Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Schriftliche Prüfung
529-0041-00S Moderne MS
Winter 2015

Vorname: _____ Name: _____

- Zeit: 60 Min. Teilen Sie sich Ihre Zeit gut ein.
  Time: 60 min, organize your time carefully.

- Sie können auf Englisch oder Deutsch antworten
  Answers are accepted in German or English.

- Es sind alle Hilfsmittel mit Ausnahme von Computern und Telekommunikation erlaubt.
  It is allowed to use all resources except for computers and communication devices.

- Unleserliche Texte, unklare Formulierungen oder unsaubere Skizzen können nicht bewertet werden. Bitte bemühen Sie sich um eine saubere Darstellung.
  Unreadable text, unclear formulations or graphs are not graded. Please try to use clear illustrations and descriptions

- Schreiben Sie jedes abzugebende Blatt einzeln mit Ihrem Namen und Vornamen an.
  Label every page with name and surname.

- Dieses Deckblatt ist ausgefüllt abzugeben. Die Aufgabenstellung ist ebenfalls einzureichen.
  Please fill in the first page. Hand in all pages including cover page and questions.

- Wir bitten Sie um Fairness und wünschen Ihnen viel Erfolg!
  We ask you for fairness and wish you good luck!

Case Study: Chromatographic Column Performance

The performances of eight commercial chromatographic columns are measured. In order to do this, three compounds were tested, and the results are denoted by a letter (P, Q, A). Four peak characteristics were measured, namely, k (which relates to elution time), N (relating to peak width) and As (asymmetry). Each measurement is denoted by a mnemonic of two halves, the first referring to the compound and the second to the nature of the test. Thus, for example, the measurement PN refers to a peak width measurement on compound P. The resulting matrix is shown in Table 1. Each row represents a measurement and each column represents a chromatographic column. The aim is to ascertain the similarities between the 8 chromatographic columns based on these 9 measured variables (related to the quality of the chromatography).

Parameter	Inertsil ODS	Inertsil ODS-2	Inertsil ODS-3	KromasilC18	KromasilC8	SymmetryC18	Supelco ABZ+	Purospher
Pk	0.25	0.19	0.26	0.3	0.28	0.54	0.03	0.04
PN	10200	6930	7420	2980	2890	4160	6890	6960
PAs	2.27	2.11	2.53	5.35	6.46	3.13	1.96	2.08
Qk	0.25	0.12	0.24	0.22	0.21	0.45	0	0
QN	12000	8370	9460	13900	16800	4170	13800	8260
QAs	1.73	1.82	1.91	2.12	1.78	5.61	2.03	2.05
Ak	2.6	1.69	2.82	2.76	2.57	2.38	0.67	0.29
AN	10700	14400	11200	10200	13800	11300	11700	7160
AAs	1.21	1.48	1.64	2.03	2.08	1.59	1.65	2.08

Figure 1

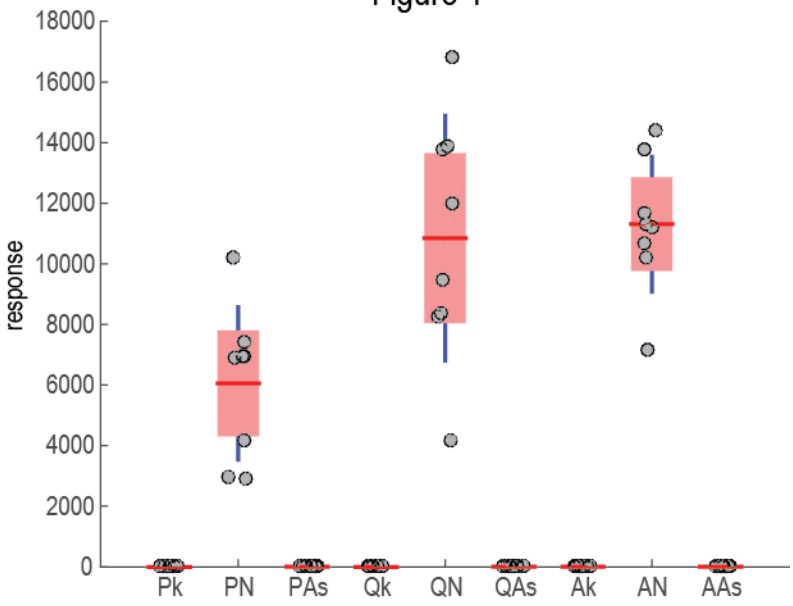
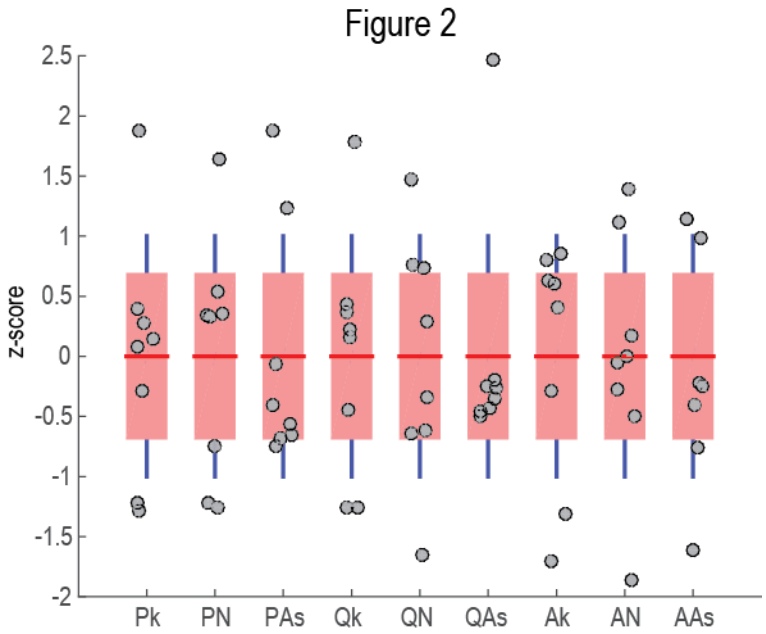


Figure 1 shows a box-plot of the raw data for the 9 measured variables. The measured points are shown together with the mean (red line) a 95% confidence interval of the standard error of the mean (red box) and ± 1 standard deviation (blue bar).

1) Explain in no more than five sentences the information you can extract from this data visualization plot.

Figure 2 shows a box-plot of the centered and scaled data.



2) Explain how centering and scaling is computed and what the advantages of such a data transformation are.

Figure 3 shows the correlation matrix for the centered and scaled data. In the diagonal you can see a histogram of the dataset. The off-diagonal elements show the regression line and Pearson's correlation coefficient employing the model of a 2-dimensional normal distribution.

3) Explain in no more than five sentences the information you can extract from this figure.
4) Explain how peak asymmetry (As) relates to retention time (k) and peak width (N) for compound Q.
5) Based on retention time (k), two of the three compounds are likely to have a similar chemical structure. Which ones? Justify your answer.

Figure 3

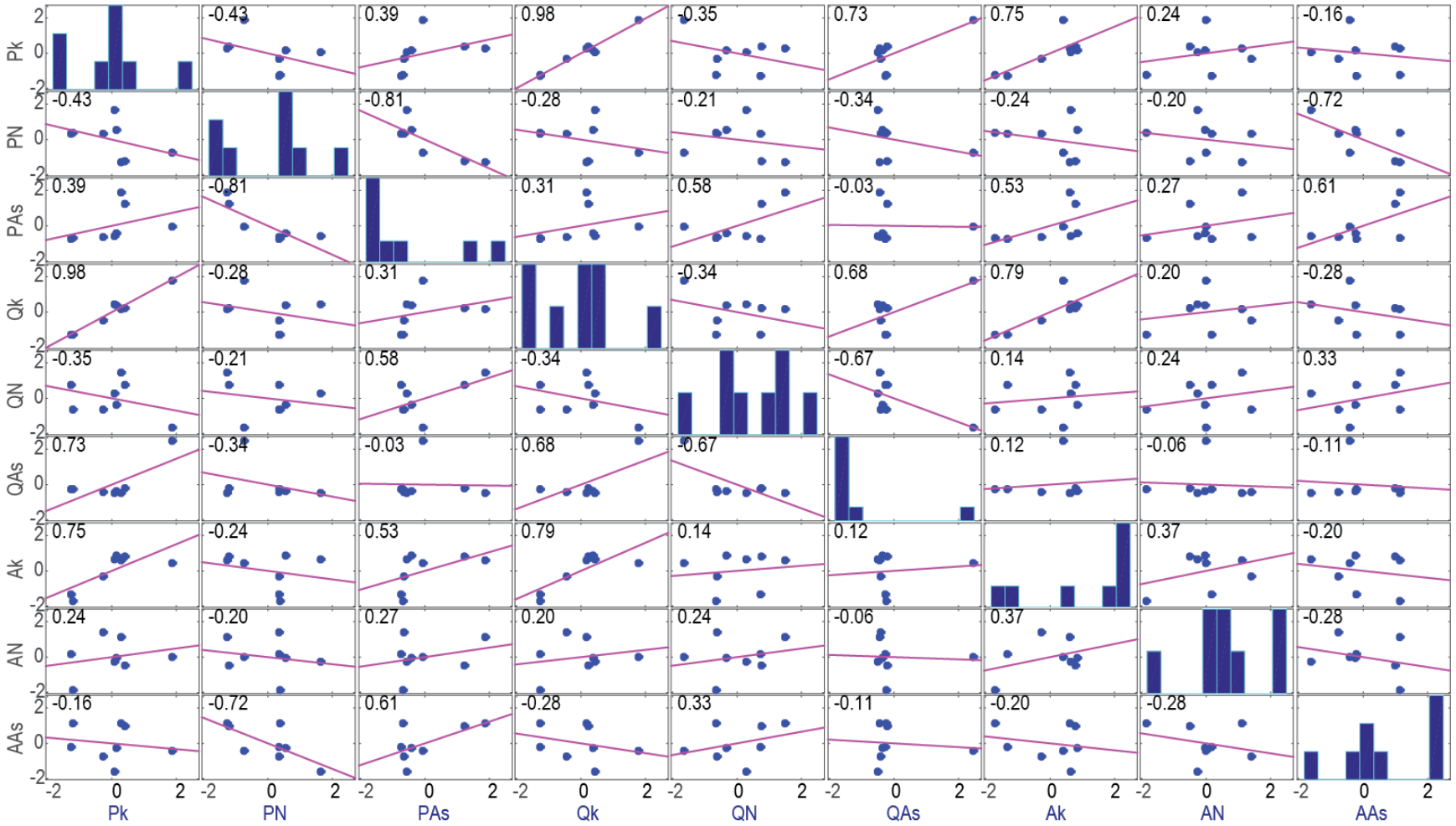

Pk
PN
PAs
Qk
QN
QAs
Ak
AN
AAs
-0.43 0.39 0.98 -0.35 0.73 0.75 0.24 -0.16
-0.43 -0.81 -0.28 -0.21 -0.34 -0.24 -0.20 -0.72
0.39 -0.81 0.31 0.58 -0.03 0.53 0.27 0.61
0.98 -0.28 0.31 -0.34 0.68 0.79 0.20 -0.28
-0.35 -0.21 0.58 -0.34 -0.67 0.14 0.24 0.33
0.73 -0.34 -0.03 0.68 -0.67 0.12 -0.06 -0.11
0.75 -0.24 0.53 0.79 0.14 0.12 0.37 -0.20
0.24 -0.20 0.27 0.20 0.24 -0.06 0.37 -0.28
-0.16 -0.72 0.61 -0.28 0.33 -0.11 -0.20 -0.28
-2 0 2

Figure 4a shows the score plot of the first two principal components after applying principal component analysis (PCA) to the matrix. Figure 4b shows the loadings plot of the second principal component.

6) Explain how PCA transforms the original data and why it is an advantageous method for multivariate problems.
7) Name another unsupervised method to visualize multivariate data.
8) Explain in no more than five sentences what information can you extract from figure 4a
9) Provide an interpretation of figure 4b. Discuss specifically the high loadings for PAs, QN and AAs.

Figure 4a

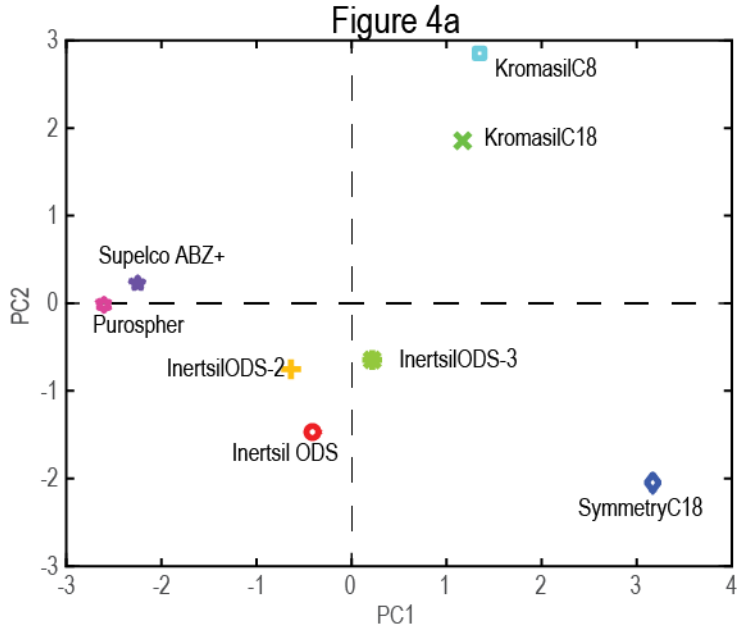
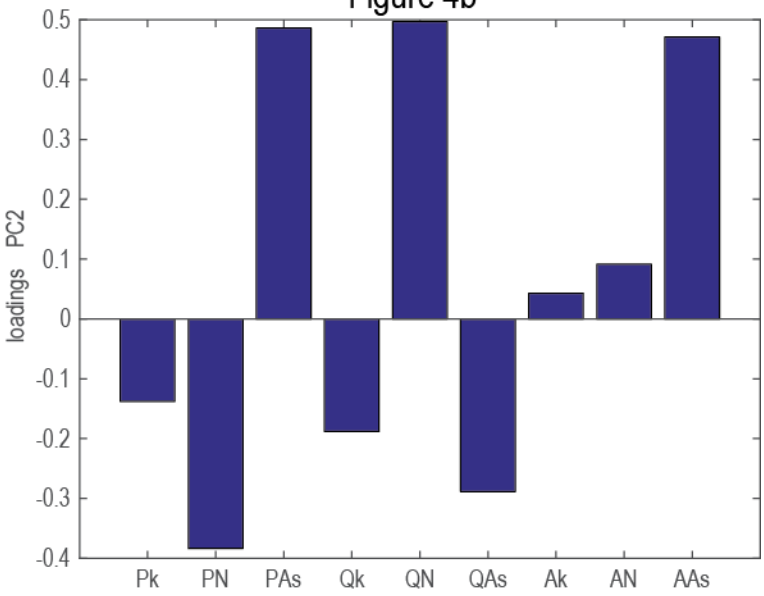


Figure 4b



2. Hard modelling

You are responsible for quality control in a pharmaceutical company. One of the products is a solution of antibiotics. The concentration of one ingredient must be 10 mg/ml with a tolerance of ±0.2 mg/ml. There is a well established and sensitive analytical method relying on fluorescence detection. The analytical instrument has to be calibrated every morning and can then be used the whole day. The signal is proportional to the concentration. The instrument is calibrated in a small range around 10 mg/ml. It is known that all the requirements of Linear Regression are sufficiently satisfied.

The following two sets of data have been measured: (for the sake of brevity all units are omitted)

Data set 1: calibration

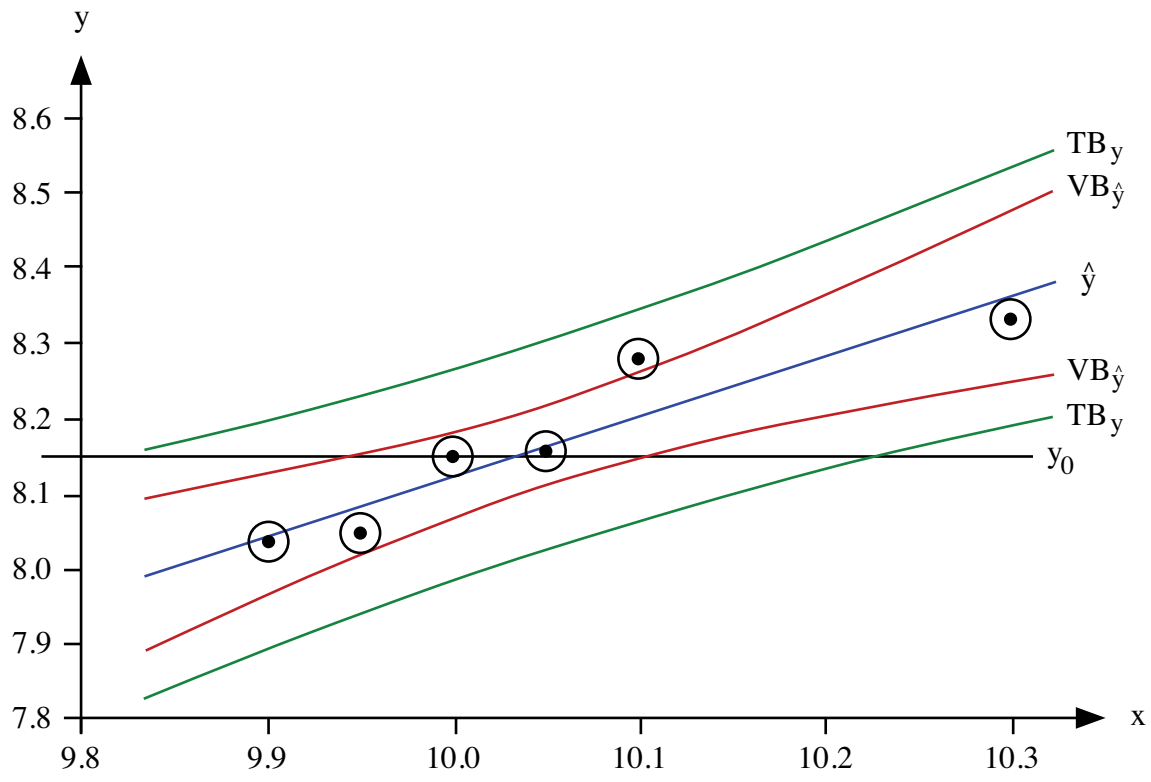
x	y	$\hat{y}$	$y-\hat{y}$ residual	leverage
9.90	8.036	8.0481	–0.0121	1.28212
9.95	8.048	8.0874	–0.0394	1.16775
10.00	8.150	8.1267	0.0233	1.11226
10.05	8.156	8.1660	–0.0100	1.09545
10.10	8.276	8.2053	0.0707	1.11226
10.30	8.330	8.3625	–0.0325	2.19089

where x: concentration
y: fluorescence intensity (arbitrary units)
$\hat{y}$, residual, leverage: see below

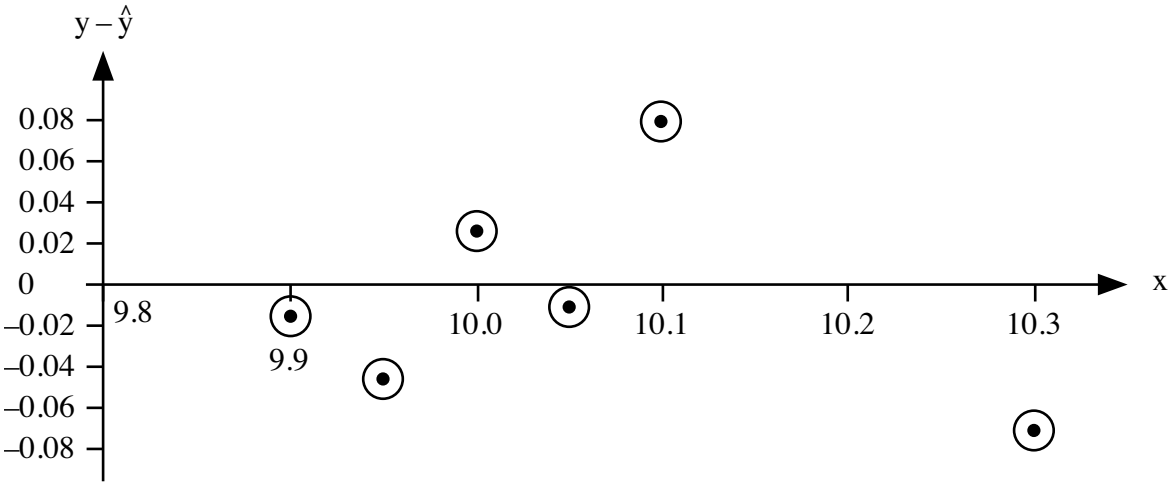
Data set 2: one single measurement of the signal from a sample of unknown concentration

$y_0 = 8.15$

Graphical representation of the data sets:



Graphical representation of the residuals:



The calculations in detail:

Excerpt of the script "Analytische Chemie III, Chemometrie Teil 1"

(Page 66 ff), known variables are replaced by their numerical value.

Estimation of the parameters

Auxiliaries: $\bar{x} = 10.05000$ $\bar{y} = 8.16600$ $\overline{x^2} = 101.0192$

$SS_x = 0.100000$ SS stands for "Sum of squares"

$SS_{xy} = 0.0786000$ (8.24)

Slope: $b = 0.7860000$ (8.25)

Intercept: $a = 0.2667000$ (8.26)

The estimates a and b are random variables. As they are estimated from the same sample, they are correlated. The correlation coefficient $\rho_{a,b}$, a number between –1 and 1, is a measure for the relation between the estimates. It is exactly known as it depends only on the independent variable. Typically it is of little relevance for the interpretation of the results. It may become important in relation to error propagation, when the error of a property has to be calculated that is a function of the parameters α and β.

$\rho_{a,b} = -0.9999175$ (8.27)

Standard deviation of the experimental errors of y:

$s = 0.04581594$ (8.28)

Number of degrees of freedom: $\nu = 4$

Confidence coefficient: 95%
t-factor for 4 degrees of freedom: $t_4 = 2.77645$

Confidence interval VB_b of the slope b:

$$s_b = 0.1448827 \tag{8.29}$$

$$VB_b = 0.7860000 \pm 0.402259 \tag{8.30}$$

Confidence interval VB_a of the intercept a:

$$s_a = 1.456191 \tag{8.31}$$

$$VB_a = 0.26670000 \pm 4.04304 \tag{8.32}$$

Estimate for the expectation value of an individual measurement:

$$\hat{y} = 0.7860000\ x + 0.2667000 \quad \text{(regression line)} \tag{8.33}$$

Confidence interval TB_y of an individual measurement:

$$TB_y = 0.7860000\ x + 0.2667000 \pm 0.127205 \sqrt{\frac{7}{6} + \frac{(x-10.05)^2}{0,1000000}} \tag{8.34}$$

Confidence interval $VB_{\hat{y}}$ of the mean $\hat{y}$:

$$VB_{\hat{y}} = 0.7860000\ x + 0.2667000 \pm 0.127205 \sqrt{\frac{1}{6} + \frac{(x-10.05)^2}{0.1000000}} \tag{8.35}$$

Leverage for the i^{th} measurement:

$$\frac{1}{\sqrt{1 - \frac{101.0192 - 2\,x_i + x_i^2}{0.1000000}}} \tag{8.36}$$

Questions

1. Calculate the unknown concentration x_0 for data set 2. Calculate its 95% confidence interval applying equation (4.16) in the appendix. Assume the standard deviation of y_0 to be the same as for the other y-values, i.e., according to equation (8.28). Don't confuse x_i in equation (4.16) with the concentration values x.

2. Are you convinced that the concentration is within the tolerance?

3. Scrutinize the two data sets. Do you find the calibration appropriate? Does it need improvement or can it be simplified by omitting some measurements? If you find the data sets inappropriate, how would you improve them?

4. Although all requirements for Linear Regression are sufficiently satisfied, could you employ a different model? What would be the advantages and disadvantages?

5. Look at the formula of x_0 that you derived in question 1. There is a t-distribution in the denominator that always covers the value zero. Speculate what the consequences are. Can you imagine a situation when this could be critical?

Appendix: Excerpt of the script "Analytische Chemie III, Chemometrie Teil 1"
(Page 19 ff)

4.6. Error propagation

It is not always possible to directly measure the interesting properties. Often they are calculated from immediately measurable properties. In order to get error limits for these calculated values, it is necessary to describe the influence of the original experimental errors on the calculated values. This topic is called error propagation (Fehlerfortpflanzung).

General:

$$z = f(x_1, x_2, ..., x_n) \qquad (4.13)$$

z : property to be calculated
f : arbitrary known Funktion
$x_1, x_2, ..., x_n$: properties with experimental error

Of every x_i it is known:

$\hat{x}_i$: estimate for x_i
s_i : estimate of the standard deviation of x_i, a measure of the random error. It is calculated with an appropriate statistical model.
Δx_i : estimated systematic error of x_i. This property is difficult to estimate. Often it is set to zero as there is no better idea.

As an estimate for z we use the value $\hat{z} = f(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n)$

Random and systematic errors propagate according to different laws. Therefore, they are treated separately and the contributions are added to the error of z. There is a problem with this procedure. Even if the random errors of all the x_i are distributed according to a normal distribution, this is not generally the case for z. In addition the number of degrees of freedom ν_i associated with the individual s_i may differ. There is no easy mathematical solution for thes difficulties. One may use the smallest occurring ν_i for all the s_i accepting the loss of information. The loss of information can be kept reasonably small by choosing a small confidence coefficient.

The final result is:

$$VB_{\hat{z}} = \hat{z} \pm (s_z t + \Delta z) \qquad (4.14)$$

s_z und Δz are calculated with the following prodedures.

4.6.2. Propagation of random errors

General: Taylor series:

$$s_z^2 = \left(\frac{\partial f}{\partial x_1}\right)^2 s_1^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 s_2^2 + ... + 2\left(\frac{\partial f}{\partial x_1}\right)\left(\frac{\partial f}{\partial x_2}\right)\rho_{1,2}\, s_1\, s_2 + ... \qquad (4.16)$$

The terms linear in s^2 and the pairwise cross terms are taken into account. In case the x_i are uncorrelated, i.e., the correlation coefficients $\rho_{i,j}$ are zero, the series is considerably simplified. However, estimates of parameters are in general correlated if they have been determined with the same sample. The correlation coefficients can most often be calculated in terms of the applied statistical model. Estimates using different samples are generally uncorrelated. The factor of 2 in front of the cross terms take into account the symmetry $\rho_{i,j} = \rho_{j,i}$. Therefore, only terms for $i < j$ are taken.