

**Sönke Albers
Bernd Skiera**

Regressionsanalyse

Vorabversion des Beitrags:

Albers, S. / Skiera, B. (1999), "Regressionsanalyse", in: Herrmann, A. / Homburg, C. (Hrsg.), "Marktforschung. Grundlagen - Methoden - Anwendungen", Wiesbaden, S. 205-236.

Februar 1998

Prof. Dr. Sönke Albers, Lehrstuhl für Betriebswirtschaftslehre, insbesondere Marketing, Christian-Albrechts-Universität zu Kiel, Olshausenstr. 40, 24098 Kiel, Tel.: 0431/880-1541, Fax: 0431/880-1166, E-Mail: albers@bwl.uni-kiel.de, <http://www.bwl.uni-kiel.de/bwlinstitute/Marketing/index.html>

Prof. Dr. Bernd Skiera, Lehrstuhl für Betriebswirtschaftslehre, insbesondere Electronic Commerce, Johann Wolfgang Goethe-Universität Frankfurt am Main, Mertonstr. 17, 60054 Frankfurt am Main, Tel. 069/798-22378, Fax: 069/798-28973, E-Mail: skiera@wiwi.uni-frankfurt.de, URL: <http://www.ecommerce.wiwi.uni-frankfurt.de/>

Inhaltsverzeichnis

| | |
|---|----|
| 1 Zielsetzung | 1 |
| 2 Mathematisch-statistische Erläuterung des Verfahrens | 1 |
| 2.1 Problemstellung | 1 |
| 2.2 Zielfunktion und Schätzung der Regressionskoeffizienten | 2 |
| 2.3 Anpassungsgüte | 4 |
| 2.4 Signifikanzprüfungen | 5 |
| 2.5 Standardisierung der Parameter | 7 |
| 2.6 Interpretation der Ergebnisse | 8 |
| 2.7 Ergebnisse des Beispiels | 8 |
| 2.8 Annahmen | 10 |
| 3 Vorgehensweise | 11 |
| 3.1 Transformation der Variablen | 11 |
| 3.2 Effizienz der Schätzer | 14 |
| 3.3 Prüfung auf Multikollinearität | 14 |
| 3.4 Prüfung auf Autokorrelation | 17 |
| 3.5 Prüfung auf Heteroskedastizität | 20 |
| 3.6 Identifizierung von Ausreißern | 22 |
| 4 Implikationen der Analyse | 23 |
| 5 Software | 24 |
| 6 Ausblick | 25 |
| 7 Literatur | 26 |

1 Zielsetzung

Die lineare Regressionsanalyse ist eines der am häufigsten angewendeten statistischen Analyseverfahren (Hair et al. 1992, Backhaus et al. 1997). Sie untersucht die lineare Abhängigkeit zwischen einer metrisch skalierten abhängigen Variablen (auch endogene Variable, Prognosevariable oder Regressand genannt) und einer oder mehreren metrisch skalierten unabhängigen Variablen (auch exogene Variablen, Prädiktorvariablen sowie Regressoren genannt). Mit Hilfe der linearen Regressionsanalyse können somit Zusammenhänge aufgedeckt und Prognosen erstellt werden. Deswegen wird die Regressionsanalyse im Marketing vor allem für die Schätzung des Zusammenhangs zwischen der Absatzmenge oder dem Marktanteil und solchen Marketing-Instrumenten wie dem Preis und dem Werbebudget (sogenannte Reaktionsfunktion) eingesetzt. Daneben findet man häufig aber auf vielfältige andere Einsatzmöglichkeiten für die Regressionsanalyse wie z.B. für die Schätzung der Abhängigkeit des Image eines Produkts von Einstellungen bestimmter Zielgruppen oder des Zusammenhangs zwischen Markenloyalität oder Kauf Erfahrungen und demographischer Merkmalen von Konsumenten. In diesem Beitrag stellen wir die Schätzung von Reaktionsfunktionen in den Vordergrund. Dafür greifen wir auf ein Zahlenbeispiel zurück, an dem das Verfahren und die Vorgehensweise beim Einsatz der Regressionsanalyse erläutert wird.

2 Mathematisch-statistische Erläuterung des Verfahrens

2.1 Problemstellung

Die Grundidee der linearen Regressionsanalyse wird nachfolgend an den in Tabelle 1 dargestellten Daten eines (fiktiven) Unternehmens erläutert. Die Tabelle 1 enthält die Absatzmengen einer Periode von 16 zufällig aus insgesamt 100 Verkaufsbezirken eines Unternehmens ausgewählten Bezirken, die sich nur hinsichtlich der Höhe der eingesetzten Marketing-Instrumente Preis, Werbebudget und Anzahl der Außendienstmitarbeiter (kurz ADM) unterscheiden. Der Bezirk 17 und das Marketing-Instrument der Mailings werden zunächst nicht berücksichtigt. Die Marketing-Managerin beauftragt ihren Assistenten mit der Analyse der Daten, wobei sie sich insbesondere dafür interessiert, ob:

- der Preis richtig gesetzt ist,
- die Investitionen in Werbung und den Außendienst sinnvoll sind und

- die Aufteilung des Budgets zwischen den beiden Instrumenten der Werbung und des Außendienstes gut gewählt ist,

wenn das Unternehmen mit Stückkosten in Höhe von 30 DM rechnet und ein Außendienstmitarbeiter Kosten in Höhe von 120.000 DM verursacht.

Tabelle 1: Verkaufsinformationen

| Bezirk | Absatzmenge | Anzahl ADM | Preis | Werbebudget | Anzahl Mailings |
|--------|-------------|------------|-------|-------------|-----------------|
| 1 | 81.996 | 7 | 49 | 228.753 | 7.106 |
| 2 | 91.735 | 5 | 46 | 370.062 | 4.733 |
| 3 | 70.830 | 4 | 50 | 297.909 | 3.734 |
| 4 | 101.192 | 6 | 45 | 271.884 | 6.152 |
| 5 | 78.319 | 6 | 51 | 299.919 | 5.734 |
| 6 | 105.369 | 7 | 47 | 367.644 | 6.640 |
| 7 | 68.564 | 3 | 47 | 241.362 | 3.115 |
| 8 | 95.523 | 7 | 46 | 244.575 | 6.859 |
| 9 | 88.834 | 7 | 49 | 296.100 | 6.905 |
| 10 | 89.511 | 5 | 46 | 372.498 | 5.142 |
| 11 | 107.836 | 6 | 45 | 359.511 | 6.196 |
| 12 | 83.310 | 7 | 50 | 324.837 | 6.801 |
| 13 | 67.817 | 4 | 50 | 288.303 | 3.965 |
| 14 | 59.207 | 6 | 54 | 289.470 | 5.830 |
| 15 | 81.410 | 6 | 52 | 363.501 | 6.124 |
| 16 | 71.431 | 3 | 46 | 361.974 | 2.509 |
| 17 | 119.000 | 3 | 45 | 240.000 | - |

2.2 Zielfunktion und Schätzung der Regressionskoeffizienten

Mit der Regressionsanalyse wird nachfolgend der Einfluß der metrisch skalierten unabhängigen Variablen (hier Anzahl ADM, Preis und Werbebudget) auf die metrisch skalierte abhängige Variable (hier Absatzmenge) untersucht. Im Falle der in diesem Beitrag im Vordergrund stehenden linearen Regressionsanalyse sieht die Regressionsgleichung wie folgt aus, wobei die Bezirke nachfolgend zur Verallgemeinerung als Beobachtungen bezeichnet werden:

$$(1) \quad y_i = b_0 + \sum_{k \in K} b_k \cdot x_{i,k} + e_i \quad (i \in I),$$

wobei:

b_0 : Konstante der Regressionsfunktion,

| | |
|-------------|---|
| b_k : | Regressionskoeffizient zur Abbildung des Einflusses der k-ten unabhängigen Variablen, |
| e_i : | Residualgröße der i-ten Beobachtung, |
| I: | Indexmenge der Beobachtungen, |
| K: | Indexmenge der unabhängigen Variablen, |
| $x_{i,k}$: | Wert der i-ten Beobachtung für die k-te unabhängige Variable, |
| y_i : | Wert der i-ten Beobachtung für die abhängige Variable. |

Die Werte der abhängigen Variablen y_i und der unabhängigen Variablen $x_{i,k}$ sind beobachtbar (hier in Form der Werte für die in Tabelle 1 dargestellten Bezirke), während die Parameter der Regressionsfunktion b_0 und b_k ($k \in K$) sowie aller Residualgrößen e_i ($i \in I$), die mitunter auch als Residuen, Störgrößen oder Fehlerterme bezeichnet werden, zu schätzen sind. Die Residualgröße e_i beschreibt dabei die Abweichung zwischen dem tatsächlichen Wert der abhängigen Variablen y_i für die i-te Beobachtung und dem auf Basis der Parameter der Regressionsfunktion geschätzten Wert der abhängigen Variablen \hat{y}_i .

$$(2) \quad \hat{y}_i = b_0 + \sum_{k \in K} b_k \cdot x_{i,k} \quad (i \in I).$$

Das Ziel der Regressionsanalyse besteht darin, die Parameter der Regressionsfunktion b_0 und b_k ($k \in K$) so zu schätzen, daß die Summe der quadrierten Abweichungen zwischen dem tatsächlichen Wert der i-ten Beobachtung y_i und deren geschätzten Wert \hat{y}_i , also die Summe der quadrierten Residualgrößen, minimiert wird. Dies wird auch als Methode der kleinsten Quadrate bezeichnet (Hansen 1993, S. 53, Schneeweiß 1990, S. 21). Es ergibt sich somit folgende Zielfunktion:

$$(3) \quad \sum_{i \in I} e_i^2 = \sum_{i \in I} (y_i - \hat{y}_i)^2 = \sum_{i \in I} \left(y_i - b_0 - \sum_{k \in K} b_k \cdot x_{i,k} \right)^2 \rightarrow \min!$$

Die Betrachtung der quadrierten Abweichungen bietet den Vorteil, daß große Abweichungen eine höhere Bedeutung erfahren als kleinere Abweichungen und die Lösung der Zielfunktion (3) algorithmisch einfach zu ermitteln ist (Pindyck/Rubinfeld 1991, S. 6). Da die Schätzung der Regressionsfunktion in den allermeisten Fällen mit Hilfe eines der in Kapitel 5 dargestellten Software-Programme geschieht und wenig zusätzliche Einsichten in das Problem gestattet, wird an dieser Stelle auf die Herleitung der Lösung der Zielfunktion (3) verzichtet. Ausführliche und didaktisch gelungene Beschreibungen

des Lösungsverfahrens finden sich beispielsweise in Pindyck/Rubinfeld (1991), Hansen (1993) oder Koutsoyannis (1977).

2.3 Anpassungsgüte

Zur Beurteilung der Anpassungsgüte der linearen Regressionsanalyse läßt man sich von der Überlegung leiten, daß ohne die Kenntnis der unabhängigen Variablen die beste Schätzung des zu erwartenden Werts der abhängigen Variablen durch die Bestimmung des Mittelwerts der abhängigen Variablen erfolgt. Die Güte einer Regressionsanalyse wird nun daran gemessen, um wieviel sich die Aussage durch die Betrachtung von unabhängigen Variablen gegenüber der ausschließlichen Betrachtung der abhängigen Variablen und der damit verbundenen "einfachen" Schätzung in Form des Mittelwerts verbessert. Gemessen wird dies durch das Bestimmtheitsmaß R^2 , das den Anteil der durch die Regressionsgleichung erklärten Varianz an der Varianz der "einfachen" Aussage in Form des Mittelwerts erfaßt:

$$(4) \quad R^2 = \frac{\sum_{i \in I} (\hat{y}_i - \bar{y})^2}{\sum_{i \in I} (y_i - \bar{y})^2}$$

Mit der Regressionsgleichung wird immer eine Varianz erklärt, die mindestens so hoch wie die Varianz der "einfachen" Schätzung ist, weil die Regressionsgleichung im ungünstigsten Fall den Einfluß der unabhängigen Variablen unberücksichtigt läßt und mit der Konstanten der Regressionsgleichung weiterhin "nur" den Mittelwert der abhängigen Variablen schätzt. Üblicherweise wird durch das Heranziehen der unabhängigen Variablen eine Verbesserung der Schätzung erreicht, die im Extremfall dazu führt, daß die geschätzten Werte der unabhängigen Variablen den beobachteten Werten entsprechen. In diesem Fall wird die gesamte Varianz der "einfachen" Schätzung erklärt. Das Bestimmtheitsmaß kann also zwischen 0% und 100% liegen. Negative Werte für das Bestimmtheitsmaß können sich lediglich dadurch ergeben, daß, wie von den meisten Statistikprogrammen angeboten, in der Regressionsgleichung auf die Konstante verzichtet wird, da in einem solchen Fall nicht zwangsläufig mit der "einfachen" Schätzung des Mittelwerts gleichgezogen werden kann.

Diese Überlegungen sollten auch deutlich machen, daß mit der Aufnahme einer zusätzlichen unabhängigen Variablen niemals eine Verschlechterung des Bestimmtheitsmaßes erfolgen kann, da der Erklärungsgehalt der zusätzlichen Variablen im schlechtesten Falle

Null ist. Im Extremfall ergibt sich ein lineares Gleichungssystem, in dem die Anzahl der Gleichungen (hier die Anzahl der Beobachtungen) der Anzahl der zu schätzenden Parameter (hier Konstante b_0 plus die Anzahl der Regressionskoeffizienten b_k) entspricht. Es ergibt sich dann mit 100% der größtmögliche Wert für das Bestimmtheitsmaß. Um diesen Effekt des stets ansteigenden Bestimmtheitsmaßes zu vermeiden, kann das folgende korrigierte Bestimmtheitsmaß herangezogen werden (Pindyck/Rubinfeld 1991, S. 78):

$$(5) \quad R_{\text{kor}}^2 = R^2 - \frac{|K| \cdot (1 - R^2)}{|I| - |K| - 1},$$

wobei:

$|I|$: Anzahl der Elemente der Indexmenge der Beobachtungen,
 $|K|$: Anzahl der Elemente der Indexmenge der unabhängigen Variablen
(entspricht der Anzahl der Regressionskoeffizienten).

Da sowohl Zähler als auch Nenner des Bruchs positiv sind, kann das korrigierte Bestimmtheitsmaß bestenfalls genau so groß wie das (unkorrigierte) Bestimmtheitsmaß sein.

2.4 Signifikanzprüfungen

Daten liegen häufig nicht für die Grundgesamtheit, sondern nur für eine Stichprobe vor. Dies gilt beispielsweise für einen Großteil der von Marktforschungsunternehmen durchgeführten Befragungen sowie der in Panels und Labor- oder Feldexperimenten erhobenen Daten. Auch in unserem Fall stellen die 16 Bezirke nur eine Stichprobe dar. In einem solchen Fall interessiert, inwieweit die auf Basis der Stichprobe festgestellten Ergebnisse auch für die Grundgesamtheit Gültigkeit haben. Für eine derartige Prüfung muß eine Verteilungsannahme für die Residualgrößen unterstellt werden. Im Falle der Regressionsanalyse (und vieler anderer Verfahren) ist dies die Normalverteilung. Diese Annahme beruht auf der Aussage des zentralen Grenzwertsatzes, der besagt, daß das (gewogene) Mittel von stochastisch unabhängigen Zufallsstichproben aus einer beliebigen Verteilung mit zunehmender Anzahl an Zufallsstichproben wiederum normalverteilt ist (Hansen 1993, S. 68).

Unter Zugrundelegung dieser Annahme können Signifikanzprüfungen für die Regressionskoeffizienten durchgeführt werden. Dazu werden die Irrtumswahrscheinlichkeiten dafür errechnet, daß die betrachteten Regressionskoeffizienten ungleich Null sind. Dies bedeutet dann, daß der Zusammenhang zwischen den betrachteten unabhängigen Va-

riablen und der abhängigen Variablen auch in der Grundgesamtheit mit einer bestimmten Wahrscheinlichkeit, der Differenz zwischen Eins und der Irrtumswahrscheinlichkeit, besteht. Die Signifikanzprüfung erfolgt dann dadurch, daß die ermittelte Irrtumswahrscheinlichkeit mit dem vorgegebenen Signifikanzniveau (häufig 5%) verglichen wird. Wenn die Irrtumswahrscheinlichkeit kleiner als dieses vorgegebene Signifikanzniveau ist, dann spricht man von einem signifikanten Einfluß.

Die Signifikanzprüfung des in der Grundgesamtheit vorliegenden Zusammenhangs zwischen allen unabhängigen Variablen und der abhängigen Variablen wird dabei mit Hilfe des folgenden F-Tests mit $|K|$ und $(|I| - |K| - 1)$ Freiheitsgraden durchgeführt:

$$(6) \quad F_{\text{emp}} = \frac{\frac{R^2}{|K|}}{\frac{1 - R^2}{|I| - |K| - 1}},$$

Für diesen empirischen F-Wert wird auf der Basis der F-Verteilung die Irrtumswahrscheinlichkeit dafür errechnet, daß der F-Wert von Null verschieden ist. Ist diese Wahrscheinlichkeit kleiner als das vorgegebene Signifikanzniveau (z.B. 5%), so liegt ein signifikanter Einfluß vor.

Wenn mit Hilfe des F-Tests ein signifikanter Zusammenhang zwischen der Gesamtheit der unabhängigen Variablen und der abhängigen Variablen festgestellt wird, sollte noch geprüft werden, ob auch jeder einzelne Regressionskoeffizient signifikant ist. Dies kann mit Hilfe des folgenden t-Tests mit $(|I| - |K| - 1)$ Freiheitsgraden erfolgen, der auch dahingehend erweitert werden kann, daß der Wert des Regressionskoeffizienten nicht gegen den Wert Null, sondern gegen einen anderen Wert geprüft wird (Pindyck/Rubinfeld 1991, S. 76 ff.).

$$(7) \quad t_{k,\text{emp}} = \frac{b_k}{s_k} \quad (k \in K),$$

wobei:

s_k : Geschätzter Standardfehler des k-ten Regressionskoeffizienten,

$t_{k,\text{emp}}$: Empirischer t-Wert für den k-ten Regressionskoeffizienten.

Für diesen empirischen t-Wert wird auf der Basis der t-Verteilung die Irrtumswahrscheinlichkeit dafür errechnet, daß der t-Wert von Null verschieden ist. Ist diese Wahrscheinlichkeit kleiner als das vorgegebene Signifikanzniveau (z.B. 5%), so ist der Einfluß signifikant.

2.5 Standardisierung der Parameter

Häufig möchte man die Einflußstärke der unabhängigen Variablen miteinander vergleichen. Dazu bietet sich der unmittelbare Vergleich der Parameterwerte nur in den seltensten Fällen an, da die unabhängigen Variablen meistens unterschiedliche Größenordnungen aufweisen. Deshalb wird häufig auf eine Betrachtung der standardisierten Regressionskoeffizienten ausgewichen, da so eine Korrektur der unterschiedlichen Größenordnungen der Variablen vorgenommen wird. Diese standardisierten Regressionskoeffizienten werden durch die Multiplikation des (unstandardisierten) Regressionskoeffizienten b_k mit der Standardabweichung der dazugehörigen unabhängigen Variablen σ_{x_k} und anschließender Division mit der Standardabweichung der abhängigen Variablen σ_y errechnet:

$$(8) \quad \text{beta}_k = b_k \cdot \frac{\sigma_{x_k}}{\sigma_y} \quad (k \in K).$$

Wären vor der Durchführung der Regressionsanalyse die abhängige und die unabhängigen Variablen standardisiert worden, dann würden sich die Regressionskoeffizienten beta_k und b_k gleichen. Der Vergleich der absoluten Werte aller standardisierten Regressionskoeffizienten zeigt, wie stark der Einfluß der einzelnen unabhängigen Variablen ist. Hohe absolute Werte deuten dabei einen stärkeren Einfluß an.

Die Betrachtung der standardisierten Regressionskoeffizienten wird kritisiert, wenn die Standardabweichungen der unabhängigen Variablen beeinflussbar sind. So würde unser Unternehmen, wenn es Preise stark, Werbung aber nur wenig variiert hätte, aufgrund der hohen Standardabweichung der Preise (siehe Gleichung (8)) mit Hilfe der standardisierten Regressionskoeffizienten einen starken Einfluß des Preises feststellen. Wenn es dagegen Preise nur wenig, Werbung aber stark variiert hätte, hätte es tendenziell das gegenteilige Ergebnis festgestellt (Wittink 1988, S. 237 ff.). Deswegen ist der Vergleich des Einflusses der unabhängigen Variablen durch die Ermittlung der Elastizitäten der unabhängigen Variablen auf die abhängige Variable vielfach sinnvoller. Diese Elastizitäten sind dimensionslos und geben an, welche relative Veränderung sich bei der abhängigen Variablen durch eine relative Veränderung der unabhängigen Variablen ergibt. Im Falle einer linearen Regressionsgleichung ist die Elastizität als:

$$(9) \quad \varepsilon_{y,x_k} = \frac{\frac{\partial y}{y}}{\frac{\partial x_k}{x_k}} = \frac{\partial y}{\partial x_k} \cdot \frac{x_k}{y} = b_k \cdot \frac{x_k}{y}$$

definiert. Diese Elastizität kann offensichtlich mit der Höhe der Werte der Variablen variieren. Zur Berechnung der durchschnittlichen Elastizität bietet sich daher die Betrachtung der Mittelwerte der jeweiligen Variablen an (Koutsoyannis 1977, S. 66, Pindyck/Rubinfeld 1991, S. 86).

2.6 Interpretation der Ergebnisse

Bei der Interpretation der Ergebnisse müssen zunächst nicht statistische, sondern inhaltliche Kriterien im Vordergrund stehen (Koutsoyannis 1977). Als erstes ist zu überlegen, ob alle relevanten Variablen in der Regressionsgleichung berücksichtigt und in einen sinnvollen funktionalen Zusammenhang gebracht worden sind. Danach bietet es sich an zu prüfen, ob die Vorzeichen der Regressionskoeffizienten plausibel sind. So sollten in unserem Beispiel die Wirkungen der Werbung und des Außendienstes positiv und die des Preises negativ sein (Assmus/Farley/Lehmann 1984, Tellis 1988). Anschließend ist die Größe der Parameter zu betrachten. Hierbei hilft häufig eine Betrachtung der Elastizitäten. So machen Werbeelastizitäten mit Werten größer als Eins vielfach wenig Sinn. Gleiches gilt für Preiselastizitäten mit absoluten Werten kleiner als Eins. Erst nach dieser inhaltlichen Betrachtung sollten statistische Kriterien wie die Betrachtung des Bestimmtheitsmaßes, der Signifikanz der Regressionsgleichung (F-Test) und der Regressionskoeffizienten (t-Test) herangezogen sowie die Überprüfung der nachfolgend noch dargestellten Annahmen der Regressionsgleichung durchgeführt werden (Koutsoyannis 1977).

2.7 Ergebnisse des Beispiels

Die Schätzung der Regressionsfunktionen erfolgt in diesem Beitrag mit dem Software-Programm SPSS 6.0 und ergibt im Falle einer multiplen Regressionsanalyse mit den drei Marketing-Instrumenten ADM, Preis und Werbung als unabhängige Variablen und der Absatzmenge als abhängige Variable ("Dependent Variable") das in Abbildung 1 im Layout an den Ausdruck des SPSS-Programms angelehnte Ergebnis:

Abbildung 1: Ergebnisse der linearen Regression

| | | | | | |
|-------------------|----------------------|-------------|----------|--------|-------|
| Equation Number 1 | Dependent Variable.. | ABSATZ | | | |
| R Square | ,91885 | | | | |
| Adjusted R Square | ,89856 | | | | |
| F = | 45,29245 | Signif F = | ,0000 | | |
| Variable | B | SE B | Beta | T | Sig T |
| ADM | 6723,477516 | 840,997139 | ,665423 | 7,995 | ,0000 |
| PREIS | -3832,503312 | 444,012774 | -,725207 | -8,632 | ,0000 |
| WERBUNG | ,069194 | ,023839 | ,242142 | 2,903 | ,0133 |
| (Constant) | 210159,44371 | 23729,90947 | | 8,856 | ,0000 |

Das Bestimmtheitsmaß R^2 ("R Square") und das korrigierte Bestimmtheitsmaß ("Adjusted R Square") weisen Werte von 91,89% und 89,86% auf. Die aufgrund des F-Tests und der t-Tests ermittelten Irrtumswahrscheinlichkeiten ("Signif F" und "Sig T") sind geringer als das hier von uns vorgegebene Signifikanzniveau von 5%, so daß alle drei Marketing-Instrumente einen signifikanten Einfluß ausüben. So bedeutet beispielsweise der Wert 0,013 in der Spalte „Sig T“ für das Werbebudget, daß lediglich mit einer Wahrscheinlichkeit von 1,3% in der Grundgesamtheit kein Zusammenhang zwischen dem Werbebudget und dem Umsatz besteht.

Die Werte für die Parameter der Regressionsfunktion stehen in der Spalte "B", deren Standardfehler in der Spalte "SE B" und die Beta-Werte in der Spalte "Beta". Entsprechend den Erwartungen haben die Anzahl der Außendienstmitarbeiter sowie das Werbebudget einen positiven Einfluß und der Preis einen negativen Einfluß auf die Absatzmenge. Die Stärke des Einflusses kann in diesem Ausdruck aber nicht aus der Höhe der Parameterwerte, sondern bestenfalls aus einer Betrachtung der Beträge der Beta-Werte ermittelt werden. Hier stellt sich heraus, daß der Preis einen etwas höheren Einfluß als die Anzahl der Außendienstmitarbeiter und beide Marketing-Instrumente wiederum deutlich höhere Einflüsse als das Werbebudget haben. Die Ermittlung der Elastizitäten anhand der Gleichung (9) ergibt eine Außendienstelastizität von 0,45, eine Preiselastizität von -2,21 und eine Werbeelastizität von 0,26. Alle Elastizitäten weisen sowohl plausible Vorzeichen als auch plausible Größenverhältnisse auf (vgl. dazu auch Hanssens/Parsons/Schultz 1990, Mauerer 1995, die Meta-Analysen von Tellis (1988), Ass-

mus/Farley/Lehmann 1984 und Lodish et al. 1995 sowie die in Skiera 1996, S. 100 ff. dargestellten Ergebnisse einer Befragung).

2.8 Annahmen

Bislang haben wir uns auf die Darstellung des Verfahrens und die damit erreichbaren Ergebnisse konzentriert und sind nicht auf die Annahmen eingegangen, die bei der Anwendung der Regressionsanalyse vorliegen müssen. Diese Annahmen, die sich entweder auf die Residualgrößen, die Beziehung zwischen abhängiger und unabhängigen Variablen, die Beziehung zwischen den unabhängigen Variablen oder die Anzahl der Beobachtungen beziehen, stehen in diesem Abschnitt im Vordergrund.

Annahmen hinsichtlich der Residualgrößen

- Normalverteilung der Residualgrößen e_i ,
- Erwartungswert von Null für alle Residualgrößen $E(e_i)=0$,
- Gleiche Varianz für alle Residualgrößen (Homoskedastizität), d.h. $E(e_i^2)=\delta^2$,
- Keine Korrelation zwischen den Residualgrößen (fehlende Autokorrelation), d.h. $E(e_i \cdot e_j)=0$.

Inhaltlich bedeutet dies im wesentlichen, daß die mit der Regressionsgleichung verbundenen Residualgrößen weder von der Größe der betrachteten Variablen (Homoskedastizität) noch von den anderen Residualgrößen, bei der Betrachtung von Zeitreihen insbesondere nicht von der Residualgröße der Vorperiode (fehlende Autokorrelation), abhängen. In unserem Beispiel sollten die Werte der Residualgrößen also unabhängig von den Ausprägungen der betrachteten drei Marketing-Instrumente sein.

Annahmen hinsichtlich des Zusammenhangs zwischen abhängiger und unabhängigen Variablen

- Erfassung aller relevanten unabhängigen Variablen,
- Linearität des Zusammenhangs.

Um sinnvolle Aussagen tätigen zu können, müssen alle relevanten unabhängigen Variablen erfaßt werden. Anderenfalls besteht die Gefahr, daß sich der Einfluß der nicht erfaßten Variablen in den Regressionskoeffizienten der erfaßten unabhängigen Variablen niederschlägt. In unserem Beispiel sollten also neben den drei betrachteten Marketing-

Instrumenten keine weiteren relevanten Einflußfaktoren auf den Umsatz, z.B. strukturelle Unterschiede zwischen den Bezirken, vorhanden sein. Weiterhin geht die lineare Regressionsanalyse von einem proportionalen, d.h. linearen Zusammenhang zwischen den unabhängigen und der abhängigen Variablen aus.

Annahmen hinsichtlich des Zusammenhangs zwischen den unabhängigen Variablen

Es wird davon ausgegangen, daß keine lineare Abhängigkeit zwischen den unabhängigen Variablen besteht (fehlende Multikollinearität). So ergibt sich beispielsweise im Falle einer hohen Korrelation zwischen zwei unabhängigen Variablen das Problem, daß keiner der beiden Variablen der Einfluß auf die abhängige Variable eindeutig zugesprochen werden kann.

Anzahl der Beobachtungen

Damit das in Gleichung (1) beschriebene Modell überhaupt schätzbar ist, muß die Anzahl der Beobachtungen mindestens so groß wie die Anzahl der zu schätzenden Parameter (Konstante b_0 plus die Anzahl aller Regressionskoeffizienten b_k) sein. Signifikante Einflüsse und die damit verbundene Übertragung der Ergebnisse der Stichprobe auf die Grundgesamtheit können aber nur festgestellt werden, wenn die Anzahl der Beobachtungen deutlich größer als die Anzahl der zu schätzenden Parameter ist. Generelle Aussagen über dieses Verhältnis können kaum getroffen werden, da dies stets vom vorliegenden Datensatz abhängt. Wenn die Anzahl der Beobachtungen jedoch nicht zumindest dreimal, besser sogar fünfmal so groß wie die Anzahl der zu schätzenden Parameter ist, so besteht nur eine geringe Chance zur Ermittlung signifikanter Zusammenhänge.

3 Vorgehensweise

Nachfolgend wird die Vorgehensweise beim Einsatz der linearen Regressionsanalyse beschrieben. Es wird auf die häufig notwendige Transformation der Variablen zur Erstellung eines linearen Modells und die Effizienz der Schätzer eingegangen. Außerdem werden die Probleme der Multikollinearität, der Autokorrelation, der Heteroskedastizität und der Ausreißer erörtert.

3.1 Transformation der Variablen

Häufig sollen nichtlineare Zusammenhänge, z.B. der Zusammenhang in unserem Zahlenbeispiel zwischen der Absatzmenge und den einzelnen Marketing-Instrumenten, mit Hilfe der nachfolgend dargestellten multiplikativen Absatzreaktionsfunktion geschätzt werden:

$$(10) \quad \text{ABSATZ} = \alpha \cdot \text{ADM}^\delta \cdot \text{PREIS}^\eta \cdot \text{WERBUNG}^\beta$$

Diese multiplikative Absatzreaktionsfunktion weist gegenüber der linearen Reaktionsfunktion den Vorteil von Wirkungsinteraktionen zwischen den Marketing-Instrumenten sowie veränderbarer Grenzerträge für die einzelnen Marketing-Instrumente auf (ein ausführlicher Vergleich unterschiedlicher Funktionsverläufe erfolgt in Hruschka 1996, S. 18 ff). Zur Schätzung dieses nichtlinearen Zusammenhangs kann vielfach auch die lineare Regressionsanalyse herangezogen werden, da diese nur einen linearen Zusammenhang in der zu schätzenden Regressionsgleichung unterstellt. Ein nichtlinearer Zusammenhang zwischen abhängiger und unabhängigen Variablen kann berücksichtigt werden, wenn er linearisierbar ist. Beispielsweise kann die obige multiplikative Absatzreaktionsfunktion durch Logarithmierung linearisiert werden:

$$(11) \quad \ln(\text{ABSATZ}) = \ln(\alpha) + \delta \cdot \ln(\text{ADM}) + \eta \cdot \ln(\text{PREIS}) + \beta \cdot \ln(\text{WERBUNG}) .$$

Die zu schätzende lineare Regressionsgleichung würde dann folgendes Aussehen haben:

$$(12) \quad Q_i = a + \delta \cdot A_i + \eta \cdot P_i + \beta \cdot W_i + e_i \quad (i \in I),$$

wobei die Variablen Q_i , A_i , P_i und W_i folgendermaßen definiert sind:

$$Q_i = \ln(\text{ABSATZ}_i),$$

$$A_i = \ln(\text{ADM}_i),$$

$$P_i = \ln(\text{PREIS}_i),$$

$$W_i = \ln(\text{WERBUNG}_i).$$

In vergleichbarer Form kann eine ganze Reihe weiterer nichtlinearer Modelle linearisiert werden (Pindyck/Rubinfeld 1991, S. 102, Schneeweiß 1990, S. 52, Hair et al. 1992, S. 52 f.). Die lineare Regressionsanalyse liefert für die so linearisierte, ursprünglich multiplikative Reaktionsfunktion (10) die nachfolgenden Ergebnisse:

Abbildung 2: Ergebnisse für die multiplikative Absatzreaktionsfunktion

| | | | | | |
|-------------------|----------------------|----------|------------|--------|-------|
| Equation Number 1 | Dependent Variable.. | LN_ABS | | | |
| R Square | | ,92758 | | | |
| Adjusted R Square | | ,90947 | | | |
| F = | 51,23087 | | Signif F = | ,0000 | |
| Variable | B | SE B | Beta | T | Sig T |
| LN_ADM | ,399516 | ,046643 | ,674064 | 8,565 | ,0000 |
| LN_PRE | -2,339480 | ,247816 | -,748741 | -9,440 | ,0000 |
| LN_WER | ,216893 | ,081927 | ,207453 | 2,647 | ,0213 |
| (Constant) | 16,980661 | 1,493204 | | 11,372 | ,0000 |

Die Bestimmtheitsmaße und der F-Wert weisen für diese multiplikative Absatzreaktionsfunktion hohe Werte auf. Das Bestimmtheitsmaß ist jedoch nur für das logarithmierte Modell (11) und nicht für das Ausgangsmodell (10) aussagekräftig. Deswegen empfiehlt es sich, das Bestimmtheitsmaß des Ausgangsmodells für die geschätzten Parameterwerte im Ausgangsmodell (10) zu berechnen. Es ergibt sich in diesem Fall ein Bestimmtheitsmaß von 93,94%. Vergleichbar sind die Signifikanzniveaus der einzelnen Variablen. Ein großer Vorteil besteht bei der Verwendung einer multiplikativen Reaktionsfunktion darin, daß die Regressionskoeffizienten die Elastizitäten der jeweiligen Marketing-Instrumente darstellen und so eine unmittelbare Interpretation der Ergebnisse erleichtern. Auch bei dieser Analyse weisen alle Marketing-Instrumente das erwartete Vorzeichen und plausible Werte auf. Es ergibt sich also für die in Gleichung (10) dargestellte multiplikative Absatzreaktionsfunktion das folgende (gerundete) Ergebnis:

$$\begin{aligned}
 \text{ABSATZ} &= \exp(16,98) \cdot \text{ADM}^{0,40} \cdot \text{PREIS}^{-2,34} \cdot \text{WERBUNG}^{0,22} \\
 (13) \qquad &= 23.676.653 \cdot \text{ADM}^{0,40} \cdot \text{PREIS}^{-2,34} \cdot \text{WERBUNG}^{0,22}
 \end{aligned}$$

Aufgrund ihren plausibleren Eigenschaften wird in der Literatur häufig auf die multiplikative und nicht auf die lineare Funktion zurückgegriffen (vgl. beispielsweise Tellis 1988). Wenn wir bei den nachfolgenden Rechnungen weiterhin auf die lineare Funktion zurückgreifen, so hat das lediglich didaktische Gründe.

3.2 Effizienz der Schätzer

In der Ökonometrie wird der geschätzte Parameter als effizienter ("efficient") Schätzer bezeichnet, wenn er erwartungstreu, also unverzerrt ("unbiased estimator") ist, und gleichzeitig den geringsten Schätzfehler aller unverzerrt geschätzten Parameter ("best estimator") aufweist (Koutsoyannis 1977, S. 101 ff., Schneeweiß 1990, S. 71 ff.). Die mit Hilfe der Methode der kleinsten Quadrate geschätzten Parameter sind aber nur dann effizient, wenn die in Kapitel 2.8 dargestellten Annahmen erfüllt sind. Anderenfalls muß entweder eine andere Schätzmethode gewählt oder die Regressionsgleichung in einer anderen Art und Weise formuliert werden.

Bei der Prüfung der Annahmen sollte insbesondere das Vorliegen von Multikollinearität, Autokorrelation und Heteroskedastizität untersucht werden. Die Normalverteilung der Residualgrößen ist dagegen von vergleichsweise untergeordneter Bedeutung, da sie aufgrund des zentralen Grenzwertsatzes bei einer "hinreichend" großen Anzahl an Beobachtungswerten stets erfüllt ist (Hansen 1993, S. 68, Pindyck/Rubinfeld 1991, S. 126). Koutsoyannis (1977, S. 197) erwähnt, daß dies selbst bei Stichproben mit nur 10-20 Beobachtungswerten häufig schon der Fall ist. Außerdem führt das Nichtvorliegen einer Normalverteilung lediglich dazu, daß die in Kapitel 2.4 erörterten F- und t-Tests nicht sinnvoll anwendbar sind. Die ermittelten Parameterwerte sind allerdings weiterhin effizient (Koutsoyannis 1977, S. 197). Es können also bei nicht normalverteilten Residualgrößen lediglich keine Aussagen über die Signifikanzniveaus der betrachteten Zusammenhänge getätigt werden.

3.3 Prüfung auf Multikollinearität

Multikollinearität liegt vor, wenn die unabhängigen Variablen untereinander linear abhängig sind. Sie führt üblicherweise zu hohen Standardabweichungen der Regressionskoeffizienten und dazu, daß diese Regressionskoeffizienten nur unzureichend interpretiert werden können. Multikollinearität in der Form der linearen Abhängigkeit zwischen zwei unabhängigen Variablen kann durch das Betrachten der Korrelationsmatrix aufgedeckt werden. Hohe absolute Werte für die Korrelationen, also Werte nahe -1 oder +1, deuten auf Multikollinearitätsprobleme hin. Multikollinearität in der Form der linearen Abhängigkeit zwischen mehr als zwei unabhängigen Variablen kann dadurch erkannt werden, daß mehrere lineare Regressionen gerechnet werden, bei denen jede der ursprünglich unabhängigen Variablen durch die anderen unabhängigen Variablen erklärt wird. Die Differenz zwischen Eins und dem Bestimmtheitsmaß für eine derartige Re-

gression wird als Toleranz der Variablen und der Kehrwert aus dieser Differenz als "Variance Inflation Factor" (VIF-Wert) bezeichnet. Nur wenn die Bestimmtheitsmaße dieser Regressionen niedrig sind, kann von einer linearen Unabhängigkeit der Variablen ausgegangen werden. Dies spiegelt sich dann in Toleranz- und VIF-Werten nahe Eins wider. Niedrigere Toleranz- und höhere VIF-Werte weisen dagegen auf Multikollinearitätsprobleme hin (für weitere Tests zur Aufdeckung der Multikollinearität siehe Koutsoyannis 1977, S. 238 ff.).

In unserem Beispiel weist die in Abbildung 3 dargestellte Korrelationsmatrix durchgehend niedrige Korrelationen zwischen den unabhängigen Variablen (unterstrichene Werte) auf, so daß keine linearen Abhängigkeiten zwischen zwei unabhängigen Variablen vorliegen. Da außerdem die ebenfalls in Abbildung 3 beschriebenen Toleranz- und VIF-Werte ("Tolerance" und "VIF") Werte nahe Eins aufweisen, kann zudem davon ausgegangen werden, daß auch keine linearen Abhängigkeiten zwischen mehreren unabhängigen Variablen bestehen.

Abbildung 3: Prüfung der Multikollinearität

| - - Correlation Coefficients - - | | | | |
|---------------------------------------|-----------|---------------|---------------|---------------|
| | ABSATZ | ADM | PREIS | WERBUNG |
| ABSATZ | 1,0000 | ,5422 | -,6683 | ,3032 |
| ADM | ,5422 | <u>1,0000</u> | <u>,1430</u> | <u>-,0804</u> |
| PREIS | -,6683 | <u>,1430</u> | <u>1,0000</u> | <u>-,1579</u> |
| WERBUNG | ,3032 | <u>-,0804</u> | <u>-,1579</u> | <u>1,0000</u> |
| ----- Variables in the Equation ----- | | | | |
| Variable | Tolerance | VIF | | |
| ADM | ,976115 | 1,024 | | |
| PREIS | ,957963 | 1,044 | | |
| WERBUNG | ,971656 | 1,029 | | |

Multikollinearität stellt also bei der Betrachtung der drei bislang betrachteten Marketing-Instrumente kein Problem dar. Dies ändert sich, wenn nun auch das bislang nicht betrachtete Marketing-Instrument der Mailings (siehe Tabelle 1) betrachtet wird. Die Anzahl der Mailings korreliert hoch (0,9895) mit der Anzahl der Außendienstmitarbeiter, so daß die in Abbildung 4 dargestellten Ergebnisse der Regressionsanalyse mit diesem zu-

sätzlichen Marketing-Instrument einen völlig veränderten Regressionskoeffizienten für die Anzahl der Außendienstmitarbeiter ergibt. Deren Einsatz übt nun einen negativen Einfluß auf die Absatzmenge aus (Elastizität von -0,09), während die Anzahl der Mailings stark positiv wirkt (Elastizität 0,52). Gleichzeitig ist die Standardabweichung des Regressionskoeffizienten des Außendiensteinsatzes etwa sieben Mal so groß wie bei der Regressionsanalyse ohne Betrachtung der Mailings.

Abbildung 4: Ergebnisse der Regressionsanalyse bei Multikollinearitätsproblemen

| | | | | | | |
|-------------------|----------------------|------------|-----------|--------|--------|--------|
| Equation Number 1 | Dependent Variable.. | | ABSATZ | | | |
| R Square | ,93121 | | | | | |
| Adjusted R Square | ,90620 | | | | | |
| F = | 37,22832 | Signif F = | ,0000 | | | |
| Variable | B | SE B | Tolerance | VIF | T | SIG T |
| ADM | -1372,882311 | 5815,151 | ,018879 | 52,968 | -,236 | ,8177 |
| PREIS | -3711,737290 | 435,531 | ,920701 | 1,086 | -8,522 | 0,0000 |
| WERBUNG | ,078382 | ,024 | ,898621 | 1,113 | 3,288 | ,0072 |
| MAILING | 7,948870 | 5,654 | ,018881 | 52,962 | 1,406 | ,1874 |
| (Constant) | 203009,25162 | 23379,27 | | | 8,683 | 0,0000 |

Eine ökonometrische Lösung der Multikollinearitätsprobleme ist schwierig (Ansatzpunkte bietet allenfalls die Ridge Regression, vgl. Hair et al. 1992, S. 49, Wittink 1988, S. 101), so daß andere Lösungsmöglichkeiten angestrebt werden müssen. So bietet sich eine Erhöhung der Anzahl der Beobachtungen an, sofern in den zusätzlich hinzugefügten Beobachtungen die Variablen, hier die Anzahl der Außendienstmitarbeiter und der Mailings, weniger stark linear voneinander abhängen. Alternativ dazu können entweder die linear voneinander abhängigen Variablen zu einer einzigen Variablen zusammengefaßt werden (z.B. mit Hilfe der Faktorenanalyse) oder eine bzw. mehrere dieser Variablen aus der Regressionsgleichung eliminiert werden. So führt beispielsweise die Eliminierung der Anzahl der Außendienstmitarbeiter dazu, daß die Elastizität der Mailings auf einen Wert von 0,41 zurückgeht. Richtig befriedigend können Multikollinearitätsprobleme jedoch häufig nicht behoben werden, da die Anzahl der Beobachtungen meistens fest vorgegeben ist oder gerade die Betrachtung der Einflüsse ganz bestimmter Variablen von Interesse ist. So kann in unserem Zahlenbeispiel nicht festgestellt werden, welchen Einfluß die Anzahl der Außendienstmitarbeiter und der Mailings alleine auf den Absatz

nehmen. Dies würde unsere Marketing-Managerin jedoch sicherlich gerne wissen, da sie dann eine entsprechende Optimierung des Marketing-Mix vornehmen könnte. Zur Ermittlung derartiger Einflüsse müßten die Anzahl der Außendienstmitarbeiter und der Mailings so variiert werden, daß nicht mehr eine derartig hohe Korrelation auftritt. Dies könnte beispielsweise durch entsprechend aufgebaute Experimente erreicht werden.

3.4 Prüfung auf Autokorrelation

Autokorrelation bedeutet, daß eine Korrelation zwischen den Störgrößen besteht. Eine derartige Korrelation tritt häufig bei der Analyse von Zeitreihen auf, wenn die zyklischen Schwankungen der Zeitreihe nicht adäquat von den unabhängigen Variablen erfaßt werden. Dies führt üblicherweise dazu, daß einige Perioden lang die beobachteten Werte zunächst überschätzt und dann einige Perioden lang unterschätzt werden. Eine Reihe von negativen Residualgrößen wechseln sich also mit einer Reihe von positiven Residualgrößen ab. Autokorrelation bewirkt, daß die Standardfehler der Regressionskoeffizienten unterschätzt und damit das Signifikanzniveau der t-Tests überschätzt werden (die Standardabweichung erscheint im Zähler des in Gleichung (7) dargestellten t-Tests). Die geschätzten Regressionskoeffizienten bleiben unverzerrt. Sie sind aber nicht mehr effizient, da der Standardfehler nicht korrekt ermittelt wird (Pindyck/Rubinfeld 1991, S. 138).

Die meisten Tests auf Autokorrelation untersuchen die Autokorrelation erster Ordnung, d.h. die Korrelation zwischen zwei zeitlich aufeinanderfolgenden Residualgrößen (Koutsoyannis 1977, S. 200). Neben einer graphischen Betrachtung der Residualgrößen wird häufig der Durbin-Watson-Test verwendet (Pindyck/Rubinfeld 1991, S. 143):

$$(14) \quad dw = \frac{\sum_{i \in I'} (e_i - e_{i-1})^2}{\sum_{i \in I} e_i^2}$$

wobei:

dw: Wert des Durbin-Watson-Tests,

I': Indexmenge der Beobachtungswerte ohne den ersten Beobachtungswert.

Wenn die Differenz zwischen den Residualgrößen zweier aufeinanderfolgender Beobachtungswerte sehr klein (groß) ist, so liegt positive (negative) Autokorrelation vor und der Zähler der Gleichung (14) nimmt kleine (große) Werte an. Dies führt dazu, daß der Durbin-Watson-Wert d gegen den Wert Null (vier) strebt. Ein Wert von zwei zeigt an,

daß keine Autokorrelation erster Ordnung vorliegt. Für eine genaue Darstellung des Signifikanzniveaus des Durbin-Watson-Tests sei auf weiterführende Literatur verwiesen (z.B. Pindyck/Rubinfeld 1991, S. 143 ff., Schneeweiß 1990, S. 187 ff., Koutsoyannis 1977, S. 212 ff.).

Vielfach wird bei der Betrachtung von Zeitreihen der Vorperiodenwert der abhängigen Variablen als unabhängige Variable, häufig dann als Lag-Variable bezeichnet, herangezogen. In diesem Falle sollte beachtet werden, daß der Durbin-Watson-Test nicht zur Aufdeckung der Autokorrelation geeignet ist. Statt dessen sollte auf Durbin's h-Test ausgewichen werden (Pindyck/Rubinfeld 1991, S. 147 ff.).

Die Betrachtung der Autokorrelation ist bei unserem Zahlenbeispiel bedeutungslos, da es sich um Querschnittsdaten handelt und somit keine zeitliche Korrelation, sondern allenfalls eine räumlichen Autokorrelation (vgl. dazu Cliff/Ord 1973) vorliegen kann, von deren Betrachtung wir an dieser Stelle absehen möchten.

Autokorrelation kann ein großes Problem darstellen, da es vor allem ein Indiz für fehlende unabhängige Variablen ist. Zwei weitere, in Tabelle 2 dargestellte Zahlenbeispiele verdeutlichen dieses Problem. Die Spalten der x- und y-Werte mögen die Beziehung zwischen x- und y-Werten in der folgenden Form beschreiben:

$$(15) \quad y = 5 + 2 \cdot x$$

Die Spalten "Error1" und "Error2" mögen die Residualgrößen in jeweils einem der beiden Beispiele im Falle einer Autokorrelation erster Ordnung beschreiben. Diese beiden Residualgrößen unterscheiden sich nur durch ihr Vorzeichen. Durch Addition der Werte von "Error1" und "Error2" zu den y-Werten ergeben sich jeweils die unabhängigen Variablen y_1 und y_2 der beiden Beispiele, z.B. $y_1 = y + \text{Error1}$.

Tabelle 2: Zahlenbeispiele zur Darstellung der Problematik der Autokorrelation

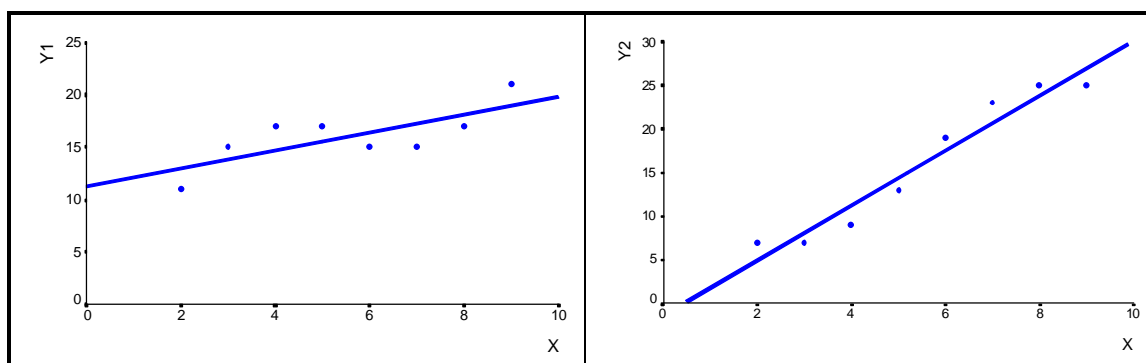
| Fall | x | y | y1 | y2 | Error1 | Error2 |
|------|---|----|----|----|--------|--------|
| 1 | 2 | 9 | 11 | 7 | 2 | -2 |
| 2 | 3 | 11 | 15 | 7 | 4 | -4 |
| 3 | 4 | 13 | 17 | 9 | 4 | -4 |
| 4 | 5 | 15 | 17 | 13 | 2 | -2 |
| 5 | 6 | 17 | 15 | 19 | -2 | 2 |
| 6 | 7 | 19 | 15 | 23 | -4 | 4 |
| 7 | 8 | 21 | 17 | 25 | -4 | 4 |
| 8 | 9 | 23 | 21 | 25 | -2 | 2 |

Die Regressionen für die beiden Beispiele mit den abhängigen Variablen y1 und y2 und der unabhängigen Variablen x ergeben die in Tabelle 3 und Abbildung 5 dargestellten Ergebnisse. Es wird deutlich sichtbar, daß die Residualgrößen in beiden Beispielen systematisch unter- bzw. überschätzt werden und die geschätzten Regressionsfunktionen den in Gleichung (15) dargestellten tatsächlichen Zusammenhang nicht widerspiegeln. Die geschätzten Regressionskoeffizienten sind aber dennoch erwartungstreu (d.h. unverzerrt), da bei einer Vielzahl an Regressionen mit den Zufallsstichproben, die auch den in Tabelle 2 dargestellten Datensätzen unterliegen, sich die in Gleichung (15) dargestellten Parameterwerte ergeben. Dies kann bereits in den beiden Beispielen daran erkannt werden, daß die Mittelwerte der Konstanten und der Regressionskoeffizienten der beiden Regressionsfunktionen für die beiden Zufallsstichproben den Parametern des in Gleichung (15) unterstellten Zusammenhangs entsprechen ($\frac{11,29-1,29}{2} = 5$ bzw. $\frac{0,86+3,14}{2} = 2$).

Tabelle 3: Ergebnisse der Regressionen im Falle von Autokorrelation

| | const1 | par1 | const2 | par2 |
|--------------------------------|--------|------|--------|------|
| Wert | 11,29 | 0,86 | -1,29 | 3,14 |
| Standardabweichung | 1,88 | 0,32 | 1,88 | 0,32 |
| Signifikanzniveau des t-Tests | 0,03 | 0,00 | 0,52 | 0,00 |
| R ² | 0,55 | | 0,98 | |
| Signifikanzniveaus des F-Tests | 0,03 | | 0,00 | |
| Durbin-Watson-Wert | 1,27 | | 1,27 | |

Abbildung 5: Graphische Darstellung der Regressionsfunktionen im Falle der Autokorrelation



Autokorrelationsprobleme können durch die Erfassung der Einflußgrößen, die für die zeitlichen Schwankungen verantwortlich sind, behoben werden. Erst wenn dies nicht möglich ist, sollte die Behebung des Problems mit Hilfe ökonomischer Verfahren, z.B. die Cochrane-Orcutt-Prozedur oder die Hildreth-Lu-Prozedur (Pindyck/Rubinfeld 1991, S. 138 ff., für weitere Verfahren siehe Hansen 1993, S. 97 ff.), angestrebt werden. Diese ökonomischen Verfahren können aber nur zum Ziel führen, wenn die Regressionsgleichung korrekt formuliert ist.

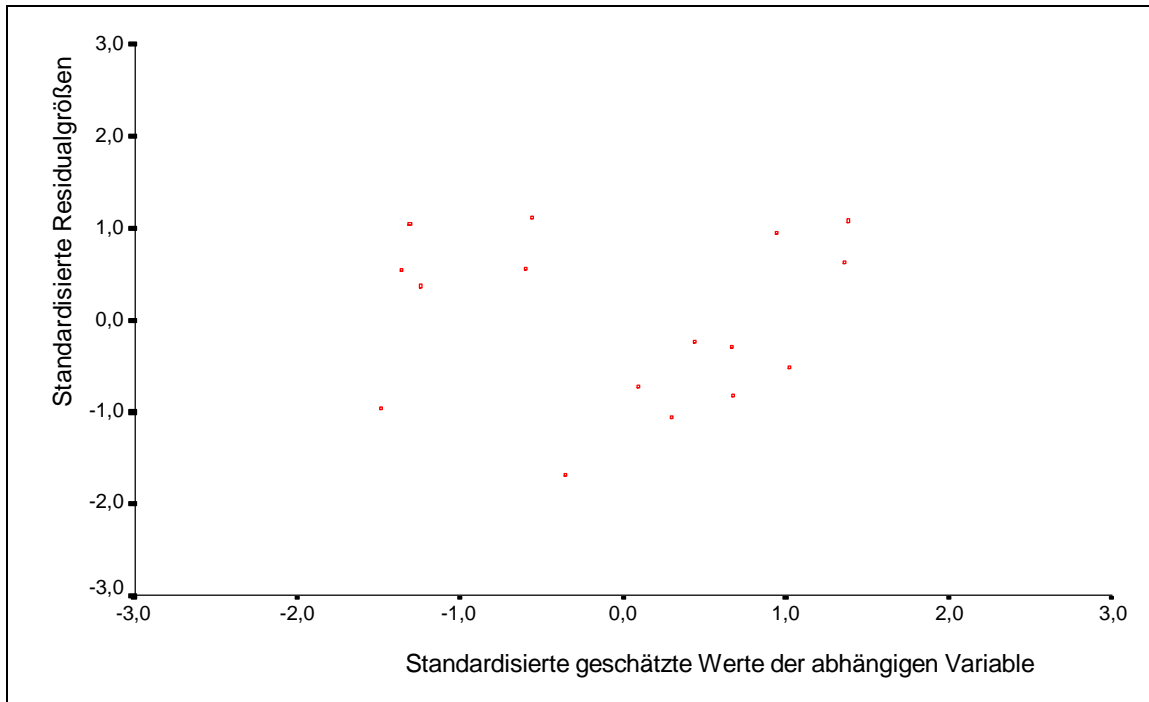
3.5 Prüfung auf Heteroskedastizität

Heteroskedastizität bedeutet, daß nicht alle Residualgrößen die gleiche Varianz aufweisen. Im Gegensatz zur Autokorrelation tritt das Problem der Heteroskedastizität seltener bei der Betrachtung von Zeitreihen, sondern eher bei der Betrachtung von Querschnittsdaten auf (Pindyck/Rubinfeld 1991, S. 127). So ist es beispielsweise naheliegend, daß die Schätzung der Marktanteile für Unternehmen mit einem großen Marktanteil einen größeren erwarteten Fehler aufweist als diejenige für Unternehmen mit einem kleinen Marktanteil (ähnlich argumentiert Koutsoyannis 1977, S. 183). Heteroskedastizität führt dazu, daß die Methode der kleinsten Quadrate nicht mehr alle Beobachtungswerte quasi gleich behandelt, sondern mehr Wert auf eine gute Prognose der Werte mit einer hohen Varianz legt und damit implizit eine höhere Gewichtung dieser Beobachtungswerte vornimmt. Dies führt zwar wiederum zu erwartungstreuen, aber nicht mehr effizienten Schätzern, da sie nicht die kleinsten Schätzfehler aufweisen. (Pindyck/Rubinfeld 1991, S. 128).

Heteroskedastizität kann durch eine graphische Gegenüberstellung der Residualgrößen mit der abhängigen oder einer der unabhängigen Variablen sowie die Anwendung des

Goldfeldt-Quandt-Tests, des Breusch-Pagan-Tests oder des White-Tests erkannt werden (Pindyck/Rubinfeld 1991, S. 132 ff.). In unserem Beispiel läßt die graphische Gegenüberstellung in Abbildung 6 keinen Zusammenhang zwischen den standardisierten Residualgrößen und den standardisierten geschätzten Werten der abhängigen Variablen erkennen.

Abbildung 6: Graphische Prüfung der Heteroskedastizität



Die Vermeidung der Heteroskedastizität kann häufig nicht durch die Erhebung zusätzlicher Variablen behoben werden, weil vielfach naheliegende Erklärungen, z.B. die oben aufgeführte Schätzung der Marktanteile für große und kleine Unternehmen, für das Auftreten der Heteroskedastizität vorliegen. Deswegen muß das Problem normalerweise mit Hilfe ökonometrischer Verfahren gelöst werden. Dies kann entweder auf Basis inhaltlicher Erwägungen durch eine geeignete Transformation der Regressionsgleichung, z.B. die Division der Regressionsgleichung durch die für das Auftreten der Heteroskedastizität verantwortliche unabhängige Variable (Koutsoyannis 1977, S. 187), erfolgen. Alternativ dazu kann auch auf Basis ökonometrischer Erwägungen eine gewichtete lineare Regressionsanalyse ("weighted least squares"), die quasi die durch die Heteroskedastizität implizit vorgenommene Gewichtung der Beobachtungswerte rückgängig macht, vorgenommen werden (Pindyck/Rubinfeld 1991, S. 129 ff.).

3.6 Identifizierung von Ausreißern

Normalerweise möchte man bei der Regressionsanalyse, daß alle Beobachtungswerte einen vergleichbaren Einfluß auf das Ergebnis haben, und nicht, daß einzelne Beobachtungen das Ergebnis sehr stark beeinflussen. Eine derartige unverhältnismäßig hohe Beeinflussung kann aber durch sogenannte Ausreißer auftreten, die sich üblicherweise dadurch auszeichnen, daß deren Beobachtungswerte weit von den anderen Beobachtungswerten abweichen. Deswegen sollten Datensätze stets auf das Vorliegen von Ausreißern geprüft werden. Dafür existiert neben der visuellen Inspektion der Verteilung der Beobachtungswerte oder der Verteilung der Residuen eine ganze Reihe an statistischen Verfahren. Vergleichsweise häufig findet man "Mahalanobis Distance" oder "Cook's Distance". Die Mahalanobis Distance baut auf den quadrierten standardisierten Werten der unabhängigen Variablen auf, während mit Cook's Distance die Veränderung der Residuen aller anderen Beobachtungswerte erfaßt wird, wenn der betrachtete Beobachtungswert aus der Regressionsgleichung entfernt wird. Einen gut verständlichen Überblick über diese und weitere statistische Verfahren geben Chatterjee/Hadi (1986).

In unserem Zahlenbeispiel ergeben sich bei der Betrachtung der Residuen, der Mahalanobis Distance und Cook's Distance keine auffälligen Werte. Der bislang nicht betrachtete Bezirk 17 weist jedoch im Gegensatz zu den anderen Bezirken sehr hohe Absatzzahlen auf. Wenn dieser Bezirk 17, deren Werte für Cook's Distance und Mahalanobis Distance gerade eben auf einen Ausreißer hindeuten, mit in die Regressionsanalyse aufgenommen wird, so kann eine deutliche Veränderung gegenüber dem bisherigen Ergebnis festgestellt werden (Abbildung 7). So sinkt nicht nur das Bestimmtheitsmaß auf 63,7% ab, sondern es ändern sich auch die Regressionskoeffizienten und deren Signifikanzniveaus. Es ergeben sich jetzt Elastizitäten für den Außendienst, den Preis und die Werbung in Höhe von 0,27, -2,64 und 0,01.

Abbildung 7: Ergebnisse des Regressionsanalyse beim Vorliegen eines Ausreißers

| | | | | | |
|-------------------|----------------------|-------------|----------|--------|-------|
| Equation Number 1 | Dependent Variable.. | ABSATZ | | | |
| R Square | ,63705 | | | | |
| Adjusted R Square | ,55329 | | | | |
| F = | 7,60575 | Signif F = | ,0035 | | |
| Variable | B | SE B | Beta | T | Sig T |
| ADM | 4310,501531 | 1868,546668 | ,399066 | 2,307 | ,0382 |
| PREIS | -4721,008371 | 1024,570054 | -,795895 | -4,608 | ,0005 |
| WERBUNG | ,001420 | ,053040 | ,004494 | ,027 | ,9790 |
| (Constant) | 289393,66905 | 51273,67876 | | 5,644 | ,0001 |

Wenn die Ausreißer nicht auf Eingabefehler zurückzuführen sind, dann kann für die Behandlung von Ausreißern keine eindeutige Anweisung gegeben werden. Vielmehr hängt deren Behandlung stark davon ab, welche Aussagen mit den Ergebnissen der durchgeführten Regressionsanalyse getroffen werden sollen. Wenn Aussagen für alle Beobachtungswerte getätigt werden sollen, dann ist eine Eliminierung von Ausreißern natürlich wenig befriedigend. Normalerweise sollen jedoch Aussagen für die Mehrzahl der Beobachtungswerte angestrebt werden, so daß eine Eliminierung der Ausreißer vielfach angebracht ist. In unserem Beispiel ist die Eliminierung des Ausreißers (Bezirk 17) vermutlich deswegen sinnvoll, weil die Marketing-Managerin stärker an Aussagen interessiert sein dürfte, die für die Mehrzahl der betrachteten Bezirke gelten. Gleichzeitig sollte die Marketing-Managerin jedoch überlegen, warum sich der Bezirk 17 von den anderen Bezirken so stark unterscheidet.

4 Implikationen der Analyse

Wenn an dieser Stelle von dem aus didaktischen Gründen eingeführten Marketing-Instrument der Mailings abgesehen wird, so sollte der Assistent die von der Marketing-Managerin in Kapitel 2.1 skizzierten Fragestellungen mit Hilfe der bereits in (13) und hier nochmals aufgeführten multiplikativen Absatzreaktionsfunktion beantworten.

$$(13) \quad \text{ABSATZ} = 23.676.653 \cdot \text{ADM}^{0,40} \cdot \text{PREIS}^{-2,34} \cdot \text{WERBUNG}^{0,22}$$

Der optimale Preis ergibt sich aus der weithin bekannten Amoroso-Robinson-Relation (vgl. beispielsweise Simon 1992, S. 163):

$$(16) \quad p^* = \frac{\varepsilon_{q,p}}{1 + \varepsilon_{q,p}} \cdot k = \frac{-2,34}{1 - 2,34} \cdot 30 \text{ DM} = 1,75 \cdot 30 \text{ DM} = 52,50 \text{ DM} .$$

Da sowohl von der Anzahl der Außendienstmitarbeiter als auch von dem Werbebudget positive Einflüsse auf die Absatzmenge ausgehen, sind Investitionen in beide Marketing-Instrumente grundsätzlich sinnvoll. Mit Hilfe des Dorfman-Steiner-Theorems (Dorfman/Steiner 1954) kann eine Empfehlung hinsichtlich der Aufteilung der Budgets zwischen den beiden Instrumenten gegeben werden. Dieses Theorem besagt, daß bei einem vorgegebenen Gesamtbudget die Budgets proportional zu ihren Elastizitäten aufgeteilt werden sollen:

$$(17) \quad \frac{\text{ADM}}{\text{ADM} + \text{WERBUNG}} \stackrel{!}{=} \frac{\delta}{\delta + \beta} = \frac{0,40}{0,40 + 0,22} = \frac{0,40}{0,62} = 64,5\% ,$$

$$(18) \quad \frac{\text{WERBUNG}}{\text{ADM} + \text{WERBUNG}} \stackrel{!}{=} \frac{\beta}{\delta + \beta} = \frac{0,22}{0,40 + 0,22} = \frac{0,22}{0,62} = 35,5\% .$$

Bei dem gegenwärtigen durchschnittlichen Gesamtbudget pro Bezirk von etwa 982.000 DM (310.000 DM für die Werbung und 5,6 Außendienstmitarbeiter à 120.000 DM) sollten also 64,5% (633.390 DM) davon für Außendienstmitarbeiter und 35,5% (348.6108 DM) für Werbung ausgegeben werden. Eine Optimierung der sich aus der Absatzreaktionsfunktion (13) ergebenden Gewinnfunktion (19)

$$(19) \quad \text{GEWINN} = (\text{PREIS} - 30) \cdot 23.676.653 \cdot \text{ADM}^{0,40} \cdot \text{PREIS}^{-2,34} \cdot \text{WERBUNG}^{0,22} - \text{ADM} - \text{WERBUNG}$$

ergibt optimale Budgets für die Werbung und den Außendienst in Höhe von 405.535 DM und 737.335 DM.

5 Software

Software zur Schätzung der Regressionsanalyse liegt in vielfältiger Form vor. Erste Analysen erlauben fast alle Tabellenkalkulationsprogramme (z.B. Microsoft Excel, Lotus 1-2-3), wobei sich diese Programme insbesondere hinsichtlich der Möglichkeiten zur graphischen Darstellung der Ergebnisse und der Anzahl der statistischen und ökonometrischen Tests unterscheiden. Die Grenzen dieser preisgünstigen Programme liegen in dem angebotenen Funktionsumfang zur einfachen Transformation der Daten, dem Durchspielen unterschiedlicher Varianten der Regressionsanalyse, den Möglichkeiten zur Aufdeckung von Heteroskedastizität und Autokorrelation, der Identifizierung von

Ausreißern und dem Anwenden der nichtlinearen Regressionsanalyse. Mit dem preislich deutlich teureren Statistikprogramm SPSS können derartige Analysen leicht durchgeführt werden. Es weist zudem den Vorteil einer hohen Bedienerfreundlichkeit auf. Allerdings sind auch in SPSS nicht alle hier erwähnten Verfahren zur Diagnostik und Bewältigung von Autokorrelation und Heteroskedastizität, z.B. Durbin's h-Statistik, enthalten. Einen höheren Funktionsumfang weisen andere Statistikprogramme wie beispielsweise SAS und RATS auf, allerdings bei einer tendenziell niedrigeren Bedienerfreundlichkeit. Prinzipiell alle ökonomischen Verfahren können bei matrixorientierten Programmen wie GAUSS oder MATLAB vorgenommen werden, da der Anwender dort direkt Matrizen bearbeitet. Die Einarbeitungszeit in diese nicht immer sehr bedienerfreundlichen Programme ist allerdings nicht zu unterschätzen. Als Fazit kann festgehalten werden, daß für das gelegentliche Rechnen einer Regressionsanalyse die meisten Tabellenkalkulationsprogramme völlig ausreichend sind. Für ausführlichere Analysen sollte aber auch auf Statistikprogramme wie beispielsweise SPSS, SAS, RATS zurückgegriffen werden. Erst wenn diese Programme nicht mehr die gewünschten Analysen ermöglichen, sollten matrixorientierte Programme wie z.B. GAUSS oder MATLAB herangezogen werden.

6 Ausblick

Häufig entsteht bei der Darstellung statistischer Verfahren wie beispielsweise der Regressionsanalyse der Eindruck, daß bei entsprechend sorgfältiger Vorbereitung die Berechnung der Ergebnisse binnen kürzester Zeit erfolgen kann. Dies ist auch aus rein statistischer Sicht richtig, da die im Kapitel 5 dargestellte Software eine derartige schnelle Berechnung gut unterstützt. Aus inhaltlicher Sicht ergibt sich jedoch meistens die Schwierigkeit, daß selbst bei sorgfältiger Vorbereitung die Ergebnisse der ersten Analyse neben einer möglichen Verletzung der Annahmen der Regressionsanalyse auch zumeist auf Probleme bei der Spezifikation der Regressionsgleichung hinweisen. So treten mitunter positive Preiselastizitäten oder negative Werbeelastizitäten auf. Dies erfordert dann beispielsweise eine in der Regel zeitaufwendige Rücksprache mit den für die Daten Verantwortlichen, die Erhebung zusätzlicher Daten oder eine andere Formulierung der Regressionsgleichung. Deswegen sollten bei empirischen Erhebungen unbedingt auch schon mit kleinen Datensätzen erste statistische Analysen durchgeführt werden, weil diese trotz ihrer geringen Größe bereits Defizite in den Daten andeuten können.

Die in diesem Beitrag dargestellten Zahlenbeispiele bezogen sich entweder auf Querschnittsdaten (Tabelle 1) oder auf Längsschnittsdaten (Tabelle 2). Häufig liegen im Marketing, insbesondere bei der Betrachtung von Paneldaten, aber auch sogenannte ge-

poolte Daten vor, die eine Kombination von Längs- und Querschnittsdaten darstellen. Deren Auswertung erfordert einige Besonderheiten bei der Auswertung, z.B. die Berücksichtigung struktureller Unterschiede zwischen den einzelnen Querschnittsdaten, die in diesem Beitrag nicht behandelt werden konnten. Für deren Behandlung wird auf Maddala (1977) und Hsiao (1986) verwiesen.

Zu guter Letzt sei auf den zunehmenden Trend zur Anwendung nichtlinearer Regressionen hingewiesen. Dies hängt sicherlich auch mit der Verfügbarkeit derartiger Analyse-möglichkeiten in mittlerweile allen gängigen Statistikprogrammen zusammen und bietet den Vorteil, daß nun vergleichsweise einfach auch nichtlineare und nichtlinearisierbare Funktionen geschätzt werden können. Dadurch können noch stärker inhaltliche Erwägungen bei der Formulierung der Regressionsgleichung berücksichtigt werden.

7 Literatur

- Assmus, G. / Farley, J.W. / Lehmann, D.R. (1984), How Advertising Affects Sales: A Meta Analysis of Econometric Results, *Journal of Marketing Research*, 21, 65-74.
- Backhaus, K. / Erichson, B. / Plinke, W. / Weiber, R. (1996), *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung*, 8. Auflage, Berlin et al.
- Chatterjee, S. / Hadi, A.S. (1986), Influential Observations, High Leverage Points, and Outliers in Linear Regressions, *Statistical Science*, 1, 379-416.
- Cliff, A.D. / Ord, J.K. (1973), *Spatial Autocorrelation*, London.
- Dorfman, R. / Steiner, P.O. (1954), Optimal Advertising and Optimal Quality, *American Economic Review*, 44, 826-836.
- Fahrmeir, L. / Kaufmann, H. / Kredler, C. (1984), Regressionsanalyse, in: Fahrmeir, L. / Haberle, A. (Hrsg.), *Multivariate statistische Verfahren*, Berlin et al, 83-154.
- Hair, J.F. / Anderson, R.E. / Tatham, R.L. / Black, W.C. (1992), *Multivariate Data Analysis*, New York et al.
- Hansen, G. (1993), *Quantitative Wirtschaftsforschung*, München.
- Hruschka, H. (1996), *Marketing-Entscheidungen*, München.
- Hsiao, C. (1986), *Analysis of Panel Data*, Cambridge et al.
- Koutsoyannis, A. (1977), *Theory of Econometrics*, 2. Auflage, Houndsmill.
- Lodish, L.L. / Abraham, M.M. / Kalmenson, S. / Livelsberger, J. / Lubetkin, B. / Richardson, B. / Stevens, M.E. (1995), Advertising Works: A Meta-Analysis of 389 Real World Split Cable T.V. Advertising Experiments, *Journal of Marketing Research*, 32, 125-139.
- Maddala, G. S. (1977), *Econometrics*, New York et al.
- Mauerer, N. (1995), *Die Wirkung absatzpolitischer Instrumente. Metaanalyse empirischer Forschungsarbeiten*, Wiesbaden.

Pindyck, R.S. / Rubinfeld, D. (1991), *Econometric Models and Econometric Forecasts*, New York et al.

Schneeweiß, H. (1990), *Ökonometrie*, 4. Auflage, Heidelberg.

Simon, H. (1992), *Preismanagement - Analyse, Strategie, Umsetzung*, Wiesbaden.

Skiera, B. (1996), *Verkaufsgebietseinteilung zur Maximierung des Deckungsbeitrags*, Wiesbaden.

Tellis, G.J. (1988), The Price Sensitivity of Selective Demand: A Meta-Analysis of Econometric Models of Sales, *Journal of Marketing Research*, 25, 391-404.

Wittink, D. R. (1988), *The Application of Regression Analysis*, Needham Heights (Mass.).