# Errata (und Veränderungen zur nächsten Auflage) zu:
# Parametrische Statistik
# –
# Verteilungen, *maximum likelihood* und GLM in R
# 2te Auflage

Carsten F. Dormann

Biometrie & Umweltsystemanalyse
Universität Freiburg

17. Februar 2024

## 1 Unstatistisches

Im Vorwort, Seite V, habe ich den Satz „Traue keiner Statistik, die Du nicht selbst gefälscht hast“ unkritisch aus meinem Gedächtnis Churchill zugeschrieben. Das ist zwar üblich, aber falsch. Tatsächlich gibt es keinen Hinweis darauf, dass Churchill das jemals gesagt hat, noch findet es sich in seinen Schriften.

Glücklicherweise gibt es auch keine Belege, dass Goebbels diesen Satz Churchill in den Mund gelegt hat, auch wenn das vorgeschlagen wurde. Weitere Details hierzu unter `https://de.wikipedia.org/wiki/Liste_gefl%C3%BCgelter_Worte/T#Traue_keiner_Statistik,_die_du_nicht_selbst_gef%C3%A4lscht_hast`. Vielen Dank an Thomas Hestermann für diesen Hinweis!

## 2 Tippfehler

Davon gibt es, leider, immer zu viele, um sie hier einzeln aufzuführen.

Dank vor allem an Juliane Wüllenweber, die alleine auf 20 Seiten solche Tippfehler (und manchmal mehr) gemeldet hat. (Interessanterweise sind die nicht gleichverteilt übers Buch. Liegt das daran, wie ich gearbeitet habe, oder daran, wie Frau Wüllenweber gelesen hat? Wer weiß.)

## 3 Inhaltliche Fehler/Korrekturen

**Einleitung, S. xiii f.** Der Editor JGR lässt sich inzwischen deutlich einfacher aufrufen, ohne Installationen, direkt mit R-Paketen. Entsprechend habe ich diesen Abschnitt überarbeitet. Zudem benutzt JGR jetzt **ggplot2** statt bisher eigene plot-Befehle, was es natürlich attraktiver macht.

**Tabelle 1.1, S. 2** In der Tabelle der Eschenumfänge sind Zeilen und Spalten vertauscht. Die Tabelle muss richtig so aussehen:

| Gruppe 1 | Gruppe 2 | Gruppe 3 | Gruppe 4 | Gruppe 5 | Gruppe 6 | Gruppe 7 | Gruppe 8 | Gruppe 9 | Gruppe 10 |
|---|---|---|---|---|---|---|---|---|---|
| 38 | 37 | 40 | 51 | 12 | 59 | 59 | 96 | 25 | 46 |
| 51 | 41 | 77 | 38 | 74 | 50 | 45 | 52 | 70 | 58 |
| 32 | 79 | 46 | 61 | 17 | 44 | 71 | 59 | 54 | 35 |
| 53 | 37 | 46 | 20 | 41 | 31 | 68 | 48 | 21 | 67 |
| 37 | 46 | 48 | 56 | 28 | 20 | 38 | 43 | 60 | 26 |
| 85 | 42 | 64 | 19 | 35 | 28 | 94 | 49 | 47 | 29 |
| 64 | 44 | 49 | 44 | 92 | 70 | 45 | 66 | 79 | 79 |
| 68 | 44 | 49 | 39 | 50 | 38 | 22 | 92 | 35 | 30 |
| 58 | 60 | 36 | 37 | 24 | 22 | 47 | 71 | 41 | 58 |
| 84 | 46 | 44 | 49 | 17 | 87 | 54 | 74 | 31 | 42 |

Vielen Dank an Joana Ries für die Korrektur!

**Abschnitt 3.3.3, S. 49** Oberste Formel: Dort ist beim Ableiten das $-$ verloren gegangen. Das ist zwar irrelevant für die Herleitung, da ja das Produkt $0 = -0$ ergeben muss, aber es fehlt halt trotzdem: $\frac{d\,\ell}{d\mu} = 0 = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$.

**Abschnitt 3.4.10, S. 62** Der Korrekturfaktor für die Beschränkung ist falsch. Es muss heißen:

$$\text{trPois}(k > 0; \lambda) = \text{Pois}(k > 0; \lambda) \cdot (1/1 - \text{Pois}(k = 0; \lambda)) = \text{Pois}(k > 0; \lambda) \cdot \frac{1}{1 - e^{-\lambda}}$$

Vielen Dank an Dr. Susanne Knies, Freiburg, für die Korrektur!

**Abschnitt 6.1, S. 101** Der Satz "Mittels der Option `continuity=T` (für Yates' *continuity correction*) kann man den korrigieren lassen." hat dort nichts zu suchen (er bezieht sich auf den $\chi^2$-Test später)! Er muss ersatzlos gestrichen werden. Die Fußnote mit Verweis auf Datenbindungskorrekturen in zwei anderen Paketen ist hingegen korrekt.

**Abschnitt 8.3, S. 119** **VGAM** gibt inzwischen P-Werte an, die ich entsprechend jetzt eingefügt habe. Zudem wird eine „Hauck-Donner"-Diagnostik durchgeführt, die ich (S. 120) erkläre: „Ein „Hauck-Donner-Effekt" in der letzten Zeile des **VGAM**-outputs würde die Gültigkeit der P-Werte in Frage stellen. Er tritt u.a. auf, wenn binäre Daten perfekt vom Modell in 0- und 1-Fits separiert werden können, oder wenn ein Parameter am Rande des Definitionsbereichs liegt (etwa eine Standardabweichung von 0)."

**Abschnitt 16.1.2, Seite 257 f.** Was immer mich geritten haben mag, hier den Schätzer für die Steigung als „*est*" zu notieren! Das habe ich jetzt ersetzt durch „$\widehat{\beta}$", ebenso auf der nächsten Seite und in der Fußnote.

# 4 Neue Abschnitte/Inhalte

## 5.2.1 Der $\chi^2$-Test und die hypergeometrische Verteilung

Die hypergeometrische Verteilung beschreibt die Wahrscheinlichkeit, $n$ rote Kugeln ohne Zurücklegen aus einer Urne mit $N$ Kugeln zu ziehen, wenn $M$ Kugeln rot sind. Das entspricht einer Zeile oder Spalte unser obigen Kontingenztabelle: wir ziehen Raucher aus der Gesamtheit, z.B. $n = 1$ männlicher Raucher aus einer „Urne" mit $N = 53$ Personen, von denen $M = 15$ Männer sind. Diese Wahrscheinlichkeit wird durch die hypergeometrische Verteilung beschrieben.

$$P(X = k|M, N, n) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \tag{1}$$

Entsprechend können wir die erwartete Wahrscheinlichkeit, genau einen männlichen Raucher anzutreffen, wenn Männer so häufig rauchen wir Frauen, direkt ausrechnen. Dabei bilden dann Männer plus Frauen die Grundgesamtheit, aus der wir die Stichprobe der Männer ziehen. Es sind also $N = 53$ (Anzahl Personen in der Grundgesamtheit), $M = 13$ (Anzahl RaucherInnen) und $n = 15$ (Anzahl „gezogener“ Männer). $P(X = 1|M = 13, N = 53, n = 15) = \frac{\binom{13}{1}\binom{53-13}{15-1}}{\binom{53}{15}} = 0.048$.

Der $\chi^2$-Test gibt die Wahrscheinlichkeit an, die in der $2 \times 2$-Tabelle gefundenen Werte (oder extremere) zu beobachten, während die hypergeometrische Verteilung die Wahrscheinlichkeit jedes einzelnen Zellenwertes berechnet. Sie sind also eng verwandt, aber quantifizieren unterschiedliche Zielgrößen; sie sind also, wie man sagt, unterschiedliche Statistiken.

Mir sind keine Beispiele aus der Umweltforschung gegenwärtig, in der diese Verteilung eine prominente Rolle spielte. Wichtiger ist sie für Lottospieler, denn sie beschreibt die Wahrscheinlichkeit, $k = 0, 1, 2, \ldots$ „Richtige“ bei der Ziehung von 6 aus 49 zu haben.

### 6.4.1 Die hypergeometrische Verteilung

Wenn wir die Wahrscheinlichkeit einer einzelnen Zelle ausrechnen wollten, dann würden wir uns der hypergeometrischen Verteilung bedienen. Leider variieren die Buchstaben und sogar die Bedeutung der Verteilungsparameter von Buch zu Buch. Die hier vorgestellte Wahrscheinlichkeitsfunktion folgt der von Wikipedia,[1] während die entsprechende Funktion in R anders gestrickt ist. Hier deshalb die Übersetzungshilfe:

$$P(X = k|M, N, n) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \tag{2}$$

wird in R zu:

```
> dhyper(x=1, m=13, n=40, k=15)

[1] 0.04826773
```

Dabei ist $x \mathrel{\widehat{=}} k$, $m \mathrel{\widehat{=}} M$, $n \mathrel{\widehat{=}} (N - M)$ (!) und $k \mathrel{\widehat{=}} n$.

Zur Not können wir diese Formel auch mit den Binomialkoeffizienten direkt ausrechnen:

```
> choose(13, 1) * choose(53-13, 15-1) / choose(53, 15)

[1] 0.04826773
```

Und um das Lottobeispiel aus dem vorigen Kapitel aufzugreifen: eine Zahl richtig zu haben ist wahrscheinlich:

```
> dhyper(1, 6, 43, 6)

[1] 0.4130195
```

Alle 6 Zahlen richtig zu tippen hingegen unwahrscheinlich:

```
> dhyper(6, 6, 43, 6)

[1] 7.151124e-08
```

also etwa 1 zu 14 Millionen. Solange aber mehrere Millionen Menschen jede Woche Lotto spielen, wird es auch immer wieder glückliche Gewinner geben.

[1] https://de.wikipedia.org/wiki/Hypergeometrische_Verteilung

# 5 Exkurs: Die $\chi^2$-Verteilung

[Neue Box in Kapitel 11, gegen Ende] Die $\chi^2$-Verteilung (mit Parameter „Freiheitsgrade“ $= n$) ist von zentraler Bedeutung in der schließenden (=P-Werte produzierenden) Statistik. Sie ergibt sich, wenn wir (standard-)normalverteilte Zufallsvariablen quadrieren und summieren. Wieso sollten wir das tun? Nun, das geschieht, wenn wir in einer Regression die Abweichungsquadrate berechnen und summieren. Die SS der ANOVA sind (nach Division durch die Standardabweichung der Residuen) eben gerade quadrierte Zufallsvariablen aus $\mathcal{N}(\mu = 0, \sigma = 1)$.

Mit anderen Worten: Die (standardisierten) Abweichungsquadrate der Regression sind $\chi^2$-verteilt, und entsprechend gibt dieser $\chi^2$-Wert uns Aufschluss über die Qualität der Regression (*goodness-of-fit*): je kleiner der $\chi^2$-Wert, desto besser der Fit. Die Standardisierung erfolgt typischerweise, indem (wie beim $\chi^2$-Test auf Seite **??** dargestellt) die Abweichungsquadrate durch die Erwartungswerte geteilt werden.

Schließlich sei noch der Zusammenhang zwischen F- und $\chi^2$-Verteilung erwähnt. So ist für zwei unterschiedliche SS $r_1$ und $r_2$ (etwa die zwischen und die innerhalb von Gruppen in der ANOVA): $F = \frac{\chi^2(r_1)/df_1}{\chi^2(r_2)/df_2}$. Der F-Wert berücksichtigt damit, dass die Varianz auf Grundlage der Residuen *geschätzt* wird, während er ja bei den anderen GLMs als fixer Wert angenommen wird. Das ist auch der Grund, weshalb für normal- und „quasi“- verteilte Daten die F-Verteilung benutzt wird, und nicht der $\chi^2$-Verteilung.

Außerdem ist die $\chi^2$-Verteilung mit Parameter $n$ ein Spezialfall der $\gamma$-Verteilung: $\chi^2(n) = \gamma(\frac{n}{2}, \frac{1}{2})$.

## 13.2.1 Kochrezept für den Test von Hypothesen

[Fußnote zu „nil null hypotheses“ hinzugefügt][2]

## [In Kapitel 13:] Ziele der Statistik

Nach diesem langen Kapitel zum Thema Testen mag der Eindruck entstanden sein, dass Hypothesentests das Hauptziel der Statistik sei. Dem ist nicht so. Tatsächlich gibt es drei Hauptziele von Statistik: Datenexploration, Hypothesentests und Vorhersagen.

**Datenexploration** Hier ist das Ziel, in den vorliegenden Daten nach möglichen Mustern zu suchen, nach Zusammenhängen oder Auffälligkeiten. Im Kreislauf der wissenschaftlichen Methode hat die Datenexploration den Zweck der induktiven **Theoriebildung**. Die Daten sind wie ein unbekanntest Land, das ich systematisch aber **ergebnisoffen** durchforsche. Aus diesem Bereich stammt die deskriptive Statistik des Kapitel **??**, aber

[2] Nur weil es ein Rezept gibt, bedeutet das nicht, dass das Essen schmeckt. Genauso ist es auch bei Hypothesen. Viele, vermutlich die meisten, publizierten Hypothesen sind trivial, also von vorn herein wahr (oder falsch). Ein häufiger Fall ist der Vergleich von zwei Pflanzen- oder Tierarten hinsichtlich einer Eigenschaft: wir wissen *vor* jedem Experiment, dass die beiden Arten unterschiedlich sind! Hier interessiert uns nicht *ob* ein Unterschied vorliegt, sondern ob dieser Unterschied groß oder klein ist. Das wäre aber eine ganz andere Hypothese, die genauere Erwartungen erfordert.
Ebenso verhält es sich mit Computersimulationen. Wenn wir etwa Populationen mit und eines ohne Dichteabhängigkeit simulieren, dann wissen wir ja bereits, dass wir das Modell geändert haben, und das es deshalb eine Unterschied gibt. Hier ergibt ein Hypothesentest auf Unterschied überhaupt keinen logischen Sinn. (Nebenbei können wir hier auch so viele Simulationen durchführen, dass wir beliebige P-Werte generieren können.) Im Englischen werden solche sinnlosen Nullhypothesen gelegentlich als “nil null hypotheses” bezeichnet (Good & Hardin 2012).

auch die Korrelationen aus Kapitel **??**. Ein P-Wert macht hier keinen Sinn (auch wenn er häufig angegeben wird), da es keine zugrundeliegenden Hypothesen gibt.

**Hypothesentest** Hier ist das Ziel, mit den vorliegenden Daten gezielt vorher formulierte **Hypothesen zu überprüfen**. Die Details haben wir uns ja auf den vorigen Seiten angesehen.

**Vorhersage** Hier ist das Ziel, den Wert der Antwortvariable für neue Werte der Prädiktoren abzuschätzen. Wir unterscheiden zwischen **Interpolation**, d.h. Vorhersagen im Wertebereich der vorliegenden Daten, und **Extrapolation**, d.h. über den Wertebereich der vorliegenden Daten hinaus. Am offensichtlichsten sind Beispiele wie das Interpolieren von Temperaturen zwischen Messstationen (etwa auf Wetterkarten), oder das Extrapolieren des Wetters für die nächsten Tage. Hierbei verlassen wir den Wertebereich der Zeit. Weniger offensichtlich, aber häufiger und generelle, bezieht sich Inter- und Extrapolation aber auf den Prädiktorraum: Wenn wir für einen Halsbandschnäpper der Attraktivität 6 eine Vorhersage machen, obwohl in unserem Datensatz der Maximalwert 5 ist, dann ist das natürlich auch eine Extrapolation. Extrapolationen sind problematisch, da wir nicht wissen können, ob der Zusammenhang zwischen Prädiktor und Antwortvariable außerhalb unser Daten so aussieht wie innerhalb. Interpolation hingegen ist typischerweise konzeptionell unproblematisch.

Alle drei sind seriöse Ziele, allein: sie vertragen sich nicht miteinander. Die Methoden sind (leicht) unterschiedlich, aber vor allem darf ich die drei Ziele nicht am gleichen Datensatz verfolgen.

Vor jeder statistischen Analyse müssen wir uns klar sein, welches dieser drei Ziele wir verfolgen: nur eines geht!

### Datenexploration verbietet Hypothesentests

Ein häufiges, aber statistisch zutiefst unseriöses Vorgehen ist das Vermischen von explorativer und Hypothesen-testender Statistik. Wenn wir in den Daten erst einmal schauen, was wohl so für Muster dort zu finden sind, uns dann eine Hypothese dazu überlegen, und diese dann *mit den gleichen Daten testen*, dann haben wir einen logischen Zirkelschluss gemacht. Im Englischen wird dieser mit *data snooping* bezeichnet. Solches „in die Daten reinschnuppern“ verzerrt die P-Werte dramatisch, da wir ja schon wissen, dass das Muster vorliegt. Das ist als würden wir ein Weihnachtsgeschenk schon vorher aufmachten, um dann zur Bescherung eine „Hypothese“ vorzuschlagen, was wir wohl bekommen.

Es ist verlockend, schon mal grob nach Mustern zu schauen, vor allem wenn es mehrere Hypothesen gibt, die wir mit den Daten testen können. Oder wenn wir eine vorformulierte Hypothese nicht bestätigt finden, aber dadurch feststellen, dass eine andere Idee dafür vielversprechend aussieht. Aber so geht das nicht: Daten stehen für einen Hypothesentest zur Verfügung, danach sind sie „verbrannt“ (für alle, die sie oder diese Ergebnisse schon gesehen haben). Für Statistiker ist das ein alter Hut, und in guten Statistikbüchern auch dargelegt (Harrell 2001).

Die Konsequenz dieser „Datenexploration verbietet Hypothesentest“-Problematik ist vielen Wissenschaftlern nicht bewusst. Wenn ich mit großem Aufwand Daten erhebe, etwa durch ein nationales Monitoring von Fledermäusen, dann ist eine explorative Datenanalyse der Tod für alle nachfolgenden Hypothesentests. Wir dürfen nicht zunächst, explorativ, ein hübsches Muster auftun, etwa dass Fledermäuse gerne Wald mögen, um dann die Hypothese zu tes-

ten, dass sich Wald auf das Vorkommen von Fledermäusen auswirkt.[3] Die Daten sind nach der Exploration für Hypothesentests verloren! Und deshalb ist gerade bei großen Forschungsprojekten so wichtig, dass vernünftige, quantitative Hypothesen formuliert sind, bevor die Daten erhoben werden. Im Nachhinein finden wir immer eine Hypothese, die sich mit den Daten bestätigen lässt!

### Vorhersageverfahren verbieten Hypothesentests

Hypothesentests sind sehr empfindlich, da sie an ganz klare Annahmen geknüpft sind. So impliziert eine (Hypothesen-testende) Regressionsanalyse, dass wir uns genau für diesen im GLM formulierten Zusammenhang interessieren. Das gilt aber nicht für Vorhersagen.

Wir haben bei Vorhersagen nicht den Anspruch, den richtigen, wahren Zusammenhang zwischen Prädiktoren und Antwort zu entdecken, sondern „nur" einen Zusammenhang, der uns gute Vorhersagen macht. Da wir dafür z.B. Prädiktoren ausprobieren können (das schauen wir uns im Kapitel **??** an), betreiben wir de facto eine Art „Vorhersageexploration". Vollkommen klar, dass das resultierende Modell alle möglichen „Hypothesen" implizit getestet hat, und deshalb nicht mehr als Hypothesentest zur Verfügung stehen kann. Um im Bild des Weihnachtsgeschenks zu bleiben wäre dies, als dürften wir das verpackte Geschenk wiegen, abmessen, schütteln, biegen, zu Boden werfen, um dann am Ende die „Hypothese" zu formulieren, dass es wohl Socken sind.

### Vorhersagen

Vorhersagen spielen in diesem Buch keine große Rolle, hauptsächlich deshalb, weil sie wissenschaftlich selten wirklich interessant sind. Vorhersagen werden häufig in einem Ingenieurskontext eingesetzt: Wenn die Wasserflöhe auf eine toxische Substanz reagieren, was ist dann ein sicherer Grenzwert für die Konzentration dieser Substanz im Seewasser? Andererseits sind Vorhersagen öffentlichkeitswirksam und politisch relevant. Deshalb hier ein paar kurze Kommentare zur Statistik der Vorhersage.

Für Vorhersagen gilt *anything goes*. Wir können irgendeinen wilden statistischen[4] Ansatz benutzen, um eine Vorhersage zu machen, mit absurden Prädiktoren und unsinnigen Verteilungsannahmen. Wichtig ist nur, dass die Vorhersage gut ist. Das wiederum schätzen wir durch **Kreuzvalidierung** ab. Das funktioniert, indem wir den Datensatz in $k$ (zumeist 5-10) Teile teilen, dann unsere Analyse mit $k-1$ Teilen durchführen (*training*), und auf den nicht benutzten Teil vorhersagen (*testing*). Das können wir nun $k$-Mal wiederholen, jeweils einen Teil nicht für das Fitten benutzen, sondern um die Güte der Vorhersage zu messen.

Vorhersagen in Zeit und Raum sind problematischer. Das hat zwei Gründe. Zum einen sind Daten in Zeit und Raum selten unabhängig, d.h. wir können uns sehr gut denken, was morgen für Wetter ist, wenn wir das Wetter von heute kennen. Diese „Nicht-Unabhängigkeit" verletzt aber die Annahmen unser *maximum likelihood*-Methode, da wir dort die Wahrscheinlichkeitsdichte je Datenpunkt nur dann aufmultiplizieren dürfen, wenn die Datenpunkte unabhängig voneinander sind (siehe Box **??**)!

Zum Anderen können wir nicht einfach zufällig Datenpunkte zur Validierung in den $k$ Teildatensätzen der Kreuzvalidierung zuweisen. Wenn der Datenpunkt „letzter Dienstag" fürs

[3] Das ist so offensichtlich, und trotzdem findet genau diese Art von „Wissenschaft" fortwährend statt; achte einmal in Publikationen darauf!

[4] Oder auch nicht-statistischen, wie etwas Teeblattlesen oder Tarotkarten; meistens sind diese Methoden nachweislich schlechter. Natürlich nur dann, wenn das System überhaupt zu einem gewissen Teil vorhersagbar ist (Smith & Donoghue 2010).

Trainieren des Modells benutzt wird, dann ist ja klar, dass deshalb auch der Datenpunkt „letzter Mittwoch“ gut vorhergesagt werden kann: der ist ja nicht unabhängig von den Trainingsdaten! Die Konsequenz ist, dass jetzt die Kreuzvalidierung „geblockt“ stattfinden muss (Roberts et al. 2017). Wenn wir das nicht tun, sehen unsere Vorhersagen zwar super aus, treffen aber schlicht nicht zu (Ploton et al. 2020).

Schließlich sind die Methoden, die zur statistischen Vorhersage benutzt werden, häufig aufwändige Computeralgorithmen, die keine strenge parametrische Grundlage haben. Sie werden unter *machine learning* zusammengefasst (Bishop 2007; Kuhn & Johnson 2013). Der Zugang zu diesem fortgeschrittenen Thema ist leichter, wenn wir die traditionellen, parametrischen Hintergrund verstanden haben (Murphy et al. 2011).

## 15.4 Modellselektion

[Kommentar zum *dredging* hinzugefügt:] Schließlich sei nochmals darauf hingewiesen, dass wir mit AIC und Freunden nur Modelle vergleichen können, die die gleiche Antwortvariable haben. Wenn sich der Datenumfang ändert (etwas weil für manche Prädiktoren ein paar Werte fehlen), dann können wir diese Modelle *nicht* mittels AIC vergleichen. Eine Änderung der Verteilung ist aber zulässig: wir können, wenn wir wollten, normal, Poisson und negativ binomiale Verteilung mittels AIC vergleichen, nicht aber mit einer Normalverteilung der log-transformierten Antwortvariable (da diese ja andere, nämlich log-transformierte Werte beschreibt)!

### 16.1.4 Standardisierte Regressionskoeffizienten

[neu:] Können wir an den geschätzten Regressionskoeffizienten nicht auch ablesen, wie wichtig eine Variable ist? Schließlich sollte die Steigung ja umso steiler sein, je wichtiger ein Prädiktor ist. Dafür müssen die Prädiktoren allerdings in ihrem Wertebereich vergleichbar sein. Wenn $x_1$ von 0 bis 1 geht, aber $x_2$ von 4 bis 100, dann würde sich ja der Effekt von $x_2$ auch auf einen größeren Wertebereich verteilen, selbst wenn beide den Wert von $y$ zum Beispiel verdoppeln würden.

Tatsächlich erfolgt der Vergleich des Einflusses über den *standardisierten Regressionskoeffizient*. Ihn zu berechnen geht auf zwei Arten:

1. Indem wir den geschätzten Regressionskoeffizienten $b_j$ für die Variabilität in $\mathbf{y}$ und $\mathbf{x}$ korrigieren. Der standardisierte Regressionskoeffizient $\beta_j$ von Prädiktor $\mathbf{x}_j$ ist dann $\beta_j = b_j \frac{s_{x_j}}{s_y}$.
2. Indem wir *vor* der Analyse, alle kontinuierlichen Prädiktoren standardisieren, d.h. ihren jeweiligen Mittelwert abziehen und durch ihre Standardabweichung teilen: $\mathbf{x}_j^s = \frac{\mathbf{x}_j^s - \bar{x}_j^s}{s_{x_j}}$. Die Schätzer für die Effekte von $\mathbf{x}_k^s$ sind jetzt auch standardisiert und direkt vergleichbar.

Standardisierung ist entsprechend nur für numerische Prädiktoren möglich, nicht für kategoriale. Um eine Vergleichbarkeit mit kategorialen Variablen herzustellen, rät Gelman (2008) zur Standardisierung durch Division durch $\mathbf{2}s_{x_j}$.[5]

Das Standardisieren der Prädiktoren hat einen weiteren, positiven Nebeneffekt. Stellen wir uns einmal vor, im Modell ist neben $x_1$ und $x_2$ auch die Interaktion $\mathbf{x}_{1\times3} = \mathbf{x}_1 \cdot \mathbf{x}_2$.

[5] In seinem Blog-Beitrag, Jahre später, rät Gelman jetzt dazu, binäre Prädiktoren als $-1$ und 1 zu kodieren, und wie gewohnt durch $1s_{x_j}$ zu teilen (`https://statmodeling.stat.columbia.edu/2009/06/09/standardization/`). Egal wie, der Vergleich zwischen numerischen und kategorialen Prädiktoren bleibt lästig.

Diese wird ja als Produkt der Werte berechnet, und entsprechend werden die resultierenden Werte von $\mathbf{x}_{1\times 3}$ ja von dem betragsgrößeren $\mathbf{x}_j$ dominiert. Somit ist die Interaktion korreliert mit dem Prädiktor mit den größeren Werten. Diese Korrelation kann zu Problemen beim Schätzen der Parameter führen (siehe nächster Abschnitt). Viele praktische Statistikbücher empfehlen deshalb, die Prädiktoren *immer* zu standardisieren, sobald eine Interaktion oder ein polynomialer Effekt im Modell ist (Harrell 2001; Faraway 2005; Crawley 2007; Zuur et al. 2007).

Ist ein (oder mehrere) standardisierter Regressionskoeffizient im Betrag größer als 1, so deutet dies darauf hin, dass die $\mathbf{x}_j$ korreliert sind, was uns direkt zum nächsten Thema führt.

## 16.3 Modellselektion

[Kommentar zum *dredging* hinzugefügt:] Bei so vielen Modellen besteht natürlich die Gefahr, dass wir mit dem „besten“ Modell auch das Rauschen beschreiben, nicht nur das Signal. So ein *overfitting* zu verhindern ist die Aufgabe der Kreuzvalidierung (siehe **??**).

## Literatur


Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer, 2nd printing 2011 edition. 7

Crawley, M. J. (2007). *The R Book*. Chichester, UK: John Wiley & Sons. 8

Faraway, J. J. (2005). *Linear Models with R*. Boca Raton: Chapman & Hall/CRC. 8

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15), 2865–2873. 7

Good, P. I. & Hardin, J. W. (2012). *Common Errors in Statistics (And How to Avoid Them)*. Wiley, 4th edition edition. 4

Harrell, F. E. (2001). *Regression Modeling Strategies - with Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer. 5, 8

Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modeling*. Berlin: Springer. 7

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., & Stainforth, D. A. (2011). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, 430, 768–772. 7

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C. F., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., & Pélissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11, 4540. 7

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40, 913–929. 7

Smith, S. A. & Donoghue, M. J. (2010). Combining historical biogeography with niche modeling in the caprifolium clade of lonicera (caprifoliaceae, dipsacales). *Systematic Biology*, 59(3), 322–341. 6

Zuur, A. F., Ieno, E. N., & Smith, G. M. (2007). *Analysing Ecological Data*. Berlin: Springer. 8