TU Dortmund

Fallstudien I - Sommersemester 2025

Projekt 5: Logistische Regression und Odds Ratios am Beispiel von COVID19-Impfungen

Dozenten:

JProf. Dr. Dennis Dobler

Loreen Sabel

Dr. Susanne Frick

Prof. Dr. Philipp Doebler

Verfasser: Quang Vinh Nguyen

Gruppenmitglieder:
Alain Kauth, Ibrahim Rahman

13. Juli 2025

Inhaltsverzeichnis

1. Einleitung

Die COVID19-Pandemie stellte nicht nur eine globale Gesundheitskrise dar, sondern führte auch zu tiefgreifenden gesellschaftlichen Spannungen hinsichtlich Vertrauen in Politik, Wissenschaft und kollektive Schutzmaßnahmen. Eine zentrale Maßnahme zur Bekämpfung der Pandemie war die Impfung gegen das Coronavirus. Trotz breiter Impfangebote zeigten sich in der Bevölkerung deutliche Unterschiede in der Impfbereitschaft (Lazarus et al., 2021), die unter anderem mit Vertrauen in Institutionen und die Risikowahrnehmung zusammenhängen (Betsch et al., 2020).

Ziel dieses Projekts ist es daher zu untersuchen, welche Faktoren beeinflussen, ob sich jemand gegen COVID19 impfen lässt. Methodisch kommt ein logistisches Regressionsmodell zum Einsatz, um den individuellen Impfstatus in Abhängigkeit verschiedener erklärender Variablen zu modellieren. Die Prädiktoren des Modells werden anhand des bayesschen Informationskriteriums ausgewählt. Zur Bewertung der Modellgüte wird die Konfusionsmatrix herangezogen. Die Analyse basiert auf Daten der 11. Runde der European Social Survey (ESS ERIC, 2025) und konzentriert sich auf Deutschland und Länder mit vergleichbarem Zugang zu Gesundheitssystemen, wie Österreich, Belgien und Frankreich. Die Ergebnisse zeigen, dass Alter sowie Vertrauen besonders mit dem individuellen Impfstatus assoziiert sind. Ältere Personen und solche mit höherem Vertrauen lassen sich mit höherer Wahrscheinlichkeit gegen COVID19 impfen.

Der Bericht beginnt mit der Beschreibung des verwendeten Datensatzes in Kapitel 2, wobei ein besonderer Fokus auf die Variablen gelegt wird, die zur Entwicklung des logistischen Regressionsmodells berücksichtigt werden. Kapitel 3 stellt die methodischen Grundlagen der logistischen Regression dar, einschließlich Schätzverfahren, Modellvoraussetzungen und Interpretation. In Kapitel 4 wird der Modellentwicklungsprozess beschrieben. Danach wird das finale Modell auf Schätzbarkeit geprüft, hinsichtlich seiner Güte bewertet und inhaltlich interpretiert. Abschließend fasst Kapitel 5 die Ergebnisse zusammen und diskutiert deren gesellschaftliche und politische Relevanz.

2. Problemstellung

Dieses Kapitel beginnt mit der detaillierten Beschreibung der Datenstruktur sowie der Erläuterung der Variablen. Im Folgenden werden die Projektziele und die Vorgehensweise ausführlich dargelegt.

2.1. Datenmaterial

Die Daten stammen aus der 11. Runde der European Social Survey (ESS ERIC, 2025), einer regelmäßigen länderübergreifenden Erhebung zu gesellschaftlichen Einstellungen, Verhalten und soziodemografischen Merkmalen in Europa. Alle Details zur Methodik, Kodierung und Übersetzung sind auf der offiziellen Website[1] der European Social Survey (ESS) dokumentiert.

Der verwendete Datensatz umfasst Informationen zu unterschiedlichen Themen wie Demografie, Verhaltenseinstellungen, usw. von 46 162 Befragten aus 28 europäischen Ländern. Eine Übersicht aller in die Analyse einbezogenen Variablen findet sich in Tabelle 1.

Tabelle 1: Übersicht über die berücksichtigten Variablen

Variable	**Bedeutung**	**Variablentyp**	**Ausprägungen**
Zielvariablen (abhängige Variablen)			
`Impfstatus`	Impfstatus mit mindestens einer Impfdosis	dichotom	{nein, ja}
potenzielle Prädiktoren (unabhängige Variablen)			
`Infektion`	COVID-Infektion seit Anfang 2020	nominal	{ja, vermutet, nein}
`Bildungsgrad`	Höchstes Bildungsniveau	ordinal	{1,...,7, Andere}
`Einkommen`	Dezil des Haushaltseinkommens der befragten Person	ordinal	{1,...,10}
`Haushaltsgröße`	Anzahl der Personen im Haushalt der befragten Person	numerisch	natürliche Zahlen
`Internetnutzung`	Häufigkeit der Internetnutzung	ordinal	{1,...,5}
`Geschlecht`	Geschlecht der befragten Person	nomial	{männlich, weiblich}
`Familienstand`	Kodierung des rechtlichen Familienstands	nomial	{1,...,6}
`Wohnort`	Art des Domizils nach abnehmendem Urbanitätsgrad	ordinal	{1,...,5}
`Alter`	Alter der befragten Person	numerisch	natürliche Zahlen
`Vertrauensindex`	Mittelwert der Vertrauensvariablen	numerisch	[0,10]

Familienstand: 1=verheiratet, 2=in einer Lebenspartnerschaft, 3=getrennt lebend, 4=geschieden, 5=verwitwet, 6=andere

Die meisten Variablen werden als separate Faktoren behandelt. Fünf Variablen - die das Vertrauen in staatliche Institutionen wie Parlament, Rechtssystem, Polizei, Politiker und politische Parteien abbilden - werden durch Mittelwertbildung zu einer aggregierten Variable - dem Vertrauensindex zusammengefasst. Die ursprünglichen Vertrauensvariablen sind alle ordinal skaliert mit Ausprägungen als Ganzzahlen von null bis zehn. Aufgrund der Mittelwertbildung weist der Vertrauensindex einen Wertebereich von null bis zehn auf, wird jedoch als eine metrische Variable interpretiert. In generalisierten linearen Modellen werden ordinal skalierte Variablen bei Verwendung von Dummy-Kodierungen wie nominale Variablen behandelt. Die Bestimmung des Skalenniveaus kategorialer Variablen als nominal oder ordinal in Tabelle 1 beeinflusst daher nicht die Schätzung, sondern lediglich die Interpretation der Ergebnisse.

[1] `https://www.europeansocialsurvey.org/`

Die Methodik der ESS ist international abgestimmt und sichert hohe Vergleichbarkeit und Zuverlässigkeit der Angaben (vgl. ESS ERIC, 2025). Im Datensatz treten fehlende Werte auf, da Befragte berechtigt sind, Antworten zu verweigern oder keine Angabe zu machen. Aufgrund des großen Stichprobenumfangs bleibt die Aussagekraft der Daten dennoch hoch. Zu beachten ist, dass der verwendete Datensatz keine Informationen zum individuellen Zugang zur Gesundheitsversorgung oder zur Qualität nationaler Gesundheitssysteme enthält, was die Kontextualisierung von Impfentscheidungen einschränkt.

2.2. Ziele des Projekts

Das vorliegende Projekt hat zum Ziel, Faktoren zu identifizieren, die Einfluss auf die Entscheidung für oder gegen die COVID19-Impfung haben. Da es sich beim Impfstatus um eine dichotome Zielvariable handelt, gilt die logistische Regression als eine geeignete Methode. Der Vollständigkeit halber werden alle Variablen des Datensatzes berücksichtigt. Allerdings besteht zwischen Vertrauensvariablen in der Regel eine gewisse Korrelation. Deshalb kommt der Vertrauensindex als eine repräsentative Variable zum Einsatz. Darüber hinaus sollte der Einfluss des Zugangs zur Gesundheitsversorgung auf den Impfstatus nicht vernachlässigt werden. Majcherek et al. (2024) erwähnte die bestehenden Ungleichheiten beim Zugang zur Gesundheitsversorgung zwischen den europäischen Ländern und ordnete Deutschland zusammen mit Österreich, Belgien, Tschechien (nicht im Datensatz enthalten) und Frankreich in eine Gruppe (EU-Länder mit moderatem Zugang zur Gesundheitsversorgung) ein. Um Verzerrungen aufgrund Unterschiede in Versorgungsbedingungen zu vermeiden, wird die Analyse daher nur auf diese Länder beschränkt.

Die Auswahl der Prädiktoren für das logistische Regressionsmodell erfolgt auf Grundlage des bayesschen Informationskriteriums (BIC), wobei das Modell mit dem niedrigsten BIC-Wert als das optimalste angesehen wird. Anschließend wird die Schätzbarkeit des finalen Modells überprüft. Dazu wird getestet, ob eine perfekte Trennung der Zielvariable durch einzelne Prädiktoren vorliegt – bei kategorialen Variablen mithilfe von Kontingenztafeln, bei numerischen Variablen anhand von Boxplots. Zusätzlich wird potenzielle Multikollinearität durch Berechnung des Varianzinflationsfaktors (VIF) geprüft. Theoretisch sollte die Schätzbarkeit für alle Kandidatenmodelle gegeben sein. In der Praxis genügt es jedoch, diese für das ausgewählte Modell sicherzustellen. Abschließend wird das finale Modell anhand der Konfusionsmatrix hinsichtlich seiner Trennschärfe bewertet und inhaltlich interpretiert.

3. Statistische Methoden

Die Datenverarbeitung und Anwendung der hier aufgeführten Methoden erfolgt mittels der Software *R* mit Version 4.4.2 (R Core Team, 2024). Zum Erstellen des Diagramms wird zusätzlich das Paket *ggplot2* (Wickham et al., 2016) verwendet. Andere erweiterte Aufgaben werden von Zusatzpaketen ausgeführt, wie etwa der umfassende Modellvergleich (Bartoń, 2025) oder Erstellung von Konfusionsmatrizen (Kuhn und Max, 2008).

Der Kürze und Konsistenz halber werden in der Beschreibung aller Methoden folgenden Notationsregeln verwendet, ohne Erläuterungen explizit zu wiederholen. Zufallsvariablen (inkl. Teststatistik) werden in Großbuchstaben und Realisierungen in Kleinbuchstaben dargestellt. Vektoren sind in fetter Kursivschrift, Matrizen auch in fetter, aber aufrechter Schrift notiert. Die griechischen Buchstaben werden einheitlich klein geschrieben, ihre Bedeutung lässt sich jedoch aus den lateinischen Buchstaben im Ausdruck erschließen.

3.1. Binäre Regressionsmodelle

Logistische Regressionsmodelle werden eingesetzt, um den Zusammenhang zwischen unabhängigen Variablen und einer binären Zielgröße zu analysieren. Die lineare Regressionsanalyse ist für diesen Zweck ungeeignet, da sie eine numerische Zielvariable erfordert und Werte außerhalb des gültigen Bereichs $[0, 1]$ liefern kann. Das logistische Modell löst dieses Problem durch eine nichtlineare Transformation des linearen Prädiktors und stellt sicher, dass die vorhergesagten Wahrscheinlichkeiten stets im Intervall $[0, 1]$ liegen.

Grundlage des logistischen Modells ist die Annahme, dass die Zufallsvariablen $Y_i \in \{0, 1\}$ für $i = 1, \ldots, n$ bei gegebenen Kovariaten $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})$ bedingt unabhängig und Bernoulli-verteilt sind. Die Kovariaten x_{ij} müssen mindestens intervallskaliert sein, wenn sie als metrisch betrachtet werden. kategoriale Variablen müssen entsprechend Dummykodiert werden (Toutenburg, 2003, S.71). Die Wahrscheinlichkeit $\pi_i = \mathbb{P}(Y_i = 1 \mid \boldsymbol{x}_i)$ für das Eintreten des Ereignisses $Y_i = 1$ wird durch den sogenannten linearen Prädiktor $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ und die logistische Linkfunktion als

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}. \tag{1}$$

modelliert (Fahrmeir et al., 2009, S.192), wobei $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ der Parametervektor ist. Die geschätzten Wahrscheinlichkeiten π_i liegen definitionsgemäß stets im Intervall $[0, 1]$.

Die Parameterschätzung in logistischen Modellen erfolgt über das Maximum-Likelihood-Verfahren, das durch Übergang zur Log-Likelihood-Funktion

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)]$$

vereinfacht wird (Fahrmeir et al., 2009, S.198-199). Diese Funktion wird numerisch maximiert, da keine geschlossene Lösung existiert. Übliche Optimierungsverfahren sind Newton-Raphson- oder Fisher-Scoring-Verfahren, die im Fall der logistischen Regression identisch sind (Fahrmeir et al., 2009, S.202). Das Fisher-Scoring-Verfahren ist ein iteratives Schätzverfahren zur Bestimmung des Maximum-Likelihood-Schätzers $\hat{\boldsymbol{\beta}}$. Ausgehend von einem beliebigen Startwert $\hat{\boldsymbol{\beta}}^{(0)}$ werden die Parameter in jedem Schritt nach folgender Aktualisierungsregel angepasst:

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \mathbf{F}^{-1}(\hat{\boldsymbol{\beta}}^{(k)}) \cdot \boldsymbol{s}(\hat{\boldsymbol{\beta}}^{(k)}), \quad k = 0, 1, 2, \ldots,$$

wobei $\boldsymbol{s}(\boldsymbol{\beta}) = \frac{\partial \ell(\beta)}{\partial \beta}$ die Score-Funktion und $\mathbf{F}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\beta)}{\partial \beta \, \partial \beta^\top}$ die Fisher-Informationsmatrix darstellt. Die Iteration wird beendet, sobald ein Abbruchkriterium wie $\frac{\|\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}\|}{\|\hat{\boldsymbol{\beta}}^{(k)}\|} \leq \varepsilon$ erfüllt ist. Der Vektor $\hat{\boldsymbol{\beta}}^{(k)}$ wird dann als Konvergenzschätzer angenommen. Damit das Fisher-Scoring-Verfahren konvergiert, muss die Fisher-Informationsmatrix $\mathbf{F}(\boldsymbol{\beta})$ im jeweiligen Iterationsschritt invertierbar sein. Diese Voraussetzung ist erfüllt, wenn die Designmatrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ vollen Rang p besitzt (Fahrmeir et al., 2009, S.202), das heißt, wenn die Spaltenvektoren von $\mathbf{X}$ linear unabhängig sind. Wenn eine oder mehrere Kovariablen die Zielvariable vollständig trennen, d.h. es existiert eine lineare Trennfunktion, die alle Beobachtungen mit $y_i = 1$ von denen mit $y_i = 0$ fehlerfrei unterscheidet (perfekte Separation), dann existiert kein endlicher Maximum-Likelihood-Schätzer $\hat{\boldsymbol{\beta}}$, da die Likelihoodfunktion im Unendlichen maximiert wird Silvapulle (1981).

Durch Umformung der Ungleichung (1) (Fahrmeir et al., 2009, S.192) ergibt sich die sogenannte Logit-Darstellung:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

Dies bedeutet, dass das logarithmierte Verhältnis der Wahrscheinlichkeit für ein positives Ereignis zur Gegenwahrscheinlichkeit (die sogenannten Odds) linear durch die Kovariaten erklärt wird. Jeder Regressionskoeffizient β_j lässt sich als Veränderung des Log-Odds interpretieren, wenn die Kovariate x_{ij} um eine Einheit erhöht wird. Wird β_j exponen-

tiiert, ergibt sich der sogenannte Odds-Ratio, also der multiplikative Effekt von x_j auf die Chancen des Eintretens von $Y = 1$ (Fahrmeir et al., 2009, S.194). Die Interpretation der Modellparameter erfolgt in zwei Schritten: Zunächst wird der lineare Effekt jeder Kovariate auf den Prädiktor η_i betrachtet. Anschließend wird dieser Effekt durch die logistische Funktion in eine nichtlineare Wirkung auf die Wahrscheinlichkeit transformiert. Ein positiver Koeffizient bedeutet, dass mit wachsendem x_j die Wahrscheinlichkeit für $Y = 1$ steigt; ein negativer Koeffizient deutet auf einen gegenteiligen Zusammenhang hin. Ist $\beta_j = 0$, hat die Kovariate keinen Einfluss.

3.2. Wald-Test

Der Wald-Test stellt das standardmäßige Verfahren zur Prüfung der statistischen Signifikanz einzelner Regressionskoeffizienten in logistischen Modellen dar. Er basiert auf der asymptotischen Normalverteilung der Maximum-Likelihood-Schätzer (Fahrmeir et al., 2009, S.205) und prüft die Nullhypothese $H_0 : \beta_j = 0 \quad \text{gegen} \quad H_1 : \beta_j \neq 0$. Die Wald-Statistik für einen einzelnen Regressionskoeffizienten berechnet sich als

$$z_j = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)}$$

(Fahrmeir et al., 2009, S.204-205), wobei $\hat{\beta}_j$ der geschätzte Regressionskoeffizient ist und $\mathrm{SE}(\hat{\beta}_j)$ den zugehörigen Standardfehler bezeichnet. Der Standardfehler ergibt sich als Quadratwurzel des j-ten Diagonalelements der geschätzten Kovarianzmatrix:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\left[\mathbf{F}^{-1}(\hat{\boldsymbol{\beta}})\right]_{jj}}.$$

Unter der Nullhypothese folgt die Teststatistik asymptotisch einer Standardnormalverteilung: $Z_j \sim \mathcal{N}(0, 1)$. Die Testentscheidung erfolgt entweder durch Vergleich mit dem kritischen Wert $z_{\alpha/2}$, oder durch Berechnung des zweiseitigen p-Werts: $p = 2 \cdot [1 - \Phi(|z_j|)]$, wobei $\Phi(\cdot)$ die Verteilungsfunktion der Standardnormalverteilung bezeichnet.

Der Wald-Test weist jedoch bei kleinen Stichprobenumfängen und extremen Koeffizientenwerten gewisse Schwächen auf. Insbesondere kann er konservativ werden, wenn die Koeffizienten stark von null abweichen, da die asymptotische Normalverteilung in solchen Fällen nur unzureichend approximiert wird.

4. Statistische Auswertung

Die initiale Untersuchung beschränkt sich auf Deutschland und Länder mit vergleichbarem Zugang zur Gesundheitsversorgung, darunter Österreich, Belgien, Tschechien und Frankreich (gemäß der Klassifizierung von Majcherek et al., 2024). Da jedoch Daten aus Tschechien im Datensatz nicht erfasst sind, wird der Analyseumfang auf die verbleibenden vier Länder eingegrenzt. Da zudem eine Interpolation durch Ersetzen fehlender Werte durch den Mittelwert oder Median für viele kategoriale Variablen nicht sinnvoll ist bzw. die Analyseergebnisse verzerren kann, werden Beobachtungen mit unvollständigen Daten ausgeschlossen. Nach dem Filterprozess umfasst der für die Analyse verwendete Datensatz die verbleibenden 6 895 Beobachtungen.

4.1. Deskriptive Statistik

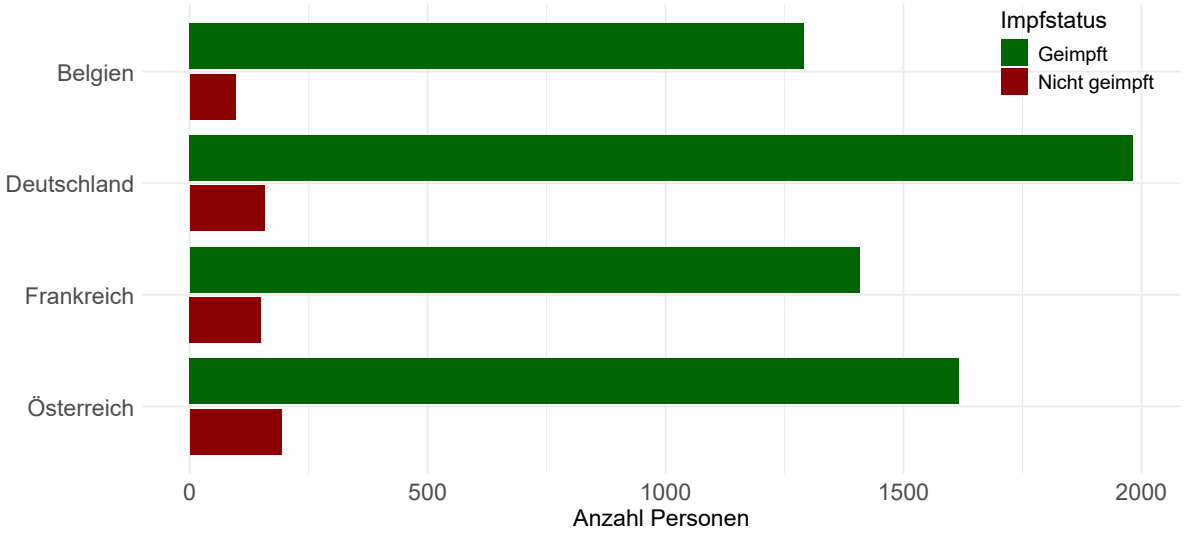

Abbildung 1: Balkendiagramm zum Impfstatus in Deutschland und Ländern mit vergleichbaren medizinischen Versorgungsbedingungen im Jahr 2023

Das Balkendiagramm in Abbildung 1 zeigt hohe Impfraten in allen vier untersuchten Ländern (vgl. die Impfdaten aus Tabelle 4). Belgien führt mit 93.1% (1 291 von 1 387 Personen), gefolgt von Deutschland mit 92.7% (1 983 von 2 140), Frankreich mit 90.4% (1 408 von 1 558) und Österreich mit 89.3% (1 617 von 1 810). Das Verhältnis zwischen geimpften und ungeimpften Personen bewegt sich zwischen 8.4/1 (Österreich) und 13.4/1 (Belgien), was auf eine hohe Vergleichbarkeit der Impfbereitschaft hinweist.

4.2. Modellentwicklung

Die Modellauswahl erfolgt durch einen erschöpfenden Vergleich der Modelle aller möglichen Kombinationen potenzieller Prädiktoren. Tabelle 2 zeigt die fünf Modelle mit den niedrigsten BIC-Werten und die in diesen Modellen vorkommenden Prädiktoren. Das beste Modell enthält lediglich Alter und den Vertrauensindex als erklärende Variablen. Auch in den übrigen Top-Modellen sind diese beiden Prädiktoren durchgehend enthalten. Im Gegensatz dazu erhöhen Variablen wie Haushaltsgröße, Haushaltseinkommen bzw. Geschlecht einerseits den BIC-Wert des Modells. Andererseits deutet ihr inkonsistentes Auftreten in den Top-Modellen darauf hin, dass es sich um instabile Prädiktoren handelt. Die Erklärungskraft dieser Variablen ist schwach und hängt stark von der Modellkonfiguration ab.

Tabelle 2: Modellselektion nach BIC

Modell	**Prädiktoren**					**BIC**
	`Alter`	`VI`	`HG`	`Geschlecht`	`Einkommen`	
1.	+	+				3 832.3
2.	+	+	+			3 839.7
3.	+	+		+		3 839.8
4.	+	+			+	3 842.1
5.	+	+	+		+	3 845.8

VI = Vertrauensindex, HG = Haushaltsgröße (vgl. Tabelle 1)

In Boxplots für das Alter (Abbildung 2) sowie für den Vertrauensindex (Abbildung 3) lässt sich eine deutliche Überlappung der Wertebereiche zwischen der geimpften und der ungeimpften Gruppe beobachten, was bedeutet, dass die Zielvariable nicht perfekt von jeder der einzelnen Prädiktorvariablen getrennt ist. Das Streudiagramm (Abbildung 4) ermöglicht die Betrachtung sämtlicher Linearkombinationen von Alter und Vertrauensindex, wobei viele Kombinationen bei geimpften und ungeimpften Gruppen gleichzeitig vorkommen (schwarze Punkte). Die perfekte Trennung beim Modell mit dem kleinsten BIC-Wert ist somit ausgeschlossen. In diesem Modell sind die VIF-Werte von Alter und Vertrauensindex beide gleich 1.009. Dadurch kann das Problem der Multikollinearität auch beseitigt werden. Zusammenfassend wird das Modell mit den zwei Variablen Alter und Vertrauensindex als optimal und gültig angesehen. Für nachfolgende Interpretationen wird dieses Modell verwendet.

4.3. Modellinterpretation

Die Schätzung von Koeffizienten in einem Regressionsmodell bzw. das Testen von Hypothesen setzt eine ausreichend große Stichprobe voraus, damit eine asymptotische Normalverteilung der Schätzer angenommen werden kann. Der Umfang der vorliegenden Analyse umfasst 6 895 Beobachtungen, was als ausreichend erachtet wird, um ein valides Ergebnis zu gewährleisten.

Tabelle 3: Ergebnisse der logistischen Regression zum Einfluss von Alter und Vertrauen auf den Impfstatus

Schätzter	**Nullhypothese**	**Teststatistik**	**p-Wert**	H_0
$\beta_0 = -0.07$	$H_0 : \beta_0 = 0$	-0.462	0.644	beibehalten
$\beta_{\text{Alter}} = 0.02$	$H_0 : \beta_{\text{Alter}} = 0$	8.226	<0.001	abgelehnt
$\beta_{\text{VI}} = 0.31$	$H_0 : \beta_{\text{VI}} = 0$	13.676	<0.001	abgelehnt

Gemäß den Ergebnissen in Tabelle 3 ist der Achsenabschnitt β_0 nicht signifikant ($p = 0.644$), was allerdings keine inhaltliche Relevanz hat, da er lediglich den Log-Odds-Wert repräsentiert, wenn das Alter und der Vertrauensindex beide den Wert Null annehmen. Dagegen zeigen sowohl das Alter ($\beta_{\text{Alter}} = 0.02$, $p < 0.001$) als auch der Vertrauensindex ($\beta_{\text{VI}} = 0.02$, $p < 0.001$) einen hochsignifikanten positiven Einfluss auf den Impfstatus. Im Rahmen der praktischen Interpretation werden die Exponentialwerte der geschätzten Koeffizienten oft im Zusammenhang Odds Ratio herangezogen. Im vorliegenden Modell beträgt dieser Wert $e^{0.02} \approx 1.02$ für das Alter, was die Odds Ratio, also des Verhältnisses zwischen der Wahrscheinlichkeit einer Impfung und der Wahrscheinlichkeit, nicht geimpft zu sein, um rund 2% je Lebensjahr erhöht. Pro eine Einheit höherer Vertrauensindex steigen die Odds Ratio um 36% ($e^{0.31} \approx 1.36$). Diese positiven Einflüsse dieser beiden Faktoren auf den Impfstatus spiegeln sich auch in den Boxplots wider. Sowohl beim Alter als auch beim Vertrauensindex sind die Medianwerte für die geimpfte Gruppe höher als für die ungeimpfte.

Mit einer Genauigkeit von 91.36% (siehe. Tabelle 6) wirkt das Modell zunächst gut. Die Konfusionsmatrix in Tabelle 5 zeigt jedoch kritische Schwächen des Modells - es kann ungeimpfte Personen nicht identifizieren. Die Sensitivität von 100% bei gleichzeitiger Spezifität von 0% offenbart, dass das Modell alle Fälle als "geimpft"klassifiziert. Das Modell zeigt somit keine praktische Vorhersagekraft für den Impfstatus.

5. Zusammenfassung

Ziel dieses Projekts war es, auf Basis der Daten der 11. Runde der European Social Survey (ESS ERIC, 2025) Faktoren zu identifizieren, die die Entscheidung für oder gegen eine COVID-19-Impfung beeinflussen. Die Analyse konzentriert sich auf Länder mit vergleichbarem Zugang zur Gesundheitsversorgung, um strukturelle Verzerrungen zu vermeiden.

Die Analyse des logistischen Regressionsmodells zur Erklärung des Impfstatus zeigt gemischte Ergebnisse. Das beste Modell (BIC = 3 832.3) enthält Alter und Vertrauensindex als Prädiktoren. Die Modellvalidierung bestätigt die Schätzbarkeit dieses Models - perfekte Separation und Multikollinearität sind ausgeschlossen. Mittels logistischer Regression werden Alter und ein aggregierter Vertrauensindex als signifikante Prädiktoren identifiziert. Höheres Alter und stärkeres Vertrauen in staatliche Institutionen erhöhen die Wahrscheinlichkeit, geimpft zu sein. Weitere Variablen wie Geschlecht, Haushaltsgröße oder Einkommen lieferten keine konsistente zusätzliche Erklärungskraft.

Jedoch offenbart die Güte-Bewertung kritische Schwächen: Das Modell klassifiziert alle Fälle als "geimpft"(Sensitivität 100%, Spezifität 0%), was zu einer täuschend hohen Accuracy von 91.36% führt, die der Baseline entspricht. Die AUC von 0.686 zeigt nur schwache Diskriminationsfähigkeit. Die einseitige Klassifikation des gewählten Modells kann auf eine zu unausgewogene Datenlage zurückzuführen sein. Bei unausgewogenen Daten orientiert sich das Modell häufig an der Gruppe der Mehrheit, um die Wahrscheinlichkeit auf Grundlage der beobachteten Daten zu maximieren.

Insgesamt bestätigt die Analyse bekannte Muster der Impfbereitschaft – insbesondere den Zusammenhang mit Alter und Vertrauen. Künftige Analysen sollten gezielt Maßnahmen zur Adressierung unausgeglichener Zielvariablen einbeziehen (z.B. Rebalancing-Techniken) und idealerweise um Informationen zum Zugang zur Gesundheitsversorgung ergänzt werden, um kontextuelle Faktoren besser zu berücksichtigen.

Literatur

Bartoń, Kamil. *MuMIn: Multi-Model Inference*, 2025. URL `https://CRAN.R-project.org/package=MuMIn`. R package version 1.48.11.

Betsch, C., Wieler, L. H., und Habersaat, K. Monitoring behavioural insights related to COVID-19. *The Lancet*, 395(10232):1255–1256, 2020.

European Social Survey European Research Infrastructure Consortium (ESS ERIC), . ESS Round 11 - 2023: Social Inequalities in Health, Gender in Contemporary Europe, 2025. URL `https://doi.org/10.21338/ess11-2023`.

Fahrmeir, L., Kneib, T., und Lang, S. Generalisierte lineare modelle. In Fahrmeir, L., Kneib, T., und Lang, S., editors, *Regression. Statistik und ihre Anwendungen*, pages 145–222. Springer, Berlin, Heidelberg, 2009.

Kuhn, und Max, . Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008. doi: 10.18637/jss.v028.i05. URL `https://www.jstatsoft.org/index.php/jss/article/view/v028i05`.

Lazarus, J. V., Ratzan, S. C., Palayew, A., Gostin, L. O., Larson, H. J., Rabin, K., Kimball, S., und El-Mohandes, A. A global survey of potential acceptance of a COVID-19 vaccine. *Nature Medicine*, 27(2):225–228, 2021.

Majcherek, D., Hegerty, S. W., Kowalski, A. M., Lewandowska, M. S., und Dikova, D. Opportunities for healthcare digitalization in Europe: Comparative analysis of inequalities in access to medical services. *Health Policy*, 139, 2024.

Silvapulle, M. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(3):310–313, 1981.

Team, R Core. R: A language and environment for statistical computing, 2024. URL `https://www.R-project.org/`.

Toutenburg, H. *Lineare Modelle: Theorie und Anwendung*. Springer-Verlag, Heidelberg, 2. edition, 2003. neu bearbeitete und erweiterte AUflage.

Wickham, H., Navarro, D., und Pedersen, T. L. *ggplot2: Elegant Graphics for Data Analysis (3e)*. Springer-Verlag New York, work-in-progress 3rd edition, 2016. URL `https://ggplot2-book.org/`.

A. Zusätzliche Tabellen

Tabelle 4: Anzahl der geimpften und ungeimpften Personen in Deutschland und Ländern mit vergleichbaren medizinischen Versorgungsbedingungen im Jahr 2023

Land	**Geimpft**	**Nicht geimpft**	**Impfrate**
Belgien	1 291	96	93.1%
Deutschland	1 983	157	92.7%
Frankreich	1 408	150	90.4%
Österreich	1 617	193	89.3%

Tabelle 5: Konfusionsmatrix des finalen logistischen Modells

		Tatsächlicher Impfstatus	
		Nicht geimpft	**Geimpft**
Vorhergesagt	**Nicht geimpft**	0	0
	Geimpft	596	6299

Tabelle 6: Modellgüte-Kennzahlen des finalen logistischen Modells

Kennzahl	**Wert**
Genauigkeit	0.9136
Sensitivität	1.0000
Spezifität	0.0000

Genauigkeit: Anteil aller korrekt klassifizierten Fälle
Sensitivität: Anteil korrekt erkannter positiver Fälle
Spezifität: Anteil korrekt erkannter negativer Fälle

B. Zusätzliche Grafiken

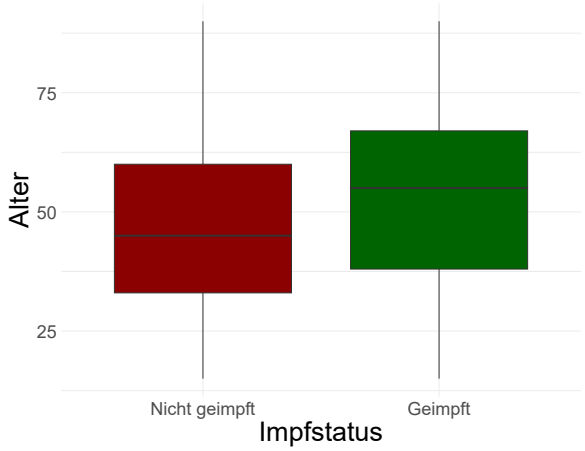

Abbildung 2: Boxplot für das Alter der Befragten

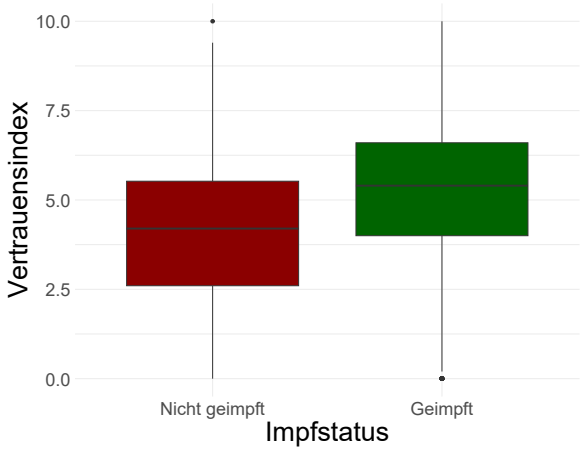

Abbildung 3: Boxplot für den Vertrauensindex der Befragten

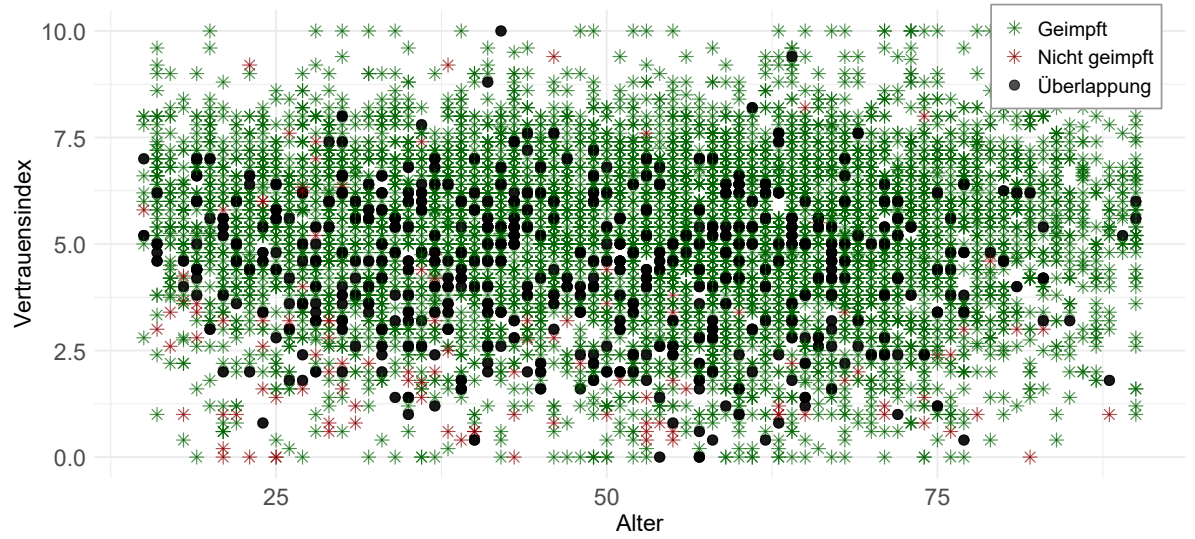

Abbildung 4: Streudiagramm der Prädiktoren Alter und Vertrauensindex nach Impfstatus