MDS – Decision Trees

Aufgabe 1

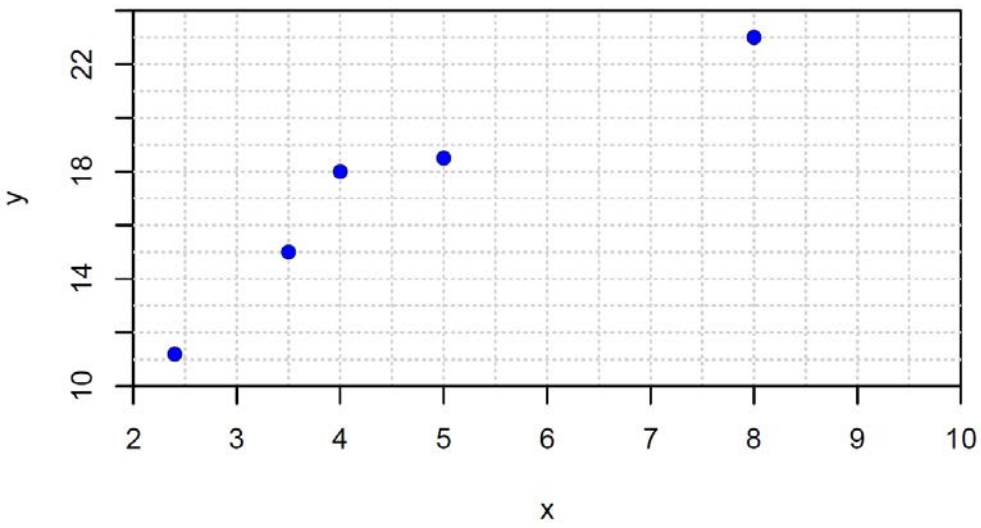
Wir wollen in dieser Aufgabe von Hand einen Regressionsbaum trainieren (rechnen). Dazu verwenden wir einen einfachen Datensatz mit einer Input-Variablen x und der Zielvariable y. Unser Datensatz hat nur $n = 5$ Beobachtungen und sieht folgendermassen aus:

x	y
2.4	11.2
3.5	15
4	18
5	18.5
8	23

a) Der Wurzelknoten (engl. *Root Node*) ist der oberste Knoten im Baum und entspricht einem Baum ohne Splits. Für diesen Baum ohne Splits würde für jede Beobachtung der Durchschnitt $\bar{y}$ im Datensatz vorhergesagt werden. Berechnen Sie basierend auf obigem Datensatz den Durchschnitt $\bar{y}$ sowie die Summe der quadrierten Residuen (SQR), die sich daraus ergibt.

b) Sie möchten nun den ersten optimalen Split im Baum berechnen. Da unser Datensatz nur eine erklärende Variable x hat, müssen wir lediglich verschiedene Splitpunkte s für die erklärende Variable x betrachten. Als mögliche Splitpunkte infrage kommen die jeweiligen Mittelpunkte zwischen zwei Beobachtungen: 2.95, 3.75, 4.5, 6.5. Berechnen Sie für jeden der vier möglichen Splitpunkte die Vorhersagen in den resultierenden zwei Regionen und die Summe der quadrierten Residuen (SQR). Welcher Split ist optimal?

c) Zeichnen Sie den optimalen Split ins untenstehende Diagramm ein. Zeichnen Sie auch die resultierenden Vorhersagen in den beiden Regionen ein. Achtung: es handelt sich unten nicht um den Input-Space mit zwei Input-Variablen, sondern um ein Diagramm mit der Zielvariable auf der y-Achse und der Input-Variable auf der x-Achse.



d) Berechnen Sie nun den optimalen Split in der rechten Region, die aus dem ersten Split resultiert. Welcher Splitpunkt ist optimal? Zeichnen Sie diesen nächsten Split ebenfalls in das Diagramm oben ein.

e) Verwenden Sie den folgenden R-Code, um die von Ihnen berechneten ersten zwei Splitpunkte zu überprüfen. Kommt R auf dasselbe Resultat wie Sie?

```
# Lade tree Paket
library(tree)

# Datenpunkte generieren
x <- c(2.4, 3.5, 4, 5, 8)
y <- c(11.2, 15, 18, 18.5, 23)

# Rechne Regressionsbaum ohne Begrenzungen
reg.tree <- tree(y ~ x, control = tree.control(nobs = 5, minsize = 3, mindev = 0.0))

# Wie sieht der resultierende Baum aus?
reg.tree

# Plot des Baums
plot(reg.tree)
text(reg.tree, pretty = 1)
```