Introduction to the linear model (LM)

Data Analysis for Psychology in R 2

dapR2 Team

Department of Psychology
The University of Edinburgh

Weeks Learning Objectives

1. Understand the link between models and functions.
2. Understand the key concepts (intercept and slope) of the linear model
3. Understand what residuals represent.
4. Be able to specify a simple linear model (labs)

What is a model?

- Pretty much all statistics is about models.
- A model is a formal representation of a system.
- Put another way, a model is an idea about the way the world is.

A model as a function

- We tend to represent mathematical models as functions.
    - which can be very helpful.
- It allows for the precise specification about what is important (arguments) and what those things do (operations)
    - This leads to predictions
    - And these predictions can be tested.

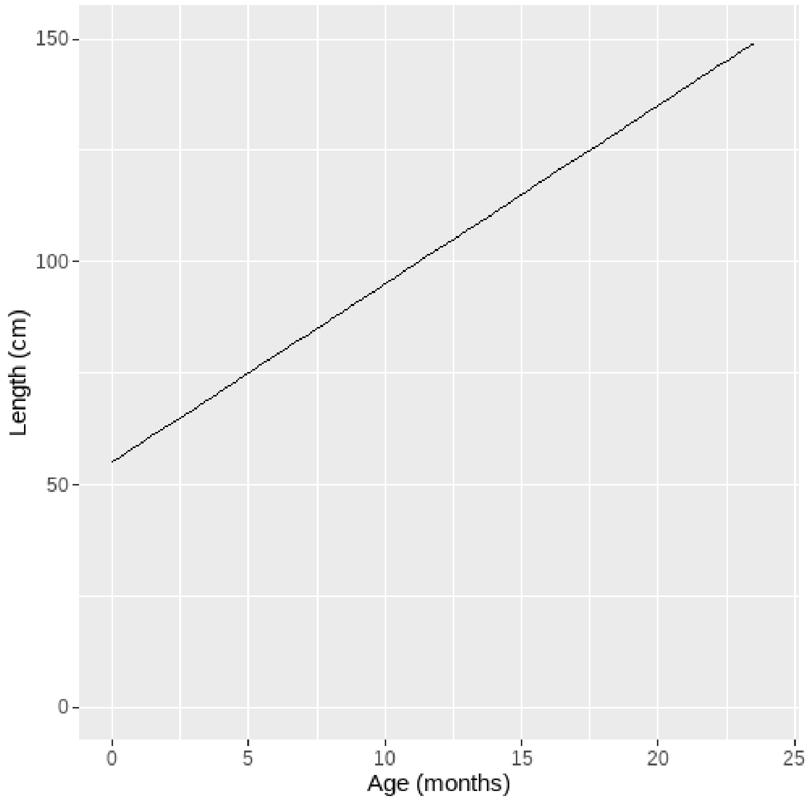
An Example

- To think through these relations, we can use a simpler example.
- Suppose I have a model for growth of babies.[1]

$$Length = 55 + 4 * Month$$

[1] Length is measured in cm.

Visualizing a model



- The black line represents our model
- The x-axis shows `Age` (x)
- The y-axis values for `Length` our model predicts

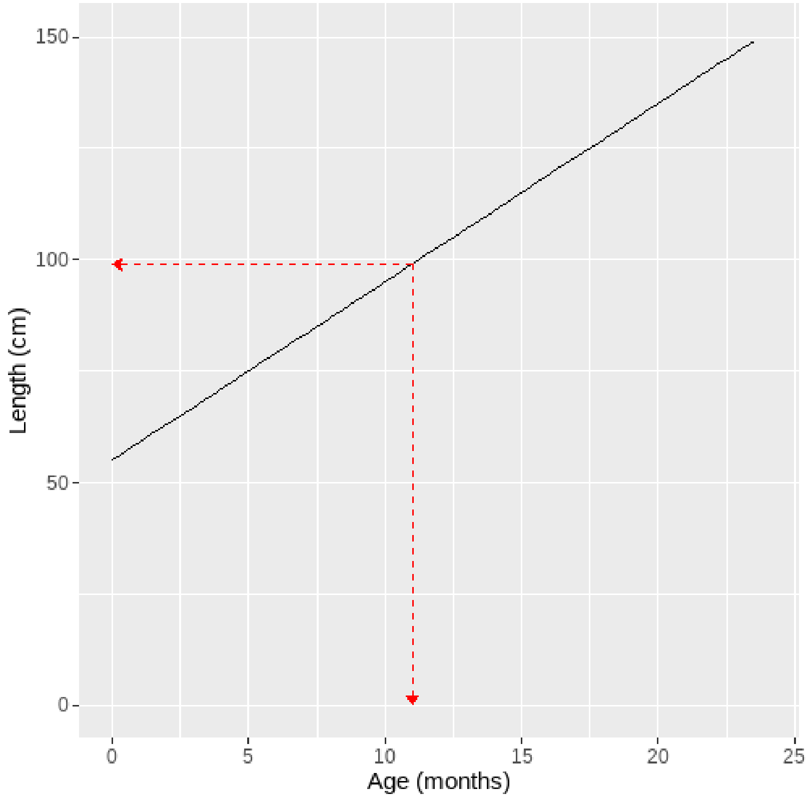
Models as "a state of the world"

- Let's suppose my model is true.
    - That is, it is a perfect representation of how babies grow.
- What are the implications of this?

Models and predictions

- My models creates predictions.
- **IF** my model is a true representation of the world, **THEN** data from the world should closely match my predictions.

Predictions and data



Age	Prediction
10.00	95
10.25	96
10.50	97
10.75	98
11.00	99
11.25	100
11.50	101
11.75	102
12.00	103

Predictions and data

- Consider the predictions when the children get a lot older...
- What do you think this would mean for our actual data?
- Will the data fall on the line?

Age	Year	Prediction	Prediction_M
216	18	919	9.19
228	19	967	9.67
240	20	1015	10.15
252	21	1063	10.63
264	22	1111	11.11
276	23	1159	11.59
288	24	1207	12.07
300	25	1255	12.55

How good is my model?

- How might we judge how good our model is?

    1. Model is represented as a function

    2. We see that as a line (or surface if we have more things to consider)

    3. That yields predictions (or values we expect if our model is true)

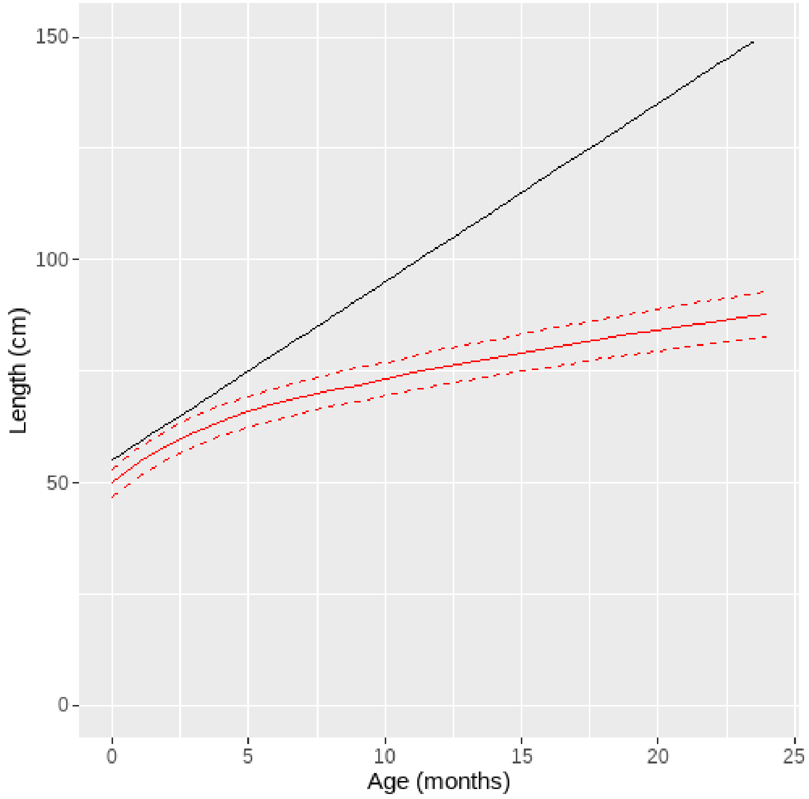
    4. We can collect data

    5. If the predictions do not match the data (points deviate from our line), that says something about our model.

Models and Statistics

- In statistics we (roughly) follow this process.
- We define a model that represents one state of the world (probabilistically)
- We then collect data to compare to it.
- These comparisons lead us to make inferences about how the world actually is, by comparison to a world that we specify by our model.

Length & Age is non-linear

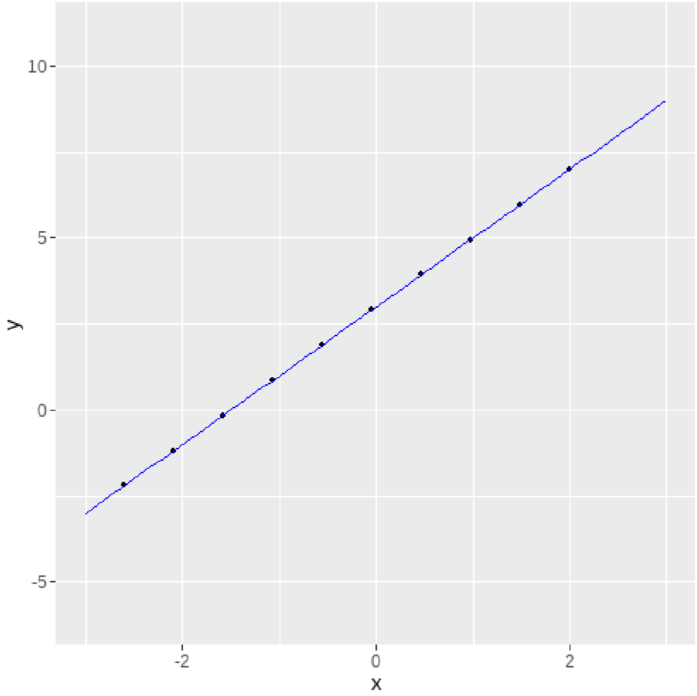


- Our red line is plotted based on the mean length for different ages real data

Deterministic vs Statistical models

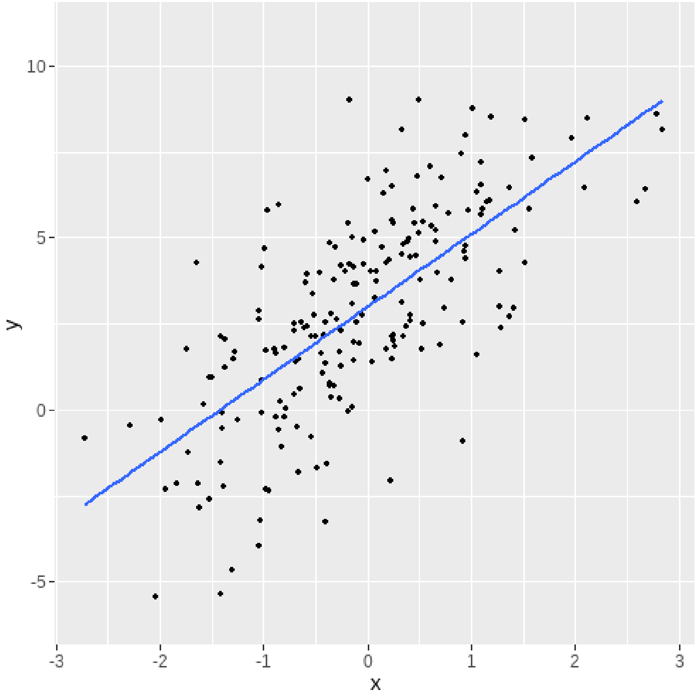
A deterministic model is a model for an **exact** relationship:

$$y = \underbrace{3 + 2x}_{f(x)}$$



A statistical model allows for case-by-case **variability**:

$$y = \underbrace{3 + 2x}_{f(x)} + \epsilon$$



Time to take a breath. Questions...

Linear model

- What we will focus on for the majority of the course is how we move from the idea of an association, to estimating a model for the relationship.
- This model is the **linear model**
- When using a linear model, we are typically trying to explain variation in an **outcome** (Y, dependent, response) variable, using one or more **predictor** (x, independent, explanatory) variable(s).

Example

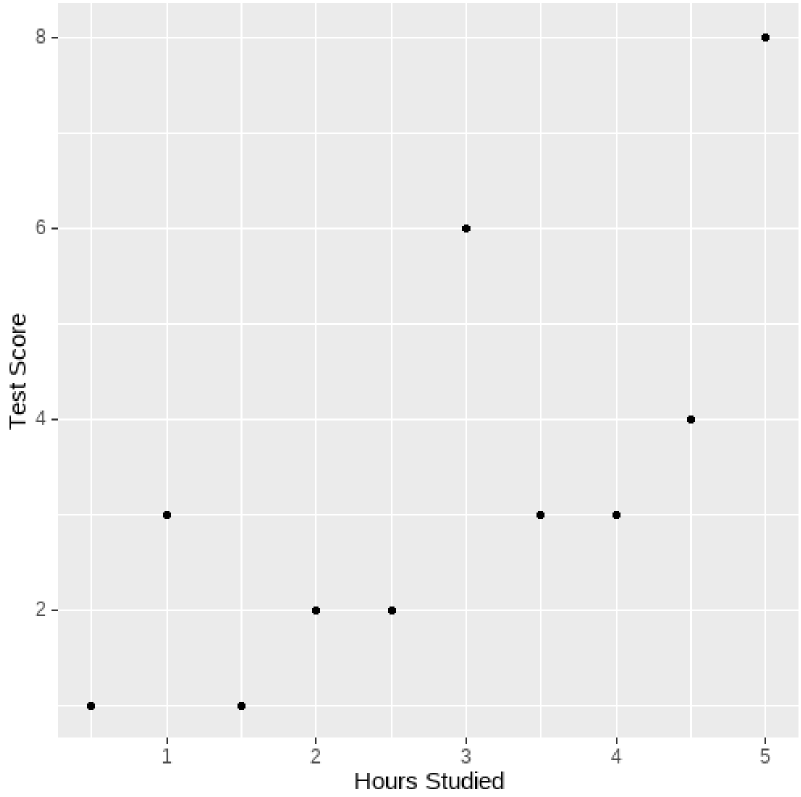
student	hours	score
ID1	0.5	1
ID2	1.0	3
ID3	1.5	1
ID4	2.0	2
ID5	2.5	2
ID6	3.0	6
ID7	3.5	3
ID8	4.0	3
ID9	4.5	4
ID10	5.0	8

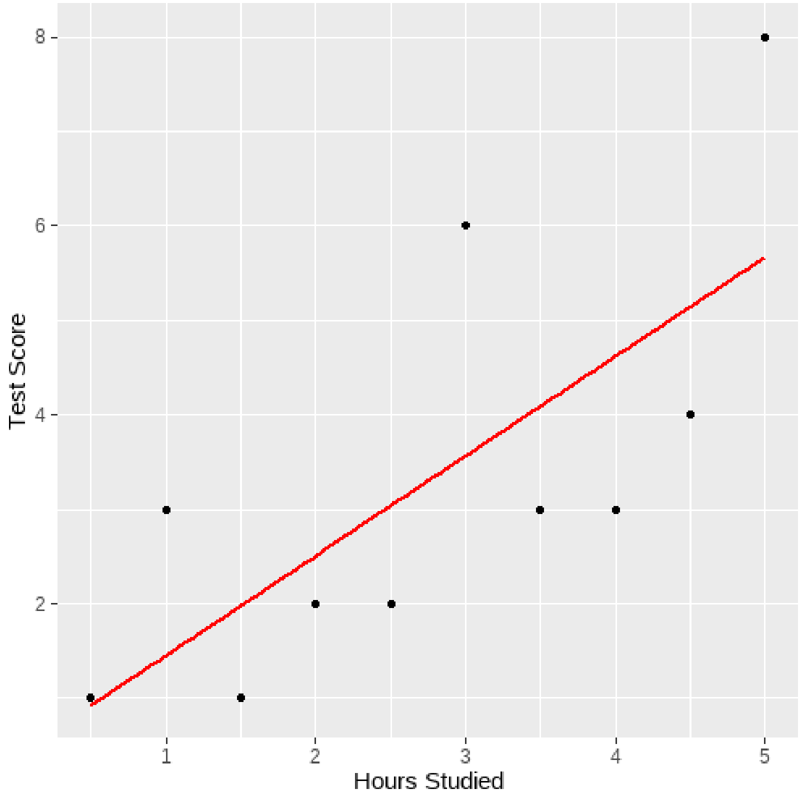
Simple data

- `student` = ID variable unique to each respondent
- `hours` = the number of hours spent studying. This will be our predictor (x)
- `score` = test score (y)

Question: Do students who study more get higher scores on the test?

Scatterplot of our data

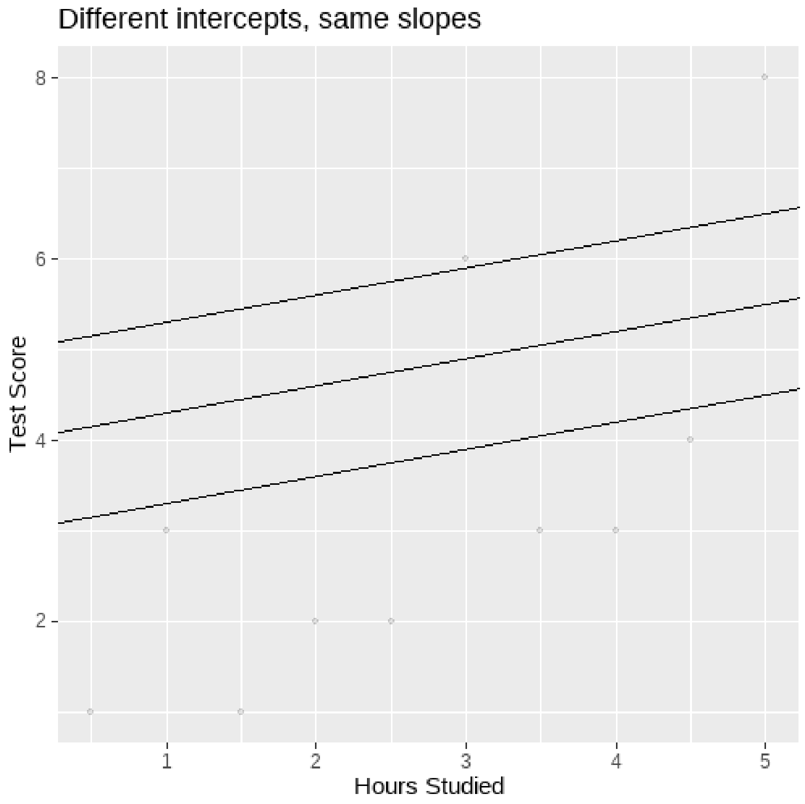




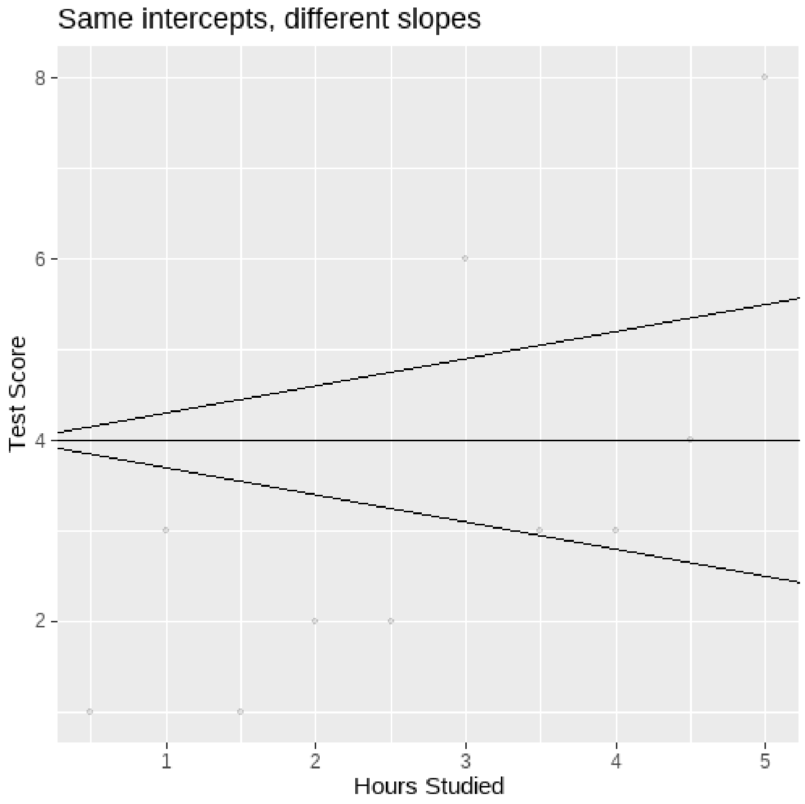
Definition of the line

- The line can be described by two values:
- **Intercept**: the point where the line crosses y, and $x = 0$
- **Slope**: the gradient of the line, or rate of change

Intercept and slope



Different intercepts, same slopes



Same intercepts, different slopes

How to find a line?

- The line represents a model of our data.
    - In our example, the model that best characterizes the relationship between hours of study and test score.
- In the scatterplot, the data is represented by points.
- So a good line, is a line that is "close" to all points.

Linear Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- y_i = the outcome variable (e.g. `score`)
- x_i = the predictor variable, (e.g. `hours`)
- β_0 = intercept
- β_1 = slope
- ϵ_i = residual (we will come to this shortly)
    - where $\epsilon_i \sim N(0, \sigma)$ independently.
    - σ = standard deviation (spread) of the errors
    - The standard deviation of the errors, σ, is constant

Linear Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- **Why do we have i in some places and not others?**

- i is a subscript to indicate that each participant has their own value.

- So each participant has their own:

  - score on the test (y_i)
  - number of hours studied (x_i) and
  - residual term (ϵ_i)

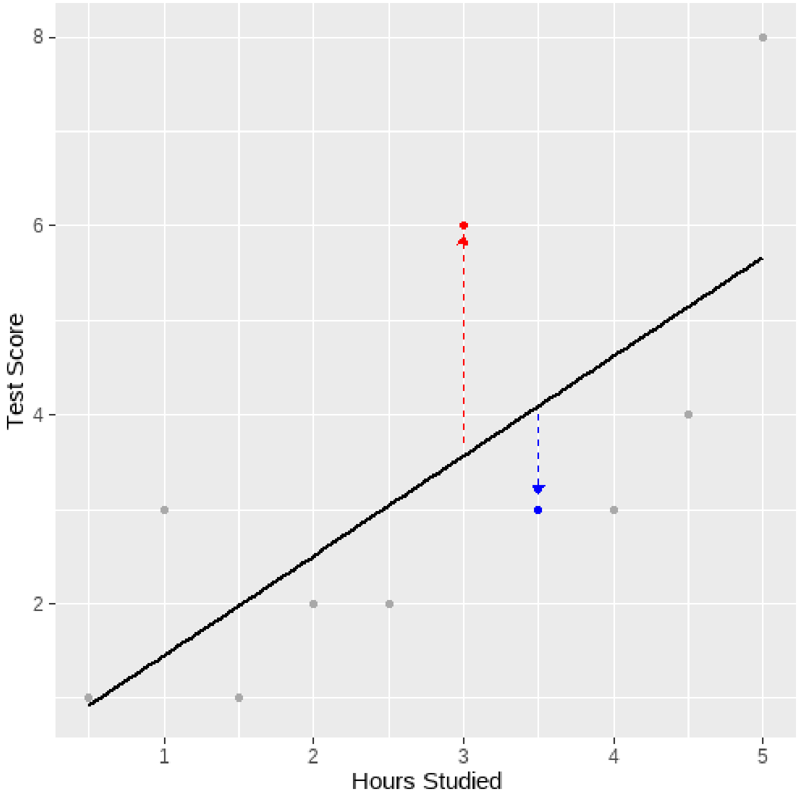
- **What does it mean that the intercept (β_0) and slope (β_1) do not have the subscript i?**

- It means there is one value for all observations.

  - Remember the model is for **all of our data**

What is ϵ_i?

- ϵ_i, or the residual, is a measure of how well the model fits each data point.
- It is the distance between the model line (on y-axis) and a data point.
- ϵ_i is positive if the point is above the line (red in plot)
- ϵ_i is negative if the point is below the line (blue in plot)



Summary

- Take home points...

    1. In statistics, we are building models that describe how a set of variables relate.
    2. The **linear model** is one such model we will use in this course.
    3. The linear model describes our data based on an intercept and a slope(s)
    4. From this model (line) we can make predictions about peoples scores on an outcome
    5. The degree to which our predictions differ from the observed data = residual = error = how good (or bad) the model is

- The majority of this course is going to revolve around getting a deeper understanding of these 5 points.

That is all for this week