Implementing Multiple Regression Models in R

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

Overview

Implement multiple regression in R

Interpret results of a multiple regression

Carry out multiple regression in R to include categorical variables

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

Interpreting Results of a Multiple Regression

Adjusted R^2

Residuals

F-statistic

Standard Errors of coefficients

R^2

Interpreting Results of a Multiple Regression

Adjusted R^2

Residuals

F-statistic

Standard Errors of coefficients

R^2

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

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)

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

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)

Interpreting Results of a Multiple Regression

Adjusted R^2

Residuals

F-statistic

Standard Errors of coefficients

R^2

Interpreting Results of a Multiple Regression

Adjusted R^2

Residuals

F-statistic

Standard Errors of coefficients

R^2

Population and Sample

Population

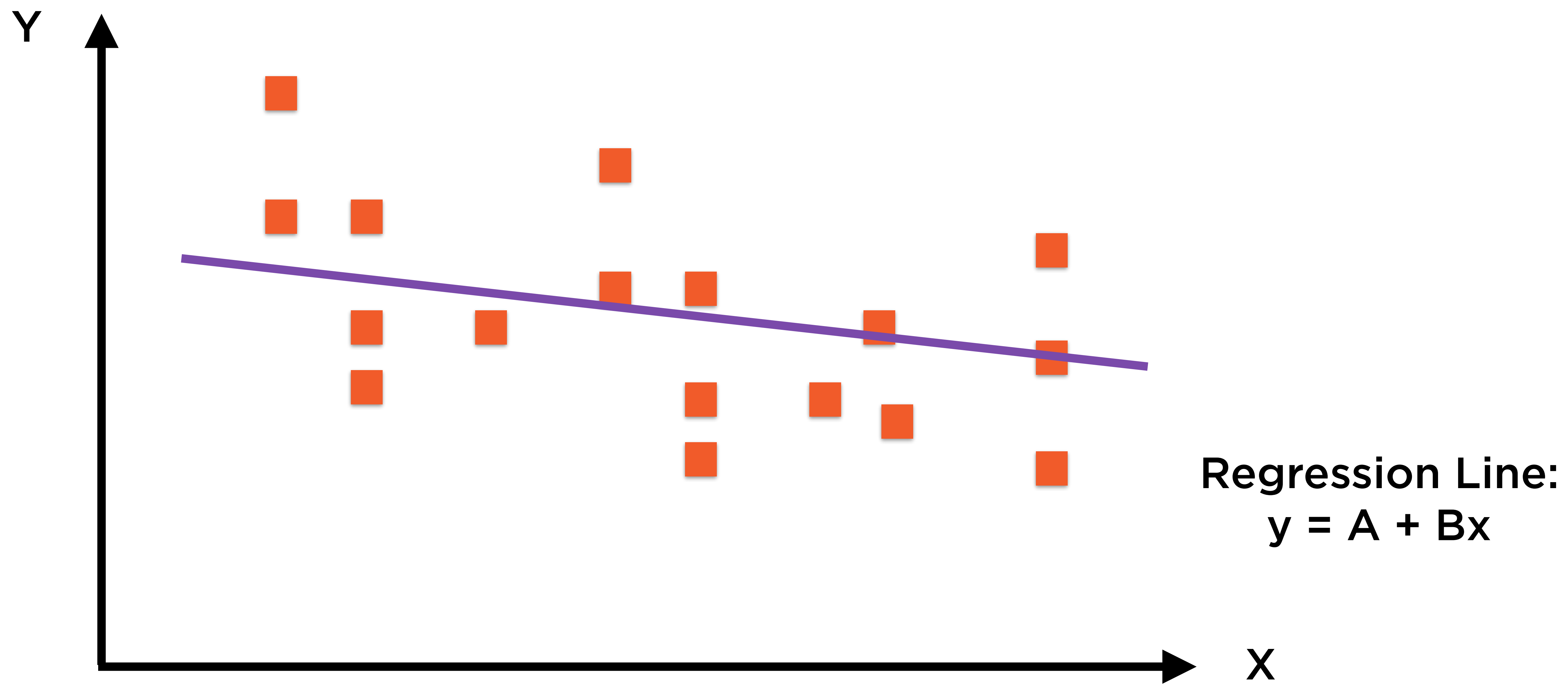
All data points out there in the universe

Sample

A subset of the population

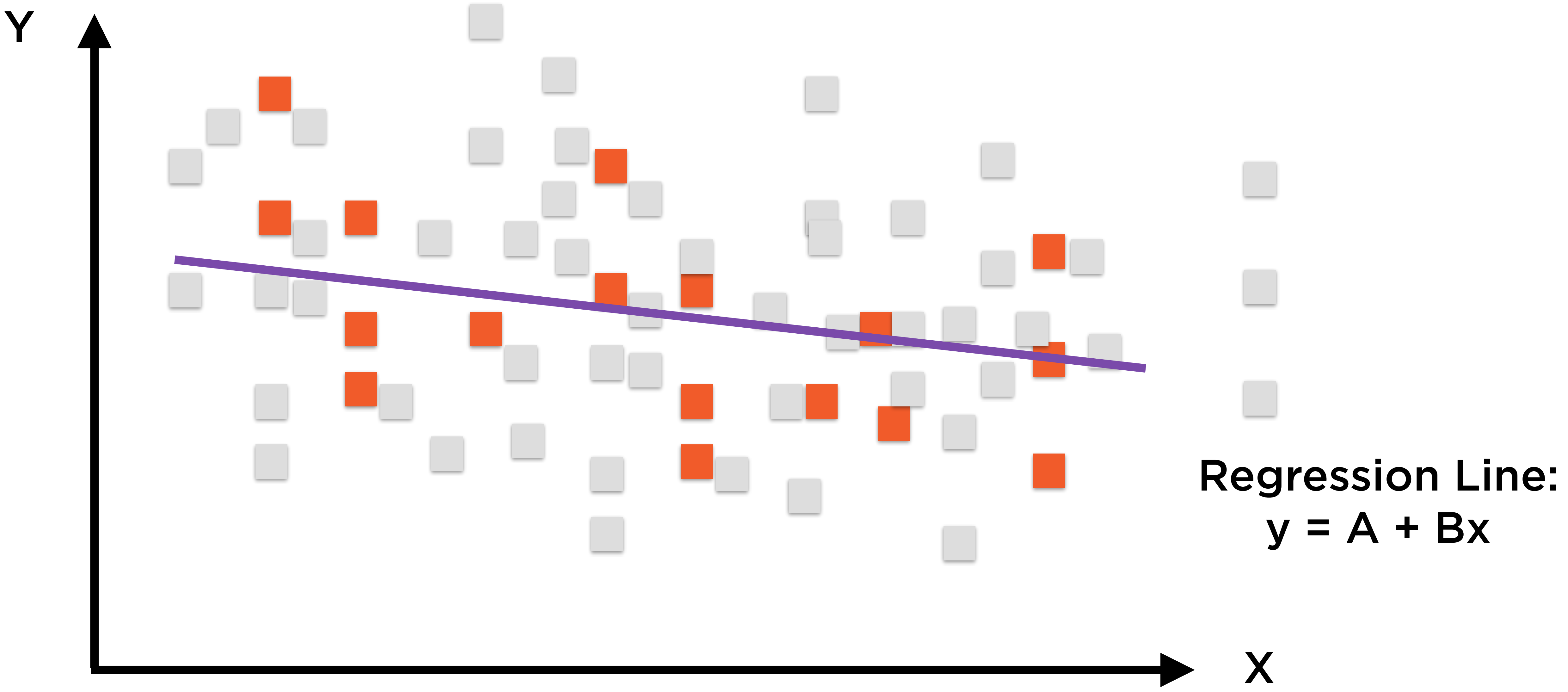
Representative Samples

Regression Works on Samples



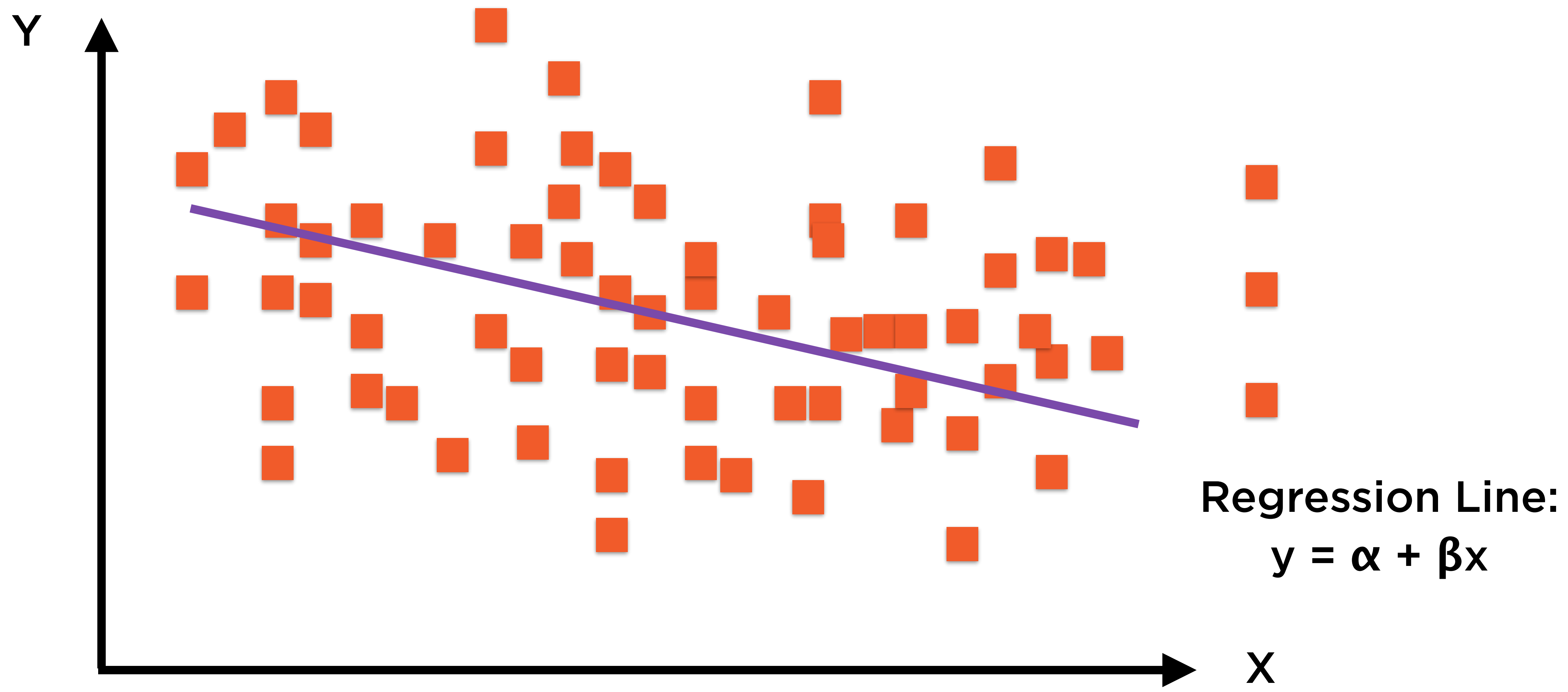
The regression line is based on a sample, not on the population

Regression Works on Samples



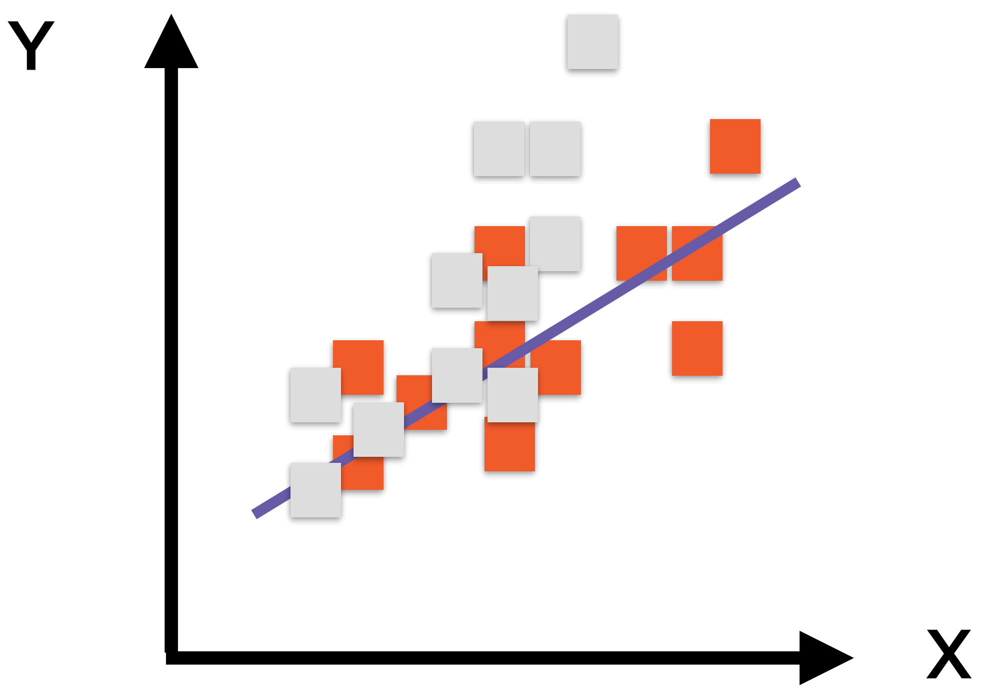
The regression line is based on a sample, not on the population

Regression Works on Samples



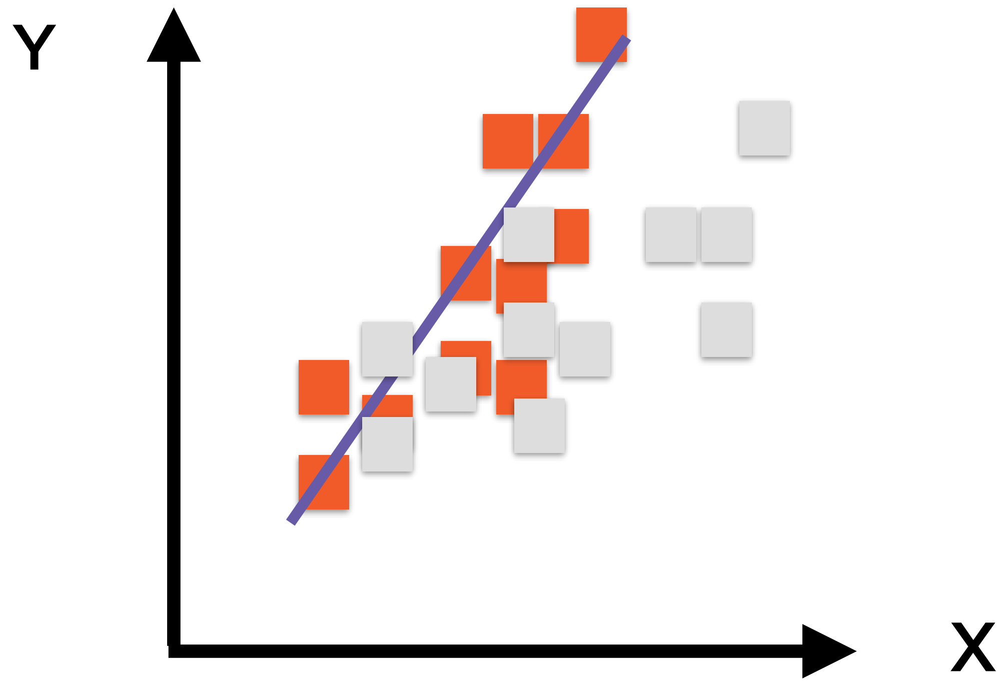
The regression line is based on a sample, not on the population

Different Samples, Different Fits



Sample 1

$$y = A_1 + B_1 x$$

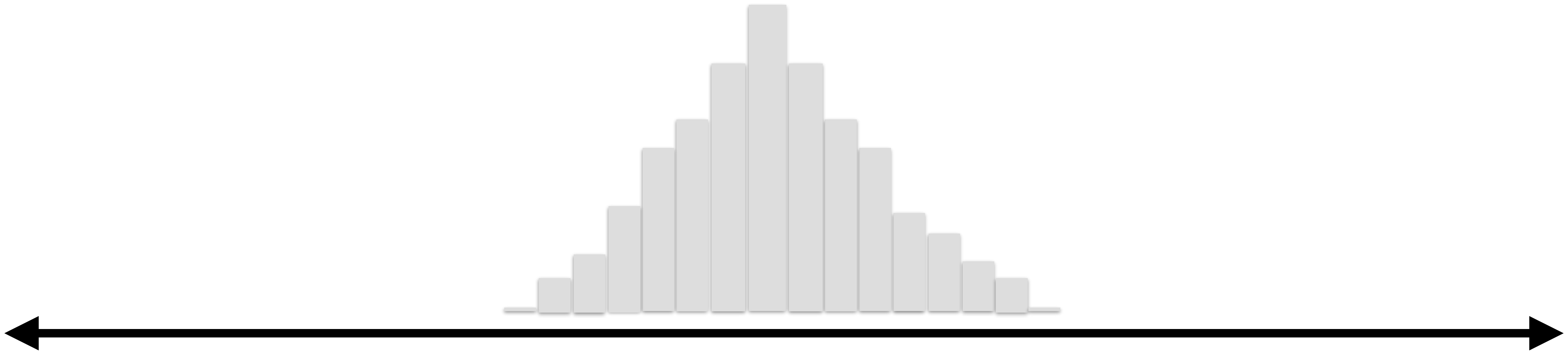


Sample 2

$$y = A_2 + B_2 x$$

Conducting regression on different samples will yield different values of A and B

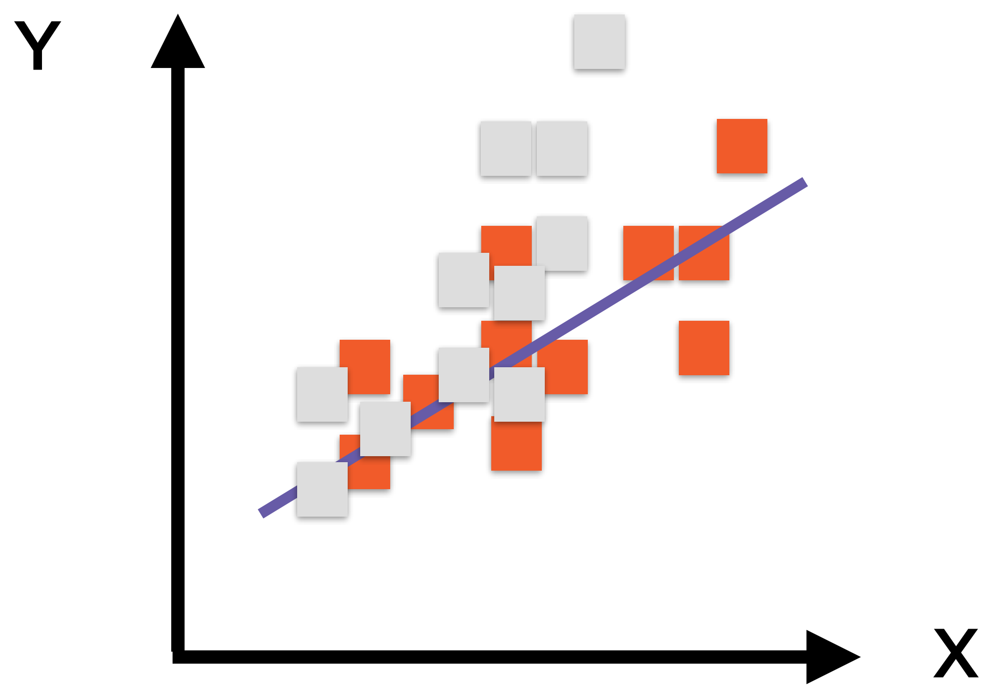
Sampling Distributions



Plotting A (or B) from millions of samples yields a bell curve

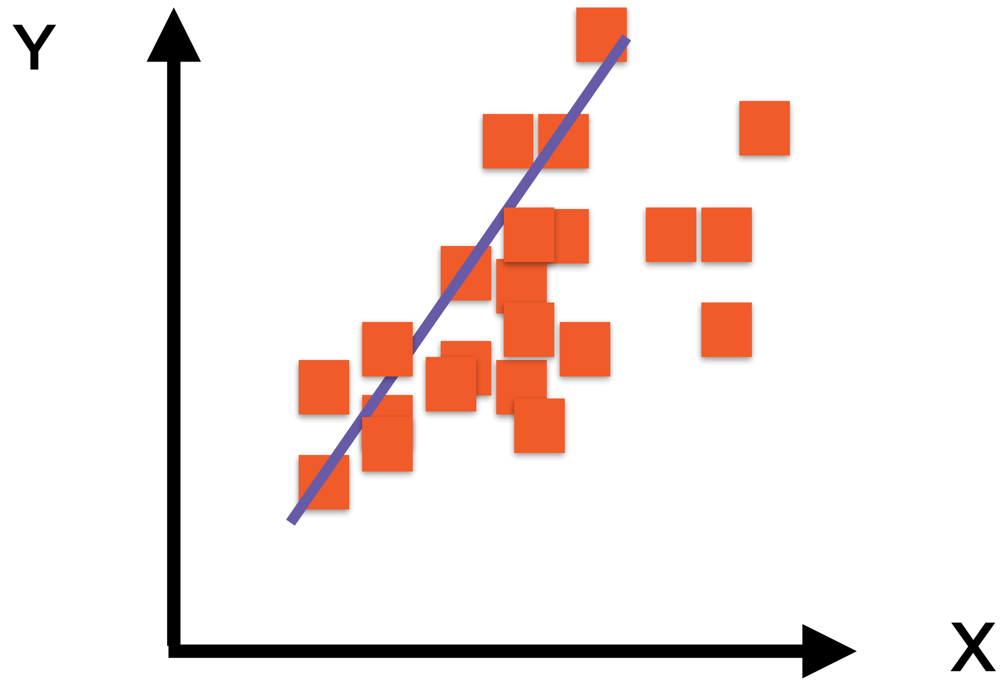
This is known as the **sampling distribution**

Different Samples, Different Fits



Sample Regression Line

$$y = A + Bx$$

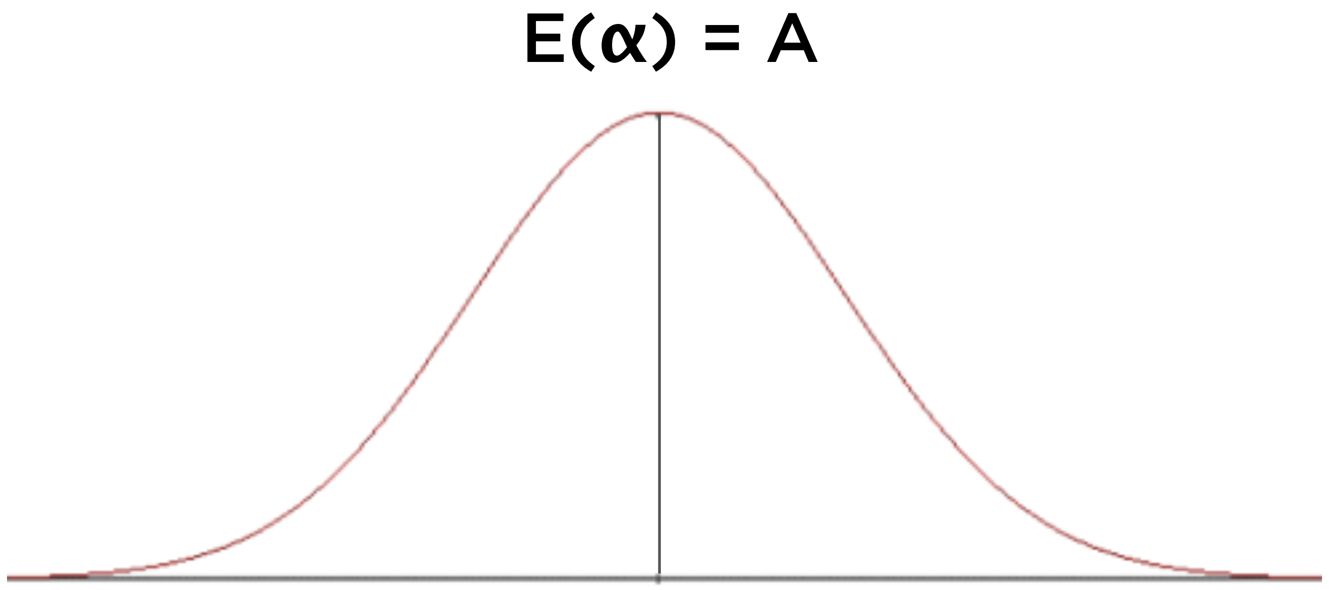


Population Regression Line

$$y = \alpha + \beta x$$

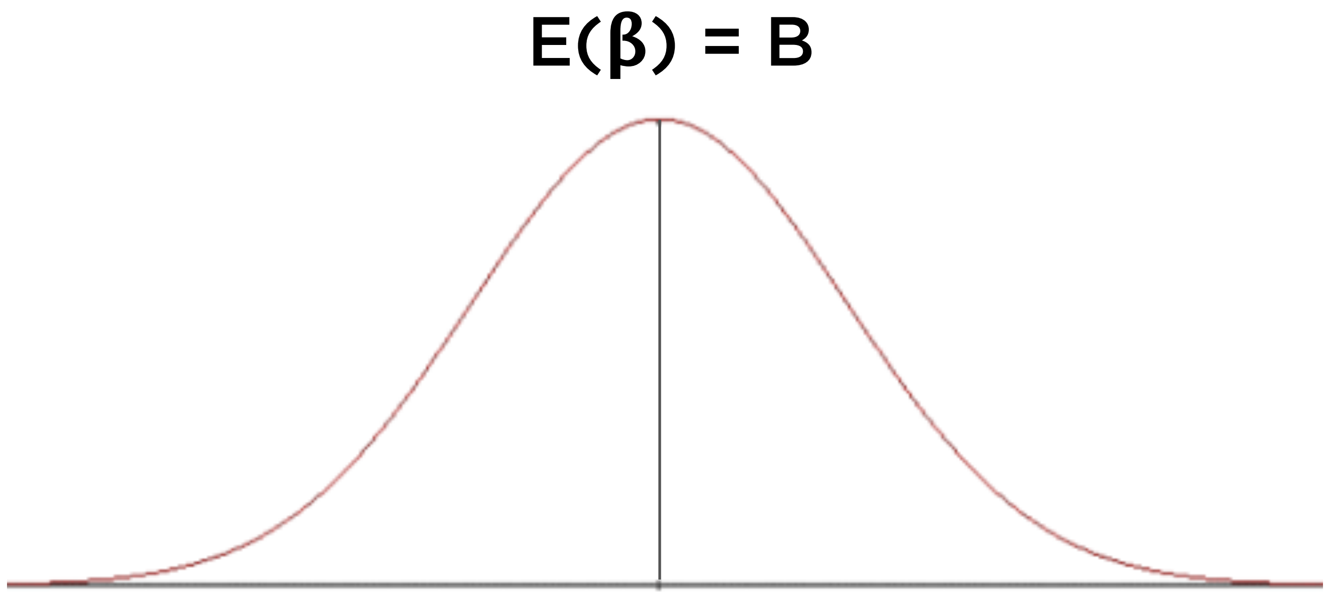
We will never know the values of the population parameters α and β

Sampling Distributions



Sampling Distribution of A

α is the population parameter, A is the sample parameter

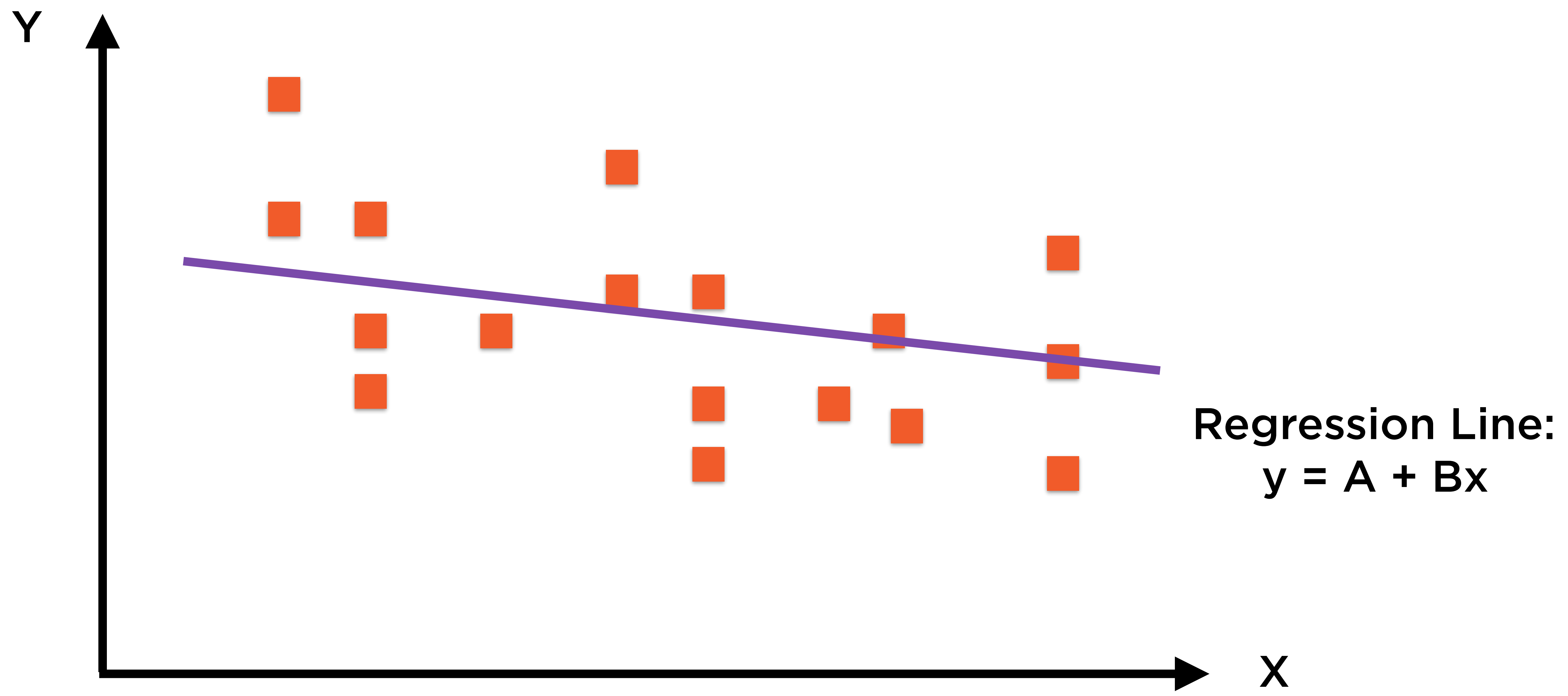


Sampling Distribution of B

β is the population parameter, B is the sample parameter

The sampling distributions are normal, and population mean is equal to sample mean

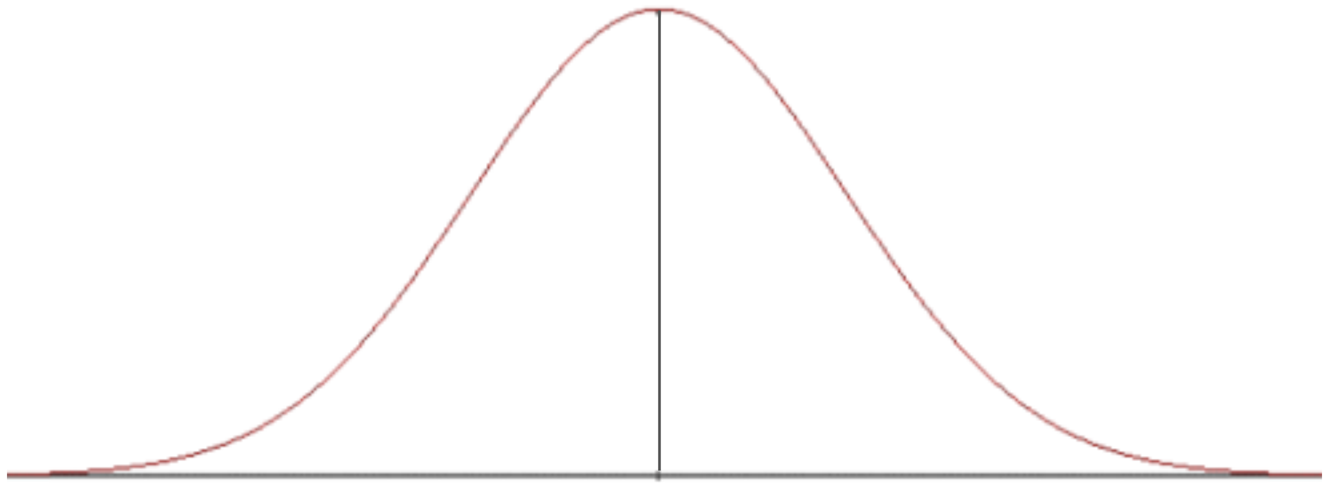
Regression Works on Samples



The sample parameters A and B are our 'best' estimates for population parameters α and β

Sampling Distributions

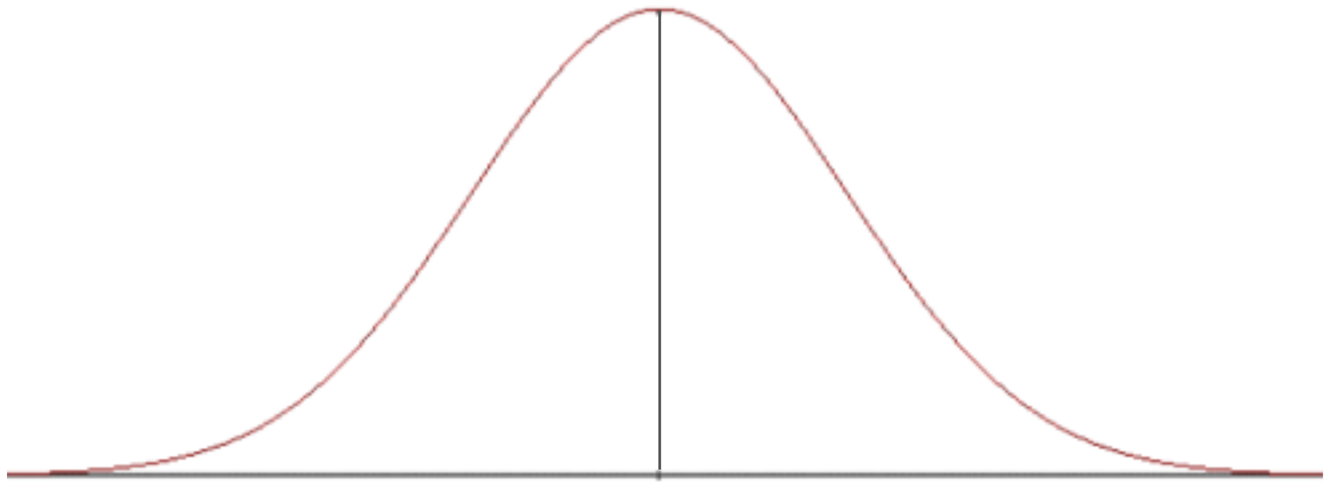
$E(\alpha) = A$



Sampling Distribution of A

α is the population parameter, A is the sample parameter

$E(\beta) = B$

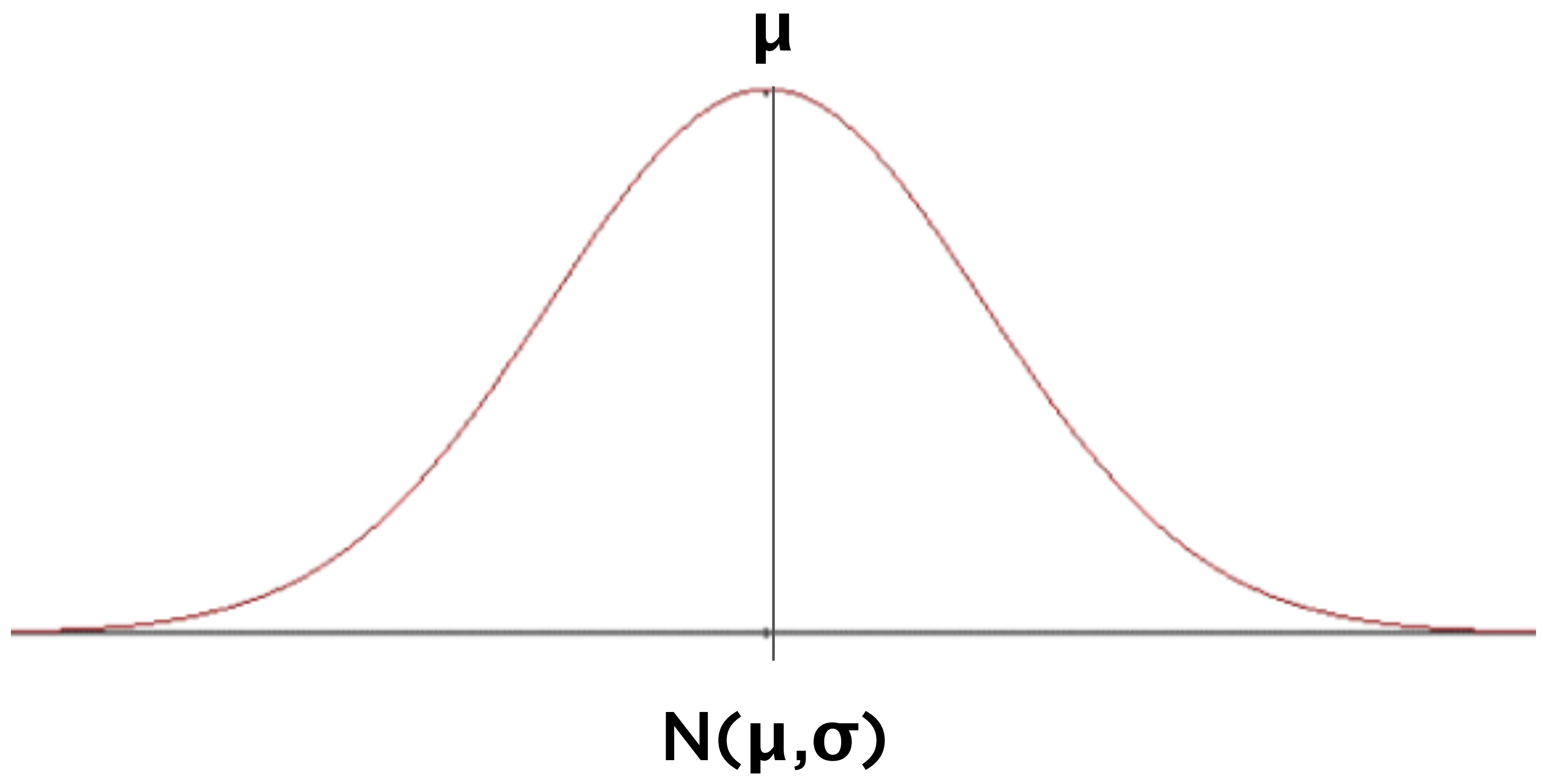


Sampling Distribution of B

β is the population parameter, B is the sample parameter

The sampling distributions are normal, and population mean is equal to sample mean

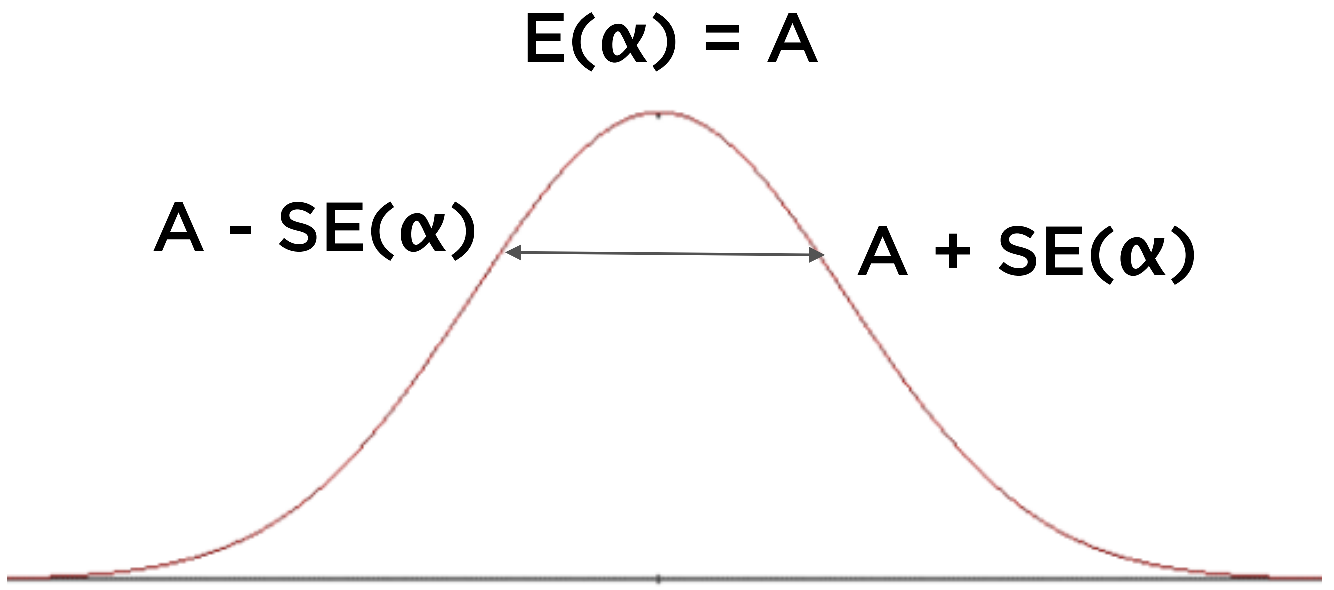
Normal Distribution



Average (mean) is μ

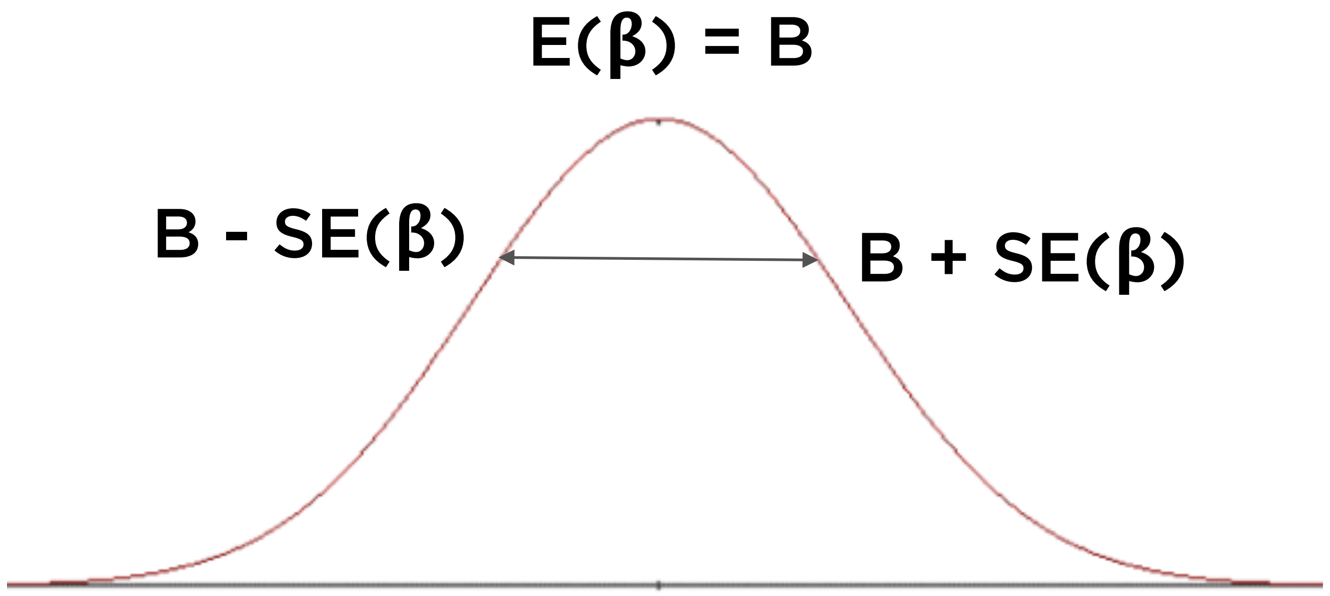
Standard deviation is σ

Standard Errors



Standard Error of A

Standard deviation of the sampling distribution of A

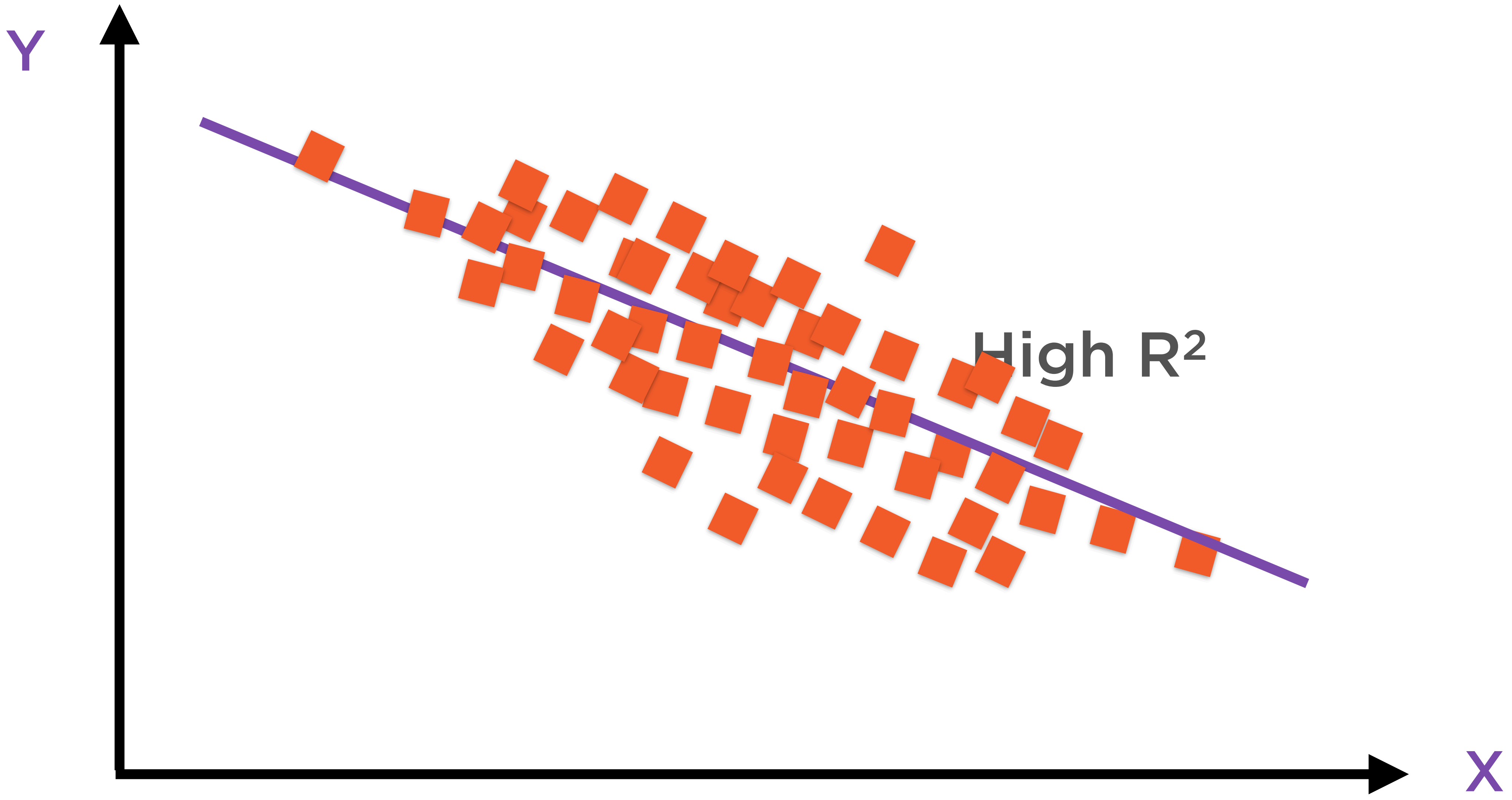


Standard Error of B

Standard deviation of the sampling distribution of A

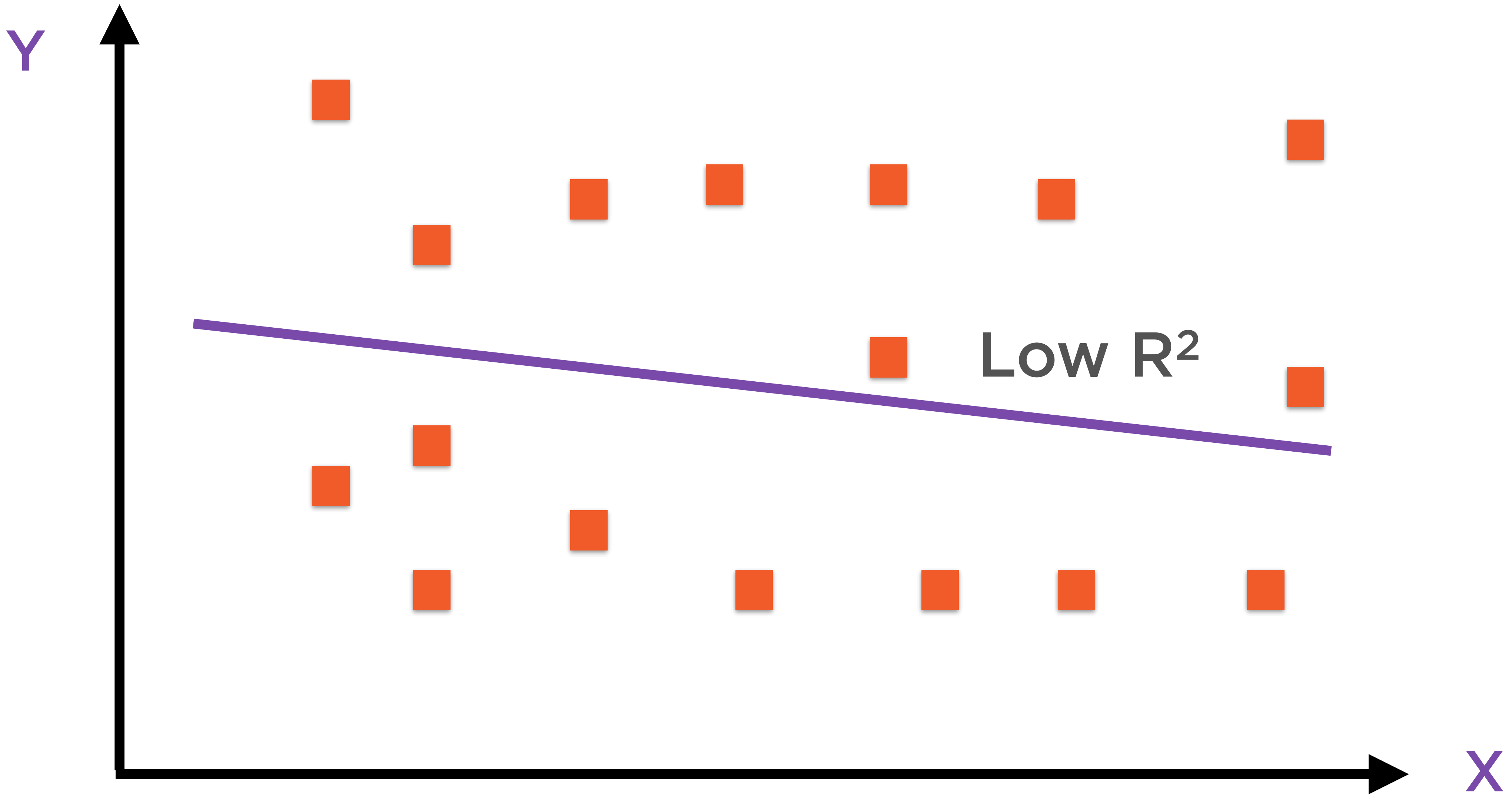
Standard error of a regression parameter is the standard deviation of the sampling distribution

Strong Cause-Effect Relationship



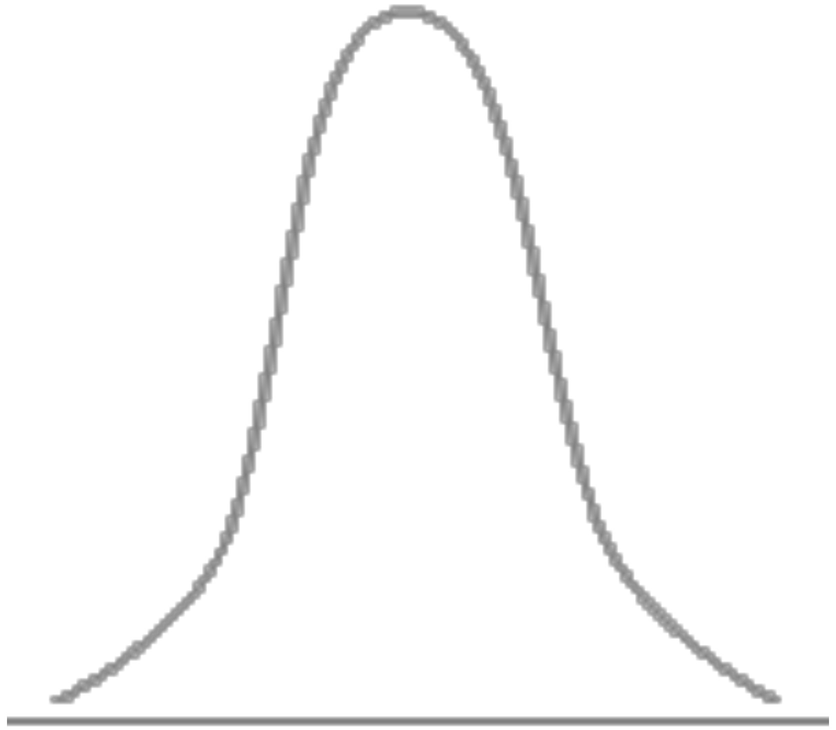
Residuals are small, standard errors are small

Weak Cause-Effect Relationship



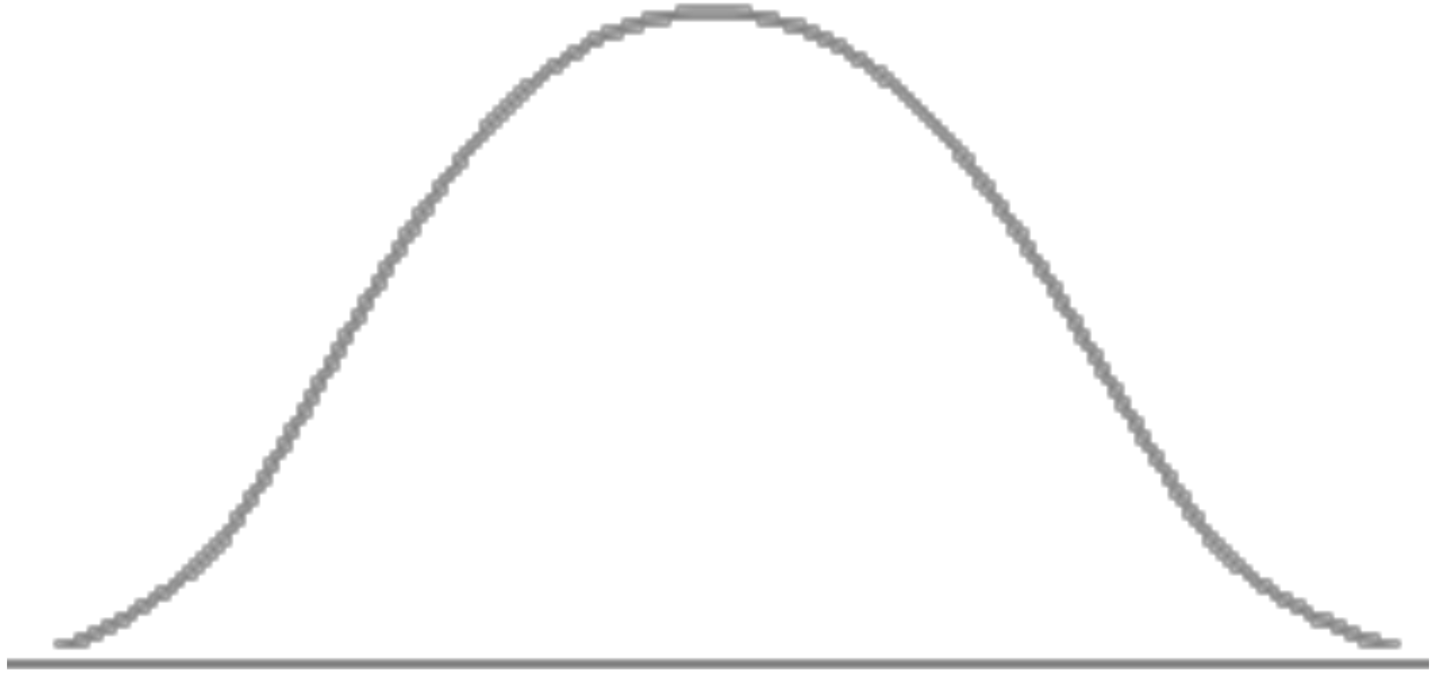
Residuals are large, standard errors are large

Standard Errors and Residuals



Low Standard Error

High confidence that parameter coefficient is well estimated



High Standard Error

Low confidence that parameter coefficient is well estimated

The smaller the residuals, the smaller the standard errors and the better the quality of the regression

Sample Regression Line

Regression Equation:

$$y = A + Bx$$

Residuals

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

RSS = Variance(e)

Residual Variance *(RSS)*

Easily calculated from regression residuals

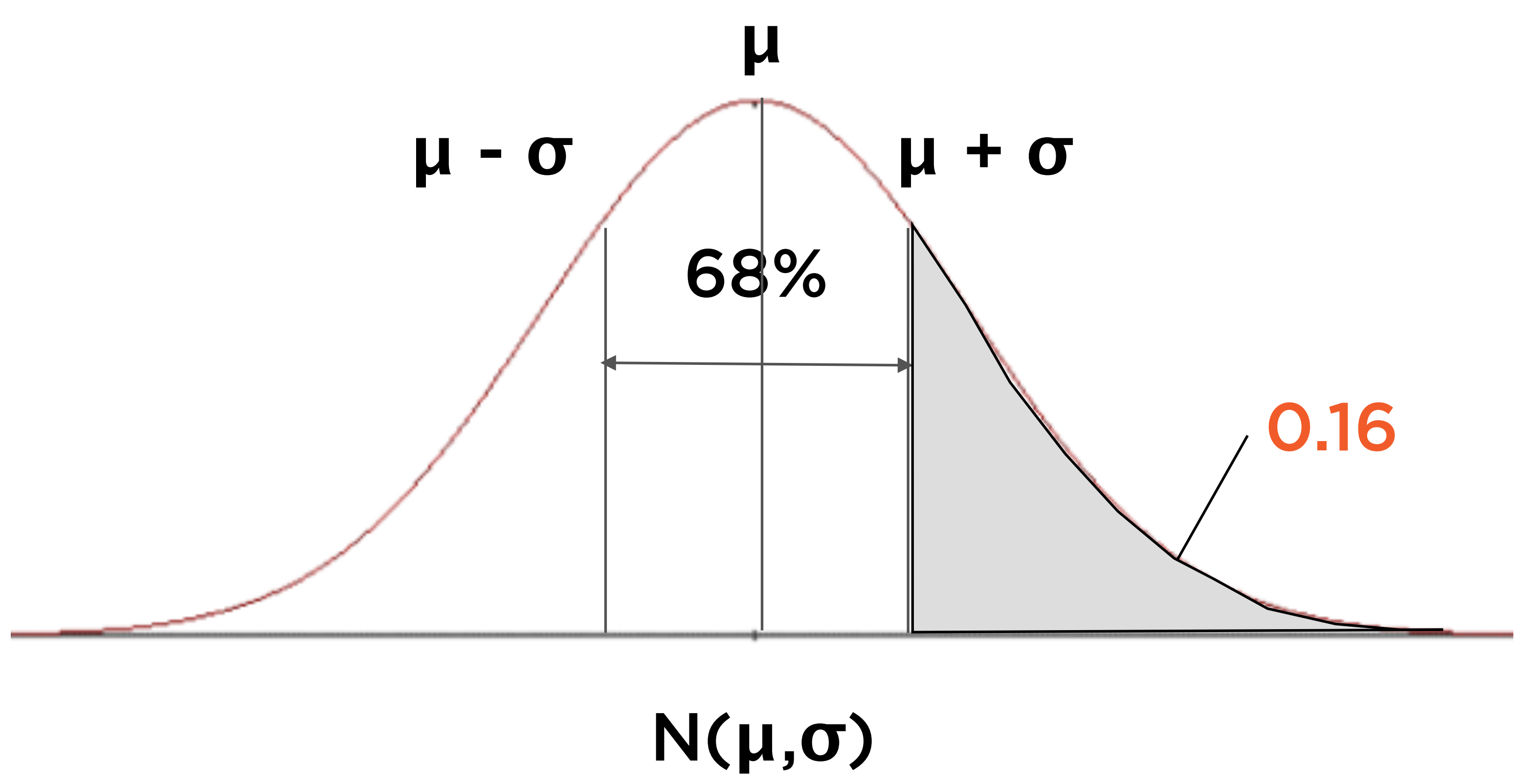
SE(α), SE(β) can be found from RSS

Estimate Standard Errors from RSS

Exact formulae are not important - reported by Excel, R…

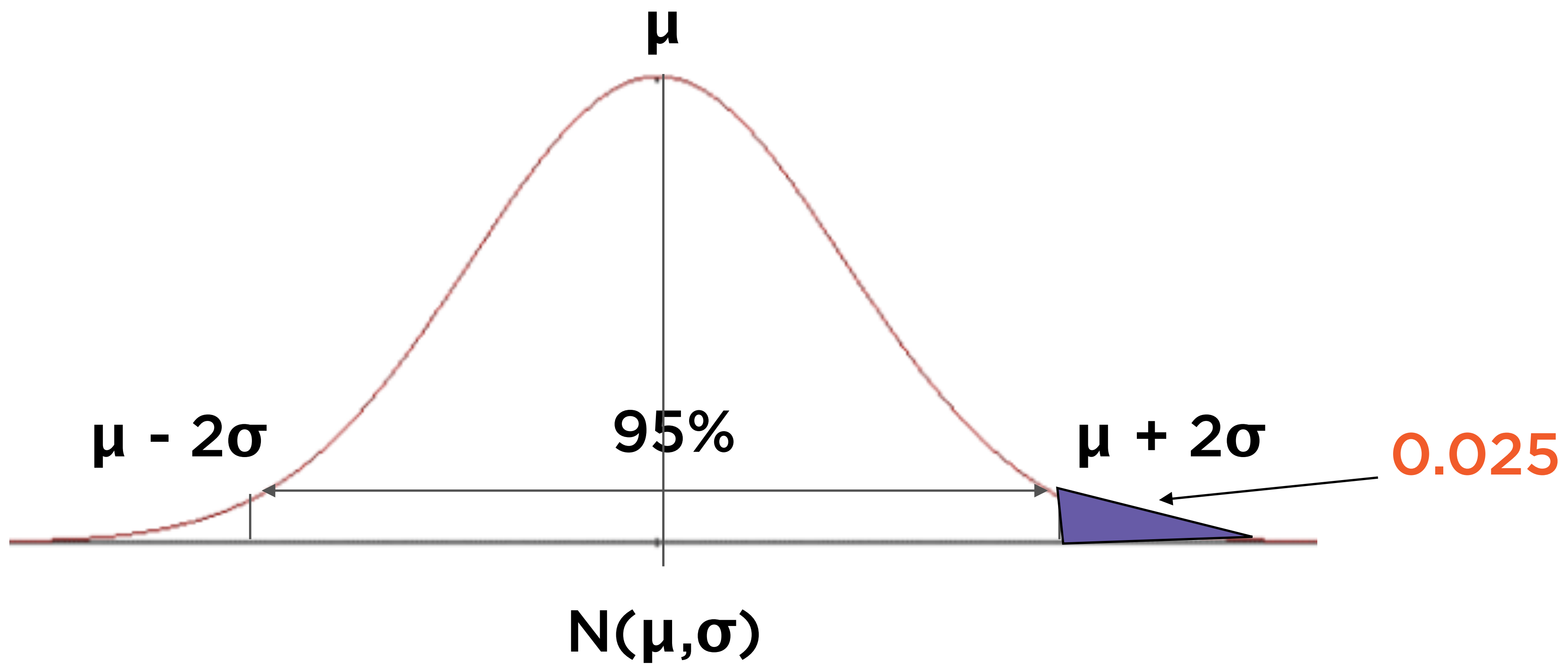
The smaller the residuals, the smaller the standard errors and the better the quality of the regression

Probability of Occurrence



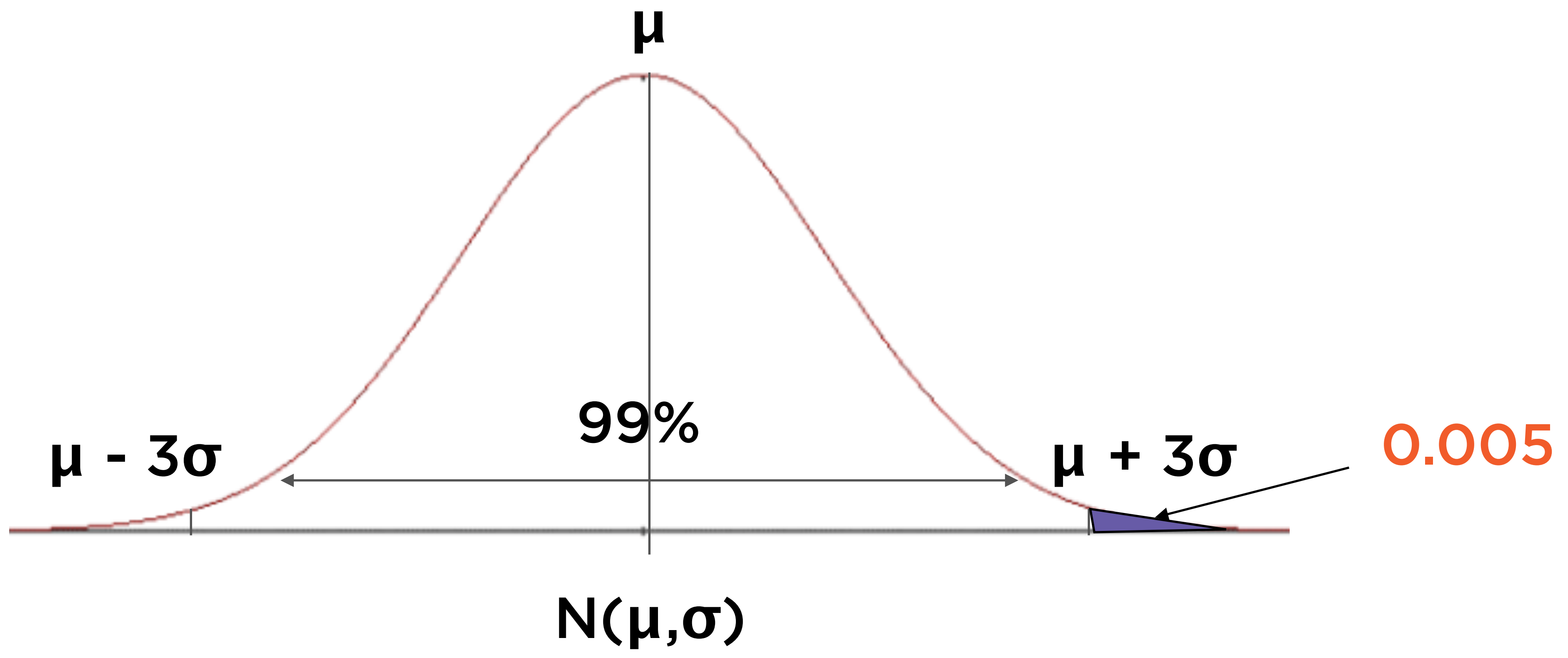
68% within 1 standard deviation of mean

Probability of Occurrence



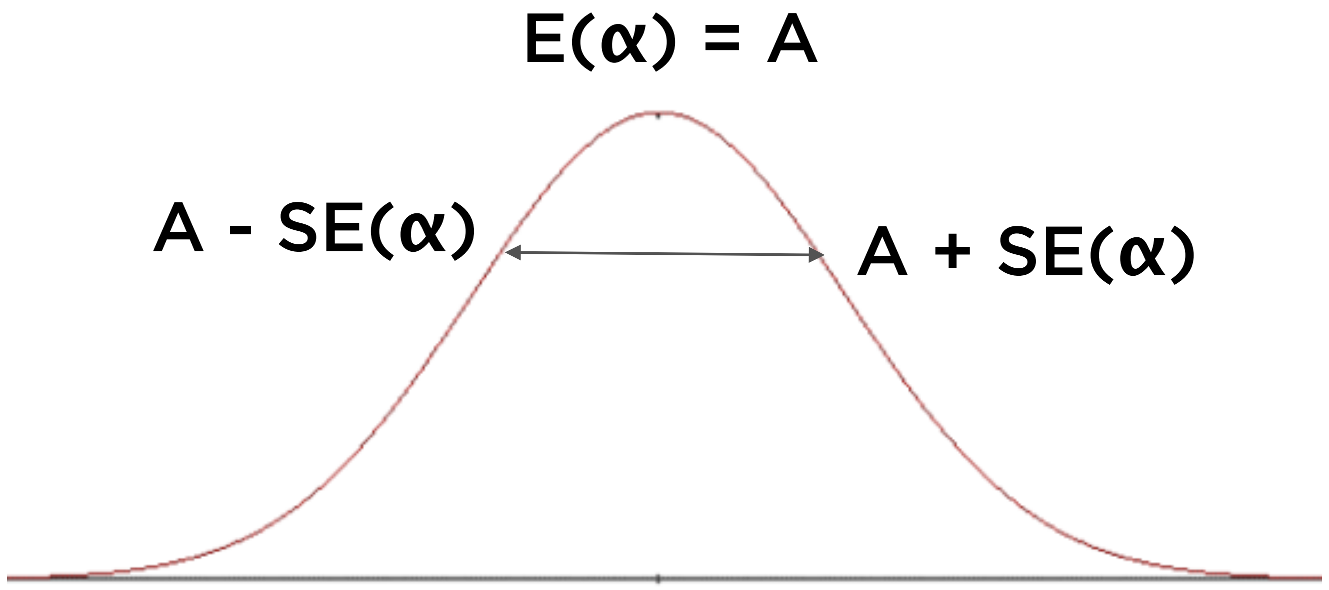
95% within 2 standard deviations of mean

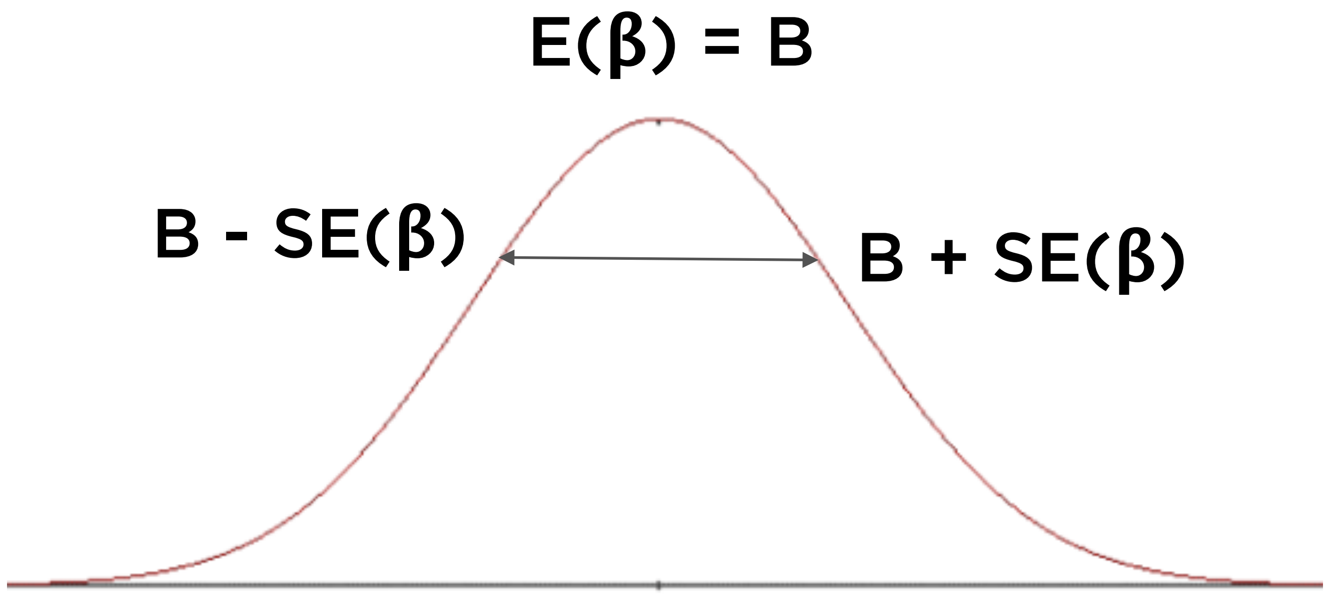
Probability of Occurrence



99% within 3 standard deviations of mean

Standard Errors





Standard Error of A

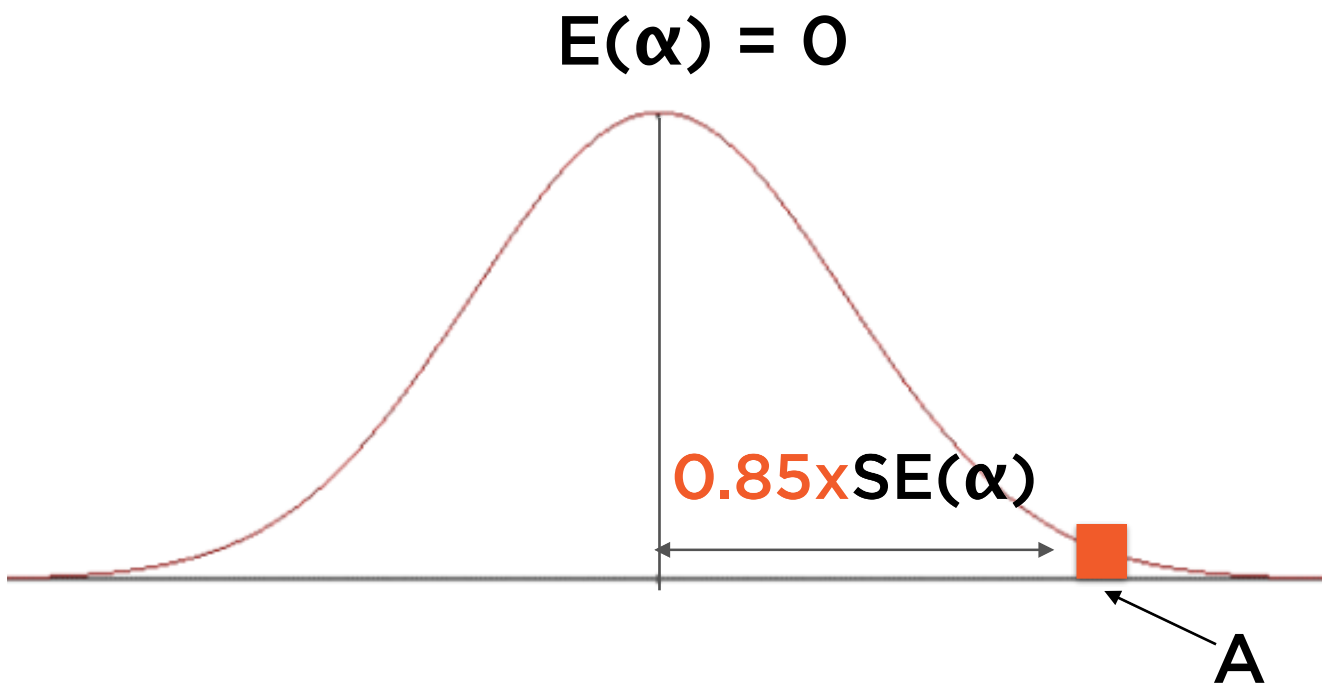
Standard deviation of the sampling distribution of A

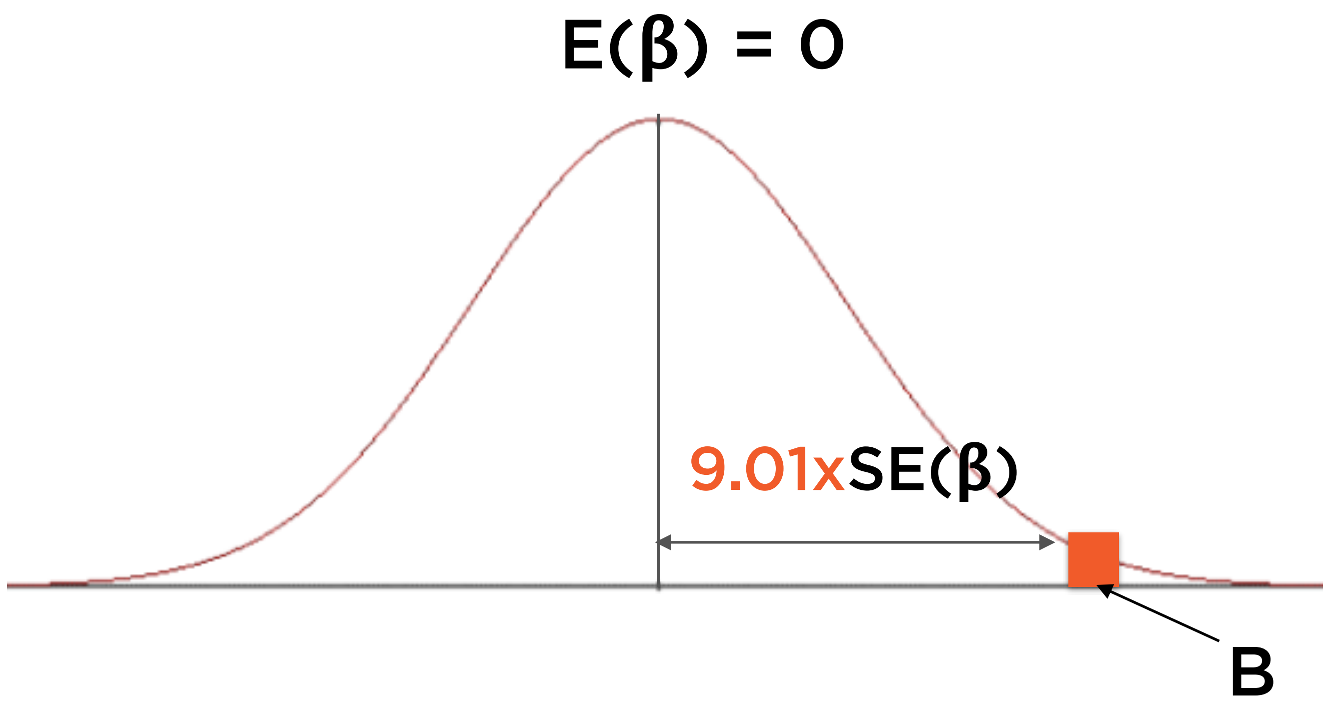
Standard Error of B

Standard deviation of the sampling distribution of A

Standard error of a regression parameter is the standard deviation of the sampling distribution

t-Statistics





t-stat(A) = 0.85

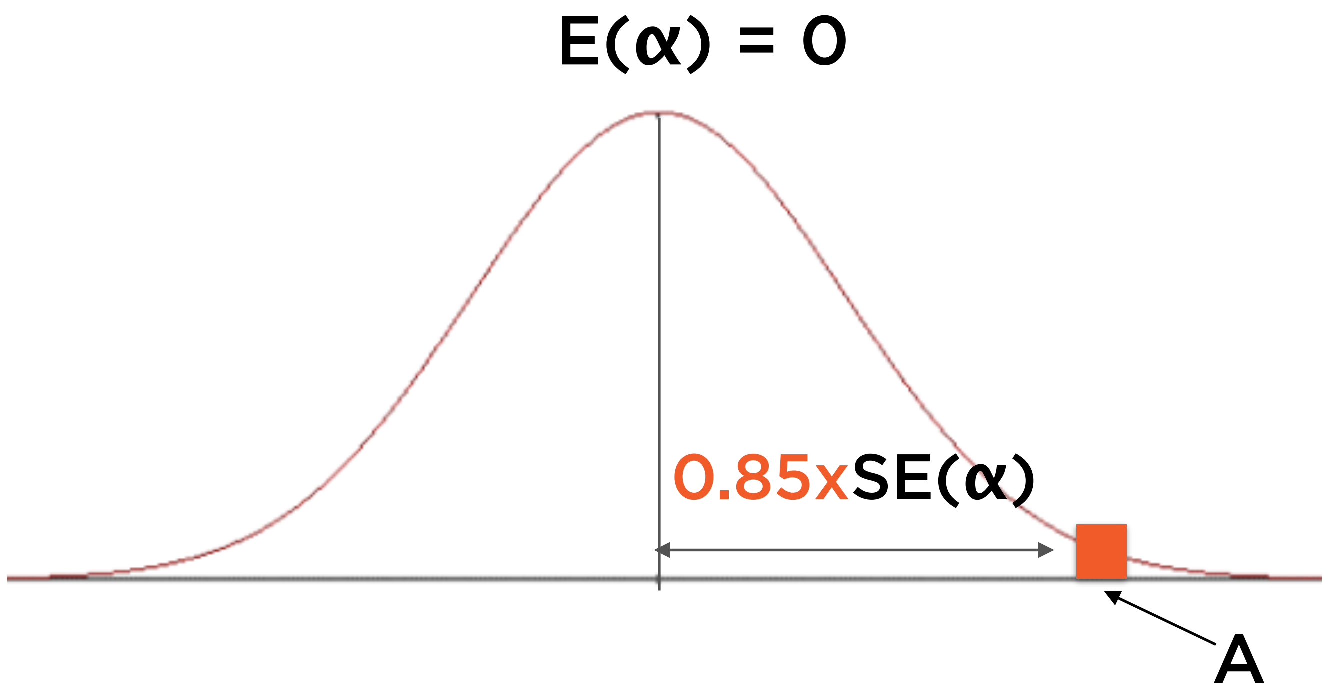
t-stat(A) = A/SE(α)

t-stat(B) = 9.01

t-stat(B) = B/SE(β)

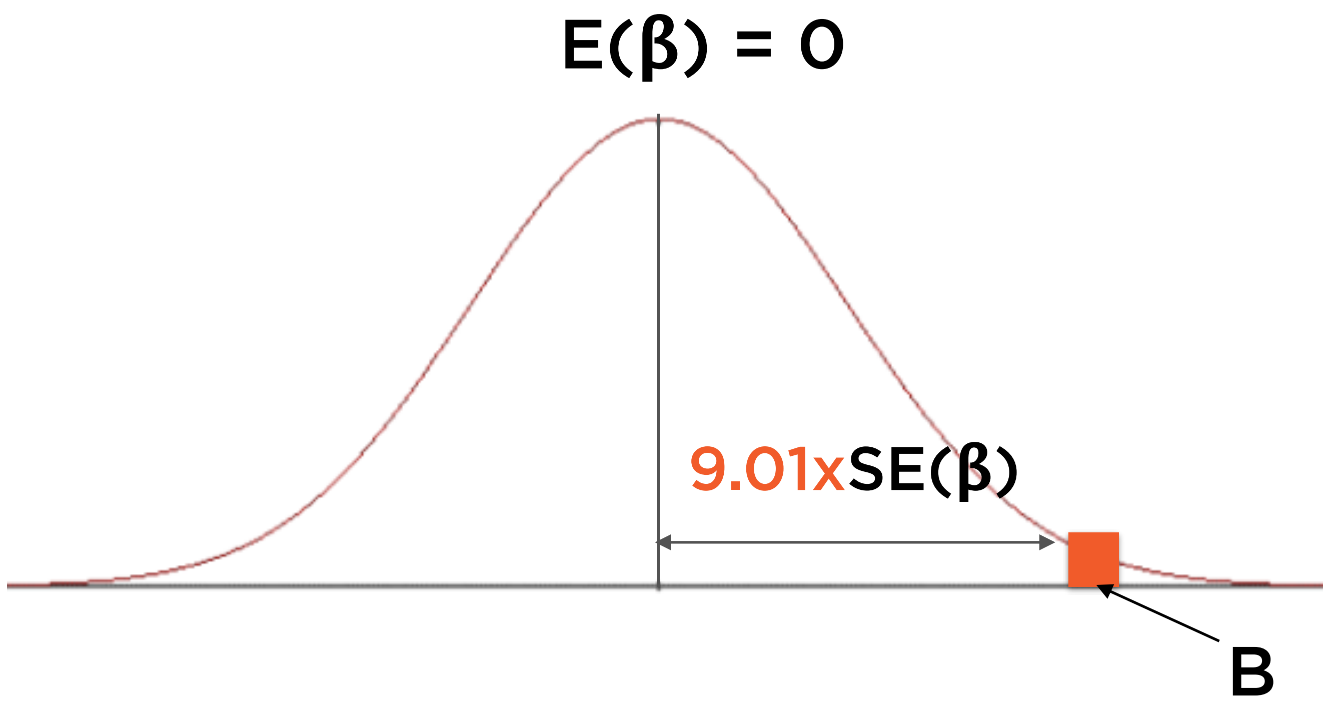
The probability distributions here are not normal, rather they follow a t-distribution

t-Statistics



t-stat(A) = 0.85

t-stat(A) = A/SE(α)



t-stat(B) = 9.01

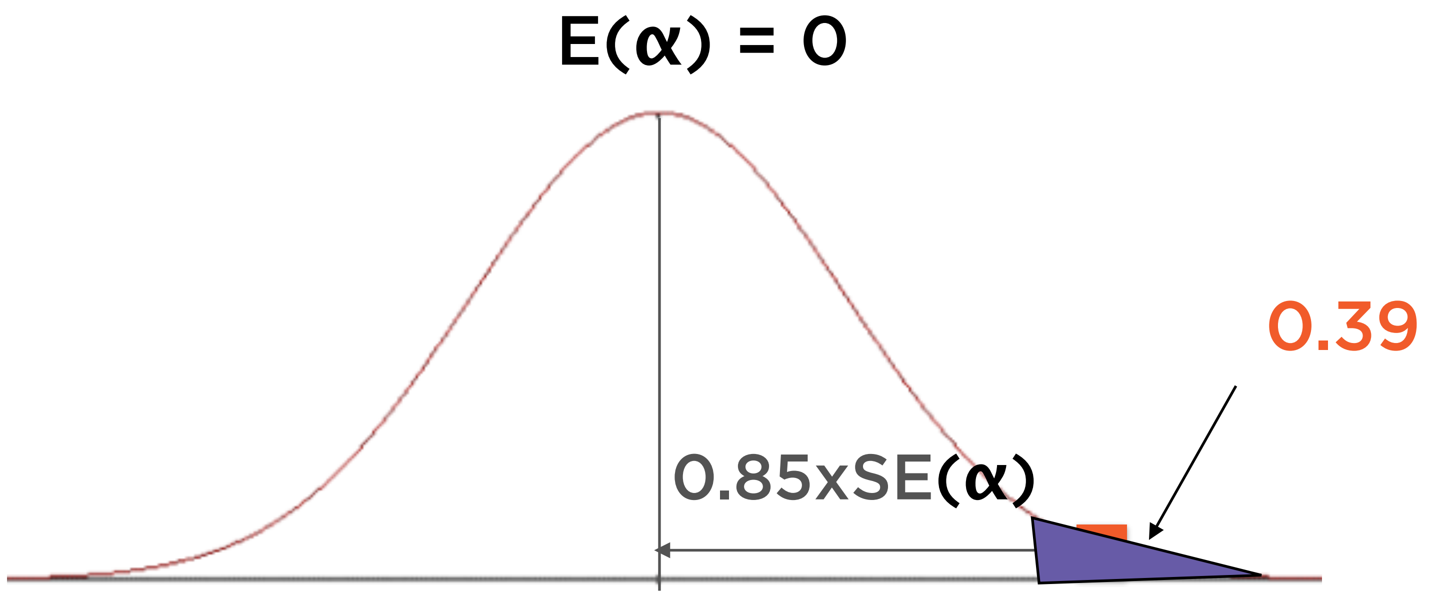
t-stat(B) = B/SE(β)

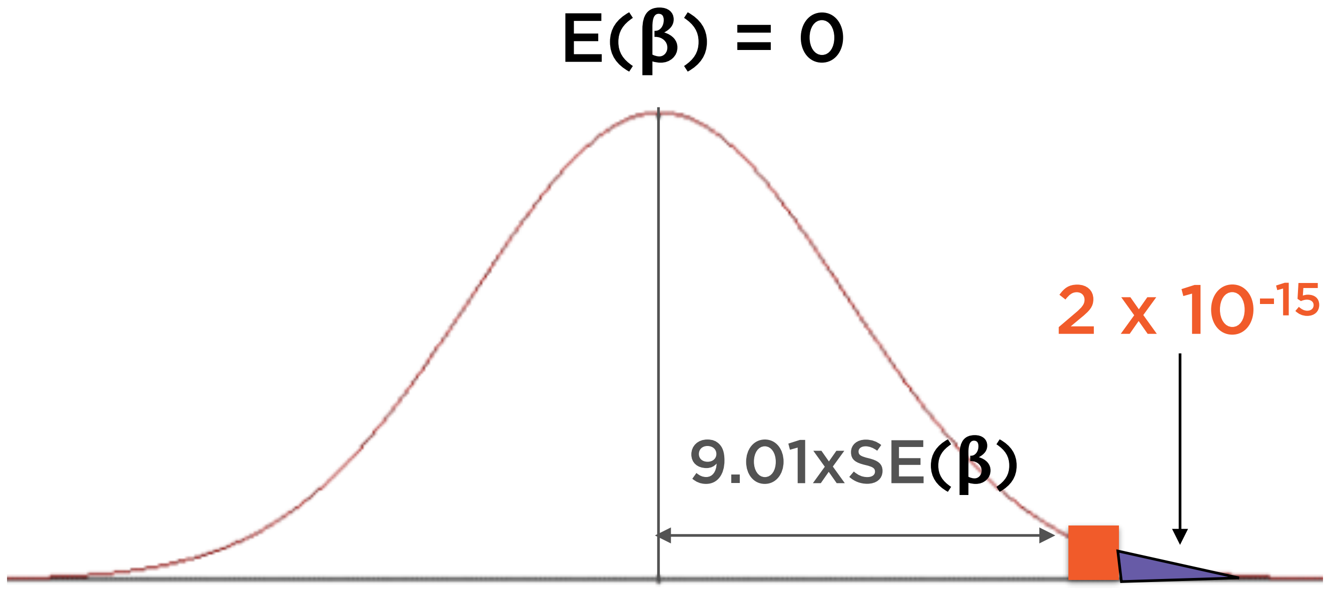
Is an individual estimate of A or B ‘adding value’ at all?

High t-statistic => Yes

The higher the t-statistic of a coefficient, the higher our confidence in our estimate of that coefficient

p-Values





p-value(A) = 0.39

Low t-stat, high p-value

p-value(B) = $2 \times 10^{-15} \sim 0$

High t-stat, low p-value

Is an individual estimate of A or B 'adding value' at all?

low p-value => Yes

The lower the p-value of a coefficient, the higher our confidence in our estimate of that coefficient

Interpreting Results of a Multiple Regression

Adjusted R^2

Residuals

F-statistic

Standard Errors of coefficients

R^2

Interpreting Results of a Multiple Regression

Adjusted R^2

Residuals

F-statistic

Standard Errors of coefficients

R^2

Sample Regression Line

Regression Equation:

$$y = A + Bx$$

Residuals

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

RSS = Variance(e)

Residual Variance *(RSS)*

Easily calculated from regression residuals

Population Regression Line

Regression Equation:

$$y = \alpha + \beta x$$

Errors

$$
\begin{aligned}
y_1 &= \alpha + \beta x_1 + \varepsilon_1 \\
y_2 &= \alpha + \beta x_2 + \varepsilon_2 \\
y_3 &= \alpha + \beta x_3 + \varepsilon_3 \\
\ldots & \qquad \ldots \\
y_n &= \alpha + \beta x_n + \varepsilon_n
\end{aligned}
$$

σ^2 = Variance(ε)

Error Variance

Can not be calculated - like all population parameters, can only be estimated from sample

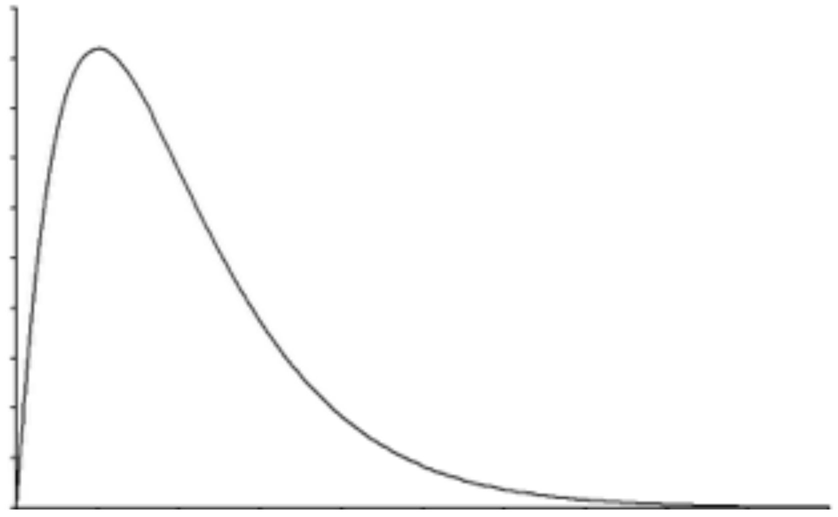
$$SER = \sqrt{\frac{RSS}{n-2}}$$

Standard Error of Regression *(SER)*

n is the number of points in the regression.

SER provides an unbiased estimator of error variance σ^2

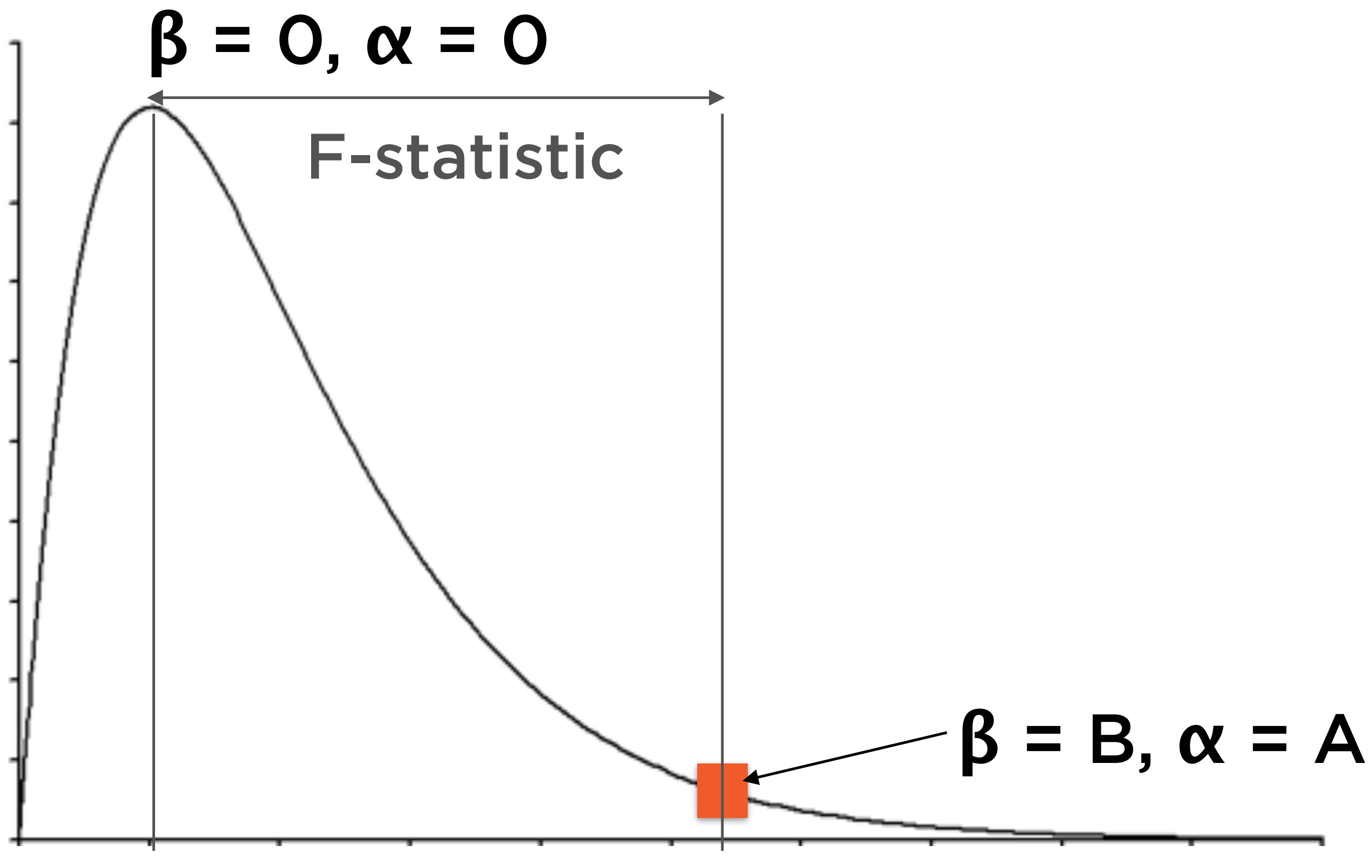
$$\frac{RSS}{\sigma^2} \sim \chi^2$$



χ^2 Distribution with n-2 Degrees of Freedom

Easily calculated from regression residuals

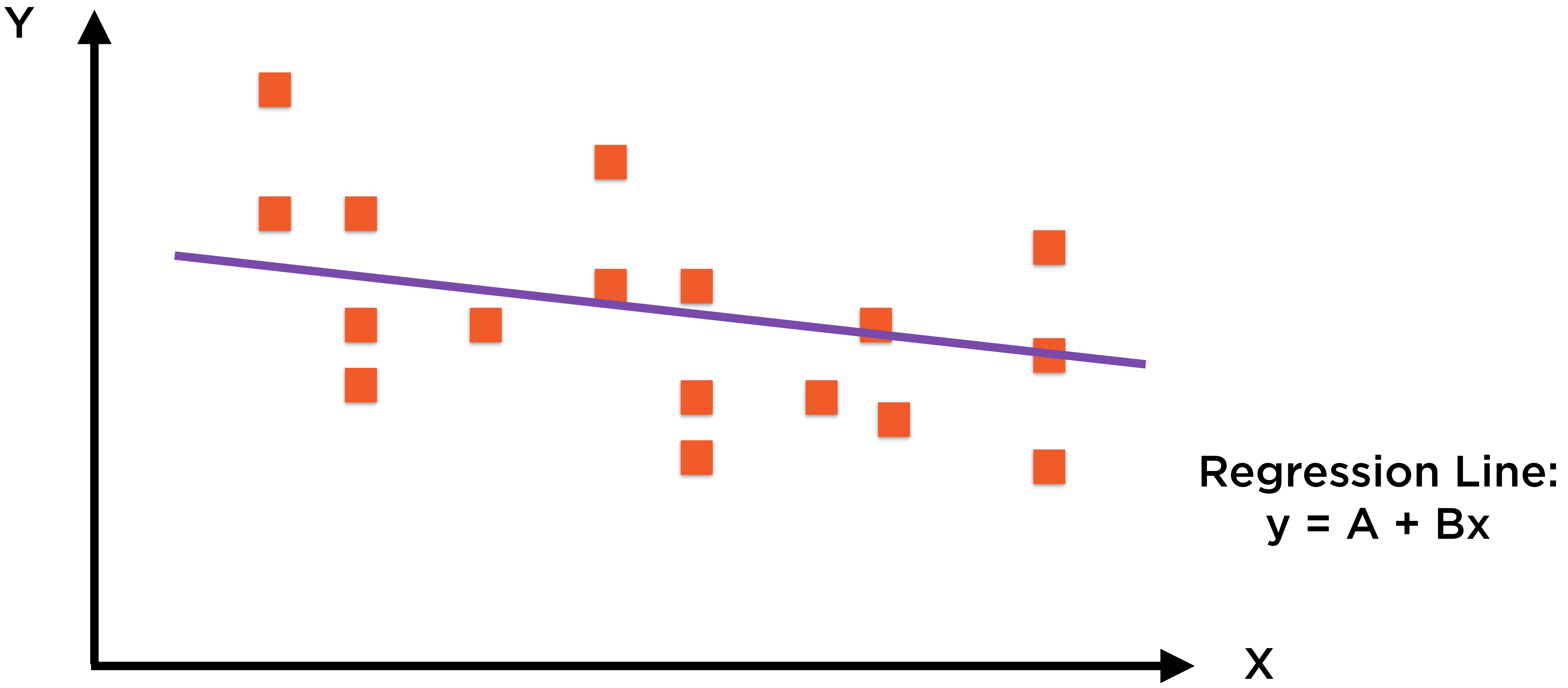
F-Statistic



Does our regression as a whole 'add value' at all?

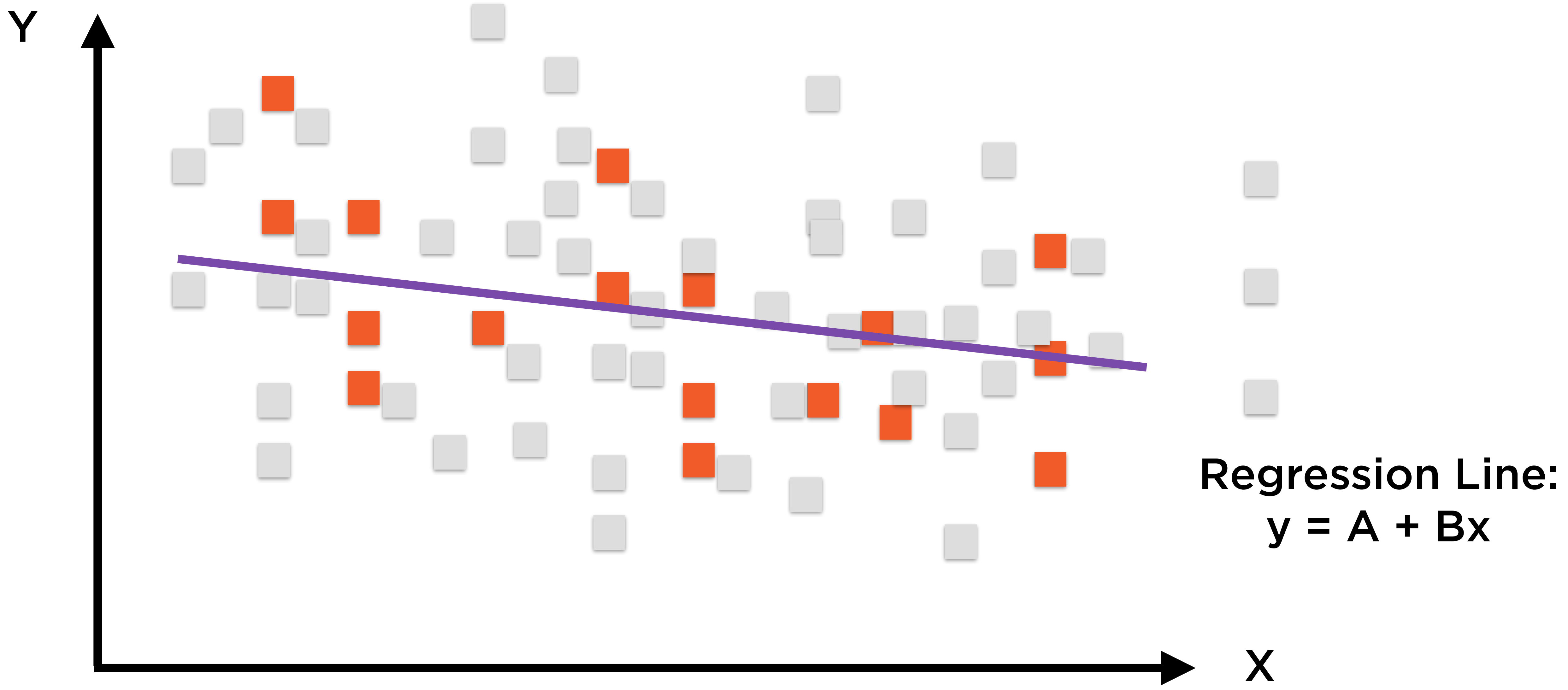
High F-statistic => Yes

Regression Works on Samples



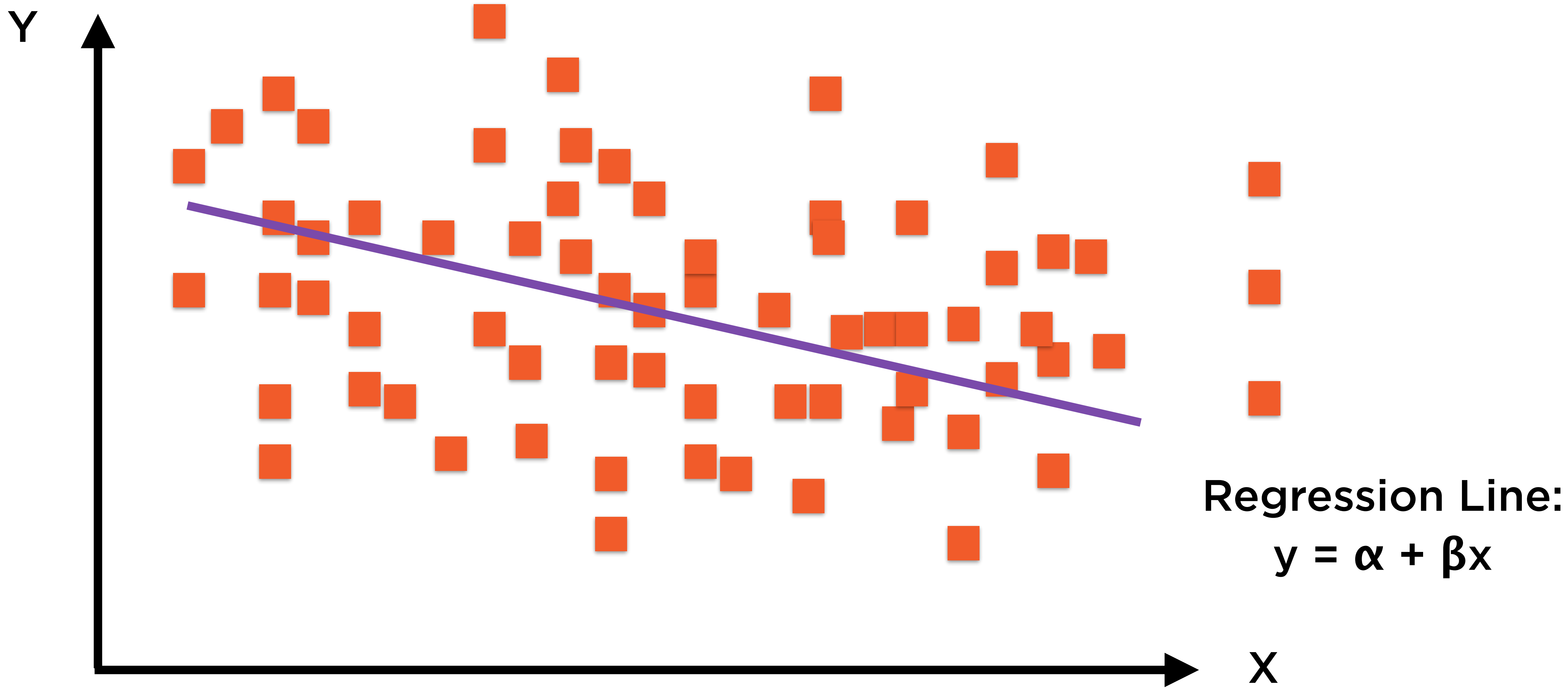
The regression line is based on a sample, not on the population

Regression Works on Samples



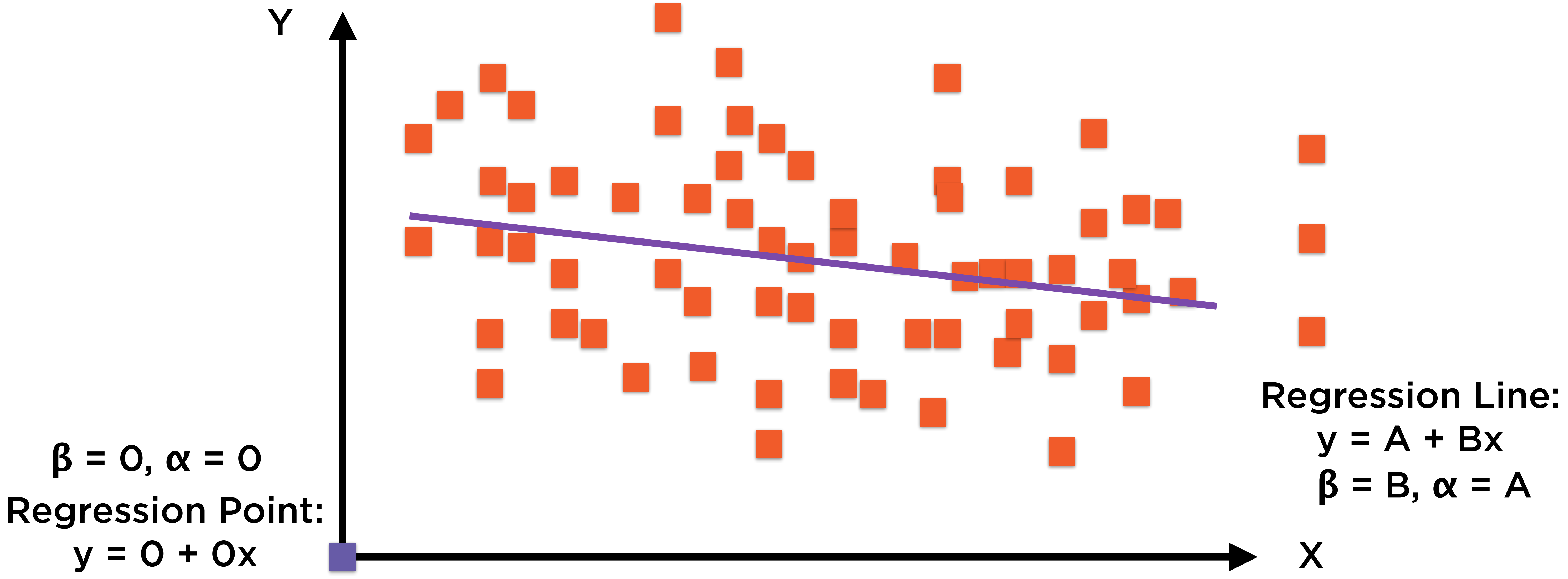
The regression line is based on a sample, not on the population

Regression Works on Samples



The regression line is based on a sample, not on the population

Regression Works on Samples



The regression line is based on a sample, not on the population

p-values and t-statistics tell us whether individual parameter coefficients are ‘good’

The F-statistic tells us whether a entire regression line is ‘good’

Interpreting Results of a Multiple Regression

Adjusted R^2

Residuals

F-statistic

Standard Errors of coefficients

R^2

Interpreting Results of a Multiple Regression

Adjusted R^2

Residuals

F-statistic

Standard Errors of coefficients

R^2

Demo

Implement multiple regression in R

Extending Multiple Regression to Categorical Variables

A Simple Regression

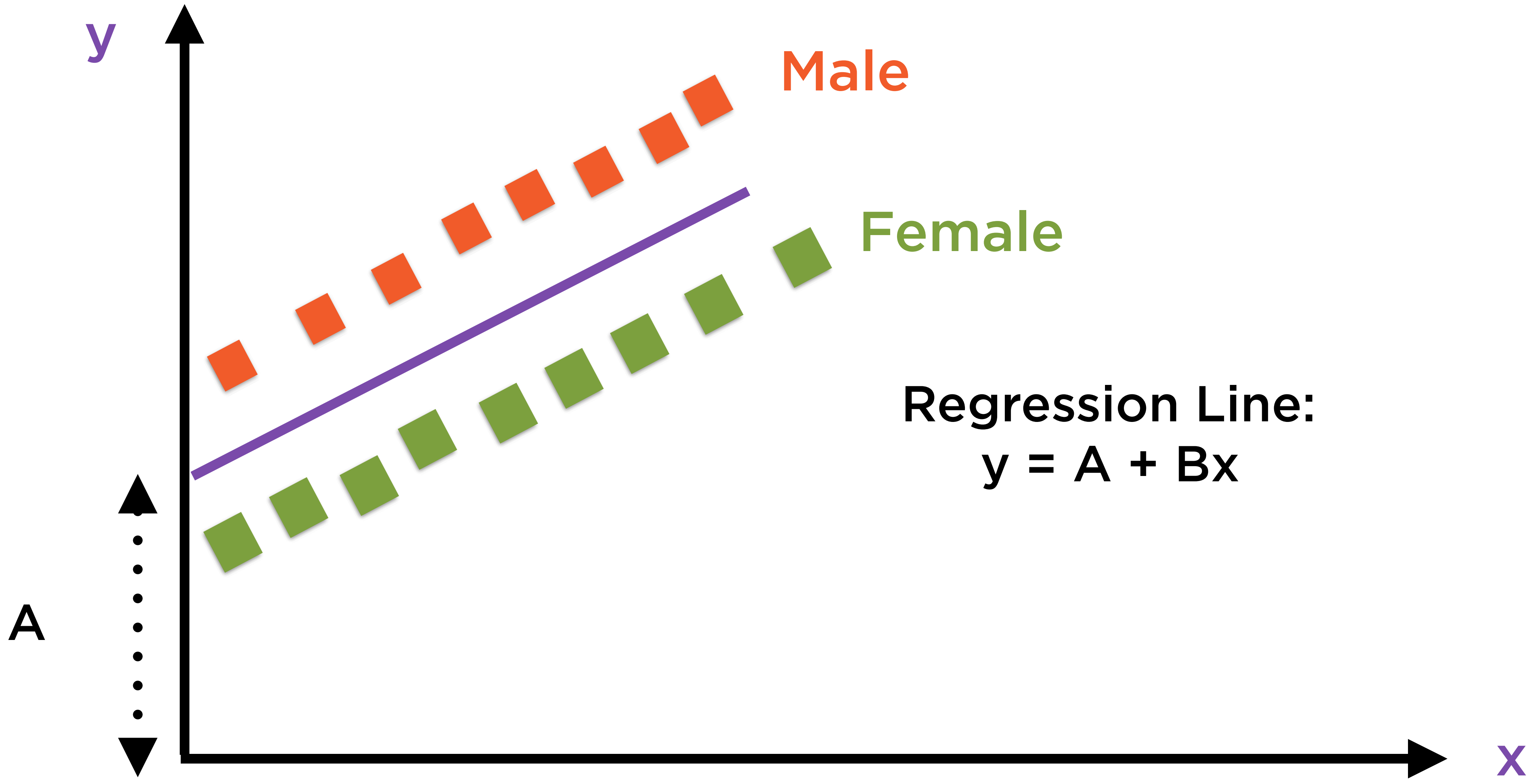
Proposed Regression Equation:

$$y = A + Bx$$

Height of individual

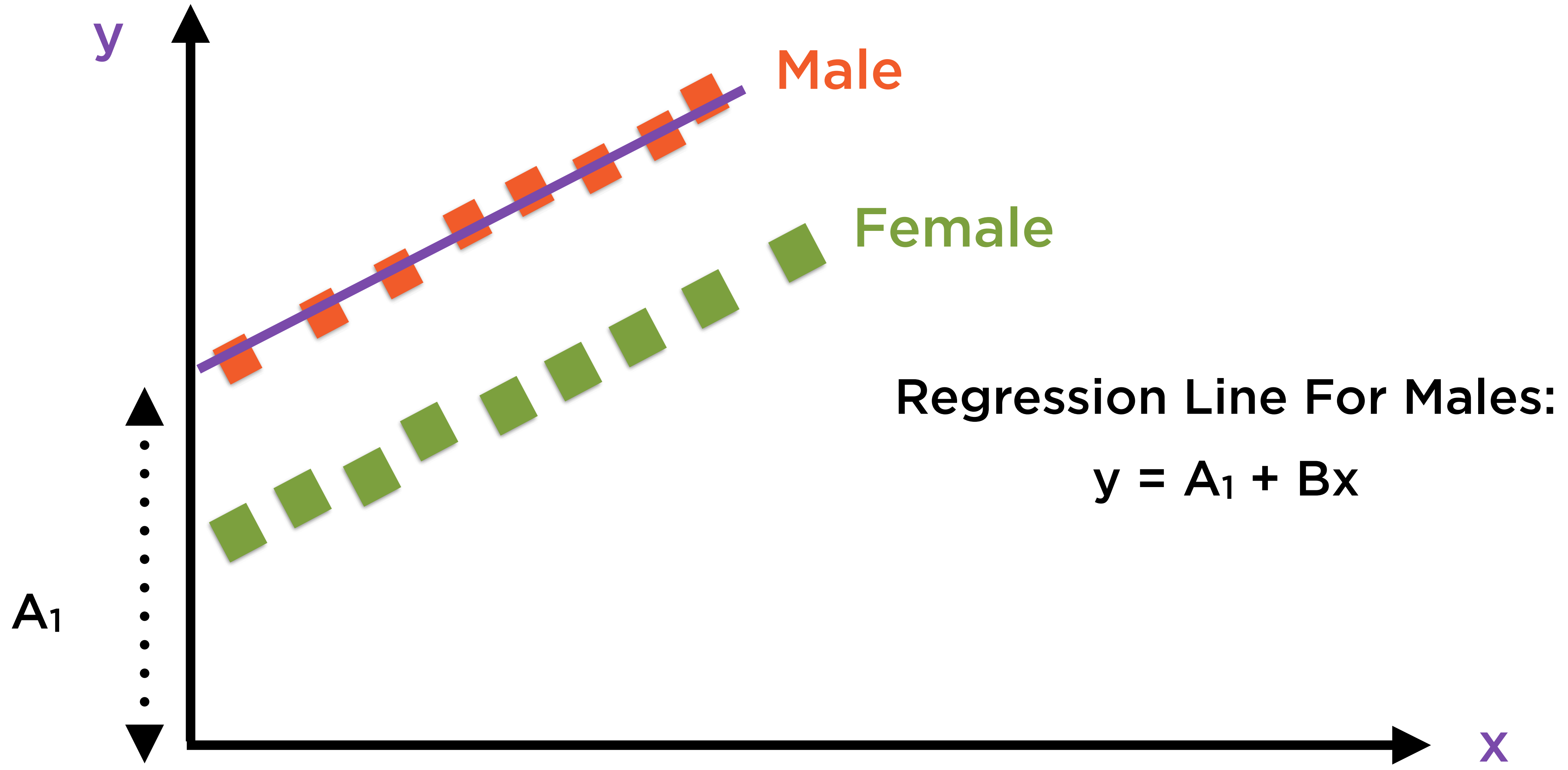
Average height of parents

A Simple Regression



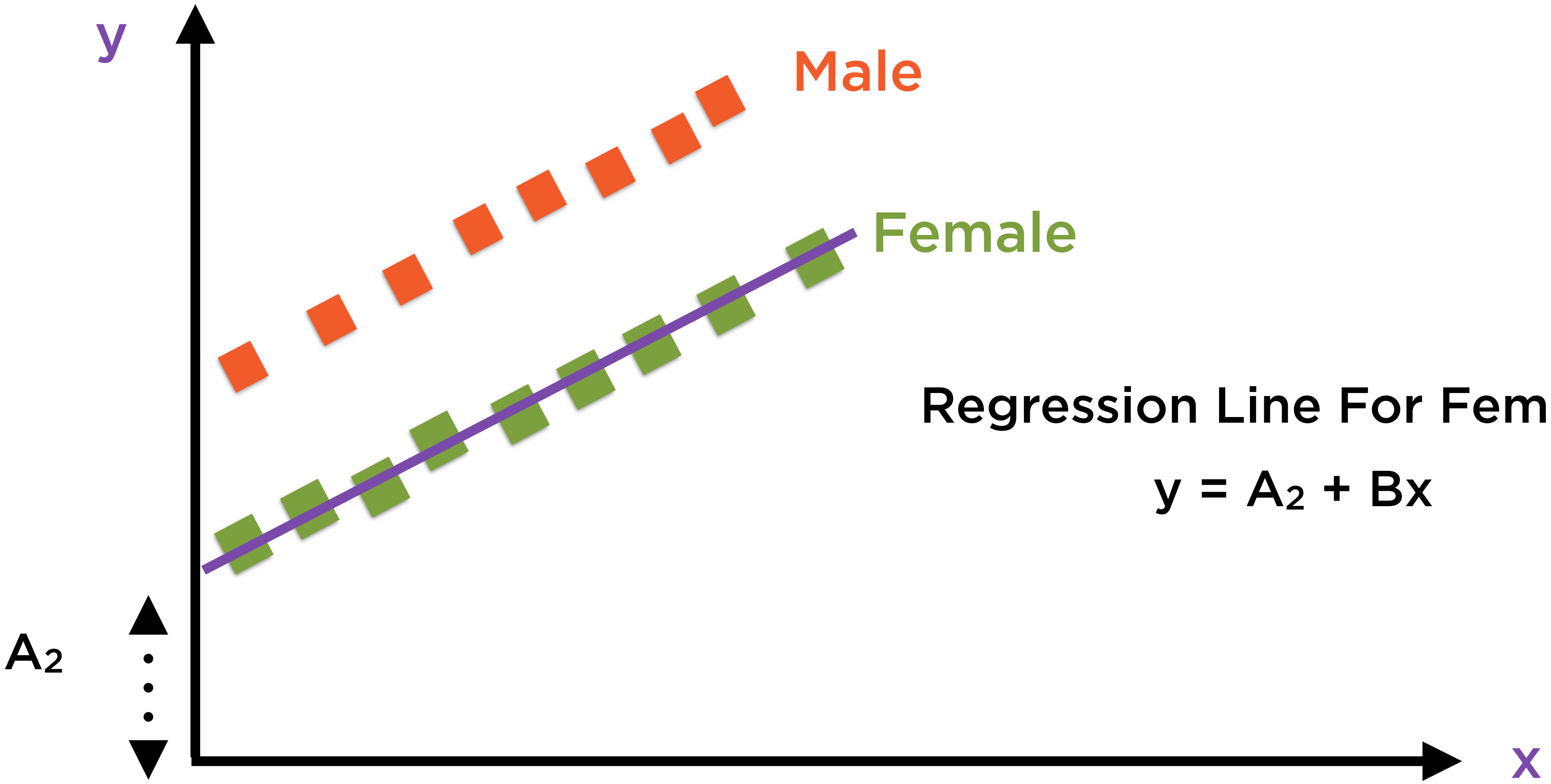
Not a great fit - regression line is far from all points!

A Simple Regression



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

A Simple Regression

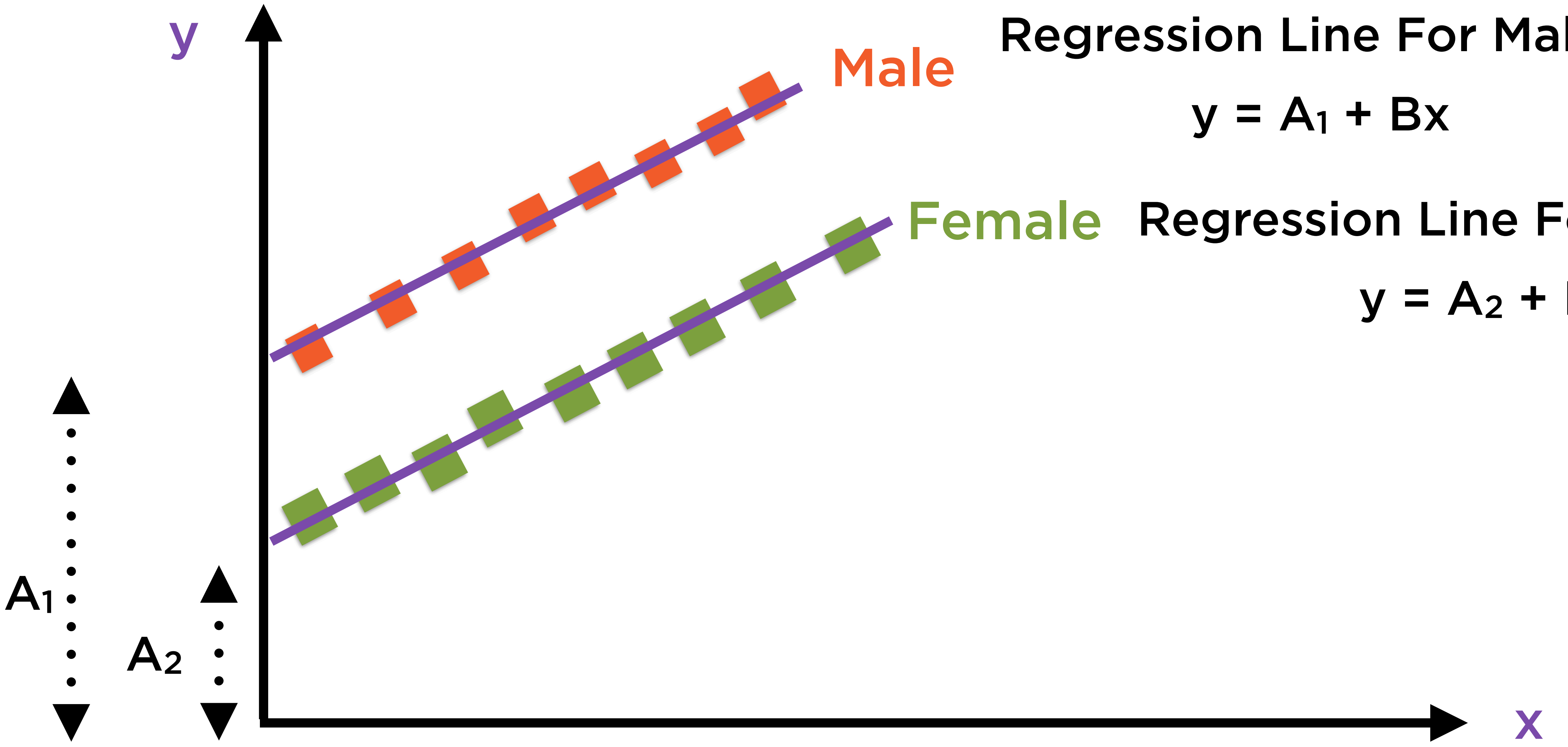


Regression Line For Females:

$$y = A_2 + Bx$$

...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

Adding A Dummy Variable

Regression Line For Males:

$y = A_1 + Bx$

Regression Line For Females:

$y = A_2 + Bx$

Combined Regression Line:

$$y = A_1D_1 + A_2D_2 + Bx$$

$D_1 = 1$ for males

$= 0$ for females

$D_2 = 1$ for females

$= 0$ for males

Adding A Dummy Variable

Regression Line For Males:

$y = A_1 + Bx$

Regression Line For Females:

$y = A_2 + Bx$

Combined Regression Line:

$$y = A_1D_1 + A_2D_2 + Bx$$

$D_1 = 1$ for males

$D_2 = 0$ for males

$$y = A_1 x 1 + \cancel{A_2 0} + Bx$$

$$= A_1 + Bx$$

Adding A Dummy Variable

Regression Line For Males:

$y = A_1 + Bx$

Regression Line For Females:

$y = A_2 + Bx$

Combined Regression Line:

$$y = A_1D_1 + A_2D_2 + Bx$$

$D_1 = 0$ for females

$D_2 = 1$ for females

$$y = \cancel{A_1 x 0} + A_2 x 1 + Bx$$

$$= A_2 + Bx$$

Adding A Dummy Variable

Original Regression Equation:

$$y = A + Bx$$

Height of individual

Average height of parents

Combined Regression Line:

$$y = A_1D_1 + A_2D_2 + Bx$$

$D_1 = 1$ for males

$= 0$ for females

$D_2 = 1$ for females

$= 0$ for males

Given data with k groups, set up k-1 dummy variables and an intercept, <u>or</u> **k dummy variables with no intercept**

Demo

Perform regression with categorical variables in R

Multiple Regression Using R

$$y = A + B_{S\&P500}x_1$$

y = Returns on Exxon stock (XOM)

x_1 = Returns on S&P 500

Multiple Regression Using R

$$y = A + B_{NASDAQ}x_1$$

y = Returns on Exxon stock (XOM)

x_1 = Returns on NASDAQ

Multiple Regression Using R

$$y = A + B_{S\&P500}x_1 + B_{NASDAQ}x_2$$

y = Returns on Exxon stock (XOM)

x_1 = Returns on S&P 500

x_2 = Returns on NASDAQ

Multiple Regression Using R

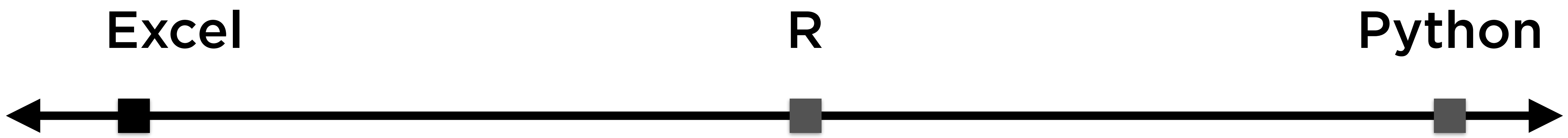
$$y = A + B_{S\&P500}x_1 + B_{USO}x_2$$

y = Returns on Exxon stock (XOM)

x_1 = Returns on S&P 500

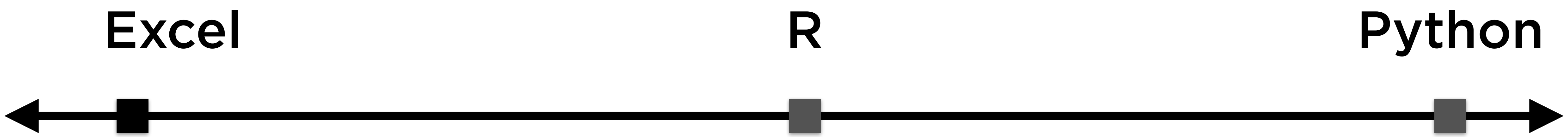
x_2 = Returns of oil prices (USO)

Ease of Prototyping



Excel is an awesome prototyping tool

Robustness and Reuse



R

Use **R for regression**: It makes sense whatever your use-case

Summary

Implemented multiple regression in R

Interpreted results of a multiple regression

Carried out multiple regression in R to include categorical variables