Explorative Untersuchung von Paneldaten für sozial- und arbeitswissenschaftliche Fragestellungen

–

Ein Vergleich klassischer und fortgeschrittener Machine Learning Methoden

Master-Thesis

Autor: Michael Schmid
Haggenstrasse 86, CH – 9014 St. Gallen
+41 76 443 58 72
michael.schmid@stud.hslu.ch

Referent: Prof. Adrian Stämpfli
Institut für Modellbildung und Simulation
Rosenbergstrasse 59, CH – 9000 St. Gallen
+41 58 257 12 14
adrian.staempfli@ost.ch

Korreferent: Prof. Dr. Harold Tiemessen
Institut für Modellbildung und Simulation
Rosenbergstrasse 59, CH – 9000 St. Gallen
+41 58 257 12 31
harold.tiemessen@ost.ch

Michael Schmid (ISD Nr. L1485026), Frühlingssemester 2021

Hochschule Luzern, Departement Wirtschaft

Master of Science in Applied Information and Data Science

St. Gallen, 03. Juni 2021

Management Summary

In aktuellen Entwicklungen rund um die Früherkennung und Prävention psychosozialer Risiken, gewinnt die Interdependenz zwischen strukturellen Arbeitsbelastungen und subjektiv empfundenen Arbeitsbeanspruchungen zunehmend an Bedeutung. Dabei stellt die Untersuchung von Wechselwirkungen zwischen arbeitsorganisatorischen, individuellen, biologischen und soziokulturellen Bedingungen ein zentrales Forschungsobjekt im SNF-Forschungsprojekt "Psychosoziale Risiken in der Arbeitswelt" der OST (Ostschweizer Fachhochschule) dar.

Ein generischer Zugang zur Untersuchung solcher Interdependenzen besteht in der Analyse und Modellierung von Paneldaten. Anhand des Swiss Household Panels, einer longitudinalen, national repräsentativen Erhebung zur sozialen Entwicklung in der Schweiz, können individuelle zeitliche Verläufe von spezifischen Lebenslagen, Arbeitsbelastungen und Arbeitsbeanspruchungen auf mögliche Interdependenzen untersucht werden.

Die Wahl der richtigen Methode zur Analyse von Paneldaten ist speziell bei sozialwissenschaftlichen Untersuchungen wie dieser nicht offensichtlich. Die gängige Literatur liefert diverse Methoden aus der klassischen Ökonometrie zur Analyse von Paneldaten. Durch die zunehmende Verfügbarkeit fortgeschrittener Machine Learning Methoden aus der Open Source Gemeinschaft, existieren neue Potenziale zur Ergänzung oder gar Verbesserung des Erkenntnisgewinns bei der Paneldatenanalyse sozialwissenschaftlicher Fragestellungen.

Im Rahmen dieser Master-Thesis werden fünf State-of-the-Art Methoden[1] der Paneldatenanalyse mit fünf fortgeschrittenen Machine Learning Methoden [2] verglichen, um folgende Forschungsfrage zu beantworten: "*Kann der empirische Erkenntnisgewinn von State-of-the-Art-Methoden der longitudinalen Paneldatenanalyse sowohl in der Tiefe[3] als auch in der Breite[4] durch fortgeschrittene Machine Learning Methoden ergänzt werden*?"

Die Forschungsfrage kann mit "Ja" beantwortet werden. Fortgeschrittene Methoden bieten eine sinnvolle Ergänzung von State-of-the-Art-Methoden, da sie neue Blickwinkel auf Paneldaten erlauben und somit zusätzliche Erkenntnisse liefern oder das Anwendungsspektrum klassischer Methoden erweitern. Darüber hinaus wird festgestellt, dass fortgeschrittene Methoden die State-of-the-Art-Methoden nicht ersetzen können.

Im Zentrum dieser Master-Thesis steht die Feststellung, dass *die richtige M*ethode zur Analyse von Paneldaten aktuell nicht existiert. Vielmehr stellt die gegenseitige Ergänzung mehrerer Methoden einen vielseitigen Erkenntnisprozess dar, den es zu verstehen, beschreiben und weiter zu operationalisieren gilt. Das A und O eines solchen Erkenntnisprozesses besteht in der iterativen Vereinigung empirischer Erkenntnisse mit theoriegeleiteten Ansätzen des klassischen Rationalismus.

Der daraus entstehende Handlungsbedarf richtet sich primär and die sozialwissenschaftliche und mathematisch/methodologische Community. Die zukünftige Berücksichtigung und Würdigung des Erkenntnisprozesses bei der Paneldatenanalyse, ist aus derzeitiger Sicht wünschenswert und sollte für zukünftige Entwicklungen in der Sozialwissenschaft und Mathematik/Statistik Beachtung finden.

---

[1] Gepooltes Modell, variable Koeffizienten, Fixed Effects, First Differences und Random Effects.

[2] Random Effects mit Kontextvariablen, Longitudinales Clustering, Instrumentelle Variablen, Gemeinsame Mittelwert-Kovarianz Modellierung, baumbasierte variable Koeffizienten Regression

[3] Erschliessung von präziseren Erkenntnissen aufgrund verbesserter, umfassenderer Methoden

[4] Erschliessung einer neuen Art von Erkenntnissen durch die Einnahme neuer Blickwinkel

Vorwort

"Sozialwissenschaft, wie wir sie heute kennen, ist vor allem durch ihre starke empirische Ausrichtung gekennzeichnet" (Wolf & Best, 2011, S.3). So stellte René König, einer der bekanntesten deutschen Soziologen, in den ersten Nachkriegsjahrzenten fest, dass Wissenschaft, und damit auch die Soziologie, "letztlich nur als empirische Forschung möglich" ist (König, 1962). "Mittlerweile hat sich diese Einsicht durchgesetzt und die quantitativ-empirische Forschung ist zum Standard in den Sozialwissenschaften geworden" (Wolf & Best, 2011).

"Während sich Probleme der klassischen Soziologie zuvorderst auf Unterschiede zwischen Personen (*Individuen*) beziehen, versucht die moderne Soziologie zusätzlich, Auswirkungen von intraindividuellen Differenzen bzw. von Ereignissen (Heirat, Geburt eines Kindes, Scheidung, Arbeitslosigkeit etc.) innerhalb individueller Lebensläufe zu bestimmen". (Giesselmann & Windzio, 2012, S.10)

Durch die Betrachtung individueller Lebensläufe gewinnen Paneldaten (allgemein auch als Längsschnittdaten bezeichnet) vermehrt an Bedeutung, da mit deren Hilfe Längsschnittfragestellungen effizient und sinnvoll untersucht werden können. Giesselmann und Windzio (2012) definieren ein Panel als Datenstruktur, bei der für mehrere Untersuchungseinheiten jeweils mindestens zwei Messungen vorliegen und zusätzlich die Zeitintervalle zwischen den Messpunkten bei allen Untersuchungseinheiten identisch sind.

Die Techniken zur Analyse von Längsschnittdaten wurden in der Ökonomie entwickelt, wo zumeist eine *large t, small n*- Struktur vorherrscht (wenige Untersuchungseinheiten mit jeweils vielen zeitlichen Messpunkten). Die Daten sozialwissenschaftlicher Panels weisen häufig eine *small t, large n*- Struktur auf (viele Untersuchungseinheiten mit jeweils wenigen zeitlichen Messpunkten). Dieser Unterschied und die Tatsache, dass biographische Variablen einer Person oft andere Verläufe aufweisen als typische wirtschaftswissenschaftliche Methoden voraussetzen, erfordern alternative Analysemethoden sozialwissenschaftlicher Panels. (Giesselmann & Windzio, 2012, S.11)

Die Tatsache, dass menschliches Verhalten in seiner vollen Komplexität nicht durch ein einfaches Modell erklärt werden kann, motiviert die Einnahme verschiedener Blickwinkel zur ansatzweisen Beantwortung sozialwissenschaftlicher Fragestellungen. Im Rahmen dieser Master-Thesis werden verschiedene Methoden zur Paneldatenanalyse betrachtet und bezüglich ihrer sozialwissenschaftlichen Aussagekraft bewertet. Dieser Prozess ist per se geprägt durch seine quantitativ-empirische Ausrichtung gemäss König (1962). Während dieses Prozesses möchte ich u.a. feststellen:

- inwiefern die rein empirische Forschung in der Sozialwissenschaft zielführend ist und
- inwiefern unterschiedliche Methoden der Paneldatenanalyse den sozialwissenschaftlichen Erkenntnisgewinn sowohl in der Tiefe[5] als auch in der Breite[6] vorantreiben und sich ggf. ergänzen.

[5] Erschliessung von präziseren Erkenntnissen aufgrund verbesserter, umfassenderer Methoden

[6] Erschliessung einer neuen Art von Erkenntnissen durch die Einnahme neuer Blickwinkel

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

HSLU	Hochschule Luzern
IFSAR-OST	Institut für Soziale Arbeit und Räume der Ostschweizer Fachhochschule
IMS-OST	Institut für Modellbildung und Simulation der Ostschweizer Fachhochschule
OLS	ordinary least squares
FE	fixed effects
RE	random effects

bspw.	beispielsweise
bzw.	beziehungsweise
CHF	Schweizer Franken
dt.	Deutsch
d. h.	das heisst
etc.	et cetera
exkl.	exclusive
inkl.	inklusive
mind.	mindestens
o.ä.	oder ähnliche(s)
resp.	respektive
S.	Seite
u.a.	unter anderem
usw.	und so weiter
v.a.	vor allem
vgl.	vergleiche
z.B.	zum Beispiel

– Teil 1 –

Einleitung

1 Stand der Forschung

1.1 Gefährdungsbeurteilung psychosozialer Risiken

"In der derzeitigen Diskussion um die Zukunft und Entwicklung der Arbeitswelt gilt es als eine der größten Herausforderungen, die psychische Gesundheit von Erwerbstätigen zu erhalten" (Paulus, 2019, S.141). Zentral hierbei ist die Bewertung der Gefährdungskonstellationen von Belastungen am Arbeitsplatz und im Privatleben unter Berücksichtigung subjektiver Bewältigungsstrategien. "Zur Erklärung möglicher Gefährdungskonstellationen der Schnittstelle Arbeit-Mensch-Gesundheit/Krankheit wird in der Arbeitswissenschaft und in angrenzenden Disziplinen, wie Arbeitssoziologie, -psychologie, auf das theoretische Belastungs-Beanspruchungs-Konzept nach Rohmert (1984) zurückgegriffen" (Paulus, 2019, S.143). Dieses besagt, dass Belastungen (wie z.B. Arbeitsmenge, Zeitdruck, Digitalisierung) personenbezogene Beanspruchungen resp. Beanspruchungsfolgen (wie z.B. Stress, Schlafstörungen, Depression) mit sich bringen, die wesentlich durch individuelle Bewältigungsstrategien beeinflusst werden (vgl. auch Böhle 2010, S. 451–481; Oppolzer, 2010, S. 13–22).

Abbildung 1 zeigt eine eigene Darstellung des Belastungs-Beanspruchungs-Konzept nach Rohmert (1984), unter der Erweiterung um eine zusätzlich subjektwissenschaftliche Perspektive. Darin werden Belastungen am Arbeitsplatz und im Privatleben einer Person (hellgrün), sowie deren subjektive Empfindung (dunkelgrün) als Ursprung einer Gesamtbelastung betrachtet. Zur Bewältigung dieser kombinierten Gesamtbelastung stehen individuelle Ressourcen (gelb) zur Verfügung, wobei die Person gleichzeitig an subjektive, gesellschaftliche und betriebliche Rahmenbedingungen (orange) gebunden ist.

Aufgrund der vorhandenen Belastungen, Rahmenbedingungen und Ressourcen entwickelt diese Person eine individuelle Bewältigungsstrategie (blau) anhand derer auf die Belastungen reagiert wird. Die zeitliche Entwicklung von Belastungen, Rahmenbedingen und Ressourcen kann als dynamischer Prozess verstanden werden, wodurch sich eine Bewältigungsstrategie über die Zeit ändern kann. Als Folge der Reaktion auf die gegebenen Belastungen entstehet eine individuelle Beanspruchung, die individuelle (positive/negative sowie kurzfristige/langfristige) Beanspruchungsfolgen mit sich bringt.

Aufgrund einschränkender Bedingungskonstellationen und fehlender Möglichkeiten zur Umgestaltung dieser Bedingungen, können selbstschädigende Muster und Einschränkungen der individuellen Handlungsfähigkeit entstehen. Die individuelle Bewältigungsstrategie einer Person wird dadurch zum interessanten Untersuchungsobjekt von Sozialforschenden. Aus Sicht der Datenwissenschaft sind individuelle Bewältigungsstrategien einerseits schwierig zu quantifizieren und ebenso umständlich in der Erhebung. Die zugrundeliegende Komplexität dieser und ähnlicher Grössen macht deren Erfassung in Fragebögen beinahe unmöglich, weshalb viele Paneldatensätze keine solche Informationen enthalten.

Die Problematik fehlender Informationen zu Bewältigungsstrategien eines Individuums schliesst die Möglichkeit jedoch nicht aus, datenbasierte Erkenntnisse zur Interdependenz zwischen Belastungen und Beanspruchungen zu gewinnen. So scheint die Idee naheliegend, dass eine Gruppe von Belastungen/Belastungsarten in engerer Verbindung mit einer Beanspruchung steht als andere. Sofern spezifische Muster in Belastungs-Bewältigungsstrategie-Beanspruchungs-Kombinationen existieren, kann die Betrachtung von Belastungs-Beanspruchungs-Kombinationen Rückschlüsse auf Gefährdungskonstellationen und Belastungsfaktoren ermöglichen.

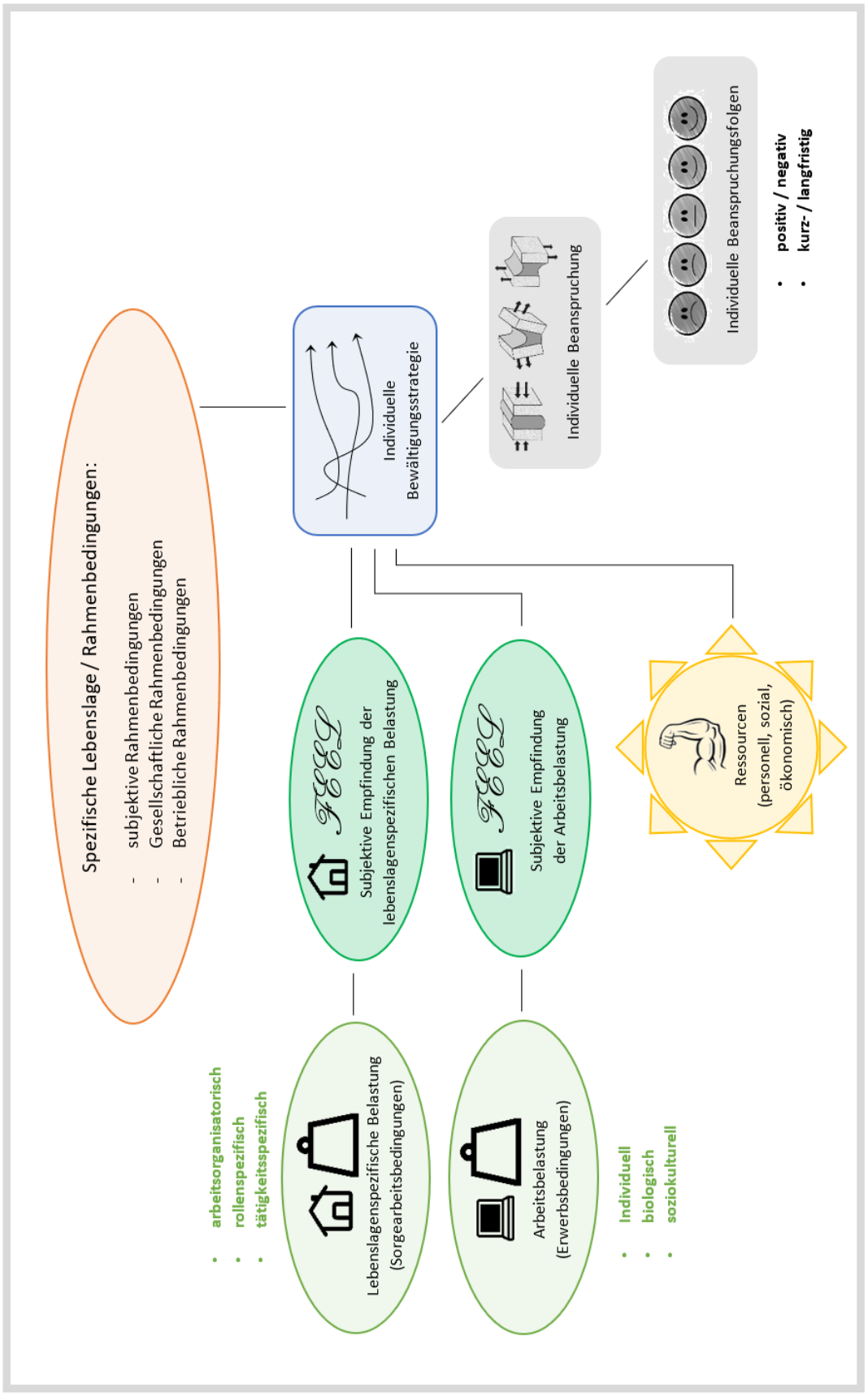


Abbildung 1: Erweiterte Darstellung des Belastungs-Beanspruchungs-Konzepts, eigene Darstellung

Die Identifikation solcher Verbindungen mittels Paneldaten hat das Potential, individuelle (gesundheitsgefährdende und -erhaltende) Verläufe von Erwerbstätigen besser zu verstehen. Zur Unterstützung eines solchen Erkenntnisgewinns werden in dieser Arbeit Methoden der Paneldatenanalyse am Beispiel soziologischer Paneldaten angewandt.

1.2 SNF-Forschungsprojekt "Psychosoziale Risiken in der Arbeitswelt"

Die Gefährdungsbeurteilung psychosozialer Risiken wird in einem laufenden Forschungsprojekt "Psychosoziale Risiken in der Arbeitswelt" des Schweizer Nationalfonds (SNF) bearbeitet. Im Fokus dieses Projekts liegt sowohl die arbeitswissenschaftliche als auch subjektwissenschaftliche Betrachtung psychosozialer Risiken (vgl. Abbildung 1). Die beteiligten Institutionen dieses interdisziplinären Grundlagenforschungsprojekts sind das Institut für Soziale Arbeit und Räume (IFSAR-OST), das Institut für Modellbildung und Simulation der Ostschweizer Fachhochschule (IMS-OST) sowie das ICAS[7]. Das Projekt fokussiert auf folgende drei Forschungsfragen (IFSAR, 2019):

1. Welche Konstellationen von Arbeitsbelastungen in Wechselwirkung mit spezifischen Lebenslagen ergeben gesundheitsgefährdende und -erhaltende Arbeitsbeanspruchungen?
2. Wie sehen gesundheitsgefährdende und -erhaltende Bewältigungsstrategien aus?
3. Wie können Gefährdungsbeurteilungen auf der Grundlage multifaktorieller Wirkungszusammenhänge prognostisch berechnet und dargestellt werden?

Die Tätigkeiten innerhalb des Projekts zielen darauf ab, die Früherkennung und Prävention psychosozialer Risiken im Spannungsfeld zwischen Arbeit und spezifischer Lebenslage zu fördern. An dieser Stelle sei zu erwähnen, dass der Begriff *Arbeit* (Arbeitsbelastung) im Rahmen dieses SNF-Forschungsprojekts sowohl die *Erwerbsarbeit* (Erwerbsarbeitsbelastung) als auch die *Sorgearbeit* (Sorgearbeitsbelastung) anspricht. Für die kommenden Ausführungen wird diese Begriffsdefinition übernommen.

Als geplante Ergebnisse dieses mehrjährigen Forschungsprojekts entstehen (IFSAR, 2019):

1. eine Multifaktorenanalyse der Interdependenzen zwischen strukturellen Arbeitsbelastungen und subjektiv empfundenen Arbeitsbeanspruchungen. Zweck dieser Analyse ist es,
2. Auswirkungen des Spannungsfeldes von psychosozialen Risikofaktoren und damit einhergehende gesundheitsgefährdende sowie gesundheitserhaltende Bewältigungsmuster zu erkennen, um
3. ein theoretisches Simulationsmodell abzuleiten, welches die Grundlage für eine evidenzbasierte Gefährdungsbeurteilung herstellt, die prognostisch risikoreiche Konstellationen erkennt und bewertet.

Diese Master-Thesis hat den Anspruch, im Rahmen dieses Forschungsprojekts, die Beantwortung der ersten Forschungsfrage und die Erreichung des ersten Forschungsergebnisses zu unterstützen. Eine inhaltliche Verknüpfung zwischen Master-Thesis und Forschungsprojekt ist durch die Projektpartner durchaus gewünscht, jedoch nicht zwingend erforderlich. Die Erkenntnisse dieser Arbeit fliessen direkt in den Erkenntnisgewinn des Projekts ein und werden dort weiterverwertet.

[7] http://www.icas-eap.com/

1.3 Stand der Forschung in der Arbeitswissenschaft

Die folgenden Ausführungen beschreiben drei Forschungsdesiderate aus Sicht der Arbeitswissenschaft. In Kurzform lässt sich behaupten, dass diese Master-Thesis eine Forschung anstrebt, die nach Paulus (2019, S. 148) "noch bislang unbearbeitet ist":

> *"Eine Forschung, welche auf die Bearbeitung der Wechselwirkungen und Abhängigkeitskonstellationen von arbeitsorganisatorischen, individuellen, biologischen und soziokulturellen Bedingungen fokussiert, ist bislang unbearbeitet."*

1) Zum Stand von Analysetools psychosozialer Risiken hält Paulus (2019, S. 142) fest:

   "Insgesamt haben sich Mess- und Bewertbarkeit psychosozialer Risiken besonders im Kontext des Burnout-Syndroms zu einem Forschungsbereich entwickelt. Weiterhin besteht ein erheblicher Forschungsbedarf aus Sicht der Arbeitswissenschaft und den daran beteiligten Disziplinen, wie Arbeitspsychologie, des Betrieblichen Gesundheitsmanagements oder der Betrieblichen Sozialen Arbeit, Gefährdungsbeurteilungen zu systematisieren und anwendungsorientierte bzw. dynamische Analysetools zu entwickeln… Janetzke und Ertel (2016) resümieren in ihrer Studie zur Gefährdungsbeurteilung im europäischen Vergleich, dass die bisherigen Instrumente der Gefährdungsbeurteilung künftig um eine stärkere Prozesssicht zu erweitern sind. D. h., Prozesse und Modelle der Beurteilung sollen verbessert werden (ebd.: 81)."

2) Zusätzlich impliziert Paulus (2019, S. 144) einen Bedarf zur lebenslagenspezifischen Erweiterung des Belastungs-Beanspruchungs-Konzepts:

   "Die Ergebnisse der geschlechtersensiblen Arbeitsforschung erweitern das Belastungs-Beanspruchungs-Konzept insofern, dass nicht ausschließlich Erwerbsarbeitsbelastungen Arbeitsbeanspruchungen erzeugen, sondern dass Wechselwirkungen mit lebenslagenspezifischen Belastungsfaktoren Gesundheitsbeeinträchtigungen potenzieren können."

3) Des Weiteren wird festgestellt, dass "strukturelle und familiäre Bedingungen sowie psychosoziale Belastungen zusammen analysiert und Untersuchungen auf multifaktorielle Belastungen ausgerichtet werden sollen." (Vanis et al., 2017, zit. in Paulus, 2019, S. 144) und dass "Erwerbsarbeitsbelastungen in einem Zusammenhang mit lebenslagen-, rollen- und geschlechtsspezifischen Aspekten stehen" (Paulus, 2019).

Zusammenfassend lässt sich festhalten, dass, nach aktuellem Stand der Forschung, ein besseres Verständnis der wirkenden Zusammenhänge in Situationen mit Erwerbsarbeitsbelastungen, lebenslagenspezifischen Belastungsfaktoren und subjektiv empfundenen Arbeitsbeanspruchungen erwünscht ist. Die Modellierung solcher Zusammenhänge, basierend auf Paneldaten, soll am Ende dieser Master-Thesis dabei helfen, das inhaltliche Verständnis dieser Zusammenhänge zu fördern und passende Methoden für entsprechende Fragestellungen vorzuschlagen.

2 Problemstellung

2.1 Zentrale Forschungsfrage

Unter Berücksichtigung der fachlichen Ausrichtung im Studiengang "Applied Information and Data Science" an der Hochschule Luzern (HSLU), wird die zentrale Forschungsfrage dieser Master-Thesis bewusst mit fachlichem statt inhaltlichem Fokus definiert.

> *Kann der empirische Erkenntnisgewinn von State-of-the-Art-Methoden der longitudinalen Paneldatenanalyse sowohl in der Tiefe[8] als auch in der Breite[9] durch fortgeschrittene Machine Learning Methoden[10] ergänzt werden?*

Der Fokus auf Längsschnittfragen liegt darin begründet, dass aus sozialwissenschaftlicher Sicht sowohl Arbeitsbelastungen und Arbeitsbeanspruchungen als auch spezifische Lebenslagen, eine zeitliche Variabilität aufweisen. Aus diesem Grund wird vermutet, dass die longitudinale Betrachtung einen Mehrwert gegenüber der Querschnittsbetrachtung bietet.

Querschnittsfragen sind damit zwar nicht ausgeschlossen, haben jedoch zweitrangigen Charakter. In der Diskussion (Kapitel 9) wird genauer darauf eingegangen, inwiefern Querschnitts- und Längsschnittbetrachtungen für die inhaltlichen Ansprüche des vorliegenden Datensatzes geeignet sind und inwiefern sich diese ergänzen.

2.2 Teilfragen und Ergebnisse

Zur Beantwortung der zentralen Forschungsfrage werden Teilfragen unter den Stichworten *Untersuchungsszenario*, *Datenaufbereitung* und *Modellierung* definiert. Diese helfen bei der Kanalisierung des generellen Erkenntnisgewinns und legen gleichzeitig die Strukturierung im Hauptteil dieser Arbeit fest.

Untersuchungsszenario:

Welche Art von Untersuchungsszenario ist im Kontext der inhaltlichen Rahmenbedingungen für den geplanten Methodenvergleich geeignet?

Als Ergebnis dieser Teilfrage wird ein Szenario definiert, das dem sozialwissenschaftlichen Anspruch gerecht wird, gleichzeitig potenzielle Gefährdungskonstellationen von spezifischen Lebenslagen und Arbeitsbelastungen (Erwerbsarbeit- & Sorgearbeit) zu beleuchten, um ihre Auswirkung auf mögliche Beanspruchungsfolgen zu untersuchen. Ein Szenario wird beschreiben durch ein Set von Beobachtungsmerkmalen, d.h. ein abhängiges Merkmal und eine Menge von unabhängigen Merkmalen.

[8] Erschliessung von präziseren Erkenntnissen aufgrund verbesserter, umfassenderer Methoden

[9] Erschliessung einer neuen Art von Erkenntnissen durch die Einnahme neuer Blickwinkel

[10] In Kapitel 3.2 "Methoden der Paneldatenanalyse" werden State-of-the-Art-Methoden für die Analyse von Paneldaten sowie fortgeschrittene Machine Learning Methoden, die im Rahmen dieser Arbeit betrachtet werden, beleuchtet.

Datenaufbereitung:

Wie können die Sphären *spezifische Lebenslage*, *Arbeitsbelastung* und *Arbeitsbeanspruchung* anhand des verfügbaren Datensatzes operationalisiert werden?
Welchen Mehrwert bezüglich des Erkenntnisgewinns bieten explorative Datenanalysen, ohne die explizite Verwendung von Machine Learning Methoden?

Als Ergebnis dieser Teilfragen wird ein bereinigter Paneldatensatz erstellt, der die drei Sphären als Menge von operationalisierten Variablen abbildet. Zusätzlich werden Methoden zur explorativen Datenanalyse eines solchen Paneldatensatzes vorgestellt und bezüglich ihrem Mehrwert bewertet.

Modellierung:

Kann der empirische Erkenntnisgewinn von State-of-the-Art-Paneldatenanalysen des Untersuchungsszenarios sowohl in der Tiefe als auch in der Breite durch fortgeschrittene Machine Learning Methoden ergänzt werden?

Die Beantwortung der letzten Teilfrage geht einher mit der Beantwortung der zentralen Forschungsfrage anhand des Untersuchungsszenarios.

2.3 Ziel

Das fachliche Ziel dieser Master-Thesis ist der Vergleich von Methoden zur Analyse von sozialwissenschaftlichen Paneldaten. Die Art des Vergleichs ist von qualitativer Natur und soll dem sozialwissenschaftlichen Analysten als Hilfestellung bei der Methodenauswahl dienen. Der Vergleich basiert auf folgenden Kriterien:

- Anwendbarkeit ("Wie gut lässt sich Methode X auf das Untersuchungsszenario anwenden?")
- Erkenntnistheoretischer Mehrwert im sozialwissenschaftlichen Kontext bezüglich:
    - Breite ("Erschliesst Methode X eine neue Art von Erkenntnissen durch die Einnahme neuer Blickwinkel?")
    - Tiefe ("Erschliesst Methode X von präziseren Erkenntnissen aufgrund verbesserter, umfassenderer Methoden?")

Die fachliche Formulierung der zentralen Forschungsfrage hat zur Folge, dass der inhaltliche Teil (Beantwortung sozialwissenschaftlicher Fragestellungen) nicht das zentrale Forschungsobjekt dieser Master-Thesis darstellt. Die Relevanz des sozialwissenschaftlichen Kontextes ist nach wie vor gegeben, da Methoden der Paneldatenanalyse anhand *sozialwissenschaftlicher* Paneldaten beurteilt und verglichen werden.

Das sozialwissenschaftliche Ziel dieser Master-Theis ist die Beurteilung von Abhängigkeiten zwischen spezifischer Lebenslage, Arbeitsbelastung und Arbeitsbeanspruchung. Potenzielle Abhängigkeiten dieser Grössen werden nicht durch sozialwissenschaftliche Hypothesen aus dem SNF-Forschungsprojekt oder gängiger Theorie vorgegeben. Aus diesem Grund wird der Findungsprozess solcher Abhängigkeiten als stark explorativ betrachtet. Folglich soll der geplante Methodenvergleich aus sozialwissenschaftlicher Sicht dabei helfen, herauszufinden:

- welche zusätzlichen Erkenntnisse eine Methode in diesem explorativen Modellierungsvorgang hervorbringt und
- ob sich Erkenntnisse mehrerer Methoden gegenseitig ergänzen oder vervollständigen.

D.h.: Diese Master-Thesis soll u.a. einer sozialforschenden Person dabei helfen, den wahren Abhängigkeiten auf den Grund zu kommen und neue Hypothesen aufzustellen. Ein Wunschziel hierfür ist die Erarbeitung eines Prozesses oder Workflows, um Sozialforschenden innerhalb des SNF-Forschungsprojekts verschiedene Methoden und deren Eignung in Bezug auf die konkrete Fragestellung vorzustellen.

Nicht-Ziele dieser Master-Thesis sind:

- Die Entwicklung neuer Methoden zur Analyse von Paneldaten.
- Die Bewertung einzelner Methoden im Kontext allgemeiner sozialwissenschaftlicher Paneldaten. Sämtliche Aussagen der Arbeit bewegen sich im Kontext der betrachteten Paneldaten (vgl. Kapitel 4.2 "Datengrundlage")

2.4 Überblick

In Kapitel 3 und 4 werden die allgemeine Methodik dieser Arbeit, die verwendeten Methoden zur Paneldatenanalyse sowie die Grundlagen der Paneldatenanalyse erarbeitet. Die in Kapitel 2.2 definierten Teilfragen zu den Themen *Untersuchungsszenario*, *Datenaufbereitung* und *Modellierung* werden jeweils in den Kapiteln 5, 6 und 7 bearbeitet. Die fachlichen Ergebnisse inkl. Methodenvergleich werden in Kapitel 8 präsentiert. Kapitel 9 fasst die inhaltlichen Erkenntnisse aus Sicht der Sozialwissenschaft zusammen und gibt eine kritische Reflexion der gemachten Erfahrungen.

3 Methodik

Die folgenden Kapitel geben eine Überblick der in dieser Master-Thesis verwendeten Methoden.

3.1 Epicycle of Analysis

Die explorative Datenanalyse des SHP-Datensatzes wird gemäss dem "Epicycle of Analysis" (Peng et al., 2015) durchgeführt. Dieser beschreibt fünf Schritte der Datenanalyse, die nicht linear, sondern in einem iterativen Prozess durchlaufen werden. In jedem Schritt wird neue Information gewonnen, wodurch andere Schritte wiederum von neuem gestartet werden sollten oder können. Die fünf Schritte sind:

1. Festlegen und Verfeinern der Fragestellung
2. Untersuchung der Daten („exploring the data“)
3. Bildung formaler, statistischer Modelle
4. Interpretation der Resultate
5. Kommunikation der Resultate

Abbildung 2: Epicycle of Analysis, (Peng et al., 2015)

Ein iteratives Durchlaufen der fünf Kernaktivitäten unterstützt einen sinnvollen Erkenntnisgewinn.

3.2 Methoden der Paneldatenanalyse

Im Modellierungsteil dieser Master-Thesis werden fünf State-of-the-Art Methoden (vgl. Kapitel 7.1 - 7.5) und fünf fortgeschrittenen Methoden (vgl. Kapitel 7.7 - 7.11) betrachtet. In Kapitel 7.6 werden zusätzliche Tests vorgestellt, um ein Untersuchungsszenario auf spezifische Modellierungseigenschaften zu untersuchen.

Die State-of-the-Art Methoden werden in Kapitel 4.1.2 und 4.1.3 gemäss (Giesselmann & Windzio, 2012) eingeführt. Diese sind:

- Gepooltes Modell
- Varaible Coefficients Modell (VCM)
- First Differences Modell (FD)
- Fixed Effects Modell (FE)
- Random Effects Modell (RE)

Die fortgeschrittenen Methoden werden direkt bei deren Anwendung in Kapitel 7.7 - 7.11 eingeführt. Diese sind:

- Random Effects Modell mit Kontextvariablen (RE-KV) – vgl. Giesselmann & Windzio (2012)
- Longitudinales Clustering (LC) – vgl. Genolini et al. (2015, 2016)
- Instrumentelle Variablen (IV) – vgl. Pokropek (2016)
- Gemeinsame Mittelwert-Kovarianz Modellierung (JMCM) – vgl. Pan & Pan (2017)
- Baumbasierte variable Koeffizienten Regression (TVCM) – vgl. Bürgin & Ritschard (2017)

3.3 Triangulation

Als Methode zur Bewertung des Erkenntniszuwachses bei der Verwendung verschiedener Methoden wird die Triangulation gemäss Oelerich und Otto (2011) angewandt.

"Triangulation beinhaltet die Einnahme unterschiedlicher Perspektiven auf einen untersuchten Gegenstand. Diese Perspektiven können sich in unterschiedlichen Methoden, die angewandt werden, und/oder unterschiedlichen gewählten theoretischen Zugängen konkretisieren… Durch die Triangulation (etwa verschiedener Methoden oder verschiedener Datensorten) sollte ein prinzipieller Erkenntniszuwachs möglich sein, dass also bspw. Erkenntnisse auf unterschiedlichen Ebenen gewonnen werden, die damit weiter reichen, als es mit einem Zugang möglich wäre" (Flick, 2008, zit. in Oelerich & Otto, 2011, S. 324).

3.4 R/RStudio

Die zentrale Forschungsfrage legt die Notwendigkeit fundierter und flexibler Datenanalysen am vorhandenen Datensatz dar. Ein geeignetes Tool für solche Datenanalysen ist die Programmiersprache R (RStudio, 2020; The R Foundation, 2020).

"R stellt eine Umgebung zur Verfügung, in der statistische Analysen und Grafiken erzeugt werden können… R ist designt, um Resultate von statistischen Prozeduren weiter zu verarbeiten… Die Tatsache, dass die Programmiersprache R auf einer formalen Computersprache basiert, gibt ihr enorme Flexibilität." (Dalgaard, 2008, S. vi, Eigene Übersetzung)

Für die rasche Aufbereitung von statistischen Analysen bietet die Programmiersprache R zwei Interfaces an, die das iterative Reporting der Thesis unterstützen:

1. RMarkdown ist eine Markdown Language zur niederschwelligen Erzeugung von Berichten. Der Vorteil von RMarkdown liegt in der Einbettung von Text, Grafiken und R-Code in einem Dokument, wodurch Erkenntnisse rasch aufgearbeitet und präsentiert werden können. (R Markdown, 2020)
2. Shiny ist ein R-Paket zur Erstellung interaktiver Apps. Es ist denkbar, dass für spezifische Fragestellungen der Master-Thesis kleinere Shiny Apps erzeugt werden. (Shiny, 2020)

Für die Bearbeitung der dritten Teilfrage werden Methoden der longitudinalen Paneldatenanalyse sowie fortgeschrittene Machine Learning Methoden vom Comprehensive R Archive Network (CRAN) in Form von R-Paketen bezogen. Eine Hilfestellung für passende Methoden liefert das Journal of Statistical Software (*J. Stat. Softw.*, 2020).

4 Grundlagen

4.1 Einführung in die Paneldatenanalyse

In diesem Kapitel wird die Nomenklatur zur Beschreibung einzelner Methoden der Paneldatenanalyse nach Giesselmann & Windzio (2012) eingeführt. Darauf aufbauend werden die zwei ursprünglichen Regressionsmethoden (OLS-Modell und gepooltes Modell) sowie weiterführende Regressionsmethoden vorgestellt. Am Ende dieses Kapitels wird auf das Thema Endogenität und Exogenität eingegangen, mit einem Fokus auf den Annahmen der beschriebenen Modelle.

4.1.1 Nomenklatur

Ein Paneldatensatz besteht aus Messungen verschiedener Untersuchungseinheiten (z.B. *Länder* oder *Personen*) $\boldsymbol{i}$ über mehrere Zeitpunkte (*Messpunkte*) $\boldsymbol{t}$ hinaus. Die Zeitintervalle zwischen den Messpunkten müssen bei allen Untersuchungseinheiten identisch sein.

Bei linearen Regressionsmodellen unterscheiden Giesselmann & Windzio (2012) zwei Gruppen unabhängiger Variablen: Variablen, denen ein zeitlich unveränderliches Merkmal (z.B. *ethnische Herkunft* oder *Geschlecht*) zugrunde liegt, werden durch den Buchstaben $\boldsymbol{z}$ gekennzeichnet. Die zweite Gruppe bilden Variablen, die im zeitlichen Verlauf einer Einheit unterschiedliche Ausprägungen annehmen können (z.B. *Familienstand* oder *Erwerbsstatus*) – diese werden mit den Buchstaben $\boldsymbol{x}$ gekennzeichnet.

Die abhängige Variable wird mit $\boldsymbol{y}$ bezeichnet und Koeffizienten, die den Effekt einer unabhängigen Variablen auf die abhängige Variable beschreiben, werden mit $\boldsymbol{b}$ (oder β) gekennzeichnet. Der kumulierte Effekt aller nicht in einem Modell integrierten Eigenschaften (*unbeobachtete Eigenschaften*) wird durch den Fehlerterm $\boldsymbol{w}$ repräsentiert. Dieser Fehlerterm wird weiter aufgeteilt in einen zeitkonstanten Fehlerterm $\boldsymbol{u}$ (*Einheiteneffekt* oder *Personeneffekt*) und einen über die Zeit veränderlichen Fehlerterm $\boldsymbol{e}$ (*idiosynkratischer Fehler*), welcher sich ausschliesslich auf Abweichungen innerhalb der einheitenspezifischen Zeitreihe bezieht. Die folgende Tabelle gibt eine Übersicht der verwendeten Notationen.

Tabelle 1: Übersicht der verwendeten Notation

Term	Interpretation
$\boldsymbol{y_{it}}$	Abhängige Variable (Wert der abhängigen Variable von Einheit i zum Zeitpunkt t)
$\boldsymbol{\bar{y}_i}$	Abhängige Variable - einheitenspezifisches Mittel (zeitliches Mittel der abhängigen Variable von Einheit i)
$\boldsymbol{x_{it}}$	Unabhängige, zeitveränderliche Variable (Wert der unabhängigen Variable von Einheit i zum Zeitpunkt t)
$\boldsymbol{\bar{x}_i}$	Unabhängige Variable - einheitenspezifisches Mittel (zeitliches Mittel der unabhängigen Variable von Einheit i)
$\boldsymbol{z_i}$	Unabhängige, zeitkonstante Variable (Wert der unabhängigen Variable von Einheit i). Es gilt: $\boldsymbol{z_i = \bar{z}_i}$
$\boldsymbol{w_i}$	Fehlerterm (Kumulierter Effekt aller nicht integrierten Merkmale von Einheit i)
$\boldsymbol{w_{it}}$	Fehlerterm (Kumulierter Effekt aller nicht integrierten - zeitkonstanten und zeitveränderlichen - Merkmale von Einheit i zum Zeitpunkt t). Diesen Fehlerterm teilen wir in die zeitkonstanten und zeitveränderlichen Komponenten auf: Es gilt: $\boldsymbol{w_{it} = e_{it} + u_i}$
$\boldsymbol{u_i}$	Fehlerterm (Kumulierter Effekt aller nicht integrierten zeitinvarianten Eigenschaften) → *Einheiten-* oder *Personeneffekt*
$\boldsymbol{e_{it}}$	idiosynkratischer Fehler Dieser Fehlerterm bezieht sich auf die einheitenspezifische (um Einheiteneffekte bereinigte) Messreihe und wird somit nur in Kombination mit dem Fehlerterm u_i verwendet.

4.1.2 State-of-the-Art: Regressionsmethoden ohne Transformation

Die Regressionsgleichung jedes Regressionsmodells stellt den Einfluss von Merkmalen (unabhängige Variablen) auf ein abhängiges Merkmal (Variable) dar und ist dadurch das statistische Abbild eines Zusammenhangs. Ein solcher Zusammenhang wird je nach Forschungsschwerpunkt (Ökonomie, Sozialwissenschaft, Biologie, Physik, …) durch sehr unterschiedliche Merkmale beschrieben und entsprechend unterschiedlich interpretiert. Im sozialwissenschaftlichen Kontext unserer Untersuchungen wird das abhängige Merkmal stets durch eine messbare Variable der *Arbeitsbeanspruchung* beschrieben. Die unabhängigen Variablen werden durch messbare Variablen der *spezifischen Lebenslage* und der *Arbeitsbelastung* beschreiben.

Zur intuitiven Beschreibung der State-of-the-Art Methoden der Regressionsanalyse verwenden wir folgendes Gedankenbeispiel:

Eine sozialwissenschaftlich forschende Person möchte für eine gegebene Population den Einfluss der unabhängigen Variable x: Kalenderwoche auf die abhängige Variable y: Zufriedenheit von Personen (Untersuchungseinheiten) untersuchen und dadurch verstehen, wie sich die allgemeine Zufriedenheit dieser Population über die Zeit entwickelt. Hierfür werden sechs Personen zweimal wöchentlich zur individuellen *Zufriedenheit* befragt und ein Paneldatensatz erstellt.

OLS Modell:

Um das obige Beispiel mit einem OLS-Modell zu beschrieben, wird die forschende Person die entsprechende OLS-Regressionsgleichung aufstellen.

$$y_i = a + b \cdot x_i + w_i \ , \quad w_i = e_i \tag{1}$$

Der Index i bezeichnet in diesem Modell eine einzelne Messung (x_i, y_i) aus einer Menge von Messpunkten (und nicht eine Person). Gemäss den Annahmen der OLS-Regression wird der Fehlerterm w_i allein durch die Zufallsvariable e_i definiert und darf keinerlei Systematik[11] aufweisen. Abbildung 3 zeigt die Messpunkte inkl. der entsprechenden Regressionsgeraden.

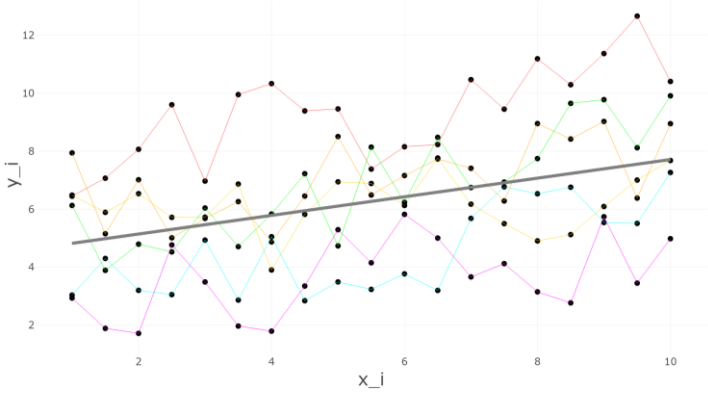


Abbildung 3: OLS-Modell (2D) inkl. verborgener Panelstruktur in Farbe

[11] Eine Systematik in den Fehlertermen wie z.B. Heteroskedastizität (Korrelation zwischen der Variation des Fehlerterms und den unabhängigen Variablen) ist ein Zeichen dafür, dass die Annahmen des Regressionsmodells verletzt sind. Gründe hierfür sind z.B. unbeobachtete Heterogenität, Gruppenunterschiede, Messfehler oder umgedrehte Kausalität.

Im OLS-Modell wird jede Messung als eigenständig betrachtet, d.h. unabhängig von der Zugehörigkeit der Messung zur Untersuchungseinheit. Der vorhandene Paneldatensatz wird folglich als Querschnitt betrachtet, wodurch die OLS-Regression blind ist für die Panelstruktur resp. die Gruppierung von Messungen (vgl. Farbcodierung in Abbildung 3). Heben wir diese Panelstruktur des Datensatzes (Zugehörigkeit einer Messung zu einer Einheit) als dritte Dimension hervor (vgl. Abbildung 4) wird ersichtlich, dass das OLS-Modell für sämtliche Einheiten identische Voraussagen macht, obwohl offensichtlich einheitenspezifische Unterschiede existieren.

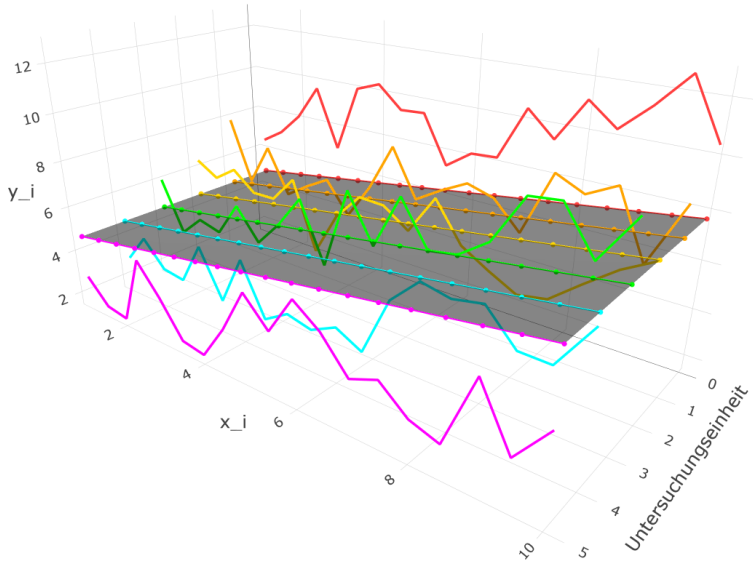


Abbildung 4: OLS-Modell (3D) inkl. verborgener Panelstruktur in Farbe

Gepooltes Regressionsmodell:

Um die implizite Struktur des Paneldatensatzes adäquat abzubilden, wird im gepoolten Regressionsmodell die Zugehörigkeit einer Messung zu einer Person mitberücksichtigt, d.h. jede Messung wird eindeutig einer Person i und einem Zeitpunkt t zugeordnet (x_{it}, y_{it}). Die Bedeutung des Index i verändert sich im Vergleich zum OLS-Modell daher massgeblich.

$$y_{it} = b_1 \cdot x_{it} + b_2 \cdot z_i + w_{it} \quad , \quad w_{it} = e_{it} + u_i \tag{2}$$

Aufgrund dieser Neuformulierung können zusätzlich unabhängige, zeitkonstante Variablen z_i betrachtet werden; die einheitenspezifischen Effekte. Anhand dieser kann die im OLS-Modell noch unbeobachtete einheitenspezifische Heterogenität nun integriert und dadurch kontrolliert werden. Der Fehlerterm w_{it} bezeichnet im gepoolten Modell sämtliche nicht berücksichtigten Einheiteneffekte u_i sowie den zeitveränderlichen idiosynkratischen Fehler e_{it}.

Im verwendeten Gedankenbeispiel kann z_i beispielsweise als "*mittlere Anzahl schlechter Erfahrungen pro Woche*" betrachtet werden, die für jede Person unterschiedlich ist und die abhängige Variable y_{it} für wachsende Werte negativ beeinflusst. Abbildung 5 zeigt, wie das gepoolte Modell somit zusätzlich personenspezifische Effekte mitberücksichtigt.

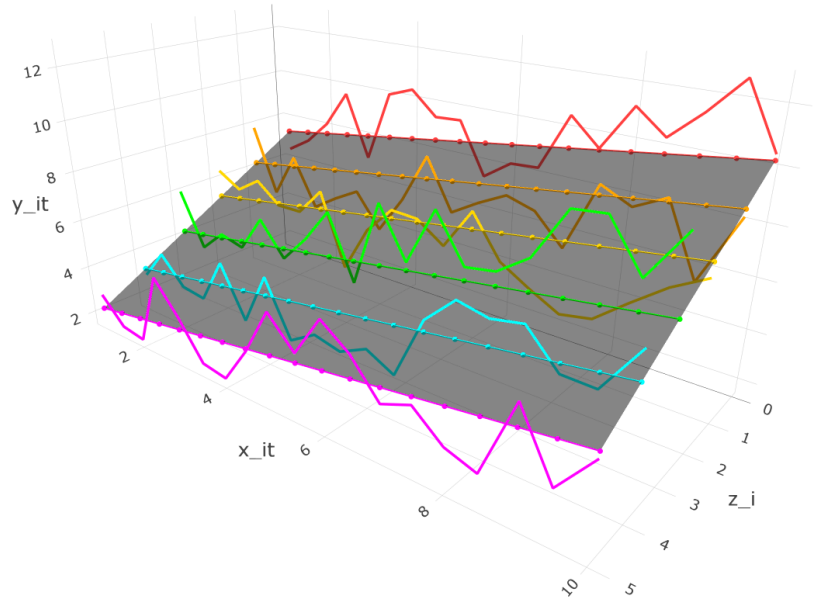


Abbildung 5: Gepooltes Regressionsmodell (3D)

Für ein gepooltes Modell mit insgesamt n Variablen (1 abhängige Variable, n-1 unabhängige Variablen) legen die Parameter $b1$ und $b2$ die jeweiligen Steigungen einer (n-1)-dimensionalen Hyperebene im n-dimensionalen Raum fest. Daraus folgt, dass ein gepooltes Modell einheitenspezifische Effekte zulässt und diese explizit in das Modell überführt.

4.1.3 State-of-the-Art: Regressionsmethoden mit Transformation

Da einheitenspezifische Effekte (Niveauunterschiede) zwischen den Untersuchungseinheiten je nach Fragestellung der forschenden Person von Interesse sind, werden sie vollständig integriert, teilweise integriert oder vollständig vernachlässigt, wodurch verschiedene Methoden der Paneldatenanalyse hervorgehen. All diese Methoden unterscheiden sich von einfachen OLS-Regressionen dadurch, dass sie die Daten vor der Anwendung von OLS transformieren. Durch diese Transformationen werden spezifische (längs- oder quer-) Eigenschaften der Paneldaten vermehrt in den Fokus der Modellierung gerückt. Tabelle 2 gibt eine Übersicht der für diese Master-Thesis relevanten Methoden, wobei die grün markierten Methoden tatsächlich Verwendung finden werden.

Tabelle 2: Wichtige Methoden der Paneldatenanalyse nach Giesselmann & Windzio (2012)

Methode	**Umgang mit einheitenspezifischen Effekten z_i bei der Modellierung durch Transformation**
First Differences (FD-Modelle)	Vernachlässigung durch Verwendung erster Differenzen $x_{it} - x_{i(t-1)}$ → $z_{it} - z_{i(t-1)} = z_i - z_i = 0$ $(y_{it} - y_{i(t-1)}) = b_1 \cdot (x_{it} - x_{i(t-1)}) + \sout{b_2 \cdot (z_i - z_i)} + w_{it}$
Fixed Effects (FE-Modelle)	Vernachlässigung durch Entmittlung (engl. "demeaning") $x_{it} - \bar{x}_i$ → $z_i - \bar{z}_i = 0$ $(y_{it} - \bar{y}_i) = b_1 \cdot (x_{it} - \bar{x}_i) + \sout{b_2 \cdot (z_i - \bar{z}_i)} + w_{it}$

Least Square Dummy Variables (LSDV-Modelle)	Berücksichtigung durch Spezifikation zusätzlicher Dummy-Variablen $c_i \cdot D_i$, wobei $D_i \in \{0,1\}$ eine separate Variable ist, die Messungen von Einheit i eindeutig spezifiziert. Beispiel: $D_4 = [0,0,0,0,1,1,1,0,0,0, \dots]$ spezifiziert, dass die 5., 6. Und 7. Messung zur Einheit $i = 4$ gehören. → c_i greift den Einfluss unbeobachteter zeitinvarianter Variablen (z_i) auf. $y_{it} = b_1 \cdot x_{it} + \sum_{i=1}^{n-1} c_i \cdot D_i + w_{it}$
OLS mit Kontextvariablen (OLS-KV-Modelle)	Berücksichtigung durch Spezifikation von Kontextvariablen $\bar{x}_i$ → $\bar{x}_i$ misst den Kohorteneffekt, z_i misst den residualen Einheiteneffekt. $y_{it} = b_1 \cdot x_{it} + b_2 \cdot \bar{x}_i + b_3 \cdot z_i + w_{it}$
Random Effects (RE-Modelle)	Teilweise Berücksichtigung durch teilweise Entmittlung $x_{it} - \lambda_i \cdot \bar{x}_i$ wobei $\lambda_i \in [0,1]$. → Verwendung eines verbesserten Schätzers für einheitenspezifische Effekte. $y_{it} - \lambda_i \cdot \bar{y}_i = b_1 \cdot (x_{it} - \lambda_i \cdot \bar{x}_i) + b_2 \cdot (z_i - \lambda_i \cdot \bar{z}_i) + w_{it}$
Hybrides Modell (RE-KV-Modelle)	Berücksichtigung durch Spezifikation von Kontextvariablen $\bar{x}_i$ und gleichzeitige Anwendung der RE-Transformation.

LSDV-Modelle bringen identische b_1-Koeffizienten hervor wie FE-Modelle, da einheitenspezifische Effekte zwar explizit ins Modell integriert, jedoch vollständig in die Dummy-Variablen D_i überführt werden. Die Bestimmung der entsprechenden "Dummy-Effekte" c_i ist bei grossen Datensätzen äusserst zeitaufwändig. Da die explizite Bestimmung von einheitenspezifischen Effekten ohnehin nicht im Zentrum unserer sozialwissenschaftlichen Fragestellungen liegt, werden LSDV-Modelle nicht weiter berücksichtigt. Stattdessen werden longitudinale Effekte vollständig anhand von FE-Modellen bestimmt, was deutlich effizienter und gleich präzise ist.

In FD- und FE-Modellen werden einheitenspezifische Effekte z_i ausdifferenziert resp. ausgemittelt. OLS-KV-Modelle erlauben hingegen, trotz Kontrolle von unbeobachteter Heterogenität in $\bar{x}_i$, die Verwendung zeitkonstanter Merkmale z_i. Gleichzeitig gehen dabei jedoch statistische Nachteile einher, welche Giesselmann & Windzio (2012) folgendermassen beschreiben:

1. "Die statistische Bedingung für unverzerrte Schätzer der Koeffizienten ist in OLS-KV-Modell erfüllt, diejenige für korrekte Berechnungen der Standardfehler dagegen nicht." (S.55)
2. "Aus diesem Grund ist die Relevanz des OLS KV-Verfahrens in der empirischen Praxis begrenzt. OLS-KV wird deshalb eher für die Modellierung von Querschnittsfragen verwendet." (S.55)
3. "Da die Kontextvariablen explizit als statistisches Instrument fungieren und nicht aufgrund theoretischer Überlegungen in das Modell integriert werden, ist die inhaltliche Aufarbeitung der entsprechenden Koeffizienten im konkreten empirischen Fall nebensächlich." (S.56)

Aufgrund der verzerrten Schätzer von Standardfehlern und der empirischen Ausrichtung auf Querschnittsfragen wird das OLS-KV-Modell in dieser Arbeit nicht weiter berücksichtigt. Die Anwendung von Kontextvariablen findet jedoch eine wichtige Anwendung bei der hybriden Methode mittels Kombination mit dem RE-Modell (vgl. Kapitel 7.7).

Weil die Modellierung der zeitlichen Entwicklung von Arbeitsbeanspruchungen aufgrund spezifischer Lebenslage und Arbeitsbelastung die Sicht von Längsschnittfragen aufweist, werden folgende Panel-Methoden, die ausschliesslich für Querschnittfragen geeignet sind, ebenfalls ausgeschlossen:

- OLS mit korrigiertem Standardfehler
- Between Regression

4.1.4 Endogenität

Ein statistisches Hindernis für die korrekte Bestimmung linearer Regressionskoeffizienten ist Endogenität. Die Schätzung von Regressionskoeffizienten mittels OLS basiert auf der Annahme, dass die unabhängigen Variablen in keinem Zusammenhang mit den Fehlertermen stehen. Diese Exogenitätsannahme wird über die Kovarianz definiert:

Gewöhnliches OLS-Modell:	$y_i = b \cdot x_i + \varepsilon_i$	$Cov(x_i, \varepsilon_i) = 0$	(3)
Gepooltes OLS-Modell:	$y_{it} = b_1 \cdot x_{it} + b_2 \cdot z_i + u_i + e_{it}$	$Cov(x_{it}, u_i) = 0$	(4)

Ist diese Annahme verletzt, liegt Endogenität vor und es gilt $Cov(x_i, \varepsilon_i) \neq 0$ resp. $Cov(x_{it}, u_i) \neq 0$. Durch den Zusammenhang zwischen unabhängigen Variablen und dem Fehlerterm wird eine ungewollte Abhängigkeit implizit im Modell verwendet, wodurch die Schätzung der Regressionskoeffizienten inkonsistent wird. Mögliche Gründe für Endogenität sind unbeobachtete Heterogenität, Gruppenunterschiede, Messfehler oder umgedrehte Kausalität (Collischon & Eberl, 2020).

Für lineare Regressionsmodelle im sozialwissenschaftlichen Kontext ist unbeobachtete Heterogenität eine allgegenwärtige Quelle eines Bias. Vernachlässigte Drittvariablen (z.B. aufgrund fehlender Erhebung) können oft dazu führen, dass "unsichtbare" Effekte durch die unabhängigen Variablen mittransportiert werden und dadurch verzerrte Regressionskoeffizienten hervorbringen (engl. *omitted variable bias*). Die potenzielle Gefahr von Endogenität kann durch Hinzunahme vieler erklärender Variablen vermindert werden. Andererseits verlieren Modelle mit zu vielen unabhängigen Variablen ihre Erklärungskraft und die Interpretation wird zunehmend komplexer.

Das Endogenitätsproblem motiviert die Verwendung sinnvoller, erklärender Variablen mit möglichst grosser Erklärungskraft. Die Auswahl und Verwendung von erklärenden Variablen im Kontext der gegebenen Fragestellung wird in Kapitel 5 behandelt.

4.2 Datengrundlage

Für sämtliche Analysen wird der Schweizer Haushaltspanel (SHP) der FORSbase als Datengrundlage verwendet.

> *"FORSbase ist die virtuelle Plattform von FORS, die es … ermöglicht, auf Daten sozialwissenschaftlicher Projekte aus der Schweiz zuzugreifen und Informationen zu diesen Projekten zu beziehen." (FORSbase, 2020)*

Der SHP enthält vielfältige Informationen zu Personen und Haushalten seit 1999 und ist eine laufende, einzigartige, grossräumige, national repräsentative longitudinale Erhebung in der Schweiz. Die Daten des SHP stellen eine reichhaltige Informationsquelle bereit, um soziale Veränderungen in der Schweiz über signifikante Zeitperioden zu verschiedenen Themen zu untersuchen. (Tillmann et al., 2020)

Der aktuell zugängliche SHP besteht aus 21 Wellen, die in den Jahren 1999-2019 erhoben wurden. Die Wellen werden in drei Zeiträume eingeteilt (Tillmann et al., 2020):

Tabelle 3: Erhebungszeiträume des SHP (Tillmann et al., 2020):

Bezeichnung	Zeitraum	Befragte Haushalte pro Jahr	Befragte Personen pro Jahr
SHP_I	1999 – 2003	5'074	7'799
SHP_II	2004 – 2012	2'538	3'654
SHP_III	2013 – 2019	3'989	6'090

Die Befragungen des SHP werden meist per Telefon durchgeführt und bestehen aus drei Fragebögen zur Haushaltszusammensetzung ("household composition"), dem Haushalt selbst und den Personen. Sämtliche Mitglieder eines Haushaltes älter als 14 Jahre sind berechtigt, den individuellen Fragebogen zu beantworten. Zusätzlich existiert ein Stellvertretungs-Fragebogen für Mitglieder des Haushaltes unter 14 Jahren, abwesende Mitglieder oder Mitglieder, die aufgrund von Krankheit oder Beeinträchtigung nicht in der Lage sind für sich selbst zu antworten. (Tillmann et al., 2020)

Pro Welle werden ca. 200 Variablen pro Haushalt (1999: 171 Variablen, 2019: 225 Variablen) bewertet und ca. 500 Variablen pro Person (1999: 430 Variablen, 2019: 541 Variablen). Diese Variablen geben Aufschluss zu Themen wie Wohnort, Ausbildung, Alter, Herkunft, Lebenssituation, Lebensstandard, politische und religiöse Einstellung, Arbeitssituation, Einkommen, sozialem Umfeld, Gesundheit usw.

Variablen zur spezifischen Lebenslage sowie zu verschiedenen Rahmenbedingungen am Arbeitsplatz und im Privatleben liefern gute Indikatoren für die entsprechenden Spähern dieser Master-Thesis. Die adäquate, datenbasierte Bewertung der Arbeitsbeanspruchung wird bei der Behandlung der ersten Teilfrage im Detail behandelt.

Der SHP Datensatz bildet insgesamt eine passende Datenquelle für die Untersuchung der Forschungsfrage, da sowohl Quantität als auch Qualität der Daten einem hohen Anforderungsniveau entsprechen.

– Teil 2 –

Analyse und Modellierung

5 Untersuchungsszenario

Wie lässt sich der Zusammenhang zwischen der *Arbeitsbeanspruchung* einer Gruppe von Personen und der individuellen, *spezifischen Lebenslage* sowie Stressoren aus (Erwerbs- und Sorge-)*Arbeitsbelastung* datenbasiert modellieren und bewerten?

Zur Beschreibung von Zusammenhängen verwendet ein Grossteil der gängigen Modellierungsmethoden [12] eine (Regressions-) Gleichung der Form $y \sim x_1 + x_2 + \ ... \ + x_n$. Diese verknüpft eine abhängige Variable y mit einer Menge von n unabhängigen Variablen x_i, wodurch unweigerlich:

1. sämtliche x_i definiert und
2. sämtliche Zusammenhänge zwischen den x_i bestimmt werden

Mit der Angabe einer Regressionsgleichung definiert die forschende Person folglich ihr Untersuchungsszenario. In der Praxis werden oft mehrere Untersuchungsszenarien resp. Regressionsgleichungen erstellt und die daraus errechneten Modelle bezüglich ihrer Erklärungskraft verglichen. Im Kontext der vorliegenden Fragestellung scheint die Definition mehrerer Untersuchungsszenarien sinnvoll. So könnten vermutete Zusammenhänge zwischen den unabhängigen Variablen variiert und ein "Nullszenario" entsprechend weiterentwickelt werden, um einen fortlaufenden Prozess der explorativen Hypothesengenerierung zu forcieren. Die Betrachtung mehrerer Untersuchungsszenarien ist dagegen hinderlich für den objektiven Vergleich von Machine Learning Methoden, da:

- der Fokus der Untersuchungen zu sehr auf inhaltliche Fragestellungen aus der Arbeitswissenschaft gelegt wird
- die Gefahr inhaltlicher Verzettelung aufgrund wachsender Komplexität deutlich vergrössert wird

Aus diesen Gründen wird bewusst ein möglichst repräsentatives Untersuchungsszenario forciert, das sämtliche Sphären (*spezifische Lebenslage*, *Arbeitsbelastung* und *Arbeitsbeanspruchung*) berücksichtigt. Dies erlaubt eine gesamtheitliche Betrachtung der sozialwissenschaftlichen Ansprüche und den gleichzeitigen Methodenvergleich auf der gewünschten Flughöhe.

Die Auswahl relevanter unabhängiger Variablen ist ein anspruchsvoller Prozess, bei dem sozialwissenschaftliches Wissen und Erfahrung einen grossen Mehrwert bieten. In einem iterativen Prozess innerhalb des SNF-Forschungsprojekts wurden sozialwissenschaftliche sowie datenbezogene Aspekte berücksichtigt und dadurch 18 relevante, unabhängige Variablen identifiziert. Als abhängige Variable wurde das Merkmal $depression$ identifiziert. Dieses wird für die geplanten Untersuchungen als am besten geeignet angesehen, um die Sphäre *Arbeitsbeanspruchung* repräsentativ abzubilden. Auf den detaillierten Auswahlprozess wird an dieser Stelle nicht weiter eingegangen. Abbildung 6 zeigt die verwendeten Variablen inkl. deren Zuordnung zu den Sphären *spezifische Lebenslage*, *Arbeitsbelastung* und *Arbeitsbeanspruchung*. Die Namen der einzelnen Variablen sind selbsterklärend. Weitere Erklärungen zur Verwendung der Variablen werden in Kapitel 6.1 "Import und Operationalisierung" behandelt.

[12] z.B. Regressionsanalyse, Baummodelle, Support Vector Machines, Neuronale Netzwerke

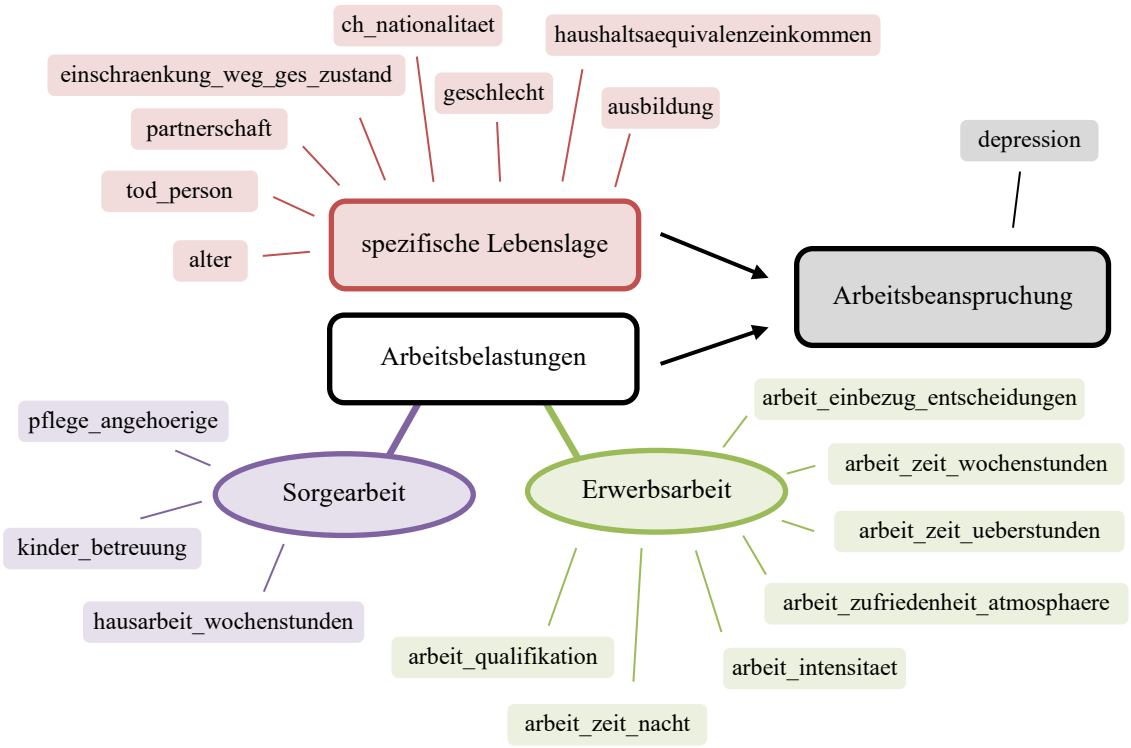


Abbildung 6: Merkmale des Untersuchungsszenarios inkl. Zurodnung zu Sphären

Die vollständige Definition des Untersuchungsszenarios erfordert die Festlegung der Zusammenhänge zwischen den unabhängigen Variablen. Regressionsmethoden erlauben im Allgemeinen die Integration komplexer Zusammenhänge in Form zusätzlicher Interaktionsterme ($x_i * x_j := x_i + x_j + x_i{:}x_j$). Da solche Interaktionsterme nicht für alle Methoden zulässig oder sinnvoll sind (z.B. für baumbasierte Methoden), wird das Untersuchungsszenario ohne Interaktionstherme definiert. Hinzu kommt, dass dem Verfasser keine Interaktionen bekannt sind, die gemäss sozialwissenschaftlicher Literatur a priori als gegeben angenommen werden sollten. Um zusätzliche Komplexität zu vermeiden, werden daher sämtliche unabhängigen Variablen als alleinstehende Terme definiert.

$$\begin{aligned} depression \sim\ & ausbildung + alter + geschlecht + \text{ch}\_nationalitaet \\ & + einschraenkung\_weg\_ges\_zustand + haushaltsaequivalenzeinkommen \\ & + partnerschaft + tod\_person + arbeit\_einbezug\_entscheidungen \\ & + arbeit\_qualifikation + arbeit\_zeit\_wochenstunden + arbeit\_zeit\_ueberstunden \\ & + arbeit\_zeit\_nacht + arbeit\_intensitaet + arbeit\_zufriedenheit\_atmosphaere \\ & + hausarbeit\_wochenstunden + kinder\_betreuung + pflege\_angehoerige \end{aligned} \quad (5)$$

Formel (5) definiert damit das Untersuchungsszenario dieser Master-Thesis. Sämtliche Methoden werden anhand dieses Untersuchungsszenarios bewertet und verglichen.

6 Datenaufbereitung

6.1 Import und Operationalisierung

Im Folgenden werden die Annahmen und daraus folgende Eingrenzungen für den Import des SHP-Paneldatensatzes erläutert. Zusätzlich wird die Operationalisierung (Transformation numerischer Variablen und Codierung von Faktorvariablen) beschrieben, die das Untersuchungsszenario auf möglichst repräsentative Weise abzubilden versucht.

Sämtliche Variablen des SHP-Datensatzes können eindeutig einer Person oder einem Haushalt zugeordnet werden. Die Zuordnung von Personen zu Haushalten ist ebenfalls eindeutig. Für das definierte Untersuchungsszenario werden ausschliesslich Personen als Untersuchungseinheiten betrachtet, weshalb Merkmale auf Ebene Haushalt wie z.B. *haushaltsaequivalenzeinkommen* auf die Personen des entsprechenden Haushaltes übertragen werden.

Zeitliche Eingrenzung

Da für unterschiedliche Erhebungszeiträume die Anzahl befragter Personen variiert (vgl. Kapitel 4.2, Tabelle 3) und die Anzahl erhobener Merkmale ebenfalls Schwankungen aufweist, wird ausschliesslich der Zeitraum von 2004 bis 2019 betrachtet. Diese Einschränkung erlaubt eine konsistentere Betrachtung von Personen und erhöht die Balanciertheit (engl. *balancedness*) des importierten Paneldatensatzes.

Räumliche Eingrenzung

Eine räumliche Eingrenzung wird nicht vorgenommen.

Demographische Eingrenzung

Aufgrund der inhaltlichen Betrachtung von Arbeitsbelastungen im Erwerbsleben, werden ausschliesslich Personen zwischen 15 und 65 Jahren betrachtet. Personen, die im zeitlichen Verlauf eine dieser Altersgrenzen überschreiten, werden ebenfalls berücksichtigt - es werden jedoch ausschliesslich Messungen innerhalb der Altersgrenzen berücksichtigt. Dies vergrössert die Datengrundlage, vermindert jedoch die Balanciertheit des importierten Datensatzes in geringem Masse.

Arbeitsspezifische Eingrenzung

Wegen der Betrachtung von Arbeitsbelastungen im Erwerbsleben, werden die Personen zusätzlich nach dem Beschäftigungsgrad selektioniert. Dabei werden ausschliesslich Personen berücksichtigt, die einen Status von 1, 2, 3, 5 oder 6 aufweisen:

1. Bezahlte Erwerbstätigkeit, Vollzeit (reguläre Arbeitszeit 37 Stunden pro Woche oder mehr)
2. Bezahlte Erwerbstätigkeit, Teilzeit (reguläre Arbeitszeit 5-36 Stunden pro Woche)
3. Bezahlte Erwerbstätigkeit, Teilzeit (1 - 4 Stunden / Woche)
4. In Ausbildung (Lehrling, Schüler/In, Student/In)
5. Mitarbeit im Familienbetrieb
6. Arbeit in geschützter Werkstatt (für Beeinträchtigte, Personen mit Problemen)
7. Kind/Frau/Mann im Haushalt (nur bis maximal 64 bzw. 65 Jahre)
8. Rentner/Innen (AHV)
9. Rentner/Innen (IV usw.)

10. Arbeitslos
11. Andere Tätigkeit (Weiterbildung, unbezahlter Urlaub).

Operationalisierung

Um Faktorvariablen in geeigneter Granularität verwenden zu können, werden die Levels einiger Faktorvariablen neu codiert. Die Motivation hinter diesen Neucodierungen ist die Annahme, dass zu viele Ausprägungen einer Faktorvariable grobe Zusammenhänge bei der Modellierung verschleiern. Aus diesem Grund werden Faktorvariablen teilweise in grösseren Gruppen zusammengefasst. Numerische Variablen werden teilweise aus bestehenden Variablen erzeugt oder transformiert. Eine Übersicht der vorgenommenen Anpassungen während dem Import befindet sich in Anhang A.

Das importierte Paneldatensatz ist somit in der Lage, die drei Sphären *spezifische Lebenslage*, *Arbeitsbelastung* und *Arbeitsbeanspruchung* adäquat darzustellen und die entsprechenden unabhängigen Variablen aus dem Untersuchungsszenario verfügbar zu machen. Weitere Informationen zum Datensatz folgen in Kapitel 6.3 "Explorative Datenanalyse".

6.2 Skalierung und Randomisierung

Bei der Anwendung verschiedener Machine Learning Methoden kann die Skalierung numerischer Variablen sinnvoll sein, damit keine Überbewertung einzelner Merkmale aufgrund grosser Zahlenwerte induziert wird. Die Skalierung eines Datensatzes hat im Allgemeinen den inhaltlichen Nachteil, dass sie die Interpretierbarkeit von Variablen, Effekten und Interaktionen erschwert. Dieser Nachteil hat jedoch zweitrangigen Charakter, da die optimale Verwendung von Machine Learning Methoden grundsätzlich vorgeht. Der vorliegende Paneldatensatz wird dennoch *nicht* skaliert[13], weil:

1. die Vergleichbarkeit verschiedener Methoden erschwert wird, sobald die Daten skaliert werden[14].
2. der Einfluss der Skalierung gering ist, da sämtliche numerische Variablen des vorliegenden Datensatz auf ähnlichen Grössenordnungen liegen[15].

Ausserhalb des Vergleich-Kontextes dieser Master-Thesis sollte die forschende Person eine Skalierung der Daten in Betracht ziehen, oder zumindest deren Verwendung oder Vernachlässigung begründen.

Um am Ende eines Modellierungsprozesses inhaltlich repräsentative Schlüsse ziehen zu können, muss der vorliegende Paneldatensatz in genügendem Masse randomisiert sein. Aufgrund der vorliegenden Datenbasis ist die externe Validität auf die Schweizer Bevölkerung eingegrenzt. Innerhalb dieser Rahmenbedingung sollte sichergestellt sein, dass die Schweizer Bevölkerung auf adäquate Weise in den

[13] Einzige Ausnahme bilden Machine Learning Methoden, die explizit eine Skalierung voraussetzen.
[14] Es kann nicht davon ausgegangen werden, dass sämtliche angewandten Methoden die Skalierung von Daten in gleicher Weise berücksichtigen. Einige Methoden nehmen für optimale Performance bereits selbstständig Skalierungen vor. Somit könnten Verzerrungen der Ergebnisse erzeugt werden, die nicht oder nur schwierig rückverfolgbar sind.
[15] Ein Vergleich der Modellierung mit den unveränderten Paneldaten und den normalisierten Paneldaten hat gezeigt, dass die erzeugen Koeffizienten von Pooling-, FE- und RE-Modell beinahe deckungsgleich sind.

Daten repräsentiert ist. Da der SHP-Datensatz als geschichtete Zufallsstichprobe (engl. *stratified random sampling*) erhoben wird (Antal & Rothenbühler, 2015), gehen wir in dieser Master-Arbeit von einem ausreichenden Masse der Randomisierung aus. Während des Importprozesses findet nebst der gewollten Selektion nach Zeit, Demographie und Arbeitsstatus eine ungewollte Selektion von Individuen aufgrund datenbezogener Eigenschaften statt. Da der ungewollte Selektionsprozess nicht *per se* zufällig ist, muss damit gerechnet werden, dass die Randomisierung in geringem Masse verletzt wird. Diese Unschärfe wird für die kommenden Untersuchungen hingenommen.

6.3 Explorative Datenanalyse

Zu Beginn eines Modellierungsprozesses stellt die Datenanalyse einen wichtigen Vorbereitungsschritt dar, um das allgemeine Verständnis der Daten zu fördern. In den folgenden Unterkapiteln werden verschiedene Methoden zur explorativen Datenanalyse vorgestellt und an Beispielen anhand des vorliegenden Paneldatensatzes demonstriert. Am Ende wird auf den Mehrwert solcher Analysen im Allgemeinen und anhand der Beispiele eingegangen.

6.3.1 Überblick

Der vorliegende Paneldatensatz besteht aus insgesamt 53'473 Beobachtungen. Der Datensatz enthält die abhängige Variable $depression$, 18 unabhängige Variablen (vgl. Kapitel 5) und die drei Variablen $year$, id (Identifikation einer Person) und $person\_haushalt$ (Identifikation eines Haushaltes). Sämtliche Beobachtungen können somit insgesamt 5'694 Personen und 4'380 Haushalten zugeordnet werden. Die Beobachtungen decken gemäss Import jährliche Messungen zwischen 2004 und 2019 ab.

Der Paneldatensatz ist nicht balanciert, d.h. es existieren nicht gleich viele zeitliche Beobachtungen pro Person. Die Balanciertheit (engl. *balancedness)* nach Ahrens & Pincus (1981) wird mit den Werten $\gamma = 0.66$ und $\nu = 0.79$ bemessen[16]. Abbildung 7 zeigt, dass nur ca. 19% aller Personen (1'087 von 5'694) über den maximal möglichen Zeitraum von 16 Jahren eine Beobachtung besitzen. Diese Tatsache wird für die kommenden Analysen hingenommen.

[16] Ein Wert von 1 entspricht bei beiden Massen einem balancierten Panel. Je näher die Werte gegen null streben, desto mehr " unbalanciert" ist das Panel. Siehe auch R-Funktion: `plm:: punbalancedness()`

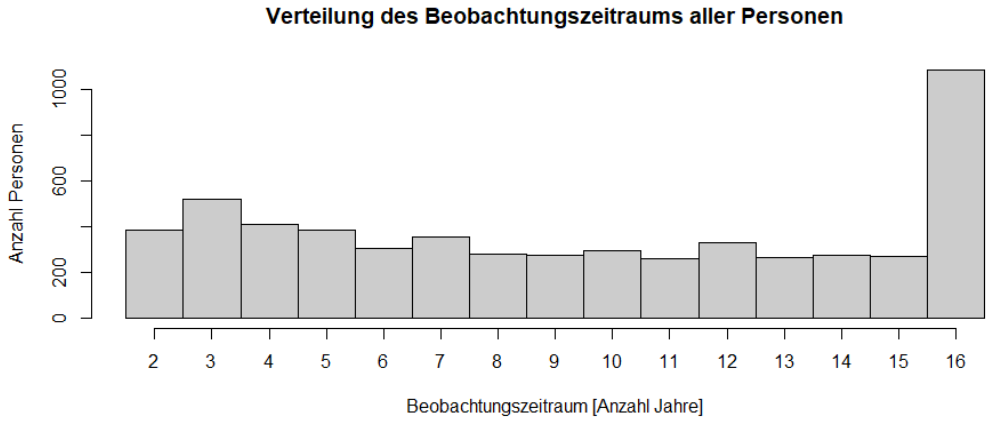


Abbildung 7: Verteilung des Beobachtungszeitraums aller Personen

Von den insgesamt 1'229'879 Messpunkten (53'473*23) sind ca. 10% ungültig (NA's). Eine Übersicht ungültiger Messwerte pro Variable befindet sich in Anhang B. Die verwendeten Methoden der Paneldatenanalyse verwenden in R/RStudio die Option `na.action = na.omit` wodurch Beobachtungen mit ungültigen Werten vernachlässigt werden. Dadurch werden für die meisten Modelle nur 24'053 der insgesamt 53'473 Beobachtungen verwendet.

Von den 22 Variablen sind 12 numerisch und 10 als Faktorvariablen codiert. Die Ausprägungen der numerischen Variablen folgen im Allgemeinen einer Normalverteilung oder Poissonverteilung. Eine Übersicht sämtlicher Verteilungen findet sich im Anhang C.

Es zeigt sich, dass die abhängige Variable "depression" vielmehr einer Poissonverteilung als einer Normalverteilung folgt. Dies scheint durchaus sinnvoll, da wir erwarten, dass eine durchnittliche Person wenige Herausforderungen bewältigen muss, die sich direkt in der "depression" niederschlagen würden. Für die Modellierung stellt diese Tatsache eine Hürde dar, weil viele Panelmethoden gewöhnliche OLS-Schätzer verwenden, die implizit eine normalverteilte abhängigen Variable annehmen. Die Verwendung von verbesserten GLS-Schätzern (Generalized least squares) wird in den späteren Kapiteln diskutiert. Für Modelle, welche eine normalverteilte abhängige Variable annehmen wird diese Unschärfe hingenommen.

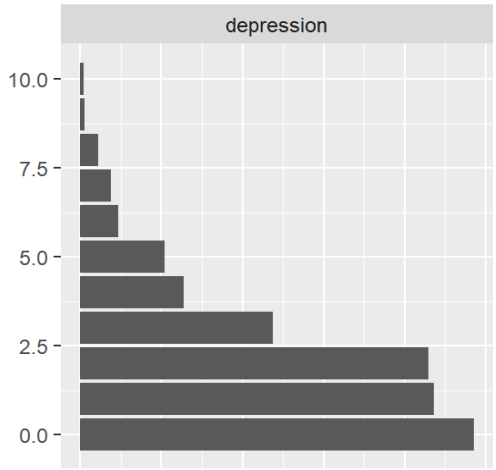


Abbildung 8: Relative Verteilung der abhängigen Variable "depression"

Die zeitliche Entwicklung von "depression" flutktuiert teilweise stark über die Zeit. Abbildung 9 gibt eine Intution für diesen Sachverhalt und zeigt deutlich, dass nicht jede Person über den gesamten Zeitraum von 2004 bis 2019 gültige Messpunkte besitzt.

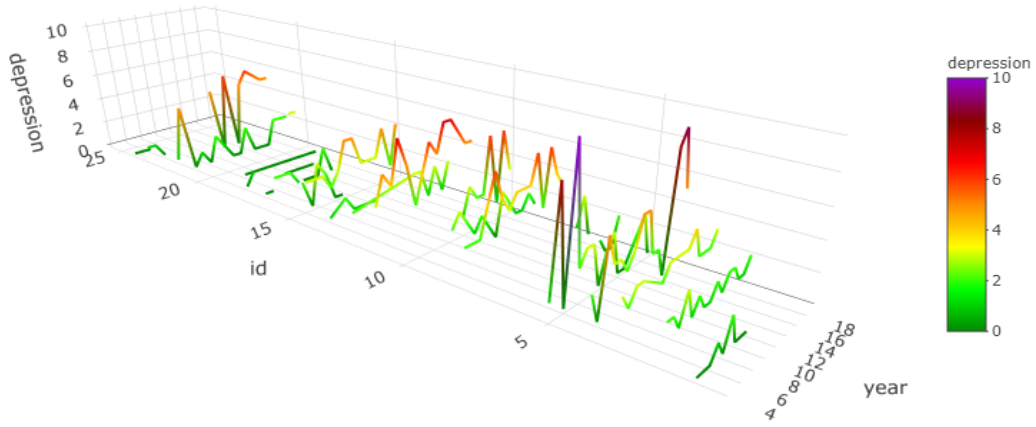


Abbildung 9: Zeitliche Entwicklung von "depression" der ersten 25 Personen

6.3.2 Multikollinearität

Die Aussagekraft von linearen Regressionsmodellen wird durch korrelierende, unabhängige Variablen negativ beeinflusst. Ist die Korrelation zwischen zwei oder mehreren unabhängigen Variablen gross, werden die Schätzungen von Effekten ineffizient, wodurch diese ihre Aussagekraft verlieren können. (Farrar & Glauber, 1967)

Abbildung 10 zeigt die Korrelationsstruktur der numerischen, unabhängigen Variablen im vorliegenden Datensatz. Nebst unsinnigen Korrelationen, wie diejenige zwischen "`id`" (Person) und "`person_haushalt`" (Haushalt), sind folgende durchaus wichtigen Strukturen hervorzuheben:

1. `Cor(arbeit_zeit_wochenstunden, arbeit_intensitaet) = 0.29`
2. `Cor(arbeit_zeit_wochenstunden, arbeit_zeit_ueberstunden) =` 0.38
3. `Cor(arbeit_zeit_wochenstunden, hausarbeit_wochenstunden) = -0.48`

Personen, die grundsätzlich viel Erwerbsarbeit leisten (hohe Anzahl Wochenstunden) neigen zu einer ebenfalls hohen Arbeitsintensität mit vielen Überstunden und gleichzeitig weniger Arbeitsstunden im Haushalt. Diese Erkenntnis ist sinnvoll, da sie eine intuitiv nachvollziehbare Gegebenheit quantitativ belegt. Für kommende Modellierungen wird bewusst darauf verzichtet, einzelne dieser Variablen aus Korrelationsgründen zu vernachlässigen. Eine Sammlung der entsprechenden Korrelationswerte befindet sich in Anhang D.

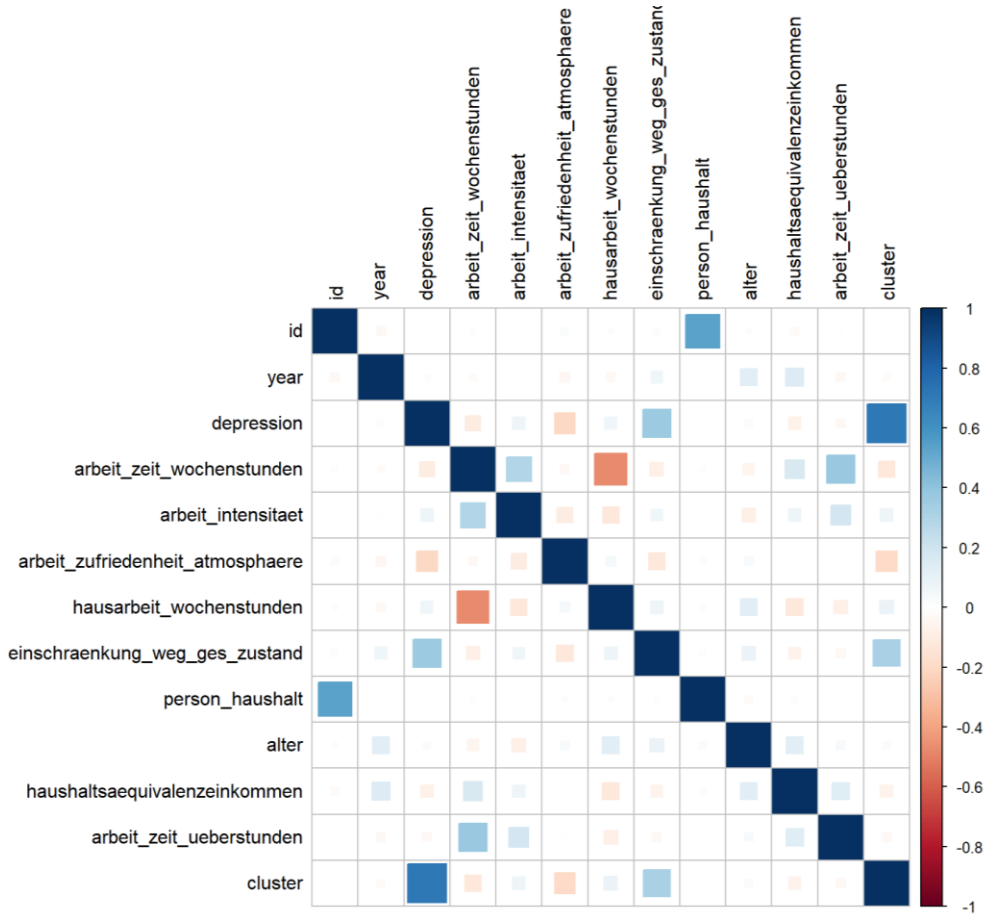


Abbildung 10: Korrelationsstruktur der numerischen unabhängigen Variablen

6.3.3 Varianz

Zur Untersuchung der Varianz einzelner Merkmale eines Paneldatensatzes wird der ursprüngliche Begriff von Varianz [17] erweitert. Aufgrund der intrinsischen Zuordnung einzelner Beobachtungen zu Untersuchungseinheiten, kann für Paneldaten untersucht werden, welcher Anteil der gesamten Varianz durch Unterschiede zwischen Untersuchungseinheiten (*between*-Varianz) und durch Unterschiede innerhalb von Untersuchungseinheiten (*within*-Varianz) zustande kommen. Aus diesen Überlegungen geht der Intraclass Correlation Coefficient (IC) hervor, der das Verhältnis von *within*- oder *between*-Varianz zur gesamten Varianz darstellt (Campbell et al., 2001). Tabelle 4 zeigt den ICC der abhängigen Variablen $depression$. Dabei stellt Zeile "id" den Anteil der *between*-Varianz und Zeile "Residual" den Anteil der *within*-Varianz dar. Daraus ist erkennbar, dass die Varianz von Depression zwischen den Untersuchungseinheiten einen leicht grösseren Anteil zur Gesamtvarianz beiträgt (56%) als die Varianz innerhalb von Untersuchungseinheiten (44%). Da der ICC eine gemittelte Sicht über alle

[17] Im herkömmlichen Sinne bezeichnet die empirische Varianz das Mass der Streuung um den Mittelwert einer Population $i \in \{1,2,\dots,N\}$ als: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$, wobei $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$.

Beobachtungseinheiten darstellt, sei darauf hingewiesen, dass trotzdem Untersuchungseinheiten einer *within*-Varianz von null vorkommen können.

Tabelle 4: ICC für "depression"

Variable	Sigma	ICC
id	2.15	0.56
Residual	1.67	0.44

Die visuelle Aufbereitung solcher Varianzeigenschaften bietet eine intuitive Möglichkeit, die Zusammensetzung eines Paneldatensatzes besser zu verstehen. Als Beispiel werden die zuvor dargelegten Varianzeigenschaften von *depression* in Abbildung 11 und Abbildung 12 visuell dargestellt. Zur Erstellung dieser Abbildungen wird der individuelle Mittelwert pro Untersuchungseinheit bestimmt (= *between*-Komponente) und sämtliche Beobachtungen einer Untersuchungseinheit um diesen Mittelwert zentriert (= *within*-Komponente). Der Vergleich mit einer optimal gefitteten Normalverteilung[18] erlaubt die Identifikation von Ausreissern sowie die Untersuchung der *between*- oder -*within*-Komponenten auf Normalität mittels Q-Q-Plots.

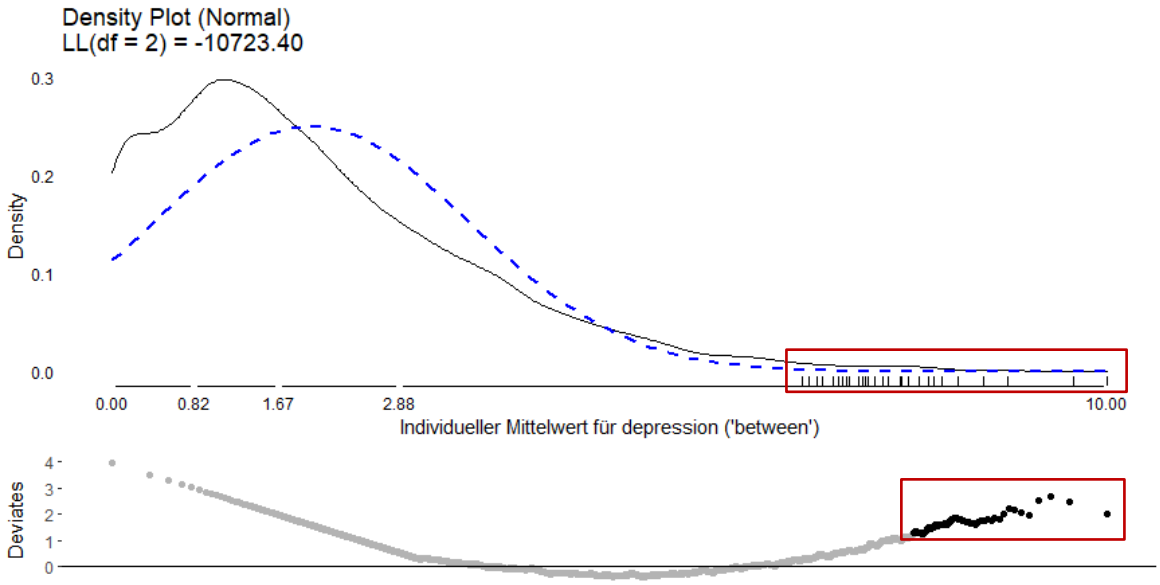


Abbildung 11: Dichteverteilung der mittleren Depression inkl. Testvergleich mit Normalverteilung (blau) und entsprechendem Q-Q-Plot (unten). Das Signifikanzniveau für Ausreisser (rot) liegt bei 0.001.

Abbildung 11 bestätigt die bereits aus Abbildung 8 gewonnene Erkenntnis, dass die abhängige Variable nicht *per se* normalverteilt ist. Zusätzlich können nun Untersuchungseinheiten identifiziert werden, die auf einem 0.001-Signifikanzniveau der gefitteten Normalverteilung als Ausreisser gelten. Eine weiterführende Analyse dieser Teilgruppe mit hohen mittleren Werten für Depression wäre ein sinnvoller Ansatzpunkt für weitere Erkenntnisse bezüglich Depression.

[18] Die hier verwendete R-Funktion `JWileymisc::testDistribution()` erlaubt den Vergleich mit weiteren Verteilungen wie z.B. Beta, Chi-Square, Gamma , Binomial, Poisson etc.

Abbildung 12 zeigt, dass die individuellen Verläufe von Depression grundsätzlich symmetrisch um den individuellen Mittelwert einer Untersuchungseinheit verlaufen. Die Verteilung selbst zeigt im Vergleich zu einer Normalverteilung eine überproportionale Anhäufung um den individuellen Mittelwert. Dies lässt vermuten, dass individuelle Verläufe von Depression eine Tendenz zur "Konstanthaltung" aufweisen. Die untersuchten Personen neigen folglich dazu, weniger von ihrer individuellen Norm abzuweichen als man in einem normalverteilten Prozess erwarten würde. Trotzdem lassen sich zahlreiche Ausreisser auf dem 0.001-Signifikanzniveau identifizieren. Diese bieten wiederum interessante Ansatzpunkte, um spezifische Untergruppen auf Eigenheiten der individuellen Verläufe von Depression zu untersuchen. An dieser Stelle wird jedoch davon abgesehen.

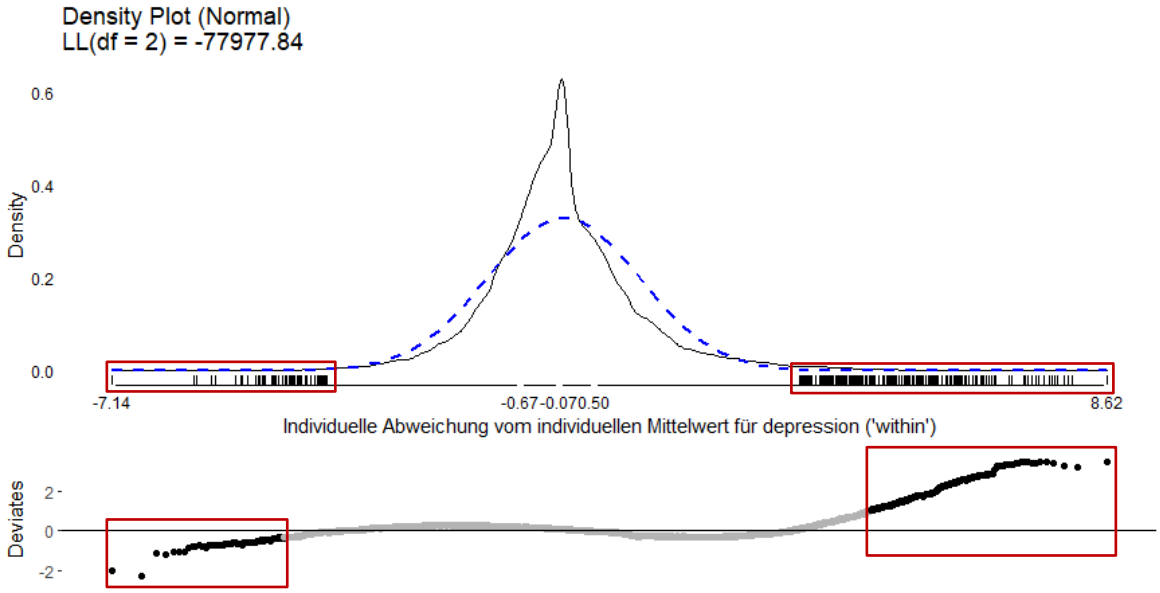


Abbildung 12: Dichteverteilung der Abweichung von der mittleren Depression inkl. Testvergleich mit Normalverteilung (blau) und entsprechendem Q-Q-Plot (unten). Das Signifikanzniveau für Ausreisser liegt bei 0.001.

Die Analyse der Varianz numerischer Merkmale bietet ein einfaches und doch mächtiges Werkzeug bei der explorativen Analyse eines Paneldatensatzes. Die entsprechenden Untersuchungen der restlichen numerischen Merkmale anhand von ICC und Verteilung von *between*- und *within*-Komponenten befindet sich in Anhang E.

6.3.4 Mehrdimensionale Verteilungen

Nebst eindimensionalen Verteilungen, welche die Analyse von Lage- und Streumassen einer Variable erlauben, bieten mehrdimensionale Verteilungen eine effiziente Methode zur Darstellung von Zusammenhängen zwischen Variablen. Im Folgenden werden zwei Anwendungen mehrdimensionaler Verteilungen anhand konkreter Beispiele vorgestellt.

Bedingte Verteilungen

Im Fall des vorliegenden Paneldatensatzes liegt eine abhängige Variable $Y = depression$ mit wenigen diskreten Ausprägungen $y \in \{0,1, \dots 10\}$ vor. Diese Tatsache erlaubt den niederschwelligen Vergleich

sämtlicher Ausprägungs-Gruppen y in Bezug auf eine zweite Variable X. Dabei bietet die bedingte Verteilung $f(x|y)$ die Grundlage eines solchen Vergleichs[19].

$$f(x|y) = \frac{P(X = x, Y = y)}{P(Y = y)} \tag{6}$$

In Abhängigkeit der Ausprägungen von X werden drei verschiedene Darstellungsarten vorgestellt, die fallspezifisch einen intuitiven Vergleich sämtlicher Ausprägungs-Gruppen y erlauben.

Fall 1: X ist ein nominales, ordinales oder diskretes Merkmal mit wenigen Ausprägungen.

Für solche Situationen bietet sich die Darstellung der bedingten Verteilung als "gehäufte Anteile" an. Dabei wird pro Ausprägung y der relative Anteil der Ausprägungen x in einem Balkendiagramm dargestellt. Die Länge eines Balkens entspricht so immer 1 und sämtliche Gruppen von Y sind vergleichbar. Die bedingte Verteilung $f(partnerschaft|depression)$ in Abbildung 13 zeigt auf, dass der Anteil der Singles höher ist bei höherem Level der Depression. Daraus kann geschlossen werden, dass ein möglicher Zusammenhang zwischen dem Partnerschaftsstatus und Depression existiert. Ob nun ein direkter Zusammenhang vom Partnerschaftsstatus auf Depression vorhanden ist, kann aufgrund dieser Darstellung jedoch nicht geschlossen werden, da versteckte Drittvariablen[20] im Spiel sein könnten.

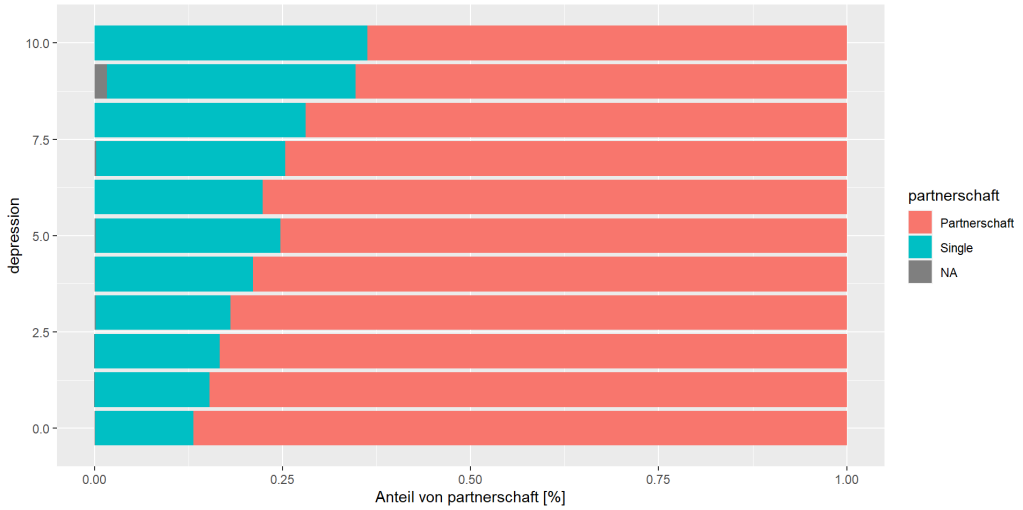


Abbildung 13: Bedingte Verteilung von Partnerschaft gegeben Depression

Fall 2: X ist ein nominales, ordinales oder diskretes Merkmal mit vielen Ausprägungen.

Darstellungen gemäss Fall 1 neigen zu Unübersichtlichkeit, sobald X viele Ausprägungen aufweist. In diesen Situationen können die relativen Anteile einer Gruppe y als separate Balken dargestellt werden. Abbildung 14 zeigt ein entsprechendes Beispiel für $f(einschraenkung\_weg\_ges\_zustand|depression)$. Die Summe sämtlicher Balken einer Depressions-Gruppe (z.B. $y = 0$) entspricht wiederum 1. Eine mögliche Interpretation dieser Grafik ist, dass Personen mit tiefen Werten für Depression öfters in der Gruppe mit wenigen Einschränkungen aufgrund des Gesundheitszustandes vorkommen als in der Gruppe mit hohen Einschränkungen aufgrund des Gesundheitszustandes. Für Personen mit hohen Werten für

[19] Umgangssprachliche Definition von $f(x|y)$: *Die bedingte Verteilung von X, gegeben Y = y.*
[20] engl. *confounder*

Depression scheint diese Tendenz umgekehrt zu sein. Dies ist folglich ein Indiz dafür, dass die Einschränkung aufgrund des Gesundheitszustands in einem Zusammenhang mit Depression steht.

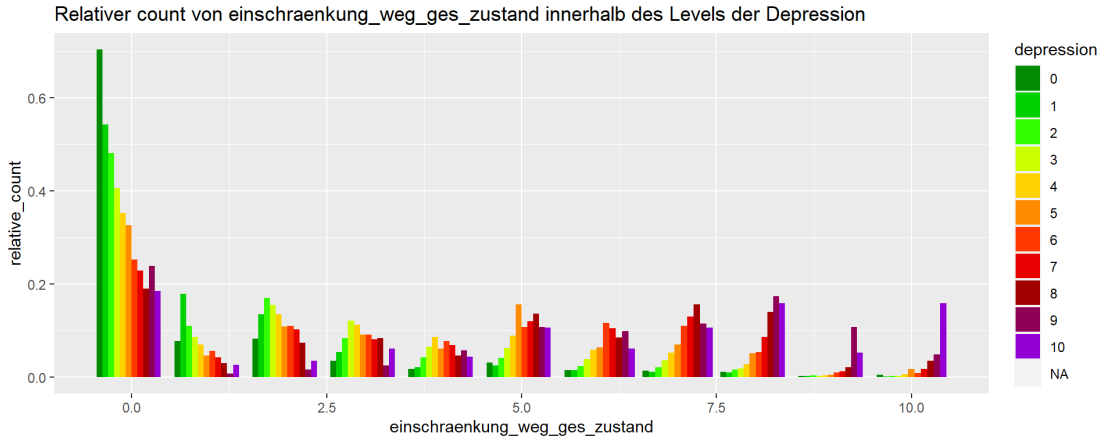


Abbildung 14: Bedingte Verteilung von "Einschränkung wegen Gesundheitszustand" gegeben Depression

Fall 3: X ist ein nominales, ordinales, diskretes Merkmal mit sehr vielen Ausprägungen oder ein oder stetiges Merkmal.

Darstellungen gemäss Fall 2 neigen zu Unübersichtlichkeit, sobald X sehr viele Ausprägungen (>20) aufweist oder ein stetiges Merkmal ist. In diesen Situationen können die relativen Anteile einer Gruppe y als Dichteverteilung dargestellt werden. Abbildung 15 zeigt ein entsprechendes Beispiel für $f(haushaltsaequivalenzeinkommen|depression)$ Die Fläche unter der Dichtefunktion einer Depressions-Gruppe (z.B. $y = 0$) entspricht wiederum 1. Aus der Grafik lässt sich schliessen, dass Personen mit hohen Werten für Depression ein tendenziell tieferes Haushaltsäquivalenzeinkommen besitzen als Personen mit tiefen und mittleren Werten für Depression. Die forschende Person sollte dabei den zugrundeliegenden Datengenerierungsprozess einer solchen Grafik stets im Hinterkopf halten. Aus Abbildung 8 (Relative Verteilung der abhängigen Variable "depression") ist bekannt, dass die Anzahl der Beobachtungen mit hohen Werten für Depression klein ist und damit die Unsicherheiten der gezeichneten Dichteverteilungen entsprechend grösser sind.

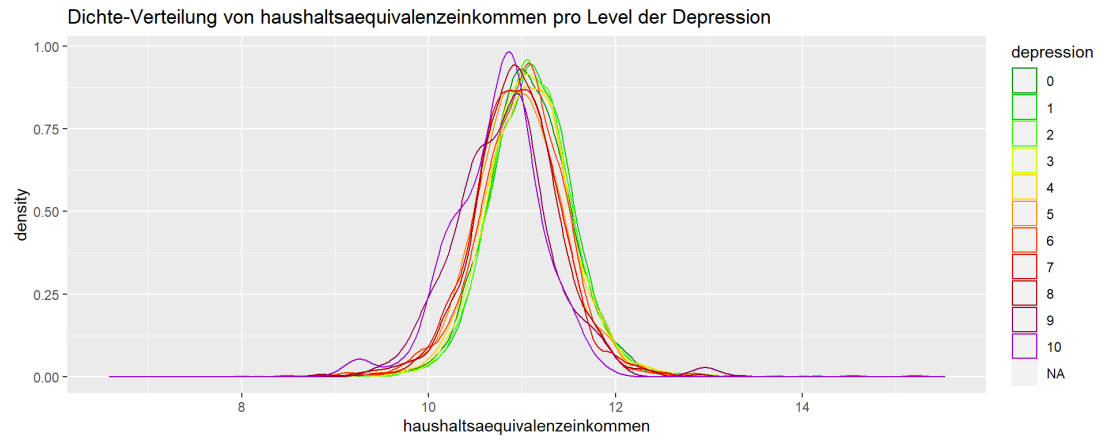


Abbildung 15: Bedingte Dichteverteilung von Haushaltsäquivalenzeinkommen gegeben Depression

Sämtliche Analysen zu bedingten Verteilungen der unabhängigen Merkmale befinden sich in Anhang F.

zweidimensionale Verteilungen

Zur gleichzeitigen Betrachtung des Zusammenhangs zwischen abhängiger Variable Y und zwei unabhängigen Variablen X_1 und X_2, bieten zweidimensionale Verteilungen eine sinnvolle Darstellungsart für den raschen Erkenntnisgewinn[21]. Dabei werden die unabhängigen Variablen in diskrete Gruppen eingeteilt und die abhängige Variable nach einer gewünschten Metrik (z.B. Mittelwert, Median, Minimum, Maximum) pro Gruppe zusammengefasst. Abbildung 16 (unten) zeigt den Mittelwert von $Y = depression$ nach den zweidimensionalen Gruppen $X_1 = einschraenkung\_weg\_ges\_zustand$ und $X_2 = ausbildung$. Es zeigt sich, dass Kombinationen der beiden unabhängigen Variablen unterschiedliche Werte der Depression (resp. unterschiedliche horizontale oder vertikale Verläufe) hervorbringen. Somit kann geschlossen werden, dass mögliche Interaktionen dieser zwei Terme durchaus vorhanden sind. Auch für diese Art summarischer Darstellung sollte die forschende Person den zugrundeliegenden Datengenerierungsprozess und damit verknüpfte Unsicherheiten verstehen und bei der Interpretation berücksichtigen. Beispielsweise zeigt Abbildung 16 (oben) einen abnehmenden Umfang der Datenbasis für grössere Werte von $X_1 = einschraenkung\_weg\_ges\_zustand$.

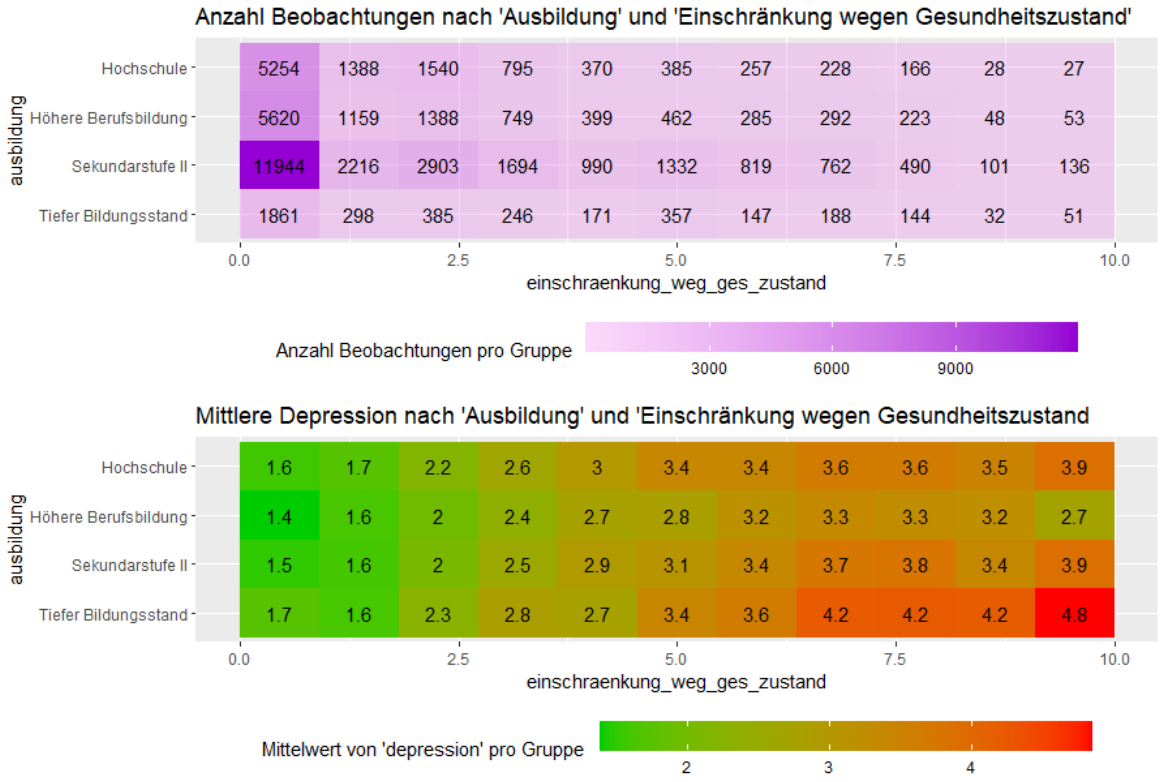


Abbildung 16: zweidimensionale Verteilungen nach diskreten Gruppen für "Ausbildung" und "Einschränkung wegen Gesundheitszustand" (oben: Anzahl Beobachtungen pro Gruppe, unten: Mittelwert von Depression pro Gruppe)

[21] Durch die explizite Darstellung der abhängigen Variable als dritte Dimension (anstatt einer Farbe), könnten ebenfalls dreidimensionale Balkendiagramme erstellt werden.

6.3.5 Mehrwert explorativer Datenanalysen

Die Methoden der explorativen Datenanalyse bieten breite Möglichkeiten zur (Vor-)Untersuchung eines Paneldatensatzes. Die Betrachtung sämtlicher Beobachtungen als einzelne Messungen bringt erste Erkenntnisse über globale Zusammenhänge von Variablen. Dabei ist zu beachten, dass die Vernachlässigung der impliziten Panelstruktur[22] eine Verzerrung der beobachteten Zusammenhänge herbeiführen kann[23]. Unter Berücksichtigung der Panelstruktur können Zusammenhänge zwischen Untersuchungseinheiten (Querschnitt / *between*-Komponente) oder innerhalb der Untersuchungseinheiten (Längsschnitt / *within*-Komponente) genauer erforscht werden.

Obwohl bei der explorativen Datenanalyse – im Gegensatz zur Modellierung von Paneldaten - keine Effektstärken quantifiziert werden, sind potenzielle Effekte resp. Zusammenhänge erkennbar. Diese geben Indizien zu linearen oder nichtlinearen Abhängigkeiten und Interaktionen zwischen mehreren Variablen und damit brauchbare Hinweise für spätere Modellierungsprozesse[24]. Des Weiteren können im Kontext der Sozialwissenschaften Erkenntnisse aus dem Querschnitt sogar dafür verwendet werden, die spätere Modellierung im Längsschnitt zu variieren[25].

Nach wie vor sollte sich der Forschende der Tatsache bewusst sein, dass explorative Datenanalysen meist auf beschreibender Statistik und damit auf summarischen oder gemittelten Betrachtungen beruhen. Von Rückschlüssen auf individuelle Untersuchungseinheiten ist deshalb abzuraten.

Die hier vorgestellten Methoden zeigen einen kleinen Ausschnitt der breiten Möglichkeiten einer explorativen Datenanalyse. Der Mehrwert einer solchen Analyse, im Kontext von Paneldaten, ist aufgrund der obigen Beschreibungen eindeutig gegeben. Die explorative Datenanalyse stellt ein zentrales Element jedes Modellierungsprozesses dar. Sie sollte grundsätzlich immer durchgeführt werden und abhängig vom Detaillierungsgrad der gewünschten Erkenntnisse iterativ vertieft werden[26].

---

[22] Die Zuordnung von Beobachtungen zu Untersuchungseinheiten

[23] Untersuchungseinheiten mit vielen Beobachtungen erhalten automatisch eine stärkere Gewichtung als Untersuchungseinheiten mit wenigen Beobachtungen. Abbildung 3 gibt eine Intuition dieses Effekts.

[24] Die Aufdeckung von Interaktionseffekten wie in Abbildung 16 kann bspw. die Verwendung eines Interaktionsterms in einer Regressionsgleichung motivieren.

[25] In Abbildung 13 wurde im Querschnitt festgestellt, dass Personen mit hohen Werten für Depression vermehrt mit dem Partnerschaftsstatus "Single" in Verbindung stehen. Die Annahme, dass dieser Effekt ebenfalls im Längsschnitt vorhanden sein könnte, ist durchaus berechtigt (d.h. "*Personen, die vom Partnerschaftsstatus "Beziehung" auf "Single" wechseln, erleben im Mittel eine Erhöhung des Wertes für Depression*") und könnte bei der späteren Modellierung untersucht werden.

[26] Das Minimalbeispiel einer explorativen Datenanalyse kann bspw. eine 5-Punkte-Zusammenfassung sein. Bereits diese niederschwellige Methode gibt wichtige Erkenntnisse über einen Datensatz.

7 Modellierung

Im Folgenden werden sämtliche State-of-the-Art-Methoden sowie die fortgeschrittenen Methoden zur Paneldatenanalyse anhand des Untersuchungsszenarios am vorliegenden Paneldatensatz angewandt.

7.1 Gepooltes Modell

Das in Kapitel 4.1.2 beschriebene gepoolte Modell stellt für jeden Modellierungsprozess mit Paneldaten eine wichtige Grundlage dar. Die daraus gewonnenen Koeffizienten liefern eine erste Quantifizierung der gesuchten Effekte, wobei sowohl die zeitliche Struktur als auch die Zuordnung von Messungen zu Untersuchungseinheiten berücksichtigt wird. Die Koeffizienten des gepoolten Modells wiederspiegeln ein möglichst neutrales resp. gewichtetes Bild sämtlicher Effekte (optimale Platzierung der (n-1)-dimensionalen Hyperebene im n-dimensionalen Raum). Abbildung 17 zeigt die Schätzung der Effekte für das definierte Untersuchungsszenario und den importierten Paneldatensatz.

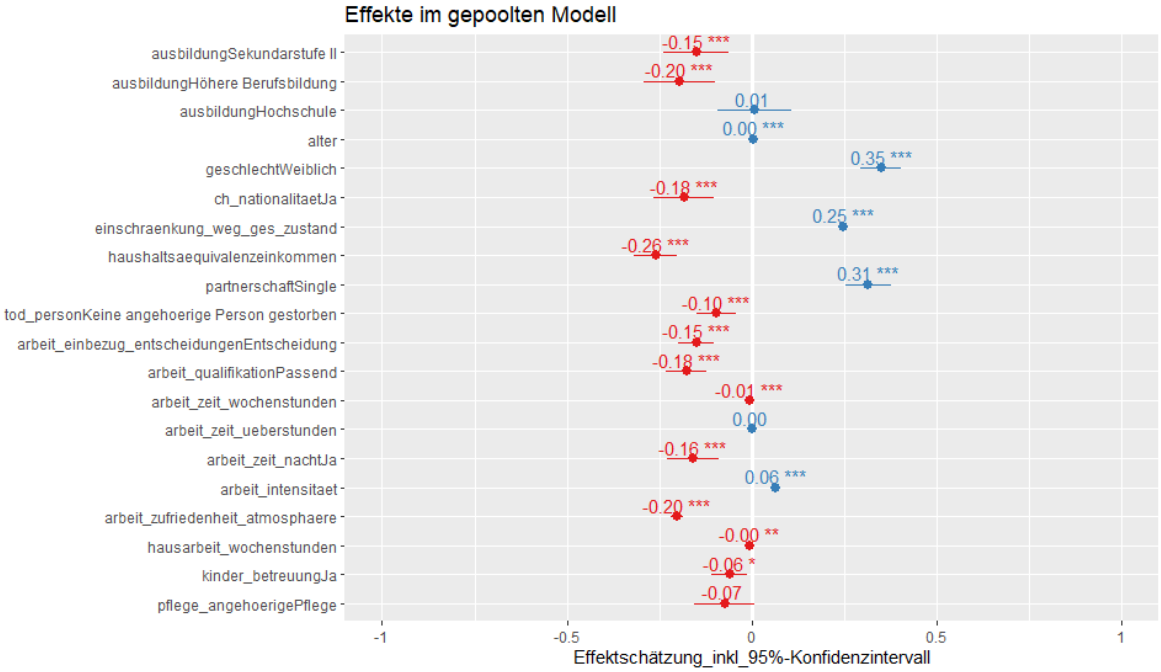


Abbildung 17: Effektschätzung im gepoolten Modell (blau > 0, rot < 0)

Der Effekt vieler unabhängiger Variablen wird auf dem 0.05-Signifikanzniveau als signifikant eingestuft. Das gepoolte Modell erklärt insgesamt 18% der Varianz in den Daten (R-Squared = 0.181, Adjusted R-Squared = 0.180) und kann grundsätzlich als signifikant eingestuft werden (p.value < 2.22e-16). Die Ausgaben von R/RStudio zum gepoolten Modell befinden sich in Anhang G.

Der Mehrwert dieses Modells liegt in der intuitiven Interpretierbarkeit und seiner neutralen resp. gemittelten Betrachtungsweise. Man könnte argumentieren, dass das Modell ungeeignet ist, da einheitenspezifische Niveauunterschiede weder implizit oder explizit berücksichtigt noch durchgängig vernachlässigt werden. Für den Erkenntnisgewinn im sozialwissenschaftlichen Kontext sind die Koeffizienten des gepoolten Modells zumindest ein Fingerzeig in Richtung der wahren Gegebenheiten und können dabei helfen, Hypothesen aufzustellen.

7.2 VCM: Variables-Koeffizienten-Modell

Ein hilfreiches Werkzeug bei der Bewertung von *individuellen* Effekten in Paneldaten ist das "variable coefficient model" (VCM). Im Gegensatz zu allen anderen State-of-the-Art-Methoden (gepoolt, FD, FE und RE) wird bei diesem Modell der *Effekt pro Untersuchungseinheit* bestimmt und nicht ein genereller Effekt über die gesamte Population. Die statistische Aufarbeitung dieser Schar von Effekten (Koeffizienten) gibt Aufschlüsse über die Verteilung und Varianz von Effekten der Untersuchungseinheiten einer Population.

Bei der Modellierung anhand des VCM sind zwei Beeinträchtigungen hervorzuheben:

1. Die Koeffizienten jedes Individuums wurden aufgrund einer sehr kleinen Stichprobe bestimmt, wodurch die Sicherheit resp. die statistische Signifikanz der Koeffizienten kleiner ist als bei Modellen, die Informationen der gesamten Population verwenden.
2. Aufgrund der kleineren Stichprobe ist die wählbare Komplexität für das Modell, also die Grösse der Regressionsgleichung, nach oben beschränkt.

Der zweite Punkt hat im Falle des vorliegenden Paneldatensatzes eine direkte Auswirkung. Das Untersuchungsszenario besteht aus 18 unabhängigen Variablen. Pro Untersuchungseinheit stehen jedoch maximal 16 Datenpunkte zur Verfügung, oft sogar weniger. Aus diesem Grund ist das VCM nicht geeignet zur gesamtheitlichen Modellierung des Untersuchungsszenarios. Alternativ bietet VCM in diesem Fall die Möglichkeit, den Zusammenhang zwischen der abhängigen Variable $y = \text{depression}$ und jeder unabhängigen Variable x_i separat zu beurteilen ($y \sim x_i$). Bezüglich Interpretation sind hier zwei Fälle zu unterscheiden:

Fall1: Die unabhängige Variable x_i ist numerisch:

In diesem Fall bezeichnet der *Regressionskoeffizient pro Untersuchungseinheit* den direkt messbaren Effekt resp. die Steigung der entsprechenden Regressionsgerade. Abbildung 18 zeigt die summarische Darstellung der individuellen Regressionskoeffizienten sämtlicher Untersuchungseinheiten für die unabhängige Variable $x_i = \textit{Zufriedenheit mit der Atmosphäre am Arbeitsplatz}$. Es zeigt sich eine verhältnismässig symmetrische Verteilung mit einem Median nahe bei null. Trotzdem fällt auf, dass vermehrt Untersuchungseinheiten mit einem negativen linearen Zusammenhang $y \sim x_i$ vorkommen. Diese Beobachtung ist konsistent zum geschätzten Koeffizienten für x_i im gepoolten Modell (-0.20***).

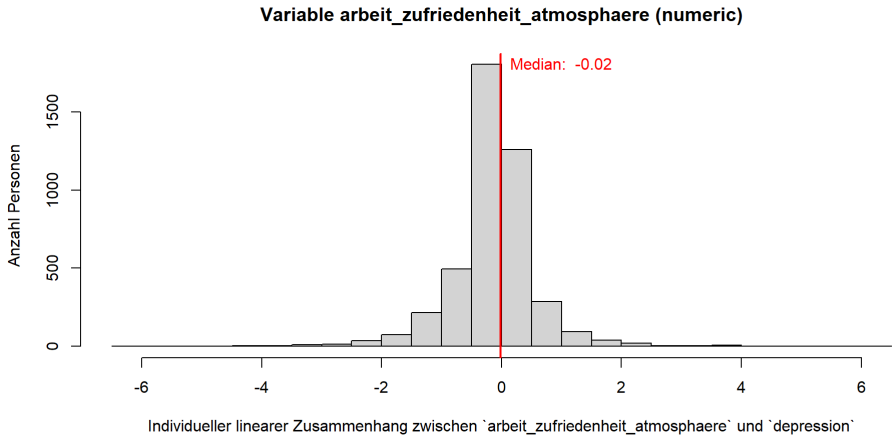


Abbildung 18: VCM für eine numerische unabhängige Variable

Fall 2: Die unabhängige Variable ist eine Faktorvariable:

In diesem Fall bezeichnet der *Regressionskoeffizient pro Untersuchungseinheit* den absoluten Offset zwischen den einzelnen Faktorlevels – analog zu ANOVA (Analysis of Variance). Hierbei ist zu beachten, dass nur Koeffizienten erzeugt werden können, wenn mindestens zwei Faktorlevels innerhalb eines Individuums vorhanden sind. Diese Tatsache kann als Nachteil gesehen werden, da alle Untersuchungseinheiten ausselektiert werden, die keinen Wechsel der unabhängigen Variable erleben. Für die forschende Person kann diese Tatsache ebenso als Vorteil gesehen werden, da nun explizit der Wechsel der Faktorvariable untersucht werden kann. Abbildung 19 zeigt die summarische Darstellung des absoluten Offsets von $y = depression$ für Untersuchungseinheiten, die einen Wechsel der unabhängigen Variable $x_i =$ Schweizer Nationalitäterfahren haben. Es ist zu beachten, dass die Anzahl Beobachtungen deutlich kleiner ist als in Fall 1, da wenige Untersuchungseinheiten einen Nationalitätswechsel erlebt haben. Im Mittel lässt sich kein Zusammenhang zwischen x_i und y herstellen. Abbildung 19 zeigt, dass ein Nationalitätswechsel sowohl positive als auch negative lineare Zusammenhänge mit "depression" aufweist und keine Tendenz hin zur oder weg von der Schweizer Nationalität existiert.

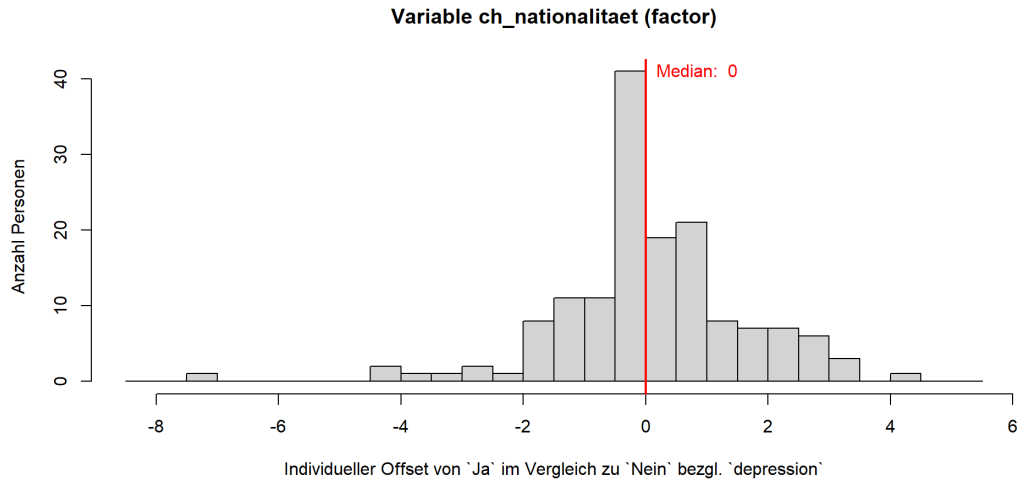


Abbildung 19: VCM für eine unabhängige Faktorvariable

Die Modellierung mittels VCM ist nicht zur Analyse grösserer Modelle resp. Regressionsgleichungen geeignet. Sie kann jedoch im Rahmen kleinerer Modelle verwendet werden, um ein Gefühl für die Variation linearer Zusammenhänge über die Untersuchungseinheiten zu erhalten. So können beispielsweise Ausreisser(-Gruppen) oder spezielle Verteilungsmuster (linksschief, rechtsschief, bimodal) in den linearen Zusammenhängen aufgedeckt werden, die bei grossen Modellen im Mittelwert untergehen (da dort ein einzelner Koeffizient geschätzt wird). Eine Zusammenstellung sämtlicher VCM-Analysen am vorhanden Paneldatensatz befindet sich in Anhang H.

Für die Erstellung allgemeingültiger Hypothesen ist das VCM folglich nicht geeignet. Möchte die forschende Person jedoch einzelne Gruppen oder Muster finden, um diese individuell zu analysieren, ist VCM eine empfehlenswerte Methode.

7.3 FD: First Differences Modell

Das FD-Modell bietet die Möglichkeit, unmittelbare Effekte von Veränderungen der unabhängigen Variablen auf die Veränderung der abhängigen Variable zu modellieren. Wie in Kapitel 4.1.3 beschrieben, vernachlässigt dieses Modell einheitenspezifische Niveauunterschiede durch die Bildung erster Differenzen gemäss Formel (7). D.h. die Gemeinsamkeit zweier jeweils aufeinanderfolgender Messungen wird eliminiert und die Effekte von absoluten Änderungen modelliert.

$$y_{it} - y_{i(t-1)} = b_1 \cdot (x_{it} - x_{i(t-1)}) + b_2 \cdot (z_i - z_i) + w_{it}, \quad w_{it} = e_{it} - e_{i(t-1)} \tag{7}$$

Da der vorliegende Paneldatensatz als Zeiteinheit ganze Jahre betrachtet, modelliert das FD-Modell entsprechend den unmittelbaren Effekt der unabhängigen Variablen auf $y = depression$ im nachfolgenden Jahr. Abbildung 20 zeigt die geschätzten Effekte für den vorliegenden Paneldatensatz.

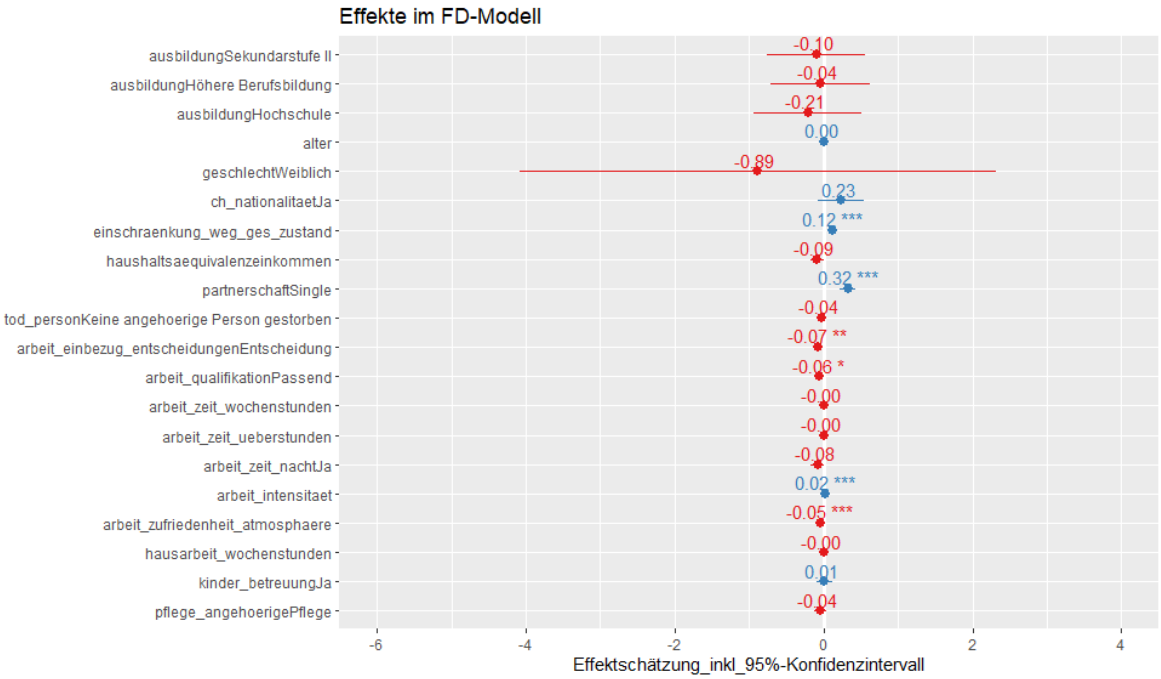


Abbildung 20: Effektschätzung des FD-Modells (blau > 0, rot < 0)

Das FD-Modell kann grundsätzlich als signifikant eingestuft werden (p.value < 2.22e-16), erklärt jedoch nur 3% der Varianz (R-Squared = 0.034, Adjusted R-Squared = 0.033). Sämtliche Ausgaben von R/RStudio zum FD-Modell befinden sich in Anhang I. Die deutlich tiefere Erklärungskraft des FD-Modells im Vergleich zum gepoolten Modell zeigt, dass sich mögliche Zusammenhänge mit Depression deutlich schlechter erklären lassen, wenn ausschliesslich Informationen aus dem unmittelbaren Vorjahr verwendet werden. In diesem Zusammenhang bieten FE-Modelle eine passende Alternative (vgl. nächstes Kapitel).

Bezüglich Interpretation von FD-Koeffizienten ist Vorsicht geboten. Ein Vergleich mit den Koeffizienten des gepoolten Modells zeigt grosse Unterschiede, wobei einzelne Effekte sogar die Richtung ändern. Hierbei soll kurz auf die unterschiedlichen Fragestellungen hinter den Modellen hingewiesen werden:

- Gepooltes Modell: "Wie hat sich y im Allgemeinen verändert, wenn sich x im Allgemeinen um eine Einheit verändert hat?"

- FD-Modell: "Wie hat sich y im letzten Jahr verändert, wenn sich x im letzten Jahr um eine Einheit verändert hat?"

Da Effekte nicht immer unmittelbar wirken, ist die Aussage gerechtfertigt, dass FD-Modelle nicht in der Lage sind, langfristige oder zeitlich verzögerte Effekte aufzugreifen. So kann die Geburt eines Kindes im Jahr γ beispielsweise eine positiven/negativen Effekt auf den Gemütszustand auf die Eltern haben, jedoch einen negativen/positiven Effekt auf den Gemütszustand im Jahr $\gamma + 3$ oder $\gamma + 10$.

Für die Generierung allgemeingültiger Hypothesen, die zeitlich langfristige Effekte darlegen, ist das FD-Modell somit tendenziell weniger geeignet als ein gepooltes Modell. Trotzdem kann die Beobachtung signifikanter, unmittelbarer Effekte im FD-Modell ein wichtiger Ansatzpunkt für die forschende Person sein, um weitere Analysen durchzuführen. So lässt beispielsweise der Koeffizient `partnerschaftSingle = 0.32***` schliessen, dass Untersuchungseinheiten, die eine Beziehung im Jahr γ aufgeben, signifikant höhere Werte für Depression im Jahr $\gamma + 1$ angeben. Die unmittelbare Wirkung des Partnerschaftsstatus auf Depression ist somit ein Ansatzpunkt für die forschende Person, weitere Untersuchungen anzustellen.

Wenn zeitliche Effekte, die länger als eine Zeiteinheit wirken, modelliert und gleichzeitig einheitenspezifische Niveauunterschiede kontrolliert werden sollen, bietet das FE-Modell eindeutige Vorteile gegenüber dem FD-Modell.

7.4 FE: Fixed-Effects Modell

Eines der berühmtesten Modelle zur longitudinalen Analyse von Paneldaten ist das aus der Ökonometrie stammende FE-Modell. Im Gegensatz zum FD-Modell werden bei diesem Modell die einheitenspezifischen Niveauunterschiede nicht durch Subtraktion des Vorjahreswertes, sondern durch Subtraktion der einheitenspezifischen Mittelwerte eliminiert. Der transformierte Datensatz gibt somit an, inwiefern ein Merkmal x_i von seinem zeitlichen Mittelwert abweicht, wodurch sämtliche Untersuchungseinheiten in eine ähnliche Vergleichsbasis transformiert werden.

$$y_{it} - \bar{y}_i = b_1 \cdot (x_{it} - \bar{x}_i) + b_2 \cdot (z_i - \bar{z}_i) + w_{it}, \quad w_{it} = e_{it} \tag{8}$$

Weil aufgrund der Entmittlung sämtlicher Variablen die Regressionsgerade durch den Schwerpunkt jeder einheitenspezifischen Messreihe läuft, beträgt die durchschnittliche idiosynkratische Abweichung auf Ebene der Untersuchungseinheiten null. Der Fehlerterm besteht im Rahmen der FE-Regression also ausschliesslich aus einer Realisation der Zufallsvariablen e_{it} (Giesselmann & Windzio, 2012, S. 44).

Zur Veranschaulichung dieses Prinzips wird auf das Gedankenbeispiel aus Kapitel 4.1.2 zurückgegriffen. Abbildung 21 zeigt, wie einheitenspezifische Niveauunterschiede im gepoolten Modell nicht berücksichtigt werden. Durch die Entmittlung, anschliessende OLS-Regression und erneutes Hinzufügen der einheitenspezifischen Mittelwerte, ist das FE-Modell in der Lage diese Niveauunterschiede mitzuberücksichtigen. Die Bestimmung der jeweiligen Effekte wird auf den entmittelten Daten vorgenommen, weshalb die Niveauunterschiede keinen Einfluss auf die geschätzten Effekte haben. Die

Entmittlung ermöglicht somit eine bessere Fokussierung auf Längsschnitt-Fragestellungen als die bisher betrachteten Modelle.

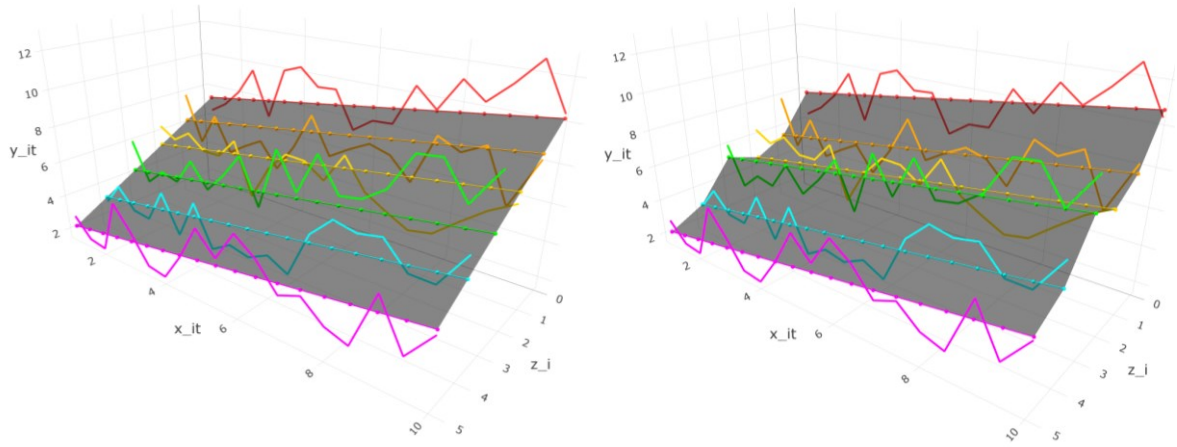


Abbildung 21: Vergleich von gepooltem Modell (links) und FE-Modell (rechts)

Abbildung 22 zeigt die entsprechende Schätzung aller Effekte für das definierte Untersuchungsszenario. Sämtliche Ausgaben von R/RStudio des FE-Modells befinden sich in Anhang J. Das FE-Modell kann zwar als signifikant eingestuft werden (p.value < 2.22e-16), erklärt jedoch die Varianz der Daten nicht (R-Squared = 0.047, Adjusted R-Squared = -0.161). Diese Erkenntnis ist äusserst wichtig für unseren Paneldatensatz. Das FE-Modell ist eindeutig besser als die Modellierung eines einfachen Durchschnitts von Depression. Dennoch kann das Modell die Varianz der Daten im Längsschnitt nicht annähernd erklären.

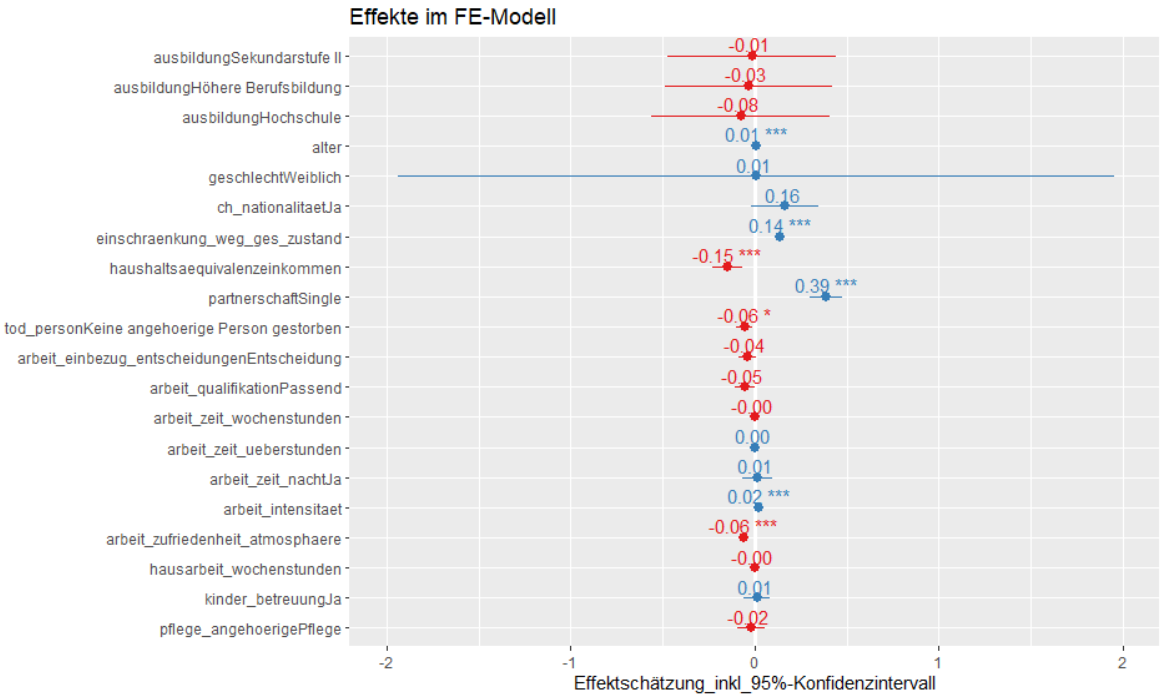


Abbildung 22: Effektschätzung des FE-Modells (blau > 0, rot < 0)

Eine mögliche Ursache hierfür sind die kleinen Schätzungen für die Effekte. In Abbildung 23 wird ersichtlich, inwiefern die Vorhersagen des FE-Modells die tatsächliche Varianz der echten Rohdaten ungenügend abbilden können. Dadurch erklärt sich der tiefe Wert für R-Squared[27] im gesamten Modell. Es ist vorstellbar, dass die Fluktuation der abhängigen Variable einem scheinbar zufälligen Muster folgt, das die gegebenen Variablen nicht zu erklären vermögen.

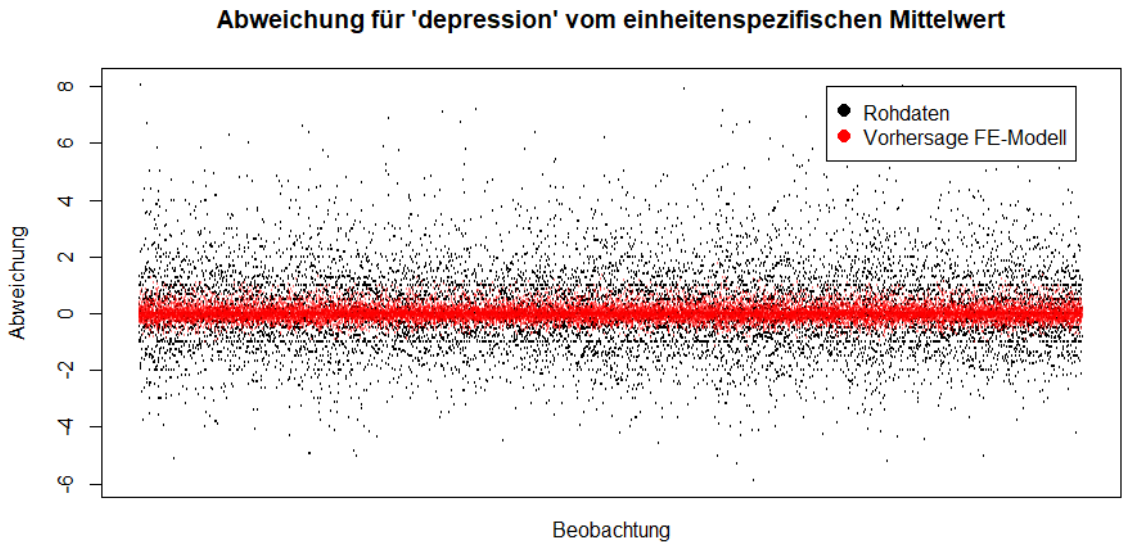


Abbildung 23: Abweichung vom einheitenspezifischen Mittelwert bei Rohdaten und FE-Modell

Aufgrund der geringen Erklärungskraft des gesamten Modells, sollten die geschätzten Effekte des FE-Modells in ihrer Aussagekraft für den vorliegenden Paneldatensatz nicht überbewertet werden. Trotzdem bleibt an dieser Stelle zu erwähnen, dass die Richtung und Signifikanz einzelner Effekte im FE-Modell konsistent ist zu den Aussagen des gepoolten oder FD-Modells.

Die Aussage, ein FE-Modell fokussiert auf die longitudinalen Eigenschaften eines Paneldatensatzes, ist grundsätzlich korrekt. Anhand des obigen Beispiels wird dennoch aufgezeigt, dass im FE-Modell nach wie vor der Effekt über eine Menge von Untersuchungseinheiten *im Mittel* bewertet wird. D.h. wenn bei 100 Untersuchungseinheiten die Depression über die Zeit zunimmt und bei 100 ähnlichen Untersuchungseinheiten die Depression in ähnlichem Masse über die Zeit abnimmt, so ist auch ein FE-Modell nicht in der Lage, diese Problematik zu entschlüsseln. Die Art von "Mittelung über alle Untersuchungseinheiten" ist eine allgemeine Schwäche von Panelmodellen, die im vorliegenden Fall schwierig zu überwinden ist.

Weil die Effektschätzungen in der Längsschnittbetrachtung aus statistischer Sicht Zweifel aufbringen, wird im nächsten Kapitel auf das RE-Modell eingegangen, das grundsätzlich auf Querschnittfragestellungen ausgerichtet ist.

[27] R^2 ist definiert als das Verhältnis aus der vom Modell erklärten Varianz und der totalen Varianz der Daten. $R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{y}_i - \bar{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} = 1 - \frac{RSS}{TSS} = \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$, mit ESS = Explained Sum of Squares, RSS = Residual Sum of Squares und TSS = total Sum of Squares.

7.5 RE: Random-Effects Modell

Basierend auf der Annahme, dass der einheitenspezifische Mittelwert $\bar{y}_i$ kein guter Schätzer für den eigentlichen Einheiteneffekt darstellt, wird im RE-Modell der Schätzer dieses Effekts mit dem Gesamtmittelwert aller Beobachtungen $\bar{y}_{it}$ gewichtet. Dieses Verfahren impliziert gemäss Giesselmann & Windzio (2012, S.83) die Idee, dass man "in dem Maße, in dem der einfache Schätzer des einheitenspezifischen Mittelwertes mit Unsicherheit assoziiert ist, auf Informationen der anderen Stichprobeneinheiten zurückgreift."

$$\bar{y}_i^{re} = \lambda_i \cdot \bar{y}_i + (1 - \lambda_i) \cdot \bar{y}_{it} \tag{9}$$

$$\bar{x}_i^{re} = \lambda_i \cdot \bar{x}_i + (1 - \lambda_i) \cdot \bar{x}_{it} \tag{10}$$

$$\bar{z}_i^{re} = \lambda_i \cdot \bar{z}_i + (1 - \lambda_i) \cdot \bar{z}_{it} \tag{11}$$

Dadurch ergibt sich folgende Transformationsformel für das RE-Modell:

$$y_{it} - \bar{y}_i^{re} = b_1 \cdot (x_{it} - \bar{x}_i^{re}) + b_2 \cdot (z_i - \bar{z}_i^{re}) + w_{it}, \quad w_{it} = e_{it} - \lambda_i \cdot \bar{e}_i + (1 - \lambda_i) \cdot u_i \tag{12}$$

Der Faktor λ_i stellt somit das Mass der Sicherheit dar, wie zuverlässig die einheitenspezifischen Mittelwerte die wahren Einheiteneffekte schätzen. Für $\lambda_i = 1$ wird der einheitenspezifische Mittelwert als sicherer Schätzer für den wahren Einheiteneffekt betrachtet, wodurch das FE-Modell reproduziert wird. Je stärker λ_i gegen null konvergiert, desto grösser wird die Annäherung des Schätzers an den Gesamtmittelwert $\bar{y}_{it}$. Der Faktor λ_i wird u.a. bestimmt durch das Verhältnis von intraindividueller (idiosynkratischer) Varianz $var(e_{it})$ und die mit der Zeit gewichteten gesamten Fehlervarianz $T \cdot var(u_i)$.

$$\lambda_i = 1 - \sqrt{\frac{var(e_{it})}{T \cdot var(u_i) + var(e_{it})}} \tag{13}$$

Die linke Grafik in Abbildung 24 zeigt schematisch, warum der einheitenspezifische Mittelwert bei kleiner intraindividueller Varianz einen durchaus sicheren Schätzer darstellt und das RE-Modell deshalb das FE-Modell reproduziert.

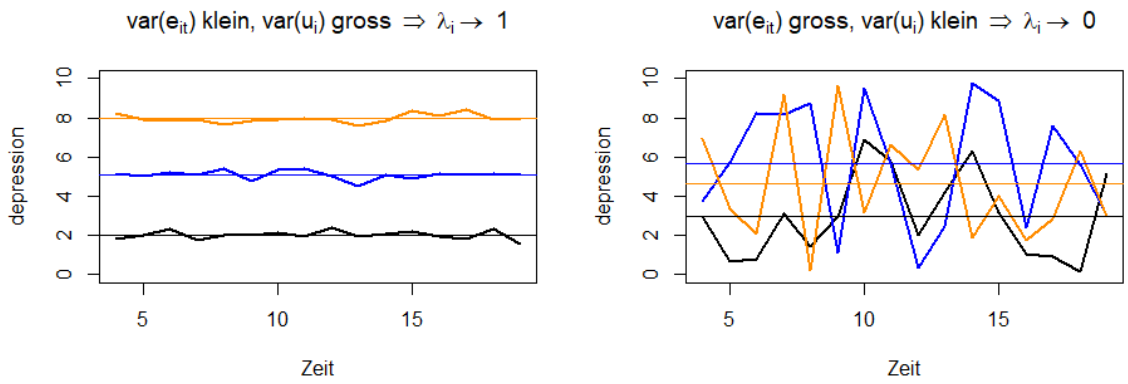


Abbildung 24: Schematische Darstellung zeitlicher Verläufe mit unterschiedlicher Varianzstruktur

Das Gedankenbeispiel aus Kapitel 4.1.2 dient in Abbildung 25 wiederum als Veranschaulichung des RE-Modells. Aufgrund schöner Eigenschaften des Gedankenbeispiels sind die Werte für λ_i bei allen Individuen nahe bei 1, wodurch das RE-Modell (rechts) ähnliche einheitenspezifischen Mittelwerte verwendet wie das FE-Modell (links). Die schwarzen, gestrichelten Hilfslinien helfen zu erkennen, dass die einheitenspezifischen Mittelwerte im RE-Modell leicht ausgeglichener (resp. weniger "hügelig") sind als im FE-Modell. Grund hierfür ist die Gewichtung mit dem globalen Mittelwert $\bar{y}_{it}$.

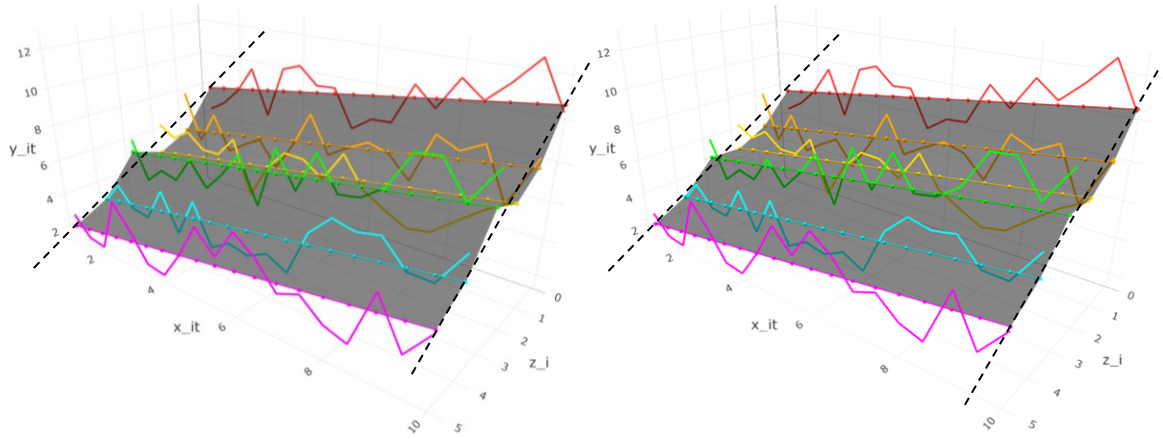


Abbildung 25: Vergleich von FE-Modell (links) und RE-Modell (rechts)

Aus Abbildung 9 (S.25) ist bekannt, dass der vorliegende Paneldatensatz tendenziell eine Struktur gemäss der rechten Grafik aus Abbildung 24 aufweist. Aus diesem Grund werden beim RE-Modell andere Effektschätzungen erwartet als beim FE-Modell (vgl. Abbildung 26). Die vollständigen Modellparameter inkl. weiterer Angaben befinden sich in Anhang K.

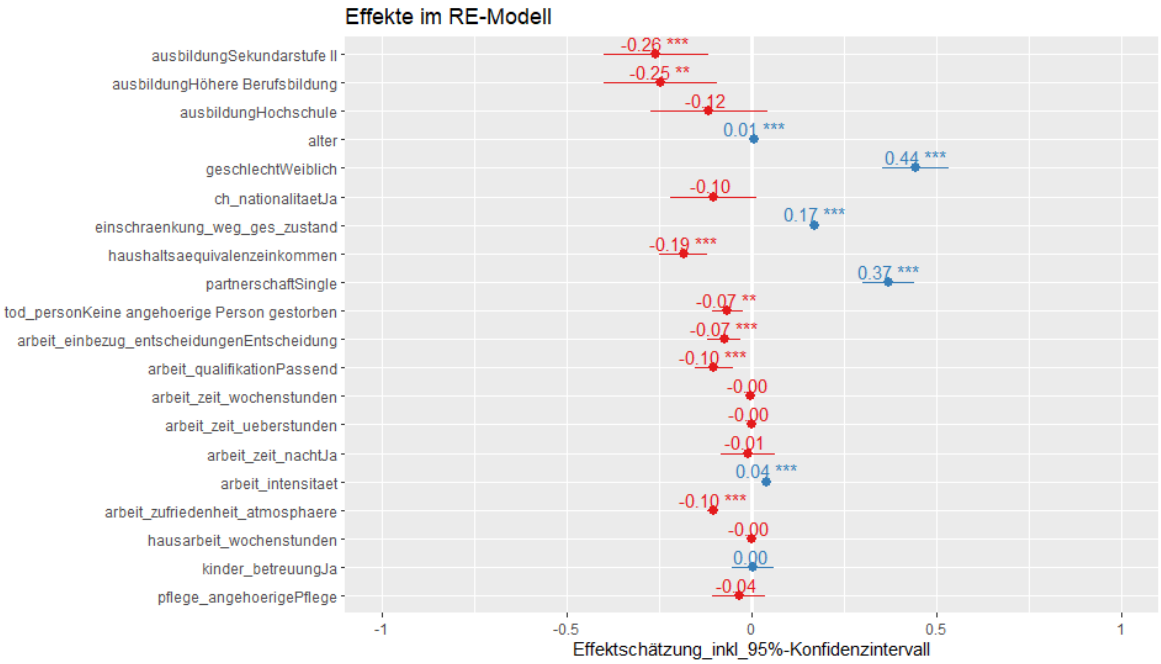


Abbildung 26: Effektschätzung des RE-Modells (blau > 0, rot < 0)

Das RE-Modell kann als signifikant eingestuft werden (p.value < 2.22e-16). Es erklärt ca. 11% der Varianz in den Daten (R-Squared = 0.109, Adjusted R-Squared = 0.109) – deutlich mehr als das FE-

Modell. Durch die gleichzeitige Berücksichtigung der Längs- und Querstruktur des Paneldatensatzes ist das RE-Modell in der Lage, den "Tradeoff" der entsprechenden Effekte zu quantifizieren. Gemäss den Modellangaben stammen ca. 54% der Varianz von idiosynkratischen Effekten und 46% von einheitenspezifischen Effekten.

Aufgrund der teilweisen Entmittlung bleiben im RE-Modell gewisse Querschnittseffekte vorhanden, die im FE-Modell vollständig entfernt wurden. Aus diesem Grund ist das RE-Modell besser in der Lage zeitlich konstante (between-)Effekte wie z.B. *Geschlecht* oder *Ausbildung* zu schätzen, wobei die Längsstruktur der Daten trotzdem berücksichtigt wird. Es sei jedoch darauf hingewiesen, dass RE-Koeffizienten die Effekte sämtlicher im Modell nicht kontrollierten, jedoch mit der abhängigen Variablen korrelierter zeitkonstanter Merkmale mittransportieren. Aufgrund dieser Unschärfe raten Giesselmann & Windzio (2012, S.99-100) davon ab, RE-Modelle für Längsschnittfragen zu verwenden.

Sollen in einem Modell gleichzeitig zeitkonstante (z_i) und zeitveränderliche (x_{it}) unabhängige Variablen betrachtet werden, bietet das RE-Modell eine Alternative zum bereits verwendeten gepoolten Modell. Für den vorliegenden Paneldatensatz ist schwierig zu beurteilen, inwiefern das RE-Modell dem FE-Modell für den vorliegenden Paneldatensatz vorzuziehen ist. Um Fragen der Modellwahl besser anzugehen, können verschiedene Tests hinzugezogen werden.

7.6 Tests für Panelmodelle

Um ein lineares Panelmodell auf seine Eigenschaften zu untersuchen, werden im Folgenden mehrere Tests aus dem R-Paket `plm`[28] vorgestellt. Diese geben durch den Vergleich verschiedener Modelle (gepoolt, VCM, FE, RE) mittels F- und Chi-Quadrat-Tests Aufschluss darüber, inwiefern das verwendete Untersuchungsszenario einheitenspezifische und zeitliche Effekte erkennt und quantifiziert.

Diese Tests sind nicht über alle Zweifel erhaben und sollen mit Vorsicht interpretiert werden. Die Autoren des `plm`-Pakets bemerken, dass ein Mehrwert solcher Tests darin besteht, eine Annäherung oder Entfernung von der Nullhypothese bei verschiedenen Modellen festzustellen (Croissant & Millo, 2008).

7.6.1 Test auf Poolbarkeit

Sind die geschätzten Koeffizienten des VCM-Modells für alle Untersuchungseinheiten gleich, so ist ein gepooltes Modell passend für die Modellierung der vorliegenden Daten. Der Vergleich von VCM- und gepoolten Koeffizienten anhand eines F-Tests überprüft die Nullhypothese, dass die VCM-Koeffizienten identisch sind[29].

Für die erstellten VCM-Modelle in Kapitel 7.2 wird die Nullhypothese auf dem 0.05-Signifikanziveau durchgängig abgelehnt. Dies ist nicht weiter erstaunlich, da jedes dieser Modelle ausschliesslich eine unabhängige Variable betrachtet und wir durchaus unterschiedliche Zusammenhänge dieser Variablen mit $depression$ erwarten für verschiedene Untersuchungseinheiten.

[28] vgl. Coprehensive R Archive Network: https://cran.r-project.org/web/packages/plm/index.html
[29] vgl. R-Funktion `plm::pooltest`

Derselbe Test liesse sich ebenfalls für den Vergleich von VCM-Modellen mit FE-Modellen verwenden. Wird die Nullhypothese in diesem Fall angenommen, so würden die individuell geschätzten VCM-Effekte gut durch ein FE-Modell (ein um Niveauunterschiede bereinigtes Modell) geschätzt.

7.6.2 Test der beobachteten einheitenspezifischen und/oder zeitlichen Effekte

In einem optimalen FE-Modell können Unterschiede zwischen Untersuchungseinheiten und/oder unterschiedlichen Zeitpunkten vollständig erklärt werden.

Durch den Vergleich von gepoolten und FE-Modellen anhand eines F-Tests[30] kann erörtert werden, inwiefern zeitliche oder einheitenspezifische Effekte vorhanden sind. Die Nullhypothese geht davon aus, dass die Koeffizienten beider Modelle identisch sind. Wird die Nullhypothese verworfen, so hat die Kontrolle von zeitlichen und/oder einheitenspezifischen Effekten im FE-Modell einen Einfluss auf die geschätzten Effekte. Für das vorliegende Untersuchungsszenario wird die Nullhypothese ausschliesslich bei zeitlichen Effekten angenommen (p-Wert: 0.08), d.h. die Kontrolle von zeitlichen Fixed-Effects hat keinen signifikanten Einfluss auf die geschätzten Koeffizienten. Aus diesem Grund scheint ein FE-Modell ausschliesslich mit einheitenspezifischen Fixed-Effects sinnvoller. Detaillierte Ausgaben befinden sich in Anhang L.

Ein ähnlicher Test überprüft die Residuen des gepoolten Modells auf zeitliche und/oder einheitenspezifische Effekte anhand eines Lagrange Multiplier Tests[31]. Für das vorliegende Untersuchungsszenario wird die Nullhypothese wiederum für zeitliche Effekte angenommen (p-Wert: 0.13), was die Aussage des F-Tests untermauert. Weitere Angaben befinden sich in Anhang L.

7.6.3 Test auf unbeobachtete einheitenspezifische oder zeitliche Effekte

Obwohl ein Panelmodell einheitenspezifische oder zeitliche Effekte erkennen und quantifizieren kann, ist nicht garantiert, dass sämtliche Effekte erkannt wurden. So können in den Residuen eines Modells weitere einheitenspezifische oder zeitliche Variationen verborgen sein.

Ein Test zur Identifikation solcher unbeobachteten Effekte ist der "Wooldridge's Test for Unobserved Effects in Panel Models"[32]. Dieser semi-parametrische Test prüft die Nullhypothese, dass keine Korrelation zwischen den Fehlern innerhalb einer Untersuchungseinheit (oder eines Zeitpunktes) existiert. Wird die Nullhypothese verworfen, so ist dies ein Zeichen für unbeobachtete Heterogenität auf individueller oder zeitlicher Ebene. Für das vorliegende Untersuchungsszenario wird die Nullhypothese bei zeitlichen Effekten angenommen (p-Wert: 0.47). Weitere Angaben befinden sich in Anhang L.

Die Erkenntnisse der letzten zwei Kapitel deuten darauf hin, dass die zeitliche Variation (beobachtet oder unbeobachtet) im vorliegenden Paneldatensatz nur wenig vorhanden ist oder durch das definierte Untersuchungsszenario bereits gut kontrolliert wird. Diese Aussage ist nicht erstaunlich, da man annehmen kann, dass die Werte für $depression$ pro Jahr im Mittel ungefähr konstant bleiben aufgrund gegenseitiger Kompensation.

[30] vgl. R-Funktion `plm::pFtest`
[31] vgl. R-Funktion `plm::plmtest`
[32] vgl. R-Funktion `plm::pwtest`

7.6.4 Test auf serielle Korrelationen

Eine weitere Möglichkeit zur Bewertung der Qualität eines Panelmodells ist die Untersuchung serieller Korrelationen. Wird in den Residuen eines Panelmodells serielle Korrelation (Autokorrelation, autoregressive Korrelation) festgestellt, ist dies ein Indiz für unbeobachtete Effekte, die durch das Modell nicht vollständig erklärt werden können. Verschiedene Tests zur Bewertung serieller Korrelationen eines longitudinalen Panelmodells sind in den Funktionen `plm::pbsytest`, `plm::pbltest`, `plm::pbgtest`, `plm::pdwtest`, `plm::pwartest` implementiert. Basierend auf verschiedenen Kriterien der ökonometrischen Literatur können so Modelle untersucht und miteinander verglichen werden.

Für das vorliegende Untersuchungsszenario wird die Nullhypothese in allen untersuchten Fällen auf dem 0.05-Signifikanzniveau verworfen (vgl. Anhang L). Dies stützt die Aussage, dass das Untersuchungsszenario nicht in der Lage ist, die abhängige Variable $depression$ vollständig durch die unabhängigen Variablen zu erklären. Hierbei ist zu erwähnen, dass der Anspruch einer vollständigen Erklärung durch die unabhängigen Variablen im sozialwissenschaftlichen Kontext äusserst hoch ist.

7.6.5 Hausman Test

Ist die Exogenitätsannahme[33] erfüllt, sind FE- & RE-Koeffizienten konsistente Schätzungen eines Modells, wobei das RE-Modell effizienter ist als das FE-Modell. Ist die Exogenitätsannahme verletzt, sind die RE-Koeffizienten gegenüber den FE-Koeffizienten verzerrt. Der Hausman Test vergleicht die FE- und RE-Koeffizienten eines Modells und bewertet somit, ob im definierten Modell Endogenität vorliegt (Chi-Quadrat-Test mit Nullhypothese H_0: $Cov(x_{it}, u_i) = 0$). Für das gegebene Untersuchungsszenario und den vorliegenden Paneldatensatz ergibt der Hausman Test einen p-Wert von 2.2e-16 (vgl. Abbildung 27) und zeigt somit an, dass erstens Endogenität vorliegt und zweitens FE- und RE-Koeffizienten inkonsistent sind.

```
## Hausman Test
## data:  base_formula
## chisq = 575.81, df = 20, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

Abbildung 27: Hausmann Test des Untersuchungsszenarios mittels RE

Daraus könnte geschlossen werden, dass kein RE-Modell verwendet werden sollte, um die vorliegenden Daten zu modellieren. Diese Interpretation ist fehlerhaft, da sie impliziert, dass eine FE-Modellierung mittels reiner Dummy-Variablen resp. Fixed Effects "richtig" oder "besser" ist als die RE-Modellierung. Die korrekte Interpretation ist die folgende: Beide Modelle kommen nicht zum selben Schluss und das RE-Modell bringt in dieser Form verzerrte Schätzungen hervor. An dieser Stelle ist die Verwendung eines erweiterten Modelles (RE-KV-Modell, vgl. Kapitel 7.7) angezeigt.

Der entsprechende Hausman Test für das RE-KV Modell (Abbildung 28) zeigt grundsätzlich dasselbe Bild, jedoch ist eine Annäherung hin zur Nullhypothese sichtbar (p-Wert von 1.579e-05). Somit wird deutlich, dass das RE-KV Modell die vorhandenen Effekte tendenziell besser erfasst, mit einer geringeren Verletzung der Exogenitätsannahme.

```
## Hausman Test
```

[33] Exogenitätsannahme: $Cov(x_{it}, u_i) = 0$. Wenn Endogenität vorliegt, gilt: $Cov(x_{it}, u_i) \neq 0$

```
## data:  formula_mc
## chisq = 57.749, df = 20, p-value = 1.579e-05
## alternative hypothesis: one model is inconsistent
```

Abbildung 28: Hausmann Test des normierten Untersuchungsszenarios mittels RE-KV

7.7 RE-KV: Hybrides Modell

Die Verbindung von Kontextvariablen und RE-Transformation bringt in der RE-KV-Methode ein neues Werkzeug hervor, das mit den Verletzungen der Exogenitätsannahme[34] im RE-Modell besser umgehen kann.

Kontextvariablen $\bar{x}_i$ ermöglichen die Verwendung zeitkonstanter Merkmale z_i bei gleichzeitiger Kontrolle unbeobachteter Heterogenität. Das RE-KV bietet gemäss Giesselmann & Windzio (2012. S.102-103) folgenden Vorteil gegenüber dem RE-Modell:

> *"Die Kontextvariablen-Regression eliminiert gerade den Anteil unbeobachteter Heterogenität, welcher mit den unabhängigen Variablen korreliert ist, während die Random Effects-Transformation den Teil des Einheiteneffektes absorbiert, der für die intraindividuelle Korrelation der Fehlerterme verantwortlich ist. OLS KV produziert also unverzerrte Schätzer, während RE die Validität von Standardfehlern und Teststatistik sicherstellt."*

Bei der technischen Umsetzung von RE-KV wird jede numerische Variable x_{it} in zwei Teile gesplittet. Einen einheitenspezifischen Mittelwert $\bar{x}_i$ (die Kontextvariable) und die zeitlichen resp. einheitenspezifische longitudinale Abweichung von diesem Mittelwert $(x_{it} - \bar{x}_i)$. Dieser Prozess wird oft als *group mean centering* oder *Mundlak Device* (Mundlak, 1978) bezeichnet.

$$x_{it} = \bar{x}_i + (x_{it} - \bar{x}_i) \tag{14}$$

Im entsprechenden RE-KV-Modell bezeichnen wir diese zwei Teile jeweils mit x_i_mn ("mean") und x_i ("deviation"). Die geschätzten Effekte im RE-KV-Modell erlauben somit eine gleichzeitige Quantifizierung von Querschnittseffekten ("Kohorteneffekt") als auch Längsschnitteffekten (einheitenspezifische Effekte) und fassen somit die Eigenschaften eines Paneldatensatzes in möglichst generischer Form auf. Abbildung 29 zeigt die Effektschätzungen des RE-KV-Modells. Weitere Modellangaben befinden sich in Anhang M.

Das RE-KV-Modell kann ca. 13% der Varianz erklären (R-Squared = 0.132, Adjusted R-Squared = 0.131) – leicht mehr als das RE-Modell mit ca. 11%. Dieser Unterschied ist gering, zeigt jedoch, dass die Erweiterung des RE-Modells mit Kontextvariablen im konkreten Beispiel eine Verbesserung der Erklärungskraft hervorbringt.

Wie in Kapitel 7.6.5 aufgezeigt, schneidet der Hausman Test des RE-KV Modells besser ab als derjenige des RE-Modells. Nach wie vor sollte im Hinterkopf behalten werden, dass im Falle des gegebenen Untersuchungsszenarios die Exogenitätsannahme mit grosser Wahrscheinlichkeit verletzt ist. Bei einer derart anspruchsvollen Fragestellung, welche Grössen sich auf die abhängige Variable 'depression'

[34] $Cov(x_{it}, u_i) = 0$

auswirken, kann davon ausgegangen werden, dass diverse zeitkonstante unabhängige Variablen nicht berücksichtigt werden, die einheitenspezifische Unterschiede weiter erklären könnten.

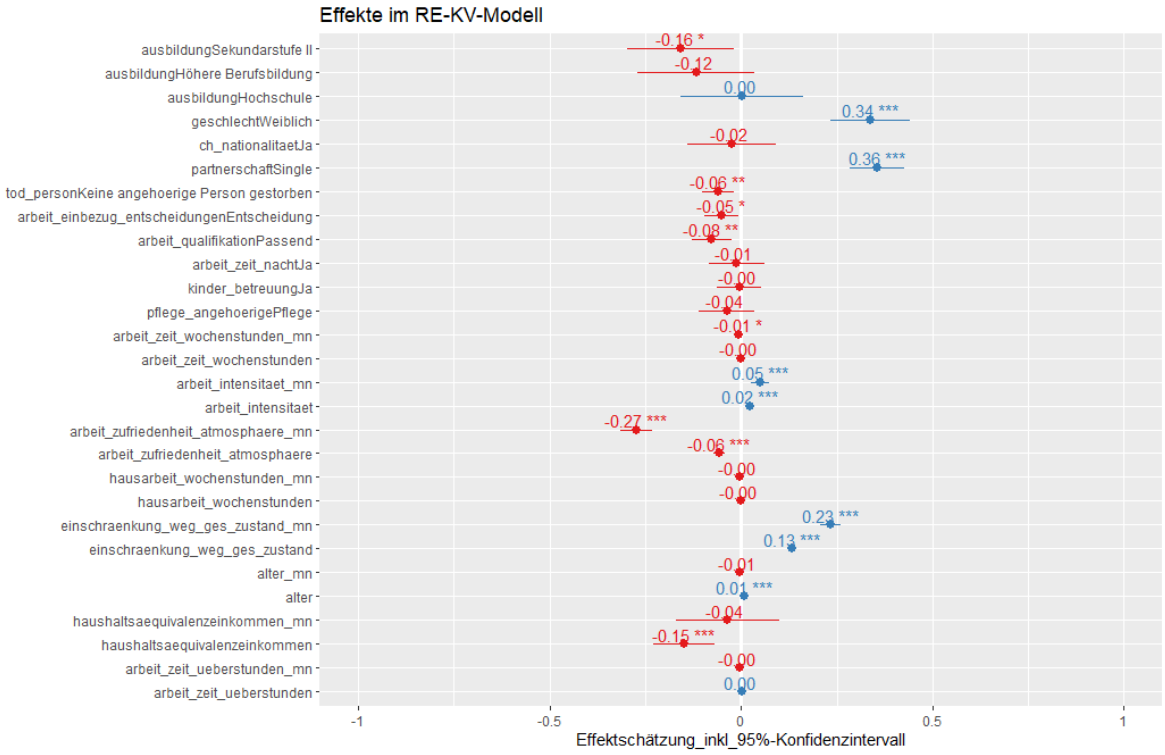


Abbildung 29: Effektschätzung des RE-KV-Modells (blau > 0, rot < 0)

7.8 LC: Longitudinales Clustering

Nebst der Modellierung longitudinaler Effekte in Paneldaten ist die forschende Person bei gewissen Fragestellungen an der Gruppenstruktur innerhalb der Paneldaten interessiert. Hierbei steht die Frage im Zentrum, ob Gruppen mit ähnlichen Eigenschaften existieren und wie sich diese zusammensetzen. Für das Clustering von Daten ohne Längsstruktur existieren diverse Partitionierungsmethoden[35] um Cluster – ähnlich lokalisierte Punkte im n-dimensionalen Raum – zu identifizieren. Liegen jedoch Paneldaten vor, so ist die forschende Person interessiert an der Identifikation von Gruppen (z.B. Personengruppen) mit ähnlichen zeitlichen Verläufen (Trajektorien, engl. *trajectories*). Für diese Art von Fragestellungen bietet das longitudinale Clustering hilfreiche Methoden zur Identifikation von Trajektorien-Gruppen.

Die R-Pakete *kml*[36], *kml3d*[37] und *kmlShape*[38] bieten eine vollumfängliche Implementierung longitudinaler Clusteringmethoden basierend auf k-means Clustering. Das Prinzip dieser Methoden basiert auf der Betrachtung einer Menge S von n Untersuchungseinheiten $i \in \{1, \dots n\}$. Die Messpunkte einer Untersuchungseinheit $i \in S$ werden definiert durch $y_{ij\Omega}$ wobei $j \in \{1, \dots, t\}$ den Zeitpunkt und $\Omega \in$

[35] bspw. *hierarchisches Clustering*, *k-means* oder *PAM (partitioning around meodis)*
[36] https://cran.r-project.org/web/packages/kml/index.html
[37] https://cran.r-project.org/web/packages/kml3d/index.html
[38] https://cran.r-project.org/web/packages/kmlShape/index.html

$\{A, B, C, \dots, M\}$ das gemessene Merkmal bezeichnen. Die Trajektorie $y_{i\,.\,.}$ der Untersuchungseinheit i wird somit definiert als eine Matrix:

$$y_{i\,.\,.} = (y_{i1\Omega}, y_{i2\Omega}, \dots, y_{it\Omega}) = \begin{pmatrix} y_{ijA} \\ \cdots \\ y_{ijM} \end{pmatrix} = \begin{pmatrix} y_{i1A} & \cdots & Y_{itA} \\ \vdots & \ddots & \vdots \\ y_{i1M} & \cdots & y_{itM} \end{pmatrix} \quad (15)$$

Durch eine Minimierung der Distanz[39] zwischen einzelnen Trajektorien wird eine Aufteilung der Menge S in k möglichst homogene Untergruppen erzeugt (Genolini et al., 2015).

Abbildung 30 zeigt den zeitlichen Verlauf der abhängigen Variable *depression* für sämtliche 5'694 Untersuchungseinheiten sowie die mit *kml* errechneten Cluster bei k = 5. Die Cluster A, B und E zeigen konstante Verläufe auf unterschiedlich Levels für *depression*. In Cluster C kann ein abnehmender Verlauf und in Cluster D ein zunehmender Verlauf festgestellt werden. Jede Untersuchungseinheit ist einem dieser fünf Cluster zugeteilt, wodurch klar abgegrenzte Untergruppen definiert und individuell untersucht werden können.

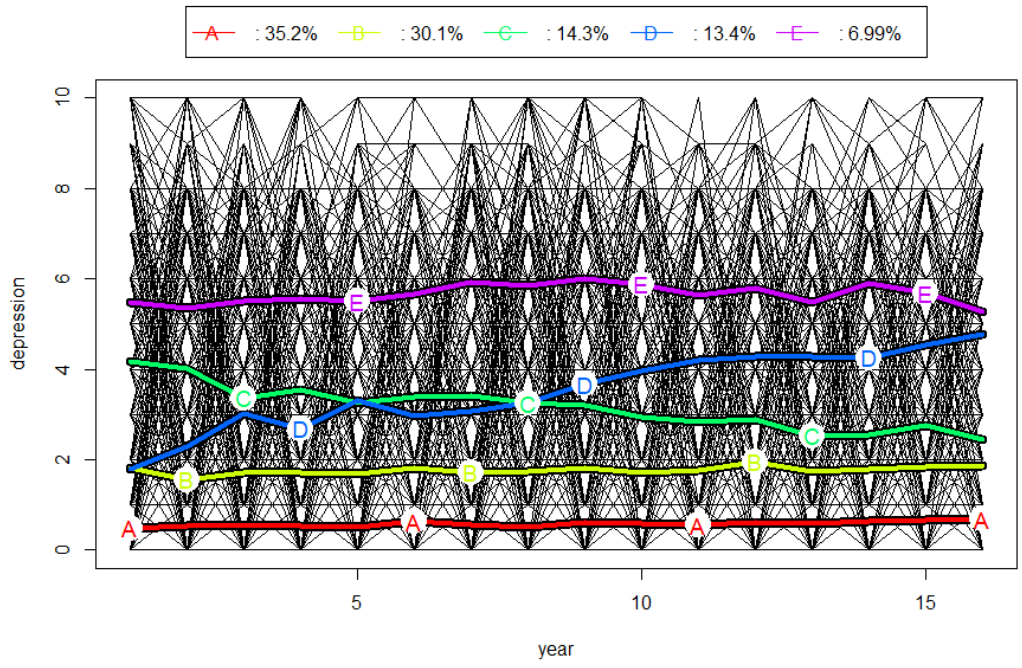


Abbildung 30: Longitudinales Clustering für "depression" mit kml für k = 5.

Die mit *kml* erzeugten Cluster berücksichtigen die zeitliche Entwicklung eines Merkmals, d.h. eine Zeile der in Formel (*15*) definierten Matrix. Die Funktionalität zur gleichzeitigen Berücksichtigung mehrerer Merkmale (mehrerer Zeilen der definierten Matrix) wird in *kml3d* bereitgestellt. Die entsprechenden Cluster berücksichtigen mehr Informationen und geben folglich Hinweise auf komplexere Zusammenhänge. Abbildung 31 zeigt die mit *kml3d* errechneten Cluster bei k = 5 für die Variablen *depression* und *alter*. Darin werden keine abnehmenden oder zunehmenden Verläufe für *depression*

[39] Die Distanz von Matrizen kann anhand verschiedener Distanzmasse (*kml*: Minkowski Distanz, *kml3d*: euklidische Distanz) und verschiedener Qualitätskriterien für *within-cluster-compactness-index* und *between-cluster-spacing-index* bewertet werden (z.B. Calinski & Harabasz Kriterium, Ray & Turi Kriterium, Davies & Bouldin Kriterium). Detaillierte Angaben hierzu liefert (Genolini et al., 2015, Kapitel 2.3 & 2.5)

identifiziert, dafür eine grobe Zuordnung zu verschiedenen Altersgruppen. Anhand dieser Zuordnung liesse sich die Vermutung aufstellen, dass gewisse Altersgruppen stärker gefährdet sind als andere. Diese Vermutung ist per se nicht korrekt.

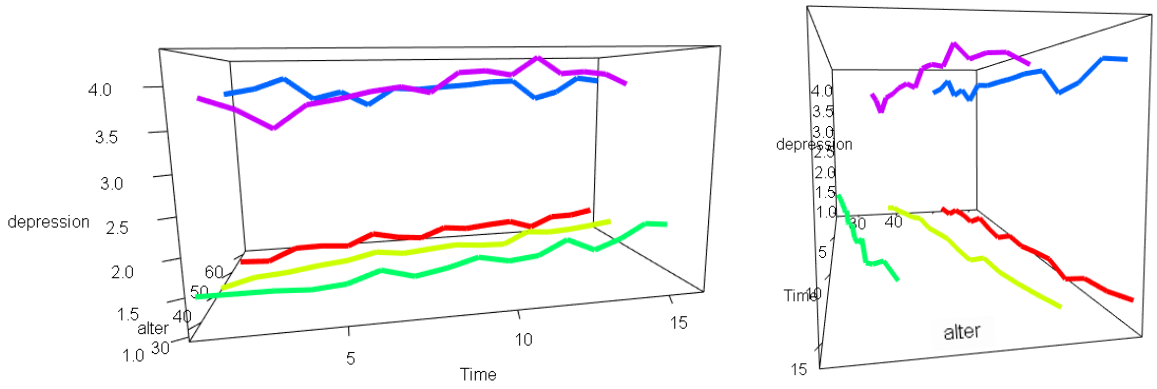


Abbildung 31: Longitudinales Clustering für "depression" und "alter" mit kml3d für k = 5

Die dargestellten Cluster in Abbildung 31 stellen den gemittelten Verlauf mehrerer Untersuchungseinheiten dar und sollten deshalb in Ihrer Bedeutung nicht überbewertet werden. Die Zuordnung aller Beobachtungseinheiten zu ihrem jeweiligen Cluster (Abbildung 32) zeigt bei der Gegenüberstellung mit den gemittelten Verläufen (Abbildung 31 - rechts), dass hohe Werte für Depression in keiner offensichtlichen Weise an die Variable $alter$ geknüpft sind.

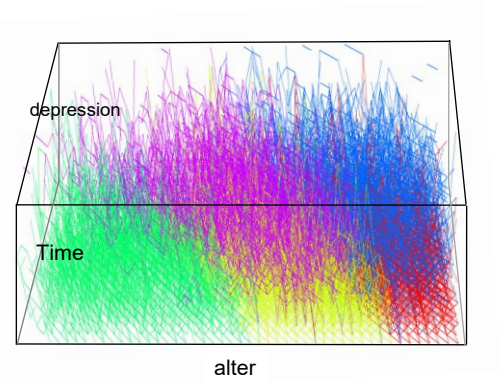


Abbildung 32: Zuordnung longitudinaler Verläufe aller Beobachtungseinheiten zu ihrem Cluster gemäss kml3d für k=5

Trotz der zuvor genannten Fallstricke bietet Clustering mittels *kml3d* die Möglichkeit, zielgerichtet und basierend auf frei wählbaren Variablen ähnliche Gruppen zu identifizieren. Dadurch kann sowohl die Hypothesengenerierung ("*Welche Gruppen oder Muster sind erkennbar?*"), als auch die Hypothesenüberprüfung ("*Können die erwarteten Gruppen oder Muster datenbasiert validiert werden?*") auf hilfreiche Weise unterstützt werden.

Sowohl *kml* als auch *kml3d* gruppieren Trajektorien, die sich lokal nahe sind. Es kann jedoch sein, dass die forschende Person nicht primär an ähnlicher Lokalisierung im n-dimensionalen Raum interessiert ist, sondern an ähnlichen Verläufen oder "Routen" durch den n-dimensionalen Raum. In diesem Fall ist die Form (engl. s*hape*) der Trajektorien wichtiger als ihre absolute Verortung in der Zeit. Das R-Paket *kmlShape* ist darauf ausgelegt die Form von Trajektorien zu berücksichtigen und entsprechend Untersuchungseinheiten mit ähnlichen "Routen" zu identifizieren – unabhängig von der zeitlichen Lokalisierung einer Trajektorie. Abbildung 33 erklärt diesen Sachverhalt schematisch. Zur Bestimmung der Ähnlichkeit von vier Verläufen $i_1 - i_4$ wird gemäss klassischer Distanzmasse (*kml*) die Zeit als weitere Dimension berücksichtigt (vgl. mittlere Grafik) und Verläufe u.a. entsprechend ihrer zeitlichen Verortung gruppiert. Aufgrund erweiterter Distanzmasse in *kmlShape* wird die zeitliche Verortung weniger stark gewichtet und damit Verläufe gruppiert, die sich vermehrt aufgrund der Form ähneln und nicht aufgrund von Gleichzeitigkeit (vgl. rechte Grafik).

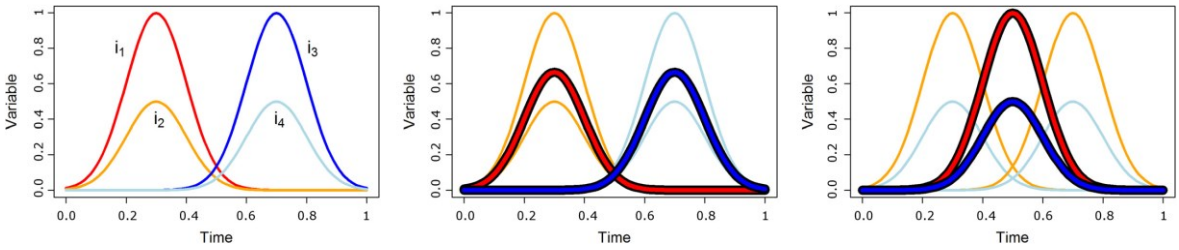


Abbildung 33: Vergleich von Clustering kml (mitte) und kmlShape (rechts), Quelle: (Genolini et al., 2016)

Zur Analyse der vorliegenden Paneldaten ist die Erweiterung in *kmlShape* hilfreich, da nicht primär Untersuchungseinheiten mit *gleichzeig ähnlichen* Verläufen identifiziert werden sollen, sondern vielmehr Untersuchungseinheiten mit *allgemein ähnlichen* Verläufen – unabhängig von Gleichzeitigkeit. Die entsprechenden Cluster des vorliegenden Paneldatensatzes für $k = 5$ sind in Abbildung 34 dargestellt[40]. Diese liefern zusätzliche Hinweise für Muster in den zeitlichen Verläufen von Depression, welche vom Verfasser folgendermassen interpretiert werden.

- Rot: Hoher Grundlevel mit Krise und anschliessender Erholung
- Grün: Mehrstufige Akkumulation (je nach Interpretation 2- oder 3-stufig)
- Dunkelblau: Erholung
- Hellblau: Tiefer Grundlevel mit Krise und anschliessender Erholung
- Violet: Konstant tiefer Level für Depression

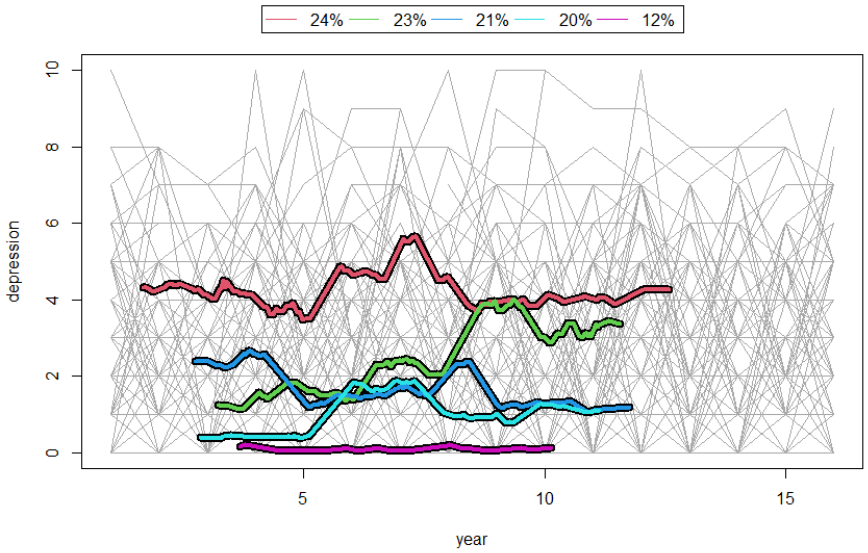


Abbildung 34: Longitudinales Clustering für "depression" mit kmlShape für k = 5

Im Allgemeinen unterstützen longitudinale Clustermethoden die Mustererkennung von zeitlichen Verläufen massgeblich. Die Methoden zur Bewertung von Ähnlichkeit verschiedener Trajektorien werden in den Paketen *kml, kml3d* und *kmlShape* gut erfasst und sind individuell anpassbar durch die Wahl von Parametern, unterschiedliche Distanz- und Qualitätskriterien. Die gegenseitige Ergänzung

[40] Weil der Suchalogrithmus von *kmlShape* für zu umfängliche Datensätze nicht konvergiert, wurde für die erstellte Abbildung eine Teilmenge von 300 Untersuchungseinheiten verwendet.

dieser Methoden ermöglicht die Einnahme verschiedener Blickwinkel auf dasselbe Untersuchungsobjekt, weshalb sie ein wichtiges Werkzeug bei der Generierung und Überprüfung von Hypothesen darstellen.

7.9 IV: Instrumentelle Variablen

Eine fortgeschrittene Methode bei der Modellierung kausaler Inferenzen ist das Prinzip der instrumentellen Variablen (IV). Da grossräumig und langzeitlich angelegte Paneldatensätze ("observational data") grundsätzlich Endogenität aufweisen und nicht in einem experimentellen Setting - unter Kontrolle möglicher Einflüsse - erhoben werden, stellt sich die Frage, wie die forschende Person adäquat mit dieser Endogenität umgehen kann. Unbeobachtete Heterogenität aufgrund fehlender Drittvariablen [41] kann eine Regressionsanalyse massgeblich beeinflussen [42], ohne Möglichkeit den Ursprung der unbeobachteten Heterogenität zu identifizieren. Für Situationen, in denen weitere Drittvariablen nicht zur Verfügung stehen, um die bedingten Zusammenhänge zu modellieren, stellen IV's eine passende Alternative dar. (Pokropek, 2016)

Das Prinzip von IV's beruht auf der Annahme, dass eine Variable Z – das sogenannte *Instrument* – den exogenen (zufälligen) Teil der Variabilität eines endogenen Prädiktors X bestimmt. Der endogene Teil der Variabilität von X wird im Modell nicht verwendet, da dieser durch unbeobachtete Heterogenität (vernachlässigte Drittvariablen U) beeinflusst ist. Durch die Bestimmung von X anhand von Z wird nur der Teil der Variabilität von X verwendet, der nicht durch unbeobachtete Heterogenität U beeinflusst ist. Eine mögliche Verletzung der Exogenitätsannahme wird dadurch behoben und ein konsistenter Zusammenhang zwischen abhängiger und unabhängigen Variablen geschaffen, wodurch die Schätzung von OLS-Koeffizienten konsistent wird.

Ein gültiges Instrument Z darf folglich keinen direkten Effekt auf Y haben, sondern nur indirekte Effekte über die Variable X (vgl. Abbildung 35). Sofern Z eine Verbindung zu U aufweist, ist das Instrument ebenfalls beeinflusst und gemäss Definition nicht gültig.

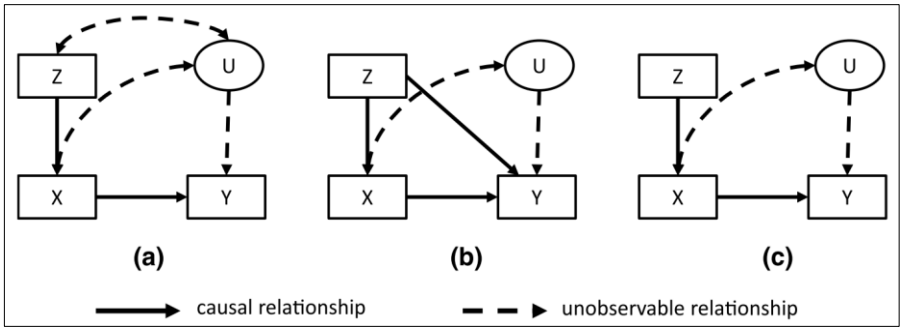


Abbildung 35: Beispiele eines ungültigen Instruments Z (a,b) und eines gültigen Instruments (c) (Pokropek, 2016)

Für das Untersuchungsszenario kann beispielsweise angenommen werden, dass die Variable a*usbildung* keinen direkten Einfluss auf die abhängige Variable d*epression* (Y) ausübt, jedoch als Instrument (Z) für die unabhängige Variable *arbeit_einbezug_entscheidungen* (X) eingesetzt werden kann. Diese Annahme

[41] omitted variable bias

[42] In diesem Fall wird ein *bias* im Modell eingeführt, in dem eine Variable im Modell den Effekt einer unbeobachteten Drittvariable transportiert.

ist durchaus berechtigt, da im Allgemeinen eine höhere Schulbildung mit mehr Entscheidungskompetenz im Berufsleben einhergeht.

Die Modellierung mittels instrumenteller Variablen geschieht anhand eines zweistufigen Regressionsverfahrens (2SLS – two stage least squares). Dabei wird im ersten Schritt jede unabhängige Variable X_i anhand der Instrumente Z_j und der anderen Kovariaten $X_1, X_2, \dots X_{i-1}, X_{i+1}, \dots X_n$ anhand einer linearen Regression modelliert. Dadurch wird sichergestellt, dass keine endogenen Einflüsse aufgrund unbeobachteter Heterogenität (Drittvariablen U) in die Prädikatoren X einfliessen. Im zweiten Schritt wird die abhängige Variable Y durch die zuvor modellierten abhängigen Variablen X_i modelliert.

Als Anwendungsbeispiel werden für das Untersuchungsszenario folgende Instrumente Z_j definiert: *ausbildung, alter, geschlecht, ch_nationalitaet, haushaltsaequivalenzeinkommen, arbeit_qualifikation, arbeit_zeit_wochenstunden, arbeit_zeit_nacht, hausarbeit_wochenstunden, kinder_betreuung* und *pflege_angehoerige.* Abbildung 36 zeigt die geschätzten Effekte der Kovariaten X_i basierend auf dem Modell mit den Instrumenten Z_i. Weitere Details zum Modell befinden sich in Anhang N. Dieses Modell scheint zwar gesamthaft signifikant (p-Wert = 9.83e-8), jedoch kann das Modell die Varianz der Daten nicht erklären (R-Squared = 0.020, Adjusted R-Squared = -0.192).

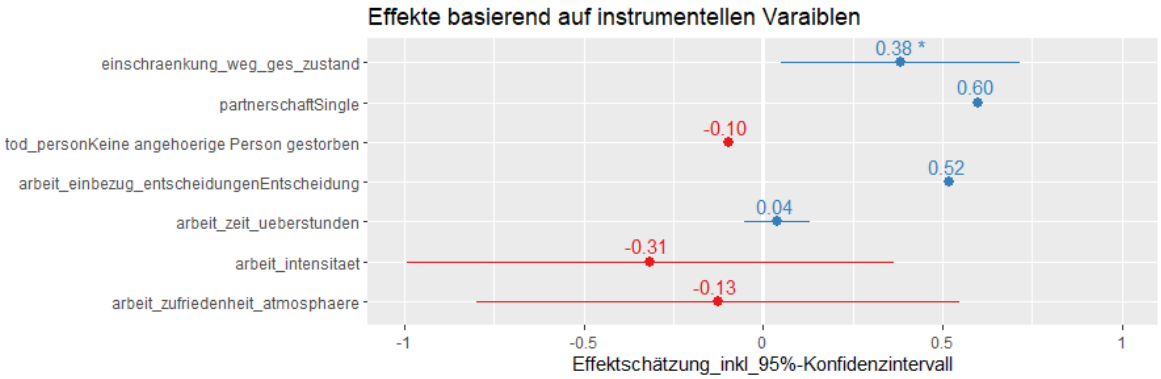


Abbildung 36: Effektschätzung anhand instrumenteller Variablen (blau > 0, rot < 0)

Die schwache Aussagekraft des Modells liegt darin begründet, dass die gewählten Instrumente sehr schwach sind. Wenn die gewählten Instrumente Z_j eine ungenügende Verbindung zu den Kovariaten X_i aufweisen, ist die Modellierung in Schritt 1 unpräzise und entsprechend die Modellierung in Schritt 2. Es scheint glaubhaft, dass die gewählten Instrumente nicht oder nur ungenügend in der Lage sind, die Kovariaten X_i adäquat zu modellieren. Eine inhaltliche Diskussion über mögliche Zusammenhänge und die Hinzunahme weiterer Instrumente kann bei dieser Problematik Abhilfe schaffen.

Die Herausforderung dieser Methode liegt in der Bestimmung valider Instrumente Z. Bei der Verwendung von instrumentellen Variablen sollte die forschende Person die Stärke der Instrumente überprüfen. Hanck et al. (2020, Kapitel 12.3) empfehlen als Faustregel eine F-Statistik von 10 bei der Überprüfung der Nullhypothese H_0: *alle Instrumente* Z_i *haben Koeffizient null* in Schritt 1 der 2SLS.

7.10 JMCM: Gemeinsame Mittelwert-Kovarainz Modellierung

Eine fortgeschrittene Methode zur Analyse von Paneldaten ist die Betrachtung der zeitlichen Entwicklung einer Variablen in Form einer Zeitreihe. Das Verständnis der inhärenten Dynamik einer Zeitreihe kann wichtige Erkenntnisse darüber liefern, ob die zeitliche Entwicklung der Variable von Interesse ein Muster im Sinne eines AR- oder MA-Prozesses[43] aufweist. Das R-Paket `jmcm` [44] (Pan & Pan, 2017) – *joint mean-covariance modeling framework* - liefert ein neueres Werkzeug zur longitudinalen Datenanalyse unter Berücksichtigung von Panelstrukturen resp. wiederholten Messungen (repeated measures).

Das Grundprinzip der Mittelwert-Kovarianz Modellierung beruht auf der Betrachtung eines abhängigen Merkmals $Y_{ij,\ i\in\{1,\dots,n\},j\in\{1,\dots,m\}}$ von n Untersuchungseinheiten, über m Zeitschritte. Die Menge aller Beobachtungen lässt sich beschreiben als Matrix $\boldsymbol{Y}$, die aus einer Menge von Vektoren $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{im})^T$ besteht.

$$\boldsymbol{Y} = (Y_1, Y_2, \dots Y_i, \dots, Y_n) = \ Y_{ij} = \begin{pmatrix} Y_{11} & \cdots & Y_{n1} \\ \vdots & \ddots & \vdots \\ Y_{1m} & \cdots & Y_{nm} \end{pmatrix} \tag{16}$$

Im Mittelwert-Kovarianz Framework für longitudinale Daten wird jedes Element Y_i beschrieben durch eine multivariate Normalverteilung um den einheitenspezifischen Mittelwert $\mu_i = (\mu_{i1}, \mu_{i1}, \dots, \mu_{im})^T$ und die $m \times m$ "within-subject" Kovarianzmatrix Σ_i. Der einheitenspezifische Mittelwert wird dabei modelliert durch eine lineare Regression, basierend auf weiteren Kovariaten X_i. (Pan & Pan, 2017)

$$Y_i \sim \mathcal{N}(\mu_i, \Sigma_i), \quad \mu_i = X_i\beta \tag{17}$$

Das Ziel der Mittelwert-Kovarianz Modellierung ist die Findung der zugrundeliegenden Kovarianz-Struktur Σ_i. Diese gibt Aufschluss darüber, wie sich die Gesamtheit der zeitlichen Messungen entwickelt und lässt Rückschlüsse auf die Struktur dieser Entwicklungen zu.

Für das Untersuchungsszenario und den vorhanden Paneldatensatz ist bereits vor einer Modellierung mit `jmcm` offensichtlich, dass ein Modell über den gesamten Datensatz keine verwertbaren Resultate hervorbringen wird. Aufgrund weniger Ausprägungen für Depression (0-10) und der Mannigfaltigkeit der (steigenden, fallenden und neutralen) Verläufe wird eine Mittelwert-Kovarianz-Modellierung auf Schwierigkeiten stossen bei der Identifikation generischer Muster. Aus diesem Grund wird die `jmcm`-Modellierung direkt auf den einzelnen Clustern aus Kapitel 7.8 durchgeführt (vgl. Abbildung 30). Aufgrund ähnlicher zeitlicher Verläufe von Depression pro Cluster, ist die Wahrscheinlichkeit für die Enthüllung interessanter Kovarianz- resp. AR- und MA-Strukturen somit höher.

Abbildung 37 und Abbildung 38 zeigen die errechneten Mittelwert-Kovarianz-Modelle für Cluster 1 und 4 der vorliegenden Paneldatensatzes. Für beide Modelle wird jeweils nur die abhängige Variable $depression$ betrachtet. Weitere Kovariaten X_i werden in diesen Modellen nicht berücksichtigt, damit die Kovarianz-Struktur der abhängigen Variable isoliert untersucht werden kann. Der obere Teil der Abbildungen zeigt jeweils die beobachteten Werte von $depression$ über die Zeit inkl. der Mittelwert-

[43] AR: autoregressiv, MA: moving average

[44] Vgl. Coprehensive R Archive Network: https://cran.r-project.org/web/packages/jmcm/index.html

Modellierung. Der untere Teil der Abbildungen zeigt jeweils die modellierte Entwicklung der logarithmierten Innovationsvarianz[45] (links) und eine Schätzung autoregressiver Koeffizienten (rechts).

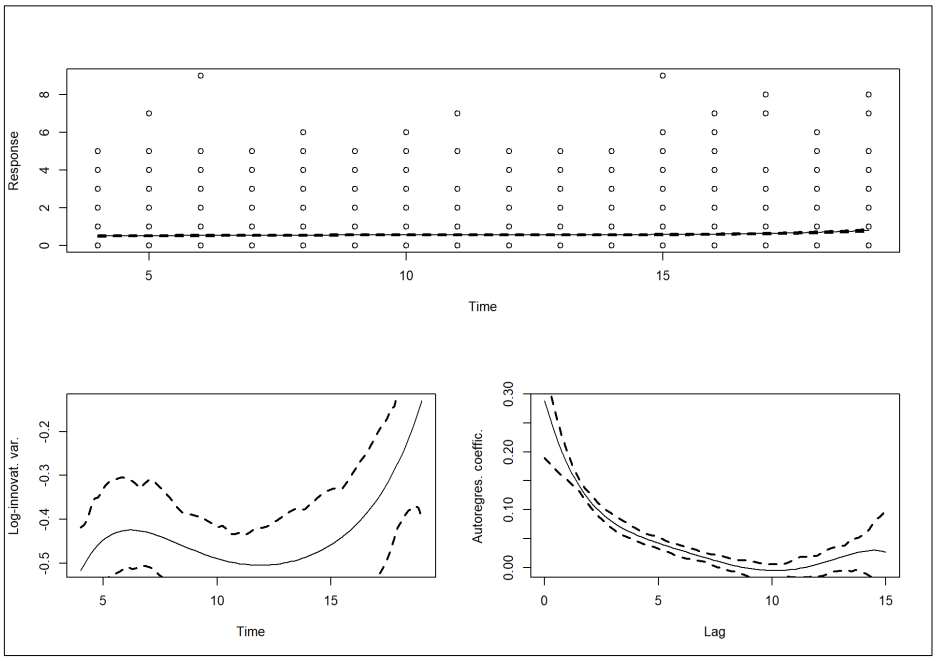


Abbildung 37: Mittelwert-Kovarianz-Modell für Cluster 1 des Paneldatensatzes

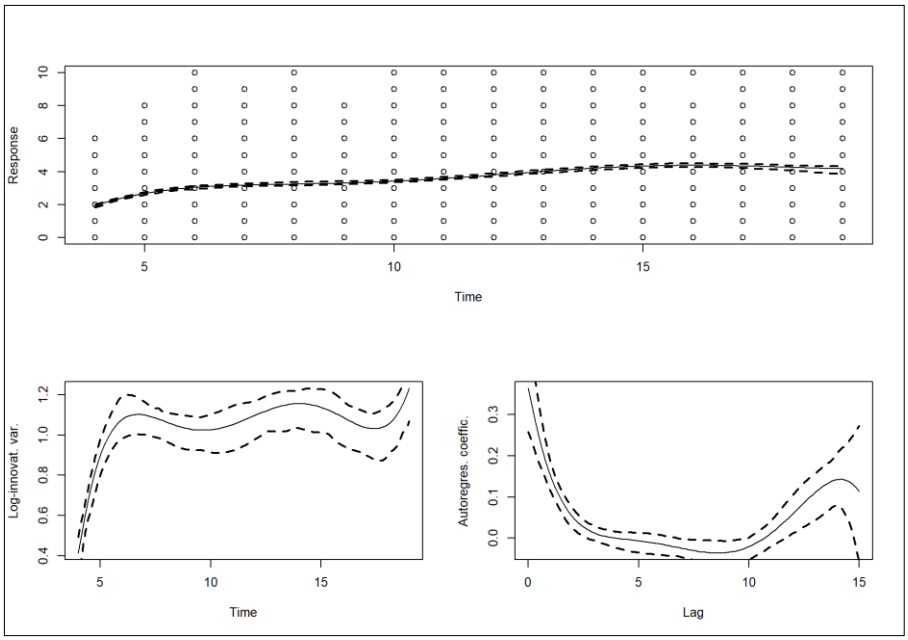


Abbildung 38: Mittelwert-Kovarianz-Modell für Cluster 4 des Paneldatensatzes

[45] Die Innovation bezeichnet in der Zeitreihenanalyse den Teil einer Zeitreihe, der als neue Information hinzukommt. Oft wird hierfür der Ausdruck E_t verwendet. Für ein AR(p)-Modell gilt dann: $\Phi(B) \cdot X_t = E_t$ und für ein MA(q)-Modell: $X_t = \Theta(B) \cdot E_t$

Aus dem Vergleich von Cluster 1 und 4 ist anhand von Abbildung 37 und Abbildung 38 ersichtlich, dass die mittlere Innovationsvarianz in Cluster 1 deutlich tiefer ist als in Cluster 4. Dies ist ein eindeutiges Indiz dafür, dass Untersuchungseinheiten mit höheren mittleren Werten für Depression ebenfalls höhere mittlere Schwankungen aufweisen. Zusätzlich ist erkennbar, dass die Innovationsvarianz in Cluster 4 relativ konstant bleibt, wohingegen in Cluster 1 eine Zunahme in den späteren Jahren (2015 – 2019) angezeigt wird. Dies könnte jedoch ein Scheineffekt sein, da die Innovationsvarianz auf polynomialen Schätzungen [46] basiert. Das autoregressive Verhalten ist in beiden Fällen ähnlich. Grössere AR-Koeffizienten kommen ausschliesslich für Lag 1 bis ungefähr Lag 5 vor, was darauf hindeutet, dass bei $depression$ im Allgemeinen eine Abhängigkeit zu den Vorjahren (ca. 1 – 5) existiert. Diese Erkenntnis ist nicht weiter erstaunlich, da wir erwarten, dass die zeitliche Entwicklung kein zufälliger Prozess wie etwa ein *random walk* ist, sondern sehr wohl mit früheren Werten von Depression verknüpft ist.

Die Modellierung mit dem gemeinsamen Mittelwert-Kovarianz-Ansatz liefert Einblicke in die "gemittelte Zeitreihen-Sicht" von Paneldaten. Diese neue Art der Betrachtung eröffnet eine passende Ergänzung zu bestehenden Regressions-Methoden. Durch die Integration von Kovariaten X_i können viele potenzielle Korrelationsstrukturen untersucht und bewertet werden. Eine wichtige Annahme des `jmcm`-Pakets (Normalverteilung der abhängigen Variable) ist für den vorliegenden Paneldatensatz tendenziell verletzt, weshalb die Interpretation der Resultate abgesichert und nicht überstrapaziert werden sollte.

7.11 TVCM: Baumbasierte variable Koeffizienten Regression

Wird in einem linearen Regressionsmodell der Effekt einer unabhängigen Variablen X auf die abhängige Variable Y durch eine Drittvariable beeinflusst, so wird diese Drittvariable im Allgemeinen als Störvariable betrachtet, sofern diese nicht explizit durch die forschende Person erfasst werden kann (vgl. Term U in Kapitel 7.9 "IV: Instrumentelle Variablen"). Kann diese Drittvariable jedoch erfasst werden, besteht die Möglichkeit, diese als zusätzliche Kovariate in das Regressionsmodell zu integrieren. Eine weitere Möglichkeit ist die Betrachtung dieser Drittvariable als Moderatorvariable (auch *intervenierende Variable* oder *Mediatorvariable*).

Eine Moderatorvariable Z_i bezeichnet eine Variable, die den Einfluss von X_i auf Y beeinflusst und dadurch den Regressions-Koeffizienten von X_i beeinflusst resp. moderiert. Im Kontext eines GLS[47] lässt sich das Prinzip der Moderatorvariable gemäss Bürgin & Ritschard (2017, S. 1) anhand von Formel (18) beschreiben. Dabei bezeichnet g eine beliebige Linkfunktion und X_p die p Prädikatoren. Der Vektor $\boldsymbol{Z_P} = \left(Z_{p1}, \dots, Z_{pL_p}\right)^T$ wird assoziiert mit dem Koeffizienten β_p und bezeichnet die L_p potenziellen Moderatorvariablen für diesen Koeffizienten. Jeder Koeffizient β_p ist somit eine Funktion, die abhängig ist von ihren entsprechenden Moderatorvariablen $\boldsymbol{Z_p}$.

$$g\big(E(Y|\cdot)\big) = X_1\beta_1(\boldsymbol{Z_1}) + X_2\beta_2(\boldsymbol{Z_2}) + \ \dots \ + X_p\beta_p\big(\boldsymbol{Z_p}\big) \tag{18}$$

[46] Für sämtliche Modellierungen in Abbildung 37 und Abbildung 38 wurden Polynome vom Grad 5 zugelassen.
[47] GLS: Generalized Least Square Modell – eine Verallgemeinerung des OLS Modells für nicht-normalverteilte abhängige Variablen.

Werden sämtliche Moderatorvariablen eines Koeffizienten als konstant betrachtet ($\boldsymbol{Z_p} = 1$), so führt dies zu einem "nicht-variierenden" Koeffizienten β_p. Für einen konstanten Prädiktor ($X_p = 1$) wird der Koeffizient $\beta_p(\boldsymbol{Z_p})$ zu einem variierenden Achsenabschnitt ("varying intercept"), der den direkten Effekt der Moderatoren $\boldsymbol{Z_p}$ auf $E(Y|\cdot)$ schätzt. Durch diese Formulierung können exogene Variablen als Prädikator, Moderator oder beides fungieren, wodurch grosse Freiheiten bei der Erkundung möglicher Moderatorvariablen entstehen. (Bürgin & Ritschard, 2017)

Das R-paket `vcrpart`[48] stellt ein intuitives Tool zur Analyse von Moderatorvariablen zur Verfügung. Der Ansatz des Pakets beruht auf einer Kombination aus linearer Modellierung und rekursiver Partitionierung ("recursive partitioning"). Dabei wird einerseits der Werteraum der möglichen Kombinationen von Moderatorvariablen $\boldsymbol{Z_p}$ aufgespannt und danach mittels rekursiver Partitionierung diejenigen Gruppen identifiziert, die ähnliche Regressionskoeffizienten hervorbringen. Die Darstellung des Einflusses einzelner Moderatorvariablen als Baumstruktur macht dieses Paket zum hilfreichen Analysetool für den Anwender.

Die Nomenklatur zur Modellierung mit `vcrpart` basiert auf der gängigen Formelstruktur in R unter der Verwendung eines zusätzlichen Terms $vc()$ um Moderatorvariablen zu identifizieren. Tabelle 5 zeigt die Verwendung dieses Terms, um den direkten Effekt von Moderatorvariablen auf die abhängige Variable zu bestimmen, oder die Moderation anderer Variablen (Beeinflussung des Effekts von x_i auf y).

Tabelle 5: Nomenklatur für Moderatorvariablen in vcrpart

Formel	Bedeutung
$y \sim x_1 + vc(z_1, z_2)$	Modell zur Bestimmung des Effekts von x_1 auf y. Aus den Moderatorvariablen z_1 und z_2 wird ein Baum erstellt, der aufzeigt welche Kombinationen von z_1 und z_2 gruppenweise unterschiedliche direkten Effekte auf *y* haben. → "varying intercept"
$y \sim x_1 + vc(z_1, z_2, by = x_1)$	Modell zur Bestimmung des Effekts von x_1 auf y. Aus den Moderatorvariablen z_1 und z_2 wird ein Baum erstellt, der aufzeigt welche Kombinationen von z_1 und z_2 gruppenweise unterschiedliche Einwirkungen auf den Effekt von x_1 auf *y* haben. → "varying slope"

Abbildung 39 zeigt den generierten Baum unter Verwendung des vorliegenden Paneldatensatzes und Formel (19). Gemäss dieser Formel wird ein leeres Modell, ohne unabhängige Variable x_i betrachtet, bei dem der Achsenabschnitt ("intercept") vernachlässigt wird.

```
depression ~ -1 + vc(ausbildung, arbeit_zeit_ueberstunden)        (19)
```

Durch die Berücksichtigung der zwei Moderatorvariablen $z_1 = ausbildung$ und $z_2 = arbeit\_zeit\_ueberstunden$ entsteht ein Baum, dessen Blätter gruppenweise ähnliche Kombination der zwei Moderatorvariablen identifizieren und den direkten Effekt dieser Gruppen auf die abhängige Variable bestimmen. Abbildung 39 zeigt, dass $ausbildung$ früher im Split verwendet wird und damit eine wichtigere Rolle spielt bei der Bildung ähnlicher Gruppen (Cluster) als $arbeit\_zeit\_ueberstunden$. Weiter kann bspw. geschlossen werden, dass $arbeit\_zeit\_ueberstunden$ vor allem zu Unterschieden

[48] Vgl. Coprehensive R Archive Network: https://cran.r-project.org/web/packages/vcrpart/index.html

bei Personen führt, die eine Sekundar-II oder Hochschulausbildung erfahren haben (und dies nur, wenn mehr als 3 Überstunden pro Woche geleistet werden).

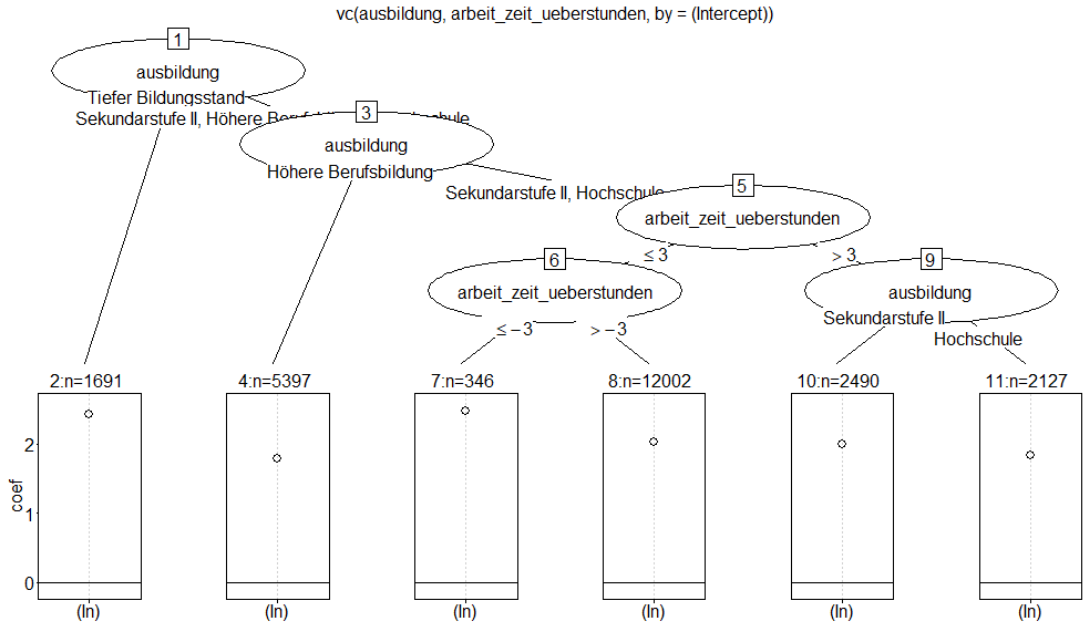


Abbildung 39: TVCM – Minimalbeispiel mit direkten Moderatoren "ausbildung" und "arbeit_zeit_ueberstunden"

Für das deutlich grössere Untersuchungsszenario kann analog vorgegangen werden, indem die gesamte Regressionsgleichung inklusive frei wählbarer Moderatorvariablen eingesetzt wird. Als Beispiel wird in Formel (20) der direkte Einfluss von $z_1 = partnerschaft$ auf die abhängige Variable sowie der Einfluss von z_1 und $z_2 = ausbildung$ auf den Effekt von $x_{11} = arbeit\_zeit\_wochenstunden$ untersucht.

```
depression ~ ausbildung + alter + geschlecht + ch_nationalitaet +
    einschraenkung_weg_ges_zustand + haushaltsaequivalenzeinkommen +
    partnerschaft + tod_person + arbeit_einbezug_entscheidungen +
    arbeit_qualifikation + arbeit_zeit_wochenstunden + arbeit_zeit_ueberstunden +
    arbeit_zeit_nacht + arbeit_intensitaet + arbeit_zufriedenheit_atmosphaere +
    hausarbeit_wochenstunden + kinder_betreuung + pflege_angehoerige +
    vc(partnerschaft) +
    vc(partnerschaft, ausbildung, by = arbeit_zeit_wochenstunden)
```
(20)

Mittels eines GLS-Modells wird für sämtliche Kovariaten x_i der Effekt geschätzt und die entsprechende Baumstruktur für die Moderatorvariablen erzeugt[49]. Abbildung 40 zeigt, dass der direkte Effekt von $partnerschaft$ auf $depression$ grösser ist bei Singles als bei Untersuchungseinheiten in einer Partnerschaft. Der moderierende Einfluss von $partnerschaft$ auf den Effekt von $arbeit\_zeit\_wochenstunden$ zeigt einen entgegengesetzten Effekt. Gemäss Abbildung 41 wird der Effekt von $arbeit\_zeit\_wochenstunden$ auf $depression$ tendenziell kleiner für Singles als für Untersuchungseinheiten in einer Partnerschaft (dies gilt für sämtliche Ausprägungen von $ausbildung$). Anhand dieses Beispiels kann folglich die Hypothese aufgestellt werden, dass das "Single-sein" zwar

[49] Ohne die Verwendung von Moderatorvariablen könnte die Regressionsgleichung einem gewöhnlichen GLS-Algorithmus (z.B. `stats::glm`) übergeben werden. Dies würde dieselben Effektschätzungen hervorbringen.

einen negativen Zusammenhang mit Depression hat, jedoch andere Effekte durch das "Single-sein" in eine positive Richtung bezüglich Depression moderiert werden. Die statistische Signifikanz dieser Effekt-Beeinflussung ist auf dem 0.05-Signifikanzniveau in den meisten Fällen abgesichert (weitere Details befinden sich in Anhang O).

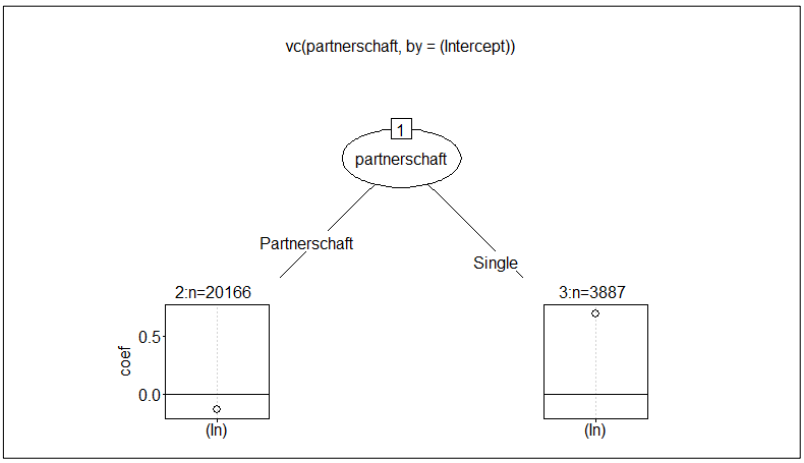


Abbildung 40: TVCM - Direkter Einfluss von "partnerschaft".

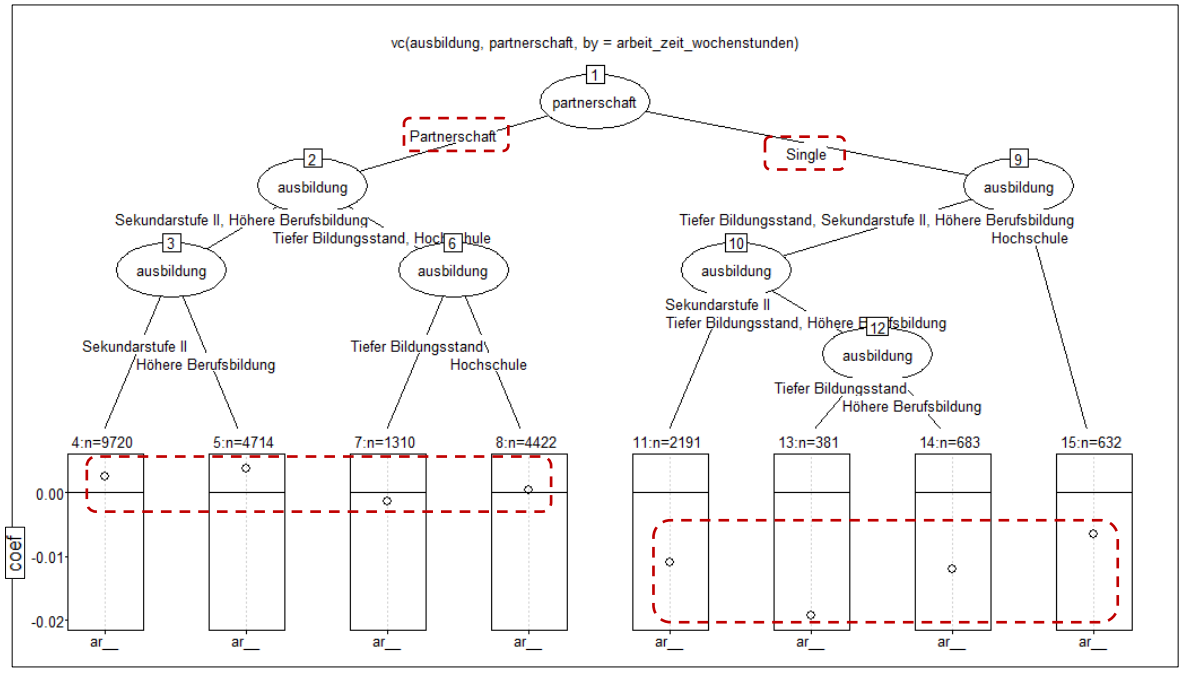


Abbildung 41: TVCM - Moderierender Einfluss von "partnerschaft" und "ausbildung" auf "arbeit_zeit_wochenstunden"

Das baumbasierte, variable Koeffizienten Regression bietet ein generisches und vielseitig einsetzbares Werkzeug zur Untersuchung von Moderatorvariablen bei der linearen Regression. Es ist gut geeignet für die Unterstützung eines Modellierungsprozesses und zur Erarbeitung oder Überprüfung von Hypothesen. Aufgrund der GLS-basierten Regressionslogik können weitere Link-Funktionen verwendet werden, was

den Anwendungsbereich des `vcrpart`-Pakets auf sinnvolle Art erweitert[50]. Eine explizite Betrachtung von Längsschnitten wird in diesem Paket nicht unterstützt. Durch die Gewichtung einzelner Messungen kann der zusätzliche Informationsgehalt von Beobachtungen derselben Untersuchungseinheit zwar kontrolliert werden, ein Koeffizientenvergleich ergab jedoch sehr kleine Unterschiede für gewichtete Effektschätzungen. Obwohl Erweiterungen für die Verwendung von FE- und RE-Termen in Regressionsgleichungen existieren, wird aufgrund mangelnder Dokumentation von deren Verwendung abgeraten, solange die Modelle nicht von Grund auf verstanden sind.

[50] In den hier verwendeten Modellen wurde aus Gründen der Interpretierbarkeit eine normalverteilte abhängige Variable angenommen. Gemäss Kapitel 6.3.1 wäre die Betrachtung einer Poisson-verteilten abhängigen Variable sinnvoll und die Verwendung des Logarithmus als Link-Funktion angezeigt.

– Teil 3 –

Ergebnis

8 Beantwortung der Forschungsfrage

8.1 Methodenvergleich: Erkenntnistheoretischer Mehrwert

Sind fortgeschrittene Machine Learning Methoden nun tatsächlich in der Lage, den Erkenntnisgewinn von State-of-the-Art-Methoden der longitudinalen Paneldatenanalyse sowohl in der Tiefe als auch in der Breite zu ergänzen? Die Erfahrungen aus der Anwendung von 10 Methoden erlauben die Beantwortung der Forschungsfrage mit "Ja".

Aufgrund der mehrdimensionalen Struktur von Paneldaten gibt es nicht das *richtige Modell* für die Analyse eines solchen Datensatzes. Speziell im Kontext sozialwissenschaftlicher Paneldaten hat sich gezeigt, dass der zugrundeliegende Datengenerierungsprozess (der Mensch) die Komplexität der zur Verfügung stehenden Modelle überschreitet. Jedes Modell hat seine Vor- und Nachteile für die Betrachtung spezifischer Aspekte eines Paneldatensatzes.

Aus diesem Grund bietet die komplementäre Sicht aus verschiedenen Blickwinkeln eine gute Strategie, um den gesamten Erkenntnisprozess zu fördern. Die Triangulation (Einnahme verschiedener Blickwinkel), welche als Methode zur Bewertung des zusätzlichen Erkenntnisgewinns eingesetzt werden sollte, stellt sich somit selbst als *die "richtige Methode"* heraus, um den Erkenntnisgewinn der sozialwissenschaftlichen Paneldatenanalyse in einem sinnvollen Prozess voranzutreiben.

Doch was ist der Blickwinkel auf einen Paneldatensatz und wie ist dieser definiert? Abbildung 42 zeigt eine Veranschaulichung dieses Konzepts. Darin setzt sich der Blickwinkel auf einen Paneldatensatz aus dem Erkenntnisinteresse der forschenden Person, der Panelsicht (längs vs. quer) und dem Bezugsobjekt (gesamte Population, Teilpopulation oder Individuum) zusammen.

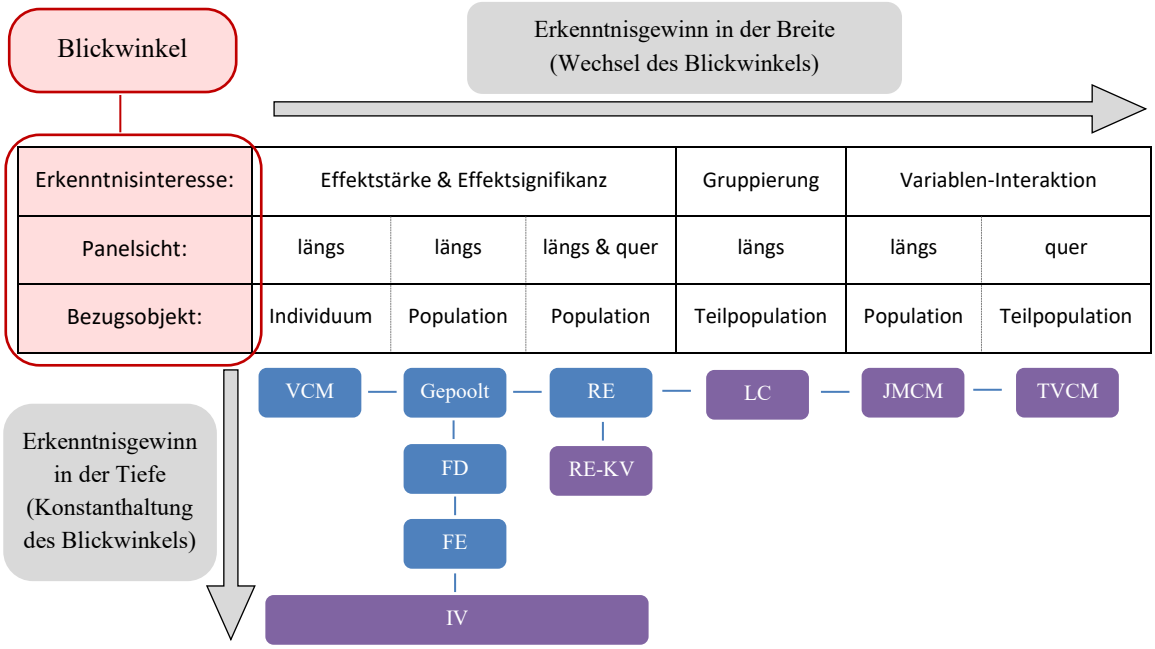


Abbildung 42: Gegenüberstellung der verwendeten Methoden nach ihrem Blickwinkel und Erkenntnisgewinn (blau: state-of-the-Art Methoden, violet: fortgeschrittene Methoden), eigene Darstellung.

Gemäss Abbildung 42 lassen sich die 10 verwendeten Methoden in Gruppen mit ähnlichen resp. unterschiedlichen Blickwinkeln einteilen. Dadurch wird grafisch erklärt, dass fortgeschrittene Methoden sowohl in der Tiefe als auch in der Breite eine Ergänzung des Erkenntnisgewinns darstellen. Die Positionierung einer Methode auf der Ordinatenachse ist nicht als fixiert anzusehen. D.h. eine Methode, die in der Grafik weiter unten positioniert ist, sollte nicht als "besser" oder "wertvoller" als eine Methode interpretiert werden, die denselben Blickwinkel hat, jedoch weiter oben positioniert ist. Somit bringt bspw. ein FE-Modell nicht zwingend bessere Erkenntnisse hervor als ein FD-Modell, sondern ist vielmehr eine Ergänzung des FD-Modells für denselben Blickwinkel. Weil das IV-Modell (resp. der IV-Ansatz) bei Längs- und/oder Querschnittbetrachtungen eingesetzt werden kann, stellt dieses eine Ergänzungen in mehreren Bereichen dar.

Abbildung 42 ist nicht abschliessend und kann durch die Betrachtung zusätzlicher Methoden weiter ergänzt werden. Des Weiteren ist anzumerken, dass die Verortung einer Methode in dieser Abbildung nicht zwingend eindeutig ist. So kann longitudinales Clustering (LC) bspw. nicht nur als Methode zum "isolierten Erkenntnisgewinn in der Breite" angesehen werden, sondern aufgrund von errechneten Clustern den Erkenntnisgewinn einer anderen Methode (z.B. FE) komplementär in der Tiefe ergänzen.

Methoden, die Effekte schätzen, können basierend auf den Regressionskoeffizienten direkt miteinander verglichen werden. Eine Tabelle der Koeffizienten (Gepoolt, FD, FE, RE, RE_KV, IV) befindet sich in Anhang P. Unterscheiden sich die Regressionskoeffizienten zwischen den Methoden bezüglich Richtung und Signifikanz, anerkennen sie folglich unterschiedliche Eigenschaften des Paneldatensatzes. In diesen Fällen ist schwer zu bewerten, welchen Koeffizienten mehr Bedeutung eingeräumt werden sollte. Als Qualitätskriterien könnte das "Mass der Verletzung von Modellannahmen" verwendet werden (vgl. Kapitel 7.6 "Tests für Panelmodelle"). So würden wir beispielsweise aufgrund des Hausman Tests den Koeffizienten des RE-KV Modells mehr Glauben schenken als jenen des RE-Modells. In jedem Fall sind unterschiedliche Regressionskoeffizienten nicht als Problem, sondern als komplementäre Sichtweisen desselben Sachverhaltes anzusehen, die es zu verstehen gilt.

Bezüglich Anwendbarkeit lassen sich die Methoden kaum unterscheiden. Sobald ein entsprechender Paneldatensatz vorhanden ist, liefern alle Methoden (resp. die entsprechenden Pakete) niederschwellige Schnittstellen zur direkten Anwendung in R. Unterschiede bezüglich Rechenzeit sind für den vorliegenden Paneldatensatz feststellbar, jedoch liegen alle Methoden in einem annehmbaren Rahmen von wenigen Sekunden bis max. 5 Minuten[51].

Die durchgeführten Untersuchungen zeigen auf, dass fortgeschrittene Machine Learning Methoden eine passende Ergänzung herkömmlicher Methoden darstellen. Im Rahmen weiterführender Untersuchungen könnten folgende Methoden die Blickwinkel auf einen Paneldatensatz noch weiter ergänzen:

- Strukturgleichungsmodelle[52] (SEQ: Structural Equation Models)
- Mehrebenenmodellierung[53] (Fokus auf *random slopes* und *cross-classification*)
- Neuronale Netzwerke[54] (NN)

[51] Die einzige Ausnahme bildet das *kmlShape*-Paket, welches für den vollständigen Paneldatensatz nicht konvergiert.

[52] vgl. R-Paket *lavaan*: https://cran.r-project.org/web/packages/lavaan/index.html

[53] vgl. R-Pakete *lme4* und *brms:* https://cran.r-project.org/web/packages/lme4/index.html, https://cran.r-project.org/web/packages/brms/index.html

[54] vgl. R-Paket *neuralnet:* https://cran.r-project.org/web/packages/neuralnet/index.html

8.2 Fallgruben und Chancen

Im Folgenden werden die wichtigsten Erkenntnisse aus den Modellierungsprozessen zusammengefasst. Diese sollen der sozialwissenschaftlich forschenden Person als Hinweise für eigene Analysen dienen.

Fallgrube: longitudinal ≠ individuell

Das Wort "longitudinal" impliziert in vielen Formulierungen, dass Individuen einzeln betrachtet werden und die errechneten Effekte individuelle "within"-Verläufe darstellen. Diese Aussage ist nur teilweise korrekt, da die errechneten Effekte nach wie vor einen *Mittelwert individueller Effekte* darstellen. So kann ein FE-Modell trotzdem einen Effekt auf null schätzen, obwohl signifikante Effekte vorhanden sind. Zur Identifikation von Gruppen mit ähnlichen Effekten bietet sich longitudinales Clustering an.

Fallgrube: State-of-the-Art-Methoden sind aussagekräftig

Die mathematische Komplexität hinter den gepoolten, FD-, FE-, RE-Modellen ist hoch und sollte nicht unterschätzt werden. Für den konkreten Anwendungsfall ist die korrekte Interpretation von Regressionskoeffizienten entscheidend. Bevor die State-of-the-Art-Methoden nicht getestet und die Resultate sinnvoll interpretiert wurden, ist von der Verwendung weiterführender Methoden abzuraten.

Chance: White Box

Sämtliche verwendeten Methoden erfordern die Definition eines vermuteten Zusammenhangs in der Art "$y \sim x_1 + x_2 + \ ...$". Dadurch wird implizit eine Vorstellung über die zu modellierenden Zusammenhänge erwartet, wodurch Anwender gezwungen werden, inhaltliche Überlegungen anzustellen. Methoden, die blind Daten entgegennehmen und Resultate ausgeben sind speziell im sozialwissenschaftlichen Kontext gefährlich und sollten mit Bedacht eingesetzt werden.

Chance: Mehrwert im Modellvergleich

Der direkte Vergleich von Modellen anhand von ANOVA, F-Tests, Chi-Quadrat-Tests etc. bietet für einen explorativen Modellierungsprozess ein wichtiges Tool zur Bewertung möglicher Stossrichtungen. Gleichzeitig geben Modelle, die ähnliche Aussagen erzeugen mehr Sicherheit für die Interpretation.

Fallgrube: Querschnitt vs. Längsschnitt

Methoden, die einheitenspezifische Heterogenität nicht vollständig kontrollieren, unterliegen immer der Gefahr einer Vermischung von Quer- und Längsschnittfragen. Beispielsweise betrachtet LC zwar longitudinale Trajektorien, die resultierenden Cluster sind jedoch per Definition durch Unterschiede im Querschnitt definiert. Die Beantwortung von Längsschnittfragen über mehrere Cluster hinweg wäre deshalb nicht sinngemäss.

Fallgrube: Nicht normalverteilte Variablen

Viele der verwendeten Modelle basieren auf OLS und somit auf der Annahme normalverteilter abhängiger Merkmale. Die forschende Person sollte sich der Modellannahmen bewusst sein und diese ggf. überprüfen.

9 Diskussion und Ausblick

9.1 Sozialwissenschaftliche Erkenntnisse

In diesem Kapitel werden sozialwissenschaftliche Resultate der erstellten Modelle kurz vorgestellt und interpretiert.

Beim Vergleich der Effekte im gepoolten, FD-, FE-, RE- und RE_KV-Modell (vgl. Anhang P und Abbildung 43) fällt auf, dass sich sämtliche Modelle bei einigen Grössen durchwegs einig sind. So hat der Partnerschaftsstatus $Single$ in allen Modellen eine verstärkende Auswirkung auf $depression$. Umgekehrt hat die Zufriedenheit mit der Arbeitsatmosphäre oder die passende Qualifikation für die Erwerbsarbeit eine senkende Auswirkung auf Depression. Für andere Variablen werden teils unterschiedliche Effekte festgestellt, die sich durch die unterschiedlichen Transformationen begründen lassen. So schätzt das FD-Modell beispielsweise einen ungewohnt tiefen Wert für $depression$ beim weiblichen Geschlecht. Dieser Ausreisser liegt darin begründet, dass das FD-Modell bei der Berechnung des Regressionskoeffizienten für $geschlecht$ ausschliesslich erste Differenzen und somit nur Personen betrachtet, die einen Wechsel des Geschlechts vorgenommen haben. Aufgrund der äusserst kleinen Population mit solchen Eigenschaften ist das entsprechende Konfidenzintervall gross und die Schätzung des Koeffizienten nicht signifikant.

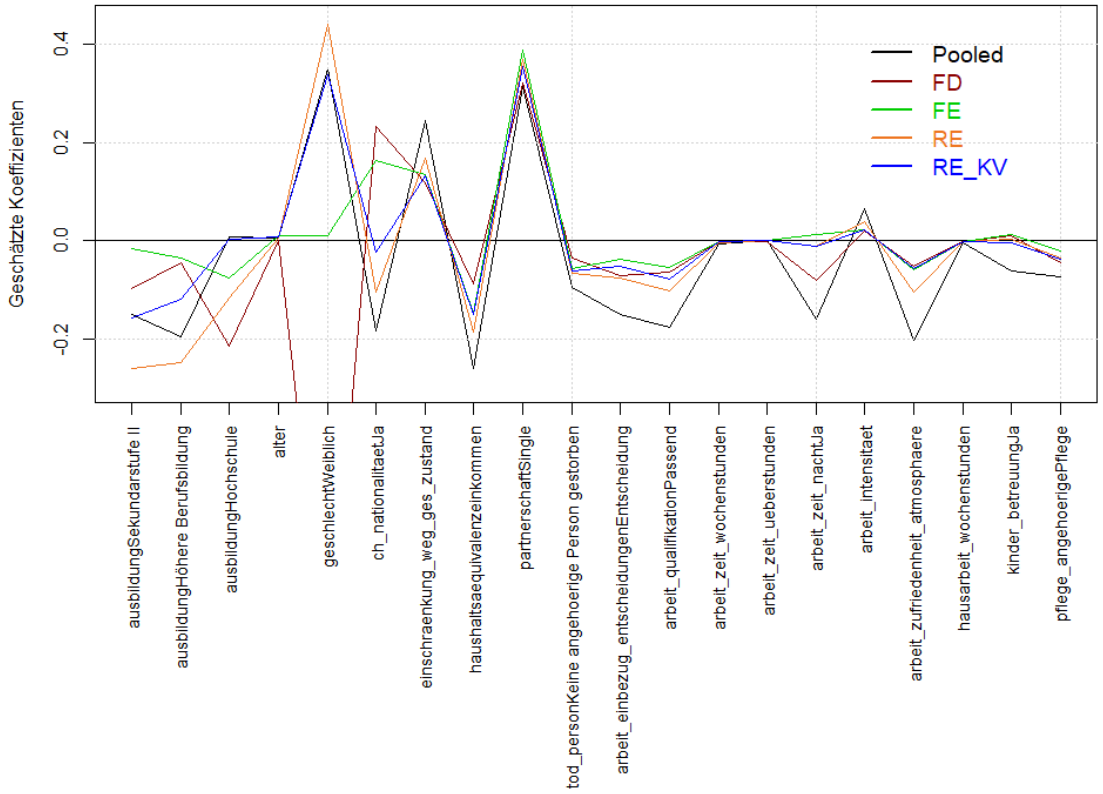


Abbildung 43: Geschätzte Koeffizienten von gepooltem, FD, FE, RE und RE_KV Modell

Des Weiteren fällt auf, dass das RE-KV-Modell oft einen "Mittelweg" zwischen dem FE- und RE-Modell darstellt. Diese Tatsache erstaunt nicht, da bereits Giesselmann & Windzio (2012) hervorheben, dass das hybride RE-KV-Modell die Eigenschaften von Längs- und Quermethoden vereint und somit ein

gewichtetes Bild von Längs- und Querschnittseffekten wiedergibt. Ob ein solcher Mittelweg gewünscht ist oder nicht, bleibt der forschenden Person überlassen.

Die gewonnenen Erkenntnisse zu direkten Zusammenhängen zwischen unabhängigen Variablen und $depression$, liefern ein erstes Bild und eine erste quantitative Baseline für die forschende Person. Zur besseren Vergleichbarkeit wurde in Kapitel 5 die Definition eines allgemeingültigen Untersuchungsszenarios gerechtfertigt. Für weitere Analysen des vorliegenden Paneldatensatzes bietet sich die Variation des Untersuchungsszenarios in Form von Transformationen, Interaktionstermen, potenzierter Terme oder Regression Splines an. Die Betrachtung von Interaktionstermen kann auf inhaltlicher Ebene viele Möglichkeiten eröffnen, um gleichzeitige Wirkungszusammenhänge zwischen den Sphären *spezifische Lebenslage*, *Arbeitsbelastung* und *Arbeitsbeanspruchung* zu erschliessen. Nach inhaltlichen Überlegungen zu den gewonnenen Erkenntnissen bieten sich auch andere Schritte an, um den datenbasierten Erkenntnisgewinn weiterzuführen, wie beispielsweise:

- Anpassung der selektierten unabhängigen Variablen
- Anpassung der Operationalisierung
- Mitberücksichtigung von Multikollinearität bei der Regression
- Reduktion der Untersuchungsszenarien in kleinere Themenblöcke

9.2 Paneldatenanalyse im sozialwissenschaftlichen Kontext

Aus Sicht des Data Scientists können alle bisher verwendeten Modelle in ihrer Aussagekraft kritisiert werden. Verletzungen von Modellannahmen, Endogenität oder unbeobachtete Heterogenität sind oftmals schwierig oder unmöglich zu eliminieren und werden zur unangenehmen Realität der Sozialwissenschaft. Dennoch liefern die verwendeten Modelle verwertbare Indizien, um einen explorativen Modellierungsprozess zu unterstützen. Durch die Identifikation signifikanter Effekte oder Personengruppen mit ähnlichen Merkmalen, können so weitere Untersuchungen motiviert werden.

Auf der anderen Seite bietet die Identifikation dieser "mittleren Effekte" die Möglichkeit, Ausreisser fernab der Norm zu identifizieren. Als Nebenprodukt der (globalen) Paneldatenanalyse können somit Spezialfälle identifiziert werden, die für sozialwissenschaftliche Teilfragen von grossem Interesse sein können.

Wie in Kapitel 8.1 aufgezeigt, ist ein Zugang aus mehreren Blickwinkeln ein wichtiges Werkzeug für die mit Paneldaten agierenden Forscher. Eine wichtige Prämisse zur Einnahme verschiedener Blickwinkel, ist die fachlich/inhaltliche Expertise von sozialforschenden Personen. Erst durch die vage Vorstellung, die konkrete Idee oder die fertig ausformulierte Hypothese, kann ein Prozess angestossen werden, in dem iterativ neue Konzepte erarbeitet werden. Ein solcher Prozess oder Workflow wird in Abbildung 44 vorgeschlagen. Zur erfolgreichen Erarbeitung von Erkenntnissen oder Hypothesen durchläuft die forschende Person mehrere Stufen, in welchen iterativ Paneldaten (Empirie) und sozialwissenschaftliche Expertise in Form zugrundeliegender Theorien verwendet werden (Rationalismus). Dieser Workflow wird im Optimalfall von beiden Seiten in iterativer Weise gelenkt, damit sich rationalistische und empirische Überlegungen immerzu überlagern. Dadurch wird sichergestellt, dass Modellierungsresultate kritisch hinterfragt und für die agile Weiterentwicklung von Ideen verwendet werden, und nicht simple Datenanalysen ohne konkrete Absichten durchgeführt werden.

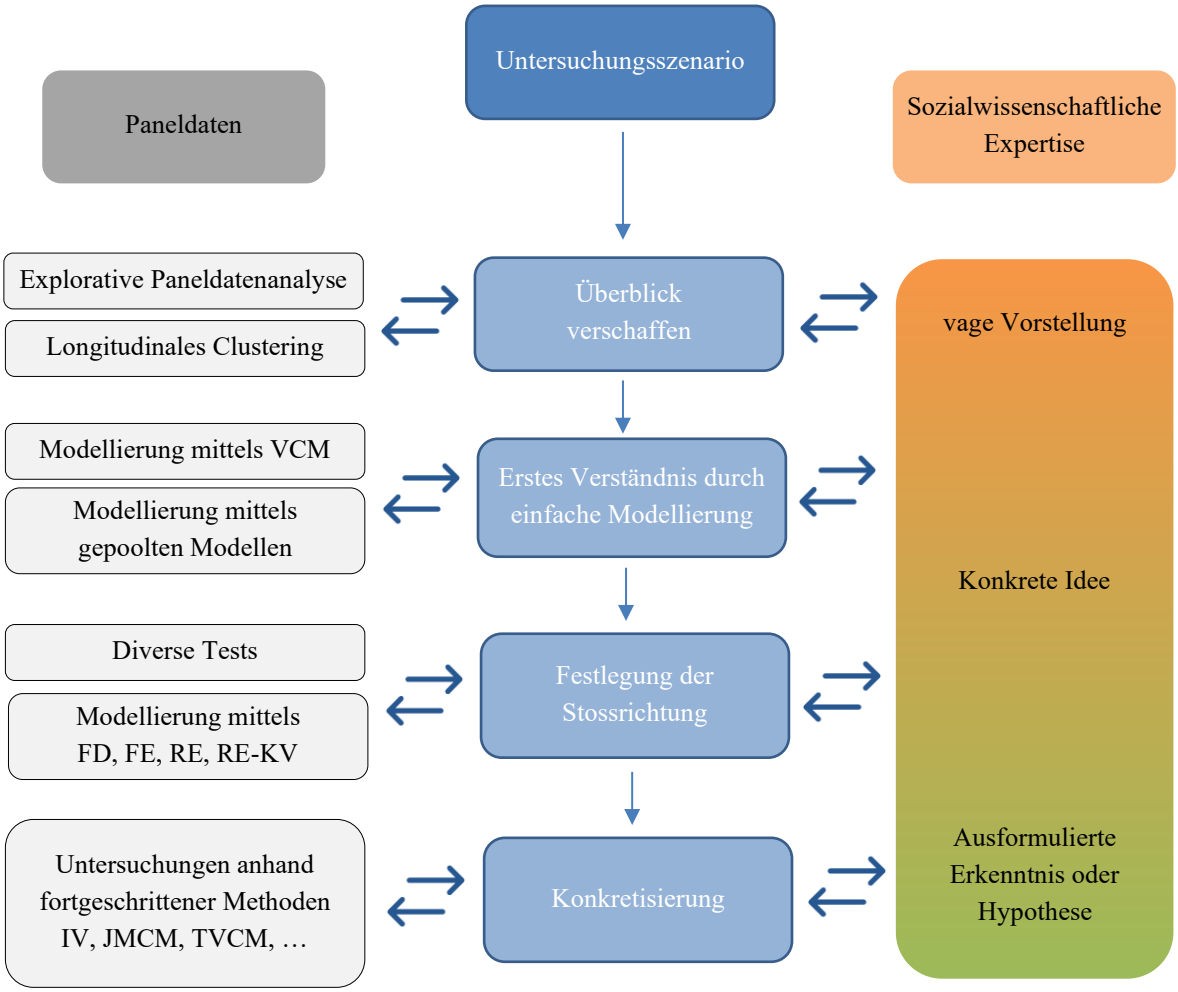


Abbildung 44: Workflow für Sozialforschende im Zusammenhang mit Paneldaten

Ein explorativer Zugang zur Hypothesengenerierung gemäss Abbildung 44 wird in der gängigen Literatur dennoch wenig gewürdigt. In Anbetracht einer realen Analysesituation, in der verschiedene Wege eingeschlagen werden und Methoden sowie Datensätze adaptiert werden, stellt der Prozess resp. der Workflow an sich ein entscheidendes Element für den Erkenntnisgewinn dar. Der vermehrte Miteinbezug von "Best Practices" in Form von *Blueprints für iteratives Denken und Modellieren in der Paneldatenanalyse* ist daher eine wünschenswerte Ergänzung der gängigen Methoden-Literatur.

9.3 Zusammenfassung und Ausblick

Die quantitativ-empirische Ausrichtung sozialwissenschaftlicher Forschung gemäss König (1962) wird durch State-of-the-Art sowie fortgeschrittene Methoden der Paneldatenanalyse unterstützt. Aufgrund von freiem Zugang zu Open Source Software und der engen Verknüpfung zwischen Forschungswelt und Open Source, stehen neue Methoden niederschwellig und zeitnah zur Verfügung.

Die daraus entstehende Sammlung an technischen Möglichkeiten bei der Paneldatenanalyse bietet der sozialforschenden Person die Möglichkeit, einen datenbasierten Erkenntnisgewinn sowohl in Tiefe[55] als auch in der Breite[56] besser abzustützen. Eine rein empirische Ausrichtung der sozialwissenschaftlichen Forschung birgt speziell im Kontext der aktuellen Methodenverfügbarkeit die Gefahr eines radikalen Empirismus (Coveney et al., 2016). Aufgrund zunehmend fortgeschrittener Methoden steigt die Gefahr, dass Schlussfolgerungen nicht mehr aufgrund struktureller Erklärungen zustande kommen, sondern aufgrund modellierter Korrelationen und Interpolationen, welche für Anwender schwer nachzuvollziehen sind.

Die durchgeführten Untersuchungen motivieren die Aussage, dass aktuelle Erklärungsmechanismen nicht ausreichen, um das komplexe System Mensch aufgrund mathematischer Modelle zu erklären. Vielmehr folgt die Erkenntnis, dass die Einnahme verschiedener Blickwinkel sowie komplementäre Methoden den Wissenszuwachs bei der Paneldatenanalyse unterstützen. Für das Gelingen eines solchen Wissenszuwachses, benötigen Forschende nicht nur die quantitativ-empirische Ausrichtung, sondern ebenfalls die des Rationalismus nach Coveney et al. (2016). Im Prozess der Bildung sinnvoller Erkenntnisse spielt die gemeinsame Evolution von Empirie und Rationalismus deshalb eine wichtige Rolle.

Künftige Untersuchungen sollten darauf abzielen, Methoden der Paneldatenanalyse einem breiten sozialwissenschaftlichen Publikum näher zu bringen und mathematische Komplexität in sinnvoller Art und Weise auf zentrale und einfach verständliche Konzepte herunterzubrechen. Gleichzeitig sollten Ansprüche an einen sozialwissenschaftlichen Modellierungs-Workflow sauber definiert und bei der Implementierung von Methoden aktiv mitberücksichtigt werden. So könnte, durch die Zusammenführung von Methodenexperten ("Mathematiker/innen") und den Fachexperten ("Sozialforschende"), die unreflektierte Verwendung von Methoden vermieden und ein optimaler Erkenntnisprozess unterstützt werden.

[55] Erschliessung von präziseren Erkenntnissen aufgrund verbesserter, umfassenderer Methoden
[56] Erschliessung einer neuen Art von Erkenntnissen durch die Einnahme neuer Blickwinkel

10 Danksagung

Ein grosser Dank geht an meinen Referenten Adrian Stämpfli, der sich sämtlichen Herausforderungen während der Verfassung dieser Master-Thesis rasch und kompetent angenommen hat. Aufgrund seiner Arbeit im Rahmen des SNF-Forschungsprojekts, konnte ich viel implizites Wissen und Code für meine Arbeiten verwenden. Ebenso geht ein Dank an Stefan Paulus, Myriel Ravagli und Thomas Schmid vom IFSAR-OST für den einfachen und spannenden Austausch zu sämtlichen sozialwissenschaftlichen Fragen.

Literaturverzeichnis

Ahrens, H., & Pincus, R. (1981). On Two Measures of Unbalancedness in a One-Way Model and Their Relation to Efficiency. *Biometrical Journal*, *23*(3), 227–235. https://doi.org/10.1002/bimj.4710230302

Antal, E., & Rothenbühler, M. (2015). *Weighting in the Swiss Household Panel Technical report*. 1–23. http://forscenter.ch/wp-content/uploads/2014/12/Weighting_technical_report.pdf

Böhle, F. (2010). Kapitel VIII Subjekt und Arbeitskraft: Arbeit und Belastung. In *Handbuch Arbeitssoziologie* (pp. 451–481). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92247-8_15

Bürgin, R., & Ritschard, G. (2017). Coefficient-wise tree-based varying coefficient regression with vcrpart. *Journal of Statistical Software*, *80*(6). https://doi.org/10.18637/jss.v080.i06

Campbell, M. K., Mollison, J., & Grimshaw, J. M. (2001). Cluster trials in implementation research: estimation of intracluster correlation coeecients and sample size. In *STATISTICS IN MEDICINE Statist. Med* (Vol. 20). https://doi.org/10.1002/1097-0258(20010215)20:3

Collischon, M., & Eberl, A. (2020). Let's Talk About Fixed Effects: Let's Talk About All the Good Things and the Bad Things. *Kolner Zeitschrift Fur Soziologie Und Sozialpsychologie*, *72*(2), 289–299. https://doi.org/10.1007/s11577-020-00699-8

Coveney, P. V., Dougherty, E. R., & Highfeld, R. R. (2016). Big data need big theory too. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2080). https://doi.org/10.1098/rsta.2016.0153

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, *27*(2), 1–43. https://doi.org/10.18637/jss.v027.i02

Dalgaard, P. (2008). *Introductory Statistics with R*. https://doi.org/10.2307/2987227

Farrar, D. E., & Glauber, R. R. (1967). *Multicollinearity in Regression Analysis: The Problem Revisited* (Vol. 49, Issue 1). https://about.jstor.org/terms

Flick, U. (2008). *Triangulation: Eine Einführung. 2 - Google Scholar*. https://scholar.google.com/scholar?hl=de&as_sdt=0,5&cluster=4186263148982760103

FORSbase. (2020). https://forsbase.unil.ch/

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). Kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, *65*(4), 1–34. https://doi.org/10.18637/jss.v065.i04

Genolini, C., Ecochard, R., Benghezal, M., Driss, T., Andrieu, S., & Subtil, F. (2016). kmlShape: An efficient method to cluster longitudinal data (Time-Series) according to their shapes. *PLoS ONE*, *11*(6), 1–24. https://doi.org/10.1371/journal.pone.0150738

Giesselmann, M., & Windzio, M. (2012). Regressionsmodelle zur Analyse von Paneldaten. In *Regressionsmodelle zur Analyse von Paneldaten*. https://doi.org/10.1007/978-3-531-18695-5

Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2020). *Introduction to Econometrics with R*. https://www.econometrics-with-r.org/12-3-civ.html

IFSAR. (2019). *SNF Projekt: Psychosoziale Risiken in der Arbeitswelt - IFSAR-Blog*. Www.Ifsa.Ch. https://www.ifsa.ch/?p=6201

Journal of Statistical Software. (2020). https://www.jstatsoft.org/index

König, R. (1962). *Handbuch der empirischen Sozialforschung* (Vol. 1). https://search.proquest.com/docview/1299514607?pq-origsite=gscholar&fromopenview=true

Mundlak, Y. (1978). *On the Pooling of Time Series and Cross Section Data* (Vol. 46, Issue 1). https://about.jstor.org/terms

Oelerich, G., & Otto, H.-U. (2011). Empirische Forschung und Soziale Arbeit – Einführung. In *Empirische Forschung und Soziale Arbeit*. https://doi.org/10.1007/978-3-531-92708-4_1

Oppolzer, A. (2010). Psychische Belastungsrisiken aus Sicht der Arbeitswissenschaft und Ansätze für die Prävention. In *Fehlzeiten-Report 2009* (pp. 13–22). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-01078-1_2

Pan, J., & Pan, Y. (2017). Jmcm: An R package for joint mean-covariance modeling of longitudinal data. *Journal of Statistical Software*, *82*(9). https://doi.org/10.18637/jss.v082.i09

Paulus, S. (2019). Gefährdungsbeurteilungen von psychosozialen Risiken in der Arbeitswelt. Zum Stand der ForschungRisk assessment of psychosocial risks in labor. State of the art in research. *The Hans Böckler Foundation*, *73*(2), 141–152. https://doi.org/10.1007/s41449-018-0117-8

Pokropek, A. (2016). Introduction to instrumental variables and their application to large-scale assessment data. *Large-Scale Assessments in Education*, *4*(1). https://doi.org/10.1186/s40536-016-0018-2

R Markdown. (2020). *R Markdown*. https://rmarkdown.rstudio.com/

Rohmert, W. (1984). Das Belastungs-Beanspruchungs-Konzept. *Zeitschrift Für Arbeitswissenschaft*, *38*(4), 193–200.

RStudio. (2020). *RStudio, Open source & professional software for data science teams*. https://rstudio.com/

Shiny, R. (2020). *R Shiny*. https://shiny.rstudio.com/

The R Foundation. (2020). *R: What is R?* https://www.r-project.org/about.html

Tillmann, R., Boris, W., Oliver, L., Ursina, K., Valérie-Anne, R., Marieke, V., Erika, A., Lebert, F., Monsch, G.-A., Dasoki, N., & Klaas, H. S. (2020). *Swiss Household Panel, FORSbase*. https://forsbase.unil.ch/project/study-public-overview/16970/1/

Wolf, C., & Best, H. (2011). Wolf, Christof, und Henning Best (Hrsg.). Handbuch der sozialwissenschaftlichen Datenanalyse. In *Politische Vierteljahresschrift* (Vol. 52, Issue 3). https://doi.org/10.5771/0032-3470-2011-3-571

Anhang

Anhang A: Operationalisierung verschiedener Variablen

Transformation von numerischen Variablen:

Neue numerische Variable	Transformation aus bestehenden Variablen
arbeit_zeit_ueberstunden =	arbeit_zeit_wochenstunden - arbeit_zeit_wochenstunden_vereinbart
Haushaltsaequivalenzeinkommen =	log(haushaltsaequivalenzeinkommen)

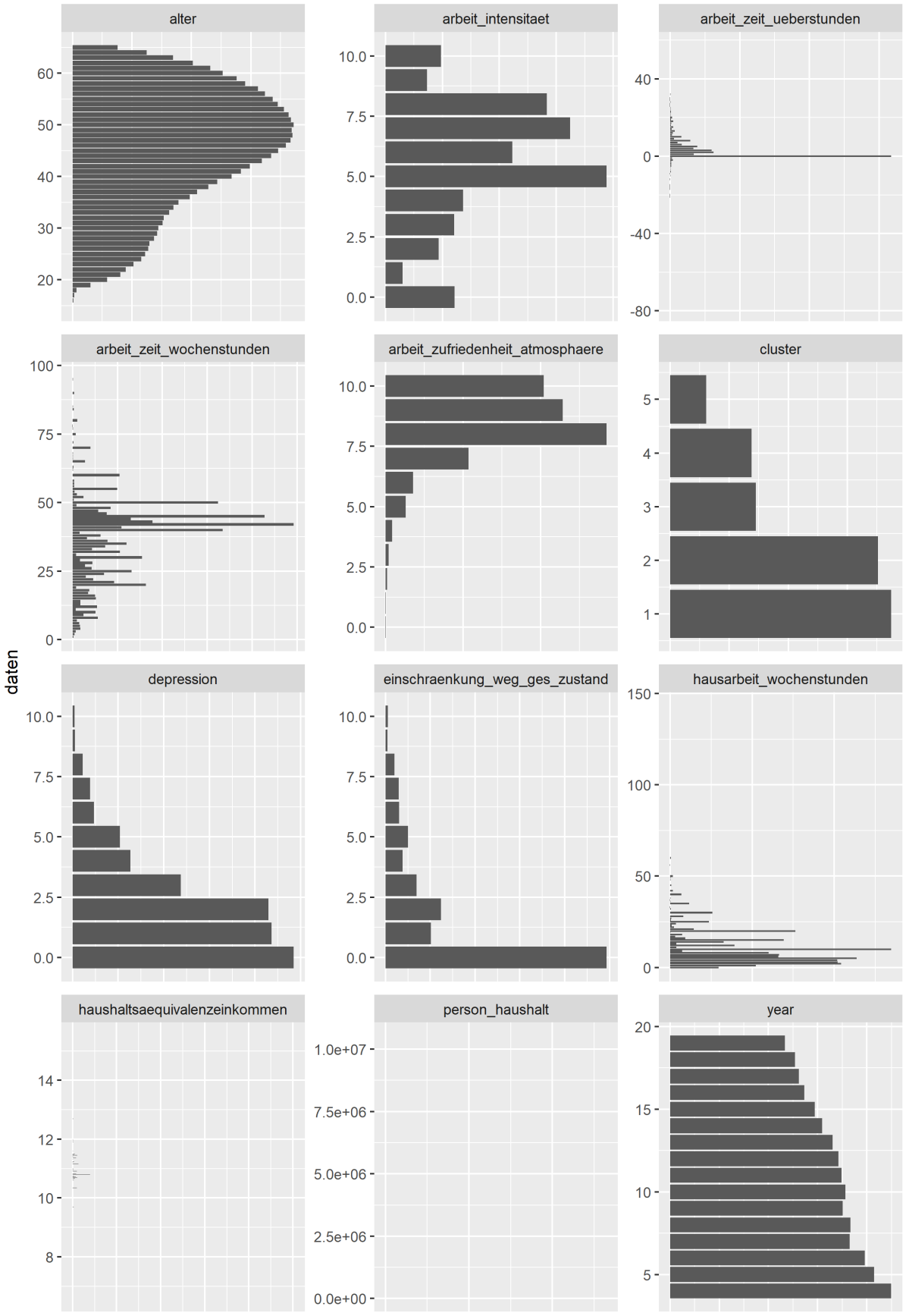
Recodierung von Faktor-Variablen:

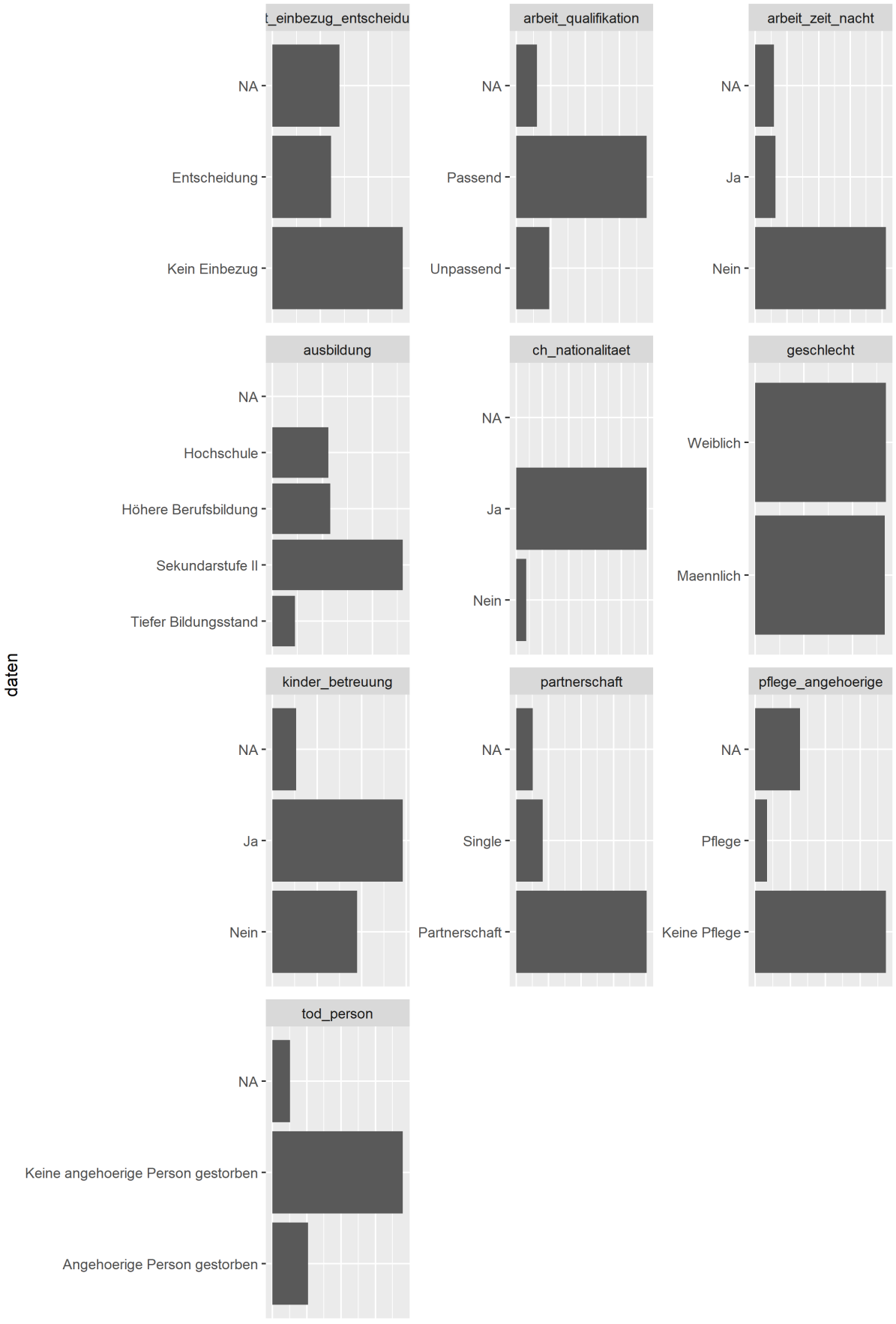
Faktor-Variable	Faktorlevel alt	Faktorlevel neu
ch_nationalitaet	drei mögliche Nationalitäten: NAT_1, NAT_2, NAT_3	Ja, wenn eine der drei möglichen Nationalitäten = "Schweiz" ist. Nein, sonst.
kinder_betreuung basierend auf SHP-Variable "haushaltstyp_mikrozensus"	• 11 "Nicht verheiratetes Paar mit Kindern" • 12 "Ehepaar mit Kindern" • 13 "Nicht mehr verheiratetes Ehepaar mit Kindern" • 21 "Nicht verheiratetes Paar ohne Kinder" • 22 "Ehepaar ohne Kinder" • 23 "Nicht mehr verheiratetes Paar ohne Kinder" • 31 "Alleinerziehender, nie verheiratet gewesener Elternteil mit Kindern" • 32 "Verheirateter Elternteil mit Kindern" • 33 "Nicht mehr verheirateter Elternteil mit Kindern" • 41 "Nie verheiratet gewesene, allein lebende Person" • 42 "Verheiratete allein lebende Person" • 43 "Nicht mehr verheiratete, allein lebende Person" • 44 "Andere Situation"	Ja, wenn Faktorlevel ∈ {11,12,13,31,32,33} Nein, wenn Faktorlevel ∈ {21,22,23,41,42,43,44}
pflege_angehoerige	• pflege_extern: Ja & Nein • pflege_wer_1: Personen-ID o. NA • pflege_wer_2: Personen-ID o. NA • pflege_wer_3: Personen-ID o. NA • pflege_wer_4: Personen-ID o. NA • pflege_wer_5: Personen-ID o. NA	Ja, wenn pflege_extern == Ja & die pflegende Person tatsächlich die externen pflegerischen Tätigkeiten übernimmt (pflege_wer_1 == id | pflege_wer_2 == id | pflege_wer_3 == id | pflege_wer_3 == id | pflege_wer_3 == id) Nein, sonst.
partnerschaft	• ja, zusammenlebend • ja, aber leben nicht zusammen • nein	Nein, wenn nein Ja, sonst
ausbildung	• 0 "keine abgeschlossene obligatorische Ausbildung" • 1 "obligatorische Ausbildung, Anlehre" • 2 "haushaltslehrjahr, 1 Jahr handelsschule" • 3 "allgemeinbildende Schule") • 4 "berufslehre (EFZ)" • 5 "vollzeitberufsschule " • 6 "bachelor/matura" • 7 "berufsprüfung mit Meisterdiplom, Eidgenössischer Fachausweis" • 8 "techniker- oder fachschule" • 9 "höhere Fachschule HTL etc." • 10 "universität, universitäre Hochschulen, PH, FH")	Tiefer Bildungsstand, wenn Faktorlevel ∈ {0,1,2,3} Sekundarstufe II, wenn Faktorlevel ∈ {4,5,6} Höhere Berufsbildung, wenn Faktorlevel ∈ {7,8,9} Hochschule, wenn Faktorlevel ∈ {10}
arbeit_einbezug_entscheidungen	• Ja, Entscheide • Ja, Meinung • Nein	Entscheidung, wenn Faktorlevel = Ja, Entscheide Kein Einbezug, sonst
arbeit_qualifikation	• 1 "Ihre Qualifikation reicht nicht für Ihre Arbeit" • 2 Ihre Qualifikation entspricht den Anforderungen von Ihrer Arbeit" • 3 "Sie sind überqualifiziert" • 4 "Ihre Qualifikation hat nichts mit Ihrer Arbeit zu tun"	Unpassend, wenn Faktorlevel ∈ {1,3,4} Passend, wenn Faktorlevel ∈ {2}

Anhang B: Anteil ungültiger Messungen pro Variable

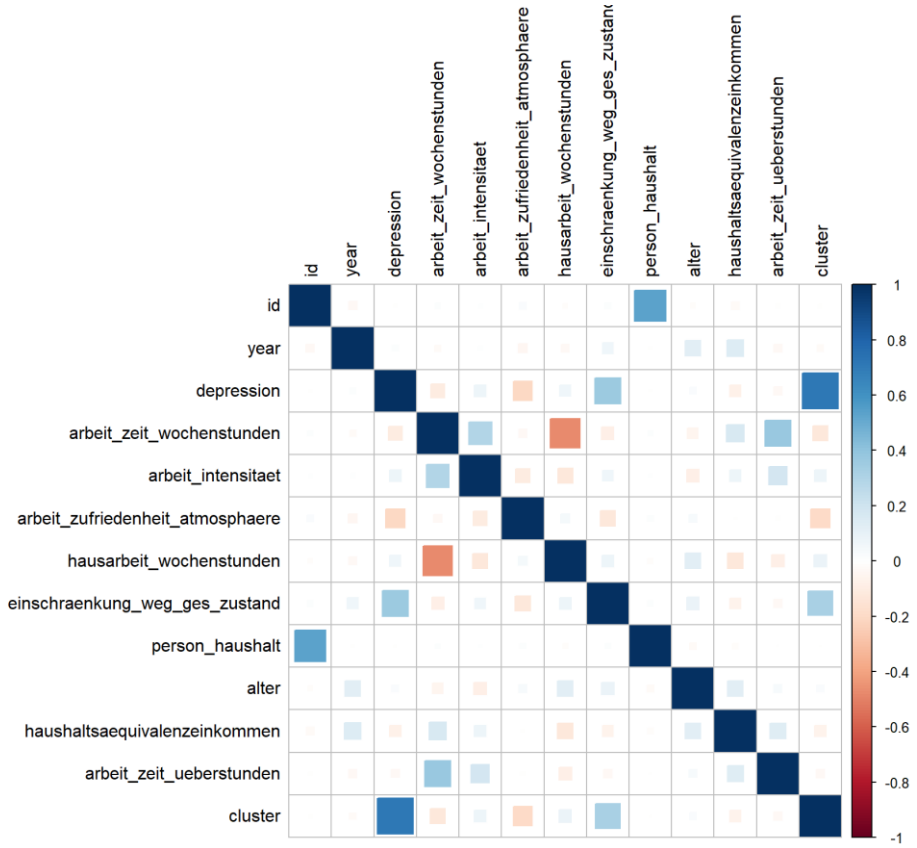
Variable	Absolute Anzahl NA's	Relativer Anteil NA's [%]
id	0	0.0
year	0	0.0
depression	5081	9.5
arbeit_zeit_wochenstunden	9273	17.3
arbeit_zeit_nacht	5939	11.1
arbeit_qualifikation	6069	11.3
arbeit_intensitaet	6165	11.5
arbeit_einbezug_entscheidungen	13990	26.2
arbeit_zufriedenheit_atmosphaere	8322	15.6
hausarbeit_wochenstunden	5863	11.0
einschraenkung_weg_ges_zustand	5090	9.5
ausbildung	6	0.0
partnerschaft	5099	9.5
tod_person	5067	9.5
person_haushalt	0	0.0
geschlecht	0	0.0
alter	0	0.0
ch_nationalitaet	1	0.0
haushaltsaequivalenzeinkommen	9654	18.1
kinder_betreuung	5302	9.9
arbeit_zeit_ueberstunden	19172	35.9
pflege_angehoerige	12790	23.9
cluster	0	0.0

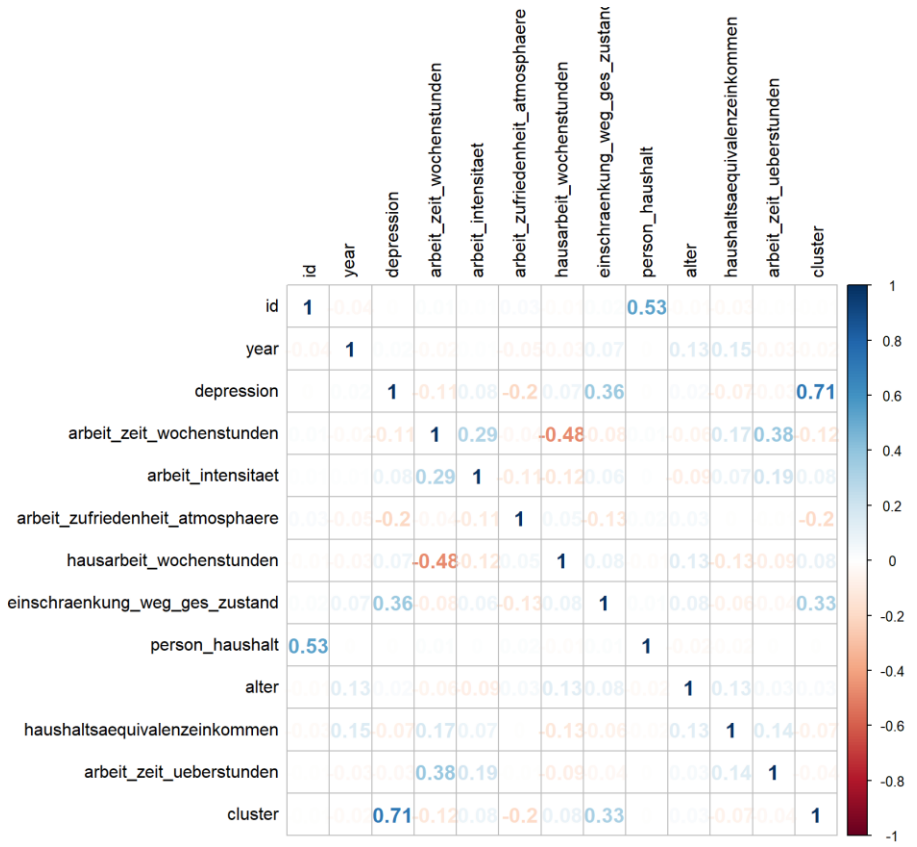
Anhang C: Verteilung von numerischen und Faktorvariablen




t_einbezug_entscheidu
NA
Entscheidung
Kein Einbezug
arbeit_qualifikation
NA
Passend
Unpassend
arbeit_zeit_nacht
NA
Ja
Nein
ausbildung
NA
Hochschule
Höhere Berufsbildung
Sekundarstufe II
Tiefer Bildungsstand
ch_nationalitaet
NA
Ja
Nein
geschlecht
Weiblich
Maennlich
kinder_betreuung
NA
Ja
Nein
partnerschaft
NA
Single
Partnerschaft
pflege_angehoerige
NA
Pflege
Keine Pflege
tod_person
NA
Keine angehoerige Person gestorben
Angehoerige Person gestorben
daten

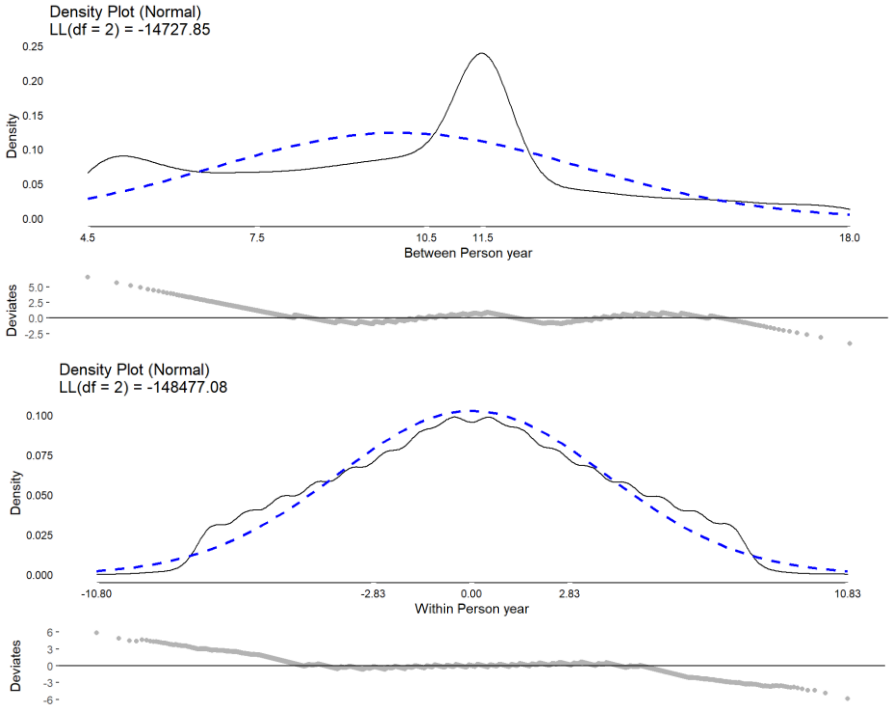
Anhang D: Korrelationsstruktur numerischer unabhängiger Variablen



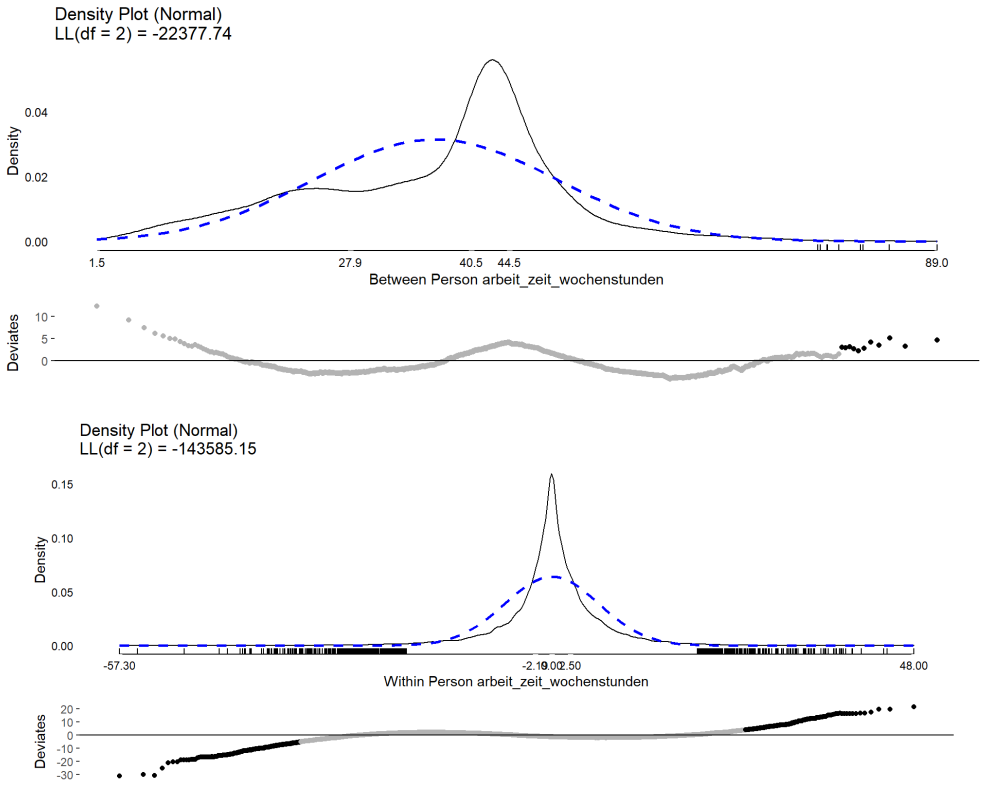


Anhang E: Varianzanalyse unabhängiger numerischer Variablen

Var	*Sigma*	*ICC*
id	4.757111	0.2158887
Residual	17.277899	0.7841113

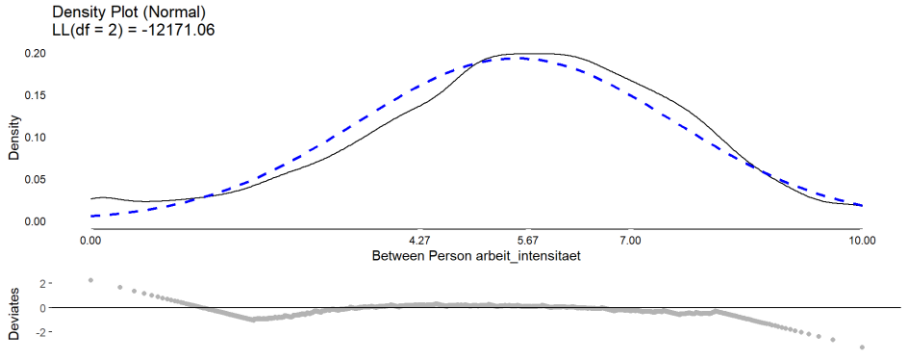


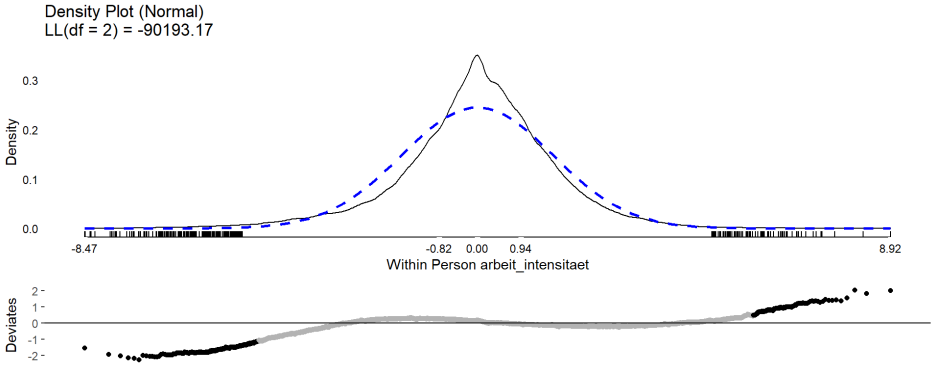
Var	*Sigma*	*ICC*
id	149.00635	0.7698155
Residual	44.55477	0.2301845



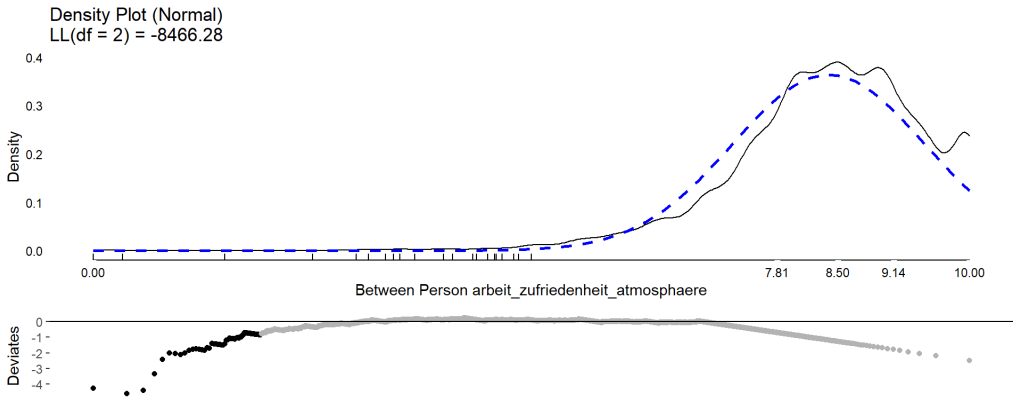
Var	*Sigma*	*ICC*

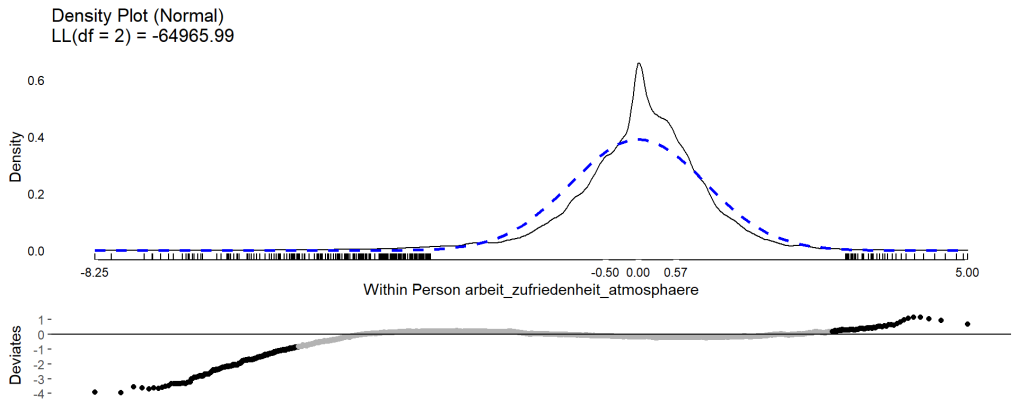
id	3.529122	0.5390107
Residual	3.018284	0.4609893





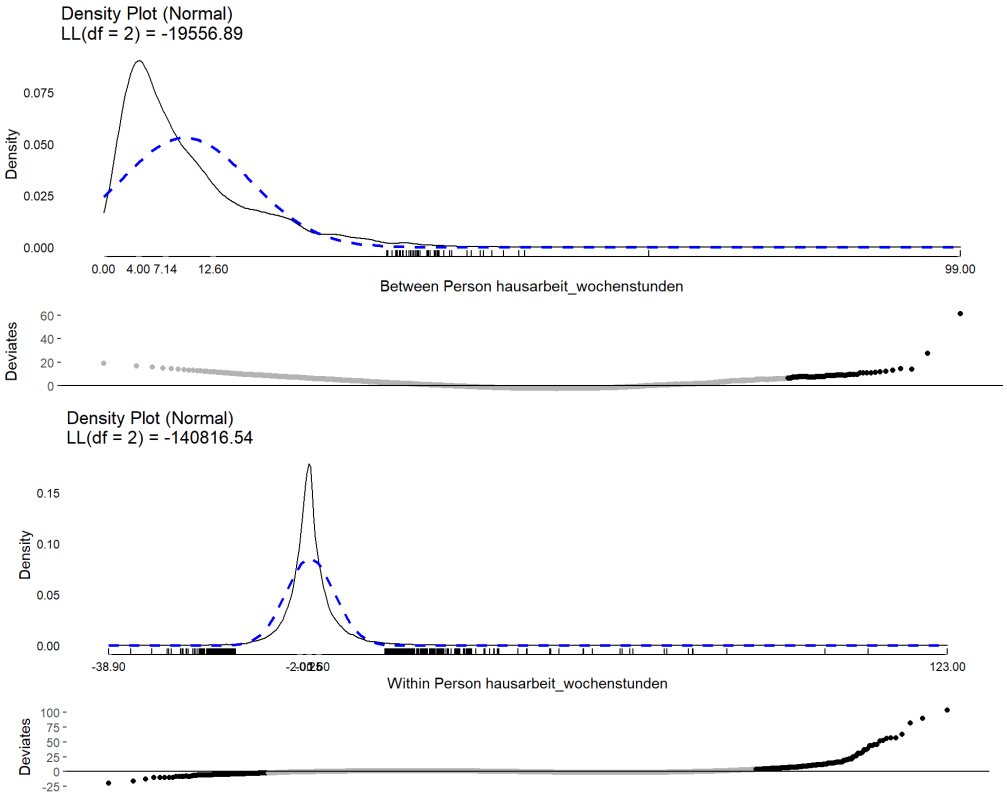
Var	***Sigma***	***ICC***
id	0.8756325	0.423446
Residual	1.1922404	0.576554



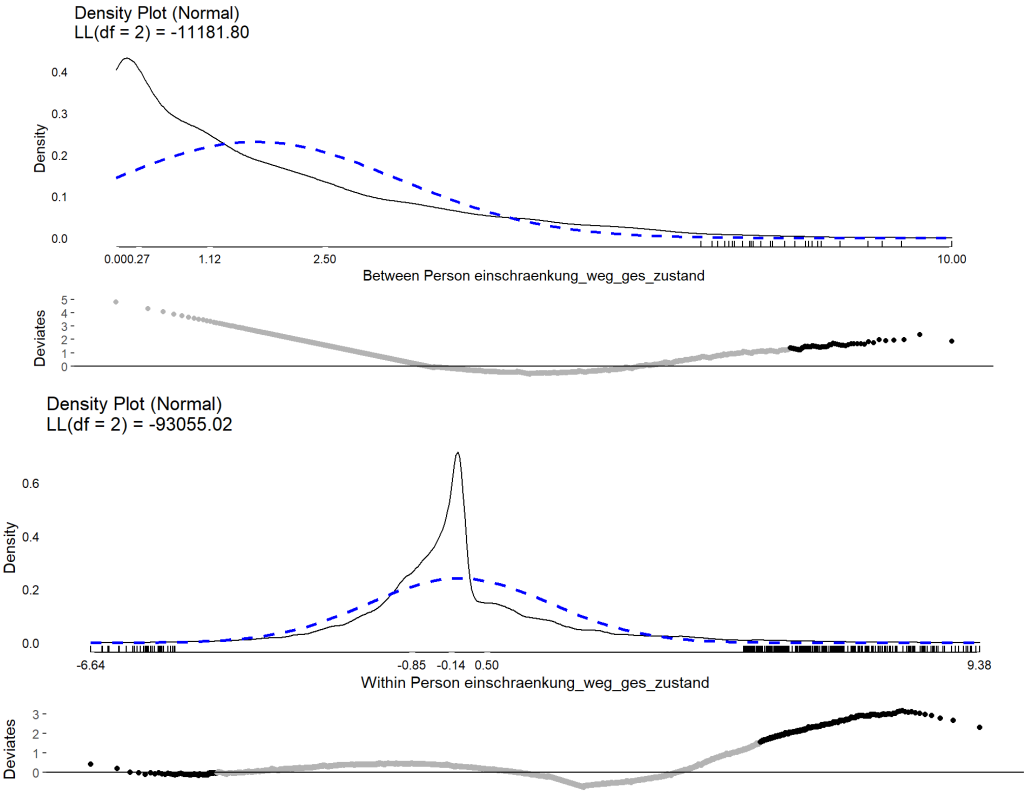


Var	***Sigma***	***ICC***

id	50.57806	0.6720316
Residual	24.68337	0.3279684

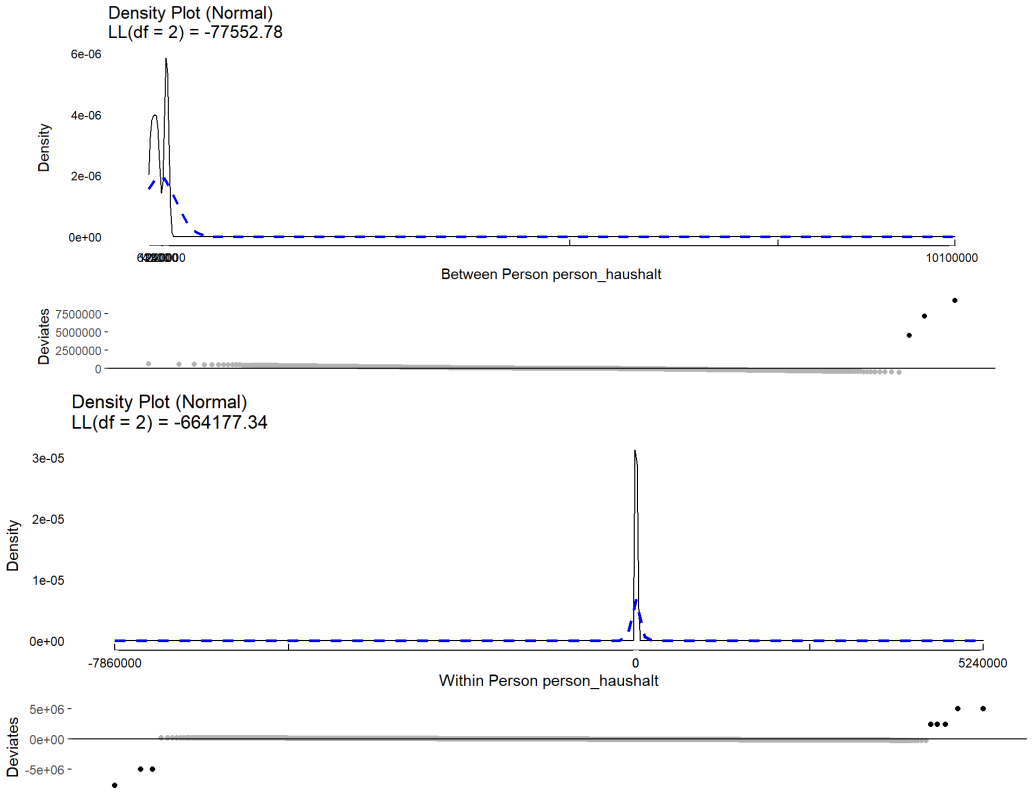


Var	***Sigma***	***ICC***
id	2.317196	0.4267902
Residual	3.112160	0.5732098

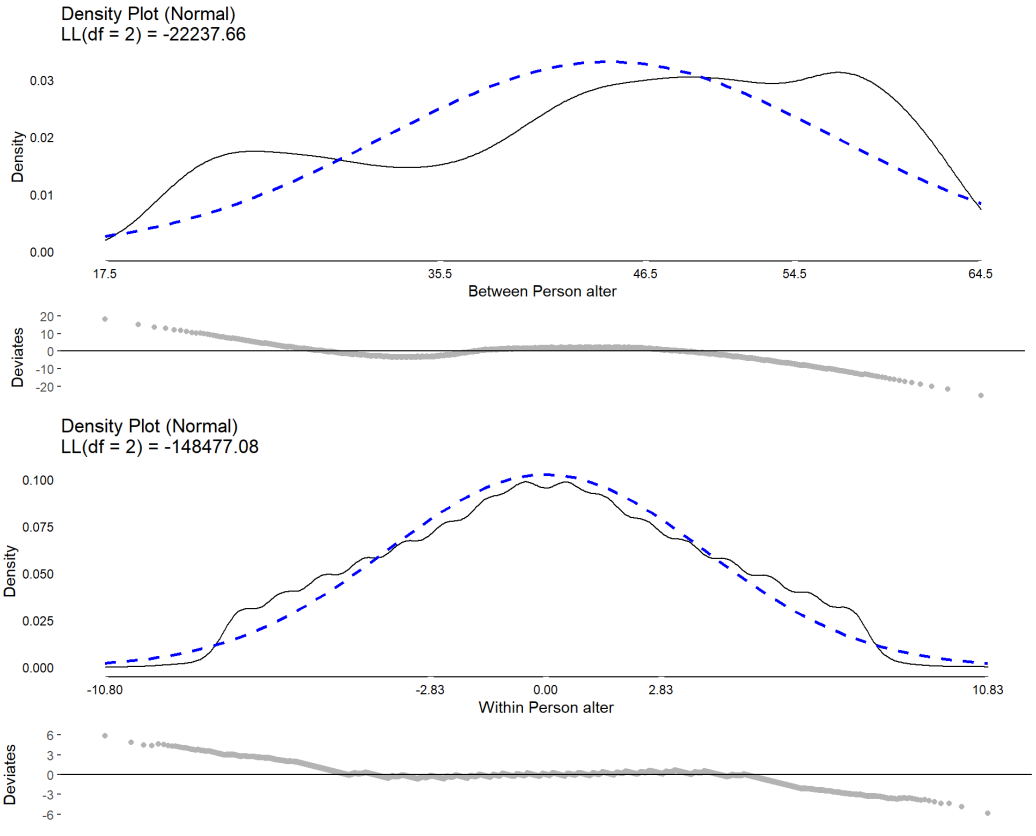


Var	***Sigma***	***ICC***

id	38067279232	0.9042123
Residual	4032654721	0.0957877

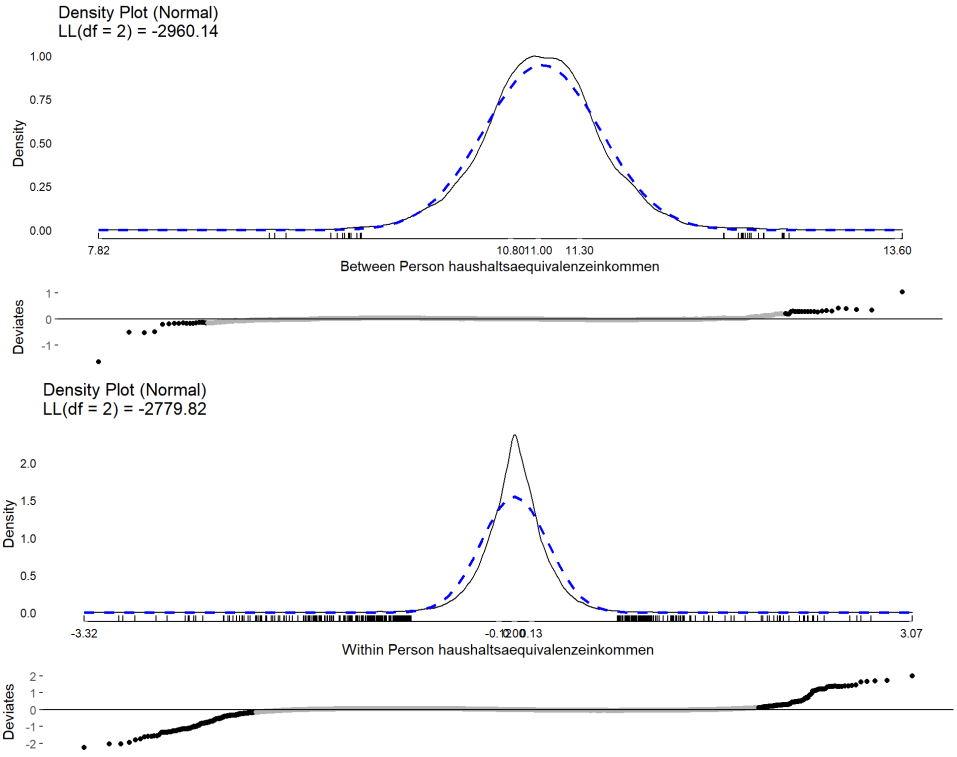


Var	***Sigma***	***ICC***
id	140.22675	0.8923105
Residual	16.92343	0.1076895

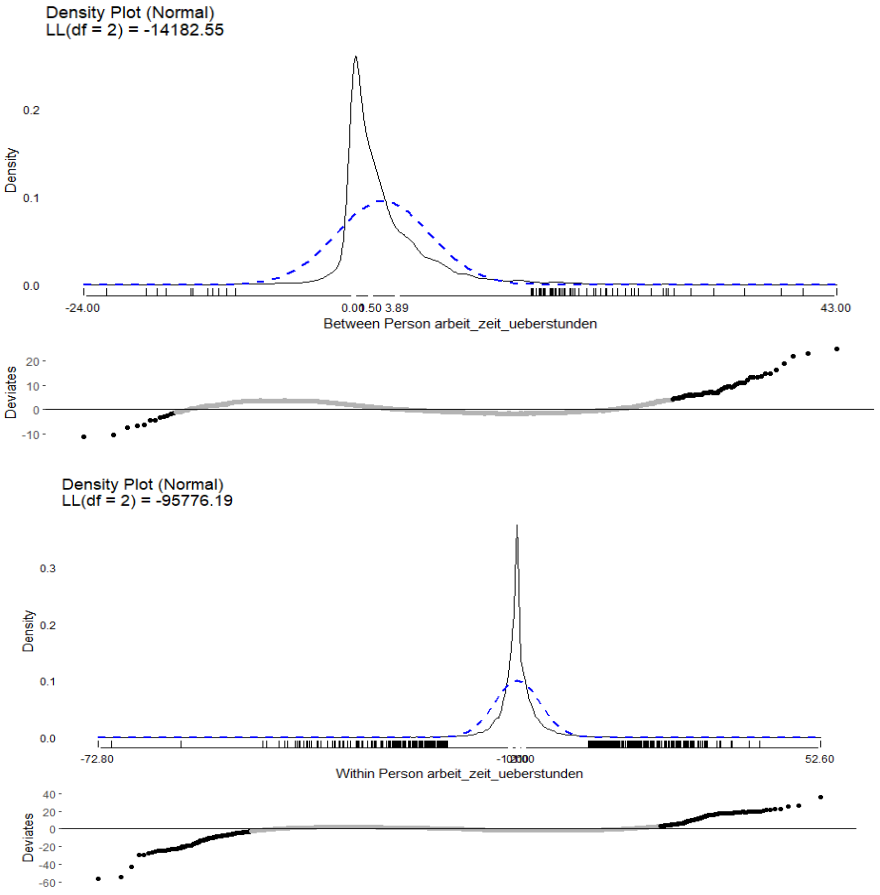


Var	***Sigma***	***ICC***
id	0.1579364	0.6758925

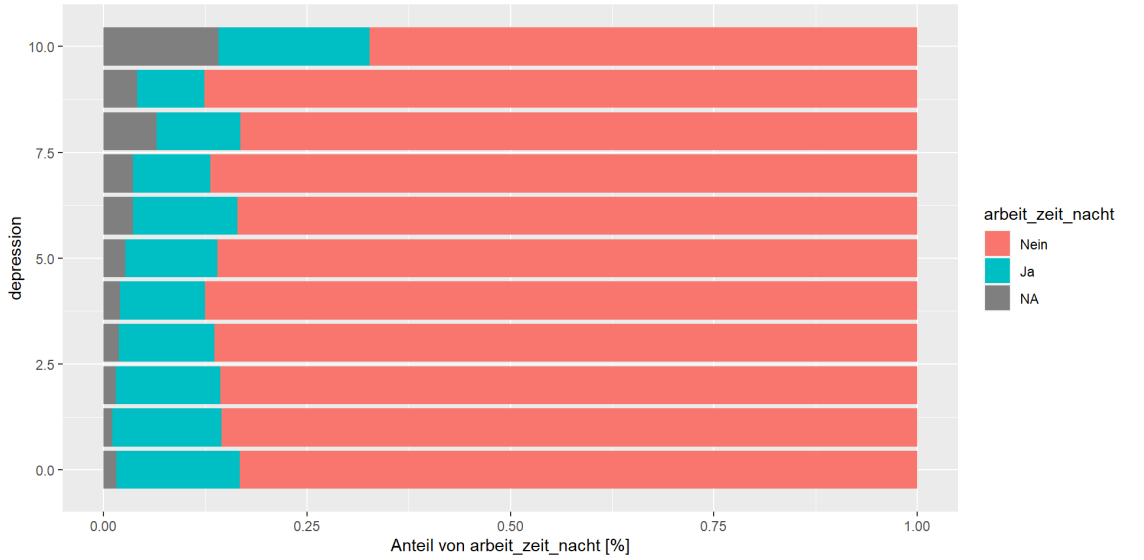
Residual	0.0757345	0.3241075

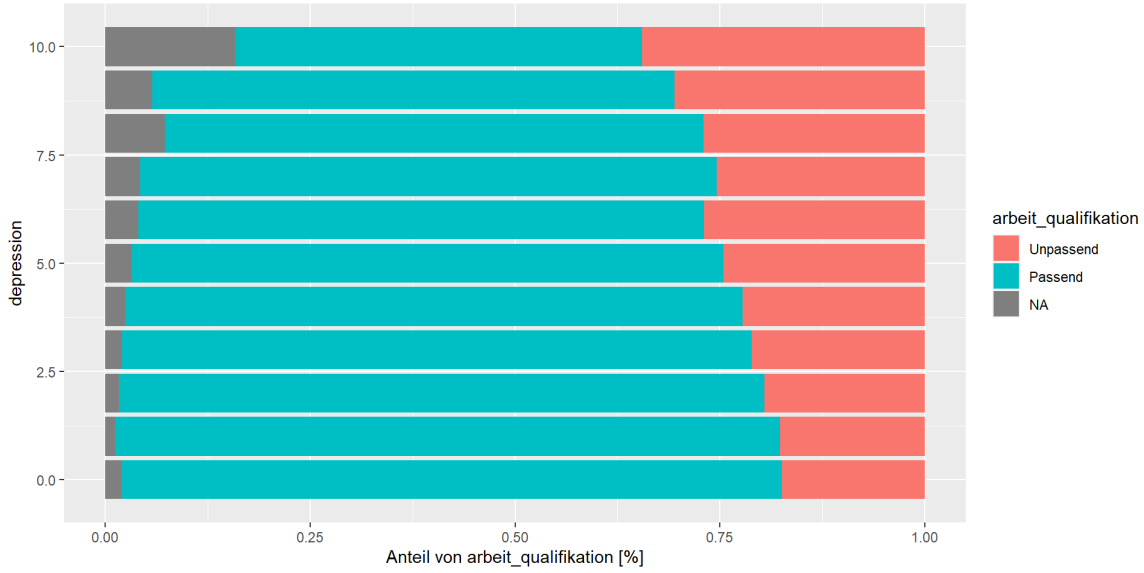


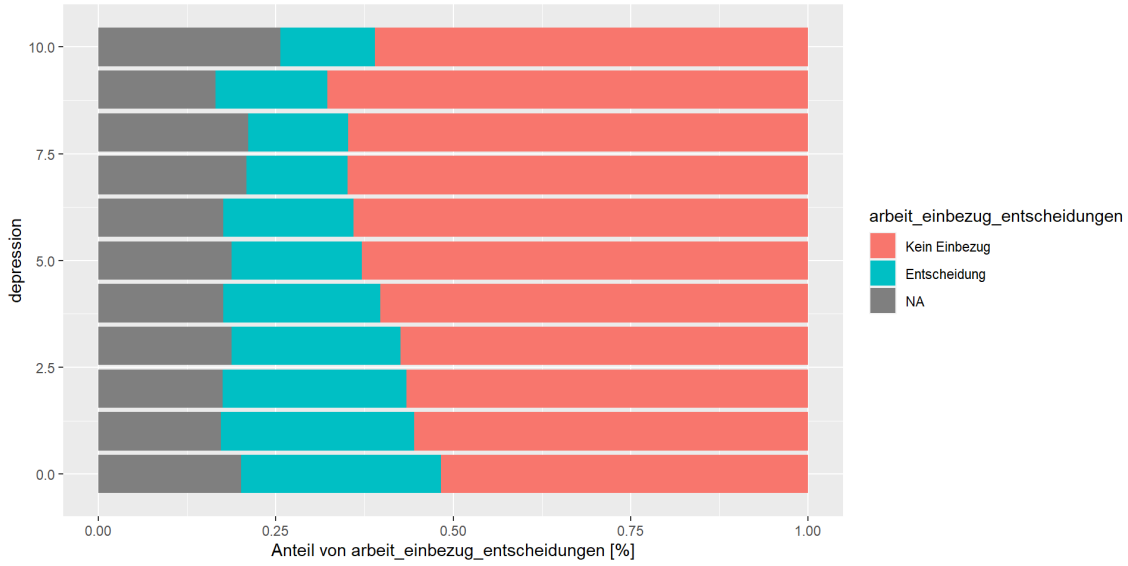
Var	*Sigma*	*ICC*
id	11.34746	0.3824499
Residual	18.32298	0.6175501

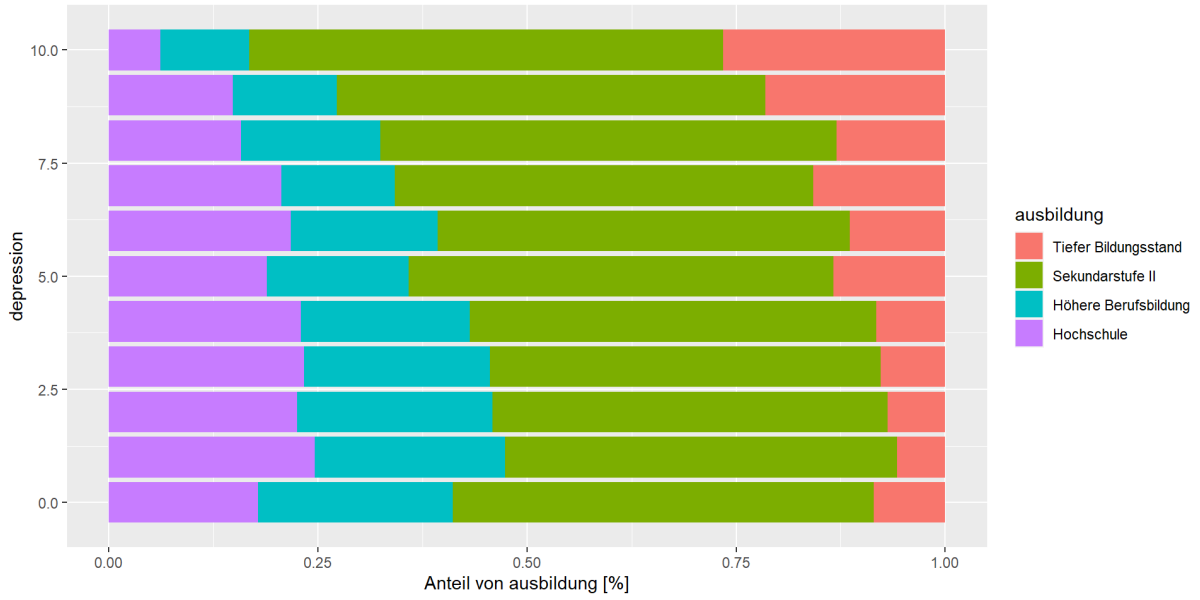


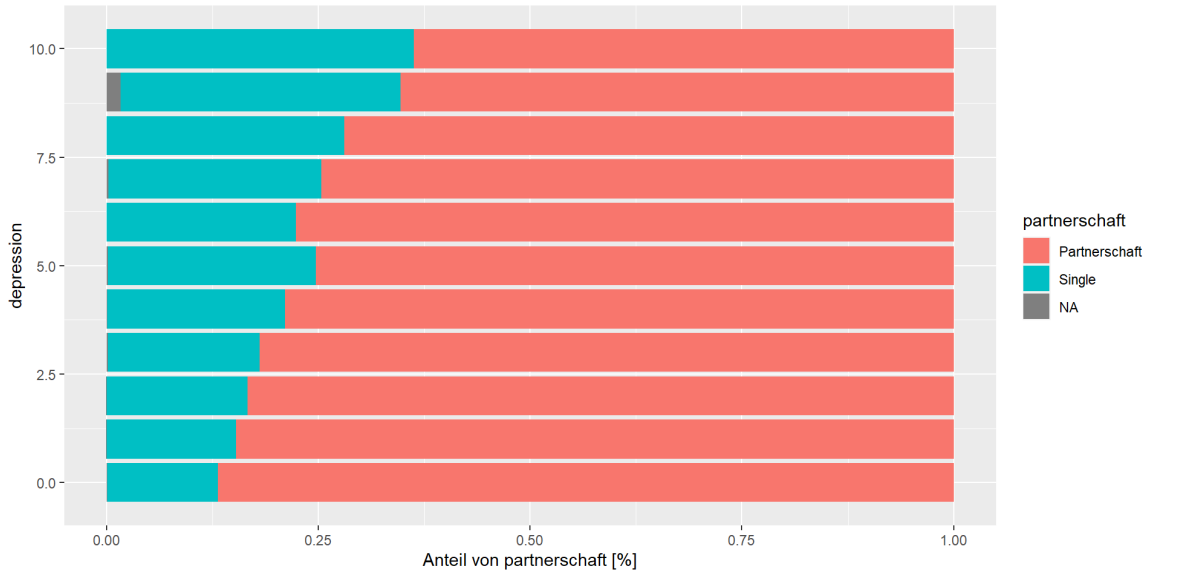
Anhang F: Bedingte Verteilungen

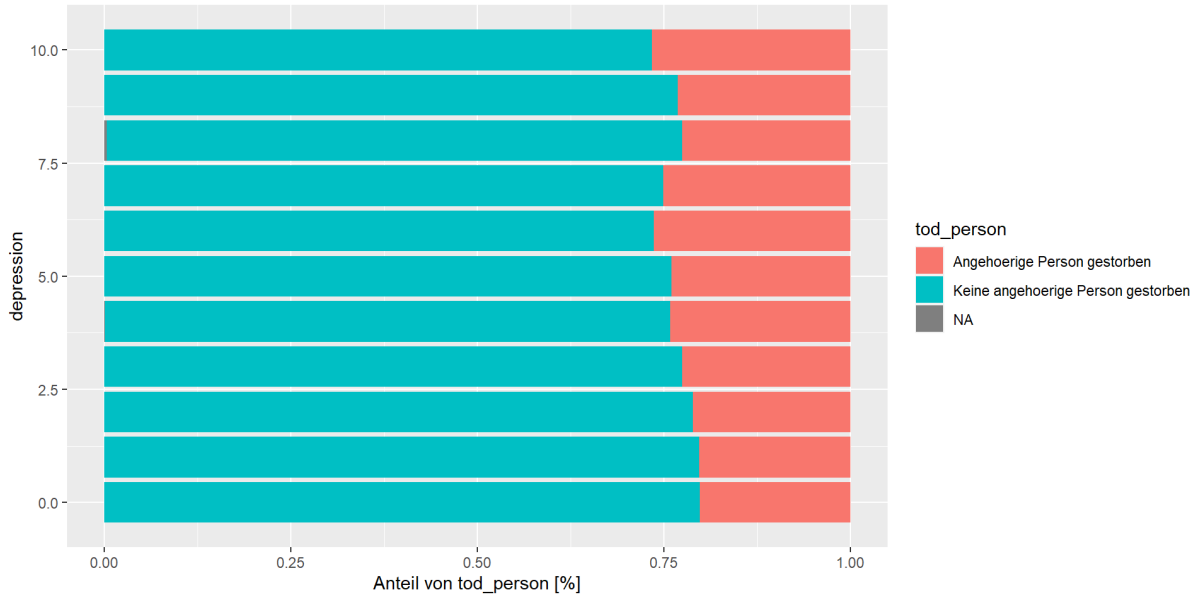


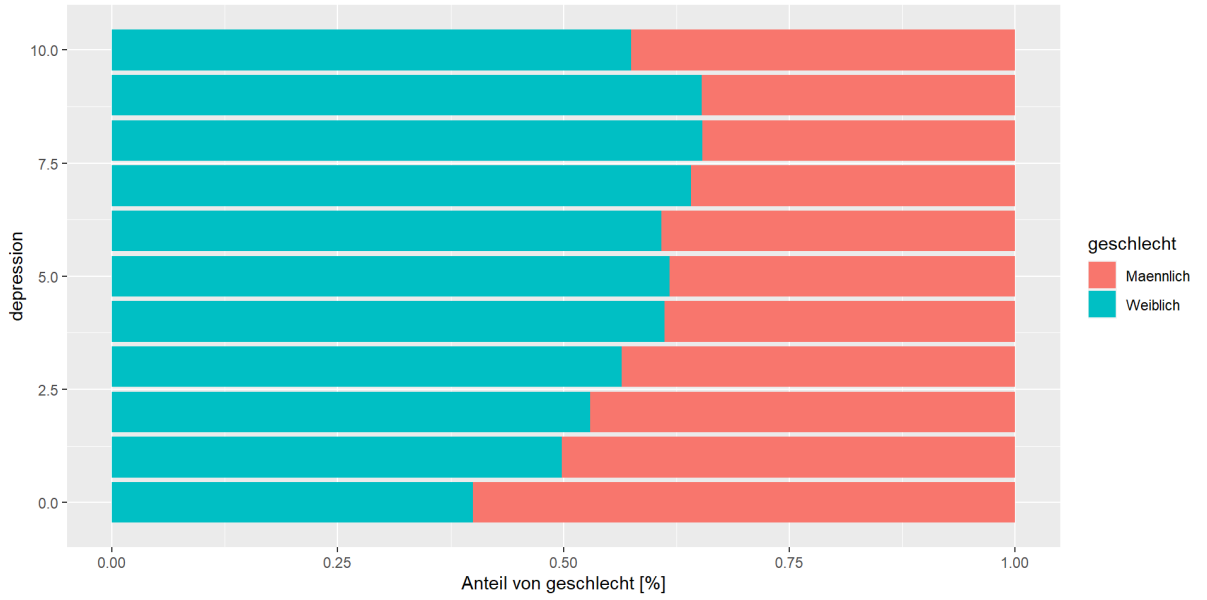


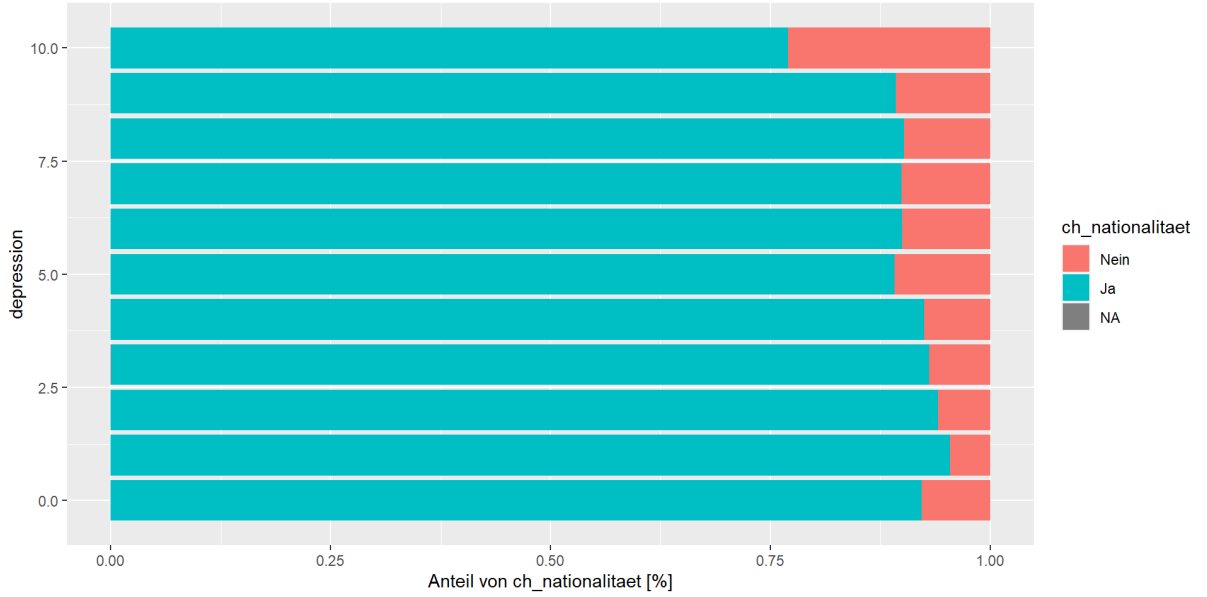


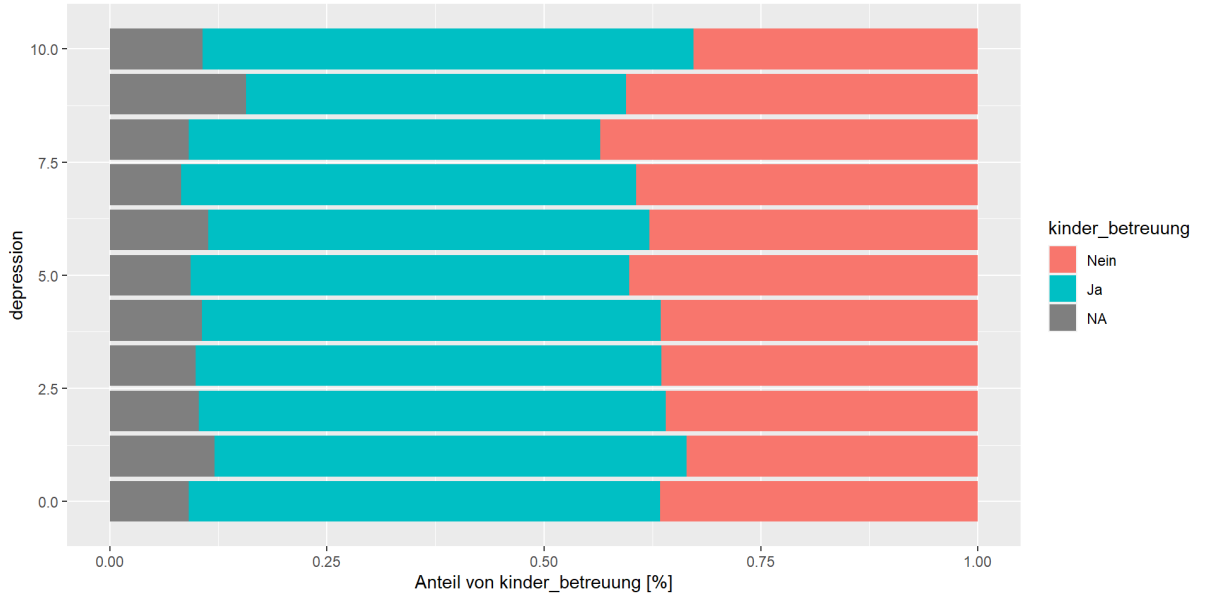

depression
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
Anteil von ausbildung [%]
ausbildung
Tiefer Bildungsstand
Sekundarstufe II
Höhere Berufsbildung
Hochschule

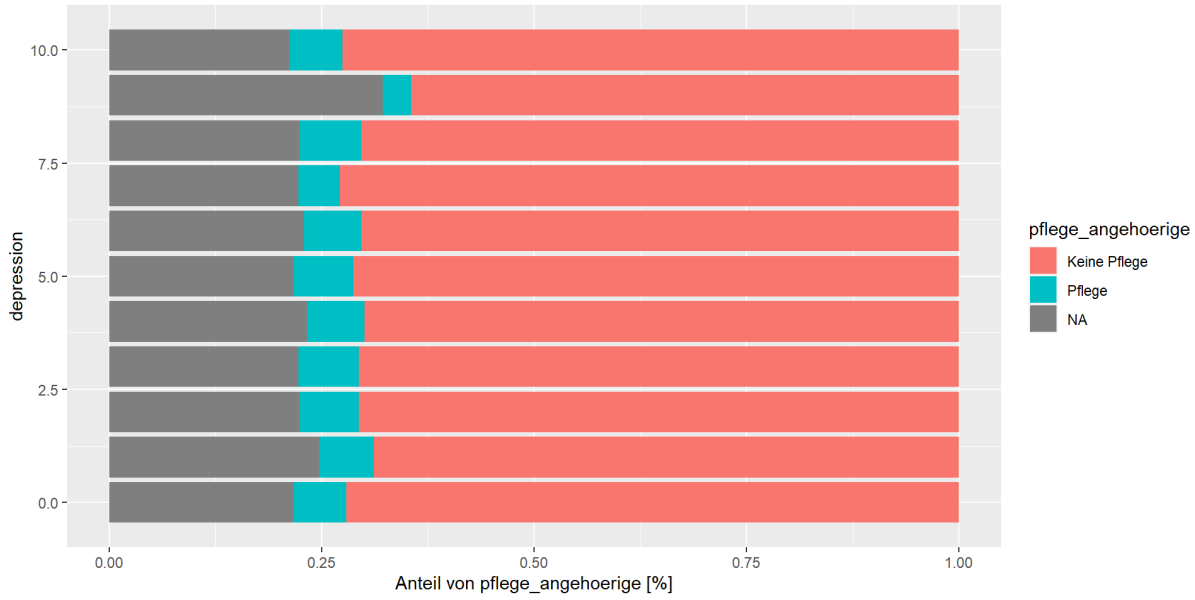

depression
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
Anteil von partnerschaft [%]
partnerschaft
Partnerschaft
Single
NA

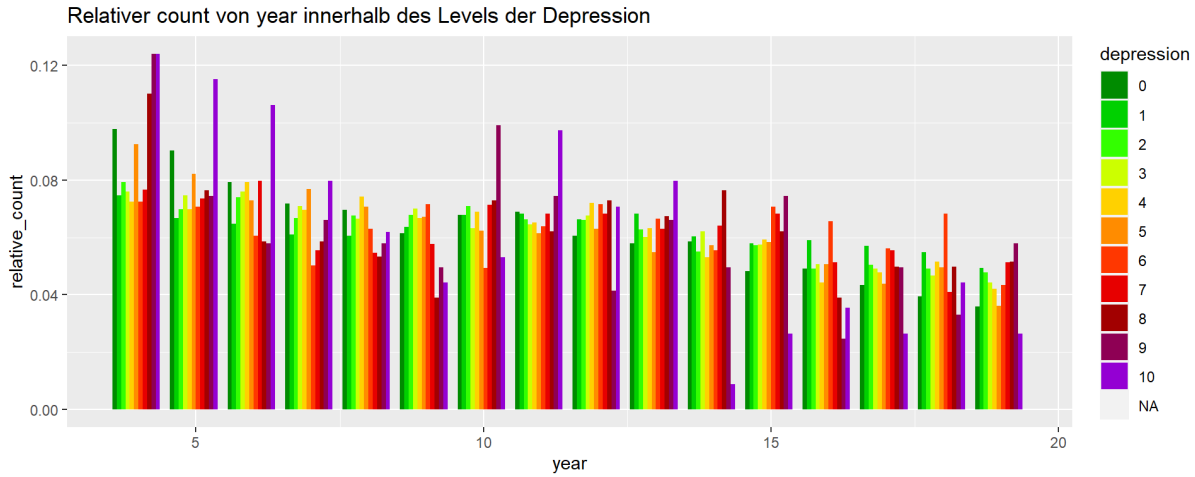

depression
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
Anteil von tod_person [%]
tod_person
Angehoerige Person gestorben
Keine angehoerige Person gestorben
NA

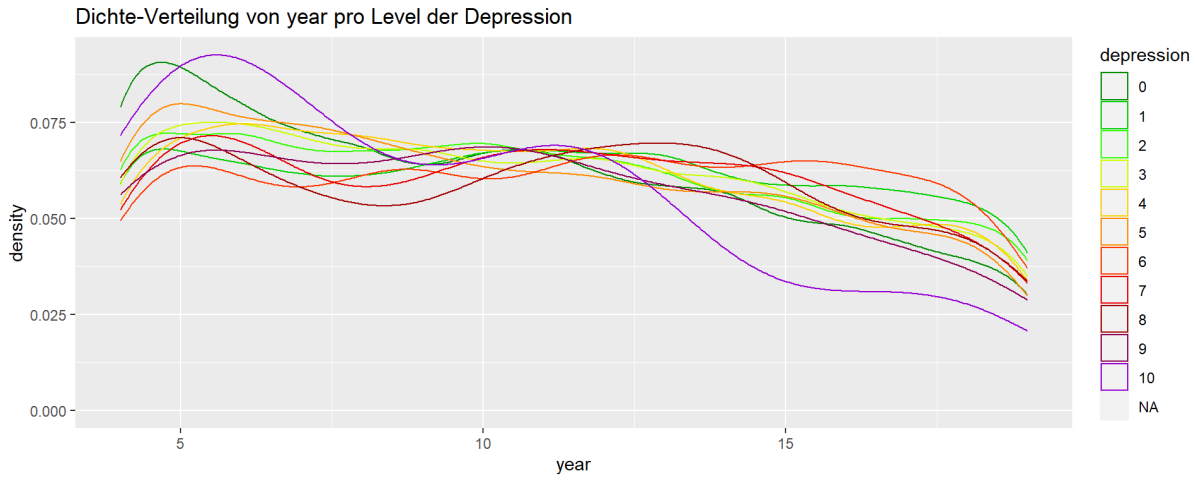

depression
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
Anteil von geschlecht [%]
geschlecht
Maennlich
Weiblich

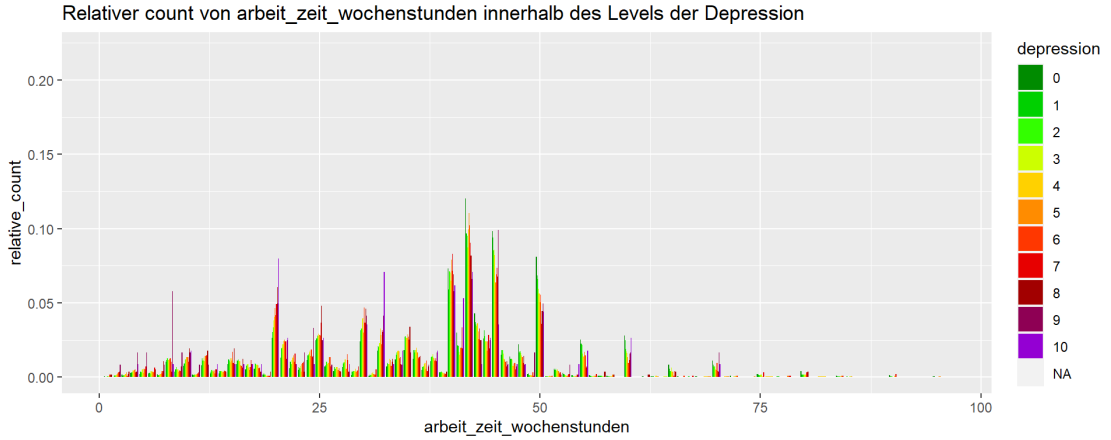

depression
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
Anteil von ch_nationalitaet [%]
ch_nationalitaet
Nein
Ja
NA

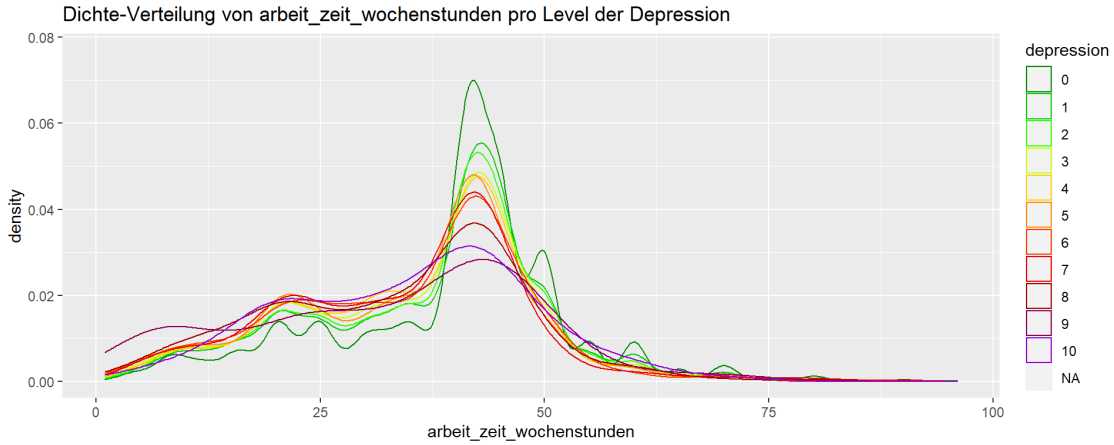

depression
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
Anteil von kinder_betreuung [%]
kinder_betreuung
Nein
Ja
NA

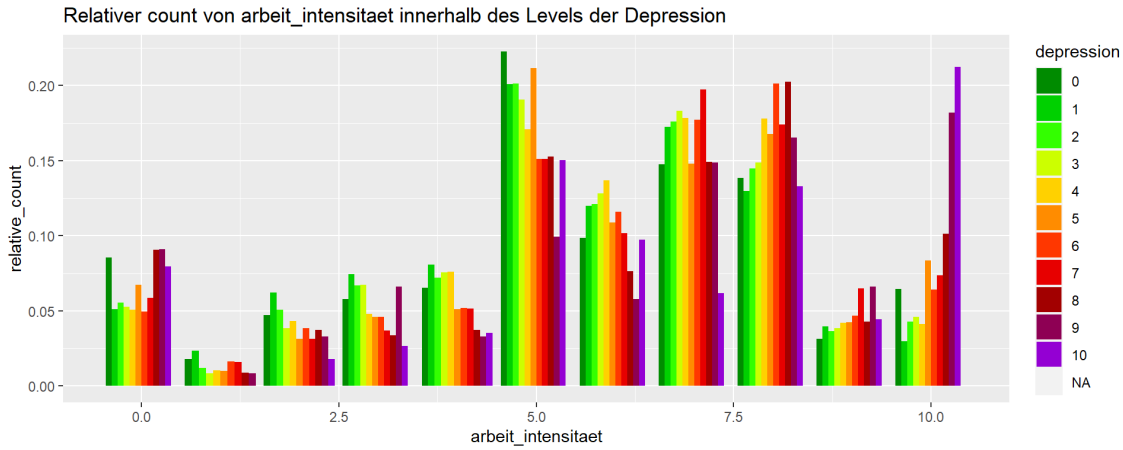

depression
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
Anteil von pflege_angehoerige [%]
pflege_angehoerige
Keine Pflege
Pflege
NA

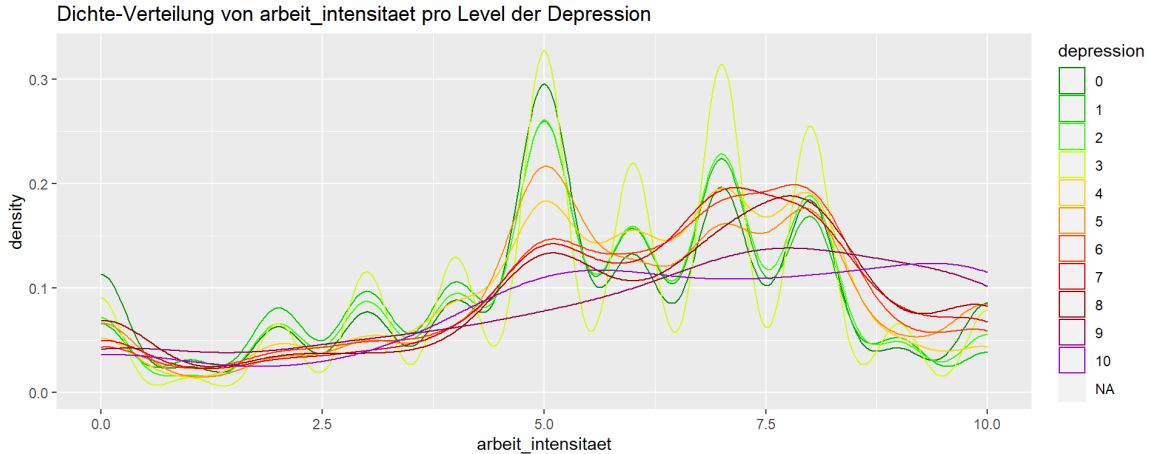

Relativer count von year innerhalb des Levels der Depression
relative_count
0.12
0.08
0.04
0.00
5
10
15
20
year
depression
0
1
2
3
4
5
6
7
8
9
10
NA

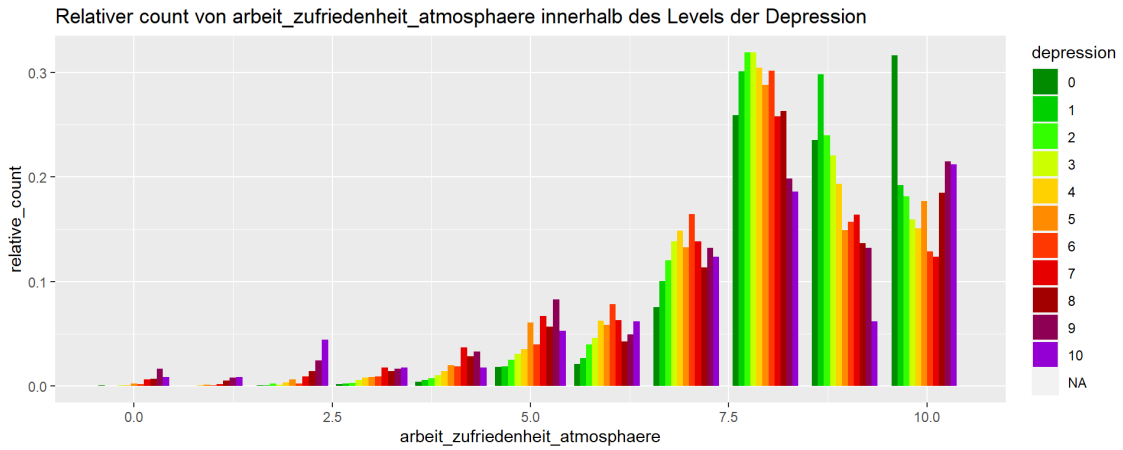

Dichte-Verteilung von year pro Level der Depression
density
0.075
0.050
0.025
0.000
5
10
15
year
depression
0
1
2
3
4
5
6
7
8
9
10
NA

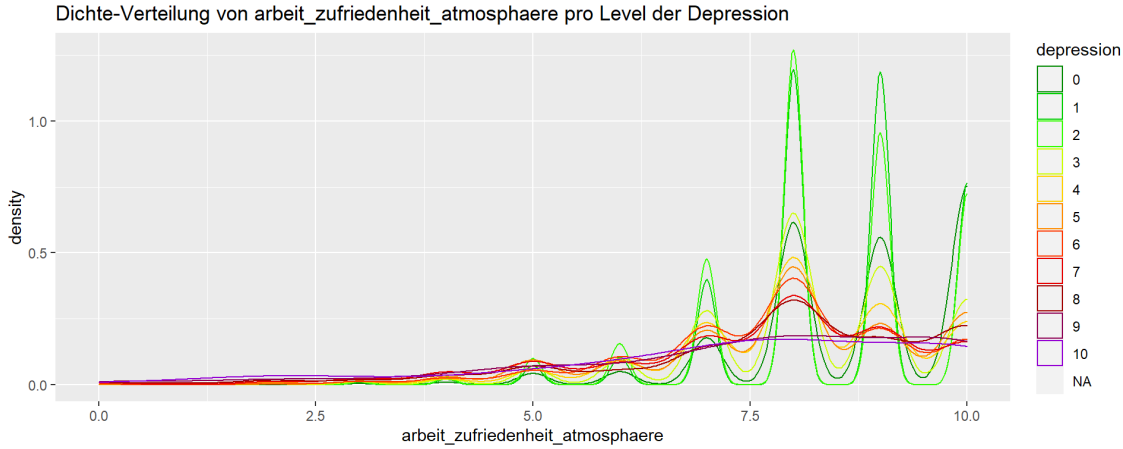

Relativer count von arbeit_zeit_wochenstunden innerhalb des Levels der Depression
relative_count
0.20
0.15
0.10
0.05
0.00
0
25
50
75
100
arbeit_zeit_wochenstunden
depression
0
1
2
3
4
5
6
7
8
9
10
NA

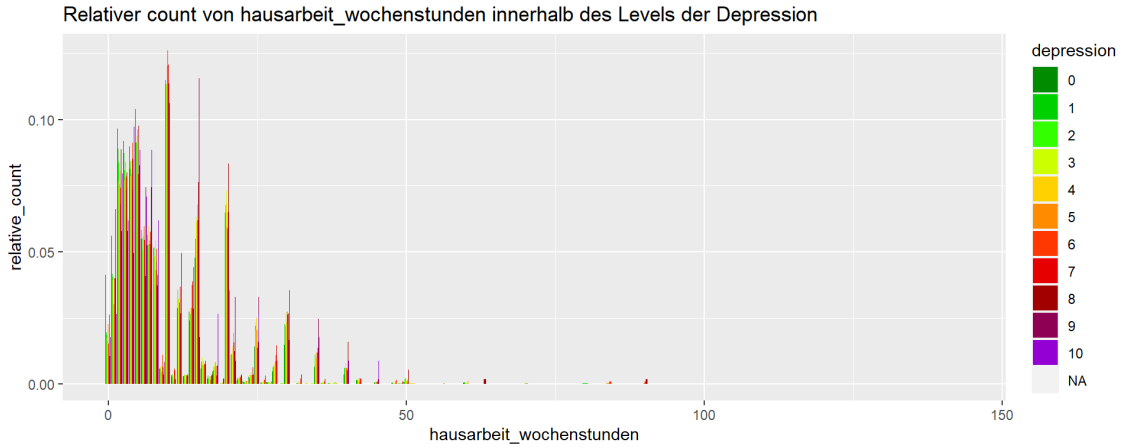

Dichte-Verteilung von arbeit_zeit_wochenstunden pro Level der Depression
density
0.08
0.06
0.04
0.02
0.00
0
25
50
75
100
arbeit_zeit_wochenstunden
depression
0
1
2
3
4
5
6
7
8
9
10
NA

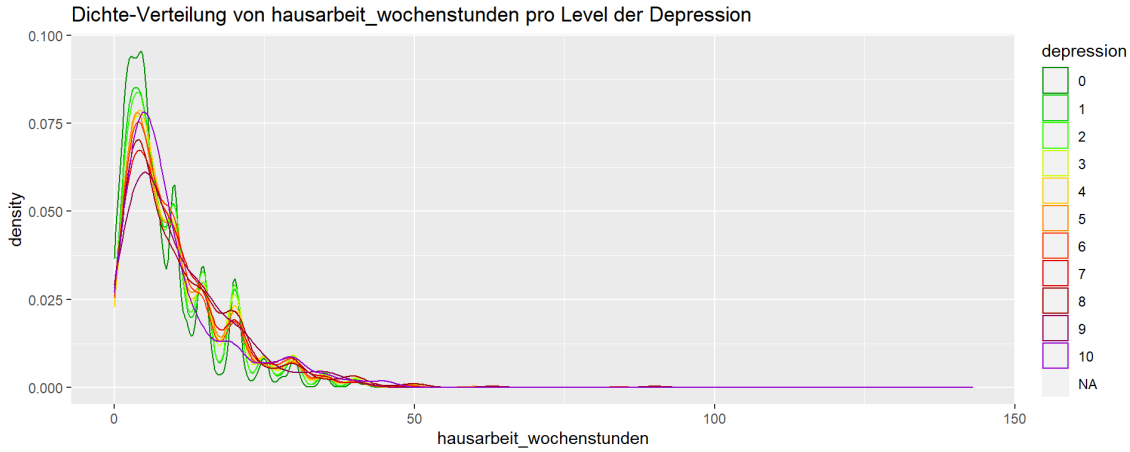

Relativer count von arbeit_intensitaet innerhalb des Levels der Depression
relative_count
0.20
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
arbeit_intensitaet
depression
0
1
2
3
4
5
6
7
8
9
10
NA

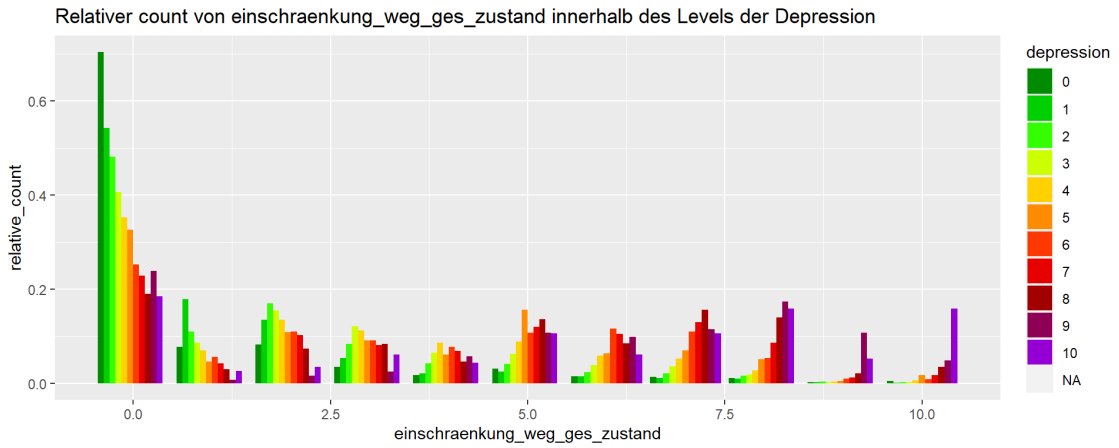

Dichte-Verteilung von arbeit_intensitaet pro Level der Depression
density
0.3
0.2
0.1
0.0
0.0
2.5
5.0
7.5
10.0
arbeit_intensitaet
depression
0
1
2
3
4
5
6
7
8
9
10
NA

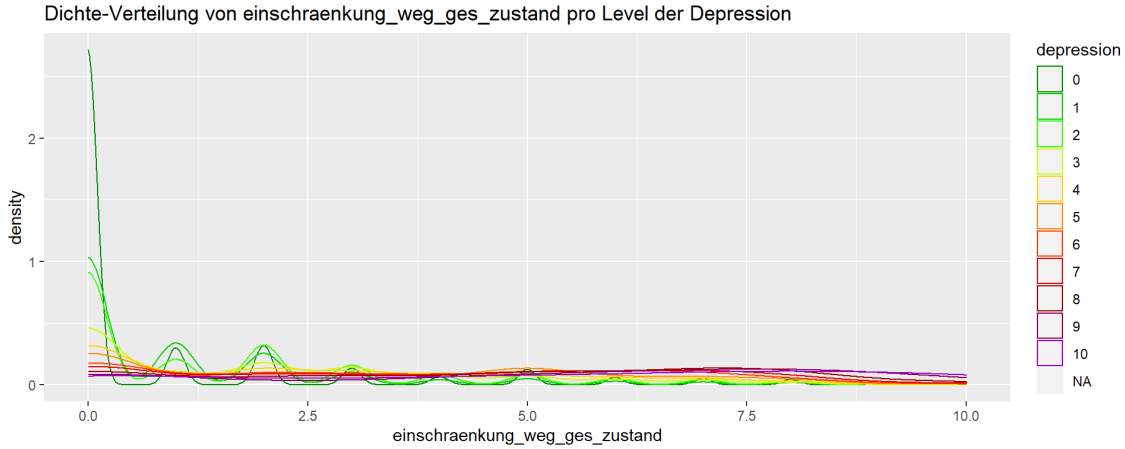

Relativer count von arbeit_zufriedenheit_atmosphaere innerhalb des Levels der Depression
relative_count
0.0
0.1
0.2
0.3
0.0
2.5
5.0
7.5
10.0
arbeit_zufriedenheit_atmosphaere
depression
0
1
2
3
4
5
6
7
8
9
10
NA

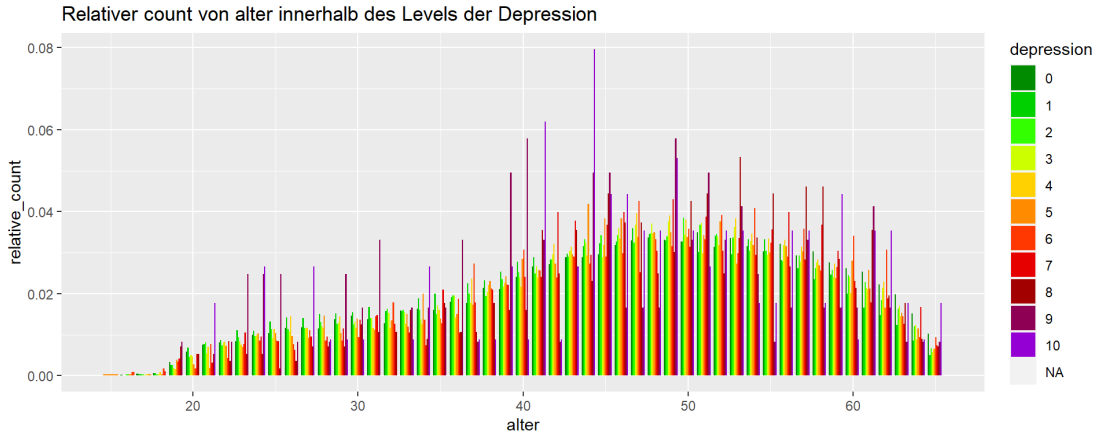

Dichte-Verteilung von arbeit_zufriedenheit_atmosphaere pro Level der Depression
density
0.0
0.5
1.0
0.0
2.5
5.0
7.5
10.0
arbeit_zufriedenheit_atmosphaere
depression
0
1
2
3
4
5
6
7
8
9
10
NA

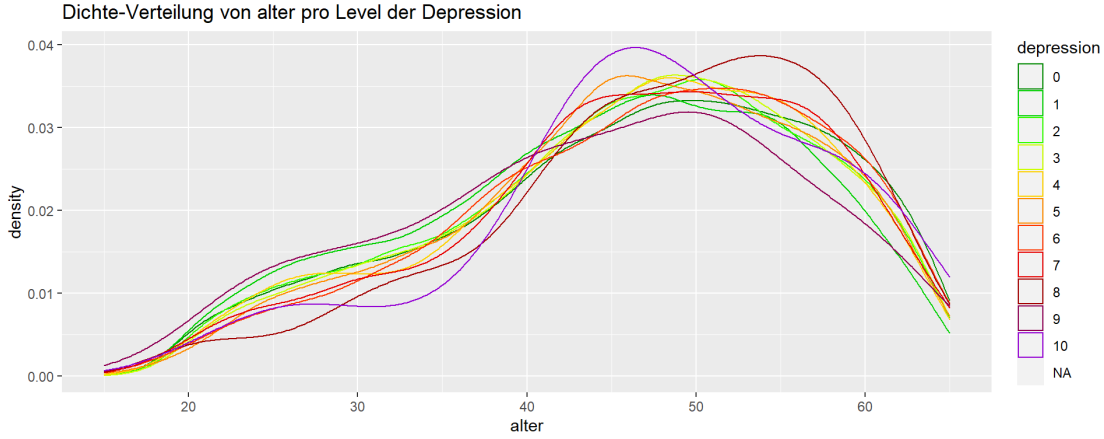

Relativer count von hausarbeit_wochenstunden innerhalb des Levels der Depression
relative_count
0.00
0.05
0.10
0
50
100
150
hausarbeit_wochenstunden
depression
0
1
2
3
4
5
6
7
8
9
10
NA

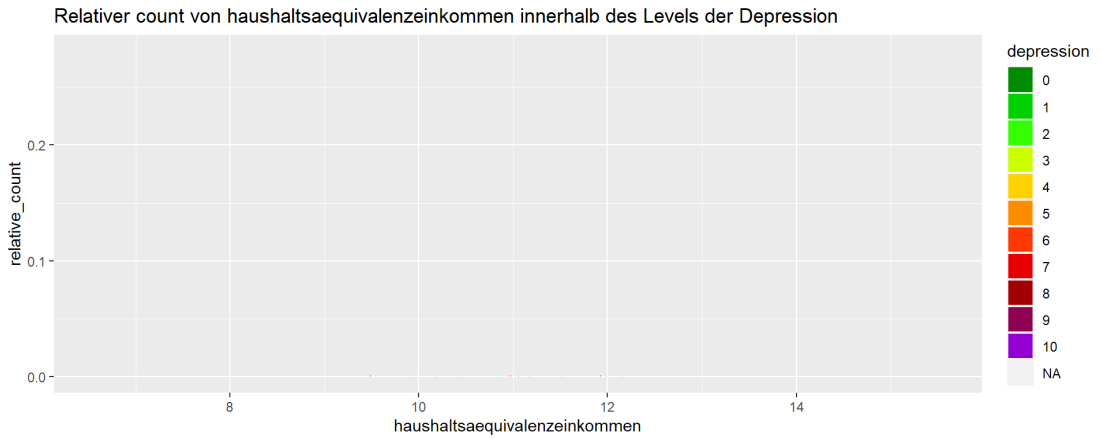

Dichte-Verteilung von hausarbeit_wochenstunden pro Level der Depression
density
0.000
0.025
0.050
0.075
0.100
0
50
100
150
hausarbeit_wochenstunden
depression
0
1
2
3
4
5
6
7
8
9
10
NA

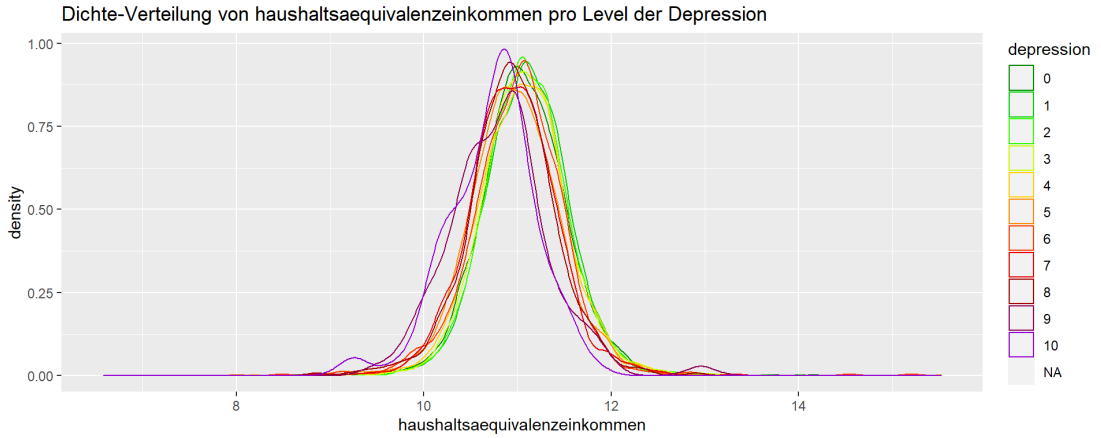

Relativer count von einschraenkung_weg_ges_zustand innerhalb des Levels der Depression
relative_count
0.0
0.2
0.4
0.6
0.0
2.5
5.0
7.5
10.0
einschraenkung_weg_ges_zustand
depression
0
1
2
3
4
5
6
7
8
9
10
NA

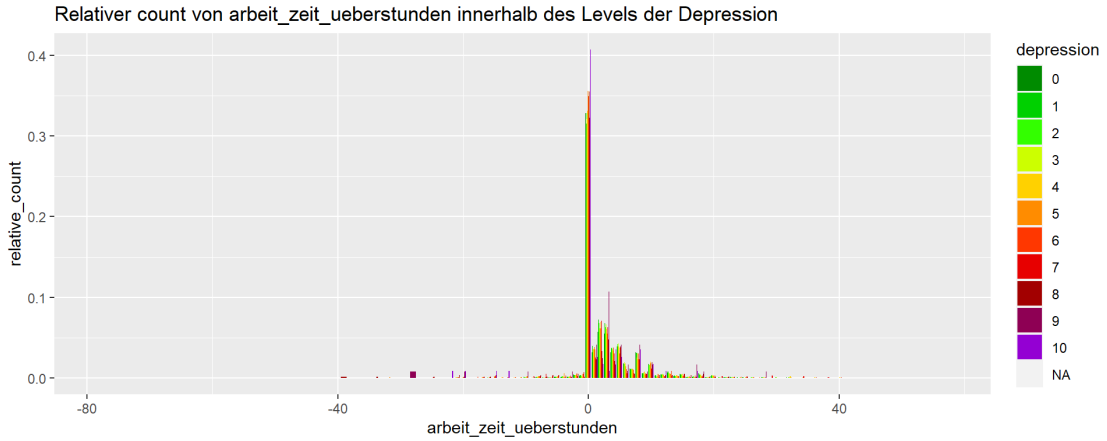

Dichte-Verteilung von einschraenkung_weg_ges_zustand pro Level der Depression
density
0
1
2
0.0
2.5
5.0
7.5
10.0
einschraenkung_weg_ges_zustand
depression
0
1
2
3
4
5
6
7
8
9
10
NA

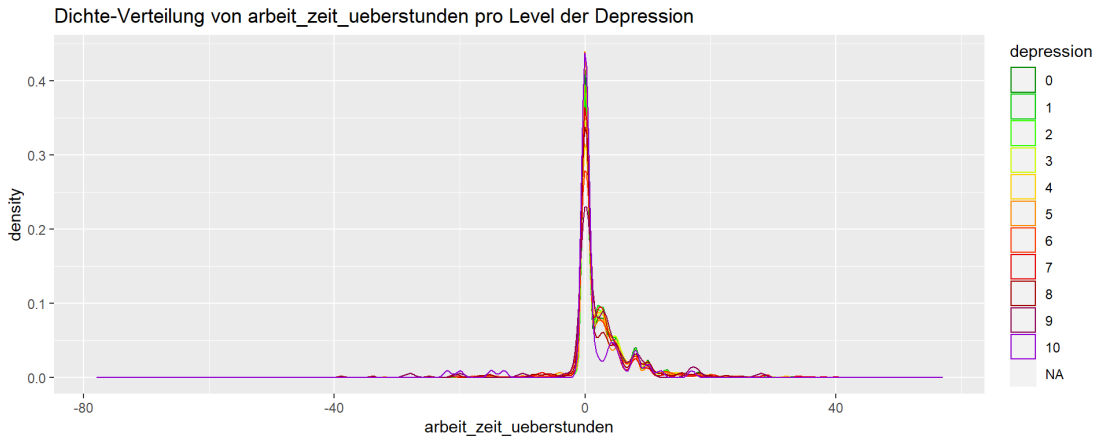

Relativer count von alter innerhalb des Levels der Depression
relative_count
0.00
0.02
0.04
0.06
0.08
20
30
40
50
60
alter
depression
0
1
2
3
4
5
6
7
8
9
10
NA


Dichte-Verteilung von alter pro Level der Depression
density
0.00
0.01
0.02
0.03
0.04
20
30
40
50
60
alter
depression
0
1
2
3
4
5
6
7
8
9
10
NA


Relativer count von haushaltsaequivalenzeinkommen innerhalb des Levels der Depression
relative_count
0.0
0.1
0.2
8
10
12
14
haushaltsaequivalenzeinkommen
depression
0
1
2
3
4
5
6
7
8
9
10
NA


Dichte-Verteilung von haushaltsaequivalenzeinkommen pro Level der Depression
density
0.00
0.25
0.50
0.75
1.00
8
10
12
14
haushaltsaequivalenzeinkommen
depression
0
1
2
3
4
5
6
7
8
9
10
NA

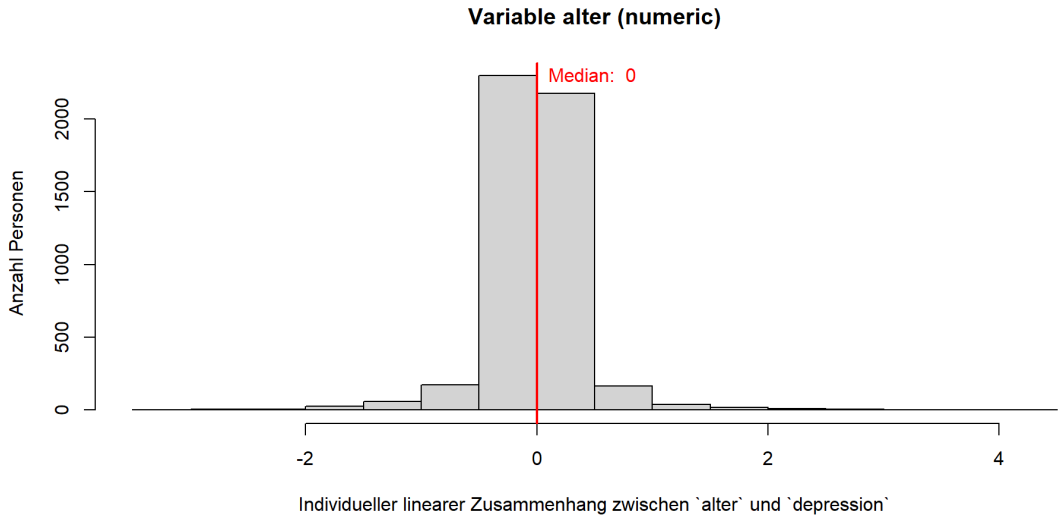

Relativer count von arbeit_zeit_ueberstunden innerhalb des Levels der Depression
relative_count
0.0
0.1
0.2
0.3
0.4
-80
-40
0
40
arbeit_zeit_ueberstunden
depression
0
1
2
3
4
5
6
7
8
9
10
NA

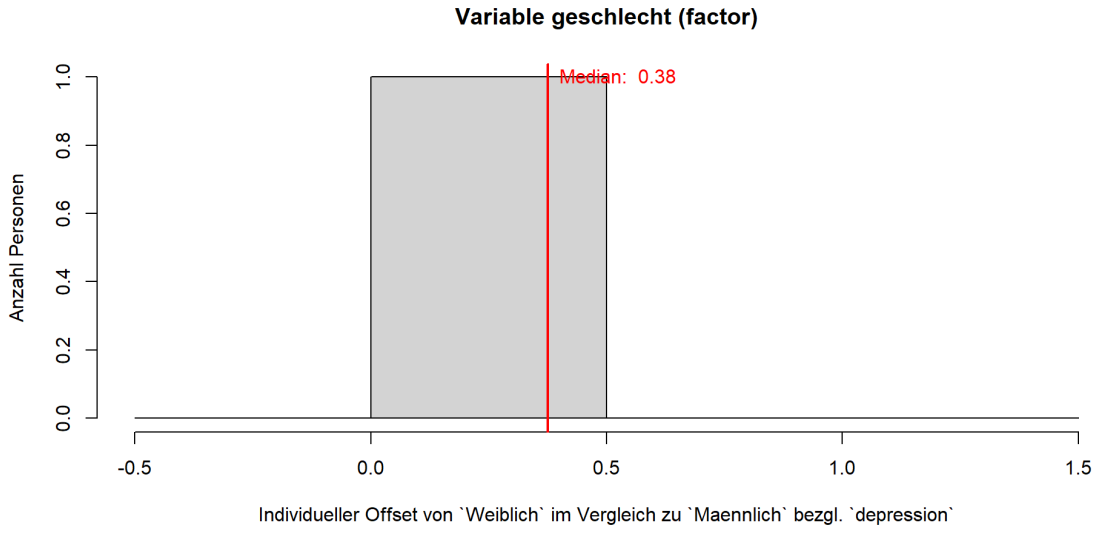

Dichte-Verteilung von arbeit_zeit_ueberstunden pro Level der Depression
density
0.0
0.1
0.2
0.3
0.4
-80
-40
0
40
arbeit_zeit_ueberstunden
depression
0
1
2
3
4
5
6
7
8
9
10
NA

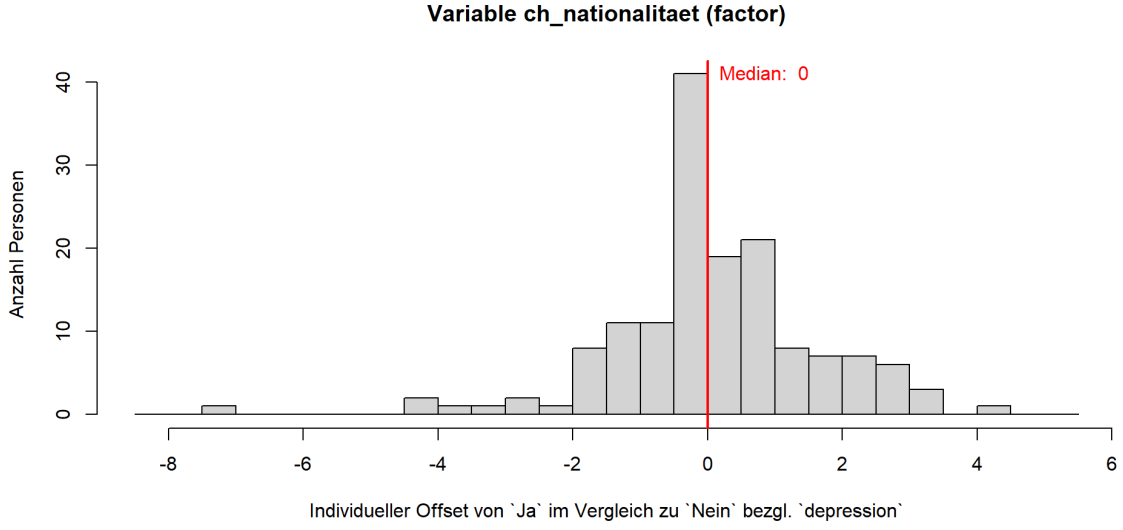
Anhang G: Gepooltes Modell – Ausgabe R/RStudio

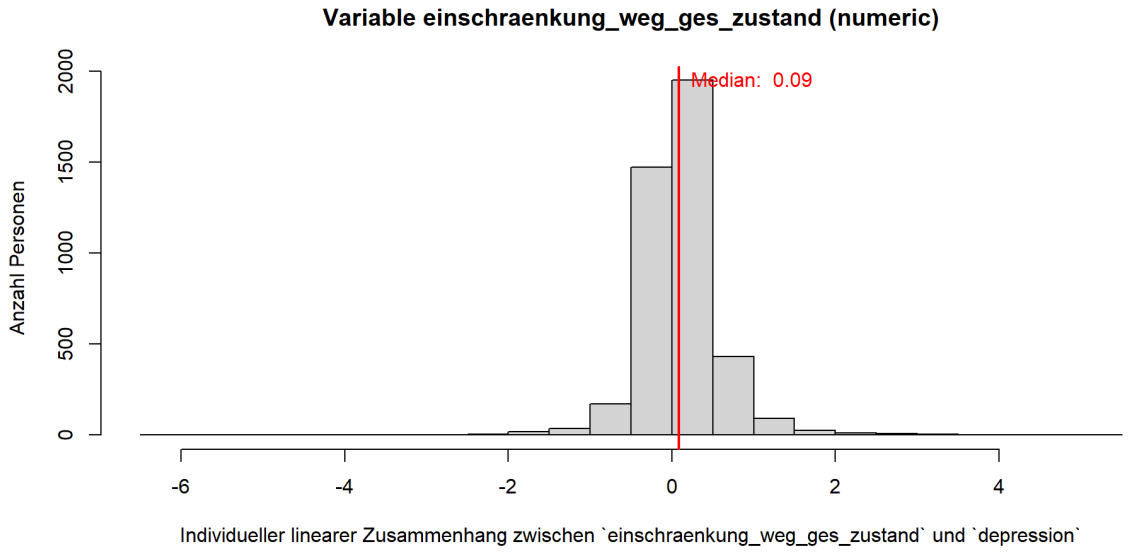
```
## Pooling Model
##
## Call:
## plm(formula = base_formula, data = df_mt, model = "pooling",
##     index = c("id", "year"))
##
## Unbalanced Panel: n = 4290, T = 1-15, N = 24053
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -5.45383 -1.17958 -0.28212  0.85399  8.86322
##
## Coefficients:
##                                              Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)                                6.26557821  0.33616868  18.6382 < 2.2e-16 ***
## ausbildungSekundarstufe II                -0.15050883  0.04510734  -3.3367 0.0008491 ***
## ausbildungHöhere Berufsbildung            -0.19604378  0.04937705  -3.9703 7.198e-05 ***
## ausbildungHochschule                       0.00739579  0.05064249   0.1460 0.8838916
## alter                                      0.00461946  0.00113583   4.0670 4.777e-05 ***
## geschlechtWeiblich                         0.34867182  0.02856149  12.2078 < 2.2e-16 ***
## ch_nationalitaetJa                        -0.18403697  0.04220694  -4.3603 1.304e-05 ***
## einschraenkung_weg_ges_zustand             0.24548948  0.00499579  49.1393 < 2.2e-16 ***
## haushaltsaequivalenzeinkommen             -0.26054171  0.02953499  -8.8215 < 2.2e-16 ***
## partnerschaftSingle                        0.31493059  0.03142142  10.0228 < 2.2e-16 ***
## tod_personKeine angehoerige Person gestorben -0.09533080  0.02712291  -3.5148 0.0004409 ***
## arbeit_einbezug_entscheidungenEntscheidung -0.15006208  0.02478075  -6.0556 1.420e-09 ***
## arbeit_qualifikationPassend                -0.17657054  0.02793808  -6.3201 2.660e-10 ***
## arbeit_zeit_wochenstunden                  -0.00729450  0.00141904  -5.1404 2.763e-07 ***
## arbeit_zeit_ueberstunden                   0.00054047  0.00239383   0.2258 0.8213775
## arbeit_zeit_nachtJa                        -0.15932397  0.03639735  -4.3774 1.206e-05 ***
## arbeit_intensitaet                         0.06431640  0.00480840  13.3759 < 2.2e-16 ***
## arbeit_zufriedenheit_atmosphaere           -0.20215808  0.00793422 -25.4793 < 2.2e-16 ***
## hausarbeit_wochenstunden                   -0.00438970  0.00169399  -2.5913 0.0095662 **
## kinder_betreuungJa                         -0.06080113  0.02524721  -2.4082 0.0160375 *
## pflege_angehoerigePflege                   -0.07405461  0.04230957  -1.7503 0.0800786 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    85703
## Residual Sum of Squares: 70184
## R-Squared:      0.18107
## Adj. R-Squared: 0.18039
## F-statistic: 265.689 on 20 and 24032 DF, p-value: < 2.22e-16
```

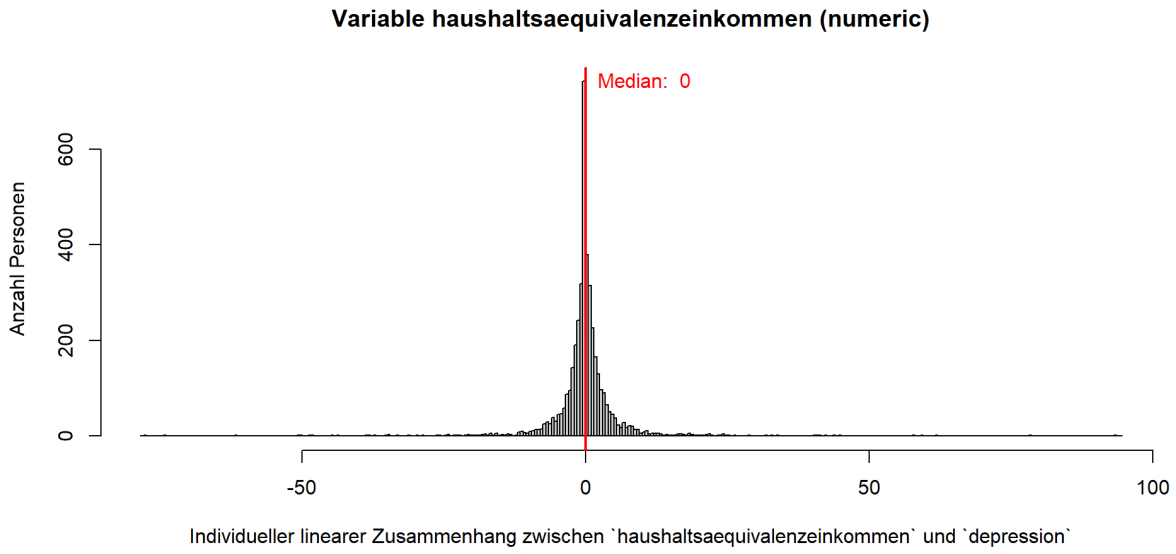
Anhang H: VCM-Koeffizienten relevanter unabhängiger Variablen

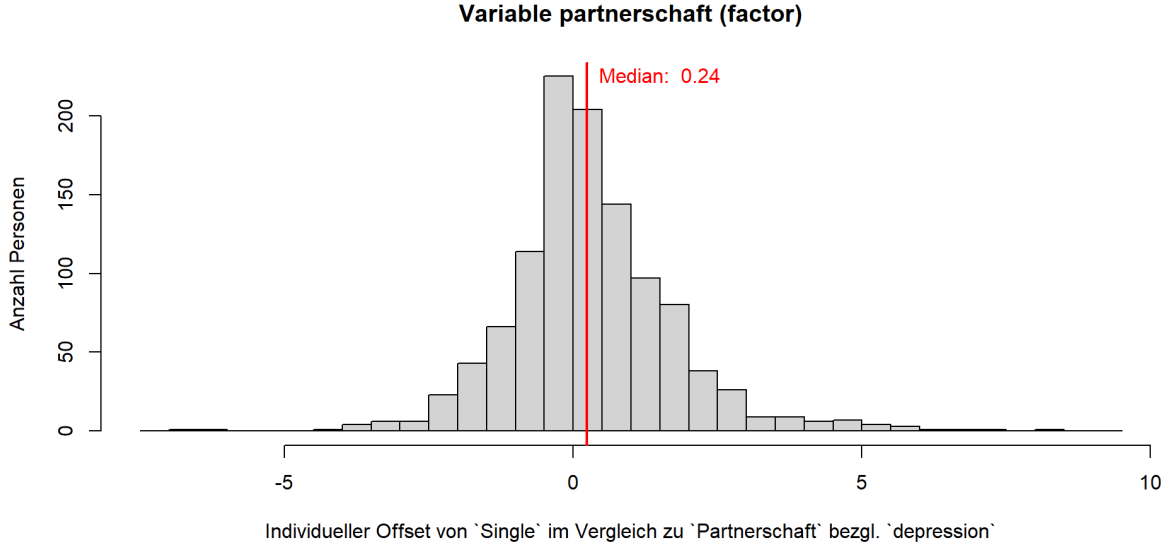


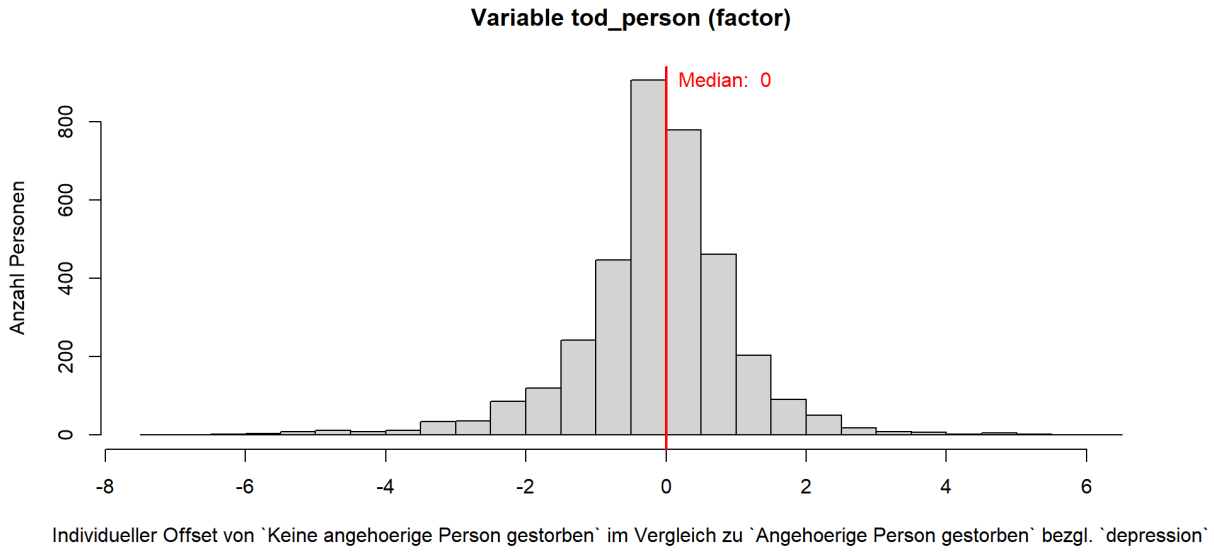


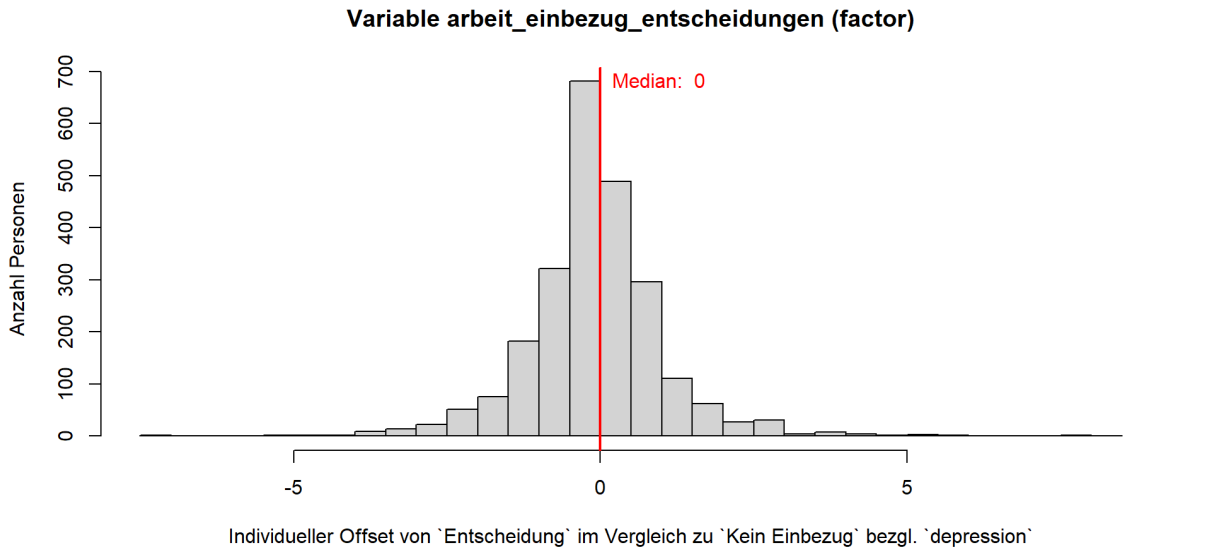


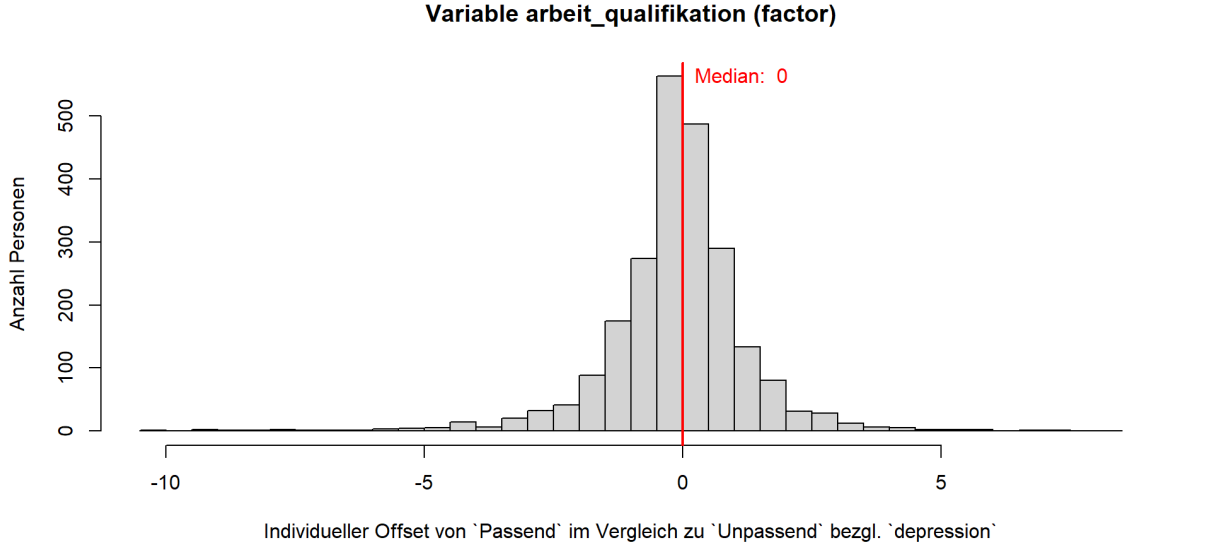

Variable einschraenkung_weg_ges_zustand (numeric)
Median: 0.09
Anzahl Personen
0
500
1000
1500
2000
-6
-4
-2
0
2
4
Individueller linearer Zusammenhang zwischen `einschraenkung_weg_ges_zustand` und `depression`

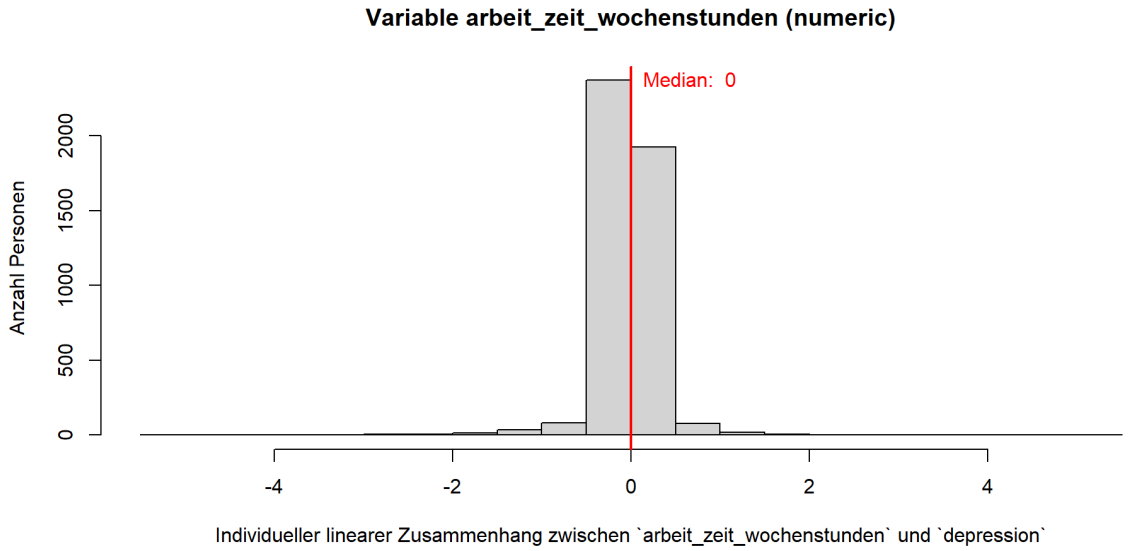

Variable haushaltsaequivalenzeinkommen (numeric)
Median: 0
Anzahl Personen
0
200
400
600
-50
0
50
100
Individueller linearer Zusammenhang zwischen `haushaltsaequivalenzeinkommen` und `depression`

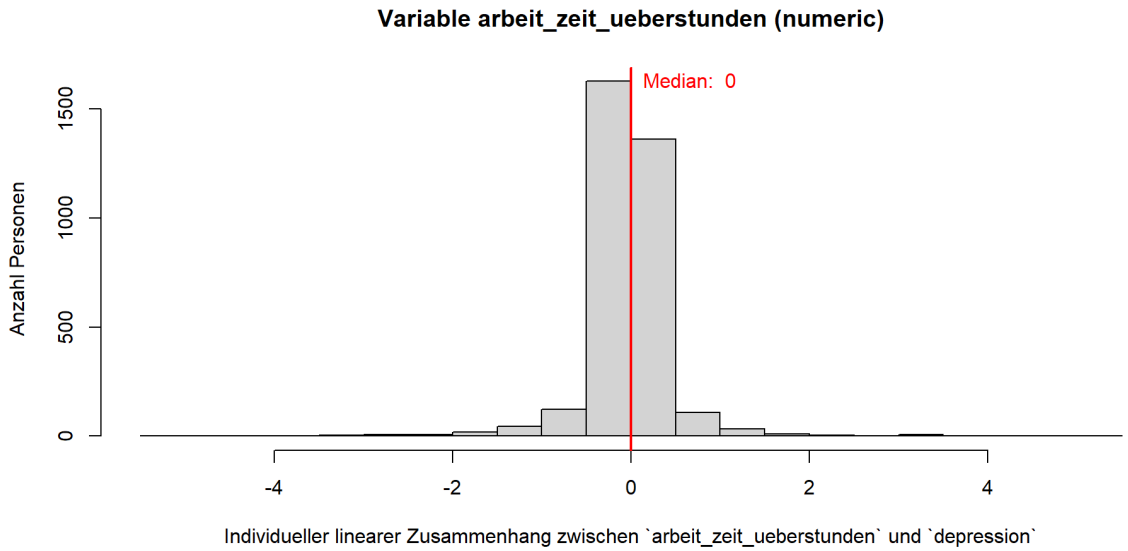

Variable partnerschaft (factor)
Median: 0.24
Anzahl Personen
0
50
100
150
200
-5
0
5
10
Individueller Offset von `Single` im Vergleich zu `Partnerschaft` bezgl. `depression`

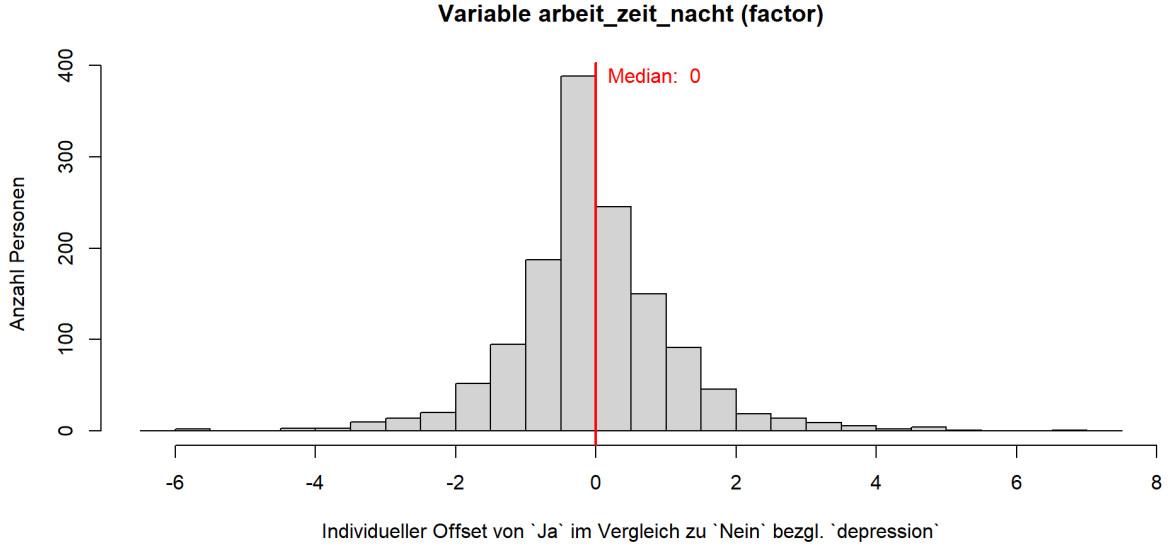

Variable tod_person (factor)
Median: 0
Anzahl Personen
0
200
400
600
800
-8
-6
-4
-2
0
2
4
6
Individueller Offset von `Keine angehoerige Person gestorben` im Vergleich zu `Angehoerige Person gestorben` bezgl. `depression`

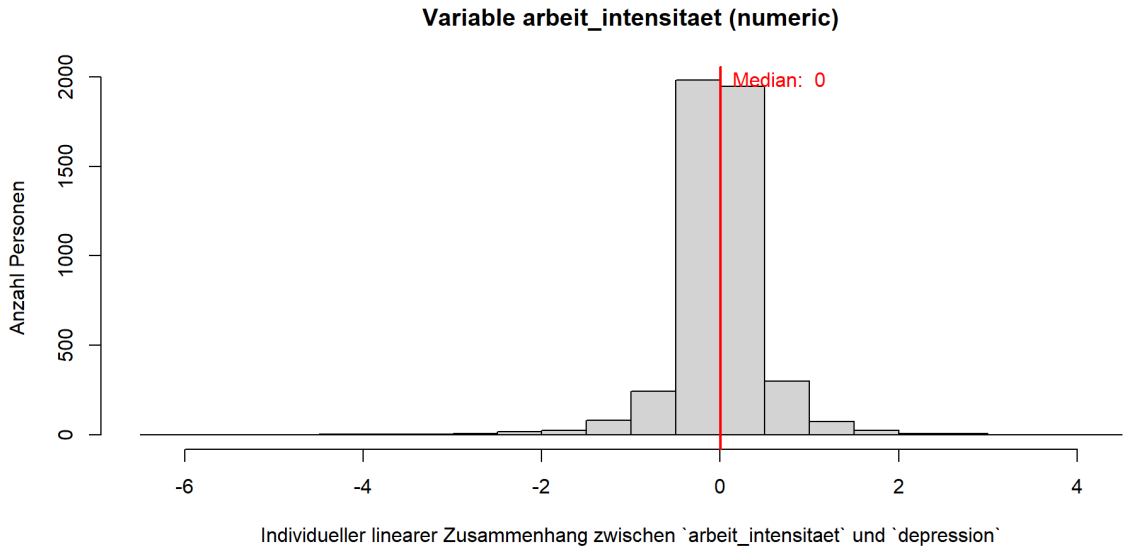

Variable arbeit_einbezug_entscheidungen (factor)
Median: 0
Anzahl Personen
0
100
200
300
400
500
600
700
-5
0
5
Individueller Offset von `Entscheidung` im Vergleich zu `Kein Einbezug` bezgl. `depression`

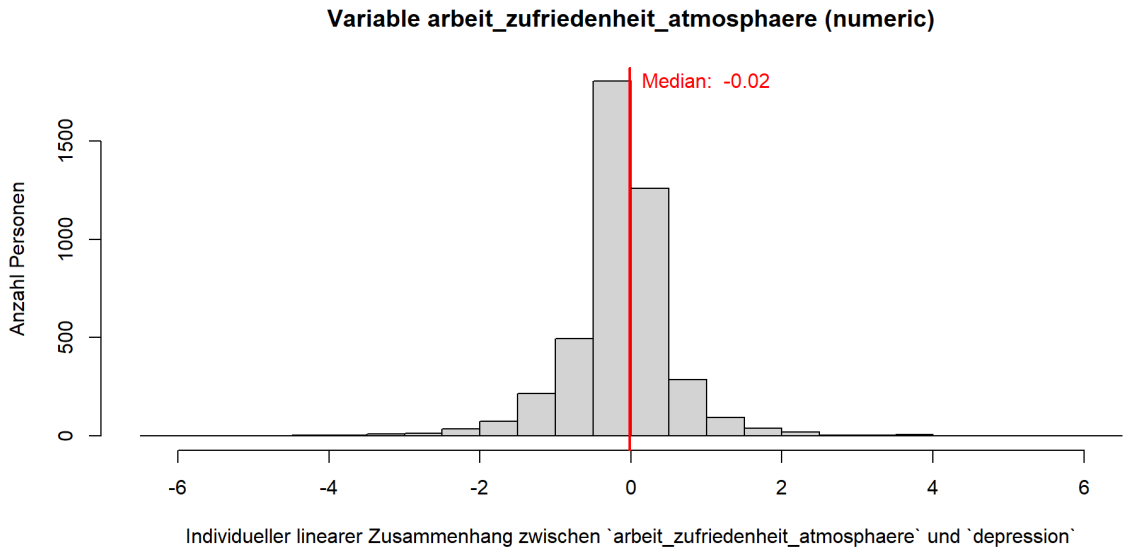

Variable arbeit_qualifikation (factor)
Median: 0
Anzahl Personen
0
100
200
300
400
500
-10
-5
0
5
Individueller Offset von `Passend` im Vergleich zu `Unpassend` bezgl. `depression`

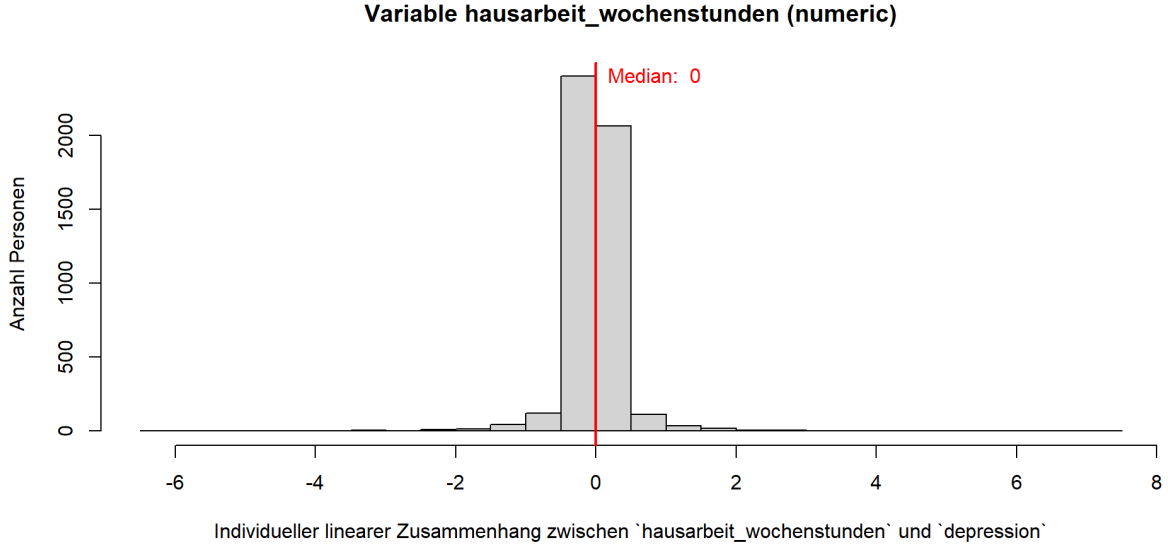

Variable arbeit_zeit_wochenstunden (numeric)
Median: 0
Anzahl Personen
0
500
1000
1500
2000
-4
-2
0
2
4
Individueller linearer Zusammenhang zwischen `arbeit_zeit_wochenstunden` und `depression`

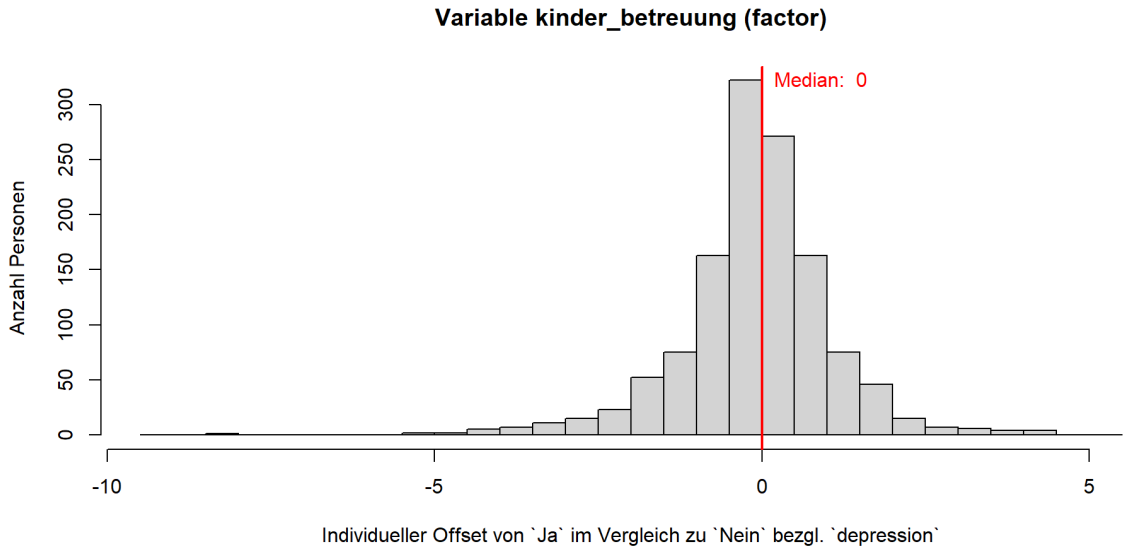

Variable arbeit_zeit_ueberstunden (numeric)
Median: 0
Anzahl Personen
0
500
1000
1500
-4
-2
0
2
4
Individueller linearer Zusammenhang zwischen `arbeit_zeit_ueberstunden` und `depression`

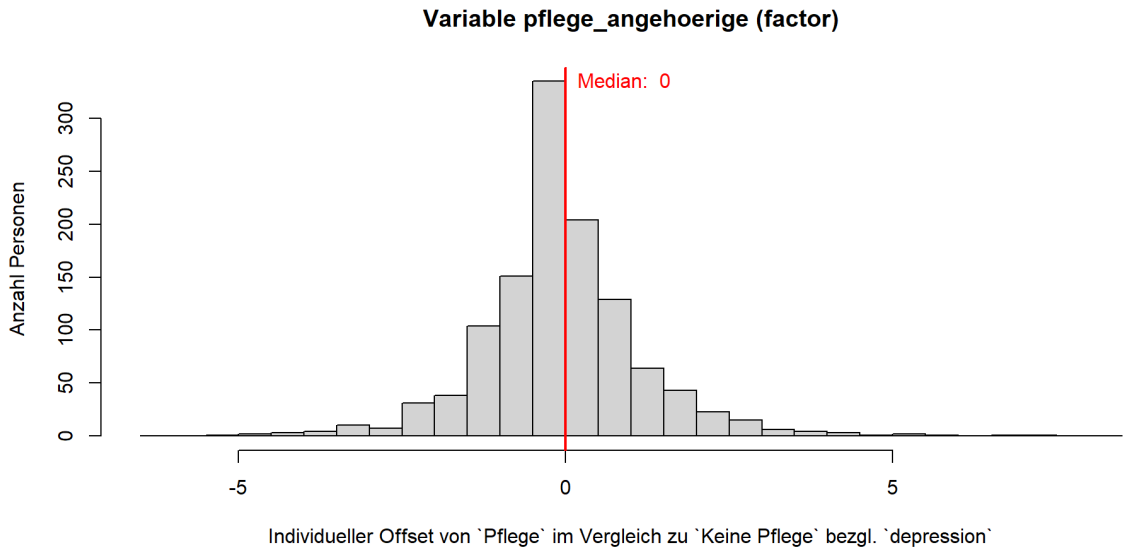

Variable arbeit_zeit_nacht (factor)
Median: 0
Anzahl Personen
0
100
200
300
400
-6
-4
-2
0
2
4
6
8
Individueller Offset von `Ja` im Vergleich zu `Nein` bezgl. `depression`


Variable arbeit_intensitaet (numeric)
Median: 0
Anzahl Personen
0
500
1000
1500
2000
-6
-4
-2
0
2
4
Individueller linearer Zusammenhang zwischen `arbeit_intensitaet` und `depression`


Variable arbeit_zufriedenheit_atmosphaere (numeric)
Median: -0.02
Anzahl Personen
0
500
1000
1500
-6
-4
-2
0
2
4
6
Individueller linearer Zusammenhang zwischen `arbeit_zufriedenheit_atmosphaere` und `depression`


Variable hausarbeit_wochenstunden (numeric)
Median: 0
Anzahl Personen
0
500
1000
1500
2000
-6
-4
-2
0
2
4
6
8
Individueller linearer Zusammenhang zwischen `hausarbeit_wochenstunden` und `depression`


Variable kinder_betreuung (factor)
Median: 0
Anzahl Personen
0
50
100
150
200
250
300
-10
-5
0
5
Individueller Offset von `Ja` im Vergleich zu `Nein` bezgl. `depression`


Variable pflege_angehoerige (factor)
Median: 0
Anzahl Personen
0
50
100
150
200
250
300
-5
0
5
Individueller Offset von `Pflege` im Vergleich zu `Keine Pflege` bezgl. `depression`

Anhang I: FD-Modell – Ausgabe R/RStudio

```
## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = base_formula, data = df_mt, model = "fd", index = c("id",
##     "year"))
##
## Unbalanced Panel: n = 4290, T = 1-15, N = 24053
## Observations used in estimation: 19763
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -10.255295  -0.893510  -0.016989   0.878021  11.304288
##
## Coefficients:
##                                              Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept)                                 0.01803880  0.01987771  0.9075 0.3641593
## ausbildungSekundarstufe II                 -0.09741310  0.33773517 -0.2884 0.7730203
## ausbildungHöhere Berufsbildung             -0.04426777  0.34275311 -0.1292 0.8972375
## ausbildungHochschule                       -0.21482667  0.36680066 -0.5857 0.5580994
## alter                                       0.00063410  0.01147044  0.0553 0.9559147
## geschlechtWeiblich                         -0.88896674  1.63152670 -0.5449 0.5858504
## ch_nationalitaetJa                          0.23243507  0.15532731  1.4964 0.1345599
## einschraenkung_weg_ges_zustand              0.11739055  0.00501524 23.4068 < 2.2e-16 ***
## haushaltsaequivalenzeinkommen              -0.08816599  0.04620001 -1.9084 0.0563599 .
## partnerschaftSingle                         0.32213548  0.05079118  6.3424 2.312e-10 ***
## tod_personKeine angehoerige Person gestorben -0.03584722  0.02107955 -1.7006 0.0890398 .
## arbeit_einbezug_entscheidungenEntscheidung  -0.07090212  0.02335356 -3.0360 0.0024003 **
## arbeit_qualifikationPassend                 -0.06372110  0.02833019 -2.2492 0.0245089 *
## arbeit_zeit_wochenstunden                   -0.00237708  0.00217773 -1.0915 0.2750486
## arbeit_zeit_ueberstunden                    -0.00053392  0.00238456 -0.2239 0.8228333
## arbeit_zeit_nachtJa                         -0.08145530  0.04348975 -1.8730 0.0610863 .
## arbeit_intensitaet                          0.01940974  0.00536282  3.6193 0.0002961 ***
## arbeit_zufriedenheit_atmosphaere            -0.05120090  0.00838463 -6.1065 1.037e-09 ***
## hausarbeit_wochenstunden                    -0.00130377  0.00192072 -0.6788 0.4972756
## kinder_betreuungJa                          0.01035268  0.05295969  0.1955 0.8450174
## pflege_angehoerigePflege                    -0.04468306  0.03772817 -1.1843 0.2362920
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    54272
## Residual Sum of Squares: 52437
## R-Squared:      0.033801
## Adj. R-Squared: 0.032822
## F-statistic: 34.532 on 20 and 19742 DF, p-value: < 2.22e-16
```

Anhang J: FE-Modell – Ausgabe R/RStudio

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = base_formula, data = df_mt, model = "within", index = c("id",
##     "year"))
##
## Unbalanced Panel: n = 4290, T = 1-15, N = 24053
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -5.848518 -0.598057 -0.040534  0.486340  8.552956
##
## Coefficients:
##                                                Estimate  Std. Error t-value  Pr(>|t|)
## ausbildungSekundarstufe II                   -0.01472870  0.23447432 -0.0628  0.949914
## ausbildungHöhere Berufsbildung               -0.03414492  0.23246905 -0.1469  0.883229
## ausbildungHochschule                         -0.07587155  0.24646523 -0.3078  0.758208
## alter                                         0.00987256  0.00244588  4.0364 5.448e-05 ***
## geschlechtWeiblich                            0.01022922  0.99319622  0.0103  0.991783
## ch_nationalitaetJa                            0.16267683  0.09373442  1.7355  0.082666 .
## einschraenkung_weg_ges_zustand                0.13516142  0.00507427 26.6366 < 2.2e-16 ***
## haushaltsaequivalenzeinkommen                -0.14811559  0.04153303 -3.5662  0.000363 ***
## partnerschaftSingle                           0.38758174  0.04481798  8.6479 < 2.2e-16 ***
## tod_personKeine angehoerige Person gestorben -0.05554241  0.02226170 -2.4950  0.012605 *
## arbeit_einbezug_entscheidungenEntscheidung   -0.03666588  0.02415061 -1.5182  0.128976
## arbeit_qualifikationPassend                  -0.05450813  0.02813112 -1.9376  0.052681 .
## arbeit_zeit_wochenstunden                    -0.00112243  0.00187421 -0.5989  0.549257
## arbeit_zeit_ueberstunden                      0.00013544  0.00239030  0.0567  0.954816
## arbeit_zeit_nachtJa                           0.01325642  0.04220450  0.3141  0.753449
## arbeit_intensitaet                            0.02202549  0.00530261  4.1537 3.285e-05 ***
## arbeit_zufriedenheit_atmosphaere             -0.05954836  0.00818275 -7.2773 3.533e-13 ***
## hausarbeit_wochenstunden                     -0.00110403  0.00187778 -0.5879  0.556575
## kinder_betreuungJa                            0.01335722  0.03572164  0.3739  0.708464
## pflege_angehoerigePflege                     -0.02089608  0.03824479 -0.5464  0.584813
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    31713
## Residual Sum of Squares: 30216
## R-Squared:      0.04721
## Adj. R-Squared: -0.16074
## F-statistic: 48.9124 on 20 and 19743 DF, p-value: < 2.22e-16
```

Anhang K: RE-Modell – Ausgabe R/RStudio

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = base_formula, data = df_mt, model = "random", index = c("id",
##     "year"))
##
## Unbalanced Panel: n = 4290, T = 1-15, N = 24053
##
## Effects:
##                 var std.dev share
## idiosyncratic 1.530   1.237 0.544
## individual    1.285   1.134 0.456
## theta:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2627  0.5614  0.6582  0.6195  0.6995  0.7288
##
## Residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -4.5629 -0.7980 -0.1995 -0.0001  0.5944  8.5282
##
## Coefficients:
##                                                Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept)                                  4.44770288  0.37476490  11.8680 < 2.2e-16 ***
## ausbildungSekundarstufe II                  -0.25948044  0.07212933  -3.5974 0.0003214 ***
## ausbildungHöhere Berufsbildung              -0.24843976  0.07790457  -3.1890 0.0014275 **
## ausbildungHochschule                        -0.11724988  0.08083666  -1.4505 0.1469319
## alter                                        0.00610336  0.00149230   4.0899 4.316e-05 ***
## geschlechtWeiblich                           0.44149933  0.04614700   9.5672 < 2.2e-16 ***
## ch_nationalitaetJa                          -0.10352845  0.05945849  -1.7412 0.0816505 .
## einschraenkung_weg_ges_zustand               0.16838737  0.00472638  35.6271 < 2.2e-16 ***
## haushaltsaequivalenzeinkommen               -0.18573047  0.03342345  -5.5569 2.746e-08 ***
## partnerschaftSingle                          0.36901514  0.03649614  10.1111 < 2.2e-16 ***
## tod_personKeine angehoerige Person gestorben -0.06609804  0.02189519  -3.0188 0.0025375 **
## arbeit_einbezug_entscheidungenEntscheidung   -0.07494856  0.02277104  -3.2914 0.0009969 ***
## arbeit_qualifikationPassend                  -0.10268956  0.02631481  -3.9023 9.526e-05 ***
## arbeit_zeit_wochenstunden                    -0.00295279  0.00156811  -1.8830 0.0596964 .
## arbeit_zeit_ueberstunden                     -0.00012800  0.00221391  -0.0578 0.9538958
## arbeit_zeit_nachtJa                          -0.01162585  0.03771350  -0.3083 0.7578787
## arbeit_intensitaet                           0.03798928  0.00480947   7.8989 2.815e-15 ***
## arbeit_zufriedenheit_atmosphaere            -0.10478433  0.00757635 -13.8304 < 2.2e-16 ***
## hausarbeit_wochenstunden                    -0.00083443  0.00170359  -0.4898 0.6242720
## kinder_betreuungJa                           0.00225870  0.02958654   0.0763 0.9391468
## pflege_angehoerigePflege                    -0.03549716  0.03697552  -0.9600 0.3370462
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    42711
## Residual Sum of Squares: 38037
## R-Squared:      0.10943
## Adj. R-Squared: 0.10869
## Chisq: 2158.32 on 20 DF, p-value: < 2.22e-16
```

Anhang L: Resultate verschiedener Tests am Paneldatensatz

plm::pFtest

```
##
##  F test for individual effects
##
## data:  base_formula
## F = 6.0888, df1 = 4289, df2 = 19743, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
##
##  F test for time effects
##
## data:  base_formula
## F = 1.5527, df1 = 14, df2 = 24018, p-value = 0.08425
## alternative hypothesis: significant effects
```

```
##
##  F test for twoways effects
##
## data:  base_formula
## F = 6.0805, df1 = 4302, df2 = 19730, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

plm::plmtest

```
## plm test for individual effect:
##
##  Lagrange Multiplier Test - (Honda) for unbalanced panels
##
## data:  base_formula
## normal = 127.09, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
## plm test for time effect:
##
##  Lagrange Multiplier Test - time effects (Honda) for unbalanced panels
##
## data:  base_formula
## normal = 1.116, p-value = 0.1322
## alternative hypothesis: significant effects
```

```
## plm test for twoways effect:
##
##  Lagrange Multiplier Test - two-ways effects (Honda) for unbalanced panels
##
## data:  base_formula
## normal = 90.657, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

plm::pwtest

```
##
##  Wooldridge's test for unobserved individual effects
##
## data:  formula
## z = 18.875, p-value < 2.2e-16
## alternative hypothesis: unobserved effect
```

```
##
##  Wooldridge's test for unobserved time effects
##
## data:  formula
## z = 0.72961, p-value = 0.4656
## alternative hypothesis: unobserved effect
```

Anhang M: RE-KV-Modell – Ausgabe R/RStudio

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = formula_mc, data = df_mt_mc, model = "random",
##     index = c("id", "year"))
##
## Unbalanced Panel: n = 4290, T = 1-15, N = 24053
##
## Effects:
##                 var std.dev share
## idiosyncratic 1.530   1.237  0.55
## individual    1.253   1.119  0.45
## theta:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2585  0.5569  0.6543  0.6156  0.6961  0.7256
##
## Residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -4.7031 -0.7614 -0.1917 -0.0022  0.5838  8.1453
##
## Coefficients:
##                                                Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept)                                  6.02644043  0.66458869   9.0679 < 2.2e-16 ***
## ausbildungSekundarstufe II                  -0.15706424  0.07110322  -2.2090 0.0271773 *
## ausbildungHöhere Berufsbildung              -0.11765633  0.07748784  -1.5184 0.1289175
## ausbildungHochschule                         0.00252939  0.08184268   0.0309 0.9753449
## geschlechtWeiblich                           0.33785680  0.05283567   6.3945 1.611e-10 ***
## ch_nationalitaetJa                          -0.02392730  0.05882789  -0.4067 0.6842034
## partnerschaftSingle                          0.35560946  0.03621020   9.8207 < 2.2e-16 ***
## tod_personKeine angehoerige Person gestorben -0.06005169  0.02165579  -2.7730 0.0055541 **
## arbeit_einbezug_entscheidungenEntscheidung   -0.05202837  0.02269191  -2.2928 0.0218586 *
## arbeit_qualifikationPassend                 -0.07675512  0.02604709  -2.9468 0.0032110 **
## arbeit_zeit_nachtJa                         -0.01139316  0.03731124  -0.3054 0.7600960
## kinder_betreuungJa                          -0.00496897  0.02942372  -0.1689 0.8658940
## pflege_angehoerigePflege                    -0.03789855  0.03658433  -1.0359 0.3002380
## arbeit_zeit_wochenstunden_mn                -0.00731565  0.00338324  -2.1623 0.0305935 *
## arbeit_zeit_wochenstunden                   -0.00155601  0.00182792  -0.8512 0.3946323
## arbeit_intensitaet_mn                        0.04969369  0.01254956   3.9598 7.501e-05 ***
## arbeit_intensitaet                           0.02297692  0.00524136   4.3838 1.166e-05 ***
## arbeit_zufriedenheit_atmosphaere_mn         -0.27303217  0.02135930 -12.7828 < 2.2e-16 ***
## arbeit_zufriedenheit_atmosphaere            -0.05709396  0.00809852  -7.0499 1.790e-12 ***
## hausarbeit_wochenstunden_mn                 -0.00412190  0.00436809  -0.9436 0.3453542
## hausarbeit_wochenstunden                    -0.00048988  0.00185333  -0.2643 0.7915302
## einschraenkung_weg_ges_zustand_mn            0.23334087  0.01374055  16.9819 < 2.2e-16 ***
## einschraenkung_weg_ges_zustand               0.13275660  0.00502742  26.4065 < 2.2e-16 ***
## alter_mn                                    -0.00523778  0.00310255  -1.6882 0.0913691 .
## alter                                        0.00862257  0.00234163   3.6823 0.0002311 ***
## haushaltsaequivalenzeinkommen_mn            -0.03558760  0.06921281  -0.5142 0.6071286
## haushaltsaequivalenzeinkommen               -0.14925577  0.04033079  -3.7008 0.0002149 ***
## arbeit_zeit_ueberstunden_mn                 -0.00473305  0.00631427  -0.7496 0.4535080
## arbeit_zeit_ueberstunden                     0.00067884  0.00237214   0.2862 0.7747454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    42905
## Residual Sum of Squares: 37222
## R-Squared:      0.13246
## Adj. R-Squared: 0.13145
## Chisq: 2848.96 on 28 DF, p-value: < 2.22e-16
```

Anhang N: IV-Modell – Ausgabe R/RStudio

```
## Oneway (individual) effect Within Model
## Instrumental variable estimation
##
## Call:
## plm(formula = instrumental_formula2, data = df_mt, model = "within",
##     index = c("id", "year"))
##
## Unbalanced Panel: n = 4290, T = 1-15, N = 24053
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -8.06231 -0.74447  0.00000  0.68958  9.65469
##
## Coefficients:
##                                              Estimate Std. Error z-value Pr(>|z|)
## einschraenkung_weg_ges_zustand                0.381929   0.169489  2.2534  0.02423 *
## partnerschaftSingle                           0.598736   0.617505  0.9696  0.33224
## tod_personKeine angehoerige Person gestorben -0.096342   3.164991 -0.0304  0.97572
## arbeit_einbezug_entscheidungenEntscheidung    0.518971   1.030931  0.5034  0.61468
## arbeit_zeit_ueberstunden                      0.037883   0.046473  0.8152  0.41498
## arbeit_intensitaet                           -0.313717   0.345845 -0.9071  0.36435
## arbeit_zufriedenheit_atmosphaere             -0.125339   0.342936 -0.3655  0.71475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    31713
## Residual Sum of Squares: 40778
## R-Squared:      0.020149
## Adj. R-Squared: -0.19292
## Chisq: 45.7366 on 7 DF, p-value: 9.8368e-08
```

Anhang O: TVCM – Ausgabe R/RStudio

```
## Tree-Based Varying Coefficients Model
##   Family: gaussian identity
##  Formula: depression ~ ausbildung + alter + geschlecht + ch_nationalitaet + einschraenkung_weg_ges_zustan
d + haushaltsaequivalenzeinkommen + partnerschaft + tod_person + arbeit_einbezug_entscheidungen + arbeit_qua
lifikation + arbeit_zeit_wochenstunden + arbeit_zeit_ueberstunden + arbeit_zeit_nacht + arbeit_intensitaet +
 arbeit_zufriedenheit_atmosphaere + hausarbeit_wochenstunden + kinder_betreuung + pflege_angehoerige + vc(pa
rtnerschaft) + vc(ausbildung, partnerschaft, by = arbeit_zeit_wochenstunden)
##     Data: df
##
## Fixed Effects:
##                                                  Estimate  Std. Error     z value
## (Intercept)                                 6.3675954785 0.353829819  17.9962093
## ausbildungSekundarstufe II                 -0.3255673031 0.128328773  -2.5369782
## ausbildungHöhere Berufsbildung             -0.4165128124 0.148231037  -2.8098894
## ausbildungHochschule                       -0.1268417268 0.145225811  -0.8734104
## alter                                       0.0043078491 0.001138899   3.7824686
## geschlechtWeiblich                          0.3540624294 0.028605684  12.3773453
## ch_nationalitaetJa                         -0.1817281814 0.042331656  -4.2929618
## einschraenkung_weg_ges_zustand              0.2448166294 0.004996859  48.9941003
## haushaltsaequivalenzeinkommen              -0.2482643335 0.029683038  -8.3638453
## partnerschaftSingle                                   NA          NA          NA
## tod_personKeine angehoerige Person gestorben -0.0942586628 0.027110377  -3.4768480
## arbeit_einbezug_entscheidungenEntscheidung -0.1507771736 0.024850407  -6.0673925
## arbeit_qualifikationPassend                -0.1758896374 0.027930039  -6.2975078
## arbeit_zeit_wochenstunden                  -0.0076942696 0.001436401  -5.3566290
## arbeit_zeit_ueberstunden                    0.0006077333 0.002427359   0.2503681
## arbeit_zeit_nachtJa                        -0.1623245941 0.036424834  -4.4564265
## arbeit_intensitaet                          0.0643421119 0.004806105  13.3875801
## arbeit_zufriedenheit_atmosphaere           -0.2022503350 0.007933694 -25.4925806
## hausarbeit_wochenstunden                   -0.0036592472 0.001707150  -2.1434829
## kinder_betreuungJa                         -0.0624554541 0.025295986  -2.4689867
## pflege_angehoerigePflege                   -0.0749374210 0.042286151  -1.7721505
##
## Varying Coefficient A: vc(partnerschaft, by = (Intercept))
## [1] root
## |   [2] partnerschaft in Partnerschaft
## |                      Estimate   Std. Error      z value
## |       (Intercept)  -0.13310855   0.02094503  -6.35513656
## |   [3] partnerschaft in Single
## |                     Estimate  Std. Error    z value
## |       (Intercept)  0.6905755   0.1086641  6.3551366
##
## Varying Coefficient B: vc(ausbildung, partnerschaft, by = arbeit_zeit_wochenstunden)
## [1] root
## |   [2] partnerschaft in Partnerschaft
## |   |   [3] ausbildung in Sekundarstufe II, Höhere Berufsbildung
## |   |   |   [4] ausbildung in Sekundarstufe II
## |   |   |                                    Estimate   Std. Error      z value
## |   |   |       arbeit_zeit_wochenstunden  0.002604037  0.001141494  2.281253743
## |   |   |   [5] ausbildung in Höhere Berufsbildung
## |   |   |                                    Estimate   Std. Error      z value
## |   |   |       arbeit_zeit_wochenstunden  0.003867298  0.001932308  2.001387983
## |   |   [6] ausbildung in Tiefer Bildungsstand, Hochschule
## |   |   |   [7] ausbildung in Tiefer Bildungsstand
## |   |   |                                     Estimate    Std. Error       z value
## |   |   |       arbeit_zeit_wochenstunden  -0.001410870   0.003422594  -0.412222453
## |   |   |   [8] ausbildung in Hochschule
## |   |   |                                    Estimate   Std. Error      z value
## |   |   |       arbeit_zeit_wochenstunden  0.000450484  0.001919815  0.234649673
## |   [9] partnerschaft in Single
## |   |   [10] ausbildung in Tiefer Bildungsstand, Sekundarstufe II, Höhere Berufsbildung
## |   |   |   [11] ausbildung in Sekundarstufe II
## |   |   |                                     Estimate    Std. Error       z value
## |   |   |       arbeit_zeit_wochenstunden  -0.010922674   0.002963274  -3.686015471
## |   |   |   [12] ausbildung in Tiefer Bildungsstand, Höhere Berufsbildung
## |   |   |   |   [13] ausbildung in Tiefer Bildungsstand
## |   |   |   |                                      Estimate    Std. Error       z value
## |   |   |   |       arbeit_zeit_wochenstunden  -0.019322816   0.004408492  -4.383090087
## |   |   |   |   [14] ausbildung in Höhere Berufsbildung
## |   |   |   |                                    Estimate   Std. Error      z value
## |   |   |   |       arbeit_zeit_wochenstunden  -0.01206634   0.00362316  -3.33033684
## |   |   [15] ausbildung in Hochschule
## |   |                                  Estimate    Std. Error       z value
## |   |       arbeit_zeit_wochenstunden  -0.006567405   0.029574178  -0.222065497
```

Anhang P: Koeffizientenschätzung linearer Regressionsmodelle

	Pooled	FD	FE	RE	RE_KV	IV
(Intercept)	6.266***	0.018		4.448***	6.026***	
	(0.336)	(0.020)		(0.375)	(0.665)	
ausbildungSekundarstufe II	-0.151***	-0.097	-0.015	-0.259***	-0.157**	
	(0.045)	(0.338)	(0.234)	(0.072)	(0.071)	
ausbildungHöhere Berufsbildung	-0.196***	-0.044	-0.034	-0.248***	-0.118	
	(0.049)	(0.343)	(0.232)	(0.078)	(0.077)	
ausbildungHochschule	0.007	-0.215	-0.076	-0.117	0.003	
	(0.051)	(0.367)	(0.246)	(0.081)	(0.082)	
alter	0.005***	0.001	0.010***	0.006***	0.009***	
	(0.001)	(0.011)	(0.002)	(0.001)	(0.002)	
geschlechtWeiblich	0.349***	-0.889	0.010	0.441***	0.338***	
	(0.029)	(1.632)	(0.993)	(0.046)	(0.053)	
ch_nationalitaetJa	-0.184***	0.232	0.163*	-0.104*	-0.024	
	(0.042)	(0.155)	(0.094)	(0.059)	(0.059)	
einschraenkung_weg_ges_zustand	0.245***	0.117***	0.135***	0.168***	0.133***	0.382**
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.169)
haushaltsaequivalenzeinkommen	-0.261***	-0.088*	-0.148***	-0.186***	-0.149***	
	(0.030)	(0.046)	(0.042)	(0.033)	(0.040)	
partnerschaftSingle	0.315***	0.322***	0.388***	0.369***	0.356***	0.599
	(0.031)	(0.051)	(0.045)	(0.036)	(0.036)	(0.618)
tod_personKeine angehoerige Person gestorben	-0.095***	-0.036*	-0.056**	-0.066***	-0.060***	-0.096
	(0.027)	(0.021)	(0.022)	(0.022)	(0.022)	(3.165)
arbeit_einbezug_entscheidungenEntscheidung	-0.150***	-0.071***	-0.037	-0.075***	-0.052**	0.519
	(0.025)	(0.023)	(0.024)	(0.023)	(0.023)	(1.031)
arbeit_qualifikationPassend	-0.177***	-0.064**	-0.055*	-0.103***	-0.077***	
	(0.028)	(0.028)	(0.028)	(0.026)	(0.026)	
arbeit_zeit_wochenstunden	-0.007***	-0.002	-0.001	-0.003*	-0.002	
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	
arbeit_zeit_ueberstunden	0.001	-0.001	0.000	0.000	0.001	0.038
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.046)
arbeit_zeit_nachtJa	-0.159***	-0.081*	0.013	-0.012	-0.011	
	(0.036)	(0.043)	(0.042)	(0.038)	(0.037)	
arbeit_intensitaet	0.064***	0.019***	0.022***	0.038***	0.023***	-0.314
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.346)
arbeit_zufriedenheit_atmosphaere	-0.202***	-0.051***	-0.060***	-0.105***	-0.057***	-0.125
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.343)
hausarbeit_wochenstunden	-0.004***	-0.001	-0.001	-0.001	0.000	
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	
kinder_betreuungJa	-0.061**	0.010	0.013	0.002	-0.005	
	(0.025)	(0.053)	(0.036)	(0.030)	(0.029)	
pflege_angehoerigePflege	-0.074*	-0.045	-0.021	-0.035	-0.038	
	(0.042)	(0.038)	(0.038)	(0.037)	(0.037)	
arbeit_zeit_wochenstunden_mn					-0.007**	
					(0.003)	
arbeit_intensitaet_mn					0.050***	
					(0.013)	
arbeit_zufriedenheit_atmosphaere_mn					-0.273***	
					(0.021)	
hausarbeit_wochenstunden_mn					-0.004	
					(0.004)	
einschraenkung_weg_ges_zustand_mn					0.233***	
					(0.014)	
alter_mn					-0.005*	
					(0.003)	
haushaltsaequivalenzeinkommen_mn					-0.036	
					(0.069)	
arbeit_zeit_ueberstunden_mn					-0.005	
					(0.006)	
Num.Obs.	24053	19763	24053	24053	24053	24053
R2	0.181	0.034	0.047	0.109	0.132	0.020
R2 Adj.	0.180	0.033	-0.161	0.109	0.131	-0.193

Eidesstattliche Erklärung

Ich erkläre hiermit,

- dass ich die vorliegende Arbeit selbstständig und ohne Mithilfe Dritter verfasst habe,
- dass ich alle verwendeten Quellen sowie alle verwendete Literatur angegeben habe,
- dass ich das Vertraulichkeitsinteresse des Auftraggebers wahren und die Urheberrechts-bestimmungen der Hochschule Luzern respektieren werde.“

St. Gallen, 3.Juni 2021

Ort, Datum

Unterschrift