Vorlesung Statistik – WiSe 2016/17

Teil 2
Korrelation und Regression
10.11.2016

Prof. Dr. Peter Martus
UKT Tübingen

Studiengänge
Molekulare Medizin FS 3(BSc)
Biochemie FS 3(BSc)
Medizininformatik FS 3 (BSc)

Vorlesung 2

- Einführendes Beispiel
- Exkurs über Geraden und ihre Gleichungen
- Regressionsanalyse, Formeln und Rechenbeispiel
- Korrelation und Bestimmtheitsmaß
- Fehlerquellen

Teil 1
Einführendes Beispiel

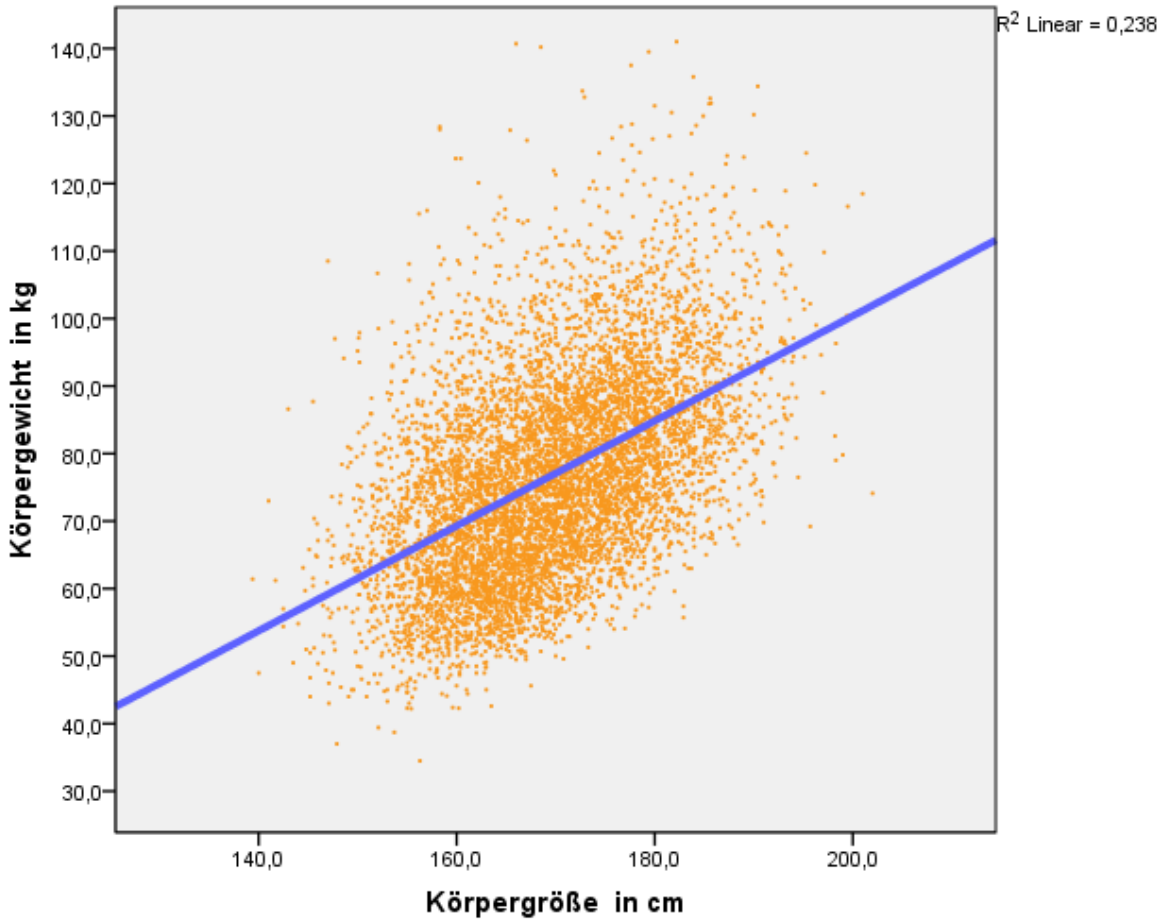
Einführung

In der letzten Vorlesung haben wir das Streudiagramm kennengelernt.

In der heutigen Vorlesung werden wir uns mit der Frage befassen, wie wir in dieses Diagramm eine Gerade legen können, die den Zusammenhang von x und y auf eine einfache Art und Weise beschreibt.

Diese Gerade heißt „Regressionsgerade“ (außerhalb der Statistik manchmal auch „Ausgleichsgerade“).

Streudiagramm mit Regressionsgerade



Quelle: BGS 1998, Robert-Koch Institut Berlin, SUF

Die Regressionsgerade

Gewicht [kg] = -55 + 0.78 • Körpergröße [cm]

$y = b_0\,[kg] + b_1\,[kg/cm] \bullet x\,[cm]$

y = x-Achsenabschnitt + Steigung • x

y = Konstante + Steigung • x

Für 180 cm Körpergröße:

-55 + 0.78*180 = 85.4 Kg

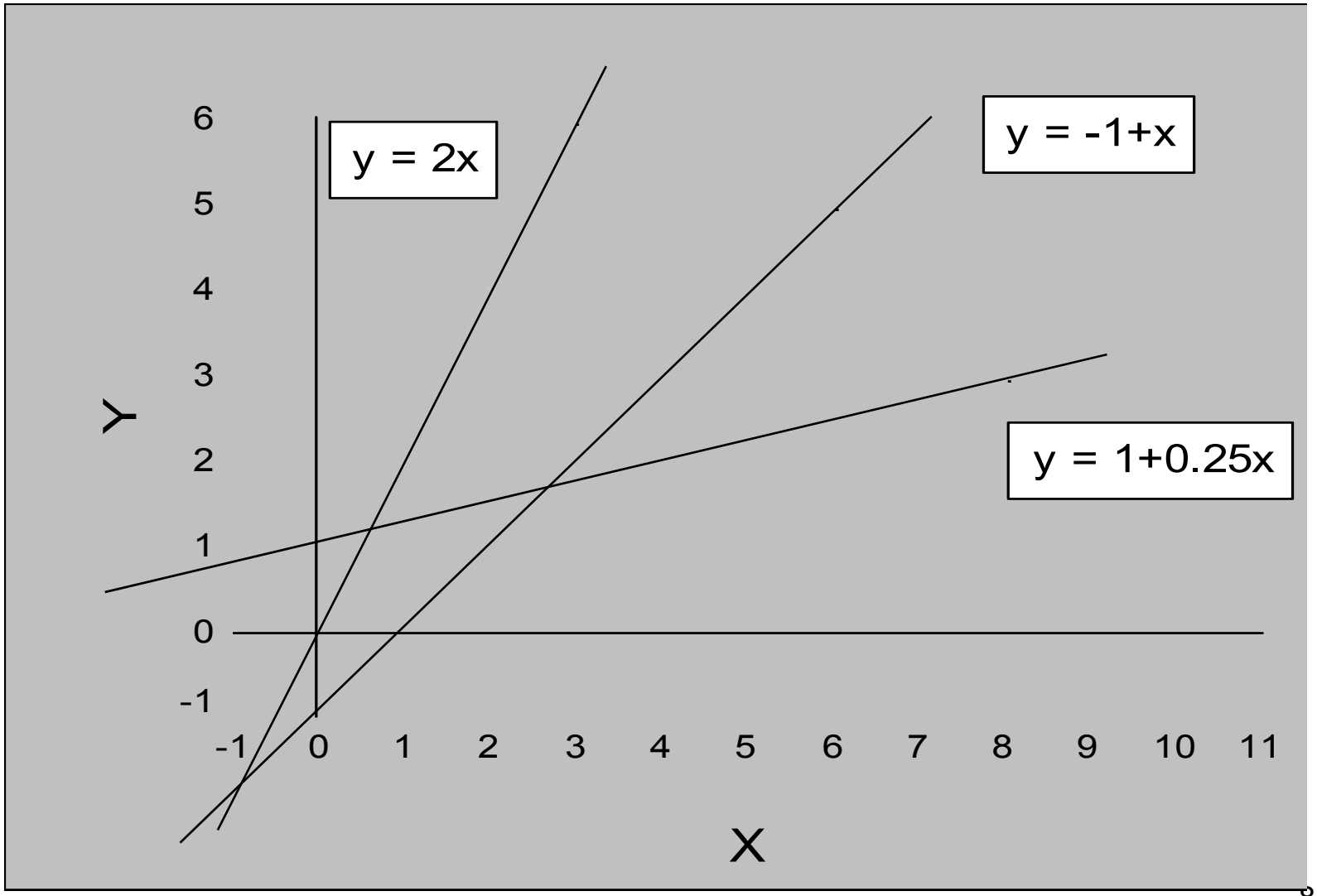
für 190 cm Körpergröße:

-55 + 0.78*190 = 85.4+0.78*10 = 93.2 Kg.

b_0 hat immer die physikalische Einheit von y, hier [kg], die Einheit von y/x, hier [kg/cm].

Teil 2
Exkurs: Geraden und Ihre Gleichungen

Exkurs: Beispiele für Geraden



Gerade und Gleichung, Sonderfälle

Identifikation von b_0 und b_1 in der Grafik:

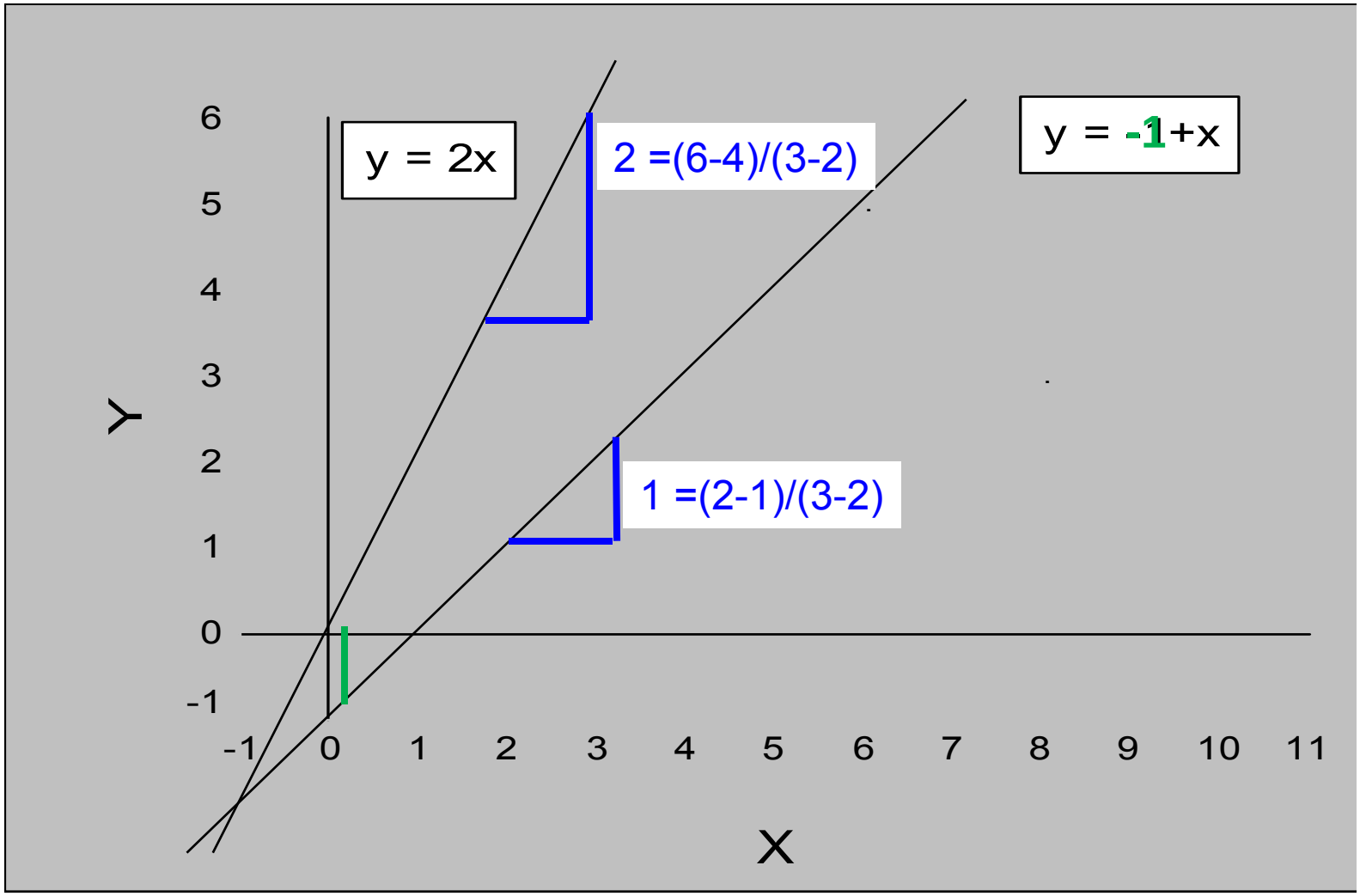
b_0 ist der Schnittpunkt der Geraden mit der y-Achse.
b_1 lässt sich durch Steigungsdreiecke bestimmen.

Sonderfälle:

$b_0 = 0$, b_1 beliebig	Gerade durch den Ursprung (0,0) mit Steigung b
$b_1 = 0$, b_0 beliebig	Waagerechte Gerade durch (0,a)
$b_0 = 0$, $b_1 = 1$	Winkelhalbierende mit y = x

Geraden senkrecht zur x-Achse sind nicht darstellbar.

Konstante und Steigung von Geraden



Teil 3
Regressionsanalyse
Formeln und Rechenbeispiel

Korrelations- und Regressionsanalyse

In realen Daten sieht man **nie** eine **exakt** lineare Abhängigkeit sondern nur eine Punktwolke, die „mehr oder weniger gut“ zu einer Geraden passt.

Daraus ergeben sich die folgenden Fragestellungen:

1. Welche Gerade passt am besten zur Punktwolke und wie wird diese Gerade bestimmt?
2. Wie gut ist diese Anpassung?

Die Regressionsanalyse beantwortet die erste, die Korrelations-analyse die zweite Frage.

Ob eine Punktwolke überhaupt für eine Gerade „geeignet“ ist, diskutieren wir am Ende der Vorlesung und in den Übungen.

Die Methode der kleinsten Quadrate

Die Gerade wird mit Hilfe der Methode der kleinsten Quadrate angepasst:

Die Summe der quadrierten Abweichungen der Messwerte von den laut Geradengleichung vorhergesagten Werten,

$$\sum_i \left[y_i - \left(b_0 + b_1 \cdot x_i \right) \right]^2$$

soll minimiert werden. Dies entspricht grafisch der Minimierung der Abstände der Punkte auf die Geraden in **y-Richtung.**

Die vorhergesagten y-Werte, $a_{YX} + b_{YX} \bullet x$, werden mit $\hat{y}$ bezeichnet, die Abweichungen $y - \hat{y}$ heißen „**Residuen**“.

Berechnung der Koeffizienten b_0 und b_1

Zur Berechnung wird die Kovarianz von x und y benötigt.

$$Cov(x,y) = s_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Setzt man x anstelle von y in den letzten Faktor ein, sieht man, dass die Kovarianz von x mit x gerade die Varianz von x ist.

Ganz analog zur Abkürzung der Kovarianz, s_{xy}, bezeichnet man die Varianz von x auch mit s_{xx} und die Streuung von x mit s_x.

Berechnung der Koeffizienten b_0 und b_1

Mittels Kurvendiskussion lässt sich zeigen, dass die Werte

$$\hat{b}_1 = \frac{s_{yx}}{s_{xx}}, \qquad \hat{b}_0 = \bar{y} - \hat{b}_1 \cdot \bar{x}$$

das Kleinstquadrate-Kriterium erfüllen.

b_1 heißt auch „Regressionskoeffizient von y auf x“.

Achtung:

Wenn man x und y vertauscht kommt man auf eine andere Gleichung als wenn man die Regressionsgleichung der Analyse von y auf x nach x auflöst. Man muss also unterscheiden, ob die Regression von y auf x oder von x auf y erfolgt.

Beispiel zur Regressionsanalyse

Diagnostische Studie

Bei 600 Freiwilligen beiderlei Geschlechts, die mindestens 70 Jahre alt waren, wurde die Nierenfunktion (Glomuläre Filtrationsrate, GFR) durch Analyse der Eliminationskurve einer Markersubstanz (Iohexol) genau gemessen.

Zusätzlich wurde der Serumwert des Kreatinins bestimmt.

Mit Hilfe einer Regressionsanalyse sollten die GFR Werte aus den Kreatininwerten geschätzt werden.

Die Kreatininmessung ist nicht aufwändig (eine Blutentnahme, Routinemessung) während die GFR Messung die Injektion einer Markersubstanz sowie 6 Blutentnahmen über 3 Stunden und jeweils aufwändige Messungen der Iohexolkonzentration erfordert.

Wir betrachten hier zehn Beispielmessungen und zeigen im Detail, wie die Regressionsgerade berechnet wird.

In der eigentlichen Studie war die Situation natürlich komplizierter: Bei 570 verwertbaren Messungen wurde in die Formel Alter und Geschlecht integriert, die GFR Messungen wurden auf die Körperoberfläche normiert und es wurde ein doppelt logarithmisches Modell verwendet (Schäffner et al., Ann Intern Med. 2012; 157:471-81).

Rohdaten und deskriptive Analyse zum Vergleich von Serum-Kreatinin (Krea) und der Glomerulären Filtrationsrate (GFR)

Nr.	GFR	GFR-$\overline{GFR}$	Quadrat	Krea	Krea-$\overline{Krea}$	Quadrat
1	63.95	10.93	119.39	1.14	0.02	0.00
2	67.51	14.49	210.06	0.95	-0.17	0.03
3	55.02	2.00	4.00	0.84	-0.28	0.08
4	40.67	-12.35	152.58	0.97	-0.15	0.02
5	77.36	24.35	592.68	0.66	-0.46	0.21
6	33.29	-19.73	389.20	1.88	0.76	0.58
7	45.75	-7.27	52.90	1.31	0.19	0.04
8	35.46	-17.56	308.23	1.44	0.32	0.10
9	52.53	-0.49	0.24	0.95	-0.17	0.03
10	58.66	5.64	31.79	1.03	-0.09	0.01
Σ	530.19	0	1861.08	11.17	0	1.09
$\Sigma/10$	**53.02**			**1.12**		
$\Sigma/9$			**206.79**			**0.122**
$\sqrt{\Sigma/9}$			**14.38**			**0.349**

Kleine Abweichungen aufgrund von Rundungsfehlern möglich

Berechnung der von Serum-Kreatinin (Krea) und der Glomerulären Filtrationsrate (GFR)

Nr.	GFR	Krea	GFR-$\overline{GFR}$	Krea-$\overline{Krea}$	(GFR-$\overline{GFR}$)*(Krea-$\overline{Krea}$)
1	63.95	1.14	10.93	0.02	0.25
2	67.51	0.95	14.49	-0.17	-2.42
3	55.02	0.84	2.00	-0.28	-0.55
4	40.67	0.97	-12.35	-0.15	1.82
5	77.36	0.66	24.35	-0.46	-11.13
6	33.29	1.88	-19.73	0.76	-15.05
7	45.75	1.31	-7.27	0.19	-1.40
8	35.46	1.44	-17.56	0.32	-5.67
9	52.53	0.95	-0.49	-0.17	0.08
10	58.66	1.03	5.64	-0.09	-0.49
Σ	530.19	11.17	0	0	-34.57
Σ/10	**53.02**	**1.12**			
Σ/9					**-3.84**

Kleine Abweichungen aufgrund von Rundungsfehlern möglich

Ergebnisse der Regressionsanalyse der Glomulären Filtrationsrate (GFR) auf Serum-Kreatin (Krea)

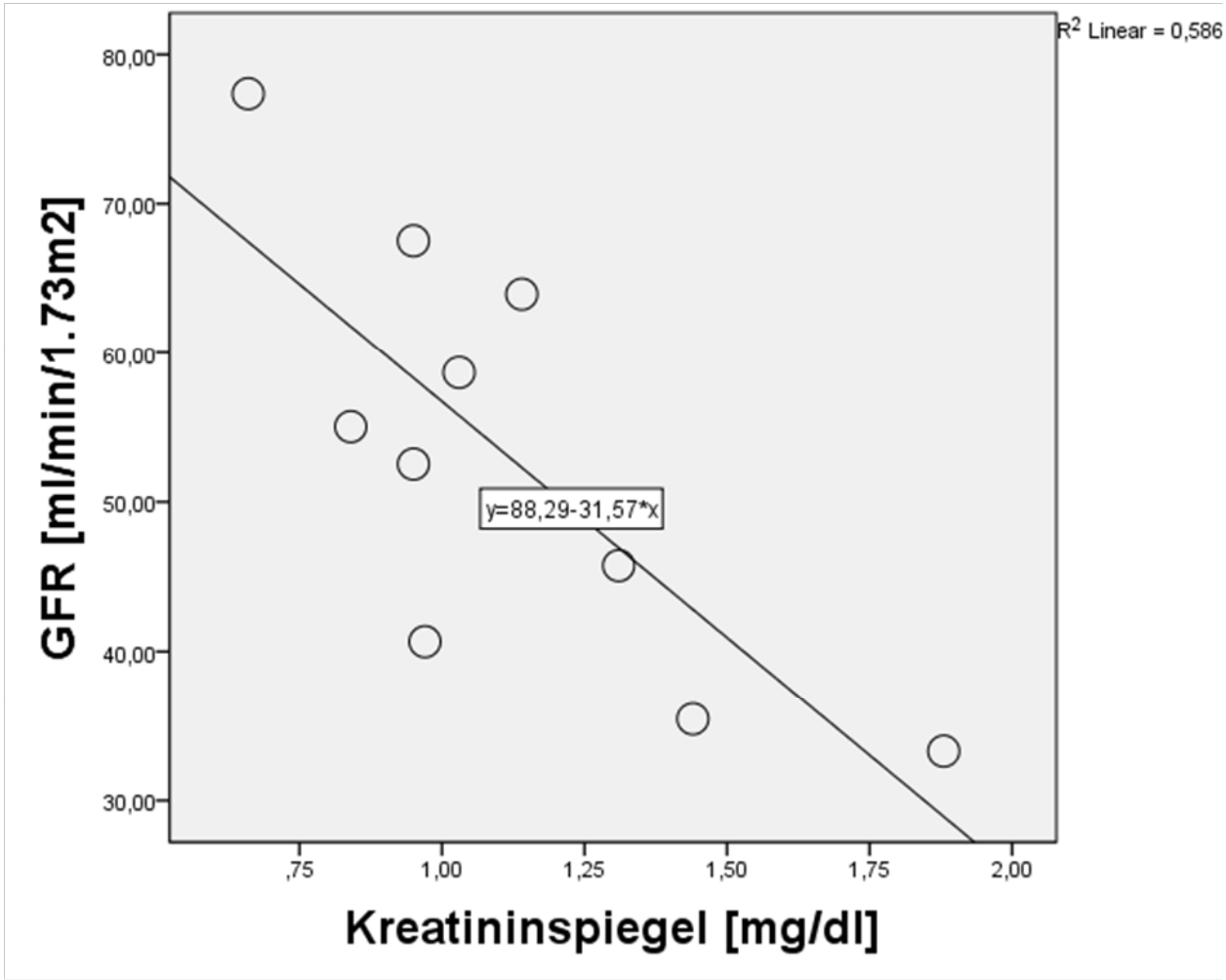
$$b_1 = s_{YX} / s^2_x = -3.84 / 0.122 = -31.58$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 53.02 - (-31.58 \cdot 1.12) = 88.29$$

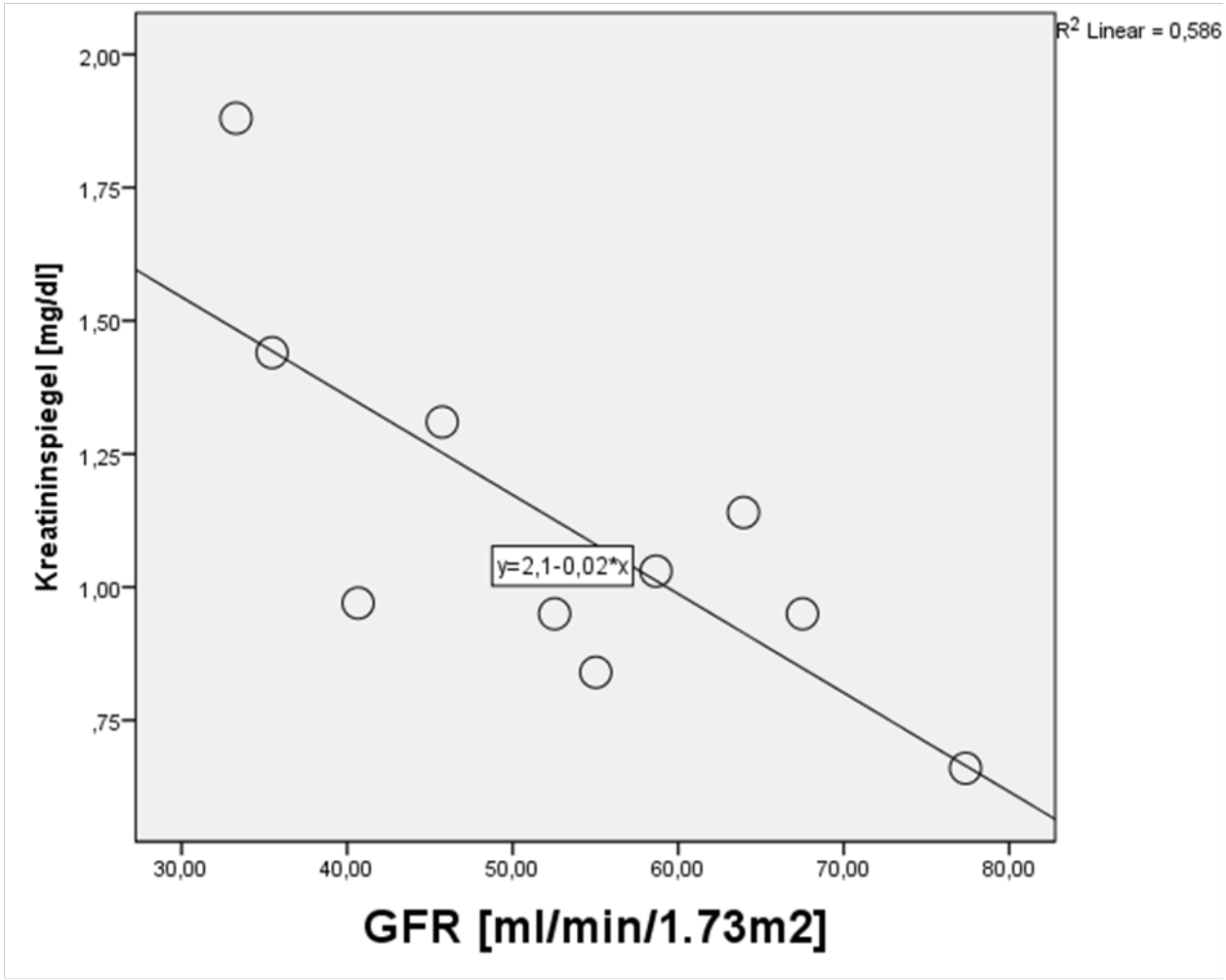
$$\hat{GFR} = 88.29 - 0.31.58 \cdot Krea$$

Kleine Abweichungen aufgrund von Rundungsfehlern möglich

Vergleich von Kreatinin (Krea) und der Glomerulären Filtrationsrate (GFR)



Regression von x (Krea) auf y (GFR)



Übung

Berechnen Sie die Gleichung der linearen Regression von Kreatinin auf die Glomeruläre Filtrationsrate (Ergebnis bereits in Graphik).

Vergleichen Sie diese Gleichung mit der nach x aufgelösten ursprünglichen Regressionsgleichung der GFR auf Krea.

Inhaltlich ist diese Gleichung allerdings nicht besonders hilfreich, warum?

Teil 4
Korrelation und Bestimmtheitsmaß

Der Korrelationskoeffizient als Effektmaß der Stärke des Zusammenhangs zwischen X und Y

Ein Maß für die Güte des Zusammenhangs zwischen zwei quantitativen Variablen ist der

Pearsonsche Korrelationskoeffizient.

Dieser Koeffizient ist der Quotient aus der Kovarianz s_{xy} und den beiden Streuungen s_x und s_y. Er wird mit r_{xy} bezeichnet.

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

Übung

Berechnen Sie den Korrelationskoeffizienten zwischen Kreatinin und der Glomerulären Filtrationsrate.

Warum spielt hier die Reihenfolge von x und y keine Rolle?

Das Bestimmtheitsmaß als Effektmaß für die Stärke des Zusammenhangs zwischen X und Y

Ein weiterer Parameter heißt r-Quadrat oder „Bestimmtheits-maß“ und ist auch tatsächlich das Quadrat des Pearson‘ schen Korrelationskoeffizienten.

Man kann zeigen, dass sich dieser Parameter alternativ als Quotient aus der „erklärten Varianz“ und der Gesamtvarianz der y-Daten berechnen lässt.

$$r^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2},$$

Regressionskoeffizient, Korrelationskoeffizient und Bestimmtheitsmaß

Vorzeichen:

Die Vorzeichen von Regressionskoeffizient und Korrelationskoeffizient stimmen immer überein. Ist es positiv, steigt die Regressionsgerade, ist es negativ, fällt sie. Das Bestimmtheitsmaß ist immer positiv oder Null.

Wertebereich:

Der Wertebereich des Regressionskoeffizienten umfasst alle reellen Zahlen. Der Korrelationskoeffizient liegt im Intervall [-1,1], Das Bestimmheitsmaß liegt im Intervall [0,1].

Physikalische Einheiten:

Die physikalische Einheit des Regressionskoeffizienten ist der Quotient aus der Einheit der y-Werte und der x-Werte. Korrelationskoeffizient und Bestimmtheitsmaß sind dimensionslos. Das Bestimmtheitsmaß lässt sich als Prozentwert interpretieren.

Teil 5
Fehlerquellen

Fehlerquellen bei Regression und Korrelation

Kausalität:

Korrelations- und Regressionsanalysen können niemals beweisen, dass die y-Variable durch die x-Variable beeinflusst wird. Es kann umgekehrt sein oder eine nicht gemessene Variable beeinflusst beide (Störche und Geburten, Wahrnehmung der Arbeitsbelastung und Burnout).

Ungeeignete Werteverteilung:

Ausreißer oder heterogene Stichproben können Korrelationen und Regressionsgleichungen verfälschen (sowohl zu scheinbar großen als auch zu scheinbar kleinen Zusammenhängen).

Nichtlineare Zusammenhänge:

Im allgemeinen überprüft man quadratische und kubische Zusammenhänge, ggf. logarithmische Transformationen.

Die letzten beiden Fehlerquellen kann man grafisch erkennen, für die Kausalität muss man inhaltlich und vom Studiendesign her denken.