SOEP The Socio-Economic Panel

888

SOEP Survey Papers

Series C – Data Documentation

SOEP – The German Socio-Economic Panel at DIW Berlin 2020

Gewichtung der SOEP-CoV-Studie 2020

Rainer Siegers, Hans Walter Steinhauer, Sabine Zinn

DIW SOEP

Running since 1984, the German Socio-Economic Panel (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing.

The SOEP Survey Papers is comprised of the following series:
Series A – Survey Instruments (Erhebungsinstrumente)
Series B – Survey Reports (Methodenberichte)
Series C – Data Documentation (Datendokumentationen)
Series D – Variable Descriptions and Coding
Series E – SOEPmonitors
Series F – SOEP Newsletters
Series G – General Issues and Teaching Materials

The SOEP Survey Papers are available at **http://www.diw.de/soepsurveypapers**

Please cite this paper as follows:

Rainer Siegers, Hans Walter Steinhauer, Sabine Zinn. 2020. Gewichtung der SOEP-CoV-Studie 2020. SOEP Survey Papers 888: Series C. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)

DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin
Germany

soeppapers@diw.de

Gewichtung der SOEP-CoV-Studie 2020

Rainer Siegers, Hans Walter Steinhauer, Sabine Zinn

Zuletzt aktualisiert am 06.08.2020

Zusammenfassung
Dieses Papier stellt das Stichprobendesign, Ergebnisse der Feldarbeit, Analysen zur Selektivität sowie Informationen zur Randanpassung der Studie Sozio-ökonomische Faktoren und Folgen der Verbreitung des Coronavirus in Deutschland vor. Die Stichprobe setzt sich aus einer zufälligen Auswahl von knapp 12.000 Haushalten aus dem Sozio-Oekonomischen Panel (SOEP) zusammen und wurde in neun Tranchen unterteilt, die zeitgleich aufeinander folgen. Beginnend ab 1. April, wurden im Zeitverlauf erste Tranchen größeren Umfangs eingesetzt. Ab Ende Mai wurden die Tranchen bezüglich ihrer Fallzahlen reduziert. Ausgehend von der normalen SOEP-Stichprobe bis hin zu den realisierten Interviews wurden unterschiedliche Auswahlschritte berücksichtigt und für diese im Rahmen der Gewichtung korrigiert.

1 Übersicht

Das Sozio-Oekonomische Panel (SOEP) ist eine Längsschnittstudie am Deutschen Institut für Wirtschaftsforschung (DIW Berlin), die beginnend im Jahr 1984, eine jährliche Befragung von Haushalten und deren Haushaltsmitgliedern in Deutschland durchführt. Somit können anhand von Daten des SOEP Verläufe und Veränderungen durch externe Einflüsse sehr gut beschrieben und analysiert werden. Im Frühjahr 2020 wurden die Haushalte des SOEP neben der regulären, persönlichen Befragung zusätzlich telefonisch (d.h. in einem CATI) zu ihren Erfahrungen durch die Corona-Krise befragt. Weitere Informationen zum Design und der Inhalte der SOEP-CoV-Studie bieten Kühne, Kroh, Liebig und Zinn (2020). Die Ergebnisse und Spotlights der Studie werden unter www.soep-cov.de gesammelt.

Der zuletzt veröffentlichte Scientific Use File (SUF) des SOEP in der Version 35 umfasst die Erhebungsjahre von 1984 bis einschließlich 2018. Die Daten aus dem Erhebungsjahr 2019 liegen der SOEP Abteilung im DIW Berlin ebenfalls vor, sind allerdings noch nicht aufbereitet und veröffentlicht. Aufgrund von Zuzügen zu und Auszügen aus Haushalten, ebenso wie durch Geburten neuer Personen in den Haushalt und das Sterben von Haushaltsmitgliedern verändert sich die Zusammensetzung der Haushalte über die Zeit. Zudem ist es möglich, dass Haushalte oder einzelne Haushaltsmitglieder in einem Erhebungsjahr ihre Teilnahme aussetzen. Aufgrund all dieser Veränderungen in Haushaltsstrukturen, werden für die SOEP-CoV-Studie diejenigen Haushalte ausgewählt, die in den Erhebungsjahren 2018 und 2019 mindestens an einer Erhebung teilgenommen haben und bis zum Feldbeginn 2020 nicht explizit ihre Teilnahme verweigert haben. Von den verbleibenden Haushalten werden darüber hinaus folgende Haushalte ausgeschlossen:

- Haushalte der Geflüchteten-Stichproben M3, M4 und M5. Diese werden im Rahmen einer gesonderten Befragung unter der Verantwortung des Instituts für Arbeitsmarkt und Berufsforschung (IAB) telefonisch zu ihren Erfahrungen in der Corona-Krise befragt.
- Haushalte der Stichproben, die 2019 erstmals befragt wurden (d.h. die Teilstichproben P und Q), um deren Teilnahmebereitschaft an der regulären 2. Welle nicht zu gefährden.
- Haushalte der sogenannten „zentralen Bearbeitung“. Die „zentrale Bearbeitung“ des SOEP Erhebungsinstituts (Kantar Public) kümmert sich um Haushalte, die über die üblichen Kontaktwege des SOEP (nämlich über Interviewer) nicht kontaktiert werden wollen oder können. Die Befragten der „Zentralen Bearbeitung“ werden in der Regel telefonisch kontaktiert und füllen den Fragebogen selbständig oder telefonisch assistiert aus. Somit handelt es sich bei den zentral bearbeiteten Haushalten um Haushalte, die bereits im Rahmen der regulären SOEP-Befragung eine große Neigung der Nichtteilnahme zeigen. Diese Haushalte sollen durch Sonderbefragungen nicht zusätzlich belastet werden.
- Haushalte ohne gültige Telefonnummer, da diese im Rahmen der SOEP-CoV-Studie nicht telefonisch befragt werden können.

Die Stichprobe der verbleibenden Haushalte wurde hinsichtlich ihrer Zusammensetzung und der Kontaktinformationen durch das Erhebungsinstitut des SOEP auf den Zeitpunkt März 2020 aktualisiert und als Bruttostichprobe für die CoV-Studie an das SOEP zurückgespielt. Diese wurde zufällig auf insgesamt neun Tranchen verteilt. Diese werden zeitlich aufeinander folgend befragt. Dabei sind die Tranchen so konstruiert, dass ihre Stichpro-

bengröße über die Zeit abnimmt. Dieser Ansatz hat der Tatsache Rechnung getragen, dass die Menschen in Deutschland in den ersten Wochen des kompletten Lockdowns (und somit während der Feldzeit der ersten vier Tranchen) den größten Herausforderungen und somit Änderungen im alltäglichen Leben gegenüberstanden.

Die ersten vier Tranchen sind die größten mit einem Befragungszeitraum von jeweils zwei Wochen. Die restlichen fünf Tranchen fallen kleiner aus und ihr Befragungszeitraum erstreckt sich auf eine Woche. Einzelne Interviews konnten erst mit einigen Tagen Verzögerung realisiert werden, so dass die tatsächlichen Befragungszeiträume zwischen den Tranchen nicht disjunkt sind. Die Befragungszeiträume und Stichprobenumfänge sind in Tabelle 1 nach den einzelnen Tranchen ausgewiesen. Der Feldstart der SOEP-CoV-Studie

Tabelle 1: Befragungszeiträume und Stichprobenumfänge nach Tranchen.

	Befragungszeitraum		Status der Haushalte in der Stichprobe		
Tranche	Feldstart	Feldende	Eingesetzt	Erreicht	Realisiert
1	01.04.2020	18.04.2020	2.756	2.068	1.689
2	14.04.2020	02.05.2020	3.296	2.450	1.932
3	27.04.2020	16.05.2020	1.767	1.310	978
4	11.05.2020	30.05.2020	1.183	871	632
5	25.05.2020	06.06.2020	608	443	309
6	02.06.2020	13.06.2020	629	450	303
7	08.06.2020	20.06.2020	578	409	288
8	15.06.2020	27.06.2020	598	433	298
9	22.06.2020	04.07.2020	584	405	265
1-9	01.04.2020	04.07.2020	11.999	8.839	6.694

war am 01.04.2020 und der letzte Tag der Befragung von Tranche 9 war am 04.07.2020. In den Tranchen 1 bis 9 wurden 11.999 Haushalte eingesetzt, wovon 8.839 Haushalte telefonisch erreicht werden konnten und schließlich 6.694 an der SOEP-CoV-Studie teilgenommen haben.

Eine grafische Aufbereitung der Stichprobenumfänge nach Status (Kontaktierbarkeit sowie Teilnahmebereitschaft) und Tranchen findet sich in Abbildung 1. Der linke Teil der Abbildung zeigt dabei Verteilung nach Tranche Kontakt- bzw. Teilnahmestatus in absoluten Fallzahlen, der rechte Teil in Anteilen. Der rechten Abbildung ist zu entnehmen, dass der Anteil der teilnehmenden Haushalte im Zeitverlauf leicht, aber stetig, gesunken ist. (Hier liegt die Vermutung nahe, dass das anfänglich große Interesse in der Bevölkerung am Thema „Corona“ über Zeit abgenommen hat). Der Anteil an Haushalten, die nicht erreicht werden konnten, ist über die Tranchen hinweg indes nahezu unverändert.

2 Ablauf der SOEP-CoV-Gewichtung

Die Gewichtung der SOEP-CoV-Studie verlief in weiten Teilen analog zur Gewichtung des SOEP-Core. Diese wird detailliert von Kroh, Siegers und Kühne (2015) beschrieben und ist für die aktuelle Version 35 dokumentiert in Siegers, Belcheva und Silbermann (2020).

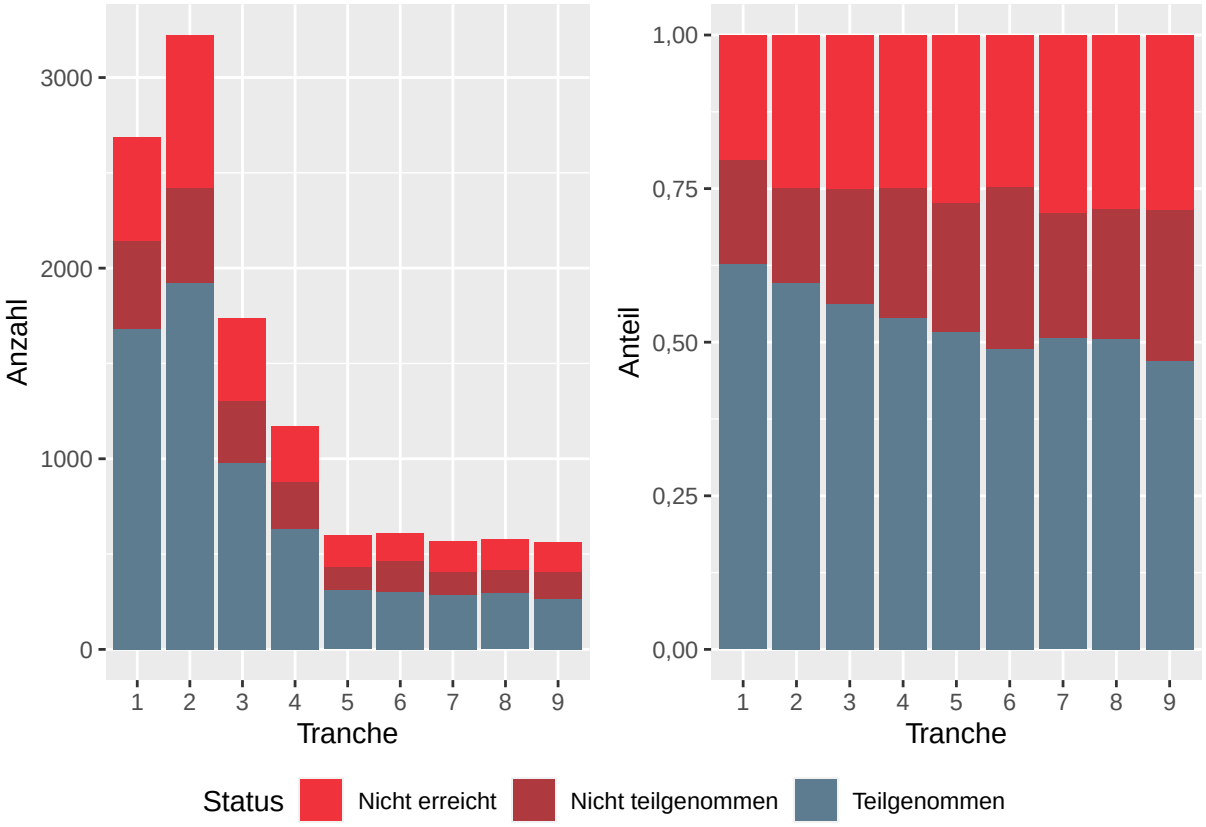

Abbildung 1: Einsatzstichproben nach Tranche und Status.

Als Ausgangsgewicht für die Gewichtung der Haushalte in der SOEP-CoV-Studie diente das Haushaltsgewicht (hhrf) ihrer letzten realisierten Befragung bis 2018, also in der Regel der SOEP-Welle *bi* (aus der SOEP SUF Version v35). Dieses wurde für die Haushalte der SOEP-CoV-Studie für aufeinanderfolgende Ausfallschritte auf Haushaltsebene adjustiert und bezüglich verschiedener Populationsverteilungen, die dem Mikrozensus 2018 entnommen wurden, randangepasst.

Ausgehend von diesen Haushaltsgewichten wurden über einen weiteren Randanpassungsschritt Gewichte für alle Personen in den teilnehmenden Haushalten generiert. Für diejenige Person des Haushalts, die an der CATI-Befragung teilgenommen hat, wurde ein weiterer Gewichtungsschritt durchgeführt, der auftretende Selektionseffekte korrigiert.

Die nachfolgende Abbildung 2 zeigt schematisch den Ablauf der Gewichtung. Konkret wurden in einem ersten Schritt die Ausgangsgewichte für die Veränderungen zwischen der Zusammensetzung des SOEP im Jahr 2018 und 2020 korrigiert. In diesem Zusammenhang wurden die 2018er SOEP Haushaltsgewichte angepasst um Zugänge zu (Zuzug in bestehende Haushalte, Neugeborene) und Abgänge (Verstorbene, Verweigerer) aus der Stichprobe.

Im darauffolgenden Schritt wurde für die Haushalte korrigiert, die von vornherein von der Teilnahme an der SOEP-CoV-Studie ausgeschlossen waren (siehe Abschnitt 1).

Für eine zeitnahe Verwendung der Daten wurde die Stichprobe der SOEP-CoV-Studie nach Abschluss bestimmter Tranchen jeweils gemeinsam gewichtet. Hierbei wurde der tranchenweise Einsatz berücksichtigt und die Haushalte jeweils auf die Grundgesamtheit

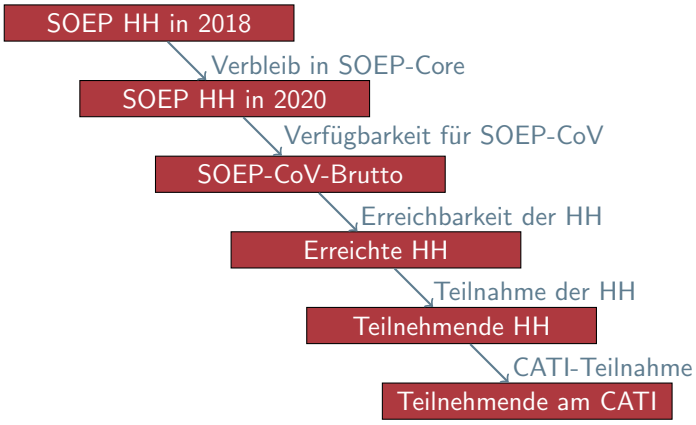

Abbildung 2: Schematischer Ablauf der Gewichtung für die SOEP-CoV-Studie (HH: Haushalte.).

hochgerechnet. Insbesondere der Einsatz der Teilstichproben M1 und M2 (Migrationsstichproben), der erst ab der zweiten Tranche stattfand, fand in diesem Schritt Berücksichtigung.

Um eine möglichst heterogene Zahl von verschiedenen Haushaltsmitgliedern zu erreichen, wurden alle Haushalte zu verschiedenen Tageszeiten von 7 Uhr morgens bis 21 Uhr abends angerufen. Generell wurde auch davon ausgegangen, dass aufgrund der Ausgangsbeschränkungen und des erhöhten Anteils an Personen, die durch die Krise im Home Office arbeiteten, Befragungspersonen telefonisch besser zu erreichen sind als vor der Krise. Die entsprechende Verteilung der Anrufe nach Wochentag, Uhrzeit und Anschluss ist in Abbildung 3 dargestellt. Dennoch verbleiben zwischen 25 und 31 Prozent der Haushalte, die im jeweiligen Befragungszeitraum nicht erreicht werden konnten (vgl. hierzu Abbildung 1 weiter oben). Im dritten Schritt der Gewichtung wurde daher für die Kontaktierbarkeit der Haushalte innerhalb der jeweiligen Befragungszeiträume korrigiert.

Im vierten Schritt wurde schließlich für die Bereitschaft der Haushalte korrigiert, an der SOEP-CoV-Befragung teilzunehmen. Für die SOEP-CoV-Studie konnten innerhalb der einzelnen Tranchen zwischen 69 und 75 Prozent der eingesetzten Haushalte erreicht werden. Über die Tranchen 1 bis 9 hinweg wurden 73 Prozent erreicht. Von den erreichten Haushalten konnten innerhalb der einzelnen Tranchen zwischen 65 und 82 Prozent der Haushalte realisiert werden. Über die Tranchen 1 bis 9 hinweg wurden 72 Prozent realisiert. Somit ergibt sich eine Response Rate nach AAPOR (The American Association for Public Opinion Research, 2016) von RR1 = 0,558. Innerhalb der einzelnen Tranchen schwankt sie zwischen 0,454 und 0,613. Diesem Schritt folgt eine Randanpassung auf eine Vielzahl an Populationsverteilungen, siehe Abschnitt 5, der die Gewichtung auf Haushaltsebene abschließt.

Anschließend wurden auf Basis der Haushaltsgewichte über einen weiteren Randanpassungsschritt Hochrechnungsfaktoren für die einzelnen Haushaltsmitglieder erstellt. Das Verfahren und die hierfür verwendeten Randverteilungen sind im Abschnitt 5 genauer beschrieben.

Auf Basis dieses Personengewichts erzeugen wir in einem letzten Schritt Hochrechnungsfaktoren für die auskunftgebende Person eines teilnehmenden Haushalts. In diesem Schritt wird für die selektive (Selbst-)Auswahl der Auskunftsperson bei Haushalten mit mindes-

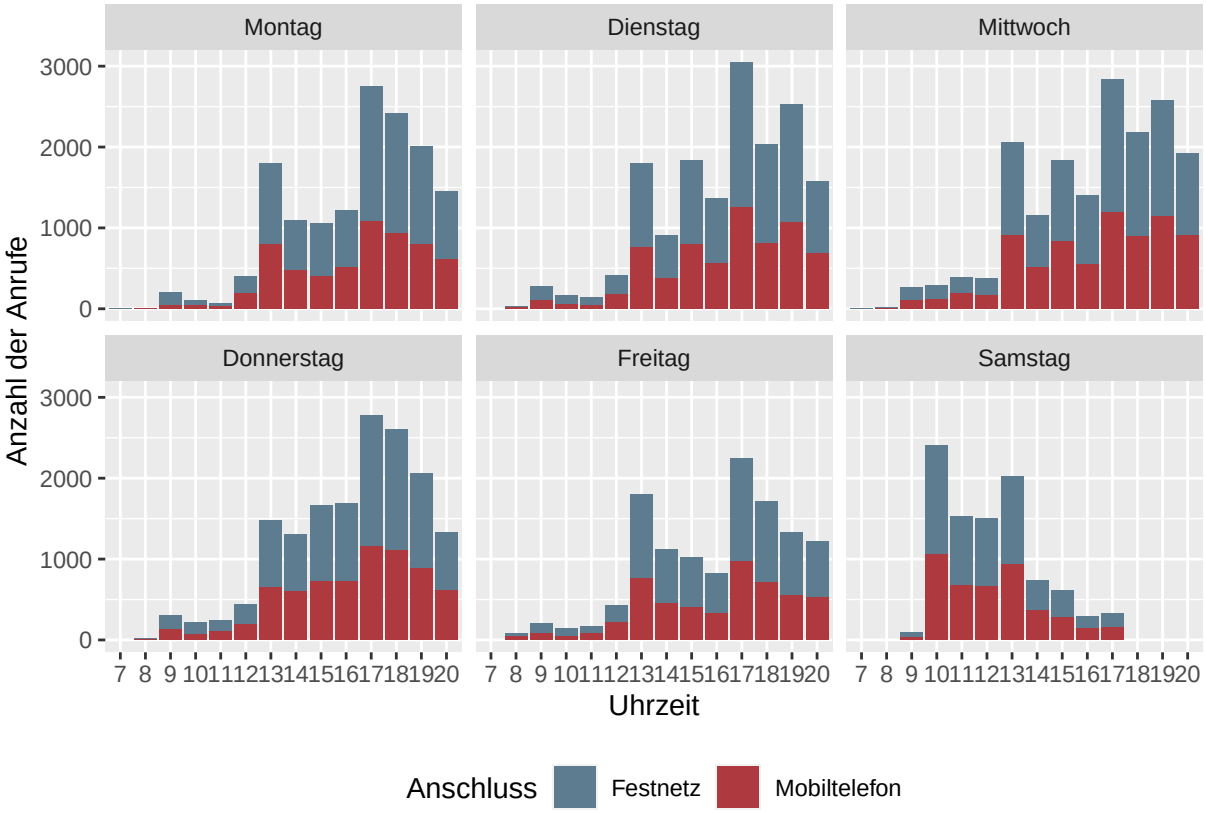

Abbildung 3: Anzahl der Anrufe nach Uhrzeit, Wochentag und Anschluss.

tens zwei Erwachsenen korrigiert.

3 Merkmale für die Gewichtung

In die Ausfallmodelle (cloglog Regressionen) der SOEP-CoV-Gewichtung gingen über 400 Merkmale auf Haushalts- und Personenebene ein. Der Großteil der Merkmale entstammt den vorangegangenen Wellen der *Paneldaten des SOEP*. Insgesamt flossen Variablen aus zahlreichen Befragungsgebieten des SOEP ein wie beispielsweise Demographie, Arbeit, Gesundheit, Bildung, Familie, Finanzen, Persönlichkeit, Migration oder auch politische Einstellung. Zudem wurden in den Ausfallmodellen, soweit sinnvoll und möglich, Personenmerkmale aggregiert auf Haushaltsebene berücksichtigt. Eine Liste mit Merkmalen, die für die Gewichtung des SOEP-Core Version 35 verwendet wurden findet sich in Siegers et al. (2020, S. 63f, 70ff).

Auch Informationen über den Kontaktverlauf gingen in die Gewichtung ein. Von Seiten des Erhebungsinstitutes wurden dem SOEP für insgesamt 86.069 Anrufe die *Kontaktprotokolle der Telefonverläufe* zur Verfügung gestellt. Diese umfassen Informationen zu erfolgreichen und erfolglosen Kontaktversuchen. Zusätzlich enthalten sie Informationen zu Datum und Uhrzeit eines Kontaktversuchs, ob über eine Festnetz- oder Mobiltelefonnummer angerufen wurde und den Rücklaufcode zu dem jeweiligen Kontaktversuch. Aus diesen Informationen haben wir weitere Variablen gebildet, die bspw. angeben, über welchen Telefonanschluss (Festnetz, Mobiltelefon, beide) ein Haushalt kontaktiert wurde oder

wie oft ein Haushalt zu bestimmten Tageszeiten kontaktiert wurde.

Des weiteren wurden die jeweils *tagesaktuellen Corona-Fallzahlen* (Anzahl der Erkrankten, Verstorbenen, Genesenen) auf Kreisebene zum Tag des Kontaktversuchs bzw. Interviews verwendet. Die entsprechenden Daten werden vom Robert Koch-Institut öffentlich zugänglich gemacht.[1] Mit Hilfe der vom Statistischen Bundesamt bereitgestellten Bevölkerungszahlen auf Kreisebene wurde zusätzlich zu den obigen Größen die Corona-Inzidenz auf Kreisebene berechnet.[2] Auch diese Inzidenz war Teil der Gewichtungsvariablen.

Ebenso flossen kleinräumige Informationen unterhalb der Kreisebene, überwiegend zur *Sozialstruktur von Nachbarschaften*, in die Ausfallmodellierung ein. Entsprechende Daten werden von Microm bereitgestellt.

Eine Zusammenfassung der Variablen, die in den verschiedenen Ausfallmodellen auf ihren Einfluss hinsichtlich einer Einschluss in die Stichprobe, Erreichbarkeit oder Teilnahme hin geprüft wurden, findet sich in Tabelle A.1 im Onlinematerial.[3]

Nicht alle Variablen fließen in jedes Ausfallmodell ein. Der Grund hierfür ist offensichtlich: unter den über 400 verfügbaren Merkmalen haben erwartungsgemäß viele keinen Einfluss auf die zu erklärende Variable (d.h. die Einschluss in die Stichprobe, die Kontaktierbarkeit oder die Teilnahme) und/oder sind miteinander hoch korreliert. Nimmt man unnötig viele erklärende Variablen in ein Modell auf, erzeugt dies eine große Streuung in den zu erzeugenden Gewichtungsfaktoren (die sich aus dem Inversen der vorhergesagten Einschluss-, Kontakt- und Teilnahmewahrscheinlichkeiten ergeben). Dies sollte aus Gründen der Stichprobeneffizienz in jedem Fall vermieden werden.

Daher wurden vor jeglicher multivariaten (Ausfall-)Modellierung alle Variablen einzeln auf ihren Zusammenhang mit der zu erklärenden Variable (d.h. Einschluss in die Stichprobe, Kontaktierbarkeit und Teilnahme) geprüft. Nur wenn dieser Zusammenhang signifikant ($p<0.05$) war, wurde die entsprechende Variable in die vorläufige Menge der erklärenden Variablen für das entsprechende Ausfallmodell aufgenommen. Aus Gründen der Modelleffizienz wurden aus der Menge der erklärenden Variablen zudem noch stark korrelierte Merkmale ausgeschlossen. Hierfür wurde die Korrelation aller erklärenden Variablen untereinander bestimmt. Von Merkmalen, die eine betragsmäßige Korrelation von größer als 0,95 aufwiesen, floss nur jenes in das Ausfallmodell ein, das den größten (signifikanten) Einfluss auf die zu erklärende Variable (d.h. die Einschluss in die Stichprobe, die Kontaktierbarkeit oder die Teilnahme) hatte. So ergaben sich für die verschiedenen Ausfallmodelle unterschiedliche Mengen an erklärenden Variablen.

In einem letzten Schritt fand nun noch eine Variablenselektion anhand des bayesianischen Informationskriteriums (BIC) statt. Hierbei wurden dem jeweiligen Modell iterativ Variablen entnommen bzw. wieder hinzugefügt, wenn diese Veränderung im Modell zu einem niedrigeren BIC und somit zu einer besseren Modellgüte führte. Dieses hier beschriebene dreistufige Verfahren zur Variablenselektion fand für jedes der Ausfallmodelle Anwendung, die im Rahmen der SOEP-CoV-Gewichtung geschätzt wurden.

[1] Die jeweils aktuellsten Daten können heruntergeladen werden unter https://opendata.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0.csv.

[2] Die Daten können über GENESIS-ONLINE (https://www-genesis.destatis.de/genesis/online) als Tabelle 12411-0015 heruntergeladen werden. Stand der Daten ist der 31.12.2018.

[3] Das Onlinematerial ist unter https://soep-cov.de/Gewichtung/ verfügbar.

4 Geschätzte Gewichtungsmodelle

Dieser Abschnitt präsentiert die Modelle, die für die oben aufgeführten Gewichtungsschritte geschätzt wurden.[4] Die Ergebnisse werden in Form von Koeffizientenplots präsentiert. Auf der y-Achse sind die Merkmale abgetragen, die als erklärende Variablen in das jeweilige Gewichtungsmodell eingeflossen sind. Parallel zur x-Achse sind die Werte der geschätzten Koeffizienten (roter Punkt) samt ihres 95%-Konfidenzintervalls (rote Balken mit vertikalen Enden) dargestellt. Die gestrichelte, vertikale Linie markiert den Wert 0. Die geschätzten Koeffizienten sind dabei vom kleinsten (oben links) hin zum größten (unten rechts) sortiert. Merkmale, deren Koeffizientenschätzer links der grau gestrichelten Linie liegen, weisen auf einen negativen Einfluss hin. Merkmale, deren Koeffizientenschätzer rechts der grau gestrichelten Linie liegen weisen auf einen positiven Einfluss hin.[5]

4.1 Ausfälle zwischen 2018 und der Bruttostichprobe SOEP-CoV

Abbildung 4 zeigt die geschätzten Koeffizienten und deren Konfidenzintervalle für das Modell mit cloglog-Link, das genutzt wurde, um für die Ausfälle zwischen der 2018er SOEP-Welle *bi* und der Bruttostichprobe an Haushalten im Jahr 2020 zu korrigieren. Wir finden, dass die Nichtteilnahme im Erhebungsjahr 2018 einen deutlich negativen Effekt auf die Bleibewahrscheinlichkeit im SOEP 2020 hat. Weiter beeinflussen der Einsatz von Übersetzungshilfen in den Migrationsstichproben im Rahmen der letzten Erhebung sowie die Zugehörigkeit zur den Migrationsstichproben M1 und M2 die Teilnahmebereitschaft negativ. Haushalte mit sehr jungen Haushaltsmitgliedern weisen ebenso wie Haushalte mit alten Haushaltsvorständen eine deutlich geringere Bleibewahrscheinlichkeit auf.[6] Auch das Nichtvorhandensein eines Internetanschlusses im Haushalt wirkt sich negativ aus auf die Wahrscheinlichkeit im SOEP zu verbleiben. Lebt mindestens eine Person im Haushalt, die angibt besonders heimatverbunden zu sein, findet sich ein negativer Effekt auf die Bleibewahrscheinlichkeit. Das Gleiche gilt für Merkmale, die in Bezug zu fehlenden Werten (konkret: partial unit nonrespone und ein hoher Anteil an item nonresponse auf Haushaltsebene) stehen. Schließlich wirkt sich auch der Umstand, dass das letzte Interview spät in der Feldphase durchgeführt wurde, negativ auf den Verbleib im SOEP aus.

Positiv auf die Bleibewahrscheinlichkeit wirken sich hingegen das Vorhandensein einer Parteipräferenz sowie ein starkes politisches Interesse bei mindestens einem Haushaltsmitglied aus. Ebenfalls positiv wirkt es sich aus, wenn eine der Personen im Haushalt ledig oder mindestens eine Person im Haushalt einen systemrelevanten Job hat. Haushalte, in denen zwei Erwachsene ohne Kinder leben und Haushalte in denen in der letzten Erhebung das Zusatzinstrument für die Mutter-Kind-Befragung ausgefüllt wurde haben eine höhere Wahrscheinlichkeit im SOEP zu verbleiben als Haushalte mit mehr als 2

[4] Zur Schätzung der Modelle wird die `glm` Funktion der freien Statistiksoftware R in der Version: 4.0.2 verwendet (R Core Team, 2020). Zur Aufbereitung der Ergebnisse werden darüber hinaus die Pakete broom (Robinson & Hayes, 2020), gridExtra (Auguie, 2017), kableExtra (Zhu, 2019) und tidyverse (Wickham et al., 2019)) verwendet. Dieses Dokument wurde mit Hilfe von rmarkdown erstellt (Xie, Allaire & Grolemund, 2018).

[5] Generell gilt, dass ein Koeffizientenschätzer, dessen Konfidenzintervall die Null einschließt, keinen signifikanten Einfluss auf die zu erklärende Variable hat. Durch die hier angewandte Methode zur Variablenselektion haben jedoch alle erklärenden Variablen in den finalen Ausfallmodellen der SOEP-CoV-Gewichtung einen signifikanten Einfluss.

[6] Haushaltsvorstand ist diejenige Person, die sich mit den Belangen des Haushalt am besten auskennt bzw. diejenige Person, die den Haushaltsfragebogen bereits beim letzten Interview ausgefüllt hat.

Personen in denen keine Kinder leben und Haushalte, in denen dieses Instrument nicht ausgefüllt wurde. Angehörige der Teilstichproben L3, die zum Zeitpunkt der Ziehung nur die Familientypen Alleinerziehende und Mehrkindfamilien enthielten, haben ebenfalls eine höhere Bleibewahrscheinlichkeit.

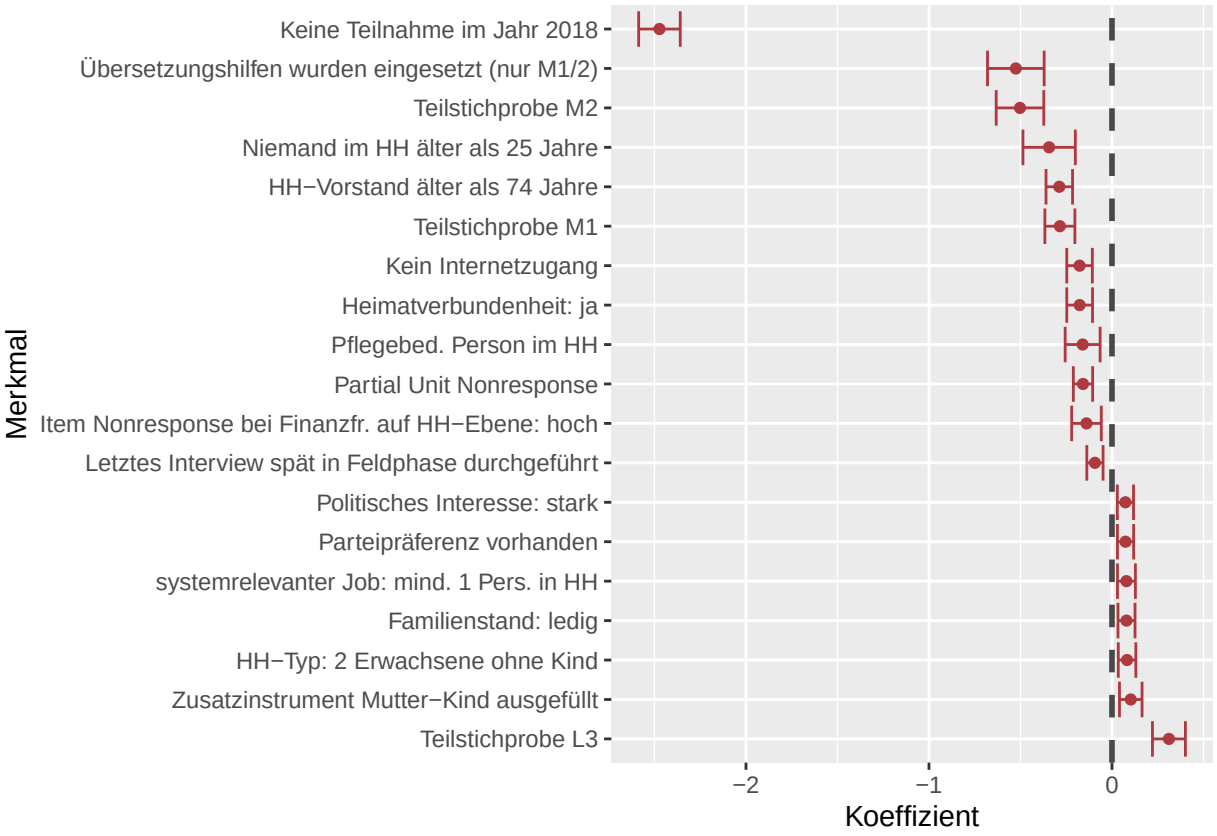

Abbildung 4: Koeffizientenplot des Modells zur Korrektur von Ausfällen zwischen der Befragung 2018 und der SOEP-CoV-Studie. (HH: Haushalt.)

4.2 Tranchenweise eingesetzte Fälle

Für die Befragung im Rahmen der SOEP-CoV-Studie kamen nur Haushalte in Frage, für die eine aktuelle Telefonnummer vorlag und die zuletzt nicht durch die „zentrale Bearbeitung“ des Erhebungsinstituts betreut wurden, siehe Abschnitt 1. Die (potentielle) selektive Verzerrung des Ausgangsbrutto für SOEP-CoV im Vergleich zur SOEP-Stichprobe 2018 wird im folgenden Modell (anhand von Informationen aus der SOEP-Befragung 2018) untersucht und quantifiziert.

Abbildung 5 zeigt die geschätzten Koeffizienten und deren 95%-Konfidenzintervalle für das zugehörige Ausfallmodell mit cloglog-Link. Auch in diesem Fall sind die Merkmale, deren Koeffizientenschätzer links der grau gestrichelten Linie liegen, relativ weniger im Ausgangsbrutto von SOEP-CoV vorhanden als im Gesamt-SOEP. Die Nichtteilnahme an der SOEP-Erhebung im Jahr 2018 ebenso wie Haushalte mit jungen (jünger als 35 Jahre) Haushaltsvorständen, sind relativ weniger im Ausgangsbrutto enthalten. Gleiches gilt für Haushalte, aus denen mindestens eine Person seit 2018 ausgezogen ist und Haushalte in Ostdeutschland (Haushalte in Thüringen und Sachsen-Anhalt und Haushal-

te der Teilstichprobe C, welche das Ausgangs-Sample für Haushalte Ostdeutschland aus dem Jahr 1990 bildet). Ein hohes Niveau an Item Nonresponse auf Haushalts-, wie auch auf Personenebene, führt zu einer geringeren Wahrscheinlichkeit. Auch die Zugehörigkeit zu den Teilstichproben A (Ausgangs-Sample Westdeutschland; 1984) und O (Haushalte in Gebieten der Sozialen Stadt; 2018) führt zu einer niedrigeren Wahrscheinlichkeit im Ausgangsbrutto zu verbleiben. Schließlich sind auch Haushalte mit zwei Erwachsenen ohne Kinder und „andere“ Haushaltszusammensetzungen relativ weniger wahrscheinlich. Schließlich wirkt sich auch die Unzufriedenheit mit dem Familienleben negativ auf den Verbleib im Ausgangsbrutto aus.

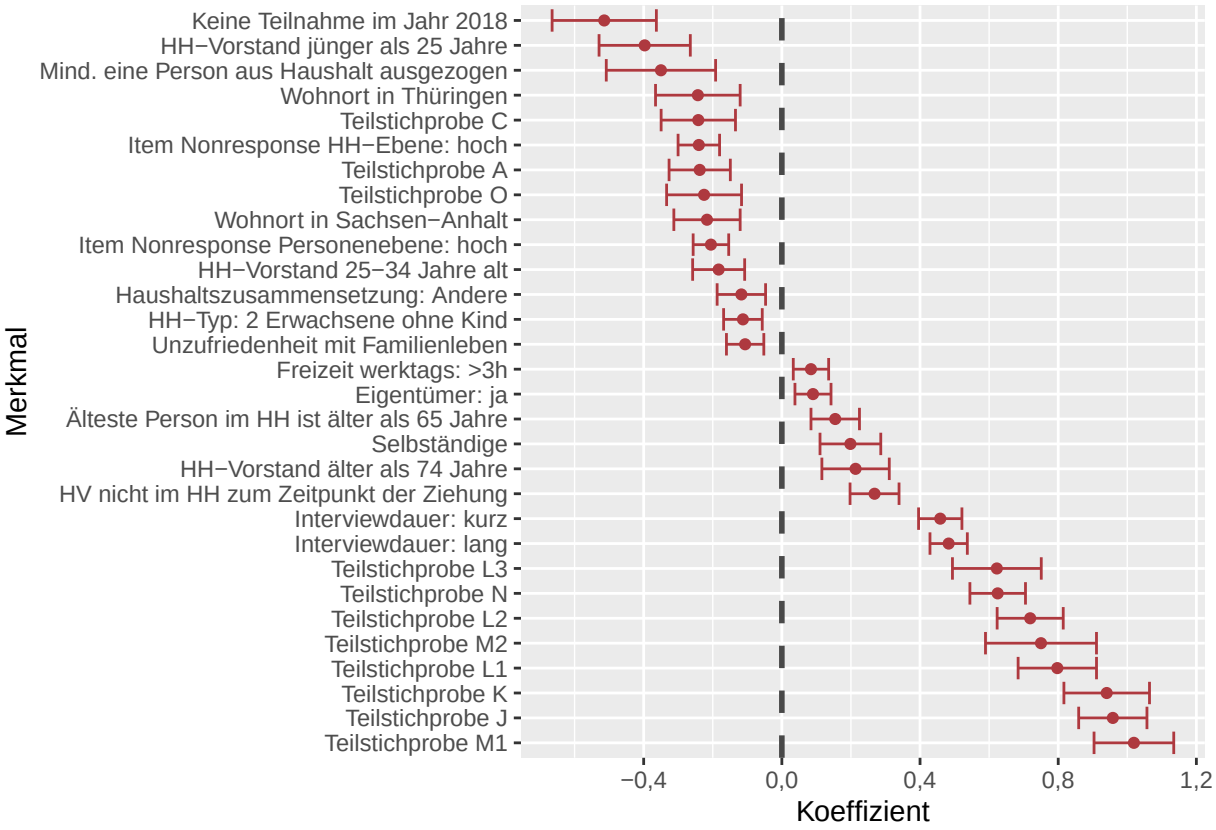

Abbildung 5: Koeffizientenplot des Modells zur Korrektur des designbedingten Verzichts auf Haushalte der „zentralen Bearbeitung“ oder ohne bekannte Telefonnummer. (HH: Haushalt.)

Relativ häufiger hingegen wurden Haushalte eingesetzt, in denen mindestens eine Person mehr als 3 Stunden Freizeit werktags hat, deren ältestes Haushaltsmitglied älter als 65 Jahre ist, in denen mindestens eine Person selbstständig ist, deren Haushaltsvorstand älter als 74 Jahre ist und deren Haushaltsvorstand zum Zeitpunkt der Stichprobenziehung noch nicht im Haushalt lebte. Ebenfalls überproportional im Ausgangsbrutto der Stichprobe enthalten sind Haushalte, bei denen das Interview der letzten Befragung besonders lang (4. Quartil der Verteilung der Befragungsdauer) oder kurz (1. Quartil der Verteilung der Befragungsdauer) gedauert hat. Ebenfalls häufiger im Ausgangsbrutto verblieben sind Haushalte der Teilstichproben J (Aufstockung aus dem Jahr 2011), K (Aufstockung aus dem Jahr 2012), Teilstichproben aus den Jahren 2010 und 2011 mit Fokus auf unterschiedliche Familientypen L1 (Geburtskohorten von 2007 bis 2010), L2 (Niedrigeinkommen, Alleinerziehend, Mehrkindfamilien) und L3 (Alleinerziehend, Mehrkindfamilien). Gleiches

gilt für die Migrationsstichproben M1 aus 2013 und M2 aus dem Jahr 2015, sowie für die Teilstichprobe N (Aufstockung aus dem Jahr 2017).

4.3 Telefonische Erreichbarkeit der Haushalte

Im Gegensatz zur bisherigen Befragung des SOEP, die für gewöhnlich mittels eines persönlichen computergestützten (CAPI) oder schriftlichen (PAPI) Interviews durchgeführt wird, wurde diese Studie als telefonische Umfrage (CATI) durchgeführt. Hierbei waren Haushalte aus unterschiedlichen Gründen nicht erreichbar, bspw. wegen falscher Telefonnummern, Nummern von Firmen- oder Fax-Anschlüssen oder auch weil die Personen des Haushalts zwischenzeitlich verstorben oder ins Ausland verzogen waren. Darüber hinaus wies ein kleiner Teil der Stichprobe einen Sperrvermerk für telefonische Befragungen beim ADM (Verband für Interessensvertretung, Selbstregulierung und Standards in der deutschen Markt- und Sozialforschung, www.adm-ev.de) auf und durfte daher nicht auf telefonischem Wege kontaktiert werden. Andere Haushalte konnten aus sonstigen Gründen während der Befragungszeit der jeweiligen Tranchen nicht erreicht werden.

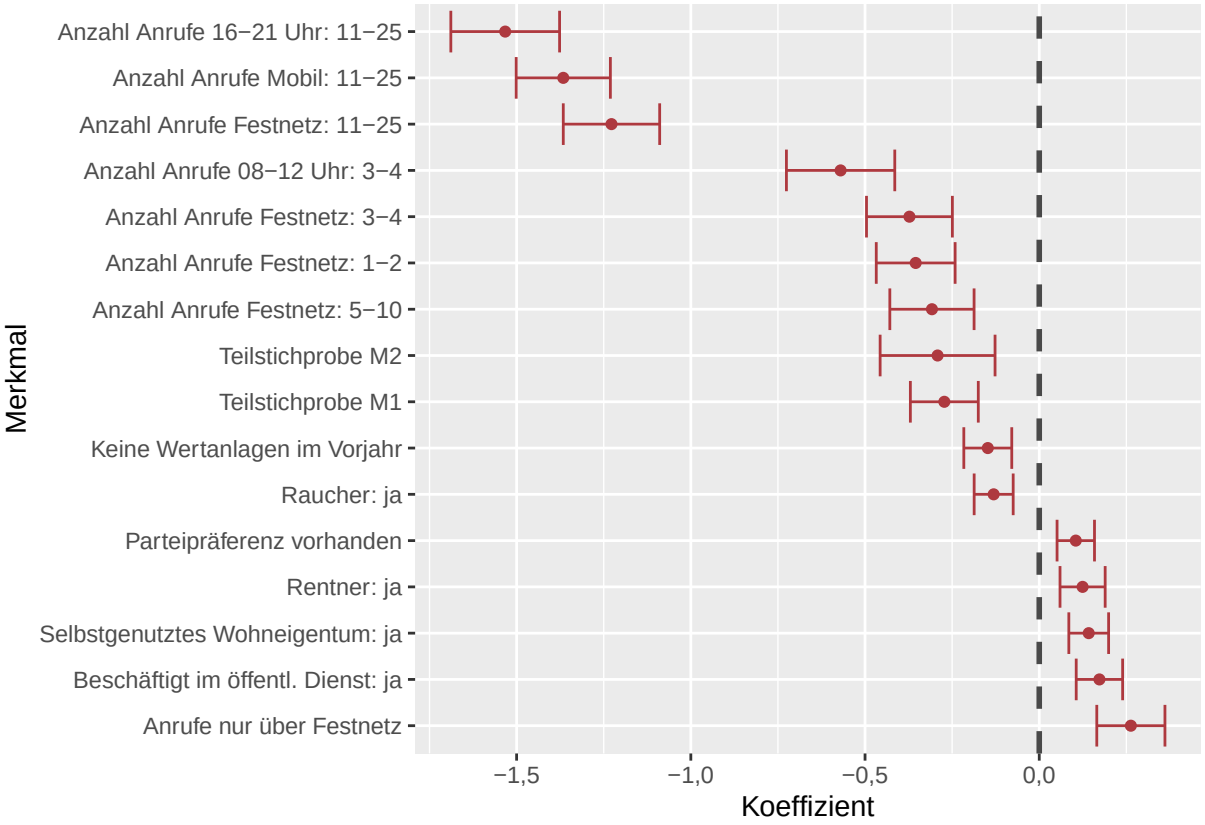

Abbildung 6: Koeffizientenplot des Modells zur Korrektur der Nichterreichbarkeit der Haushalte in der SOEP-CoV-Studie. (HH: Haushalt.)

Abbildung 6 zeigt die geschätzten Koeffizienten und deren Konfidenzintervalle für das Modell mit cloglog-Link, das genutzt wird, um für die Erreichbarkeit der Haushalte zu kontrollieren. Um die Kontaktierbarkeit von Haushalten zu beschreiben, wurden Angaben zu Uhrzeiten und Häufigkeit telefonischer Kontakte genutzt, siehe auch Abbildung 3. Einige Haushalte waren besonders schwer zu erreichen und wurden daher oft (11-25 Anrufe) auf Festnetz und Mobiltelefon angerufen, ebenso wie überwiegend nachmittags bis

abends. Auch Haushalte, die weniger oft über das Festnetz angerufen wurden bzw. zu anderen Zeiten waren zum Teil schwierig zu erreichen. Gleiches gilt für Haushalte der Migrationsstichproben M1 und M2. Haushalte, die im Vorjahr keine Wertanlagen hatten und in den mindestens eine Person raucht weisen ebenfalls eine niedrigere Wahrscheinlichkeit auf erreicht zu werden.

Eine erhöhte Wahrscheinlichkeit Haushalte telefonisch zu erreichen, liegt bei Haushalten vor, die in denen mindestens eine Person eine Parteipräferenz für eine bestimmte Partei hat. Auch Haushalte, in denen mindestens eine verrentnete Person lebt waren leichter zu erreichen. Ebenfalls sind Haushalte die selbstgenutztes Wohneigentum bewohnen und Haushalte mit mindestens einer im öffentlichen Dienst beschäftigten Person leichter zu erreichen. Schließlich weisen auch Haushalte, die ausschließlich über einen Festnetzanschluss kontaktiert wurden, eine höhere Erreichbarkeit auf.

4.4 Teilnahme der Haushalte an der SOEP-CoV-Studie

Die Haushalte, die während der jeweiligen Befragungszeiträume telefonisch erreicht werden konnten, entschieden sich dann schließlich für bzw. gegen die Teilnahme an der SOEP-CoV-Studie. Abbildung 7 zeigt die geschätzten Koeffizienten und deren Konfidenzintervalle für das Modell mit cloglog-Link, das genutzt wurde, um für Verweigerung der Teilnahme an der SOEP-CoV-Studie zu korrigieren.

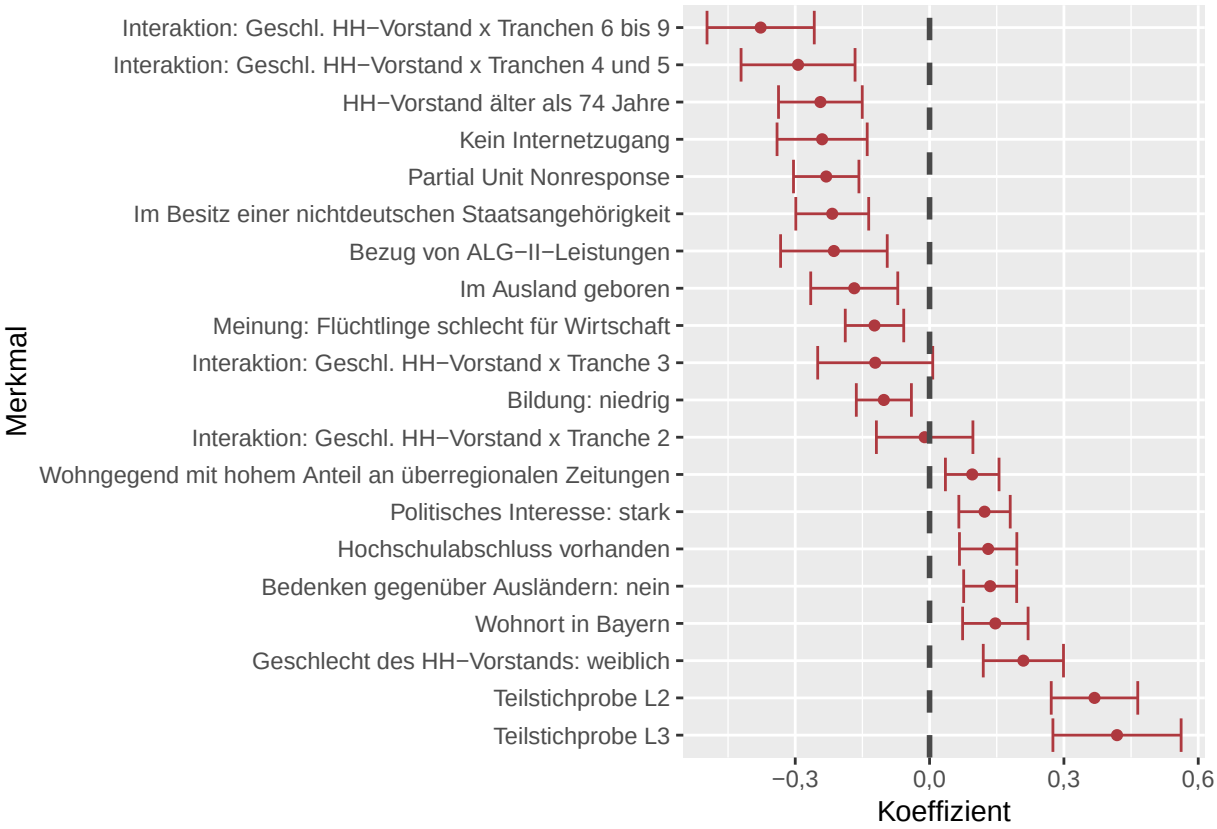

Abbildung 7: Koeffizientenplot des Modells zur Korrektur der Teilnahmeverweigerung von kontaktierten Haushalten. (HH: Haushalt.)

Unter den Faktoren, die die Teilnahmeentscheidung der Haushalte negativ beeinflussten,

sind das Nichtvorhandensein eines Internetanschlusses im Haushalt, teilweise Nichtteilnahme von Befragungspersonen am letzten SOEP-Interview im Haushalt und dass der Haushaltsvorstand älter als 74 Jahre die Prädiktoren mit dem stärksten Einfluss hat. Zudem beobachten wir eine niedrigere Teilnahmewahrscheinlichkeit für Haushalte, mit mindestens einer Person nichtdeutscher Staatsangehörigkeit oder in denen mindestens eine Person der Meinung ist, dass Flüchtlinge schlecht für die Wirtschaft seien. Gleiches gilt für Haushalte in denen mindestens eine Person im Ausland geboren wurde und in denen mindestens eine Person Arbeitslosengeld II bezieht. Schließlich verringert sich die Teilnahmewahrscheinlichkeit, wenn mindestens eine Person im Haushalt keinen Schulabschluss hat.

Positiv hingegen wirkte sich hier aus, wenn der Haushalt in einer Wohngegend mit hohem Anteil an überregionalen Zeitungen wohnt oder mindestens eine Person im Haushalt ein starkes politisches Interesse hat. Auch Haushalte, in denen mindestens eine Person keine Bedenken gegenüber Ausländern hat oder in denen mindestens eine Person einen Hochschulabschluss besitzt, haben eine höhere Teilnahmewahrscheinlichkeit. Ein Wohnort in Bayern wirkt sich ebenfalls positiv auf die Teilnahmewahrscheinlichkeit aus. Besonders teilnahmebereit waren auch Haushalte mit einem weiblichen Haushaltsvorstand. In Interaktion mit der Tranchierung finden sich hingegen negative Effekte, die dadurch zu erklären sind, dass in den späteren Tranchen gezielt nach männlichen Teilnehmern für das Telefoninterview gefragt wurde. Schließlich wirkt sich auch die Zugehörigkeit zu den Teilstichproben L2 (Familientypen: Niedrigeinkommen, Alleinerziehend, Mehrkindfamilien) und L3 (Familientypen: Alleinerziehend, Mehrkindfamilien) positiv auf die Teilnahmeentscheidung aus.

4.5 Kontaktperson beim Telefoninterview

Bei SOEP-CoV wurde je Haushalt nur eine Person befragt, die auch einige Proxy-Informationen über die anderen Haushaltsmitglieder angegeben, aber zu großen Teilen über sich selbst berichtet hat. Die Auswahl der Kontaktperson war dabei nicht systematisch, sondern war davon abhängig wer zur angerufenen Zeit ans Telefon ging und bereit war, an der Befragung teilzunehmen. Generell wurde über den ganzen Tag verteilt angerufen, vermehrt allerdings am späten Nachmittag und abends, um auch berufstätige Personen befragen zu können, siehe auch Abbildung 3. Um eine Verzerrung hinsichtlich des Geschlechts der befragten Person zu verringern, wurde einerseits sowohl nach dem Haushaltsvorstand als auch regelmäßig nach einem männlichen Haushaltsmitglied gefragt. Da für die Teilnahme am CATI der SOEP-CoV-Studie erforderlich war, dass die zu befragende Person zum Zeitpunkt der Befragung mindestens 18 Jahre alt war, gingen auch nur SOEP-Haushaltsmitglieder in die Modellierung ein, die dieses Kriterium erfüllten. Außerdem wurden zur Modellierung nur Personen aus Haushalten berücksichtigt, in denen mindestens zwei volljährige Personen leben, da in erfolgreich kontaktierten 1-Personen- oder Alleinerziehendenhaushalten eindeutig ist, welche Person die Fragen beantwortet.

Abbildung 8 zeigt die geschätzten Koeffizienten und deren Konfidenzintervalle für das Modell mit cloglog-Link, das genutzt wurde, um hinsichtlich Verzerrungen auf Personenebene zu korrigieren. Mit Blick auf die Selektion innerhalb der teilnehmenden Mehrpersonenhaushalte zeigt sich, dass Personen im Alter von 18 bis 24 Jahren seltener an der CATI-Befragung teilnehmen als Personen höheren Alters. Ebenso weisen Personen mit

Abitur und Personen der Altersgruppen „65 bis 69“ und „70 Jahre und älter“ eine niedrigere Teilnahmewahrscheinlichkeit auf als Personen ohne Abitur bzw. Personen im Alter von 25 bis 68 auf. Gleiches gilt für Männer sowie für vollzeiterwerbstätige Personen.

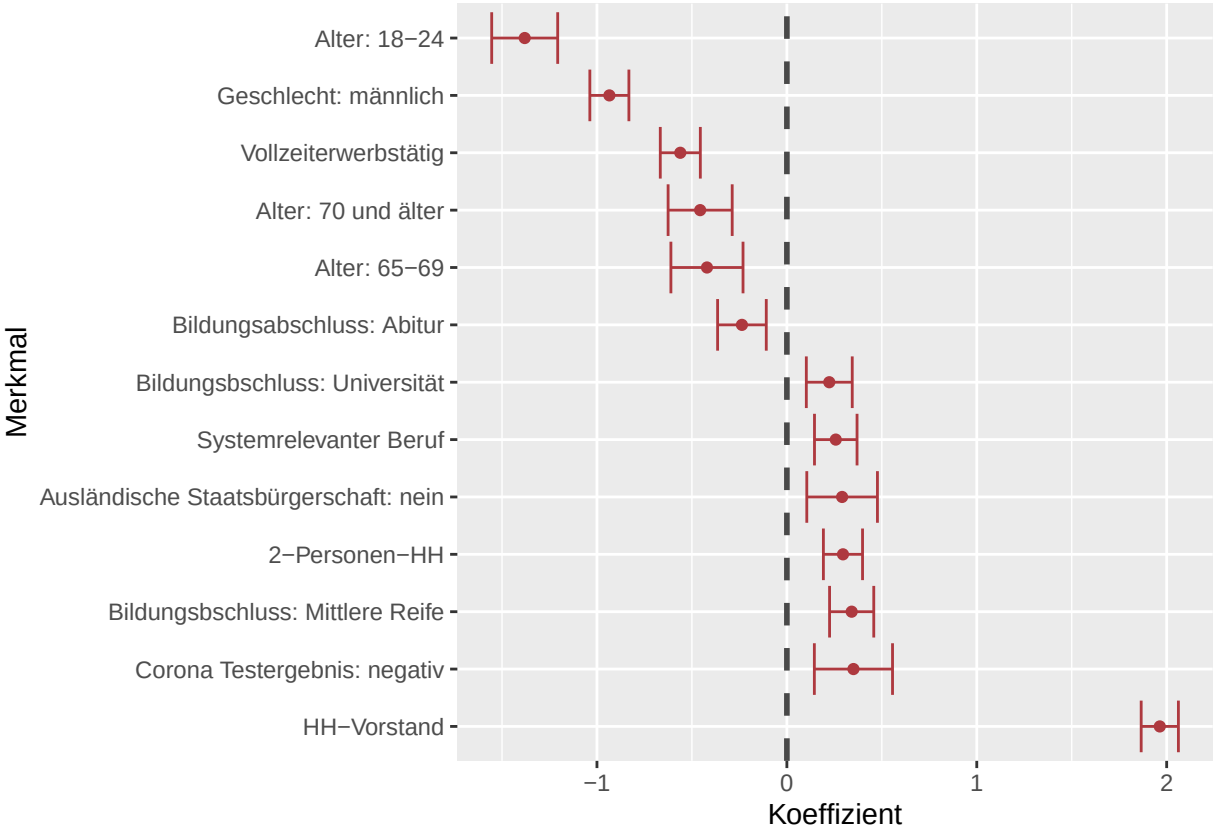

Abbildung 8: Koeffizientenplot des Modells zur Korrektur an der CATI-Teilnahme in der SOEP-CoV-Studie. (HH: Haushalt.)

Hingegen nehmen Personen mit Universitätsabschluss oder systemrelevanten Berufen mit einer höheren Wahrscheinlichkeit am CATI teil. Das gleiche gilt für Personen, die in einem 2-Personen Haushalt leben im Vergleich zu Personen, die in Haushalten mit mehr als 2 Personen leben. Personen mit einer mittleren Reife nehmen ebenso mit einer höheren Wahrscheinlichkeit am CATI teil, wie Personen, die bereits auf Covid-19 getestet wurden und deren Ergebnis negativ ausgefallen ist. Schließlich nimmt übermäßig häufig der Haushaltsvorstand der Befragung von 2018 an der CATI-Befragung teil.

5 Trimmen und Randanpassung

Mit dem Ziel die statistische Effizienz von gewichteten Analysen zu verbessern, wurden die Gewichte getrimmt. Durch das Trimmen der Gewichte wird die Varianz reduziert und somit einer möglichen Verzerrung gewichteter Analysen durch einzelne Beobachtungen mit großen Gewichten entgegengewirkt. Die Gewichte wurden hierbei nicht bei einem bestimmten Wert gekappt, sondern es findet eine Umverteilung der Gewichte nach der „Weight Distribution“ Methode statt (Potter, 1990).

Dieser Methode liegt die parametrische Annahme zugrunde, dass die Gewichte w einer

inversen Beta-Verteilung mit Verteilungsfunktion F_w folgen. Die beiden Parameter der Verteilung werden aus den Gewichten geschätzt und es wird ein Maximalwert τ berechnet, so dass $1 - F_w(\tau) = 0,99$. Gewichte, die diesen Wert τ überschreiten, werden an diesem Maximalwert getrimmt und die überschüssige Masse wird auf die übrigen Gewichte verteilt. Nun wird für die derart getrimmten Gewichte, analog zum obigen Vorgehen, ein neuer Maximalwert $\hat{\tau}$ berechnet. Liegen nun Gewichte vor, die größer sind als $\hat{\tau}$, werden diese am neuen Maximalwert getrimmt und die verbleibende Masse wird wiederum auf alle Gewichte kleiner $\hat{\tau}$ umverteilt. Dieses Verfahren wird iterativ so lange wiederholt, bis keines der getrimmten Gewichte mehr größer ist als der neue Maximalwert oder anders ausgedrückt bis $\tau = \hat{\tau}$. Das Trimmen der Gewichte kam zum einen auf der Haushaltsebene und zum anderen auf Ebene der Personen im CATI-Gewichtungsschritt zur Anwendung.

Um Stichprobenfehler und Undercoverage auszugleichen, werden alle Gewichte in einem letzten Schritt an bekannte Randverteilungen angepasst. Hierzu wurde die in Deville, Särndal und Sautory (1993) beschriebene Raking Prozedur angewandt. Da für das Jahr 2020 noch keine Randverteilungen vom Statistischen Bundesamt bereitgestellt werden können (z.B. durch den entsprechenden Mikrozensus), wurden für die Randanpassungen auf Haushaltsebene und für alle Personen des Haushaltes die letzten vorhandenen Randverteilungen des Mikrozensus' aus dem Jahr 2018 verwendet. Ein dritter zur Verfügung gestellter Gewichtungsfaktor rechnet nur die Kontaktpersonen hoch. Da es sich dabei ausschließlich um erwachsene Personen handelt und uns für diese Population keine Ränder des Mikrozensus vorlagen, wurden die entsprechenden Randverteilungen für Erwachsene auf Basis der SOEP-Daten von 2018 geschätzt.

Auf Haushaltsebene wurden Verteilungen zur Anzahl der Haushalte nach Bundesland, Haushaltsgröße, Gemeindegrößenklasse, selbstbewohntem Eigentum, Haushaltstyp sowie zum letzten Zuzugsjahr eines Haushaltsmitglieds aus dem Ausland zur Randanpassung genutzt. Der entsprechende Randanpassungsschritt erfolgte nach dem Gewichtungsschritt, der Verzerrungen auf Haushaltsebene bei einem realisierten Interview in einem Haushalt ausgleicht, und dem Trimmen der Gewichte. Die Ränder auf Haushaltsebene samt ihrer Ausprägungen und der zugehörigen Häufigkeiten sind in Tabelle A.2 im Onlinematerial aufgeführt.[3]

Auf der Personenebene wurden Verteilungen zur Anzahl der Personen in der Grundgesamtheit nach Alter, Geschlecht, Staatsbürgerschaft (Deutsch vs. andere) zur Randanpassung der Gewichte herangezogen. Diese Randanpassung erfolgte an den Personengewichten, für alle Haushaltsmitglieder in einem realisierten Haushalt. Die Ränder auf Personenebene in realisierten Haushalten samt ihrer Ausprägungen und der zugehörigen Häufigkeiten sind in Tabelle A.3 im Onlinematerial unter aufgeführt.[3] Für die Randanpassung im Anschluss an den CATI-Gewichtungsschritt werden die Ränder zur Anpassung verwendet, wie sie in Tabelle A.4 im Onlinematerial zu finden sind.[3]

6 Zusammenfassung der Gewichte

Tabelle 2 weist für die einzelnen Tranchen die Anzahl der Haushalte und der Personen aus, die an der SOEP-CoV-Studie teilgenommen haben. Da je Haushalt nur eine Person interviewt wurde, ist die Zahl der am CATI teilnehmenden Personen identisch mit der Zahl der Haushalte. Darüber hinaus enthält die Tabelle Angaben dazu, wie viele Haushalte und darin lebende Personen ein Gewicht mit dem Wert 0 aufweisen. Da je Haushalt nur

eine Person am CATI teilnimmt, weisen die CATI-Gewichte für die übrigen Personen im Haushalt ebenfalls den Wert 0 auf. Gewichte mit dem Wert 0 treten auf, da in der Teilstichprobe D (1994/5 Migration (1984-1994, West)) ein Schneeballverfahren zur Anwendung kam. Aufgrund dessen können für bestimmte Haushalte keine Inklusionswahrscheinlichkeiten und somit auch keine Gewichte berechnet werden. Hierbei sind Haushaltsgewichte mit hhrf gekennzeichnet, Gewichte für alle Haushaltsmitglieder mit phrf und die Gewichte von Personen, die im Rahmen der SOEP-CoV-Studie mittels CATI befragt werden konnten, mit phrf_cati.

Tabelle 2: Zusammenfassende Informationen zu den Gewichtungsdaten.

	Anzahl der		Anzahl der Gewichte mit Wert 0		
Tranche	Haushalte	Personen	hhrf	phrf	phrf_cati
1	1.689	4.126	7	14	2.444
2	1.932	4.947	9	21	3.024
3	978	2.443	1	1	1.466
4	632	1.584	1	4	953
5	309	723	0	0	414
6	303	756	3	5	456
7	288	750	1	3	463
8	298	722	5	11	429
9	265	665	0	0	400
1-9	6.694	16.716	27	59	10.049

Die nachfolgende Tabelle 3 zeigt die Verteilung der verschiedenen Gewichte (phrf, phrf und phrf_cati) für die in Tabelle 2 berichteten Fallzahlen. Bei der Berechnung der entsprechenden Statistiken wurden Gewichte mit dem Wert 0 ausgeschlossen.

Tabelle 3: Verteilung der verschiedenen Gewichte nach Tranche.

Tranche	Minimum	Median	Mittelwert	Maximum	Standardabw.	Summe
			Gewicht: hhrf			
1	48	3.697	6.279	62.921	7.595	10.562.046
2	8	3.193	5.473	59.144	6.563	10.524.192
3	35	3.931	6.371	62.995	7.741	6.224.776
4	80	3.688	6.537	58.421	8.154	4.125.110
5	131	3.713	6.894	56.348	8.812	2.130.310
6	49	3.521	6.098	38.746	7.227	1.829.350
7	18	3.630	6.745	49.683	8.130	1.935.906
8	20	4.436	7.372	51.321	8.691	2.159.963
9	77	3.617	7.118	65.067	9.037	1.886.347
1-9	8	3.581	6.206	65.067	7.592	41.378.000
			Gewicht: phrf			
1	43	2.692	4.956	75.018	6.798	20.378.307
2	6	2.449	4.250	77.311	5.579	20.936.930
3	29	2.900	5.165	54.870	6.904	12.613.619
4	74	2.667	5.237	76.366	7.663	8.274.771
5	107	2.916	5.655	57.986	7.883	4.088.392
6	46	2.539	4.722	49.384	6.330	3.545.887
7	17	2.770	5.534	64.162	7.644	4.133.597
8	16	3.157	5.797	60.224	7.634	4.121.793
9	63	2.571	5.293	57.744	7.532	3.519.703
1-9	6	2.648	4.900	77.311	6.727	81.613.000
			Gewicht: phrf_cati			
1	60	5.674	10.254	92.106	12.478	17.246.453
2	3	5.463	9.574	98.090	11.601	18.410.223
3	49	6.263	10.790	92.106	12.925	10.542.234
4	159	6.131	10.754	80.088	12.865	6.785.625
5	206	6.118	11.460	98.090	13.668	3.540.989
6	48	5.954	10.487	65.970	12.309	3.146.210
7	26	6.186	11.244	92.106	13.966	3.227.107
8	38	7.091	11.622	72.292	13.117	3.405.221
9	127	6.909	12.082	72.292	13.812	3.201.754
1-9	3	5.862	10.425	98.090	12.552	69.505.815

7 Ableiten eigener Gewichtungsfaktoren

Mit den SOEP-CoV-Daten ist eine Vielzahl von Analysen an unterschiedlichsten Analysemengen möglich. Für jede potentielle Analysemenge eigene Gewichte zur Verfügung zu stellen, übersteigt den Rahmen des Machbaren. Dennoch sollen und müssen die zur Verfügung gestellten Gewichte der gesamten SOEP-Cov Stichprobe für statistische Auswertungen, die auf Populationsaussagen abzielen, genutzt werden; wenn auch nur um zu prüfen, ob die Gewichte relevant für die Berechnung von Populationsstatistiken sind (z.B. durch den simplen Vergleich von gewichteten und ungewichteten Statistiken). Die SOEP-CoV-Gewichte wurden für die gesamte Stichprobe (der neun SOEP-CoV-Tranchen) an Haushalten bzw. Personen, die an der CATI-Befragung teilgenommen haben, erzeugt. Somit stellen sie Hochrechnungsfaktoren für genau diese Stichprobe bzw. für eine Zufallsauswahl aus dieser Stichprobe dar. Das bedeutet, dass für jede Analysemenge, die diese Voraussetzung nicht erfüllt, Adjustierungsfaktoren berechnet werden müssen, damit Hochrechnungen auf die Grundgesamtheit der SOEP-CoV-Stichprobe möglich sind.

- Um in einem ersten Schritt zu prüfen, ob die SOEP-CoV-Gewichte für eine Teilstichprobe der SOEP-CoV-Stichprobe verwendet werden können und — falls dies nicht ohne weiteres möglich ist — entsprechende Adjustierungsfaktoren abzuleiten, muss eine Selektivitätsanalyse durchgeführt werden:
- Hierbei müssen mindestens alle Variablen, die in die geplante Analyse aufgenommen werden sollen, als erklärende Variablen in ein logistisches Regressionsmodell (oder eine probit oder cloglog Regression) einfließen.
- Die abhängige Variable dieses Selektionsmodells ist ein Indikator (kodiert auf 0 und 1), der angibt ob im Vergleich zur gesamten SOEP-CoV-Stichprobe eine Datenzeile Teil der Analysemenge ist (y = 1) oder nicht (y = 0).
- Das Selektionsmodell umfasst somit genauso viele Datenzeilen wie es in SOEP-CoV Beobachtungen gibt.
- Zeigt nun keine der Analysevariablen einen signifikanten (d.h. $p < 0{,}05$) und gleichzeitig bedeutungsvollen Effekt (d.h. $\beta > 0,01$) hinsichtlich der Zuordnung zur Analysemenge, ist die betrachtete Teilstichprobe eine im Hinblick auf die Analysevariablen zufällige Auswahl aus der gesamten SOEP-CoV-Stichprobe. Die originalen SOEP-CoV-Gewichte können zur Hochrechnung dieser Teilstichprobe auf die Grundgesamtheit genutzt werden. Hierbei gilt zu beachten, dass gewichtete Angaben dann in Summe natürlich nicht die gesamte Populationsgröße ergeben, sondern eben nur auf die Teilpopulation, auf die sich die Analyse bezieht.
- Ergibt die Selektivitätsanalyse allerdings Verzerrungen der Teilstichprobe hinsichtlich der Analysevariablen (d.h. gibt es signifikante und bedeutungsvolle Effekte in der Selektivitätsanalyse), ist eine Korrektur der SOEP-CoV-Gewichte erforderlich, bevor sie zu Hochrechnungszwecken herangezogen werden können. Diese Korrektur der SOEP-CoV-Gewichte erfolgt über die Multiplikation mit einem Adjustierungsfaktor, der sich wiederum aus der durchgeführten Selektivitätsanalyse ergibt.
- Konkret heißt das: Alle Analysevariablen, die sich als signifikant und gleichzeitig bedeutungsvoll herausgestellt haben, fließen in eine neue Selektivitätsanalyse ein. Analysevariablen, die in der zuvor berechneten Selektivitätsanalyse nicht signifikant und/oder bedeutungsvoll waren, werden hierbei außer Acht gelassen (um eine unnötige Varianzerhöhung in den zu erzeugenden Adjustierungsfaktoren zu vermeiden). Die abhängige Variable der neuen Selektivitätsanalyse ist identisch mit der der zuvor berechneten, auch die Stichprobengröße bleibt unverändert.

- Auf Basis der geschätzten (neuen) Selektivitätsanalyse müssen nun für jede Datenzeile Wahrscheinlichkeiten geschätzt (bzw. vorhergesagt) werden der Analysemenge anzugehören. Das kann in Stata mit dem Befehl `predict pr` getan werden und in R mit dem Befehl `predict()` unter Berücksichtigung des Arguments `type = "response"`. Nun werden der Analysemenge die vorhergesagten Wahrscheinlichkeiten für eine Zugehörigkeit zur originalen SOEP-CoV-Stichprobe zugespielt. Die Inverse dieser Wahrscheinlichkeiten gibt den Adjustierungsfaktor an, der mit den SOEP-CoV-Gewichten zu multiplizieren ist, um für Verzerrungen im Vergleich zur gewichteten Ausgangsstichprobe der SOEP-CoV-Studie zu korrigieren. Mit anderen Worten, durch die Multiplikation der SOEP-CoV-Gewichte, die zur Analysemenge gehören, mit der inversen vorhergesagten Wahrscheinlichkeit ergibt sich das gesuchte adjustierte Gewicht, das zur Berechnung von Populationsstatistiken hergezogen werden kann.
- *Anmerkung*: Es ist in jedem Fall angeraten, zu überprüfen wie gut das berechnete Selektionsmodell zwischen Zugehörigkeit und Nicht-Zugehörigkeit zur Analysemenge diskriminieren kann, z.B. durch die Nutzung entsprechender Boxplots: ein Boxplot gibt die Verteilung der (vorhergesagten) Wahrscheinlichkeiten für die Analysemenge an und ein Box-Plot zeigt die (vorhergesagten) Wahrscheinlichkeiten für den Teil der SOEP-CoV-Stichprobe, der nicht Teil der Analysemenge ist. Generell sollte der erste Boxplot eine Verteilung nahe der 1 anzeigen, der zweite eine Verteilung nahe der 0 und die Inter-Quartile-Ranges beider Boxplots sollten möglich wenig Überschneidungen in ihrem Wertebereich aufweisen. Ist dies nicht der Fall, diskriminiert das verwendete Modell nicht gut und die Hinzunahme weiterer erklärender Variablen, die den Selektionsmechanismus (besser) beschreiben, der die Analysemenge erzeugt hat, ist sinnvoll.

Literatur

Auguie, B. (2017). gridextra: Miscellaneous functions for "grid"graphics [Software-Handbuch]. Zugriff auf https://CRAN.R-project.org/package=gridExtra (R package version 2.3)

Deville, J.-C., Särndal, C.-E. & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, *88* (423), 1013–1020. doi: 10.1080/01621459.1993.10476369

Kroh, M., Siegers, R. & Kühne, S. (2015). Gewichtung und Integration von Auffrischungsstichproben am Beispiel des Sozio-oekonomischen Panels (SOEP). In J. Schupp & C. Wolf (Hrsg.), *Nonresponse bias: Qualitätssicherung sozialwissenschaftlicher umfragen* (S. 409–444). Wiesbaden: Springer Fachmedien Wiesbaden. Zugriff auf https://doi.org/10.1007/978-3-658-10459-7_13 doi: 10.1007/978-3-658-10459-7_13

Kühne, S., Kroh, M., Liebig, S. & Zinn, S. (2020, Jun.). The Need for Household Panel Surveys in Times of Crisis: The Case of SOEP-CoV. *Survey Research Methods*, *14* (2), 195-203. Zugriff auf https://ojs.ub.uni-konstanz.de/srm/article/view/7748 doi: 10.18148/srm/2020.v14i2.7748

Potter, F. J. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the american statistical association, section on survey research methods* (S. 225-230). Zugriff auf http://www.asasrms.org/Proceedings/papers/1990_034.pdf

R Core Team. (2020). R: A language and environment for statistical computing [Software-Handbuch]. Vienna, Austria. Zugriff auf https://www.R-project.org/

Robinson, D. & Hayes, A. (2020). broom: Convert statistical analysis objects into tidy tibbles [Software-Handbuch]. Zugriff auf https://CRAN.R-project.org/package=broom (R package version 0.5.6)

Siegers, R., Belcheva, V. & Silbermann, T. (2020). *SOEP-Core v35 Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP) (1984 until 2018)* (SOEP Survey Papers Nr. 826). Berlin: DIW/SOEP. Zugriff auf https://www.diw.de/documents/publikationen/73/diw_01.c.745900.de/diw_ssp0826.pdf

The American Association for Public Opinion Research. (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9. Aufl.). AAPOR.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4* (43), 1686. doi: 10.21105/joss.01686

Xie, Y., Allaire, J. & Grolemund, G. (2018). *R markdown: The definitive guide*. Boca Raton, Florida: Chapman and Hall/CRC. Zugriff auf https://bookdown.org/yihui/rmarkdown (ISBN 9781138359338)

Zhu, H. (2019). kableextra: Construct complex table with 'kable' and pipe syntax [Software-Handbuch]. Zugriff auf https://CRAN.R-project.org/package=kableExtra (R package version 1.1.0)