BIFIE-SPSS-Makros – v1.6

Stand vom 13. Juli 2017; Konrad Oberwimmer

Features

- **BIFIE-konforme Berechnung der Schätzer und Standardfehler** gängiger statistischer Kennwerte in den Datensätzen des BIFIE.
- Läuft stabil bei wiederholten Durchgängen, lässt bisher geöffnete Daten unberührt, läuft mit nur einem INSERT FILE Befehl, Output ergänzt den bereits vorhandenen Output im Viewer.
  → **durchgängige Analysesyntax möglich**
- Alle Zwischenschritte werden wieder aufgeräumt und nicht auf der Festplatte gespeichert.
  → **bestmögliche Datensicherheit**
- Vollständig in SPSS-Syntax implementiert. Zentrale Makros steuern die Spezifikationen für verschiedene Datensätze (bspw. Anzahl der Replikationen, Anzahl der Imputation etc.).
  → **einfache und unmittelbare Verwendung für alle SPSS-Besitzer/innen**

ACHTUNG: Ein SPSS-Fehler in Version 21.0.0.0 führt zu fehlerhaften Berechnungen. Es muss – nur in dieser Version! - ein Patch (wenigstens v21.0.0.1) installiert werden, bevor die Makros richtig arbeiten.

Abdeckungsgrad

In Version 1.6 können die BIFIE-SPSS-Makros mit den Daten der folgenden Erhebungen umgehen:

Erhebung	Kürzel	Technische Details
Bildungsstandards Baseline 8. Schulstufe (erhoben 2009)	B809I B809S	Replicate Weights (JKK) und multiple Imputation (nicht nur PVs; in vers. Dateien)
Bildungsstandards Baseline 4. Schulstufe (erhoben 2010)	B410I B410K B410S	Replicate Weights (JKK) und multiple Imputation (nicht nur PVs; in vers. Dateien)
TIMSS/PIRLS 2011 (Schüler)	TP11I	Jackknife-Zonen und multiple Imputation für Plausible Values (PVs) und Proficiency Levels (PLs) (in einer Datei)
PISA 2012 (Schüler)	PS12I	Replicate Weights (BRR, fay=0.5) und multiple Imputation für PVs und PLs (in einer Datei)
Bildungsstandards-Überprüfung M8 2012	M812I M812S	multiple Imputation (nicht nur PVs; in vers. Dateien)
Bildungsstandards-Überprüfung M4 2013	M413I M413K M413S	multiple Imputation (nicht nur PVs; in vers. Dateien)
Bildungsstandards-Überprüfung E8 2013	E813I E813S	multiple Imputation (nicht nur PVs; in vers. Dateien)

Bildungsstandards-Überprüfung D4 2015	D415I D415S	multiple Imputation (nicht nur PVs; in vers. Dateien)
Bildungsstandards-Überprüfung D8 2016	D816I D816S	multiple Imputation (nicht nur PVs; in vers. Dateien)

Und es sind folgende Verfahren implementiert:

- Makro !UNIVAR: Mittelwert und Standardabweichung
- Makro !FREQ: Häufigkeiten, Anteilswerte und gültige Prozente
- Makro !CROSSTAB: (Bivariate) Kreuztabelle
- Makro !CORREL: Korrelation (listenweise Korrelationsmatrix)
- Makro !LINREG: (Multiple) Lineare Regression
- Makro !MEANDIFF: Mittelwertsunterschied bei zwei Gruppen

Verwendung der Makros

Grundlagen der Verwendung

INSERT FILE="C:\meinPfad\mcr_BIFIE_1_4.sps".

```
!UNIVAR
  dataset=B410I/
  file=B410I_S_IMP$_intern/
  path=C:\meinPfad/
  quick=yes/
  vars=[Variablennamen]/
  grp=[gruppierende Variable]/.
```

Das Makro !UNIVAR steht hier stellvertretend für die anderen Makros, da bei allen Makros die grundlegende Verwendung gleich ist.

- Mit dem INSERT-Befehl wird die Makro-Datei geladen. Das muss in einer Syntax (bzw. einer SPSS-Session) nur einmal gemacht werden.
- Alle Makros beginnen mit einem Rufzeichen, um sie von anderen SPSS-Befehlen zu unterscheiden.
- Auch ein Makro-Aufruf muss am Ende mit einem Punkt abgeschlossen werden.
- Alle Parameter sind mit einem Schrägstrich abzuschließen.
- Ansonsten sind Reihenfolge und Anordnung (Zeilenumbruch oder Leerzeichen-getrennt) der Parameter frei wählbar.

Parameter für alle statistischen Prozeduren:

Name	Beschreibung	Beispiele/Ausprägungen
dataset	Spezifiziert die Art des Datensatzes für die Analyse (siehe Kürzel auf S. 1)	Bsp.: B410I, TP2011
file	Dateiname der einen Datendatei oder gleichbleibender Dateiname für mehrere Imputationsdatensätze, wobei das $-Zeichen als Platzhalter für die Zählung des Imputationsdurchlaufes fungiert. Die Dateiendung .sav ist optional, muss also nicht angegeben werden.	Bsp.: B410I_S_IMP$_intern
path	[optional] Working Directory für die Analysen	Bsp.: P:\meineDaten
quick	[optional] Gibt an, ob die Sampling Variance nur vom ersten	no, yes

	Imputationsdatensatz geschätzt werden soll. Default: no	
showvariances	[optional] Zeigt bei ‚yes' die Sampling- und Imputationsvarianz im Output an. Default: no	no, yes
debug	[optional] Arbeitet bei ‚yes' mit voller Befehls-Ausgabe und entfernt temporäre Datasets nicht Default: no	no, yes
wgt	[optional] Statt dem üblichen Gewicht, das durch den Parameter dataset bestimmt wird, kann ein alternatives Gewicht angegeben werden.	wgtstud_D4S

Oben sind diejenigen Parameter beispielhaft verwendet, die für alle Makros relevant sind, während die für !UNIVAR speziellen Parameter *vars* und *grp* nur angedeutet sind.

- Mit dem *dataset*-Parameter wird spezifiziert, auf welcher Art Datensatz die Berechnung erfolgen wird. Bislang gibt es „B410I", „TP2011" und „PISA2012" als mögliche Einstellungen, womit die Schülerdatensätze von Baseline 4. Schulstufe, TIMSS/PIRLS 2011 und PISA 2012 gemeint sind.
  Der Parameter organisiert die dahinterliegenden Einstellungen, sodass man diese nicht selbst vornehmen muss.
  So wird im Beispiel (für Baseline 4.SSt.) implizit festgelegt: Jackknife-Methode, 132 Replicate Weights (w_fstr1 – w_fstr132), ein Final Student Weight (wgtstud) und 10 Imputationsdatensätze. Im Output werden diese impliziten Einstellungen zur Information ausgegeben.
- Mit *file* wird der Dateiname der einen Datendatei oder – im Fall von multiplen Imputationsdateien – der mehreren gleichbenannten Dateien angegeben. Das $-Zeichen im Namen ist dabei Platzhalter für eine laufende Nummer[1] bei Imputationsdatensätzen. Die Dateiendung (.sav) kann aber muss nicht angegeben werden. Das Makro öffnet also im Beispiel die Dateien „**B410I_S_IMP**1**_intern**.sav", „**B410I_S_IMP**2**_intern**.sav" usw.

Optional können noch folgende Parameter gesetzt werden:

- Das Makro setzt das Arbeitsverzeichnis auf den mit dem Parameter *path* spezifizierten Dateipfad. Dort sollten sich dann die Datensätze befinden. Es ist aber gleichwertig möglich, den Dateipfad direkt bei *file* zu spezifizieren oder das Arbeitsverzeichnis in der Syntax zunächst händisch zu setzen.
- Der Parameter *quick* ist ebenso für alle Makros relevant und ist standardmäßig auf den Wert „no" gesetzt. (Der Parameter kann also ausgelassen werden.) Wenn *quick* – wie im Beispiel – auf „yes" gesetzt wird, so wird das Makro die Sampling-Varianz nur aus dem ersten Imputationsdatensatz berechnen. Dieses Vorgehen wird zwar von OECD/IEA vorgeschlagen (vgl. OECD, 2009), entspricht aber nicht der methodisch korrekten Berechnung der Gesamtvarianz des Schätzers (vgl. Rubin, 1987). Eine Notwendigkeit, diese Einstellung vorzunehmen, kann aus der langen Berechnungszeit des SPSS-Makros entstehen; so auch die Begründung im PISA Data Analysis Manual.

ACHTUNG: Es wird empfohlen, in der freien Analysephase die Quick-Funktion zu nutzen und für die Berichtslegung alle relevanten Ergebnisse nochmals ohne Quick-Funktion zu replizieren, um BIFIE-konforme Standardfehler angeben zu können.

Es existieren noch zwei weitere allgemeine Parameter, die für die praktische Anwendung nicht relevant sind:

- Mit *showvariances=yes* kann eine gesonderte Ausgabe von Sampling- und Imputationsvarianz im Output angefordert werden. Diese sind für jeden Schätzer (*estimate*)

[1] Ohne führende Nullen: 1, 2, 3, …, 10, 11, …

im Output immer mit *varmi* (Varianz durch multiple Imputation) und *varrep* (Varianz durch Replicate Weights) bezeichnet. Diese Information ist in der üblichen Berichtslegung nicht relevant.

- Mit *debug=yes* wird das expandierte SPSS-Makro in einem separaten Output Viewer angezeigt und temporäre Datasets beibehalten. Wie der Name des Parameters andeutet, ist dies nur zur Auffindung von Fehlern relevant.

Hinweise zur Angabe von Variablennamen:

- Groß- und Kleinschreibung spielt in SPSS keine Rolle.
- Werden Listen von Variablen als Parameter übergeben, so sind die Elemente mit Leerzeichen abzutrennen (keine Beistriche, Strichpunkte, kein SPSS-Schlüsselwort TO)
- Befinden sich mehrfach imputierte Variablen (bspw. PVs) in einer Datendatei in mehreren Spalten, so ist das $-Zeichen als Platzhalter für die Zählung zu verwenden. Bspw: PV$MATH für PV1MATH, PV2MATH, … PV5MATH in PISA

HINWEISE: Während der Laufzeit der Makros öffnet und schließt SPSS mehrere Fenster, es werden mitunter auch Warnhinweise angezeigt, die ignoriert werden können (siehe Anhang A: Ignorierbare Warnhinweise).

Es wird empfohlen, in dieser Zeit SPSS nicht weiter zu verwenden. Allerdings ist es problemlos möglich, einstweilen mit anderen Programmen zu arbeiten.

Ist das Makro fertig, steht in der Statusleiste rechts unten wieder „IBM SPSS Statistics Prozessor bereit" und SPSS hat alle temporären Fenster geschlossen. Die Ausgabe wurde dem Output angefügt, der vor dem Makro-Aufruf aktuell war. Dieser Output wird vom Makro mit „mcr_bifie_prev" benann.

Auch der zuvor geöffnete Datensatz bleibt geöffnet und wird mit „mcr_bifie_prev" benannt. Wenn es einen solchen nicht gibt, weil nie ein Datensatz geöffnet war, dann erscheint im Output eine irrelevante Warnmeldung.

Makro !UNIVAR: Mittelwerte und Standardabweichungen (für Subgruppen)

```
!UNIVAR
  dataset=B410I/
  file=B410I_S_IMP$_intern/
  path=C:\meinPfad/
  vars=PVD PVM/
  grp=geschlecht/.
```

Laufzeit[2] normal: 76 sek.
Laufzeit quick: 34 sek.

Spezielle Parameter für !UNIVAR:

Name	Beschreibung
vars	Variablennamen der auszuwertenden Variablen (metrisch)
grp	[optional] Variablennamen der gruppierenden Variablen (kategorial)

Im Beispiel werden die Mittelwerte und Standardabweichungen für Deutsch und Mathematik getrennt nach Geschlecht ausgegeben. Das SPSS-Makro bedarf nur weniger Parameter ergänzt zu den allgemeinen Einstellungen, die oben beschrieben wurden.

- Mit dem Parameter *vars* können beliebig viele Variablen genannt werden, für die Mittelwerte und Standardabweichung (sowie deren Standardfehler) berechnet werden sollen.
  Die unter *vars* genannten Variablen müssen in allen Imputationsdatensätzen vorhanden sein. Werden eigene Variablen durch Rekodierung oder Umrechnung kreiert (bspw. Zusammenfassung von Werten zu Klassen), so ist diese Rekodierung auch in allen Datensätzen anzuwenden. Zur Unterstützung bei dieser mühseligen Aufgabe gibt es das Makro !DEPLOY, siehe S. 13.
- Mit dem Parameter *grp* – der optional ist – können diese Mittelwerte und Standardabweichungen für alle Variablen getrennt nach einer oder mehreren gruppierenden Variablen ausgegeben werden. Es kommt bei Angabe von *grp* aber zu keinem

[2] Laufzeiten wurden unter folgenden Bedingungen ermittelt: SPSS 22, Intel Core i7-3520M CPU @ 2x2,9GHz, 8GB Arbeitsspeicher, 64bit Windows 7
Als Beispieldatensatz wurde stets der Schülerdatensatz der Bildungsstandards Baseline 4. Schulstufe verwendet: 10 Imputationsdatensätze, 132 Gewichte, 9478 Fälle. Berechnungen an den nationalen Datensätzen der internationalen Studien gehen vergleichsweise schnell, da nur 5 PVs, 75-80 Gewichte und etwa die halbe Fallzahl.

Gesamtergebnis im Output, weswegen dafür u.U. ein erneuter Makro-Aufruf ohne *grp* notwendig ist.

Der Output enthält neben der allgemeinen Information zur Berechnung zwei wesentliche Tabellen. Die erste ist mit „Weighted and unweighted case count" beschriftet und gibt Informationen über die grundlegenden Fallzahlen im Datensatz (bzw. den einzelnen Gruppen) aus. Siehe dazu S. 9.
Die zweite ist mit „Estimates and standard errors" benannt und enthält das eigentliche Ergebnis. Für alle Gruppen – sofern vorhanden – und für jede Variable (pro Gruppe) sind in je einer Zeile angeführt:

1. Gewichtete …
2. … und ungewichtete Zahl gültiger Fälle
3. Name der Variable und des berechnten Kennwerts (MEAN oder SD)
4. *Estimate*: finale Schätzung des Kennwerts über alle Imputationsdatensätze
5. *SE*: Standardfehler des Schätzers
6. *df*: Freiheitsgrade des Schätzers nach der Formel von Rubin (1987); wenn keine Imputationsvarianz vorliegt (bei nicht-imputierten oder vollständigen Variablen), fehlt dieser Wert und Konfidenzintervalle und Signifikanzwerte (s.u.) werden anhand einer Standardnormalverteilung gebildet
7. *lower*: untere Grenze eines 95%-Konfidenzintervalls um den Punktschätzer
8. *upper*: obere Grenze eines 95%-Konfidenzintervalls um den Punktschätzer
9. *p_one*: einseitiger Signifikanzwert auf der t-Verteilung mit den entsprechenden Freiheitsgraden (siehe Rubin, 1987)
10. *fracmiss*: fraction of missing information = Anteil fehlender Information an der Gesamtinformation zur Bildung des Schätzers; wenn keine Imputationsvarianz vorliegt (bei nicht-imputierten oder vollständigen Variablen), fehlt dieser Wert
11. Die Spalten 3-10 erneut für die Standardabweichung.

HINWEIS: Die Werte SE, df, lower, upper, p_one und fracmiss werden im Folgenden zusammenfassend als „Inferenzwerte" des Schätzers bezeichnet.

Die weiteren Elemente der Ausgabe (wie etwa die Titel „Zusammenfassung" und die Tabellen „Verarbeitete Fälle") lassen sich in der Erstellung leider nicht unterdrücken und können einfach ignoriert werden.

```
!UNIVAR
  dataset=B410I/
  file=B410I_S_IMP$_intern/
  path=C:\meinPfad/
  vars= FE404RM FE40102C/
  grp=geschlecht/.
```

Laufzeit normal: 6 min. 40 sek.
Laufzeit quick: 1 min. 20 sek.

Spezielle Parameter für !FREQ:

Name	Beschreibung
vars	Variablennamen der auszuwertenden Variablen (kategorial)
grp	[optional] Variablennamen der gruppierenden Variablen (kategorial)

Das Makro berechnet einfache Häufigkeiten (cnt), Anteilswerte pro Variable (pct) und gültige Prozente (valid_pct) für alle angegebenen Variablen. Die Parametrisierung ist mit *vars* und *grp* genau gleich zu den univariaten Statistiken, siehe Ausführungen bei !UNIVAR auf S. 5. Im Beispiel werden also diese Kennwerte nach Geschlecht getrennt für die Schulbildung der Mutter (FE404RM) und das Herkunftsland der Mutter (FE40102C) berechnet.

Der Output enthält neben der allgemeinen Information zur Berechnung wenigstens zwei wesentliche Tabellen. Die erste ist mit „Weighted and unweighted case count“ beschriftet und gibt Informationen über die grundlegenden Fallzahlen im Datensatz (bzw. den einzelnen Gruppen) aus. Siehe dazu S. 9. Die weiteren Tabellen – beschriftet mit „Counts, percentages and valid percentages for … (per group)“ – gelten pro Variable und geben für diese jeweils die berechneten Kennwerte – sowie die zugehörigen Inferenzwerte[3] – aus.
Dabei ist in der Spalte „category“, die nach den gruppenbildenden Spalten kommt, der Wert jeder im Datensatz vorkommenden Merkmalsausprägung angeführt. Dieser Wert ist – aufgrund der Art der Berechnung im Makro – ohne das zugehörige Label angeführt. Es bedarf also des Wissens um die Bedeutung jeden Wertes. Weiters wird im Output nicht zwischen gültigen und fehlenden Werten unterschieden – außer implizit bei den gültigen Prozent.

[3] Da ein Nullhypothesentest für (relative) Häufigkeiten sinnlos ist, wird der einseitige Signifikanzwert (p_one) hier nicht ausgegeben.

```
!CROSSTAB
  dataset=B410I/
  file=B410I_S_IMP$_intern/
  path=C:\meinPfad/
  varrow=FE404RM/
  varcol=FS41501/.
```

Laufzeit normal: 1 min. 58 sek.
Laufzeit quick: 33 sek.

Spezielle Parameter für !CROSSTAB:

Name	Beschreibung
varrow	1 Variablenname der kategorialen Variable in den Zeilen
varcol	1 Variablenname der kategorialen Variable in den Spalten
sel	[optional] Filtervorschrift zum Auswählen von Fällen in Klammern

Im Beispiel wird die kreuztabularische Verteilung zwischen der höchsten abgeschlossenen Ausbildung der Mutter (FE404RM) und des voraussichtlichen weiteren Schulbesuchs nach der Volksschule (FS41501) berechnet.

- Mit *varrow* wird die zeilenbildende (kategoriale) Variable spezifiziert. Auch wenn der Output nachher nicht tatsächlich kreuztabularisch angeordnet ist, hat dies eine Bedeutung, da neben den Zellhäufigkeiten auch die zeilenweisen Prozent und die damit zu vergleichende Randverteilung ausgegeben werden. Es macht also einen Unterschied, welche der beiden Variablen hier angeführt wird und welche bei *varcol*.
- Der Parameter *varcol* spezifiziert die spaltenbildende (kategoriale) Variable. In einem typischen Anwendungsfall empfiehlt es sich, die Variable, die als abhängige Variable gedacht wird, in die Spalten zu geben, während die unabhängige Variable in die Zeilen kommt.
- Sowohl *varrow* als auch *varcol* akzeptieren nur eine Variable.
- Das Makro akzeptiert mit dem Parameter *sel* auch eine vorausgehenden Fallauswahl, die beim Makro !CORREL (S. 9) erklärt wird. Im Beispiel wird diese optionale Funktion nicht genutzt.

Der Output enthält neben der allgemeinen Information zur Berechnung vier wesentliche Tabellen.
Die erste ist mit „Weighted and unweighted case count“ beschriftet und gibt Informationen über die grundlegenden Fallzahlen im Datensatz (bzw. den einzelnen Gruppen) aus. Siehe dazu S. 9.
Die zweite heißt „Cell count“ und gibt für alle Kombinationen der Merkmalsausprägungen der beiden Variablen den finalen Schätzer für die Fallzahl (estimate) sowie dessen Inferenzwerte[4] aus. (siehe zu den Details die Ausführungen zum Makro !UNIVAR).
Die dritte heißt „Row percentages“ und liefert die Zeilenprozentwerte (wiederum inkl. Inferenzwerte). Die Zeilenprozentwerte summieren pro Ausprägung der Variable, die in der Tabelle links steht, zu 100%.
Die vierte heißt „Marginal distribution“ und gibt die von der anderen Variable unabhängigen relativen Häufigkeiten der spaltenbildenden Variable an (sowie deren Inferenzwerte). In einer typischen Argumentation zur unterschiedlichen Verteilung der Spalten-Variable je nach Ausprägung der Zeilen-Variable wird man die Zeilenprozent mit dieser Randverteilung vergleichen.

[4] Da ein Nullhypothesentest für kombinierte (relative) Häufigkeiten sinnlos ist, wird der einseitige Signifikanzwert (p_one) hier nicht ausgegeben.

Makro !CORREL: (Listenweise) Korrelationen unter mehreren Variablen

```
!CORREL
  dataset=B410I/
  file=B410I_S_IMP$_intern/
  path=C:\meinPfad/
  vars=PVD PVM FSHISEI/
  sel=(geschlecht=1)/.
```

Laufzeit ohne Selektion von Fällen: 21 min. 26 sek.
Laufzeit quick (ohne Selektion): 2 min. 10 sek.

Spezielle Parameter für !CORREL:

Name	Beschreibung
vars	Variablennamen der auszuwertenden Variablen (metrisch)
sel	[optional] Filtervorschrift zum Auswählen von Fällen in Klammern

Im Beispiel werden die Korrelationen zwischen den Plausible Values in Deutsch, Mathematik und dem HISEI berechnet. Und zwar speziell für Mädchen:

- *vars* gibt eine beliebig lange Liste an Variablen an, für die dann Korrelationen berechnet werden. Gegen eine starke Nutzung dieser Funktion (bspw. zum Erhalt einer vollen Korrelationsmatrix unter allen Variablen im Datensatz) spricht die lange Laufzeit des Makros. Es gilt die allgemeine Bedingung, dass alle diese Variablen in allen Imputationsdatensätzen vorhanden sein müssen.
- Mit dem optionalen Parameter *sel* kann eine Einschränkung auf bestimmte Fälle des Datensatzes vorgenommen werden. Wird der in Klammer stehende Ausdruck für einen Fall wahr, so wird dieser Fall in die Berechnung aufgenommen, ansonsten nicht. Im Beispiel wird so auf die Schülerinnen eingeschränkt, für die die Geschlecht-Variable den Wert 1 hat.
  Mit diesem Parameter werden indirekt auch Gruppenvergleiche möglich, es muss das Makro dazu mehrfach mit der Selektion von jeweils einer Gruppe aufgerufen werden.

HINWEIS: Das Makro berechnet die Korrelationsmatrix mit listenweisem Fallausschluss. Das bedeutet, dass ein Fall für alle paarweisen Korrelationen ausgeschlossen wird, wenn er nur bei einer der angeführten Variablen einen fehlenden Wert hat. Alle Korrelationen beruhen damit auf der gleichen Fallzahl.

Allerdings verliert man Fälle bei Korrelationen, für die u.U. weniger Missings vorliegen als in der gesamten Variablenliste. Um dies zu umgehen, ist das Makro mehrfach aufzurufen, mit jeweils zwei der Variablen für deren Korrelation. Bei 4 Variablen sind das etwa 6 Aufrufe, bei 5 Variablen 10 Aufrufe usw.

Der Output enthält neben der allgemeinen Information zur Berechnung zwei wesentliche Tabellen. Die erste ist mit „Weighted and unweighted case count“ beschriftet und gibt Informationen über die grundlegenden Fallzahlen im Datensatz (bzw. den einzelnen Gruppen) aus. Siehe dazu S. 9.
Die zweite ist mit „Correlations (listwise)“ benannt und enthält das eigentliche Ergebnis, die Korrelationen zwischen allen Variablen. Dabei wird jede Korrelation mit „r“ beginnend benannt, dann folgt der erste Variablenname, dann ein Unterstrich („_“) und dann der zweite Variablenname. Wie bei der univariaten Statistik sind für jede Korrelation der finale Schätzwert (*estimate*) sowie dessen Inferenzwerte angeführt. (siehe zu den Details die Ausführungen zum Makro !UNIVAR).

```
!LINREG
  dataset=B410I/
  file=B410I_S_IMP$_intern/
  path=C:\meinPfad/
  dep=PVM/
  pre=geschlecht FSHISEI FS40101C/.
```

Laufzeit normal (mit standard. Koeffizienten): 20 min. 30 sek.
Laufzeit normal (ohne standard. Koeffizienten): 5 min. 45 sek.
Laufzeit quick (mit standard. Koeffizienten): 3 min. 15 sek.
Laufzeit quick (ohne standard. Koeffizienten): 1 min. 15 sek.

Spezielle Parameter für !LINREG:

Name	Beschreibung
dep	1 Variablenname der abhängigen Variable (metrisch)
pre	Variablennamen der unabhängigen Variablen (metrisch)
sel	[optional] Filtervorschrift zum Auswählen von Fällen in Klammern
stdcoeff	[optional] Gibt an, ob standardisierte Koeffizienten (Beta) ausgegeben werden sollen. Wenn nicht, läuft das Programm merklich schneller durch. Default: yes; Alternativwert: no

Im Beispiel wird eine lineare Regression auf das Ergebnis in Mathematik ausgeführt, wobei drei Variablen als Prädiktoren/Regressoren verwendet werden: Geschlecht, HISEI und ob der/die Schüler/in im Ausland geboren ist.

- Der Parameter *dep* erwartet genau einen Variablennamen der abhängigen Variable.
- Der Parameter *pre* kann beliebig viele Variablen aufnehmen, welche die unabhängigen Variablen der linearen Regression sind. Wie immer gilt, dass alle Variablen in allen Imputationsdatensätzen vorhanden sein müssen.
- Das Makro akzeptiert mit dem Parameter *sel* auch eine vorausgehenden Fallauswahl, die beim Makro !CORREL (S. 9) erklärt wird. Im Beispiel wird diese optionale Funktion nicht genutzt.
- Eine weitere Möglichkeit dieses Makros ist, zur Beschleunigung der Berechnung die standardisierten Regressionskoeffizienten (Beta-Gewichte) abzuschalten. Dies geschieht mit *stdcoeff=no* und wird im Beispiel nicht genutzt.

Der Output enthält neben der allgemeinen Information zur Berechnung zwei wesentliche Tabellen. Die erste ist mit „Weighted and unweighted case count“ beschriftet und gibt Informationen über die grundlegenden Fallzahlen im Datensatz (bzw. den einzelnen Gruppen) aus. Siehe dazu S. 9.
Die zweite heißt „Regression coefficients“. Nach dem üblichen Schema (estimate und Inferenzwerte; siehe Details bei !UNIVAR, S. 5) werden folgende Kennwerte ausgegeben:

- b0: Die Regressionskonstante (Intercept)
- b[PRE]: unstandardisierte Regressionskoeffizienten für jede unabhängige Variable
- beta[PRE]: sofern nicht durch „stdcoeff=no/“ abgewählt, die standardisierten Regressionskoeffizienten für jede unabhängige Variable
- rsq: R^2 der linearen Regression

```
!LINREG
  dataset=B410I/
  file=B410I_S_IMP$_intern/
  path=C:\meinPfad/
  dep=PVM/
  pre=Geschlecht/
  diff=0 1/
  grp=FS40101C/.
```

Laufzeit: 2 min. 14 sek.
Laufzeit quick: 35 sek.

Spezielle Parameter für !MEANDIFF:

Name	Beschreibung
dep	1 Variablenname der abhängigen Variable (metrisch)
pre	1 Variablenname der unabhängigen Variable (kategorial)
diff	2 Zahlen (durch Leerzeichen getrennt) mit den Ausprägungen der zu vergleichenden Gruppen
grp	[optional] Variablennamen der gruppierenden Variablen (kategorial)

Das Makro für Mittelwertsdifferenzen bei zwei Gruppen hat vier spezifische Parameter:

- Der Parameter *dep* erwartet genau einen Variablennamen der Variable, für die Mittelwerte berechnet werden. Es ist zum momentanen Entwicklungszeitpunkt nicht möglich, hier mehr als eine Variable anzugeben.
- Mit dem Parameter *pre* kann ebenfalls nur eine Variable spezifiziert werden, nach der die zu vergleichenden Gruppen gebildet werden.
- Durch Angabe von zwei Zahlen beim Parameter *diff* werden die zu vergleichenden Gruppen spezifiziert. Dieser Parameter ist verpflichtend! Auch bei dichotomen Variablen kann SPSS die Gruppen nicht automatisch bilden sondern es bedarf der Angabe der Kodierungen.
- Mit dem Parameter *grp* – der optional ist – können die Mittelwertsdifferenzen getrennt für alle Merkmale/Merkmalskombinationen einer oder mehrerer gruppierender Variablen berechnet werden. Es kommt bei Angabe von *grp* aber zu keinem Gesamtergebnis im Output, weswegen dafür u.U. ein erneuter Makro-Aufruf ohne *grp* notwendig ist.

Das Beispiel läuft darauf hinaus, dass die Mittelwertsdifferenz im Lesen nach Geschlecht (0=männlich, 1=weiblich) berechnet wird; und zwar getrennt für die Gruppe der in Österreich geborenen Kinder (FS40101C=1) und der nicht in Österreich geborenen Kinder (FS40101C=2).

Der Output enthält neben der allgemeinen Information zur Berechnung drei wesentliche Tabellen. Die erste ist mit „Weighted and unweighted case count“ beschriftet und gibt Informationen über die grundlegenden Fallzahlen im Datensatz (bzw. den einzelnen Gruppen) aus. Siehe dazu S. 9.
Die zweite Tabelle mit Namen „Unstandardized mean difference(s)“ gibt die Mittelwerte und Standardabweichungen der abhängigen Variable für die beiden Vergleichsgruppen – u.U. aufgeteilt nach verschiedenen Gruppen – aus. In der Zeile für die jeweils höhere Ausprägung der Vergleichsgruppen folgen die unstandardisierte Mittelwertsdifferenz (aus Sicht dieser Vergleichsgruppe) und deren inferenzstatistische Kennwerte. Es gilt hier zu beachten, dass der angegebene Signifikanzwert von einer einseitigen Testung ausgeht.
In der dritten Tabelle „Standardized mean difference(s), Cohens d“ wird nach dem selben Schema die standardisierte Mittelwertsdifferenz (Effektstärkemaß Cohens d) ausgegeben und inferenzstatistisch getestet. Cohens d wird dabei über die gepoolte Varianz der beiden Vergleichsgruppen gebildet.

Ausgabe von Fallzahlen

Die SPSS-Makros liefern standardmäßig im Output eine Zählung von Fällen, die sowohl gewichtet (*n* bzw. *nvalid*) als auch ungewichtet (*nu* bzw. *nuvalid*) erfolgt. Mit *..valid* wird jene Fallzahl bezeichnet, bei denen die beteiligten Variablen keine fehlenden Werte aufweisen. Man sieht dies in der Tabelle „Weighted and unweighted case count".

Für die Makros !UNIVAR, !FREQ und !MEANDIFF sind die Fallzahlen nach Gruppen getrennt angeführt, wenn eine Gruppierung angefordert wurde, und es gibt eine Übersichtstabelle für die (gewichteten und ungewichteten) Gruppengrößen. Damit kann berechnet werden, wie viele Fälle (gewichtet und ungewichtet) pro Gruppe fehlen.

Die Makros !UNIVAR und !FREQ sind die einzigen, bei denen die Fallzahlen, wie auch die Kennwerte, für die Variablen einzeln berechnet werden. Es kommt also nicht zum listenweisen Ausschluss von Fällen, wenn ein Fall wenigstens einen Missing-Wert aufweist.

Bei allen anderen Makros sind die Kennwerte über listenweisen Ausschluss von Fällen berechnet. Die in der Übersichtstabelle genannten *valid*-Fallzahlen beziehen sich also auf die Fälle, die tatsächlich für die Berechnung verwendet wurden und sind für alle Kennwerte gleich. Die Angaben *n* und *nu* beziehen sich auf die (gewichtete und ungewichtete) Zahl von Fällen, die nach Anwendung eines Selektors (Parameter *sel*) noch verblieben sind. Dies erlaubt die Berechnung wie viel Prozent der gewählten (!) Fälle fehlen.

Hilfsmakro !DEPLOY

Liegen die multiplen Imputationen in verschiedenen Datenfiles vor, so wie es bei BIST-BL der Fall ist, muss bei eigenen Rekodierungen oder Umrechnungen von Variablen Sorge getragen werden, dass diese in allen Imputationsdatensätzen existieren, bevor sie zur Analyse mit den Makros verwendet werden können.
Diese Vorgang unterstützt das Makro !DEPLOY. Hier ein Beispiel:

```
!DEPLOY
  file1=B410I_S_IMP$_intern/
  nmi=10/
  path=C:\meinPfad/
  syntax=meineRekodierung.sps/
  outfile=B410I_S_IMP$_KOPIE/
  outpath=C:\meinNeuerPfad/.
```

Es ist folgendermaßen parametrisiert:

- *file1* bestimmt die Dateinamen wie bei allen anderen Makros auch ($-Zeichen als Platzhalter für Imputationszählung und Endung .sav ist optional). Damit dem Makro bekannt ist, wie viele Imputationsdateien vorliegen, ist hier der Parameter *nmi* auch zu setzen.
- *path* (optional) gibt in gewohnter Weise den Pfad an, der als Arbeitsverzeichnis verwendet wird und in dem die Datendateien erwartet werden.
- Mit *syntax* wird auf eine Datei verwiesen, welche Rekodierungs- und Berechnungssyntax enthält, die für alle Imputationsdatensätze gilt. Der Inhalt dieser Datei könnte etwa sein:

  ```
  NUMERIC geschlecht (F1.0).
  RECODE Geschlecht (0=2) (1=1) (ELSE=SYSMIS) INTO geschlecht.
  VALUE LABELS geschlecht 1 'weiblich' 2 'männlich'.
  EXECUTE.
  ```

  Hier wird die Variable Geschlecht zu geschlecht rekodiert (wenngleich sich am Informationsgehalt nichts ändert), mit Labels versehen und die anstehende Transformation verbindlich ausgeführt.
  Die Syntax-Datei kann sowohl relativ zu *path* – wie im Beispiel – angegeben sein oder auch durch einen absoluten Dateipfad. Die Endung .sps ist wiederum optional!
- *outfile* (optional) bestimmt die Dateinamen der zu speichernden Dateien. Er ist zu verwenden, wie Dateiangaben bei allen anderen Makros auch ($-Zeichen als Platzhalter für Imputationszählung und Endung .sav ist optional). Fehlt dieser Parameter (und auch outpath, s.u.), werden die unter *file1* angegebenen Dateien überschrieben.
- *outpath* (optional) bestimmt, in welchen Ordner die Dateien beim Speichern abgelegt werden sollen. Wenn er angegeben wird, wechselt das Arbeitsverzeichnis in diesen Ordner. Ist *outpath* verschieden zu *path*, so entsteht eine Kopie der Dateien.

WARNUNG: SPSS ist bei der Interpretation von Syntax aus einem Makro heraus strenger als im Direkt-Betrieb. So ist es bspw. notwendig, alle Kommentare mit Punkt abzuschließen!

HINWEIS: Es wird dringend empfohlen, für alle Rekodierungen gemeinsam eine Rekodierungssyntax anzulegen, welche auf dem Datensatz auch repetitiv lauffähig ist (kein Überschreiben von Originalvariablen), und keinerlei Speicherbefehle enthält.

So eine Rekodierungssyntax kann an einem einzelnen Imputationsdatensatz ausprobiert werden, um dann mit !DEPLOY auf allen Imputationsdatensätzen ausgebracht zu werden.

Dabei ist es empfehlenswert, ein Suffix für die Kopie anzuhängen, sodass die Rohdaten stets unverändert bleiben. Dies erhöht die Nachvollziehbarkeit von Analysen.

Für die Aufgaben der Aggregierung und des Mergings auf unterschiedlichen Ebenen (Schüler, Klassen, Schulen) steht ein weiterer Parameter zur Verfügung:

- *file2* bestimmt nach üblichem Muster die Dateinamen einer zweiten Datendatei, die jeweils parallel (also mit der gleichen Imputationszählung) zur ersten geöffnet wird. Die erste Datendatei (*file1*) wird in das SPSS-Dataset deploy_file1 geladen und die zweite Datei in deploy_file2. Zunächst ist das SPSS-Dataset deploy_file1 aktiv.

Durch das parallele Öffnen einer zweiten Datei aus dem gleichen Imputationsdurchlauf und die Konvention zur Benennung der Datasets ist es möglich eine Syntax zu schreiben, die etwa Schülermittelwerte (des jeweiligen Imputationsdurchlaufes) an den Klassendatensatz (des gleichen Imputationsdurchlaufs) anhängt.

Literatur

OECD (2009). PISA Data Analysis Manual. SPSS 2nd edition. OECD Publishing, Paris.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York.

Anhang A: Ignorierbare Warnhinweise

Nr.	Text
o.N.	Unbekanntes Daten-Set mcr_bifie_prev.
3211	Der Wert der Gewichtungsvariablen war für mindestens einen Fall null oder negativ oder der Wert fehlte. Solche Fälle sind für statistische Prozeduren und für statistische Grafik, die eine Fallgewichtung mit einem positiven Wert erfordern, nicht sichtbar, aber die Fälle verbleiben in der Datei und werden von nicht statistischen Funktionen wie z. B. LIST und SAVE verarbeitet.
5281	SPSS Statistics wird im Unicode-Codierungsmodus ausgeführt. Diese Datei ist in einer ländereinstellungsspezifischen (Codepage) Codierung codiert. Die definierte Breite einer Zeichenfolgevariablen wird automatisch verdreifacht, um einen möglichen Datenverlust zu vermeiden. Mit ALTER TYPE können Sie die Länge von Zeichenfolgevariablen auf die Länge des längsten beobachteten Werts der einzelnen Zeichenfolgevariablen setzen.