

Wenn Ausreißer keine Ausreißer sind: Ein Methodenartefakt im European Social Survey (ESS), Welle 11

Peer Ole Liliental (Lehrstuhl für Sozialwissenschaftliche Datenanalyse)



Ausgangspunkt:

In einer explorativen Analyse des politischen Nachrichtenkonsums (nwspol) im European Social Survey (ESS), Welle 11, fällt in Deutschland eine kleine Gruppe mit ungewöhnlich hohen Werten auf. Histogramme, Boxplots und Dichtekurven zeigen eine stark rechtsschiefe Verteilung mit einer zweiten, deutlich abgesetzten Gruppe. Dieses Muster zeigt sich auch in Italien, Ungarn und der Ukraine, in Deutschland ist es aber am ausgeprägtesten. Die folgende Analyse konzentriert sich deshalb auf den deutschen Datensatz.

Erste Herangehensweise: Ausreißergrenze

Für eine erste Herangehensweise wird die Verteilung in zwei Gruppen eingeteilt. Dafür wird mit folgender Formel eine Grenze berechnet.

$$Q_3 + 1.5 \cdot IQR$$

Diese Grenze liegt in Deutschland bei 480 Minuten täglicher Nachrichtenkonsum. Mittels einer dichotomen Variable werden die Fälle in eine Normalkonsum- und eine Hochkonsum-Gruppe getrennt.

Der Anteil der Fälle, die sich in der Hochkonsum-Gruppe befinden, liegt bei 0,298. Mittels einer Tabelle, die die Anteile der Hochkonsum-Gruppe in den einzelnen Ausprägungen von ausgewählten Variablen angibt, lässt sich erkennen, bei welcher Ausprägung es überdurchschnittlich bzw. unterdurchschnittlich viele Fälle der Hochkonsum-Gruppe gibt.

Tab. 1: Anteil der Hochkonsum-Gruppe nach ausgewählten Merkmalen (ESS Welle 11, Deutschland, N = 2417). Hochkonsum definiert als nwspol > Q3 + 1,5 · IQR (480 Min.). Es werden nur die oberen 4 und die unteren 4 Zeilen der Tabelle dargestellt.

Variable	Ausprägung	Absolute Anzahl der Fälle in der Hochkonsum-Gruppe	Anteil der Fälle in der Hochkonsum-Gruppe
Region	Sachsen-Anhalt	43	0.683
Region	Berlin	37	0.451
Region	Brandenburg	32	0.444
Region	Mecklenburg-Vorpommern	15	0.441

...	...	...	...
Region	Thüringen	11	0.186
Vertrauen in die Polizei	Sehr (8 – 9)	80	0.186
Region	Baden-Württemberg	51	0.154
Region	Niedersachsen	24	0.121

Bereits die deskriptive Übersicht legt nahe, dass einige ostdeutsche Regionen überproportional häufig in der Hochkonsum-Gruppe vertreten sind. Aufgrund teils kleiner Fallzahlen ist dies jedoch vorsichtig zu interpretieren.

Abgleich mit der Interviewstartzeit

Da die vorherigen Herangehensweisen keine plausiblen Erklärungen liefern, ergab eine Kontaktaufnahme mit dem ESS Datenarchiv der Norwegian Agency for Shared Services in Education and Research und der Abteilung für Social Surveys des Leibniz Instituts für Sozialwissenschaften, dass die Werte der Hochkonsumgruppe mit der Interviewstartzeit zusammenhängen. Zur Identifikation der betroffenen Fälle wird die Interviewstartzeit „ainws“ in Gesamtminuten umgerechnet und mit dem erfassten Wert für „nwspol“ verglichen: Zunächst wird die Interviewstartzeit in Gesamtminuten umgerechnet.

$$t_{ainws} = h_{ainws} \times 60 + m_{ainws}$$

wobei h_{ainws} die Stunde und m_{ainws} die Minute der Interviewstartzeit bezeichnen. Anschließend wird die Differenz zwischen dem gemessenen Nachrichtenkonsum-Wert nwspol und der umgerechneten Startzeit t_{ainws,i} für jeden Fall i berechnet:

$$d_i = nwspol_i - t_{ainws,i}$$

Abschließend werden die Differenzen d_i, die innerhalb eines Toleranzbereichs von 60 Minuten liegen und somit vermutlich fehlerhaft sind (Startzeit statt Antwort erfasst), markiert.

$$error_i = \begin{cases} 1, & \text{wenn } -60 \leq d_i \leq 60 \\ 0, & \text{sonst} \end{cases}$$

Die binäre Variable errori kennzeichnet somit Fälle, bei denen der nwspol-Wert vermutlich fälschlicherweise die Interviewstartzeit statt der tatsächlichen Nachrichtenkonsum-Angabe widerspiegelt. Das Toleranzfenster von ±60 Minuten soll mögliche Verzögerungen zwischen Zeiteintrag und tatsächlichem Interviewbeginn berücksichtigen. In der deutschen Stichprobe der Welle 11 lassen sich mit dieser Variable 555 Fälle identifizieren, was 22,9 Prozent der gesamten Stichprobe entspricht. Durch diese Variable können nahezu alle Fälle oberhalb der eingangs genannten Grenze von 480 Minuten erklärt werden.

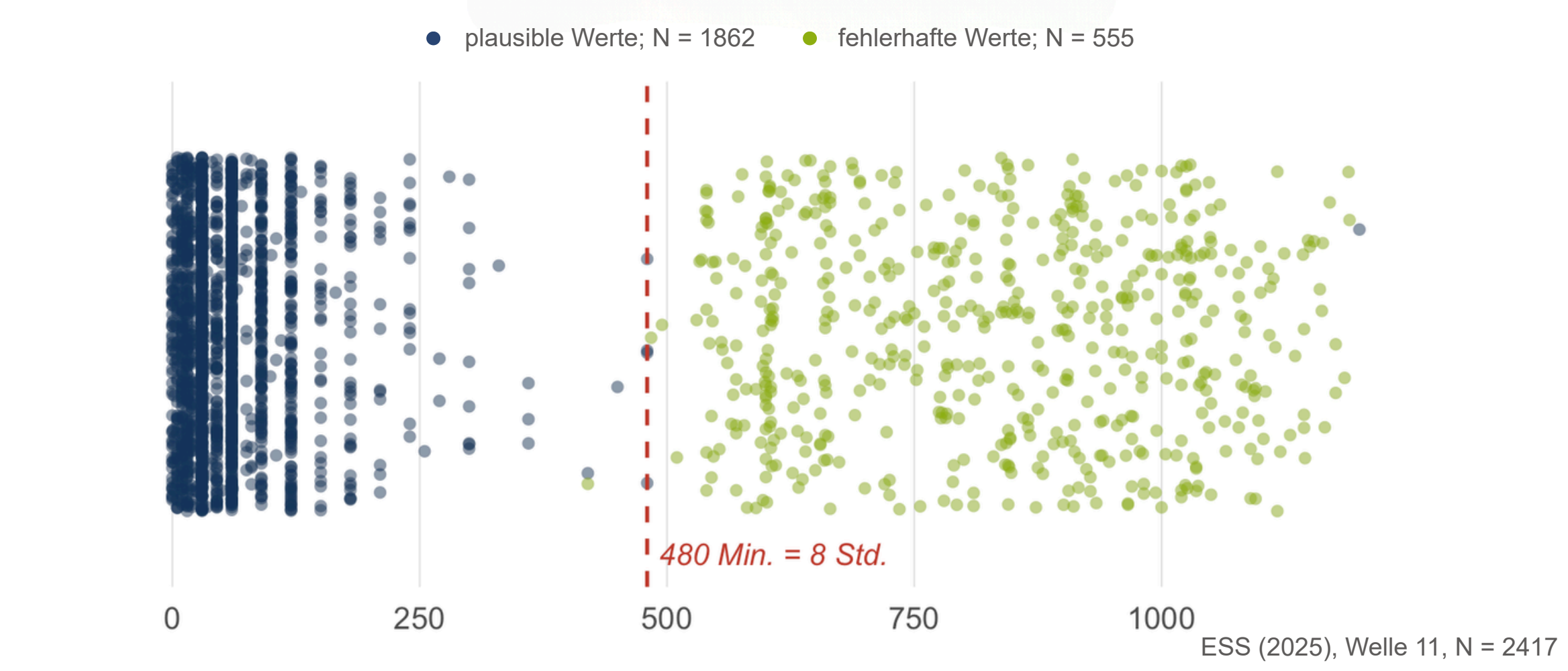
Ergebnis:

Die extremen Ausreißer im Nachrichtenkonsum-Item (nwspol) in Welle 11 lassen sich auf einen systematischen Erhebungsfehler zurückführen: Bei der computergestützten Face-to-Face-Befragung war nwspol die erste Frage im Fragebogen. Bevor die Interviewenden mit dem Fragebogen begonnen haben, sollten sie die Interviewstartzeit in das CAPI-System eintragen. In zahlreichen Fällen haben die Interviewenden versehentlich die Startzeit erneut anstelle der tatsächlichen Antwort in das Feld für das Nachrichtenkonsum-Item eingetragen. Durch diesen Interviewer-Effekt kam es zu den unplausibel hohen Werten. Ein Abgleich mit der Variable ainws (Interviewbeginn) bestätigt diese Annahme. Bundesweit sind 555 Fälle (22,9 Prozent der deutschen Stichprobe) betroffen und international 1.564 Fälle in 25 Ländern (3,1 Prozent der Gesamtstichprobe). Der Effekt verzerrt die Ergebnisse erheblich, da der deutsche Mittelwert für nwspol von 234 auf 56,6 Minuten nach Bereinigung der betroffenen Fälle sinkt. Da die korrekten Werte retrospektiv nicht rekonstruierbar sind, sollten die fehlerhaften Werte in einer neuen Version des ESS korrigiert und mit einem Gewichtungsfaktor ergänzt werden.

*Die Ost-West-Variable ist nicht unter den 36 Prädiktoren der LASSO-Regression, wird jedoch trotzdem zum Vergleich in das Modell aufgenommen, da die vorherige erste Herangehensweise, zeigt, dass in Welle 11 die Regionen in Bezug auf den Nachrichtenkonsum stark herausfallen.

Verteilung des Nachrichtenkonsums

in Minuten pro Tag in Deutschland, Welle 11



Zweite Herangehensweise: LASSO-Regression

Da die erste Herangehensweise kein plausibles Ergebnis für die hohen Werte des Nachrichtenkonsums liefert, wird nun in der zweiten Herangehensweise eine LASSO-Regression (Least Absolute Shrinkage and Selection Operator) verwendet. Die LASSO-Regression dient zur Identifikation relevanter Prädiktoren für den Nachrichtenkonsum. LASSO ist ein Regressionsverfahren, welches im Gegensatz zur klassischen OLS-Regression einen Strafterm (L1-Penalty) auf die Koeffizienten legt, wodurch irrelevante Prädiktoren automatisch auf null gesetzt werden. Die LASSO-Regression wird ausschließlich auf Basis der Welle 10 durchgeführt, da diese im Vergleich zur Welle 11 keine Hochkonsumgruppe besitzt.

$$\beta_{LASSO} = argmin_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

Ergebnis und Weiterverarbeitung:

Das lambda.min-Modell selektierte 36 Prädiktoren mit von null verschiedenen Koeffizienten, darunter politisches Interesse, Altersgruppe und Internetnutzung als stärkste Effekte. Von den ausgewählten Variablen wurden die stärksten anschließend in einem gewichteten OLS-Modell (Post-LASSO-Ansatz) für die Wellen 9, 10 und 11 verwendet, um interpretierbare Koeffizienten mit Standardfehlern und Signifikanzniveaus zu erhalten. Die ausgewählten Variablen erklären den Nachrichtenkonsum in Welle 9 und 10 mit einem R² von 0,100 bzw. 0,205 ausreichend gut, jedoch sinkt das R² in Welle 11 auf 0,082 ab. Außerdem ändern sich die Werte sowie Vorzeichen der einzelnen Variablen und die Signifikanzen nehmen ab. Die zusätzlich eingefügte Ost-West-Variable zeigte in Welle 9 keine Signifikanz und in Welle 10 nur eine schwache Signifikanz (p < 0,05) bei geringem Effekt. In Welle 11 hingegen wird sie hochsignifikant (p < 0,001) mit einem massiv erhöhten Koeffizienten.

Tab. 2: Gewichtete OLS-Regression (Post-LASSO) für nwspol nach ESS-Welle. Standardfehler in Klammern.

	Welle 9	Welle 10	Welle 11
(Intercept)	92.070** (33.344)	131.293*** (12.037)	-29.455 (133.676)
Ost- und Westdeutschland	-1304 (7.092)	5.085* (2.576)	149.131*** (25.182)
Politisches Interesse	-16.952*** (3.512)	-28.665*** (1.368)	14547 (12.869)
Altersgruppe	9.050*** (2.025)	9.376*** (0.719)	15.532* (7.822)
Internetnutzung (Min./Tag)	0.039** (0.015)	0.058*** (0.005)	-0.034 (0.059)
Scham bei homosex. Familienmitglied	-6.068 (3.171)	-5.105*** (1.169)	14.715 (13.220)
Vertrauen in Massenmedien	-6.858 (4.005)	-3.729* (1.522)	8.827 (15.690)
Sicherheitsgefühl abends	8.240* (3.405)	2.676* (1.227)	13.612 (12.714)
Wohnort-Typ	-2.554 (2.543)	0.068 (0.897)	-47.648*** (9.627)
Erwerbstätig	-14.355* (5.631)	-6.001** (2.093)	-34.482 (21.419)
Kinder im Haushalt	13.854 (7.444)	-3.335 (2.533)	122.054*** (28.332)
Vertrauen in UN	-3.055** (1.178)	-1.267** (0.414)	6254 (4.420)
Adoptionsrecht gleichgeschl. Paare	2.164 (2.561)	-1.827 (1.036)	9.105 (11.266)
Num.Obs.	1147	4551	1336
R ²	0.100	0.205	0.082
R ² Adj.	0.090	0.203	0.073

N: Welle 9 = 1147; Welle 10 = 4551; Welle 11 = 1336 | * p < 0.05, ** p < 0.01, *** p < 0.001

Datenquelle:

European Social Survey European Research Infrastructure (ESS ERIC) (2023 - 2025) ESS11 - integrated file, edition 4.1 [ESS11e04], ESS10 - integrated file, edition 3.3 [ESS10e03_3], ESS9 - integrated file, edition 3.3 [ESS9e03_3] Sikt - Norwegian Agency for Shared Services in Education and Research.