Checkliste zur Interpretation von GLMs und Statistik allgemein

Prof. Dr. Roland Schäfer (FSU Jena)

Die besonders wichtigen Dinge (im Sinne einer Chekliste, die Sie in der Klausur abarbeiten sollten) sind grün markiert. Der Rest sind Hinweise und Hilfestellungen. Nur ausdrücklich als optional markierte Punkte können Sie ignorieren, wenn Sie möchten.

Allgemeine Darstellung von Modellen (GLMs)

Theoretischer Rahmen (vor allem Schäfer 2020)

1. Was macht probabilistische Grammatik „probabilistisch“? Was ist der Unterschied zu einer klassischen (aristotelischen, algebraischen, diskreten, …) Grammatik?
2. Wie und wieso werden im Rahmen von probabilistischen Modellen sog. grammatische Zweifelsfälle als Alternationen modelliert?
3. Spezifisch für meine Ansätze: Was ist eine Prototypentheorie mit gewichteten Merkmalen (*cue validities*), und was leistet sie im Rahmen eines probabilistischen Modells?

Beschreibung des konkreten Phänomens (Schäfer 2018, 2019; Schäfer & Sayatz 2016)

1. Um welches grammatische Phänomen geht es? Bitte möglichst mit einfachen, gut nachvollziehbaren Beispielen für die Alternanten (z.B. starke/schwache Flexion bei schwachen Maskulina ; *obwohl* mit V2 und verschiedenen Interpunktionsmöglichkeiten ; Maßangabe mit Apposition und Pseudo-Partitiv).
2. Wie wird das Phänomen modelliert? Konkret z.B.: Wie werden die Prototypen für die (meistens zwei) Alternanten definiert? Aus welchen Theorien (z.B. Grammatikalisierungsmodelle, Interpunktionstheorien) werden die Einflussfaktoren abgeleitet?
3. Was sind die Einflussfaktoren (= die unabhängigen Variablen = die Regressoren) und wie werden sie aus zu dem/den Prototypen in Beziehung gesetzt? Oft geschieht Letzteres direkt. Z.B.: Die paradigmatischen Verhältnisse machen eine Vorhersage darüber, ob in bestimmten Kasusformen schwache Maskulina erwartungsgemäß häufiger stark flektiert werden als in anderen. Der Regressor „Kasus“ ergibt sich damit direkt. Manchmal geht das nicht so einfach: Bei den Maßangaben wird zum Beispiel der Einflussfaktor „Standardsprachlichkeit“ mit anderen konkreten Regressoren nur approximiert, weil es kein vernünftiges Maß für Standardsprachlichkeit gibt.

Interpretation des Modells

Beschreibung der Modellstruktur

1. Welche Regressoren sind im Modell enthalten? Konzentrieren Sie sich in der Vorbereitung auf nominal skalierte Einflussfaktoren. Numerische bzw. intervallskalierte Regressoren spielen in der Prüfung keine Rolle.

2. Evtl. ein kurzes allgemeines Wort zu dummy-kodierten Regressoren. Warum gibt es z.B. für den konzeptionell drei- oder vierwertigen Regressor Kasus zwei, drei oder vier einzelne Regressoren im Modell? (Das steht auf den Folien, aber auch in meinem Handbuch-Artikel zu gemischten Modellen.)
3. Was liegt auf dem Intercept? Speziell bei meinen Modellen, weil sie entsprechend von mir gebaut werden: Inwiefern modelliert der Intercept den (proto-)typischen Fall?
4. Hierarchische/Gemischte Modelle haben wir nicht ausreichend besprochen. Dementsprechend entfällt die Beschreibung dieses Teils der Modellstruktur.

Interpretation der geschätzten Parameter

1. Welche Präferenz (für welche der Alternanten) modelliert der Intercept? Hierbei ist nur die Interpretation des Vorzeichens zielführend, weil (wie besprochen) der Intercept und alle anderen Koeffizienten auf der Logit-Skala liegen, deren Werte bzw. Größenordnung bezüglich ihres Einflusses auf die Wahrscheinlichkeit nicht linear interpretierbar sind. Für den Intercept gilt: Wenn er 0 ist, modelliert er eine Wahrscheinlichkeit von exakt 0.5 für das Ereignis (= die modellierte Alternante, z.B. starke Flexion eines schwachen Maskulinums). Das entspricht linguistisch einer völlig freien Alternation, also einer Chance von 1:1 für beide Alternanten.[1]
2. **Optional** (wirklich nur für Bonuspunkte): Woran liegt das? Um welchen Wert sind die Logits zentriert, und um welchen Wert sind die Wahrscheinlichkeiten zentriert? Warum müssen die Logits transformiert werden, um Wahrscheinlichkeiten zu ergeben? Wieso sind die Logits vor der Transformation bezüglich ihres Einflusses auf die Wahrscheinlichkeit nicht linear interpretierbar? Dazu ist die Betrachtung der Kurve der inversen Logit-Funktion zielführend (s. meine Folien und Backhaus et al.).
3. Allgemeine Beschreibung, was Koeffizienten und ihre zugehörigen Standardfehler, z-Werte und p-Werte sind.
4. Beachten Sie: Die Nullhypothese für den Test jedes Koeffizienten ist, dass er 0 ist (H0: $\beta=0$). Wir testen also, ob der zugehörige Regressor überhaupt einen Effekt hat oder nicht. Für den Intercept gilt das Gleiche. Die Nullhypothese ist, dass er 0 ist, was der in Punkt (1) beschriebenen völlig freien Alternation entspricht.
5. Für jeden Koeffizienten testen wir unter der Annahme, dass der gesamte Rest des Modells gleich bleibt und korrekt ist. Das ist natürlich problematisch, weil wir nicht sicher sein können, ob der Rest korrekt ist. Es gibt Möglichkeiten, das etwas besser abzusichern, aber diese Möglichkeiten haben wir nicht besprochen. Behalten Sie das aber evtl. im Hinterkopf, falls Sie irgendwann nochmal etwas mit solchen Modellen zu tun haben.
6. Auf der allgemeinen Beschreibung basierend: Welche Koeffizienten wurden signifikant getestet? (Wenn *sig* nicht spezifiziert ist, gilt *sig=0.05* als Standard.)
7. **Optional**: Wenn für die geschätzten Koeffizienten ein Konfidenzintervall (im Standardfall ein 95 %-Intervall) angegeben wurde, entspricht es einem erfolgreichen (signifikanten) Test (im Standardfall bei *sig=0.05*), wenn das Konfidenzintervall die 0 nicht einschließt. Das KI sagt uns, in welchem Bereich 95 % aller Werte liegen, wenn der Koeffizient den geschätzten Wert hat. Schließt es 0 mit ein, ist 0 als wahrer Wert des Koeffizienten nicht in den extremen 5 % aller Werte. Das ist äquivalent zu einem wie in Punkt (4) beschriebenen Test bei *sig=0.05*, der die Abweichung von der Nullhypothese H0: $\beta=0$ testet.
8. **Optional**: Das Chancenverhältnis (*odds ratio* ; OR) quantifiziert den Einfluss eines Koeffizienten auf die **Chance**, dass das Ereignis eintritt. Für jeden Koeffizienten β gilt $OR(\beta)=e^{\beta}$ (e ist die Eulersche Zahl ; wir benötigen also die ordinäre Exponentialfunktion). Diese Chancenverhältnisse können Sie linear interpretieren. Zum Beispiel: Eine OR von 4 für Dativ gegenüber einer OR von 2 für Genitiv bedeutet, dass sich die Chance, dass das Ereignis eintritt, doppelt so stark erhöht, wenn der Kasus Dativ ist, verglichen mit dem Fall, wenn es ein Genitiv ist. (Dafür können Sie die ORs nicht wie die Logits addieren.) Das ist zwar verglichen mit den Koeffizienten schöner, aber man erkauft es sich um den Preis, dass man über Chancen statt über Wahrscheinlichkeiten redet. Die Chance (*odds* ; *o*) ist das

[1] Das gilt nur, wenn eine echte Zufallsstichprobe genommen wurde. Sobald stratifizierte Stichproben zum Einsatz kommen, kann der Intercept uninterpretierbar werden. Das spielt in der Prüfung keine Rolle, aber denken Sie daran, falls Sie selbst einmal Modelle spezifizieren und interpretieren oder entsprechende Literatur lesen.

Verhältnis von Wahrscheinlichkeit und Gegenwahrscheinlichkeit (bei $p=0.7$ ist $o=0.7/0.3=2.33$; in Worten: es ist 2.33 Mal so wahrscheinlich, dass das Ereignis eintritt wie dass es nicht eintritt). Eine OR ist damit ein Verhältnis zweier Verhältnisse. Und wenn Sie zwei ORs vergleichen, reden Sie über ein Verhältnis eines Verhältnisses eines Verhältnisses. Wie intuitiv Sie das finden, überlasse ich Ihnen. Auf den Folien und bei Backhaus et al. wird es erklärt.

9. Nur für signifikant getestete Koeffizienten: In welche Richtung verschieben die einzelnen Koeffizienten die Wahrscheinlichkeit relativ zum Intercept? Hier ist wieder nur das Vorzeichen allgemein interpretierbar. Wenn Sie allerdings z.B. einen Koeffizienten für Dativ und einen für Akkusativ haben, und einer der beiden ist größer als der andere (z.B. für Akkusativ 1.0 und für Dativ 2.0), dann modelliert der größere durchaus einen stärkeren Einfluss auf die geschätzte Wahrscheinlichkeit. (Im Beispiel: Der Dativ erhöht die Wahrscheinlichkeit, dass das modellierte Ereignis eintritt stärker als der Akkusativ.) Sie können das allerdings nur als größer bzw. kleiner interpretieren, nicht linear als „doppelt so groß“ usw.
10. Passen die geschätzten Koeffizienten zu den theoretischen Vorhersagen? In einigen meiner Artikel gibt es dazu noch weitere Tabellen und Grafiken, die die Ergebnisse des Modells ggf. übersichtlicher Darstellen, und die Sie sich ansehen sollten.

Hinweise zur Interpretation von statistischen Tests

Diese Hinweise betreffen sowohl einzelne Testverfahren wie den t-Test oder den χ^2-Test als auch Tests im Rahmen der Modellinterpretation (dort bei GLMs meistens z-Tests).

1. Das Wort „beweisen“ ist im Kontext statistischer Inferenz **immer** fehl am Platz! Mit statistischen Verfahren versuchen wir, Nullhypothesen zurückzuweisen, indem wir eine zufriedenstellende Abweichung von der Nullhypothese in den Daten finden. Zufriedenstellend sind Abweichungen, die mit dem angemessenen Signifikanzniveau (Standardfall *sig=0.05*) getestet werden.
2. **Signifikanz ist immer und ausschließlich eine Eigenschaft eines Tests!** In der wirklichen Welt gibt es einen Effekt, oder es gibt ihn nicht, und es gibt also keine „signifikanten Effekte“ (nur existierende und nicht existierende). Wenn der Test ein signifikantes Ergebnis hat, dann gehen wir (bis auf Weiteres) davon aus, dass es einen Effekt gibt. (Wir haben ein gewisses Maß an Evidenz für einen Effekt gefunden.) Bewiesen haben wir nichts, denn wir können jederzeit auch mit ein bisschen Pech ein falsch-positives Ergebnis erzielt haben.
3. Wir errechnen mit dem Test **nicht** die Wahrscheinlichkeit, dass wir uns geirrt haben! Wenn wir die H0 zurückweisen, obwohl sie korrekt ist, haben wir uns mit $p=1$ geirrt. (= Ein eingetretenes Ereignis hat die Wahrscheinlichkeit 1.) Wenn wir sie zurückweisen, wenn sie nicht korrekt ist, haben wir uns mit $p=0$ (also nicht) geirrt. (= Ein nicht eingetretenes Ereignis hat die Wahrscheinlichkeit 0.)
4. **Wir wissen nach dem Test nicht, wie wahrscheinlich es ist, dass es einen Effekt gibt!**
5. Der p-Wert sagt uns, **wie wahrscheinlich es vor dem Experiment war**, das im Experiment dann tatsächlich gefundene konkrete Ergebnis (oder ein noch stärker von der Nullhypothese abweichendes Ergebnis) zu erhalten, falls die Nullhypothese stimmt. Insofern sagt uns der p-Wert, wie überraschend ein solches Ergebnis wäre, falls die Nullhypothese stimmt. (Denken Sie an die Tea Tasting Lady, denn dieses Beispiel illustriert sehr gut den Inferenzprozess.)
6. Eine Testung mit erfolgreich erreichtem *sig*-Niveau kann so interpretiert werden, dass die Nullhypothese nicht korrekt ist **oder ein hinreichend seltenes Ereignis eingetreten ist**.
7. Achtung! Bei *sig=0.05* reicht $p=0.05$. Das heißt, dass in einem von zwanzig Fällen auch bei korrekter Nullhypothese (= kein Effekt) ein solches oder ein extremeres Ereignis eintritt. So extrem selten ist das auch wieder nicht. (Daher plädieren viele auch für niedrigere *sig*-Niveaus. In der Kernphysik gilt 5σ statt wie in vielen Wissenschaften üblich 1.96σ bzw. 2σ, also fünf Standardfehler statt knapp zwei. Das entspricht *sig=0.0000015*.)
8. Da wir aber nicht wissen, ob es nicht doch ein seltenes Ereignis war, wissen wir über die Realität nicht viel mehr als vorher. Wir arbeiten nur erst einmal unter der Annahme weiter, dass die Nullhypothese nicht stimmt. Erst eine lange Serie von wiederholten Testungen führt dazu, dass sich substantielle Hypothesen (z.B. „Kognitive Grammatiken sind probabilistisch.“) als sehr gut gesichert etablieren. **Wie gut** gesichert die Hypothese ist, kann und will man in frequentistischer Statistik nicht quantifizieren.

In Bayesianischer Statistik (basierend auf, aber nicht „nach“ Thomas Bayes) kann man das angeblich, aber das ist umstritten, und ich bin nicht überzeugt davon.

9. Der p-Wert wird bei gleichbleibenden Unterschiedsmaßen mit steigender Stichprobengröße kleiner. Z.B. beim t-Test für zwei Stichproben: Wenn die gemessene Differenz der Mittelwerte gleich bleibt, aber die Stichprobengröße steigt wird, wird der p-Wert kleiner. (= Es gibt mehr Evidenz dafür, dass der Mittelwertunterschied tatsächlich existiert, weil es mit größeren Stichproben leichter ist, existierende Unterschiede zu finden.)
10. Der p-Wert wird bei gleichbleibender Stichprobengröße und steigendem Unterschiedsmaß kleiner. Z.B. beim t-Test für zwei Stichproben: Wenn die Stichprobengröße gleich bleibt, aber der gemessene Unterschied der Mittelwerte größer wird, wird der p-Wert kleiner. (= Es gibt mehr Evidenz dafür, dass der Mittelwertunterschied tatsächlich existiert, weil ein größerer Unterschied unabhängig von der Datenmenge leichter zu finden ist.)
11. Für Tests, die von einem Standardfehler abhängen (z.B.: z-Test, t-Test) gilt: Bei gleicher Stichprobengröße und gleichem Unterschiedsmaß (z.B. Differenz der Mittelwerte) steigt der p-Wert mit steigender Varianz, weil mit steigender Varianz der Standardfehler größer wird. (= Die Stichproben weichen bei gleicher Größe im Mittel mehr vom wahren Wert ab.)

Sonstige Hinweise zur Prüfung

1. Für die Aufwärmfragen: Schauen Sie sich z-Test, t-Test mit einer und zwei Stichproben, χ^2-Test, Fisher-Test und zumindest konzeptionell die einfache ANOVA an. Die Berechnung des z- und t-Tests (auch mit zwei Stichproben) sollten Sie **bei gegebenem Standardfehler** auswendig kennen (= die Formel in Form eines sehr einfachen Bruchs). Darüberhinaus rechnen Sie bitte nicht damit, irgendwas selber berechnen zu müssen. Sie sollten für die genannten Tests aber wissen, was man mit ihnen testet (z.B. Mittelwertunterschiede, Unterschiede in Zähldaten), und wie man die Ergebnisse interpretiert.
2. Merken Sie sich die Zahl 1.96 und ihre Bedeutung im Rahmen von Tests, die auf Normalverteilungen und verwandten Verteilungen (wie t) basieren. „Rechnen“ können Sie auch mit $2 \approx 1.96$.