Analyse von Cluster- und Paneldaten

Statistische Analyseverfahren, die mit den Begriffen Cluster- oder Paneldaten in Verbindung stehen, werden in verschiedenen Forschungsdisziplinen mit unterschiedlichen Synonymen bezeichnet. Gängige Bezeichnungen in der Psychologie, Medizin und empirischen Sozialforschung sind zum Beispiel Mehrebenenanalyse (multilevel modeling), Hierarchische lineare Modellierung (hierarchical linear models) und Analyse gemischter Modelle (mixed model analysis). In der Ökonometrie und den Sozialwissenschaften, vor allem in der Längsschnittforschung, wird vorrangig der Begriff der Paneldatenanalyse verwendet. In den meisten Fällen bezieht sich die Bezeichnung auf die Struktur der vorliegenden Daten.

fu:stat bietet regelmäßig Schulungen für Hochschulangehörige sowie für Unternehmen und weitere Institutionen an. Die Inhalte reichen von Statistikgrundlagen (Deskriptive, Testen, Schätzen, lineare Regression) bis zu Methoden für Big Data. Es werden außerdem Kurse zu verschiedenen Software-Paketen gegeben. Auf Anfrage können wir auch gerne individuelle Inhouse-Schulungen bei Ihnen anbieten.

Inhaltsverzeichnis

- Datenstruktur
  - Beispiel 1
  - Beispiel 2
- Auswertungsmöglichkeiten
  - Aggregieren
  - Auswertung mit gepoolten Daten
  - Mehrebenenanalyse
- Modellwahl
- Umsetzung in Statistik Software

Datenstruktur

In einem Datensatz können Beobachtungseinheiten in mehreren Ebenen enthalten sein. Ein Beispiel mit zwei Ebenen aus der Bildungsforschung ist die Untersuchung des Lernerfolgs von Schülern (Ebene 1) aus verschiedenen Klassen (Ebene 2). Hierbei bilden die Beobachtungen der Schüler die erste Ebene und die verschiedenen Klassen die zweite Ebene. Es können Charakteristika (Variablen) auf Schülerebene (z. B. Testergebnisse aus Leistungskontrollen, Geschlecht) und auf Klassenebene (z.B. Klassengröße, Eigenschaften des Klassenlehrers) im Datensatz enthalten sein. Die Abbildung Beispiel 1 zeigt eine schematische Darstellung dieser Datenstruktur. Der Abbildung steht eine Datentabelle gegenüber, die verdeutlicht, wie die Daten organisiert sein müssen, um eine Mehrebenenanalyse durchführen zu können. Die Farben in den Tabellenüberschriften korrespondieren zu den Farben in der Abbildung und verdeutlichen, auf welcher Ebene die Variablen gemessen werden. Die dargestellte Datenstruktur wird auch als *Long Format* bezeichnet. In einigen Panel-Datensätzen, wie zum Beispiel im Sozioökonomischen Panel (SOEP), ist jedoch das *Flat Format* üblich. Hier wird jeder Person nur eine Datenzeile zugeordnet. Für Beobachtungen unterschiedlicher Zeitpunkte werden neue Variablen generiert. Um eine Mehrebenenanalyse durchführen zu können, müssen die Daten in das *Long-Format* transformiert werden. Dies sollte in keinem Fall per Hand geschehen, da eine händische Transformation sehr aufwändig und fehleranfällig ist. In gängigen Statistikprogrammen (STATA, R, SPSS, SAS, JMP) gibt es Funktionen, die die Transformation vom *Flat-Format* zum *Long-Format* (und umgekehrt) durchführen.

Literaturhinweise

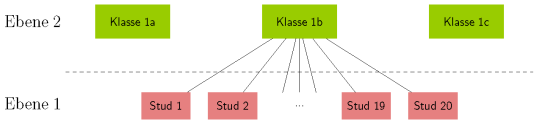
Mehrebenenanalyse

- Gelman, Andrew / Jennifer Hill (2007): Data Analysis Using Regression and Multilevel /Hierarchical Models, New York

Panelanalyse

- Giesselmann, Marco / Michael Windzio (2012): Regressionsmodelle zur Analyse von Paneldaten, Wiesbaden

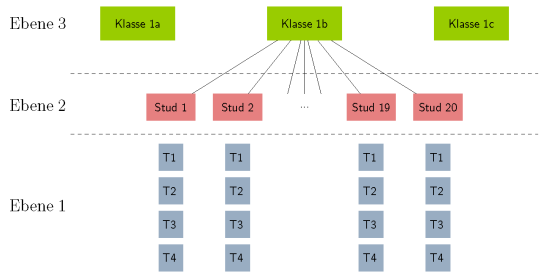
Beispiel 1



Klasse	Student	Klassengröße	IQ	Leistungstest
1a	1	25	110	95
1a	2	25	112	80
…				
1b	1	28	115	70
1b	2	28	100	82
…				
1c	1	21	117	91
1c	2	21	110	78

Diese Datenstruktur kann auf Beispiele aus anderen Forschungsdisziplinen übertragen werden: In der Paneldatenanalyse entsprechen die Individuen der Ebene 2 und die Zeitpunkte der Ebene 1; in klinischen Studien mit Patienten aus unterschiedlichen Kliniken entsprechen die Kliniken der zweiten Ebene und die Patienten der ersten Ebene. Das obige Beispiel aus der Bildungsforschung lässt sich auf mehr als zwei Ebenen erweitern. Werden die Schüler zum Beispiel über mehrere Zeitpunkte beobachtet, kommt eine weitere Ebene hinzu.

Beispiel 2



Klasse	Student	Zeitpunkt	Klassengröße	IQ	Leistungstest
1a	1	T1	25	110	95
1a	1	T2	25	110	93
1a	2	T1	25	112	80
1a	2	T2	25	112	85
…					
1b	1	T1	28	115	70
1b	1	T2	28	115	72
1b	2	T1	28	100	82
1b	2	T2	28	100	90
…					
1c	1	T1	21	117	91
1c	1	T2	21	117	90
1c	2	T1	21	110	78
1c	2	T2	21	110	80

Analog zu dieser Erweiterung könnten in der Paneldatenanalyse die Individuen aus unterschiedlichen Regionen stammen. Die Regionen würden in diesem Fall Ebene 3, die Individuen Ebene 2 und die Zeitpunkte Ebene 1 entsprechen. Im Beispiel der Klinischen Studie könnten die Patienten über mehrere Zeitpunkte beobachtet werden, wobei die Kliniken dann Ebene 3, die Patienten Ebene 2 und die Zeitpunkte Ebene 1 entsprächen.

Die Datenstruktur bestimmt die Abhängigkeitsstruktur oder auch Clusterung in den Daten. Als Cluster werden allgemein Beobachtungen bezeichnet, die sich aufgrund von Gemeinsamkeiten ähneln. Im Beispiel wird die Abhängigkeit durch die Klassenzugehörigkeit bestimmt. Es ist zu erwarten, dass die Ergebnisse der Schüler innerhalb einer Klasse ähnlicher sind als die Ergebnisse im Vergleich zwischen den Klassen. Dies ist bedingt durch messbare klassenspezifische Faktoren, wie Klassengröße oder Geschlecht des Lehrers, aber auch durch nicht messbare Faktoren, wie Zusammengehörigkeitsgefühl oder Beziehung zum Lehrer. Werden die Schüler zusätzlich über verschiedene Zeitpunkte beobachtet, kommt eine zeitliche Abhängigkeit hinzu. Es ist zu erwarten, dass sich die Ergebnisse eines Schülers über die Zeit ähnlicher sind als die Ergebnisse unterschiedlicher Schüler. Dies ist bedingt durch beobachtbare schülerspezifische Eigenschaften, wie Geschlecht oder sozioökonomische Kennzahlen des Elternhauses, aber auch nicht beobachtbaren Eigenschaften, wie Motivation für das Fach oder Begabung. Diese Abhängigkeit, die durch Zugehörigkeit in den verschieden Ebenen bedingt ist, wird allgemein als Intraklassenkorrelation (geläufige Abkürzung ICC) bezeichnet.

Auswertungsmöglichkeiten

Hierarchisch organisierte Daten können auf verschiedene Weise analysiert werden. Analysemethoden unterscheiden sich hinsichtlich Ihrer Komplexität, wobei die Wahl der Methode an die Fragestellung gekoppelt sein sollte.

Für diesen Abschnitt soll folgendes gelten: Die abhängige Variable liegt auf Individualebene (Ebene 1) vor und wird mit $\{y\}$ gekennzeichnet; erklärende Variablen können auf Individualebene, aber auch auf höheren Ebenen vorliegen, wobei erklärende Variablen der Ebene 1 mit $\{x\}$ und erklärende Variablen der Ebene 2 mit $\{z\}$ gekennzeichnet werden. Erklärende Variablen aus höheren Ebenen geben Aufschluss über den übergeordneten Kontext, aus dem die Individualbeobachtungen stammen, und werden deshalb häufig als Kontextvariablen bezeichnet.

Aggregieren

In einem Datensatz mit zwei Ebenen werden die in Ebene 1 gemessen Variablen auf Ebene 2 zusammengefasst. Für die Daten aus Beispiel 1 hieße das, dass Mittelwerte über die Schüler innerhalb der einzelnen Klassen gebildet werden. Ist die Variation der Schüler innerhalb der Klassen gering und sind laut Fragestellung ausschließlich die Unterschiede zwischen den Klassen von Interesse, kann das Aggregieren der Daten eine geeignete Methode sein. Ein Regressionsmodell für die Ergebnisse im Leistungstest auf Ebene 2 mit IQ und Klassengröße als erklärenden Variablen lautet

> Dieses Vorgehen entspricht in der Paneldatenanalyse der "between regression", in der die zeitlichen Mittelwerte der Individuen betrachtet werden.

$$\bar{y}_{i} = \alpha_0+ \alpha_1 \bar{x}_{i} +\alpha_2 z_{i} + e_{i}$$

$$e_{ij} \sim N(0,\sigma_e^2)$$

Modellparameter

- \(\alpha_0\): Regressionskonstante
- \(\alpha_1\): Effekt von \(x\) auf \(y\)
- \(\alpha_2\): Effekt der Kontextvariable \(z\) auf \(y\)
- \(\sigma_e^2\): Varianz des Fehlerterms

\(\bar{y}_{i}\) entspricht dem mittleren Testergebnis, \(\bar{x}_{i}\) dem mittleren IQ in Klasse \(i\) und \(z_i\) der Klassengröße von Klasse \(i\); \(e_{i}\) kennzeichnet den Fehlerterm des aggregierten Regressionsmodells. Dieses Vorgehen hat zur Folge, dass die Anzahl der verwendeten Beobachtungen auf die Anzahl der in Ebene 2 beobachteten Einheiten reduziert wird. Im Beispiel entspricht das der Anzahl der beobachteten Schulklassen.

> **Hinweis zur Indizierung**
>
> Es ist üblich für jede Datenebene einen Index eizuführen. Bei Daten mit Messwiederholungen wird für die Zeitebene häufig der Index \(t\) gewählt.

Auswertung mit gepoolten Daten

Die hierarchische Datenstruktur wird ignoriert und die Daten werden behandelt als wären sie unabhängig. Ein gepooltes Regressionsmodell für die Ergebnisse im Leistungstest auf Ebene 1 mit IQ und Klassengröße als erklärenden Variablen lautet

$$y_{ij} = \alpha_0+ \alpha_1 x_{ij} +\alpha_2 z_{i} + e_{ij}.$$

$$e_{ij} \sim N(0,\sigma_e^2)$$

Modellparameter

- \(\alpha_0\): Regressionskonstante
- \(\alpha_1\): Effekt von \(x\) auf \(y\)
- \(\alpha_2\): Effekt der Kontextvariable \(z\) auf \(y\)
- \(\sigma_e^2\): Varianz des Fehlerterms

Der Index \(ij\) kennzeichnet Individuum \(j\) in Klasse \(i\). Werden in ein Regressionsmodell Kontextvariablen \(z\) aufgenommen (die Aufnahme von Kontextvariablen ist optional), kann ein Teil der Abhängigkeit zwischen den Beobachtungen aufgefangen werden, da \(z\) eine Ursache für die Abhängigkeit sein kann. Es ist denkbar, dass größere Schulklassen tendenziell etwas schlechtere Ergebnisse aufweisen als kleinere Schulklassen. In diesem Fall wäre die Klassengröße eine Ursache für die Abhängigkeit in den Daten. Ein Teil der Abhängigkeit bleibt erhalten, wenn im Fehlerterm \(e\) unbeobachtete klassenspezifische Eigenschaften enthalten sind, die Einfluss auf die Zielgröße \({y}\) haben. Diese unbeobachteten Einflussgrößen (auch als unbeobachtete Heterogenität bezeichnet) sorgen dafür, dass die Beobachtungen weiterhin abhängig sind. In einer OLS Regression, die Unabhängigkeit der Beobachtungen voraussetzt, könnte dieses Vorgehen eine fehlerhafte Inferenz zur Folge haben. Die Standardfehler würden tendenziell unterschätzt werden, was die Wahrscheinlichkeit erhöht, Effekte fälschlicherweise als signifikant anzunehmen.

Mehrebenenanalyse

Modelle der Mehrebenenanalyse sind darauf ausgerichtet, die Abhängigkeitsstruktur durch geeignete Modellanpassungen zu berücksichtigen. Die unbeobachtete Heterogenität, die als Ursache für die Abhängigkeit der Beobachtungen angenommen wird, kann zum einen direkt als fester Effekt geschätzt werden. Zum Anderen kann die unbeobachtete Heterogenität als Teil des Fehlerterms aufgefasst werden (zufälliger Effekt). Hierbei wird die Varianz des Fehlerterms in verschiedene Komponenten zerlegt, die Intraklassenkorrelation wird direkt geschätzt und bei der Schätzung des Intercepts und der Steigungsparameter des Regressionsmodells berücksichtigt.

Modell mit festen Effekten (fixed effects model)

> In der Paneldatenanalyse entspricht dieses Vorgehen der fixed effects Regression mit einer within tranformation.

Die unbeobachtete Heterogenität zwischen den Elementen der Ebene 2 wird direkt mit Hilfe von Dummy-Variablen berücksichtigt (d.h. es wird eine Indikatorvariable für jede Klasse aufgenommen). Das bedeutet, dass spezifische Regressionskonstanten für jedes Element der Ebene 2 geschätzt werden. Abhängigkeiten, die auf unbeobachtete Heterogenität zwischen den Klassen zurückzuführen ist, können so berücksichtigt werden. Das Regressionsmodell für die Ergebnisse im Leistungstest mit IQ und Klassengröße als erklärenden Variablen und festen klassenspezifischen Konstanten lautet

$$y_{ij} = \alpha_{0}+ \mu_i+ \alpha_1 x_{ij} + e_{ij}.$$

Modellparameter

$$ e_{ij} \sim N(0,\sigma_e^2)$$

- \(\alpha_{0}\): Regressionskonstante
- \(\mu_{i}\): klassenspezifische Verschiebung
- \(\alpha_1\): Effekt von \(x\) auf \(y\)
- \(\sigma_e^2\): Varianz des Fehlerterms

Die klassenspezifische Regressionskonstante setzt sich folglich aus der Regressionskonstanten und der klassenspezifische Verschiebung zusammen: \(\alpha_i = \alpha_0 + \mu_i\). Die Regressionskonstante der Referenzklasse ist \(\alpha_0\). Die zusätzlich in das Modell aufgenommen Dummy-Variablen \(\mu_i\) absorbieren die gesamte Heterogenität (beobachtet und unbeobachtet) zwischen den Klassen. Die Fehlerterme können nun als unabhängig angenommen werden. Eine OLS Regression, die Unabhängigkeit der Beobachtungen voraussetzt, ist wieder zulässig. Dieser Modelltyp ist attraktiv aufgrund seiner Einfachheit, bringt jedoch auch Nachteile mit sich: Effizienzverlust, da für jedes Element der Ebene 2 eine spezifische Regressionskonstante geschätzt wird, was die Anzahl der Freiheitsgrade verringert (In einer Studie, in der \(k\) Klassen untersucht werden, müssten \(k-1\) zusätzliche Parameter geschätzt werden); Kontextvariablen (\(z\)) können nicht aufgenommen werden, da die Heterogenität zwischen den Klassen vollständig durch die Dummy-Variablen aufgefangen wird.

Modelle mit festen Effekten bieten sich insbesondere dann an, wenn die Zahl der beobachteten Einheiten nicht sehr groß ist und Vorhersagen nur für die betrachteten Einheiten in der Stichprobe getroffen werden sollen. Werden in einer Panelanalyse zum Beispiel die Mitgliedsstaaten der EU untersucht, mit dem Ziel, Vorhersagen für Entwicklungen in der EU zu treffen, ist ein Modell mit festen Effekten geeignet. Ist die Zahl der Beobachtungseinheiten sehr groß und es sollen Rückschlüsse von der Stichprobe auf die Grundgesamtheit gezogen werden (z. B. bei Haushaltspanels), sind Modelle mit festen Effekten ungeeignet. In diesem Fall bieten sich Modelle mit zufälligen Effekten an.

Modell mit zufälliger Regressionskonstante (random intercept model)

Die unbeobachtete Heterogenität zwischen den Klassen wird als zufällig ausgefasst. Wie bei den festen Effekten werden die Niveauunterschiede zwischen den einzelnen Klassen mit Hilfe von klassenspezifischen Intercepts dargestellt. Im Gegensatz zu den festen Effekten werden die Intercepts nicht direkt geschätzt, sie werden hingegen als Teil des zufälligen Fehlerterms angesehen. Der Fehlerterm setzt sich nun aus einer klassenspezifischen und einer individuenspezifischen zufälligen Komponente zusammen. Modelle mit zufälligen Effekten werden auch gemischte Modelle genannt, weil es neben dem zufälligen zusammengesetzten Fehlerterm auch einen systematischen Teil mit festen Effekten gibt.

Das Regressionsmodell mit zusammengesetztem Fehlerterm lautet

$$y_{ij} = \underbrace{\alpha_{0} + \alpha_1 x_{ij} +\alpha_2 z_{i}}_{feste\, \, Effekte} + \underbrace{u_{i}+ e_{ij}}_{Fehlerterm}$$

$$ e_{ij} \sim N(0,\sigma_e^2), \,\, u_{i} \sim N(0,\sigma_u^2)$$

Modellparameter

- \(\alpha_{0}\): Regressionskonstante
- \(\alpha_1\): Effekt von \(x\) auf \(y\)
- \(\alpha_2\): Effekt der Kontextvariable \(z\) auf \(y\)
- \(\sigma_e^2\): Varianz des individuellen Fehlerterms
- \(\sigma_u^2\): Varianz der gruppenspezifischen Konstante

Die klassenspezifische Regressionskonstante setzt sich aus der \(\alpha_{0}\) und dem klassenspezifischen Fehlerterm zusammen: \(\alpha_i = \alpha_0 + u_{i}\). Es wird angenommen, dass \(u_{i}\) und \(e_{ij}\) unabhängig voneinander sind und die Niveauunterschiede zwischen Klassen (\(u_{i}\)) normalverteilt sind und um einem Erwartungswert von Null mit einer bestimmten Varianz streuen. Um das Modell zu schätzen, muss im Gegensatz zum Modell mit festen Effekten nur ein weiterer Parameter geschätzt werden: Die Varianz von \(u_{i}\). Das Mehrebenenmodell mit zufälligen Effekten ist folglich wesentlich sparsamer als das mit festen Effekten. Es basiert allerdings auf der Annahme der Normalverteilung beider Komponenten des Fehlerterms. Aus diesem Grund ist hier eine Diagnose der Residuen besonders wichtig (z. B. durch Betrachtung von QQ-Plots der Residuen aus beiden Ebenen).

Mit den geschätzten Varianzkomponenten für \(\sigma_u^2\) und \(\sigma_e^2\) kann die Intraklassenkorrelation (ICC) geschätzt werden:

$$ICC =\sigma_u^2/(\sigma_u^2 +\sigma_e^2)$$

Die ICC ist ein Maß für die Abhängigkeit zwischen den Individuen innerhalb einer Klasse. Wenn die ICC sehr klein ist, oder sich als insignifikant herausstellt, ist die Abhängigkeit in den Daten eventuell nicht gravierend, trotz einer hierarchischen Datenstruktur. In diesem Fall kann es ratsam sein, zu einer weniger komplexen OLS Regression zurückzukehren, die unter der Annahme unabhängiger Beobachtungen valide Ergebnisse liefert.

Modell mit zufälligem Steigungsparameter (random slope model)

Nicht nur die Regressionskonstante, sondern auch Steigungsparameter können zwischen den Klassen variieren. Es wäre denkbar, dass aufgrund eines bestimmten klassenspezifischen Lernkonzepts die Schüler insgesamt sehr erfolgreich im Leistungstest sind, unabhängig vom gemessenen IQ des einzelnen Schülers. In dieser Klasse wäre der Effekt von der \(x\)-Variable IQ auf das Ergebnis im Leistungstest wahrscheinlich weniger stark, der Steigungsparameter von IQ wäre kleiner als in Klassen mit anderen Lernkonzepten. So könnte der Steigungsparameter \(\alpha_1\) zufällig zwischen den Klassen variieren. Ein solches Regressionsmodell mit klassenspezifischen zufälligen Steigungsparametern und Konstanten lautet

$$y_{ij} =\underbrace{\alpha_{0} + \alpha_1 x_{ij} +\alpha_2 z_{i} }_{feste\, \, Effekte} + \underbrace{c_{i} x_{ij} +u_i + e_{ij}}_{Fehlerterm}$$

$$ e_{ij} \sim N(0,\sigma_e^2), \,\, u_{i} \sim N(0,\sigma_u^2), \,\, c_{i} \sim N(0, \sigma_c^2)$$

$$Cov(u_i,c_i)=\rho$$

Modellparameter

- \(\alpha_{0}\): Regressionskonstante
- \(\alpha_1\): Effekt von \(x\) auf \(y\)
- \(\alpha_2\): Effekt der Kontextvariable \(z\) auf \(y\)
- \(\sigma_e^2\): Varianz des individuellen Fehlerterms
- \(\sigma_u^2\): Varianz der gruppenspezifischen Konstante
- \(\sigma_c^2\): Varianz des gruppenspezifischen Steigungsparameters

Der klassenspezifische Steigungsparameter setzt sich aus dem globalen Parameter \(\alpha_1\) und dem klassenspezifischen Fehlerterm zusammen: \(\alpha_{1i} = \alpha_1 + c_{i}\). Um das Modell zu schätzen, müssen zwei weitere Parameter geschätzt werden: Der Varianzparameter \(\sigma_c^2\) und \(\rho\), die Kovarianz zwischen \(c_{i}\) und \(u_{i}\).

Das bisher besprochene Zwei-Ebenen-Modell kann um mehrere Ebenen erweitert werden. Für jede Ebene können zufällige Regressionskonstanten und Steigungsparameter in das Modell aufgenommen werden. Mit jeder weiteren Ebene und jedem weiteren zufälligen Parameter, der in das Modell aufgenommen wird, steigert sich die Komplexität des zusammengesetzten Fehlerterms, und entsprechend die Anzahl der zu schätzenden Varianzkomponenten.

Modellwahl

Um sich für ein Modell zu entscheiden, sollte zunächst eine Variablenselektion durchgeführt werden. Hierbei sollten die erklärenden Variablen im Modell bleiben, die einen relevanten Teil der Variabilität der abhängigen Variablen erklären. Siehe hierzu auch das Kapitel zu Modellselektion in diesem Wiki.

Nachdem die Variablen für den systematischen Teil des Regressionsmodells ausgesucht wurden, können nach und nach zufällige Effekte in das Modell mit aufgenommen werden. Konkurrierende Modelle mit unterschiedlich komplexem Fehlerterm können mit Hilfe des Akaike information criterion (AIC) oder dem Bayesian information criterion (BIC) verglichen werden.

Umsetzung in Statistik Software

- Für R-User ist der R-Blogger Eintrag "Getting Started with Multilevel Modeling in R" eine nützliche Quelle, um einen Einstieg in die Mehrebenenanalyse zu finden. Hier wird die Analyse mit Hilfe des lme4 Pakets durchgeführt.
- Für Stata-User gibt es ein sehr hilfreiches Video vom STATA Youtube Channel. Hier werden die Befehle vorgestellt und der Output interpretiert.

Introduction to multilevel linear models in Stata

- Für JMP-User ist der On-Demand Webcast "Advanced Mastering JMP: Linear Mixed Models Using JMP® Pro" eine nützliche Quelle, um einen Einstieg in die Mehrebenenanalyse zu finden. Hier gibt es zwei Videos ("Overview and Random Coefficients Models" und "Repeated Measures and Panel Data Models"), die in die Mehrebenenanalyse mit JMP einführen.

- Für SAS-User können mit den Prozeduren *Proc Panel* und *Proc Mixed* Mehrebenenanalysen durchführen.

Einführung in Proc Panel

Einführung in Proc Mixed

- Es gibt spezielle Programme, die auf Mehrebenenanalyse spezialisiert ist: MLwiN, MPlus.
- Das Centre for Multilevel Modelling bietet eine umfassende Übersicht über mögliche Programme für die Anwendung von Mehrebenenanalysen.