Implementing Simple Regression Models in Excel

Vitthal Srinivasan
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Build regression models in Excel

Understand and test the regression assumptions

Use simple regression models in Excel

- to explain variance
- to make forecasts

Avoid some common regression pitfalls

Applying Simple Regression

Simple Regression

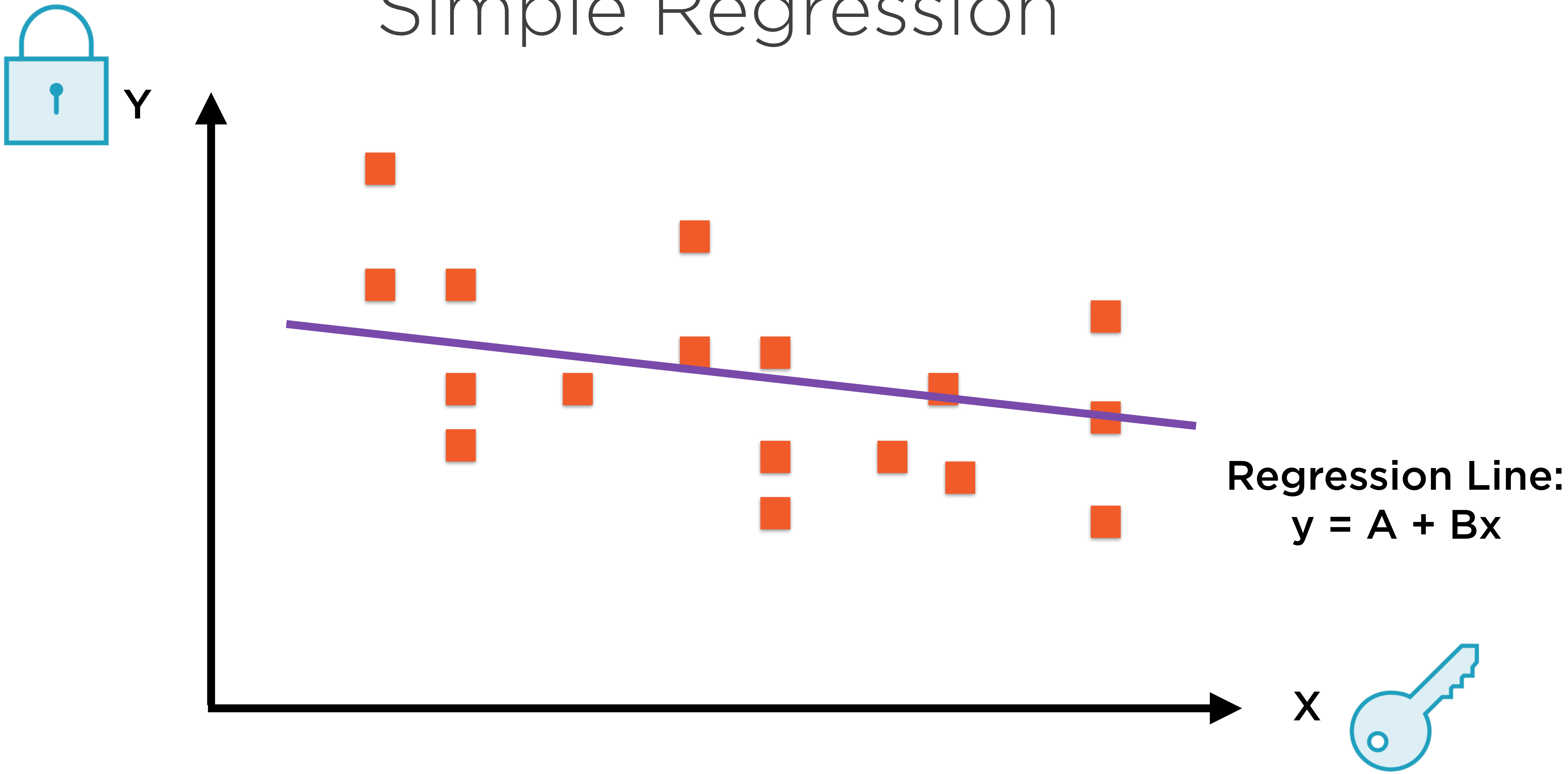
Cause

Changes in Dow Jones equity index

Effect

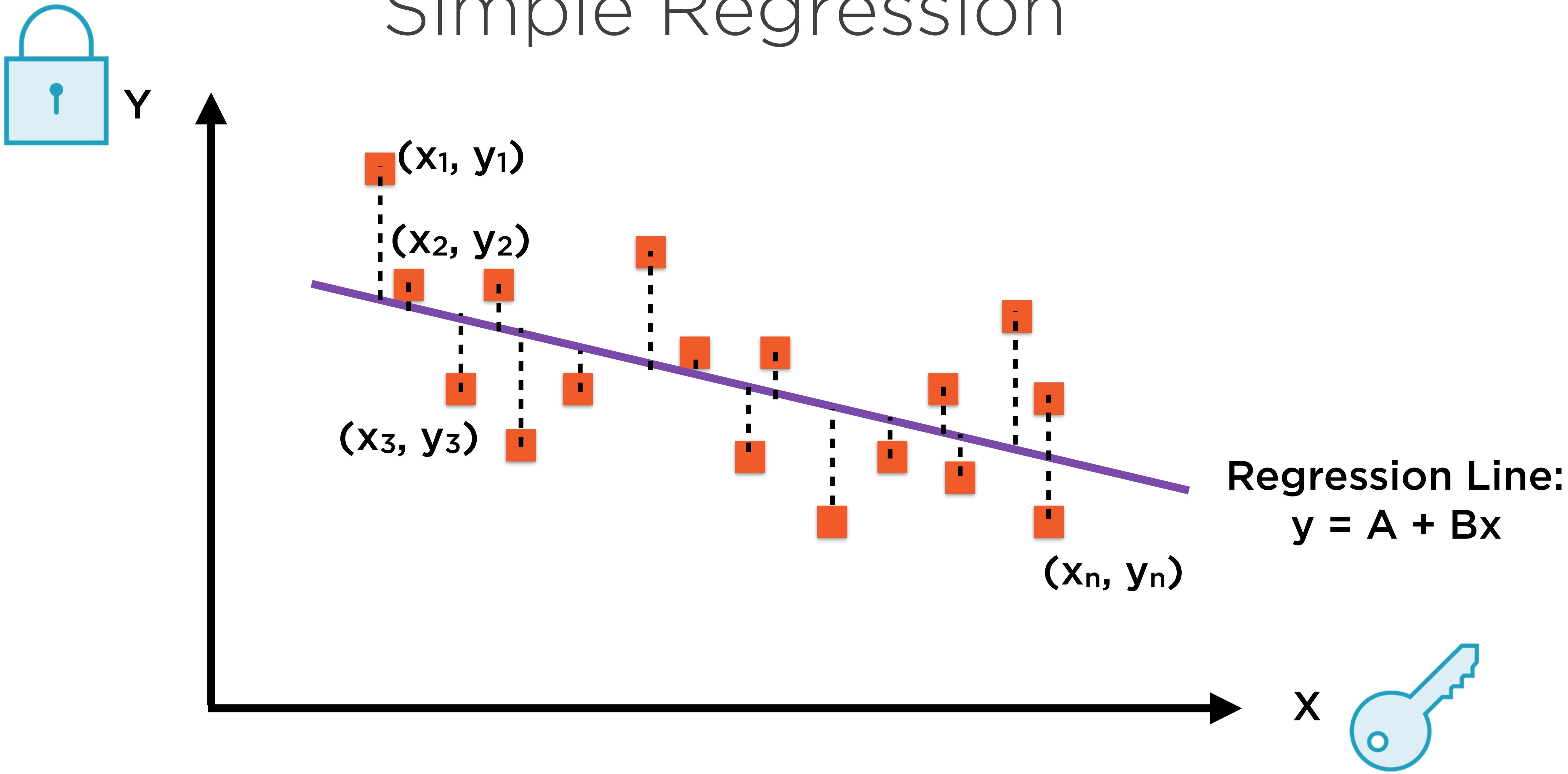
Changes in price of Exxon Stock

Simple Regression



Find the equation of the regression line, measure goodness-of-fit

Simple Regression



Represent all n points as (x_i, y_i), where i = 1 to n

Simple Regression

Regression Equation:

$$y = A + Bx$$

$$
\begin{aligned}
y_1 &= A + Bx_1 \\
y_2 &= A + Bx_2 \\
y_3 &= A + Bx_3 \\
&\ldots \\
y_n &= A + Bx_n
\end{aligned}
$$

Simple Regression

Regression Equation:

$$y = A + Bx$$

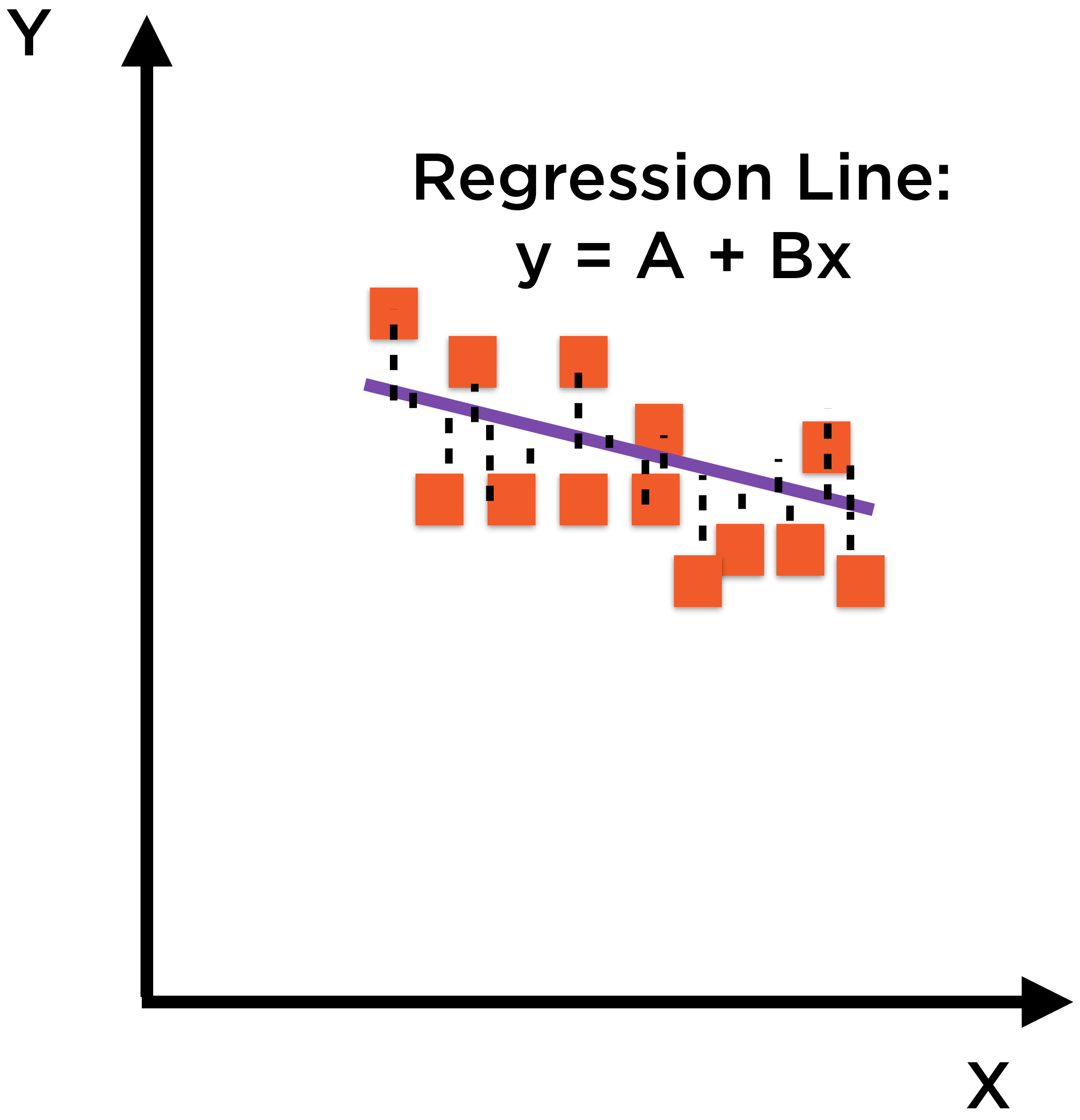
$$
\begin{aligned}
y_1 &= A + Bx_1 + e_1 \\
y_2 &= A + Bx_2 + e_2 \\
y_3 &= A + Bx_3 + e_3 \\
&\dots \\
y_n &= A + Bx_n + e_n
\end{aligned}
$$

Simple Regression

Regression Equation:

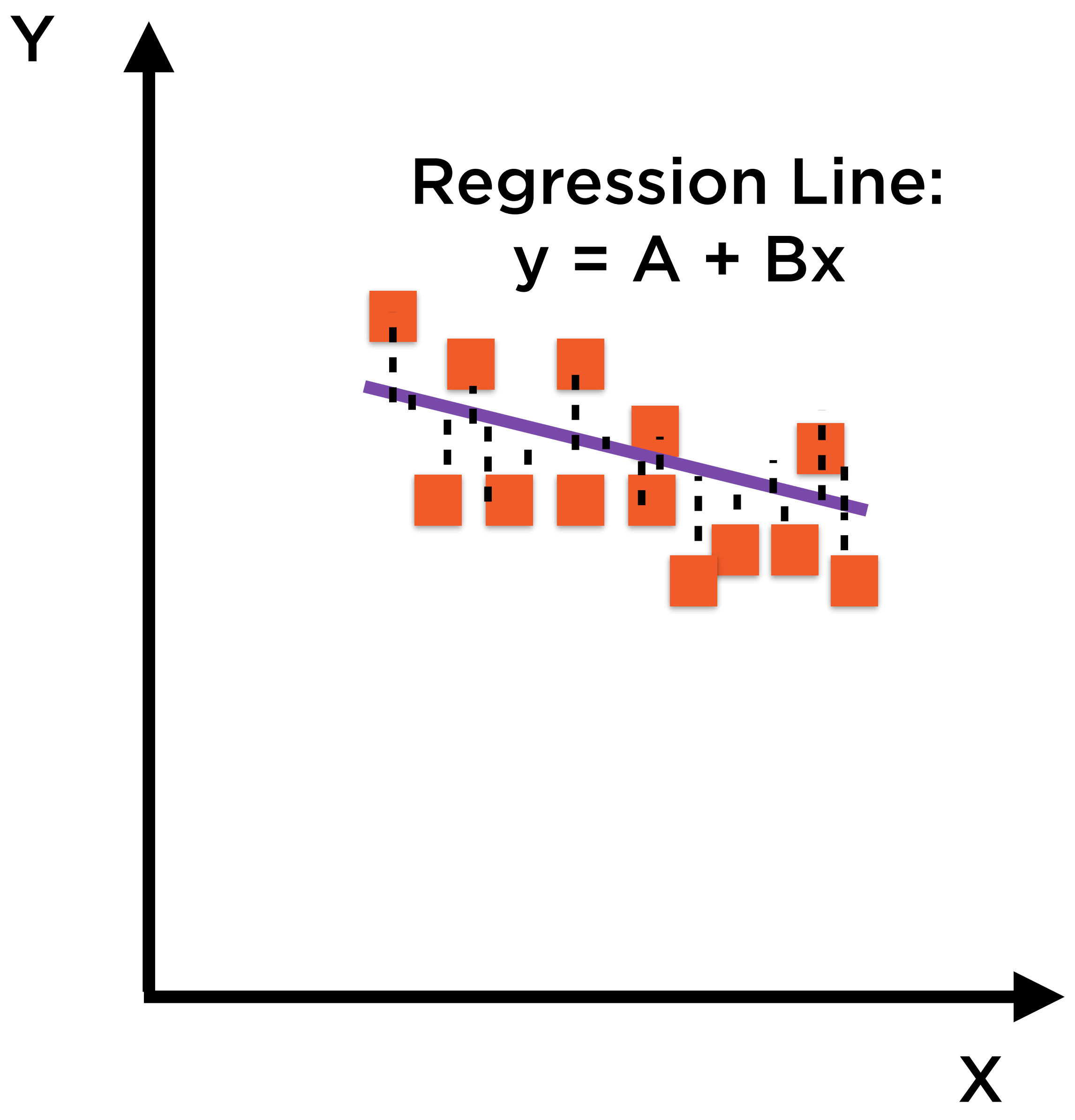
$$y = A + Bx$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} + B \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_n \end{bmatrix}$$



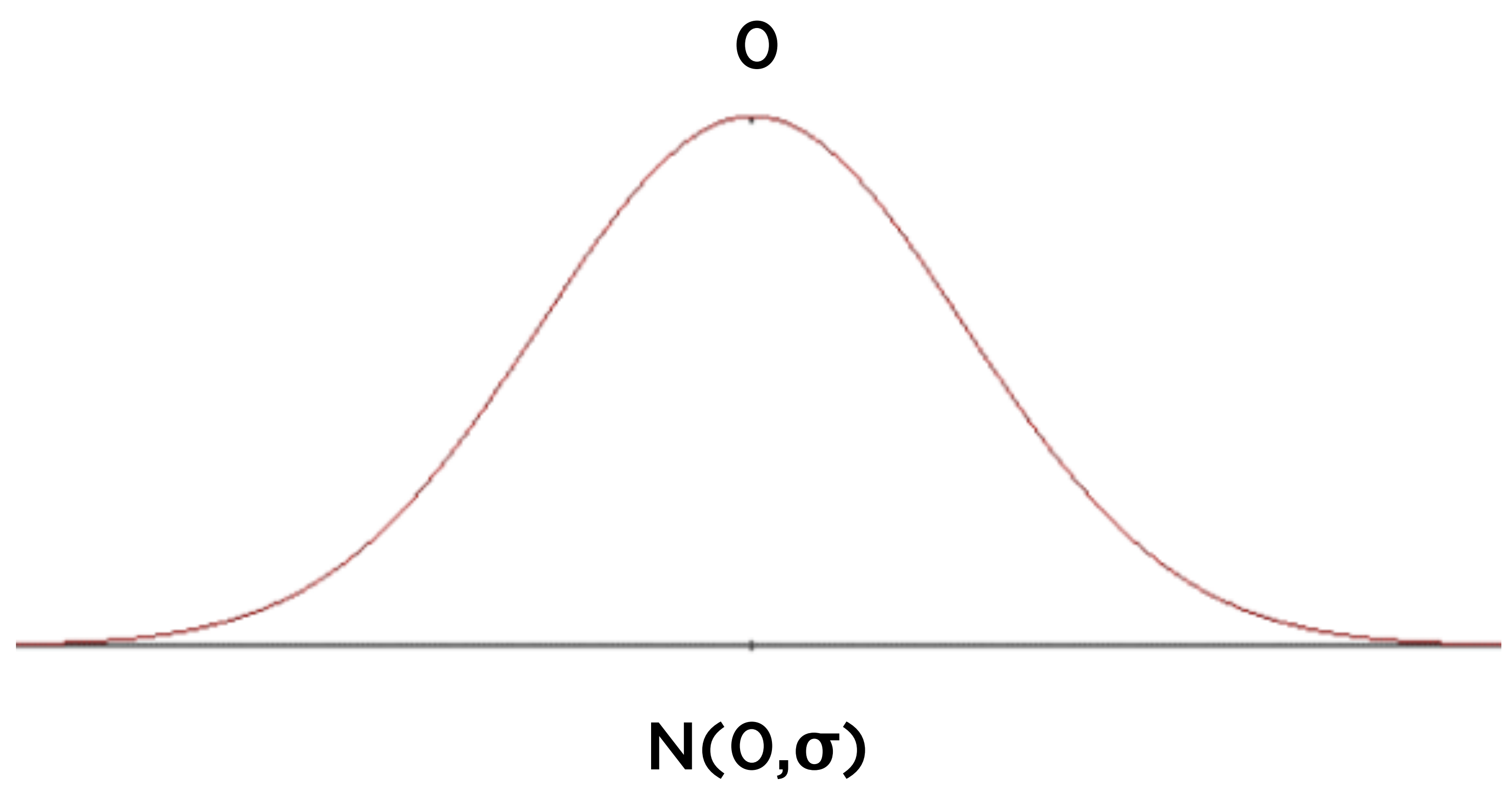
Ideally, residuals should

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



Ideally, residuals should

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



$e \sim N(0,\sigma^2)$

Zero-mean, Common Variance, Normal

Three assumptions relate to probability distribution of residuals

$e = y - y'$

$\Rightarrow \quad y = y' + e$

$\Rightarrow \quad \text{Mean}(y) = \text{Mean}(y') + \text{Mean}(e)$

$\Rightarrow \quad \text{Mean}(y) = \text{Mean}(y')$

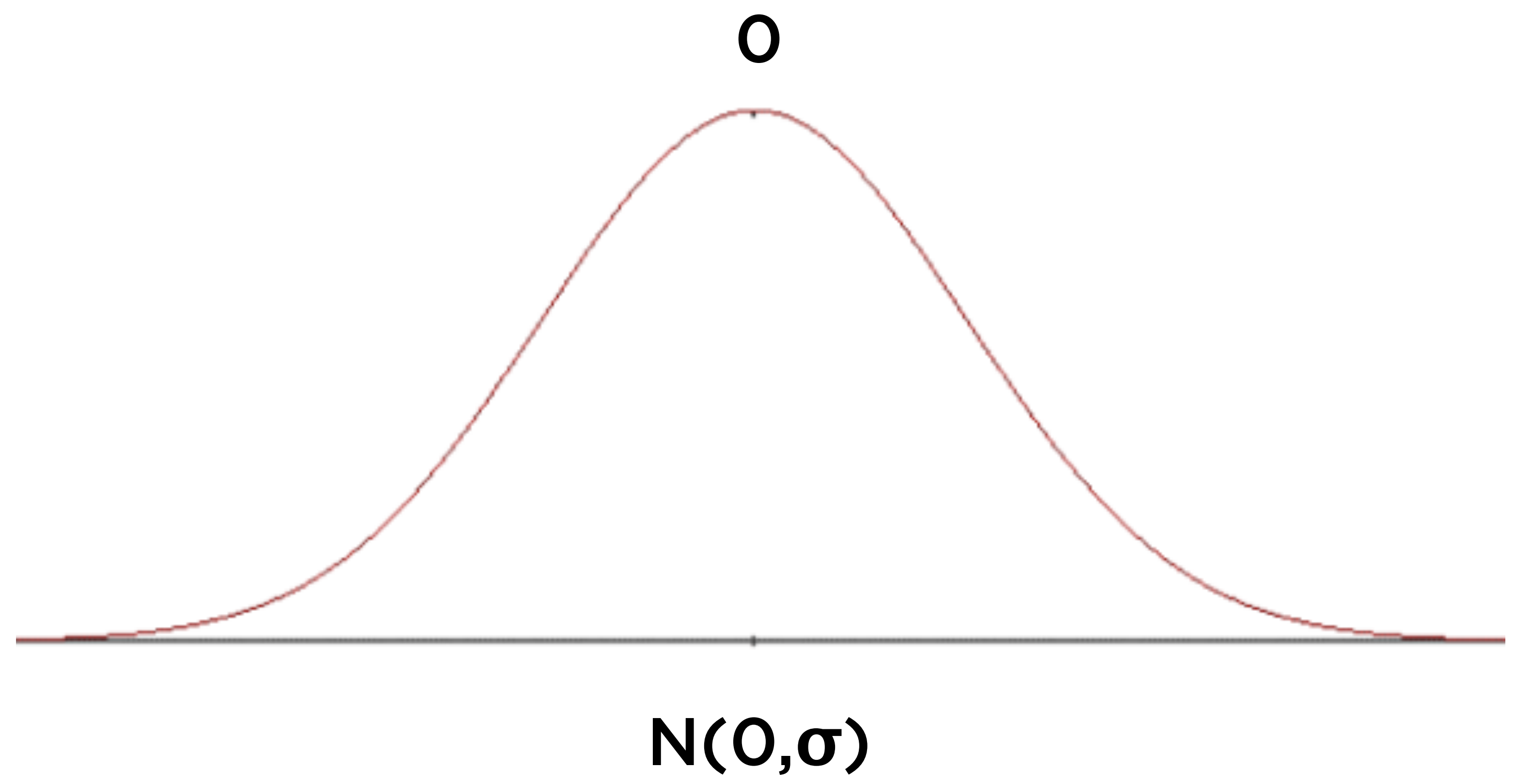
Zero-mean: Always Satisfied

The procedure of least-squares ensures this - no need to check

Mean(y) = Mean(y')

Sample Mean = Regression Mean

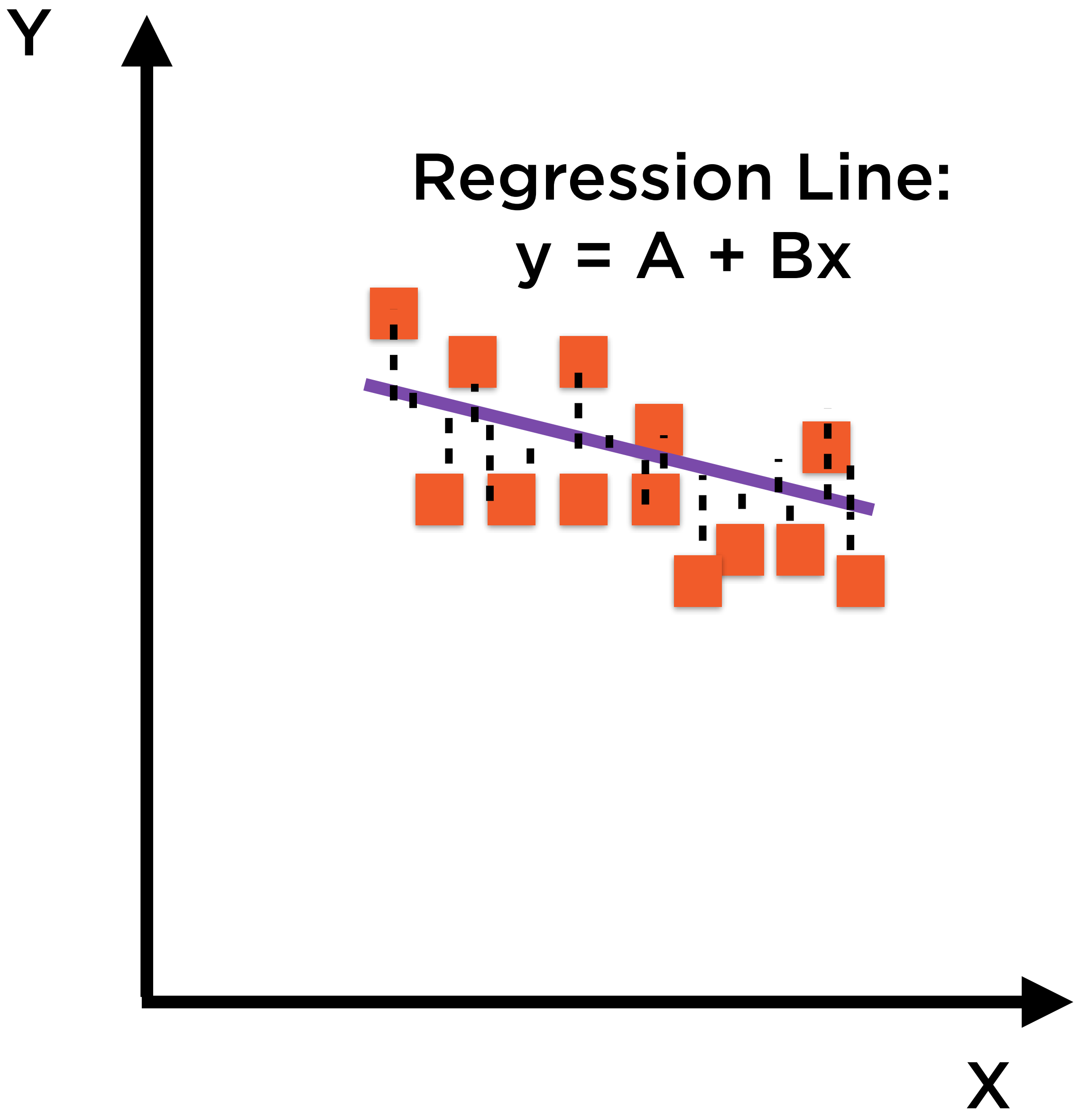
The procedure of least-squares ensures this - no need to check



$e \sim N(0,\sigma^2)$

Common Variance, Normal: Harder to Check

Hard to check directly - usually indirectly checked



Ideally, residuals should

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

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

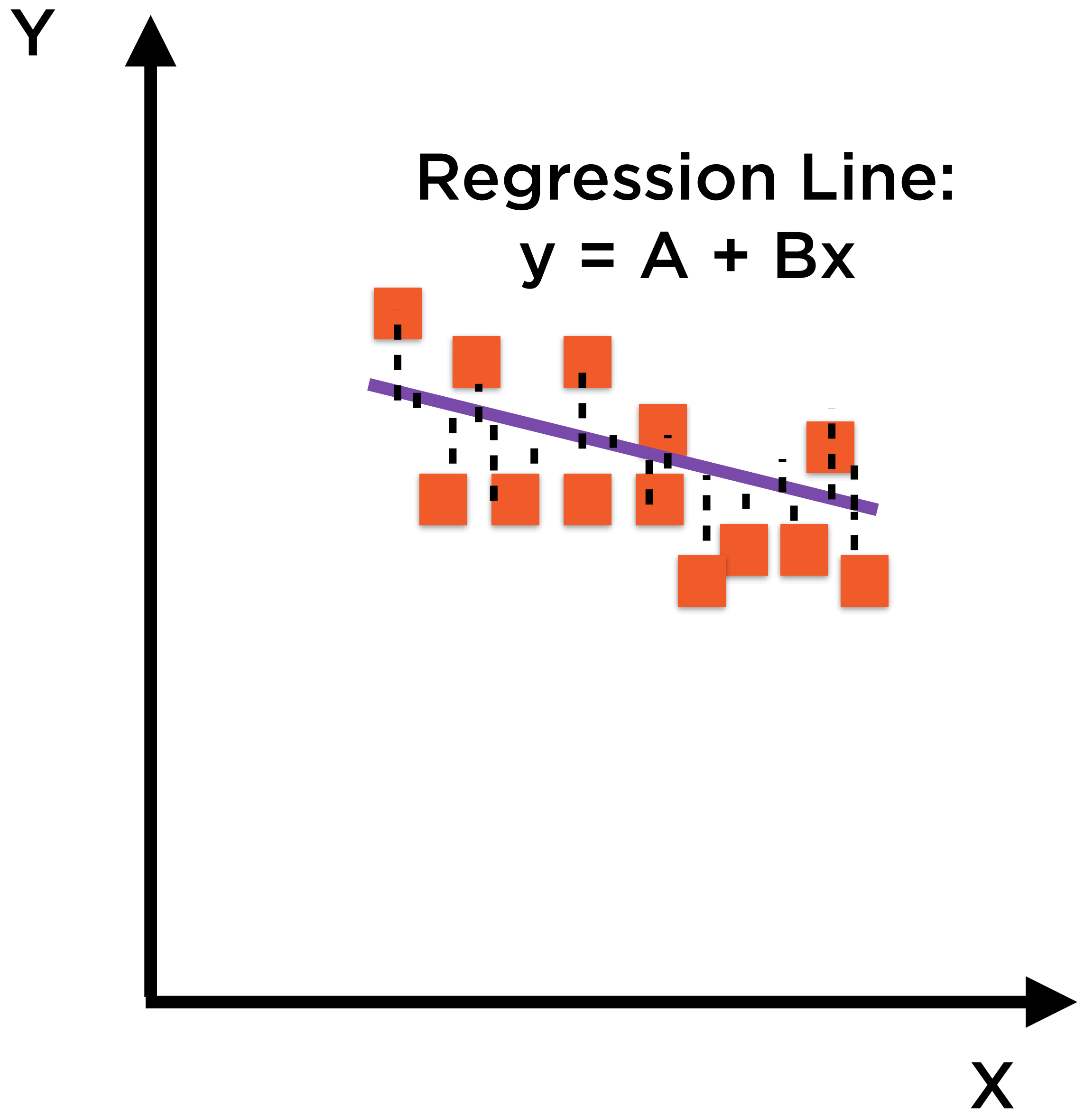
$e^1 = [e_1, e_2, e_3 \ldots e_{n-1}]$

$e^2 = [e_2, e_3, e_4 \ldots e_n]$

$\text{correl}(e^2, e^1) = 0$

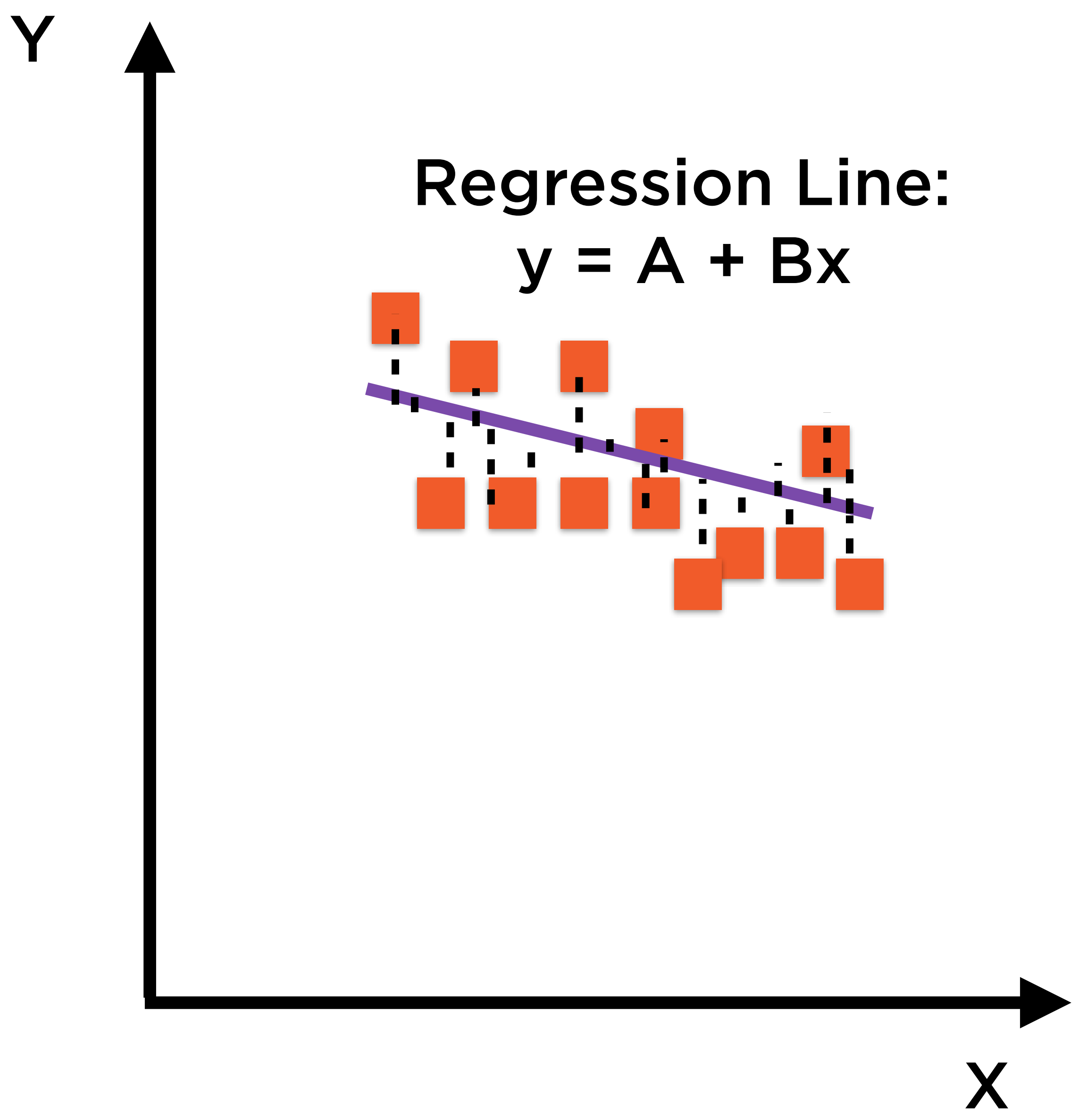
Self-Independence => Zero Auto-correlation

Shift residuals by 1,2... and measure correlation with self



Ideally, residuals should

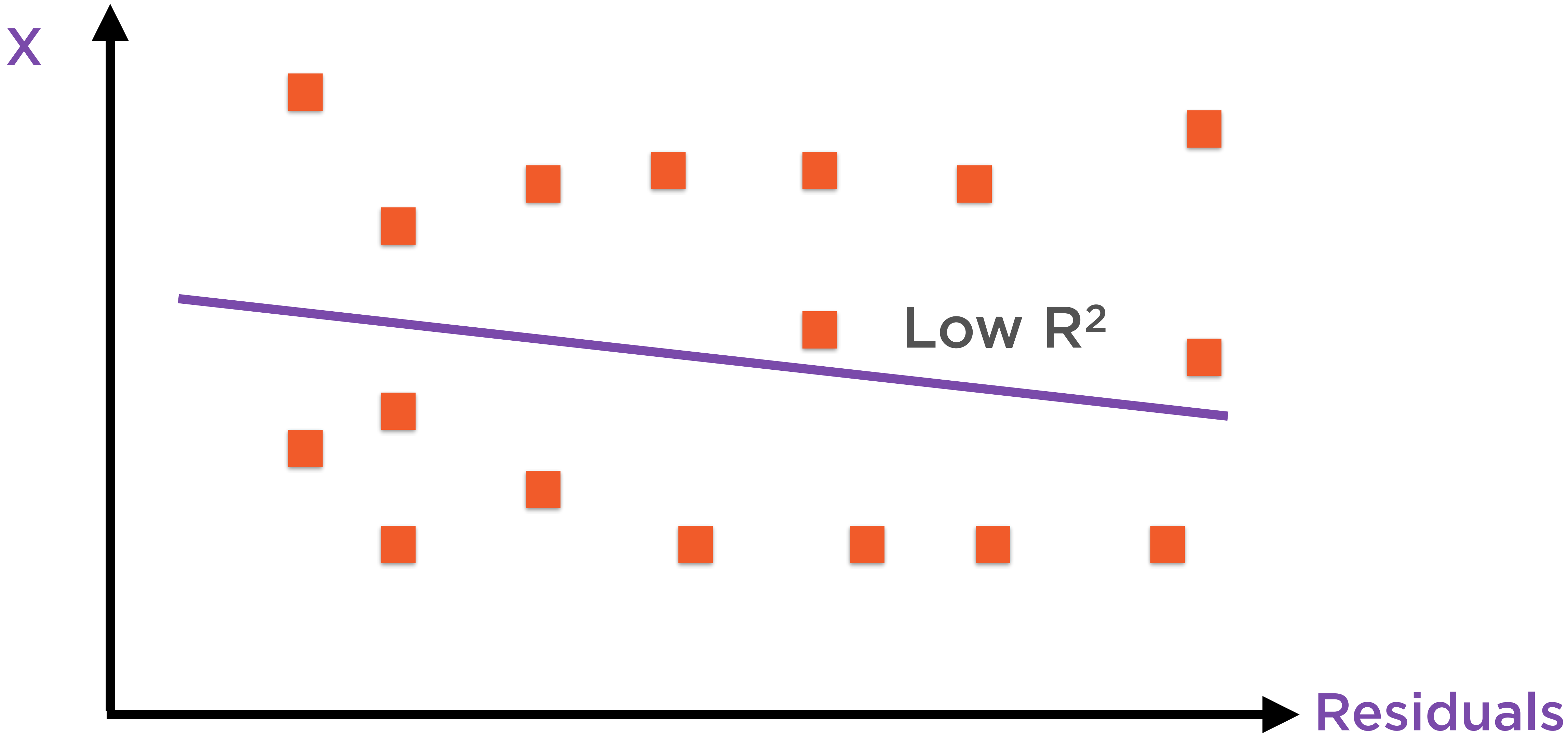
- have zero mean
- common variance
- be independent of each other
- be independent of x
- be normally distributed



Ideally, residuals should

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

Independence from X



Residuals are independent of X

Violations of Regression Assumptions

Risks in Simple Regression

No cause-effect relationship

Regression on completely unrelated data series

Mis-specified relationship

Non-linear (exponential or polynomial) fit

Incomplete relationship

Multiple causes exist, we have captured just one

Risks in Simple Regression

No cause-effect relationship

Regression on completely unrelated data series

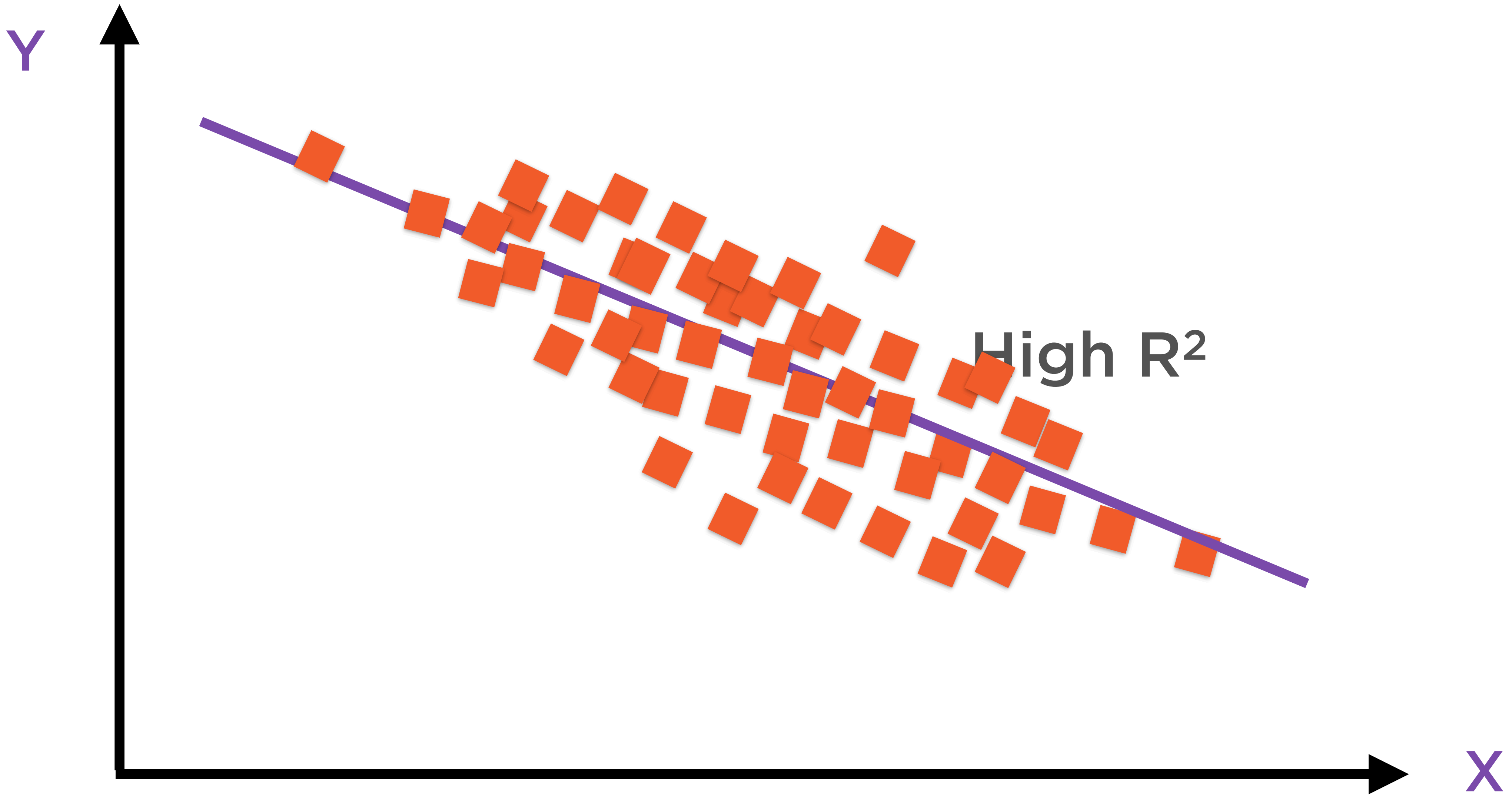
Mis-specified relationship

Non-linear (exponential or polynomial) fit

Incomplete relationship

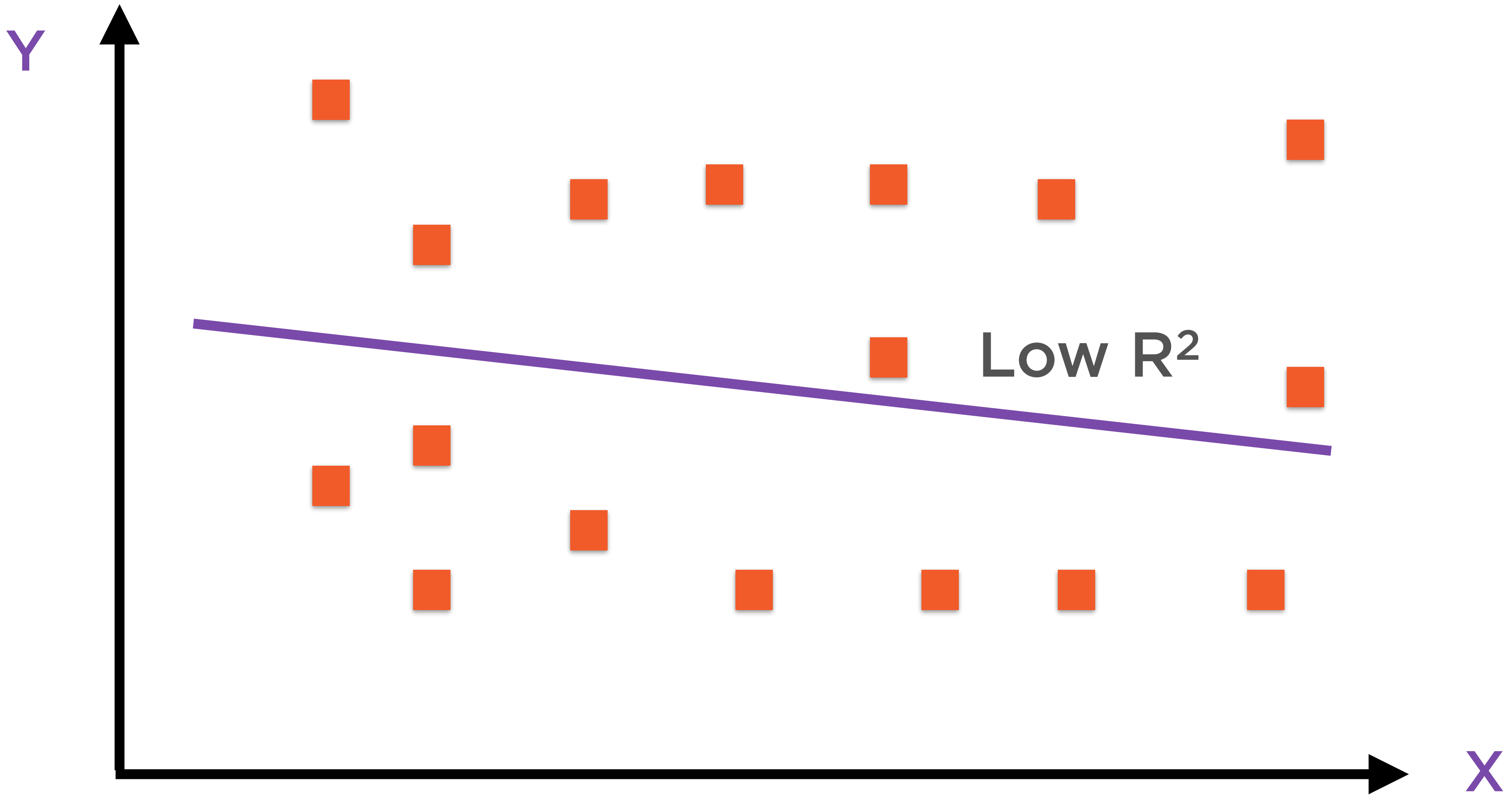
Multiple causes exist, we have captured just one

Strong Cause-Effect Relationship



Scatter plot of X and Y

Weak Cause-Effect Relationship



Abandon this model, go back to the data

Risks in Simple Regression

No cause-effect relationship

Regression on completely unrelated data series

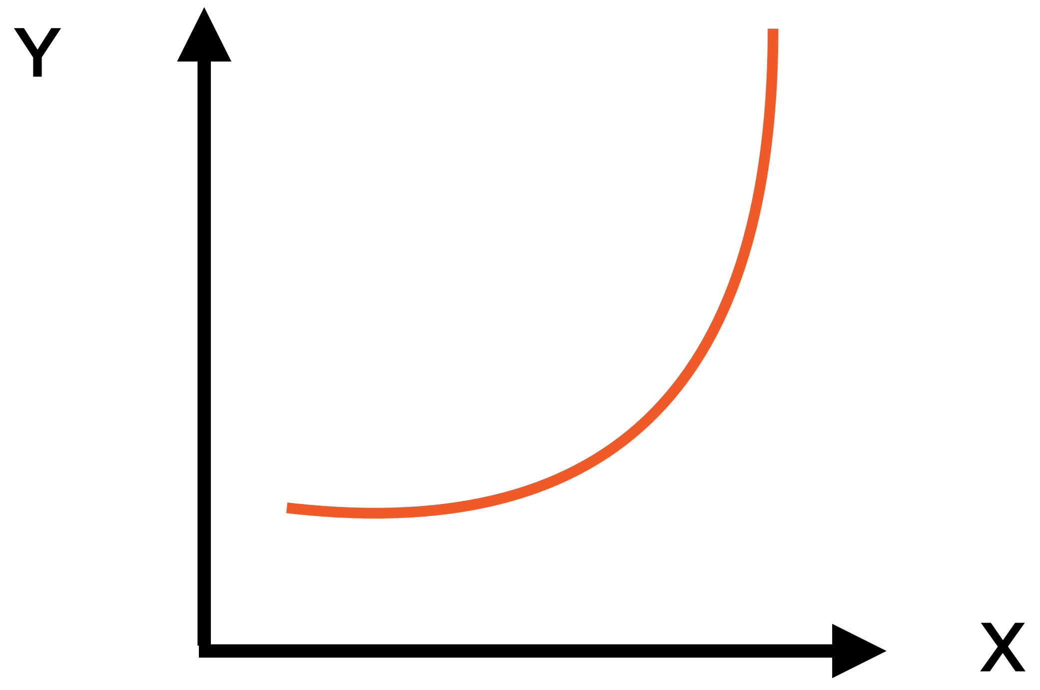
Mis-specified relationship

Non-linear (exponential or polynomial) fit

Incomplete relationship

Multiple causes exist, we have captured just one

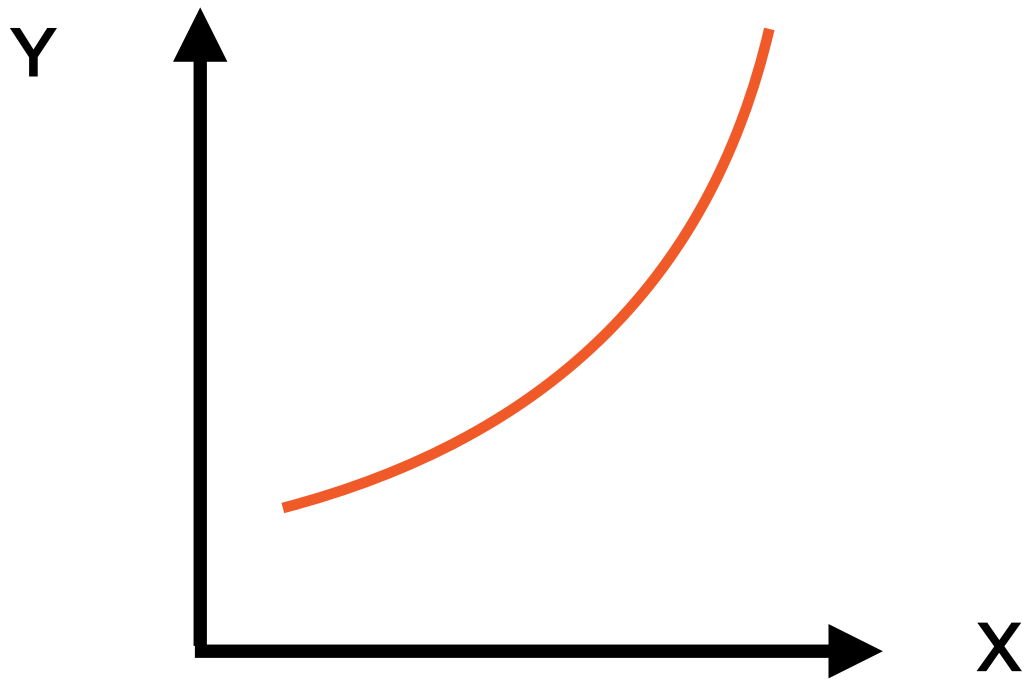
Transform Non-linear Data



Exponential

$y = A + Be^x$

Transform using logarithms

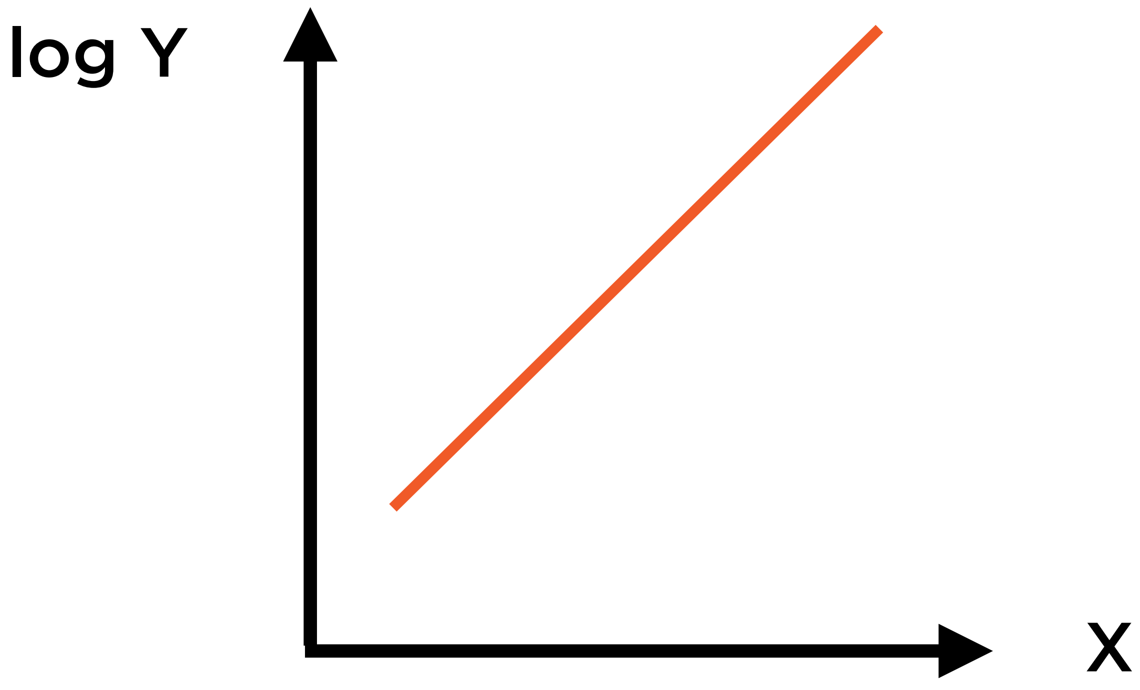


Polynomial

$y = A + Cx^2$

Transform using logarithms or simply regress on x^2

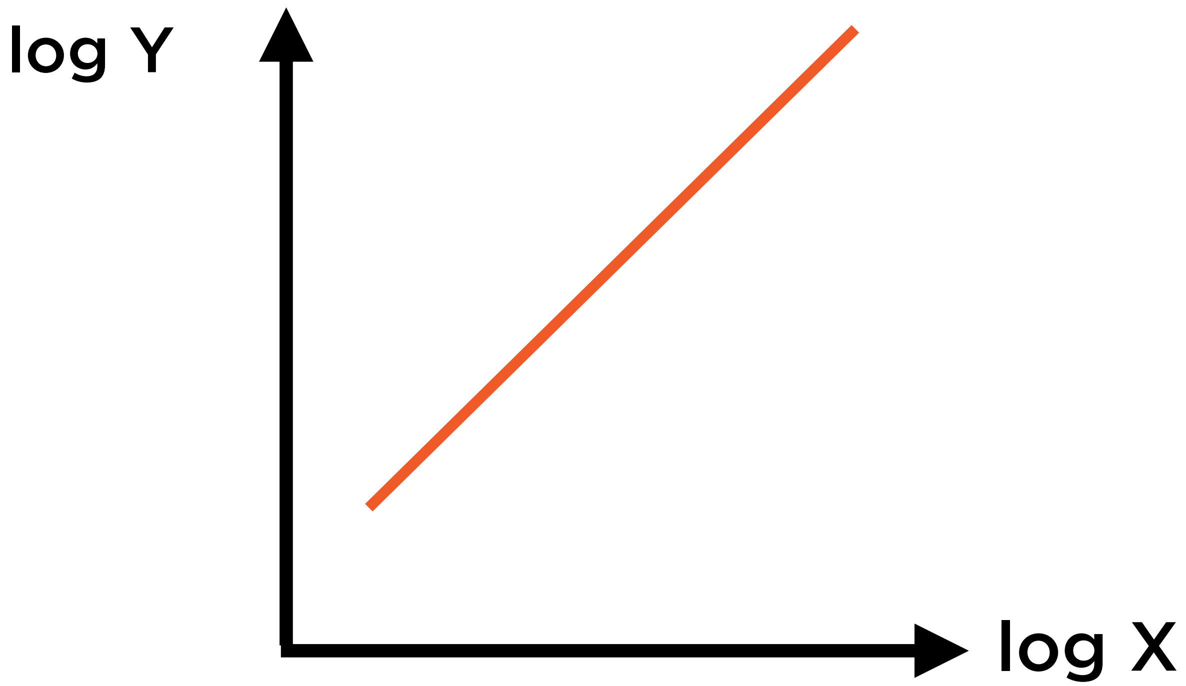
Transform Non-linear Data



Exponential

$$\log y = C + Dx$$

Now regress log y on x

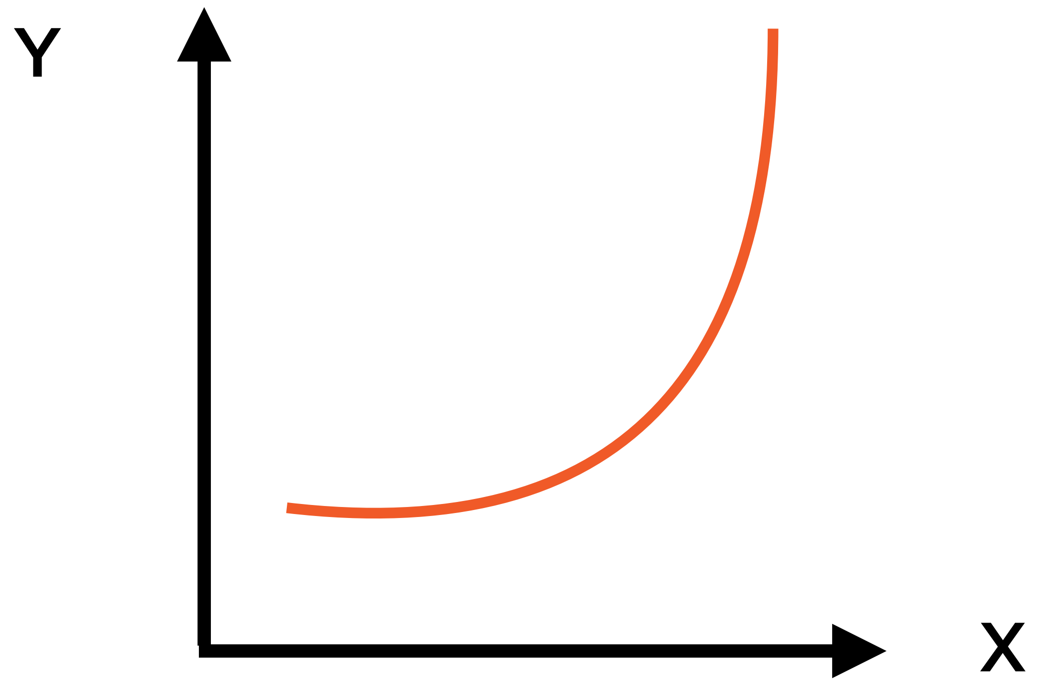


Polynomial

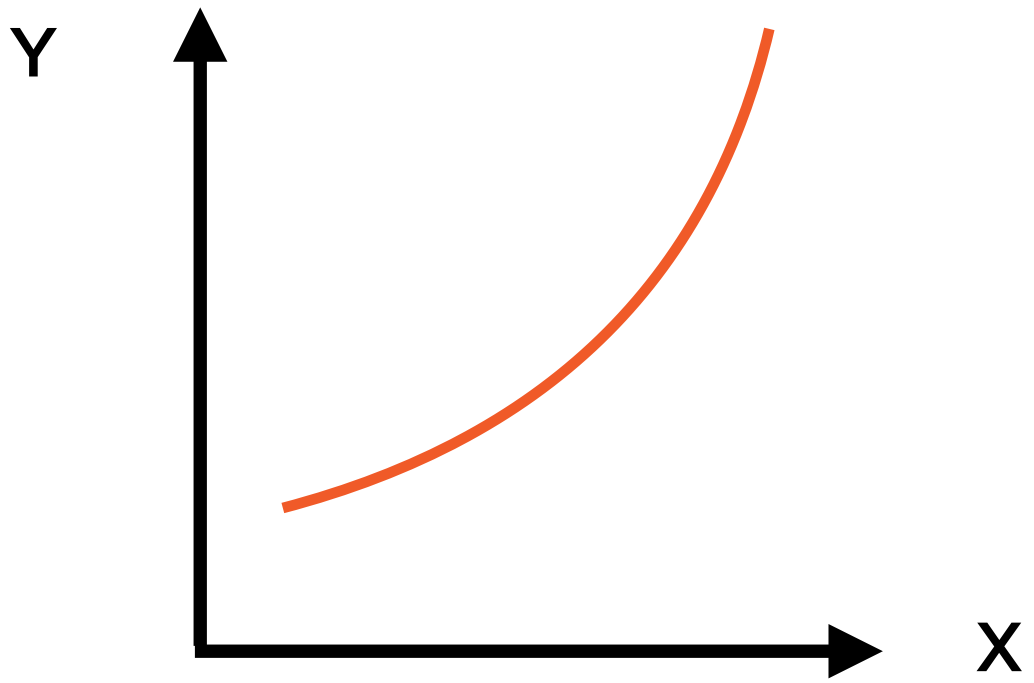
$$\log y = C + D \log x$$

or simply regress y on x^2

Never Regress Non-Stationary Data



Exponential



Polynomial

Smoothly trending data will lead to poor quality regression models

First Differences

$y'_{12} = \log y_2 - \log y_1$

$x'_{12} = \log x_2 - \log x_1$

Regress y' and x'

Log Differences

$y'_{12} = (y_2 - y_1)/y_1$

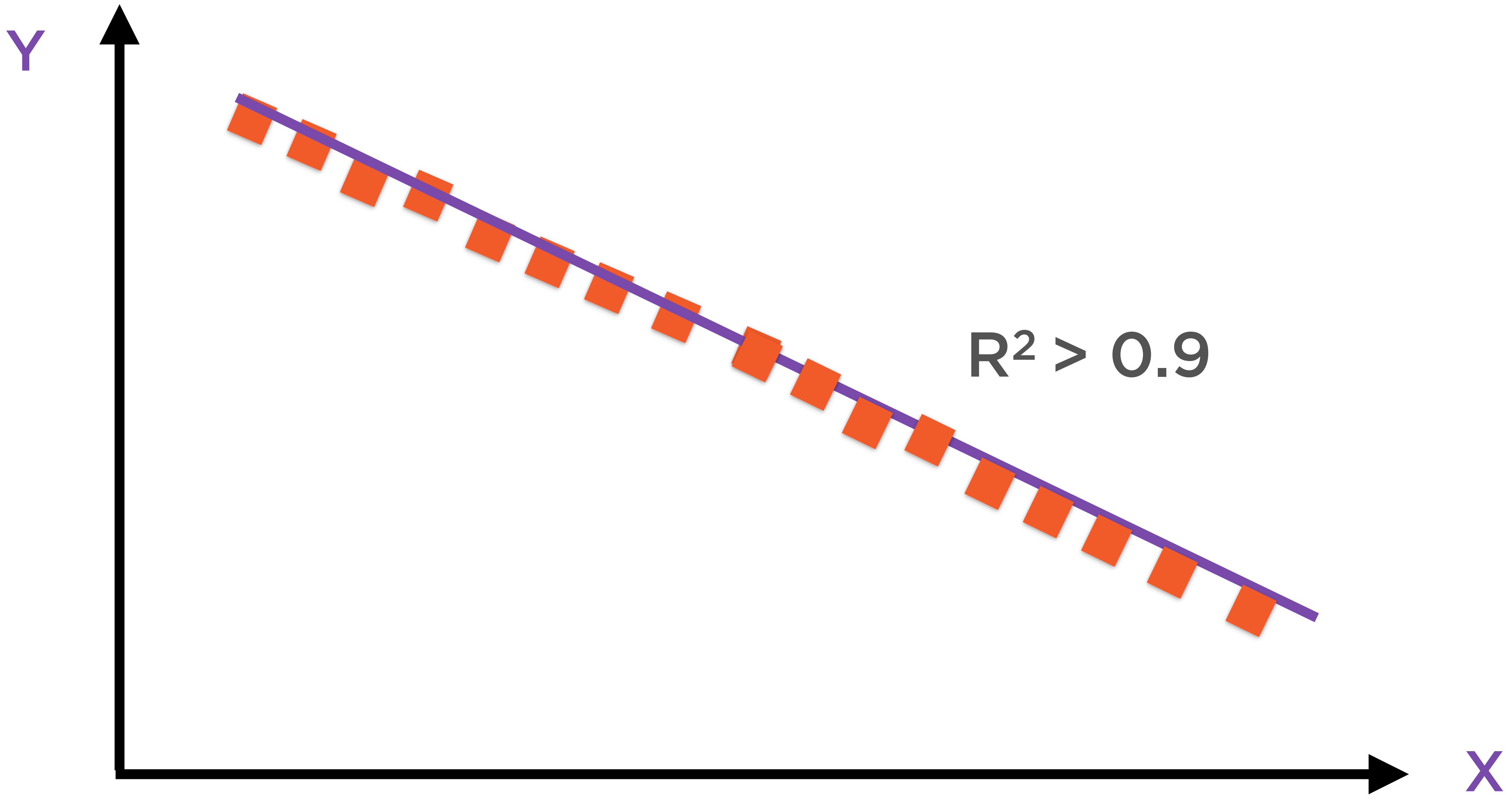
$x'_{12} = (x_2 - x_1)/x_1$

Regress y' and x'

Returns

Take first differences of smooth data converting either to log differences or returns

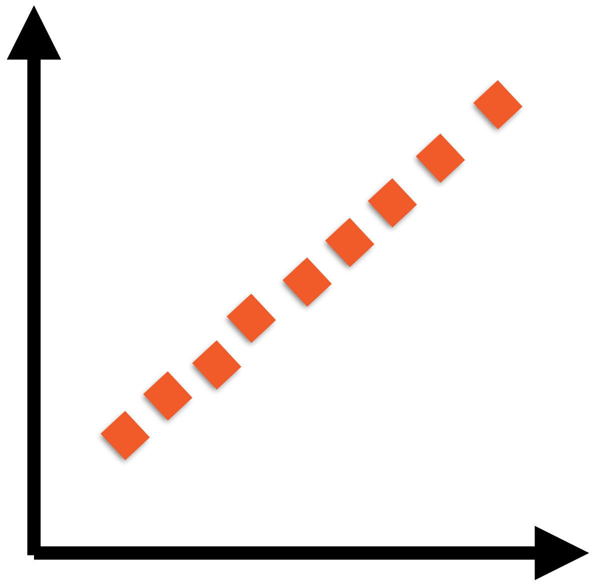
Beware of Perfect Fits



Scrutinize residuals for independence

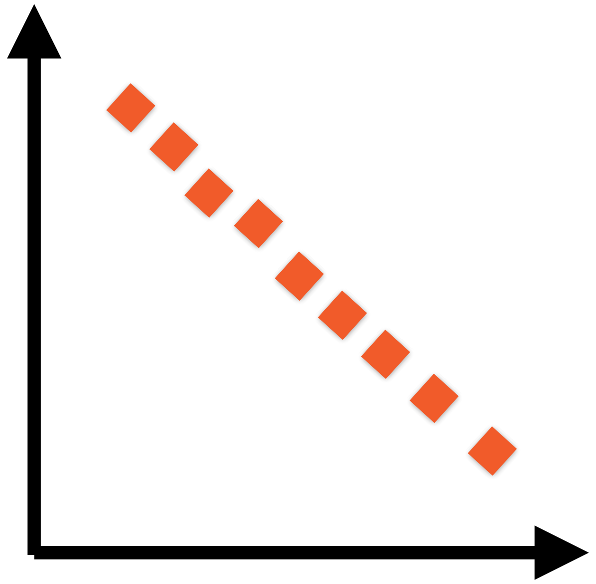
Independence is hard to quantify, so we measure **correlation** instead

Zero Correlation Usually Implies Independence



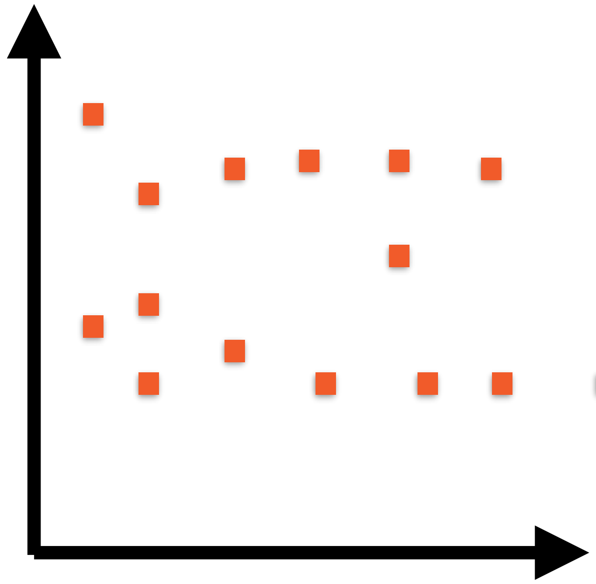
Correlation = +1

As X increases, Y increases linearly



Correlation = -1

As X increases, Y decreases linearly



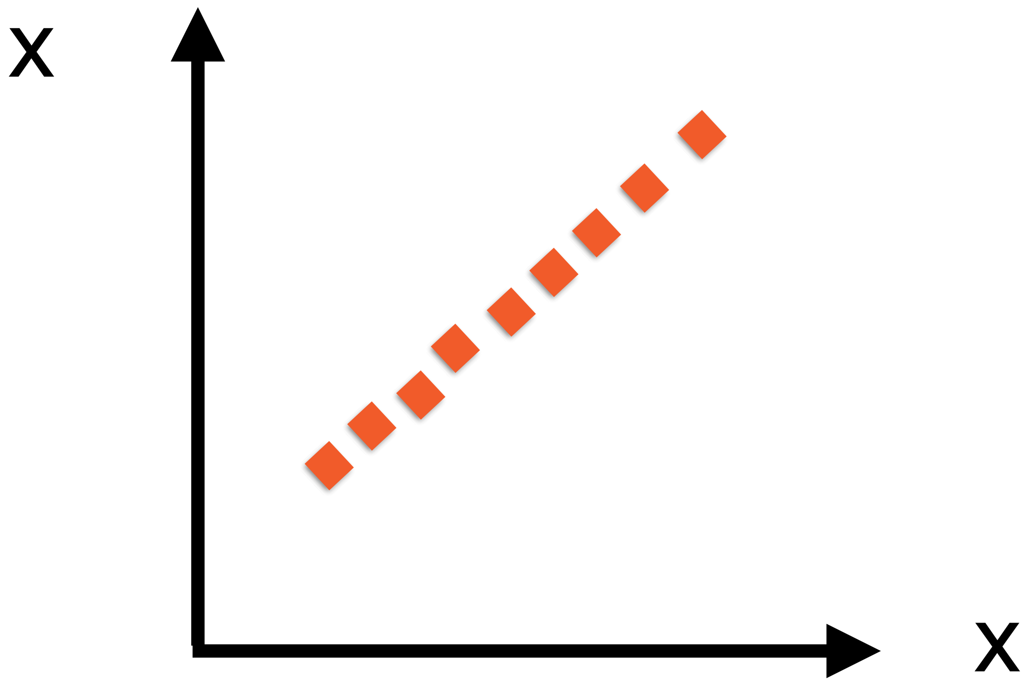
Correlation = 0

Changes in X independent of changes in Y

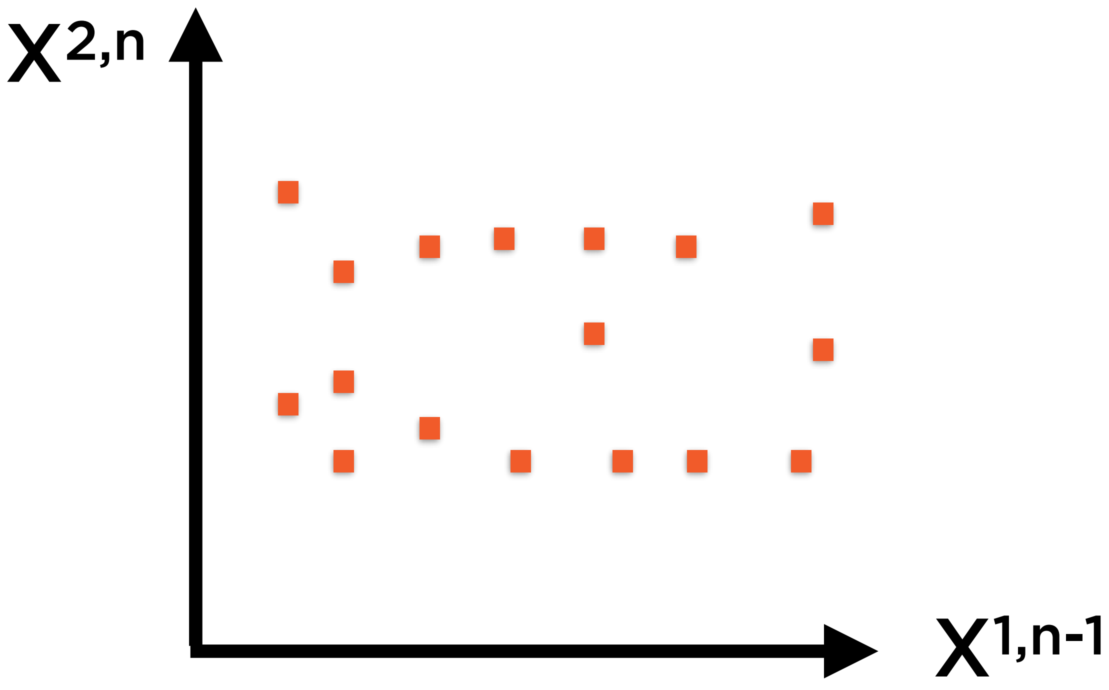
Lag-1 Autocorrelation

X	X_1	X_2	X_3	X_4	…	X_n
$X^{2,n}$		X_2	X_3	X_4	…	X_n
$X^{1,n-1}$	X_1	X_2	X_3	…	X_{n-1}	

Lag-1 Autocorrelation



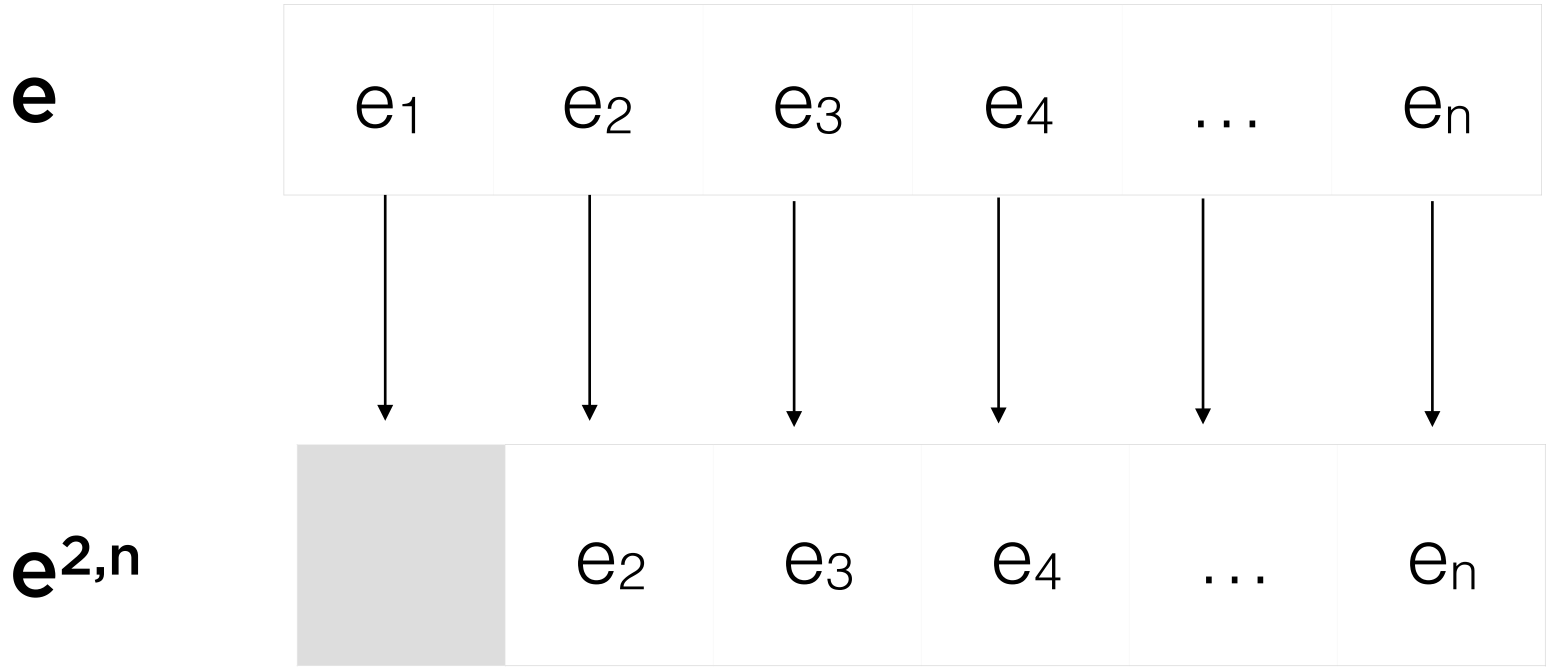
Correlation with self



Lag-1 Autocorrelation

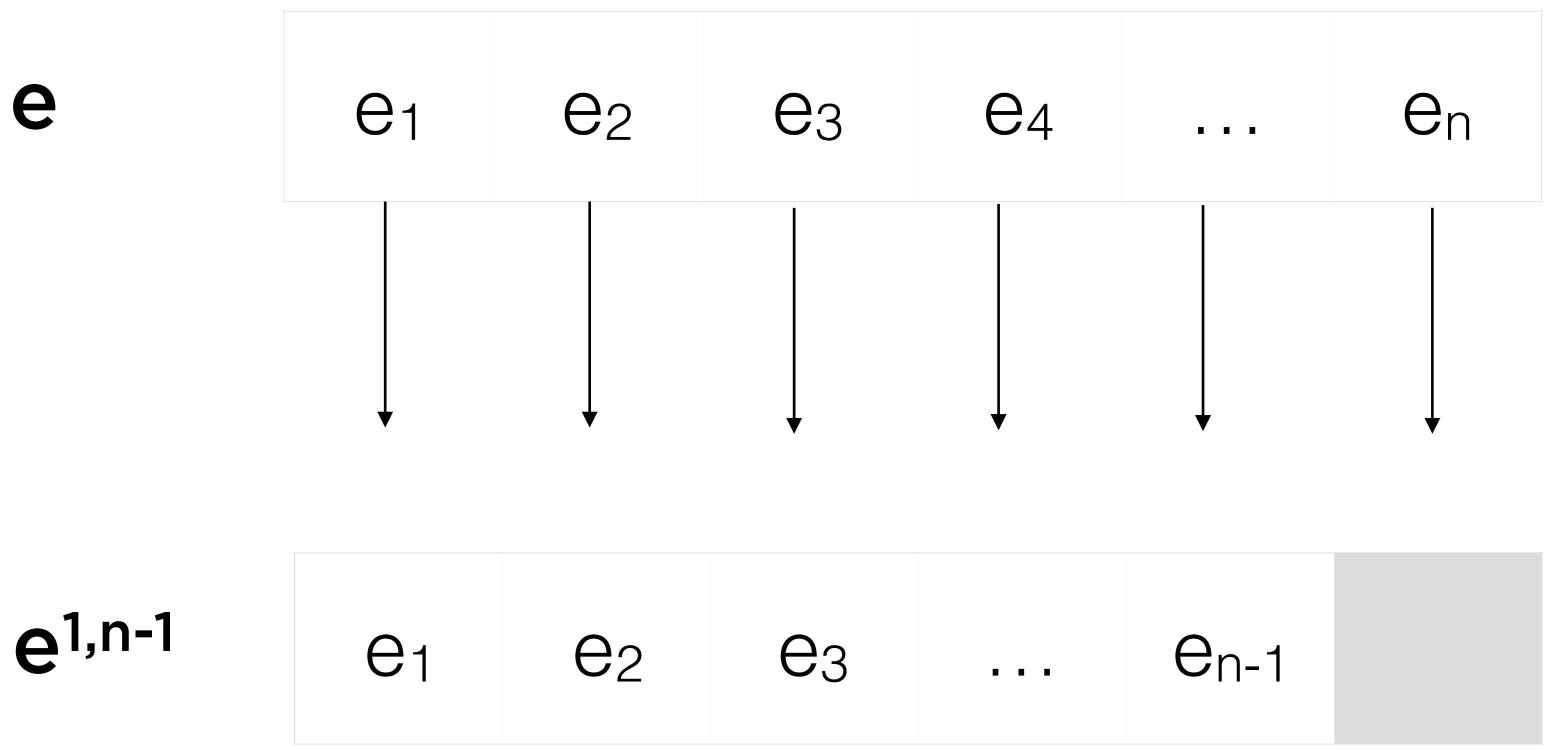
Correlation of any series with itself is always +1, so measure lag-1 autocorrelation instead

Lag-1 Autocorrelation of Residuals



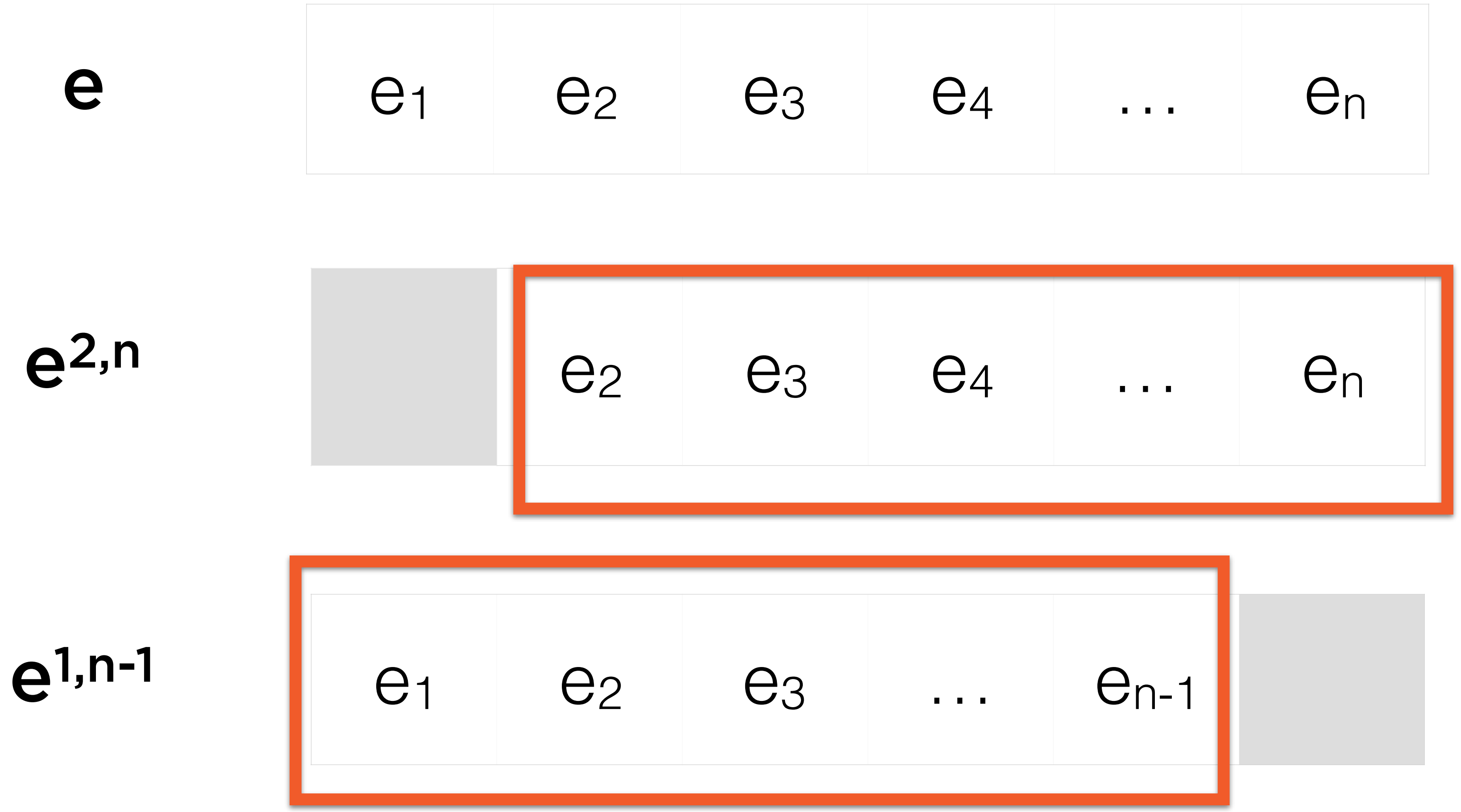
$e^{2,n}$ = Exclude value 1, include values 2 to n

Lag-1 Autocorrelation of Residuals



$e^{1,n-1}$ = Include values 1 to n-1, exclude value n

Lag-1 Autocorrelation of Residuals



$$\mathrm{correl}(e^{1,n-1}, e^{2,n}) = 0$$

Correlation of these two vectors should be zero

Risks in Simple Regression

No cause-effect relationship

Regression on completely unrelated data series

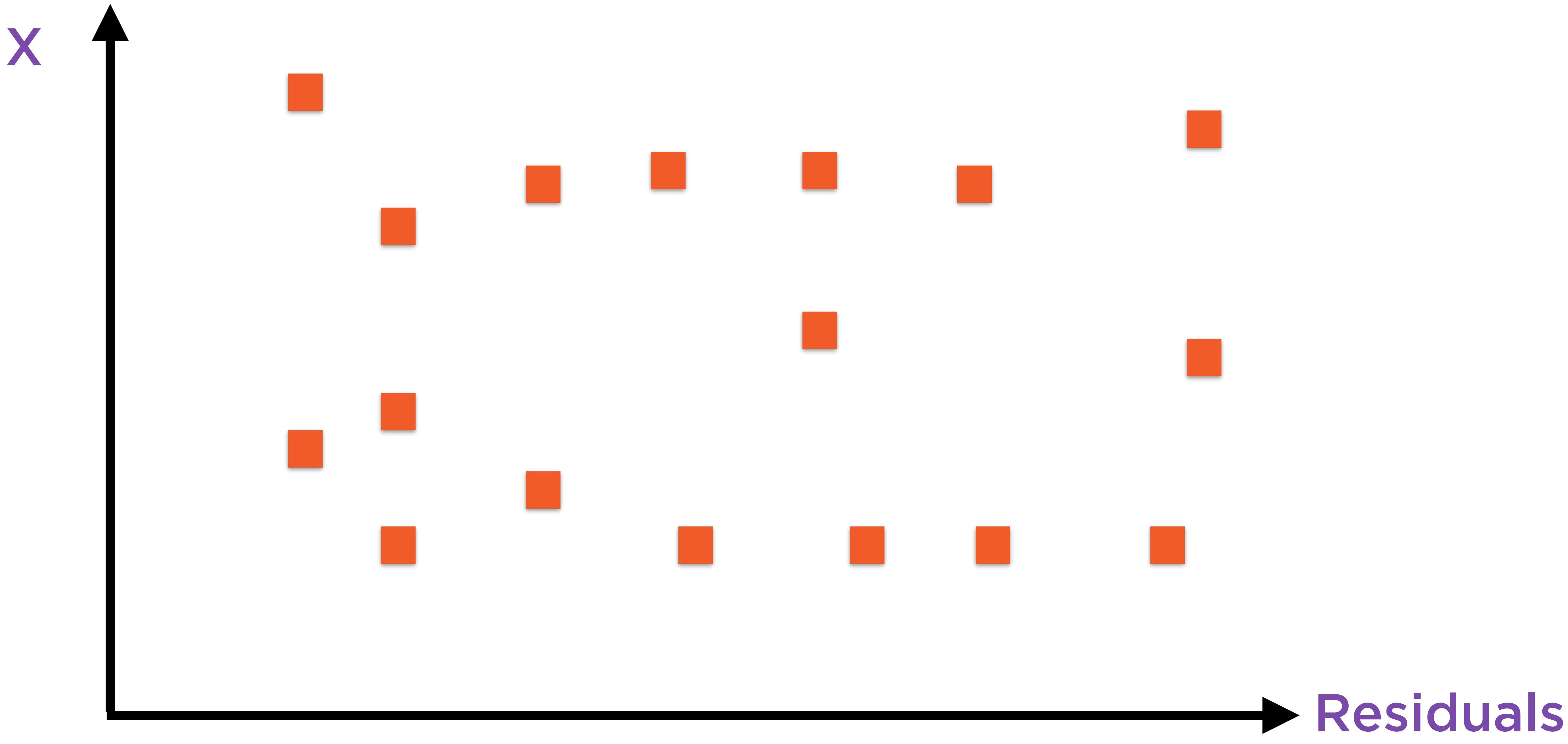
Mis-specified relationship

Non-linear (exponential or polynomial) fit

Incomplete relationship

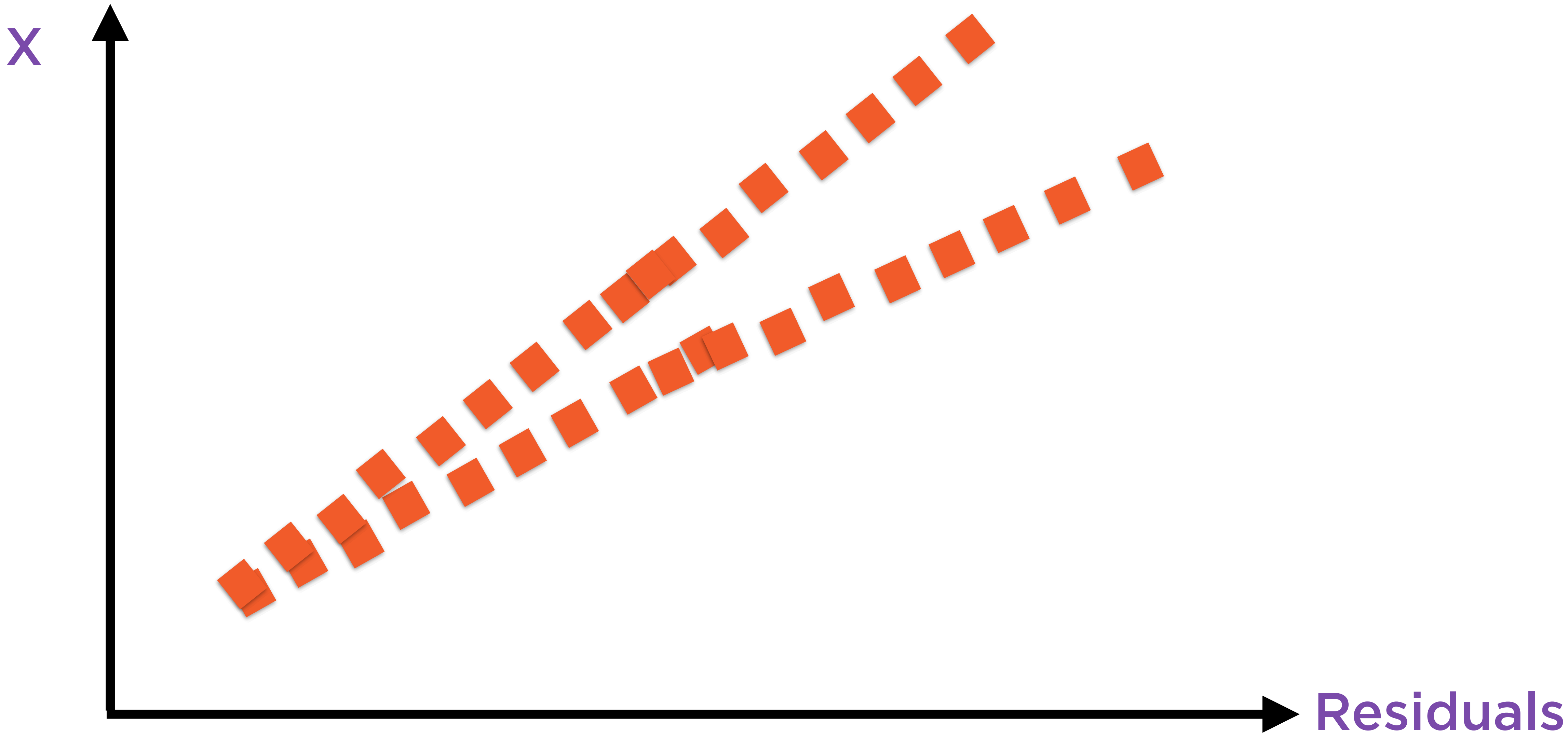
Multiple causes exist, we have captured just one

“Good” Residuals

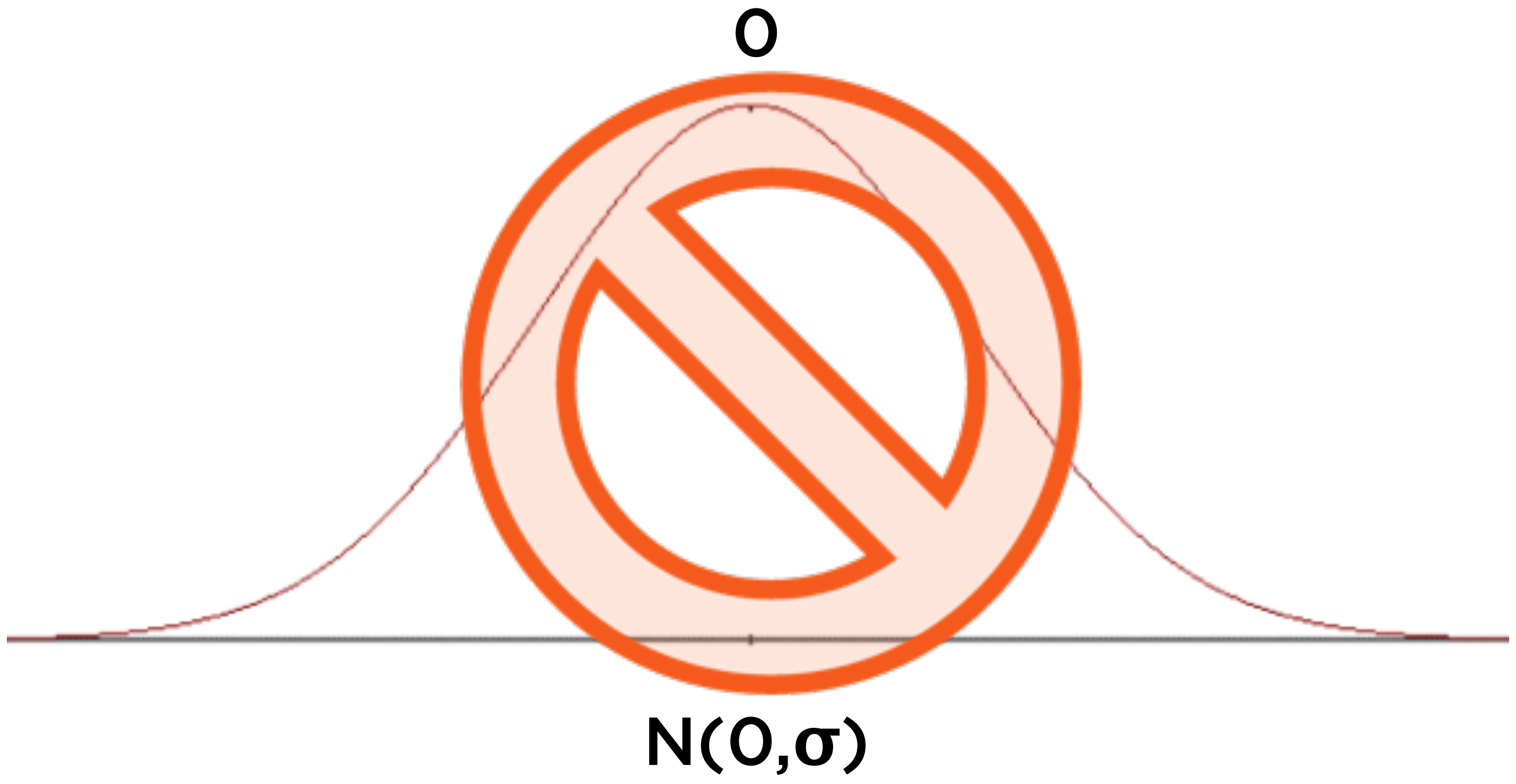


Residuals are independent of X

“Bad” Residuals



Clear relationship between residuals and X



$e \nsim N(0,\sigma^2)$

"Bad" Residuals ~ Heteroskedasiticity

Possible causes vary, but missing x-variables is an important one

Residuals drawn from a distribution with non-constant variance are said to be **heteroskedastic**

Diagnosing Risks in Simple Regression

No cause-effect relationship

low R^2, plot of X ~ Y has no pattern

Mis-specified relationship

high R^2, residuals are not independent of each other

Incomplete relationship

low R^2, residuals are not independent of x

Mitigating Risks in Simple Regression

No cause-effect relationship

Wrong choice of X and Y - back to drawing board

Mis-specified relationship

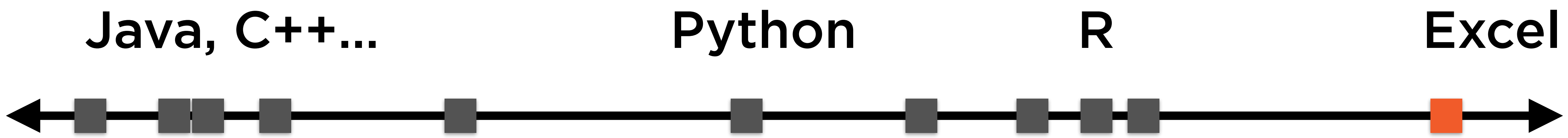
Transform X and Y - convert to logs or returns

Incomplete relationship

Add X variables (move to multiple regression)

Excel for Simple Regression

Ease of Prototyping



Excel is the fastest prototyping tool out there

Robustness and Re-use

No free lunches

Applying Simple Regression

Sanity Check

Scatter of X and Y

Eyeball for linear fit

Perform Regression

Find slope, intercept, R^2

Excel functions available

Residuals In Isolation

Check independence with self

Autocorrelation not present

Residuals and X

Scatter of X and residuals

No pattern, no linear fit

Explain Variation

Interpret slope, intercept, R^2

Safe to use regression results

Forecast

Predict Y for new X

Excel function available

Demo

Download data from Yahoo Finance

Regression plots in one step

Regression coefficients

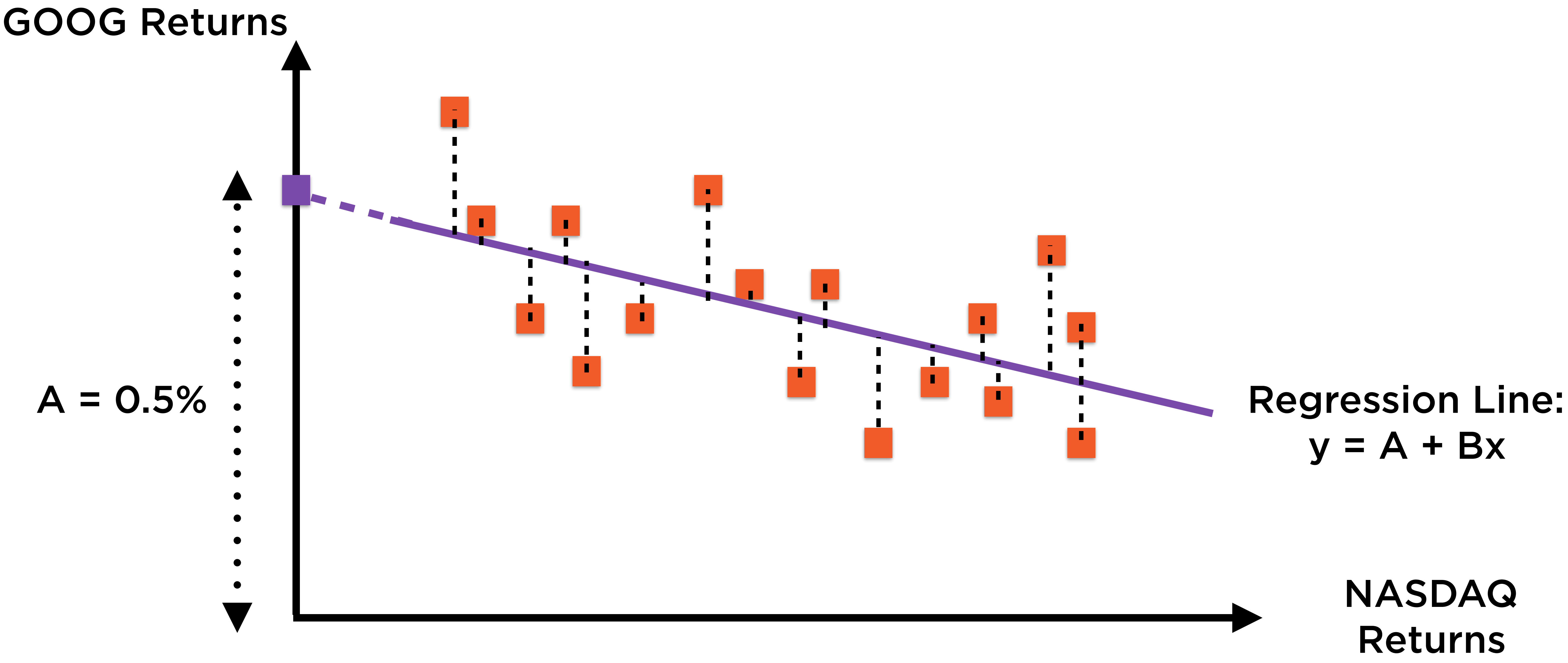
- Slope
- Intercept

Sanity checking residuals

Forecasting

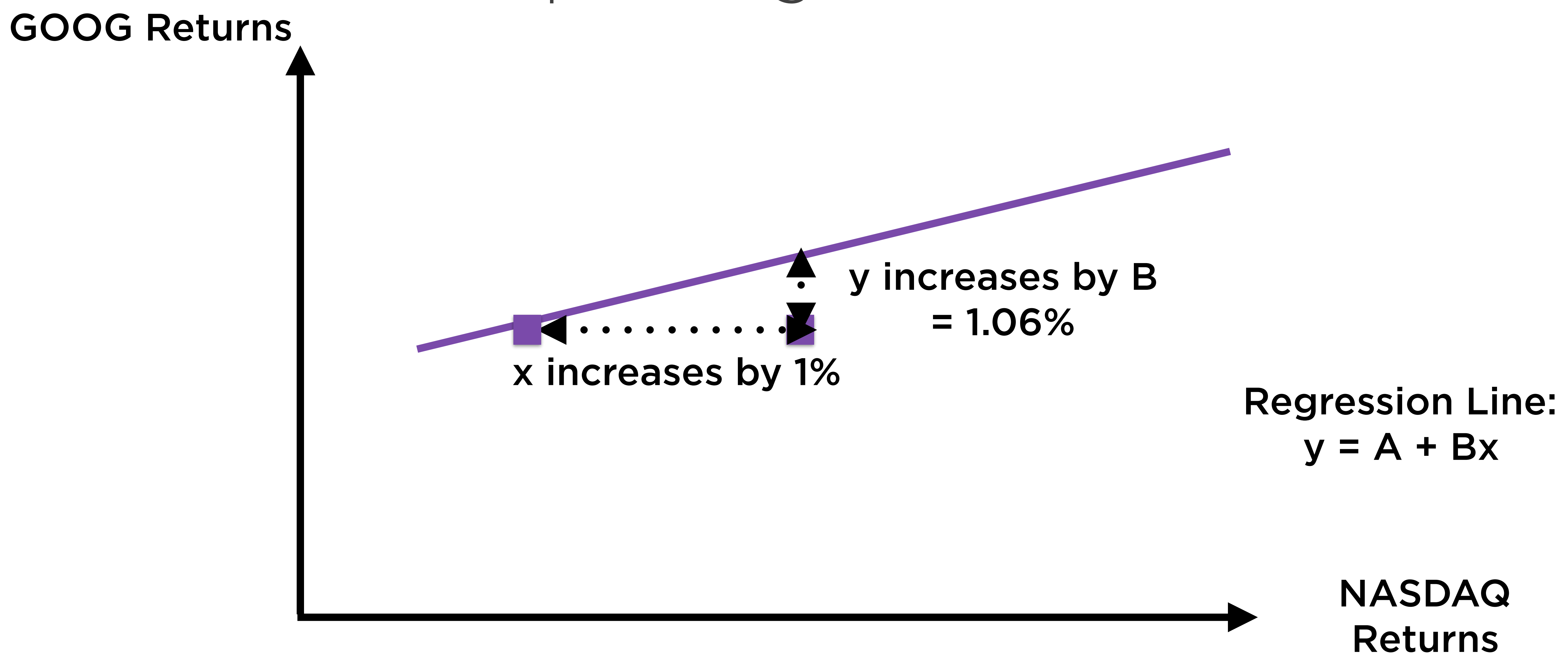
Misuse of regression

Explaining Variation



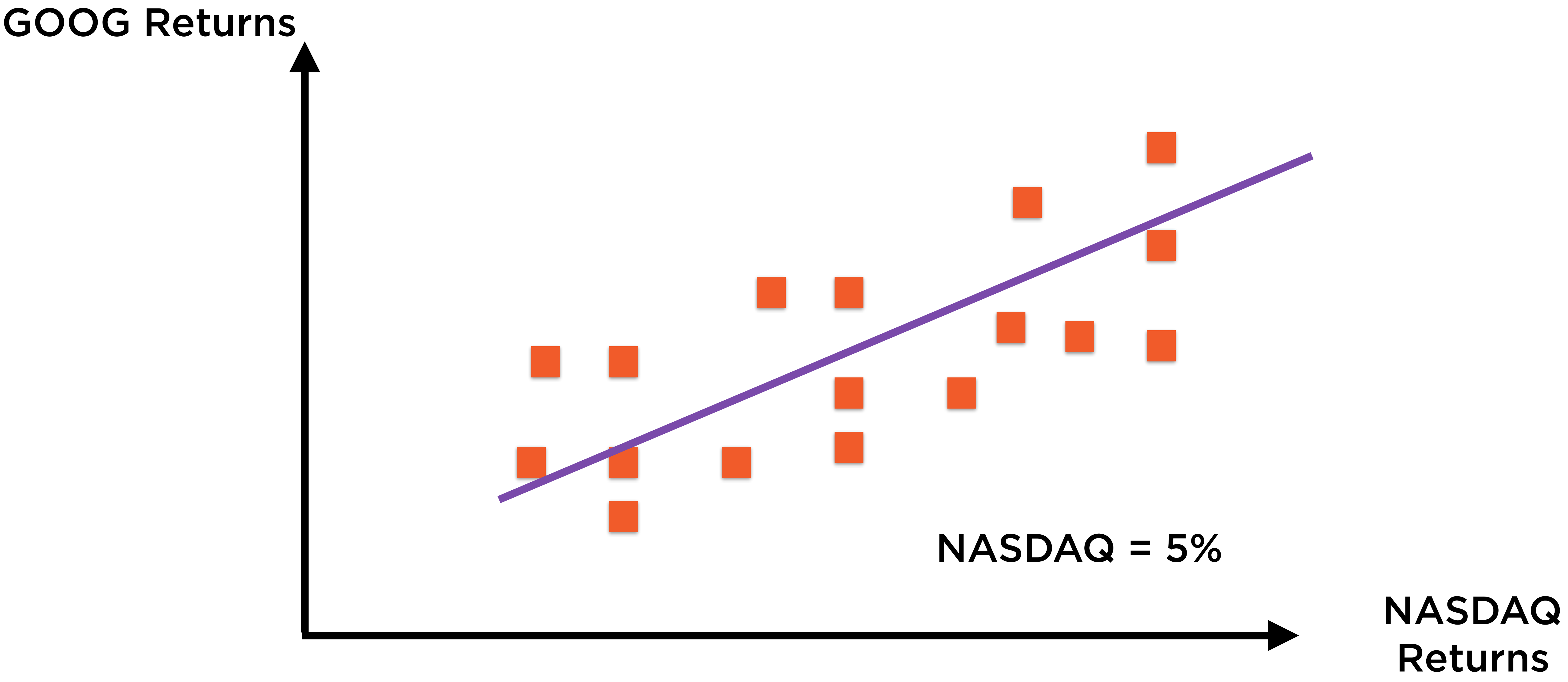
GOOG has an in-sample alpha of 0.5% per month over the NASDAQ

Explaining Variation



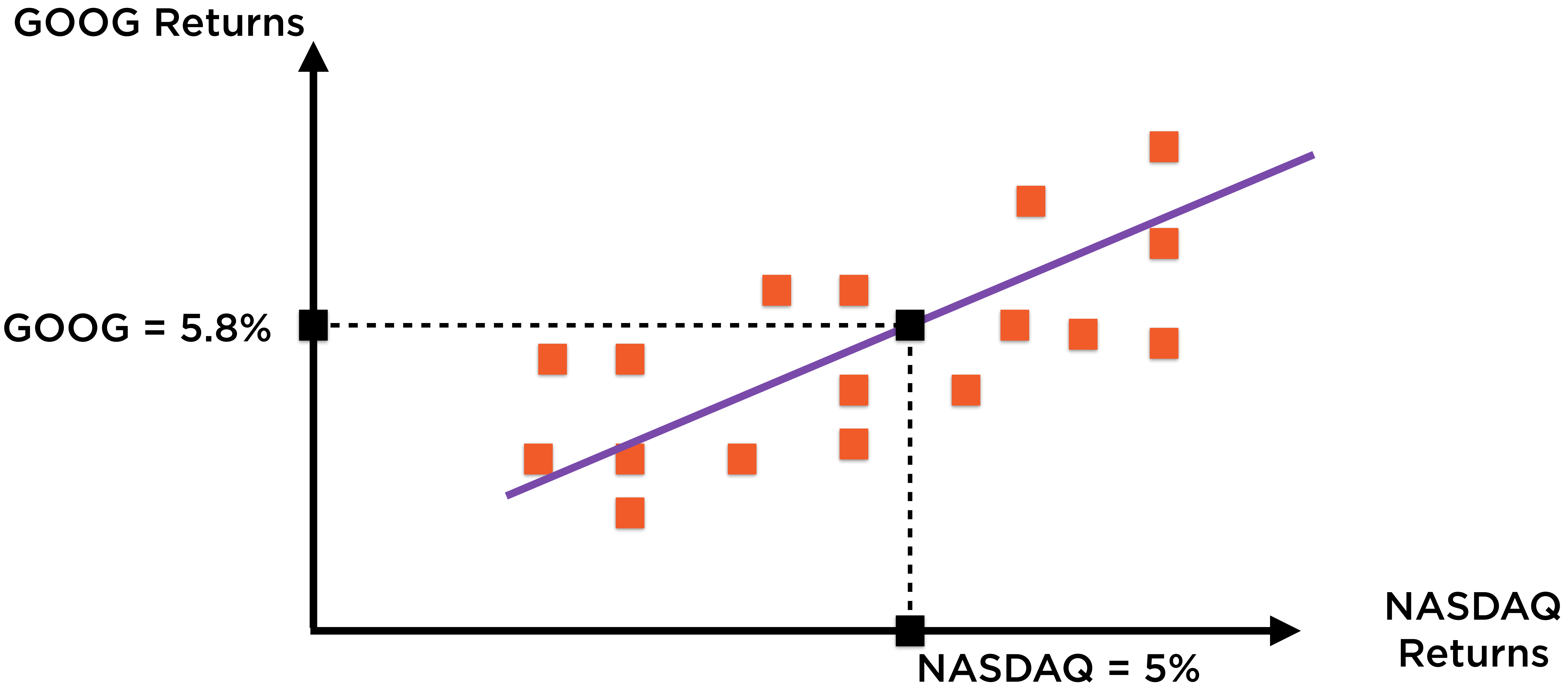
GOOG has an in-sample beta of 1.06 with the NASDAQ

Prediction Using Regression



Find the regression line - the line with the "best fit"

Prediction Using Regression



Find the regression line - the line with the "best fit"

Poorly Specified Regression Models

Overview

Build regression models in Excel

Understand and test the regression assumptions

Use simple regression models in Excel

- to explain variance
- to make forecasts

Avoid some common regression pitfalls

Summary

Built regression models in Excel

Avoided some common regression pitfalls

Use simple regression models in Excel

- to explain variance
- to make forecasts