

## 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!

Standardisierungen von Regressionskoeffizienten werden durchgeführt, um die Stärke der Koeffizienten innerhalb eines Regressionsmodells untereinander zu vergleichen, unabhängig von der Messeinheit/Messskala. Der Vergleich zwischen Modellen sowie eine direkte Interpretation oder eine Prognose für die abhängige Variable sind mit standardisierten Koeffizienten jedoch nicht möglich. Dabei wird folgendermaßen vorgegangen: der unstandardisierte Regressionskoeffizient ( $b$ ) einer unabhängigen Variable ( $x$ ) wird mit der Standardabweichung der Variable ( $s_x$ ) multipliziert und durch die Standardabweichung der abhängigen Variable ( $s_y$ ) geteilt (siehe (1)).

$$b^* = b * \frac{s_x}{s_y} \quad (1)$$

Beispiel: in einem Regressionsmodell mit den (metrischen) Variablen Alter und Einkommen kann nur mit unstandardisierten partiellen Regressionskoeffizienten nicht bestimmt werden, welche Variable einen stärkeren Einfluss auf die abhängige Variable hat, da die unabhängigen Variablen in verschiedenen Einheiten gemessen werden und je nach Skalierung auch unterschiedlich hoch ausfallen können. Bei standardisierten Koeffizienten besteht dieses Problem nicht und es kann eine Aussage darüber getroffen werden, ob das Alter oder das Einkommen einen stärkeren Einfluss auf die abhängige Variable hat.

### 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]

```
#Datensatz einlesen
allbus <- read_sav("C:/Users/Anna/Documents/Uni S/Statistische-Modellbildung-II-WS1819-master/data/allbus2014.sav")
#Subdatensatz erstellen
allb_sub <- select(allbus, v84, v86, v81, v420)

#Relevante Variablen raussuchen und umbenennen
allb_sub <- rename(allb_sub, alter = v84,
  bildung = v86,
  geschl = v81,
  einkommen = v420)
#Niedrigstes Alter auf 0 setzen, Bildung aufbereiten mit 6 und 7 als missings, Geschlecht
dummyscodieren
allb_sub <- mutate(allb_sub, alter0 = alter - 18,
  bildung_rec = ifelse(bildung == 6 | bildung == 7, NA, bildung - 1),
  geschl_rec = ifelse(geschl == 2, 0, 1))

# Subdatensatz erstellen für z-standardisierte Variablen und Variablen auswählen
allb_sub_z <- allb_sub
  select(einkommen, alter, alter0, geschl_rec, bildung_rec)
# Variablen z-standardisieren
allb_sub_z <- mutate(allb_sub_z, alter_z = scale(alter),
  alter0_z = scale(alter0),
  bildung_z = scale(bildung_rec),
  einkommen_z = scale(einkommen))
```

**2.a) Vergleichen Sie die Zahlenwerte, Mean und die Standardabweichung von alter\_z und alter0\_z und erklären Sie Ihre „Beobachtung“.**

```
allb_sub_z <- select(allb_sub_z, alter_z, alter0_z)
descr(allb_sub_z)
```

```
## Basic descriptive statistics
```

var	type	label	n	NA.prc	mean	sd	se	md	trimmed	range	skew
alter_z	numeric	alter_z	3468	0.09	0	1	0.02	0.03	-0.01	4.17	(-1.8-2.37) 0.06
alter0_z	numeric	alter0_z	3468	0.09	0	1	0.02	0.03	-0.01	4.17	(-1.8-2.37) 0.06

allb_sub_z		
	alter_z	alter0_z
1	-0.93911214	-0.93911214
2	0.03197083	0.03197083
3	0.37470599	0.37470599
4	0.66031863	0.66031863
5	0.54607357	0.54607357
6	0.37470599	0.37470599
7	0.94593127	0.94593127
8	-1.39609236	-1.39609236
9	0.48895105	0.48895105
10	0.20333841	0.20333841
11	0.94593127	0.94593127
12	-0.25364181	-0.25364181
13	0.08909335	0.08909335

Betrachtet man den Datensatz in R (siehe Tabelle links), sieht man, dass die Zahlenwerte von alter\_z und alter0\_z identisch sind. Ebenso verhält es sich mit dem arithmetischen Mittel und der Standardabweichung (gelb hervorgehoben im Output oben).

Daran kann man erkennen, dass es für die Standardisierung keine Rolle spielt, in welcher Einheit oder auf welcher Skala eine Variable, hier das Alter, gemessen wurde: man erhält das gleiche Ergebnis für das Alter, das mit den „wahren“ Alterswerten der Beobachtungen beschrieben wird und für das Alter0, wo von allen „wahren“ Alterswerten der Wert 18 abgezogen wurde.

**2.b) Führen Sie eine Regression von Einkommen auf Alter\_0 und Bildung (Modell 1) und eine Regression von Einkommen auf alter0\_z und bildung\_z (Modell 2) durch und vergleichen Sie die b-Koeffizienten.**

```
#Datensatz mit allen benötigten Variablen erstellen
allb_sub_z <- select(allb_sub, alter0, einkommen, bildung_rec)
allb_sub_z <- mutate(allb_sub_z,
  alter0_z = scale(alter0),
  bildung_z = scale(bildung_rec),
  einkommen_z = scale(einkommen))

# Modelle bilden
mod2 <- lm(einkommen ~alter0 + bildung_rec, data = allb_sub_z)
mod3 <- lm(einkommen_z ~alter0_z + bildung_z, data = allb_sub_z)

#Beide Modelle in einer Tabelle darstellen
texreg(list(mod2, mod3))
```

Tabelle 1: Einflüsse auf Einkommen

	Modell 1 (unstandardisiert)	Modell 2 (standardisiert)
(Intercept)	7.17 *** (0.28)	0 (0.02)
Alter0	0.04 *** (0.01)	0.14 *** (0.02)
Bildung	1.20 *** (0.07)	0.29 *** (0.02)
R <sup>2</sup>	0.08	0.08
Adj. R <sup>2</sup>	0.08	0.08
n	3039	3039
RMSE	4.74	0.96

\*\*\* p<0.001, \*\* p<0.01, \*p<0.05

Während die b-Koeffizienten in Modell 1 eine Interpretation ermöglichen, sind die b\*-Koeffizienten in Modell 2 nur untereinander vergleichbar. Die standardisierten Werte kommen sich „näher“, da sie immer zwischen -1 und 1 liegen, während die unstandardisierten Werte auch außerhalb liegen, was man am unstandardisierten Wert für Bildung, 1,2 erkennt.

## 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.

Die b\*-Werte in Modell 2 sagen aus, dass der Einfluss von Bildung auf Einkommen etwas höher ist als der von Alter auf Einkommen, bzw. dass die Erklärkraft von Bildung bezüglich der Varianz von Einkommen etwas höher ist als die von Alter. Das sieht man daran, dass der Wert 0,29 (Bildung) höher ist als der Wert 0,14 (Alter).

(Keine Werte von b in Modell 2 vorhanden, da sie in Modell 1 sind?)

## 3. Erstellen Sie ein multivariates Regressionsmodell mit Y=Einkommen. Versuchen Sie dabei den R<sup>2</sup>-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.**

#Strategien für großes R<sup>2</sup> (die normalerweise berücksichtigt/vermieden werden sollten):

- #1) Viele UVs
- #2) UVs, die kausal nachgeordnet sind
- #3) UVs mit inhaltlicher Nähe zu AV
- #4) Beobachtungen, die linearer Beziehung widersprechen, ausschließen
- #5) X-Variablen mit höchster Varianz
- #6) AV mit kleinster Residualvarianz

#1), 2), 3) und 5):

# Einkommen (Nettoeinkommen v417). Einflüsse: Geschlecht v81, Alter (metrisch gemessen) v84, Schulabschluss v86, Berufsstatus ISCED 2011 v102, Arbeitsstunden pro Woche v118, Geburt in D v377, Erhebungsgebiet westost v7

# Nachgeordnete UVs/inhaltliche Nähe (vermutet): Zahl der Bücher im Haushalt v524, Wohnfläche in m<sup>2</sup> v594, Sprechanlage v843

```
allbus14 <- read_sav("C:/Users/Anna/Documents/Uni S/Statistische-Modellbildung-II-WS1819-master/data/allbus2014.sav")
```

```

allbus14 <- rename(allbus14, Einkommen=V417, Weiblich=V81, Alter=V84, Abschluss=V86, IS-
CED2011=V102, hArbeit=V118, GeburtD = V377, Osttdt=V7, Buecher=V524, Flaeche=V594, Sprech=V843)
allbus14_s <- select(allbus14, Einkommen, Weiblich, Alter, Abschluss, ISCED2011, hArbeit, Ge-
burtD, Osttdt, Buecher, Flaeche, Sprech, Einkommen2)

# Beobachtungen mit missings löschen
allbus14_s<-allbus14_s[!(allbus14_s$Einkommen==99997 | allbus14_s$Einkommen==99990),]
allbus14_s<-allbus14_s[!(allbus14_s$Alter==999),]
allbus14_s<-allbus14_s[!(allbus14_s$Abschluss==6 | allbus14_s$Abschluss==7 | allbus14_s$Ab-
schluss==99),]
allbus14_s<-allbus14_s[!(allbus14_s$ISCED2011==94 | allbus14_s$ISCED2011==99),]
allbus14_s<-allbus14_s[!(allbus14_s$hArbeit==999.6 | allbus14_s$hArbeit==999.9),]
allbus14_s<-allbus14_s[!(allbus14_s$GeburtD==9),]
allbus14_s<-allbus14_s[!(allbus14_s$Osttdt==9),]
allbus14_s<-allbus14_s[!(allbus14_s$Buecher==98 | allbus14_s$Buecher==99),]
allbus14_s<-allbus14_s[!(allbus14_s$Flaeche==9998 | allbus14_s$Flaeche==9999),]
allbus14_s<-allbus14_s[!(allbus14_s$Sprech==8),]
allbus14_s<- allbus14_s[complete.cases(allbus14_s), ]

# Variablen neu codieren (Dummies)
allbus14_s <- mutate (allbus14_s, Weiblich = ifelse(Weiblich == 2, 1, 0),
                      GeburtD = ifelse(GeburtD ==1, 1, 0),
                      Osttdt= ifelse(Osttdt == 2, 1, 0),
                      Sprech = ifelse(Sprech == 1, 1, 0))

#Regressionsmodell
mod4 <- lm(Einkommen ~ Weiblich + Alter + Abschluss + ISCED2011 + hArbeit + GeburtD + Osttdt +
Buecher + Flaeche + Sprech, data = allbus14_s)
texreg(list(mod4))

#R² ist 0,17.

\hline
& Model 1 \\\
\hline
(Intercept) & $-1516.08^{***}$ \\\
& $(322.54)$ \\\
Weiblich & $-585.75^{***}$ \\\
& $(99.59)$ \\\
Alter & $18.31^{***}$ \\\
& $(3.95)$ \\\
Abschluss & $35.49$ \\\
& $(59.13)$ \\\
ISCED2011 & $164.84^{***}$ \\\
& $(38.06)$ \\\
hArbeit & $31.64^{***}$ \\\
& $(4.51)$ \\\
GeburtD & $110.84$ \\\
& $(144.67)$ \\\
Osttdt & $-310.40^{**}$ \\\
& $(100.11)$ \\\
Buecher & $99.58^{**}$ \\\
& $(32.35)$ \\\
Flaeche & $3.04^{***}$ \\\
& $(0.81)$ \\\
Sprech & $79.21$ \\\
& $(90.94)$ \\\
\hline
R$^2$ & 0.17 \\\
Adj. R$^2$ & 0.17 \\\
Num. obs. & 1529 \\\
RMSE & 1733.24 \\\
\hline
\multicolumn{2}{l}{\scriptsize{$^{***}$p<0.001$, $^{**}$p<0.01$, $^*$p<0.05$}}

```

```

#4): Lineare Beziehung zwischen einzelnen Variablen anschauen
plot (Einkommen~Alter, data=allbus14_s)
plot (Einkommen~Abschluss, data=allbus14_s)
plot (Einkommen~hArbeit, data=allbus14_s) #Notiz: alle über 70h löschen, die verdienen wieder
weniger?
plot (Einkommen~Buecher, data=allbus14_s)
plot (Einkommen~Flaeche, data=allbus14_s) #Notiz: alle über 300m² löschen, die verdienen weni-
ger?

plot (Einkommen~Weiblich, data=allbus14_s)
plot (Einkommen~GeburtD, data=allbus14_s)
plot (Einkommen~Ostdt, data=allbus14_s)
plot (Einkommen~Sprech, data=allbus14_s)
# Dummies sagen nicht so viel über Linearität

#Ausreißer löschen, auch von AV (#6)
allbus14_s<-allbus14_s[!(allbus14_s$Einkommen>30000),]
allbus14_s<-allbus14_s[!(allbus14_s$hArbeit>70),]
allbus14_s<-allbus14_s[!(allbus14_s$Flaeche>300),]
allbus14_s<- allbus14_s[complete.cases(allbus14_s), ]

#Nochmal Modell probieren
mod5 <- lm(Einkommen ~ Weiblich + Alter + Abschluss + ISCED2011 + hArbeit + GeburtD + Ostdt +
Buecher + Flaeche + Sprech, data = allbus14_s)
texreg(list(mod5))

#R² bei 0,4!

\hline
& Model 1 \\\
\hline
(Intercept) & $-1326.56^{***}$ \\\
& $(175.64)$ \\\
Weiblich & $-521.44^{***}$ \\\
& $(53.13)$ \\\
Alter & $17.37^{***}$ \\\
& $(2.09)$ \\\
Abschluss & $33.19$ \\\
& $(31.28)$ \\\
ISCED2011 & $208.99^{***}$ \\\
& $(20.14)$ \\\
hArbeit & $28.99^{***}$ \\\
& $(2.55)$ \\\
GeburtD & $178.71^{*}$ \\\
& $(76.43)$ \\\
Ostdt & $-456.11^{***}$ \\\
& $(53.15)$ \\\
Buecher & $41.22^{*}$ \\\
& $(17.23)$ \\\
Flaeche & $2.30^{***}$ \\\
& $(0.51)$ \\\
Sprech & $120.06^{*}$ \\\
& $(48.18)$ \\\
\hline
R$^2$ & 0.40 \\\
Adj. R$^2$ & 0.40 \\\
Num. obs. & 1507 \\\
RMSE & 911.20 \\\
\hline
\multicolumn{2}{l}{\scriptsize{$^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$}}

#Schauen, ob kategorisiertes Einkommen besser ist: (alles neu weil kategorisiertes und direk-
tes Einkommen sich ausschließen wegen missings)

allbus14 <- read_sav("C:/Users/Anna/Documents/Uni S/Statistische-Modellbildung-II-WS1819-mas-
ter/data/allbus2014.sav")
allbus14 <- rename(allbus14, Weiblich=V81, Alter=V84, Abschluss=V86, ISCED2011=V102, hAr-
beit=V118, GeburtD= V377, Ostdt=V7, Buecher=V524, Flaeche=V594, Sprech=V843, Einkommen2=V418)
allbus14_s <- select(allbus14, Weiblich, Alter, Abschluss, ISCED2011, hArbeit, GeburtD, Ostdt,
Buecher, Flaeche, Sprech, Einkommen2)

```

```
# Beobachtungen mit missings löschen
allbus14_s<-allbus14_s[!(allbus14_s$Einkommen2==95 | allbus14_s$Einkommen2==99),]
allbus14_s<-allbus14_s[!(allbus14_s$Alter==999),]
allbus14_s<-allbus14_s[!(allbus14_s$Abschluss==6 | allbus14_s$Abschluss==7 | allbus14_s$Ab-
schluss==99),]
allbus14_s<-allbus14_s[!(allbus14_s$ISCED2011==94 | allbus14_s$ISCED2011==99),]
allbus14_s<-allbus14_s[!(allbus14_s$hArbeit==999.6 | allbus14_s$hArbeit==999.9),]
allbus14_s<-allbus14_s[!(allbus14_s$GeburtD==9),]
allbus14_s<-allbus14_s[!(allbus14_s$Ostdt==9),]
allbus14_s<-allbus14_s[!(allbus14_s$Buecher==98 | allbus14_s$Buecher==99),]
allbus14_s<-allbus14_s[!(allbus14_s$Flaeche==9998 | allbus14_s$Flaeche==9999),]
allbus14_s<-allbus14_s[!(allbus14_s$Sprech==8),]
allbus14_s<- allbus14_s[complete.cases(allbus14_s), ]

# Variablen neu codieren (Dummies)
allbus14_s <- mutate (allbus14_s, Weiblich = ifelse(Weiblich == 2, 1, 0),
                     GeburtD = ifelse(GeburtD ==1, 1, 0),
                     Ostdt= ifelse(Ostdt == 2, 1, 0),
                     Sprech = ifelse(Sprech == 1, 1, 0))
mod6 <- lm(Einkommen2 ~ Weiblich + Alter + Abschluss + ISCED2011 + hArbeit + GeburtD + Ostdt +
Buecher + Flaeche + Sprech, data = allbus14_s)
texreg(list(mod6))
```

```
\hline
& Model 1 \\
\hline
(Intercept) & $3.42$ \\
& $(1.89)$ \\
Weiblich & $-2.38^{***}$ \\
& $(0.58)$ \\
Alter & $0.02$ \\
& $(0.03)$ \\
Abschluss & $0.77^{*}$ \\
& $(0.38)$ \\
ISCED2011 & $0.67^{**}$ \\
& $(0.23)$ \\
hArbeit & $0.08^{***}$ \\
& $(0.02)$ \\
GeburtD & $2.22^{*}$ \\
& $(0.93)$ \\
Ostdt & $-3.18^{***}$ \\
& $(0.58)$ \\
Buecher & $0.00$ \\
& $(0.19)$ \\
Flaeche & $0.00$ \\
& $(0.01)$ \\
Sprech & $-0.17$ \\
& $(0.54)$ \\
\hline
R$^2$ & 0.42 \\
Adj. R$^2$ & 0.39 \\
Num. obs. & 209 \\
RMSE & 3.75 \\
\hline
\multicolumn{2}{l}{\scriptsize{$^{***}$p<0.001$, $^{**}$p<0.01$, $^{*}$p<0.05$}}
#R² ist 0,42.
```