

# Data Fitting Module Student Guide

## Activity 1: Fitting a Straight Line By Hand (20 minutes)

Often data have two variables, such as the magnitude of the force  $F$  exerted on an object and the object's acceleration  $a$ . In this Module we will examine some ways to determine how one of the variables, such as the acceleration, depends on the other variable, such as the force.

Say we have collected data for the acceleration  $a$  of a cart of mass  $M$  for a constant applied force  $F$ . We want to determine how the acceleration depends on the force. The acceleration is some function of the force:

$$a=f(F) \quad \text{Eq.[1]}$$

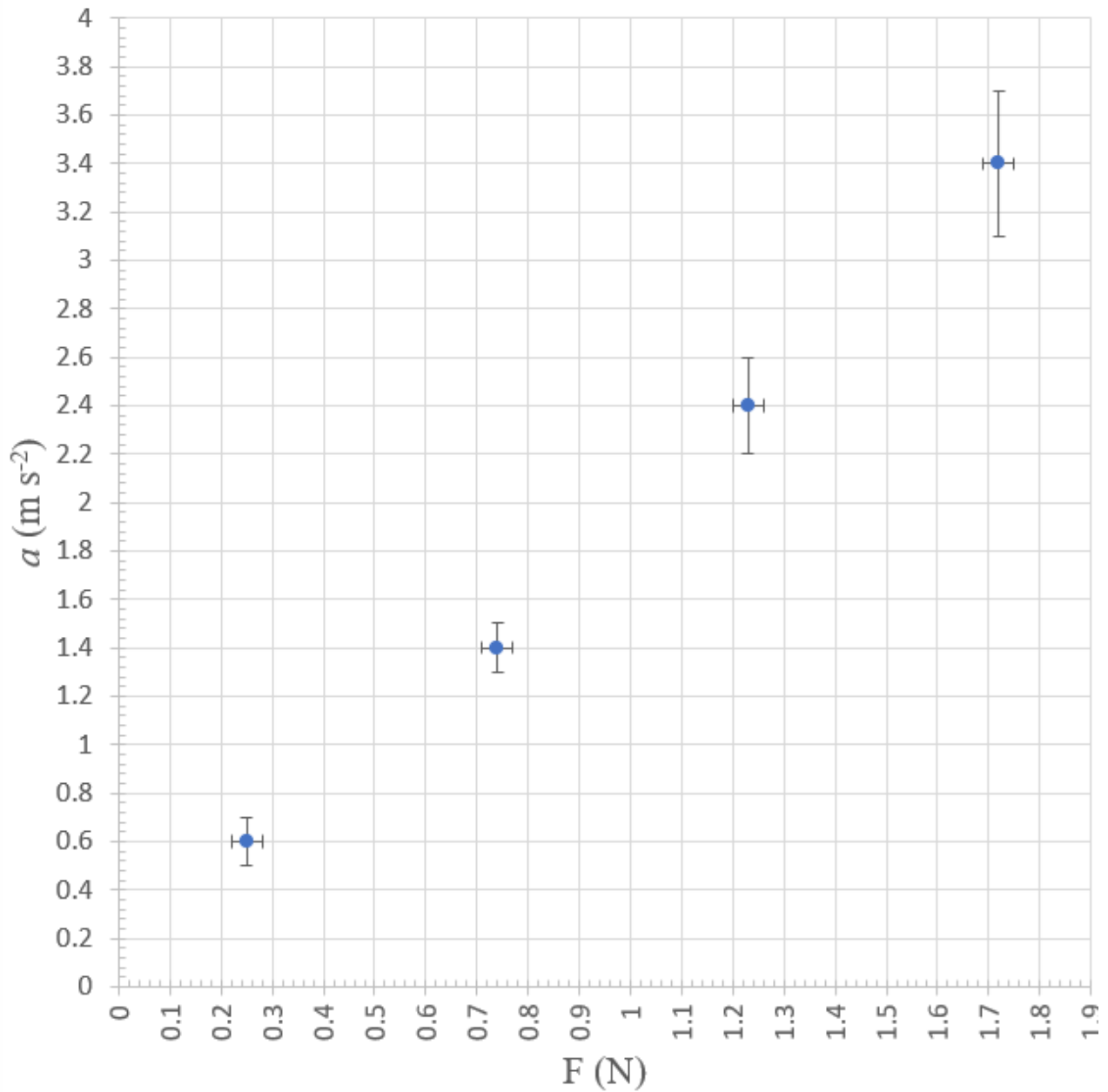
In this case the variable  $F$  is called the **independent variable**, it is the quantity that is being experimentally changed by changing the mass  $m$ . Then the variable  $a$  is called the **dependent variable**, and its value depends on the value of the independent variable.

Table 1 shows the data for the experiment, for which four different constant forces were applied, and in each case the acceleration was measured. The graph is on the next page.

$F$ (N)	$a$ (m s <sup>-2</sup> )
0.25 ± 0.03	0.6 ± 0.1
0.74 ± 0.03	1.4 ± 0.1
1.23 ± 0.03	2.4 ± 0.2
1.72 ± 0.03	3.4 ± 0.3

**Table 1**

## Force - Acceleration Data



The dots are at the values of the force and acceleration, and the length of the bars through the dot indicate the values of the uncertainties. These indicators of uncertainty are historically called “error bars”, but would better be called “uncertainty bars”.

If we assume Newton's 2<sup>nd</sup> Law is correct for the data, then for a frictionless cart:

$$a = \frac{1}{M}F \quad \text{Eq. [2]}$$

Eq. 2 is called a **model** of the physical system. From the equation, the slope of a straight line through the data points is equal to  $1/M$ .

- A. Draw the best straight line that you can through all the data points. A straight line has the equation  $y = mx + b$ , but you have an extra constraint that  $b$  must be zero, since the acceleration should be exactly zero when there is no applied force. Considering that the uncertainties in the values of the data are saying that the experimenter believes that the actual value has a 68% probability of being within the range given by the uncertainties, does the line have to go through all of the rectangles defined by the uncertainties or only most of them? Explain. Find the best-fit slope  $m$  of the line, including its units. Calculate the mass  $M$ , including units.
- B. In finding the "best" straight line, you may have noticed that you can wiggle the ruler around a bit and still account pretty well for the data within the experimental uncertainties. Determine how much you can wiggle the ruler and still account for the data. Remember the line with the maximum or the minimum slope only has to go through about 68% of the error bars. The amount of wiggle you can do with the ruler and still account for the data determines the **range** of the slope, and the **uncertainty** in the slope is half this range. Determine what that uncertainty in the slope is. Present your experimental determination of the mass  $M$  including its uncertainty. Include an image of your plot with the lines you drew. Recall:
- If a quantity is raised to a power,  $z = x^n$ , then the uncertainty in  $z$  is given by  $u(z) = |nx^{(n-1)}u(x)|$ . Here  $M = m^{-1}$ , i.e.  $n = -1$ .
  - Uncertainties should be specified to two significant figures, at most. The most precise column in the number for the uncertainty should also be the most precise column in the number for the value. For example, if a value is 8.10136 and the uncertainty is 0.015632, then the measurement should be specified as  $8.101 \pm 0.016$ .

## Activity 2: Least-Squares Fitting with Python (30 minutes)

In Activity 1 you fit the data of Table 1 to a model by hand. The model was given by Eq. 2, and you used the graph of the data to perform the fit and determine the value and uncertainty in the parameter  $1/M$ .

Often we use computers to do such fits numerically. The most common type of fitting is to a polynomial model:

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + \dots \quad \text{Eq.[3]}$$

For example, if the model is a straight line,  $y = mx + b$ , then  $a_0$  is the intercept  $b$ ,  $a_1$  is the slope  $m$ , and all other of the parameters  $a_i$  are zero. If the model is a parabola,  $y = cx^2$ , then the only non-zero parameter is  $a_2$  which is  $c$  in the model. In general, the fit determines the values of the parameters  $a_i$  that are non-zero.

Say we are fitting to an arbitrary model:

$$y = f(x) \quad \text{Eq.[4]}$$

We have a series of values of the data:  $(x_1, y_1)$ ,  $(x_2, y_2)$ ,  $\dots$ ,  $(x_N, y_N)$ . For each datapoint, the fitted value of the dependent variable,  $y_{i,\text{fit}}$ , is given by:

$$y_{i,\text{fit}} = f(x_i) \quad \text{Eq.[5]}$$

However, the experimental value of  $y_i$  is unlikely to be exactly equal to  $y_{i,\text{fit}}$ . We define the **residual**  $r_i$  to be:

$$r_i = y_i - y_{i,\text{fit}} = y_i - f(x_i) \quad \text{Eq.[6]}$$

A perfect fit the sum of the residuals for all the data is zero. However, the **sum of the squares of the residuals** is not zero. It is a measure of how much the model differs from the data. The most common technique for computer fitting of data to a model is called **least-squares**. The name is because it finds the values of the fitted parameters for which the sum of the squares of the residuals is a minimum.

There is a famous quartet of  $(x, y)$  pairs devised by Anscombe [[F.J. Anscombe, American Statistician 27 \(Feb. 1973\), pg. 17](#)]. Here is a Python program that loads the usual libraries, defines the four datasets, and does some analysis of the first dataset:

```
from pylab import polyfit, plot, show
from numpy import mean, var

# Anscombe's first dataset
```

```

A1x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
A1y = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26,
10.84, 4.82, 5.68]

# The second
A2x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
A2y = [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13,
7.26, 4.74]

# The third
A3x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
A3y = [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15,
6.42, 5.73]

# The fourth
A4x = [8, 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]
A4y = [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56,
7.91, 6.89]

print ("First dataset:")
print ("Mean of x:", mean(A1x))
print ("Variance of x:", var(A1x, ddof = 1))
print ("Mean of y:", mean(A1y))
print ("Variance of y:", var(A1y, ddof = 1))
print ("Straight line fit:", polyfit(A1x, A1y, 1, full = True))

```

Note that the first of the four datasets consists of the values of  $x$  in  $A1x$ , and the values of  $y$  in  $A1y$ . The other three datasets are similarly named except for the number in the variable name.

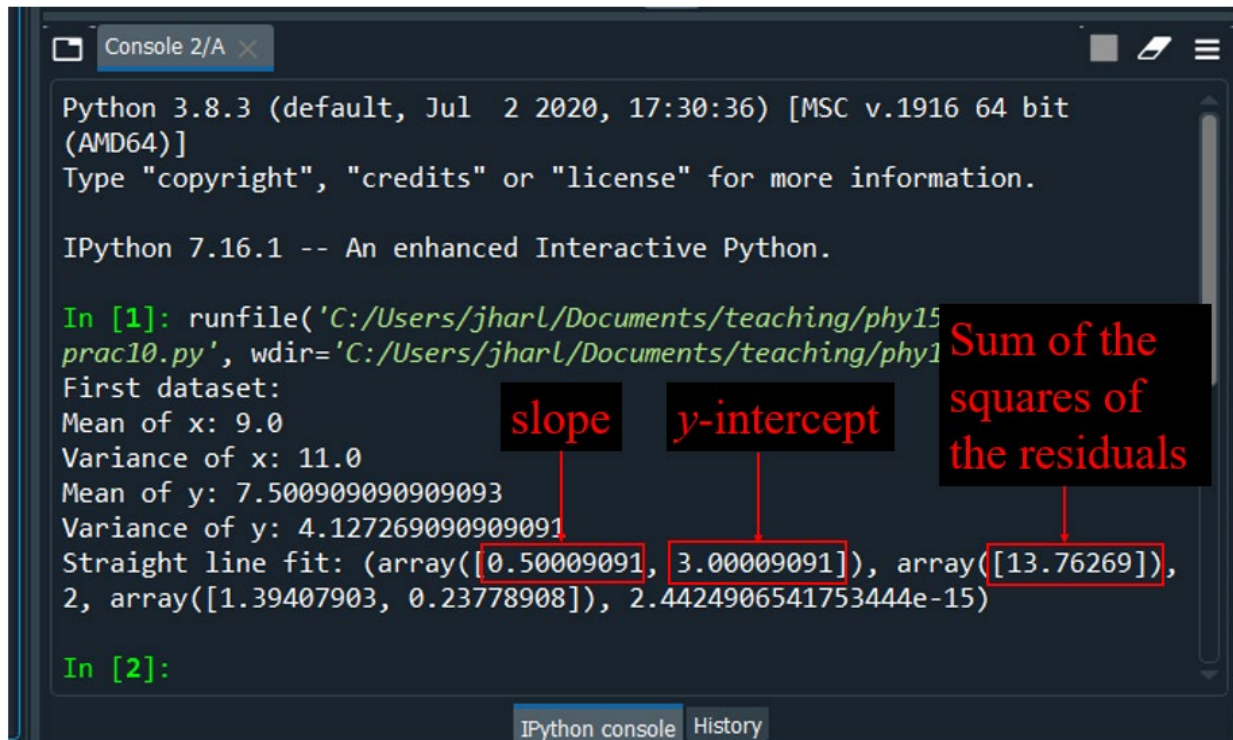
The program computes the means and variances of the  $x$  and  $y$  variables. The last line fits the data to a straight line. Run the program. The output window will look like the image below, except for the red boxes and labels, which have been added.

```
Console 2/A x
Python 3.8.3 (default, Jul 2 2020, 17:30:36) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 7.16.1 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/jharl/Documents/teaching/phy15
prac10.py', wdir='C:/Users/jharl/Documents/teaching/phy15
First dataset:
Mean of x: 9.0
Variance of x: 11.0
Mean of y: 7.500909090909093
Variance of y: 4.127269090909091
Straight line fit: (array([0.50009091, 3.00009091]), array([13.76269]),
2, array([1.39407903, 0.23778908]), 2.4424906541753444e-15)

In [2]:
```



There is a lot of information in the results of the fit, but we will concentrate on the slope,  $y$ -intercept, and sum of the squares of the residuals.

Copy and paste the program lines into your spyder so that it computes the means, variances, and does the fit and change the copied lines so that the program does the calculation on the 2<sup>nd</sup>, 3<sup>rd</sup>, and 4<sup>th</sup> datasets.

- A. Report the approximate values for the fits for the four datasets: slope,  $y$ -intercept, and the sum of the squares of the residuals of the fit. From these fit parameters, what might you conclude about whether or not the four datasets are almost identical?

Plot the first dataset with:

```
plot(A1x, A1y, 'bo')
show()
```

- B. Look at the plot, and then plot the other three datasets. Include the plots in your presentation, as well as the best-fit lines. Now what do you think about the similarity of the four datasets? Is a straight line model appropriate for each of the datasets? If not, explain what is wrong.

### Activity 3: Evaluating the Quality of a Fit (15 Minutes)

Imagine we are fitting some data to a straight line:  $y = mx + b$ . If there is only one datapoint, then no such fit is possible: any line going through the datapoint is equivalent to any other line going through the datapoint. If there are exactly two datapoints, then there is no doubt about the values of the slope and intercept: they are the slope and intercept of the line connecting the two points. However, if there are three or more datapoints, then we can imagine a range of slopes and intercepts of lines that more-or-less are consistent with the data.

The **degrees of freedom** of a fit are the number of datapoints minus the number of parameters to which we are fitting, which is two for a straight line. Fits with negative degrees of freedom are impossible. Fits with zero degrees of freedom are exact.

- A. Imagine you are fitting the data of Table 1 to a straight line with an added parabolic term:

$$a = mF + b + cF^2$$

What is the number of degrees of freedom of the fit?

In Activity 2 we learned the graphs are an important tool in evaluating fits. Now we will learn about some quantitative ways of evaluating a fit.

The sum of the squares of the residuals of a fit,  $ss$ , is:

$$ss = \sum_{i=1}^N [y_i - f(x_i)]^2$$

Eq.[7]

where we have fit the data to the model  $y = f(x)$  and there are  $N$  datapoints. It measures the “goodness” of the fit, with smaller values meaning a better fit. But there is no objective way to determine if the value of  $ss$  is “small” or “large.”

However, if the data have uncertainties in the dependent variable,  $u(y_i)$ , then we can weight each residual by 1 over that uncertainty, and form the sum of the squares of the weighted residuals. This sum is called the **chi-squared**,  $\chi^2$ . ( $\chi$  is the Greek letter “chi” which starts with a hard “k”-sound, and rhymes with the word “eye.”)

$$\chi^2 = \sum_{i=1}^N \left[ \frac{y_i - f(x_i)}{u(y_i)} \right]^2$$

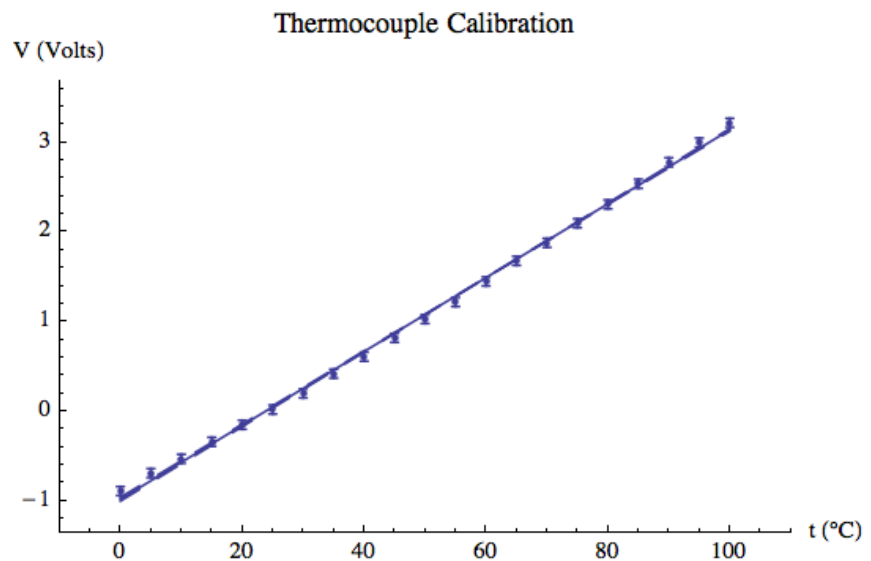
Eq.[8]

Now a “least-squares” fit finds the minimum in the  $\chi^2$ , which may be for different values of the fitted parameters than the values found by minimizing the sum of the squares of the residuals.

If the data are correct and the model is reasonable, the  $\chi^2$  should be roughly equal to the number of degrees of freedom. If the  $\chi^2$  is much larger than the number of degrees of freedom, the fit is poor. If the  $\chi^2$  is much less than the number of degrees of freedom, the fit is too good to be true.

### Real Data Example

A thermocouple is a device that emits a voltage that depends on its temperature. Thermocouples are often used as thermometers. The figure shows some student-collected data on calibrating a thermocouple that was presented by Bevington [Philip R. Bevington, *Data Reduction and Analysis* (McGraw-Hill, 1969), pg. 138]. The student assigned an

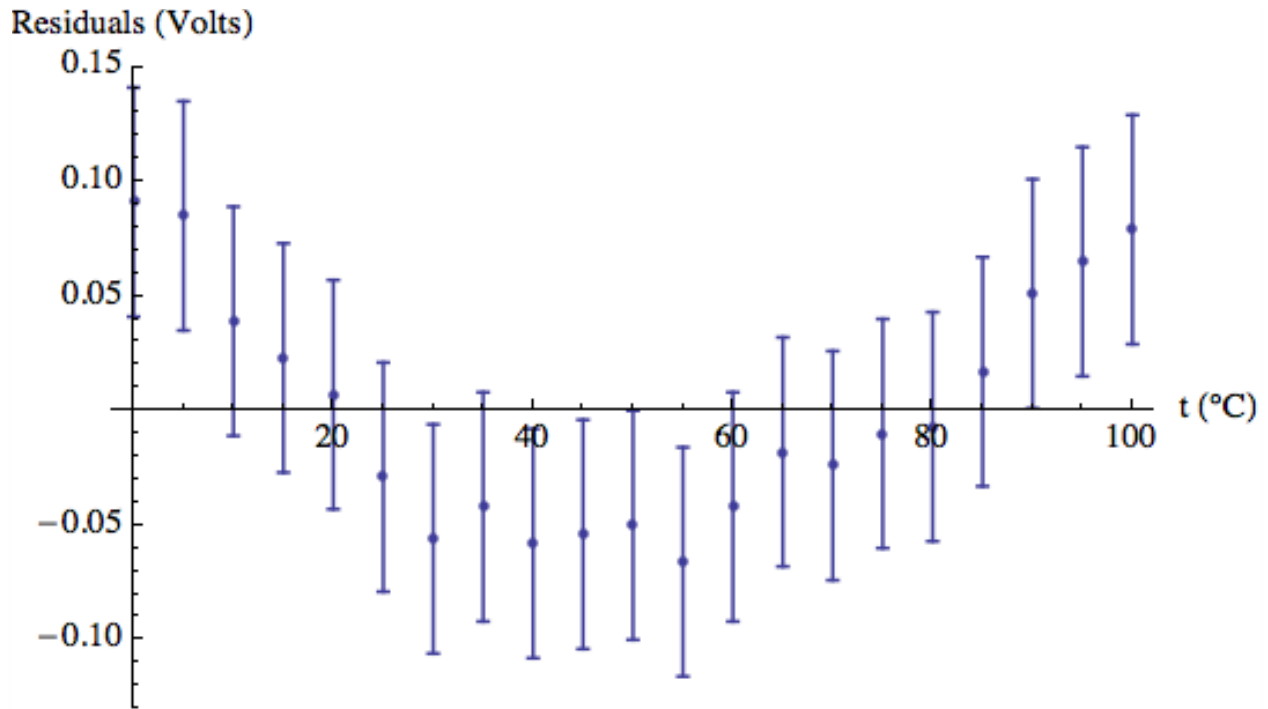


uncertainty to the voltage, but not to the temperature. Also shown in the figure is the result of fitting the data to a straight line. The results of the fit were:

```
slope: 0.0412 ± 0.0004
intercept: -0.98 ± 0.02
chi-squared: 21.05
degrees of freedom: 19
```

- B. Just from the numerical results of the fit and from looking at the figure above, is this a good fit to a reasonable model?
- C. The Figure below shows a plot of the residuals. Now what do you think of the fit?



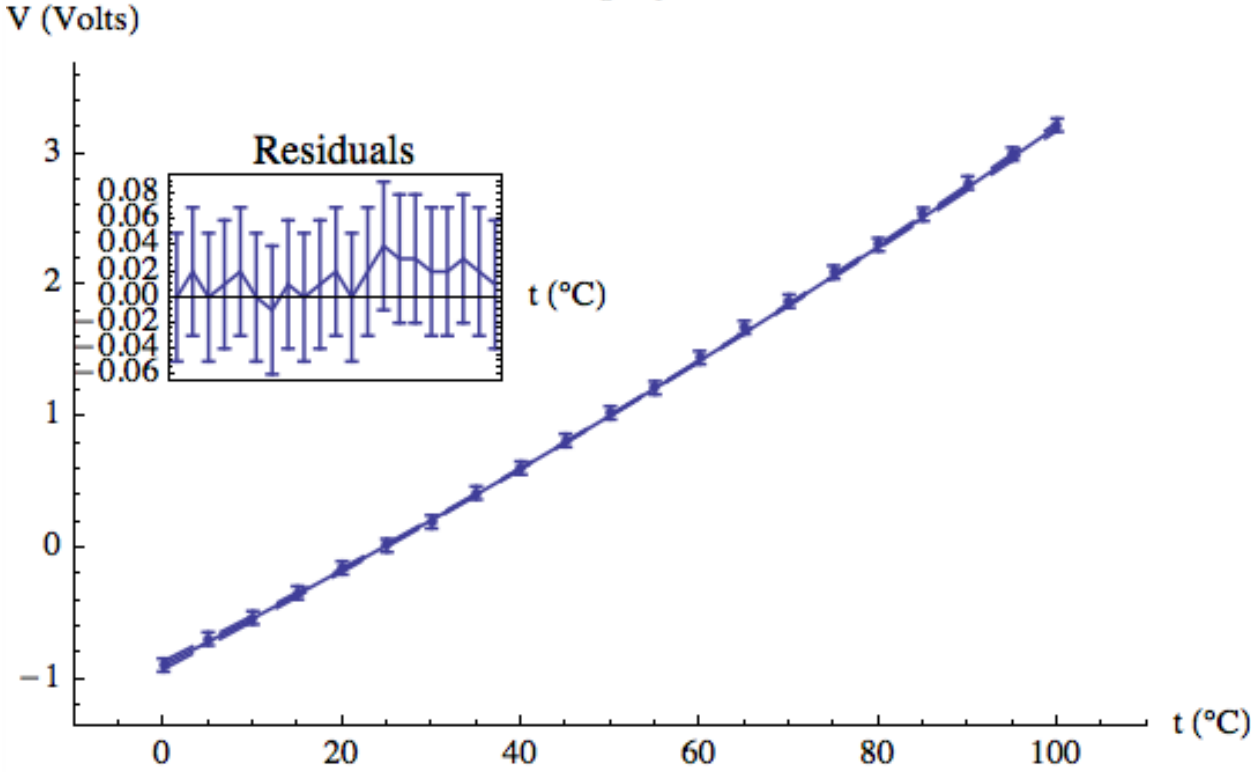


D. We add a quadratic term to the fit, so we are fitting to:  $V = mt + b + ct^2$ . The numerical results of the fit are:

slope:  $m = 0.035 \pm 0.001$   
intercept:  $b = -0.89 \pm 0.03$   
quadratic term:  $c = -.00006 \pm 0.00001$   
chi-squared: 1.007  
degrees of freedom: 18

The graphical result of the fit including a plot of the residuals as an insert is shown in the Figure below. Is this a good fit? Are there any problems with it? If so, what are they and how can they be explained?

# Calibrating to a 2nd order polynomial



This Guide was originally written by David M. Harrison, Dept. of Physics, Univ. of Toronto, September 2013. It was updated for at-home learning by Jason Harlow March 2021.