Automatische Verfahren zur Prädiktorauswahl in Regressionsmodellen

Literaturarbeit vorgelegt von
Markus Graf (markus.graf@uzh.ch)
am Psychologisches Institut der Universität Zürich
Betreut durch Dr. Christina Werner

Zusammenfassung

Eigener Abstract: -
Themenvorgabe: In vielen psychologischen Bereichen geht es darum, Kriteriumsvariablen durch Prädiktorvariablen möglichst gut vorherzusagen. Wenn viele potentielle Prädiktorvariablen in Frage kommen und es keine theoretischen Gründe gibt, die nur ganz bestimmte Prädiktorvariablen nahelegen, werden in Anwendungssituationen oft automatische Verfahren der Auswahl von Prädiktorvariablen verwendet, um mit möglichst wenigen Prädiktoren eine möglichst gute Vorhersage des Kriteriums zu erreichen, beispielsweise die sog. "Stepwise"-Methode in multiplen Regressionsmodellen. Die Literaturarbeit soll einen Überblick über verschiedene existierende Möglichkeiten zur Selektion von Prädiktorvariablen in Regressionsmodellen geben, und deren Eignung für psychologische Anwendungen kritisch diskutieren.

Inhaltsverzeichnis

Einführung

Das Standardverfahren um den quantitative Zusammenhang zwischen einer abhängigen und einer unabhängigen Variablen zu beschreiben stellt die Regressionsanalyse dar. Begründet wurde dieses Verfahren durch Carl Friedrich Gauss in seiner Schrift, in der er, mit Hilfe der Methode der kleinsten Quadrate, die Bewegung der Himmelskörper um die Sonne im Kegelschnitt beschrieb (Gauss, 1809).

Im Unterschied zur einfachen linearen Regression, werden in einem multiplen Regressionsmodell mehrere unabhängige Variablen mit einbezogen. Es resultiert eine Regressionsgleichung welche zur Vorhersage einer Kriteriumsvariable aufgrund mehrerer Prädikatorvariablen genutzt wird (Bortz & Schuster, 2011, S. 448). Als Gretchenfrage stellt sich nun welche Prädikatoren nun ein Modell am besten erklären. Zu beginn der Psychologischen Forschung mussten Modelle von Hand berechnet werden. Zwangsläufig wurden wenige Prädikatoren erhoben und einfache Modelle Gerechnet. Friedman analysierte beispielsweise 1944 die Langlebigkeit von Turbinenschaufeln in Abhängigkeit von Stress, Temperatur und einigen Legierungsparametern. Zwar wurde die Berechnung nicht mehr von Hand durchgeführt, doch benötigte eine Regressionsschätzung inklusive Berechnung der Teststatistiken rund 40 Stunden (Armstrong, 2011, p.2). Jeder durchschnittliche Computer erledigt dies heutzutage Sekundenbruchteile.

Mit dem technische Fortschritt einhergehend wurden Verfahren entwickelt, welche möglichen Kombinationen von Prädiktoren berücksichtigen und gegeneinander testen. In einem ersten Teil dieser Arbeit werden die wichtigsten Verfahren dargestellt.

Es können beliebig viele potentiell erklärende Variablen erhoben werden um sich komplexe Modell generieren zu lassen. Menschen tendieren zu glauben, dass komplexe Probleme komplexe Lösungen benötigen. Die Forschung zeigt jedoch, dass gerade das Umgekehrte der Fall ist (Armstrong, 2011, p.3). Insbesondere Gigerenzer demonstrierte eindrucksvoll wie mit einfachen Rekognitionsheuristiken bessere Vorhersagen gemacht werden konnten als mit komplexen statistischen Modellen

(Borges, Goldstein, Ortmann & Gigerenzer, 1999). Komplexe Modelle können sehr gute Vorhersagen liefern innerhalb des Trainingsdatensatz, doch oft scheiter die Vorhersage beim Versuch, diese zu generalisieren. Der zweite Teil befasst sich mit dem Problem der Überanpassung komplexer Modelle und diskutiert mögliche Lösungsansätze.

Automatische Modellwahlverfahren

In der psychologischen Forschung kommen oft Ex-post-facto-Designs zum Einsatz. Dies insbesondere weil viele Daten mit geringem Aufwand erhoben werden können. Oft werden Daten gleich für mehrere Studien erhoben und in anderen Studien verwendet. Daraus resultieren viele potenzielle Prädikatorvariablen für neue Fragestellungen. Es gilt das "beste" Modell berechnen, oft ohne das es theoretische Gründe gibt nach denen die Auswahl der Prädiktoren einzuschränken ist. "Alle Modelle sind falsch, aber einige sind nützlich" (Box, 1979, p.202). Box will damit hervorheben, dass obschon in der Literatur oft vom "besten" oder "wahren" Modell gesprochen wird, dies nur ein Approximation darstellt (Weakliem, 2004, p.172).

Exhaustiv Regression

Eine naive Herangehensweise ist alle möglichen Modelle welche mit p Prädiktoren möglich sind durch zurechnen. Zur Beurteilung der Modellgüte kann die mittlere quadratische Abweichung herangezogen werden. Das "beste" Modell hat die kleinste mittlere quadratische Abweichung (S_p-Kriterium) (Thompson, 1978, p. 5). Da alle Möglichen Kombinationen durch gerechnet werden, wird das "beste" Modell, also das welches den Referenzdatensatz am besten Vorhersagen kann, gefunden. Entsprechend wird in der Literatur auch oft vom optimalen Modell gesprochen. Es liegt jedoch auf der Hand, dass dies zwar die interne Validität erhöht, aber die externe Validität gefährdet. Das optimale Modell kann, insbesondere bei hoher Komplexität, zu fest auf die Referenzdaten behaftet sein und dadurch schlecht generalisierbar sein. Auf diesen Effekt der Überanpassung wird

im zweiten Teil dieser Arbeit noch näher eingegangen. Thompson (1978, p.6) sieht einzig den Nachteil darin, dass der Rechenaufwand exponentiell mit der Anzahl zu berücksichtigender Prädikatoren steigt. Es müssen immer $2^p - 1$ Modelle berechnet werden, bei 5 Prädikatoren sind dies 31 Modelle, bei 10 bereits 1023 usw. Während früher eingeschränkte Rechenkapazität oft ein ökonomischer Faktor war - es musste Rechenzeit in einem Rechenzentrum reserviert werden, spielt die Rechengeschwindigkeit auf modernen Systemen eine untergeordnete Rolle, da insbesondere in der Psychologie oft nur eine Handvoll Prädikatoren durch gerechnet werden müssen.

Schrittweise Regression

Das optimale Modell beinhaltet jeden Prädikator, der die Voraussage bezüglich des getesteten Datensatzes auch nur minimal verbessert. Es stellt sich die Frage ob diese minimale Verbesserung noch nützlich ist. Die "stepwise"-Verfahren arbeiten wesentlich liberaler, in dem Prädikatoren hinzugefügt oder eliminiert werden, je nach deren Relevanz für die Modellgüte. Es werden Kriterien festgelegt, nach welchen ein Modell als angemessen zu betrachten ist. Dies hat gegenüber der exhaustiven Verfahren den Vorteil, dass nicht alle Modelle berechnet werden müssen und entsprechend schneller Lösungen gefunden werden. Im Schnitt müssen xxxx Modelle berechnet werden, um eine adäquate Lösung zu finden (?, ?).

Innerhalb der schrittweisen Verfahren unterscheidet man zwischen *forward selection* und *backward elimination*. Ausgehend vom leeren Modell werden in der ersten Variante schrittweise weitere Variable der Nützlichkeit nach in das Modell integriert, bis eine Abbruchbedingung erfüllt ist.

In der zweiten Variante werden alle Prädikatoren in das Modell integriert und schlechte entfernt, wiederum bis das Kriterium erreicht ist.

Die Aufnahme einer neuen Variable kann dazu führen, dass eine bereits im Modell vorhandene Variable obsolet wird. Um diesem Umstand Rechnung zu tragen werden oft forward selection und backward elimination kombiniert (Bortz & Schuster, 2011, p. 461).

In seltenen Fällen kann es vorkommen, dass zwei Variablen für sich das Kriterium in die Regressionsgleichung aufgenommen zu werden nicht erfüllen, jedoch zusammen zum Vorhersagepotential einen substantiellen Beitrag leisten (Jacob Cohen, West, Cohen & West, 2003, p.261). Schrittweise Verfahren sind entsprechend nicht in der Lage solche Effekte mit zu berücksichtigen.

Entgegen dem exhaustiven Verfahren besteht bei schrittweisen Verfahren das Problem, dass unter Umständen nicht das optimale Modell gefunden wird. Da die Nützlichkeitsunterschiede, welche oft nur geringe statistische Bedeutung haben, das Modell bestimmen, bezeichnet Bortz und Schuster (2011, p. 462) dieses Verfahren eher zu den explorativen gehörend. Harrell (2001, p. 56ff) lehnt das Verfahren gar ab und führt ins Feld, dass sämtliche statistischen Prinzipien verletzt würden. Berk (1978) zeigte jedoch in einem Vergleich, dass die durchschnittliche Differenz der Fehlerquadratsummen zwischen exhaustiven und schrittweiser Regression kaum 7% übertrifft.

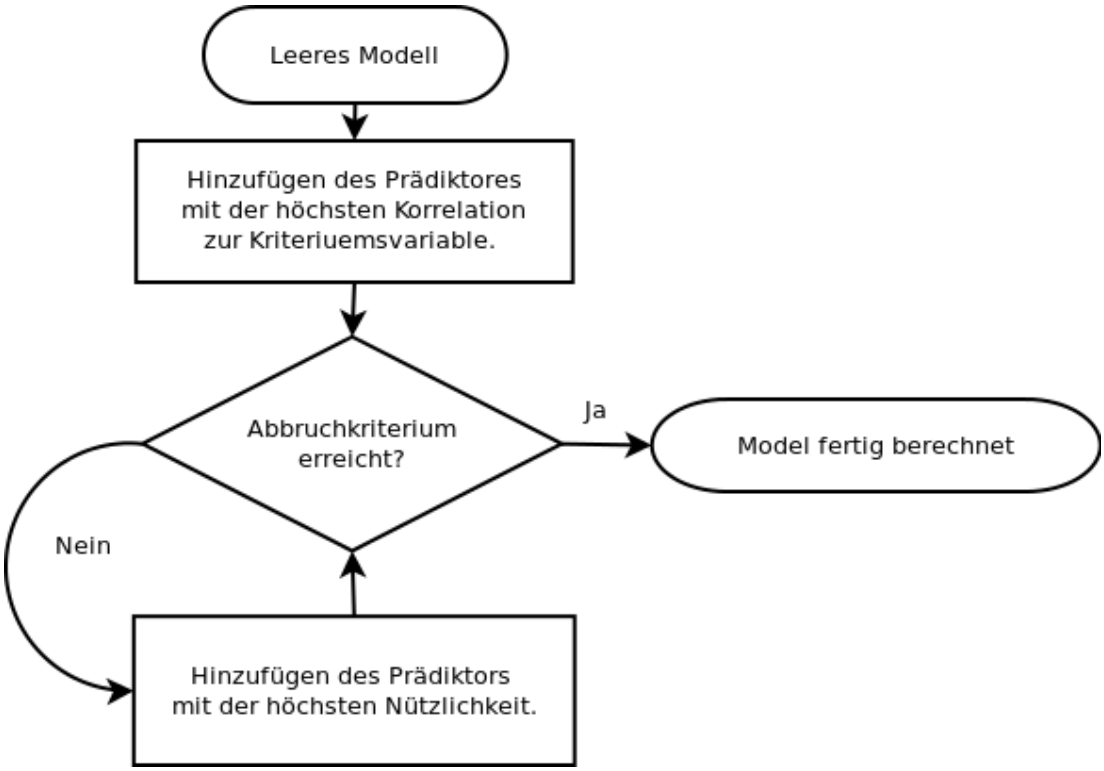

Abbildung 1. : Forward Selection

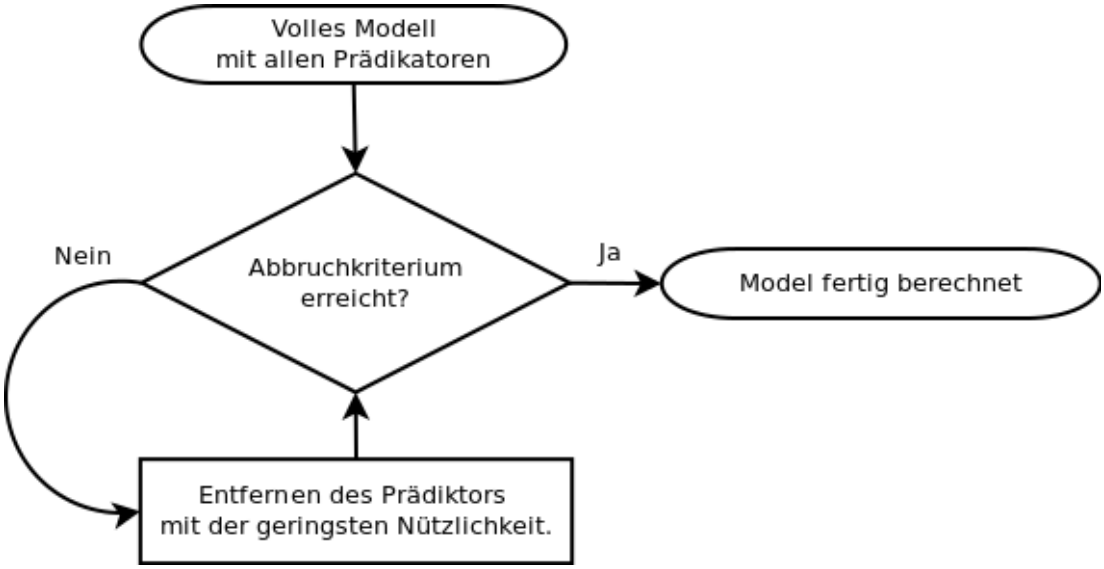

Abbildung 2. : Backward Elimination

Zentrales Element der schrittweisen Regression ist das Mass zur Beurteilung der Modellanpassung, welche besagt, weshalb und wann ein Modell als akzeptabel zu betrachten ist. Als Folge dessen wird damit auch die Anzahl relevanter Prädiktoren bestimmt.

Bestimmtheitsmass Das Quadrat des multiplen Korrelationskoeffizienten R^2 besagt wie viel systematische Varianz aufgeklärt wird. Je grösser die Zahl der unabhängigen Variablen ist, desto grösser wird das Bestimmtheitsmass, weshalb eine Korrektur vorgenommen werden muss. Insbesondere bei kleinem R^2 sollte auf Signifikanz getestet werden, entsprechend wird in schrittweisen Verfahren Prädikatoren nicht einzig aufgrund von R^2 selektiert.

Signifikanztest Das Verfahren wird beendet, wenn kein Prädikator mehr vorhanden ist, der das Vorhersagepotential signifikant erhöht (Bendel & Afifi, 1977, p.48). Das vergleichen zweier Regressionsgleichungen mittels Signifikanztest bedingt, dass diese geschachtelt sein müssen, das kleinere Modell muss im grösseren enthalten sein (Jacob Cohen et al., 2003, p. 508).

Das gewählte Signifikanzniveau ist eigentlich unbegründet gewählt (Weakliem, 2004, p. 174). Derksen und Keselman (1992, p. 269) diskutieren mehrere Empfehlungen für Signifikanzniveaus und weisen darauf hin, das sich über mehrere Tests der α-Fehler kumuliert. In Simulationen mit artifiziellen Daten zeigen Mundry und Nunn (2009) das Problem multipler Tests beispielhaft auf. Daraus resultierend lehnen sie die Verwendung der schrittweisen Regression gar ab.

Ein weitere Schwäche des Hypothesentestens ist der Einfluss der Stichtprobengrösse. So kann bei genügend grossem Umfang nahezu jede Hypothese abgelehnt werden, was wiederum zu komplexen und überangepassten Modellen führt (Weakliem, 2004, p.173).

Multikollinearität

Multikollinearität, als hohe Korrelationen zwischen mehreren Prädikatoren, führt zu Problemen bei der automatischen Modellwahl. In schrittweisen Verfahren ist es in solchen Fällen häufig vom Zufall abhängig welche der beteiligten Variablen als erste weggelassen beziehungsweise aufgenommen wird. Ein Anstieg der korrelation zwischen Prädiktoren hat zur Folge, dass (a) die p-Werte bei Signifikanztests sinken, (b) schwache Prädiktoren entgegen den starken eher fälschlicherweise ausgeschlossen werden, (c) korrekt hoch signifikante Prädiktoren werden eher ausgeschlossen wenn die korrelation zwischen einem konfundierenden Prädiktor und dem Kriterium steig und (d) selbst wenn die korrelation zwischen konfundierendem Prädiktor und Kriterium klein ist, besteht die Gefahr, dass korrekt schwach signifikante Prädiktoren nicht signifikant werden (Graham, 2003, p. 2810).

Grundsätzlich gilt die Unabhängigkeit der Prädikatoren als Voraussetzung für die multiple Regression, doch gerade in der psychologischen Forschung lassen sich Korrelationen meist schlecht vermeiden. Um bereits im Vorfeld Hinweise auf potentielle Kollinearität zu erhalten, empfiehlt es sich die Kovarianzmatrix zwischen allen Prädikatoren vor der eigentlichen Auswahl zu betrachten. Eventuell wurde das selbe Merkmal mehrmal erhoben. Der Zusammen hang zwischen den Prädikatorvariablen und der vorhergesagten Kriteriumsvariable lässt sich mit dem

Strukturkoeffizienten c_i ausdrücken. So kann auf die Prädikatoren eingeschränkt werden, welche am besten das Kriterium vorhersagen (Bortz & Schuster, 2011, S. 453).

Überanpassung - Der einfachste Ausdruck verwickelter Probleme ist immer der beste

Der Mensch strebt nach Kontrolle und verfällt dabei oft der Illusion, alles bis ins letzte Detail kontrollieren zu müssen. Entsprechend wurden im Zuge der technischen Revolution hoch komplexe Systeme und Maschinen entwickelt, welche jeder Anwendung gerecht werden sollten. Insbesondere im Bereich des Produktdesigns hat in den letzten Jahren ein radikales umdenken statt gefunden: "Simplicity is about subtracting the obvious, and adding the meaningful." (Maeda, 2006). Einfachheit wurde gelebt was zahlreiche unumstritten grossartige Produkte hervor brachte.

Die Illusion der Komplexität ist auch in der Statistik anzutreffen (Armstrong, 2011, p. 3). Im Kontext der Modellwahl äussert sich dies in Form der Überanpassung, wenn also das Modell zu sehr an die Testdaten angepasst ist. Insbesondere Modelwahlverfahren, welche die Anzahl der Prädikatoren nicht abstrafen, sind davon betroffen. Als Einflussfaktoren seitens der Daten sind Repräsentativität und Stichprobengrösse zu nennen. Mit steigendem Stichprobenumfang und höheren Repräsentativität sinkt die Überanpassung und steigt die stabilität der Vorhersage.

Informationskriterium in schrittweisen Regressionen

EINFüHREND (Weakliem, 2004, p.170)

Das Akaikes Informationskriterium (AIC) und Bayessche Informationskriterium (BIC) basieren auf der Maximum-Likelihood-Methode, Regressionsgleichungen müssen nicht ineinander geschachtelt sein und der Kennwert berücksichtigen die Komplexität des Modells anhand der Prädiktoranzahl. Die beiden Kenn-

werte strafen dem Prinzip der Sparsamkeit entsprechend Komplexität, was der Überanpassung entgegen wirkt.

Das schrittweise Verfahren stoppt, wenn die Grösse AIC / BIC nicht mehr abnimmt.

Kreuzvalidierungsverfahren

Die Generalisierbarkeit eines Modelles lässt sich durch Vergleiche mit neuen, von der Referenzstichprobe unabhängigen, Stichproben ermitteln. Zu diesem Zweck kommen sogenannte Kreuzvalidierungsverfahren zum Einsatz.

Die Idee hinter der Kreuzvalidierung liegt darin, die Daten in eine Referenzstichprobe, anhand der die Gleichung geschätzt wird, und in eine (Validierung) oder mehrere (Kreuzvalidierung) zusätzliche Teststichproben anhand dererer die Stabilität als durchschnittliche Fehlerquote berechnet wird (Arlot & Celisse, 2010, p. 3).

Es frag sich welchen Platz die Kreuzvalidierung in der Modellselektion einnimmt. Kreuzvalidierungen können über ein Set von n Regressionsgleichungen durchgeführt werden, beispielsweise potentielle Modelle nach einer schrittweisen Regression (Arlot & Celisse, 2010, p. 12). n potentiellen Modelle werden validiert, wobei unter Umständen das stabilste nicht gleich dem vielversprechendsten aus der vorangegangenen Selektion ist. Überangepasste Modelle können somit eliminiert werden und an deren Platz rücken einfachere und stabilere Modelle. Der Vorteil dieses Vorgehens liegt darin das (a) die Validierung komplett von der Modellselektion getrennt werden kann und (b) es nur n Durchgänge benötigt. In der Modellselektion wird jedoch die stabilität nicht berücksichtigt. Bei der schrittweisen Regression haben wir gesehen, dass das "beste" Modell nicht zwangsläufig gefunden wird. Entsprechendes gilt für die Stabilität, was zur Konsequenz führen kann, dass zwar gute Modelle gefunden werden, diese jedoch allesamt nicht stabil genug sind oder das stabilste nicht gefunden wird. Arlot und Celisse (2010, p. 12) nennen noch die Möglichkeit die Stabilität in die Modellselektion zu integrieren und als Kriterium zu berücksichtigen. Zu jeder Modellschätzung wird deren Sta-

bilität berechnet und Modelle welche keine genügenden Werte aufweisen werden abgewiesen.

Kreuzvalidierungsverfahren unterscheiden sich in erster Linie anhand der Strategie, mit der die Daten "getrennt" werden. In der Regel wird dafür ein genug grosser Datensatz herangezogen und unterteilt, wobei vorausgesetzt wird, dass die Teilstichproben unabhängig und gleich verteilt sind.

Zusammenfassung / Diskussion

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig ob ich schreibe: »Dies ist ein Blindtext« oder »Huardest gefburn«?. Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muß keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie »Lorem ipsum« dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Literatur

Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Armstrong, J. (2011). Illusions in regression analysis. *Available at SSRN 1969740*.

Bendel, R. B. & Afifi, A. A. (1977). Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical Association*, *72* (357), 46–53.

Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics*, *20* (1), 1–6.

Borges, B., Goldstein, D. G., Ortmann, A. & Gigerenzer, G. (1999). Can ignorance beat the stock market. *Simple heuristics that make us smart*, *59*, 72.

Bortz, J. & Schuster, C. (2011). *Statistik für Human-und Sozialwissenschaftler* (Bd. 6). Springer.

Box, G. E. (1979). *Robustness in the strategy of scientific model building.* (Bericht). DTIC Document.

Derksen, S. & Keselman, H. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45* (2), 265–282.

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium.*

Graham, M. H. (2003). Confronting multicollinearity in ecological multiple regression. *Ecology*, *84* (11), 2809–2815.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis.* Springer.

Jacob Cohen, P. C., West, L. S. A., Cohen, J. & West, S. G. A. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences/.* Mahwah, NJ, USA: Lawrence Erlbaum Associates.

Maeda, J. (2006). The laws of simplicity (simplicity: Design, technology, business, life).

Mundry, R. & Nunn, C. L. (2009). Stepwise model fitting and statistical inference: turning noise into signal pollution. *The American Naturalist*, *173* (1), 119–123.

Thompson, M. L. (1978). Selection of variables in multiple regression: Part I. A review and evaluation. *International Statistical Review/Revue Internationale de Statistique*, 1–19.

Weakliem, D. L. (2004). Introduction to the special issue on model selection. *Sociological Methods & Research; Sociological Methods & Research.*

Anhang

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig ob ich schreibe: »Dies ist ein Blindtext« oder »Huardest gefburn«?. Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muß keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie »Lorem ipsum« dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Selbstständigkeitserklärung

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig ob ich schreibe: »Dies ist ein Blindtext« oder »Huardest gefburn«?. Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muß keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie »Lorem ipsum« dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.