# FITTING DATA BY THE METHOD OF LEAST SQUARES

REFERENCE [Taken from [1, p. 25.]]

EXERCISE Given $n$ experimental data points $[x_k, y_k]$ and a *fitting equation*, *aka* model equation, $f[x_k]$ with unknown coefficients $\{\alpha_k\}$ with $1 \leq k \leq n$, the method of least squares consists on minimizing the *square of the root mean square, rms, error*

$$e = \sum_k {\epsilon_k}^2 = \sum_k (f_k - y_k)^2$$

with respect to the coefficients $\{\alpha_k\}$.

For instance, consider that a theory predicts the data in table 1 decreasing with increasing $x$ as a quadratic in $1/x$.

Begin by defining the model to fit the data:

$$f[x] = \alpha_1 + \alpha_2 \frac{1}{x} + \alpha_3 \frac{1}{x^2} \,. \tag{0.1}$$

Then, find the square of the rms error of the model and the data:

$$e = \sum_k \left( \alpha_1 + \alpha_2 \frac{1}{x} + \alpha_3 \frac{1}{x^2} - y_k \right)^2 .$$

Minimize $e$ with respect to the coefficients $\{\alpha_k\}$:

$$\begin{aligned}
\partial_1 e &= 2 \sum_k \left( \alpha_1 + \alpha_2 \frac{1}{x_k} + \alpha_3 \frac{1}{x_k^2} - y_k \right) = 0 \,, \\
\partial_2 e &= 2 \sum_k \frac{1}{x_k} \left( \alpha_1 + \alpha_2 \frac{1}{x_k} + \alpha_3 \frac{1}{x_k^2} - y_k \right) = 0 \,, \\
\partial_3 e &= 2 \sum_k \frac{1}{x_k^2} \left( \alpha_1 + \alpha_2 \frac{1}{x_k} + \alpha_3 \frac{1}{x_k^2} - y_k \right) = 0 \,.
\end{aligned}$$

| $x_k$ | $y_k$ |
|---|---|
| 1.3 | 5.42 |
| 2.2 | 4.28 |
| 3.7 | 3.81 |
| 4.9 | 3.62 |

Table 1 Data with inverse power of $x$ decrease

Distribute the sums in every term, perform algebra and replace the values of table 1 to have

$$\begin{aligned}4.000\alpha_1 + 1.698\alpha_2 + 0.913\alpha_3 &= 17.130\,,\\ 1.698\alpha_1 + 0.913\alpha_2 + 0.577\alpha_3 &= 7.883\,,\\ 0.913\alpha_1 + 0.577\alpha_2 + 0.400\alpha_3 &= 4.52\,.\end{aligned}$$

Solve the system of equations to find $\{\alpha_1 = 3.261, \alpha_2 = 1.480, \alpha_3 = 1.722\}$. The model thus becomes

$$f[x] = 3.261 + 1.480\frac{1}{x} + 1.722\frac{1}{x^2}\,,$$

with a fitting error

$$e = \sum_{k=1}^{4}\left(3.261 + 1.480\frac{1}{x} + 1.722\frac{1}{x^2} - y_k\right)^2 = 0.001\,. \qquad \square$$

DATA ANALYSIS AND REGRESSION ANALYSIS

*Procedure*

Follow the procedure to fit data:

- plot data to see trends, NaNs, missing values, outliers, *&c.*. Consider fitting equations if there is not a subjacent theoretical equation for the data;
- preprocess data: deal with missing values (interpolation), carefully remove NaNs and outliers, consider filtering and detrending data if appropriate;
- summarize data with descriptive statistics;
- normalize or standardize data to find more accurate fitting coefficients;
- perform regression analysis with the norm. or std. data: find the fitting equation, perform error analysis (correlation coefficients, residuals);
- denormalize or destandardize the coefficients and the fitting equation to give them the original dimensions;
- analyze dimensions, specially for the fitting equation coefficients; *i.e.*, give context to abstract fitting equations; non-dimensionalyze the results if possible;
- present the fitting equation and its coefficients with proper dimensions, limits of applicability and errors. Follow Sonin's advice for result presentation, [2, p. 23].

*Example revisited*

Let's revisit the reference exercise, but this time let's apply the procedure to analyze data and then perform regression analysis.

To give a physical context to the data, assume that the impulse is temperature and the response mass density; *i.e.*, $x$ becomes temperature $\theta/°\mathrm{C}$ and $y$ mass density $\rho/\mathrm{g\,cm^{-3}}$.

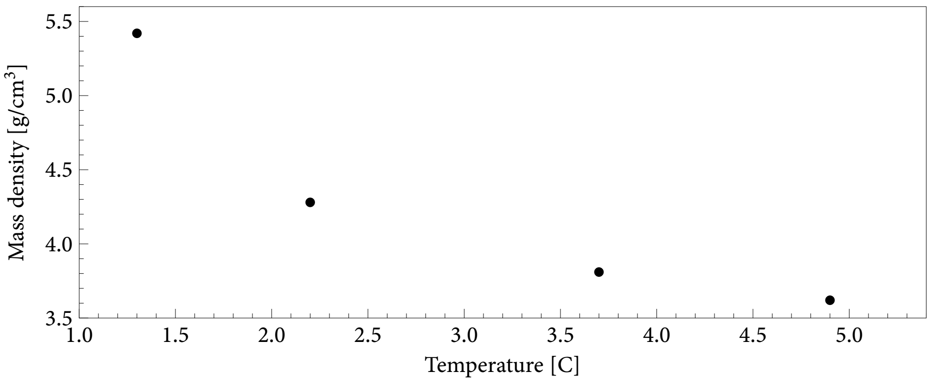

Figure 1 Unprocessed experimental data

*Data preprocessing*

To begin with, plot the unprocessed data to see trends, NaNs and so on.

As seen in fig. 1, there are no NaNs, missing values or outliers. There is no need to filter, nor to detrend data either.

Since there is a « theory » behind data, assume the fitting equation eq. (0.1); *i.e.*,

$$\rho[\theta] = \alpha_1 + \alpha_2 \frac{1}{\theta} + \alpha_3 \frac{1}{\theta^2} \,.$$

*Descriptive statistics*

*Data scaling*

Scale (normalize and non-dimensionalize) data by dividing the values by their maximum; *i.e.*, by dividing temperature values by 4.9 °C and mass density values by 5.42 g/cm$^3$; *i.e.*, transform data as

$$\bar{\theta} = \frac{\theta}{\theta_{\max}} \qquad \text{and} \qquad \bar{\rho} = \frac{\rho}{\rho_{\max}} \,, \tag{0.2}$$

where the bars represent scaled quantities.

The result is presented in ...

*Regression analysis*

Perform the regression analysis as explained in the *Reference exercise* section with the scaled data to find:

$$\bar{\alpha}_1 = 0.601703 \,, \qquad \bar{\alpha}_2 = 0.055723 \qquad \text{and} \qquad \bar{\alpha}_3 = 0.0132309 \,,$$

where the bars represent scaled coefficients.

*Rescaling model equation*

Rescale the model equation by applying the inverse transform of eq. (0.2):

$$\frac{\rho}{\rho_{\max}} = \bar{\alpha}_1 + \bar{\alpha}_2 \frac{\theta_{\max}}{\theta} + \bar{\alpha}_3 \frac{\theta^2_{\max}}{\theta^2} \,.$$

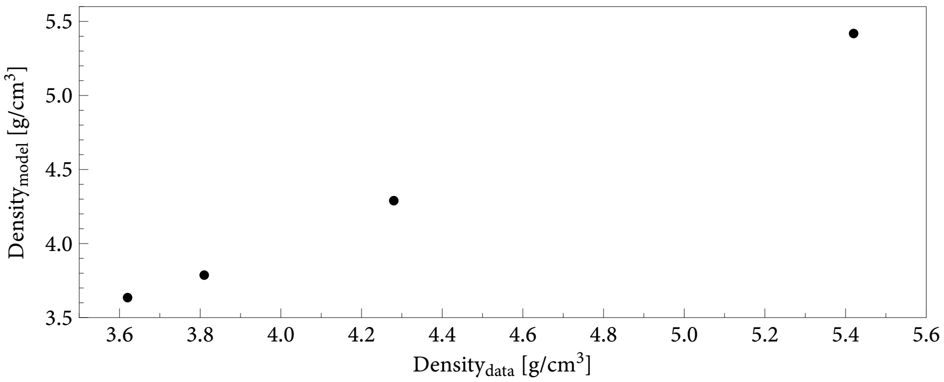

Figure 2 Cross-check between data and model prediction values

By rescaling the model, the coefficients become

$$\alpha_1 = 3.2612306\,, \qquad \alpha_2 = 1.479891434 \qquad \text{and} \qquad \alpha_3 = 1.72179258678\,.$$

Finally, find the rms error as in table ...

$$e = 0.001\,.$$

Figure 2 depicts a cross check between data and model prediction. The straight line indicates agreement between the two; *i.e.*, the model fits experimental data.

*Results*

Regression of the data presented in ... results in the model

$$\rho = \alpha_1 + \frac{\alpha_2}{\theta} + \frac{\alpha_3}{\theta^2}\,, \qquad \begin{cases} \alpha_1/\mathrm{g\,cm^{-3}} & = 3.2612306\,, \\ \alpha_2/\mathrm{g\,°C\,cm^{-3}} & = 1.479891434\,, \\ \alpha_3/\mathrm{g\,°C^2\,cm^{-3}} & = 1.72179258678\,, \end{cases}$$

which is valid for $1.3 \leq \theta/°\mathrm{C} \leq 4.9$ and $3.62 \leq \rho/\mathrm{g\,cm^{-3}} \leq 5.42$, with an rms error of 0.001.

Finally, the model equation can be presented in dimless form:

$$\bar{\rho} = 0.602 + \frac{0.0557}{\bar{\theta}} + \frac{0.0132}{\bar{\theta}^2} \qquad \begin{cases} \bar{\rho} = \rho/5.42\,\mathrm{g\,cm^{-3}}\,, \\ \bar{\theta} = \theta/4.9\,°\mathrm{C}\,, \end{cases}$$

which is valid for $0.265 \leq \bar{\theta} \leq 1.000$ and $0.668 \leq \bar{\rho} \leq 1.0000$, with an rms error of 0.001.

# REFERENCES

[1] Harold Cohen. *Numerical Approximation Methods*. Springer, 2010.

[2] Ain A. Sonin. *The Physical Basis of Dimensional Analysis*. MIT Press, 2001.