Identifikation und Visualisierung von Interaktionen in dünnbesetzten additiven Modellen- Eine Anwendung auf Überlebensdaten nach Nierentransplantation

Fabian Rasch

Fachbereiche Mathematik und Naturwissenschaften & Informatik, Hochschule Darmstadt

The data reported here have been supplied by the United Network for Organ Sharing as the contractor for the Organ Procurement and Transplantation Network. The interpretation and reporting of these data are the responsibility of the author and in no way should be seen as an official policy of or interpretation by the OPTN or the U.S. Government. The entire analysis is based on OPTN data as of June 20, 2020.

Motivation

Die Dialysebehandlung bei versagenden Nieren ist ein starke Einschränkung der Lebensqualität. Mehrmals wöchentliche, stundenlange, ambulante Termine und eine strenge Diät führen zu hohem psychischem Stress. Ein dauerhafte Wiederherstellung der Lebensqualität ist nur durch eine Nierentransplantation möglich.

Im Rahmen dieser Thesis wurden Nierentransplantationsdaten vom U.S. Department of Health and Human Services untersucht. Das Ziel war es mittels verschiedener Selektionsverfahren relevante Variablen zu identifizieren. Als besonderer Fokus wurden Interaktionen mit dem KDPI untersucht. Der KDPI steht für die Qualität der Spenderniere und spielt eine zentrale Rolle im Allokationsprozess. Es wurde untersucht ob der KDPI bei allen Patienten den gleichen Einfluss hat oder ob mangelnde Nierenqualität in Interaktion mit Merkmalen des Empfängers zu einer anderen Prognose führt.

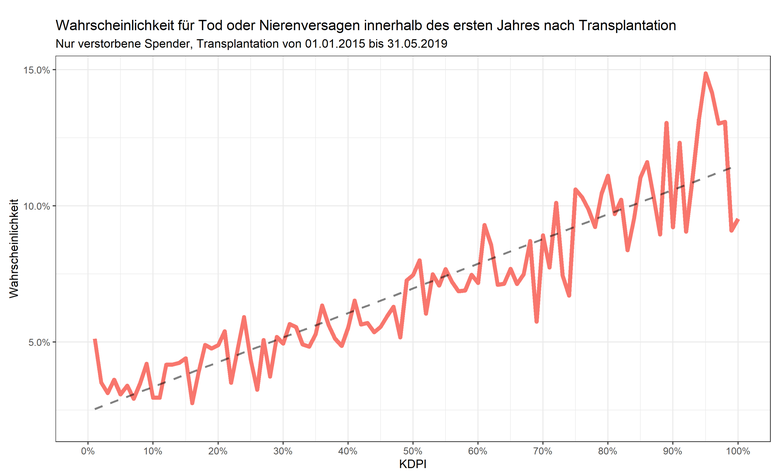

Abbildung 1: Einfluss des KDPI.

Die Erklärbarkeit von Prognosemodellen spielt eine wesentliche Rolle bei Entscheidungen, die das Leben eines Menschen gravierend beeinflussen können. Daher wurden spezielle, regularisierte additive Modelle trainiert und verglichen, um damit einflussreiche Variablen und Interaktionen zu einer ausgewählten Variable (hier: KDPI) zu identifizieren. Auf Basis der Variablenselektion, wurde ein gut erklärbares Regressionsmodell trainiert, um ein besseres Verständnis für Einflussfaktoren bei Nierentransplantationen entwickeln zu können. Die Ergebnisse wurden in einem interaktiven R-Shiny Dashboard verständlich dargestellt.

Zusätzlich wurden die Methoden auf simulierten Daten geprüft, um methodische Unterschiede bei der Modellierung mit Interaktionen zu einer ausgewählten Variable festzustellen. In der Simulationsstudie sollten die Methoden Haupt- und Interaktionsterme mit verschiedenen Kombinationen an Effektstärken unter Einwirkung vieler Störvariablen identifizieren.

Methoden

Bei den Methoden handelt es sich um Regressionsmodelle mit Lasso Regularisierung. Bei den Unterschieden der Methoden sind im Hinblick auf Interaktionen insbesondere die Einhaltung hierarchischer Bedingungen festzuhalten. Durch hierarchische Bedingungen können Interaktionen von Variablen nur in ein Modelle selektiert werden, wenn einer bzw. beide Haupteffekte im Modell enthalten sind. Zudem unterscheiden sich die Modelle durch die Möglichkeit Variablen in Gruppen zu regularisieren und nicht-lineare Effekte zu modellieren.

Group Lasso ist eine Erweiterung der Lasso Regularisierung. Dabei kann die Designmatrix X in J Gruppen eingeteilt werden. Group Lasso löst:

$$\min_{\beta\in\mathbb{R}^{\mathrm{p}}}\left\{\left\|\mathrm{Y}-\sum_{\mathrm{j}=1}^{\mathrm{J}}\mathrm{X}_{\mathrm{j}}\beta_{\mathrm{j}}\right\|_2^2+\lambda\sum_{\mathrm{j}=1}^{\mathrm{J}}\|\beta_{\mathrm{j}}\|_2\right\}$$

Durch die ℓ^2 Norm im zweiten Teil können einzelne Elemente innerhalb einer Gruppe nicht gleich null gesetzt werden und somit nur gesamte Gruppen ein- bzw. ausgeschlossen werden.

GLINTERNET trainiert Modelle mit paarweisen Interaktionen unter Beachtung der starken Hierarchiebedingung. Auch hier können Gruppen für unabhängige Variablen definiert werden. Mit Hilfe einer speziellen Parametrisierung von β_1, β_2 und $\beta_{1:2}$ lässt sich das Optimierungsproblem, beispielhaft mit zwei Variablen X_1 und X_2, umschreiben zu:

$$\underset{\mu,\beta}{\operatorname{argmin}}\frac{1}{2}\|\mathrm{Y}-\mu\cdot 1-\mathrm{X}_1\boldsymbol{\beta}_1-\mathrm{X}_2\boldsymbol{\beta}_2-\mathrm{X}_{1:2}\boldsymbol{\beta}_{1:2}\|_2^2+\lambda(\|\boldsymbol{\beta}_1\|_2+\|\boldsymbol{\beta}_2\|_2+\|\boldsymbol{\beta}_{1:2}\|_2)$$

Pliable bietet eine Variante der Lasso Regularisierung mit modifizierenden Variablen. Hierbei wird zwischen der typischen Designmatrix der Variablen X und der Designmatrix der modifizierenden Variablen Z unterschieden. Es besteht eine schwache Hierarchiebedingung zwischen X und Z. Pliable löst:

$$\underset{\beta_0,\theta_0,\beta,\Theta}{\operatorname{argmin}}\frac{1}{2\mathrm{N}}\sum(\mathrm{y_i}-\hat{\mathrm{y}_i})^2+(1-\alpha)\lambda\sum_{\mathrm{j}=1}^{\mathrm{p}}\left((\|\beta_{\mathrm{j}},\theta_{\mathrm{j}}\|_2+\|\theta_{\mathrm{j}}\|_2\right)+\alpha\lambda\sum_{\mathrm{j,k}}|\theta_{\mathrm{j,k}}|_1$$

Durch den Tuning-Parameter $\alpha\in(0,1)$ kann die Berücksichtigung der Interaktionseffekte kontrolliert werden. Die Koeffizienten der Hauptterme sind in β_{j} und die der Interaktionsterme in $\theta_{\mathrm{j,k}}$ enthalten.

GAMSEL trainiert und regularisiert generalisierte additive Modelle. Die Idee ist angelehnt an Smoothing Splines und bietet die Möglichkeit nicht-lineare Effekte von Variablen zu untersuchen. GAMSEL löst:

$$\underset{\alpha_0,\{\boldsymbol{\alpha}_{\mathrm{j}}\},\{\beta_{\mathrm{j}}\}}{\operatorname{argmin}}\frac{1}{2}\left\|\mathrm{y}-\alpha_0-\sum_{\mathrm{j}=1}^{\mathrm{p}}\alpha_{\mathrm{j}}\mathrm{x_j}-\sum_{\mathrm{j}=1}^{\mathrm{p}}\mathrm{U}_{\mathrm{j}}\boldsymbol{\beta}_{\mathrm{j}}\right\|_2^2+\lambda\sum_{\mathrm{j}=1}^{\mathrm{p}}\left(\gamma|\alpha_{\mathrm{j}}|+(1-\gamma)\|\boldsymbol{\beta}_{\mathrm{j}}\|_{\mathrm{D}_{\mathrm{j}}^*}\right)+\frac{1}{2}\sum_{\mathrm{j}=1}^{\mathrm{p}}\psi_{\mathrm{j}}\boldsymbol{\beta}_{\mathrm{j}}{}^{\mathrm{T}}\mathrm{D}_{\mathrm{j}}\boldsymbol{\beta}_{\mathrm{j}}.$$

Hierbei ist β_{j} der nicht-lineare Anteil der Variable X_{j} und U_{j} ist die Auswertung der Basisfunktionen. Durch den Parameter γ lassen sich lineare oder nicht-lineare Effekte bevorzugen. Der zweite Teil des Strafterms sorgt für die Einhaltung der vorher definierten Freiheitsgrade.

Die wichtigsten Unterschiede zusammengefasst:

Verfahren	Hierarchie	Variablen gruppieren	Nicht-Lineare Effekte
Group Lasso	Keine	Ja	Nein
GLINTERNET	Stark	Ja	Nein
Pliable	Schwach/Stark	Nein	Nein
GAMSEL	Keine	Nein	Ja

Tabelle 1: Die wichtigsten Unterschiede der Methoden

Ergebnisse

Mit Hilfe der Selektionsverfahren konnte die Anzahl an Variablen von 28 Haupttermen und entsprechend 27 Interaktionstermen mit dem KDPI auf 17 Hauptterme und 3 Interaktionsterme reduziert werden. Insgesamt verbesserten die Interaktionen die Modelle nur marginal. Durch ein interaktives Dashboard konnten die Effekte verständlich dargestellt werden.

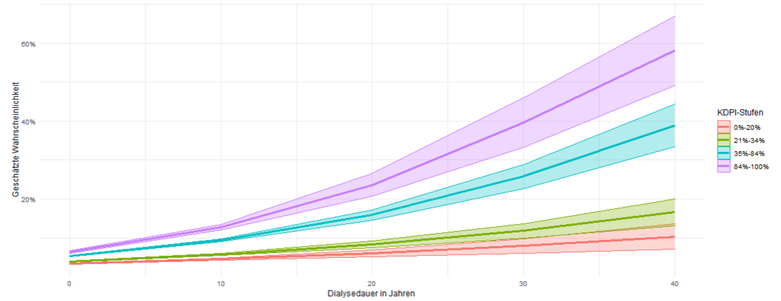

Abbildung 2: Interaktion zwischen KDPI und der Dialysedauer in Jahren.

Beim systematischen Methodenvergleich anhand des Anwendungsfall und einer Simulationsstudie konnten erhebliche Unterschiede in den Methoden festgestellt werden. GGLasso und Pliable zeigten erhebliche Probleme bei der Identifikation kategorieller Variablen und stetigen Variablen mit wenigen Ausprägungen. GAMSEL fiel durch eine sehr hohe Varianz in den Modellen auf und schien insgesamt ungeeignet bei der Modellierung von Interaktionen. GLINTERNET zeigte die beste Performance der Modelle und eignete sich für den speziellen Anwendungsfall der Interaktionen mit einer ausgewählten Variable am besten.

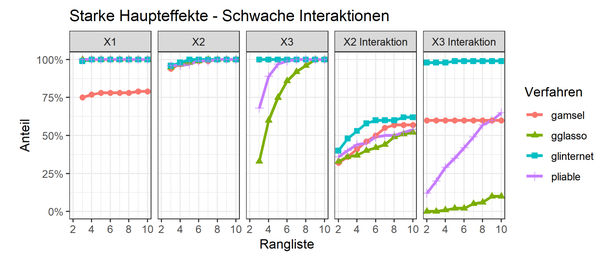

Abbildung 3: Auszug der Simulationsstudie.

Referenzen

[1] Alexandra Chouldechova and Trevor Hastie.
Generalized additive model selection, 2015.

[2] Michael Lim and Trevor Hastie.
Learning interactions through hierarchical group-lasso regularization, 2013.

[3] Robert Tibshirani and Jerome Friedman.
A pliable lasso, 2018.

[4] Yi Yang and Hui Zou.
A fast unified algorithm for solving group-lasso penalize learning problems, 2014.