

1. 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: *ueb816173.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 ( $x_1, x_2, \dots$ ). `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  $x_1$  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  $x_1, x_4, x_6, x_{11}$  .
- (c) Nehmen Sie an, dass das beste Modell ausschliesslich die Variablen  $x_{10}, x_8, x_9$  (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  $x_{11}, x_2, x_4, x_5$  .
- (e) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit backward selection). Dieses Modell besteht aus den Variablen  $x_1, x_3, x_4, x_6, x_{10}, x_{11}$  .

2. 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: *ueb685877.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 ( $x_1, x_2, \dots$ ). `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  $x_4$  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  $x_2, x_3, x_5, x_6$  .
- (c) Nehmen Sie an, dass das beste Modell ausschliesslich die Variablen  $x_1, x_2, x_3, x_6$  (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  $x_2, x_3, x_6$  .
- (e) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit backward selection). Dieses Modell besteht aus den Variablen  $x_2, x_3, x_4, x_5, x_6$  .

3. 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: *ueb409640.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 ( $x_1$ ,  $x_2$ , 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  $x_1$  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  $x_1$   $x_4$  .
- (c) Nehmen Sie an, dass das beste Modell ausschliesslich die Variablen  $x_1$   $x_6$  (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  $x_1$   $x_4$  .
- (e) Suchen Sie das Subset der Variablen, welches das beste Modell (bzgl. BIC) produziert (mit backward selection). Dieses Modell besteht aus den Variablen  $x_1$   $x_4$   $x_7$   $x_{10}$  .