Datenschutzhandbuch

Das Forschungsdatenzentrum (FDZ) der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung (IAB) ermöglicht externen Wissenschaftlerinnen und Wissenschaftlern den Zugang zu Mikrodaten für die nicht-kommerzielle Forschung im Bereich der Sozialversicherung und der Arbeitsmarkt- und Berufsforschung. Bei den vom FDZ bereitgestellten Daten handelt es sich um Sozialdaten, die den Datenschutzbestimmungen des Sozialgesetzbuches X (SGB X) bzw. den Vorgaben der statistischen Geheimhaltung unterliegen.

Neben der Erstellung und Dokumentation von Datenprodukten für die Arbeitsmarkt- und Berufsforschung besteht die Hauptaufgabe des FDZ darin, Datennutzende bei der Sicherstellung der Geheimhaltung von Informationen über statistische Einheiten wie Personen oder Betriebe zu unterstützen. Um dies zu gewährleisten muss nach Auswertungen per kontrollierter Datenfernverarbeitung vor Herausgabe von Ergebnisdateien deren Deanonymisierungspotenzial geprüft werden.

Die Datennutzung sollte in der Regel in Stata erfolgen. Sind vereinzelte Auswertungen mit Stata nicht durchführbar, besteht die Möglichkeit, dafür mit R oder Octave zu arbeiten. Kontaktieren Sie hierfür bitte im Vorfeld das FDZ.

Die Nutzung der schwach anonymisierten Daten des FDZ erfolgt mittels kontrollierter Datenfernverarbeitung oder im Gastaufenthalt. Für beide Zugangswege setzt das FDZ die Browser-basierte Software JoSuA[1] (Job Submission Application) ein. Bitte nutzen Sie für die Datenaufbereitung[2], die Vorbereitung von Gastaufenthalten und zum Testen der Lauffähigkeit der Auswertungsprogramme den Internal Use Modus in JoSuA. Die für Sie sichtbaren Ergebnisse werden hier einer automatischen Datenschutzprüfung unterzogen und zensiert. Die Ergebnisse können in der JoSuA-Webschnittstelle betrachtet, aber nicht heruntergeladen werden[3]. Die im Internal Use Modus in Ihrem Nutzerverzeichnis gespeicherten Datensätze stehen Ihnen anschließend auch im Presentation/Publication Modus zur Verfügung.

Die Programme der ausgewählten Ergebnisse, die Sie im Rahmen einer Präsentation oder Publikation veröffentlichen wollen, laden Sie bitte im Presentation/Publication Modus hoch. Stellen Sie bitte zuvor im Internal Use Modus sicher, dass Ihre Do-Files lauffähig und die den Ergebnissen zugrundeliegenden Fallzahlen ausreichend sind (weitere Informationen zu der erforderlichen Anzahl an Beobachtungen finden Sie im weiteren Verlauf des Datenschutzhandbuches). Die im Presentation/Publication Modus hochgeladenen Programme und Ergebnisdateien werden einer Datenschutzprüfung unterzogen.

---

[1] Nähere Informationen zu JoSuA und den unterschiedlichen Modi sind den Vorgaben des FDZ zur Nutzung von Datenfernverarbeitung und Gastaufenthalt zu entnehmen: http://doku.iab.de/fdz/access/Vorgaben_DAFE.PDF

[2] Bei Abschluss des Projektes können die Datenaufbereitungsprogramme über den Presentation/Publication Modus zum Nachweis bei Gutachtenden von beispielsweise Zeitschriften hochgeladen werden.

[3] Die im Internal Use Modus generierten Ergebnisse dürfen nicht abgeschrieben, fotografiert, ausgedruckt oder kopiert werden. Es ist außerdem untersagt, Dritten Einsicht zu gewähren. Jede Verwendung dieser Ergebnisse außer zur Weiterentwicklung Ihrer Programme stellt eine Vertragsverletzung dar.

Im Zuge der Datenschutzprüfung läuft zunächst automatisch ein eigens dafür entwickeltes Prüfskript über die von den Nutzenden produzierten Ergebnisdateien und durchsucht diese bei ausgewählten Stata-Befehlen auf kleine Fallzahlen und führt ggf. erste Löschungen und Zensierungen durch[4]. Anschließend erfolgt eine manuelle Datenschutzprüfung der hochgeladenen Do-Files, Ado-Files und die Ergebnisdateien durch einen wissenschaftlichen Mitarbeitenden des FDZ und es werden ggf. manuell weitere Löschungen und Zensierungen vorgenommen. Nach erfolgreicher Prüfung durch das FDZ werden Ihnen die Programme und Ergebnisdateien zum Download zur Verfügung gestellt. Trotz der Prüfroutinen im FDZ sind die Datennutzenden verpflichtet, die übermittelten Ergebnisse auch nochmal selbst zu prüfen. Bei Zweifeln dürfen die Ergebnisse nicht veröffentlicht werden, sondern müssen dem FDZ angezeigt werden. Zudem verpflichten sich die Datennutzenden im Datennutzungsvertrag auf Geheimhaltung, was unter anderem umfasst, dass keine Rückrechnungen vorgenommen werden, um gelöschte Werte zu ermitteln. Bei Verstößen wird der Nutzer bzw. die Nutzerin bzw. in schweren Fällen auch das gesamte Institut bzw. die gesamte Universität für die Dauer von bis zu zwei Jahren von jeder weiteren Datennutzung ausgeschlossen. Des Weiteren können die im Nutzungsvertrag genannten Strafen zur Anwendung kommen. Informationen über Verstöße gegen die o.g. Verpflichtungen werden zudem an andere Forschungsdatenzentren weitergegeben.

Folgende allgemeine Kriterien müssen bei der Generierung von Auswertungsprogrammen beachtet werden, damit ein Auswertungsprogramm den Richtlinien des FDZ[5] entspricht:

1. Ausführliche Dokumentation der Auswertungsschritte:

Um die Datenschutzprüfung durchführen zu können, muss es dem, in einem täglich rotierenden Verfahren wechselnden, prüfenden FDZ-Mitarbeitenden möglich sein, sich in den Programmen zurecht zu finden, ohne dass Kenntnisse über den Projektinhalt vorhanden sind. Aus diesem Grund ist eine ausreichende Kommentierung der Do-Files notwendig.

Achten Sie dabei darauf,

- Zwischenüberschriften zu verwenden (z. B. "Aufbereitung der Variablen X", "Aggregation auf Betriebsebene"),
- die Inhalte und Funktion von Schleifen zu beschreiben und Tabellenbeschreibungen in Schleifen zu programmieren (z. B. mit dem *display* Befehl),
- für alle wichtigen Variablen, die neu gebildet wurden, Variablenlabels und ggf. Wertelabels zu vergeben,

[4] Zu beachten ist hierbei, dass das automatisierte Programmskript bei bestimmten Stata-Befehlen wie *summmarize* oder *tabulate* auch aggregierte Ausgaben, beispielsweise auf Kreisebene, trotz einer ausreichenden Anzahl an Personen bzw. Betrieben innerhalb der Kreise löscht bzw. zensiert. In solchen Fällen würden wir Sie bitten Rücksprache mit dem FDZ zu halten, um eine solche Löschung von Ergebnissen durch eine angepasste Programmierung zu vermeiden.

[5] Vgl. auch Hochfellner et al. (2012): Datenschutz am Forschungsdatenzentrum. FDZ-Methodenreport 06/2012 und die Vorgaben des FDZ zur Nutzung von Datenfernverarbeitung und Gastaufenthalten (http://doku.iab.de/fdz/access/Vorgaben_DAFE.PDF).

- möglichst sprechende Variablennamen zu verwenden
- selbst generierte Variablen direkt vor Auszählungen kurz zu beschreiben und

  den Inhalt einer jeden Abbildung und komplexer Tabellen zu beschreiben.

2. Syntaxgestützte Programmierung und Speichern von Ergebnisdateien

Zunächst muss ein Master-Do-File angelegt werden, um aus diesem alle weiteren Do-Files, die zu einem Job gehören, auf einmal in der richtigen Reihenfolge zu starten. Zudem sollte in der Master-Datei kurz der Inhalt der einzelnen aufgerufenen Do-Files erläutert werden. Template Do-Files finden Sie auf unserer Homepage unter https://fdz.iab.de/de/FDZ_Data_Access/FDZ_Remote_Data_Access.aspx.

Damit für den prüfenden FDZ-Mitarbeitenden ersichtlich wird, welches Do-File das jeweilige Log-File generiert, muss für jedes Do-File zu Beginn eine entsprechend benannte Ausgabedatei (Log-File) angelegt werden.

Die Auswertungsprogramme (wie auch eingeschickte Ado-Files) dürfen keine Daten oder Fallzahlen enthalten. Externe aggregierte Datensätze müssen separat eingeschickt werden. Beachten Sie hierbei bitte unsere Vorgaben zum Import aggregierter Datensätze in einem späteren Kapitel. Ausgenommen hiervon sind Klassifikationen und Umrechnungsfaktoren (z. B. Beitragsbemessungsgrenzen, Euro-Umrechnung, BIP-Deflatoren, Verbraucherpreisindizes, Umschlüsselung von Wirtschaftszweigen, Kreisen oder Arbeitsagenturbezirken). Bitte kommentieren Sie, welche Klassifikationen und Umrechnungsfaktoren Sie in Ihrer Programmierung verwenden.

3. Prüfung der Ergebnisdatei im Internal Use Modus

Vor dem Hochladen der Do-Files im Presentation/Publication Modus in JoSuA ist im Internal Use Modus eine eigene Kontrolle[6] hinsichtlich der Einhaltung des Datenschutzes durchzuführen. Es soll insbesondere darauf geachtet werden, dass die Fallzahlen der einzelnen Auswertungen ausreichend sind. Nähere Informationen zur erforderlichen Anzahl an Beobachtungen und weitere Hinweise zum Datenschutz sind in den folgenden Kapiteln zu finden.

Im Folgenden sind die wichtigsten Richtlinien des FDZ zur datenschutzrechtlichen Prüfung der Analyseergebnisse dargestellt:

1. Anzahl der Beobachtungen

Sämtliche Ergebnisse müssen aus datenschutzrechtlichen Gründen auf mindestens 20 Beobachtungen basieren. Dies gilt für alle statistischen Kennzahlen, Kreuztabellen und auch multivariate Analysen. In Kreuztabellen muss die Fallzahl jeder einzelnen Zelle > 20 sein, die Gesamtzahl der Tabelle ist hier nicht ausreichend.

Kritisch eingestuft werden Ergebnisse, die auf weniger als 20 Beobachtungen basieren. Wenn also Datenabfragen auf einer zu kleinen Beobachtungszahl basieren, werden vom FDZ Löschungen in der

[6] Wir behalten uns vor, Ergebnisse, die offensichtlich nicht zuerst im Internal Use Modus auf Lauffähigkeit und Einhaltung des Datenschutzes getestet wurden, nicht freizugeben.

Ergebnisdarstellung vorgenommen. Ziel der angewandten Routinen ist die Veränderung der Ergebnisse durch Löschungen, sodass kein Deanonymisierungspotenzial mehr vorhanden ist.

Zunächst müssen alle Werte (Fallzahlen, Verteilungsparameter und Regressionskoeffizienten), die auf einer Beobachtungszahl unter 20 basieren, gelöscht werden. Die Mindestanforderung von 20 Beobachtungen der betrachteten Einheit gilt sowohl für die Betriebs- als auch die Personendaten, die das FDZ anbietet. Um die Möglichkeit des Rückrechnens von gelöschten Werten durch Zwischen- oder Randsummen zu verhindern, müssen ggf. weitere Werte gelöscht oder gerundet werden.

Beispiel 1:

Ostdeutschland

	Betriebsrat		
Anz. svpfl. Besch.	**Ja**	**Nein**	**Total**
1 1-4 SVB	43	1380	1423
2 5-9 SVB	39	547	586
3 10-99 SVB	594	1322	1916
4 100-499 SVB	573	175	748
5 500-999 SVB	142	16	158
Total	1391	3440	4831

Vorher

	Betriebsrat		
Anz. svpfl. Besch.	**Ja**	**Nein**	**Total**
1 1-4 SVB	43	1380	1423
2 5-9 SVB	3*	54*	586
3 10-99 SVB	594	1322	1916
4 100-499 SVB	573	175	748
5 500-999 SVB	14*	/	158
Total	1391	3440	4831

Nachher

In Beispiel 1 wird das Vorgehen des FDZ bei der Datenschutzprüfung von deskriptiven Kreuztabellen kurz erläutert. Zunächst wurden hier Werte < 20 aufgrund der zu geringen Fallzahl durch „/“ ersetzt (Primärsperrung). Damit keine Rückrechnung des gelöschten Wertes mit Hilfe der Randsumme möglich ist, wurden zudem weitere Werte durch ein „*“ einer Sekundärsperrung unterzogen.

Beispiel 2:

Westdeutschland

	Betriebsrat		
Anz. svpfl. Besch.	**Ja**	**Nein**	**Total**
1 1-4 SVB	64	2461	2525
2 5-9 SVB	54	847	901
3 10-99 SVB	859	1985	2844
4 100-499 SVB	793	255	1048
5 500-999 SVB	198	22	220
Total	1968	5570	7538

Vorher

	Betriebsrat		
Anz. svpfl. Besch.	**Ja**	**Nein**	**Total**
1 1-4 SVB	64	2461	2525
2 5-9 SVB	5*	84*	901
3 10-99 SVB	859	1985	2844
4 100-499 SVB	793	255	1048
5 500-999 SVB	19*	2*	220
Total	1968	5570	7538

Nachher

Gesamtdeutschland

Anz. svpfl. Besch.	Betriebsrat Ja	Nein	Total
1 1-4 SVB	107	3841	3948
2 5-9 SVB	93	1394	1487
3 10-99 SVB	1453	3307	4760
4 100-499 SVB	1366	430	1796
5 500-999 SVB	340	38	378
Total	3359	9010	12369

Vorher

Anz. svpfl. Besch.	Betriebsrat Ja	Nein	Total
1 1-4 SVB	107	3841	3948
2 5-9 SVB	93	1394	1487
3 10-99 SVB	1453	3307	4760
4 100-499 SVB	1366	430	1796
5 500-999 SVB	340	38	378
Total	3359	9010	12369

Nachher

Zudem muss ausgeschlossen werden, dass eine Rückrechnung über mehrere Tabellen hinweg möglich ist. Dies wird in Beispiel 2 deutlich. In Westdeutschland ist die Fallzahl in den einzelnen Zellen zwar jeweils >=20, da jedoch neben Ost- und Westdeutschland auch die Tabelle für Gesamtdeutschland ausgewiesen ist, wäre eine Rückrechnung der gelöschten Werte von Ostdeutschland (aus Beispiel 1) möglich. Aus diesem Grund werden auch in der Tabelle von Westdeutschland die entsprechenden Zellen zensiert (Sekundärsperrung: „*“).

Aufgrund des Deanonymisierungsrisikos ist es notwendig, Auszählungen nach allen Subgruppen auszuweisen, sobald die Möglichkeit des Rückrechnens besteht. In obenstehenden Beispiel ist es daher notwendig sowohl die Tabelle nach West- als auch nach Ostdeutschland auszuweisen, da durch die Werte für Gesamtdeutschland auf Ostdeutschland rückgerechnet werden kann. Wollen Sie hingegen beispielsweise nur das Baugewerbe analysieren, ist keine Ausgabe aller Wirtschaftszweige notwendig, da keine Rückrechnung mit Hilfe der Ergebnisse für die Gesamtwirtschaft möglich ist.

Die zugrundeliegenden Fallzahlen sind bei allen statistischen Kennzahlen oder multivariaten Analysen auszuweisen. Bei Fehlen der Fallzahlen werden die betreffenden Ergebnisse gelöscht.

2. Statistische Kennzahlen

Statistische Kennzahlen lassen auf den ersten Blick keinen Rückschluss auf die zugrunde gelegten Fallzahlen zu. Das heißt aber nicht, dass die Ausgabe statistischer Kennzahlen, wie z.B. Mittelwerte, nicht als problematisch zu sehen ist. Auch hier gilt das Prinzip, dass die angesprochenen Kennzahlen nur dann als sicher eingestuft werden, wenn die Berechnungsgrundlage mindestens 20 Beobachtungen umfasst. Als Spezialfall ist die Ausgabe statistischer Kennzahlen bei Dummys zu sehen. Bei binär kodierten Merkmalen verteilen sich die Ausprägungen dieser auf lediglich zwei Kategorien. Auch wenn die Gesamtanzahl der Beobachtungen einer Dummy-Variable größer als 20 ist, kann es sein, dass durch eine schiefe Verteilung nur drei Personen oder Betriebe auf eine der Ausprägungen fallen. In diesem Fall wird die Auswertung als unsicher angesehen, auch wenn nicht direkt über die Gesamtanzahl der Beobachtungen auf eine geringe Besetzung einer Kategorie geschlossen werden kann. Denn dies kann mit Hilfe des Mittelwerts einfach herausgefunden werden. Um diese Fälle identifizieren und prüfen zu können, ist es erforderlich, dass bei Mittelwerten neben der Fallzahl immer auch Minimum, Maximum und Standardabweichung ausgegeben werden.

Beispiel 3:

Variable	Obs	Mean	Std. Dev.	Min	Max
r60	201	2.373134	.9192794	1	3
r61	140	.0857143	.2809469	0	1
r62a	73	2.219178	2.340742	1	15

Vorher

Variable	Obs	Mean	Std. Dev.	Min	Max
r60	201	2.373134	.9192794	1	3
r61	140	/	/	/	/
r62a	73	2.219178	2.340742	1	15

Nachher

Bei der Dummy-Variable *r61* mussten alle Werte zensiert werden, da aufgrund der schiefen Verteilung nur 12 Betriebe (140*0.0857143) auf eine Ausprägung fallen.

Werden Quoten dargestellt, muss die Anzahl der gültigen Werte in den einzelnen Subgruppen ebenfalls ausgewiesen werden. Auf relative Häufigkeiten sollte verzichtet werden, besser ist die Ausgabe von absoluten Zahlen.

3. Perzentile/ Quantile

Bei der Ausgabe von Perzentilen muss darauf geachtet werden, dass mindestens 20 Beobachtungen im jeweiligen Perzentil enthalten sind. Bei einer detaillierten Ausgabe (1% Perzentile) müssen folglich insgesamt mindestens 2000 Beobachtungen in die Ausgabe mit einfließen, um die Datenschutzvorgaben des FDZ zu gewährleisten. Grundsätzlich gilt, je mehr Information man erhalten will, umso mehr Beobachtungen müssen für die gesamte Verteilung vorhanden sein:

- ➔ Mind. 20 Beobachtungen für die Ausgabe von Mittelwerten (Ausnahme Dummys)
- ➔ Mind. 40 Beobachtungen für die Ausgabe von 50%-Perzentilen
- ➔ Mind. 80 Beobachtungen für die Ausgabe von 25%- oder 75%-Perzentilen
- ➔ Mind. 200 Beobachtungen für die Ausgabe von 10%- oder 90%-Perzentilen
- ➔ Mind. 400 Beobachtungen für die Ausgabe von 5%- oder 95%-Perzentilen
- ➔ Mind. 2000 Beobachtungen für die Ausgabe von 1%- oder 99%-Perzentilen

Werden Quantilswerte ausgegeben, so ist immer der geringste Abstand zwischen den gewählten Prozentsätzen (auch zur Null und zu 100) ausschlaggebend – werden z.B. die Perzentile 10 – 15 – 30 ausgegeben, muss die Stichprobe mindestens (15-10)/100 * x ≥ 20 → x ≥ 400 Beobachtungen betragen, um die notwendigen 20 Beobachtungen in jedem Quantil zu gewährleisten.

4. Gewichtung

Bei Verwendung jeglicher Hochrechnungs- oder Gewichtungsfaktoren in deskriptiven Analysen sind auch immer analoge ungewichtete Ergebnisse auszuweisen. Die gewichteten und die dazugehörigen ungewichteten Ergebnistabellen müssen dabei immer direkt aufeinander folgen. Dies muss auch in Schleifen entsprechend berücksichtigt werden. Löschungen in den ungewichteten Tabellen werden in die entsprechenden gewichteten Tabellen übertragen. Fehlt die ungewichtete Ausgabe, wird die gewichtete Tabelle komplett gelöscht.

Bei gewichteten Regressionen ist die Angabe der ungewichteten Fallzahl notwendig, wenn iweights oder fweights verwendet werden.

5. Grafiken

Zu jeder Grafik ist nachzuweisen, auf wie vielen Beobachtungen die einzelnen dargestellten Werte basieren. Dabei ist es nicht ausreichend die Gesamtfallzahl auszuweisen, sondern es muss die Beobachtungszahl jedes einzelnen dargestellten Balkens bzw. Punktes aufgelistet werden. Dies kann über die Angabe der Fallzahlen entweder direkt in den Grafiken oder in Tabellen, die sich unmittelbar vor oder nach der Erstellung der Grafiken befinden, erfolgen. Auch für Grafiken gilt die Fallzahlgrenze von mindestens 20 Beobachtungen. Somit sind z.B. Scatterplots auf Individualebene nicht erlaubt, da hinter den einzelnen ausgegebenen Datenpunkten weniger als 20 Beobachtungen stehen. Während bei Histogrammen (frequency/percentage) die Fallzahlen für jeden einzelnen Balken ausgegeben werden müssen, ist bei Density Plots die Gesamtfallzahl ausreichend. Koeffizienten- bzw. Marginsplots sind unbedenklich, wenn die Fallzahl in der zugehörigen Regressionstabelle ausreichend ist. Zudem ist die Verwendung der asis-Option beim Speichern der Grafik im gph-Format wichtig, um sicherzustellen, dass die Grafik in der derzeitigen Form gespeichert wird und keine nachträgliche Bearbeitung mehr möglich ist.

Beispiel: Ausgabe von Grafiken

```
* Ausgabe von Grafiken im Stata-Format
graph ... , ... saving($log\mygraph.gph, replace asis)
* oder
graph save $log\mygraph.gph, replace asis
* Ausgabe von Grafiken im png-Format (Internal Use Mode)
graph export $log\mygraph.png, replace
```

6. LaTeX-Output

Stata-Ergebnisdateien müssen entweder die Endung „.log“ oder „.txt“ tragen. Ergebnisse, die in separate Dateien außerhalb der von Stata angelegten Log-Files geschrieben wurden, können nicht bereitgestellt werden. Sie müssen wieder in das Log-File mit „type *Pfadangabe*" eingebunden werden. Zur Datenschutzprüfung müssen die Ergebnisse unmittelbar vor dem eingebundenen Text nochmals in lesbarer Form dargestellt werden. Bei deskriptiven Tabellen werden die eingebundenen Tabellen komplett gelöscht, sobald eine Zelle zu geringe Fallzahlen aufweist.

Beispiel: Einbindung von Ergebnissen in das Log-File

```
* Inhalt einer csv- oder txt-Datei im Log-File ausgeben:
[…] // Ausgabe als Tabelle/Liste
type $log/ergebnisse.csv
type $log/ergebnisse.txt
* LaTeX-Code direkt ins Log-File, nicht in separate Datei schreiben:
[…] // Ausgabe der Regressionen
esttab A B, c(b se) tex // Beispiel für Regressionen A und B
```

7. Aggregierte Datensätze

Import aggregierter Datensätze in das Projekt

Externe Kenngrößen auf aggregiertem Niveau (z. B. Arbeitslosenquoten nach Kreisen) dürfen den Daten zugespielt werden, wenn sie den Datenschutzrichtlinien des FDZ entsprechen. Jeder aggregierte Wert muss auf mindestens 3 Beobachtungen beruhen. Das Zusammenspielen von Einzelbeobachtungen auf Personen- oder Betriebsebene ist nicht möglich. Der Datensatz muss Variablen enthalten, die die Anzahl der zugrundeliegenden Beobachtungen pro Variable angeben. Bitte geben Sie ausschließlich absolute Häufigkeiten an (z.B. Anzahl Männer und Anzahl Frauen anstatt dem Frauenanteil). Relative Häufigkeiten lassen eine Überprüfung der Einhaltung des Datenschutzes nicht ohne Weiteres zu. Die Berechnung von Anteilswerten sollte erst in der FDZ-Umgebung erfolgen.

Die entsprechenden Kenngrößen müssen dem FDZ mit einer Beschreibung des Datensatzes (inkl. aussagekräftige Merkmalsbeschreibung, der Aggregationsebene und der Quellenangabe) als Datensatz im Stata Format[7] übermittelt werden (iab.fdz@iab.de). Die Übermittlung aggregierter Kennzahlen innerhalb von Do-Files ist nicht zulässig.

Nach Prüfung und Freigabe durch das FDZ werden die externen Datensätze im Verzeichnis orig zur Verfügung gestellt. Bitte stimmen Sie die Zuspielung von Aggregatdaten frühzeitig mit dem FDZ ab und schicken Sie externe Aggregatdaten mindestens drei Arbeitstage vor Beginn des Gastaufenthalts per E-Mail an das FDZ.

Export/Transfer aggregierter Datensätze aus dem Projekt

Falls Sie aggregierte Datensätze als Ergebnisdatei erhalten möchten, sprechen Sie die Vorgehensweise bitte **vorher**, im Idealfall bereits bei der Antragsstellung, mit uns ab. Neben dem Zweck der Aggregation muss geklärt werden, auf welcher Ebene aggregiert wird und welche Variablen in welchem Aggregatzustand (Summe, Mittelwert, etc.) enthalten sein sollen.

Für die Überprüfung des Datenschutzes muss für jede aggregierte Variable eine zusätzliche Variable mit der Anzahl der zugrundeliegenden gültigen Fälle gebildet werden. Es reicht nicht aus, eine Variable mit der Anzahl der Fälle für den ganzen aggregierten Datensatz zu erzeugen, da die Anzahl der gültigen Fälle je nach Variable stark variieren kann (z.B. wenn eine Variable sehr viele fehlende Werte hat, wie der Umsatz im IAB-Betriebspanel). Ist die Fallzahl in einzelnen Zellen des Datensatzes < 20 müssen diese auf Missing gesetzt werden. Zudem muss sichergestellt sein, dass keine Rückrechnung dieser gelöschten Werte möglich ist. Ist beispielsweise die Gesamtzahl der Beschäftigten, die Anzahl an Frauen und die Anzahl an Männern nach Wirtschaftszweigen ausgewiesen, muss auch der Wert der Anzahl der Männer im entsprechenden Wirtschaftszweig auf Missing gesetzt werden, wenn die Anzahl der Frauen in diesem Wirtschaftszweig < 20 ist.

Die Generierung der aggregierten Daten kann mit den Testdaten vorbereitet werden und sollte während eines Gastaufenthalts erfolgen. Die Programme müssen anschließend nochmals mit JoSuA gestartet werden. Bitte verwenden Sie dafür den Modus *Presentation/Publication* und vermerken Sie

[7] Eingeschickte Datensätze im Excel-Format können nicht zur Verfügung gestellt werden.

im Kommentarfeld, dass Sie einen generierten Datensatz erhalten möchten und den entsprechenden Dateinamen. Speichern Sie diesen Datensatz immer im Unterverzeichnis data.

Für die Dokumentation der aggregierten Datensätze gelten die Vorgaben für aggregierte Variablen entsprechend:

- Es muss je aggregierter Variable eine zusätzliche Variable gebildet werden, welche die Anzahl der entsprechenden gültigen Werte (Personen und Betriebe) im Aggregat enthält (z. B. `bysort Aggregat: egen N_var1 = count(var1)`).
- Ist das gewünschte Aggregat eine Quote, darf lediglich die Anzahl der gültigen Werte in den einzelnen Subgruppen des Aggregats ausgegeben werden (z. B. Anzahl Männer in Bundesland X und Anzahl Frauen in Bundesland X für die Frauenquote in Bundesland X). Die Berechnung der Quoten soll außerhalb der FDZ Umgebung erfolgen.
- Bei Verwendung des Befehls `collapse` muss die Anzahl der nicht fehlenden Werte je Variable berechnet werden:

  `bys persnr: var1 = _n==_N` (setze eine 1 für den letzten Spell einer Person und sonst Null)

  ```
  collapse (count) N_var1 = var1 (count) N_var2 = var2.
  ```

  Es ist nicht ausreichend, mittels `collapse (count) persnr` lediglich die Anzahl der Zeilen zu zählen.

Alle im Anschluss an eine Aggregation erstellten Ergebnisdateien müssen am Anfang in der Kommentierung folgende Informationen enthalten:

- auf welche Ebene aggregiert wurde,
- in welchem Programmschritt aggregiert wurde (bitte auch in der Master-Datei kennzeichnen),
- wie viele Personen und Betriebe mindestens pro Datenzeile eingegangen sind (z. B.: Zellen mit <20 Betrieben wurden gelöscht) und
- Name der Variablen, welche die Beobachtungsanzahl pro Datenzeile enthält.

Fehlen die entsprechenden Hinweise auf die Aggregation zu Beginn der Ergebnisdatei, wird bei der Datenschutzprüfung eventuell zu viel gelöscht. Die Anzahl der aggregierten Tabellen sollte im Sinne der Datensparsamkeit auf ein Minimum beschränkt sein.

Für jedes Projekt kann **nur einmal** ein solcher Datensatz weitergegeben werden. Dies gilt auch, wenn der aggregierte Datensatz in einem anderen FDZ-Projektverzeichnis zur Zuspielung bereitgestellt werden soll. Die Übermittlung des Datensatzes erfolgt per E-Mail.

8. Schätzungen

Regressionsoutput

Die Gesamtzahl der zur Schätzung verwendeten Beobachtungen muss bei Regressionen mindestens 20 betragen. Mit Hilfe von Regressionen lassen sich allerdings auch indirekt kleine Fallzahlen ermitteln. Deshalb reicht es nicht aus, nur die Gesamtanzahl der Beobachtungen zu beachten. Problematisch sind Regressionen, aus deren Koeffizienten sich der unbedingte Mittelwert ergibt, wie beispielsweise bei Regressionen mit einer einzelnen unabhängigen, nicht-metrischen Variable oder einer Regression, die lediglich Dummys, beispielsweise für die Bundesländer, enthält. Ebenso kritisch sind Regressionen in denen lediglich zwei Variablen und deren Interaktionsterme enthalten sind. Das Gleiche gilt für mehr als zwei Variablen in voll interagierten Modellen. Hier muss sichergestellt werden, dass die einzelnen Ausprägungen der dichotomen oder kategorialen Variable mindestens 20 Beobachtungen beinhalten. Dies soll durch eine Auszählung der einzelnen Variable nach dem Regressionsoutput unter Verwendung von *if e(sample)* gezeigt werden. Jede weitere Variable im Modell lässt keine Rückschlüsse mehr zu.

Ereignisdatenanalyse

Bei Anwendung des `sts list`-Befehls im Rahmen der Ereignisdatenanalyse, um sich die Ergebnisse des Kaplan-Meier-Schätzer auszuweisen, ist der Wert zu Beginn der Tabelle (population at risk) für die Datenschutzprüfung relevant. Darüber hinaus muss bei der Ausweisung von Gruppen in Grafiken der `sts list`-Befehl ebenfalls für jede Gruppe erfolgen.

Bei Weitergabe von Ergebnissen aus Survivor-Analysen muss lediglich der Anfangswert >= 20 sein, nicht jeder einzelne Schritt.

Sequenzmusteranalyse

Bei der Sequenzmusteranalyse ist sicherzustellen, dass graphische Darstellungen auf einer ausreichend großen Anzahl von Beobachtungen basieren. Hierfür gibt es zwei zulässige Möglichkeiten. Unbedenklich ist es, Sequenzmuster pro Zeiteinheit als Anteile der Zustände anzugeben (z. B. im Monat 1 sind 30 % der Personen in Arbeitslosigkeit und 70 % in Erwerbstätigkeit). Des Weiteren ist es möglich, aggregierte Sequenzmuster für Gruppen darzustellen, wenn jede Gruppe mindestens 3 Beobachtungen beinhaltet. Sequenzmuster für einzelne Individuen graphisch darzustellen ist nicht zulässig.