Aufgaben zur Kreuzvalidierung

- Wiederholen Sie die Modellselektion aus dem Beispiel mit leave-one-out Kreuzvalidierung!
- Überlegen Sie hierzu, wie systematisch die nötigen Trainings- und Testdatensätze erzeugt werden können.
- Für jede Aufteilung muss das Modell neu geschätzt werden und für den jeweils ausgelassenen Punkt eine Prognose erstellt werden.
- Schließlich müssen die so gewonnenen Prognosefehler gemittelt werden.
- Lösung hier.

Modellselektion mittels leave-one-out CV und Prognosefehler

```
prognosefehler <- rep(NA, length(xdata))
for (leftout in 1:length(xdata))
  {
    training <- seq(1,length(xdata))[-leftout]
    x <- xdata[training]
    y <- ydata[training]
    test <- leftout
    prognosefehler[leftout] <-
      (predict(lm(y ~ x), data.frame(x=xdata[test])) -
                                          ydata[test])^2
  }
cat("Mittlerer quadratischer Prognosefehler",
                                mean(prognosefehler))
```

```
## linear lm(y ~ x)
Mittlerer quadratischer Prognosefehler 7.454557
### quadratisch lm(y ~ x + I(x^2))
Mittlerer quadratischer Prognosefehler 6.947941
### kubisch lm(y ~ x + I(x^2) + I(x^3))
Mittlerer quadratischer Prognosefehler 6.468842
### 4. Grades lm(y ~ x + I(x^2) + I(x^3) + I(x^4))
Mittlerer quadratischer Prognosefehler 18.53203
```

Resampling und Hypothesentests

- Literatur: Introduction to the Bootstrap World, D.D. Boos, Statistical Science , 2003, Vol 18, No 2, 168-174
- Sowohl der Bootstrap als auch der Jackknife sind verteilungsunabhängige Verfahren.
- Mit beiden können über die Resampling Stichproben prinzipiell beliebige Statistiken und ihre Standardfehler geschätzt werden.
- Ein Spezialfall ist allerdings das Testen von Hypothesen und damit verbunden das Herleiten von empirischen p-Werten über Resampling.
- Der Fallstrick beim Resampling von p-Werten liegt in der entscheidenden Voraussetzung des Testens: Die Stichproben müssen unter der Nullhypothese H_0 gezogen werden!

Beispiel Bootstrap-Test

- Angenommen man hat zwei unabhängige Stichproben $X_1, \ldots, X_m$ und $Y_1, \ldots, Y_n$ und möchte die Differenz der Mittelwerte der beiden Stichproben $\mu_X - \mu_Y$ gegen Gleichheit testen.

- Für ein nicht-parametrisches Konfidenzintervall zöge man einfach Stichproben mit Zurücklegen gemäß der empirischen Verteilungsfunktionen von X und Y.

- Testet man aber die Hypothese $H_0 : \mu_X - \mu_y = 0$ aus Bootstrap-Stichproben, z.B. über die gepoolte t-Statistik

$$t_p = \frac{(\bar{X} - \bar{Y})}{\sqrt{s_p^2(\frac{1}{m} + \frac{1}{n})}},$$

wobei

$$s_p^2 = \frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}$$

und S_X^2, S_Y^2 die üblichen erwartungstreuen Varianzschätzer der beiden Stichproben X und Y, so wird die Bedingung der Nullhypothese, nämlich übereinstimmende Mittelwerte, in der Regel nicht eingehalten!

- Um eine zulässige Resampling-Methode für Hypothesentests zu erhalten, muss also eine Methode gefunden werden, um Stichproben unter der gültigen Nullhypothese zu generieren.

- Eine Möglichkeit wäre ein Bootstrapping von Stichproben X^* vom Umfang m und Y^* vom Umfang n aus der gepoolten Stichprobe $(X_1, \ldots, X_m, Y_1, \ldots, Y_m)$.

- Das Zusammenfassen der Teilstichproben erzwingt die unabhängig identische Verteilung aller Beobachtungen der Stichproben X^* und Y^*!

- Tatsächlich würden so sogar Stichproben unter der viel strengeren Nullhypothese $F_X = F_Y$ gezogen!

- Eine elegante Formulierung für Tests unter dieser Nullhypothese bilden die sogenannten Permutationstests, die im weiteren Verlauf der Veranstaltung vorgestellt werden.

- Tatsächlich lässt sich aber auch ein Bootstrap-Test formulieren, der nur auf der deutlich schwächeren Annahme $H_0 : \mu_X = \mu_Y$ beruht!

- Es ist also immer wichtig, exakt zu überlegen, welche Hypothese und welche Alternative man trennen möchte!

- Der Zweistichproben-Welchtest mit der Teststatistik

$$t_W = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}}$$

beispielsweise wäre für diese Hypothese angemessen, wenn nichts weiter als die Existenz der zweiten Momente der Stichprobenverteilungen voraus gesetzt wird.

- Das adäquate Bootstrapverfahren unter der Nullhypothese würde dann getrennte (Bootstrap-)Stichproben aus $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_m - \bar{X})$ und $(Y_1 - \bar{Y}, Y_2 - \bar{Y}, \ldots, Y_n - \bar{Y})$ erzeugen und wiederholt die Teststatistik berechnen.

- Aus diesem Prinzip konstruierte Tests heißen auch Monte-Carlo Tests.

Der Bootstrap p-Wert

- Angenommen T_0 ist der Wert einer Teststatistik aus einer gegebenen Stichprobe. Dann ist $P(T > T_0|H_0)$ die Definition des p-Wertes für eine Nullhypothese, bei der große Werte von T gegen die Hypothese sprechen. (Achtung: Wie bei normalen Tests ist die Formulierung wichtig!)

- Ist die Verteilung von T eine diskrete Gleichverteilung auf k möglichen Ausprägungen $t_1, \ldots, t_k$, so ist der p-Wert genau der Anteil der $t_i \geq T_0$.

- Entsprechend, wenn man die diskreten Realisierungen T_i^* von T der nicht-parametrischen Bootstrapstichproben unter H_0 betrachtet, kann man den Bootstrap-p-Wert definieren als

$$p_B = \frac{\text{Anz. } T_i^* \geq T_0}{\text{Anz. Replikationen} \, (=: B)}.$$

- In der Literatur findet sich manchmal auch die Definition

$$p_B = \frac{Anz.T_i^* \geq T_0 + 1}{B + 1}.$$

  Mit wachsender Replikationszahl verschwinden die Unterschiede der Definitionen.

- Im Falle des parametrischen Bootstrap aus einer stetigen Verteilung ist die Voraussetzung der diskreten Gleichverteilung der T_i^* sogar exakt erfüllt!

- Der p-Wert p_B ist in diesem Falle exakt diskret gleichverteilt auf den Werten $\left\{0, \frac{1}{B}, \frac{2}{B}, \ldots, \frac{B}{B}\right\}$, weshalb gilt, dass das Niveau α des Tests exakt eingehalten wird, wenn $(B+1)\alpha \in \mathcal{Z}$ gilt.

- Dies gilt, da unter der Nullhypothese $T_0, T_1^*, \ldots, T_B^*$ i.i.d. sind und deshalb jede der $(B+1)$! Permutationen dieser Werte gleichwahrscheinlich ist.

- Aus diesem Grund ist es geschickt B so zu wählen, dass die Ganzzahligkeit von $\alpha(B+1)$ gegen ist.

- Beispiel: $\alpha = 0.05,\ B = 19, 39$ oder 99.

- Die asymtotische Zulässigkeit dieses Verfahrens auch für den parametrischen Bootstrap wurde ebenfalls untersucht. (Hall 1986)

- Was die Konvergenz der so simulierten p-Werte betrifft, so gibt es Resultate, die unter erträglich strengen Regularitätsbedingungen die asymtotische Gleichverteilung der p-Werte zeigen. Insgesamt scheint die Verwendung von Bootstrap-p-Werten angemessen, wenn man im Hinterkopf behält, dass es sich nicht um exakte p-Werte handelt!