Divide et calcula - Ein SAS-Makropaket zur systematischen explorativen Berechnung von Subgruppen

Jörg Sahlmann
iOMEDICO
Ellen-Gottlieb-Straße 19
79106 Freiburg
joerg.sahlmann@iomedico.com
post@joerg-sahlmann.de

Zusammenfassung

Im Rahmen der Auswertung von Beobachtungsstudien kommt es regelmäßig zu Rückfragen, wie sich denn ein bestimmter Parameter in dieser oder jener Subgruppe verhält. Nicht immer sind die Subgruppen im Beobachtungsplan oder im SAP so vorgegeben gewesen. Häufig werden sie auch erst beim Review der Tabellen relevant.
Das Makro SUBGEX generiert für eine Menge von Faktoren das Kreuzprodukt dieser Faktoren und berechnet für diese Subgruppen deskriptive Statistiken.
Die Subgruppen werden als Scatterplot (SVG, skalierbar) dargestellt mit der Gruppengröße auf der x-Achse und einem Lagemaß auf der y-Achse. Die x-Achse schneidet den Parameterwert für die Gesamtgruppe.
Die deskriptiven Statistiken werden als filterbare Exceltabelle ausgegeben, bei der jede Subgruppe durch eine Tabellenzeile repräsentiert wird.
Der Output des Makros ermöglicht einen schnellen Überblick über die Subgruppensituation und dient der weiteren Hypothesengenerierung für künftige Studien.
Die Idee zu diesem Makro kommt aus einem Workshop-Vortrag bei der GMDS 2019, bei der eine R Shiny-App mit einer entsprechenden Funktionalität von Susanne Lippert, Bodo Kirsch und Christoph Muysers vorgestellt wurde[1]. Es wurde um die zeitliche Komponente erweitert.

Schlüsselwörter: Makro, Subgruppen

1 Ausgangslage

Wir haben eine Menge von Merkmalsträgern (z. B. Patienten), bei denen eine abhängige Zielvariable (z. B. ein Fragebogenscore, ein zeitabhängiges Event oder ein sonstiges Effektmaß) gemessen wurde.
Für diese Zielvariable wird ein Mittelwert oder ein anderes angemessenes Lagemaß berechnet.
Wenn weitere unabhängige Merkmale für den Merkmalsträger erfasst worden sind, stellt sich in der Regel die Frage, welche Merkmale in welcher Kombination einen Einfluss auf die Zielvariable haben. Viele Beobachtungspläne halten sich hier vornehm zurück und verweisen auf den SAP.

Der Statistiker kann und will nicht alle möglichen Kombinationen von Merkmalen und ihren Ausprägungen durchprobieren.

2 Lösungsansatz

Wir nehmen n unabhängige kategoriellen Merkmale und bilden alle möglichen Kombinationen für Teilmengen von 1 bis n Elementen.

Für vier Merkmale a, b, c und d ergeben sich die folgenden Kombinationen.

Kein Merkmal von vier Merkmalen: 1 (aufgelistet nur der Vollständigkeit halber)
Ein Merkmal von vier Merkmalen: 4 (a, b, c, d)
Zwei Merkmale von vier Merkmalen: 6 (ab, ac, ad, bc, bd, cd)
Drei Merkmale von vier Merkmalen: 4 (abc, abd, acd bcd)
Vier Merkmale von vier Merkmalen: 1 (abcd)

Die Anzahl für k Merkmale aus n Merkmalen ergibt sich aus dem Binomialkoeffizienten (n über k).
Die Gesamtanzahl ergibt sich aus der Zweierpotenz 2 hoch n.

Die Zahl der Subgruppen für jede Kombination von Merkmalen ergibt sich aus dem Produkt der Zahl der Merkmalsausprägungen.

Für jede Subgruppe wird das entsprechende Lagemaß berechnet. Das Ergebnis wird gegen die Gruppengröße in einem Scatterplot aufgetragen.
Die Gesamtgruppe hat dabei den höchsten Wert auf der x-Achse. Die Subgruppen verteilen sich dann links davon.

3 Beispiel

In einer klinischen Studie wird die Lebensqualität per Fragebogen erhoben und in Form eines Summenscores abgebildet.
Weiterhin werden das Alter, das Geschlecht, das Krankheitsstadium und ein Aktivitätsindex erhoben.
Für das Makropaket wird das Alter kategorisiert.

3.1 Übersichtsdarstellung

Für einen Testdatensatz ergibt sich im ersten Makrolauf (Makro %sgstart) die Abbildung 1. Eine Referenzlinie zeigt den Mittelwert des Summenscores für die Gesamtgruppe an. Die einzelnen Marker stehen dann jeweils für eine Subgruppe. Es werden zunächst alle Subgruppen dargestellt, um einen ersten Eindruck von den Subgruppengrößen und der Streuung zu bekommen.

Der Scatterplot wird auch als SVG-Grafik bereitgestellt. Per Mouseover und Tooltip wird hier die Modellnummer und die Subgruppengröße dargestellt.

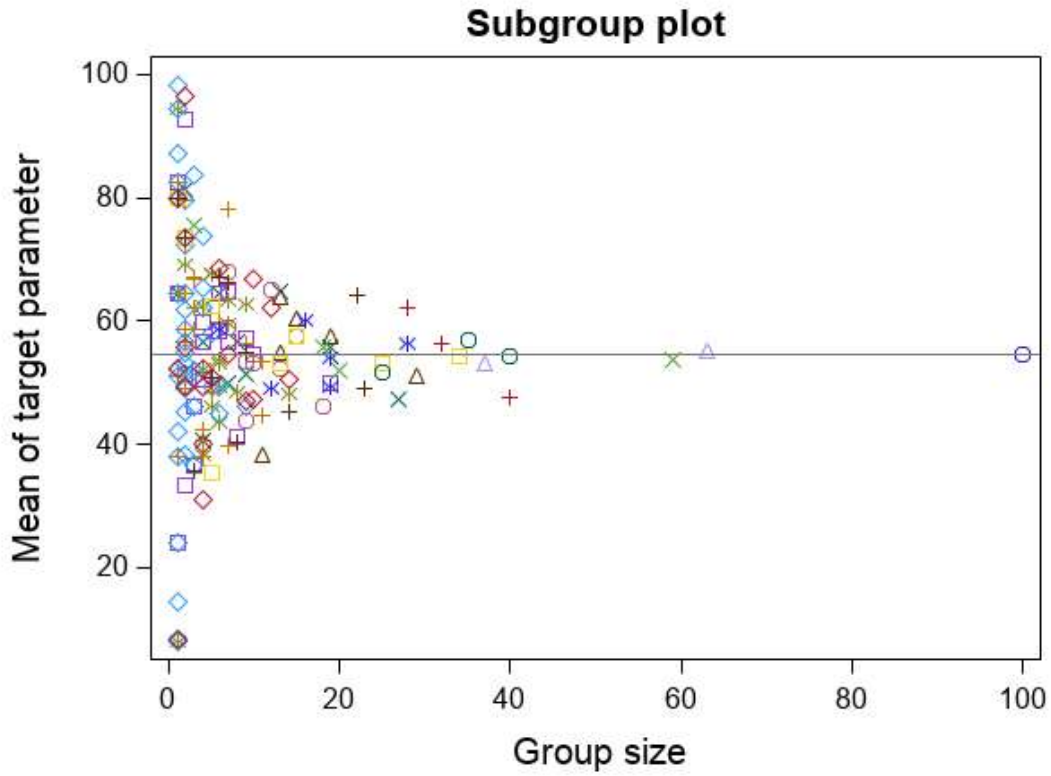

Abbildung 1: Überblick über alle Modelle und Subgruppen (Tootip nicht dargestellt)

3.2 Modellliste

Mit dem Makro %sgmodellist wird eine nummerierte Liste erzeugt, damit man per Index auf ein bestimmtes Modell zugreifen kann. Das brauchen wir, um nach dem Modell zu filtern.

Tabelle 1: Modellliste

Number	Model
0	NA
1	agecat
2	agecat ecog
3	agecat gender
4	agecat gender ecog
5	agecat gender stage
6	agecat gender stage ecog
7	agecat stage
8	agecat stage ecog
9	ecog
10	gender

11	gender ecog
12	gender stage
13	gender stage ecog
14	stage
15	stage ecog

Wie bereits oben angesprochen, haben wir bei vier Merkmalen 16 mögliche Modelle. Diese werden in Tabelle 1 aufgelistet.

3.3 Zoomen und Labeln

Mit dem Makro %sglabel kann in den Scatterplot hineingezoomt werden und die Marker werden mit der Modellnummer gelabelt.

Abbildung 2 zeigt einen Ausschnitt aus der Abbildung 1. Per Mouseover und Tooltip wird hier die Modellnummer und die Subgruppengröße dargestellt.

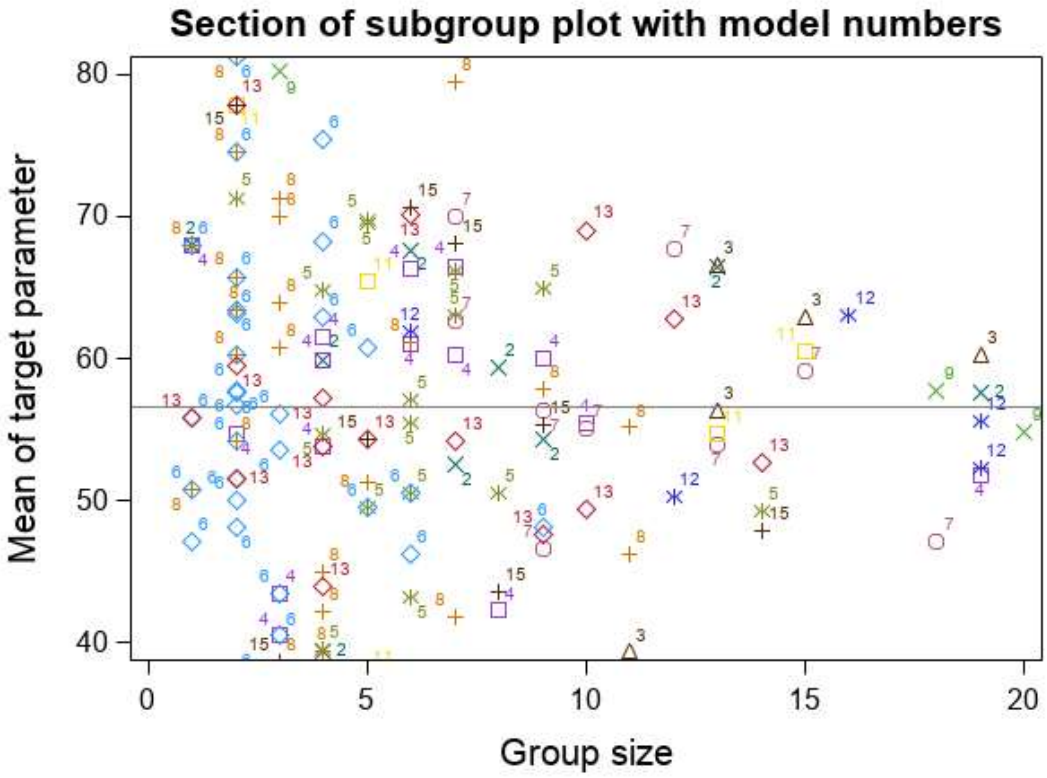

Abbildung 2: Ausschnitt mit Labeln (Tooltip nicht dargestellt)

3.4 Filtern

Mit dem Makro %sgfilter kann nach einem bestimmten Modell gefiltert werden. Hierzu wird die Modellnummer aus der Modellliste aus Parameter übergeben. Der Tooltip zeigt hier die Modellnummer und die Ausprägungen der Merkmale an.

Die Abbildung 3 zeigt eine nach Modellnummer 7 gefilterte Darstellung.

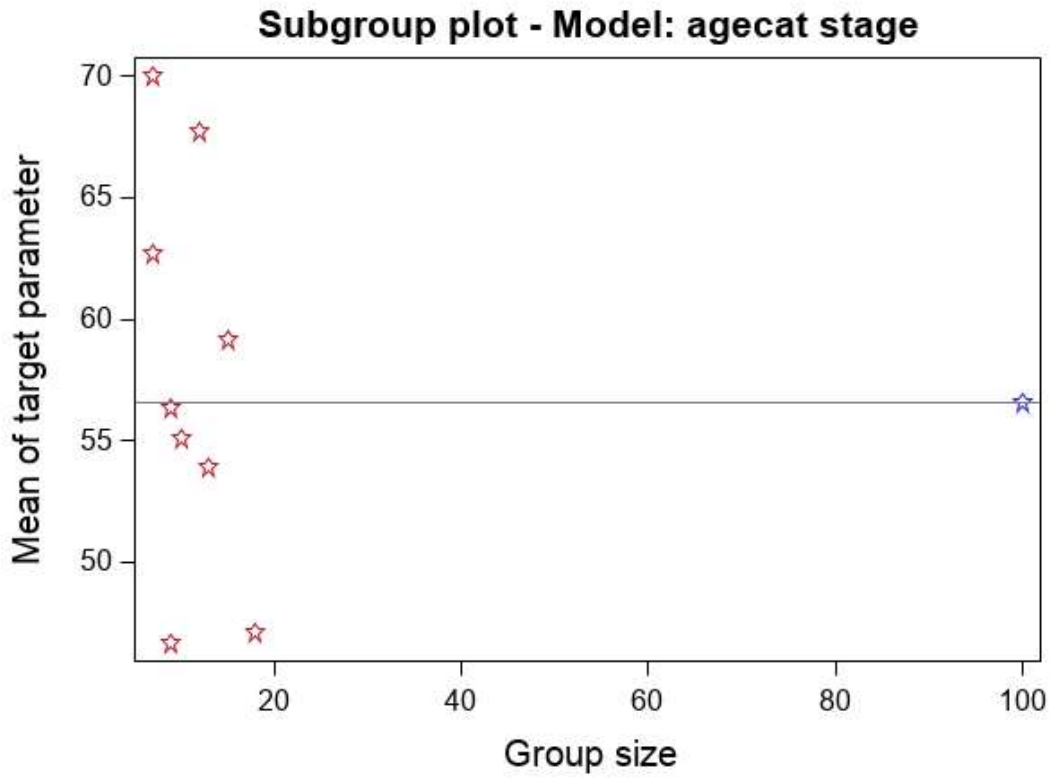

Abbildung 3: Filterung nach Modellnummer 7: agecat stage (Tooltip nicht dargestellt)

3.5 Zeitverläufe

Mit dem Makro %sgtimeshift kann nach einem bestimmten Modell gefiltert werden und die Bewegung zwischen zwei Zeitpunkten dargestellt werden. Hierzu wird die Modellnummer aus der Modellliste aus Parameter übergeben. Der Tooltip zeigt hier die Modellnummer und die Ausprägungen der Merkmale an.

Die Abbildung 4 zeigt eine nach Modellnummer 7 gefilterte Darstellung mit Vektoren, die die Richtung der Entwicklung darstellen.

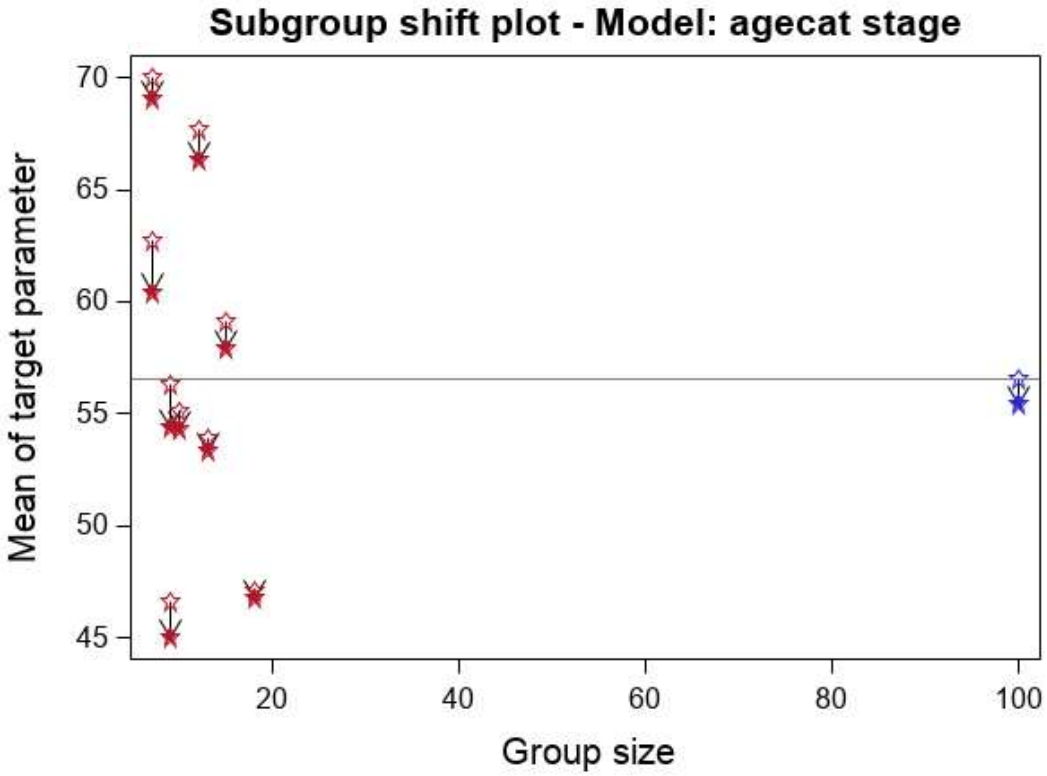

Abbildung 4: Filterung und zeitliche Entwicklung (Tooltip nicht dargestellt)

Wie in Abbildung 5 gezeigt, lassen sich auch mehrere Zeitpunkte darstellen.

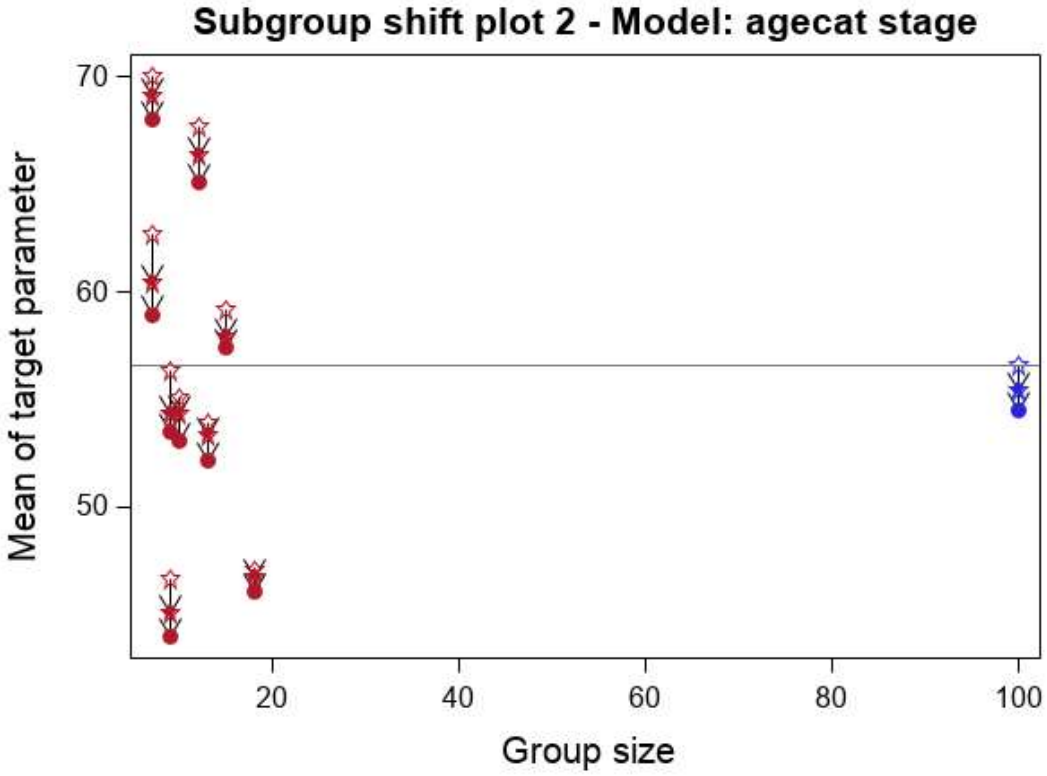

Abbildung 5: Zeitliche Entwicklung über mehrere Zeitpunkte (Tooltip nicht dargestellt)

4 Diskussion

Eine Darstellung von Subgruppen in dieser Form ist zweischneidig.
Auf der einen Seite liefert sie eine gute Übersicht über die Verteilung des Lagemaßes der Zielvariablen unter verschiedenen Bedingungen.
Auf der anderen Seite ist bei unkritischer Betrachtung der Ergebnisse und des Nichtwahrhabens der Problematik des multiplen Testens eine Fehlinterpretation der Ergebnisse möglich.
Wir betrachten diese Form der Darstellung als gute Möglichkeit zur Hypothesengenerierung für kommende Studien.

Literatur

[1] B. Kirsch, S. Lippert, C. Muysers: Subgroup Explorer zur systematischen Analyse von Subgruppen – Einsetzbar bei der Nutzenbewertung? Meeting abstract: Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie. 64. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS). Dortmund, 08.-11.09.2019. Düsseldorf: German Medical Science GMS Publishing House; 2019. DocAbstr. 144.

Anhang A Quellcode

Den Zugriff auf den Quellcode für die SAS-Makros schicken wir auf Anforderung gerne zu.