LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Programmieren mit statistischer Software

Eva Endres, M.Sc.

Institut für Statistik
Ludwig-Maximilians-Universität München

Datenaufbereitung

The Forbes 2000 Ranking of the World's Biggest Companies I

- Paket installieren/laden

```
> # install.packages("HSAUR2")
> library(HSAUR2)
```

```
Warning: package 'HSAUR2' was built under R version 3.3.3
Loading required package: tools
```

- Datensätze im Paket HSAUR2

```
> data(package = "HSAUR2")
```

- Daten laden

```
> data("Forbes2000", package = "HSAUR2")
> ls()
```

```
[1] "Forbes2000"
```

- Hilfeseite zu den Daten öffnen

```
> # ?Forbes2000
> help("Forbes2000", package = "HSAUR2")
```

The Forbes 2000 Ranking of the World's Biggest Companies II

- Ausgabe der Klasse und der ersten Beobachtungen für jede Variable

```
> str(Forbes2000)

> str(Forbes2000, vec.len = 1) # verkuerzter Ausschnitt

'data.frame': 2000 obs. of  8 variables:
 $ rank       : int  1 2 ...
 $ name       : chr  "Citigroup" ...
 $ country    : Factor w/ 61 levels "Africa","Australia",..: 60 60 ...
 $ category   : Factor w/ 27 levels "Aerospace & defense",..: 2 6 ...
 $ sales      : num  94.7 ...
 $ profits    : num  17.9 ...
 $ assets     : num  1264 ...
 $ marketvalue: num  255 ...
```

The Forbes 2000 Ranking of the World's Biggest Companies III

- Daten anschauen

```
> print(Forbes2000)   # alle Zeilen

> head(Forbes2000)    # die ersten 6 Zeilen
```

```
  rank                 name        country                category  sales
1    1            Citigroup  United States                 Banking  94.71
2    2     General Electric  United States         Conglomerates 134.19
3    3 American Intl Group  United States             Insurance  76.66
4    4          ExxonMobil  United States Oil & gas operations 222.88
5    5                   BP United Kingdom Oil & gas operations 232.57
6    6      Bank of America  United States                 Banking  49.01
  profits  assets marketvalue
1   17.85 1264.03      255.30
2   15.59  626.93      328.54
3    6.46  647.66      194.87
4   20.96  166.99      277.02
5   10.27  177.57      173.54
6   10.81  736.45      117.55
```

The Forbes 2000 Ranking of the World's Biggest Companies IV

```
> tail(Forbes2000)    # die letzten 6 Zeilen
     rank                      name        country
1995 1995                      AMEC United Kingdom
1996 1996            Siam City Bank       Thailand
1997 1997         Yokogawa Electric          Japan
1998 1998       Hindalco Industries          India
1999 1999                    Nexans         France
2000 2000 Oriental Bank of Commerce          India
                        category sales profits assets marketvalue
1995                Construction  5.17    0.02   2.62        1.53
1996                     Banking  0.48    0.02  11.27        1.47
1997 Business services & supplies  2.78   -0.22   2.96        3.29
1998                   Materials  1.35    0.14   2.47        2.76
1999               Capital goods  5.09    0.00   2.71        0.88
2000                     Banking  0.81    0.10   7.16        1.17

> head(Forbes2000, n=2) #Ausgabe der ersten beiden Zeilen
> tail(Forbes2000, n=10) #Ausgabe der ersten 10 Zeilen
```

The Forbes 2000 Ranking of the World's Biggest Companies V

- Dimension des Datensatzes

```
> nrow(Forbes2000)     # Anzahl an Zeilen
[1] 2000
> ncol(Forbes2000)     # Anzahl an Spalten
[1] 8
> dim(Forbes2000)      # Anzahl an Zeilen und Spalten
[1] 2000    8
> # dim(Forbes2000)[1]
> # dim(Forbes2000)[2]
```

- Variablennamen

```
> names(Forbes2000)
[1] "rank"        "name"        "country"     "category"    "sales"
[6] "profits"     "assets"      "marketvalue"
```

Simple Summary Statistics I

- Summary: 5-Punkte-Zusammenfassung plus Mean für numerische Variablen, absolute Häufigkeiten für Faktoren

Simple Summary Statistics II

```
> summary(Forbes2000)
      rank            name                     country
 Min.   :   1.0   Length:2000        United States :751
 1st Qu.: 500.8   Class :character   Japan         :316
 Median :1000.5   Mode  :character   United Kingdom:137
 Mean   :1000.5                      Germany       : 65
 3rd Qu.:1500.2                      France        : 63
 Max.   :2000.0                      Canada        : 56
                                     (Other)       :612
                  category        sales            profits
 Banking                 : 313   Min.   :  0.010   Min.   :-25.8300
 Diversified financials: 158   1st Qu.:  2.018   1st Qu.:  0.0800
 Insurance               : 112   Median :  4.365   Median :  0.2000
 Utilities               : 110   Mean   :  9.697   Mean   :  0.3811
 Materials               :  97   3rd Qu.:  9.547   3rd Qu.:  0.4400
 Oil & gas operations    :  90   Max.   :256.330   Max.   : 20.9600
 (Other)                 :1120                     NA's   :5
     assets          marketvalue
 Min.   :   0.270   Min.   :  0.02
 1st Qu.:   4.025   1st Qu.:  2.72
 Median :   9.345   Median :  5.15
 Mean   :  34.042   Mean   : 11.88
 3rd Qu.:  22.793   3rd Qu.: 10.60
 Max.   :1264.030   Max.   :328.54
```

Subsets I

- Auswahl/Ausschluss von Zeilen

```
> Forbes2000[1:3, ]
  rank                 name       country      category  sales profits
1    1           Citigroup United States       Banking  94.71   17.85
2    2    General Electric United States Conglomerates 134.19   15.59
3    3 American Intl Group United States     Insurance  76.66    6.46
   assets marketvalue
1 1264.03      255.30
2  626.93      328.54
3  647.66      194.87
> Forbes2000[-(4:2000),]
  rank                 name       country      category  sales profits
1    1           Citigroup United States       Banking  94.71   17.85
2    2    General Electric United States Conglomerates 134.19   15.59
3    3 American Intl Group United States     Insurance  76.66    6.46
   assets marketvalue
1 1264.03      255.30
2  626.93      328.54
3  647.66      194.87
```

Subsets II

- Auswahl/Auschluss von Spalten

```
> Forbes2000[, 2] # nach Index
> Forbes2000$name # nach Name
> Forbes2000[, "name"] # nach Name
> Forbes2000[,c("name", "country")]
> Forbes2000[,-c(3:ncol(Forbes2000))] # nach Index
> # Forbes2000[, -"name"] # funktioniert nicht
> Forbes2000[,-which(names(Forbes2000) == "name")]
> Forbes2000[,-which(names(Forbes2000) %in% c("name", "country"))]
```

- Kombination aus beidem

```
> vars <- c("name", "sales", "profits", "assets")
> Forbes2000[1:3, vars]
                  name  sales profits  assets
1           Citigroup  94.71   17.85 1264.03
2    General Electric 134.19   15.59  626.93
3 American Intl Group  76.66    6.46  647.66
```

Subsets III

- Logische Vektoren

```
> Forbes2000$assets > 1000
```

```
> table(Forbes2000$assets > 1000)
```

```
FALSE  TRUE
 1997     3
```

- Auswahl über logische Bedingung (alle Einträge mit `TRUE` werden ausgewählt)

Subsets IV

```
> Forbes2000[Forbes2000$assets > 1000, "name"]
[1] "Citigroup"        "Fannie Mae"       "Mizuho Financial"
> # mit which Zugriff auf den Index
> Forbes2000[which(Forbes2000$assets > 1000), "name"]
[1] "Citigroup"        "Fannie Mae"       "Mizuho Financial"
> Forbes2000[Forbes2000$country == "Australia" & Forbes2000$rank < 100, ]
   rank                name   country category sales profits assets
86   86 Natl Australia Bank Australia  Banking 15.34    2.69 269.94
   marketvalue
86       36.51
> Forbes2000[Forbes2000$country == "Australia" | Forbes2000$rank < 100, ]
```

Subsets V

- Mit der `subset()`-Funktion

```
> # ?subset
> subset(x=Forbes2000, subset=Forbes2000$assets > 1000, select=vars)
                  name sales profits  assets
1           Citigroup 94.71   17.85 1264.03
9          Fannie Mae 53.13    6.48 1019.17
403 Mizuho Financial 24.40  -20.11 1115.90
```

Missing Values I

- Fehlende Werte sind mit NA kodiert ('not available' or 'missing value')

```
> summary(Forbes2000$profits)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
-25.8300   0.0800   0.2000   0.3811   0.4400  20.9600        5
```

- Der Umgang mit NAs kann in vielen Funktionen angegeben werden.

```
> mean(Forbes2000$profits)
[1] NA
> # ?mean
> mean(Forbes2000$profits, na.rm = TRUE)
[1] 0.3811328
> cor(Forbes2000$profits, Forbes2000$sales)
[1] NA
> # ?cor
> cor(Forbes2000$profits, Forbes2000$sales, use="pairwise.complete.obs")
[1] 0.4042672
```

Missing Values II

- `is.na()` identifiziert die fehlenden Werte

```
> which(is.na(Forbes2000$profits))
[1]  772 1085 1091 1425 1909
> # which(is.na(Forbes2000))    # nicht so hilfreich
> which(is.na(Forbes2000), arr.ind=TRUE)
      row col
772   772   6
1085 1085   6
1091 1091   6
1425 1425   6
1909 1909   6
> # Den Index der gueltigen Werte einer Variable erhaelt man mit
> which(!is.na(Forbes2000$profits))
```

- Vollständige Fälle

```
> complete.cases(Forbes2000$profits)  # vollstaendige Eintr\"age
> complete.cases(Forbes2000)          # vollstaendige Beobachtungen (Zeile
```

Missing Values III

- Löschen aller Beobachtungen mit fehlenden Werten (eher unüblich)

```
> # ?na.omit
> Forbes2000cc <- na.omit(Forbes2000)
> dim(Forbes2000cc)
[1] 1995    8
> mean(Forbes2000cc$profits)
[1] 0.3811328
```

Basic Data Manipulations I

- Erzeugen einer neuen Variable mit:
  `Datensatz$Variablenname <- ...`

```
> Forbes2000cc$costs <- Forbes2000cc$sales - Forbes2000cc$profits
```

- Löschen einer Variable mit:
  `Datensatz$Variablenname <- NULL`

```
> Forbes2000cc$category <- NULL
> dim(Forbes2000cc)
[1] 1995    8
```

- Gleichzeitiges Bearbeiten mehrerer Variablen
  - Manche Funktionen können gleichzeitig auf mehrere Spalten angewendet werden (hier Standardisieren)

```
> Forbes2000cc[, 4:7] <- scale(Forbes2000cc[, 4:7])
```

Basic Data Manipulations II

- Andere Funktionen können nur auf Vektoren angewendet werden

```
> median(Forbes2000cc[, 4:7])
```

```
Error in median.default(Forbes2000cc[, 4:7]): need numeric
data
```

```
> c(median(Forbes2000cc[, "sales"]),
+   median(Forbes2000cc[, "profits"]))
```

```
[1] -0.2967792 -0.1025984
```

- Alternative mit `apply` bzw. `sapply`
  (Details später im Kapitel zur Vekorisierung)

```
> # ?apply
> apply(X=Forbes2000cc[, 4:7], MARGIN=2, FUN=median)
```

```
      sales     profits      assets marketvalue
 -0.2967792  -0.1025984  -0.2478837  -0.2756262
```

```
> # ?sapply
> sapply(X=Forbes2000cc[, 4:7], FUN=median)
```

```
      sales     profits      assets marketvalue
 -0.2967792  -0.1025984  -0.2478837  -0.2756262
```

Basic Data Manipulations III

- Simultane Analyse einer Zeile - schwierig, da meist Variablen unterschiedlicher Klassen

```
> Forbes2000cc[722, ]
```

```
    rank              name country      sales   profits    assets
722  722 Skandia Insurance  Sweden 0.04555149 -0.499097 0.2117475
    marketvalue costs
722   -0.287061 11.03
```

- Manipulation von Zellen

```
> Forbes2000cc[722, "sales"] <- 0.05
> Forbes2000cc[722, ]
```

```
    rank              name country sales   profits    assets marketvalue
722  722 Skandia Insurance  Sweden  0.05 -0.499097 0.2117475   -0.287061
    costs
722 11.03
```

Saving data I

- Speichern eines einzelnen R-Objekts in einer Datei

```
> # ?saveRDS
> saveRDS(Forbes2000cc, file = "Forbes2000cc.rds")
> rm(Forbes2000cc)
```

- Laden des gespeicherten Objekts

```
> Forbes2000cc <- readRDS("Forbes2000cc.rds")
```

- Alternativen:

```
> ?write.table  # zum Speichern von Datensaetzen als Text-Datei
> ?save         # zum Speichern eines oder mehrerer Objekte
> ?save.image   # zum Speichern des gesamten Arbeitsverzeichnises
>
> setwd("...")
> save.image(file="workspace.RData")
```