Dr. G. Eichner
Math. Inst.
JLU Gießen

Übungen zu
R1: Grundlagen der Datenanalyse mit R
Blatt 4

SoSe 2024
9. 5. 2024
Abgabe: ≤22. 5. 2024, 14:00 Uhr

1. Der in base-**R** in einem Data Frame namens `airquality` „eingebaute" Datensatz enthält (länger zurückliegende) tägliche Messungen zur Luftqualität in New York. Machen Sie sich auf der zugehörigen **R**-Hilfeseite etwas mit der Bedeutung der Variablen in seinen sechs Spalten vertraut!

   Gelegentlich ist es z. B. aus rechentechnischen oder statistischen Gründen sinnvoll bis nötig, dass Stichprobendaten, die von numerischen Variablen stammen, welche in stark unterschiedlichen Wertebereichen „leben", vergleichbar zu machen bzw. auf ähnliche Größenordnungen zu bringen, indem sie *zentriert* und *standardisiert* werden. (Beides zusammen heißt auch *normiert* oder *normalisiert* zu werden.) Dabei bedeutet Werte zu zentrieren i. d. R., dass von ihnen ihr arithmetisches Mittel abgezogen wird, und Werte zu standardisieren i. d. R., dass sie nach Zentrierung (!) durch ihre Standardabweichung dividiert werden.

   Normieren Sie die Werte der vier Variablen in `airquality`, die Messwerte der Luftqualität enthalten!

2. Für eine Stichprobe vom Umfang n seien Daten zu $p \geq 2$ verschiedenen, reellwertigen Variablen in jeweils n-dimensionalen Spaltenvektoren, also in $\mathbf{x}_i \in \mathbb{R}^n,\ i = 1, \ldots, p$, gespeichert. Diese Spaltenvektoren seien in der Datenmatrix $\mathbf{M} = (\mathbf{x}_1|\cdots|\mathbf{x}_p)$ zusammengefasst abgelegt.

   a) Implementieren Sie zu $\mathbf{M}$ die empirische Kovarianzmatrix $\widehat{\mathrm{Cov}}(\mathbf{M}) \equiv (\hat{\sigma}(\mathbf{x}_i, \mathbf{x}_j))_{1\leq i,j\leq p}$ und die empirische Korrelationsmatrix $\widehat{\mathrm{Cor}}(\mathbf{M}) \equiv (\hat{\rho}(\mathbf{x}_i, \mathbf{x}_j))_{1\leq i,j\leq p}$ unter *ausschließlicher* Verwendung des Matrix-Vektor-Kalküls (und insbesondere ohne Schleifen o. Ä.)!
   Dabei bezeichne $\hat{\sigma}(\mathbf{x}_i, \mathbf{x}_j)$ für $1 \leq i, j \leq p$ die empirische Kovarianz der Elemente der Vektoren $\mathbf{x}_i$ und $\mathbf{x}_j$ und $\hat{\rho}(\mathbf{x}_i, \mathbf{x}_j) \equiv \hat{\sigma}(\mathbf{x}_i, \mathbf{x}_j)/(\hat{\sigma}(\mathbf{x}_i) \cdot \hat{\sigma}(\mathbf{x}_j))$ deren empirische Pearsonsche Korrelation, wobei wie üblich $\hat{\sigma}^2(\mathbf{x}_i) \equiv \hat{\sigma}(\mathbf{x}_i, \mathbf{x}_i)$.
   Beachten Sie für die Formel der empirischen Kovarianz den Hinweis „Zur Erinnerung" in §2.8.11 im Skript!

   Tipp zum Matrix-Vektor-Kalkül: $\hat{\sigma}(\mathbf{x}_i, \mathbf{x}_j) \stackrel{!}{=} \dfrac{1}{n-1}\mathbf{x}_i'(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n')\mathbf{x}_j$, wobei $\mathbf{I}_n$ die $(n \times n)$-Einheitsmatrix und $\mathbf{1}_n$ den n-dimensionalen Einsenvektor bezeichnen. (Das „!" über dem „=" sollte als Aufforderung gedeutet werden, den Vektor-Matrix-Kalkül, sprich das Rechnen mit Vektoren und Matrizen aus der Linearen Algebra zu rekapitulieren und die Aussage dieses Tipps zu beweisen.)

   b) Vergleichen Sie Ihre Lösung(en) für die $p = 2$ Beispielvektoren `faithful$waiting` und `faithful$eruptions` aus der Old-Faithful-Aufgabe von Blatt 2 zur Kontrolle mit den Ergebnissen, die die **R**-Funktionen `cov` und `cor` liefern.

   c) Ergänzung: Welche Korrelationswerte erhalten Sie, wenn Sie die Berechnungen *separat* für kurze und nicht kurze Eruptionsdauern (und ihre jeweiligen Wartezeiten) ausführen und wie deuten Sie diese Ergebnisse im Vergleich zu der erhaltenen Korrelation für die Gesamtdaten (unter Berücksichtigung der Tatsache, dass der Pearsonsche Korrelationskoeffizient die Stärke des *linearen* Zusammenhangs zwischen zwei metrisch skalierten Variablen quantifiziert)?

3. Machen Sie sich (wie im Skript am Anfang von Abschnitt 2.10 gezeigt) mit Hilfe der Funktion `data` den Data Frame `cu.summary` des **R**-Paketes `rpart` verfügbar und kontrollieren Sie (mit `objects()` oder `ls()` oder im RStudio-Environment) die „Existenz“ von `cu.summary`. (Falls das Paket `rpart` in Ihrer **R**-Installation noch nicht vorhanden ist, installieren Sie es wie im Skript in Abschnitt 1.7 beschrieben vorher mit `install.packages("rpart")`.)

   a) Extrahieren Sie mithilfe der Funktion `subset` aus `cu.summary` den (Teil-)Data Frame, der nur die Daten zu deutschen Fahrzeugen enthält.

   b) Im Resultat von eben ist die Spalte `Country` nun offenbar überflüssig. Eliminieren Sie sie bereits bei der Extraktion des (Teil-)Data Frames im Aufruf von `subset`.

   c) In den verbliebenen `Factor`-Spalten treten einige ihrer `Factor`-Levels nicht mehr auf. Eliminieren Sie diese überflüssigen Levels.

4. (Technische Forts. von Aufgabe 3.)

   a) Extrahieren Sie unter Verwendung der Funktion `is.na`, die in §2.11.2 vorgestellt wurde, aus `cu.summary` einen Data Frame `cu2`, dem die Zeilen von `cu.summary` fehlen, in denen sich `NA`-Einträge befinden.

      Vergleichen Sie Ihr Vorgehen mit der Anwendung von `na.omit` auf `cu.summary`.

   b) Führen Sie dies entsprechend für die Spalten von `cu.summary` durch und speichern Sie das Ergebnis in `cu3`. Tun Sie es letztlich auch für Zeilen *und* Spalten von `cu.summary` und legen Sie das Resultat in `cu4` ab.

      Ist für `cu4` die Reihenfolge der Elimination von Zeilen und Spalten relevant? Ist die Elimination von Spalten sinnvoll?