ELSEVIER

Chemometrics and Intelligent Laboratory Systems 41 (1998) 249–252

Chemometrics and intelligent laboratory systems

Statistical aspects on fitting the Arrhenius equation

Rolf Sundberg

Mathematical Statistics, Stockholm University, S-10691 Stockholm, Sweden

Received 16 February 1998; accepted 23 April 1998

Abstract

Motivated by a recent mathematical paper, we discuss statistical parameter estimation in the Arrhenius equation, that relates kinetic reaction rates to temperature. In opposition to the paper in question, we argue theoretically for the appropriateness of using ordinary least squares on log-transformed data and supply some empirical support in this direction.

Keywords: Reaction kinetics; Regression analysis; Transformation of data

1. Introduction

Recently in this journal, Klička and Kubáček [1] published a detailed mathematical study of how statistical properties can be affected when the Arrhenius equation is linearized by log-transformation of the kinetic rate determinations. The Arrhenius equation states that the kinetic rate constant k depends on thermodynamic temperature T through the relation

$$k = k(T) = Ae^{-E_a/RT}, \qquad (1)$$

where the pre-exponential (or frequency) factor A and the activation energy E_a are parameters whose values are to be determined (estimated) from data, and R is the gas constant. The assumed setup of Klička and Kubáček is that for a series of temperature values T the corresponding k-values are measured with independent and normally distributed random measurement errors, with the same constant error variance for each of them, i.e., irrespective of T or k (homoscedasticity). When the Arrhenius relation is made linear in $1/T$ by the formation of ln k, the resulting measurements are not exactly normally distributed, their statistical mean values are then no longer exactly the logarithms of Eq. (1), and their variance will depend on T.

There is certainly a pedagogic value in pointing out these facts, as done in Ref. [1], and in particular that ordinary least squares need not be the most efficient method after a linearization. However, there is a danger that the procedure and the adjustments as advocated in Ref. [1] will be interpreted by the reader as valid in some absolute sense, as being 'the procedure to be followed'. As a matter of fact, the results in Ref. [1] are highly dependent on the validity of their statistical model, and the authors do not present any data or any other experience or arguments to support their statistical assumptions. Quite the contrary, I will argue here that the statistical model adopted in Ref. [1] is not likely to be particularly realistic, and I will present theoretical argumentation and supporting empirical data for the much simpler model of an ordinary (constant variance) linear regression for the log-transformed rate constants.

0169-7439/98/$19.00
PII S0169-7439(98)00052-5

2. Statistical model discussion

Like in Ref. [1], we assume that relation (1) would hold for error-free measurements. With measurement noise ϵ added we have measurements (indexed by i)

$$Y_i = k_i + \epsilon_i = A \exp(-E_a/RT_i) + \epsilon_i,$$

where it is assumed in Ref. [1], as if it were obvious, that the noise ϵ_i is $N(0,\sigma)$ distributed, and in particular of constant variance, $\mathrm{Var}(Y_i) = \mathrm{Var}(\epsilon_i) = \sigma^2$. An alternative model could be formed for example by assuming a constant coefficient of variation (relative standard deviation) for Y_i, that is, $\mathrm{Var}(Y_i) = (\sigma k_i)^2$ for some other σ. With this particular model, we would say that we have multiplicative noise in the kinetic rate determinations, and we could write

$$Y_i = k_i(1 + \epsilon_i'), \tag{2}$$

with $\epsilon_i' = \epsilon_i/k_i$ and $\mathrm{Var}(\epsilon_i') = \sigma^2$. Now note that log-transformation of the y-values from Eq. (2) would yield additive noise:

$$Y_i' = \ln k_i + \epsilon_i'',$$

where $\epsilon_i'' = \ln(1 + \epsilon_i') \approx \epsilon_i'$ also has constant variance. Thus, since $\ln k_i$ is linear in $1/T_i$, the multiplicative noise model (2) with normally distributed noise is approximately equivalent to a simple linear regression of $\ln Y$ on $1/T$ with normally distributed noise of constant variance.

Which model is correct? This is of course the wrong question. All models are more or less good approximations to a reality that we cannot completely control. Even the deterministic relation (1) is only an approximation in itself, when derived from kinetic theory (see, e.g., Ref. [2], Chap. 5). We should rather ask which statistical model is likely to fit data better. The answer to this question can only be tentative until we have the real data. However, a theoretical argument for the multiplicative noise model, that is for a constant coefficient of variation, can be constructed and is given in Appendix A below. This derivation is made under some simplifying assumptions but is valid much more generally. Furthermore, the conclusion that the noise contribution is likely to be multiplicative is supported by real kinetics data. Here is such an example.

Example: Figs. 1 and 2 show empirical kinetic rate constants plotted against inverse temperature for a

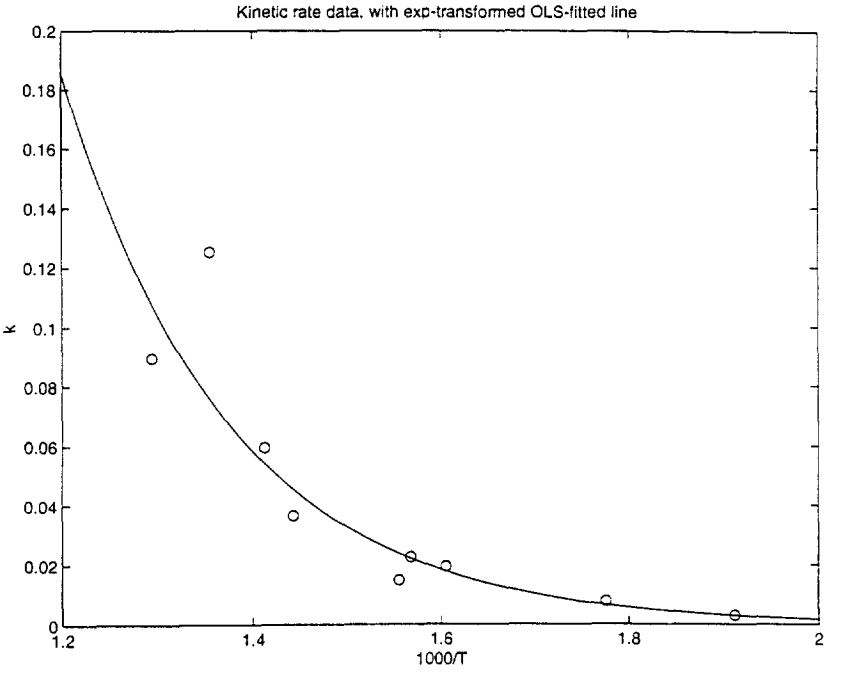

Fig. 1. Kinetic rate data, with exp-transformed OLS-fitted line.

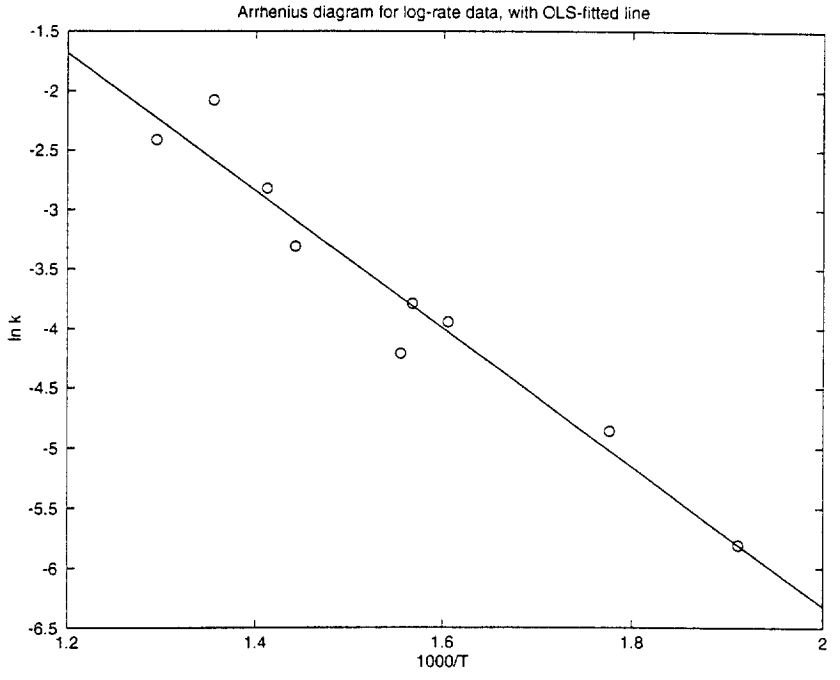

Fig. 2. Arrhenius diagram for log-rate data, with OLS-fitted line.

simple oxidation of methanol. I have taken these data from a course project at the Royal Institute of Technology in Stockholm some years ago. The rate constant of the reaction was determined by the integration method, i.e., by solving the kinetic differential equation and fitting the solution to observed concentrations (see, e.g., Ref. [2], Chap. 2.4). The observed k-values ranged from $3 \cdot 10^{-3}$ to $125 \cdot 10^{-3}$, for a series of temperatures between 523°K and 773°K. Fig. 1 shows the empirical rate constants themselves plotted against $10^3/T$, and Fig. 2 shows the same data after a log-transformation of the rate constants (that is a so-called Arrhenius diagram). Also shown in Fig. 2 is the regression line from an ordinary least squares fit, and in Fig. 1 the exponential curve obtained from this fitted line by the inverse transformation. It is clear from both diagrams that the model fits the data in its mean value structure, and from Fig. 2 that the variance around the line does not show any considerable trend with T. The smaller residuals in the right part of Fig. 2 may be naturally explained by the fact that these few data points are much more influential (have a higher leverage) than individual points in the left part of the diagram. It is also clear from Fig. 1 that the model of Ref. [1], assuming constant variance in the original k-values, is far from realistic.

This is but one example, but it demonstrates at least that we shall not follow categorically the conclusion formulated in Ref. [1], that "if the linear regression of the log-transformed Arrhenius equation is preferred to nonlinear methods, the ordinary least squares estimator should be avoided". On the contrary, the example indicates that we should better work with ordinary least squares on log-transformed data. This is formulated as a conclusion in Section 3.

3. Conclusion

From the theoretical argument given, together with the empirical evidence of the example, we draw the following conclusion. When fitting the Arrhenius equation to data, it is reasonable to try as statistical model a simple linear regression of $\ln Y$ on $1/T$ with additive noise of constant variance. The model fit to data should of course be checked, but only if the assumption of constant variance in this model seems

unrealistic need we try alternative variance structures and fit a weighted regression, or perhaps even a non-linear regression with additive noise for Y itself, that is the very model assumed in Ref. [1].

Appendix A. An argument that determinations of k have a constant coefficient of variation

For simplicity we think of a first-order reaction, in which the concentration c of a reactant consumed in the reaction is followed as it decreases exponentially over time,

$$c(t) = c_0 e^{-kt}. \tag{3}$$

For further simplicity we also assume that the initial concentration c_0 is so well-controlled that we may neglect the error in c_0. This assumption implies that k is the single unknown parameter of relation (3). During the reaction the concentration $c(t)$ is measured at timepoints $t_1, \ldots, t_r$. We are concerned with the precision in k as determined from such data. This will depend on the precision in the measurements of $c(t_i)$. For simplicity we assume that the absolute precision σ_c^2 in determinations of c does not depend on c. This is not at all critical, but should be reasonable at least for analytical error, if concentration is not going too low, whereas it is more difficult to speculate about other sources of error, as for example the difficulty in controlling the reaction times t_i. In order to estimate k we fit a proportional linear regression to the log-transformed relative concentrations $\ln c/c_0$. A constant variance σ_c^2 for c corresponds to a constant coefficient of variation for $\ln c$, according to the laws of propagation of error, that is

$$\mathrm{Var}(\ln c) \approx \sigma_c^2/c^2.$$

This implies that the most efficient estimator of k would be the weighted regression estimator

$$\hat{k} = \sum_1^r w_i t_i \ln c_i / \sum_1^r w_i t_i^2,$$

where c_i stands for the measured $c(t_i)$-values and the weights should be proportional to the inverse variance, i.e. $w_i \propto c(t_i)^2$. It follows that the precision in $\hat{k}$ is

$$\mathrm{Var}(\hat{k}) \approx \frac{\sigma_c^2}{\sum t_i^2 c(t_i)^2} = \frac{\sigma_c^2/c_0^2}{\sum t_i^2 e^{-2kt_i}}.$$

Formula (4) shows that there is relatively little information very early and very late in the reaction, but on a time-scale depending on k. Usually the observation times t_i will also be selected according to what k-value is expected, in order to guarantee that there is a suitable range of c-values in each individual experiment. Suppose we aim at selecting the $t_i = t_i(k)$ such that the set of concentrations c_i will be about the same, irrespective of k, and that we succeed. Then we should have $t_i(k) \propto 1/k$. Insertion in formula (4) now shows that $\mathrm{Var}(\hat{k}) \propto k^2$, i.e., as asserted $\hat{k}$ has a constant coefficient of variation.

References

[1] R. Klička, L. Kubáček, Statistical properties of linearization of the Arrhenius equation via the logarithmic transformation, Chemom. Intell. Lab. Syst. 39 (1997) 69–75.

[2] F. Wilkinson, Chemical Kinetics and Reaction Mechanisms, Van Nostrand-Reinhold, New York, 1980.