Universität Hamburg
Fachbereich Psychologie
46-02.020 Seminar: Datenanalyse, A
Seminarleiter: Ingmar Böschen
Wintersemester 2010/ 2011

13.03.11

Prüfungsleistung Quantitative Methoden II, WS10/11

Rick Bode Email: rick.bode@studium.uni-hamburg.de

Die EAM (European Association of Movement)-Ergebnisse von 2000 zeigten, dass deutsche Kinder, verglichen mit denen aus anderen europäischen Ländern, sich unterdurchschnittlich wenig bewegen. Die beliebtesten Freizeitaktivitäten seien Fernsehen und Computerspielen. Dies sorge kurzfristig für schlechte motorische Leistungen, langfristig aber würde Deutschland damit wahrscheinlich immer weniger Medaillen bei den olympischen Spielen gewinnen und außerdem würden die Kosten für das Gesundheitssystem exponentiell steigen. Diese erschreckenden Ergebnisse sorgen dafür, dass die bereits existierenden Schulkonzepte radikal überdacht werden müssen.

Einige Pilotprojekte wurden bereits gestartet und müssen nun bewertet werden. Unter anderem in der Gemeinde Erpel in Rheinland-Pfalz. In Erpel gibt es zwei Schulen, eine staatliche („Hans-Wolfgang-Erpel-Gesamtschule“) und eine private („Mens sana in corpore sano“-Schule), in der der Unterricht sich dadurch unterscheidet, dass die Kinder in Letzterer ein tägliches Bewegungsprogramm absolvieren (zum Beispiel: „Laufen mit der Maus“, „Yoga mit Hans“). Alle Kinder besuchten bis zur sechsten Klasse die örtliche Grundschule („Halbe-Erpel-Grundschule“) und sind nun seit einem halben Jahr auf die zwei Schulen verteilt. Ich überprüfe die motorischen Fähigkeiten von Kindern im Alter von zwölf bis dreizehn Jahren.

Verschiedene Einflüsse können nun eine Rolle auf die Entwicklung der motorischen Fähigkeiten der Kinder haben. So wurden die Schulform der Kinder $x_1;\ \begin{matrix} x_1 = 0 \rightarrow staatlicheSchule \\ x_1 = 1 \rightarrow privateSchule \end{matrix}$, das Alter der Kinder x_2, wobei der Eintritt ins zwölfte Lebensjahr als Nullpunkt oder Referenz gesetzt wurde, sodass die Abweichungen in Monaten angegeben sind. Des weiteren wurde die Körpergröße in Zentimetern x_3 und die von Kindern und Eltern geschätzte wöchentliche Schlafdauer der Kinder in Stunden x_4 registriert.

In diesem Aufsatz möchte wird mittels Regressionsanalyse überprüfen, ob es einen Zusammenhang zwischen diesen vier Faktoren und dem Testergebnis y gibt und wenn ja welche

der Faktoren ein ideales Modell vermitteln. Bei dem Test handelt es sich um den MoBS (Motorisches-Bewegungs-Standard-Inventar für Kinder und Jugendliche).

Daraus erbgeben sich folgende statistische Hypothesen für die einzelnen Faktoren:

$$H_0: \alpha_0=0 \;;\; \beta_0=0 \qquad H_1: \alpha_0=0 \;;\; \beta_0=0$$
$$\alpha_{x1}=\alpha_0 \;;\; \beta_{x1}=\beta_0 \qquad \alpha_{x1}\neq\alpha_0 \;;\; \beta_{x1}\neq\beta_0$$
$$\alpha_{x2}=\alpha_0 \;;\; \beta_{x2}=\beta_0 \qquad \alpha_{x2}\neq\alpha_0 \;;\; \beta_{x2}\neq\beta_0$$
$$\alpha_{x3}=\alpha_0 \;;\; \beta_{x3}=\beta_0 \qquad \alpha_{x3}\neq\alpha_0 \;;\; \beta_{x3}\neq\beta_0$$
$$\alpha_{x4}=\alpha_0 \;;\; \beta_{x4}=\beta_0 \qquad \alpha_{x4}\neq\alpha_0 \;;\; \beta_{x4}\neq\beta_0$$

Ziel ist es ein Modell zu finden, dass möglichst wenige dieser Variablen verwendet, aber gleichzeitig für eine möglichst große Aufklärung von y sorgt. Alle Tests werden dabei auf einem Signifikanzniveau von $p=0.05$ durchgeführt.

Für die Erhebung konnte ich insgesamt $n=97$ Versuchspersonen gewinnen. In Tabelle 1 findet sich eine Übersicht zu den entsprechenden Lage- und Verteilungsmaße der einzelnen Variablen.

	MoBS	Alter	Körpergröße	Schlaf pro Woche
Minimum	39.53	-0.858	126.7	32.50
1. Quartil	153.05	2.952	141.7	43.70
Median	367.12	5.103	146.6	52.50
Mean	311.97	4.943	146.8	51.64
3. Quartil	432.08	6.780	153.1	59.20
Maximum	530.82	12.932	165.9	69.10
sd	142.61	2.800	7.531	10.036

Tabelle 1: deskreptive Beschreibung

Insgesamt gibt es $H(x_1=0)=36$ Schüler, die die staatliche Schule besuchen und $H(x_1=1)=61$ Schüler, die die private Schule besuchen. Dabei befanden sich die Schüler zum Zeitpunkt der Erhebung in einem durchschnittlichen Alter von knapp zwölf Jahren und fünf Monaten. Über das Alter hinweg sind die Schüler mit einer Standardabweichung von etwas weniger als drei Monaten normal verteilt. Eine Normalverteilung zeigt sich ebenso bei der Körpergröße der Schüler. Diese variiert zwischen $x_{3\min}=1.267\,m$ und $x_{3\max}=1.659\,m$. Die wöchentliche Schlafdauer liegt zwischen $x_{4\min}=32.5\,\mathrm{h}$ und $x_{4\max}=69.1\,\mathrm{h}$, was einer durchschnittlichen

täglichen Schlafdauer von $4.64-9.87h$ entspricht. Die starke Varianz lässt sich wohl vor allem durch die mit der Pubertät verbundenen Änderungen erklären. Im Durchschnitt schlafen die Teilnehmer noch $x_{4\text{mean}}=7.38h$ täglich, was wohl dem Normalzustand in diesem Alter entspricht. Die entsprechenden Verteilungen dieser Werte zeigen sich auch noch einmal in Abbildung 1.

Mithilfe der Boxplots (Abbildung 2) lassen sich Ausreißer feststellen. Dabei zeigt sich, dass eine der Versuchspersonen ein stark abweichendes Alter hat. Hierbei handelt es sich um einen Teilnehmer, der erst ein Jahr später eingeschult wurde, er besucht jedoch genau so lange wie alle anderen Schüler die Schule. Außerdem wird in dem für Erzeugung der Grafiken verwendeten Programm „R“ zur Hilfe der Konstruktion von Boxplots die hinges (oben links), statt der Quartile genutzt (Groß, 20 $Y_i=277.9649\cdot x1_i+3.4114\cdot x2_i$ 10, S. 60, 70-71). In einer Darstellung, die die Quartile (oben rechts) verwendet ist kein Ausreißer mehr zu sehen. Ich gehe also davon aus, dass dieser Wert nicht für eine Verzerrung der Ergebnisse sorgt.

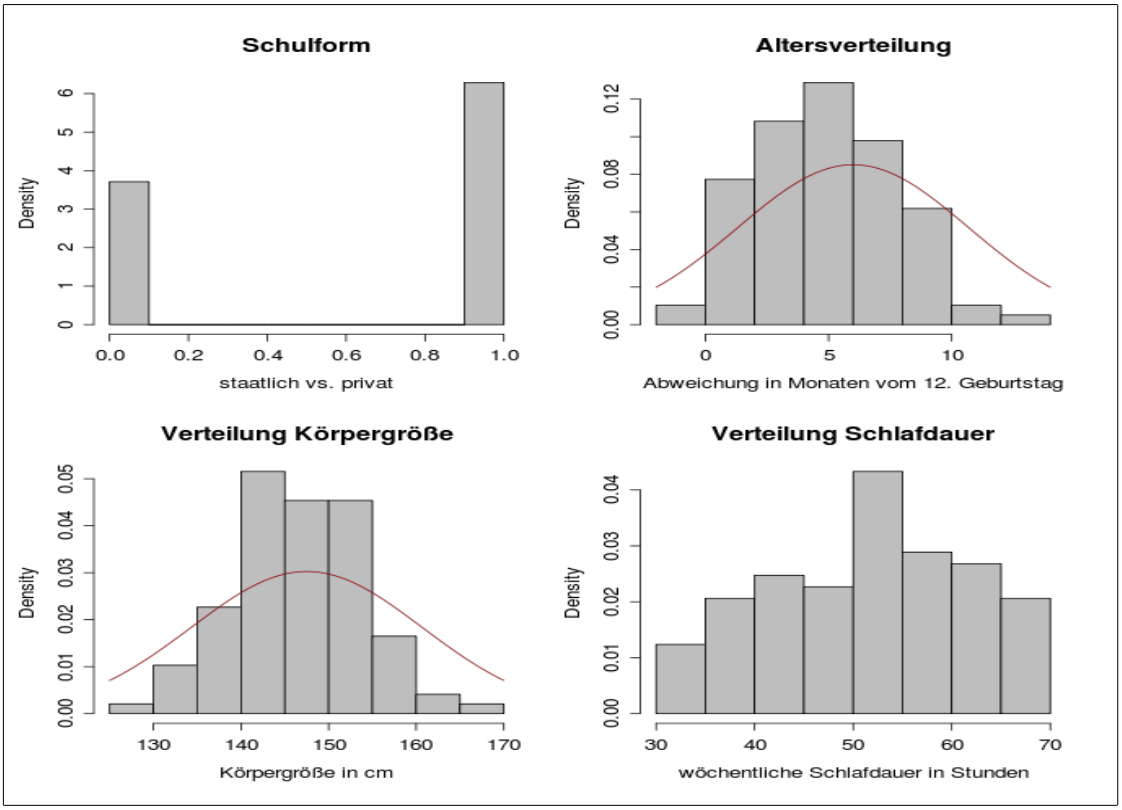

Abbildung 1: Histogramme

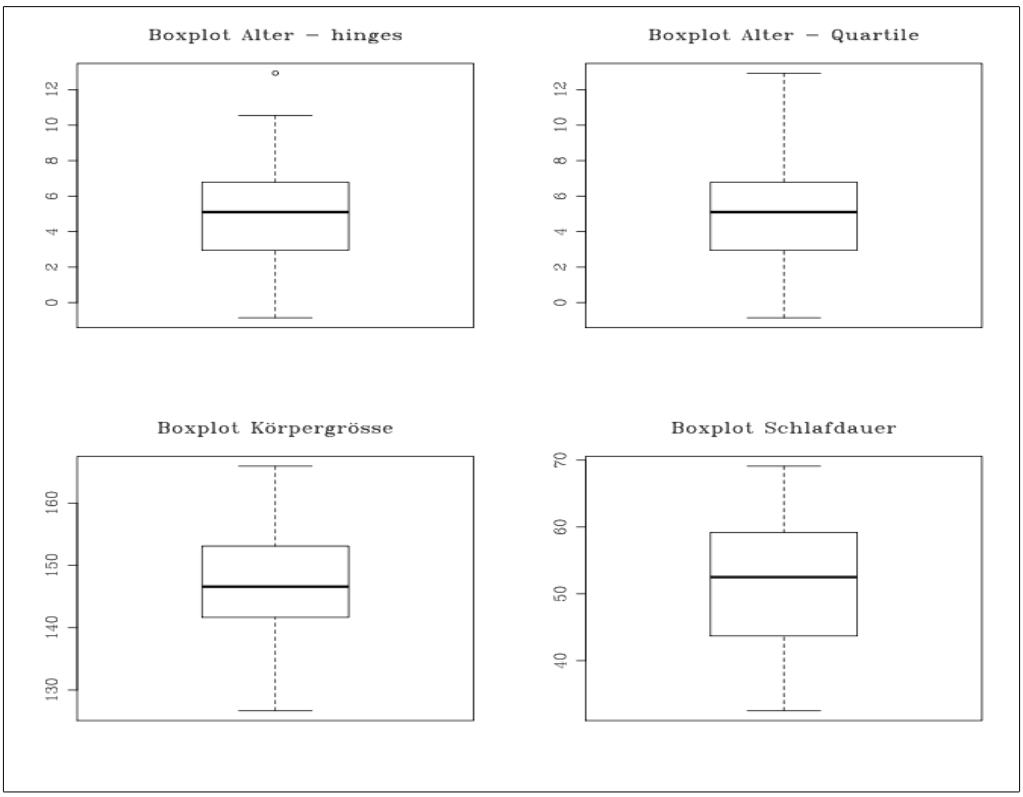

Abbildung 2: Boxplots

Bei der Überprüfung der einfachen linearen Modelle auf die Gauss-Markov-Annahmen zeigt sich, dass lediglich für die Regression zwischen der Schulform und dem Testwert im MoBS die nötigen Voraussetzungen für die Berechnung einer solchen erfüllt sind. In Abbildung 3 ist für dieses Modell zum einen die Verteilung der Residuen (links) und zum anderen, das lineare Regressionsmodell (rechts), inklusive entsprechender Punktewolke aufgeführt.

```
Call:
lm(formula = dat$y ~ dat$x1)

Residuals:
    Min      1Q  Median      3Q     Max
-121.222  -35.062   -2.072   36.248  116.678

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  138.852      8.319   16.69   <2e-16 ***
dat$x1       275.290     10.491   26.24   <2e-16 ***
---
Residual standard error: 49.92 on 95 degrees of freedom
Multiple R-squared: 0.8788,   Adjusted R-squared: 0.8775
F-statistic: 688.6 on 1 and 95 DF,  p-value: < 2.2e-16
```

Tabelle 2: Regression Schulform - MoBS

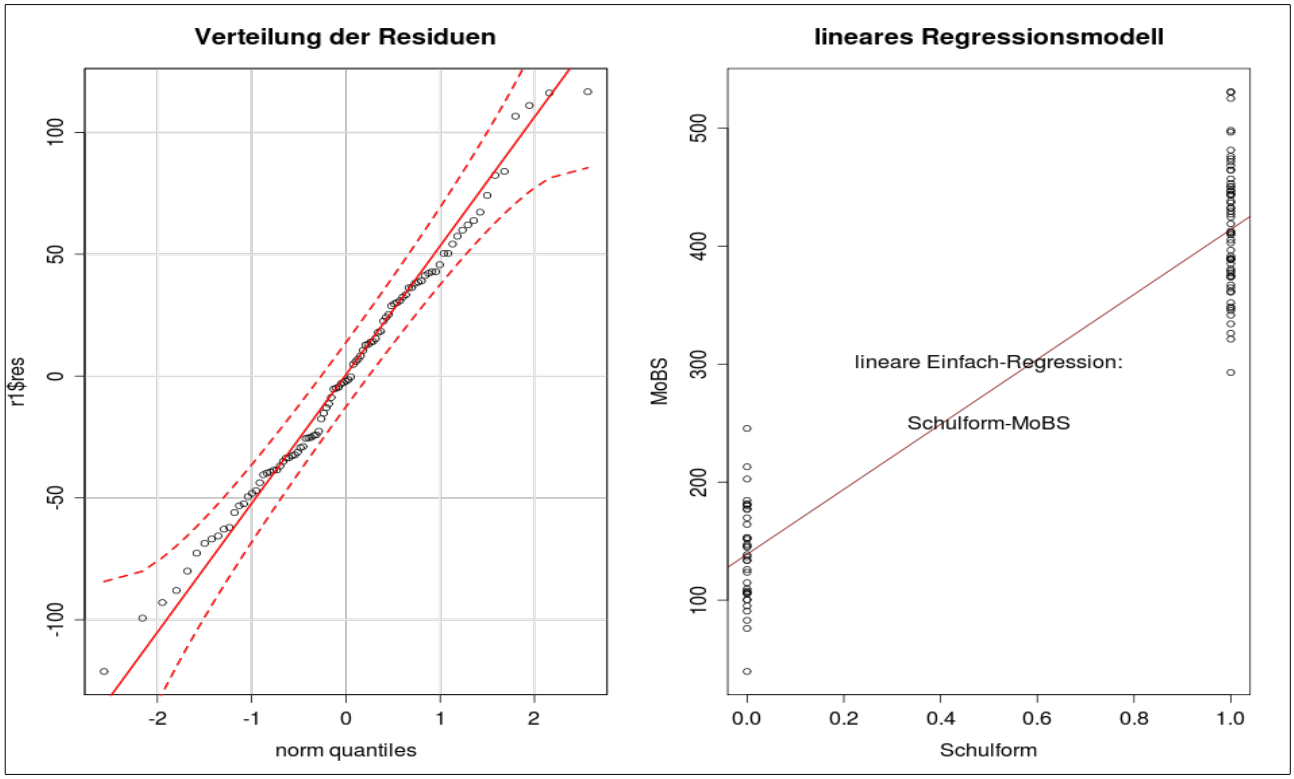

Abbildung 3: Regression 1

In Tabelle 2 lässt sich eine Erklärung dieses Modells sehen. Mittels „formula“, lässt sich noch einmal nachvollziehen welches Modell der Berechnung zu Grunde liegt. Bei den „Residuals“ handelt es sich um die in Abbildung 3 (links) bereits dargestellten Abweichung. Die berechneten Werte stimmen hierbei mit der Grafik überein. Aus dem Punkt „Coefficients“ lässt sich nun unsere lineare Funktionsgleichung ablesen. Unter $Pr(>|t|)$ lässt sich erkennen, dass die erhaltenen Werte signifikant sind. Aus dem linearen Modell ergibt sich nun ein $\alpha = 138.852$ und ein $\beta = 275.290$. Damit lässt sich das lineare Modell wie folgt formulieren: $Y_i = 138.852 + x_i \cdot 275.290$, wobei x_i der Schulform und Y_i dem Testergebnis des MoBS entspricht. Die Güte des Modells vermittelt das Bestimmtheitsmaß R^2. Als konservativeres Kriterium wähle ich das adjustierte Bestimmtheitsmaß. Es gibt an, dass das Modell $R^2_{adj} = 0.8775$ der Varianz zwischen abhängiger und unabhängiger Variable aufklären kann.

Des weiteren lässt sich nun untersuchen inwiefern ein zwei- oder mehrfaktorielles Regressionsmodell sinnvoll ist. Eine Möglichkeit zur Bewertung bietet hierfür das AIC (Groß,

2010, S. 214-218) nach Akaike, welches auf dem Maximum-Likelihood Ansatz beruht. Die Funktion $step()$ ermöglicht schrittweise eine Anpassung des Modells und eine Berechnung des jeweilig zugehörigen AIC's. Bei Ausführung erhalte ich folgende Funktion als Bestes Modell:

$$Step: AIC = 619.08 \\ dat\$ y \sim x1 + x4 + x3 + x2$$

. Die Frage die sich nun stellt ist, inwiefern ein solches Modell Sinn macht, da recht viele Faktoren verwendet werden und die AIC-Werte sich bei Zunahme der letzten drei Faktoren nur geringfügig ändern. Bei der Überprüfung der Voraussetzungen lässt sich außerdem feststellen, dass die Voraussetzungen für die Modelle bei denen das Alter und die Körpergröße zu Hilfe genommen wurde nicht erfüllt sind. Lediglich für das zweifaktorielle Regressionsmodell, indem der Wert des MoBS durch die Schulform und die wöchentliche Schlafdauer erklärt wird, sind die Voraussetzungen erfüllt.

```
Call:
lm(formula = dat$y ~ dat$x1 + dat$x4)

Residuals:
     Min       1Q   Median       3Q      Max
-94.52317 -23.69097   0.07415  23.99634 110.70521

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.9928    20.2247  -1.928   0.0569 .
dat$x1      277.9649     7.6492  36.339  < 2e-16 ***
dat$x4        3.4114     0.3701   9.217  8.5e-15 ***
---
Residual standard error: 36.37 on 94 degrees of freedom
Multiple R-squared: 0.9363,   Adjusted R-squared: 0.935
F-statistic:   691 on 2 and 94 DF,  p-value: < 2.2e-16
```

Tabelle 3: Regression Schulform - wöchentlicher Schlaf - MoBS

In Tabelle 3 sind die entsprechenden Daten dieses Modells dargestellt. Der Wert für den Achsenabschnitt wird dabei nicht signifikant, ich kann daher nicht davon ausgehen, dass er verscheiden von Null ist. Die beiden Beta-Koeffizienten hingegen weisen ein signifikantes Ergebnis auf. Damit lässt sich das Regressionsmodell wie folgt formulieren:

$Y_i = 277.9649 \cdot x1_i + 3.4114 \cdot x2_i$. Das adjustierte Bestimmtheitsmaß gibt eine Varianzaufklärung von $R^2_{adj} = 0.935$. Dieses Modell sorgt also für eine geringfügig bessere Aufklärung der Varianz, als das einfache lineare Regressionsmodell. Eine entsprechende Darstellung des Modells findet sich in Abbildung 4.

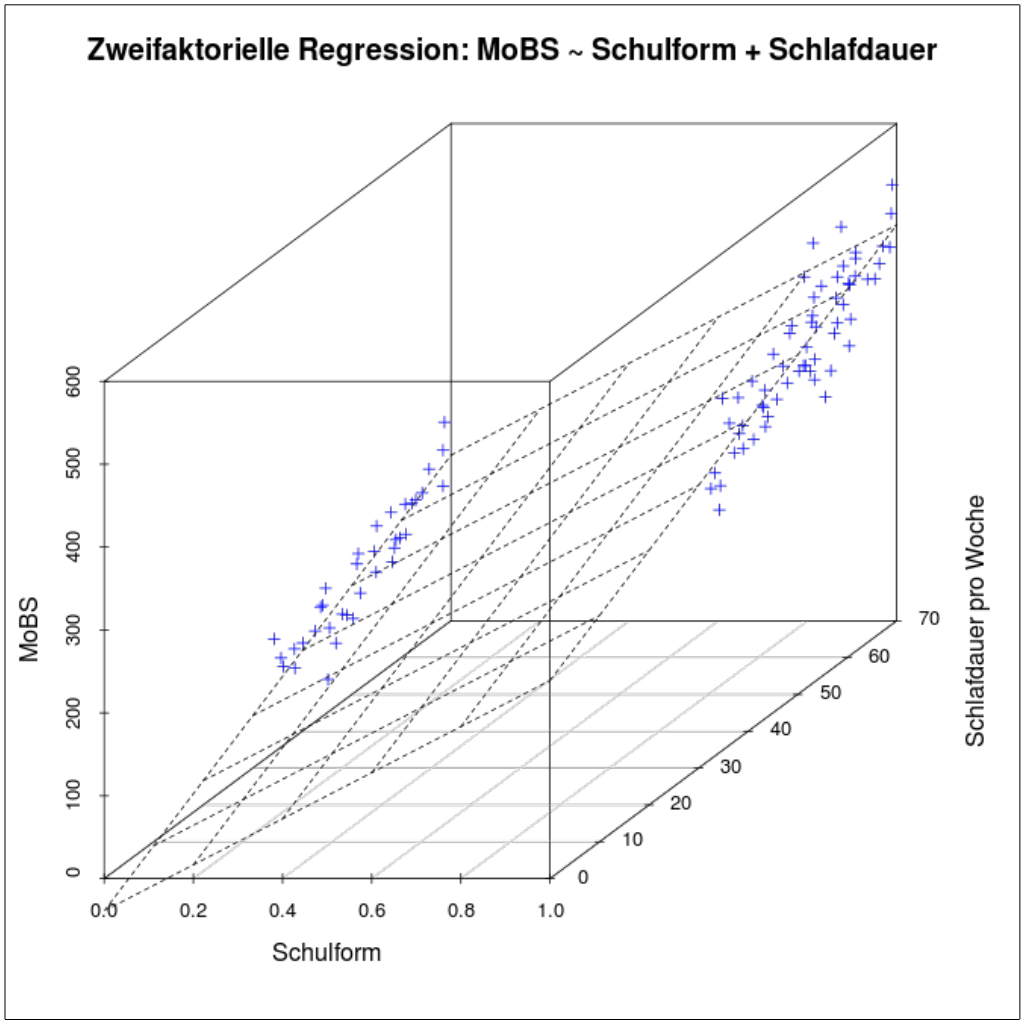

Abbildung 4: lineare Regression Schulform - Schlafdauer - MoBS

Zusammenfassend habe ich zwei mögliche Regressionsmodelle erhalten. In dem einfacheren der beiden zeigt sich, wie das Testergebnis im MoBS durch die Schulform ausgedrückt werden kann. Für unsere Untersuchung lässt sich damit die entsprechende Nullhypothese verwerfen und die Alternativhypothese annehmen. Zu beachten ist jedoch bei allen Modellen, dass sich keine Kausalität ergibt. Die Alternativhypothese könnte wie folgt lauten: Kinder von Schulen mit einem täglichen Bewegungsprogramm, haben bessere Testergebnisse im MoBS. Es lässt sich dabei nicht

feststellen ob das tägliche Bewegungsprogramm für bessere motorische Fähigkeiten sorgt oder ob beispielsweise Kinder mit besseren motorischen Fähigkeiten lieber auf eine Schule mit täglichen Bewegungsprogramm gehen, bzw. ihre Eltern sie dadurch fördern wollen.
Alle anderen Nullhypothesen zur linearen Einfachregression muss ich beibehalten, da die entsprechenden Modellvoraussetzungen nicht erfüllt waren.

Ich konnte eine Kombination von unabhängigen Variablen ergründen, so gibt es einen kumulativen Zusammenhang zwischen der Schulform und der wöchentlichen Schlafdauer auf der einen, sowie dem Testergebnis im MoBS auf der anderen Seite. Die Aussage die getroffen wird, könnte wie folgt lauten: Kinder von einer Schule mit einem täglichen Bewegungsprogramm und hoher wöchentlicher Schlafdauer, haben bessere Testergebnisse im MoBS. Auch hier gilt wie im einfachen Modell, dass keine Kausalität feststellbar ist. Dieses Modell ist komplexer und hat eine leicht bessere Vorhersage-kraft. Gewünscht ist natürlich ein möglichst große Vorhersage-kraft, wichtig ist aber auch die Verwendung eines möglichst einfachen Modells. Ich würde die Entscheidung für eines der beiden Modelle grundsätzlich von der Forschungsfrage abhängig machen, würde mich aber wahrscheinlich für das einfache lineare Regressionsmodell entscheiden, da das zweite meine Aussagekraft nur minimal verbessert.

Um weitere Peinlichkeiten bei zukünftigen olympischen Spielen zu vermeiden, würde ich dem Kultusministerium empfehlen, die bereits vorhandenen Ergebnisse noch einmal experimentell zu überprüfen. Sollten diese Studien meine Ergebnisse bestätigen, würde ich weiter empfehlen, zusätzliche Bewegungsprogramme an allen Schulen flächendeckend einzuführen.

Literatur

Groß, J. (2010). *Grundlegende Statistik mit R: Eine anwendungsorientierte Einführung in die Verwendung der Statistik Software R* (1. Aufl.). Vieweg+Teubner.

```
#Daten einlesen
setwd("/home/rick/Dokumente/Uni/Psychologie/Datenanalyse I/pruefung")
dat=read.table("DatReg32.txt", header=T)

#geladene Bibliotheken
library(Hmisc)
library(car)
library(gmodels)
library(gplots)
library(grDevices)

#deskriptive Statistik
summary(table(dat$y,dat$x1,dat$x2,dat$x3,dat$x4))
max(dat$vpnr)
sd(dat)
sum(dat$x1)



par(mfrow=c(2,2),family="HersheySerif", ps=14)
boxplot(dat$x2, main="Boxplot Alter - hinges")
boxplot(dat$x2, main="Boxplot Alter - Quartile",range=0)
boxplot(dat$x3, main="Boxplot Körpergröße")
boxplot(dat$x4, main="Boxplot Schlafdauer")
boxplot(dat$y[dat$x1==0],dat$y[dat$x1==1], xlab="staatlich vs. privat")

shapiro.test(dat$x2)
shapiro.test(dat$x3)
shapiro.test(dat$x4)
#Histogramme
par(mfrow=c(2,2),family="HersheySerif", ps=12)
hist(dat$x1, main="Schulform",xlab="staatlich vs. privat",border="black",col="navyblue",angle=45,
density=30, freq=F)
dat$x2->x
hist(x, main="Altersverteilung",xlab="Abweichung in Monaten vom 12.
Geburtstag",border="black",col=c("saddlebrown","navyblue","red3"), bg="grey",angle=45, density=30,
freq=F)
curve(dnorm(x, mean=mean(x), sd=sd(x)), add=T, col="darkred",lwd=2)
dat$x3->x
hist(x, main="Verteilung Körpergröße",xlab="Körpergröße in
cm",border="black",col=c("saddlebrown","navyblue","red3"),angle=45, density=30, freq=F)
curve(dnorm(x, mean=mean(x), sd=sd(x)), add=T, col="darkred",lwd=2)
dat$x4->x
hist(x, main="Verteilung Schlafdauer",xlab="wöchentliche Schlafdauer in
Stunden",border="black",col=c("saddlebrown","navyblue","red3"),angle=45, density=30, freq=F)



r1=lm(dat$y~dat$x1)
r2=lm(dat$y~dat$x2)
r3=lm(dat$y~dat$x3)
r4=lm(dat$y~dat$x4)

shapiro.test(r1$res) #erfüllt Voraussetzungen
shapiro.test(r2$res) #erfüllt Voraussetzungen nicht
shapiro.test(r3$res) #erfüllt Voraussetzungen nicht
shapiro.test(r4$res) #erfüllt Voraussetzungen nicht

par(mfrow=c(1,2), family="HersheySerif", ps=10)
qqPlot(r1$res, main="Verteilung der Residuen")
plot(dat$x1,dat$y, xlab="Schulform",pch=13,col="darkblue", ylab="MoBS", main="lineares
Regressionsmodell")
abline(r1, col="dark red");text(0.5,350,"lineare");text(0.5,300," Einfach-
Regression:");text(0.5,250,"Schulform-MoBS")


summary(r1)
r12=lm(dat$y~dat$x1+dat$x2)
```

```
r13=lm(dat$y~dat$x1+dat$x3)
r14=lm(dat$y~dat$x1+dat$x4)

library(scatterplot3d)
s3d<-scatterplot3d(dat$x1,dat$x4,dat$y)
s3d$plane3d(r14)

lm.all<-lm(dat$y~dat$x1+dat$x2+dat$x3+dat$x4)
AIC(lm.all)
step(lm.all,direction="both")

inde<-dat[,2:6]
lm.null<-lm(dat$y~1,data=inde)
lm.forw <- step(lm.null, scope=formula(inde),direction="both")

r143<-lm(dat$y~dat$x1+dat$x4+dat$x3)
shapiro.test(r143$res)
qqPlot(r143$res)

r1432<-lm(dat$y~dat$x1+dat$x2+dat$x3+dat$x4)
shapiro.test(r1432$res)
qqPlot(r1432$res)
```