Musterlösung: Übungsaufgabe 3

Statistische Modellbildung II

17.November 2017

Aufgabe 1

Was ist unter einer Dummy-Variablen zu verstehen, wozu wird sie benötigt und wie ist sie zu interpretieren?

Dummy-Variablen sind binär codierte Variablen, die i.d.R. die Ausprägungen 0 und 1 haben. Sie werden für die Berücksichtigung der Effekte von Variablen, die eine nominale oder ordinale Skalierung aufweisen, wie z.B. Geschlecht oder Bildung, verwendet. Der Wert 1 beschreibt dabei das Vorliegen des bestimmten Merkmals, während ein Fall den Wert 0 erhält, wenn das Merkmal nicht vorliegt. Für die Erstellung von DummyVariablen aus einer Variable mit mehreren, nicht metrischen Ausprägungen wird die Ausgangsvariable in k-1 dichotome Variablen zerlegt, d.h. eine Variable weniger, als Ausprägungen vorliegen. Die übrig gebliebene Ausprägung stellt die Referenzkategorie für die anderen Dummy-Variablen dar.

Aufgabe 2

Erstellen Sie aus der Bildungsvariablen "eine" Dummyvariable. Führen Sie eine Regression von Einkommen auf Alter, Bildung und Geschlecht durch:

- `0` OHNE ABSCHLUSS
- `1` VOLKS-,HAUPTSCHULE
- `2` MITTLERE REIFE
- `3` FACHHOCHSCHULREIFE
- `4` HOCHSCHULREIFE
- `NA` ANDERER ABSCHLUSS
- `NA` NOCH SCHUELER
- `NA` KEINE ANGABE

```
allb_sub <- allb_sub %>%
  mutate(ohne_abschluss = ifelse(bildung_rec == 0, 1, 0),
         hauptschule    = ifelse(bildung_rec == 1, 1, 0),
         realschule     = ifelse(bildung_rec == 2, 1, 0),
         fachhoch       = ifelse(bildung_rec == 3, 1, 0),
         hochschul      = ifelse(bildung_rec == 4, 1, 0))

dummod <- lm(einkommen ~ ohne_abschluss +
           realschule + fachhoch + hochschul +
           alter0 + geschl_rec, data = allb_sub)

comp <- lm(einkommen ~ bildung_rec +
           alter0 + geschl_rec, data = allb_sub)

texreg(list(dummod,comp),
            float.pos ="ht!")
```

	Model 1	Model 2
(Intercept)	6.20***	5.15***
	(0.26)	(0.28)
ohne_abschluss	−1.40*	
	(0.65)	
realschule	1.53***	
	(0.21)	
fachhoch	3.39***	
	(0.32)	
hochschul	3.69***	
	(0.22)	
alter0	0.04***	0.04***
	(0.00)	(0.00)
geschl_rec	3.56***	3.56***
	(0.16)	(0.16)
bildung_rec		1.24***
		(0.07)
R^2	0.21	0.21
Adj. R^2	0.21	0.21
Num. obs.	3039	3039
RMSE	4.39	4.40

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1: Statistical models

Aufgabe 2a

Interpretieren Sie die Koeffizienten für "Abitur" und "Realschulabschluss". Achten Sie genau auf Ihre Formulierung.

Eine 18-jährige weibliche Person mit einem Volks- oder Hauptschulabschluss verdient laut Modell 1 durschnittlich 6,20 Einkommenseinheiten. Eine 18-jährige weibliche Person mit einem Realschulabschluss verdient 1,53 Einkommenseinheiten mehr als eine gleichaltrige weibliche Person mit einem Volks- oder Hauptschulabschluss (b=1.53, p<0.001), unter Kontrolle des Alters und Geschlecht. Eine 18-jährige weibliche Person mit Abitur verdient 3,69 Einkommenseinheiten mehr als eine gleichaltrige weibliche Person mit einem Volks- oder Hautpschulabschluss (b=3.69, p<0.001), unter Kontrolle des Alters und Geschlecht.

Aufgabe 2b

Interpretieren Sie R^2 und vergleichen dies mit den Ergebnissen aus der letzten Woche (Aufgabenblatt 2).

Durch die Kontrolle des Einflusses der verschiedenen Bildungsabschlüsse, dem Geschlecht und dem Alter auf das Einkommen können 21% der Varianz im Modell gebunden werden. Im Vergleich mit dem multivariaten Modell von Aufgabenblatt 1 fällt auf, dass in Modell 5 ebenso 21% der Varianz statistisch erklärt werden können. Das liegt daran, dass die Modelle (Bildung als kontinuierliche oder dummy-Variable) beinahe identisch sind und das selbe abbilden.

Aufgabe 3

Was ist unter einer Moderatorvariable zu verstehen und auf welche Weise(n) kann der Einfluss eines Moderatoreffektes untersucht werden? Erklären Sie kurz das Vorgehen hierbei.

Eine Moderatorvariable ist eine Variable z welche einen Einfluss auf den Effekt zwischen einer x und einer y Variable ausübt. Also in dem hier angewandten Beispiel hat die Moderatorvariable "Geschlecht" einen

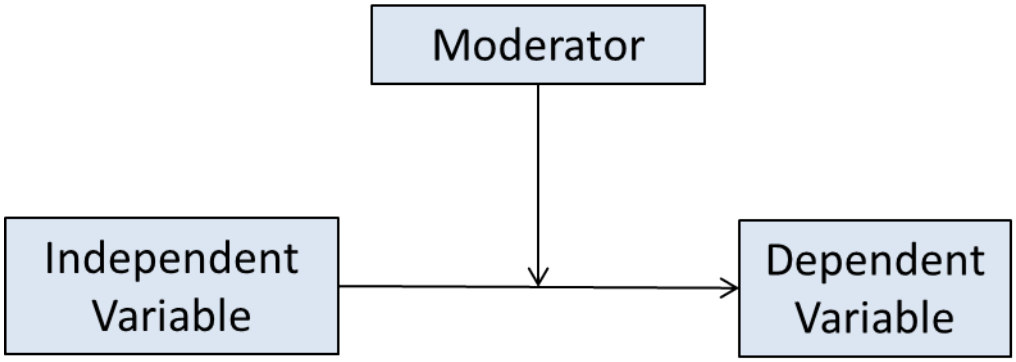


Figure 1: Moderatoreffekt

Einfluss darauf, wie Alter auf Einkommen wirkt, also auf den Effekt von Alter auf Einkommen. Eine Moderatorvariable beeinflusst also nicht die Ausprägung einer Variable, sondern sie beeinflusst die Beziehung zwischen x und y. Denn da wo z (also die Moderatorvariable) variiert, dort variiert auch der Effekt von Alter auf Einkommen.

Ein Moderatoreffekt kann auf zwei verschiedene Weisen betrachtet werden. Eine Lösung wäre, eine *Interaktionsvariable* zu erstellen. Das heißt man multipliziert die z (Moderator-) Variable und die x (UV) Variable miteinander und bezieht diese mit in das Regressionsmodell. Durch die Multiplikation kann die Interaktionsvariable nur dann einen Wert annehmen, wenn beide Variablen einen Wert angenommen haben sonst wäre die Interaktionsvariable schließlich 0. Nun kann man beispielsweise prüfen, ob der Alterseffekt (auf Einkommen) sich in Abhängigkeit vom Geschlecht, also der Moderatorvariable, verändert. Hier ist also wichtig, dass nur, wenn auch die Ausgangsvariablen berücksichtigt werden, der Interaktionseffekt auch etwas aussagen kann. Problematisch kann hier sein, dass dann die Signifikanz der Originalvariablen abfällt, bedingt durch Multikollinearität.

Deshalb gibt es noch eine zweite Lösung: *Einen Multigruppenvergleich.* Man splittet des Datensatz anhand der Moderatorvariable, z.B. nach weiblich und männlich, rechnet die Regression für beide Gruppen und vergleicht dann. Problematisch hier kann werden, dass die Signifikanz der Moderatoreffekte nur auf umständliche Weise geprüft werden kann.

Aufgabe 4

Erstellen Sie eine Interaktionsvariable zwischen Geschlecht und Alter und reduzieren Sie vor den folgenden Regressionsanalysen den Datensatz, um die Fälle, bei denen die Geburtenentscheidungen keine besondere Bedeutung mehr für die Gehaltsentwicklung haben sollten (Alter unter 46 Jahren -> SELE IF Alter_0 < 28). Modell 1 enthält dann Alter und Geschlecht, in Modell 2 kommt die Interaktionsvariable hinzu.

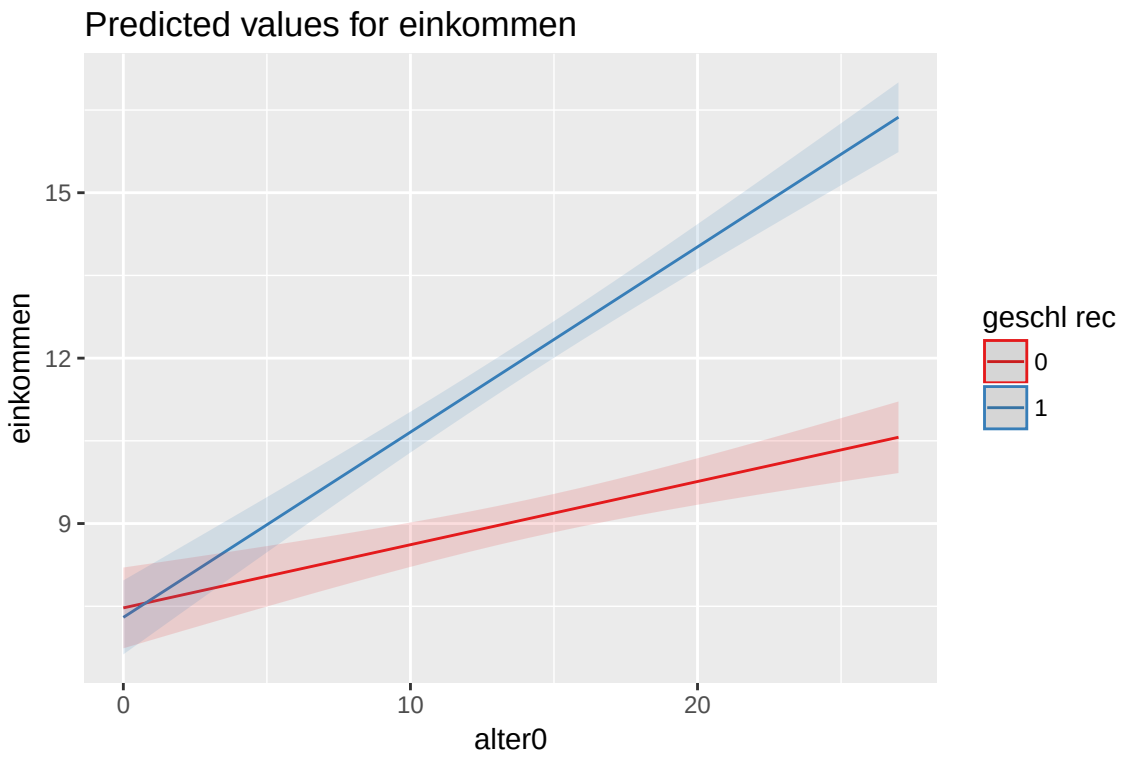
```
# Filter erstellen
allb_sub_old <- allb_sub %>%
  filter(alter0 < 28)

mod1 <- lm(einkommen ~ alter0 + geschl_rec, data = allb_sub_old)
mod2 <- lm(einkommen ~ alter0 + geschl_rec + alter0 * geschl_rec, data = allb_sub_old)

texreg(list(mod1,mod2),
            float.pos ="ht!")
```

```
plot_model(mod2, type = "int")
```

	Model 1	Model 2
(Intercept)	5.76***	7.47***
	(0.29)	(0.37)
alter0	0.23***	0.11***
	(0.02)	(0.02)
geschl_rec	3.00***	−0.17
	(0.25)	(0.51)
alter0:geschl_rec		0.22***
		(0.03)
R^2	0.22	0.25
Adj. R^2	0.22	0.25
Num. obs.	1224	1224
RMSE	4.36	4.27

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Statistical models



Aufgabe 4a

Berechnen Sie anhand von Modell 1 und Modell 2 jeweils das prognostizierte Einkommen eines Mannes und einer Frau im Alter von jeweils 30 Jahren.

```
#Modell 1
intercept_1 <- 5.76
alter0_1 <- 0.23
geschl_rec_1 <- 3.00

#Modell 1: Mann 30 Jahre
```

```r
intercept_1 + alter0_1 * 12 + geschl_rec_1
```

```
## [1] 11.52
```

```r
#Modell 1: Frau 30 Jahre
intercept_1 + alter0_1 * 12
```

```
## [1] 8.52
```

```r
#Modell 2
intercept_2 <- 7.47
alter0_2 <- 0.11
geschl_rec_2 <- -0.17
int <- 0.22

#Modell 2: Mann 30 Jahre
intercept_2 + 12 * alter0_2 + geschl_rec_2 + 12 * int
```

```
## [1] 11.26
```

```r
#Modell 2: Frau 30 Jahre
intercept_2 + 12 * alter0_2
```

```
## [1] 8.79
```

Aufgabe 4b

Was ist dabei der Interaktionseffekt und wie lässt er sich inhaltlich begründen?

Der Interaktionseffekt besteht darin, dass der Zusammenhang zwischen Alter und Einkommen bei Frauen weniger stark positiv ist als bei Männern. Das Geschlecht moderiert also den Effekt von Alter auf Einkommen. Spezifiziert man keine Interaktionsvariable wie in Model 1, wird der Effekt von Alter nicht zwischen den Gruppen separiert und der Effekt ist nicht ersichtlich. Interpretiert werden könnte dies so, dass Frauen häufiger einen weniger kontinuierlichen Berufsweg haben, vor allem durch Schwangerschaft und Kindererziehung.