SM II Abgabe 1

Maren Meinzinger

01.November 2018

Aufgabe 1

Aufgabe 1a

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

Unter Auspartialisierung versteht man, dass die Anteile bzw. der Einfluss von anderen Variablen (in einem multivariaten Modell) auf die abhängige Variable versucht wird rauszunehmen. Denn die unabhängigen Variablen können Effekte aufeinander haben, die dann die Messung einer UV auf die abhängige Variable beeinflusst, ohne dass sie es sollte. Genauer gesagt bedeutet eine Auspartialisierung die Berechnung von einem Koeffizienten und ergibt sich während der Partialisierung von multivariaten Modellen. Beim sequenziellen Vorgehen werden mehrere Modelle nacheinander spezifiziert, indem zusätzliche Variablen eingefügt werden. So sollen die Kontrolleffekte von Variablen identifiziert werden (verstärkt sich der Einfluss oder wird er geringer?). Somit können Prädiktorvariablen Einfluss auf andere Prädiktorvariablen haben und damit indirekt auch die Stärke der Messung auf die abhängige Variable haben. Deshalb sollten immer mehrere Prädiktorvariablen berücksichtigt werden, um diese „Verbindung“ zwischen den unabhängigen Variablen zu kappen und damit die Messung der abhängigen Variable genauer wird.

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?

Wie schon zuvor erklärt, können die unabhängigen Variablen untereinander korrelieren und damit Einfluss auf Y, also die abhängige Variable haben. Denn die unabhängigen Variablen, die mit anderen UV korrelieren partialisieren Anteile der UV, selbst wenn sie nicht mit Y, also der abhängigen Variable korreliert sind (bzw. keinen kausalen Effekt auf Y ausüben). Dies nennt man den Suppressoreffekt. Durch das Auspartialisieren werden nur noch die „wahren“ Effekte, also die Anteile, die mit Y korrelieren in die Messung mit eingenommen. Somit wird der „wahre“ Effekt besser gemessen.

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_sav("data/allbus2014.sav")

head(allbus)
```

2. Schritt: relevante Variablen identifizieren

- V84 ALTER: BEFRAGTE
- V86 ALLGEMEINER SCHULABSCHLUSS
- V81 GESCHLECHT, BEFRAGTE
- V420 NETTOEINKOMMEN<OFFENE+LISTENANGABE>,KAT.

```
# binoculaR(allbus)



# sollte auskommentiert sein wenn man das Dokument "knitted"
# Fehler kann ignoriert werden: object 'datatables_html' not found
```

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

Tipp: Nutze `select`

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


allb_sub
```

```
## # A tibble: 3,471 x 4
##    V84       V86       V81       V420
##    <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl>
##  1 33        5         2         14
##  2 50        3         2         " 9"
##  3 56        3         1         17
##  4 61        3         1         " 8"
##  5 59        3         2         " 9"
##  6 56        3         1         21
##  7 66        2         1         12
##  8 25        5         2         NA
##  9 58        4         1         13
## 10 53        3         1         19
## # ... with 3,461 more rows
```

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

Tipp: Nutze `rename`

```
allb_sub <- rename(allb_sub, ALTER = V84,
                   ALLGEMEINERSCHULABSCHLUSS = V86,
                   GESCHLECHT = V81,
                   NETTOEINKOMMEN = V420)

allb_sub
```

```
## # A tibble: 3,471 x 4
##    ALTER     ALLGEMEINERSCHULABSCHLUSS GESCHLECHT NETTOEINKOMMEN
##    <dbl+lbl> <dbl+lbl>                 <dbl+lbl>  <dbl+lbl>
##  1 33        5                         2          14
```

```
##  2 50        3                          2          " 9"
##  3 56        3                          1          17
##  4 61        3                          1          " 8"
##  5 59        3                          2          " 9"
##  6 56        3                          1          21
##  7 66        2                          1          12
##  8 25        5                          2          NA
##  9 58        4                          1          13
## 10 53        3                          1          19
## # ... with 3,461 more rows
```

5. Schritt: Als nächstes Rekodieren wir die Variablen

Tipp: `mutate` und `ifelse` machen die Aufgabe einfacher :)

```
mutate(allb_sub,
       alter0 = ALTER - 18,
       bildung_rec = ifelse(ALLGEMEINERSCHULABSCHLUSS == 6|
      ALLGEMEINERSCHULABSCHLUSS == 7, NA, ALLGEMEINERSCHULABSCHLUSS - 1),
      geschl_rec = ifelse(GESCHLECHT == 2, 0, 1))
```

```
## # A tibble: 3,471 x 7
##    ALTER ALLGEMEINERSCHU~ GESCHLECHT NETTOEINKOMMEN alter0 bildung_rec
##    <dbl> <dbl+lbl>        <dbl+lbl>  <dbl+lbl>      <dbl+>       <dbl>
##  1 33    5                2          14              15              4
##  2 50    3                2          " 9"            32              2
##  3 56    3                1          17              38              2
##  4 61    3                1          " 8"            43              2
##  5 59    3                2          " 9"            41              2
##  6 56    3                1          21              38              2
##  7 66    2                1          12              48              1
##  8 25    5                2          NA              " 7"            4
##  9 58    4                1          13              40              3
## 10 53    3                1          19              35              2
## # ... with 3,461 more rows, and 1 more variable: geschl_rec <dbl>
```

```
allb_sub
```

```
## # A tibble: 3,471 x 4
##    ALTER     ALLGEMEINERSCHULABSCHLUSS GESCHLECHT NETTOEINKOMMEN
##    <dbl+lbl> <dbl+lbl>                 <dbl+lbl>  <dbl+lbl>
##  1 33        5                         2          14
##  2 50        3                         2          " 9"
##  3 56        3                         1          17
##  4 61        3                         1          " 8"
##  5 59        3                         2          " 9"
##  6 56        3                         1          21
##  7 66        2                         1          12
##  8 25        5                         2          NA
##  9 58        4                         1          13
## 10 53        3                         1          19
## # ... with 3,461 more rows
```

Aufgabe 3

Berechnen Sie folgende (sequentielle) Regressionsmodelle:

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	10.18***	8.05***	16.27***	5.48***	10.51***
	(0.27)	(0.25)	(0.26)	(0.41)	(0.45)
ALTER	0.02***			0.04***	0.04***
	(0.01)			(0.01)	(0.00)
ALLGEMEINERSCHULABSCHLUSS		0.93***		1.09***	1.14***
		(0.07)		(0.07)	(0.07)
GESCHLECHT			−3.47***		−3.56***
			(0.17)		(0.16)
R^2	0.00	0.05	0.12	0.07	0.20
Adj. R^2	0.00	0.05	0.12	0.07	0.20
Num. obs.	3064	3063	3065	3062	3062
RMSE	4.95	4.83	4.65	4.78	4.44

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

Table 1: Statistical models

- *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.*

`lm` ist die Funktion für lineare Regression

```
modela <- lm(NETTOEINKOMMEN ~ ALTER, data = allb_sub)
modelb <- lm(NETTOEINKOMMEN ~ ALLGEMEINERSCHULABSCHLUSS, data = allb_sub)
modelc <- lm(NETTOEINKOMMEN ~ GESCHLECHT, data = allb_sub)
modelab <- lm(NETTOEINKOMMEN ~ ALTER + ALLGEMEINERSCHULABSCHLUSS, data = allb_sub)
modelabc <- lm(NETTOEINKOMMEN ~ ALTER + ALLGEMEINERSCHULABSCHLUSS + GESCHLECHT, data = allb_sub)

# FABIO: texreg statt htmlreg benutzen wenn man ein .pdf erstellt :)
texreg(list(modela, modelb, modelc, modelab, modelabc))
```

Aufgabe 3a

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

Da es sich nicht um standardisierte Koeffizienten handelt, können die Regressionskoeffizienten über die Modelle verglichen werden. Man sieht, dass die Regressionskoeffizienten bei der Variablen Alter sich nur minimal verändert, von 0,02 in Modell a auf 0,04 auf Modell 5. Damit kann der Effekt von Alter auf Einkommen pro Zeitraum erklärt werden. Anders gesagt, auf einer Skala von zehn Jahren nehmen die Werte von Alter auf Einkommen um 0,2 bzw. 0,4 zu. Der Effekt von Alter auf Einkommen ist positiv und zudem höchst signifikant, da p < 0,05. Durch die Auspartialisierung, also durch das Hinzunehmen der Kontrollvariablen Allgemeiner Schulabschluss wird der Effekt von Alter stärker (je älter, desto gebildeter, desto mehr Einkommen).

Die Variable „Allgemeiner Schulabschluss“ hat einen großen und positiven Effekt auf das Nettoeinkommen, der Regressionskoeffizient ist sehr stark und der Zusammenhang damit sehr stark. Die Variable ist ebenfalls stark signifikant (p < 0,001). Über die Modelle steigt der Effekt sogar noch von 0.93 auf 1.14, wenn die Kontrollvariable Geschlecht mit einbezogen wird. Das zeigt also, je gebildeter eine Person ist, desto höher ist ihr Einkommen.

Bei der Variablen Geschlecht kann ein sehr großer Effekt und eine hohe Signifikanz beobachtet werden. Die negativen Werte in der Tabelle sind so zu interpretieren, dass aufgrund der Zuordnung der Skalenwerte, der Effekt von Geschlecht auf Einkommen negativ ist, wenn die Person eine Frau ist. Andersherum wäre der Effekt also positiv, wenn es sich bei der Person um einen Mann handeln würde. Auch hier nehmen die Werte über die Modelle zu, bei Hinzunahme der anderen Kontrollvariablen (Modell abc, Einkommen auf Alter, Bildung und Geschlecht). Anders ausgedrückt, nimmt der Effekt von Geschlecht auf Einkommen über die Jahre zu, wenn es sich um eine Frau handelt.

Aufgabe 3b

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

Anhand der Tabelle sieht man, dass Rˆ2 über die Modelle größer wird. Dies ist allerdings gewöhnlich der Fall, denn je mehr Variablen es gibt, bzw. dazu kommen, desto größer wird Rˆ2. Bei Rˆ2 geht es um die Varianz. Das heißt, man teilt die ausgeschöpfte Varianz durch die zu beobachtende Varianz, um den linearen oder nicht linearen Zusammenhang zwischen x und y erklären zu können. Im vorliegenden Modell nun steigt Rˆ2 von 0,0 auf lediglich 0,20, d.h. der geschätzte Wert verbessert sich nur um 20% im Model 5. Das bedeutet, dass nur 20% der Streuung um den Mittelwert von Y kann mit Hilfe der Regressionsgeraden erklärt werden. Es handelt sich also um einen niedrigen Werte („gute“ Rˆ2 Werte sind über den Bereich von 0,7 bzw. 70%). Anders ausgedrückt, durch die Hinzunahme der Variablen Geschlecht im Modell abc wird eine bessere Erklärung der Varianz von Einkommen ermöglicht.