Daniel Wollschläger

Grundlagen der Datenanalyse mit R, 2. Auflage – Errata

Stand: 25. September 2014

Inhaltlich relevante Korrekturen

- Abschn. 2.7.1, S. 64, Fußnote 19:
  "Wegen $0! = 1$ erzeugt `prod(numeric(0))` ebenso wie `factorial(`~~`numeric(`~~`0`~~`)`~~`)` das Ergebnis 1."
- Abschn. 2.12.2, S. 103:
  "Die Entsprechung zwischen variablen Feldern und Objekten wird über deren Reihenfolge hergestellt,"
- Abschn. 2.12.4, S. 107:
  "Das Ergebnis eignet sich besonders, um mit der `substring()` Funktion weiterverarbeitet zu werden,"
- Abschn. 2.12.5, S. 108:
  "indem dem Ergebnis von `substring()` ein passender Vektor von Zeichenketten zugewiesen wird"
- Abschn. 6.8.3, S. 206:
  "`> abline(h=0.5, v=-b0/`~~`n1`~~`b1, col="gray")`"
- Abschn. 7.1.2, S. 211:
  "`> absDiff <- abs(DV - ave(DV, IV, `~~`Fun`~~`FUN=median))  # abs. Abweichungen`
- Abschn. 7.4.1.2, S. 232:
  "`> SSEtot <- anRes["id", "Sum Sq"] + anRes["`~~`id:IV`~~`IV:id", "Sum Sq"]`
- Abschn. 8.5.2, S. 315:
  "`(pLess <- `~~`1-`~~` pbinom(N-obs `~~`-1`~~`, N, 0.5))    # linksseitig`"
  "`[1] `~~`0.994091s`~~` 0.02069473`"

  "`> (pTwoSided <- 2* `~~`(1-pbinom(max(obs,N-obs)-1,N,0.5))`~~` min(pLess, pGreater))`"
- Abschn. 8.5.7, S. 320:
  "Der Rangsummentest nach Friedman dient der Analyse von Daten einer ~~kategorialen~~ stetigen ordinalen Variable, die in p abhängigen Stichproben erhoben wurde."
- Abschn. 9.1.1, S. 332:
  "dessen Breite durch die $\alpha/2$- und $1-\alpha/2$-Quantile der Werte von $t^{\star} = (\hat{\theta}^{\star} - \hat{\theta})/S^{\star}_{\hat{\theta}}$ multipliziert mit der Streuung von $\hat{\theta}^{\star}$ definiert ist."
- Abschn. 9.1.3, S. 337:
  "`> (pValBS <-(sum(Fstar >= Fbase) + 1) / (length(Fstar) + 1))`"
  Der Vergleich der Bootstrap $F^{\star}$-Werte mit dem F-Wert der Basisstichprobe über `>=` ist nicht tolerant gegenüber Problemen der numerischen Repräsentation von Gleitkommazahlen. Es ist daher besser, neben `>` nur auf ungefähre Gleichheit zu testen:

  ```
  tol     <- .Machine$double.eps^0.5
  FsIsGEQ <- (Fstar > Fbase) | (abs(Fstar-Fbase) < tol)
  (pValBS <- (sum(FsIsGEQ) + 1) / (length(Fstar) + 1))
  ```

- Abschn. 9.2.1, S. 340:
  "`> (pVal <- sum(resDM <= diffM) / length(resDM))`"
  Vgl. Hinweis zu Abschn. 9.1.3, S. 337:
  ```
  tol      <- .Machine$double.eps^0.5
  DMsIsLEQ <- (resDM < diffM) | (abs(resDM-diffM) < tol)
  (pVal    <- sum(DMsIsLEQ) / length(resDM))
  ```

- Abschn. 9.2.2, S. 342:
  "`> (pVal <- sum(resMD <= mean(DVd)) / length(resMD))`"
  Vgl. Hinweis zu Abschn. 9.1.3, S. 337:
  ```
  tol      <- .Machine$double.eps^0.5
  MDsIsLEQ <- (resMD < mean(DVd)) | (abs(resMD-mean(DVd)) < tol)
  (pVal    <- sum(MDsIsLEQ) / length(resMD))
  ```

- Abschn. 10.1.1, S. 344, Fußnote 3:
  "Eine (1×1)-Diagonalmatrix $\boldsymbol{X}$ mit dem einzigen Element $x \neq 1$ kann mit dem zweiten Argument `nrow` als `diag(x, nrow=1)` ~~müsste daher mit `matrix(x)`~~ erzeugt werden."

- Abschn. 10.2, S. 361:
  "Die Hauptkomponenten selbst werden durch `prcomp()` ~~nicht berechnet~~ in der Komponente `x` der erzeugten Liste ausgegeben, und durch `princomp()` ~~liefert sie aber~~ in der Komponente `scores` ~~der erzeugten Liste~~."

- Abschn. 10.2, S. 362:
  "Für die Datenmatrix $\boldsymbol{X}$ gilt dann $\boldsymbol{X} = \cancel{\boldsymbol{B}\boldsymbol{H}_s}\ \boldsymbol{H}_s\boldsymbol{B}^t + \boldsymbol{c}$."

- Abschn. 10.6.2, S. 377:
  "Die Modellformel ist multivariat wie mit `lm(⟨Modellformel⟩)` zu formulieren"
  "Außerdem ist T^2 gleich dem $(n_1 + n_2 - \cancel{1}2)$-fachen der Hotelling-Lawley-Spur."

- Abschn. 10.8, S. 383:
  "`> Ydf1 <- data.frame(IVman, DV1=Ym1[ , 1], DV2=Ym1[ , 2])`"

Weitere Hinweise

- Abschn. 6.1, S. 171:
  Mit der Funktion `r.test()` aus dem Paket `psych` lassen sich auch Hypothesen darüber testen, ob zwei theoretische Korrelationskoeffizienten aus unabhängigen oder abhängigen Stichproben identisch sind.

- Abschn. 6.3.1, S. 180:
  Eine empfehlenswerte Alternative zu `scatterplot3d()` ist die Funktion `scatter3d()` aus dem Paket `car`, die als Argument dieselbe Modellformel wie `lm()` akzeptiert. Die hier mit `scatter3d(weight ~ height + age)` zu erstellende Grafik enthält bereits die Vorhersageebene und Residuen. Zudem lässt sich die dargestellte Perspektive durch Klicken und Ziehen mit der Maus interaktiv ändern, was den 3D-Eindruck fördert (vgl. Abschn. 11.8.2).

- Abschn. 6.3.4, S. 185:
  Die Funktion `sim.slopes()` aus dem Paket `QuantPsyc` eignet sich nur für den dargestellten Fall einer moderierten Regression mit einem Prädiktor und einem Moderator. In Modellen mit weiteren Prädiktoren ist das ausgegebene Ergebnis nicht korrekt.

- Abschn. 6.6.1, S. 191:
  Da Extremwerte die Streuungen mit beeinflussen, sollten für die Diagnose von Extremwerten evtl. robuste Schätzer für Varianzen und Kovarianzmatrizen in Betracht gezogen werden (vgl. `covMcd()` aus dem Paket `robustbase` und Abschn. 2.7.4). So geschätzte Kovarianzmatrizen können etwa an das Argument `cov` von `mahalanobis()` übergeben werden.

- Abschn. 8.5.9, S. 323:
  "Der Basisumfang von von R stellt für den [Bowker-] Test keine eigene Funktion bereit, weshalb eine manuelle Rechnung notwendig ist."
  Die in Abschn. 8.5.10 vorgestellte Funktion `mcnemar.test()` berechnet automatisch den Bowker-Test, wenn Kontingenztafeln mit mehr als zwei Zeilen und Spalten als Argument übergeben werden (Dank an Andri Signorell für den Hinweis).

- Abschn. 9.1.3, S. 336:
  Eine Variante des model-based resampling ist der sog. *wild bootstrap* für Situationen, in denen Heteroskedastizität vorliegt. Hier wird $E^{\star}$ nicht aus den Residuen E gebildet, sondern aus dem Produkt $E/\sqrt{1-h} \cdot U$. Dabei ist h die Variable Hebelwert (vgl. Abschn. 6.6.1) und U eine unabhängige Zufallsvariable mit $E(U) = 0$ und $E(U^2) = 1$.

  - Eine Wahl für U sind dichotome Variablen mit sog. F_1-Verteilung, die den Wert $-(\sqrt{5}-1)/2$ mit Wahrscheinlichkeit $p = (\sqrt{5}+1)/(2\sqrt{5})$ annehmen und den Wert $(\sqrt{5}+1)/2$ mit Wahrscheinlichkeit $1 - p = (\sqrt{5}-1)/(2\sqrt{5})$. Bei n Residuen:
    ```
    sample(c(-(sqrt(5) - 1) / 2, (sqrt(5) + 1) / 2), size=n,
           prob=c((sqrt(5) + 1) / (2*sqrt(5)), (sqrt(5) - 1) / (2*sqrt(5))))
    ```

  - Eine alternative Wahl für U sind ebenfalls dichotome Variablen mit sog. F_2- bzw. Rademacher-Verteilung: Sie nehmen die Werte -1 und 1 jeweils mit Wahrscheinlichkeit $1/2$ an, drehen also das ursprüngliche Vorzeichen jedes Residuums zufällig um. Bei n Residuen:
    ```
    sample(c(-1, 1), size=n, prob=c(0.5, 0.5))
    ```

- Abschn. 9.2.1, S. 340:
  Anstatt die Parameter der Normalverteilung aus der Permutationsverteilung zu schätzen, hätte auch die unter H_0 gültige Verteilung der Mittelwertsdifferenz $\mathcal{N}(0, \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2})$ eingezeichnet werden können.

- Abschn. 10.2, S. 359:
  Zusätzlich erlaubt es `princomp()`, die Kovarianzmatrix der Daten über das Argument `covmat` separat zu spezifizieren. Dies könnte etwa für robuste Schätzungen der theoretischen Kovarianzmatrix genutzt werden (vgl. `covMcd()` aus dem Paket `robustbase`), die auch Grundlage der robusten Hauptkomponentenanalyse im Paket `pcaPP` sind.

Tipp- und Druckfehler

- Abschn. 2.7.3, S. 66:
  "Für die Berechnung des Modalwertes, also des am häufigsten vorkommenden Wertes eines Vektors,"

- Abschn. 3.3.3.2, S. 132:
  "> subset(myDf1, (~~myDf1$~~sex == "m") & (~~myDf1$~~rating > 2))"
  "> subset(myDf1, (~~myDf1$~~IQ < 90) | (~~myDf1$~~IQ > 110))"

- Abschn. 4.2.5.1, S. 159:
  "Für einen detaillierten Vergleich der Arbeit mit R, SAS und SPSS vgl. Muenchen (2011), der auch den Datenaustausch zwischen den Programmen behandel~~n~~t."

- Abschn. 7.4.1, S. 230:
  "`> DV_t4 <- round(rnorm(N, 0.4, 1), 2)      # AV zu t`~~3~~`4`"

- Abschn. 7.5.1, S. 237:
  "`> aov(⟨AV⟩ ~ ⟨UV1⟩ + ⟨UV2⟩ + ⟨UV1⟩:⟨UV2⟩`~~)~~`, data=⟨Datensatz⟩)`"
  "`> aov(⟨AV⟩ ~ ⟨UV1⟩*⟨UV2⟩`~~)~~`, data=⟨Datensatz⟩)`"

- Abschn. 8.1, S. 282:
  "auch wenn sich der β-Fehler so nicht exakt begrenz~~t~~en lässt."

- Abschn. 8.1.5, S. 290:
  "dass der von `chisq.test()` ausgegebene p-Wert nicht die Reduktion der Freiheitsgrade widerspiegelt und deshalb für diesen Test nicht der richtige ist."

- Abschn. 8.2.6.1, S. 298:
  "Das F-Maß als harmonisches Mittel von Präzision und recall wird bisweilen als integriertes Gütemaß für eine Klassifikation herangezogen."

- Abschn. 9.1, S. 330:
  "hat das erste Element des Vektors $\hat{\theta}^{\star}$ und das zweite Element der plug-in-Schätzer $\hat{\sigma}^{2\star}_{\hat{\theta}}$ der theoretischen Varianz $\sigma^2_{\hat{\theta}}$ zu sein."

- Abschn. 10.1.7, S. 355:
  "weshalb der Koordinatenvektor $\boldsymbol{y}$ des orthogonal auf V projizierten Vektors $\boldsymbol{x}$ bzgl. der Basis $\boldsymbol{A}$ durch $(\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}^t\boldsymbol{x}$ gegeben ~~sind~~ ist."

- Abschn. 10.6.2, S. 377:
  "Lineare Strukturgleichungsmodelle werden durch die Pakete `sem` (Fox & Byrnes, 2011) ..."

- Abschn. 10.6.2, S. 377:
  "~~`facT`~~`IVht     1    0.6073    11.235    2    37    0.0001539 ***`"

- Abschn. 10.8, S. 382:
  "Für weitere Klassifikationsverfahren wie Varianten der ~~der~~ Clusteranalyse, CART-Modelle"

- Abschn. 11.2.1, S. 424:
  "Alternativ lässt sich mit der `hexbin()` Funktion aus dem gleichnamigen Paket ein Diagramm erstellen, das~~s~~ die Diagrammfläche in hexagonale Regionen einteilt"