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:

Diese Lösung ist einer von vielen Wegen, wie man zu einem guten Modell kommt. Durch die Wahl anderer Transformationen, ein anderes Vorgehen im Lösen der Multikollinearität oder einer anderen Variablenselektion besteht durchaus die Möglichkeit, dass Sie ein ebenso gutes oder vielleicht auch besseres Modell erhalten.

Einlesen der Daten:

```
load("data/senic.rda")
```

(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?

Als erstes entfernen wir die Variable `id` aus dem Modellierungsdatensatz, da dies eine blosse Durchnummerierung der Spitäler ist und daher ohne prädiktiven Wert.

```
senic <- senic[,-1]
```

Als nächstes überprüfen wir, ob der Datensatz fehlende Werte enthält:

```
any(is.na(senic)) # keine fehlenden Werte
```

```
[1] FALSE
```

Aus `summary(senic)` sieht man, dass der Datensatz keine ungewöhnlichen Werte für die einzelnen Variablen beinhaltet.

Weiter gilt es sicherzustellen, dass die kategoriellen Variablen als Faktoren gespeichert sind und zu überprüfen, ob alle Levels mit genügend Beobachtungen vorhanden sind:

```
senic$school <- factor(senic$school, labels=c("ja", "nein"))
table(senic$school)
```

```
  ja nein
  17   96
```

```
senic$region <- as.factor(senic$region)
table(senic$region)
```

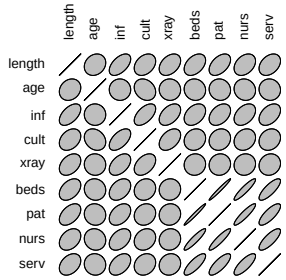
```
NE  N  S  W
28 32 37 16
```

Zur einfacheren Interpretation wandeln wir die erklärende Variable `serv` um. Die erklärende Variable weist den Prozentanteil an angebotenen Dienstleistungen aus. Wir wandeln sie in die Zahl der angebotenen Leistungen um:

```
senic$serv <- senic$serv / 100 * 35
```

Als letztes betrachten die Korrelationen:

```
library(ellipse, quietly = TRUE)
# Korrelation nur von stetigen Variablen:
senic2 <- senic[,!colnames(senic) %in% c("id","school","region")]
plotcorr(cor(senic2))
```

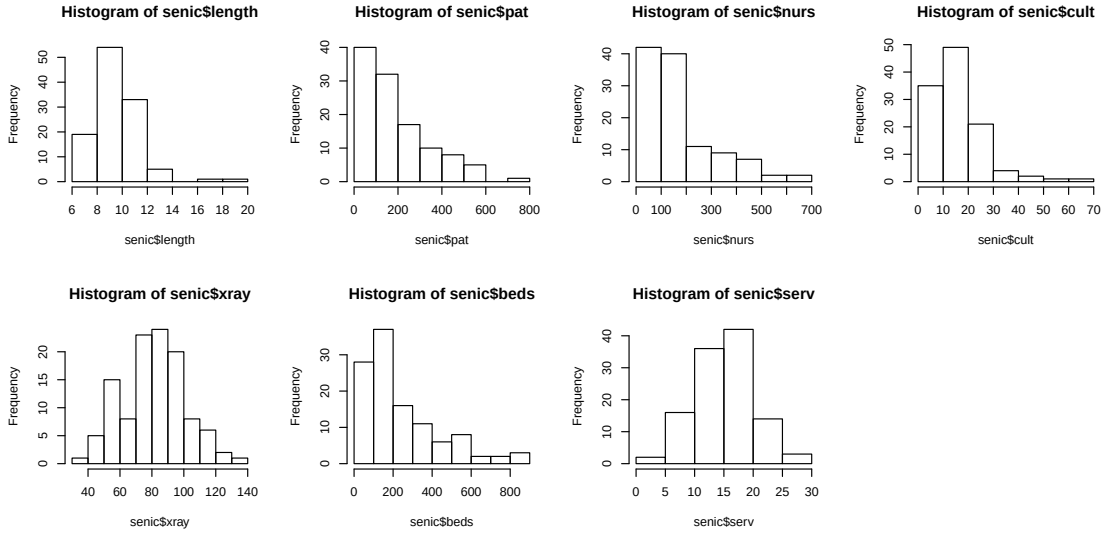


Die erklärenden Variablen `beds`, `pat`, `nurs` und `serv` korrelieren miteinander. Es könnten Multikollinearitätsprobleme beim Anpassen auftreten.

(b)

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

```
par(mfrow = c(2,4))
hist(senic$length)
hist(senic$pat)
hist(senic$nurs)
hist(senic$cult)
hist(senic$xray)
hist(senic$beds)
hist(senic$serv)
```



Aufgrund ihrer Definition und/oder rechtschiefe Verteilung sind die Variablen `length`, `pat` und `nurs` Kandidaten für eine Log-Transformation beziehungsweise `cult`, `xray`, `beds` und `serv` für die Wurzeltransformation (Tukey's First Aid Transformationen). Diese Transformationen sind **ein Vorschlag** der Dozentin. Es gibt durchaus andere Möglichkeiten (beispielsweise nur die Logarithmus Transformation auf rechtschiefe Verteilungen anzuwenden), die ebenfalls zu einem guten und gültigen Modell führen können.

(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.

```
# Datensatz zur Modellierung, damit anschliessend im lm() Befehl die
# Kurzschreibweise verwendet werden kann.
# Nicht transformierte Variablen:
senic2 <- senic[,c("age","inf","school","region")]
# Transformierte Variablen:
senic2$Loglength <- log(senic$length)
senic2$LogPat <- log(senic$pat)
senic2$LogNurs <- log(senic$nurs)
```

```
senic2$SqCult <- sqrt(senic$cult)
senic2$SqXray <- sqrt(senic$xray)
senic2$SqBeds <- sqrt(senic$beds)
senic2$SqServ <- sqrt(senic$serv)
```

Wir passen das volle Regressionsmodell mit Transformationen an:

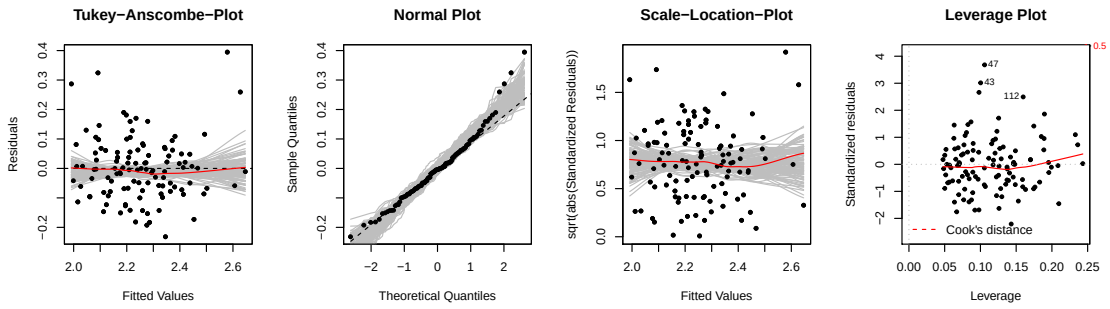
```
senic.lm1 <- lm(Loglength ~ ., data = senic2)
summary(senic.lm1)
```

```
Call:
lm(formula = Loglength ~ ., data = senic2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23192 -0.06872 -0.00394  0.05565  0.39474

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.3240637  0.2715829   4.875 4.09e-06 ***
age          0.0084075  0.0027390   3.070 0.002760 **
inf          0.0475954  0.0128782   3.696 0.000358 ***
schoolnein  -0.0651276  0.0377616  -1.725 0.087670 .
regionN     -0.0766583  0.0320482  -2.392 0.018627 *
regionS     -0.1174039  0.0317554  -3.697 0.000356 ***
regionW     -0.2016211  0.0427950  -4.711 7.96e-06 ***
LogPat       0.1419051  0.0583328   2.433 0.016764 *
LogNurs     -0.0593083  0.0403368  -1.470 0.144613
SqCult      -0.0028705  0.0130088  -0.221 0.825811
SqXray       0.0219524  0.0118210   1.857 0.066245 .
SqBeds      -0.0008259  0.0074391  -0.111 0.911819
SqServ      -0.0451765  0.0298513  -1.513 0.133337
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1134 on 100 degrees of freedom
Multiple R-squared:  0.6373,    Adjusted R-squared:  0.5938
F-statistic: 14.64 on 12 and 100 DF,  p-value: < 2.2e-16
```

Der globale F-Test ist signifikant. Das Modell enthält also mindestens eine Variable, die etwas zur Erklärung von der mittleren Hospitalisationslänge beiträgt.

Überprüfen der Gültigkeit des Modells mit grafischer Residuenanalyse:

```
load("data/resplot.rda")
par(mfrow = c(1,4))
resplot(senic.lm1)
```



Die Plots weisen auf keine groben Verletzungen der Annahmen hin, abgesehen von 4 Ausreissern. Für eine erste Anpassung ist dies ok, da diese Ausreisser im Leverage-Plot nicht als einflussreich klassifiziert werden.

(d)

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

```
# Die erste Spalte zeigt die VIF-Werte
vif(senic.lm1)[,1]
```

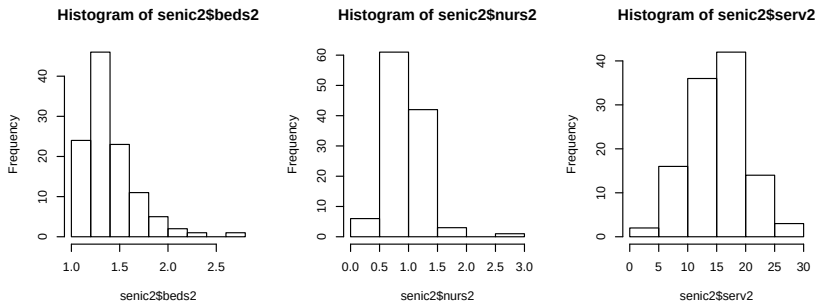
```
      age       inf    school    region    LogPat   LogNurs    SqCult
 1.300012  2.559825  1.600742  1.934902 19.400734  9.839187  2.206864
   SqXray    SqBeds    SqServ
 1.461710 14.974570  4.033445
```

Gewisse VIF-Werte übersteigen 5 deutlich. Es liegt hier also ein Multikollinearitäts-Problem vor, welches man angehen muss. Da sowohl Grösse, wie auch Auslastung, Betreuungsverhältnis und Breite der angebotenen Services eine Rolle für die Aufenthaltsdauer spielen dürften, werden wir hier das Multikollinearitätsproblem nicht über Amputation, sondern durch Generierung neuer Variablen lösen. Wir transformieren dazu die Grössen so, dass die Anzahl Patienten als Referenzgrösse gegeben ist und die anderen Grössen die Relation dazu geben:

```
# Generieren der relativen Variablen
senic2$beds2 <- senic$beds / senic$pat
senic2$nurs2 <- senic$nurs / senic$pat
senic2$serv2 <- senic$serv
```

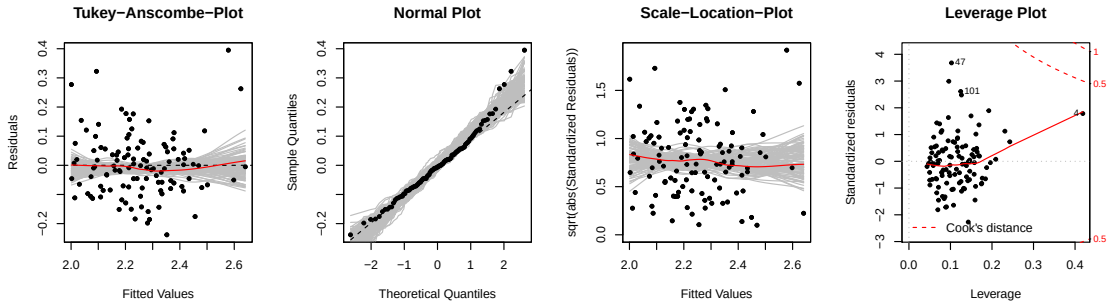
Nun überprüfen wir noch kurz die Rechschiefe in der Verteilung, um zu entscheiden, ob es besser ist, die neuen Variablen mit oder ohne Transformation ins Modell aufzunehmen.

```
par(mfrow = c(1,4))
hist(senic2$beds2)
hist(senic2$nurs2)
hist(senic2$serv2)
```



Bei den neuen Variablen zeigt nur noch für `beds2` eine Rechtschiefe. Daher ist das neue Modell:

```
senic2$SqBeds2 <- sqrt(senic2$beds2)
senic.lm2 <- lm(Loglength ~ age + inf + SqCult + SqXray + SqBeds2 +
                  school + region + LogPat + nurs2 + serv2, data = senic2)
par(mfrow = c(1,4))
resplot(senic.lm2)
```



Die Residuenplots des neue Modells `senic.lm2` zeigen ein fast identisches Bild zum vorangehenden Modell `senic.lm1`, das noch ohne die kollinearitätsbedingte Variablenänderung formuliert wurde. Mit Hilfe der VIFs prüfen wir, ob die Multikollinearität behoben werden konnte:

```
vif(senic.lm2)[,1]
```

```
     age      inf   SqCult   SqXray  SqBeds2   school   region
1.266347 2.509667 2.186628 1.468383 1.868661 1.523624 1.987177
  LogPat    nurs2    serv2
5.732534 1.710724 4.168489
```

Die VIFs sehen sehr viel besser aus und liegen nun fast alle unter dem Schwellwert von 5. Die verbleibende Multikollinearität versuchen wir über die Variablenselektion aufzulösen.

(e)

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

Wir führen hier eine schrittweise Variablenselektion basierend auf AIC durch.

```
senic.full <- lm(Loglength ~ age + inf + SqCult + SqXray + SqBeds2 +
                   school + region + LogPat + nurs2 + serv2, data = senic2)
senic.empty <- lm(Loglength ~ 1, data = senic2)
```

```
lm3.AIC <- step(senic.full, trace = FALSE,
                scope = list(lower = senic.empty, upper = senic.full))
lm3.AIC$anova
```

```
       Step Df     Deviance Resid. Df Resid. Dev       AIC
1           NA           NA       100   1.285144 -479.8464
2 - SqBeds2  1 0.0002581055       101   1.285402 -481.8237
3  - SqCult  1 0.0008498506       102   1.286252 -483.7490
```

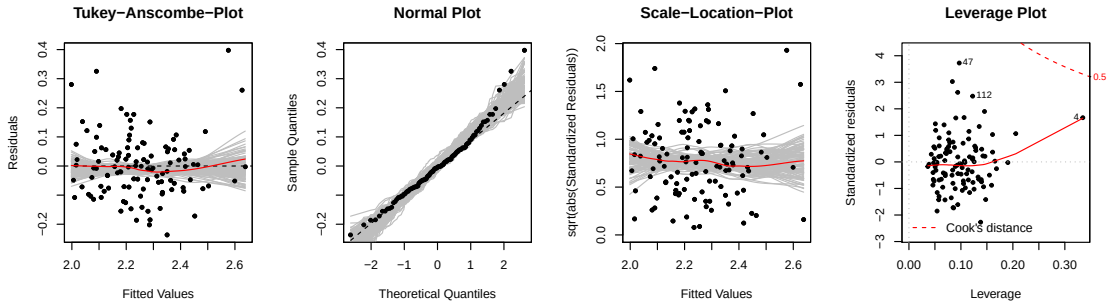
Es werden nur 2 Variablen aus dem Modell entfernt.

Wir passen das selektierte Modell an:

```
senic.lm3 <- lm(Loglength ~ age + inf + SqXray + school + region
                + LogPat + nurs2 + serv2, data = senic2)
```

Erneut führen wir eine Residuenanalyse durch, um die Gültigkeit dieses Modells zu verifizieren. In den Plots zeigt sich, dass das Modell korrekt ist und die Annahmen weitestgehend erfüllt. Auffallend sind wieder die 4 Ausreisser, die jedoch weiterhin ohne gefährlichen Einfluss auf die Modellanpassung sind.

```
par(mfrow = c(1,4))
resplot(senic.lm3)
```



Multikollinearität?

```
vif(senic.lm3)[,1]
```

```
     age      inf    SqXray   school   region   LogPat    nurs2
1.114385 1.862324 1.410470 1.486516 1.648524 5.032742 1.563888
   serv2
4.058354
```

Das Problem der Multikollinearität ist behoben. Alle VIF-Werte sind kleiner als 5.

(f)

Interpretieren Sie Ihr selektiertes Modell aus (e).

```
summary(senic.lm3)$adj.r.squared
```

```
[1] 0.6018679
```

```
coef(senic.lm3)
```

```
 (Intercept)          age          inf       SqXray   schoolnein
 1.304770724  0.008765008  0.044684119  0.022012636 -0.066651788
     regionN      regionS      regionW       LogPat        nurs2
-0.074110896 -0.117088099 -0.193940384  0.080659106 -0.070318281
       serv2
-0.006701707
```

```
drop1(senic.lm3, test = "F")
```

```
Single term deletions

Model:
Loglength ~ age + inf + SqXray + school + region + LogPat + nurs2 +
    serv2
       Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>              1.2862 -483.75
age     1   0.15370 1.4400 -472.99 12.1883  0.000712 ***
inf     1   0.21290 1.4991 -468.44 16.8830 8.045e-05 ***
SqXray  1   0.04623 1.3325 -481.76  3.6663  0.058326 .
school  1   0.04316 1.3294 -482.02  3.4227  0.067200 .
region  3   0.36337 1.6496 -461.63  9.6051 1.213e-05 ***
LogPat  1   0.09482 1.3811 -477.71  7.5195  0.007210 **
nurs2   1   0.03662 1.3229 -482.58  2.9036  0.091428 .
serv2   1   0.03508 1.3213 -482.71  2.7822  0.098385 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aus dem Summary-Output lässt sich entnehmen, dass das Modell 60% der Gesamtstreuung erklärt.

Als vergrössernd für die Aufenthaltszeit wirken folgende Prädiktoren: höheres Alter der Patienten (signifikant auf 5%), höheres Infektionsrisiko (signifikant auf 5%) und die Spitalgrösse, also die höhere Patientenzahl (signifikant auf 5%). Vermutlich ebenfalls zu einer leicht vergrössten Aufenthaltszeit führt die höhere Anzahl Röntgen. Dies ist aber auf 5% nicht signifikant.

Senkend auf die Hospitalisationslänge wirkt ein hohes Betreuungsverhältnis (d.h. viel Personal pro behandelten Patient) und auch eine hohe Anzahl an angebotenen Services aus. Allerdings können beide Effekte nicht als signifikant bestätigt werden, da die beiden zugehörigen t-Test je einen p-Wert grösser als 0.05 haben.

Ebenfalls nicht nachzuweisen sind Unterschiede zwischen normalen Spitälern und Unispitälern.

Jedoch bestehen erhebliche regionale Unterschiede in Bezug auf die Aufenthaltsdauer, wobei diese vor allem im S und W geringer ausfällt als im N und NE. Der F-Test der katiegoriellen Variable `region` ist hochsignifikant auf 5% (p-Wert ist deutlich kleiner als 0.05).