9. Regressionsanalyse

- Einfache lineare Regression
- Multiple lineare Regression
- Hypothesentests in der linearen Regression
  - Varianzzerlegung und Bestimmtheitsmaß
  - F-Test auf Modellgüte
  - F-Test auf Einfluss von Parametergruppen
  - Tests für die Regressionskoeffizienten
  - Konfidenz- und Prognoseintervalle
- Überprüfung der Modellannahmen
  - Motivation
  - Residuenanalyse
- Indikatorvariablen

9 Regressionsanalyse [1,4,5]

Ziel: Erklärung/Prognose von Variablen durch andere Variablen

- **Regressand** Y: zu erklärende, abhängige Zufallsvariable
- **Regressoren** $X_1, \ldots, X_k$: erklärende, unabhängige Zufallsvariablen

Beispiele

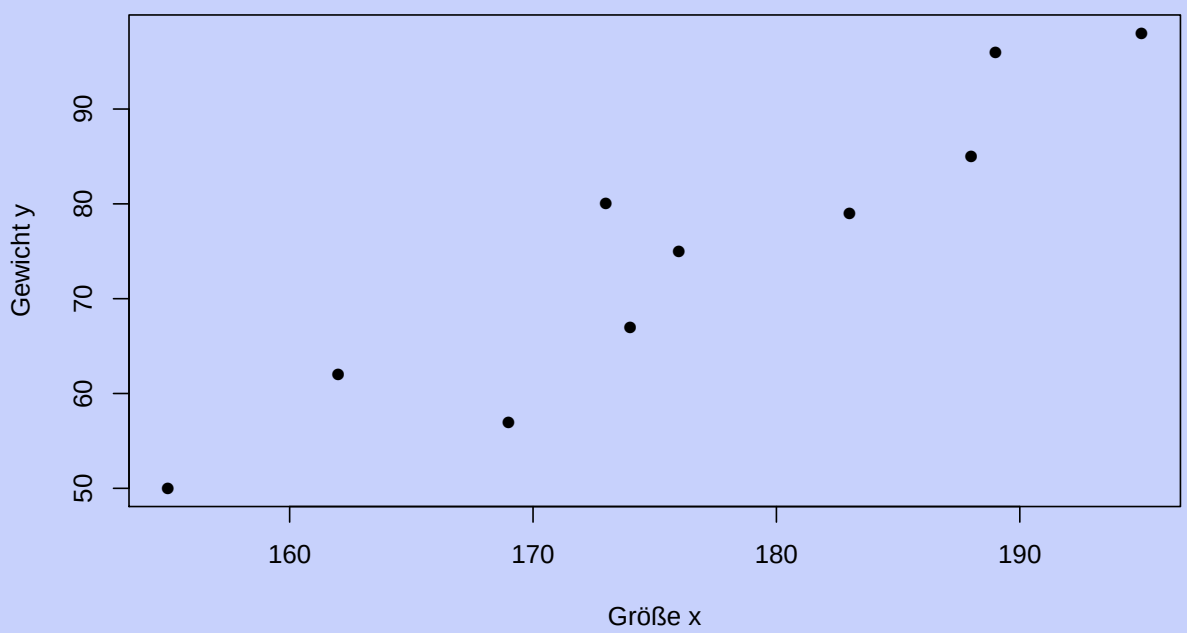
- ↪ Mietspiegel: Y=qm-Kaltmiete, X_1=Wohnfläche, X_2= Lage, X_3=Baujahr,...
- ↪ Gebrauchtwagen: Y =Marktwert, X_1 =Alter, X_2 =TKM, X_3 =Zustand,...
- ↪ Paket-Service: Y =Tagesfahrtzeit eines Wagens, X_1=Zahl der Anlieferungen, X_2=Routenlänge, X_3=Zielregion...
- ↪ Kunden-Umsätze: Y=Wert der Aufträge, X_1=Einkommen, X_2=Familienstand, X_3=Berufsgruppe, X_4=Anzahl der Aufträge,...

In der Praxis liegen Datensätze für diese Variablen/Merkmale vor.

Beispiel 1: Größe und Gewicht

Von 10 Personen wurden Körpergröße x und Körpergewicht y gemessen:

	A	B	C
1	Person	x	y
2	1	162	62
3	2	173	80
4	3	169	57
5	4	189	96
6	5	176	75
7	6	188	85
8	7	155	50
9	8	174	67
10	9	183	79
11	10	195	98



Lässt sich ein rechnerischer Zusammenhang zwischen x,y finden, der die Daten „gut“ beschreibt? Wie lässt sich das Gewicht einer Person „vorhersagen“, welche 181 cm groß ist?

Beispiel 2: Lieferzeitenanalyse (aus [4])

Ein Unternehmen füllt Getränke-/Süßigkeitenautomaten nach. In der Vergangenheit wurden einige Liefervorgänge protokolliert:

Y: Lieferzeit (Minuten), `delTime`

X_1: Anzahl nachzufüllender Produkte, `n.prod`

X_2: Fußweg (ft), `distance`

	n.prod	distance	delTime		n.prod	distance	delTime		n.prod	distance	delTime
1	7	560	16.68	9	30	1460	79.24	17	6	200	15.35
2	3	220	11.50	10	5	605	21.50	18	7	132	19.00
3	3	340	12.03	11	16	688	40.33	19	3	36	9.50
4	4	80	14.88	12	10	215	21.00	20	17	770	35.10
5	6	150	13.75	13	4	255	13.50	21	10	140	17.90
6	7	330	18.11	14	6	462	19.75	22	26	810	52.32
7	2	110	8.00	15	9	448	24.00	23	9	450	18.75
8	7	210	17.83	16	10	776	29.00	24	8	635	19.83
								25	4	150	10.75

Für die Tourenplanung wird eine Formel gesucht, mit der die Lieferzeit möglichst gut anhand eines konkreten Auftrags prognostiziert werden kann.

Datenquelle in R: `library(robustbase),data(delivery)`

Beispiel 3: Lebensdauer eines Werkzeugs (aus [5])

Für 20 Werkzeuge wurden gemessen:

- Y: Lebensdauer des Schneidwerkzeuges
- X_1: Umdrehungen pro Minute
- X_2: Werkzeugtyp

	y	x1	x2		y	x1	x2
1	18.73	610	A	11	30.16	670	B
2	14.52	950	A	12	27.09	770	B
3	17.43	720	A	13	25.40	880	B
4	14.54	840	A	14	26.05	1000	B
5	13.44	980	A	15	33.49	760	B
6	24.39	530	A	16	35.62	590	B
7	13.34	680	A	17	26.07	910	B
8	22.71	540	A	18	36.78	650	B
9	12.68	890	A	19	34.95	810	B
10	19.32	730	A	20	43.67	500	B

Wie wirken sich Drehzahl und Werkzeugtyp auf die Lebensdauer aus?

Ziel: Erklärung/Prognose von Variablen durch andere Variablen

- **Regressand** Y: zu erklärende, abhängige Zufallsvariable
- **Regressoren** $X_1, \ldots, X_k$: erklärende, unabhängige Zufallsvariablen

$\hookrightarrow$ Kausalmodell $Y = f(X_1, \ldots, X_n)$:

- Dabei liegt f in vorgegebener „passender“ Funktionsklasse $\mathcal{F}$, z.B. $f(x_1, \ldots, x_k) = a_0 + a_1 x_1 + \cdots + a_k x_k$
- Dieser Idealfall ist bei stochastischen Daten unrealistisch.

$\hookrightarrow$ Annahme: „gestörtes“ Kausalmodell $Y = f(X_1, \ldots, X_k) + \epsilon$ (f, ϵ unbekannt).

$\hookrightarrow$ Ziel: Finde „optimales“ f in Funktionsklasse $\mathcal{F}$.

$\hookrightarrow$ Lösung mit der KQ-Methode:

- theoretisch: $E((Y - f(X_1, \ldots, X_k))^2) \stackrel{!}{=} \min\limits_{f \in \mathcal{F}}$
- empirisch: $\sum\limits_{i=1}^{n} (y_i - f(x_{i1}, \ldots, x_{ik}))^2 \stackrel{!}{=} \min\limits_{f \in \mathcal{F}}$ mit Daten x_{ij}, y_i

Memo: (Regeln für Erwartungswerte, Varianzen und Kovarianzen)

$$
\begin{aligned}
(*) \quad & E(X^2) = var(X) + (E(X))^2 \\
(**) \quad & var(X + c) = var(X) \text{ für } c \in \mathbb{R} \\
(***) \quad & var(X + Y) = var(X) + 2cov(X, Y) + var(Y)
\end{aligned}
$$

Beispiel: Minimierungsproblem für lineare Funktionen $f(x) = ax + b$

Für bekannte $\mu_1 = E(X), \mu_2 = E(Y), \sigma_1^2 = var(X), \sigma_2^2 = var(Y), \rho = cor(X, Y)$:

$$
\begin{aligned}
E(Y - (aX + b))^2) &\overset{(*)}{=} var(Y - (aX + b)) + (E(Y - (aX + b))^2 \\
&\overset{(**)}{=} var(Y - aX) + (E(Y) - aE(Y) - b))^2 \\
&\overset{(***)}{=} \underbrace{\sigma_2^2 - 2a\, cov(X, Y) + a^2\sigma_1^2}_{\text{min. für } a = \frac{cov(X,Y)}{\sigma_1^2} = \frac{\sigma_2}{\sigma_1}\rho} + \underbrace{(\mu_2 - a\mu_1 - b)^2}_{\text{min. für } b = \mu_2 - a\mu_1}
\end{aligned}
$$

d.h. beste lin. Vorhersage ist $f(X) = \frac{\sigma_2}{\sigma_1}\rho X + (\mu_2 - \frac{\sigma_2}{\sigma_1}\rho\mu_1) = \mu_2 + \frac{\sigma_2}{\sigma_1}\rho(X - \mu_1)$

Ohne Vorgaben an f: Problem wird durch bedingten Erwartungswert gelöst.

Übung: Löse das Minimierungsproblem $E((Y - f(X))^2) \stackrel{!}{=} \min$ für

1. $f(x) = a$, mit $a \in \mathbb{R}$. Minimiere hierzu $E((Y - a)^2) = \ldots$ in a.

$$\begin{aligned} E((Y-a)^2) &= E(Y^2 - 2aY + a^2) \\ &= E(Y^2) - 2aE(Y) + a^2 \end{aligned}$$

Ein Minimum dieses Terms ergibt sich durch Ableiten nach a und Nullsetzen: $-2E(Y)+2a = 0 \Leftrightarrow a = EY$. Der Erwartungswert EY ist also die beste Konstante, um Y zu erklären.

2. $f(x) = ax$ mit $a \in \mathbb{R}$. Minimiere hierzu $E((Y - aX)^2) = \ldots$ in a

$$\begin{aligned} E((Y-aX)^2) &= E(Y^2 - 2aXY + a^2X^2) \\ &= E(Y^2) - 2aE(XY) + a^2E(X^2) \end{aligned}$$

Ein Minimum dieses Terms ergibt sich durch Ableiten nach a und Nullsetzen: $-2E(XY) + 2aE(X^2) = 0 \Leftrightarrow a = \frac{E(XY)}{E(X^2)}$

Memo: Bedingte Verteilung/Bedingter Erwartungswert

↪ Hier bivariater Fall: Zufallsvektor (X, Y) mit

- diskreter/stetiger Dichte $f_{X,Y}(x, y)$
- Randdichten $f_X(x)$, $f_Y(y)$.

↪ **Bedingte Verteilung** $\mathcal{L}(Y|X = x)$: ist gegeben durch **bedingte Dichte**:

$$f_{Y|X=x}(y) = \begin{cases} f_{X,Y}(x, y)/f_X(x) & \text{falls Nenner} > 0 \\ f_Y(y) & \text{sonst} \end{cases}$$

(Bayes-Formel) „bedingte WS in y bilden bei festem x eine WS-Verteilung“.

↪ **Bedingter Erwartungswert**: $f(x) = E(Y|X = x)$:

- Erwartungswert der bedingten Verteilung $\mathcal{L}(Y|X = x)$.
- Schreibweise für $f(X)$: $E(Y|X) = f(X)$
- Der bed. EW löst das (obige) Problem $E((Y - f(X))^2) \stackrel{!}{=} \min$.

↪ Konzept übertragbar auf Zufallsvektoren:
$\mathcal{L}(Y_1, \ldots, Y_m|X_1 = x_1, \ldots, X_k = x_k)$ bzw. $E(Y|X_1 = x_1, \ldots, X_n = x_n)$

Beispiel (DuW): Bivariate Normalverteilung

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right)}$$

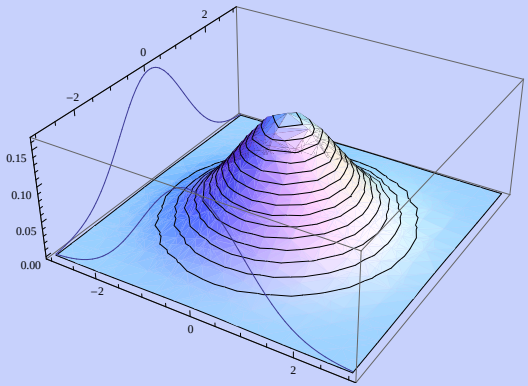
mit $\mu_1, \mu_2 \in \mathbb{R}, \sigma_1, \sigma_2 > 0$, $-1 < \rho < 1$.

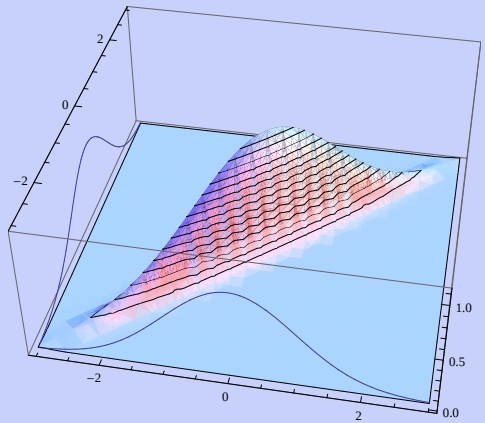
$\hookrightarrow$ Randverteilungen sind $\mathcal{L}(X) = \mathcal{N}(\mu_1, \sigma_1^2)$ und $\mathcal{L}(Y) = \mathcal{N}(\mu_2, \sigma_2^2)$

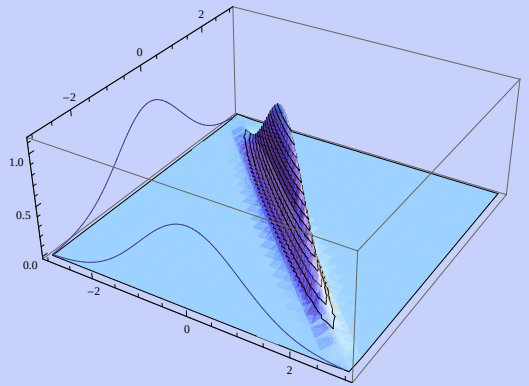
$\hookrightarrow$ ρ ist Pearson-Korrelation von X, Y.

Illustration: `https://www.geogebra.org/m/whgyqhmf`

Für $\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1$ und $\rho = 0 / \rho = 0.99 / \rho = -0.99$







Jeweils: $f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right)$ und $\mathcal{N}(0,1)$-Randverteilungen.

↪ Betrachte speziell $\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1$.Bedingte Dichte ist

$$\begin{aligned} f_{Y|X=x}(y) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right)\right) / \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right) - \frac{x^2}{2}\right) \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2 - (1-\rho^2)x^2\right)\right) \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(y^2 - 2\rho yx + \rho^2 x^2\right)\right) \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right) \end{aligned}$$

d.h. $\mathcal{L}(Y|X = x) = \mathcal{N}(\rho x, 1 - \rho^2)$ und $E(Y|X = x) = \rho x$

↪ Für allgemeine μ_i, σ_i^2:

- $\mathcal{L}(Y|X = x) = \mathcal{N}(\mu_2 + \frac{\sigma_2}{\sigma_1} \cdot \rho(x - \mu_1), \sigma_2^2(1 - \rho^2))$
- $E(Y|X = x) = \mu_2 + \frac{\sigma_2}{\sigma_1} \cdot \rho(x - \mu_1)$

„In (multivariaten) Normalverteilungsmodellen ist die beste Prognose linear.“

Memo: Regeln für bedingte Erwartungswerte

↪ Linearität: $E(a + bY_1 + cY_2|X) = a + bE(Y_1|X) + cE(Y_2|X)$

↪ Totale Wahrscheinlichkeit: $E(Y) = E(E(Y|X))$

↪ Faktorisierung: $E(Y \cdot g(X)|X) = g(X) \cdot E(Y|X)$

↪ Substituieren/Eliminieren (SE): Wenn X, Y st.u. sind, dann gilt:

(SE1) $P(h(X, Y) \in B|X = x)) = P(h(x, Y) \in B)$

(SE2) $E(h(X, Y)|X = x) = E(h(x, Y))$

↪ Varianzzerlegung: Mit $var(Y|X = x) = E((Y - E(Y|X = x))^2|X = x) = E(Y^2|X = x) - (E(Y|X = x))^2$ gilt

$$var(Y) = var(E(Y|X)) + E(var(Y|X))$$

- linker Summand: Variabilität in Y erklärt durch X
- rechter Summand: Nicht durch X erklärte Variabilität.

Aussagen jeweils „fast sicher“ und unter Annahme existierender Erwartungswerte

Allgemeines Regressionsmodell mit Fehlerterm

- $\hookrightarrow$ $Y = f(X_1, \ldots, X_k) + \sigma V$, $f \in \mathcal{F}$ (vorgegebene Klasse von Funktionen)
- $\hookrightarrow$ σV: nicht beobachtbarer - Fehler mit $E(V) = 0$, $var(V) = 1$, σ unbekannt.
- $\hookrightarrow$ Grundannahme an den Fehler: $cor(f(X_1, \ldots, X_k), V) = 0$. Damit lassen sich KQ-Schätzer für f, σ gewinnen, aber (noch) keine Bereichsschätzer/Tests.
- $\hookrightarrow$ Stärkere Grundannahme $(X_1, \ldots, X_k)$ und V sind st.u.
  Dann wird $\mathcal{L}(Y|X_1, \ldots, X_k)$ durch $f \in \mathcal{F}$, $\mathcal{L}(V)$ und σ bestimmt:

$$P(Y \leq y | X_1 = x_1, \ldots, X_k = x_k) \overset{(SE1)}{=} P(f(x_1, \ldots, x_k) + \sigma V \leq y)$$

$$= P(V \leq \frac{y - f(x_1, \ldots, x_k)}{\sigma})$$

- $\hookrightarrow$ $\mathcal{L}(X_1, \ldots, X_k)$ und damit $\mathcal{L}(Y, X_1, \ldots, X_k)$ bleiben i.d.R. unspezifiziert. Aussagen (Tests, Schätzer, Bereichsschätzer) werden fast ausschließlich bedingt an den Werten $X_1 = x_1, \ldots, X_k = x_k$ getroffen.
- $\hookrightarrow$ Fast ausschließlich verwendet: klassisches Regressionsmodell: $\mathcal{L}(V) = \mathcal{N}(0, 1)$.

Regressionsmodell zu einer Stichprobe

- ↪ Es liegt eine u.i.v.-Stichprobe $(Y_i, X_{i1}, \ldots, X_{ik})$, $i = 1, \ldots, n$, vor.
- ↪ $Y_i = f(X_{i1}, \ldots, X_{ik}) + \sigma V_i$ mit unbekannter Funktion $f \in \mathcal{F}$, d.h.
  - □ bekannter „Typ“ der Regressionsfunktion, keine vollständige Spezifikation.
  - □ nicht beobachtbarer Fehler $\epsilon_i = \sigma V_i$
- ↪ $E(\epsilon_i)$=0, $var(\epsilon_i) = \sigma^2$ (Homoskedastizität)
- ↪ ggf. Normalverteilungsannahme für ϵ_i.

Statistische Aufgaben im Regressionsmodell

- ↪ Schätzung von f, σ^2, bei parametrischem Typ auch Konfidenzintervalle.
- ↪ Schätzung durch KQ/OLS-Methode: $\sum_{i=1}^{n}(Y_i - f(X_{i1}, \ldots, X_{ik}))^2 \stackrel{!}{=} \min_{f \in \mathcal{F}}$
- ↪ Unter Normalverteilungsannahme:
  - □ Schätzung durch ML-Methode
  - □ Prüfung von Hypothesen zu f.

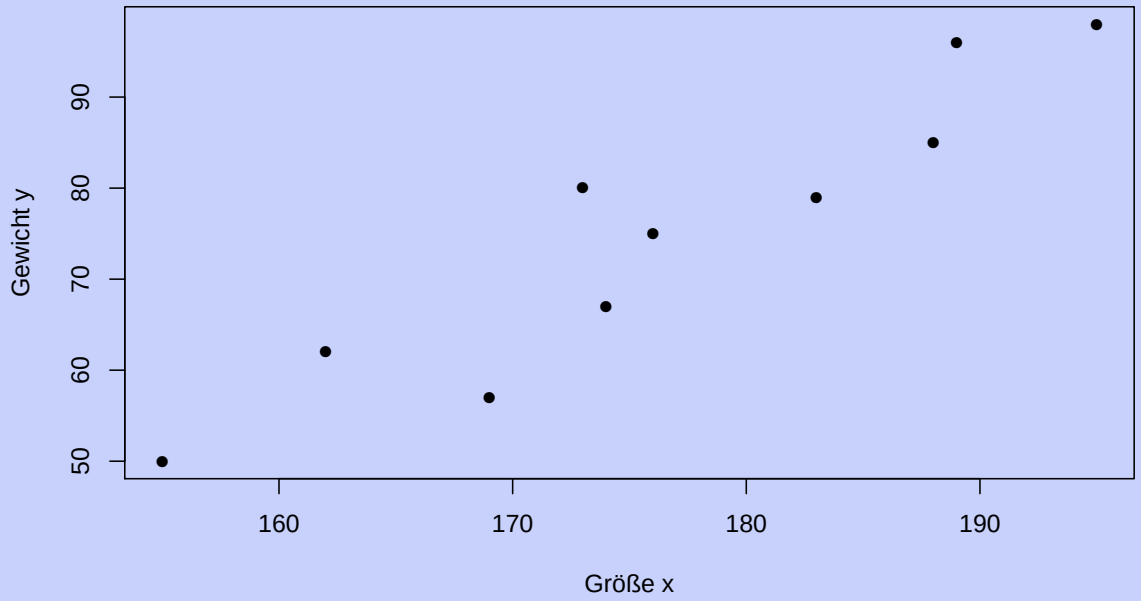
9.1 Einfache lineare Regression

Ein Regressor, f linear, $f(x) = \beta_0 + \beta_1 x$ mit $\beta_0, \beta_1 \in \mathbb{R}$

$\hookrightarrow$ $Y = \beta_0 + \beta_1 X + \epsilon$ mit st.u. ϵ, X

- $E(Y|X = x) = \beta_0 + \beta_1 x$,
- $var(Y|X = x) = var(\beta_0 + \beta_1 x + \epsilon) = var(\epsilon) = \sigma^2$

Größe und Gewicht



	A	B	C	D	E
1	beta	30	0,1		
2	Person	const	x	y	ydach
3	1	1	162	62	46,2
4	2	1	173	80	47,3
5	3	1	169	57	46,9
6	4	1	189	96	48,9
7	5	1	176	75	47,6
8	6	1	188	85	48,8
9	7	1	155	50	45,5
10	8	1	174	67	47,4
11	9	1	183	79	48,3
12	10	1	195	98	49,5

Gesucht: Eine Funktion $f(x) = \beta_0 + \beta_1 x$, welche die Daten möglichst gut beschreibt.

↪ Schätzung von β_0, β_1 anhand u.i.v.-Stichprobe $(x_1, y_1), \ldots, (x_n, y_n)$. Dabei mindestens zwei verschiedene x_i angenommen.

↪ **KQ-Methode**: $K(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \stackrel{!}{=} \min_{\beta_0, \beta_1}$. Notwendig: $\nabla K = \bar{0}$

$$\left.\begin{aligned} \text{(I)} \quad & \frac{\partial K}{\partial \beta_0} = -2n(\bar{y} - \beta_0 - \beta_1 \bar{x}) \stackrel{!}{=} 0 \\ \text{(II)} \quad & \frac{\partial K}{\partial \beta_1} = -2\Big(\sum_{i=1}^{n} x_i y_i - n\beta_0 \bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2\Big) \stackrel{!}{=} 0 \end{aligned}\right\} \quad \text{(Normalgleichungen)}$$

↪ Auflösen von (I) nach β_0 und Einsetzen in (II)

$$\begin{aligned} 0 &= \sum_{i=1}^{n} x_i y_i - n(\bar{y} - \beta_1 \bar{x})\bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - \beta_1 \Big(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\Big) \end{aligned} \quad \Rightarrow \quad \begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{cov(x,y)}{s_x^2} = \frac{s_y}{s_x} r_{xy} \end{aligned}$$

↪ Prognostizierte Werte: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

↪ $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$, $i = 1, \ldots, n$ heißen **Residuen**. Nutzen:

- Erwartungstreue KQ-Schätzung von σ^2 durch $\frac{1}{n-2} \sum e_i^2$
- Überprüfung der Modellgüte durch $R^2 = 1 - \sum e_i^2 / \sum (y_i - \bar{y})^2$.

	x	y	x^2	xy	y^2	$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$	$e = y - \hat{y}$	e^2
1	162	62	26244	10044	3844	57.74	4.26	18.13
2	173	80	29929	13840	6400	70.85	9.15	83.74
3	169	57	28561	9633	3249	66.08	-9.08	82.50
4	189	96	35721	18144	9216	89.91	6.09	37.05
5	176	75	30976	13200	5625	74.42	0.58	0.33
6	188	85	35344	15980	7225	88.72	-3.72	13.85
7	155	50	24025	7750	2500	49.40	0.60	0.36
8	174	67	30276	11658	4489	72.04	-5.04	25.41
9	183	79	33489	14457	6241	82.76	-3.76	14.17
10	195	98	38025	19110	9604	97.06	0.94	0.88
Σ	1764	749	312590	133816	58393	749.00	0.00	$SS_{Res} = 276.41$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{133816 - 10 \times 176.4 \times 74.9}{312590 - 10 \times 176.4^2} = \frac{1692.4}{1420.4} \approx 1.1915$$

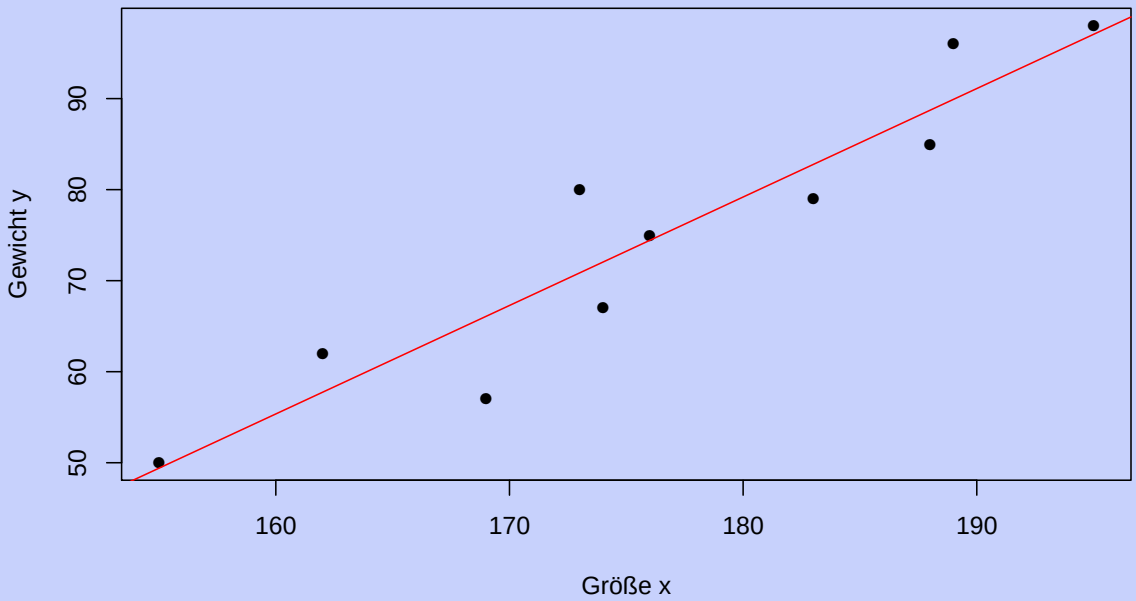
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 74.9 - 1.1915 \cdot 176.4 \approx -135.2798$$

$$SS_T = \Sigma y_i^2 - n\bar{y}^2 = 58393 - 10 \times 74.9^2 = 2292.9$$

$$R^2 = 1 - SS_{Res}/SS_T = 1 - \frac{276.41}{2292.9} \approx 0.879$$

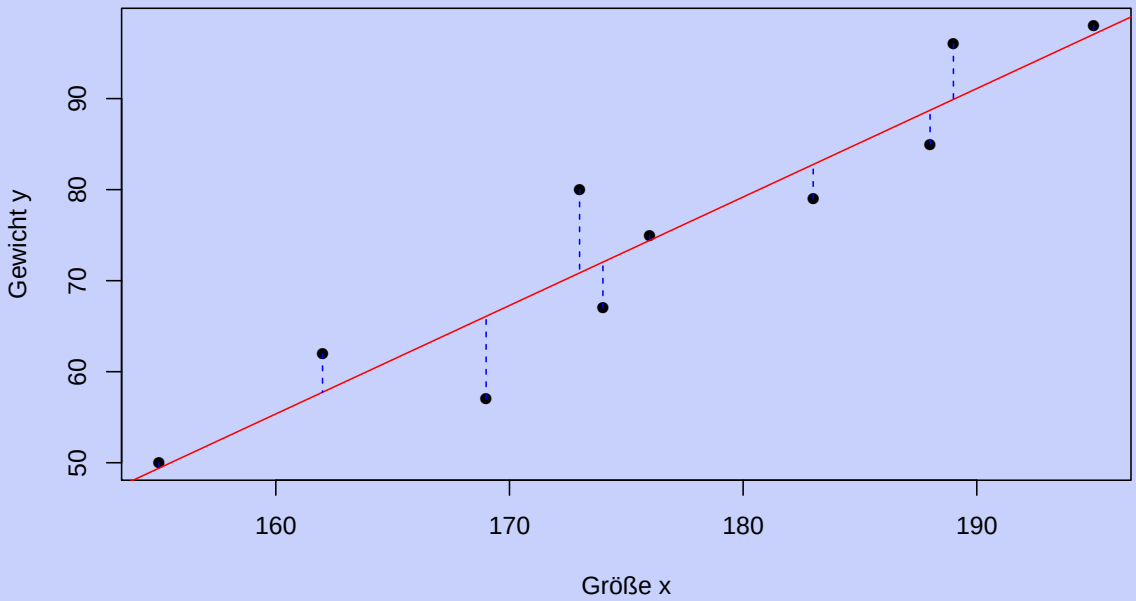
R^2 besagt, dass sich etwa 87,9% der Gewichts-Variabilität durch die (unterschiedlichen) Größen erklären lassen, dazu später mehr.

Größe: $x =$	162	173	169	189	176	188	155	174	183	195
Gewicht: $y =$	62	80	57	96	75	85	50	67	79	98



$$\hat{y} = -135.28 + 1.19 \cdot x$$

Größe: $x =$	162	173	169	189	176	188	155	174	183	195
Gewicht: $y =$	62	80	57	96	75	85	50	67	79	98



$$\hat{y} = -135.28 + 1.19 \cdot x, e = y - \hat{y}$$

Umsetzung in R:

```
persdata=data.frame(
  x=c(162,173,169,189,176,188,155,174,183,195),
  y=c(62,80,57,96,75,85,50,67,79,98)
)
#pdf(file="LM_Bsp3.pdf",width=8,height=5)
lm.persdata<-lm(y~x,data=persdata) #die Funktion zur Regression.
    Standardausgabe: Koeffizienten
plot(persdata,xlab="Groesse x",ylab="Gewicht y",main=NULL,pch=16)
abline(lm.persdata,col="red") # setzt die Koeffizienten in die
    Regressionsgerade um

sapply(1:NROW(persdata),function(i){lines(rep(persdata$x[i],2),c(
    persdata$y[i],predict(lm.persdata)[i]),col="blue",lty="dashed")
    }) # zeichnet die Residuen ein
#dev.off()
```

Übung: Der Inhaber einer Kette von 5 freien Tankstellen möchte wissen, ob und wie sich der Tagesgewinn aus den Kraftstoffumsätzen erklären lässt. An einem Tag beobachtet er folgende Werte (in €):

Ums. K. x	6000	2500	8500	6500	9500
Gewinn y	3000	4000	2000	3000	3500

Führe eine einfache lineare Regression durch.

	x	y	x^2	xy	y^2	$\hat{y}$	$e = y - \hat{y}$	e^2
1	6000	3000	36000000	18000000	9000000	3193.49	-193.49	37439.60
2	2500	4000	6250000	10000000	16000000	3738.87	261.13	68188.95
3	8500	2000	72250000	17000000	4000000	2803.94	-803.94	646316.88
4	6500	3000	42250000	19500000	9000000	3115.58	-115.58	13359.24
5	9500	3500	90250000	33250000	12250000	2648.12	851.88	725705.60
Σ	33000	15500	247000000	97750000	50250000	15500.00	0.00	1491010.27

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{97750000 - 5 \times 6600 \times 3100}{247000000 - 5 \times 6600^2} = \frac{-4550000}{29200000} \approx -0,1558$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3100 + 0,1558 \times 6600 = 4128.425$$

$$\hat{y} = -0.1558x + 4128.425$$

$$SS_T = \Sigma y^2 - n\bar{y}^2 = 50250000 - 5 \times 3100^2 = 2200000$$

$$R^2 = 1 - SS_{Res}/SS_T = 1 - 1491010.27/2200000 \approx 0,322$$

Der Wert des Bestimmtheitsmaßes besagt, dass sich im bestimmten Modell nur etwa 32,2% der Variabilität in den Gewinndaten durch den Kraftstoffumsatz erklären lassen. Das spricht im vorliegenden Fall eher gegen eine Brauchbarkeit des einfachen linearen Modells.

↪ Vergleich der empirischen und theoretischen Lösungen

- empirisch: $\sum_{i=1}^{n}(y_i - (ax_i + b))^2 \stackrel{!}{=} \min_{a,b}$

$$\begin{aligned} a &= \frac{s_y}{s_x} r_{xy} \\ b &= \bar{y} - a\bar{x} \end{aligned}$$

- theoretisch (s.o.): $E((Y - (aX + b))^2) \stackrel{!}{=} \min_{a,b}$

$$\begin{aligned} a &= \frac{\sigma_Y}{\sigma_X} \rho_{XY} \\ b &= E(Y) - aE(X) \end{aligned}$$

- Unter Normalverteilungsannahme (s.o.) $E((Y - f(X))^2) \stackrel{!}{=} \min_{f}$

$f(x) = E(Y|X = x) = ax + b$ mit

$$\begin{aligned} a &= \frac{\sigma_Y}{\sigma_X} \rho_{XY} \\ b &= E(Y) - aE(X) \end{aligned}$$

↪ In allen Fällen ergibt sich dieselbe Lösung (im empirischen Fall mit empirischen Kennzahlen zu Lage, Streuung, Zusammenhang).

↪ Die empirische Lösung ergibt sich aus der theoretischen, wenn man als Modell die Stichprobenverteilung der y_i, x_i annimmt.

Übung: Oben wurde die Funktion $K(\beta_0, \beta_1) = \sum(y_i - \beta_0 - \beta_1 x_i)^2$ minimiert. Dabei war (I) $\frac{\partial K}{\partial \beta_0} = -2n(\bar{y} - \beta_0 - \beta_1 \bar{x})$ (II) $\frac{\partial K}{\partial \beta_1} = -2(\sum x_i y_i - n\beta_0 \bar{x} - \beta_1 \sum x_i^2)$
Prüfe mittels Hesse-Matrix, ob die berechnete Lösung minimal ist.

$$H_K(\beta_0, \beta_1) = \begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum x_i^2 \end{pmatrix}$$

hat die Hauptminoren

- $2n > 0$
- $4n\sum x_i^2 - 4n^2\bar{x}^2 = 4n(\sum x_i^2 - n\bar{x}^2) = 4n\sum(x_i - \bar{x})^2 \geq 0$

Nach dem erweiterten Determinantenkriterium für 2×2-Matrizen ist H_K stets positiv semidefinit, also ist K konvex. Der kritische Punkt ist deshalb Stelle eines globalen Minimums.