

14. Multiple Regression

Problemstellung (bei Zweifachregression):

Für eine Größe z , die von zwei unabhängigen Variablen x und y abhängt, liege ein dreidimensionales Streudiagramm $\{(x_i, y_i, z_i), i = 1, \dots, n\}$ vor.

Nähere diesen Befund durch eine zweidimensionale Regressionsfunktion $\hat{z}(x, y)$ an!

Lineare Zweifach-Regressionsfunktion:

$$\hat{z}(x, y) = a + bx + cy$$

Bestimmung der Koeffizienten:

$$F = \sum_{i=1}^n (z_i - a - bx_i - cy_i)^2 \stackrel{!}{=} \min!$$

$$\Rightarrow \frac{\partial F}{\partial a} = 0, \quad \frac{\partial F}{\partial b} = 0, \quad \frac{\partial F}{\partial c} = 0$$

\Rightarrow lineares System für die drei Unbekannten a, b und c .

15. Maßzahlen der bivariaten Analyse

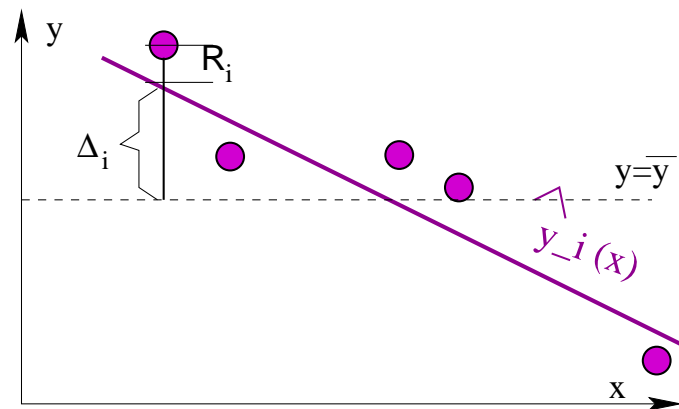
15.1 Bestimmtheits- und Unbestimmtheitsmaß

Das **Bestimmtheitsmaß** B bzw. das **Unbestimmtheitsmaß** $U = 1 - B$ einer Regression geben Antwort auf die Frage: "wieviel Prozent der tatsächlich beobachteten Schwankungen von Y_i werden durch die Regressionsfunktion $\hat{y}(x)$ erklärt bzw. nicht erklärt?"

Dazu spaltet man die Abweichungen der abhängigen Variablen y_i vom Mittelwert \bar{y} auf in eine durch die Regression

- **erklärte Abweichung** $\Delta_i = (\hat{y}_i - \bar{y})$,
- und die verbleibende **Residualabweichung** $R_i = (y_i - \hat{y}_i)$

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) = R_i + \Delta_i$$



Das Unbestimmtheitsmaß misst die *Quadratsumme* der *nicht-erklärten* Abweichung relativ zur *Quadratsumme* der *Gesamtabweichung*:

$$U = \frac{s_R^2}{s_y^2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad B = 1 - U$$

15.2 Bestimmtheitsmaß im linearen Fall

Im Falle einer linearen, einfachen Regressionsfunktion $\hat{y}(x) = a + bx$ gilt folgende wichtige Beziehung

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

bzw.

$$s_y^2 = s_{\Delta}^2 + s_R^2$$

Die **Gesamtvarianz** s_y^2 ist die Summe der (durch die Regression) **erklärten Varianz** s_{Δ}^2 und der **Residualvarianz** s_R^2 .

Damit kann man im linearen Fall das Bestimmtheitsmaß auch schreiben als

$$B = 1 - U = \frac{s_y^2 - s_R^2}{s_y^2} = \frac{s_{\Delta}^2}{s_y^2}$$

15.2(b) Aufgaben zum Bestimmtheitsmaß

1. Leiten Sie die (nur bei linearer Regression gültige) Formel $s_y^2 = s_R^2 + s_\Delta^2$ her. (nichttrivial! aus Summen folgen normalerweise nicht Quadratsummen!)

[Lösung: www.mtreiber.de/statistik1/folien9_exkurse.pdf]

2. Zeigen Sie, dass man bei linearer Regression das Bestimmtheitsmaß errechnen kann durch

$$B = \frac{s_{xy}^2}{s_x^2 s_y^2} = r_{xy}^2$$

3. Berechne Sie die Unbestimmtheits- und Bestimmtheitsmaße für einige Regressionsaufgaben des vorigen Kapitels

15.3 Korrelationskoeffizienten

Während einer nur Zahlen im Kopf hat, kann er nicht auf den Kausalzusammenhang kommen

Der empirische **Maßkorrelationskoeffizient nach Pearson** gibt das Maß der *linearen(!)* Abhängigkeit in einem Streudiagramm der beiden kardinalskalierten Variablen X und Y an:

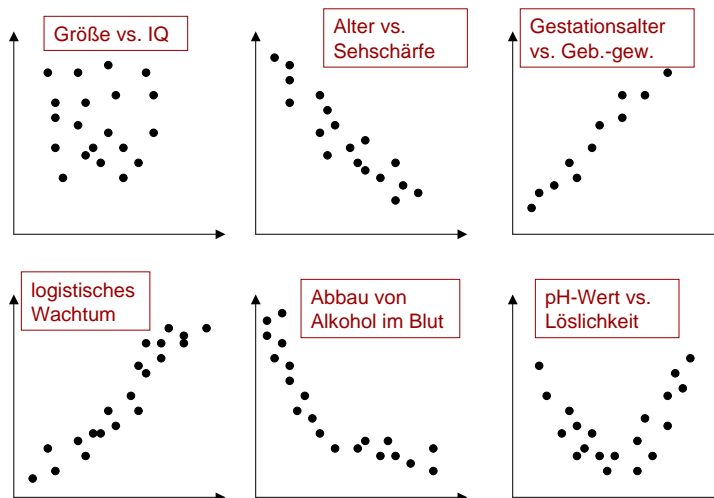
$$r_{xy} = r_{yx} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Wertebereich $-1 \leq r_{xy} \leq 1$
- Vorzeichen: Positiv bei ansteigender Tendenz, negativ bei abfallender Tendenz.
- Betrag: Nahe 1 \Rightarrow deutlicher Trend bzw. Zusammenhang; nahe Null \Rightarrow kein wesentlicher linearer Zusammenhang.

15.3(b) Aufgaben und Beispiele

1. Formen Sie den Ausdruck für den Korrelationskoeffizienten so um, dass man zum Berechnen nur noch \bar{x} , \bar{y} sowie die Summen von x_i^2 , y_i^2 und $x_i y_i$ braucht.
2. Wie könnte man aus klassierten Daten (Kreuztabelle) den Korrelationskoeffizienten berechnen?
3. Schätzen Sie die Korrelationskoeffizienten in folgenden Scatter-Plots ab!

Scatterplot - Beispiele



(Beispiel aus

<http://www.imise.uni-leipzig.de/~ingo/lehre/Biomathematik/zusammenhangsanalyse.pdf>)

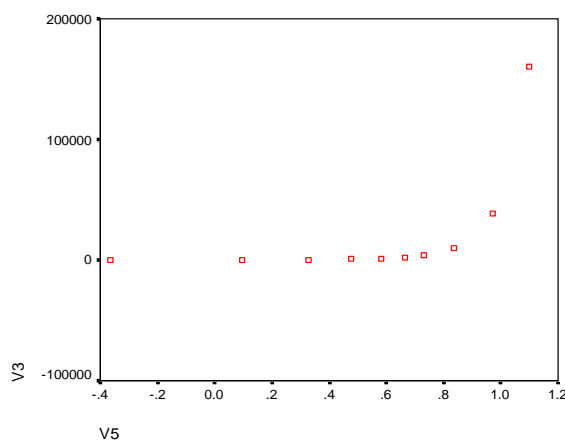
15.3(c) Rangkorrelationskoeffizient nach Spearman

Mit dem Maß

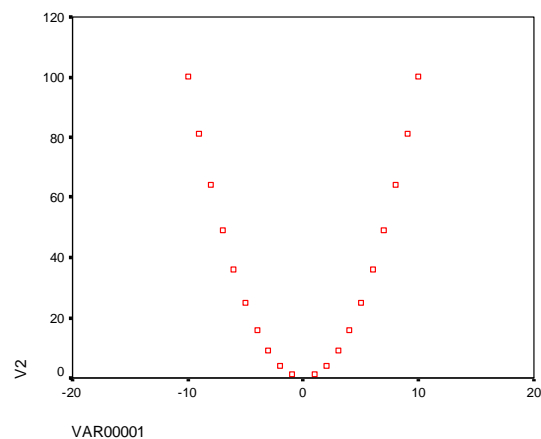
$$r_s = 1 - \frac{6 \sum_{i=1}^n (R_i^x - R_i^y)^2}{n(n^2 - 1)}$$

kann die lineare Abhängigkeit auch bei lediglich ordinalskalierten Daten quantifiziert werden. Dabei geben die R_i^x , R_i^y die *Rangziffern* von 1 bis n an, wobei keine Rangziffer doppelt vorkommen darf.

Arten von Korrelationskoeffizienten (Beispiele)



Pearson: 0.558 (p=0.094)
Spearman: 1



Pearson: 0
Spearman: 0

15.3(d) Bemerkungen

- Im Gegensatz zur Regressionsanalyse sind bei der Korrelationsanalyse die beiden Variablen gleichwertig
- Der Rangkorrelationskoeffizient r_s lässt sich natürlich auch für kardinalskalierte Daten berechnen. Er ist dann nichts weiter als der Maßkorrelationskoeffizient r_{xy} , angewandt auf die Rangziffern der Daten.



Aufgabe: Leiten Sie für kardinalskalierte Daten die Definition von r_s aus der für r_{xy} her. (Lösung: Siehe Exkurse zu dieser Vorlesung, www.mtreiber.de/statistik1/folien9_exkurse.pdf)

- Im Falle einer linearen Regression gilt: Das Bestimmtheitsmaß ist das *Quadrat* der Korrelationsfunktion:

$$B = r_{xy}^2$$

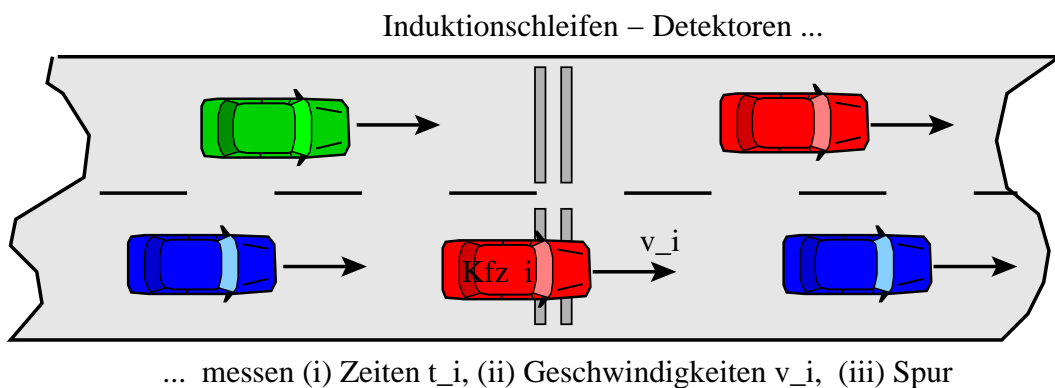
Aufgabe: Zeigen Sie dies! (Lösung: www.mtreiber.de/statistik1/folien9_exkurse.pdf)

Im linearen Fall tritt also der konzeptionelle Unterschied zwischen der Regressions- und der Korrelationsanalyse in den Hintergrund.

- Selbst für nominalskalierte Daten gibt es ein Zusammenhangsmaß, welches hier allerdings nicht weiter betrachtet wird: Das Kontingenzmaß nach Cramér's. (Vgl. z.B. Eckstein, Kap. 5.1.1).

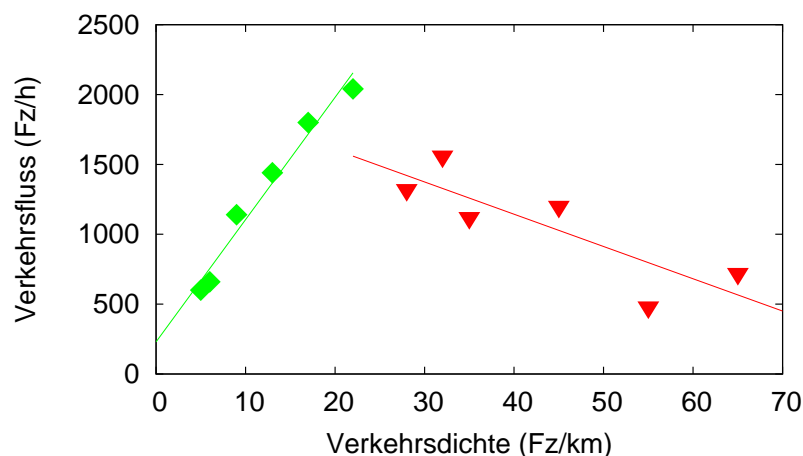
15.3(e) Komplettes Übungsbeispiel

Eines der wichtigsten und einfachsten Hilfsmittel zur Analyse von Verkehrs-Strömen und zur Stauprognose ist das sogenannte **Fundamentaldiagramm**, bei dem der Verkehrsfluss Q (Fahrzeuge pro Zeiteinheit, die eine feste Stelle überqueren) in Abhängigkeit der Verkehrsdichte ρ (Fahrzeuge pro Strecke) aufgetragen ist. Üblicherweise werden die Daten für jede Spur getrennt mittels *Doppel-Induktionsschleifen* gewonnen und pro Minute ein Wert für die Dichte und den Fluss ausgegeben.



Als einfaches Beispiel seien nun für freien und gestauten Verkehr je 6 Datenpunkte (=6 Minuten) gegeben:

Dichte (frei)	5	9	17	13	6	22
Fluss (frei)	600	1140	1800	1440	660	2040
Dichte (gestaut)	28	55	65	45	32	35
Fluss (gestaut)	1320	480	720	1200	1560	1120



15.3(e) Übungsbeispiel II

1. (\Rightarrow *Statistik I*) Ein Messschleifen-Detektor einer gegebenen Spur misst in jeder Minute die Zahl der über ihn gefahrenen Fahrzeuge sowie alle Geschwindigkeiten. Wie würde man daraus die Werte in der Tabelle bestimmen, d.h. den Fluss Q und eine bestmögliche Schätzung für die Dichte ρ ?
2. Berechnen Sie, jeweils getrennt für die sechs Punkte freien und gestauten Verkehrs, die linearen Regressionen.
3. Bestimmen Sie für den freien Verkehr die mittlere "Wunschgeschwindigkeit" V_0 durch Vergleich der Regression mit der Beziehung

$$Q_{\text{frei}}(\rho) = V_0 \rho.$$

4. Im Kolonnenverkehr wird zum Vorderfahrzeug je nach Fahrer ein gewisser zeitlicher Abstand T gehalten (Fahrschule: "Abstand=halber Tacho"). Bestimmen Sie aus der linearen Regression des gestauten Verkehrs die mittlere Folgezeit T der Fahrer sowie die maximale Verkehrsdichte ρ_{max} bei stehendem Verkehr. Vergleichen Sie dazu das Ergebnis der Regression mit der Beziehung (wie kommt man drauf?)

$$Q_{\text{stau}}(\rho) = \frac{1}{T} \left(1 - \frac{\rho}{\rho_{\text{max}}} \right).$$

Wie verhält sich die Empfehlung der Fahrschule dazu?

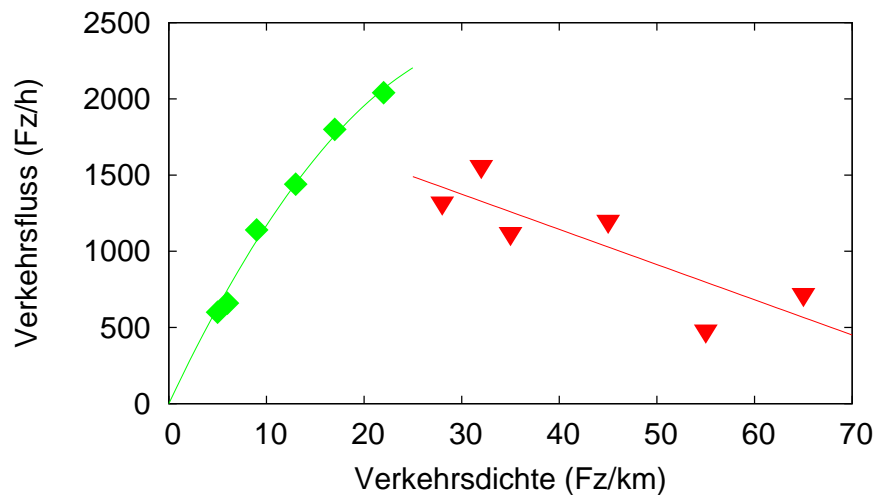
5. Berechnen Sie Korrelationen und Bestimmungsmaß für die linearen Regressionen des freien und gestauten Verkehrs.

15.3(e) Übungsbeispiel III

6. Warum ist die lineare Regression für den freien Verkehr etwas fragwürdig? betrachten Sie dazu den Wert der Regression an der Stelle $\rho = 0$. Führen Sie eine nichtlineare Regression mit der Regressionsfunktion

$$\hat{Q}(\rho) = b\rho + c\rho^2$$

durch. Welchen Wert für die freie Wunschgeschwindigkeit V_0 erhält man mit diesem verbesserten Ansatz?

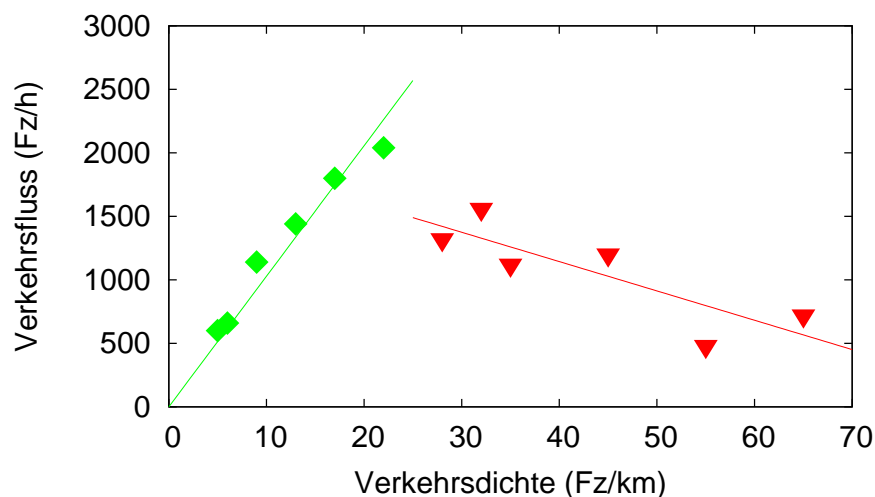


15.3(e) Übungsbeispiel IV

7. Man kann auch bei der lineare Regression berücksichtigen, dass aus äußeren Gründen der Punkt $(0, 0)$ fest gegeben ist, indem man den konstanten Term der Regressionsfunktion weglässt:

$$\hat{Q}(\rho) = b\rho$$

Zeichnen Sie auch für diesen Fall die Regressionsfunktion und bestimmen Sie die Geschwindigkeit im freien Verkehr.



Merke: Das Ergebnis hängt vom Ansatz ab; mit einem ungeeigneten Ansatz (hier $\hat{Q}(\rho) = a + b\rho$) erhält man falsche Aussagen; besser ist hier $\hat{Q}(\rho) = b\rho$, am besten die nichtlineare Regression.

15.4(a) Interpretation der Korrelation

Eine **“signifikante”** Korrelation zwischen X und Y kann u.a. bedeuten:

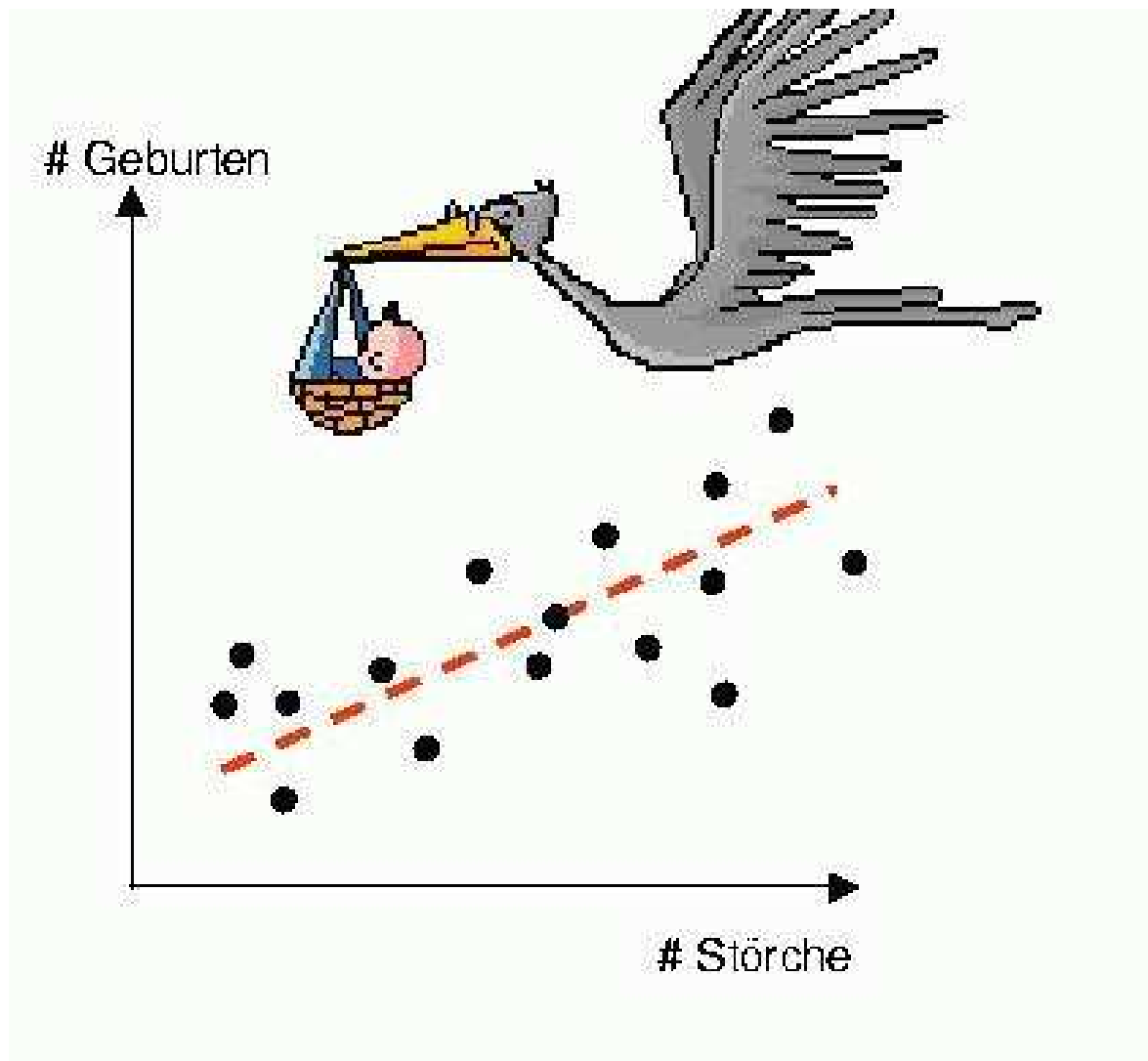
1. **kausaler Zusammenhang** zwischen X und Y
2. gemeinsame **dritte Ursache** Z
3. nicht neutrale Wahl der Grundgesamtheit: **“selektive Wahrnehmung”**
4. **Inhomogenität** der Stichprobe
5. **Ausreißer** in den Daten

Beispiele:

- (a) Die Anzahl der Störche ist i.A. *signifikant positiv korreliert* mit der Geburtenrate
- (b) Auf Autobahnen mit niedriger Verkehrsdichte *steigt* die gemessene mittlere Geschwindigkeit zunächst mit der Verkehrsdichte an
- (c) Die Unfallhäufigkeit steigt mit dem Alkoholisierungsgrad der Fahrer
- (d) Das mittlere Gehalt von Arbeitnehmern in Deutschland ist *signifikant* mit der Schuhgröße korreliert

Aufgabe: Ordnen Sie den Beispielen (a) bis (d) je eine der Interpretationen 1. bis 5. zu!

15.4 (b) Beispiele von Scheinkorrelationen



“Je größer die Storchendichte, desto größer die Geburtenrate”

“Je häufiger der Arztbesuch, desto größer die Wahrscheinlichkeit, im nächsten Jahr zu sterben”

“Je massiver der Einsatz der Feuerwehr, desto größer der Schaden”