Musterlösung: Übungsaufgabe 1

Statistische Modellbildung II

5.November 2017

Aufgabe 1

Aufgabe 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?

Es ist wichtig, mehrere Prädiktoren zu verwenden, da die unabhängigen Variablen häufig Wechselwirkungen untereinander aufweisen. Unter Auspartialisierung versteht man das "Herausrechnen" der Effekte der anderen unabhängigen Variablen, sodass Y nur auf den Teil der Varianz von X zurückgeführt wird, der nicht von den anderen unabhängigen Variablen beeinflusst wird. Hierbei werden die Effekte der anderen unabhängigen Variablen berücksichtigt und konstant gehalten. Auch wenn unabhängige X-Variablen keinen Einfluss auf die abhängige Y-Variable aufweisen, so können diese andere X-Variable beeinflussen und so den Einfluss einer X-Variable auf Y verfälschen bzw. den wahren Effekt verbergen (z.B. Suppressoreffekt). Daher ist es stets wichtig auf andere Prädiktorvariablen zu berücksichtigen, auch wenn diese keinen Einfluss auf Y nehmen.

Aufgabe 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?

Das kann aufgrund von sogenannten *Suppressoreffekten* der Fall sein. Im bivariaten Modell kann der Einfluss der unabhängigen Variable auf abhängige Variablen durch Varianzanteile überlagert sein, welche nicht mit der abhängigen Variablen zusammenhängen. Daher können im bivariaten (und somit unbereinigten) Fall insignifikante Ergebnisse zustande. Im multiplen Regressionsmodell wird dieser Effekt kontrolliert und bereinigt, sodass im Gesamtmodell signifikante Ergebnisse entstehen können. Die enstprechende Varianz der unabhängigen Variable, welche keinen Einfluss auf die abhängige Variable ausübt, wird auspartialisiert und somit wird der eigentlich Effekt sichtbar.

Aufgabe 2

Bevor Sie die Analysen durchführen, suchen Sie im Codebuch (o. Variablenliste) Ihres Datensatzes (ALLBUS 2014) am besten Mittels STRG+F (aufrufen der "Suchenfunktion" in nahezu allen Programmen) die folgenden Variablen heraus: Alter, Geschlecht, Schulabschluss und individuelles Nettoeinkommen in der Fassung "Offene Angaben+Listeangaben".

Kodieren Sie dann diese Variablen wie folgt:

- *Alter: Startwert auf 0 setzen; 18=0, 48=30*
- *Schulabschluss- bzw. Schuldbildung: 5 Ausprägungen; 0=kein Schulabschluss, 1=HS, 2=RS, 3=FHR, 4=Abi; Rest=-1 bzw. Missing*
- *Geschlecht: 0=weiblich; 1=männlich*

1. Schritt: Datensatz einladen

```
allbus <- read_spss("allbus2014.sav")
```

2. Schritt: relevante Variablen identifizieren

```
var_names(allbus, "alter") # V84 ALTER: BEFRAGTE<R>
var_names(allbus, "schulabschluss") # V86 ALLGEMEINER SCHULABSCHLUSS
var_names(allbus, "geschl") # V81 GESCHLECHT, BEFRAGTE<R>
var_names(allbus, "eink") # V420 NETTOEINKOMMEN<OFFENE+LISTENANGABE>,KAT.
```

3. Schritt: Jetzt wählen wir die Variablen und erstellen ein Subset!

```
allb_sub <- select(allbus, V84, V86, V81, V420)
```

4. Schritt: Als nächstes benennen wir die Variablen um!

```
allb_sub <- rename(allb_sub, alter=V84, bildung = V86, geschl = V81, einkommen = V420)
```

5. Schritt: Als nächstes Rekodieren wir die Variablen

```
allb_sub <- mutate(allb_sub,
                   alter0 = alter - 18,
                   bildung_rec = ifelse(bildung == 6 | bildung == 7, NA, bildung-1),
                   geschl_rec = ifelse(geschl == 2, 0, 1))

#ODER mit dem Recode() Befehl aus dem car package

allb_sub <- mutate(allb_sub,
                    alter0 = alter - 18,
                    bildung_rec = Recode(bildung-1,
                                         "5 = NA;
                                          6 = NA"),
                    geschl_rec = ifelse(geschl == 2, 0, 1))
```

Bonus: Alles mit dem pipe operator %>%

```
# all together now!
allb_sub <- allbus %>%
  select(V84, V86, V81, V420) %>%
  rename(alter=V84, bildung = V86, geschl = V81, einkommen = V420) %>%
  mutate(alter0 = alter - 18) %>%
  mutate(bildung_rec = ifelse(bildung == 6 | bildung == 7, NA, bildung-1)) %>%
  mutate(geschl_rec = ifelse(geschl == 2, 0, 1))

head(allb_sub)
```

```
## # A tibble: 6 x 7
##       alter   bildung    geschl einkommen    alter0 bildung_rec geschl_rec
##   <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl>       <dbl>      <dbl>
## 1        33         5         2        14        15           4          0
## 2        50         3         2         9        32           2          0
## 3        56         3         1        17        38           2          1
## 4        61         3         1         8        43           2          1
## 5        59         3         2         9        41           2          0
## 6        56         3         1        21        38           2          1
```

Aufgabe 3

Berechnen Sie folgende (sequentielle) Regressionsmodelle:

- *Modell a: Einkommen auf Alter;*
- *Modell b: Einkommen auf Bildung;*
- *Modell c: Einkommen auf Geschlecht;*
- *Modell ab: Einkommen auf Alter und Bildung;*
- *Modell abc: Einkommen auf Alter, Bildung und Geschlecht.*

```
modell_a <- lm(einkommen ~ alter0, data = allb_sub)
modell_b <- lm(einkommen ~ bildung_rec, data = allb_sub)
modell_c <- lm(einkommen ~ geschl_rec, data = allb_sub)
modell_ab <- lm(einkommen ~ alter0 + bildung_rec, data = allb_sub)
modell_abc <- lm(einkommen ~ alter0 + bildung_rec + geschl_rec, data = allb_sub)

#Modelle anzeigen
texreg(list(modell_a,
          modell_b,
          modell_c,
          modell_ab,
          modell_abc),
       caption = "Modelle 1 - 5: Unstandartisierte Koeffizienten",
       custom.coef.names = c("(Intercept)", "Alter", "Bildung", "Geschlecht (0/1)"),
       float.pos = "ht!")
```

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	10.53***	8.75***	9.33***	7.17***	5.15***
	(0.19)	(0.19)	(0.12)	(0.28)	(0.28)
Alter	0.02***			0.04***	0.04***
	(0.01)			(0.01)	(0.00)
Bildung		1.05***		1.20***	1.24***
		(0.07)		(0.07)	(0.07)
Geschlecht (0/1)			3.47***		3.56***
			(0.17)		(0.16)
R^2	0.00	0.07	0.12	0.08	0.21
Adj. R^2	0.00	0.06	0.12	0.08	0.21
Num. obs.	3064	3040	3065	3039	3039
RMSE	4.95	4.78	4.65	4.74	4.40

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

Table 1: Modelle 1 - 5: Unstandartisierte Koeffizienten

Für standartisierte Koeffizenten können wir zunächst alle Variablen mit `mutate_all` und der Funktion `scale` z-standartisieren und mittelwertzentrieren.

```
allb_sub_scale <- mutate_all(allb_sub, scale)

modell_a_beta <- lm(einkommen ~ alter0, data = allb_sub_scale)
modell_b_beta <- lm(einkommen ~ bildung_rec, data = allb_sub_scale)
modell_c_beta <- lm(einkommen ~ geschl_rec, data = allb_sub_scale)
modell_ab_beta <- lm(einkommen ~ alter0 + bildung_rec, data = allb_sub_scale)
modell_abc_beta <- lm(einkommen ~ alter0 + bildung_rec + geschl_rec, data = allb_sub_scale)
```

```r
#Modelle anzeigen
texreg(list(modell_a_beta ,
          modell_b_beta,
          modell_c_beta,
          modell_ab_beta,
          modell_abc_beta),
       caption = "Modelle 1 - 5: Standartisierte Koeffizienten",
       custom.coef.names = c("(Intercept)", "Alter", "Bildung", "Geschlecht (0/1)"),
       float.pos = "ht!")
```

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	−0.00	0.01	−0.01	0.00	−0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Alter	0.07***			0.14***	0.14***
	(0.02)			(0.02)	(0.02)
Bildung		0.26***		0.29***	0.30***
		(0.02)		(0.02)	(0.02)
Geschlecht (0/1)			0.35***		0.36***
			(0.02)		(0.02)
R^2	0.00	0.07	0.12	0.08	0.21
Adj. R^2	0.00	0.06	0.12	0.08	0.21
Num. obs.	3064	3040	3065	3039	3039
RMSE	1.00	0.96	0.94	0.96	0.89

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

Table 2: Modelle 1 - 5: Standartisierte Koeffizienten

Aufgabe 3a

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

In Model 1 wird ersichtlich, dass mit jedem zusätzlichen Lebensjahr das Einkommen um 0.02 Einheiten steigt (statistisch hoch signifikant mit $p < 0.001$). Betrachtet man zusätzlich auch die Bildung (Model 4), wird der Koeffizient für Alter etwas größer ($b = 0.04$, $p < 0.001$). Dies deutet auf einen Surpressoreffekt hin. Inhaltlich könnte dies als Lebenszykluseffekt interpretiert werden: Insbesondere viele Personen mit Abitur befinden sich durch ein Studium während jungem Lebensalter noch in Ausbildung und haben daher ein geringeres Einkommen, als die reine Kenntnis des Bildungsabschlusses dies vorhersagen würde. Durch Hinzunahme des Geschlechts ändert sich der Koeffizient nicht sichtbar im Vergleich zu Model 4 (Model 5: $b = 0.04$, $p < 0.001$).

In Model 2 wird ersichtlich, dass mit jedem höheren Bildungsabschluss das Einkommen um 1.05 Einheiten steigt ($p < 0.001$). Durch Hinzunahme des Alters wird der Effekt betragsmäßig geringfügig größer (Model 4: $b = 1.20$, $p < 0.001$). Dies deutet ebenfalls auf den oben diskutierten Surpressor-Effekt hin. Der Koeffizient für Bildung wird geringfügig größer, wenn Geschlecht ebenfalls im Model enthalten ist (Modell 5: $b = 1.24$, $p < 0.001$).

Männer haben durchschnittlich 3.47 Einheiten mehr Einkommen as Frauen (Model 3). Der Koeffizienten wird etwas größer, wenn ebenfalls Alter und Bildung im Modell enthalten sind (Model 5: $b = 3.56$, $p < 0.001$).

Aufgabe 3b

Vergleichen Sie R^2 über die Modelle und erläutern Sie was hier festzustellen ist!

Für die drei bivariaten Modelle ist festzustellen, dass das die Kontrolle des Einflusses von Geschlecht auf das Einkommen die höchste Erklärungskraft hat, und hier 12% der Varianz des Einkommens statistisch erklärt werden können. Durch Einbezug des Bildungsstatus kann 6% der Varianz des Einkommens statistisch erklärt werden. Alter allein kann die Varianz von Einkommen nicht erklären. Für das multivariate Modell, das sowohl das Alter als auch den Bildungsstatus miteinbezieht, ist festzustellen, dass durch die Hunzunahme von Alter zum Bildungsstatus nun eine kleine Steigerung der Varianzerklärung stattgefunden hat und diese nun 8% beträgt. Das multivariate Modell, dass alle drei unabhängigen Variabeln miteinbezieht, kann 21% der Varianz des Einkommens statistisch erklären und hat somit die höchste Erklärungskraft.

Bonus:

Für die Visualisierung des Modell können wir `plot_model` aus dem `sjPlot` package benutzen.

```
lubripack::lubripack("sjPlot", silent = T)

plot_model(modell_abc,
           type = "slope", #zeigt Regressionsline und Loess-Kurve
           show.data = T) #zeigt Datenpunkte
```