5th exercises for SIM'2021

Ex. 1
Continue from exercise 4.1. Download data `linear-model-data-1.dat` from the course webpage. Complete the following task by writing yourself the functions/procedures needed. At this point, do not use existing regression analysis packages that might be available on your computing platform. From the fitted model $\hat{y} = 2.874 - 5.08162x + 0.763641x^2$,
a) Compute standardized residuals (from Eq. (3.27)) and plot against x.
b) Form diagnostic parameter table from model variables as in page 3-13 in the lecture material. What can be deducted regarding the model variables from the parameter table?

Ex. 2
We have two sets of observations, `linear-model-data-2a.dat` and `linear-model-data-2b.dat` (on course webpage). We know that both should have dependency $y = \beta_0 + \beta_1 x^2$, but we do not know if the constant β_0 is the same for both groups. Study this and make decision based on joined linear model for both sets together, with categorical variable included to separate the sets, i.e. $y_i = \beta_0 + \beta_c g_i + \beta_1 x_i^2$, where $g_i = 0/1$ is the categorical variable marking the group, and β_c is its coefficient.

Ex. 3
Load observations from `linear-model-data-3.dat`. First two columns are explanatory variables x_1, x_2, third is dependent variable y. Find suitable model between y and x_i. You can test functions of x_i's and interaction. Test at least two different models and choose best according to model selection criterion of your choice. Check if parameters in the model are significant and that residuals against predicted $\hat{y}$ seem unbiased and homoscedastic. You can use ready-made regression package or write your own procedures.

Ex. 4
Fit nonlinear model to observed degree of linear polarization of comet Hale-Bopp in red wavelength filter. Download data from course webpage (`dataRed.dat`). Use trigonometric model

$$y_i = \mathrm{f}_i(\boldsymbol{\beta}) = \beta_1 \sin(x_i)^{\beta_2} \cos(x_i/2)^{\beta_3} \sin(x_i - \beta_4). \tag{1}$$

a) Program function $S(b_1, b_2, b_3, b_4) = \sum^n (y_i - \mathrm{f}_i(b_1, b_2, b_3, b_4))^2$ in your computing environment. Use minimization procedure to find best estimates b_1, b_2, b_3, b_4. (If you cannot use minimization, test few choices for parameters yourself and choose the best ones).
b) Plot data together with the best fit function.
c) Compute sum of squared residuals, SSE, and residual variance s^2.
d) Compute test statistics $t_i = \frac{b_i}{s\sqrt{m^{ii}}}$ for the parameters. For $m^{ii} = [M^{-1}]_{ii}$ you need the matrix $\mathbf{F}(\boldsymbol{b})$ as in Eq. (4.15). Partial derivatives of f are given in the end of this page.
e) Which parameter is the most uncertain, i.e., has smallest value of test statistics? Test its p-value for the hypothesis H_0 that it could be removed from the model.

$$\frac{\partial \mathrm{f}}{\partial b_1} = -\sin(x)^{b_2}\ \cos\left(\frac{x}{2}\right)^{b_3}\ \sin(b_4 - x) \tag{2}$$

$$\frac{\partial \mathrm{f}}{\partial b_2} = -b_1\ \sin(x)^{b_2}\ \cos\left(\frac{x}{2}\right)^{b_3}\ \sin(b_4 - x)\ \log\left(\sin(x)\right) \tag{3}$$

$$\frac{\partial \mathrm{f}}{\partial b_3} = -b_1\ \sin(x)^{b_2}\ \cos\left(\frac{x}{2}\right)^{b_3}\ \sin(b_4 - x)\ \log\left(\cos\left(\frac{x}{2}\right)\right) \tag{4}$$

$$\frac{\partial \mathrm{f}}{\partial b_4} = -b_1\ \sin(x)^{b_2}\ \cos\left(\frac{x}{2}\right)^{b_3}\ \cos(b_4 - x) \tag{5}$$