Kapitola V.

REGRESE A KALIBRACE.

Luděk Dohnal

Volný překlad práce (8).

1. Úvod

Kalibrace je nutná k docílení konsistence měření. Obvykle je její součástí zjišťování závislosti mezi odezvou přístroje a jednou nebo více referenčními (vztažnými) hodnotami. Jakmile je tato závislost určena, slouží k predikci (předpovědi, určení) hodnoty (např. koncentrace) z odezvy přístroje. Tento článek pojednává o nejčastěji používané statistické metodě k modelování závislosti mezi odezvou a koncentrací, t.j. o lineární regresi metodou nejmenších čtverců. Lineární regresní metody se obecně používají k určení nejen přímkových závislostí. Článek se soustřeďuje na praktické použití lineární regrese a interpretaci regresních statistik. Pro ty, kteří se chtějí dozvědět víc o teorii a regresních výpočtech, existují výborné práce, např. (1 - 3).

Poznámka: Každému, kdo zamýšlí používat lineární regresi metodou nejmenších čtverců, lze doporučit používání statisticko-grafického počítačového programu. Tím se m.j. urychlí vytváření grafů potřebných k ověření validity regresních statistik. Ve spreadsheetech (tabulkových kalkulátorech, např. Excel, Lotus) bývají zabudovány funkce, které je rovněž možno používat. Posledně zmíněné lze ovšem připustit pouze za předpokladu, že byly tyto rutiny validovány na přesnost, např. s použitím standardních datových sad (4) a pokud jsou jimi používané algoritmy výpočtu dostatečně podrobně známy.

2. Co to je regrese?

Ve statistice se tento termín používá k vymezení skupiny metod, které popisují stupeň závislosti mezi jedou proměnnou (nebo sadou proměnných) a druhou proměnnou (nebo sadou proměnných). Většinou se používá dnes běžný statistický postup - regrese metodou nejmenších čtverců. Je to nejjednodušší způsob, jak proložit čáru experimentálními body tak, aby byl součet čtverců (druhých mocnin) odchylek experimentálních bodů od této čáry co nejmenší, tedy čáru nejlepší shody (best fit).

Obrázek V.1 ukazuje výsledky typických kalibrací zpracovaných lineární regresí s přímkovým modelem. Plná přímka představuje čáru nejlepší shody proloženou experimentálními body (kroužky). Dále jsou zde uvedeny anebo zakresleny tyto údaje: korelační koeficient (correlation coefficient), odchylky (residuals), směrnice (slope), úsek na ose závisle proměnné (intercept), meze spolehlivosti regresní přímky (confidence limits for the regression line) a meze spolehlivosti predikce (confidence limits for the prediction). Co všechny tyto pojmy znamenají?

3. Korelační koeficient

Ať už se použije kalkulačka, spreadsheet nebo statistický program, většinou pracovníci posuzující výsledky analýz si nejdříve vypočítají korelační koeficient, který bývá označován r. Korelační koeficient se může pohybovat v intervalu od -1 (dokonalá negativní závislost) do +1 (dokonalá positivní závislost). Nula znamená žádnou závislost, viz obrázky V.2a až 2c. Korelační koeficient je mírou POUZE LINEÁRNÍ ZÁVISLOSTI mezi dvěma sadami dat. Korelační koeficient blízký jedničce (nebo minus jedné) je nutnou nikoliv však postačující podmínkou "dobré" lineární závislosti. V praxi bývá velmi často špatně interpretován především z tohoto hlediska (5). Obrázky V.2d a V.2e ukazují případy, kdy samotná hodnota korelačního koeficientu může vést ke zcela mylnému závěru. Je proto nezbytné vždycky zkonstruovat a vizuálně posoudit graf vzájemné závislosti proměnných (tzv. korelační graf).

Podobně jako u t-testů (6), je statistická významnost hodnoty korelačního koeficientu závislá na množství dat n. Ke zjištění, zda konkrétní hodnota r ukazuje na statisticky významnou závislost, můžeme použít testování Pearsonova korelačního koeficientu (7) - viz obrázek V.3. Máme-li jen čtyři body, pak počet stupňů volnosti (number of degrees of freedom, D.F.) je 2 a lineární korelační koeficient (correlation coefficient) metodou nejmenších čtverců r = -0.94 (r^2 = 0.884) není významný na hladině spolehlivosti 95% (95% confidence level). Jestliže ale máme více než 60 bodů, pak např. hodnota r = 0.26 (r^2 = 0.0676) ukazuje na významnou, byť ne příliš silnou, positivní lineární závislost. Je dobré si uvědomit, že závislost, která je statisticky významná, nemusí mít význam praktický. Tento test samotný nedokazuje linearitu nebo dobrou shodu. V praxi je většinou nejlepší vizuálně posoudit data vynesená do grafu a to, jak moc přiléhají k regresní čáře, která je pomocí nich vypočtena.

Je rovněž důležité upozornit, že samotná statisticky významná korelace dvou proměnných nemůže být hodnocena jako indikátor kauzality (příčinné souvislosti) mezi těmito proměnnými. U automobilu existuje např. negativní korelace mezi celkovým počtem ujetých kilometrů a účinností katalyzátoru. Avšak počet ujetých kilometrů není příčinou snižování účinnosti katalyzáto-

ru. Příčinou je usazování sloučenin síry a fosforu, které postupně otráví katalyzátor. Obecně se kauzalita dokazuje jen velmi těžko. Je třeba provést řada systematických experimentů, při nichž je třeba nezávisle na sobě měnit kritické parametry (pokud možno v každém experimentu změnit pouze jeden parametr) a měřit odezvu systému na každou změnu.

4. Směrnice (slope) a úsek na ose závisle proměnné (intercept)

Při lineární regresi se většinou předpokládá vyjádření vztahu mezi X a Y pomocí přímky Y = a + bX (viz obrázek V.1). Přímkový model je vhodný pouze tehdy, když data aspoň přibližně splňují předpoklad lineární závislosti. To můžeme posoudit třeba tak, že data vyneseme do grafu a posuzujeme křivost (např. obrázek V.2d). Anebo vyneseme residua proti predikovaným hodnotám eventuelně proti hodnotám Y (změřeným) - viz obrázek V.4. Residua (residuals) by měla mít normální rozložení, což lze posoudit vizuálně (tvoří přibližně rovnoměrně hustý a pravidelný "mrak" bodů). Jejich normalitu je možno rovněž testovat statisticky (8).

Ačkoliv může být v konkretním případu známo, že závislost je nelineární a je popsána jinou funkcí, např. exponencielou, provádí se někdy "linearizace" dat pomocí transformace proměnné Y a/nebo proměnné X. Typická je např. transformace logaritmická nebo mocninná. (Poznámka o tom, že tato operace pro docílení dobré shody může vyžadovat váženou regresi, je uvedena níže.)

5. Residua a směrodatná odchylka residuí (residual standard error = residual standard deviation)

Residuum je rozdíl mezi predikovanou hodnotou a aktuální (naměřenou) hodnotou (viz obrázek V.1). Graf závislosti residuí vs. predikované (nebo naměřené) hodnoty je mocným diagnostickým nástrojem. Umožňuje odhalit v datech eventuální periodicitu nebo jinou závislost eventuelně možnou odlehlou hodnotu (possible outlier) - obrázek V.4. Je ho možno též použít k odhalení vlivných bodů (viz bias, "leverage" a odlehlé hodnoty).

Směrodatná odchylka residuí (RSE = residual standard error) je statistickou mírou odchýlení dat od regresní čáry. RSE se používá při výpočtu řady užitečných regresních statistik, např. intervalů spolehlivosti a při testování na odlehlé hodnoty.

$$RSE = s_y \sqrt{\frac{(n-1)(1-r^2)}{(n-2)}}$$

6. Intervaly spolehlivosti

Podobně jako většina statistik jsou též směrnice (b) a úsek na ose závisle proměnné (a) odhady, které jsou vypočtené z konečně velkého vzorku, takže jsou zatíženy určitou nejistotou. Přesně řečeno tato, nejistota je důsledkem náhodné variability mezi jednotlivými výběrovými soubory (sadami dat). Mohou existovat též jiné typy nejistoty, např. plynoucí z bias (nesprávnosti) měření. Ty však přesahují rámec tohoto sdělení. Nejistota je ve většině statistických postupů kvantifikována pomocí mezí spolehlivosti a dalších statistik, jako jsou směrodatná odchylka průměru a p-hodnoty. Příklady takových statistik jsou v tab. V.1.

Tabulka V.1
Statistiky získané pomocí funkce Regresní analýza (Excel 97) aplikované na data použitá pro konstrukci kalibračního grafu na obr. V.1.

	koeficienty	"standard error"	t hodnota	p hodnota	95% interval spolehlivosti	
					dolní mez	horní mez
intercept	-0.0460000	0.0396488	-1.1601853	0.2794236	-0.137430	0.0454305
slope	0.1123636	0.0063900	17.5843202	1.1176E-07	0.097628	0.1270990

Všimněte si, na jak velký počet číslic jsou výsledky uváděny. Ve skutečnosti u žádné z výše uvedených hodnot není zaručeno více než 3 platné číslice.
P-hodnota je v tomto případě pravděpodobnost, že nenulová hodnota veličiny vznikla náhodně, jestliže pravá hodnota veličiny je nula. Z tabulky vidíme, že existuje 28% pravděpodobnost, že nenulový intercept je pouze náhoda a současně pouze 0.00001% pravděpodobnost, že nenulový slope je pouze náhoda. Existuje konvence, že p-hodnota menší nebo rovna 0.05 indikuje významně nenulovou statistiku. Takže při posuzování např. výsledků z tab. V.1 vidíme, že není důvod k zamítnutí hypotézy, která říká, že intercept (úsek na svislé ose) je nulový, a současně že je důvod k zamítnutí hypotézy, která říká, že slope (směrnice) je nulová. Jinými slovy, intercept je nevýznamně odlišný od nuly (tedy prakticky nulový), a slope je významně odlišný od nuly (v daném případě něco kolem 0.11).

Interval spolehlivosti pro regresní přímku může být znázorněn pro všechny body podél vodorovné osy. Na obr. V.1 jsou jeho hranice znázorněny tečkovaně. Jak vidíme, prakticky to znamená, že model je určitější v prostřední části, směrem k okrajům jeho určitost klesá. To je důležitý poznatek především pro případ, kdybychom uvažovali o extrapolaci.

Po zkonstruování kalibračního modelu pro analýzu pomocí regrese je kalibrační graf obvykle používán k predikci (předpovídání, určování) hodnot koncentrace (vodorovná osa) z odezvy přístroje (svislá osa). Tato predikce má rovněž svou nejistotu vyjádřenou intervalem spolehlivosti, na obrázku V.1 jsou jeho hranice znázorněny čárkovaně.

$$X_{predicted} = \left(\frac{(\overline{Y} - a)}{b} \right)$$

Interval spolehlivosti predikce je

$$X_{predicted} \pm \left(\frac{t(RSE)}{b} \right) \sqrt{\frac{1}{n} + \frac{1}{n} + \frac{(\overline{Y} - y)^2}{b^2 (n-1) s_x^2}}$$

Jestliže chceme zmenšit velikost intervalu spolehlivosti predikce, můžeme tak učinit několikerým způsobem.

a) Zajistíme, aby se hodnoty odezvy přístroje pro měření, které chceme z kalibrace odečítat, vyskytovaly poblíž středu kalibrační čáry, tedy blízko hodnot $\overline{x}$ a $\overline{y}$. Jinými slovy, když potřebujeme malý konfidenční interval (interval spolehlivosti) pro nízké hodnoty x, potom standardy (kalibrátory, referenční vzorky) použité ke kalibraci je třeba vybrat tak, aby měly hodnoty v oblasti nízkých hodnot x. V analytické chemii je typickou sadou koncentrací kalibrátoru např. 0.05, 0.1, 0.2, 0.4, 0.8, 1.6. Tedy pouze jeden nebo dva použité kalibrátory jsou pro vyšší koncentrace, jestliže potřebujeme co nejužší interval spolehlivosti pro stanovované koncentrace v oblasti cca 0.05 až 0.4. Uvedená posloupnost kalibrátorů vede na jedné straně sice k užšímu intervalu spolehlivosti, na druhé straně je ale kalibrační model stává náchylnější ke vyniku chyb typu "leverage" tedy zkreslení koncovými body (viz tab. V.1).

b) Zvětšíme počet bodů kalibrace (n). Použitím více než 10 kalibračních bodů dosahujeme už pouze malých zlepšení. Proto použití další kalibračních bodů lze obecně doporučit pouze tehdy, jestliže je jejich získání dostatečně jednoduché (rychlé a levné).

c) Zvětšíme počet paralelních stanovení (m) neznámého vzorku. Typicky to bývá 2 až 5 replikátů. Další zvýšení jejich počtu přináší už jen velmi malé zvýšení spolehlivosti.

d) Zvětšíme kalibrační rozmezí pro potvrzení toho, že kalibrační závislost je i nadále lineární.

7. "Bias" (vychýlení, strannost), "leverage" (zkreslení koncovými body) a "outliers" (odlehlé hodnoty)

Vlivné body (influence points), které mohou ale nemusejí být odlehlými hodnotami, mohou mít značný vliv na regresní model a tím také na jeho predikční schopnost. Jestliže bod leží někde uprostřed vzhledem k hodnotám X ale je odlehlý svou hodnotou Y, bude v důsledku jeho přítomnosti posunuta regresní přímka směrem nahoru resp. dolů. Takový bod tedy výrazně ovlivní tzv. off-set nebo bias ve směru osy y a tím ovlivní predikované hodnoty (viz obrázek V.2f). Jestliže ale bod leží při jednom nebo druhém konci regresní přímky ve směru osy x a je odlehlý svou hodnotou Y, dochází jeho vlivem ke změně sklonu regresní přímky. Takový značně vychýlený koncový bod je na obrázku V.2g. Ovlivnění sklonu regresní přímky může být hlavním problémem, jestliže jeden nebo více odlehlých bodů jsou hodně vzdáleny od ostatních ve směru osy x a jsou tedy zkreslujícími koncovými body. Body s "high leverage" zjišťujeme vizuálně.

Pro posouzení, zda je bod odlehlý (vzhledem k intervalu spolehlivosti predikce, který vyplývá z regresního modelu) můžeme použít následující test na odlehlé hodnoty.

$$hodnota\ testu = \frac{\left| residual_{\max} \right|}{RSE \sqrt{1 + \frac{1}{n} + \frac{(y_i - \overline{y})^2}{(n-1) s_y^2}}}$$

Například pro bod podezřelý z odlehlosti (possible outlier) na obrázku V.4 je hodnota tohoto testu 1.78 a kritická hodnota pro 95% spolehlivosti je 2.37 (z tab. V.2 pro n = 10). Přestože nám tento bod připadá odlehlý, z výsledku testu vyplývá, že je rozumné brát jeho vychýlenost jako dílo náhody a že tedy patří do dané datové sady.

Tabulka V.2
Kritické hodnoty pro testování odlehlých hodnot pro jednoduchou lineární regresi metodou nejmenších čtverců.

rozsah souboru n	kritické hodnoty pro spolehlivost 95%	99%	rozsah souboru n	kritické hodnoty pro spolehlivost 95%	99%
5	1,74	1,75	25	2.88	3.25
6	1.93	1.98	30	2.96	3.36
7	2.08	2.17	35	3.02	3.40
8	2.20	2.23	40	3.08	3.43
9	2.29	2.44	45	3.12	3.47
10	2.37	2.55	50	3.16	3.51
12	2.49	2.70	60	3.23	3.57
14	2.58	2.82	70	3.29	3.62
16	2.66	2.92	80	3.33	3.68
18	2.72	3.00	90	3.37	3.73
20	2.77	3.06	100	3.41	3.78

8. Extrapolace a interpolace

Už bylo zmíněno, že regresní čára je nositelem určité nejistoty a že tato nejistota se zvyšuje směrem k jejím koncům. Jestliže se regresní čáru snažíme extrapolovat mimo meze, v nichž leží získaná reálná data, mohou vznikat poměrně velké chyby predikce. Z toho vyplývá, že při konstrukci kalibračního grafu musí kalibrátory pokrývat větší rozsah koncentrací, než je předpokládaný rozsah koncentrací v analyzovaných vzorcích. Jako alternativu je možno použít několik kratších kalibračních grafů, přičemž se vždy rozsahy dvou sousedních částečně překrývají.

9. Vážená lineární regrese a jiné metody kalibrace

V praxi se často stává, že přesnost se mění v závislosti na koncentraci. Nejčastěji se absolutní hodnota směrodatné odchylky zvětšuje se zvětšující se měřenou hodnotou viz obrázek V.5a, závislost odezvy (response) na koncentraci (concentration). Na příslušném grafu residuí - odchylky jednotlivých měření (residuals) vs. predikované hodnoty (predicted value) resp. hodnoty měřené, je to ještě lépe patrné (obrázek V.5b). Jestliže jsou změny přesnosti statisticky významné, pak je vhodné ke konstrukci kalibrační závislosti použít vážené lineární regrese.

Většina statistických programových balíků umí váženou regresi počítat. Vytvořený regresní model je obvykle podobný modelu nevážené regrese pro "dobrá" data, tedy taková data, která nevykazují významné změny přesnosti se vzrůstající měřenou hodnotou. Interval spolehlivosti predikce je ovšem jiný.

Jiné často používané metody kalibrace jsou jednobodová kalibrace a tzv. "bracketing". Tyto postupy jsou popsány např. v ISO 11095 (3).

10. Závěrem

1) Vždycky vyneste data do grafu. Nespoléhejte se při posuzování závislosti pouze na regresní statistiky. Kupříkladu samotný korelační koeficient není spolehlivým ukazatelem dobré shody.

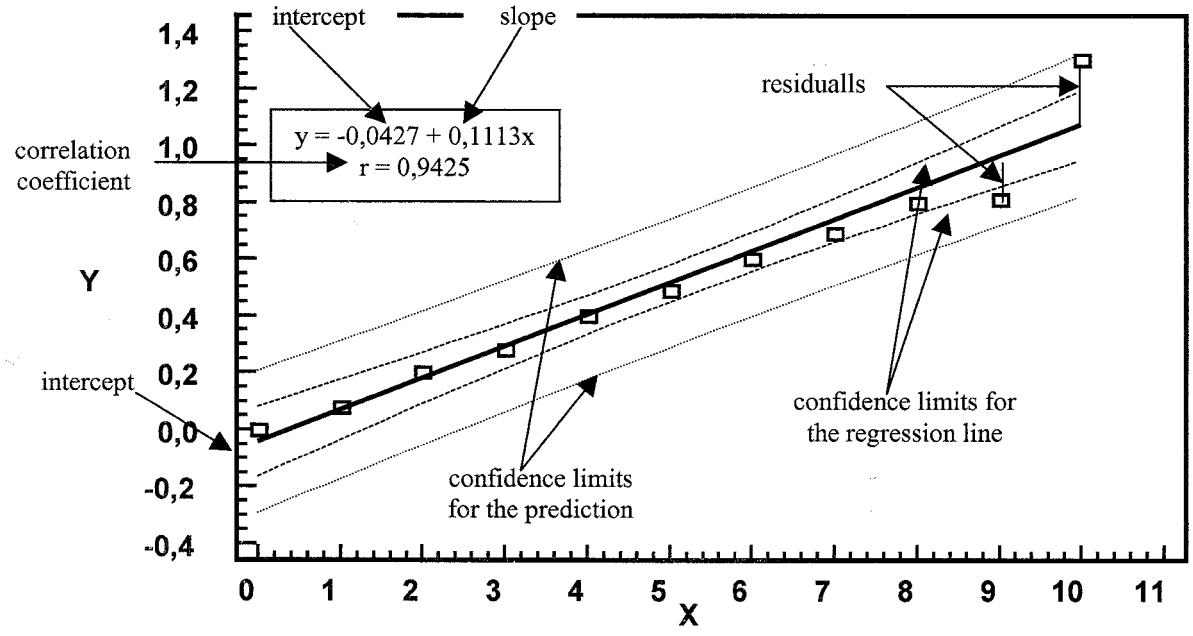
2) Vždy zkonstruujte graf residuí a vezměte jej do úvahy při posuzování závislosti. Je to cenný diagnostický nástroj.

3) Vyloučete vlivné body (typu "leverage", typu "bias" a odlehlé body) pouze tehdy, jestliže lze určit příčiny jejich vzniku.

4) Uvědomte si, že regresní čára je čárou, která nejlépe vyjadřuje závislost dat. Je třeba vzít v úvahu, že tato čára má svoji určitou nejistotu resp. neurčitost.

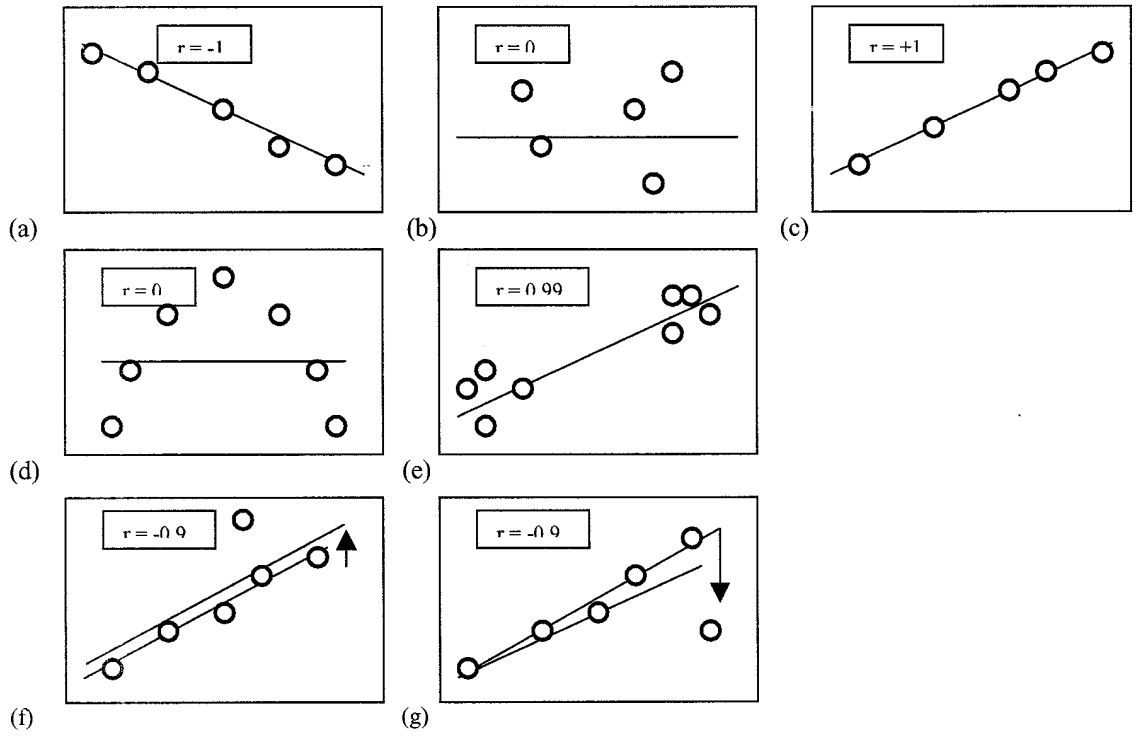
LITERATURA

1. Snedecor, G.W., Cochran, W.G.: Statistical Methods, The Iowa State University Press, USA, 6th edition (1967)
2. Draper, N., Smith, H.: Applied Regression Analysis, J.Wiley & Sons, New York, 2nd edition (1981)
3. BS ISO 11095; Linear calibration using reference materials (1996)
4. Statistical Software Qualification: Reference data sets, Ed. Butler, B.P., Cox, M.G., Elison, S.L.R., Hardcastle, W.A., The Royal Society of Chemistry. ISBN 0 85404 422 1 (1996)
5. Sahai, H., Singh, R.P.: The use of R^2 as a measure of goodness of fit: An overview, Virginia J. Sci. 40(1), 5-9 (1989)
6. Burke, S.: Statistics in context: Significance testing, VAM Bulletin, 17, 18-21, Automn 1997

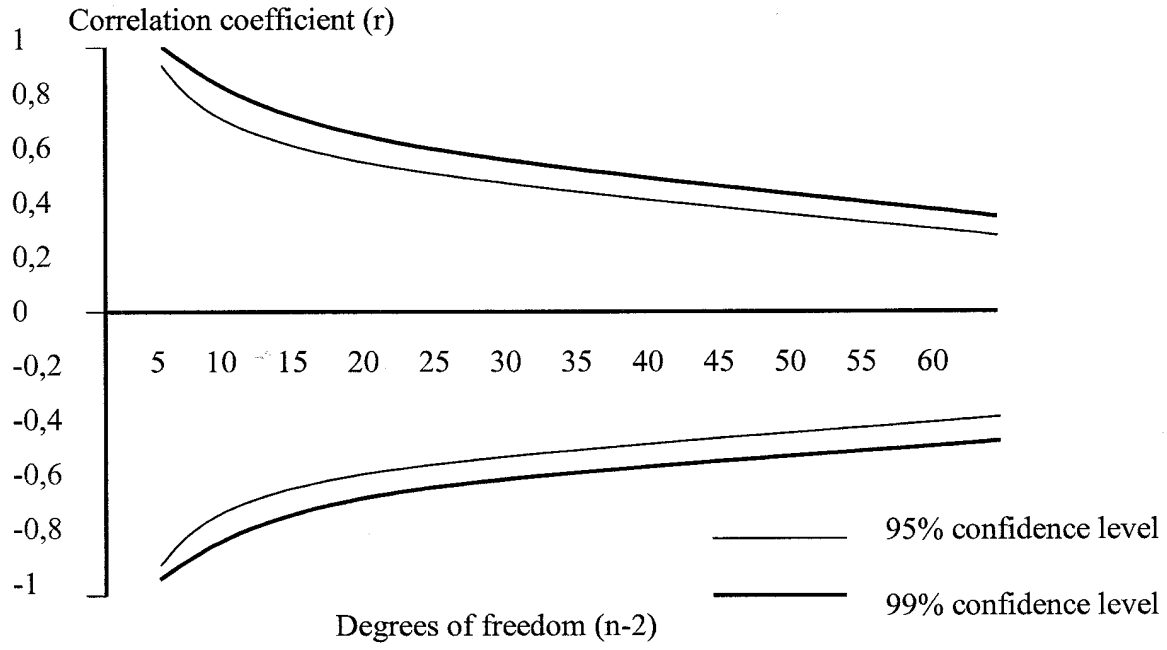


Obrázek V.1 Typický kalibrační graf s popisem součástí

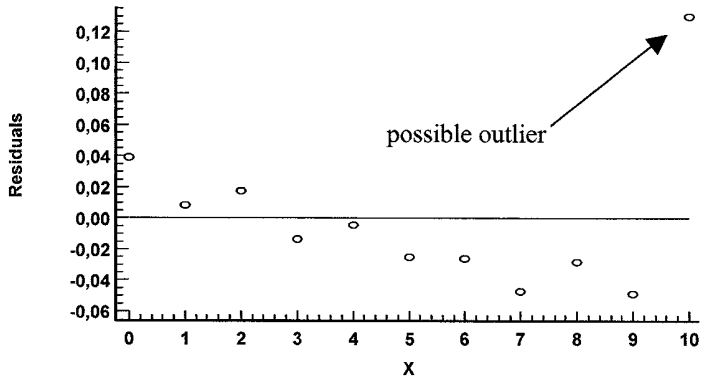
7. Snedecor, G.W., Cochran, W.G.: Statistical Methods, Table A11, pp. 557, The Iowa State University Press, USA, 6th edition (1967)
8. Burke,S.: Statistics in context: Regression and calibration, VAM bulletin (Valid Analytical Measurement), publication of Laboratory of the Government Chemist, issue No. 18, 18-21, Spring 1998
9. BS 2846 (ISO 5479) Statistical interpretation of data. Part 7: Tests for departure from normality. British Standards Institution (1984)



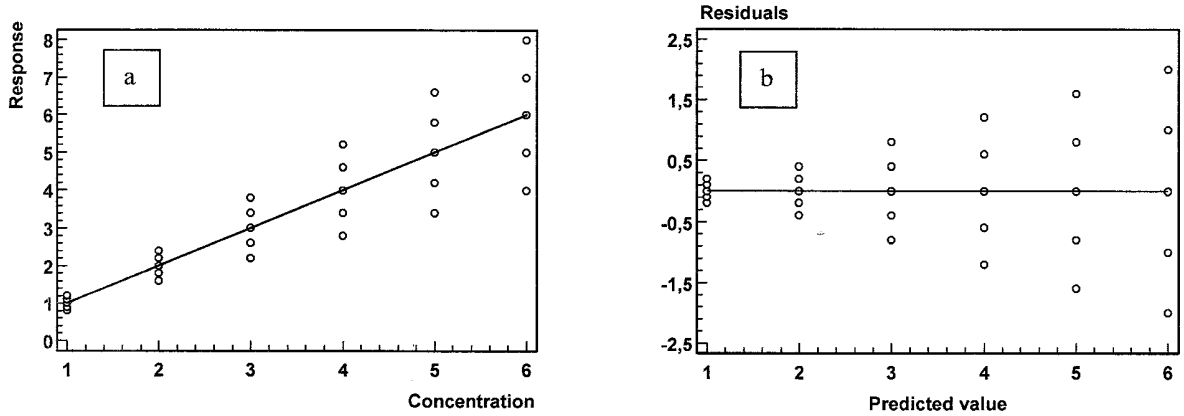
Obrázek V.2 Korelační koeficienty a stupeň shody (goodnes of fit).



Obrázek V.3 Statistická významnost korelačního koeficientu.



Obrázek V.4 Graf residuí.



Obrázek V.5 Grafy typické závislosti odezvy přístroje na koncentraci.

Přehled symbolů

symbol	význam
X	nezávislá (vysvětlující) proměnná, např. koncentrace sady kalibrátorů
Y	závislá (vysvětlovaná) proměnná, např. odezva přístroje
x_i nebo x_j	hodnota x pro i-tý nebo j-tý bod
y_i nebo y_j	hodnota y pro i-tý nebo j-tý bod
$\bar{\bar{y}}$	průměr všech y hodnot kalibrace
$\bar{y}$	průměrná hodnota odezvy přístroje pro m replikátů jednoho měřeného vzorku
s_x	výběrová směrodatná odchylka hodnot proměnné x v kalibraci
s_y	výběrová směrodatná odchylka hodnot proměnné y v kalibraci
a	intercept regresní čáry (hodnota Y pro X=0)
b	slope (sklon, směrnice) regresní čáry
m	počet paralelních stanovení (replikátů) pro jednu hodnotu x
n	počet bodů (párů x_i, y_i)
r	korelační koeficient regrese metodou nejmenších čtverců
RSE	směrodatná odchylka residuí (residual standard error)
$residual_{max}$	největší residuum
t	kritická hodnota získaná z tabulek t-hodnot pro n-2 stupně volnosti