Teil: lineare Regression

1 Einführung

2 Prüfung der Regressionsfunktion

3 Die Modellannahmen zur Durchführung einer linearen Regression

4 Dummyvariablen

1 Einführung

- Eine statistische Methode um Zusammenhänge zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen zu untersuchen. Alle einbezogenen Variablen müssen **metrisch** skaliert sein. Es besteht die Option dichotome Variablen auch metrisch zu interpretieren (Dummy-Variablen).

- Die Entscheidung, welche Variable als abhängige (AV) und welche als unabhängige Variable (UV) in die Analysen einbezogen werden, muss vorab aus einem **theoretischen Bezugsrahmen** abgeleitet werden.

- Es geht bei der Regression nicht nur um die Feststellung der Stärke eines Zusammenhangs. Die zugrunde liegenden Annahmen gehen über die einer Korrelationsanalyse hinaus. Bei der Zuordnung von UV und AV werden **kausale Beziehungen** unterstellt.

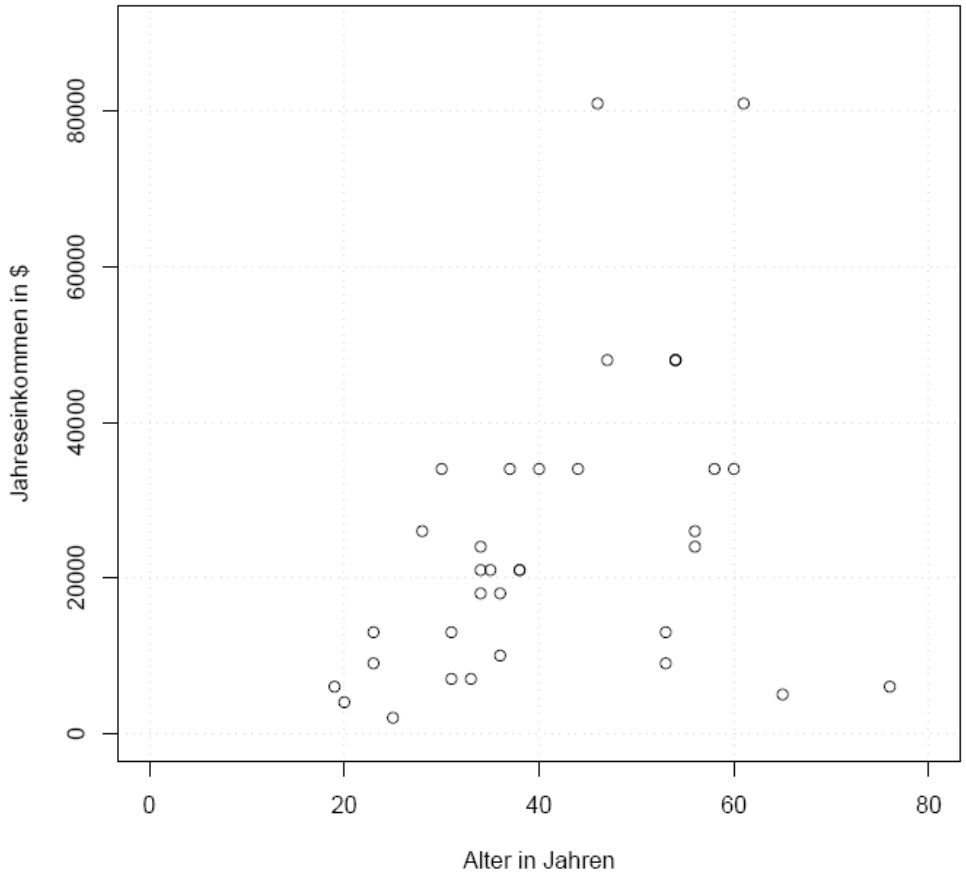
- Anwendung: **Prognosen** und **Kausalanalysen**

- Bezeichnung der **UV**: exogene Variable, erklärende Variable, Prädiktorvariable

- Bezeichnung der **AV**: endogene Variable, erklärte Variable, Prognosevariable

Ansatz der einfachen linearen Regression

Punktwolke der AV und UV.



- Dabei soll x die UV sein und y die AV.

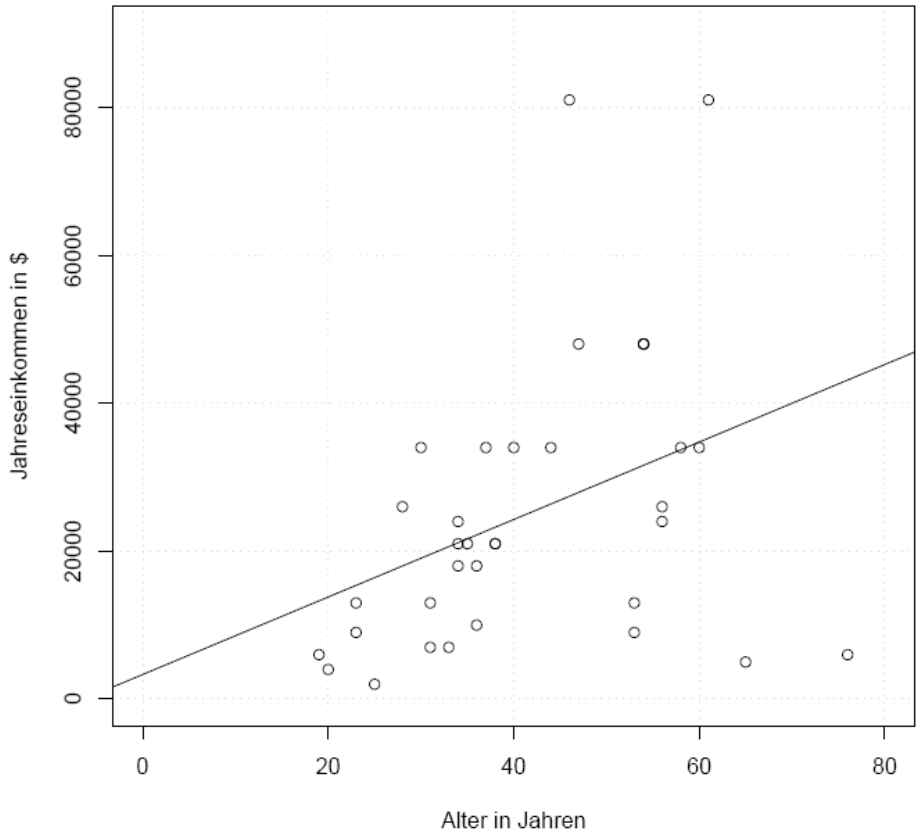
- Linear soll in diesem Kontext bezogen auf x und y heißen:

$$\frac{\Delta x}{\Delta y} = konstant$$

$$y = ax + b$$

Ziel der Regression

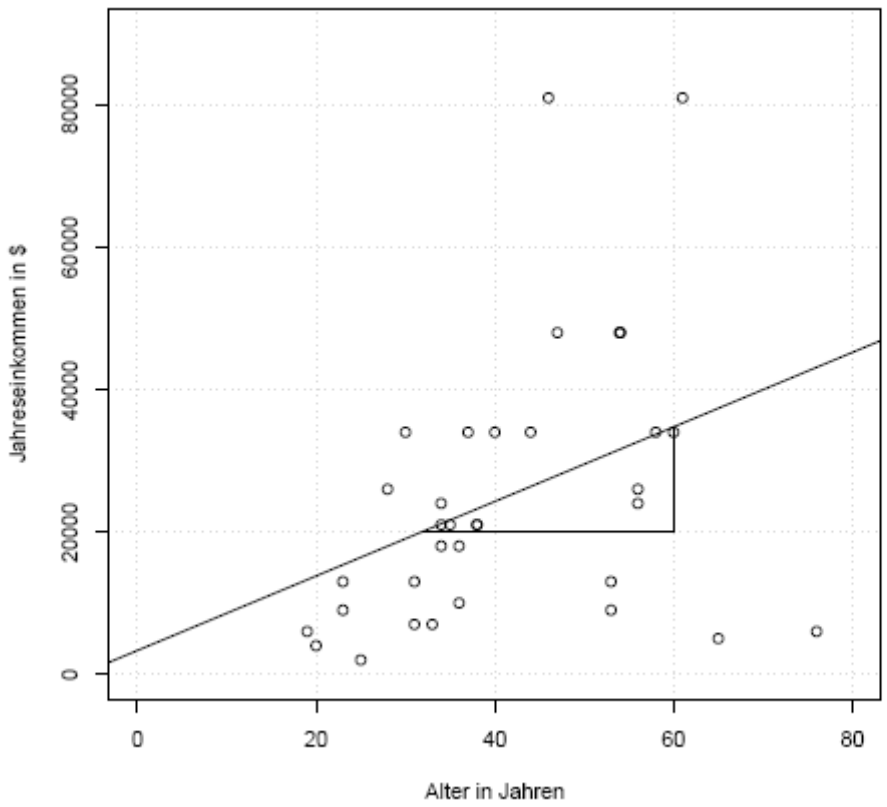
- Zurückführen von y auf x.
- Ziel ist es nun, mit Hilfe eines Modellansatzes, die Punktwolke zu beschreiben.



Parametersuche

Die grundsätzliche Frage besteht nun darin, wie findet man die beiden relevanten Parameter a und b und wie findet man die optimale Regressionsgerade durch die Punktwolke?

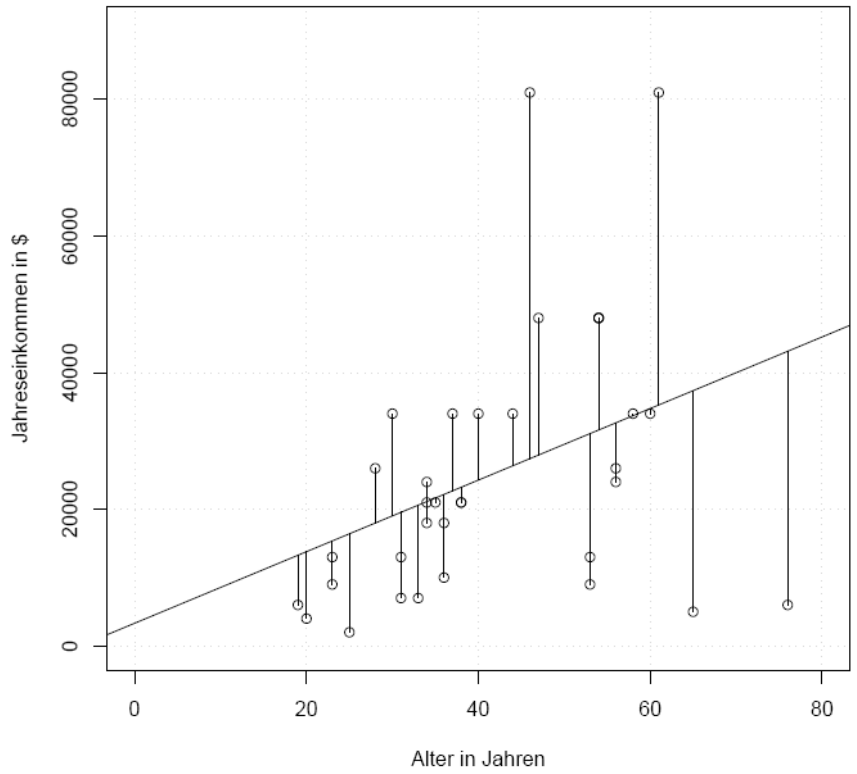
Lösung: a und b optimal bestimmen!



- WIE: Methode der kleinsten Quadrate

Methode der kleinsten Quadrate

Mit dieser Methode versucht man die Summe der quadratischen Abweichungen in Bezug auf die beiden Parameter a und b zu minimieren!



Allgemeiner Ansatz

$$y_i = b + ax_i + e_i$$

$$e = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 = \sum_{i=1}^{n} e_i \rightarrow min!$$

y_i	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
62,81	5,0	-3,3	10,9	-70,69	233,277
105,16	7,0	-1,3	1,7	-28,34	36,842
68,98	4,0	-4,3	18,5	-64,52	277,436
82,12	5,0	-3,3	10,9	-51,38	169,554
117,44	7,0	-1,3	1,7	-16,06	20,878
86,18	5,0	-3,3	10,9	-47,32	156,156
100,11	6,0	-2,3	5,3	-33,39	76,797
124,05	7,0	-1,3	1,7	-9,45	12,285
146,64	8,0	-0,3	0,1	13,14	-3,942
74,72	1,00	-7,3	53,3	-58,78	429,094
115,98	6,00	-2,3	5,3	-17,52	40,296
153,36	10,00	1,7	2,9	19,86	33,762
101,86	6,00	-2,3	5,3	-31,64	72,772
97,71	5,00	-3,3	10,9	-35,79	118,107
124,44	7,00	-1,3	1,7	-9,06	11,778
142,13	7,00	-1,3	1,7	8,63	-11,219
169,08	12,00	3,7	13,7	35,58	131,646
…	…	…	…	…	…
			∑ 571,4		**∑ 4539,228**

$\bar{y} = 133{,}5 \quad \bar{x} = 8{,}3$

y = Einkommen/100

x = Berufserfahrung in Jahren

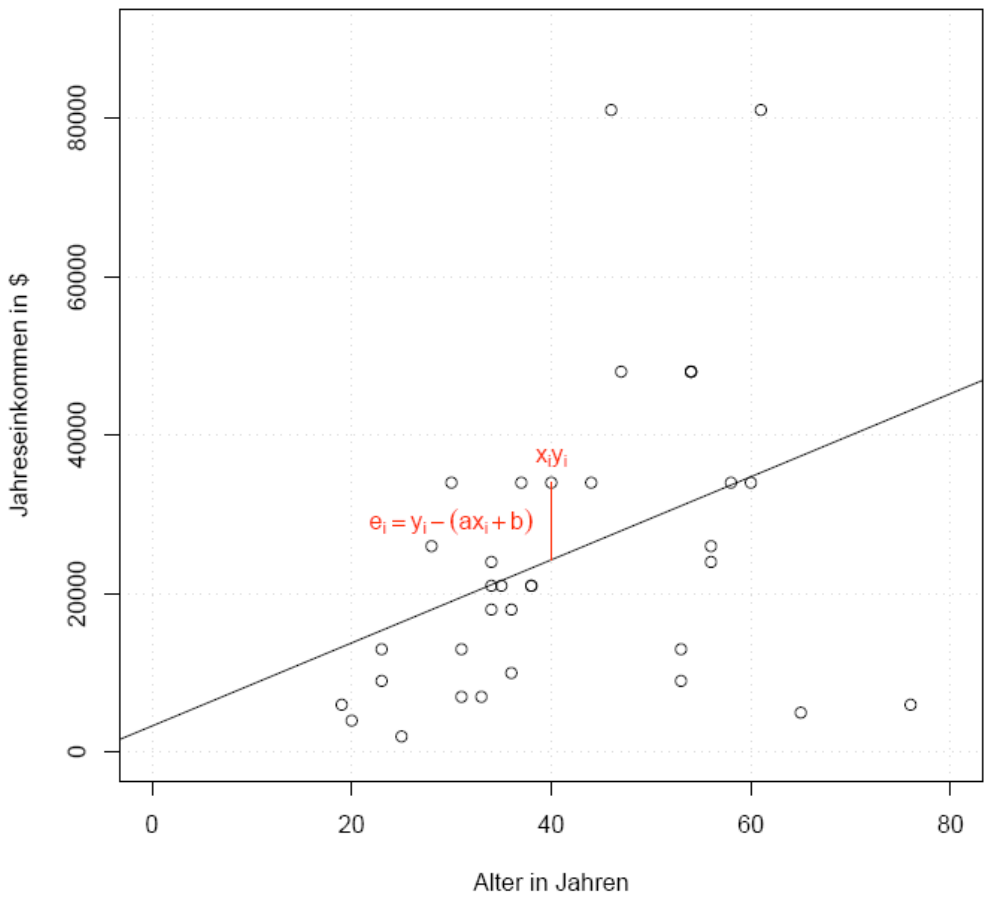
$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{4539{,}228}{571{,}4} = 7{,}944$$

$$a = \bar{y} - b\bar{x} = 133{,}5 - (7{,}944 * 8{,}3) = 67{,}6$$

$$y = 67{,}6 + (7{,}944 * x) + e$$

Person mit 5 Jahren Berufserfahrung:

$$\hat{y} = 67{,}6 + (7{,}944 * 5) = 67{,}6 + 39{,}72 = 107{,}32$$



- *b* ist die Konstante
- *a* der Regressionskoeffizient

Formales Vorgehen

1. Ermittlung eines theoretischen Bezugsrahmens
2. Formulierung von Hypothesen
3. Operationalisierung
4. Formulierung des Modells
5. Schätzung der Regressionsfunktion
6. Prüfung der Regressionsfunktion (Modellannahmen, Parameter, Verallgemeinerbarkeit)

Durchführung

Fragestellung: Gibt es einen Zusammenhang von Alter und Einkommen?

1. **Ermittlung eines theoretischen Bezugsrahmens**
   - Humankapitaltheorie

2. **Formulierung von Hypothesen**
   - Mit zunehmendem Alter steigt die Berufserfahrung, was sich direkt in steigendem Einkommen
   - bemerkbar macht. Ein Zugewinn an Bildung führt über eine gesteigerte Produktivität zu mehr Einkommen.

3. **Operationalisierung**

4. **Formulierung des Models**

   - *Einkommen = b + a*Bildungsjahre*

5. **Schätzung der Funktion**

Analysieren → Regression → Linear

6. **Prüfung der Regressionsfunktion und Annahmen**

2 Prüfung der Regressionsfunktion

Ausgehend von der Schätzung der Parameter mit Hilfe der Methode der kleinsten Quadrate stellt sich jetzt die Frage, wie gut das *Modell* zur Erklärung der *Realität* beiträgt. Dabei wird in 2 grundsätzliche Bereiche unterschieden:

- **Globale Prüfung der Regressionsfunktion:** Wird die abhängige Variable durch die in das Modell einbezogene(n) Variable(n) erklärt? Wenn ja, wie gut ist diese Erklärung?
- **Prüfung der einzelnen Regressionskoeffizienten** Tragen die Variablen einzeln zur Erklärung bei? Wenn ja, wie gut?

Was bedeutet Güte eines Modells: Welcher Anteil der Streuung der AV kann mit dem verwendeten Regressionsmodellaufgeklärt werden?

Bestimmtheitsmaß (Determinationskoeffizient) R^2

Der Anteil, der durch die Regression aufgeklärten Streuung ist das Bestimmtheitsmaß **R^2**. Es misst die Güte der Anpassung der Regressionsfunktion an die empirischen Daten (*goodness of fit*).

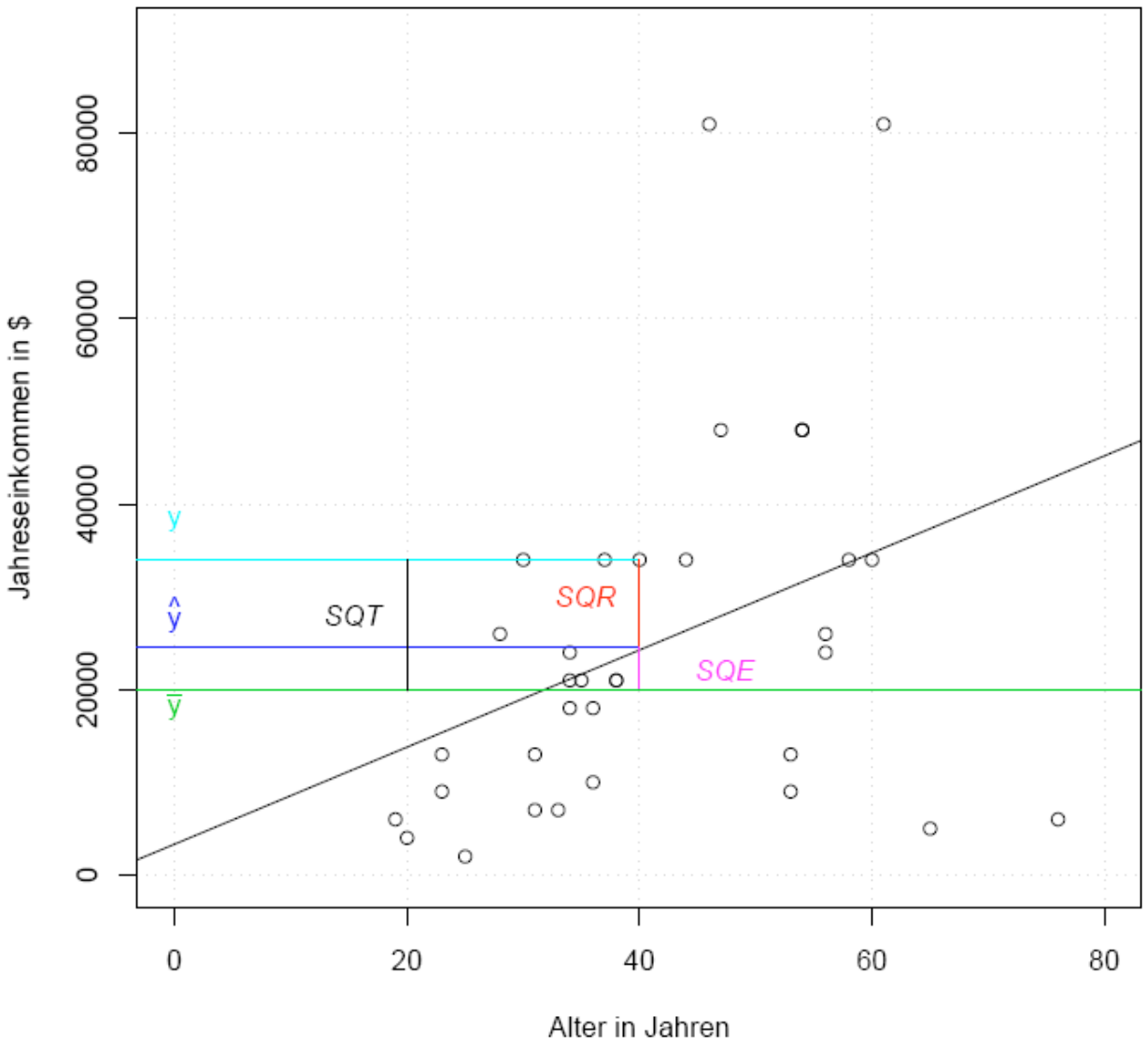
Die gesamte Streuung **SQT** von y_i (**s**um of s**q**uares **t**otal) setzt sich wie folgt zusammen:

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

$$SQT = SQE + SQR$$

$$\textbf{s}\text{um of s}\textbf{q}\text{uares }\textbf{t}\text{otal} = \text{...}\textbf{e}\text{plained} + \text{...}\textbf{r}\text{esidual}$$

$$\text{Gesamtstreuung} = \begin{matrix}\text{durch die Regression} \\ \text{erkläre Streuung}\end{matrix} + \text{Reststreuung}$$


Jahreseinkommen in $
0
20000
40000
60000
80000
0
20
40
60
80
Alter in Jahren
y
ŷ
ȳ
SQT
SQR
SQE

$$r^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT} = \frac{y\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{(y_i - \overline{y})^2}$$

$$0 \leq r^2 \leq 1$$

Interpretationsbeispiel

- Wird zur Erklärung oder Vorhersage der AV Einkommen nicht nur der Mittelwert sondern auch die UV Alter verwendet, wird die Summe des quadrierten Vorhersagefehlers um 14,4 % gesenkt, als wenn nur der Mittelwert genutzt würde.
- Alter erklärt etwa 14 % der Streuung von Einkommen (in der untersuchten Population).

Prüfung der einzelnen Regressionskoeffizienten

Regressionskoeffizienten

- Geben das Ausmaß der Steigerung der AV an, für den Fall, dass die UV um eine Einheit steigt.

Weitere Maße

Korrigiertes Bestimmtheitsmaß

- Aus einer „Gewinn maximierenden Perspektive“ wäre es sinnvoll, R^2 zu erhöhen, und damit immer mehr Varianz der AV zu erklären. Der einfachste Weg besteht in der Berücksichtigung weiterer UVs.
- Das korrigierte Bestimmtheitsmaß berücksichtigt die Anzahl der Regressoren. Damit besteht auch die Möglichkeit eines sinkenden R^2.

$$r^2_{KORR} = 1 - \frac{n-1}{n-k}(1-r^2)$$

n = Anzahl der Beobachtungen

k = Anzahl der Regressoren

Standardisierte Regressionskoeffizienten

Bei der Verwendung unterschiedlicher metrischer Variablen werden in der Regel auch unterschiedliche Skalen in das Modell eingehen. Um die Regressionskoeffizienten vergleichen zu können, kann man auf eine Standardisierung zurückgreifen.

$$\beta_{k_{STAND}} = \beta_k \frac{s_{X_k}}{s_Y}$$

Interpretation: wird die UV um eine Standardabweichung erhöht, gibt der Standardisierte Regressionskoeffizient an, um wie viele Standardabweichungen die AV steigt (oder bei negativem Vorzeichen sinkt).

Anwendung

 *datensatz14.sav*

Variablenname	**Bedeutung**
Alter	Alter in Jahren
Bildung	Bildungsjahre
Einkommen	Einkommen in Jahren

1. Schätzen Sie eine Regressionsfunktion von Einkommen auf Alter und Bildungsjahre.
2. Interpretieren Sie die Ergebnisse.

Zusatz: Überzeugen Sie sich über den Zusammenhag anhand einer Grafik (z.B. Streudiagramm).

Von der Grundgesamtheit zur Stichprobe

F Statistik

- Werden Daten aus Stichproben verwendet, stellt sich die Frage, in wie weit das geschätzte Modell auch **Gültigkeit für die Grundgesamtheit** hat. Aufgrund der Stichprobendaten enthalten die geschätzten Parameter **zufällige Streuung**.

Der p-Wert der F-Statistik gibt Auskunft über die Wahrscheinlichkeit der Daten, unter der Annahme der H0 (alle Regressionskoeffizienten sind = 0. In der Grundgesamtheit besteht kein Zusammenhang der UVs mit der AV).

t-Test

- Die F-Statistik, testet das Gesamtmodell, der t-Test die einzelnen Koeffizienten auf Zufälligkeit.
- Der p-Wert des t-Tests, gibt Auskunft über die Wahrscheinlichkeit der Daten, unter der Annahme der H0 (Der einzelne Regressionskoeffizient = 0).

Was beinhaltet der Fehlerterm?

- Messfehler
- Stichprobenfehler
- Nicht aufgeklärte Varianz

3 Die Modellannahmen zur Durchführung einer linearen Regression

Die Verwendung des stochastischen Modells der Regressionsanalyse, beruht auf einer Reihe von Annahmen.

- **Linearität**: Das Modell ist linear in den Regressionskoeffizienten. Die Anzahl der UV´s muss kleiner sein als die Anzahl der Fälle im Datensatz.
- **Unabhängigkeit der Fehler:** Es darf keine Autokorrelation bestehen.
- **Homoscedastizität**: Die Störgrößen haben konstante Varianzen
- **keine Multikollinearität**: Zwischen den UV´s besteht keine perfekte lineare Abhängigkeit
- **Vollständigkeit des Modells**: Die einzelnen Fehler haben den Erwartungswert
- **Normalverteilung**: Die Fehler sind normal verteilt.

→ Gegenstand von Vertiefungskursen

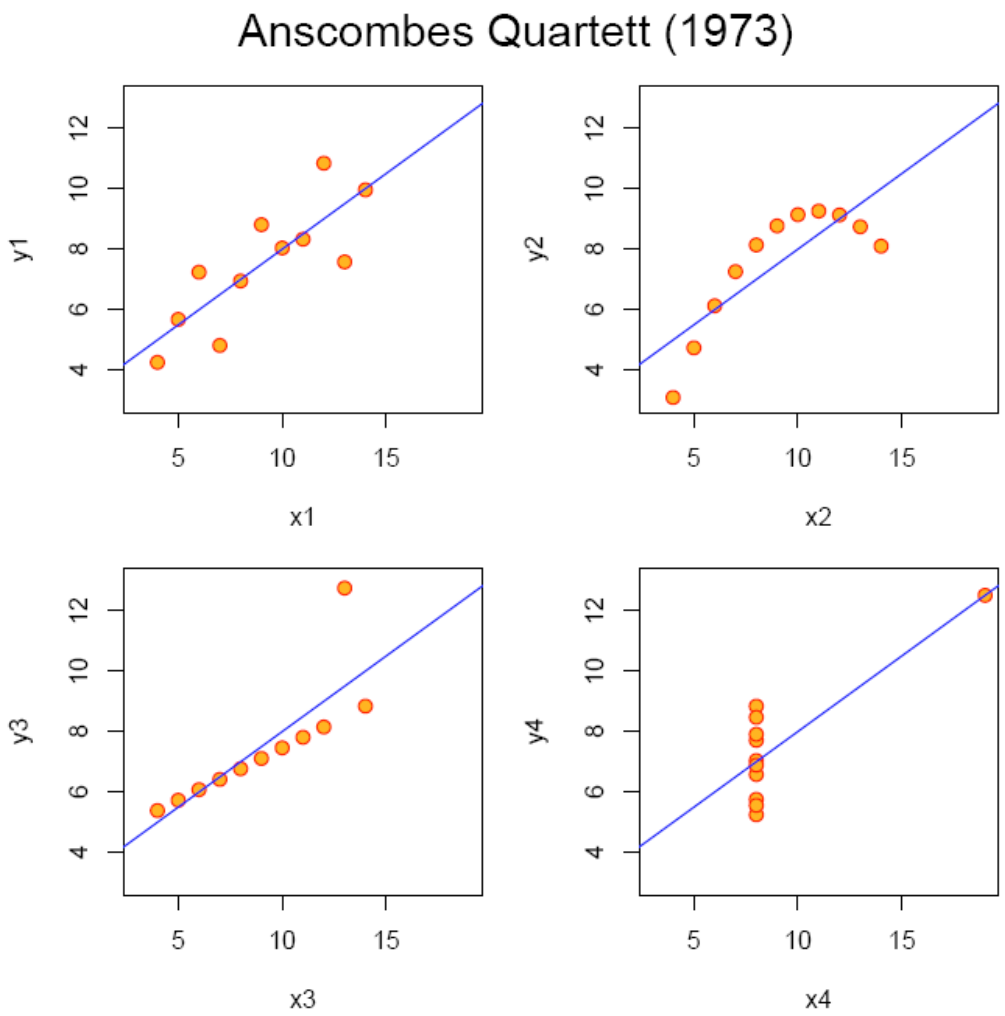
Diagnostik der Modellannahmen

- **Linearität**: Vor allem über die Plots der einzelnen UV´s gegen die AV.
- **Unabhängigkeit der Fehler**: Durbin Watson Test
- **Homoscedastizität**: Standardmethode hierfür sind die Plots der Residuen gegen die vorhergesagten Werte
- **keine Multikollinearität**: Toleranzwerte
- **Vollständigkeit des Modells**: overfitting versus underfitting
- **Normalverteilung**: Nur bei sehr kleinen Stichproben relevant, visuelle Überprüfung

→ Gegenstand von Vertiefungskursen

Ausgewählte Lösungsmöglichkeiten bei Verletzungen der Annahmen

Suche nach einflussreichen Fällen (Ausreißern)


Anscombes Quartett (1973)
y1
y2
y3
y4
x1
x2
x3
x4
4
6
8
10
12
5
10
15

a) Analysieren → deskriptive Statistik → explorative Datenanalyse

4 Dummyvariablen

Vorteil:

- Es können auch kategoriale Variablen im Modell berücksichtigt werden
- Es können nicht lineare Zusammenhänge aufgedeckt werden

Eine mögliche Strategie zur Anpassung des Modells an nichtlineare Beziehungen in Daten ist die Konstruktion von Dummy-Variablen. Dabei müssen immer n¡-1 Variablen im Modell integriert werden (n = Anzahl der Kategorien).

Interpretation:

- Bei dichotomen Variablen (z.B. Geschlecht 0 = männlich und 1 = weiblich) steht der Regressionskoeffizient für den Unterschied zwischen den beiden Ausprägungen.
- Bei mehr als 2 Ausprägungen wird eine als Referenzkategorie deklariert. Diese wird nicht mit ins Modell einbezogen → Gegenstand von Vertiefungskursen

Anwendung

 *datensatz14a.sav*

Variablenname	**Bedeutung**
Alter	Alter in Jahren
Bildung	Bildungsjahre
Einkommen	Einkommen in Jahren
Geschlecht	Männlich oder weiblich

1. Schätzen Sie eine Regressionsfunktion von Einkommen auf Geschlecht. Wie interpretieren Sie den Regressionskoeffizienten?
2. Verwenden Sie nun auch die anderen Variablen zur Erklärung von Einkommen. Diskutieren Sie das Modell.