Übungsaufgabe Nummer 2

Universität Stuttgart 09.11.2018
Institut für Sozialwissenschaften, SOWI IV
Seminar: Statistische Modellbildung II
Wintersemester 2018/2019
Dozent: Thomas Krause, M. A.
Abgabedatum: 12.11.2018 Anke Daiber

---

1. Erläutern Sie exemplarisch wozu b* benutzt wird und wie man diesen inhaltlich interpretiert!

Der Regressionskoeffizient b* entsteht durch die Standardisierung des Regressionskoeffizienten b. Bei bivariaten Zusammenhängen entspricht b* Pearson's r. Berechnet wird b*, indem der Regressionskoeffizient b mit dem Quotienten des Divisors der Standardabewichung von X und des Dividenden der Standardabweichung von Y multipliziert wird. Die Regressionskoeffizienten b* werden innerhalb eines Modells hinsichtlich ihrer Stärke vergleichbar. Dadurch kann der einflussreichste Zusammenhang ausgemacht werden. Dies ist möglich, da durch die Standardisierung eine Betrachtung unabhängig der individuellen Skalenbreite entsteht. Die Interpretation des Regressionskoeffizienten b* ist inhaltlich jedoch nicht unbedingt sinnvoll, aber möglich. Durch die zu verwendende Einheit der Standardabweichung wird die Interpretation intuitiv nur schwer verständlich. Der Wertebereich von b* reicht im Normalfall von -1 bis 1, bei Multikollinearitäten können hingegen auch größere Werte vorkommen.

Beispielhaft lässt sich die Benutzung von b* am Einfluss von Alter, Bildung und Geschlecht auf Einkommen erkennen. Im Modell 3 der Tabelle 1 ist zu erkennen, dass das Geschlecht mit 0,359 den stärksten Effekt auf das Einkommen hat. Der schwächste Effekt liegt mit 0,14 beim Alter. Theoretisch: Steigt das Alter um eine Standardabweichung an, dann erhöht sich das Einkommen der betreffenden Person um 0,14 Standardabweichungen des Einkommens. Da dies inhaltlich aber schwer intuitiv begreifbar ist, spielt die Bedeutung hauptsächlich losgelöst der absoluten Werten eine große Rolle. Die standardisierten Regressionskoeffizienten b* von Modell 3 lassen sich aber nicht mit den b* von Modell 2 vergleichen – hierfür müssten die unstandardisierten Regressionskoeffizienten genutzt werden.

Tabelle 1: Vergleich der Modelle

	Modell 1		Modell 2		Modell 3	
	b	b*	b	b*	b	b*
Konstante	10,529		7,165		5,151	
Alter	**0,019*** **	**0,068*** **	**0,039*** **	**0,135*** **	**0,04*** **	**0,14*** **
Bildung			**1,199*** **	**0,291*** **	**1,244*** **	**0,3*** **
Geschlecht					**3,558*** **	**0,359*** **
N						
R^2	**0,0046*** **		**0,082*** **		**0,211*** **	
Korr. R2 / Sig. Gesamtmodell	**0,00427*** **		**0,081*** **		**0,210*** **	
Änderung in R2 Modellverbesserung			0,077		0,129	

2. **Führen Sie eine z-Standardisierung für die Originalaltersvariable und die auf Null gesetzte Altersvariable sowie für „unsere“ Bildungsvariable durch.**
   a. **Vergleichen Sie die Zahlenwerte, Mean und die Standardabweichung von alter_z und alter_0z und erklären Sie Ihre „Beobachtung“.**

Weder zwischen den Zahlenwerten oder beim Mean, noch bei der Standardabweichung sind Unterschiede zwischen den beiden z-standardisierten Altersvariablen zu erkennen, obwohl bei den ursprünglichen Variablen das Minimum um 18 Einheiten weiter oben liegt. Der Mean liegt nun bei beiden Variablen bei 0 und die Standardabweichung jeweils bei 1. Begründet werden kann das mit der Z-Standardisierung.

Die Zahlenwerte hängen von der Skalierung der Variable ab. Unterschiedlich skalierte Variablen können aber nicht gut miteinander verglichen werden, daher werden sie z-standardisiert. Z-transformierte Werte werden erstellt, indem von jedem Messwert das arithmetische Mittel subtrahiert und die Differenz durch die Standardabweichung dividiert werden. Die Abweichungen der Messwerte vom arithmetischen Mittel, also 0, werden in Standardabweichungen ausgedrückt. Die Form der Verteilung wird dabei nicht beeinflusst, aber die Metrik.

   b. **Führen Sie eine Regression von Einkommen auf Alter_0 und Bildung (Modell 1) und eine Regression von Einkommen auf alter_0z und bildung_z durch und vergleichen Sie die b-Koeffizienten.**

Die b-Koeffizienten liegen in Modell 1 beim Alter bei 0,04 und bei der Bildung bei 1,2. Beide Koeffizienten sind dabei hochgradig signifikant. Mit jedem Lebensjahr erhöht sich das Einkommen um 0,04 Kategorien. Mit einem eine Kategorie höheren Bildungsabschluss erhöht sich das Einkommen um 1,2 Kategorien. Die Bildung hat dabei mit einem b* von 0,29 einen stärkeren Effekt als das Alter (b*=0,14).

In Modell 2 wird die gleiche Regression mit den z-standardisierten Variablen durchgeführt. Die b-Koeffizienten verändern sich dadurch: das b von alter_0z liegt bei 0,68 und das b der Bildung liegt nun bei 1,45. Die b-Koeffizienten liegen jetzt also etwas höher und sind nach wie vor hochgradig signifikant. Die standardisierten Regressionskoeffizienten b* verändern sich im Vergleich zum ersten Modell hingegen nicht.

   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.**

Wenn anstelle des Einkommens das z-standardisierte Einkommen verwendet wird (Modell 3), verändern sich die standardisierten Regressionskoeffizienten im Vergleich zu den ersten beiden Modellen nicht. Die unstandardisierten Regressionskoeffizienten b hingegen nehmen im Modell 3 die gleichen Werte wir die standardisierten Regressionskoeffizienten aller Modelle an. Wenn alle beteiligten Variablen vor der Regressionsanalyse z-standardisiert werden, werden direkt interpretierbare, standardisierte Regressionskoeffizienten berechnet. Durch die Veränderung der Skalenbreite in Modell 2, wenn also nur die unabhängigen Variablen z-standardisiert werden, verändern sich die unstandardisierten Regressionskoeffizienten im Vergleich zum Modell 1 ohne Z-Transformation, nicht aber die standardisierten Regressionskoeffizienten.

Um untereinander vergleichbare Regressionskoeffizienten zu erhalten, macht es demnach keinen Unterschied, ob vor der Regressionsanalyse (Z-Standardisierung) oder danach (Berechnung von b*) standardisiert wird.

3. **Erstellen Sie ein multivariates Regressionsmodell mit Y=Einkommen. Versuchen Sie dabei den R^2-Wert so groß wie nur irgendwie möglich zubekommen. Fügen Sie die entsprechenden Teile des Outputs in Ihre Abgabe ein.**

R^2 = 0,9237

```
Call:
lm(formula = V420 ~ age + edu + sex + V425 + V70 + V71 + V727 +
    V730 + V521 + V96 + V95 + schicht + fdp + V38 + V8 + V417 +
    V491 + V128, data = allbus)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5079 -0.6017  0.2129  0.7042  2.2318 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.4755431  1.7592831   3.112 0.002511 ** 
age         -0.0051002  0.0133579  -0.382 0.703534    
edu          0.2233145  0.1712340   1.304 0.195623    
sex          0.1235525  0.2933806   0.421 0.674696    
V425         0.3751761  0.2795881   1.342 0.183124    
V70         -0.0880348  0.0752958  -1.169 0.245522    
V71         -0.0041388  0.0018469  -2.241 0.027574 *  
V727        -0.1095245  0.1151640  -0.951 0.344223    
V730        -0.0643848  0.0853989  -0.754 0.452927    
V521        -0.3020370  0.3122292  -0.967 0.336047    
V96          0.3904727  0.4365341   0.894 0.373531    
V95          0.1422545  0.4977538   0.286 0.775716    
schicht      0.3784234  0.2816540   1.344 0.182580    
fdp          1.1809497  0.7622133   1.549 0.124923    
V38          0.5951784  0.3132481   1.900 0.060743 .  
V8          -0.0983705  0.2061608  -0.477 0.634449    
V417         0.0047829  0.0002308  20.727  < 2e-16 ***
V491        -0.0005387  0.0001535  -3.510 0.000713 ***
V128        -0.0029038  0.0016606  -1.749 0.083885 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.262 on 87 degrees of freedom
  (3365 observations deleted due to missingness)
Multiple R-squared:  0.9237,	Adjusted R-squared:  0.9079 
F-statistic: 58.48 on 18 and 87 DF,  p-value: < 2.2e-16
```

Anhang: R-Code

```
#install.packages("foreign")

library("foreign")

#install.packages("survey")

library("survey")

#install.packages("psych")

library("psych")

#install.packages("dplyr")

library("psych")

#install.packages("lm.beta")

library("lm.beta")


allbus<-read.spss ("C:/Users/Anke Daiber/Documents/Uni/Mastersemester 2/Statistik Krause/Übungsaufgaben/Uebungsaufgabe 1/Allbus2014.sav",

        to.data.frame=T, use.value.labels = FALSE,reencode=T)


#Alter Befragter v84: 18 auf 0

allbus$age<-allbus$V84-18


#Geschlecht 0 =weiblich 1 =männlich

allbus$sex<-allbus$V81

allbus$sex[allbus$V81==2]<-0


#Schulabschluss v86

#5 Ausprägungen; 0=kein Schulabschluss, 1=HS, 2=RS, 3=FHR, 4=Abi; Rest=-1 bzw. Missing --> soll sein

allbus$education<-allbus$V86-1

allbus$education[allbus$education>=5]<- NA


#Z-Standardisierung

allbus$age.z<-(allbus$age-mean(allbus$age,na.rm=T))/sd(allbus$age,na.rm=T)

allbus$V84.z<-(allbus$V84-mean(allbus$V84,na.rm=T))/sd(allbus$V84,na.rm=T)

allbus$education.z<-(allbus$education-mean(allbus$education,na.rm=T))/sd(allbus$educa-
tion,na.rm=T)

allbus$V420.z<-(allbus$V420-mean(allbus$V420,na.rm=T))/sd(allbus$V420,na.rm=T)
```

```
#Aufgabe2b
describe(allbus$age.z)
describe(allbus$V84.z)
allbus$age.z
allbus$V84.z

#Multivariate Regression
#Modell 1
fit2b<-lm(V420~age+education, data=allbus)
lm.beta(fit2b)
summary(fit2b)
#Modell 2
fit2b2<-lm(V420~age.z+education.z, data=allbus)
lm.beta(fit2b2)
summary(fit2b2)
#Modell 3 (Aufgabenteil c)
fit2c<-lm(V420.z~age.z+education.z, data=allbus)
lm.beta(fit2c)
summary(fit2c)

#Rekodierung
allbus$age<-allbus$V84-18
allbus$sex<-ifelse(allbus$V81==2,0,1)
allbus$edu<-allbus$V86-1
allbus$edu[allbus$edu>=5]<-NA
allbus$schicht<-ifelse(allbus$V172==6,NA,allbus$V172)
allbus$fdp<-ifelse(allbus$V729==3,1,0)

#Multivariate Regression: Einkommen auf Alter, Bildung und Geschlecht
fitabc<-
lm(V420~age+edu+sex+V425+V70+V71+V727+V730+V521+V96+V95+schicht+fdp+V38+V
8+V417+V491+V128, data=allbus)
lm.beta(fitabc)
summary(fitabc)
```