POLISH HYPERBARIC RESEARCH 2(95)2026
Journal of Polish Hyperbaric Medicine and Technology Society

ASSESSMENT OF THE CONTENT OF THE MAIN COMPONENTS OF BREATHING MIX. PART I

Ryszard Kłos

Department of Underwater Work Technology, Naval Academy in Gdynia

ABSTRACT

In quantitative instrumental analysis, measurement results are most often obtained in the form of relative analytical signals, which require in the calibration process1 reference to measurement standards[2] in order to determine the value of the measured quantity[3] with the accuracy and precision needed. The measurement results are influenced by a number of interfering factors that are difficult to define unambiguously and simultaneously[4] and which shape the magnitude of systematic errors. These are mitigated by carrying out measurements on test samples and reference samples under conditions that are as close to identical as possible, which allows for the determination of appropriate corrections to the measurement results obtained from the conversion of relative analytical signals. The measurement process, which takes such corrections into account, is considered to be under statistical control, i.e. on the assumption that random components are the dominant component of measurement uncertainty.
The articles present a method for estimating measurement uncertainty based on measurement standards, used in instrumental analysis, which is under statistical control of the process. It is illustrated by the determination of the content of the main components of breathing mixtures in diving with the separation of components by gas chromatography using a thermal conductivity detector[5].
Keywords: decompression, diving safety, underwater works, decompression tables, dive computers.

ARTICLE INFO

PolHypRes 2026 Vol. 95 Issue 2 pp. 43 – 100
ISSN: 1734-7009 **eISSN:** 2084-0535
DOI: 10.2478/phr-2026-0009
Pages: 38, figures: 0, tables: 2
page www of the periodical: www.phr.net.pl

Publisher
Polish Hyperbaric Medicine and Technology Society

Original article

Submission date: 14.02.2026 r.
Acceptance for print: 25.02.2026 r.

INTRODUCTION

From the point of view of ensuring diver safety, the composition of the breathing gas should be determined together with its adequate uncertainty[6], allowing the certifying body to approve the breathing gas for use during a defined diving process.

The basis for this study is my doctoral dissertation [1]. Although it was written many years ago, it has not lost its relevance in terms of the approach to measuring the composition of breathing mix in diving technology.

The approach described here is still significantly narrowed in many respects in relation to existing needs. The focus here is on estimating confidence intervals when determining the quantitative content of basic components of breathing mix for divers using instrumental methods, primarily gas chromatography separation using a thermoconductometric TDC sensor. The linear regression approach is general in nature and can be used in other instrumental methods, such as spectroscopy [2,3].

Methods for eliminating gross and systematic errors have not been taken into account here, assuming that during the implementation of the measurement method, every effort was made to ensure that the measurement process was under statistical control, hence confidence intervals are estimated only for random errors [4].

There exist quite a few textbooks that explain how to calculate uncertainty using instrumental analysis methods. A comprehensive work devoted to the processing of measurement results using physicochemical instrumental analysis methods is worth recommending [5]. Most textbooks do not describe the derivations of the models used, which may raise objections as to their reliability, as editorial errors should always be taken into account. An example of this is the Polish translation of a monograph which contains a significant number of editorial errors [6]. Quite often, when analysing confidence intervals, the lack of information on the simplifications and assumptions underlying the derivations can lead to interpretation problems.

In the approach proposed in this paper, certain differences can be observed between the approach to uncertainty budgeting based on statistical error propagation and the metrological approach [7]. The proposed approach differs from that presented in monographs on metrology, because the functions that capture the metrological uncertainty budget usually take into account a significant number of parameters, hence the uncertainty value calculated using these methods seems to be overestimated. This is a consequence of the fact that this approach assumes the accumulation of errors owing to the fact that they all have their maximum estimate and the same signs[7]. The probability that errors arising from different parameters/arguments will have the same sign decreases rapidly as their number increases. A comparison of the key differences between the two approaches is summarised in Table 1.

INSTRUMENTAL ANALYSIS

The professional literature lists measuring devices recommended for determining the composition of breathing mixes in diving operations. However, it should be noted that diving conditions may vary and thus affect the requirements for measuring devices. It follows that, before applying the relevant methodologies—such as verification, adjustment, calibration, analytical procedures, etc.—their suitability must be assessed [8].

In analytical measurements, instrumental analysis methods are of the greatest importance. Their use shortens the time needed to perform a single analysis, which means that it can be repeated multiple times or even used to monitor the composition of the breathing mix. This allows for statistical validation of the measurements and estimation of the confidence intervals of the analyses [9].

Tab. 1

Comparison of maximum error calculation with statistical methods [5].

	Metrological error calculation	Statistical methods
range	assessment of random error of a single measurement and systematic errors for indirect measurements	assessment of random errors for direct and indirect errors for counts $n > 1$
absolute error	additive, named upper bound Δ for absolute error: $x \pm \Delta x$	additive, dimensionless (unnamed) variance from the samples[2] and non-additive, named standard deviation of the mean values$s_{\bar{x}}$: $\bar{x} \pm t(\alpha, \nu) \cdot s_{\bar{x}}$
relative error	for the product additive upper bound of relative error $\delta = \frac{\Delta x}{\bar{x}}$	non-additive coefficient of variation $v = \frac{\hat{\sigma}}{\hat{\mu}}$
composite error	$\Delta f(x_1..x_n) = \sum_i \lvert\frac{\partial f}{\partial x_i}\rvert \cdot \Delta x_i$	$\sigma_f^2 = \sum_i (\frac{\partial f}{\partial x_i})^2 \cdot \sigma_i^2$

At present, classical chemical analysis methods are of very limited significance. However, it is commonly believed that they are just as accurate as instrumental chemical analysis methods. This belief is generally mistaken and stems solely from the inability to account for all errors in classical chemical analysis methods. However, it should not be assumed that instrumental chemical analysis methods are always superior simply because they often employ technologically advanced equipment. Most instrumental analysis methods continue to require instrument calibration using standards. For precise or critical measurements, this procedure is required prior to each measurement. It can be assumed that in quantitative instrumental analysis, almost all instruments require calibration using special standards. This is because the measurement result consists of relative analytical signals, which must be related to standards that reproduce absolute physicochemical values. Moreover, the values of the measured quantities depend on many elusive factors that influence or interfere with the analytical signals. The effect of these factors can be mitigated by performing a measurement under the same conditions as on the reference sample, thereby estimating the appropriate corrections.

MASS SPECTROSCOPY

Mass spectroscopy MS enables the characterisation of the mass of ions and ionised molecules based on the deflection of their trajectories in magnetic and electric fields. [10]. Mass spectroscopy is one of the most sensitive methods of instrumental analysis, providing invaluable services, particularly in the qualitative analysis of organic compounds. It is a destructive method, but requires extremely small samples. For quantitative measurements, mass spectroscopy requires calibration with spectrally pure reference substances. MS can only be used to analyse gaseous samples or samples that can be converted into a highly rarefied vapour, usually in the range from 0.01 mPa to100 mPa. The analytical result is a relative mass spectrum, the interpretation of which is not straightforward, and currently the processing of the spectrum is based entirely on libraries and computer software.

The mass spectrum is characteristic of a given chemical compound. The intensities of the individual ion beams, whether obtained from pure substances or mixtures, depend on a number of factors, primarily the method of ionisation. To obtain reproducible results, the spectra of the reference and analysed substances must be recorded under the same conditions. The calculation of the composition of mixtures involves solving a system of equations for selected bands.

The information obtained from the spectrum MS is not unambiguous. This is caused by multiple ionisation of molecules, which manifests itself in the fact that, in addition to the band for ions with the expected mass-to-charge ratio m/e, bands are also observed in other places: m/2e ,m/3e, etc. Another reason for the complexity of the spectra is the fact that during the decay of different substances, ions with the same mass-to-charge ratio may be formed, meaning that the bands will overlap. Even the spectrum of a mixture of simple substances such as: : CO_2, CO, O_2, N_2, shows strong band overlap [11].To a large extent, the inconvenience in interpreting spectra is eliminated in MS by coupling a mass spectrometer with a gas chromatograph GC. The separated bands of individual components in the GC chromatographic process are identified using a mass spectrometer MS. Coupling a gas chromatograph with a mass spectrometer GC/MS seems to be one of the best solutions for laboratory

measuring stations used to determine contaminants in breathing mix. This method can also be used to determine the composition of reference mixtures for other analytical methods.

INFRARED SPECTROSCOPY

Infrared spectroscopy IR is a recognised and valued method of instrumental analysis used for qualitative determination of organic compounds. Infrared spectra of molecules show slight broadening of absorption bands in the gas phase [12]. This method can be used to detect trace amounts of organic compounds. Gases such as: CO_2 and CO, produce clear and well-distinguishable absorption bands, which only partially overlap. In the wave number range from 800 cm^{-1} do 2000 cm^{-1} the spectrum is clear. This is the so-called "fingerprint" range, in which the characteristic absorption bands of many organic compounds occur. Just as fingerprints distinguish people, the bands in this range allow us to distinguish between chemical compounds, provided we can interpret them correctly, which is not a simple task. Similarly, in the wavelength range around 3000 cm^{-1} the spectrum is clear. Within this range, the absorption bands of hydrocarbons occur.

In the wave number ranges from 800 cm^{-1} to 2000 cm^{-1} and around 3000 cm^{-1}, organic compounds that may be present in the breathing atmosphere as metabolic products show high IR absorption activity. They can be identified together with typical pollutants such as CO_2 i CO.

Quantitative measurements using this method cannot always be performed with sufficient accuracy and reproducibility due to significant blurring of the IR bands, their overlapping in the case of mixtures, the inability to separate and integrate them accurately, the curvature of the signal-concentration[8] relationship characteristics, cumbersome calibration procedures, etc. [13,14,15]. Due to the widespread use of IR spectroscopy methods, the equipment is available in a specialised form for measurements in gas mixtures. With the help of IR spectroscopy, it is possible to measure the composition of breathing mixtures during the flow of the tested medium. This method cannot be used to determine homonuclear and atomic gases, therefore it is impossible to determine the argon Ar content with its use. Apart from that, the rest of the typical impurities can be determined [16,17,18]. The water vapour content causes significant difficulties in IR measurements owing to the very broad absorption bands of this compound.

GAS CHROMATOGRAPHY

Gas chromatography GC involves the separation of the components of the analysed sample as a result of specific adsorption on the chromatographic column filling and then their determination using a detector. The separation of the sample components is achieved by changing the operating parameters of the chromatographic[9] column. Adequate reproducibility and repeatability of measurements is achieved by selecting the detector and its[10] operating conditions. Descriptions of fillings can be found in the catalogues of companies that manufacture them and in chromatography journals[11]. The basics of the use of fillings and detectors in gas analysis can be found in monographic publications [19,20].

Using GC, it is possible to separate almost all vapours and gases that may contaminate the breathing gas [16].

ELECTROCHEMICAL ANALYSERS

Electrochemical methods are often used to measure the content of contaminants in gas mixtures. The principle of measurement is based on the penetration of gas particles through a selective or almost selective barrier into the interior of an electrochemical cell, where they dissolve in the electrolyte[12].

As a result of a chemical reaction with the electrolyte solution, ions are formed, which undergo electrochemical reactions on the electrodes of the cell. The electrode reactions lead to the formation of a current signal, which, after electronic processing, constitutes an analytical signal.

These methods require calibration. The use of ion-selective electrodes, although less accurate than spectroscopic methods, has the advantage that the measurement is relatively simple and can be carried out in field conditions [21]. Ion-selective electrodes can be used to measure the content of CO_2 i CO, NH_3, NO_x, CH_4 and pairs of other hydrocarbons. Hydrocarbon and CO analysers show low selectivity in relation to the compounds being measured, which means that the measurement may be disturbed by other components of the tested breathing atmosphere. The selection of appropriate electrodes is troublesome, thus it should be performed by a specialist analytical chemist for each analytical problem.

CHEMICAL METHODS

At present, the only chemical method commonly used to measure the content of pollutants in breathing mixtures is indicator tubes, e.g. [22]. The principle of measurement using indicator tubes is based on a characteristic chemical reaction with appropriate chemical reagents applied to the filling in a transparent tube, which is most often silica gel. The reaction takes place along the tube in accordance with the flow of the gas medium. By measuring the distance of the reaction front per unit of gas passed, it is possible to roughly estimate the concentration of the chemical compound being measured. The flow can be forced by: a bellows pump with a known volume, pressure reducing devices under, for example, rotametric flow control, various types of constant flow pumps or diffusion pumps. The analytical tubes are sealed at both ends. The tube must be unblocked before measurement. Most often, glass tubes are sealed at the ends, and unblocking them involves breaking off the ends of the tube. Analytical reagents are applied to a carrier or are contained in internal ampoules. Before or after adsorption of the contaminant, these ampoules should be broken and saturate the filling with the poured reagent. The ampoules are broken with a piercer or by breaking the inner tube[16]. The indicator tubes have an appropriate scale and are supplied with instructions for measuring and estimating the results. An important part of the instructions is to provide analytical reactions and a list of substances that may interfere with the measurement. If measuring tubes are used as indicators for a given measurement issue for the first time, the applicability of a specific tube should be analysed. Measuring tubes for a selected type of chemical substance can be made for

different measurement ranges and for different analytical reactions. It is best if the type of analytical tube is selected by a specialist analytical chemist. Since there are different analytical situations, it is very important to select the appropriate tube and procedure for its use, e.g. the use of tubes for pre-conditioning or protective tubes.

Measuring methods using indicator tubes as very simple methods have found wide application as threshold indicators for swift, approximate assessment of contaminant content.

SPECIAL AND SPECIALISED ANALYSERS

Some chemical compounds display specific properties. These properties form the basis for the development of special analysers. Such analysers are often highly resistant to interference, which means that chemical compounds with these specific properties can be determined with good repeatability and reproducibility, regardless of the other components of the analysed sample. Nevertheless, before using such an analyser, it is always necessary to check for the possibility of measurement interference. An example of this is the magneto-thermo-conductometric method of measuring oxygen content, which utilises the paramagnetic properties of oxygen. This method is used with good results for nitrogen-oxygen and helium-oxygen mixtures.

Often, universal measurement methods are narrowed down to specific measurements. This is how specialised measuring devices are created. As a rule, the use of universal laboratory measurement methods is not possible in field conditions. The creation of specialised devices reduces their capabilities and, as a rule, their accuracy, but allows measurements to be taken in field conditions. Typical specialised analysers include gas chromatographs used to measure exhaust gas composition. Other examples include specialised IR spectrometers for measuring CO_2 emissions or specialised thermoconductometers for measuring hydrogen emissions.

CONCLUSION

It would appear that the widespread adoption of instrumental analysis methods in diving technology provides sufficient control over the composition of the breathing gas, subject to their correct metrological operation. Unfortunately, even minor oversights in this regard can lead to significant errors. Regrettably, a high degree of reliance on instrumental analysis methods often goes hand in hand with a relaxation of rigorous metrological supervision, on account of the costs involved and the time-consuming nature of the process. This paper aims to make users of instrumental analysis methods aware that such methods continue to rely on calibration; otherwise, they may lead to serious errors and endanger the health of divers.

1. LINEAR REGRESSION

Regression analysis is one of the most commonly used models for studying covariation. This model defines the influence of independent variables[17] $\forall_{i\in[1;n]}\ x_i : y = f(x_1..x_n)$ on the dependent variable[18] y. According to the paradigm of stochastic determinism[19] under observation conditions, the regression function is disturbed by a confounding factor ϵ. Thus, the regression model can be written down as an equation:

$$y = f(x_1..x_n) + \epsilon \tag{1.1}$$

where: $y-$ dependent variable, $f-$ regression function, x_i- independent variable, $i \in [1;n]$, $\epsilon-$ confounding factor.

Often, the regression model used will take the form of a linear combination:

$$y = a_0 + a_1 \cdot x_1 + .. + a_n \cdot x_n \tag{1.2}$$

where: a_i – linear regression coefficients[20], $i \in [0;n]$

The confounding factor ϵ is treated in (1.1) as a random variable with a standard normal distribution[21] $\mathbb{N}(\bar{x}, \sigma)$ with a mean of zero $\bar{x} = 0$, and a standard deviation $\sigma = 1$. It is important that this assumption is validated when analysing the problem situation.

REGRESSION

The word *regression*[22] has many meanings in different fields of knowledge. It was introduced into statistics by F. Galton and used in biological observations to describe changes in green pea populations[23] [23].

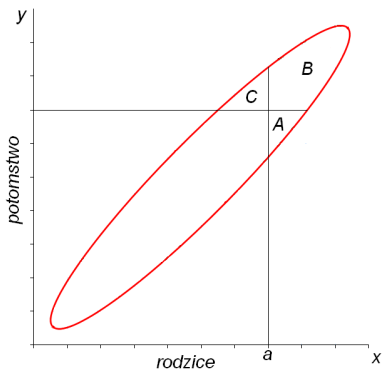


Fig. 1.1 Concentration ellipse for linear regression observed in a green pea population.

The average value of the pea radius x for offspring shows regression[24], which can be modelled by a straight line. Of course, this trend applies to the mean value $\bar{x}$,, as individual seeds show an almost constant spread around this value $\bar{x}$ compensating for the observed regression[25]. This phenomenon can be traced using the example of a concentration ellipse – Fig. 1.1 (24).

In Fig. 1.1, for the diameter of a $x \geq a$ pola, the fields A and B represent large parents, for which only the fraction B represents equally large offspring.

The fraction C represents large offspring that are the generation of not quite large parents. The relatively large offspring represented by the field B+C is approximately the same as the large parents represented by the field A+B. It can be seen from this that the natural spread compensates for the observed regression of the mean value $\bar{x}$ so that the widespread so-called "ferality" of green pea crops is not observed.

Research on the green pea population contributed to the establishment of the concept of linear regression in statistics, which is conceptually distant from its original meaning.

LINEARISATION

An important feature of the linear regression function is that it is the sum of factors a_i multiplied by variables x_i,, where a_i are unknown coefficients, and x_i are the results of measurements of independent variables X. The origin of the values x_i is irrelevant. The factors a_i of linear regression for predicting[26] the values $\hat{y}$ are obtained by solving a system of linear equations for the observations made:

$$\forall_{j\in[1;m]\ \wedge\ i\in[1;n]} \begin{cases} y_1 = a_0 + a_1 \cdot x_{1,1} + .. + a_n \cdot x_{n,1} + \epsilon_1 \\ .. \\ y_m = a_0 + a_1 \cdot x_{1,m} + .. + a_n \cdot x_{n,m} + \epsilon_m \end{cases} \tag{1.3}$$

where: $y_j -$ is the realisation of the dependent variable $j \in [1;m]$, $f -$ is the regression function, $x_{i,j} -$ is the realisation of the independent variable for which $i \in [1;n]$,$\epsilon_j -$ is the confounding factor, assuming that $\epsilon_1 \ldots \epsilon_m$ are an independent random sample from the same general population with a standard normal distribution N(0,1).

A linear relationship does not have to be an adequate mathematical model of the phenomena occurring, but it is a useful form of approximation used in practice[27]. The assumptions made should be validated for the developed model:

- testing the differences between the values of the dependent variable calculated from measurements and predicted by the model,
- testing the differences between the full model and simpler linear combinations, which are usually preferred,
- testing the distribution of residual values, etc.

EXPECTED VALUE

Certain characteristic numerical values are associated with the distribution of a random variable X. These characteristics are called *the parameters of the distribution of a random variable* X. The most important parameters of random variables include: expected value EX and variance D^2X.

The *expected value* EX[28] is the number around which the values of the random variable X cluster when the experiment is repeated multiple times. The expected value of a random variable X with a discrete distribution $\{X \leftarrow x_i; P(x_i) |\ x_i \in \in S\}$ is the number $E\ X$, defined by the following formula[29]:

$$\forall_{x_i \in X} \quad EX = \sum_i x_i \cdot P(x_i) \big|\ x_i \in S \tag{1.4}$$

where; $EX -$ is the expected value, $X -$ is a random variable, $x_i -$ are the realizations of the random variable X , $P(x_i) -$ is the probability of the occurrence of the event x_i of the random variable $X\ : X \leftarrow x_i$, $S -$ is the set of results of the draws

The expected value has the following properties:

- $E(c) = c$
- $E(c \cdot X) = c \cdot EX$
- $E(I) = 1$ where I denotes the identity function 1
- if X and Y are defined in the same probability space $(\Omega, \mathcal{F}, P)$, then $E(X + Y) = EX + EY$

According to the above rules, the expected value for the linear function $Y = a \cdot X + b$ will be:

$$E(Y = a \cdot X + b) = a \cdot EX + b \tag{1.5}$$

The expected value EX of the random variable X with a zero-one distribution $X \in \{0; 1\}$, for which the occurrence of the event in question will be marked with number one $\{X \leftarrow x\} = 1$, and the absence of this event with a zero $\{X \leftarrow x\} = 0$ and the probability of the event occurring $\{X \leftarrow x = 1\}$ will be p, then according to relation (1.4), the expected value EX will be:

$$EX = 1 \cdot p + 0 \cdot (1 - p) = p \tag{1.6}$$

The expected value ES_N of the random variable S_N, which is the sum N of the random variables $S_N = X_1 + .. + X_N$, denoting the number of successes in N independent $Bernoulli$ trials, will be:

$$ES_N = E(X_1 + .. + X_N) = EX_1 + .. + EX_N = N \cdot P(x_i) \tag{1.7}$$

Where: $P(x_i)$ – is the probability of occurrence of event x_i of random variable X :$X \leftarrow x_i$ with distribution $X \leftarrow x_i \in \{0; 1\}$

VARIANCE

$Variance\ of\ a\ random\ variable$ is a measure of the dispersion of the values of the random variable X around the expected value EX, denoted by the symbol D^2X and defined by the formula:

$$D^2X \equiv E(X - EX)^2 = EX^2 - (EX)^2 \quad | \quad EX^2 = \sum x_i^2 \cdot P(x_i) \tag{1.8}$$

where: D^2X – variance of a random variable X

with properties[30]:

- $D^2c = 0$
- $D^2(c \cdot X) = c^2 \cdot D^2X$
- for independent random variables X and Y variance:
  $D^2(X + Y) = D^2X + D^2Y$

(1.9)

where: c – is a constant.

According to definition (1.8), the variance D^2X is a measure of the difference between the value of the random variable X and the expected value EX, and squaring this difference gives a non-negative number[31].

To calculate the variance D^2X for the $Bernoulli$ distribution, we can use the aforementioned property of the variance of the sum of independent random variables $D^2(X + Y) = D^2X + D^2Y$, which is equal to the sum of their variances:

$$D^2(X_1 + .. + X_N) = D^2X_1 + .. + D^2X_N \tag{1.10}$$

Performing calculations according to formula (1.8), similarly to relation (1.6), it can be shown as[32]:

$$\forall_{X \leftarrow x \in \{0;1\}}\ D^2X = EX^2 - (EX)^2 = [0^2 \cdot (1 - p) + 1^2 \cdot p] - p^2 = p \cdot (1 - p) \tag{1.10}$$

hence, the variance of the random variable$S_N = X_1 + .. + X_N$ with a $Bernoulli$ distribution will be:

$$D^2S_N = D^2(X_1 + .. + X_N) = D^2X_1 + .. + D^2X_N = p \cdot (1 - p) + .. + p \cdot (1 - p) = N \cdot p \cdot (1 - p) \tag{1.11}$$

The $Bernoulli$ distribution described here in more detail is important in the validation process, but it does not play a significant role in this study.

COVARIANCE

A generalisation of variance (1.8) to the two-dimensional case is *covariance,* which describes the co-variability of two features:

$$cov(X, Y) = \sum_i (x_i - EX) \cdot p_i(X) \cdot (y_i - EY) \cdot p_i(Y)\ cov(X, Y) = \frac{1}{n} \sum_i [(x_i - \bar{x}) \cdot (y_i - \bar{y})] \tag{1.12}$$

with the following properties:

- $cov(X, a) = 0 \ \ | \ a = const$
- $cov(a \cdot X, b \cdot Y) = a \cdot b \cdot cov(X, Y) \ \ | \ a, b = const$
- $cov(a + X, b + Y) = cov(X, Y)$
- $cov(a \cdot X + b \cdot Y, c \cdot W + d \cdot V) = a \cdot c \cdot cov(X, W) + a \cdot d \cdot cov(X, V) + +b \cdot c \cdot cov(Y, W) + b \cdot d \cdot cov(Y, V) \ | \ a, b, c, d = const$

$$cov(X, X) = var(X) \tag{1.13}$$

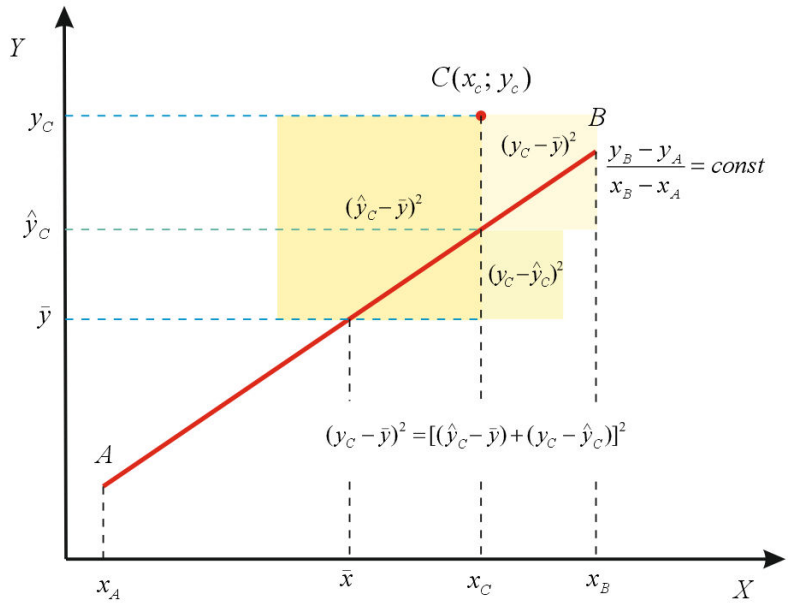


Figure 1.2 Variance decomposition and analysis for a single value and a linear regression model; where:(y_c-y¯)^2- total variance,(y^_C-y¯)^2- variance explained deterministically by the linear model. (y_C-y^_C)^2- unexplained residual variance.

Covariance is a measure of the simultaneous dispersion of two quantities from their mean values[34], hence:

VARIANCE ANALYSIS

Many statistical methods are based on the analysis of deviations[35] calculated from a sample. Figure 1.2 shows an example mathematical model of the phenomenon represented by a segment $\overline{AB}$ of a straight line described by the equation $\frac{y_B - y_A}{x_B - x_A} = const$, and the distribution of deviations from the model for the point $C(x_C;\ y_C)$.

The total observed variance $s = \frac{1}{N}\sum_i (y_i - \bar{y})^2$ consists of the variability explained by the linear model $s_r = \frac{1}{N}\sum_i (\hat{y}_i - \bar{y})^2$ and the variability unexplained[36] $s_\epsilon = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$: $s = s_r + s_\varepsilon$. After dividing both sides by the total variability s , we obtain:

$$\frac{s_r}{s} + \frac{s_\epsilon}{s} = 1 \Big|\ r^2 = \frac{s_r}{s};\ \ \varphi^2 = \frac{s_\epsilon}{s} \quad (1.14)$$

where: r^2 – coefficient of determination, φ^2 – coefficient of convergence,s – total variance, s_r – deterministic share in the observed variance explained by the model, s_ϵ – random share in the observed variance.

The coefficient of determination r^2 from equation (1.14) represents the fraction[37] of explained variance[38]. Its complement, the coefficient of convergence $\varphi^2 = 1 - r^2$, represents the fraction of variance not explained by the linear regression model [25].

VARIANCE FRACTIONS

For $j-$ multiple sampling of an n –element sample consisting of values $y_{n,j}$, which are the realisation of a random variable $Y = f(X)$:$y_{n,j} \leftarrow Y$, it is possible, in accordance with the second paradigm of physics dealing with the statistical analysis of averaged quantities, to assume as a mathematical model of the value of each element$y_{n,j}$, a linear combination:

$$y_{n,j} = \mu + A_{n,j} + \varepsilon_{n,j} \quad (1.15)$$

where: $y_{n,j}$ – is the result *n*–repetition of this measurement, μ – is the true value, $A_{n,j}$ – is the difference between the result $y_{n,j}$ and the true value μ ,$\varepsilon_{n,j}$ – is the factor related to the stochastic dispersion n – of this result in the jth this sample

For each sample with number $n \in [1,..,N]$, the mean $\bar{y}_j$ can be calculated. The mean calculated from all values $y_{n,j}$, denoted as $\bar{y}$, is considered a natural estimator of the true value $\hat{\mu}$:$\hat{\mu} \triangleq \bar{y}$, where the symbol$\triangleq$ should be equated with the term "corresponds to" or "is proportional to"[39].

The values $A_{n,j}$ of the difference between the result $y_{n,j}$ and the true value μ are considered here as a systematic component[40] and the dispersion $\varepsilon_{n,j}$ (as a $stochastic$ component[41] of the uncertainty of each dependent variable $y_{n,j} = f(x_j)$.

In accordance with the central limit theorem, it is most often assumed that the dispersion factor $\varepsilon_{n,j}$ has a normal distribution with variance σ^2 and an expected value equal to zero[42]: $\varepsilon_{n,j} \sim N(0, \sigma)$.

Without prejudice to the generality of the considerations, it is assumed here that for each $j-th$ sample of size n, the sum of the differences $S(A_{n,j})$ between each element of the sample $y_{n,j}$ and the actual value μ should be equal to zero: $\mu = 0$ with an accuracy of a constant[43]: $\forall_i\ S(A_i) = \sum_{n=1}^{N} A_{n,i} = 0$.

The estimator of the total systematic variance $\hat{A}$ can be the difference between the mean value $\bar{y}_j$ of the measurement results for *j-th* of this repetition of sampling with size n and the total mean value $\bar{y}$:$\hat{A} \triangleq \bar{y}_j - \bar{y}$. The natural unbiased estimator of the square of the total

systematic variance $\hat{A}^2$ will be: $\hat{A}^2 = \frac{1}{j-1} \cdot \sum_{j=1}^{I} (\bar{y}_j - \bar{y})^2$. The denominator contains the number of degrees of freedom $j-1$ due to the following relationship: $\bar{y} = \frac{1}{j} \sum_{j=1}^{I} \bar{y}_j$.

The sum of squares $S(A)$ of the estimators $\widehat{A}^2$ of deterministic variance after all J trials with accuracy to a constant can be written as[44]:

(1.16)

$$S(A) = \sum_{j=1}^{J} \sum_{j=1}^{N} (\bar{y}_j - \bar{y})^2 ,$$

where: $S(A)$ – sum of squares of systematic deviations.

Similarly, for the dispersion value ε, the sum of squares of statistical dispersion $S(\varepsilon)$ can be written as:

(1.17)

$$S(\varepsilon) = \sum_{j=1}^{J} \sum_{j=1}^{N} (y_{i,j} - \bar{y}_j)^2 ,$$

where: $S(\epsilon)$ – sum of squares of statistical dispersion.

The total value of the squares of deviations S can be calculated in the same way as before:

(1.18)

$$S = \sum_{i=1}^{I} \sum_{n=1}^{N} (y_{n,i} - \bar{y})^2$$

where: S – global sum of squares of deviations.

To show that the total value of deviations S from (1.18) is the sum of the values given by equations (1.16) and (1.17): $S = S(A) + S(\epsilon)$, we can extend the sum of squares of deviations S from relation (1.18) by the means $\bar{y}_i$ for individual samples, and then use the properties of the quadratic function:

(1.19)

$$\begin{aligned} S &= \sum_i \sum_n (y_{n,i} - \bar{y})^2 = \sum_j \sum_n (y_{n,i} - \bar{y}_i + \bar{y}_i - \bar{y})^2 = \\ &= \sum_i \sum_n [(y_{n,i} - \bar{y}_i)^2 + 2 \cdot (y_{n,i} - \bar{y}_i) \cdot (\bar{y}_i - \bar{y}) + (\bar{y}_i - \bar{y})^2] = \\ &= \sum_i \sum_n [(y_{n,i} - \bar{y}_i)^2 + \sum_i \sum_n (\bar{y}_i - \bar{y})^2 + 2 \cdot \sum_i \sum_n [(y_{n,i} - \bar{y}_i) \cdot (\bar{y}_i - \bar{y})] = \\ &= S(\epsilon) + S(A) + 2 \cdot \sum_i \sum_n [(y_{n,i} - \bar{y}_i) \cdot (\bar{y}_i - \bar{y})] \end{aligned}$$

The sums of the products $\sum_i \sum_n [(y_{n,i} - \bar{y}_i) \cdot (\bar{y}_i - \bar{y})]$ from formula (1.19) can be written as:

(1.20)

$$\begin{aligned} &\sum_i \sum_n [(y_{n,i} - \bar{y}_i) \cdot (\bar{y}_i - \bar{y})] = \\ &= (y_{1,1} - \bar{y}_1) \cdot (\bar{y}_1 - \bar{y}) + \cdots + (y_{N,1} - \bar{y}_1) \cdot (\bar{y}_1 - \bar{y}) + \\ &+ (y_{1,2} - \bar{y}_2) \cdot (\bar{y}_2 - \bar{y}) + \cdots + (y_{N,2} - \bar{y}_2) \cdot (\bar{y}_2 - \bar{y}) + \\ &\ldots \\ &+ (y_{1,I} - \bar{y}_i) \cdot (\bar{y}_I - \bar{y}) + \cdots + (y_{N,I} - \bar{y}_I) \cdot (\bar{y}_I - \bar{y}) = \\ &= N \cdot \sum_i \{(y_i - \bar{y}) \cdot [\sum_n (y_{n,i} - \bar{y}_i)]\} \end{aligned}$$

In relation (1.20), the average values $\bar{y}_i$ in the summation $\sum_n (y_{n,i} - \bar{y}_i)$ do not depend on the sample size N , therefore:

$$\sum_{n=1}^{N} (y_{ni} - \bar{y}_i) = \sum_{n=1}^{N} y_{n,i} - N \cdot \bar{y}_i$$

According to the definition of the mean value, we can write:

(1.22)

$$\forall_{N=const} \quad \bar{y}_j = \frac{1}{N} \cdot \sum_{n=1}^{N} y_{n,i} \quad \Rightarrow \quad \sum_{n=1}^{N} y_{n,i} = N \cdot \bar{y}_i$$

From relations (1.21) and (1.22) it follows that: $\sum_{n=1}^{N} (y_{n,i} - \bar{y}_i) = 0$. Therefore, the expression $\sum_i \sum_n [(y_{n,i} - \bar{y}_i) \cdot (\bar{y}_i - \bar{y})]$ from equation (1.20) is identically equal to zero:

Na podstawie zależności: (1.23) i (1.19) można stwierdzić, że:

(1.23)

$$\sum_i \sum_n [(y_{n,i} - \bar{y}_i) \cdot (\bar{y}_i - \bar{y})] = 0$$

(1.24)

$$S = S(A) + S(\epsilon) \quad \blacksquare$$

(1.17)

CONCLUSION

Equation (1.24) shows the sum of the squares of the deviations associated to the difference between the result and its actual value, which leads to the conclusion that the total variance of the regression S is the sum of the partial variances $S = S(A) + S(\epsilon)$. This property is often used to evaluate the maximum variance using regression as a reference curve.

1. LINEAR CORRELATION OF TWO VARIABLES

It often happens that the results of measurements of the dependence of the dependent variable Y on the independent variable X should theoretically lie along a straight line. Since the measurement results are subject to random errors, it is not possible to select the coefficients of the equation so that the straight line graph passes exactly through all the measurement points. Therefore, the coefficients of the equation $\{a; b\}$ should be selected so that the resulting straight line $y = a + b \cdot x$ describes the measurement results $(x_i; y_i)$ as accurately as possible. One of the methods for calculating the coefficients $\{a; b\}$ of a straight line equation $y = a + b \cdot x$ that meets this requirement is the least squares method. In this method, the value of the slope coefficient b and the intercept a in the equation of a straight line $y = a + b \cdot x$ are selected in accordance with $Gauss's$ postulate so that the sum of the squares of the deviations[i] is minimised. If we write the equation of a straight line in the form $y = a + b \cdot x$, the sum of the squares of the deviations S will be expressed by the formula:

(2.1)

$$S = \sum_i (y_i - a - b \cdot x_i)^2 \equiv \sum_i (y_i - \hat{y}_i)^2$$

where: S – is the sum of squares of deviations, a – is the intercept of the regression line , b – is the slope of the regression line, $\hat{y}_i$ – is the estimated value y_i for the observed value x_i for a linear regression.

LEAST SQUARES METHOD

The least squares method can be based on a standard search for the minimum of function (2.1) by calculating partial derivatives with respect to the slope coefficient b with a constant intercept a and with respect to the intercept a with a constant slope coefficient b, and setting them equal to zero:

(2.2)

$$\begin{cases} \left(\frac{\partial S}{\partial a}\right)_b = -2 \cdot \sum_i (y_i - a - b \cdot x_i) = 0 \\ \left(\frac{\partial S}{\partial b}\right)_a = -2 \cdot \sum_i x_i \cdot (y_i - a - b \cdot x_i) = 0 \end{cases}$$

By transforming the equation $-2 \cdot \sum_i (y_i - a - b \cdot x_i) = 0$, we obtain: $\sum_i y_i - -N \cdot a - b \cdot \sum_i x_i = 0 | n = \{1..N\}$. For the equation $-2 \cdot \sum_i x_i \cdot (y_i - a - -b \cdot x_i) = 0$, we obtain: $\sum_i x_i \cdot y_i - a \cdot \sum_i x_i - b \cdot \sum_i x_i^2 = 0$. For the value , which is the midpoint of the regression line, the regression equation can be written as: $\overline{y} = a + b \cdot \bar{x}$. From this, we can calculate the value of the slope coefficient $a = \bar{y} - b \cdot \bar{x}$ and transform the system of equations (2.2) to the form:

(2.3)

$$\begin{cases} \hat{a} = \bar{y} - \hat{b} \cdot \bar{x} \\ \sum_i x_i \cdot y_i - \hat{a} \cdot \sum_i x_i - \hat{b} \cdot \sum_i x_i^2 = 0 \end{cases}$$

where: $\bar{y}$ – is the arithmetic mean of the measurements of the values y_i, $\bar{x}$ – is the arithmetic mean of the measurements of the values x_i.

The parameters of the regression equation can be easily calculated from the system of equations (2.3). By solving the system of equations (2.3) by substituting the first equation into the second, we obtain: $\sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i = b \cdot (\sum_i x_i^2 - \bar{x} \cdot \sum_i x_i)$. Thus, the estimates of the coefficients of the regression line $\{\hat{a}; \hat{b}\}$ can be written as:

(2.4)

$$\forall_{y=\hat{a}+\hat{b}\cdot x} \begin{cases} \hat{a} = \bar{y} - \hat{b} \cdot \bar{x} \\ \hat{b} = \frac{\sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i}{\sum_i x_i^2 - \bar{x} \cdot \sum_i x_i} \end{cases} \Bigg| \; \bar{x} = \frac{1}{N} \cdot \sum_i x_i \wedge \bar{y} = \frac{1}{N} \cdot \sum_i y_i ,$$

where: N – is the number of points for determining the regression equation, $\hat{a}$ – is the estimate of the intercept of the simple regression equation, $\hat{b}$ – is the estimate of the slope of the simple regression equation.

The estimates (2.4) are the best unbiased estimators of the linear regression parameters[46]: $E(a) = \hat{a}$ and $E(b) = \hat{b}$ (26) .

The system of equations (2.4) shows that for all values known from measurements[47] $\sum_i x_i \cdot y_i$,$\sum_i x_i$,$\sum_i x_i^2$,$\bar{y}$ and $\bar{x}$, the equation of the regression line can be estimated from an N – element sample: $\hat{y} = \hat{a} + \hat{b} \cdot x$.

The equation of a straight line regression can sometimes be used even when the graph in the Cartesian coordinate system is non-linear, but it can be linear in the semi-logarithmic system[48] $\log \hat{Y} = \log a + b \cdot X$ or logarithmic[49] $\log \hat{Y} = \log a + b \cdot \log X$.

In the system of equations (2.4), the denominator $\sum_i x_i^2 - \bar{x} \cdot \sum_i x_i$ of the estimation of the directional coefficient $\hat{b}$ is the sum of squares S_{XX} for independent variables x_i:

(2.5)

$$S_{XX} = \sum_i x_i^2 - \bar{x} \cdot \sum_i x_i = \sum_i (x_i - \bar{x})^2 ..$$

Using relations (2.5) and (2.6), the estimation of the directional coefficient from the system of equations (2.4) can be written in a simpler form:

(2.6)

$$S_{XY} = \sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i = \sum_i y_i \cdot (x_i - \bar{x}).$$

Using relations (2.5) and (2.6), the estimation of the directional coefficient from the system of equations (2.4) can be written in a simpler form:

(2.7)

$$\hat{b} = \frac{S_{XY}}{S_{XX}} = \sum_i c_i \cdot y_i \quad \Big| \quad c_i = \frac{(x_i - \bar{x})}{S_{XX}} .$$

For a straight line regression passing through the origin of the coordinate system, equation (2.2) takes the form: $\frac{\partial S}{\partial b} = -2 \cdot \sum_i x_i \cdot (y_i - b \cdot x_i) = 0$, hence the slope coefficient for such a straight line regression can be written as:

(2.8)

2.3)

$$\forall_{y=\hat{b}\cdot x} \quad \hat{b} = \frac{\sum_i x_i \cdot y_i}{\sum_i x_i^2} .$$

F-TEST

The distribution of the random variable F is used to compare population parameters. The shape of its probability density distribution is similar to the distribution χ^2 and is called the F–Snedecor distribution (27) . The test value F is obtained by constructing the ratio of two values χ_1 and χ_2 and the corresponding degrees of freedom ν_1 and ν_2: $\forall_{\nu_1 > \nu_2} F(\alpha, \nu_1, \nu_2) = \frac{\chi_1}{\nu_1} / \frac{\chi_2}{\nu_2}$, where α represents the significance level of inference[ii] and the variables (χ_1, χ_2) have the distribution χ^2. Note that the order of the values in the formulated proportion is important. The mean value μ for the distribution F is $\forall_{\nu_1 > 2}\ \mu = \frac{\nu_2 \nu_2}{\nu_2 - 2}$ and the variance $\forall_{\nu_1 > 4}\ \sigma^2 = (\frac{\nu_2}{\nu_2 - 2})^2 \cdot \frac{2 \cdot (\nu_1 + \nu_2 - 2)}{\nu_1 \cdot (\nu_2 - 4)}$. The density of the distribution L does not depend on the variance σ^2, but is a function of the number of degrees of freedom ν for the distributions of the variable χ and reaches its maximum at the point $L_{max} = \frac{\nu_1 - 2}{\nu_1} \cdot \frac{\nu_2}{\nu_2 + 2}$.

2.4)

Fisher's statistical test F will be based on F – Snedecor distribution to assess the difference between the variances between the regression equation $\hat{y} = a + b \cdot \hat{x}$ and passing through the origin of the coordinate system $\hat{y} = b \cdot \hat{x}$. The test proportion will be the value:
where: $\hat{\sigma}^2_{y=b\cdot x}$ – variance for the simplified equation $y = \hat{b} \cdot x$,$\hat{\sigma}^2_{y=a+b\cdot x}$ – variance for the full equation $y = \hat{a} + \hat{b} \cdot x$,

(2.9)

$$\forall\ F = \frac{\hat{\sigma}^2_{y=b\cdot x}}{\hat{\sigma}^2_{y=a+b\cdot x}} \hat{\sigma}_1^2 > \hat{\sigma}_2^2$$

where: $\hat{\sigma}^2_{y=b\cdot x}$ – variance for the simplified equation $y = \hat{b} \cdot x$, $\hat{\sigma}^2_{y=a+b\cdot x}$ – variance for the full equation $y = \hat{a} + \hat{b} \cdot x$,

which should be compared with the critical value $F_{kr}(\bar{P}, \nu_1 = n-1, \nu_2 = n-2)$ of the F-Snedecor distribution. If $F_{kr} \geq F$, then it should be assumed that the variances of both equations do not differ from a statistical point of view, hence the simplified regression equation $y = b \cdot x$ can be adopted, which is preferred from a theoretical point of view (5).

Example 1.

For the results of measurements of the field under the bands, for different volumes of injected oxygen,

To calculate the residual variance, the relationship (3.7) derived in the next chapter was used: $\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{\nu}$. Hence, the test value F will be: $\forall_{\hat{\sigma}_1^2 > \hat{\sigma}_2^2}\ F = \frac{\hat{\sigma}_1^2(y=b\cdot x)}{\hat{\sigma}_2^2(y=a+b\cdot x)}$.

F-Fishe inference based on F-Snedecor distribut n shows that the following inequality holds: $F_{kr} \equiv F(P, \nu_1, \nu_2) \geq F$. It can therefore be assumed that the variances of both linear regression equations: $y = b \cdot x$ and $y = a + b \cdot x$ do not differ from a statistical point of view. Hence, as a model of observed variability, a simplified regression equation $y = b \cdot x$ can be adopted, which is preferred from a theoretical point of view.

The results of the calculations are summarised in the table below:

$$S_{XX} = \sum_i x_i^2 - \bar{x} \cdot \sum_i x_i = \sum_i (x_i - \bar{x})^2$$
$$S_{XY} = \sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i = \sum_i y_i \cdot (x_i - \bar{x})$$
$$\hat{\sigma}^2 = \frac{S_{YY} - \hat{b} \cdot S_{XY}}{\nu}$$

$$\forall_{\hat{\sigma}_1^2 > \hat{\sigma}_2^2}\ F = \frac{\hat{\sigma}_1^2(y=b\cdot x)}{\hat{\sigma}_2^2(y=a+b\cdot x)}$$

$$F_{kr}(P, \nu_1 = n-1, \nu_2 = n-2) \geq F \rightarrow y = \hat{b} \cdot x$$

x_i	y_i	x_i	x_i^2	y_i	y_i^2	$x_i \cdot y_i$
0.005	0.07968	0.005	0.000	0.080	0.006	0.000
0.010	0.17585	0.010	0.000	0.176	0.031	0.002
0.015	0.26242	0.015	0.000	0.262	0.069	0.004
0.020	0.34459	0.020	0.000	0.345	0.119	0.007
0.025	0.42393	0.025	0.001	0.424	0.180	0.011
0.040	0.66741	0.040	0.002	0.667	0.445	0.027
0.050	0.83517	0.050	0.003	0.835	0.698	0.042
0.080	1.20019	0.080	0.006	1.200	1.440	0.096
0.100	1.50058	0.100	0.010	1.501	2.252	0.150
0.150	2.09007	0.150	0.023	2.090	4.368	0.314
0.200	2.78012	0.200	0.040	2.780	7.729	0.556
0.250	3.31059	0.250	0.063	3.311	10.960	0.828
0.300	3.87806	0.300	0.090	3.878	15.039	1.163
sumy		1.245	0.237	17.549	43.337	3.199
średnia	0.10	1.35				

liczba punktów pomiarowych	$n =$	13	
równanie regresji		$y=a+b\cdot x$	$y=b\cdot x$
współczynnik kierunkowy prostej regresji	$\hat{b} =$	12.9043	13.5038
wyraz wolny prostej regresji	$\hat{a} =$	0.1141	
wariancja pomiędzy wartościami zmierzonymi (y_i) i obliczonymi (Y_i)	$\hat{\sigma}^2 =$	0.00526	0.00482
poziom ufności	$P =$	0.95	
wybór równania prostej regresji przy pomocy testu F dla wariancji	$F =$	0.9167	
wartość krytyczna testu F	$F(P, \nu_1, \nu_2) =$	2.7876	
decyzja		0	1

COEFFICIENT OF DETERMINATION

The coefficient of determination r^2 corresponds numerically to the value of the correlation coefficient r after squaring it. The greater the ratio of the variance from the sample for residual values $s_\epsilon = \frac{1}{N} \cdot \sum_i (y_i - \hat{y}_i)^2$ around the regression line to the total variance[51] from the sample $s = \frac{1}{N} \cdot \sum_i (y_i - \bar{y})^2$, the better the quality *of the prediction*[52] $: r^2 = \frac{s_\epsilon}{s} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$ – Figure 1.2. In the absence of any relationship between the variables X and Y, the coefficient of determination r^2, which is the ratio of residual variance s_ϵ to total variance s, would be equal to unity: $\varphi^2 = \frac{s_\epsilon}{s} = 1$. If the variables X and Y were strictly functionally dependent, then the residual variance s_ϵ would be equal to zero $s_\epsilon = 0$ and the convergence coefficient would be $\varphi^2 = \frac{s_\epsilon}{s} = 0$.

Most often, the convergence coefficient φ^2 is in the range [0; 1]. Based on equation (1.14), we can record:

$$r^2 = 1 - \varphi^2 = 1 - \frac{s_\epsilon}{s} = \frac{s - s_\epsilon}{s} = \frac{s_r}{s} \Big|\ s = \frac{1}{N} \cdot \sum_i (y_i - \bar{y})^2 \wedge s_\epsilon = \frac{1}{N} \cdot \sum_i (y_i - \hat{y}_i)^2 \tag{2.10}$$

where: r^2 – coefficient of determination, φ^2 – coefficient of variation, s_ϵ – residual variance, s – total variance, s_r – variance explained by the simple regression model

Equation (2.10) defines the coefficient of determination r^2, which is an indicator of the quality of the model's fit to the measurement data[53]. It informs us what part of the total variance Y is explained by the regression model.

The coefficient of determination r^2 can be defined as the ratio (2.10) of the explained variance[54] s_r to the total variance[55] s. A higher value of the coefficient of determination r^2 indicates a better fit to the model. For example, the value of the coefficient of determination $r^2 = 0{,}4$ indicates that the variance of the values Y around the regression line is 60%[56]. In other words, 40% of the

observed variability Y was explained by regression, and 60% remained in residual variance.

For a small sample, the value of r^2 may be overestimated. This disadvantage is not present in the corrected coefficient of determination $\hat{r}^2$, where the corresponding sums of squares are divided by the number of corresponding degrees of freedom[iii]:

$$\hat{r}^2 = 1 - \frac{s_\epsilon}{s} \cdot \frac{\nu}{\nu_\epsilon} \quad \Big| \quad \nu_\epsilon = n - k - 1 \ \wedge \ \nu = n - 1 \tag{2.11}$$

where: $\hat{r}^2$ – is the corrected coefficient of determination, n – is the number of variables for the model, k – is the number of model coefficients, ν – is the total number of degrees of freedom for the variance.

Example 2.

For the results of measurements of the field under the bands for different volumes of injected oxygen, collected in Table 4.3, simple regression calculations were performed according to the system of equations (2.4) and using equation (2.8).

Then, using relation (2.10), the coefficient of determination r^2 was calculated.

The results of the calculations are summarised in the table below, which shows that over 99% of the observed variability was explained by the regression line: $y \cong 12{,}9043 + 0{,}1141$.

i	x_i	y_i	x_i^2	$x_i \cdot y_i$	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	$(y_i - \bar{y})^2$
1	0.005	0.079680	0.00003	0.00040	0.17858	0.00978	1.61345
2	0.010	0.175846	0.00010	0.00176	0.24311	0.00452	1.37839
3	0.015	0.262424	0.00023	0.00394	0.30763	0.00204	1.18260
4	0.020	0.344589	0.00040	0.00689	0.37215	0.00076	1.01064
5	0.025	0.423929	0.00063	0.01060	0.43667	0.00016	0.85742
6	0.040	0.667410	0.00160	0.02670	0.63023	0.00138	0.46579
7	0.050	0.835166	0.00250	0.04176	0.75928	0.00576	0.26495
8	0.080	1.200193	0.00640	0.09602	1.14641	0.00289	0.02241
9	0.100	1.500584	0.01000	0.15006	1.40449	0.00923	0.02271
10	0.150	2.090068	0.02250	0.31351	2.04971	0.00163	0.54785
11	0.200	2.780115	0.04000	0.55602	2.69492	0.00726	2.04553
12	0.250	3.310592	0.06250	0.82765	3.34013	0.00087	3.84433
13	0.300	3.878057	0.09000	1.16342	3.98535	0.01151	6.39159
suma	1.245	17.549	0.237	3.199	17.549	0.05781	19.648
średnia	0.096	1.350					

$b =$ 12.9043

$a =$ 0.1141

$$\begin{cases} a = \bar{y} - b \cdot \bar{x} \\ b = \dfrac{\sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i}{\sum_i x_i^2 - \bar{x} \cdot \sum_i x_i} \end{cases}$$

$r^2 =$ 0.99705763

$$r^2 = 1 - \frac{s_\varepsilon}{s} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

The correction takes into account the fact that the coefficient of determination r^2 is calculated from a sample and therefore has an overestimated value[58] when generalising the results to the general population. The corrected coefficient of determination $\hat{r}^2$ determines how well the obtained regression equation would fit a sample of infinite size $n \to \infty$ from the same population N. Hence, the correction $\frac{\nu}{\nu_\varepsilon}$ for a large n sample size is negligible, since using relation (2.11) we can write: $\lim_{n\to\infty} \frac{\nu_\varepsilon}{\nu} = \lim_{n\to\infty} \frac{n-k-1}{n-1} = \lim_{n\to\infty} 1 - \frac{k}{n-1} = 1$.

As mentioned earlier, the key issue in the least squares method is to estimate the value of the sum $S = \sum_i (y_i - a - b \cdot x_i)^2 \equiv \sum_i (y_i - \hat{y}_i)^2$ of the squares of the deviations (2.1) of the empirical measurement points y_i from their estimates obtained in the prediction[59] $\hat{y}_i$. When linear regression completely explains the variability, the sum S from relation (2.1) is identically equal to zero: $S \equiv 0$. This case is also referred to as volatility elimination by linear regression. By inserting the regression equation $\hat{y}_i = a + b \cdot x$ into equation (2.1) and transforming the first expression $a = \bar{y} - b \cdot \bar{x}$ from the system of equations (2.3), we can write that $S = \sum_i (y_i - \bar{y} - b \cdot \bar{x} - b \cdot x_i)^2 = \sum_i (y_i - \bar{y}_i)^2 - b^2 \cdot \sum_i (x_i - \bar{x}_i)^2$, and hence

$$S = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \bar{y}_i)^2 - b^2 \cdot \sum_i (x_i - \bar{x}_i)^2 \tag{2.12}$$

Conclusion

The chapter describes the least squares method for estimating linear regression coefficients and the use of F–Fisher based on F–Snedecor distribution to assess the homogeneity of variance (σ) for two versions of linear regression equations: simplified $y = b \cdot x$ and full $y = a + b \cdot x$. The expected result of the test is to conclude that there is no statistical difference between the variances, and therefore the simplified regression equation $y = b \cdot x$ can be adopted, which is preferred from a metrological point of view[60] [6]. The preference for the equation $y = b \cdot x$ is related to the fact that when establishing the relationship between the values indicated by the measuring device reproduced for the standard[61], only a model in the form of a simplified regression equation should be adopted, for an intercept equal to zero: $\hat{a} \equiv 0$, as theoretically, when a blind sample is given[62], no signal generated by the measuring device should be observed.

The chapter also discusses the coefficient of determination r^2, which illustrates the extent to which the regression equation explains the observed variability during the calibration process.

2. REGRESSION ANALYSIS

Based on variance analysis, it is assumed that the best estimate $\hat{\sigma}^2$ of the actual variance σ^2 from the sample is the sum of squares of deviations from the mean (2.1) $S = \sum_i (y_i - \hat{y}_i)^2$, divided by the number of degrees of freedom ν, $s^2 = \frac{S}{\nu} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\nu}$. The standard deviation divided by the square root $\sqrt{\nu}$ of the number of degrees of freedom[63] ν is proportional to the estimation of the confidence interval $\Delta\hat{y}$ for the value $\hat{y}$:$\Delta\hat{y} = \pm\frac{s}{\sqrt{\nu}} \cdot t(P.\nu)$ – see further **(3.15).**

VARIANCE OF THE SLOPE COEFFICIENT

Using the properties of variance, for independent variables:

(1.9) $D^2(c \cdot X) = c^2 \cdot D^2 X$

(1.10) $D^2(X_1 + .. + X_N) = D^2 X_1 + .. + D^2 X_N$

and using the relationship (2.7)$\hat{b} = \frac{S_{XY}}{S_{XX}} = \sum_i c_i \cdot y_i \,|c_i = \frac{(x_i - \bar{x})}{S_{XX}}$, the variance of the slope coefficient of a simple regression can be written as:

$$D^2(\hat{b}) = D^2(\sum_i c_i \cdot y_i) = D^2(y_i) \cdot \sum_i c_i^2 \quad \Big| \quad c_i = \frac{(x_i - \bar{x})}{S_{XX}} \tag{3.1}$$

Assuming that $D^2(y_i) \equiv \sigma$ and using the relationship (2.5) $S_{XX} = \sum_i x_i^2 - \bar{x} \cdot \sum_i x_i = \sum_i (x_i - \bar{x})^2$, equation (3.1) can be transformed into the form:

$$D^2(\hat{b}) \triangleq s_{\hat{b}}^2 = \hat{\sigma}^2 \cdot \sum_i c_i^2 = \hat{\sigma}^2 \cdot \frac{\sum_i (x_i - \bar{x})^2}{S_{XX}^2} = \frac{\hat{\sigma}^2}{S_{XX}} \tag{3.2}$$

VARIANCE OF THE INTERCEPT

To estimate the variance of the intercept, proof is needed that the covariance $cov(\bar{y}, \hat{b})$ of the mean value of the dependent variable $\bar{y}$ and the estimated slope coefficient $\hat{b}$ is equal to zero: $cov(\bar{y}, \hat{b}) = 0$.
Using the relationship:

(2.4) $$\forall_{y=\hat{a}+\hat{b}\cdot x} \begin{cases} \hat{a} = \bar{y} - \hat{b} \cdot \bar{x} \\ \hat{b} = \frac{\sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i}{\sum_i x_i^2 - \bar{x} \cdot \sum_i x_i} \end{cases} \Bigg| \; \bar{x} = \frac{1}{n} \cdot \sum_i x_i \;\wedge$$
$$\bar{y} = \frac{1}{n} \cdot \sum_i y_i$$

(2.5) $S_{XX} = \sum_i x_i^2 - \bar{x} \cdot \sum_i x_i = \sum_i (x_i - \bar{x})^2$

The following transformations can be performed:

$$\hat{b} = \frac{\sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i}{\sum_i x_i^2 - \bar{x} \cdot \sum_i x_i} = \frac{n \cdot (\sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i)}{n \cdot (\sum_i x_i^2 - \bar{x} \cdot \sum_i x_i)} =$$

$$= \frac{n^2 \cdot \left(\frac{1}{n} \sum_i x_i \cdot y_i - \frac{1}{n} \bar{y} \cdot \sum_i x_i\right)}{n \cdot (\sum_i x_i^2 - \bar{x} \cdot \sum_i x_i)} = \frac{n \cdot \left(\frac{1}{n} \sum_i x_i \cdot y_i - \frac{1}{n} \sum_i y_i \cdot \bar{x}\right)}{\sum_i x_i^2 - \bar{x} \cdot \sum_i x_i} =$$

$$= \frac{\sum_i x_i \cdot y_i - \bar{x} \cdot \sum_i y_i}{S_{XX}} = \frac{\sum_i (x_i \cdot y_i - \bar{x} \cdot y_i)}{S_{XX}} = \frac{\sum_i [y_i \cdot (x_i - \bar{x})]}{S_{XX}}$$

Based on the above transformations, where $\hat{b} = \frac{\sum_i [y_i \cdot (x_i - \bar{x})]}{S_{XX}}$ and from the covariance property (1.12): $cov(a \cdot X, b \cdot Y) = a \cdot b \cdot cov(X, Y) \;\; | \;\; a, b = const$, the expression $cov(\bar{y}, \hat{b}) = cov\left(\frac{\sum_i y_i}{n}, \frac{\sum_i y_i \cdot (x_i - \bar{x})}{S_{XX}}\right)$ can be transformed into the form: $cov(\bar{y}, \hat{b}) = \frac{1}{n \cdot S_{XX}} \cdot cov[\sum_i y_i, \sum_i y_i \cdot (x_i - \bar{x})]$, and hence further[iv] to: $cov(\bar{y}, \hat{b}) = \frac{1}{n \cdot S_{XX}} \cdot \sum_i (x_i - \bar{x}) \cdot cov(y_i, y_i)$. Given the relationship (1.14) $cov(Y, Y) = var(Y) \equiv \sigma^2$, we can write down the following:

$$cov(\bar{y}, \hat{b}) = \frac{\sum_i (x_i - \bar{x})}{n \cdot S_{XX}} \cdot \sigma^2 \tag{3.3}$$

Using the fact that[65] $\sum_i (x_i - \bar{x}) \stackrel{\text{def}}{=} 0$, relation (3.3) can be written as the covariance value equal to zero (28):

$$cov(\bar{y}, \hat{b}) = 0 \tag{3.4}$$

Using relation (2.4), which is an estimate of the directional coefficient of linear regression $\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$ using the least squares method, the variance $D^2\hat{a}$ for the intercept $\hat{a}$ can be written as: $D^2\hat{a} = D^2(\bar{y} - \hat{b} \cdot \bar{x})$. Using the properties of variance[66] (1.8): $D^2(X \pm Y) = D^2 X + D^2 Y \pm 2 \cdot cov(X, Y)$, we can write: $D^2\hat{a} = D^2\bar{y} + D^2(\hat{b} \cdot \bar{x}) - 2 \cdot cov(\bar{y}, \hat{b} \cdot \bar{x})$. Using the properties of variance (1.8) $D^2(c \cdot X) = c^2 \cdot D^2 X$, we can further write: $D^2\hat{a} = D^2\bar{y} + \bar{x}^2 \cdot D^2\hat{b} - 2 \cdot \bar{x} \cdot cov(\bar{y}, \hat{b})$.

Using relation (3.4) $cov(\bar{y}, \hat{b}) = 0$, the expression can be further simplified as: $D^2\hat{a} = D^2\bar{y} + \bar{x}^2 \cdot D^2\hat{b}$. The variance value for the mean value $D^2\bar{y}$ is $D^2\bar{y} = \frac{\hat{\sigma}^2}{n}$, which can be further transformed to: $D^2\hat{a} = \frac{\hat{\sigma}^2}{n} + \bar{x}^2 \cdot D^2\hat{b}$. Using the relation (3.2) $D^2(\hat{b}) = \frac{\hat{\sigma}^2}{S_{XX}}$, we can write that the variance of the estimation of the intercept $D^2\hat{a}$ is:

$$D^2\hat{a} \triangleq s_{\hat{a}}^2 = \hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right). \tag{3.5}$$

VARIANCE OF THE ESTIMATED VALUE

In prediction, two cases should be distinguished, namely when the prediction is made for:

- the reference population used to estimate the parameters of the simple regression,
- another population[67] used to draw conclusions about it based on simple regression from the reference population.

The first case concerns the determination of the confidence interval for simple regression$y = \hat{a} + \hat{b} \cdot x + \epsilon$. The second concerns the determination of the confidence

interval for the use of simple regression to predict the dependent variable $\hat{y}$ or the independent variable $\hat{x}$, for a population from another sample, for example using simple regression as a regression curve characterising the detector response for the analysis of an unknown sample[68].

VARIANCE FOR LINEAR REGRESSION

In linear regression, it is assumed that there exists a linear relationship between any realisation of the independent variable $X \leftarrow x_i$ and the corresponding realisation of the dependent variable $Y \leftarrow y_i : E(y|x) = a + b \cdot x$. However, in reality, there occur deviations ϵ from the estimated regression line $y = \hat{a} + \hat{b} \cdot x$, hence the model taking into account deviations from the estimated regression line will take the form:

$$y = \hat{a} + \hat{b} \cdot x + \epsilon \tag{3.6}$$

where: ϵ – represents a random error with a normal distribution $\mathbb{N}(0, \sigma)$ with a mean value of zero: $\mu = 0$ and an unknown variance σ.

For the estimated regression line (3.6) $\forall_{i \in \mathbb{N}}$ $y_i = \hat{a} + \hat{b} \cdot x_i + \epsilon_i$, the sum of squares of deviations S_ϵ from the linear regression can be written analogously to equation (2.1): $S_\epsilon = \sum_i \epsilon_i^2 = \sum_i (y_i - \hat{a} - \hat{b} \cdot x_i)^2 = \sum_i (y_i - \hat{y}_i)^2$. On this basis, the unbiased estimator of the unknown variance σ of the random error distribution ϵ can be estimated as:

$$\hat{\sigma}_\epsilon^2 = \frac{S_\epsilon}{\nu} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\nu} = \frac{S_{YY} - \hat{b} \cdot S_{XY}}{\nu} \tag{3.7}$$

where: ν –-represents the number of degrees of freedom $\nu = n - 1$ for the simplified equation $y = \hat{b} \cdot x$ and $\nu = n - 2$ for the equation $y = \hat{a} + \hat{b} \cdot x$, n – the number of measurements.

In some situations, it is convenient to transform regression equation (3.6) into the form:

$$\begin{aligned} y_i &= \hat{a} + \hat{b} \cdot (x_i - \bar{x}) + \hat{b} \cdot \bar{x} + \epsilon \\ &= \left(\hat{a} + \hat{b} \cdot \bar{x}\right) + \hat{b} \cdot (x_i - \bar{x}) + \epsilon \\ &= \hat{a}' + \hat{b} \cdot (x_i - \bar{x}) + \epsilon \end{aligned} \tag{3.8}$$

Redefining the explanatory variable[69] in equation (3.8), in relation to relationship (3.6), caused the intersection point with the axis OX to shift from zero $y = 0$ to the value $y = \bar{x}$. Comparing the first equation from the system of equations (2.4) $\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$ with equation (3.8), it can be seen that $\hat{a}'$ can be expressed as: $\hat{a}' = \bar{y} = \hat{a} + \hat{b} \cdot \bar{x}$. During the transformation, the value of the determined slope coefficient of the regression line $\hat{b}$ does not change. Therefore, the regression equation (3.8) can be written in a simpler form as:

$$y_i = \bar{y} + \hat{b} \cdot (x_i - \bar{x}) + \epsilon. \tag{3.9}$$

The main application of the regression model is to estimate the expected response $E(y)$ for a specific value of the variable x. Let x_0 be the value of the variable $X \leftarrow x_0$ occurring in any formula adopted to determine the regression line for which we want to read the average response: $E(\hat{y}|x^0)$, where x^0 is any value of the independent random variable X within the range of the original data $x_1^0 .. x_i^0$ used to determine the regression line model. The unbiased point estimator $E(\hat{y}|x^0)$ can be found based on the fitted model as $E(\hat{y}|x^0) = \hat{\mu}_{y|x^0} = \hat{a} + \hat{b} \cdot x^0$. The variance $D^2\left(\hat{\mu}_{\hat{y}|x^0}\right)$ can be written as $D^2\left(\hat{\mu}_{\hat{y}|x^0}\right) = D^2(\hat{a} + \hat{b} \cdot x_i^0)$.

Using the equations:

$$\begin{aligned} &(1.8) \quad D^2(c \cdot X) = c^2 \cdot D^2 X \\ &(2.3) \quad \hat{a} = \bar{y} - \hat{b} \cdot \bar{x} \\ &(3.2) \quad D^2\left(\hat{b}\right) = \frac{\sigma^2}{S_{XX}} \\ &(3.4) \quad cov\left(\bar{y}, \hat{b}\right) = 0 \\ &(3.9) \quad \hat{y} = \bar{y} + \hat{b} \cdot (x_i - \bar{x}) + \epsilon \end{aligned}$$

$D^2\left(\hat{\mu}_{y|x^0}\right) = \frac{s^2_{\hat{y}|x^0}}{n} = D^2(\hat{a} + \hat{b} \cdot x^0)$ can be written as[70], as:

$$\begin{aligned} D^2(\hat{\mu}_{\hat{y}|x^0}) &= D^2\left(\hat{a} + \hat{b} \cdot x^0\right) = D^2[\bar{y} + \hat{b} \cdot (x^0 - \bar{x})] = \\ &= D^2\bar{y} + (x^0 - \bar{x})^2 \cdot D^2\hat{b} - 2 \cdot (x^0 - \bar{x}) \cdot cov\left(\bar{y}, \hat{b}\right) = \\ &= \frac{\sigma^2}{n} + (x^0 - \bar{x})^2 \cdot \frac{\sigma^2}{S_{XX}} = \sigma^2 \cdot \left(\frac{1}{n} + \frac{(x^0 - \bar{x})^2}{S_{XX}}\right). \end{aligned} \tag{3.10}$$

As can be seen from equation (3.10), the value of the estimated variance during the calibration process $\frac{s^2_{\hat{y}|x^0}}{n}$ depends on the value of the standard x^0 and reaches its minimum for its mean value[71] $x^0 = \bar{x}$. It should also be noted that there is no summation sign in the numerator of the second fraction.

Equation (3.10) is used to determine the confidence interval Δy for each point corresponding to the reference value x^0 for the determined regression line $\hat{y} = \hat{a} + \hat{b} \cdot x^0$ of the calibration process. The confidence interval is bounded on both sides by hyperbolas, and their distance is smallest for the mean value $\bar{x}$. The variance value (3.10) differs from the variance value for the reference curve calculated for all points.

VARIANCE OF OUTLIERS

When deciding whether measurement points can be added to the set of points used to determine the regression line for inference based on the existing regression line[72] (3.6), the variance[73] must first be determined: $D^2\left(\hat{\mu}_{y|y_0}\right) = D^2(\hat{y}_0 - y_0)$ for the regression line $y_0 = a_0 + b_0 \cdot x_0 + \epsilon$, when the value of the dependent variable is fixed and equals $y_0 = f(x_0, \epsilon)$, and the estimated value for x_0 is: $\hat{y}_0 = \hat{a}_0 + \hat{b}_0 \cdot x_0$. The error ϵ has a normal distribution $\epsilon \in \mathbb{N}(0, \sigma^2)$ with a central value of zero $\mu = 0$ and variance σ^2. Hence, we can write:

$$\begin{aligned} D^2(\hat{y}_0 - y_0) &= D^2[\hat{a}_0 + \hat{b}_0 \cdot x_0 - (a_0 + b_0 \cdot x_0 + \epsilon)] = \\ &= D^2\left(\hat{a}_0 + \hat{b}_0 \cdot x_0 - \epsilon\right) = D^2(\hat{a}_0) + x_0^2 \cdot D^2\left(\hat{b}_0\right) + D^2\epsilon + \\ &+ 2 \cdot x_0 \cdot cov\left(\hat{a}_0, \hat{b}_0\right) - 2 \cdot x_0 \cdot cov(\hat{a}_0, \epsilon) - 2 \cdot x_0 \cdot cov(\hat{b}_0, \epsilon) \end{aligned} \tag{3.11}$$

The covariances $cov(\hat{a}_0, \epsilon) \stackrel{\text{def}}{=} 0$ and $cov(\hat{b}_0, \epsilon) \stackrel{\text{def}}{=} 0$ are by definition equal to zero, as they concern the interdependence between independent variables. Using:

(1.14) $cov(X,X) = var(X) = D^2(X)$

(2.3) $\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$

(3.2) $D^2(\hat{b}) = \frac{\hat{\sigma}^2}{S_{XX}}$

(3.4) $cov(\bar{y}, \hat{b}) = 0$

the transformation can be written as: $cov(\hat{a}_0, \hat{b}_0) = cov(\bar{y} - \hat{b}_0 \cdot \bar{x}, \hat{b}_0)$. And further: $cov(\bar{y} - \hat{b}_0 \cdot \bar{x}, \hat{b}_0) = cov(\bar{y},\ \hat{b}_0) - \bar{x} \cdot cov(\hat{b}_0, \hat{b}_0) = -\bar{x} \cdot D^2(\hat{b}_0)$.
Hence, ultimately, we can written [26]:

$$cov(\hat{a}_0, \hat{b}_0) = -\bar{x} \cdot \frac{\hat{\sigma}^2}{S_{XX}} \tag{3.12}$$

Therefore, since

(1.8) $D^2(c \cdot X) = c^2 \cdot D^2 X$

(1.9) $D^2 c = 0$

(3.2) $D^2 \hat{b} = \frac{\hat{\sigma}^2}{S_{XX}}$

(3.5) $D^2 \hat{a} \triangleq s_{\hat{a}}^2 = \hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)$

(3.7) $D^2 \epsilon = \sigma$

(3.12) $cov(\hat{a}_0, \hat{b}_0) = -\bar{x} \cdot \frac{\hat{\sigma}^2}{S_{XX}}$

the relationship (3.11) can be written as:

$$\begin{aligned} D^2(\hat{y}_0 - y_0) &= \hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} + \frac{x_0^2}{S_{XX}} + 1 - 2 \cdot x_0 \cdot \bar{x} \cdot \frac{\hat{\sigma}^2}{S_{XX}}\right) = \\ &= \hat{\sigma}^2 \cdot \left(1 + \frac{1}{n} + \frac{\bar{x}^2 + x_0^2 - 2 \cdot x_0 \cdot \bar{x}}{S_{XX}}\right) = \\ &= \hat{\sigma}^2 \cdot \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right). \end{aligned} \tag{3.13}$$

PREDICTION VARIANCE

The system of equations (2.4) can be used to predict the value of the independent variable $\bar{x}_z = \frac{\bar{y}_z - \hat{a}}{\hat{b}}$ based on the regression line and the mean values of the dependent variable $\bar{y}_z = \frac{\sum_i y_i}{k}$ [6]:

$$\sigma_{\bar{x}_z}^2 = \frac{\sigma^2}{\hat{b}^2} \cdot \left(\frac{1}{k} + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\hat{b}^2 \cdot S_{XX}}\right) \tag{3.14}$$

This situation occurs when measuring the gas content in a breathing mixture based on a reference curve, using gas chromatography and a TCD detector.

The prediction variance $\sigma_{\bar{x}_z}^2$ (3.14) can be estimated using the variance estimate for error propagation (5.7).

The regression equation (3.9) expressed by the $\hat{y} = \bar{y} + \hat{b} \cdot (x_i - \bar{x})$ means can be written as $x = \bar{x} + \frac{1}{\hat{b}} \cdot (\hat{y} - \bar{y})$. Since y, $\bar{y}$ and $\hat{b}$ are independent of each other, the variance can be estimated from the error propagation modelling relationship (1.25). Expanding the function for first-order derivatives[74] [5] $\sigma_x^2 = [\frac{\partial x(y,\bar{y},\hat{b})}{\partial \hat{y}}]^2 \cdot \sigma_{\hat{y}}^2 + [\frac{\partial x(y,\bar{y},\hat{b})}{\partial \bar{y}}]^2 \cdot \sigma_{\bar{y}}^2 + [\frac{\partial x(y,\bar{y},\hat{b})}{\partial \hat{b}}]^2 \cdot \sigma_{\hat{b}}^2$ and substituting: $\sigma_{\hat{y}}^2 = \frac{\sigma^2}{m}$, $\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}$, $\sigma_{\hat{b}}^2 = \frac{(x_i - \bar{x})^2}{S_{XX}}$ we obtain equation (3.14).

CONFIDENCE INTERVAL FOR REGRESSION

Variance estimates can be used to determine confidence intervals for linear regression parameters estimated by the least squares method (3.9) and confidence intervals for predictions based on simple regression in a similar way to that used to estimate the true value μ from a small sample using T–Student distribution:

$$\Delta\hat{y} = y \pm t(\bar{P}, \nu) \cdot \frac{s}{\sqrt{\nu}} \tag{3.15}$$

where: $t(\bar{P}, \nu)$ – expansion coefficient being the quantile for the T-Student distribution for the confidence level $\bar{P} = 1 - \alpha$ and the number of degrees of freedom[75] ν, α – significance, s – standard deviation

The difference between the observed value of the dependent variable y_i and the value estimated by the least squares method $\hat{y}_i$ can be considered as the residual unexplained by the regression line ϵ_i, as given in equation (3.7).

The uncertainties in determining the directional coefficient Δb and the intercept Δa can be written similarly to the calculation of the uncertainty in determining the means from a small sample[76] using variance analysis [6,26,29,30]:

$$\begin{cases} \Delta\hat{a} = t(\bar{P}, \nu) \cdot s_a \\ \Delta\hat{b} = t(\bar{P}, \nu) \cdot s_b \end{cases} \tag{3.16}$$

where: $\Delta\hat{a}$ – uncertainty in estimating the intercept of the linear regression, $\Delta\hat{b}$ – uncertainty in estimating the slope coefficient of the linear regression, $t(\bar{P}, \nu)$ – coefficient for T-Student's two-tailed distribution at a confidence level $\bar{P}$ and number of degrees of freedom[77] ν, s_a – variance from intercept measurements for simple regression, s_b – variance from slope coefficient measurements for the regression line

Using relation (3.5) $s_{\hat{a}}^2 = \hat{\sigma}^2 \cdot (\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}})$ and relations from the system of equations (3.16), confidence intervals for the intercept can be estimated as:

$$\hat{a} - t(\bar{P}, \nu = n-2) \cdot \sqrt{\hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)} \le a \le \hat{a} + t(\bar{P}, \nu) \cdot \sqrt{\hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)} \tag{3.17}$$

where: $t(\bar{P}, \nu = n-2)$ – quantile for the two-tailed T-Student distribution for significance $\alpha = 1 - \bar{P}$ and $\nu = n - 2$ degrees of freedom, α – significance, $\bar{P} = 1 - \alpha$, $S_{XX} = \sum_i x_i^2 - \bar{x} \cdot \sum_i x_i == \sum_i (x_i - \bar{x})^2$ according to (2.5), $\hat{\sigma}^2$ – variance for linear regression $\hat{\sigma}^2 = \frac{S_\epsilon}{\nu} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\nu} = \frac{S_{YY} - \hat{b} \cdot S_{XY}}{\nu}$ according to (3.7).

Using relation (3.2) $s_{\hat{b}}^2 = \frac{\hat{\sigma}^2}{S_{XX}}$ and relations from the system of equations (3.16), confidence intervals for the intercept can be estimated as:

(3.18)

$$\hat{b} - t(\bar{P}, \nu) \cdot \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}} \leq b \leq \hat{b} + t(\bar{P}, \nu) \cdot \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}$$

where: $t(\bar{P}, \nu)$ – quantile for the two-tailed T-Student for significance $\alpha = 1 - \bar{P}$ and ν degrees of freedom, α – significance $\bar{P} = 1 - \alpha$,$S_{XX} = \sum_i x_i^2 - \bar{x} \cdot \sum_i x_i = \sum_i (x_i - \bar{x})^2$ according to (2.5), $\hat{\sigma}^2$ – variance for linear regression $\hat{\sigma}^2 = \frac{S_\varepsilon}{\nu} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\nu} = \frac{S_{YY} - \hat{b} \cdot S_{XY}}{\nu}$ according to (3.7), ν – degrees of freedom[78].

Example 3.

The equation of the regression line was determined by the least squares method from the averages of oxygen band integrations observed after injecting different volumes of oxygen– Table 4.3.

Tab 4.3.

i	x_i	y_i	x_i^2	$x_i \cdot y_i$	$\hat{y}_i$
1	0,005	0,0796803	0,00003	0,0004	0,179
2	0,01	0,175846	0,00010	0,0018	0,243
3	0,015	0,2624235	0,00023	0,0039	0,308
4	0,02	0,3445894	0,00040	0,0069	0,372
5	0,025	0,4239291	0,00063	0,0106	0,437
6	0,04	0,6674096	0,00160	0,0267	0,630
7	0,05	0,83516585	0,00250	0,0418	0,759
8	0,08	1,2001927	0,00640	0,0960	1,146
9	0,1	1,500583567	0,01000	0,1501	1,404
10	0,15	2,0900684	0,02250	0,3135	2,050
11	0,2	2,78011545	0,04000	0,5560	2,695
12	0,25	3,3105919	0,06250	0,8276	3,340
13	0,3	3,8780566	0,09000	1,1634	3,985
Suma	1,245	17,549	0,237	3,1987	17,549
Średnia	0,096	1,350			0,179

Using the system of equations (2.4), we can calculate the following:

$$\begin{cases} \hat{a} \cong 1{,}350 - 12{,}9 \cdot 0{,}096 \cong 0{,}110 \\ \hat{b} \cong \frac{3{,}1987 - 1{,}350 \cdot 1{,}245}{0{,}237 - 0{,}096 \cdot 1{,}245} \cong 12{,}9 \end{cases}.$$

Using equation (2.8), the slope coefficient will be:

$$\hat{b} \cong \frac{3{,}1987}{0{,}237} \cong 13{,}5 \pm 0{,}5.$$

The calculations were performed on a calculator with an accuracy of three significant digits. They are consistent with the results obtained from calculations performed in EXCEL according to the system of equations (3.16). Using the equations:

(3.16) $$\begin{cases} \Delta\hat{a} = t(\bar{P}, \nu) \cdot s_a \\ \Delta\hat{b} = t(\bar{P}, \nu) \cdot s_b \end{cases}$$

(3.2) $$s_{\hat{b}}^2 = \frac{\hat{\sigma}^2}{S_{XX}}$$

.(2.5) $$S_{XX} = \sum_i x_i^2 - \bar{x} \cdot \sum_i x_i = \sum_i (x_i - \bar{x})^2$$

(3.16) $$\Delta\hat{b} = \pm t(\bar{P}, \nu = n - 1) \cdot \sqrt{\frac{\hat{\sigma}^2}{\sum_i x_i^2 - \bar{x} \cdot \sum_i x_i}}$$

the uncertainty $\Delta\hat{b}$ of determining the slope coefficient $\hat{b}$ was estimated.

$$S_{YY} = \sum_i y_i^2 - \bar{y} \cdot \sum_i y_i = \sum_i (y_i - \bar{y})^2$$

$$S_{XX} = \sum_i x_i^2 - \bar{x} \cdot \sum_i x_i = \sum_i (x_i - \bar{x})^2$$

$$S_{XY} = \sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i = \sum_i y_i \cdot (x_i - \bar{x})$$

$$\hat{\sigma}^2 = \frac{S_{YY} - \hat{b} \cdot S_{XY}}{\nu}$$

$$s_{\hat{b}}^2 = \frac{\hat{\sigma}^2}{S_{XX}} \qquad \Delta\hat{b} = t(\alpha, \nu) \cdot s_b$$

$$s_{\hat{a}}^2 = \hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right) \qquad \Delta\hat{a} = t(\alpha, \nu) \cdot s_a$$

x_i	y_i	x_i	x_i^2	y_i	y_i^2	$x_i \cdot y_i$
0.005	0.07968	0.005	0.000	0.080	0.006	0.000
0.010	0.17585	0.010	0.000	0.176	0.031	0.002
0.015	0.26242	0.015	0.000	0.262	0.069	0.004
0.020	0.34459	0.020	0.000	0.345	0.119	0.007
0.025	0.42393	0.025	0.001	0.424	0.180	0.011
0.040	0.66741	0.040	0.002	0.667	0.445	0.027
0.050	0.83517	0.050	0.003	0.835	0.698	0.042
0.080	1.20019	0.080	0.006	1.200	1.440	0.096
0.100	1.50058	0.100	0.010	1.501	2.252	0.150
0.150	2.09007	0.150	0.023	2.090	4.368	0.314
0.200	2.78012	0.200	0.040	2.780	7.729	0.556
0.250	3.31059	0.250	0.063	3.311	10.960	0.828
0.300	3.87806	0.300	0.090	3.878	15.039	1.163
sumy		1.245	0.237	17.549	43.337	3.199
średnia		0.10		1.35		

liczba punktów pomiarowych	$n =$	13	
równanie regresji		$y = \hat{a} \cdot \hat{b} \cdot x$	$y = \hat{b} \cdot x$
współczynnik kierunkowy prostej regresji	$\hat{b} =$	12.9043	13.5038
wyraz wolny prostej regresji	$\hat{a} =$	0.1141	
wariancja pomiędzy wartościami zmierzonymi (y_i) i obliczonymi (Y_i)	$\hat{\sigma}^2 =$	0.00526	0.00482
wariancja wspólczynnika kierunkowego prostej regresji	$s_{\hat{b}}^2 =$	0.04467	0.04095
wariancja wyrazu wolnego prostej regresji	$s_{\hat{a}}^2 =$	0.00081	
poziom ufności	$P =$	0.95	
błąd współczynnika kierunkowego prostej regresji	$\Delta\hat{b} =$	0.465	0.441
błąd wyrazu wolnego prostej regresji	$\Delta\hat{a} =$	0.063	
wybór równania prostej regresji przy pomocy testu F dla wariancji	$F =$	0.9167	
wartość krytyczna testu F	$F\left(\frac{\alpha}{2}, \nu_1, \nu_2\right) =$	2.7876	
decyzja		0	1

For the determined regression line $y = \hat{a} + \hat{b} \cdot x$, its confidence interval $\pm\Delta$ can be determined, i.e. the values of the dependent variable y based on the given values of the independent variable x. Due to unavoidable errors ϵ, when estimating the value $\hat{y}$ based on the regression line $y = \hat{a} + \hat{b} \cdot x + \epsilon$, consisting of the uncertainty in determining the intercept $\Delta\hat{a}$ and the slope $\Delta\hat{b}$ the estimated value $\hat{y}$ should be treated as a random variable Y. Taking into account the variance $D^2(\epsilon) = \sigma^2$ from relation (3.10) and the uncertainty[v] ϵ the confidence interval for the expected value of the dependent variable Ey for the value x_0 from the pool of standards already used to determine the regression line can be written as:

$$\hat{a} + \hat{b} \cdot x_i^0 - t(\bar{P}, \nu) \cdot \sqrt{\hat{\sigma}^2 \cdot \left[\frac{1}{n} + \frac{(x_i^0 - \bar{x})^2}{S_{XX}}\right]} \leq \leq Ey \leq \hat{a} + \hat{b} \cdot x_i^0 + t(\bar{P}, \nu) \cdot \sqrt{\hat{\sigma}^2 \cdot \left[\frac{1}{n} + \frac{(x_i^0 - \bar{x})^2}{S_{XX}}\right]} \tag{3.19}$$

Similarly, based on the variance (3.13), the confidence interval $\Delta\hat{y} = t(\bar{P}, \nu) \cdot \sqrt{\hat{\sigma}^2 \cdot \left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}\right]}$ for rejecting an additional point x_i can be estimated, or it can be added to the estimation of the regression line[80] as:

$$\hat{a} + \hat{b} \cdot x_i - t(\bar{P}, \nu) \cdot \sqrt{\hat{\sigma}^2 \cdot \left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}\right]} \leq \leq \hat{y} \leq \hat{a} + \hat{b} \cdot x_i + t(\bar{P}, \nu) \cdot \sqrt{\hat{\sigma}^2 \cdot \left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}\right]}. \tag{3.20}$$

Example 4.

The confidence interval (3.19) was calculated for simple regressions of the type $y = \hat{b} \cdot x$ from the mean band integrations for oxygen observed after injecting different volumes collected in Table 4.3.

The confidence limits determined according to formula (3.19) show which straight line equations can be drawn within these limits. In the case under consideration, the regression equation after calibration must correspond to the equation $y = \hat{b} \cdot x$ (as follows from the inference based on F-Fisher distribution), pass through the point representing the mean values, and fall within the limits determined by formula (3.19).

TLEN 2023-08-14

$S_{YY} = \sum_i y_i^2 - \bar{y} \cdot \sum_i y_i = \sum_i (y_i - \bar{y})^2$

$S_{XX} = \sum_i x_i^2 - \bar{x} \cdot \sum_i x_i = \sum_i (x_i - \bar{x})^2$

$S_{XY} = \sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i = \sum_i y_i \cdot (x_i - \bar{x})$

$\hat{\sigma}^2 = \frac{S_{YY} - \hat{b} \cdot S_{XY}}{v}$

$s_{\hat{b}}^2 = \frac{\hat{\sigma}^2}{S_{XX}}$ $\quad \Delta\hat{b} = t(P, v) \cdot s_b$

$s_{\hat{a}}^2 = \hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)$ $\quad \Delta\hat{a} = t(P, v) \cdot s_a$

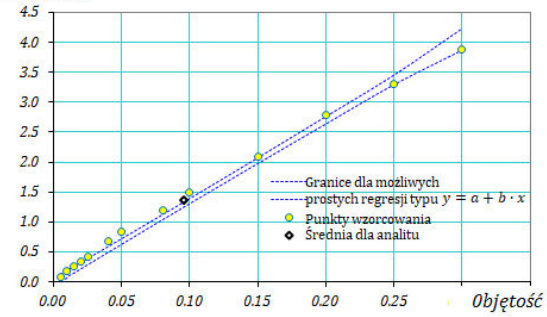
$\Delta\hat{y} = \hat{a} + \hat{b} \cdot x \pm t(\bar{P}, v) \cdot \sqrt{D^2Y}$ $\quad D^2Y = \hat{\sigma}^2 \cdot [\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}}]$

$\Delta\hat{x} = \frac{\frac{\sum_i y_i}{k} - \hat{a}}{\hat{b}} \pm t(\bar{P}, v) \cdot \sqrt{D^2X}$ $\quad D^2X = \frac{\hat{\sigma}^2}{\hat{b}^2} \cdot [\frac{1}{n} + \frac{1}{m} + \frac{(\frac{\sum_k y_k}{m} - \bar{y})^2}{\hat{b}^2 \cdot S_{XX}}]$

WZORCOWANIE									
x_i	y_i	x_i	x_i^2	y_i	y_i^2	$x_i \cdot y_i$	$\hat{y}_i$	$\hat{y}_i + \Delta\hat{y}_i$	$\hat{y}_i - \Delta\hat{y}_i$
0.005	0.07968	0.005	0.000	0.080	0.006	0.000	0.068	0.125	-0.094
0.010	0.17585	0.010	0.000	0.176	0.031	0.002	0.135	0.192	0.079
0.015	0.26242	0.015	0.000	0.262	0.069	0.004	0.203	0.258	0.148
0.020	0.34459	0.020	0.000	0.345	0.119	0.007	0.270	0.324	0.216
0.025	0.42393	0.025	0.001	0.424	0.180	0.011	0.338	0.390	0.285
0.040	0.66741	0.040	0.002	0.667	0.445	0.027	0.540	0.589	0.492
0.050	0.83517	0.050	0.003	0.835	0.698	0.042	0.675	0.722	0.629
0.080	1.20019	0.080	0.006	1.200	1.440	0.096	1.080	1.123	1.038
0.100	1.50058	0.100	0.010	1.501	2.252	0.150	1.350	1.392	1.308
0.150	2.09007	0.150	0.023	2.090	4.368	0.314	2.026	2.074	1.977
0.200	2.78012	0.200	0.040	2.780	7.729	0.556	2.701	2.763	2.639
0.250	3.31059	0.250	0.063	3.311	10.960	0.828	3.376	3.456	3.296
0.300	3.87806	0.300	0.090	3.878	15.039	1.163	4.051	4.232	3.870
sumy		1.245	0.237	17.549	43.337	3.199			
średnia	0.10	1.35							

liczba punktów pomiarowych	$n =$	13	
równanie regresji		$y = \hat{a} \cdot \hat{b} \cdot x$	$y = \hat{b} \cdot x$
współczynnik kierunkowy prostej regresji	$\hat{b} =$	12.9043	13.5038
wyraz wolny prostej regresji	$\hat{a} =$	0.1141	
wariancja pomiędzy wartościami zmierzonymi (y_i) i obliczonymi (Y_i)	$\hat{\sigma}^2 =$	0.00526	0.00482
wariancja współczynnika kierunkowego prostej regresji	$s_{\hat{b}}^2 =$	0.04467	0.04095
wariancja wyrazu wolnego prostej regresji	$s_{\hat{a}}^2 =$	0.00081	
poziom ufności	$P =$	0.95	
błąd współczynnika kierunkowego prostej regresji	$\Delta\hat{b} =$	0.465	0.441
błąd wyrazu wolnego prostej regresji	$\Delta\hat{a} =$	0.063	
wybór równania prostej regresji przy pomocy testu F dla wariancji	$F =$	0.9167	
wartość krytyczna testu F	$F(P, v_1, v_2) =$	2.7876	
decyzja		0	1

Integracja



This type of calculated limits does not provide any significant information for the analyst. More interesting is the confidence interval calculated according to formula (3.20). The results of calculations for simple regressions of the type $y = \hat{b} \cdot x$ from the average band integrations for oxygen, observed after injecting different volumes of it, collected in Table 4.3, are shown below.

TLEN 2023-08-14

$S_{YY} = \sum_i y_i^2 - \bar{y} \cdot \sum_i y_i = \sum_i (y_i - \bar{y})^2$

$S_{XX} = \sum_i x_i^2 - \bar{x} \cdot \sum_i x_i = \sum_i (x_i - \bar{x})^2$

$S_{XY} = \sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i = \sum_i y_i \cdot (x_i - \bar{x})$

$\hat{\sigma}^2 = \frac{S_{YY} - \hat{b} \cdot S_{XY}}{v}$

$s_{\hat{b}}^2 = \frac{\hat{\sigma}^2}{S_{XX}}$ $\quad \Delta\hat{b} = t(P, v) \cdot s_b$

$s_{\hat{a}}^2 = \hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)$ $\quad \Delta\hat{a} = t(P, v) \cdot s_a$

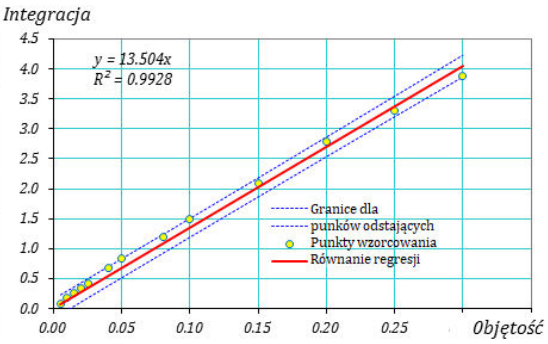
$\Delta\hat{y} = \hat{a} + \hat{b} \cdot x \pm t(\bar{P}, v) \cdot \sqrt{D^2Y}$ $\quad D^2Y = \hat{\sigma}^2 \cdot [\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}}]$

$\Delta\hat{x} = \frac{\frac{\sum_i y_i}{k} - \hat{a}}{\hat{b}} \pm t(\bar{P}, v) \cdot \sqrt{D^2X}$ $\quad D^2X = \frac{\hat{\sigma}^2}{\hat{b}^2} \cdot [\frac{1}{n} + \frac{1}{m} + \frac{(\frac{\sum_k y_k}{m} - \bar{y})^2}{\hat{b}^2 \cdot S_{XX}}]$

WZORCOWANIE									
x_i	y_i	x_i	x_i^2	y_i	y_i^2	$x_i \cdot y_i$	$\hat{y}_i$	$\hat{y}_i + \Delta\hat{y}_i$	$\hat{y}_i - \Delta\hat{y}_i$
0.005	0.07968	0.005	0.000	0.080	0.006	0.000	0.068	0.229	-0.094
0.010	0.17585	0.010	0.000	0.176	0.031	0.002	0.135	0.296	-0.026
0.015	0.26242	0.015	0.000	0.262	0.069	0.004	0.203	0.363	0.042
0.020	0.34459	0.020	0.000	0.345	0.119	0.007	0.270	0.431	0.110
0.025	0.42393	0.025	0.001	0.424	0.180	0.011	0.338	0.498	0.178
0.040	0.66741	0.040	0.002	0.667	0.445	0.027	0.540	0.699	0.381
0.050	0.83517	0.050	0.003	0.835	0.698	0.042	0.675	0.833	0.517
0.080	1.20019	0.080	0.006	1.200	1.440	0.096	1.080	1.237	0.923
0.100	1.50058	0.100	0.010	1.501	2.252	0.150	1.350	1.507	1.193
0.150	2.09007	0.150	0.023	2.090	4.368	0.314	2.026	2.184	1.867
0.200	2.78012	0.200	0.040	2.780	7.729	0.556	2.701	2.864	2.537
0.250	3.31059	0.250	0.063	3.311	10.960	0.828	3.376	3.547	3.205
0.300	3.87806	0.300	0.090	3.878	15.039	1.163	4.051	4.232	3.870
sumy		1.245	0.237	17.549	43.337	3.199			
średnia	0.10	1.35							

liczba punktów pomiarowych	$n =$	13	
równanie regresji		$y = \hat{a} \cdot \hat{b} \cdot x$	$y = \hat{b} \cdot x$
współczynnik kierunkowy prostej regresji	$\hat{b} =$	12.9043	13.5038
wyraz wolny prostej regresji	$\hat{a} =$	0.1141	
wariancja pomiędzy wartościami zmierzonymi (y_i) i obliczonymi (Y_i)	$\hat{\sigma}^2 =$	0.00526	0.00482
wariancja współczynnika kierunkowego prostej regresji	$s_{\hat{b}}^2 =$	0.04467	0.04095
wariancja wyrazu wolnego prostej regresji	$s_{\hat{a}}^2 =$	0.00081	
poziom ufności	$P =$	0.95	
błąd współczynnika kierunkowego prostej regresji	$\Delta\hat{b} =$	0.465	0.441
błąd wyrazu wolnego prostej regresji	$\Delta\hat{a} =$	0.063	
wybór równania prostej regresji przy pomocy testu F dla wariancji	$F =$	0.9167	
wartość krytyczna testu F	$F(P, v_1, v_2) =$	2.7876	
decyzja		0	1

Integracja



This type of calculated limits provides information for the analyst as to whether there are any outliers in the reference curve. The confidence interval calculated according to formula (3.20) for the best-fit least squares method of $y = \hat{b} \cdot x$ regression, estimated from the average band integrations for oxygen observed after injecting different volumes collected in Table 4.3, does not contain any outliers, as all points lie within the limits set by formula (3.20).

Example 5.

For the results of measurements of the field under the bands for different volumes of injected oxygen collected in Table 4.3, simple regression calculations were performed in Example 1. In Example 2, it was established that there is no statistical difference between the variance for the simplified equation $y = \hat{b} \cdot x$ and the full equation $y = \hat{a} + \hat{b} \cdot x$, hence the simplified equation $y = \hat{b} \cdot x$ was adopted to predict the volume of the analyte in the injected sample. In Example 3, the confidence interval for the slope coefficient $\hat{b} \cong 13{,}5 \pm 0{,}5$ was established.

Using equation (3.20), limits for outliers were plotted, showing that all points used to estimate the calibration curve did not contain any outliers.

Under the same conditions, the analysed sample was injected three times, obtaining signals for oxygen at the level of $y_i \in \{3{,}33;\ 3{,}00; 3{,}34\}$. The oxygen volume value V_{O_2} in the analysed sample was $V_{O_2} \equiv \hat{x}_A \cong (0{,}2451 \pm 0{,}007)cm^3$.

Using the syringe dosing method, the uncertainty of the determined concentration ΔC, according to (5.9), will be: $\forall_{C=\frac{\hat{x}}{V}}\ \Delta C = \frac{1}{V} \cdot \Delta\hat{x} + \frac{\hat{x}}{V^2} \cdot \Delta V$. Since the volume of the analysed sample was $V \cong (1{,}00 \pm 0{,}02)cm^3$, the oxygen content in the analysed sample was: $C \cong (24{,}6 \pm 0{,}8)$.

		ANALIZA y_i
		3.33000
		3.30000
		3.34000
	suma	9.97000
	średnia	3.32333
liczba punktów pomiarowych	$m=$	3
		$y = \hat{a} + \hat{b} \cdot x$ / $y = \hat{b} \cdot x$
wariancja	$D^2\hat{x} =$	0.000012 / 0.000010
błąd oznaczenia	$\Delta\hat{x} =$	0.0078 / 0.0070
objętość tlenu	$\hat{x} =$	- / 0.2461
Objętość próbki	$V =$	1
Niepewność objętości	$\Delta V =$	0.002
Stężenie	$C=$	- / **24.6**
Niepewność	$\Delta C =$	- / **0.8**

$C = \frac{\hat{x}}{V}$; $\Delta C = |\frac{1}{V}|\Delta\hat{x} + |\frac{\hat{x}}{V^2}|\Delta V$

TLEN 2023-08-14

$S_{YY} = \sum_i y_i^2 - \bar{y} \cdot \sum_i y_i = \sum_i (y_i - \bar{y})^2$

$S_{XX} = \sum_i x_i^2 - \bar{x} \cdot \sum_i x_i = \sum_i (x_i - \bar{x})^2$

$S_{XY} = \sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i = \sum_i y_i \cdot (x_i - \bar{x})$

$\hat{\sigma}^2 = \frac{S_{YY} - \hat{b} \cdot S_{XY}}{\nu}$

$s_{\hat{b}}^2 = \frac{\hat{\sigma}^2}{S_{XX}}$ $\quad \Delta\hat{b} = t(P, \nu) \cdot s_b$

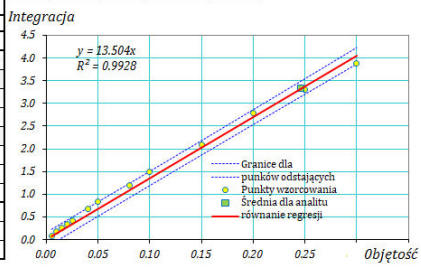
$s_{\hat{a}}^2 = \hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)$ $\quad \Delta\hat{a} = t(P, \nu) \cdot s_a$

$\Delta\hat{y} = \hat{a} + \hat{b} \cdot x \pm t(\bar{P}, \nu) \cdot \sqrt{D^2Y}$ $\quad D^2Y = \hat{\sigma}^2 \cdot [\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}}]$

$\Delta\hat{x} = \frac{\frac{\sum_i y_i}{m} - \hat{a}}{\hat{b}} \pm t(\bar{P}, \nu) \cdot \sqrt{D^2X}$ $\quad D^2X = \frac{\hat{\sigma}^2}{\hat{b}^2} \cdot [\frac{1}{n} + \frac{1}{m} + \frac{(\frac{\sum_i y_i}{m} - \bar{y})^2}{\hat{b}^2 \cdot S_{XX}}]$

WZORCOWANIE										
x_i	y_i	x_i	x_i^2	y_i	y_i^2	$x_i \cdot y_i$	$\hat{y}_i$	$\hat{y}_i + \Delta\hat{y}_i$	$\hat{y}_i - \Delta\hat{y}_i$
0.005	0.07968	0.005	0.000	0.080	0.006	0.000	0.068	0.229	-0.094
0.010	0.17585	0.010	0.000	0.176	0.031	0.002	0.135	0.296	-0.026
0.015	0.26242	0.015	0.000	0.262	0.069	0.004	0.203	0.363	0.042
0.020	0.34459	0.020	0.000	0.345	0.119	0.007	0.270	0.431	0.110
0.025	0.42393	0.025	0.001	0.424	0.180	0.011	0.338	0.498	0.178
0.040	0.66741	0.040	0.002	0.667	0.445	0.027	0.540	0.699	0.381
0.050	0.83517	0.050	0.003	0.835	0.698	0.042	0.675	0.833	0.517
0.080	1.20019	0.080	0.006	1.200	1.440	0.096	1.080	1.237	0.923
0.100	1.50058	0.100	0.010	1.501	2.252	0.150	1.350	1.507	1.193
0.150	2.09007	0.150	0.023	2.090	4.368	0.314	2.026	2.184	1.867
0.200	2.78012	0.200	0.040	2.780	7.729	0.556	2.701	2.864	2.537
0.250	3.31059	0.250	0.063	3.311	10.960	0.828	3.376	3.547	3.205
0.300	3.87806	0.300	0.090	3.878	15.039	1.163	4.051	4.232	3.870
		1.245	0.237	17.549	43.337	3.199			

sumy			
średnia		0.10	1.35
liczba punktów pomiarowych	$n =$	13	
równanie regresji		$y = \hat{a} + \hat{b} \cdot x$	$y = \hat{b} \cdot x$
współczynnik kierunkowy prostej regresji	$\hat{b} =$	12.9043	13.5038
wyraz wolny prostej regresji	$\hat{a} =$	0.1141	
wariancja pomiędzy wartościami zmierzonymi (y_i) i obliczonymi (Y_i)	$\hat{\sigma}^2 =$	0.00526	0.00482
wariancja współczynnika kierunkowego prostej regresji	$s_{\hat{b}}^2 =$	0.04467	0.04095
wariancja wyrazu wolnego prostej regresji	$s_{\hat{a}}^2 =$	0.00081	
poziom ufności	$P =$	0.95	
błąd współczynnika kierunkowego prostej regresji	$\Delta\hat{b} =$	0.465	0.441
błąd wyrazu wolnego prostej regresji	$\Delta\hat{a} =$	0.063	
wybór równania prostej regresji przy pomocy testu F dla wariancji	$F =$	0.9167	
wartość krytyczna testu F	$F(P, \nu_1, \nu_2) =$	2.7876	
decyzja		0	1



CONFIDENCE INTERVAL FOR THE PREDICTION OF AN INDIVIDUAL VALUE

Based on the measured mean values of the measurand $\bar{y}_z = \frac{\sum_i y_i}{m}$ for the analysed sample, the confidence interval for the independent variable $\bar{x}_z = \frac{\bar{y} - \hat{a}}{\hat{b}}$ can be estimated as:

$$\frac{\bar{y}_z - \hat{a}}{\hat{b}} - \frac{t(\bar{P}, \nu = n-1)}{\hat{b}^2} \cdot \sqrt{\hat{\sigma}^2 \cdot [\frac{1}{m} + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}]} \leq$$
$$\leq \bar{x}_z \leq \frac{\bar{y}_z - \hat{a}}{\hat{b}} - \frac{t(\bar{P}, \nu = n-1)}{\hat{b}^2} \cdot \sqrt{\hat{\sigma}^2 \cdot [\frac{1}{m} + \frac{1}{n} + \frac{(\frac{\sum_i y_i}{m} - \bar{y})^2}{S_{XX}}]} \tag{3.21}$$

In equation (3.21), the number of degrees of freedom ν was assumed to be $\nu = n - 1$ corresponding to a regression line passing through the origin of the coordinate system $y = \hat{b} \cdot x$, as only such a regression line is consistent with the theoretical assumptions [6].

Tab. 3.1

Basic mathematical relationships used in regression analysis.

Name	Formula	No
Means	$\bar{x} = \frac{\sum_n x_i}{n}$ $\bar{y} = \frac{\sum_n y_i}{n}$	(2.4)
Sums of squares	$S_{XX} = \sum_i x_i^2 - \bar{x} \cdot \sum_i x_i = \sum_i (x_i - \bar{x})^2$ $S_{YY} = \sum_i y_i^2 - \bar{y} \cdot \sum_i y_i = \sum_i (y_i - \bar{y})^2$	(2.5)
	$S_{XY} = \sum_i x_i \cdot y_i - \bar{y} \cdot \sum_i x_i = \sum_i y_i \cdot (x_i - \bar{x})$	(2.6)
Linear regression model	$Y = a + b \cdot x$ $Y = b \cdot x$	
Estimators	$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$	(2.4)
	$\hat{b} = \frac{S_{XY}}{S_{XX}}$	(2.7)
	Residual	
Residua of	$\hat{\sigma}^2 \triangleq s^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{\nu} = \frac{S_{YY} - \hat{b} \cdot S_{XY}}{\nu}$	(3.7)
intercept	$D^2 \hat{a} \triangleq s_{\hat{a}}^2 = \sigma^2 \cdot (\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}})$	(3.5)
slope coefficient	$D^2(\hat{b}) \triangleq s_{\hat{b}}^2 = \hat{\sigma}^2 \cdot \sum_i c_i^2 = \frac{\sigma^2 \cdot \sum_i (x_i - \bar{x})^2}{S_{XX}^2} = \frac{\hat{\sigma}^2}{S_{XX}}$	(3.2)
estimation Y	$D^2 Y = \hat{\sigma}^2 \cdot (\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}})$	(3.10)
estimation X	$D^2 X = \frac{\hat{\sigma}^2}{\hat{b}^2} \cdot (\frac{1}{n} + \frac{1}{m} + \frac{(\frac{\sum_k y_k}{m} - \bar{y})^2}{\hat{b}^2 \cdot S_{XX}})$	(3.14)
	Confidence interval for estimation of value	
$\hat{a}$	$\Delta \hat{a} = t(P, \nu) \cdot \hat{s}_a$	(3.16)
$\hat{b}$	$\Delta \hat{b} = t(P, \nu) \cdot \hat{s}_b$	
Y	$\Delta Y = \hat{a} + \hat{b} \cdot x \pm t(\bar{P}, \nu) \cdot \sqrt{D^2 Y}$	(3.15)
X	$\Delta X = \frac{\frac{\sum_k y_k}{m} - \hat{a}}{\hat{b}} \pm t(\bar{P}, \nu) \cdot \sqrt{D^2 X}$	

Where: n – number of calibration measurements, x_i – known concentration or other independent variable, y_i – measured signal for the independent variable x_i, $\bar{x}$ – mean for the independent variable x_i ,$\bar{y}$ – mean for the measured values y_i, S_{XX} – sum of squares of deviations from the mean $\bar{x}$ for independent variables x_i, S_{YY} – sum of squares of deviations from the mean $\bar{y}$ for independent variablesy_i, S_{XY} – adjusted sum for the product of independent variables x_i and dependent variables y_i, a – intercept of the regression line, b – slope coefficient of the regression line, $t(\bar{P}, \nu)$ – T-Student coefficient, m – number of replicates of analyte signal measurements in the analysed sample, Y – expected signal value for the analyte when estimating the calibration curve, X – estimated signal value for the analyte in the analysed sample.

CONCLUSION

This chapter presents methods for estimating instrumental analysis results, including confidence intervals for analyte concentrations. In particular, it provides a theoretical introduction to the use of linear regression as a method for processing results using a calibration curve in gas chromatography with a thermoconductometric detector TDC. The error in determining the concentration decreases quite rapidly with the number m of analyses performed, therefore it is recommended that it be at least $m \geq 5$ and the calibration curve should contain at least $n \geq 7$ points within the linear range of TCD the sensor.

Table 3.1 lists useful relationships used in practice for linear regression [30].

Example 6.

In the examples analysed, an EXCEL spreadsheet was used to simplify the calculations. In this way, the tedious calculations that must be performed during calibration and measurements become an almost imperceptible burden, not to mention that they support the process at every stage.

For instance, performing calculations on an ongoing basis makes it possible to eliminate outliers immediately as they appear.

The structure of the spreadsheet, with references to the relationships derived in the text of this study, is shown below.

		ANALIZA	
		y_k	
		y_k	
	suma	$\sum_k y_k$	
	średnia	$\bar{y}_k$	
liczba punktów pomiarowych	$m=$	m	
		$y=\hat{a}\cdot\hat{b}\cdot x$	$y=\hat{b}\cdot x$
wariancja	$D^2X=$	(3.14)	(3.14)
błąd oznaczenia	$\Delta\hat{x}=$	(3.21)	(3.21)
objętość tlenu	$\hat{x}=$	(2.7)	(2.8)
Objętość próbki	$V=$	V	
Niepewność objętości	$\Delta V=$	0.002	
Stężenie	$c=$	(5.9)	(5.9)
Niepewność	$\Delta C=$	(5.9)	(5.9)

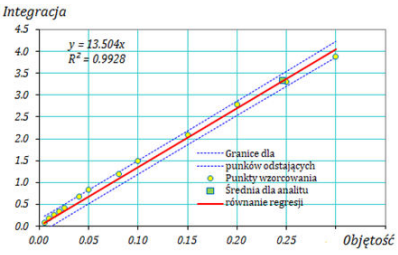
$C=\frac{\hat{x}}{\hat{p}^2}$; $\Delta C=|\frac{1}{\hat{p}}|\Delta\hat{x}+|\frac{\hat{x}}{\hat{p}^2}|\Delta V$

Próbka Data

$$S_{YY}=\sum_i y_i^2-\bar{y}\cdot\sum_i y_i=\sum_i(y_i-\bar{y})^2$$
$$S_{XX}=\sum_i x_i^2-\bar{x}\cdot\sum_i x_i=\sum_i(x_i-\bar{x})^2$$
$$S_{XY}=\sum_i x_i\cdot y_i-\bar{y}\cdot\sum_i x_i=\sum_i y_i\cdot(x_i-\bar{x})$$
$$\hat{\sigma}^2=\frac{S_{YY}-\hat{b}S_{XY}}{\nu}$$
$$s_{\hat{b}}^2=\frac{\hat{\sigma}^2}{S_{XX}} \qquad \Delta\hat{b}=t(P,\nu)\cdot s_b$$
$$s_{\hat{a}}^2=\hat{\sigma}^2\cdot\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right) \qquad \Delta\hat{a}=t(P,\nu)\cdot s_a$$
$$\Delta\hat{y}=\hat{a}+\hat{b}\cdot x\pm t(\bar{P},\nu)\cdot\sqrt{D^2Y} \qquad D^2Y=\hat{\sigma}^2\cdot\left[\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{XX}}\right]$$
$$\Delta\hat{x}=\frac{\frac{\sum_k y_k}{m}-\hat{a}}{\hat{b}}\pm t(\bar{P},\nu)\cdot\sqrt{D^2X} \qquad D^2X=\frac{\hat{\sigma}^2}{\hat{b}^2}\cdot\left[\frac{1}{n}+\frac{1}{m}+\frac{(\frac{\sum_k y_k}{m}-\bar{y})^2}{\hat{b}^2\cdot S_{XX}}\right]$$

WZORCOWANIE									
x_i	y_i	x_i	x_i^2	y_i	y_i^2	$x_i\cdot y_i$	$\hat{y}_i$	$\hat{y}_i+\Delta\hat{y}_i$	$\hat{y}_i-\Delta\hat{y}_i$
x_i	y_i	x_i	x_i^2	y_i	y_i^2	$x_i\cdot y_i$	$\hat{y}_i$	(3.20)	(3.20)
sumy		$\sum_m x_i$	$\sum_m x_i^2$	$\sum_m y_i$	$\sum_m y_i^2$	$\sum_m x_i\cdot$			
średnia		(2.4)	(2.5)						

liczba punktów pomiarowych	$n=$	n	
równanie regresji		$y=\hat{a}\cdot\hat{b}\cdot x$	$y=\hat{b}\cdot x$
współczynnik kierunkowy prostej regresji	$\hat{b}=$	(2.4)	(2.8)
wyraz wolny prostej regresji	$\hat{a}=$	(2.4)	
wariancja pomiędzy wartościami zmierzonymi (y_i) i obliczonymi (Y_i)	$\hat{\sigma}^2=$	(3.7)	(3.7)
wariancja współczynnika kierunkowego prostej regresji	$s_{\hat{b}}^2=$	(3.2)	(3.2)
wariancja wyrazu wolnego prostej regresji	$s_{\hat{a}}^2=$	(3.5)	
poziom ufności	$P=$	0.95	
błąd współczynnika kierunkowego prostej regresji	$\Delta\hat{b}=$	(3.16)	(3.16)
błąd wyrazu wolnego prostej regresji	$\Delta\hat{a}=$	(3.16)	
wybór równania prostej regresji przy pomocy testu F dla wariancji	$F=$	(2.9)	
wartość krytyczna testu F	$F(P,\nu_1,\nu_2)=$	=ROZKŁAD.F.ODW(a;n-1;n-2)	
decyzja		0	1



REFERENCES

1. Allen T.T. 2006. *Introduction to engineering statistics and Six Sigma.* London: Springer Ltd., 2006.
2. Alpert N.D., Keiser W.E., Szymanski H.A. 1974. *Infrared spectroscopy* Warsaw: Państwowe Wydawnictwo Naukowe, 1974.
3. Automotive Industry Task Force. 2010. *Measurement systems analysis Reference Manual.* Fourth Edition. No place: Chrysler Group LLC, Ford Motor Company, General Motors Corporation, 2010. ISBN 978-1-60-534211-5.
4. Balicki A., Makać W. 2000. *Methods of statistical inference.* Gdańsk: Gdańsk University Press, 2000. ISBN 83-7017-727-1.
5. Barrow G.M. 1963. *The structure of molecules.* New York: W.A.Benjamin Inc., 1963.
6. BIML. 2000. *International vocabulary of terms in legal metrology.* GUM 1996. Paris: Bureau International de Métrologie Légale, 2000. Translation: International vocabulary of basic and general terms in metrology.
7. Birkes D., Dodge Y. 1993. *Alternative methods of regression.* Hoboken: Jon Wiley & Sons, Inc., 1993. ISBN: 9780471568810; DOI: 10.1002/9781118150238.
8. Borzymiński J. 2015. *International Vocabulary of Terms in Legal Metrology, translation of OIML V 1 Edition 2013 (E/F). Excerpts from PN-EN ISO/IEC 17000:2006 cited.* Warsaw: Main Office of Measures Publishing House, 2015. ISBN 978-83-940756-1-3.
9. Burdick R.K., Borrow C.M., Montgomery D.C. 2005. *Design and Analysis of Gauge R&R Studies: Making Decisions with Confidence Intervals in Random and Mixed ANOVA Models.* Philadelphia: Society for Industrial and Applied Mathematics, 2005. ISBN 0-89871-588-1.
10. Camman K. 1977. *The use of ion-selective electrodes.* Warsaw: Scientific and Technical Publishing House, 1977.
11. Cowper C.J., J.De Rose. 1989. *Chromatographic analysis of gases.* Warsaw: Scientific and Technical Publishing House, 1989.
12. Czermiński J.B., Iwasiewicz A., Paszek Z., Sikorski A. 1992. *Statistical methods for chemists.* Warsaw: PWN Scientific Publishers, 1992. ISBN: 83-01-06166-9.
13. Danzer K., Than E., Molch D. 1980. *Analytics.* Warsaw: Scientific and Technical Publishing House, 1980.
14. Dettmer-Wilde K., Engewald W. 2014. *Practical Gas Chromatography— A Comprehensive Reference.* Heidelberg: Springer, 2014. ISBN 978-3-642-54639-6 ISBN 978-3-642-54640-2 (eBook) DOI 10.1007/978-3-642-54640-2.
15. Doerffel K. 1989. *Statistics for Analytical Chemists.* Warsaw: Scientific and Technical Publishing House, 1989. ISBN 83-204-1113-0.
16. Ewing W. 1980. *Instrumental methods in chemical analysis.* Warsaw: Państwowe Wydawnictwo Naukowe, 1980.
17. Galton F. 1886. Regression towards mediocrity In hereditary stature. *Journal of the Anthropological Institute.* 1886, Volume 15, pages 246-263.
18. Jakus B., Kłos R. 2007. Measurement process capability. *Polish Hyperbaric Research 19(2007).* 2007, Volume 19, pp. 33-45.
19. Jaworski J., Morawski R., Olędzki J. 1992. *Introduction to metrology and experimental techniques.* Warsaw: Scientific and Technical Publishing House, 1992. ISBN 83-204-1463-6.
20. Kęcki Z. 1998. *Fundamentals of molecular spectroscopy.* Warsaw: Wydawnictwo Naukowe PWN, 1998. ISBN 83-01-10503-8.
21. Kłos R. 2014. *Helium saturation diving – theoretical foundations for diving and training.* 2nd edition (revised). Gdynia: Polish Society of Hyperbaric Medicine and Technology, 2014. ISBN 978-83-938322-1-7.
22. —. 1990. *Method of measuring the composition of breathing mixtures in hyperbaric diving complexes.* Gdynia: Naval Academy, 1990. Doctoral thesis.
23. —. 1990. *Methodology for measuring the composition of breathing mixtures in hyperbaric diving complexes.* Gdynia: Naval Academy, 1990.
24. —. 2006. Uncertainty of measurement results in diving technology. *Polish Hyperbaric Research.* 2006, Vol. 14, 1, pp. 31-48.
25. —. 2015. Assessment of the reliability of the measurement system. *Polish Hyperbaric Research.* 2015, Vol. 51, 2.
26. —. 2008. *Life support systems on submarines.* Gdynia: Polish Society of Hyperbaric Medicine and Technology, 2008. p. 163. ISBN 978-83-924989-4-0.
27. —. 2007. *The use of statistical methods in diving technology – Lecture notes.* Gdynia: Polish Society of Hyperbaric Medicine and Technology, 2007. ISBN 978-83-924989-26.
28. Kłos R.:. 2025. *Rebreathers in Diving Science: Semi-Closed and Closed Circuit Diving Apparatuses.* Abingdon, Oxon: CRC Press (Taylor & Francis Group, LLC), 2025. ISBN: 978-1-032-59608-2 (hbk), ISBN: 978-1-032-64919-1 (pbk), ISBN: 978-1-032-64920-7 (ebk), DOI: 10.1201/9781032649207.
29. Meier P.C., Zund R.E. 2000. *Statistical Methods in Analytical Chemistry.* Second Edition. Chichester: John Wiley & Sons, Inc, 2000. ISBN 0-471-29363-6.
30. Miller J.N., Miller J.C. 2010. *Statistics and Chemometrics for Analytical Chemistry.* Sixth Edition. Harlow: Pearson Education Limited, 2010. ISBN: 978-0-273-73042-2.
31. —. 2016. *Statistics and Chemometrics in Analytical Chemistry.* Warsaw: Wydawnictwo Naukowe PWN, 2016. ISBN: 978-83-01-18303-5.
32. Montgomery D.C. 2009. *Statistical Quality Control.* 6th Edition. New York: John Wiley & Sons Inc., 2009. ISBN 978-0-470-16992-6.
33. Montgomery D.C., Peck E.A., Vining G.G. 2013. *Solutions Manual to Accompany Introduction to Linear Regression Analysis.* Fourth Edition. Hoboken: Wiley & Sons Inc., 2013. ISBN 978-1-118-47146-3.
34. OIMC Geneva. 1996. *International vocabulary of BASIC and general terms in metrology.* Warsaw: Central Office of Measures, 1996. VIM 1984.
35. Ryan A.G. 2013. *Solutions Manual to Accompany Introduction to Linear Regression Analysis Fifth Edition.* Hoboken: John Wiley & Sons, Inc., 2013. ISBN 978-1-118-47146-3.
36. Sadtler Laboratory. Sadtler Standard Spectra.
37. Sheskin D.J. 1997. *Handbook of parametric and nonparametric statistical procedures.* Boca Raton: CRC Press, 1997. ISBN 0-8493-3119-6.

38. Taylor J.R. 1995. *Introduction to measurement error analysis.* Warsaw: Wydawnictwo Naukowe PWN, 1995. ISBN 83-01-11820-2.
39. Thompson B. 1977. *Fundamentals of gas analysis by gas chromatography*. No place: Varian Associates Inc., 1977.
40. UNIPHOS. 2022. Gas Detector Tube Handbook. [Online] 2022. http://WWW.UNIPHOS.EU.
41. Varian. 1979. *CDS 111C Operators Manual.* Palo Alto: Varian Associates Inc., 1979. 85-001142-00.
42. Volk W. 1973. *Applied Statistics for Engineers*. Warsaw: Scientific and Technical Publishing House, 1973. 519.27:54.007.1.

dr hab. inż. Ryszard Kłos, prof. nadzw. AMW
Akademia Marynarki Wojennej im. Bohaterów Westerplatte
Zakład Technologii Prac Podwodnych
81 – 103 Gdynia 3,
ul. Śmidowicza 69
Tel.: +58 626 27 46

[1] Calibration is a set of operations that, under specified conditions, establishes the relationship between the values of a measured quantity indicated by a measuring instrument or measuring system, or the values represented by a measurement standard or reference material, and the corresponding values of the quantity realised by [33,34,35], measurement standards
[2] etalons
[3] measurand
[4] in most cases, it is not possible to carry out the measurement process for the standard and the measurand simultaneously, so the interfering factors in both processes may also be slightly different
[5] Thermal Conductivity Detector TCD
[6] confidence levels are specified by relevant standards and regulations
7 worst-case scenario method
8 deviations from *Lambert-Beer's* law
[9] e.g.: carrier gas flow rates, carrier gas type, measurement temperature, length, diameter and type of column filling, etc.
[10] e.g. Supelco SA, Machery-Nagel GmbH
[11] e.g. J.Chrom.; J.Chrom.Sci.; Gas Chrom.; Chromatographia
[12] as a rule, these are concentrated solutions of various coloured salts
[13] colour
[14] through an appropriate number of suctions
15 reaction substrate
[16] in this case, the outer tube is divided and sealed with elastic tape
[17] enforcement, explanatory variables
[18] system response, explanatory variable
[19] in short: we cannot measure anything accurately
20 linear regression parameters
[21] The probability density function of the normal distribution $\mathbb{N}$ with mean μ and standard deviation σ, or variance σ^2, is given by the formula: $f_{\mu,\sigma}(x) = \frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}\exp\frac{-(x-\mu)^2}{2\cdot\sigma^2}$. The fact that the random variable X has a normal distribution with expected value μ and variance σ^2 is often written as $X\sim\mathbb{N}(\mu,\sigma^2)$, although the notation $X\sim\mathbb{N}(\mu,\sigma)$ is also encountered. If $\mu = 0$ and $\sigma = 1$, then the distribution is called the standard normal distribution, and its density function simplifies to the form: $\phi_{\mu=0,\sigma=1}(x) = \frac{1}{\sqrt{2\cdot\pi}}\exp -\frac{x^2}{2}$
[22] regression (in biology) – a decrease in the diversity and abundance of a group of animals or plants
[23] studies of changes in the average diameter of peas in intergenerational parents□offspring relationships
[24] decrease in the average diameter of peas in offspring
[25] defined by Galton as "regression towards the mean"
[26] formulation of a judgement
[27] linear regression is of great practical importance due to its broad theoretical basis – for example, in linear programming
[28] mathematical hope
[29] the symbol E comes from the French word $espérance$, meaning hope
[30] for (1.8): $D^2(c\cdot X) \equiv E(c\cdot X)^2 - [E(c\cdot X)]^2 = E(c^2\cdot X^2) - (c\cdot EX)^2 = c^2\cdot EX^2 - c^2\cdot(EX)^2 =$
$= c^2\cdot[EX^2 - (EX)^2] = c^2\cdot D^2X$
for (1.9): $\forall_{EX\equiv\mu}\ \ D^2X \equiv E(X-EX)^2 \equiv E(X-\mu)^2 = E(X^2 - 2\cdot\mu\cdot X + \mu^2) =$
$= EX^2 - 2\cdot\mu\cdot EX + \mu^2\cdot E(I) = EX^2 - 2\cdot\mu^2 + \mu^2 \equiv EX^2 - (EX)^2$
[31] In addition to the symbol D^2X, variance is also denoted as: σ_x^2 or $varX$
[32] The variance for N of *Bernoulli* trials can also be calculated in another way. According to the properties of the expected value, we can write: $ES_N^2 = E\left(\sum_i^N X_i^2 + 2\cdot\sum_{i\neq j}^N\sum_{j\neq i}^N X_i\cdot X_j\right) = \sum_i^N EX_i^2 +$
$+2\cdot\sum_{i\neq j}^N\sum_{j\neq i}^N E(X_i\cdot X_j)$. The random variable X_i^2 has the same distribution as the random variable X :$\{(1;\rho),(0;1-\rho)\}$, and it follows that: $EX_i^2 = EX_i = \rho_i$. The random variable $X_i\cdot X_j$ takes the value 1 only when: $X_i = X_j = 1$, i.e. it is determined by the product of independent events: $\{\omega: X_i(\omega) = 1\}\cap$ $\cap\{\omega: X_j(\omega) = 1\}$. These events are independent, thus: $P(X_i\cdot X_j = 1) = P(X_i = 1)\cdot P(X_j = 1)$. Hence: $X_i\cdot X_j$ has the distribution:$\{(1;\rho^2),(0;1-\rho^2)\}$, so $P(X_i\cdot X_j = 1) = 1\cdot\rho^2 + 0\cdot(1-\rho) = \rho^2$. There are $\frac{\{N\cdot(N-1)\}}{2}$ different products of $X_i\cdot X_j$, therefore the above equation takes the form: $ES_N^2 = \sum_i^N\rho_i + 2\cdot\frac{N\cdot(N-1)}{2}\cdot\cdot$ $\sum_i^N\rho_i^2 = N\cdot\rho + N\cdot(N-1)\cdot\rho^2$.
[33] As can be seen from the figure, for a single point in the field, the sums do not add up $(y_C - \bar{y})^2 \neq (\hat{y}_C - \bar{y})^2 + (y_C - \hat{y}_C)^2$. Only the sums for all points for which the regression was calculated can be globalised according to the relationship (1.24) discussed below
[34] for linear regression, the covariance is correlated with the slope of the line: $\forall_{Y=a+b\cdot X}\ \ a = \frac{cov(X,Y)}{var(X)}$
[35] residual
[36] variance
[37] the part of the whole, most often measured by a fractional share or percentage content
[38] removed by correlation
[39] sometimes used to denote proportionality the tilde symbo l~, which is used here to express "has a distribution"
[40] deterministic
[41] stochastic – random; this term is used to distinguish it from the term statistical, which also has a meaning outside mathematical statistics
[42] the tilde ~ is used here to express "has a distribution"
[43] when the sum $S(A_j)$ is equal to some constant value C, a scale shift can be applied where the systematic factor $A_{n,j}$ from equation (**2**) will take the form: $A'_{n,j} = A_{n,j} - C$, and hence the sum $S(A_j)$ will still be equal to zero
[44] the values of the difference estimator $\hat{A}^2$ do not depend on the number of repetitions N for the sample, hence summing n gives N the values $\hat{A}^2 : S(A) = n\cdot\hat{A}^2 = \frac{N}{i-1}\sum_{i=1}^I(\bar{y}_i - \bar{y})^2$
[45] squares of the distances of the measurement points from the line drawn
[46] proof can be found, for example, in [26]
[47] often a different set of values is used based on a different system of equations (2.4), but this is not significant, as the calculations are usually performed using various types of software, e.g. Excel
[48] after logarithming both sides of the exponential function: $\hat{Y} = a\cdot 10^{b\cdot X}$
[49] after taking the logarithm of both sides of the power function: $\hat{Y} = a\cdot X^b$
[50] $\alpha = 1 - \bar{P}$, where$\bar{P}$ represents the confidence level

[51] of total variability
[52] formulation of the adjudicating judgement
[53] r^2 close to 1 indicates that almost all of the variability of the dependent variable can be explained by the independent variables included in the model
[54] sum of squares of deviations of the model from the mean value
[55] total sum of squares of deviations of experimental values from the mean value
[56] because, based on equation (**2.10**) $s_\epsilon = (1 - r^2) \cdot s$
[57] the corrected coefficient of determination $\hat{r}^2$ is always smaller than the coefficient of determination $r^2 : \hat{r}^2 < r^2$
[58] it is biased
[59] a scientific method of inference used to predict how the values of important parameters of a process or event will develop in the future
[60] for the regression equation $y = b \cdot x$, which is a calibration curve, a zero concentration corresponds to a zero response from the measuring device
[62] a blind test is used to exclude the expectation effect, for example, during a clinical trial, it consists in keeping the subjects unaware of whether they are taking the drug or a neutral substance (placebo)
[63] $\nu = n - 1$ for the simplified equation $y = \hat{b} \cdot x$ and $\nu = n - 2$ for the equation $y = \hat{a} + \hat{b} \cdot x$
[64] by multiplying the terms of the sum in sequence, we obtain: $cov\{[\sum_i y_i, \sum_i [y_i \cdot (x_i - \bar{x})]\} = \sum_i [(x_i - \bar{x}) \cdot cov(y_i, y_i)] + + \sum_i [(x_i - \bar{x}) \cdot cov(y_1, y_i)] + \sum_i \{[(x_i - \bar{x}) \cdot cov(y_2, y_i)] + .. + \sum_i [(x_i - \bar{x}) \cdot cov(y_n, y_i)]$, for mixed terms $cov(y_i, y_j) \stackrel{\text{def}}{=} 0 \mid i \neq j$, because the values of the variable $Y \leftarrow y$ are independent (there is no covariance), we can write: $cov\{[\sum_i y_i, \sum_i [y_i \cdot (x_i - \bar{x})]\} = \sum_i [(x_i - \bar{x}) \cdot cov(y_i, y_i)]$
[65] $cov(\bar{y}, \hat{b}) = \frac{\sigma^2}{n \cdot S_{XX}} \cdot \sum_i (x_i - \bar{x}) = \frac{\sigma^2}{S_{XX}} \cdot \frac{\sum_i (x_i - \bar{x})}{n} = \frac{\sigma^2}{S_{XX}} \cdot \frac{n \cdot \bar{x} - \sum_i x_i}{n} = \frac{\sigma^2}{S_{XX}} \cdot (\bar{x} - \bar{x}) = 0$; of course, the sum of deviations $\sum_i (x_i - \bar{x})$ is by definition equal to zero $\sum_i (x_i - \bar{x}) \stackrel{\text{def}}{=} 0$
[66] the derivation will be shown only for the sum, because for the difference the argument is analogous:
$D^2(X + Y) = E[(X + Y - E(X + Y))^2] = E[(X + Y - EX - EY)^2] = E\{[(X - EX) + (Y - EY)]^2\} =$
$= E[(X - EX)^2 + 2 \cdot (X - EX) \cdot (Y - EY) + (Y - EY)^2 = D^2X + D^2Y + 2 \cdot cov(X, Y)$
[67] Here, the measurement population
[68] as in the case of physicochemical analysis of gases using the chromatographic method
[69] regressor
[70] in equation (3.10), the factor ϵ representing random error was not taken into account, as the considerations concern a regression curve without deviations.
[71] Hence, the value of the "uncertainty corridor" is narrowed down to the mean value
[72] characterising, for example, the TCD detector
[73] The question must be answered as to whether the obtained value can still be included in the set that has been used so far to determine the regression line, or whether it should be rejected. It is therefore interesting to consider the average deviation for a single measurement y_i from the mean representing the value Ey. That is, the tolerance limits must be found when the variance of the difference sought $\Delta_i = |y_i - Ey|$. Hence, the variance will be the sum of the variances: $\sigma_\Delta^2 = \sigma^2 + \sigma^2 \cdot [\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}] = \sigma^2 \cdot [1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}]$. It can be seen from this that the derivation presented in the text can be obtained more quickly by this alternative method of reasoning Określono nieprawidłowe źródło. .
[74] for mixed derivatives, the values are zero due to the independence of the parameters: y, $\bar{y}$ and $\hat{b}$
[75] $\nu = n - 1$ for the simplified equation $y = \hat{b} \cdot x$ and $\nu = n - 2$ for the equation $y = \hat{a} + \hat{b} \cdot x$
[76] Using T-Student distribution
[77] $\nu = n - 1$ for the simplified equation$y = \hat{b} \cdot x$ and $\nu = n - 2$ for the equation $y = \hat{a} + \hat{b} \cdot x$
[78] $idem$
[79] errors in determining the intercept $\Delta\hat{a}$ and the slope coefficient $\Delta\hat{b}$
[80] constituting, for example, the characteristics of the TCD detector