TU Dortmund

Fallstudien I - Sommersemester 2025

Projekt 4: Ein lineares Modell für die mittlere Lebenserwartung in NRW

Dozenten:

JProf. Dr. Dennis Dobler

Loreen Sabel

Dr. Susanne Frick

Prof. Dr. Philipp Doebler

Verfasser: Quang Vinh Nguyen

Gruppenmitglieder:
Alain Kauth, Ibrahim Rahman

22. Juni 2025

Inhaltsverzeichnis

1. Einleitung

Ein langes Leben zu führen zählt für viele Menschen zu den größten Wünschen. Die mittlere Lebenserwartung gilt dabei als ein zentraler Indikator für die Gesundheit und Lebensqualität einer Bevölkerung. Entsprechend erfahren jene Faktoren, die die menschliche Lebenserwartung beeinflussen können, seit jeher große Aufmerksamkeit – sowohl in der Öffentlichkeit als auch in der wissenschaftlichen Gemeinschaft. Rau und Schmertmann (2020) zeigten, dass ökonomische Faktoren, die sich auf die Armen konzentrieren, wie etwa die Arbeitslosenquote, in engerer Relation zur Lebenserwartung stehen als Makrofaktoren wie Bevölkerungsdichte oder Durchschnittseinkommen.

Im Rahmen des vorliegenden Projekts werden zwei lineare Modelle zur Erklärung der mittleren Lebenserwartung für Männer und für Frauen in nordrhein-westfälischen (NRW) Kreisen entwickelt. Als Kandidaten-Prädiktoren werden drei Prädiktoren untersucht: der German Index of Socioeconomic Deprivation (GISD) als Maß für sozioökonomische Benachteiligung (Michalski et al., 2024), die Siedlungsfläche pro Kopf als Indikator für individuelle Raumverfügbarkeit sowie der Anteil der Leistungsempfänger:innen als Ausdruck sozialer Bedürftigkeit. Unter Abwägung von Erklärungskraft und Modellkomplexität umfassen die finalen Modelle (sowohl für Männer als auch für Frauen) lediglich zwei Prädiktoren: den GISD-Score und die Siedlungsfläche pro Kopf, auch ohne Interaktionseffekte. Der Zusammenhang zwischen GISD-Score und Lebenserwartung ist jedoch deutlicher. Insgesamt zeigt sich ein negativer Einfluss sozioökonomischer Benachteiligung und ein positiver Effekt der Siedlungsfläche pro Kopf auf die Lebenserwartung. Die Richtung und Stärke der Effekte sind in beiden Modellen vergleichbar, fallen jedoch im Modell für Männer etwas ausgeprägter aus.

In Kapitel 2 wird der in diesem Projekt verwendete Datensatz erörtert, wobei ein besonderer Fokus auf die Variablen gelegt wird, die im linearen Modell berücksichtigt werden. Im Kapitel 3 wird die LOESS-Kurve als ein Werkzeug zur explorativen Untersuchung von Zusammenhängen zwischen Variablen vorgestellt. Des Weiteren werden das lineare Modell, die Schätzung der Modellparameter anhand der Kleinste-Quadrate-methode sowie Verfahren zur Reduktion der Modellkomplexität durch Hypothesentests erklärt. Darüber hinaus werden Modellbewertungsmethoden wie der Varianzinflationsfaktor und die Cooks Distanz behandelt. Kapitel 4 beschreibt den Modellentwicklungsprozess. Die wichtigsten Ergebnisse und eine ausführliche Diskussion finden sich in Kapitel 5.

2. Problemstellung

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

2.1. Datenmaterial

Die Analyse basiert auf einer Beobachtungsstudie unter Verwendung öffentlich zugänglicher Sekundärdaten. Die Lebenserwartungsschätzungen stammen aus der Studie von Rau und Schmertmann (2020) und der German Index of Socioeconomic Deprivation (GISD) aus der Veröffentlichung von Michalski et al. (2024). Weitere sozioökonomische und flächenbezogene Angaben werden dem Regionalstatistikportal der amtlichen Statistik (Stichtag 31.12.2016) entnommen.

Grundlage ist eine Vollerhebung aller 53 Kreise und kreisfreien Städte in NRW, wobei jede Beobachtungseinheit einem Kreis bzw. einer kreisfreien Stadt entspricht. Der zugrunde liegende Datensatz enthält verschiedene demografische und raumbezogene Merkmale. Zu den zentralen Variablen zählen der Kreisschlüssel und der Name des Kreises, der GISD-Score für 2016, die Gesamtbevölkerung ausgehend vom Zensus 2011, die Anzahl der Sozialleistungsempfänger:innen, sowie die Flächenangaben zu Gesamtfläche, Verkehrsfläche und Siedlungsfläche in Hektar. Darüber hinaus sind die bayesianische Punktschätzungen der Lebenserwartung bei Geburt für Frauen und Männer im Datensatz enthalten, jeweils zusammen mit anderen Variablen wie Rangwerten und Konfidenzgrenzen für die Lebenserwartung. In diesem Projekt werden jedoch ausschließlich Variablen berücksichtigt, die beim Erstellen linearer Modelle zum Einsatz kommen (siehe Tabelle 1). Dabei sind zwei Variablen nicht direkt im ursprünglichen Datensatz enthalten, sondern werden aus vorhandenen Größen berechnet. Der Leistungsempfängeranteil ergibt sich demnach aus der Anzahl der Sozialleistungsempfänger:innen im Verhältnis zur Gesamtbevölkerung. Die Siedlungsfläche pro Kopf wird ermittelt, indem die Siedlungsfläche durch die Gesamtbevölkerung dividiert wird.

Fehlende Werte treten im Datensatz nicht auf. Da die Informationen aus amtlichen Quellen stammen, ist von einer hohen Datenqualität und Messgenauigkeit auszugehen. Der GISD als zusammengesetzter Index bringt naturgemäß eine gewisse Unsicherheit durch Indikatorauswahl und Gewichtung mit sich, bleibt aber ein etabliertes Maß zur regionalen Ungleichheitsmessung.

Tabelle 1: Übersicht über die Modellvariablen

Variable	Bedeutung	Skalenniveau	Ausprägungen
Zielvariablen (unabhängige Variablen)			
`Le_F`	Punktschätzung der Lebenserwartung für Frauen	verhältnisskaliert	positive Zahlen
`Le_M`	Punktschätzung der Lebenserwartung für Männer	verhältnisskaliert	positive Zahlen
potenzielle Prädiktoren (abhängige Variablen)			
`GISD`	German Index of Socioeconomic Deprivation	verhältnisskaliert	[0, 1]
`LEA`	Leistungsempfängeranteil	verhältnisskaliert	[0, 1]
`SFK`	Siedlungsfläche pro Kopf (ha/Person)	verhältnisskaliert	positive Zahlen

2.2. Ziele des Projekts

Das Ziel dieses Projekts besteht in der Entwicklung linearer Modelle, die die regionale mittlere Lebenserwartung in NRW-Kreisen durch gleichzeitige Berücksichtigung mehrerer unabhängiger Variablen angemessen abbilden. Als Zielvariablen des linearen Modells werden die bayesianischen Punktschätzungen der Lebenserwartungen bei Geburt herangezogen, da sie eine zentrale und vergleichbare Maßzahl für die Gesundheitslage einer Bevölkerung darstellen. Alternativen wie Rangwerte oder Intervallgrenzen sind davon abgeleitet und daher weniger geeignet. Die Modellierung erfolgt getrennt für Männer und Frauen, da Unterschiede zwischen den Geschlechtern nicht vernachlässigt werden sollten. Für potenzielle Prädiktoren stehen drei erklärende Variablen im Fokus. Die Auswahl orientiert sich an theoretischen Überlegungen zu sozialer Ungleichheit und Umweltfaktoren. Der GISD-Score dient als Maß für die sozioökonomische Lage einer Region. Der Anteil der Leistungsempfänger:innen an der Gesamtbevölkerung kann als Indikator für soziale Bedürftigkeit interpretiert werden. Die Siedlungsfläche pro Kopf fungiert als Indikator für die individuelle Raumverfügbarkeit.

Die Daten werden zunächst explorativ mit Streudiagrammen und LOESS-Kurven analysiert, um lineare und nichtlineare Zusammenhänge zwischen den erklärenden Variablen und den Zielgrößen zu identifizieren. Zur besseren Modellierung werden die Variablen gegebenenfalls transformiert. Das lineare Modell wird schrittweise entwickelt, beginnend mit allen drei Prädiktoren und anschließender Selektion der wichtigsten Prädiktoren sowie Prüfung möglicher Interaktionen zwischen den verbleibenden Einflussgrößen. Modellvergleiche werden mittels Hypothesentests durchgeführt. Die Bewertung der Modelle sowie die Prüfungen zentraler Annahmen erfolgen mithilfe des Varianzinflationsfaktors, der Cooks Distanz und grafischer Diagnosen. Am Ende dient das finale Modell der Interpretation der Haupteffekte auf die regionale Lebenserwartung in den NRW-Kreisen.

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.

Obwohl das lineare Modell durch Dummy- bzw. Effektkodierung kategoriale Variablen verarbeiten kann (Toutenburg, 2003, S.71), ist die folgende Beschreibung aller Mothoden in diesem Kapitel auf den Fall kontinuierlicher abhängiger und unabhängiger Variablen ausgerichtet.

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. Lineares Modell

In der Realität zeigt sich häufig, dass Zielvariablen nicht von einer einzigen unabhängigen Variable, sondern von mehreren davon abhängen. Zur Erklärung der Zielvariablen durch mehrere unabhängige Variablen wird in der Regel ein multiples lineares Modell (auch multiple Regression genannt) verwendet.

Die multiplen Beziehungen zwischen den n Zielvariablen $Y_i, i = 1, \ldots, n$ und k Prädiktoren (oder auch Regressoren genannt) $X_{i1}, \ldots, X_{ik}$ werden nach Rencher und Schaalje (S.150-151) modelliert als

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, \ldots, n$$

oder in Matrixform als

$$\underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{\boldsymbol{Y},\, n\times 1} = \underbrace{\begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix}}_{\mathbf{X},\, n\times(k+1)} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{\boldsymbol{\beta},\, (k+1)\times 1} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon},\, n\times 1}. \tag{1}$$

In Gleichung (1) werden der Vektor $\boldsymbol{Y}$ als Zielvektor und die Matrix $\mathbf{X}$ als Regressormatrix benannt. Die Regressormatrix $\mathbf{X}$ lässt sich gegebenenfalls durch zusätzliche Spalten für Interaktionen und höhere Potenzen erweitern, um komplexere Zusammenhänge abzubilden. Der Vektor $\boldsymbol{\beta}$ enthält die unbekannten Regressionskoeffizienten $\beta_0, \ldots, \beta_k$ und wird als Parametervektor bezeichnet. Da ein perfekter Parametervektor, der die Gleichung mit beobachteten Daten $\boldsymbol{y} = \mathbf{x}\boldsymbol{\beta}$ exakt erfüllt, nahezu nie existiert, werden die Abweichungen durch die Residuen $\varepsilon_1, \ldots, \varepsilon_n$ im Residuenvektor $\boldsymbol{\varepsilon}$ dargestellt.

Zur Schätzung des Parametervektors $\boldsymbol{\beta}$ wird in der Regel ein Vektor gesucht, der die Quadratsumme der Abweichungen $\hat{\varepsilon}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}, i = 1, \ldots, n$ minimiert. Der Kleinste-Quadrate-Schätzer $\hat{\boldsymbol{\beta}}_{KQ}$ ist eindeutig bestimmt, wenn $\operatorname{rank}(\mathbf{x}) = k+1$. Laut Rencher und Schaalje (S.142) gilt in diesem Fall

$$\hat{\boldsymbol{\beta}}_{KQ} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}\right)^2 = \left(\mathbf{x}^\top \mathbf{x}\right)^{-1} \mathbf{x}^\top \boldsymbol{y}. \quad (2)$$

Wenn die quadrierten Abweichungen mit bestimmten Gewichten $w_i > 0, i = 1, \ldots, n$ in der Parameterschätzung verbunden sind, kommt der gewichtete Kleinste-Quadrate-Schätzer $\hat{\boldsymbol{\beta}}_{GKQ}$ zum Einsatz. Der GKQ-Schätzer

$$\hat{\boldsymbol{\beta}}_{GKQ} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} w_i \left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}\right)^2 = \left(\mathbf{x}^\top \mathbf{w} \mathbf{x}\right)^{-1} \mathbf{x}^\top \mathbf{w} \boldsymbol{y} \quad (3)$$

verallgemeinert den klassischen KQ-Schätzer $\hat{\boldsymbol{\beta}}_{KQ}$, bei dem alle Gewichte gleich eins sind (Wells und Krakiwsky, 1971, S.96-97). Dabei ist $\mathbf{w} = \operatorname{diag}(w_1, \ldots, w_n)$ eine Diagonalmatrix mit den vorgegebenen Gewichten. Es ist zu beachten, dass nur wenn die Regressormatrix $\mathbf{x}$ den Rang $k+1$ hat, die beiden Matrizen $\mathbf{x}^\top \mathbf{x}$ in (2) und $\mathbf{x}^\top \mathbf{w} \mathbf{x}$ in (3) invertierbar sind und der Schätzer $\hat{\boldsymbol{\beta}}_{KQ}$ bzw. $\hat{\boldsymbol{\beta}}_{GKQ}$ die einzige Lösung für das entsprechende Minimierungsproblem ist.

Gemäß dem Gauß-Markov-Theorem weist der KQ-Schätzer unter allen unverzerrten linearen Schätzwerten eine minimale Varianz auf (engl. Best Linear Unbiased Estimator - BLUE), sofern die folgenden drei Annahmen erfüllt sind (Rencher und Schaalje, 2008, S.138,146). Erstens sollen die Fehlerterme im Mittel null sein, also $\mathbb{E}(\varepsilon_i) = 0$ für alle $i = 1, \ldots, n$. Zweitens wird vorausgesetzt, dass sie eine konstante Varianz aufweisen, d. h. $\operatorname{Var}(\varepsilon_i) = \sigma^2$ für alle i. Drittens müssen die Fehlerterme unkorreliert sein, also $\operatorname{Cov}(\varepsilon_i, \varepsilon_j) = 0$ für alle $i \neq j$.

3.2. LOESS-Kurve

LOESS (Locally Estimated Scatterplot Smoothing) ist ein nichtparametrisches Verfahren zur Darstellung von Zusammenhängen zwischen Variablen, auch wenn diese nicht linear sind. Bei zwei Variablen erzeugt LOESS eine glatte Kurve (sog. LOESS-Kurve), die zusammen mit den Datenpunkten in einem Streudiagramm visualisiert wird.

Im Folgenden wird die Erstellung der LOESS-Kurve skizziert. Eine ausführliche Darstellung mit Beispielen findet sich bei Jacoby (siehe S.608-611). Gegeben seien n Datenpunkte $(x_1, y_1), \ldots, (x_n, y_n)$. Um den Zielwert y^* an einer beliebigen Stelle x^* zu schätzen, wird im LOESS-Verfahren eine lokale gewichtete Regression durchgeführt. Dabei erhalten Datenpunkte, die näher an x^* liegen, ein höheres Gewicht. Zunächst werden zwei Parameter festgelegt: der Glättungsparameter $\delta \in (0, 1]$, der den Anteil der für die lokale Regression genutzten Datenpunkte bestimmt, sowie der Polynomgrad $\lambda \in \mathbb{N}$, der die Komplexität der lokalen Anpassung steuert. Sei d_{lim} die Distanz zwischen x^* und dem $\lfloor \delta n \rfloor$-nächstgelegenen Punkt, ergibt sich das Gewicht für den Datenpunkt $(x_i, y_i), i = 1, \ldots, n$ als $w_i = W\left(\frac{|x_i - x^*|}{d_{\text{lim}}}\right)$, wobei die Tricube-Gewichtsfunktion definiert ist durch $W(u) = (1 - u^3)^3$ für $0 \leq u < 1$ und $W(u) = 0$ für $u \geq 1$. Die Regressormatrix $\mathbf{x}$ und die Gewichtsmatrix $\mathbf{w}$ werden formuliert als

$$\mathbf{x} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^\lambda \\ 1 & x_2 & x_2^2 & \cdots & x_2^\lambda \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^\lambda \end{bmatrix} \quad \text{und} \quad \mathbf{w} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}.$$

Nach (3) lautet der Schätzer der Koeffizienten $\hat{\boldsymbol{\beta}}_{GKQ} = \left(\mathbf{x}^\top \mathbf{w} \mathbf{x}\right)^{-1} \mathbf{x}^\top \mathbf{w} \mathbf{y}$. Der zugehörige Funktionswert an $x*$ wird durch Auswertung des lokal angepassten Polynoms berechnet, also $\hat{y}^* = \begin{bmatrix} 1 & x^* & (x^*)^2 & \cdots & (x^*)^\lambda \end{bmatrix} \hat{\boldsymbol{\beta}}_{GKQ}$. Die LOESS-Kurve entsteht, indem lokal geschätzte Werte an vielen eng beieinanderliegenden Stellen berechnet und im Scatterplot zu einer glatten Linie verbunden werden. Die Anzahl der Auswertungspunkte ist unerheblich, sofern die Variabilität der Zielvariable Y hinreichend dargestellt wird (Jacoby, 2000, S.583).

Die LOESS-Kurve unterstützt das visuelle Erkennen und Schätzen möglicher Zusammenhänge zwischen Variablen. Ihre Form hängt jedoch von den Parametern ab: Kleine δ- und hohe λ-Werte liefern detailreiche, aber empfindliche Kurven; große δ- und niedrige λ-Werte erzeugen glattere, robustere Verläufe.

3.3. Hypothesentest für einen Koeffizienten

Beim Aufbau eines linearen Modells wird angestrebt, die Zielvariable möglichst genau zu schätzen und gleichzeitig ein Modell mit wenigen, aber relevanten Prädiktoren zu verwenden. Hypothesentests ermöglichen den formalen Vergleich eines vollständigen mit einem reduzierten Modell zur Auswahl eines geeigneten Modells.

Um ein exaktes Ergebnis des Hypothesentests für einen Regressionskoeffizienten zu erreichen, werden die Annahmen des Gauß-Markov-Theorems um die Normalverteilung der Residuen erweitert. Alle Annahmen lassen sich als $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ zusammenfassen (vgl. die Darstellung von Rencher und Schaalje S.185). Ein Hypothesentest für einen Koeffizienten $\beta_j, j = 1, \ldots, k$ im Parametervektor $\boldsymbol{\beta}$ prüft, ob dieser signifikant von null abweicht. Demnach wird die Nullhypothese $H_0 : \beta_j = 0$ gegen die Alternativhypothese $H_1 : \beta_j \neq 0$ getestet. Unter H_0 ergibt sich folgende Teststatistik mit zugehöriger Verteilung (Rencher und Schaalje, 2008, S.205):

$$T = \frac{\hat{\beta}_j}{\sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \cdot \left[(\mathbf{X}^\top\mathbf{X})^{-1}\right]_{jj}}} \sim t_{n-k-1},$$

wobei $\hat{\beta}_j$ das j-te Element des KQ-Schätzers $\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}$ aus Gleichung (2), $\left[(\mathbf{X}^\top\mathbf{X})^{-1}\right]_{jj}$ das j-te Diagonalelement der Inversen des Produkts $\mathbf{X}^\top\mathbf{X}$ und t_{n-k-1} die t-Verteilung mit $n-k-1$ Freiheitsgraden bezeichnet. Die Nullhypothese H_0 wird verworfen, wenn der p-Wert kleiner als das vorgegebene Signifikanzniveau α ist, also $2 \cdot \mathbb{P}(T \geq |t|) < \alpha$.

3.4. Modellbewertung

Güte der Anpassung: Das Bestimmtheitsmaß R^2 gibt an, wie gut ein Regressionsmodell die Streuung der abhängigen Variable erklärt. Es wird berechnet mit der Formel $R^2 = 1 - \frac{SS_{\mathrm{res}}}{SS_{\mathrm{tot}}}$ (Toutenburg, 2003, S.145-146). Dabei ist $SS_{\mathrm{res}} = \sum_{i=1}^{n}\varepsilon_i^2$ die Quadratsumme der Abweichungen und $SS_{\mathrm{tot}} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ die Gesamtquadratsumme, also die Streuung der tatsächlichen Werte um ihren Mittelwert $\bar{y}$. Ein R^2-Wert von 1 bedeutet, dass das Modell die gesamte Streuung perfekt erklärt.

Multikollinearität: Multikollinearität (Toutenburg, 2003, S.112-114) bedeutet, dass die Information eines Prädiktors durch andere Prädiktoren erklärbar ist. Dies macht das lineare Modell instabil, da die Effekte nicht eindeutig zugeordnet werden können. Im Extremfall ist ein Prädiktor eine lineare Kombination anderer, die Designmatrix $\mathbf{x}$ hat

dann keinen vollen Rang, und die Produkte $\mathbf{x}^\top\mathbf{x}$ in (2) und $\mathbf{x}^\top\mathbf{w}\mathbf{x}$ in (3) sind nicht invertierbar. Für jeden Regressor $X_j, j = 1, \ldots, k$ wird VIF_i definiert als $\text{VIF}_j = \frac{1}{1-R_j^2}$ (Toutenburg, 2003, S.118-119) Dabei ist der multiplen Korrelationskoeffizient R_j^2 das Bestimmtheitsmaß von X_j auf die anderen Variablen. Ein VIF von 4 entspricht einem R_j^2-Wert von 0.75 und kann daher als Schwellenwert zur Erkennung von Multikollinearität verwendet werden.

Einflussreiche Beobachtungen: Einflussreiche Beobachtungen können die Schätzung eines Regressionsmodells erheblich verzerren. Cooks Distanz dient dazu, den Einfluss einzelner Beobachtungen auf das gesamte Regressionsmodell zu quantifizieren. Sie misst, wie stark sich die geschätzten Regressionskoeffizienten ändern würden, wenn eine bestimmte Beobachtung aus dem Datensatz entfernt wird. Laut Rencher und Schaalje (siehe S.237) lässt sich Cooks Distanz über den Unterschied zwischen den Vorhersagevektoren des Modells mit und ohne Beobachtung i in folgender Matrixform darstellen als $D_i = \frac{(\hat{\mathbf{y}}-\hat{\mathbf{y}}_{(i)})^\top(\hat{\mathbf{y}}-\hat{\mathbf{y}}_{(i)})}{(k+1)\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2}$. Hierbei steht $\hat{\mathbf{y}}$ für den Vektor der geschätzten Werte auf Basis des vollständigen Modells, und $\hat{\mathbf{y}}_{(i)}$ für den entsprechenden Vektor, der durch Schätzung ohne die i-te Beobachtung entsteht. Ein hoher Wert der Cooks Distanz weist darauf hin, dass die entsprechende Beobachtung einen starken Einfluss auf das Regressionsmodell hat. Als Faustregel gilt ein Wert größer als $\frac{4}{n-k-1}$ als potenziell auffällig, was eine genauere Prüfung der betreffenden Beobachtung nahelegt.

4. Statistische Auswertung

4.1. Deskriptive Statistik

Die deskriptive Analyse mittels Streudiagrammen mit LOESS-Kurven zeigt für beide Geschlechter einen klar negativen Zusammenhang zwischen dem GISD-Score und der Lebenserwartung (siehe Abbildung 1 und Abbildung 2). Die Siedlungsfläche pro Kopf steht tendenziell in positivem Zusammenhang mit der Lebenserwartung, wobei dieser bei Frauen leicht gekrümmt und weniger stabil erscheint, während er bei Männern flacher und gleichmäßiger verläuft. Beim Anteil der Leistungsempfänger:innen zeigt sich für Frauen ein schwacher, inkonsistenter negativer Trend, für Männer hingegen ein deutlicherer negativer Zusammenhang. Insgesamt zeigt sich ein klarer linearer Zusammenhang zwischen dem GISD-Score und der Lebenserwartung. Bei den beiden übrigen Prädikto-

ren ist der Zusammenhang nicht eindeutig linear, es lässt sich jedoch kein bestimmtes Muster erkennen.

4.2. Modellentwicklung

Für die Erstellung der linearen Modelle werden alle Prädiktoren standardisiert, um vergleichbare Regressionskoeffizienten zu erhalten. Darüber hinaus verhindert Standardisierung, dass einzelne Variablen bei der Analyse von Interaktionseffekten das Modell unverhältnismäßig dominieren

Tabelle 2: Hypothesentests zu den Modellen zur Erklärung der Lebenserwartung bei Frauen

Modell	**Nullhypothese**	**Teststatistik**	**p-Wert**
1	$H_0 : \beta_0 = 0$	1633.567	< 0.0001
1	$H_0 : \beta_{\text{GISD}} = 0$	-8.390	< 0.0001
1	$H_0 : \beta_{\text{SFK}} = 0$	1.713	0.093
1	$H_0 : \beta_{\text{LEA}} = 0$	-0.718	0.476
2	$H_0 : \beta_0 = 0$	1641.545	< 0.0001
2	$H_0 : \beta_{\text{GISD}} = 0$	-8.888	< 0.0001
2	$H_0 : \beta_{\text{SFK}} = 0$	2.153	0.036
3	$H_0 : \beta_0 = 0$	1613.915	< 0.0001
3	$H_0 : \beta_{\text{GISD}} = 0$	-6.253	< 0.0001
3	$H_0 : \beta_{\text{SFK}} = 0$	2.086	0.042
3	$H_0 : \beta_{\text{Int}} = 0$	-0.359	0.721

Die Tabelle 2 dokumentiert den schrittweisen Modellentwicklungsprozess für die weibliche Lebenserwartung mit standardisierten Prädiktoren. Das Vollmodell (Modell 1) testet alle drei Variablen, wobei nur der GISD-Score signifikant ist, während Siedlungsflächenanteil und Leistungsempfängeranteil nicht signifikant sind. Modell 2 eliminiert den nicht-signifikanten Leistungsempfängeranteil und behält nur GISD-Score und Siedlungsflächenanteil bei, die beide nun Signifikanz erreichen. Modell 3 erweitert das reduzierte Modell um eine Interaktion zwischen GISD und Siedlungsflächenanteil, die sich jedoch als nicht signifikant erweist. Dieser Entwicklungsprozess identifiziert Modell 2 als optimal, da es nur signifikante Prädiktoren ohne überflüssige Komplexität enthält.

Auch die Entwicklung des Modells für Männer folgt einem ähnlichen Ansatz (siehe Tabelle 3): Das Vollmodell (Modell 1) mit allen drei Prädiktoren zeigt Signifikanz für GISD und SFK, während LEA nicht signifikant ist. Das reduzierte Modell (Modell 2) behält nur die signifikanten Variablen bei. Modell 3 erweitert das reduzierte Modell um eine GISD×SFK-Interaktion, die keine Signifikanz erreicht. Modell 2 stellt die optimale Balance zwischen Erklärungskraft und Sparsamkeit dar.

Tabelle 3: Hypothesentests zu den Modellen zur Erklärung der Lebenserwartung bei Männern

Modell	**Nullhypothese**	**Teststatistik**	**p-Wert**	H_0
1	$H_0 : \beta_0 = 0$	1259.577	<0.001	abgelehnt
	$H_0 : \beta_{\text{GISD}} = 0$	8.895	<0.001	abgelehnt
	$H_0 : \beta_{\text{SFK}} = 0$	2.453	0.018	abgelehnt
	$H_0 : \beta_{\text{LEA}} = 0$	0.397	0.693	beibehalten
2	$H_0 : \beta_0 = 0$	1270.327	<0.001	abgelehnt
	$H_0 : \beta_{\text{GISD}} = 0$	9.363	<0.001	abgelehnt
	$H_0 : \beta_{\text{SFK}} = 0$	2.836	0.007	abgelehnt
3	$H_0 : \beta_0 = 0$	1249.366	<0.001	abgelehnt
	$H_0 : \beta_{\text{GISD}} = 0$	6.023	<0.001	abgelehnt
	$H_0 : \beta_{\text{SFK}} = 0$	2.838	0.007	abgelehnt
	$H_0 : \beta_{\text{Int}} = 0$	0.394	0.695	beibehalten

4.3. Modellinterpretation

Der Vergleich der additiven Modelle zur Erklärung der mittleren Lebenserwartung zeigt, dass die Richtung und Stärke der Effekte in beiden Geschlechtern ähnlich sind. In beiden Modellen wirkt sich ein höherer GISD-Score negativ auf die Lebenserwartung aus, während eine größere Siedlungsfläche pro Kopf positiv korreliert ist. Der Effekt der sozioökonomischen Benachteiligung fällt bei Männern etwas stärker aus (–0.585 gegenüber –0.458), ebenso der Einfluss der Siedlungsfläche (0.177 gegenüber 0.111). Auch das Bestimmtheitsmaß R2R2̂R2 liegt bei Männern etwas höher (0.669) als bei Frauen (0.637), was auf eine geringfügig bessere Modellanpassung hindeutet. Die Varianzinflationsfaktoren (VIF) liegen in beiden Fällen bei rund 1.008 und deuten auf keine relevante Multikollinearität hin. Insgesamt zeigen sich weitgehend konsistente Muster, wobei der Einfluss der Prädiktoren im Modell für Männer etwas ausgeprägter ausfällt.

Tabelle 4: Vergleich der additiven Modelle zur Erklärung der Lebenserwartung

	Frauen	**Männer**
Koeffizienten		
Intercept	83.436	78.268
GISD-Score	-0.458	-0.585
Siedlungsfläche pro Kopf	0.111	0.177
Modellgüte		
R^2	0.637	0.669
VIF-Werte		
GISD-Score	1.008	1.008
Siedlungsfläche pro Kopf	1.008	1.008

5. Zusammenfassung

Ziel dieses Projekts war es, mithilfe linearer Modelle zu untersuchen, welche sozioökonomischen und raumbezogenen Faktoren mit der mittleren Lebenserwartung in nordrhein-westfälischen Kreisen (NRW) in Zusammenhang stehen. Grundlage der Analyse war ein Datensatz mit öffentlich zugänglichen Indikatoren, insbesondere dem GISD-Score zur Messung sozioökonomischer Benachteiligung, der Siedlungsfläche pro Kopf und dem Anteil der Sozialleistungsempfänger:innen. Die Lebenserwartung bei Geburt – getrennt für Männer und Frauen – diente als Zielgröße.

Die Analyse zeigte in beiden Modellen (für Männer und Frauen), dass der GISD-Score ein starker negativer Prädiktor ist, während die Siedlungsfläche pro Kopf positiv mit der Lebenserwartung assoziiert ist. Der Anteil der Leistungsempfänger:innen wies keinen signifikanten zusätzlichen Effekt auf. Interaktionstests zwischen GISD-Score und Siedlungsfläche pro Kopf ergaben keine signifikanten Wechselwirkungen, weshalb auf einen Interaktionsterm im finalen Modell verzichtet wurde.

Die Ergebnisse deuten darauf hin, dass sozioökonomische Benachteiligung auf regionaler Ebene stark mit geringerer Lebenserwartung einhergeht. Dies bestätigt frühere Studien und legt nahe, dass Maßnahmen zur Verringerung sozialer Ungleichheit gesundheitliche Effekte haben könnten. Zugleich sollte bei der Interpretation beachtet werden, dass es sich um eine Beobachtungsstudie handelt; kausale Schlüsse sind daher nicht möglich. Zudem besteht ein Risiko der Scheinkorrelation durch unbeobachtete Einflussgrößen.

Für zukünftige Untersuchungen wären Analysen mit längsschnittlichen Daten oder feineren regionalen Auflösungen wünschenswert. Auch der Einbezug weiterer sozialer, infrastruktureller oder umweltbezogener Indikatoren könnte helfen, ein differenzierteres Bild von regionalen Einflussfaktoren auf die Lebenserwartung zu zeichnen.

Literatur

Jacoby, W. G. Loess: A Nonparametric, Graphical Tool for Depicting Relationships Between Variables. *Electoral Studies*, 19(4):577–613, 2000.

Michalski, N., Soliman, O., Reis, M., Tetzlaff, F., Nowossadeck, E., und Hoebel, J. German Index of Socioeconomic Deprivation (GISD), 2024. Berlin.

Rau, R. und Schmertmann, C. P. Lebenserwartung auf Kreisebene in Deutschland. *Dtsch Arztebl Int*, 117:493–499, 2020.

Regionalstatistikportal, . URL `https://www.regionalstatistik.de`. (Codes 12411, 22121, 33111).

Rencher, A. C. und Schaalje, G. B. *Linear Models in Statistics*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2nd edition, 2008. ISBN 978-0-471-75498-5.

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.

Wells, D. E. und Krakiwsky, E. J. *The Method of Least Squares*. Department of Geodesy and Geomatics Engineering, University of New Brunswick, Fredericton, N.B., Canada, 1971. Latest reprinting February 1997.

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/`.

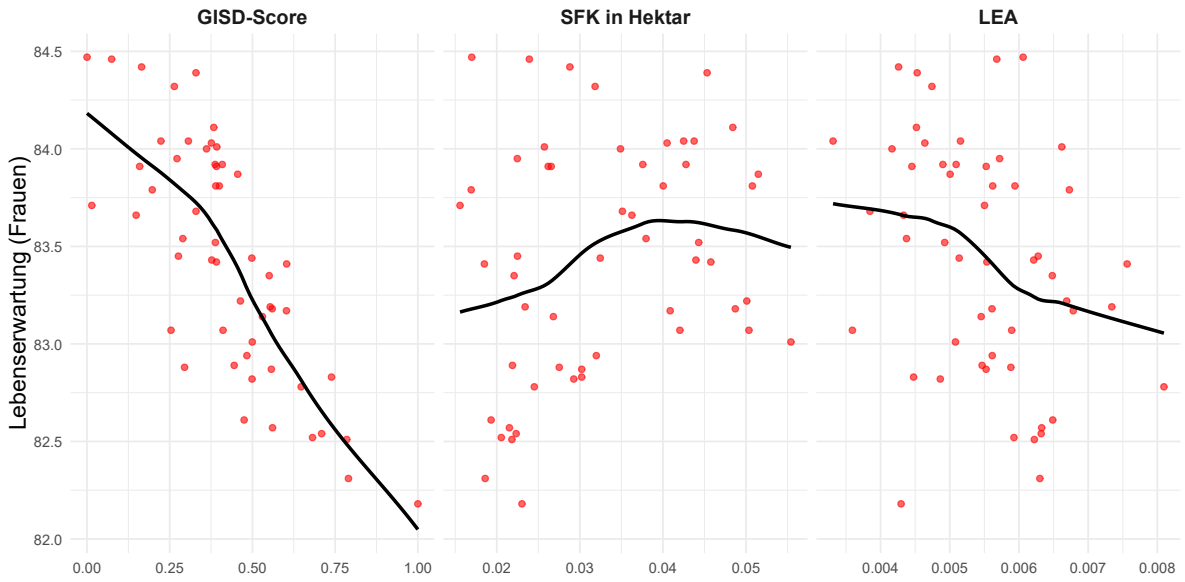

Abbildung 1: Zusammenhang zwischen erklärenden Variablen und Lebenserwartung – Frauen

A. Zusätzliche Grafiken

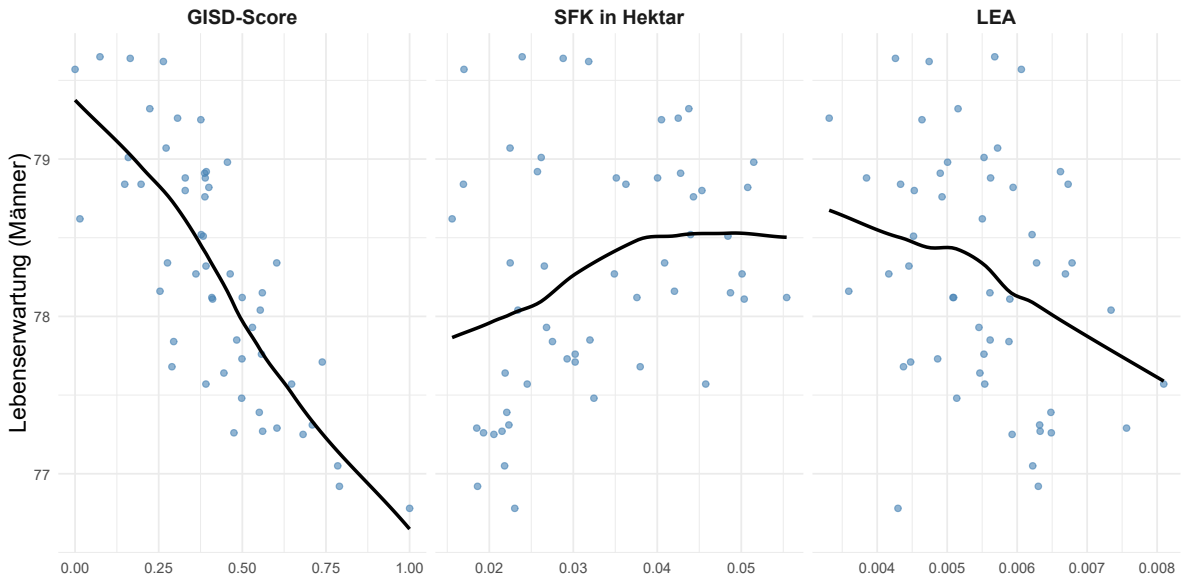

Abbildung 2: Zusammenhang zwischen erklärenden Variablen und Lebenserwartung – Männer

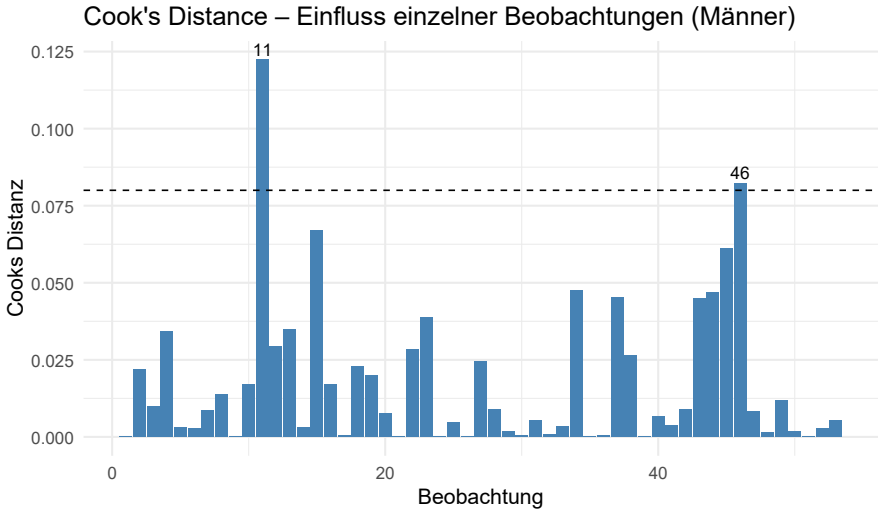

Abbildung 3: Enter Caption

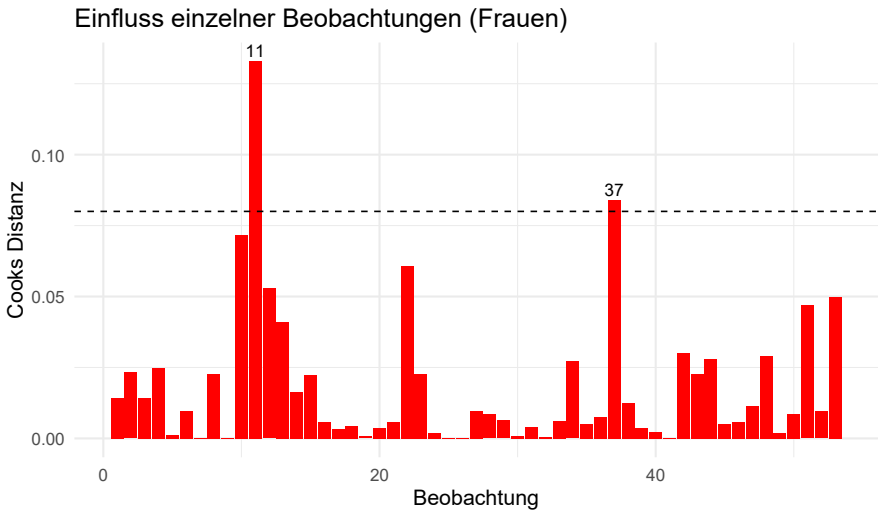

Abbildung 4: Einfluss einzelner