Statistik erleben: Regression und Korrelation im Schulalltag

Wolfgang Riemer

Vorbemerkung

Als ich im dritten Semester eine Vorlesung zur Einführung in die Wahrscheinlichkeitsrechnung und Statistik belegen wollte, wurde ich postwendend in der den Nachbarhörsaal zur Maßtheorie geschickt...
Einige Jahre später, als ich dann selber – der soliden Hochschulausbildung nacheifernd – mit „leuchtenden Augen“ einen Merksatz über Ereignisalgebra festhalten wollte, landete ein Papierflieger an der Tafel.: Auch ich musste erkennen: Stochastik kann auf der Schule (noch viel weniger als Analysis oder lineare Algebra) als verkleinertes Abbild einer systematisch aufgebauten Hochschul - Stochastik unterrichtet werden, wenn man einen nennenswerten Prozentsatz der Schüler erreichen möchte. Im Laufe der folgenden Jahre wurde mein Stochastik-Unterricht dann immer mehr geprägt durch die folgenden Paradigmen:

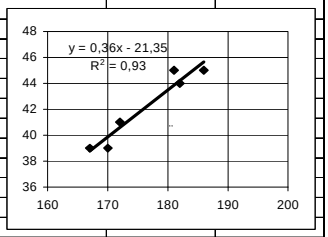
- Stochastikunterricht lebt vom Experimentieren.
  Wer nicht hin und wieder experimentieren möchte, weil „Experimentieren nicht zur Mathematik gehört“, lasse die Finger von der Stochastik.
- Die Experimente müssen echte Fragen beantworten.
- Die Fragen sollte man **vor** Durchführung der Experimente als (möglicherweise konkurrierende) Hypothesen formulieren. Dadurch
- entsteht eine emotionale Bindung an „das Problem“
- nähert man sich Grundideen der beurteilenden Statistik (mitunter wachsen sogar beschreibende und beurteilende Statistik zusammen)
- wird deutlich, was „Modellbildung“ meint.

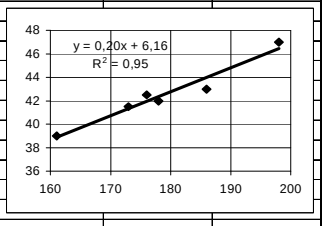
Am Beispiel der Korrelations- und Regressionsrechnung, die neuerdings in NRW in der Jahrgangsstufe 11 unterrichtet wird, soll punktuell angedeutet werden, wie sich diese Leitlinien auch hier umsetzen lassen. Dabei wird Bezug genommen auf das Buch [1] von dem ein Auszug mit freundlicher Genehmig des Klett-Verlages in diesem Manuskript abgedruckt wurde. Ein erprobter Handzettel für einen „Steilkurs“ mit diesem Lehrwerk mag die Darstellung abrunden.

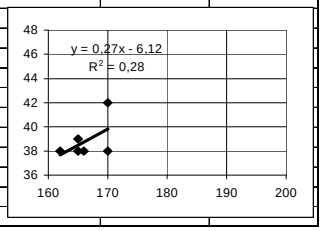
1 Einführung und grundlegende Begriffe - das „black-box-white-Box“ Prinzip

(Datei: GK11-Mittwoch-Länge-Schuhgröße.xls, Regression-Korrelation.xls, Lehrbuch: S. 60, S. 76, S. 58)
Ein Kurs wird in Sechsergruppen unterteilt. In jeder Gruppe messen die Teilnehmer Körpergröße X und Schuhgröße Y. Die Daten werden „online“ in ein vorbereitetes Kalkulationsblatt (Abb. 1) eingetragen und zusammen mit der Punktwolken-Grafik projiziert.

	A	B	C	D	E	F	G	H
1	GK11-Mittwo	Länge	Schuhgröße					
2	Gruppe 1	xi	yi	xi-xquer	yi-yquer	(xi-xquer)²	(yi-yquer)²	(xi-xquer)(yi-yquer)
3	Wolfgang	172	41	-4,3	-1,2	18,8	1,4	5,1
4	Mikey	181	45	4,7	2,8	21,8	8,0	13,2
5	Tobi	182	44	5,7	1,8	32,1	3,4	10,4
6	Jessica	167	39	-9,3	-3,2	87,1	10,0	29,6
7	Alice	170	39	-6,3	-3,2	40,1	10,0	20,1
8	Marc	186	45	9,7	2,8	93,4	8,0	27,4
9	Summe	1058	253			293,3	40,8	105,7
10	Mittelwert	176,3	42,2					
11	Varianz					48,9	6,8	
12	Standardabweichung					7,0	2,6	
13	Kovarianz							17,6
14	Gruppe 2	xi	yi	xi-xquer	yi-yquer	(xi-xquer)²	(yi-yquer)²	(xi-xquer)(yi-yquer)
15	Heino	173	41,5	-5,7	-1,0	32,1	1,0	5,7
16	Dani	198	47	19,3	4,5	373,8	20,3	87,0
17	Andi	186	43	7,3	0,5	53,8	0,3	3,7
18	Anastasia	161	39	-17,7	-3,5	312,1	12,3	61,8
19	Markus	178	42	-0,7	-0,5	0,4	0,3	0,3
20	Vitali	176	42,5	-2,7	0,0	7,1	0,0	0,0
21	Summe	1072	255			779,3	34,0	158,5
22	Mittelwert	178,7	42,5					
23	Varianz					129,9	5,7	
24	Standardabweichung					11,4	2,4	
25	Kovarianz							26,4
26	Gruppe 3	xi	yi	xi-xquer	yi-yquer	(xi-xquer)²	(yi-yquer)²	(xi-xquer)(yi-yquer)
27	Anne	170	42	3,7	3,2	13,4	10,0	11,6
28	Sandra	170	38	3,7	-0,8	13,4	0,7	-3,1
29	Dominika	162	38	-4,3	-0,8	18,8	0,7	3,6
30	Sybille	165	39	-1,3	0,2	1,8	0,0	-0,2
31	Maggi	166	38	-0,3	-0,8	0,1	0,7	0,3
32	Katharina	165	38	-1,3	-0,8	1,8	0,7	1,1
33	Summe	998	233			49,3	12,8	13,3
34	Mittelwert	166,3	38,8					
35	Varianz					8,2	2,1	
36	Standardabweichung					2,9	1,5	
37	Kovarianz							2,2







Anhand dieser Tabelle, die dank der eingetragenen Schülernamen mit höchstem Interesse studiert wird, werden die Begriffe Mittelwert, Varianz, Standardabweichung, Kovarianz und Regressionsgerade (Trendgerade) erarbeitet. Dabei werden die Formeln und die Zwischenrechnungen als „black box“ im Kalkulationsblatt vorgegeben. Die Aufgabe der Schüler ist es, diese Formeln und die numerischen Ergebnisse mit Inhalt zu füllen. Natürlich kann man auch Spalten im Tabellenblatt ausblenden. Wer das Blatt für eigene Unterrichtszwecke verwenden möchte, lösche die Messwerte und trage die eigenen ein. Wie Excel die Regressionsgerade („Trendgerade“) mit und das Bestimmtheitsmaßes r^2 berechnet, wird erst später thematisiert, wenn diese Größen ihre Nützlichkeit und Aussagekraft unter Beweis gestellt haben.

Man erkennt in Abb. 1:

- Je größer der Mittelwert $\bar{x}$ der Körpergröße X, desto weiter rechts liegt die Punktwolke
- Je größer der Mittelwert $\bar{y}$ der Schuhgröße Y, desto weiter oben liegt die Punktwolke

  So liegt die Punke der dritten Gruppe „unten links“. Die Teilnehmer sind verhältnismäßig „klein“ und sie tragen „kleine Schuhe“. Tatsächlich handelt es sich um eine Mädchengruppe.
- Je größer Varianz und Standardabweichung ist, desto mehr streuen die Daten in der zugehörigen Richtung. Bei Gruppe 2 streut die Körpergröße, bei Gruppe 1 streut die Schuhgröße besonders stark.
- Die Trendgerade beschreibt in allen Fällen den Zusammenhang: Je größer die Person, desto größer sind tendenziell auch die Schuhe. Je cm Körpergröße nimmt die Schuhgröße um „ca. 0,3“ zu.
- Die von Excel gezeichneten Trendgeraden verlaufen durch den Mittelpunkt ($\bar{x}$; $\bar{y}$) der Punktwolke
- Bei Sandra und Sybille ergeben die Produkte $(x_i - \bar{x})(y_i - \bar{y})$ negative Beiträge zur Kovarianz. Was das inhaltlich bedeutet, finden Schüler nach einer engagierten Diskussion selber heraus:
  Bei Sandra ist der zweite Faktor negativ, sie ist größer als ihr Gruppenmittel, trägt aber kleinere Schuhe als ihr Gruppenmittelwert „vorschreibt“
  Bei Sandra ist der erste Faktor negativ. Sie ist kleiner als ihr Gruppenmittel, trägt aber „für ihre Körpergröße zu große Schuhe“. Immer dann, wenn ein Datenpunkt links oben oder rechts unten vom Mittelpunkt der Punktwolke liegt, ist sein Beitrag zur Kovarinz negativ. Positive Kovarianz bedeutet also, dass die Punktwolke „steigende Tendenz“ besitzt
- Das von Excel zusammen mit der Formel der Trendgeraden angebotene Bestimmtheitsmaß R^2 liegt um so näher bei 1 je weniger die Punkte um die Trendgerade streuen.

Es macht keinen Sinn, die nützlichen und aussagekräftigen Begriffe Varianz, Standardabweichung, Kovarianz, und Regressionsgerade „problemorientiert herleiten“ zu wollen. In der Tat ist dies kaum möglich, da die tiefere Bedeutung dieser Begriffe sich erst im Rahmen einer Beschäftigung mit der Normalverteilung ergibt. Ehrlicher und spannender ist es, diese Formeln vorzugeben und deren Bedeutung inhaltlich anhand eigener authentischer Daten zu erschließen und mit Inhalt zu füllen, also auf diesem Wege erste Lichtstrahlen in die „black box“ hineinfließen zu lassen.

2 Exkurs

Daten sollte man in der Stochastik stets in Teilgruppen erheben! Man „erlebt“ dadurch die mit Stichproben verbundenen Zufallsschwankungen und die Bedeutung des Gesetzes der großen Zahl erschließt sich beim Zusammenfassen einzelner Stichproben. Wer würde daran zweifeln, dass die Trendgerade aus Abb. 2, der 57 Datenpaare zugrunde liegen, die Beziehung zwischen Körpergröße und Schuhgröße treffender beschreibt als die einzelnen Diagramme aus Abb. 1, der jeweils nur 6 Datenpaare zugrunde liegen? Man kommt zum Schluss, dass die Schuhgröße je cm Körpergröße im Schnitt um ca. 0,28 zunimmt.

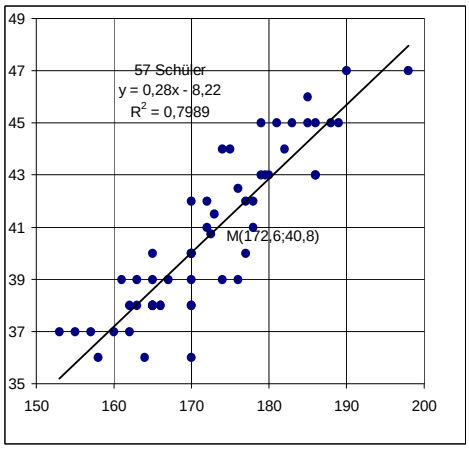


Abb. 2

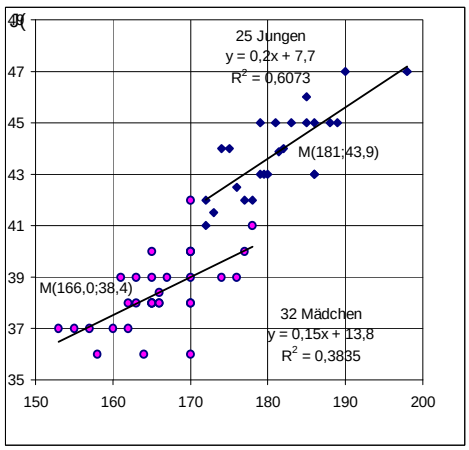


Abb. 3

Ein ganz anderes Bild ergibt sich, wenn man die Punktwolken für Mädchen und Jungen trennt. Die Steigungen der beiden Regressiensgeraden sind bei beiden Geschlechtern deutlich kleiner, durch Zusammenlegen der beiden Teilstichproben entsteht ein falsches Bild von der Wirklichkeit.
Man erkennt an diesem Beispiel, welche Vorsicht beim Schlussfolgern aus Daten walten sollte.

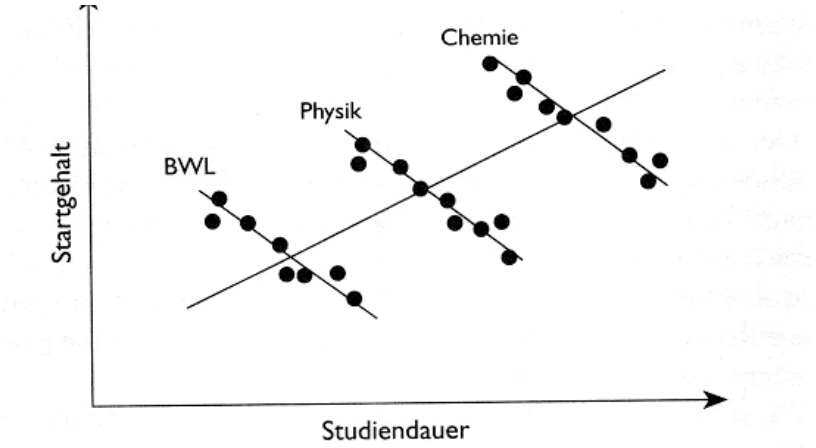


Abb. 4

Das gleiche Phänomen ist als Simpson Paradoxon bekannt, das Krämer in seinem Büchlein „Denkste!“ [2;S. 151] anhand von Abb. 4 wie folgt erläutert: Insgesamt (beim Zusammenfassen aller Studienfächer BWL Physik, Chemie) zeigt Abb. 4 mit steigender Studiendauer den Trend zu wachsendem Startgehlt beim Berufseinstieg, obwohl in jedem einzelnen Studienfach das Anfangsgehalt mit der benötigten Studiendauer sinkt. Chemiker studieren eben deutlich länger als Betriebswirtschaftler.

3 Lineare stochastische Modelle

(Datei Schuhe.xls, Lehrbuch S. 63)
Hinter der Regressionsrechnung steck die Annahme (Hypothese), dass zwischen zwei Zufallsgrößen X und Y ein linearer Zusammenhang der Form $Y = \alpha \cdot X + \beta + D$ besteht. Dabei ist D eine Störgröße mit einer bestimmten Störvarianz V_D und dem Erwartungswert μ_D=0. Ein Paradebeispiel dafür ist die Beziehung zwischen Schuhgröße X und Schuh (innen) Länge, die durch Produktionsvorschriften festgelegt ist. D beschreibt, ob die betrachteten Schuhe bei gleicher Schuhgröße groß oder klein ausfallen.
In einer spekulativen Phase werden verschiedene lineare Modelle aufgestellt. Als Steigungsfaktor vermuten Schüler Werte zwischen 0,5 und 1,5. Ob der y-Achsenabschnitt β den Wert 1 hat? Gelten möglicherweise für Damen und Herrenschuhe andere Zusammenhänge? Ob der Zusammenhang überhaupt linear ist? Das Experiment wird mit Spannung erwartet. Erfahrungsgemäß reicht die Datenbasis eines Kurses, um dem „wirklichen“ Steigungsfaktor α=2/3 und dem y-Achsenabschnitt $\beta = 0$ sehr nahe zu kommen. Verwendet man die Schuhaußenlänge, dann liegt der Wert von β in der Nähe von 1 (cm). Zum Umwandeln von Englischen/Amerikanischen auf die bei uns gebräuchlichen Französischen Maße nutze man die Tabelle auf der Marginalie von S. 64.

X (Größe)	Y (cm)
42	25,7
34	22
35	22,1
33	19,5
35	21,8
33	20,5
34	22
33	20,8
34	21,5
34	21
33	20,5
39	24,5
39	24,8
39	24,3
42	25,8
42	26
19	12,2
21	13
Mittelwerte	
34,5	21,6
Varianzen	
36,806	13,916
Kovarianz	22,478
Regression	0,611

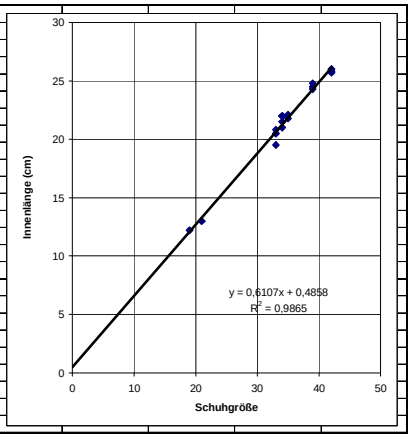


Abb. 5 Schuhe

Es wird deutlich, dass es bei der Regressionsrechnung darum geht, mit Hilfe der Trendgeraden mögliche lineare Zusammenhänge zu erkennen, die durch den Schleier empirischen Datenwolken in der Regel nur ungenau

erkennbar sind. Die „Verschleierung“ ist das Resultat individueller Zufallseinflüsse. Bekleidungsstücke „fallen manchmal etwas größer, manchmal etwas kleiner aus“. Auch Messungenauigkeiten hinterlassen ihre Spuren. Diese Einflüsse werden durch die Störgröße D modelliert.
Neben dem Schuhgrößen-Beispiel sind auch die Broka-Formeln „Normalgewicht = Körpergröße –100“ oder „Idealgewicht=0,9*Körpergröße-90“ sehr geeignet, um den Begriff des „linearen stochastischen Modells“ zu erarbeiten. Wie man die Broka-Formel auch für Kartoffeln, Erdbeeren oder andere „Naturprodukte“ aufstellen kann, wird aus S. 66, in Aufgabe 10 angedeutet.
a) Man wählt eine „repräsentative Standardkartoffel“. Die lineare Funktion ordnet der gemessenen Länge X das gemessene Gewicht Y zu.
b) Man ermittelt mit der Kreisformel aus dem Umfang die Querschnittsfläche in der Mitte der Kartoffel und bestimmt damit das Gewicht einer Kartoffel, die durch Einschieben einer 1cm dicken Scheibe in der Kartoffelmitte entsteht – und hat damit einen Wert für den Steigungsfaktor α.
Die Beschäftigung mit Kartoffeln o. ä. hat den Vorteil, dass es um weniger sensible Daten geht als bei dem eigenen Körpergewicht. Die aufgestellten Broka-Formeln sind durchaus geeignet, um die Zusammenhänge zwischen Länge und Gewicht bei Kartoffeln einer bestimmten Sorte mit Länge zwischen 3 und 10 cm treffend zu beschreiben. Doch ihr Wert besteht auch darin, die Grenzen linearer Modelle aufzuzeigen. Kartoffeln der Länge 0 müssten ein deutlich von 0 verschiedenes Gewicht besitzen. Wer für Kartoffeln Broca-Formeln aufgestellt und getestet hat, wird sich auch noch lange nach dem Abitur an den Mathematikunterricht erinnern.

4 Regressionskoeffizient

Wenn der Begriff der Kovarianz (vgl. 1 oder den unten stehenden Exkurs) zur Verfügung steht, kann man die Formel für den Regressionskoeffizienten inhaltlich verstehen, auch ohne Rückgriff auf die artifiziell anmutenden Minimalitätsargumente: Größere Kartoffeln (x- $\bar{x}$ >0) sind tendenziell schwerer (y- $\bar{y}$ >0). Wir nehmen dem linearen Modell entsprechend an, dass Gewichts- und Größenzunahme proportional sind, dass also im Mittel gilt $(y_i - \bar{y}) \approx \alpha \cdot (x_i - \bar{x})$. ($\alpha$ ist der unbekannte – zu schätzende – theoretische Regressionskoeffizient). Dann gilt

$$c_{XY} = \frac{1}{n}((x_1 - \bar{x}) \cdot (y_1 - \bar{y}) + (x_2 - \bar{x}) \cdot (y_2 - \bar{y}) + ... + (x_n - \bar{x}) \cdot (y_n - \bar{y}))$$
$$\approx \frac{1}{n}(\alpha \cdot (x_1 - \bar{x})^2 + \alpha \cdot (x_2 - \bar{x})^2 + ... + \alpha \cdot (x_n - \bar{x})^2)$$
$$= \alpha \cdot \frac{1}{n}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2) = \alpha \cdot V_X$$

damit ergibt sich $\alpha \approx \frac{c_{XY}}{V_X}$.

Der unbekannte theoretische Regressionskoeffizient α des linearen Modells lässt sich folglich durch den aus der Stichprobe ermittelten empirischen Regressionskoeffizienten $a = \frac{c_{XY}}{V_X}$ schätzen. Es gilt $\alpha \approx a$. Damit ist eine weiter Ecke der Excel - black-box weiß gemacht worden.

Exkurs: Wenn man Freude am Experimentieren hat und die Zeit für ein kleines Projekt reicht, bietet sich folgende Fragestellung an, um die Bedeutung der Kovarianz in Erinnerung zu rufen: (Datei Rollweiten.xls, Lehrbuch S. 58.). Jeder Schüler bekommt bei Start mit dem gleichen Fahrrad aus dem Stand eine halbe Pedalumdrehung Schwung und muss damit möglichst weit rollen. Auch Fünftklässler werden zum Experiment eingeladen. Rollen die schwereren Teilnehmer tendenziell weiter (weil sie dank ihres Körpergewichts mehr Schwung bekommen) oder rollen die leichteren weiter, weil sie wegen des kleineren Körpergewichts weniger Rollreibungs-Verluste besitzen?). Wieder hat man zwei konkurrierende Hypothesen, über die man trefflich streiten kann. Die experimentellen Daten (S. 58) besitzen positive Kovarianz. Das bedeutet: Schwere Leute rollen tendenziell weiter.
Hintergrund-Notiz: Wenn das Fahrrad kein Eigengewicht hätte, dann dürfte die Rollweite nicht von dem Gewicht des „Testpiloten“ abhängen. Sowohl die Energie als auch die Rollreibung sind nämlich dem Gewicht des Radler proportional – und die Einflüsse müssten sich exakt kompensieren. Das Eigengewicht des Radels sorgt aber dafür, dass Schwere Leute weiter rollen, weil man im Gegensatz zu leichteren das Gewicht des Radlers eher vernachlässigen kann. Eine genauere Analyse des Experimentes findet sich in [3] (und auf S. 163). Die Theorie zeigt, dass die Beziehung zwischen Rollweite und Körpergewicht in Wirklichkeit nicht linear ist..

5 Korrelation

(S. 70, Datei Minimalität).
Die Varianz V_Y kann man zerlegen in $V_Y = a^2 V_X + V_D$. Die entsprechenden Standardabweichungen lassen sich deuten als Kathetenlängen in einem rechtwinkligen Dreieck mit s_D als Katheten- und s_Y als Hypotenusenlänge.

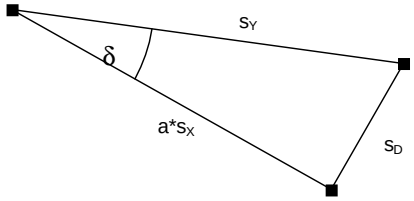


Abb. 6: „der Satz des Pythagoras in der Stochastik“

Je kleiner der Anteil der Störvarinz V_D an V_Y, desto „flacher“ wird das Dreieck. Den Sinus des zugehörigen Winkels bezeichnet man als Korrelationskoeffizienten. Es gilt $r = \cos(\delta)$. Man erhält die Formel $r = \frac{s_{X;Y}}{s_X \cdot s_Y}$.

Die Nützlichkeit des Korrelationskoeffizienten ergibt sich daraus, dass man mit seiner Hilfe die Standardabweichung s_D der nicht direkt beobachtbaren Störgröße D nach der Formel $s_D = \sqrt{1-r^2}s_Y$ berechnen kann. Damit ist auch die letzt Ecke der Excel – black box erleuchtet worden, „in der das Bestimmtheitsmaß r^2 als Quadrat des Korrelationskoeffizienten wohnt“.

Ein lohnender Abschluss der Unterrichtsreihe entsteht durch einen „Sprung“ in die beurteilende Statistik, wenn man nämlich die Standardabweichung der Störgröße benutzt, um mit der σ -Regel eine die Genauigkeit von Prognosen zu quantifizieren. Der interessierte Leser sei auf S. 72 und die Datei „Vorschulkinder.xls verwiesen.. Hier wird exemplarisch gezeigt, wie genau man aus der Körpergröße das Alter von Vorschulkindern vorhersagen kann.
Interessante Projekte [1; 162 f] zum Hören von Tongeschlechtern Dur / Moll und dem Rollwiderstand von Fahrrad-Dynamos runden die Thematik ab.

Literatur

[1] Lambacher-Schweizer 11 (NRW). Stuttgart 2001. Klett-Verlag No 73221.
[2] W. Krämer. Denkste! Frankfurt 1995, Campus Verlag.
[3] W. Riemer. „Wie weit rollt dein Fahrrad?“ MNU 51/7 S. 426 und 51/8 S. 475.

Handzettel: Steilkurs durch die Regressionsrechnung mit dem Buch Lambacher-Schweizer -11 und Excel
Lehrbücher sind linear aufgebaut. Im Unterricht dagegen kann man springen und – unter Ausnutzung des black-box-Verfahrens eigene Akzente setzen. Der folgende Unterrichtsgang hat sich bewährt:

Block A Grundlegende Begriffe

- Daten, z. B. „Körpergröße X – Schuhgröße Y“ in Sechsergruppen erheben, Punktdiagrame mit Trendgeraden und Bestimmtheitsmaßen zeichnen lassen und inhaltlich interpretieren.
- Bedeutung von Mittelwert Standardabweichung, Kovarianz im gewählten Kontext erleben.
- Nachrechnen: $M(\bar{x};\bar{y})$ liegt stets auf Excels Trendgeraden. Mittelwerte, Varianz, Standardabweichung, Kovarianz an einfachen Beispielen berechnen
  Material:
  Datei GK11-Mittwoch-Länge-Schuhgröße.xls
  Datei Regression-Korrelation.xls
  Übungen und Experimente aus LS, Abschnitte 1 bis 4 einstreuen

Block B Regression

- Begriff des linearen stochastischen Modells erarbeiten, Formel für den empirischen Regressionskoeffizienten a = $\frac{c_{XY}}{V_X}$ erarbeiten.
- Regressionsgeraden mit y=a(x- $\bar{x}$)+ $\bar{y}$ per Hand berechnen und zeichnen lassen
  Material
  Datei: Schuhe.xls
  LS Abschnitt 5

Block C Korrelation

- Störvarianz in Abhängigkeit von der Steigung a studieren
- Additivität $V_Y=a^2V_X+V_D$ entdecken und begründen.,
  Korrelation definieren als $r = \frac{erklärteVarianz}{beobachteteVarianz} = \frac{c_{X;Y}}{s_X \cdot s_Y}$
  Mit Darstellung an einem rechtwinklingen Dreieck.
  Datei Minimalität-der-Regressionsgeraden
  LS Abschnitte 6 und 7, insbes. Marginalie auf S. 71

Block D Experimente Wirklichkeit erkennen.

- Besonders hilfreich für das begriffliche Verständnis: Eine Versuchsperson würfelt Zufallszahlenpaare nach einem geheimgehaltenen linearen Modell, der Rest des Kurses versucht, das lineare Modell mit Hilfe von Excel zu rekonstruieren.
  Material:
  Datei Regression-Korrelation.xls
  Datei Regression-Korrelation-Simulation-Step
  LS11 Abschnitt 8

Eine sinnvolle minimale Lernsequenz ergibt sich schon aus Block A und D.