Musterlösung: Übungsaufgabe 5

Statistische Modellbildung II

30.November 2017

Aufgabe 1

Erstellen Sie eine Regression von Einkommen auf Bildung, Geschlecht und Alter sowie der Dummyvariablen Zugang zu tertiärer Bildung (bild_tert), die null kodiert ist, wenn der betreffende Befragte einen niedrigeren Schulabschluss als Fachhochschulreife hat und eins, wenn Umgekehrtes der Fall ist. Hinzu kommen die Interaktionsvariablen zwischen Geschlecht und Alter (gesch_alter) sowie zwischen Alter und Zugang zu tertiärer Bildung (alt_tert).

- 0 OHNE ABSCHLUSS
- 1 VOLKS-,HAUPTSCHULE
- 2 MITTLERE REIFE
- 3 FACHHOCHSCHULREIFE
- 4 HOCHSCHULREIFE

```
allb_sub <- allb_sub %>%
  mutate(bild_tert = ifelse(bildung_rec > 2, 1 ,0))

fit1 <- lm(einkommen ~ bildung_rec + geschl_rec +
               alter0 + bild_tert +
               geschl_rec * alter0 +
               alter0 * bild_tert, data = allb_sub)

texreg(fit1, float.pos = "!h")
```

	Model 1
(Intercept)	6.88***
	(0.41)
bildung_rec	1.06***
	(0.16)
geschl_rec	2.55***
	(0.33)
alter0	−0.00
	(0.01)
bild_tert	−1.84***
	(0.47)
geschl_rec:alter0	0.03**
	(0.01)
alter0:bild_tert	0.08***
	(0.01)
R^2	0.23
Adj. R^2	0.23
Num. obs.	3039
RMSE	4.35

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

Table 1: Statistical models

Aufgabe 1a

Berechnen Sie das Konfidenzintervall für die Variablen bild_tert und Alter mittels der Koeffizienten und interpretieren Sie diese.

$$KI_{95} = b \pm t_{df} \times SE_b$$

Für $df > 120$ und 95% Signifikanzniveau ist der kritische Wert $t = 1.96$

Für $df > 120$ und 99% Signifikanzniveau ist der kritische Wert $t = 2.58$

```
# confint(fit1) #Konfidenzintervalle anzeigen lassen

coefs <- tidy(fit1) %>%
        filter(term == "bild_tert" | term == "alter0") %>%
        mutate(estimate = round(estimate,3),
               std.error = round(std.error,3)) %>%
        select(term, estimate, std.error)



coefs$low.se.95 <- round(coefs$estimate - 1.96 * coefs$std.error, 3)
coefs$high.se.95 <- round(coefs$estimate + 1.96 * coefs$std.error, 3)

coefs$low.se.99 <- round(coefs$estimate - 2.58 * coefs$std.error, 3)
coefs$high.se.99 <- round(coefs$estimate + 2.58 * coefs$std.error, 3)

kable(coefs)
```

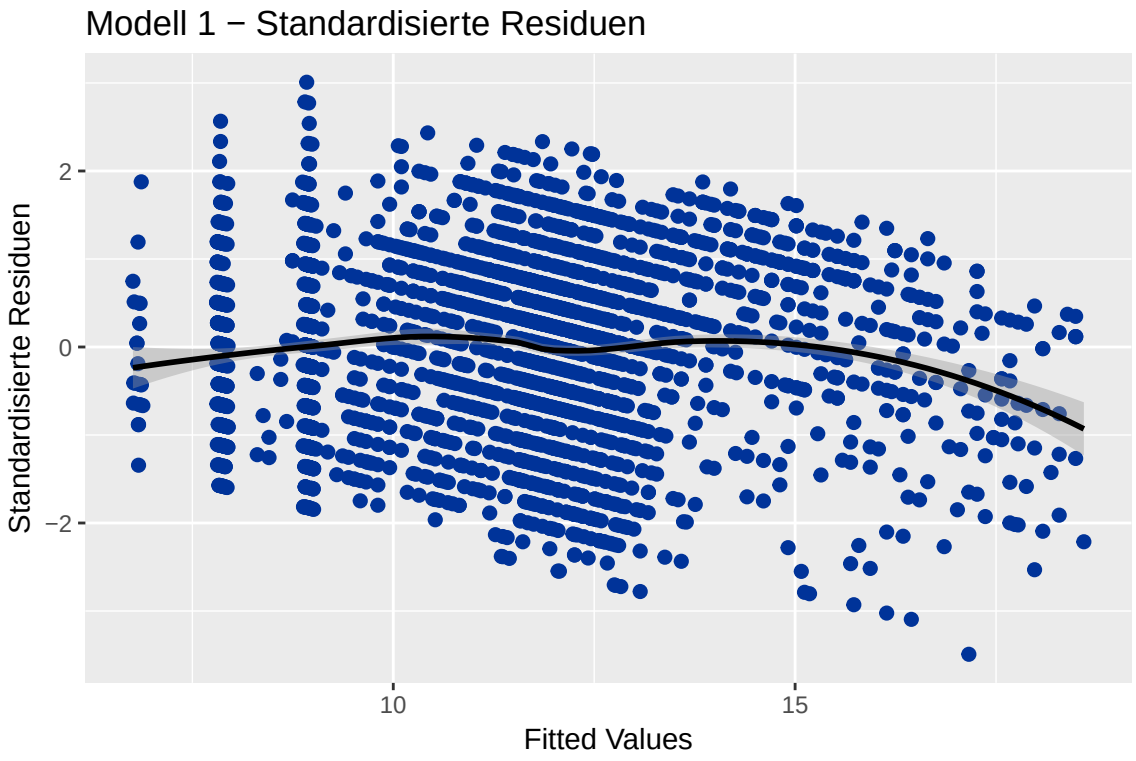
term	estimate	std.error	low.se.95	high.se.95	low.se.99	high.se.99
alter0	-0.002	0.008	-0.018	0.014	-0.023	0.019
bild_tert	-1.836	0.468	-2.753	-0.919	-3.043	-0.629

Aufgabe 1b

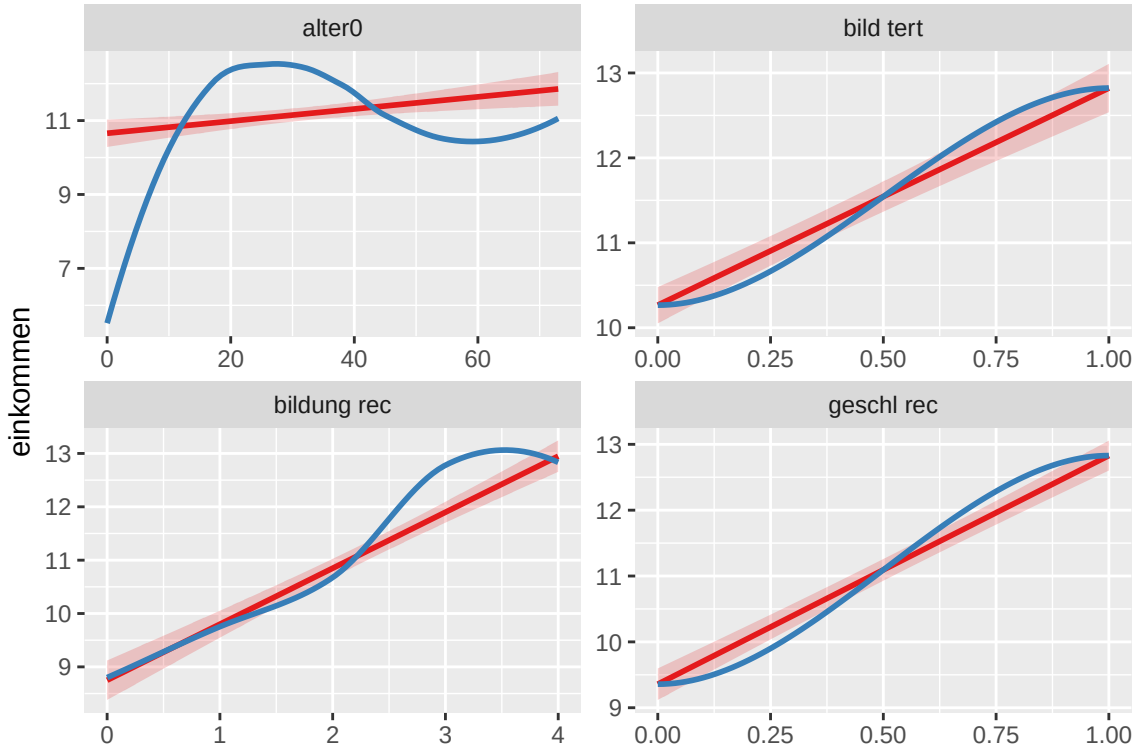
Testen Sie das Gesamtmodell auf Linearität.

```
diag_mod <- augment(fit1)

diag_mod %>% #Datensatz einladen
  select(.fitted, .std.resid) %>% #Variablen auswählen (x als erstes, y als zweites)
  sjplot(fun = "scatter",      #definiere den Plot als Scatterplot
         jitter.dots = T,      #füge jitter hinzu (Streuung der Punkte)
         fit.line = T,         #zeige eine Regressionsgerade
         fitmethod = "loess", #"loess" statt "lm"
         show.ci = T,          #zeige zusätzliche das Konfidenzintervall
         title = "Modell 1 - Standardisierte Residuen",   #Titel der Grafik
         axis.titles = c("Fitted Values",
                         "Standardisierte Residuen")) #Ein Vektor mit den Namen der Achsen
```




```
plot_model(fit1, type = "slope")
```



Durch eine visuelle Untersuchung der Residuen und der vorhergesagten Werte und einer Loess-Kurve kann vermutet werden, dass ein nicht-linearer Zusammenhang besteht. Diese Nicht-Linearität kommt höchstwahrscheinlich von dem quadratischen Effekt von Alter.

Aufgabe 2

Was ist unter Multikollinearität zu verstehen, warum ist es ein Problem, wenn diese in einer Modellschätzung vorliegt und wie kann das Vorliegen derselben diagnostiziert werden?

Eine Multikollinerarität liegt vor, wenn unabhängige Variablen untereinander korrelieren. Problematisch ist das, weil bei einer zu hohen Korrelation zwischen den unabhängigen Variablen die Effekte nicht mehr auf einzelne unabhängige Variablen zurückzuführen sind und die b-Koeffizienten sowie Standardfehler stark verzerrt werden. Ob Multikollinearität vorliegt, kann man zum Beispiel durch das Erstellen einer Korrelationsmatrix, in welcher alle bivariaten Zusammenhänge aller unabhängigen Variablen berechnet werden, überprüfen. Eine weitere Möglichkeit ist die Berechnung der Toleranzwerte (oder VIF-Werte) für die X-Variablen. So wird angegeben, wie hoch die eigenständige Erklärungskraft einer X-Variable ist, also wie viel Anteil an Varianz sie selbst erklären kann. Sinkt dieser Wert unter 0,2 (über 5 VIF), hat die Variable wenig Eigenerklärungskraft.

Aufgabe 3

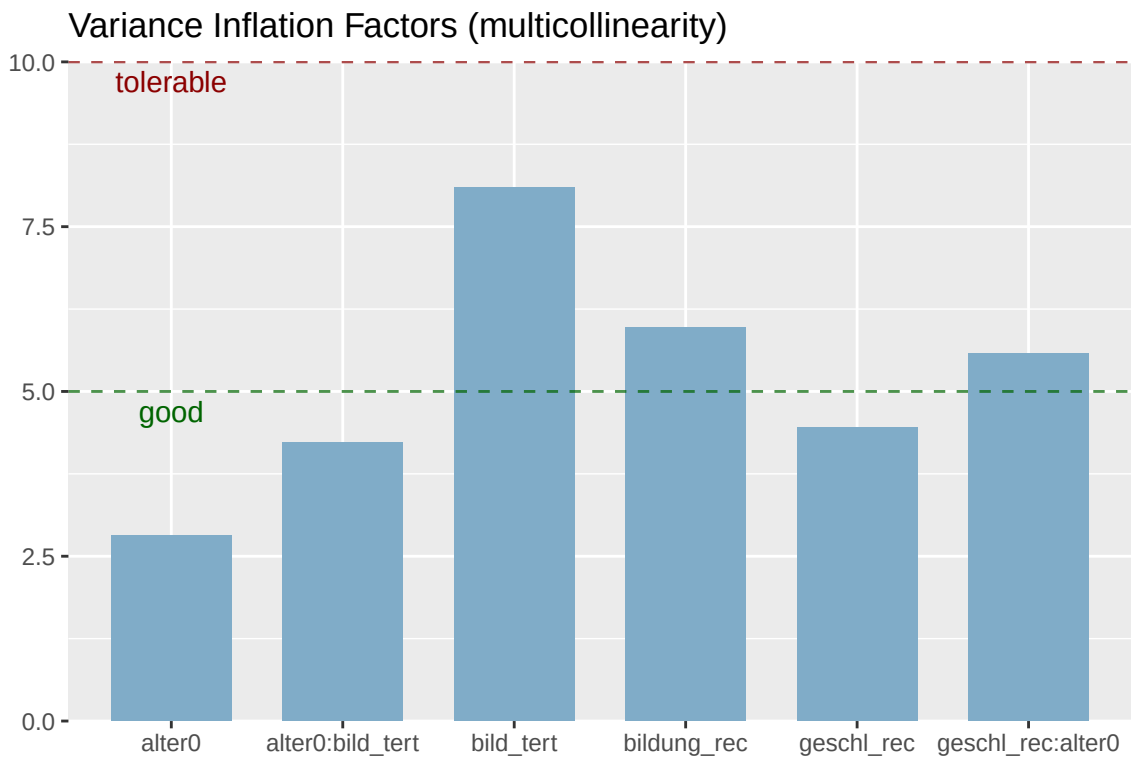
Wie ausgeprägt ist die Multikollinearität im Regressionsmodell von Aufgabe 1? Welche Gründe (inhaltliche) lassen sich für die Multikollinearität identifizieren?

```
(1/vif(fit1)) %>%
  tidy() %>%
  kable()
```

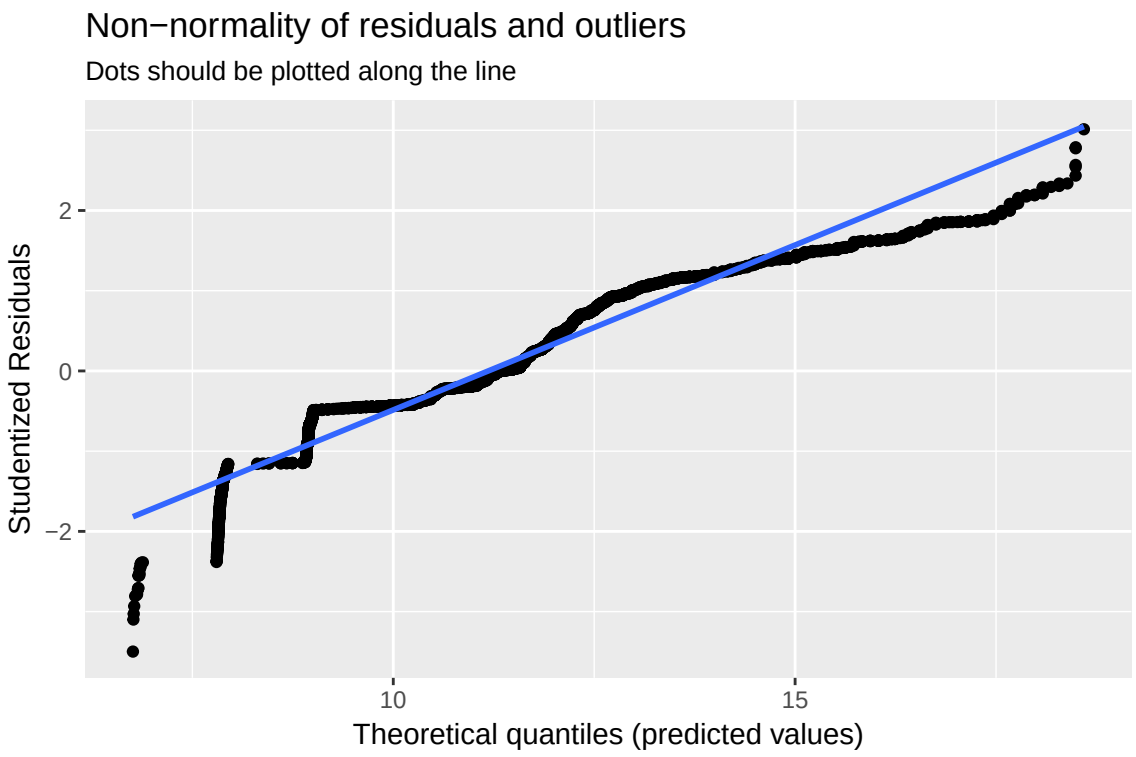
names	x
bildung_rec	0.1674532
geschl_rec	0.2243787
alter0	0.3554997
bild_tert	0.1233993
geschl_rec:alter0	0.1791200
alter0:bild_tert	0.2361686

```
plot_model(fit1, type = "diag")
```

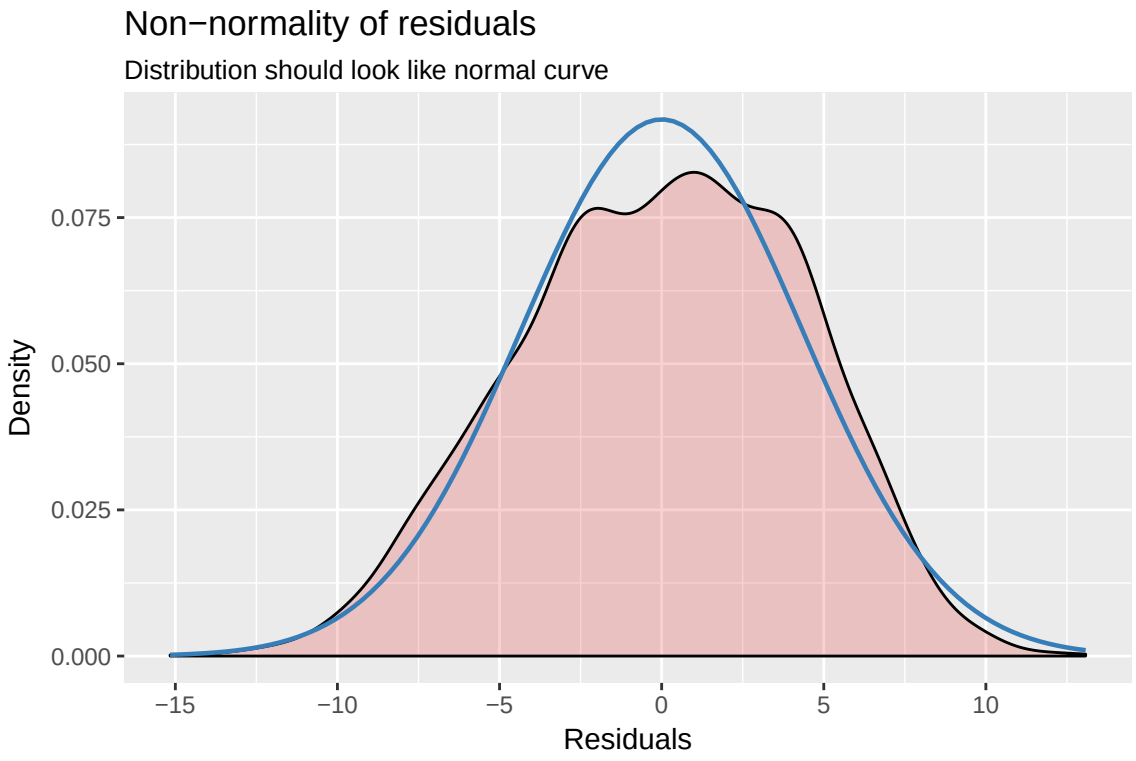
[[1]]


Variance Inflation Factors (multicollinearity)
10.0
7.5
5.0
2.5
0.0
tolerable
good
alter0
alter0:bild_tert
bild_tert
bildung_rec
geschl_rec
geschl_rec:alter0

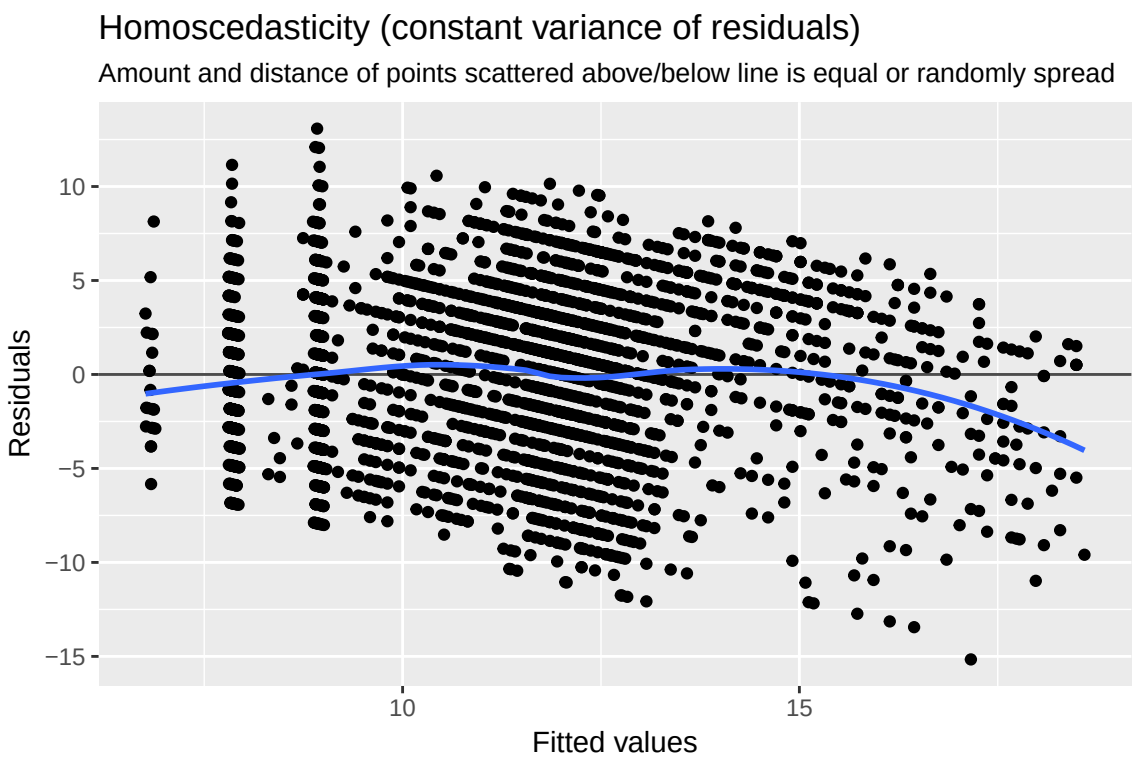
[[2]]


Non−normality of residuals and outliers
Dots should be plotted along the line
Studentized Residuals
2
0
−2
10
15
Theoretical quantiles (predicted values)

[[3]]



[[4]]



Die VIF-Werte sollten den Grenzwert von 5 nicht überschreiten (bzw. Toleranzwerte über 0,2). Dies tun allerdings sowohl die eigentliche Bildungs-Variable, als auch die Dummy-Bildungs-Variable, als auch die Interaktionsvariable von Geschlecht und Alter.

Die Multikollinearität kann in diesem Modell ganz klar darauf zurück geführt werden, dass die Dummy-Variable `bild_tert` sich mit der Bildungsvariable informationstechnisch überschneidet sowie Interaktionsvariablen mit aufgenommen wurden.