# Understanding and Implementing Dummy Coding


Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

# Overview

Regression with dummy variables

Limitations of one-hot encoding

The dummy variable trap

Overcoming the limitations of one-hot encoding with dummy encoding

Performing dummy or treatment coding in regression analysis

One-hot encoding: k columns for k categories

Dummy coding: k-1 columns for k categories

# The Dummy Trap in Linear Regression

# A Simple Regression

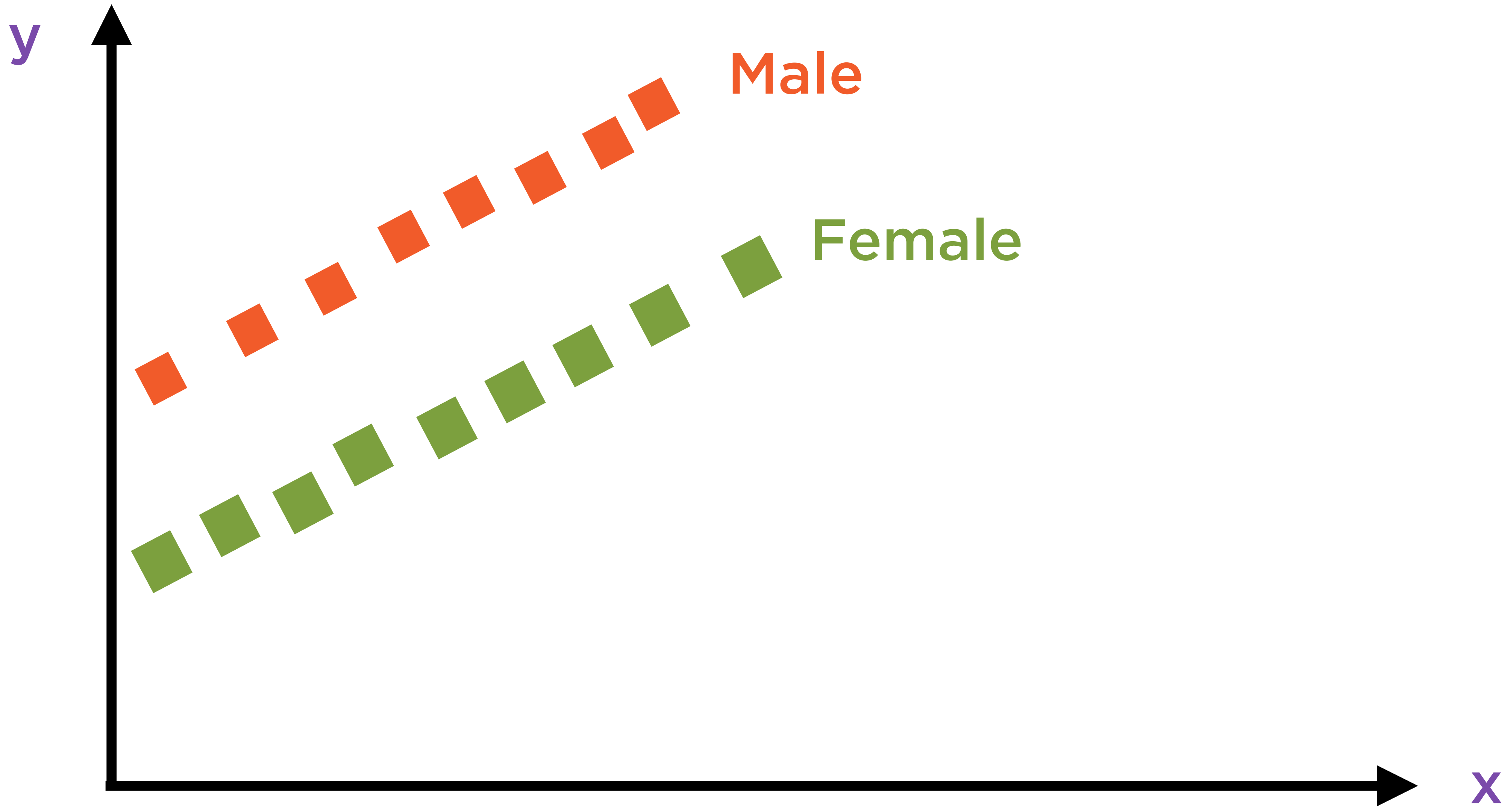
**Proposed Regression Equation:**

$$y = A + Bx$$

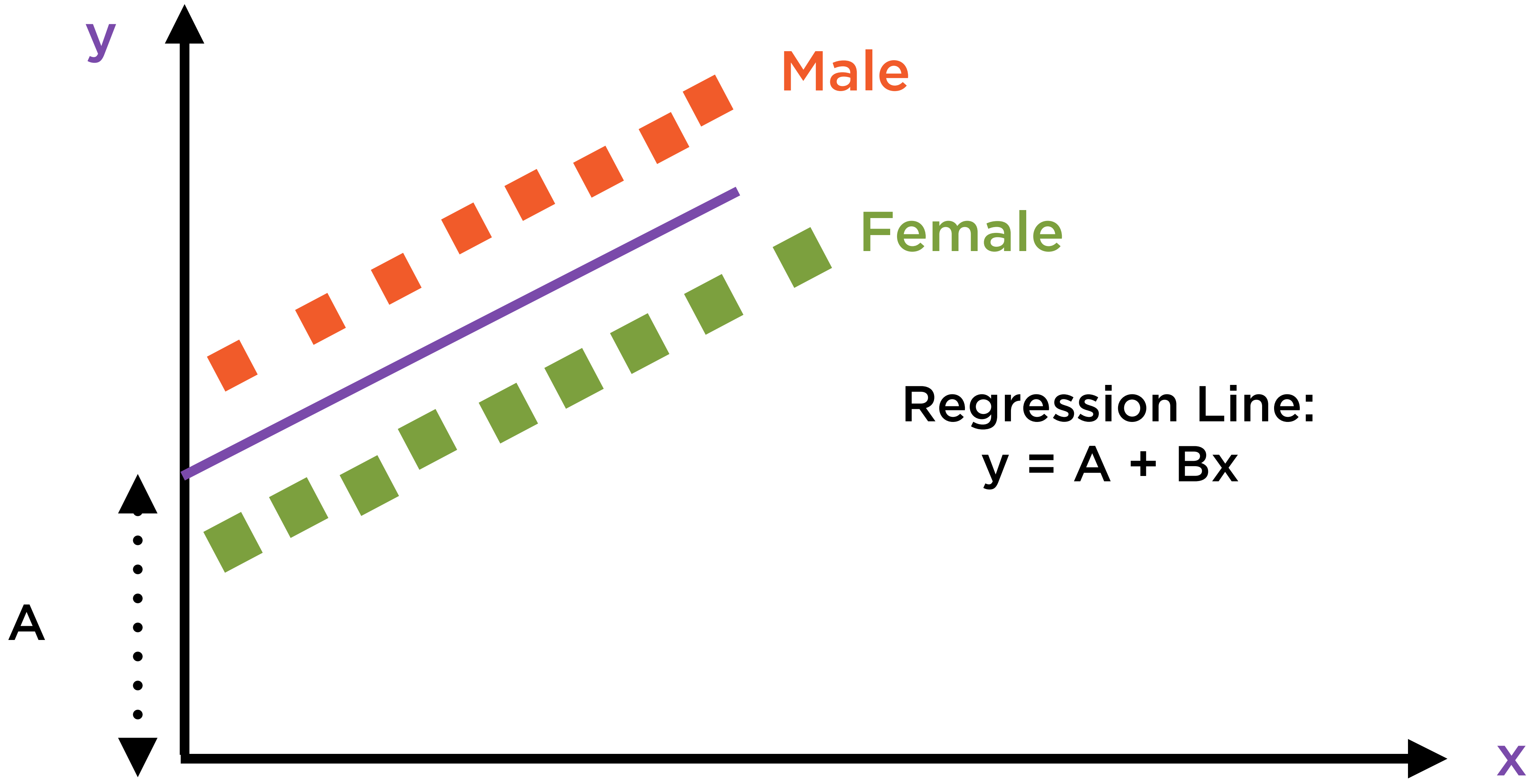
Height of individual

Average height of parents

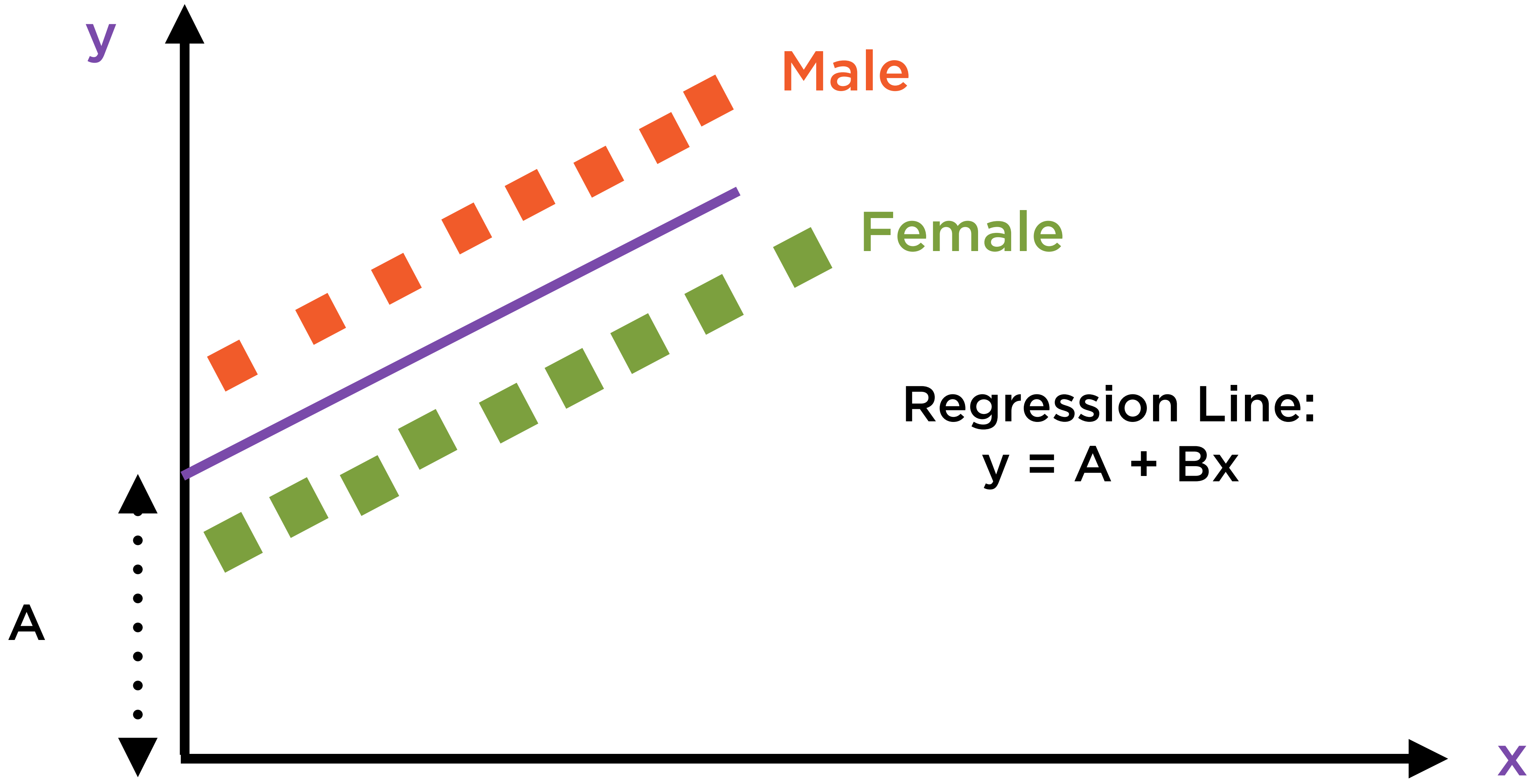
# A Simple Regression



# A Simple Regression

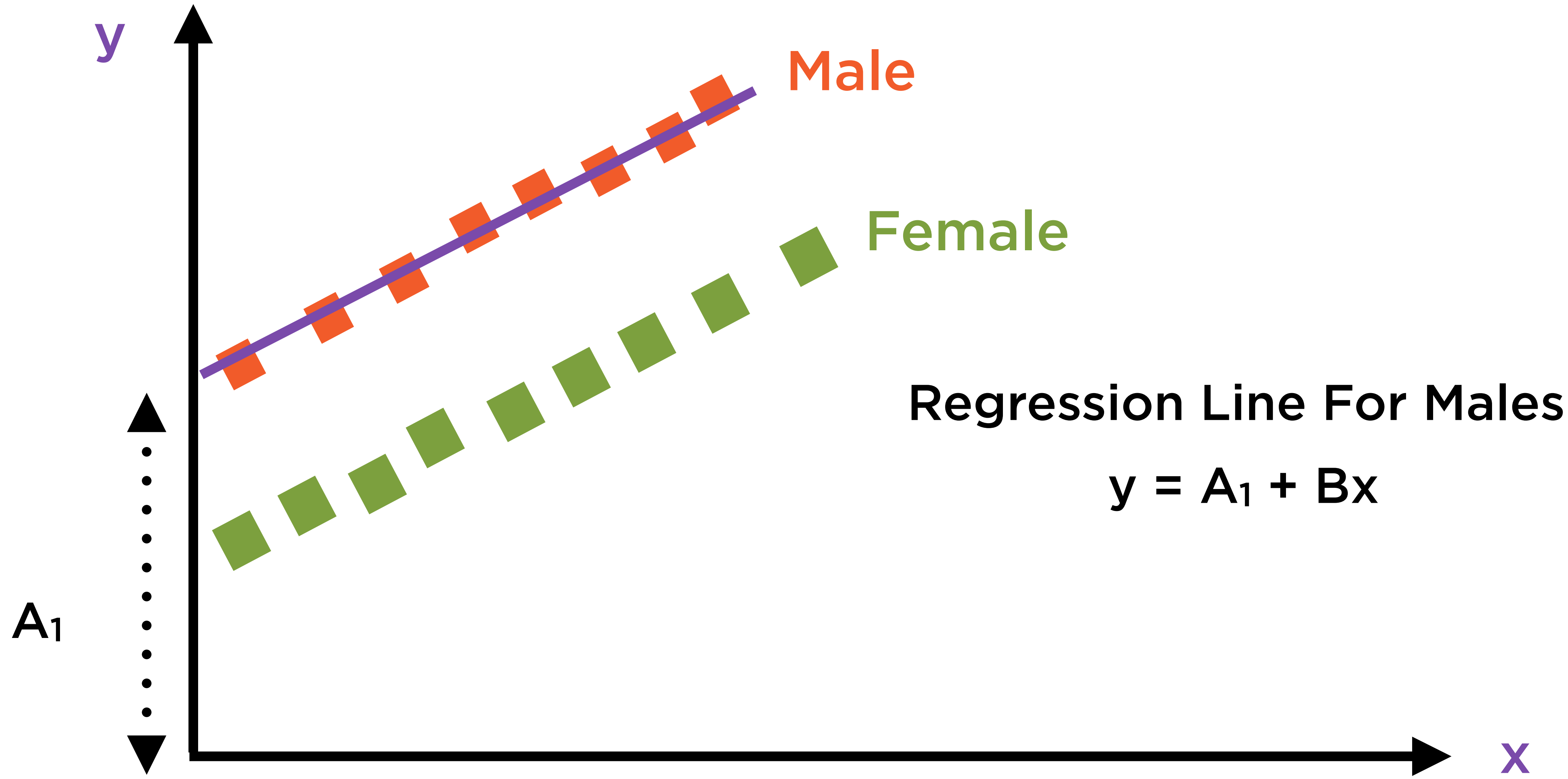


# A Simple Regression




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

# A Simple Regression

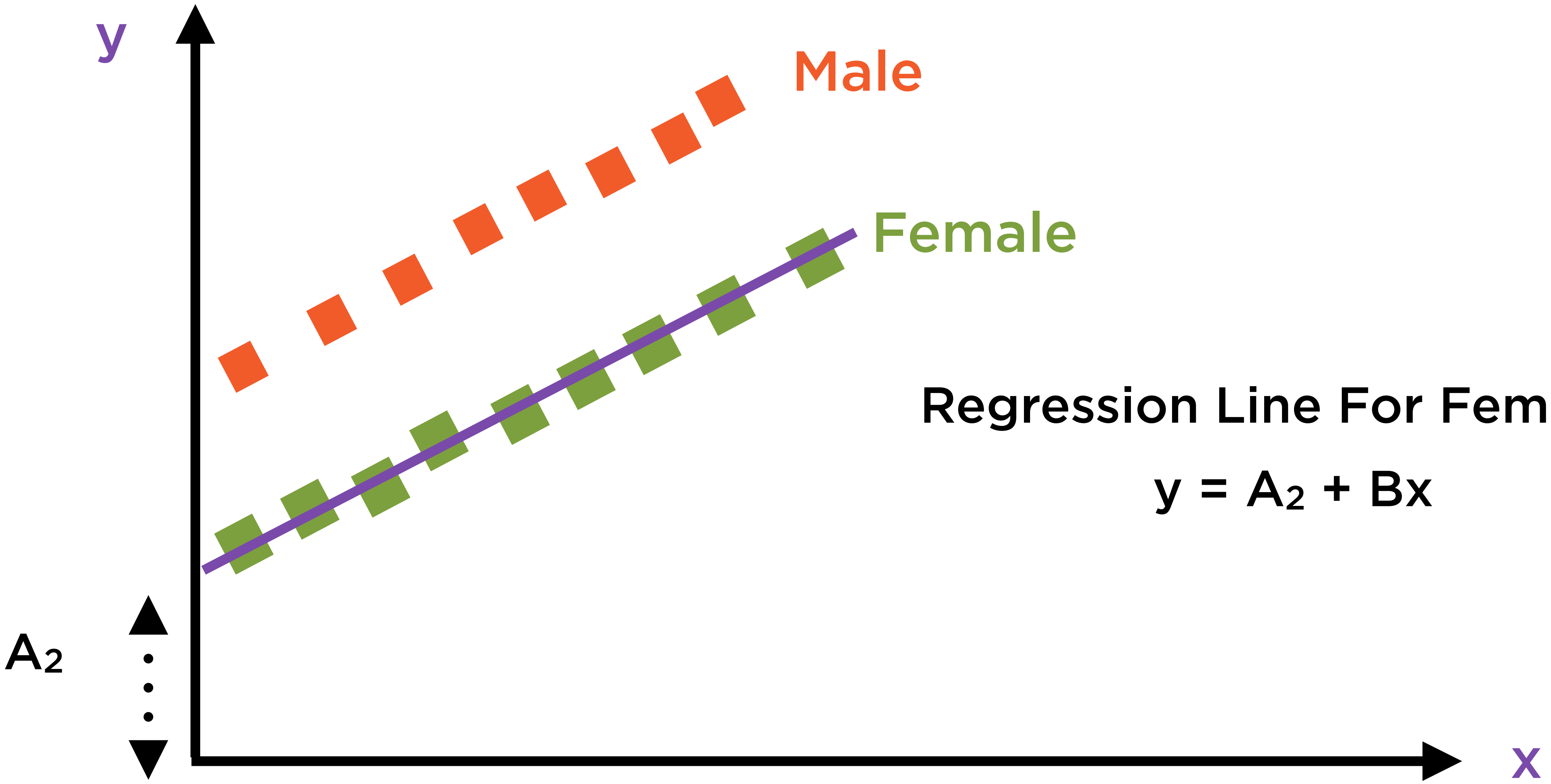



Regression Line For Males:

$y = A_1 + Bx$

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

# A Simple Regression

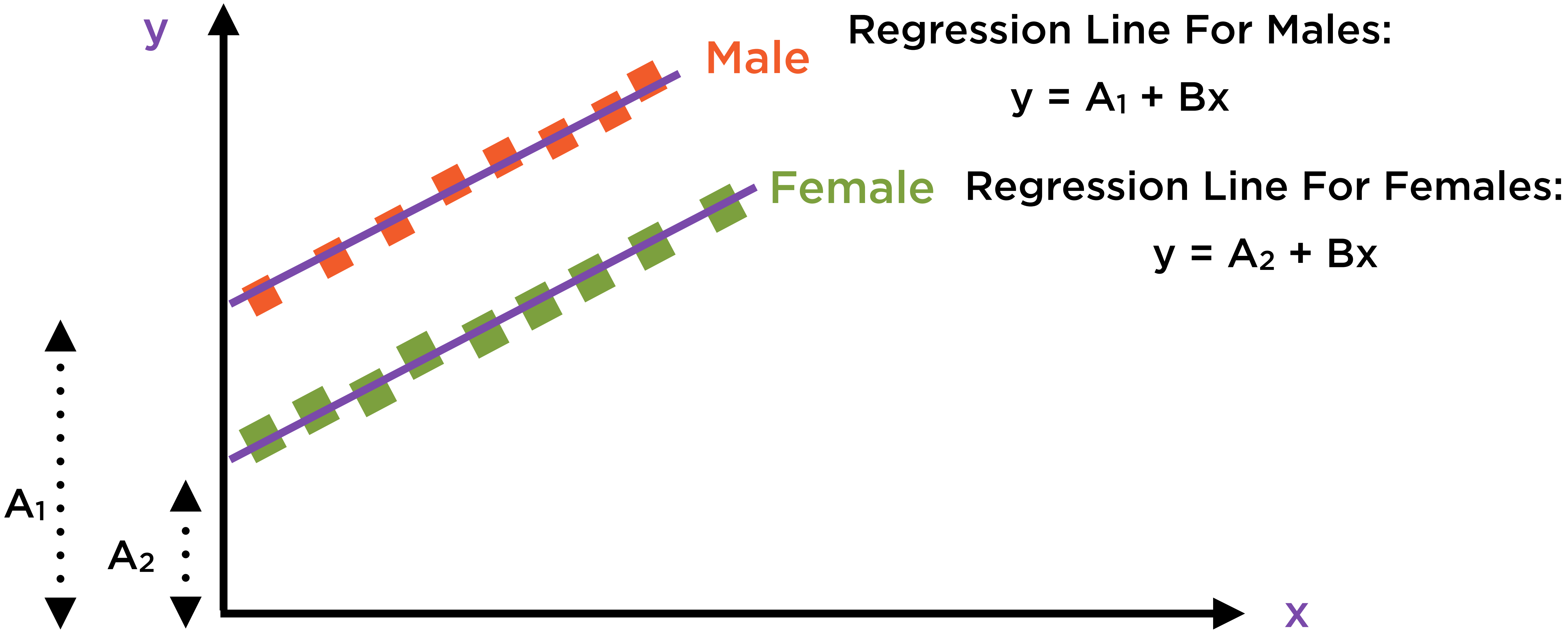



**Regression Line For Females:**

$y = A_2 + Bx$

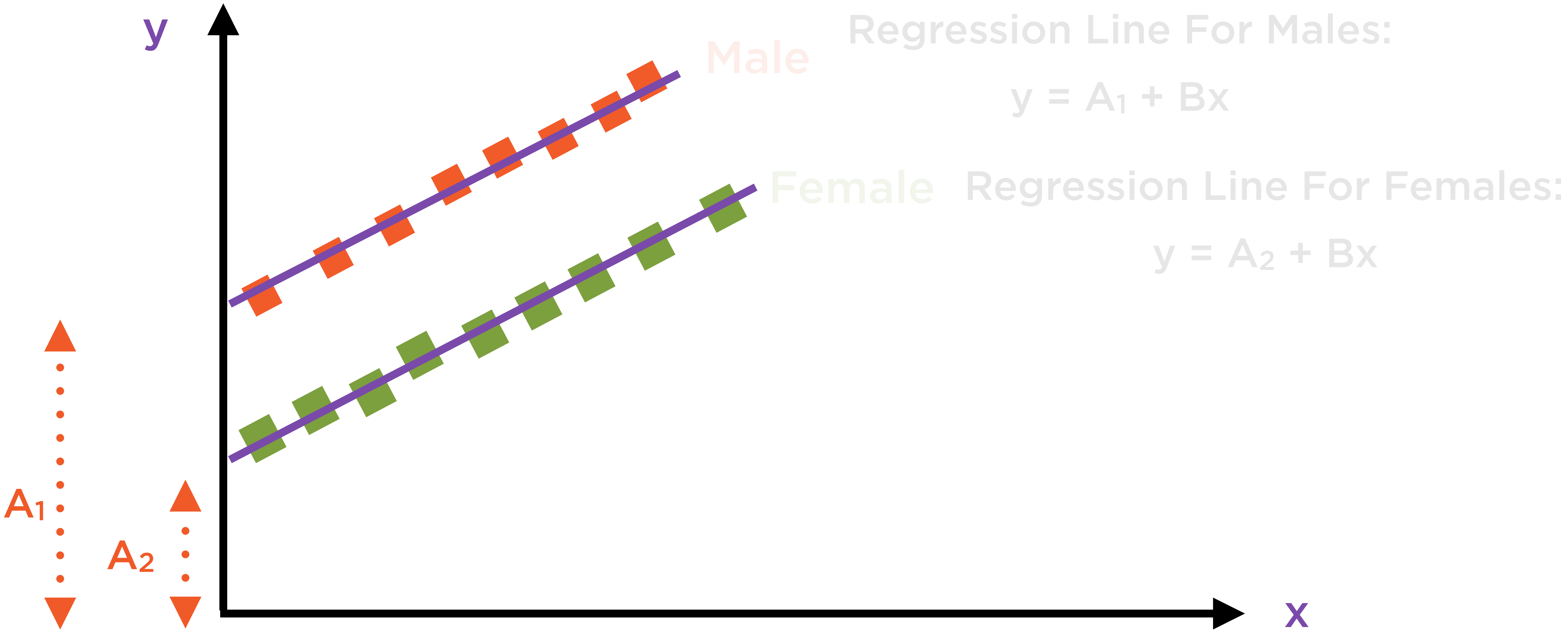
**...and another great fit for females**

# A Simple Regression




**Two lines - same slope, different intercepts**

# A Simple Regression




**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

$= 1$ 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 levels (groups), so we added 1 dummy variable and kept the intercept**

# The Dummy Trap

**Combined Regression Line:**

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

$D = 0$ for males

$= 1$ for females

Adding 2 dummy variables here would have led us into the Dummy Trap

# Dummy Variable Trap

**If a categorical variable is used as a feature (x-variable) in linear regression**

**And if that categorical variable has k levels**

**Trap: Using k dummy variables and an intercept**

**Causes multi-collinearity and an unstable regression model**

Dummy Trap: Using k dummy variables <u>and</u> an intercept

**Unstable regression model**

# Avoiding the Dummy Trap

# Dummy Variable Trap

If a categorical variable is used as a feature (x-variable) in linear regression

And if that categorical variable has k levels

Trap: Using k dummy variables and an intercept

Causes multi-collinearity and an unstable regression model

# Avoiding the Dummy Variable Trap

**Use either**

- k dummy variables and exclude the intercept
- k-1 dummy variables and include the intercept

**In either case, k levels need k variables (including the intercept)**

Avoid the Dummy Variable Trap:
k levels need k variables
(including the intercept)

# Avoiding the Dummy Variable Trap

**Use either**

- k dummy variables and exclude the intercept
- k-1 dummy variables and include the intercept

**In either case, k levels need k variables (including the intercept)**

# Avoiding the Dummy Variable Trap

Use either

- k dummy variables and exclude the intercept
- k-1 dummy variables and include the intercept

In either case, k levels need k variables (including the intercept)

# Avoiding the Dummy Variable Trap

**Using k-1 variables and including an intercept is the usual choice**

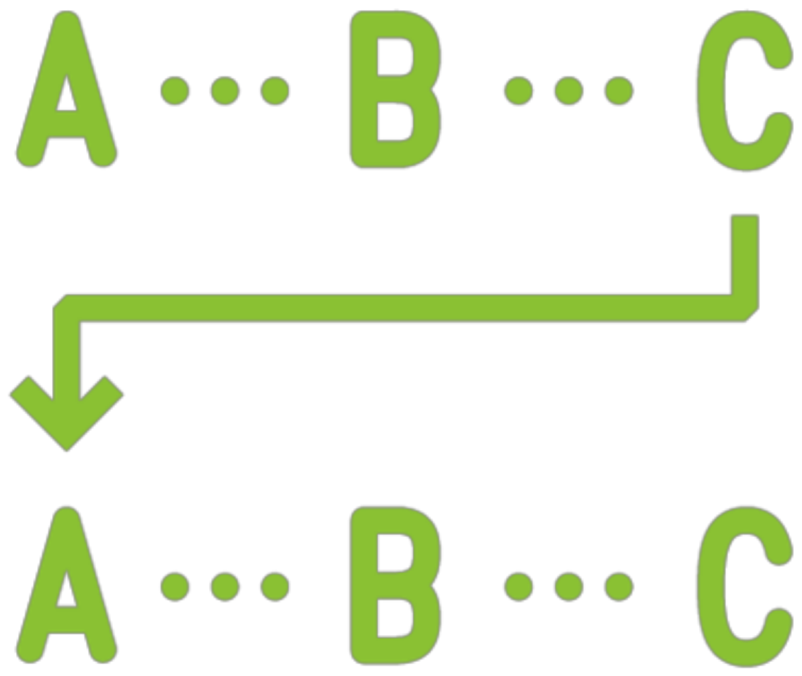
**The excluded level is called the reference level**

# Avoiding the Dummy Variable Trap

Using k-1 variables and including an intercept is the usual choice

The excluded level is called the reference level

# Reference Level




Represented by the intercept in the regression

The coefficients of other levels are expressed in terms of the reference level

# 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

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 |

# 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 |

# 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 |

# 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 |

# Overcoming the Limitations of One-hot Coding

Avoid using one-hot encoded categories with intercept - this leads to the dummy trap

# Dummy Variable Trap

If a categorical variable is used as a feature (x-variable) in linear regression

And if that categorical variable has k levels

Trap: Using k dummy variables and an intercept

Causes multi-collinearity and an unstable regression model

# One-hot Encoded Cities

| Category | New York | London | Paris | Bangalore |
|---|---|---|---|---|
| New York | 1 | 0 | 0 | 0 |
| London | 0 | 1 | 0 | 0 |
| Paris | 0 | 0 | 1 | 0 |
| Bangalore | 0 | 0 | 0 | 1 |

k categories and k columns to represent
k categories

# One-hot Encoded Cities

| Category | New York | London | Paris | Bangalore |
|---|---|---|---|---|
| New York | 1 | 0 | 0 | 0 |
| London | 0 | 1 | 0 | 0 |
| Paris | 0 | 0 | 1 | 0 |
| Bangalore | 0 | 0 | 0 | 1 |

**Cannot use directly if performing regression with intercept**

# Avoiding the Dummy Variable Trap

Use either

- k dummy variables and exclude the intercept
- k-1 dummy variables and include the intercept

In either case, k levels need k variables (including the intercept)

Solution: use one-hot encoding but **drop** one category column

**Dummy encoding**

# Dummy Encoded Cities

| Category | New York | London | Paris |
|---|---|---|---|
| New York | 1 | 0 | 0 |
| London | 0 | 1 | 0 |
| Paris | 0 | 0 | 1 |
| Bangalore | 0 | 0 | 0 |

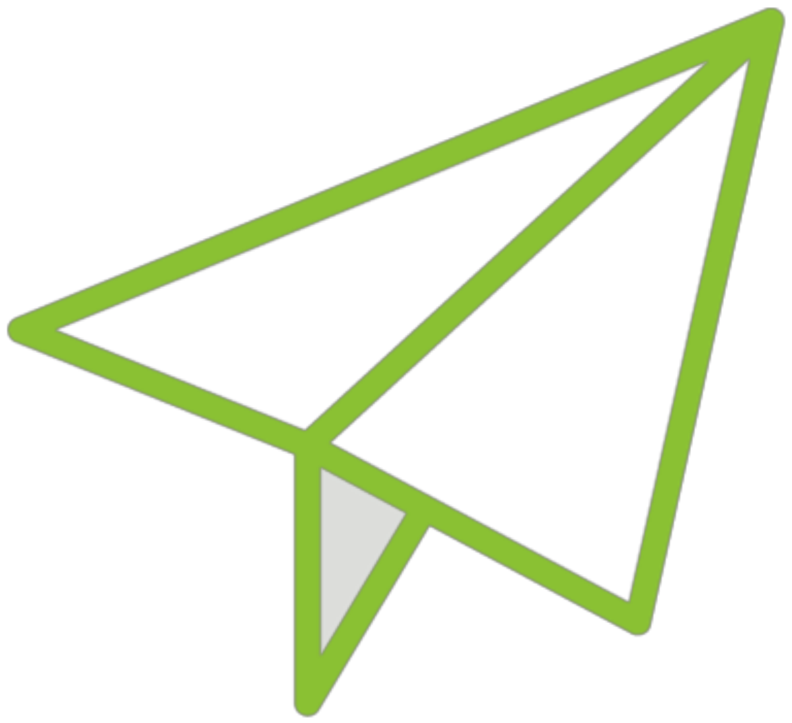
**k categories and k-1 columns to represent k categories**

# Dummy Encoded Cities

| Category | New York | London | Paris |
|---|---|---|---|
| New York | 1 | 0 | 0 |
| London | 0 | 1 | 0 |
| Paris | 0 | 0 | 1 |
| Bangalore | 0 | 0 | 0 |

**Bangalore is the reference level or category**

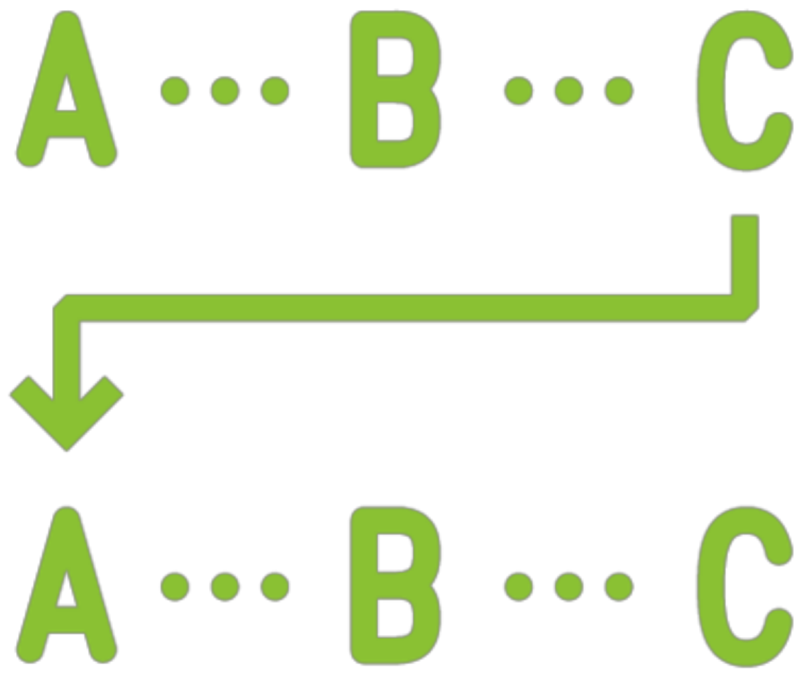
# Dummy Coding



Name used for scheme with k-1 dummy variables along with intercept

Excluded level is called the reference level

# Reference Level




Represented by the intercept in the regression

The coefficients of other levels are expressed in terms of the reference level

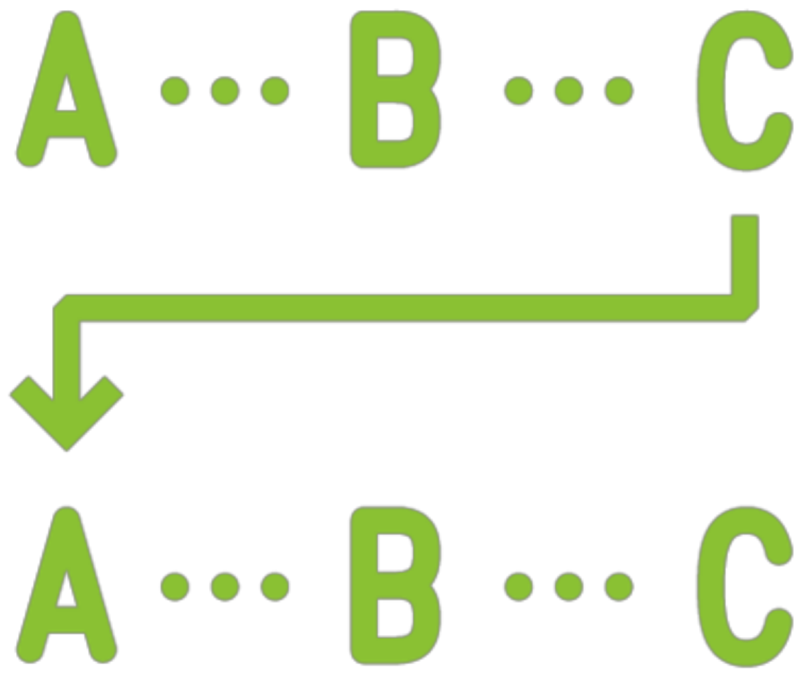
# Dummy Coding with Linear Regression

| | |
|---|---|
| Application | Compare other levels to reference |
| Intercept | Mean of y-values of reference level |
| Coefficient for level(i) | Mean of y-values of level(i) - mean of y-value for reference level |

If **no information** available for a data point  i.e. all coefficients are zero

The y-value for that point is assumed to be the average y-value for the **reference** level

# Intercept Value for Dummy Coding




**Intercept (constant) will be the mean y-value for reference level**

**Coefficients of other dummies will be in terms of reference level too**

**Coefficient of each included variable = Mean of y-values of that level - Mean of y-values of reference level**

# A Simple Regression

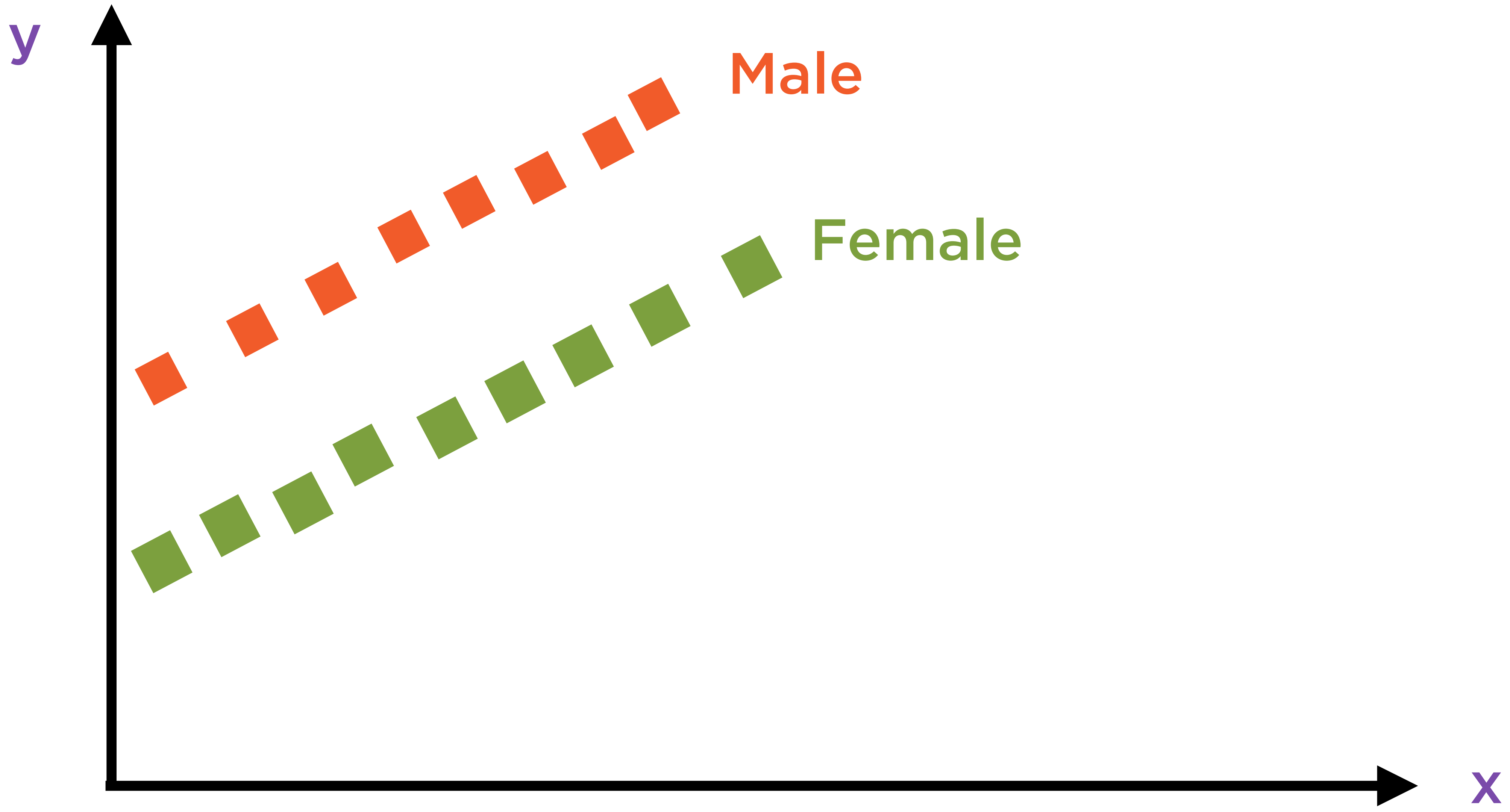
**Proposed Regression Equation:**

$$y = A + Bx$$

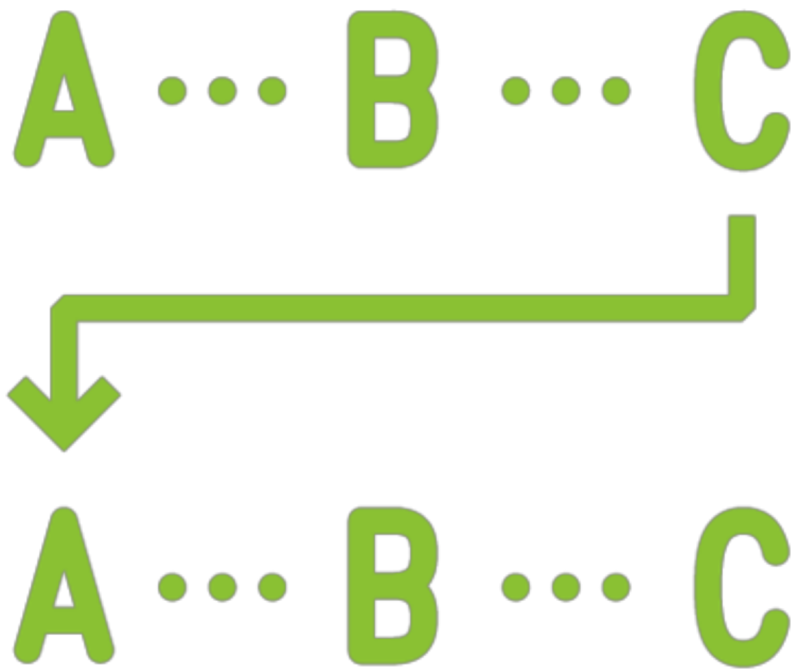
Height of individual

Average height of parents

# A Simple Regression



# Intercept Value for Dummy Coding

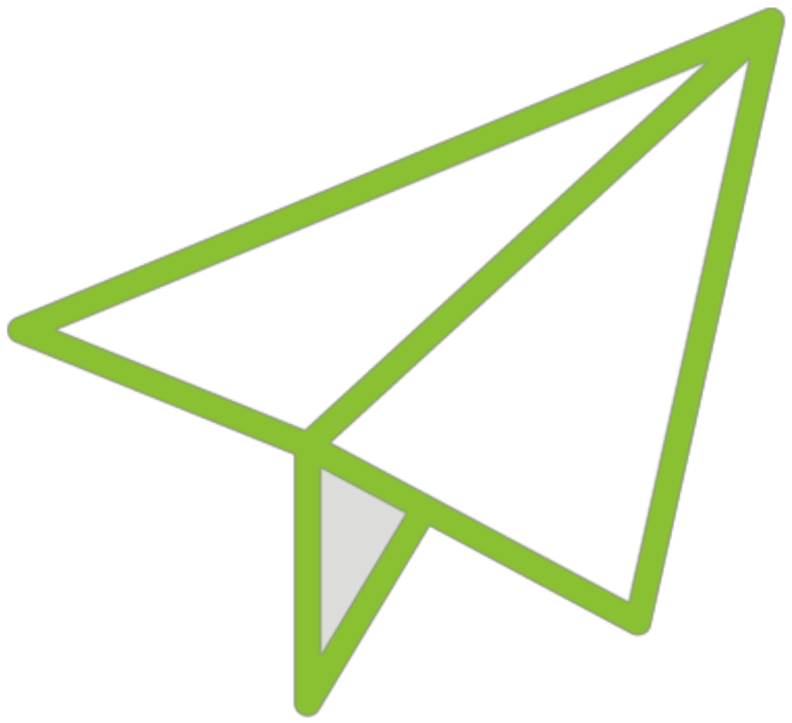



Reference level = father (males)

Intercept = mean height of fathers (males) in data

Coefficient for height of mother = mean height of mothers (females) - mean height of fathers (males)

# Assumptions in Dummy Coding



Dummy coding does not assume independence of coefficients

ANOVA assumes independent coefficients but linear regression does not

Which is why dummy coding is most often used with linear regression

# Demo

Performing linear regression using dummy encoding

# Summary

Regression with dummy variables

Limitations of one-hot encoding

The dummy variable trap

Overcoming the limitations of one-hot encoding with dummy encoding

Performing dummy or treatment coding in regression analysis