Von kleinen Werten mit großem Einfluss - Sum Of Squares

Dr. Peter Paul Heym – Sum Of Squares

Jeder kennt diese Situation aus der eigenen, alltäglichen Arbeit: Es sind manchmal die kleinen Dinge, die einen großen Einfluss auf das Ergebnis unserer Arbeit haben können. Manchmal gewinnen unscheinbare Faktoren auf einmal an Bedeutung, wenn ihnen die entsprechende Aufmerksamkeit geschenkt wird – oder geschenkt werden muss. Einer dieser unscheinbaren Faktoren kann gerade bei der analytischen Methodenvalidierung eine große Rolle spielen. Oft wird dieser Faktor, diese eine Zahl, übersehen, denn ihr Wert spielt bei der Validierung auf den ersten Blick keine große Rolle. Der Name dieser Zahl ist die sogenannte *residual sum of squares*, kurz *RSS*, oder zu Deutsch die „Summe der Fehlerquadrate". Was es mit dieser *residual sum of squares* auf sich hat und wie man diese Zahl interpretiert, werden wir in diesem Beitrag am Beispiel der analytischen Methodenvalidierung klären und damit dieser Zahl die Bedeutung zukommen lassen, die sie verdient. Ebenso werden die Vor- und Nachteile der *RSS* besprochen, sowie Ansätze diskutiert, mit denen man potentielle Fehlinterpretationen der *residual sum of squares* vorbeugen kann.

Analytische Methodenvalidierung und die ICH Q2(R1)

Für die analytische Methodenvalidierung ist ein Dokument von Bedeutung, in dem mehrere Punkte einer Methode geprüft werden müssen, um sie als *fit-for-purpose* zu deklarieren. *Fit-for-purpose* bedeutet, dass die Methode den Zweck erfüllt, für den sie gedacht ist. Neben den Eigenschaften der *specificity* und *range*, *accuracy* und *precision*, sowie dem Bestimmen der *limit of detection* (LOD) und *limit of quantification* (LOQ), ist auch die Linearität der Methode zu beurteilen. Diese Eigenschaften sind in der Guideline "**Validation of Analytical Procedures: Text and Methology Q2(R1)"** der ICH, des *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use*, beschrieben [1]. Für jeden dieser Punkte gelten Vorschriften, nach denen bestimmt wird, ob eine Methode *fit-for-purpose* ist oder nicht. Bei der Prüfung, ob eine Methode für Ihren Zweck geeignet ist, orientieren sich die entsprechenden Behörden an diesem Dokument. Bei dem Thema der Linearität muss dabei nachgewiesen werden, ob es innerhalb des *range* der Methode einen linearen Zusammenhang zwischen der Analytenkonzentration und einem Signal, abhängig von der verwendeten Messmethode, gibt. Die Grundlage für die Bestimmung der Linearität ist dabei das Aufstellen eines mathematischen Zusammenhangs zwischen (Analyten-) Konzentration und Signal. Der mathematische Zusammenhang kann dabei mithilfe der linearen Regression hergestellt werden, und wir können diesen anschaulich durch einfache Grafiken, wie in Abbildung 1, darstellen.

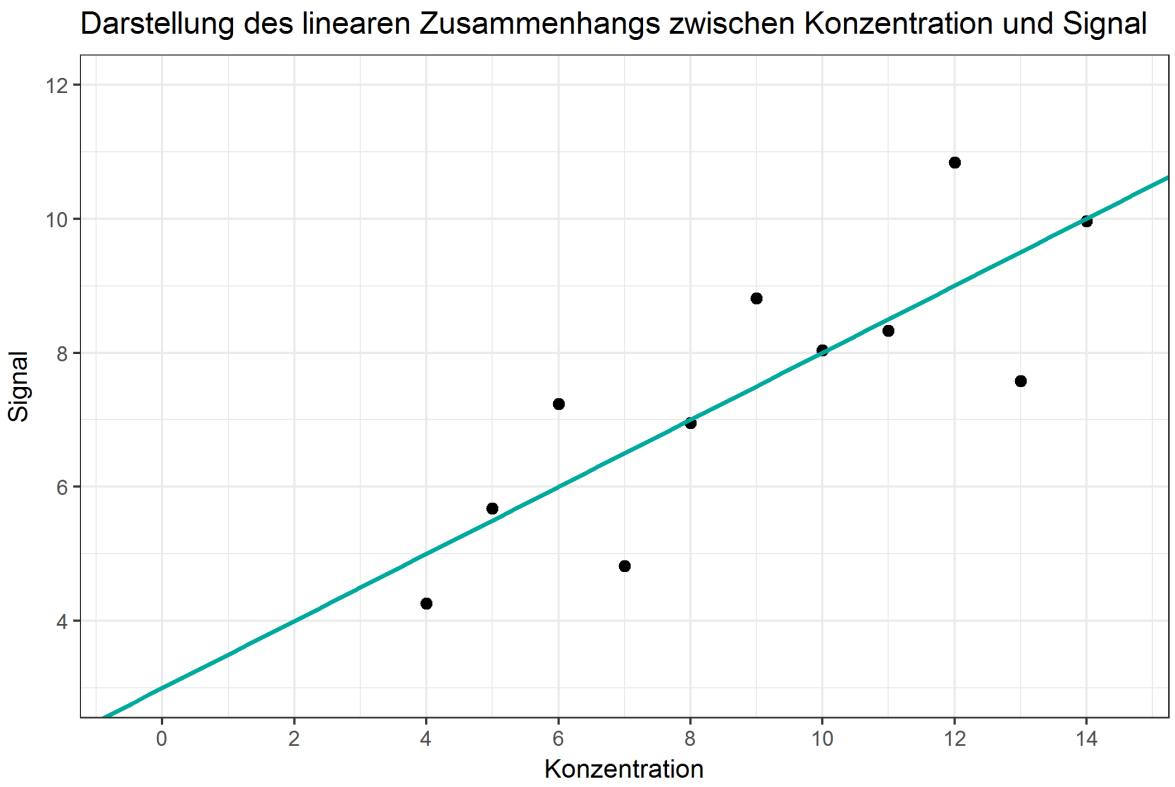

Abbildung 1: Darstellung des linearen Zusammenhangs zwischen Analytenkonzentration und Signal

Mit Hilfe solcher Darstellungen ist der Zusammenhang beider Größen erkennbar. Mit steigender Konzentration gewinnt auch das Signal an Intensität. Mathematisch wird dieser Zusammenhang durch den Anstieg der Geraden, oder auf Englisch *slope*, beschrieben. Was ebenfalls von Bedeutung ist, ist die Höhe der Geraden, bei der sie die y-Achse schneidet, beziehungsweise, das Signal, das bei einer Konzentration von 0 detektiert würde. Dieser Wert wird der Y-Achsenabschnitt (engl. *y-intercept*) genannt. Vorliegendes Beispiel weist einen Anstieg von $0{,}5$ und einen Achsenschnittpunkt von $3{,}0$ auf. Das bedeutet, bei einer Konzentration von 0 würde die Methode noch ein Signal von 3 detektieren und für jede Steigerung der Konzentration um eine Einheit würde sich das Signal um den Wert $0{,}5$ erhöhen. Mathematisch lautet dieser Zusammenhang: $\text{Signal} = 3 + 0{,}5 * \text{Konzentration}$ beziehungsweise $\mathrm{y} = 3 + 0{,}5 * \mathrm{x}$. Statistikprogramme, oder Excel sind in der Lage, diese Analysen zu tätigen, darauf wird später kurz eingegangen.

Mit Blick auf die Daten stellt man fest, dass man bei einer Konzentration von 9 ein Signal von $8{,}81$ gemessen hat. Würde man die Konzentration $x = 9$ in die Gleichung einsetzen, würde man jedoch einen Wert von $3 + (0{,}5 * 9) = 7{,}5$ erhalten. Die Regressionsgerade sagt also etwas Anderes vorher, als der Wirklichkeit entspricht. Sie beschreibt den bestmöglichen linearen Zusammenhang zwischen den Daten. Dennoch ist die Gerade nicht in der Lage, jeden Wert genau vorherzusagen, denn jeder einzelne gemessene Datenpunkt weicht von der berechneten Gerade ab. Diese Unterschiede, diese Schwankungen in den Daten, sorgen dafür, dass die Gerade den Zusammenhang nicht zu 100% erfassen kann. Sie kann in diesem Fall die Schwankungen in den gemessenen Werten zu ca. $66{,}7\%$ erklären. Dieser Wert entspricht dem berühmten R^2 Wert – dieser sagt aus, dass mit der berechneten Regressionsgeraden $66{,}7\%$ der Variabilität in den Daten erklärt werden können. Die restlichen $33{,}3\%$ an Variabilität können durch diese lineare Regression nicht erklärt werden.

Reicht eine Grafik, um der ICH gerecht zu werden?

Nun sind mittels linearer Regression der lineare Zusammenhang zwischen Konzentration und Signal, als auch Anstieg und Achsenschnittpunkt, sowie die Güte des Zusammenhanges ermittelt. Reichen diese Maße aus, um unsere Methode als *fit-for-purpose* zu beschreiben? In der Guideline der ICH heißt es dazu:

*"A **linear relationship** should be evaluated across the range … of the analytical procedure. … Linearity should be evaluated by visual inspection of a **plot of signals** as a function of **analyte concentration** or content. If there is a linear relationship, test results should be evaluated by appropriate statistical methods, for example, by **calculation of a regression line** … . The **correlation coefficient, y-intercept, slope** of the regression line and **residual sum of squares** should be submitted."*

Die Regressionsgerade, zusammen mit dem *slope* und *y-intercept*, sind bereits bestimmt worden. Grafisch wurden die Daten ebenfalls dargestellt. Die Güte des Zusammenhanges wurde durch den R^2 Wert ausgedrückt. Die Wurzel des R^2 Wertes, die mit *r* bezeichnet wird, entspricht dem Korrelationskoeffizienten (*correlation coefficient*), der durch die ICH gefordert wird. … Wäre da nicht der letzte Satz, in dem von den *residual sum of squares* die Rede ist.

Von Einzelwerten zu *residual sum of squares*

Was haben die *residual sum of squares* mit der Geraden und den Daten zu tun und was sagen diese eigentlich aus? Wie bereits festgestellt wurde, sagt die Gerade für die Konzentration von 9 ein Signal von $7{,}5$ vorher, obwohl die eigentliche Messung ein Signal von $8{,}81$ ergab. Da die Varianz in den Daten laut R^2 „nur" zu knapp 67% erklärt werden kann, bleiben noch rund 33% unerklärte Varianz übrig. Einen Teil dieser Variabilität, der Abweichung unserer Messwerte vom mathematisch idealen Zusammenhang, spiegelt sich in jedem einzelnen Messwert wider. So „verschätzt" sich das mathematische Modell bei der Konzentration von 9 um $(8{,}81 - 7{,}50) = y - \hat{y} = 1{,}31$ Signaleinheiten. Das entspricht dem Fehler, den das Regressionsmodell an diesem Punkt begeht, es *unter*schätzt den tatsächlich gemessenen Wert von $8{,}81$ um $1{,}31$ Einheiten. Für die gemessene Konzentration von 8 beträgt der Unterschied zwischen Messwert und Vorhersage des Signals $(6{,}95 - 7{,}00) = -0{,}05$. Für diese Konzentration *über*schätzt die Gerade die tatsächliche Konzentration - wenn auch um nur $0{,}05$ Signaleinheiten. Diesen Unterschied nennt man Fehler, Rest oder Residuum, auf Englisch *error*, *residuum* oder *residual* – er wird meistens mit ε (epsilon) bezeichnet. Den Fehler für den ersten Datenpunkt mit Konzentration $x = 4$ würde man daher folgendermaßen beschreiben: $\varepsilon_1 = y_1 - \hat{y}_1 = 4{,}26 - 5{,}00 = -0{,}74$. Die Abweichungen für jeden Messwert kann man in der gleichen Grafik darstellen, indem man die entsprechenden Fehler durch graue Linien darstellt. Die Länge jeder grauen Linie entspricht dem Fehler ε, dem *residual* jedes Datenpunktes. Um nun die Güte des gesamten linearen Zusammenhanges zu beschrieben, liegt es auf der Hand, diese Fehler für alle Messwerte zu addieren, denn das entspräche der totalen Abweichung des linearen Modells von den gemessenen Werten:

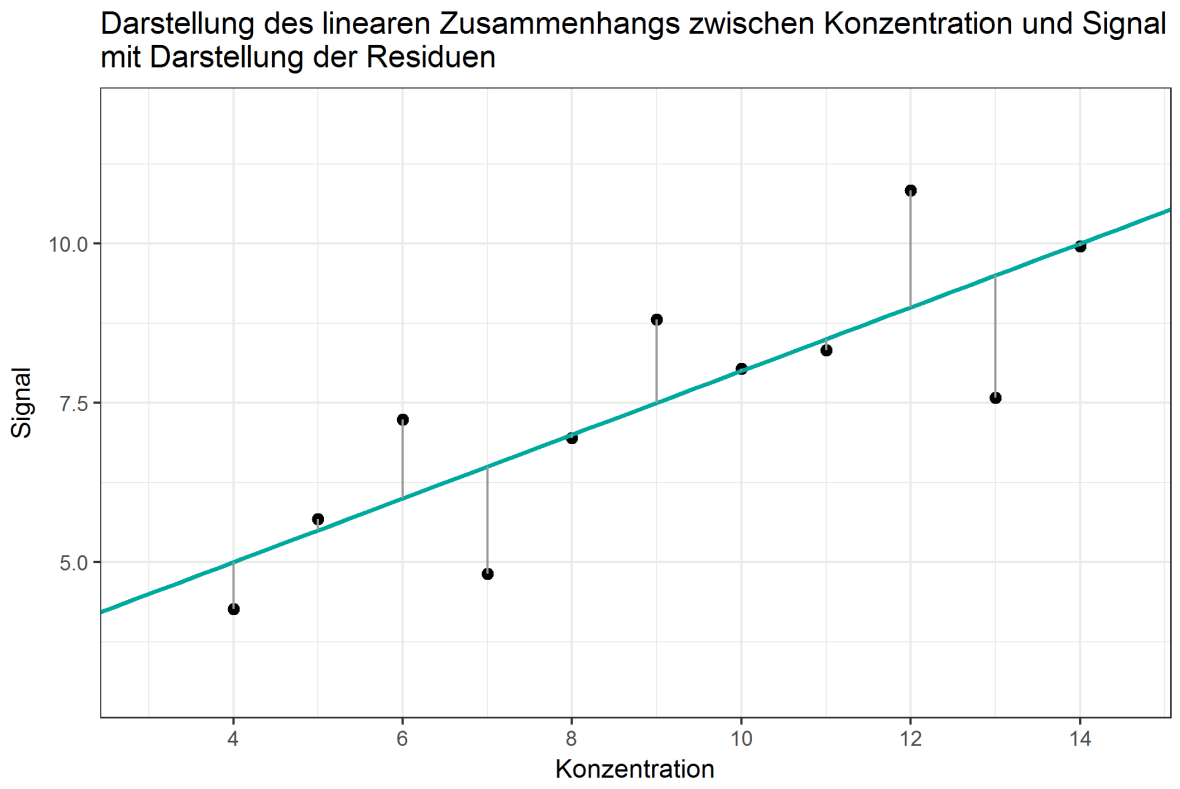

Abbildung 2: Darstellung des linearen Zusammengangs zwischen Konzentration und Signal mit Darstellung der Residuen

Wie in Tabelle 1 nachzurechnen, ist die Summe der Fehler (die *sum of residuals*) genau 0. Da die Abweichungen sowohl negativ, als auch positiv sein können, gleichen sich diese genau aus - dies ist eine Eigenschaft der linearen Regression. Obwohl man sehen und nachrechnen kann, dass sich das Modell in jedem Datenpunkt „verschätzt“, ist die Summe dieser Schätzfehler 0, und damit kein aussagekräftiger Wert, was die Qualität unserer Methode angeht.

Tabelle 1: Residuendarstellung

X-Werte	4	5	6	7	8	9	10	11	12	13	14	Summe
Residuen	-0,740	0,179	1,239	-1,680	-0,050	1,309	0,039	-0,171	1,838	-1,921	-0,041	0

Um Aussagen über die Qualität machen zu können, bedient man sich daher des „Tricks“, alle Fehlerwerte zu quadrieren. Das hat zwei Vorteile: Zum einen ist das Quadrat einer Zahl immer positiv. Und zum anderen wird aus einem Wert, der kleiner als 1 ist, ein noch kleinerer Wert (z.B. $0{,}5^2 = 0{,}25$), wohingegen Werte, die größer sind als 1, durch das Quadrieren noch größer werden (z.B. $2{,}5^2 = 6{,}25$). Kleine Abweichungen von der Regressiongerade werden daher „belohnt“, große Abweichungen „bestraft“.

Tabelle 2: X-Werte, Residuen und Residuenquadrate des Datensatzes

X-Werte	4	5	6	7	8	9	10	11	12	13	14	Summe
Residuen	-0,740	0,179	1,239	-1,680	-0,050	1,309	0,039	-0,171	1,838	-1,921	-0,041	0
Residuen-quadrate	0,547	0,032	1,535	2,822	0,002	1,713	0,001	0,029	3,378	3,690	0,001	13,754

Man quadriert daher zuerst die Abweichungen, die Residuen, und summiert diese anschließend. … *First, we square the residuals and sum them up afterwards*. Das Ergebnis ist die Summe der Fehlerquadrate, Residuenquadratsumme oder engl. *residual sum of squares*. Dies ist der Wert, der beschreibt, wie groß die komplette quadratische Abweichung der Messwerte von den idealen Werten der Regressionsgerade ist. Da alle Fehler quadriert und damit positiv sind, ist auch die Summe immer positiv – kleine Summen der Fehlerquadrate sollten folglich immer gute mathematische Zusammenhänge beziehungsweise gute Kalibriergeraden, wie in diesem Beispiel, repräsentieren.

Doch was sagt der Wert über die Daten und Güte des linearen Zusammenhanges aus? Je kleiner dieser Wert ist, desto besser sollte der Zusammenhang zwischen X- und Y Werten sein, zwischen Analytenkonzentration und Signal, sein. Oder etwa doch nicht? Welche Informationen stehen in der Guideline der ICH dazu? In der Q2(R1) findet sich keine Aussage bezüglich der (z.B. maximal erlaubten) Größe der *residual sum of squares*. Warum? Sollte es nicht Grenzen geben, die vorschreiben, was ein gutes oder erlaubtes Maß für Abweichungen ist und was nicht? Ab welchen Werten ist die Methode nicht mehr *fit-for-purpose* bezüglich der Linearität? Warum gibt diese Guideline keine Vorschläge oder Hinweise bezüglich der Bewertung der *RSS*?

Um das zu beantworten, muss man nochmals untersuchen, was die *residual sum of squares* bedeuten. Die *sum of squares* repräsentieren Abweichungen der Messwerte von den „idealen“ Werten der Regressionsgeraden. In diesem Beispiel misst man, wie in Grafik 1, das Signal in Litern. Was passiert, wenn das Signal stattdessen in Gallonen (1 Gallone = 4,54609 L) gemessen würde?

Tabelle 3: Auftretende Unterschiede in Kennzahlen der linearen Regression bei Änderung der Einheit des Signals

	Anstieg der Geraden	Achsenabschnitt	RSS	R^2
Liter	0,500	3,000	13,754	0,667
Gallonen	**0,109**	**0,660**	**0,666**	**0,667**
Umrechnung	$0{,}5 = 0{,}109 * 4{,}54609$	$3{,}0 = 0{,}66 * 4{,}54609$	$13{,}754 = 0{,}666 * 4{,}54609^2$	

Der Wechsel der Einheit bewirkt mehrere Änderungen in Werten, die bei der Methodenvalidierung angegeben werden müssen. Der Achsenabschnitt verringert sich ebenso wie der Anstieg der Geraden. Die *RSS* reduzieren sich von über 13 auf unter 1 - auf unter 5% des originalen Wertes (!). Lediglich der R^2 Wert bleibt gleich. Wie kommt dies zu Stande? Die Antwort liegt in der Umrechnung von Liter auf Gallonen. Der Umrechnungsfaktor von $4{,}54609$ verursacht nicht nur eine Änderung der Regressionsgeraden, sondern verändert auch die *residual sum of squares*, obwohl die Gesamtqualität des Modelles bezogen auf den R^2 Wert, wie in Tabelle 3 beschrieben, die Gleiche bleibt. Eine entsprechende Grafik bliebe bis auf die Achseneinteilung ebenfalls unverändert.

Dies bedeutet, man könnte jede beliebige (z.B. von der ICH festgelegte) Obergrenze für die *RSS* mit Hilfe einer einfachen Änderung der Einheit elegant umgehen. Die *RSS* können daher immer nur in Verbindung mit der Messmethode, dem R^2 und den Eigenschaften der Regressionsgeraden begutachtet werden. Die alleinige *RSS* hat ohne Bezug zum Rest der Methode keine Aussagekraft. Es macht also durchaus Sinn, dass die ICH die Angabe des *RSS* verlangt **ohne** eine Grenze für die Güte des Modelles festzulegen. Da die *RSS* nur mit allen anderen Angaben zum Regressionsmodell Sinn macht, ist auch der folgende Satz der ICH völlig berechtigt:

*"The … **residual sum of squares** should be submitted. A **plot of the data** should be included. In addition, an **analysis of the deviation of the actual data points from the regression line may also be helpful** for evaluating linearity."*

Dieser Satz ist von großer Bedeutung, auch wenn das Wort *may* aus der Guideline nicht sofort darauf hindeutet. Denn selbst bei kleinen *RSS* ist es theoretisch möglich, dass wenige Datenpunkte für einen Großteil der Abweichungen verantwortlich sind und damit die Qualität unter Umständen negativ beeinflussen. Ebenso können vermeintlich große *RSS* zu hervorragenden Regressionen gehören, je nachdem in welcher Einheit man misst. Dazu kommt, dass die *RSS* nur eine Angabe über die Abweichungen *aller* Datenpunkte, also z.B. der Kalibrierung über den *gesamten* Konzentrationsbereich, macht. Dabei müsste man darüber hinaus bewerten, wie die Beiträge der *einzelnen* Datenpunkte in Bezug auf die Gesamtabweichung sind. Daher wünscht sich die ICH zusätzliche Maßnahmen, die neben der *RSS* auch die einzelnen Datenpunkte auf Einfluss untersuchen.

Um die Methode als *fit-for-purpose* zu deklarieren, ist es nötig, nachzuweisen, dass alle aufgenommenen Datenpunkte dem tatsächlichen chemischen oder physikalischen Zusammenhang entsprechen, daher sollten in dem Datensatz keine ungewöhnlichen Abweichungen der einzelnen Datenpunkte zu der Regressionsgeraden auftreten. Weder kleine *residual sum of squares*, noch hohe R^2 Werte allein sind Hinweise auf *fit-for-purpose* Methoden. Daher muss man von Eigenschaften des Gesamtdatensatzes zu den Eigenschaften der Einzelwerte übergehen, zur ***„analysis of the deviation of the actual data points from the regression line"***, wie es die ICH fordert.

Von den *residual sum of squares* zu den Einzelwerten

Da die *RSS* nur den gesamten Datensatz betrachtet, könnte man in einem ersten Schritt für jeden Datenpunkt dessen Anteil an der *residual sum of squares* berechnen und dies tabellarisch oder grafisch darstellen:

Tabelle 4: Anteile der residual sum of squares bezogen auf alle Einzeldatenpunkte

X-Werte	4	5	6	7	8	9	10	11	12	13	14	Summe
Residuen	-0,740	0,179	1,239	-1,680	-0,050	1,309	0,039	-0,171	1,838	-1,921	-0,041	0
Residuen-quadrate	0,547	0,032	1,535	2,822	0,002	1,713	0,001	0,029	3,378	3,690	0,001	13,754
Anteil an RSS (%)	3,98	0,23	11,16	**20,52**	0,01	12,45	0,01	0,21	**24,56**	**26,82**	0,01	100

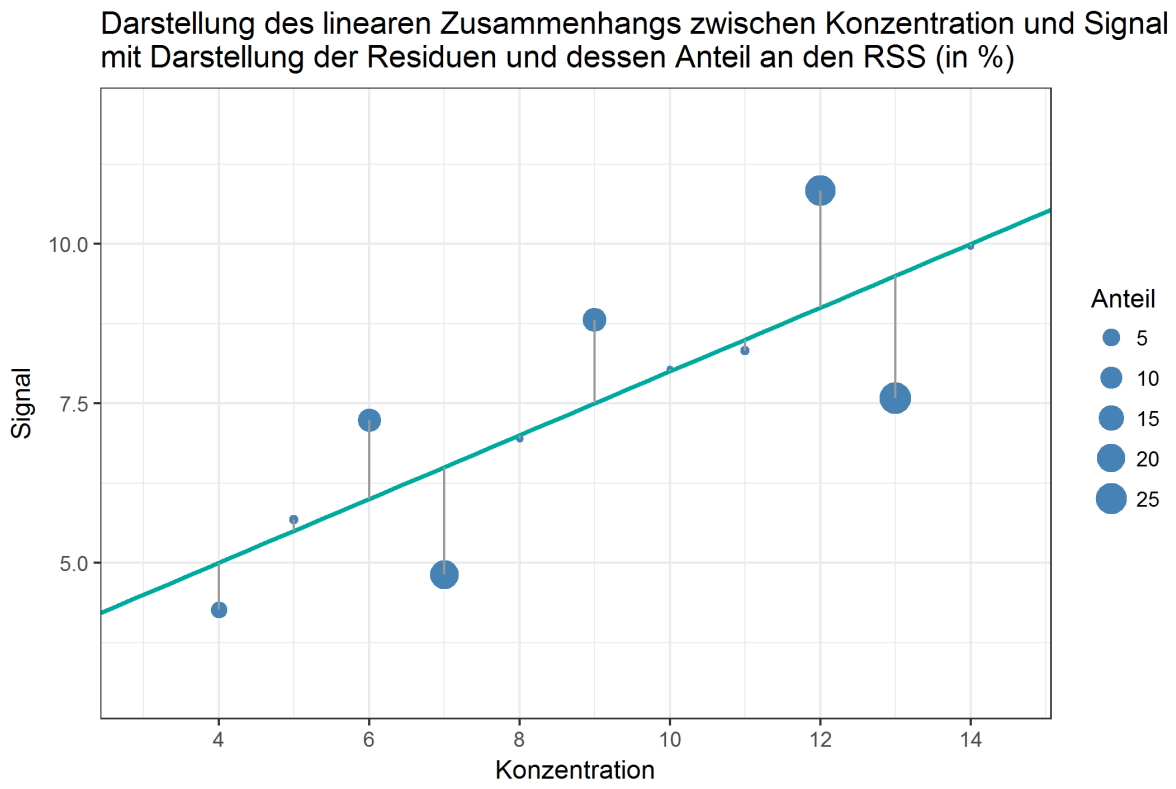

Abbildung 3: Linearer Zusammenhang zwischen Konzentration und Signal, mit Darstellung der Anteile an der residual sum of squares

In Abbildung 3 beschreibt die Größe jedes Datenpunktes den Anteil an der gesamten *RSS*. Doch leider gibt diese Grafik keine Aufschlüsse darüber, welchen Einfluss jeder Datenpunkt auf die Qualität des linearen Models und damit die Qualität der analytischen Methode hat. Denn obwohl bereits 3 Datenpunkte (Konzentration 7, 12 und 13) mehr als 70% der gesamten *residual sum of squares* ausmachen, kann man daraus nicht schlussfolgern, dass diese einen großen Einfluss, beziehungsweise andere Datenpunkte keinen Einfluss auf das Modell haben.

Hat Values und *Cooks Distance* – was beeinflusst die lineare Regression tatsächlich?

Der Einfluss eines einzelnen Datenpunktes lässt sich erst damit bestimmen, wenn man untersucht, ob das Entfernen des Datenpunktes die Regressionsgerade stark verschieben würde. Es ist durchaus möglich, dass Datenpunkte mit großem Anteil der *RSS* die Regressionsgerade kaum verschieben würden, wenn sie entfernt würden. Dies ist abhängig davon, in welcher Entfernung sich der Datenpunkt zu dem Rest aller anderen Datenpunkte befindet. Folgende Grafik zeigt den jeweiligen Einfluss der Datenpunkte auf die Regressionsgerade, würde man diese entfernen:

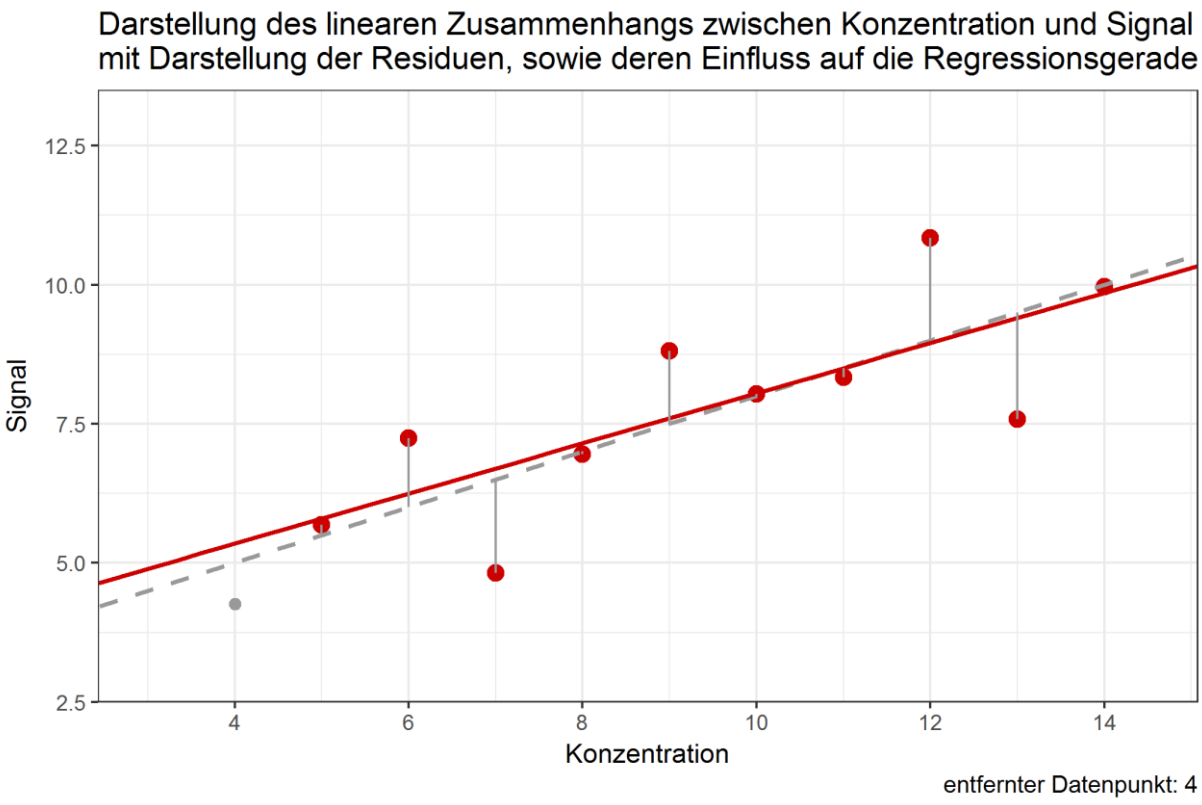

Abbildung 4: Verschiebung der Regressionsgeraden durch Weglassen einzelner Datenpunkte

Obwohl der Wert 9 einen Anteil von über 12% an den *residual sum of squares* besitzt, würde ein Weglassen dieses Wertes die Gerade kaum verschieben. Im Gegenzug dazu würde das Weglassen von Datenpunkt 4 die Gerade ähnlich stark beeinflussen, obwohl dessen Anteil an der *RSS* nur knapp 4% beträgt. Dies ist ein Beispiel dafür, dass Werte, die (in Bezug auf die Konzentration) in größerer Entfernung zu allen anderen Punkten liegen, einen potentiell größeren Einfluss auf die Verschiebung der Regressionsgeraden haben können. Diesen Einfluss nennt man *leverage*. Datenpunkte mit großem *leverage* werden als *high-leverage points* bezeichnet. Das deutsche Äquivalent zu *leverage* lässt sich z.B. mit dem Wort „aushebeln“ beschreiben – *high-leverage points* sind in der Lage, die Regressionsgerade auszuhebeln, beziehungsweise deren Lage stark zu ändern. Datenpunkte an den Randbereichen der X-Werte sind dazu stärker in der Lage, als Datenpunkte, die nahe dem Zentrum der Datenpunkte liegen. In der Methodenvalidierung betrifft das vornehmlich Datenpunkte an der Bestimmungsgrenze, welche einen hohen *leverage* aufweisen können. Doch gerade von Werten in diesem Konzentrationsbereich muss eine hohe Genauigkeit erwartet werden. Umso wichtiger ist es, den Einfluss dieser Punkte genauestens zu untersuchen.

Die Grenze, ab der man einen *high-leverage point* kennzeichnet, liegt in diesem Beispiel bei $2 * (2/n) = 0{,}363$, wobei $n = 11$ die Anzahl der Datenpunkte beschreibt. In Abbildung 5 sind die Datenpunkte in Bezug auf deren *leverage* gekennzeichnet. Synonyme für *leverage* sind die sogenannten *hat values*, oder H Werte.

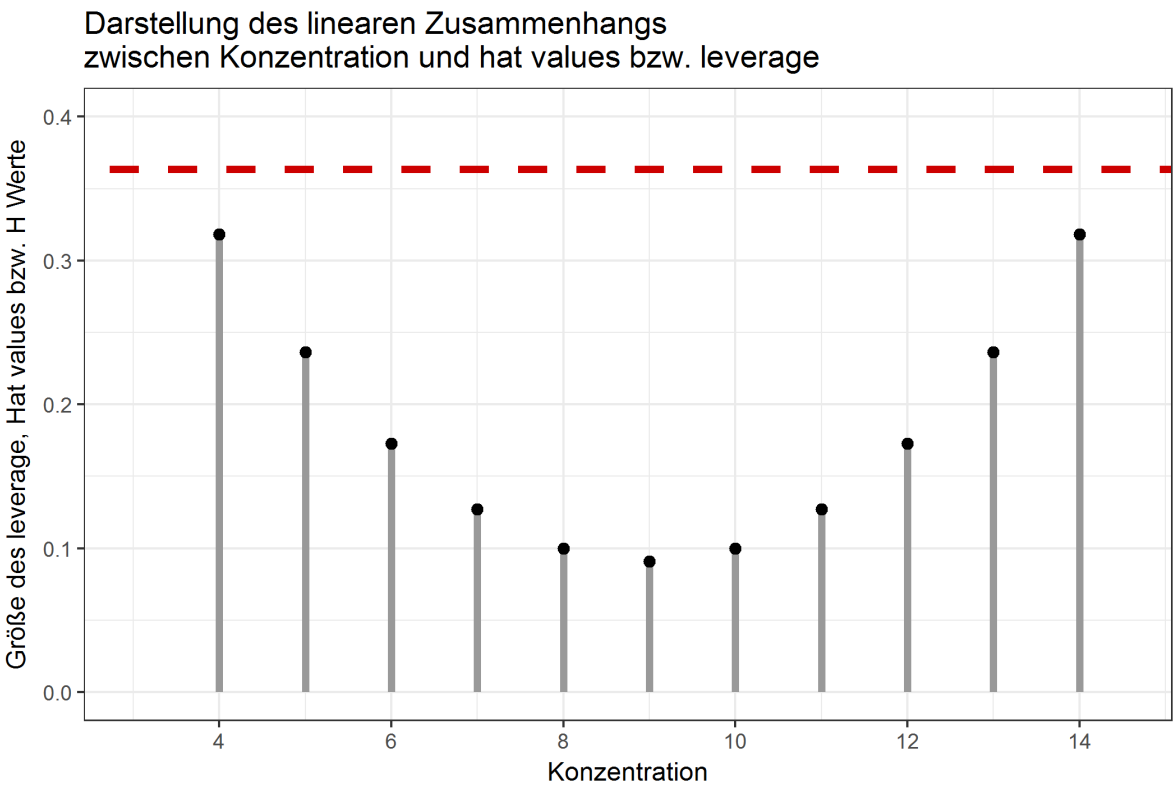

Abbildung 5: leverage der Einzeldatenpunkte in Bezug auf die Grenze von 0,363

Die Punkte für Konzentrationen von 4, 5, als auch 13 und 14 haben hohe *leverage* Werte, da sie, in Bezug auf alle anderen Datenpunkte, relativ weit entfernt sind. Die *leverage* Werte liegen jedoch unter der Grenze von $0{,}363$, so dass in diesem Datensatz keine *high-leverage points* vorhanden sind. Hohe *leverage* Werte sind nicht automatisch Datenpunkte mit großem Einfluss, da sie nur abhängig von der X-Koordinate (bzw. der Konzentration) sind. Die Y-Werte, also in diesem Falle die Signalintensität, spielen für die Berechnung des *leverage* keine Rolle. Daher kann es ebenso vorkommen, dass Datenpunkte mit geringen *leverage*-Werten dennoch einen starken Einfluss auf die Regressionsgerade haben und damit die Qualität der Methode besonders stark beeinflussen. Um diese Punkte zu identifizieren, benötigt man ein weiteres Maß, welches den Fokus auf die Y-Koordinaten der Daten legt.

Begutachtet man sowohl den Einfluss der X- und Y-Koordinaten jedes Einzelwertes auf die Regressionsgerade, ist es möglich, sogenannte *influencial observations*, also einflussreiche Datenpunkte, auszumachen. Ein solches Maß für den tatsächlichen Einfluss ist die sogenannte *Cooks Distance* oder die sogenannten D Werte [2]. Diese sollten ebenfalls möglichst gering sein, und sollten den Wert von, in diesem Falle, $4/(n-2)$ nicht überschreiten, wobei n die Anzahl der Datenpunkte beschreibt. In Abbildung 6 sind die Datenpunkte nach *Cooks Distance* aufgetragen. Der Grenzwert für diesen Datensatz hier liegt bei $4/9 = 0{,}445$. In diesem Beispiel wird der Datenpunkt mit $X = 13$ als *influencial observation* gekennzeichnet, da er die Regressionsgerade besonders stark zu sich verschiebt - und dass, obwohl er nicht den größten *leverage* aufweist.

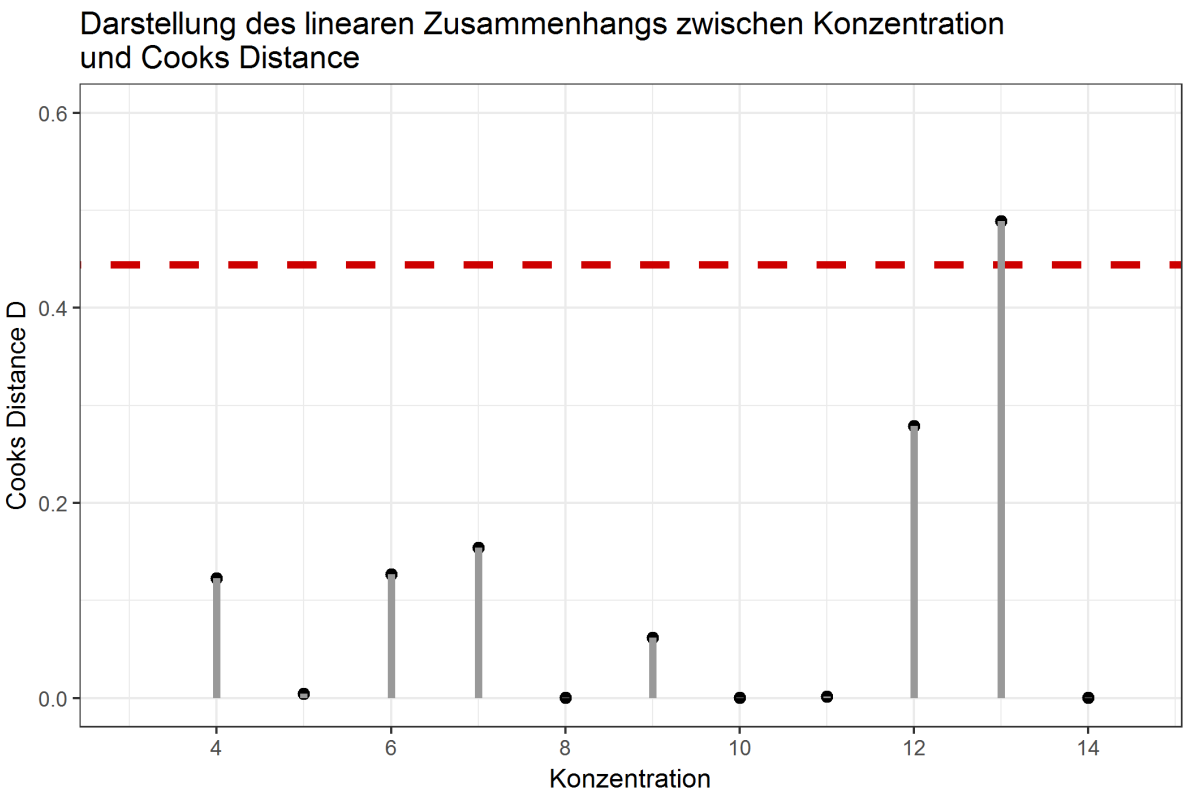

Abbildung 6: Cooks Distance zur Identifizierung von influencial observations

Die Werte von *Cooks Distance* sind eine "Mischung" aus den jeweiligen Anteilen der *RSS* und den *leverage* Werten. Damit kombinieren sie die Eigenschaften aller Einzelwerte: Sie untersuchen zum einen die Abweichung jedes Wertes in Y-Koordinaten (also des Signals, der *RSS*), als auch die Abweichung in X-Koordinaten (also Analytenkonzentration, bzw. *leverage*) in Bezug auf alle anderen Punkte. Erst damit lässt sich bestimmen, wie groß der Einfluss jedes Einzelwertes auf die Regressionsgerade, und damit auf die Güte des linearen Zusammenhanges, tatsächlich ist. Eine Grafik, in der die Daten, zusammen mit *Cooks Distance*, beschrieben sind, könnte folgendermaßen aussehen: Die Daten und die Regressionsgerade, werden wie bisher dargestellt, und die Farbe der Datenpunkte gibt an, ob die *Cooks Distance* jedes Punktes den Wert von 0,445 übersteigt. Damit werden automatisch alle einflussreichen Datenpunkte gekennzeichnet.

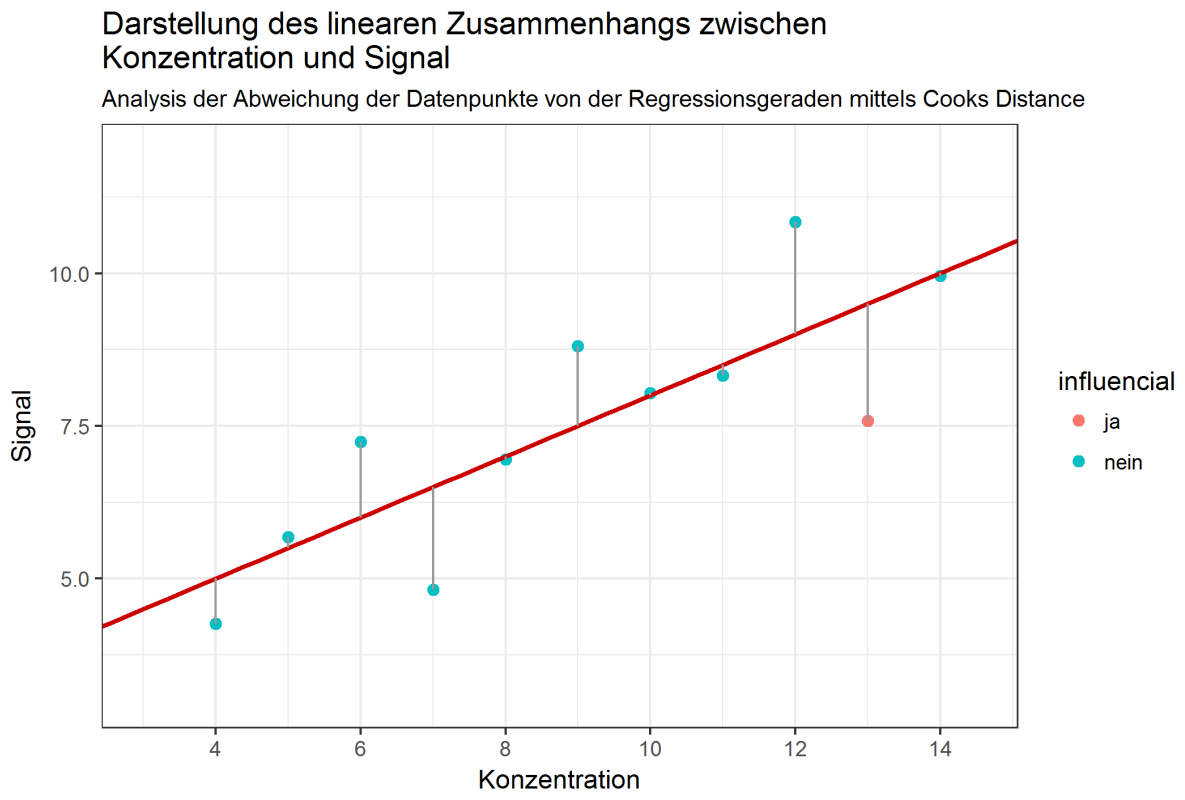

Abbildung 7: Lineare Regression und mittels Cooks Distance identifizierter einflussreicher Datenpunkt

Was ist also zu tun bei der Untersuchung der Linearität?

Die von der ICH geforderte ***„analysis of the deviation of the actual data points from the regression line“*** kann aus gutem Grund nicht nur in der visuellen Begutachtung der Datenpunkte liegen. Das Berechnen des Einflusses jedes Datenpunktes offenbart die eigentliche Qualität der linearen Regression. Die Angaben der Merkmale wie *r* oder R^2, Anstieg und Achsenschnittpunkt, oder der *RSS* zeigen nur eine Hälfte der Medaille. Doch da weder die *residual sum of squares*, noch der R^2 oder die Eigenschaften der Regressionsgeraden Hinweise auf den Einfluss von Einzelwerten geben können, sollte die ***„Analysis der Abweichung der eigentlichen Datenpunkte von der Regressionsgerade“*** immer durchgeführt werden.

Datenpunkte mit großem Einfluss auf die Regressionsgerade können ein Indiz dafür sein, dass der lineare Zusammenhang über die *gesamte* Breite des Konzentrationsbereiches nicht überall gleich „gut“ sein könnte, so wie es die ICH fordert. Hinweise darauf finden sich häufig nicht durch visuelle Inspektion der Daten, sondern durch Analysen wie der Berechnung der *Cooks Distance*. Findet man jedoch potentielle *influencial observations*, muss in solchen Fällen anschließend überprüft werden, in welchen Konzentrationsbereichen diese Abweichungen auftreten und ob diese Abweichungen „normal“ sind. Gegebenenfalls muss geklärt werden, mit welchen Maßnahmen diese Schwankungen zu kontrollieren und zu minimieren sind.

Fazit

Die *residual sum of squares* sind ein statistischer Wert, der z.B. bei der linearen Regression Anwendung findet. Seine Bedeutung wird oft vernachlässigt, gleichzeitig kann er bei falscher Interpretation für Missverständnisse sorgen. So wie der Mittelwert die Eigenschaften vieler Datenpunkte auf einen Wert zusammenfasst, gehen gleichzeitig die Information über die Einzelwerte verloren. Bei den *RSS* gehen die einzelnen Beiträge der Fehlerquadrate verloren. Diese müssen daher separat untersucht werden, um die *RSS* korrekt interpretieren zu können. Statistische Kennzahlen wie die *hat values* bzw. *leverage*, oder *Cooks Distance* helfen, die Beiträge der Einzelwerte korrekt einzuordnen und somit auch der richtigen Einschätzung der *residual sum of squares*.

Bei der analytischen Methodenvalidierung müssen neben den allgemeinen Kennzahlen auch die *RSS* angegeben werden, wobei in entsprechenden Guidelines einerseits (zurecht) keine Angaben zu Obergrenzen der *RSS* gemacht werden, andererseits die Verwendung der Residuenanalyse mit dem Wort *may* nicht als zwingend erforderlich dargestellt wird. Die grafische Analyse der Residuen kann in offensichtlichen Fällen ausreichend sein, jedoch ist sie nicht in der Lage, potentielle einflussreiche Datenpunkte zu ermitteln. Doch gerade diese *influencial observations* können die *fitness-for-purpose* in Bezug auf die Linearität der Methode unter Umständen in Gefahr bringen, da diese den aufgestellten linearen Zusammenhang zwischen Analytenkonzentration und Signal verzerren können. Visuelle Analysen haben immer subjektiven Charakter - „*scientific justification*“ hingegen kann nur mit objektiven Kriterien wie z.B. *Cooks Distance* oder ähnlichen Residuenanalyseverfahren erreicht werden.

Die *residual sum of squares* stellen somit bei der Auswertung von (z.B. linearen) Regressionsverfahren den Startpunkt einer interessanten Reise dar, an dessen Ende der Erkenntnisgewinn steht, was der (z.B. lineare) Zusammenhang tatsächlich Wert ist. Damit sind die *residual sum of squares* nicht nur für analytische Methodenvalidierung von großer Bedeutung.

RSS, leverage und Cooks Distance in Excel 2016

Vorbereitungen

In Excel 2016 lassen sich oben beschriebene Vorgehensweisen nachrechnen: Oben besprochenes Beispiel enthält die Werte aus der Publikation von Francis Anscombe [3]. Diese beinhalten das berühmte *Anscombe-Quartett*, das aus 4 Datensätzen besteht, die interessanterweise allesamt die gleichen statistischen Merkmale bezüglich der linearen Regression aufweisen. Verwendet sind hier die Daten aus dem ersten der vier Datensätze (Tabelle 5).

Tabelle 5: Für das Beispiel verwendeter Datensatz

Beobachtung	X	Y
1	4	4,26
2	5	5,68
3	6	7,24
4	7	4,82
5	8	6,95
6	9	8,81
7	10	8,04
8	11	8,33
9	12	10,84
10	13	7,58
11	14	9,96

Die X und Y Werte können leicht nach Excel kopiert werden. Werden beispielsweise in Spalte A die X Werte (oder Konzentration) und in Spalte B die Y Werte (Signal) eingetragen, kann man über die Registerkarte „**Daten**“ und anschließend mit Klick auf „**Datenanalyse**“ im sich öffnenden Fenster „**Regression**“ auswählen. Dort trägt man die X- und Y-Bereiche der Werte ein. Anschließend setzt man noch den Haken für die Angabe der Residuen und bestätigt mit „OK“. In dem sich öffnenden neuen Sheet befinden sich folgende Angaben in folgenden Zellen:

Tabelle 6: Angabe der Merkmale in Excel zur linearen Regression

X-Werte	Zelle	Wert	Formel / Symbol
Multipler Korrelationskoeffizient (*correlation coefficient*)	B4	0,816	r
Bestimmtheitsmaß	B5	0,666	R^2 $(R^2 = r * r)$
Anzahl Beobachtungen	B8	11	n
***RSS* (*residual sum of squares*)**	B14	13,762	$RSS = \sum_i \varepsilon_i^2 = (y_i - \hat{y}_i)^2$
Achsenschnittpunkt (*y-intercept*)	B17	3,000	Y Wert für $X = 0$, β_0
Anstieg der Geraden (*slope*)	B18	0,500	β_1
Schätzwerte für Y	B25:B35		$\hat{y}$
Residuen (*residuals*)	C25:C35		$\varepsilon = y - \hat{y}$

Es bietet sich nun an, die Ausgabe der Regression mit den Originaldaten zusammenzufügen, z.B. durch Anhängen der Originaldaten zu der Ausgabe von Excel:

	A	B	C	D	E
24	*Beobachtung*	**Schätzwert Y**	**Residuum**	**Konzentration**	**Signal**
25	1	5,000	-0,740	4	4,26
26	2	5,501	0,179	5	5,68
27	3	6,001	1,239	6	7,24
28	4	6,501	-1,681	7	4,82
29	5	7,001	-0,051	8	6,95
30	6	7,501	1,309	9	8,81
31	7	8,001	0,039	10	8,04
32	8	8,501	-0,171	11	8,33
33	9	9,001	1,839	12	10,84
34	10	9,501	-1,921	13	7,58
35	11	10,001	-0,041	14	9,96

Abbildung 8: Ausgangsformat der Daten für lineare Regression und Residuenanalyse

So stehen nun die Schätzwerte für Y in Spalte B, die Residuen in Spalte C, und die originalen X und Y Werte in Spalten D und E.

Um die D Werte bzw. *Cooks Distance* Werte für alle Datenpunkte zu berechnen, benötigt man folgende Formel: $D_i = \frac{isr_i^2}{p} * \frac{h_i}{1-h_i}$. Dabei fällt auf, dass man zwei Zwischenschritte benötigt, um D berechnen zu können. Im ersten Schritt benötigt man die Werte für den *leverage*, h_i, und anschließend die quadrierten isr Werte. Grundlage dafür sind jedoch die *residuals*, aus denen sich auch die *residual sum of squares* berechnen lassen.

Berechnung der *Residual Sum of Squares*

Der Fehler, der *residual*, für den ersten Datenpunkt kann auch manuell über die Formel: E25 - B25 berechnet werden. Das Fehlerquadrat ergibt sich entsprechend als (E25 – B25)^2. Im Beispiel können so die manuell erstellen Residuen in Zellen F25:F35 und die Residuenquadrate in G25:G35 berechnet werden (siehe Abbildung 9). Die Summe der Residuenquadrate, die *RSS* beziehungsweise die *residual sum of squares* sind in Zelle F37 gespeichert. Der Wert von 13,76 stimmt mit dem Wert in Zelle B14 überein.

Berechnung der *leverage* bzw. H Werte

Für die weiteren Berechnungen benötigt man noch den Mittelwert der X Werte ($\bar{x}$), sowie die Standardabweichung der X Werte (s_x). Den Mittelwert der X Werte erhält man durch MITTELWERT(D25:D35), dieser wird in Zelle D37 gespeichert. Die Standardabweichung der X Werte erhalten wir durch STABW.S(D25:D35) und Speichern in Zelle D38. Diese Werte benötigen wir für die Berechnung des *leverage* bzw. der *hat values*:

Die Formel für die Berechnung lautet: leverage oder $h_i = \frac{1}{n} + \frac{1}{n-1} * \left(\frac{x_i - \bar{x}}{s_x}\right)^2$. In Excel nutzt man die eben berechneten Werte für $\bar{x}$ und s_x, so lautet die Formel für die Berechnung des *leverage* in Excel für den ersten Datenpunkt in Zelle D25: (1/$B$8) + 1/($B$8-1) * (($D25 - $D$37) / $D$38)^2. Auch hier erkennt man, dass der *leverage* nur von den originalen X Werten (aus der D-Spalte) und von der Lage der restlichen X Werte (aus Zellen D37 und D38) abhängig ist.

H25 =(1/$B$8)+1/($B$8-1)*(($D25-$D$37)/$D$38)^2

	A	B	C	D	E	F	G	H
24	*Beobachtung*	Schätzwert Y	Residuum	Konzentration	Signal	residuals	residuenquadrat	leverage
25	1	5,000	-0,740	4	4,26	-0,740	0,548	0,318
26	2	5,501	0,179	5	5,68	0,179	0,032	0,236
27	3	6,001	1,239	6	7,24	1,239	1,536	0,173
28	4	6,501	-1,681	7	4,82	-1,681	2,825	0,127
29	5	7,001	-0,051	8	6,95	-0,051	0,003	0,100
30	6	7,501	1,309	9	8,81	1,309	1,714	0,091
31	7	8,001	0,039	10	8,04	0,039	0,002	0,100
32	8	8,501	-0,171	11	8,33	-0,171	0,029	0,127
33	9	9,001	1,839	12	10,84	1,839	3,381	0,173
34	10	9,501	-1,921	13	7,58	-1,921	3,691	0,236
35	11	10,001	-0,041	14	9,96	-0,041	0,002	0,318
36								
37			Mean X	9		RSS	13,763	
38			SD X	3,317				

Abbildung 9: Schritte für die Berechnung des leverage

Berechnung der *Cooks Distance* bzw. *D* Werte

Den *leverage* benötigt man nun, um den nächsten Schritt für die Berechnung der *Cooks Distance* zu tätigen. Der Schritt umfasst die Berechnung eines sogenannten isr Wertes, den man für die *Cooks Distance* benötigt. Dieser ergibt sich aus der Formel: $\frac{\varepsilon_i}{s_E\sqrt{1-h_i}}$. Die Werte ε_i für sind bereits in Spalte F bzw. Spalte C vorhanden. Die Werte für h_i sind die *leverage* Werte in Spalte H. Den Wert s_E kann man mit WURZEL(($B$8-1)/($B$8-2)*STABW.S(F25:F35)^2) bestimmen und diesen in Zelle D39 speichern. Die Formel für den ersten Datenpunkt (Reihe 25) ist daher für den isr: F25/$D$39/WURZEL(1-$H25), alle isr Werte werden in Spalte I gespeichert.

Daraus können nun im letzten Schritt die D Werte aus *Cooks Distance* berechnet werden. Die *Distance* D für Datenpunkt i ergibt sich nun als: $D_i = \frac{isr_i^2}{p} * \frac{h_i}{1-h_i}$. Für Datenpunkt 1 aus Reihe 25 ergibt sich dafür folgende Formel: $I25^2 / 2 * $H25 / (1-$H25). Die Werte werden in Spalte J gespeichert. Damit hat man nun alle Daten zusammen, um nicht nur die Regression, sondern auch die Analysis zu den Abweichungen der einzelnen Datenpunkte vornehmen zu können.

Folgende Grafik gibt nochmals eine Übersicht über alle Berechnungsschritte:

	A	B	C	D	E	F	G	H	I	J	K	L	M
22	AUSGABE: RESIDUENPLOT			Kopierte Originaldaten		(1/$B$8)+1/($B$8-1)*(($D25-$D$37)/$D$38)^2				F25/$D$39/WURZEL(1-$H25)			
23													
24	*Beobachtung*	Schätzwert Y	Residuum	Konzentration	Signal	residuals	residuenquadrat	leverage	isr	Cooks Distance			
25	1	5,000	-0,740	4	4,26	-0,740	0,548	0,318	-0,7252	0,1227		$I25^2/2*$H25/(1-$H25)	
26	Ergebnis aus Anwendung der Regression in Excel 2	5,501	0,179	5	5,68	0,179	0,032	0,236	0,1661	0,0043			
27	3	6,001	1,239	6	7,24	1,239	1,536	0,173	1,1019	0,1268			
28	4	6,501	-1,681	7	4,82	-1,681	2,825	0,127	-1,4549	0,1543			
29	5	7,001	-0,051	8	6,95	-0,051	0,003	0,100	-0,0433	0,0001			
30	6	7,501	1,309	9	8,81	1,309	1,714	0,091	1,1103	0,0616			
31	7	8,001	0,039	10	8,04	0,039	0,002	0,100	0,0332	0,0001			
32	8	8,501	-0,171	11	8,33	-0,171	0,029	0,127	-0,1481	0,0016			
33	9	9,001	1,839	12	10,84	1,839	3,381	0,173	1,6349	0,2790			
34	10	9,501	-1,921	13	7,58	-1,921	3,691	0,236	-1,7779	0,4892			
35	11	10,001	-0,041	14	9,96	-0,041	0,002	0,318	-0,0405	0,0004			
36													
37	MITTELWERT(D25:D35)		Mean X	9		RSS	13,763						
38	STABW.S(D25:D35)		SD X	3,317									
39	WURZEL(($B$8-1)/($B$8-2)*STABW.S(F25:F35)^2)		S_E	1,237			SUMME(G25:G35)						

Auswertung erfolgt von links nach rechts: von der Berechnung der Residuen, über die Residuenquadrate, zu *leverage*, isr und schließlich zu *Cooks Distance*

Abbildung 10: Übersicht über alle verwendeten Formeln und Zwischenschritte zur Berechnung der Cooks Distance, ausgehend von den Originaldaten, sowie dem Ergebnis der Regression in Excel 2016

Quellen

[1] **Q2(R1) Validation of Analytical Procedures: Text and Methodology:** http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q2_R1/Step4/Q2_R1__Guideline.pdf

[2] **Cooks Distance:** Cook, R. Dennis (February 1977). "Detection of Influential Observations in Linear Regression". Technometrics. American Statistical Association. 19 (1): 15–18.

[3] **Anscombes Quartett:** F. J. Anscombe: *Graphs in Statistical Analysis*. In: *American Statistician*. 27, Nr. 1, 1973, S. 17–21.

Über den Autor

Dr. Peter P. Heym hat als Gastautor diesen Beitrag für die Lösungsfabrik geschrieben. Er hat Bioinformatik studiert und am Leibnitz-Institut für Pfanzenbiochemie Halle mit dem Thema „*In silico* characterisation of *At*PARP1 and virtual screening for *At*PARP inhibitors to increase resistance to abiotic stress“ (computergestütztes Inhibitordesign) promoviert. Er ist der Inhaber von Sum Of Squares - Statistical Consulting (www.sumofsquares.de), einem Dienstleistungsunternehmen, welches sich auf statistische Beratung für Studenten, Privatpersonen, Firmen und Unternehmen spezialisiert hat. Neben statistischer Beratung runden Unterstützung bei universitären Abschlussarbeiten, Betreuung und Auswertung von Umfragen, Workshops (z.B. zur Programmiersprache R), Seminare, Fortbildungen, auch im GMP-Bereich, das vielfältige Angebot ab.