Dr. Michael Höfler

Modul MA-HPSTS-7: Advanced Multivariate Statistics

Fehlende Werte

Der rote Faden

1.Fehlende Werte führen zu unnötig **großem Zufallsfehler** (= Standardfehler, Konfidenzintervalle zu breit), wenn man die Individuen mit unvollständigen Werten in einer Analyse weglässt.

2. … und zu **Bias** - falls diejenigen mit unvollständigen Werten sich im Hinblick auf den zu schätzenden Parameter, z.B. Mittelwert oder Assoziation, unterscheiden.

3. Es gibt **drei Mechanismen** fehlender Werte: a) Werte fehlen **völlig zufällig („missing completely at random“)**; b) das Fehlen **hängt nur von beobachteten Variablen ab („missing at random“),** und Bias kann mit beobachteten Variablen verhindert werden; c) das Fehlen **hängt auch von unbeobachteten Variablen ab („missing not at random“),** und Bias kann nicht verhindert werden.

4. Fehlende Werte können **ersetzt (imputiert)** werden. Bei **einheitlicher Ersetzung** ersetzt man jeden fehlenden Wert mit demselben Wert, z.B. dem Mittelwert (aller vorhandenen Werte). Hierbei unterschätzt man den Standardfehler (weil man nur Information vervielfältigt) und unterschätzt Zusammenhänge.

5.Bei **individueller Ersetzung** ersetzt man bei unterschiedlichen Individuen mit unterschiedlichen Werten (z.B. vorhergesagte Werte aus Regressionsmodell).

6. Bei **deterministischer Ersetzung** ist jeder eingesetzte Wert durch die Daten eindeutig vorgegeben. (Jede einheitliche Ersetzung ist deterministisch, aber nicht umgekehrt.) Hiermit unterschätzt man die Standardfehler (weil man nur Information vervielfältigt).

7. Probabilistische Ersetzung zieht stattdessen für jeden zu ersetzenden Wert eine Zufallszahl aus einer Verteilung (z.B. Normalverteilung, die sich aus Vorhersage für einen fehlenden Wert einer Person mittels linearer Regression ergibt).

8. Problem: Das Ergebnis der probabilistischen Ersetzung ist **vom Zufall abhängig**.

9. Lösung: Man **ersetzt mehrmals zufällig** (**multiple Imputation**), führt die geplante Analyse dann in jedem aufgefüllten Datensatz durch und kombiniert die Ergebnisse zu einem Gesamtergebnis.

10. Breit anwendbares Verfahren der multiplen Imputation, bei dem Prädiktoren fehlender Werte selbst fehlende Werte haben können: **iterative chained equations.**

11. Die Ersetzung fehlender Werte funktioniert umso besser, je bessere Prädiktoren man für fehlende Werte hat. Je schlechtere Prädiktoren, umso stärker unterschätzt man meist Zusammenhänge.

Auswirkungen fehlender Werte

Auswertbare Stichprobe verkleinert sich

Complete-Case-Analyse: lasse jedes Individuum weg, für das mindestens der Wert einer für die Analyse notwendigen Variable fehlt.

Standardvorgehen aller Statistikprogramme

- **Beispiel 1**: berechne Mittelwert der Variable „kognitive Geschwindigkeit“ → alle Individuen werden rausgelassen, bei denen Wert von kognitive Geschwindigkeit fehlt.
- **Beispiel 2**: berechne Korrelation zwischen „kognitive Geschwindigkeit“ und „verbale Intelligenz“ → alle rausgelassen, wo Wert von kognitive Geschwindigkeit oder verbale Intelligenz fehlt.

Beispiel 3: Fünf benötigte Variablen haben je 10% Missings, die unabhängig voneinander auftreten. N = 1000.

- Wahrscheinlichkeit für komplette Information = $0.9^5 = 0.59$.
- → über 40% der Stichprobe verloren (0.59*1000 = 590 auswertbare Fälle)

Folge 1 also: mehr zufälliger Fehler

Tests

Die auswertbare Stichprobe reduziert sich → **geringere statistische Power** (= „größerer Beta-Fehler“)

Vorhandene Unterschiede können mit geringerer Wahrscheinlichkeit gefunden werden.

Schätzungen

Größere Standardfehler, breitete Konfidenzintervalle

→ Schätzungen weniger zuverlässig, weniger Erkenntnisgewinn

Folge 2: unterschiedliche auswertbare Stichproben

- Beispiel: In Stichprobe von N = 100 fehlen bei „kognitiver Geschwindigkeit“ 10 Werte.
- → N = **90** zur Berechnung des Mittelwerts von „kognitiver Geschwindigkeit“
- Bei 30 Individuen fehlt aber der Wert von „kognitiver Geschwindigkeit“ oder „verbaler Intelligenz“
- N = **70** zur Berechnung der Korrelation zwischen „kognitiver Geschwindigkeit“ und verbaler Intelligenz
- Problem nun: **unterschiedliche Stichproben** (N = 90 vs. 70), möglicherweise nicht vergleichbar!
- Ansatz: beide Auswertungen auf N = 70 reduzieren
- Aber: Stichprobe dadurch für Mittelwertsberechnung unnötig reduziert

Folge 3: systematischer Fehler (Bias*)

- Individuen, bei denen Werte fehlen, repräsentieren u.U. nicht alle Individuen
- z.B. könnten Individuen mit geringer „kognitiver Geschwindigkeit“ mit geringerer W.keit an entsprechendem Test teilnehmen
- → Überschätzung des Mittelwerts von „kognitiver Geschwindigkeit“
- **Bias = E(Schätzung) – wahrer Wert**, E(.) = Erwartungswert, mittelt zufälligen Fehler heraus (wenn man ganz oft auf dieselbe Weise Stichproben zöge).

Bias hängt immer davon ab, was man schätzen möchte. Z.B. können fehlende Werte größeren Bias in Schätzung der durchschnittlichen kognitiven Geschwindigkeit mit sich bringen als in deren Korrelation mit Alter.

Mechanismen fehlender Werte

1. Missing completely at random (MCAR)

- Ob ein Wert fehlt oder nicht, ist rein zufällig und hängt weder mit beobachteten, noch unbeobachteten Variablen zusammen.
- Die Stichprobe derer mit vollständigen Werten ist eine Zufallsstichprobe der Gesamtstichprobe.
- Praktische Implikation: Trotz fehlender Werte kommt bei Auswertung nichts systematisch anderes heraus → kein Bias, aber Stichprobe verkleinert (Analyse „ineffizient“)
- **Beispiel:** Es sind einige Fragebögen auf dem Postweg verlorengegangen (oder: unsystematische Eingabefehler, Ausfall eines Messgeräts aufgrund von Stromausfall).

2. Missing at random (MAR)

- Fehlende Werte treten nicht rein zufällig auf
- … aber ihr Auftreten hängt nur von beobachteten Variablen ab.
- (Technische Erklärung: stratifiziert man die Stichprobe nach diesen Variablen (Zellen bilden z.B. nach Alter und Geschlecht), hat man **in jeder Zelle eine Zufallsstichprobe** der Gesamtstichprobe, d.h. MCAR.)
- Man kann Bias sicher verhindern, indem man die Daten in Abhängigkeit von diesen Variablen auffüllt.
- **Simples Beispiel:** nur bei Männern im Alter von <20 sind Fragebögen verloren gegangen (weil sie nach Alter und Geschlecht sortiert waren).
- Oder: Je nach Alter und Geschlecht sind unterschiedlich viele Fragebögen verloren gegangen. (Innerhalb von Alters/-Geschlechts-Gruppen hat man Fragebögen rein zufällig verloren.)

3. Missing not at random (MNAR)

- Fehlende Werte treten nicht zufällig auf
- Ihr Auftreten hängt (auch) von unbeobachteten Variablen ab.
- Bias kann durch statistische Verfahren nicht garantiert vermieden werden!
- Konkret kommt es für Bias darauf an, was genau man schätzt.

Beispiel: Fehlen der Daten hängt neben Alter und Geschlecht auch von der unbeobachteten Variable „körperliches Handicap“ ab (Probanden mit Handicap haben Fragebögen häufiger nicht verschickt). Handicap hänge a) mit körperlicher Gesundheit zusammen, aber nicht mit b) Assoziation zwischen körperlicher Gesundheit und Geschlecht.
→ MNAR führt in a) zu Bias, aber nicht in b) (nachdem man die Daten mittels Alter und Geschlecht aufgefüllt hat).)

Wie sollte man sich praktisch verhalten?

- MCAR lässt sich leicht widerlegen: untersuche, ob (mindestens) eine beobachtete Variable mit dem Fehlen einer anderen Variable (kodierbar als 0 = vorhanden, 1 = fehlt) assoziiert ist.
- Ob jedoch MAR oder MNAR vorliegt, ist empirisch nicht eindeutig entscheidbar.
- Beispielsweise, weil in den Variablen, die man zum Auffüllen nutzt, oft wiederum Werte fehlen (→ unklar, ob Variablen das Fehlen von Werten vollständig erklärt).
- **Im Hinblick auf Bias** bei der Schätzung eines konkreten Parameters: Man kann nie mit Gewissheit sagen, ob sich ein Zusammenhangsmuster, mittels dessen man Daten auffüllt, von vorhandenen auf fehlende Daten übertragen lässt (implizite Annahme beim Datenauffüllen!, s.u.).
- **Beispiel:** Man füllt fehlender Werte von „körperlicher Gesundheitszustand“ mittels Regression und x-Variablen Alter, Geschlecht … auf (vollständige Daten). Unter denen, wo „körperlicher Gesundheitszustand“ fehlt (unbeobachtet), sind die β anders.

- Man kann jedoch mittels logistischer Regression ein multiples Modell zur **Vorhersage des Fehlens des Wertes einer Variable** (dichotome Variable) berechnen.
- Gutes Maß für Erklärungswert eines solchen Modells: AUC = Fläche unter der ROC-Kurve*
- **Heuristik**: bei gutem Erklärungswert des Modells ist man nahe an MAR ⟶ Daten auffüllen und aufgefüllte Daten auswerten.

* AUR = area under the curve = hier Wahrscheinlichkeit, dass ein Individuum mit (tatsächlich) fehlendem Wert eine größere Modellwahrscheinlichkeit hat (einen fehlenden Wert zu haben) als ein Individuum mit (tatsächlich) vorhandenem Wert. Range: 0.5 (kein Erklärungswert) – 1 (vollständiger Erklärungswert).

Ersetzung fehlender Werte

Einheitliche versus individuelle Ersetzung

- **Einheitliche Ersetzung**: jeden fehlenden Wert mit demselben Wert (z.B. Mittelwert der vorhandenen Werte) ersetzen
- **Individuelle Ersetzung**: jeden fehlenden Wert mit individuellem Wert (z.B. durch Regression aufgrund von individuellen Kovariablenwerten vorhergesagt, s.u.) ersetzen

Einheitliche Ersetzung mittels Mittelwert/Median/häufigstem Wert

Schätzung eines Verteilungsparameters (z.B. Mittelwert) einer Variable

MCAR: kein Bias

M(N)AR: Bias (Unterschätzung/Überschätzung möglich)

Schätzung eines Unterschieds/Zusammenhangs (z.B. Korrelation) zwischen zwei Variablen (**X** und **Y**), **X** und/oder **Y** hat fehlende Werte

MCAR: Bias (Zusammenhang unterschätzt*)

M(N)AR: Bias (Zusammenhang unterschätzt*)

* Grund für Unterschätzung: Ersetzt man jeden fehlenden Wert z.B. von **Y** durch denselben Wert, nimmt man an, dass **Y** von **X** unabhängig sei. Bei den Individuen, wo man die Werte von **Y** aufgefüllt hat, sind diese von **X** unabhängig!

Individuelle Ersetzung fehlender Werte in einzelnen Items eines Fragebogens

- Situation: Fragebogen aus Items, die zu einem Gesamtscore verrechnet werden sollen
- Hat eine Person die meisten Items beantwortet (z.B. mind. 80%), kann man Fragebogenscore dieser Person dennoch aus den beantworteten Items berechnen („schätzen“).

Beispiel: Person hat 8 von 10 Items beantwortet.

- **Mittelwertsscore:** bilde Mittelwert aus den 8 Items
- **Summenscore:** multipliziere Mittelwertsscore mit 10 (Gesamtanzahl der Items)
- Zulässig, wenn Fragebogen gute psychometrische Eigenschaften hat (insbesondere ähnliche Itemschwierigkeit), man nimmt an, dass bei einer Person Items rein zufällig fehlen.

Individuelle Ersetzung - mittels Regression

Ersetze z.B. fehlende Werte der Variable „verbale Intelligenz“ durch vorhergesagten Wert aus linearem Regressionsmodell

Y = verbale Intelligenz, **x** = Alter, Geschlecht, Bildung**.**

Vorhergesagter Wert von **Y** = verbale Intelligenz in Person i

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{Alter,i} + \hat{\beta}_2 x_{Geschlecht,i} + \hat{\beta}_3 x_{Bildungsjahre,i}$$

Schätzungen der Regressionskoeffizienten
(berechnet aus Teilstichprobe derer mit vollständigen Daten)

Werte der x-Variablen bei Person i
(verfügbar auch, wenn **Y**-Wert fehlt)

Vorgehen

1. Man schätzt unter allen Fällen, wo die Werte von allen vier Variablen vorhanden sind (complete cases), die β.

2. Es gibt Personen, für die die Werte der **x**-Variablen, aber nicht die von **Y** vorliegen.

3. Für diese kann man die **x**-Werte in die Gleichung einsetzen und damit **Y** vorhersagen.

4. Damit werden die fehlenden Werte von **Y** ersetzt.

Offenes Problem: Was tun, wenn auch Werte in **x**-Variablen fehlen (s.u.)?

Allgemeines Vorgehen bei **deterministischer Ersetzung**

- Bestimme aus vorhandenen Daten, welcher Wert eingesetzt werden soll
- Ersetze dann einen fehlenden Wert mit genau diesem Wert
- **Merkmal jeder deterministischer Ersetzung**: Durch gegebene Daten steht zu 100% fest, welcher Wert eingesetzt wird.

- **Problem:** Information nur von vorhandenen Werten auf fehlende Werte übertragen, damit bloß vervielfältigt
- Es wird so getan, als ob ein fehlender Wert mit Sicherheit derjenige wäre, der eingesetzt wird.
- Zusätzliche Varianz, die es in fehlenden Werten geben würde, bleibt damit unberücksichtigt.
- → **Unterschätzung der Varianz** in aufgefüllten Daten (→ Standardfehler einer Schätzung, z.B. Mittelwert oder Assoziation, unterschätzt; Konfidenzintervall zu schmal).

Weitere Verfahren

Kodiere fehlende Werte in kategorialen Variablen als eigene Kategorie

- Z.B. Bildungsgrad einfach, mittel, hoch, **keine Angabe**
- Kann nützlich sein, um die Daten deskriptiv darzustellen oder wenn „keine Angabe“ eine inhaltliche Bedeutung hat
- Tatsächlich aber mischen sich unter „keine Angabe“ Probanden mit einfacher, mittlerer und hoher Bildung in unbekanntem Verhältnis.
  → Ergebnisse nicht interpretierbar

Hotdeck-Methode

- **Idee**: Ersetze fehlende Werte anhand der Verteilung derer, die in **x**-Variablen ähnliche Ausprägungen haben
- **Beispiel verbale Intelligenz**: Teile Stichprobe in Strata (Schichten) gemäß Alter, Geschlecht und Bildung. Ersetze fehlenden Wert eines Probanden z.B. durch Mittelwert seines Stratums.
- Methode deterministisch oder probabilistisch (s.u.) verwendbar
- Probleme:

  - **x**-Variablen dürfen wiederum keine Missings haben

  - Alle **x**-Variablen müssen kategorial sein oder kategorisiert werden (Informationsverlust).

  - In jedem Stratum muss es genügend Fälle geben, um fehlende Werte reliabel ersetzen zu können → man kann nur wenige **x**-Variablen berücksichtigen

Probabilistische Ersetzungen

- Ersetze nicht mit festem Wert, sondern mit **Zufallszahl aus der Verteilung des vorhergesagten Wertes**

- **Beispiel:** Variable ist normalverteilt; schätze aus vollständigen Daten Mittelwert (Erwartungswert) mit 1.3 und SD mit 2.1

  → ersetze mit Zufallszahl aus Normalverteilung N(1.3 , 2.1)

- **Lineare Regression liefert normalverteilte Vorhersagen:** z.B. Mittelwert je nach Ausprägung von **x**-Variablen als 1.3, 1.5, 1.7 …. , SD (der Vorhersage) immer = 2.0

  → Zufallszahl aus N(1.3, 2.0) bzw. N(1.5, 2.0) bzw. N(1.7, 2.0) .. ...

Problem bei probabilistischer Ersetzung

Die Imputation (Ersetzung) mit einem einzigen Wert aus einer Verteilung berücksichtigt zwar die Unsicherheit im fehlenden Wert, **aber nicht die Varianz in der Ersetzung** (Varianz der Verteilung, aus der man eine Zufallszahl zieht).

→ **Unterschiedliche Ergebnisse**, wenn Verfahren wiederholt würde (Datensatz mehrfach aufgefüllt)!

Damit **unerwünschte Variation** in den Ergebnissen einer Analyse, die sich dem Ersetzen der Daten anschließt!

Lösung: Multiple Imputation (MI)

- Ersetze fehlende Werte **mehrfach** (multipel).
- Erzeuge somit **mehrere aufgefüllte Datensätze.**
- Statistische Auswertung erfolgt a) in jedem dieser Datensätze getrennt, und b) Ergebnisse werden aggregiert.
- → so viele aufgefüllte Datensätze, dass aggregiertes Ergebnis kaum noch vom Zufall abhängt, üblich: 20

Rubin-Formeln zum Aggregieren der Ergebnisse multipler Imputation (MI)

$\hat{Q}_j$ bezeichnet die Schätzung des interess. Parameters (z.B. Regressionskoeffizient) im aufgefüllten Datensatz j. Gesamtschätzung über **m** aufgefüllte Datensätze:

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j.$$

Mittlerer Standardfehler *innerhalb* der m Imputationen:

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U_j.$$

Mittlerer Standardfehler *zwischen* den m Imputationen:

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{Q}_j - \bar{Q}\right)^2.$$

Gesamtvaria

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B.$$ mit

$$df = (m-1)\left(1 + \frac{m\bar{U}}{(m+1)B}\right)^2.$$

Nur lernen:

Gesamtschätzung = Mittelwert der einzelnen Schätzungen

Gesamtvarianz setzt sich zusammen aus a) **Varianz innerhalb der aufgefüllten Datensätze** und b) **Varianz zwischen aufgefüllten Datensätzen**

Ergänzung:

Nicht klausurrelevant!

Full information maximum likelihood-Schätzung

- Komplexe statistische Schätzverfahren erlauben es, in einem Regressionsmodell die Parameter **unter Verwendung aller Angaben aller Individuen** zu schätzen.
- V.a. für fehlenden Werten in **Längsschnittsdaten** verfügbar
- Kein explizites (eigenes) Auffüllen der Daten nötig! Gleichungssystem mit fehlenden Werten gelöst
- Numerische Methode: „EM-Algorithmus“ (Estimation, Maximation)
- Ergebnisse wohl ähnlich wie wenn man vorhandene Messungen nutzt, um fehlende Messungen zu ersetzen (Simulationen)*.
- Beispiel: „Mixed-/random effects-/multivel-Modelle“ für Longitudinaldaten wie Interventionsstudien mit Dropout (Nichtteilnahme ab best. Messzeitpunkt)

* Wood, AM, Hillsdon M, Carpenter J. Comparison of imputation and modeling methods in the analysis of physical activity trial with missing outcomes. *International Journal of Epidemiology* 2005; 34, 89-99.

Derzeit **praktikabelste Methode der multiplen Imputation:** Iterative Chained Equations (**ICE**)

Situation

- Eine Analyse benötigt verschiedene Variablen $\mathbf{Y_1}$, $\mathbf{Y_2}$, … $\mathbf{Y_k}$, die jeweils fehlende Werte haben*
- $\mathbf{Y_1}$, $\mathbf{Y_2}$, … $\mathbf{Y_k}$ können unterschiedliches Skalenniveau und unterschiedlichen Verteilungstyp (Normalverteilung, Poisson-Verteilung etc.) aufweisen.

* inklusive Variablen, die man zum Ersetzen dieser Variablen braucht, die ihrerseits fehlende Werte haben können usw. (diese sind hiermit auch gemeint)

Grundidee

- Stelle für jede Variable $\mathbf{Y_j}$ mit fehlenden Werten eine **Regressionsgleichung** auf, in der $\mathbf{Y_j}$ abhängige Variable ist
- Regressionsmodell entspricht Skalenniveau (z.B. logistische Regression für binäre Variablen) und Verteilungstyp* von $\mathbf{Y_j}$.
- Ergibt ein Gleichungssystem, das numerisch gelöst und in dem am Ende jedes $\mathbf{Y_j}$ ersetzen werden kann

* Siehe Thema „verallgemeinerte lineare Modelle“ (generalized linear models) im Kurs von Rudolf.

Konkreter Ablauf (nicht klausurrelevant)

- Sortiere die $\mathbf{Y_j}$ so dass die erste Variable die wenigsten Missings hat, die zweite die zweitwenigsten usw. → neue Variablenreihenfolge $\mathbf{Y_1'}$, $\mathbf{Y_2'}$, … $\mathbf{Y_k'}$
- Grund: $\mathbf{Y_1'}$ ist am besten aufzufüllen und ´damit wahrscheinlich besserer Prädiktor für andere aufzufüllende Variablen als $\mathbf{Y_2'}$, … $\mathbf{Y_k'}$. Usw.
- Ersetze die Missings in $\mathbf{Y_1'}$ (durch vorhergesagten Wert aus Regressionsmodell)
- Dazu müssen die fehlenden Werte in $\mathbf{Y_2'}$, …, $\mathbf{Y_k'}$ zunächst einfach ersetzt werden (z.B. durch Mittelwert).
- Ersetze die Missings in $\mathbf{Y_2'}$
- …
- Ersetze die Missings in $\mathbf{Y_k'}$.
- Nun ist ein Durchlauf (Cycle) beendet.

- Man hat jetzt alle fehlenden Werte in $\mathbf{Y_1'}$, $\mathbf{Y_2'}$, … $\mathbf{Y_k'}$ ersetzt und kann diese verwenden, um im nächsten Durchlauf bessere Ersetzungen zu erhalten.
- Es erfolgen nun so viele Durchläufe, bis sich die geschätzten Regressionskoeffizienten der Gleichungen zur Ersetzung der $\mathbf{Y_1'}$, $\mathbf{Y_2'}$, … $\mathbf{Y_k'}$ nicht mehr nennenswert ändern („die Werte konvergieren“).

- Nun kann für jeden fehlenden Wert jeder aufzufüllenden Variable $\mathbf{Y_1}$, $\mathbf{Y_2}$, … $\mathbf{Y_k}$ die Verteilung bestimmt werden, aus der (dem jew. Modell nach) der fehlende Wert stammt (gemäß Modellvorhersage).
- Aus diese Verteilungen werden jetzt **Zufallszahlen** gezogen.
- Jede Ziehung (aller fehlenden Werte; alle Variablen, alle Individuen) ergibt einen aufgefüllten Datensatz (multiple Imputation).

Bedeutung dieses Verfahrens

- Am breitesten anwendbare Methode, um fehlende Werte aufzufüllen
- Gibt es prinzipiell inzwischen auch in SPSS (→ „analysieren“ → „multiple Imputation“)
- Modul in SAS, das sich sogar die Modelle für die **Y** selbst sucht
www.ats.ucla.edu/stat/sas/seminars/missing_data/mi_new_1.htm

Anwendung auf DEGS-Daten

Im Datensatz von N = 4483 sollen drei neuropsychologische Variablen aufgefüllt werden. Die entspr. neuropsychologischen Tests wurden nicht bei den Probanden durchgeführt, die nur ein Telefoninterview gemacht haben, weisen daher (viele) Missings auf:

- **LDS**: Letter digit test, kognitive Geschwindigkeit (N=3867, hohe Werte = hohe kognitive Geschwindigkeit)
- **TMT-A**: Trail marking test, # Fehler bei visueller Suche (N=3884)
- **TMT-B**: Trail marking test, # Fehler bei visueller Aufmerksamkeit/Target switching (N=3839)

Andere zur Ersetzung verwendete Variablen

Variable	Beschreibung	
hznp12	Verbale Intelligenz (Wortschatztest)	
hznp5	Verbales Arbeitsgedächtnis	
sex	Geschlecht	
age age_sq age_cu	Alter in Jahren linear, quadratisch, kubisch	
gkpol4	Politische Gemeindeklassengröße (wie bisher)	
SDses	Soziökonomischer Status (wie bisher)	
SDcasmin	Bildungsgrad (wie bisher)	
MIausl	Migrationshintergrund (wie bisher)	
fam	Familienstand (wie bisher)	
partner	Zusammenleben mit Partner (wie bisher)	
beruf	Berufstätigkeit (wie bisher)	
ost	Ost-/Westdeutschland (0 = West, 1 = Ost)	
yuany8_di	irgendeine affektive Störung in den letzten 12 Monaten (1 = ja, 0 = nein)	
yuany5di	irgendeine Angststörung in den letzten 12 Monaten	
yuany3_di	irgendeine Substanzstörung in den letzten 12 Monaten	
analog d*	Lebenszetdiagnosen	
anz12 anz12_sq	Anzahl Einzeldiagnosen psychischer Störungen in den letzten 12 Monaten, linear/quadratisch	
anzlt anzlt_sq	Anzahl Einzeldiagnosen psychischer Störungen im Leben, linear/quadaratisch	

Beispiel Modellbildung für LDS (kognitive Geschwindigkeit) mittels schrittweiser Selektion

```
. #delimit;
delimiter now ;
. stepwise , pr(.05) pe(.01) :
>  regress LDS age age_sq sex duany8_di yuany8_di duany5di yuany5di  (_ISDses_2 _ISDses_3) (_ISDcasmin_2 _ISDc
> asmin_3) hznp12 hznp5
>             beruf partner anz12 anz12_sq anzlt anzlt_sq (_Igkpol4_2 _Igkpol4_3 _Igkpol4_4) (_Ifam_2 _Ifam_3
>  _Ifam_4 _Ifam_5) ost  ;
                      begin with full model
p = 0.8309 >= 0.0500  removing yuany5di
p = 0.6699 >= 0.0500  removing anz12
p = 0.4495 >= 0.0500  removing yuany8_di
p = 0.4055 >= 0.0500  removing _Ifam_2 _Ifam_3 _Ifam_4 _Ifam_5
p = 0.3677 >= 0.0500  removing duany8_di
p = 0.2064 >= 0.0500  removing anz12_sq
p = 0.1258 >= 0.0500  removing ost
p = 0.1470 >= 0.0500  removing _Igkpol4_2 _Igkpol4_3 _Igkpol4_4
p = 0.0685 >= 0.0500  removing age_sq
```

Lineare Regression mit schrittweiser Variablenselektion
(kombinierte Rückwärts-/Vorwärtsselektion)

Beispiel LDS

Source	SS	df	MS
Model	113805.501	13	8754.26931
Residual	120247.163	3746	32.1001502
Total	234052.664	3759	62.2646086

Number of obs = 3760
F(13, 3746) = 272.72
Prob > F = 0.0000
R-squared = 0.4862
Adj R-squared = 0.4845
Root MSE = 5.6657

LDS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.2593687	.0076263	-34.01	0.000	-.2743209	-.2444165
beruf	.9184922	.2208453	4.16	0.000	.4855035	1.351481
sex	3.435023	.1903925	18.04	0.000	3.06174	3.808306
anzlt	-.6662899	.1710456	-3.90	0.000	-1.001641	-.3309382
partner	.9314402	.2254985	4.13	0.000	.4893285	1.373552
duany5di	.7300287	.341398	2.14	0.033	.0606845	1.399373
anzlt_sq	.0545891	.0244957	2.23	0.026	.006563	.1026153
_ISDses_2	1.687363	.2953035	5.71	0.000	1.108391	2.266334
_ISDses_3	2.675499	.405619	6.60	0.000	1.880243	3.470754
_ISDcasmin_2	.8286424	.2452363	3.38	0.001	.3478327	1.309452
_ISDcasmin_3	1.011617	.3766311	2.69	0.007	.2731949	1.750039
hznp12	.2715704	.0210785	12.88	0.000	.2302439	.3128969
hznp5	.5951455	.0531344	11.20	0.000	.4909703	.6993206
_cons	29.62587	.6823671	43.42	0.000	28.28802	30.96371

Analog wird nun für jede Variable mit Missings ein Modell gebildet (aber teilweise ein generalisiertes lineares Modell verwendet) ...

Imputation dann mittels Befehl

`mi ice`

Dieser Befehl gehört, anders als die anderen „mi …“-Befehle, nicht zum Standardpaket von Stata, kann aber (wie viele andere) übers Netz installiert werden:

findit mi_ice

Im aufgehenden Fenster auf den Link
mi_ice from `http://www.homepages.ucl.ac.uk/~ucakjpr/stata`
klicken,

im nächsten Fenster auf (click here to install) .

```
. mi set flong

.
. // aufgefüllte Var. registrieren:
. mi register imputed LDS TNT_a TNT_b SDses  SDcasmin
(692 m=0 obs. now marked as incomplete)
```

Beim Auffüllen soll gleich Datensatz in langem Format entstehen

für jeden Probanden 1 (ursprüngliche Werte) + 20 (aufgefüllte Werte) Zeilen

Aufzufüllende Variablen müssen als solche „registriert“ werden

```
. /// hier o.- und m.-variablen zusätzlich angeben, sonst gibt's unten eine fehlermeldung
> xi i.SDses i.SDcasmin
i.SDses           _ISDses_1-3         (naturally coded; _ISDses_1 omitted)
i.SDcasmin        _ISDcasmin_1-3      (naturally coded; _ISDcasmin_1 omitted)
```

Mehrkategoriale Variablen müssen hier aufgeführt werden, damit Dummy-Variablen berechnet werden
(hier wird Syntax **ibX.** nicht unterstützt)

```
delimiter now ;
. mi ice
> /* ordinale variablen mit o. angeben, qualitative mit m. - bei den gleichungen dann i. */
>
> /* Kovariablen */
> age age_sq age_cu sex out2_duany8_di anz12  hznp12 hznp5
>
> /* Outcomes, o.: ordinal, m.: nominal */
> LDS TMT_a  TMT_b o.SDses o.SDcasmin
>
>
> /* optionen */
>         ,
>         add(20)
>         seed(464364753)
>
> /* Liste Regressionsarten je nach Outcome */
>          cmd(  LDS:    regress
>              , TMT_a : nbreg
>              , TMT_b : nbreg
>                  , i.SDses: ologit
>                  , i.SDcasmin: ologit
>                  )
>
> /* Gleichungen */
> eq(
> LDS: age age_sq sex duany8_di  _ISDses_2 _ISDses_3 _ISDcasmin_2 _ISDcasmin_3 hznp12 hznp5
> TMT_a: age age_sq anz12 beruf_ hznp5 hznp12
> TMT_b: age age_sq anz12 _ISDses_2 _ISDses_3  hznp12 hznp5
> SDses: age age_sq sex age_cu  sex_age sex_age_sq age_cu
> SDcasmin: age age_sq sex age_cu
> hznp12: sex age sex*age age_sq i.SDcasmin
> hznp5: age  i.SDcasmin
>    )
```

Komplexe Syntax zur Durchführung der Imputation

#missing values	Freq.	Percent	Cum.
0	3,723	83.05	83.05
1	113	2.52	85.57
2	46	1.03	86.59
3	450	10.04	96.63
4	49	1.09	97.72
5	101	2.25	99.98
7	1	0.02	100.00
Total	4,483	100.00	

Dann Ausgabe:

Auflistung, wieviele Probanden Missings in wievielen Variablen haben

Variable	Command	Prediction equation
age		[No missing data in estimation sample]
age_sq		[No missing data in estimation sample]
age_cu		[No missing data in estimation sample]
sex		[No missing data in estimation sample]
duany8_di		[No missing data in estimation sample]
anz12		[No missing data in estimation sample]
SDses	ologit	age age_sq age_cu sex duany8_di anz12 hznp12 hznp5 LDS TMT_a TMT_b _ISDcasmin_2 _ISDcasmin_3
_ISDses_2	ologit	[Passively imputed from (SDses==2)]
_ISDses_3	ologit	[Passively imputed from (SDses==3)]
SDcasmin	ologit	age age_sq age_cu sex duany8_di anz12 hznp12 hznp5 LDS TMT_a TMT_b _ISDses_2 _ISDses_3
_ISDcasmi~2	ologit	[Passively imputed from (SDcasmin==2)]
_ISDcasmi~3	ologit	[Passively imputed from (SDcasmin==3)]
hznp5	regress	age age_sq age_cu sex duany8_di anz12 hznp12 LDS TMT_a TMT_b _ISDses_2 _ISDses_3 _ISDcasmin_2 _ISDcasmin_3
hznp12	regress	age age_sq age_cu sex duany8_di anz12 hznp5 LDS TMT_a TMT_b _ISDses_2 _ISDses_3 _ISDcasmin_2 _ISDcasmin_3
TMT_a	nbreg	age age_sq age_cu sex duany8_di anz12 hznp12 hznp5 LDS TMT_b _ISDses_2 _ISDses_3 _ISDcasmin_2 _ISDcasmin_3
LDS	regress	age age_sq sex duany8_di _ISDses_2 _ISDses_3 _ISDcasmin_2 _ISDcasmin_3 hznp12 hznp5 TMT_a
TMT_b	nbreg	age age_sq age_cu sex duany8_di anz12 hznp12 hznp5 LDS TMT_a _ISDses_2 _ISDses_3 _ISDcasmin_2 _ISDcasmin_3

Auflistung aller Regressions-gleichungen

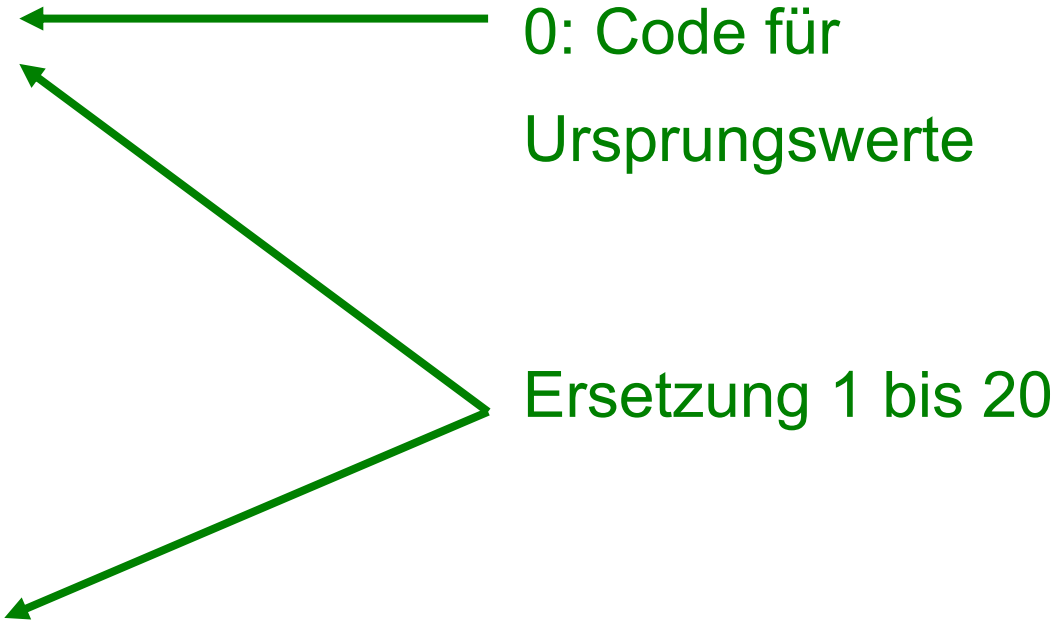
```
Imputing .........1.........2.........3.........4.........5.........6.........7.........8.........9..
> ........10.........11.........12.........13.........14.........15.........16.........17.........18..
> ........19.........20
file C:\Users\Hoefler\AppData\Local\Temp\ST_2e000001.tmp saved

[note: imputed dataset now loaded in memory]
(20 imputations added; M=20)
```

20 imputierte Datensätze erzeugt

Die nach der Imputation automatisch angelegte Variable *_mi_m* gibt an, welche Zeile im Datensatz zu welcher Imputation gehört

```
. tab _mi_m
```

_mi_m	Freq.	Percent	Cum.
0	4,483	4.76	4.76
1	4,483	4.76	9.52
2	4,483	4.76	14.29
3	4,483	4.76	19.05
4	4,483	4.76	23.81
5	4,483	4.76	28.57
6	4,483	4.76	33.33
7	4,483	4.76	38.10
8	4,483	4.76	42.86
9	4,483	4.76	47.62
10	4,483	4.76	52.38
11	4,483	4.76	57.14
12	4,483	4.76	61.90
13	4,483	4.76	66.67
14	4,483	4.76	71.43
15	4,483	4.76	76.19
16	4,483	4.76	80.95
17	4,483	4.76	85.71
18	4,483	4.76	90.48
19	4,483	4.76	95.24
20	4,483	4.76	100.00
Total	94,143	100.00	



Multipel imputierte Datensätze können dann mittels Präfix `mi estimate` und den üblichen Regressionsbefehlen ausgewertet werden

```
. mi estimate: regress LDS ib1.fam sex age
(system variable _mi_id updated due to changed number of obs.)
(imputed variables TNT_a TNT_b SDses SDcasmin hznp5 hznp10a hznp12 beruf_ partner
_ISDcasmin_2 _ISDcasmin_3 unregistered because not in m=0)
(143 m=0 obs. now marked as complete)

Multiple-imputation estimates                     Imputations     =         20
Linear regression                                 Number of obs   =       4478
                                                  Average RVI     =     0.1436
                                                  Largest FMI     =     0.1788
                                                  Complete DF     =       4471
DF adjustment:   Small sample                     DF:     min     =     525.40
                                                          avg     =    1204.88
                                                          max     =    2515.17
Model F test:       Equal FMI                     F(   6, 2755.3) =     337.02
Within VCE type:          OLS                     Prob > F        =     0.0000
```

LDS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fam						
2	.5910888	.9047747	0.65	0.514	-1.183091	2.365268
3	-.536041	.3842913	-1.39	0.163	-1.290041	.2179588
4	-1.80473	.4451631	-4.05	0.000	-2.677806	-.9316533
5	-1.495175	.3175572	-4.71	0.000	-2.118645	-.8717053
sex	3.198128	.2150499	14.87	0.000	2.775665	3.620591
age	-.274939	.0080987	-33.95	0.000	-.2908326	-.2590454
_cons	45.15405	.4991559	90.46	0.000	44.17417	46.13392

Hängt kognitive Geschwindigkeit vom **Familienstand** ab, wenn man Alter und Geschlecht berücksichtigt?

Verwitwete und nie Verheiratete haben niedrigere kogn. Geschwindigkeit als Verheiratete

Literatur

- Carpenter J., Kenward M. *Multiple Imputation and its Application (Statistics in Practice),* John Wiley & Sons, 2013
- Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*. 2006; 59:1087-91.
- Royston P. 2007. Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring. *Stata Journal* 7: 445-464.
- Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research* 1999; 8: 3-15
- (Seaman S et al, What is meant by „missing at random“? *Statistical Science* 2013; 28:257-268.)

Aufgaben

1a) Untersuchen Sie mit dem nichtaufgefüllten Datensatz (`Fehlende Werte Daten.dta`) und linearer Regression, wie **LDS = kognitive Geschwindigkeit** mit **beruf = Berufstätigkeit** zusammenhängen. Verwenden Sie dazu lineare Regression und keine Kontrollvariablen.

b) Ersetzen Sie nun die fehlenden Werte von kognitive Geschwindigkeit, indem sie für alle den Mittelwert aller vorhandenen Werte einsetzen. Am einfachsten geht dies mit der Syntax:

```
summarize LDS
replace LDS =r(mean) if LDS==.
```

Vergleichen Sie den geschätzten Regressionskoeffizienten mit dem Ergebnis in a. und geben Sie eine kurze Erklärung für den Unterschied.

2a) Untersuchen Sie nun mit dem aufgefüllten Datensatz „Fehlende Werte Daten aufgefüllt.dta“, wie **LDS = kognitive Geschwindigkeit** mit **beruf = Berufstätigkeit** zusammenhängt. Warum liegt das Ergebnis für den Regressionskoeffizienten zwischen den Ergebnissen in a) und b)?

b) Welche(s) der drei Konfidenzintervalle, glauben Sie, aus 1a), 1b) und 2a) ist/sind falsch? Inwiefern ist es/sind sie falsch und warum (jeweils kurze Begründung)?

3) Beurteilen Sie die Aussage: „Ob man Bias in aufgefüllten Daten hat, kann man untersuchen, wenn man das Ergebnis (hier: Schätzung des Regressionskoeffizienten von „beruf“) der aufgefüllten Daten mit dem der ursprünglichen, nichtaufgefüllten Daten vergleicht.“ Glauben Sie, dass diese Aussage (allgemein) stimmt (Begründung)?

Aufgabenlösungen

1a) Mit nichtaufgefülltem Datensatz:

```
regress LDS beruf
```

Berufstätige haben im Durchschnitt eine um 5,8 (95%KI = 5,4-6,3) größere kognitive Geschwindigkeit

b)

```
summarize LDS
replace LDS=r(mean) if LDS==.
regress LDS beruf
```

Berufstätige haben im Durchschnitt nur noch eine um 5,1 (95%KI = 4,7-5,5) größere kognitive Geschwindigkeit. **Erklärung**: LDS-Werte wurden unter der Annahme ersetzt, dass sie unabhängig von allen anderen Variablen, also auch Berufstätigkeit seien. Dadurch wird der Zusammenhang kleiner.

2a)
Mit nichtaufgefülltem Datensatz:

```
mi estimate: regress LDS beruf
```

Berufstätige haben im Durchschnitt eine um 5,6 (95%KI = 5,1-6,0) größere kognitive Geschwindigkeit. Das Ergebnis in 1a. beschreibt den Zusammenhang zwischen den *beobachteten* Werten von kognitive Geschwindigkeit und Berufstätigkeit, 1b. ist die denkbar schlechteste Ersetzung.

Nach multipler Imputation sollte der Zusammenhang größer erscheinen als in 1b), aber kleiner als im nichtaufgefüllten Datensatz:

```
mi estimate: regress LDS beruf sex age ib1.fam
```

b)
Das Konfidenzintervall in 1b) ist zu schmal, weil es die statistische Unsicherheit in der (deterministischen) Ersetzung nicht berücksichtigt. (Die in 1a) und 1c) sind korrekt, aber 1a) ist ineffizient (unnötig breit).)

3)

Es kann sein, dass durch fehlende Werte Bias in den nichtaufgefüllten Daten entsteht (complete-case-Analyse), dass also die Individuen mit fehlenden Werten nicht alle Individuen repräsentieren (MAR/MNAR). Bei der bisherigen Betrachtung haben wir implizit Complete-Case-Analyse als „Gold Standard“ zur Untersuchung von Bias durch das Auffüllen der Daten betrachtet. Begründung dafür ist, dass das Auffüllen der Daten in der Praxis nie perfekt ist, wodurch i.d.R. zusätzlicher Bias entsteht. Dabei unterstellten wir heuristisch, dass der Bias in der Complete-case-Analyse am geringsten ist. Dies muss aber nicht so sein, und wir können dies nicht empirisch überprüfen, weil wir die wahren fehlenden Werte nicht kennen.