Astronomy 337
Spring 2009
Homework 4: Curve fitting
Due Tuesday, April 21

Least Squares Curve Fitting

Download from the class website's downloads directory the ASCII file *cepheids2.dat*. That file contains a table of data from Sebo et al (2002, ApJS, 142, 71) for 605 Cepheid variable stars in the Large Magellanic Cloud (LMC), with the following columns:

RAh	Hour of Right Ascension (J2000)
RAm	Minute of Right Ascension (J2000)
RAs	Second of Right Ascension (J2000)
DEd	Degree of Declination (J2000)
DEm	Arcminute of Declination (J2000)
DEs	Arcsecond of Declination (J2000)
DelV	Calculated Vmag error
Period	Cepheid period in days
Bmag	The mean phase-weighted B band magnitude
Vmag	The mean phase-weighted V band magnitude
Rmag	The mean phase-weighted R band magnitude
Imag	The mean phase-weighted I band magnitude
Name	Cepheid name

You can use these data to re-derive the famous Cepheid period-luminosity relation (PLR) pioneered by Henrietta Leavitt nearly 100 years ago.

Suppose you don't know the distance to the LMC, but assume that it is sufficiently far from us that all the stars in it are at the same distance. Then the apparent magnitude V should show the same PLR as the absolute magnitude, but offset by the distance modulus $DM = V - M_V$ to the LMC.

Assume that that relation between period P and apparent magnitude V can be given by the form

$$V = c_1 \log P + c_2$$

or

$$y = c_1 x + c_2$$

with $x = \log P$, and $y = V$.

1. Start by making a plot of V (on the x-axis) vs. log(P) (on the y-axis). Be sure to plot V with brightness increasing upwards! Now make an educated guess for what you think would be reasonable values of c_1 and c_2, assuming a linear fit, based on your by-eye estimate of the plot's slope and y-intercept. Write your guess down.

2. Now find c_1 and c_2 using the least squares method, as follows.

   Let the residuals to the fit $r_i = y_i - y'_i$, where y'_i is the *fitted* value of y at the value $x = x_i$, and $i = 1 : n$ with $n =$ the number of data points. Then let

   $$\chi^2 = \sum_i (\frac{r_i}{\sigma_{y,i}})^2 = \sum_i (\frac{y_i - y'_i}{\sigma_{y,i}})^2$$

   as usual. Assume $\sigma_{y,i} = \sigma_V = 0.2$ mag for all values of $y(= V)$.

   Write an IDL program or programs that help you to do the following:

   (a) Calculate χ^2 explicitly (with the formula above) for any combination of c_1 and c_2 input by hand. Use trial and error with at least 10 different attempts for each coefficient to find best-fitting values of c_1 and c_2, i.e. values that minimize χ^2. You will probably need to iterate, i.e. fix c_2 and find a good c_1, then fix c_1 at the new value and vary c_2, then switch again, etc. Report all your values of c_1, c_2, and χ^2. Include a plot of $x = \log P$ and $y = V$ showing your chosen fit.

   Sample IDL conceptual flowchart:

   - read data into variable arrays
   - define new variable for $\sigma_{y,i}$, i.e. σ_V
   - take log of P and store in new variable array
     IDL hint: $IDL > \ logP \ = \ alog10(P)$
   - plot V vs. $\log(P)$ (for V, make sure bright = up!)
   - estimate $c1, c2$ by eye from plot, input in IDL, and calculate V_{fit}
   - calculate residuals (difference between V array and V_{fit} array)
   - calculate χ^2
   - print out c_1, c_2, χ^2
   - change c_1 or c_2, calculate new χ^2. Is it smaller? Larger? Iterate to minimize χ^2.

   (b) Calculate the best fit coefficients using the standard IDL function, "poly_fit", for both a linear fit (polynomial order $n = 1$, straight line, as above) and a quadratic fit ($n = 2$) of the form $y = c_1x^2 + c_2x + c_3$. Report both sets of your values of c_n and χ^2 and plot.

   Sample IDL conceptual flowchart:

   - find number of lines in each array
     IDL hint: $IDL > \ nstars = n\_elements(V)$

- define new array of measurement errors ($= \sigma_V$) for V mag.
  IDL hint: $IDL > \; measure\_errors = REPLICATE(sigmaV, nstars)$
- calculate polynomial coefficients for 1st order (linear) fit using IDL's least-squares fitting function, $poly\_fit$
  IDL hint: $IDL > \; coeffs = poly\_fit(Plog, V, 1, MEASURE\_ERRORS = measure\_errors, chisq = chisquare\_idl)$
- print our coefficients and χ^2

(c) For each of your two fits computed in part 2b, make (and turn in) a plot of the residuals r_i vs. $\log P$. Describe the residuals plot. Does it look like a scatter plot, or are there systematic bumps or wiggles that could be well-fit with a higher-order polynomial? How different are the residuals and the value of χ^2 between the two fits?

3. Application of PLR:

   Now let's calibrate and apply the PLR. Say you find a Cepheid variable star with period $P = 39.81 \pm 0.01$ days and apparent magnitude $V = 7.29 \pm 0.1$ in the nearby globular cluster 47 Tuc, whose distance you know from main sequence fitting to be 4.15 ± 0.07 kpc.

   (a) Calculate the distance modulus to 47 Tuc, and then the absolute magnitude M_V of the Cepheid variable you have discovered.

   (b) Then use the M_V and the period P for this single star combined with the best-fit PLR coefficients you derived in part 2b to calculate the distance modulus to the LMC.

   (c) Finally, calculate the distance to the LMC in kpc.

   (d) Find a recent published value of the distance to the LMC; how does your calculated value compare to it? Are they consistent within the errors you have estimated?

Turn in your programs and all plots with your written answers.