

Lineare Regression (1)

- Einführung I -

Mittels Regressionsanalysen und komplexeren, auf Regressionsanalysen basierenden Verfahren können scheinbar verschiedene, jedoch ineinander überführbare Fragen untersucht werden:

- Wie eng ist und welche Richtung hat der **Zusammenhang** zwischen zwei Variablen?
- **Unterscheiden** sich die Mittelwerte einer Variablen in verschiedenen (Teil-) Gruppen?
- Wie gut lassen sich die Werte einer Variablen durch eine oder mehrere andere Variablen **vorhersagen** (d.h. auf die Werte einer oder mehrerer anderer Variablen zurückführen)?
- Mit welchen Modellen können **Zusammenhänge zwischen mehreren Variablen** am besten beschrieben werden?
- Wie gut **passt ein theoretisches Modell** des Zusammenhangs zweier oder mehrerer Variablen zu den Daten?

Lineare Regression (2)

- Einführung II -

Die grundlegenden Prinzipien der Regressionsanalyse lassen sich anhand der im Folgenden betrachteten **einfachen linearen Regression** darstellen.

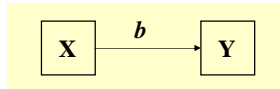
Bei der einfachen linearen Regression

- kennt man den Ausprägungsgrad einer Variablen (z.B. X) und möchte den Ausprägungsgrad einer anderen Variable (z.B. Y) vorhersagen.
- Die Variable, die vorhergesagt werden soll, wird als **abhängige Variable (AV)** oder als **Kriterium** bezeichnet (üblicherweise Y).
z.B. Gewalteinrichtungen, Erfolg von Interventionsmaßnahmen, beruflicher Erfolg
- Die Variable, die zur Vorhersage dient, wird als **unabhängige Variable (UV)** oder als **Prädiktor** bezeichnet (üblicherweise X).
z.B. Gewalterfahrung im Elternhaus, Einbindung in Peergroups, Arbeitsmotivation
- Es soll diejenige **lineare** Funktion gefunden werden, die den Zusammenhang zwischen X und Y optimal beschreibt.

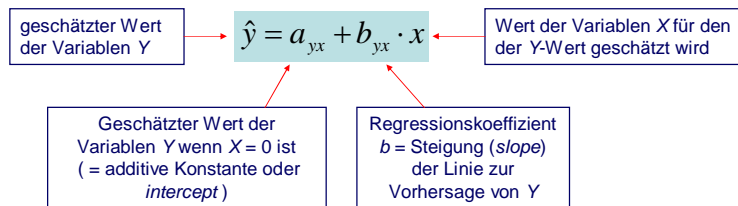
Lineare Regression (3)

- Einführung III -

Das **konzeptuelle Modell** zur Vorhersage einer abhängigen Variablen Y durch eine unabhängige Variable X ist:



Die einfachste mathematische Funktion zur Vorhersage ist eine **lineare Gleichung**, in der die X-Werte zur Schätzung der Y-Werte linear transformiert werden:

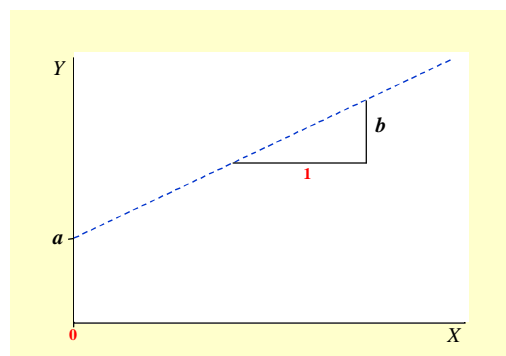


Die Reihenfolge der Indizes „yx“ an den Koeffizienten a und b gibt an, dass sie der Vorhersage der Y-Werte anhand der X-Werte dienen: Es handelt sich hier um eine **Regression von Y auf X**.

Lineare Regression (4)

- Einführung IV -

Der Wert b_{yx} bedeutet inhaltlich die durchschnittliche Veränderungsrate der Y-Werte pro Zunahme einer Einheit von X-Werten, während der Wert a_{yx} denjenigen Y-Wert angibt, an dem die Regressionslinie die y-Achse schneidet:



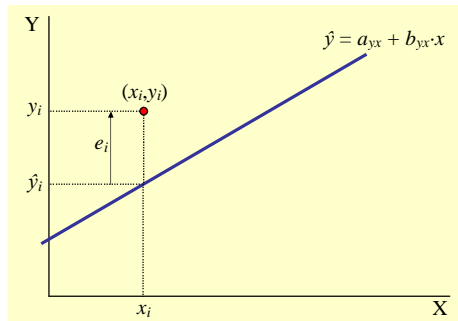
Regressionsgerade mit Steigung b und additiver Konstanter a

Lineare Regression (5)

- Einführung V -

Vorhersage eines Y-Wertes anhand der Regressionsgeraden:

Für einen gegebenen Wert x_i wird parallel zur Y-Achse eine Linie zur Regressionsgeraden und von dort parallel zur X-Achse bis zur Y-Achse gezogen, der Schnittpunkt gibt den geschätzten Y-Wert \hat{y}_i :



Obwohl die Regressionsgerade dem Trend der Punkte im Streudiagramm am besten entspricht (nach einem bestimmten Kriterium, s.u.), werden i.d.R. nicht alle Punkte genau auf der Geraden liegen: Der **Schätzfehler** bei der Vorhersage eines bestimmten Y-Wertes ist als Differenz zwischen dem tatsächlichen und dem vorhergesagten Y-Wert definiert: $e_i = y_i - \hat{y}_i$.

Lineare Regression (6)

- Einführung VI -

Da alle anhand der Regressionsgleichung vorhergesagten Werte auf der Regressionsgeraden liegen, ist ein **Gesamtmaß für den Schätzfehler** der Regressionsgleichung nach dem **Kriterium der kleinsten Quadrate (least squares)** die Summe der quadrierten Abweichungen der vorhergesagten Werte \hat{Y} von den tatsächlichen (gegebenen) Werten.

Die Koeffizienten b_{yx} und a_{yx} der Regressionsgeraden werden so gewählt, dass diese Summe **minimal** ist:

$$\sum (y - \hat{y})^2 = \min$$

Diese Art der linearen Regression wird wegen des Kriteriums der kleinsten Quadrate im Englischen auch als **ordinary least squares regression** oder **OLS-regression** bezeichnet.

Lineare Regression (7)

- Einführung VII -

Die Gleichungen, mit denen die Koeffizienten b_{yx} und a_{yx} nach dem Kriterium der kleinsten Quadrate bestimmt werden können (mathematisch mit Mitteln der Differentialrechnung ableitbar), sind:

Steigung (slope)

→

$$b_{yx} = \frac{\sum_{i=1}^n ((x_i - \bar{x}) \cdot (y_i - \bar{y}))}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

=

$$\frac{\text{COV}_{xy}}{s_x^2}$$

=

$$r_{xy} \frac{s_y}{s_x}$$

„Konstante“ (intercept)

→

$$a_{yx} = \bar{y} - b_{yx} \cdot \bar{x}$$

D.h., zuerst wird b_{yx} und dann a_{yx} bestimmt. Die erste Variante der Gleichung für b_{yx} wird benutzt, wenn Rohdaten vorliegen, die zweite und dritte Variante können benutzt werden, wenn die Kennziffern der Kovarianz und die Varianz von X oder der Korrelation und der Standardabweichungen von Y und X verfügbar sind.

Lineare Regression (8)

- Beispiel I -

i	$x[\text{cm}]$	$y[\text{kg}]$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	187	72	8	2	16	64	4
2	170	60	-9	-10	90	81	100
3	180	73	1	3	3	1	9
4	184	74	5	4	20	25	16
5	178	72	-1	2	-2	1	4
6	180	70	1	0	0	1	0
7	172	62	-7	-8	56	49	64
8	176	70	-3	0	0	9	0
9	186	80	7	10	70	49	100
10	177	67	-2	-3	6	4	9
Summe:	1790	700	0	0	259	284	306

$$n = 10 \quad \bar{x} = \frac{1790}{10} = 179.0 \quad \bar{y} = \frac{700}{10} = 70.0 \quad s_{xy} = \frac{259}{10} = 25.9$$

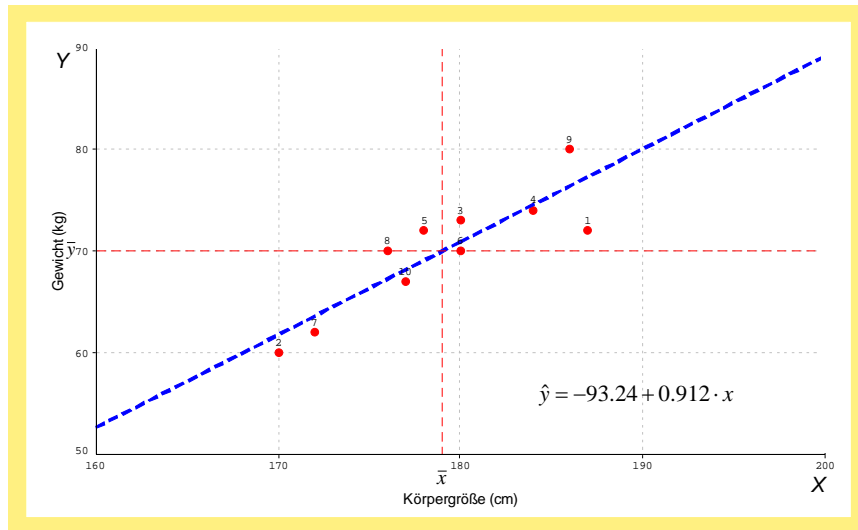
$$s_x^2 = \frac{284}{10} = 28.4 \quad s_y^2 = \frac{306}{10} = 30.6 \quad r_{xy} = \frac{25.9}{5.329 \cdot 5.532} = 0.879$$

$$s_x = \sqrt{28.4} = 5.329 \quad s_y = \sqrt{30.6} = 5.532$$

$$b_{yx} = 0.879 \cdot \frac{5.532}{5.329} = 0.912 \quad a_{yx} = 70.0 - 0.912 \cdot 179.0 = -93.243 \quad \hat{y} = -93.243 + 0.912 \cdot x$$

Lineare Regression (9)

- Beispiel II -



Lineare Regression (10)

- Merkmale der linearen Regression I -

- Gibt es keine Information, auf der die Vorhersage basieren kann, wird der gleiche Schätzwert – das *arithmetische Mittel der abhängigen Variablen Y* – für jeden vorherzusagenden Fall benutzt.

D.h., wenn die Korrelation zwischen der unabhängigen Variablen und der abhängigen Variablen Null ist, wird aus den Formeln für b_{yx} und a_{yx} :

$$b_{yx} = 0 \cdot \frac{s_y}{s_x} = 0$$

$$a_{yx} = \bar{y} - 0 \cdot \bar{x} = \bar{y}$$

Selbstverständlich würde niemand eine Regressionsgleichung zur Vorhersage benutzen, wenn Prädiktor (UV) und Kriterium (AV) unkorreliert wären.

Lineare Regression (11)

- Merkmale der linearen Regression II -

- Werden alle Werte als z-Werte ausgedrückt, ist der z-Wert der durch die Gleichung vorhergesagten Variablen Y gleich dem z-Wert der Prädiktorvariablen X multipliziert mit der Produkt-Moment-Korrelation von Y und X.

Da die Standardabweichungen der z-Werte Eins sind und die Mittelwerte Null, folgt algebraisch, dass

$$z_{\hat{y}_i} = r_{xy} \cdot z_{x_i}$$

Hiermit ergibt sich eine wichtige Eigenschaft der Regression:

Da z_x mit einem Wert multipliziert wird (der Korrelation), der in *realen* Daten numerisch immer kleiner dem Betrag von Eins ist, wird der z-Wert der vorhergesagten Variablen immer näher am arithmetischen Mittel liegen als der z-Wert der Variablen, auf dem die Vorhersage basiert. Dies ist die Basis des sogenannten **Regressionseffekts** (Regression zur Mitte von Y).

Unter anderem demonstriert die Gleichung des Weiteren, dass die **Regressionsgerade durch den Punkt (\bar{x}, \bar{y}) verläuft**: Ist $z_x = 0$ (d.h. gleich dem Mittelwert), dann ist der z-Wert der vorhergesagten Variablen ebenfalls = 0.

Lineare Regression (12)

- Merkmale der linearen Regression III -

- Je näher die Korrelation zwischen X und Y an Null liegt, desto größer ist der Vorhersagefehler.

Auch wenn die Regressionsgleichung den Vorhersagefehler minimiert, kann er doch zu groß sein. Ist die Korrelation zwischen Prädiktor (UV) und Kriterium (AV) numerisch klein (z.B. +0.07), würde die Regression einen beträchtlichen Vorhersagefehler produzieren, weil die Korrelation beider Variablen so gering ist.

Das bedeutet, dass eine Korrelation nahe Null für praktische Vorhersagezwecke nutzlos ist, auch wenn die Korrelation statistisch signifikant sein sollte (was mit jeder genügend großen Stichprobe der Fall sein kann).

Das **Vorzeichen** des Korrelationskoeffizienten hat jedoch **nichts** mit dem Vorhersagefehler zu tun: Korrelationen von z.B. +0.55 und -0.55 sind gleich gut für Vorhersagen, weil die **Stärke** des Zusammenhangs in beiden Fällen gleich ist.

Lineare Regression (13)

- Merkmale der linearen Regression IV -

- *Die Korrelation zwischen Y und jeder linearen Transformation von X ist numerisch gleich dem Betrag von r_{xy} . Da \hat{Y} ebenfalls nichts anderes als eine lineare Trans-formation von X darstellt, ist auch die der Betrag der Korrelation zwischen X und Y gleich dem Betrag der Korrelation zwischen \hat{Y} und Y .*

Wenn also der lineare Zusammenhang zwischen X und Y schwach ist (r_{xy} liegt nahe Null), werden die vorhergesagten \hat{Y} -Werte größtenteils ungenau und ziemlich verschieden von den tatsächlichen Y -Werten sein ($r_{\hat{Y}Y}$ wird genauso nahe bei Null liegen). Ist dagegen der Zusammenhang zwischen X und Y stark (r_{xy} ist numerisch groß), werden die vorhergesagten \hat{Y} -Werte präzise und konsistent mit den tatsächlichen Y -Werten sein ($r_{\hat{Y}Y}$ wird numerisch genauso groß sein).