Laboratory 3: Method of Least Squares

Introduction

Consider the graph of experimental data in Figure 1. In this experiment x is the independent variable and y the dependent variable. Clearly they are correlated with each other and the correlation seems to be linear. We would like to find the line of the form

$$y(x) = a + bx \tag{1}$$

that best fits this data, [Which in truth actually consist of pairs of measurements (x_i, y_i)]. While we can eyeball the line it would be nice to have an objective well established method to determine the values for a and b. The most common method to do this is called the *method of least squares*.

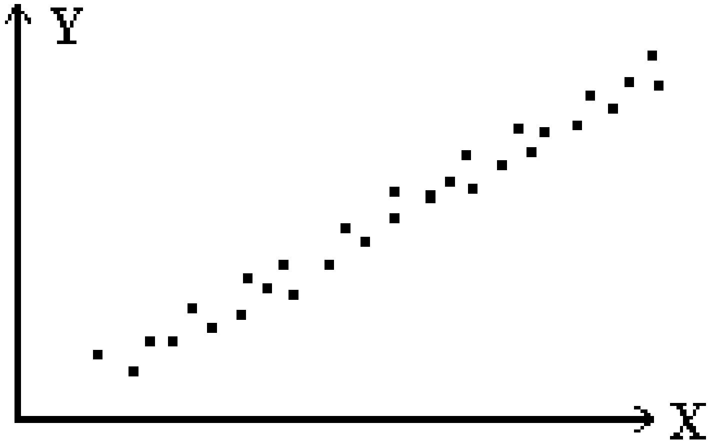

Figure 1: Plot of Y versus X

Our task is to determine the coefficients *a* and *b* in such a way that the discrepancy between the values of our measurements y_i, and the corresponding fitted values $y(x_i)$ are *minimized*. We cannot determine the coefficients exactly with only a finite number of observations, but we do want to extract from these data the *most probable* estimates for the coefficients. Note that this technique assumes that the uncertainties are all in the y variable and follow Gaussian statistics as is often the case for uncertainties arising from fluctuations in repeated readings of the instrumental scale caused by settings are not exactly reproducible. These uncertainties are called instrumental regardless of whether they are due to imperfections in the equipment or to human imprecision.

Method of Maximum Likelihood

Our data consist of a sample of observations extracted from a parent population which determines the probability of making a particular observation. Let us define parent coefficients a_o and b_o such that the *actual* linear relationship between y and x is given by

$$y_0(x) = a_o + b_0 x \tag{2}$$

For any given value x_i we can calculate the probability density $\Omega_i(a, b)$ for making the observed measurement y_i, by assuming a Gaussian distribution about the actual value $y_o(x_i)$ with a standard deviation $\sigma_i = \sigma$, i.e.

$$P(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(y_i - y_0(x_i))^2}{2\sigma^2}} \tag{3}$$

The probability of making simultaneous measurements of all N y_i is the product of these probabilities

$$\Omega(a_0, b_0) = \prod_{i=1}^{N} \Omega(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\sum_{i=1}^{N} \frac{(y_i - y_0(x_i))^2}{2\sigma^2}} \tag{4}$$

Of course, we do not know the parent distribution values for a_o and b_o, but for any *estimated* values of the coefficients *a* and *b*, we can calculate the probability density for making the observed set of measurements as

$$\Omega(a, b) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\sum \frac{(\Delta y_i)^2}{2\sigma^2}} \tag{5}$$

where $\Delta y_i = y_i - (a + b x_i)$.

The method of maximum likelihood consists of making the assumption that by maximizing equation (5) for the observed set of measurements we are most likely to obtain the best estimates for a_o and b_o. Maximizing the probability $\Omega(a, b)$ is equivalent to *minimizing* the sum in the exponential. We define the quantity chi squared, χ^2, to be the sum in the exponential:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{\Delta y_i}{\sigma} \right)^2 \tag{6}$$

We have used the same symbol χ, defined in Experiment 2, because this is essentially the same definition in a different context. Our method for finding the optimum fit to the data will be to minimize this sum of *squared* deviations and, hence, to find the fit which produces the smallest χ^2.

Note: Our development assumes the error associated with any single measurement is the *same* for all measurements. Modifications to the development must be made when this is not so.

Minimizing χ^2

In order to find the values of the coefficients *a* and *b* which yield the minimum value for χ^2 we use the methods of the calculus, i.e.

$$\frac{\partial \chi^2}{\partial a} = 0 \tag{7}$$

And

$$\frac{\partial \chi^2}{\partial b} = 0 \tag{8}$$

Rearranged these equations yields a pair of simultaneous equations to be solved for the coefficients *a* and *b*. This will give us the values of the coefficients for which χ^2 is minimized. This is done with the determinants below. In these equations be sure to distinguish the difference between the *square of the sum* of x_i, $(\sum x_i)^2$, and the *sum of the squares* of x_i, $\sum x_i^2$:

$$\begin{gathered} a = \frac{1}{\Delta}\left(\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i\right) \\ b = \frac{1}{\Delta}(N \sum x_i y_i - \sum x_i \sum y_i) \\ \Delta = N \sum x_i^2 - (\sum x_i)^2 \end{gathered} \tag{9}$$

Estimation of Errors

In order to find the uncertainty in the estimation of the coefficients *a* and *b* in our fitting procedure, we refer to our discussion of the propagation of errors in Laboratory 1. Each of our data points y_i have been used in the determination of the parameters, and each has contributed some fraction of its own uncertainty to the uncertainty in our final determination. Ignoring systematic errors which would introduce correlations between the uncertainties, the standard deviation σ_z of the determination of any parameter z is given by

$$\sigma_z^2 = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial z}{\partial y_i} \right)^2 = \sigma_i^2 \sum_{i=1}^{N} \left(\frac{\partial z}{\partial y_i} \right)^2 \qquad (10)$$

If we assume that the uncertainties are instrumental and all the same, they can be estimated from the data. Our definition in Laboratory 1 of the sample variance s^2, which approximates σ^2, is the sum of the squares of deviations of the data points from the calculated mean divided by the number of degrees of freedom. In this case, the number of degrees of freedom is the number of data points N minus the number of parameters (two) which we determined before calculating s^2. Thus, our estimated parent standard deviation $\sigma_i = \sigma$ is

$$\sigma^2 \cong s^2 = \frac{1}{N-2} \sum_{i=1}^{N} (y_i - a - bx_i)^2 \qquad (11)$$

Note that it is this common uncertainty, σ, which we have minimized by our least-squares fitting procedure. The derivatives in equation (10) can be evaluated by taking the derivatives of equations (9), and we can find an expression for the uncertainties in parameters *a* and *b*, i.e.

$$\boxed{\begin{aligned} \sigma_a^2 &\cong \frac{\sigma^2}{\Delta} \sum x_i^2 \\ \sigma_b^2 &\cong \frac{N\sigma^2}{\Delta} \end{aligned}} \qquad (12)$$

It may not be obvious from these forms, but the larger the number of data points the smaller is the error in the quantities *a* and *b*.

Using R to graph and do least squares regression.

The two most important commands are xyplot for graphing and lm for fitting a linear model. At their most basic they have the syntax: xyplot(Y~X, data=data_frame) and lm(Y~X, data=data_frame). An example is useful, we will use the rubberband dataset that comes as part of R. First we must bring fastR and rubberband into our current session:

- require(fastR)
- data(rubberband)

Next we will look at the structure of the dataset:

- head(rubberband)

Graphing

From this you will see that there are two variables: Stretch and Distance. Stretch is the one we have control over so it is the independent variable and Distance is the dependent. Use xyplot to graph the data, use the variable names separated by a tilde and specify the data as coming from the data_frame rubberband.

- xyplot(….)

We will now complicate the xyplot command by adjusting the type. By default we have been using type ‘p’ (for point). We can make this explicit:

- xyplot(Y~X, data=data_frame, type=c(‘p’))

Using this command should have left your graph unchanged. There are other types as well, the other one that is relevant is type ‘r’ for regression.

- xyplot(Y~X, data=data_frame, type=c(‘r’))

What is most useful is to make both on the same graph:

- xyplot(Y~X, data=data_frame, type=c(‘p’, ‘r’))

Fitting

It is useful to see the least-squares regression line graphed with the data to make decisions about how meaningful the fit is, but it is useful to know the coefficients in the equation for the line

$$Distance = a + b\, Stretch. \tag{13}$$

Try using lm:

- lm(Distance~Stretch, data=rubberband)

This gives the most basic information, the intercept (a) and the coefficient that multiplies Stretch (b). We can extract more information if we look in more detail at the linear model structure, first by storing the model and then by looking at its summary:

- flight.model<-lm(Distance~Stretch, data=rubberband)
- summary(flight.model)

This gives a variety of information including estimates for the coefficients, uncertainties in the coefficents, probabilities that the coefficients values could be explained by chance instead of a real correlation.

The last thing we need is an estimate of the variance in the parent distribution, s. This can be estimated from the difference between the linear model and the data points, namely the residuals. The residual of each data point is held by the linear model variable residuals

- flight.model$residuals

and also returned by the function resid(linearModel)

- resid(flight.model)

From Equation (11) we want the sum of the residuals squared divided by the number of degrees of freedom, which is also a model variable. Casting Eq. (11) in the syntax of R

- s<-sqrt(sum(resid(flight.model)^2)/flight.model$df.residual)

If you look at the summary of our model again you will see that this quantity was already reported as the *residual standard error*. While you are looking at that part of the summary you will also find the R^2 value. It is worth noting that this gives the percent of the variance in the data that is explained by the model.

Experimental Procedure for the General Case

Given the data shown in Table 1 below:

Table 1: Experimental data for temperature versus position along a rod

Trial	X_i (cm)	T_i (^oC)
1	1.0	15.6
2	2.0	17.5
3	3.0	36.6
4	4.0	43.8
5	5.0	58.2
6	6.0	61.6
7	7.0	64.2
8	8.0	70.4
9	9.0	98.8

1. Determine the best parameters *a* and *b* to the equation $T_i = a + bx_i$. Note that we are assuming that all of the error is associated with a measurement of temperature, *not* length.

2. Determine the standard deviation of the temperature data.
3. Determine the errors associated with *a* and *b*.
4. Express the thermal gradient in a manner suitable for reporting; i.e., the slope of the line.
5. Plot the data and fit.
6. Use the standard deviation s to draw an error bar on each data point of temperature ± s. An error bar about a data point reflects the probability that another measurement would reproduce the first value within one standard deviation 66% of the time. To do this in R, you can use the following xyplot() command, adjusting X, Y, s, and data to fit your situation.

- xyplot(Y~X, data, lb=data$Y-s,ub=data$Y+s,
- panel=function(x, y, lb, ub, …){
- panel.xyplot(x, y, type=c('b','r'), …)
- panel.segments(x0=x, x1=x, y0=lb, y1=ub, …)
- }
-)

7. These data were made up, but a normal thermometer might have a ΔT = 0.5°C. How does this compare to s?

Specific Case for Constraint *a* = 0

An important subset of the general problem discussed above is the linear function constrained to pass through the origin, i.e.

$$y(x) = c\,x \tag{14}$$

This is a linear problem but it represents a function which has a y intercept of zero. We follow the same approach as above starting with the definition of χ^2, and we proceed as in the previous section and find the minimum of χ^2 with respect to c by letting:

$$\frac{\partial \chi^2}{\partial c} = 0 \tag{15}$$

The standard deviation associated with a data point y_i in this case is

$$\sigma^2 \cong s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - cx_i\right)^2 \tag{16}$$

where the N - 1 factor appears instead of N - 2 since we have only used the data once to determine c.

To find the error in the slope we find the change in c with respect to y_i, i.e. $\frac{\partial c}{\partial y_i}$, then we obtain

$$\sigma_c = \frac{\sigma}{\sqrt{\sum_{i=1}^{N} x_i^2}} \tag{17}$$

Thus, equations (15), (16), and (17) provide us with the linear least squares fit of data constrained to go through the origin.

Experimental Procedure for the Specific Case of Data Passing Through the Origin

Part One:

You are given a set of aluminum disks with different diameters.

1. Measure and record the diameter of each disk and label these data points d_i.
2. Devise a technique for measuring the circumference of the disk and label these data points as $circ_i$.
3. Determine the best parameter c fit to the equation $circ_i = c\ diam_i$. In R this is done by giving the intercept an explicit value of 0
   - Disk.model<-lm(circ~diam+0, data=Disk)

   Note that we are choosing to treat the diameter as the independent variable, this is arbitrary and is equally valid the other way around.
4. Determine the standard deviation associated with each C_i.
5. Determine the error associated with c.
6. Use your model to predict a circumference based on the model for a diameter of 10cm.
   - p<-predict(Disk.model, newdata=data.frame(diam=10))
7. Use your model to predict a circumference for all diameters in your dataset.
8. Use the results of 7 to plot the data and the intercept=0 fit,
   - xyplot(circ+p~diam,data=Disk, type='b')
9. Record the best fit parameter for the slope as well as for the estimated error in the slope. Compare to the known value of π and discuss the significance of your results for the goodness of your data.

Part Two:

You are given a voltage supply, a decade resistor box, an analog current meter, and a digital volt meter. Set the analog current meter to the 250 milliamp scale and the decade resistor box to 10 ohms. Use the digital meter to record the resistance of the decade resistor box at this setting when it is not yet connected to anything else. Now, connect the voltage supply, the resistor box, and the current meter into a simple series circuit. Place the digital volt meter in parallel across the terminals of the voltage supply and set it to measure DC voltages up to 200 mV. Using this set up:

1. Measure and record the current through the resistor as well as the voltage of the power supply for at least 10 different voltages between 20 and 200 mV. (**NOTE:** be sure not to exceed 200 mV so as not to damage the current meter).
2. Plot the voltage versus current data you collected using R and find what it determines to be the best fit parameter for the slope as well as the estimated uncertainly in the slope. Make sure to constrain the intercept to zero when fitting the line to your data.
3. Using Ohm's Law (V=IR), compare what you found from the graph for the resistance of the decade resistor box to what the digital meter found it to be. If they do not agree within the level of your experimental precision, discuss possible reasons for the differences. Where possible site specific measurements or estimated values taken from the laboratory that support your suppositions for the sources of the error (i.e. do not just say human error in reading the meters and leave it at that.)

WHEN YOU FINISH: LEAVE THINGS AS YOU FOUND THEM!