Universität Stuttgart 11.11.2018
Institut für Sozialwissenschaften (SOWI IV)
Veranstaltung: Statistische Modellbildung II
Thomas Krause (M.A.)
Wintersemester 2018/19 Vanessa Schwaiger

Übungsaufgabe 2

1. Wozu werden Standardisierungen durchgeführt und wie wird dabei vorgegangen? Erläutern Sie zudem exemplarisch wozu b* benutzt wird und wie man diesen interpretiert!

Standardisierte Regressionskoeffizienten bezeichnen das Ausmaß der Veränderung einer abhängigen y-Variablen auf einer Standardskala. Durch die Standardisierung werden Skaleneffekte eliminiert und die Werte innerhalb eines Modells können vergleichbar gemacht werden (dieses Vorgehen ist jedoch auch umstritten, da es keine vollständige Unabhängigkeit von standardisierten Koeffizienten und empirischen Messeinheiten gibt).

b* ist ein standardisierter Regressionskoeffizient, der sich durch die Standardisierung des Regressionskoeffizienten b ergibt. Durch die Standardisierung liegt b* im Wertebereich zwischen -1 und +1. Die Formel hierfür sieht vor, die Standardabweichung von x (jeder Variable) durch die Standardabweichung von y zu dividieren und diesen Wert mit dem Regressionskoeffizient b zu multiplizieren. Bei bivariaten Zusammenhängen ist b* mit Pearsons r gleich zu setzen. Mit b* kann die einflussreichste Variable identifiziert werden. So können die Regressionskoeffizienten innerhalb eines Modells nur verglichen werden, wenn sie standardisiert sind, da die Standardisierung eine Betrachtung unabhängig der individuellen Skalenbreite darstellt. Eine inhaltliche Interpretation ist, wie beim Regressionskoeffizient b möglich (in Standardabweichungen), jedoch nicht unbedingt sinnvoll da die Angabe mit Standardabweichungen nicht „einfach“ verständlich und vorstellbar ist.

Multivariates Modell

	Modell 1		Modell 2		Modell 3	
	b	b*	b	b*	b	b*
Konstante	10,529		7,165		5,151	
Alter	**0,019*****	**0,068*****	**0,039*****	**0,135*****	**0,04*****	**0,14*****
Bildung			**1,199*****	**0,291*****	**1,244*****	**0,3*****
Geschlecht					**3,558*****	**0,359*****
R^2	0,0046		0,082		0,211	
Korr. R2 / Sig. Gesamtmodell	**0,00427*****		**0,081*****		**0,210*****	
Änderung in R2 Modellverbesserung			0,077		0,129	

Exemplarisch kann hier der Einfluss von Alter, Bildung und Geschlecht auf Einkommen herangezogen werden. In Modell 3 lässt sich anhand des standardisierten Koeffizienten b* erkennen, dass Geschlecht den stärksten (signifikanten) Einfluss hat. Der Regressionskoeffizient b kann hingegen nicht verglichen werden, da die individuelle Skalenbreite jeweils berücksichtigt werden muss.

2. Führen Sie eine z-Standardisierung für die Originalaltersvariable (alter_z) und die auf Null gesetzte Altersvariable (alter_0z) sowie für „unsere“ Bildungsvariable (0 bis 4). [Daten: ALLBUS 2014]
2.a) Vergleichen Sie die Zahlenwerte, Mean und die Standardabweichung von alter_z und alter_0z und erklären Sie Ihre „Beobachtung“.

	Alter_z	Alter_0z
Mean	0	0
Standardabweichung	1	1

Die Zahlenwerte liegen jeweils zwischen -1,8 und 2,37, haben also einen relativ kleinen Wertebereich (im Gegensatz zu den vorherigen Skalen) und viele Dezimalstellen. Diese z-Werte geben an, um wie viele Einheiten (Standardabweichungseinheiten) ein Messwert oberhalb oder unterhalb vom Durchschnitt/Mittelwert liegt.
Der Mean sowie die Standardabweichung weisen die gleichen Werte bei beiden z-standardisierten Variablen auf. Der Mean liegt auf 0 und die Standardabweichung bei 1. Dies liegt daran, dass die z-Werte eine standardisierte Skala bilden und immer einen Mittelwert von 0 und eine Standardabweichung von 1 aufweisen, ganz egal welche Skala die Variablen zuvor hatten. Durch die Standardisierung können einzelne Messwerte vergleichbar gemacht werden.

2.b) Führen Sie eine Regression von Einkommen auf Alter_0 und Bildung (Modell 1) und eine Regression von Einkommen auf alter_0z und bildung_z (Modell 2) durch und vergleichen Sie die b-Koeffizienten.
2.c) Wie erklären Sie die Werte b und b* in Modell 2? TIPP: Verwenden Sie bei Modell 2 das z-transformierte Einkommen als abhängige Variable.

	Modell 1 (Einkommen)		Modell 2 (Einkommen.z)	
	Alter	Bildung	Alter.z	Bildung.z
b	0,0387***	1,1986***	0,1367***	0,293***
b*	0,1353***	0,291***	0,1353***	0,291***

Die b-Koeffizienten in Modell 1 und 2 sind unterschiedlich. Während in Modell 1 das Einkommen durch die Zunahme um ein Lebensjahr um 0,0387 Einheiten im Einkommen steigt, sind es im Modell 2 0,1367 Einheiten. Weil die Skalen unterschiedlich sind, können diese Werte jedoch nicht miteinander verglichen werden. In beiden Modellen hat die Bildung einen stärkeren Einfluss auf Einkommen als das Alter. Auffällig ist, dass die b-Koeffizienten von Modell 2 (nahezu) identisch mit den beta-Koeffizienten von Modell 1 sowie den beta-Koeffizienten von Modell 2 sind.
Der b*-Wert in Modell 1 gibt an, wie viele Standardabweichungen das Einkommen zunimmt, wenn das Alter/die Bildung um 1 Standardabweichung zunimmt. Da Standardabweichungen die Maßeinheit für z-Werte sind, stellt die Steigung (der b-Wert) in Modell 2 denselben Wert wie b* in Modell 1 dar. Dieser Wert verändert sich nicht durch die Berechnung des standardisierten Koeffizienten b* in Modell 2, da die Steigung ohnehin auf einer standardisierten Skala angegeben ist. Für die Berechnung von b* werden die Standardabweichungen von x und y dividiert (beides mal 1) sowie mit dem b-Koeffizienten multipliziert: (1/1)*0,1367=0,1367 (bzw. aufgrund Rundungsfehlern in der oben gezeigten Tabelle 0,1353).

3. Erstellen Sie ein multivariates Regressionsmodell mit Y=Einkommen. Versuchen Sie dabei den R^2-Wert so groß wie nur irgendwie möglich zu bekommen. Jeder schmutzige Trick der Sozialforschung ist erlaubt (und in diesem Fall erwünscht). Fügen Sie die entsprechenden Teile des SPSS-Outputs in Ihre Abgabe ein.

Einzige Einschränkung: Keine Regression von Y auf Y.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.1306442  1.5968791   3.213 0.001837 **
age         -0.0047746  0.0132818  -0.359 0.720096
education    0.2397835  0.1669818   1.436 0.154552
sex          0.1291530  0.2918565   0.443 0.659197
V425         0.3882952  0.2770093   1.402 0.164510
V128        -0.0030138  0.0016373  -1.841 0.069038 .
V70         -0.0863541  0.0748825  -1.153 0.251953
V38          0.5896529  0.3116573   1.892 0.061781 .
V71         -0.0041065  0.0018375  -2.235 0.027963 *
V727        -0.1053625  0.1143282  -0.922 0.359268
V730        -0.0665668  0.0849014  -0.784 0.435117
V521        -0.3035296  0.3108404  -0.976 0.331503
V96          0.3874296  0.4345679   0.892 0.375077
V95          0.1438899  0.4955530   0.290 0.772222
V417         0.0048021  0.0002262  21.226  < 2e-16 ***
V491        -0.0005347  0.0001526  -3.504 0.000723 ***
Schicht      0.3683301  0.2796234   1.317 0.191179
FDP          1.1702492  0.7585327   1.543 0.126473
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.256 on 88 degrees of freedom
  (3365 observations deleted due to missingness)
Multiple R-squared:  0.9235,    Adjusted R-squared:  0.9087
F-statistic: 62.45 on 17 and 88 DF,  p-value: < 2.2e-16
```

Es wurden möglichst viele Variablen hinzugefügt. Von diesen sind lediglich zwei hochsignifikant, trotzdem liegt das R^2 bei 0,9235. Die Variablen wurden anhand inhaltlicher Nähe aber auch Plausibilität ausgewählt.

Anhang R-Code

```
#install.packages("foreign")
library("foreign")
#install.packages("survey")
library("survey")

allbus<-read.spss ("C:/Users/Vanes/Desktop/Allbus.sav",
                   to.data.frame=T, use.value.labels = FALSE,reencode=T)

#relevante Variablen:
#V84   ALTER: BEFRAGTE<R>
#V86   ALLGEMEINER SCHULABSCHLUSS
#V81   GESCHLECHT, BEFRAGTE<R>
#V420 NETTOEINKOMMEN<OFFENE+LISTENANGABE>,KAT.


#Alter Befragter v84
table(allbus$V84)
allbus$age<-allbus$V84-18
allbus$age
```

```
#Geschlecht 0 =weiblich 1 =männlich
#Geschlecht v81 bis jetzt: 1= männlich 2 =weiblich
allbus$V81
allbus$sex<-allbus$V81
allbus$sex[allbus$V81==2]<-0
allbus$sex

#Schulabschluss v86
#5 Ausprägungen; 0=kein Schulabschluss, 1=HS, 2=RS, 3=FHR, 4=Abi; Rest=-1 bzw.
Missing --> soll sein
allbus$V86

#allbus$education = ifelse(allbus$V86 == 6 | allbus$V86 == 7, NA, allbus$education - 1)
allbus$education<-allbus$V86-1
allbus$education[allbus$education>=5]<- NA
allbus$education

#Einkommen: V420
#z-Standardisierung
allbus$age.z<-(allbus$age-mean(allbus$age,na.rm=T))/sd(allbus$age,na.rm=T)
allbus$V84.z<-(allbus$V84-mean(allbus$V84,na.rm=T))/sd(allbus$V84,na.rm=T)
describe(allbus$age.z)
describe(allbus$V84.z)
allbus$age.z
allbus$V84.z

allbus$education.z<-(allbus$education-
mean(allbus$education,na.rm=T))/sd(allbus$education,na.rm=T)
allbus$V420.z<-(allbus$V420-mean(allbus$V420,na.rm=T))/sd(allbus$V420,na.rm=T)

#Regression
library("lm.beta")
fit1<-lm(V420~age + education, data=allbus)
summary(fit1)
lm.beta(fit1)

#mit standardisiertem Einkommen
fit3 <- lm(V420.z ~age.z+ education.z, data=allbus)
summary(fit3)
lm.beta(fit3)

#Aufgabe 3:
#FDP-Wähler
allbus$FDP<-ifelse(allbus$V729==3,1,0)
fit<-lm(V420~age + education + sex + V425+ V128 + V70 + V38 + V71+ V727 + V730 +
V521+ V96 + V95+ V417 + V491 + Schicht+ FDP, data=allbus)
summary(fit)
lm.beta(fit)
```