Statistische Modellbildung

Detlef Steuer steuer@hsu-hh.de

FT 2021

Struktur der Veranstaltung (11.4.2022)

Vorlesung und Übung

- Es wird ein Skript geben. Das Skript wird jeweils (hoffentlich) spätestens Freitag Abend online auf der Seite der Veranstaltung hochgeladen.
- Dort findet sich auch Material für die Übungen.
- Die Besprechung der Vorlesung und möglicher Übungsaufgaben findet in der vorgesehenen Zeit statt: Jeweils Montag 12:00-13:30 Uhr. Tendentiell werden während der Veranstaltung ebenfalls Übungsaufgaben bearbeitet.

Verwendete Werkzeuge

- Die Veranstaltung nutzt `R`, `RStudio` und `RMarkdown`.
- `RStudio` ist optional.

Warum RMarkdown?

- RMarkdown ist ein Dialekt von Markdown.
- In der Veranstaltung werden Dokumente in RMarkdown geschrieben und dann in PDF oder HTML umgewandelt.
- RMarkdown verbindet Markdown mit R, indem man R Programmcode und Formatierungsanweisungen mischen kann.
- Zur Veröffentlichung wird mittels pandoc der Quelltext z.B. in LaTeXübersetzt.
- Dadurch wird sogenanntes Literate Programming möglich!

Die Kombination von R und LaTeX ergibt Reproducible Research!

- Der Ursprung liegt im *literate programming* von Donald Knuth bereits um 1980.
  Programs are meant to be read by humans, and only incidentally for computers to execute. (Donald Knuth)

- Im Kern geht es bei *Reproducible Research* darum, Analysen in ihrer Gesamtheit einem zukünftigen Rezipienten an die Hand zu geben. Die Wiederholbarkeit wird erreicht indem sowohl die Daten *vollständig* vorhanden sind, als auch die Analyse *vollständig* und *vollständig dokumentiert* vorliegt. Die Analyse umfasst (idealerweise) auch die das *computational environments*, welches zur Analyse genutzt wird.

Der Mechanismus von Literate Programming

- Ausgangspunkt ist ein Dokument, welches Code und Kommentare enthält.
- Die einzelnen Codeteile werden *chunks* genannt.
- Zur Übersetzung des Dokuments wird es in seine Teile zerlegt. Die Codechunks werden in einer `R` Sitzung ausgeführt.
- Die einzelnen Teile werden in der richtigen Reihenfolge wieder zusammengesetzt, die Codechunks aber jeweils durch das Ergebnis ihrer Auswertung ersetzt.
- In `R` ist dafür das Paket `knitr` zuständig.
- Wir sehen das alles gleich in Aktion!

Warum Reproducible Research?

Non-reproducible single occurrences are of no significance to science. (Karl Popper)

- Im Jahr 2012 wurde eine Studie von Begley und Ellis in Nature veröffentlicht, die ein Jahrzehnt Forschung untersuchte. Diese Studie fand, dass 47 von 53 Papieren zu medizinischer Krebsforschung nicht reproduzierbar waren.
- In letzter Zeit wird das Problem auch bei vielen Studien zu AI festgestellt.
- Es gibt eine wirkliche Krise der wissenschaftlichen Methode.
- In der Pharmakologie ist die Wiederholbarkeit der Schlüssel zur Zulassung neuer Medikamente.

Die Werkzeuge (standing on the shoulders of giants)

- TEX bzw. LATEX (Donald Knuth)
- pandoc (John MacFarlane)
- `R` (The R-core team)
- RStudio (Hadley Wickham, ggplot, dplyr, devtools etc.)
- `knitr` (Yihui Xie)
- SWeave (Fritz Leisch)
- git (Linus Torvalds)
- Jabref (Team aus D: Oliver Kopp (ab 2011), Stefan Kolb (ab 2015), Matthias Geiger (ab 2015), Tobias Diez (ab 2015), Christoph Schwentker (ab 2016), Linus Dietz (ab 2017), Carl Christian Snethlage (ab 2020), Dominik Voigt (ab 2020)) oder
- zotero (Ray Rosenzweig Center for History and New Media, George Mason University)

RMarkdown, ein erstes Dokument

- RMarkdown hat ebenfalls einen Header. Dort finden sich z.B. das Zielformat, der Autor etc.
- Das Format ist das sog. YAML (yet another markup language).

```
---
title: "MarkdownIntro"
author: "Detlef Steuer"
date: "20 1 2021"
output: html_document
---
```

- Danach folgt in RMarkdown formatierter Klartext.
- Beschreibung der Möglichkeiten unter http://rmarkdown.rstudio.com.
- Dateiendung: .Rmd

Der knitr-Zauber

- knitr kann nicht nur `R`, sondern auch python, Julia, sql, bash oder C Programmblöcke
- knitr verwaltet die Grafiken
- knitr cached die Ergebnisse von Rechnungen
- knitr erlaubt Verweise auf Codeblocks an anderer Stelle im Dokument
- knitr controlliert den Ablauf all der beteiligten externen Programme!

Übung 11.4.2022

Lesen Sie
https://rmarkdown.rstudio.com/articles_intro.html und
spielen Sie das dortige Beispiel durch.

Ziel der Veranstaltung

- Herausarbeiten der speziellen statistischen Herangehensweise an die Modellbildung.
- Einbeziehung der Prinzipien des Reproducible Research.
- Bewertungskriterien von konkurrierenden Modellen.
- Schwerpunkte: Multiple Regression, Nichtlineare Regression, Varianzanalyse, Diskriminanzanalyse, Saisonale Modelle, Modellauswahl, Informationsmaße

Literatur

- Hain, Statistik mit R, RRZN Hannover 2011 (über die Uni erhältlich)
- Dalgaard, Introductory statistics with R, Springer (elektronisch über die Bibliothek verfügbar)
- Faraway, Linear Models in R, Chapmann and Hall
- Ligges, Programmieren in R, Springer (elektronisch über die Bibliothek verfügbar)
- Literatur für den ersten Teil der Vorlesung, Beispiele sind dort zum Teil entnommen
- Yihui Xie (2012). knitr: A general-purpose package for dynamic report generation in R. R package version 0.6. http://CRAN.R-project.org/package=knitr
- Statistische Modellbildung, Walter Gruber, 2019
- Reichhaltige Informationen im Netz!

Was kennzeichnet Statistische Modellbildung?

- Donald Knuth hat die Datenanalyse als Kunst bezeichnet, um diese von der reinen Wissenschaft zu unterscheiden. (Programming as an art, Knuth 2008)
- Die statistische Modellbildung strebt nicht hauptsächlich danach, wissenschaftliche Wahrheiten zu entdecken, sondern danach, die vorhandenen Daten zu strukturieren und durch einfache Formel(n) die Zusammenhänge zwischen den beobachteten Variablen zu beschreiben.
- Durch diese Beschreibung wird eine Abstraktion weg vom Einzelfall geschaffen, was ermöglicht, gewinnbringend Folgerungen aus Daten zu ziehen.
- Diese Abstraktion rechtfertigt den Begriff des Modells!

Beispiel für Statistische Modellbildung

- Beispiel Maskenpflicht in der Pandemie. Der genaue physikalische Zusammenhang zwischen Virusgröße, Viruslast, Filtermaß, Passform, korrekter Sitz, etc. ist extrem schwer zu modellieren, wenn man die "Wahrheit" sucht. Zählt man jedoch einfach die Infektionen in Gruppen mit und ohne Masken, so ist eine Wirkung klar.
- Natürlich ist man hocherfreut, wenn man klar interpretierbare Modelle findet.
- Es reicht in der Statistik ein Modell zu finden, dass für eine gegebene Fragestellung gut genug ist!
- Für die kinetische Energie beim Zusammenprall zweier Fahrzeuge ist es völlig ok, keine relativistischen Effekte mit einzurechnen, auch wenn die "ganze Wahrheit" dies erfordern würde.

Kontrast zum Theoriegetriebenen Modell

- Gerade in der VWL sind Modelle, die aus einer theoretischen Begründung folgen sehr verbreitet.
- Das Modell wird in diesen Fällen ohne Daten postuliert.
- Aus Statistikersicht ist das völlig ok, solange die Empirie nicht gegen das Modell spricht. Dann hat die Theorie einen Fehler ("Schwachpunkt")!
- Hier geht es nicht zuerst um Nützlichkeit, sondern um wissenschaftliche Erkenntnis, also einen Wahrheitsanspruch.
- Wenn die Theorie der Empirie zu stark entgegensteht, verlangt der westliche Wissenschaftsbegriff (Popper) eine Korrektur der Theorie! (Beispiel Mindestlohn!)

Der Begriff des Statistischen Modells

- Herkunft aus der Mathematischen Statistik: Es ist ein Tripel aus einer Menge von Elementarereignissen, eine σ-Algebra dieser Ereignisse und einer Menge von möglichen Wahrscheinlichkeitsmaßen über diesem Mengensystem.
- Die Modellbildung versucht in gewisser Weise ein bestes Wahrscheinlichkeismaß aus der Menge der möglichen Maße zu identifizieren.
- Im Modell stecken Annahmen über den Daten generierenden Prozess.
- Sind die Daten stetig, diskret? Kennt man eine parametriesierte Verteilung oder muss man nichtparametrisch arbeiten? Habe ich Vermutungen über die Struktur des Zusammenhangs oder nicht?

Mögliche Strukturannahmen

- Das lineare Modell unterstellt metrische Ziel- und Einflussgrößen und normalverteilte, unabhängige Fehler
- Ein Modell für einen Würfelwurf die Gleichverteilung über die Würfelseiten
- Ein neuronales Netz verzichtet auf ein interpretierbares Modell zugunsten eines riesigen Parameterraumes, um eine möglichst gute Prognose zu erreichen.
- Nichtparametrische Glätter verzichten auf eine interpretierbare funktionale Beschreibung, überwinden dafür aber die Nutzung desselben Zusammenhanges über den ganzen Beobachtungsraumes.
- Von black-box Modellen spricht man, wenn auf eine Erklärung des Ergebnisses durch die Modellannahmen verzichtet wird.
- Großes Problem für ADM!

Der Prozess der Modellbildung

- “Bildung” ist nichts fertiges, sondern eine aktive Tätigkeit
- Zu Beginn der Modellbildung steht die Fragestellung, die mit dem Modell beantwortet werden soll. Im Optimalfall sogar vor der Datensammlung.
- Die Datensammlung! Wo schaue ich nach welchen Daten? Ist die Qualität der Daten gut genug für den geplanten Zweck? (Luca-App fail!)
- Die deskriptive und explorative Datenanaylse, als das Anschauen der Daten! Wichtig: Die Hypothese, welches Modell zu den Daten passen könnte, sollte nicht hier entwickelt werden, da ansonsten die p-Werte nicht zuverlässig sind. Passiert aber trotzdem, da keine anderen Daten vorhanden! Eigentliches Ziel ist ein Desastercheck.
- Statistische Inferenz

Worum es in der Vorlesung nicht geht

- ML machine learning
- AI artificial intelligence (auch eigentlich nur Statistik)
- Kollege Adämmer hält dieses Semester die entsprechende Vorlesung!

Laborbuch für die Veranstaltung (25.4.22)

- Die Live-Programmierung passiert in einer R-Markdown-Datei.
- Jeder lege sich ein Verzeichnis und eine Datei an.
- Sinnvollerweise wird von jeder Software die Versionsnummer notiert!

Ein erster Datensatz

- Ziel ist zunächst die Untersuchung der Daten. Qualität und Geeignetheit.
- Dazu gehört natürlich die, soweit möglich, "Reparatur" der Daten im Sinne der Korrektheit und Einheitlichkeit.
- Beispieldaten aus Faraway, Linear Models in R

```
> install.packages("faraway")
### besser:
### if (! ("faraway" %in% installed.packages())){
### install.packages("faraway") }
> require(faraway)
> data(pima)
```

- Studie des National Institute of Diabetes and Digestive and Kidney Diseases an 768 erwachsenen Frauen der Pima Indianer.

Hintergrundinformation

- Wegweisende Studie über den Zusammenhang von Diabetes mit genetischen Ursachen.
- Pima Indianer haben die weltweit höchste Diabetesrate.
- Sie sind in der Nähe von Phoenix/Arizona beheimatet.
- Infos über die Daten

  ```
  > help(pima)
  ```

Der erste Blick auf die Daten

- `str(pima)`

```
'data.frame': 768 obs. of  9 variables:
 $ pregnant : int  6 1 8 1 0 5 3 10 2 8 ...
 $ glucose  : int  148 85 183 89 137 116 78 115 197 125
 $ diastolic: int  72 66 64 66 40 74 50 0 70 96 ...
 $ triceps  : int  35 29 0 23 35 0 32 0 45 0 ...
 $ insulin  : int  0 0 0 94 168 0 88 0 543 0 ...
 $ bmi      : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3
 $ diabetes : num  0.627 0.351 0.672 0.167 2.288 ...
 $ age      : int  50 31 32 21 33 30 26 29 53 54 ...
 $ test     : int  1 0 1 0 1 0 1 0 1 1 ...
```

- Man könnte 768 Beobachtungen noch einzeln durchgucken. Man kann es sich aber auch leichter machen!
- Handarbeit ist schlecht! Und widerspricht auch den Prinzipien des Reproducible Research.

Exploration der Daten

```
> summary(pima)
    pregnant         glucose        diastolic         triceps
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
    insulin           bmi           diabetes           age
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00
 Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
 Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
      test
 Min.   :0.000
 1st Qu.:0.000
 Median :0.000
 Mean   :0.349
 3rd Qu.:1.000
 Max.   :1.000
```

Was fällt auf?

- 17 Schwangerschaften ist ungewöhnlich, aber nicht ausgeschlossen!
- Blutdruck 0 ist ungesund, ebenso BMI 0 . . .

```
> pima$diastolic
```

- Wie viele sind es nun genau? Achtung: wichtiger Trick!

```
> sum(pima$diastolic == 0)
[1] 35
```

- Vermutlich sind in der Studie fehlende Werte als 0 festgehalten worden. Die Daten müssen curatiert werden.

Daten “reparieren”

- Veränderungen werden immer nur an Arbeitskopien vorgenommen!

  ```
  > lpima <- pima # Arbeitskopie!
  ```

- In der Arbeitskopie werden die fehlenden Werte durch die korrekte Kodierung `NA` ersetzt.

  ```
  lpima$diastolic[ lpima$diastolic == 0 ] <- NA
  lpima$glucose[ lpima$glucose == 0 ] <- NA
  lpima$triceps [ lpima$triceps == 0 ] <- NA
  # or use replace
  lpima$insulin <- replace(lpima$insulin,
   lpima$insulin == 0 , NA)
  lpima$bmi [ lpima$bmi == 0 ] <- NA
  ```

Oder eleganter

```
lpima <- within ( lpima , { # Rekodieren der NAs
diastolic <- replace(diastolic, diastolic == 0, NA),
glucose <- replace( glucose, glucose == 0, NA)
triceps <- replace( triceps, triceps == 0, NA)
insulin <- replace( insulin, insulin == 0, NA)
bmi <- replace ( bmi, bmi == 0, NA)
})
```

Summary der korrigierten Daten

```
summary(lpima)
    pregnant         glucose       diastolic         triceps
 Min.   : 0.000   Min.   : 44.0   Min.   : 24.00   Min.   : 7.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:22.00
 Median : 3.000   Median :117.0   Median : 72.00   Median :29.00
 Mean   : 3.845   Mean   :121.7   Mean   : 72.41   Mean   :29.15
 3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:36.00
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
  NA's   :5       NA's   :35       NA's   :227
    insulin            bmi          diabetes          age
 Min.   : 14.00   Min.   :18.20   Min.   :0.0780   Min.   :21.00
 1st Qu.: 76.25   1st Qu.:27.50   1st Qu.:0.2437   1st Qu.:24.00
 Median :125.00   Median :32.30   Median :0.3725   Median :29.00
 Mean   :155.55   Mean   :32.46   Mean   :0.4719   Mean   :33.24
 3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
 Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
 NA's   :374      NA's   :11
      test
 Min.   :0.000
 1st Qu.:0.000
 Median :0.000
 Mean   :0.349
 3rd Qu.:1.000
 Max.   :1.000
```

Weiterer Schwachpunkt der Daten:

- Die Variable `test` wird in der Zusammenfassung als numerischer Wert behandelt, obwohl es sich um eine kategorielle Variable handelt.
- In R werden solche Variablen *factor* genannt und können außerdem beschreibende Werte (Faktorstufen , factor levels) erhalten.

```
> lpima$test <- factor(pima$test)
> levels(lpima$test) <- c("negativ", "positiv")
> summary(lpima$test)
negativ positiv
500
268
```

- Allein an der Zusammenfassung der Daten kann man nunmehr keine Unregelmäßigkeiten mehr entdecken.

Andere Werkzeuge zur Datenbereinigung (2.5.2022)

- Die Daten waren schon aufbereitet, also fast klinisch perfekt. Nur Kleinkram musste noch nachgearbeitet werden.
- Das Vorbereiten der Daten ist normalerweise sehr viel aufwändiger.
- In `R` ist `readLines()` die Funktion, um Textdaten unstrukturiert einzulesen und dann in `R` aufzubereiten. Allerdings für große Daten relativ langsam.
- Es ist in der Praxis sehr hilfreich (mindestens) eine Skriptsprache neben `R` zu kennen, um Textfiles zu bearbeiten.
- Empfehlung: awk, relativ leicht zu lernen, oder Perl, sehr mächtig, oder moderener Python
- Allen ist gemein, dass sie speziell für sehr große Textdateien entwickelt wurden.

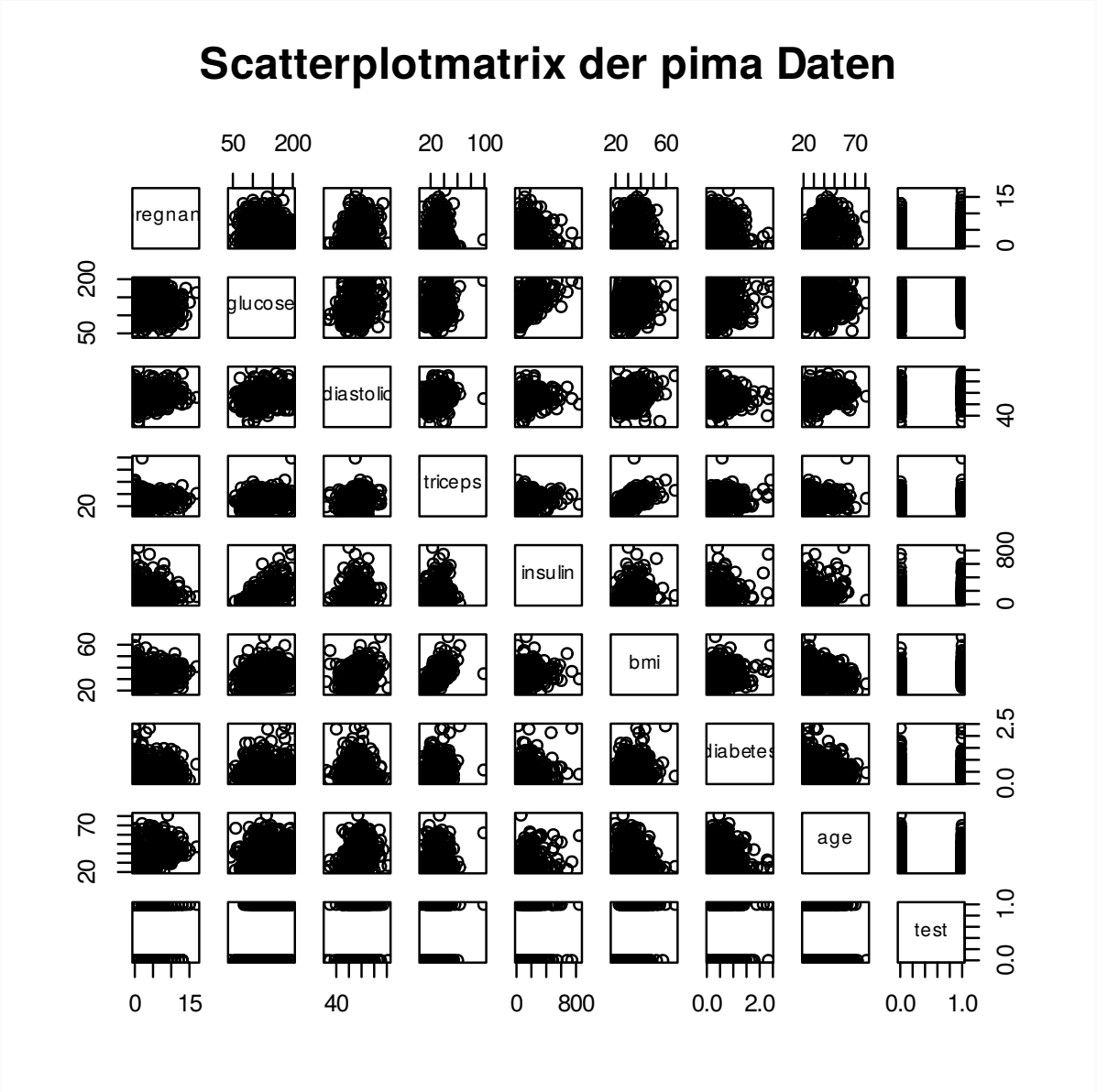
Grafische Exploration der Daten

- Analog zu `summary()` ist `pairs()` oder `plot()` immer ein erster guten Schritt.

```
plot(lpima, main="Scatterplotmatrix der pima Daten",
     pch=".")
```

- Alle Variablen auf einen Blick
- Ausreißer oft gut sichtbar

Scatterplotmatrix



Identifikation des Ausreißers

- Der maximale Wert von Triceps liegt weit außerhalb aller anderen Messungen
- ```
  # Identifikation der Beobachtung
  which.max(lpima$triceps)
  # und entfernen aus dem Datenpool
  # evtl auch 99 als NA Kodierung?
  lpima <- lpima[-580,]
  ```
  ```

Das complete.cases() Kommando

- Fehlende Werte machen viele Berechnungen schwierig und müssen immer beachtet werden.
- Um nur Beobachtungen eines Dataframes ohne NA zu betrachten kann der Befehl complete.cases() genutzt werden.
- Z.B.

```
cpima <- lpima[complete.cases(lpima),]
cor(cpima)
```

Korrelationsmatrix

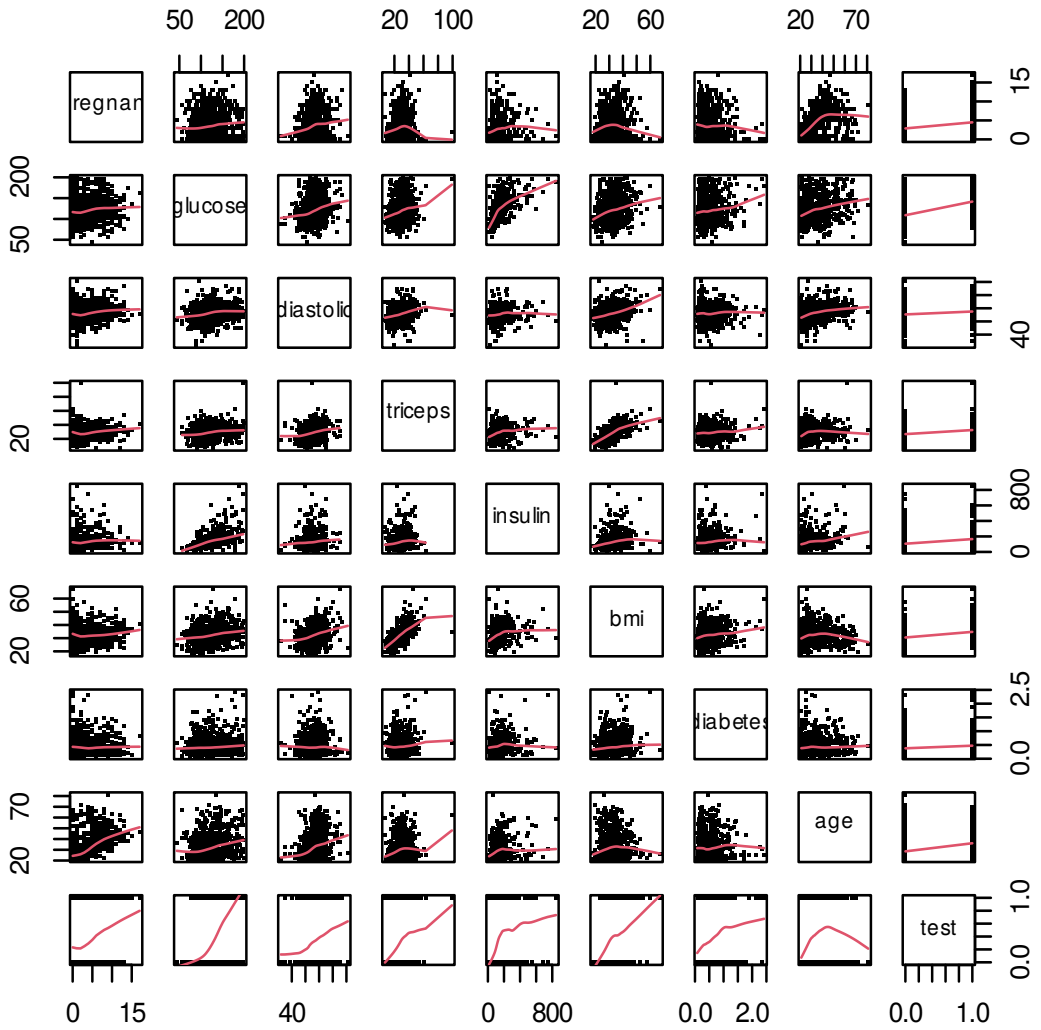
```
#           pregnant   glucose  diastolic   triceps    insulin
pregnant   1.000000000 0.1982910  0.2133548 0.0932094 0.07898363
glucose    0.198291043 1.0000000  0.2100266 0.1988558 0.58122301
diastolic  0.213354775 0.2100266  1.0000000 0.2325712 0.09851150
triceps    0.093209397 0.1988558  0.2325712 1.0000000 0.18219906
insulin    0.078983625 0.5812230  0.0985115 0.1821991 1.00000000
bmi       -0.025347276 0.2095159  0.3044034 0.6643549 0.22639652
diabetes   0.007562116 0.1401802 -0.0159711 0.1604985 0.13590578
age        0.679608470 0.3436415  0.3000389 0.1677611 0.21708199
test       0.256565956 0.5157027  0.1926733 0.2559357 0.30142922
#                 bmi     diabetes        age      test
pregnant  -0.02534728  0.007562116 0.67960847 0.2565660
glucose    0.20951592  0.140180180 0.34364150 0.5157027
diastolic  0.30440337 -0.015971104 0.30003895 0.1926733
triceps    0.66435487  0.160498526 0.16776114 0.2559357
insulin    0.22639652  0.135905781 0.21708199 0.3014292
bmi        1.00000000  0.158771043 0.06981380 0.2701184
diabetes   0.15877104  1.000000000 0.08502911 0.2093295
age        0.06981380  0.085029106 1.00000000 0.3508038
test       0.27011841  0.209329511 0.35080380 1.0000000
```

Streudiagramme mit Glättungskurve

- Zusätzlich zu den Punkten wird eine nichtparametrische, glatte Kurve durch die Punktwolken gelegt. Die Linie dient lediglich dem Auge, um evtl. Muster in den Daten zu erkennen.
- Als Scatterplotmatrix mittels der pairs() Funktion

```
pairs( lpima,
  panel= function( x, y) {
  panel.smooth( x, y, span= 2/3, pch=".") })
```

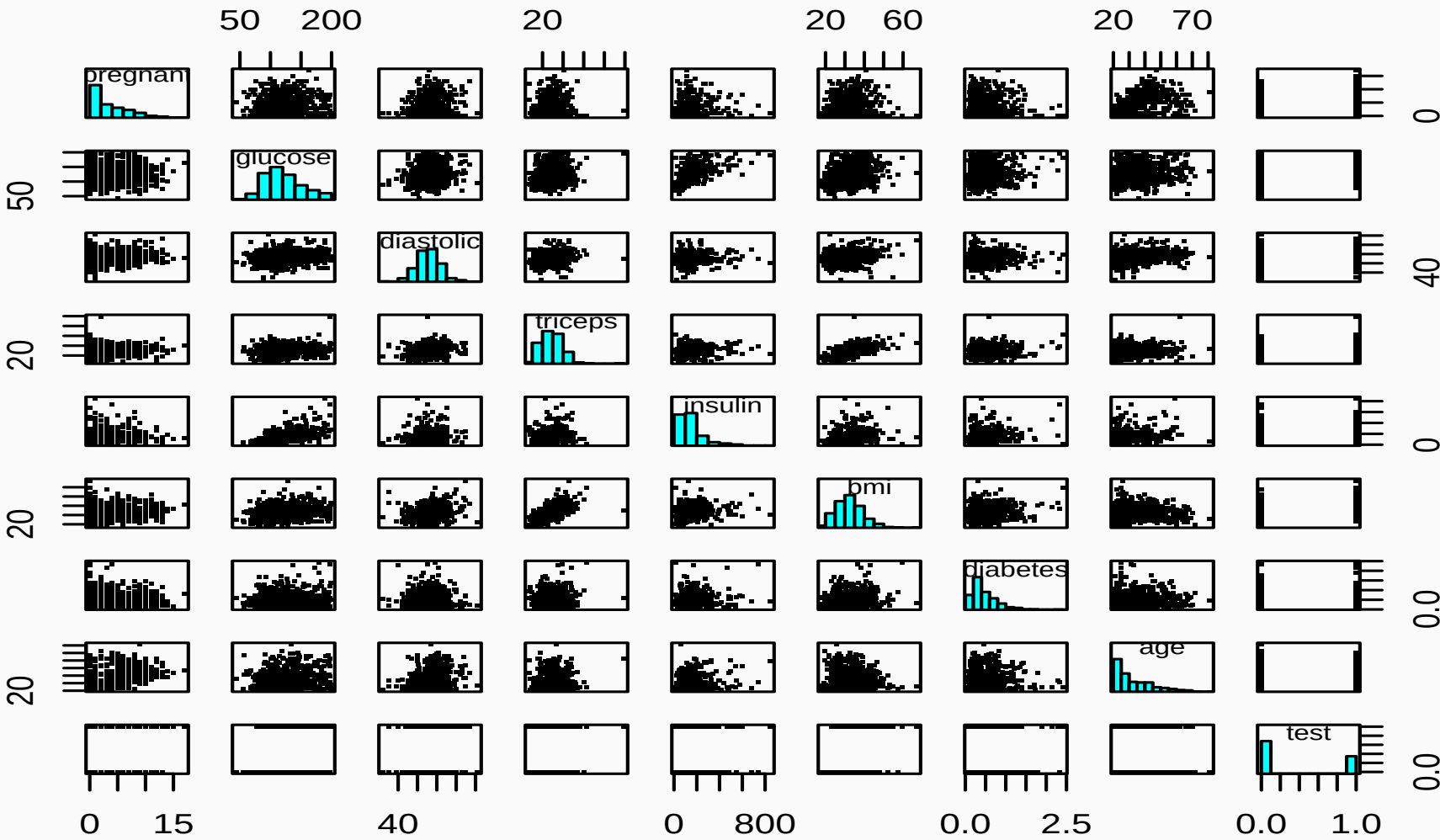
Scatterplotmatrix mit Glätter



Scatterplotmatrix mit Histogram auf der Diagonalen

```
panel.hist <- function(x, ...) {
    usr <- par("usr")
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}
pairs( lpima, panel= function( x, y) {
  panel.smooth( x, y, span= 2/3, pch=".")}
  diag.panel=panel.hist , pch=".")
```

Scatterplotmatrix mit Histogram

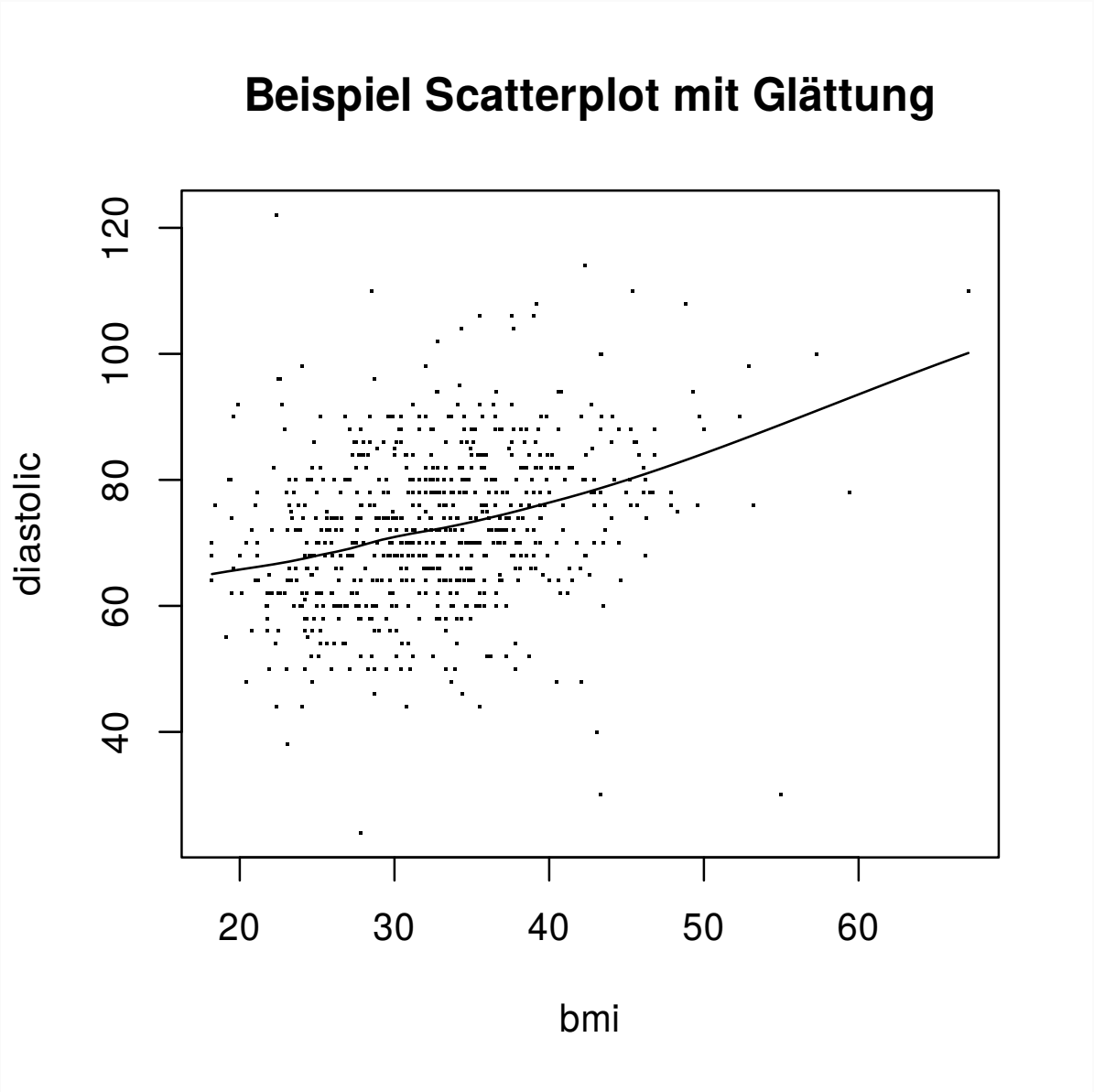


Zoom auf die Variablen bmi und diastolic

- Für ein Variablenpaar

```
attach(lpima)
scatter.smooth(bmi, diastolic,
  main="Beispiel Scatterplot mit Glättung", pch=".")
detach("lpima")
```

Scatterplot mit Glätter



Grundlagen: Warum funktioniert Statistische Modellbildung?

- Es gibt zwei große Sätze, die einen großen Teil der Statistik erst ermöglichen.
- Das Gesetz der großen Zahlen besagt im Wesentlichen, dass sich die Verteilung des Mittelwertes einer Stichprobe unter sehr schwachen Vorraussetzungen einer Normalverteilung nähert.
- Der Satz von Gliwenko-Cantelli, auch Hauptsatz der Mathematischen Statistik genannt, besagt, dass die empirische Verteilungsfunktion einer Stichprobe sich der theoretischen Verteilungsfunktion beliebig nähert.

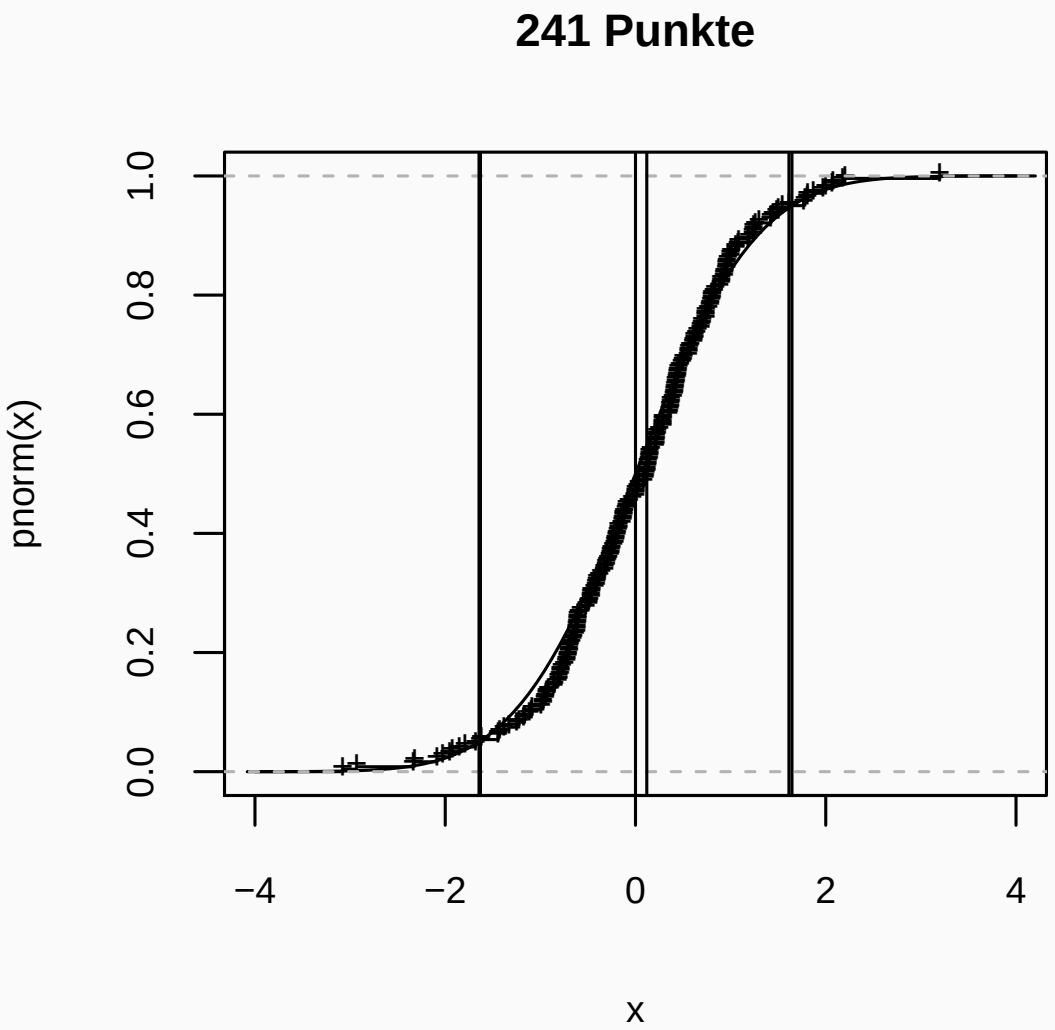
Exkurs: Simulation Gliwenko-Cantelli

- Die Ordnungsstatistiken $x_{(i)}$
  Zu jeder Stichprobe paarweise verschiedener $x_i, i = 1, \ldots, n$ gehört die Folge der Ordnungsstatistiken $x_{(i)}, i = 1, \ldots, n$, die die aufsteigend sortierte Folge der Beobachtungen bezeichnet. Die erste Ordnungsstatistik $x_{(1)}$ ist gleich dem Minimum der Beobachtungen, $x_{(n)}$ gleich dem Maximum.
- Es gilt empirisch $\hat{F}(x_{(i)}) = \frac{i}{n}$.
- Es lassen sich natürlich entsprechende Zufallsvariablen $X_{(i)}$ für die Ordungsstatistiken definieren.
- Nach dem Satz von Gliwenko-Cantelli konvergiert die empirische Verteilungsfunktion $\hat{F}$ der Stichprobe $x_i, i = 1, \ldots, n$ an jeder Stetigkeitsstelle von F, der Verteilung der X_i, mit dem Stichprobenumfang n gegen die wahre Verteilung F.
- Ohne dieses Ergebnis gäbe es nur deskriptive Statistik!

Demo der Konvergenz in Gliwenko-Cantelli

```
par(ask=TRUE)
samplesize <- 1
while ( samplesize <250) {
  if (samplesize > 100) par(ask=FALSE)
  curve(pnorm(x), -4,4, main=paste(samplesize, "Punkte"))
  sample <- sort(rnorm(samplesize))
  lines(ecdf(sample), pch="+")
  abline(v=qnorm(c(0.05,0.5,0.95)))
  abline(v=c(sample[round(samplesize/20)],
   median(sample),
   sample[round(19*samplesize/20)], col = "red"  ))
  samplesize <- samplesize +10
}
dev.copy2pdf(file="approxdemo.pdf", width=4, height=4)
```

Demo der Approximation



Zusammenhang zur Idee des Q-Q Plots

- Damit gilt für hinreichend große n: $x_{(i)} \approx F^{-1}(\frac{i}{n})$.
- Für den Fall, dass die Ordnungsstatistiken $x_{(i)}$ mit den $\frac{i}{n}$ -Quantilen der korrekten Verteilung abgetragen werden, liegen die Punkte $(x_{(i)}, F^{-1}(\frac{i}{n}))$ auf der Hauptdiagonalen des Scatterplots.
- Diese Eigenschaft wird ausgenutzt, um einen "grafischen Anpassungstest", den Q-Q-Plot zu entwickeln.
- Je näher die empirischen Quantile an einer Geraden liegen, desto besser sind Verteilungsannahmen erfüllt.

Übungen 2.5.2022

- Beginnen Sie ein Markdown Dokument für dieses Trimester.
- Vollziehen Sie die Exploration des `pima` Datensatzen in diesem Markdown-Dokument nach.
- Bringen Sie die Simulation bei sich zum Laufen.
- Vergleichen Sie QQ-Plots von Normalverteilten Zufallszahlen mit t-Verteilungen mit verschiedenen Freiheitsgeraden.

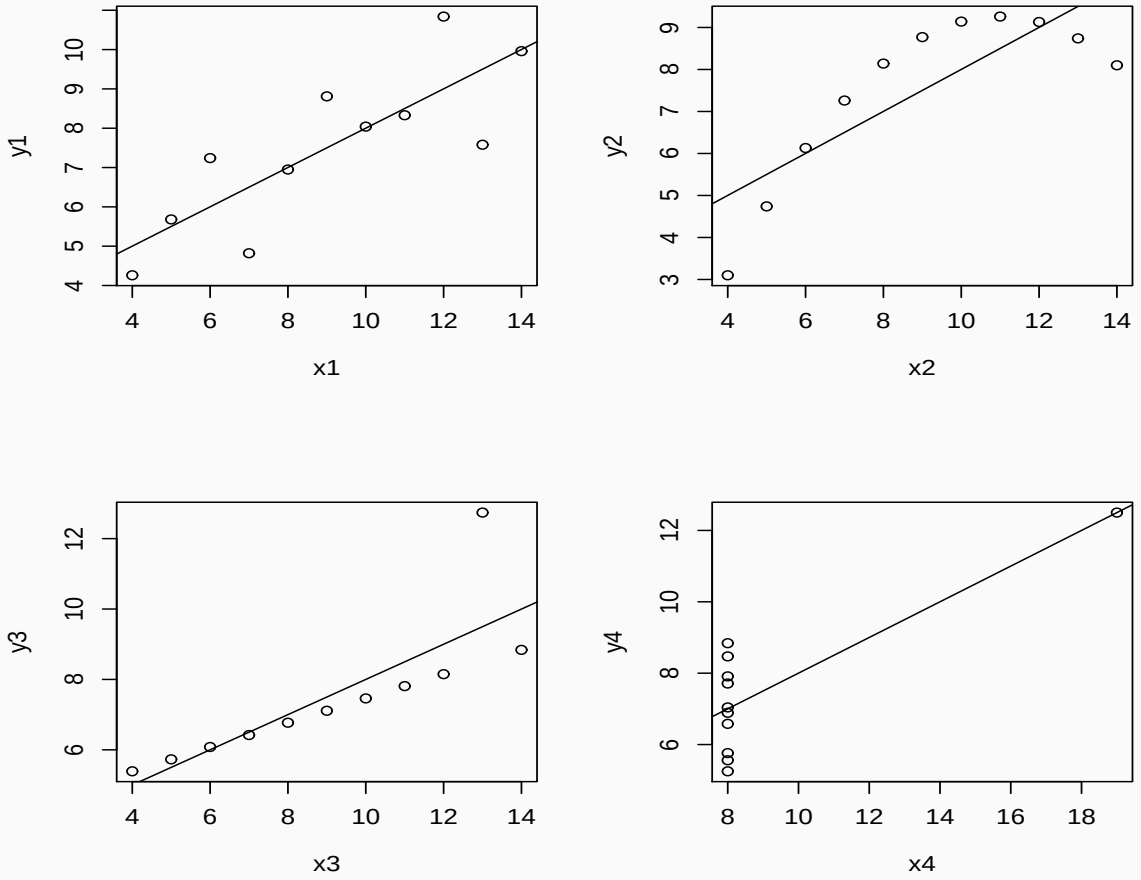
Regression

- Von Regression spricht man immer, wenn man eine Zielgröße $Y \in \mathcal{R}$ (abhängige Variable, Antwortvariable, response, output, dependent) durch eine oder mehrere Einstellgrößen $X_1, X_2, \ldots, X_p$ (unabhängige Variable, Einstellgröße, erklärende Variable, predictor, input, independent) durch einen unterstellten funktionalen Zusammenhang $Y = f(X)$ erklären oder modellieren möchte. Bei $p = 1$ spricht man von *einfacher Regression*, bei $p > 1$ von *multipler Regression*.
- Gibt es mehr als eine Zielgröße Y, so spricht man von *multivariater Regression*.
- Sind X und Y reellwertig, so liegt eine einfache Regression, wie in Statistik I+II, vor.
- Ist ein X_i qualitativ, so gelangt man zur (Ko-)varianzanalyse (ANOVA).

Das lineare Modell

- Ganz allgemein wird ein funktionaler Zusammenhang $Y = f(X_1, \ldots, X_p) + \varepsilon$ postuliert.
- Normalerweise ist f nicht bekannt und folglich nicht schätzbar.
- Beschränkung auf lineare Modelle $Y = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p + \varepsilon$
- Linear bezieht sich darauf, dass der Einfluß der β_i linear ist, nicht auf die Einflußgrößen selbst. Z.B. ist $Y = \beta \log(X) + \varepsilon$ ein lineares Modell oder auch $Y = \beta X^2 + \varepsilon$.
- Die Einschränkung auf lineare Modell ist in der Praxis nicht sehr streng. Manche Funktionen können in eine lineare Form transformiert werden und bei hinreichend glatten Funktionen ist die lineare Form oft eine gute Approximation (Taylor-Approximation).

Optischer Check: *Anscombe's quartet*



Alle Datensätze haben dieselben Werte für Mittel, Varianz und sogar dieselben Regressionsgeraden! (`data(anscombe)`)

Einfache lineare Regression

- Regression, also die Erklärung einer metrischen *Zielgröße*, auch abhängige Variable, durch eine *Einflußgröße* (auch Einstellgröße, Unabhängige) ist sicherlich die Methode der Statistik schlechthin.
- Generalvoraussetzung ab jetzt: $(x_1, y_1), \ldots, (x_n, y_n)$ sind eine gegebene Stichprobe vom Umfang n. Hierbei bezeichnet X die Einflußgröße und Y die Zielgröße, jeweils aus $\mathcal{R}$.
- Theorie bekannt aus Statistik, hier die Umsetzung in `R`.
- Beispiel `pima`-Daten. Aus der Scatterplotmatrix ist z.B. der Zusammenhang von `diastolic` und `bmi` interessant.

Lineare Regression in R

- Ziel: Schätzung von Parametern im linearen Modell

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

- Ergebnis: Modell (*fit*, *Anpassung*)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$$

- Residualvektor: $\varepsilon = (y - \hat{y})_{i=1,\ldots,n}$ mit der Fehlervarianz $\hat{\sigma}_\varepsilon$.
- β_0 heißt Achsenabschnitt, die β_i heißen Regressionskoeffizienten, β der Koeffizientenvektor.
- $X = (1, X_1, \ldots, X_p)$ heißt *Designmatrix*.
- Das Ganze in R: `lm( Y ~ X [, dataframe])`

Streuungszerlegung im linearen Modell

- Seien $SQT = \sum_1^n (y_i - \bar{y})^2$ *sum of squares total* oder Gesamtstreuung,
- $SQE = \sum_1^n (\hat{y}_i - \bar{y})^2$ *sum of squares explained* oder erklärte Streuung sowie
- $SQR = \sum_1^n (y_i - \hat{y}_i)^2$ *sum of squared residuals* oder Reststreuung.
- Dann gilt:

$$SQT = SQE + SQR!$$

(Übung!)

Einfache lineare Regression in R ($p = 1$)

```
> lm(diastolic ~ bmi, pima)
Call:
lm(formula = diastolic ~ bmi, data = pima)
Coefficients:
(Intercept)          bmi
    55.4869       0.5199
```

- Kommando `lm()` (linear model)
- Liefert die bekannten Schätzer (und mehr)
- Die Anzeige ist nicht das Ergebnis der Regression in R, sondern die Methode `print()` angewendet auf ein Objekt vom Typ Regression.
- Das Ergebnis eines `lm()` Aufrufs ist ein *Objekt der Klasse* `lm` !

Ein Regressionsobjekt in R

```
> result <- lm(diastolic ~ bmi, pima)
> names(result)
 [1] "coefficients"  "residuals"   "effects"     "rank"
 [5] "fitted.values" "assign"      "qr"     "df.residual"
 [9] "na.action"     "xlevels"     "call"        "terms"
[13] "model"
> result$coefficients
(Intercept)         bmi
 46.9717586   0.6918389
```

Schneller Überblick zur Regression

- `summary()` eine der wichtigsten Methoden in `R`

```
> summary(result)
Call: lm(formula = diastolic ~ bmi, data = pima)
Residuals:
     Min       1Q   Median       3Q      Max
-54.0807  -7.6278  -0.3313   7.2619  54.8676
Coefficients:
    Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.48694    2.11810  26.197  < 2e-16 ***
bmi          0.51989    0.06382   8.147 1.63e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 11.86 on 727 degrees of freedom
  (39 observations deleted due to missingness)
Multiple R-squared: 0.08365,Adjusted R-squared: 0.08239
F-statistic: 66.37 on 1 and 727 DF,  p-value: 1.630e-15
```

Interpretation der `summary`

- Zuerst steht die angewendete Modellgleichung.
- Dann die *five-number-summary* des Fehlervektors.
- Danach eine Tabelle mit je einer Zeile je geschätztem Parameter $\hat{\beta}_i$.
- Für jeden Parameter steht in der Zeile der Variablenname, die Schätzung $\hat{\beta}_i$, die Standardabweichung dieses Schätzers, die Teststatistik, die sich daraus ergibt, und der p-Wert unter der Nullhypothese $\beta_i = 0$.
- In der letzten Spalte finden sich die "Sternchen". Dort kann man für die üblichen Niveaus 10%, 5%, 1% und 0.1% direkt die Signifikanz eines entsprechenden Tests ablesen.

Interpretation der `summary`

- Weiter wird der Schätzer $\hat{\sigma}_\varepsilon$ mit den zugehörigen Freiheitsgraden angegeben und
- es wird auf die Anzahl von *missing values* hingewiesen.
- Abschließend sind noch das (multiple) Bestimmtheitsmaß R^2 bzw. R^2_{adj} und die F-Statisik zum sogenannte *Goodness-of-fit-test* angegeben.

Der p-Wert

- Die Spalte `Pr(> |t|)` gibt den sogenannte p-Wert zur Teststatistik an.
- Zur Erinnerung: Bei einem statistischen Test wird eine Hypothese $\mathcal{H}_0$ verworfen, wenn für eine realisierte Teststatistik T gilt, dass unter der der Nullhypothese die Wahrscheinlichkeit einer Realisierung in der gemessenen Größenordnung kleiner oder gleich dem festgelegten Niveau α ist. Dazu vergleicht man das zur Hypothese gehörende Quantil mit der beobachteten Teststatistik und entscheidet entsprechend.
- Dabei geht die Information verloren, wie nah die Realisierung an der kritischen Grenze beobachtet wurde.
- Der `p-Wert` gibt nun genau das Niveau eines Testes an, bei dem Teststatistik und kritischer Wert exakt zusammen fallen würden.

Das Bestimmtheitsmaß R^2

- In der einfachen Regression (eine Einflußgröße) ist das Bestimmtheitsmaß R^2 definiert als

$$R^2 = 1 - \frac{SSR}{SST}.$$

- Man kann zeigen: $R^2 = r_{XY}^2$, wobei r_{XY} den empirischen Korrelationskoeffizienten bezeichnet.
- Werte liegen zwischen 0 (Modell erklärt keinen Varianzanteil) und 1 (Modell erklärt die Varianz vollständig)
- Multiples und adjustiertes R^2 werden bei der multiplen Regression betrachtet.

Der *Goodness-of-fit*-Test

- Heißt auch der *Overall-F-Test*.
- Überprüft wird die Hypothese H_0

$$H_0 : \beta_i = 0 \text{ für alle } i \quad \text{gegen} \quad H_1 : \beta_j \neq 0 \text{ für mindestens ein } j.$$

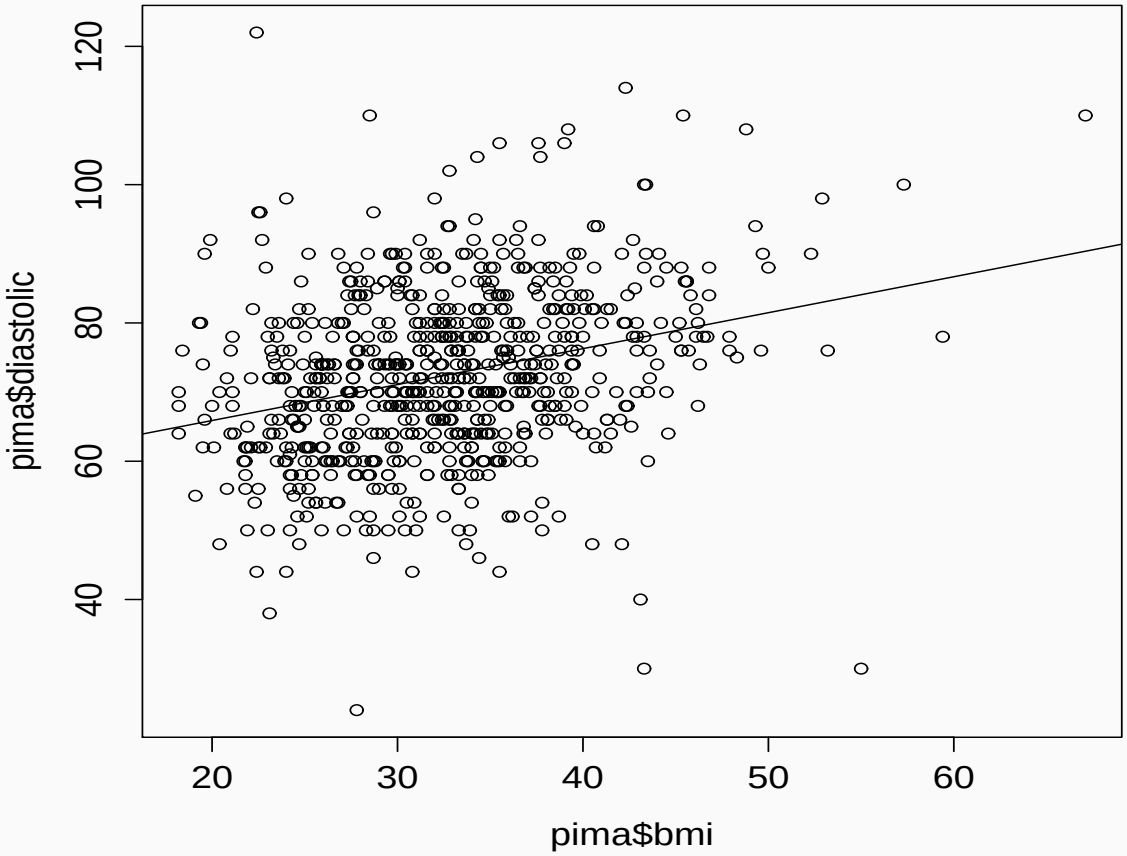
- Die Teststatistik ist in diesem Fall:

$$F = \frac{R^2}{1-R^2}\frac{n-p-1}{p} = \frac{SQE}{SQR}\frac{n-p-1}{p} \sim F(p, n-p-1) \text{ unter } H_0.$$

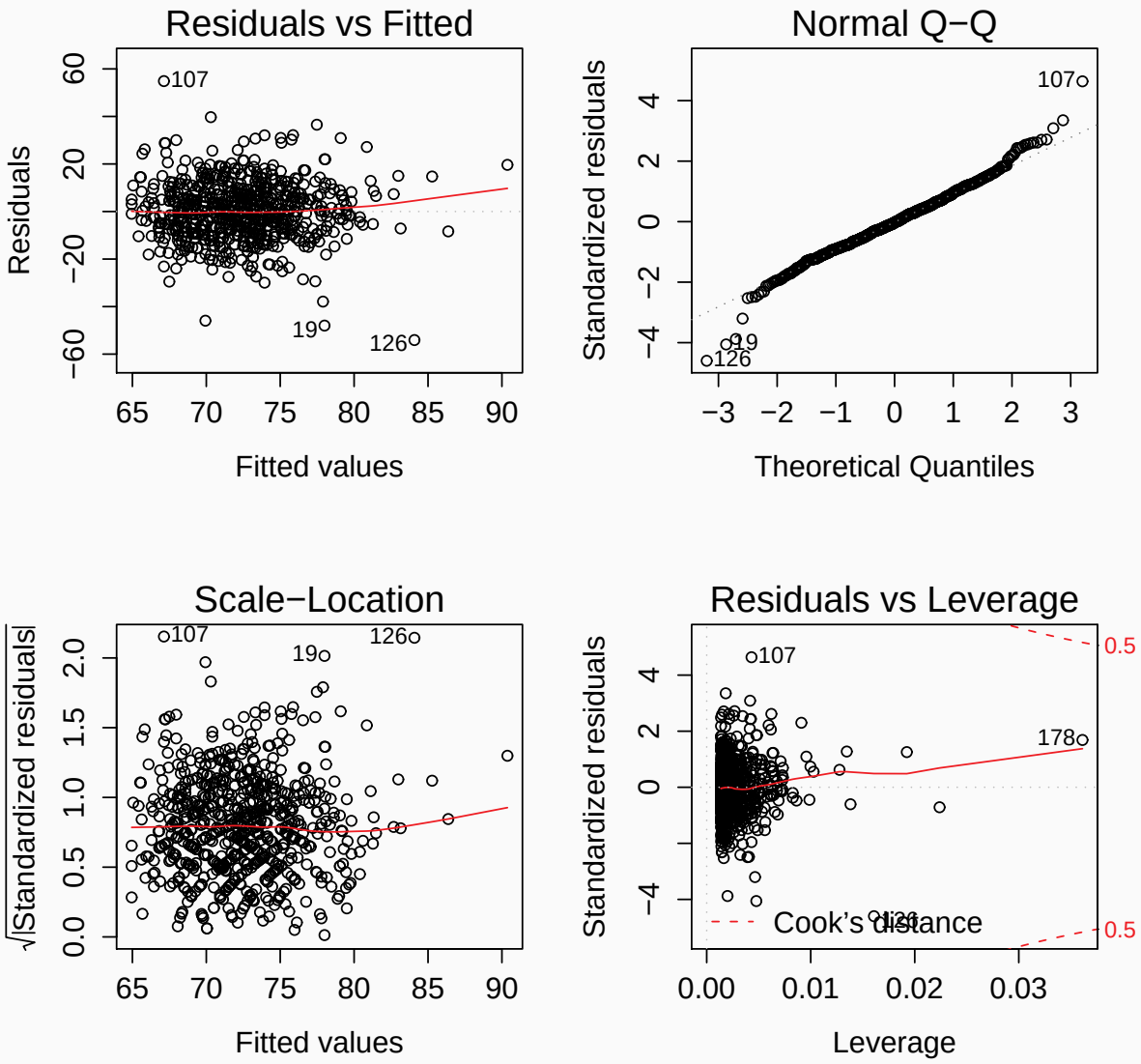
- Etwas irreführender Name! Es wird getestet, ob irgendeiner der Regressoren signifikanten Einfluß hat.

Grafiken zur Regression in R

```
> plot(pima$bmi, pima$diastolic) ; abline(result)
```



Residualanalyse – diagnostische Plots



Residualanalyse – diagnostische Plots

- Ergebnis von `plot(result)` sind *diagnostische Plots* zur optischen Beurteilung der Angemessenheit der Regression.
- In der linke Spalte sind Plots zur Beurteilung der Homoskedastizität (oben der sog. Tukey-Anscombe-Plot).
- Rechts oben ein sogenannter Q-Q Plot.
- Rechts unten ein *leverage* Plot (Einfluß einer Beobachtung).

Tukey-Anscombe-Plot

- Plot zur Beurteilung der Heteroskedastizität
- Insbesondere Hilfreich, wenn R^2 hoch, da sonst die y-Skala die Abweichungen von der Regressionsgeraden maskieren kann.
- x-Achse enthält die $\hat{y}_i$
- y-Achse die Residuen $y_i - \hat{y}_i$
- In R wird zusätzlich die lowess Kurve eingezeichnet
- Im Fall der Homoskedastizität sollen die Punkte des Plots in einem Band parallel zur x-Achse um den y-Wert 0 streuen.
- R markiert einige Punkte als Ausreißer.

Scale-Location-Plot

- Ebenfalls zur Beurteilung der Heteroskedastizität
- Varianzschätzer sind noch empfindlicher gegen Ausreißer als Mittelwertschätzer.
- x-Achse wieder $\hat{y}_i$
- y-Achse wird als Wurzel der standardisierten Residuen berechnet. Das ist nötig, da die beobachteten Fehler $\epsilon_i = y_i - \hat{y}_i$ als Zufallsvariablen betrachtet nicht i.i.d. sind.
- Es gilt $Var(R_i) = (1 - H_{ii})\sigma^2$, wobei $H_{ii} = \frac{1}{n} + \frac{(x_1 - \bar{x})}{\sum_1^n (x_i - \bar{x})^2}$ die sogenannte *leverage-/(Hebelarm-)Funktion ist. Deshalb definiere /standardisierte Residuen* als $\tilde{R}_i = \frac{R_i}{\sqrt{\hat{\sigma^2}(1 - H_{ii})}}$ mit identischer Varianz für alle i.

Residuals vs Leverage

- Ebenfalls ein Plot um Heteroskedastizität festzustellen
- Da die leverage eine Funktion des Abstands von x_i von $\bar{x}$ ist, sollten die Residuen zu einem x_i mit hohem leverage klein sein.
- *Klein* wird mit der sogenannten *Cook's-distance* gemessen.
- In R im Plot mit der gestrichelten Linie eingezeichent. Punkte außerhalb dieser Linie sind verdächtig bzw. deuten Modellverletzung an.

Der Q-Q Plot

- Beim Q-Q Plot werden die theoretischen Quantile einer Verteilung und die empirischen Quantile einer Stichprobe gegeneinander geplottet. Unter der Nullhypothese bildet dieser Graph eine Gerade.
- Erster Plot:

```
x <- rnorm(100, mean=2, sd=3)
qqnorm(x)
qqline(x)
```

- Zweiter Plot:

```
x <- rcauchy(100)
qqnorm(x)
qqline(x)
```

Vergleich der Q-Q-Plots

- links normalverteilt, rechts cauchy verteilt