# Role of Range and Precision of the Independent Variable in Regression of Data

**Neima Brauner**
School of Engineering, Tel-Aviv University Israel, Tel-Aviv 69978, Israel

**Mordechai Shacham**
Dept. of Chemical Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

*Regression of the experimental data of one independent variable, y vs. a linear combination of functions of an independent variable of the form $y = \sum \beta_j f_j(x)$ is considered. Inherent collinearity among the terms of such functions may prevent obtaining a model of a desired accuracy. Traditional collinearity indicators, condition number of the normal matrix, variance inflation factor, and a new indicator (truncation-error-to-noise ratio) are used to investigate the effects of the range and precision of the independent-variable data on collinearity among functions in a regression model. Statistical confidence intervals are used to demonstrate harmful effects of collinearity. The harmful effects increase by reducing the range of the independent variable data and/or its precision. Using only independent variable data, the new collinearity indicator allows the identification of the point where the number of terms in a particular regression model becomes larger than can be justified on statistical grounds. The use of the new criterion can improve experimental design in order to minimize the harmful effects of collinearity and enable a rapid screen of correlations published in the literature for identifying those that include more parameters than can be justified.*

## Introduction

Precise modeling and regression of experimental data becomes increasingly important as computer-based modeling and design of chemical processes become more widespread. The accuracy of the model of a particular process critically depends on the accuracy of models that are used to predict physical and thermodynamic properties. There are ongoing projects, such as the Design Institute for the Physical Property Data (DIPPR) project of the AIChE (Daubert and Danner, 1987), where experimental data from the literature is being critically reviewed and regressed with selected empirical equations.

In regression of experimental data, the objective is to obtain an equation that predicts values within the experimental error. In order to achieve this objective, equations containing a large number of constants relating to different functions of the same independent variable have been developed. Wagner (1973), for example, listed vapor-pressure models as a function of temperature, using equations containing from four to ten parameters. While increasing the number of parameters will usually reduce the sum of squares of errors (when the error is defined as the difference between the calculated and experimental value of the dependent variable), the use of too many terms related to the same independent variable has several critical drawbacks. For instance, the normal equation, which is used to calculate the parameters, becomes ill-conditioned, whereby adding or removing experimental points from the data set may drastically change the parameter values. There is no statistical justification to the inclusion of some of the terms in the correlation because the respective parameter values are not significantly different from zero. The derivatives of the dependent variable are not represented correctly, and extrapolation outside the region, where the measurements were taken, yields absurd results even for a small range of extrapolation.

Nowadays, statistical analysis (F-test, confidence intervals) are routinely used to test whether there is a justification for including a particular term or a function in a model equation. But, many regression models that were published in the literature can be suspected of having more terms and constants than can probably be justified on statistical grounds.

Correspondence concerning this article should be addressed to N. Brauner.

In this article we focus on regression of physical and thermodynamic properties' data. Typically a model for such data will contain both theory-based and empirical terms. The stepwise regression procedure is used most often for determining the number of terms that should be included in the model. In this procedure, the theory-based terms are included first, and then new terms are added as long as their inclusion can be justified (Wagner, 1973). There are also more recent procedures, which allow optimization of the structure of the regression equation where newly added terms can replace those that were already included in the equation, if such a replacement is statistically justifiable (Setzman and Wagner, 1989). All the commonly used techniques require calculation of the variance, which can be carried out only when information (measured values and error estimates) on the dependent variable is available.

The aim of this article is to investigate the connection between the range and precision of the independent variables and the maximal number of parameters that can be included in a model. The range and precision of the independent variables are limited by physical constraints (a liquid, for example, cannot exist below the temperature of the triple point or above the temperature of the critical point) as well as by limitations of the experimental apparatus. These limitations are known in advance. Therefore, establishing such a connection helps to identify the appropriate structure of a correlation before experiments are actually carried out. This can lead to a better and more representative experimental design.

The discussion is limited to linear regression, in particular to the case where the regression model contains several different functions of the same independent variable. It is shown that inherent collinearity among the terms of such functions may prevent obtaining a model of a desired accuracy. Traditional collinearity indicators, which are condition number of the normal matrix, variance inflation factor, and a new indicator—the truncation-error-to-noise ratio—are used to investigate the effects of the range and precision of the independent-variable data on collinearity among functions in a regression model. The extent of collinearity among these functions is used to establish an upper limit on the number of terms that can be included in a model. Only models whose general structure (including the order in which the different terms are added) is known *a priori* from theoretical analysis or previous experience will be considered. Many such models are in use in engineering. For example, models of physical properties as a function of temperature (Daubert and Danner, 1987), dimensionless heat, mass, and momentum transfer coefficients as a function of other relevant dimensionless groups, and so on.

Two examples (regression of viscosity and vapor-pressure data) are presented to demonstrate the application of the proposed criteria. Most of the calculations related to those examples were carried out using the regression program of the POLYMATH 4.0 package (Shacham and Cutlip, 1996). Calculations related to matrices were done using MATLAB (Math Works, 1992).

## Linear Regression with Models Comprising of Functions of One Independent Variable

Let us assume that there is a set of $N$ data points of a dependent variable (measured variable, such as vapor pressure, viscosity, heat capacity, etc.), $y_i$ vs. an independent variable (controlled variable, such as temperature, concentration, pressure) $x_i$, $i = 1, 2, \ldots, N$. A regression model comprising a linear combination of $n$ different functions of the independent variable is considered. Thus, the regressors are $x_1 = f_1(x)$, $x_2 = f_2(x)$, $\ldots$, $x_n = f_n(x)$.

A linear model fitted to the data is the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \epsilon_i \tag{1}$$

where $\beta_0, \beta_1, \ldots, \beta_n$ are the parameters of the model and $\epsilon_i$ is a measurement error in $y_i$. It is assumed that $\epsilon_i$ is independently and identically distributed (i.i.d.).

The vector of estimated parameters $\hat{\beta}^T = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n)$ is usually calculated using the least-squares error approach, by solving the normal equation:

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}, \tag{2}$$

where $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{A}$ is the normal matrix.

One of the assumptions of the least-squares error approach is that there is no error in the independent variables. This is rarely true, however. Most reports on experimental measurements include estimated error in the independent variable. If such an estimate is not included, a lower limit on the error can be estimated from the number of decimal digits in which the data are reported. (For example, if the temperature is reported with one digit after the decimal point in K, then the error is at least $\pm 0.05$ K). Thus, the value of an independent variable can be represented by

$$x_i = \tilde{x}_i + \delta x_i, \tag{3}$$

where $\tilde{x}_i$ is the expected measured value and $\delta x_i$ is the error (or uncertainty) in its value.

The error in the independent variable is, of course, carried over to its functions. The error in the different functions can be estimated from

$$|\,\delta x_{ji}\,| \le \left|\frac{\partial f_j}{\partial x}\right|_{x = x_i} |\,\delta x_i\,|, \tag{4}$$

where $\delta x_{ji}$ is the estimated error in the $j$th function of $x_i$.

The errors $\delta \boldsymbol{x}$, $\delta \boldsymbol{x}_1$, $\ldots$, $\delta \boldsymbol{x}_n$ result from the limited precision of the measurement and control devices, and they will be denoted "natural errors" or "noise." The natural errors have an important role in determining the number of terms and parameters that can be included in a regression model.

### *Collinearity and its diagnostics*

Collinearity among the different variables can severely limit the accuracy of a regression model. A typical consequence of collinearity is that adding or removing a single data point may cause very large changes in the calculated parameter values. This effect is usually called "ill-conditioning" of the regression equation. Proper diagnostics of collinearity as the cause of ill-conditioning is very important since once detected, appropriate measures to reduce the ill-conditioning can be taken.

A collinearity is said to exist among the columns of $X = [x_1, x_2, \ldots, x_n]$ if for a suitable small predetermined $\eta > 0$ there exist constants $c_1, c_2, \ldots, c_n$, not all of which are zero, such that (Gunst, 1984)

$$c_1 x_1 + c_2 x_2 + \ldots + c_n x_n = \Delta; \quad \text{with } \| \Delta \| < \eta \cdot \| c \|, \quad (5)$$

where $\| \cdot \|$ indicates the norm of a matrix or a vector. This definition cannot be used directly for diagnosing collinearity because it is not known how small $\eta$ should be so that the harmful effects of collinearity will show.

Belsley (1991) lists several criteria and procedures that can be used to detect collinearity. In general those criteria can be divided into two groups: those that are based only on the independent-variables data (e.g., condition number of the normal matrix, variance inflation factor), and those that also require data of the dependent variable (e.g., confidence intervals). Diagnosing collinearity using only the independent variable data has many advantages, since it allows a distinction between collinearity and the measurement error (in the dependent variable) as causes of limited accuracy of a regression model. Such a distinction is required for a better experimental design. Two widely used criteria of the first group—the condition number of the normal matrix and the variance inflation factor (VIF)—will be briefly reviewed. A new criterion (of the same group), "the truncation-error-to-noise ratio" will be introduced.

### *Diagnosis based on the condition number of the normal matrix*

A higher level of collinearity makes the normal matrix more ill-conditioned, whereby the errors in the measured values of $y_i$ or $x_i$ will be amplified when calculating the vector of parameter values, $\hat{\beta}$.

Denoting $A = X^T X$ (normal matrix) and $b = X^T y$, the errors in the calculated parameter values, $\delta\hat{\beta}$, are bounded by (Dahlquist et al., 1974, p. 176):

$$\kappa(A) \frac{\| \delta A \|}{\| A \|} \geq \frac{\| \delta \hat{\beta} \|}{\| \hat{\beta} + \delta \hat{\beta} \|}, \quad (6)$$

where $\kappa(A)$ is the condition number of the normal matrix and $\delta A$ is the matrix of errors in $A$. A similar equation relates the error in $b$, $\delta b$, to the error $\delta\hat{\beta}$:

$$\frac{\| \delta \hat{\beta} \|}{\| \hat{\beta} \|} \leq \kappa(A) \frac{\| \delta b \|}{\| b \|}. \quad (7)$$

The condition number is the ratio of the largest to the smallest eigenvalue of $A$. A strong collinearity results in a higher condition number, thereby amplifying both $\delta b$ and $\delta A$. The former represents measurement errors in both the dependent and independent variables, while $\delta A$ represents the errors in the independent variables.

The condition number of the normal matrix has been extensively used for collinearity diagnosis. However, no critical value for the condition number has been established to indicate harmful collinearity. Therefore, the condition number is not a quantitative measure for collinearity. Furthermore, it has been shown that certain data transformations can reduce the value of the condition number without affecting the level of collinearity (Belsley, 1984).

### *Diagnosis based on the variance inflation factor*

The VIF measures how much collinearity has increased the variance of the model parameters. It can be defined as (e.g., Belsley, 1991, sec. 2.3)

$$\text{var}\left(\hat{\beta}_j\right) = \sigma^2 a_{jj} = \frac{\sigma^2}{\Sigma_i (x_{ji} - \bar{x}_j)^2} \text{VIF}_j \quad (8)$$

$$\text{VIF}_j = \frac{1}{1 - R_j^2}; \quad R_j^2 = \frac{\Sigma_i (\hat{x}_{ji} - \bar{x}_j)^2}{\Sigma_i (x_{ji} - \bar{x}_j)^2}; \quad a_{jj} = (X^T X)^{-1}_{jj}, \quad (9)$$

where $\sigma^2$ is the variance of the dependent variable; $\hat{x}_{ji}$ is the calculated value of $x_{ji}$ when it is regressed on the other independent variables; and $R_j^2$ is the corresponding multiple correlation coefficient.

As can be seen from its definition, the VIF can be calculated from the independent variable data only. A high VIF value indicates a value of $R_j^2$ near to unity and a case of collinearity. As for the condition number, there is no well-defined critical value for VIF, although some authors (see Chatterjee and Price, 1991) suggest 10 as a threshold value to indicate harmful collinearity. It should be noted that for non-centered data, there is a different definition for VIF that yields higher values. However, the trend indicated by both of them is the same. In the examples, Eqs. 8 and 9 are used.

### *Diagnosis based on the truncation-error-to-noise ratio*

Let us consider Eq. 5, which was used to define collinearity. This equation can be divided by, say, $c_j$ to yield

$$c_{1,j} x_1 + c_{2,j} x_2 + \ldots + x_j + \ldots + c_{n,j} x_n = \Delta_j,$$

where $c_{k,j} = c_k / c_j$; $k = 1, 2, \ldots, n$. When the coefficients $c$ are obtained by regressing $x_j$ as a function of the other independent variables, $\Delta_j$ is the residual of this representation and is denoted as the "truncation error" (notation is further clarified in the Appendix). Since the independent variables are subject to an error, the value of $\Delta_j$ is also subject to an error. An upper limit for this error can be estimated from the general error propagation formula:

$$\delta_j = |c_{1,j}| \, |\delta x_1| + |c_{2,j}| \, |\delta x_2| + \cdots + |\delta x_j| + \cdots + |c_{n,j}| \, |\delta x_n|, \quad (10)$$

where $|\delta x_k|$, $k = 1, \ldots, n$ are the natural errors in the independent variables and are obtained by Eq. 4. The values, $\delta_j$, can be used to judge the significance of the truncation errors, $\Delta_j$. When $\delta_j \gg \Delta_j$, the truncation error is within the experimental "noise" level, and thus is practically zero. In this case, harmful collinearity exists. A diagnostic criterion that compares the truncation error to the noise level can be established based on the ratio of the norms of $\Delta_j$ and $\delta_j$:

$$Er_j^I = \frac{\| \Delta_j \|}{\| \delta_j \|}. \quad (11a)$$

The value of $Er_j$ expresses how much of the variation of the truncation error about the zero (mean) value (as represented by the residual plot) can be attributed to experimental noise. When $Er_j \leq 1$, the truncation error is practically zero, so there is a harmful collinearity among the regressors used in the model, whereas a value of $Er_j \gg 1$ indicates that the truncation error is much larger than the noise level, so harmful effects of collinearity are not expected. In between these two extremes, the numerical experimentation should determine the critical values of $Er_j$.

It should be noted that $\boldsymbol{\delta}_j$ as obtained by Eq. 10 represents an upper limit on the noise level, and hence, a lower estimate on $Er$. An estimate for the lower limit for the noise level (which can often be more realistic) is to use $\boldsymbol{\delta}_j = |\,\delta \boldsymbol{x}_j\,|$, whereby the noise level is taken as the natural error in the value of $x_j$. The corresponding $Er$ reads:

$$Er_j^u = \frac{\|\,\boldsymbol{\Delta}_j\,\|}{\|\,\delta \boldsymbol{x}_j\,\|}. \tag{11b}$$

This definition of $Er$ provides an upper estimate on the truncation-error-to-noise ratio.

### *Criterion based on confidence intervals*

A frequently used statistical indicator to determine whether a particular term should be included in a regression model is the confidence interval. This interval is defined by

$$\hat{\beta}_j - t(\nu, \alpha) s\sqrt{a_{jj}} \leq \beta_i < \hat{\beta}_j + t(\nu, \alpha) s\sqrt{a_{jj}}, \tag{12}$$

where $t(\nu, \alpha)$ is the statistical $t$ distribution corresponding to $\nu$ degrees of freedom ($\nu = N-(n+1)$), and a desired confidence level, $\alpha$, and $s$ is the standard error of the estimate.

Clearly, if $\hat{\beta}_j$ is smaller in its absolute value than the term $t(\nu, \alpha) s\sqrt{a_{jj}}$, then the zero value is included inside the confidence interval, $\pm t(\nu, \alpha) s\sqrt{a_{jj}}$. Thus, there is no statistical justification to include the associated term in the regression model. If the independent variables are strongly correlated, most confidence intervals will be larger (in absolute values) than the respective parameter values. Thus, the confidence interval test may be insufficient to pinpoint which of the terms should be removed from the model due to collinearity. In such cases, it may be necessary to repeat the confidence intervals test using uncorrelated orthogonalized data to validate the results.

The confidence interval test relies on more information than is required for the previous tests. In particular, it depends on $s$, which reflects the measurement errors in the dependent variable (and also, indirectly, the error in the independent variable) and the magnitude of the diagonal elements of $\boldsymbol{A}^{-1}$, which strongly relate to $\kappa(\boldsymbol{A})$.

Confidence intervals are useful for evaluating the statistical significance of a regression model. However, the calculation of the confidence intervals requires carrying out first the experiments (for obtaining the values of the independent variables) and then the regression calculations. Also, since the values of the confidence intervals depend on several factors, the effect of collinearity cannot be distinguished from the effect of the experimental errors in the dependent-variable data.

### *Effect of range and precision of the independent variable on collinearity*

Equation 10 provides a basis for assessing the effect of the precision and the range of the independent variable on collinearity.

The effect of precision is included in the denominator of Eq. 10. Higher precision of $x$ results in a smaller value of $|\,\boldsymbol{\delta}_j\,|$, thus increasing $Er_j$.

The norm of the truncation error (numerator of Eq. 11) depends on the range of the independent variable data. Recalling that $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are actually different functions of a single variable $\boldsymbol{x}$, a local approximation around $x = x_0$ (where $x_0$ is a point located within the measurement interval) can be used to represent $x_j = f_j(x)$ in terms of the other $n-1$ functions used in the regression model, of the form

$$x_j = \sum_{\substack{k=1 \\ k \neq j}}^{n} c_{k,j} x_k.$$

The values of $c_{k,j}$ can be obtained, for example, by requiring that $f_j(x)$ and its first $n-2$ derivatives are matched at $x = x_0$ (see a more detailed explanation in the Appendix). Obviously, the truncation error associated with such an approximation increases by increasing the range of $x$, ($\max |\,x_i - x_0\,|$). Thus increasing the range of the independent variable affects an increase of $\|\,\boldsymbol{\Delta}_j\,\|$, and consequently an increase of $Er_j$.

In the remainder of the article two examples are presented to demonstrate the use of the various collinearity indicators for detecting the harmful effects of collinearity in regression, and the influence of the range and precision of the independent variables on collinearity.

## Example 1: Regression of Liquid-Viscosity Data

Table 1 shows the viscosity of the liquid hydrogen bromide as a function of temperature. These data are from Viswanath and Natarajan (1989), who cite Steele et al. (1906) as the source of the data. The reported precision of the temperature data is $\pm 0.1$ K and the reported accuracy of the viscosity data is $\pm 0.5$ %.

The range of the temperature is very narrow (186.8–199.4 K) because of the small gap between the melting point (186.2 K) and normal boiling point (206.4 K). Dauber and Danner (1989) propose to use the three-parameter equation:

$$\ln(\eta) = A' + B'/T + C' \ln T, \tag{13}$$

where $\eta$ is the viscosity in (Pa·s).

It is a common practice to carry out the calculations using normalized, dimensionless variables. Using such variables re-

**Table 1. Viscosity Data for Hydrogen Bromide**

| | Temp. (K) | Vis. (Pa·s)×10³ |
|---|---|---|
| 1 | 186.8 | 0.911 |
| 2 | 188.8 | 0.902 |
| 3 | 190.8 | 0.89 |
| 4 | 193.7 | 0.877 |
| 5 | 197.3 | 0.857 |
| 6 | 199.4 | 0.851 |

From: Steele et al. (1906).

**Table 2. Collinearity Indicators for Two- and Three-Parameter Models***

| No. | No. of Parameters | Condition No. | VIF | $Er^l$ | $Er^u$ |
|---|---|---|---|---|---|
| 1 | 2 | $7.4583\times10^3$ | 1 | 44.58 | 44.58 |
| 2 | 3 | $1.4267\times10^8$ | 12,890 | 0.197 | 0.393 |

*Normalized independent variable data of Example 1.

duces the condition number of the normal matrix, leading to a more accurate calculation of the model parameters. The temperature is normalized by dividing it by the maximal temperature; thus, $\tau = T/199.4$, and the viscosity is normalized by dividing it by the viscosity at the highest temperature; thus, $\mu = \eta/0.851$. Note that mean centering of temperature data is impossible when using Eq. 13, since $\ln(\tau)$ is undefined for negative values of $\tau$ and $1/\tau$ is undefined at $\tau = 0$.

First, we compare the 2-parameter model ($\ln(\mu) = A + B/\tau$) with the 3-parameter model ($\ln(\mu) = A + B/\tau + C \ln \tau$) using the various collinearity indicators that are based only on the independent variable data.

Table 2 shows $\kappa(A)$, VIF, and the two values of $Er{:}Er^l$ are calculated based on $\delta_j$ (Eq. 11a) and $Er^u$, which is based on the natural error error in $\ln \tau$ (Eq. 11b).

All collinearity indicators given in Table 2 indicate that the effects of collinearity are much more severe in the three-parameter than in the 2-parameter equation. The condition number and the VIF are larger by more than four orders of magnitude, for the 3-parameter model. The value of $Er$ is much greater than one (44.58) for the 2-parameter equation and becomes smaller than one for the 3-parameter model. Since in this model $Er < 1$, collinearity prevents using the 3-parameter model for data of such a range and precision.

The graphical interpretation of $Er$ in this example is given in Figure 1. This figure shows the truncation error $\Delta_{ij}$ as calculated from Eq. 5a for the three-parameter model in comparison to the natural error ($\delta[\ln(\tau)] = \pm 1/|\tau| \delta T/199.4|$). It can be seen that the truncation errors are much smaller than the natural errors; thus, all the accurate information in $\ln(\tau)$ is already included in its linear representation by $1/\tau$.

To verify that the 3-parameter equation is indeed inappropriate for representing these data, the parameters of the model representing $\ln(\mu)$ as a function of $1/\tau$ and $\ln(\tau)$ have been calculated. The calculated values are shown in Table 3. The calculations related to Table 3 were carried out using normalized data. The results were verified using orthogonalized data (not shown). For the 3-parameter model, all confidence intervals are much larger than the parameter values, indicating that this model is inappropriate. For the two-parameter model, the confidence intervals are smaller than the parameter values and the variance is smaller than for the 3-parameter model.

Figure 2 shows the relative error of the viscosity calculated using the two-parameter model. The error is randomly distributed and it is less than 0.3%, which is less than the experimental error in the viscosity data (0.5%).

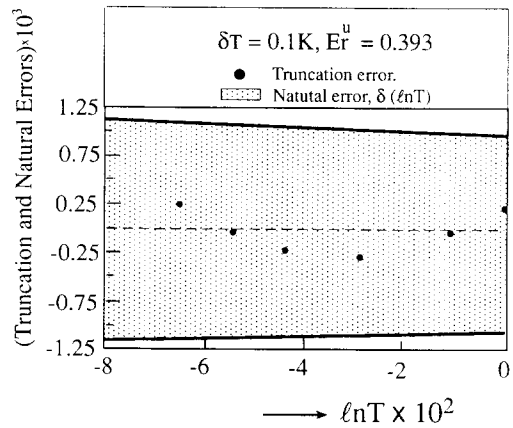


**Figure 1. Truncation error in representing $\ln(\tau)$ as a linear function of $1/\tau$ in comparison to the natural error.**

It can be concluded that because of the narrow range of temperature where experimental data are available, the viscosity of liquid hydrogen bromide can be represented by the 2-parameter model within the accuracy of the data. This model is statistically valid. There is no statistical justification to include one more term (and parameter) in the viscosity equation, since such an addition leads to a less accurate correlation with absurdly large confidence intervals. The collinearity indicator, $Er$, allows prediction of the maximal number of terms that can be included in the particular model using the independent variable data only.

## Example 2. Regression of Vapor-Pressure Data

Table 4 shows vapor pressure data vs. temperature of 1-propanethiol as published by Osborn and Douslin (1966). The temperature, which is considered the independent variable, is reported with five decimal digit accuracy, and Osborn and Douslin (1966) indicate that the precision of the temperature measurement is 0.001 K. The vapor pressure (dependent variable) is also reported in five decimal digits.

The following form of the Riedel equation is considered for correlating these data:

$$\ln(P) = A' + \frac{B'}{T} + C'\ln(T) + D'T^2. \qquad (14)$$

For regression, the variables are brought into dimensionless forms $\tau = (T + 273.15)/375.238$ and $\pi = P/2026.0$. Thus Eq. 14 becomes

$$\ln(\pi) = A + \frac{B}{\tau} + C\ln(\tau) + D\tau^2. \qquad (14a)$$

**Table 3. Regression of $\ln(\mu)$ with the Model $\ln(\mu) = A + B/\tau + C \ln \tau$**

| Correlation | $A$* | $B$ | $C$ | $s^2$ |
|---|---|---|---|---|
| 2 parameter | $-1.04422 \pm 0.08467$ | $1.04266 \pm 0.0818$ | | $2.997\times10^{-6}$ |
| 3 parameter | $-1.58808 \pm 12.26$ | $1.58665 \pm 12.263$ | $0.561924 \pm 12.67$ | $3.969\times10^{-6}$ |

*Confidence interval values rounded to four significant digits.

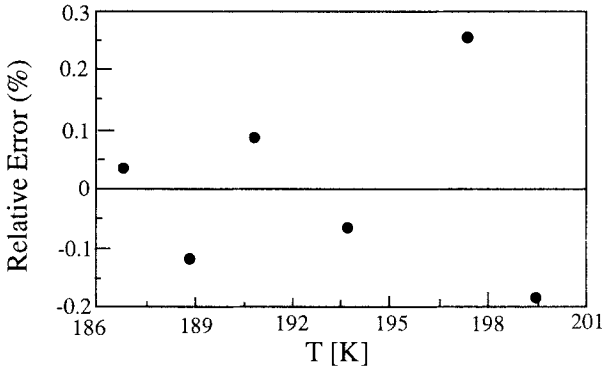


**Figure 2. Relative error in viscosity values calculated using a two-parameter model.**

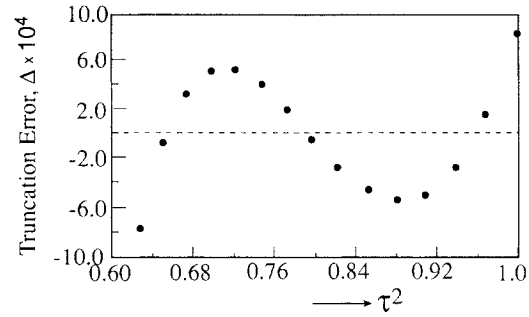


**Figure 3. Residual plot for representation of $\tau^2$ as a linear function of $1/\tau$ and $\ln(\tau)$.**

This set of data and the model (Eq. 14a) are used to investigate the effects of the range and precision on collinearity and on the number of parameters that can be used in the model. Four different data sets are generated. The first data set is the basic data set: temperature range is 24.275°C to 102.088°C with 15 data points, where the estimated precision of the temperature data is 0.001°C. In the second data set, the last eight data points have been removed to yield a temperature range of 24.275°C to 56.605°C. In the third data set, eight data points (No.'s 2,4,6,7,8,10,12, and 14) have been removed without altering the temperature range of the basic data set. In the fourth data set, the basic set was used except that the last two digits of the temperature values were rounded to render an estimated precision of ±0.05°C in the temperature values.

To calculate the various collinearity indicators we first regress $\tau^2$ as function of $1/\tau$ and $\ln(\tau)$. Thus

$$\hat{\tau}^2 = a + \frac{b}{\tau} + c\ln(\tau), \tag{15}$$

where $a$, $b$, and $c$ are parameters of the regression equation. For the basic data set, the following values for the parameters are obtained: $a = -1.83139 \pm 0.1097$, $b = 2.83057 \pm 0.1102$, and $c = 4.77925 \pm 0.1239$ (confidence intervals rounded to four significant digits).

The truncation error in representing $\tau^2$ by Eq. 15, for the basic data set, is shown in the residual plot (Figure 3). It can be seen that the error in this representation is in the interval of $[-10^{-3}, 10^{-3}]$.

Table 5 shows $\kappa(A)$, the VIF, $Er^l$, and $Er^u$ for the 4-parameter equation of the vapor pressure obtained with the various data sets. For the basic data set, both the condition number and VIF are very large compared to accepted standards in the statistical literature, but the $Er$ values ($\gg 1$) imply that collinearity does not prevent the use of the 4-parameter regression equation.

Narrowing the range of the independent variable data (set No. 2) affects an increase of $\kappa(A)$ and VIF by two orders of magnitude. The value of $Er^l$ is very close to one and $Er^u$ is less than 10, indicating that this is a borderline case, where collinearity may prevent the use of a 4-parameter equation. All collinearity measures indicate that reducing the number of data points without changing the range of the data (set No. 3) has an insignificant effect on collinearity.

Reducing the precision of the temperature data (set No. 4) affects very small changes in $\kappa(A)$ and VIF (in comparison to the basic data set), but reduces very significantly both $Er^u$ and $Er^l$. Since $Er^l \sim 0.3$ and $Er^u \sim 2$, it is very probable that collinearity prevents the use of the 4-parameter equation in this case.

Table 6 shows the parameter values (including confidence intervals) and variances of the regression equations obtained for $\ln(\tau)$ with three- and four-parameter versions of Eq. 14a using the various data sets. The calculations related to Table 6 were carried out using normalized data. The results were further verified using orthogonalized data (not shown).

**Table 4. Vapor-Pressure Data for 1-Propanethiol**

| | Temp. (°C) | Pres. (mm Hg) |
|---|---|---|
| 1 | 24.275 | 149.41 |
| 2 | 29.563 | 187.57 |
| 3 | 34.891 | 233.72 |
| 4 | 40.254 | 289.13 |
| 5 | 45.663 | 355.22 |
| 6 | 51.113 | 433.56 |
| 7 | 56.605 | 525.86 |
| 8 | 62.139 | 633.99 |
| 9 | 67.719 | 760.00 |
| 10 | 73.341 | 906.06 |
| 11 | 79.004 | 1,074.6 |
| 12 | 84.710 | 1,268.0 |
| 13 | 90.464 | 1,489.1 |
| 14 | 96.255 | 1,740.8 |
| 15 | 102.088 | 2,026.0 |

From: Osborn and Douslin (1966).

**Table 5. Collinearity Indicators for Various Data Sets for Example 2**

| Data Set* | Condition No. | VIF | $Er^l$ | $Er^u$ |
|---|---|---|---|---|
| No. 1 | $5.0969 \times 10^8$ | $6.5098 \times 10^4$ | 15.66 | 94 |
| No. 2 | $5.4085 \times 10^{10}$ | $1.46275 \times 10^6$ | 1.484 | 8.9 |
| No. 3 | $3.7423 \times 10^8$ | $6.0156 \times 10^4$ | 18.3 | 110 |
| No. 4 | $5.1054 \times 10^8$ | $6.5098 \times 10^4$ | 0.313 | 1.89 |

*Data set description: 1. Temp. range 24.275–102.088°C, 15 data points, full precision; 2. Temp. range 24.275–56.605°C, 7 data points, full precision; 3. Temp. range 24.275–102.088°C, 7 data points, full precision; 4. Temp. range 24.3–102.1°C, 15 data points, last two digits rounded.

Table 6. Regression of $\ln(\pi)$ with 3- and 4-Parameter Versions of Eq. 14a for Various Data Sets

| Data Set No.* | No. of Parameters | $A$ | $B$ | $C$ | $D$ | $s^2$ |
|---|---|---|---|---|---|---|
| 1 | 3 | $13.3515 \pm 0.1085$ | $-13.3523 \pm 0.1090$ | $-3.81954 \pm 0.1226$ | — | $2.476 \times 10^{-7}$ |
| 1 | 4 | $15.1438 \pm 0.1770$ | $-16.1226 \pm 0.2728$ | $-8.49691 \pm 0.4601$ | $0.978684 \pm 0.09619$ | $5.798 \times 10^{-9}$ |
| 2 | 3 | $15.6726 \pm 0.0996$ | $-15.6726 \pm 0.0997$ | $-4.315 \pm 0.105$ | — | $2.629 \times 10^{-9}$ |
| 2 | 4 | $15.1829 \pm 3.270$ | $-14.9813 \pm 4.615$ | $-3.22244 \pm 7.292$ | $-0.201684 \pm 1.346$ | $3.258 \times 10^{-9}$ |
| 3 | 3 | $13.3278 \pm 0.2539$ | $-13.3284 \pm 0.2549$ | $-3.79101 \pm 0.2871$ | — | $4.463 \times 10^{-7}$ |
| 3 | 4 | $15.0737 \pm 0.4833$ | $-16.0201 \pm 0.7430$ | $-8.33002 \pm 1.251$ | $0.946858 \pm 0.2606$ | $1.308 \times 10^{-8}$ |
| 4 | 3 | $13.5712 \pm 0.1730$ | $-13.5726 \pm 0.1737$ | $-4.0633 \pm 0.1954$ | — | $6.292 \times 10^{-7}$ |
| 4 | 4 | $14.3191 \pm 1.861$ | $-14.7285 \pm 2.869$ | $-6.01491 \pm 4.839$ | $0.408323 \pm 1.0117$ | $6.404 \times 10^{-7}$ |

*See description of data sets in Table 5.

For the basic data set, the 4-parameter model is clearly superior to the 3-parameter model. The variance is smaller by almost two orders of magnitude. The confidence intervals for all parameters of both models are small relative to the parameter values, thus all parameters are significantly different from zero.

The adequacy of the 4-parameter model is further emphasized by comparing the residual plots for the 3-parameter model (Figure 4) and the 4-parameter model (Figure 5). In Figure 4, a clear pattern can be recognized and the error range is $[-0.8 \times 10^{-3}, 1.2 \times 10^{-3}]$. The 4-parameter model yields a random error distribution (Figure 5) and the error range is smaller by an order of magnitude $[-1.8 \times 10^{-4}, 1.2 \times 10^{-4}]$.

The results for data set 2 in Table 6 show that reducing the range of the data makes the 4-parameter model inadequate for representing the data. The variance of this model is larger than that of the 3-parameter model and the coefficients of $\ln(\tau)$ and $\tau^2$ are not significantly different from zero. The various collinearity indicators in Table 5 (in particular $Er$) predict these effects of reducing the range of temperature where measurements are available.

Reducing the number of data points without changing the range (data set 3) results in an increased variance (due to less degrees of freedom), but other than that it does not affect significant changes in the 3- or 4-parameter model representation of the data.

Reducing the precision of the independent variable data (set No. 4) renders the 4-parameter model inadequate. The variance of the 4-parameter model is larger than that of the 3-parameter model, and parameter $D$ in the four-parameter model is not significantly different from zero. The collinearity indicator $Er$ predicts the effect of reduced precision, while both $\kappa(A)$ and VIF remained unaffected by this change.

The effects of the range and precision of the independent variable in regressing vapor-pressure data are summarized in Figures 6 and 7. Figure 6 shows the values obtained for $Er$, $\kappa(A)$, VIF, $s$ [the standard deviation of $\ln(\pi)$], and $\Delta D/D$ (the relative confidence interval on parameter $D$ in Eq. 14a vs. the normalized temperature range (the value 1 corresponds to the full range).

Reducing the range of the data has an insignificant effect on the standard deviation, but changes considerably the variables associated with collinearity: ($\kappa(A)$, VIF) and $\Delta D/D$ increase by orders of magnitude when the temperature range is reduced by 60% (from 1 to 0.4). The most meaningful indication of the harmful effects of collinearity is manifested in the value of $\Delta D/D$, which exceeds the value of 1 at the relative temperature range of 0.6 (meaning that for a smaller range of temperature, parameter $D$ is no longer significantly different from zero). These ill effects of collinearity are predicted by $Er$, which approaches a value of 1.

In Figure 7, the same variables are plotted vs. the error in the temperature, $\delta T$ (introduced by rounding out the last digits). As expected, the traditional collinearity indicators $\kappa(A)$ and VIF (which use only information of the independent variables) are not affected significantly by the change of precision. However, the standard deviation and $\Delta D/D$ increase significantly with increasing $\delta T$, which is accompanied by a decrease of $Er$. Both $\Delta D/D$ and $Er^1$ cross the threshold

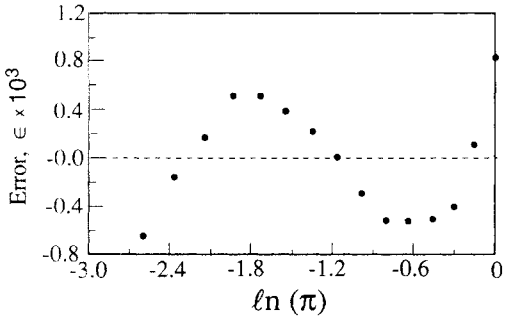


Figure 4. Residual plot of the vapor-pressure data represented by a three-parameter equation.

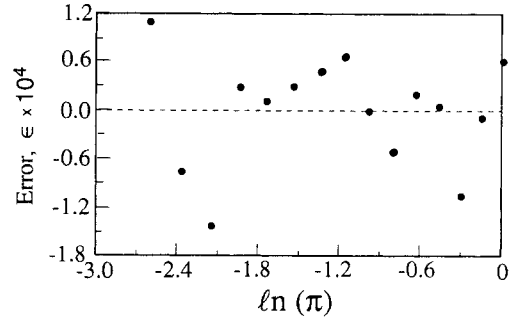


Figure 5. Residual plot of the vapor-pressure data represented by a four-parameter equation.

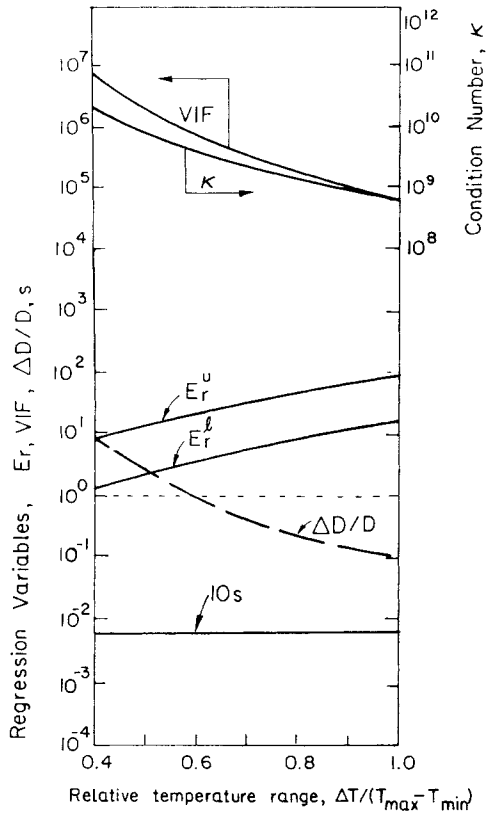


**Figure 6. Variation of regression-related variables with the range of the independent variable (for the vapor-pressure data).**

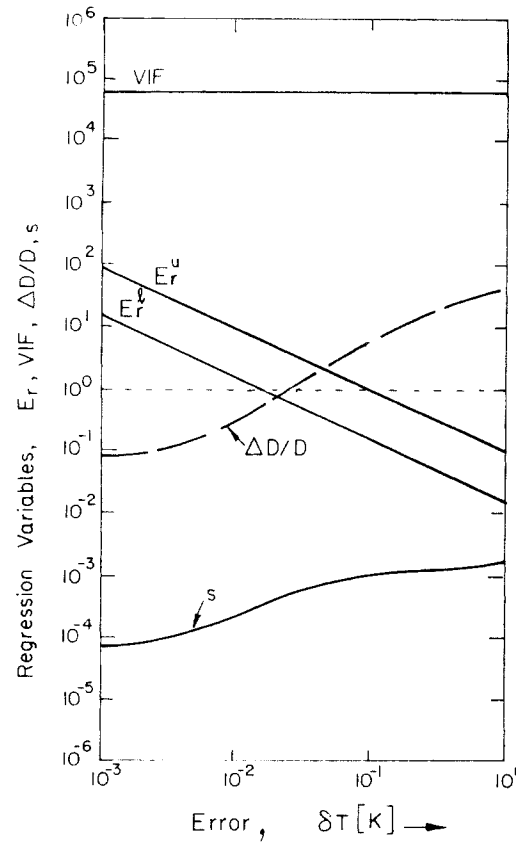


**Figure 7. Variation of regression-related variables with the precision of the independent variable (for the vapor-pressure data).**

value of 1 at about $\delta T = 0.05$ K, indicating that for such (and lower) precision, collinearity prevents the use of the 4-parameter regression equation.

## Conclusions

It has been shown that when correlating a dependent variable with a linear combination of functions of the same independent variable, collinearity among those functions must be considered. The extent of the collinearity depends on the various functional terms included in the correlation and the range of the independent variable. The impact of this collinearity on the quality of the correlation depends on the precision of the independent variable.

All collinearity indicators, which are based on independent variable data only, showed that reducing the range causes an increase of the collinearity among the different terms of a regression model. However, only the new indicator introduced here, the truncation-error-to-noise ratio, provides a critical threshold value that indicates when collinearity reaches a level that requires a reduction of the number of terms in a regression equation. Statistical confidence interval values verified the conclusions reached based on the truncation-error-to-noise ratio criterion.

Traditional collinearity indicators, such as the condition number of the normal matrix and the variance inflation factor, are practically unaffected by the precision of the independent variable. Indeed, it can be shown that there is a relationship between the truncation error and VIF (see, for example, Belsley, 1991, pp. 27–29):

$$\mathrm{VIF}_j = \frac{\Sigma_i (x_{ji} - \bar{x}_j)^2}{\Sigma_i \Delta_{ji}^2},$$

which does not include an implicit reference to the error in $x_{ji}$. This is probably the reason for poor diagnostics provided by the VIF in many cases (see Example 2). Both the truncation-error-to-noise ratio and confidence-interval values indicate that a limited precision of the independent variable may render the use of a regression model with too many parameters statistically unjustifiable and numerically unstable.

The new collinearity indicator *Er* allows setting an upper limit on the number of functions and parameters in a particular regression model even before the experiments are carried out. If the model with the maximal number of parameters cannot represent the dependent-variable value at the desired accuracy, then the range and/or the precision of the independent-variable values must be increased to allow the addition of more terms to the regression equation. In some cases, transformation of the variables can reduce the harmful effects of collinearity (Shacham and Brauner, 1997). Marquardt and Snee (1975) recommend the use of "ridge regression," where a small positive value is added to the diagonal elements of the normal matrix, to reduce the harmful effects of

collinearity. While this method may allow inclusion of more terms in a regression equation, it introduces a certain bias in the estimated parameter values. The method of "partial least squares (PLS)" (Wold et al., 1984) has also been recommended as a means for reducing the ill effects of collinearity, but this method is applicable mainly when there is a large number of independent variables and yields very complicated model equations (a drawback when modeling thermophysical data).

For a particular type of correlation (say Riedel's equation for vapor pressure), an approximate relationship can be established between the maximal number of parameters and the range and precision of the independent variable. This relationship can be used for evaluating the validity of equations published in the literature. If an equation contains more (or significantly less) terms than expected, its statistical validity and accuracy are questionable.

The theoretical analysis and examples provided in this article include only one independent variable. However, the extension to cases of several independent variables is straightforward.

## Literature Cited


Belsley, D. A., "Demeaning Conditioning Diagnostics Through Centering," *Amer. Stat.*, **38**(2), 73 (1984).

Belsley, D. A., "Condition Diagnostics, Collinearity and Weak Data in Regression," Wiley, New York (1991).

Chatterjee, S., and B. Price, *Regression Diagnostics*, Wiley, New York (1991).

Dahlquist, A., A. Björk, and N. Anderson, *Numerical Methods*, Prentice Hall, Englewood Cliffs, NJ (1974).

Daubert, T. E., and R. P. Danner, *Physical and Thermodynamic Properties of Pure Chemicals: Data Compilation*, Hemisphere, New York (1989).

Gunst, R. F., "Toward a Balanced Assessment of Collinearity Diagnostics," *Amer. Stat.*, **38**(2), 79 (1984).

Marquardt, D. W., and D. R. Snee, "Ridge Regression in Practice," *Amer. Stat.*, **29**(2), 3 (1975).

Math Works, Inc., The Student Edition of MATLAB, Prentice Hall, Englewood Cliffs, NJ (1992).

Osborn, A. G., and D. R. Douslin, "Vapor Pressure Relations of 36 Sulfur Compounds Present in Petroleum," *J. Chem. Eng. Data*, **11**, 502 (1966).

Setzmann, U., and W. Wagner, "A New Method for Optimizing the Structure of Thermodynamic Correlation Equations," *Int. J. Thermophysics.*, **10**(6), 1103 (1989).

Shacham, M., and N. Brauner, "Minimizing the Effects of Collinearity in Polynomial Regression," *Ind. Eng. Chem. Res.*, **36**(10), 4405 (1997).

Shacham, M., and M. B. Cutlip, *POLYMATH 4.0 User's Manual*, CACHE Corporation, Austin, TX (1996).

Steele, B. D., D. McIntosh, and E. H. Archibald, "The Vapour Pressures, Densities, Surface Energies and Viscosities of Pure Solvents," *Philos. Trans. Roy. Soc. (London)*, **A205**, 99 (1906).

Viswanath, D. S., and A. Natarajan, *Data Book on the Viscosity of Liquids*, Hemisphere Publishing, New York (1989).

Wagner, W., "New Vapor Pressure Measurements for Argon and Nitrogen and a New Method for Establishing Rational Vapor Pressure Equations," *Cryogenics*, **13**, 470 (1973).

Wold, S., A. Ruhe, H. Wold, and W. J. Dunn, III, "The Collinearity Problem in Linear Regression, The Partial Least Squares (PLS) Approach to Generalized Inverse," *SIAM J. Sci. Stat. Comput.*, **5**(3), 735 (1984).


## Appendix: Example for Derivation of an Expression for the Truncation Error as a Function of $(x - x_0)$

Considering, for example, a model of the form:

$$y = \beta_0 + \frac{\beta_1}{x} + \beta_2 \ln x + \beta_3 x^2. \tag{A1}$$

At a vicinity of a point $x_0$ within the measurement interval, $x^2$ can be expressed as a linear function of $1/x$ and $\ln(x)$, $x^2 \simeq c_1 + c_2/x + c_3 \ln(x)$, which can be written in the following form:

$$x^2 \simeq c'_1 x_0^2 + c'_2 x_0^3/x + c'_3 x_0^2 \ln(x/x_0), \tag{A2}$$

where $c'_3 = c_3/x_0^2$, $c'_2 = c_2/x_0^3$, and $c'_1 = (c_1 + c_3 \ln x_0)/x_0^2$. Requiring that the function, its first, and second derivatives on both sides of Eq. A2 are equal at $x = x_0$, results in three equations with three unknowns ($c'_1$, $c'_2$, $c'_3$):

$$\begin{aligned} x_0^2 &= c'_1 x_0^2 + c'_2 x_0^2 \\ 2x_0 &= -c'_2 x_0 + c'_3 x_0 \\ 2 &= 2c'_2 - c'_3, \end{aligned} \tag{A3}$$

which yields $c'_1 = -3$, $c'_2 = 4$, $c'_3 = 6$. Any point within the measurement interval can be used for $x_0$. Then truncation error at point $x_i$ is defined by

$$\Delta_i = c'_1 x_0^2 + c'_2 x_0^3/x_i + c'_2 x_0^2 \ln(x_i/x_0) - x_i^2. \tag{A4}$$

Such an analysis can be used to estimate the truncation error (and $Er$) without carrying out any regression. However, when the coefficients of the linear representation of $x^2$ as a function of $1/x$ and $\ln(x)$ are obtained by regression, the resulting coefficient values satisfy Eqs. A3 with an "optimal" value of $x_0$ in the least-squares sense (in Example 2, the value of $x_0 \equiv T_0 = 334.24$ K is obtained).

In the Taylor series expansion of $\Delta$ (Eq. A4) around $x = x_0$, the terms containing the function, its first, and second derivatives vanish, and the error term is proportional to $|d^3\Delta/dx^3|_{x=x_0} = (x - x_0)^3/3!$. Thus, the truncation error in expressing $x^2$ as a linear function of $1/x$ and $\ln(x)$ is proportional to $(x - x_0)^3$, which is bounded by $(x_{max} - x_{min})^3$. Increasing the range of the independent variable increases the truncation error.

*Manuscript received Dec. 26, 1996, and revision received Nov. 21, 1997.*