Datenergänzung mit multipler Imputation und predictive mean matching – nichts Neues, nur eine Erinnerung

in R mit *mice*

Tagung IBS, Ökologie, DVFFA-Biometrie

Axel Albrecht

Forstliche Versuchs- und Forschungsanstalt Baden-Württemberg (Freiburg)

12.-13. März 2020, Erlangen

```
   age  bmi hyp chl
1    1   NA  NA  NA
2    2 22.7   1 187
3    1   NA   1 187
4    3   NA  NA  NA
5    1 20.4   1 113
6    3   NA  NA 184
7    1 22.5   1 118
8    1 30.1   1 187
9    2 22.0   1 238
10   2   NA  NA  NA
11   1   NA  NA  NA
12   2   NA  NA  NA
13   3 21.7   1 206
14   2 28.7   2 204
15   1 29.6   1  NA
16   1   NA  NA  NA
17   3 27.2   2 284
18   2 26.3   2 199
19   1 35.3   1 218
20   3 25.5   2  NA
21   1   NA  NA  NA
22   1 33.2   1 229
23   1 27.5   1 131
24   3 24.9   1  NA
25   2 27.4   1 186
> 
```

Fehleigenschaften

(types of missingness, missingness = R)

- MCAR: missing completely at random, $P(R|Y_{com}) = P(R)$

  Verteilung der Fehlwerte ist **weder** abhängig von beobachteten Variablen, **noch** von nicht beobachteten aber interessierenden Parametern. Ist fast nie der Fall.

- MAR: missing at random, $P(R|Y_{com}) = P(R|Y_{obs})$

  Verteilung der Fehlwerte in einer Variablen kann abhängig von beobachteten Werten sein, **aber nicht** von den fehlenden Werten.

- MNAR: missing not at random, $P(R|Y_{com}) = P(R|Y_{obs}, Y_{mis})$

  Verteilung der Fehlwerte in einer Variablen kann abhängig von beobachteten Werten sein, **und auch** von den fehlenden Werten

  (z. B. Personen mit hohem Einkommen meiden Angaben)

  Entspricht systematischem Fehlen

Welche Fehleigenschaften sind problematisch?

MCAR: keine Probleme

- complete case analysis möglich, nicht effizient aber verzerrungsfrei, Datenverlust

MAR

- $P(Y_{obs};\theta) = \int P(Y_{com};\theta)dY_{mis}$: ist keine gültige Stichprobenverteilung, nur gültig bei Y_{mis} = MCAR
- Likelihood jedoch gültig

→ Parameterschätzungen sind unverzerrt, aber Korrelationen und andere Inferenzmaße nicht

MNAR: likelihood und Stichprobenverteilung ungültig

- Explizite Verteilung für R nötig, was stark stört
- $P(Y_{obs}, R;\theta,\xi) = \int P(Y_{com};\theta)P(R|Y_{com};\xi)dY_{mis}$

Welche Fehleigenschaften sind problematisch?

MCAR: keine Probleme

- complete case analysis möglich, nicht effizient aber verzerrungsfrei, Datenverlust

MAR

- $P(Y_{obs};\theta) = \int P(Y_{com};\theta)dY_{mis}$: ist keine gültige Stichprobenverteilung, nur gültige Likelihood = MCAR
- Likelihood jedoch gültig
  → Parameterschätzungen sind unverzerrt, aber Korrelationen und andere Inferenzmaße nicht

Datenergänzung sinnvoll *(engl. imputation)*

MNAR: likelihood und Stichprobenverteilung ungültig

- Explizite Verteilung für R nötig, was stark stört
- $P(Y_{obs}, R;\theta,\xi) = \int P(Y_{com};\theta)P(R|Y_{com};\xi)dY_{mis}$

Wie identifiziere ich die Fehleigenschaften?

- Gar nicht
  - Naja – bisschen schon: Argumentativ
  - Aber nicht formal quantifizierbar
- Es gibt geplante Fehlwerte (z. B. „nur messen wenn...“): MAR ist dann korrekte Annahme

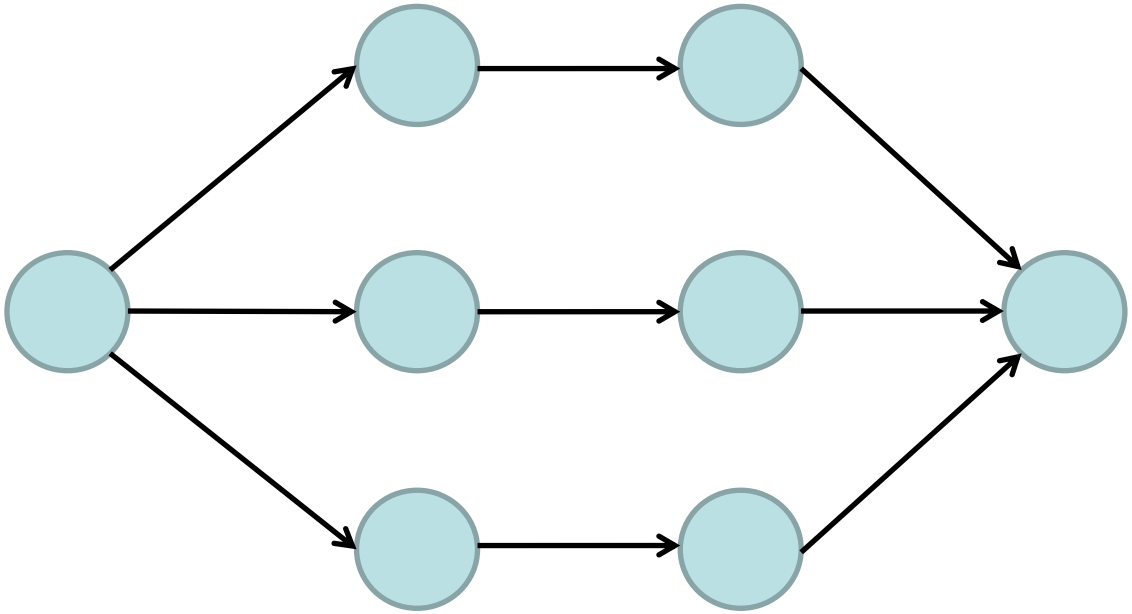
Übersicht Umgang mit Fehlwerten

- **Keine Ergänzung**, nur komplette Datensätze auswerten: Complete case analysis
  - Nur gültig bei MCAR (verzerrungsfrei aber ineffizient)
  - Ab MAR: Bias
  - In einigen nicht MCAR-Fällen möglich: Gewichtung der vollständigen Beobachtungen (Little und Rubin 2002)
  - **Problem: ev. erheblicher Datenverlust**
- **Datenergänzung** (engl.: imputation)
  - **Single** imputation: einmaliges Einsetzen / Ergänzen
  - **Multiple** imputation: wiederholtes, mehrmaliges Einsetzen / Ergänzen

Übersicht Datenergänzungsverfahren

- **Singuläre Imputation**
  - Mittelwert einsetzen: Problem: zu kleine Standardfehler: $\bar{y} \pm 1.96\sqrt{\frac{S^2}{N}}$
  - Einfaches Einsetzen (Zufallsziehung) aus vorhandenen Beobachtungen: (Ergänzung aus unbedingter Verteilung): erhält die Varianz, verzerrt aber Korrelationen.
  - Bedingte Mittelwertergänzung: „halbintelligent“, da fehlende Y aus Regressionsmodell Y~ f(X) für bekannte Y-X-Paare ergänzt werden (Interpolation, Nearest Neighbor, Most Similar Neighbor)
  - Ersetzen durch bedingte Verteilung (Regression, maximum likelihood Schätzer). Das beste der schlechten Verfahren (Unsicherheit wird immer noch unterschätzt).
- **Multiple Imputation**
  - Schätzwert einsetzen durch Regression mit vorhandenen anderen Prädiktoren
  - Häufiges Durchführen (30+?)
  - Weitere Auswertungen müssen den Wiederholvorgang berücksichtigen

Konzept der Multiplen Imputation



Unvollständige Daten | Ergänzte Daten | Analyse-ergebnisse | Gepoolte Ergebnisse

Klasse	**mids**	**mild**	**mira**	**mipo**
Name	imp	idl	fit	est
erstellt durch	mice()	complete()	with()	pool()
Beschreibung	multipel ergänzter Datensatz	multipel ergänzte Liste von Daten	wiederholte Analyse multipel ergänzter Daten	gepoolte Ergebnisse multipel ergänzter Daten
	multiply imputed dataset	multiply imputed list of data	multiple imputation repeated analyses	multiple imputation pooled results

Multiple Imputation – das „Poolen“

Rubin 1987

$$\bar{Q} = m^{-1} \sum_{j=1}^{m} \hat{Q}^j$$

Mittelwert (Q: Koeffizienten)

$$\bar{U} = m^{-1} \sum_{j=1}^{m} U^j$$

Varianz innerhalb einer Imputation

$$B = (m-1)^{-1} \sum_{j=1}^{m} \left[\hat{Q}^j - \bar{Q}\right]^2$$

Varianz zwischen den Imputationen

$$T = \bar{U} + (1 + m^{-1})B$$

Gesamtvarianz (Var-Cov von $\bar{Q}$)

Mice - Ergänzungsmethoden

Method	Description	Scale type	Default
pmm	Predictive mean matching	numeric	Y
norm	Bayesian linear regression	numeric	
norm.nob	Linear regression, non-Bayesian	numeric	
mean	Unconditional mean imputation	numeric	
2l.norm	Two-level linear model	numeric	
logreg	Logistic regression	factor, 2 levels	Y
polyreg	Polytomous (unordered) regression	factor, >2 levels	Y
lda	Linear discriminant analysis	factor	
sample	Random sample from the observed data	any	

Predictive mean matching: *6 Schritte*

1. Lineares Modell *[auch GLM]* zur Schätzung der Fehlwerte, anhand vollständiger Beobachtungen (cc)
2. Zufallsziehung aus der posterioren Vorhersageverteilung *(posterior predictive distribution)* von $\hat{\beta}$ für neue Koeffizienten: β^* → dies wird bei allen Ergänzungsalgorithmen zur Schaffung von Streuung in den ergänzten Werten getan.
3. Berechnung fehlender y_i mit β^*, vorhandener y_i mit $\hat{\beta}$
4. Finde für jeden Fall mit fehlendem Y die nächsten z. B. drei *vorhergesagten* Werte unter den Fällen mit beobachtetem Y
5. Ziehe zufällig einen dieser drei „nahen“ Fälle und setze für den fehlenden Wert y_i den *beobachteten* Wert dieses nahen Falls ein
6. Wiederhole >= 5 mal *(multiple imputation)*

Quelle der Variabilität wiederholter Datenergänzung

„data driven methods“ Wahrung des Datenrahmens

Beispielanwendung

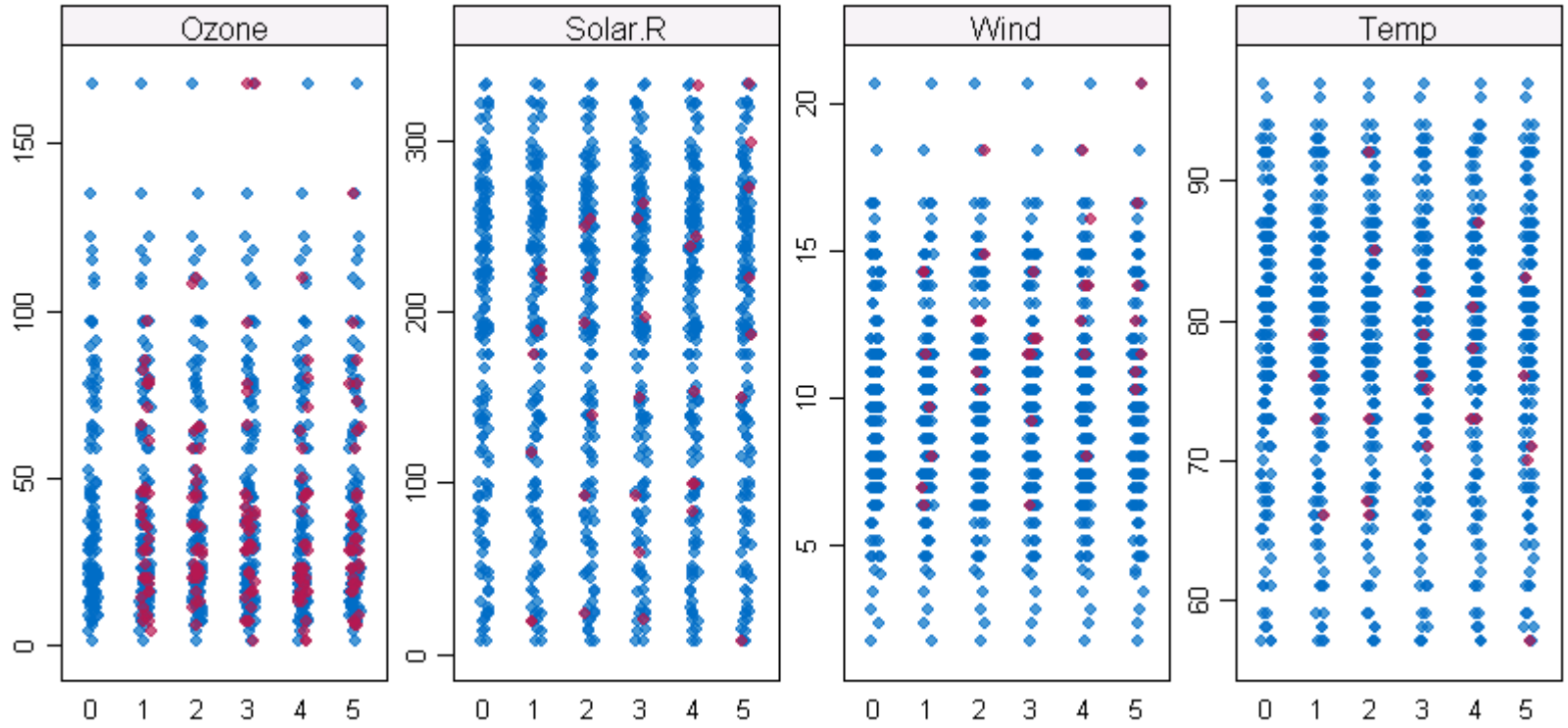
Daten

```
> library(mice)
> library(lattice)
> library(data.table)
> library(ggplot2)
> library(corrgram)
> #### 1) Eingangsdaten, sukzessive "amputieren" ####
> data <- airquality
> data[4:10, 3] <- rep(NA, 7) # beim Wind 7 Fehlwerte hinzufügen
> data[1:5, 4] <- NA # bei der Temperatur zusätzlich noch vier
> data <- data[-c(5, 6)] # Monat und Tag brauchen wir nicht
> summary(data)
     Ozone           Solar.R          Wind             Temp      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :57.00  
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:73.00  
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
 Mean   : 42.13   Mean   :185.9   Mean   : 9.806   Mean   :78.28  
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
 NA's   :37       NA's   :7       NA's   :7        NA's   :5      
> 
```

Beispielanwendung

Erste Ergänzung (mice()) mit ‚pmm‘

```
> #### 2) Multiple Ergänzung mit predictive mean matching (pmm) ####
> tempData <- mice(data, m = 5, maxit = 50, meth = 'pmm', seed = 1234, print = F)
>
> #### 3) Ergänzung visualisieren - teils zeitaufwändig ####
> stripplot(tempData, pch = 20, cex = 1.2) # bei unseren Daten der sprechendste Plot
>
```



Beispielanwendung

Wiederholtes Analyse-LM, Poolen

```
> #### 4) wiederholt eine Regression anfertigen ####
> modelFit1 <- with(tempData, lm(Temp ~ Ozone + Solar.R + Wind))
> summary(modelFit1) # fünf sets an Koeffizienten werden errechnet
# A tibble: 20 x 5
   term        estimate std.error statistic  p.value
   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)  71.7      2.57        27.8   6.86e-61
 2 Ozone         0.173    0.0229       7.55  4.13e-12
 3 Solar.R       0.0105   0.00661      1.59  1.15e- 1
 4 Wind         -0.259    0.199       -1.30  1.95e- 1
 5 (Intercept)  74.2      2.76        26.9   4.33e-59
 6 Ozone         0.147    0.0249       5.93  2.05e- 8
 7 Solar.R       0.0146   0.00700      2.08  3.90e- 2
 8 Wind         -0.471    0.211       -2.23  2.71e- 2
 9 (Intercept)  73.5      2.45        30.1   3.72e-65
10 Ozone         0.148    0.0201       7.38  1.01e-11
11 Solar.R       0.0138   0.00648      2.13  3.46e- 2
12 Wind         -0.409    0.190       -2.15  3.29e- 2
13 (Intercept)  74.1      2.46        30.2   2.60e-65
14 Ozone         0.164    0.0219       7.49  5.62e-12
15 Solar.R       0.0112   0.00638      1.75  8.19e- 2
16 Wind         -0.446    0.186       -2.39  1.79e- 2
17 (Intercept)  73.0      2.41        30.3   1.47e-65
18 Ozone         0.170    0.0207       8.23  8.72e-14
19 Solar.R       0.0120   0.00624      1.91  5.75e- 2
20 Wind         -0.420    0.179       -2.35  2.03e- 2
>
> # und dann die je 5 Koeffizienten "poolen"
> summary(pool(modelFit1))
               estimate   std.error statistic        df      p.value
(Intercept) 73.30727107 2.781694608 26.353458  64.38216 0.000000e+00
Ozone        0.16070257 0.025842335  6.218578  37.52957 3.001339e-07
Solar.R      0.01240081 0.006819786  1.818357 112.40653 7.167231e-02
Wind        -0.40084488 0.213696580 -1.875766  60.38760 6.552344e-02
> |
```

Beispielanwendung

Weitere Ergänzungen

```
> #### 5) nun zum Vergleich single imputation als mean imputation ####
> data_si <- mice(data, method = "mean", maxit = 1, m = 1)

 iter imp variable
  1   1  Ozone  Solar.R  Wind  Temp
> modelFit2 <- with(data_si, lm(Temp ~ Ozone + Solar.R + Wind))
>
> # und single imputation mit pmm #
> data_si_pmm <- mice(data, method = "pmm", maxit = 1, m = 1)

 iter imp variable
  1   1  Ozone  Solar.R  Wind  Temp
> modelFit3 <- with(data_si_pmm, lm(Temp ~ Ozone + Solar.R + Wind))
>
> #### 6) weiter noch der complete case Ansatz ####
> lm_cc <- lm(Temp ~ Ozone + Solar.R + Wind, data = data)
>
```

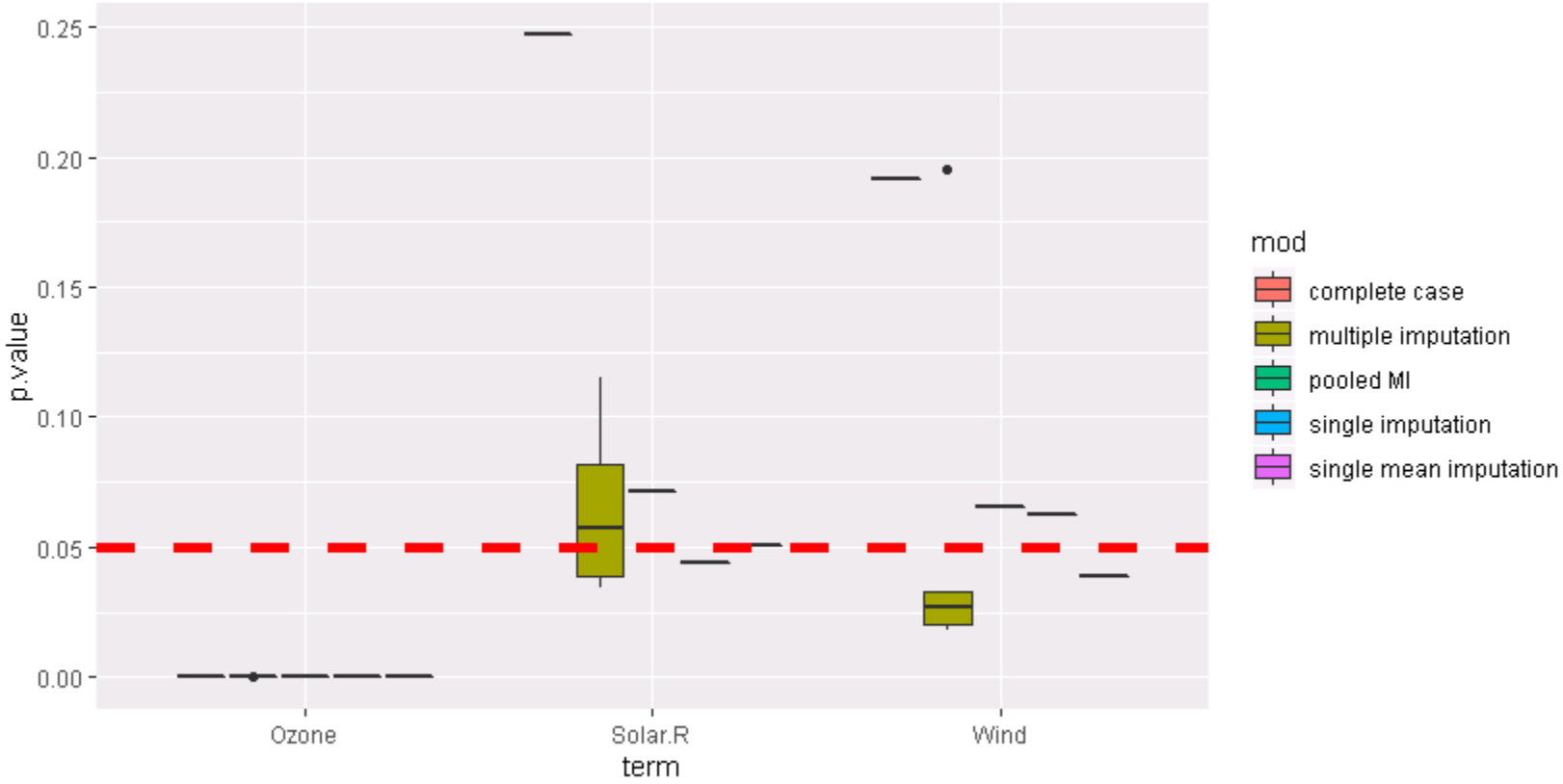
Beispiel-anwendung

Zusammenführung und Grafiken

```
> #### 7) und schließlich graphischer Vergleich ####
> ressi <- as.data.table(summary(modelFit2))
> ressi[, mod := "single mean imputation"]
>
> ressipmm <- as.data.table(summary(modelFit3))
> ressipmm[, mod := "single imputation"]
>
> resmi <- as.data.table(summary(modelFit1))
> resmi[, mod := "multiple imputation"]
>
> rescc <- as.data.table(coef(summary(lm_cc)))
> rescc[, term := rownames(coef(summary(lm_cc)))]
>
> setnames(rescc, old = c("Estimate", "Std. Error", "t value", "Pr(>|t|)"), new = c("estimate", "std
.error", "statistic", "p.value"))
> rescc[, mod := "complete case"]
>
> resmi_pool <- as.data.table(summary(pool(modelFit1)))
> resmi_pool[, term := rownames(summary(pool(modelFit1)))]
> resmi_pool[, mod := "pooled MI"]
>
> resmi_pool[, df := NULL]
> names(resmi_pool)
[1] "estimate"  "std.error" "statistic" "p.value"   "term"      "mod"
>
> res <- rbind(ressi, ressipmm, resmi, rescc, resmi_pool)
> res
          term     estimate   std.error statistic      p.value                    mod
 1: (Intercept) 73.249283874 2.748946466 26.646312 1.493801e-58 single mean imputation
 2:       Ozone  0.157544325 0.025075509  6.282797 3.467824e-09 single mean imputation
 3:     Solar.R  0.013766196 0.006986292  1.970458 5.063923e-02 single mean imputation
 4:        Wind -0.425146881 0.204079126 -2.083245 3.893707e-02 single mean imputation
 5: (Intercept) 73.376447783 2.619597037 28.010586 3.002409e-61      single imputation
 6:       Ozone  0.151458623 0.021873069  6.924434 1.219284e-10      single imputation
 7:     Solar.R  0.013338770 0.006567367  2.031068 4.402604e-02      single imputation
 8:        Wind -0.379025925 0.202214599 -1.874375 6.283649e-02      single imputation
 9: (Intercept) 71.651271685 2.574951563 27.826260 6.864247e-61    multiple imputation
10:       Ozone  0.173163580 0.022948439  7.545767 4.125420e-12    multiple imputation
11:     Solar.R  0.010480524 0.006609764  1.585612 1.149473e-01    multiple imputation
12:        Wind -0.258509009 0.198750418 -1.300672 1.953789e-01    multiple imputation
13: (Intercept) 74.217610292 2.757499104 26.914827 4.328603e-59    multiple imputation
14:       Ozone  0.147403889 0.024868255  5.927392 2.054808e-08    multiple imputation
15:     Solar.R  0.014577303 0.007001448  2.082041 3.904855e-02    multiple imputation
16:        Wind -0.471027324 0.211046423 -2.231866 2.711611e-02    multiple imputation
17: (Intercept) 73.540310135 2.445549836 30.071074 3.719206e-65    multiple imputation
18:       Ozone  0.148498701 0.020112765  7.383306 1.012403e-11    multiple imputation
19:     Solar.R  0.013809390 0.006476312  2.132292 3.462215e-02    multiple imputation
20:        Wind -0.408916084 0.189843251 -2.153967 3.285079e-02    multiple imputation
21: (Intercept) 74.137492855 2.458526997 30.155249 2.599720e-65    multiple imputation
22:       Ozone  0.163960517 0.021890251  7.490116 5.616006e-12    multiple imputation
23:     Solar.R  0.011184479 0.006384594  1.751792 8.186705e-02    multiple imputation
24:        Wind -0.446066203 0.186391735 -2.393165 1.794877e-02    multiple imputation
25: (Intercept) 72.989670374 2.409695619 30.289996 1.467619e-65    multiple imputation
26:       Ozone  0.170486175 0.020718716  8.228607 8.719943e-14    multiple imputation
27:     Solar.R  0.011952350 0.006244884  1.913943 5.754469e-02    multiple imputation
```

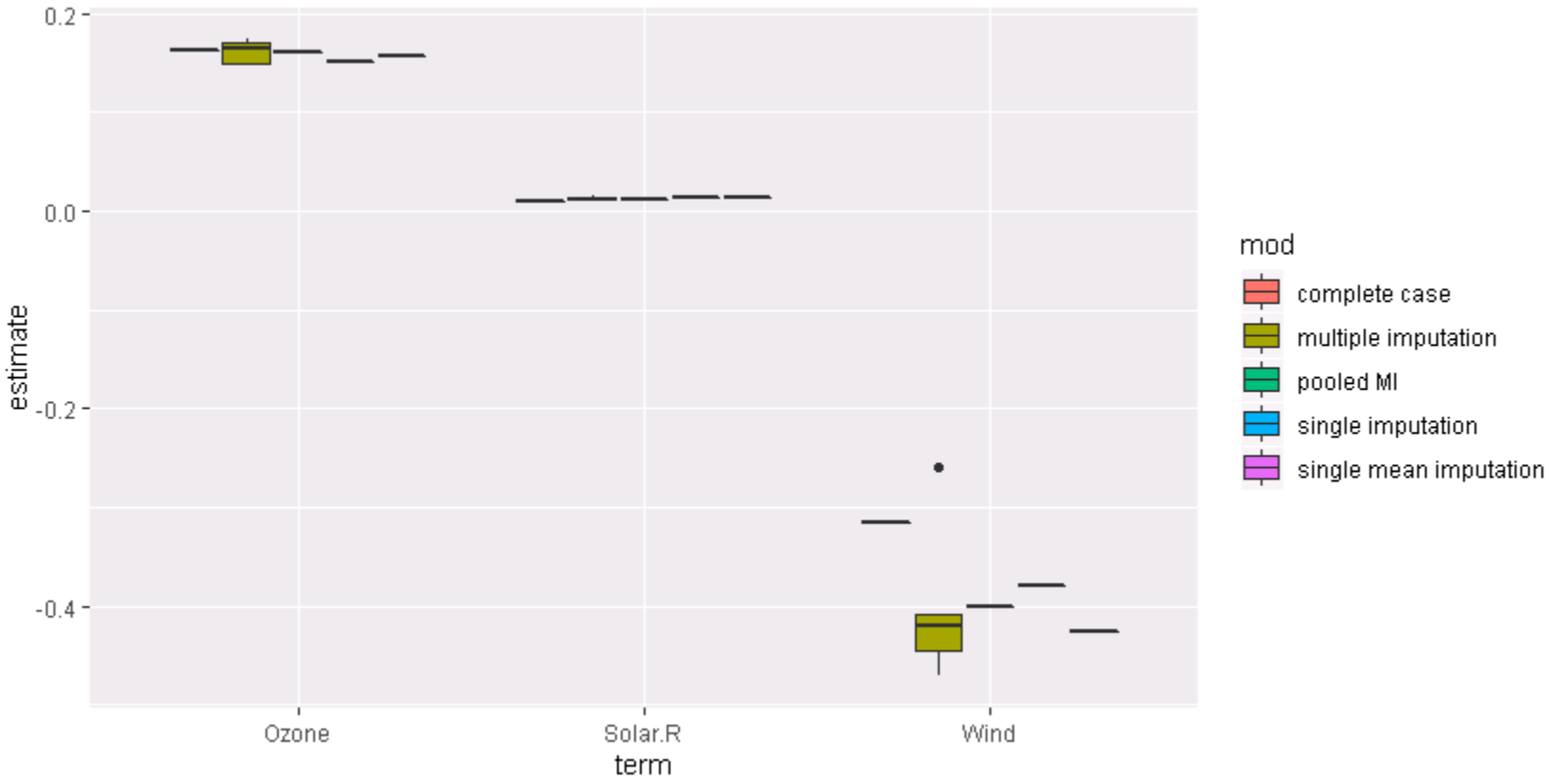
Beispielanwendung

Ergebnisgrafiken: P-Wert



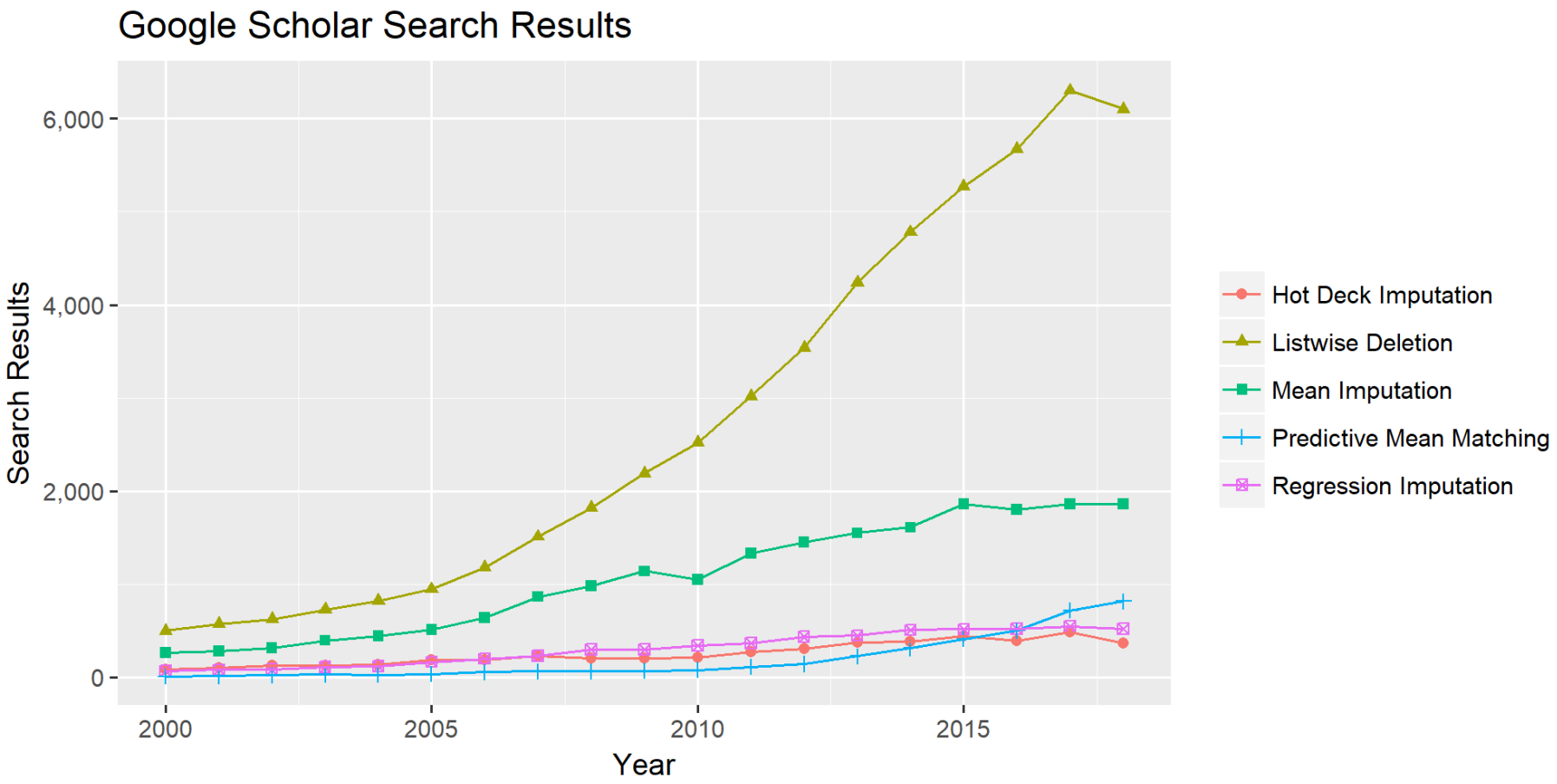
Beispielanwendung

Ergebnisgrafiken: Koeffizienten



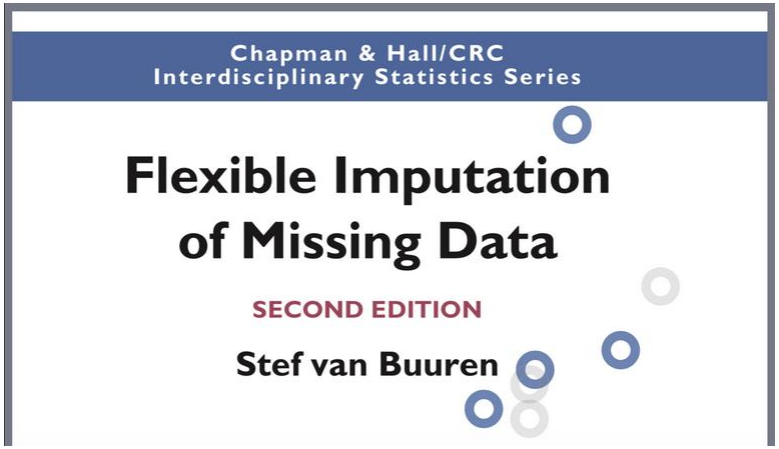
Zusammenfassung

- Fehleigenschaften (MCAR, MAR, MNAR)
- Complete case, Singuläre Imputation (<5% univariat und MCAR), sonst: Multiple Imputation
- Einsatzbereiche MI
  - Voraussetzung: Missing at random
  - Antwort- und erklärende Variablen ergänzbar
  - Nominal-, intervallskalierte und ordinale Variablen
- Vorgehen:
  1. mehrmals ergänzen (Standard: 5-10 mal, oder siehe Graham et al. 2007)
  2. Geplante Analysen / Auswertungen mit mehreren Datensätzen
  3. daraus gepoolte Verteilungsparameter / Koeffizienten schätzen



https://statisticsglobe.com/imputation-methods-for-handling-missing-data/

Literatur



Online-Ressourcen:

**** https://stefvanbuuren.name/fimd/ ****

Missing Data / Fehlwerte:

- Schafer JL and Graham JW (2002) Missing data: our view of the state of the art. Psychol Methods 7:147-177
- Little RJA and Rubin D (2002) Statistical Analysis with Missing Data. 408 pp.

Multiple Imputation / Datenergänzung:

- Rubin D (1987) Multiple Imputation for Nonresponse in Surveys. Wiley, New York
- Schafer JL (1999) Multiple imputation: a primer. Stat Methods Med Res 8:3-15
- Rubin DB (1996) Multiple Imputation after 18+ Years. Journal of the American Statistical Association 91:473-489, doi: 10.1080/01621459.1996.10476908
- Graham JW, Olchowski AE and Gilreath TD (2007) How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci 8:206-213, doi: 10.1007/s11121-007-0070-9
- Graham JW (2009) Missing data analysis: making it work in the real world. Annu Rev Psychol 60:549-576, doi: 10.1146/annurev.psych.58.110405.085530
- www.multiple-imputation.com

MI (etc.) in R:

- Su Y-S, Gelman A, Hill J and Yajima M (2011) Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. Journal of Statistical Software 45:1-27
- Buuren Sv and Groothuis-Oudshoorn K (2011) mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 45:1-67