https://doi.org/10.3311/PPch.18793



**Periodica Polytechnica Chemical Engineering**, 66(2), pp. 325–339, 2022

# Testing the Fit of Regression Models Estimated with Extremely Small Samples: Application in Pharmaceutical Stability Studies

Máté Mihalovits[1*], Sándor Kemény[1]

[1] Department of Chemical and Environmental Process Engineering, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, H-1111 Budapest, Műegyetem rkp. 3., Hungary
[*] Corresponding author, e-mail: mihalovits.mate@vbk.bme.hu

Received: 18 June 2021, Accepted: 13 August 2021, Published online: 03 January 2022

**Abstract**

Pharmaceutical stability studies are conducted to estimate the shelf life, i.e. the period during which the drug product maintains its identity and stability. In the evaluation of process, regression curve is fitted on the data obtained during the study and the shelf life is determined using the fitted curve. The evaluation process suggested by ICH considers only the case of the true relationship between the measured attribute and time being linear. However, no method is suggested for the practitioner to decide if the linear model is appropriate for their dataset. This is a major problem, as a falsely selected model may distort the estimated shelf life to a great extent, resulting in unreliable quality control. The difficulty of model misspecification detection in stability studies is that very few observations are available. The conventional methods applied for model verification might not be appropriate or efficient due to the small sample size. In this paper, this problem is addressed and some developed methods are proposed to detect model misspecification. The methods can be applied for any process where the regression estimation is performed on independent small samples. Besides stability studies, frequently performed construction of single calibration curves for an analytical measurement is another case where the methods may be applied. It is shown that our methods are statistically appropriate and some of them have high efficiency in the detection of model misspecification when applied in simulated situations which resemble pre-approval and post-approval stability studies.

**Keywords**

trend analysis, detection of model misspecification, stability study, pharmaceutical quality control

## 1 Introduction

According to the definition of FDA (U.S. Food and Drug Administration), shelf life is the period during which a drug product remains suitable for the intended use. The length of this period determined during registration of the drug product is referred to as claimed shelf life. It is required for the products from the on-going manufacturing process to have a period of shelf life the same or longer than that of the claimed shelf life. If this requirement is not fulfilled, a deviation procedure must be initialized by the producer company and the root cause is to be found. Also, regulatory authorities such as the FDA might issue a recall of the concerned products from the market.

Stability studies are conducted to estimate shelf life by monitoring different attributes of the drug products. In the phase of the drug registration, studies are called pre-approval stability studies, while studies in the phase of on-going production are called on-going or post-approval stability studies. When a batch (products manufactured in one process circle) is considered for a stability study, samples are collected from it right after the production. The samples are held under regulated conditions (temperature and humidity) throughout the study, and relevant chemical, biological, and physical attributes are measured at certain time points. The conditions under which the samples are to be held and the measuring frequency can be found in the ICH Q1A Guide [1]. The ICH Q1A Guide [1] considers pre-approval stability studies only, but the general principles can be and convenient to be applied for on-going stability studies as well.

The data collected for each attribute during a stability study are evaluated to estimate the shelf life. The statistical methods that should be used in the evaluation process are

**Cite this article as:** Mihalovits, M., Kemény, S. "Testing the Fit of Regression Models Estimated with Extremely Small Samples: Application in Pharmaceutical Stability Studies", Periodica Polytechnica Chemical Engineering, 66(2), pp. 325–339, 2022. https://doi.org/10.3311/PPch.18793

specified in ICH Q1E Guide [2]. Two statistical methods are considered in the guide: calculation of confidence band for the estimation of shelf life and application of ANCOVA for the test of poolability of regression lines from different stability studies. Mihalovits and Kemény [3] suggested a third method to detect out-of-trend (outlier) points in stability studies. In all three methods, it is required that the regression model fitted on the stability data is appropriate. If the fitted model is not adequate, that is model misspecification is present, the methods are biased, resulting in an unreliable estimation of shelf life and thus unreliable quality control.

The Q1E Guide generally considers lines as appropriate trends, and the statistical methods suggested there assume the adequacy of the linear trends. The Guide mentions the possibility of application of other models:

"The relationship can be represented by a linear or non-linear function on an arithmetic or logarithmic scale. In some cases, a non-linear regression can better reflect the true relationship" ([2]:p.7).

However, it does not advise the practitioners on how to test whether the linear model is appropriate and how to proceed if the linear model is found to be not appropriate. In stability studies, the connection between the measured attribute and time is usually assumed to be linear. That is because it is the simplest model and might be appropriate in a lot of cases. When the measured attribute is the active pharmaceutical ingredient (API) and it is known that its degradation follows zero-order kinetic, the linear trend is appropriate. However, for other attributes, there might not be any theoretical idea to support any initial model to be fitted. In this case, sound statistical methods are required to verify the appropriateness of linear or any other type of model. Although model verification methods (discussed in Section 2) are well settled in the literature, those methods assume a relatively large sample size. In the case of stability studies the sample size is extremely small and the appropriateness and efficiency of the widely applied methods are questionable.

In this paper, new methods are suggested to detect model misspecification. These methods are applicable even when only a few observations are available in a sample. The methods use information from historical samples (e.g. data from historical stability studies) and thus they need to be available. In general, the methods may be applied for any process where the regression estimation is performed on independent small samples. Besides stability studies, frequently performed construction of single calibration curves for an analytical measurement is another case where the methods may be useful. The independent variable is chosen to be time in this paper as it is in the case of stability studies, but it may be substituted with any other quantity depending on the problem (e.g. concentration in calibration curve fitting). It should be noted, that model selection methods are omitted from the discussion, as the aim here is not to select the seemingly best model from candidate models, but to test whether the linear or any other initially fitted model is appropriate (accordance with the established practice in the evaluation of pharmaceutical stability studies).

## 2 Model verification techniques in the literature

In model verification, it is tested whether a candidate model is an appropriate fit for the data. It is checked whether there are detectable discrepancies between the observations and the fitted model at a given probability. Acceptance of the null hypothesis of no discrepancy means that no evidence was found against the adequacy of the model at the given significance level. Rejection of the null hypothesis means that there is (at least one) evidence against the adequacy of the fitted model. A model with linear function can be written as:

$$Y = \beta_0 + \beta_1 x, \tag{1}$$

where $\beta_0$ and $\beta_1$ are the true intercept and slope, respectively, and $Y$ is the expected value at $x$. In this paper, this model is referred to as a linear model (although in statistics linear model is a family of models (that are linear in the parameters), which includes the model defined in Eq. (1)).

Model verification is generally performed by residual analysis. Discussion about residual analysis can be found for example in the works of Pagan and Hall [4], Draper and Smith [5] and Neter et al. [6]. One of the aims of the residual analysis is to test the assumptions of the errors. If the assumption of independent and normally distributed errors is fulfilled (i.e. no evidence is found against it), the fitted model may be considered appropriate. The errors ($\varepsilon_i$) are obtained as $Y - y_i$, where $y_i$ are the measured observations. As the true model, and thus $Y$ is not known and estimated by a sample, only the estimates of the errors, the ordinary residuals ($\hat{\varepsilon}_i$) may be tested. Beside the ordinary residuals, other types of residuals can be used in the analysis. The generally tested residuals are:

- ordinary residuals,
- standardized residuals,
- internally studentized residuals,
- externally studentized residuals.

A brief discussion of these residuals is given in the following. Pierce and Kopecky [7] showed that as the sample size approaches infinity, the distribution of the ordinary residuals approaches the distribution of the errors. Thus, testing the assumptions of the errors by testing the ordinary residuals is appropriate when the sample size is sufficiently large. However, when the sample size is small the distribution of the residuals may differ from that of the error terms to a great extent. The ordinary residuals are not identically distributed as their variance depends on the allocation of the values of the independent variable ($x_i$). When the observation is closer to the centrum of the bulk of the data they have a larger variance, while observations closer to the edge have a smaller variance. Behnken and Draper [8] called it the ballooning effect. Besides not having common distribution, the residuals are correlated [9]. Accordingly, in small samples, ordinary residuals should not be used in the residual analysis.

The standardized residuals ($\hat{\varepsilon}_i^*$) are obtained by dividing the ordinary residuals with the residual standard deviation ($s$):

$$\hat{\varepsilon}_i^* = \frac{\hat{\varepsilon}_i}{s}, \tag{2}$$

where:

$$s = \sqrt{\left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2\right) / (n-p)}, \tag{3}$$

and $n$ is the number of observations used to fit the regression line and $p$ is the number of parameters in the fitted model, equals two in the linear case. If the ordinary residuals were independent and normally distributed variables, standardization would result in a standard normal distributed random variable which would be convenient for outlier detection. However, as the ordinary residuals are not independent and normally distributed variables, the standardized residuals are not standard normal distributed either. Gray and Woodall [10] showed based on Shiffler's work [11] that the standardized residuals are bounded and the bounds depend on the sample size. The bounded distribution of the standardized residuals in a small sample differs to a great extent from the normal distribution of the errors. Thus, the application of standardized residuals in the general residual analysis is inappropriate for small samples.

The internally studentized residuals are obtained by adjusting the standardized residuals in a way that they have unit standard deviation:

$$r_i = \frac{\hat{\varepsilon}_i}{s\sqrt{1-h_i}}, \tag{4}$$

$h_i$ is called the leverage or potential of the $i^{\text{th}}$ observation and obtained by the formula [5]:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{5}$$

where $n$ is the number of observations used to fit the regression line, $x_i$ is the value of the independent variable of the $i^{\text{th}}$ observation, and $\bar{x}$ is the average of $x_i$ values. The presented formula for calculation of $h_i$ is appropriate only if the fitted model is linear. The general calculation of leverage, which includes matrix calculations can be found in the work of Draper and Smith [5]. The internally studentized residuals are bounded (similarly to the standardized residuals) and tau distributed [12]. The bounded tau distribution differs to a great extent from the normal distribution of the errors when the sample size is small. Thus, internally studentized residuals should not be used in residual analysis when the sample size is small.

The fourth type of the generally considered residuals is the externally studentized residuals. The distribution of the externally studentized residuals is more familiar for the practitioners; they follow student's $t$-distribution. The externally studentized residuals ($t_i$) are calculated as:

$$t_i = \frac{\hat{\varepsilon}_{i(i)}}{s_{(i)}\sqrt{1+h_{i(i)}}}, \tag{6}$$

where $s_{(i)}$ is the residual standard deviation estimated from the fit without the $i^{\text{th}}$ observation (hence the $(i)$ notation) and calculated according to Eq. (3) while the given point is left out from the fitting. $\hat{\varepsilon}_{i(i)}$ is the difference between the $i^{\text{th}}$ observation and the regression curve estimated without the $i^{\text{th}}$ observation. $\hat{\varepsilon}_{i(i)}$ is usually called deleted residual. The $\sqrt{1+h_{i(i)}}$ term in Eq. (6) is the adjustment of the deleted residuals so they have a common, unit standard deviation. $h_{i(i)}$ is obtained by the formula:

$$h_{i(i)} = \frac{1}{n_{(i)}} + \frac{(x_i - \bar{x}_{(i)})^2}{\sum_{k=1,k\neq i}^{n}(x_k - \bar{x}_{(i)})^2}, \tag{7}$$

where $n_{(i)}$ is the number of observations used to fit the regression line ($i^{\text{th}}$ observation excluded), $x_i$ is the value of the independent variable of the $i^{\text{th}}$ observation, and $\bar{x}_{(i)}$ is the average of the values of the independent variable of the

observations used in the fitting ($i^{th}$ observation excluded). While the externally studentized residuals are correlated, they have normal distribution which is convenient for residual analysis.

The distribution and behavior of the discussed types of residuals depend on the sample size. Methods in residual analysis which generally assume a sufficiently large sample size may not be appropriate and may be ineffective for small samples. Razali and Wah [13] studied the power of the most widely applied test for normality including the Shapiro-Wilk test [14]. It was shown, that the tests have rather low effectiveness when the sample size is the smallest that is investigated, 10. Different plotting techniques may also be used to visually investigate the behavior of the residuals (e.g. QQ plot [15] and standardized PP plot [16]). However, visual inspections are subjective; any visual discrepancies found by the practitioner should only raise suspicion for the presence of potential discrepancies. Therefore, the decision should not be based on visual inspections only.

Statistical tests which partition the initial dataset and compare certain statistics of the partitions to detect model misspecification usually applied in the literature. The original lack of fit test was developed by Fisher [17]. The test requires to have repeated measurements at any given $x_i$. Instead of including every repetition in the model fitting, the lack of fit test applies only the average of the repeated measurements. This way, two independent estimators of the residual variance can be obtained: one from the fitting (residual sample variance) and one from the repetitions. The latter is referred to as pure error and is independent of the fitted model. If the fitted model is appropriate, the ratio of the residual sample variance and the pure error follows an *F*-distribution. The expected value of the ratio is one, and a significantly greater null statistic indicates lack of fit, i.e. model misspecification. Another test for model misspecification was developed by Utts [18] which is called the rainbow test. The test aims to detect whether curvature is needed to describe the data. In the method, the initial dataset is partitioned into two datasets: one with the observations located around the center of the bulk of the points ( $\bar{x}$ ), and the other with observations on the edges of the bulk of the points. If the fitted model is correct the residual sample variance estimated from each dataset is an unbiased estimator of the residual variance. The general *F* test utilizing the extra sum of squares [6] is applied in the test. When the extra sum of squares method is applied generally, a misspecified model is considered to be incomplete. That is, the fitted model is considered to be a part of the true model and at least one parameter is left out from the fitted model. Due to the assumption of this nested nature between the falsely selected and the true model, the *F* test is one-sided.

Robust statistics may also be applied to detect model misspecification. In most cases, robust statistics are used to overcome (to some extent) the effect of present of outliers or heteroskedasticity (non-constant error variance) and deviation from the assumed error distribution which would result in biased estimators [19]. In regression analysis, robust standard error may be used to calculate inferences for the model, e.g. confidence intervals for the parameters and confidence band for the curve. The most-widely used formulas for the robust standard error are based on the work of White [20]. While using the robust standard error to calculate inferences is not a solution when a falsely selected model is fitted, it can be used to test whether model misspecification is present at all [21]. The information matrix test developed by White [22] compare the robust standard error with the standard error obtained using least-squares estimation. The test statistic follows a chi-squared distribution as the sample size approach infinity. For small sample sizes different approaches were suggested [23–25] to obtain critical values for the information matrix test. However, the study of the appropriateness of the different approaches do not concern sample sizes smaller than 25. Thus, the application of the test for sample sizes even smaller is questionable. An extensive discussion about the application and misconceptions about the robust standard error is provided by King and Roberts [21].

Outlier detection is also part of the residual analysis, however, it is not used for model verification. In outlier detection one usually assumes that the form of the fitted model is correct, but the estimation of the parameters in the model might be biased due to outlier points. Some of the developed methods in this paper (discussed in Section 3) are inspired by the technique briefly discussed in the following.

Snedecor and Cochran [26] proposed a test to check if the expected value of a deleted observation at $x_i$ agrees with that of the fitted model at $x_i$. This method tests whether the expected value of the externally studentized residuals is zero. In the approach, an observation is detected as an outlier when:

$$|t_i| > t(1-\alpha/2, \nu), \tag{8}$$

where $t_i$ is the externally studentized residual calculated at $x_i$ and $t(1-\alpha/2, \nu)$ is the upper critical *t*-value with degrees of freedom $\nu$ at $\alpha$ significance level. To detect

outlier observations in a dataset, each point is to be tested. The externally studentized residuals are not independent of each other, and the significance level used in each test is to be adjusted accordingly. The technique that is applied is based on the Bonferroni inequality [27], which states that for $q$ test with $\alpha$ significance level, the probability of falsely rejecting the null hypothesis is increased from the desired value of $\alpha$ to a maximum of $\alpha'$, where $\alpha' = q\alpha$. If the desired overall significance level of outlier detection is 0.05 ($\alpha'$) and there are five observations (five tests to be performed), the critical $t$ values are to be taken at $\alpha = 0.01$.

## 3 Developed methods for detection of model misspecification

The presented methods are applicable in situations in which fitting is performed on extremely small independent samples. For example, if there are three samples (e.g. three stability studies or calibration datasets) each with six observations, the proposed methods can be applied to test the adequacy of the fitted model. It is assumed that the true models of every sample have the same form (e.g. they are all linear models) but it is not required that the parameters of the models agree. In the case of linear models, it means that the true slope or true intercept can vary from sample to sample. Also, the residual variances ($\sigma^2$) need to agree in every sample (does not necessarily mean that estimates of the residual variances ($s^2$) agree) and each sample is required to include at least four points. The assumption of common residual variance can be justified for example when the same analytical measurement (instrument) is used for obtaining observations in each sample, and it also may be tested with Bartlett's test [28].

When the form of the fitted model is not appropriate all the observations are actually outliers compared to the falsely selected model. Accordingly, the goal of the investigation is to detect outlier nature. However, outlier nature can also be present when the model is correct. That is, two situations may be distinguished:

- Outlier nature is present while the form of the fitted model agrees with the form of the true model but the expected values of some observations are biased. This is the theoretical situation when outlier detection methods are used. Observations with such an outlier nature are referred to as ordinary outlier here.
- Outlier nature is present while the form of the fitted model DOES NOT agree with the form of the true model therefore the expected values of the observations are all biased compared to the falsely selected model. That is, the model is misspecified and the outlier nature would not occur if the appropriate model were fitted.

When the sample size is extremely small it is rather difficult to distinguish between cases of ordinary outlier nature and misspecified model.

The idea in the developed methods is the following: observations at each value of the independent variable from the different samples are grouped (observations from every sample at each $x_i$ form a group). For example, in stability studies, observations from the different studies at each time point are grouped, while in the calibration process, observations at the same concentrations are to be grouped. Then, it is tested for each group whether it is an outlier, based on certain statistics derived from the data in the group. If the model is misspecified, each group is in fact outlier, and finding outlier nature by the test is desired. However, it is a possibility that the model is correct but a group is an outlier due to a single ordinary outlier observation in the group. Finding of outlier nature of the group in this situation is undesired. The influence of ordinary outlier observation on the group is smaller and more negligible when the number of non-outlier observations increases in the group. To decrease the chance of false alarm due to ordinary outlier observations it is suggested to consider a group for testing only if at least three observations are available in that group. If one considers only the groups with an even greater number of observations, the chance of detection of model misspecification when it is only an ordinary outlier that is present, further decreases. Fig. 1 illustrates the grouping method: each blue circle marks a group eligible for testing.

The proposed methods are presented for the case of stability studies. Based on ICH Q1A Guide [1] the time points in a stability study are considered to be $x = \{0,3,6,9,12,18,24,36\}$ months. It should be noted that the zero time point is mathematically not appropriate if the fitted model is logarithmic as the logarithm of 0 is not defined. For these cases, it is suggested to adjust the zero time point to the exact time point when the observation is obtained. Another solution is an arbitrary choice of a sufficiently small (greater than zero) time point. For such cases, it is suggested to take the zero time point to be 0.05 months, which equals 1.5 days.

In the following $k(=1,2,\ldots,8)$ represents the $k^{\text{th}}$ element of the values of $x=\{0.05,3,6,9,12,18,26,36\}$. The groups are to be formed and the test statistics are to be calculated for each $k$, provided that there are at least three

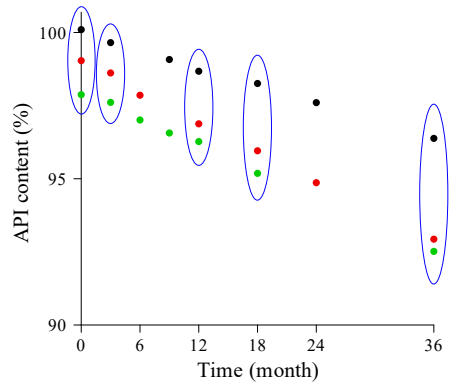


**Fig. 1** Construction of groups: each blue ellipse marks a group eligible for testing

observations at a given time point. In Fig. 1 the eligible groups are those at $x_1$, $x_2$, $x_5$, $x_6$ and $x_8$ ($k=\{1,2,5,6,8\}$).

### 3.1 Testing the averages of the adjusted deleted residuals

The adjusted deleted residuals ($d_k$) are obtained in each sample by the following formula:

$$d_k = \frac{\hat{\varepsilon}_{k(k)}}{\sqrt{1+h_{k(k)}}}, \tag{9}$$

where $h_{k(k)}$ is calculated according to Eq. (7) and $\hat{\varepsilon}_{k(k)}$ is the deleted residual. The adjusted deleted residuals follow a normal distribution with zero mean and residual variance $\sigma^2$ if the fitted model is correct (Section 2). The test statistic is formulated by averaging $d_k$ values at $k$ over the samples and dividing the average by the sample standard error. As the deleted residuals are obtained by excluding the corresponding points from the fits, two independent estimates of the standard error can be obtained. It can be estimated by the sample variance of the adjusted deleted residuals or the residual sample variance obtained in the fitting when observations at $k$ are excluded. When the sample standard error is estimated by the sample variance of the adjusted deleted residuals, the test statistic is obtained as:

$$D_k = \frac{\bar{d}_k}{s_{d_k} / \sqrt{m_k}}, \tag{10}$$

while when it is estimated from the fits, the test statistic is calculated as:

$$D_k^* = \frac{\bar{d}_k}{S_{(k)} / \sqrt{m_k}}, \tag{11}$$

where $\bar{d}_k$ is the average of the adjusted deleted residuals at $k$, $m_k$ is the number of observations at $k$ and $s_{d_k}$ is the sample standard deviation of the adjusted deleted residuals at $k$. When there is no observation in a given sample at $k$, the sample does not contribute to the calculation of $\bar{d}_k$, $m_k$, and $s_{d_k}$. $S_{(k)}$ in Eq. (11) is the square root of the pooled residual sample variance. $S_{(k)}$ is obtained by pooling the residual sample variances obtained from all the fits when observations at $k$ are excluded:

$$S_{(k)} = \sqrt{\frac{\sum_{l=1}^{u}\left[\left(n_{l(k)}-p\right)s_{l(k)}^2\right]}{\sum_{l=1}^{u}\left(n_{l(k)}-p\right)}}, \tag{12}$$

where $u$ is the number of samples (studies), $n_{l(k)}$ is the number of observations in the $l^{\text{th}}$ sample without the observation at $k$, $p$ is the number of parameters to be estimated in the model ($p=2$ for linear models) and $s_{l(k)}^2$ is the residual sample variance in the $l^{\text{th}}$ sample when observation at $k$ is excluded. When there is no available observation in a given sample at $k$, $n_{l(k)}=n_l$ and $s_{l(k)}^2 = s_l^2$.

If the fitted model is correct both $s_{d_k}^2$ and $S_{(k)}^2$ estimate the true error variance ($\sigma^2$) independently from $\bar{d}_k$ and are $\chi^2\sigma^2/\nu\left(s_{d_k}^2\right)$ and $\chi^2\sigma^2/\nu\left(S_{(k)}^2\right)$ distributed, respectively. $\bar{d}_k$ values are normally distributed with expected value of zero and variance of $\sigma^2/m_k$. Accordingly, $D_k$ is a ratio of a normally distributed and a $\sqrt{\chi^2\sigma^2/\nu\left(s_{d_k}^2\right)}$ distributed random variable, while $D_k^*$ is a ratio of a normally distributed and a $\sqrt{\chi^2\sigma^2/\nu\left(S_{(k)}^2\right)}$ distributed random variable. Therefore, both statistics follow a Student's $t$-distribution:

$$D_k \sim t\left(\nu\left(s_{d_k}^2\right) = m_k - 1\right), \tag{13}$$

$$D_k^* \sim t\left(\nu\left(S_{(k)}^2\right) = \sum_{l=1}^{u}\left(n_{l(k)}-p\right)\right). \tag{14}$$

The tests of $D_k$ and $D_k^*$ are two-tailed tests. When $D_k$ or $D_k^*$ exceeds the critical $t$-values with degrees of freedom of $m_k-1$ and $\sum_{l=1}^{u}\left(n_{l(k)}-p\right)$ respectively at a given significance level, the group at $k$ is marked as an outlier, that is model misspecification is present. At each time point where at least three observations are available, the test statistics are to be calculated. The decision is to be made by comparing each of the values of $D_k$ or $D_k^*$ to the critical values. The critical $t$-values may vary from time point to time point due to the varying value of $m_k$ and $n_{l(k)}$. Similarly to the adjusted deleted residuals, the test statistics obtained at different time points are not independent. Considering the Bonferroni adjustment, the critical values

are $t(\alpha'/2q,v)$ and $t(1-\alpha'/2q,v)$, lower and upper values respectively, where $q$ is the number of tests to be performed (the number of available $D_k$ (or $D_k^*$) values). We propose the application of 0.01 overall significance level ($\alpha'$), which results in 0.01 false alarm rate. That is, the probability of falsely rejecting an appropriate model - based on $D_k$ (or $D_k^*$ if that is the applied test statistic) exceeding the critical values in at least one group - is 1 %.

### 3.2 Testing the extra sum of squares when grouped observations are excluded

The extra sum of squares method can be applied to measure the reduction in the error sum of squares when one or more observations are deleted from the fit. The test statistic $ESS_k$ at a given $k$ (provided that $m_k \geq 3$) corresponds to the situation when observations at $k$ are deleted in every sample. $ESS_k$ has an $F$-distribution and is obtained as:

$$ESS_k = \frac{\left(SS - SS_{(k)}\right)/m_k}{SS_{(k)}/\sum\left(n_{l(k)}-p\right)} \sim F\left(m_k, \sum\left(n_{l(k)}-p\right)\right), \quad (15)$$

where $SS$ is the sum of the residual sum of squares from every fit when all observations are included. $SS_{(k)}$ is the sum of the residual sum of squares from every fit when observations at $k$ are excluded, $n_{l(k)}$ is the number of observations in the $l^{\text{th}}$ sample when observations at $k$ are excluded, $m_k$ is the number of samples in which observation is present at $k$ and $p$ is the number of parameters to be estimated in the model. $SS$ is obtained by fitting a regression line on each sample separately and summing the sum of squared residuals from the fits:

$$SS = \sum SS_l. \quad (16)$$

$SS_{(k)}$ is obtained by fitting a regression line in each sample, while the observations at $k$ are excluded and summing the sum of squared residuals from the fits:

$$SS_{(k)} = \sum SS_{l(k)}. \quad (17)$$

When there is no available observation in a sample at $k$, $SS_{l(k)}=SS_l$. Contrary to the generally applied extra sum of squares method (discussed in Section 2), this test is not one-sided. That is because in this case it is not assumed that the falsely selected model is nested into the true model. The test is more general here: it is to be tested whether the form of the fitted model is correct. This more general approach requires a two-sided test. $ESS_k$ values are to be compared to the upper and lower critical $F$-values with degrees of freedom of $m_k$ and $\sum\left(n_{l(k)}-2\right)$ at $\alpha'/q$ significance level.

### 3.3 Over-samples rainbow test

The developed over-samples rainbow test is based on Utts's rainbow test (Section 2). Each sample is divided into two subsamples: one with observations at the earliest and the latest time points (wing observations) and another with observations at the mid-range of the time points (mid-range observations). The idea behind the method is that if the linear model is not correct and there is a curvature, the linear fit on the mid-range points will provide a better fit than the linear fit on the whole range of the points. The test statistic ($ORT$) is obtained as:

$$ORT = \frac{\left(SS - SS_D\right)/\sum\left(n_{lW}\right)}{SS_D/\sum\left(n_l - n_{lW} - p\right)}, \quad (18)$$

and it follows an $F$-distribution:

$$ORT \sim F\left(\sum\left(n_{lW}\right), \sum\left(n_l - n_{lW} - p\right)\right), \quad (19)$$

where $SS$ is the sum of the residual sum of squares from every fit when all observations are included and $SS_D$ is the sum of the residual sum of squares from every sample when the fittings are performed on the mid-range observations. $n_l$ is the number of observations in the $l^{\text{th}}$ sample, $n_{lW}$ is the number of wing observations in the $l^{\text{th}}$ sample and $p$ is the number of parameters to be estimated in the model. When the model is misspecified and there is a curvature, the expected value of the test statistic increases. Accordingly, a one-sided test is to be applied, and the test statistic is to be compared with the upper critical $F$-value at $\alpha$ significance level. The significance level is not required to be adjusted as only one test statistic ($ORT$) is calculated for a dataset.

The selection of the mid-range observation in each sample can be performed in different ways. Two ways are considered here:

- mid-range observations are obtained from the initial sample sets by excluding two observations: one at the earliest and one at the latest time point in each sample; for that, each sample must contain at least 5 observations, so that, at least three mid-range observations are available and fitting can be performed on them. The test statistic is referred to as $ORT$.
- mid-range observations are obtained from the initial sample sets by excluding three observations: one at the earliest and two at the latest two time points in each sample; for that each sample must contain at least 6 observations, so that, at least three mid-range observations are available and fitting can be performed on them. The test statistic is referred to as $ORT^*$.

## 4 Efficiency study of the developed tests

The statistical tests presented in Section 3 may be applied to detect model misspecification. Decisions based on statistical tests are always exposed to the possibility of false decisions. The significance level defines the chance of falsely rejecting the null hypothesis (type I error). In the developed methods 0.01 overall significance level are suggested. This means that when one of the proposed test statistics is applied, the chance of detecting model misspecification when the model is actually appropriate (that is the probability of type I error) is 1 %. The other case of false decision occurs when model misspecification is present but it cannot be detected and the model is accepted as appropriate. This is the type II error and $\beta$ is the rate of its occurrence. The efficiency of the candidate test statistics can be compared by their power which is the probability of detecting model misspecification when it is present. The power is obtained as $1-\beta$. The higher is the power of a test for a given situation the more efficient the test statistic is, i.e. the more frequently it can detect model misspecification when it is present. Power depends on the dataset (number of samples, number of data in each sample, residual variance), the form of the fitted model and the form of the true model as well. Accordingly, a comparison study, which answers the question of which test is the most efficient (has the highest power) in any situation, may not be performed. However, some general insight may be gained about the efficiency of the proposed methods by studying situations that resemble those that occur in the practice. The time points of the measurements are considered to be $x=\{0.05,3,6,9,12,18,24,36\}$ months (based on ICH Q1A Guide [1] with adjustment of 0 month time point to 0.05 month). Two main scenarios are considered in the study:

- Registration: there is a single measurement at each time point, and there are three samples (studies) available. That is, there are 24 observations, 8 in each sample. This situation occurs in the practice at the registration phase of the drug product.
- Post-approval: the number of observations in each sample is limited to six in order to make the situation resembling post-approval stability studies, where missing observations are usually present. Observation at 0.05 month is always available in each sample as it is in the practice, while the time points of the other five observations are randomly generated from the remaining time points (3,6,9,12,18, 24,36 months). Only one observation is allowed at a given time point in a sample. The number of samples considered in the investigation is three and four.

Two situations are considered within each of the described scenarios:

1. A situation that resembles monitoring an active pharmaceutical ingredient (API). The true model is: $Y=97-\log(x)$ for each sample and the fitted model is linear in each sample. Fig. 2 shows the true model, the interval within which 95 % of the observations fall (coverage) when the residual variance ($\sigma^2$) is 1, and the fitted mean line (in the case when there are observations at every time point). The mean line gives the trend of the falsely selected linear model.
2. A situation that resembles monitoring pH. The true model is $Y=6+\exp(-0.06x)$ for each sample and the fitted model is linear in each sample. Fig. 3 shows the true model, the interval within which 95 % of the observations fall (coverage) when the residual variance is 0.05, and the fitted mean line (in the case when there are observations at every time point).

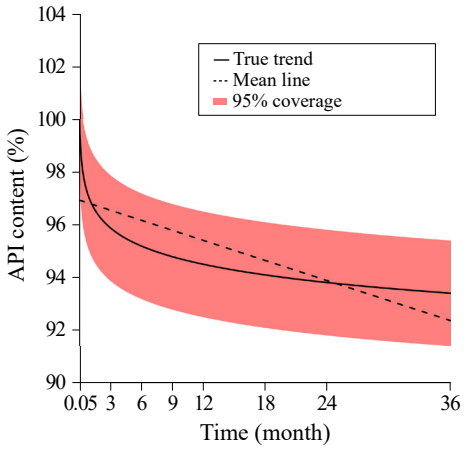


**Fig. 2** True trend curve in the API case with 95 % coverage of potential observations and the fitted mean line

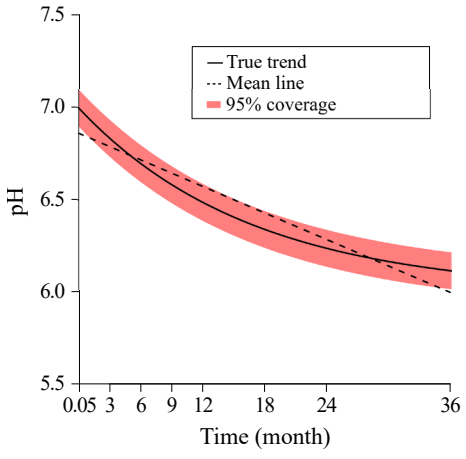


**Fig. 3** True trend curve in the pH case with 95 % coverage of potential observations and the fitted mean line

The effect of the residual variance ($\sigma^2$) is studied in the following way:

- In the API cases, the residual standard deviation ($\sigma$) is considered to be 1 and 0.5.
- In the pH cases, the residual standard deviation ($\sigma$) is considered to be 0.05 and 0.01.

The magnitude of the residual standard deviation was chosen to resemble the practice: in analytical measurements usually, the upper bound for relative standard deviation (RSD) is 1 %. Based on that bound, the standard deviation of the measurements in the range of 90–100 (API case) and 6–7 (pH case) is 1 and ~0.05, respectively. It should be noted that the variation of the observations is caused by the analytical measurement and the sampling as well. Therefore, an even larger variation of observations might be observed in the practice.

Each situation was simulated 100,000 times and the tests based on the test statistics defined in Section 3 ($D_k$, $D_k^*$, $ESS_k$, $ORT$ and $ORT^*$) were performed in each situation with an overall significance level of 0.01 in each test. The power is calculated for each test by counting the cases in which the test was able to detect model misspecification and dividing it by the number of all cases, 100,000. Besides the developed test statistics, the Shapiro-Wilk test on the residuals at 0.01 significance level was also performed in each situation. The test was chosen as it is the most widely applied method in the practice to check the assumption of the normality of the errors, and thus the appropriateness of the fitted model. Also, Dixon's Q test [29] is performed in each situation. Dixon's Q test is a non-parametric method for outlier detection in which the gap between the two largest values (or two smallest values if their gap is larger) is compared to the range of the dataset. The aim is not to detect a single outlier residual, but to detect an outlier group at one of the time points. Thus, the test is performed on the $\bar{d}_k / \sqrt{m_k}$ values, that is the means of the adjusted deleted residuals corrected with $\sqrt{m_k}$, so they have equal variances (this is only required when there are 4 samples). Only the time points at which $m_k \geq 3$ are considered. The critical values for the tests are taken from the work of Rorabacher [30].

### 4.1 Simulation results

An investigation before the efficiency study was carried out to show that the suggested methods are statistically adequate. A method is adequate when the test statistic truly follows its theoretically defined distribution. It can be checked by comparing the significance level with the probability of type I error that is observed in a simulation study. If the observed type I error agrees with the significance level, the method is appropriate. The found probability of type I error is obtained as the number of cases in which the test detects an appropriate model as misspecified, divided by the number of total cases. The type I error depends only on the significance level and it does not depend on the true model, the number of samples and observations, or the value of the residual variance. Three samples were generated in each step, each with the true model being $Y=100+x$ and $\sigma^2=1$. Table 1 shows the probability of type I error found for the different methods, based on 1,000,000 simulated samples of three, which is a large enough sample for the values to be accurate to the first decimal place.

It should be noted that application of Bonferroni inequality (in the cases of $D_k$, $D_k^*$ and $ESS_k$) ensures only that the applied overall significance level is not greater than the nominal one. That is, the expected value of the found probability of type I error may be actually smaller than the nominal one. Despite the presence of this phenomenon, the found probabilities are sufficiently close to the nominal levels, that is the methods are statistically appropriate.

The simulation results of the efficiency study (the probability of detection of model misspecification when it is present, i.e. power (%)) for the registration scenario are shown in Table 2, while for the post-approval scenarios are shown in Tables 3 and 4. In the tables, API refers to the situation in which the true trend of the active pharmaceutical ingredient is $Y=97-\log(x)$, while pH refers to the situation in which the true trend of pH is $Y=6+\exp(-0.06x)$. For example, in the registration phase (i.e. all the 8 data are available at time points of 0.05, 3, 6, 9, 12, 18, 24 and 36 months) at the API case, when $\sigma=1$, the probability of detecting that the falsely selected linear model is not appropriate is 96.4 % when $D_k^*$ is applied (Table 2).

Generally, the Shapiro-Wilk test (SW) is less efficient than the developed methods, and also it is rather ineffective. The only exception is in the registration phase of API when $\sigma=0.5$ (87.6 %). There are cases where the power is smaller than the nominal significance level (1 %). This behavior of the test, which is the result of the residuals

**Table 1** Found type I error of the developed tests

| Nominal (%) | $D_k$ | $D_k^*$ | $ESS_k$ | $ORT$ | $ORT^*$ |
|---|---|---|---|---|---|
| 5 | 4.87 | 4.88 | 4.90 | 4.99 | 5.01 |
| 1 | 1.00 | 0.99 | 1.01 | 1.01 | 1.00 |
| 0.1 | 0.095 | 0.099 | 0.100 | 0.100 | 0.107 |

**Table 2** Power (%) of the tests in the registration case

| Registration | | | | |
|---|---|---|---|---|
| | API | | pH | |
| sigma | 1 | 0.5 | 0.05 | 0.025 |
| $D_k$ | 8.2 | 27.1 | 14.1 | 43.1 |
| $D_k^*$ | 96.4 | >99.9 | 74.5 | 99.4 |
| $ESS_k$ | 70.3 | >99.9 | 15.0 | 28.2 |
| *ORT* | 82.1 | >99.9 | 70.6 | 99.5 |
| *ORT** | 53.7 | 99.7 | 61.7 | 99.0 |
| *SW* | 8.9 | 87.6 | 4.2 | 37.8 |
| *Q* | 2.7 | 0.4 | <0.0 | <0.0 |

**Table 3** Power (%) of the tests in the post-approval case with 3 sample

| Post-approval: 3 samples - 6 observations | | | | |
|---|---|---|---|---|
| | API | | pH | |
| sigma | 1 | 0.5 | 0.05 | 0.025 |
| $D_k$ | 12.3 | 35.5 | 10.5 | 21.5 |
| $D_k^*$ | 88.4 | >99.9 | 32.3 | 45.9 |
| $ESS_k$ | 45.9 | 99.0 | 5.1 | 6.9 |
| *ORT* | 39.0 | 96.1 | 26.8 | 57.6 |
| *ORT** | 8.7 | 35.0 | 9.9 | 28.9 |
| *SW* | 0.9 | 14.8 | 0.7 | 3.5 |
| *Q* | <0.0 | <0.0 | <0.0 | <0.0 |

**Table 4** Power (%) of the tests in the post-approval case with 4 sample

| Post-approval: 4 samples - 6 observations | | | | |
|---|---|---|---|---|
| | API | | pH | |
| sigma | 1 | 0.5 | 0.05 | 0.025 |
| $D_k$ | 28.3 | 78.6 | 22.5 | 44.3 |
| $D_k^*$ | 97.5 | >99.9 | 60.4 | 79.4 |
| $ESS_k$ | 62.7 | >99.9 | 11.1 | 19.2 |
| *ORT* | 59.0 | 99.8 | 42.2 | 82.5 |
| *ORT** | 14.6 | 61.0 | 16.3 | 52.6 |
| *SW* | 1.4 | 16.5 | 0.4 | 0.7 |
| *Q* | <0.0 | <0.0 | <0.0 | <0.0 |

being correlated, was also observed by Weisberg [31]. The very same behavior can be observed regarding Dixon's Q test ($Q$). Even though 5 % significance level is used in these tests, in most cases, the test has undetectable power. The reason for that is presumably the correlation between the residuals (more specifically the correlation between the means of the adjusted deleted residuals), which seems to result in bounded Q test statistics that is smaller than the critical value in most cases. It should be noted, that the Q test is developed for uncorrelated values, and the behavior of the test when correlation between the observations is present, is not investigated in the literature to the best knowledge of the authors of this paper. The Q test may be used on the means of the ordinary residuals (instead of the means of the adjusted residuals) at the different time points as well, however, besides being correlated, the ordinary residuals have different variances, resulting in a Q test statistic with even more distorted behavior.

In the registration scenario (Table 2), it can be concluded that $D_k$ is inefficient in most cases, while $D_k^*$ almost always detect model misspecification in the API case. *ORT* and $ESS_k$ can be considered efficient in the API case with at least 82.1 % and 70.3 % probability of detection of model misspecification, respectively. In the pH case, the tests are generally less efficient. $D_k^*$ and *ORT* perform reasonably well with 74.5 % and 70.6 % probability of detection respectively when $\sigma = 0.05$. When $\sigma = 0.025$ the efficiency of $D_k^*$, *ORT* and *ORT** is more than 99 %. It should be noted that while $ESS_k$ is efficient in the API case, it is inefficient in the pH case. It is difficult to name a single factor which is responsible for the test statistics being more efficient in the API case. The nature of the true curve (logarithmic, exponential, polynomial, etc.), how well the curvature agrees with the trend of the fitted line and how each observation contributes to the shape of the curve might be considered as possible factors.

By comparing the results of the registration case with those of the post-approval case it can be stated that the tests are less efficient in the post-approval situation. It is the effect of having only six observations instead of eight in each sample. More observation makes the discrepancy between the observations and the falsely selected model more detectable. When three samples are available (Table 3) the test statistics have poor efficiency in the pH case. $D_k$ and *ORT** are also inefficient in the API case, while $D_k^*$ performs well with 88.4 % and >99.9 % power. Overall, when four samples are available in post-approval situation (Table 4), the efficiency is higher compared to the cases of three samples. In the case of four samples *ORT** is still inefficient or have low efficiency, while $D_k$ can be considered efficient (78.6 %) in the API case when $\sigma = 0.5$. Generally, in the pH case, the test statistics are inefficient or have low efficiency. However, when four samples are available and $\sigma = 0.025$ $D_k^*$ and *ORT* can be considered efficient with ~80 % detection of model misspecification.

An interesting fact should be pointed out: in the four samples case of the pH situation, at smaller sigma $D_k^*$ is more efficient than *ORT*, while at larger sigma *ORT* is

the more effective one. This phenomenon shows that the power of the different tests is influenced by the nature of the data in different ways, and thus it is difficult to make a general conclusion about the efficiency of the methods.

In this study, the aim was to investigate the effectiveness of the presented tests in extreme situations (i.e. small sample size and rather few samples). It can be concluded that in the investigated situations $D_k^*$ and *ORT* perform reasonably well, and these methods are more effective in the detection of model misspecification, than the Shapiro-Wilk test and Dixon's Q test.

## 5 Application in a stability study of an API content

The application of $D_k^*$ on a real dataset is demonstrated and compared to the results of the generally applied residual analysis in the following. $D_k^*$ is chosen as it was found to be one of the two best methods ($D_k^*$ and *ORT*) in Section 4. The data are obtained in numerous stability studies of a certain API. The values of the API content represent the percent of the nominal content. The data can be found in Table 5.

The linear model is typically the initial one that is fitted by the statistical software used in stability studies to evaluate the stability data. Thus, a linear model was fitted to the data in each study. The coefficients of the fitted models are summarized in Table 6. The observations with the fitted lines are plotted in Figs. 4 and 5. In most cases, the lines seem to be appropriate to describe the observations. The goodness of the fit is not convincing in Study 5 (Fig. 4), however, it should raise concern regarding that sample and not the goodness of the linear fit. The same conclusion can be made regarding the goodness of the linear fits based on the adjusted $R^2$ values (Table 7).

**Table 5** API content (%) from the stability studies

| Time point (month) | Study 1 | Study 2 | Study 3 | Study 4 | Study 5 |
|---|---|---|---|---|---|
| 0 | 98.2 | 99.7 | 99.8 | 101.8 | 99.1 |
| 3 | – | – | – | – | – |
| 6 | 94.3 | 96.9 | 95.8 | 98.2 | 95.3 |
| 9 | – | – | – | – | – |
| 12 | 91.8 | 94.0 | 92.5 | 96.1 | 91.0 |
| 18 | 92.1 | 94.9 | 91.5 | 96.5 | 89.5 |
| 24 | 89.7 | 91.2 | 89.2 | 93.3 | 89.8 |
| 36 | 87.1 | 87.7 | 86.0 | 90.4 | 87.4 |

**Table 6** Coefficients of the fitted linear models

| Coefficient | Study 1 | Study 2 | Study 3 | Study 4 | Study 5 |
|---|---|---|---|---|---|
| Intercept | 96.7 | 99.1 | 98.3 | 100.8 | 96.9 |
| Slope | −0.283 | −0.318 | −0.367 | −0.295 | −0.306 |

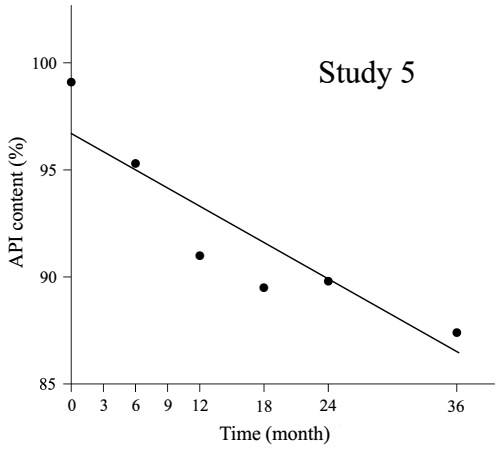


**Fig. 4** Linear fit on the dataset of Study 5

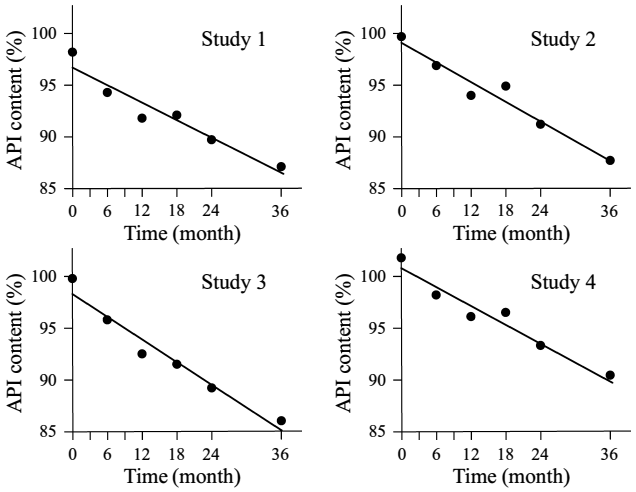


**Fig. 5** Linear fits on the datasets of Study 1–4

Plotting the residuals against the predicted values might help realize the inappropriateness of the fitted model. The plot of the residuals against the predicted values (Fig. 6) shows no sign of discrepancy between the observations and the fitted model; the residuals scatter around zero in a near constant range. The plot of the externally studentized residuals against the predicted values (Fig. 7) however, might raise suspicion that the fitted model is not appropriate or the assumption of the homoscedastic might not be fulfilled. However, the evidence based on eyeballing is not convincing and the decision should not be based only on these plots.

The p-values of the Durbin-Watson test (*DW*) to detect autocorrelation and the Shapiro-Wilk test (*SW*) to detect non-normality of the errors (both performed on

**Table 7** Adjusted $R^2$ and p-values of *SW* and *DW*

| | Study 1 | Study 2 | Study 3 | Study 4 | Study 5 |
|---|---|---|---|---|---|
| adjusted $R^2$ | 0.90 | 0.94 | 0.94 | 0.93 | 0.79 |
| *DW* | 0.442 | 0.462 | 0.074 | 0.952 | 0.002 |
| *SW* | 0.84 | 0.79 | 0.42 | 0.77 | 0.60 |

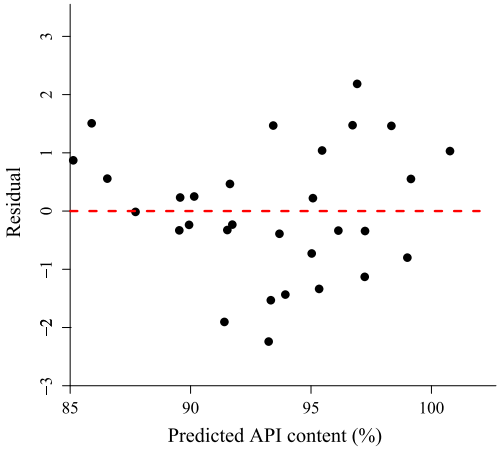


**Fig. 6** Residuals against the predicted values

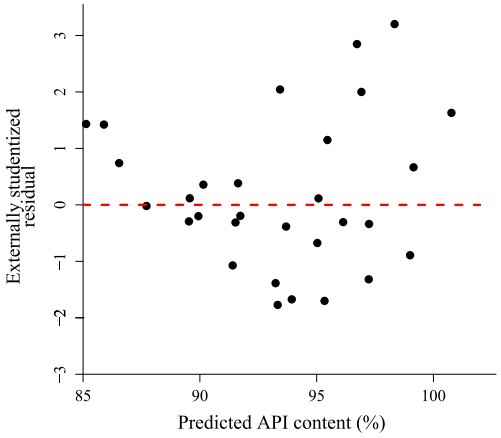


**Fig. 7** Studentized residuals against the predicted values

the ordinary residuals) can be found in Table 7. The small p-value of the DW test confirms that deviation is present in Study 5. Based on the other p-values, deviations (e.g. model misspecification) cannot be detected.

The normality test can be performed on the whole dataset of the residuals and the externally studentized residuals as well. On the normal probability plots (Fig. 8), no discrepancy can be observed between the residuals and the theoretical normal distribution. The Shapiro-Wilk test performed on the residuals and the externally studentized residuals result in p-values of 0.83 and 0.25, respectively. Thus, model misspecification cannot be detected.

Considering the investigation of the residuals, there is no strong evidence against the appropriateness of the fitted linear model. However, further investigation of Study 5 might be required.

Application of $D_k^*$ provides another way to test the appropriateness of the linear models. There are six eligible time points with at least three measurements, and thus six $D_k^*$ values are calculated. The calculated values can be found in Table 8.

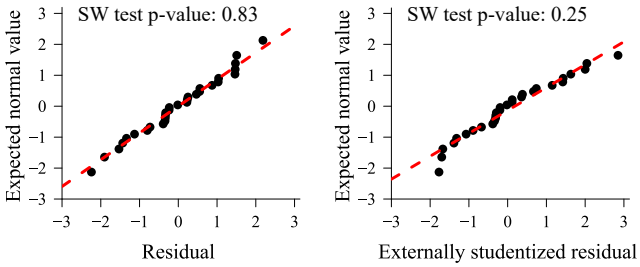


**Fig. 8** Normal probability plots of residuals and studentized residuals

**Table 8** Calculated test statistics

| Time point (month) | 0 | 6 | 12 | 18 | 24 | 36 |
|---|---|---|---|---|---|---|
| $D_k^*$ | 4.04 | −0.70 | −3.29 | 0.31 | −0.35 | 1.81 |

The applied overall significance level is 0.01. To maintain this level, the significance level used in the test of $D_k^*$ in each time is to be corrected according to the Bonferroni adjustment. Thus, the significance level used in each test is 0.01/6. The degrees of freedom in each test is calculated as $\sum_{l=1}^{5}\left(n_{l(k)}-2\right)$, where $n_{l(k)}$ equals five in each case. Thus, the two-sided critical *t*-values with degrees of freedom of 15 and at 0.01/6 significance level equal to ±3.82. $D_k^*$ at the time point of 0 month exceeds the critical value, that is, model misspecification is detected, and a more appropriate model is to be fitted. In reality, there are numerous more investigated stability studies connected to the presented dataset from which it can be deduced (even from the tests of normality of the residuals) that the trends are in fact not linear, but exponential. The aim of this study was to show that in some cases, the suggested methods in this paper may detect more efficiently if model misspecification is present. If the false selection of the linear model is not detected, the estimation of the shelf life, the ANCOVA applied in the estimation of shelf life and the methods used to detect OOT nature are biased, resulting in unreliable quality control.

## 6 Further perspective

It was found in the simulation study that $D_k^*$ and *ORT* are the most efficient tests with generally good efficiency in most of the investigated situations. The results obtained in the study are valid when only the corresponding test is used for a given dataset. It is a reasonable idea to think that by using more than one test statistic simultaneously for checking model misspecification (for example one applies both $D_k^*$ and *ORT*) higher power might be accomplished. However, this is not necessarily true. To uphold the overall significance level when more than one test is used, the individual significance level used in each test (which might have been already adjusted according to the Bonferroni correction) is to be adjusted. The outcome of the adjustment

is increased individual significance level which results in lower power for the given test. That is if more than one test is applied for detection of model misspecification the individual power decreases. However, the overall power is comprised of these individual powers, which might be greater than the power of any single test. The idea of simultaneous application of test statistics might be beneficial but it requires further investigation to obtain a general idea of which tests should be applied together.

The presence of model misspecification and ordinary outlier nature interfere with each other's statistical investigation. Both of the methods assume that the other phenomenon is not present. That is, in model misspecification it is assumed that ordinary outliers are not present, while in outlier detection it is assumed that the form of the fitted model is appropriate. In the presented methods, to decrease the impact of potential ordinary outlier nature, it was recommended to calculate test statistics for a given $k$ only if at least three observations are available at that time point. It is a reasonable idea to investigate how the efficiency and appropriateness of the proposed methods change when ordinary outlier observations are present in the dataset. It is a possibility that a test that now seems less efficient is more robust to the presence of outliers and therefore may be a better choice for detection of model misspecification when it cannot be reasonably assumed that outliers are not present.

Additionally, the modeling of the power of the developed tests could be developed using statistical techniques such as the response surface method. It could help to understand and optimize the response by refining the determinations of relevant factors (e.g. number of observations in a sample, number of samples, allocation of the observations) using the developed model [32]. Also, novel techniques such as machine learning or artificial intelligence may be used [33–35]. On the other hand, it would be of interest to incorporate stochastic modeling, in case of extending the analysis to a greater number of independent variables, in order to have a robust system that allows generating predictions of the response in front of partiality in the independent variables [36]. The results of the mentioned modeling techniques, would allow for redefinition of the allocation of the points and the number of the batches to be measured (in the pre-approval studies) that are defined by the ICH Q1A Guide [1], in order to increase the probability of the detection of model misspecification.

## 7 Conclusion

Pharmaceutical guide Q1E lacks statistical support regarding the detection of model misspecification. It is a difficult task to detect a discrepancy between a falsely selected model and the observations in stability studies due to the small sample size. To address this problem, different statistical tests were presented which may be applied to test the appropriateness of the fitted model. These tests are shown to be statistically appropriate (i.e. found type I error agrees with the nominal level) and some of them are showed to have sufficient efficiency in the investigated situations. The tests based on $D_k^*$ and $ORT$ are found to have reasonably high efficiency. This is especially true when applied in the registration phase of the drug product. Application of these test statistics are of importance in evaluation of stability studies. The bias that can be caused in the estimation of the shelf life by model misspecification is disadvantageous and should be prevented. The authors believe that the presented methods (or those with found to have high efficiency) should be considered as a general step in the evaluation of stability data.

## Acknowledgement

The work by M. Mihalovits was supported by Richter Gedeon Talentum Foundation.

## References

[1] International Conference on Harmonization, Expert Working Group "ICH Harmonised Tripartite Guideline: Stability testing of new drug substances and products Q1A(R2)", In: International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, Brussels, Belgium, 2003, pp. 1–17. [online] Available at: https://database.ich.org/sites/default/files/Q1A%28R2%29%20Guideline.pdf [Accessed: 01 June 2021]

[2] International Conference on Harmonization "ICH Harmonised Tripartite Guideline: Evaluation for stability data Q1E", In: International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, Brussels, Belgium, 2003, pp. 1–8. [online] Available at: https://database.ich.org/sites/default/files/Q1E%20Guideline.pdf [Accessed: 01 June 2021]

[3] Mihalovits, M., Kemény, S. "Regression control chart with unknown parameters for detection of out-of-trend results in pharmaceutical on-going stability studies", Journal of Pharmaceutical and Biomedical Analysis, 188, Article number: 113375, 2020. https://doi.org/10.1016/j.jpba.2020.113375

[4] Pagan, A. R., Hall, A. D. "Diagnostic tests as residual analysis", Econometric Reviews, 2(2), pp. 159–218, 1983. https://doi.org/10.1080/07311768308800039

[5] Draper, N. R., Smith, H. "Applied Regression Analysis", John Wiley & Sons, New York, USA, 1998. https://doi.org/10.1002/9781118625590

[6] Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W. "Applied Linear Statistical Models", McGraw-Hill/Irwin, Chicago, IL, USA, 1996.

[7] Pierce, D. A., Kopecky, K. J. "Testing goodness of fit for the distribution of errors in regression models", Biometrika, 66(1), pp. 1–5, 1979. https://doi.org/10.1093/biomet/66.1.1

[8] Behnken, D. W., Draper, N. R. "Residuals and Their Variance Patterns", Technometrics, 14(1), pp. 101–111, 1972. https://doi.org/10.1080/00401706.1972.10488887

[9] Cook, R. D., Weisberg, S. "Residuals and Influence in Regression", Chapman and Hall, New York, NY, USA, 1982.

[10] Gray, J. B., Woodall, W. H. "The Maximum Size of Standardized and Internally Studentized Residuals in Regression Analysis", The American Statistician, 48(2), pp. 111–113, 1994. https://doi.org/10.1080/00031305.1994.10476035

[11] Shiffler, R. E. "Maximum Z Scores and Outliers", The American Statistician, 42(1), pp. 79–80, 1988. https://doi.org/10.1080/00031305.1988.10475530

[12] Thompson, W. R. "On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation", The Annals of Mathematical Statistics, 6(4), pp. 214–219, 1935. https://doi.org/10.1214/aoms/1177732567

[13] Razali, N. M., Wah, Y. B. "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-darling tests", Journal of Statistical Modeling and Analytics, 2(1), pp. 21–33, 2011.

[14] Shapiro, S. S., Wilk, M. B. "An analysis of variance test for normality (complete samples)", Biometrika, 52(3/4), pp. 591–611, 1965. https://doi.org/10.2307/2333709

[15] Gan, F. F., Koehler, K. J. "Goodness-of-Fit Tests Based on P-P Probability Plots", Technometrics, 32(3), pp. 289–303, 1990. https://doi.org/10.1080/00401706.1990.10484682

[16] Wilk, M. B., Gnanadesikan, R. "Probability plotting methods for the analysis for the analysis of data", Biometrika, 55(1), pp. 1–17, 1968. https://doi.org/10.1093/biomet/55.1.1

[17] Fisher, R. A. "The Goodness of Fit of Regression Formulae, and the Distribution of Regression Coefficients", Journal of the Royal Statistical Society, 85(4), pp. 597–612, 1922. https://doi.org/10.2307/2341124

[18] Utts, J. M. "The rainbow test for lack of fit in regression", Communications in Statistics-Theory and Methods, 11(24), pp. 2801–2815, 1982. https://doi.org/10.1080/03610928208828423

[19] Huber, P. J. "Robust Statistics", John Wiley & Sons, New York, NY, USA, 2004.

[20] White, H. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity", Econometrica, 48(4), pp. 817–838, 1980. https://doi.org/10.2307/1912934

[21] King, G., Roberts, M. E. "How Robust Standard Eerrors Expose Methodological Problems They Do Not Fix, and What to Do about It", Political Analysis, 23(2), pp. 159–179, 2015. https://doi.org/10.1093/pan/mpu015

[22] White, H. "Maximum likelihood estimation of misspecified models", Econometrica, 50(1), pp. 1–25, 1982. https://doi.org/10.2307/1912526

[23] Chesher, A., Spady, R. "Asymptotic expansions of the information matrix test statistic", Econometrica, 59(3), pp. 787–815, 1991. https://doi.org/10.2307/2938228

[24] Davidson, R., MacKinnon, J. G. "A new form of the information matrix test", Econometrica, 60(1), pp. 145–157, 1992. https://doi.org/10.2307/2951680

[25] Horowitz, J. L. "Bootstrap-based critical values for the information matrix test", Journal of Econometrics, 61(2), pp. 395–411, 1994. https://doi.org/10.1016/0304-4076(94)90092-2

[26] Snedecor, G. W., Cochran, W. G. "Statistical Methods", The Iowa State University Press, Ames, IA, USA, 1968.

[27] Galambos, J., Simonelli, I. "Bonferroni-type Inequalities with Applications", Springer, New York, NY, USA, 1996.

[28] Bartlett, M. S. "Properties of Sufficiency and Statistical Tests", Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences, 160(901), pp. 268–282, 1937. https://doi.org/10.1098/rspa.1937.0109

[29] Dean, R. B., Dixon, W. J. "Simplified Statistics for Small Numbers of Observations", Analytical Chemistry, 23(4), pp. 636–638, 1951. https://doi.org/10.1021/ac60052a025

[30] Rorabacher, D. B. "Statistical treatment for rejection of deviant values: critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level", Analytical Chemistry, 63(2), pp. 139–146, 1991. https://doi.org/10.1021/ac00002a010

[31] Weisberg, S. "Some Large-Sample Tests for Nonnormality in the Linear Regression Model: Comment", Journal of the American Statistical Association, 75(369), pp. 28–31, 1980. https://doi.org/10.2307/2287374

[32] Toro, N., Saldaña, M., Castillo, J., Higuera, F., Acosta, R. "Leaching of Manganese from Marine Nodules at Room Temperature with the Use of Sulfuric Acid and the Addition of Tailings", Minerals, 9(5), Article number: 289, 2019. https://doi.org/10.3390/min9050289

[33] Unnikrishnan, S., Donovan, J., Macpherson, R., Tormey, D. "In-process analysis of pharmaceutical emulsions using computer vision and artificial intelligence", Chemical Engineering Research and Design, 166, pp. 281–294, 2021. https://doi.org/10.1016/j.cherd.2020.12.010

[34] Caldararu, O., Blundell, T. L., Kepp, K. P. "Three Simple Properties Explain Protein Stability Change upon Mutation", Journal of Chemical Information and Modeling, 61(4), pp. 1981–1988, 2021. https://doi.org/10.1021/acs.jcim.1c00201

[35] Jiao, Z., Hu, P., Xu, H., Wang, Q. "Machine Learning and Deep Learning in Chemical Health and Safety: A Systematic Review of Techniques and Applications", ACS Chemical Health & Safety, 27(6), pp. 316–334, 2020.
https://doi.org/10.1021/acs.chas.0c00075

[36] Saldaña, M., González, J., Jeldres, R. I., Villegas, Á., Castillo, J., Quezada, G., Toro, N. "A Stochastic Model Approach for Copper Heap Leaching through Bayesian Networks", Metals, 9(11), Article number: 1198, 2019.
https://doi.org/10.3390/met9111198