SM II Abgabe 2

Robin Gerl

11.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!

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

Standardisierungen dienen dazu die Skalierungen der verschiedenen Variablen anzupassen um die Effektstärke letztendlich zwischen verschiedenen Variablen in einem Modell vergleichen zu können. Dabei wird zuerst die Standardabweichung von x geteilt durch die Standardabweichung von y. das Produkt wird mit dem b-Wert multipliziert und gibt letztendlich die Stärke des Zusammenhanges 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 <- allb_sub %>%
  mutate(alter_z= scale(alter, center = F, scale = T)) %>%
  mutate(alter_0z= scale(alter0, center = F, scale = T)) %>%
  mutate(bild_z=scale(bildung_rec, center = F, scale = T)) %>%
  mutate(einkommen_z=scale(einkommen, center = F, scale = T))
```

Aufgabe 2a

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

Die Zahlenwerte bewegen sich bei alter_z zwischen 0,34 und 1,73. Da alter_0z schon umgerechnet wurde, beginnen hier die Werte mit 0 rangieren jedoch bis 2,03. Das arithmetische Mittel ist bei alter_z höher (mean=0,94) als bei alter_0z (mean=0,87). Dieser Unterschied lässt sich auf die vorherige Berechnung der Variable zurückführen. Die Standardabweichung liegt bei 0,33 im Falle von alter_z und bei 0,49 für die Variable alter_0z. Diese Unterschiede kommen zustande da durch die Berechnung von alter auf alter0 das mean sich verringert hat (von 49,44 auf 31,44). Die z-standardisierung wird anhand der arithmetischen Mittel und der Standardabweichung berechnet weswegen die z-Werte sich ebenfalls unterscheiden.

```
allb_sub %>%
  describe()
```

```
##             vars    n  mean    sd median trimmed   mad   min   max range
## alter          1 3468 49.44 17.51  50.00   49.25 20.76 18.00 91.00 73.00
## bildung        2 3466  3.36  1.25   3.00    3.33  1.48  1.00  7.00  6.00
## geschl         3 3471  1.49  0.50   1.00    1.49  0.00  1.00  2.00  1.00
## einkommen      4 3065 11.15  4.96  11.00   11.11  5.93  1.00 22.00 21.00
## alter0         5 3468 31.44 17.51  32.00   31.25 20.76  0.00 73.00 73.00
## bildung_rec    6 3427  2.33  1.21   2.00    2.31  1.48  0.00  4.00  4.00
```

```
## geschl_rec     7 3471  0.51  0.50   1.00    0.51  0.00  0.00  1.00  1.00
## alter_z        8 3468  0.94  0.33   0.95    0.94  0.40  0.34  1.73  1.39
## alter_0z       9 3468  0.87  0.49   0.89    0.87  0.58  0.00  2.03  2.03
## bild_z        10 3427  0.89  0.46   0.76    0.88  0.57  0.00  1.52  1.52
## einkommen_z   11 3065  0.91  0.41   0.90    0.91  0.49  0.08  1.80  1.72
##               skew kurtosis   se
## alter         0.06    -0.89 0.30
## bildung       0.34    -1.03 0.02
## geschl        0.03    -2.00 0.01
## einkommen     0.03    -0.86 0.09
## alter0        0.06    -0.89 0.30
## bildung_rec   0.25    -1.30 0.02
## geschl_rec   -0.03    -2.00 0.01
## alter_z       0.06    -0.89 0.01
## alter_0z      0.06    -0.89 0.01
## bild_z        0.25    -1.30 0.01
## einkommen_z   0.03    -0.86 0.01
```

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.

```
altbi <- lm(einkommen ~ alter0 + bildung_rec, data = allb_sub)
altbiz <- lm(einkommen_z ~ alter_0z + bild_z, data = allb_sub)

screenreg(list(altbi,altbiz))
```

```
##
## ====================================
##               Model 1      Model 2
## ------------------------------------
## (Intercept)     7.17 ***     0.59 ***
##                (0.28)       (0.02)
## alter0          0.04 ***
##                (0.01)
## bildung_rec     1.20 ***
##                (0.07)
## alter_0z                     0.11 ***
##                             (0.02)
## bild_z                       0.26 ***
##                             (0.02)
## ------------------------------------
## R^2             0.08         0.08
## Adj. R^2        0.08         0.08
## Num. obs.    3039         3039
## RMSE            4.74         0.39
## ====================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

Bildung besitzt einen größeren Zusammenhang mit der Variable Einkommen als die Variable für Alter. Die Kennziffer für Alter wird durch die Standardisierung höher. Durch die Standardisierung sinkt der Wert für Bildung.

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-Werte unterscheiden sich zwischen den Modellen erheblich. Während die b-Werte von Model 1 Aussagen über die Inhaltlichen Zusammenhang zwischen den Variablen treffen, geben die b*-Werte Aussagen über die Stärke des Zusammenhanges an. So besitzt die Bildung einen größeren Effekt auf das Einkommen als das Alter.

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 :)

```
allb <- allbus %>%
      select(V84, V86, V81, V420, V64, V63, V70, V209, V274, V279 )  %>%
      rename(alter = V84,
           bildung = V86,
           geschl  = V81,
        einkommen  = V420,
        Elektro = V64,
        Metal = V63,
        Fernsehen = V70,
        PInt = V209) %>%
      mutate(alter0 = alter - 18,
        bildung_rec = ifelse(bildung == 6 | bildung == 7, NA, bildung - 1),
         geschl_rec = ifelse(geschl == 2, 0, 1))

max <- lm(einkommen~bildung_rec+alter+geschl+Elektro+Metal+Fernsehen+PInt, data = allb)

screenreg(max)
```

```
##
## ========================
##              Model 1
## ------------------------
## (Intercept)    13.25 ***
##                (0.56)
## bildung_rec     1.14 ***
##                (0.07)
## alter           0.04 ***
##                (0.01)
## geschl         -3.29 ***
##                (0.17)
## Elektro        -0.03
##                (0.07)
## Metal          -0.20 **
##                (0.07)
## Fernsehen       0.02
```

```
##                (0.04)
## PInt           -0.38 ***
##                (0.08)
## ------------------------
## R^2             0.22
## Adj. R^2        0.22
## Num. obs.    3030
## RMSE            4.38
## ========================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

Mit diesem Modell lässt sich letztendlich nur 22% der Varianz erklären. Im Kern ist es jedoch ziemlich egal welche Variablen in die Regression eingeführt werden. Der Koeffizient R^2 steigt mit jeder zusätzlichen Variable an. Eine andere Variante zur Erhöhung des R^2 ist die Einbeziehung von Variablen die inhaltliche Nähe mit der Y Variable besitzen. Ein Beispiel wäre zum Beispiel das Haushaltseinkommen miteinzubeziehen.