Modellselektion (AIC, BIC, Pseudo R^2...)

Nach einer explorativen Analyse der Daten und der Wahl einer passenden Modellklasse, geht es darum, das bestmögliche Modell zu den vorliegenden Daten zu finden. Daher stellt sich die Frage, was "bestmögliches" Modell bedeutet und wie ein solches bestimmt werden kann. In diesem Zusammenhang wird der Gedanke aufgegriffen, dass mit keinem Regressionsmodell die Realität eins zu eins abgebildet werden kann. Nimmt man zu viele erklärende Variablen auf, läuft man Gefahr, das Modell zu "overfitten" (überanpassen). Ein überangepasstes Modell erklärt die zum Schätzen verwendete abhängige Variable meist sehr gut, schneidet jedoch in der Vorhersage von Daten außerhalb der verwendeten Stichprobe häufig schlecht ab. Auf der anderen Seite kann ein Modell auch "underfitted" sein, d.h. die aufgenommenen unabhängigen Variablen können die abhängige Variable nur sehr unzureichend erklären.

Das Thema der Modellselektion ist ein allgegenwärtiges in der Statistik/Regressionsanalyse. Dennoch gibt es keine absoluten, objektiven Kriterien, anhand derer entschieden werden kann, ob das eine oder das andere Modell gewählt werden sollte. Vielmehr existieren viele verschiedene Verfahren, die versuchen, zwischen möglichst viel Erklärungsgehalt des Modells und möglichst wenig Komplexität (siehe dazu Ockhams Rasiermesser) abzuwägen.

AIC (Akaike-Information-Criterion)

Das AIC dient dazu, verschiedene Modellkandidaten zu vergleichen. Dies geschieht anhand des Wertes der log-Likelihood, der umso größer ist, je besser das Modell die abhängige Variable erklärt. Um nicht komplexere Modelle als durchweg besser einzustufen, wird neben der log-Likelihood noch die Anzahl der geschätzten Parameter als Strafterm mitaufgenommen.

\[AIC(P)=-2\hat{l}_P+2|P|\]

In der Formel steht \(P\) für die Anzahl der im Modell enthaltenen Parameter und \(\hat{l}\) für den Wert der log-Likelihoodfunktion. Das Modell mit dem kleinsten AIC wird bevorzugt.

Das AIC darf nicht als absolutes Gütemaß verstanden werden. Auch das Modell, welches vom Akaike Kriterium als bestes ausgewiesen wird, kann eine sehr schlechte Anpassung an die Daten aufweisen. Die Anpassung ist lediglich besser als in den Alternativmodellen.

Im klassischen Regressionsmodell unter Normalverteilungsannahme der Störterme kann das AIC auch folgendermaßen dargestellt werden:

\[AIC(P)=n\ln(\hat{\sigma }^{2})+2|P|\]

\(\hat{\sigma }^{2}\) steht für die geschätzte Varianz der Fehler \(\epsilon_i\) des Modells. Mithilfe des Akaike Informationskriteriums können auch nicht geschachtelte Modelle, d.h. Modelle mit unterschiedlichen erklärenden Variablen verglichen werden. Bei geschachtelten Modellen hingegen finden sich alle Prädiktoren eines kleineren Modelles in größeren Vergleichsmodellen wieder. Berechnet man das AIC von Modellen mit gleicher Parameterzahl, so entspricht die Auswahl nach dem kleinsten AIC der Auswahl nach der kleinsten Residuenquadratsumme.

BIC (Bayesian-Information-Criterion)

Das BIC (auch SIC, Schwarz Information Criterion, genannt) ist dem AIC sehr ähnlich. Zur Bewertung der Modellgüte wird der Wert der log-Likelihood herangezogen. Davon wird als Strafterm die Anzahl der geschätzten Parameter multipliziert mit dem natürlichen Logarithmus der Anzahl der Beobachtungen abgezogen. Im Gegensatz zum Akaike Kriteriuim passt sich der Strafterm an die Größe der Stichprobe an. Schon ab einer Stichprobengröße von acht \((\ln(8)=2,07944>2)\) bestraft das BIC komplexere Modelle stärker als das AIC.

\[BIC(P)=-2\hat{l}_P+|P|\ln(n)\]

In der Formel steht \(P\) für die Anzahl der im Modell enthaltenen Parameter und \(\hat{l}\) für den Wert der log-Likelihoodfunktion. Das Modell mit dem kleinsten BIC wird bevorzugt. Auch für das BIC gilt, dass das Modell mit dem kleinsten Wert des Informationskriteriums eine bessere Anpassung aufweist als die Alternativmodelle. Dennoch kann der Gesamterklärungsgehalt des Modells gering sein.

Im klassisches Regressionsmodell unter Normalverteilungsannahme der Störterme kann das BIC auch folgendermaßen dargestellt werden:

$$BIC(P)=n\ln(\hat{\sigma }^{2})+|P|\ln(n)$$

$\hat{\sigma }^{2}$ steht für die geschätzte Varianz der Fehler ϵ_i des jeweiligen Modells. Mithilfe des Bayesianischen Informationskriteriums können anders als beispielsweise mit dem Likelihood-Quotiententest oder F-Test nicht nur geschachtelte, sondern auch nicht geschachtelte Modelle, d.h. Modelle mit unterschiedlichen Prädiktoren verglichen werden. Berechnet man das BIC von Modellen mit gleicher Parameterzahl, so entspricht die Auswahl nach dem kleinsten BIC der Auswahl nach der kleinsten Residuenquadratsumme.

In der Praxis finden beide Auswahlkriterien Anwendung und werden oft sogar zusammen verwendet. Insgesamt ist das AIC jedoch gebräuchlicher als das BIC.

McFaddens Pseudo R^2

Im Fall einer metrischen abhängigen Variable in einem linearen Regressionsmodell werden zur Bewertung der Modellgüte oft die Bestimmtheitsmaße R^2 und adjustiertes $\bar{R}^2$ herangezogen. Bei Modellen mit nominal- oder ordinalskalierten abhängigen Variablen gibt es keine direkte Entsprechung, da die Varianzzerlegung und somit das R^2 nicht berechnet werden könnnen. Aus diesem Grund gibt es verschiedene Pseudo-Bestimmtheitsmaße (allgemein Pseudo-R^2) mit unterschiedlichen Ansätzen. Sie sind so konstruiert, dass sie dem üblichen Bestimmtheitsmaß in Interpretation und Anwendung ähneln. Die Werte der Pseudo-Bestimmheitsmaße sind auf den Bereich 0 bis 1 festgelegt, wobei ein Wert nahe 1 auf eine bessere Modellanpassung hinweist als ein Wert nahe 0.

Hier wird das auf der log-Likelihood basierende McFaddens Pseudo R^2 vorgestellt, da es in der Praxis oft Anwendung findet:

$$R_{\text{McFadden}}^2=1-\frac{\ln(L_1)}{\ln(L_0)}$$

McFaddens korrigiertes Pseudo R^2

$$\bar{R}_{\text{McFadden}}^2=1-\frac{\ln(L_1)-k}{\ln(L_0)}$$

$\ln(L_1)$ steht für die log-Likelihood des geschätzten Modells mit erklärenden Variablen, $\ln(L_0)$ ist die log-Likelihood des Nullmodells, das nur eine Konstante enthält, und k gibt im adjustierten $\bar{R}^2$ die Anzahl der im Modell enthaltenen unabhängigen Variablen an.

Neben diesem Bestimmheitsmaß gibt es noch weitere, die von gängigen Statistikprogrammen im Standardoutput angegeben werden. Beispiele hierfür sind die R^2 von Nagelkerke und von Cox & Snell. Bei diesen beiden Maßen wird zur Berechnung auf die Likelihood zurückgegriffen, bei McFadden hingegen die log-Likelihood verwendet. Da die Werte innerhalb eines Modells stark variieren können, sollten die unterschiedlichen Bestimmheitsmaße für verschiedene Stichproben nicht verglichen werden.

Vorwärts- und Rückwärtsselektion

Ein Problem, das das Ergebnis von Vorwärts- und Rückwärtsselektion stark beeinflussen kann, ist das Vorliegen von Kollinearität. Sind Kovariate zu stark korreliert, kann das dazu führen, dass ihr Einfluss auf die erklärte Variable durch das verwendete Modellwahlkriterium nicht erkannt wird. Sie würden dem zufolge möglicherweise nicht in das Modell aufgenommen bzw. daraus entfernt werden. Daher ist es wichtig, Daten auf Kollinearität zu prüfen, bevor ein Verfahren zur Variablenselektion eingesetzt wird.

Vorwärtsselektion

1. Als Grundmodell wird das "kleinstmögliche" Modell, bestehend nur aus einer Konstanten verwendet.
2. Anschließend wird die Variable in das Modell übernommen, welche die größte Verbesserung bringt. Die Verbesserung wird anhand eines vorher festgelegten Modellwahlkriteriums gemessen. Üblicherweise werden dazu AIC, BIC oder sogar beide Kriterien verwendet.
3. In den Folgeschritten wird jeweils eine erklärende Variable zusätzlich aufgenommen und das Modell somit schrittweise komplexer. Dabei folgt man immer dem in Schritt 2 festgelegten Verfahren und wählt die Einflussgröße, mit der AIC oder BIC minimal werden.
4. Stoppregel: Das Vorgehen wird dann beendet, wenn durch die Aufnahme weiterer Kovariaten keine Verbesserung des Modellwahlkriteriums mehr erreicht werden kann.
5. Das resultierende Modell kann dann zur weiteren Analyse verwendet werden.

Rückwärtsselektion

1. Als Ausgangsmodell wird bei der Rückwärtsselektion das volle Modell gewählt, d.h. es werden alle zur Verfügung stehenden erklärenden Variablen aufgenommen
2. In den folgenden Schritten wird jeweils die Kovariate entfernt, die den schlechtesten Wert des Modellwahlkriteriums (z.B. höchstes AIC, BIC) liefert.
3. Stoppregel: Das Vorgehen wird dann beendet, wenn durch Herausnahme einer weiteren Einflussgröße keine Verbesserung im Wert des Auswahlkriteriums mehr erzielt werden kann.
4. Das ermittelte Modell ist das beste im Sinne des jeweils verwendeten Kriteriums und kann anschließend zur Schätzung/Prognose verwendet werden.

Die beiden Methoden der Vorwärts- und Rückwärtsselektion können auch kombiniert durchgeführt werden. Hierbei wird bei jedem Schritt des Verfahrens sowohl auf das Hinzufügen, als auch auf das Entfernen von Variablen aus dem Modell getestet. Es kann dazu kommen, dass die Methoden unterschiedliche Ergebnisse liefern. Diese können wieder anhand von Auswahlkriterien verglichen werden.

Oft wird bei den vorgestellten Methoden gleichzeitig mit dem Akaike und dem Bayesianischen Informationskriterium gearbeitet. Sie unterscheiden sich durch den Strafterm für Modellkomplexität und können somit zu unterschiedlichen Ergebnissen kommen. Deshalb muss im Einzelfall entschieden werden, welches Kriterium sinnvoller ist. Die verschiedenen Vorgehensweisen der Variablenselektion sind in modernen Statistikprogrammen im Standardpaket enthalten und müssen also nicht von Hand durchgeführt werden.

Betrachtung aller möglichen Modelle

Mit den Verfahren der Vorwärts- und Rückwärtsselektion werden maximal $k(k-1)/2$ Modelle (bei k vorliegenden möglichen Prädiktoren) verglichen. Es gibt also keine Garantie, das beste Modell zu finden. Als Alternative zum schrittweisen Vorgehen bietet sich die Methode an, alle 2^k möglichen Modelle zu vergleichen. Es wird das Modell identifiziert, welches das gewählte Anpassungsmaß maximiert oder das Informationskriterium minimiert.