Building Linear Models Using StatsModels

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Ordinary least squares regression makes many assumptions about data

Sometimes get very restrictive

Different variations to get around these

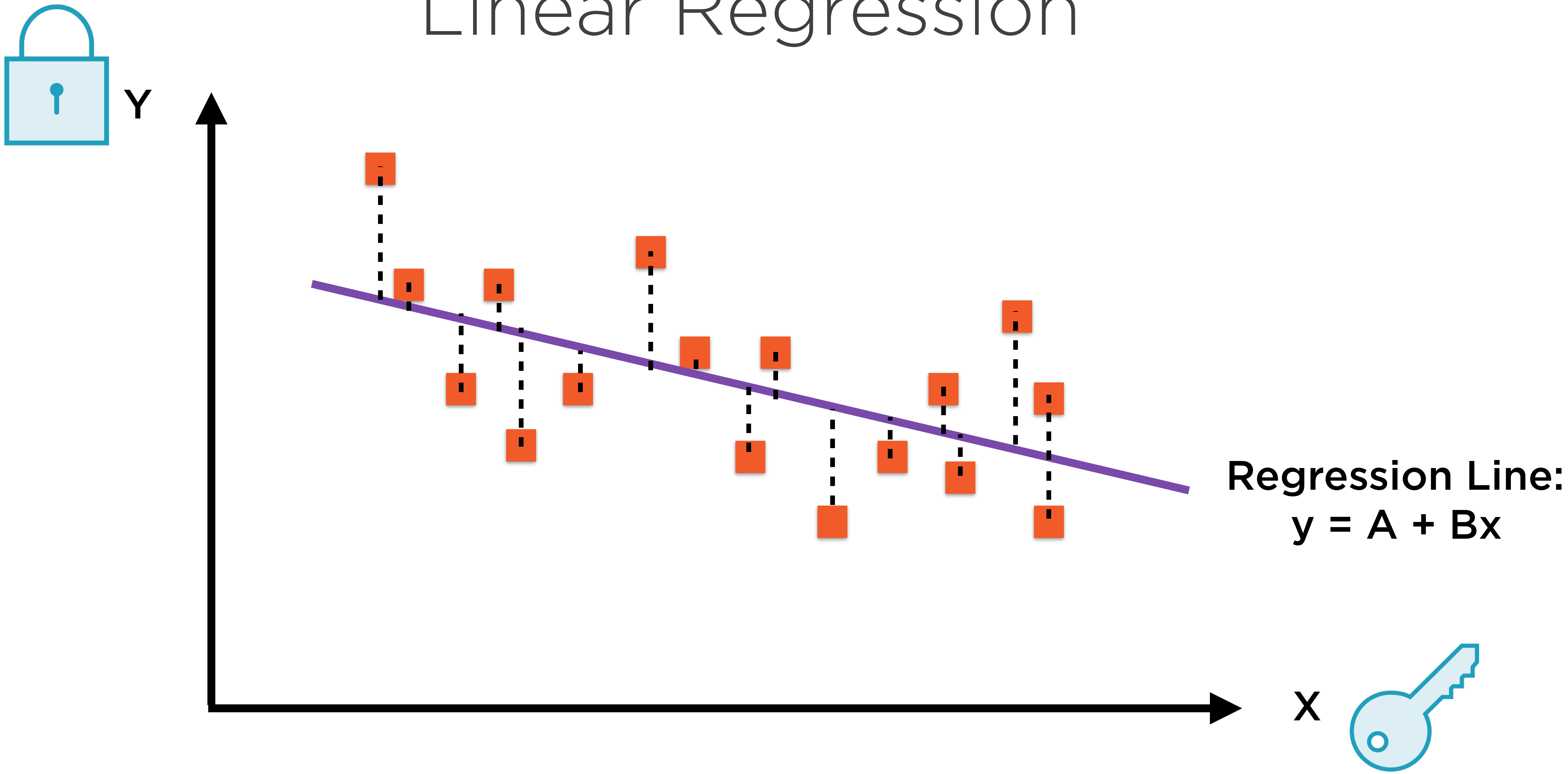
Generalized or weighted least squares for heteroscedasticity

Generalized linear models for non-normal y variables

Robust linear models to cope with outliers

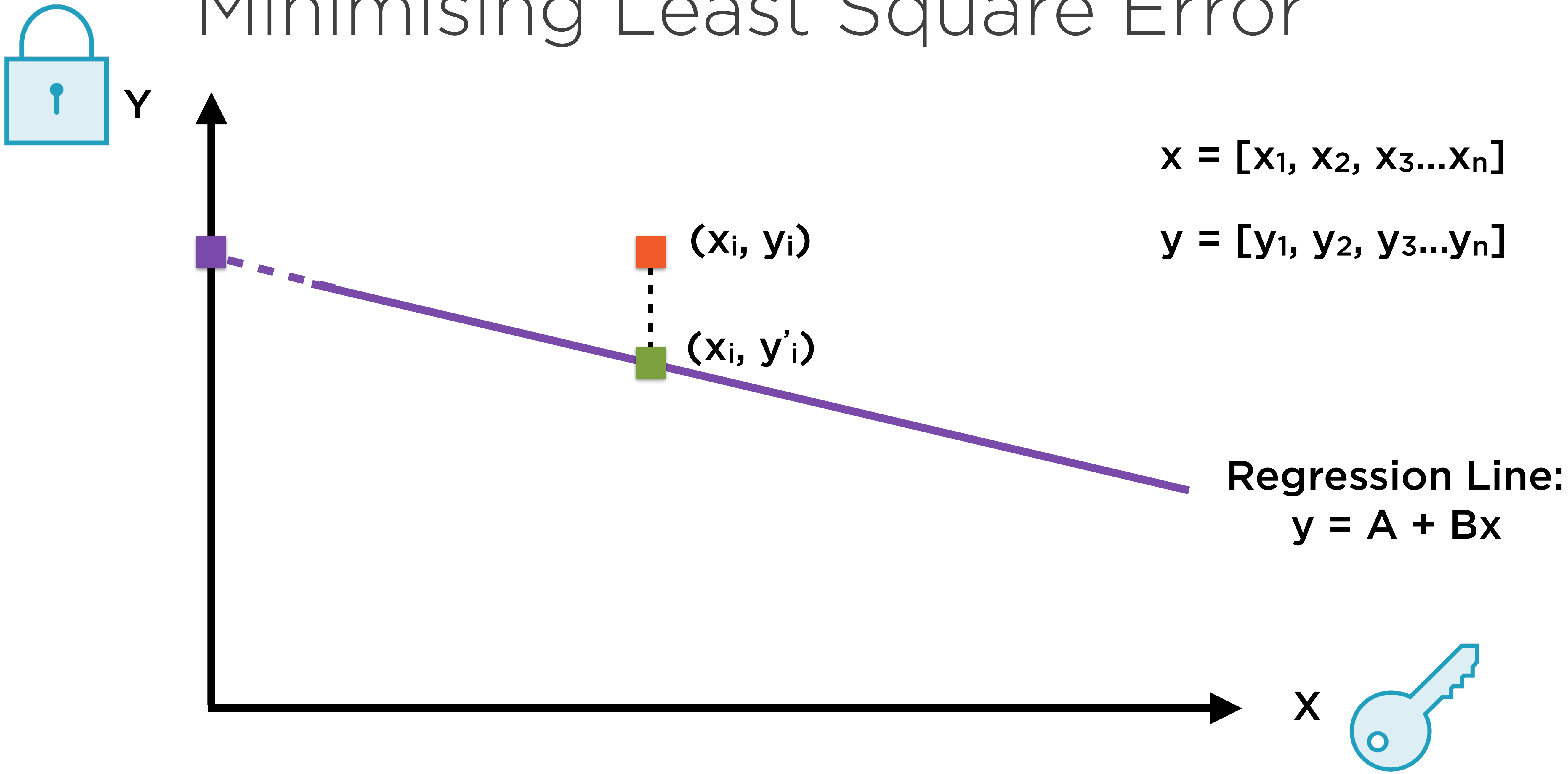
Regression Assumptions

Linear Regression



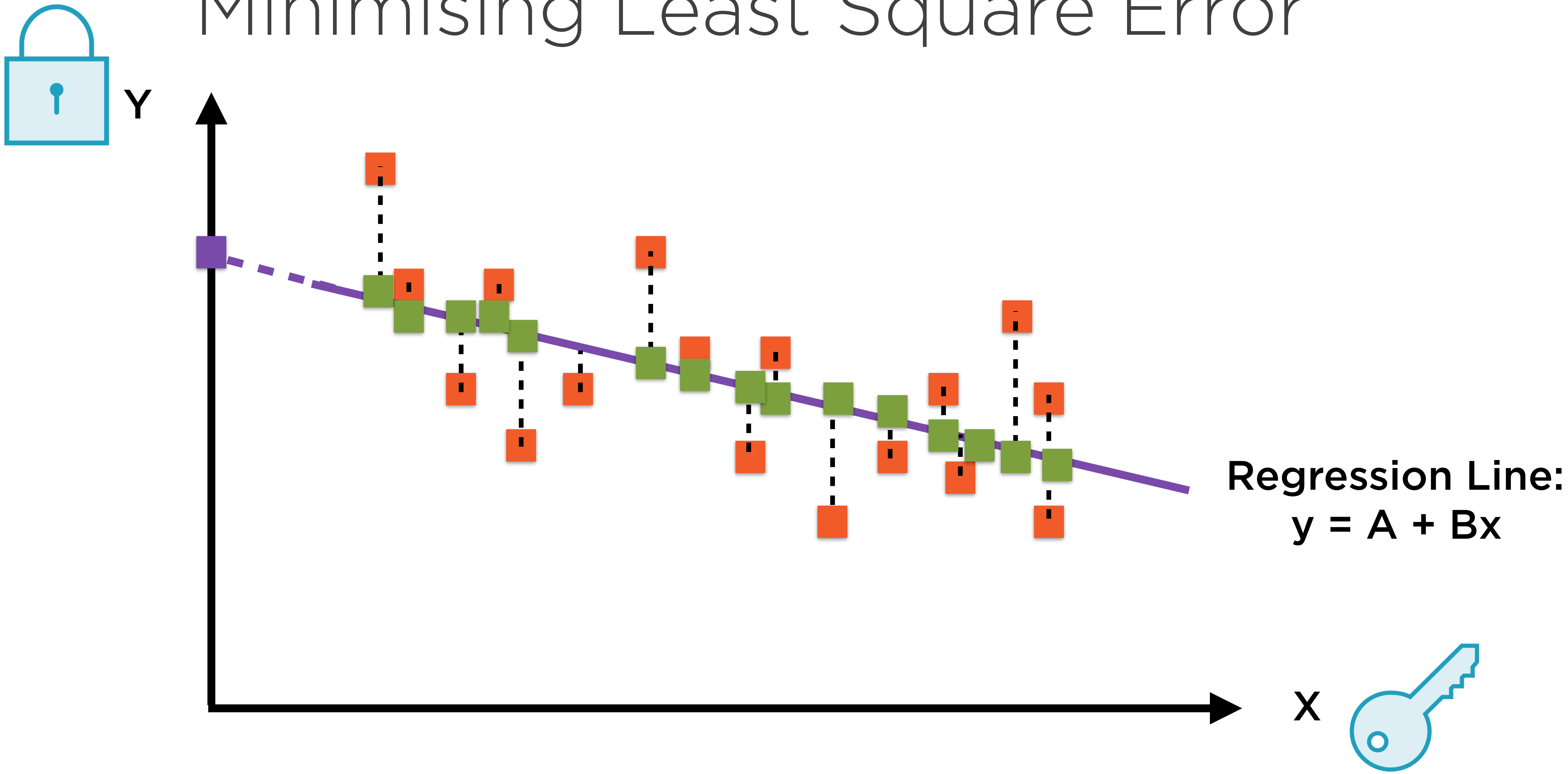
The "best fit" line is called the regression line

Minimising Least Square Error



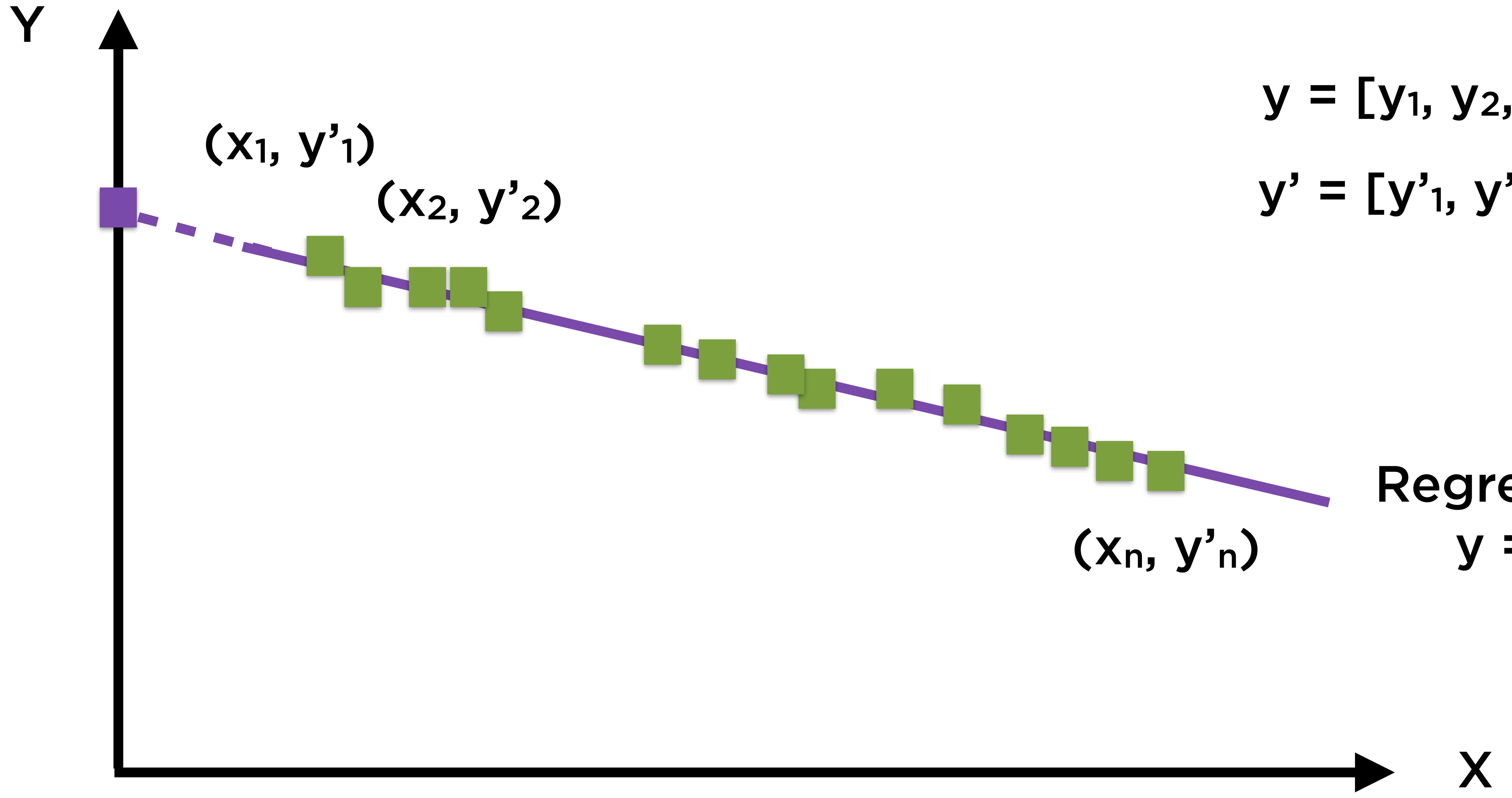
Each point (x_i, y_i) has a corresponding point (x_i, y'_i) on the regression line

Minimising Least Square Error



Find all such points (x_i, y'_i) on the regression line

Minimising Least Square Error



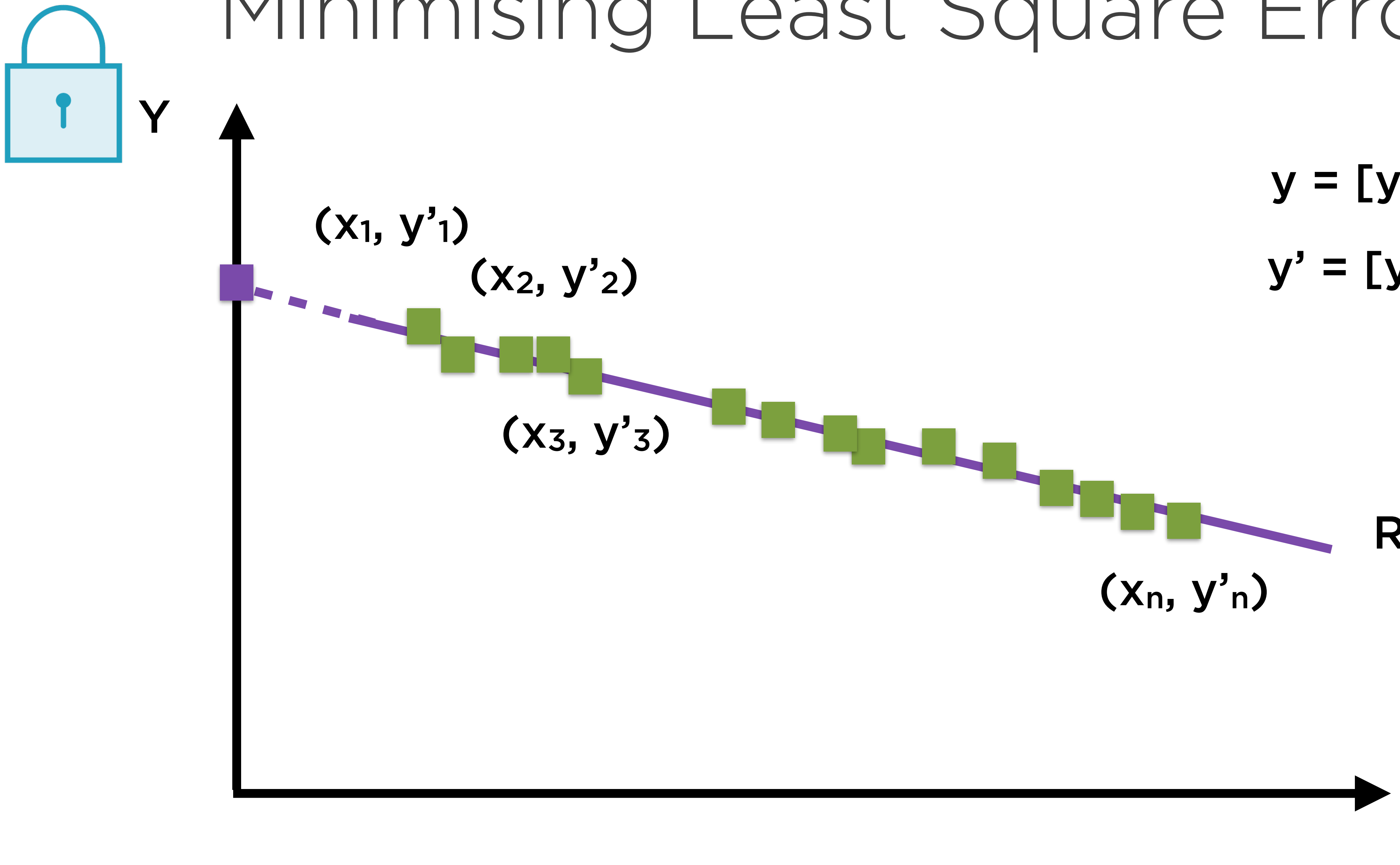
$y = [y_1, y_2, y_3 \ldots y_n]$

$y' = [y'_1, y'_2, y'_3 \ldots y'_n]$

Regression Line:
$y = A + Bx$

Find all such points (x_i, y'_i) on the regression line

Minimising Least Square Error



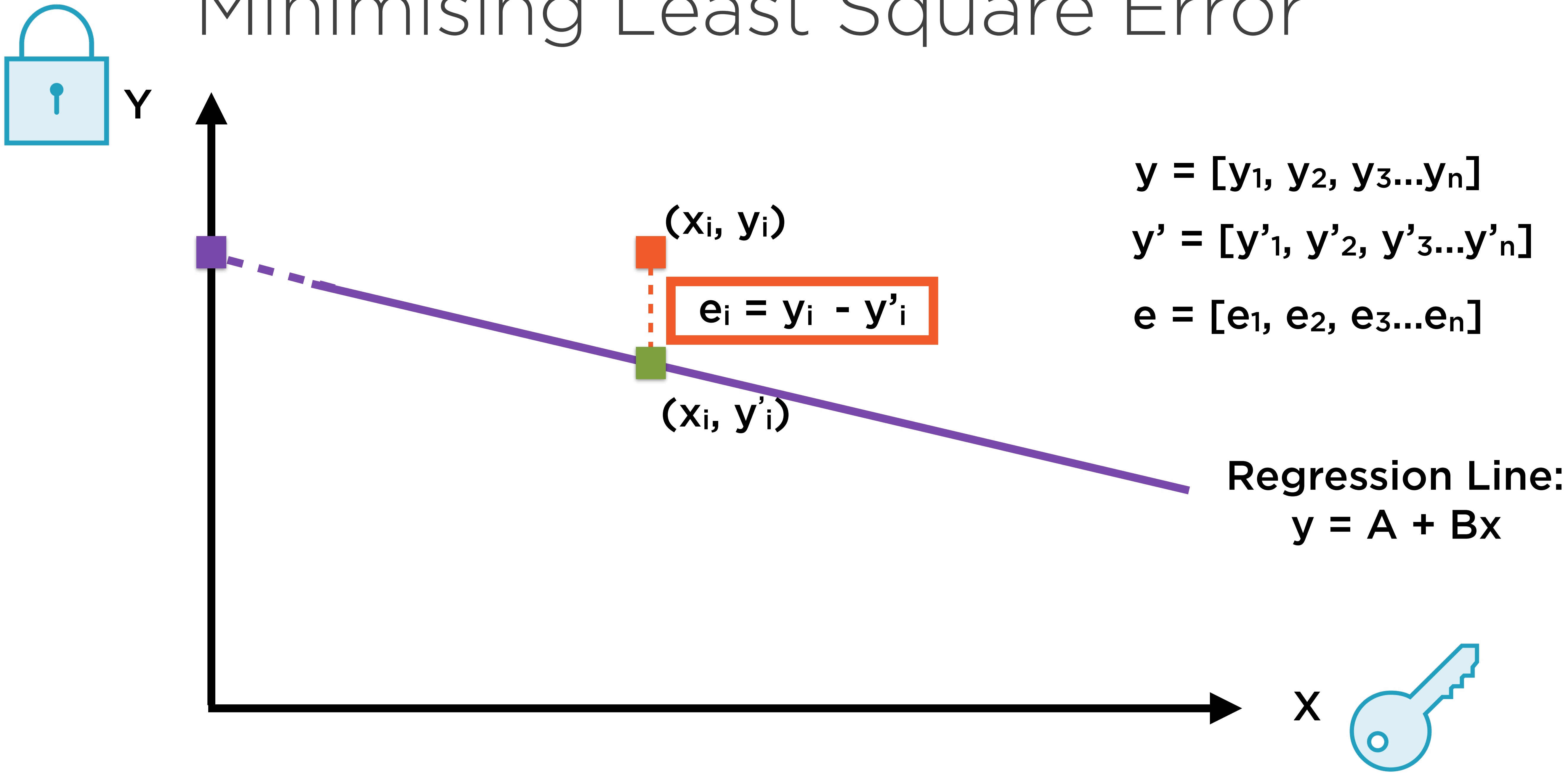
$y = [y_1, y_2, y_3 \ldots y_n]$

$y' = [y'_1, y'_2, y'_3 \ldots y'_n]$

Regression Line:
$y = A + Bx$

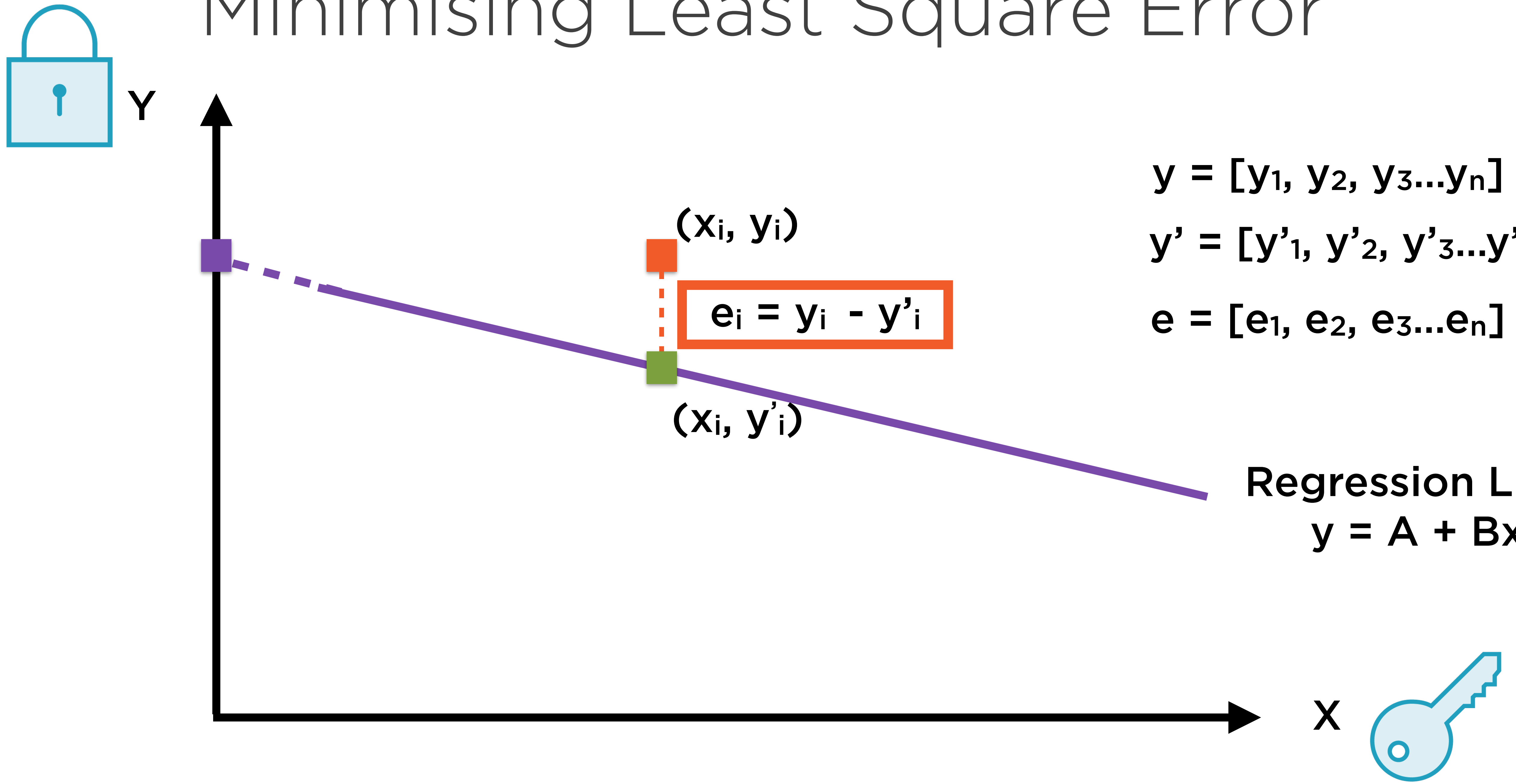
The corresponding values of y'_i are called the fitted values

Minimising Least Square Error



For each point, the difference between y_i and y'_i is called e_i, the residual or the error

Minimising Least Square Error



$y = [y_1, y_2, y_3 \ldots y_n]$

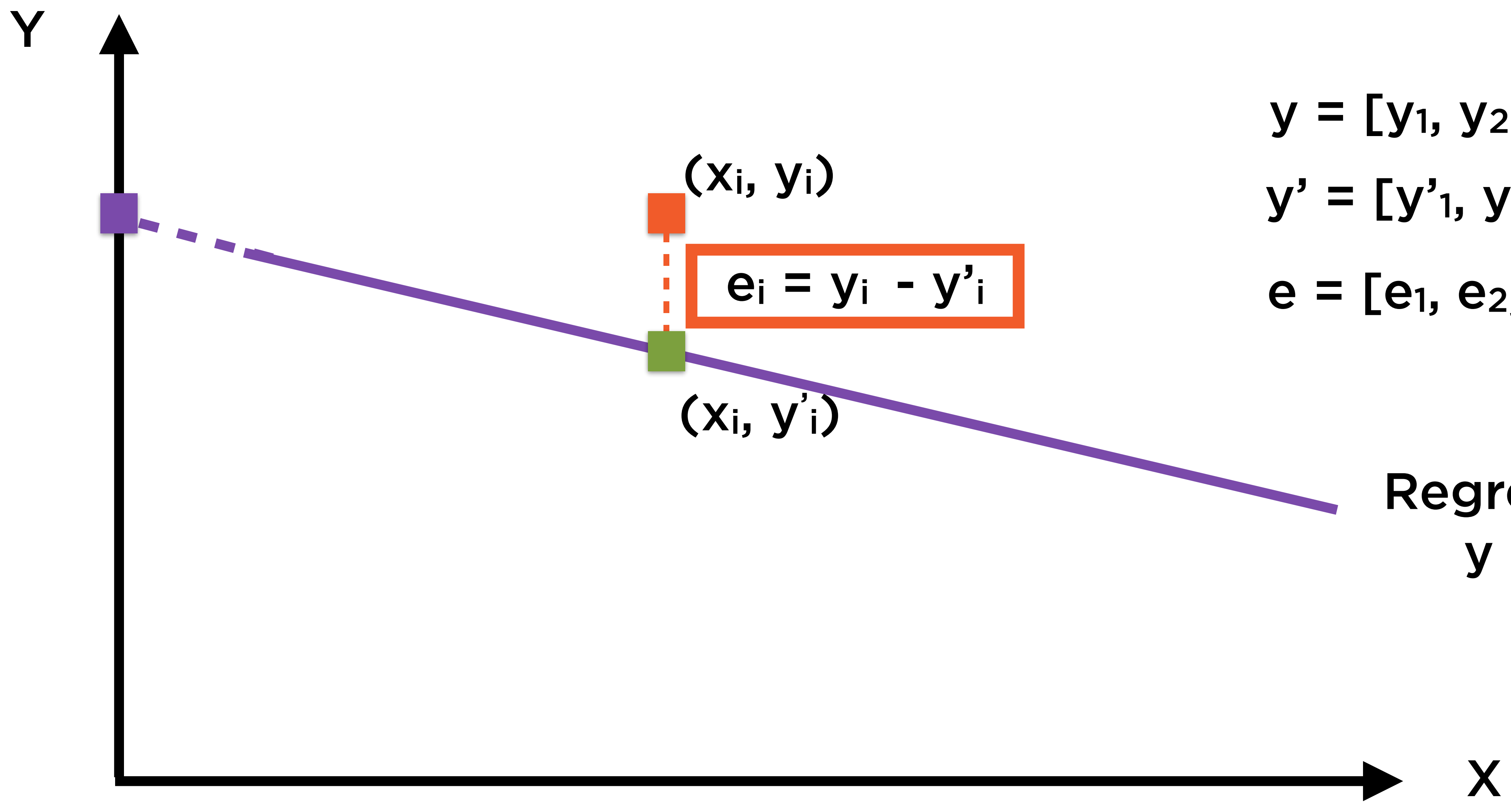
$y' = [y'_1, y'_2, y'_3 \ldots y'_n]$

$e = [e_1, e_2, e_3 \ldots e_n]$

Regression Line:
$y = A + Bx$

Residuals of a regression are the difference between actual and fitted values of the dependent variable

Minimising Least Square Error



$y = [y_1, y_2, y_3 \ldots y_n]$

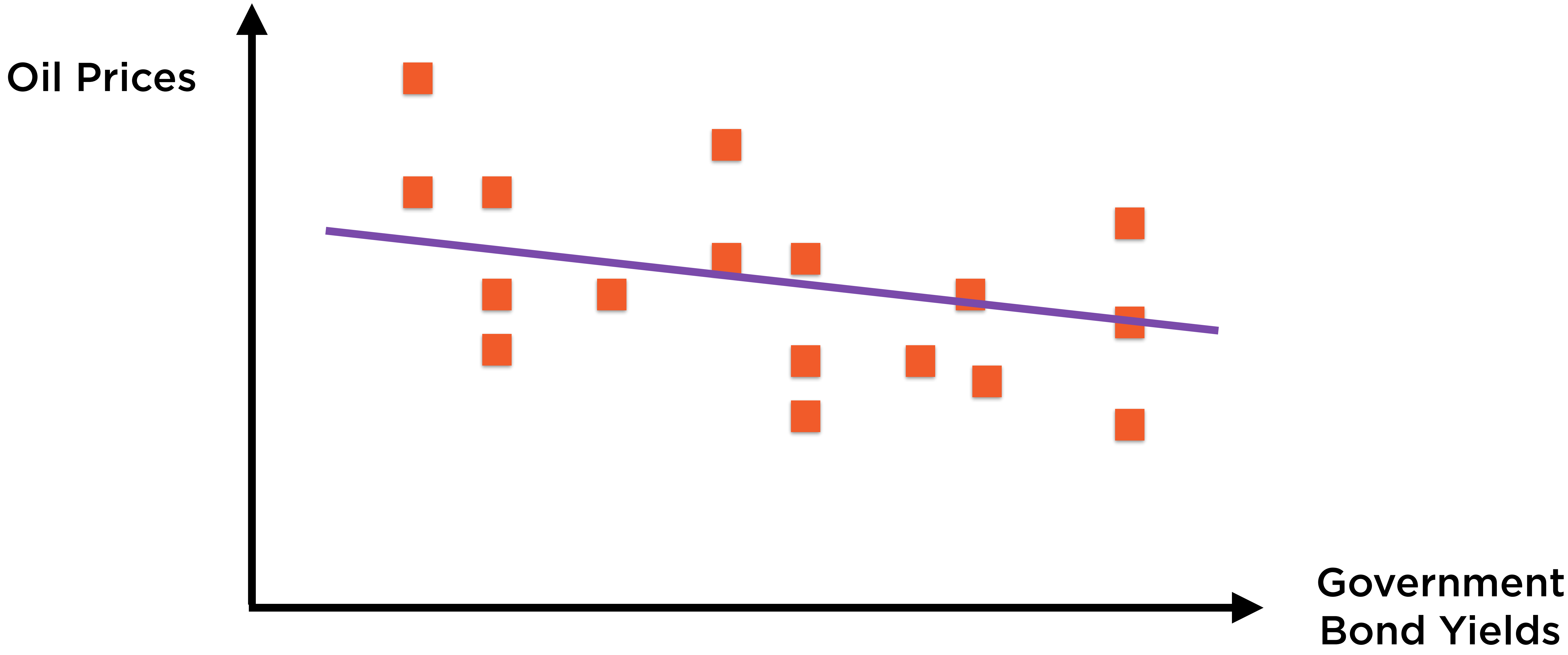
$y' = [y'_1, y'_2, y'_3 \ldots y'_n]$

$e = [e_1, e_2, e_3 \ldots e_n]$

Regression Line:
$y = A + Bx$

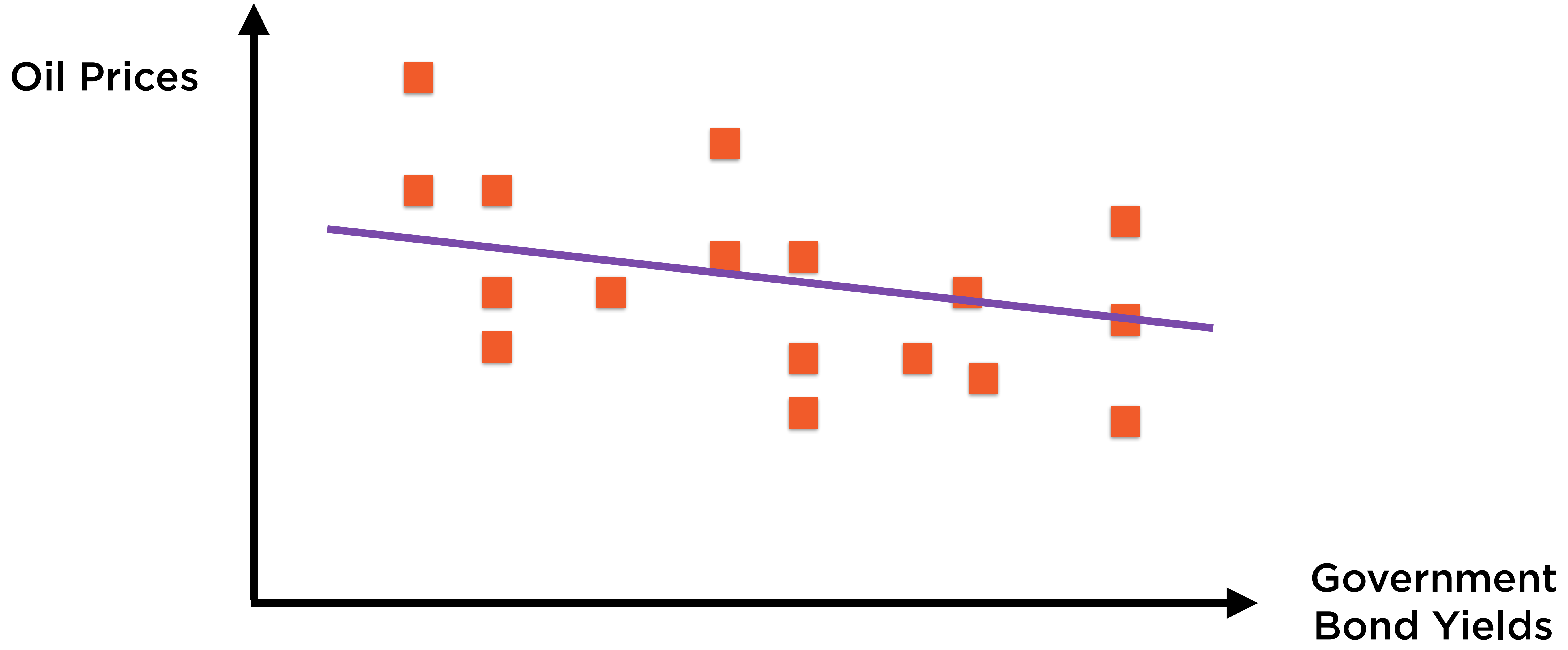
For each point, the difference between y_i and y'_i is called e_i, the residual or the error

Linear Regression



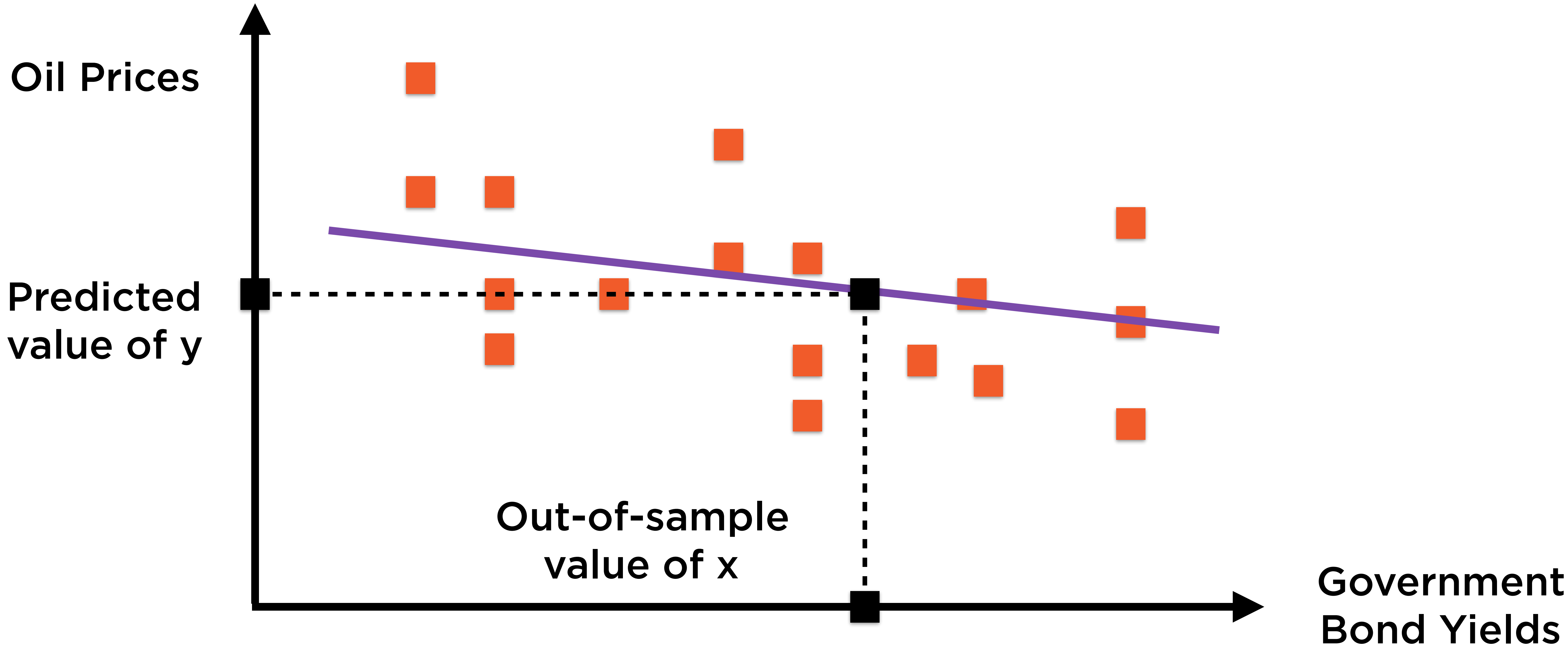
Finding the "best" such straight line is
called Linear Regression

Linear Regression



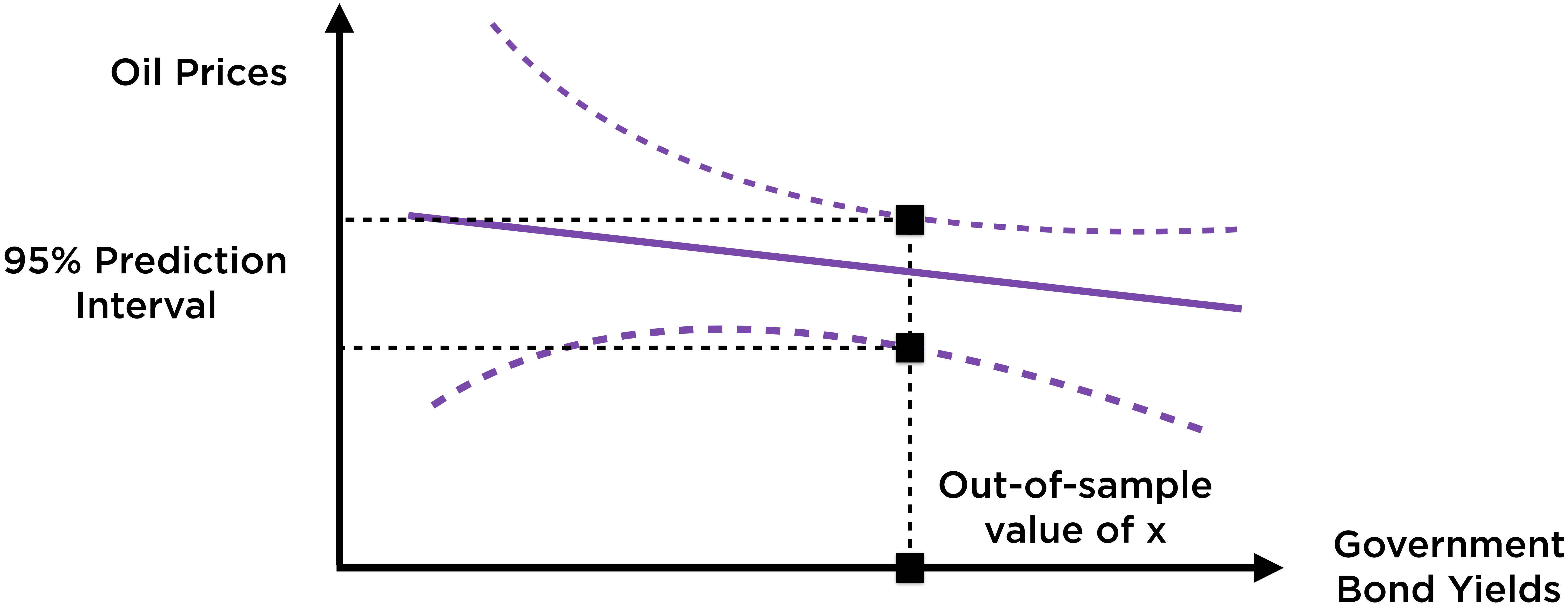
Regression not only gives us the equation of this line, it also signals how reliable the line is

Prediction Using Regression



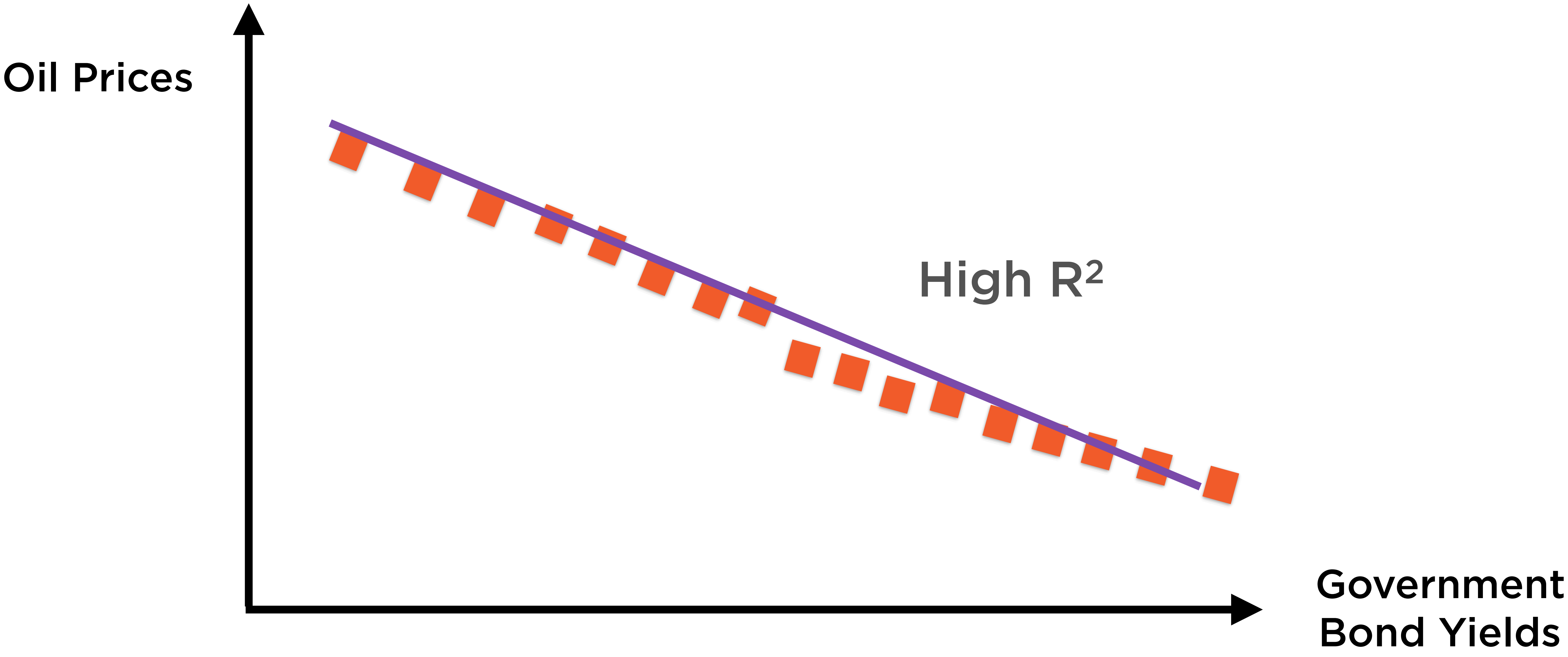
Given a new value of x, use the line to predict the corresponding value of y

Prediction Using Regression



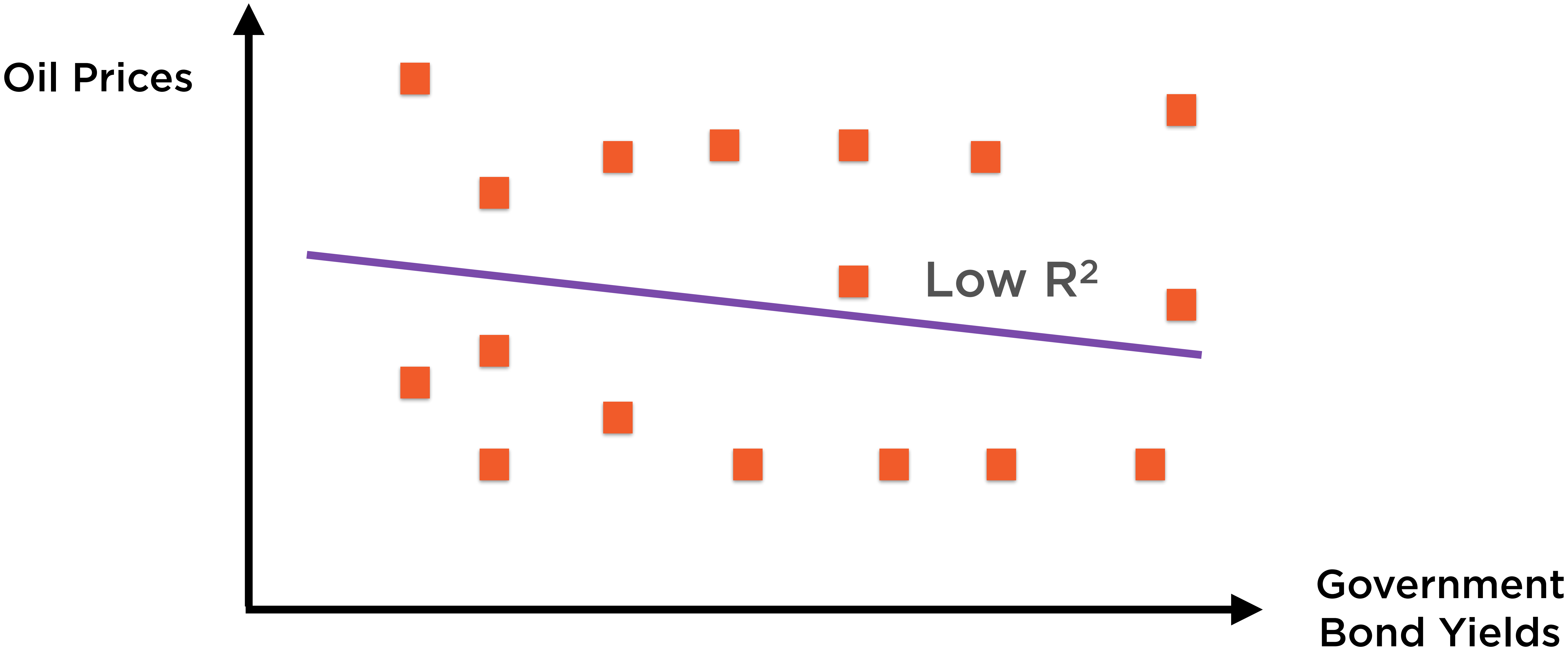
Regression also allows you to specify prediction intervals (similar to confidence intervals) around this point estimate

Linear Regression



High quality of fit

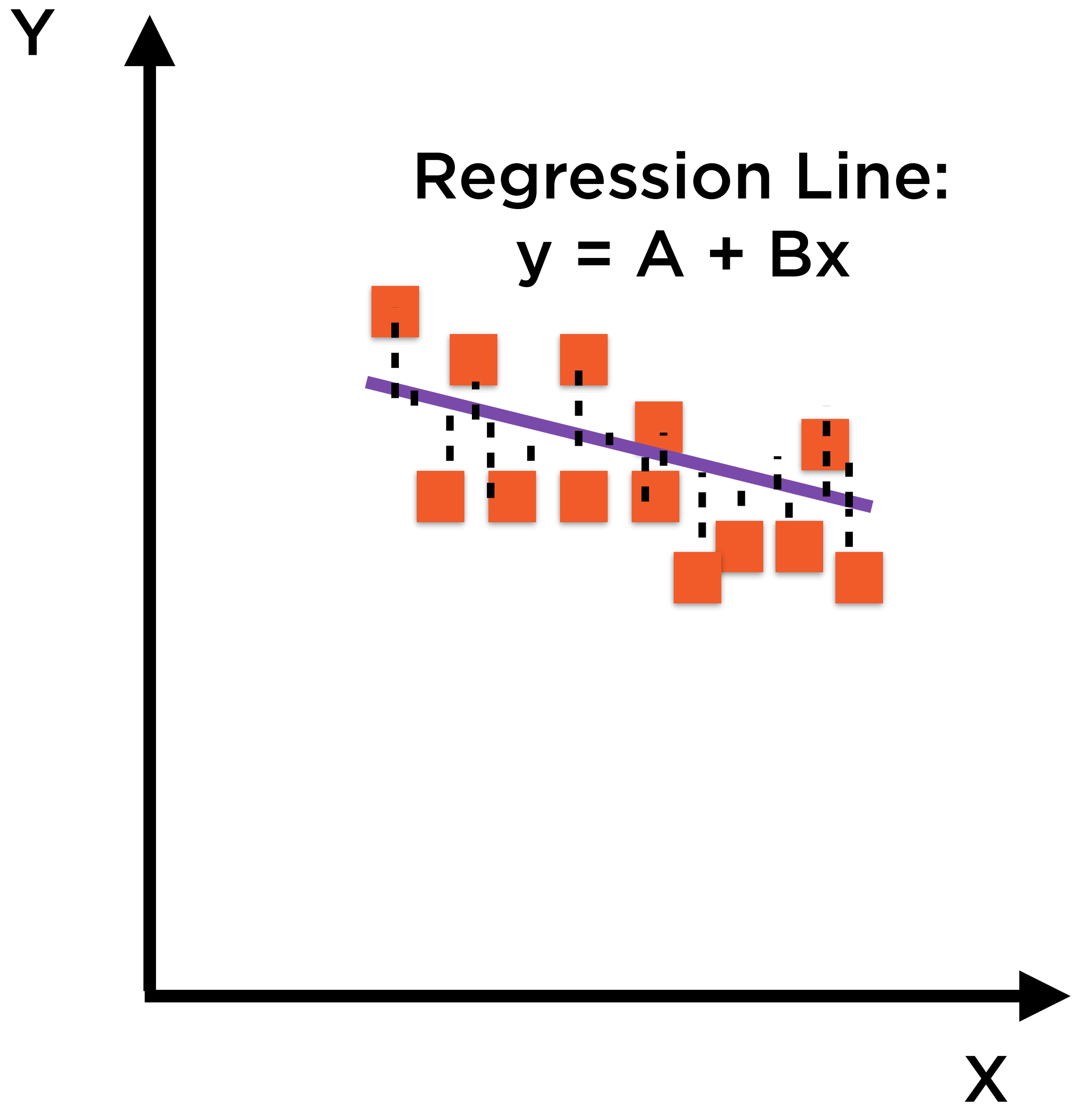
Linear Regression



Low quality of fit

To find the “best fit” line we need to make some assumptions about regression error

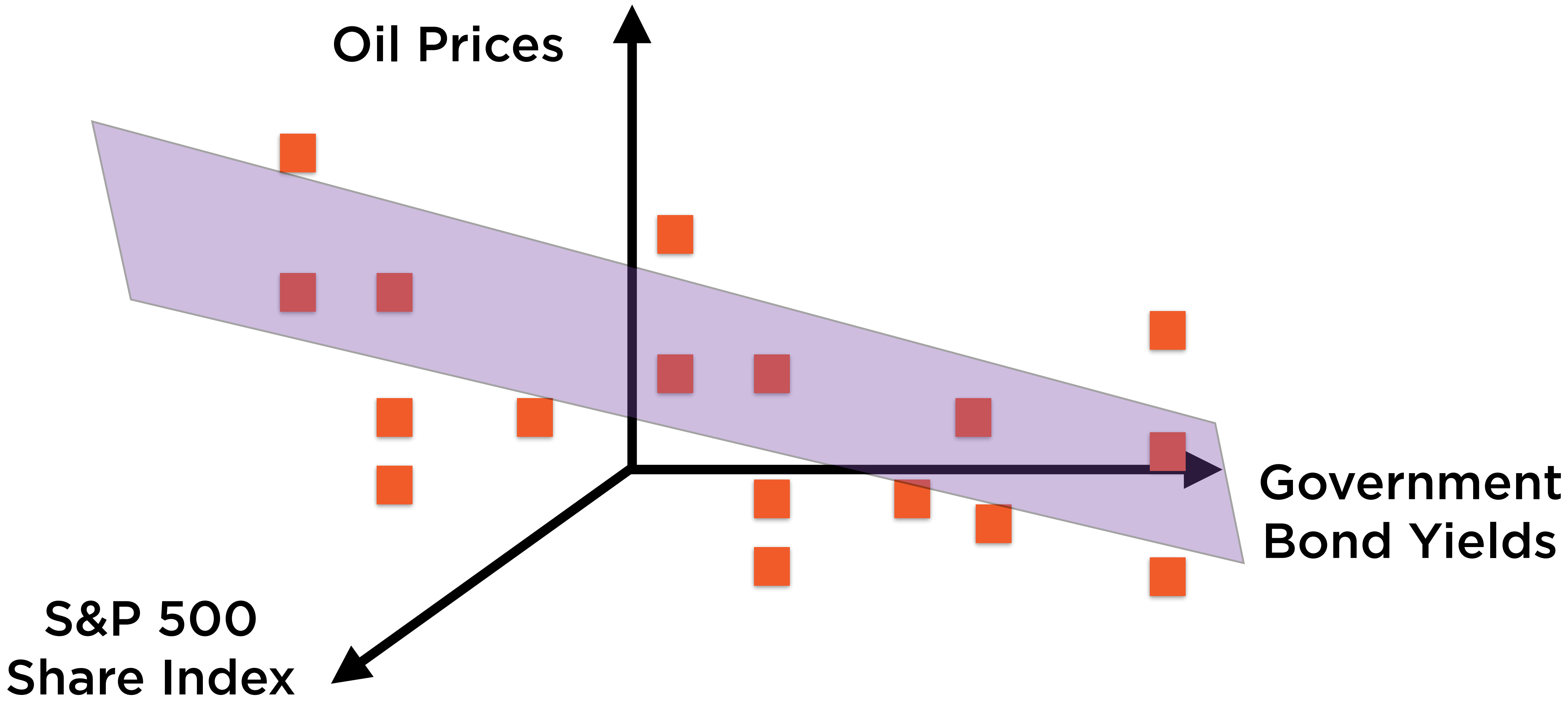
There is a fine distinction between errors and residuals - but we can ignore it



Ideally, residuals should

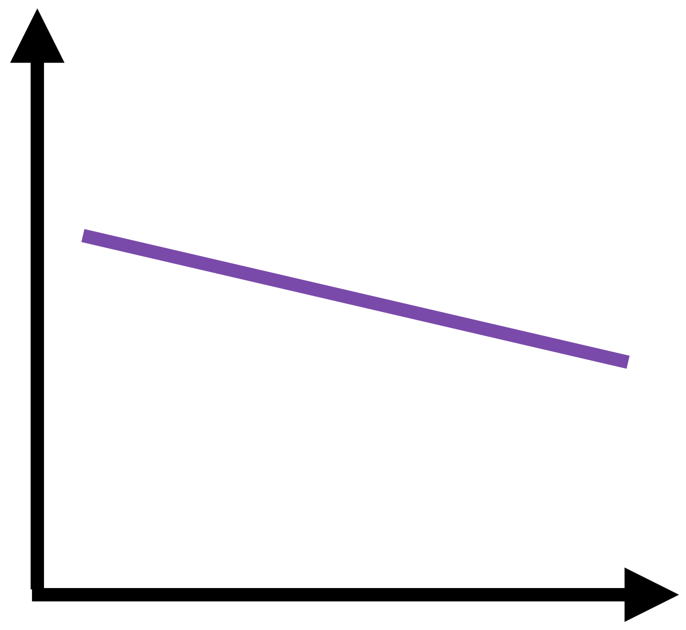
- have zero mean
- have constant variance
- be independent of each other
- be independent of x
- be normally distributed

Multiple Regression



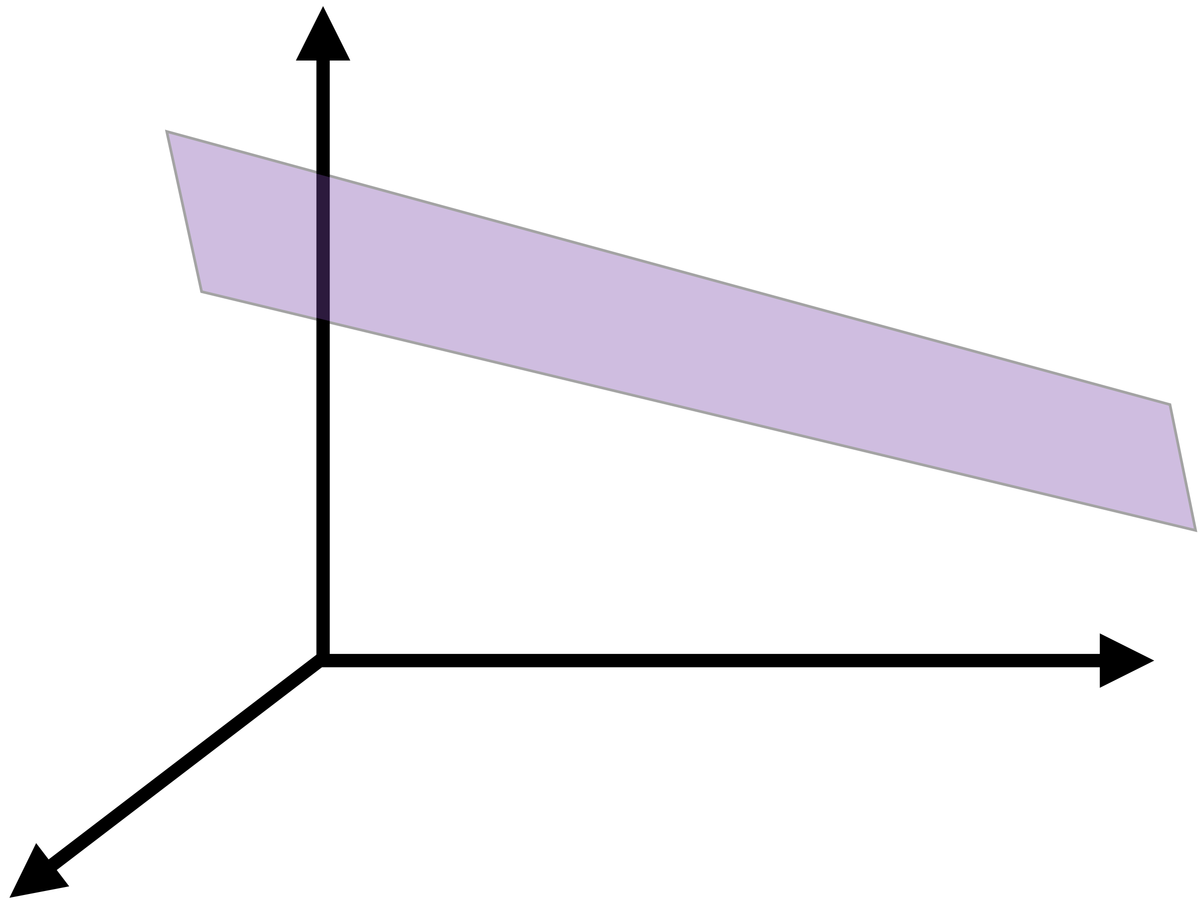
Linear Regression can easily be extended to n-dimensional data

Simple and Multiple Regression



Simple Regression

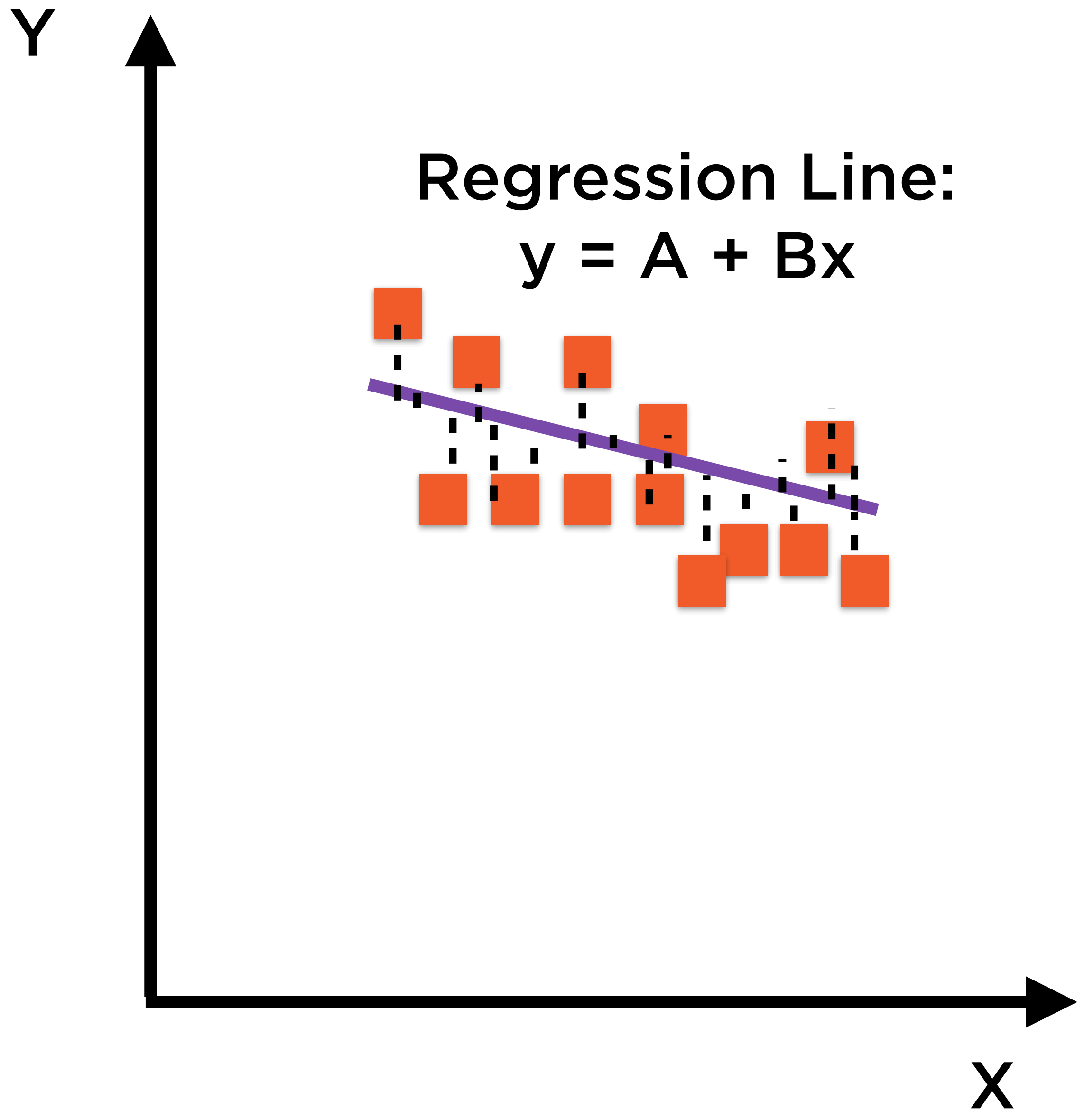
Data in 2 dimensions



Multiple Regression

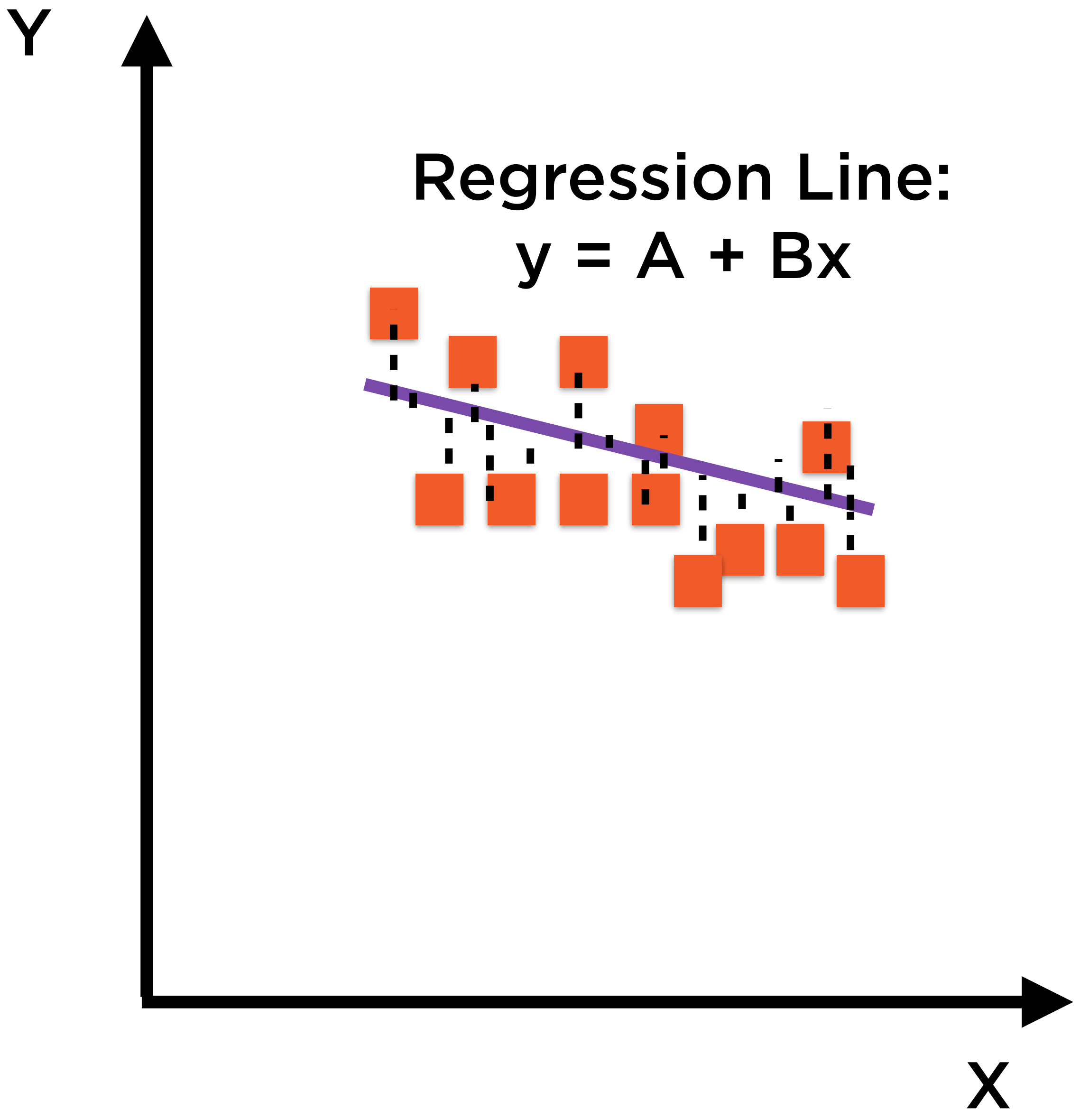
Data in > 2 dimensions

Heteroscedasticity



Ideally, residuals should

- have zero mean
- have constant variance
- be independent of each other
- be independent of x
- be normally distributed



Ideally, residuals should

- have zero mean
- have constant variance
- be independent of each other
- be independent of x
- be normally distributed

Heteroscedasticity:
Non-constant variance

This can be a serious problem in building regression models

Heteroscedasticity

Detection

Implications

Solutions

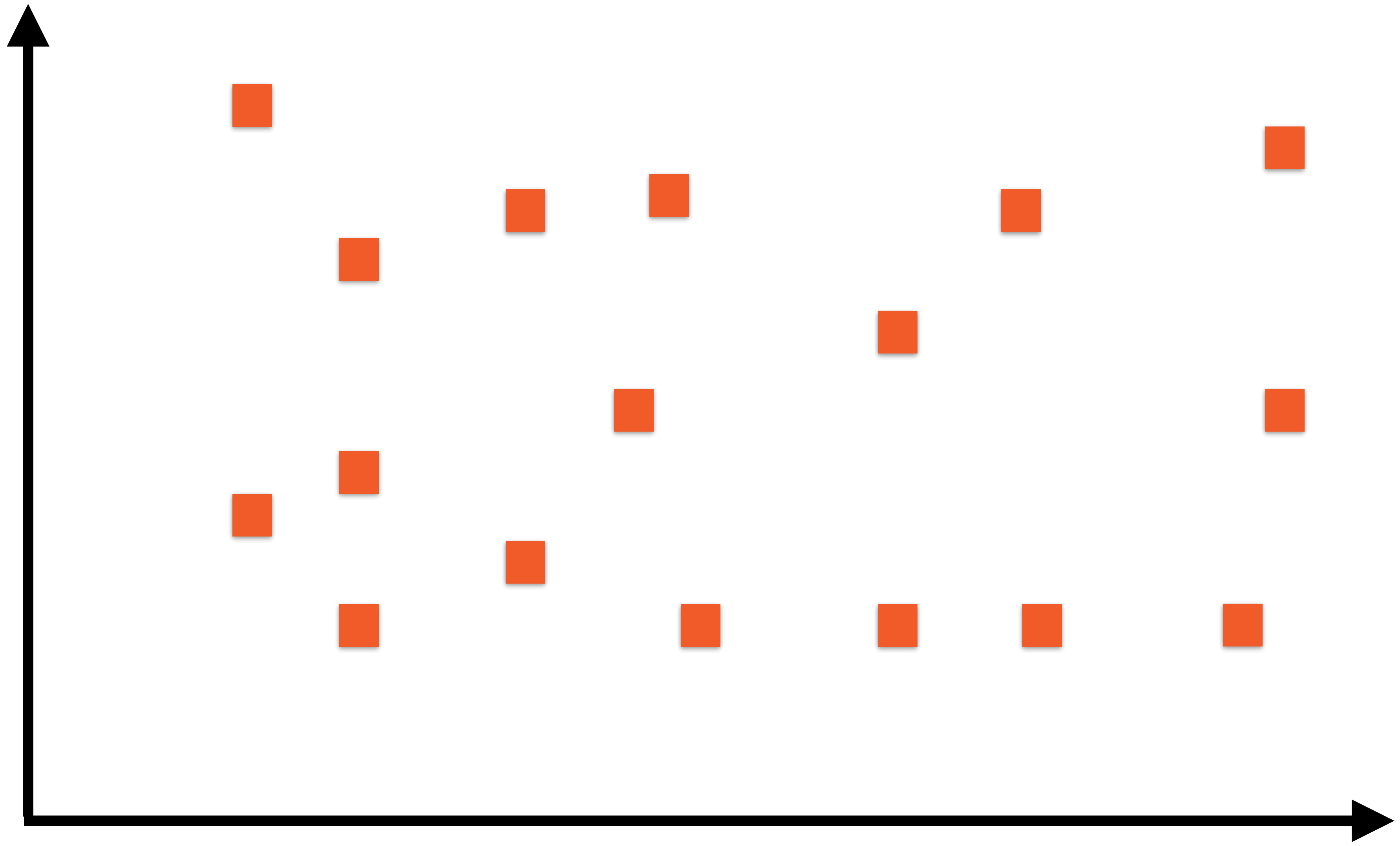
Heteroscedasticity

Detection

Implications

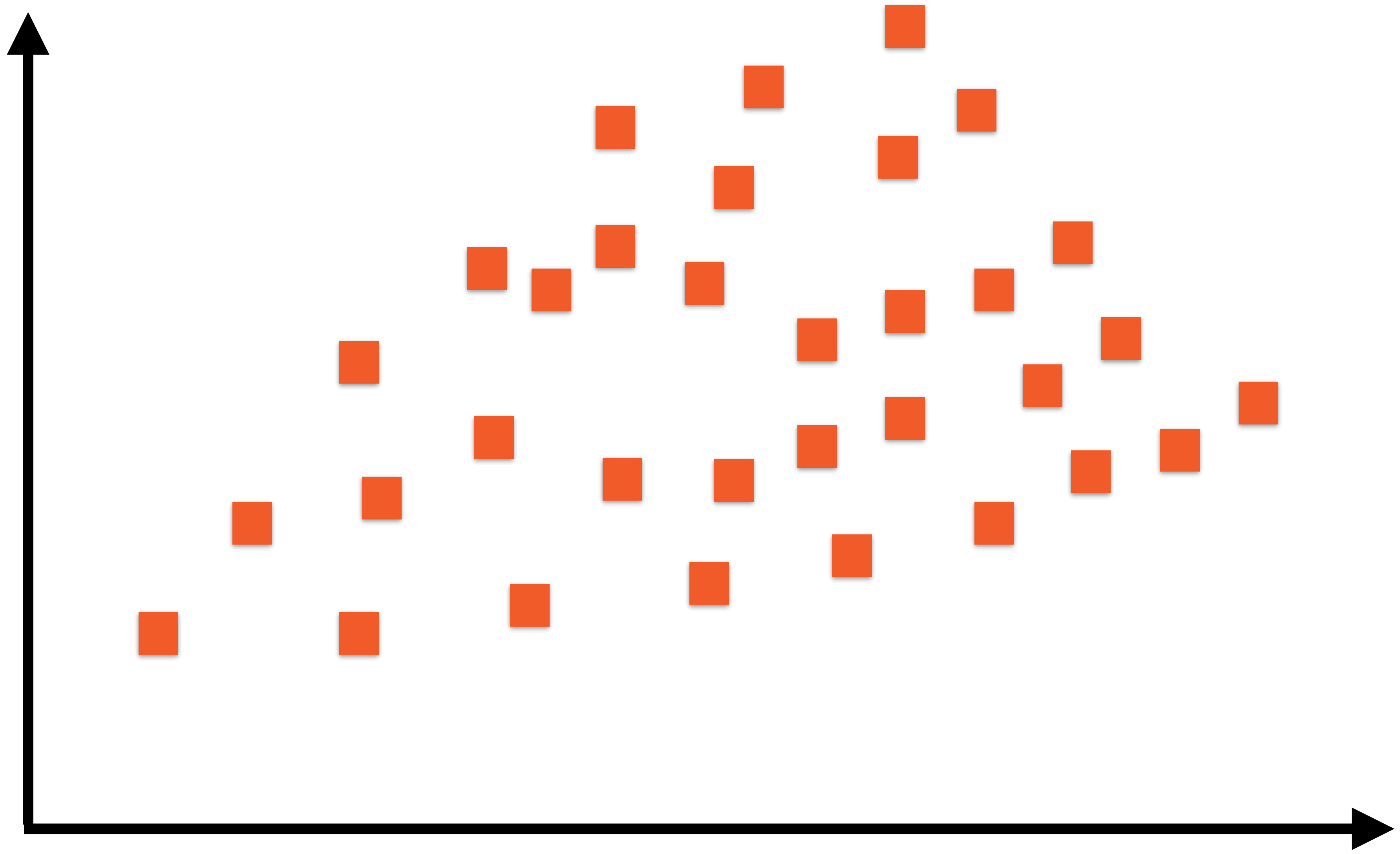
Solutions

Scatter Plot of Residuals



No clear pattern to the residuals

Scatter Plot of Residuals



Residual fan-out in a clear pattern

Detecting Heteroscedasticity

Scatter plot of residuals

- look for tell-tale fan shape

R^2 too good to be true

- non-stationary data in regression

Tests for Heteroscedasticity

- Anscombe test
- Breusch-Pagan test
- many others

Detecting Heteroscedasticity

Scatter plot of residuals

- always plot residuals

R^2 too good to be true

- any $R^2 > 80\%$ is worth a second look
- use returns, not prices

Tests for Heteroscedasticity

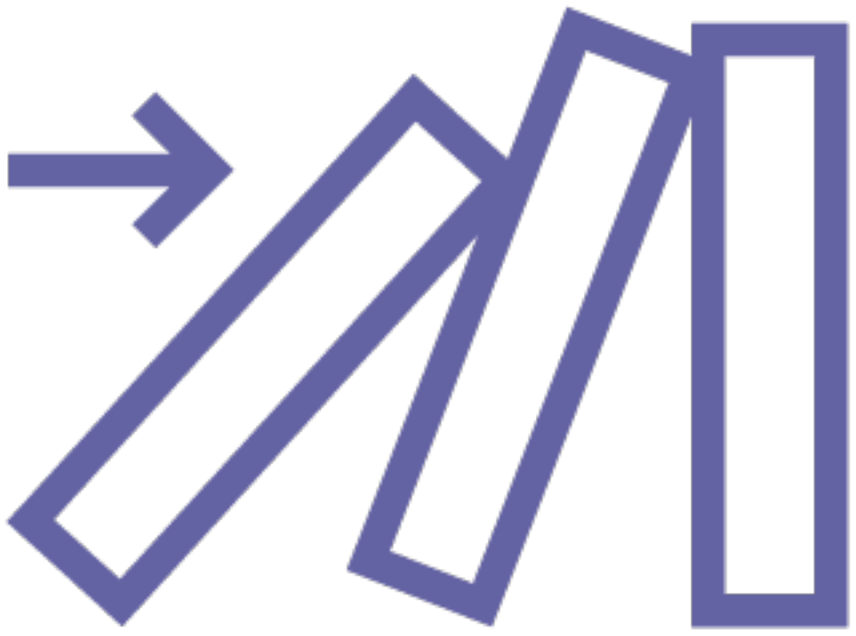
- seldom used in practice

Heteroscedasticity

Detection

Implications

Solutions



Implications of Heteroscedasticity

Overall regression equation is still unbiased

However estimates of regression parameters now biased

Confidence intervals may be worse than they appear

Using regression for prediction could be risky

Heteroscedasticity

Detection

Implications

Solutions

Solutions of Heteroscedasticity

Use transformed data

- use of log returns

Use different regression model

- weighted least squares
- generalized least squares

Generalized Least Squares (GLS)

A technique for fitting a “better” regression line between the residuals in an OLS model when they exhibit heteroscedasticity

Weighted Least Squares (WLS)

Weighted least squares (WLS) is a specialization of GLS regression

Weighted Least Squares

OLS minimizes Mean Square Error (MSE)

WLS minimizes weighted MSE

What weights to use?

Need to specify - major drawback of WLS

Weighted Least Squares Use Cases

Data is hetereoscedastic

Regression should concentrate on specific data points

Not all data points are equal

The linear regression is part of another non-linear procedure

Weighted Least Squares Drawbacks

What weights to use?

Need to specify - major drawback of WLS

Need very precise weight estimates

Sensitive to outliers

Transforming data is a more commonly used way of dealing with heteroscedasticity than using GLS or WLS

Demo

Implementing weighted least squares regression

Generalized Linear Models

Generalized Linear Models

A flexible generalization of ordinary linear regression that allows for non-normal y-variables

Generalized Linear Models

A flexible generalization of ordinary linear regression that allows for non-normal y-variables

$$Y_t = c + \sum_{i=1}^{p} \phi_i X_t + \epsilon_t$$

General Form of Linear Model

Same equation as OLS, but now Y can be non-normal, even categorical

Possible Y Distributions

Binomial

Poisson

Gamma

Possible Y Distributions

Binomial

Poisson

Gamma

Binomial

Binomial Y Variables

Categorical data: discrete values

Binary: 0 or 1, True or False

Binomial: sum of binary variables in each category

Elements of a GLM

Probability distribution of Y

Normal, Binomial, Categorical and more

Mean function

Relationship between regression parameters and mean of Y

Link function

Transformation to make X-Y relationship linear

Logistic Regression as Example

Probability distribution of Y

Binary categorical

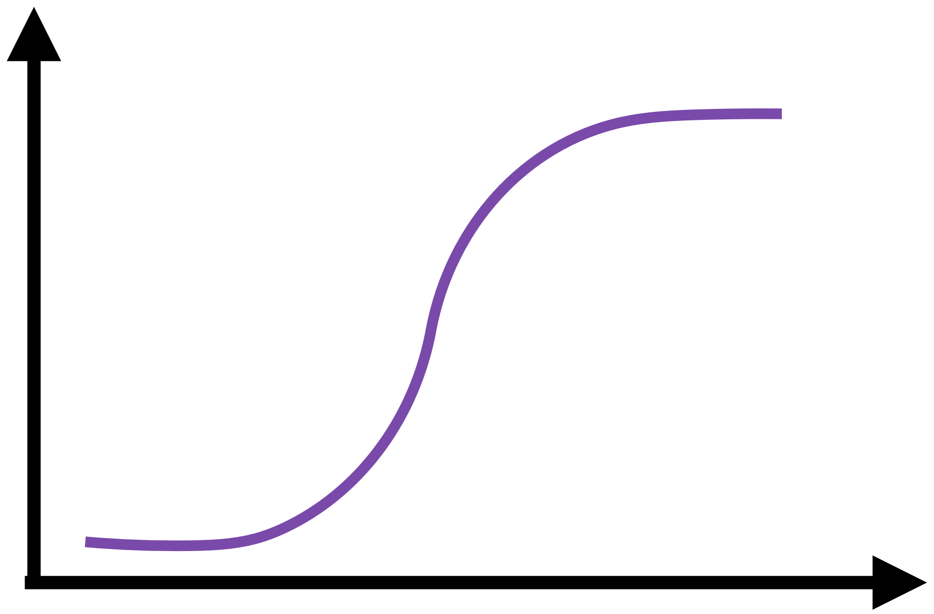
Mean function

S-curve equation

Link function

Logit function

Logistic regression can be performed using GLM



Logistic Regression

S-curves are widely studied, well-understood

$$y = \frac{1}{1 + e^{-(A+Bx)}}$$

Logistic regression uses S-curve to estimate probabilities

Logistic Regression

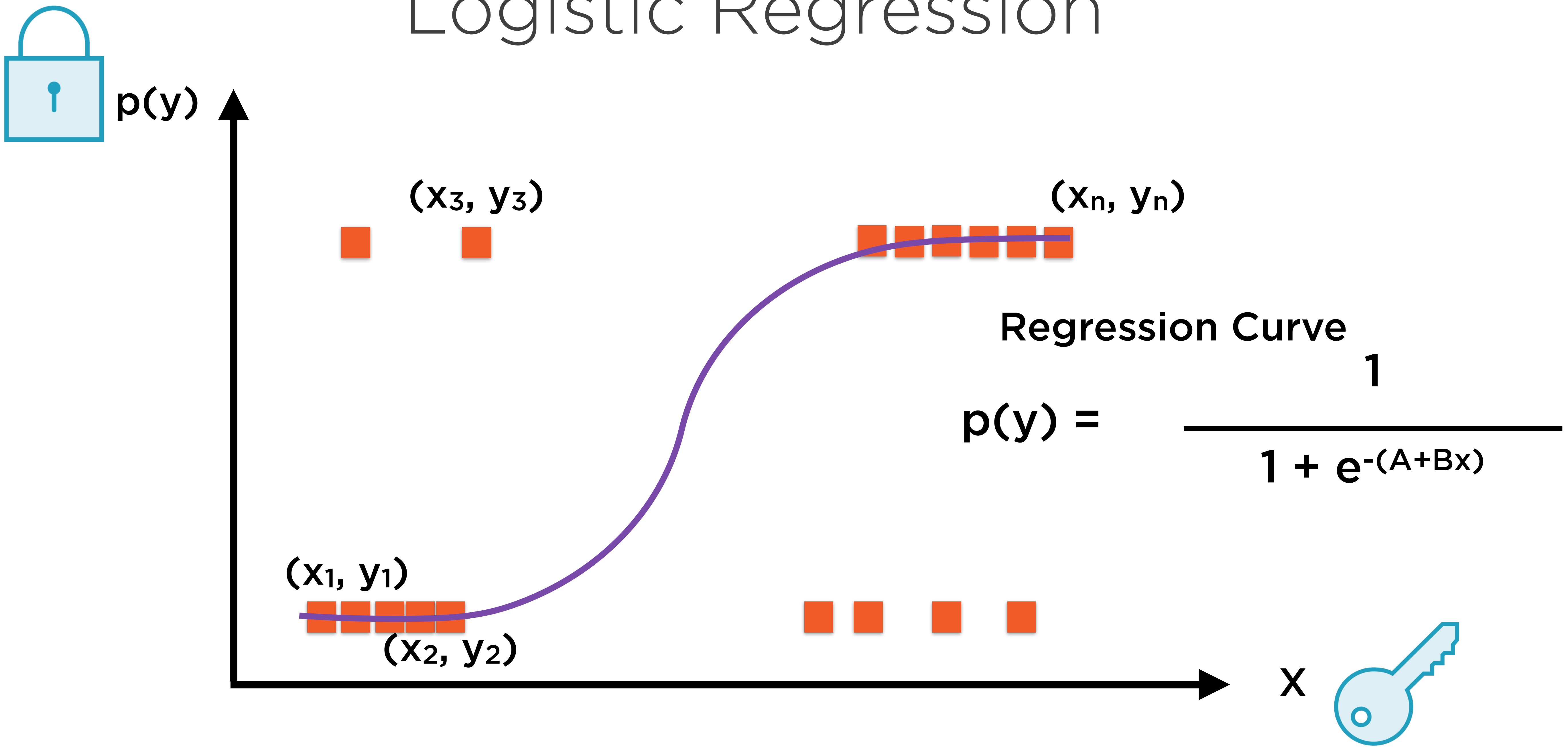
Logistic Regression

Regression Equation:

$$p(y_i) = \frac{1}{1 + e^{-(A+Bx_i)}}$$

Given a set of points where x "predicts" probability of success in y, use logistic regression

Logistic Regression



“It just works”: GLMs are a great way to fit linear models to binary or multinomial data without going deep into math

Demo

Implement generalised linear regression for a binomial Y distribution

Robust Linear Models

Regression using OLS works well when the **basic assumptions** about the underlying data are true

OLS regression is highly sensitive to outliers

Robust Linear Models

Modified regression algorithms that perform better than OLS in the presence of outliers (and also in cases of heteroscedasticity)

Robust Regression

Usually superior to OLS regression

Still not as popular

- complex to understand
- multiple competing algorithms
- computationally intensive
- not supported in Excel and other popular tools

Demo

Implement robust linear regression

Summary

Ordinary least squares regression makes many assumptions about data

Generalized or weighted least squares for heteroscedasticity

Generalized linear models for non-normal y variables

Robust linear models to cope with outliers