Arbeitsblatt 7

Multiple lineare Regression (Variablenselektion)

Aufgabe 1: Spitäler

In einer Studie über die Kontrolle des Infektionsrisikos in US-amerikanischen Spitälern wurden bei einer Zufallsstichprobe von 113 Spitälern folgende Variablen erfasst:

- `id`: Identifikationsnummer des Spitals
- `length`: mittlere Hospitalisationslänge aller PatientInnen (in Tagen)
- `age`: mittleres Alter der PatientInnen
- `inf`: mittleres Infektionsrisiko (in Prozent)
- `cult`: Anzahl bakteriologischer Tests pro nichtsymptomatische PatientIn x 100
- `xray`: Anzahl Röntgenuntersuchungen pro nichtsymptomatische PatientIn x 100
- `beds`: Anzahl Betten
- `school`: Universitätsklinik 1=ja 2=nein
- `region`: geographische Region 1=NE 2=N 3=S 4=W
- `pat`: mittlere Anzahl PatientInnen im Spital pro Tag
- `nurs`: mittlere Anz. vollbeschäftigter, ausgebildeter Krankenschwestern und Pfleger
- `serv`: Prozentsatz angebotener Dienstleistungen aus insgesamt 35 möglichen

Die Daten befinden sich im File `senic.rda`. Das Ziel dieser Aufgabe ist es, ein gutes Modell für die mittlere Hospitalisationslänge (`length`) zu finden. Gut heisst dabei, dass es einerseits korrekt, andererseits leicht interpretierbar und ohne überflüssige Variablen ist. Folgen Sie dabei den folgenden Schritten:

(a)

Zuerst bereiten wir die Daten vor. Folgen Sie dabei den folgenden Schritten:

- Gibt es im Datensatz fehlende Werte? (`is.na(var)`)
- Gibt es Variablen mit ungewöhnlichen Werten? (`summary(var)`)
- Sind alle kategoriellen Variablen als `factor` gespeichert?
- Transformieren Sie die Variable `serv` zur einfacheren Interpretation in die Zahl der angebotenen Leistungen (`= senic$serv / 100 * 35`) um.
- Gibt es stark korrelierte Variablen? Wäre das ein Problem für die Anpassung?

(b)

Betrachten Sie die einzelnen Variablen und Ihre Verteilung. Wo würden Sie Tukey's First Aid Transformationen anwenden?

(c)

Passen Sie das volle Modell mit ihren vorbereiteten Daten aus (a) und (b) an und überprüfen Sie, ob es grobe Verletzungen der Modellannahmen gibt.

(d)

Prüfen Sie das angepasste Modell auf Multikollinearität. Falls ein Problem vorliegt, versuchen Sie dieses mittels Amputation oder Generierung neuer Variablen zu lösen.

(e)

Suchen Sie nach einem geeigneten Modell mittels Variablenselektion und überprüfen Sie die Gültigkeit des selektierten Modells mittels Residuenanalyse.

(f)

Interpretieren Sie Ihr selektiertes Modell aus (e).