

2 Regression

Der einfachste Ansatz, die in Abschnitt 1 beschriebene Problematik, mathematisch zu betrachten ist die lineare Regression. Dabei wird eine oder mehrere abhängige Variable von einer Gruppe unabhängiger Variablen beschrieben. Der Zusammenhang zwischen den abhängigen und unabhängigen Variablen wird dabei als linear vorausgesetzt.

2.1 Deskriptive lineare Regression

Zunächst klammern wir die stochastische Natur der Daten vollständig aus. Die dabei entwickelten Konzepte können in Folge auch in den aufbauenden stochastischen Modellen verwendet werden. Wir nehmen an, wir kennen T Beobachtungen von y und von x_1, \dots, x_k und wollen y über den Ansatz

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (2.1)$$

möglichst gut beschreiben. Dabei gilt die Notation $\mathbf{y} = (y_1, \dots, y_T)' \in \mathbb{R}^T$, $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_k) \in \mathbb{R}^{T \times (k+1)}$, $\mathbf{x}_i = (x_{i1}, \dots, x_{iT_i})' \in \mathbb{R}^T$, $\mathbf{1} = (1, \dots, 1)$ und $\mathbf{u} = (u_1, \dots, u_T)' \in \mathbb{R}^T$.

Die Voraussetzung, einen linearen Ansatz zu verwenden, ist nicht so gravierend wie es im ersten Moment erscheint. Transformationen von \mathbf{X} ermöglichen die Betrachtung komplexerer Modelle durch Lösen der linearen Regression. Im Fall eines polynomialen Trends vom Grad p wird die Matrix auf die Dimension $\mathbb{R}^{T \times p \cdot k + 1}$ erweitert und erhält folgende Gestalt.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{11}^2 & \dots & x_{11}^p & x_{12} & \dots & x_{1k}^p \\ \vdots & \vdots & \ddots & & \ddots & \ddots & & \vdots \\ \vdots & \vdots & \ddots & & \ddots & \ddots & & \vdots \\ 1 & x_{T1} & x_{T1}^2 & \dots & x_{T1}^p & x_{T2} & \dots & x_{Tk}^p \end{pmatrix}$$

Die Normalgleichungen

Im Rahmen der deskriptiven linearen Regression sollen Parameter $\boldsymbol{\beta}$ gewählt werden, sodass $\mathbf{X}\boldsymbol{\beta}$ möglichst gut \mathbf{y} beschreibt. Die Qualität der Wahl der $\boldsymbol{\beta}$ soll im Allgemeinen über die Methode der kleinsten Quadrate bestimmt werden.

$$S(\boldsymbol{\beta}) = \mathbf{u}'\mathbf{u} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (2.2)$$

$\boldsymbol{\beta}$ ist also optimal gewählt, wenn es $S(\boldsymbol{\beta})$ minimiert.

$$\begin{aligned}
S(\beta) &= \mathbf{u}'\mathbf{u} \\
&= \mathbf{y}'\mathbf{y} - \beta' \mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\beta + \beta' \mathbf{X}'\mathbf{X}\beta \\
&= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\beta + \beta' \mathbf{X}'\mathbf{X}\beta \\
\frac{\delta S(\beta)}{\delta \beta} &= 0 - 2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta = 0
\end{aligned}$$

Durch einfache Umformungen und Nullsetzen der Ableitung ergeben sich aus der letzten Zeile direkt die Normalgleichungen.

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y} \quad (2.3)$$

Die obigen Umformungen besagen lediglich, dass ein β , das die Normalgleichungen erfüllt, $S(\beta)$ minimiert. In Schönfeld (1969) wird auf die Lösungen der Gleichungen inklusive ausgiebiger Beweisführung genau eingegangen. Hier sollen nur die wichtigsten Ergebnisse zusammengefasst werden. Zunächst sind die Normalgleichungen immer lösbar, da $\mathbf{X}'\mathbf{y} \in \text{im}(\mathbf{X}'\mathbf{X})$ gilt. Es gilt auch, dass $\hat{\beta}$ genau dann eine Lösung der Normalgleichungen ist, wenn es die Funktion $S(\beta)$ minimiert. Ist $\mathbf{X}'\mathbf{X}$ regulär so ist $\hat{\beta}$ eindeutig durch $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ bestimmt.

2.1.1 Geometrische Interpretation

Wir definieren die Kleinsten-Quadrate-Residuen $\hat{\mathbf{u}}$ als $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta}$. Diese Residuen erfüllen zwei grundlegende Orthogonalitätsrelationen. Es gelten

$$\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0} \text{ und } \hat{\mathbf{y}}'\hat{\mathbf{u}} = 0.$$

Die erste Relation folgt aus den Normalgleichungen wie aus den Umformungen $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{X}'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\beta}$ sofort zu sehen ist.

Jedes Element \mathbf{y} aus \mathbb{R}^T lässt sich also darstellen als $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{u}}$. Dabei stammt $\hat{\mathbf{y}}$ offensichtlich aus dem Spaltenraum von \mathbf{X} und $\hat{\mathbf{u}}$ aus seinem orthogonalem Komplement. Die Regression kann deshalb als eine Projektion von \mathbf{y} auf $\text{sp}(\mathbf{X})$ interpretiert werden. Die auftretenden Projektoren in unserem Modell sind $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ als Projektion von \mathbb{R}^T auf $\text{sp}(\mathbf{X})$ und \mathbf{y} auf $\hat{\mathbf{y}}$. Also muss $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ eine Projektion auf das orthogonale Komplement des von \mathbf{X} erzeugten Spaltenraumes sein.

2.1.2 Bestimmtheitsmaß

Durch elementare Umformungen unserer quadratischen Fehlersumme ist erkennbar, dass die Quadratsumme der Fehler $S(\hat{\beta}) = \hat{\mathbf{u}}'\hat{\mathbf{u}}$ darstellen lässt als Differenz der Quadratsumme der Beobachtungswerte von \mathbf{y} und der Quadratsumme der Projektion von \mathbf{y} auf $\text{sp}(\mathbf{X})$. Das ist gleichwertig zu der Aussage, dass sich die Variation der Beobachtungswerte aus der Summe der Variation der Projektionen $\hat{\mathbf{y}}$ und der Residuen zusammensetzt. Daher gibt das Verhältnis $\frac{\hat{\mathbf{y}}'\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{y}}$ jenen Anteil der Variation an, der durch $\hat{\mathbf{y}}$ bzw. innerhalb des Spaltenraums von \mathbf{X}

erklärt werden kann. Darüber hinaus kann durch Zentrieren von \mathbf{y} die Streuung eindeutig in erklärte und nicht erklärte Streuung zerlegt werden.

$$\mathbf{y}'\mathbf{y} - T\bar{y}^2 = \bar{\mathbf{y}}'\bar{\mathbf{y}} - T\bar{\bar{y}}^2, \quad (2.4)$$

$$s_y^2 = s_{\bar{y}}^2 + s_{\bar{\bar{y}}}^2 \quad (2.5)$$

Dabei ist $\bar{\bar{y}} = \frac{1}{T} \sum \hat{y}_i$ und $\bar{y} = \frac{1}{T} \sum y_i$.

Das am häufigsten Verwendete Qualitätsmaß einer linearen Regression, der multiple Korrelationskoeffizient R^2 ist nun definiert als $R^2 = \frac{s_{\bar{y}}^2}{s_y^2} = 1 - \frac{s_{\bar{\bar{y}}}^2}{s_y^2}$. Auf der Suche nach einem optimalen Modell wird die Qualität nicht nur über den Anteil der beschriebenen Streuung beschrieben sondern auch über die Übersichtlichkeit des Modells. Da eine Vergrößerung des $sp(\mathbf{X})$ aber immer einen größeren oder zumindest den gleichen Anteil von \mathbf{y} erklären wird, würde eine Entscheidung rein auf dem multiplen Korrelationskoeffizienten immer zu einem größtmöglichen Modell führen. Ein Ausweg wäre das korrigierte Bestimmtheitsmaß, das eine Strafe für größere Modelle einbezieht. Das korrigierte Bestimmtheitsmaß ist unter anderem in Backhaus, Erichson, Plinke und R.Weiber (2008) ausführlich erklärt. Wir gehen erst zu einem späteren Zeitpunkt auf die Modellentwicklung und Variablenselektion ein.

2.2 Multiple lineare Regression

Im letzten Kapitel wurde die stochastische Natur der Daten ausgeklammert. Wir gehen nun davon aus, dass das Modell $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ vorliegt und wollen gute, zumeist lineare Schätzer für $\boldsymbol{\beta}$ sowie die Varianzen des Modells herleiten. Strukturell und methodisch sind die Ausführungen dieses Abschnitts an Deistler (2002) angelehnt.

Im klassischen linearen Regressionsmodell werden folgende Annahmen getroffen:

(a1) \mathbf{X} ist nichtstochastisch

(a2) $\mathbf{X}'\mathbf{X}$ ist regulär

(b1) $E\mathbf{u} = \mathbf{0}$

(b2) $E\mathbf{u}\mathbf{u}' = \sigma^2\mathbf{I}$

(c) $(\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^k \times \mathbb{R}^+$

Die bedeutendste Einschränkung in diesem Modell ist (b2). Hier wird gefordert, dass die Fehler unkorreliert sind. Ist die Varianz-Kovarianzmatrix eine Diagonalmatrix mit konstanten Varianzen spricht man auch von Homoskedastie. Der Grund warum dieses eingeschränkte Modell betrachtet wird ist, dass Modelle mit heteroskedastischem Fehler günstig transformiert werden können.

2.2.1 Heteroskedastischer Fehler

Wird Bedingung (b2) des klassischen linearen Regressionsmodell abgeschwächt auf eine neue Bedingung

(b2*) $E\mathbf{u}\mathbf{u}' = \sigma^2\mathbf{\Omega}$, mit $\mathbf{\Omega}$ ist Varianz-Kovarianz-Matrix, σ^2 unbekannt, $\mathbf{\Omega} \in \mathbb{R}^{T \times T}$ bekannt,

so spricht man von einem verallgemeinerten linearen Regressionsmodell. Dabei muss zusätzlich noch eine Normierung von $\mathbf{\Omega}$ erfolgen um σ^2 eindeutig zu bestimmen. Da $\mathbf{\Omega}$ eine Varianz-Kovarianz-Matrix ist, ist $\mathbf{\Omega}$ symmetrisch und positiv definit. Der Beweis dafür ist in diversen Lehrbüchern, unter anderem in Putanen, Styan und Isotalo (2011) zu finden. Darüber hinaus wissen wir, dass für positive symmetrische Matrizen reguläre Matrizen \mathbf{R} und \mathbf{P} existieren, mit $\mathbf{\Omega} = \mathbf{R}\mathbf{R}'$ und $\mathbf{\Omega}^{-1} = \mathbf{P}'\mathbf{P}$. Solche Zerlegungen können etwa über die Spektralzerlegung $\mathbf{\Omega} = \mathbf{O}\mathbf{\Lambda}\mathbf{O}'$ gefunden werden, indem $\mathbf{R} = \mathbf{O}\sqrt{\mathbf{\Lambda}}$ definiert wird.

Das verallgemeinerte Regressionsmodell lässt sich nun über

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}^* \quad (2.6)$$

mit $\mathbf{y}^* = \mathbf{P}\mathbf{y}$, $\mathbf{X}^* = \mathbf{P}\mathbf{X}$ und $\mathbf{u}^* = \mathbf{P}\mathbf{u}$ in ein klassisches lineares Regressionsmodell transformieren. Die Eigenschaft (b2) wird wiederhergestellt, da $\mathbf{P}\mathbf{\Omega}\mathbf{P}' = \mathbf{P}\mathbf{R}\mathbf{R}'\mathbf{P}' = \mathbf{I}$ gilt, was direkt aus der Konstruktion von \mathbf{P} und \mathbf{R} folgt. Die restlichen Bedingungen sind trivialerweise erfüllt. Dass lineare Transformationen von Schätzern für $\boldsymbol{\beta}$ im transformierten Modell optimal bleiben zeigen wir im folgenden Abschnitt. Wir sehen also, dass es genügt, das klassische lineare Modell zu schätzen, vorausgesetzt wir kennen $\mathbf{\Omega}$ und können die Transformation durchführen. Da $\mathbf{\Omega}$ im Allgemeinen unbekannt ist, wird auch hier eine Schätzung notwendig sein.

2.2.2 Satz von Gauss Markov

Das Ziel dieses Abschnittes ist es, den besten linearen unverzerrten Schätzer für $\boldsymbol{\beta}$ zu finden. Der beste Schätzer wird hier als jener mit kleinster Varianz definiert.

Satz 1 (Gauss Markov). *Unter den Annahmen (a1), (a2), (b1), (b2) und (c) ist $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ der beste lineare unverzerrte Schätzer für $\boldsymbol{\beta}$.*

Der Beweis erfolgt in mehreren Schritten. Zunächst definieren wir die Klasse der unverzerrten linearen Schätzer $\tilde{\boldsymbol{\beta}} = \mathbf{D}\mathbf{y} + \mathbf{d}$. Aufgrund der Erwartungstreue wird verlangt, dass $\boldsymbol{\beta} = E(\mathbf{D}\mathbf{y} + \mathbf{d}) = \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{E}\mathbf{u} + \mathbf{d}$ für alle $\boldsymbol{\beta}$ gilt. Wählen wir $\boldsymbol{\beta} = \mathbf{0}$, so folgt aus (b1) direkt, dass der Schätzer $\tilde{\boldsymbol{\beta}}$ genau dann erwartungstreu ist, wenn $\mathbf{D}\mathbf{X} = \mathbf{I}$ und $\mathbf{d} = \mathbf{0}$ gelten. Um den besten Schätzer, jenen mit minimaler Varianz zu erhalten, muss $\Sigma_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}}$ minimiert werden im Sinne der folgenden Ordnungsrelation.

$$\mathbf{A} \leq \mathbf{B} \iff \mathbf{B} - \mathbf{A} \geq \mathbf{0}$$

Um diese Minimierung durchzuführen werden zwei Zwischenergebnisse benötigt.

Lemma 1. *Sei $\mathbf{z} = \mathbf{C}\mathbf{y} + \mathbf{c}$, wobei \mathbf{y} eine Zufallsvariable mit $E\mathbf{y}^2 = \Sigma_{yy}$ ist, so gilt $\Sigma_{zz} = \mathbf{C}\Sigma_{yy}\mathbf{C}'$.*

Lemma 2 (Zerlegungslemma). *Es gelte $\mathbf{C}\mathbf{X} = \mathbf{L}$ und \mathbf{X} habe vollen Spaltenrang, so lässt sich $\mathbf{C}\mathbf{C}'$ folgendermaßen zerlegen:*

$$\mathbf{C}\mathbf{C}' = \mathbf{L}\mathbf{X}^+(\mathbf{L}\mathbf{X}^+)' + (\mathbf{C} - \mathbf{L}\mathbf{X}^+)(\mathbf{C} - \mathbf{L}\mathbf{X}^+)' \text{ mit } \mathbf{X}^+ = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Lemma 1 wird durch einfaches Einsetzen und umformen gezeigt. Der Beweis wird etwa in Schmid und Trede (2006) durchgeführt. Das Zerlegungslemma folgt sofort durch Ausmultiplizieren der rechten Seite.

Aus $\tilde{\beta} = D\mathbf{y} + \mathbf{d}$ und Lemma 1 folgt $\Sigma_{\tilde{\beta}\tilde{\beta}} = D\Sigma_{yy}D' = \sigma^2 DD'$. $\Sigma_{\tilde{\beta}\tilde{\beta}}$ wird also genau dann minimiert, wenn DD' minimiert wird unter den Bedingungen $DX = I$ und $\mathbf{d} = \mathbf{0}$. Aus dem Zerlegungslemma folgt aber $DD' = X^+X^{+'} + (D - X^+)(D - X^+)'$. Der erste Teil ist konstant, der zweite Teil verschwindet genau dann, wenn $D = X^+$ gilt. Daraus folgt, der Schätzer $\tilde{\beta} = D\mathbf{y} + \mathbf{d}$ ist genau dann der beste lineare erwartungstreue Schätzer, wenn $D = X^+$ und $\mathbf{d} = \mathbf{0}$ gelten, womit der Satz von Gauss Markov für das klassische lineare Regressionsmodell bewiesen ist.

Die Varianz von $\hat{\beta}$ kann leicht über Lemma 1 bestimmt werden.

$$\Sigma_{\hat{\beta}\hat{\beta}} = \sigma^2 (X'X)^{-1} X'X (X'X)^{-1} = \sigma^2 (X'X)^{-1}$$

Um den Satz von Gauss Markov auch für das verallgemeinerte Modell anwenden zu können, muss nur das transformierte lineare Modell geschätzt werden.

$$\hat{\beta}^* = (X^{*'}X^*)^{-1} X^{*'}\mathbf{y}^* = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1}\mathbf{y}$$

Der transformierte Schätzer heißt verallgemeinerter kleinste Quadrate Schätzer oder Aitkenschatzen und wird als $\tilde{\beta}$ geschrieben. Er ist der beste lineare unverzerzte Schätzer des verallgemeinerten Regressionsmodells. Der Beweis erfolgt analog zum Beweis von Gauss Markov.

Multivariate statistische Analyse von Gesundheitsdaten
österreichischer Sozialversicherungsträger

Ortner, Th.

2015, XI, 67 S. 26 Abb., Softcover

ISBN: 978-3-658-08395-3