Musterlösung: Übungsaufgabe 2

Fabio Votta

14.November 2018

Aufgabe 1

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

$$b^* = b * \frac{s_x}{s_y}$$

Durch die Verwendung der absoluten (also unstandardisierten) Werte hängt der nummerische Zahlenwert eines Zusammenhangs von der Skalierung der Variablen ab. Die Standardisierung dient dazu, trotz unterschiedlicher Skalen eine Aussage über die relative Stärke des Einflusses treffen zu können.

Um b^* zu erhalten wird die Kovarianz durch die Standardabweichung von X und Y dividiert wird, sodass der Wert zwischen -1 und 1 rangiert. b^* wird benutzt, um die Stärke des Effektes von im selben Modell enthaltenen Prädiktoren zu vergleichen. Es kann somit der einflussreichste Zusammenhang identifiziert werden. Interpretiert wird dieser wie folgt: Wenn X um eine Standardabweichung steigt, steigt y um b^* Standardabweichungen. Zwischen mehreren Variablen im gleiche Modell zeigen größere (bzw. weiter von Null entfernte) Werte einen stärkeren Einfluss an.

Aufgabe 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]

```
allb_sub_z <- allb_sub %>%
  select(einkommen, alter, alter0, geschl_rec, bildung_rec) %>%
  mutate(alter_z = scale(alter),
         alter0_z = scale(alter0),
         bildung_z = scale(bildung_rec),
         einkommen_z = scale(einkommen))
```

Aufgabe 2a

Vergleichen Sie die Zahlenwerte, Mean und die Standardabweichung von alter$_z$ und alter_0z und erklären Sie Ihre "Beobachtung".

```
allb_sub_z %>%
  select(alter_z, alter0_z) %>%
  describe() %>%
  select(-vars, -range, -trimmed, -mad,  -skew, -kurtosis,  -se) %>%
  kable()
```

	n	mean	sd	median	min	max
alter_z	3468	0	1	0.0319708	-1.79595	2.373994
alter0_z	3468	0	1	0.0319708	-1.79595	2.373994

Die Zahlenwerte, Mean und Standardabweichung von alter_z und alter0_z sind aufgrund der Standardisierung gleich ungeachtet dessen, dass sie davor eine andere Skalenbreite hatten.

Aufgabe 2b

Führen Sie eine Regression von Einkommen auf alter_0 und bildung (Modell 1) und eine Regression von einkommen_z auf alter_0z und bildung_z (Modell 2) durch und vergleichen Sie die b-Koeffizienten.

```
mod1 <- lm(einkommen ~ alter0 + bildung_rec, data = allb_sub_z)
mod2 <- lm(einkommen_z ~ alter0_z + bildung_z, data = allb_sub_z)

tbl_std(mod2, type = "latex")
```

Table 2:

	Dependent variable:	
	einkommen_z	
	b	std.b
	(1)	(2)
alter0_z	0.137***	0.135***
	(0.018)	(0.018)
bildung_z	0.293***	0.291***
	(0.018)	(0.018)
Constant	0.005	0.000
	(0.017)	(0.017)
Observations	3,039	3,039
R^2	0.082	0.082
Adjusted R^2	0.081	0.081
Residual Std. Error (df = 3036)	0.955	0.955
F Statistic (df = 2; 3036)	135.439***	135.439***
Note:	*p<0.1; **p<0.05; ***p<0.01	

```
texreg(list(mod1, mod2), float.pos = "ht!")
```

	Model 1	Model 2
(Intercept)	7.17***	0.00
	(0.28)	(0.02)
alter0	0.04***	
	(0.01)	
bildung_rec	1.20***	
	(0.07)	
alter0_z		0.14***
		(0.02)
bildung_z		0.29***
		(0.02)
R^2	0.08	0.08
Adj. R^2	0.08	0.08
Num. obs.	3039	3039
RMSE	4.74	0.96

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Statistical models

```
summary(lm.beta(mod1))
```

```
##
## Call:
## lm(formula = einkommen ~ alter0 + bildung_rec, data = allb_sub_z)
##
## Residuals:
## <Labelled double>
##      Min       1Q   Median       3Q      Max
## -12.0444  -3.6715   0.1664   3.6227  11.8613
##
## Labels:
##  value              label
##      0     KEIN EINKOMMEN
##      1     UNTER 200 EURO
##      2     200 - 299 EURO
##      3     300 - 399 EURO
##      4     400 - 499 EURO
##      5     500 - 624 EURO
##      6     625 - 749 EURO
##      7     750 - 874 EURO
##      8     875 - 999 EURO
##      9   1000 - 1124 EURO
##     10   1125 - 1249 EURO
##     11   1250 - 1374 EURO
##     12   1375 - 1499 EURO
##     13   1500 - 1749 EURO
##     14   1750 - 1999 EURO
##     15   2000 - 2249 EURO
##     16   2250 - 2499 EURO
##     17   2500 - 2749 EURO
##     18   2750 - 2999 EURO
##     19   3000 - 3999 EURO
```

```
##     20   4000 - 4999 EURO
##     21   5000 - 7499 EURO
##     22 7500 EURO UND MEHR
##     97         VERWEIGERT
##     99       KEINE ANGABE
##
## Coefficients:
##             Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept) 7.165316     0.000000   0.282215  25.390  < 2e-16 ***
## alter0      0.038743     0.135298   0.005168   7.496 8.56e-14 ***
## bildung_rec 1.198558     0.291016   0.074333  16.124  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.741 on 3036 degrees of freedom
##   (432 observations deleted due to missingness)
## Multiple R-squared:  0.08191,    Adjusted R-squared:  0.08131
## F-statistic: 135.4 on 2 and 3036 DF,  p-value: < 2.2e-16
```

Bei Modell 1 hat das unstandardisierte Alter einen sehr geringen positiven Einfluss von 0,039 Einheiten auf das Einkommen. Die Bildungsvariable dafür einen positiven Einfluss von 1,199. Bei Modell 2 hat sich das deutlich verändert. Die standardisierte alter0 Variable hat einen positiven Einfluss von 0,135 Standardeinheiten auf das standardisierte Einkommen. Die Bildung hier nun mit einem positiven Einfluss von 0,291 Standardeinheiten. So lässt sich nun sehen, dass Bildung einen fast doppelt so starken Einfluss auf das Einkommen ausübt als das Alter.

Aufgabe 2c

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 und b* Werte sind im Modell mit dem standardisierten Einkommen (beniahe) identisch. Das lässt darauf schließen lässt, dass die Standardisierung der UVs und AV im vorhinein dazu führt, dass auch der unstandardisierte Wert als standardisierter Koeffizient zu interpretieren wäre.

Aufgabe 3

Erstellen Sie ein multivariates Regressionsmodell mit Y=Einkommen. Versuchen Sie dabei den R2-Wert so groß wie nur irgendwie möglich zu bekommen. Jeder schmutzige Trick der Sozialforschung ist erlaubt (und in diesem Fall erwünscht).

- *Einzige Einschränkung: Keine Regression von Y auf Y.*

```r
allb_r <- allbus %>%
  select(V84, V86, V81, V420, V98, V118, V269, V103, V7, V13, V14,
       V16, V20, V21, V22, V25, V30,V31, V494, V9, V209, V279,
       V71, V711, V216, V215, V495, V513, V514, V377) %>%
  rename(alter = V84,
         bildung = V86,
         geschl = V81,
         einkommen = V420,
         arbeitsstd = V118,
         keineberufsausbildung = V98,
         beruf = V103,
```

```
          westost = V7,
          internet = V14,
          computer = V16,
          essen = V20,
          besuchfreunde = V21,
          besuchfamilie = V22,
          kunst = V25,
          theater = V30,
          museum = V31,
          haushaltseinkommen = V494,
          wirtschaftslage = V9,
          fernsehenmin = V71,
          dauerbildung = V711,
          demzufr = V216,
          linksrechts = V215,
          prokopfeink = V495 ,
          krankengeldhh = V513,
          elterngeldhh = V514,
          gebd = V377) %>%
  na.omit() %>%
  mutate(alter0 =alter -18,
           alter0quad =alter0*alter0,
           bildung_rec = ifelse(bildung ==6|bildung ==7,0, bildung -1),
           geschl_rec = ifelse(geschl ==2,0,1),
           ganztags = ifelse(beruf ==1,1,0),
           halbtags = ifelse(beruf ==2,1,0),
           west = ifelse(westost ==1,1,0),
           immigrant = ifelse(gebd ==2,1,0))


highr2  <- lm(einkommen~geschl_rec +alter0 +alter0quad +bildung_rec +
           keineberufsausbildung +arbeitsstd +halbtags +west +
           internet +computer +essen +besuchfreunde +besuchfamilie +kunst +
           theater +museum +fernsehenmin +
           haushaltseinkommen +wirtschaftslage +
           dauerbildung +demzufr +linksrechts +
           prokopfeink + krankengeldhh +
           elterngeldhh +immigrant,data =allb_r)

texreg(highr2, float.pos ="ht!")
```

	Model 1
(Intercept)	−0.84
	(1.49)
geschl_rec	1.92***
	(0.22)
alter0	0.30***
	(0.03)
alter0quad	−0.00***
	(0.00)
bildung_rec	0.25*
	(0.12)
keineberufsausbildung	−2.71***
	(0.41)
arbeitsstd	0.06***
	(0.01)
halbtags	−2.74***
	(0.33)
west	1.53***
	(0.23)
internet	−0.27**
	(0.09)
computer	0.13
	(0.08)
essen	−0.32*
	(0.12)
besuchfreunde	−0.02
	(0.12)
besuchfamilie	−0.22*
	(0.10)
kunst	0.14
	(0.10)
theater	0.12
	(0.18)
museum	0.10
	(0.19)
fernsehenmin	−0.00
	(0.00)
haushaltseinkommen	0.18***
	(0.03)
wirtschaftslage	−0.41**
	(0.15)
dauerbildung	0.14***
	(0.04)
demzufr	0.05
	(0.09)
linksrechts	−0.02
	(0.06)
prokopfeink	0.00***
	(0.00)
krankengeldhh	−1.99
	(1.82)
elterngeldhh	0.88
	(0.62)
immigrant	−0.24
	(0.33)
R^2	0.68
Adj. R^2	0.67
Num. obs.	764
RMSE	2.54

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$