Der Versuch einer Veranschaulichung von „Kointegration“

Der Aufsatz von M. Beenstock et al. „Polynomial cointegration tests of anthropogenic impact on global warming“ wird im Folgenden als BEE abgekürzt. Er beschäftigt sich mit der statistischen Analyse von Zeitreihen. Da hierbei Begriffe verwendet werden, die den meisten Leser(inne)n wohl nicht vertraut sind, wird im Folgenden eine Veranschaulichung versucht. Mathematische Strenge ist dabei weder möglich noch beabsichtigt. Ist man hieran doch interessiert, bietet die in BEE angegebene Literatur ausreichend Material. Die nachfolgend besprochenen Begriffe sind

- Zeitreihen,
- stationär,
- instationär,
- erste und zweite Differenz,
- Trend,
- I(0, I(1), I(2),
- Kointegration und schließlich
- polynomiale Kointegration.

1. Stationäre und instationäre Zeitreihen

Statistik beschäftigt sich mit zufällig verteilten Größen. Diese sind in BEE die Elemente von Zeitreihen, wie z.B. Jahresmitteltemperaturen, jährliche Konzentrationen von Treibhausgasen oder Sonnenparameter wie z.B. jährliche Strahlungsstärken. Es sind Messwerte, die, den Gesetzen des Zufalls folgend, schwanken. Eine Zeitreihe wird als *stationär* bezeichnet, wenn ihre maßgebende Kenngrößen, wie z.B. Mittelwert oder Varianz, nicht von der Zeit abhängen. Anderenfalls ist sie *instationär*. Gegenüber dieser formalen Definition ist die folgende, nicht ganz korrekte Definition anschaulicher: Trägt man eine Zeitreihe, z.B. Jahresmitteltemperaturen, graphisch auf, kann man erkennen, ob sie dazu tendiert, eine bestimmte Schwankungsbreite dauerhaft zu verlassen oder nicht. Wenn die Zeitreihe nun grundsätzlich immer wieder in einen bestimmten Schwankungsbereich zurückkehrt, ist sie *stationär*. Wenn sie dagegen langfristig „davonläuft“ und niemals wieder zurückkehrt, ist sie *instationär*.

Falls Temperaturreihen keinen langfristigen Trend enthalten, sind sie nach heutigem Kenntnisstand *stationär*. Als einen externen – manchmal auch als „deterministisch“ bezeichneten - Trend kann man sich das stete Ansteigen von anthropogenen (menschgemachten) Treibhausgasen vorstellen. Er könnte für eine stetig zunehmende Erwärmung und somit ein „Weglaufen“ der Temperaturreihe sorgen. Damit würde die Temperaturreihe *instationär* werden. Ein externer Trend muss aber keineswegs einen anthropogenen Ursprung aufweisen. Es könnte auch eine neu auftretende, ungewöhnliche Sonnenaktivität oder irgend etwas anderes Unbekanntes sein – wie sagt man so schön „Nichts Genaues weiß man nicht“. Infolge der unterschiedlichen „Stärken“ von Instationarität werden in BEE Temperaturreihen mit externem Trend als „schwach instationär“ bezeichnet.

2. Integrierte Zeitreihen, I(0), I(1), I(2), ...

Die Forscher Box und Jenkins stellten 1970 die interessante Frage, ob die Ableitung einer instationären Zeitreihe ebenfalls instationär ist. Die Ableitung einer Zeitreihe ist natürlich wieder eine Zeitreihe und sie ist natürlich ebenfalls wieder zufälligen Schwankungen unterworfen. Ist sie aber auch wieder instationär? Bei Temperaturreihen ist dies offensichlich nicht der Fall, nichtlineare Trends ausgenommen. Bildet man fortlaufend die Temperaturdifferenzen zweier aufeinanderfolgenden Jahre und dividiert man durch den zugehörigen Zeitabschnitt (hier ein Jahr), erhält man die Differenzenquotienten der Zeitreihe, die den Ableitungen entspricht. Es zeigt sich nun, dass im Falle von Temperaturreihen nach „Differenzierung“ stationäre Zeitreihen der Differenzenquotienten entstehen. Dies ist anschaulich auch verständlich. Wie bereits betont, sind Temperaturreihen von Natur aus stationär (genauer, sie sind stationär, wenn ihr Hurst-Exponen kleiner als 1 ist), nur ein externer Trend, oder aber Hurst-Exponenten > 1 können sie instationär werden lassen. Betrachtet man eine stationäre Temperaturreihe ohne externen Trend, so ergibt sich nach Differenzieren natürlich wieder eine stationäre Zeitreihe. Bei einer Temperaturreihe mit externem, linearen Trend kommt durch das Differenzieren dagegen der Trend zum Vorschein - mit den zufälligen Schwankungen um die Trendlinie. Da die differenzierte Temperaturreihe die Nachbarschaft der Trendlinie nicht verlässt, ist sie gemäß Definition zu einer stationären Zeitreihe geworden. Das Differenzieren einer Temperaturreihe erzeugt also stets eine stationäre Zeitreihe, vorausgesetzt, der evtl. vorhandene externe Trend ist linear. Nichtlineare Trends sind Spezialfälle, deren Erläuterung hier zu weit führen würde. In BEE wird auch der nichtlineare Fall berücksichtigt.

Da bei Temperaturzeitreihen einmaliges Differenzieren, oder umgekehrt einmaliges Integrieren der Differenzenreihe, aus der instationären eine stationäre Zeitreihe bzw. umgekehrt macht, wird Temperaturreihen die Eigenschaft I(1) zugeschrieben. Daraus ergibt sich folgende Definition: Eine instationäre Zeitreihe heißt „integriert von der Ordnung p“, kurz I(p), wenn sie nach p-maligem Differenzieren stationär wird. Ein Gegenbeispiel ist weißes Rauschen, eine Zeitreihe von reinen Zufallswerten ohne Trend. Diese Reihe kann man integrieren oder differenzieren so oft man will, es ergibt sich stets wieder weißes Rauschen. Anthropogene Treibhausgasemissionen haben, wie BEE zeigen, dagegen die Eigenschaft I(2), d.h. erst nach zweimaligem Differenzieren wird die anfänglich instationäre Zeitreihe stationär. Zeitreihen, die die nach p-maligem Differenzieren stationär werden, wobei p mindestens 2 ist, werden im Zusammenhang mit der nachfolgend beschriebenen Kointegration mit dem Begriff „polynomial“ gekennzeichnet.

3. Kointegration

Schon relativ früh wurde beobachtet, dass übliche statistische Verfahren Abhängigkeiten zwischen Zeitreihen aufzeigen können, die nicht real sind. Hierfür ist in der englischsprachigen Fachliteratur die Bezeichnung „spurious“ (unecht, scheinbar) gebräuchlich. Die Methode, „spurious“-Phänomene zu vermeiden, bestand nunmehr in einer neuen Methode, der Kointegration, die auf der I(p)-Eigenschaft von Zeitreihen aufbaut. Zwei Zeitreihen mit **gleichem** I(p) werden voneinander abgezogen und die Eigenschaften der Differenzen-Zeitreihe untersucht. Von der unabdingbaren Voraussetzung eines gleichen I(p) beider Ausgangsreihen stammt die Bezeichnung „**Ko“** im Begriff Kointegration“. Polynomiale Kointegration ist darüberhinaus Kointegration für p größer oder gleich 2. Es sei an dieser Stelle bereits auf den entscheidenden Punkt eines gleichen I(p) für beide verglichenen Zeitreihen hingewiesen. Zwei Zeitreihen, von denen die eine beispielsweise die Eigenschaft I(2), die zweite aber I(1) aufweist, können statistisch nichts miteinander zu tun haben. Ausnahmen davon werden in der Literatur und in BEE behandelt, gehen aber über unsere einfache Erläuterung hinaus. In BEE ist bei diesem Sonderfall unter 2.3 von „anthropogenic anomaly“ die Rede.

Die entscheidende Frage lautet nun: Gegeben sind zwei Zeitreihen mit gleichem I(p). Wann haben sie statistisch etwas miteinander zu tun? Man beginnt jetzt schon zu ahnen, wohin die Reise geht. BEE untersucht mit Hilfe der „Kointegration“, ob die Zeitreihen anthropogener Treibhausgase etwas mit Temperaturzeitreihen statistisch zu tun haben. Nur in diesem Fall wäre der Mensch – zumindest statistisch nachweisbar - für die jüngste Erwärmung verantwortlich. Ich möchte hier schon das wichtigste Ergebnis von BEE vorwegnehmen: Da anthropogene Treibhausgas-Zeitreihen vom Typ I(2), Temperaturreihen aber vom I(1) sind, haben sie nichts miteinander zu tun. Die eigentliche Arbeit in BEE bestand darin, die I(p)-Eigenschaften der verglichenen Zeitreihen als tatsächlich vorhanden nachzuweisen sowie alle Sonderfälle zu untersuchen.

Damit es nicht zu trocken wird und um das Konzept der Kointegration zu veranschaulichen, betrachten wir zwei Betrunkene, die nach durchzechtem Abend nach Hause taumeln. Unter welchen Umständen haben ihre Wege (Orts-Zeitreihen) statistisch etwas miteinander zu tun? Gehen wir zunächst davon aus, dass beide zu sehr alkoholisiert sind, um den jeweils anderen überhaupt zu bemerken. Der Weg jedes einzelnen Betrunkenen wird durch das sog. Random-Walk-Modell beschrieben. Es ist übrigens sehr leicht zu programmieren. Für jeden Schritt eines Betrunkenen werden zwei Zufallszahlen benötigt. Die eine gibt die Schrittlänge in x-Richtung der Ebene, die andere in y-Richtung an. Die **individuellen** Schrittlängen sind zwar rein zufällig, aber naturgemäß stationär, denn ein Mensch kann nur Schritte bis zu einer bestimmten Maximallänge machen. Es ergibt sich für jeden der beiden Betrunkenen ein Weg, der sich proportional zur Wurzel(N) (N = Schrittzahl) vom Ausgangspunkt - hier die Bar, wo es den Schnaps gab - entfernt. Damit ist der Random-Walk beider Betrunkener instationär vom Typ I(1). Differenzierung liefert die Zeitreihen der individuellen Schrittweiten, die begrenzt und damit stationär sind.

Die Anwendung der Kointegration kann nun darin bestehen, die Abstandszeitreihe zwischen den beiden Betrunkenen zu analysieren. Man kann sofort sagen, dass sie ebenso wie die Einzelschritte der beiden Betrunkenen zufällig ist. Ist sie aber instationär oder stationär geworden? Und die wichtigste Frage im hier betrachteten Zusammenhang: Kann aus der Abstandszeitreihe einen statistischen Zusammenhang zwischen den eingeschlagenen Wegen beider Betrunkenen (Ursprungszeitreihen) hergeleitet oder erkannt werden? In unserem Beispiel ganz offensichtlich nicht. Es existiert kein statistischer Zusammenhang zwischen den Torkelwegen der beiden Zecher. In komplexeren Anwendungen können sich aber „spurious“-Zusammenhänge zeigen. In unserem Beispiel würde die Kointegration zeigen, dass die Abstandszeitreihe wieder rein zufällig und instationär ist. Genauso, wie sich beide Betrunkenen stetig von ihrem Ausgangspunkt entfernen, entfernen sie sich auch stetig voneinander. Alle in Rede stehenden Zeitreihen sind daher instationär und voneinander völlig unabhängig.

Wir nehmen nun ein verändertes, interessanteres Szenario an: Die beiden Betrunkenen mögen zwei Wohnungen im gleichen Haus haben, aber nur einer von beiden hat noch den separaten Haustürschlüssel in der Tasche. Der andere hat ihn im Suff verloren. Der Betrunkene ohne Schlüssel ruft dem anderen Betrunkenen immer wieder zu, er möge sich nicht zu weit entfernen. Er möchte schließlich ohne Haustürschlüssel noch ins Haus kommen, um nicht im Freien zu übernachten. Ist die Abstandszeitreihe immer noch instationär? Es kommt darauf an. Wenn es dem Betrunkenen ohne Haustürschlüssel gelingt, seinen Saufkumpanen durch Zurufe innerhalb eines maximalen Abstands zu halten, ist die Wegdifferenz zwischen beiden zwar immer noch zufällig sie ist aber nun stationär – die Wegzeitreihen beider Betrunkenen sind kointegriert. Entfernen sie sich dagegen immer weiter voneinander, liegt wie im oben beschrieben einfachen Fall weiterhin Instationarität vor. Mathematisch ist die Bedingung der Stationarität (Kointegration) im vorliegenden Fall zwar nicht mehr ganz anschaulich, aber dafür noch ein wenig „großzügiger". Es muss nicht maximale Abstandswahrung vorliegen, es reicht, wenn ein beliebiger linearer Zusammenhang zwischen den beiden Zeitreihen-Wegen der beiden Betrunkenen existiert.

4. Nullhypothese und Regressionsanalyse

Man kann nun den Kern von BEE verstehen. Es geht hier darum, die reale statistische Abhängigkeit zwischen zwei Variablen - hier anthropogene Treibhausgase und Temperaturen - herauszufinden. Hierzu bedient man sich nicht nur der Methode der Kointegration, man verwendet zudem noch die sog. Nullhypothese. Man setzt einen statistischen Zusammenhang voraus und versucht, unter Einhaltung einer bestimmten statistischen Schwelle nachzuweisen, dass der angenommene Zusammenhang nicht existiert. Ist dies der Fall, gilt die Nullhypothese als zurückgewiesen. In BEE wird als Nullhypothese angenommen, dass kein Zusammenhang zwischen anthropogenen Treibhausgasen und Temperaturverläufen besteht. Nebenbei: Zurückweisung einer Nullhypothese bedeutet nicht unbedingt, dass das Gegenteil zutrifft. Man ist bei Zurückweisung lediglich nicht mehr gezwungen, die Nullhypothese zu akzeptieren. Die Kointegration ist die Methode der Wahl, um der „Spurious-Falle" zu entgehen. Sie vermeidet die üblichen Standardverteilungen und verwendet dagegen die richtigen Verteilungen.

An dieser Stelle soll die anschauliche Erläuterung von BEE beendet werden, denn die in BEE angewandten Methoden und Vorgehensweisen sind tiefergehender, ausführlicher und komplizierter als hier beschrieben. Hier kann nur auf den Originalaufsatz verwiesen werden. Schlussendlich ist daran zu erinnern, dass sich BEE nicht mit der Physik der untersuchten Phänomene befasst. Es ist eine rein statistische Betrachtung mit all ihren Stärken und Schwächen. Mit Statistik kann man keine physikalischen Modelle erstellen, man kann allenfalls Modelle bestätigen oder widerlegen. Wenn man so will, ist Statistik „modellvernichtend".

Prof. Dr. Horst-Joachim Lüdecke
Jan. 2013