

13. Lineare (Regressions-) Modelle

$$\hat{y}(x)$$

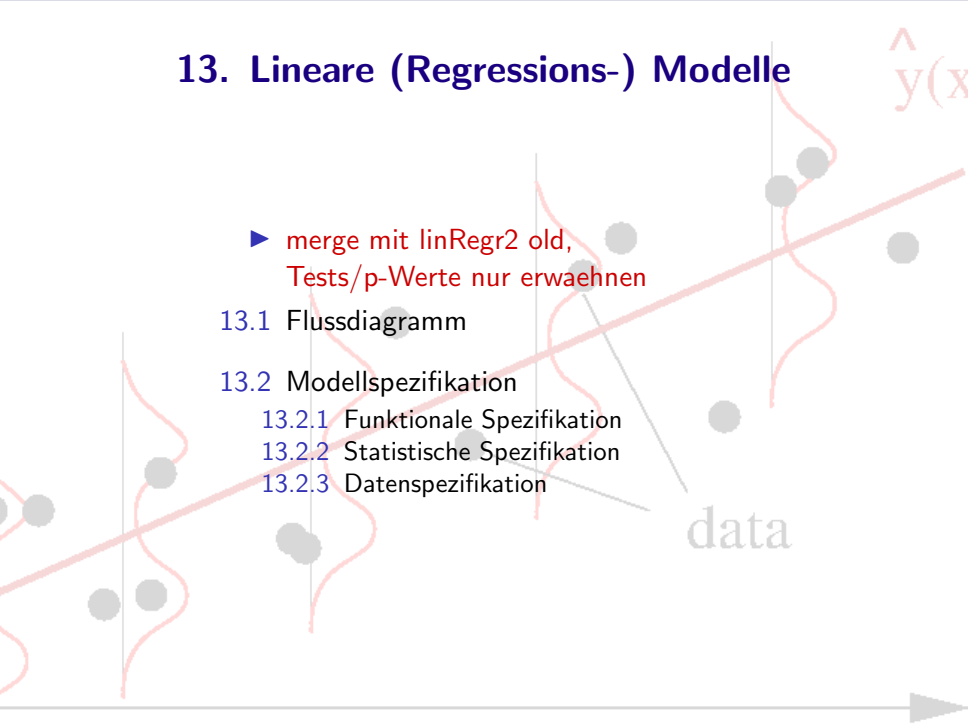
- ▶ merge mit linRegr2 old,
Tests/p-Werte nur erwahnen

13.1 Flussdiagramm

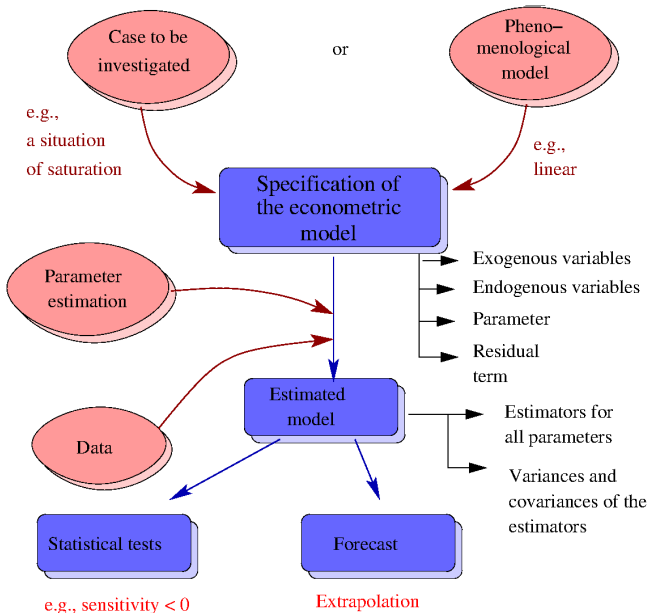
13.2 Modellspezifikation

- 13.2.1 Funktionale Spezifikation
- 13.2.2 Statistische Spezifikation
- 13.2.3 Datenspezifikation

data



13.1 Flussdiagramm



13.2 Modellspezifikation

Modellspezifikation: *vollständige strukturelle Spezifikation* des Modells und seine Konsistenz mit den verfügbaren Daten. Es gibt drei Ebenen:

- ▶ **Funktionale Spezifikation:** Exogene und endogene Variable und die funktionale Form, wie sie im Modell auftauchen. Insbesondere die Transformation der exogenen Variablen \tilde{x} in lineare **Faktoren** $x_j = g_j(\tilde{x})$ durch feste, i.A. nichtlineare Funktionen $g_j(\cdot)$
- ▶ **Statistische Spezifikation:** Wie sind z.B. *Fehlerterme* verteilt und miteinander korreliert?
- ▶ **Datenspezifikation:** Passt das Modell zu den Daten? Gibt es eine ausreichende Zahl an Datensätzen? Sind alle exogenen *und* endogenen Variablen vorhanden?

WARNUNG

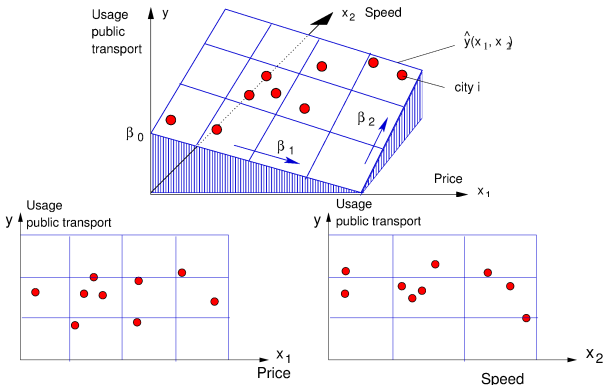
Fehlspezifizierte Modelle führen zu Fehlern aller Art, von unproblematisch bis desaströs

- ▶ **irrelevant:** Einige Fehlspezifikationen werden automatisch durch die Kalibrierungsprozedur “entdeckt”, z.B. in dem sie “Null/Null” Fehler oder singuläre Matrizen produziert
- ▶ **mild:** keine automatische Detektierung durch die Kalibrierungsmethode, aber milde Folgen: Schätzer ist nach wie vor unverzerrt/effizient, aber die induktive Statistik führt zu unkorrekten Ergebnissen
- ▶ **mittel:** Immer noch unverzerrte Schätzer, aber dieser ist nicht mehr effizient und es gibt starke inferentielle Fehler, insbesondere wird eine zu hohe Signifikanz vorgegaukelt
- ▶ **desaströs:** ie Ergebnisse sind in einer unvorhersehbaren Weise verzerrt

Junk in, junk out!

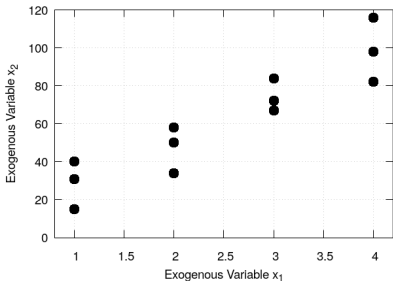
Es gibt Lügen, verdammte
Lügen und **Statistik!**

13.2.1 Funktionale Spezifikation 1: relevante Faktoren

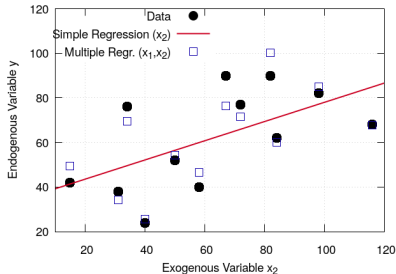
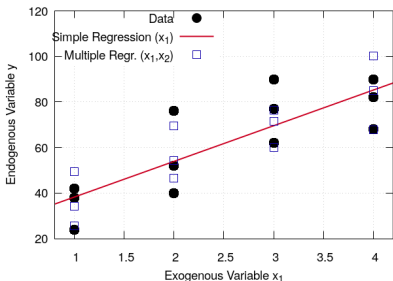


- ▶ Berücksichtige alle relevanten Faktoren (oben), lasse keinen aus! (unten)
- ▶ Folgen fehlender Faktoren: i.A. Verzerrung: **“junk in, junk out”**
- ▶ Folgen überflüssiger Faktoren: **keine Verzerrung, aber unnötig hohe Schätzfehler**
- ▶ Lösung: Tests, z.B. **F-test**: *ökonometrische Expertise nötig!*

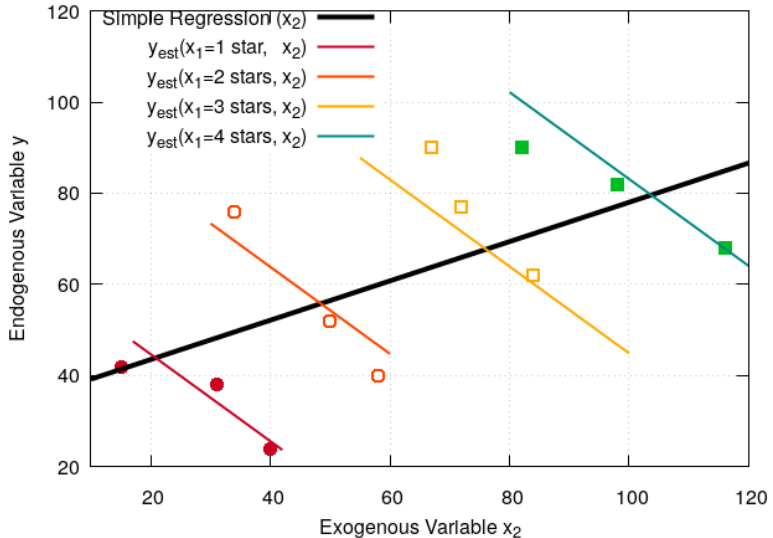
Beispiel: Auslastung von Hotels



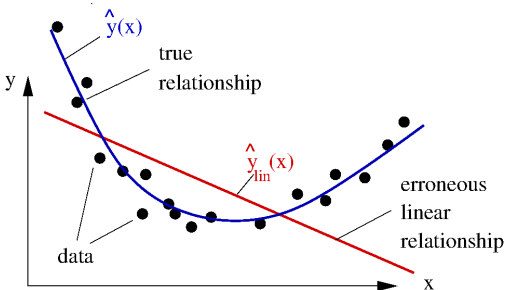
- ▶ $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ mit den Faktoren $x_0 = 1$, x_1 : Proxy für Qualität [# stars]; x_2 : Preis [€/Nacht].
- ▶ Die exogenen Faktoren sind (nicht vollständig) korreliert: ✓
- ▶ Endogene Variable: Auslastung [%]
- ▶ Nachfrage ist positiv mit Qualität und Preis (!) korreliert



Effekt der Korrelationen zwischen den exogenen Variablen



Funktionale Spezifikation 2: Linearität



- ▶ Das Modell sollte linear sein (hier nicht erfüllt)
- ▶ **Konsequenz: "junk in, junk out"**
- ▶ **Lösung:** Nichtlineare Transformation der exogenen Variablen in **Faktoren**, bezüglich derer das Modell linear ist, hier z.B. $x'_0 = 1, x'_1 = 1/x, x'_2 = x^2$ or $x'_0 = 1, x'_1 = x, x'_2 = x^2$.

Beispiel: Kraftstoffverbrauch

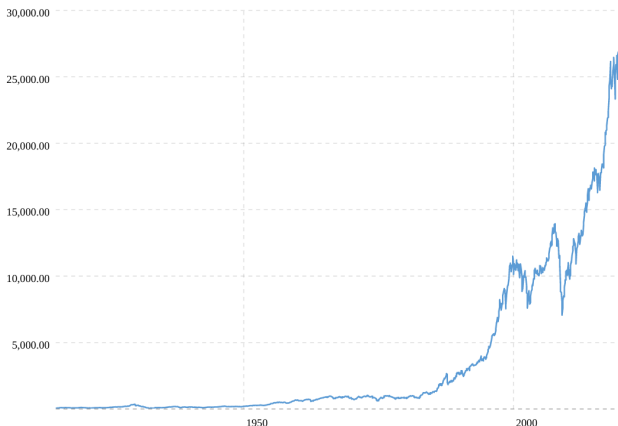
Für einen konstanten Wirkungsgrad chemische \rightarrow mechanische Energie ist der für die Fortbewegung benötigte Verbrauch y (pro 100 km) proportional zum *Fahrwiderstand* mit den additiven Anteilen

- ▶ Rollreibung: Unabhängig von der Geschwindigkeit \tilde{x}_1 , proportional zur Masse \tilde{x}_2 .
- ▶ Luftwiderstand: proportional zur quadrierten Geschwindigkeit \tilde{x}_1^2 , unabhängig von der Masse
- ▶ Steigung: proportional zum Produkt aus Masse und Steigung \tilde{x}_3

Zusätzlich gibt es eine Basisverbrauchsrate (≈ 0.5 Liter/h) für Leerlauf, Heizung, Lüftung, Licht etc \Rightarrow Anteil proportional zu $1/\text{Geschwindigkeit} \Rightarrow$ Model

$$y(\mathbf{x}) = \sum_{j=1}^4 \beta_j x_j + \epsilon, \quad x_1 = \tilde{x}_2, \quad x_2 = \tilde{x}_1^2, \quad x_3 = \tilde{x}_2 \tilde{x}_3, \quad x_4 = \frac{1}{\tilde{x}_1}$$

Transformation der endogenen Variable I



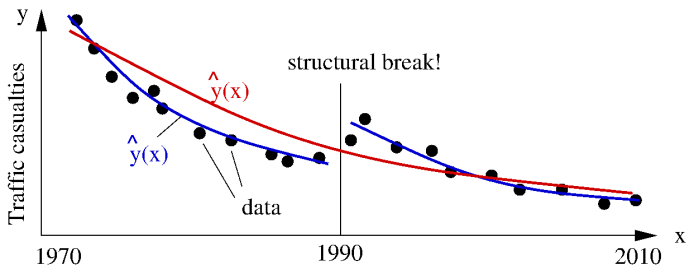
Eine Zeittransformation $\tilde{x} \rightarrow x = \exp(\tilde{x})$ würde zwar linearisieren, aber die Fluktuationen (Residualterme) sind nicht mehr i.i.d (\Rightarrow statistische Spezifikation)

Transformation der endogenen Variable II



Transformation der *endogenen* Variable $y \rightarrow u = \ln(y)$, $x = \tilde{x}$
gibt hingegen ein korrekt spezifiziertes Modell $u(x) = \beta_0 + \beta_1 x + \epsilon$, $\epsilon \sim \text{i.i.d.}$

Funktionale Spezifikation 3: Homogenität der Grundgesamtheit



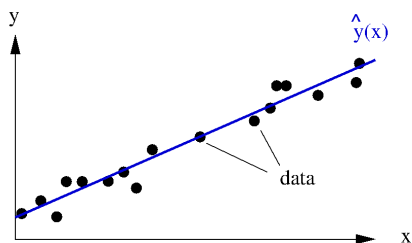
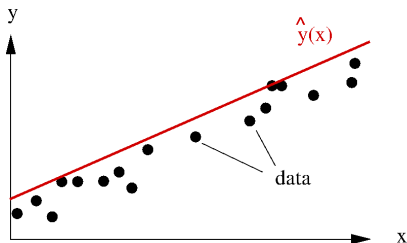
- ▶ **Folgen einer Verletzung:** ein nichtberücksichtigter **Strukturbruch** im Raum der exogenen Variablen führt zu einer **Verzerrung: junk in, junk out**
 - ▶ **Lösung:** *Dummyvariable* mit Werten 0 und 1 vor/nach dem Strukturbruch
- ? Möglicher Strukturbruch in obigem Plot?



1. neue Datenbasis (DDR+Westdeutschland → Deutschland); 2. Umdefinitionen, z.B. "ernsthaft verletzt" bedeutete vor dem Strukturbruch stationärer, danach ambulanter+stationärer Krankenhausaufenthalt

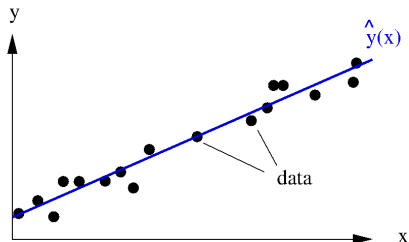
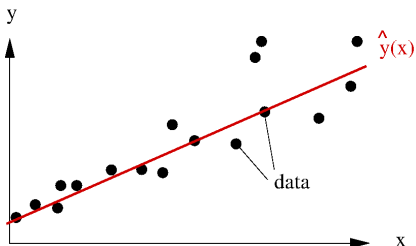
13.2.2 Statistical Spezifikation

1. Residuum ϵ hat Erwartungswert null



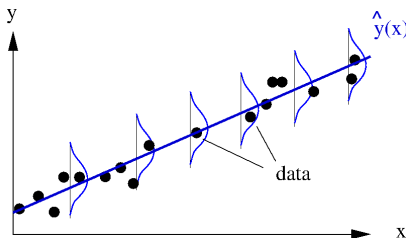
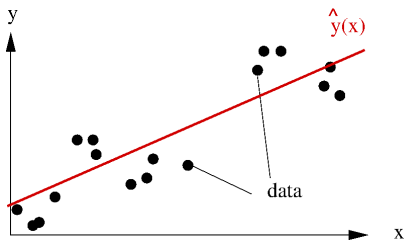
- ▶ **Bedingung:** $E(\epsilon) = 0$.
- ▶ **Konsequenz:** **keine:** Die OLS- (KQ-) Kalibrierung/Schätzmethode berücksichtigt dies automatisch. Bei der diskreten Wahltheorie ist es sogar überhaupt nicht relevant (**Warum?**)

Statistische Spezifikation 2: Homoskedastizität



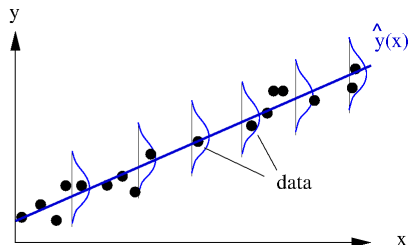
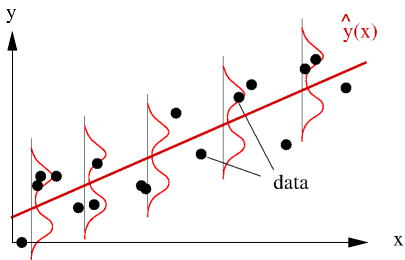
- ▶ **Bedingung:** ϵ soll konstante Varianz haben (Homoskedastizität, rechts), keine variable (Heteroskedastizität, links)
- ▶ **Konsequenz:** bei Verletzung bleibt der KQ-Schätzer **unverzerrt, ist aber nicht mehr effizient** (ein relativ "milder" Fehler).
- ▶ **Lösung:** Fortgeschrittene Methoden wie gewichteter KQ-Schätzer. Ggf Transformation der abhängigen Variable (→ Dow-Jones-Beispiel)

Statistische Spezifikation 3: Korrelationsfreiheit der Residuen



- ▶ **Bedingung:** ϵ ist nicht bezüglich der x_i und/oder y korreliert (rechts). Die linke Modell-Daten-Kombination ist fehlspezifiziert
- ▶ **Konsequenz:** relativ **mild:** (KQ-Schätzer nicht effizient; Unterschätzung der Schätzfehler, evtl unbedeutende Verzerrung).
- ▶ **Lösung:** Identifiziere aus dem Sachverhalt einen nichtberücksichtigten systematischen Einfluss, z.B. Periodizität

Statistische Spezifikation 4: Gaußverteilte Residuen



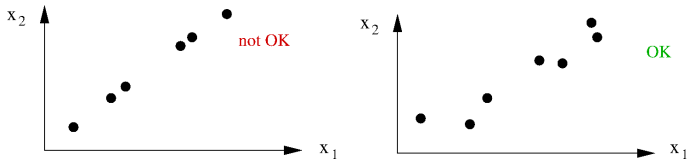
- ▶ **Bedingung:** $\epsilon \sim N(0, \sigma^2)$ (rechts), nicht anders verteilt (links).
- ▶ **Konsequenz:** eine Verletzung hat **sehr milde** Konsequenzen. Der KQ-Schätzer bleibt unverzerrt *und* effizient, nur die üblichen Fehlerabschätzungen und Tests sind falsch, vor allem bei kleinen Fehlerwahrscheinlichkeiten α
- ▶ Alle vier statistischen Spezifikationen zusammen:

$\epsilon \sim \text{i.i.d.} N(0, \sigma^2)$ i.i.d.: **identical independent distributions**

Datenspezifikation 1: genug Daten

- ▶ Jeder Datensatz muss *alle* endogenen und exogenen Variablen enthalten und es muss mehr Datensätze als Parameter geben: $n > J + 1$. Die Daten müssen das Modell also *überbestimmen*
- ▶ **Konsequenz einer Verletzung:** Falls $n = J + 1$, kann das Modell *exakt* ohne Fehler an die Daten angepasst werden: $\epsilon_i = 0$ bzw. *overfitting*. Dennoch **harmlos**, da der KQ-Schätzer und andere Tests einen “0/0-Fehler” auswerfen. Bei $n < J + 1$ wird darüberhinaus eine Inversion einer singulären Matrix versucht
- ▶ **Konsequenz einer Erfüllung “gerade so”:** Falls es nur wenig mehr Datensätze als Parameter gibt, also wenig **Freiheitsgrade** verbleiben, ist der Schätzer unverzerrt und effizient, aber hat **großen Schätzfehler**
- ▶ **Lösung:** Mehr Daten ...

Datenspezifikation 2: Keine Multi-Kolarität



- ▶ Multikolarität → Mindestens eine exogene Variable kann *in allen Datensätzen* als Linearkombination anderer Variablen angegeben werden. Die Datenmatrix ist damit *singulär*.
- ▶ Nichtperfekte Korrelationen hingegen sind ausdrücklich erlaubt
- ▶ Nichtperfekte Korrelationen sind häufig, z.B. Preis vs. Qualität (**Vorzeichen?**)
- ▶ **Konsequenz einer verletzung:** der KQ-Schätzer **“entdeckt” für Sie eine Verletzung automatisch:** “0/0”-Fehler. Nahezu perfekte Kolarität ⇒ **große Schätzfehler**

Sind alle Kriterien aus allen drei Spezifikationen erfüllt, sagt man, dass das ökonometrische Problem die **Gauß-Markov Annahmen** erfüllt

Wie entdeckt man Multi-Kolinearität?

- ▶ Gegeben: n Datensätze $\{x_{i0}, \dots, x_{ij}, \dots, x_{iJ}\}$, $i = 1, \dots, n$ (die Datensätze enthalten auch y_i , aber das ist hier nicht relevant)
- ▶ x_{ij} ist der j^{th} Faktor im Datensatz i
- ▶ Multikollinearität ist gegeben, wenn man einen Faktor x_k als Linearkombination aus allen anderen Faktoren *für alle Datensätze* ausdrücken kann:

$$x_{ik} = \sum_{j \neq k} c_j x_{ij} \quad \forall i = 1, \dots, n, \quad \text{konstante Koeffizienten } c_j$$

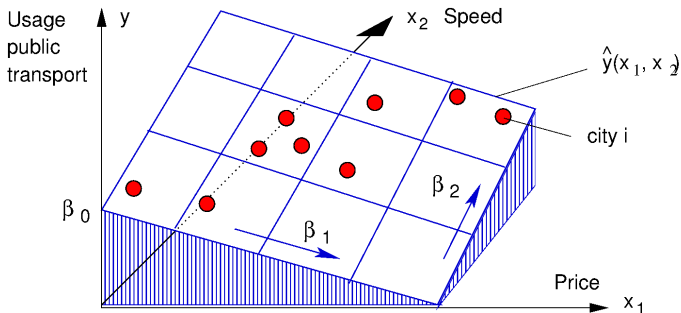
- ▶ Bei zwei Variablen ist dies leicht: $x_2 = c_0 x_1$ bzw Korrelationskoeffizient ist $=1$ oder $= -1$. Bei mehreren Variablen gilt das nicht.
- ▶ Lösung: Check, ob die **deskriptive Varianz-Kovarianz-Matrix**

$$S_{jk} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

den vollen Rang $J + 1$ hat bzw. $\det \mathbf{S} \neq 0$

- ▶ Für $n < J + 1$ ist das trivialerweise nicht erfüllt

Datenspezifikation 2: Beispiel



- ▶ Die Nachfrage y_i für den ÖPNV in den Städten i hängt vom Preis x_{i1} und der Qualität x_{i2} (Proxy: Haustür-zu-Haustür-Geschwindigkeit) ab.
- ▶ Parameter: Achsabschnitt (*intercept*) β_0 , Preissensitivität β_1 , Qualitätsbewusstsein β_2
- ▶ Preis und Qualität sind (*wie bei fast allen Produkten und Dienstleistungen*) korreliert, aber unvollständig: OK