Understanding Multiple Regression Models

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

Overview

Extend regression analysis to multiple explanatory variables

Interpret the results of a multiple regression

Mitigate the risks that accompany multiple regression

Apply multiple regression to include categorical variables

Introducing Multiple Regression

“A butterfly flapping its wings in Brazil can cause a tornado in Texas”

Butterfly Effect

The butterfly effect is the concept that small causes can have large effects. In chaos theory, the butterfly effect is the sensitive dependence on initial conditions in which a small change in one state of a deterministic nonlinear system can result in large differences in a later state.
Wikipedia

Simple Regression

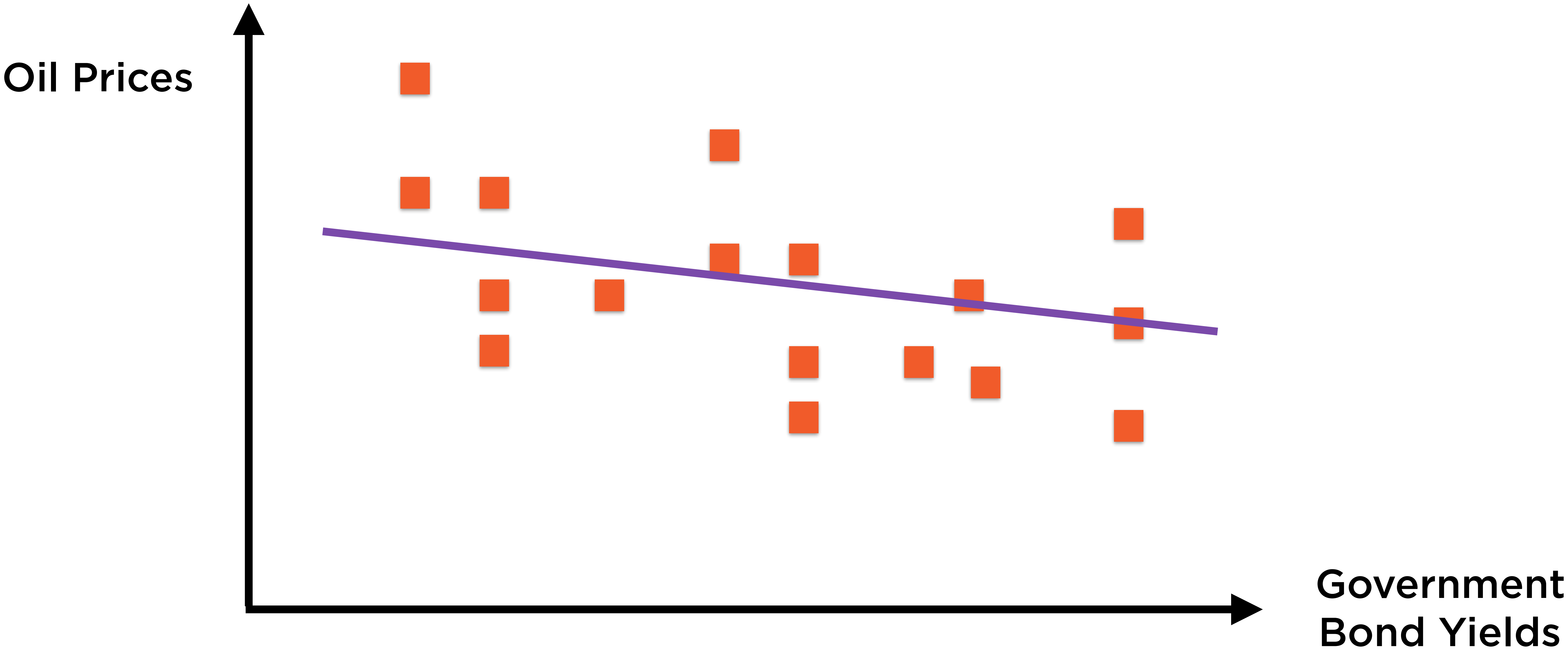
Cause

Independent variable

Effect

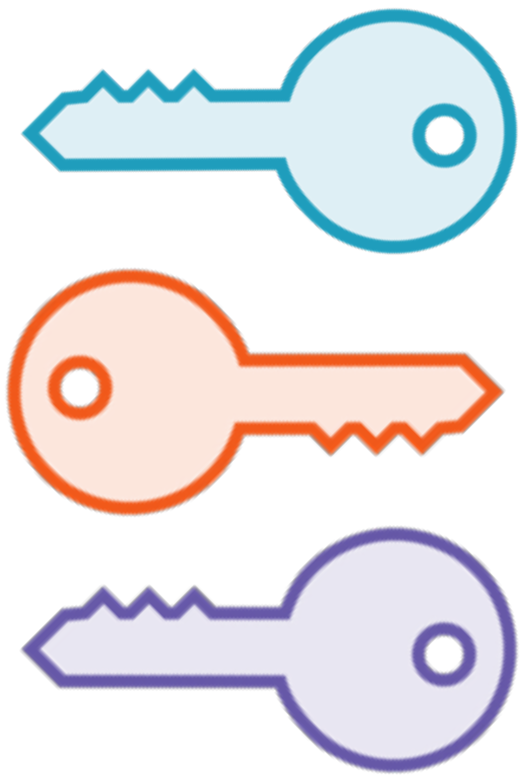
Dependent variable

Simple Regression



One cause, one effect

Multiple Regression



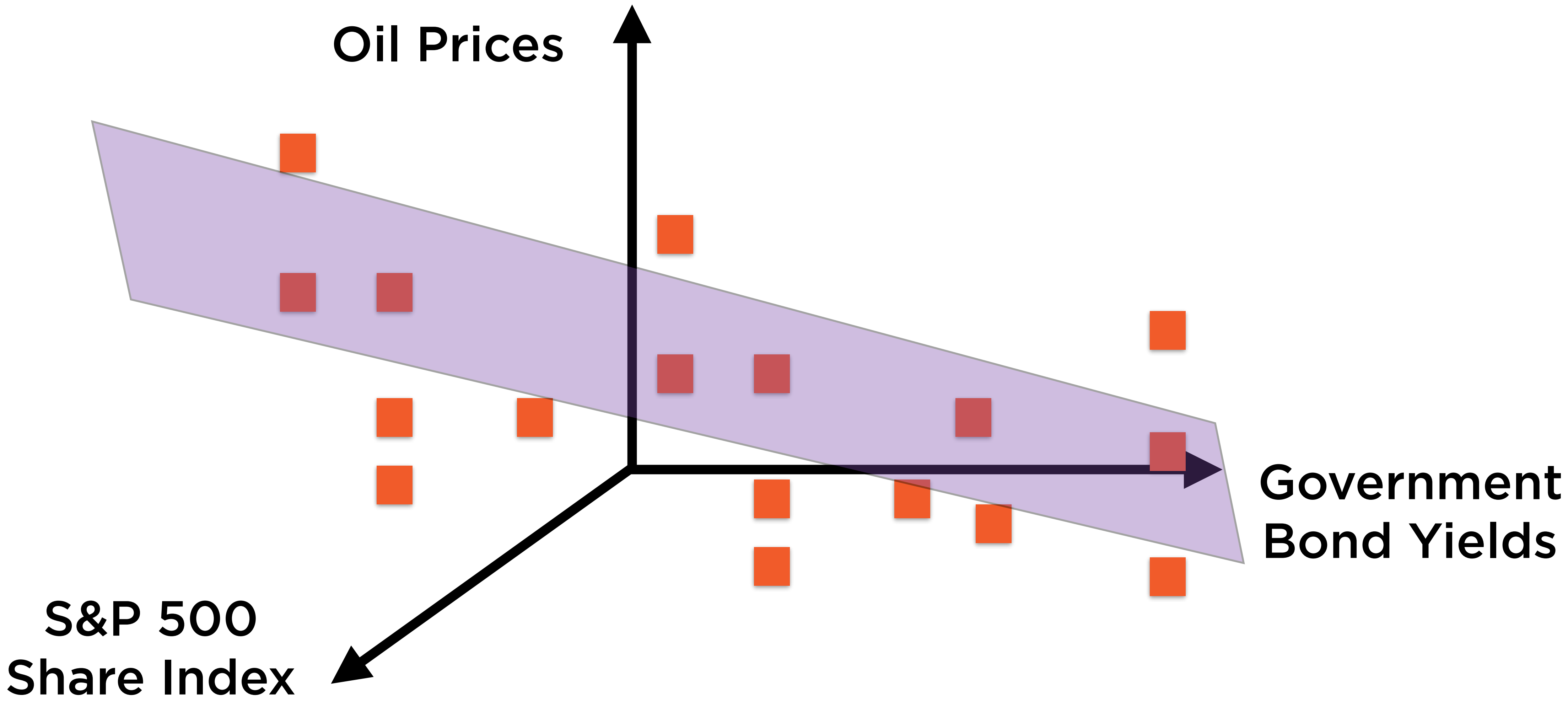
Causes

Independent variables

Effect

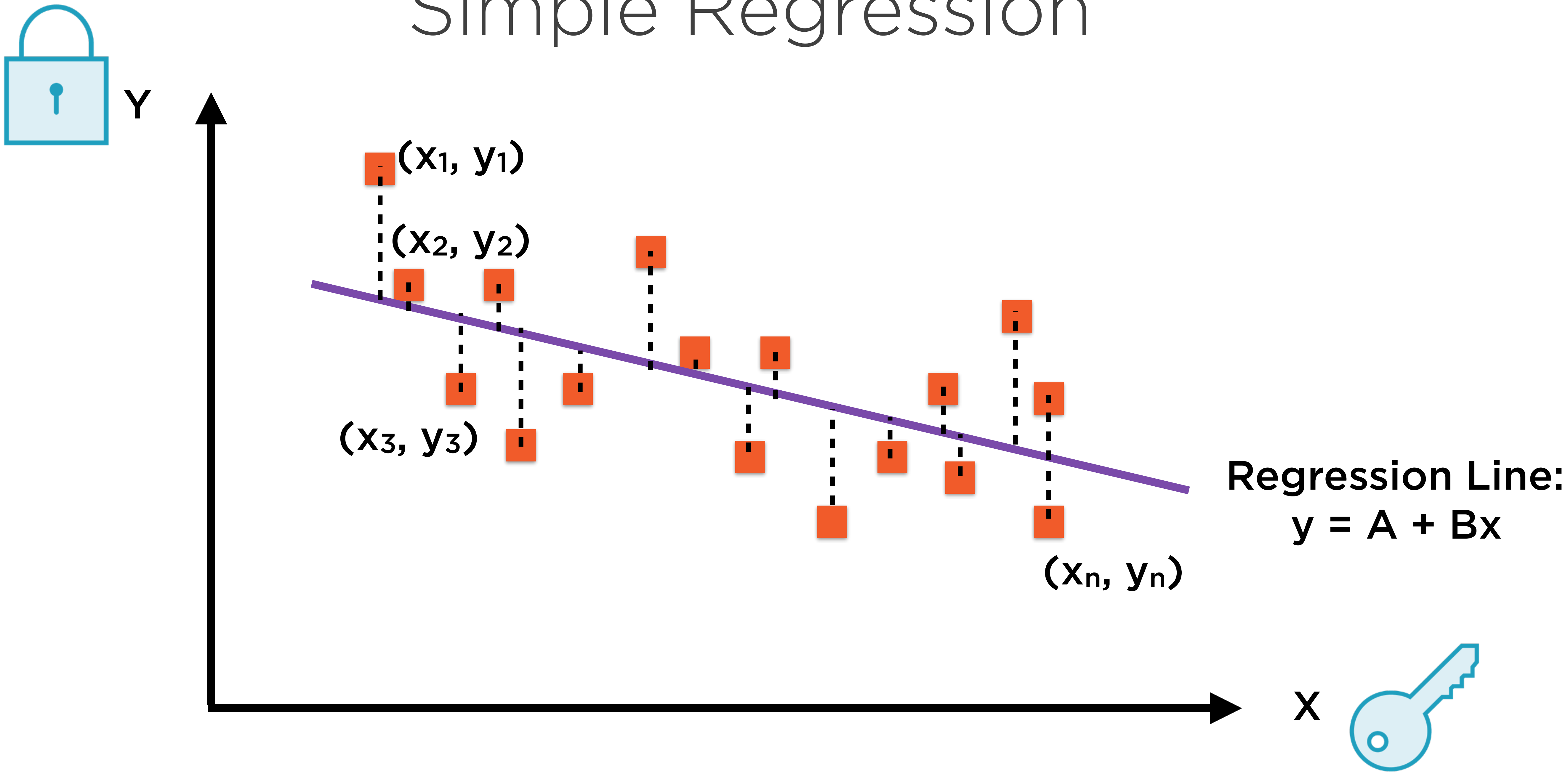
Dependent variable

Multiple Regression



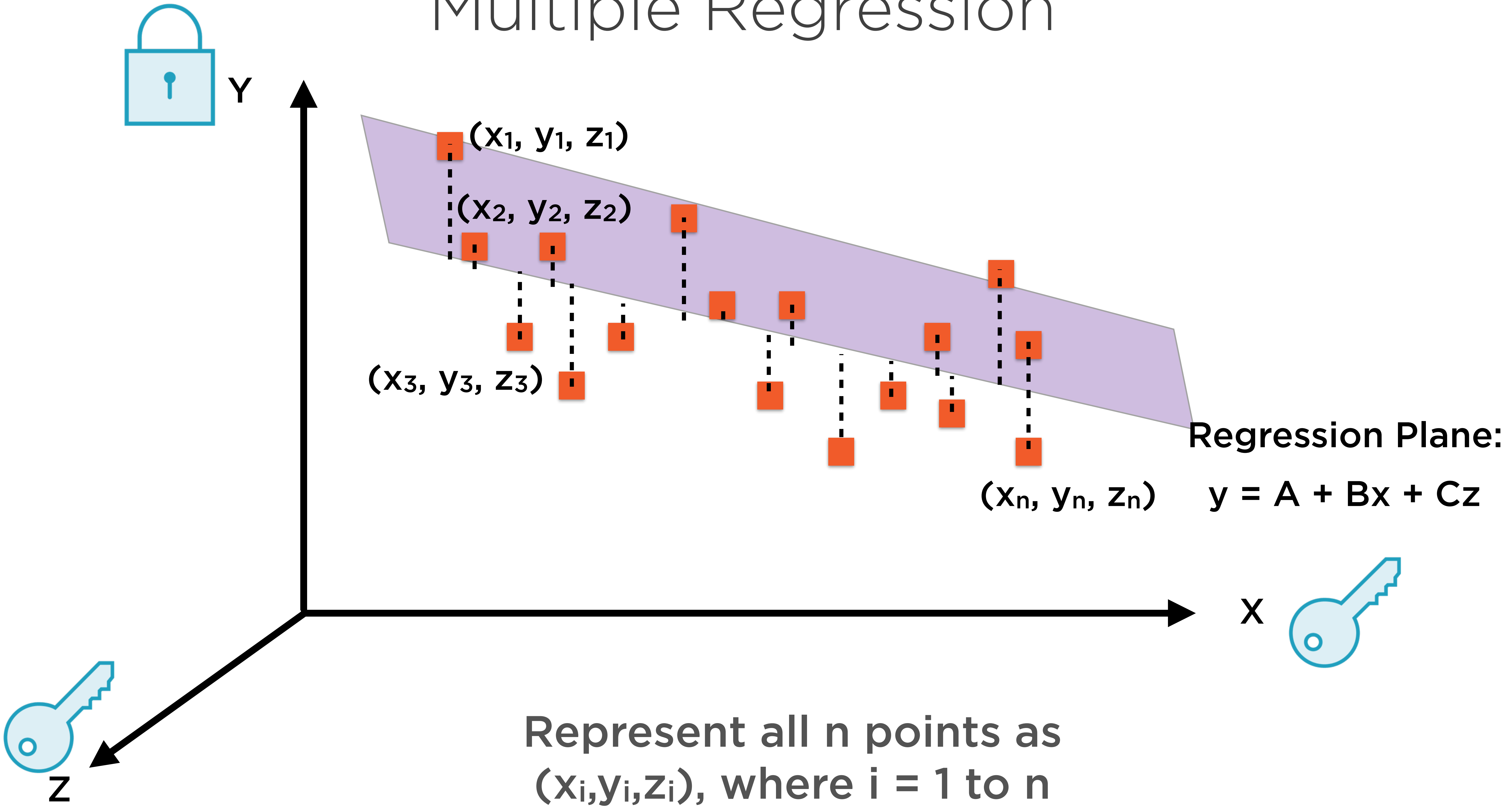
Many causes, one effect

Simple Regression



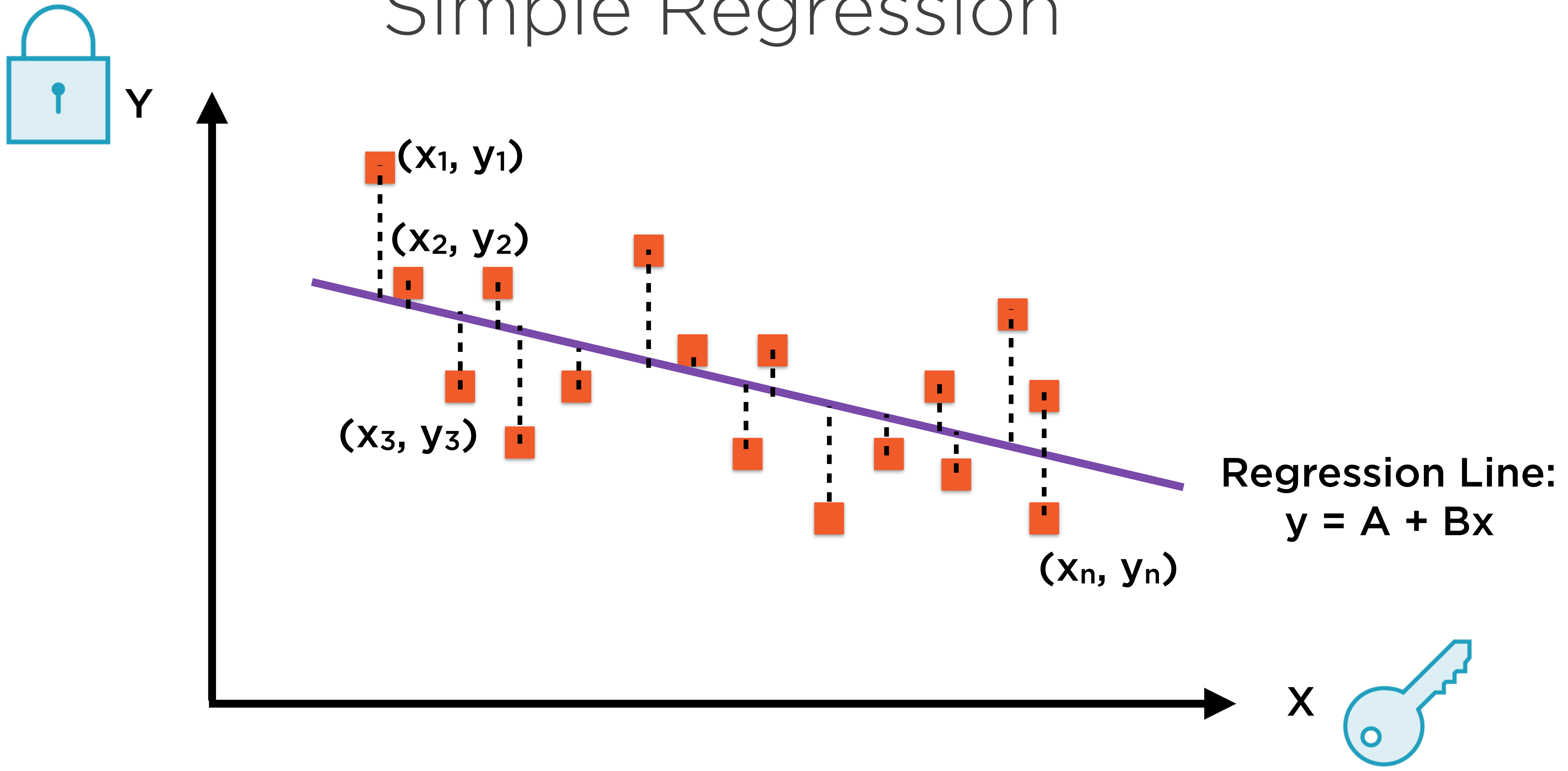
Represent all n points as
(x_i, y_i), where i = 1 to n

Multiple Regression



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

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:

$$y = A + Bx$$

$$y_1 = A + Bx_1 + e_1$$

$$y_2 = A + Bx_2 + e_2$$

$$y_3 = A + Bx_3 + e_3$$

$$\ldots \qquad \ldots$$

$$y_n = A + Bx_n + e_n$$

Simple Regression

Regression Equation:

$$y = A + Bx$$

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

Simple Regression

Regression Equation:

$$EXXON_t = A + B\,DOW_t$$

$$\begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \cdots \\ E_n \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} + B \begin{bmatrix} D_1 \\ D_2 \\ D_3 \\ \cdots \\ D_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \cdots \\ e_n \end{bmatrix}$$

E_i = % return on Exxon stock on day i

D_i = % return of Dow Jones index on day i

Simple Regression

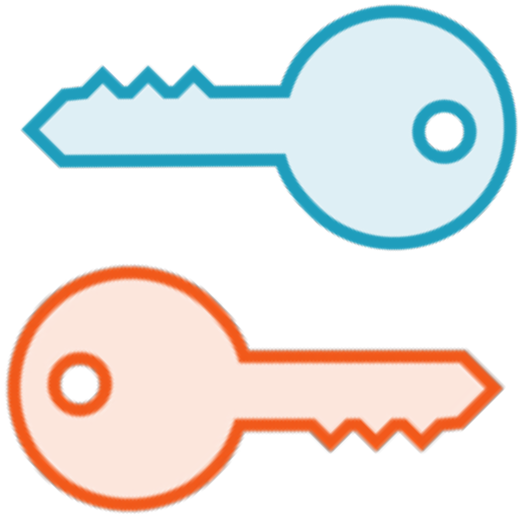
Cause

Changes in Dow Jones equity index

Effect

Changes in price of Exxon Stock

Multiple Regression



Causes

Dow Jones index,
price of oil

Effect

Exxon stock

Multiple Regression

Regression Equation:

$$EXXON_t = A + B\ DOW_t + C\ OIL_t$$

$$\begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \cdots \\ E_n \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} + B \begin{bmatrix} D_1 \\ D_2 \\ D_3 \\ \cdots \\ D_n \end{bmatrix} + C \begin{bmatrix} O_1 \\ O_2 \\ O_3 \\ \cdots \\ O_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \cdots \\ e_n \end{bmatrix}$$

E_i = % return on Exxon stock on day i

D_i = % return of Dow Jones index on day i

O_i = % change in price of oil on day i

Multiple Regression

Regression Equation:

$$y = A + Bx + Cz$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} + B \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_n \end{bmatrix} + C \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \cdots \\ z_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \cdots \\ e_n \end{bmatrix}$$

Multiple Regression

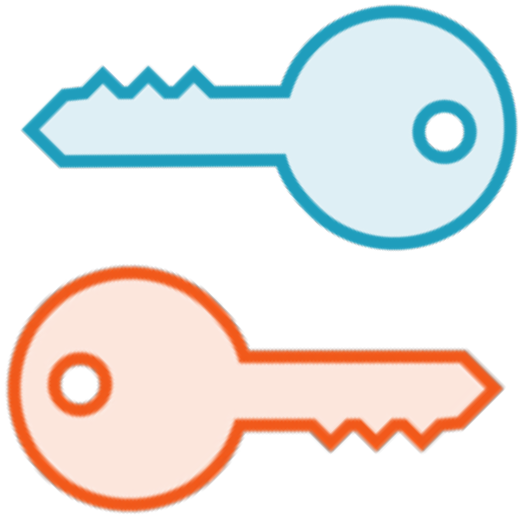
Regression Equation:

$$y = A + Bx + Cz$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ 1 & x_3 & z_3 \\ \ldots & \ldots & \ldots \\ 1 & x_n & z_n \end{bmatrix} * \begin{bmatrix} A \\ B \\ C \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \ldots \\ e_n \end{bmatrix}$$

n Rows, 1 Column	n Rows, 3 Columns	3 Rows, 1 Column	n Rows, 1 Column

Multiple Regression



2 Causes

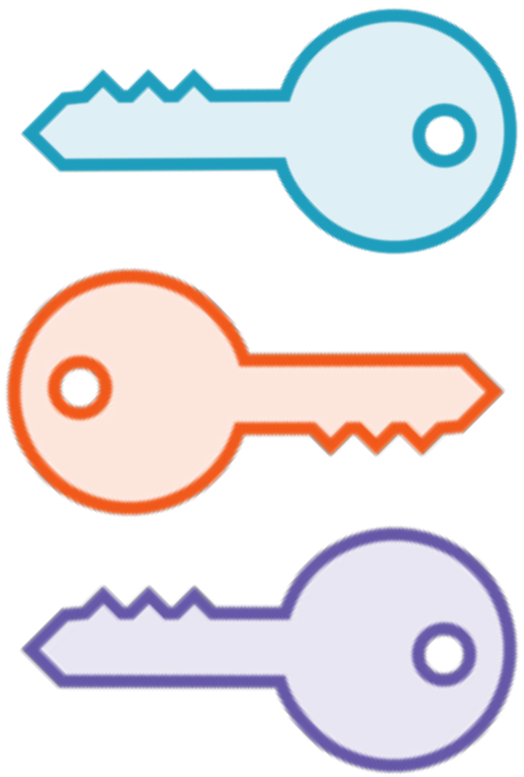
Dow Jones index,

price of oil

1 Effect

Exxon stock

Multiple Regression



k Causes

Dow Jones index,

price of oil, bond yields...

1 Effect

Exxon stock

Multiple Regression

Regression Equation:

$$y = C_1 + C_2x_1 + \dots + C_{k+1}x_k$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix} = C_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} + C_2 \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ \dots \\ x_{n1} \end{bmatrix} + \dots C_{k+1} \begin{bmatrix} x_{1k} \\ x_{2k} \\ x_{3k} \\ \dots \\ x_{nk} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_n \end{bmatrix}$$

Multiple Regression

Regression Equation:

$$y = C_1 + C_2x_1 + \ldots + C_{k+1}x_k$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & & x_{1k} \\ 1 & x_{21} & & x_{2k} \\ 1 & x_{31} & \ldots & x_{3k} \\ \ldots & \ldots & & \ldots \\ 1 & x_{n1} & & x_{nk} \end{bmatrix} * \begin{bmatrix} C_1 \\ C_2 \\ \ldots \\ C_{k+1} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \ldots \\ e_n \end{bmatrix}$$

n Rows,
1 Column

n Rows,
k Columns

k Rows,
1 Column

Multiple Regression

Regression Equation:

$$y = C_1 + C_2x_1 + \ldots + C_{k+1}x_k$$

Multiple regression involves finding k+1 coefficients, k for the explanatory variables, and 1 for the intercept

Estimation Methods in Multiple Regression

Method of moments

Method of least squares

Maximum likelihood estimation

The method of least squares works for multiple regression too

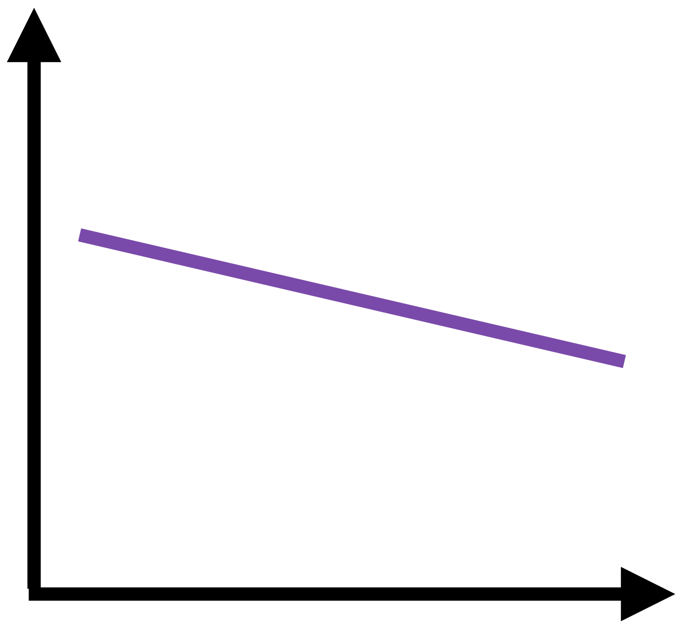
Regression Equation:

$$y = C_1 + C_2x_1 + \ldots + C_{k+1}x_k$$

The "best fit" line is the one where the sum of the squares of the lengths of the errors is minimised

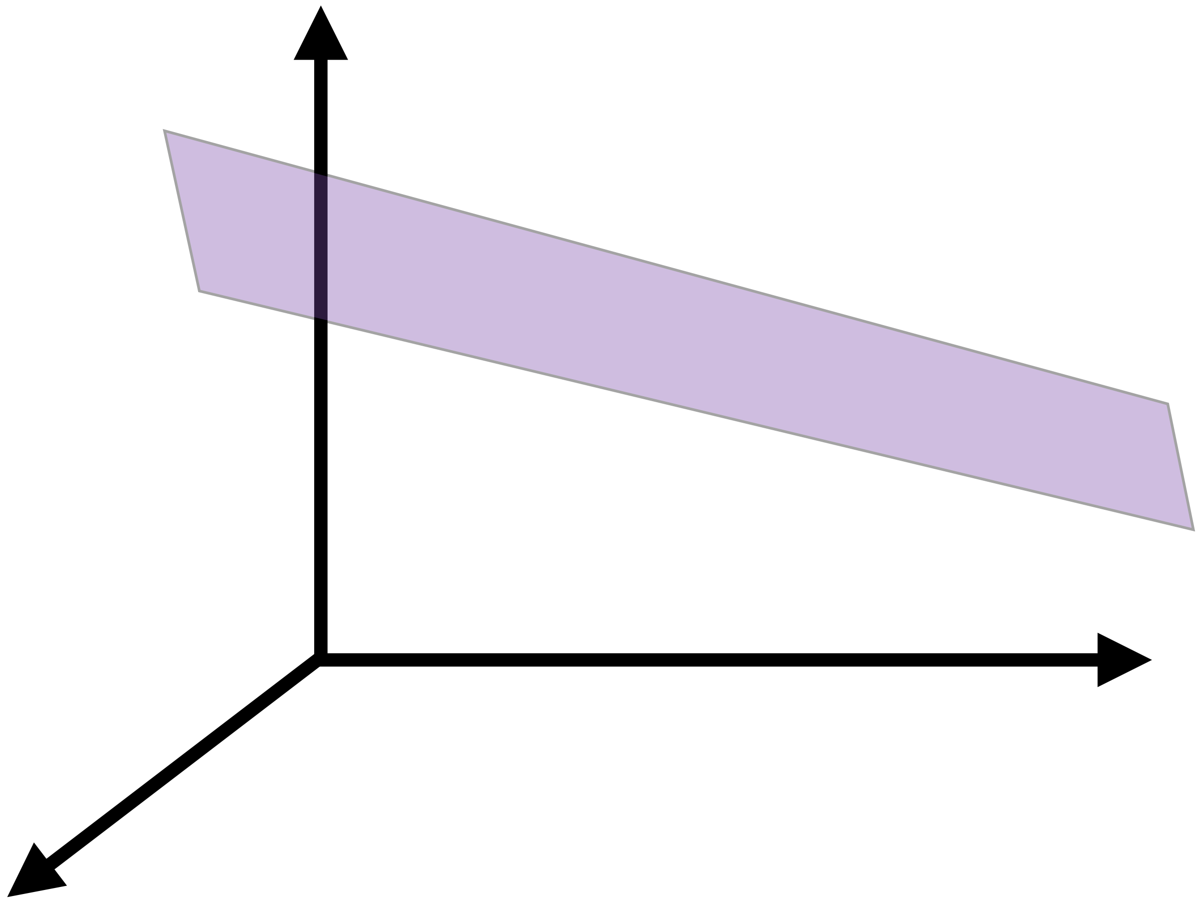
Risks in Multiple Regression

Simple and Multiple Regression



Simple Regression

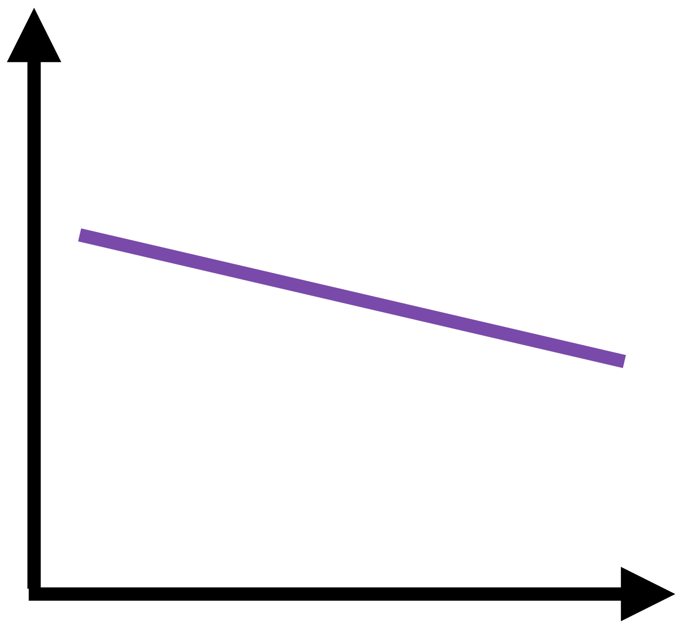
Data in 2 dimensions



Multiple Regression

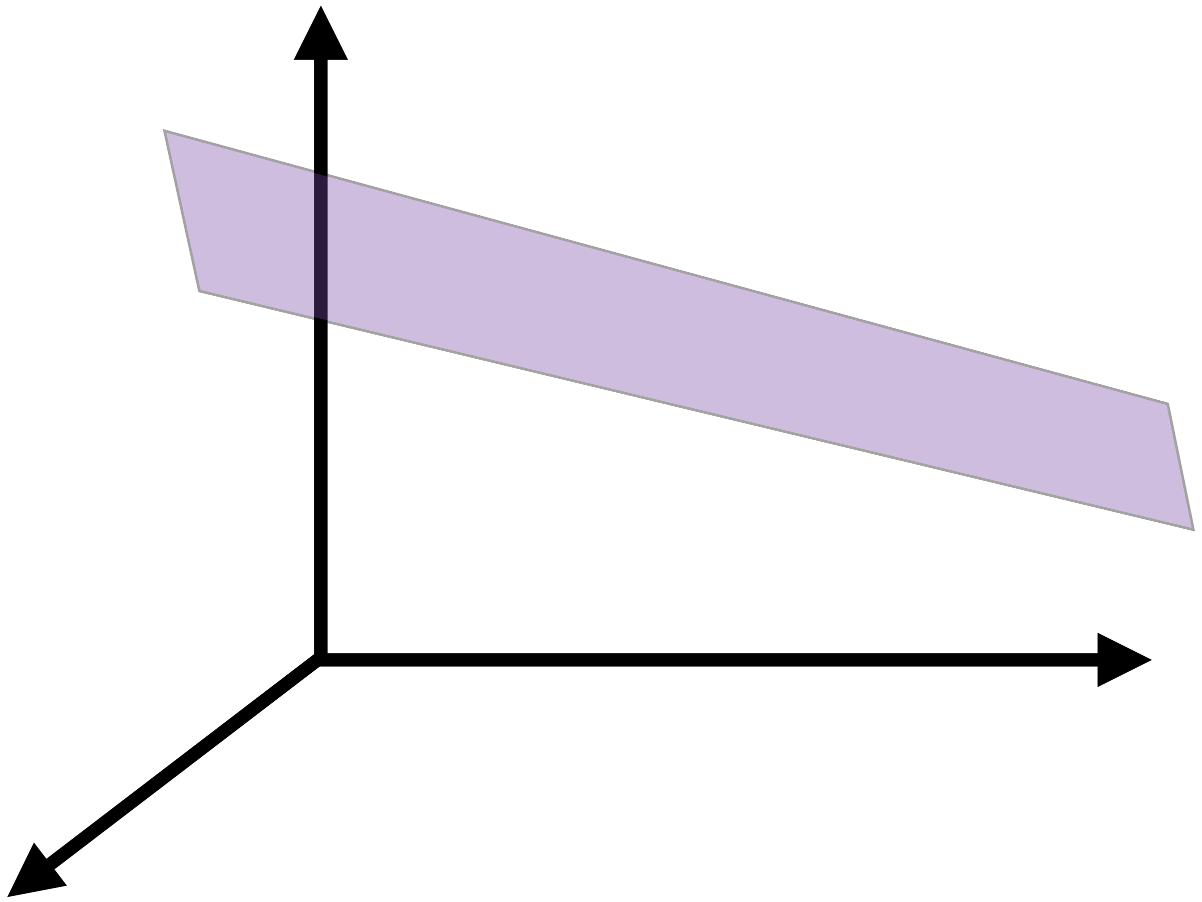
Data in > 2 dimensions

Simple and Multiple Regression



Simple Regression

Risks exist, but can usually be mitigated analysing R^2 and residuals



Multiple Regression

Risks are more complicated, require interpreting regression statistics

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

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)

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

"If everyone is thinking alike, then somebody isn't thinking."

General Patton

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

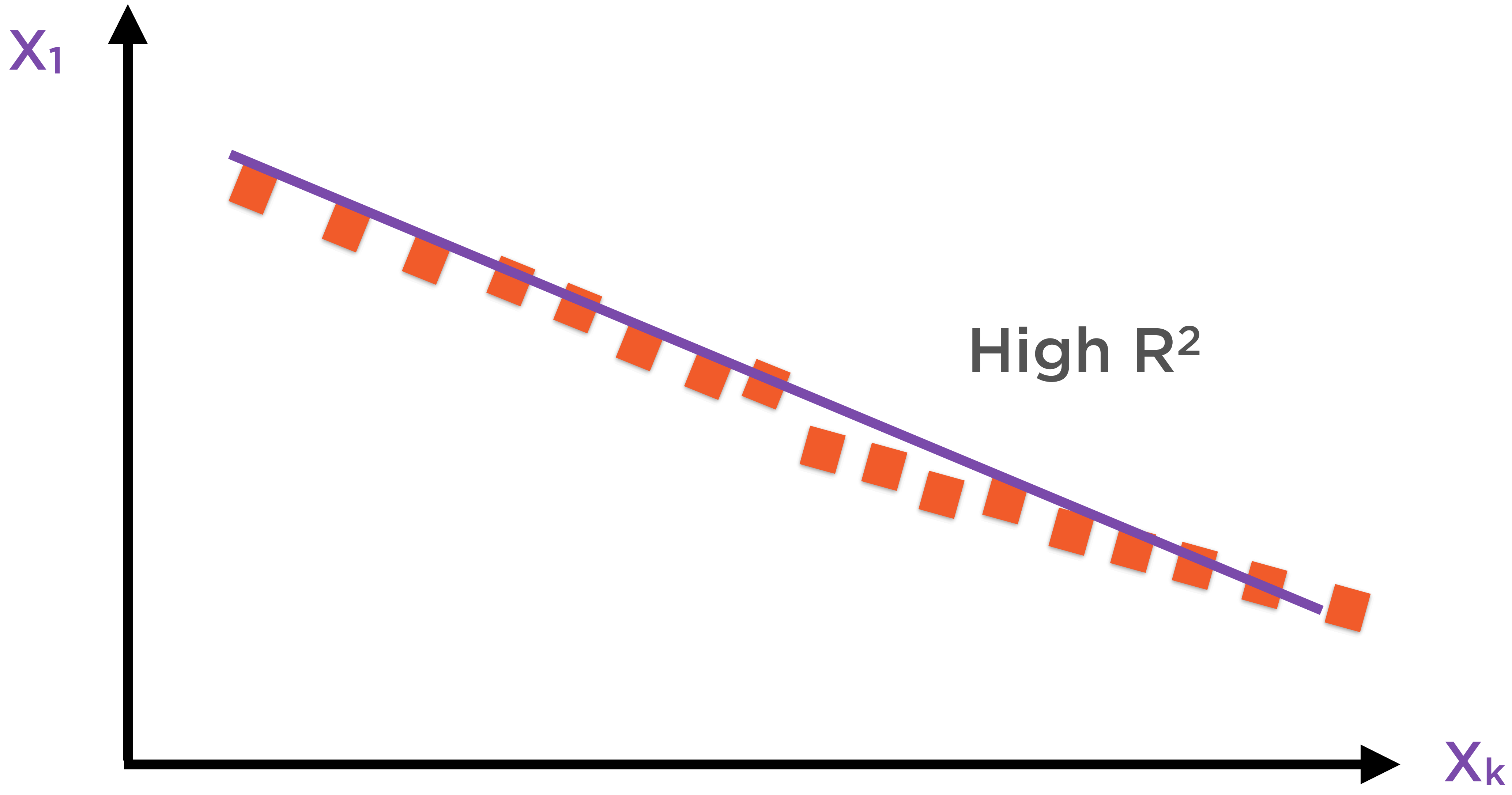
Multiple Regression

Regression Equation:

$$y = C_1 + C_2 x_1 + \ldots + C_k x_{k-1}$$

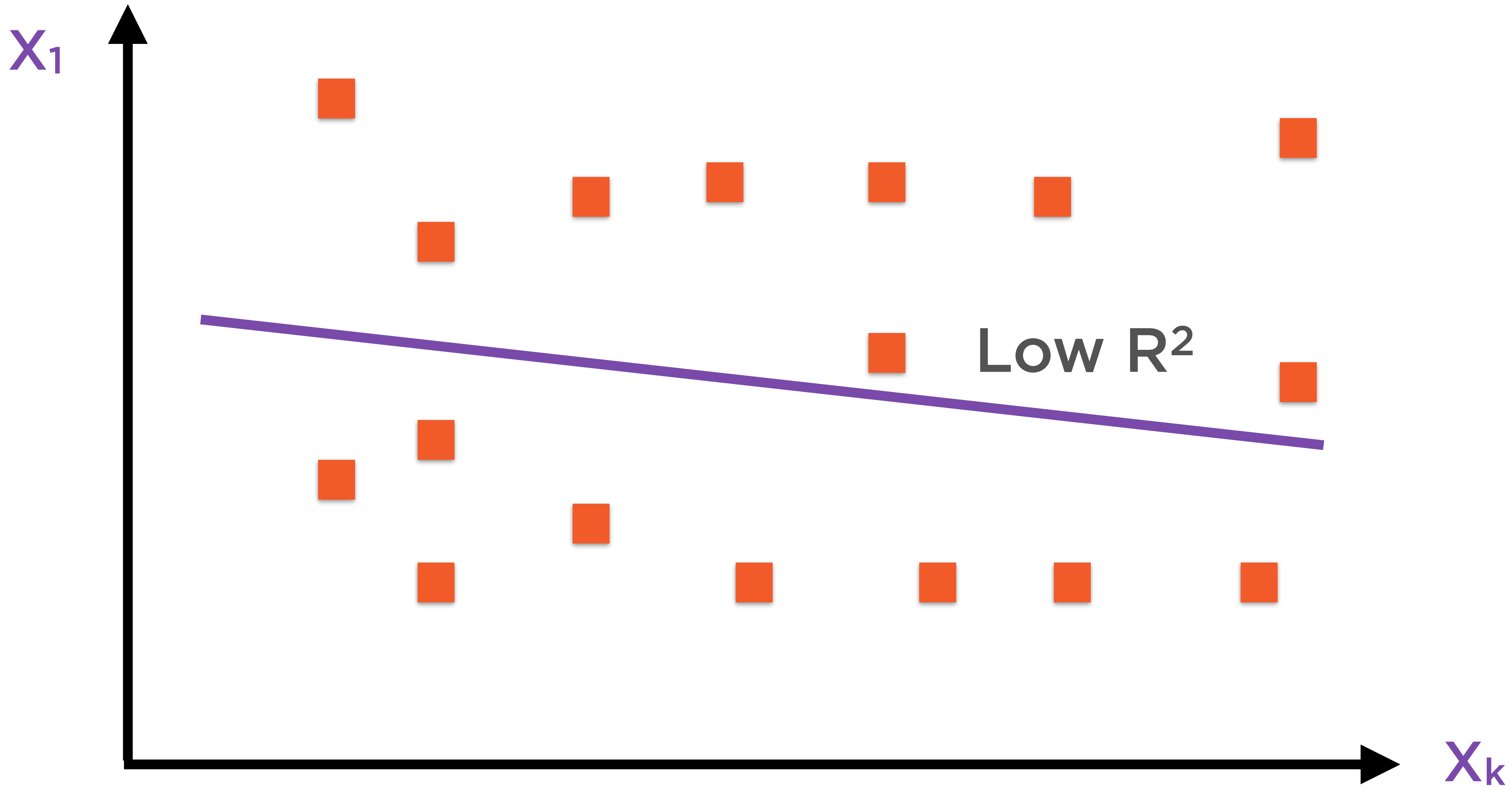
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \ldots \\ y_n \end{bmatrix} = \begin{array}{c} \\ \left[\begin{array}{ccccc} 1 & \boxed{x_{11}} & & \boxed{x_{1k-1}} \\ 1 & x_{21} & & x_{2k-1} \\ 1 & x_{31} & \ldots & x_{3k-1} \\ \ldots & \ldots & & \ldots \\ 1 & x_{n1} & & x_{nk-1} \end{array}\right] \\ \quad\quad X_1 \quad\quad\quad X_k \end{array} * \begin{bmatrix} C_1 \\ C_2 \\ \ldots \\ C_k \end{bmatrix}$$

Bad News: Multicollinearity Detected



Highly correlated explanatory variables

Good News: No Multicollinearity Detected



Uncorrelated explanatory variables

Multiple Regression

Regression Equation:

$$EXXON_t = A + B\ DOW_t + C\ OIL_t$$

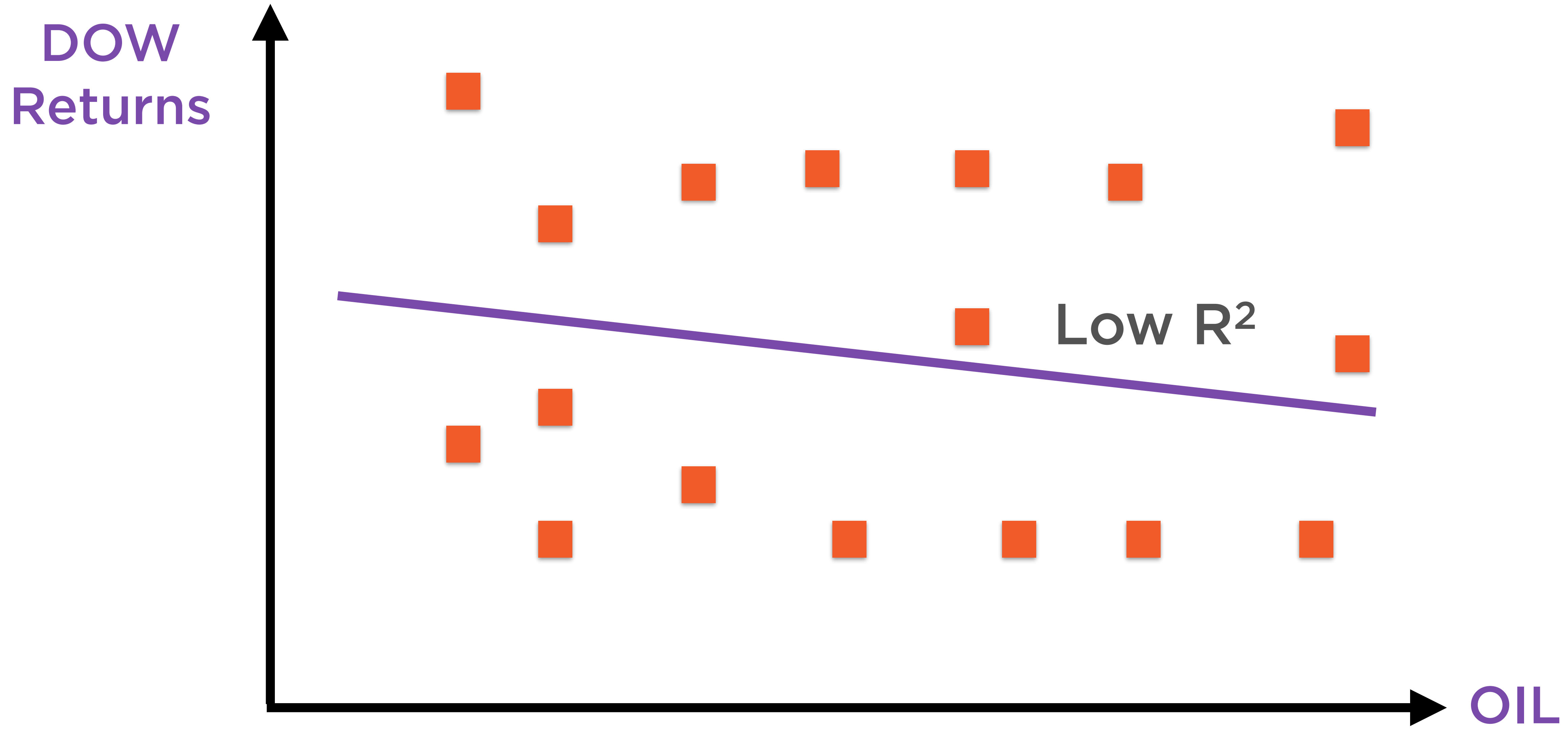
$$\begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \dots \\ E_n \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} + B \begin{bmatrix} D_1 \\ D_2 \\ D_3 \\ \dots \\ D_n \end{bmatrix} + C \begin{bmatrix} O_1 \\ O_2 \\ O_3 \\ \dots \\ O_n \end{bmatrix}$$

E_i = % return on Exxon stock on day i

D_i = % return of Dow Jones index on day i

O_i = % change in price of oil on day i

Good News: No Multicollinearity Detected



Uncorrelated explanatory variables

Multiple Regression

Regression Equation:

$$EXXON_t = A + B\ DOW_t + C\ NASDAQ_t$$

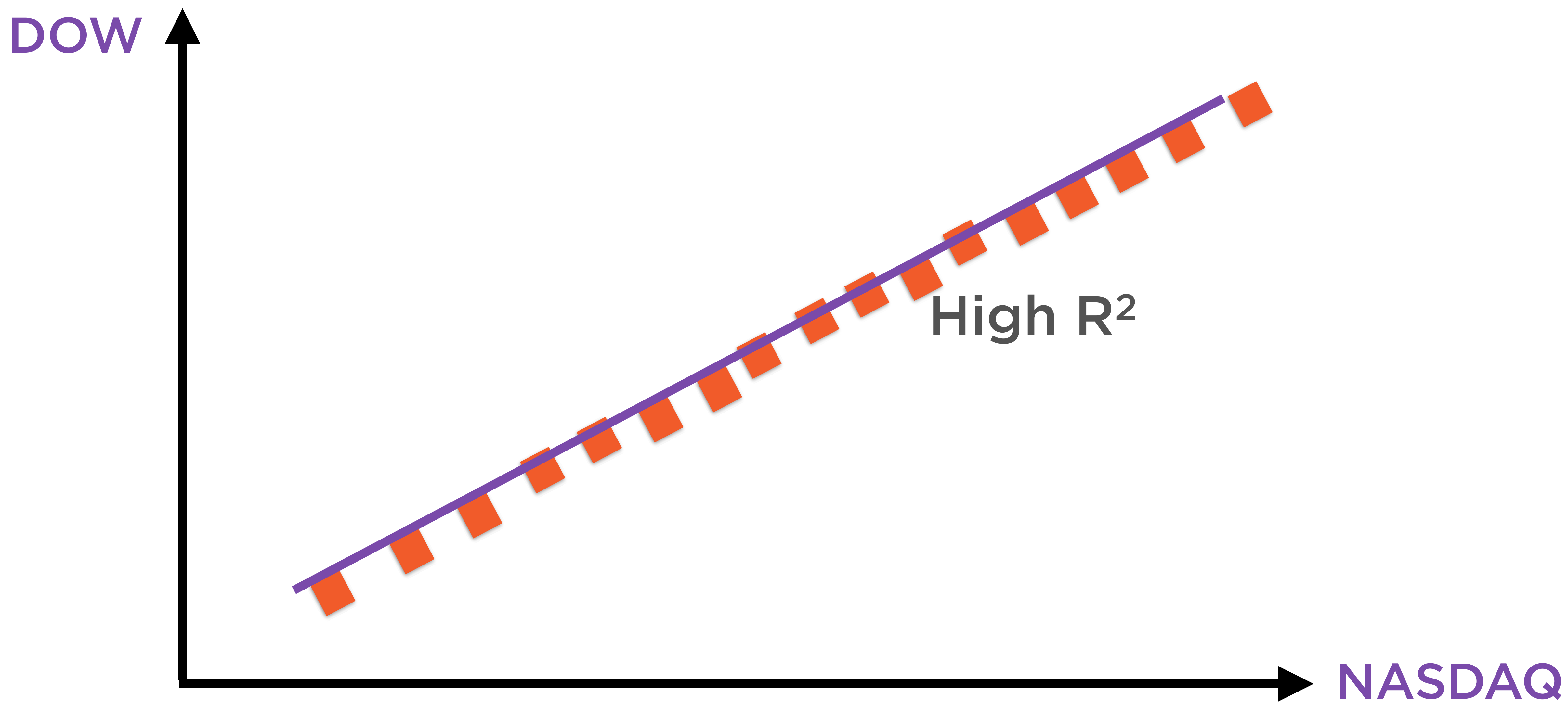
$$\begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \cdots \\ E_n \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} + B \begin{bmatrix} D_1 \\ D_2 \\ D_3 \\ \cdots \\ D_n \end{bmatrix} + C \begin{bmatrix} N_1 \\ N_2 \\ N_3 \\ \cdots \\ N_n \end{bmatrix}$$

E_i = % return on Exxon stock on day i

D_i = % return of Dow Jones index on day i

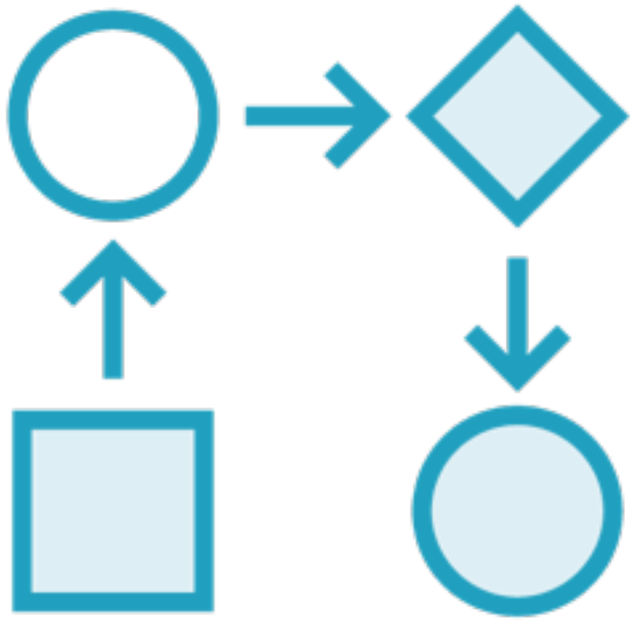
N_i = % return of NASDAQ index on day i

Bad News: Multicollinearity Detected



Highly correlated 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)

Common Sense

Think deeply about each x variable

Eliminate closely related ones

Dig down to underlying causes

Multiple Regression

Proposed Regression Equation:

$$EXXON_t = A + B\ DOW_t + C\ NASDAQ_t + D\ OIL_t$$

% return on Exxon stock on day i

% return of Dow Jones index on day i

% return of NASDAQ index on day i

% change in price of oil on day i

Common Sense

Proposed Regression Equation:

$$EXXON_t = A + B\ DOW_t + C\ NASDAQ_t + D\ OIL_t$$

Dow Jones Industrial Average	NASDAQ 100 Index	Oil Prices
30 Large-cap US stocks	100 large tech stocks	Price of barrel of oil

Common Sense

Proposed Regression Equation:

$$EXXON_t = A + B\ DOW_t + C\ NASDAQ_t + D\ OIL_t$$

Dow Jones Industrial Average

30 Large-cap US stocks

NASDAQ 100 Index

100 large tech stocks

Oil Prices

Price of barrel of oil

Do we really need both Dow and NASDAQ returns as explanatory variables?

Common Sense

Proposed Regression Equation:

$$EXXON_t = A + B\ DOW_t + C\ NASDAQ_t + D\ OIL_t$$

Dow Jones Industrial Average

30 Large-cap US stocks

NASDAQ 100 Index

100 large tech stocks

Oil Prices

Price of barrel of oil

If yes - consider keeping one, and constructing a new explanatory variable of their difference

Common Sense

Proposed Regression Equation:

$$EXXON_t = A + B\ DOW_t + C\ OIL_t$$

Dow Jones Industrial Average

30 Large-cap US stocks

Oil Prices

Price of barrel of oil

What underlying factors drive both US large-cap stocks and the price of oil?

Common Sense

GDP growth

Interest rates

US dollar strength

Seasonality

What underlying factors drive both US large-cap stocks and the price of oil?

Common Sense

GDP growth	Interest rates
US dollar strength	Seasonality

What underlying factors drive both US large-cap stocks and the price of oil?

Multiple Regression

Original Regression Equation:

$$EXXON_t = A + B\ DOW_t + C\ NASDAQ_t + D\ OIL_t$$

Revised Regression Equation:

$$EXXON_t = A + B\ DOW_t + C\ INTEREST_t + D\ GDP_t$$

Nuts and Bolts

‘Standardise’ the variables

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

Set up dummy variables right

Distribute lags

Nuts and Bolts

‘Stepwise regression’ - use with care

Automated selection of x variables

Variants

- Backward elimination
- Forward selection
- Iterative Elimination

Heavy Lifting

Find underlying factors that drive the correlated x variables

Principal Component Analysis (PCA) is a great tool

Multiple Regression

Proposed Regression Equation:

$$HOME_t = A + B\ 5\text{-}yr_t + C\ 10\text{-}year_t + D\ 2\text{-}year_t + E\ 1\text{-}year_t + F\ 3\text{-}month_t + G\ 1\text{-}day_t + \ldots$$

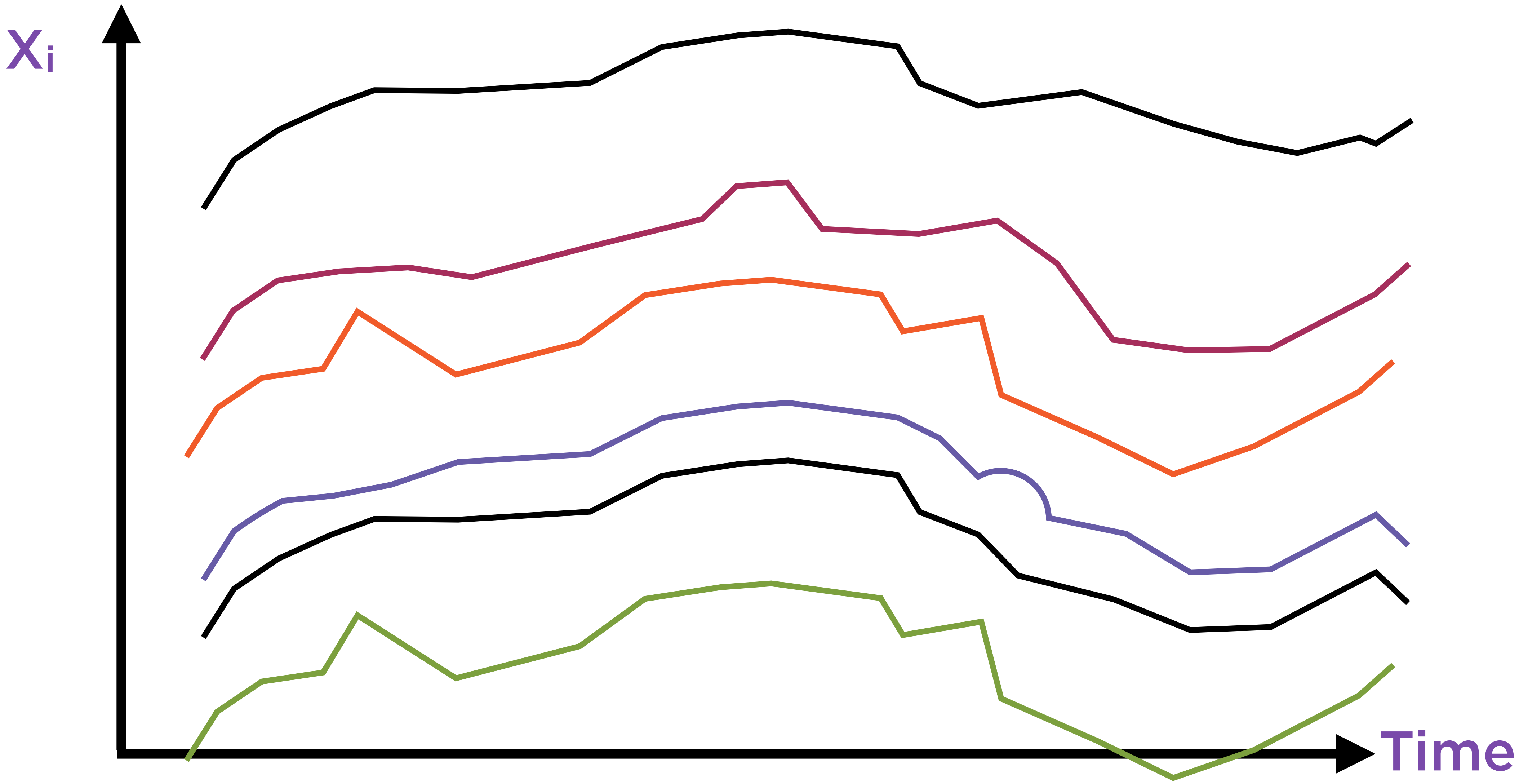
% change in home prices in month i

yield on 5-year bond in month i

yield on 10-year bond in month i

...

Bad News: Multicollinearity Detected



Highly correlated explanatory variables

Factor Analysis

3-month government bonds	1-year government bonds	5-year government bonds
1-day (overnight) money market	30-year government bonds	5-year swap rate (inter-bank)

Interest rates on a wide variety of fixed-income instruments

Factor Analysis on Interest Rates

Level

How high are interest rates?

Slope

How steep is the yield curve?

Twist

How convex is the yield curve?

Three uncorrelated factors explain most variation in all interest rates

Factor Analysis

The factors identified are guaranteed to be uncorrelated

However, they may not have an intuitive interpretation

Principal Component Analysis is one procedure for factor analysis

Dimensionality Reduction via Factor Analysis

20-dimensional data

1 column for each interest rate out there

Factor Analysis

3-dimensional data

3 columns, 1 for each factor

Dimensionality Reduction via Factor Analysis

$$\begin{bmatrix} x_{11} & & x_{1k-1} \\ x_{21} & & x_{2k-1} \\ x_{31} & \dots & x_{3k-1} \\ \dots & & \dots \\ x_{n1} & & x_{nk-1} \end{bmatrix} \longrightarrow$$

1 column for each interest rate out there

Factor Analysis

$$\longrightarrow \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \\ \dots & \dots & \dots \\ f_{n1} & f_{n2} & f_{n3} \end{bmatrix}$$

3 columns, 1 for each factor

Factor Analysis is a dimensionality-reduction technique to identify a few underlying causes in data

Multiple Regression

Proposed Regression Equation:

$HOME_t = A + B\ 5\text{-}yr_t + C\ 10\text{-}year_t + D\ 2\text{-}year_t +$
$E\ 1\text{-}year_t + F\ 3\text{-}month_t + G\ 1\text{-}day_t + \ldots$

Principal Component Analysis

Revised Regression Equation:

$$HOME_t = A + B\ LEVEL_t + C\ SLOPE_t + D\ TWIST_t$$

Benefits of Multiple Regression

Simple Regression Is a Great Tool

Powerful

Perfectly suited to two common use-cases

Versatile

Easily extended to non-linear relationships

Deep

The first “crossover hit” from Machine Learning

Multiple Regression Is Even Better

Powerful

Also controls for effects different causes

Versatile

Also works with categorical data

Deep

Especially if combined with factor analysis

Multiple Regression Is Even Better

Powerful

Also controls for effects different causes

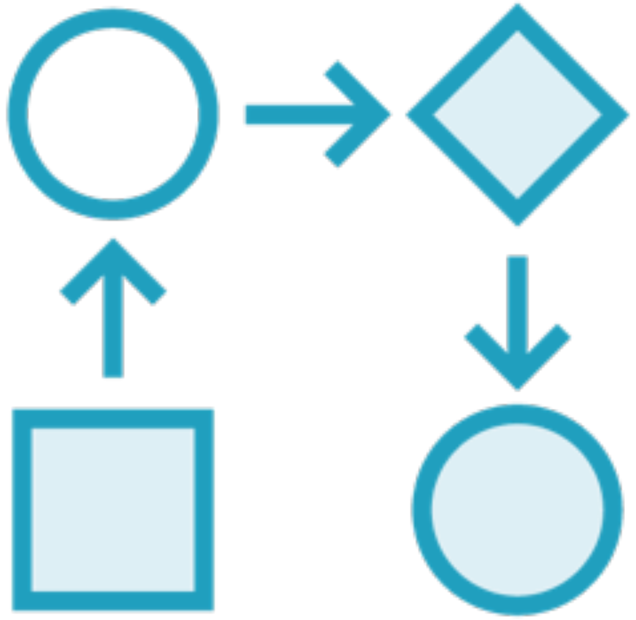
Versatile

Also works with categorical data

Deep

Especially if combined with factor analysis

Two Common Applications of Regression



Explaining Variance

How much variation in one data series is caused by another?

Making Predictions

How much does a move in one series impact another?

Controlling For Different Causes

Proposed Regression Equation:

$$EXXON_t = A + B\ DOW_t + C\ OIL_t$$

% return on Exxon stock on day i

% return of Dow Jones index on day i

% change in price of oil on day i

All else being equal, how much will Exxon stock move by if oil prices increase by 1%?

Controlling For Different Causes

Proposed Regression Equation:

$$EXXON_t = A + B\ DOW_t + C\ OIL_t$$

All else being equal, how much will Exxon stock move by if oil prices increase by 1%?

Controlling For Different Causes

$$EXXON_t = A + B\ DOW_t + C\ OIL_t$$

$$EXXON_? = A + B\ DOW_t + C\ (OIL_t + 1\%)$$

All else being equal, how much will Exxon stock move by if oil prices increase by 1%?

Controlling For Different Causes

$$EXXON_t = A + B\ DOW_t + C\ OIL_t$$

$$EXXON_? = A + B\ DOW_t + C\ (OIL_t + 1\%)$$

$$\text{Change in } EXXON = EXXON_t - EXXON_?$$

All else being equal, how much will Exxon stock move by if oil prices increase by 1%?

Controlling For Different Causes

$$\text{EXXON}_t = \cancel{A} + B\ \text{D}\cancel{\text{O}}\text{W}_t + C\ \text{O}\cancel{\text{I}}\text{L}_t$$

$$-\quad \text{EXXON}_? = \cancel{A} + B\ \text{D}\cancel{\text{O}}\text{W}_t + C\ (\text{O}\cancel{\text{I}}\text{L}_t + 1\%)$$

All else being equal, how much will Exxon stock move by if oil prices increase by 1%?

Controlling For Different Causes

$$EXXON_t = A + B\ DOW_t + C\ OIL_t$$

$$-\quad EXXON_? = A + B\ DOW_t + C\ (OIL_t + 1\%)$$

$$\text{Change in EXXON} = C$$

All else being equal, how much will Exxon stock move by if oil prices increase by 1%?

Regression coefficients tell how much y changes for a unit change in each predictor, **all others being held constant**

Multiple Regression Is Even Better

Powerful

Also controls for effects different causes

Versatile

Also works with categorical data

Deep

Especially if combined with factor analysis

Multiple Regression Is Even Better

Powerful

Also controls for effects different causes

Versatile

Also works with categorical data

Deep

Especially if combined with factor analysis

Multiple Regression Is Even Better

Powerful

Also controls for effects different causes

Versatile

Also works with categorical data

Deep

Especially if combined with factor analysis

Interpreting the Results of a Regression Analysis

Interpreting Results of a Simple Regression

R^2

Measures overall quality of fit - the higher the better (up to a point)

Residuals

Check if regression assumptions are violated

Standard errors of individual coefficients are usually of little significance

Interpreting Results of a Multiple Regression

Adjusted R^2

Residuals

F-statistic

Standard Errors of coefficients

R^2

e = y - y’

=> y = y’ + e

=> Variance(y) = Variance(y’ + e)

=> Variance(y) = Variance(y’) + Variance(e) + Covariance(y’,e)

A Not-Very-Important Intermediate Step

Variance of the dependent variable can be decomposed into variance of the regression fitted values, and that of the residuals

e = y' - y

=> y = y' + e

=> Variance(y) = Variance(y' + e)

Always = 0

=> Variance(y) = Variance(y') + Variance(e) + $\boxed{\text{Covariance(y',e)}}$

A Leap of Faith

This is important - more on why in a bit

Variance(y) = Variance(y’) + Variance(e)

Variance Explained

Variance of the dependent variable can be decomposed into variance of the regression fitted values, and that of the residuals

$\boxed{\text{Variance(y)}}$ = Variance(y') + Variance(e)

Total Variance (*TSS*)

A measure of how volatile the dependent variable is, and of much it moves around

TSS = $\boxed{\text{Variance(y')}}$ + Variance(e)

Explained Variance *(ESS)*

A measure of how volatile the fitted values are - these come from the regression line

TSS = Variance(y)

$TSS = ESS + \boxed{Variance(e)}$

Residual Variance *(RSS)*

This the variance in the dependent variable that can not be explained by the regression

TSS = Variance(y)   ESS = Variance(y')

TSS = ESS + RSS

Variance Explained

Variance of the dependent variable can be decomposed into variance of the regression fitted values, and that of the residuals

TSS = Variance(y)   ESS = Variance(y') RSS = Variance(e)

$R^2 = ESS / TSS$

R^2

The percentage of total variance explained by the regression. Usually, the higher the R^2, the better the quality of the regression (upper bound is 100%)

R^2 = ESS / TSS

R^2

In multiple regression, adding explanatory variables always increases R^2, even if those variables are irrelevant and increase danger of multicollinearity

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)

Nuts and Bolts

‘Stepwise regression’ - use with care

Automated selection of x variables

Variants

- Backward elimination
- Forward selection
- Iterative Elimination

Extending Multiple Regression to Categorical Variables

A Simple Regression

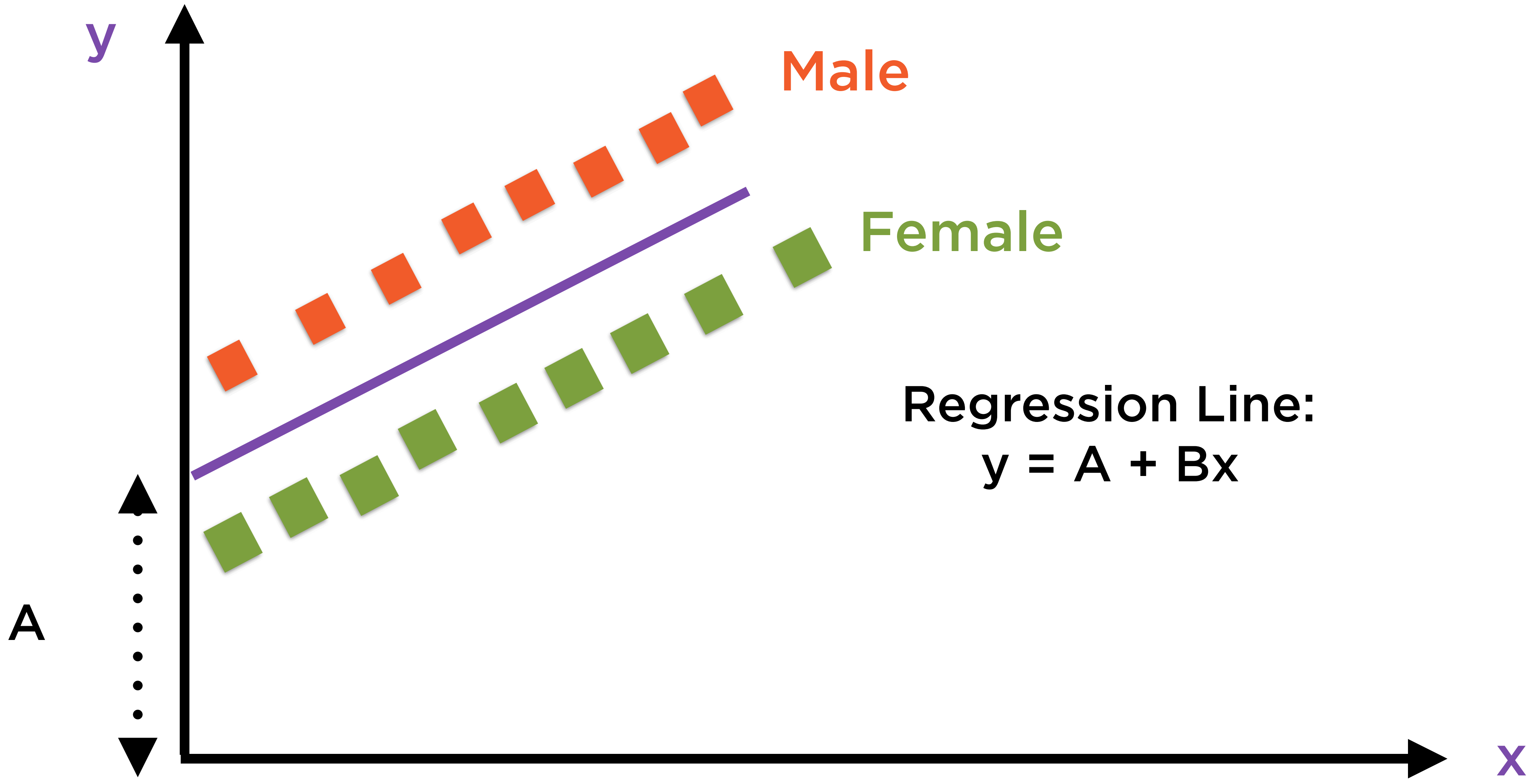
Proposed Regression Equation:

$$y = A + Bx$$

Height of individual

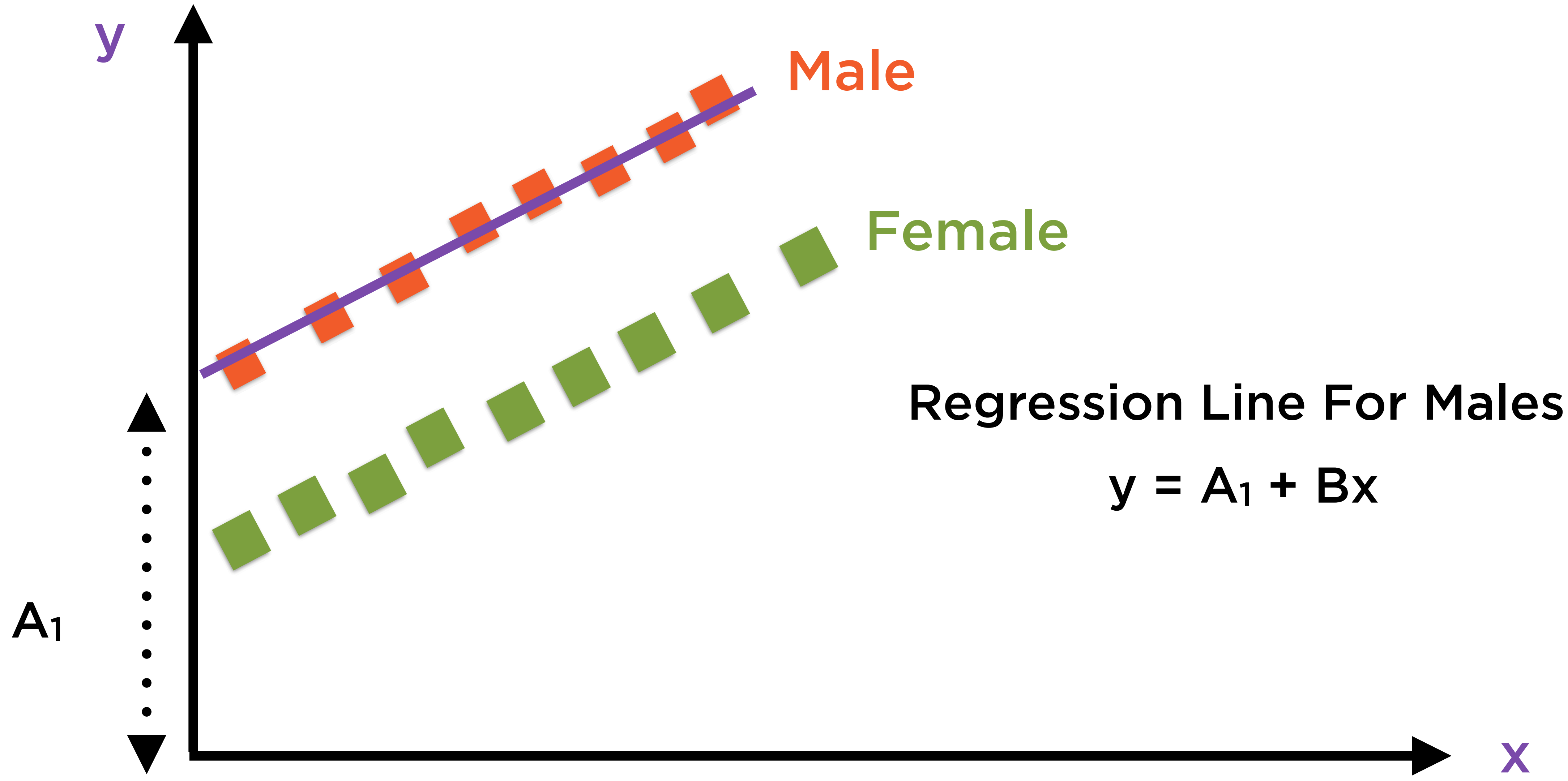
Average height of parents

A Simple Regression



Not a great fit - regression line is far from all points!

A Simple Regression

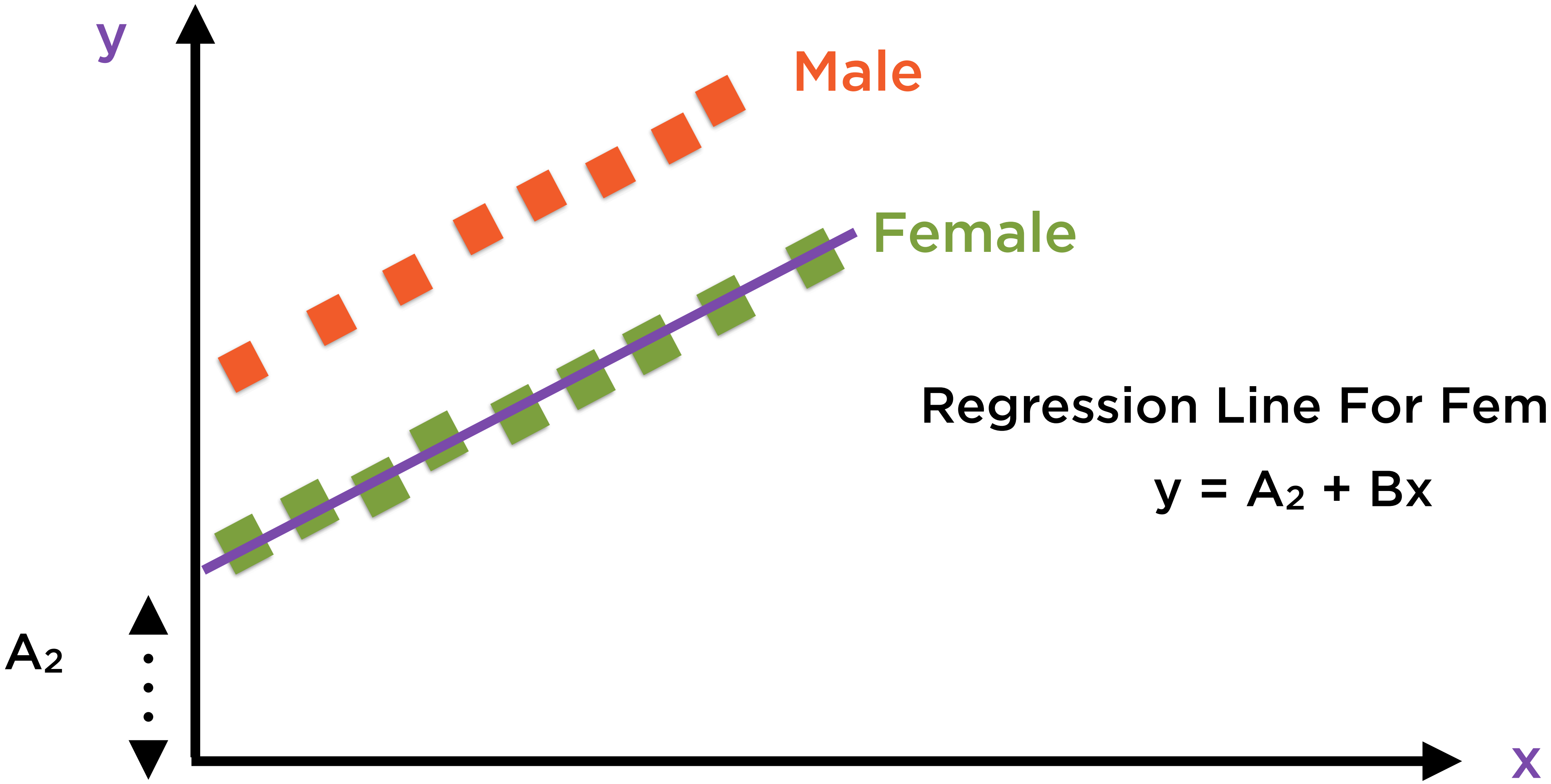


Regression Line For Males:

$$y = A_1 + Bx$$

We can easily plot a great fit for males...

A Simple Regression

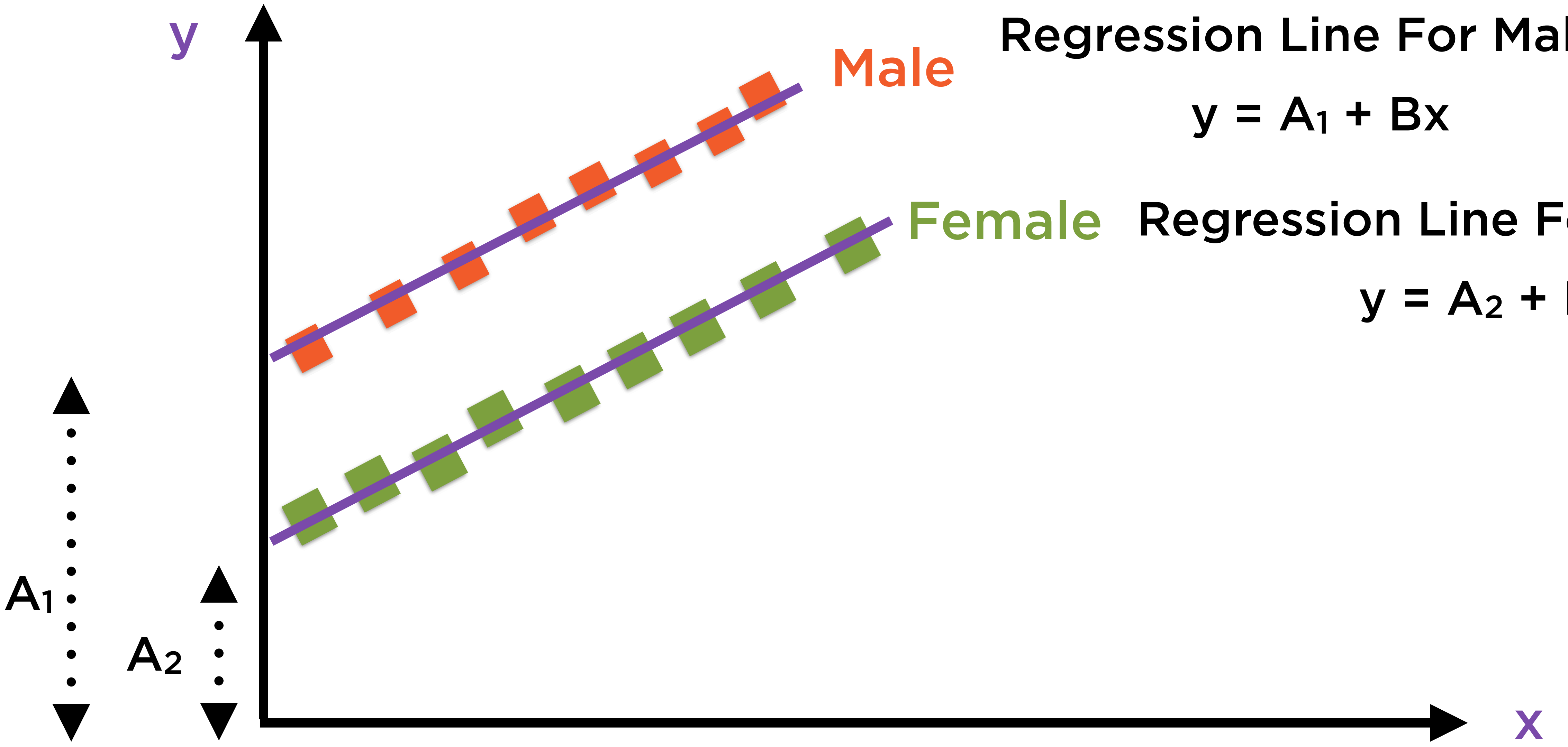


Regression Line For Females:

$$y = A_2 + Bx$$

...and another great fit for females

A Simple Regression



Regression Line For Males:

$$y = A_1 + Bx$$

Regression Line For Females:

$$y = A_2 + Bx$$

Two lines - same slope, different intercepts

Adding A Dummy Variable

Regression Line For Males:

$y = A_1 + Bx$

Regression Line For Females:

$y = A_2 + Bx$

Combined Regression Line:

$$y = A_1 + (A_2 - A_1)D + Bx$$

$D = 0$ for males

$= 1$ for females

Adding A Dummy Variable

Regression Line For Males:

$y = A_1 + Bx$

Regression Line For Females:

$y = A_2 + Bx$

Combined Regression Line:

$$y = A_1 + (A_2 - A_1)D + Bx$$

$D = 0$ for males

$$y = A_1 + \cancel{(A_2 - A_1)D} + Bx$$

$$= A_1 + Bx$$

Adding A Dummy Variable

Regression Line For Males:

$y = A_1 + Bx$

Regression Line For Females:

$y = A_2 + Bx$

Combined Regression Line:

$$y = A_1 + (A_2 - A_1)D + Bx$$

$D = 1$ for females

$$y = \cancel{A_1} + (A_2 \cancel{- A_1}) + Bx$$

$$= A_2 + Bx$$

Adding A Dummy Variable

Original Regression Equation:

$$y = A + Bx$$

Height of individual

Average height of parents

Combined Regression Line:

$$y = A_1 + (A_2 - A_1)D + Bx$$

$D = 0$ for males

$= 1$ for females

Adding A Dummy Variable

Combined Regression Line:

$$y = A_1 + (A_2 - A_1)D + Bx$$

$D = 0$ for males

$= 1$ for females

The data contained 2 groups, so we added 1 dummy variable

Given data with k groups, set up k-1 dummy variables, else multicollinearity occurs

Dummy and Other Categorical Variables

Dummy Variables

Binary - 0 or 1

Categorical Variables

Finite set of values - e.g. days of week, months of year...

To include non-binary categorical variables, simply add more dummies

Testing for Seasonality

Proposed Regression Equation:

$$y = A + BQ_1 + CQ_2 + DQ_3$$

Average stock returns

Quarter of the year

The data contains 4 groups, so we added 3 dummy variables

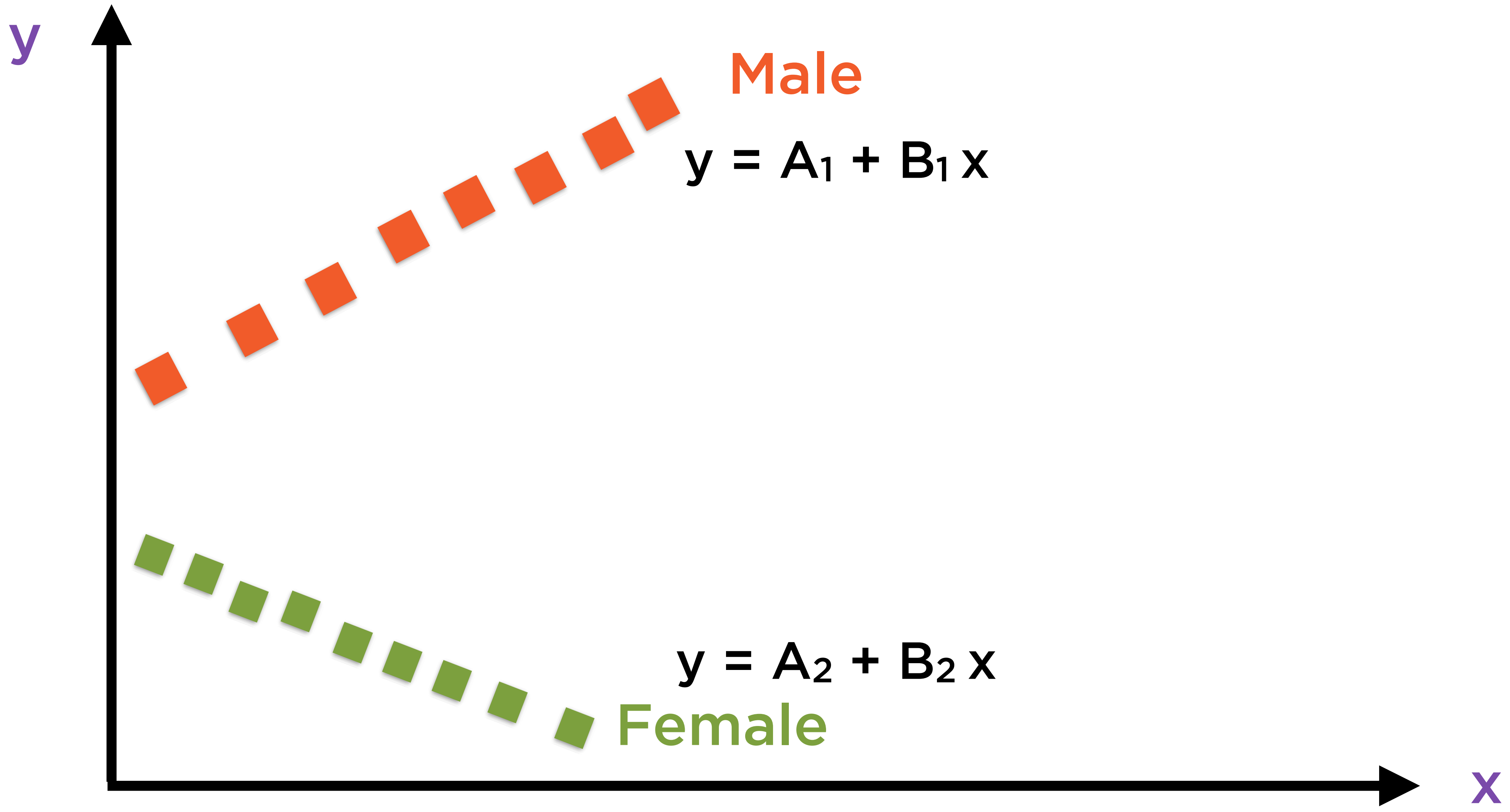
Testing for Seasonality

$$y = A + BQ_1 + CQ_2 + DQ_3$$

The data contains 4 groups, so we added 3 dummy variables

Q_1	$= 1$	for Jan, Feb, Mar
	$= 0$	for other quarters
Q_2	$= 1$	for Apr, May, Jun
	$= 0$	for other quarters
Q_3	$= 1$	for July, Aug, Sep
	$= 0$	for other quarters

Different Groups, Different Slopes



Dummy variables can also be extended for use where groups have different slopes

Adding A Dummy Variable

Regression Line For Males:

$y = A_1 + B_1 x$

Regression Line For Females:

$y = A_2 + B_2 x$

Combined Regression Line:

$$y = A_1 + (A_2 - A_1)D_1 + B_1 x + (B_2 - B_1)D_2$$

$D_1 = 0$ for males

$= 1$ for females

$D_2 = 0$ for males

$= x$ for females

Adding A Dummy Variable

Regression Line For Males:

$y = A_1 + B_1 x$

Regression Line For Females:

$y = A_2 + B_2 x$

For males:

$D_1 = 0$

$$y = A_1 + (A_2 - A_1)\boxed{D_1} +$$
$$B_1 x + (B_2 - B_1)\boxed{D_2}$$

$D_2 = 0$

$$= A_1 + B_1 x$$

Adding A Dummy Variable

Regression Line For Males:

$y = A_1 + B_1 x$

Regression Line For Females:

$y = A_2 + B_2 x$

For females:

$D_1 = 1$

$$y = A_1 + (A_2 - A_1)\boxed{(1)} + B_1 x + (B_2 - B_1)\boxed{x}$$

$D_2 = x$

$$= A_1 + (A_2 - A_1) + B_1 x + (B_2 - B_1)x$$

$$= A_2 + B_2 x$$

Dummy Variables

X

Linear regression

Y

Logistic regression

Summary

Understood the formidable benefits of multiple regression

Mitigated the significant risks that come with those benefits

Understood the utility of Adjusted-R^2

Used multiple regression to work with categorical variables