ASTR469 Lecture 18: Statistics and Fitting

Assess yourself/study guide after lecture & reading (without peeking at notes)...

1. Assume a and b are constants, x and y are variables with standard deviations σ_x and σ_y. Assume covariance is zero. Find the expression for the standard deviation for f in each of the below cases.

   (a) $f = x + a$

   (b) $f = x + y$

   (c) $f = ax + by$

   (d) $f = axy$

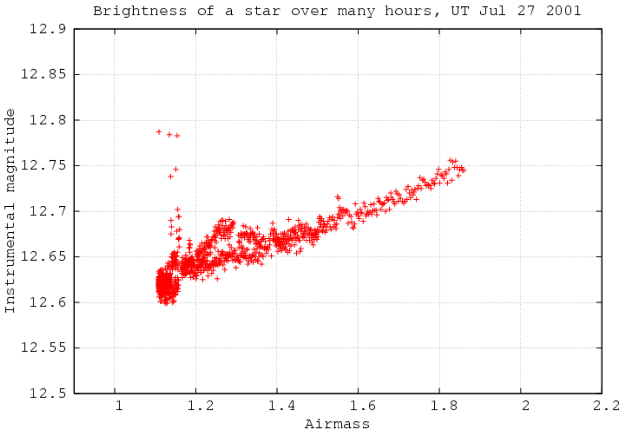
   (e) $f = ax^b$

   (f) $f = ae^{bx}$

2. Try using GNUPLOT to instead fit a quadratic function to the spectral line data I gave you (using function `f(x) = a*x**2 + b*x + c`). What values does it come up with?

3. If you tried the previous point, you would find that a ends up being consistent with zero, with a huge error on the measurement of a (something like 0.001 +/- 5.0). The $\chi^2_{\rm red}$ value quoted for the quadratic fit is similar to that of your linear function. Ask yourself: which function was a better choice: linear or quadratic? Why?

4. Let's say you measured the magnitude of a star over a few hours one evening, and found the magnitude vs. airmass plot looked like a plot we have seen before:

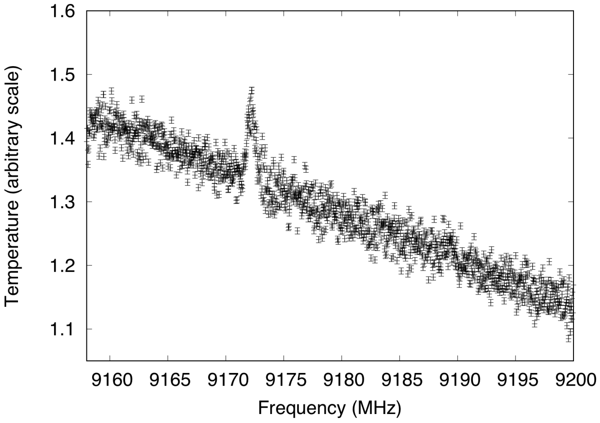


   How exactly would you go about determining precise values and measurement errors for the k value and unextincted magnitude of this star?

1 The Reason for Astro-statistics

You know have the theoretical background relevant to understanding actual observations: how to set them up and how to understand the measurements that come out. After you get your data, you can do actual science! But to do that, often we need to use some basic statistical tools for the analysis and interpretation of our data.

Example: Here is a simple spectral line observation.



Information you might want to know about this line:

- What's the peak frequency and width of this spectral line?
- What's the S/N of this detection?
- What is the shape/slope of the background (non-line) emission?
- What are the measurement errors in the measurements listed above?
- (More complex question) Is this line made up of distinct, contributing gas clouds with slightly different velocity distributions, or does it seem to come from only one cloud?

These are all fairly basic things to know, and certainly you could do some of them in a very loose way by estimating from the plot. But we are scientists, and need to quantify measurements a bit more precisely; we also require the study of uncertainties to determine the reliability of results. There is uncertainty (measurement error) associated with all measurements. It is absolutely necessary to understand and accurately depict your uncertainties when presenting your data.

This is not just true for fitting a spectral line; you might want to fit a line to a scatter plot, estimate the uncertainty in a single measurement, or any number of other things. So let's discuss a few very common methods in astronomical data analysis and a bit of background theory.

2 Very Basic Statistics

Let's say we are studying the relationship $F = ma$. And let's assume we know a absolutely (assume it is due to gravity). We have measured m 100 times and plotted a histogram, and it looks like a Gaussian distribution. This is the "normal distribution" or "bell curve" that we saw in the photometry lecture. The general form is:

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

The parameter μ in this definition is the mean of the distribution (and also its mode). The parameter σ is the standard deviation. A random variable with a Gaussian distribution is said to be **normally distributed** and is called a **normal deviate**.

The stuff out front sets the amplitude of the distribution's peak, which happens when $x = \mu$. The form above is normalized to 1 (area under the curve is 1). For our example, $f = F$ and $x = m$.

2.1 Mean, Standard Deviation, Variance, and RMS

As we saw in the photometry lecture, σ is the "standard deviation" that encompasses 68% of a Gaussian curve. This is what we usually quote as the measurement error on some measurement (assuming that the measurement uncertainty is Gaussian-distributed, which is usually true in astronomy—recall we previously saw the Poissonian distribution for count rates, and that the Poissonian distribution is close to Gaussian with high counts).

We can actually measure our standard deviation by taking multiple measurements and plotting a histogram of the results, then finding the best-fit Gaussian curve by varying μ and σ. We can also compute the μ and σ for a list of N measurements. The mean is easy since it's just the average of all the values:

$$\mu = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2}$$

In statistical analysis one often also uses the *variance* of a list of N measurements, which is the square root of the standard deviation and given by:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{3}$$

There is one other quantity often quoted, which is the "root mean squared." This is computed as:

$$\text{RMS} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2} \tag{4}$$

As you can see by looking at the last two equations, RMS $= \sigma$ when the mean of the data is zero (that is, the root-mean-square will equal the standard deviation if your data is completely noise-like, and contains no "signal").

2.2 Error propagation

Now assume we have some function $f(x, y, z, ...)$ and are separately measuring multiple parameters $(x, y, z...)$ within this function. We might ask ourselves: What is the error on f based on the measurement errors on the separate parameters?

We can Taylor expand the function if the uncertainties are small:

$$f - \bar{f} \approx \frac{\partial f}{\partial x}(x - \bar{x}) \tag{5}$$

or

$$\delta f \approx \left| \frac{\partial f}{\partial x} \right| \delta x \tag{6}$$

where under normal circumstances $\delta x = \sigma_x$. The variance is therefore

$$\sigma_f^2 = \frac{1}{N-1}\sum_{i=1}^{N}(f_i - \bar{f})^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\partial f}{\partial x}(x_i - \bar{x})\right)^2 \tag{7}$$

If we have two variables, x and y,

$$f - \bar{f} \approx \frac{\partial f}{\partial x}(x - \bar{x}) + \frac{\partial f}{\partial y}(y - \bar{y}) \tag{8}$$

which gives us

$$\sigma_f^2 = \frac{1}{N-1}\sum_{i=1}^{N}(f_i - \bar{f})^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left[\frac{\partial f}{\partial x}(x_i - \bar{x}) + \frac{\partial f}{\partial y}(y_i - \bar{y})\right]^2 \tag{9}$$

after some algebra, this leads to

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\sigma_{xy}^2 \tag{10}$$

If the variables are independent, the uncertainty in the function *f(x,y,z,. . . .)*, σ is

$$\sigma = \left[\left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial f}{\partial z}\right)^2 \sigma_z^2 \ldots\right]^{0.5}$$

Where the various σ terms are the uncertainties on the subscripted variables.

This formula is only strictly true for when the σ values are small compared to the partial derivatives. If the uncertainties are not independent, there are "covariate" terms, but we will mostly exclude those from our discussion.

Example formulas

Function	Standard deviation
$f = aA$	$\sigma_f = a\sigma_A$
$f = aA + bB$	$\sigma_f = \sqrt{a^2\sigma_A^2 + b^2\sigma_B^2 + 2ab\sigma_{AB}}$
$f = aA - bB$	$\sigma_f = \sqrt{a^2\sigma_A^2 + b^2\sigma_B^2 - 2ab\sigma_{AB}}$
$f = AB$	$\sigma_f \approx \lvert f\rvert\sqrt{\left(\frac{\sigma_A}{A}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2 + 2\frac{\sigma_{AB}}{AB}}$
$f = \frac{A}{B}$	$\sigma_f \approx \lvert f\rvert\sqrt{\left(\frac{\sigma_A}{A}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2 - 2\frac{\sigma_{AB}}{AB}}$
$f = aA^b$	$\sigma_f \approx \lvert abA^{b-1}\sigma_A\rvert = \left\lvert\frac{fb\sigma_A}{A}\right\rvert$
$f = ae^{bA}$	$\sigma_f \approx \lvert f(b\sigma_A)\rvert$

We can generally assume the covariance (σ_{AB}) is zero.

3 Regression and Goodness-of-fit

Let's now discuss fitting a model to data. We will do a theoretical treatment here, but the real work will come when we do the projects in the second half of this semester.

Most observations are compared with some sort of model, and often this comparison involves fitting a function to the data (via *regression*). We then need to quantify the agreement between the model and the data (by quantifying a *goodness-of-fit*).

Consider the following situations:

- You want to remove or measure the continuum emission from the spectral line data above.
- You want to fit a Gaussian to the spectral line above to see how wide and tall it is, and at what frequency the peak is ($\mu = \nu_{\text{peak}}$ in this case).
- You need to predict values you haven't measured, based on values you have measured that seem to follow a (yet unknown) functional shape. You could then fit a model to the data and interpolate to find other values. In general, you need as many data points as there are variables.

All of these can be solved with regression fitting, coupled with a goodness-of-fit check.

Important note: for any regression method to work, you need to have more data points than model parameters that you are fitting for ($n_{\text{data}} > n_{\text{var}}$). For instance, in a linear function, $y = mx + b$, you have two variables: m and b. In a quadratic function, $y = ax^2 + bx + c$, you have three variables (a, b, c) even if you set one of those variables is zero. In these cases, the highest order term sets the complexity. **The simplest model that fits the data is always best!** Put simply, *you should generally never try to fit two data points with a quadratic.*[1]

[1] I am still waiting for this advice to show up on a fortune cookie.

3.1 Least Squares Regression

The earliest form of regression was the method of least squares, which was published by Legendre in 1805, and by Gauss in 1809.

Let's say you have x and y values for your data, and you want to fit a function $y_{\mathrm{model}}(x)$ to the data. The least squares method minimizes the sum of the residuals squared (sum squared residuals, SSR), and thus you want to minimize the value of the SSR as given by:

$$\mathrm{SSR} = \sum_{i=0}^{N} R_i^2 \tag{11}$$

and

$$R_i = (y_{i,\mathrm{data}} - y_{i,\mathrm{model}}). \tag{12}$$

You can see that under perfect conditions the summed squared residuals will be zero. It's not too difficult to actually optimize the fitted function (by varying all its input parameters over a range) to minimize the summed squared residuals, but I won't make you do it by hand. We'll use computers for that.

Although the unsquared sum of distances might seem a more appropriate quantity to minimize, use of the absolute value results in discontinuous derivatives which cannot be treated analytically. The square deviations from each point are therefore summed, and the resulting residual is then minimized to find the best fit line. This procedure results in outlying points being given disproportionately large weighting.

In class, we tried a linear least squares fit to the above line spectrum using the software GNUPLOT. If you want to replicate that, the data is under the Fitting and Error Analysis lecture heading here:
`https://sarahspolaor.faculty.wvu.edu/classes/astr469`

The inputs we tried were...
Plot the data, column 1 as x column 2 as y:

```
plot 'astr469line.dat' using 1:2 title "Data"
```

Set a function and some very approximate seed values for its variables:

```
f(x) = m*x + b
m = -1.0
b = 3000
```

Have a look at your initial guess (it will probably not fit the data at all):

```
plot 'astr469line.dat' using 1:2 title "Data", f(x) title "Model"
```

Fit your function to your data, and tell it which variables to try varying over:

```
fit f(x) 'astr469.dat' using 1:2 via m,b
```

At this point, GNUPLOT iterated over some range of values in m and b and found one that it thought minimized the SSR. As it went through the iterations, it printed a few statistical values. At the end it printed the values it thought were right, along with some errors associated with those values. The error percentage tells you approximately how much to believe the fit. Most seed values should work fine, but if you start orders of magnitude away from the real values, you might converge on a function that does not fit well. To check the function, replot it over the data; the m and b values will already be automatically set by GNUPLOT's fitting process:

```
plot 'astr469line.dat' using 1:2 title "Data", f(x) title "Model"
```

Side note: GNUPLOT is a great quick-plotting tool in general. I highly recommend it when you want to have a very quick look at data or do an easy, quick fit (as above).

3.2 Reduced χ^2 Goodness-of-fit Assessment

Regression analysis itself, and the value of the SSR, don't tell us whether the fit was good or bad, nor does it tell us whether a linear model fit better than a quadratic model, or some other model. To assess those things statistically, we need a metric for the goodness-of-fit of the model. The most commonly used in many areas of astrophysics is the reduced chi-squared method.

One way in which a measure of goodness of fit statistic can be constructed, in the case where the standard deviation of the measurement error is known, is to construct a weighted sum of squared errors:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_{i,\text{data}} - y_{i,\text{model}})^2}{\sigma_i^2} \tag{13}$$

where σ again is the measurement error on data point i. This definition is only useful when one has estimates for the error on the measurements, but it leads to a situation where a chi-squared distribution can be used to test goodness of fit, provided that the errors represent a Gaussian distribution.

Have a look at the equation; you can see immediately that as the standard deviation goes up, χ^2 goes down. This is saying that the goodness of fit is less stringent. We also see that as the number of data points goes up, the χ^2 goes up. This seems strange and wrong, so let's have a look at the *reduced* χ^2 metric.

The reduced χ^2 is simply the chi-squared divided by the number of degrees of freedom:

$$\chi^2_{\text{red}} = \frac{\chi^2}{\mathcal{F}} = \frac{1}{\mathcal{F}} \sum_{i=1}^{N} \frac{(y_{i,\text{data}} - y_{i,\text{model}})^2}{\sigma_i^2} \tag{14}$$

Here, $\mathcal{F}$ is the number of degrees of freedom, usually given by $n_{\text{data}} - n_{\text{var}} - 1$. The advantage of the reduced chi-squared is that it already normalizes for the number of data points and model complexity.

As a rule of thumb:

- $\chi^2_{\rm red} \gg 1$ indicates a poor model fit (look at the equation... lots of data with small errors far from model, or very few degrees of freedom, will give you a large $\chi^2_{\rm red}$. This may also happen if the error variance has been underestimated.

- $\chi^2_{\rm red} \simeq 1$ indicates that the extent of the match between observations and estimates is in accord with the error variance.

- $\chi^2_{\rm red} \ll 1$ indicates that the model is 'over-fitting' the data (either the model is improperly fitting noise, or the error variance has been overestimated).

You can use $\chi^2_{\rm red}$ to therefore compare a few different models of different shape. The one with the $\chi^2_{\rm red}$ closest to 1 will be the best model, with the same previous caveat before that if you have two models with similar $\chi^2_{\rm red}$ values, the least complex one is usually the ideal descriptor for your data.