Statistical Data Analysis
Problem sheet #8

Exercise 1: A set of measurements y_i, $i = 1, \ldots, n$ are treated as independent Gaussian distributed values with given standard deviations σ_i. Each measurement corresponds to a known value of a control variable x_i. Suppose the expectation value of y is given as a function of x by

$$f(x; \alpha, \beta) = \alpha x + \beta x^2 \ ,$$

where α and β are unknown parameters.

1(a) Write down the log-likelihood function for the parameters and show that the maximum-likelihood (ML) estimators are in this case the same as the least-squares (LS) ones, found from the minimum of

$$\chi^2(\alpha, \beta) = \sum_{i=1}^{n} \frac{(y_i - f(x_i; \alpha, \beta))^2}{\sigma_i^2} \ .$$

1(b) Show that the least-squares estimators for α and β can be found from the solution of a system of equations of the form

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} e \\ f \end{pmatrix} \ ,$$

and find the quantities a, b, c, d, e and f as functions of the x_i, y_i and σ_i.

1(c) Make a qualitative sketch of a contour of $\chi^2(\alpha, \beta)$ in the (α, β) plane about the LS estimate, supposing the estimators are negatively correlated. Indicate how this can be used to find the standard deviations of $\hat{\alpha}$ and $\hat{\beta}$.

Would the statistical error in $\hat{\alpha}$ change if β were to be known exactly? Justify your answer.

1(d) Explain how the value of $\chi^2(\alpha, \beta)$ at its minimum may be used to quantify the goodness-of-fit.

Suppose we nave $n = 5$ data points and the value of $\chi^2(\alpha, \beta)$ at its minimum is 12.0. What conclusion can one draw from this about possible errors for the estimates $\hat{\alpha}$ and $\hat{\beta}$?

Exercise 2: For this problem refer to the root macro `SimpleFit.C` and the data file `testData.txt` from the course website. The code is basically C++, but it is executed through the program root rather than being run as an independent program. First run root and at the prompt, type

```
.L simpleFit.C
simpleFit()
```

The first command loads the contents of the file `simpleFit.C` and thus defines the functions contained in it. The second command calls the function `simpleFit()`. This prompts the user for a data file containing columns of numbers representing here the usual ingredients of a least-squares fit, x, y and σ.

The macro contains a fit function, which is currently set up as a polynomial,

$$f(x; \boldsymbol{\theta}) = \sum_{k=0}^{n} \theta_k x^k \ .$$

After reading in the data, the parameters of the polynomial are fitted using the method of least squares. The results are extracted and displayed, including the fitted parameter values, their standard deviations, the minimized χ^2, the corresponding p-value, the covariance matrix $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$ and its inverse.

2(a) Try different orders for the polynomial by changing the variable `npar` (i.e., to obtain an order $n =$ `npar - 1`). What is the smallest order giving a p-value greater than 0.1?

2(b) Consider the cases of order $n = 2$, 3 and 4 (i.e., 3, 4 and 5 parameters). By using the polynomial together with the fitted parameters, find the predicted value of the function at $x = 5$, $x = 6$ and $x = 10$. The program contains an object called `f` of type `TF1*` that you can use to access the fit function. You will need (in a loop over the parameters)

```
f->SetParameter(i, thetaHat[i]);
```

to set the parameter values to the fitted ones and

```
f->Eval([x[i]);
```

to evaluate the function at the point `x[i]`. Please refer to the root documentation for more details.

Using error propagation, find the standard deviations of the differences

$$\begin{aligned} d_1 &= f(x = 6; \hat{\boldsymbol{\theta}}) - f(x = 5; \hat{\boldsymbol{\theta}}) \, , \\ d_2 &= f(x = 10; \hat{\boldsymbol{\theta}}) - f(x = 5; \hat{\boldsymbol{\theta}}) \, . \end{aligned}$$

Comment on how one expects the variance in the difference to behave as the difference between the x values decreases.

2(c) Suppose for the case of order $n = 3$, a certain model predicts the parameter values: $\theta_0 = -0.75$, $\theta_1 = 2.5$, $\theta_2 = -0.5$ and $\theta_3 = 0.026$. By using the inverse of the covariance matrix found in the macro (variable `Vinv`), compute the χ^2 statistic comparing your fitted values with the model predictions,

$$\chi^2 = \sum_{i,j=0}^{n} (\theta_{\mathrm{mod},i} - \hat{\theta}_i)(V^{-1})_{ij}(\theta_{\mathrm{mod},j} - \hat{\theta}_j)$$

Note that here the estimators $\hat{\boldsymbol{\theta}}$ are treated as a set of 4 measured quantities and they are compared to 4 fixed model predictions. Find the corresponding p-value. Is the model in acceptable agreement with the estimated parameter values?

For purposes of comparison, try computing the χ^2 using only the diagonal elements of the covariance matrix (note this is incorrect!). That is, use the formula

$$\chi^2_{\mathrm{bad}} = \sum_{i=0}^{n} \frac{(\theta_{\mathrm{mod},i} - \hat{\theta}_i)^2}{V[\hat{\theta}_i]} \, .$$

Find the corresponding p-value under the (false) assumption that the statistic follows a chi-square pdf and notice that it leads to a completely erroneous conclusion.

Finally compute the χ^2 by comparing the measured (x, y, σ) values to the prediction $f(x; \boldsymbol{\theta}_{\mathrm{mod}})$. This will not give exactly the same value as obtained from the fitted parameters but it should lead to the same conclusion.