Übungsaufgabe Nummer 1

Universität Stuttgart
Institut für Sozialwissenschaften, SOWI IV
Seminar: Statistische Modellbildung II
Wintersemester 2018/2019
Dozent: Thomas Krause, M. A.
Abgabedatum: 29.10.2018

26.10.2018

Anke Daiber

1a. Was ist unter Auspartialisierung zu verstehen und wieso ist es aufgrund der beteiligten Mechanismen wichtig immer mehrere Prädiktorvariablen zu berücksichtigen, auch wenn diese ggf. keinen Einfluss auf die abhängige Variable haben?

Bei einer Auspartialisierung werden die Werte jeder X-Variable um diejenigen Anteile bereinigt, die von anderen X-Variablen beeinflusst werden. Übrig bleiben die „reinen" Werte, welche nun für die Regression der Y-Variable auf die jetzt bereinigten X-Variablen genutzt werden können. Die Regressionskoeffizienten werden daher in multiplen Regressionen auch „partielle Regressionskoeffizienten" genannt.

Es ist wichtig mehrere Prädiktorvariablen zu berücksichtigen, da Supressorvariablen (x2) den „wahren" Zusammenhang zwischen einer unabhängigen (x1) und einer abhängigen (y) Variablen unterdrücken können. Der „wahre" Zusammenhang wird erst dann freigegeben, wenn die Varianzanteile von x2 aus x1 auspartialisiert werden, die keinen Einfluss auf y haben. Auch wenn die Supressorvariable keinen Einfluss auf die Y-Variable hat, kann sie Einfluss auf den Effekt einer anderen X-Variablen nehmen. Außerdem sind mehrere Prädiktorvariablen wichtig, da sie Effekte verringern, aufheben oder die Richtung ändern können.

1b. Wieso können unabhängige Variablen (xi) im multiplen Regressionsmodell einen Einfluss auf Y haben, obwohl die bivariate Korrelation zwischen ihnen und Y nicht signifikant ist?

Unabhängige Variablen können dennoch einen Einfluss auf Y haben, da unterschiedliche Wirkungsverläufe je nach Gruppe bei der x-Variablen aus der bivariaten Analyse vorhanden sein können. Wenn bei einer sequentiellen Modellierung eine andere x-Variable hinzugenommen wird, werden die unterschiedlichen Wirkungsverläufe aufgeschlüsselt und sichtbar. Ein vormals verdeckter Effekt wird nun sichtbar. Der Effekt wird kontrolliert, das heißt, dass er auf den eigenen Erklärungsbereich reduziert wird. Die unabhängige Variable aus der bivariaten Analyse kann eine Proxy-Variable sein, das heißt, dass eigentlich eine andere unabhängige Variable über z. B. soziale Prozesse wirkt.

3a Vergleichen Sie die Regressionskoeffizienten über die Modelle und erläutern Sie was hier festzustellen ist!

Wie in Tabelle 1 zu sehen ist, zeigt sich, dass das Einkommen sehr schwach mit dem Alter korreliert (b*=0,0678). Der Zusammenhang ist signifikant. Mit jedem zusätzlichen Altersjahr steigt das Einkommen um durchschnittlich 0,0194 Kategorien an.

Tabelle 1: Vergleich der Modelle a, b und c

Modelle	Einkommen ~ Alter	Einkommen ~ Bildung	Einkommen ~ Geschlecht
B	0,0194***	1,0517***	3,4737***
B*	0,0678	0,2553	0,3496
R^2	0,0046***	0,0652***	0,1222***
Korr. R^2	0,0043***	0,0649***	0,122***

*** p<0,001

Das Einkommen korreliert ebenfalls mit der Bildung, der Zusammenhang ist hier jedoch eher schwach (b*=0,2553) und ebenfalls signifikant. Ein eine Kategorie höherer Bildungsabschluss führt zu einem 1,0517 Kategorien höheren Einkommen. Das Einkommen korreliert mäßig stark mit dem Geschlecht (b*=0,3496). Der Zusammenhang ist signifikant. Männer haben im Vergleich zu Frauen ein um 3,4737 Kategorien höheres Einkommen. Vergleicht man die signifikanten Effekte hinsichtlich ihrer Stärke anhand des standardisierten Regressionskoeffizienten B*, zeigt sich, dass die Effektstärke beim Einfluss von Geschlecht am höchsten ist, gefolgt von der Bildung und die geringste Effektstärke hat das Alter. Wenn sich das Alter um eine Standardabweichung nach oben verändert, dann ändert sich das Einkommen um 0,0678 Standardabweichungen. Wenn sich die Bildung um eine Standardabweichung nach oben verändert, dann ändert sich das Einkommen um 0,2553 Standardabweichungen. Wenn sich das Geschlecht um eine Standardabweichung nach oben verändert, dann ändert sich das Einkommen um 0,3496 Standardabweichungen. Da Drittvariablen die bivariaten Zusammenhänge verzerren können, wird eine multivariate lineare Regression geschätzt. Das Ziel ist dabei, den Einfluss der jeweiligen unabhängigen Variablen zu isolieren.

Im zweiten Modell (vgl. Tabelle 2) ist zu sehen, dass durch die Hinzunahme der Bildung als unabhängiger Variable der Regressionskoeffizient b von Alter von 0,019 auf 0,039 ansteigt. Durch die Auspartialisierung der gemeinsamen Varianz von Alter und Bildung ist demnach der Zusammenhang von Alter und Einkommen weiter freigelegt worden. Durch die Hinzunahme von Geschlecht als unabhängiger Variable in Modell 3 bleibt b von Alter nahezu gleich mit b =

0,04. Der Effekt ist in jedem Modell signifikant. Der Regressionskoeffizient b von Bildung liegt im zweiten Modell bei b = 1,199 und steigt im dritten Modell auf b = 1,244 leicht an. Die Effekte sind dabei alle signifikant. Der Regressionskoeffizient b liegt in Modell 3 beim Geschlecht bei 3,558. Männer haben ein um 3,558 Kategorien höheres Einkommen als Frauen. Im zweiten Modell hat die Bildung mit b* = 0,291 einen stärkeren Effekt auf das Einkommen als das Alter (b*=0,135). In Modell 3 hat das Geschlecht den stärksten Effekt (b*=0,359), direkt gefolgt von der Bildung (b*=0,3) und deutlich schwächer das Alter (b*=0,14). Im multivariaten Modell 3 ist zu sehen, dass die Konstante bei 5,151 Kategorien im Einkommen liegt. Eine befragte Person, die bei allen unabhängigen Variablen den Wert null aufweist, liegt kurz über der fünften Kategorie beim Einkommen.

3b Vergleichen Sie R2 über die Modelle und erläutern Sie was hier festzustellen ist!

Durch Kenntnis des Alters lassen sich 0,43% der Varianz aufklären. Die Varianzaufklärung von Bildung liegt bei 6,49%. Durch Kenntnis des Geschlechts lassen sich 12,2% der Varianz aufklären. Durch Hinzunahme der Bildung lässt sich 7,7% mehr Varianz im zweiten Modell im Vergleich zu Modell 1 aufklären (R^2 = 0,081). Die Varianzaufklärung steigt im dritten Modell durch die Hinzunahme des Geschlechts um 12,9%-Punkte im Vergleich zu Modell 2 an. Die Gesamtvarianz liegt bei R^2 = 0,21. Alle Modelle sind insgesamt signifikant.

Tabelle 2: Vergleich der Modelle a, ab und abc

	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	

Anhang: R-Code zur Lösung der Übungsaufgabe

```
library("foreign")
library("survey")
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)

#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
table(allbus$V420)

#Modell a: Einkommen auf Alter
fita<-lm(V420 ~ age, data=allbus)
lm.beta(fita)
summary(fita)
#Modell b: Einkommen auf Bildung
fitb<-lm(V420 ~ edu, data=allbus)
lm.beta(fitb)
summary(fitb)
#Modell c: Einkommen auf Geschlecht
fitc<-lm(V420~sex, data=allbus)
lm.beta(fitc)
summary(fitc)
#Modell ab: Einkommen auf Alter und Bildung
fitab<-lm(V420~age+edu, data=allbus)
lm.beta(fitab)
summary(fitab)
#Modell abc: Einkommen auf Alter, Bildung und Geschlecht
fitabc<-lm(V420~age+edu+sex, data=allbus)
lm.beta(fitabc)
summary(fitabc)
```