SM II Abgabe 2

Ingrid Espinoza de Sämann

12.November 2018

Aufgabe 1

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

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

Eine gaengige Art und Weise, Standarisierungen durchzufuehren, ist mithilfe der z-Transformation. Z-Werte, auch genannt Standarwerte, werden in einer Form umgewandelt, die es erlaubt, sie mit Werten derselben oder einer anderen Verteilung zu vergleichen. Man erzeugt sie, indem man von jedem Messwert (Rohwert) das arithmetische Mittel substrahiert und die Diferrenz durch die Standarabweichung dividiert. Jede Verteilung von z-Werten hat ein arithmetisches Mittel von Null und eine Standardabweichung von Eins. Die Abweichungen der Messwerte vom aritmetischen Mittel werden in Standardabweichungseinheiten augedrueckt. Z-transformierten Variablen sind bei der Bildung Indizes aus Variblen unterschiedlicher Skalierung sehr praktisch.

b^* ist ein standarisierte Regressionkoeffizient. Standarisiert wird der Slope durch die Bereinigung um die Metrik von Y (der Divisor (Sy)) und die Gewichtung um die Metrik von X (der Dividend (Sx)). $b^* = b(Sx/Sy)$ Es gibt Auskunft über die Staerke eines Zusammenhaenges. Mit deren Hilfe kann das 'einflusstaerkste' Zusammenhang identifiziert werden. Er ist so zu interpretieren, dass bei Veraenderung von X um eine Standardabweichung, sich Y um b^*-Standarabweichungen ändert. Da der Koeffizient standarisiert ist, kann man innerhalb eines Modells die verschiedenen Koeffizienten auf ihren jeweiligen Einfluss vergleichen. Da b^* stichprobenabhängig ist, sollte er nicht zwischen Modelle verglichen werden.

Aufgabe 2

Fuehren Sie eine z-Standardisierung fuer die Originalaltersvariable (alter_z) und die auf Null gesetzte Altersvariable (alter_0z) sowie fuer "unsere" Bildungsvariable (0 bis 4). [Daten: ALLBUS 2014]

```
allb_sub_z <- allb_sub %>%
  mutate(alter_z = scale(alter, center = F, scale = T)) %>%
  mutate(alter0_z = scale(alter0, center = F, scale = T)) %>%
  mutate(bildung_z = scale(bildung, center = F, scale = T)) %>%
  mutate(einkommen_z = scale(einkommen, center = F, scale = T))

allb_sub_z
```

```
## # A tibble: 3,471 x 11
##    alter bildung geschl einkommen alter0 bildung_rec geschl_rec alter_z
##    <dbl> <dbl+l> <dbl+> <dbl+lbl> <dbl+>       <dbl>      <dbl>   <dbl>
##  1 33    5       2      14        15              4          0   0.629
##  2 50    3       2      " 9"      32              2          0   0.953
##  3 56    3       1      17        38              2          1   1.07
##  4 61    3       1      " 8"      43              2          1   1.16
##  5 59    3       2      " 9"      41              2          0   1.12
```

```
##  6 56    3       1      21        38              2          1   1.07
##  7 66    2       1      12        48              1          1   1.26
##  8 25    5       2      NA        " 7"            4          0   0.477
##  9 58    4       1      13        40              3          1   1.11
## 10 53    3       1      19        35              2          1   1.01
## # ... with 3,461 more rows, and 3 more variables: alter0_z <dbl>,
## #   bildung_z <dbl>, einkommen_z <dbl>
```

Aufgabe 2a

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

Dadurch, dass beide Variablen stadarisiert wurden, weisen sie sehr aenhliche Werte auf, denn sie sind praktisch gleich skaliert.

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

	n	mean	sd	median	min	max
alter	3468	49.4403114	17.5062282	50.0000000	18.0000000	91.000000
bildung	3466	3.3620889	1.2525966	3.0000000	1.0000000	7.000000
geschl	3471	1.4923653	0.5000137	1.0000000	1.0000000	2.000000
einkommen	3065	11.1491028	4.9622879	11.0000000	1.0000000	22.000000
alter0	3468	31.4403114	17.5062282	32.0000000	0.0000000	73.000000
bildung_rec	3427	2.3262329	1.2123214	2.0000000	0.0000000	4.000000
geschl_rec	3471	0.5076347	0.5000137	1.0000000	0.0000000	1.000000
alter_z	3468	0.9425297	0.3337386	0.9531996	0.3431519	1.734823
alter0_z	3468	0.8735960	0.4864256	0.8891475	0.0000000	2.028368
bildung_z	3466	0.9369585	0.3490779	0.8360503	0.2786834	1.950784
einkommen_z	3065	0.9134701	0.4065710	0.9012538	0.0819322	1.802508

Aufgabe 2b

Fuehren 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.

Es macht keinen Unterschied, wenn man vor der Regression standarisiert ober ob man es in der Regression macht. Die Werte von Modell 2 sind die standarisierten Koeffizienten von Modell 1 (s. lmbeta)

```
mod1 <- lm(einkommen ~ alter0 + bildung, data = allb_sub_z)

mod2 <- lm(einkommen_z ~ alter0_z + bildung_z, data = allb_sub_z)

screenreg(list(mod1, mod2))
```

```
##
## ====================================
##              Model 1      Model 2
```

```
## -------------------------------------
## (Intercept)     6.21 ***     0.51 ***
##                (0.34)       (0.03)
## alter0          0.04 ***
##                (0.01)
## bildung         1.09 ***
##                (0.07)
## alter0_z                     0.12 ***
##                             (0.02)
## bildung_z                    0.32 ***
##                             (0.02)
## -------------------------------------
## R^2             0.07         0.07
## Adj. R^2        0.07         0.07
## Num. obs.    3062         3062
## RMSE            4.78         0.39
## =====================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

```
#install.packages("lm.beta")
library(lm.beta)
lm.beta(mod1)
```

```
##
## Call:
## lm(formula = einkommen ~ alter0 + bildung, data = allb_sub_z)
##
## Standardized Coefficients::
## (Intercept)      alter0     bildung
##   0.0000000   0.1427816   0.2688033
```

Aufgabe 2c

Wie erklaeren Sie die Werte b und b in Modell 2? TIPP: Verwenden Sie bei Modell 2 das z-transformierte Einkommen als abhaengige Variable.*

B Werte geben Auskunft über die Größe eines Einflusses (EInkommen). Standarisierte Werte geben Auskunft über die Stärke eines Effekts (wie viel Varianz wird gebunden?) Die Werte von Modell 2 sind die standarisierten Koeffizienten von Modell 1 (s. lmbeta)

Aufgabe 3

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

- Einzige Einschraenkung: Keine Regression von Y auf Y.

```
#binoculaR(allbus)

#cor_matrix <- allbus %>%
  #select(V420, V9, V70, V71, V81, V116, V190, V214, V279, V377, V448, V456, V491, V492, V496) %>%
  #cor()
```

```r
#corrplot(cor_matrix, type = "lower", order = "hclust", tl.srt = 45)

#Regression

mod3 <- lm(V420 ~ V9 + V70 + V71 + V81 + V116 + V190 + V214 + V279 + V377 + V448 + V456 + V491 + V49 + V

screenreg(mod3)
```

```
##
## ========================
##               Model 1
## ------------------------
## (Intercept)  -239.47
##              (294.80)
## V9             -1.67 *
##                (0.80)
## V70            -0.06
##                (0.35)
## V71             0.00
##                (0.01)
## V81            -4.64 **
##                (1.43)
## V116           -2.12
##                (1.59)
## V190           -1.57
##                (1.22)
## V214            0.04
##                (0.83)
## V279            0.08
##                (0.15)
## V377           -0.79
##                (1.94)
## V448           -0.14
##                (0.09)
## V456            0.13
##                (0.15)
## V491            0.00
##                (0.00)
## V49             0.29
##                (1.12)
## V496            0.30
##                (1.26)
## ------------------------
## R^2             0.67
## Adj. R^2        0.47
## Num. obs.      38
## RMSE            3.33
## ========================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

Zwar ist kaum ein Koeffizient signifikant, mein Rš ist aber schoen hoch! <3