DATA SCIENCE
CLASS 2: LINEAR REGRESSION

0. BASIC FORM
II. CATEGORICAL VARIABLES

0. BASIC FORM

Q: What is the motivation for learning about linear regression?

- widely used
- runs fast
- easy to use (not a lot of tuning required)
- highly interpretable
- basis for many other methods

	continuous	categorical
supervised	????	?????
unsupervised	????	?????

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

Q: What is a regression model?
A: A functional relationship between input & a continuous response variable.

Q: What is a regression model?
A: A functional relationship between input & a continuous response variable.

The simple linear regression model captures a linear relationship between a single input variable x and a response variable y:

Q: What is a regression model?
A: A functional relationship between input & a continuous response variable.

The simple linear regression model captures a linear relationship between a single input variable x and a response variable y:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Q: What do the terms in this model mean?

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Q: What do the terms in this model mean?

$$y = \beta_0 + \beta_1 x + \varepsilon$$

A: y = response variable (the one we want to predict)

Q: What do the terms in this model mean?

$$y = \beta_0 + \beta_1 x + \varepsilon$$

A: y = response variable (the one we want to predict)

x = input variable (the one we use to train the model)

Q: What do the terms in this model mean?

$$y = \beta_0 + \beta_1 x + \varepsilon$$

A: y = response variable (the one we want to predict)

x = input variable (the one we use to train the model)

β_0 = intercept (where the line crosses the y-axis)

Q: What do the terms in this model mean?

$$y = \beta_0 + \beta_1 x + \varepsilon$$

A: y = response variable (the one we want to predict)

x = input variable (the one we use to train the model)

β_0 = intercept (where the line crosses the y-axis

β_1 = regression coefficient (the model parameter)

Q: What do the terms in this model mean?

$$y = \beta_0 + \beta_1 x + \varepsilon$$

A: y = response variable (the one we want to predict)

x = input variable (the one we use to train the model)

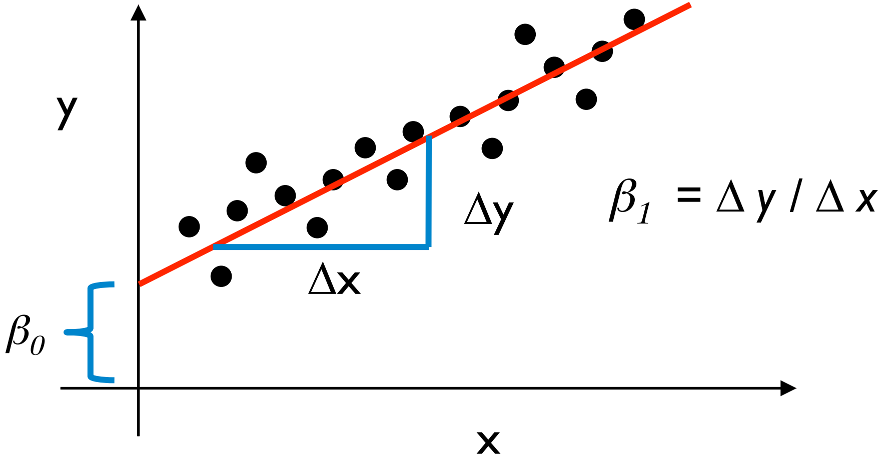
β_0 = intercept (where the line crosses the y-axis

β_1 = regression coefficient (the model parameter)

ε = residual (the error)

Q: What do the terms in this model mean?

$$y = \beta_0 + \beta_1 x + \varepsilon$$



We can extend this model to several input variables, giving us the multiple linear regression model:

We can extend this model to several input variables, giving us the multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

II. CATEGORICAL VARIABLES

Q: How do we deal with categorical variables? (i.e., with k levels)

Major (k=4)
Computer Science
Engineering
Business
Literature
Business
Engineering

Q: How do we deal with categorical variables? (i.e., with k levels)

A: Create a k-1 binary ("dummy") variables.

Major (k=4)
Computer Science
Engineering
Business
Literature
Business
Engineering

Engineering	Business	Literature
0	0	0
1	0	0
0	1	0
0	0	1
0	1	0
1	0	0

Computer Science is the reference

Q: Why k-1 and not k?

A: Because k-1 captures all possible outputs, and to avoid multicollinearity.

Q: Why k-1 and not k?
A: Because k-1 captures all possible outputs, and to avoid multicollinearity.

Q: Does it matter which factor level I leave out?
A: Yes, this is the reference point for all other factor levels.

Q: Is this the only way to represent categorical data?

A: This is the conventional way to represent nominal data, however, ordinal data can be represented with integers.

Q: Is this the only way to represent categorical data?
A: This is the conventional way to represent nominal data, however, ordinal data can be represented with integers.

Q: What does this mean?
A: Categories that can be ranked (i.e., strongly disagree, disagree, neutral, agree, strongly agree) can be represented as 1, 2, 3, 4, 5.