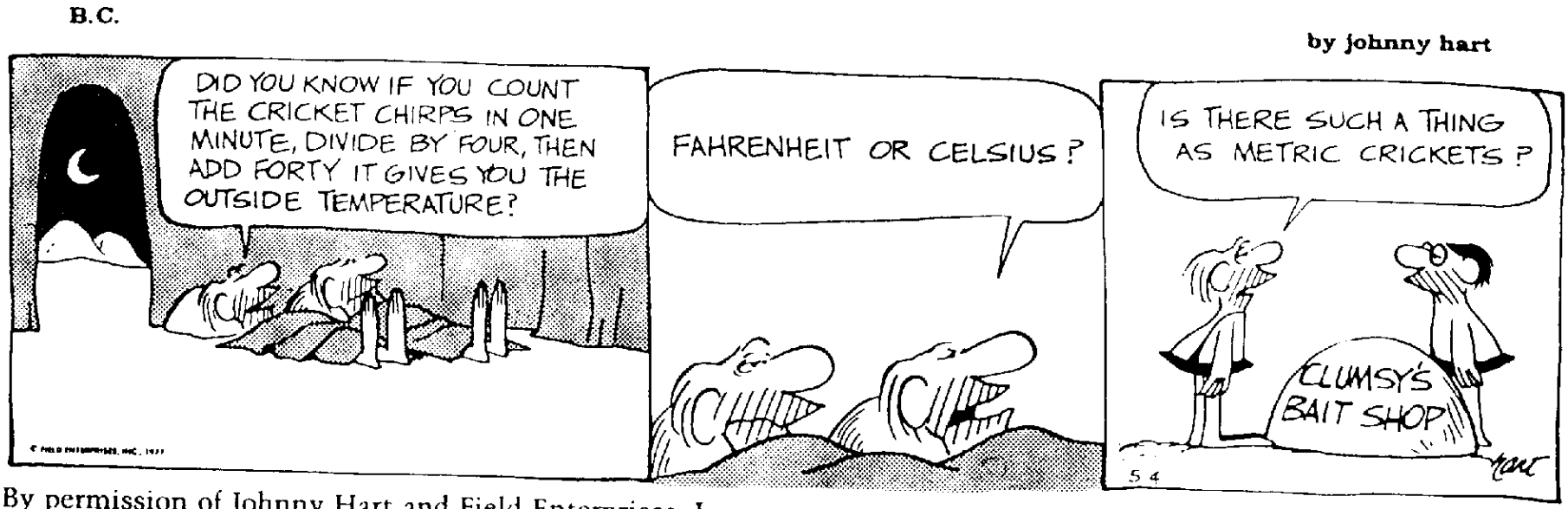
Crashkurs
Einführung Biostatistik

Prof. Burkhardt Seifert

Abteilung Biostatistik, ISPM
Universität Zürich

- Deskriptive Statistik
- Wahrscheinlichkeitsrechnung, Versuchsplanung
- Statistische Inferenz
  - Prinzip statistischer Tests
  - Konfidenzintervalle
  - Stichprobengrösse, Power
- **Korrelation und einfache lineare Regression**

Korrelation und einfache lineare Regression



By permission of Johnny Hart and Field Enterprises, Inc.

- Korrelation (Definition, Tests, Rangkorrelation, Gefahren)
- Einfache lineare Regression (Modell, Kleinste Quadrate, erklärte Varianz, Tests)
- Multiple Regression (Idee)

Korrelation und Regression

- Analyse des Zusammenhangs von **zwei stetigen Variablen (bivariate Daten)**

Beispiele:

Beziehung Gewicht zu Grösse

Zusammenhang zwischen Körperfett und BMI

Fragestellungen:

1. Welcher Zusammenhang besteht zwischen zwei Variablen x und y ?

2. Lässt sich die Variable y aus der Variablen x vorhersagen?

Bivariate Daten

- Beobachtung von zwei **stetigen** Variablen (x, y) an der selben Beobachtungseinheit

$\longrightarrow$ **paarweise** Beobachtungen $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Beispiel: Anteil Körperfett bestimmt durch Wiegen unter Wasser

$n = 241$ Männer im Alter von 22 bis 81 Jahren

$y =$ Anteil Körperfett (in %)

$x =$ BMI = Gewicht/Höhe^2 (in kg/m^2)

Nr	BMI	bodyfat
1	23.65	12.3
2	23.36	6.1
3	24.69	25.3
4	24.91	10.4
...	...	...
240	27.01	26
241	29.8	31.9

- alle Beobachtungen eines Individuums in der selben Zeile
- Beobachtungen verschiedener Individuen untereinander

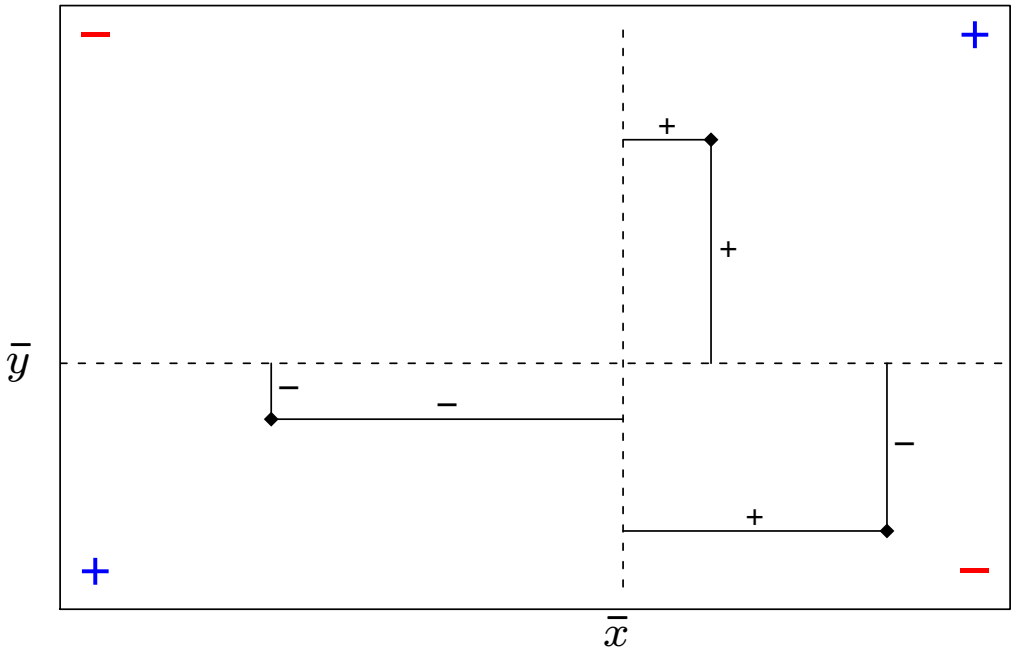
Korrelation

Pearsons Produkt–Moment Korrelation $r = \frac{s_{xy}}{s_x\, s_y}$

mit Kovarianz

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

Zähler s_{xy}:



Nenner: Standardisierung

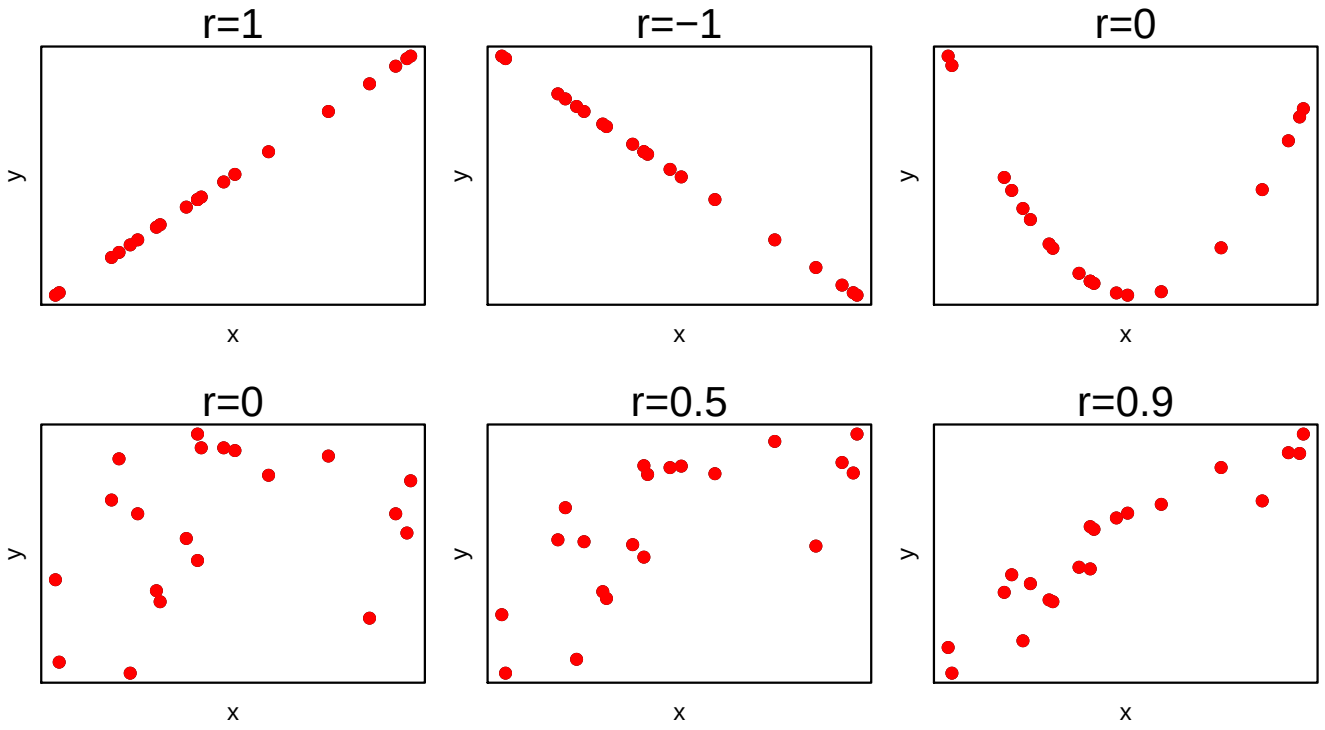
- Vorzeichen zeigt Richtung des linearen Zusammenhangs
- Grösse zeigt Intensität des linearen Zusammenhangs

$-1 \leq r \leq 1$

$r = 1 \quad \longrightarrow \quad$ deterministisch positiver linearer Zusammenhang

$r = -1 \quad \longrightarrow \quad$ deterministisch negativer linearer Zusammenhang

$r = 0 \quad \longrightarrow \quad$ kein linearer Zusammenhang



Tests auf linearen Zusammenhang

Besteht ein linearer Zusammenhang jenseits des Zufalls?

wissenschaftliche Hypothese: wahre Korrelation $\rho \neq 0$

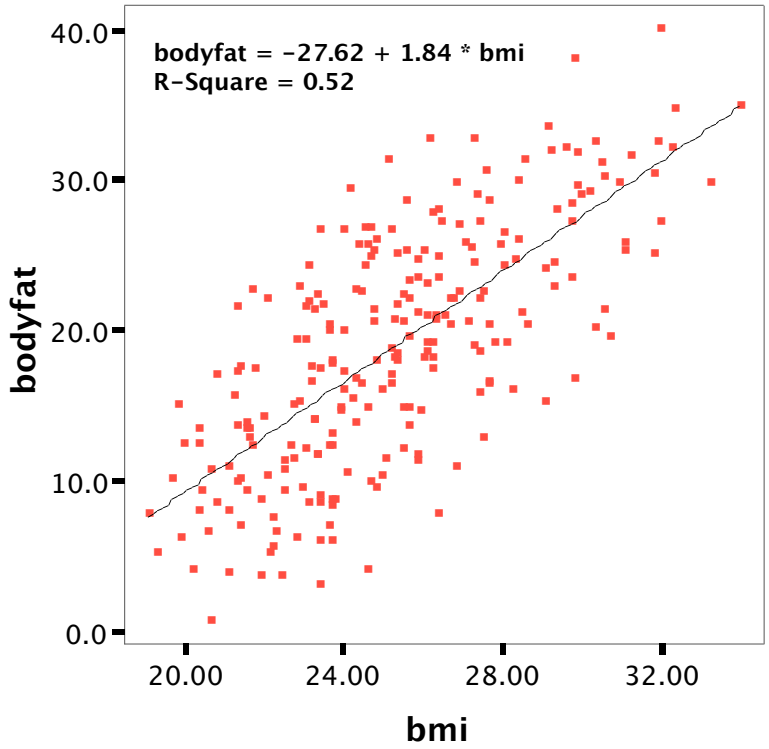
Nullypothese: wahre Korrelation $\rho = 0$

Annahmen: (x, y) gemeinsam **normalverteilt**, Paare **unabhängig**

Correlations

		bodyfat	bmi
bodyfat	Pearson Correlation	1	.718**
	Sig. (2-tailed)		.000
	N	241	241
bmi	Pearson Correlation	.718**	1
	Sig. (2-tailed)	.000	
	N	241	241

**. Correlation is significant at the 0.01 level (2-tailed).



Spearmans Rangkorrelation

Behandlung von Ausreissern?

Testen ohne Normalverteilung?

Idee: Ähnlich zu Mann-Whitney Mittelwerts–Vergleich

Vorgehen: Man korreliert die Ränge miteinander anstatt die Zahlen selber.

Beispiel: Pearson Korrelation 0.72

Correlations

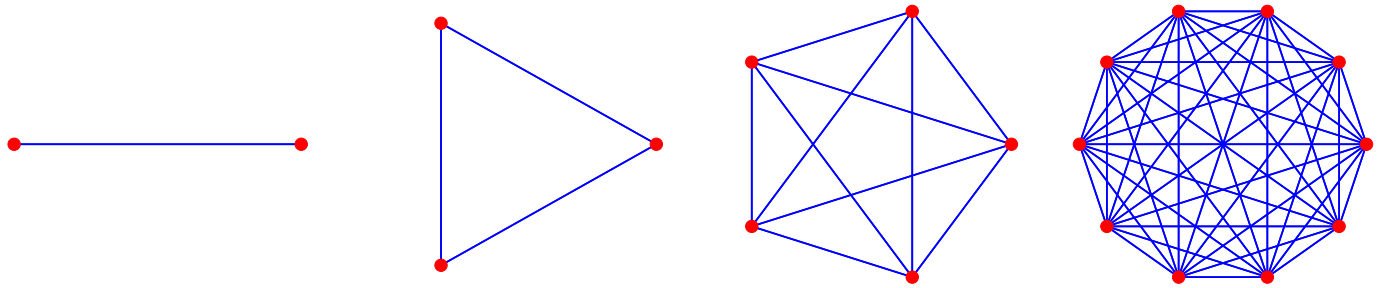
			bodyfat	bmi
Spearman's rho	bodyfat	Correlation Coefficient	1.000	.705**
		Sig. (2-tailed)	.	.000
		N	241	241
	bmi	Correlation Coefficient	.705**	1.000
		Sig. (2-tailed)	.000	.
		N	241	241

**. Correlation is significant at the 0.01 level (2-tailed).

- Für normalverteilte Daten sehr ähnlich zu Pearson-Korrelation

Gefahren der Fehlinterpretation von Korrelationen

- Wahrscheinlichkeit einer falschen Signifikanz 5% für **einen** statistischen Test



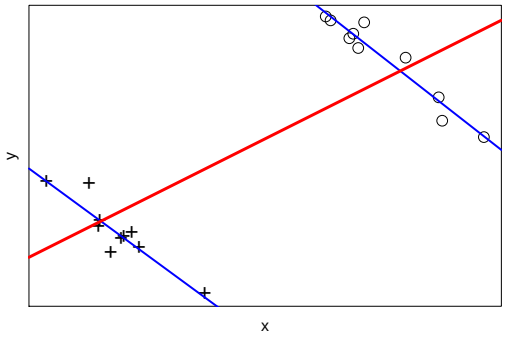
Anzahl Variable	2	3	5	10
Anzahl Korrelationen	1	3	10	45
Wahrsch. falscher Signif.	0.05	0.14	0.40	0.91

- Anzahl der Paare steigt dramatisch mit der Anzahl Variablen

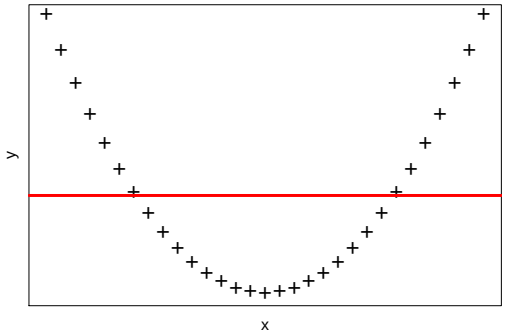
$\longrightarrow$ erhöhte Wahrscheinlichkeit für falsche Signifikanzen

Gefahren der Fehlinterpretation von Korrelationen (contd)

- Heterogenitätskorrelation
  (kein oder sogar umgekehrter
  Zusammenhang in den Gruppen)



- nichtlinearer Zusammenhang
  (starker Zusammenhang, aber $r = 0$,
  d. h. r nicht aussagekräftig)



- Scheinkorrelationen über Zeit (gemeinsamer Trend)
- allgemeiner: Konfundierung durch 3. Variable

$\longrightarrow$ Anzahl der Störche und Geburten in einem Gebiet sind korreliert

Lineare Regression

- Korrelation misst Zusammenhang symmetrisch
- Regressionsanalyse ist statistische Analyse der Wirkung einer Variablen auf eine andere $\longrightarrow$ gerichtete Beziehung

y = abhängige Variable, Zielvariable

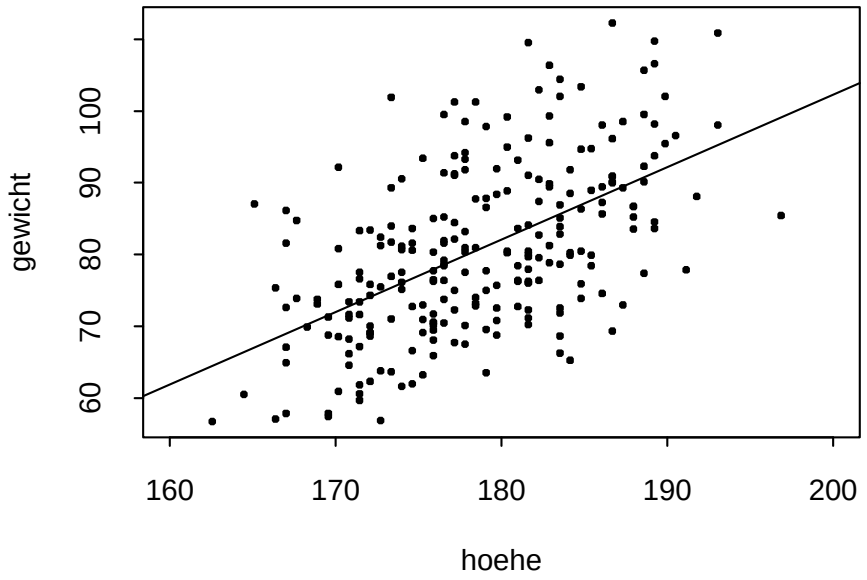
x = unabhängige Variable, Prädiktor
oft **nicht zufällig** (Zeit, Alter) und **nicht stetig** (Geschlecht, Gruppe)

Ziel: quantitatives Gesetz finden wie sich y ändert, wenn x sich ändert

Beispiel: Ist das Gewicht ein gutes Mass für Übergewicht ?

Regression y = Gewicht, x = Höhe ($n = 241$ Männer)

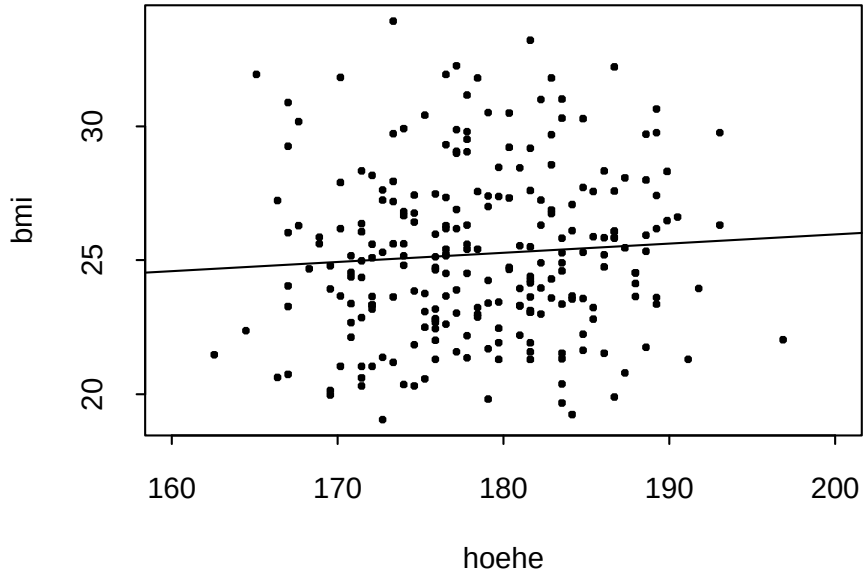
$$y = -99.7 + 1.01 \times x, \quad r^2 = 0.31, \quad p < 0.0001$$



Das Gewicht hängt signifikant von der Körpergrösse ab.

Regression y = BMI, x = Höhe

$$y = 19.2 + 0.034 \times x, \quad r^2 = 0.005, \quad p = 0.27$$



- Der BMI hängt nicht von der Körpergrösse ab, ist also ein besseres Mass für Übergewicht als Gewicht.

Statistisches Modell der linearen Regression

$y_i = a + b\,x_i + \varepsilon_i \quad i = 1, \ldots, n$

$a + b\,x$ = „wahre“ Regressionsfunktion

a = Achsenabschnitt (Wert der Regressionsfunktion für $x = 0$)

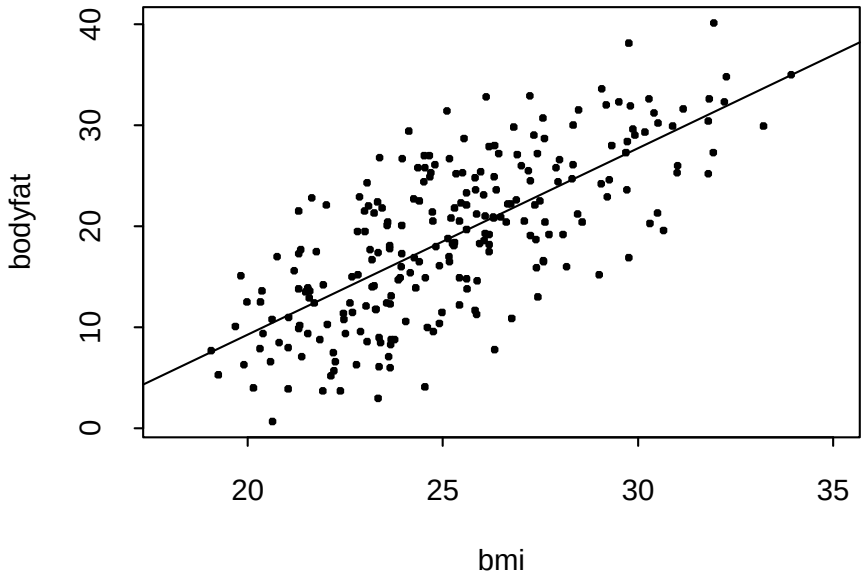
b = Steigung

ε = unbeobachtbare zufällige Residuen

Annahmen zu den ε_i :

- wahres Mittel = 0
- Varianz konstant
- für Tests und Konfidenzintervalle: **unabhängig und normalverteilt**

Beispiel: Ist der BMI ein gutes Mass für Körperfett ?



Geradengleichung: bodyfat $= -27.6 + 1.84 \times$ bmi

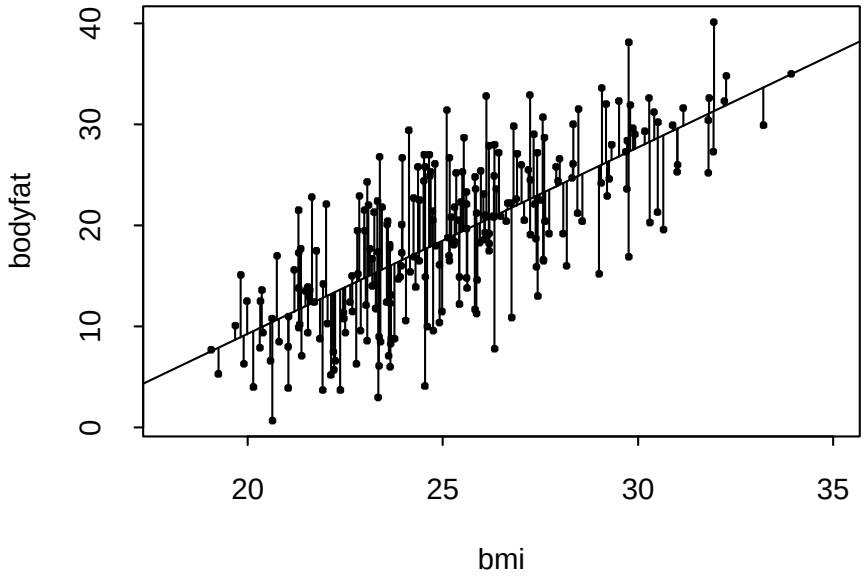
Interpretationen:

1. Männer mit einem BMI von 26 kg/m^2 haben im Mittel 20.2% Körperfett.

2. Männer mit einem um 1 kg/m^2 erhöhten BMI haben im Mittel 1.8% mehr Körperfett.

Die Methode der kleinsten Quadrate

- Wie legt man Gerade am besten durch die Daten ?



Gefitteter Wert (Vorhersage): $\widehat{y}_i = \widehat{a} + \widehat{b}\, x_i$

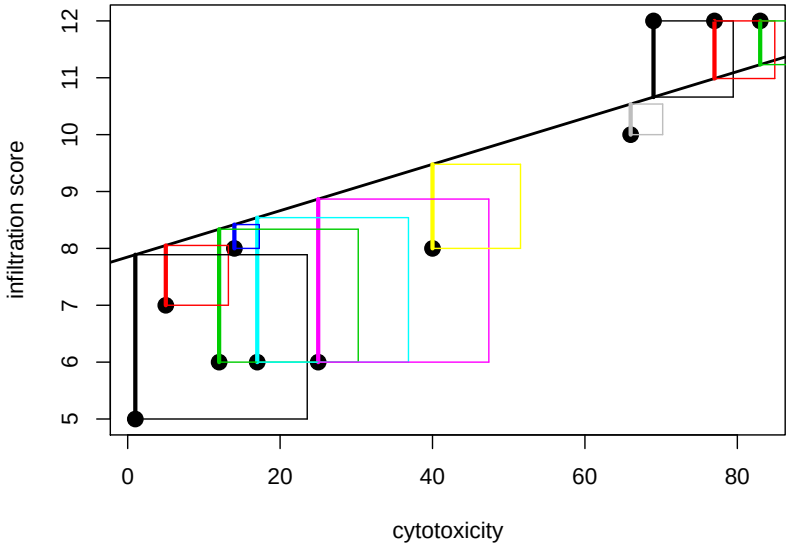
Wähle Parameter so, dass quadratische Abweichung

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

minimal wird

Die Methode der kleinsten Quadrate

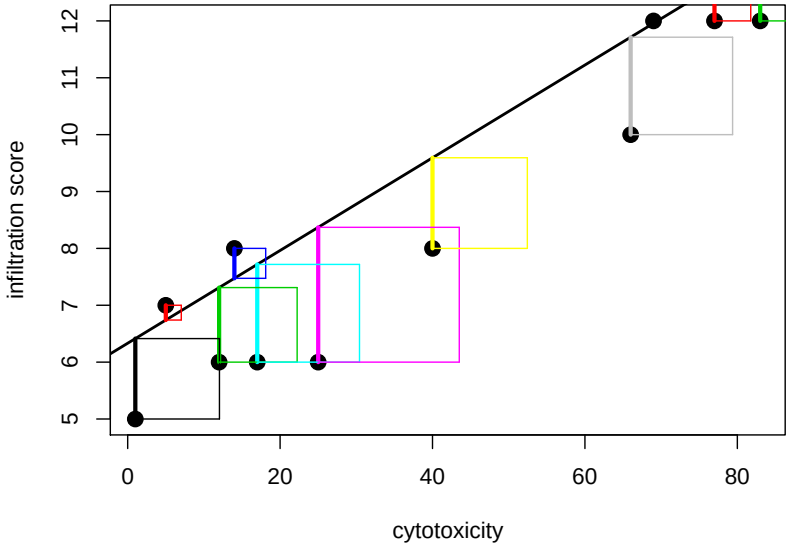
Es wäre möglich, die Gerade iterativ zu bestimmen:



Quadratsumme: 36

Die Methode der kleinsten Quadrate

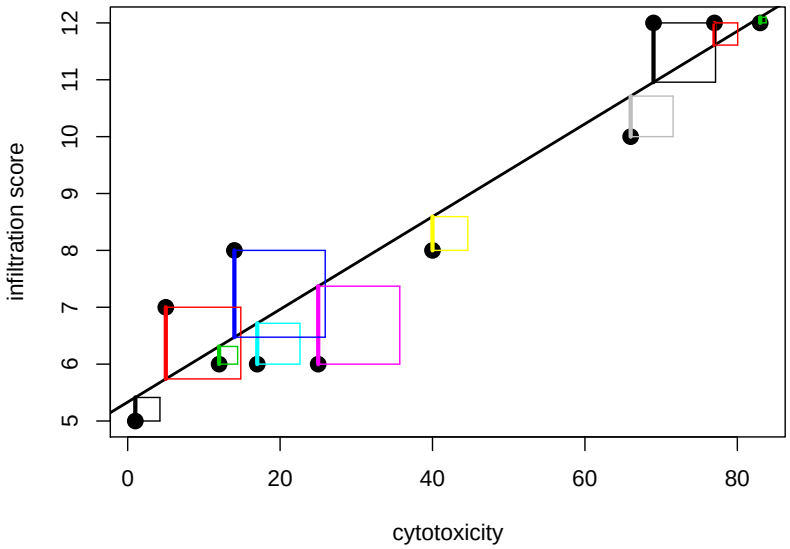
Es wäre möglich, die Gerade iterativ zu bestimmen:



Quadratsumme: 20

Die Methode der kleinsten Quadrate

Es wäre möglich, die Gerade iterativ zu bestimmen:



Quadratsumme: 9

- Es gibt aber eine explizite Formel:

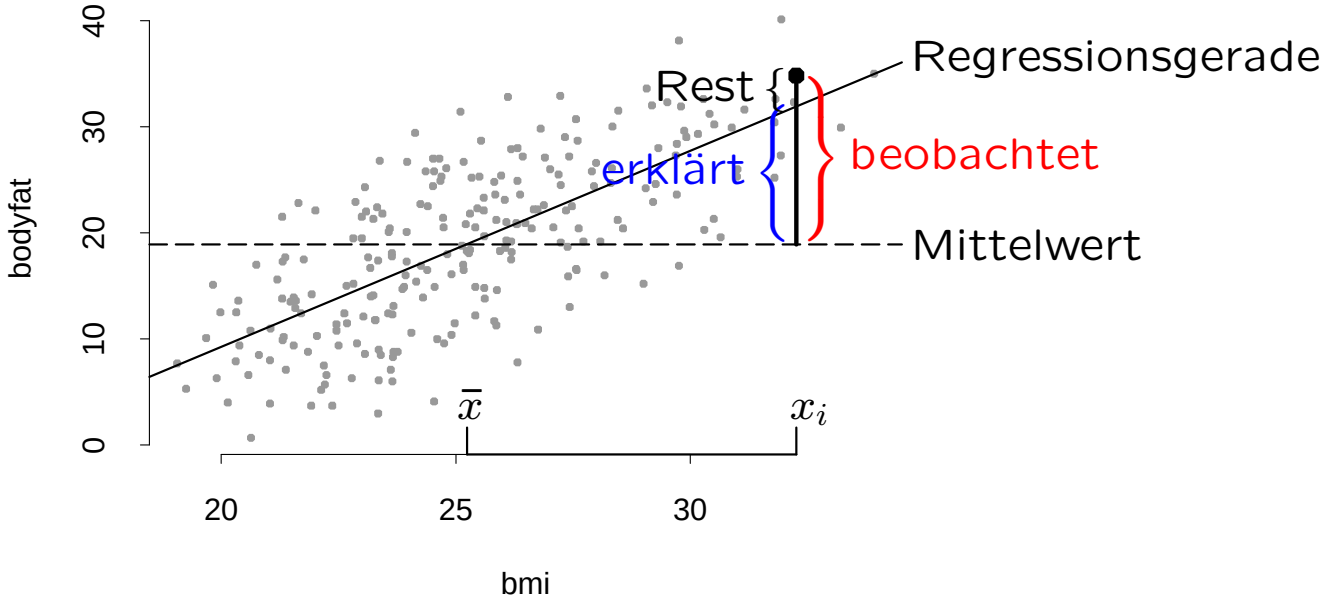
  Steigung: $\widehat{b} = \frac{s_{xy}}{s_x^2} = r\,\frac{s_y}{s_x}$

  Achsenabschnitt: $\widehat{a} = \bar{y} - \widehat{b}\,\bar{x}$

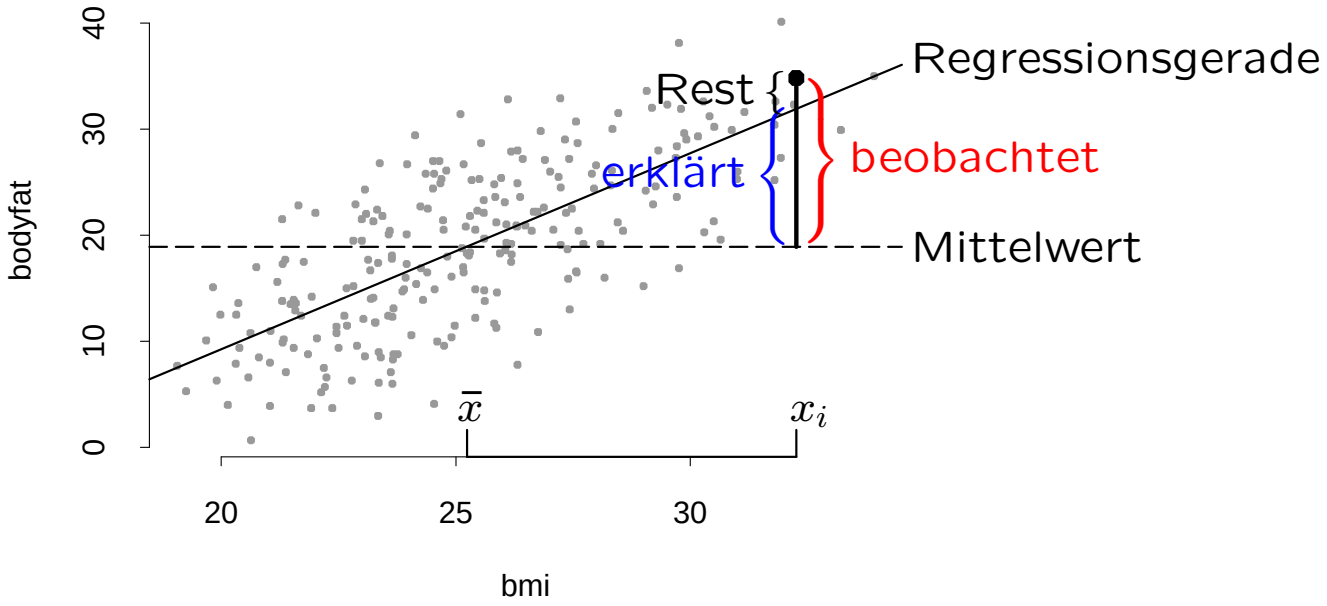
Durch die Regression erklärte Varianz

Frage: Kann die Regression von y auf x für die Vorhersage genutzt werden?

- Wieviel Varianz von y kann durch Regressionsgerade, d. h., durch Kenntnis von x, erklärt werden ?



- beobachtet = erklärt + Rest
- Der Abstand der Regressionsgeraden vom Mittelwert ist der Teil des Abstands der Beobachtung vom Mittelwert, der durch die Regression erklärt wird.



- Varianzen sind additiv, nicht Standardabweichungen (hier: in Stichprobe)

  Var(beobachtet) = Var(erklärt) + Var(Rest)

- Var(beobachtet) = Var(y) = s_y^2

- **erklärte Varianz:** Var(erklärt) = $r^2\, s_y^2$ für Interessierte: erklärt = $\widehat{b}(x_i - \bar{x})$

$$s^2_{\text{erklärt}} = \widehat{b}^2\, s_x^2 = \left(r\, \tfrac{s_y}{s_x}\right)^2 s_x^2 = r^2\, s_y^2$$

r^2 = Var(erklärt)/s_y^2 = **Teil der Varianz von y, der durch x erklärt wird**

Residualvarianz (nicht erklärt):

$$\text{Var(Rest)} = s_y^2 - r^2 s_y^2 = (1 - r^2)\, s_y^2$$

d. h., Beobachtungen streuen um Regressionsgerade mit SD

$$\text{SD(Rest)} = \sqrt{1 - r^2}\, s_y$$

r	0.3	0.5	0.7	0.9	0.99
$s_{\text{res}}/s_y = \sqrt{1 - r^2}$	0.95	0.87	0.71	0.44	0.14
Gewinn$= 1 - \sqrt{1 - r^2}$	5%	13%	29%	56%	86%

Beispiel :

	Mittelwert Körperfett ist	18.9
	Standardabweichung Körperfett s_y ist	8.0
	Korrelation(BMI,Körperfett) r ist	0.72
$\longrightarrow$	individuelles Körperfett variiert	
–	um Mittelwert 18.9 mit SD $s_y =$	**8.0**
–	um vorhergesagte Werte $\hat{y}$ mit SD $\sqrt{1 - r^2}\, s_y =$	**5.5**
	Gewinn:	**30%**

Test auf linearen Zusammenhang

Wissenschaftliche Hypothese: y ändert sich mit x systematisch ($b \neq 0$)

Nullhypothese: wahre Steigung $b = 0$

• falls (x, y) normalverteilt $\longrightarrow$ gleicher Test wie auf Korrelation $\rho = 0$

In Regressionsanalyse:

• alle Analysen **bedingt** für gegebene Werte $x_1, \ldots, x_n$

$\longrightarrow$ Verteilung von x nebensächlich, kann auch deterministisch sein

$\longrightarrow$ einfacher als Analysen der Korrelation

$\longrightarrow$ exaktes Konfidenzintervall für b

Output SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.718[a]	.516	.514	5.5472

a. Predictors: (Constant), bmi

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7835.191	1	7835.191	254.621	.000[a]
	Residual	7354.493	239	30.772		
	Total	15189.68	240			

a. Predictors: (Constant), bmi

b. Dependent Variable: bodyfat

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-27.617	2.939		-9.398	.000	-33.407	-21.828
	bmi	1.844	.116	.718	15.957	.000	1.617	2.072

a. Dependent Variable: bodyfat

Multiple Regression

Regression mit mehreren erklärenden Variablen

Gründe für multiple Regressionsanalyse:

1. mögliche Effekte von zusätzlichen „Stör“–Variablen in einer Studie mit einer Einflussgrösse eliminieren.

   Beispiel: Häufige Störgrösse ist das Alter. y = Blutdruck, x_1 = Dosierung Hypertensivum, x_2 = Alter.

2. mögliche Prognosefaktoren erforschen, von denen wir nicht wissen, ob sie wichtig oder redundant sind.

   Beispiel: y = Stenose, x_1 = HDL, x_2= LDL, x_3 = BMI, x_4 = Rauchen, x_5 = Triglyceride.

3. Formel zur besseren Vorhersage aus erklärenden Variablen entwickeln.

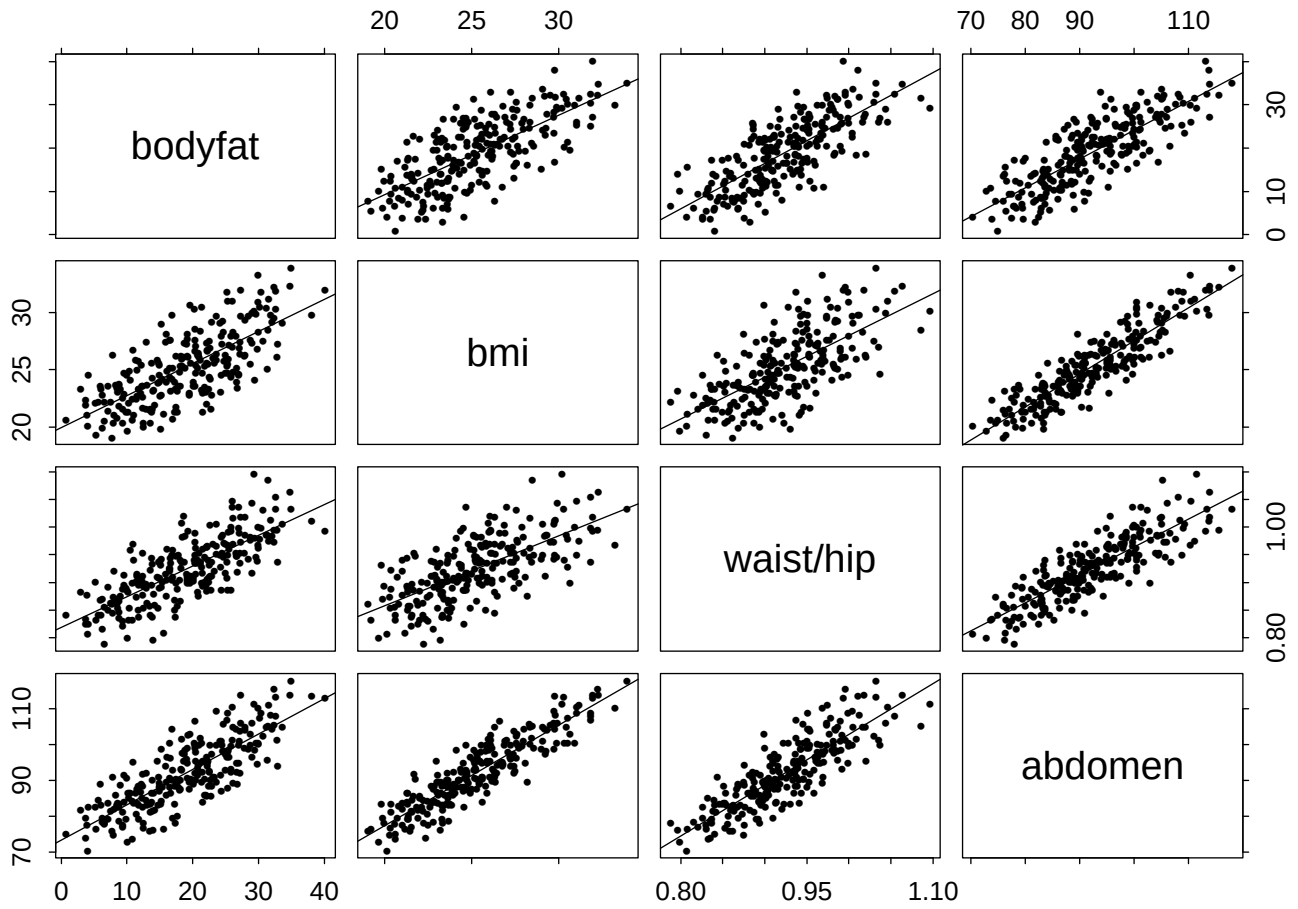
   Beispiel: y = Erwachsenengrösse, x_1 = Grösse als Kind, x_2 = Grösse der Mutter, x_3 = Grösse des Vaters.

4. Wirkung einer Variablen x_1 auf eine andere Variable y studieren, wobei Einfluss weiterer Variablen $x_2, \ldots, x_k$ berücksichtigt wird.

Beispiel:

y = Anteil Körperfett (in %)

in Abhängigkeit von x = BMI, Quotient Bauch– zu Hüftumfang, Bauchumfang



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.825[a]	.681	.677	4.5229

a. Predictors: (Constant), abdomen, waist/hip, bmi

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10341.53	3	3447.176	168.514	.000[a]
	Residual	4848.157	237	20.456		
	Total	15189.68	240			

a. Predictors: (Constant), abdomen, waist/hip, bmi

b. Dependent Variable: bodyfat

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-60.045	5.365		-11.192	.000	-70.615	-49.476
	bmi	.123	.236	.048	.519	.605	-.343	.588
	waist/hip	38.468	10.262	.280	3.749	.000	18.251	58.684
	abdomen	.438	.105	.533	4.183	.000	.232	.644

a. Dependent Variable: bodyfat

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-60.045	5.365		-11.192	.000	-70.615	-49.476
	bmi	.123	.236	.048	.519	.605	-.343	.588
	waist/hip	38.468	10.262	.280	3.749	.000	18.251	58.684
	abdomen	.438	.105	.533	4.183	.000	.232	.644

a. Dependent Variable: bodyfat

- Vorhersageformel:
  bodyfat = -60 + 0.12 * BMI + 38.5 * waist/hip + 0.44 * abdomen
- Regressionskoeffizienten, Konfidenzintervalle und P–Werte für einzelne Variable gelten bei festgehaltenen anderen Covariablen
- BMI nicht signifikant, wenn waist/hip und abdomen bekannt, Konfidenzintervall eng
- schrittweise Regression nötig