

Regressorenauswahl mittels graphentheoretischer Unterstützung

Dietrich Eherler

Lehrstuhl für Wirtschafts- und Sozialstatistik
Wirtschaftswissenschaftliche Fakultät
Friedrich-Schiller-Universität Jena
Carl-Zeiß-Straße 3
07443 Jena
Tel. 03641 943304
Fax 03641 943302
E-mail: d.eherler@wiwi.uni-jena.de

Abstract

Bei der Anwendung von automatisierten Auswahlverfahren von Regressoren in der Regressionsanalyse (z.B. Stepwise Regression) kann es zu systematischen Fehlern kommen. Dies begründet sich in der Insuffizienz des Regressionskriteriums, das besagt, dass eine Variable X Regressor wird, falls bedingte Abhängigkeit mit dem Regressanden Y gegeben der übrigen bereits einbezogenen Regressoren vorliegt. Beim Auftreten dieses Fehlers ist die geläufige Interpretationsmöglichkeit von Regressionskoeffizienten bei nicht standardisierten Variablen nicht mehr gegeben (Cox und Wermuth 1993). Diesen Fehler kann das sogenannte Adjazenzkriterium (Glymour et al. 1997) aufdecken. Es besagt, daß eine Variable X Regressor wird, wenn es keine Teilmenge der betrachteten Regressoren gibt, für die der Regressor X und der Regressand Y bedingt unabhängig sind gegeben einer solchen Teilmenge. Seine Herkunft und seinen Namen hat das Kriterium aus der graphischen Modellierung. Dieser Beitrag stellt ein SAS Programm zu dieser Problematik vor.

1. Einleitung

Die Anwendung von Regressionsmodellen dient letztendlich der Erstellung von Prognosen. Dabei kann in zwei grundsätzlich verschiedene Arten von Prognosen unterschieden werden, die mit regressionsanalytischen Methoden erstellt werden können. Zum einen handelt es sich um *diagnostische* Fragestellungen: Gegeben sind Beobachtungen von den Regressoren und von Interesse ist eine Vorhersage für die Werte des Regressanden. Zum anderen besteht häufig das Interesse, den Wert einer Variablen vorherzusagen, falls andere Variablen auf einen Wert festgesetzt werden. Um letztere Art von *manipulativen* Prognosen zu erstellen, müssen substantielle Kenntnisse über den datengenerierenden Prozess vorliegen. Der Unterschied wird in den Abschnitten 2 und 3 an einfachen Beispielen deutlich gemacht.

Bei beiden Arten der Prognose steht ein Anwender der Fragestellung gegenüber, welche Regressoren in einen Ansatz mit einzubeziehen sind. Zum einen kann es sehr aufwendig sein, eine große Anzahl von Regressoren zu ermitteln und zu messen. Redundante Regressoren sollten vermieden werden. Außerdem, falls ein Regressionsmodell auf Trainingsdaten gelernt und zur Prognose auf einen anderen Datensatz angewandt wird, kann eine kleinere Anzahl von Regressoren zu qualitativ besseren Ergebnissen führen, weil das Problem des Overfitting für das auf den Trainingsdaten ermittelte Modell verringert werden kann. Dies trifft natürlich nicht zu, wenn mit dem selben Datensatz eine Prognose erstellt werden soll. Hier kann mit einer Vergrößerung der Anzahl der Regressoren stets eine Verbesserung der Güte, gemessen an einer Verringerung der Fehlervarianz, erreicht werden. Zum dritten muß klar sein, welche Interpretation den Regressionskoeffizienten zukommen soll. In vielen Anwendungen stehen manipulative Prognosen im Vordergrund. D.h. Regressionskoeffizienten werden in einer Art und Weise interpretiert, dass eine interventive Kontrolle der Regressoren auf der rechten Seite der Gleichung (exogene Variablen) die durch den Regressionskoeffizienten quantifizierte Auswirkung auf die den

Erwartungswert des Regressanden auf der rechten Seite Gleichung (endogene Variable) hat (z.B. Spirtes et al. 1993, S. 1-2).

Zur Lösung des Auswahlproblems der Regressoren werden häufig automatische Verfahren der Regressorenauswahl angewendet wie beispielsweise Schrittweise Regression. Dieser Beitrag liefert einen Ansatz basierend auf graphentheoretischen Überlegungen, wie eine automatische Regressorenauswahl erfolgen kann, bei der systematische Fehler, die bei einer Anwendung der Stepwise Regression auftreten können, aufgedeckt werden können. Die Ausführungen stützen sich dabei ausschließlich auf den für Anwendungen wichtigsten Fall, dass eine multivariate Normalverteilung vorliegt. Für andere Regressionsansätze sind die Überlegungen an entsprechender Stelle zu modifizieren.

2. Stochastische Regression

Gegeben sei ein $n+1$ dimensionaler, multivariat normalverteilter Zufallsvektor $\mathbf{X}=(X_1, \dots, X_n, Y)'$. Dieser ist charakterisiert durch seinen Erwartungswertvektor μ und seine Kovarianzmatrix Σ (vgl. Whittaker 1990, S. 158). Dabei sei die Komponente Y die zu erklärende Variable. Im Fall einer multivariaten Normalverteilung ist die lineare Regression durch

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + U \quad (1)$$

gegeben. Dabei gilt, dass

$$\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

gerade die bedingte Erwartung von Y gegeben X_1, \dots, X_n darstellt. Die Variable U in Gleichung (1) ist somit definiert innerhalb der betrachteten Variablen X_1, \dots, X_n, Y durch

$$U = Y - E(Y|X_1, \dots, X_n) \quad (2)$$

und stellt die Abweichung der bedingten Erwartung von Y dar. Es ist bekannt, dass mit dieser Festlegung die Variable U die folgenden Eigenschaften besitzt, die im klassischen Regressionsmodell explizit vorausgesetzt werden (z.B. Greene 1997, S. 231-5).

Diese Eigenschaften sind (vgl. Steyer und Eid 1993, S. 356-7)

- (i) $E(U|X_1, \dots, X_n) = 0$
- (ii) $E(U) = 0$
- (iii) $\text{Cov}(E(Y|X_1, \dots, X_n), U) = 0$
- (iv) $\text{Var}(Y) = \text{Var}(E(Y|X_1, \dots, X_n)) + \text{Var}(U)$.

Unter den hier gegebenen Voraussetzungen lässt sich ein Regressionsmodell für jede andere Variable als endogene Variable herleiten und immer sind die klassischen Eigenschaften erfüllt. Dies folgt aus elementaren Eigenschaften der bedingten Erwartung (z.B. Spanos 1999).

Die Regressionskoeffizienten $\beta_0, \beta_1, \dots, \beta_n$ sind durch die Kovarianzmatrix Σ darstellbar. Im einzelnen gilt mit $\mathbf{R}_i = \{X_1, \dots, X_n\} \setminus \{X_i\}$

$$\beta_i = \frac{\text{Cov}(Y, X_i | \mathbf{R}_i)}{\text{Var}(X_i | \mathbf{R}_i)} \quad 1 \leq i \leq n \quad (3)$$

$$\beta_0 = E(Y) - \sum_{i=1}^n \beta_i E(X_i) \quad (4)$$

Im allgemeinen werden nicht alle β_i von Null verschieden sein ($0 \leq i \leq n$). Aus Gleichung (3) ist ersichtlich, dass β_i genau dann gleich Null ist, wenn die bedingte Kovarianz gleich Null ist ($1 \leq i \leq n$). In der hier betrachteten Situation einer multivariaten Normalverteilung ist dies der Fall, wenn Y und X_i gegeben den restlichen Variablen bedingt unabhängig sind. Damit ist schon das Regressionskriterium hergeleitet. Eine Variable taucht auf der rechten Seite von Gleichung (1) als Regressor auf, falls diese Variable bedingt abhängig ist vom Regressanden gegeben den restlichen Variablen. Auf dieser Tatsache baut die Schrittweise Regression auf. Die speziellen Auswahlstrategien gehen dabei wie folgt vor (z.B. SAS Institute 1989, S. 1397-8).

1. Vorwärtsselektion

Ausgehend von der 'leeren' Regressionsgleichung wird anhand eines Gütekriteriums (z.B. R^2) derjenige Regressor aufgenommen, der dieses Kriterium optimiert. In einem nächsten Schritt wird ein Regressor aufgenommen, der wiederum das spezifizierte Gütekriterium optimiert. Dieses Vorgehen bricht ab, falls die Verbesserung des Gütekriteriums eine vorgegebene Schranke nicht mehr überschreitet.

2. Rückwärtsselektion

Ausgehend vom kompletten Regressionsmodell wird derjenige Regressor auf der rechten Seite von Gleichung (1) entfernt, der wiederum ein spezifiziertes Gütekriterium nicht mehr als eine vorgegebene Schranke verschlechtert. Dies ist der Fall, wenn eine Variable gegeben den restlichen noch verbleibenden Variablen unabhängig vom Regressanden. Diese Streichung wird solange fortgeführt bis eine Entfernung einer weiteren Variablen das Gütekriterium mehr als die vorgegebene Schranke verschlechtert.

Mit dem hier geschilderten Ansatz lässt sich jede Art der Prognose bearbeiten, die darauf aufbaut, bei beobachteten Regressoren den Erwartungswert, also eine Punktprognose, für den Regressanden zu erstellen. Es bedarf dazu lediglich der Annahme der multivariaten Normalverteilung. Eine Abhängigkeit, bzw. im hier betrachteten Fall, eine Korrelation kann auch eine Scheinkorrelation sein. Wenn es darum geht, aus den gegebenen Informationen über die Variablen X_1, \dots, X_n die Variable Y zu prognostizieren, liefert der Ansatz in Gleichung (1) gemessen am Gütekriterium des L2 Abstands immer die beste Prognose (z.B. Spanos 1999, S. 365). Dabei kann durchaus ein verschwindend geringes Bestimmtheitsmaß (quadrierter multipler Korrelationskoeffizient (z.B. Backhaus et al. 1996, S. 25)) auftreten. Dennoch ist die Regression (1) bei den hier getroffenen Annahmen, bewertet am L2-Abstand das beste, um Y aus den Variablen X_1, \dots, X_n vorherzusagen.

Um den Unterschied zu den Ausführungen im folgenden Abschnitt darzustellen, sollen dies zwei extreme Beispiele verdeutlichen.

Beispiel 2.1

1. Sei X die Anzahl grauer Haare und Y die Sehschärfe. Gleichung (1) besitzt bei diesen Variablen die Gestalt

$$Y = \beta_0 + \beta_1 X + U. \quad (6)$$

Eine Beobachtung einer bestimmten Anzahl grauer Haare bei einer Person lässt bei einer unterstellten positiven Korrelation, Schlussfolgerungen auf die zu erwartende Sehschärfe dieser Person zu. Ein Eingriff, bei dem einer Person die Haare gefärbt werden, hätte dagegen keinerlei Auswirkungen auf die Sehschärfe der Person. Umgekehrt kann eine Regression aufgestellt werden mit der Gleichung

$$X = \alpha_0 + \alpha_1 Y + V. \quad (7)$$

In diesem Fall kann eine Beobachtung der Sehschärfe bei einer Person benutzt werden, um den Erwartungswert von X unter der beobachteten Bedingung zu bestimmen. Würde die Sehschärfe beispielsweise durch eine Laseroperation 'manipuliert', kann keine begründete Aussage über die Anzahl grauer Haare getroffen werden.

2. Ein anderes Beispiel für diesen Sachverhalt ist der folgende. Sei X der Absatz an Badeartikeln und Y der Speiseeiskonsum in einer Periode. Es ist plausibel anzunehmen, dass die beiden Größen stark positiv korrelieren. Für Regressionsansätze analog wie zu Gleichung (6) und (7) hieße dies, dass die entsprechenden Koeffizienten β_1 und α_1 positiv wären. Eine Beobachtung der einen Größe lässt Schlüsse zu auf den Erwartungswert der anderen Größe gegeben dieser Beobachtung. Käme jedoch ein Speiseeishersteller auf die Idee, die betrachtete Population mit Badeartikeln zu 'penetrieren' auf den Wert $x + \Delta x$, dürfte dies, trotz der hohen positiven Korrelation, keine wesentlichen Veränderungen des Eiskonsums nach sich ziehen. Um die Steigerung des

Erwartungswerts des Speiseeiskonsums von $y = \beta_0 + \beta_1 x$ auf den Wert $y + \beta_1 \Delta x$ real zu sehen, müsste der Speiseeishersteller einfach warten, bis er für den Absatz an Badeartikeln den entsprechenden Wert $x + \Delta x$ beobachten kann.

Einen Zugang, wann Regressionskoeffizienten die Bedeutung besitzen, dass eine Festsetzung der Regressoren die über die Koeffizienten quantifizierte Auswirkung auf den Regressanden haben, wird im nächsten Abschnitt vorgestellt. Außer rein stochastischen Annahmen sind dazu weitere Anforderungen zu stellen.

3. Manipulation in linearen Strukturgleichungsmodellen

Bereits Strotz und Wold (1960) geben eine Bedingung an, wann ein Regressionskoeffizient 'kausal' interpretiert werden kann. Dieser Ansatz baut auf lineare Strukturgleichungsmodelle auf. In vielen Anwendungen auf den Gebieten der Ökonometrie, der Psychometrie und Soziologie wird die Annahme der Linearität der Zusammenhänge getroffen. In den letzten Jahren wurde dieser Ansatz aufgegriffen und verallgemeinert im Zusammenhang mit dem gestiegenen Interesse an der graphischen Modellierung von Wahrscheinlichkeitsverteilungen. Die bekanntesten Vertreter dieser Forschungsrichtung sind Spirtes et al. (1993) und Pearl (2000). In dieser Arbeit ist das Hauptaugenmerk auf den Spezialfall linearer Systeme gerichtet.

Definition 3.1

Seien $\mathbf{U}=(U_1, \dots, U_n)'$ unabhängige, multivariat normalverteilte Zufallsvariablen auf einem Wahrscheinlichkeitsraum (Ω, Σ, P) . Weiterhin seien $\mathbf{F}=(f_1, \dots, f_n)'$ lineare Funktionen, so dass Zufallsgrößen $\mathbf{V}=(X_1, \dots, X_n)'$ durch ein lineares Gleichungssystem

$$X_i = \sum_{j \neq i} \alpha_{ij} X_j + U_i \quad 1 \leq i \leq n \quad (8)$$

wohldefiniert sind. Das Tupel $M=(\mathbf{U}, \mathbf{V}, \mathbf{F})$ erhält die Bezeichnung *lineares Strukturgleichungsmodell*. Die Matrix $\mathbf{A}=(\alpha_{ij})$ $i, j=1, \dots, n$ heißt *Koeffizientenmatrix* zu M .

In einem Strukturgleichungsmodell werden die interessierenden Größen X_i in Abhängigkeit von ihren direkten Einflussgrößen X_j und U_i über die lineare Funktion f_i dargestellt. Dabei repräsentieren die Variablen X_j den innerhalb der Variablenmenge \mathbf{V} bestimmten Einfluss auf die Variable X_i . Die Variablen U_i modellieren diejenigen Einflüsse, die nicht durch die Variablen X_j erfasst sind. Idealerweise sollen alle systematischen Einflüsse auf eine Variable X_i durch die Größen X_j erklärt sein und den Variablen U_i die Rolle von Störgrößen zukommen. Somit repräsentieren die Variablen $\mathbf{U}=(U_1, \dots, U_n)'$ konzeptionell die nicht explizit berücksichtigten Einflussfaktoren. Der stochastische Charakter dieser Variablen bestimmt die probabilistischen Charakteristika eines Strukturgleichungsmodells.

Die Momente der Zufallsvariablen $\mathbf{V}=(X_1, \dots, X_n)'$ lassen sich mit elementaren Rechenregeln für Erwartungswerte und Kovarianzen bestimmen. Als Linearkombination multivariat normalverteilter Zufallsvariablen ist der Zufallsvektor \mathbf{V} wiederum multivariat normalverteilt (z.B. Whittaker 1990, S 159). Das nächste Lemma gibt an, wie die Momente von \mathbf{V} bestimmt werden können.

Lemma 3.1

Gegeben sei ein lineares Strukturgleichungsmodell M . Sei $\mu_{\mathbf{U}}$ der Erwartungswertvektor und $\Sigma_{\mathbf{U}}$ die Kovarianzmatrix der Variablen $\mathbf{U}=(U_1, \dots, U_n)'$. Ist die Matrix $\mathbf{S}=(\mathbf{I}-\mathbf{A})$ invertierbar, sind der Erwartungswertvektor $E(\mathbf{V})=\mu_{\mathbf{V}}$ und die Kovarianzmatrix $\text{Kov}(\mathbf{V})=\Sigma_{\mathbf{V}}$ gegeben durch

$$\begin{aligned} \mu_{\mathbf{V}} &= \mathbf{S}^{-1} \mu_{\mathbf{U}} \\ \Sigma_{\mathbf{V}} &= \mathbf{S}^{-1} \Sigma_{\mathbf{U}} (\mathbf{S}^{-1})^T. \end{aligned} \quad (9)$$

Beweis:

Anwendung elementarer Rechenregeln für Momente von Zufallsvektoren.

In der bisher dargestellten Form ist ein lineares Strukturgleichungsmodell ein aus qualitativen Komponenten aufgebautes Objekt. Es stellt auf abstrakte Weise die Interaktionen der untersuchten Variablen $\mathbf{V}=(X_1, \dots, X_n)'$ dar. In praktischen Untersuchungen werden für diese Variable Beobachtungen vorliegen. Diese sind Realisierungen für eine ganz spezielle Festlegung der Parameter der Zufallsgrößen U_1, \dots, U_n und der Koeffizientenmatrix \mathbf{A} . Die Spezifizierung dieser Größen wird daher auch als *Parameterisierung* eines Strukturgleichungsmodells bezeichnet. Für eine gegebenen Parameterisierung kann ein lineares Strukturgleichungsmodell als datengenerierender Prozess aufgefasst werden. Mit diesem Generierungsprozess steht eine wichtige Information zur Verfügung. Es kann untersucht werden, wie sich die Verteilungen der Variablen ändern würden, wenn statt des 'natürlichen', d.h. des beobachteten Generierungsprozesses ein manipulativ veränderter betrachtet wird.

Genau dieses Gedankenexperiment des Eingriffs in den Generierungsprozess greift die folgende Definition einer Manipulation von Variablen auf. Zu einer übersichtlicheren Darstellung werde ohne Beschränkung der Allgemeinheit angenommen, dass es sich bei den manipulierten Variablen um die ersten k Variablen $\mathbf{X}=(X_1, \dots, X_k)'$ handelt und die restlichen Variablen $\mathbf{Y}=(X_{k+1}, \dots, X_n)'$ die sonstigen, nicht manipulierten Variablen darstellen. Dadurch ist eine Partition der Koeffizientenmatrix in vier Blöcke gegeben

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \tag{10}$$

wobei \mathbf{A}_{11} die Interaktion der manipulierten Variablen untereinander, \mathbf{A}_{12} den funktionalen Einfluss der nicht manipulierten Variablen auf die manipulierten Variablen, \mathbf{A}_{21} den funktionalen Einfluss der manipulierten Variablen auf die nicht manipulierten und \mathbf{A}_{22} die Interaktion der nicht manipulierten Variablen beschreibt. Entsprechend sei $\mu_{\mathbf{V}}=(\mu_1, \mu_2)'$ partitioniert.

Definition 3.2

Sei M ein lineares Strukturgleichungsmodell mit einer wie in Gleichung (10) partitionierten Koeffizientenmatrix. Eine *Manipulation* der Variablen X_1, \dots, X_k ist die Substitution der ersten k Gleichungen durch die Festsetzungen der Variablen X_1, \dots, X_k auf spezifizierte Realisationen x_1, \dots, x_k , kurz $\mathbf{X}'=\mathbf{x}'$. Dann ist das manipulierte Gleichungssystem unter der Manipulation $\mathbf{X}'=\mathbf{x}'$ mit $\mathbf{A}_M=\mathbf{A}_{21}$ und $\mathbf{A}_X=\mathbf{A}_{22}$ gegeben durch

$$\begin{pmatrix} X_{k+1}^x \\ \vdots \\ X_n^x \end{pmatrix} = \mathbf{A}_M \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} + \mathbf{A}_X \begin{pmatrix} X_{k+1} \\ \vdots \\ X_n \end{pmatrix} + \begin{pmatrix} U_{k+1} \\ \vdots \\ U_n \end{pmatrix}. \tag{11}$$

Als verkürzte Schreibweise wird dafür

$$\mathbf{Y}^x = \mathbf{A}_M \mathbf{x} + \mathbf{A}_X \mathbf{Y} + \mathbf{U}_x \tag{12}$$

verwendet. Die Verteilung der Variablen \mathbf{Y}^x erhält die Bezeichnung *manipulierte Verteilung*.

Es ist anzumerken, dass die manipulierte Verteilung kontrafaktischer Natur ist. Erst wenn eine der Manipulation entsprechende Intervention, d.h. eine Festsetzung der Variablen X_1, \dots, X_k tatsächlich durchgeführt würde, könnte die manipulierte Verteilung beobachtet werden. Liegen jedoch Kenntnisse über den datengenerierenden Prozess in Form des Strukturgleichungsmodells vor, kann die manipulierte Verteilung aus der ursprünglichen Verteilung bestimmt werden. Diese ist als Linearkombination von multivariat normalverteilten Zufallsgrößen und Konstanten selbst multivariat normalverteilt (z.B. Whittaker 1990, S. 159).

Satz 3.1

Sei M ein lineares Strukturgleichungsmodell. Die Verteilung der manipulierten Variablen \mathbf{Y}^x unter der Manipulation $\mathbf{X}'=\mathbf{x}'$ besitzt den Erwartungswertvektor

$$\mu_x = (\mathbf{I} - \mathbf{A}_x)^{-1} \mu_2 + \mathbf{A}_M \mathbf{x} \quad (13)$$

und die Kovarianzmatrix

$$\Sigma_x = (\mathbf{I} - \mathbf{A}_x)^{-1} \Sigma_{\mathbf{U}_x} [(\mathbf{I} - \mathbf{A}_x)^{-1}]^T \quad (14)$$

wobei \mathbf{I} die $(n-k) \times (n-k)$ Einheitsmatrix ist und $\Sigma_{\mathbf{U}_x}$ die Kovarianzmatrix von U_{k+1}, \dots, U_n .

Beweis:

Analog wie bei Lemma 3.1.

Die manipulierte Verteilung gibt Auskunft darüber, wie eine Variable auf die Festsetzung anderer Variablen reagiert. Sie spiegelt damit den eingangs erwähnten Sachverhalt einer Prognose für die abhängigen Variablen wider, falls durch eine Entscheidung oder Intervention eine Veränderung der disponierbaren Variablen herbeigeführt wird. Die folgende Definition formalisiert diese Überlegungen.

Definition 3.3

Der *erwartete Effekt* einer Manipulation $\mathbf{X}'=\mathbf{x}'$ in einem linearen Strukturgleichungsmodell auf die abhängige Variablen \mathbf{Y} ist gegeben durch $E(\mathbf{Y}^x)$ und wird symbolisiert durch $E_x(\mathbf{Y})$. Ist X

eine Variable aus \mathbf{V} und Y eine andere Variable aus \mathbf{V} , so heißt $\tau_{XY} = \frac{\partial}{\partial x} E_x(Y)$ der *totale Effekt* von X auf Y . Ist \mathbf{Z} die Menge der anderen direkten Einflussvariablen in der Gleichung für Y im Gleichungssystem (8) vereinigt mit X , so heißt $\frac{\partial}{\partial x} E_z(Y)$ der *direkte Effekt* von X auf Y .

Das folgende Beispiel illustriert die eingeführten Begriffe. Außerdem verdeutlicht es, worin der Unterschied liegt zwischen der Betrachtung einer manipulierten Verteilung und einer bedingten Verteilung, insbesondere dann, wenn eine Vorhersage der Auswirkung einer Festsetzung einer Variablen ermittelt werden soll.

Beispiel 3.1

Sei M ein lineares Strukturgleichungsmodell mit Gleichungen für die Variablen

K	Kopfschmerzen in einer Population
N	Nachfrage nach Analgetika
P	Preis für Analgetika
B	Budgetverbrauch zur Behandlung von Kopfschmerzen.

Die Koeffizientenmatrix sei gegeben als

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0,8 & 0 & -1,25 & 0 \\ 0 & 0,8 & 0 & 0 \\ 0,6 & 0 & 1 & 0 \end{pmatrix}.$$

Der Zufallsvektor \mathbf{U} sei multivariat normalverteilt mit den Parametern

$$\mu_{\mathbf{U}} = \begin{pmatrix} 5 \\ 7 \\ 4 \\ 2 \end{pmatrix} \quad \Sigma_{\mathbf{U}} = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2,56 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}.$$

Daraus ergibt sich die präintervenive (beobachtbare) Verteilung der $\mathbf{V}=(K,N,P,B)'$, die multivariat normal ist, mit

$$\mu = \begin{pmatrix} 5 \\ 3 \\ 6,4 \\ 11,4 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 4 & 1,6 & 1,28 & 3,68 \\ 1,6 & 1,89 & -0,09 & 0,87 \\ 1,28 & -0,09 & 2,56 & 1,98 \\ 3,68 & 0,87 & 1,98 & 8,19 \end{pmatrix}.$$

Mit diesem Modell soll nun prognostiziert werden, wie sich die Aufwendungen zur Linderung von Kopfschmerzen verändern, wenn der Preis für Analgetika auf 8 (Geldeinheiten) durch eine Manipulation fixiert wird. Die Parameter der manipulierten Verteilung geben Auskunft über die zu erwartenden Veränderungen. Mit Gleichung (13) und (14) resultieren

$$\mu_{P=8} = \begin{pmatrix} 5 \\ 1 \\ 13 \end{pmatrix} \quad \Sigma_{P=8} = \begin{pmatrix} 4 & 3,2 & 2,4 \\ 3,2 & 3,56 & 1,92 \\ 2,4 & 1,92 & 5,44 \end{pmatrix}.$$

Unter der Manipulation folgt, dass die Nachfrage nach Analgetika im Erwartungswert auf 1 zurückgeht, jedoch die Belastung des Budgets von 11,4 auf 13 Geldeinheiten ansteigt. Dieser Wert berechnet sich mit Gleichung (13) als

$$E_{P=8}(B) = E[0,6K + 1,0(P=8)] = 0,6E(K) + 1 \cdot 8 = 5 + 8.$$

Der totale Effekt auf den Budgetverbrauch bei einer Festsetzung von $P=p$ ist durch

$$\tau_{XY} = \frac{\partial}{\partial p} E_p(B) = \frac{\partial}{\partial p} [0,6E(K) + 1p] = 1$$

gegeben. Dieser stimmt in diesem Beispiel mit direkten Effekt

$$\frac{\partial}{\partial p} E_{pk}(B) = \frac{\partial}{\partial p} [0,6k + 1p] = 1$$

überein, da der Preis für Analgetika keinen indirekten Einfluss auf den Budgetverbrauch hat.

Würde anstatt der manipulierten Verteilung unter der Manipulation $P=8$ die bedingte Verteilung unter der Bedingung $P=8$ betrachtet, veränderte sich auch die Verteilung der Hintergrundvariablen und der Variablen, die nicht von einer Manipulation beeinflusst werden. Eine solche Vorgehensweise ist zur Prognose des Effektes einer Intervention ungeeignet (Balke und Pearl 1995). Beispielsweise verändert sich die Verteilung der Kopfschmerzerkrankten unter der Bedingung $P=8$ zu $K|(P=8) \sim N(6,69, 2,65)$, d.h. der Erwartungswert würde von 5 auf 6.69 ansteigen. Da jedoch der Effekt einer Fixierung vorhergesagt werden soll, ist es nicht adäquat anzunehmen, es gäbe im Mittel plötzlich mehr Leute mit Kopfschmerzen in der betrachteten Population.

Im obigen Beispiel wird deutlich, dass zur Bestimmung der Prognosen von Interventionen die Kenntnis der Koeffizientenmatrix von essentieller Bedeutung ist. Wie ganz spezielle Koeffizienten konsistent mit einem Regressionsansatz ermittelt werden können, soll daher im nächsten Abschnitt entfaltet werden.

4. Regression in Strukturgleichungsmodellen

Mit den in Abschnitt 3 vorgestellten linearen Strukturgleichungsmodellen können Prognosen von Interventionen vorhergesagt werden aus beobachteten Daten. Dies ist eine häufig zu bewältigende Aufgabenstellung. In diesem Abschnitt soll geklärt werden, wie eine Auswahl geeigneter Regressoren erfolgen kann. Dazu hilfreich ist die Darstellung des Strukturgleichungsmodells als gerichteter Graph.

Definition 4.1

Der zu einem Strukturgleichungssystem M *gehörende Graph* G besteht aus den Knoten $\mathbf{V}=\{X_1, \dots, X_n\}$ und den Kanten $X_j \rightarrow X_i$ für alle Indizes j , bei denen α_{ij} in Gleichung (8) von Null verschieden ist.

Der Graph zu einem Strukturgleichungsmodell stellt eine intuitive Darstellungsform dar, die komplexe Analysen mit allein graphentheoretischen Überlegungen erlaubt. Weil immer dort eine Kante auftaucht, wo der Koeffizient α_{ij} von Null verschieden ist, heißen die Elemente der Koeffizientenmatrix auch Pfadkoeffizienten (Wright 1923).

Eine erste wichtige Eigenschaft des Graphen gibt Satz 4.1 wieder. Über den graphentheoretischen Begriff der d -Separation (Pearl 1988, S. 117) ermittelte Trennungseigenschaften sind auf die probabilistische Ebene direkt übertragbar. Beispiele, die diesen Begriff erläutern erfolgen nach Vorstellung der zentralen Resultate in diesem Abschnitt.

Definition 4.2 (vgl. Pearl 1988, S. 117)

Gegeben sei ein gerichteter azyklischer Graph G . Seien X und Y zwei Knoten von G und sei \mathbf{Z} eine Teilmenge der Knoten mit $X, Y \notin \mathbf{Z}$. Dann wird X durch \mathbf{Z} von Y d -separiert (Symbol: $\langle X | \mathbf{Z} | Y \rangle$), wenn gilt:

Für jeden ungerichteten Pfad zwischen X und Y existiert ein Knoten W auf dem Pfad mit

1. W besitzt aufeinander zulaufende Kanten $\rightarrow W \leftarrow$ und W sowie seine Nachfolger gehören nicht zu \mathbf{Z} ,

oder

2. W besitzt nicht aufeinander zulaufende Kanten $\rightarrow W \rightarrow$, $\leftarrow W \rightarrow$ oder $\leftarrow W \leftarrow$ und W gehört zu \mathbf{Z} .

Zwei Mengen von Knoten \mathbf{X} und \mathbf{Y} werden durch eine Menge \mathbf{Z} d -separiert, falls \mathbf{Z} jeden Knoten $X \in \mathbf{X}$ von jedem Knoten $Y \in \mathbf{Y}$ d -separiert. Ein Pfad zwischen zwei Knoten X und Y ist durch \mathbf{Z} d -separiert, wenn auf diesem Pfad ein Knoten W liegt mit der Eigenschaft 1. oder 2.

Auch für gerichtete Graphen, die Zyklen enthalten ist der d -Separationsbegriff anwendbar. Zwei Knoten X und Y sind d -separiert durch eine Menge \mathbf{Z} , falls jeder Pfad zwischen X und Y durch \mathbf{Z} d -separiert ist.

Satz 4.1

Gegeben sei ein lineares Strukturgleichungsmodell M . Sind die Zufallsvariablen \mathbf{U} gemeinsam unabhängig, so gilt

$$\langle \mathbf{X} | \mathbf{Z} | \mathbf{Y} \rangle \Rightarrow \mathbf{X} \text{ und } \mathbf{Y} \text{ sind unabhängig gegeben } \mathbf{Z}$$

für beliebige disjunkte Teilmengen \mathbf{X} , \mathbf{Y} und \mathbf{Z} von \mathbf{V} .

Beweis:

Spirtes et al. (1998, S. 218)

Dieser Satz bildet die Grundlage für die weiteren Aussagen. Der erste Satz besagt, wie ein einzelner Koeffizient mit einem Regressionsansatz bestimmt werden kann.

Satz 4.2

Sei M ein lineares Strukturgleichungsmodell mit multivariat normalverteilten Zufallsvariablen und G der zugehörige Graph mit einer Kante $X \rightarrow Y$. Sei G_α der aus G entstehende Graph,

indem die Kante $X \rightarrow Y$ aus G entfernt wird. Der zur Kante $X \rightarrow Y$ gehörende Pfadkoeffizient α ist identifizierbar, falls eine Menge $Z \subset V \setminus \{X, Y\}$ existiert mit

1. Z enthält keine Nachfolger von Y
2. Z d-separiert X und Y in G_α

Genügt eine Menge Z diesen Bedingungen, so gilt

$$\alpha = r_{XY.Z} = \frac{\text{Cov}(Y, X | Z)}{\text{Var}(X | Z)} \tag{15}$$

d.h. α ist gleich dem bedingten Regressionskoeffizienten.

Falls Z nicht die beiden Bedingungen erfüllt, dann ist $r_{XY.Z}$ p-fast sicher kein konsistenter Schätzer für α .

Beweis:

Spirtes et al. (1998, S. 201)

Dieses Kriterium enthält einen Test, der ohne weitere Matrizen algebraische Analysen direkt auf den Graphen angewendet werden kann und eine Entscheidung darüber liefert, ob ein Effekt identifizierbar ist. Zudem beantwortet das Kriterium darüber hinaus, wie der Effekt bestimmt werden kann und wann der Effekt mit dem entsprechenden Regressionskoeffizienten übereinstimmt. Die Wahl der Menge Z kann so getroffen werden, dass bezüglich eines Kriteriums wie des Schätzfehlers optimiert wird. Eine weitere Konsequenz aus Satz 4.2 ist, dass (p-fast sicher) ein Pfadkoeffizient nicht über einen Regressionsansatz bestimmt werden kann, wenn der zugehörige Graph zu einem linearen Strukturgleichungsmodell einen Zyklus bezüglich X und Y enthält oder die Hintergrundvariablen U_X und U_Y korreliert sind (vgl. Spirtes et al. 1998, S. 201).

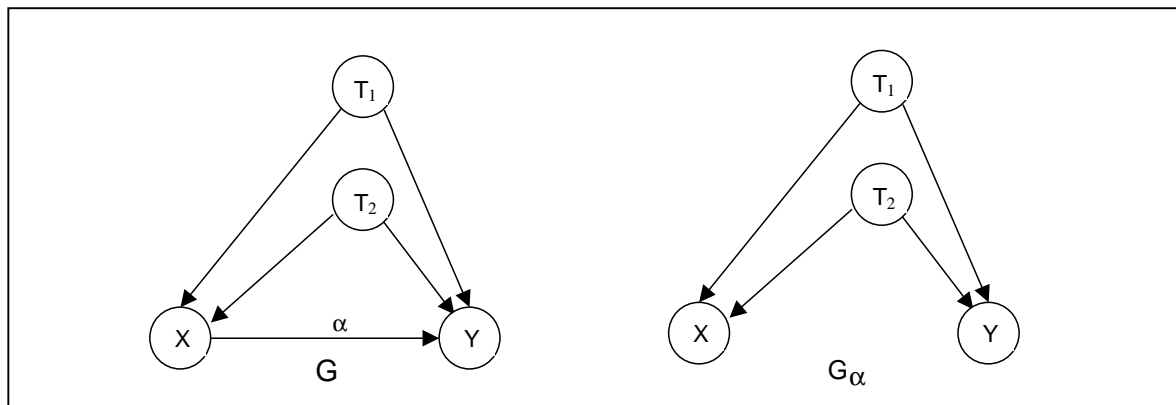
Beispiel 4.1

Sei M ein lineares Strukturgleichungsmodell und sei G in Abbildung 4.1 der zugehörige Graph. Die Hintergrundvariablen U_1, \dots, U_4 seien standard, multivariat normalverteilt.

Das Gleichungssystem sei gegeben durch

$$\begin{aligned} T_1 &= U_1 \\ T_2 &= U_2 \\ X &= \gamma_1 T_1 + \gamma_3 T_2 + U_3 \\ Y &= \alpha X + \gamma_2 T_1 + \gamma_4 T_2 + U_4 \end{aligned}$$

Abbildung 4.1



In G_α sind X und Y durch $Z = \{T_1, T_2\}$ d-separiert, denn auf jedem Pfad zwischen X und Y existiert ein Knoten, der die Eigenschaft 2. der d-Separation erfüllt. Außerdem ist weder T_1

noch T_2 ein Nachfolger von Y . Mit Satz 4.2 gilt, dass α durch $\beta_1 = r_{XY.Z}$ über einen Regressionsansatz

$$Y = \beta_0 + \beta_1 X + \beta_2 T_1 + \beta_3 T_2 + V \quad (16)$$

konsistent geschätzt werden kann. Dabei können X und T_i hochgradig abhängig sein ($i=1,2$).

Das Kriterium in Satz 4.2 ist hinreichend zur Identifikation von einzelnen Pfadkoeffizienten. Es gibt jedoch Fälle, in denen zwar nicht ein einzelner Koeffizient identifizierbar ist, jedoch der gesamte Effekt einer Manipulation auf eine Zielvariable, also einschließlich indirekter Effekte, bestimmt werden kann (vgl. Pearl 2000, S. 151). Besteht primäres Interesse in der Ermittlung des totalen Effektes, so kann das sogenannte Backdoor-Kriterium im folgenden Satz 4.3 direkt zur Anwendung kommen, ohne dass ein Augenmerk auf die indirekten, über gerichteten Pfade laufende Beeinflussungen, gelegt wird.

Satz 4.3

Sei M ein lineares Strukturgleichungsmodell mit multivariat normalverteilten Zufallsvariablen und G der zugehörige Graph. Sei G_X der aus G entstehende Graph, indem alle von X ausgehenden Kanten entfernt werden. Der (totale) Effekt von einer Variablen X auf eine Variable $Y \in V \setminus \{X\}$ ist identifizierbar, falls eine beobachtbare Menge $T \setminus \{X, Y\}$ mit den Eigenschaften

1. für alle $T \in T$ gilt: T ist kein Nachfolger von X
2. T d-separiert X von Y im Graphen G_X

existiert. Sind die zwei Bedingungen erfüllt, ist der (totale) Effekt τ_{XY} durch den bedingten Regressionskoeffizienten

$$r_{XY.T} = \frac{\text{Cov}(Y, X | T)}{\text{Var}(X | T)} \quad (17)$$

gegeben.

Beweis:

Pearl (1998, S. 22)

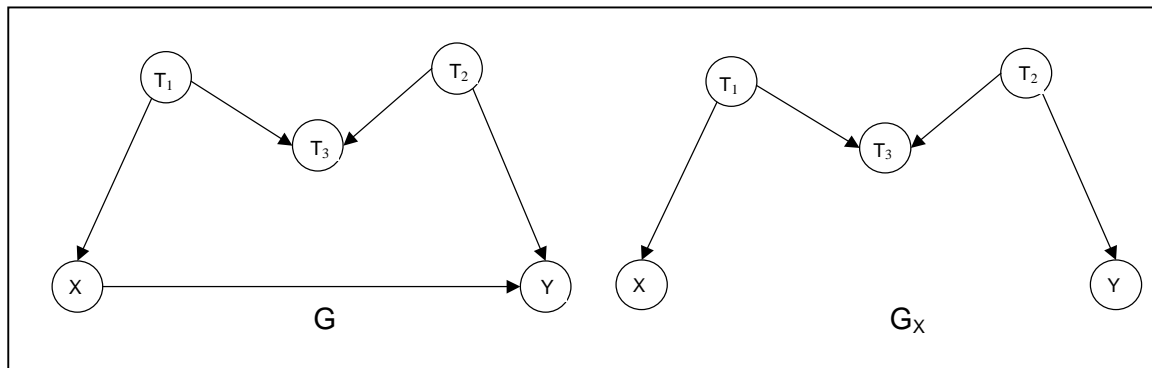
Eine Anwendung dieser Formel sorgt für eine Adjustierung der Einflüsse derjenigen Variablen, die sowohl auf X als auch auf Y einen Einfluss haben, da in G_X nur noch Pfade X und Y verbinden, die auf X sowie auf Y zulaufende Kante haben. Diese Variablen erzeugen eine Scheinabhängigkeit zwischen X und Y , die bei einer Beurteilung des Einflusses von X auf Y berücksichtigt werden muss. Genügt eine Menge T den Eigenschaften 1. und 2. in Satz 4.2, ist diese Scheinabhängigkeit 'neutralisiert'. Variablen mit der Eigenschaft, sowohl X als auch Y potenziell zu beeinflussen, werden auch als konfundierende Variablen bezeichnet, da die Tatsache ihrer Existenz zur Verwirrung über das Ausmaß eines Effekts führen kann (z.B. Simpsonsches Paradoxon (Simpson 1951), Greenland et al., 1999). Häufig wird empfohlen, zu einer Adjustierung möglichst viele Variablen zu verwenden. Ändert sich der Schätzer für den Koeffizienten durch Hinzunahme weiterer Variablen nicht wesentlich, ist davon auszugehen, dass eine adäquate Menge zu einer Adjustierung gefunden worden ist (vgl. Pratt und Schlaiffer 1988).

Dass jedoch erst durch eine Einbeziehung weiterer Variablen eine Verzerrung hervorgerufen wird, ist selten bekannt (vgl. Spirtes et al. 1998, S. 189-91). Das folgende Beispiel illustriert eine solche Situation.

Beispiel 4.2

Sei M ein lineares Strukturgleichungsmodell mit dem in Abbildung 4.2 dargestellten zugehörigen Graphen.

Abbildung 4.2



In G_X sind X und Y durch die leere Menge d -separiert. Auf dem einzigen Pfad zwischen X und Y liegt ein Knoten mit aufeinander zulaufenden Kanten und dieser Knoten darf nicht in der Separierungsmenge enthalten sein. Zur Schätzung des Ausmaßes einer Festsetzung von X auf Y kann der Ansatz

$$Y = \beta_0 + \beta_1 X + V \quad (18)$$

gewählt werden. Es gilt $\beta_1 = r_{XY}$.

In einer empirischen Untersuchung kann es durchaus der Fall sein, dass von den potenziell konfundierenden Variablen nur T_3 beobachtet ist. Da T_3 sowohl von X als auch von Y abhängig ist, würde T_3 zu einer Adjustierung verwendet. Doch sowohl Satz 4.2 als auch Satz 4.3 sagen aus, dass $\beta_1 = r_{XY \cdot T_3}$, ermittelt aus dem Ansatz

$$Y = \beta_0 + \beta_1 X + \beta_2 T_3 + V, \quad (16)$$

kein konsistenter Schätzer für eine Quantifizierung der Auswirkung einer Festsetzung von X auf Y ist.

Eine Behandlung dieser Problematik soll an dieser Stelle nicht stattfinden. Operationale Kriterien zur Feststellung, welche Adjustierungsmenge hinreichend ist zur korrekten Bestimmung eines Effekts sind in Kischka und Eherler (2000, 2001) gegeben.

5. Graphische versus Schrittweise Selektion von Regressoren

5.1 Diagnostische Prognosen

In Abschnitt 2 ist der stochastische Regressionsansatz vorgestellt worden. Sollen diagnostische Prognosen erstellt werden, die darauf abzielen, bei gegebenen Beobachtungen von Variablen eine andere vorherzusagen, so spielt es keine Rolle, ob die zur Prognose notwendige Abhängigkeit aufgrund einer Scheinkorrelation besteht oder ob ein struktureller Zusammenhang zwischen den Größen vorliegt. Zur Einsetzbarkeit des Regressionsansatzes für solche Prognosen ist allein das Vorliegen einer entsprechenden Abhängigkeit erforderlich, die gewährleistet, dass die durch Gleichung (3) bestimmten Koeffizienten von Null verschieden sind. Oder anders ausgedrückt, jede Art von Abhängigkeit ist nützlich zu einer Prognose im hier betrachteten Sinn. Zu einer Komplexitätsreduzierung in umfangreichen Problemstellungen können jedoch auch in diesem Fall Methoden der schrittweisen Auswahl oder der graphischen Modellierung eingesetzt werden (z.B. Jordan 1998). Beispiele hierfür sind Expertensysteme auf probabilistischer Basis, bei denen bestehende Unabhängigkeiten zwischen Variablen ausgenutzt werden, um effizientere Berechnungsvorschriften für Inferenzprobleme herzuleiten (Cowell et al. 1999).

5.2 Manipulative Prognosen

Eine andere Situation ist gegeben, wenn durch einen Regressionsansatz die Auswirkung einer Intervention bzw. einer Festsetzung von Variablen vorhergesagt werden soll. Die Beispiele am Ende von Abschnitt 2 und Abschnitt 3 verdeutlichen, dass in diesem Fall substantielle Überlegungen angestellt werden müssen, um zu einer adäquaten Auswahl von Regressoren zu gelangen. Zwischen Regressoren und Regressanden muss eine strukturelle Abhängigkeit vorliegen. Im Abschnitt 4 wurde dafür ein Ansatz vorgestellt, der auf Strukturgleichungsmodelle aufbaut. Zur Anwendung dieses Ansatzes muss natürlicher Weise gegeben sein, dass die betrachtete Problemstellung sich auf angemessene Weise durch ein Strukturgleichungsmodell abbilden lässt. Davon soll im folgenden ausgegangen werden, d.h. für die betrachtete Problemstellung existiert ein Strukturgleichungsmodell nach Definition 3.1.

In diesem Fall liefert der zugehörige Graph wichtige Erkenntnisse.

5.2.1 Vergleich der Auswahlkriterien

Eine zentrale Aussage zum Auswahlproblem von Regressoren liefert der folgende Satz über d-Separation. Dieser charakterisiert, unter welcher Bedingung zwei Variablen X und Y in einem Graphen adjazent sind. In die Sprache der Regressionsanalyse übersetzt geben diese Kriterien an, wann eine Variable als Regressor verwendet werden soll.

Satz 5.1

Sei G ein gerichteter azyklischer Graph mit der Knotenmenge V . Dann gilt

1. Zwei beliebige Knoten X und Y sind genau dann adjazent in G , wenn keine Teilmenge T von $V \setminus \{X, Y\}$ existiert, die X und Y d-separiert.
2. Für beliebige Knoten X , Y und Z mit der Eigenschaft, dass X und Y sowie Z und Y adjazent, jedoch X und Z nicht adjazent sind, ist genau dann $X \rightarrow Y \leftarrow Z$ ein Teilgraph von G , wenn X und Y durch keine Teilmenge Q von $V \setminus \{X, Z\}$ d-separiert sind, die Y enthält.

Beweis:

Verma und Pearl (1991).

Teil 1 des Satzes liefert ein (notwendiges) Kriterium, wann eine Variable als Regressor verwendet werden soll. Um das graphische Kriterium in die probabilistische Sprache zu übersetzen, bedarf es jedoch der sogenannten *Faithfulnessbedingung*. Diese sagt aus, dass sich strukturelle Abhängigkeit bzw. Unabhängigkeit im Graphen, repräsentiert durch d-Separation und probabilistische Abhängigkeit bzw. Unabhängigkeit eindeutig entsprechen. Es gilt immer, dass d-Separation Unabhängigkeit impliziert (vgl. Satz 4.1). Es gibt jedoch Ausnahmen, bei denen Unabhängigkeit vorliegt, obwohl eine strukturelle Abhängigkeit vorhanden ist. Spirtes et al. (1993, S. 95-96) zeigen, dass solche Ausnahmen im hier betrachteten Fall normalverteilter Modelle p-fast unmöglich sind.

Eine Folgerung aus Satz 5.1 ist das folgende Auswahlkriterium. Demgegenüber gestellt ist noch mal das schrittweise Auswahlkriterium aus Abschnitt 2.

Adjazenzkriterium (*Graphisches Selektionskriterium*):

X_i ist ein Regressor für die Variable Y , falls X_i und Y gegeben jeder beliebigen Teilmenge T der restlichen potenziellen Regressoren bedingt abhängig sind.

Regressionskriterium (*Schrittweises Selektionskriterium*):

X_i ist ein Regressor für die Variable Y , falls X_i und Y gegeben den restlichen potenziellen Regressoren bedingt abhängig sind.

Aus Satz 5.2 kann gefolgert werden, dass sich die beiden Ansätze nicht unterscheiden, wenn alle relevanten Variablen mit einbezogen sind und die Menge der potenziellen Regressoren nur aus Variablen besteht, die nicht vom Regressanden beeinflusst werden.

Satz 5.2

Sei G ein gerichteter azyklischer Graph mit der Knotenmenge V . Sei Y ein Knoten, PA_Y die Menge Elternknoten von Y und X die Menge der Nichtnachfolger von Y in G . Dann gilt, dass PA_Y den Knoten Y von $R=X \setminus PA_Y$ d-separiert, d.h. $\langle Y | PA_Y | R \rangle$.

Beweis:

Da $R=X \setminus PA_Y$ aus der Menge der Nichtnachfolger ohne die Eltern von Y besteht, verläuft jeder Pfad in G mit nicht aufeinander zulaufenden Kanten von einem Knoten aus $R \in R$ nach Y über PA_Y , d.h. auf jedem ungerichteten Pfad zwischen $R \in R$ und Y liegt ein Elternknoten. Hätte ein Elternknoten auf einem solchen Pfad aufeinander zulaufende Kanten, wäre dieser Knoten kein Elternknoten. Somit existiert auf jeder Verbindung zwischen $R \in R$ und Y mit nicht aufeinander zulaufenden Kanten ein Elternknoten, der die Bedingung 2 in Definition 4.2 erfüllt.

Folgerung 5.1

Besteht die Menge der potenziellen Regressoren nur aus Variablen, die nicht von Y beeinflusst werden, liefern graphisches und schrittweises Selektionskriterium unter Faithfulness die gleichen Regressoren.

Beweis:

Schrittweise Auswahl: Gegeben den Elternvariablen von Y sind nach Satz 5.2 alle anderen potenziellen Regressoren von Y bedingt unabhängig. Da bei einer schrittweisen Auswahl die Eltern berücksichtigt werden, werden die restlichen Variablen eliminiert, wenn alle Eltern PA_Y einbezogen sind. Folglich liefert eine schrittweise Auswahl die Menge PA_Y als Regressoren.

Graphische Auswahl: Es gibt keine Teilmenge T der potenziellen Regressoren, für die eine Variable $X_i \in PA_Y$ bedingt unabhängig ist von Y . Folglich wählt die graphische Auswahl nach Satz 5.2 ebenfalls die Menge PA_Y als Regressoren.

Sollen manipulative Prognosen erstellt werden, ist die Wahl der Regressoren von entscheidender Bedeutung. In Anwendungen kann a-priori seltenst sicher gestellt werden, dass alle relevanten Variablen berücksichtigt sind. Es kann auch nicht immer mit Sicherheit ausgeschlossen werden, dass in der Menge der Regressoren keine Variablen sind, die selbst vom Regressanden beeinflusst werden. Beispiele hierfür können sich in der Kundenzufriedenheitsanalyse ergeben. Bei einer Modellierung in diesem Themenbereich stellt sich beispielsweise die Frage, ob die Gesamtzufriedenheit aus Teilzufriedenheiten mit Qualität, Preis etc. eines Produktes erklärt werden kann oder ob umgekehrt die Gesamtzufriedenheit die Zufriedenheit mit Teilaspekten erklärt. Dabei ist es jedoch wichtig, wenn die Gesamtzufriedenheit verbessert werden soll, die Regressoren für diese zu kennen.

Das letztere Problem kann bei einer schrittweisen Auswahl von Regressoren zu systematischen Fehlern führen, die das Adjazenzkriterium im Vergleich zum Regressionskriterium aufdeckt (siehe Beispiel 5.2). Bei ausgelassenen Variablen sind weitergehende Überlegungen anzustellen, die den Umfang dieser Abhandlung sprengen. Möglichkeiten zur Bewältigung dieser Problemstellung sind beispielsweise in Glymour et al. (1997), Greenland et al. (1999) und Kischka und Eherler (2000,2001) angegeben.

5.2.2 Implementierung mit SAS

Ein SAS Programm zur graphischen Selektionsauswahl wurde unter Verwendung von PROC Corr und PROC Iml erstellt. Das Ergebnis des Programms kann mit der PROC Reg weiterverarbeitet werden. Das folgende Schema wird dabei abgearbeitet.

Eingabe: Daten, die Stichprobe einer multivariaten Normalverteilung sind, ein Regressand Y und die Menge der potenziellen Regressoren
 Ausgabe: Liste von Regressoren

Verarbeitungsschritte

1. Aufruf von PROC Corr zur Bestimmung der Kovarianzmatrix
2. Einlesen der Kovarianzmatrix in IML und Bestimmung der bedingten Kovarianzen
3. Entscheidung, ob bedingte Kovarianz signifikant von Null verschieden ist

Der zweite und dritte Schritt wird solange wiederholt, bis für jede Teilmenge der noch verbleibenden Regressoren untersucht worden ist, ob zwischen dem Regressanden und dem aktuell betrachteten Regressoren eine bedingte Abhängigkeit vorliegt.

Wiederhole

- Wähle X_i
- Wiederhole

Teste, ob Y und X_i gegeben jeder Teilmenge von $\mathbf{V}\{Y, X_i\}$ abhängig sind

Bis Y und X_i unabhängig sind gegeben einer Teilmenge oder bis jede Teilmenge untersucht worden ist

Bis alle X_i abgearbeitet sind.

Als Teststatistik in der zweiten Schleife kann Fishers Z verwendet werden.

$$Z(\hat{\rho}_{X_i Y, \mathbf{T}}, n) = \frac{1}{2} \sqrt{n - |\mathbf{T}| - 3} \cdot \ln \left[\frac{1 + \hat{\rho}_{X_i Y, \mathbf{T}}}{1 - \hat{\rho}_{X_i Y, \mathbf{T}}} \right] \quad (19)$$

Dabei ist $\hat{\rho}_{X_i Y, \mathbf{T}}$ die geschätzte partielle Korrelation von X_i und Y gegeben \mathbf{T} , $|\mathbf{T}|$ die Mächtigkeit der Bedingungs Menge und n der Stichprobenumfang. Bei multivariater Normalverteilung ist $Z(\rho_{X_i Y, \mathbf{T}}, n) - Z(\hat{\rho}_{X_i Y, \mathbf{T}}, n)$ standard normalverteilt (Anderson 1984).

Die SAS Kommandozeilen zu Schritt 1 sind wie folgt.

```
* Kovari anzmatri x;
PROC CORR COV DATA=Ei ngabe OUP=Korrmatr ix;
RUN;
```

Die Kovarianzmatrix wird um eine Spalte ergänzt, die Auskunft über die Adjazenz des jeweiligen Regressoren zum Regressanden gibt. Die erste Zeile bzw. erste Spalte der Kovarianzmatrix enthält die Kovarianzen des Regressanden mit den Regressoren.

```
* Erwei tern der Kovari anz um Adj azenzspal te;
DATA Korrmatr ix_erwei tert;
SET Korrmatr ix;
Adj azenz+1;
RUN;
```

Die SAS-Datei Korrmatr ix_erwei tert wird in PROC IML weiterverarbeitet, indem die im Kasten aufgeführten Schleifen durchlaufen werden. Sind X_i und Y gegeben einer Teilmenge unabhängig, wird der Eintrag in der Adjazenzspalte gleich Null gesetzt.

Nach Ablauf der beiden Schleifen wird die Adjanzenspalte weiterarbeitet zu folgender Ausgabe:

Liste der moeglichen Regressoren

XK
x2
x3
x8

Anhand der folgenden Beispiele sollen die beiden Auswahlstrategien verglichen werden.

Beispiel 5.1

Angenommen, eine multivariate Normalverteilung resultiere aus dem folgenden linearen Strukturgleichungsmodell M mit unabhängigen, normalverteilten Zufallsgrößen U_1, \dots, U_4 .

$$X_1 = U_1$$

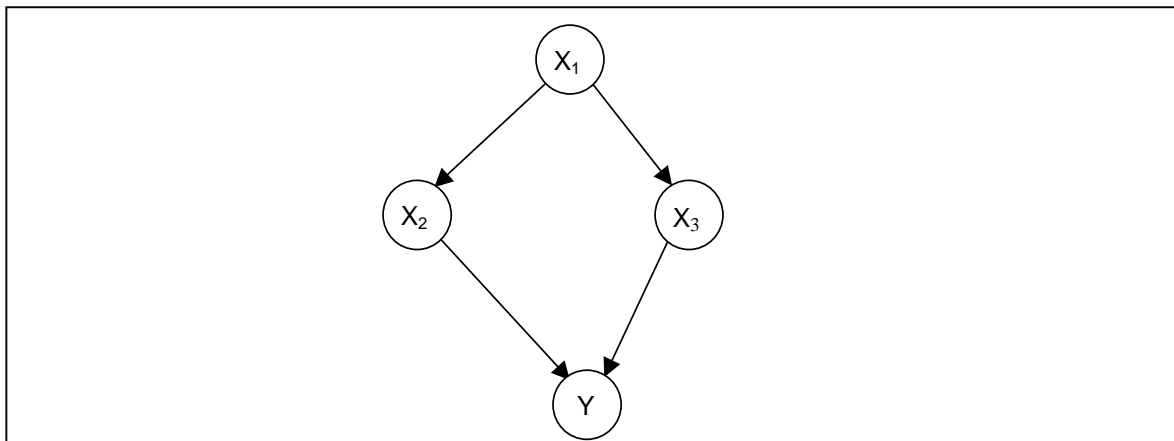
$$X_2 = \alpha_{21} X_1 + U_2$$

$$X_3 = \alpha_{31} X_1 + U_3$$

$$Y = \alpha_{42} X_2 + \alpha_{43} X_3 + U_4$$

Der zugehörige Graph ist in Abbildung 5.1 dargestellt.

Abbildung 5.1



Zur Simulation von 1000 Datensätzen wird die folgende Parameterisierung gewählt. Die Zufallsgrößen U_1, \dots, U_4 seien standardnormalverteilt und die Koeffizienten seien willkürlich auf die Werte $\alpha_{21}=0.8$, $\alpha_{31}=1.5$, $\alpha_{42}=0.5$, $\alpha_{43}=-1.0$ festgelegt. Daraus ergibt sich mit Lemma 3.1 die (symmetrische) Kovarianzmatrix $\mathbf{V}=(X_1, X_2, X_3, Y)'$

$$\text{Cov}(\mathbf{V}) = \begin{pmatrix} 1 & & & \\ 0,8 & 1,64 & & \\ 1,5 & 1,2 & 3,25 & \\ -1,1 & -0,38 & -2,65 & 3,46 \end{pmatrix}.$$

Nach Folgerung 5.1 resultieren bei graphischer oder schrittweiser Regression dieselben Regressoren, und zwar X_2 und X_3 . Diese Aussagen gelten auch auf Populationsebene, d.h. im wahren Modell mit Gleichung (3).

Der Output der graphischen Selektion hat mit einem Signifikanzniveau von 0.01 die Gestalt.

Liste der moeglichen Regressoren

XK

x2
x3

Ein schrittweises Vorgehen mit der Option Selection=Forward liefert den folgenden Output.

No other variable met the 0.0100 significance level for entry into the model.							
Summary of Forward Selection Procedure for Dependent Variable Y							
Step	Entered	Variable In	Number R**2	Partial R**2	Model C(p)	F	Prob>F
1	X3	1	0.6036	0.6036	291.4449	1521.2664	0.0001
2	X2	2	0.0897	0.6933	1.7240	291.8015	0.0001

Beispiel 5.2

Sei M ein lineares Strukturgleichungsmodell mit unabhängigen, normalverteilten Zufallsgrößen U_1, \dots, U_5 .

$$X_1 = \alpha_{12} X_2 + \alpha_{14} X_4 + U_1$$

$$X_2 = U_2$$

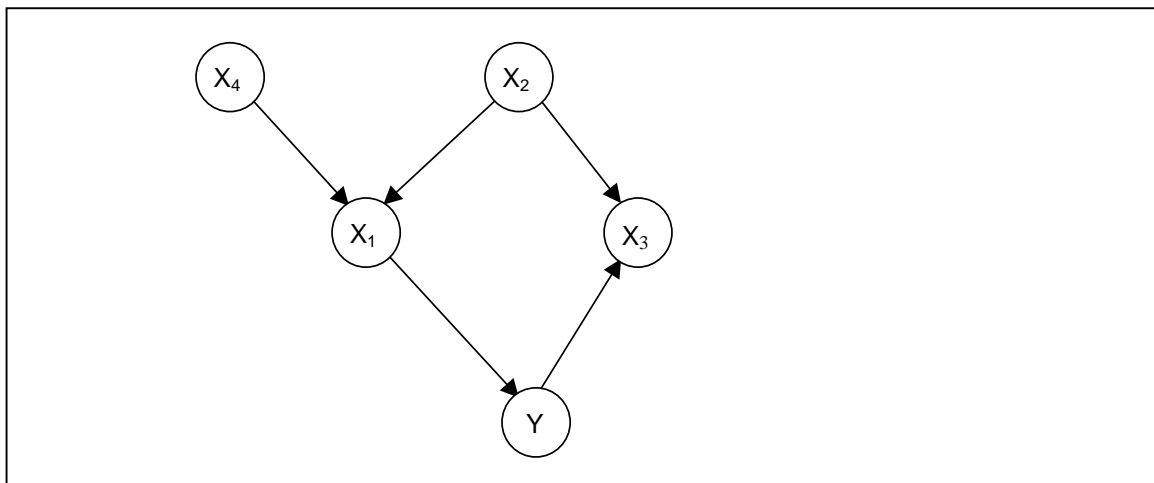
$$X_3 = \alpha_{32} X_2 + \alpha_{35} Y + U_3$$

$$X_4 = U_4$$

$$Y = \alpha_{51} X_1 + U_5$$

Der zugehörige Graph ist in Abbildung 5.2 dargestellt.

Abbildung 5.2



Zur Simulation von 1000 Datensätzen sei die Parameterisierung wie folgt spezifiziert. Die Zufallsgrößen U_1, \dots, U_5 seien standardnormalverteilt und die Koeffizienten seien willkürlich als $\alpha_{12}=0.5$, $\alpha_{14}=1.2$, $\alpha_{32}=0.75$, $\alpha_{35}=0.8$, $\alpha_{51}=1.0$ gewählt. Daraus ergibt sich mit Lemma 3.1 die Kovarianzmatrix von $\mathbf{V}=(Y, X_1, X_2, X_3, X_4)'$ zu

$$\text{Cov}(\mathbf{V}) = \begin{pmatrix} 3,69 & & & & \\ 2,69 & 2,69 & & & \\ 0,5 & 0,5 & 1,0 & & \\ 3,327 & 2,527 & 1,15 & 4,5241 & \\ 1,2 & 1,2 & 0 & 0,96 & 1,0 \end{pmatrix}.$$

Zur Erstellung von manipulativen Prognosen von Y würde bei der Einbeziehung von X_3 als Regressor nach Satz 4.2 bzw. Satz 4.3 ein systematischer Fehler begangen. Genau dies ist der Fall. PROC Reg mit der Option `Select ion=Forward` liefert als Output:

```

No other variable met the 0.0100 significance level for entry into the model.

Summary of Forward Selection Procedure for Dependent Variable Y

```

Step	Entered	Variable In	Number R**2	Partial R**2	Model C(p)	F	Prob>F
1	X1	1	0.5987	0.5987	461.9930	1490.7120	0.0001
2	X3	2	0.1049	0.7036	82.4301	353.1057	0.0001
3	X2	3	0.0214	0.7250	6.6688	77.4762	0.0001

Diese Regressoren resultieren auch auf einer Populationsbetrachtung mit Gleichung (3). Die graphische Selektion liefert mit einem Signifikanzniveau von 0.01 eine andere Liste möglicher Regressoren.

```

Liste der moeglichen Regressoren
XK
X1
x3

```

Eine divergierende Ausgabe der beiden Selektionsmethoden kann aufgrund von Folgerung 5.1 nur erfolgen, wenn die Menge der potenziellen Regressoren eine Variable enthält, die vom Regressanden beeinflusst wird. Dies ist hier der Fall. Somit ist ein systematischer Fehler der schrittweisen Auswahl aufgedeckt.

Mit weiteren Methoden, die aus dem zweiten Teil von Satz 5.1 abgeleitet sind, kann in einem solchen Fall überprüft werden, bei welcher der vom Adjazenzkriterium ermittelten potenziellen Regressoren es sich um einen Nachfolger handelt. Im Beispiel 5.2 könnte ermittelt werden, dass $X_2 \rightarrow X_3 \leftarrow Y$ ein Teilgraph des Variablensystems ist bzw. dass Y kein Nachfolger von X_3 ist, falls von einem offenen Variablensystem ausgegangen würde (Spirtes et al. 1993).

6. Zusammenfassung

Gegenstand dieses Beitrags ist der Vergleich von Verfahren zur Auswahl von Regressoren. Besteht das Interesse, mit einem Regressionsansatz Prognosen für den Regressanden zu erstellen, falls die Regressoren auf bestimmte Werte festgesetzt werden, ist es von entscheidender Bedeutung, eine adäquate Menge von Regressoren auszuwählen. Bei einer automatischen Auswahl mit klassischen Verfahren, die auf dem Regressionskriterium beruhen, können dabei systematische Fehler auftreten. Wird dagegen zur Auswahl das Adjazenzkriterium verwendet, kann bei einer Diskrepanz der Ergebnisse von Adjazenz- und Regressionskriterium ein Rückschluss auf eine Fehlspezifikation getroffen werden. Dazu wird eine Lösung mittels SAS vorgeschlagen. Zu einer Erstellung von Prognosen einer Festsetzung sind jedoch noch weitere Überlegungen anzustellen. Diese beziehen sich insbesondere darauf, ob alle relevanten Variablen in eine Untersuchung miteinbezogen sind.

Literaturverzeichnis

- Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (1996): *Multivariate Analysemethoden: eine anwendungsorientierte Einführung*, 8. Auflage, Springer, Berlin.
- Balke, A., Pearl, J. (1995): Counterfactuals and Policy Analysis in Structural Models, in P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence*, Vol. 11, Morgan Kaufmann, San Fransisco, 11-18.

- Cowell, R.G., Dawid, A.P., Lauritzen, S.L., Spiegelhalter, D.J. (1999): *Probabilistic Networks and Expert Systems*, Springer, New York.
- Cox, D.R., Wermuth, N. (1993): Linear Dependencies Represented by Chain Graphs, *Statistical Science*, Vol. 8, 3, 204-18.
- Glymour, C., Spirtes, P., Scheines, R (1997): In Place of Regression, in: Humphreys, P.(Ed.): *Patrick Suppes: Scientific Philosopher*, Vol. 1, Kluwer Academic Publishers, Dordrecht, 339-66.
- Greene, W.H. (1997): *Econometric Analysis*, Third Edition, Prentice Hall, Upper Saddle River.
- Greenland, S., Pearl, J., Robins, J.M. (1999): Causal Diagrams for Epidemiological Research, *Epidemiology*, 10, 27-48.
- Jordan, M.I. (Hrsg.) (1998): *Learning in Graphical Models*, The MIT Press, Cambridge.
- Kischka, P., Eherler, D. (2000): Unconfounded Effects Derived from Graphical Structures, *Proceedings Eighth International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems*, Volume 3, Consejo Superior De Investigaciones Cientificas, Madrid, 1278-1284.
- Kischka, P., Eherler, D. (2001): Causal Graphs and Unconfoundedness, erscheint in *Allgemeines Statistisches Archiv*, 85/2.
- Pearl, J. (1988): *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo.
- Pearl, J. (1998): Graphs, Causality and Structural Equation Models, Cognitive Systems Laboratory, UCLA, *Technical Report*, R-253, Los Angeles.
- Pearl, J. (2000a): *Causality - Models, Reasoning, and Inference*, Cambridge University Press, Cambridge.
- Pratt, J.W., Schlaiffer, R (1988): On the Interpretation and Observation of Laws, *Journal of Econometrics*, 39, 23--52.
- SAS Institute Inc. (1989): *SAS STAT User's Guide*, Version 6, Fourth Edition, Volume 2, SAS Institute, Cary.
- Simpson, E.H. (1951): The Interpretation of Interaction in Contingency Tables, *Journal of the Royal Statistical Society*, Ser. B 13, 238 - 241.
- Spanos, A. (1999): *Probability Theory and Statistical Inference*, Cambridge University Press, Cambridge.
- Spirtes, P., Scheines, R., Glymour, C. (1993): *Causation, Prediction, and Search*, LNS, Springer, New York.
- Spirtes, P., Richardson, T., Meek, C., Scheines, R., Glymour, C. (1998): Using Path Diagrams as a Structural Equation Modeling Tool, *Sociological Methods & Research*, Vol. 27, No. 2, 182-225.
- Steyer, R., Eid, M. (1993): *Messen und Testen*, Springer, Berlin.
- Strotz, R.H., Wold, H.O.A (1960): Recursive vs. Nonrecursive Systems: An Attempt of Synthesis, *Econometrica*, Vol. 28, 2, 417--27.
- Verma, T., Pearl, J. (1991): Equivalence and Synthesis of Causal Models, in P.P. Bonissone, M. Henrion, L.N. Kanal, J.F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence*, Vol. 6, Elsevier, Amsterdam, 255-268.
- Whittaker, J. (1990): *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons, Chichester.
- Wright, S. (1923): The Theory of Path Coefficients, *Genetics*, 8, 239-255.