Dr. G. Eichner
Math. Inst.
JLU Gießen

Übungen zu
R1: Grundlagen der Datenanalyse mit R
Blatt 7

SoSe 2024
30. 5. 2024
Abgabe: ≤12. 6. 2024, 14:00 Uhr

1. (Weitere Fortsetzung von Aufgabe 1 auf Blatt 6.) Weitere, unterschiedliche Mittel der EDA für die SMSA-Daten:

   a) Erstellen Sie für die Realisierungen der Variablen `Region` ein Kreisdiagramm und einen "Dot Chart", deren Beschriftungen hinreichend aufschlussreich sind!

   b) Lassen Sie für alle metrisch skalierten Variablen in `SMSA.df` Boxplots (ohne Berücksichtigung der Regionen) zeichnen, und zwar zunächst nur mithilfe von `boxplot` so, dass „parallele" Boxplots in ein und demselben Koordinatensystem entstehen! Ist das eine gute Darstellung?

   Tun Sie dann das Entsprechende unter Zuhilfenahme von `lapply`, sodass „separate" Darstellungen in eigenen Koordinatensystemen entstehen! Ist letztere eine bessere Darstellung? (Wichtig: Rufen Sie vor jeder Grafik, die Sie unter Verwendung von `lapply` erstellen, den Befehl `par(mfrow = c(2, 5)` auf!)

   c) Fertigen Sie für alle metrisch skalierten Variablen in `SMSA.df` mit `pairs` die paarweisen Streudiagramme an! Führen Sie `pairs` auch in ihrer Formelvariante aus, über die Sie notwendige Informationen auf der entsprechenden Hilfeseite finden!

   d) Wenden Sie auf einige der metrisch skalierten Variablen in `SMSA.df` die eine oder andere von Ihnen ausgewählte, streng monotone Transformation an und lassen Sie die paarweisen Streudiagramme erneut zeichnen! Setzen Sie die von Ihnen gewählten Transformationen direkt in `pairs`' Formelvariante ein, um die Leistungsfähigkeit der Formelvariante zu erkennen!

2. (Und nun eine Fortsetzung von Aufgabe 2 auf Blatt 6.) Fertigen Sie für die Zahlen der praktizierenden Ärzte und Ärztinnen, die Prozentsätze an High-School-AbsolventInnen und die privaten Gesamteinkommen der SMSA-Daten Histogramme und QQ-Plots an. Was halten Sie von der Zulässigkeit der Normalverteilungsannahme für die Variablen?

   Fertigen Sie nun dieselben Grafiken sowohl für die (beliebig) logarithmierten Daten an als auch für die Quadratwurzel der Daten. Wie verhält es sich jeweils mit der Normalverteilungsannahme für die transformierten Variablen?

3. Technische Hintergrundinformationen zur Dichteschätzung:

   Es sei F eine Verteilungsfunktion mit existierender Dichte $f = F'$. Die Zufallsvariablen $X_1, \ldots, X_n$ seien unabhängig und identisch nach F verteilt (kurz: X_i u.i.v. $\sim F$). Ein Kern-Dichteschätzer für – das unbekannte – f ist definiert durch

$$f_n(t) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{b_n} K\left(\frac{X_i - t}{b_n}\right),$$

   wobei die „Kernfunktion" K (mindestens) die Eigenschaften $K(x) \geq 0$ und $\int_{-\infty}^{\infty} K(x)dx = 1$ erfüllt und wobei für die „Bandbreite" b_n gilt, dass $b_n \to 0$ und $nb_n \to \infty$ für $n \to \infty$.

Ist (z. B.) $K(x) = \frac{1}{2} \cdot 1_{\{-1<x\leq 1\}}$ der „Rechteckskern“, erhält man

$$f_n(t) = \frac{1}{2nb_n} \sum_{i=1}^{n} 1_{\left\{-1<\frac{X_{i:n}-t}{b_n}\leq 1\right\}} = \frac{F_n(t+b_n) - F_n(t-b_n)}{2b_n},$$

also einen Differenzenquotienten der – noch nicht einmal stetigen – empirischen Verteilungsfunktion $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_{i:n}\leq t\}}$. (Memo: $X_{i:n}$, $i = 1, \ldots, n$, bezeichnet die Ordnungsstatistiken der X_i, also die aufsteigend sortierten X-Werte.)

Die eigentliche Aufgabe:

Für die $n = 201$ Mietdaten $(x_1, \ldots, x_n)$ aus Aufgabe 3 von Blatt 6 soll das obige f_n mit Rechteckskern und einer festen Bandbreite $b_n > 0$ an ausgewählten Stellen $t_1, \ldots, t_k$ bestimmt werden. D. h., es ist der Vektor $(f_n(t_1), \ldots, f_n(t_k))$ zu ermitteln.

Verwenden Sie konkret (zunächst) $b_n = 100$ sowie als t_j mit $j = 1, \ldots, k = 10$ ein maximal ausgedehntes, äquidistantes Gitter im Intervall $[500, 4000]$ und gehen Sie wie folgt vor:

a) Bilden Sie aus den Vektoren $(x_1, \ldots, x_n)$ und $(t_1, \ldots, t_k)$ die Matrix aller paarweisen Differenzen ihrer Elemente, d. h.

$$D := \left(x_{i:n} - t_j\right)_{\substack{1\leq i\leq n\\ 1\leq j\leq k}}$$

unter Verwendung der Funktion `outer` mit dem Argument `FUN = "-"`.

b) Ermitteln Sie sodann die Indikatoren-Matrix

$$Ind := \left(1_{\{-1<D_{ij}/b_n\leq 1\}}\right)_{\substack{1\leq i\leq n\\ 1\leq j\leq k}}$$

c) Berechnen Sie schließlich $(f_n(t_1), \ldots, f_n(t_k))$ unter Verwendung der Matrix Ind und *ausschließlich* der Matrix-Vektor-Multiplikation.

d) Lassen Sie sich die Werte $f_n(t_j)$, $j = 1, \ldots, k$ in Form eines Polygonzuges durch die Punkte $(t_j, f_n(t_j))$, $j = 1, \ldots, k$ mit Hilfe von `plot(x, y, type = "l")` zeichnen, wobei an `x` der Vektor der t_js und an `y` der Vektor der $f_n(t_j)$s übergeben werden müssen.

e) Erhöhen Sie k und wiederholen Sie Teile 3a bis 3d mit variierendem b_n.