Stochastik für Informatiker
Einfache lineare Regression

Hanno Gottschalk

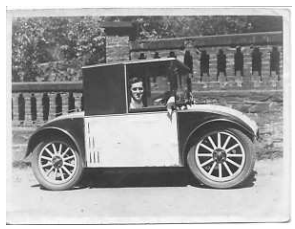
July 4, 2023

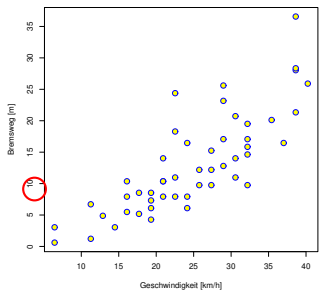
Inhaltsverzeichnis Vorlesung

Streudiagramme 3 / 32

Bremsweg - 1920

- $X : \Omega \to [0, \infty)$ Geschwindigkeit [kmh], $Y : \Omega \to [0, \infty)$ Bremsweg[m]
- Ω Bremsvorgänge von PKW 1920
- $\chi \subseteq \Omega$ Bremsvorgänge, die während einer Testkampagne 1920 gemessen wurden.





Je höher die Geschwindigkeit, desto länger der Bremsweg (aber nicht immer)

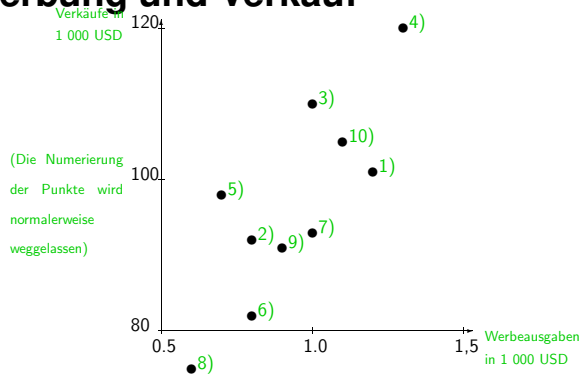
Werbung und Verkauf

Monat	1	2	3	4	5	6	7	8	9	10
Werbeausgaben	1.2	0.8	1.0	1.3	0.7	0.8	1.0	0.6	0.9	1.1
Verkäufe	101	92	110	120	90	82	93	75	91	105

Die Kontingenztabelle (2 Werte fehlen) ist im Wesentlichen leer (und damit weitgehend unbrauchbar)

W.ausg/Verk	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
1.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1.0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0.9	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.8	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0.7	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Streudiagramm Werbung und Verkauf



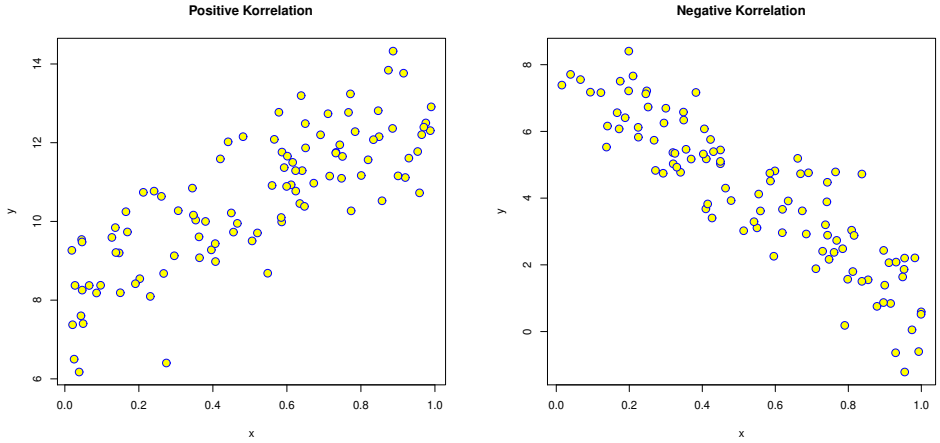
Monat	1	2	3	4	5	6	7	8	9	10
Werbeausgaben	1.2	0.8	1.0	1.3	0.7	0.8	1.0	0.6	0.9	1.1
Verkäufe	101	92	110	120	90	82	93	75	91	105

Besser, da nicht Unmengen redundanter Nullen mitgeschleppt werden . . .

Motivation

Wir sprechen von Korrelation von numerischen Merkmalen X und Y wenn

- Mit Anstieg von X steigt auch Y tendenziell an (*positive Korrelation*)
- Mit Anstieg von X fällt Y tendenziell ab (*negative Korrelation*)



Die emp. Kovarianz

Def.: Es seien in der Stichprobe $\chi \subseteq \Omega$, $\chi = \{\omega_1, \ldots, \omega_n\}$ für die Merkmale X und Y die Merkmalsausprägungen $x_1, \ldots, x_n$ bzw. $y_1, \ldots, y_n$ gemessen worden. Die emp. Kovarianz ist gegeben durch

$$\hat{\sigma}_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{n}{n-1}(\overline{xy} - \bar{x}\bar{y}) \tag{1}$$

Bem.: (i) Für $X = Y$ ist $\hat{\sigma}_{X,X} = \hat{\sigma}_X^2$

(ii) $\hat{\sigma}_{X,Y} > 0$ X und Y in der Stichprobe χ positiv korreliert

(iii) $\hat{\sigma}_{X,Y} < 0$ X und Y in der Stichprobe χ negativ korreliert

(iv) $\hat{\sigma}_{X,Y} = \hat{\sigma}_{Y,X}$

Der Korrelationskoeffizient

Die Stärke des Zusammenhanges zwischen X und Y erschließt sich nicht unmittelbar aus der Kovarianz.

Z.B. $x_i \to ax_i + b \Rightarrow \hat{\sigma}_{X,Y} = a\hat{\sigma}_{X,Y}$

$[\hat{\sigma}_{X,Y}] = [X] \times [Y]$, die Kovarianz ist nicht dimensionslos...

Def.: Der emp. Korrelationskoeffizient ist gegeben durch

$$\hat{r}_{X,Y} = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X \hat{\sigma}_Y} \tag{2}$$

($\hat{\sigma}_X, \hat{\sigma}_Y > 0$ emp. Standardabweichung von X und Y)

Falls $\hat{\sigma}_X = 0$ oder $\hat{\sigma}_Y = 0$ ist $\hat{r}_{X,Y}$ nicht definiert.

Eigenschaften des Korrelationskoeffizienten

Es gilt: (i) $\hat{r}_{X,Y}$ ist dimensionslos

(ii) $-1 \leq \hat{r}_{X,Y} \leq 1$

(iii) $|\hat{r}_{X,Y}| = 1 \Leftrightarrow$ es gibt a, b so dass $y_i = ax_i + b \; \forall i = 1, \ldots, n$

(iv) $x_i \to ax_i + b$ läßt $\hat{r}_{X,Y}$ unverändert ✓

Denn: (i) $[\hat{\sigma}_X] = [X]$ und $[\hat{\sigma}_Y] = [Y]$

(ii) Nach der Cauchy-Schwarz Ungleichung

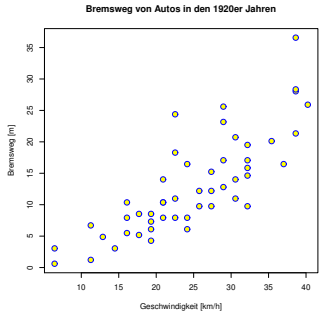
$$\left| \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \right| \leq \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{3}$$

Also $|(n-1)\hat{\sigma}_{X,Y}| \leq \sqrt{(n-1)}\hat{\sigma}_X \sqrt{(n-1)}\hat{\sigma}_y$ (iii) In (3) gilt $= \Leftrightarrow (y_i - \bar{y}) = a(x_i - \bar{x}) \Rightarrow$ setze $b = \bar{y} - a\bar{x}$

Beispiele für Korrelation

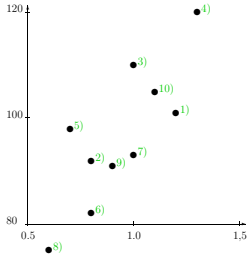
Korrelationskoeffizent Geschwindigkeit — Bremsweg

$$\hat{r}_{X,Y} = 0.807$$



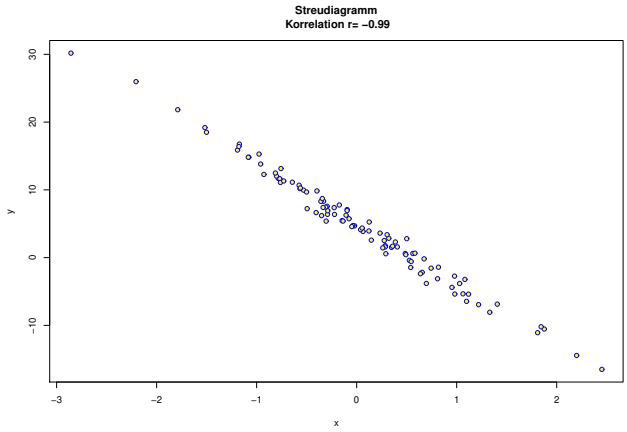
Korrelationskoeffizient Werbungskosten — Verkäufe

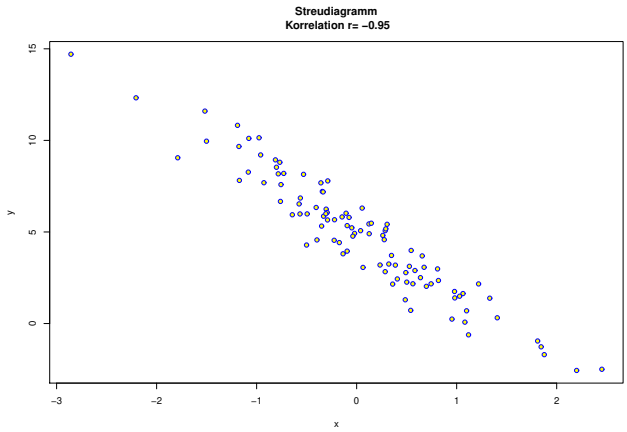
$$\hat{r}_{X,Y} = 0.875$$

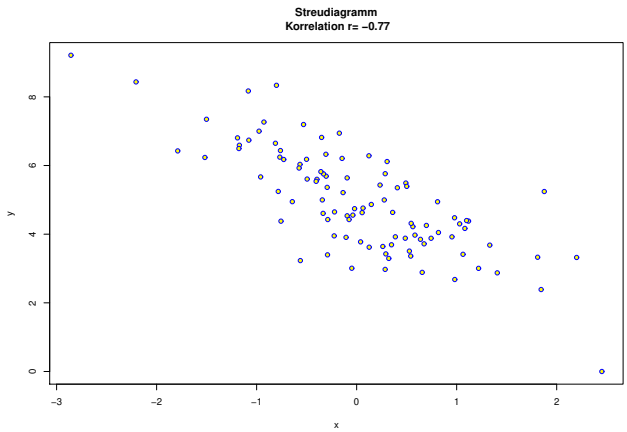


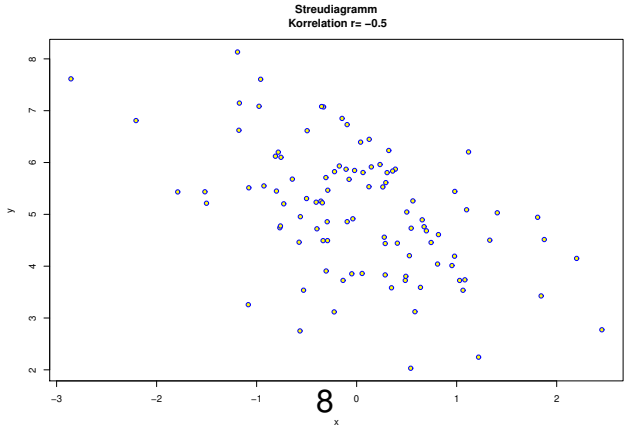
Einstufung der Korrelation

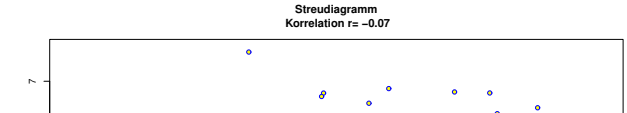
- $|\hat{r}_{X,Y}| \leq 0.3$ schwache oder keine Korrelation
- $0.3 < |\hat{r}_{X,Y}| \leq 0.9$ mittelstarke Korrelation
- $|\hat{r}_{X,Y}| > 0.9$ starke Korrelation











Kausale Abhängigkeiten

Auch der Korrelationskoeffizient ist symmetrisch (wie die Kovarianz)

$$\hat{r}_{X,Y} = \hat{r}_{Y,X}$$

Schon deshalb sagt der Korrelationskoeffizient nichts über kausale Abhängigkeit

Die kausalen Abhägigkeiten ergeben sich aus einer Interpretation, die per se nicht von den Daten her kommt.

- Mehr Werbung führt zu mehr Verkauf ✓
- Mehr Verkauf führt zu mehr Werbung **?**

Scheinkorrelation

Es gibt eine verborgene Ursache Z, so dass Zunahme von Z zur Zunahme von X und Zunahme (Abnahme) von Y führt

Dann erscheinen X und Y positiv (negativ) korreliert...

Würde man den Z- Effekt herausrechnen, wäre die Korrelation nicht vorhanden oder sogar umgekehrt!

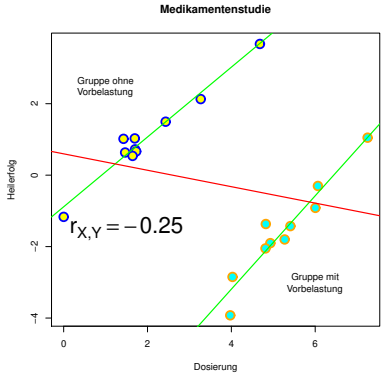
- Verkaufszahlen von Importware und lokaler Produktion sind positiv korreliert
- Je mehr Import, desto mehr lokale Produktion **?**
- Die verstekte Ursache Konjunktur treibt beide in dieselbe Richtung

Beispiel Scheinkorrelation

- Medizinische Studie
- Untersuchen Heilerfolg in Anhängigkeit von Dosierung
- Es resultiert eine negative Korrelation $\hat{r}_{\text{Dosis,Erfolg}} = -0.25$
- Medikament Mist **?**

Beispiel Scheinkorrelation II

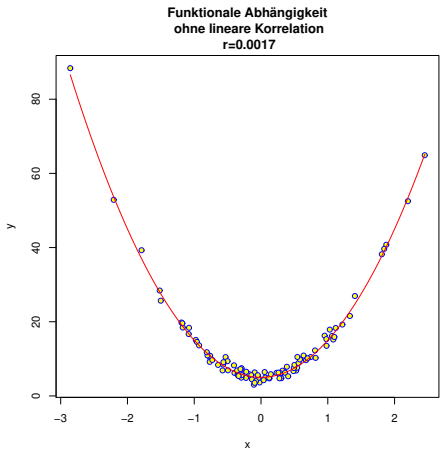
Blick auf die Daten — immer eine gute Idee:
'let the data speak'!



Der Gruppe mit Vorbelastung wurde eine höhere Dosis verabreicht - die Heilaussichten waren für diese Gruppe aber von vornherein geringer!

Abhängigkeit ohne lin. Korrelation

Merkmale X und Y können starke Abhängigkeit aufweisen, ohne linear korreliert zu sein!



Berechnung von Ausgleichsgraden (lin. Reg.) 20 / 32

Vorbemerkung - Modellierung I

Eine Perfekte funktionale Abhängigkeit von Y von X hat die Form

$$Y = f(X) \tag{4}$$

Oft verursachen jedoch noch weitere, nicht gemessene Faktoren $Z, U, V \ldots$ eine Veränderung von $Y \ldots$

Funktionales Modell der Statistik:

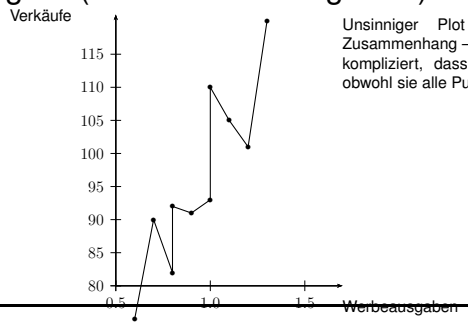
$$Y = f(X) + \text{statistische Schwankungen} \tag{5}$$
$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{6}$$

Def.: Die Werte $\epsilon_i = y_i - f(x_i)$ heißen die Residuen (oder Reststreuung) des funktionalen Modells f gegeben die Werte $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$

Vorbemerkung - Modellierung II

Aufgabe der Modellbildung:

- Auffinden einer *guten* Funktion f, die Abhängigkeit beschreibt
- einfach zu handhaben
- Erklärungskraft groß (u.a. Reststreuung klein)



Unsinniger Plot für funktionalen Zusammenhang – Die Funktion ist so kompliziert, dass sie nichts erklärt, obwohl sie alle Punkte genau trifft.

Overfitting

Bei verrauschten Daten, vermeide *overfitting*

overfitting liegt vor, wenn die funktionale Abhängigkeit stark von einzelnen Datenpunkten abhängt

Als Daumenegel: Der Datensatz sollte c.a. $10\times$ größer sein, als die Zahl der freien Parameter des Modells ...

Es sei denn, wir verwenden moderne 'Shrinkage' Methoden zur Modellanpassung (mit Bayes-Motivation)

Genaueres kann man nur mit Hilfe von Versuchsplan, Vorwissen über Rauschen in Y etc. sagen

Der Faktor 10 ist kein Dogma!

Lineare Abhängigkeit

Das einfachste Modell für die Abhängigkeit von X und Y ist die *lineare Abhängigkeit*

$$Y = \beta X + \alpha + \text{Reststreuung} \tag{7}$$

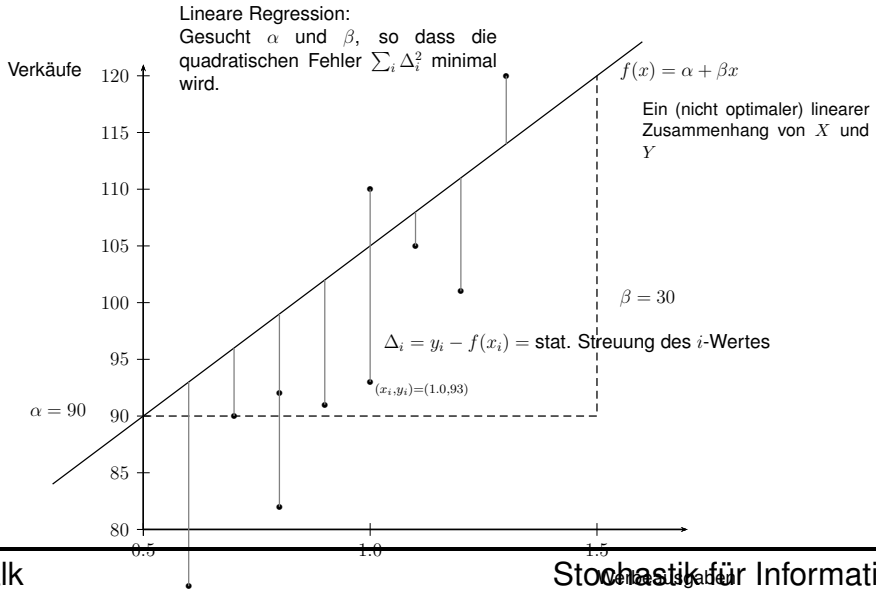
Die freien Parameter $\alpha, \beta \in \mathbb{R}$ werden so angepasst, dass die Reststreuung minimal wird.

- einfache Handhabung ✓
- gute Interpretierbarkeit von α und β ✓

Z.B.: α Verkäufe ohne Werbung, β Ertragssteigerung durch Werbeeinsatz

Wie finde ich die besten Werte $\hat{\alpha}, \hat{\beta}$ für die freien Parameter?

Kleinste Quadrate (least squares)



Lin. Reg. – Berechnung

Minimiere die Fehlerquadrate

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 \tag{8}$$

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\partial Q(\alpha, \beta)}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i)) \\ 0 &\stackrel{!}{=} \frac{\partial Q(\alpha, \beta)}{\partial \beta} \\ &= -2 \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i)) x_i = 2 \sum_{i=1}^{n} (\alpha x_i + \beta x_i^2 - x_i y_i) \end{aligned}$$

Teile beide Gleichungen durch $n \Rightarrow$

Lin. Reg. – Berechnung

Die Lösung $\hat{\alpha}, \hat{\beta}$ erfüllt also

$$\begin{aligned} 0 &= \hat{\alpha} + \hat{\beta} \bar{x} - \bar{y} \\ 0 &= \hat{\alpha} \bar{x} + \hat{\beta} \overline{x^2} - \overline{xy} \\ &= (\bar{y} - \hat{\beta} \bar{x}) \bar{x} + \hat{\beta} \overline{x^2} - \overline{xy} = \frac{n-1}{n} (\hat{\beta} \hat{\sigma}_X^2 - \hat{\sigma}_{X,Y}) \end{aligned}$$

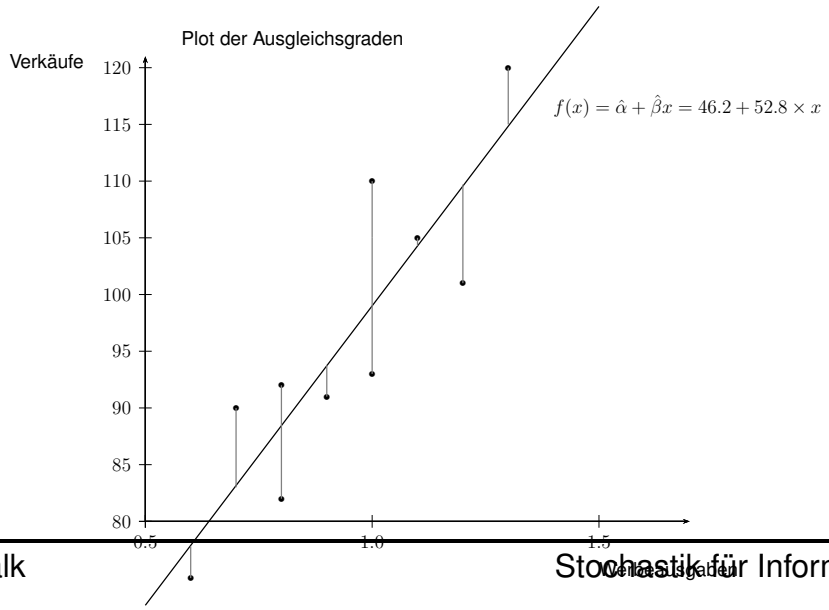
Satz: Die optimalen Koeffizienten $\hat{\alpha}$ und $\hat{\beta}$ (im Sinne der kleinsten Fehlerquadrate) sind gegeben durch

$$\hat{\beta} = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X^2} \tag{9}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \tag{10}$$

Bemerkung: $[\hat{\beta}] = [Y]/[X]$ und $[\hat{\alpha}] = [Y]$

Ausgleichsgrade Verkäufe



Effektiv Rechnen 29 / 32

Rechenbeispiel

Beispiel:

- Auf Bauernhof werden zwei Futtermittel A und B für Hühner
- A ist billiger und B ist teurer
- X der Anteil des teuren Futtermittels B
- Y Anzahl der gelegten Eier

Merkmal/Wert Nr.	1	2	3	4	5
X=Anteil von Futter B	0	1/4	1/2	3/4	1
Y=gelegte Eier	95	102	104	108	113

Rechnen mit Taschenrechner

	1	2	3	4	5	$\sum$	ø
X	0	1/4	1/2	3/4	1	2.5	0.5
Y	95	102	104	108	113	522	104.4
X^2	0	0.0625	0.25	0.5625	1	1.875	0.375
Y^2	9025	10404	10816	11664	12769	54678	10935.6
XY	0	25.5	52	81	113	271.5	54.3

a) $\bar{x} = 0.5$, $\bar{y} = 104.4$.

$s_X^2 = \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right) = \frac{5}{4}\left(0.375 - 0.5^2\right) = 0.15625$

$s_Y^2 = \frac{n}{n-1}\left(\overline{y^2} - \bar{y}^2\right) = \frac{5}{4}\left(10935.6 - 104.4^2\right) = 45.3$

b) $\mathrm{Cov}_{X,Y} = \frac{n}{n-1}\left(\overline{xy} - \bar{x}\bar{y}\right) = \frac{5}{4}\left(54.3 - 0.5 \times 104.4\right) = 2.625$

c) $\hat{\beta} = \frac{\mathrm{Cov}_{X,Y}}{s_X^2} = \frac{2.625}{0.15625} = 16.8$, daher $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 104.4 - 16.8 \times 0.5 = 96 \Rightarrow f(x) = 96 + 16.8 \times x$.

R – wie Rechnen

Mit R ist lineare Regression ein Kinderspiel!

```
R> ?cars # Daten in R vorhanden - US Einheiten!
R> myModel=lm(data=cars,dist~speed)
                    # lm für linear model
R> myModel$coefficients # gefittete Werte
R> plot(cars) # der Streuplot
R> abline(myModel) # die Ausgleichsgrade
```