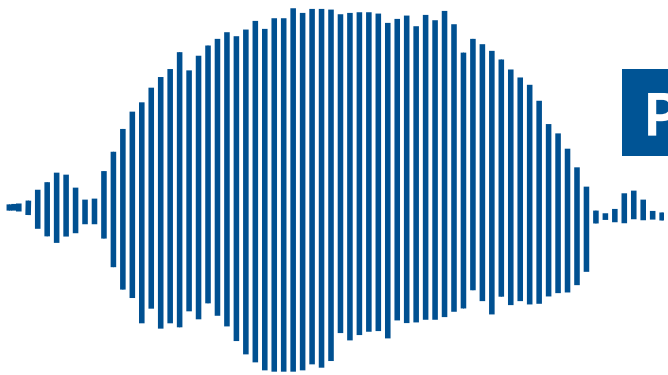
Statistische Analyse

Korrelation, Regression & Varianzanalyse

Felix Golcher

Humboldt-Universität zu Berlin

10. April 2010



Psycholinguistischer Methodenworkshop

9. und 10. April 2010

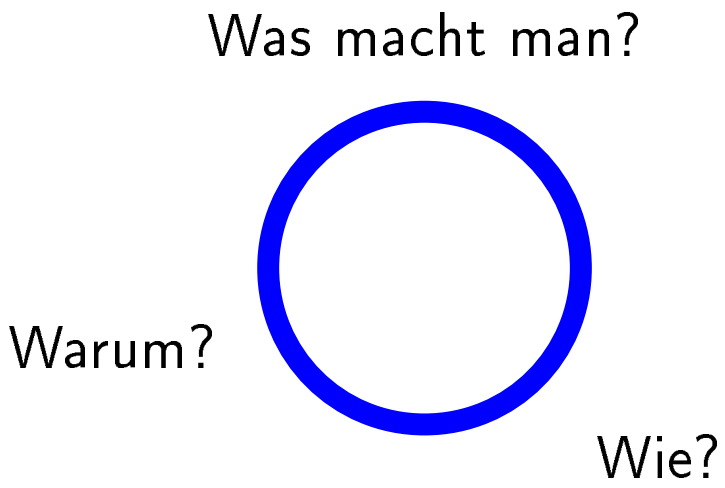
Felix Golcher[1]

- Doktorand (Korpuslinguistik)
- Ex-Physiker
- Wissenschaftlicher Mitarbeiter
  - *Beratung bei der Erhebung und Auswertung statistischer Daten*
  - Sprechstunde n.V.
  - Raum 3.333
  - 030 / 2093 9774
  - Felix.Golcher@hu-berlin.de

[1] http://www.linguistik.hu-berlin.de/staff/golcherf

Was machen wir heute?

Grundlagen, Verbindungen, Zusammenhänge & Grenzen.



Gliederung

1. Einführung
2. Grundbegriffe
3. Erinnerung: Grundlegende statistische Tests
4. Lineare Regression
5. Multiple lineare Regression
6. Das Allgemeine Lineare Modell (ALM)
7. Grenzen des ALM

Literaturempfehlungen

Bortz 2005 Das Handbuch. Der Klassiker. Ein bisschen veraltet, aber in den Grundlagen vollständig. verständlich. Viel zu dick. Zum Nachlesen, nicht zum Durchlesen. Leider absolut kein Bezug zur Linguistik.

Dalgaard 2008 R und Statistik. Inspiriert. Relativ knapp. Wenig Grundlagen. Autor ist einer der Miterfinder von R. Medizinisch orientiert. Schön zu lesen. Mit Einsichten.

Baayen 2008 R und Linguistik. Viele Beispiele. Keine Theorie. Teilweise etwas wirr. Beispiele manchmal mit Vorsicht zu genießen. Vieles nur angerissen. Eine Fundgrube, aber man muss graben.

das perfekte Buch sollte statistische Theorie (nicht zu tief, nicht zu oberflächlich) mit linguistischen Anwendungen (mehr Korpus, denn hier sind die wirklichen Fallen) mit vielen schönen Beispielen in R verbinden. Es muss noch geschrieben werden.

Vorbereitung

- Starten Sie bitte R.
  - Sie können an der Kommandozeile Kommandos eingeben und bekommen Antworten.
  - Oder sie erstellen ein Skript (eine Reihe Kommandos in einem Textfile), das Sie dann immer wieder ausführen können.
  - So ein Skript wurde für heute erstellt.
- Laden Sie das Skript `http://korpling.german.hu-berlin.de/~felix/mws/script.R` herunter.
- Gehen Sie in R auf *Skript öffnen* und öffnen Sie die gespeicherte Datei.
- mit Strg + r können Sie entweder die Zeile in der der Cursor ist oder ein selektiertes Gebiet ausführen lassen.
- Wir orientieren uns an den Tags der Form ##00X##.

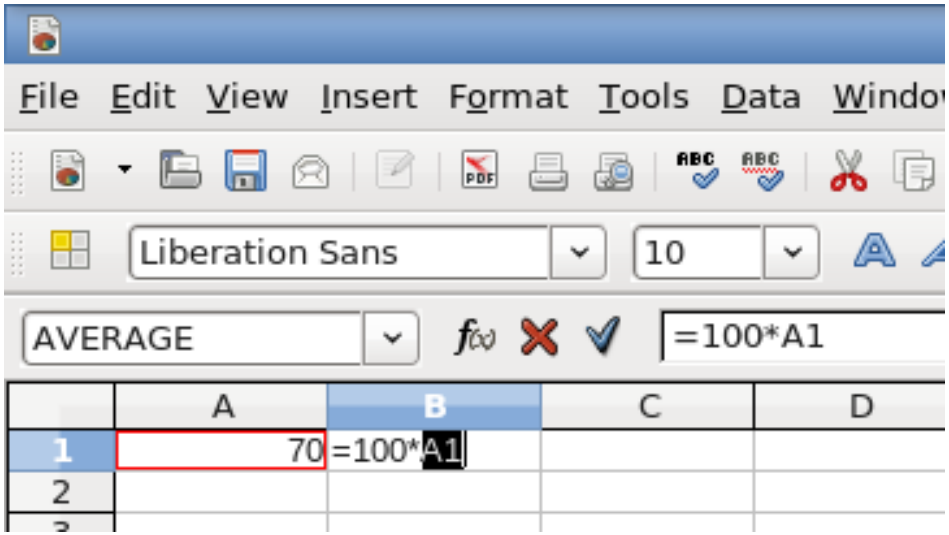
Was ist R

- eine Programmiersprache
- ein statistisches Analyseprogramm
- ein mächtiges Werkzeug zur graphischen Darstellung von Daten
- es ist frei[2].
- funktioniert unter „allen“ Betriebssystemen
- es gibt Hunderte/Tausende von Erweiterungsmodulen (packages)

[2] R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org`.

Was ist eine Variable?

Parallelbeispiel zu einem Tabellenkalkulationsprogramm:



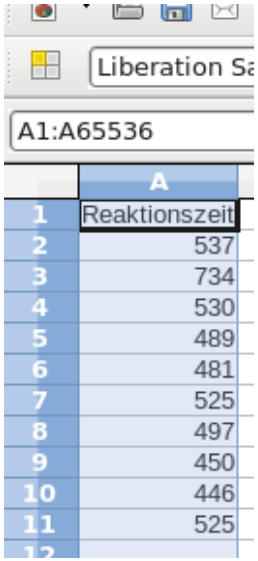
- Eine Variable (hier **A1**) ist ein Behälter für einen Wert.
- Diesen kann man ändern, dann wird überall der neue Wert verwendet.

Entsprechender R-Code wäre (Im Skript ##001##):

```
> A1 <- 70
> A1
[1] 70
> 100*A1
[1] 7000
```

Ignorieren Sie die Klammer ([1]) fürs erste.

Was ist ein Vektor? (##002##)

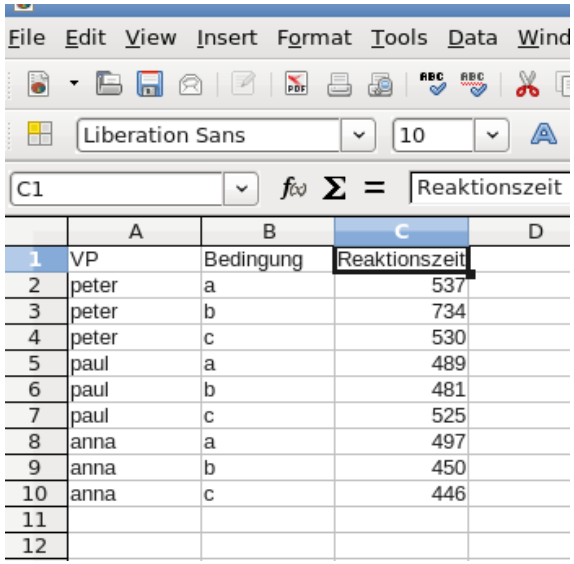


- Eine Aneinanderreihung von Werten.
- Einen Vektor (uvm.) kann man auch einer Variablen zuordnen.
- Das meiste, was man mit einer Zahl machen kann, kann man auch mit einem Vektor machen:
  - $\Rightarrow$ Anwendung auf jedes Vektorelement.

```
> Reaktionszeit <- c(537, 734, 530, 489, 481, 525, 497,
+                    450, 446, 525)
> Reaktionszeit/2
[1] 268.5 367.0 265.0 244.5 240.5 262.5
[7] 248.5 225.0 223.0 262.5
```

Damit ist dann auch die Bedeutung der Klammer ([1]) klar.

Was ist ein *data frame*? (##003##)



	A	B	C	D
1	VP	Bedingung	Reaktionszeit	
2	peter	a	537	
3	peter	b	734	
4	peter	c	530	
5	paul	a	489	
6	paul	b	481	
7	paul	c	525	
8	anna	a	497	
9	anna	b	450	
10	anna	c	446	
11				
12				

- Ein *data frame* ist das Äquivalent einer Excel-Tabelle.
- In R entweder über `data.frame(...)` oder `read.table(...)`

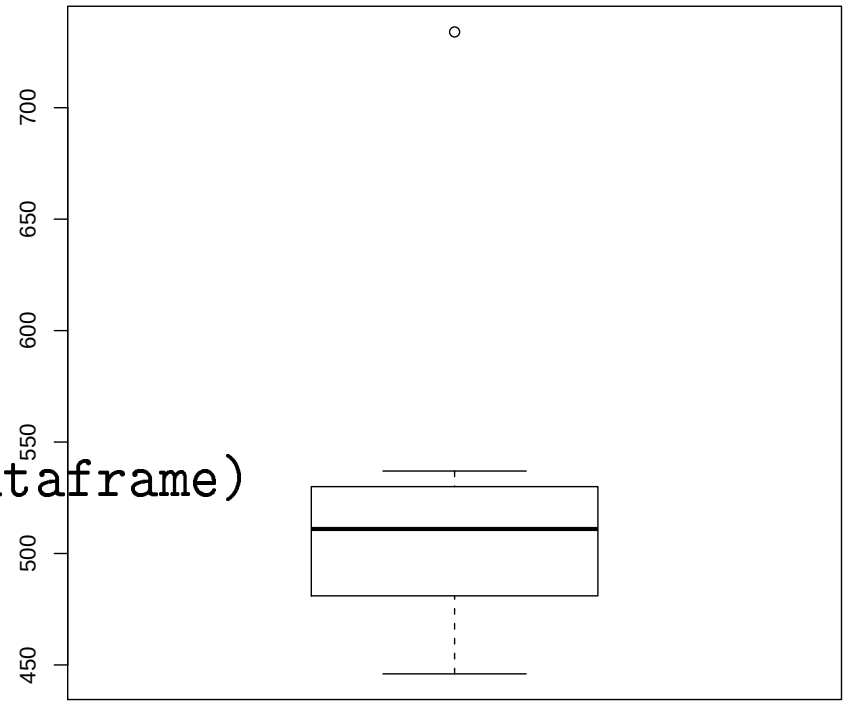
```
> mydataframe <- mydataframe <- read.table(
  "http://korpling.german.hu-berlin.de/~felix/mws/dataframe.dat")
> head(mydataframe,3)
     VP Bedingung Reaktionszeit
1 peter         a           537
2 peter         b           734
3 peter         c           530
```

Was ist eine Funktion? (##004##)

Ein Stück Programm, das Argumente bekommt und (nicht immer) etwas zurück gibt:

```
zurueck <- function(Arg1,Arg2=irgendwas)
```

```
> mean(mydataframe$Reaktionszeit)
[1] 521
> boxplot(Reaktionszeit,data=mydataframe)
```



Was ist eine Nullhypothese?

- Normalerweise möchte man zeigen, dass irgendetwas einen Unterschied macht.

Beispiel

Muttersprachler lesen schneller als Zweitsprachler.

- Die Nullhypothese H_0 ist nun die Annahme, dass es den gesuchten Unterschied nicht gibt.

Beispiel (Nullyhpothese H_0)

Muttersprachler lesen nicht schneller als Zweitsprachler.

- Die Alternativhypothese (H_1) ist im Standardfall die Annahme, dass es den erhofften Unterschied gibt.

Was ist der p-Wert?

- Irgendeinen Unterschied (z.B.) in der Lesegeschwindigkeit wird man immer messen.
- Der p-Wert ist nun

Definition (p-Wert)

die Wahrscheinlichkeit -bei Gültigkeit der Nullhypothese eine Abweichung zu finden, die mindestens so extrem wie die, die man tatsächlich misst.

$\Rightarrow$ Ein Maß dafür wie gut die Nullhypothese die Daten erklärt.

- großes $p \Rightarrow H_0$ erklärt die Daten gut $\Rightarrow$ Es spricht nichts gegen sie.
- kleines $p \Rightarrow H_0$ erklärt die Daten schlecht $\Rightarrow$ Wir zweifeln an H_0.
  $\Rightarrow$ Meist wollen wir zweifeln!

Was ist signifikant?

- Nullhypothese erklärt die Daten zu schlecht
  $\Rightarrow$ Wir erklären sie für falsch!
- Gleichbedeutend mit: p-Wert zu klein.
- „zu“ klein ist meist $< 5\%$
- Das nennt man auch das α- oder Signifikanzniveau.

α- und β-Fehler

	H_0 gilt	H_0 gilt nicht
Entscheidung für H_0	richtig	β-Fehler
Entscheidung gegen H_0	α-Fehler	richtig

- Oft betrachtet man nur den α-Fehler: Zu früh gefreut:
  $\Rightarrow$ Nichts da, obwohl es so aussieht.
- Auch interessant ist aber die Teststärke $1 - \beta$:
  $\Rightarrow$ Wie wahrscheinlich ist es, etwas zu finden, wenn etwas da ist?

Warum so kompliziert?

- Die ganze Sache mit der Nullhypothese klingt verdammt hintenrum:

  Angenommen, die Nullhypothese gilt, wie wahrscheinlich ist es dann, Daten zu bekommen, die mindestens so extrem von dieser abweichen wie...

- Wieso berechnet man nicht die Wahrscheinlichkeit dafür, dass der Effekt, der einen interessiert wirklich da ist?

  Wie wahrscheinlich ist es auf Grundlage meiner Daten, dass Muttersprachler schneller lesen?

- Die Antwort ist einfach:
  - Hat man ein Modell, kann man die Wahrscheinlichkeit für die Daten berechnen.
  - Von den Daten auf die Wahrscheinlichkeit eines Modells zu schließen, ist so gut wie unmöglich.

Eine kleine Kritik am Signifikanztest[3]

- Der Signifikanztest kommt aus einem industriellen Umfeld
- Dafür passt er. Passt er auch in die Wissenschaft?

	Autofabrik	Psycholinguistik
H_0	Autos sicher	Muttersprachler und Zweitsprachler lesen gleich schnell
H_0 ablehnen	Fabrik anhalten	Veröffentlichen
α-Fehler	Verzögerung (Geld)	spätere Falsifikation (oder ein Mythos)
β-Fehler	Rückrufaktion (mehr Geld)	Interessantes unveröffentlicht (für immer)

Vielleicht möchte man sich gar nicht entscheiden, sondern seine Ergebnisse präsentieren wie sie sind.

[3] Es gibt durchaus noch andere Kritikpunkte.

Häufige Irrtümer

1. Ein signifikantes Ergebnis beweist nicht, dass der Effekt existiert.
2. Fehlende Signifikanz beweist nicht, dass der Effekt nicht da ist.
3. Der p-Wert ist nicht die Wahrscheinlichkeit für die Nullhypothese.
4. Dass ein Artikel veröffentlicht wurde, heißt nicht, dass die Statistik stimmt.

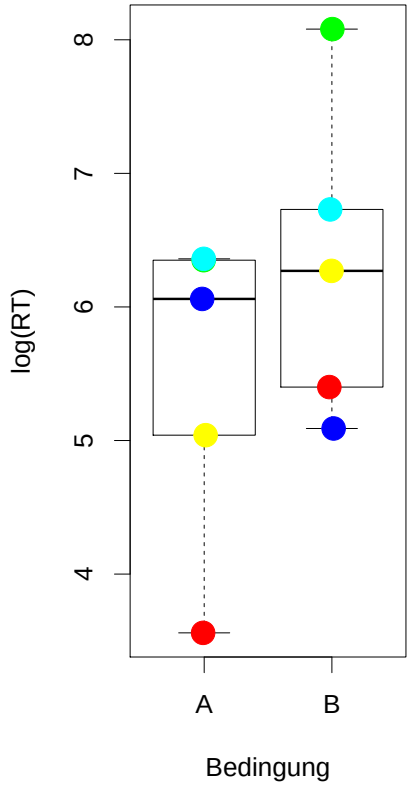
Der t-Test: Erläuterung an einem Beispiel

Den t-Test gibt es in einigen Varianten. Eine wichtige ist folgende Situation

- Sie haben zwei Stichproben, z.B. Reaktionszeiten[4]:

Versuchsperson	VP1	VP2	VP3	VP4	VP5
Bedingung A	3.56	6.06	5.04	6.35	6.36
Bedingung B	5.4	5.09	6.27	8.08	6.73

- Sie beobachten einen Unterschied. Ist dieser zufällig?



[4] Genauer: Logarithmen von (mittleren) Reaktionszeiten! Reaktionszeiten sind *nie* normalverteilt!

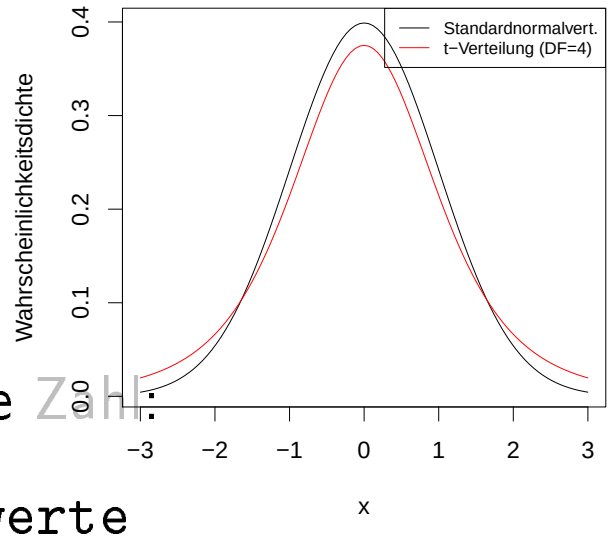
Das Verfahren

- Sie verwandeln beide Stichproben in eine einzige Zahl:

$$t = \frac{\texttt{Differenz der Mittelwerte}}{\frac{\texttt{Streuung der Stichproben}}{\sqrt{\texttt{Stichprobengröße}}}}$$

- Logik dahinter:
  - Mittelwerte sind immer normalverteilt. (Auch ihre Differenzen)
  - Wenn Sie durch die Streuung teilen, können Sie Stichproben mit großer und kleiner Streuung gleich behandeln.
  - Die Streuung eines Mittelwertes wird mit $\sqrt{\texttt{Stichprobengröße}}$ kleiner! (kommt später nochmal!)

Das Verfahren



- Sie verwandeln beide Stichproben in eine einzige Zahl.

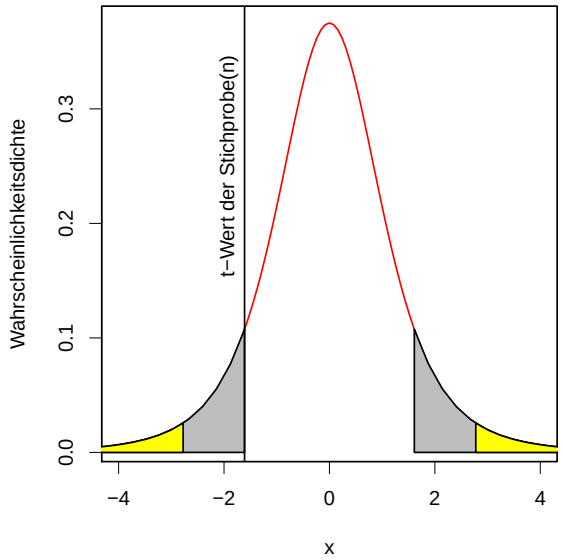
$$t = \frac{\texttt{Differenz der Mittelwerte}}{\frac{\texttt{Streuung der Stichproben}}{\sqrt{\texttt{Stichprobengröße}}}}$$

- Logik dahinter:
  - Mittelwerte sind immer normalverteilt. (Auch ihre Differenzen)
  - Wenn Sie durch die Streuung teilen, können Sie Stichproben mit großer und kleiner Streuung gleich behandeln.
  - Die Streuung eines Mittelwertes wird mit $\sqrt{\texttt{Stichprobengröße}}$ kleiner! (kommt später nochmal!)

  $\Rightarrow$ Wenn es **keinen** wirklichen Unterschied gibt (H_0),

  $<=$ Dann ist t *fast* normal verteilt, aber nicht ganz. (t-Verteilung)

Abschluss des Beispiels (##005##)



- Wenn man den t-Wert im Beispiel berechnet kommt man auf -1.61.
- Werte, die so mindestens so weit vom Mittelwert weg sind, haben zusammen eine Wahrscheinlichkeit von 0.18
  - Dies ist der p-Wert.
  - Im Bild der graue (und gelbe) Bereich.
- Erst ab wesentlich extremeren t-Werten (gelber Bereich) fällt diese Wahrscheinlichkeit unter 5%.

$\Rightarrow$ Unsere Stichprobe zeigt keinen signifikanten Unterschied.

Ein weiteres Beispiel: Simulierte Daten

- Wir gehen im Skript zu ##006##.
- Dort wird ein kleines Experiment simuliert.
- Die Nullhypothese H_0 gilt hier
  - $\Rightarrow$ Beide Gruppen haben denselben (Populations)mittelwert.
- Wer von Ihnen bekommt trotzdem ein signifikantes Ergebnis?
- Wiederholt man das Experiment oft genug stellt man fest:
  - 5% der Fälle ergeben falsch positive (α-Fehler)

Ein echtes Beispiel: Translationese[5]

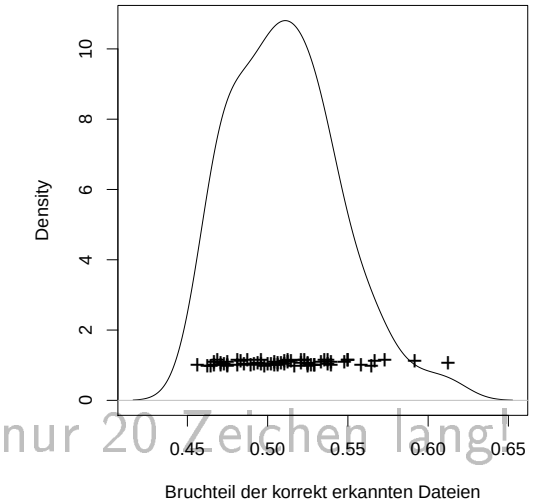
- Es wurde ein Maß für die Ähnlichkeit zweier Texte getestet.
- 813 Artikel, 569 Original italienisch, 244 Übersetzungen ins It.
- jeweils 15 Testfiles aus beiden Untermengen.
- jeweils 1 Trainingstext aus dem Rest.
- Kann man anhand der höheren Ähnlichkeit die Originale herausfinden?
- 60 mal wurde das Korpus durcheinandergewürfelt.
  - $\Rightarrow$ 60 Datenpunkte

[5] Korpus nach Baroni und Bernardini (2006)

Ergebnisse (##007##)

- Man beachte: Die Trainingstexte waren hier nur 20 Zeichen lang!

Ergebnisse (##007##)



- Man beachte: Die Trainingstexte waren hier nur 20 Zeichen lang!
- Die Daten sehen ausreichend normal aus.
  - Man kann das auch mit `shapiro.test()` testen.
- Der t-Test ergibt ein signifikantes Ergebnis ($p = 0.0117 < 0.05$)
- Wenn man Zweifel hat, ob die Daten normalverteilt sind:
  - Wilcoxon-Test (`wilcox.test()`)
  - Auch das gibt ein signifikantes Ergebnis ($p = 0.0317 < 0.05$)
- Fazit: 20 Zeichen Trainingstext reichen aus um die Chance für eine Erkennung als Original/Übersetzung über die Random-Baseline von 50% zu heben.
  - ... wenn auch nur auf etwa 51%

t-Test: eine wichtige Voraussetzung (##008##)

- Die wichtigste Voraussetzung für den t-Test:
  - Die Daten sollten normalverteilt sein.
- Was passiert eigentlich, wenn das nicht der Fall ist?
- Der Anteil der α signifikanten Stichproben verändert sich.
- Witzigerweise in eine unerwartete Richtung.
- Ein Ausweg ist der Wilcoxon-Test[6] `wilcox.test()`.
- (Es gibt noch weitere Annahmen für den t-Test, hier verschwiegen.)

[6] Er trägt verschiedene Namen.

Ein fundamentales Problem

Eins muss Ihnen klar sein:

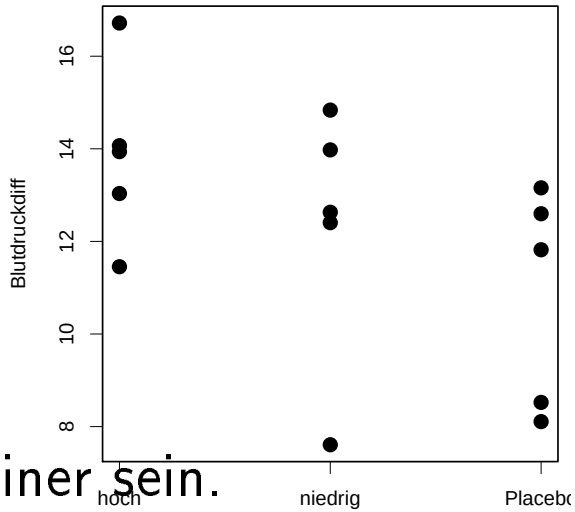
- Für fast alle statistischen Verfahren müssen die Daten normalverteilt sein.
- Ihre Daten werden nie normalverteilt sein.
- Andere Verfahren machen andere Annahmen.
- Auch diese werden Sie verletzen.
- Sie werden immer mit Näherungen arbeiten müssen.
- Wie weit darf man gehen??
  - Schauen Sie sich (als Erstes!) Ihre Daten an. Was sieht man denn?
  - Suchen Sie nach alternativen Auswertungsverfahren:
    - ★ Sie liegen nicht schlecht, wenn die Ergebnisse sich ähneln.
  - Trauen Sie keinem zu knappen Ergebnis.
  - Haben Sie Mut zur Vereinfachung, aber ...
  - ... versuchen Sie das Ausmaß der Vereinfachung zu überblicken.

Erweiterung des t-Tests auf mehr als 2 Gruppen

Wir beginnen mit einer typischen Fragestellung.

- Ein Medikament wird getestet.
- Drei Behandlungsstufen: Plazebo, geringe Dosis, hohe Dosis.
- In jeder Gruppe sind 20 Versuchspersonen.
- Wir messen jeweils die Blutdruckdifferenz vor und nach der Medikation:
  - Intervallskalierte abhängige Variable!
- Der mittlere Blutdruck pro Gruppe wird unterschiedlich sein.
- **Frage:** Kann der Zufall diese Unterschiede erklären?

ANOVA: Das Grundprinzip



- Annahme: Die drei Mittelwerte sind „eigentlich“ gleich.
- Jede der Gruppen erlaubt eine Schätzung der Varianz des Blutdrucks von Person zu Person.
- Die Varianz der Mittelwerte sollte um $1/n$ kleiner sein.
- Das heißt, es sollte ungefähr gelten:

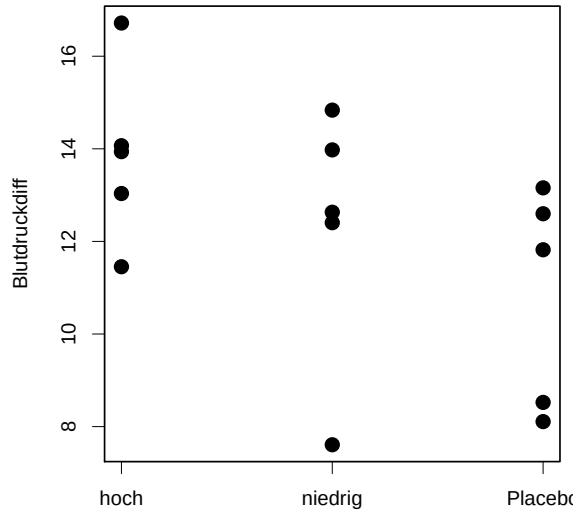
$$\texttt{Mittelwert der Varianzen} = n \cdot \texttt{Varianz der Mittelwerte}$$

oder auch

$$F = \frac{n \cdot \texttt{Varianz der Mittelwerte}}{\texttt{Mittelwert der Varianzen}} \approx 1$$

- Wenn F „zu weit“ von 1 weg ist, erklären wir den Effekt des Medikaments für signifikant.

Ein simuliertes Beispiel (##009##)



- Wir simulieren je 5 Werte wie im Eingangsbeispiel.
- Die Nullhypothese gilt nicht!

Gruppe	Blutdruckdifferenz (in *mmHg*)
Plazebo	10
niedrige Dosis	12
hohe Dosis	15

- Wir berechnen den F-Bruch manuell und mit `aov()`
- Längst nicht immer finden wir den vorhandenen Unterschied!

Ein lebensechtes Beispiel[7] (##010##)

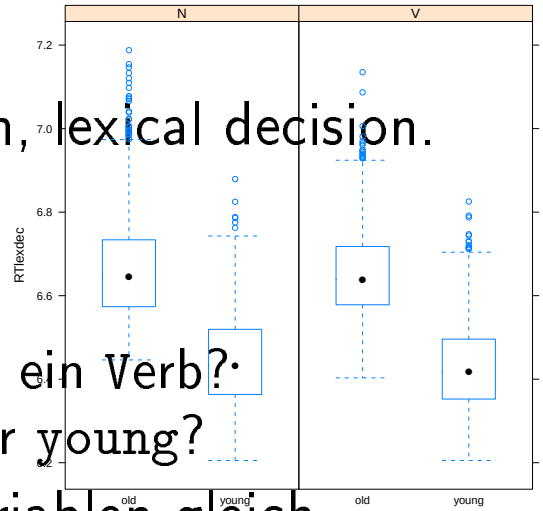
- 285 niederländische Verben.
- Drei Gruppen, je nachdem, mit welchem Hilfsverb sie vorkommen

Bezeichnung im Datensatz	Bedeutung
`hebben`	Zeigt nur das Hilfsverb *hebben*.
`zijn`	Zeigt nur das Hilfsverb *zijn*.
`zijnheb`	Kommt mit beiden Hilfsverben vor.

- `VerbalSynsets` Gibt die Zahl der Synsets an.
- Hängt die Zahl der Synsets davon ab, mit welchem Hilfsverb ein Verb vorkommt?
- Die Daten sind schwierig: Nicht normalverteilt / nicht „balanciert“.
- Wir folgen drei verschiedenen Gedanken, jeweils weisen die Ergebnisse in die selbe Richtung.
- Wir schlussfolgern Signifikanz.

[7] nach Baayen (2008)

Ein Beispiel zum Ausblick[8] (##011##)



- Datensatz mit (logarithmischen) Reaktionszeiten, lexical decision.
- 2197 Englische Worte
- 2 kategoriale Variablen:
  - `WordCategory`: Was das Wort ein `Nomen` oder ein `Verb`?
  - `AgeSubject`: War die Versuchsperson `old` oder `young`?
- Baayen führt ANOVAS u.ä. durch, die beide Variablen gleich behandeln.
  - Ist das gerechtfertigt?
- Das Alter unterscheidet sich entscheidend von der Wortart:
  - Die Wortart kann man verfeinern, man behält aber immer Kategorien.
  - Das Alter lässt sich aber als kontinuierliche Variable auffassen.
- Die Reduzierung auf eine Zweiteilung verschenkt (kurz gesagt) Information.
- Baayen selbst schreibt dazu im selben Buch ▸ folgendes.

[8] Baayen 2008, 170ff.

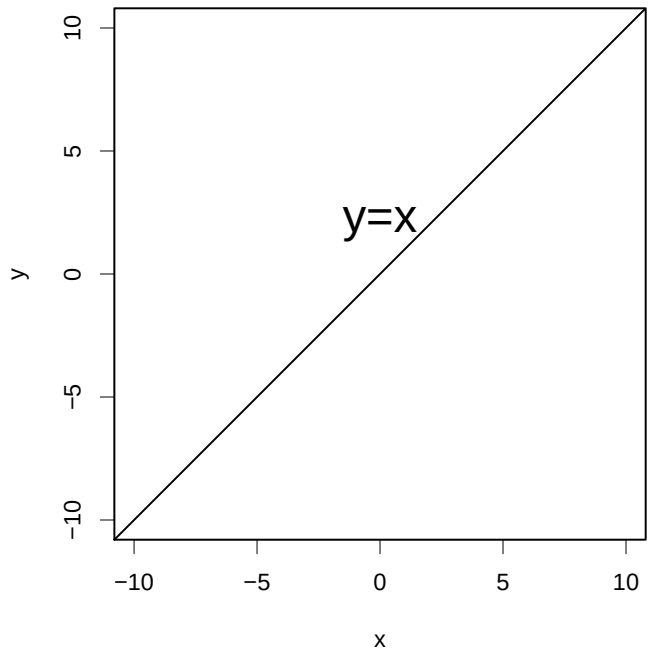
Was sind Geraden mathematisch?

Um lineare Regression verstehen zu können, müssen wir uns kurz erinnern.

- Die einfachste Gerade ist

$$y = x$$

- Die Punkte, für die diese Gleichung gilt, liegen auf der Diagonalen (##012##).



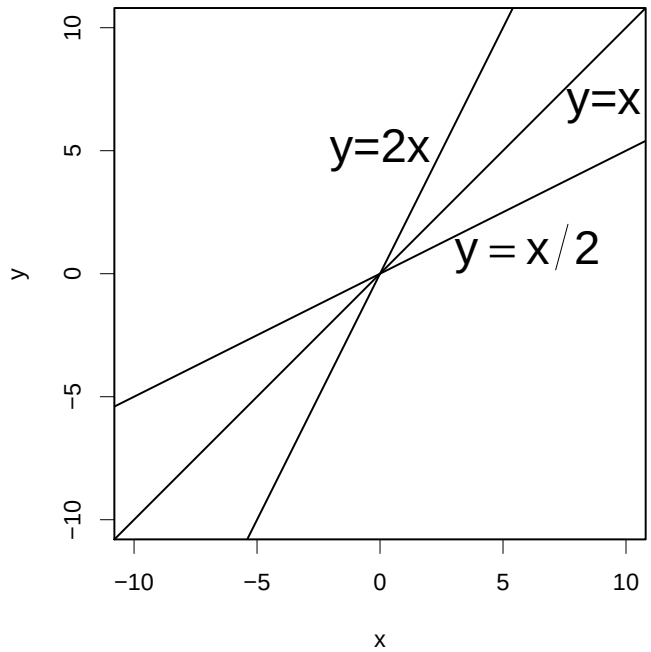
- Das ist noch nicht besonders interessant.

Wir ändern die Steigung (##013##)

- Wenn wir vor x einen Faktor schreiben (b)

$$y = b \cdot x \quad (= bx)$$

verändert sich die Steigung der Geraden



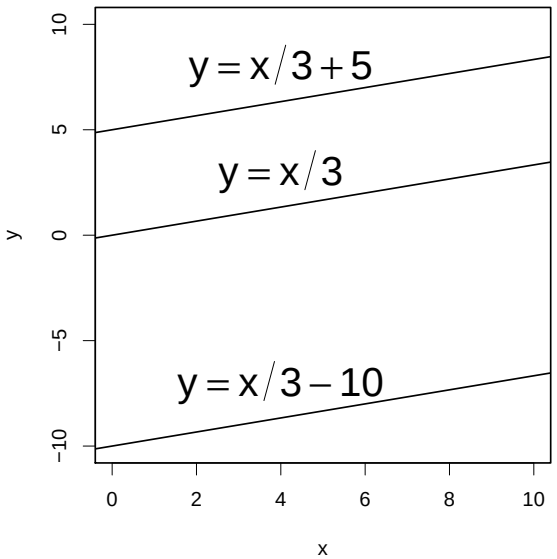
- Aber auch das ist noch nicht besonders flexibel.

Die allgemeine Geradengleichung (##014##)

- Wir fügen noch eine Konstante hinzu:

$$y = b \cdot x + c \quad (= bx)$$

- Dies verschiebt die Gerade nach oben ($c > 0$) oder unten ($c < 0$).



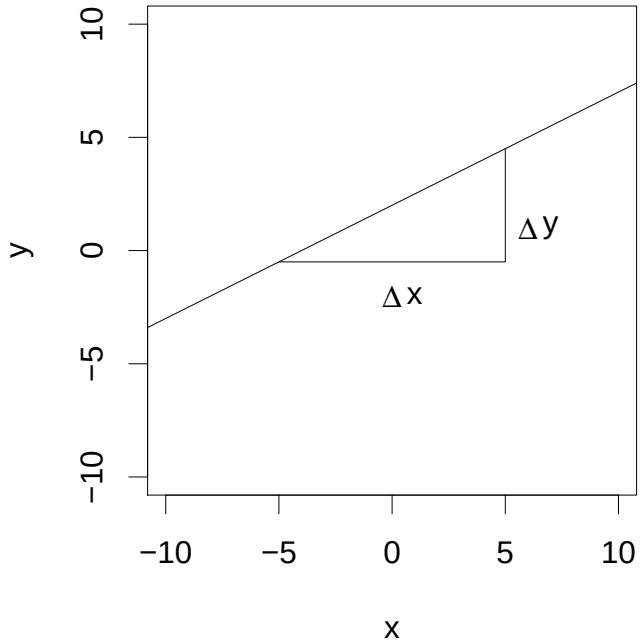
- Die Konstante entspricht dem y bei $x = 0$ (Achsenabschnitt).
- Nun können wir jede beliebige Gleichung darstellen.

Nochmal in Worten

- Beachte: Eine Gerade hat immer zwei Parameter:

  $$\texttt{abh. Variable} = \texttt{Steigung} \times \texttt{unabhängige Variable} + \texttt{Achsenabschn}$$

- Die Steigung ist das Ausmaß in dem sich y ändert, wenn man x ändert.

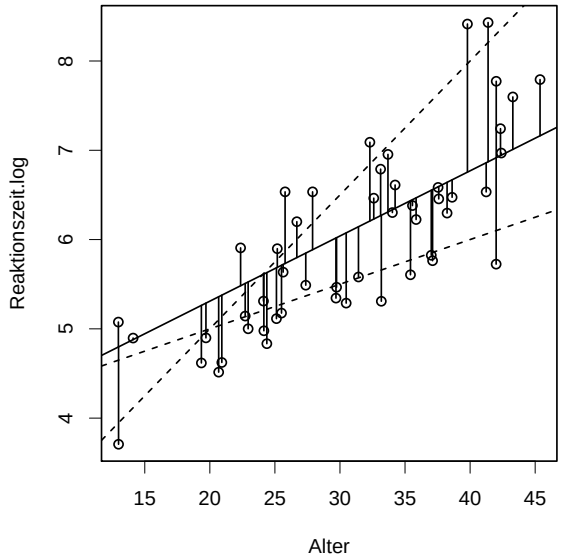


$$\texttt{Steigung} = b = \frac{\Delta x}{\Delta y}$$

- Daran werden wir uns gleich nochmal erinnern.

◂ Berechnung der Steigung

Geraden durch Punktewolken (##015##)



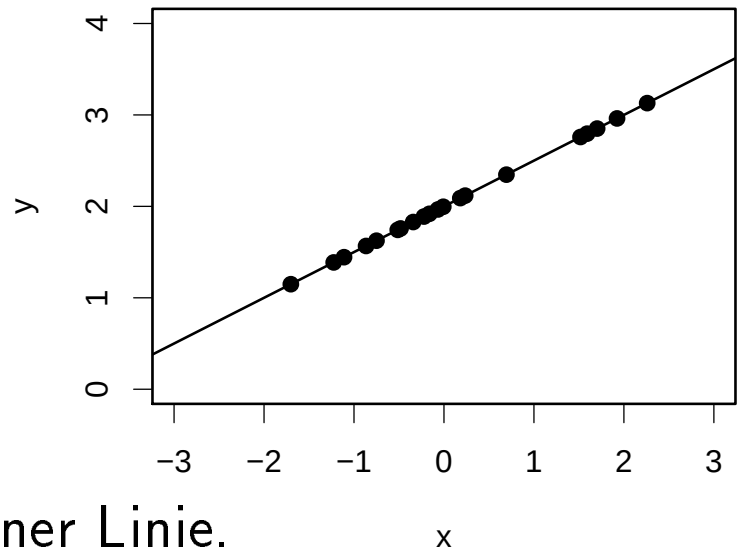
- Wir haben nun den Verdacht: Reaktionszeit nimmt mit dem Alter zu.
- Wir versuchen folgenden Ansatz:

$$\texttt{Reaktionszeit} = \texttt{Alter} \times \texttt{Steigung} + \texttt{Konstante} + \textit{Normalverteiltes}$$

- Man kann nun jede Versuchsperson als einen Datenpunkt darstellen.
- Man erhält eine Punktewolke.
- Durch diese kann man eine Menge verschiedene Geraden legen.
- Man wählt die Gerade,
  - ...für die die Summe der quadrierten Abweichungen minimal ist.
  - wie wenn ein Gummiband die Punkte mit der Geraden verbinden würde.
  - Diese Gerade macht die beobachteten Daten am wahrscheinlichsten.[9]

[9] *Maximum Likelihood*

Ein idealer Fall



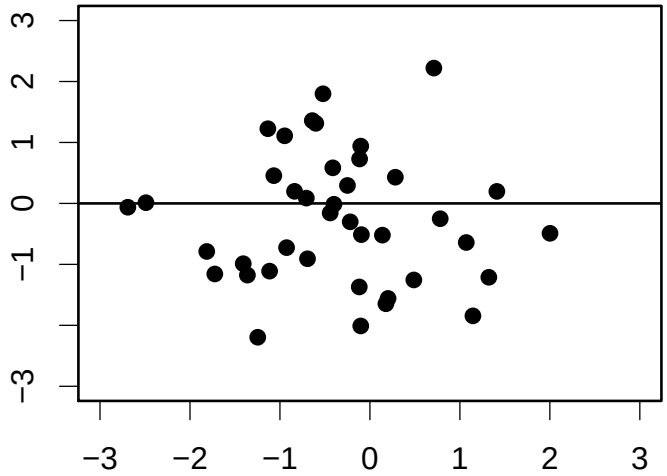
- Angenommen, wir haben Punkte auf einer Linie.
- Dann ist die Gerade klar.
- Die Steigung ist dann (Das hatten wir ja schon mal.):

$$\texttt{Steigung} = \frac{\texttt{Ausdehnung in } y}{\texttt{Ausdehnung in } x} = \frac{\texttt{Streuung in } y}{\texttt{Streuung in } x} = \frac{\texttt{Standardabw. in } y}{\texttt{Standardabw. in } x}$$

- Einfacher kann man auch schreiben

$$b = \frac{s(y)}{s(x)}$$

Auf der anderen Seite (##007##)17



- Wenn wir dagegen keine Beziehung haben zwischen x und y...
  - $\Rightarrow$ ... so wird die Steigung immer Null sein!

$$b = 0$$

- Für Fälle zwischen diesen beiden Extremfällen wird gelten

$$0 < b < \frac{s_y}{s_x}$$

- Das kann man schreiben als

$$b = r\frac{s_y}{s_x} \quad \texttt{mit} \quad 0 < r < 1$$

- Dieses r nennt man die Korrelation zwischen x und y.
  - $r = 1$ heißt: perfekter Zusammenhang zwischen x und y
  - $r = 0$ heißt: gar kein Zusammenhang zwischen x und y
  - $r = -1$ heißt: perfekter Zusammenhang, aber Steigung negativ.

Als Formel

Das nur, um zu zeigen, dass man das einfach ausrechnen kann...

$$r = \frac{cov(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\frac{\sum_{i=1}^{n}(x_i - \overline{x}^2)}{n-1} \cdot \frac{\sum_{i=1}^{n}(y_i - \overline{y}^2)}{n-1}}$$

... und um Sie zu erschrecken.

Behandlung mit R (##018##)

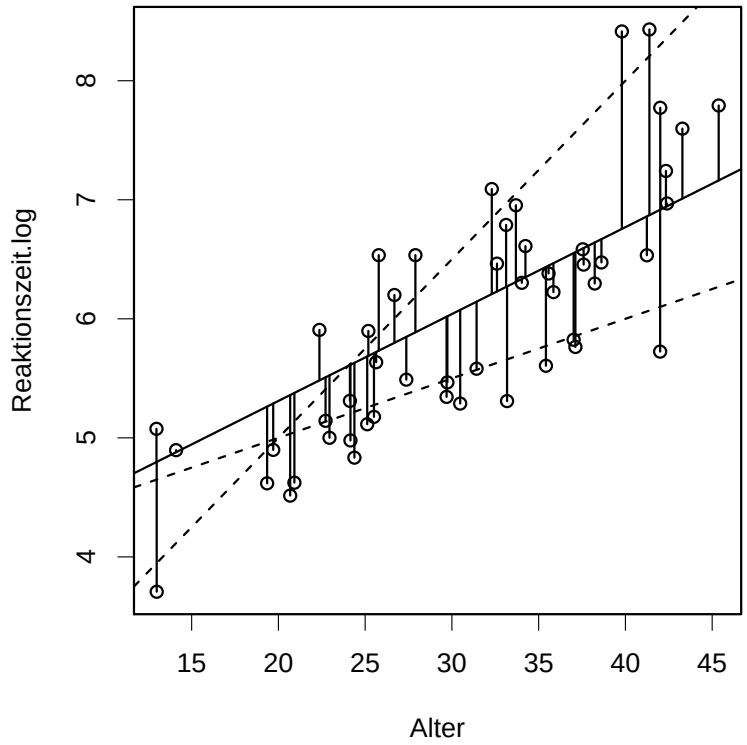
- Die Berechnung einer Regressionsgeraden in R ist denkbar einfach[10]:

  ```
  > lm(y ~ x, data=mydataframe)
  ```

- Hierbei müssen Sie für `x`, `y` und `mydataframe` jeweils Ihre Bezeichnungen eingeben.
- Sie bekommen die Geradengleichung zurück (etwas kryptisch).
- Wenn Sie das Ergebnis zwischen speichern und `summary()` darauf anwenden, bekommen Sie mehr Informationen.

[10] Es gibt noch eine wichtige Alternative: `ols()` aus dem Paket `MASS`.

Deutung der Korrelation



Zwei Deutungen sind möglich:

1. Wie genau kann ich y schätzen, wenn ich x kenne?
   - $r = 0$: Kenntnis von x ist nutzlos.
   - $r = 1$: Kenne ich x, kenne ich y.
2. Die erklärte Varianz:
   - Sei $Var(y)$ die Varianz der y-Stichprobe.
   - Sei $Var(y_v)$ die Varianz der vorhergesagten Werte.
   - Sei $Var(y_r)$ die Varianz der Differenzen, der Residuen.
   - Dann gilt

$$Var(y) = Var(y_v) + Var(y_r)$$

   - Nun ist aber $Var(y_v) = r^2$, die Varianz, die man erklären kann!
   - Das kommt uns irgendwie bekannt vor (ANOVA)...

Korrelation und Kausalität

Nur eine kleine Warnung:

- Eine hohe Korrelation heißt nicht, dass auch ein kausaler Zusammenhang besteht.
- Sie können nur die eine Variable (besser) vorhersagen, wenn Sie die anderen kennen.
- Korrelation ist symmetrisch.

Korrelation und Regression

- ... sind zwei Seiten der selben Münze.
- Ob Sie
  - eine Regressionsgerade berechnen, um eine Variable aus der anderen vorherzusagen, oder ob Sie,
  - die Korrelation zwischen zwei Variablen berechnen, um zu sehen, was die eine über die andere aussagt,
- ist egal.

Voraussetzungen für lineare Regression

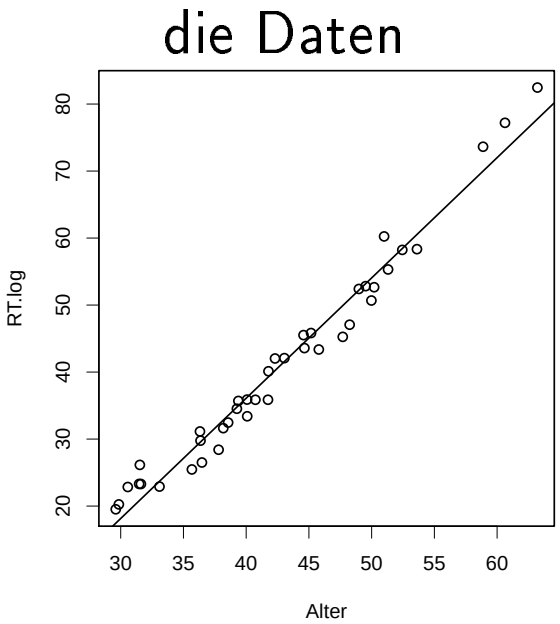
- Die x-Werte und die y-Werte sollten normalverteilt sein.
- Ein x-Wert sollte den nächsten nicht beeinflussen (keine Zeitreihe).
- Die Varianz der y-Werte sollte sich nicht mit x ändern.

Diese Annahmen sollte man überprüfen. Zum Beispiel mit einer Resiudalanalyse.

Modellkritik in Ansätzen - Ausreißer (##019##)

- Bei jedem Experiment gibt es Ausreißer:
  - Versagende Technik, verschlafene Versuchspersonen, Schreibfehler.
- Nun gibt es die naive Regel: wir schmeißen alles raus, was mehr als 2 Standardabweichungen vom Mittelwert entfernt ist.
- Damit entfernt man selbst bei perfekt normalverteilten Daten 4.5% der Daten (in R: `2*(1-pnorm(2))`)
- Meist ist das zu viel und die richtigen Punkte findet man oft trotzdem nicht.
- Das schauen wir im Skript an einem praktischen Beispiel an.

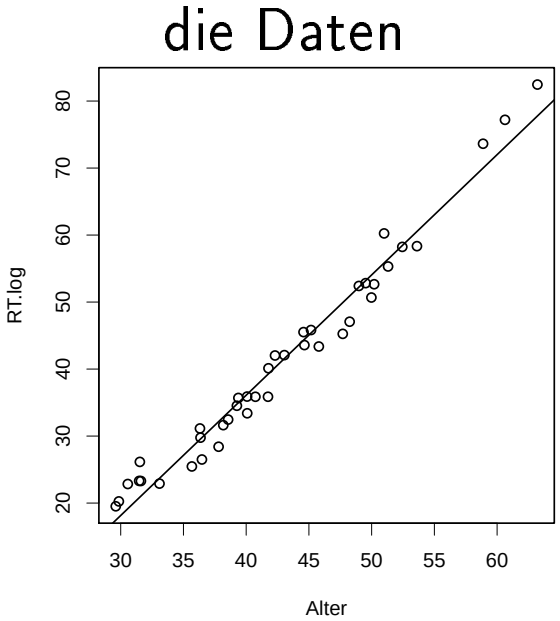
Modellkritik in Ansätzen - nichtlineares (##020##)

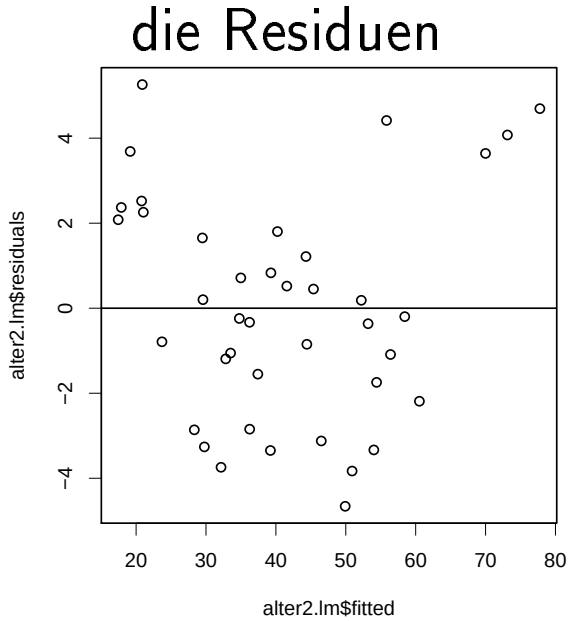


- Dieser Datensatz sieht erstmal ganz gut aus.

[11] aufgetragen über den vorhergesagten Werten

Modellkritik in Ansätzen - nichtlineares (##020##)





- Dieser Datensatz sieht erstmal ganz gut aus.
- Erst die Analyse der Residuen[11] macht misstrauisch.
- Hier könnte man einen quadratischen Term einführen, um die Nichtlinearität zu modellieren. Darauf gehen wir hier nicht ein.

[11] aufgetragen über den vorhergesagten Werten

Modellkritik in Ansätzen- vergessene Variablen (##021##)

- Wieder ein Datensatz aus Baayen (2008).
- Lexical decision Reaktionszeiten (logarithmisch).
- ca 2000 Verben und Nomen.
- Viele verschiedene Variablen, die man ausprobieren kann.
- Wir suchen erst einmal nach einem Frequenzeffekt.

Modellkritik in Ansätzen- vergessene Variablen (##021##)

- Wieder ein Datensatz aus Baayen (2008).
- Lexical decision Reaktionszeiten (logarithmisch).
- ca 2000 Verben und Nomen.
- Viele verschiedene Variablen, die man ausprobieren kann.
- Wir suchen erst einmal nach einem Frequenzeffekt.
- An den Residuen erkennt man, dass etwas nicht stimmt.
- genaueres Hinsehen: Das Alter muss in Betracht gezogen werden.

zurück

Ist die Natur linear?

- Lineare Regression bedeutet zwar nicht, durch alles eine Gerade legen zu wollen.
- Aber im Grunde eben doch. Der Vorrat an Möglichkeiten ist begrenzt.
- Die Natur ist aber nie eine Gerade. (Auch keine Parabel.)
- Die Modelle, die man gemeinhin verwendet, sind also immer falsch.
- Damit wird auch der Begriff der „Schätzung“ witzig
  - *Wir schätzen die Steigung der Geraden durch die Regressionsgleichung.*
- Man schätzt etwas, was es nicht gibt.
- Wenn eine Steigung, *signifikant* ist, heißt das nur:
  - Das Modell ohne Steigung ($b = 0$) erklärt die Daten schlecht.
- Warum macht man das dann?(!)
  - Man kann es rechnen.
  - Es ist immer eine erste Näherung (zumindest lokal). Stichwort *Taylor-Entwicklung*
  - Es gibt modernere Ansätze. Wir haben heute die Rechenkraft.

Wann ist eine Regressionsgeradensteigung *signifikant*?

- Eben haben wir von einer signifikanten Geradensteigung gesprochen.
- Was bedeutet das?
  - Wenn man viele Stichproben ziehen würde wie die, die man hat,
  - Dann berechnete sich jeweils eine leicht unterschiedliche Steigung.
  - Die Nullhypothese lautet nun: Die Steigung ist eigentlich Null.

Wann ist eine Regressionsgeradensteigung *signifikant*?

- Eben haben wir von einer signifikanten Geradensteigung gesprochen.
- Was bedeutet das?
  - Wenn man viele Stichproben ziehen würde wie die, die man hat,
  - Dann berechnete sich jeweils eine leicht unterschiedliche Steigung.
  - Die Nullhypothese lautet nun: Die Steigung ist eigentlich Null.
- Wir nehmen nun wieder an, die Nullhypothese gilt.

Wann ist eine Regressionsgeradensteigung *signifikant*?

- Eben haben wir von einer signifikanten Geradensteigung gesprochen.
- Was bedeutet das?
  - Wenn man viele Stichproben ziehen würde wie die, die man hat,
  - Dann berechnete sich jeweils eine leicht unterschiedliche Steigung.
  - Die Nullhypothese lautet nun: Die Steigung ist eigentlich Null.
- Wir nehmen nun wieder an, die Nullhypothese gilt.
- Wie wahrscheinlich es ist dann,...

Wann ist eine Regressionsgeradensteigung *signifikant*?

- Eben haben wir von einer signifikanten Geradensteigung gesprochen.
- Was bedeutet das?
  - Wenn man viele Stichproben ziehen würde wie die, die man hat,
  - Dann berechnete sich jeweils eine leicht unterschiedliche Steigung.
  - Die Nullhypothese lautet nun: Die Steigung ist eigentlich Null.
- Wir nehmen nun wieder an, die Nullhypothese gilt.
- Wie wahrscheinlich es ist dann,...
- ... eine mindestens so große Steigung zu bekommen

Wann ist eine Regressionsgeradensteigung *signifikant*?

- Eben haben wir von einer signifikanten Geradensteigung gesprochen.
- Was bedeutet das?
  - Wenn man viele Stichproben ziehen würde wie die, die man hat,
  - Dann berechnete sich jeweils eine leicht unterschiedliche Steigung.
  - Die Nullhypothese lautet nun: Die Steigung ist eigentlich Null.
- Wir nehmen nun wieder an, die Nullhypothese gilt.
- Wie wahrscheinlich es ist dann,...
- ... eine mindestens so große Steigung zu bekommen
- wie die, die man tatsächlich berechnet hat.

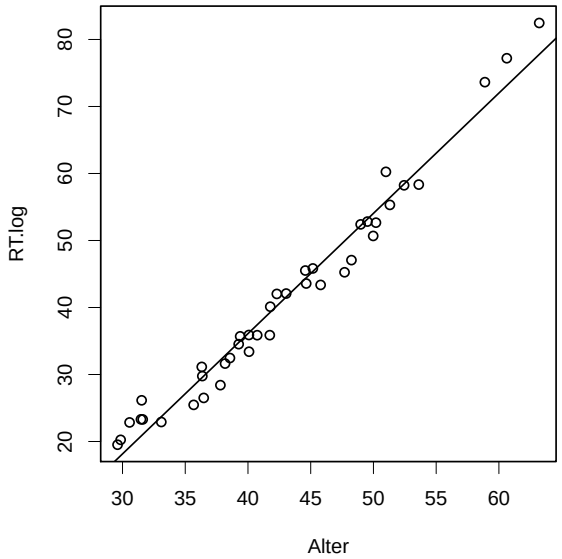
Wann ist eine Regressionsgeradensteigung *signifikant*?

- Eben haben wir von einer signifikanten Geradensteigung gesprochen.
- Was bedeutet das?
  - Wenn man viele Stichproben ziehen würde wie die, die man hat,
  - Dann berechnete sich jeweils eine leicht unterschiedliche Steigung.
  - Die Nullhypothese lautet nun: Die Steigung ist eigentlich Null.
- Wir nehmen nun wieder an, die Nullhypothese gilt.
- Wie wahrscheinlich es ist dann,...
- ... eine mindestens so große Steigung zu bekommen
- wie die, die man tatsächlich berechnet hat.
- Diese Wahrscheinlichkeit ist der p-Wert.
- Ist er kleiner als 5% erklären wir die Steigung für signifikant.

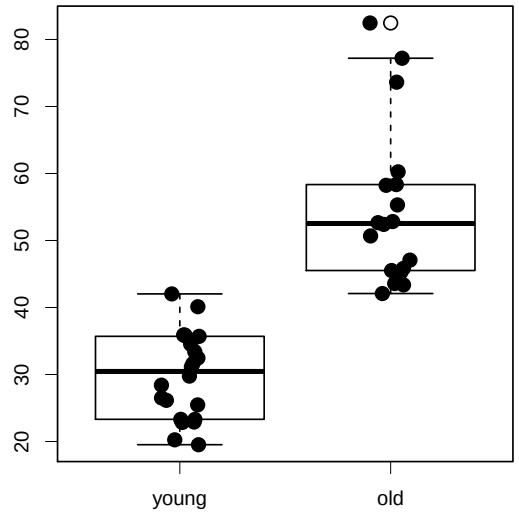
Wir gehen einen Schritt zurück (##022##)

- Aus pädagogischen Gründen tun wir etwas, was wir nie tun wollten:
  - Wir die kontinuierliche Variable `Alter` in zwei Kategorien zusammen

Alter kontinuierlich



$\Rightarrow$



Nur jung oder alt

- Man kann nun auch durch das rechte Bild eine Gerade ziehen.
- Diese ist genau dann signifikant von Null, wenn der Unterschied zwischen beiden Altersgruppen signifikant ist.
- Das ist genau dann der Fall, wenn ein t-Test signifikant ist.

Die ganze Prozedur kann man auch ◂ andersherum machen.

Wir sind wieder beim t-Test

- Natürlich geht bei der Zweiteilung des Alters Information verloren.
- Dennoch ist es nur eine gewisse Vereinfachung.
- Die Struktur des Problems bleibt die selbe.
- Deswegen sollte es nachzuvollziehen sein, dass auch im kontinuierlichen Fall die Signifikanz der Steigung mit einem t-Test berechnet werden kann.

Das heißt, die Signifikanz der Steigung...

- ist das selbe wie die Signifikanz der Korrelation.
- ist das selbe wie die Signifikanz eines Unterschiedes.
- kann mit einem t-Test überprüft werden.

Dies gilt im zweigeteilten wie im kontinuierlichen Fall.

Was macht man mit 2 unabhängigen Variablen?

- Sehr oft reicht natürlich eine einzige Variable nicht aus.
  - Wir haben das am Beispiel mit den englischen Worten gesehen.
- Die Erweiterung von einer auf zwei Variablen ist einfach.
- Statt

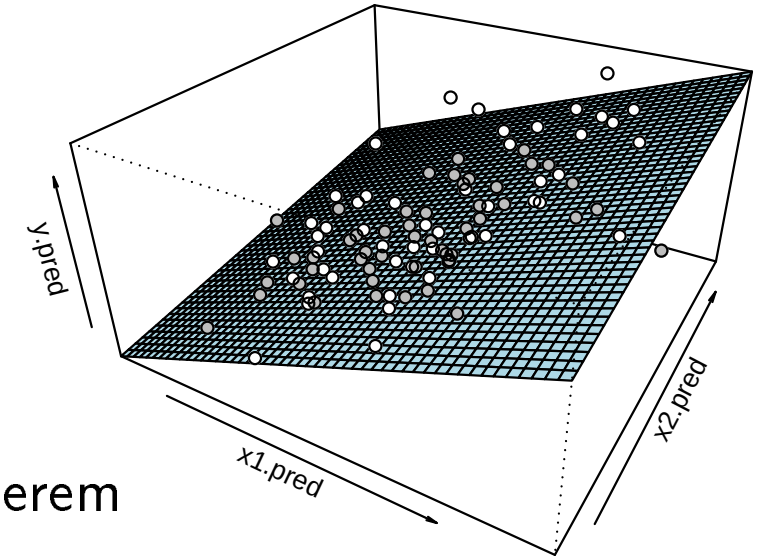
$$y = b \cdot x + c + \texttt{normalverteiltes}$$

- heißt es nun

$$y = a \cdot x_1 + b \cdot x_2 + c + \texttt{normalverteiltes}$$

- Wir gehen von einer Geraden zu einer Ebene über.

Ausprobieren in R (##023##)



- Wir simulieren Daten, die genau zu unserem nun 2-dimensionalen Modell passen.
- Die Fitfunktion (`lm`) ist die selbe.
- Unser Modell

$$y = a \cdot x_1 + b \cdot x_2 + c + \texttt{normalverteiltes}$$

wird in R als

```
y ~ x1 + x2
```

formuliert. Der Achsenabschnitt wird automatisch berechnet.

Was kann eine Ebene?

- Eine Ebene ist wiederum eine ziemlich spezielle Fläche.
- Längst nicht jede (3dimensionale) Punktewolke ist durch eine Ebene beschreibbar.
- Man macht wieder das machbare bzw berechenbare.
- Und fängt mit einer vernünftigen Näherung an. Das ist nun mal eine Ebene.
- Wir kommen nun aber zu einer wichtigen Erweiterung
  - Interaktionen.

Der Interaktionsterm

- Wir erweitern unser Modell

$$y = ax_1 + bx_2 + c + \texttt{normalverteiltes}$$

um einen weiteren Term.

- Dieser hängt vom Produkt der beiden unabhängigen Variablen ab.

$$y = ax_1 + bx_2 + dx_1x_2 + c + \texttt{normalverteiltes}$$

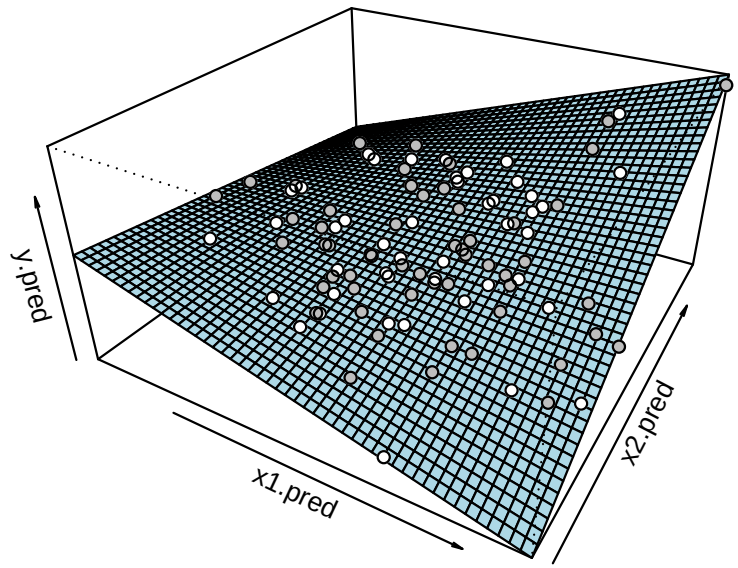
- Dies kann man auch anders formulieren:

$$y = ax_1 + (b + dx_1)x_2 + c + \texttt{normalverteiltes}$$

- Das heißt, der Effekt durch x_2 hängt jetzt von x_1 ab.
- oder umgekehrt:

$$y = (a + dx_2)x_1 + bx_2 + c + \texttt{normalverteiltes}$$

gebogene Ebenen (##024##)



- In R beschreibt man eine Interaktion zwischen x_1 und x_2 mit

```
x1:x2
```

- Dh., unser Modell wäre als

```
y ~ x1 + x2 + x1:x2
```

  zu formulieren.
- Dafür gibt es eine Abkürzung:

```
y ~ x1*x2
```

Was heißt hier eigentlich „linear“?

- Linear, da denkt man ja an Linien.
- Oder eben an Ebenen.
- Besonders eben sieht ein Modell mit Interaktion aber nicht aus.
- Das ist ein häufiges Missverständnis:
  - Linear heißt nicht linear in den Variablen x_1, x_2, ..., x_i
  - Linear heißt linear in den Parametern a, b, c, ...

Was ist ein Modell?

- Das ist eine sehr gute Frage!
- Hier vielleicht so etwas wie eine gedachte „Messwertproduktionsmaschine“.
- Man steckt Variablenwerte hinein (Alter, Frequenz, Wortart...)
- und bekommt Vorhersagen über Messwerte (Reaktionszeiten?) heraus.
- Und man kann berechnen, wie wahrscheinlich mit diesem Modell die eigenen Daten sind.
- Die Maschine hat Schrauben: Die Parameter.
- Man schraubt so lange an diesen herum, bis die Wahrscheinlichkeit für die Daten maximal ist.
- Dann erklärt/reproduziert das Modell die Daten gut.

Was ist das Allgemeine Lineare Modell (ALM/GLM)?

- Unsere Gleichung, also unser Modell

$$y = ax_1 + bx_2 + dx_1x_2 + c + \texttt{normalverteiltes}$$

- kann man beliebig erweitern
  - auf beliebige Anzahlen von x_i
  - und auch auf mehrere y_j gleichzeitig (multivariates, hier nur erwähnt)
- Diese Verallgemeinerung ist das allgemeine lineare Modell.

Zurück zur ANOVA

- Wir haben ▸ oben aus einer kontinuierlichen Variablen eine diskrete gemacht.
- Man kann natürlich auch den umgekehrten Weg gehen.
- Angenommen, wir haben eine kategoriale Variable Numerus.
  - `Singular` kodieren wir als 0.
  - `Plural` kodieren wir als 1.
- Nun rechnet man im ALM weiter, als hätte man eine kontinuierliche Variable.
- Schon hat man den gewöhnlichen t-Test ins System integriert.
- Was ist aber mit kategorialen Variablen mit mehr als zwei Ausprägungen?

Mehr als eine Variable: Integration der ANOVA

- Angenommen, wir betrachten Wortarten: `N`, `V` und `Adj`.
- Dann brauchen wir zur Kodierung zwei Variablen x_1 und x_2:

	x_1	x_2
`N`	1	0
`V`	0	1
`Adj`	0	0

- Mit diesen beiden Variablen kann man dann ein lineares Modell bestücken:

$$y = ax_1 + bx_2 + c + \texttt{normalverteiltes}$$

- Nun kann man wieder fröhlich im linearen Modell rechnen.
- Ein Interaktionsterm würde wegfallen, da $x_1 x_2$ immer Null ist.

Was gehört alles zum ALM?

- t-Test
- ANOVA
- MANOVA (multivariate ANOVA)
- χ^2-Tests
- Kovarianzanalyse
- und natürlich die lineare Regression...

Random Effects

- Die Variablen, die wir bisher behandelt haben, waren reproduzierbar:
  - Man wird andere Worte finden, die so und so häufig sind.
  - Man wird andere Worte finden, die Nomen sind.
  - Man wird andere Versuchspersonen finden, die 30 Jahre alt sind.
- Es gibt aber auch Variablen, die nicht reproduzierbar sind:
  - Man kann keine Versuchsperson finden, die sich wie Karl verhält.
    - ★ Karlheit ist keine reproduzierbare Eigenschaft.
  - Man wird kein Wort finden, dass sich genau wie *Zylinderlinse* verhält (Häufigkeit, Schriftbild, Assoziationen, etc.)
- Solche nicht-reproduzierbaren Variablen nennt man *random*.
- Sie brauchen eine spezielle Behandlung.
- Die moderne Lösung sind die *Linear Mixed Effects Models*.

Nicht normalverteilte abhängige Variablen[12]

- Alles, was wir bisher gemacht haben, nimmt normalverteilte Residuen an.
- Das gilt natürlich längst nicht für alle möglichen abhängigen Variablen:
  - Wovon hängt es ab, welche von zwei möglichen Konstruktionen ein Sprecher auswählt? (Abhängige Variable ist kategorial.)
  - Wovon hängt es ab, wie viele Fehler eine Versuchsperson macht? (Abhängige Variable ist binomialverteilt.)
- Für solche Fälle gibt es die Erweiterung vom

  General Linear Model

  auf das

  Generalized Linear Model

- Alle sind sich einig, dass das ein schlechter Name ist, aber so heißt es nun mal.

[12] Faraway 2006; Dobson und Barnett 2008; McCullagh und Nelder 1992; Firth 1991; Hinkley, Reid und Snell 1991.

Literatur I

Baayen, R. H. (2008). *Analyzing Linguistic Data – A practical introduction to statistics*. Cambridge University Press.

Baroni, Marco und Silvia Bernardini (2006). „A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text“. In: *Literary and Linguistic Computing* 21.3 (Sep. 2006), S. 259–274.

Bortz, J. (2005). *Statistik für Sozialwissenschaftler*. 6. Aufl. Heidelberg: Springer.

Dalgaard, Peter (2008). *Introductory Statistics with R*. 2. Aufl. New York: Springer.

Dobson, A.J. und A.G. Barnett (2008). *Introduction to Generalized Linear Models*. 3. Aufl. Chapman und Hall.

Faraway, Julian J. (2006). *Extending the linear model with R: generalized linear, mixed effects, and non-parametric regression models*. Chapman & Hall/CRC.

Literatur II

Firth, D. (1991). „Generalized linear models“. In: *Statistical Theory and Modelling. In Honour of Sir David Cox, FRS*. Hrsg. von D. V. Hinkley, N. Reid und E.J. Snell. London: Chapman & Hall.

Gries, Stefan Thomas (2008). *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoeck & Ruprecht.

Hinkley, D. V., N. Reid und E.J. Snell, Hrsg. (1991). *Statistical Theory and Modelling. In Honour of Sir David Cox, FRS*. London: Chapman & Hall.

McCullagh, P. und J. A. Nelder (1992). *Generalized Linear Models*. Chapman und Hall.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org`.

Dichotomizing continuous variables: a warning

Ein Zitat aus Baayen (2008). Vergleiche ◂ aber hier.

A final methodological issue that should be mentioned is the unfortunate practice in psycholinguistics of dichotomizing continuous variables. For instance, Baayen et al. [1997] studied frequency effects in visual word recognition by contrasting high-frequency words with low-frequency words. The two sets of words were matched in the mean for a number of other lexical variables. However, this dichotomization of frequency reduces an information-rich continuous variable into an information-poor two-level factor. If frequency were a treatment that we could administer to words, like raising the temperature or the humidity in an agricultural experiment, then it would make sense to maximize one's chances of finding an effect by contrasting observations subjected to a fixed very low level of the treatment with observations subjected to a fixed very high level of the treatment. Unfortunately, frequency is a property of our experimental units, it cannot...

...be administered independently, and it is correlated with many other lexical variables. Due to this correlational structure, dichotomization of linguistic variables almost always leads to factor levels with overlapping or nearly overlapping distributions of the original variable – it is nearly impossible to build contrasts for extreme values on one linguistic variable while matching for a host of other correlated linguistic variables. As a consequence, the enhanced statistical power obtained by comparing two very different treatment levels is not available. In these circumstances, dichotomization comes with a severe loss of statistical power, precise information is lost and nonlinearities become impossible to detect. Furthermore, samples obtained through dichotomization tend to be small and to get ever smaller the more variables are being matched for. Such samples are also non-random in the extreme, and hence do not allow proper statistical inference. To make matters even worse, dichotomization may also have various other adverse side effects, including spurious significance [see, e.g., Cohen, 1983, Maxwell and Delaney, 1993, MacCallum et al., 2002]. Avoid it. Use regression.

Kollinearität (##025##)

- Betrachten wir folgendes Modell

$$y = ax_1 + bx_2 + c + \texttt{normalverteiltes}$$

- Wenn nun x_1 und x_2 eng zusammenhängen, also

$$x_2 = dx_1 + e(+\texttt{normalverteiltes})$$

- Dann kann man das Modell anders schreiben:

$$\begin{aligned} y &= ax_1 + b(dx_1 + e) + c + \texttt{normalverteiltes} \\ &= (a + bd)x1 + be + c + \texttt{normalverteiltes} \end{aligned}$$

- Das ist aber nicht mehr linear!

$\Rightarrow$ Kollinearität zerstört die Linearität des Modells!