Romedius Troberg

Modellselektion

Masterarbeit

Diplom.de

Bibliografische Information der Deutschen Nationalbibliothek:

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

ISBN: 9783961164455

Romedius Troberg

Modellselektion

Diplom.de

Inhaltsverzeichnis

1 Einleitung

Eines der bekanntesten Zitate über die Validität von Statistik wird fälschlicher Weise Winston Churchill zugeordnet. Auch wenn Winston Churchill nicht der Autor des berühmten Zitats „Glaube keiner Statistik, die du nicht selber gefälscht hast“ ist, so ist die Bedeutung dieses Zitat genauso valide. Dieses Zitat wird heute oft angebracht, um Statistik als Werkzeug für die individuelle Auslegung von Daten im Sinne des Autors darzustellen, welcher versucht durch die Vorlage einer Statistik die Unwissenden – also vor allem Nicht-Statistiker – von seiner Meinung zu überzeugen.

Weit abfälliger äußerte sich Friedrich Nietzsche über das Thema Statistik und machte in einem seiner bekanntesten Zitate deutlich, was er von Statistik hält. Mit seinem Zitat, dass „die Massen […] nur in dreierlei Hinsicht einen Blick zu verdienen [scheinen]: einmal als verschwimmende Copien der grossen Männer, auf schlechtem Papier und mit abgenutzten Platten hergestellt, sodann als Widerstand gegen die Grossen und endlich als Werkzeuge der Grossen; im Uebrigen hole sie der Teufel und die Statistik!“, äußerte Nietze (1874) seine Kritik am unwissenschaftlichen Handeln vieler seiner Zeitgenossen.

In vielen wissenschaftlichen Arbeiten wird versucht eine Theorie anhand von Modellen zu erklären. Für den Beweis, dass das jeweilige Modell die Theorie bestätigt, bedienen sich viele Wissenschaftler der statistischen Analyse von Daten. Dabei existiert in der Theorie eine große Anzahl an Modellen, welche für die statistische Analyse von Rohdaten zur Verfügung stehen.

Die Auswahl eines geeigneten Modells für die Diskussion der Forschungsfrage ist für jeden Wissenschaftler essenziell, da die Wahl des richtigen Modells wesentlichen Einfluss auf die Validität der Analyse hat. Im Falle einer fehlerhaften Modellselektion führen statistische Analysen häufig zu Fehlinterpretationen und können im schlimmsten Fall aufwendige Arbeiten als wertlos erscheinen lassen.

Von besonderer Bedeutung ist die Wahl des richtigen Modells in der Wirtschaftsinformatik. Das Masterstudium Wirtschaftsinformatik der Ferdinand Porsche FernFH beschäftigt sich in mehreren Fächern mit dem Thema der

statistischen Datenanalyse im Bereich der Wirtschaftsinformatik. Besonders in den Fächern IS421 - Business & Competitive Intelligence Systems und MT422 - Methoden der Datenanalyse lernen die Studierenden die Grundlagen der Datenanalyse und Grundsätze bei der Wahl eines geeigneten Modells in der Wirtschaftsinformatik kennen.

Die Masterarbeit zum Thema „Modellselektion“ soll darlegen, was unter Modellselektion zu verstehen ist, und soll aufzeigen, welche Gütekriterien bei der Selektion geeigneter Modelle in der statistischen Analyse herangezogen werden können. Hierdurch soll die kolportierte Aussage Churchills, dass Statistik nur ein Werkzeug sei, um die Unwissenden von der eigenen Theorie zu überzeugen, widerlegt werden und es soll ein Gegenbeweis dafür geliefert werden, dass Statistik das Werk des Teufels sei.

Die Arbeit wird darlegen, dass Statistik, unter anderem aus Sicht der Wirtschaftsinformatik, eine wertvolle Wissenschaft ist, die besonders dadurch ihre Berechtigung erfährt, dass sie andere Wissenschaftsbereiche dabei unterstützt, die jeweiligen Theorien mit statistisch sinnvoll ausgewerteten Daten zu verifizieren oder falsifizieren.

1.1 Forschungsfrage

Um Das Thema der Modellselektion wissenschaftlich zu untersuchen wird die folgende Forschungsfrage in der Masterarbeit behandelt:

Welche Gütekriterien können herangezogen werden, um ein geeignetes Modell auszuwählen, damit die statistische Analyse in wissenschaftlichen Arbeiten die jeweilige Theorie valide verifiziert bzw. falsifiziert?

1.2 Methodische Vorgangsweise

Das Vorgehen in der Masterarbeit ist eine Kombination aus Empirie und nicht-empirischen Methoden. Zunächst wird ein Überblick über die vorhandene Literatur in dem Forschungsgebiet gebracht, um aufzuzeigen, wie der aktuelle Stand der Forschung in diesem Bereich ist und um folglich zu einem Schluss zu kommen, in welchem unerforschten Bereich sich das Thema der Masterarbeit befindet. Hierfür

ist es nötig, einen möglichst umfassenden, aber nicht zu detailreichen Überblick über vorhandene Literatur zu geben.

Darauffolgend wird sich die Masterarbeit in Form einer Diskussion mit der Fragestellung auseinandersetzen, inwieweit die Beantwortung der Forschungsfrage für die Forschung in dem Bereich der Modellselektion einen Mehrwert bieten kann. Danach soll eine nähere Definition von Modellselektion vorgenommen werden, um eine Abgrenzung des, in der Masterarbeit, bearbeiteten Themas machen zu können. Dies ist notwendig, da der Autor nach ausgiebiger Recherche vermutet, dass zur vollständigen Analyse des Themas „Modellselektion" der Umfang einer Masterarbeit nicht ausreichend sein wird.
Folgend werden die bekanntesten Modelle im Bereich der statistischen Datenanalyse erklärt und auf die einzelnen Gütekriterien eingegangen, welche für die Selektion eines geeigneten Modells herangezogen werden können. In diesem Bereich werden von dem Autor verschiedene Hypothesen zu den Gütekriterien und der daraus resultierenden Möglichkeit einer Modellselektion aufgestellt.

Um die aufgestellten Hypothesen zu verifizieren bzw. falsifizieren, hat sich der Autor entschieden, anhand eines Datensatzes die praktische Umsetzung zu demonstrieren und die Ergebnisse zu interpretieren.
Aufgrund der Ergebnisse der Analyse wird der Autor darauf folgend einen Bogen zu den bisherigen Erkenntnissen in der Literatur spannen, wie im Überblick über die vorhandene Literatur angeführt, und wird versuchen durch seine eigenen Ergebnisse einzelne Erkenntnisse zu verifizieren bzw. falsifizieren. Nach dieser Darstellung hat der Autor vor eine Zusammenfassung seiner Ergebnisse anzubringen und die erreichte Forschungsleistung herauszustellen.

Der Schluss der Arbeit wird nebst einer Zusammenfassung der Ergebnisse des Hauptteils auch Gedanken zur weiteren Bearbeitungsmöglichkeit des Themas darlegen, um so zukünftigen Forschern einen Denkanstoß zu bieten, in welcher Weise sie bei der Erforschung des Themas weiter vorgehen können und welche

Fragestellungen im Bereich der Modellselektion weiterhin offen sind. Der Autor wird dabei versuchen diese Forschung zu kategorisieren, indem er aufzeigen wird, welche Fragestellungen sich für weitere Masterarbeiten eignen oder welche Fragestellungen einen solch großen Umfang in der Bearbeitung haben, dass sie eher in einer Doktorarbeit beantwortet werden können.

1.3 Übersicht über die Kapitelinhalte

Die Masterarbeit zum Thema „Modellselektion“ beginnt mit einer Vorstellung der Grundlagen der Modellselektion. In diesem Kapitel werden die Theorie und die Kriterien der Modellbildung erläutert. Diese Vorstellung leitet zu dem Kapitel „Schritte der Modellformulierung“ über, in welchem aufgezeigt wird, wie ein Modell in der Statistik formuliert werden kann und was die Intentionen der Modellbildung sein können. Modelle selbst können in der Statistik in Form von parametrischen und nichtparametrischen Varianten formuliert werden. Nachdem diese Grundlagen erklärt sind, beschäftigt sich die Masterarbeit intensiv mit dem Thema der Modellselektion selbst. Hierbei werden zunächst die Intention für die Modellselektion selbst diskutiert und der Prozess der Modellselektion vorgestellt. Hierbei werden die gebräuchlichsten Ansätze, die Selektion mittels Tests und der Richtung der Modellselektion, die Selektion mittels sog. Shrinkageansätzen und die Selektion auf Basis eines Selektionskriteriums, im Bereich der Modellselektion vorgestellt. Im Kapitel Regressionsanalyse in der Statistik werden in der Folge einzelne Regressionsmodelle vorgestellt sowie deren Anwendung und Nutzen diskutiert. Besonders bei Vorliegen von Multikollinearität bietet es sich an die Ridge Regression bzw. die Lasso Regression in Betracht zu ziehen.

In jedem Abschnitt wird dabei auch auf die Möglichkeit einer Anwendung der jeweiligen Theorie mit der Statistiksoftware R vorgestellt und anhand von öffentlich verfügbaren Datensätzen deren praktische Anwendung simuliert.

2 Grundlagen der Modellselektion

Die Modellselektion ist ein breites Spektrum der Statistik. Im Zuge der Modellselektion müssen mehrere Teilbereiche bzw. Disziplinen der Statistik beachtet werden. So kommen im Zuge der Modellselektion Methoden der deskriptiven Statistik zum Einsatz, indem vorliegende Daten in geeigneter Weise aufbereitet, zusammengefasst und beschrieben werden, als auch die mathematischen Methoden der induktiven Statistik. Aber auch einige Aspekte der explorativen Statistik werden in der Modellselektion angewandt, indem mit analytischen Methoden und Datamining systematische Zusammenhänge in den Daten erforscht werden. Dabei stellen sich zu Beginn der Modellselektion einige grundsätzliche Fragen, welche analysiert und bearbeitet werden müssen, bevor die eigentliche Selektion eines Modells beginnt.
Zunächst sollten die Kriterien der Modellbildung bedacht werden. Dies ist wichtig, um auf der Suche eines Modells eine geeignete Vorgehensweise zu wählen. Als nächster Schritt muss ein Modell formuliert werden, um dieses in Form einer Regressionsanalyse zu schätzen. Ist man sich über die Prozesse der Modellselektion bewusst, so sollte man sich die Intention der Modellbildung klar machen, um zu entscheiden, um welche Art von Modell es sich grundsätzlich handeln wird.

2.1 Überblick über vorhandene Literatur im Bereich Modellselektion

Bevor man sich mit dem Thema der Modellselektion auseinandersetzt, bietet es sich an, sich einen Überblick über die vorhandene Literatur zu dem Thema zu verschaffen. Die Literatur zum Thema Modellselektion begann mit der schrittweisen Regression (Breaux (1967)) und Autometrie (Hendry und Richard (1987)). Sie ging zu fortgeschritteneren Verfahren über, von denen die bekanntesten die nicht negative Garrotte (Breiman (1995)), der Operator für die Auswahl des geringsten Winkels und der Schrumpfung (Breiman (1995)) sind. LASSO (Tibshirani (1996)) und das sichere Unabhängigkeits-Screening (Fan und Zhang (2008)) sind in

späteren Arbeiten zum Thema Modellselektion beschrieben. Fan und Lv (2010) überprüfen den größten Teil der Literatur zu linearen und verallgemeinerten Modellen.

Ein beachtlicher Teil der Literatur ist Methoden und algorithmischen Lösungen gewidmet, aber auch die optimale Wahl der Parameterstrafe findet in der Literatur Beachtung. Breheny und Huang (2009) und Huang et al. (2012) geben einen vollständigen Überblick über Auswahlverfahren für Modelle mit gruppierten Variablen. Sie erörtern in ihren Arbeiten technische Vergleiche, insbesondere im Hinblick auf die Konvergenzrate. Castle et al. (2011) vergleichen die Autometrie mit einer Vielzahl anderer Methoden, wie z.B. der schrittweisen Selektion, dem Akaike-Informationskriterium, LASSO und weiteren, hinsichtlich der Vorhersagegenauigkeit unter Orthogonalität der Regressoren. Dabei schenken sie dynamischen Modellen besondere Aufmerksamkeit. Park et al. (2015) geben einen Überblick über die Verfahren zur Variablenauswahl. Fan und Lv (2010) bieten einen umfassenden Überblick über den Kontext der unabhängigen Überprüfung der Unabhängigkeit. Fu (1998) vergleicht Bridge und LASSO sowohl theoretisch als auch empirisch unter Verwendung von Simulationen und realen Daten. Epprecht et al. (2017) vergleichen die Autometrie und LASSO gemäß der Vorhersage und Auswahlgenauigkeit. Gute Übersichten für diese Klassifizierung wurden von Blum und Langley (1997) und Saeys et al. (2007) veröffentlicht. Neuere Ansätze, einschließlich der Relation zu ökonometrischen Methoden, wurden von Mehmood et al. (2012) und Jovic´ et al. (2015) untersucht.

Die einzelnen Informationskriterien finden sich bei Mallows (1973), Akaike, (1973), Hurvich und Tsai (1990), Schwarz (1978) oder Hannan und Quinn (1979).

Die Arbeiten von Hurvich und Tsai (1990), Steyerberg et al. (1999), Whittingham et al. (2006) oder Flom und Cassell (2007) beschäftigen sich mit der schrittweisen Regressionsanalyse.

Bisherige Arbeiten zum Thema „Modellselektion“ konzentrieren sich vor allem auf die Wahl eines Modells mittels eines einzigen Verfahrens bzw. vergleichen einige Verfahren miteinander. Die vorliegende Masterarbeit richtet ein Augenmerkt auf die gesamte Theorie der Modellselektion und beschäftigt sich mit der Frage, welche

Gütekriterien bei der Modellselektion herangezogen werden können, um eine geeignete Wahl für ein Modell zu treffen. Dies ist entscheidend, da es bei Wahl eines Modells darauf ankommt, dieses so zu wählen, dass sich die Rohdaten durch das Modell aussagekräftig analysieren und folglich interpretieren lassen. Um eine Aussage darüber zu treffen, wie „gut" ein jeweiliges Modell die gegebenen Daten darstellt existieren in der Theorie diverse Gütekriterien. In dieser Arbeit werden die verschiedenen Verfahren und Gütekriterien vorgestellt. Deren praktische Anwendung wird mit der Statistik Software R erläutert. Jeder Abschnitt schließt daher, soweit möglich, mit der entsprechenden Berechnung in R ab. Für diese Berechnungen werden frei verfügbare Datensätze der R-Community[1] verwendet.

Um sich dem Thema der Modellselektion zu nähern, bietet es sich zunächst an, den Kriterien der Modellbildung einen Blick zu widmen, um zu verstehen, was die Grundlagen der Modellbildung und in Folge der Selektion eines geeigneten Modells sind.

2.2 Kriterien der Modellbildung

Bei der Modellbildung bieten sich die zwei Grundsätze, dass das gewählte Modell

(i) die Daten gut (genug) beschreiben sollte und

(ii) nicht mehr erklärende Variablen als unbedingt nötig verwenden sollte, an. (Crawley (2007))

Die zweite Bedingung (ii) ist unter dem Schlagwort Ockhams Rasiermesser bekannt. Crawley (2007) schreibt hierzu: „Entia non sunt multiplicanda praeter necessitatem" (Crawley 2007, S. 325). Diese lateinische Bedingung sagt aus, dass in dem Fall, dass zur Beschreibung eines Phänomens bsw. zwei Modelle existieren, jenes in der Wahl bevorzugt werden sollte, welches mit weniger erklärenden Variablen die Daten gut

[1] Die verwendeten Datensätze und Packages sind in R frei verfügbar; Für eine genaue Beschreibung der Datensätze und Packages vgl. https://cran.r-project.org/web/packages/available_packages_by_date.html; Der analysierte Hauptdatensatz ist der Boston Housing – Datensatz verfügbar u.a. unter https://www.kaggle.com/schirmerchad/bostonhoustingmlnd/data (abgerufen: 01.05.2020).

beschreibt. Für die Praxis kann hieraus der Schluss gezogen werden, dass jenes Modell gewählt werden sollte, welches das vermeintlich Einfachere der beiden ist.
Oftmals startet der Prozess der Modellbildung mit einer vorläufigen Formulierung eines Modells, welches dann auf die oben erwähnten Bedingungen ((i) und (ii)) hin geprüft wird. In einem iterativen Prozess, der sogenannten Modellselektion, werden die gewählten Modelle solange modifiziert, bis entweder ein Modell übrig bleibt, welches für die Analyse der Daten herangezogen werden kann oder ein Modell bleibt das alle Prädiktoren als wirkungslos (z.B. nicht signifikant) eliminiert hat und man schlussfolgern muss, dass das untersuchte Phänomen mit den angenommenen Prädiktoren nicht gut (genug) beschrieben werden kann. Im Extremfall könnte es sein, dass kein Modell zur Analyse der Daten als geeignet angesehen werden kann. In diesem Fall müsste man mit dem bestmöglichen Modell Vorlieb nehmen.

2.3 Schritte zur Modellformulierung

Den ersten Schritt der Bildung eines Modells stellt die Formulierung einer vorläufigen Modellgleichung dar, mit welcher die Verteilung einer (oder mehrerer) abhängiger/erklärter Variablen durch die lineare Kombination einer (oder mehrerer) unabhängiger/erklärender Variablen und ggf. ihrer Interaktionen beschrieben wird. Interaktion meint in diesem Zusammenhang das Zusammenwirken von zwei oder mehreren Variablen im Modell.
Eine mögliche, einfache Modellformulierung könnte folgendermaßen in zwei Schritten erfolgen:

(1) $\text{Variable}_{\text{abhängig}} \sim \text{Variable}^1_{\text{unabhängig}} + \text{Variable}^2_{\text{unabhängig}} + \text{Variable}^1_{\text{unabhängig}} : \text{Variable}^2_{\text{unabhängig}}$

(2) $\text{Vorhergesagte Variable}_{\text{abhängig}} = b_0 + b_1 \times \text{Variable}^1_{\text{unabhängig}} + b_2 \times \text{Variable}^2_{\text{unabhängig}} + b_3 \times \text{Variable}^1_{\text{unabhängig}} : \text{Variable}^2_{\text{unabhängig}}$

Die in Schritt (1) abgebildete Modellgleichung versucht die Relation zwischen einer abhängigen Variable und mehreren unabhängigen Variablen zu bilden, so dass, im Idealfall, Werte der abhängigen Variable mit Werten der unabhängigen Variablen vorhergesagt werden können.

Diese Modellgleichung wird in einem weiteren Schritt (2) durch eine Regressionsanalyse untersucht. Als Resultat einer solchen Regressionsanalyse erhält man Werte für die Koeffizienten (b1, b2, b3), welche aufzeigen, wie stark und in welche (positive oder negative) Richtung die erklärenden Variablen die erklärte Variable beeinflussen.
Darüber hinaus erhält man hierdurch eine Prüfstatistik (t-Wert) und deren Streuung (Standardfehler σ), welche dafür dienen können, einen Signifikanzwert (p-Wert) für den entsprechenden Koeffizienten zu berechnen. Der Signifikanzwert drückt hierbei aus, ob die unabhängige(n) Variable(n) einen signifikanten Einfluss auf die abhängige Variable hat (haben).

2.4 Die zwei Sichtweisen der Modellselektion

Grundsätzlich sind in der Modellwahl zwei verschiedene Sichtweisen möglich. Zum einen kann versucht werden das „wahre Modell" bzw. die „Realität" hinter den Daten zu erfassen. Dieses „wahre Modell" hängt im Extremfall von unendlich vielen Parametern ab. Das Erfassen dieses Modells ist nur durch eine mehr oder weniger gute Approximation durch endlich dimensionale Verfahren möglich. Offensichtliche Effekte lassen sich hierbei bereits mit sehr einfachen Verfahren erkennen. Mittels statistischer Verfahren lassen sich weniger offensichtliche Effekte errechnen. Sehr kleine Effekte können hingegen, so gut wie, nicht erkannt werden und werden in Regressionsmodellen meist in Form eines Fehlerterms addiert. Die Güte eines solchen Modells kann nur relativ geschätzt werden.
Zum anderen könnte sich die Realität in einem Modell mit endlich vielen Parametern darstellen lassen. In diesem Fall wäre es grundsätzlich möglich das „wahre Modell" zu erkennen. Ein Beispiel hierfür wären Computersimulationen. Die Güte dieser Modelle kann folglich absolut geschätzt werden.