SM II Abgabe 1

Ilirjana Ajazaj

30.Oktober 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 wird die Bereinigung einer unabhängigen Variablen Xi um die Einflüsse der anderen unabh??ngigen Variablen. Es werden also Teile der Varianz "herausgenommen". Auspartialierung ist also eine Möglichkeit, den Einfluss von Störvariablen zu kontrollieren, sodass die Varianz der abh??ngigen Variable um jenen Anteil, den diese mit (z.B. Drittvariable oder Störvariable) gemeinsam hat, zu bereinigen.

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?

Variablen mit einem Supressoreffekt (Kontrolleffekt von Variablen) sind solche Variablen, die mit einer anderen UV und der AV korreliert sind, bzw. einen kausalen Einfluss (scheinbare non-Korrelation) auf sie aus??ben. Sie partialisieren (wechselseitig) Anteile aus, die nicht mit Y korreliert sind bzw. keinen kausalen Einfluss auf Y haben! Wird auspartialisiert, so gehen mehr Anteile von X1 in die Regression ein, die mit Y korrelieren bzw. einen kausalen Zusammenhang auf Y aus??ben. Der "wahre" Effekt von X1 (und X2) wird (zu einem gr????eren Anteil) freigelegt. Bei einer "scheinbaren Nonkorrelation" besteht kein bivariater Zusammenhang zwischen X und Y. Kontrolliert man durch eine Drittvariable, dann erscheint ein substanzieller Zusammenhang, der durch die Supressorvariable verdeckt wurde.

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) # zum Datensatz inspizieren :)

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

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

Tipp: Nutze `rename`

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

5. Schritt: Als nächstes Rekodieren wir die Variablen

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

```
allb_a <- mutate (allb_a, alter0 = alter - 18, bildung_rec = ifelse (bildung == 6 | bildung == 7, NA, b

allb_a
```

```
## # A tibble: 3,471 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
##  7 66        2         1         12        48                1          1
##  8 25        5         2         NA        " 7"              4          0
##  9 58        4         1         13        40                3          1
## 10 53        3         1         19        35                2          1
## # ... with 3,461 more rows
```

Bonus: Alles mit dem pipe operator %>%

```
allb_a <- 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_a)
```

```
## # 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.*

`lm` ist die Funktion für lineare Regression

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

texreg(list (modell_a,
             modell_b,
             modell_c,
             modell_ab,
             modell_abc),
          caption = "Modelle mit Unstandardisierte Koeffizienten",
       custom.coef.names = c ("Intercept", "Alter", "Bildung", "Geschlecht"), 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			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 mit Unstandardisierte Koeffizienten

Aufgabe 3a

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

Hier handelt es sich um unstandatisierte b-Werte (empirische Einheit, Intercept vorhanden, b-Werte über 1). Im ersten Modell (Modell 1) wurde Einkommen auf Alter untersucht. Es zeigt sich, dass für jedes weitere (Lebens-)Jahr (x) das EInkommen (y) um 0.02 (b-Wert) Einheiten steigt und ist auf dem 99%igem Signifikanzniveau hoch signifikant. Der Einfluss von Alter wird auch im Modell 4 und Modell 5 mit Bildung und Geschlecht aufgenommen und hat einen b-Wert von 0,04, der höher als im ersten Modell (Modell 1) zu beobachten ist und ebenfalls auf dem 99%igem Signifikanzniveau hoch signifikant ist. Der Anstieg des b-Wertes ist jedoch mit Vorsicht zu interpretieren, da es sich hierbei möglicherweise auf einen Supressoreffekt handelt. Alter hat je Bildungsgruppe unterschiedliche Wirkungsverläufe und daher erst unter Kontrolle von Bilung, der um diese kontrollierte Altereffekt sichbar gemacht wird. So verringert sich in Alter der Kohorteneffekt (z.B. Rentner) und der proxy für Berufserfahrung verstärt sich unter Kontrolle von Bildung.

Im zweiten Modell (Modell 2) wurde der Effekt von Einkommen auf Bildung berechnet. Hier ist zu beobachten, dass mit jedem höheren Bildungsabschluss das Einkommen um 1,05 Einheiten steigt. Die Bildungsvariable wurde auch im Modell 4 (mit einem b-Wert von 1,20) mit Alter und 5 (mit einem b-Wert von 1,24) mit Geschlecht mit aufgenommen.

Im dritten Modell (Modell 3) wurde Einkommen auf Geschlecht untersucht. Hier ist festzustellen, dass Männer durchschnittlich um 3.47 Einheiten mehr Einkommen erhalten als Frauen. Geschlecht wird auch im Modell 5 mit aufgenommen, neben Alter und Bildung und weist einen höheren Koeffizienten als im Modell 3.

Aufgabe 3b

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

R^2 ist ein Korrelationskoeffizient mit Wertebereich von [0;1] und gibt Information über Anteil der durch X erklärte Varianz. Betrachten man die Modell 1-3 so ist zu beobachten, dass Modell 3 (Zusammehnag zwischen Geschlecht und Einkommen) am stärksten ist: Durch Kenntnis von Geschlecht lassen sich 12% (r2=0,12) der Varianz von Einkommen statistisch erklären. Durch Bildung lassen sich 7% der Varianz des Einkommens statistisch erklären, durch Alter jedoch kann die Varianz des Einkommens nicht erklärt werden. Vergleicht man die Modelle 4 und 5, so ist festzustellen, dass Modell 5 mit den Variablen Alter, Bildung und Geschlecht 21% der Varianz des Einkommens statistisch erkören können, währen im Modell 4 Alter und Bildung nur 8% der Varianz des Einkommens ausschöpfen kann.