GENERAL ASSEMBLY

DATA SCIENCE

CLASS 5: LINEAR REGRESSION

0. BASIC FORM
I. ESTIMATING COEFFICIENTS
II. CATEGORICAL VARIABLES
III. MAKING INFERENCES

0. BASIC FORM

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

Q: Where does logistic regression belong in this diagram?

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

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*

Q: What is a **regression** *model?*

A: A functional relationship between input & response variables.

Q: What is a **regression** *model?*

A: A functional relationship between input & response variables.

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 & response variables.

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)*

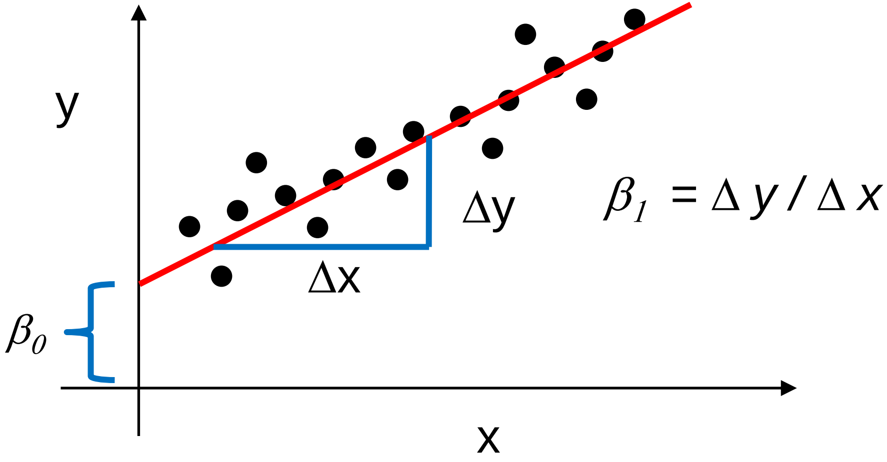
β_0 = **intercept** *(where the line crosses the y-axis)*

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:*

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

I. ESTIMATING COEFFICIENTS

*Q: How to determine the **impact** of a particular input variable on the response variable?*

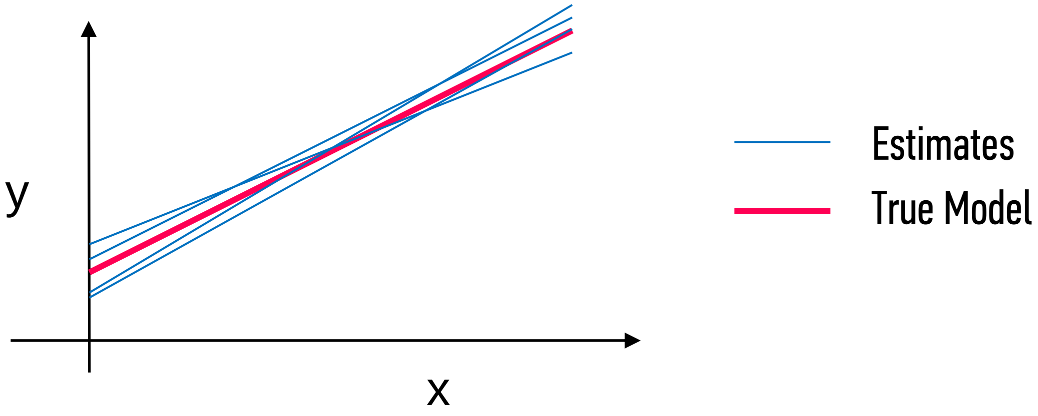
A: The coefficient estimates $(\hat{\beta})$

Q: What is meant by estimates?

A: We are making an inference based off of a sample.

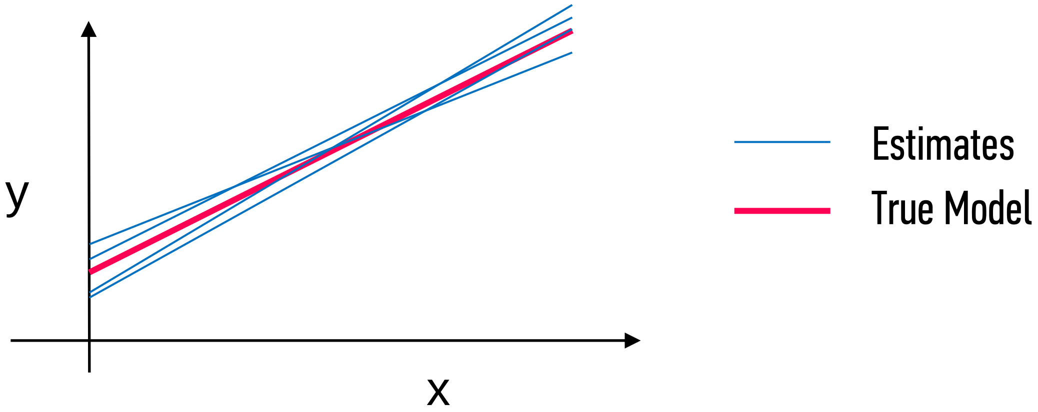
Q: What is meant by estimates?

A: We are making an inference based off of a sample.



Q: What is meant by estimates?

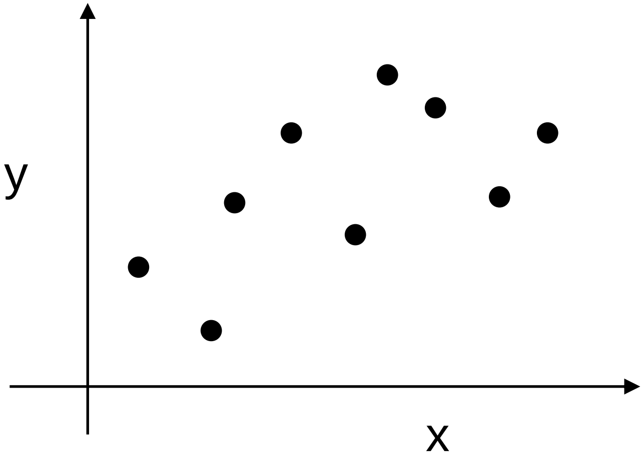
A: We are making an inference based off of a sample.



A fundamental part of statistics is quantifying our confidence that our estimates are reflective of truth.

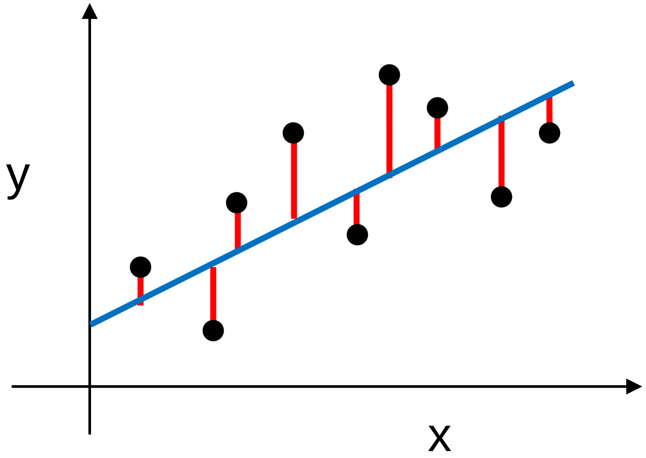
*Q: How to **estimate** coefficients for a linear model?*

A: By finding the line that minimizes the sum of squared residuals.



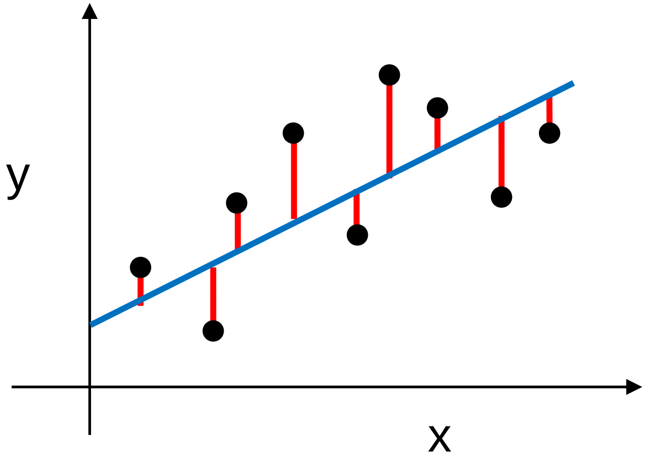
Q: How to ***estimate*** *coefficients for a linear model?*

A: By finding the line that minimizes the sum of squared residuals.



*Q: How to **estimate** coefficients for a linear model?*

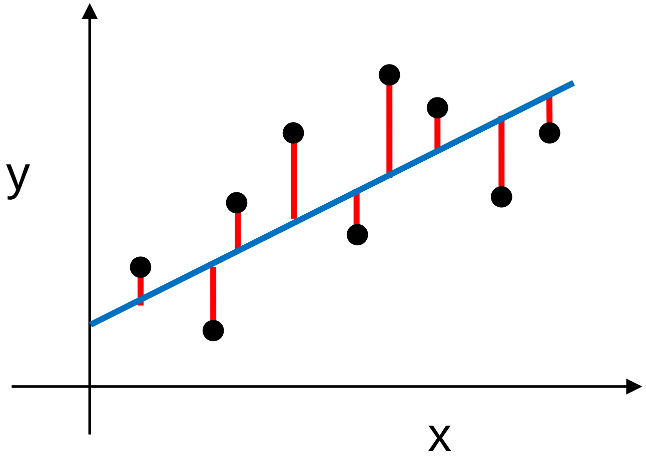
A: By finding the line that minimizes the sum of squared residuals.



$$SS_{residuals} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Q: How to ***estimate*** *coefficients for a linear model?*

A: By finding the line that minimizes the sum of squared residuals.



Model Prediction

$$SS_{residuals} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Observed Result

Q: How to calculate estimates that minimize the sum of squared errors?

A: Through calculus, it can be shown that the following equation minimizes the sum of squared errors.

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Let's walk through an trivial calculation to see how this works.

$$X = \begin{pmatrix} 1, & 3.385 \\ 1, & 0.48 \\ 1, & 1.35 \\ 1, & 465 \\ 1, & 36.33 \end{pmatrix} \qquad Y = \begin{pmatrix} 44.5 \\ 15.5 \\ 8.1 \\ 423 \\ 119.5 \end{pmatrix}$$

Along the way, we'll review some matrix math.

$$\hat{\beta} = \boxed{(X^T X)}^{-1} X^T Y$$

Transposing simply means flipping the columns and rows

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 3.385 & 0.48 & 1.35 & 465 & 36.33 \end{pmatrix} \begin{pmatrix} 1, & 3.385 \\ 1, & 0.48 \\ 1, & 1.35 \\ 1, & 465 \\ 1, & 36.33 \end{pmatrix} = \begin{pmatrix} 5 & 506.54 \\ 506.54 & 217558.38 \end{pmatrix}$$

$$\hat{\beta} = \boxed{(X^T X)}^{-1} X^T Y$$

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 3.385 & 0.48 & 1.35 & 465 & 36.33 \end{pmatrix} \begin{pmatrix} 1, & 3.385 \\ 1, & 0.48 \\ 1, & 1.35 \\ 1, & 465 \\ 1, & 36.33 \end{pmatrix} = \begin{pmatrix} 5 & 506.54 \\ 506.54 & 217558.38 \end{pmatrix}$$

$$\hat{\beta} = \boxed{(X^T X)^{-1}} X^T Y$$

Only square matrices can be inverted

$$(XX^T)^{-1} = \begin{pmatrix} 5 & 506.54 \\ 506.54 & 217558.38 \end{pmatrix}^{-1} = \begin{pmatrix} \boxed{0.26} & -6.1\cdot 10^{-4} \\ -6.1\cdot 10^{-4} & 6.0\cdot 10^{-6} \end{pmatrix}$$

Taking the inverse of a 2x2 matrix simply means swapping across diagonals, and dividing each value by the determinant.

$$\frac{217558.38}{5\times 217558.38 - 506.54\times 506.54}$$

$$\hat{\beta} = (X^T X)^{-1} \boxed{X^T Y}$$

$$X^T Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 3.385 & 0.48 & 1.35 & 465 & 36.33 \end{pmatrix} \begin{pmatrix} 44.5 \\ 15.5 \\ 8.1 \\ 423 \\ 119.5 \end{pmatrix} = \begin{pmatrix} 610.6 \\ 201205.4 \end{pmatrix}$$

$$\hat{\beta} = \boxed{(X^T X)^{-1} X^T Y}$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} 0.26 & -6.1 \cdot 10^{-4} \\ -6.1 \cdot 10^{-4} & 6.0 \cdot 10^{-6} \end{pmatrix} \begin{pmatrix} 610.6 \\ 201205.4 \end{pmatrix} = \begin{pmatrix} 37.201 \\ 0.838 \end{pmatrix}$$

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: 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 a limitation?
A: Not really, a comparison must have a baseline.

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.

II. MAKING INFERENCES

Linear modeling is a parametric technique, meaning that it relies on specific assumptions about the underlying data:

1) *Linearity and additivity of the relationship between input and response variables*
2) *Homoscedasticity of the errors*
3) *Normality of the Error Distribution*
4) *Statistical independence of the errors*

Source: http://people.duke.edu/~rnau/testing.htm

Q: How to determine the whether a coefficient estimate is significant?

A: The p-value associated with the coefficient t-value.

Q: How to determine the whether a coefficient estimate is significant?

A: The p-value associated with the coefficient t-value.

Q: What is a p-value?

A: The probability of getting the observed outcome (e.g., the coefficient estimate) if the null hypothesis were true ($p < 0.05$ is typically considered significant).

Q: What is the null hypothesis for linear regression coefficients?

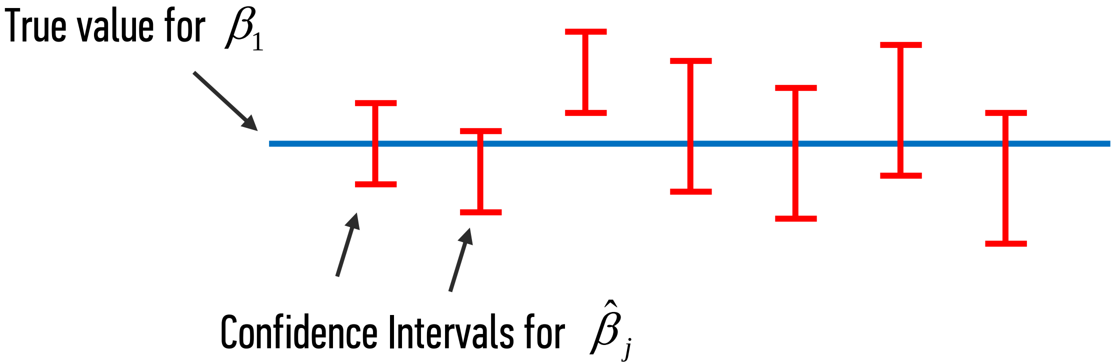
A: There is no relationship between X and Y.

$$H_0: \beta_j = 0$$

$$H_a: \beta_j \neq 0$$

Q: What does the confidence interval mean?

A: 95% of the time, the true coefficients will be in this range.



Q: What does the confidence interval mean?

A: 95% of the time, the true coefficients will be in this range.