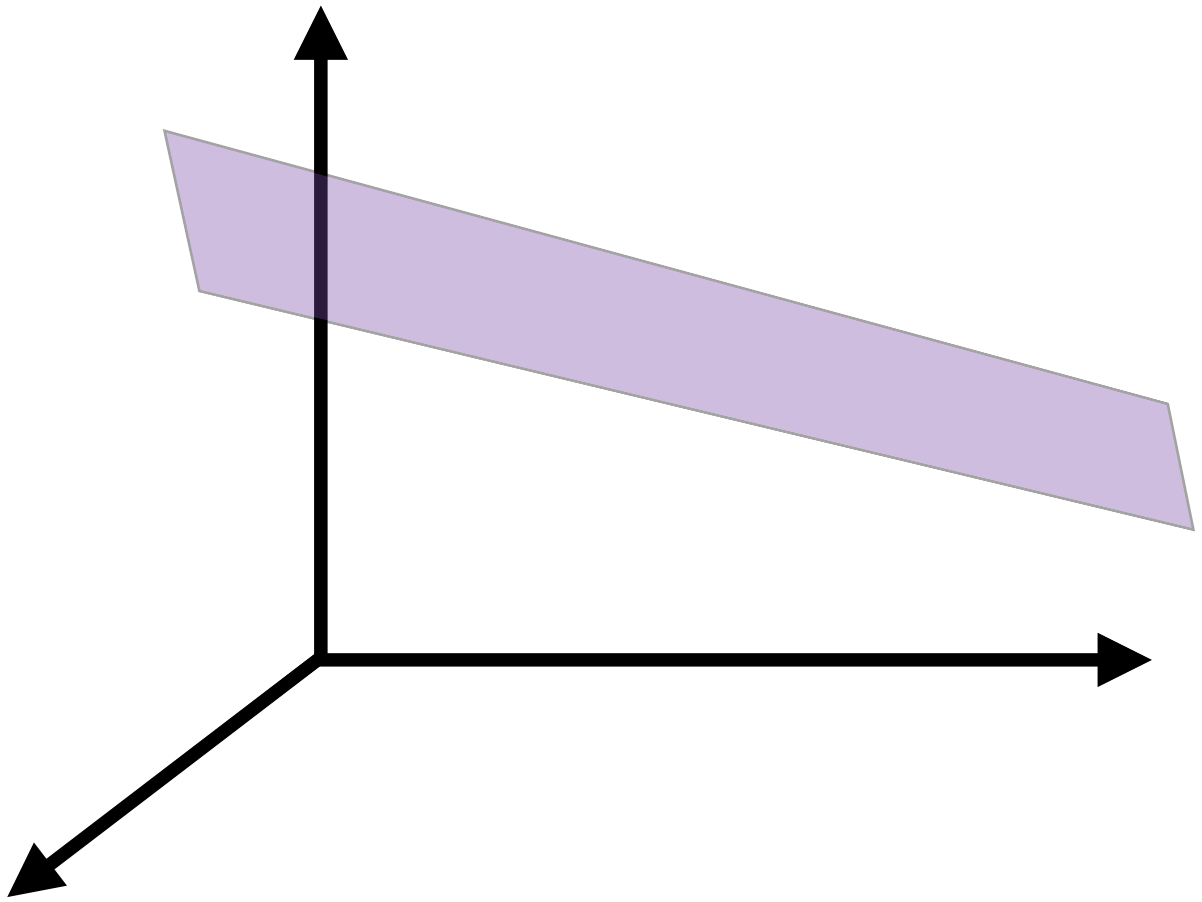
Multiple Regression

Simple and Multiple Regression



Simple Regression

Data in 2 dimensions



Multiple Regression

Data in > 2 dimensions

The big new risk with multiple regression is **multicollinearity**: X variables containing the same information

Multiple Regression

Regression Equation:

$$y = C_1 + C_2x_1 + \ldots + C_kx_{k-1}$$

$$\underset{n \times 1}{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \ldots \\ y_n \end{bmatrix}} = \underset{n \times k}{\begin{bmatrix} 1 & x_{11} & & x_{1k-1} \\ 1 & x_{21} & & x_{2k-1} \\ 1 & x_{31} & \ldots & x_{3k-1} \\ \ldots & \ldots & & \ldots \\ 1 & x_{n1} & & x_{nk-1} \end{bmatrix}} * \underset{k \times 1}{\begin{bmatrix} C_1 \\ C_2 \\ \ldots \\ C_k \end{bmatrix}}$$

n Rows,
1 Column

n Rows,
k Columns

k Rows,
1 Column

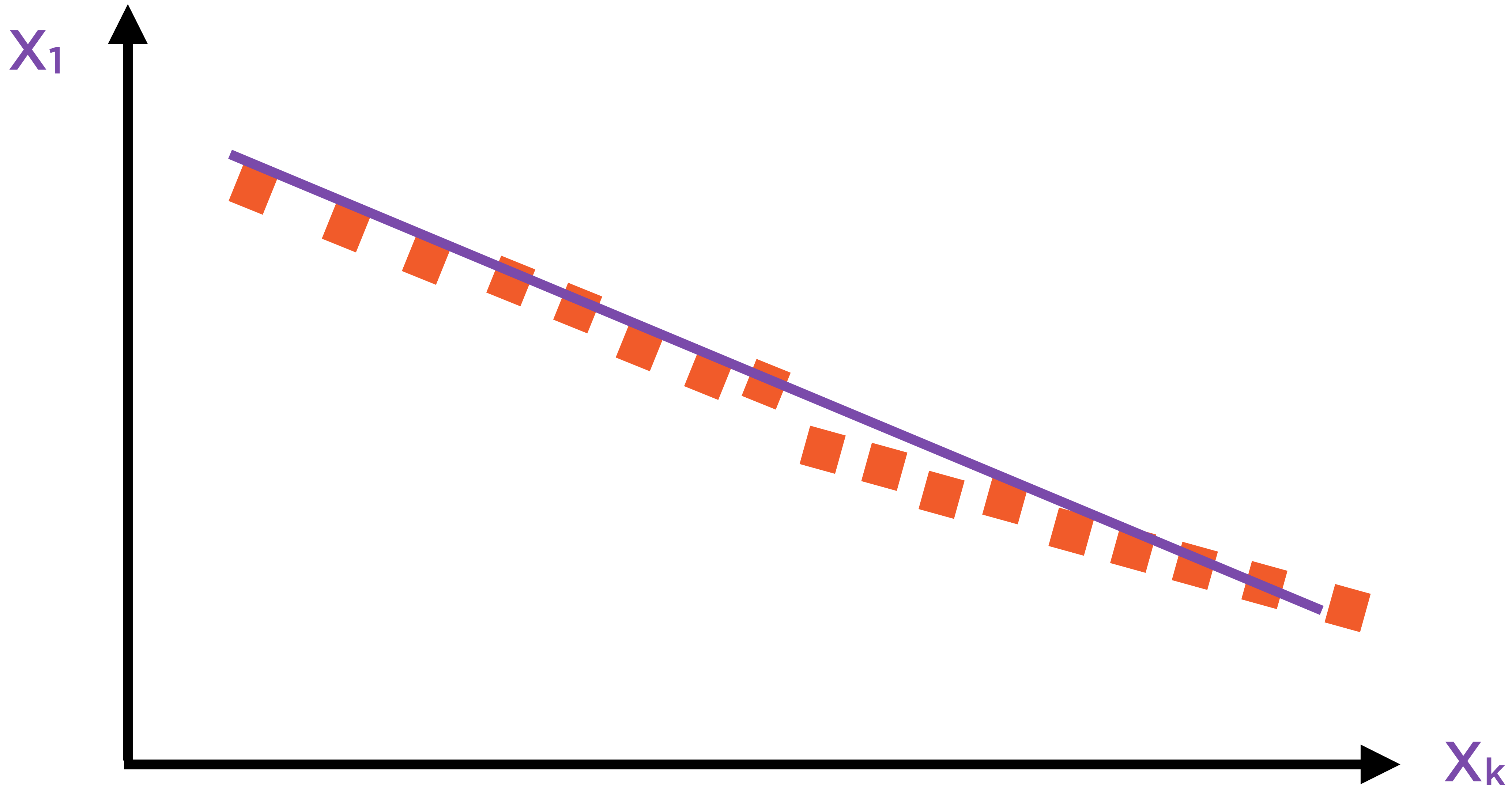
Multiple Regression

Regression Equation:

$$y = C_1 + C_2x_1 + \dots + C_kx_{k-1}$$

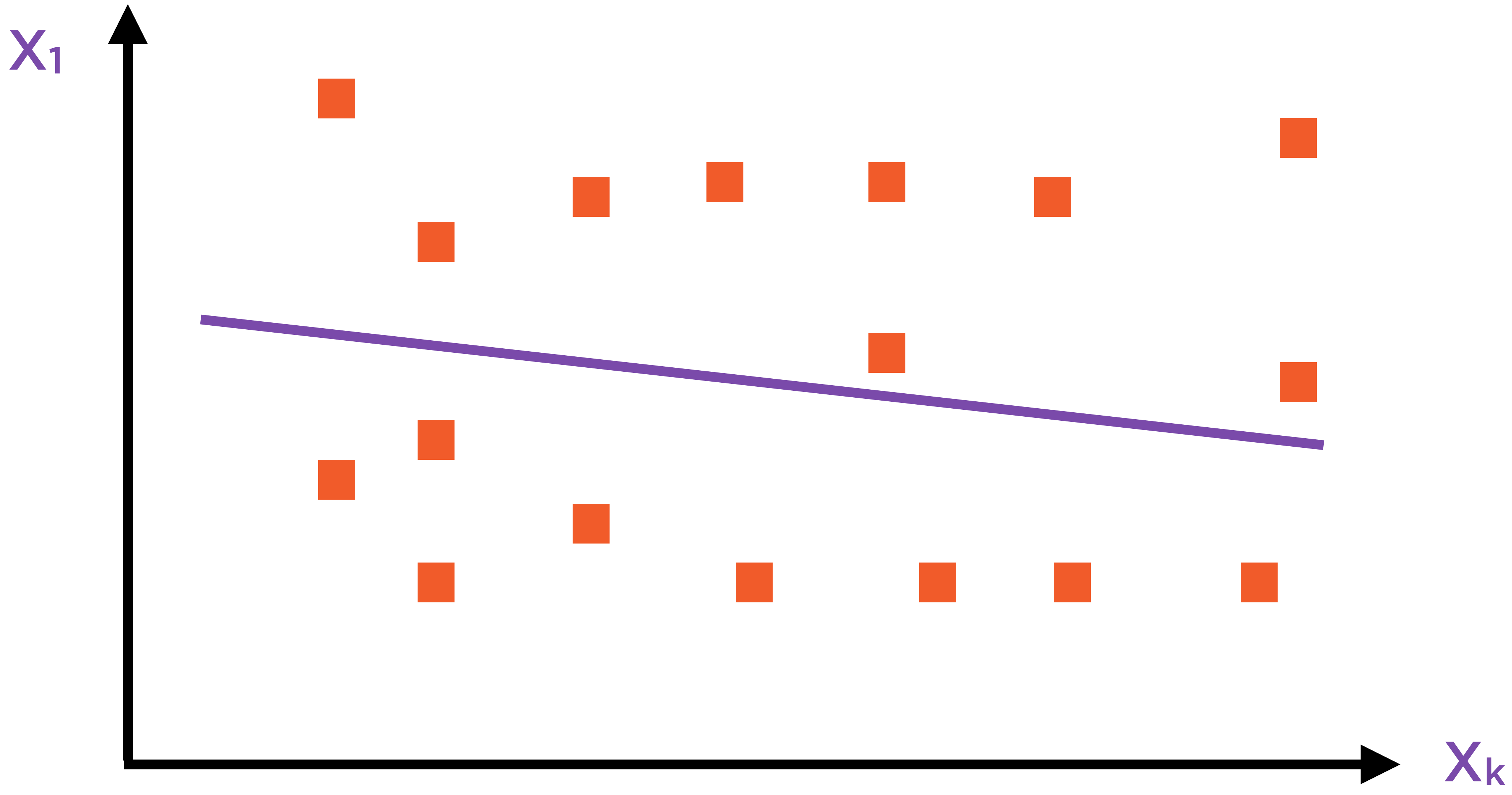
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix} = \begin{array}{c} \begin{bmatrix} 1 & x_{11} & & x_{1k-1} \\ 1 & x_{21} & & x_{2k-1} \\ 1 & x_{31} & \dots & x_{3k-1} \\ \dots & \dots & & \dots \\ 1 & x_{n1} & & x_{nk-1} \end{bmatrix} \\ \quad\quad x_1 \quad\quad\quad\quad x_k \end{array} * \begin{bmatrix} C_1 \\ C_2 \\ \dots \\ C_k \end{bmatrix}$$

Bad News: Multicollinearity Detected



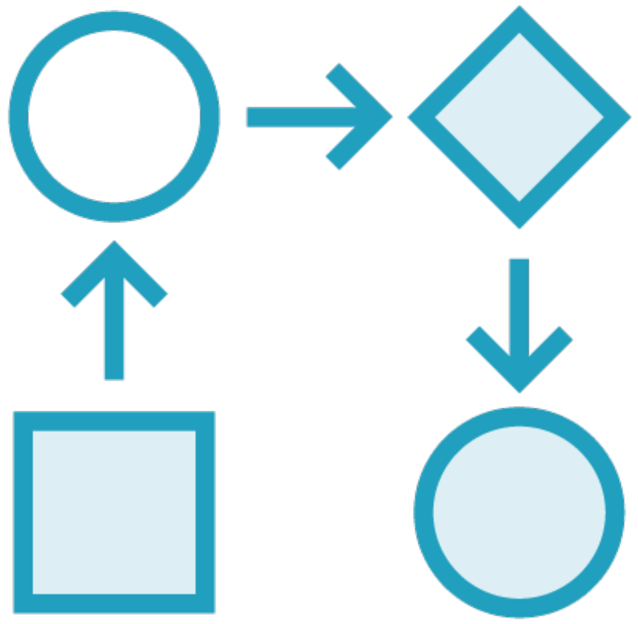
Highly correlated explanatory variables

Good News: No Multicollinearity Detected



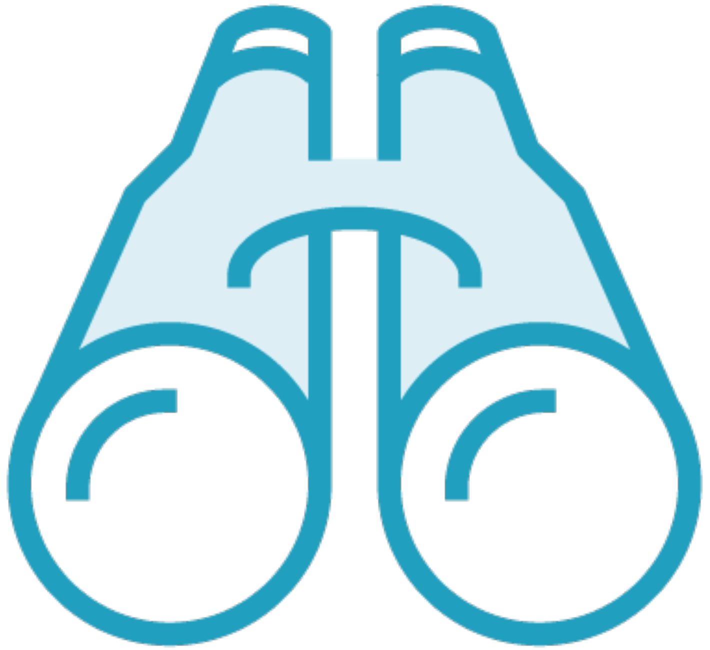
Uncorrelated explanatory variables

Multicollinearity Kills Regression's Usefulness



Explaining Variance

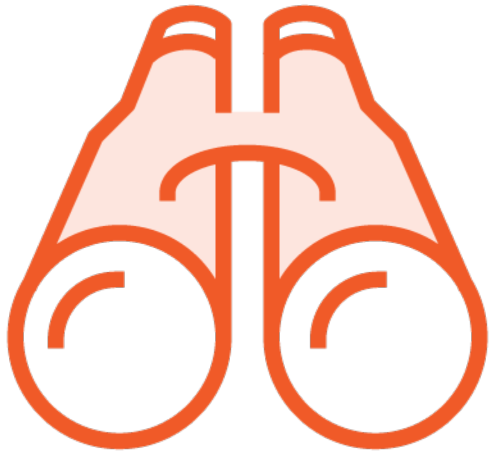
The R^2 as well as the regression coefficients are not very reliable



Making Predictions

The regression model will perform poorly with out-of-sample data

Multicollinearity: Prevention and Cure

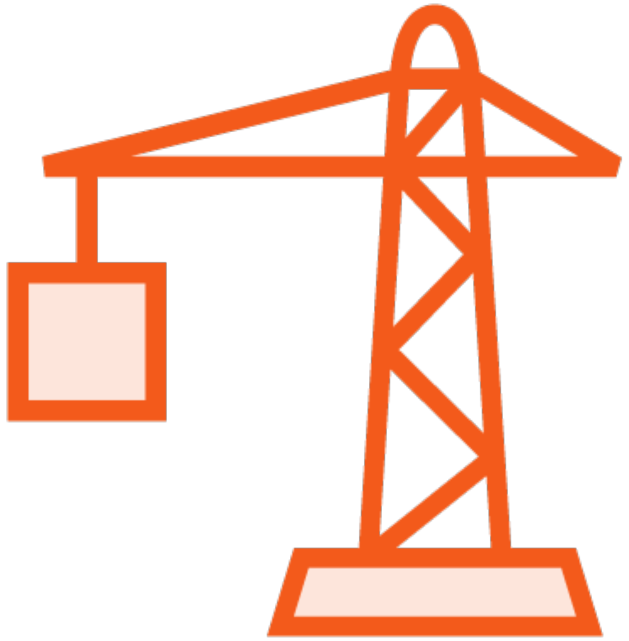


Common Sense

Big-picture understanding of the data

Nuts and Bolts

Setting up data right



Heavy Lifting

Factor analysis, principal components analysis (PCA)

R^2

The most common and popular metric for evaluating regression

Between 0 and 100%

Unfortunately, always increases by adding new x variables

Can lead to overfitting

Adjusted R^2 preferred for evaluating multiple regression

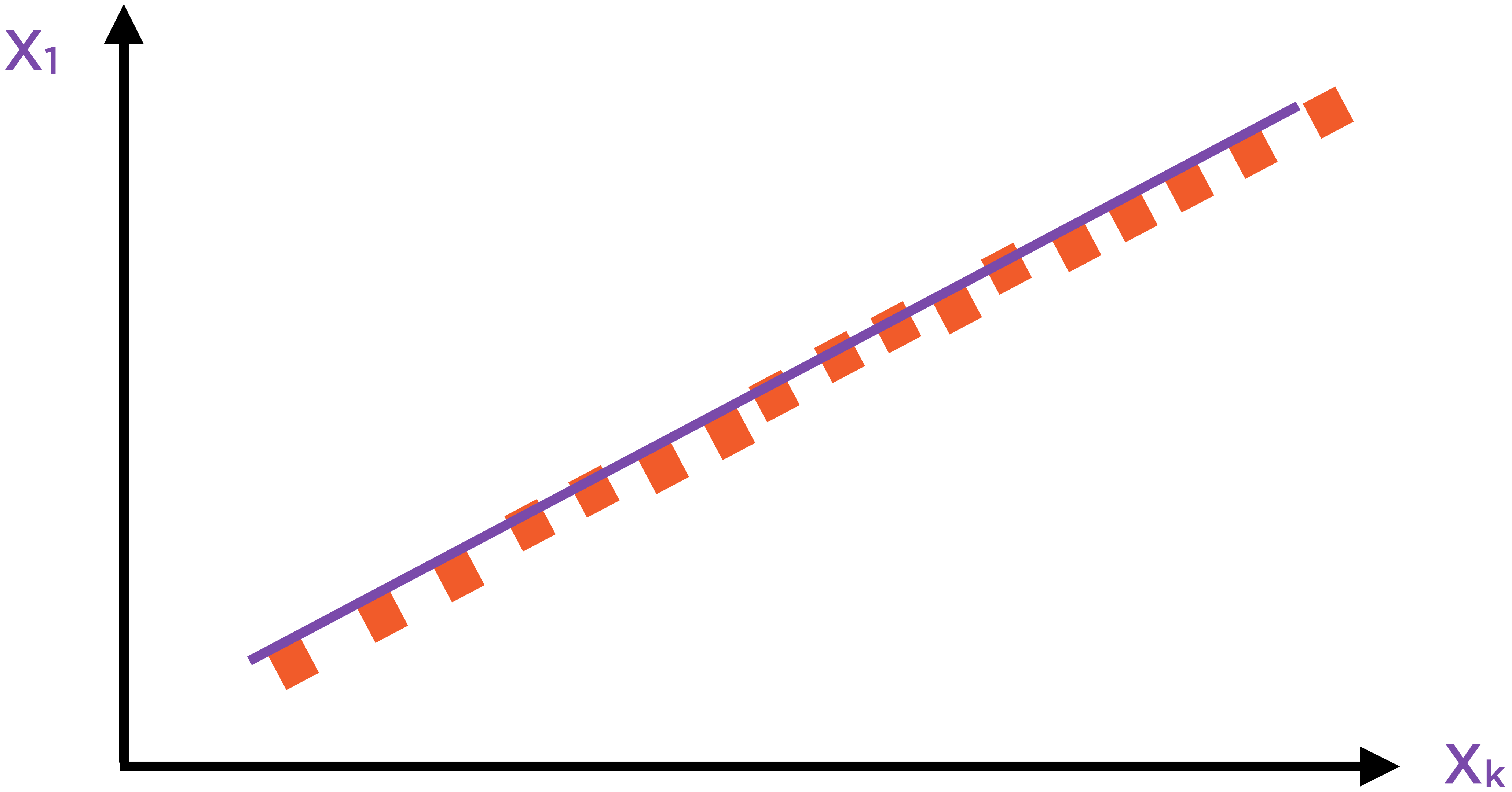
Adjusted-R^2 = R^2 x (Penalty for adding irrelevant variables)

Adjusted-R^2

Increases if irrelevant* variables are deleted

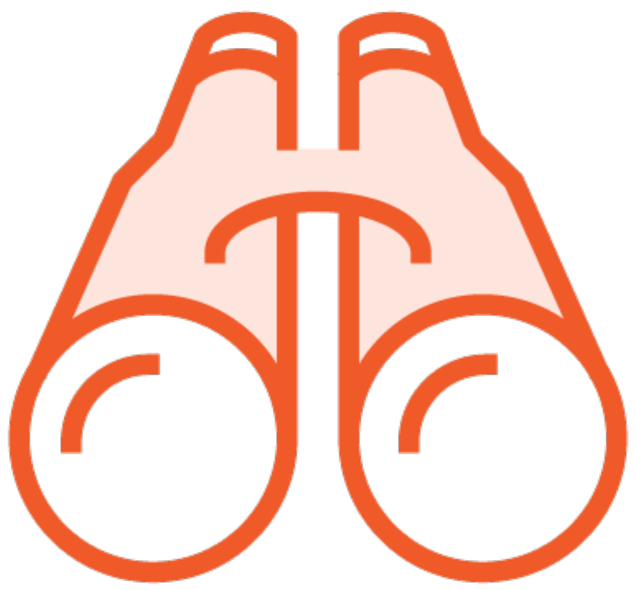
(*irrelevant variables = any group whose F-ratio < 1)

Bad News: Multicollinearity Detected



Highly correlated explanatory variables

Common Sense



Think deeply about each x variable

Eliminate closely related ones

Perform feature selection to select relevant x variables

Nuts and Bolts

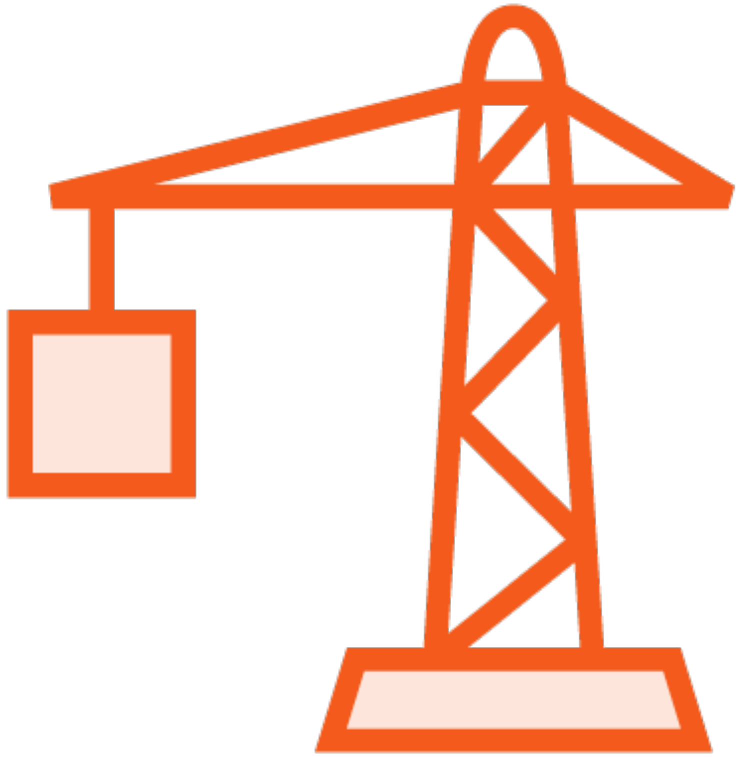
‘Standardize’ the variables

Rely on adjusted-R^2, not plain R^2

Set up dummy variables right

Distribute lags

Heavy Lifting



Find underlying factors that drive the correlated x variables

Principal Component Analysis (PCA) is a great tool

Demo

Performing simple linear regression with a single predictor using analytical and machine learning techniques

Demo

Performing multiple regression using analytical and machine learning techniques

Selecting relevant features using statistical methods

Summary

Regression to predict continuous variables

Simple and multiple regression

Multicollinearity and risks in regression

R-square and adjusted R-square

Selecting features for regression using statistical techniques