ETH Zürich
Statistik II (BIOL, HST) **Lösungen zur Serie 2 - 01**

1. **Aufgabe**
   Es soll eine Zielgrösse (y) durch mehrere erklärende Variablen (x) vorhergesagt werden. Es wird vermutet, dass nicht alle erklärenden Variablen für eine gute Vorhersage nötig sind. Das Ziel dieser Aufgabe ist es, ein Subset der erklärenden Variablen zu finden, welches eine möglichst gute Vorhersage zulässt.

   Die Daten sind in folgendem rda-File gespeichert: *ueb*816173.*rda*. Laden Sie dieses File mit dem Befehl `load` oder über das Menü in RStudio. Das rda-File enthält zwei data frames: `dat` enthält Trainingsdaten mit einer Zielgrösse `y` und mehreren erklärenden Variablen (`x1`, `x2`, etc.). `datTest` ist ein Testdatensatz mit einer Beobachtung, bei dem der Wert für die Variable `y` fehlt (`NA`).

   (a) Laden Sie die Daten. Im Trainingsdatensatz steht in Zeile 87 der Daten der Variable x1 der Wert 0.042 .

   (b) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit best subset selection). Dieses Modell besteht aus den Variablen x1 x4 x6 x11 .

   (c) Nehmen Sie an, dass das beste Modell ausschliesslich die Variablen x10 x8 x9 (und einen Achsenabschnitt) enthält. Dieses Modell sagt auf dem Testdatensatz den Wert 3.025 voraus.

   (d) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit forward selection). Dieses Modell besteht aus den Variablen x11 x2 x4 x5 .

   (e) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit backward selection). Dieses Modell besteht aus den Variablen x1 x3 x4 x6 x10 x11 .

   Lösung

```
(b)
library(leaps)
m1 <- regsubsets(y ~ ., data = dat, method = "exhaustive", nvmax = 11)
m1s <- summary(m1)
m1s$bic
(c)
fit1 <- lm(y ~ x10+x8+x9, data = dat)
predict(fit1, newdata = datTest)
(d)
m2 <- regsubsets(y ~ ., data = dat, method = "forward", nvmax = 11)
(e)
m3 <- regsubsets(y ~ ., data = dat, method = "backward", nvmax = 11)
```

   (a) **Falsch**. Der richtige Wert ist 0.768 .

   (b) **Richtig**. Das gewählte Subset der Variablen bei best-subset selection ist x1 x4 x6 x11 .

   (c) **Richtig**. Der wahre Wert ist 3.025 .

   (d) **Falsch**. Das gewählte Subset der Variablen bei forward selection ist x1 x4 x6 x11 .

   (e) **Falsch**. Das gewählte Subset der Variablen bei backward selection ist x1 x4 x6 x11 .

2. **Aufgabe**
   Es soll eine Zielgrösse (y) durch mehrere erklärende Variablen (x) vorhergesagt werden.

Es wird vermutet, dass nicht alle erklärenden Variablen für eine gute Vorhersage nötig sind. Das Ziel dieser Aufgabe ist es, ein Subset der erklärenden Variablen zu finden, welches eine möglichst gute Vorhersage zulässt.

Die Daten sind in folgendem rda-File gespeichert: *ueb*685877.*rda*. Laden Sie dieses File mit dem Befehl `load` oder über das Menü in RStudio. Das rda-File enthält zwei data frames: `dat` enthält Trainingsdaten mit einer Zielgrösse `y` und mehreren erklärenden Variablen (`x1`, `x2`, etc.). `datTest` ist ein Testdatensatz mit einer Beobachtung, bei dem der Wert für die Variable `y` fehlt (`NA`).

(a) Laden Sie die Daten. Im Trainingsdatensatz steht in Zeile 15 der Daten der Variable x4 der Wert 0.935 .

(b) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit best subset selection). Dieses Modell besteht aus den Variablen x2 x3 x5 x6 .

(c) Nehmen Sie an, dass das beste Modell ausschliesslich die Variablen x1 x2 x3 x6 (und einen Achsenabschnitt) enthält. Dieses Modell sagt auf dem Testdatensatz den Wert 6.546 voraus.

(d) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit forward selection). Dieses Modell besteht aus den Variablen x2 x3 x6 .

(e) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit backward selection). Dieses Modell besteht aus den Variablen x2 x3 x4 x5 x6 .

Lösung

```
(b)
library(leaps)
m1 <- regsubsets(y ~ ., data = dat, method = "exhaustive", nvmax = 6)
m1s <- summary(m1)
m1s$bic
(c)
fit1 <- lm(y ~ x1+x2+x3+x6, data = dat)
predict(fit1, newdata = datTest)
(d)
m2 <- regsubsets(y ~ ., data = dat, method = "forward", nvmax = 6)
(e)
m3 <- regsubsets(y ~ ., data = dat, method = "backward", nvmax = 6)
```

(a) **Falsch**. Der richtige Wert ist -0.786 .

(b) **Falsch**. Das gewählte Subset der Variablen bei best-subset selection ist x2 x3 x6 .

(c) **Richtig**. Der wahre Wert ist 6.546 .

(d) **Richtig**. Das gewählte Subset der Variablen bei forward selection ist x2 x3 x6 .

(e) **Falsch**. Das gewählte Subset der Variablen bei backward selection ist x2 x3 x6 .

3. **Aufgabe**
Es soll eine Zielgrösse (*y*) durch mehrere erklärende Variablen (*x*) vorhergesagt werden. Es wird vermutet, dass nicht alle erklärenden Variablen für eine gute Vorhersage nötig sind. Das Ziel dieser Aufgabe ist es, ein Subset der erklärenden Variablen zu finden, welches eine möglichst gute Vorhersage zulässt.

Die Daten sind in folgendem rda-File gespeichert: *ueb*409640.*rda*. Laden Sie dieses File mit dem Befehl `load` oder über das Menü in RStudio. Das rda-File enthält zwei data frames: `dat` enthält Trainingsdaten mit einer Zielgrösse `y` und mehreren erklärenden Variablen (`x1`, `x2`, etc.). `datTest` ist ein Testdatensatz mit einer Beobachtung, bei dem der Wert für die Variable `y` fehlt (`NA`).

(a) Laden Sie die Daten. Im Trainingsdatensatz steht in Zeile 206 der Daten der Variable x1 der Wert 0.346 .

(b) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit best subset selection). Dieses Modell besteht aus den Variablen x1 x4 .

(c) Nehmen Sie an, dass das beste Modell ausschliesslich die Variablen x1 x6 (und einen Achsenabschnitt) enthält. Dieses Modell sagt auf dem Testdatensatz den Wert 2.94 voraus.

(d) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit forward selection). Dieses Modell besteht aus den Variablen x1 x4 .

(e) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit backward selection). Dieses Modell besteht aus den Variablen x1 x4 x7 x10 .

Lösung

```
(b)
library(leaps)
m1 <- regsubsets(y ~ ., data = dat, method = "exhaustive", nvmax = 10)
m1s <- summary(m1)
m1s$bic
(c)
fit1 <- lm(y ~ x1+x6, data = dat)
predict(fit1, newdata = datTest)
(d)
m2 <- regsubsets(y ~ ., data = dat, method = "forward", nvmax = 10)
(e)
m3 <- regsubsets(y ~ ., data = dat, method = "backward", nvmax = 10)
```

(a) **Falsch**. Der richtige Wert ist -0.41 .

(b) **Richtig**. Das gewählte Subset der Variablen bei best-subset selection ist x1 x4 .

(c) **Richtig**. Der wahre Wert ist 2.94 .

(d) **Richtig**. Das gewählte Subset der Variablen bei forward selection ist x1 x4 .

(e) **Falsch**. Das gewählte Subset der Variablen bei backward selection ist x1 x4 .