Seminar zur Industriebetriebslehre III
Wintersemester 07/08
Dr. Claudia Höck

„Lineare und nichtlineare multiple Regressionsanalyse zur Prognose vom Absatzverläufen“

Ryutaro Karasu

Agenda

I. **Motive der Regressionsanalyse zur Prognose**

II. **Ablaufschritte der Regressionsanalyse**

III. **Anwendungsbeispiel in Excel**

IV. **Kritische Zusammenfassung**

Agenda

I. **Motive der Regressionsanalyse zur Prognose**

II. **Ablaufschritte der Regressionsanalyse**

III. **Anwendungsbeispiel in Excel**

IV. **Kritische Zusammenfassung**

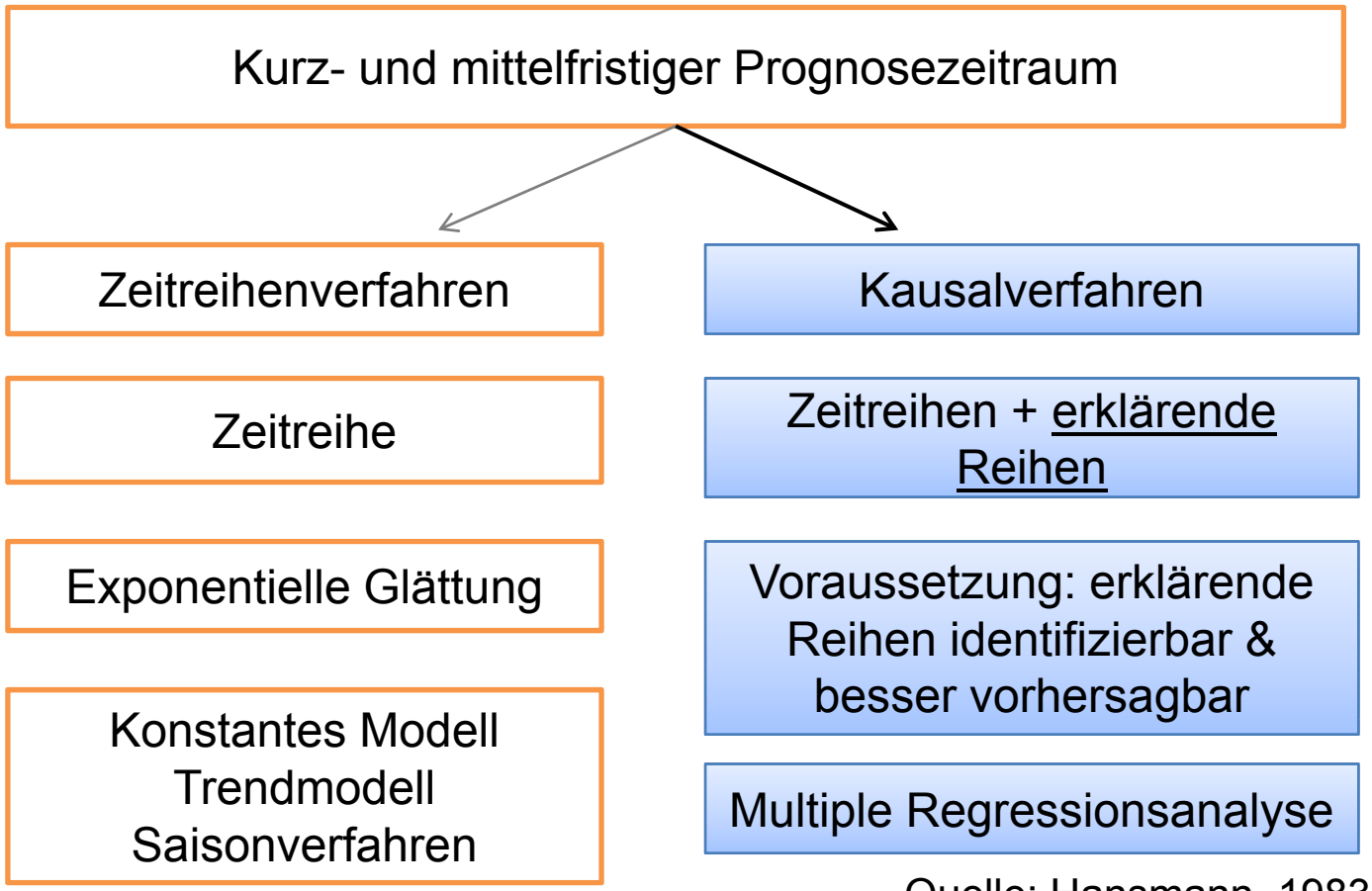
Prognosen

- Auftragsfertigung: Planung auf Basis von vorhanden und erwarteten Kundenaufträgen
  - keine Unsicherheit

- Produktion für den anonymen Markt: Planung auf Basis von Prognosen
  - Unsicherheit

- Prognosen sind dadurch gekennzeichnet:
  - aus den Vergangenheitsdaten
  - über die theoretischen Erklärungen der Vergangenheit
  - Annahmen bzw. Vorhersagen für die Zukunft abzuleiten.

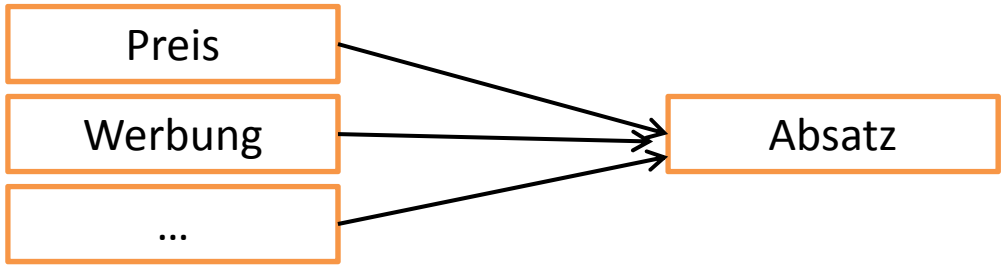
Quantitative Prognoseverfahren



Quelle: Hansmann, 1983, S.143

Regressionsanalyse

- Die Regressionsanalyse beschreibt den linearen Zusammenhang zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen.



- hilfreich wären folgende Fragestellungen:
  - Wie wirkt der Preis auf die Absatzmenge?
  - Welche Absatzmenge ist zu erwarten, wenn der Preis und gleichzeitig auch die Werbeausgaben um bestimmte Größen verändert werden?

Regressionsanalyse

- **Einfache Regression:** die endogene Variable wird durch einen exogenen Einflussfaktor erklärt.
- **Multiple Regression:** mindestens zwei unabhängige Variablen werden zur Prognose der untersuchenden Variable berücksichtigt.
- **Lineare Regressionsanalyse:** die Einflussfaktoren haben einen linearen Zusammenhang.
- **Nicht-lineare Regressionsanalyse:** die Einflussfaktoren werden in beliebiger nicht-linearer Funktionsform verknüpft.

Quelle: Höck, C. , Universität Hamburg (2007)

Agenda

I. Motive der Regressionsanalyse zur Prognose

II. Ablaufschritte der Regressionsanalyse

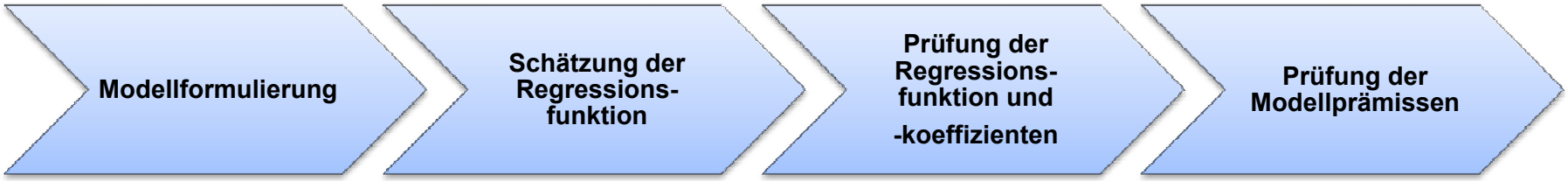
III. Anwendungsbeispiel in Excel

IV. Kritische Zusammenfassung

Ablaufschritte der Regressionsanalyse

Ablaufschritte der Regressionsanalyse

(1) Modellformulierung
(2) Schätzung der Regressionsfunktion
(3) Prüfung der Regressionsfunktion und -koeffizienten
(4) Prüfung der Modellprämissen



(1) Modellformulierung

- Was ist zu prognostizieren?
  - z.B. Absatzmenge, Umsatz, Gewinn, etc.
- Was sind die relevanten Einflussfaktoren dafür?
  - z.B. Werbung, Preis, Saison, etc.
- Welches die abhängige und welches die unabhängige(n) Variable(n)?
  - z.B. Wetter → Eisverkauf, nicht Eisverkauf → Wetter
- Prüfung der Verfügbarkeit der relevanten Daten
- Darstellung der Zielgrößenwerte im sog. zwei-dimensionalen Streudiagramm entweder über die Zeit oder in Abhängigkeit von der unabhängigen Variable
- Beobachtung der Streuung

(2) Schätzung der linearen multiplen Regressionsfunktion

Allgemein: $\hat{y}_t = b_0 + b_{1t} \cdot x_{1t} + b_{2t} \cdot x_{2t} + b_{3t} \cdot x_{3t} + \ldots + b_{jt} \cdot x_{jt} + \ldots + b_{Jt} \cdot x_{Jt}$

mit $\hat{y}_t$ = Schätzung der abhängigen Variable y für die Zeit t

b_0 = Konstantes Glied

b_j = Regressionskoeffizienten (j = 1, 2,…, J)

x_{jt} = Werte der unabhängigen Variablen

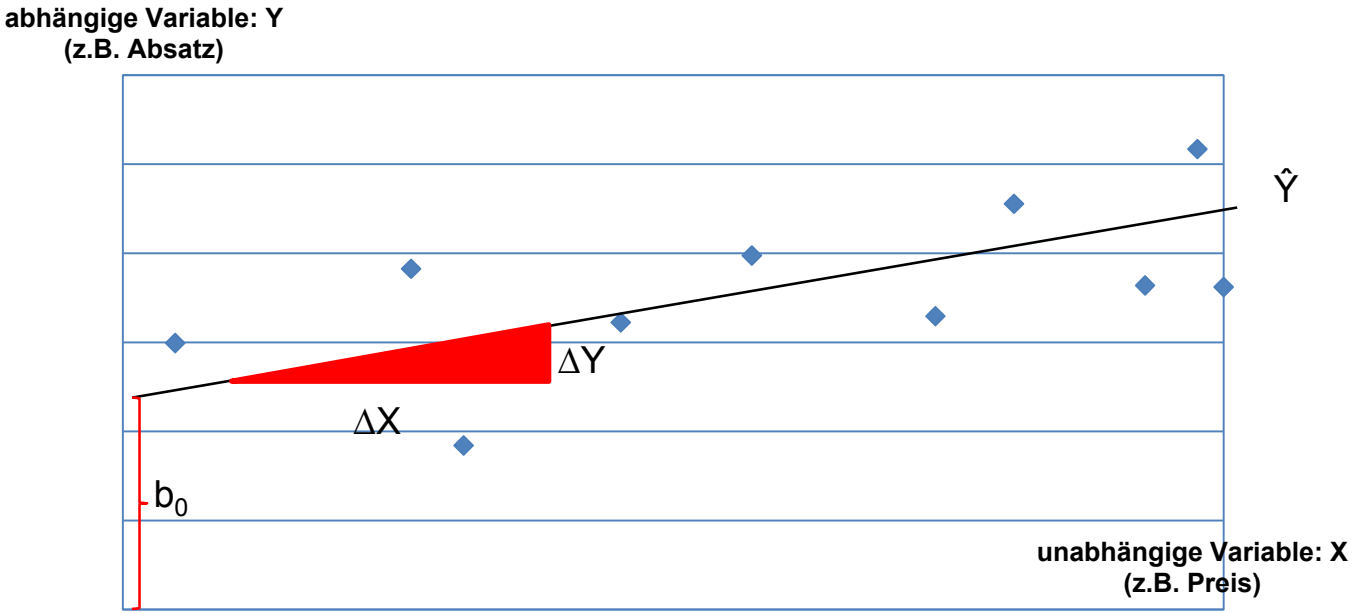
(j = 1, 2,…, J; t = 1, 2, …, T)

J = Zahl der unabhängigen Variablen

T = Zeithorizont (t = 1,……, T) bzw. Zahl der Beobachtungen

(2) Schätzung der einfachen Regressionsfunktion

Einfache lineare Regressionsfunktion: $\hat{Y}_t = b_0 + b_{1t} \cdot X_t$



Streudiagramm und Regressionsgerade (Quelle: Backhaus et al., 2003 S. 55)

(2) Zielfunktion der Regressionsanalyse

- Die Schätzung der Regressionsfunktion basiert auf der Methode der kleinsten Quadrate, um die quadrierten Prognosefehler zu minimieren.
  - tatsächlicher Wert: y_t
  - geschätzter Wert: $\hat{y}_t$
  - Prognosefehler: $e_t = y_t - \hat{y}_t$
- Zielfunktion der Regressionsanalyse:

$$\sum_{t=1}^{T} e_t^2 = \sum_{t=1}^{T} \left[y_t - \left(b_0 + b_1 x_{1t} + b_2 x_{2t} + \cdots + b_j x_{jt} + \cdots + b_J x_{Jt} \right) \right]^2 \rightarrow \min$$

- Damit können die einzelnen Regressionsparameter b_0, b_1, b_2,…, b_J ermittelt werden.

(2) Nicht-lineare Regressionsfunktion

Aber:

- Die KQ-Methode ist nur bei einer linearen Regressionsfunktion anzuwenden.
- Ein linearer Zusammenhang besteht erst dann, wenn die Punkte eng um eine „gedachte“ Gerade streuen.
- Eine nicht-lineare Beziehung ist auch denkbar für die bessere Anpassung zur Streuung.

(2) Schätzung der nicht-linearen Regressionsfunktion

- Linearisierung durch Transformation

  – z.B. eine exponentielle Funktion:
  $\hat{y} = ae^{bx}$
  – Logarithmieren:
  $\tilde{y} = \ln(y), \tilde{a} = \ln(a)$
  – Lineare Form:
  $\tilde{y} = \tilde{a} + bx \rightarrow$ einzelne Parameter zu ermitteln
  – Rücktransformation (Entlogarithmieren):
  $y = ae^{bx}$

(3) Güte der Prognose

- Fehlermaße:
  - Mittlere absolute Abweichung (MAA)
  - Mittlere quadrierte Abweichung (MQA)
  - Wurzel der MQA (WMQA)
  - Mittlere prozentuale Abweichung (MPA)

  externe Qualität

  Vergleich vom tatsächlichen mit dem geschätzten Wert

- Gütekriterien für lineare Regressionsanalysen:
  - Bestimmtheitsmaß (R^2)
  - F-Wert
  - t-Werte

  interne Qualität

  Die Qualität der Schätzung an sich

(3) Bestimmtheitsmaß R^2

- Ein Gütemaß für den linearen Zusammenhang zwischen der abhängigen und den unabhängigen Variablen,

  d.h. wie gut sich die Schätzung an die Daten anpasst.

$$\sum_{t=1}^{T}(y_t-\bar{y})^2=\sum_{t=1}^{T}(\hat{y}_t-\bar{y})^2+\sum_{t=1}^{T}(y_t-\hat{y}_t)^2$$

Gesamtstreuung = erklärte Streuung + nicht erklärte Streuung

$$R^2=\frac{\sum_{t=1}^{T}(\hat{y}_t-\bar{y})^2}{\sum_{t=1}^{T}(y_t-\bar{y})^2}=1-\frac{\sum_{t=1}^{T}(y_t-\hat{y}_t)^2}{\sum_{t=1}^{T}(y_t-\bar{y})^2}=1-\frac{\sum_{t=1}^{T}e_t^2}{\sum_{t=1}^{T}(y_t-\bar{y})^2}$$

$0 \le R^2 \le 1$

z.B. $R^2 = 0{,}78$ hat eine Aussage, dass

die 78 % der Gesamtstreuung durch das Schätzmodell erklärt werden.

(3) Korrigiertes Bestimmtheitsmaß

- Das Bestimmtheitsmaß erhöht sich durch die Zahl der unabhängigen Variablen, auch bei der Aufnahme von irrelevanten Variablen.

- Das korrigiertes Bestimmtheitsmaß ($R^2_{korr.}$) berücksichtigt zusätzlich die Zahl der Regressoren und kann auch durch Aufnahmen von weiteren Regressoren abnehmen.

(3) F-Test

- **Ziel:**

  Prüfung, ob der Wert R^2 sich zufällig ergeben hat.
- **Vorgehen:**
  - Formulierung einer Nullhypothese H_0:

    „Es besteht kein kausaler Zusammenhang zwischen der abhängigen und der unabhängigen Variablen“

    (H_0: $b_1=b_2=b_3=\ldots=b_J=0$)
  - Berechnung empirisches F-Wertes
  - Wahl einer Vertrauenswahrscheinlichkeit
  - Vergleich mit dem kritischen Wert (aus der F-Tabelle)

Überschreitet der F-Wert mit dem kritischen Wert, ist die Nullhypothese zu verwerfen, d.h. mindestens eine unabhängige Variable ist signifikant.

(3) t-Werte

- Wenn der F-Test ergeben hat, dass nicht alle Regressionskoeffizienten gleich Null sind, werden diese jetzt einzeln überprüft.

- Die Nullhypothese: (H_0: $b_j = 0$) wird auch hier wieder getestet.

- Der empirische t-Wert im Absolutbetrag ist mit dem kritischen Wert von der t-Verteilung mit der Vertrauenswahrscheinlichkeit und der Freiheitsgrade zu vergleichen.

- $|t_{emp}| > t_{tab} \rightarrow H_0$ wird verworfen $\rightarrow$ Einfluss ist signifikant

(4) Prüfung der Modellprämissen

Prämisse	Prämissenverletzung	Konsequenzen
Linearität der Parametern	Nichtlinearität	Verzerrung der Schätzwerte
Vollständigkeit der Störgrößen	Unvollständigkeit	
Homeskedastizität der Störgrößen	Heteroskedastizität	Ineffizienz
Unabhängigkeit der Störgrößen	Autokorrelation	
Keine lineare Abhängigkeit zwischen der unabhängigen Variablen	Multikollinearität	Verminderte Präzision der Schätzwerte
Normalverteilung der Störgrößen	Nicht normalverteilt	Ungültigkeit der Signifikanz (F-Test und t-Test)

Quelle: Backhaus et al., 2003, S.92

(4) Nichtlinearität

- Das lineare Regressionsmodell fordert, dass die Beziehung linear in den Parametern ist.
  - Dies lässt sich anhand bestimmter Transformation erreichen.

- Nichtlinearität besteht auch, wenn Strukturbrüche wie Niveau- und Trendänderung bestehen.
  - Diese sind durch die sog. Dummy-Variablen zu berücksichtigen.

(4) Unvollständigkeit

- Eine vorhergehende Prüfung muss feststellen, ob das Modell vollständig ist.
  - Die sachlogische Prüfung sollte theoretisch fundiert erfolgen.
  - Es muss festgestellt werden, ob alle wichtigen relevanten Variablen in das Modell aufgenommen wurden und
  - keine überflüssige irrelevante Variablen enthalten sein.

  ➢ Eine theoriegeleiteten Erhebung vermeidet Fehler bei der Auswahl der Variablen.

(4) Heteroskedastizität

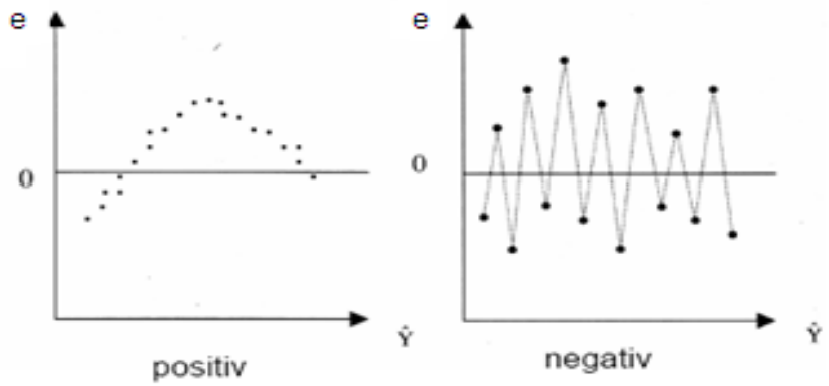
- In der Regressionsanalyse sollten alle Residualgrößen die gleiche Varianz aufweisen.

  - Zur Überprüfung:
    - Visuelle Inspektion der Residuen
    - Rechnerische Methode auch möglich
  - Konsequenz:
    - Heteroskedastizität führ zu Ineffizienz der Schätzung und verfälscht den Standardfehler der Regressionskoeffizienten (→ t-Wert).
  - Begegnung von Heteroskedastizität:
    - Transformation der abhängigen Variablen oder der gesamten Regressionsgleichung

(5) Autokorrelation

- Liegt vor, wenn Residuen nicht unkorreliert sind.
- Tritt vor allem Zeitreihen auf.

  – Zur Überprüfung:
    - Visuelle Inspektion der Residuen
    - Rechnerische Methode: Durbin-Watson-Test
  – positive und negative Autokorrelation



(6) Multikollinearität

- Liegt bei linearer Abhängigkeit der Regressoren vor.

Aber:

- immer ein gewisser Grad an linearer Abhängigkeit.

Deshalb:

- erst dann problematisch, wenn eine starke Abhängigkeit zwischen den unabhängigen Variablen besteht.
  - Zur Identifikation:
    - Korrelationsmatrix (nahe $|1| \rightarrow$ernsthafte Multikollinearität)
    - Rechnerische Methoden
  - Behebung:
    - Faktoranalyse
    - Entfernung einer belasteten Variable unter Berücksichtigung der Vollständigkeit

(7) Nicht-Normalverteilung

- Das statistische Modell der linearen Regression beruht auf der Annahme der Normalverteilung der Störgrößen.

- Ist relevant für die Durchführung der statischen Tests (F-Test, t-Test).

Agenda

I. Motive der Regressionsanalyse zur Prognose

II. Ablaufschritte der Regressionsanalyse

III. Anwendungsbeispiel in Excel

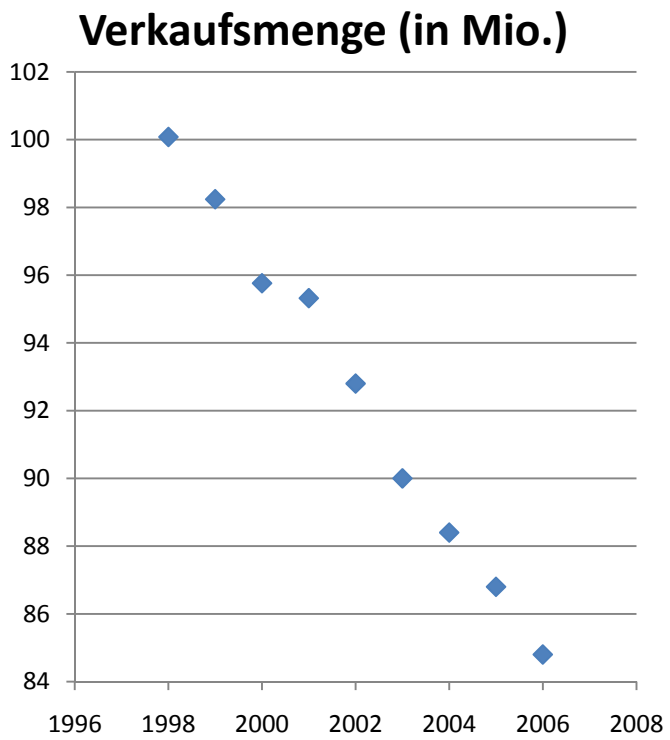
IV. Kritische Zusammenfassung

Vorgehensweise

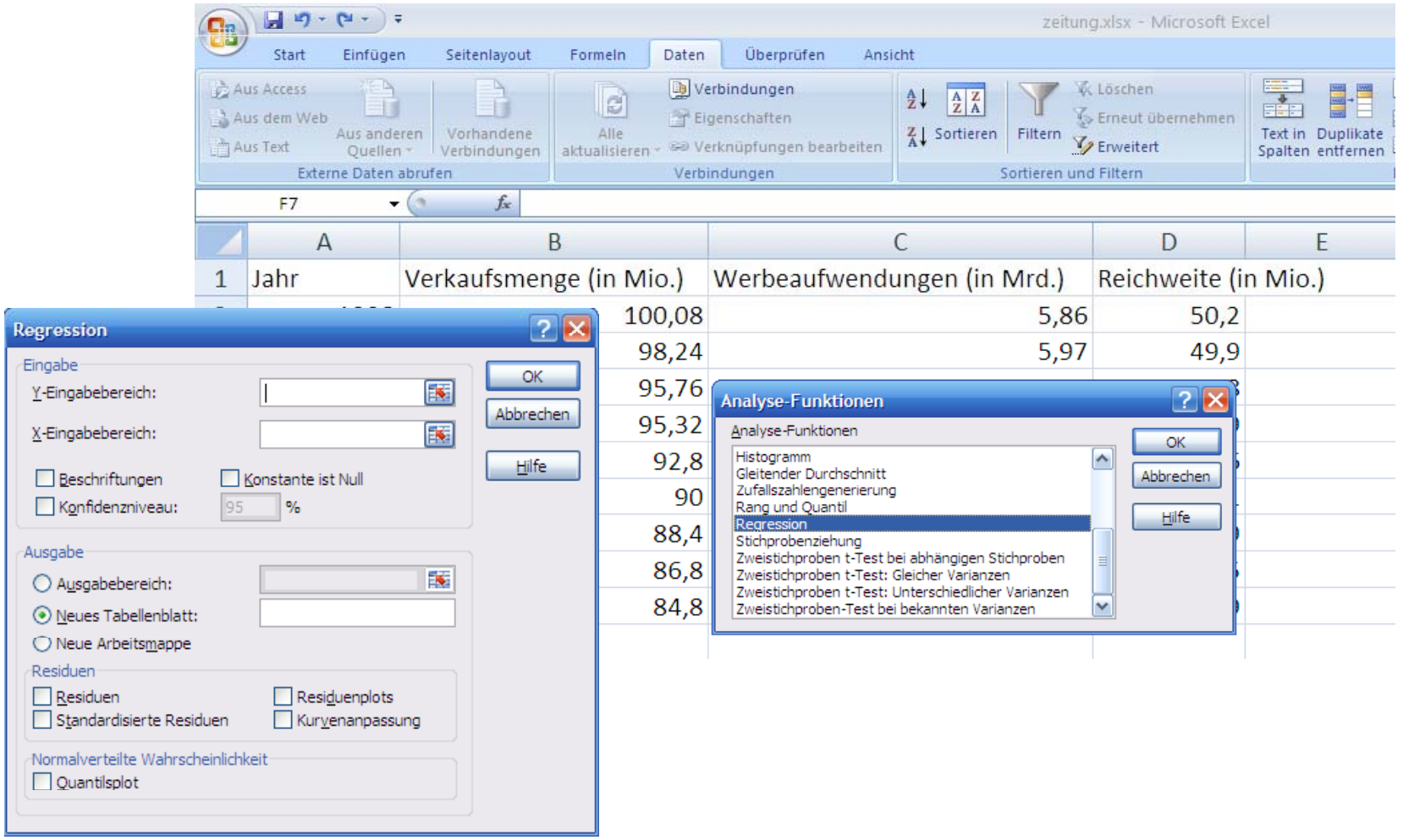
- Formulierung des Modells
  - Verkaufsmenge der Tageszeitungen in Deutschland zwischen 1998 und 2006
  - Unabhängigen Variablen: Werbeaufwendungen und Reichweiten
  - Datenerhebung anhand der Daten von Bundesverband Deutscher Zeitungsverleger e.V.
- Herstellung einer Regressionsfunktion
- Überprüfung der Kriterien und Interpretation
- Prognose

Daten

Jahr	Verkaufs-menge (in Mio.)	Werbeauf-wendungen (in Mrd.)	Reich-weite (in Mio.)
1998	100,08	5,86	50,2
1999	98,24	5,97	49,9
2000	95,76	6,52	49,8
2001	95,32	5,63	49,9
2002	92,8	4,99	49,6
2003	90	4,43	49,1
2004	88,4	4,51	49
2005	86,8	4,48	48,5
2006	84,8	4,53	47,9



Schätzung der Regressionsfunktion in Excel



Ergebnis in Excel

AUSGABE: ZUSAMMENFASSUNG

Regressions-Statistik	
Multipler Ko	0,974163859
Bestimmthe	0,948995225
Adjustiertes	0,931993633
Standardfeh	1,379623932
Beobachtung	9

ANOVA

	Freiheitsgrade (df)	*adratsummen (S*	*lere Quadratsumme (*	*Prüfgröße (F)*	*F krit*
Regression	2	212,4838268	106,2419134	55,81802231	0,00013269
Residue	6	11,42017316	1,903362194		
Gesamt	8	223,904			

	Koeffizienten	Standardfehler	t-Statistik	P-Wert	Untere 95%	Obere 95%	Untere 95,0%	Obere 95,0%
Schnittpunkt	-171,6668297	47,42912217	-3,619439321	0,011104625	-287,721711	-55,6119488	-287,721711	-55,6119488
Werbeaufwe	1,863258967	0,995674386	1,87135372	0,110471303	-0,57306848	4,29958642	-0,57306848	4,29958642
Reichweite (	5,158317993	1,042151386	4,949682034	0,002578767	2,60826542	7,70837057	2,60826542	7,70837057

Prüfung der Prämissen

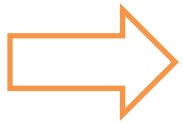
- Multikollinearität: Korrelationsmatrix

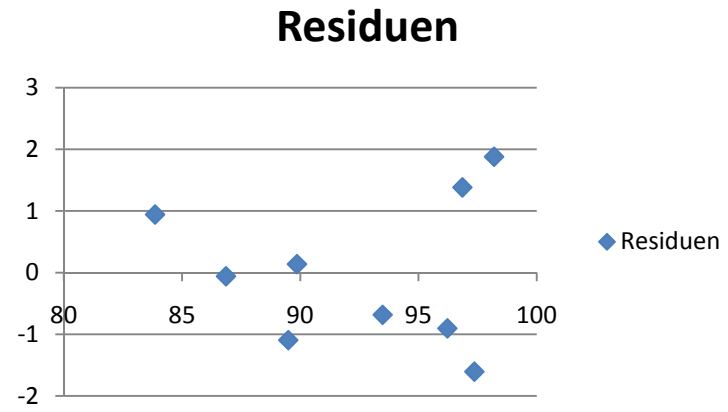
	Verkaufsmenge (in	*fwendungen*	*chweite (in Mio.)*
Verkaufsmer	1		
Werbeaufwe	0,86065744	1	
Reichweite (i	0,95876261	0,78654416	1

- Heteroskedastizität & Autokorrelation: visuelle Inspektion

AUSGABE: RESIDUENPLOT

Beobachtung	*Verkaufsme.*	*Residuen*
1	98,199431	1,88056896
2	96,8568941	1,38310587
3	97,3658548	-1,6058548
4	96,2233861	-0,9033861
5	93,4834049	-0,6834049
6	89,8608209	0,13917908
7	89,4940498	-1,0940498
8	86,8589931	-0,0589931
9	83,8571652	0,94283477





Interpretation & Prognose

- $\hat{y}_t = -171{,}66 + 1{,}86 \cdot Werbeaufb_t + 5{,}158 \cdot Reichweite_t$
- 94,9% der Gesamtstreuung werden mit diesem Modell erklärt und das ist eine gute Schätzung.
- F-Wert besagt, dass es eine signifikante Beziehung zwischen der abhängigen und unabhängigen Variablen besteht.
- Die Werbeaufwendung hat einen positiven Einfluss, aber sie ist nicht signifikant.
- Die Reichweite ist signifikant und hat einen positiven Effekt, d.h. wenn sich die Reichweite um eine Einheit erhöht, steigt die Verkaufsmenge um 5,158 Einheiten. → **plausibel**
- Bei der Fortschreibung der Werten von den unabhängigen Variablen ergibt sich die Verkaufsmenge von 83,834 in 2007.

Probleme

- Ineffizienz aufgrund:
  - Multikollinearität:
    - Die Korrelationskoeffiziente von 0,79 ist zu hoch.
  - Heteroskedastizität:
    - Mangelnde Konstanz der Varianz aller Residualgrößen
  - Autokorrelation:
    - Negative Autokorrelation

- Fortschreibung der vorjährigen Daten als Schätzwert für die unabhängigen Variablen

- Berücksichtigung der insignifikanten Variable

Agenda

I. Motive der Regressionsanalyse zur Prognose

II. Ablaufschritte der Regressionsanalyse

III. Anwendungsbeispiel in Excel

IV. Kritische Zusammenfassung

Vorteile

- Die statistische Eigenschaft
- Damit Berücksichtigung jeder Art von kausalen Beziehungen
- Entwicklung der benutzerfreundlichen Softwares
- Die theoretischen und statistischen Implikationen der Methode sind weitgehend geklärt.
- Die Probleme durch Verletzung der Grundannahmen sind durch zusätzliche Überlegungen prinzipiell lösbar.
- Für die Prognosewerte können Konfidenzintervalle angegeben werden

Nachteile

- Erklärende Reihen bekannt, identifizierbar und besser vorhersagbar
  - rechtzeitige und leichtere Erkennung als abhängige Variable
- Alle relevante Einflussfaktoren vorhanden
  - z.B. Schwierigkeit bei der Gewinnung von Konkurrenzdaten
- Kosten und Zeit für den Datenumfang
- Überwachung und Aktualisierung der Daten

Vielen Dank für Eure Aufmerksamkeit!