# CITS5502
# Workshop week 4 – Regression analysis

# Interpolation

- If we have known data points, and we wish to predict what other data points might be, we can use *regression analysis*[1]
- A simple example:
  - We have an object moving at a constant speed
  - If at t = 1 second, distance travelled = 2m, and
  - at t = 5 seconds, distance travelled = 10m ...
  - What would be the distance travelled at, say, t = 3 seconds?

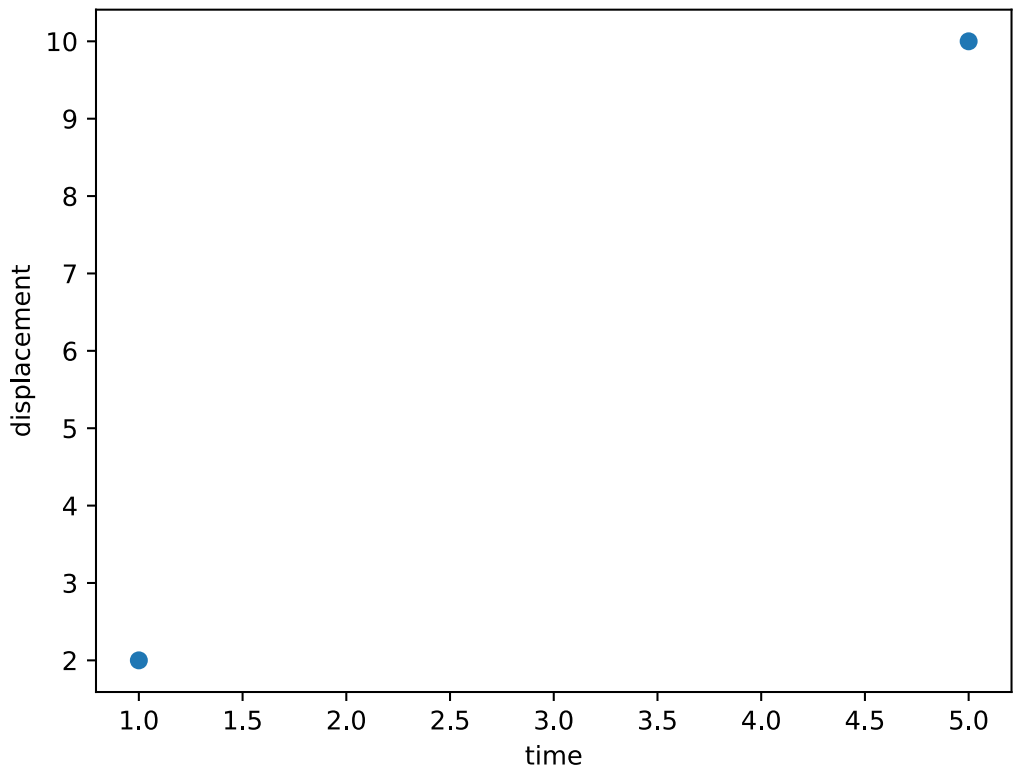
[1] Why the name "regression analysis"? It dates from the 19th C work of Francis Galton on heredity. It's not a very good name, but it's what we're stuck with.

# Simple example

- In the case of the moving object, we know that the graph of position against time will be a straight line –
  so we can use the simplest regression technique, *linear interpolation*
- (which assumes that data follows a straight line between adjacent measurements)

# Simple example, cont'd

- Assume the function between two points is a straight line
- Find equation of the line that passes through the two points
- Put a value of x in the equation to find y
- Put a value of y in the equation to find x



# Simple example, cont'd

This example is simple enough that we can do it by hand. ($s$ = displacement, $t$ = time)

Equation for a line:
$t = as + b$

Substitute in $(5, 10)$:
$10 = 5a + b$

Substitute in $(1, 2)$:
$2 = a + b$ So $10 = 5a + 5b$

So
$0 = 4b$
$b = 0$

And $a = 2$. The equation for the line is $t = 2s$.

# Steps in regression analysis

- State the problem (in this case, predict position at time)
- Identify relevant variables (position, time)
- Collect data (in this case, already provided)
- Specify a model (in this case, linear)
- Model fitting (trivial in this case)
- Model validation and critique
- Apply the model and solve the problem

# Predictor and response variables

- The input variables are called the
  - independent variables, OR
  - predictor variables, OR
  - experimental variables
- The output variable is referred to as the
  - dependent variable, OR
  - response variable, OR
  - outcome variable

# Specifying a model

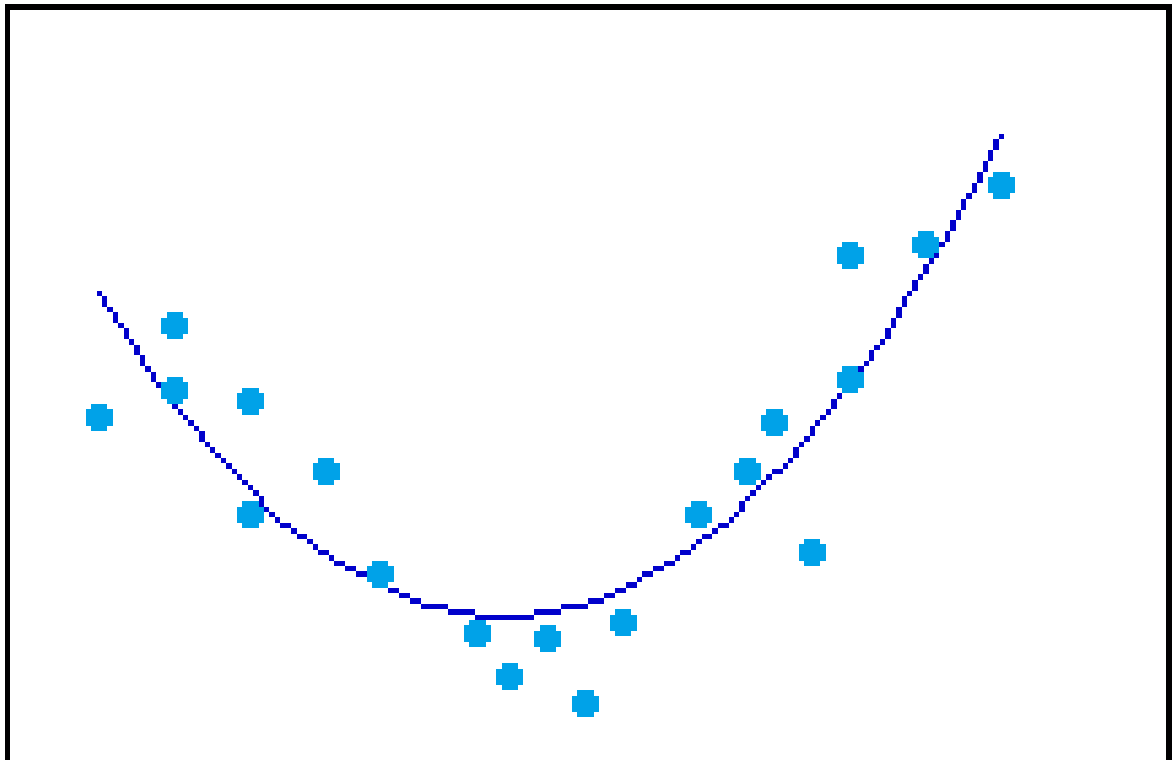
In general, we specify models in terms of *equations* relating our variables, which take (currently unknown) values for their parameters.

e.g. $t = as + b$ relates time and displacement, and takes 2 parameters ($a$ and $b$)

# Possible models

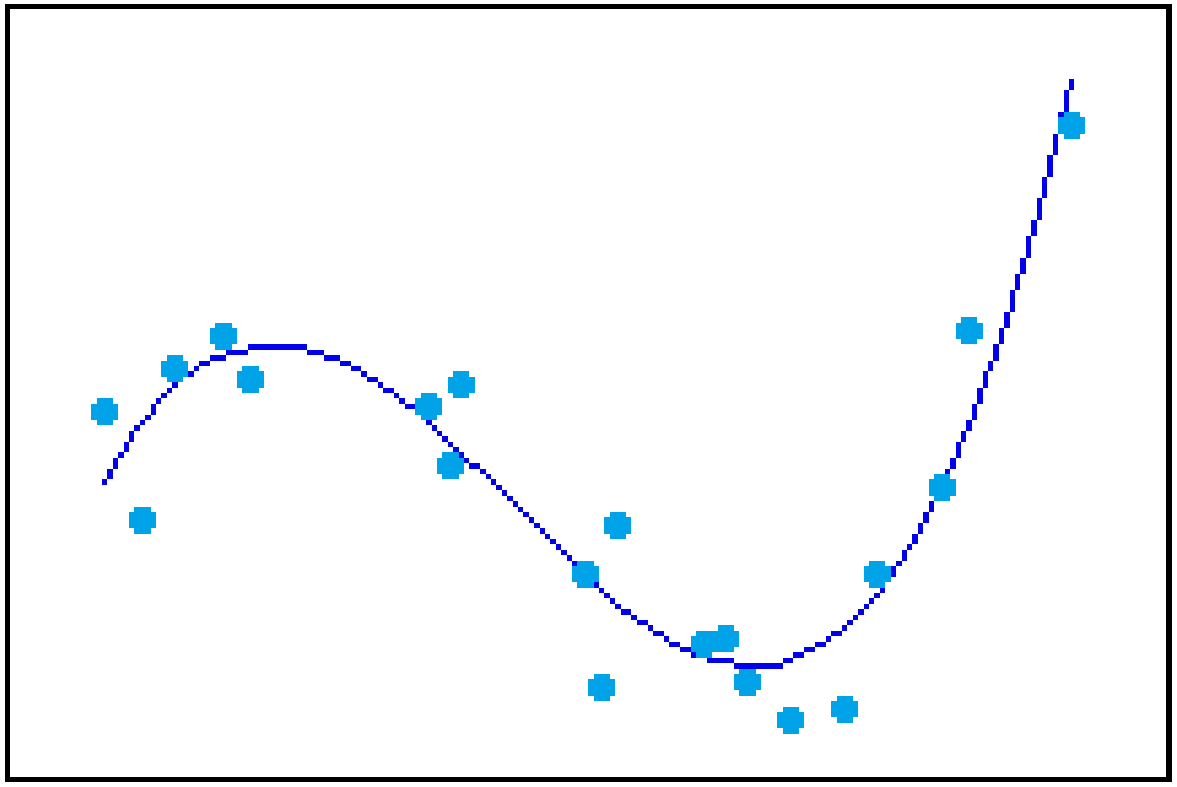
- Refer to the lecture from this week.
- Some discussion of exponential and logarithmic models: https://people.richland.edu/james/lecture/m116/logs/models.html
- Some examples of *polynomial* equations:

Quadratic:



# Possible models

Cubic:



# Possible models

- Recall that in the steps for regression analysis, one thing we do at the end is *critique* our model.
  Actually, we should be doing this all the way through.
- One question to ask is, Is this model physically plausible?

# Possible models

Example:

- I am growing tomatoes, and measuring how large they grow, depending on how much fertilizer I use.
- If I use 0.5 kg of fertilizer on my garden, my tomatoes grow with an average diameter of 3.5cm.
- If I use 0.2 kg of fertilizer on my garden, I get an average diameeter of 1cm.
- What are some plausible models here? What are some implausible ones?

# Model fitting

- Collected data always contains some degree of error or imprecision
- If we assume that all data points are accurate and we want to infer new intermediate data points – that's called *interpolation*
- *Curve fitting* or *model fitting* is used when we want to match a mathematical model to a set of measurements which may contain some error (the usual case)

# Model fitting

- In the case of the moving object, we had only 2 data points, and were using a linear model - the task of "fitting" the model to the data was trivial.
- In general, if we have $n$ data points, they can be *exactly* fitted by a polynomial of degree $n - 1$.
- For instance: 3 data points can be exactly fitted by a quadratic equation.
- But just because we *can* exactly fit a model doesn't mean we *should*.

  Very few processes in the real world give rise to high-degree polynomials.

# Model fitting

- Assuming we do think we have a sensible model, however, a typical process is:
  - Start with a rough guess as to what the parameters might be
  - Calculate the *sum of squares of deviations* of the data from this version of the model.
  - Alter the parameters a bit, and try again
  - Until we seem to have *minimized* the same of the squares.
- This is the "least squares" method.

# Automatic model fitting

- Programming languages with good statistics packages will nearly always provide facilities for doing polynomial regression (and linear regression, which is a special case of this).
- e.g. in Python using Numpy, we use the `polyfit()` function

```
numpy.polyfit(x, y, degree)
```

- The R language provides the `lm()` ("linear model") function for linear regression, and `poly()` for polynomials
- Matlab, like Numpy, has `polyfit()` function.
- In Excel, the easiest way is to use the "Data Analysis Toolpack"

# Automatic model fitting

- For assignments in CITS5502, you are welcome to use whatever language you are most familiar with, or Excel.

# Goodness of fit

- How well do our various models fit the data? Which should we choose?
- One simple measure of fit is to look at the *sum of squared error*
- However, that will be very different between different sets of data, and it can be hard to know what constitutes a "large" or "small" amount of error.
- So a typical measure is one called $R^2$, or the "coefficient of determination".
- This is defined in statistics as "the percentage of variation in the response that is explained by the model" . . .
- . . . but we can basically think of it as a number between 0 and 1 that expresses goodness of fit, with 0 being bad and 1 being good.

# Goodness of fit

- The equation for $R^2$ is:

$$1 - \frac{\text{sum of squared error}}{\text{sum of squared differences from mean of y}}$$

- So $R^2$ is basically comparing the errors of your regression model (top of the fraction) to the errors you'd get if you just used the mean of $y$ to model your data (bottom of the fraction).