University of Alberta, Faculty of Science
Department of Physics
PHYS 574: Winter 2021

Homework Assignment 2: Good Plots and Data Analysis Pt. 1

Problems Due: Friday, 17:00 Apr. 2nd, 2021
Please hand in a PDF of your completed problems via eClass.

Class Reading:

1. "Data analysis recipes: Fitting a model to data" (Hogg, Bovy, & Lang, 2010)
   Available free at https://arxiv.org/abs/1008.4686

Problems: The maximum possible grade is 100 points. Individual points for each problem are listed in brackets. I suggest using Python or Matlab for programming (but can only help with the former). Even if you are a Matlab user, you might find that Problems 3 and 4 are best done in Python given the Jupyter notebooks provided on eClass. You should include all code that you use. Given this, I suggest using a notebook interface for this homework.

1. **[40 pts.]** Monte-Carlo Simulation of Polynomial Fit

For this problem you are going to perform Monte-Carlo simulations of a second order polynomial, drawing data assuming a set model and Gaussian errors. The purpose of this problem is to explore the range in the χ^2 values returned by fitting to different simulations of the same model. For this problem, assume $y_{\rm model}(x) = 1 + 0.1234x + 0.5678x^2$ and Gaussian errors in y. Assume that values of x are drawn from a uniform distribution between 0 and 20. Assume that each datum's error σ_i is drawn from a Gaussian with $\mu = 5$ and $\sigma = 0.5$ (you should ensure that no simulated error is less than or equal to 0).

a. Create a function that simulates y values given arrays of x_i, $y_{{\rm model},i}$, and σ_i. While you will assume the error at σ_i does not change between simulations, the simulated y data ($y_{\rm sim}$) should assume that each data point $y_{{\rm sim},i}$ is drawn from a Gaussian with $\mu_i = y_{\rm model}(x_i)$ and $\sigma_i = \sigma_i$.

b. Create a function that returns the χ^2 given arrays of $y_{{\rm data},i}$, $y_{{\rm model},i}$, and σ_i.

c. In this part of the problem, use 8 data points (so that you have 5 degrees of freedom) for each simulation. Using 1000 simulations, calculate the median and 1-σ confidence interval of the 1000 values of χ^2.

d. Repeat 1c using 13 data points (so that you have 10 degrees of freedom) for each simulation.

e. Repeat 1c using 103 data points (so that you have 100 degrees of freedom) for each simulation.

f. Repeat 1c using 1003 data points (so that you have 1000 degrees of freedom) for each simulation.

g. For each of the median values of 1c to 1f, calculate the probability that the χ^2 could be above the measured median χ^2 (Hint: think about using a cumulative chi-square probability distribution function — you don't need to write your own function for this).

2. **[20 pts.]** Fitting a line — 1-D errors and Linear-Least-Squares
For this problem you will use a linear-least-squares (or polynomial) fitting routine to fit $y = mx + b$, assuming Gaussian errors for σ_{y_i} and ignoring σ_{x_i} and ρ_{xy_i}. For each of the two datasets below, make a plot that shows the data, with error-bars, and the best-fit line. Report the (1-σ) errors on the slope and intercept, as derived from the covariance matrix. Calculate the χ^2 of the fit, calculate the probability that the χ^2 could be above the measured χ^2, and discuss if this is an acceptable fit.

- **a.** For this fit, use the data from Table 1 in Hogg, Bovy, & Lang (2010), excluding IDs 2–4.
- **b.** For this fit, use the data from Table 1 in Hogg, Bovy, & Lang (2010).

3. **[20 pts.]** Fitting a line — 1-D errors and MCMC
For this problem you will use an MCMC fitting routine of your choice to fit $y = mx + b$, assuming Gaussian errors for σ_{y_i} and ignoring σ_{x_i} and ρ_{xy_i}. For each of the two datasets below, make a plot that shows the data, with error-bars, and the best-fit line. Report the (1-σ) errors on the slope and intercept, as derived from the appropriate confidence interval. For each dataset's fit, compare the values of the slope and intercept to that derived from Problem 2.

- **a.** For this fit, use the data from Table 1 in Hogg, Bovy, & Lang (2010), excluding IDs 2–4.
- **b.** For this fit, use the data from Table 1 in Hogg, Bovy, & Lang (2010).

4. **[20 pts.]** Fitting a line — 2-D errors and MCMC
For this problem you will use an MCMC fitting routine of your choice to fit $y = mx + b$, assuming Gaussian errors for σ_{x_i} and σ_{y_i} with covariance ρ_{xy_i}. For each of the two fits below, make a plot that shows the data, with error-bars, and the best-fit line. Report the (1-σ) errors on the slope and intercept, as derived from the appropriate confidence interval. For each dataset's fit, compare the values of the slope and intercept to that derived from Problems 2&3.

- **a.** For this fit, use the data from Table 1 in Hogg, Bovy, & Lang (2010), excluding IDs 2–4.
- **b.** For this fit, use the data from Table 1 in Hogg, Bovy, & Lang (2010).