Der Tanz der Residuen –

Erarbeitung statistischer Grundbegriffe mit Hilfe von EXCEL

BERND RECKELKAMM, BIELEFELD

Zusammenfassung: *Mit Hilfe eines Tabellenkalkulationssystems wie z.B. EXCEL können grundlegende Begriffe der Statistik in ganz neuer Form erarbeitet und insbesondere visualisiert werden. Dieser Aufsatz liefert zum einen Beispiele, die in das Thema „Explorative Datenanalyse" einführen. Zum anderen gibt er Anregungen für das Erarbeiten von Begrifflichkeiten und deren Zusammenhängen bei dem umfangreichen Thema „Regression und Korrelation". EXCEL erweist sich hier als ein variationsreiches Hilfsmittel, das auf eine Vielzahl unterschiedlicher Unterrichtssituationen und ihrer jeweiligen Erfordernisse elastisch reagiert. Die vorgestellten Materialien stehen im www zur Verfügung* [1]

1 Daten visualisieren

Als Einstieg bietet sich die Frage an: „Wie lang ist eine Minute?" In Zweiergruppen versuchen die SchülerInnen jeweils, die Länge einer Minute zu „erfühlen", ohne auf die Uhr zu schauen. Der jeweils andere Partner gibt dem Probanden eine Rückmeldung (faktische Dauer der „gefühlten" Minute), sodass der Proband versuchen kann, im 2. Versuch die 60 Sekunden besser zu schätzen. Wie lässt sich dieses Datenmaterial sinnvoll visualisieren? Welche Informationen ergeben sich aus den visualisierten Daten? Die folgenden Rohdaten (hier von 22 SchülerInnen) müssen dazu in eine geeignete neue Form gebracht werden.

Laufende Nummer	Name	1. Versuch		2. Versuch	
		Messwert	Abweichung von 60 sec	Messwert	Abweichung von 60 sec
1	Arne	40	-20	62	2
2	Bettina	58	-2	83	23
3	Carla	54	-6	57	-3
4	Dorian	76	16	70	10
5	Ernst	76	16	56	-4

Tab. 1: Auszug aus den Rohdaten zum Schülerexperiment ‚Schätzen von 60 Sekunden'

Als Standardveranschaulichung bietet sich ein Häufigkeitsdiagramm an, das deutlich unterschieden den ersten und den zweiten Datensatz enthält. [2] Die Klassenbreiten sind frei wählbar, im vorliegenden Fall sind es 10 Sekunden als Intervallbreite mit den Vielfachen von 10 als Intervallgrenzen. [3]

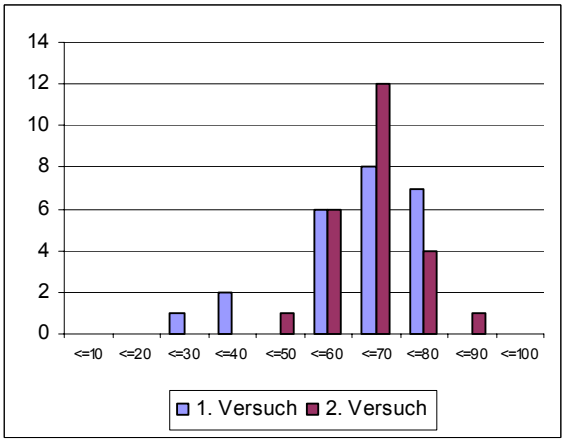


Fig. 1: Vergleich der Schätzungen aus 1. und 2. Versuch – Zielwert 60 als Klassengrenze

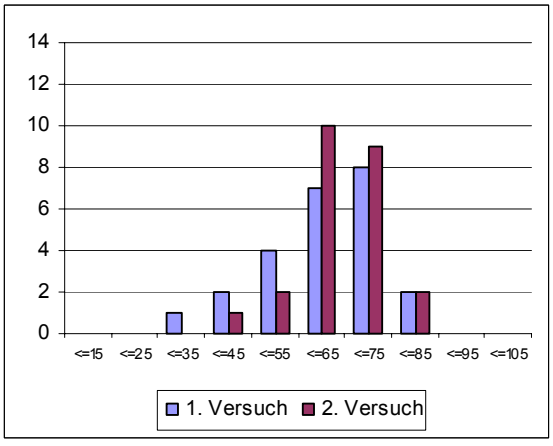


Fig. 2: Vergleich der Schätzungen aus 1. und 2. Versuch – Zielwert 60 als Klassenmitte

Eine erste Analyse kann die Symmetrie ins Auge fassen (Fig. 1). Dabei fällt auf, dass der 2. Versuch (dunklere Säulen) ein „symmetrischeres“ Bild liefert als der 1. Eine zweite Analyse deutet darauf hin, dass im 2. Versuch die Daten weniger „streuen“.

An dieser Stelle kann thematisiert werden, wie Intervallbreiten und –mitten sinnvoll zu wählen sind. Die Darstellung in Fig. 2 zeigt z.B. die Auswertung, wenn die „Zielzahl“ 60 nicht Intervallgrenze sondern Intervallmitte ist.

Als weitere Darstellungsweise bieten sich Boxplots an. Diese werden von EXCEL allerdings nicht direkt unterstützt. Im www erhält man zwar plug-ins, diese sind jedoch nicht gut in der Handhabung. Will man dennoch nicht auf dieses Mittel verzichten, so kann man die benötigten Daten von EXCEL berechnen lassen und daraus eine „handgestrickte“ Boxplot-Version erstellen (siehe Fig. 3)

Max	76	83
Q3	71,25	70
MED	63	64
Q2	54,75	59,25
Min	28	42
Spannweite	48	41

Tab. 2: Kennziffern zum Vergleich von 1. und 2. Versuch

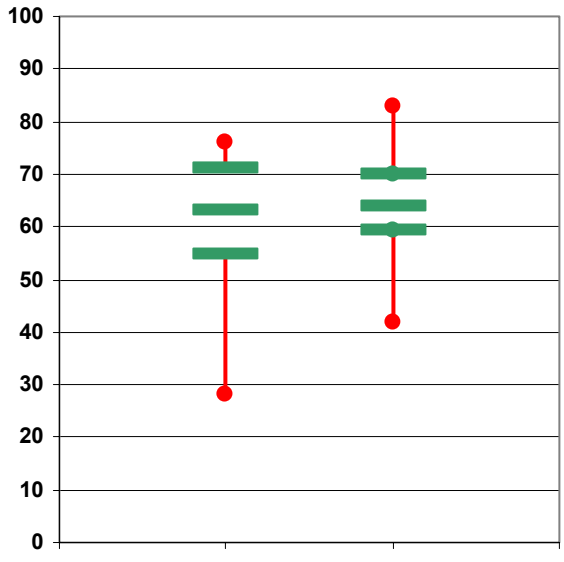


Fig. 3: Boxplot zum Vergleich von 1. und 2. Versuch

Die oben geäußerten Vermutungen werden nun deutlicher sichtbar: Der rechte (2.) Datensatz streut deutlich weniger, besonders bei seinen mittleren 50%, liegt aber mit seinem Median *weiter* von den 60 Sekunden weg. Damit ergibt sich als neue Fragestellung, wie sich diese Streuungen rechnerisch behandeln lassen.

Will man diesen Fragestellungen rund um die Visualisierung und Auswertung von Daten *intensiver* nachgehen [4], so sollte man auf eine andere Software ausweichen, z.B. FATHOM, da EXCEL bei größeren und variablen Datenmengen sowie unterschiedlichen Fragestellungen eher unhandlich ist. Dies gilt sowohl für die Arbeit mit Häufigkeitsdiagrammen als auch für die Boxplots, die in der hier angesprochenen Version keine Variationen, z.B. Tukey- oder Cleveland-Boxplots erlauben.

2 Streuungsmaße

Einen naheliegenden Zugang zum rechnerischen Vergleich von Streuungen bilden die eben angeführten Quartile. Sowohl in der reinen Zahldarstellung als auch in der graphischen Darstellung innerhalb eines Boxplots liefern sie sinnvolle Auskünfte über die beiden Verteilungen.

Einen weiteren Zugang zeigen die zwei folgenden Fragen:

a) Welches Abstandsmaß wollen wir wählen? In der Regel bieten sich an der orientierte, der absolute oder der quadrierte Abstand, jeweils gewichtet oder ungewichtet.

b) Um welche Kenngröße soll die Streuung betrachtet werden? In Frage kommen z.B. Median oder das arithmetische Mittel der Verteilung, oder in obiger Situation die „Zielzahl“ 60.

Jede Kombination aus diesen Gesichtspunkten lässt sich mit Hilfe von EXCEL schnell und komfortabel durchspielen. Allerdings sind dies in der Regel keine Standardanwendungen, die auf „Press-Button“ funktionieren. Der folgende Tabellenausschnitt (Fig. 4) zeigt jeweils drei aufsummierte Abstandswerte, ungewichtet, für den Mittelwert (xq) und den Median (xs). Solange wir nur einen einzelnen Datensatz untersuchen, gibt es keine Not, die aufsummierten Werte zu gewichten. EXCEL lässt uns – und das heißt vor allem: den SchülerInnen – die Wahl.

	A	B	C	D	E	F	G	H	I	J	K
1											
2				xq	61,3			xs	63,0		
3				0,0	227,0	3542,5		-42,0	220,0	3616,0	
4											
5	Nr.	Name	xi	xi-xq	abs(xi-xq)	(xi-xq)^2		xi-xs	abs(xi-xs)	(xi-xq)^2	
6	1	Arne	40	-21,3	21,3	451,6		-23,0	23,0	529,0	
7	2	Bettina	58	-3,3	3,3	10,6		-5,0	5,0	25,0	
8	3	Carla	54	-7,3	7,3	52,6		-9,0	9,0	81,0	
9	4	Dorian	76	14,8	14,8	217,6		13,0	13,0	169,0	

Fig. 4: Ausschnitt aus dem Tabellenblatt zur Berechnung der Abweichungen vom Mittelwert xq und vom Median xs

Variiert man die obigen Daten (bzw. eine Auswahl) eine Weile, so ergeben sich die klassischen Vermutungen:

- Beim Mittelwert addieren sich die orientierten Abweichungen immer zu Null.
- Die Summe der absoluten Abweichungen (SAA) ist beim Median kleiner als bei beliebigen Konkurrenten.
- Beim Mittelwert ergibt sich diese Eigenschaft für die Summe der quadratischen Abweichungen (SQA).

Um die letzte dieser Eigenschaften zu untersuchen, wechseln wir zu einem neuen, kleinen Datensatz. Die folgende Tabelle (Fig. 5) zeigt in B2 bzw. E2 den Mittelwert xq = 5. In F3 liest man ab, dass die Summe der Quadratischen Abweichungen von xq = 5 den Wert 88 hat, kurz SQA(5) = 88. Diese Kurzschreibweise inspiriert den Zugang über die *Funktion* SQA. Diese ordnet jeder Zahl die Summe der Quadratischen Abweichungen zu. Für x = 8 ergibt sich – s.u. – SQA(8) = 133. Offenbar hat der Graph sein Minimum an der Stelle x = 5 = xq.

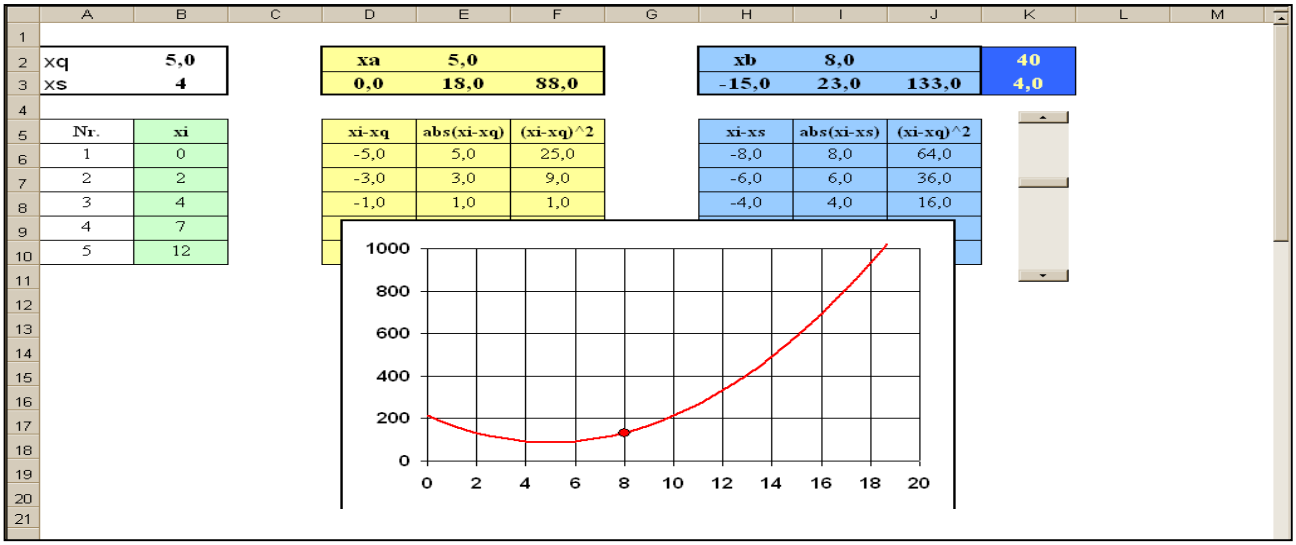


Fig. 5: Die Summe der quadratischen Abweichungen ist für den Mittelwert minimal

Im Zuge der Beschreibung der Tabellen und des Graphen ergeben sich die folgenden, vertiefenden Anschlussfragen von selbst:

- Ist der Graph zu SQA tatsächlich eine Parabel (2. Grades)?
- Kann man das Minimum auch algebraisch berechnen?

Die Wahl von nur 5 Punkten (Wertepaaren) ermöglicht es, die Gleichung der Funktion SQA konkret aufzustellen [– es *ist* tatsächlich eine quadratische Gleichung –] und anschließend das Minimum zu bestimmen [– mittels quadratischer Ergänzung und damit ohne Analysis, wenn das Thema zu Beginn der Jahrgangsstufe 11 behandelt wird –].

Je nach Unterrichtssituation kann man *nun* versuchen, den *allgemeinen* Nachweis für diese Eigenschaft des Mittelwertes zu führen. Da das Ziel und die Grundideen jetzt vorhanden sind, mag die

Durchführung mittels Summenzeichen und/ oder Pünktchen etwas mehr Aussicht auf Erfolg haben.

Eine Untersuchung zur entsprechenden Eigenschaft des Median kann sich anschließen. Vollständigkeit der Nachweise wird hier sicherlich weniger das Thema sein. Stattdessen stehen Visualisierung, Hypothesenbildung und das Erläutern von Zusammenhängen im Vordergrund.

An dieser Stelle kann die Untersuchung monovariater Verteilungen schließen und die Vorbereitung von Regressions- und Korrelationsbetrachtungen beginnen. Wir wechseln damit zu bivariaten Verteilungen. [5]

3 Das Streudiagramm

Von allen SchülerInnen des Kurses werden zunächst die persönlichen Daten *Körpergröße, Gewicht, (monatliches) Taschengeld und Schuhgröße* ermittelt und anschließend tabellarisch erfasst, s.u. (Bei Bedenken hinsichtlich der Persönlichkeitsrechte der SchülerInnen sollte man vorgefertigte Daten nehmen.)

Vergleich persönlicher Daten

		ki	gi	ti	si
Nr.	Name	Körpergröße	Gewicht	Taschengeld	Schuhgröße
1	Arne	180	66	120	45
2	Bettina	178	60	67	41
3	Carla	181	63	50	43,5
4	Dorian	162	55	40	38

Tab. 3: Auszug aus den Urdaten zu persönlichen Merkmalen der Schüler

Wie schon zuvor stellt sich auch hier die Frage nach einer angemessenen visuellen Aufbereitung der Daten. Über die Verteilungen der *einzelnen* Merkmale hinaus ist wird jetzt jedoch die Frage in den Mittelpunkt gestellt, ob es erkennbare Abhängigkeiten *zwischen* den Merkmalen gibt. Die kanonische Darstellung ist das Streudiagramm (auch: Punktplot) im klassischen Koordinatensystem (KOS). Geht man davon aus, dass die Körpergröße weitgehend unbeeinflussbar ist, kann man die anderen drei Merkmale z.B. in Abhängigkeit von der Körpergröße betrachten. Das ist beim Taschengeld sicherlich wenig sinnvoll und sollte sich entsprechend im Diagramm zeigen.

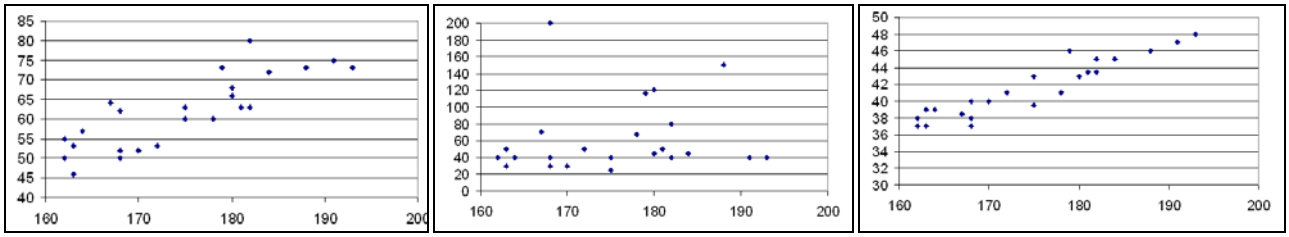


Fig. 6: Vergleich im Streudiagramm

–Körpergröße - Gewicht –Körpergröße - Taschengeld –Körpergröße - Schuhgröße

Auch ohne theoretischen Hintergrund wird in Fig. 6 deutlich, dass Bild 1 und Bild 3 einen starken linearen Zusammenhang aufzeigen, während Bild 2 (Taschengeld) eher „wolkig“ erscheint. Zudem scheint der lineare Trend in Bild 3 stärker zu sein als in Bild 1. Man beachte allerdings die jeweiligen Skalierungen, die zu Fehleinschätzungen führen können. Es ergeben sich direkt die folgenden naheliegenden Fragen:

- Kann man die Verteilung durch eine Gerade beschreiben? Ist eine Voraussage oder Berechnung von Zwischenwerten auf der Grundlage eines Geradenmodells möglich?
- Wie „gut“ gibt das Messergebnis die Stärke des linearen Zusammenhangs wieder? In welchem Maße ist eine Gerade „angemessen“?

Während die erste Frage das Thema der Regression ins Spiel bringt, verweist die zweite auf die Frage der Korrelation. Zumindest der ersten Frage gehen wir in diesem Beitrag vertieft nach.

4 Die Regressionsgerade

Wir beschränken uns auf die Untersuchung des ersten Bildes. Man kann zunächst „frei nach Augenmaß“ eine sog. Ausgleichsgerade in das Diagramm legen (hier: die untere Gerade) und diese dann um diejenige Gerade ergänzen (hier: die obere Gerade), die sich aus der Berechnung des Idealgewichts“ ergibt:

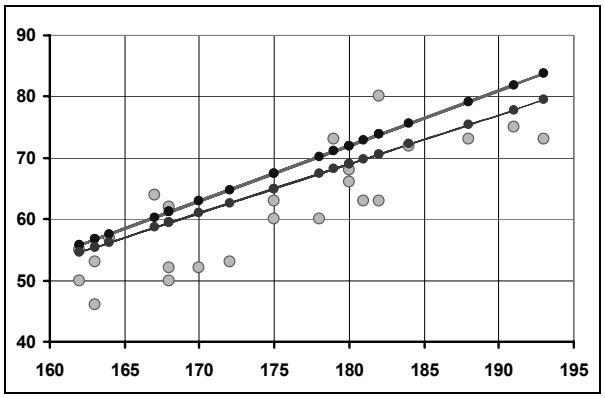


Fig. 7: Ausgleichsgerade zwischen Größe und Gewicht

Dabei verwenden wir die folgende, auch im Alltag benutzte und daher weitgehend bekannte Formel:

Idealgewicht (in kg)=
Körpergröße in kg minus 100, abzgl. 10%.

Welche Gerade beschreibt die vorliegende Punktwolke „besser“? Was heißt hier überhaupt „besser“? Und gibt es „noch bessere“ Geraden?

Zunächst liegt nahe, jeweils die „Abstände“ der Punkte von den Geraden aufzusummieren und zu vergleichen. Offen ist jedoch, ob es sich um lotrechte Abstände zur Geraden im Sinne „kürzeste Verbindung“ oder Parallelstrecken zur y-Achse [6] handeln soll. Da die Ausgleichsgerade Werte liefert, die man mit den tatsächlichen Werten vergleichen möchte (Differenz von Funktionswerten), bietet sich letzteres an. Zudem ist diese Variante auch rechnerisch deutlich einfacher als die erste. – ein nicht notwendigerweise schlagendes Argument, aber eins, das die Schüler sofort „überzeugt“ [7]

Um einen besseren anschaulichen Eindruck zu gewinnen, nehmen wir zwei Veränderungen vor:

- Wir untersuchen weitere Geraden, beschränken uns aber auf Geraden, die durch den Schwerpunkt der Verteilung verlaufen. Diese Einschränkung ist „harmlos“, allerdings führt man den Nachweis in der Schule in der Regel nicht. Der Vorteil besteht in der Beschränkung auf 1 Parameter, nämlich die Steigung der Geraden. Mit Hilfe der Punktsteigungsform lässt sich diese Gerade bequem in EXCEL plotten und variieren.
- Wir plotten die senkrechten Abstände zwischen den Punkten der Verteilung und der Ausgleichsgeraden, die sog. Residuen, in einem neuen Koordinatensystem. Der Vorteil dieser ergänzenden Darstellung besteht im senkrechten Auftreffen der Abstandslinien, die hier tatsächlich Lote sind, was im ersten Koordinatensystem i.a. nicht der Fall ist.

Drehen wir in Fig. 8 nun die Ausgleichsgerade im Schwerpunkt der Verteilung, indem wir den Steigungsparameter variieren, so erhalten wir im Residuenplot (also im rechten KOS) einen Eindruck davon, ob die Gerade jeweils „gut liegt“. Dabei ist durchaus zu diskutieren, an welchen Merkmalen „gut liegen“ festgemacht werden soll. Es bieten sich an: die Anzahl der Punkte oberhalb bzw. unterhalb der x-Achse, ein „ausbalancierter“ Gesamteindruck oder die Lage von möglichen Ausreißern.

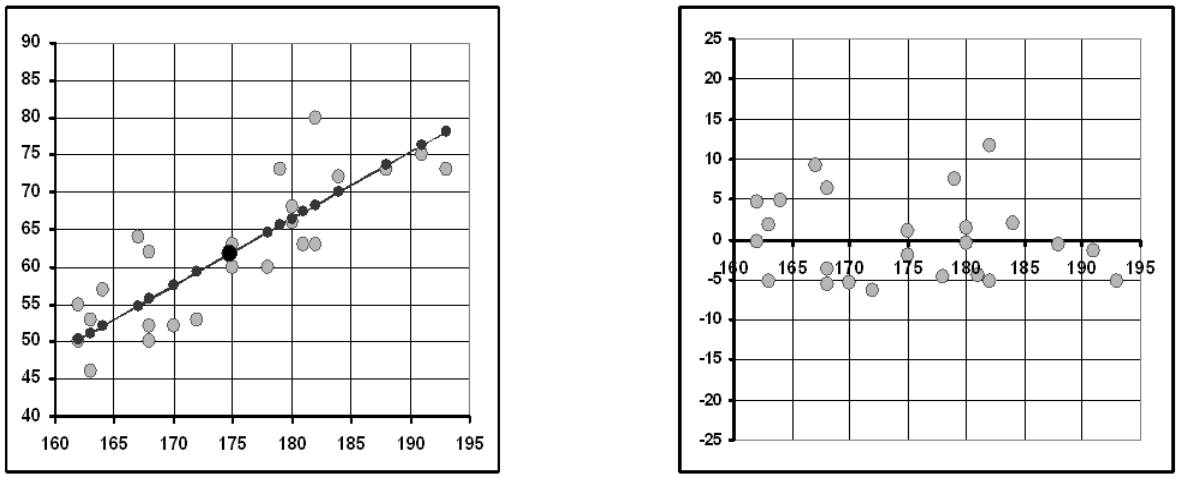


Fig. 8: Betrachtung der Residuen im ursprünglichen Koordinatennetz und im Residuenplot

Der optische Eindruck liefert zunächst einmal einen plausiblen Wert für die Steigung der optimal liegenden Ausgleichsgeraden. Damit ist der Begriff der *Regressionsgeraden* hinreichend motiviert. Mit Hilfe der Punkt-Steigungs-Form kann auch sofort eine Funktionsgleichung angegeben werden. Damit

lässt sich – noch vor aufwändigen weiteren Analysen – die Grundidee der Regressionsanalyse vertiefen: Rückführung eines Merkmals auf ein anderes. Was von der Regressionsgeraden (dem Funktionswert an einer bestimmten Stelle) abweicht, ist ein nicht zu erklärender Rest, ein *Residuum.* Dieser Zusammenhang erklärt den Ausdruck. [8]

Allerdings ist immer noch unklar, wie ein rechnerisches Kriterium für „gute“ bzw. „optimale“ Lage der Ausgleichsgeraden aussehen soll. Hier hilft der Rückblick auf die oben behandelte Optimalitätseigenschaft des Arithmetischen Mittels: Man minimiert die Summe von quadratischen Abweichungen, da der Schwerpunkt als arithmetisches Mittel berechnet wurde. Dieser Zugang ist die bekannte *Methode der kleinsten Quadrate* [9]: Zu jeder Steigung m gehört eine Gerade g_m. Diese liegt in der Regel nicht genau auf den Punkten der Verteilung. Quadriert man die senkrechten Abstände und summiert sie auf, erhält man eine neue Funktion SQA:

$$SQA: m \rightarrow SQA(m)$$

Da wir in dieser Phase zunächst eine spezielle Situation untersuchen, ist es nicht erforderlich, eine Gewichtung vorzunehmen. In der Tabelle von Fig. 9 werden diese Zusammenhänge erarbeitet: Im dritten KOS ist der Funktionswert SQA(1,02) = 888 angezeigt. Mit Hilfe des Schiebereglers oben rechts wird m variiert, entsprechend wird SQA(m) jeweils neu berechnet: Der Punkt im rechten System bewegt sich auf einer parabelähnlichen Linie.

Das Minimum der Funktion SQA lässt sich nun in der Zelle F2 *und* im rechten KOS ablesen. Damit sind wir in der Lage, den *Regressionskoeffizienten*, d.i. die Steigung der Regressionsgeraden, auf unterschiedliche Weisen, deren Zusammenspiel aber hier augenfällig bleibt, zu bestimmen:

- anschaulich durch die Lage der Residuen im 2. KOS,
- rechnerisch durch die Summenbildung (hier) in Zelle F2,
- graphentheoretisch durch das Minimum einer Parabel (wenn es denn eine ist).

Insofern ist die nachfolgende EXCEL-Anwendung der Kern der gesamten Reihe zur Regression. Wie schon oben bei den Minimierungen rund um die Kennzahlen ist jetzt die Grundlage gegeben, auf der man den Regressionskoeffizienten sinnvoll auch allgemein herleiten kann. Dies geschieht dann aber nicht mehr über die Köpfe der Schüler hinweg, sondern immer in Auseinandersetzung mit dem hier gezeigten Material. Die um die Regressionsgerade tanzenden Residuen zeigen auf besonders intuitive Weise die Kernidee des Verfahrens.

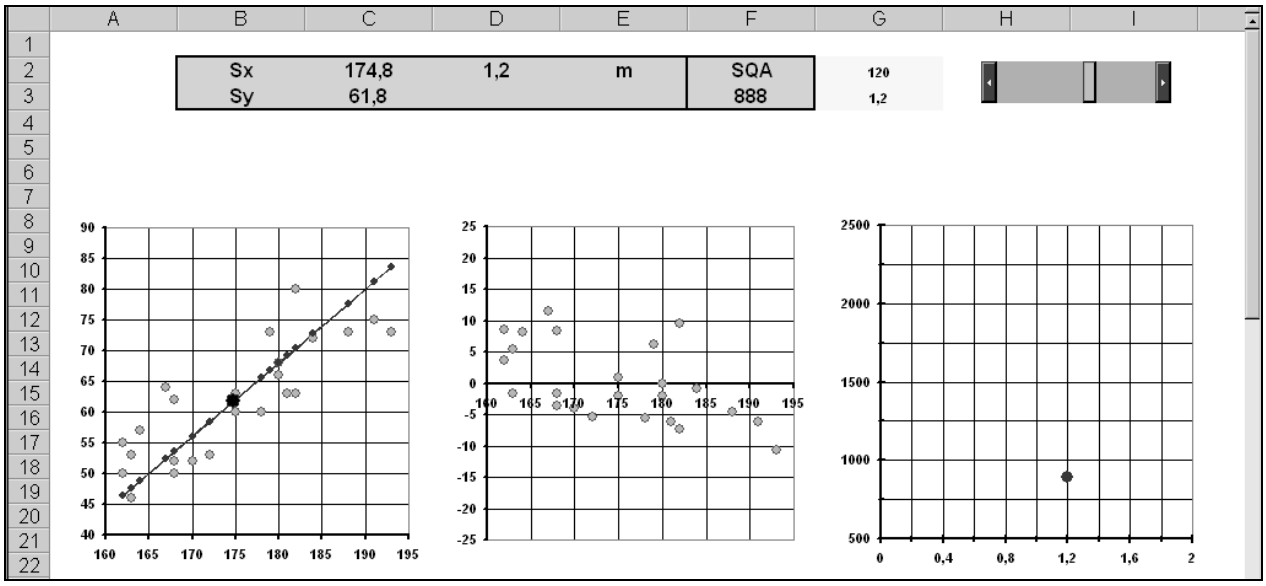


Fig. 9: Tabellenblatt zum ‚Tanz der Residuen’

Kommentar: *Bei diesem Tabellenblatt in Fig. 9 handelt es sich um den Schluss einer sich nach und nach entwickelnden Sequenz von Tabellenblättern, die in einer (1) Mappe gespeichert werden. Die Technik, Tabellen schrittweise in neuen Blättern zu erweitern, ist in hohem Maße geeignet, komplexere Situationen mit Hilfe von EXCEL zu bearbeiten und trotzdem im Nachhinein nachvollziehbar machen zu können, wie die Schlusstabelle zustande gekommen ist. Die Vorstufen zu dem obigen Blatt waren: Originaldatensatz zu Größe und Gewicht, Streudiagramm (Punktplot), Ausgleichsgerade im Streudiagramm (mit m und b als Parametern), Residuenplot, Berechnung der Summe der Abweichungsquadrate, Einführung des Schwerpunkts (nur noch m als Parameter), Einbau eines Schiebers zur „kontinuierlichen“ Änderung der Werte von m, Verschieben störender Elemente im Bild.*

5 Korrelation

Zu *jedem* Punktplot erhält man nach obigem Verfahren eine Regressionsgerade. Über den linearen Charakter der Punktwolke sagt das Ergebnis daher *nichts* aus. Ein Maß für deren „Linearität“ und damit die Güte und Anwendbarkeit der Regressionsgeraden ist die Korrelation bzw. der Korrelationskoeffizient. Einen sinnvollen Zugang bildet der Weg über die zweite Regressionsgerade: Statt die Summe der Abstandsquadrate in y-Richtung zu minimieren, minimiert man sie in x-Richtung. [10]

Als Einstieg in diese neue Betrachtung kann ein methodenkritischer Rückblick auf das bisherige Vorgehen dienen: Das Kriterium für die optimale Lage der Ausgleichsgeraden waren die senkrechten Abstände im Sinne von Verbindungslinien. Das war zwar plausibel und praktisch, aber geometrisch keineswegs zwingend. Die waagerechten Abstände sind geometrisch genauso überzeugend. Frage: Liefern sie dieselbe Regressionsgerade?

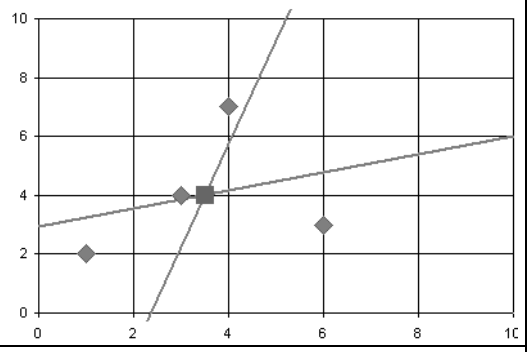
Tauscht man die x- und die y-Koordinaten aus und plottet diese Daten in einem eigenen KOS, so kann man in diesem neuen KOS den schon benutzen Regressionskalkül anwenden. Die so erhaltene 2. Regressionsgerade muss dann nur wieder in das erste KOS eingefügt werden. Da die neue Regressionsgerade durch denselben Schwerpunkt wie die erste geht, reicht es dabei aus, im alten KOS den Kehrwert des ermittelten zweiten Regressionskoeffizienten zu benutzen.

Jetzt lässt sich mit Hilfe einer geeigneten EXCEL-Tabelle (s. Fig. 10) studieren, wie unterschiedliche Punktanordnungen die Lage der beiden Regressionsgeraden beeinflussen.

Situation 1

i	xi	yi
1	1	2
2	6	3
3	3	4
4	4	7

ax =	0,308
ay =	0,286
1/ay =	3,500
r =	0,296



Situation 2

i	xi	yi
1	1	9
2	2	7
3	5	3
4	7	3

ax =	-1,033
ay =	-0,870
1/ay =	-1,149
r =	-0,948

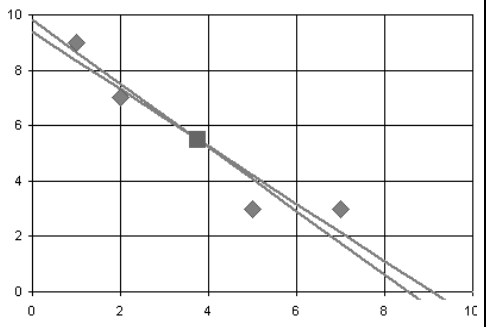


Fig. 10: Der Winkel zwischen den Regressionsgeraden ‚entspricht’ der Korrelation

***Kommentar**: Didaktische Kernidee dieser EXCEL-Mappe: Durch Variieren der 4 Ausgangspunkte lassen sich die Eigentümlichkeiten von Punktwolken und ihren Regressionsgeraden „in Echtzeit“ untersuchen. So wird z.B. deutlich, dass in Situation 2 nicht nur die zu erwartenden negativen Regressionskoeffizienten auftauchen, sondern dass auch r negativ ist und sehr nahe bei –1 liegt. Nicht abgebildet ist derjenige Teil des Tabellenblattes, in dem schrittweise die arithmetischen Mittel, die Differenzen, ihre Quadrate, deren Summen etc. berechnet werden. Alle durchgeführten Berechnungen verwenden die im Unterricht erarbeiteten Formeln. Die von EXCEL zur Verfügung gestellten Formeln werden nicht benutzt, da sie die mathematischen Zusammenhänge nicht erkennen lassen.*

Je „linearer“ der Charakter der Wolke, desto kleiner der Winkel zwischen den beiden Geraden, sodass im Extremfall die beiden Regressionsgeraden übereinstimmen. In diesem Sonderfall ergibt sich - hier in den Bezeichnung von *Griesel / Postel* - für die Regressionskoeffizienten [11]:

$$m_x = \frac{1}{m_y} \Leftrightarrow \frac{s_{xy}}{s_{xx}} = \frac{s_{yy}}{s_{xy}} \Leftrightarrow \frac{s_{xy} \cdot s_{xy}}{s_{xx} \cdot s_{yy}} = 1$$

Der entscheidende Term für das Studium der rechnerischen Zusammenhänge ist offenbar der letzte. Zieht man dort „vorsichtig“ die Wurzel, erhält man:

$$r := \frac{s_{xy}}{\sqrt{s_{xx}} \cdot \sqrt{s_{yy}}} = \frac{\text{Kovarianz}}{\text{Standardabweichungen}}$$

Mit Hilfe der genannten EXCEL-Datei kann nun zusätzlich untersucht werden, wie die Veränderungen in einer Punktwolke das Verhalten von r beeinflussen: positiv, negativ, $|r| \le 1$. Beschränkt man sich „zu Studienzwecken“ zunächst auf eine Wolke von nur vier Punkten, können sämtliche Daten tabellarisch erfasst bzw. aus den Grunddaten konkret berechnet werden. EXCEL zeigt „in Echtzeit“ die jeweils neu berechnete Situation.

Anmerkungen

[1] Die vorgestellten Materialien, insbesondere die EXCEL-Dateien sind im www zu finden auf der Homepage der Fachschaft Mathematik des Helmholtz-Gymnasiums, Bielefeld:
`www.helmholtz-bi.de` bzw.
`www.helmholtz-bi.de/uangebot/faecher/mathe/start.htm`

[2] Es handelt sich bei dieser Erhebung um einen Gruppenvergleich kontinuierlicher Daten.

[3] Die Erstellung eines Häufigkeitsdiagramms in EXCEL geschieht in zwei Schritten. Zunächst werden in einer neuen Tabelle Klassengrenzen vorgegeben, zu denen EXCEL die jeweilige Häufigkeit des Eintretens errechnet. Diese Tabelle wird dann als EXCEL-Säulendiagramm dargestellt. Näheres zur Technik: s.o. Anm. 1) und das ausführliche Beispiel in Riemer et.al. (2000), S. 46. Die Klassenmitten sind für die Rechnung ohne Belang. Sie dienen nur zur Beschriftung des Diagramms. Daher auch in der 1. Abbildung die Schreibweise „<= 10“ statt „5“.

[4] Vorschläge, wie man die Explorative Datenanalyse intensiver aufbereiten kann, um Datensätze detaillierter zu untersuchen und u.U. auch einzelne Werte zu hinterfragen, finden sich bei Biehler et.al. (2003), in mathematik lehren (1997) und bei Vogel/ Wintermantel (2003).

[5] Will man länger bei den monovariaten Datensätzen bleiben, so wäre eine mögliche ergänzende Fragestellung: Sind die Daten des 2. Datensatzes von dem des ersten abhängig?

[6] Der weiter unten skizzierte Weg zur Korrelation thematisiert die Möglichkeit, die Parallelstrecken zur x-Achse als Abstände zu nehmen.

[7] Wenn die Schüler zu Beginn des Schuljahres Koordinatengeometrie gemacht haben, ist ihnen ein rechnerischer Weg i.d.R. bekannt, wenn auch nicht geläufig.

[8] Einen Ansatz, der schon sehr früh diesen Aspekt der Residuen mit aufnimmt, findet sich in Riemer et.al. (2000) Auf der unter 1) genannten Homepage findet sich unter dem Eintrag „Gott würfelt doch!“ auch eine Anregung, ein vorgegebenes lineares Modell, das „per Zufall“ verändert wurde, mithilfe der Regressionsgeraden zu rekonstruieren. Die Residuen entsprechen dann den „zufälligen“ Abweichungen vom eigentlich linearen Ausgangsmodell.

[9] s. dazu den Aufsatz von Borovcnik (1988).

[10] Zu dem hier beschriebenen Weg vgl. Griesel, Postel (1999), S.136 und 139.

[11] Der Weg, den Griesel/ Postel(1999) hier einschlagen, nämlich r schlicht aus dem Nichts heraus zu definieren und dann als harmonisches Mittel zu deuten (s. S. 139), ist wenig überzeugend.

Literatur

Biehler, Rolf; Kombrink, Klaus; Schweynoch Stefan (2003): MUFFINS: Statistik mit komplexen Datensätzen. *Stochastik in der Schule* 23(1), 11-25

Borovcnik, Manfred (1988): Methode der kleinsten Quadrate. *Stochastik in der Schule* 8(2), 17-24

Griesel, Heinz; Postel, Helmut (Hrsg.) (1999): *Elemente der Mathematik, 11. Schuljahr*, NRW. Hannover: Schroedel

Jahnke, Thomas; Wuttke, Hans et.al. (2000): *Mathematik 11. Schuljahr*. Berlin: Cornelsen

Landesinstitut für Schule und Weiterbildung (Hrsg.) (1994): *Trends und Zusammenhänge. Materialien zur Explorativen Datenanalyse und Statistik in der Schule.* Soest

mathematik lehren, Heft 97 (Dezember 1999): Daten und Modelle

Riemer, Wolfgang et.al (2000): *Lambacher Schweizer–LS 11* Nordrhein-Westfalen. Stuttgart: Klett

Vogel, Dankwart; Wintermantel, Gertraud (2003): *explorative datenanalyse – statistik aktiv lernen.* Stuttgart: Klett

Der Verfasser

Bernd Reckelkamm, Lehrer am Helmholtz-Gymnasium in Bielefeld, seit Februar 2002 teilabgeordnet an die Universität Bielefeld
`abcrc@t-online.de`