Implementing Multiple Regression Models in Excel

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

Overview

Implement multiple regression in Excel

Interpret results of a multiple regression

Carry out multiple regression in Excel to include categorical variables

Implementing Multiple Regression In Excel

Regression Functions in Excel

Slope	**Intercept**
Rsq	**Forecast**

We already have used several different functions for simple regression and forecasting

Multiple Regression in Excel

linest

$y = A + B_1x_1 + B_2x_2$

logest

$y = A \times B_1^{x1} \times B_2^{x2}$

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

y = Returns on Exxon stock (XOM)

x_1 = Returns on S&P 500

x_2 = Returns of oil prices (USO)

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

DATE	XOM
2016-12-01	1.5%
2016-11-01	-0.9%
2006-01-01	0.5%

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

DATE	S&P 500	USO
2016-12-01	1.2%	2.5%
2016-11-01	-1.1%	-4%
2006-01-01	0.7%	2.3%

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

TRUE

If TRUE

$y = A + Bx$

else

$y = Bx$

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

TRUE

If TRUE, detailed regression statistics are displayed

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
ESS	RSS	

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

B_{USO}	$B_{S\&P500}$	$\boxed{A}$
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
ESS	RSS	

Intercept A

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
ESS	RSS	

Coefficients (in reverse order from formula)

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	**$SE_{S\&P500}$**	**SE_A**
R^2	SER	
F	d_f	
ESS	RSS	

Standard Errors

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
$\boxed{R^2}$	SER	
F	d_f	
ESS	RSS	

R^2

(not adjusted-R^2)

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	**SER**	
F	d_f	
ESS	RSS	

Standard Error of Regression

Multiple Regressing Using **linest**

=linest(known_y's,[known_x's],[const],[stats])

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

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
$\boxed{F}$	d_f	
ESS	RSS	

F-statistic

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

$$y = A + B_{S\&P500}X_1 + B_{USO}X_2$$

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	$\boxed{d_f}$	
ESS	RSS	

Degrees of freedom = n - k - 1

n = number of points

k = number of explanatory variables

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
$\boxed{\text{ESS}}$	RSS	

Explained Sum of Squares

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

$$y = A + B_{S\&P500}X_1 + B_{USO}X_2$$

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
ESS	**RSS**	

Residual Sum of Squares

Multiple Regressing Using **linest**

Array formula: Ctrl+Shift +Enter is awkward

Adjusted R^2 not reported, F-statistic not interpreted

Coefficients reported in reverse order

Missing values not handled gracefully

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

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

$$y = A + B_{S\&P500}X_1 + B_{USO}X_2$$

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
ESS	RSS	

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

$$y = A + B_{S\&P500}X_1 + B_{USO}X_2$$

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	**$SE_{S\&P500}$**	**SE_A**
R^2	SER	
F	d_f	
ESS	RSS	

Standard Errors

Population and Sample

Population

All data points out there in the universe

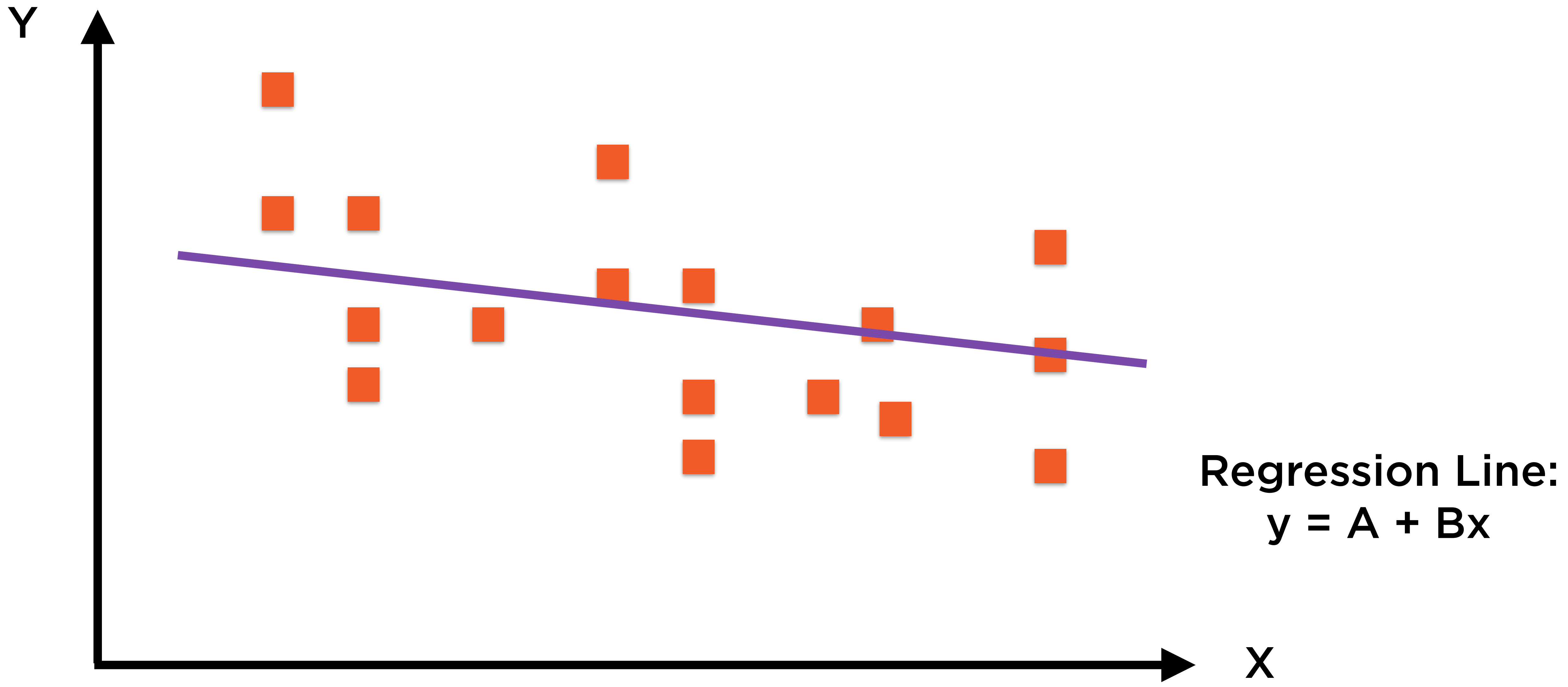
Sample

A subset of the population

Representative Samples

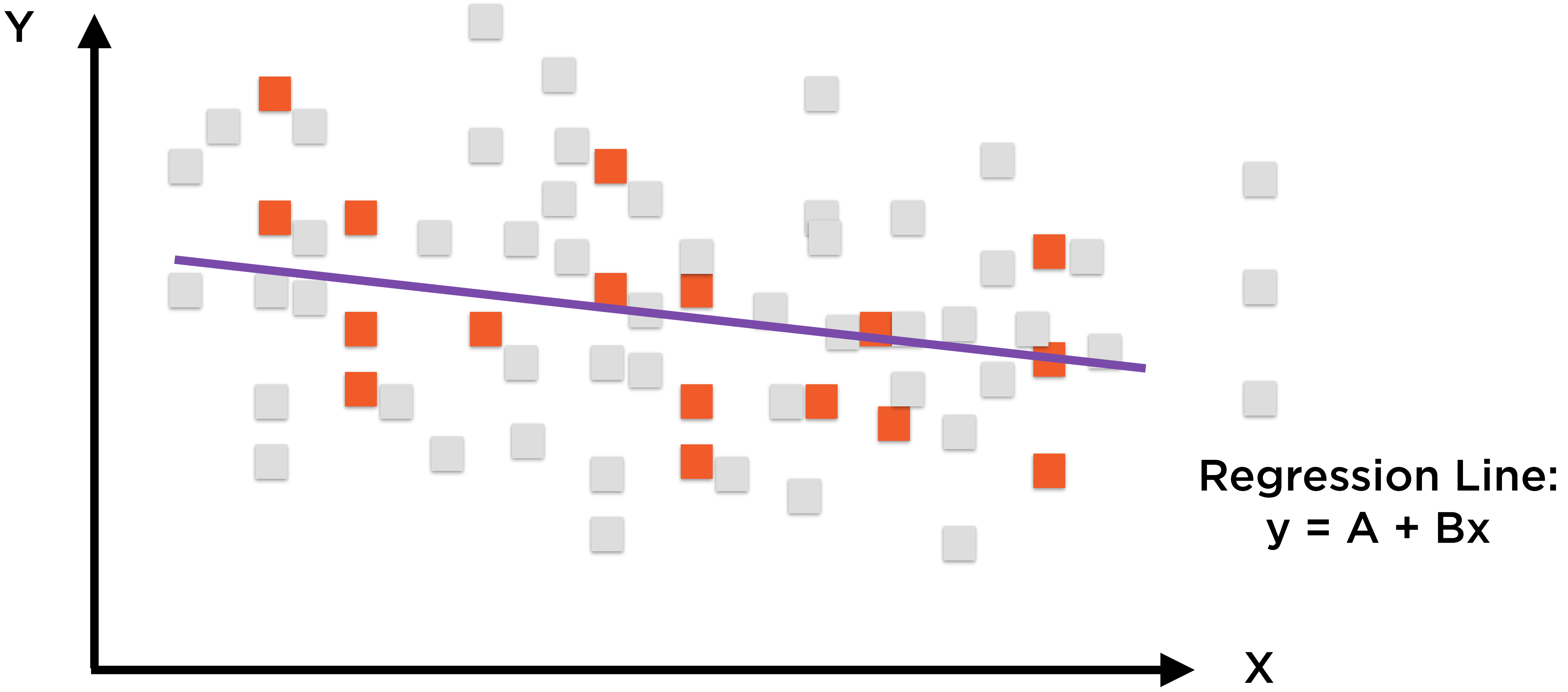
Population **Unbiased Sample** **Biased Sample**

Regression Works on Samples



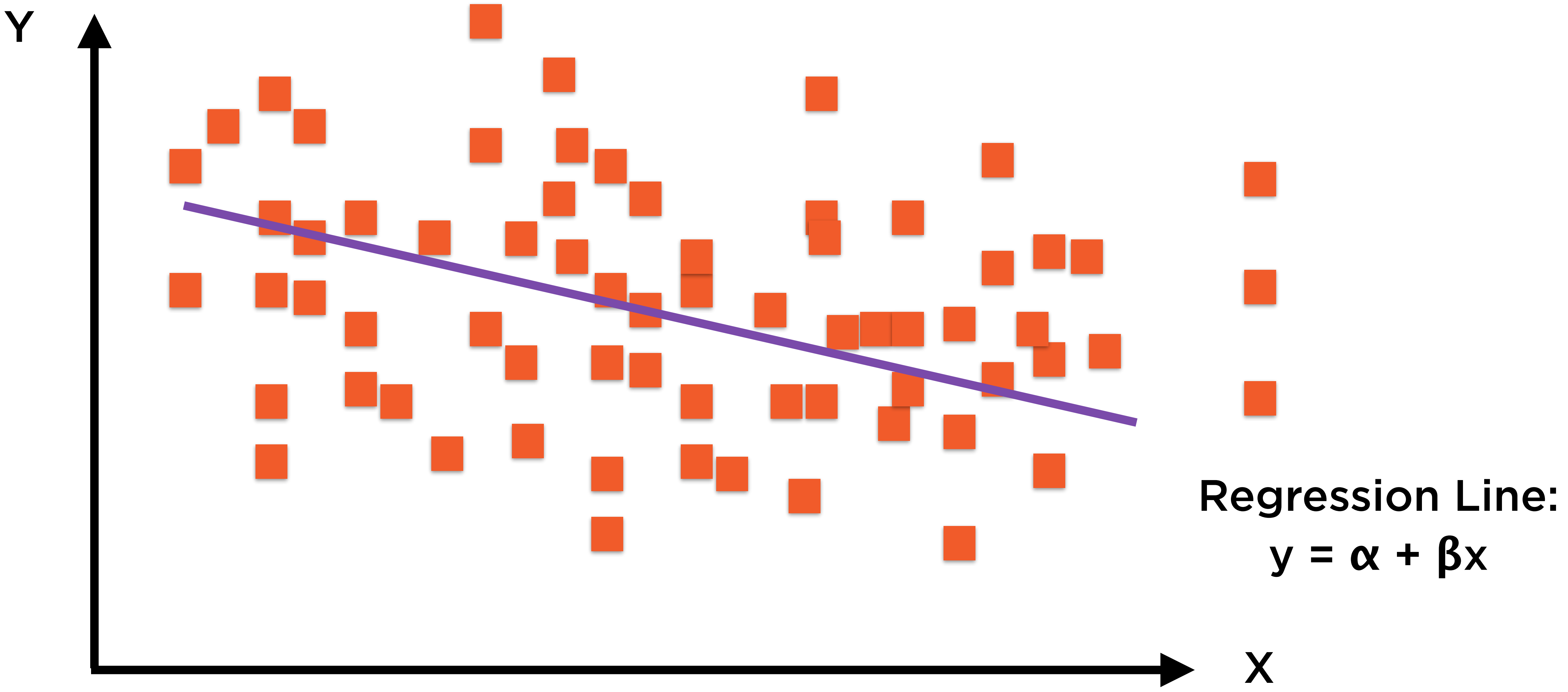
The regression line is based on a sample, not on the population

Regression Works on Samples



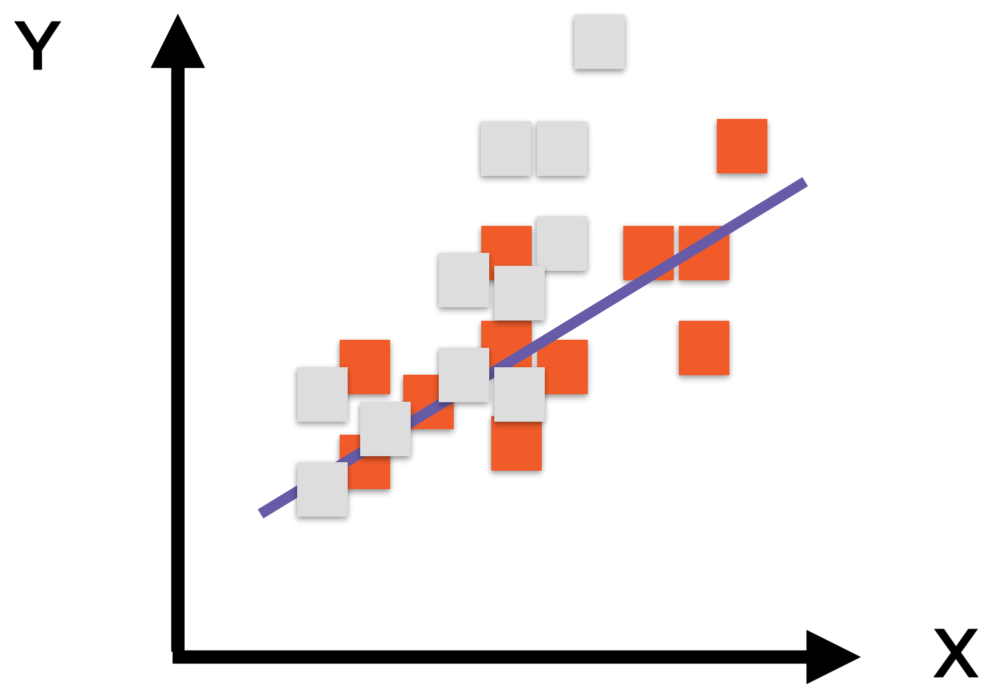
The regression line is based on a sample, not on the population

Regression Works on Samples



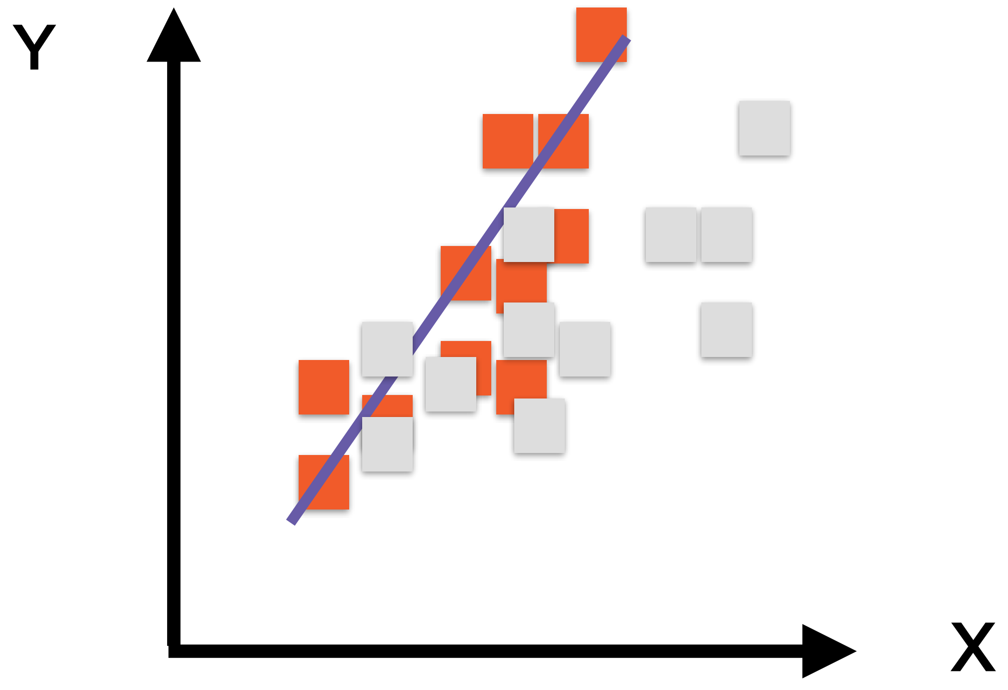
The regression line is based on a sample, not on the population

Different Samples, Different Fits



Sample 1

$$y = A_1 + B_1x$$

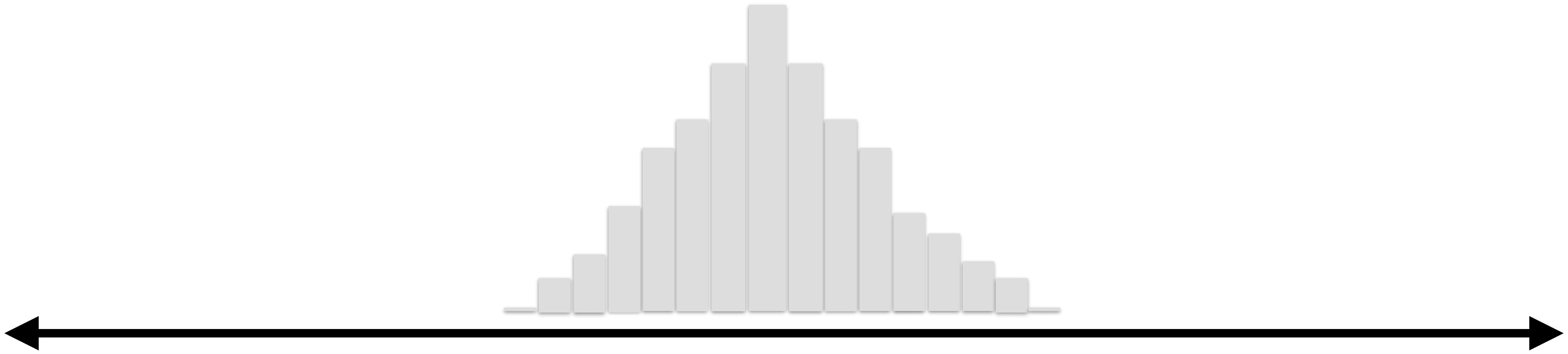


Sample 2

$$y = A_2 + B_2x$$

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

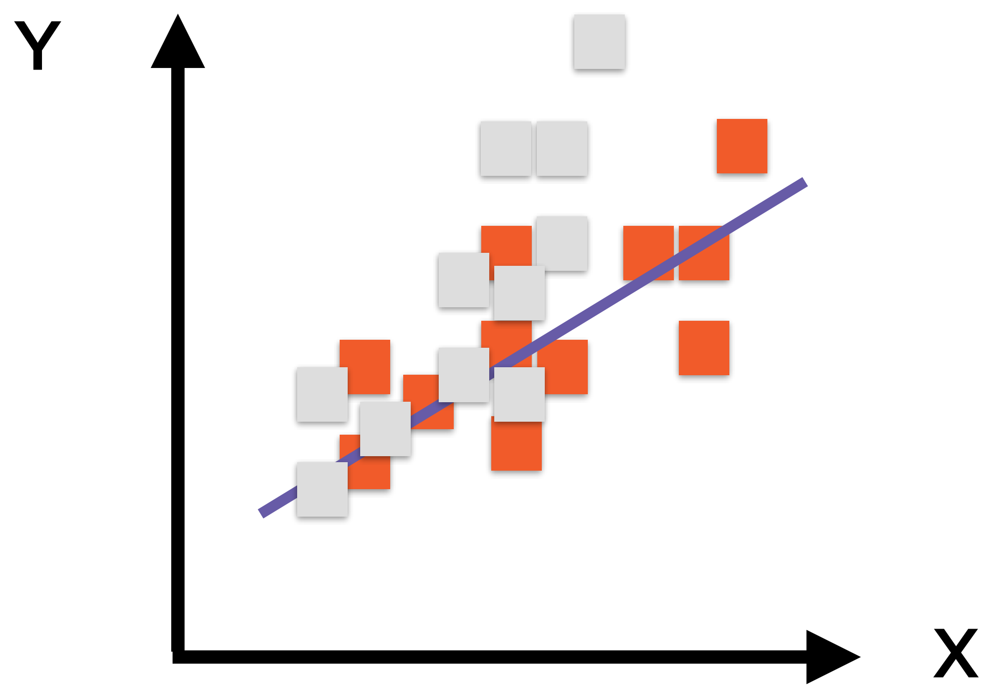
Sampling Distributions



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

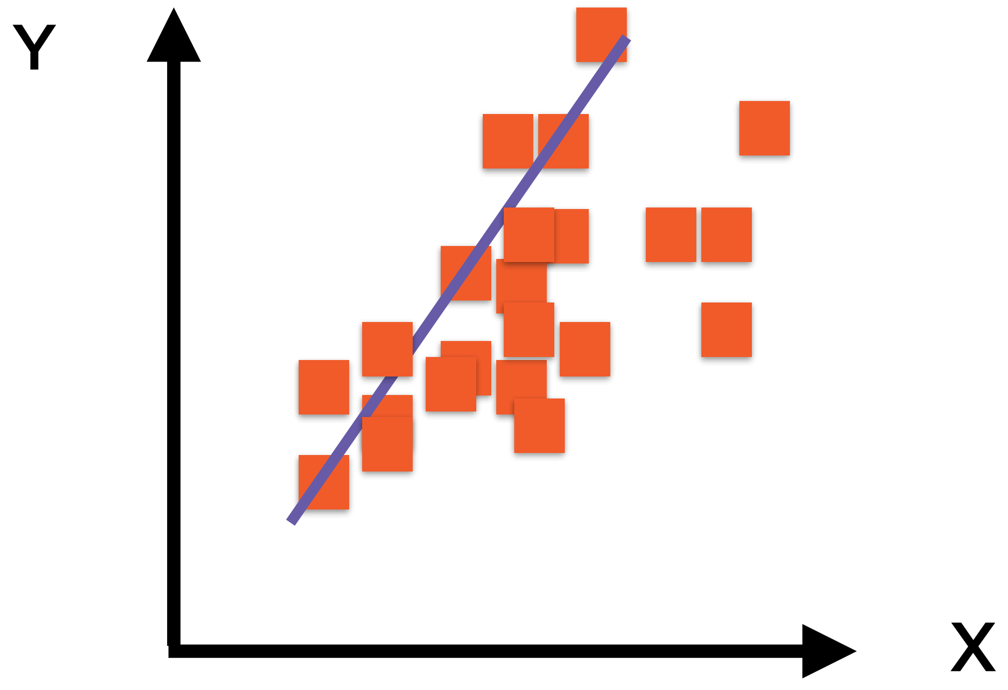
This is known as the sampling distribution

Different Samples, Different Fits



Sample Regression Line

$$y = A + Bx$$

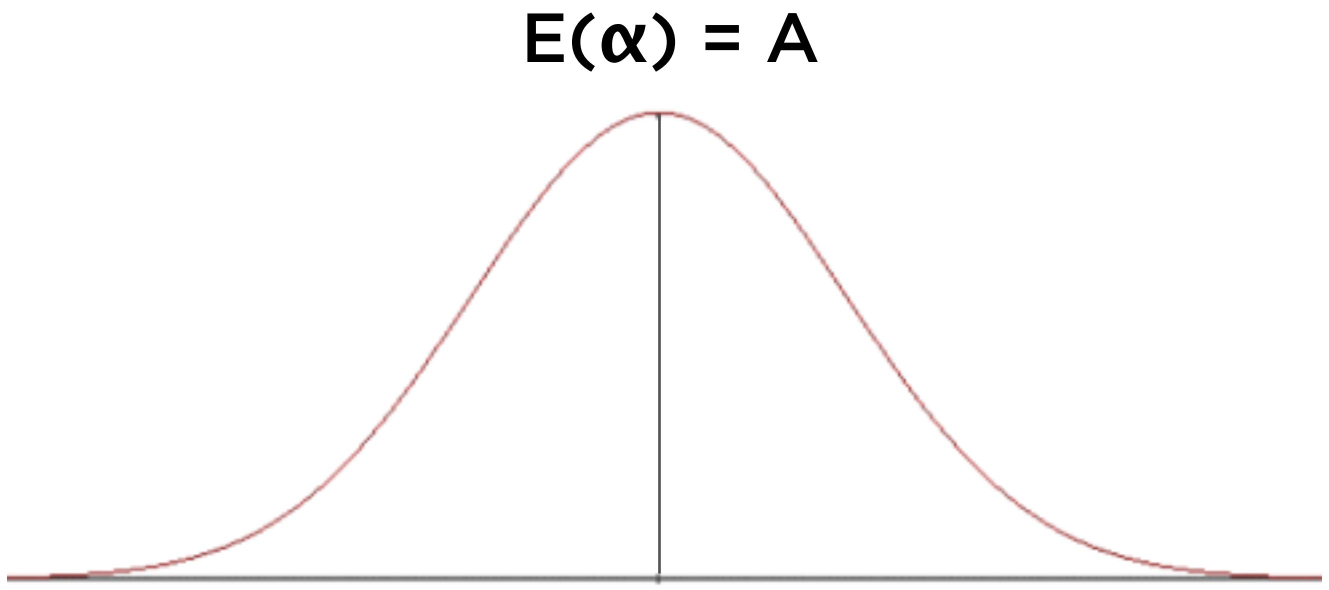


Population Regression Line

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

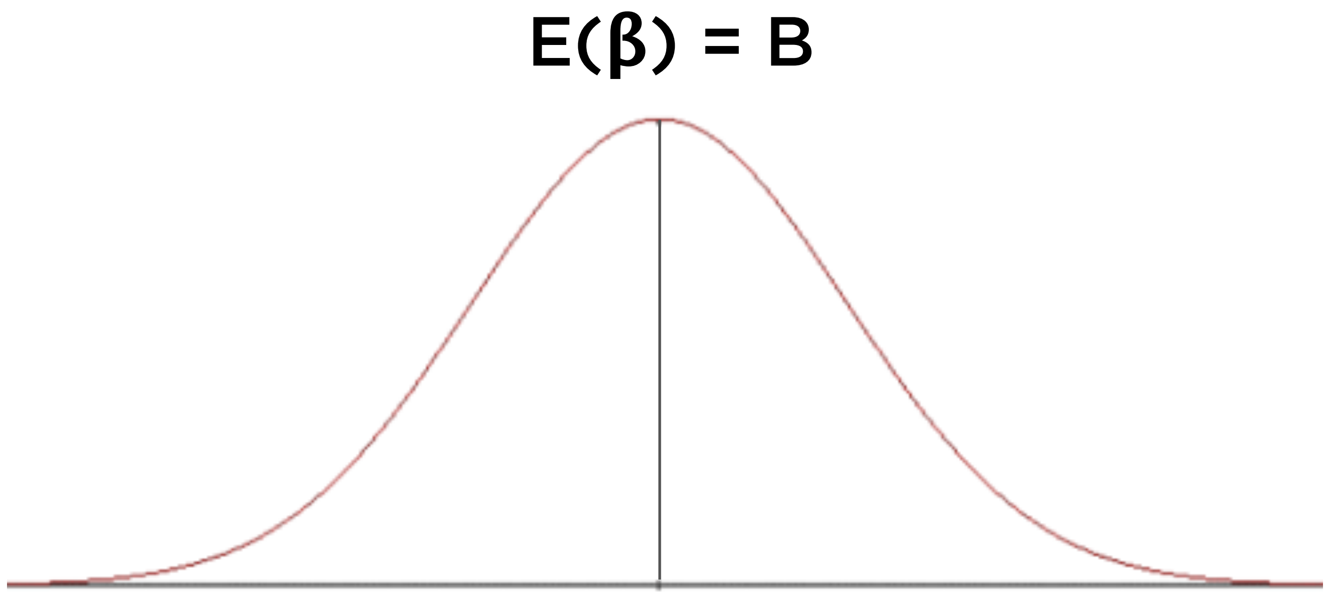
We will never know the values of the population parameters α and β

Sampling Distributions



Sampling Distribution of α

α is the population parameter, A is the sample parameter

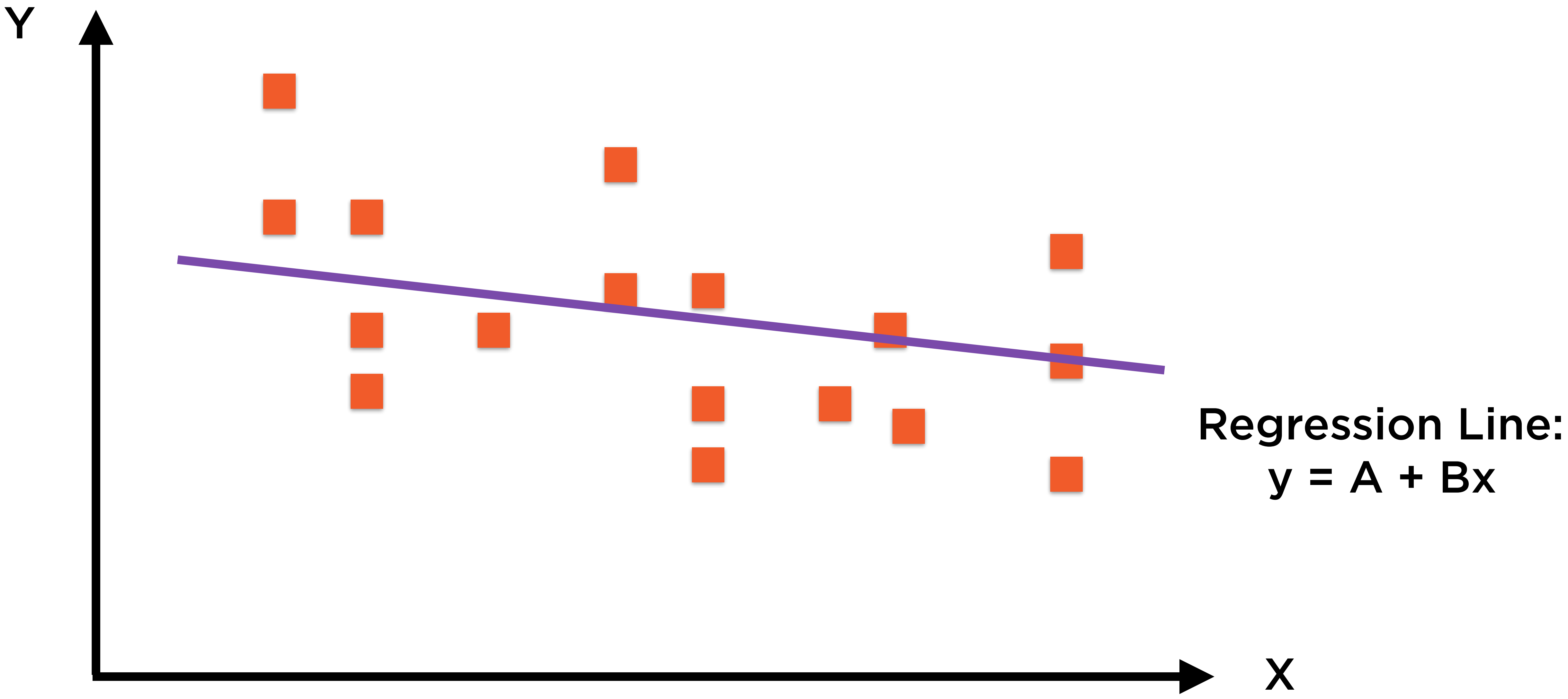


Sampling Distribution of β

β is the population parameter, B is the sample parameter

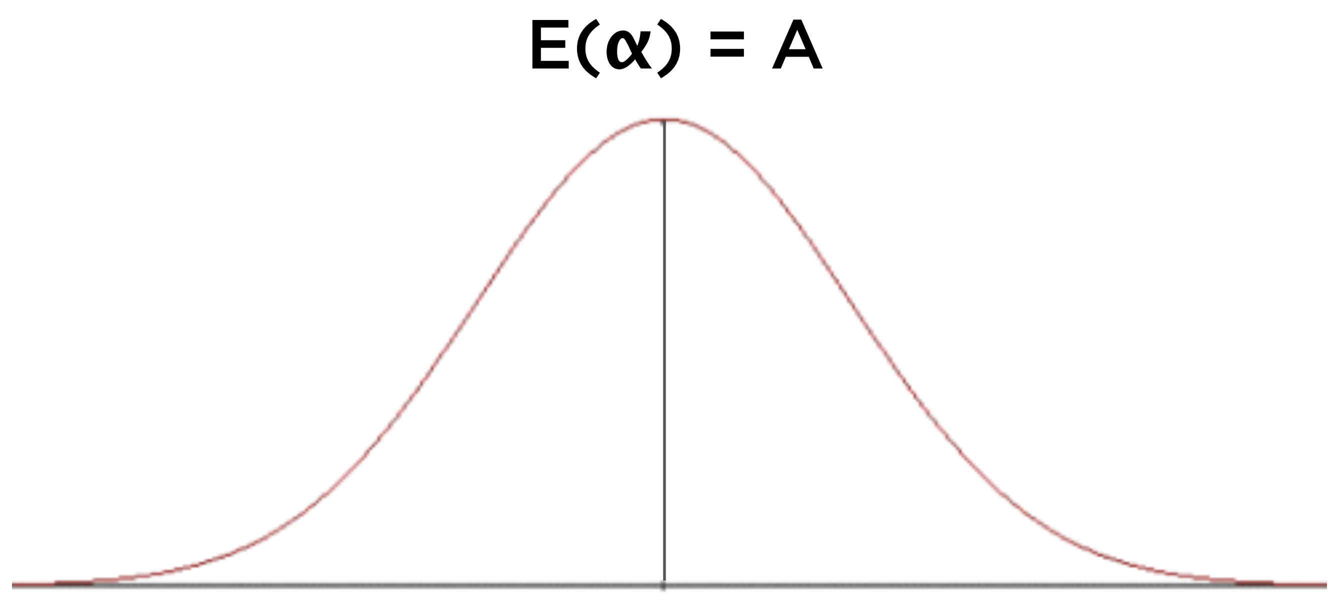
The sampling distributions are normal, and population mean is equal to sample mean

Regression Works on Samples



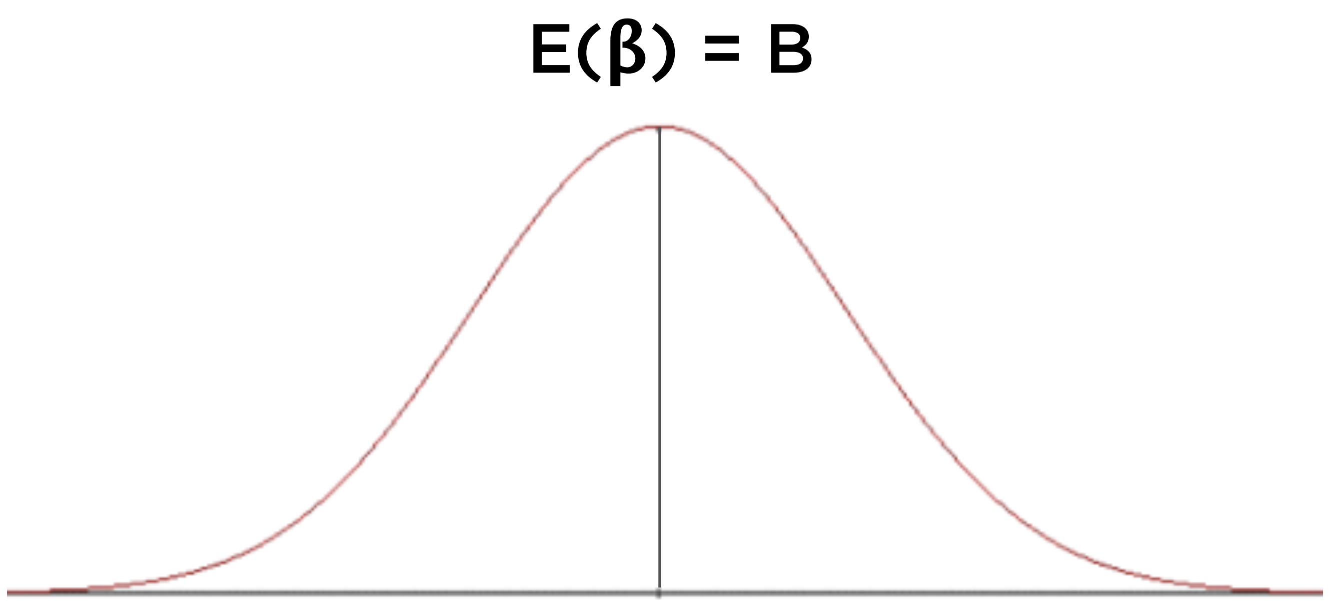
The sample parameters A and B are our 'best' estimates for population parameters α and β

Sampling Distributions



Sampling Distribution of α

α is the population parameter, A is the sample parameter

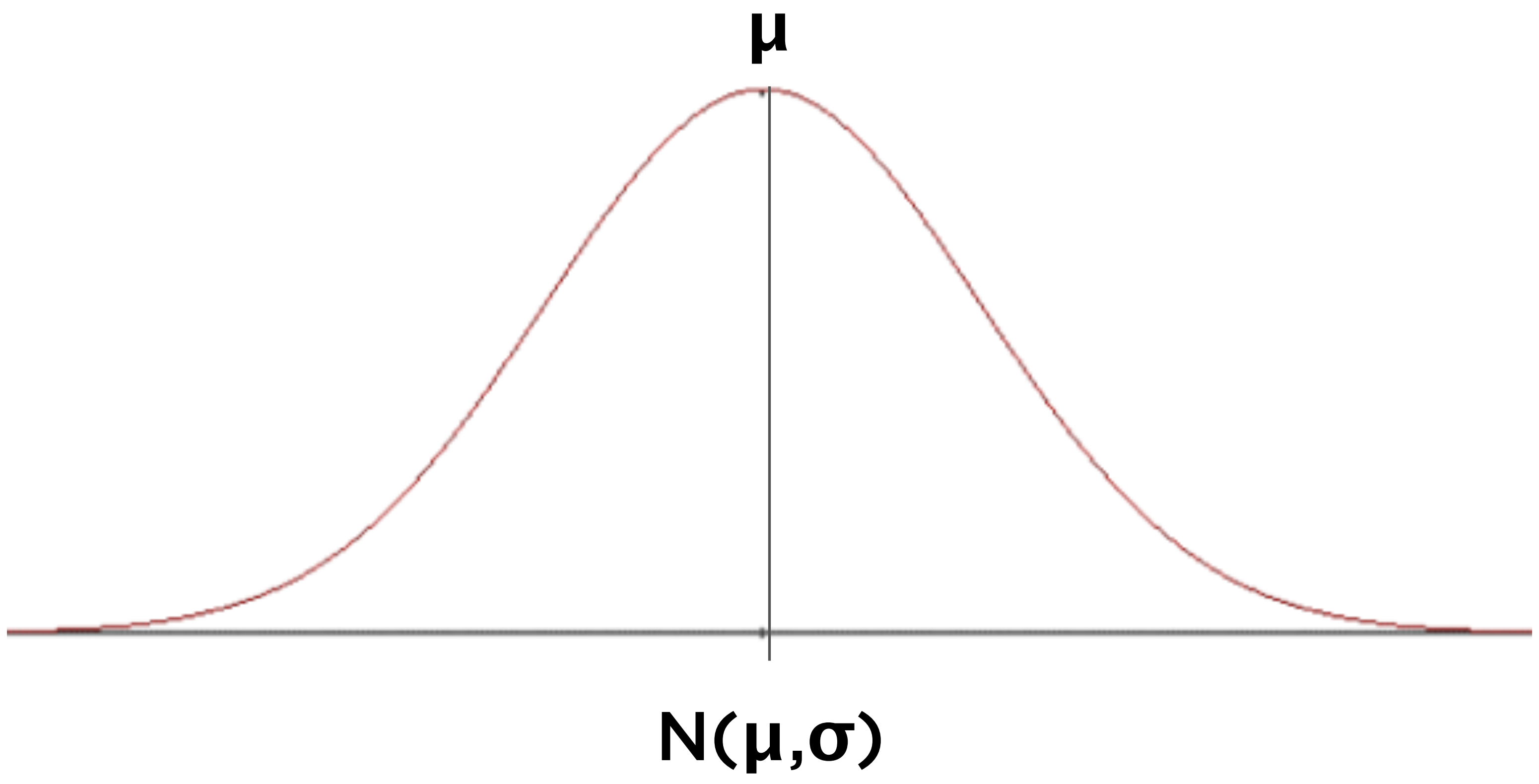


Sampling Distribution of β

β is the population parameter, B is the sample parameter

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

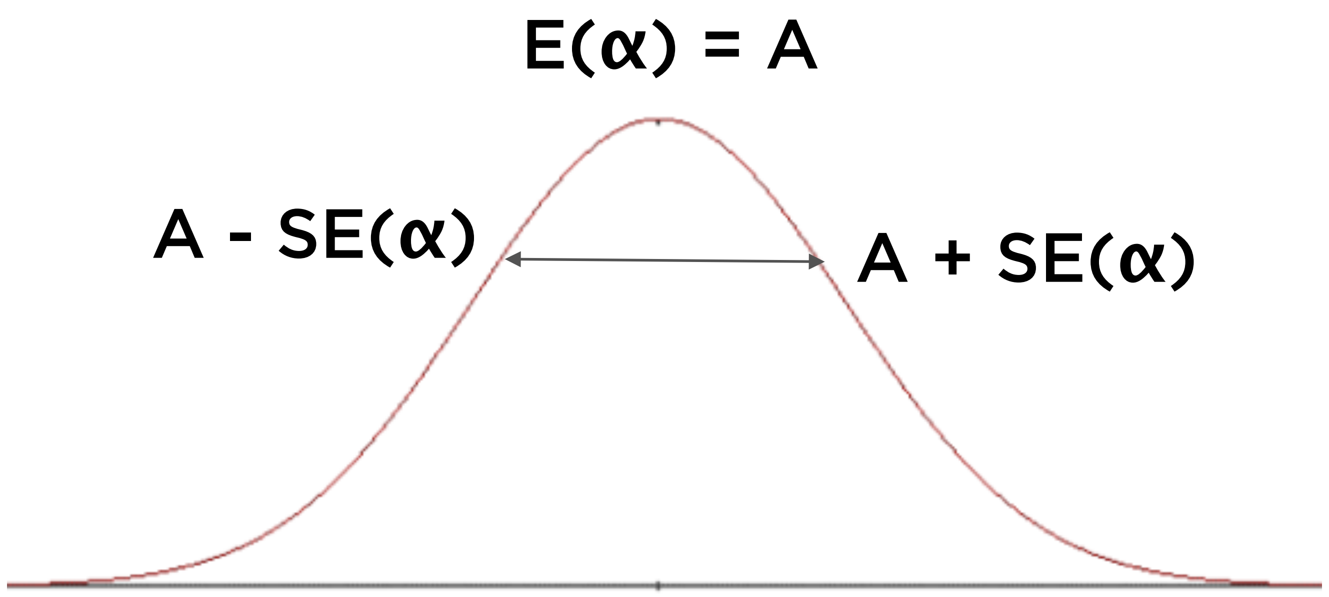
Normal Distribution



Average (mean) is μ

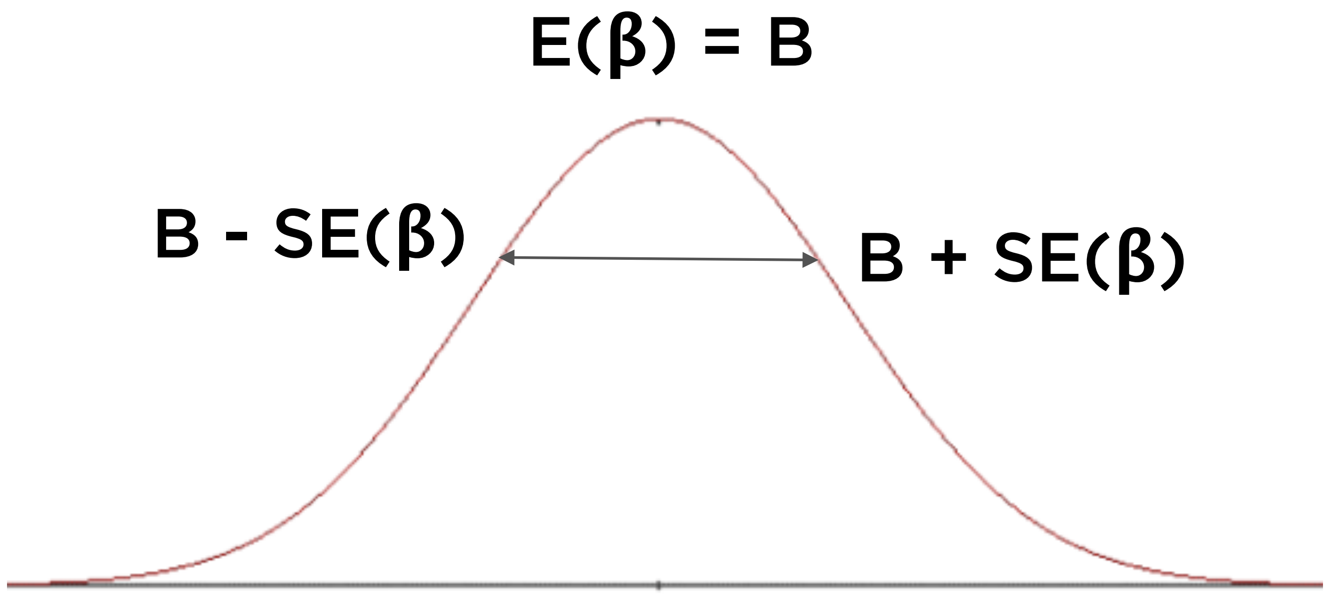
Standard deviation is σ

Standard Errors



Sampling Distribution of α

α is the population parameter, A is the sample parameter

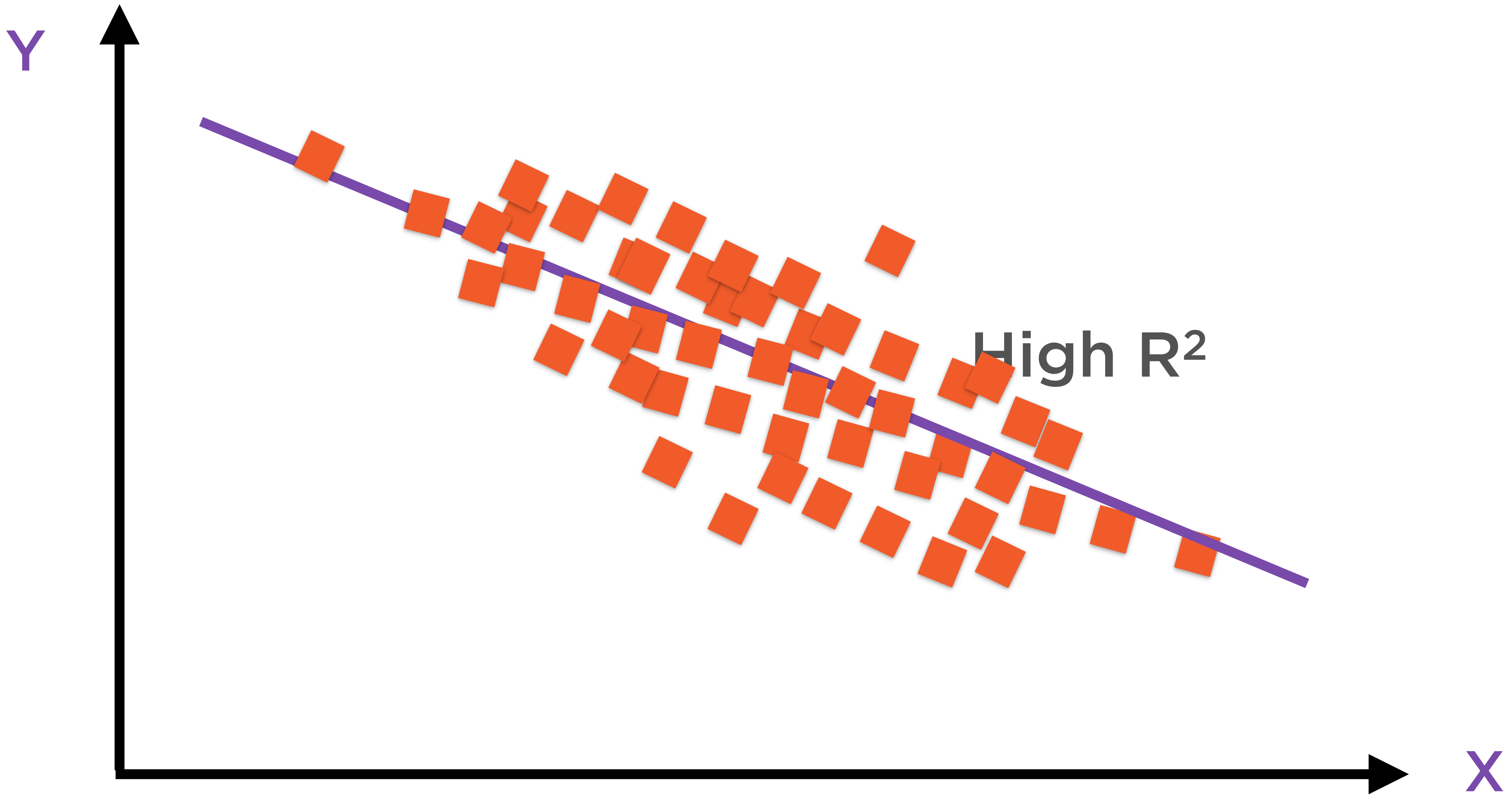


Sampling Distribution of β

β is the population parameter, B is the sample parameter

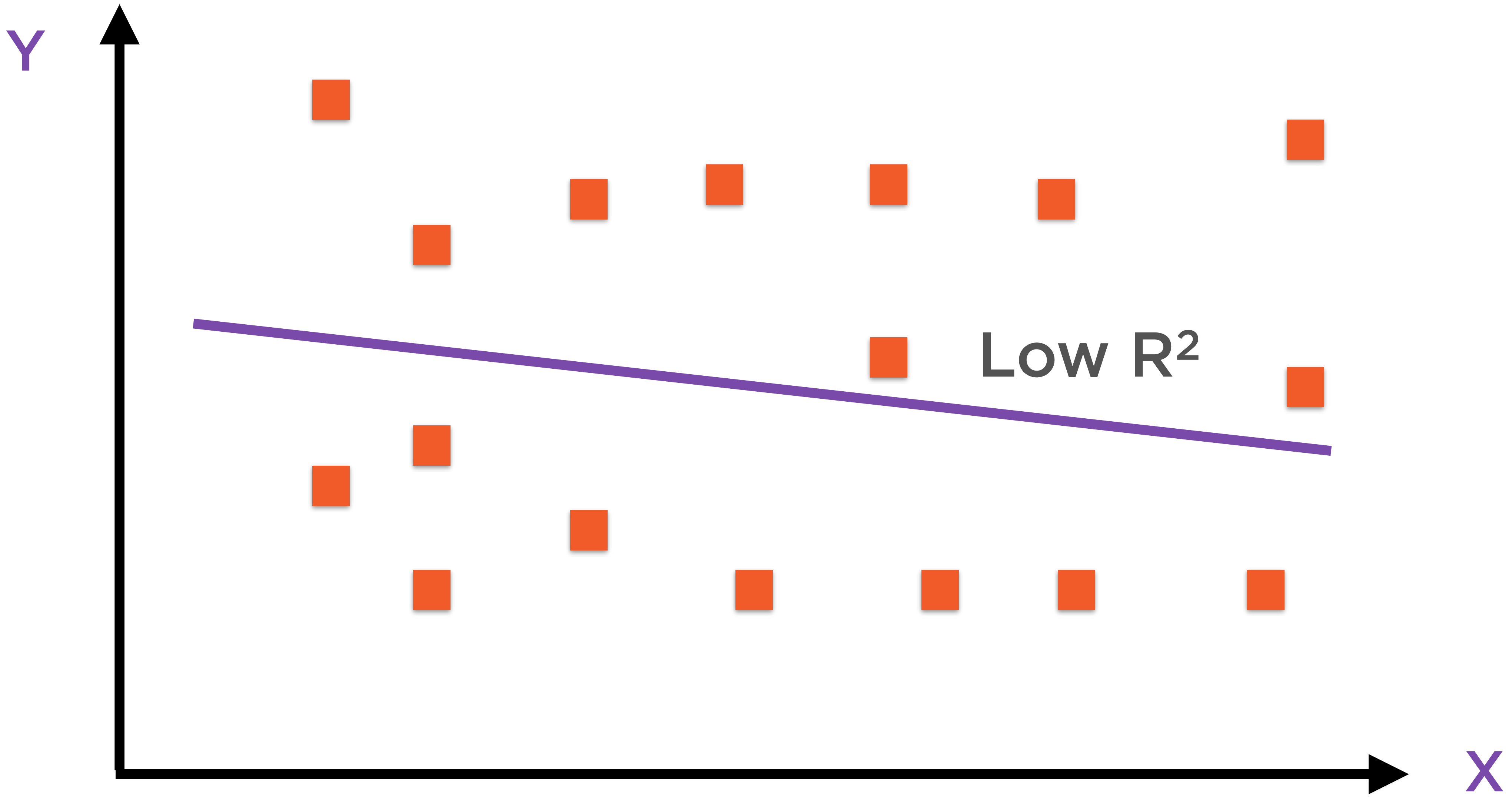
Standard error of a regression parameter is the standard deviation of the sampling distribution

Strong Cause-effect Relationship



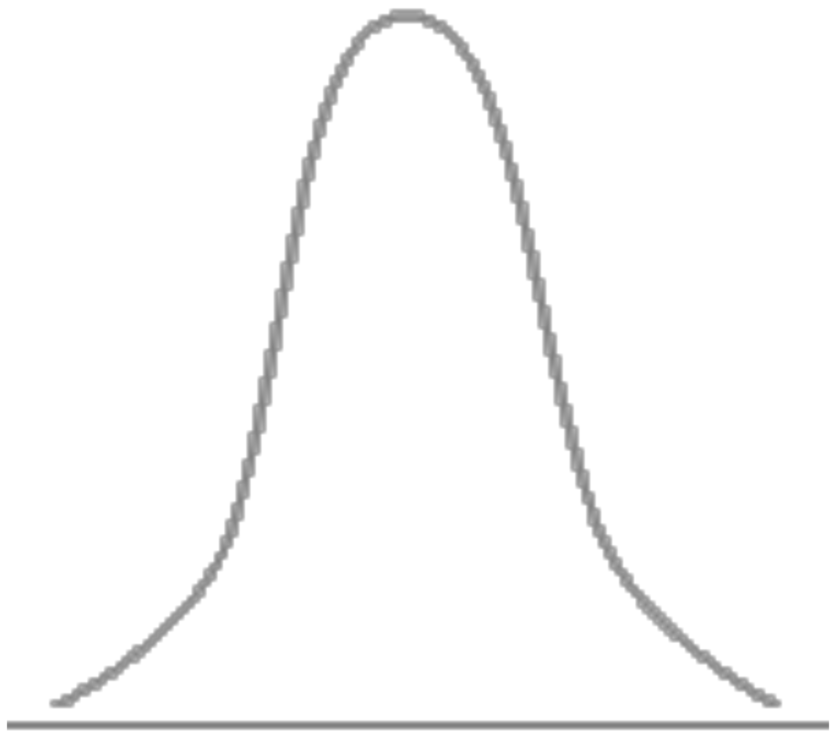
Residuals are small, standard errors are small

Weak Cause-effect Relationship



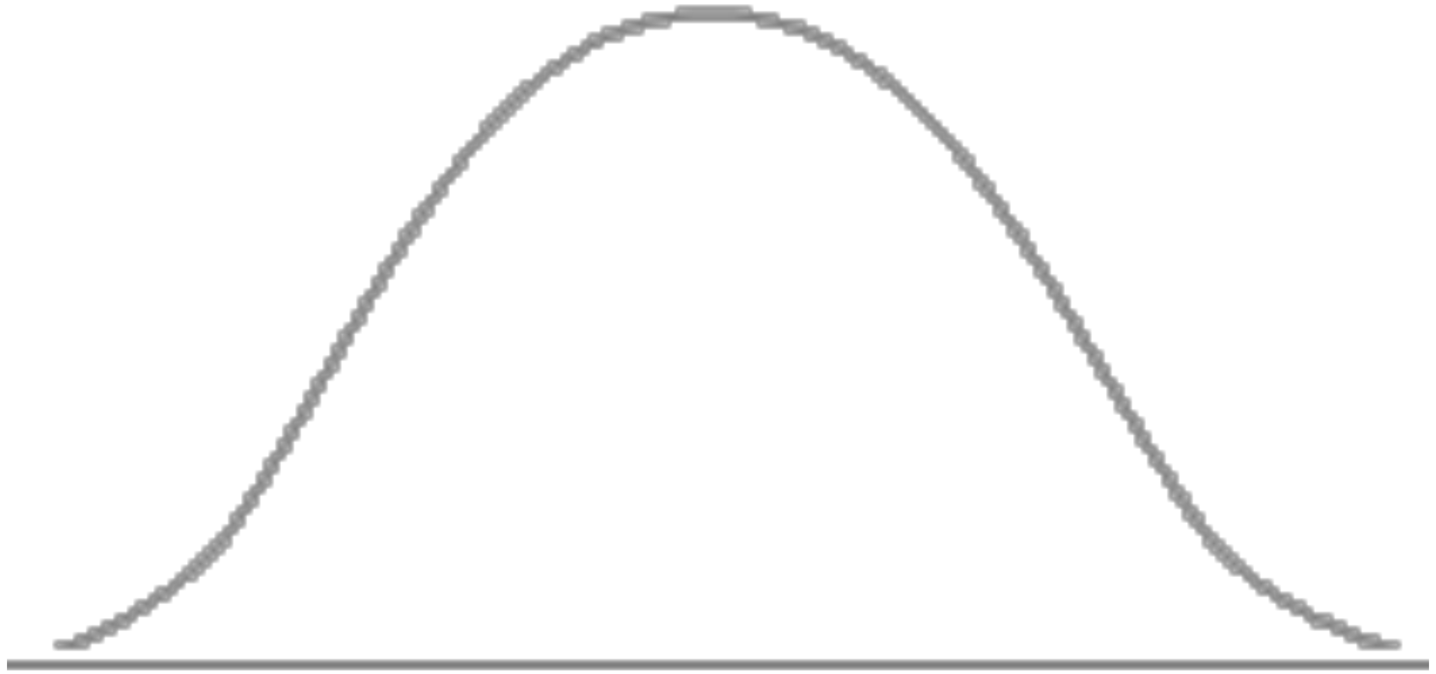
Residuals are large, standard errors are large

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:

$$y = A + Bx$$

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

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

Multiple Regressing Using **linest**

```
=linest(known_y’s,[known_x’s],[const],[stats])
```

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

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
ESS	RSS	

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

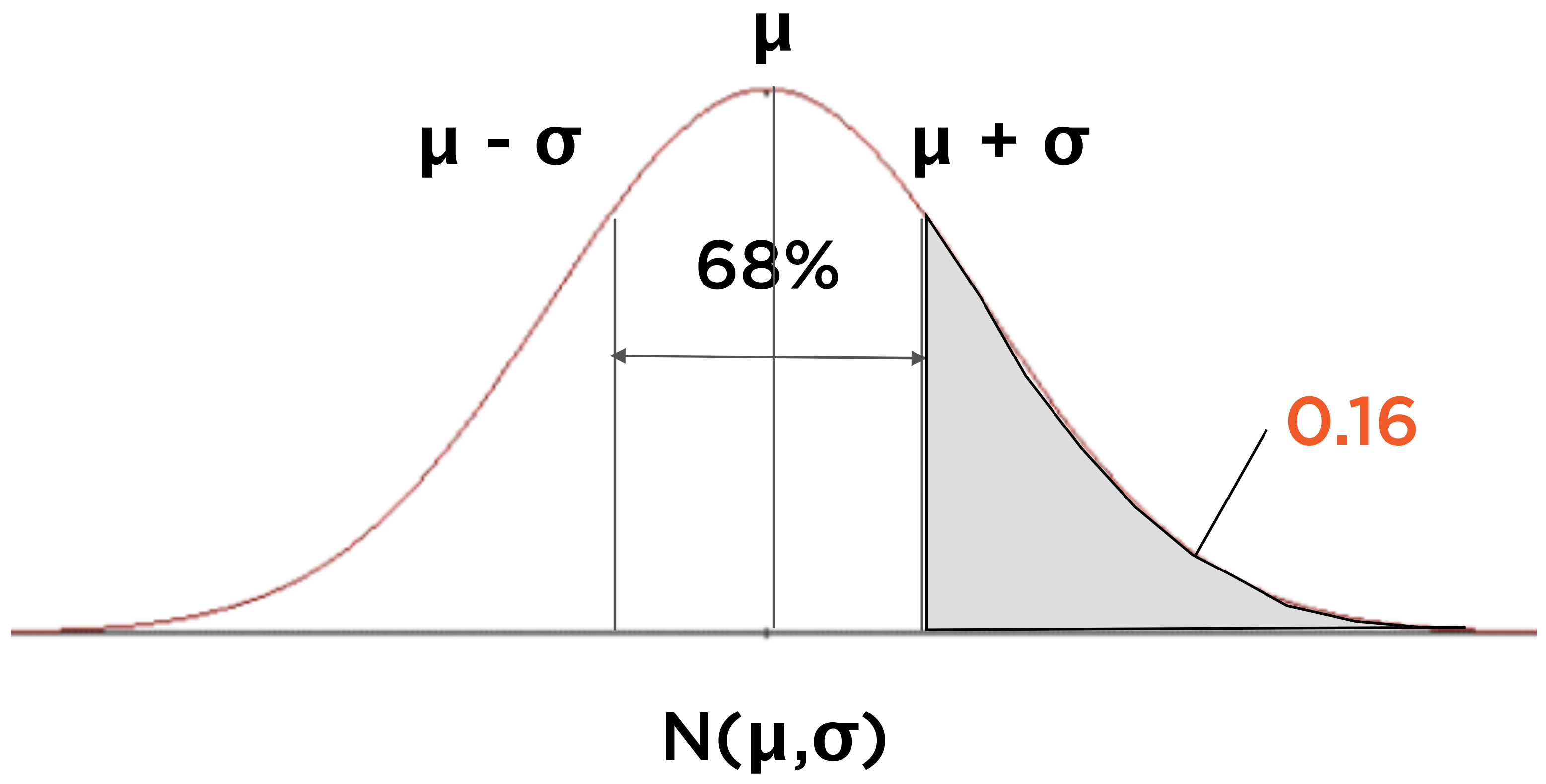
$$y = A + B_{S\&P500}x_1 + B_{USO}x_2$$

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
ESS	RSS	

t-statistics

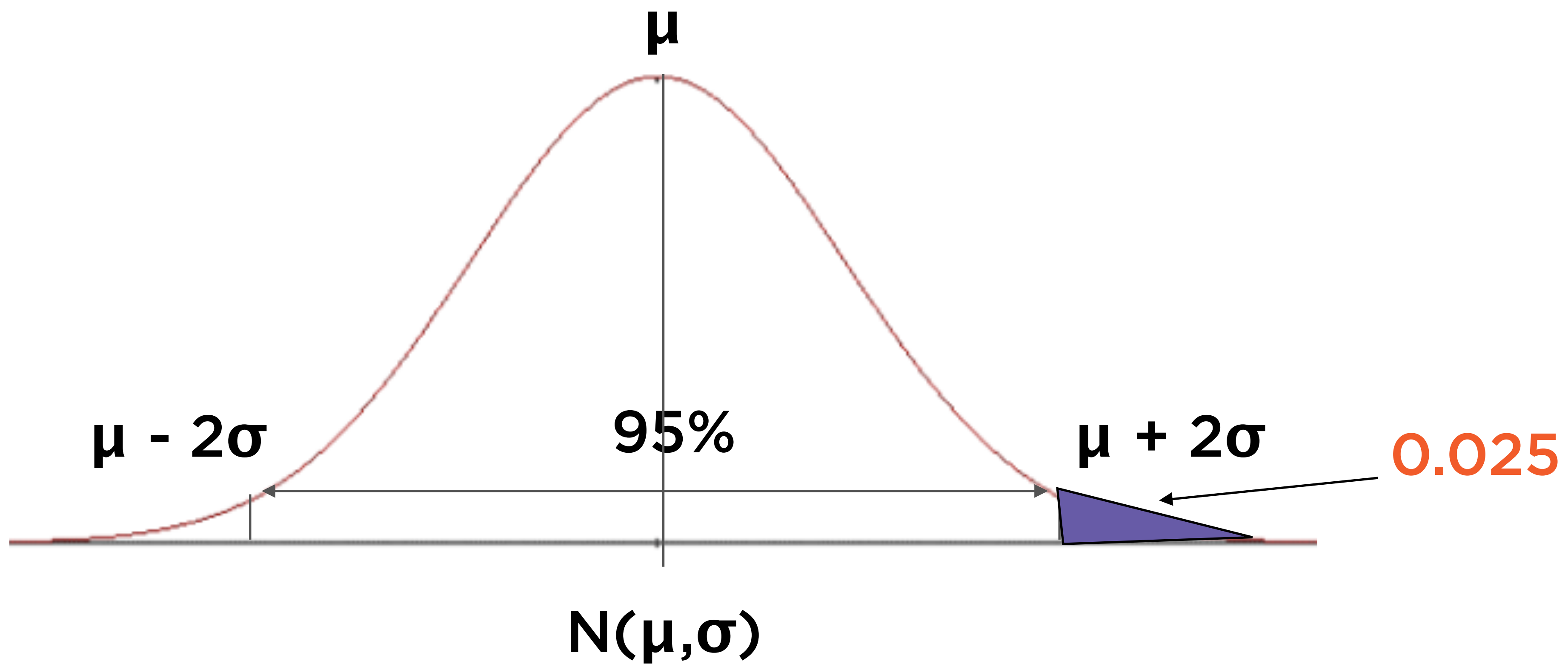
B_{USO} / SE_{USO}	$B_{S\&P500}$ / $SE_{S\&P500}$	A / SE_A

Probability of Occurrence



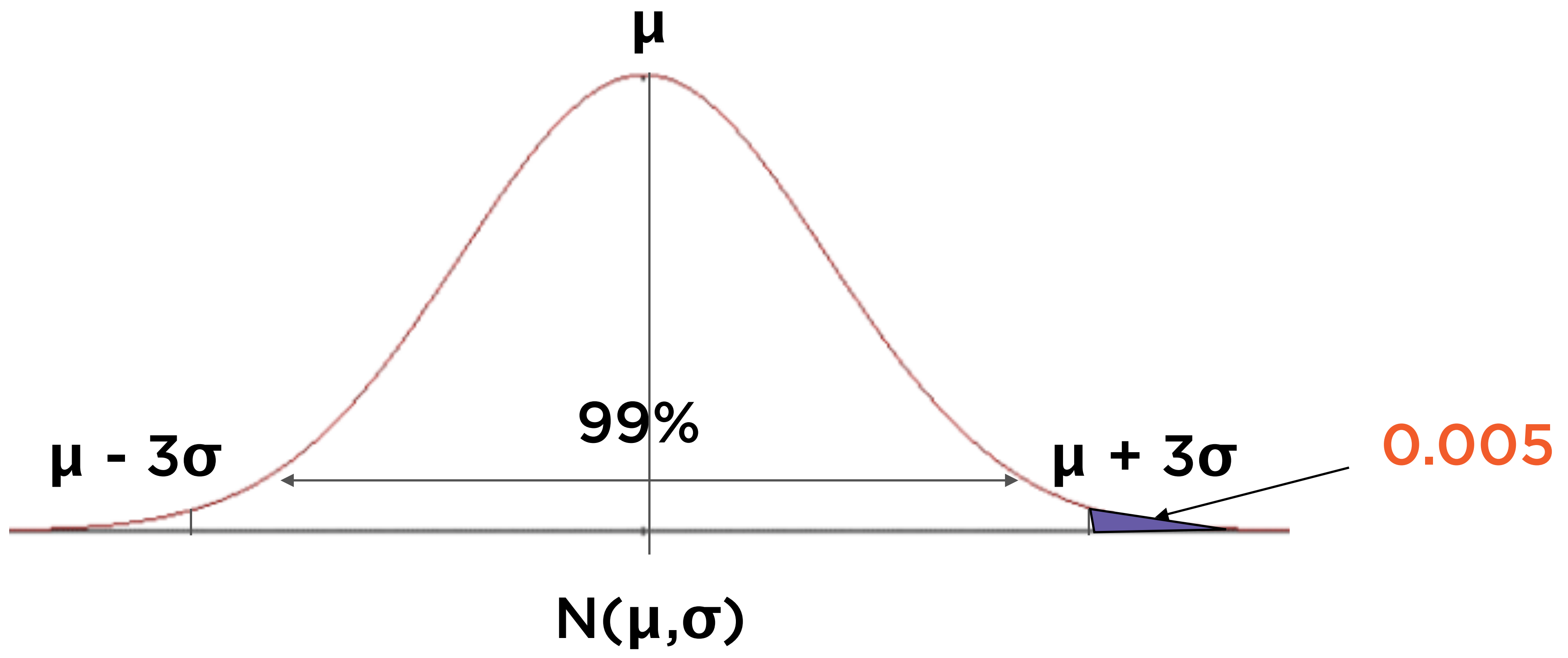
68% within 1 standard deviation of mean

Probability of Occurrence



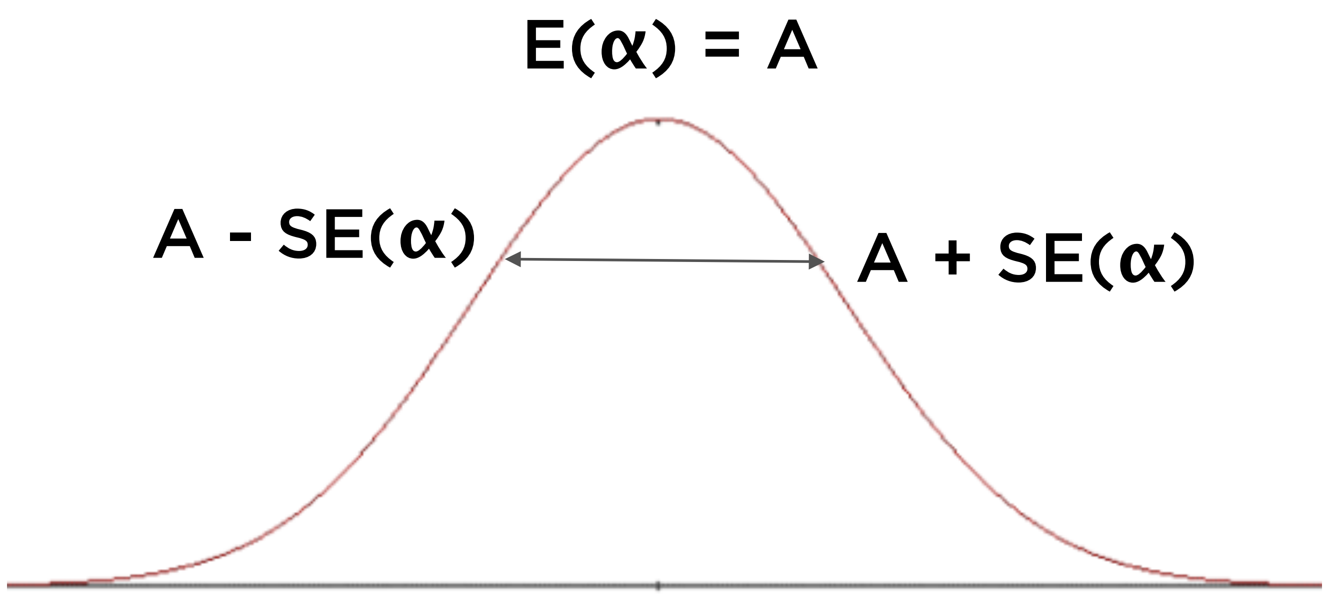
95% within 2 standard deviations of mean

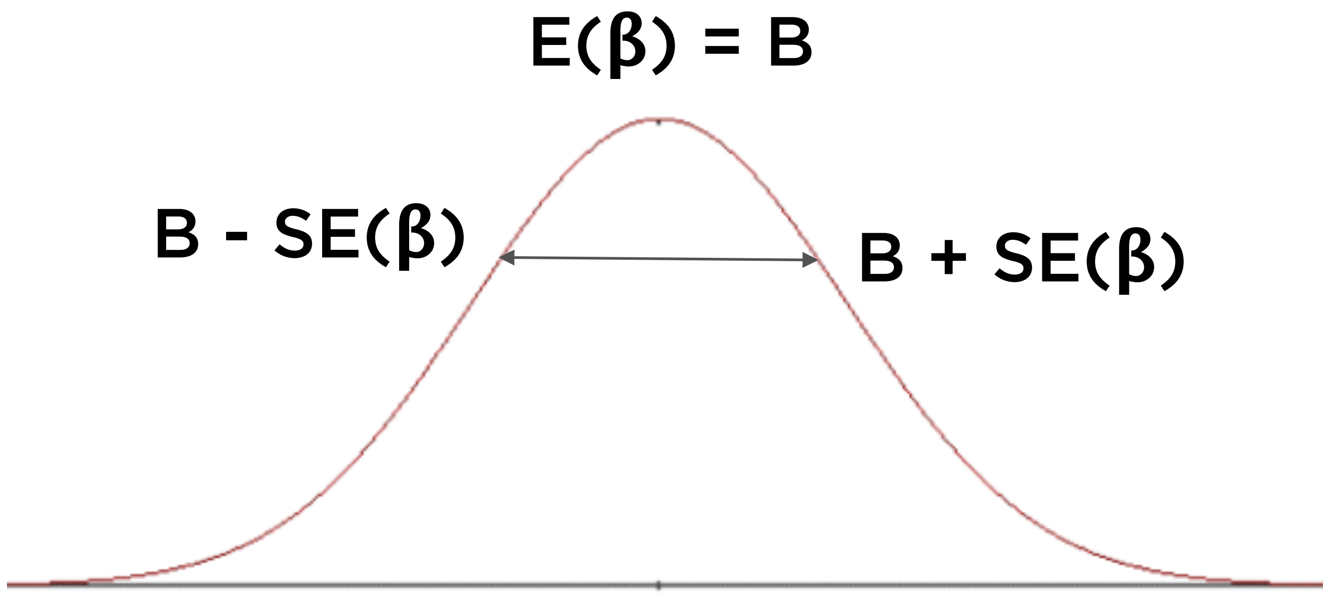
Probability of Occurrence



99% within 3 standard deviations of mean

Standard Errors





Standard Error of α

α is the population parameter, A is the sample parameter

Standard Error of β

β is the population parameter, B is the sample parameter

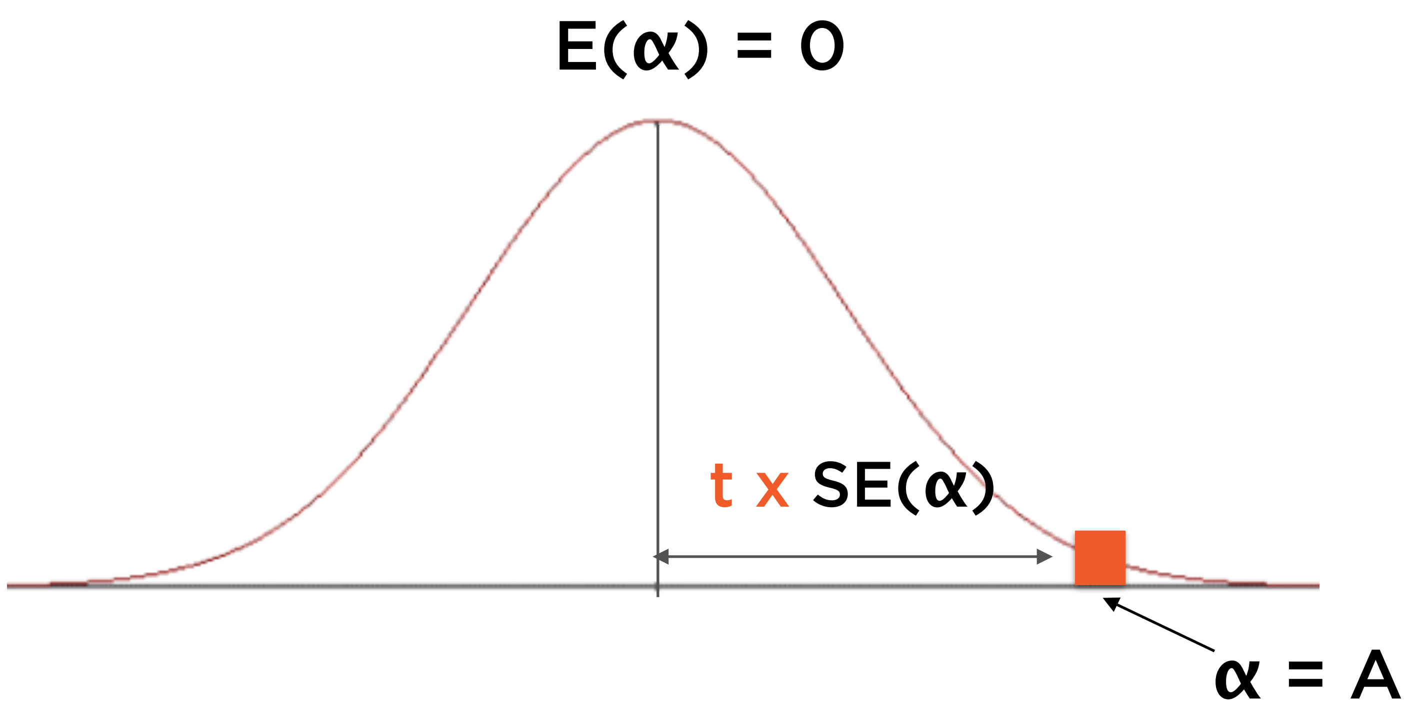
Standard error of a regression parameter is the standard deviation of the sampling distribution

Null Hypotheses

What if the population parameter α were actually zero?

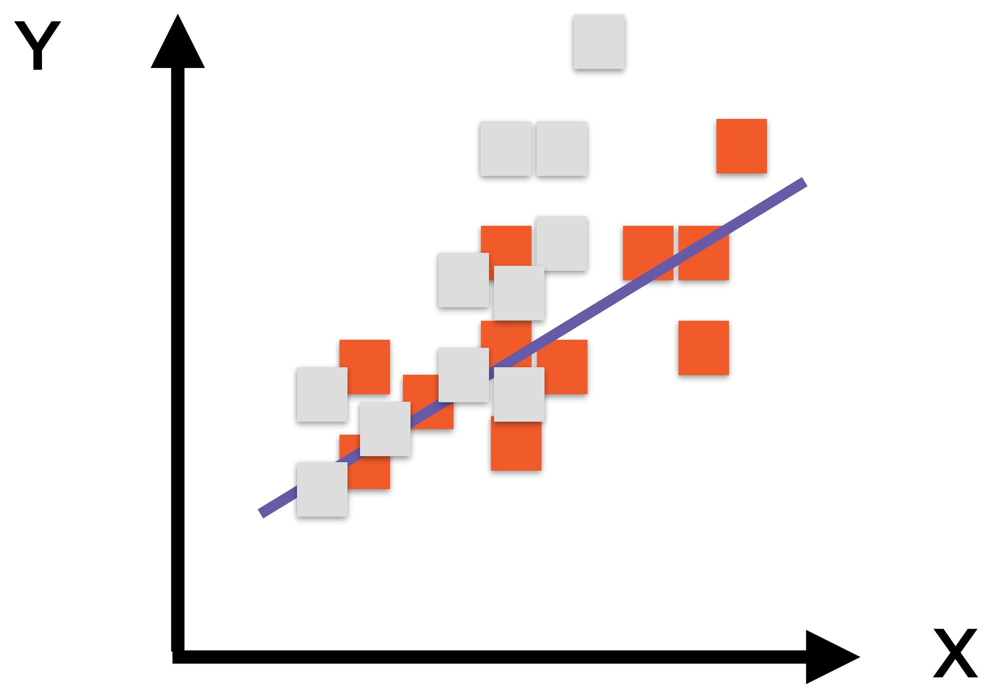
Call this the null hypotheses H_0

Null Hypotheses: α = 0



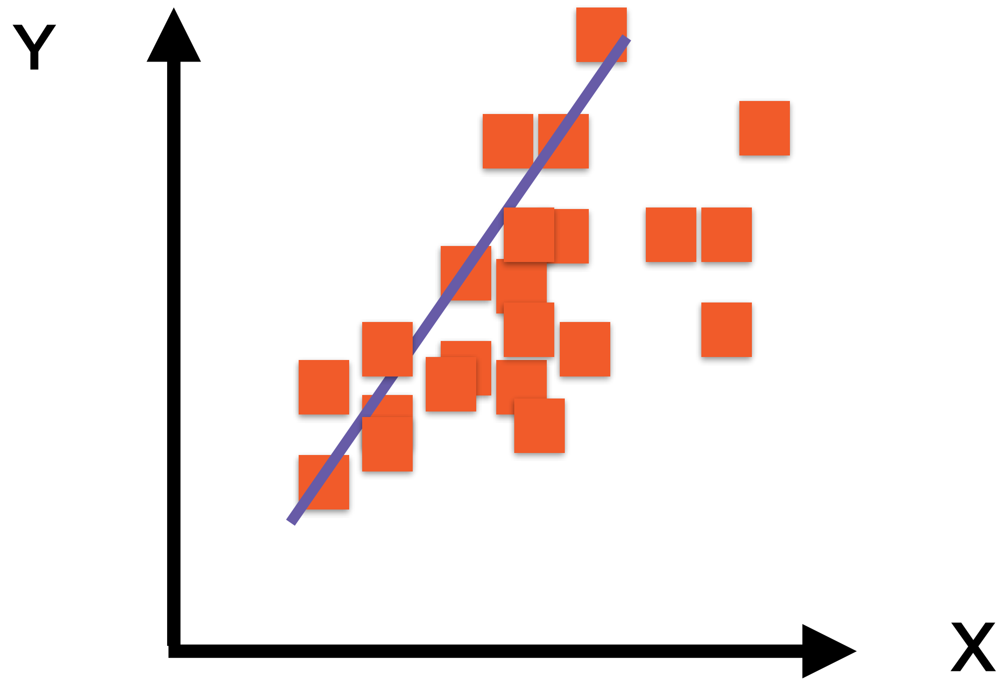
If this were actually true, how likely is it that our sample regression would yield the estimate α = A?

Why Zero?



Sample Regression Line

$$y = A + Bx$$

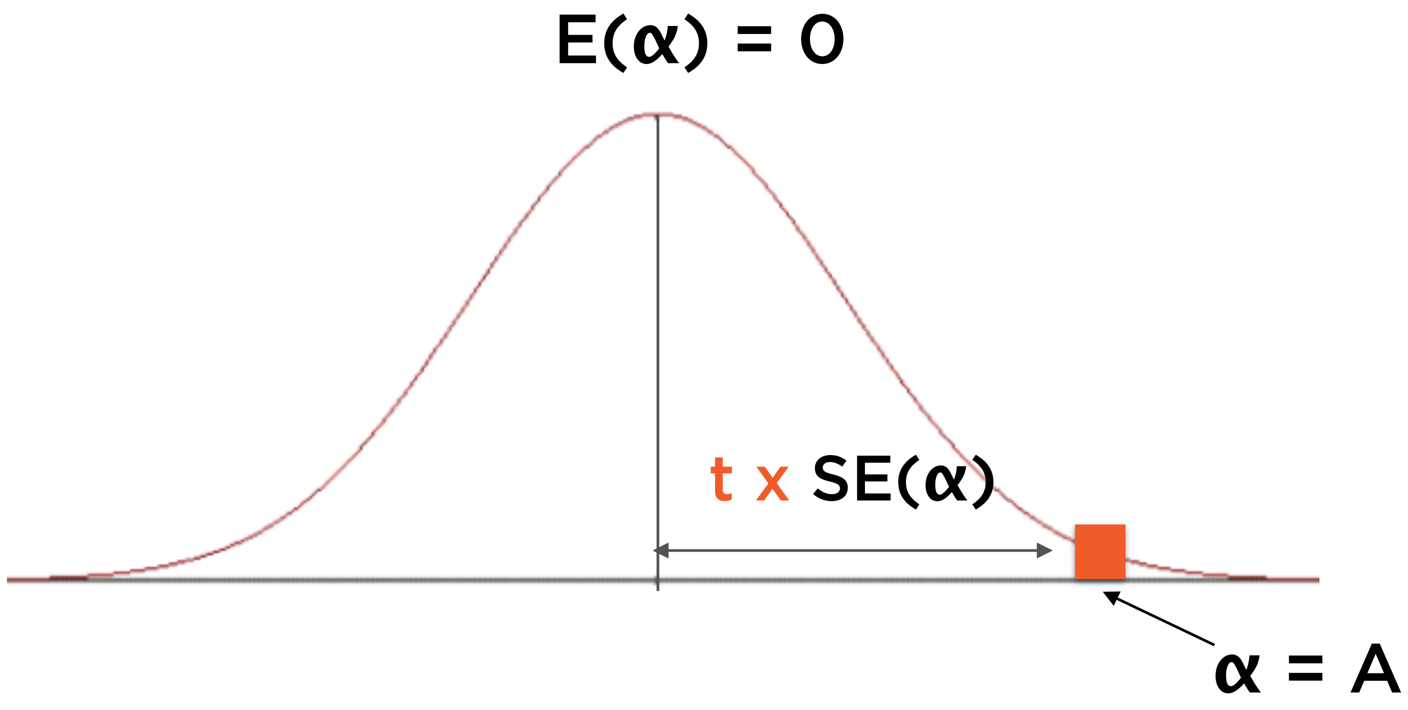


Population Regression Line

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

If $\alpha = 0$, it is adding no value in the regression line and should just be excluded

Null Hypotheses: $\alpha = 0$



The farther from the mean, the more unlikely that $\alpha = 0$

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

$$y = A + B_{S\&P500}X_1 + B_{USO}X_2$$

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
ESS	RSS	

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

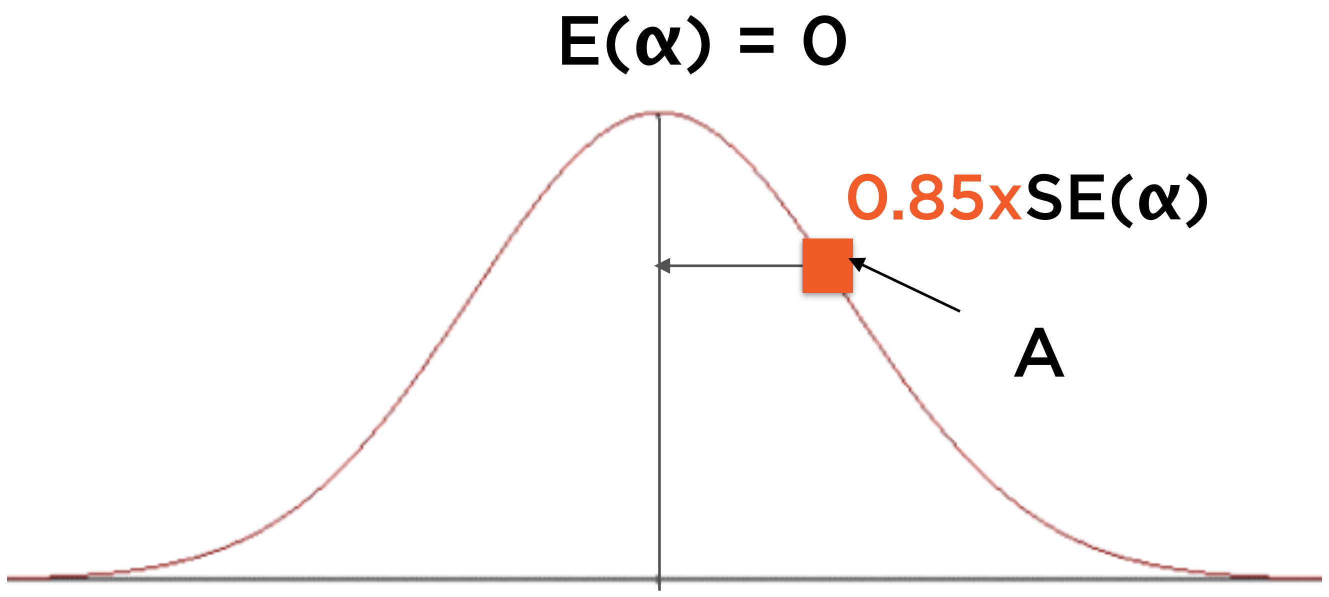
$$y = A + B_{S\&P500}x_1 + B_{USO}x_2$$

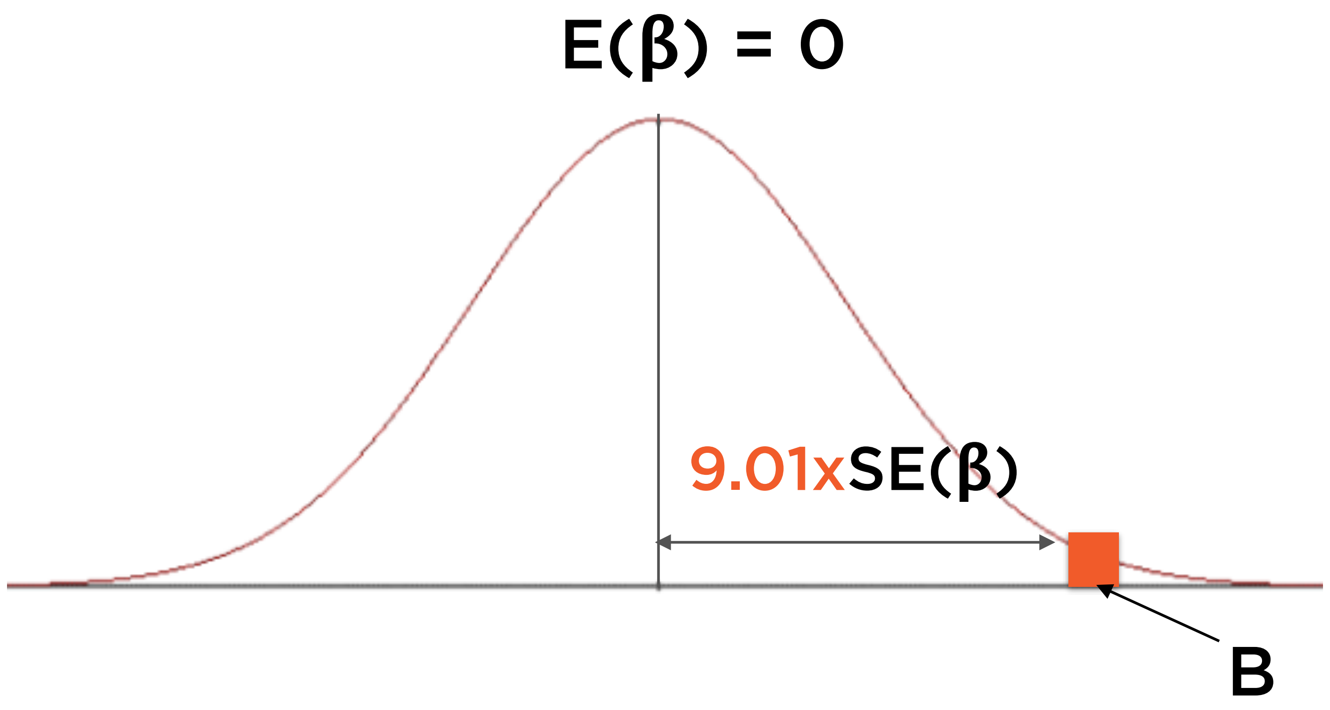
B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
ESS	RSS	

t-statistics

B_{USO} / SE_{USO}	$B_{S\&P500}$ / SE_{SP500}	A / SE_A

t-Statistics





$t\text{-stat}(\alpha) = 0.85$

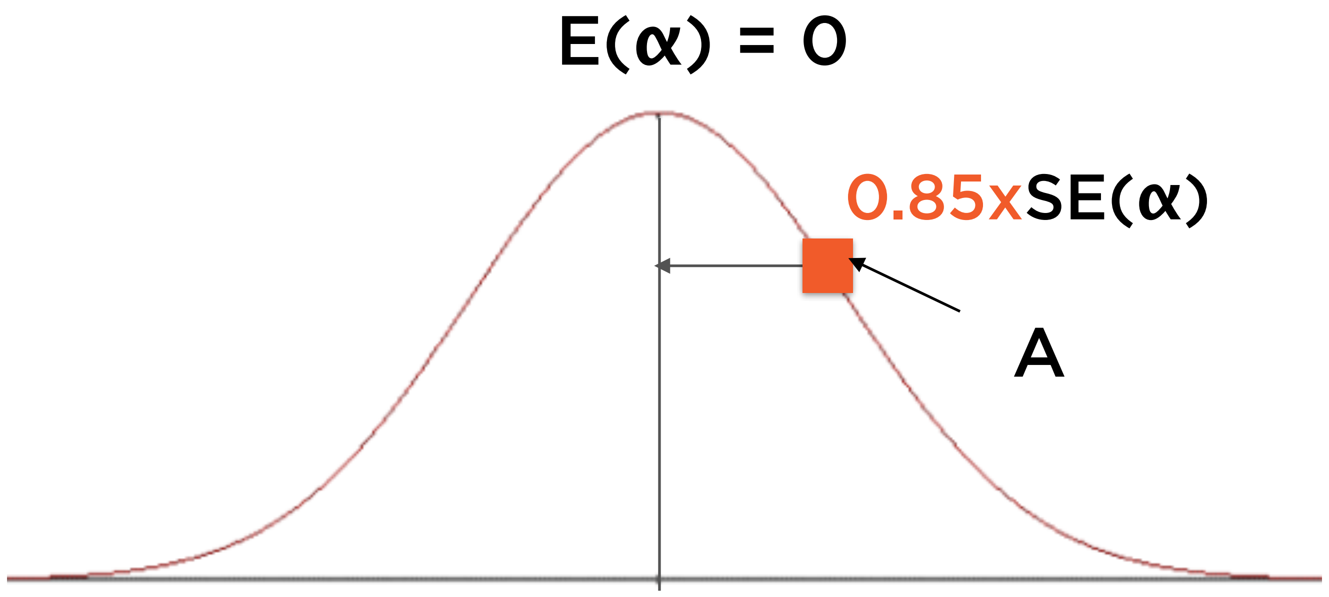
$t\text{-stat}(\alpha) = A/SE(\alpha)$

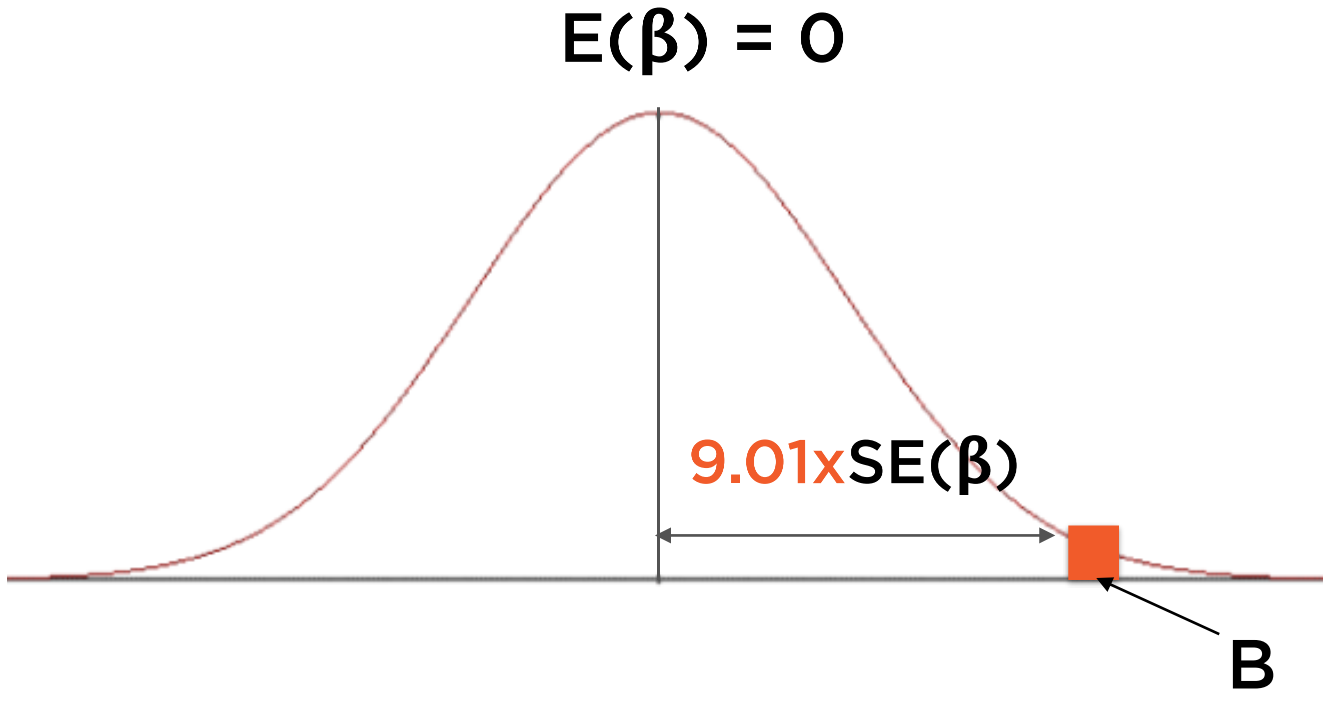
$t\text{-stat}(\beta) = 9.01$

$t\text{-stat}(\beta) = B/SE(\beta)$

We are now testing a hypothesis, that the population parameter is actually zero

t-Statistics





t-stat(α) = 0.85

t-stat(α) = A/SE(α)

t-stat(β) = 9.01

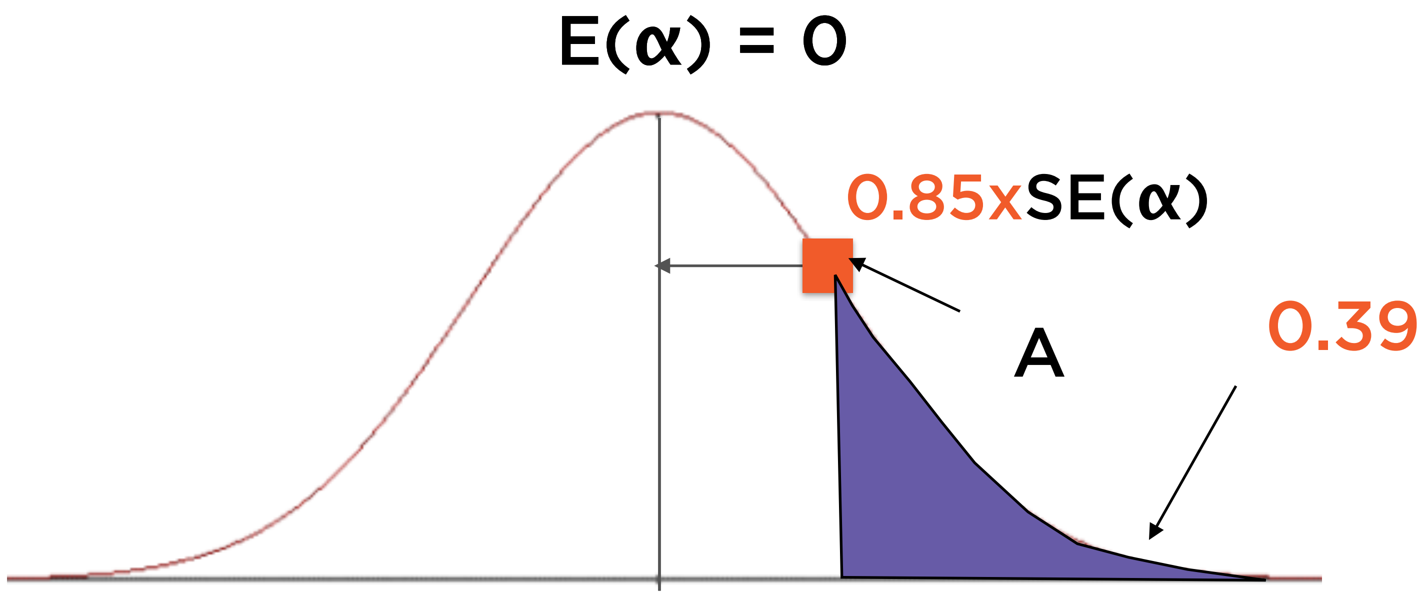
t-stat(β) = B/SE(β)

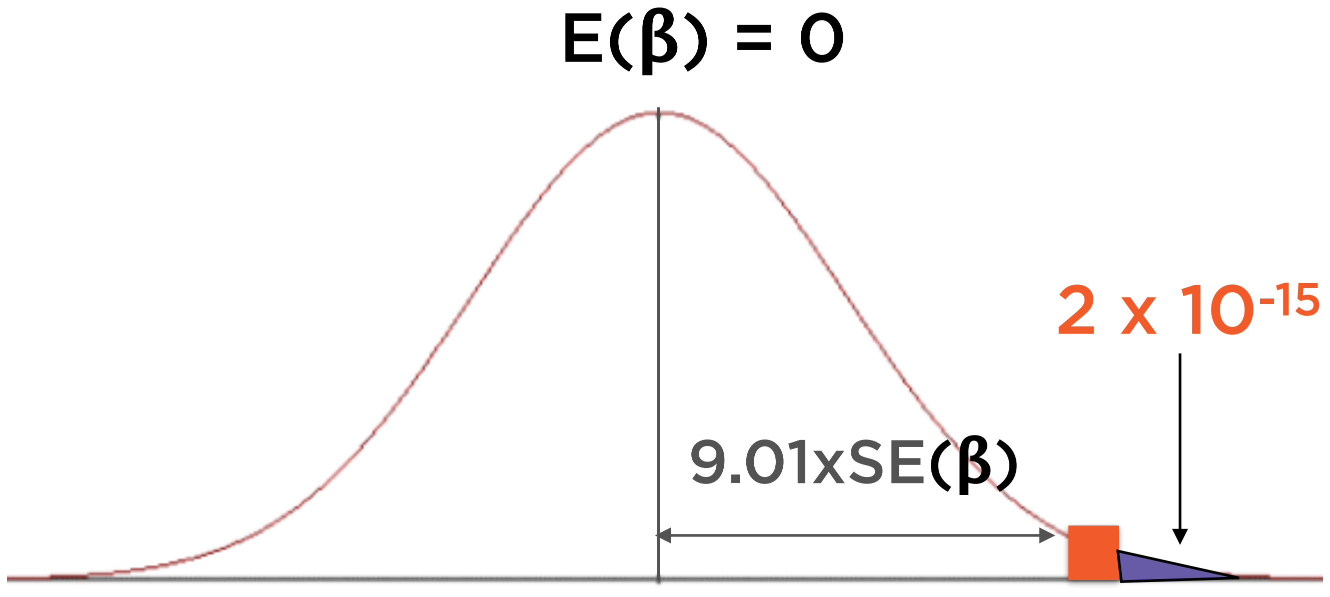
Is an individual estimate of A or B ‘adding value’ at all?

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(α) = 0.39

Low t-stat, high p-value

p-value(β) = 2 x 10^{-15} ~ 0

High t-stat, low p-value

Is an individual estimate of α or β ‘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

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
ESS	RSS	

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	**SER**	
F	d_f	
ESS	RSS	

Standard Error of Regression

Sample Regression Line

Regression Equation:

$$y = A + Bx$$

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

Sample Regression Line

Regression Equation:

$$y = A + Bx$$

Residuals

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

RSS = Variance(e)

Residual Variance *(RSS)*

Easily calculated from regression residuals

Population Regression Line

Regression Equation:

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

Errors

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

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

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
ESS	RSS	

Multiple Regressing Using **linest**

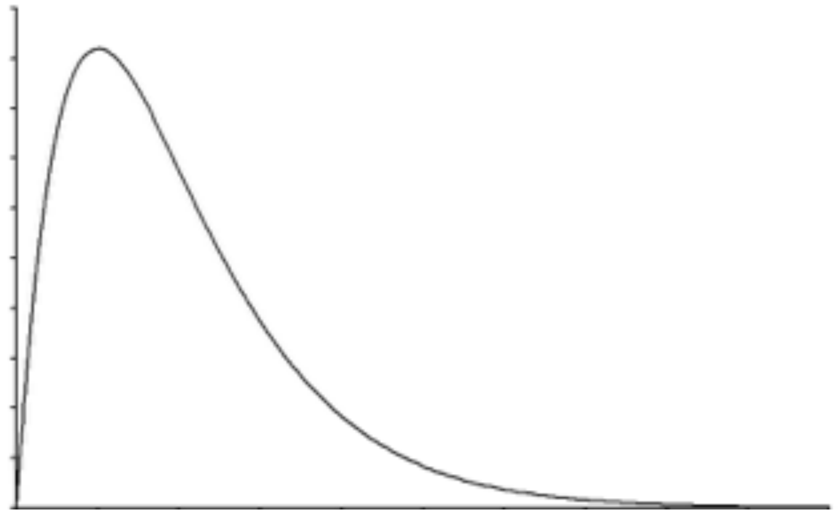
```
=linest(known_y's,[known_x's],[const],[stats])
```

$$y = A + B_{S\&P500}X_1 + B_{USO}X_2$$

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
$\boxed{F}$	d_f	
ESS	RSS	

F-statistic

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



χ^2 Distribution

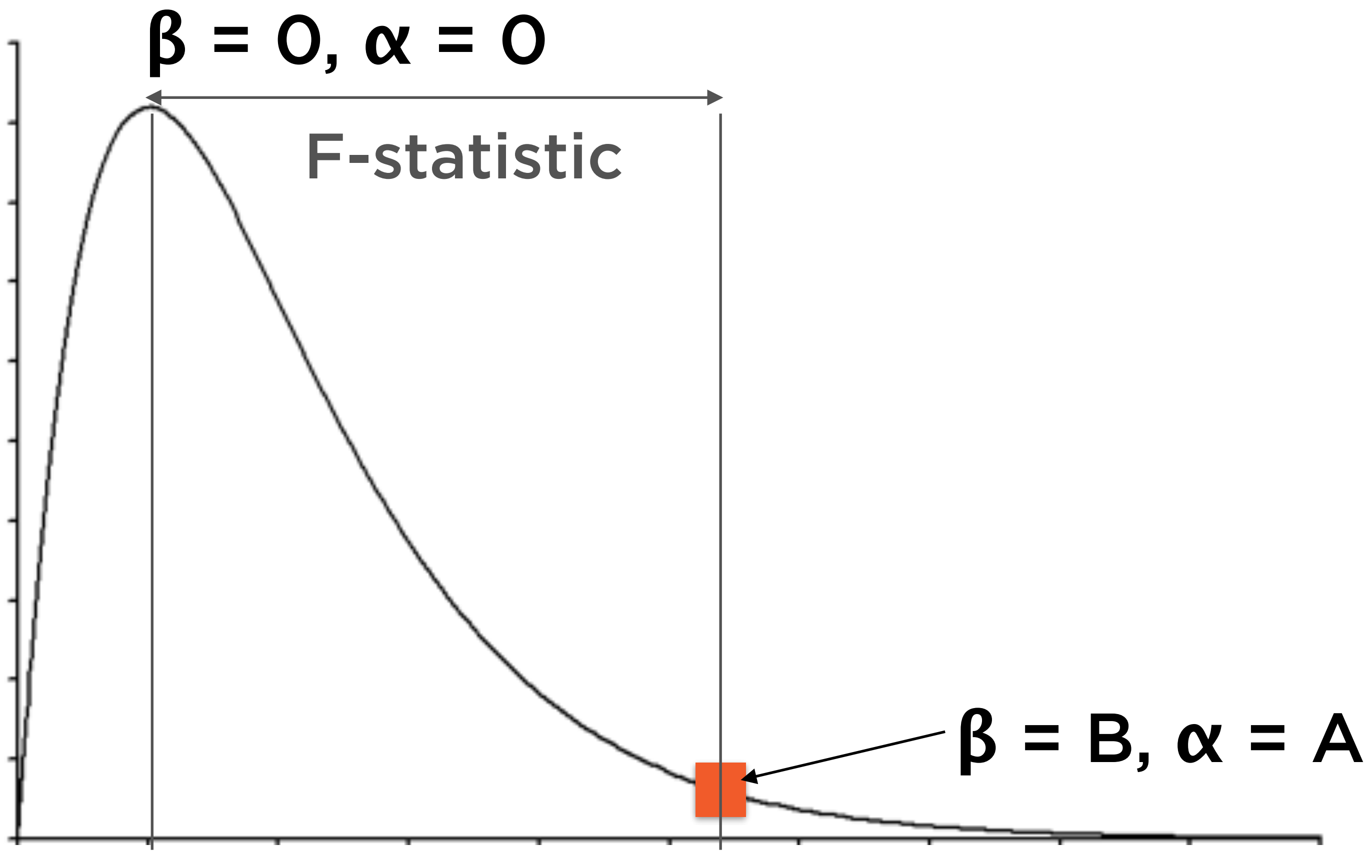
Never mind the fine print about degrees of freedom for now

Null Hypotheses

What if **all** population parameters were zero? i.e. $\beta = \alpha = 0$

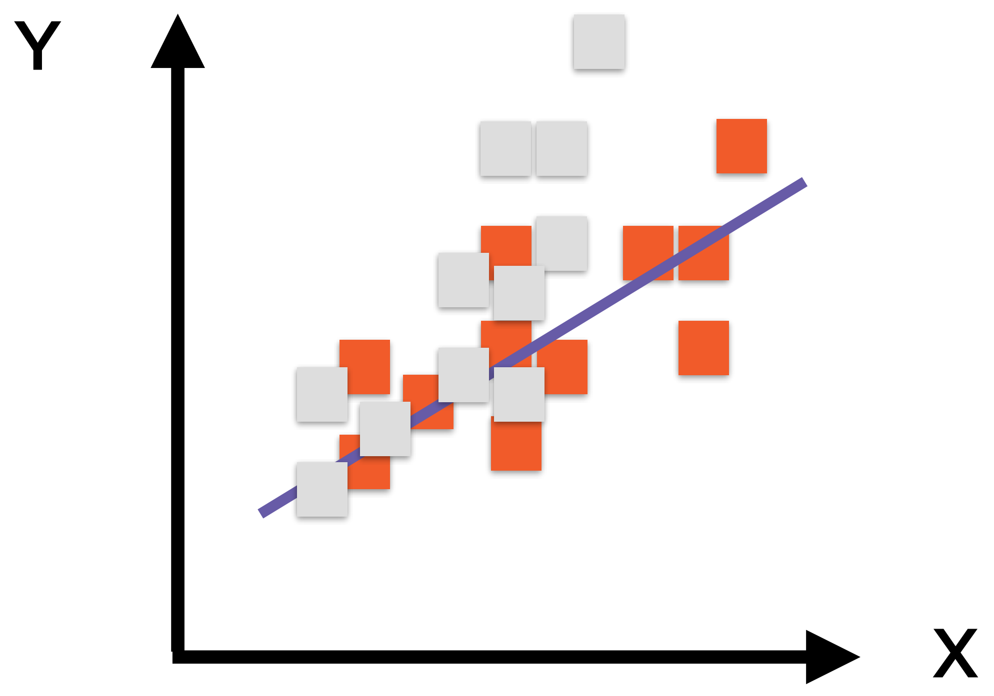
Call this the null hypotheses H_0

Null Hypotheses: $\beta = \alpha = 0$



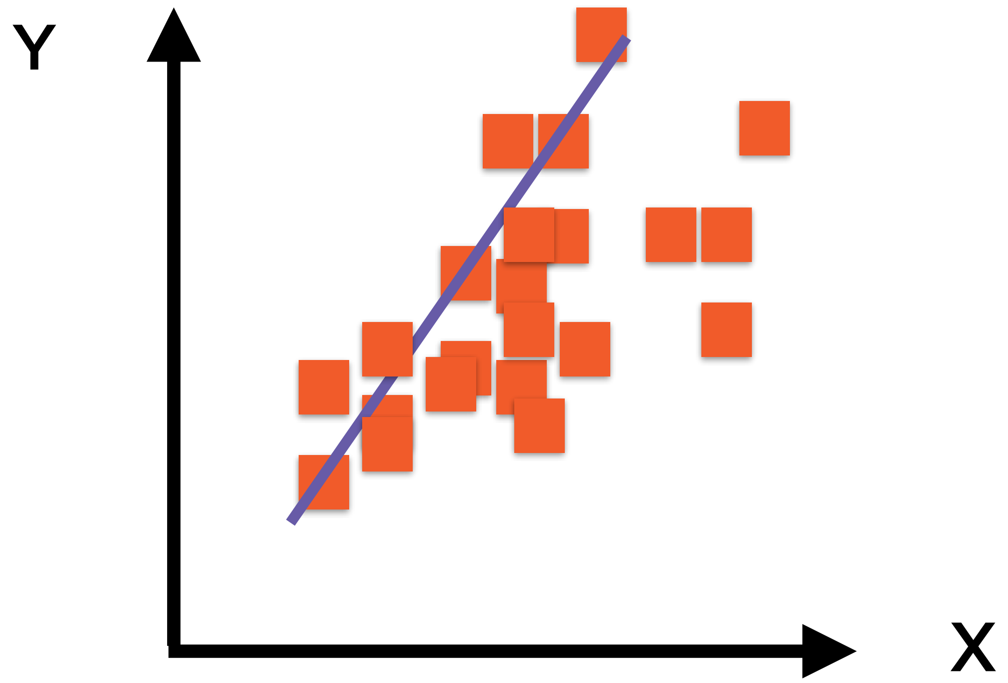
If this were actually true, how likely is it that our sample regression would yield the estimate $\beta = B, \alpha = A$?

Why Zero?



Sample Regression Line

$$y = A + Bx$$

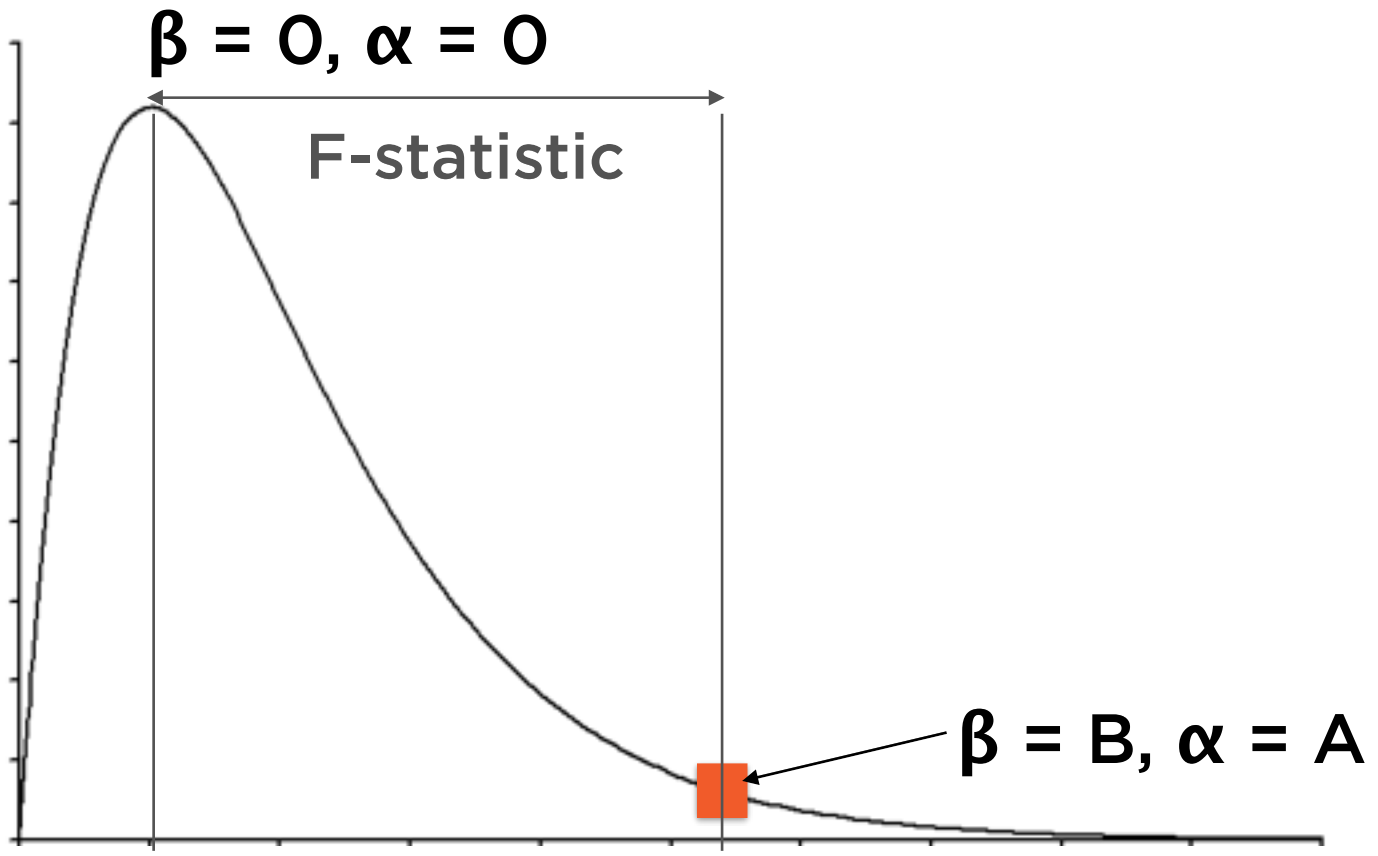


Population Regression Line

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

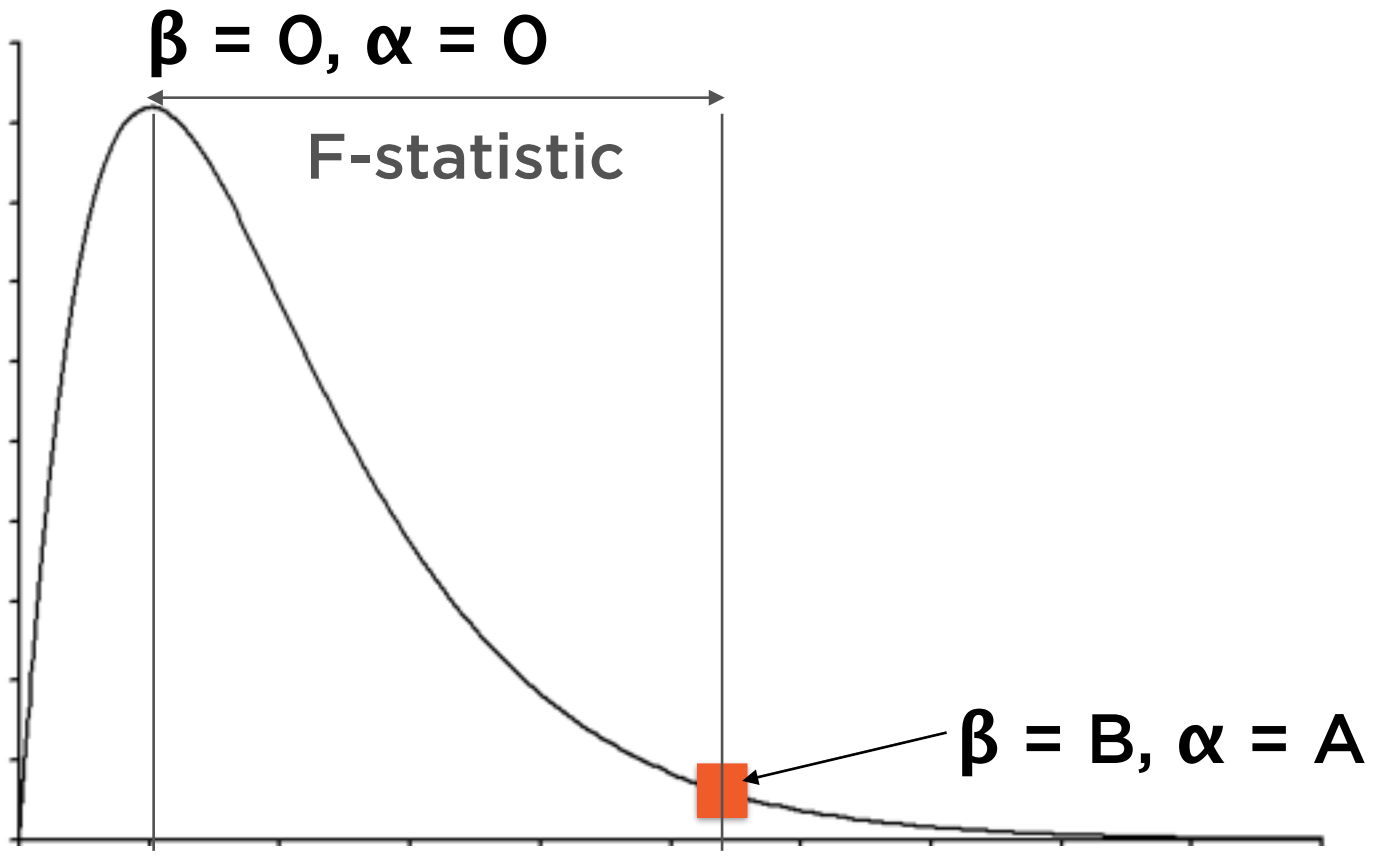
If $\alpha = \beta = 0$, our regression line is not adding any value at all

Null Hypotheses: $\alpha = 0$



The farther from the peak, the more unlikely that $\alpha = \beta = 0$

F-Statistic



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

High F-statistic => Yes

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

$$y = A + B_{S\&P500}X_1 + B_{USO}X_2$$

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	d_f	
ESS	RSS	

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
$\boxed{F}$	d_f	
ESS	RSS	

F-statistic

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

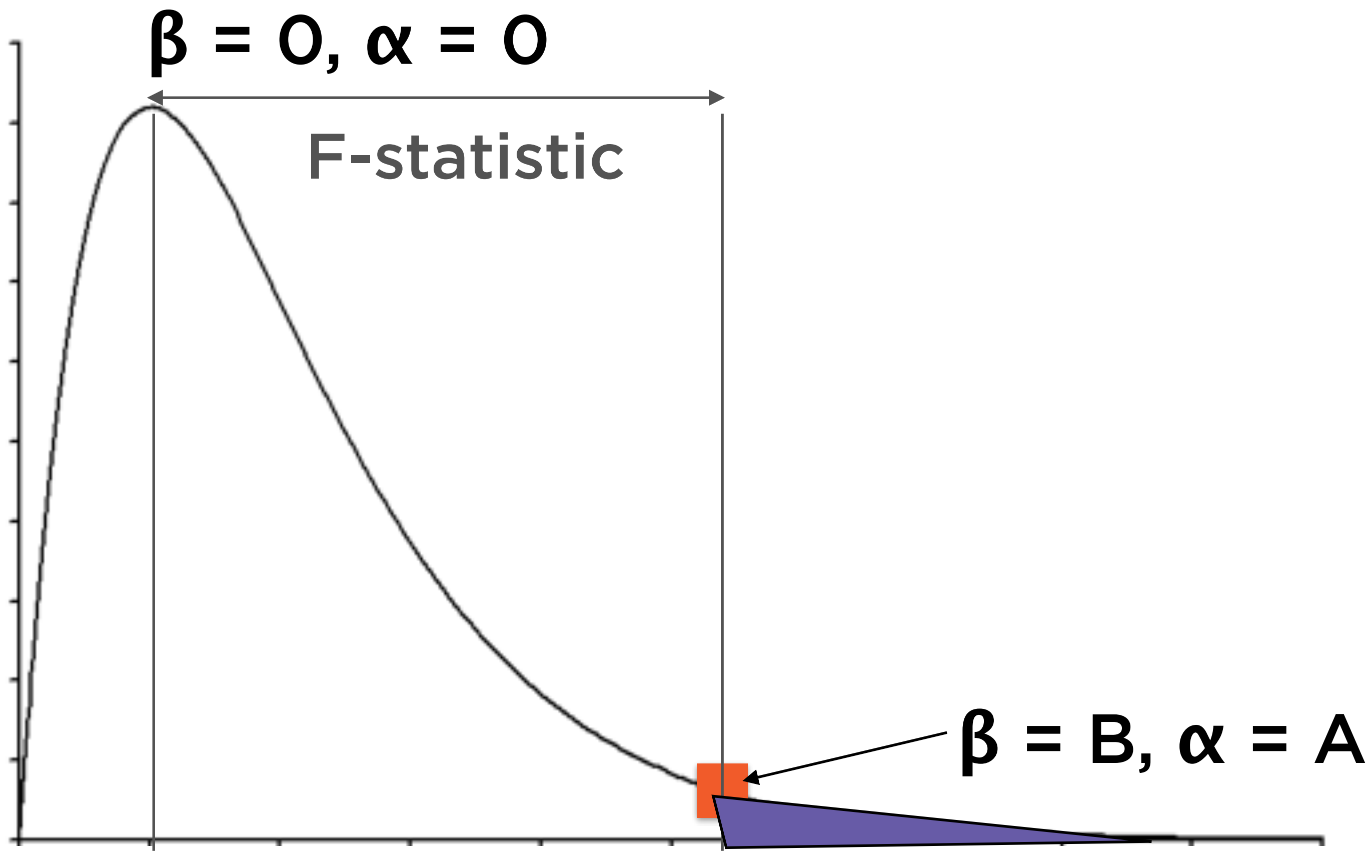
B_{USO}	$B_{S\&P500}$	A
SE_{USO}	$SE_{S\&P500}$	SE_A
R^2	SER	
F	$\boxed{d_f}$	
ESS	RSS	

Degrees of freedom = n - k - 1

n = number of points

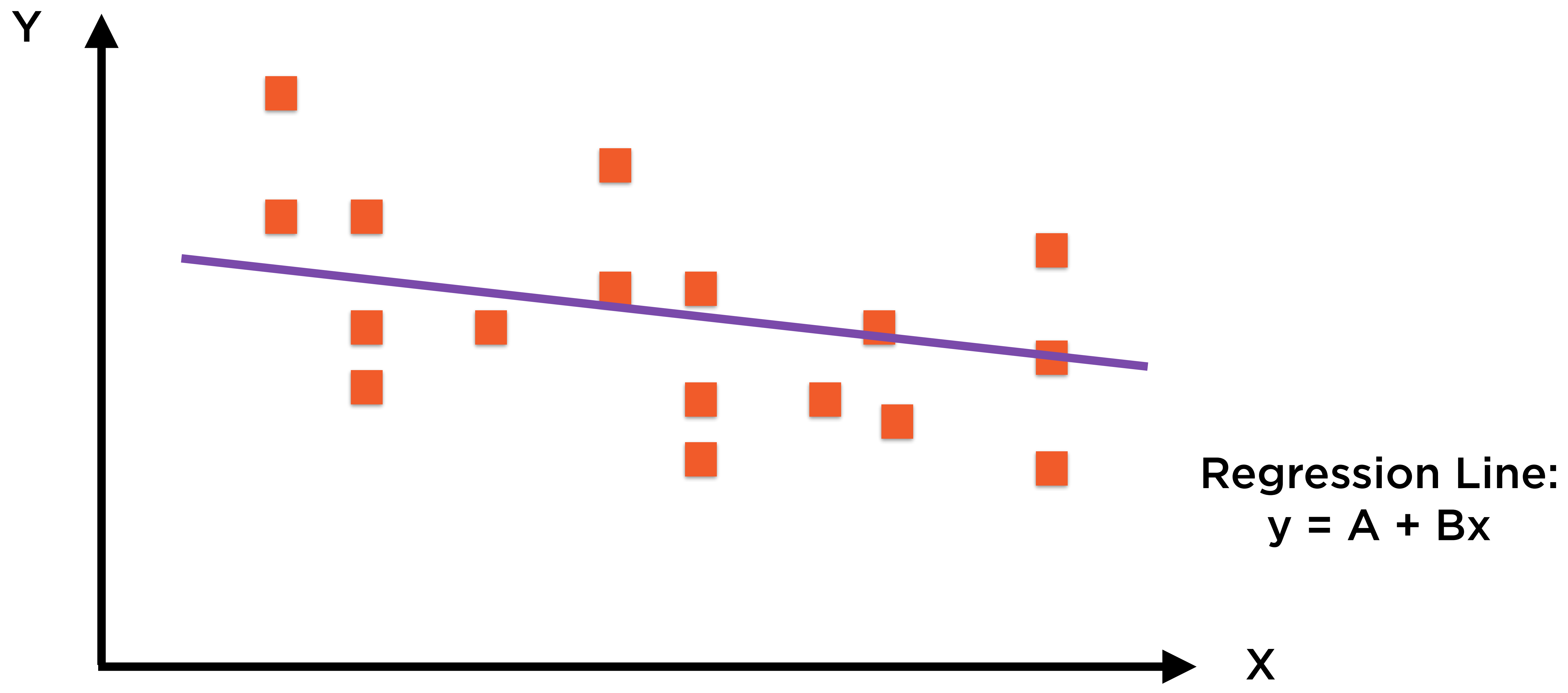
k = number of explanatory variables

F-Statistic to p-Value



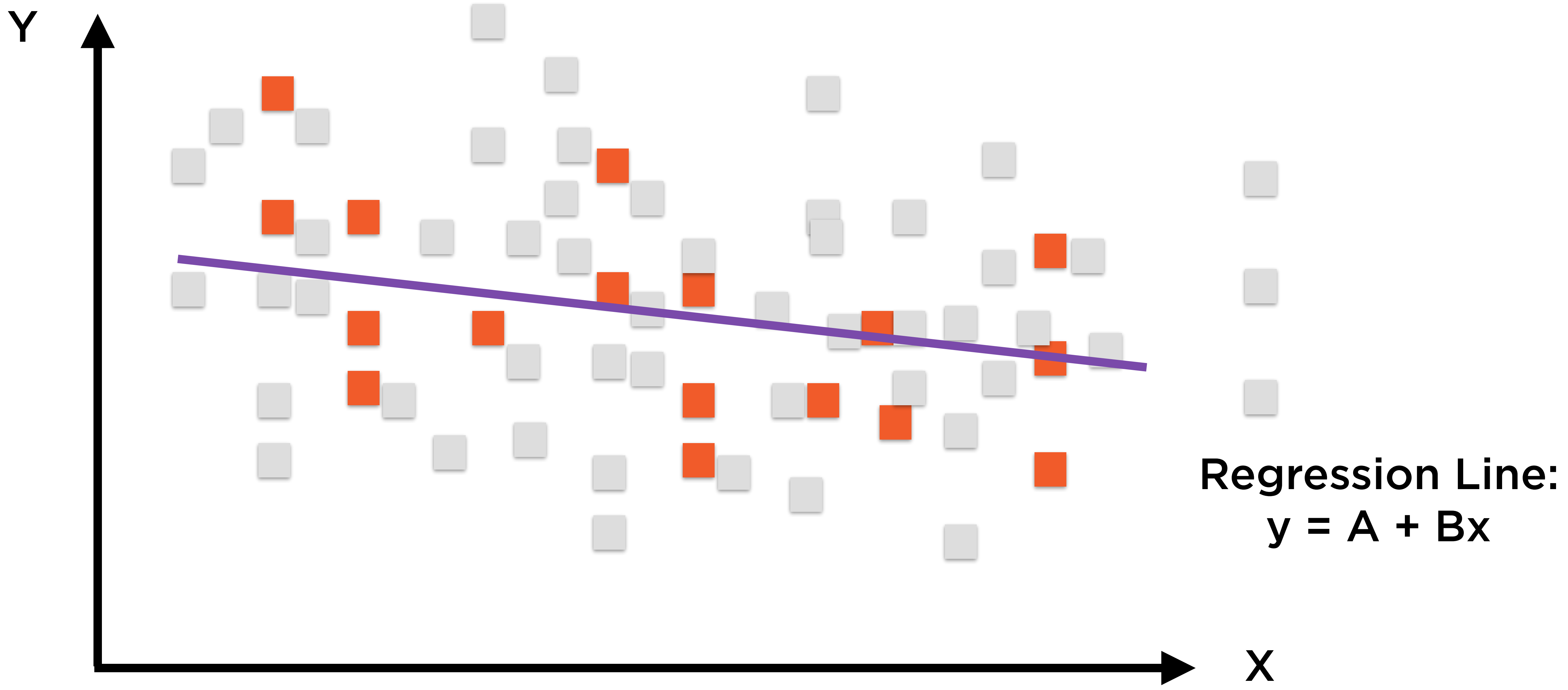
=fdist(F,n-d_f-1,d_f)

Regression Works on Samples



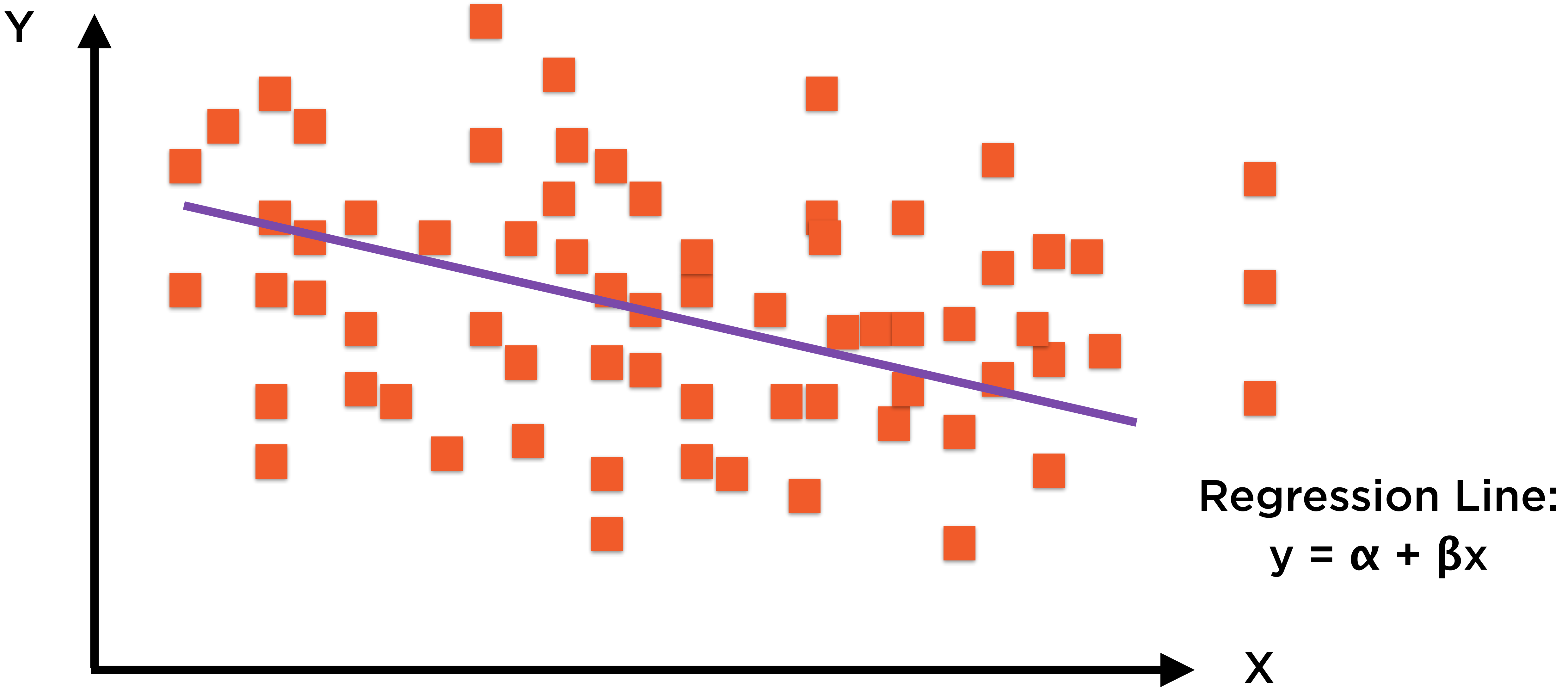
The regression line is based on a sample, not on the population

Regression Works on Samples



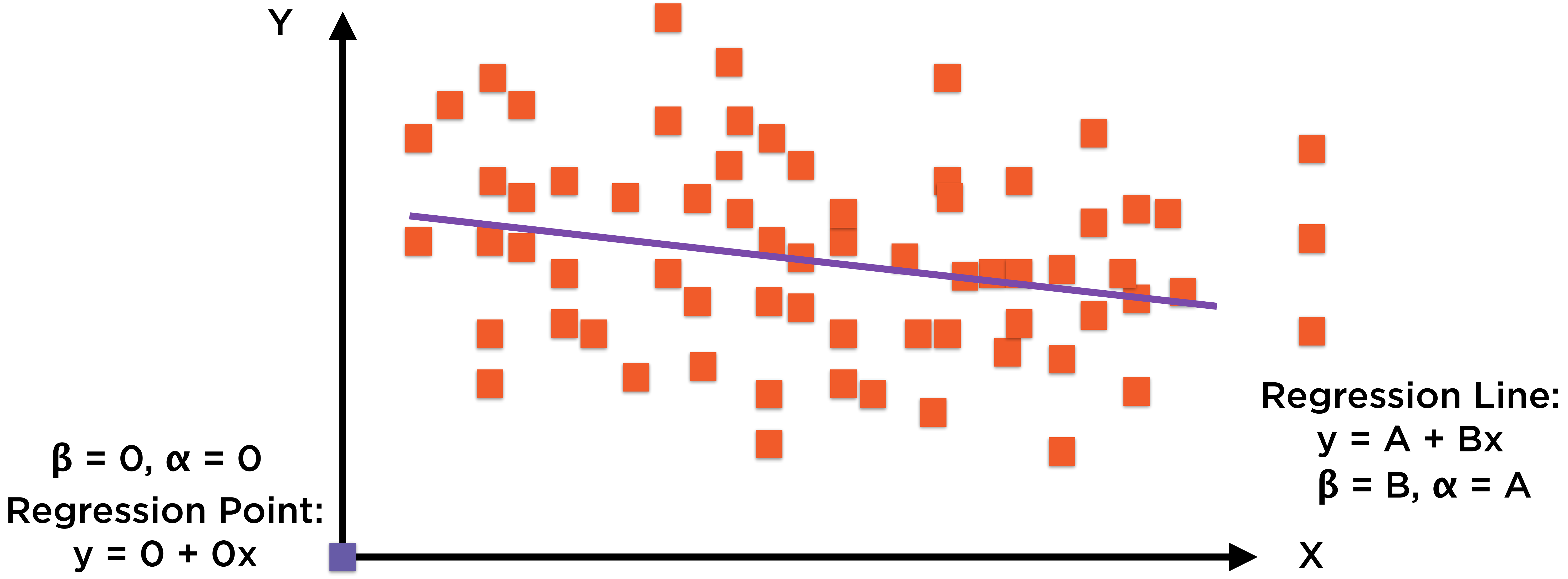
The regression line is based on a sample, not on the population

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’

Demo

Implement multiple regression in Excel

Multiple Regressing Using **linest**

```
=linest(known_y’s,[known_x’s],[const],[stats])
```

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

y = Returns on Exxon stock (XOM)

x_1 = Returns on S&P 500

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

DATE	XOM
2016-12-01	1.5%
2016-11-01	-0.9%
2006-01-01	0.5%

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

DATE	S&P 500
2016-12-01	1.2%
2016-11-01	-1.1%
2006-01-01	0.7%

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

TRUE

If TRUE

y = A + Bx

else

y = Bx

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

TRUE

If TRUE, detailed regression statistics are displayed

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

$B_{S\&P500}$	A
$SE_{S\&P500}$	SE_A
R^2	SER
F	d_f
ESS	RSS

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

$B_{S\&P500}$	$\boxed{A}$	Intercept A
$SE_{S\&P500}$	SE_A	
R^2	SER	
F	d_f	
ESS	RSS	

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

Slope

$B_{S\&P500}$	A
$SE_{S\&P500}$	SE_A
R^2	SER
F	d_f
ESS	RSS

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

$B_{S\&P500}$	A
$SE_{S\&P500}$	SE_A
R^2	SER
F	d_f
ESS	RSS

Standard Errors

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

$B_{S\&P500}$	A
$SE_{S\&P500}$	SE_A
$\boxed{R^2}$	SER
F	d_f
ESS	RSS

R^2

(not adjusted-R^2)

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

$B_{S\&P500}$	A
$SE_{S\&P500}$	SE_A
R^2	**SER**
F	d_f
ESS	RSS

Standard Error of Regression

Multiple Regressing Using **linest**

```
=linest(known_y’s,[known_x’s],[const],[stats])
```

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

$B_{S\&P500}$	A
$SE_{S\&P500}$	SE_A
R^2	SER
F	d_f
ESS	RSS

F-statistic

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

$B_{S\&P500}$	A
$SE_{S\&P500}$	SE_A
R^2	SER
F	$\boxed{d_f}$
ESS	RSS

Degrees of freedom = n - k - 1

n = number of points

k = number of explanatory variables

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

$B_{S\&P500}$	A
$SE_{S\&P500}$	SE_A
R^2	SER
F	d_f
ESS	RSS

Explained Sum of Squares

Multiple Regressing Using **linest**

```
=linest(known_y's,[known_x's],[const],[stats])
```

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

$B_{S\&P500}$	A
$SE_{S\&P500}$	SE_A
R^2	SER
F	d_f
ESS	**RSS**

Residual Sum of Squares

Extending Multiple Regression to Categorical Variables

A Simple Regression

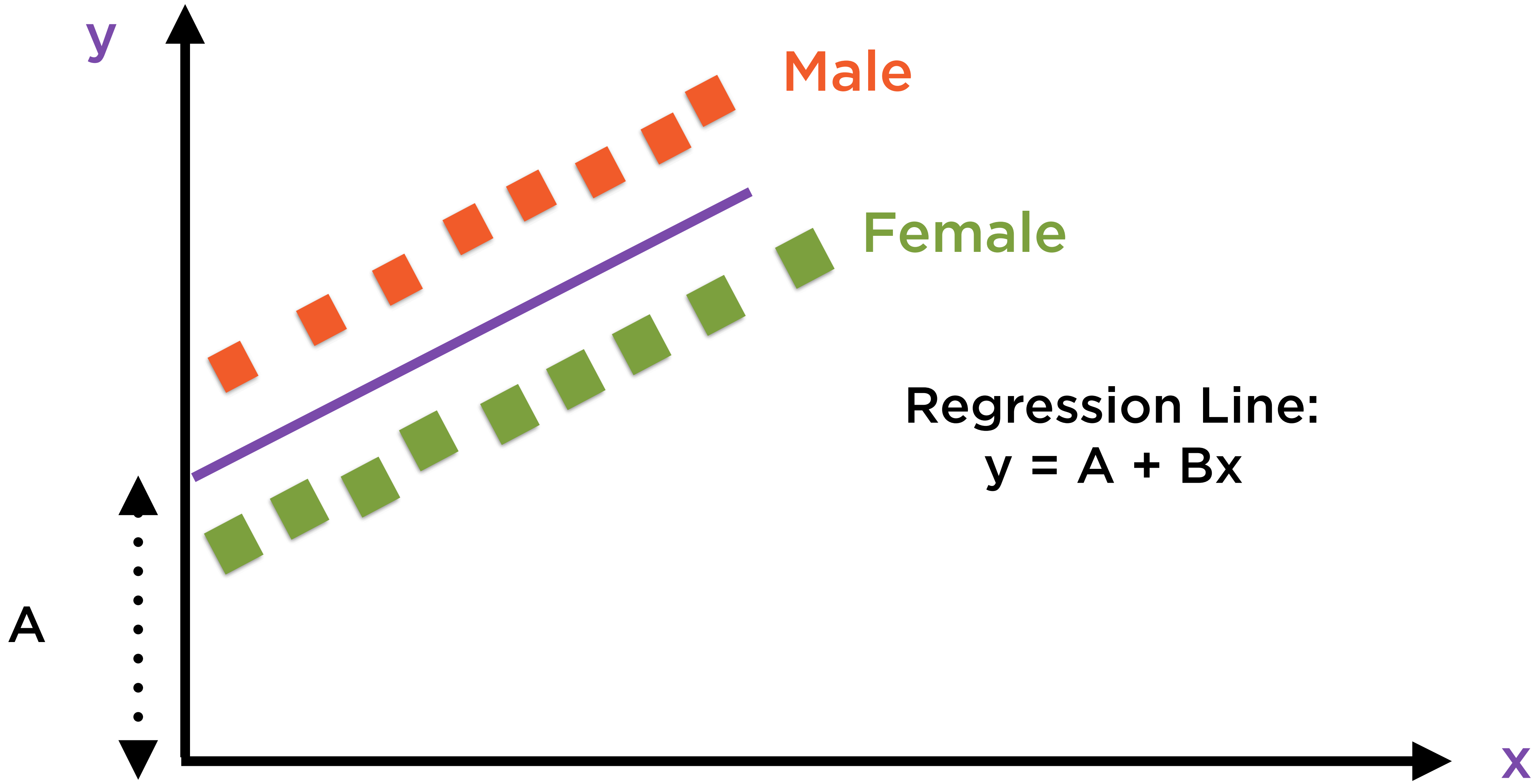
Proposed Regression Equation:

$$y = A + Bx$$

Height of individual

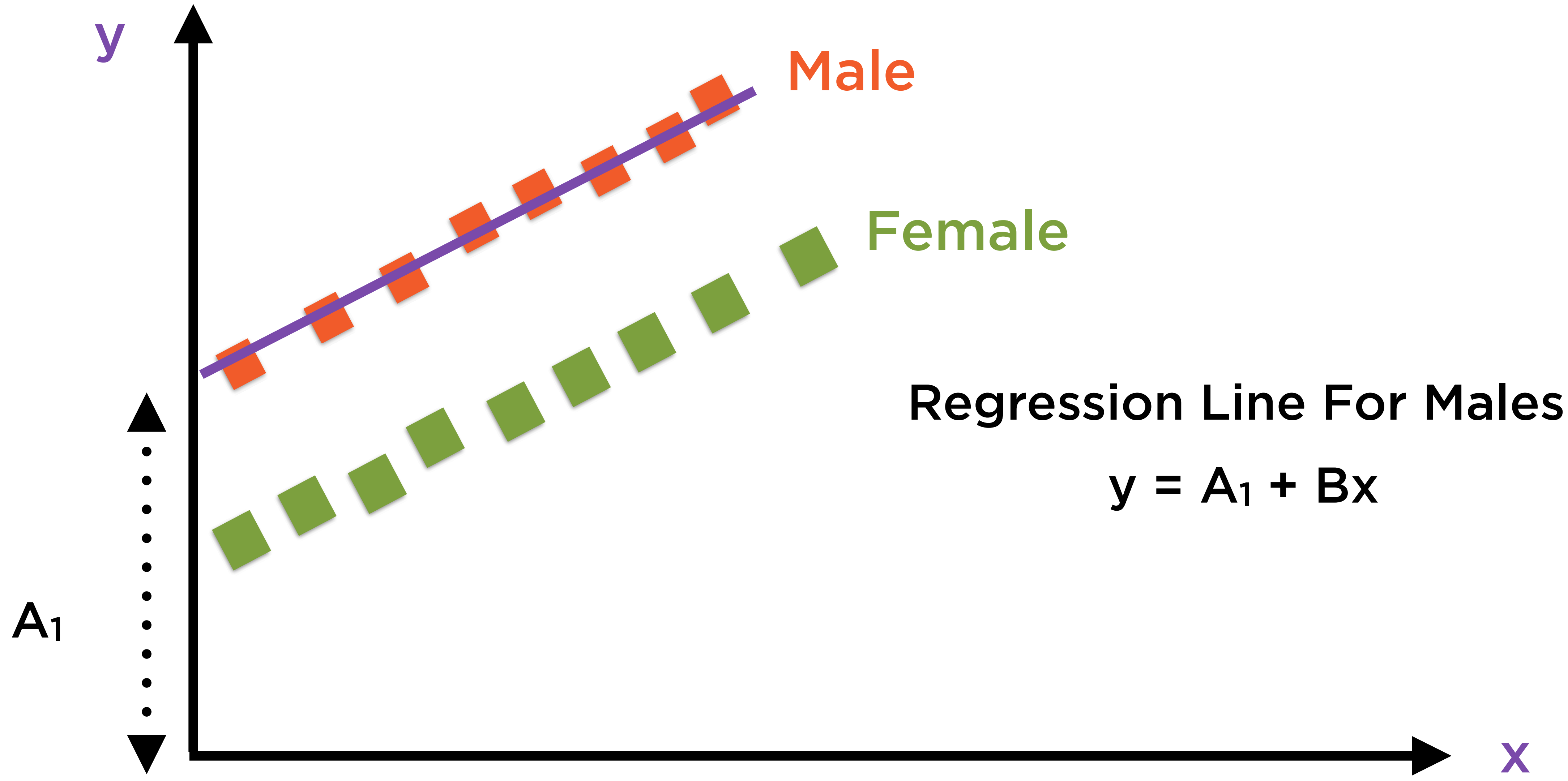
Average height of parents

A Simple Regression



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

A Simple Regression

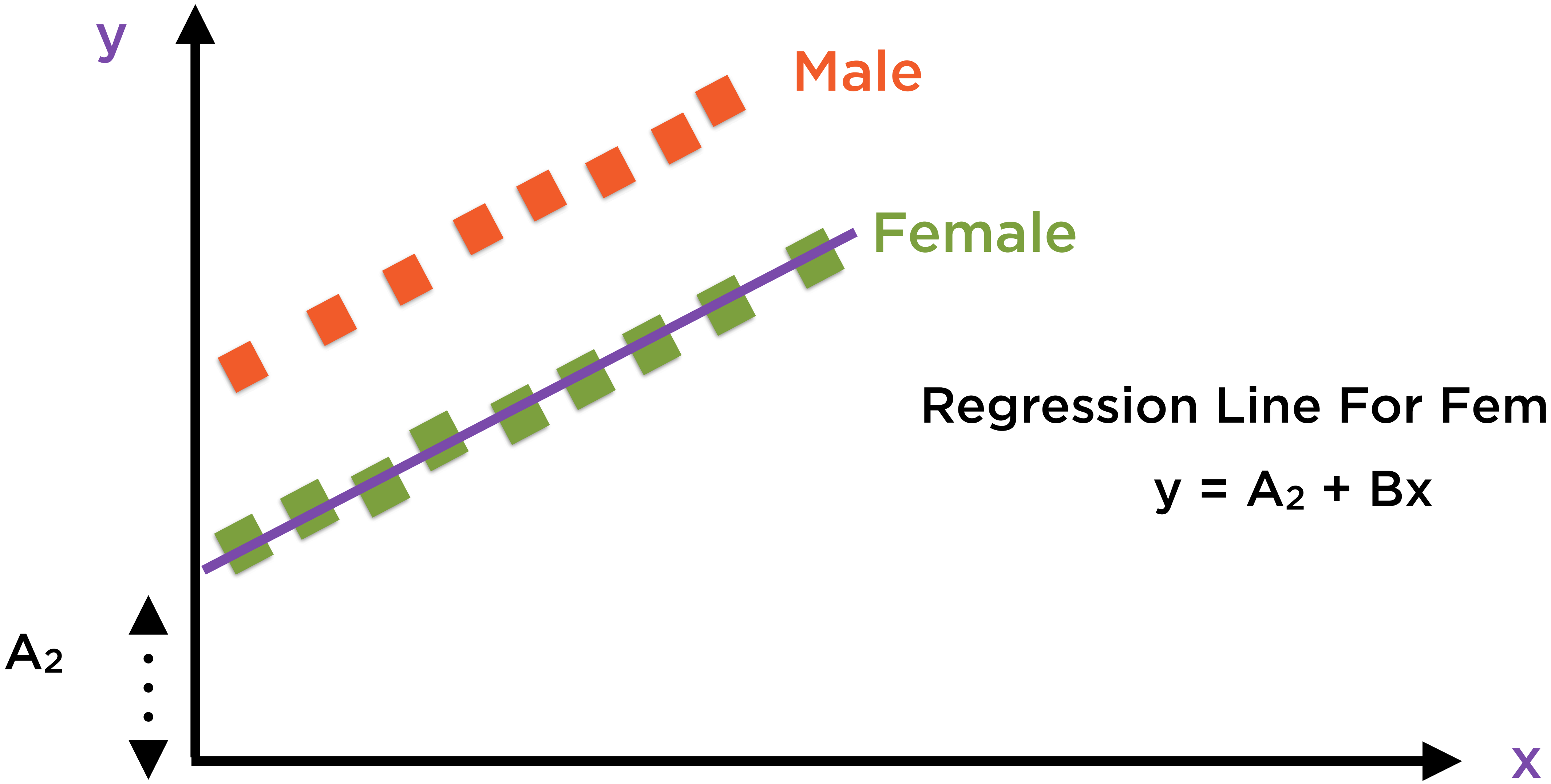


Regression Line For Males:

$$y = A_1 + Bx$$

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

A Simple Regression

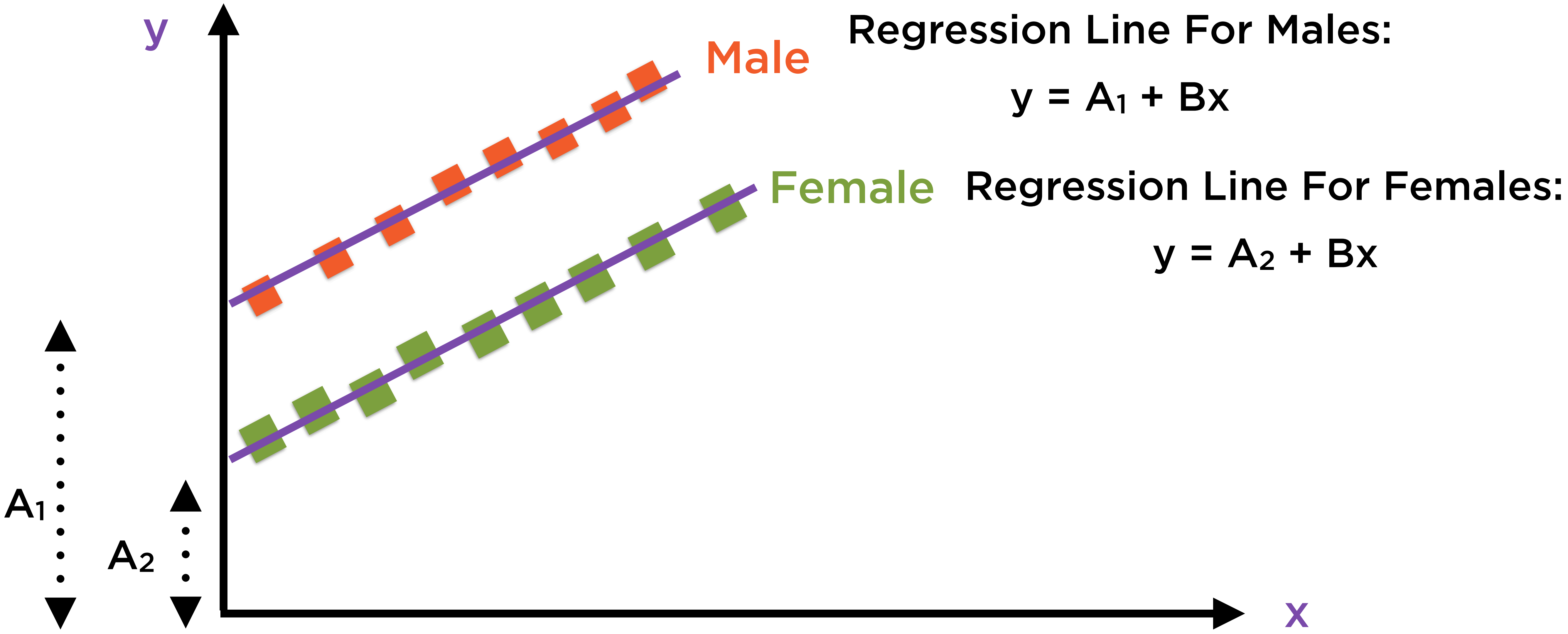


Regression Line For Females:

$$y = A_2 + Bx$$

...and another great fit for females

A Simple Regression



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:

$$y = A_1 D_1 + A_2 D_2 + Bx$$

$D_1 = 1$ 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

$$y = \cancel{A_1x0} + A_2x1 + Bx$$

$$= A_2 + Bx$$

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

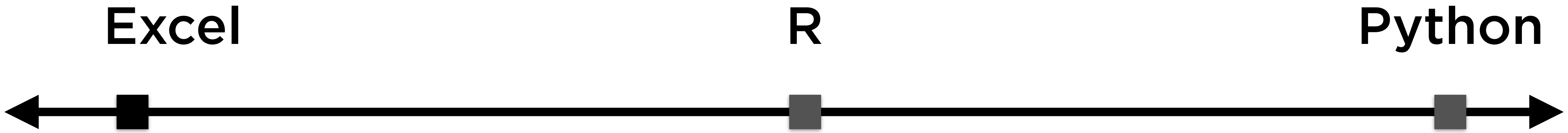
$= 0$ for males

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

Demo

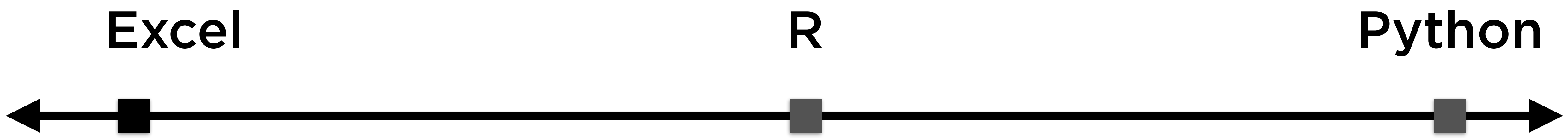
Perform regression with categorical variables in Excel

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 Excel

Interpreted results of a multiple regression

Carried out multiple regression in Excel to include categorical variables