13 Regression

13 Regression
13.1 Was bedeutet Regression?
13.2 Die Regressionsgerade
13.2.1 Anwendungen für Regressionsgeraden
13.2.2 Die Anpassung einer Regressionsgeraden an eine Punktwolke
13.2.2.1 Anpassung nach Augenmaß
13.2.2.2 Anpassung durch Berechnung der Konstanten a und b
13.2.2.2.1 Berechnung des Regressionskoeffizienten b
13.2.2.2.2 Berechnung des Achsenabschnitts a
13.2.2.2.3 Berechnung von y_T-Werten über die Konstanten a und b
13.2.2.2.4 Berechnung der Lage der Geraden über die Konstanten a und b
13.2.2.2.5 Berechnung von y_T-Werten über Dreisatzrechnung
13.3 Graphische Ermittlung von a und b
13.3.1 Grapgische Ermittlung von b
13.3.2 Graphische Ermittlung von a
13.4 Das Bestimmtheitsmaß
13.5 Die Zerlegung der Gesamtvarianz
13.6 Übungen

13.1 Was bedeutet Regression?

Den Begriff Regression prägte der britische Naturforscher Sir Francis Galton im 19.Jh. im Zusammenhang mit empirischen Untersuchungen menschlicher Eigenschaften. Ihn beschäftigten u.a. Fragen zur Vererbung der Körpergröße, etwa ob die Nachkommen großer (kleiner) Väter so groß (klein) werden wie die Väter oder sogar größer (kleiner). Zur Prüfung solcher Überlegungen untersuchte er u.a. die Körpergröße von Vätern und erwachsenen Söhnen. Wir wollen mit Beispiel 1 zeigen, wie Galton zu dem Begriff Regression kam. Dazu verwenden wir keine Originaldaten von Galton, sondern fiktive Zahlen, die dem Problem angepasst sind.

Beispiel 1

Größe der Väter in cm	Größe der Söhne in cm
X	Y
150	153
155	158
160	162
165	167
170	171
175	173
180	179
185	182
190	184
195	187
200	191

Tabelle.1

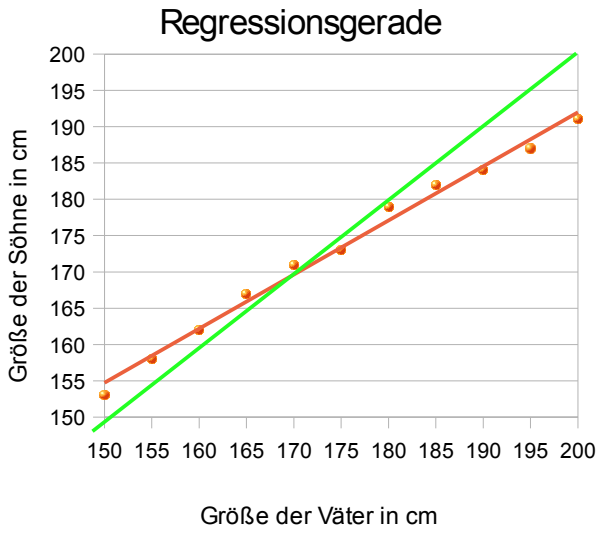


Abb.1

Abb.1 zeigt was Galton feststellte. Die grüne Linie entspricht der Annahme, dass die Söhne so groß werden wie die Väter. Der Sohn eines 155 cm großen Vaters wäre demnach 155 cm groß. Die den empirischen (hier fiktiven) Werten angepasste rote Linie zeigt aber die Tendenz, derzufolge die Söhne kleiner Väter etwas größer sind als die Väter und die Söhne großer Väter etwas kleiner als die Väter. Die Vererbung der Körpergröße führt also nicht zu Extremwerten. Vielmehr zeigt sich, dass die Größen der Söhne sich in Richtung zum Populationsmittelwert „zurück“entwickeln (**regressed**). Dabei ist dieser Rückschritteffekt um so stärker, je extremer die Größen der Väter sind. In unserem Beispiel gehen die Differenzen im mittleren

Bereich um 170 cm gegen Null. Die der Punktwolke (Punktfolge) angepasste rote Gerade nannte Galton **regression-line**. In diesem Sinne verstehen wir unter Regression eine Tendenz zur Normalisierung. Mit Hilfe der der linearen Regressionsgeraden entsprechenden Funktion f(x) = y = b * x + a (näheres siehe weiter unten) können wir rechnerisch von den Größen der Väter (X) Rückschlüsse ziehen auf die Größe der Söhne (Y). Oder allgemeiner: Die Regressionanalyse erlaubt Vorhersagen der Größe einer Variablen aus Kenntnis der Größe einer anderen Variablen.

13.2 Die Regressionsgerade

13.2.1 Anwendungen für Regressionsgeraden

Die Regressionsgerade ist das, was wir bei einer linearen Korrelation einfach Gerade genannt haben. Ausgangspunkt einer Regressionsanalyse sind die Daten, die wir in einem x/y-Diagramm als Punktwolke darstellen. Dieser Punktwolke müssen wir bei Tendenz zur Linearität eine Gerade anpassen. Bevor wir uns damit beschäftigen, wie wir diese Anpassung vornehmen, wollen wir kurz darauf eingehen, zu welchen Zwecken die Gerade bzw. die in ihr enthaltene Information angewendet werden kann. Wir nennen dazu drei Beispiele:

1. Erkennung des linearen Zusammenhangs zweier Variabler
Eine der Voraussetzungen zur Berechnung einer linearen Regressionsnalyse ist zumindest eine Tendenz zur Linearität der Punktwolke. Um dies visuell erkennen zu können, stellen wir die empirischen Daten in einem x/y-Diagramm dar. Können wir der Punktwolke eine Gerade anpassen, dann ist *diese* Voraussetzung für die Regressionsanalyse gegeben.

2. Kalibrationsgerade
Wenn z.B in der Photometrie ein linearer Zusammenhang zwischen Extinktion (E) und Konzentration (c) besteht, dann resultiert im x/y-Diagramm eine Gerade (Abb.2). Es ist ein in der Praxis übliches Verfahren, an dieser Geraden graphisch interpolierend Konzentrationen für gemessene Extinktionswerte zu ermitteln.

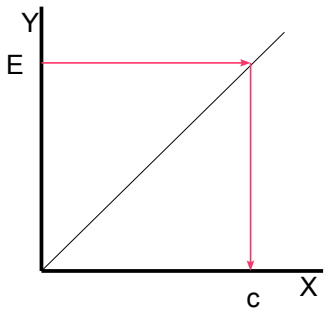


Abb.2

3. Graphische Ermittlung der Konstanten a und b
Grundlage der Regressionsanalyse ist die Geradengleichung y = b * x + a. An einer Geraden können wir graphisch die Funktionskonstanten a und b ablesen. Mit Hilfe dieser Kennzahlen der Regression lassen sich für vorgegebene x-Werte die gemäß der Funktion resultierenden y-Werte berechnen. Die Notation der Gleichung ist nicht verbindlich. Oft finden wir in der Literatur statt des a ein b und statt des b ein m. Die Gleichung lautet dann z.B. y = m * x + b. Vorsicht beim Literaturstudium!

13.2.2 Die Anpassung einer Geraden an eine Punktwolke

Beispiel 2
Bei 10 Mäusen wurden deren Masse und die Masse beider Nieren bestimmt. Tab.2 und Abb.3 zeigen die Daten. Wir interessieren uns für die folgenden Fragen:

1. Sind die Nierenmassen und die Körpermassen korreliert?

2. Um wie viel mg ändert sich die Nierenmasse, wenn die Körpermasse sich um 1 g ändert?
3. Wie stark beeinflusst die Körpermasse die Nierenmasse? Was mit „wie stark“ gemeint ist, sehen wir weiter unten.

Masse der Mäuse in g	Masse beider Nieren in mg
X	Y
34	524
34	576
27	458
29	483
30	475
33	471
25	347
26	331
25	332
23	322

Tabelle 2

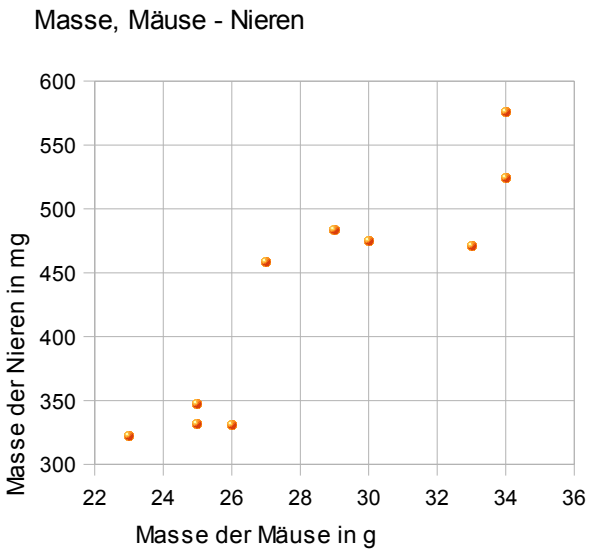


Abb.3

Der Punktwolke lässt den Schluss zu, dass eine lineare Korrelation vorliegt. Wir werden ihr daher eine Gerade anpassen. Das Ziel ist, die Gerade so zu legen, dass sie alle Punkte optimal repräsentiert. *Die* Gerade, die dieser Forderung gerecht wird, entspricht dann der Funktion y = b * x + a, der Modellvorstellung für den linearen Zusammenhang der beiden Variablen X und Y. Wir wollen hier drei Wege nennen, wie wir einer Punktwolke eine Gerade anpassen können.

1. Anpassung mit einem Tabellenkalkulationssystem (TKS)
Die ist sicher der übliche Weg, den jeder kennt, der mit einem TKS arbeitet. (Lit.: Monka, Schöneck, Voss; Statistik am PC, Hauser Verlag, 5. Auflage 2008). Da wir keine Bedienungsanleitung für den PC schreiben wollen, werden wir uns mit diesem Verfahren nicht beschäftigen. Vielmehr wollen wir zeigen, wie wir eine Gerade anpassen, „wenn der PC mal ausfällt“. So etwas soll ja schon mal vorgekommen sein.

2. Anpassung nach Augenmaß
Wir legen die Gerade nach Augenmaß so in die Punktwolke, dass wir den Eindruck haben, alle Punkte seien durch die Gerade gut repräsentiert.

3. Anpassung durch Berechnung
Wir berechnen aus empirischen Daten die Konstanten a und b der Funktion y = b * x + a und passen die Gerade nach Maßgabe der Berechnungsergebnisse der Punktwolke an.

Wir wollen den 2. und 3. Weg beschreiben.

13.2.2.1 Anpassung nach Augenmaß

Wie können wir einer Punktwolke eine Gerade nach Augenmaß optimal anpassen? Wenn mehrere Personen mit dem Lineal eine Gerade in die Punktwolke der Abb.3 legen, dann resultieren erfahrungsgemäß unterschiedliche Geraden. Und dann können wir fragen, nach welchem Kriterium die optimal liegende Gerade von den anderen, nicht so gut angepassten, zu unterscheiden ist. In Abb.4 haben wir der Punktwolke nach Augenmaß eine provisorische Gerade (blau) so angepasst, dass die empirisch ermittelten Punkte teils oberhalb und teils unterhalb der Geraden liegen. Wir messen die Abstände der Punkte von der Geraden. Diese Abstände nennen wir Residuen (Reste). Da die Werte der abhängigen Variablen Y parallel zur Y-Achse variieren, werden die Abstände parallel zur Y-Achse gemessen. Ein Kriterium für die Güte der Lage könnte sein, dass die Summe der Residuen oberhalb der Geraden (ΣO) gleich der Summe der Residuen unterhalb der Geraden (ΣU) ist so dass die Summe dieser Summen gleich 0 ist (ΣO + ΣU = 0).

Für die weiteren Überlegungen wollen wir folgendes vereinbaren. Wir bezeichnen in Abb.4 die empirischen y-Werte als y_i und die auf der Geraden liegenden y-Werte als y_T. Das T bedeutet „theoretischer“ Wert, es kennzeichnet *die* y-Werte, die nach der Modellvorstellung (y = b * x + a) zu erwarten sind, die also auf der Geraden liegen. Wenn wir in Abb.4 alle Residuen (y_i -y_T) messen und summieren, dann stellen wir bei gut liegender Geraden fest, dass $\Sigma O + \Sigma U \approx 0$ ist. Das ist verständlich, da die Differenzen oberhalb und unterhalb der Geraden unterschiedliche Vorzeichen (+ und -) haben. Ist $\Sigma O + \Sigma U \neq 0$, dann müssen wir die Lage der Geraden korrigieren und die Abstände neu vermessen. Das wird im Idealfalle solange iterativ fortgesetzt bis $\Sigma O + \Sigma U = 0$ ist. Aber: Das Ergebnis $\Sigma O + \Sigma U = 0$ erhalten wir auch dann, wenn wir die Gerade um den Schwerpunkt der Verteilung ($\bar{x}$ = 28,6; $\bar{y}$ = 432) drehen und z.B. flacher legen (rote Linie). Die Abstände werden dann größer, aber es gilt auch dann $\Sigma O + \Sigma U = 0$. Letztlich würde jede Gerade durch den Schwerpunkt zu diesem Ergebnis führen.

Da beim Vergleich der blauen mit der roten Geraden der subjektive Eindruck entsteht, die blaue läge besser, erscheint das Kriterium „Summe der Abweichungen“ nicht hinreichend um die beste Gerade sicher erkennen zu können. Wir müssen also ein anderes Verfahren anwenden um die Gerade optimal positionieren zu können. Trotz dieser Unsicherheit wird das Verfahren „Augenmaß“ in der Praxis bei orientierenden Untersuchungen oft angewendet, wenn nicht grundsätzlich mit dem PC gearbeitet wird.

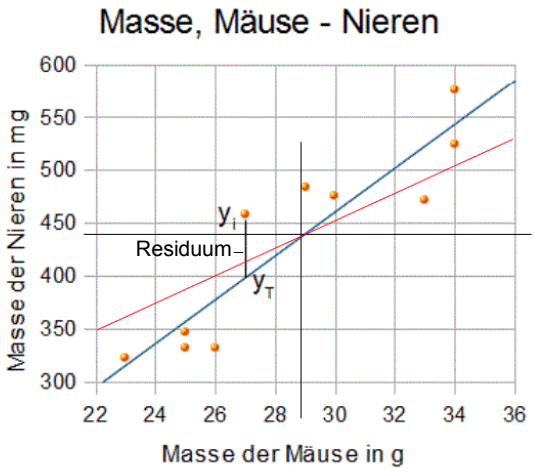


Abb. 4

13.2.2.2 Anpassung durch Berechnung der Konstanten a und b
Methode der kleinsten Quadrate

Das übliche Verfahren zur Berechnung der Lage der Geraden ist die von Gauß (1795) und Legendre (1805) entwickelte **Methode der kleinsten Quadrate**. Durch das Quadrieren der Abstände werden alle Quadrate positiv. Die Addition der Quadrate führt somit nicht zu 0. Ziel ist es, die Gerade so zu positionieren, dass die Summe der Abstandquadrate (Residuenquadrate) minimiert wird.

$\boxed{\Sigma(y_i - y_T)^2 \rightarrow \text{min!}}$ Gesprochen: Minimieren Sie die Summe der Abstandquadrate.

Die Gerade, für die die Bedingung $\Sigma(y_i - y_T)^2 \rightarrow$ min! erfüllt ist, repräsentiert die Punktwolke optimal. Mit der Aufforderung „$\Sigma(y_i - y_T)^2 \rightarrow$ min!“ kennen wir zwar das Kriterium für die optimale Lage, aber wir wissen noch nicht, wie wir sie konkret legen damit die Bedingung $\Sigma(y_i - y_T)^2 \rightarrow$ min! erfüllt ist. Wir könnten die Lösung wieder durch Iteration finden indem wir die Gerade nach Augenmaß legen, die Bedingung $\Sigma(y_i - y_T)^2 \rightarrow$ min! prüfen, gegebenenfalls die Lage korrigieren, neu messen usw. bis die Bedingung erfüllt ist. Bei diesem graphischen Verfahren führt Unsicherheit bei der Abstandmessung wieder zu unpräzisen Werten.
Die Methode der kleinsten Quadrate ist aber ein rein rechnerisches Verfahren, welches darauf hinaus läuft, dass die Lage der Geraden direkt, ohne eine zuvor erstellte Graphik, berechnet wird. Die Lage einer Geraden im Koordinatensystem ist durch die beiden Konstanten a und b der Gleichung y = b * x + a eindeutig bestimmt. a ist der Achsenabschnitt und b die Steigung der Geraden, der Regressionskoeffizient. Letzteren wollen wir zunächst berechnen.

13.2.2.2.1 Berechnung des Regressionskoeffizienten b

Mit b berechnen wir, wie steil die Gerade im Koordinatensystem liegt. (Abb.5)

Der Regressionskoeffizient b ist, da aus empirischen Daten ermittelt, ein Schätzer für den Regressions-koeffizienten ß der Grundgesamtheit. Die Voraussetzungen für die Berechnung entsprechen denen der Berechnung des Korrelationskoeffizienten r (Normalverteilung beider Variablen, Linearität, Messwerte). Wir gehen davon aus, dass diese Voraussetzungen hier vorliegen.

Wir bleiben bei den Massen der Mäuse und deren Nieren in Beispiel 2. Wenn, wie hier, X der Regressor ist, dann indizieren wir das zu berechnende b als b_x. Liegt keine Gefahr der Verwechslung (mit b_y, siehe unten) vor, dann können wir auf den Index verzichten.

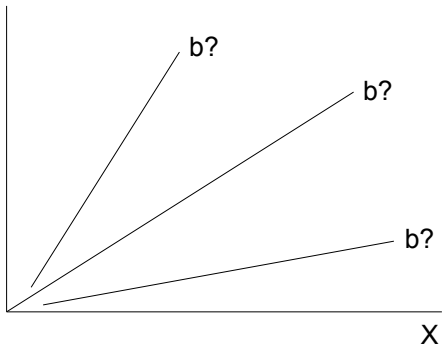


Abb. 5

Neben den unten stehenden Gleichungen finden wir in der Literatur auch andere äquivalente Formulierungen.

Regressionskoeffizient b_x

$$b_x = \frac{\text{Kovarianz}}{\text{Varianz X}} = \frac{s_{xy}}{s_x^2} \longrightarrow b = \frac{1/n\text{-}1 * \sum(x_i-\bar{x})*(y_i-\bar{y})}{1/n\text{-}1 * \sum(x_i-\bar{x})^2}$$

Wir stellen zunächst die Arbeitstabelle (Tab.3) mit den für die Berechnung notwendigen Termen zusammen: $\bar{x}$; $\bar{y}$; $\sum(x_i-\bar{x})*(y_i-\bar{y})$; $\sum(x_i-\bar{x})^2$; $\sum(y_i-\bar{y})^2$; n = Anzahl der Wertepaare.

x_i	y_i	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})*(y_i-\bar{y})$
34	524	29,16	8482,41	497,34
34	576	29,16	20764,81	778,14
27	458	2,56	681,21	-41,76
29	483	0,16	2611,21	20,44
30	475	1,96	1857,61	60,34
33	471	19,36	1528,81	172,04
25	347	12,96	7208,01	305,64
26	331	6,76	10180,81	262,34
25	332	12,96	9980,01	359,64
23	322	31,36	12078,01	615,44
$\bar{x}$	**$\bar{y}$**	**$\Sigma(x_i-\bar{x})^2$**	**$\Sigma(y_i-\bar{y})^2$**	**$\Sigma(x_i-\bar{x})*(y_i-\bar{y})$**
28,6	**431,9**	**146,4**	**75372,9**	**3029,6**
Σx	**Σy**		**$1/n\text{-}1\Sigma(y_i-\bar{y})^2$**	
286	**4319**		**8374,76**	

Tabelle 3

Nun ist

$$b_x = \frac{\text{Kovarianz}}{\text{Varianz X}} = \frac{1/n\text{-}1 * \sum(x_i - \bar{x}) * (y_i - \bar{y})}{1/n\text{-}1 * \sum(x_i - \bar{x})^2} = \frac{1/9 * 3029{,}6}{1/9 * 146{,}4} = \frac{336{,}62}{16{,}266} = 20{,}69$$

$\underline{b_x = 20{,}69}$

Es gibt einen zweiten Regressionskoeffizienten b_y, der dann berechnet wird, wenn wir Y als Regressor betrachten.

Zur Berechnung von b_y gilt

$$b_y = \frac{\text{Kovarianz}}{\text{Varianz Y}} = \frac{1/n\text{-}1 * \sum(x_i - \bar{x}) * (y_i - \bar{y})}{1/n\text{-}1 * \sum(y_i - \bar{y})^2} = \frac{1/9 * 3029{,}6}{1/9 * 75372{,}9} = \frac{336{,}62}{8374{,}76} = 0{,}0402$$

$\underline{b_y = 0{,}0402}$ runden zu 0,04

Andere Notationen für b: $b_x = b_{y/x}$, gesprochen by auf x
$b_y = b_{x/y}$, gesprochen bx auf y

Der Zusammenhang der beiden Koeffizienten b_x und b_y, wenn r = 1

Der Wert b_x = 20,69 mg/g könnte zu dazu führen, b_y aus b_x berechnen zu wollen oder umgekehrt. Und zwar nach folgender Überlegung:

b_x = 20,69 mg/g ⟶ 20,69 mg Nieren entsprechen 1 g Maus
1 mg Nieren entsprechen x g Maus
x = 1/20,69 = 0,048 (gerundet)

Dieses Ergebnis 0,048 stimmt aber nicht mit dem oben berechneten Werte b_y = 0,040 überein. Die Differenz ist nicht auf eine Messunsicherheit zurückzuführen, vielmehr ist diese Dreisatzrechnung nur dann zulässig, wenn der Korrelationskoeffizient r = |1| ist. In unserem Falle - Sie können ja mal nachrechnen – ist r = +0,91. Und damit ist diese Berechnung hier nicht erlaubt.

13.2.2.2.2 Berechnung des Achsenabschnitts a (Schätzer für α der Grundgesamtheit)

Mit dem Koeffizienten b = 20,69 kennen wir die Steigung der Geraden. Das bedeutet hier, dass pro X-Einheit, Y um 20,69 Einheiten steigt. Damit wissen wir, wie schräg die Gerade im Koordinatensystem liegt. Wir müssen aber noch berechnen, wie hoch sie liegt, wo die Gerade – wenn beide Achsen im Ursprung (0) beginnen - die Y-Achse schneidet.

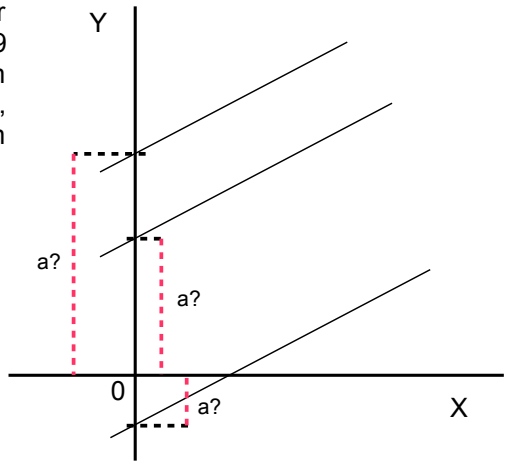


Abb. 6

Zur Berechnung der Lage der Geraden nutzen wir die Funktionsgleichung y = b * x + a, in der wir y durch y_T und x durch x_i ersetzen.

$$y_T = b * x_i + a$$

Es bedeuten

x_i	ein beliebiger x-Wert aus der Spanne der empirischen x-Werte
y_T	der für einen x_i-Wert berechnete theoretische y-Wert, der auf der Geraden erwartet werden kann.
b	der Regressionskoeffizient, die Steigung der Geraden
a	der Achsenabschnitt, der Schnittpunkt der Geraden mit der Y-Achse, wenn beide Achsen im Ursprung (0) beginnen.

Durch Umstellen der Gleichung nach a erhalten wir

$$a = y_T - b * x_i.$$

Wir ersetzen nun y_T durch $\bar{y}$ und x_i durch $\bar{x}$ und erhalten die Gleichung zur Berechnung von a.

Achsenabschnitt a $a = \bar{y} - b * \bar{x}$

Hier könnte die Frage gestellt werden, warum y_T durch $\bar{y}$ ersetzt wird und nicht durch $\bar{y}_T$. Die Antwort: $\bar{y} = \bar{y}_T$. Aus Tabelle 3 entnehmen wir $\bar{x}$ = 28,6 und $\bar{y}$ = 431,9.

Dann ist

a = 431,9 – 20,69 * 28,6
a = - 159,83

Dieser Wert zeigt an, wie hoch die Gerade im Netz liegt. Er hat nur rechentechnische Bedeutung und macht in diesem Falle praktisch keinen Sinn, da ein negatives Nierengewicht (- 159,83 mg) nicht auftreten kann. Damit wir a in einer Graphik abgelesen werden können, müssen wie in Abb.7 beide Achsen mit 0 beginnen. Gegebenenfalls müssen beide Skalen so verlängert werden, dass Werte im 2., 3. und/oder 4 Quadranten angegeben werden können. In Abb.7 resultiert graphisch ein Wert von a ≅ -160. Üblicherweise wird die Darstellung entsprechend der Abb.4 gewählt, da bei deren Skalierung die Punkte genauer eingetragen werden können und graphisch genauer interpoliert werden kann als in Abb.7.

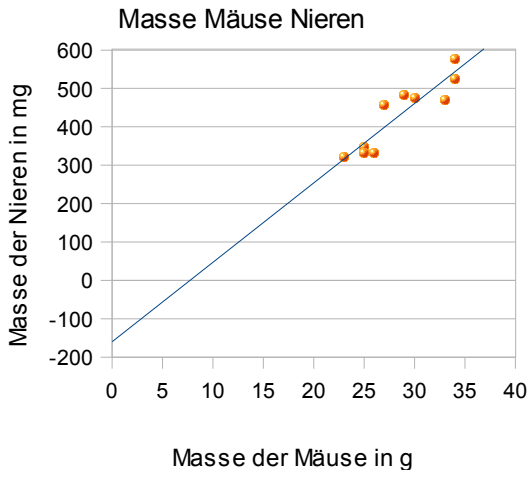


Abb.7

13.2.2.2.3 Berechnung von y_T-Werten über die Konstanten a und b

Nachdem die Kennwerte a und b bekannt sind, können wir die zu jedem x_i-Wert (aus der experimentell ermittelten Spanne) den nach der Modellvorstellung zu erwartenden y_T-Wert berechnen.

Für x_i = 31, b = 20,96 und a = -159,83 gilt

$$y_T = b * x_i + a$$
$$y_T = 20,69 * 31 + -159,83$$
$$y_T = 481,56$$

Für die Nieren einer 31 g schweren Maus sind demnach theoretisch 482 mg zu erwarten.

13.2.2.2.4 Berechnung der Lage der Geraden über die Konstanten a und b

Nach den gleichen Überlegungen wie oben können wir nun die optimale Lage der Geraden ermitteln. Dazu berechnen wir für zwei beliebige x_i-Werte, die im Bereich der empirisch gewonnenen Grenzen liegen, die y_T-Werte. Die beiden resultierenden Punkte verbinden wir linear und haben damit die berechnete, optimale Gerade. Wir wählen z.B. x_i = 23 und x_i = 34.

Dann erhalten wir nach $y_T = b * x_i + a$
für x_i = 23 $y_T = 20,69 * 23 + -159,83 = 316,04$
für x_i = 34 $y_T = 20,69 * 34 + -159,83 = 543,63$

Aus den Punkten x_{23};y_{316} und x_{34};y_{544} resultiert die Gerade in Abb.8 in der die beiden berechneten Endpunkte der Geraden gekennzeichnet sind.

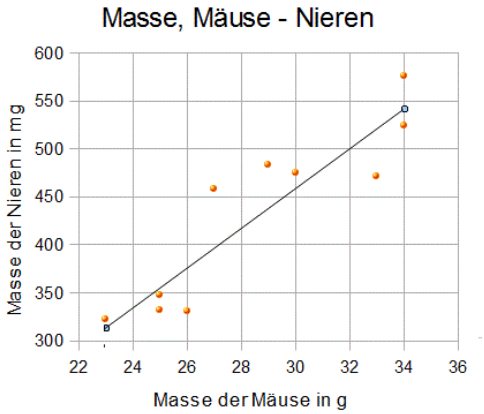


Abb.8

Bemerkungen

1. Wir hätten z.B. auch die beiden näher beieinanderliegenden x_i-Werte 28 und 30 wählen können. Die beiden resultierenden Punkte lägen dann auch näher beieinander. Es erscheint unmittelbar einsichtig, dass eine kleine Ungenauigkeit beim Verbinden der beiden Punkte (in Abb.9 übertrieben dargestellt) einen um so größeren Fehler bei der Steigung zur Folge hat, je näher die beiden Punkte beieinander liegen. Die Gerade a weicht bei gleicher Ungenauigkeit beim Verbinden der Punkte deutlich stärke von der richtigen (horizontalen) Lage der Geraden (rot) ab, als die Gerade b. Daher ist die Wahl zweier weit entfernter x-Werte zu empfehlen.

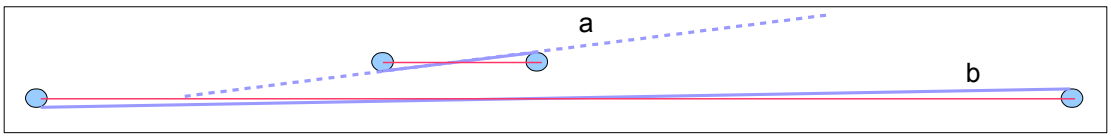


Abb.9

2. Gegen Rechenfehler ist niemand geschützt. Wird einer der beiden y_T-Werte falsch berechnet, so resultiert trotzdem eine Gerade zwischen den beiden Punkten. Dass sie falsch liegt, wird u.U. nicht unmittelbar auffallen. Um sich gegen diese Fehlermöglichkeit abzusichern, ist es sinnvoll, für einen dritten x-Wert y_T zu berechnen. Sind alle drei y_T-Werte richtig berechnet, so müssen die drei Punkte auf einer Geraden liegen. Ist ein Wert falsch berechnet, so liegen nicht alle Punkte auf einer Geraden und das fällt sofort als unplausibel auf. Welches der falsche Wert ist, das muss durch Nachrechnen geprüft werden.

13.2.2.2.5 Berechnung von y_T-Werten über Dreisatzrechnung.

Auch wenn dieses Verfahren weniger „elegant“ ist als die Berechnung über die Funktionsgleichung, wollen wir es trotzdem erklären. Der berechnete Wert b = 20,69 bedeutet, dass 20,69 Y-Einheiten 1 X-Einheit entsprechen. Auf unsere Daten angewendet heißt das,

die Zunahme der Masse bei Mäusen um 1g
entspricht
der Zunahme der Masse der Nieren um 20,69 mg

Wir entnehmen graphisch interpolierend Abb.8, dass die Nieren einer 25 g schweren Maus 360 mg wiegen. Wenn wir nun berechnen wollen, wie schwer die Nieren einer 31 g schweren Maus sind, dann gilt folgende Überlegung:

Bei 1 g Maus-Zuwachs nehmen die Nieren um 20,69 mg zu
Bei 6 g Maus-Zuwachs (31g – 25 g) nehmen die Nieren um 6 * 20,69 mg = 124,14 mg zu.

360 mg + 124 mg = 484 mg (gerundet)
Die Nieren einer 31 g schweren Maus wiegen also nach unserer Modellvorstellung 484 mg.

Dieser Wert stimmt mit der Graphik in Abb.8 in etwa überein. Natürlich könnten wir den Wert direkt an der Graphik interpolieren, wir wollten hier aber die Rechnung zeigen.

Der Koeffizient b_y = 0,04 besagt, 0,04 X-Einheiten ≙ 1 Y-Einheit. Oder: Wenn die Nierenmasse um 1 mg steigt, dann wird die entsprechende Maus um 0,04 g schwerer sein. Wir könnten also, wenn wir nach der Sektion die Nierenmasse einer Maus mit 412 mg bestimmen, berechnen, wie schwer die Maus war:
Wir entnehmen der Geraden in Abb.8: 400 mg Nierenmasse (Y) entsprechen 27,0 g Maus (X).

1 mg Nierenmasse entsprechen 0,04 g Maus
12 mg Nierenmasse entsprechen 0,48 g Maus

Die Maus mit einer Nierenmasse von 412 mg hat also eine Masse von 27,48 g → 27,5 g

Bemerkungen
1. Beachten Sie die Unsicherheit bei graphisch ermittelten Werten.
2. Bedenken Sie, dass die Berechnung von y-Werte aus X-Werten und umgekehrt nur Schätzwerte ergeben, wenn der Korrelationskoeffizient <|1| ist.
3. Es darf nur im Bereich der Werte, die durch das Experiment abgedeckt sind, interpoliert werden. Da wir nicht wissen, ob die Regression unterhalb und oberhalb der Grenzwerte von X linear verläuft, dürfen wir nicht extrapolieren.

13.3 Graphische Ermittlung von b und a

Nehmen wir an, wir müssten mit den Daten von Beispiel 2 für eine Maus von 31,4 g die zu erwartende Nierenmasse nach y = b * x + a berechnen. Und setzen wir voraus, dass nur eine nach Augenmaß gelegte Gerade vorhanden ist. Da zu deren Konstruktion a und b nicht erforderlich waren, sind uns die beiden Konstanten nicht bekannt. Wir benötigen sie aber für die geforderte Berechnung. (Natürlich könnten wir den gesuchten Wert an der Geraden graphisch interpolieren. Wir wollen hier aber die graphische Ermittlung der Konstanten zeigen.) a und b können wir an der vorhandenen Geraden graphisch ablesen. Das geht relativ einfach nach folgenden Definitionen für a und b. Siehe Abb. 10.

b = Differenzquotient $\Delta y/\Delta x = (y_2-y_1)/(x_2-x_1)$

b = tan α = Gegenkathete/Ankathete (siehe hierzu „Umrechnung von tan in den Differenzquotienten“ weiter unten.

a = Schnittpunkt der Geraden mit der Y-Achse, wenn beide Achsen im Ursprung (0) beginnen.

13.3.1 Graphische Ermittlung von b

Die im Folgenden benutzte Bezeichnung LE bedeutet Längeneinheit. Wir verwenden sie statt mm oder cm, weil bei der Darstellungen von Abbildungen am Bildschirm oder im Ausdruck je nach Zoom-Einstellung mm/cm-Werte u.U. nicht mehr stimmen würden.

1. Der Differenzquotient b = $\Delta y/\Delta x$

Abb.10 zeigt die nach Augenmaß gelegte Gerade. Wir legen zwei möglichst weit voneinander entfernte x_i-Werte fest, z.B. $x_2 = 32$ und $x_1 = 24$. Über die Lote und deren Schnittpunkte mit der Geraden ermitteln wir die beiden korrespondierenden y_i-Werte $y_2 = 495$ und $y_1 = 340$. Die Ablesung der y-Werte ist bei der gegebenen Skalierung nur als Schätzung möglich.

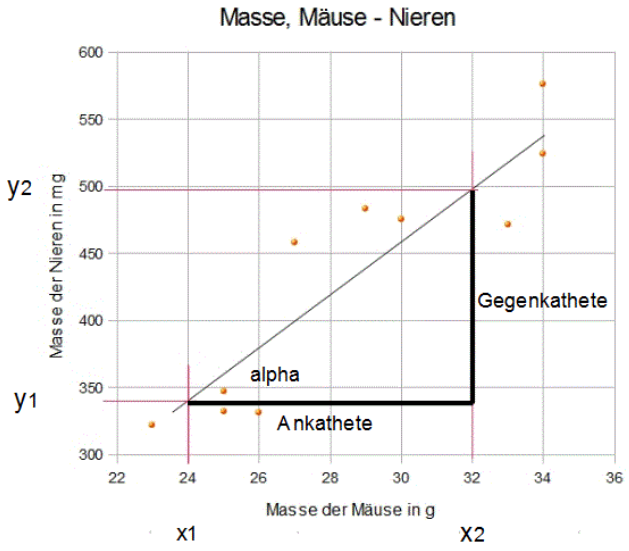


Abb.10

Nach Einsetzen der Werte in die Gleichung erhalten wir

$b = (y_2-y_1)/(x_2-x_1)$
$b = (495 – 340)/(32-24) = 155/8$
$\underline{b = 19{,}375 \rightarrow 19{,}4}$

Das bedeutet 19,4 Y-Einheiten/1 X-Einheit, also 19,4 mg Nieren pro 1 g Maus.

2. Gegenkathete / Ankathete

In Abb.10 bezeichnen wir die Kantenlänge eines Gitterquadrates als 1 LE. Wenn wir die Katheten in LE messen, dann gilt für die Gegenkathete: 3,2 LE und für die Ankathete 4 LE. Die Werte sind Schätzungen, da sie nicht exakt abgelesen werden können. Der folgende Wert für b ist damit auch eine Schätzung.

$b = GK/AK = 3{,}2/4 = 0{,}8$

Dieser Wert stimmt nicht mit dem oben errechneten b = 19,4 überein, siehe dazu weiter unten.

3. Über b = tan α

Die Definition der trigonometrischen Funktion tangens ist gleich dem Quotienten GK/AK. Durch Messung erhalten wir in Abb.10 $\alpha = 40{,}5°$ und über eine trigonometrische Tabelle oder über den Taschenrechner **tan 40,5° = 0,854**. Auch dieser Wert entspricht nicht b = 19,4. Er ist wegen der Unsicherheit bei der Winkelmessung ebenfalls eine Schätzung. Wir mitteln 0,8 und 0,854 zu 0,827. (Ob drei Nachkommastellen gerechtfertigt sind, soll hier mal keine Rolle spielen.)

Umrechnung von tan α (0,827) in den Differenzquotienten (19,4)

Der über 2. und 3. erhaltene tan-Wert 0,827 stimmt nicht überein mit dem Differenzquotienten b = 19,4. Die Werte wären dann gleich, wenn auf beiden Achsen 1 X-Einheit und 1 Y-Einheit metrisch gleich langen Strecken entsprächen. Der Grund für die Ungleichheit (19,4 und 0,827) liegt in der unterschiedlichen

Skalierung von Ordinate und Abszisse. Wir wollen zeigen, wie wir rechnerisch von dem tan-Wert 0,872 zum Differenzquotienten b = 19,4 kommen. Beachten Sie im folgenden Text den Unterschied zwischen Längen-Einheiten (LE) und X-Einheiten (Gramm) bzw. Y-Einheiten (Milligramm).

a Wir messen eine (wegen der Messunsicherheit) möglichst große Strecke der Abszisse in LE aus. Die Strecke 22 g bis 32 g entspricht 10 X-Einheiten. In Abb.10 ist sie 5 LE lang.

b Die gleiche Strecke (5 LE) tragen wir auf der Ordinate beginnend z.B. bei 300 mg nach oben ab. Sie reicht bis 550 mg, das sind 250 Y-Einheiten.

Danach gilt

5 Längeneinheiten		5 Längeneinheiten
↓		↓
10 X-Einheiten	entsprechen	250 Y-Einheiten
1 X-Einheit	entspricht	25 Y-Einheiten

c Die Anzahl Y-Einheiten, die *einer* X-Einheit entspricht, nennen wir **Skalenfaktor**. Diese hat hier den Wert 25. Nun gilt

$$b = \tan \alpha \ * \textbf{Skalenfaktor}$$

$$b = 0{,}827 * 25 = 20{,}7$$

Wenn beide Skalen im oben genannten Sinne die gleiche Einteilung hätten, dann wäre der Skalenfaktor gleich 1. Und dann wäre $b = \tan \alpha$.

Das Ergebnis 20,7 ist nicht gleich 19,4. Bedenken Sie angesichts dieser Abweichung, dass die Werte, mit denen wir gerechnet haben, graphisch ermittelt wurden. Welchem Wert 19,4 oder 20,7 für b wir den Vorzug geben, könnte davon abhängen, bei welchem Ermittlungsverfahren der geringere Ablesefehler bei der Graphik zu erwarten ist. Letztlich könnten wir auch mitteln. b = 20,7 entspricht recht gut dem berechneten Wert b = 20,69.

13.3.2 Graphische Ermittlung von a

In Abb.7 lesen wir ab: $a \approx -160$.

Wir haben gesehen, dass graphisch ermittelte Werte durch Messunsicherheit belastet sind. Die graphische Ermittlung von a und b an einer nach Augenmaß gelegten Geraden liefert relativ unsichere Werte. Es ist daher sinnvoller, der beiden Konstanten nach $y = b * x + a$ zu berechnen.

13.4 Das Bestimmtheitsmaß

In Beispiel 2 variieren die Mäusemassen (X) und die Nierenmassen (Y) um ihre Mittelwerte. Wenn wir überlegen, welches die Ursache für die Variation der Nierenmassen (abhängige Variable) ist, dann können wir uns verschiedenen Faktoren vorstellen, die einen Einfluss auf die empirisch ermittelten Massen der Nieren haben könnten, z.B.

Mögliche Faktoren (steuernde Variable)		abhängige Variable
Mäusemasse Futterqualität Haltungsbedingungen Gesundheitszustand persönliche und systematische Messfehler	}	können u.U. s_y^2 der Nierenmasse beeinflussen.

Möglicherweise haben diese Faktoren unterschiedlich starke Einflüsse auf s_y^2. Experimentell wurde nur die

Masse der Mäuse in die Untersuchung einbezogen. Daher können wir rechnerisch auch nur den Einfluß der Mäusemassen auf die Nierenmassen untersuchenn. Die Frage lautet: Für wie viel % der Varianz s_y^2 ist die Varianz der Mäusemasse die Ursache? Wie viel % von s_y^2 kann durch s_x^2 erklärt werden?

Eine Antwort auf diese Frage erhalten wir über das sogenannte Bestimmtheitsmaß B (Determinationsmaß). Es entspricht einerseits dem Produkt der beiden Regressionskoeffizienten b_x und b_y. Andererseits ist dieses Produkt identisch mit dem Quadrat des Korrelationskoeffizienten r. Es gilt also

Bestimmtheitsmaß $B = b_x * b_y = r^2$

Für Beispiel 2 gilt: $B = 20{,}69 * 0{,}040 = r^2 = 0{,}8268$

$r = \sqrt{B} = \sqrt{(b_x * b_y)}$
$r = \sqrt{(20{,}69 * 0{,}040)}$
$r = 0{,}909$

Damit ist der Korrelationskoeffizient r das geometrische Mittel ($\bar{x}_G = \sqrt{\Pi x_i}$) der beiden Regressionskoeffizienten. Der Wert r^2 beantwortet nun die Frage nach dem numerischen Einfluss der Varianz von X auf die Varianz von Y. $r^2 = 0{,}83$ (gerundet) bedeutet nach Multiplikation mit 100, dass 83 % von s_y^2 durch s_x^2 erklärt werden. Wir bezeichnen diese 83 % als sogenannte „erklärte Varianz". **Vorsicht!** Das Wort „erklärt" sagt nichts über einen kausalen Zusammenhang zwischen X und Y.

Es gilt also **Erklärte Varianz** $= s_y^2 * r^2$

Nach Tabelle 3: Erklärte Varianz $= 8374{,}76 * 0{,}83 = 6951$

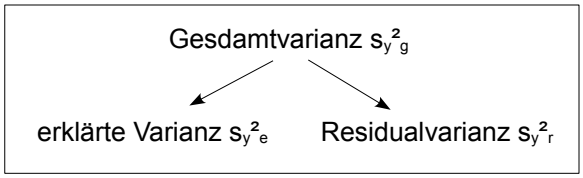
6951 sind 83 % von 8374,76.

Je größer die erklärte Varianz ist (im Idealfall = s_y^2) um so stärker ist die Korrelation. Denn bei $r = 1 \rightarrow r^2 = 1 \rightarrow$ erklärte Varianz $= s_y^2 * 1 = s_y^2$.

Die Differenz der erklärten Varianz zur Gesamtvarianz, also die restlichen 17 %, nennen wir Restvarianz oder Residualvarianz. Die Ursache für diese Varianz kann aus den vorliegenden Untersuchungsergebnissen nicht erklärt werden.

13.5 Die Zerlegung der Gesamtvarianz in erklärte Varianz und Residualvarianz.

Wir haben über das Bestimmtheitsmaß die Gesamtvarianz in zwei Teilbereiche aufgeteilt, in die erklärte Varianz und die Residualvarianz.



An den fiktiven Daten von Beispiel 3 wollen wir noch einmal auf die Zerlegung der Gesamtvarianz eingehen. Der Datensatz ist mit n = 3 an sich für eine Regressionsanalyse zu klein. Als Grundlage zur Erklärung des Rechenweges sollen die Daten aber reichen.

Beispiel 3
Daten

X	3	6	9
Y	2	8	8

Tabelle 4

Abb.11 zeigt die drei Wertepaare mit der berechneten Geraden. In Abb.12 haben wir den Bereich um Punkt $x_3;y_2$ vergrößert dargestellt um die Zerlegung der Gesamtvarianz graphisch zu visualisieren.

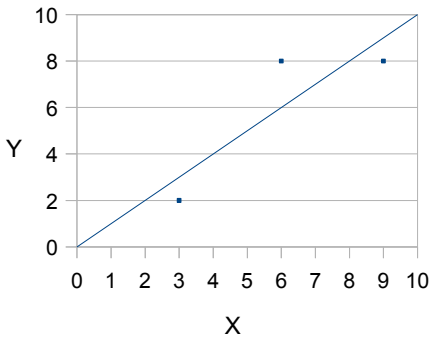
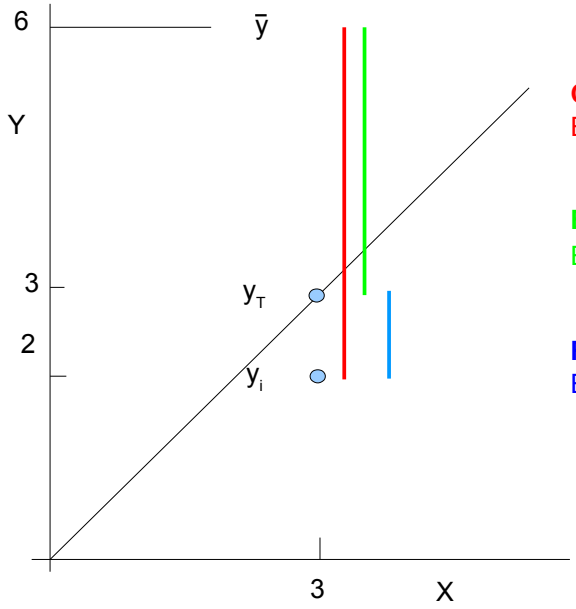


Abb.11



Gesamtvarianz $s_y^2{}_g = 1/n\text{-}1 * \sum(y_i - \bar{y})^2$
Entspricht der Differenz $y_i - \bar{y}$

Erklärte Varianz $s_y^2{}_e = 1/n\text{-}1 * \sum(y_T - \bar{y})^2$
Entspricht der Differenz $y_T - \bar{y}$

Residualverianz: $s_y^2{}_r = 1/n\text{-}1 * \sum(y_T - y_i)^2$
Entspricht der Differenz $y_T - y_i$ (Residuen)

Abb.12

Wie schon gesagt, haben wir die erklärte Varianz in Beispiel 2 über das Bestimmtheitsmaß berechnet. Wir können die drei Varianzen aber auch direkt über ihre Gleichungen berechnen, wozu wir folgende Terme $\sum(y_i - \bar{y})^2$; $\sum(y_T - \bar{y})^2$ und $\sum(y_T - y_i)^2$ benötigen.

y_i	y_T	$y_T-\bar{y}$	$(y_T-\bar{y})^2$	y_T-y_i	$(y_T-y_i)^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$
2	3	-3	9	1	1	-4	16
8	6	0	0	-2	4	2	4
8	9	3	9	1	1	2	4
$\bar{y}$	**$\bar{y}_T$**		**Summe**		**Summe**		**Summe**
6	**6**		**18**		**6**		**24**

Tabelle 5

Daraus folgt Erklärte Varianz: $1/n\text{-}1*\sum(y_T - \bar{y})^2 = ½ * 18 = 9$
Gesamtvarianz: $1/n\text{-}1*\sum(y_i - \bar{y})^2 = ½ * 24 = 12$
Residualvarianz: $1/n\text{-}1*\sum(y_T - y_i)^2 = ½ * 6 = 3$

Wir kennen jetzt die drei Varianzen und da

$$s_{y^2e} = s_{y^2g} * r^2$$

können wir aus ihnen den Korrelationskoeffitienten r berechnen:

$$r^2 = s_{y^2e}/s_{y^2g}$$
$$r^2 = 9/12 = 0{,}75$$
$$r = 0{,}867$$

Diesen Wert erhalten wir für r auch über ein TKS.

Mit den Themen Korrelation und Regression ist der erste Teil unserer Einführung in die Statistik, die deskriptive Statistik, abgeschlossen. Zu diesen Teil möchte ich ein Buch empfehlen, welches ich vor Kurzem kennenlernte: **Griffiths,D.** Statistik von Kopf bis Fuß, O'Reilly, Köln, 2009. Es ist die deutsche Übersetzung des Originals **Griffiths,D.** Head First Statistics, O'Reilly, Köln, 2009. Beide Ausgaben sind lesenswert! Mit Kapitel 14 beginnt der zweite Teil, die Inferenzstatistik.

13.6 Übungen

Übung 1
Bei 10 Buschbohnen (Phaseolus vulgaris „Valja“) wurde Länge und Breite gemessen. Die Ergebnisse stehen in Tabelle 5. Da der Regressor nicht begründet festgelegt werden kann, weisen wir die Länge arbiträr der Variablen X zu. Berechnen Sie b_x; b_y; r; B und die erklärte Varianz als %-Wert. Gehen Sie davon aus, dass die Voraussetzungen für die Berechnung gegeben sind.

X	Y
Länge	Breite
mm	mm
10,0	4,0
11,0	4,5
12,9	5,1
10,3	4,5
12,7	5,2
11,9	5,2
10,5	4,5
10,5	5,3
10,5	4,7
11,0	4,6

Tabelle 6

Ergebnisse
$b_x = 0{,}28$. $b_y = 1{,}63$. $r = 0{,}68$. $B = 0{,}46$. $s_{y^2e} = 46\%$