Statistical Methods

Yiannis Tsapras

Exercise 6 for **August 12**, 2024, 18:00

Maximum Likelihood Estimation

6.1 Planning an experiment: where to place measuring points? (partly analytical)

An experiment is to be conducted which boils down to measure the intercept of a line (with the y-axis). This is done by a least-squares fit to measured points $(x_i, y_i), i \in \{0, 1, 2, \ldots\}$. A measuring device measures all y_i-values with equal precision s so that $y_i \sim N(y_{i,\mathrm{true}}, s^2)$. The point x_0 is fixed to $x_0 = 1$ while the others can be placed freely in the experiment (within reason).

a: Each measurement is very expensive. You only obtained limited funding which allows you to measure *two points*. Where do you place point x_1 to obtain the highest possible precision for the line intercept? What is the attainable precision (in units of s)?

Hint: in the lecture the formula for the Fisher information in the case of least-squares was given for independent normally distributed uncertainties. By inverting the Fisher information for the present case one arrives (after some algebra …) at the following covariance matrix for the estimated intercept b and slope m ($\boldsymbol{\theta} \equiv (b, m)^{\mathrm{T}}$)

$$\mathrm{Cov}\left[\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{T}}\right] = \frac{s^2}{(x_1 - x_0)^2} \begin{pmatrix} (x_1^2 + x_0^2) & -(x_1 + x_0) \\ -(x_1 + x_0) & 2 \end{pmatrix}$$

b: You are lucky and found money to measure a third value, x_2. Where do you place the points x_1, x_2 now? What is the best possible precision of b now?

Hint: it would rather tedious to work-out the covariance for 3 points analytically. While you can try to do this (in fact, some UKSta students managed), the idea here is rather to write an R-script: set-up the Fisher information numerically. Then invert it to plot $\mathrm{Var}[b]$ in the (x_1, x_2)-plane, and determine the best positions graphically. Guided by the result in a), make reasonable choices of the region in which you are searching. On web site of the exercises you can find a function `mapper()` that can help to map $\mathrm{Var}[b]$ in the (x_1, x_2)-plane.

Note (once again): all the considerations above can be done *before* actual measurements have been taken. While for the present problem not vital, is is usually advantageous to know the operation characteristics of the measuring device.

6.2 Fitting a straight line with known uncertainties in y

The next exercise is taken from the paper of Hogg et al. “Fitting a straight line to data”. Have perhaps a look at their section 1 for verifying your solutions.

Using the linear algebra as given in the lecture, or the formulation in the article by Hogg et al. (on the web), fit a straight line $y = mx + b$ to the x, y, and σ_y values for data points 5 through 20 in the table in the file *hogg_table1.txt* (on the web site). That is, ignore the first four data points, and also ignore the columns for σ_x and $\mathrm{Cor}[x, y]$.

a: Write an R-program to calculate the best fitting parameters m, and b, their uncertainties, correlations (not covariances!), and χ^2 of the fit!

b: Make a plot showing the points, their uncertainties, and the best-fit line! Hint: error bars can be plotted by using arrows, e.g.:

```
arrows(x, y-sigy, x, y+sigy, length=0.05, angle=90, code=3)
```

c: Do you have an idea how to illustrate the uncertainty of the location of the fitting line?

d: Repeat the previous exercise but for all the data points in the table! Is there anything you do not like about the result?

e: Of course, as statistics oriented programming language R has build in the fitting of linear models with the command `lm`. Compare `lm` results to yours. Hint: `lm` might appear a bit cryptic. Have a look at the example below for inspiration.

```
lm(formula = y ~ x + I(x^2), data = table_hogg, subset = (5:20),
          weights=1.0/table_hogg$sigy^2)
```

Note that the “std. error” in the output of `lm()` are not the $1\,\sigma$ error bars but auxiliary quantities useful in combination with t-statistics.