Getting computing into the classroom: developing models

Erik Spence

SciNet HPC Consortium

30 October 2014

A note before we start

The following Big Data challenge for high school students has been announced:

`http://cysjournal.ca/page/bigdatachallenge`

The challenge involves students analysing a dataset of grocery-store transactions, coming up with their own results, and presenting their report to a panel of Big Data experts.

Please share with your students.

Agenda

The plan for today:

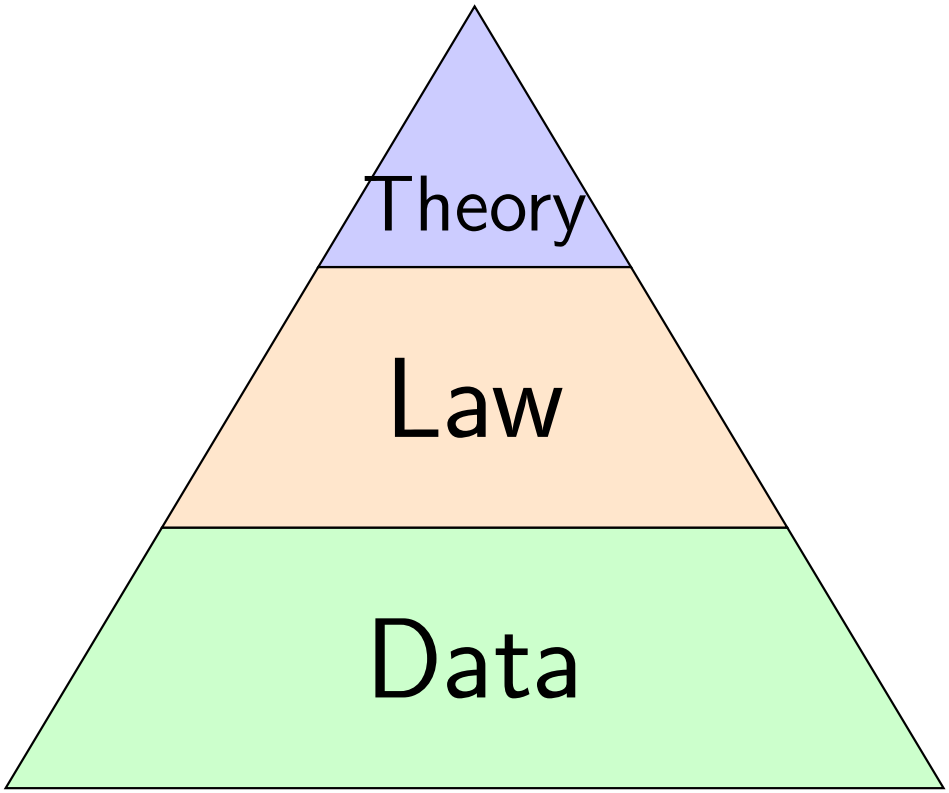
- Examine how modelling and simulation fit into the broader framework of science.
- Collect some data and build some mathematical models.
- Compare physical models with mathematical models.
- Explore some basics of mathematical modelling.
- Examine the effects of data error on model development.
- Examine regression and its use in developing models.

Has anyone NOT yet logged in to the Edmodo site?

SciNet compute • calcul CANADA

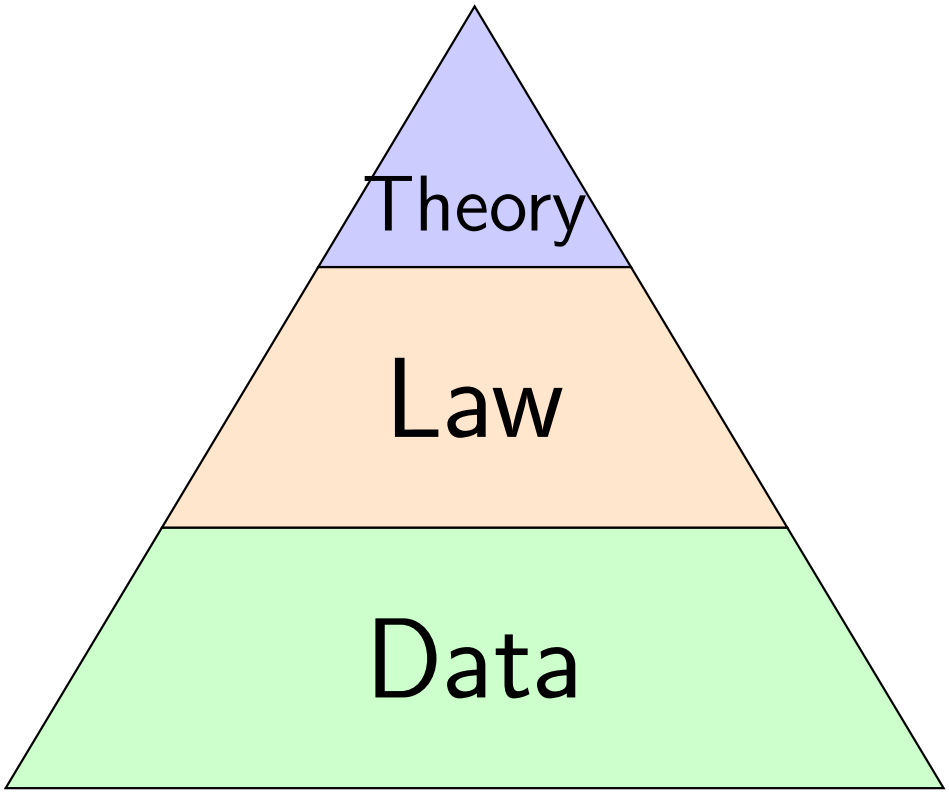
Science, in a nutshell

Behold, the structure of science!



Science, in a nutshell

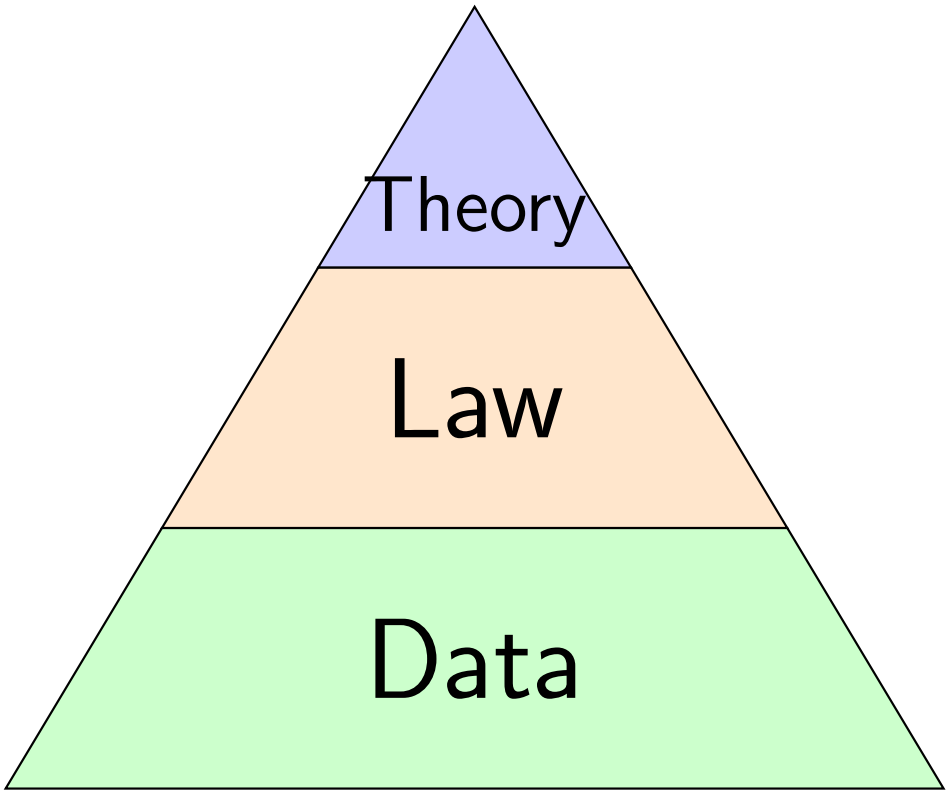
Behold, the structure of science!



Here's an example:

- Data: when you let go of a rock, it accelerates toward the ground.
- Law: it always accelerates at a rate of 9.8 m/s^2.
- Theory: gravitation.

Computation and science



How does computation fit into this picture?

- Computation is used to predict what data might look like, based on laws.
- Computation is used to test possible laws, produced from theory, against data.
- Computation is used to determine laws from data.
- Computation is used to sort and clean data.

SciNet compute • calcul CANADA

Goals for today

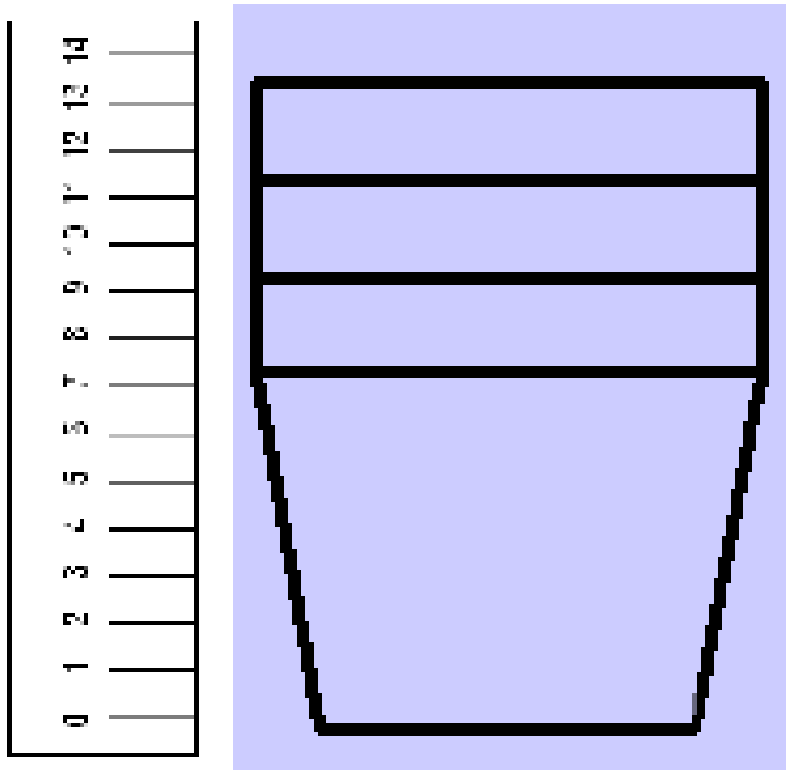
Rather than use existing laws to predict future data, as we did last visit, today we're going to look at the determination of laws from data (model development).

- Build a mathematical model from data.
- Judge the goodness-of-fit of the model to the data.
- Experience the process: Data $\longrightarrow$ Model $\longrightarrow$ Simulation
- Graphical interpretation – what does the graph tell me?

SciNet compute • calcul CANADA

Part I: Measurement data

In our first analysis, we're going to determine the "law of stacked cups". We will develop a model which predicts the height of a stack of cups.



We will use a pre-built "Just Add Data" Excelet.

Objectives

With this exercise we will:

- Collect and plot experimental data.
- Develop a mathematical model and examine its goodness of fit.
- Make predictions with the model.
- Physically interpret the model by simulation.
- Examine the inverse function of the model, and its use.
- Simulate possible experimental errors in the model.

And in the process, develop the "law of stacked cups".

SciNet compute • calcul CANADA

Collecting the data

Gather yourselves into small groups.

- You will be given a selection of cups.
- Using a metric ruler, measure the stack height to the nearest millimetre.
- Measure for 2, 4, 6, 8, 10, 12 cups. Record your results.
- Put the data into the stacking_cups.xls spreadsheet (available on the Edmodo site), under the "generate a model" tab.

Student question: what is the dependent variable in the data being collected?

What are we doing?

How does this exercise work?

- As you put the data into the spreadsheet, you'll see that a line of best fit is generated.
- This is accomplished by performing a "linear regression" on the data.
- The regression is your mathematical model for the data (your law which describes the data).
- One measure of the goodness of fit for a regression is given by r^2. A perfect fit of the model to the data would yield $r^2 = 1$.
- Be sure to put your slope and intercept in the yellow boxes on the right!!

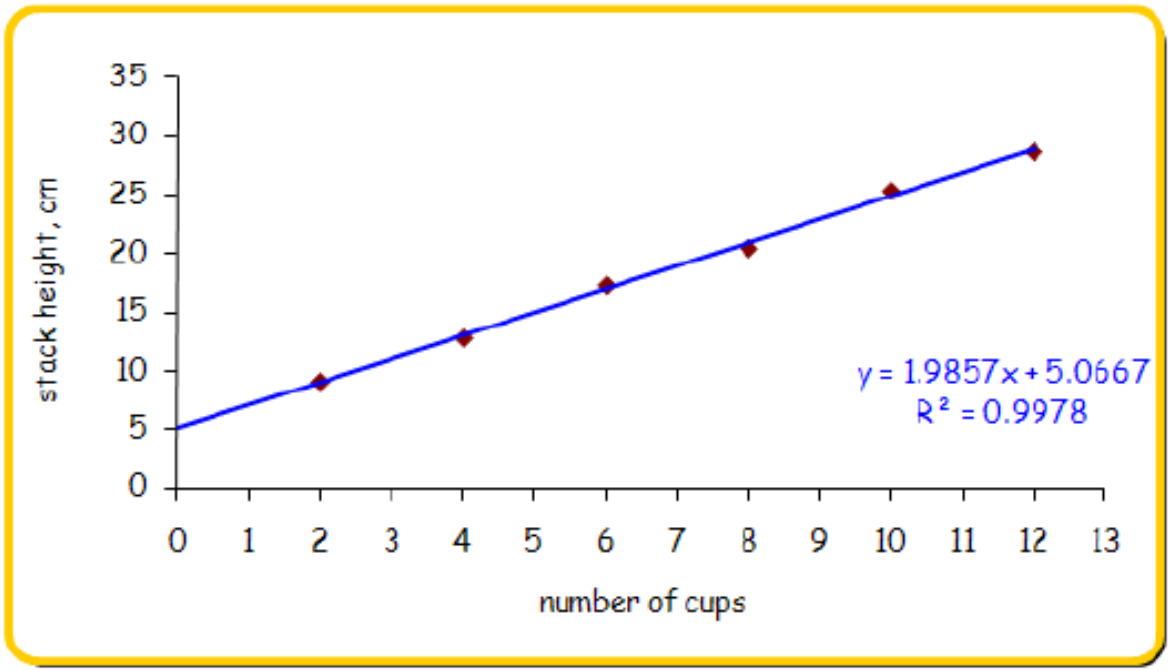
SciNet compute • calcul CANADA

The mathematical model

Measuring the Stack Height of Nested Styrofoam Cups - Building a Mathematical Model

Using a centimeter ruler, carefully measure the stacks to the nearest 0.1 cm. Record your data in the yellow cells.

number of cups	stack height
2	9.1
4	12.8
6	17.3
8	20.5
10	25.4
12	28.7



From the best-fit line record the following in the yellow cells:

What is the value of the slope?

1.99

What is the value of the intercept?

5.07

How well does the line fit the data?

R^2 value

As data are added, the points will plot and a best-fit line and the equation of that line will appear on the graph.

Student questions

Further questions for students:

- What does the equation which is generated for the data mean?
- What does the slope represent in terms of the variables being investigated?
- What are the units of the slope?
- Why is the y-intercept non-zero? What does it represent in terms of the variables investigated? What are its units?

Now make some predictions using your model:

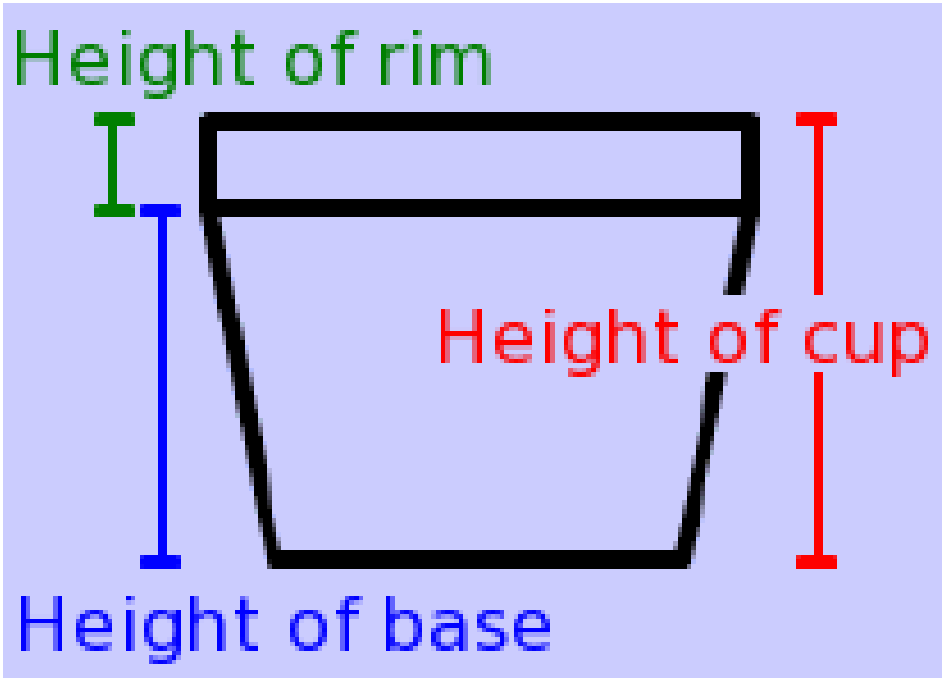
- What is the height of 150 cups?
- How many cups are required to obtain your height?

SciNet compute • calcul CANADA

Using the model

Having developed the "law of stacked cups", we will now use it to make predictions.

- Click on the "simulate with model" tab.
- Notice that we have the ability to change the "rim height" and "base height".



Student questions

Questions for consideration:

- How does changing the height of the rim influence the regression line on the graph?
- How does changing the base influence the regression line?
- Rewrite your mathematical model to incorporate base height and rim height.
- For your cups, what are the actual values of the parameters (rim and base heights), based on the regression results?
- Are these values realistic?
- What can we say about the uniformity of cup manufacturing?
- To reverse the equation, meaning figuring out how many cups would fit into, say a 61cm box, what needs to be done algebraically?

SciNet compute • calcul CANADA

Error analysis

Go to the "simulate errors" tab and you will find five different errors to investigate for this mathematical model. Errors can come from measurement or manufacturing.

How does each error influence the data and the model?

Errors	Slope of line	y-intercept	Scatter of data
Random measurement error			
Systematic measurement error			
Rim uniformity			
Base uniformity single stack			
Base uniformity multiple stacks			

An introduction to regression

So we've developed our model of the data (the "law of stacked cups"). But how did the excelet do it?

- We gave the data to the excelet.
- The excelet performed a linear regression on the data.

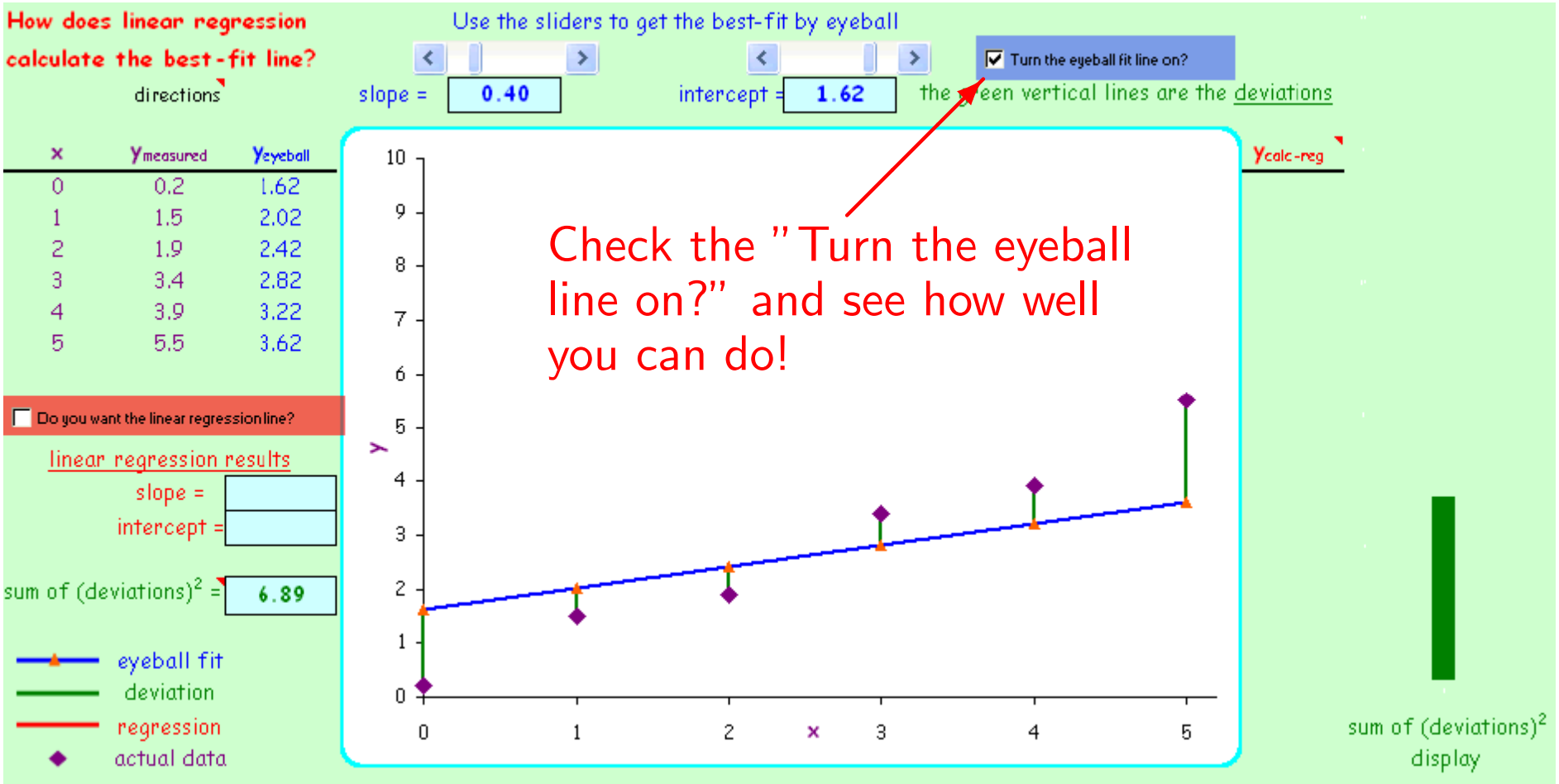
But what is that, "linear regression"?

- The goal of linear regression is to minimize the sum of the squared deviations.
- The deviations are the distances between the model value (the fit) and the actual data.

SciNet compute • calcul CANADA

A regression excelet

Open the "regression.xls" excelet.



A regression excelet, continued

The "regression.xls" excelet has bars which allow you to adjust the parameters of the model.

- Adjust both the intercept and the slope to minimize the sum of the squared deviations.
- How close can you come to the data by eyeballing it?
- Click on the red "Do you want the linear regression line?" checkbox.
- How do your intercept and slope compare to the computerized version?
- Can you do better than the computer?

Let me know if you can do better than the computer.

SciNet compute • calcul CANADA

A regression excelet, goodness-of-fit

The goodness-of-fit can be judged by the minimizing of the sum of the squared deviations.

- The smaller the sum, the better the line fits the data points.
- The value of r^2 is a numerical way of expressing the minimized squared deviations. A value of 1 is a perfect fit.
- A plot of the remaining deviations (click on the "deviations" tab) is another way of judging the goodness-of-fit.
- If the remaining deviations are not random, there should probably more structure to your model.

Part II: M&Ms modelling

In most real scientific situations, you do not know the form of the law that will explain your data. Let's perform another experiment:

- Take a cup, paper towel, 70 M&Ms.
- This is your first datapoint: trial 0, 70 M&Ms.
- Shake the cup, and dump the M&Ms on the paper towel, on the table.
- Remove any M&Ms that are "M-up".
- Record the datapoint: trial 1, X M&Ms remaining.
- Repeat 4 more times.

Don't eat the data! (Just yet, anyway.)

SciNet compute • calcul CANADA

M&Ms modelling

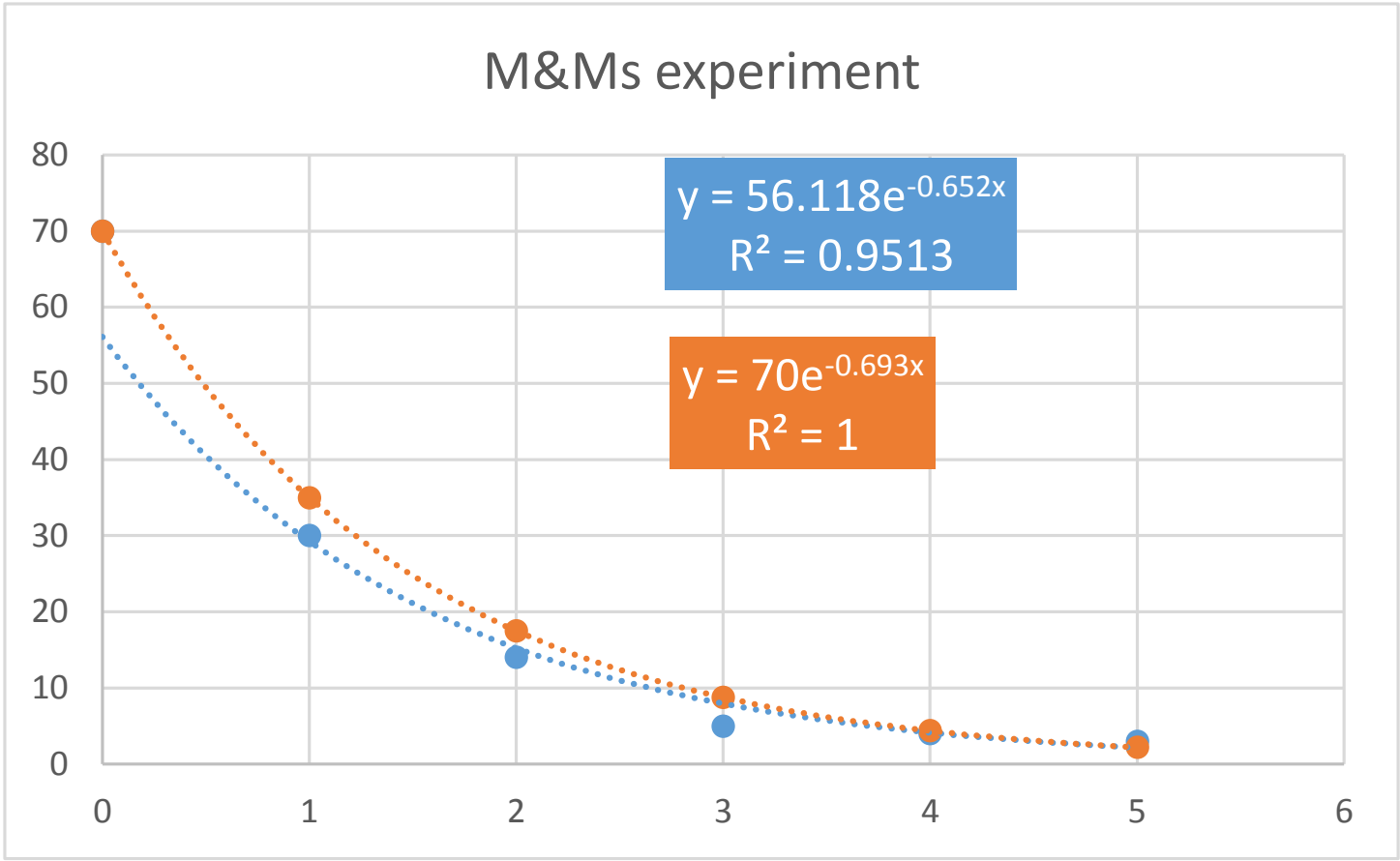
We will now develop a model to predict the removal rate of our M&Ms:

- Once you have your data, enter the data into a new Excel page.
- Make a scatter plot of the data.
  - Highlight the data to be plotted.
  - Click on "Insert Chart".
  - Select "Scatter".
  - Select "Scatter with only Markers".
- With the graph selected, click on the "Trendline" button.
- Click on "More Trendline Options", at the bottom.
- Select "Display Equation" and "Display R-squared".
- Try various types of trendlines.

Which type of trendline best models the data? How do you know?

My M&M data

Trial	Data	Theory
0	70	70
1	30	35
2	14	17.5
3	5	8.75
4	4	4.375
5	3	2.1875



M&Ms modelling, continued

Once Excel has produced an equation, we are ready to go from Law to Theory.

Can you convert your equation from numbers to symbols? If not, what is the difficulty?

M&Ms modelling, continued

Once Excel has produced an equation, we are ready to go from Law to Theory.

Can you convert your equation from numbers to symbols? If not, what is the difficulty?

- The form of the equation we're after is $f(x) = a(1/2)^{bx}$.
- However, Excel will only give you a trendline of the form $g(x) = ce^{dx}$.
- We need to convert the answer that Excel gives to the answer that we're after.
- In this case this is straightforward. If the fit is a good one, then $e^{dx} = \left(e^{d}\right)^{x} = (1/2)^{x}$.
- You should double check that your value of d leads to this result.

Student questions

Questions which might be posed to students:

- What type of trendline best fits the data?
- What do the fit parameters in the equation represent?
- Why is there a discrepancy between the experimental and theoretical value?
- If we instead started with 100 M&Ms, how many would we expect to have after 5 trials?
- What natural processes might this experiment represent?

SciNet compute • calcul CANADA

What have we learned?

What have we learned in today's class?

- Computation can be used to analyse data, to determine the laws which govern the data.
- Once such (mathematical) laws are determined, we can use them to make predictions about future measurements. This results in the "hypothesis falsifiability" which is so critical to the Scientific Method.
- Errors in the data, whether measurement or systematic, affect the resulting model, and consequently its predictions.
- M&Ms are yummy.

SciNet compute • calcul CANADA

For next class

We are still soliciting suggestions of what topics to cover. Possibilities include:

- A review of computer widgets that are available, and the topics they might pertain to.
- Playing with Big Data, whatever form that may take.
- Technologies for large-scale computation, such as in-class computer clusters.
- Other suggestions?

Email me, post to Edmodo, whatever!

SciNet compute • calcul CANADA