SM II Abgabe 1

Anna Götze

27. 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?

In ein Regressionsmodell mit einer unabhängigen Variable werden mehr unabhängige Variablen hinzugenommen: der Einfluss von X auf Y wird um die Einflüsse anderer Variablen (= derer, die hinzugenommen werden) bereinigt. Dadurch nähert sich der Wert der Regressionskoeffizieten der Variablen X (X1, X2, X3...) dem 'wahren' Wert des Einflusses der Variable.

Mehrere Prädiktorvariablen sollten berücksichtigt werden, um sicherzugehen, dass der Wert des Regressionskoeffizieten der interessierenden unabhängigen Variable so wenig verfälscht wie möglich ist.

Aufgabe 1b

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

Wenn im bivariaten Modell der Zusammenhang nicht signifikant ist, kann das daran liegen, dass der Einfluss von X auf Y von einer anderen Variable, die im Modell nicht berücksichtigt wurde, verdeckt wird. Es kann also doch eine Korrelation bestehen, die nur versteckt ist. Deshalb ist es wichtig, multiple Modelle aufzustellen, um solche Zusammenhänge aufzudecken.

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("C:/Users/Anna/Documents/Uni S/Statistische-Modellbildung-II-WS1819-master/data/allbus2014.sav")
```

2. Schritt: relevante Variablen identifizieren

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

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

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

# ist ein Subset vom Datensatz

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!

```
allb_sub <- rename(allb_sub, Alter = V84,

               Bildung = V86,

               Geschlecht = V81,

               Einkommen = V420)

allb_sub
```

Alter <S3: labelled>	Bildung <S3: labelled>	Geschlecht <S3: labelled>	Einkommen <S3: labelled>	alter0 <S3: labelled>	bildung_rec <dbl>	geschlecht_rec <dbl>
33	5	2	14	15	4	0
50	3	2	9	32	2	0
56	3	1	17	38	2	1
61	3	1	8	43	2	1
59	3	2	9	41	2	0
56	3	1	21	38	2	1
66	2	1	12	48	1	1
25	5	2	*NA*	7	4	0
58	4	1	13	40	3	1
53	3	1	19	35	2	1

Next 123456 ... 100 Previous 1-10 of 3,471 rows

5. Schritt: Als nächstes Rekodieren wir die Variablen

```
allb_sub <- mutate(allb_sub,
                 # Alter Rekodieren auf Startpunkt 0
                 alter0 = Alter - 18,
                 # Bildung Rekodieren, 6 und 7 auf "Missing" setzen und Bildung - 1
                 bildung_rec = ifelse(Bildung == 6 | Bildung == 7, NA, Bildung - 1),
                 # Geschlecht umkodieren: wenn 2, dann 0, sonst 1-
                 geschlecht_rec = ifelse(Geschlecht == 2, 0, 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: Einkommen auf Alter;*
modell_a <- lm(Einkommen ~ alter0, data = allb_sub)
screenreg(modell_a)
#       *Modell b: Einkommen auf Bildung;*
modell_b <- lm(Einkommen ~ bildung_rec, data = allb_sub)
#       *Modell c: Einkommen auf Geschlecht;*
modell_c <- lm(Einkommen ~ geschlecht_rec, data = allb_sub)
#       *Modell ab: Einkommen auf Alter und Bildung;*
modell_ab <- lm(Einkommen ~ alter0 + bildung_rec, data = allb_sub)
#       *Modell abc: Einkommen auf Alter, Bildung und Geschlecht.*
modell_abc <- lm(Einkommen ~ alter0 + bildung_rec + geschlecht_rec, data = allb_sub)

#Modelle anzeigen
texreg(list(modell_a,
          modell_b,
          modell_c,
          modell_ab,
          modell_abc))
```

	Modell a	Modell b	Modell c	Modell ab	Modell abc
Intercept	*10.53 (***)* *(0.19)*	*8.75 (***)* *(0.19)*	*9.33 (***)* *(0.12)*	*7.17 (***)* *(0.28)*	*5.15 (***)* *(0.28)*
Alter-18	0.02 (***) (0.01)			0.04 (***) (0.01)	0.04 (***) (0.00)
Bildung		1.05 (***) (0.07)		1.20 (***) (0.07)	1.24 (***) (0.07)
Geschlecht			3.47 (***) (0.17)		3.56 (***) (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
n	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

Aufgabe 3a

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

- Alter: Zuerst ein Wert von 0,02; unter Hinzunahme weiterer Variablen aber 0,04
- Bildung: Einfluss ist ebenfalls in multivariaten Modellen höher
- Geschlecht: Einfluss ist ebenfalls im multivariaten Modell höher

→ Für alle 3 unabhängigen Variablen wird in den multivariaten Modellen ein höherer Einfluss berechnet als im bivariaten Modell. Da bei multivariaten Modellen mehrere Variablen gleichzeitig berücksichtigt werden können, sind diese genauer und „besser“ als bivariate Modelle. Die unterschiedlichen Regressionskoeffizienten in bi- und multivariaten Modellen zeigen, dass die bivariaten hier nicht geeignet sind, um den Einfluss von Variablen zu berechnen.

Aufgabe 3b

Vergleichen Sie R2 über die Modelle und erläutern Sie was hier festzustellen ist!

Modell abc weist das höchste R^2 auf, hat jedoch auch am meisten Variablen. Deshalb muss das bereinigte, korrigierte R^2 (Adjusted R^2) betrachtet werden, das verwendet wird, um Modelle mit verschieden vielen Variablen zu vergleichen: auch das adjusted R^2 ist bei Modell abc am höchsten (21%), währen es bei Modell a 0,00, bei Modell b 0,06, bei Modell c 0,12 und bei Modell ab 0,08 beträgt. Die Variation der unabhängigen Variablen Alter, Bildung und Geschlecht erklärt also die Variation der abhängigen Variablen zu 21%.