UNDER PEER REVIEW

Original Research Article

Application of Regression Analysis to Explore the Relationship Between Combustible Gas in Insulating Oil of Power Transformers

ABSTRACT

This paper firstly explains the theory of regression analysis - model, least squares method, coefficient of determination, model assumptions, significance test, etc., those how to calculate through actual data to carry out what these important nouns –such as $\boldsymbol{R^2}, \boldsymbol{SSR}, \boldsymbol{SSE}, \boldsymbol{SST}, \boldsymbol{F}, \boldsymbol{and\ t}$ and derive relevant parameters to being analysis and interpretation. Secondly, for clearing understand what those paraments, author took some data from references to calculate in. The results are presented in reports through the EXCEL application software SPSS system to operating, the same time it provides analysis and interpretation. Finally, the combustible gases in the insulating oil of power transformers - hydrogen (H_2), methane (CH_4), ethane (C_2H_2), ethylene (C_2H_4), acetylene (C_2H_6) and carbon monoxide (CO), these gases were used to diagnose what condition for transformer operation its normal or abnormal, they are important roles in diagnose. Historical case data from Taiwan Power Company were selected and regression analysis was performed to understand the relationship between the various gases. During the analysis, it was found that C_2H_4 had no effect on the increase or decrease of the amount of H_2 gas. This paper attempt hopes to arouse the awareness and interest of cross-disciplinary professionals in regression analysis and thus discover new methods for oil-gas diagnosis. Based on the principle of technology sharing, the research results were written into a paper as reference by scholars and maintenance personnel in the field of power engineering.

Keywords: Variables, Regression analysis, Combustible gases, Regression Analysis Calculator, SPSS (Statistical Program for Social Sciences).

1. INTRODUCTION

Regression method is a tool for analyzing the relationship between variables. It mainly explores the linear relationship between independent variables (x) and dependent variables (y). Through the establishment of regression models, the variables (y) of interest to researchers can be inferred and predicted. This paper discusses the estimation of the regression equation ($\hat{y}_i = b_0 + b_1 \times x_i + \varepsilon_i$), how to obtain the intercept b_0 and slope b_1 values from the independent variable data through the least square method (SSE), and then calculate the total sum of squares (SST) and regression sum of squares (SSR), and then calculate the explanatory power of the coefficient of determination $R^2 = SSR/SST$, significance test and judgment as well as finally calculate ε_i, etc.

In addition, regression analysis of the relationship between the combustible gases in transformer insulating oil - hydrogen (H_2), methane (CH_4), ethane (C_2H_2), ethylene (C_2H_4), acetylene (C_2H_6) and carbon monoxide (CO) has always puzzled me until I understood the theory of regression analysis, read the literature and used the SPSS system in the EXCEL application software and online regression calculator. In actual operation, the combustible gas detection data of many years is input into the above system or calculator for analysis and discussion. The influence relationship between the five combustible gases. Its purpose is to provide equipment maintenance personnel with the coordination of regression analysis and transformer internal maintenance diagnosis. In addition to this section, this paper also includes sections such as literature review, research steps (simple linear and multiple regression analysis), combustible gas regression analysis, review, conclusion and references.

2. LITERATURE REVIEW

The earliest form of regression was the method of least squares, which was applied by Legendre in 1805 and Gauss in 1809 to develop into a statistical method for analyzing data with the aim of understanding the direction and strength of the correlation between two or more variables and the amount of change in the dependent variable when the independent variable changes. We selected papers related to this paper from numerous literature sources and recounted their characteristics as follows: [1] To investigate the correlation between two variables, one is the independent variable (X) or predictor, and the other is the dependent variable (Y) or outcome value; and to calculate the correlation coefficient of the equation using simple examples. [2][3] The meaning of each correlation coefficient in simple linear and multivariate regression analysis is described, and the significance test procedure and standards are used. The process uses practical cases to perform regression analysis calculations. [4] Explore the relationship between multiple independent variables and a dependent variable, explain how to use it and use examples, and especially explain the sequence of steps. [5] The R-squared and adjusted R-squared of regression analysis and the basis for judging whether the hypothesis test falls within the rejection region are explained. However, the significance level of 0.05 is an important discriminant data. [6] Describes the estimated standard error, which is established by the least squares regression line and the numerical value of the measured prediction error. The larger the value, the greater the error; the lower the accuracy, and vice versa. [7] Explain the relationship between the P value and the null hypothesis （H_0） or the alternative null hypothesis （H_1）and how to calculate its value. [8] Instructions for operation and interpretation of online regression analysis calculator. [9][10] Explain the combustible gases obtained by decomposing insulating oil, and its content is used as the basis for judging the internal maintenance of the transformer. [11] List the cases from previous years as supporting evidence.

3. RESEARCH METHODS

Regression analysis is a statistical method of analyzing data, which aims to understand the relationship between one or more independent variables and the corresponding dependent variables, and to establish a mathematical model to understand the changes. Generally, it can be divided into two types: simple linear regression analysis (one independent variable (X) versus one dependent variable (Y) and multiple linear regression analysis (two or more independent variables (X1, X2) versus one dependent variable (Y). The relevant parameter calculation method and result analysis are as follows:

3.1 MODEL ASSUMPTIONS

The original model (taking multiple linear regression as an example):

$$\boldsymbol{y_i = \beta_0 + \beta_1 \times x_{1I} + \beta_2 \times x_{2i} + \varepsilon_i}$$

Convert to an estimated formula:

$$\boldsymbol{\hat{y}_i = b_0 + b_1 \times x_{1I} + b_2 \times x_{2i} + \varepsilon_i}$$

However, the error term must satisfy three major assumptions: (1) normality, (2) independence, and (3) homogeneity of variance.

3.2 HYPOTHESIS TESTING

The regression analysis test is based on the F test, t test and R square coefficient, and their respective characteristics are described as follows:

3.2.1 Significance test of the estimated regression equation (F test): After confirming that the significance p value is < 0.05, the F test is performed and whether the coefficients of all independent variables $\boldsymbol{b_i}$ are 0. Only when the coefficients are not 0 does it have predictive power.

Null hypothesis

$$H_0 = b_1, b_2, \dots b_n = 0$$

Alternative hypothesis

$$H_1 = b_1, b_2, \dots b_n \neq 0$$

Statistics

$$F = MSR/MSE$$

3.2.2 Marginal test of individual regression coefficients (t test): After confirming that the significance p value is < 0.05, a marginal test is performed to explore whether the $\boldsymbol{b_i}$ coefficient of the individual independent variable is 0. When the coefficient is not 0, the independent variable has explanatory power.

Null hypothesis:

$$H_0 = b_1, b_2, \dots b_n = 0\ (i = 1,2, \dots n)$$

Alternative hypothesis:

$$H_1 = b_1, b_2, \dots b_n \neq 0\ (i = 1,2, \dots n)$$

Statistics:

$$t = b_1/S_{b1}$$

$$S_{b1} = S/\sqrt{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n}$$

3.2.3 $\boldsymbol{R^2}$ also known as the coefficient of determination, is an indicator to measure the performance of a regression model. It represents the proportion of the variation in the dependent variable Y that can be explained by the independent variable X. It can also be calculated by subtracting the residual sum of squares (SSE) from 1 and dividing it by the total variation (SST). The calculation formulas for $\boldsymbol{R^2}$, SSR, SSE, SST, etc. are as follows.

$$R^2 = 1 - SSE/SST$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SST = SSR + SSE$$

Therefore, R^2 is not an objective indicator. The number of variables must also be taken into consideration. The adjusted $\mathbf{R^2}$ can be regarded as an unbiased estimate of $\mathbf{R^2}$ and is expressed as $\boldsymbol{\bar{R}^2}$. The calculation is as follows:

$$\boldsymbol{\bar{R}^2 = 1 - (SSE/n - k - 1)/(SST/n - 1)}$$

Where n is the number of samples and k is the number of variables. From the above, we can see that $\boldsymbol{\bar{R}^2}$ (Adjusted R^2) is less than R^2.

4. PROCESS

First, collect the data of variables (including dependent variables and independent variables) and then calculate according to the calculation steps - such as the multiplication value, average value, average value, square value of the difference, etc. of each variable, and then establish a linear regression equation, and then calculate the required parameters such as intercept, slope, error value, etc. Then, the regression, residual, total sum of squares, F and t statistical values, and significant p-values are calculated according to the formula by using the SPSS system of EXCEL for analysis. For convenience, the block flow is presented as follows (Figure 1):

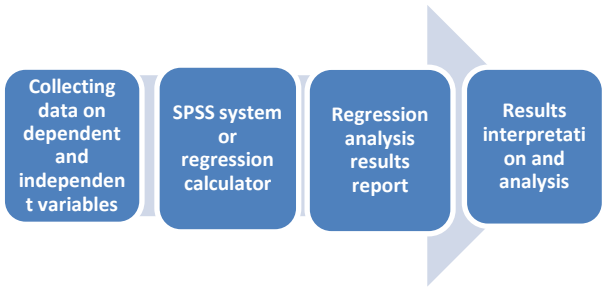

Figure 1. Block flow chart

4.1 EXAMPLE CALCULATION

A convenience store company has ten chain stores. The data on daily advertising and marketing expenses (x_i) and the number of meals sold (y_i) are shown in Table (1). The correlation between the two items is verified by estimating the regression equation to see whether it is valid and has significant predictive power.

Table 1. Advertising and selling list

Location (i)	Advertising (x_i)	selling (y_i)
1	150	156
2	160	180
3	180	190
4	160	170
5	190	198
6	210	250
7	180	189
8	160	168
9	180	191
10	260	280

The sample data is converted into a simple linear regression model such as Formula 1, and then the relevant parameter values are calculated.

$$y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i \quad (1)$$

Where $\boldsymbol{i}$=1,...,n, and $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are unknown, it first considers $\boldsymbol{\varepsilon_i}$ as 0, perform the calculation, and convert the statistical values $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ in the regression model into estimated regression equation parameters $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$, as shown in formula (2)； $\boldsymbol{b_1}$ and $\boldsymbol{b_0}$ are the regression coefficient (slope) and intercept value respectively, ranging from -∞ to +∞. The estimated regression equation parameter values can be obtained by using the least squares method to calculate formulas (3)(4)(5).

$$\hat{y}_i = b_0 + b_1 \times x_i \quad (2)$$

$$b_1 = \sum_{i=1}^{n}[(x_i - \bar{x}) \times (y_i - \bar{y})] / \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (3)$$

$$b_1 = 10924/9410 = 1.1609 \quad (4)$$

$$b_0 = \bar{y} - b_1 \times \bar{x} = -15.2434 \quad (5)$$

The residual square sum algorithm is as shown in formula (6).

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 706.01 \quad (6)$$

The total variance SST algorithm is as shown in formula (7).

$$SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = 13387.60 \quad (7)$$

The squared difference SSR algorithm is as shown in formula (8).

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 = 12681.59 \quad (8)$$

The explanatory power of the relationship between variables is $\boldsymbol{R^2}$ algorithm, as shown in formula (9).

$$R^2 = SSR/SST = 12681.59/13387.60 = 0.9473 \quad (9)$$

The mean square error (MSE) is the residual sum of squares (SSE) divided by its degrees of freedom (n-2), as shown in formula (10).

$$MSE = SSE/n - 2 = 88.2513 \quad (10)$$

The sample standard deviation (S) (estimated standard deviation) is obtained from the square root of the MSE, as shown in formula (11).

$$S = \sqrt{MSE} = \sqrt{SSE/n - 2} = 9.3942 \quad (11)$$

In this paper, the sample standard deviation (S) is considered to be consistent with the random error ($\boldsymbol{\varepsilon_I}$), so the estimated regression equation is as shown in formula (12).

$$\hat{y}_i = b_0 - b_1 \times x_i + \varepsilon_i = -15.243 + 1.161 \times x_i + 9.394 \quad (12)$$

4.2 COEFFICIENT OF DETERMINATION

The sum of squares caused by regression is also the proportion of the explained variation to the total sum of squares (total variation). The symbol $\boldsymbol{R^2}$ represents the numerical range of 0~1. The closer it is to 1, the better the explanatory power. However, its significance (p value) must be less than 0.05.

$$R^2 = SSR/SST = 12681.59/13387.60 = 0.9473$$

The significance test depends on the regression relationship between the dependent variable and the independent variable. It is necessary to conduct a hypothesis test on whether the slope $\boldsymbol{b_1}$ is equal to 0, and then calculate the t-test statistic value through the sample standard deviation.

S_{b1} is the standard deviation of the b_1 coefficient (σ_{b1}), as shown in formula (13).

$$S_{b1} = S/\sqrt{\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} (x_i)^2/n}$$

$$= 9.3942/97.0052 = 0.0968 \qquad (13)$$

$$\text{t test statistic} = b_1/S_{b1} = 1.1609/0.0968 = 11.9875$$

The degree of freedom (df) refers to the number of independent or freely changing data in the sample when the sample statistic is used to estimate the number of mothers in the population. For example, if the sample is N, the degree of freedom of the sample mean is N-1. However, $\boldsymbol{df}$ is divided into three types according to its attributes: regression term $\boldsymbol{df_r}$., error term $\boldsymbol{df_e}$ and total variation $\boldsymbol{df_t}$.

$$df_r. = 1 \text{ 與} df_e = n - k - 1 \text{ 及} df_t = n - 1$$

$$MSR = SSR/df_r \qquad (14)$$

$$MSE = SSE/df_e \qquad (15)$$

The F test statistical value algorithm is as follows:

$$F = MSR/MSE = (SSR/1)/(SSE/n - k - 1) \qquad (16)$$

$$F = 12681.59/88.2513 = 143.699$$

A high $\boldsymbol{R^2}$indicates a high degree of explanation. If too many variables are included, the coefficient estimate will become unstable. Therefore, $\boldsymbol{R^2}$ must take the number of variables into consideration, so $\boldsymbol{R^2}$ is adjusted to make it an unbiased estimate. After adjustment, it is expressed as $\boldsymbol{\bar{R}^2}$ as shown in formula (17), where n is the sample size and k is the number of variables. It is known that $\boldsymbol{\bar{R}^2 < R^2}$.

$$\bar{R}^2 = 1 - (SSE/n - k - 1)/(SST/n - 1) \qquad (17)$$

$$\bar{R}^2 = 0.9407$$

4.3 SIMPLE LINEAR REGRESSION REPORT

From EXCEL through the SPSS system, the steps are - data - data analysis - regression - input relevant data in sequence - execution results, as shown in Table (2).

Table 2. Simple estimated regression equation report

Regression Statistics								
Multiples of R	0.973274883							
R^2	0.947263998							
Adjusted R^2	0.940671998							
Standard error	9.394203676							
Number of observations	10							
ANOVA								
	Degrees of Freedom	SS	MS	F	Significant value			
Regression	1	12681.5915	12681.6	143.699	2.16103E-06			
Residual	8	706.0085016	88.2511					
Sum	9	13387.6						
	Coefficient	Standard error	t statistic	P-value	Down 95%	Up 95%	Down 95.0%	Up 95.0%
Intercept	-15.24335813	17.96940394	-0.8483	0.42093	-56.68087793	26.19416167	-56.68087793	26.19416167
Advertising	1.160892667	0.096842314	11.9875	2.2E-06	0.93757389	1.384211445	0.93757389	1.384211445

From the regression statistics and ANOVA and $\boldsymbol{b_0}$,$\boldsymbol{b_1}$ parameter values in Table (3), the test results are as follows:

1. $\boldsymbol{R^2}$ = 0.9472 means that 96.03% of the total variation in the regression model can be explained by the independent variables, and the adjusted $\boldsymbol{\bar{R}^2}$= 0.9406.
2. The significance test of the regression model calculated the F value to be 143.699 and the significance (p) value to be 0.000002161, rejecting the null hypothesis and indicating that the model has predictive ability.
3. The single t-statistic value of advertising (X1) is 11.987, and the significance (p) value is 0.000002161, rejecting the null hypothesis. Therefore, marketing advertising have a

significant impact on the dependent variable number of selling (Y).

4. Error test: The residual value must obey the normal distribution assumption, and the standard error $\boldsymbol{\varepsilon_i}$= 9.3942.

4.4 MULTIVATE ESTIMATION REGRESSION

Continuing from the previous paragraph, add another independent variable (pricing item), and the variable related data is shown in Table (3), then the simple estimation regression analysis becomes a multivariate estimation regression analysis, and the relevant steps are described as follows:

Table 3. Data for Advertising and Price and number of selling

No	Advertising(X1)	Price(X2)	Selling(Y)
1	150	150	156
2	160	140	180
3	180	125	190
4	160	137	170
5	190	120	198
6	210	99	250
7	180	119	189
8	160	145	168
9	180	124	191
10	260	96	280

Exploring the relationship between two or more (symbol X represents the number of independent variables) independent variables and one dependent variable is called multiple regression analysis. The relationship between the independent variables $(\boldsymbol{x_1, x_2, \ldots x_n})$ and the dependent variable (Y) and the error term $\boldsymbol{\varepsilon_i}$ is called a deterministic mathematical model, formula (16). Where $\boldsymbol{i}$= $\boldsymbol{1, \ldots, n}$.

$$\boldsymbol{y_i = \beta_0 + \beta_1 \times x_{1I} + \beta_2 \times x_{2i} + \varepsilon_i} \qquad (16)$$

Usually, the actual values of the parameters $(\boldsymbol{\beta_0, \beta_1, \ldots \beta_n})$ in the multiple regression model are unknown, so they are estimated by using the sample data values, and the sample statistical values $\boldsymbol{b_0, b_1, \ldots b_n}$ are used to replace the parameters $\boldsymbol{\beta_0, \beta_1, \ldots \beta_n}$ in the regression model, as shown in formula (17).

$$\boldsymbol{\hat{y}_i = b_0 + b_1 \times x_{1I} + b_2 \times x_{2i} + \varepsilon_i} \qquad (17)$$

When calculating the parameters of $\boldsymbol{b_1}$, $\boldsymbol{b_2}$ and $\boldsymbol{b_0}$, $\boldsymbol{\varepsilon_i}$ should be considered as 0.

To simplify the program calculation, several characters (such as D, E, F, etc.) will be used to represent different calculation methods, as shown below:

$\boldsymbol{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)}$ **refers as** $\boldsymbol{D}$ ◦

$\boldsymbol{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)}$ **refers as** $\boldsymbol{E}$ ◦

$\boldsymbol{\sum_{i=1}^{n}(y_i - \bar{y})}$ **refers as F** ◦

$\boldsymbol{[(D^2 \times E^2) - (D \times E)^2]}$ **refers as G** ◦

The abbreviations of the related parameter calculation formulas are as follows: Formulas (18), (19), and (20).

$$\boldsymbol{b_1 = E^2 \times D \times F - (D \times E) \times (E \times F)/G} \qquad (18)$$

$$\boldsymbol{b_2 = D^2 \times E \times F - (D \times E) \times (D \times F)/G} \qquad (19)$$

$$\boldsymbol{b_0 = \bar{y}_i - b_1 \times \bar{x}_1 - b_2 \times \bar{x}_2} \qquad (20)$$

The calculation steps for the other relevant parameter values are consistent with those in the previous paragraph, so they will not be repeated here. The execution results are shown in Table (4).

Table 4. Multivariate estimated regression equation report

Regression Statistics								
Multiples of R	0.979939996							
R^2	0.960282395							
Adjusted R^2	0.948934508							
Standard error	8.715531367							
Number of observation	10							
ANOVA								
	Degrees of Freedom	SS	MS	F	Significant value			
Regression	2	12855.9	6427.94	84.62213117	1.24865E-05			
Residual	7	531.723	75.9605					
Sum	9	13387.6						
	Coefficient	andard err	t statistic	P-value	Down 95%	Up 95%	Down 95.0%	Up 95.0%
Intercept	119.3050272	90.3774	1.32008	0.228337092	-94.40344971	333.0135041	-94.40344971	333.014
Advertising	0.845233009	0.22694	3.72455	0.007411658	0.308614896	1.381851121	0.308614896	1.38185
Price	-0.611814086	0.40391	-1.5147	0.173613266	-1.566906423	0.343278251	-1.566906423	0.34328

From the regression statistics and ANOVA and $\boldsymbol{b_0}$, $\boldsymbol{b_1}$, $\boldsymbol{b_2}$ parameter values in Table (4), the test results are as follows.

1. $\boldsymbol{R^2}$ = 0.9603 means that 96.03% of the total variance in the regression model can be explained by the independent variables. The adjusted $\overline{\boldsymbol{R}}^{\boldsymbol{2}}$=0.9489.
2. The significance test of the regression model calculated the F value to be 84.6221 and the significance (p) value to be 0.0000125, rejecting the null hypothesis and indicating that the model has predictive ability.
3. The calculated t-statistic value of advertising (X1) is 3.7245, and the significance (p) value is 0.0074, rejecting the null hypothesis.
4. The t-statistic value calculator price (X2) is -1.5147, and the significance (p) value is 0.1736, and the null hypothesis is accepted; therefore, price (X2) has no significant effect on the change in the number of selling (Y).
5. Error test: The residual value must obey the normal distribution assumption, and the standard error $\boldsymbol{\varepsilon_i}$ = 8.7155.

The multivariate estimation regression analysis equation is as shown in formula (21).

$$\widehat{y_i} = b_0 - b_1 \times x_{1I} + b_2 \times x_{2i} + \varepsilon_i$$

$$= 119.30 - 0.6118 \times x_{1i} + 0.945 \times x_{2i} + 8.7155 \ (21)$$

The t-statistic value of the above-mentioned pricing (X2) is -1.5147, and the significance (p) value is 0.1736. The null hypothesis is accepted, and there is no significant effect on the dependent variable (Y). Therefore, it is necessary to eliminate the independent variable (X2) and re-execute the result which seems to a simple linear regression analysis, which is consistent with Table (2), Formula (12) and the test analysis.

There is also a regression analysis calculator [8] that readers can refer to and compare for themselves.

4.5 SELF-MADE REGRESSION CALCULATOR

Based on the above theories and calculation techniques, the author developed a simple binary regression analysis calculator to facilitate the analysis and identification of relevant data for being easily, shown as Table (3). To create a data table, insert calculation formulas at different positions in the EXCEL application software through normal steps sequent to compile out of a results sheet; the actual execution results are shown in Table (5).

Table 5. Self-made regression analysis report

Formula=	119.305	+	0.845	*	X1	+	-0.612	*	X2
H0 or H1 Hyp.	H1								
R^2 =	0.9603								
Adjusted R^2=	0.9489								
SST=	13387.6000								
SSR=	12855.8766								
SSE=	531.7234								
MSR=	6427.9383								
MSE=	75.9605								
Standard Error=	8.7155								
F Statistic=	84.6221								
F(Significance P)=	0.00402515								
t Statistic(b1)=	3.7245								
t Statistic(b2)=	-1.5147								
t Statistic(b0)=	1.3201								
(b1)(Significance P)=	0.00089647		(b2)(Significance P)		0.00144125		(b0)(Significance P)		2.57536E-08

From the analysis and comparison of Table (4) and Table (5), only the significant values in the F test and the t test are different, but both have the ability to predict the regression impact, and the other result values are consistent. Therefore, it is proved that this self-made calculator can be applied to binary regression analysis.

5. COMBUSTIBLE GAS REGRESSION ANALYSIS

This section studies the combustible gases decomposed in transformer insulating oil - hydrogen (H_2), acetylene (C_2H_2), methane (CH_4), ethylene (C_2H_4), ethane (C_2H_6), carbon monoxide (CO) and other gases. The content of these gases is of great importance for the detection and interpretation of internal abnormalities of transformers in service. For more information, please refer to references [9]. [10]. [11]. In order to try to use regression analysis equations to analyze the relationship between combustible gases, the author specifically cited the combustible gas content detection data of 10 groups of abnormal conditions in Taiwan Power Company over the years, as shown in Table (6), and directly executed the test results using the EXCEL software SPSS system, as shown in Table (7).

Table 6. Combustible gases tested data list

No	C_2H_4	C_2H_2	CH_4	C_2H_6	CO	H_2
1	7	10.1	41	88	57	44
2	11	36	10	2	685	113
3	21	53.8	54	79	35	140
4	58	51.4	271	116	128	181

5	82	167	74	8	39	418	8	930	1446	884	314	199	1718
6	787	23.7	346	78	312	239	9	590	1.7	601	208	72	185
7	1077	0.4	694	356	36	48	10	384	1.9	211	66	411	133

Table 7. Combustible gases multivariate estimation regression equation report

Regression Statistics								
Multiples of R	0.999559802							
R^2	0.999119798							
Adjusted R^2	0.998019546							
Standard error	22.3422227							
Number of observations	10							
ANOVA								
	Degrees of Freedom	SS	MS	F	Significant value			
Regression	5	2266460.2	453292.0401	908.0825702	3.38657E-06			
Residual	4	1996.699661	499.1749153					
Sum	9	2268456.9						
	Coefficient	Standard error	t statistic	P-value	Down 95%	Up 95%	Down 95.0%	Up 95.0%
Intercept	187.1545401	16.22807501	11.53276282	0.000322818	142.0981807	232.211	142.0981807	232.2108995
C2H4	0.048250649	0.044905398	1.074495527	0.343096743	-0.076426723	0.17293	-0.076426723	0.172928022
C2H2	1.077297543	0.023817924	45.23053968	1.42893E-06	1.011168385	1.14343	1.011168385	1.1434267
CH4	0.533090756	0.099240811	5.371688809	0.005800763	0.257554091	0.80863	0.257554091	0.808627422
C2H6	-1.603078386	0.18115576	-8.849171481	0.000900416	-2.10604741	-1.1001	-2.10604741	-1.100109363
CO	-0.18433991	0.042281423	-4.359832229	0.012063719	-0.30173196	-0.0669	-0.30173196	-0.066947861

From the regression statistics and ANOVA and these $\boldsymbol{b_0}, \boldsymbol{b_1}, \boldsymbol{b_2}, \boldsymbol{b_3}, \boldsymbol{b_4}, \boldsymbol{b_5}$ parameter values in Table (7), the test results are as follows:

1. $\boldsymbol{R^2}$ = 0.9991 means that 99.91% of the total variation in the regression model can be explained by the independent variables, and the adjusted $\overline{\boldsymbol{R}}^{\boldsymbol{2}}$=0.9980.
2. The regression model significance test calculated the F value to be 908.083 and the significance (p) value to be 0.0000034, rejecting the null hypothesis and showing predictive ability.
3. The calculated t-statistic value of C_2H_4(X1) is 1.0744, and the significance (p) value is 0.343, so the null hypothesis is accepted; therefore, C_2H_4(X1) has no significant effect on Y(H_2) and needs to be eliminated.
4. Error verification: The residual value must obey the normal distribution, and the standard error $\boldsymbol{\varepsilon_i}$ = 22.3422.
5. The results of the multivariate regression analysis of transformer insulating oil test data with hydrogen (H_2) as the dependent variable (Y) and the other acetylene (C_2H_2), methane (CH_4), ethane (C_2H_6), and carbon monoxide (CO) as independent variables (X2,..,X5) are shown in formula (22).

$$\hat{y}_i = 187.155 + 1.0773 \times x_2 + 0.533 \times x_3 - 1.6031 \times x_4 - 0.1843 \times x_5 + 22.3422 \quad (22)$$

In addition, multiple regression analysis will be used to treat the six types of combustible gases as dependent variables (Y) and the other five gases as independent variables (X1,...,X5) respectively. After the regression analysis, the five groups of results executed by the SPSS system, except $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$**($H_2$), a**re summarized as shown in Table (8).

Table 8. Six types of combustible gases detection reports

Input the data in Table (6) into the multivariate estimated regression equation and use the regression analysis SPSS system to calculate the relevant parameters, as shown in the following formulas.
$\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{b_0} + \boldsymbol{b_1} \times \boldsymbol{x_{1i}} + \boldsymbol{b_2} \times \boldsymbol{x_{2i}} + \boldsymbol{b_3} \times \boldsymbol{x_{3i}} + \boldsymbol{b_4} \times \boldsymbol{x_{4i}} + \boldsymbol{b_5} \times \boldsymbol{x_{5i}} + \boldsymbol{\varepsilon_i}$
$\hat{y}_i$(C_2H_2)= -173.66 + 0.926 X x_{1i}(H_2) – 0.489 X x_{2i}(CH_4) + 1.482 X x_{3i}(C_2H_6) + 0.171 X x_{4i}(CO) + 20.715 $\boldsymbol{R^2}$= 0.999. $\overline{\boldsymbol{R}}^{\boldsymbol{2}}$= 0.997. F-statistic: 840.242, P-value: 0.00000395; The P-value of C_2H_4 does not meet the requirements and has no significant effect on the dependent variable and needs to be eliminated.
$\hat{y}_i$(CH_4)= - 307.13 + 1.647 X x_{1i}(H_2) – 1.757 X x_{2i}(C_2H_2) + 2.792 X x_{3i}(C_2H_6) + 0.305 X x_{4i}(CO) + 39.277 $\boldsymbol{R^2}$= 0.996. $\overline{\boldsymbol{R}}^{\boldsymbol{2}}$= 0.983. F-statistic: 110.32, P-value: 0.000225; The P-value of C_2H_4 does not meet the requirements and has no significant effect on the dependent variable and needs to be eliminated.
$\hat{y}_i$(C_2H_6)= 112.94 – 0.593 X x_{1i}(H_2) + 0.638 X x_{2i}(C_2H_2) + 0.0285 X x_{3i}(CH_4) - 0.1143 X x_{4i}(CO) + 13.594 $\boldsymbol{R^2}$= 0.997. $\overline{\boldsymbol{R}}^{\boldsymbol{2}}$= 0.987. F-statistic: 1743.81, P-value: 0.000133; The P-value of C_2H_4 does not meet the requirements and has no significant effect on the dependent variable and needs to be eliminated.

$\hat{y}_i$(CO)= 888.45 - 4.481 X x_{1i}(H_2) + 4.844 X x_{2i}(C_2H_2) + 2.399 X x_{3i}(CH_4) - 7.504 X x_{4i}(C_2H_6) + 110.16 $\boldsymbol{R^2}$= 0.939. $\boldsymbol{\bar{R}^2}$= 0.736. F-statistic: 6.0274, P-value: 0.05318; The P-value of C_2H_4 does not meet the requirements and has no significant effect on the dependent variable and needs to be eliminated.
Although the $\hat{y}_i(C_2H_4)$ term has $\boldsymbol{R^2}$= 0.937. $\boldsymbol{\bar{R}^2}$= 0.728. F statistic value: 5.8323, P value: 0.05615; but the P values of H_2.CH_4.C_2H_2.CO.C_2H_6 are all > 0.05, which does not meet the requirements. Therefore, the $\hat{y}_i(C_2H_4)$ term needs to be eliminated.

From Table (8), we know that each of the six combustible gases is used as the dependent variable (Y) and the other five gases are used as independent variables (X1,...,X5) to form a set of multiple regression equations. Among the total of six sets of equations, the C_2H_4 group only has a significant effect on C_2H_6, and the other independent variables have no significant effect and need to be eliminated. In the other five groups (H_2.C_2H_2.CH_4.C_2H_6.CO), the independent variable C_2H_4 had no significant effect on the dependent variable and had to be eliminated. In addition, the change of each unit number of each independent variable in the CO group will affect the change of the dependent variable. For example, the parameter of C_2H_6 in the regression equation is -7.504. When it increases by one unit, the change of CO group decreases by 7.504. The increase or decrease of the coefficients of each variable in the other groups will be similar. From the above analysis, it is found that the relationship between C_2H_4 gas and H_2 gas is very small, that is, C_2H_4 has no influence on the increase or decrease of the amount of H_2 gas.

6. DISCUSSION

The size of the $\mathbf{R^2}$ value must be consistent with rejecting the null hypothesis ($\mathbf{H_1}$: $\mathbf{b_1} \neq \mathbf{0}$) and its significance (p value) must be lower than 0.05 to have substantial explanatory power. As for $\mathbf{\bar{R}^2}$, it is the data that has been appropriately adjusted ($\mathbf{\bar{R}^2} < \mathbf{R^2}$). The t-test and F-test statistics must be consistent with rejecting the null hypothesis ($\mathbf{H_1}$: $\mathbf{b_1} \neq \mathbf{0}$) and their significance (p-value) must be less than 0.05 to have moderate explanatory power. However, the F and t statistics for simple linear regression analysis are different in size but have the same p-value. The degrees of freedom are divided into three types: the degrees of freedom of the regression term ($\mathbf{df_r}$) is 1 for a group of independent variables, the degrees of freedom of the error term ($\mathbf{df_e}$) is n-p-1, and the degrees of freedom of the total variation term ($\mathbf{df_t}$) is n-1. When two independent variables are not independent of each other, they have collinearity, which will cause the regression model to have duplicate explanatory variables, resulting in incorrect explanatory power and prediction power, which is called the variation inflation factor (VIF). For example, the formula VIF=$\mathbf{1/1 - R^2}$. Determine whether the independent variables of the multivariate estimation regression model are independent. The smaller the VIF value, the better. If it is greater than 10, it means that the independent variables are linear and one of them should be eliminated. Self-made regression calculators have been shown to be useful for binary regression analysis. The SPSS system in the EXCEL application software is slightly different from the self-made and online regression calculator in terms of usage, so the result reports are also different. If you are interested, you can go online to operate and compare. However, all calculations in this paper are done using the EXCEL software SPSS system. The difference between error and residual is the difference between the observed value relative to the population mean and the observed value relative to the sample mean, but in this paper the two are considered to be the same. Through multivariate regression analysis, we can understand the relationship between the contents of each combustible gas (C_2H_4.C_2H_2.CH_4.C_2H_6.CO) and the change in the amount of hydrogen (H_2).

7. CONCLUSION

Regression analysis is an effective tool for evaluation and forecasting as well as for supporting research data. This paper uses theoretical explanations, combined with actual data, to perform calculations and then uses application software to generate report results for analysis and interpretation. From this can see that as long as it was been collect enough variables (independent variables, dependent variables), being can carry out the above analysis and judgment. In another case, the correlation between the contents of various combustible gases in the insulating oil of power transformers was been explored through regression analysis, from which it was been found that C_2H_4 had no influence on the increase or decrease of H_2 gas volume. This paper

uses the EXCEL software SPSS system to generate various result reports for analysis and interpretation. Finally, author will write down what these experiences of study to share it with those technicians who worked in the field of electricity for reference and hopes that senior scholars will give me criticism and suggestions.

REFERENCES

1. Silvia Valcheva. Simple Linear Regression Examples. https://www.intellspot.com/Linear-Regression-Examples. Jun 5 2025. On website.
2. Simple Linear Regression Analysis, Kaohsiung University of Science and Technology. https://www2.nkust.edu.tw. 5 30 2025.on website.
3. Multiple regression analysis, Kaohsiung University of Science and Technology. https://www2.nkust.edu.tw.6 3 2025. On website.
4. Multiple linear regression analysis, Yongxi Statistics Consulting Consultant. https://www.yongxi-stat.com/multiple-regression-analysis/ 5 30 2025.on website.
5. Qiu Bingcheng, data analysis, Medium Statistics in Carrot Cheng，on Mar 5, 2022. https://medium.com/qiubingcheng/ 5 30 2025. On website.
6. Estimate the standard error of the measurement prediction error. https://drfishstats.com/regression/standard-error-of-estimate/ 5 31 2025. On website.
7. Tutorial on calculating p-value in a spreadsheet, KDAN OFFICE. https://kdan-office.kdandoc.com/ 5 30 2025. On website.
8. Multiple Linear Regression Calculator, Statistic Kingdom. https://www.statskingdom.com/multi_linear_regression.him/ 6 6 2025. On website.
9. Ming-Jong Lin, Liang-Bi Chen, Chao-Tang Yu. A Methodology for Diagnosing Faults in Oil-Immersed Power Transformers Based on Minimizing the Maintenance Cost. IEEE ACCESS (Volume: 8).17 Nov 2020.
10. H. Malik, Tarkeshwar, and R. K. Jarial, An expert system for incipient fault diagnosis and condition assessment in transformers, in Proc. Int. Conf. Computer. Intellect. Commun. Net. Gwalior, India, Oct. 2011, pp. 138142.
11. Ming Jong Lin. Diagnosing Potentially Abnormal Attribute of Power Transformers Method, Journal of Engineering Research and Reports. Volume 22, Issue 7, Page 46-56, 2022.

COMPETING INTERESTS DISCLAIMER:
Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.