Purpose: In this lab, we will learn graphical techniques for analyzing data.

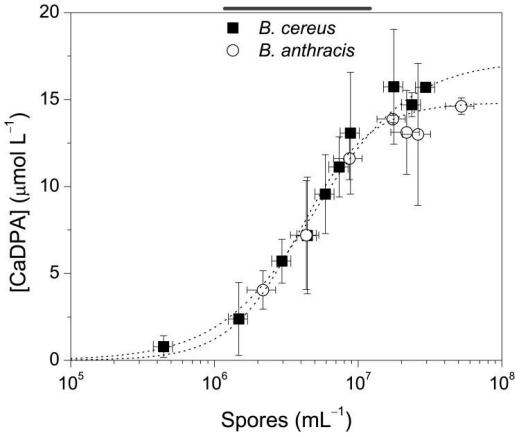
Theory:

One of the most important aspects of experimental science is the analysis of measured data. In this lab, we are going to focus on graphical analysis of measured data through a mathematical or computational procedure called curve fitting or regression, which is a tool used not just in physical science, but also in life sciences, social sciences, economics, and medicine.

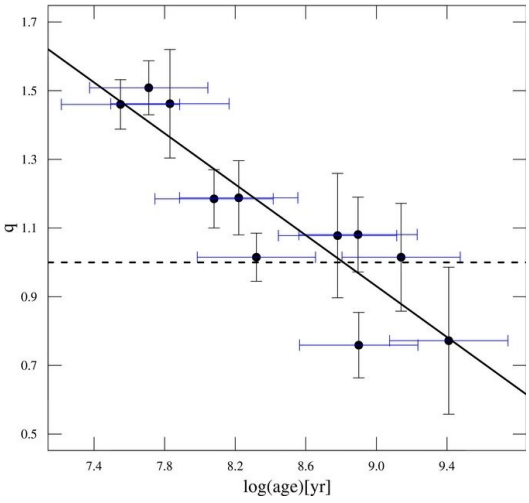
First, we will need to set up the coordinate system and graph the data points. Features of a good graph include:

- the independent variable on the horizontal axis and the dependent variable on the vertical axis
- a title for the graph that is dependent vs. independent. For example, “velocity vs time.”
- axis labels that include the variable and its unit. For example, if time is on the horizontal axis, the axis label would be t (s).
- equally spaced tick marks with at least every fifth tick mark labeled with its numerical value.
- data points marked with small, but not invisible dots.

The **independent variable** in an experiment does not depend on any other variable in that experiment. **Dependent variables** depend, by some physical or mathematical relation, on the other variables in the experiment. For example, if in a lab, we were analyzing the motion of a ball thrown vertically upward in the air, we might measure the velocity and position of the ball every 0.25 seconds. In that situation, the time doesn’t depend on anything – time is progressing forward at its own rate, which doesn’t depend on what is happening with the ball. Time is the independent variable and position and velocity are both dependent variables.



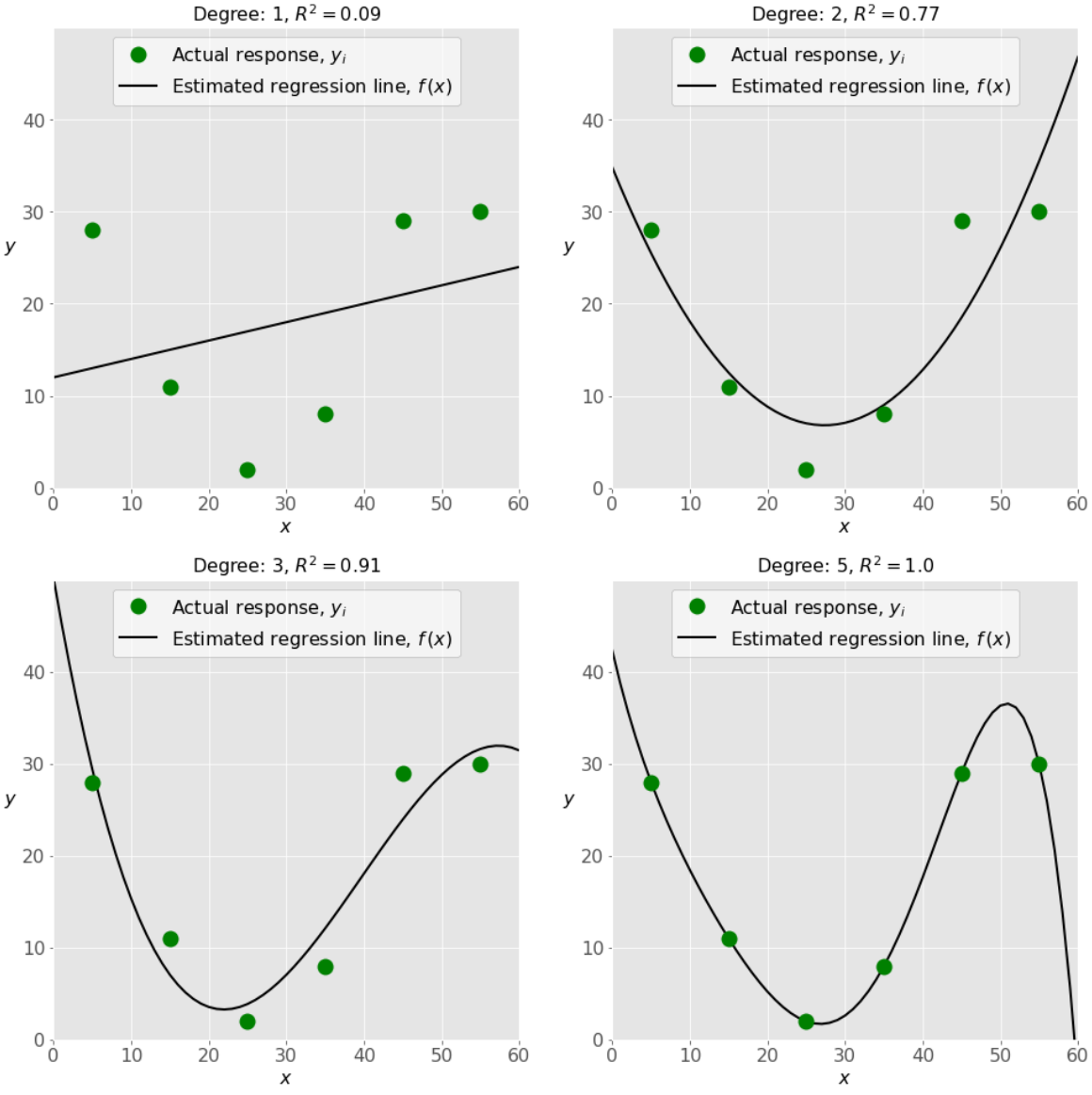
Second, we will need to include **error bars** for our data points. Any measured quantity has an uncertainty. The uncertainty in the measurements of the independent variable will determine the size of the horizontal error bars and the uncertainty in the measurements of the dependent variable will determine the size of the vertical error bars on your graph. Some examples of graphs with both horizontal and vertical error bars are shown.



Finally, we will add the **regression curve** to the graph. Regression curves can also be called regression lines, trend-lines, or lines of best fit. Regression involves fitting or correlating data to some mathematical function, where the function may be linear, quadratic, logarithmic, exponential, or something else. The end results of regression are the fit parameters of the regression curve and a goodness-of-fit parameter that tells you how well the regression curve matches the data. There are many different methods for curve-fitting and associated goodness-of-fit parameters.

In this lab, we will be using linear regression, power series regression, polynomial regression, and a **goodness-of-fit parameter** called the coefficient of determination or **R^2**.

The goodness-of-fit parameter R^2 is the proportion of the variation in the dependent variable that is predictable from the independent variable(s). If the regression curve is a perfect fit for the data, $R^2 = 1$. If the regression curve doesn't match the data at all, $R^2 = 0$. Values of R^2 between 0.9 and 1.0 are pretty good fits. In the following figures, notice that the linear (polynomial degree 1) regression curve doesn't fit the data well, while the quadratic (polynomial degree 2) and cubic (polynomial degree 3) regression curves are better. The degree-5 polynomial is able to perfectly fit the data, but this suffers from a problem called over-fitting. If you make the functional form of your regression curve complicated enough, you can always get a perfect fit with $R^2 = 1$, but it usually doesn't reflect reality. Overfitting occurs when the number of fit parameters is greater than or equal to the number of data points. In this case, a polynomial of degree 5 has the functional form $y = Ax^5 + Bx^4 + Cx^3 + Dx^2 + Ex + F$, with 6 fit parameters A, B, C, D, E, and F. There are only 6 data points, so this is overfit. Avoiding overfitting is one reason why taking lots of data is important in science.



Another way we can avoid overfitting data is to use our knowledge from theoretical physics to guide our choice of the functional form of the regression curve. For example, if we were analyzing the motion of a ball thrown vertically upward in the air, we might measure the velocity and position of the ball every 0.25 seconds. Then we would make two graphs, velocity vs time and position vs time.

For the graph of velocity vs time, we would choose a linear regression curve, because we know that one of our kinematic equations is $v = v_0 + at$, which means the velocity depends linearly on time. If our regression curve $y = mx + b$ had fit parameters $m = -9.76$ and $b = 3.02$, our measured acceleration would be $-9.76\ m/s^2$ and our measured initial velocity would be $+3.02$ m/s. Notice that the slope of the graph of velocity vs time is acceleration, which makes sense when you compare the definition of average acceleration (change in velocity over change in time) with the definition of slope (rise over run).

For the graph of position vs time, we would choose a quadratic (polynomial degree 2) regression curve, because we know that one of our kinematic equations is $y = y_0 + v_0t + \frac{1}{2} at^2$, which means the position depends quadratically on time, since the highest exponent on the independent variable is 2.

Procedure Part 1:

In this section of the lab, we will investigate the linear relationship between circumference and diameter for a cylinder.

1. Measure the diameter D and circumference C of 5 cylinders. Each measured value should have uncertainty and units, in the form _____ ± _____ cm.
    a. To measure the diameter, use a Vernier caliper.
    b. To measure the circumference, wrap a piece of string around the cylinder and then measure the length of the string using a ruler.
    c. The data point (0,0) means a cylinder of zero diameter (infinitesimally thin) would have zero circumference.

Table 1

	Diameter (cm)	Circumference (cm)
	0	0
Cylinder 1		
Cylinder 2		
Cylinder 3		
Cylinder 4		
Cylinder 5		

2. Each lab partner should separately create a graph of C vs. D by hand on the graph paper provided. The terminology C vs. D means that D is the independent variable, on the horizontal axis, and C is the dependent variable, on the vertical axis.
    a. Using graph paper, set up a coordinate system that will allow you to fit all the data onto the paper. Make sure your graph has a title, axis labels, and evenly spaced tick marks.

b. Graph the data points using small dots and the uncertainties using horizontal and vertical error bars. If the error bars are smaller than the size of your dots, you don't need to graph them.

c. Use a ruler to draw a line that comes close to the data points. For data points that don't fall on the line, about half should be above and half below the line.

d. Choose two points on your line (not necessarily data points) and calculate the slope m of your hand-drawn line using rise / run. Note the position of your vertical intercept (0,b).

Hand-drawn m = _____

Hand-drawn b = _____

e. For this graph, does the slope m have units, or is it a unitless quantity? Explain.

f. For this graph, does the vertical intercept b have units, or is it a unitless quantity? Explain.

Compare the slope calculated from your hand-drawn line to the one your lab partner found. These two values will probably be slightly different, which isn't surprising since drawing the lines involved visual estimation.

3. A more reliable slope is obtained using a algorithm called least squares fitting or even more sophisticated statistical techniques. We will now use graphing software to find the equation of the best-fit line to your data.

a. Open Google Sheets on either the lab computer or your laptop, logging in with your UM student email address. Change the name of the file from "Untitled Spreadsheet" to "Phys 223 Lab 3". Your data will save to the cloud.

b. Share the file with your lab partner (Select Share under the File Menu). You should do this for every lab that requires data tables or graphing.

c. Enter the diameter data in Column A and the circumference data in Column B.

d. Highlight your data and then select Insert, Chart, Scatter Chart to create a graph of C vs. D.

e. Investigate the settings in Google Sheets and figure out how to add the following to your graph:

- a title "Circumference vs Diameter"
- axis labels "Circumference (cm)" and "Diameter (cm)"
- vertical error bars
- a linear trendline, which is determined by a least squares algorithm. You want to label the trendline with its equation. We are choosing a linear fit in this case because we know from theory that C is linearly dependent on diameter. Note the option to show R^2. You can't display both R^2 and the error bars at the same time, but turn off the error bars temporarily to note the goodness-of-fit parameter.

R^2 = _____

f. Comparing the trendline equation with the general form of a line, y = mx + b, what are your experimental values for the slope m and the vertical intercept b? Include units, if appropriate.

Least-squares m= _____

Least-squares b = _____

4. Since the cross-section of a cylinder is a circle, the theoretical prediction for the circumference C as a function of diameter D is the same as for a circle: $C = \pi D$. Comparing that equation with the general form of a line, y = mx + b, what are your theoretical predictions for the slope m and the vertical intercept b? Include units, if appropriate.

Theoretical m = _____

Theoretical b = _____

5. Find the percent error in your hand-drawn and least-squares fit values of the slope of C vs. D. Recall that percent error compares a measured value with a theoretical value, while percent difference compares two measured values and make sure you use the correct equation to find these two percent error values. Show your work and follow the rules for the number of significant figures to report for calculated values.

Percent error in hand-drawn m = _____

Percent error in least-squares m = _____

6. Staple both graphs (hand-drawn and computer-generated) to this worksheet.

Procedure Part 2:

In this section of the lab, we will investigate the non-linear relationship between volume and radius for a sphere.

7. Measure the radius r and volume V of 5 spheres. Each measured value should have uncertainty and units. Each calculated value should follow significant figure rules.
    a. To measure the diameter, use a Vernier caliper.
    b. The radius of a sphere is half the diameter.
    c. To measure the volume, we will use the displacement method.
        - Partially fill the graduated cylinder with water, using a pipette to fine-tune the meniscus to be exactly on one of the tick-marks so you have an accurate measurement of the initial water level.
        - Carefully submerge the sphere (no splashing!) and measure the final water level.
        - Subtract (final water level – initial water level) to get the volume of the sphere.
        - Readjust the water level and repeat for each sphere.

Table 2

	Diameter (cm)	Radius (cm)	Initial Water Level (mL)	Final Water Level (mL)	Sphere volume (1 mL = 1 cm^3)
		0			
Sphere 1					
Sphere 2					
Sphere 3					
Sphere 4					
Sphere 5					

8. The unit of mL is equivalent to a cubic centimeter. Do a unit conversion calculation to prove that 1 mL = 1 cm^3, using the definition that 1000 L = 1 m^3.

9. Add another sheet to your Google Sheets document, enter your data, and create a graph of V (cm^3) vs r (cm). Make sure you include a title and axis labels (as you will in all future labs also).

10. We want to include a least-squares fit to the data. A linear trendline is NOT the best choice (which you can see from the value of the goodness-of-fit parameter R^2 if you attempt a linear fit). Explore the other "trendline" options and see which function type (linear, exponential, polynomials of different orders, power series, etc.) of regression curve seems to fit the data best without risk of overfitting. The degree (also called order) of a single-variable polynomial or power series is the highest exponent that appears. For example, a power series of order 2 has a functional form $y = Ax^2$ and a polynomial of order 2 has a functional form of $y = Ax^2 + Bx + C$.

11. Sometimes, if we have no theoretical prediction, we choose a functional form just based on goodness-of-fit (getting R^2 as close to 1 as we can without over-fitting). But often, as is the case for this lab, theory (as opposed to experiment) will guide our choice of a functional form for the regression curve. The theoretical formula for the volume V of a sphere of radius r is $V = (4/3)\pi r^3$. Based on that, which function type (linear, exponential, polynomials of different orders, power series, etc.) should we choose for the regression curve?

    _____

12. Display the equation for your "trendline" aka regression curve. What is the leading coefficient A? (Hint: the leading coefficient for a polynomial means the number in front of the term with the largest exponent.)

    Least-squares A = _____

13. Using $V = (4/3)\pi r^3$, what is the theoretical prediction of A?

    Theoretical A = _____ ≈ _____

14. Find the percent error. Show your work.

    Percent error for least-squares A = _____

15. Staple a printout of your computer-generated graph to this worksheet.

16. If you are using a lab computer, **LOG OUT** of Google Sheets.

Pre-lab **Name: _____ Section:_____**

Read over the theory and procedure for this lab before completing this pre-lab.

1. What is the purpose of this lab?

2. Imagine if there was a part 3 of this lab procedure, where you draw a set of squares, with side lengths of your choosing, and then measure their areas by counting squares on graph paper. Let the side length of the square be the variable x and the area be A.

a) Is A or x the independent variable in this hypothetical situation?

b) If you wanted to construct a graph of your data, would A or x be on the vertical axis?

c) What would be the title of your graph, Area vs Side Length or Side Length vs Area?

d) Thinking about the (well-established) theoretical equation for the area of a square $A=x^2$, which function type would you choose for a regression curve? Choices are: linear, exponential, polynomials of order 2, polynomial of order 3, power series, etc.

e) Still considering the equation $A=x^2$, what is the theoretical prediction for the leading coefficient of the regression curve?

f) Use the data shown in the table.

Side length x (cm)	Area A (cm^2)
0	0
3.00 ± 0.05	9.0 ± 0.5
4.15 ± 0.05	17.0 ± 0.5
5.85 ± 0.05	34.5 ± 0.5
7.20 ± 0.05	51.5 ± 0.5

Refer to Step 3 of the procedure for instructions on how to create a graph in Google Sheets. Create a graph of A vs. x with a title “Area vs Side Length”, axis labels “Area (cm^2)” and “Side Length (cm),” vertical error bars, a “trendline” aka regression curve, and the “trendline” labeled with its equation. Staple a printout of your computer-generated graph to this worksheet.