Tutorial 3: Fitting models

January 19, 2024

This tutorial will build on the notes from the lecture Yesterday. You can start by going to the Github repository where the notes are given in `https://github.com/dstndstn/FittingAModel2024`. Find the `FORK` button that will create a *copy* of this repository in your Github account. You can then work on your own copy. After you have `Forked` your copy of the repository, you can `git clone` it to your laptop or Symmetry's Jupyterhub.

In order to find the best line fit for our data with parameters m and b, we need to maximize the likelihood:

$$\mathcal{L}(y|x) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_{y,i}} \exp\left(-\frac{(y_{pred}(x_i, b, m) - y_i)^2}{2\sigma_{y,i}^2}\right)$$

In the lecture, we have computed the optimal parameters `b_opt` and `m_opt` that minimize the negative log-likelihood using the `optimize` function. In this tutorial, the goal is to compute error bars on optimal parameters using the `Jackknife method`.

We will be using the notebook `Notes-from-class.ipynb` from the forked repository as a starting point.

a) Why the log likelihood is more stable to optimize compared to the likelihood?

b) Discuss how the Jackknife method can be used to determine the variance and the covariance estimates for the `b` and `m` parameters. Note that these were called `var_b`, `var_m`, and `cov_bm` in the notebook 'Notes-from-class.ipynb'.

c) Create the covariance matrix:

$$\texttt{cov} = \begin{bmatrix} \texttt{var\_b} & \texttt{cov\_bm} \\ \texttt{cov\_bm} & \texttt{var\_m} \end{bmatrix}$$

Note that `[[1.,1],[1,1]]` creates a `Vector{Vector{Float64}}`, which is not the same thing as a Matrix. In Julia, the syntax for creating a matrix goes as something like `[1, 1; 1, 1]`.

Recall from class that for N data points:

$$\texttt{var\_b} = \frac{(N-1)}{N} \sum_{i=1}^{N} (b_i - \texttt{b\_opt})^2$$

$$\texttt{var\_m} = \frac{(N-1)}{N} \sum_{i=1}^{N} (m_i - \texttt{m\_opt})^2$$

$$\texttt{cov\_bm} = \frac{(N-1)}{N} \sum_{i=1}^{n} (b_i - \texttt{b\_opt})(m_i - \texttt{m\_opt})$$

where b_i and m_i are the jackknife samples.

Note: see Chapter 4 in Efron 1982: "The Jackknife, the Bootstrap and other resampling plans" if you are curious about the proof.

d) Make a plot of the `b, m` plane, showing an ellipse centered on the maximum log likelihood which corresponds to the optimal `b_opt` and `m_opt` values. This should have the same orientation as the contour map we plotted in class.

   Hint: you can generate points on a circle and multiply them with the square root of the `cov` matrix.

e) Draw random samples in `b, m` from a Gaussian distribution with the `cov` covariance matrix (and mean `b_opt, m_opt`). If you draw 100 samples and plot them, you should see that they scatter in a way that traces out the ellipses you have plotted.

f) Take 10 or 20 of those random samples in `b, m` and plot the resulting lines over the data in `x, y` space. If you look at that set of lines, do they all seem like fairly reasonable fits to the data?

g) We have seen that `b` and `m` are anti-correlated, can you describe what that implies about these okay-fit lines?