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

***Original Research Paper***

## ABSTRACT

|  |
| --- |
| This paper firstly explains the theory of regression analysis - model, least squares method, coefficient of determination, model assumptions, significance test, etc., and how to calculate through actual data to carry out what these essential nouns –such as $R^{2},SSR,SSE,SST,F,and t$ and derive relevant parameters to being analysis and interpretation. Secondly, understand what those parameters are, the author took some data from references to calculate. The results are presented in reports through the EXCEL application software SPSS system to clearly, the same time it provides analysis and interpretation. Finally, the combustible gases in the insulating oil of power transformers - hydrogen (H2), methane (CH4), ethane (C2H2), ethylene (C2H4), acetylene (C2H6) 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 The company was selected, and a regression analysis was performed to understand the relationship between the various gases. The study found that C2H4 had no effect on the increase or decrease of the amount of H2 gas. This paper attempts 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 a 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**

The regression method is a tool for analyzing the relationship between variables. It mainly looks at the linear relationship between independent variables (x) and dependent variables (y). Researchers can infer and forecast the dependent variables (y) of interest by building regression models. This paper discusses estimating the regression equation (𝒚̂𝒊 = 𝒃𝟎 + 𝒃𝟏 × 𝒙𝒊 + 𝜺𝒊), how to find the intercept 𝒃𝟎 and slope 𝒃𝟏 values from the data of the independent variable using the least squares method (SSE), and how to calculate the total sum of squares (SST) and the regression sum of squares (SSR). It also explains how to compute the coefficient of determination 𝑹𝟐 = 𝑺𝑺𝑹⁄𝑺𝑺𝑻, perform significance tests, make judgments, and finally calculate 𝜺𝒊, along with other steps.

In addition, regression analysis of the relationship between the combustible gases in transformer insulating oil - hydrogen (H2), methane (CH4), ethane (C2H2), ethylene (C2H4), acetylene (C2H6) and carbon monoxide (CO) has always puzzled me until I understood the theory of regression analysis, read the literature and used the SPSSsystem in the EXCEL application software and online regression calculator. In actual operation, the combustible gas detection data from 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 least squares method, applied by Legendre in 1805 and Gauss in 1809. It evolved into a statistical technique for analyzing data to understand the direction and strength of the correlation between two or more variables, as well as the amount of change in the dependent variable when the independent variable changes. We selected papers related to this topic from numerous sources in the literature. We summarized their characteristics as follows: [1] To investigate the correlation between two variables, one serves as the independent variable (X) or predictor, while the other is the dependent variable (Y) or outcome. We also used simple examples to calculate the correlation coefficient from the equation. [2][3] We described the meaning of each correlation coefficient in simple linear and multivariate regression analysis, along with the procedures and standards for significance testing. The process involves performing regression analysis calculations on practical cases. [4] We explored the relationship between multiple independent variables and a dependent variable, explained how to use this method with examples, and clarified the step-by-step process. [5] The R-squared and adjusted R-squared of regression analysis, along with the basis for determining whether the hypothesis test falls within the rejection region, are explained. The significance level of 0.05 is emphasized as an important threshold. [6] We described the estimated standard error, derived from the least squares regression line, which reflects the measured prediction error. A larger value indicates greater error and lower accuracy, while a smaller value suggests higher accuracy. [7] The relationship between the P-value and the null hypothesis (𝑯𝟎) or the alternative hypothesis (𝑯𝟏) and how to calculate it is explained. [8] Instructions for operating and interpreting the online regression analysis calculator. [9][10] Explain the combustible gases produced by decomposing insulating oil, and note that their content is used to assess the transformer's internal condition. [11] List relevant cases from previous years as supporting evidence.

**3. METHODOLOGY**

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):

$y\_{i}=β\_{0}+β\_{1}×x\_{1I}+β\_{2}×x\_{2i}+ε\_{i} (1)$

Convert to an estimated formula:

$$\hat{y}\_{i}=b\_{0}+b\_{1}×x\_{1I}+b\_{2}×x\_{2i}+ε\_{i} (2) $$

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-squared 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 to determine whether the coefficients of all independent variables $b\_{i} $ are 0. It has predictive power only when the coefficients are not zero (0).

Null hypothesis

$$H\_{0}=b\_{1},b\_{2},…b\_{n}=0$$

Alternative hypothesis

$$H\_{1}=b\_{1},b\_{2},…b\_{n}\ne 0$$

Statistics

$F={MSR}/{MSE}$ (3)

3.2.2 Marginal test of individual regression coefficients (t test): After confirming that the significance p value is less than 0.05, a marginal test is conducted to determine whether the 𝒃𝒊 coefficient of the individual independent variable equals 0. When the coefficient is not 0, the independent variable has explanatory power.

Null hypothesis:

$$H\_{0}=b\_{1},b\_{2},…b\_{n}=0 (i=1,2,…n)$$

Alternative hypothesis:

$$H\_{1}=b\_{1},b\_{2},…b\_{n}\ne 0 \left(i=1,2,…n\right)$$

Statistics:

$t= {b\_{1}}/{S\_{b1}}$ (4)

$S\_{b1}={S}/{\sqrt{\sum\_{i=1}^{n}x\_{i}^{2}-{\left(\sum\_{i=1}^{n}x\_{i}\right)^{2}}/{n}}}$(5)

3.2.3 $ 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 $ R^{2}$, SSR, SSE, SST, etc. are as follows.

$R^{2}=1-{SSE}/{SST}$ (6)

$SSR=\sum\_{i=1}^{n}\left(\hat{y}\_{i}-\overbar{y}\right)^{2}$ (7)

$SSE=\sum\_{i=1}^{n}\left(y\_{i}-\hat{y}\_{i}\right)^{2}$ (8)

$SST=\sum\_{i=1}^{n}\left(y\_{i}-\overbar{y}\right)^{2}$(9)

$SST=SSR+SSE$ (10)

Therefore, R² is not an objective indicator. The number of variables must also be taken into consideration. The adjusted $R^{2}$ can be regarded as an unbiased estimate of $R^{2}$ and is expressed as $\overbar{R}^{2}$. The calculation is as follows:

$\overbar{R}^{2}=1-({{SSE}/{n-k-1})}/{({SST}/{n-1)}}$(11)

Where n is the number of samples and k is the number of variables. From the above, we can see that $\overbar{R}^{2}$ (Adjusted R²) is less than R².

**4. DISCUSSION**

The size of the $R^{2}$ value must be consistent with rejecting the null hypothesis ($H\_{1}:b\_{1}\ne 0)$ and its significance (p value) must be lower than 0.05 to have substantial explanatory power. As for $\overbar{R}^{2}$, it is the data that has been appropriately adjusted ($\overbar{R}^{2}<R^{2})$. The t-test and F-test statistics must be consistent with rejecting the null hypothesis ($H\_{1}:b\_{1}\ne 0)$ 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 $(df\_{r})$ is 1 for a group of independent variables, the degrees of freedom of the error term $(df\_{e})$ is n-p-1, and the degrees of freedom of the total variation term $(df\_{t})$ 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=${1}/{1-R^{2}}$. 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 (C2H4.C2H2.CH4.C2H6.CO) and the change in the amount of hydrogen (H2).

**5. 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 C2H4 had no influence on the increase or decrease of H2 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.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc. have been used during the writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.

2.

3.

**REFERENCES**

1. Silvia Valcheva. Simple Linear Regression Examples. <https://www.intellspot.com/Linear-Regression-Examples>. Jun 5 2025. On website.
2. Anonymous, 2025. Simple Linear Regression Analysis for class handouts, Kaohsiung University of Science and Technology. <https://www2.nkust.edu.tw/~tsungo/Publish/14%20Simple%20linear%20regression%20analysis%20powerpoint.pdfwww.nkust.edu.tw>. 5 30 2025
3. Multiple regression analysis for class handouts, Kaohsiung University of Science and Technology.

<https://www2.nkust.edu.tw/~tsungo/Publish/15%20Multiple%20regression%20analysis%20powerpoint.pdf#:~:text>. 6 3 2025

1. Multiple linear regression analysis of handouts from consulting company, Yongxi Statistics Consulting Consultant.

https://www.yongxi-stat.com/multiple-regression-analysis/ 5 30 2025.on website.

1. Qiu Bingcheng, data analysis, Medium Statistics in Carrot Cheng in Chinese，on Mar 5, 2022.

<https://medium.com/qiubingcheng/> 5 30 2025.

1. Estimate the standard error of the measurement prediction error in Chinese.

[https://drfishstats.com/regression/standard- error-of-estimate/](https://drfishstats.com/regression/standard-%20%20%20error-of-estimate/) 5 31 2025.

1. Tutorial on calculating p-value in a spreadsheet, KDAN OFFIC from consulting company,

<https://kdan-office.kdandoc.com/> 5 30 2025.

1. Multiple Linear Regression Calculator, Statistic Kingdom from consulting company,

<https://www.statskingdom.com/multi_linear_regression.him/> 6 6 2025.

1. 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.
2. 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.
3. 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.