Universität Stuttgart 26.10.2018
Institut für Sozialwissenschaften (SOWI IV)
Veranstaltung: Statistische Modellbildung II
Thomas Krause (M.A.)
Wintersemester 2018/19 Vanessa Schwaiger

Übungsaufgabe 1

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?

Das Verfahren der Auspartialisierung kann auch als Bereinigungs-Verfahren bezeichnet werden. Durch die Auspartialisierung erhält man die partiellen Regressionskoeffizienten. Diese sind deshalb notwendig, weil wir den „reinen“ Effekt der unabhängigen Variablen betrachten müssen, um einen wirklichen Zusammenhang zu bestätigen. Hierfür sind zwei Regressionen pro unabhängige Variable notwendig. Zuerst wird eine Regression eingesetzt, um den Zusammenhang zwischen der ersten (x1) und der zweiten (x2) unabhängigen Variablen zu überprüfen, um in der zweiten Regression die Residuen von x1 als unabhängige Variable in die Regressionsgleichung einzusetzen. Somit erhält man b1 (den partiellen Koeffizienten b für x1 → den bereinigten Effekt).

Prädiktorvariablen sind deshalb notwendig, weil sie zum Beispiel als Suppressor-Variablen fungieren können. Das bedeutet, dass sie den Effekt anderer unabhängiger Variablen beeinflussen können. Der „wahre“ Zusammenhang kann unterdrückt werden. Um diesen freizugeben muss eine Auspartialisierung vorgenommen werden. Selbst wenn die Suppressor-Variable keinen Einfluss auf die unabhängige Variable besitzt, kann sie den Effekt einer anderen unabhängigen Variablen beeinflussen. Außerdem kann der Effekt einer unabhängigen Variablen durch Zunahme weiterer Variablen verringert/verstärkt werden und sogar ein Wechsel der Richtung des Zusammenhangs ist möglich.

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

Bei einer bivariaten Analyse wird ein Zusammenhang zwischen einer unabhängigen und einer abhängigen Variable geprüft. Dieser Effekt kann verzerrt werden, da die Wirkungsverläufe unterschiedlich ausfallen würden, wenn x1 im Hinblick auf unterschiedliche Gruppen beobachtet werden würde. Durch Hinzunahme weiterer unabhängiger Variable kann x1 in diese unterschiedlichen Gruppen aufgeschlüsselt werden und ein Effekt aufgedeckt werden, der zuvor verdeckt wurde. Hierfür wird das sequenzielle Vorgehen genutzt.

3

Regressionskoeffizienten

Modell	b	b*	R^2	Korr. R^2
Einkommen ~ Alter	0,019392***	0,06779049	0,004596	0,00427
Einkommen ~ Bildung	1,05170***	0,2553349	0,0652	0,06489
Einkommen ~ Geschlecht	3,4737***	0,3496311	0,1222	0,122

Multivariates Modell

	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*** **
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	

3a.

Die bivariaten Zusammenhänge sind jeweils auf dem 99,99%igem Signifikanzniveau signifikant. Der Zusammenhang zwischen Alter und Einkommen ist mit b*=0,068 sehr schwach. Mit jedem Lebensjahr steigt das Einkommen um 0,019 Einheiten (b=0,019). Bei jeder Zunahme der Bildungs-Variable sind dies hingegen schon 1,052 (b) und bei männlichem Geschlecht 3,474 (b) Einheiten. Der Einfluss von Geschlecht ist mit b*=0,35 mäßig stark, während der von Bildung eher schwach ist (b*=0,255). Es scheint, als ob das Einkommen mit steigendem Alter und höherer Bildung zunimmt, sowie bei Männern höher ist als bei Frauen. Bei Zunahme des Alters um eine Standardabweichung verändert sich das Einkommen um 0,068 Standardabweichungen. Bei Veränderung der Bildung um eine Standardabweichung sind es 0,255 Standardabweichungen und bei Veränderung um eine Standardabweichung des Geschlechts 0,35 Standardabweichungen.

Im multivariaten Modell zeigt sich, dass die Konstante in Modell 3 bei 5,151 Einheiten im Einkommen liegt. Ein Befragter, der bei allen unabhängigen Variablen den Wert null aufweist, würde demnach bei 5,151 Einheiten auf der Einkommensvariable liegen. Der Koeffizient b der Variable Alter nimmt mit Zunahme der Bildung in Modell 2 leicht zu (b=0,039) und ist im Modell 3 mit b=0,04 gleichbleibend. Mit jedem nächsthöheren Bildungsabschluss steigt das Einkommen um 1,199 Einheiten (Modell 2). In Modell 3 erhöht sich dieser Wert noch einmal geringfügig (b=1,244). Das Einkommen erhöht sich bei Befragten, die männlich sind um 3,558 Einheiten. Die Regressionskoeffizienten b wurden demnach durch die Zunahme der unabhängigen Variablen erhöht.

Vergleicht man die signifikanten Effekte hinsichtlich ihrer Stärke anhand des standardisierten Regressionskoeffizienten b*, zeigt sich, das Geschlecht mit 0,359 den größten Effekt auf politische Partizipation hat, dicht gefolgt von der Bildung mit 0,359. Das Alter hat hingegen einen schwächeren signifikanten Effekt (b*=0,14). Alle Effekte sind hochsignifikant (99,99%iges Signifikanzniveau). Hier kann der Vorteil eines multivariaten Modells beobachtet werden: wenn die unabhängigen Variablen durch die Betrachtung der bivariaten Analyse ausgewählt worden wären, wäre Alter wegen des geringen Effekts nicht miteinbezogen worden obwohl sich in der multivariaten Analyse ein signifikanter Effekt aufzeigt. Durch die Auspartialisierung der Varianz von Alter und Bildung, konnte der Zusammenhang zwischen Alter und Einkommen eigenständig bewertet werden und wurde weiter freigelegt.

3b.

Bei der Betrachtung von R^2 fällt auf, dass ein deutlicher Zuwachs zwischen dem Modell 2 und 3 (also durch Zunahme der Variablen Geschlecht) zu beobachten ist. 21 Prozent der Varianz im Einkommen können durch Kenntnis der drei unabhängigen Variablen erklärt werden. Das geschätzte Modell besitzt somit eine moderate Erklärungskraft. Während das Modell 1 mit R^2=0,046 keine Erklärungskraft aufweist erhöht sich diese in Modell 2 mit R^2=0,081. Die Kenntnis des Alters alleine kann nur einen sehr geringen, nicht nennenswerten Anteil an Varianz im Einkommen erklären.

Anhang R-Code:

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

allbus<-read.spss ("C:/Users/Vanes/Desktop/Allbus.sav",
            to.data.frame=T, use.value.labels = FALSE,reencode=T)


#relevante Variablen:
#V84   ALTER: BEFRAGTE<R>
#V86   ALLGEMEINER SCHULABSCHLUSS
#V81   GESCHLECHT, BEFRAGTE<R>
#V420 NETTOEINKOMMEN<OFFENE+LISTENANGABE>,KAT.


#Alter Befragter v84
table(allbus$V84)
allbus$age<-allbus$V84-18
allbus$age


#Geschlecht 0 =weiblich 1 =männlich
#Geschlecht v81 bis jetzt: 1= männlich 2 =weiblich
allbus$V81
allbus$sex<-allbus$V81
allbus$sex[allbus$V81==2]<-0
allbus$sex

#Schulabschluss v86
```

```
#5 Ausprägungen; 0=kein Schulabschluss, 1=HS, 2=RS, 3=FHR, 4=Abi; Rest=-1 bzw. Missing
--> soll sein
allbus$V86
allbus$education<-allbus$V86-1
allbus$education[allbus$education>=5]<- NA
allbus$education

#Einkommen: V420
allbus$V420

#Regression
library("lm.beta")
fit <- lm(V420 ~age , data=allbus)
summary(fit)
lm.beta(fit)


fit1<-lm(V420 ~ education, data=allbus)
summary(fit1)
lm.beta(fit1)

fit2<-lm(V420 ~ sex, data=allbus)
summary(fit2)
lm.beta(fit2)

fit3<-lm(V420~age +education, data=allbus)
summary(fit3)
lm.beta(fit3)

fit4<-lm(V420~age + education + sex, data=allbus)
summary(fit4)
lm.beta(fit4)
```