Scientific Research Publishing

Applied Mathematics, 2023, 14, 386-410
https://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

Multiple Regression and Big Data Analysis for Predictive Emission Monitoring Systems

Zinovi Krougly[1], Vladimir Krougly[2], Serge Bays[2]

[1]Department of Mathematics, Western University, London, Canada
[2]Limesoft Inc., London, Canada
Email: zkrougly@uwo.ca, info@limesoft.ca

How to cite this paper: Krougly, Z., Krougly, V. and Bays, S. (2023) Multiple Regression and Big Data Analysis for Predictive Emission Monitoring Systems. *Applied Mathematics*, **14**, 386-410.
https://doi.org/10.4236/am.2023.145023

Received: April 20, 2023
Accepted: May 28, 2023
Published: May 31, 2023

Abstract

Predictive Emission Monitoring Systems (PEMS) offer a cost-effective and environmentally friendly alternative to Continuous Emission Monitoring Systems (CEMS) for monitoring pollution from industrial sources. Multiple regression is one of the fundamental statistical techniques to describe the relationship between dependent and independent variables. This model can be effectively used to develop a PEMS, to estimate the amount of pollution emitted by industrial sources, where the fuel composition and other process-related parameters are available. It often makes them sufficient to predict the emission discharge with acceptable accuracy. In cases where PEMS are accepted as an alternative method to CEMS, which use gas analyzers, they can provide cost savings and substantial benefits for ongoing system support and maintenance. The described mathematical concept is based on the matrix algebra representation in multiple regression involving multiple precision arithmetic techniques. Challenging numerical examples for statistical big data analysis, are investigated. Numerical examples illustrate computational accuracy and efficiency of statistical analysis due to increasing the precision level. The programming language C++ is used for mathematical model implementation. The data for research and development, including the dependent fuel and independent NOx emissions data, were obtained from CEMS software installed on a petrochemical plant.

Keywords

Matrix Algebra in Multiple Linear Regression, Numerical Integration, High Precision Computation, Applications in Predictive Emission Monitoring Systems

1. Introduction

Continues Emissions Monitoring Systems (CEMS) were traditionally used to monitor emissions from stationary sources. CEMS requires a significant capital

investment and is very expensive to operate and maintain due to the high initial cost of the hardware and significant amount of labor required to maintain the equipment. Predictive Emission Monitoring Systems (PEMS) can be a cost-effective alternative, they can continuously monitor dependent parameters and estimate emissions with a close to real accuracy, greatly reducing the initial commissioning costs and ongoing maintenance.

PEMS development is based on proven scientific methods to perform statistical data analysis of the significant parameters used as inputs to a multiple regression model. Several commercial software packages, such as Minitab, MATLAB, and R, are available and provide methods and techniques for multiple linear regression models [1] [2] [3] [4]. We will focus on using the available techniques in combination with C++ programming language to develop the multiple linear and polynomial regression software that can work as a powerful PEMS model. Other used methods are: big data analysis, vectors, matrices and linear algebra, high precision numerical calculations [5] [6], as well as custom-developed classes and objects implementing high-performance scientific computing. Our PEMS model also uses numerical integration technique, including integral representation for the cumulative distribution function (CDF) of t, F and Normal distributions. These methods require big statistical data analysis for testing the significance of regression coefficients and calculation of significance levels for F statistic.

The Data Acquisition System software was provided by Limesoft Inc. [7] to input and test the PEMS mathematical model, present results and compare them in real-time and on historical trend chart with actual NOx values obtained from the CEMS gas analyzer.

The PEMS model was verified against performance specification [8] requirements. It also has all the necessary test procedures to perform PEMS model evaluation, assessment and verification, to prove that PEMS software results have required accuracy in order to be accepted by the Environmental Agency.

The remainder of the paper is organized as follows. In Section 2, a brief description of the underlying theory is given, to introduce matrix algebra formulation to multiple linear regression. Section 3 applies predictive emission monitoring with matrix notation. Sections 4 and 5 describe statistical techniques and numerical integration used in PEMS model development. In section 6, multiple polynomial regression model was tested with big data analysis. In Section 7, PEMS mathematical model was tested by actual CEMS gas analyzer. The procedures described in Section 8, provide a framework for testing PEMS model during normal engine operations. In Section 9 and Appendix, we discuss the role of high precision software in statistical computations. In Sections 10, 11 and 12, we consider comprehensive methodology for developing PEMS, limitations of PEMS model and environmental benefits of PEMS over CEMS.

2. Matrix Algebra Formulation to Multiple Linear Regression

Linear regression model specifies that the regression function is a linear function

of the regressor (independent) variables. More realistic applications require more general regression models because the regression function is not linear or because there are many regressor variables. Including polynomial terms in the regression model is a common way to achieve a better approximation to the true regression function. This leads to the term polynomial regression. Incorporating several regressor variables, with or without additional polynomial terms, is considered a multiple regression model. The most effective way to express the mathematical operations is fitting model in matrix notation.

Suppose that there are k regressor variables and n observations, $\left(x_{i1}, x_{i2}, \cdots, x_{ik}, y_i\right), i = 1, 2, \cdots, n$. The regression function can be modeled as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \cdots, n \tag{1}$$

This model is a system of n equations that can be written in matrix form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{3}$$

The vector $\boldsymbol{y}$ is the $n \times 1$ vector of the observations, the design matrix $\boldsymbol{X}$ is the $n \times (k+1)$ matrix containing the values of the input variables, the parameter $\boldsymbol{\beta}$ is $(k+1) \times 1$ vector of the regression coefficients and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of random errors. The vector of the least squares estimator $\hat{\boldsymbol{\beta}}$ is the solution of the normal equations, which can be written in matrix notation as

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \tag{4}$$

Multiplying both sides of the normal equations (4) by $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$, we obtain the least square estimate of $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{5}$$

where $\boldsymbol{X}'$ is the transpose of the matrix $\boldsymbol{X}$.

If the variables $x_1, x_2, \cdots, x_k$ are linearly independent, the matrix $\boldsymbol{X}'\boldsymbol{X}$ is always nonsingular [4], so the methods for inverting matrices can be used to find $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. In practice, we perform matrix calculation using the systematic definition of the vector and matrix classes in C++ [9].

The fitted value of the response variable at the data point $\left(x_{i1}, x_{i2}, \cdots, x_{ik}\right)$ is:

$$\hat{y}_i = \hat{b}_0 + \sum_{j=1}^{k} \hat{b}_j x_{ij}, \quad i = 1, 2, \cdots, n \tag{6}$$

The matrix form of the fitted values and residuals is:

$$\hat{y} = X\hat{\beta} \quad \text{and} \quad e = y - \hat{y} \tag{7}$$

The test for significance in regression involves the null hypothesis H_0 to determine whether a linear relationship exists between the response variable y and

any of k regressors, $x_1, x_2, \cdots, x_k$. The test for $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is based on a statistic that has a particular F distribution:

$$F_0 = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} = \frac{\text{MSR}}{\text{MSE}} \tag{8}$$

Rejection region and p value for a level α test are:

$$F_0 > f_{\alpha,k,n-k-1}, \quad p \text{ value} = 1 - f_{\alpha,k,n-k-1}(F_0) \tag{9}$$

We should reject H_0 if the computed value of F statistic in Equation (8) is greater than $f_{\alpha,k,n-k-1}$ distribution value with parameters k and $n-k-1$. Test for significance of regression involving the total sum of squares SST which is partitioned into a sum of squares, due to regression SSR, and a sum of squares due to errors SSE:

$$\text{SST} = \text{SSR} + \text{SSE}$$

The computation formula for the regression sum of squares is:

$$\text{SSR} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{10}$$

The sum of squares for error is defined by

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{y} \tag{11}$$

The strength of a regression model is measured using the coefficient of multiple determination R^2, which takes the values between 0 and 1:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{12}$$

Since R^2 always increases when regressors are added to the model, it is better to use adjusted R^2_{adj} statistic involving degrees of freedom:

$$R^2_{adj} = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} \tag{13}$$

3. Predictive Emission Monitoring with Matrix Notation

Table 1 lists eight data set examples used as tests of the numerical calculations to various types of PEMS mathematical models.

The data that appeared in Table 2 correspond to Example 1 in Table 1. The format includes the observation interval, number of observations, observation number, observation vector, and design matrix, containing the input variables of data with n = 22,200 and k = 15. Table 3 shows the results of fuel combustion analysis.

C++ software multiple regression outputs are given in Table 4 for estimated $\hat{\boldsymbol{\beta}}$ and calculated t values.

Consider matrix algebra calculations for the data set presented in Table 5.

Fitted values are calculated by Equation (14)

Table 1. Data set examples used as tests of the numerical calculations.

Example	Gas emission	n	k	Multiple regression model
1	NO	22,200	15	linear
2	NO	22,200	30	polynomial
3	NO	5485	15	linear
4	NO	5485	30	polynomial
5	NO	96	15	linear
6	NO	96	30	polynomial
7	NO	8444	15	linear
8	NO_2	8444	15	linear

Table 2. Example 1. Observation interval, number of observations, observation number, observation vector and design matrix, containing the input variables of data with $n = 22{,}200$ and $k = 15$.

Int	#	Obs	y_i	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
0 - 10	2661	5000	0.00	1	23.9	0.00	2.7	88.2	0.04	0.05	1.8	0.56	0.15	0.09	0.03	9.0	23.9	23.9	101.4
		4680	4.8	1	15.2	0.00	3297.7	88.2	0.04	0.05	1.8	0.58	0.16	0.09	0.04	9.0	15.5	15.0	127.6
10 - 20	45	6384	11.3	1	10.9	0.00	8647.8	91.5	0.02	0.02	1.1	0.18	0.05	0.03	0.67	6.4	11.5	10.2	152.3
		6385	14.5	1	10.8	0.00	8655.5	91.5	0.02	0.02	1.1	0.18	0.05	0.03	0.67	6.4	11.5	10.2	152.0
20 - 30	1691	14,338	26.7	1	4.0	28,786.5	24,585.5	87.5	0.04	0.05	1.7	0.49	0.14	0.09	0.01	9.9	4.2	3.7	175.5
		9086	29.8	1	2.5	24,504.2	22,600.1	87.6	0.04	0.05	1.7	0.48	0.14	0.08	0.02	9.9	3.0	2.1	169.7
30 - 40	5363	9722	34.7	1	3.9	9372.3	27,207.8	87.5	0.04	0.05	1.8	0.48	0.14	0.09	0.01	9.9	4.2	3.6	160.4
		10,738	38.9	1	2.7	24,836.7	30,197.3	87.6	0.04	0.05	1.8	0.47	0.14	0.09	0.01	9.9	2.7	2.6	172.3
40 - 50	7284	10,239	41.1	1	1.7	24,019.7	30,697.8	87.5	0.04	0.05	1.8	0.47	0.14	0.09	0.01	10.0	1.7	1.7	169.6
		16,792	49.8	1	2.3	25,987.4	28,039.0	88.1	0.04	0.05	1.8	0.64	0.18	0.10	0.04	9.0	2.3	2.3	163.3
50 - 60	3627	17,209	51.3	1	2.3	25,935.4	28,675.3	88.05	0.03	0.05	1.8	0.56	0.15	0.09	0.03	9.3	2.2	2.4	165.2
		381	58.1	1	2.3	26,403.1	30,325.9	89.7	0.02	0.03	1.2	0.25	0.07	0.05	0.20	8.5	3.1	1.5	169.3
60 - 72	1529	585	60.6	1	2.2	24,510.1	30,202.5	90.5	0.02	0.03	1.3	0.35	0.10	0.06	0.13	7.5	1.7	2.7	170.1
		846	61.5	1	1.8	24,547.4	30,373.9	90.7	0.02	0.03	1.3	0.32	0.09	0.05	0.23	7.2	1.6	1.9	169.7

Table 3. Results of fuel combustion analysis.

y_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7
NO	O_2	PG Fuel	FNG Fuel	CH_4 in FNG	nC_5H_{12} in FNG	iC_5H_{12} in FNG	C_2H_6 in FNG
(ppm)	(%)	(Nm^3/hr)	(Nm^3/hr)	(mol% (wet))	(mol% (wet))	(mol% (wet))	(mol% (wet))
x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
C_3H_8 in FNG	nC_4H_{10} in FNG	iC_4H_{10} in FNG	CO_2 in FNG	N_2 in FNG	O_2 in Flue Gas-1	O_2 in Flue Gas-2	Flue Gas Temperature
(mol% (wet))	(mol% (wet))	(mol% (wet))	(mol% (wet))	(mol% (wet))	(% wet)	(% wet)	(% wet)

Table 4. Estimated $\hat{\boldsymbol{\beta}}$ and calculated t value. Values $3.929e-12 \equiv 3.929\times10^{-12}$.

$\hat{\beta}_i$	Estimate	d.e. ($\hat{\beta}_i$)	s.e. ($\hat{\beta}_i$)	t value
$\hat{\beta}_0$	38.4872	5.03727	12.9562	2.97056
$\hat{\beta}_1$	50.2346	2734.87	301.8899	0.1664
$\hat{\beta}_2$	−0.000147609	3.929e−12	−12.9008	−12.901
$\hat{\beta}_3$	0.00126975	1.259e−11	0.0000	62.001
$\hat{\beta}_4$	0.362444	0.00051525	0.0000	2.766
$\hat{\beta}_5$	2374.26	287.012	0.1310	24.277
$\hat{\beta}_6$	−3682.95	249.038	91.0988	−40.428
$\hat{\beta}_7$	−31.9369	0.0123964	0.6427	−49.69
$\hat{\beta}_8$	29.1276	0.524861	4.1822	6.9648
$\hat{\beta}_9$	−1072.14	34.185	33.7519	−31.765
$\hat{\beta}_{10}$	3275.63	95.6281	56.4512	58.026
$\hat{\beta}_{11}$	1.95295	0.0064368	0.4631	4.2167
$\hat{\beta}_{12}$	−5.67656	0.00264964	0.2971	−19.103
$\hat{\beta}_{13}$	−25.4091	683.714	150.9446	−0.1683
$\hat{\beta}_{14}$	−25.3162	683.724	150.9457	−0.16772
$\hat{\beta}_{15}$	−0.0163964	3.039e−07	0.0032	−5.15233

Table 5. Given observation vector y and design matrix X, containing the input variables.

y_i	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
55.76	1	2.29	26,428.6	30,834.0	89.2	0.0255	0.0355	1.391	0.2937	0.0894	0.0612	0.1022	8.783	2.894	1.687	171.7130

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_{15} x_{i15} \tag{14}$$

Multiple linear regression model with $k = 15$ regression coefficients predicts the fitted value

$$\begin{aligned}\hat{y}_i = {} & 38.49 + 50.23x_1 - 0.0001476x_2 + 0.0012698x_3 + 0.362444x_4 \\ & + 2374.26x_5 - 3682.95x_6 - 31.94x_7 + 29.13x_8 - 1072.14x_9 \\ & + 3275.63x_{10} + 1.95x_{11} - 5.68x_{12} - 25.41x_{13} - 25.32x_{14} - 0.0163964x_{15}\end{aligned}$$

This regression model can be used to obtain the fitted values by substituting each observation into Equation (14). Thus, fitted value $\hat{y}_i$ for the data in **Table 5** is:

$$\begin{aligned}\hat{y}_i = {} & 38.49 + 50.23(2.29092) - 0.0001476(26428.6) + 0.0012698(30834) \\ & - 0.0001476(89.2179) + 0.0012698(0.025518) + 0.362444(0.035515)\end{aligned}$$

$$+2374.26(1.39106)-3682.95(0.293697)-31.94(0.089405)$$
$$+29.13(0.061245)-1072.14(0.102182)+3275.63(8.78347)$$
$$+1.95(2.89435)-5.68(1.68749)-0.0163964(171.713)$$
$$=51.09.$$

Next step is calculations of estimated regression coefficients $\hat{\boldsymbol{\beta}}$, t value, p value and significance level. Regression sum of squares for errors is calculated as

$$\mathrm{SSE}=e'e=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$$

and

$$\hat{\sigma}^2=\mathrm{MSE}=\frac{\mathrm{SSE}}{n-k-1}$$

$$\mathrm{SSE}=e'e=4.67^2+\cdots+\left(-11.50\right)^2=739266$$

Estimation of the error of variance is:

$$\hat{\sigma}^2=\mathrm{MSE}=\frac{\mathrm{SSE}}{n-k-1}=\frac{739266}{22200-15-1}=33.3243$$

and $\sigma=\sqrt{33.3243}=5.77272$.

The diagonal elements $\mathrm{d.e.}\left(\hat{\beta}_i\right)$ of the matrix $\left(X'X\right)^{-1}$ used to calculate the standard error of the estimate $\mathrm{s.e.}\left(\hat{\beta}_i\right)$.

For example

$$\mathrm{s.e.}\left(\hat{\beta}_0\right)=\sigma\times\sqrt{\mathrm{d.e.}\left(\hat{\beta}_0\right)}=5.77272\times\sqrt{5.03727}=12.9562$$

$$\mathrm{s.e.}\left(\hat{\beta}_1\right)=\sigma\times\sqrt{\mathrm{d.e.}\left(\hat{\beta}_1\right)}=5.77272\times\sqrt{2734.87}=301.89$$

$$\mathrm{s.e.}\left(\hat{\beta}_{15}\right)=\sigma\times\sqrt{\mathrm{d.e.}\left(\hat{\beta}_{15}\right)}=5.77272\times\sqrt{3.03898\times10^{-7}}=0.0032$$

The corresponding t statistic values are

$$t_0=\frac{\hat{\beta}_0}{\mathrm{s.e.}\left(\hat{\beta}_0\right)}=\frac{38.4872}{12.9562}=2.97056$$

$$t_1=\frac{\hat{\beta}_1}{\mathrm{s.e.}\left(\hat{\beta}_1\right)}=\frac{50.2346}{301.8899}=0.166401$$

…

$$t_{15}=\frac{\hat{\beta}_{15}}{\mathrm{s.e.}\left(\hat{\beta}_{15}\right)}=\frac{-0.0163964}{0.0032}=-5.15233$$

Estimated coefficients $\hat{\boldsymbol{\beta}}$, t values and p values shown in **Table 6**. The t test statistic and p values for the significance of each regression coefficient are given in the second and third columns. The significance levels are indicated in the last column. The low p values, with indicate +++, specify significance level less than 0.01, and show that corresponding parameters estimate $\hat{\beta}_i$ should be kept in the model and that they are useful in modeling the fitted values $\hat{y}_i$. Thirteen of

Table 6. Estimated coefficients $\hat{\boldsymbol{\beta}}$, t values and p values. The last column specifies the following significance levels: + corresponds to 0.1, ++ to 0.05, and +++ to 0.01.

$\hat{\beta}_i$	Estimate	t value	p value	SL
$\hat{\beta}_0$	38.4872	2.97056	0.00297	+++
$\hat{\beta}_1$	50.2346	0.1664	0.86784	–
$\hat{\beta}_2$	−0.000147609	−12.901	0.0000	+++
$\hat{\beta}_3$	0.00126975	62.001	0.0000	+++
$\hat{\beta}_4$	0.362444	2.766	0.00568	+++
$\hat{\beta}_5$	2374.26	24.277	0.0000	+++
$\hat{\beta}_6$	−3682.95	−40.428	0.0000	+++
$\hat{\beta}_7$	−31.9369	−49.69	0.0000	+++
$\hat{\beta}_8$	29.1276	6.9648	3.38e−12	+++
$\hat{\beta}_9$	−1072.14	−31.765	0.0000	+++
$\hat{\beta}_{10}$	3275.63	58.026	0.0000	+++
$\hat{\beta}_{11}$	1.95295	4.2167	2.49e−05	+++
$\hat{\beta}_{12}$	−5.67656	−19.103	0.0000	+++
$\hat{\beta}_{13}$	−25.4091	−0.1683	0.8663	–
$\hat{\beta}_{14}$	−25.3162	−0.16772	0.8668	–
$\hat{\beta}_{15}$	−0.0163964	−5.1523	2.6e−07	+++

the sixteen p values are less than 0.01. Thus, all regression coefficients, except three (β_1, β_{13} and β_{14}), are significantly different from zero at the level $\alpha = 0.01$.

The regression sum of squares is computed from Equation (10),

$$\text{SSR} = \hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 39240555.922 - \frac{862490.518^2}{22200} = 5732002.172$$

The sum of squares for error by Equation (11) is:

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{y}$$
$$= 39979813.3969 - 39240555.922 = 739257.475$$

To test the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$, we calculate by Equation (8) statistic F_0:

$$F_0 = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} = \frac{5732002.172/15}{739257.475/(22200-15-1)}$$
$$= \frac{382133.478}{33.324} = 11467.2$$

That is the highly significant result, since $F_0 > f_{0.05,15,\infty} = 1.67$ and even more as $F_0 > f_{0.01,15,\infty} = 2.04$ (or since the p value is considerably smaller than $\alpha = 0.01$). The null hypothesis should be rejected at any reasonable significance level. We conclude that there is a useful linear relationship between y and at least one of the $k = 15$ regressors in the model. This does not mean that all fifteen regressors are useful. It was shown about this relationship in Table 6, as thirteen from sixteen p values are less than 0.01. All regression coefficients except three are significantly different from zero at the level $\alpha = 0.01$.

Computationally adjusted R^2_{adj} statistic is:

$$R^2_{adj} = 1 - \frac{\mathrm{SSE}/(n-k-1)}{\mathrm{SST}/(n-1)} = 1 - \frac{739265.823/(22200-15-1)}{6471259.647/(22200-1)} = 0.886$$

where

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE} = 5732002.172 + 739257.475 = 6471259.647$$

4. Numerical Integration for Cumulative Distribution Functions. Calculations of *t* and *p* Values

To find stable, accurate, and computationally efficient methods for performing big matrix calculations and numerical integration techniques, software implementation with high precision arithmetic is needed. In [5] the numerical examples illustrate the computational accuracy and efficiency of the numerical integration technique, particularly for the direct Laplace transform and its inverse, and C++ implementation of the composite Simpson's Rule for numerical integration (direct Laplace transform). Several algorithms have been proposed in [6] and [10] for numerical integration technique and software implementation with arbirary-precision arithmetic.

Fundamental development in estimation p values, significance levels, and other statistics in ANOVA calculations involve the numerical integration program to approximate CDF of t distribution, F distribution and Normal distribution.

The CDF of a real-valued random variable X is the function given by

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(u)\mathrm{d}u, \quad -\infty < x < \infty \tag{15}$$

The probability density function (PDF) of t distribution is [4]:

$$f(x) = \frac{\Gamma\left[(\nu+1)/2\right]}{\sqrt{\nu\pi}\,\Gamma(\nu/2)} \cdot \frac{1}{\left[\left(x^2/\nu\right)+1\right]^{(\nu+1)/2}}, \quad -\infty < x < \infty \tag{16}$$

where $\nu = n-k-1$ is the number of degrees of freedom and Γ is the gamma function. The formula for the gamma function is: $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t$, for $\alpha > 0$.

The Normal distribution $N(\mu, \sigma^2)$ with parameters μ and σ has PDF:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-(x-\mu)^2/2\sigma^2} \tag{17}$$

The t distribution converges to Normal distribution as the number of degrees

of freedom ν approaches to ∞. The t distribution provides statistical analysis for testing significance levels of regression coefficients.

The F probability distribution has two shape parameters, denotes by ν_1 and ν_2. The parameter ν_1 is the number of numerator degrees of freedom, and ν_2 is the number of denominator degrees of freedom.

The formula for the PDF of F distribution [4] is:

$$f(x)=\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}x^{(\nu_1/2)-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left[\left(\frac{\nu_1}{\nu_2}\right)x+1\right]^{(\nu_1+\nu_2)/2}},\quad 0<x<\infty \tag{18}$$

It is usually abbreviated as f_{ν_1,ν_2}.

The formula for the significance level of the F distribution does not exist in a closed form. It is computed numerically. In Table 7 the significance level α of the F distribution is given as a function of different values of the shape parameters ν_1 and ν_2 and F_0 statistic = 5.

For example if $\nu_1=5$, $\nu_2=1000$ and F_0 statistic = 5, the confidence level is:

$$P\left(F_{\nu_1,\nu_2}>5\right)=1.42\times10^{-9}$$

The 3D plot in Figure 1 follows the outputs given in Table 7.

Table 8 and Figure 2 give the significance level α of the F distribution for different values of the shape parameters ν_1 and ν_2 and F_0 statistic = 10.

Let's consider a software program to perform a numeric integration. The idea is to calculate an approximation of the area under the function's curve between a and b points, as illustrated in Figure 3.

The technique consists in splitting the interval between a and b into n subintervals of equal length. Each subinterval has width $w=(b-a)/n$. The height of that subinterval varies. So we may choose the mid-point x of the subinterval and calculate the height as $h=f(x)$ (the vertical dashed line). If we multiply the height h and the width w, we obtain the area of a rectangle that comes fairly close to the actual area under the function $f(x)$ in that subinterval. Doing this repeatedly, once for each subinterval will yield a pretty good approximation of the entire area under the curve, especially if we have a large number of subintervals,

Table 7. Significance level α of the F_{ν_1,ν_2} distribution and F_0 statistic = 5.

F_0 statistic = 5	$\nu_1=2$	$\nu_1=3$	$\nu_1=5$	$\nu_1=10$	$\nu_1=15$
$\nu_2=5$	6.41e−2	5.76e−2	5.10e−2	4.48e−2	4.25e−2
$\nu_2=10$	3.12e−2	2.26e−2	1.49e−2	8.95e−3	7.07e−3
$\nu_2=100$	8.52e−3	2.84e−3	3.95e−4	7.20e−6	3.22e−7
$\nu_2=1000$	6.91e−3	1.91e−3	1.57e−4	4.10e−7	1.42e−9
$\nu_2=10000$	6.75e−3	1.83e−3	1.41e−4	2.79e−7	6.23e−10

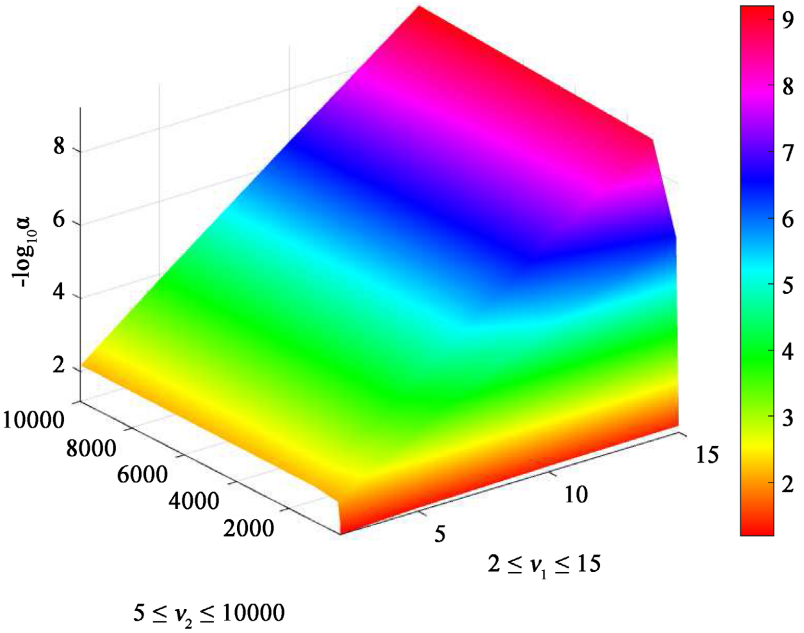


Figure 1. Significance level α of the F_{ν_1,ν_2} distribution and F_0 statistic = 5.

Table 8. Significance level α of the F_{ν_1,ν_2} distribution and F_0 statistic = 10.

F_0 statistic = 10	$\nu_1 = 1$	$\nu_1 = 2$	$\nu_1 = 4$	$\nu_1 = 5$	$\nu_1 = 10$
$\nu_2 = 5$	2.50e−2	1.79e−2	1.33e−2	1.22e−2	1.01e−3
$\nu_2 = 10$	1.01e−2	4.12e−3	1.60e−3	1.21e−3	5.72e−4
$\nu_2 = 100$	2.07e−3	1.10e−4	7.55e−7	8.83e−8	1.90e−11
$\nu_2 = 1000$	1.61e−3	5.01e−5	6.16e−8	2.37e−9	4.44e−16
$\nu_2 = 10000$	1.57e−3	4.59e−5	4.49e−8	1.46e−9	1.11e−16

i.e., if n is large.

The numerical integration program was implemented in C++ high precision software [5]. The calculations can be performed up to 1000 decimal places [11]. The Appendix introduces C++ code used to implement numerical CDF and PDF for t and F distributions in arbitrary precision.

The p value is: $p = 2P\left(X \geq |t|\right) = 2\left(1 - F\left(X\right)\right)$, where the random variable X has t distribution with ν degrees of freedom, and $F\left(X\right) = P\left(X \leq |t|\right)$ is a CDF of X. The p values, corresponding to t distribution with $\nu = n - k - 1 = 22200 - 15 - 1 = 22184$ degrees of freedom, shown in **Table 6**. For example:

$$p_0 = 2 \times P\left(X \geq |t_0|\right) = 2 \times P\left(X \geq 2.97056\right) = 2 \times \left(1 - F\left(2.97056\right)\right)$$
$$= 2\left(1 - 0.998514\right) = 0.00297$$

$$p_1 = 2 \times P\left(X \geq |t_1|\right) = 2 \times P\left(X \geq 0.1664\right) = 2 \times \left(1 - F\left(0.1664\right)\right)$$
$$= 2\left(1 - 0.566079\right) = 0.86784$$

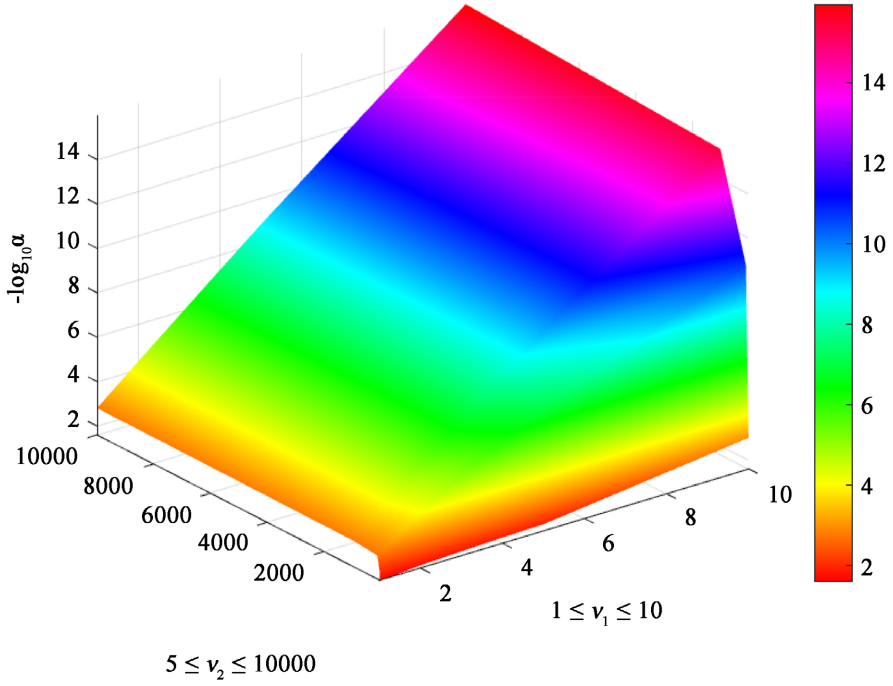


Figure 2. Significance level α of the F_{ν_1,ν_2} distribution and F_0 statistic = 10. Estimated data presented in terms of $-\log_{10}\alpha$.

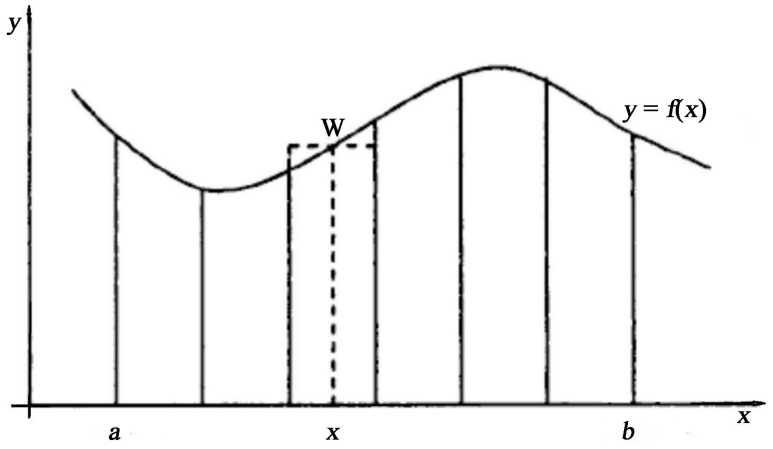


Figure 3. Numerical integration.

$$p_{15} = 2\times P\left(X \geq |t_{15}|\right) = 2\times P\left(X \geq |-5.15233|\right) = 2\times\left(1-F\left(5.15233\right)\right)$$
$$= 2\left(1-0.99999987\right) = 2.6\times 10^{-7}.$$

5. Confidence Interval and Prediction Interval

The confidence interval and prediction interval are calculated by the following equations:

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0'\left(X'X\right)^{-1}x_0} \leq \mu_{y|x_0} \leq \hat{\mu}_{y|x_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0'\left(X'X\right)^{-1}x_0} \quad (19)$$

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\left(1+x_0'\left(X'X\right)^{-1}x_0\right)} \leq Y_0 \leq \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\left(1+x_0'\left(X'X\right)^{-1}x_0\right)} \quad (20)$$

A regression model can be used to predict new or future observations on the response variable Y corresponding to the independent variables, $x_{01}, x_{02}, \cdots, x_{0k}$. If $\boldsymbol{x}_0' = (1, x_{01}, x_{02}, \cdots, x_{0k})$, a point estimate of the future observation Y_0 at the point $x_{01}, x_{02}, \cdots, x_{0k}$ is:

$$\hat{\boldsymbol{y}}_0 = \boldsymbol{x}_0' \hat{\boldsymbol{\beta}}. \tag{21}$$

A prediction interval is always wider than a confidence interval. Confidence interval expresses the error in estimating the mean of a distribution, while the prediction interval expresses the error in predicting a future observation from the distribution at point x_0.

We calculate the confidence interval and the prediction interval using CDF for t distribution and Normal distributions. Now we can do that for the t distribution with PDF Equation (16) and Normal distribution with PDF Equation (17).

6. Multiple Polynomial Regression Model

As shown in Table 6 for multiple linear regression model, all regression coefficients except three (β_1, β_{13}, and β_{14}) are significantly different from zero at the level $\alpha = 0.01$. Three regression coefficients have the p values higher than 0.01 ($\alpha > 0.01$). They are not statistically significant and indicate strong evidence for the null hypothesis.

Example 2. Consider the following multiple polynomial regression model (the second order no-interaction model). There are $k = 15$ initial regressor variables $(x_{i1}, x_{i2}, \cdots, x_{ik}, y_i)$, $i = 1, 2, \cdots, n$, and r additional regressor variables with quadratic terms:

$$\begin{aligned} y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \beta_{k+1} x_{i1}^2 + \beta_{k+2} x_{i2}^2 \\ + \cdots + \beta_{k+r} x_{ik}^2 + \epsilon, \quad i = 1, 2, \cdots, n \end{aligned} \tag{22}$$

Estimated regression coefficients $\hat{\boldsymbol{\beta}}$ are shown in Table 9. The t test statistic and p values for the significance of each regression coefficient are given in the second and third columns. The significance levels are indicated in the last column. The low p values, with indicate +++, specify significance level less than 0.01, and indicate that the corresponding regression parameter estimate $\hat{\beta}_i$ is useful in modeling the fitted values $\hat{y}_i$. Twenty five of the thirty p values are less than 0.01. Thus, all regression coefficients, except five ($\beta_1, \beta_{13}, \beta_{14}$, and β_{29}) are significantly different from zero at the level $\alpha = 0.01$.

For both linear and polynomial regression models, software output produced the desired information, which is summarized in Table 10.

Table 11 shows fitted values, and residuals for n = 5485 data points. Example 3 and Example 4 correspond to linear and polynomial regression models accordingly.

Figure 4 shows two scatter plots for the observations, observation y_i and fitted $\hat{y}_i$ values with n = 22,200 data points. Example 1 (left) and Example 2 (right) correspond to linear and polynomial regression models accordingly.

The data with n = 96 was created as a fragment of the data with n = 22,200

Table 9. Example 2. Estimated coefficients $\hat{\boldsymbol{\beta}}$, t values and p values for multiple polynomial regression model with n = 22,200 and k = 30. The last column specifies the following significance levels: + corresponds to 0.1, ++ to 0.05, and +++ to 0.01.

$\hat{\beta}_i$	Estimate	t value	p value	SL	$\hat{\beta}_i$	Estimate	t value	p value	SL
$\hat{\beta}_0$	223.419	2.68984	0.00714866	+++					
$\hat{\beta}_1$	223.264	0.820289	0.412051	–	$\hat{\beta}_{16}$	−0.00434745	−0.757042	0.449024	–
$\hat{\beta}_2$	0.0013588	46.9057	0.0000	+++	$\hat{\beta}_{17}$	−4.03292e−08	−51.8582	0.0000	+++
$\hat{\beta}_3$	−0.00019503	−3.50041	0.000464625	+++	$\hat{\beta}_{18}$	3.12665e−08	28.0213	0.0000	+++
$\hat{\beta}_4$	−4.94342	−2.15846	0.0308923	++	$\hat{\beta}_{19}$	0.0322251	2.18051	0.0292195	++
$\hat{\beta}_5$	−2706.66	−7.14531	9.04166e−13	+++	$\hat{\beta}_{20}$	55462.2	10.4717	0.0000	+++
$\hat{\beta}_6$	−7144.87	−18.8506	0.0000	+++	$\hat{\beta}_{21}$	47348.6	11.0293	0.0000	+++
$\hat{\beta}_7$	−8.77857	−2.46834	0.0135743	++	$\hat{\beta}_{22}$	−6.11989	−6.13739	8.42156e−10	+++
$\hat{\beta}_8$	−612.272	−39.1742	0.0000	+++	$\hat{\beta}_{23}$	608.313	42.6455	0.0000	+++
$\hat{\beta}_9$	1182.26	13.6697	0.0000	+++	$\hat{\beta}_{24}$	−5727.35	−22.7139	0.0000	+++
$\hat{\beta}_{10}$	6253.54	41.8127	0.0000	+++	$\hat{\beta}_{25}$	−23577.5	−27.4662	0.0000	+++
$\hat{\beta}_{11}$	17.9835	13.3965	0.0000	+++	$\hat{\beta}_{26}$	−14.5793	−8.72223	0.0000	+++
$\hat{\beta}_{12}$	3.07398	1.66984	0.0949509	+	$\hat{\beta}_{27}$	−0.213056	−1.99948	0.0455565	++
$\hat{\beta}_{13}$	−112.235	−0.824712	0.409535	–	$\hat{\beta}_{28}$	0.0153982	4.84613	1.26066e−06	+++
$\hat{\beta}_{14}$	−111.942	−0.822571	0.410752	–	$\hat{\beta}_{29}$	−0.00481319	−1.15303	0.248899	–
$\hat{\beta}_{15}$	0.061281	7.59121	3.19744e−14	+++	$\hat{\beta}_{30}$	−0.000452108	−10.8538	0.0000	+++

Table 10. Example 2. Observation interval, number of observations, observation number, fitted values, and residuals for linear and polynomial regression models. The observation numbers are in accordance with scatter plots 4.

Observations			Linear			Polynomial		
Interval	#	Obs	y_i	$\hat{y}_i$	e_i	y_i	$\hat{y}_i$	e_i
0 - 10	2661	5000	0	−0.06	0.06	0	0.156	−0.156
		4680	4.76	4.97	−0.02	4.76	4.47	0.29
10 - 20	45	6384	11.28	20.22	−8.94	11.28	10.76	0.52
		6385	14.47	20.26	−5.78	14.47	10.83	3.64
20 - 30	1691	14,338	26.73	28.74	−2.01	26.73	27.29	−0.56
		9086	29.79	29.34	0.45	29.79	29.80	−0.01
30 - 40	5363	9722	34.74	36.02	−1.28	37.74	34.76	−0.02
		10,738	38.875	36.94	1.93	38.875	38.88	0.00
40 - 50	7284	10,239	41.14	39.85	1.29	41.14	41.62	−0.48

Continued

		16,792	49.84	51.81	−1.97	49.84	48.43	1.41
50 - 60	3627	17,209	51.32	54.97	−3.59	51.32	51.31	−0.01
		381	58.12	54.40	3.72	58.12	58.39	−0.27
60 - 72	1529	585	60.56	57.06	3.50	60.56	60.84	−0.28
		846	61.55	58.67	2.88	61.55	62.44	−0.89

Table 11. Fitted values, and residuals for $n = 5485$ data points. Example 3 and Example 4 correspond to linear and polynomial regression models accordingly.

Observations Interval	Linear			Polynomial		
	y_i	$\hat{y}_i$	e_i	y_i	$\hat{y}_i$	e_i
32.5 - 37.5	35.91	43.23	−7.32	35.91	36.56	−0.65
37.5 - 42.5	40.43	46.11	−5.68	40.43	43.02	−2.59
42.5 - 47.5	45.00	45.75	−0.74	45.00	45.22	−0.22
47.5 - 52.5	50.13	50.09	0.04	50.41	50.33	0.08
52.5 - 57.5	55.00	54.17	0.83	50.00	54.56	0.44

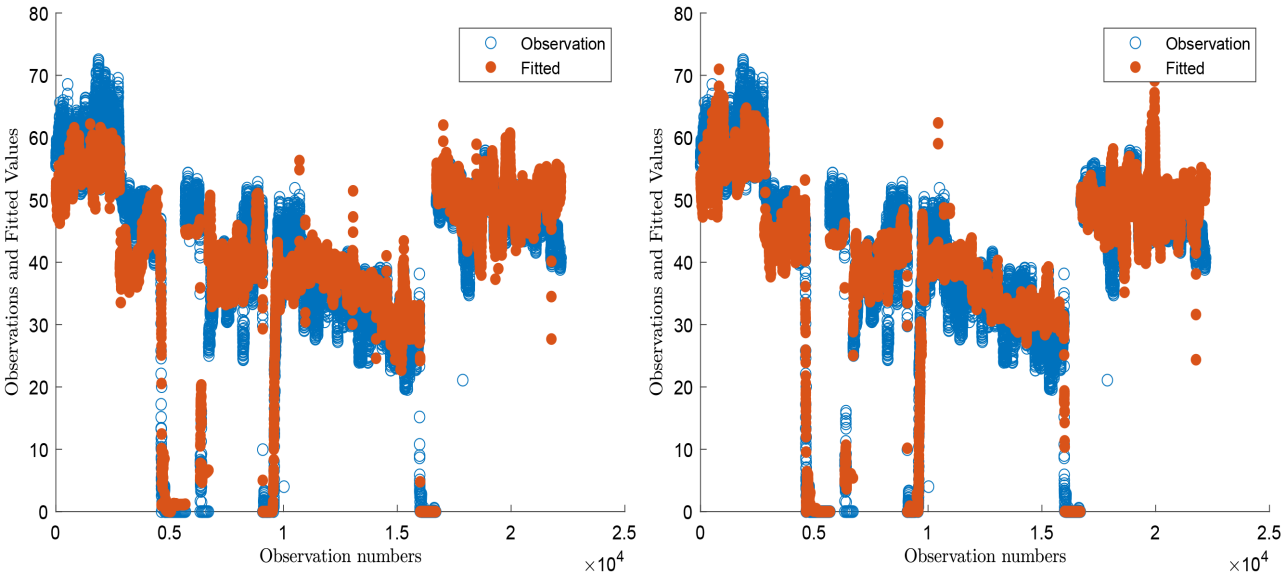


Figure 4. Two scatter plots of the observations y_i and fitted $\hat{y}_i$ values with $n = 22{,}200$ data points. Example 1 (left) and Example 2 (right) correspond to linear and polynomial regression models accordingly.

observations. Figure 5 shows two scatter plots of the observation y_i and fitted $\hat{y}_i$ values of the data with $n = 96$. Example 5 (left) and Example 6 (right) correspond to linear and polynomial regression models accordingly.

Evidently, the two scatter plots in Figure 4 and in Figure 5, are very much alike. They demonstrate similar tracking boundary movements for observed and fitted values and can illustrate whether the algorithm has succeeded in obtaining high-order accuracy or has failed due to numerical instability.

Figure 4 and Figure 5 demonstrate slight improvement, as visually observed, for multiple polynomial regression model vs linear model. These figures illustrate that the fitted and observed values are visually more overlapping on the right scatter plots corresponding to the multiple polynomial regression model.

Quantitative comparison shows slight improvements in accuracy through residual average and the coefficient of multiple determination R^2, which is an important step in building realistic regression models.

For polynomial model with n = 22,200 and k = 30 residual average is 3.825 over 4.571 for linear model (decreased on 16.34%), and adjusted R^2_{adj} statistic is 0.917 over 0.886 for linear model.

For polynomial model with n = 96 and k = 15 residual average is 0.188 over 0.275 for linear model (decreased on 31.6%), and R^2 is 0.968 over 0.943 for linear model.

For the Example 6, the number of terms in approximation was increased from k = 15 (linear model) to $k+r=15+15=30$ (polynomial model. Naturally it would seem that father increase of the number of second-order terms would improve the accuracy of the approximation since more terms seem to produce better solution results.

On the other hand, using a large number of terms will not be of benefit if the precision with which each term is calculated is insufficient.

In Figure 6, n = 5485, k = 15 (Example 3) and k = 30 (Example 4). A slight improvement is visually observed for multiple polynomial regression model over linear model. These figures illustrate that the fitted and observed values are visually more overlapping on the right scatter plot corresponding to the multiple polynomial regression model.

Figure 7 shows two scatter plots of the observation y_i and fitted $\hat{y}_i$ values with n = 8444, for gases NO (Example 7, left) and NO_2 (Example 8, right).

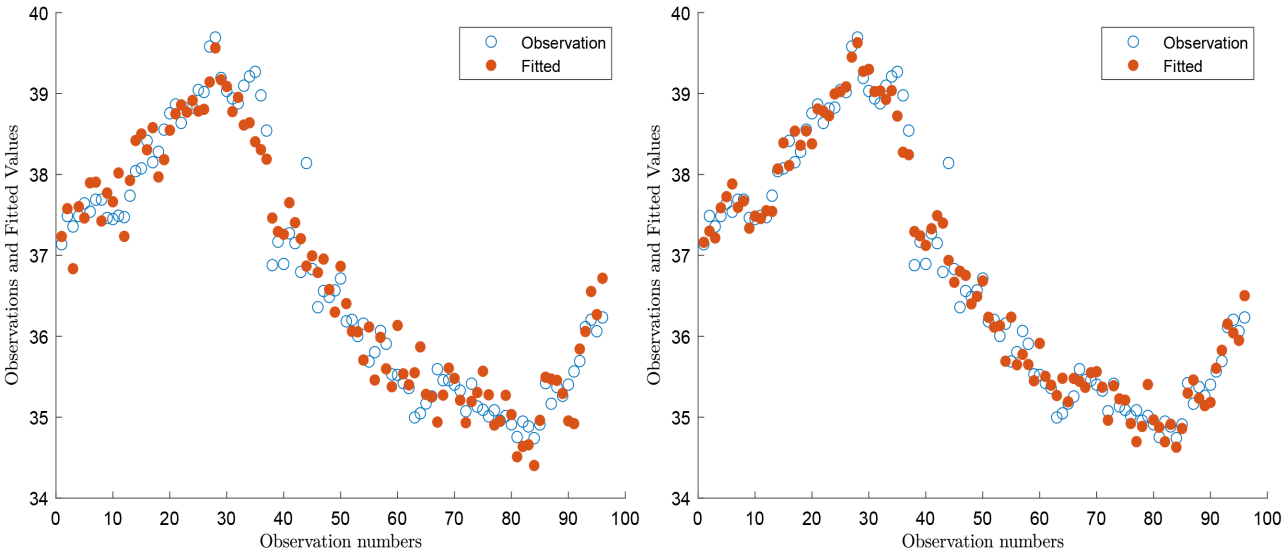


Figure 5. Two scatter plots of the observations y_i and fitted $\hat{y}_i$ values of the data with n = 96. Example 5 (left) and Example 6 (right) correspond to linear and polynomial models accordingly.

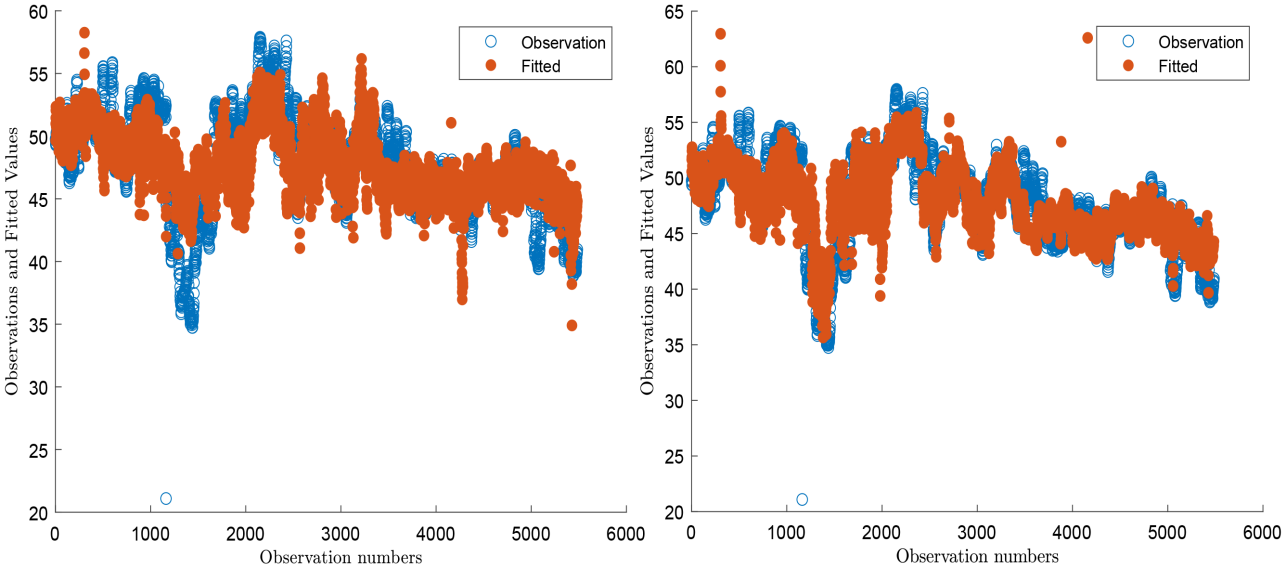


Figure 6. Two scatter plots of the observations y_i and fitted $\hat{y}_i$ values of the data with $n = 5485$ for linear (left) and multiple polynomial regression model (right).

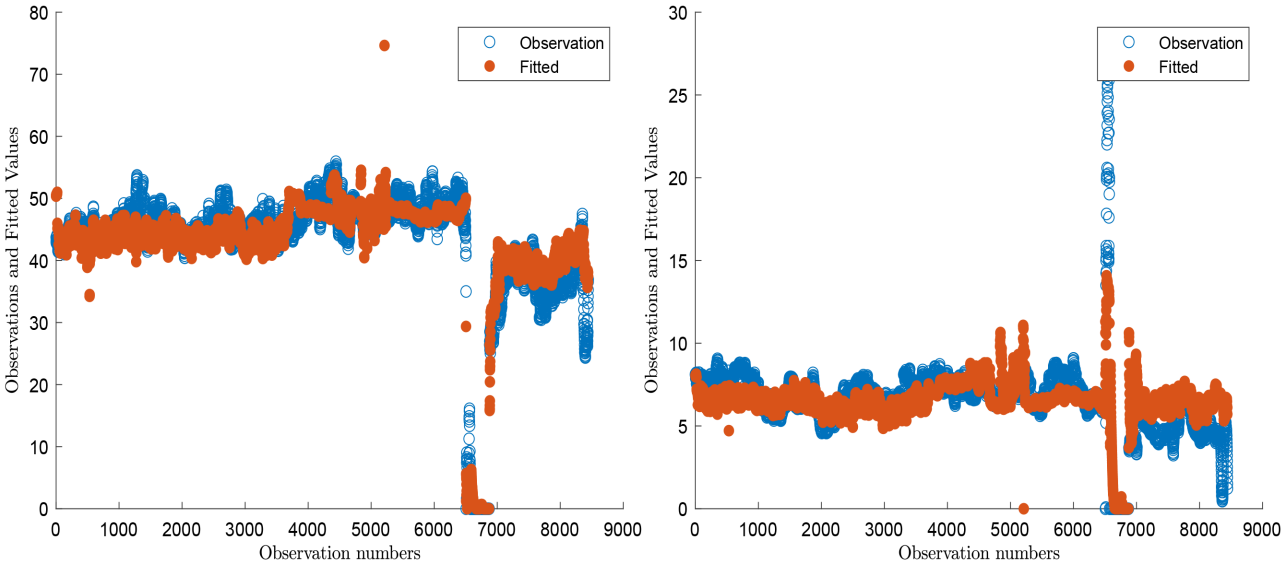


Figure 7. Two scatter plots of the observation y_i and fitted $\hat{y}_i$ values with $n = 8444$, for NO (Example 7, left) and NO_2 (Example 8, right).

Figure 8 illustrates for Example 1 the histogram of residuals using PDF scaling. The area of each bar is the relative number of observations. The sum of the bar areas is equal to 1. The histogram of residuals can be used to check whether the variance is normally distributed. A symmetric bell-shaped histogram which is evenly distributed around zero indicates that the normality assumption is likely to be true. This histogram indicates that random error is not normally distributed, it suggests that the model's underlying assumptions may work well in some areas, and show what can happen to prediction and inference when certain assumptions are violated.

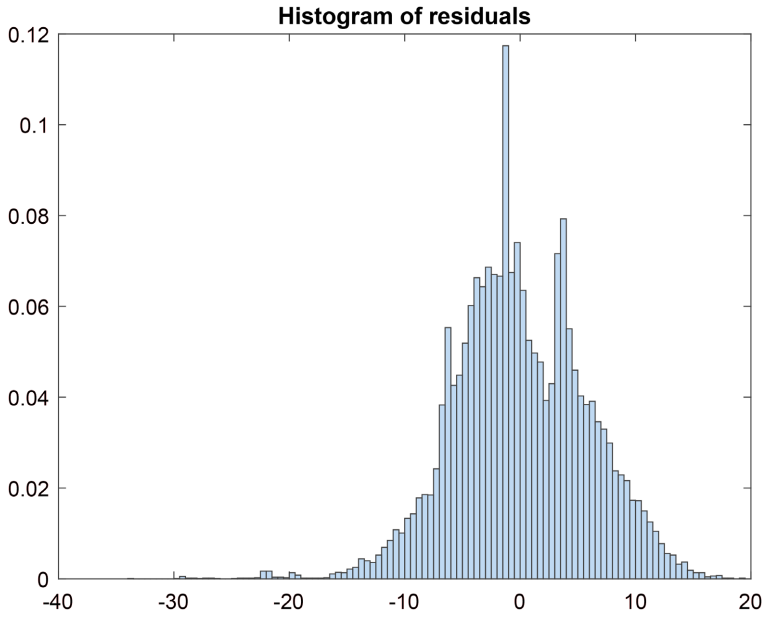


Figure 8. Histogram of residuals.

7. PEMS Mathematical Model versus Actual CEMS Gas Analyzer

PEMS mathematical model was tested by actual CEMS gas analyzer. In **Table 12**, for Example 3 with $n = 5485$, are shown estimated coefficients $\hat{\boldsymbol{\beta}}$, the t test statistic and p values. The significance level of each regression coefficient is indicated in the last column. The low p values, with indicate +++, specify significance level less than 0.01, and show that corresponding parameters estimate $\hat{\beta}_i$ should be kept in the model and that they are useful in modeling the fitted values $\hat{y}_i$. Eleven of the sixteen p values are less than 0.01. Thus, all regression coefficients, except five ($\beta_0, \beta_1, \beta_7, \beta_{13}$ and β_{14}), are significantly different from zero at the level $\alpha = 0.01$.

Figure 6, for $n = 5485$, shows the scatter plot of the observation y_i versus fitted $\hat{y}_i$.

The Data Acquisition System software was provided by Limesoft Inc. [7] to input and test the PEMS mathematical model. We present results and compare them in real-time and on a historical trend chart with actual values obtained from the CEMS gas analyzer.

Figure 9 shows screenshot of PEMS Model—Trend Chart for Example 3 with $n = 5485$. Predicted plot PEMS versus actual CEMS plot is given for 15 min average NO data. Predicted and actual NO values are visibly overlapping. PEMS mathematical model seems to be an effective way to determine emissions based on historical and real-time process data.

8. Testing and Calculation Procedures Based on Performance Specification Requirement

The procedures described in this section provide a framework for testing PEMS

Table 12. Example 3 with $n = 5485$. Estimated coefficients $\hat{\boldsymbol{\beta}}$, t values and p values. The last column specifies the following significance levels: + corresponds to 0.1, ++ to 0.05, and +++ to 0.01.

$\hat{\beta}_i$	Estimate	t value	p value	SL
$\hat{\beta}_0$	24.875	2.0048	0.045028	++
$\hat{\beta}_1$	287.09	0.81338	0.41603	–
$\hat{\beta}_2$	−0.00062175	−11.877	3.8531e−32	+++
$\hat{\beta}_3$	0.00073417	12.413	6.5675e−35	+++
$\hat{\beta}_4$	0.70596	5.2322	1.7375e−07	+++
$\hat{\beta}_5$	−3430.2	−27.226	3.4893e−153	+++
$\hat{\beta}_6$	515.75	3.5411	0.00040175	+++
$\hat{\beta}_7$	0.13826	0.18428	0.8538	–
$\hat{\beta}_8$	−181.15	−22.397	2.2488e−106	+++
$\hat{\beta}_9$	319.43	9.1941	5.2632e−20	+++
$\hat{\beta}_{10}$	1967.2	22.336	7.9487e−106	+++
$\hat{\beta}_{11}$	−15.214	−21.596	2.5086e−99	+++
$\hat{\beta}_{12}$	−10.916	−25.876	2.2745e−139	+++
$\hat{\beta}_{13}$	−143.52	−0.81322	0.41613	–
$\hat{\beta}_{14}$	−141.14	−0.79977	0.42388	–
$\hat{\beta}_{15}$	0.10409	2.7726	0.0055791	+++

model for Example 3 during normal engine operations. They are based on performance specification [8]. PEMS model must pass a relative accuracy (RA) test and accompanying statistical tests to be acceptable for use in demonstrating compliance with applicable requirements. We demonstrate those procedures for the data with $n = 96$ illustrated in the scatter plot in **Figure 5**.

First is the arithmetic mean of the differences:

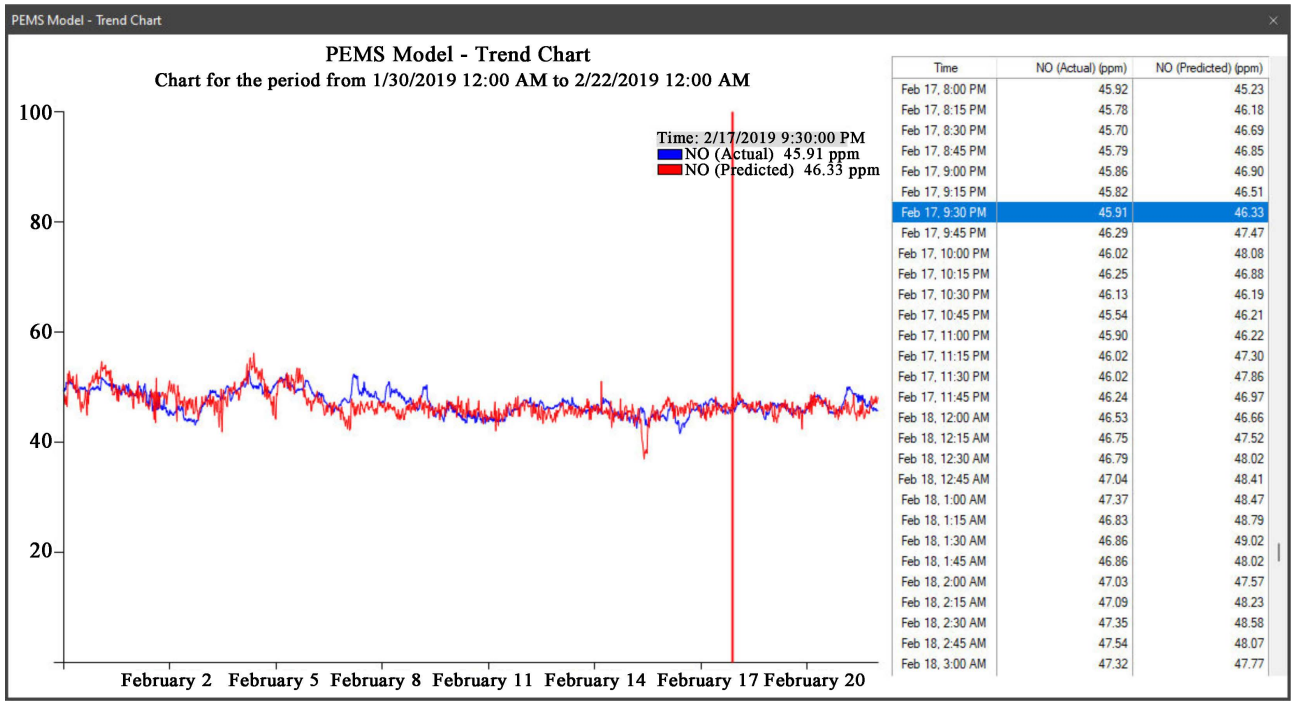
$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{96} \times 3.3899 = 0.03531, \tag{23}$$

where n is the number of data points and d_i is the difference between observation and fitted values.

Second is the standard deviation of the differences:

$$S_d = \sqrt{\frac{\sum_{i=1}^{n} d_i^2 - n * \bar{d}^2}{n-1}} = \sqrt{\frac{11.8414 - 96 * 0.03531^2}{96-1}} = 0.3513 \tag{24}$$

Third is the confidence coefficient for $\alpha = 0.025$ and for $n-1$ degrees of



Time	NO (Actual) (ppm)	NO (Predicted) (ppm)
Feb 17, 8:00 PM	45.92	45.23
Feb 17, 8:15 PM	45.78	46.18
Feb 17, 8:30 PM	45.70	46.69
Feb 17, 8:45 PM	45.79	46.85
Feb 17, 9:00 PM	45.86	46.90
Feb 17, 9:15 PM	45.82	46.51
Feb 17, 9:30 PM	45.91	46.33
Feb 17, 9:45 PM	46.29	47.47
Feb 17, 10:00 PM	46.02	48.08
Feb 17, 10:15 PM	46.25	46.88
Feb 17, 10:30 PM	46.13	46.19
Feb 17, 10:45 PM	45.54	46.21
Feb 17, 11:00 PM	45.90	46.22
Feb 17, 11:15 PM	46.02	47.30
Feb 17, 11:30 PM	46.02	47.86
Feb 17, 11:45 PM	46.24	46.97
Feb 18, 12:00 AM	46.53	46.66
Feb 18, 12:15 AM	46.75	47.52
Feb 18, 12:30 AM	46.79	48.02
Feb 18, 12:45 AM	47.04	48.41
Feb 18, 1:00 AM	47.37	48.47
Feb 18, 1:15 AM	46.83	48.79
Feb 18, 1:30 AM	46.86	49.02
Feb 18, 1:45 AM	46.86	48.02
Feb 18, 2:00 AM	47.03	47.57
Feb 18, 2:15 AM	47.09	48.23
Feb 18, 2:30 AM	47.35	48.58
Feb 18, 2:45 AM	47.54	48.07
Feb 18, 3:00 AM	47.32	47.77

Figure 9. A screenshot of C# software output. Predicted PEMS vs actual CEMS plot are given for 15 min average NO data.

freedom:

$$cc = t_{0.025}\frac{S_d}{\sqrt{n}} = 1.988\frac{0.3533}{96} = 0.071 \tag{25}$$

Fourth is the relative accuracy:

$$\text{RA} = \frac{|\bar{d}| + |cc|}{\bar{y}} \times 100 = \frac{|0.03531| + |0.071|}{36.7638} \times 100 = 0.2896\%, \tag{26}$$

where $|\bar{d}|$ is absolute value of the mean differences, $|cc|$ is absolute value of the confidence coefficient and $\bar{y} = 36.7638$ is the mean of the observation values. This is acceptible result as the RA must not exceed 20% if the PEMS measurements are between 100 ppm and 10 ppm.

Next is the test for significance of regression (with $\alpha = 0.025$). The regression sum of squares computed by Equation (10) is:

$$\text{SSR} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = 195.16939 \tag{27}$$

The sum of squares for error is defined by (11):

$$\text{SSE} = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = 11.84140 \tag{28}$$

F_0 statistic vs constant model is defined by (8):

$$F_0 = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} = \frac{195.16939/15}{11.84140/(96-15-1)} = 87.90 \tag{29}$$

and p value = 6.49×10^{-28}. This low p-value, less than 0.025, is statistically significant. It indicates strong evidence against the null hypothesis.

Coefficient of multiple determination is calculated by Equation (12):

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SSR} + \text{SSE}} = 1 - \frac{11.84140}{195.16939 + 11.84140} = 0.943 \qquad (30)$$

and multiple correlation coefficient is: $R = \sqrt{0.943} = 0.97$. The PEMS correlation is acceptable, as it must be 0.8 or grater.

9. The Role of High Precision Arithmetic in Numerical Calculations

High precision software has been implemented in C++ for calculating numerical Laplace and inverse Laplace transforms [5].

The Appendix introduces C++ code used to implement numerical CDF and PDF for t and F distributions in arbitrary precision.

Table 13 shows sum of squares, F_0 statistic and R^2 evaluated in double with precision $N = 16$ and high precision with $N = 32$ and $N = 128$. The notation * indicates the calculation error in double precision. Fist we compare Equations (11) and (28).

Double precision gives calculation error by formula $\text{sse} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} = 136.407$.

The solution by Equation (28) gives $\text{SSE} = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = 11.84$. The accurate solution $\text{SSE} = 11.72$ gives high precision software.

Next we compare Equations (10) and (27).

Double precision gives calculation error by formula

$$\text{ssr} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 70.604.$$

Table 13. Sum of squares, F statistic and R^2 evaluated in double with $N = 16$ and multiple precision level with $N = 32$ and $N = 128$. The notation * indicates the calculation error in double precision.

#	Formula	$N = 16$	$N = 32$	$N = 128$
1	$\text{ssr} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$	70.604*	195.288	195.288
	$\text{SSR} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$	195.169	195.288	195.288
2	$\text{sse} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$	136.407*	11.723	11.723
	$\text{SSE} = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$	11.841	11.723	11.723
3	$f_0 = \frac{\text{ssr}/k}{\text{sse}/(n-k-1)}$	2.760*	88.856	88.856
	$F_0 = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)}$	87.904	88.856	88.856
4	$r^2 = \frac{\text{ssr}}{\text{ssr} + \text{sse}}$	0.58*	0.94	0.94
	$R^2 = \frac{\text{SSR}}{\text{SSR} + \text{SSE}}$	0.94	0.94	0.94

The solution by Equation (27) as $\mathrm{SSR} = \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2$ gives $\mathrm{SSR} = 195.169$. High precision software gives the accurate solution $\mathrm{SSR} = 195.288$.

10. Comprehensive Methodology for Developing PEMS

PEMS development is based on techniques for statistical data analysis of significant parameters used as inputs to multiple regression model. The methodology includes the following steps.

Data collection: Obtain historical data on dependent fuel and independent NOx emissions from CEMS software installed on the industrial source. Data covers a representative range of operating conditions and emission levels.

Variable selection: Identify the most significant process parameters, such as oxygen, temperature, pressure, fuel flow, and fuel composition, affecting emissions formation. Perform correlation and regression analyses to determine the relationships between these variables and the emission levels.

Model development: Split the historical data into training and testing sets. Use the training set to develop the PEMS model and the testing set to assess its predictive accuracy. Adjust and refine the model as necessary to improve its performance.

Model evaluation: Assess the accuracy of the PEMS model by comparing real-time and historical trend charts of estimated emissions with actual NOx values obtained from CEMS gas analyzers. Verify the model against the performance specification [8] requirements and perform additional tests for PEMS model evaluation and assessment.

Model implementation: Integrate the PEMS model into the plant's control system to provide real-time emission estimates and facilitate data-driven decision-making for emission control and optimization.

Ongoing model maintenance: Periodically update and recalibrate the PEMS model to account for changes in process conditions, equipment performance, and regulatory requirements.

11. Limitations of PEMS Model

Although PEMS offers several advantages over CEMS, it has certain limitations.

Applicability: PEMS may not be suitable for all types of industrial sources, especially those with highly variable or complex emission profiles such as coal-fired power stations or alternative fuel cement plants. Industries with rapidly changing operating conditions or multiple emission sources may require more sophisticated models or hybrid monitoring approaches combining PEMS with CEMS.

Data quality: The accuracy of PEMS is highly dependent on the quality and representativeness of the historical data used for model development. Poor data quality or gaps in data coverage can result in less accurate and reliable PEMS models.

Model adaptability: PEMS models may require regular updates and recalibra-

tion to maintain accuracy as process conditions change over time. This will often require close collaboration between plant operators and PEMS vendors.

12. Environmental Benefits of PEMS over CEMS

PEMS offers several environmental benefits compared to CEMS.

Enhanced process optimization and potential emissions reduction, by providing real-time insights into emission levels. PEMS can help improve overall plant efficiency by providing continuous feedback on process performance. This information can be used to optimize fuel usage, reduce excess air, improve combustion efficiency, and facilitate timely interventions to optimize process parameters and minimize emissions. This proactive approach to emissions control can result in more effective and sustainable emission reduction strategies. Real-time feedback enables plant operators to adjust processes, ultimately reducing the release of pollutants into the environment.

Improved compliance and reporting: PEMS models can help industries maintain compliance with environmental regulations by providing accurate, real-time emissions data. This information can be used to generate reports for regulatory agencies and demonstrate ongoing compliance with emissions limits. Improved compliance can lead to fewer penalties and fines. The adoption of PEMS models signals a commitment to sustainable business practices and environmental stewardship. By demonstrating the implementation of advanced monitoring techniques and a proactive approach to emission reduction, industries can enhance their corporate social responsibility profiles and contribute to global efforts to mitigate climate change and reduce air pollution.

Reduced maintenance and downtime: PEMS models typically require less maintenance than CEMS equipment, resulting in reduced downtime for industrial facilities. This not only lowers maintenance costs but also minimizes the potential for accidental releases of pollutants during maintenance activities. By reducing the need for invasive and time-consuming maintenance procedures, PEMS contribute to a safer and more environmentally friendly workplace.

Reduced resource consumption: PEMS rely on existing process data and sensors, reducing the need for additional hardware and consumables used in CEMS. By minimizing the need for sampling equipment and consumable materials, such as calibration gases, PEMS help reduce the environmental impact associated with waste disposal and resource extraction. PEMS models typically consume less energy compared to the operation of CEMS analyzers, reducing the overall environmental impact. This benefit is particularly significant for large industrial facilities where energy consumption can be a major contributor to greenhouse gas emissions and operating costs. The reduced energy consumption also translates to cost savings for the facility.

13. Conclusion

Software-based PEMS uses data collected by the process sensors and analyzers to

learn how various process parameters, such as oxygen, temperature, pressure, fuel flow, and fuel composition affect the emissions formation. In many cases this data can be used to develop and install PEMS, to replace or to be used instead of CEMS. This development is based on a regression model and statistic methods, but also on previously collected CEMS data. To develop a stable, accurate, and computationally efficient PEMS model this study used high precision C++ software to conduct the multiple regression analysis. Computation statistics methods were used for the estimation of NO and NO_2 emissions from one of the stacks of the petrochemical refinery plant. The methods for PEMS implicated the creation of mathematical models to express the relationship between emissions and various operating and external parameters, such as flue gas temperature, excess combustion air, and heat load. The applicability of PEMS has been tested with multiple regression analysis of big statistical data. Multiple linear regression model allows the response variable to be modeled as a function of more than one input variable. The computations are considerably more complex than in simple linear regression. The most efficient way to deal with multiple linear regression mathematically is by using the matrix algebra approach. In multiple polynomial regression model, the response variable is not expressed as a linear combination of the parameters. Many ideas in the multiple polynomial regression are similar to those in linear regression. The mathematical processes in multiple regression require model fitting and making statistical inferences. We investigated the accuracy of the PEMS model by applying test procedures described in performance specification [11]. The most important result is that the PEMS model was tested and was found to be suitable for continuous emissions monitoring, provided that dependent influencing parameters would continue to operate in levels recorded and used for PEMS model development.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Akritas, M. (2016) Probability & Statistics with R for Engineers and Scientists, Pearson Education.

[2] Chapra, S.C. (2012) Applied Numerical Analysis with Matlab. 3rd Edition, McGraw-Hill.

[3] Hayter, A.J. (2012) Probability and Statistics for Engineers and Scientists, 4th Edition, Brooks/Cole, Cengage Learning.

[4] Montgomery, D.C. and Runger, G.C. (2020) Applied Statistics and Probability for Engineers. 7th Edition, John Wiley & Sons.

[5] Krougly, Z., Davison, M. and Aiyar, S. (2017) The Role of High Precision Arithmetic in Calculating Numerical Laplace and Inverse Laplace Transforms. *Applied Mathematics*, **8**, 562-589. https://doi.org/10.4236/am.2017.84045

[6] Krougly, Z. (2021) Accuracy and Precision Requirement in Probability Models. *Re-*

liability: *Theory & Applications*, **16**, 133-151.

[7] Industrial Software Solutions for Environmental and Process Monitoring. https://limesoft.ca/

[8] US EPA 40CFR 60 PS 16 (2017) Performance Specification 16—Specifications and Test Procedures for Predictive Emission Monitoring Systems in Stationary Sources, https://www.epa.gov/sites/default/files/2017-08/documents/performance_specification_16.pdf

[9] Flowers, B.H. (2000) An Introduction to Numerical Methods in C++. Oxford University Press, Oxford.

[10] Krougly, Z.L., Jeffrey, D.J. and Tsarapkina, D. (2014) Software Implementation of Numerical Algorithms in Arbitrary Precision. 2013 15*th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, Timisoara, 23-26 September 2013, 131-137. https://doi.org/10.1109/SYNASC.2013.25

[11] High-Precision Software Directory. https://www.davidhbailey.com/dhbsoftware/

Appendix: C++ Code Used to Implement Numerical CDF and PDF for *t* and *F* Distributions in Arbitrary Precision

The statistical technique and numerical integration used in PEMS model development were described in Sections 4 and 9. The calculations can be perform up to 1000 places of decimals/1000 significant digits. The C++ code used to implement numerical CDF and PDF for *t* and *F* distributions in arbitrary precision are given below.

```
// Probability density function of F distribution
mp_real f(mp_real x, mp_real u, mp_real v){
    mp_real fpdf = gamma((u + v) / 2) * pow((u / v), (u / 2)) * pow(x, ((u / 2) - 1))/
                gamma(u / 2) * gamma(v / 2) * pow(((u / v) * x + 1), ((u + v) / 2));
   return fpdf;
}
// Cumulative distribution function
mp_real integrate(mp_real(*f)(mp_real x, mp_real u, mp_real v), mp_real a, mp_real b, int n){
    mp_real x, w, fcdf = 0.0;
    w = (b - a) / n;
    x = a + w / 2;
    for (int i = 0; i < n; i++) {
        fcdf += f(x,u,v) * w;
        x += w;
    }
    return fcdf;
}
// Probability density function of t distribution
mp_real t(mp_real x, mp_real n){
   mp_real pi = 2.0 * asin(1.0);
   mp_real tpdf = (1.0 / sqrt(n * pi)) * (tgamma((n + 1.0) / 2.0) / tgamma(n / 2.0)) * pow(1.0 + x * x / n,
        -(n + 1.0) / 2.0);
    return tpdf;
}
```