Angewandte Regression — Serie 6

1. Der Datensatz `support3.csv` beinhaltet die totalen Spitalkosten von Patienten (mit gewissen Krankheiten) in amerikanischen Spitälern zwischen 1989 und 1991. Die verschiedenen Variablen sind:

`totcst`	totale Kosten
`age`	Alter
`dzgroup`	Krankheitsgruppe
`num.co`	Anzahl von Komorbidität (Mehrfachdiagonose)
`edu`	Jahre der Ausbildung
`income`	Einkommen
`scoma`	Equivalentes Mass für Glasgow-Koma-Wert
`meanbp`	Mittelwert Blutdruck
`hrt`	Puls
`resp`	Atemfrequenz
`temp`	Temperatur
`race`	Rasse
`pafi`	Verhältnis PaO2/FiO2 (Blut-Gasmischung)

Quelle: F. E. Harrall, Regression Modeling Strategies

Wir möchten die Totalkosten linear modelllieren. Für die Modellauswahl haben wir die *forward-*,*backward-* und die *subset*-Techniken kennen gelernt. In dieser Aufgabe befassen wir uns mit zwei weiteren Methoden: *Lasso* und *Ridge Regression.*

a) Gibt es sinnvolle First-Aid-Transformationen?

b) Machen Sie Modellauswahl (ohne Wechselwirkungen, ohne höhere Terme) mit den uns bekannten Techniken: forward und backward.

c) Lasso: die Lassotechnik (siehe Vorlesung 5.3.n) minimiert die Funktion

$$Q(\underline{\beta}, \lambda) = \sum_i R_i + \lambda \sum_i |\beta_i^*|,$$

wobei β_i^* die standartisierten Koeffizienten sind.
Mit `source("ftp://stat.ethz.ch/WBL/Source-WBL-2/R/lassogrp.R")` können Sie die Lasso-Prozeduren aufrufen. **Wichtig:** Da momentan diese Lasso-Prozeduren noch nicht als R-Package vorhanden sind, finden Sie unter *ftp://stat.ethz.ch/WBL/Source-WBL-2/R/lassogrp.pdf* die nötigen und weiteren Informationen.

- Starten Sie mit dem vollen Modell (ohne Wechselwirkungen, ohne höhere Terme) und verwenden Sie den Befehl `r.lasso<-lasso(`*`formula`*`,`*`data`*`)`. Was ist der Output von `r.lasso[c(1,10,20)]` (resp für irgendwelche Integer ≤ 21)?
- Was wird mit `plot(r.lasso, type=type="norms")` geplotet?
- Welches λ wählen Sie als Ihr Modell?
- Machen Sie eine Residuenanalyse. Hinweis: Benützen Sie den Befehl `e.lasso<-extract.lassogrp(r.lasso,lambda=`*`IhrLambda`*`)` und machen Sie einen `plot`.

d) Verbessern Sie Ihr Modell bei Hinzunahme von Wechselwirkungen und höheren Termen. Residuenanalyse?

e) Ridge Regression: die Ridge Regression minimiert die Funktion

$$Q(\underline{\beta}, \lambda) = \sum_i R_i + \lambda \sum_i (\beta_i^*)^2,$$

wobei β_i^* die standartisierten Koeffizienten sind.
Mit `library(MASS)` können Sie die Ridge Regression-Prozeduren und helps aufrufen.

- Starten Sie mit einem vollen Modell (ohne Interaktionen, höhere Terme) und verwenden Sie den Befehl `lm.ridge(`*`formula,data,lambda=`*`)`, zb für *lambda=seq(1,800,1)*. Benützen Sie `lm.ridge(...)$coef[,1:3]` um die Struktur von lm.ridge zu sehen.
- Was bewirkt `plot()`? Vergleichen Sie diesen Plot mit dem Plot von Lasso. Ausser den vielen Linien, was fällt auf?