SM II Abgabe 2

Abigail Alexander

12.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!

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

Wenn Regressionskoeffizienten standardisiert sind, sind sie mit einander vergleichbar innerhalb eine Regressionsanalyse unter gewissen Bindingungen. Die unterschiedlichen Skalen werden standardisiert und haben eine Mittelwert von 0 und Standardabweichung von 1. Der standardisierte Regressionskoeffizient ist gleich den unstandardisierte Regressionskoeffizient multipliziert mit der Standardabweichung der unabhängigen Variablen und dividiert durch die Standardabweichung der abhängigen Variablen.

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) und das Einkommen. [Daten: ALLBUS 2014]

```r
allb_sub<-allb_sub %>%
  mutate(alter_z = scale(alter, center = F, scale = T)) %>%
  mutate(alter_0z = 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
```

```
## # 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: alter_0z <dbl>,
## #   bildung_z <dbl>, einkommen_z <dbl>
```

Aufgabe 2a

Vergleichen Sie die Zahlenwerte, Mean und die Standardabweichung von alter_z und alter_0z und erklären Sie Ihre "Beobachtung".

```
allb_sub %>%
  select(alter_z, alter_0z) %>%
  describe() %>%
  select(-vars, -range, -trimmed, -mad,  -skew, -kurtosis,  -se) %>%
  kable()
```

	n	mean	sd	median	min	
alter_z	3468	0.9425297	0.3337386	0.9531996	0.3431519	
alter_0z	3468	0.8735960	0.4864256	0.8891475	0.0000000	
Wie erwarte	t ist d	ie Mean-wert	e höher für	die Variable	alter_z als	die der Variable alter_0z, da ihre min Wer

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)
mod2 <- lm(einkommen_z ~ alter_0z + bildung_z, data = allb_sub)
screenreg(list(mod1, mod2))
```

```
##
## ======================================
##               Model 1      Model 2
## --------------------------------------
## (Intercept)     7.17 ***     0.51 ***
##                (0.28)       (0.03)
## alter0          0.04 ***
##                (0.01)
## bildung_rec     1.20 ***
##                (0.07)
## alter_0z                     0.12 ***
##                             (0.02)
## bildung_z                    0.32 ***
##                             (0.02)
## --------------------------------------
## R^2             0.08         0.07
## Adj. R^2        0.08         0.07
## Num. obs.    3039         3062
## RMSE            4.74         0.39
## ======================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

Im Modell 2 ist der b-koeffizient für die Variable alter grösser als im Modell 1, aber die für Bildung ist kleiner. Die Standardabweichungen sind im Modell 2 gleich, obwohl im Modell 1 sind die unterschiedlich. In beiden Modellen ist der b-koeffizient für bildung grösser als für alter. Eine Schlussfolgerung wäre, die Variable bildung erklärt die Variation von Einkommen besser als die Variable 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.*

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.*

Tipp: Mit `binoculaR` könnt ihr den Datensatz nach relevanten Variablen durchsuchen :)

```
mod1 <- lm(V420 ~ V83 + V84 + V86 + V81 + V83*V84 + V135 + V136 + V135*V136 + V493 + V529 + V530 + V529
screenreg(list(mod1))
```

```
##
## =========================
##              Model 1
## -------------------------
## (Intercept)  -2478.95
##              (2602.19)
## V83              1.01
##                 (1.22)
## V84              0.39
##                (17.22)
## V86              0.87
##                 (0.60)
## V81             -3.59 **
##                 (1.21)
## V135             0.33
##                 (0.55)
## V136            -1.62
##                 (1.75)
## V493             0.00 ***
##                 (0.00)
## V529           215.01
##               (400.08)
## V530             0.23
##                 (0.39)
## V83:V84          0.00
##                 (0.01)
## V135:V136       -0.02
##                 (0.05)
## V529:V530       -0.11
##                 (0.20)
## -------------------------
## R^2              0.57
## Adj. R^2         0.45
## Num. obs.       54
## RMSE             3.86
## =========================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```