Musterlösung: Übungsaufgabe 4

Statistische Modellbildung II

23.November 2017

Aufgabe 1

Erstellen Sie ein Streudiagramm zwischen den Residuen der Regression von "prestige_befragter" auf "bildung_befragter" und der Variablen "bildung_befragter". Beschreiben Sie das Diagramm, sprich wie die fünf "Streusäulen" in Relation zueinander aussehen. Äußern Sie eine Vermutung ob dies bereits eine Verletzung der A1 ist und begründen Sie diese kurz

Eine der Prestigeskalen auswählen:

- Index für den Prestige-Rang von Berufen (SIOPS)

```
allb_sub <- allb_sub %>%
  rename(prestige_befragter = V112)

mod1 <- lm(prestige_befragter ~ bildung_rec, data = allb_sub)
texreg(mod1)
```

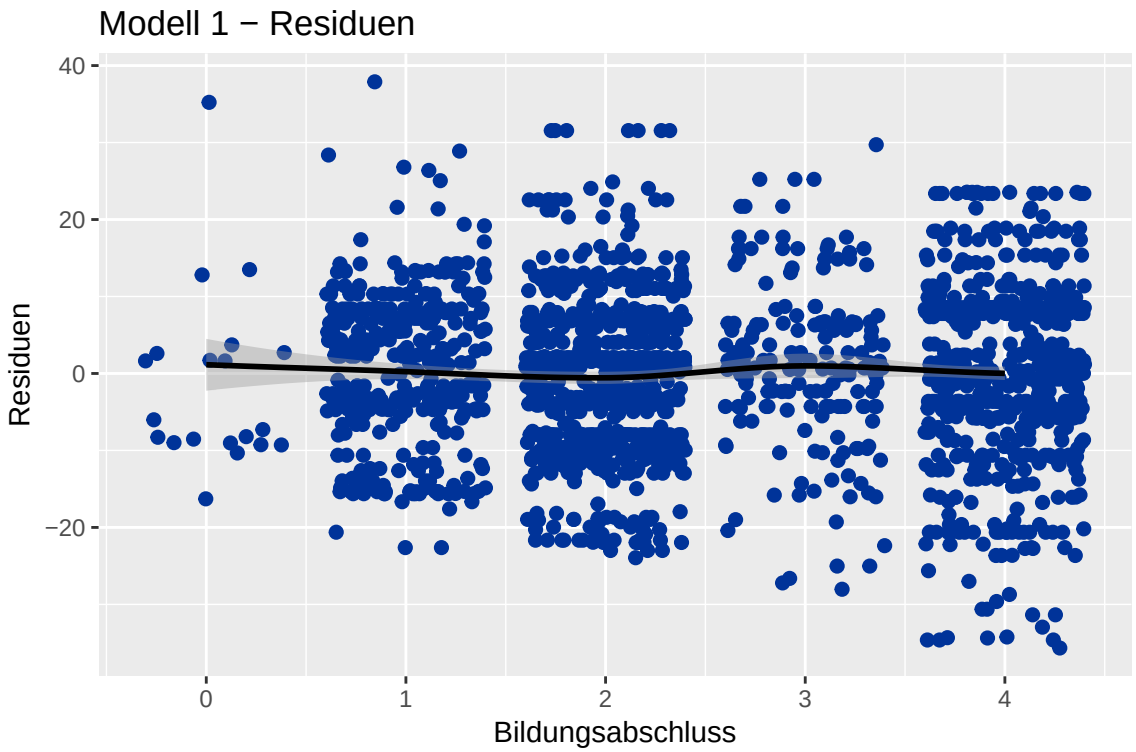
```
diag_mod <- augment(mod1)

diag_mod %>% #Datensatz einladen
  select(bildung_rec, .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 - Residuen",   #Titel der Grafik
         axis.titles = c("Bildungsabschluss",
                         "Residuen")) #Ein Vektor mit den Namen der Achsen
```

	Model 1
(Intercept)	29.29***
	(0.60)
bildung_rec	6.33***
	(0.22)
R^2	0.31
Adj. R^2	0.31
Num. obs.	1872
RMSE	10.83

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

Table 1: Statistical models



Es wird aus dem Streudiagramm ersichtlich, dass die Varianz der Residuen mit zunehmendem X tendenziell größer wird. Für die 0-Kategorie sind zu wenige Fälle vorhanden, um über die Streung der Residuen aussagekräftige Schlüsse zu ziehen. Die Kategorien 1 und 2 weisen relativ breite, aber einander ähnliche Streuungen auf. Für die 3-Kategorie ist eher weniger Varianz zu beobachten, für die 4-Kategorie scheint die Streuung am breitesten. Insbesondere die im Vergleich zu den anderen Gruppen geringe Streeung der Residuen der 3-Kategorie sind für die A1 als problematisch zu betrachten,weswegen eine Verletzung angenommen werden kann. Um mehr Sicherheit diesbezüglich zu erlangen, wäre ein Levene-Test zu empfehlen.

Aufgabe 2

Erstellen Sie eine Regression von Einkommen auf Alter (Alter_0) und speichern Sie dabei die unstandardisierten sowie die standardisierte Residuen aus. Erstellen Sie anschließend ein Streudiagramm für die beiden unterschiedlichen Residuenarten. Die UV soll hierbei das Alter der Befragten sein.

```
mod2 <- lm(einkommen ~ alter0, data = allb_sub)
texreg(mod2)
```

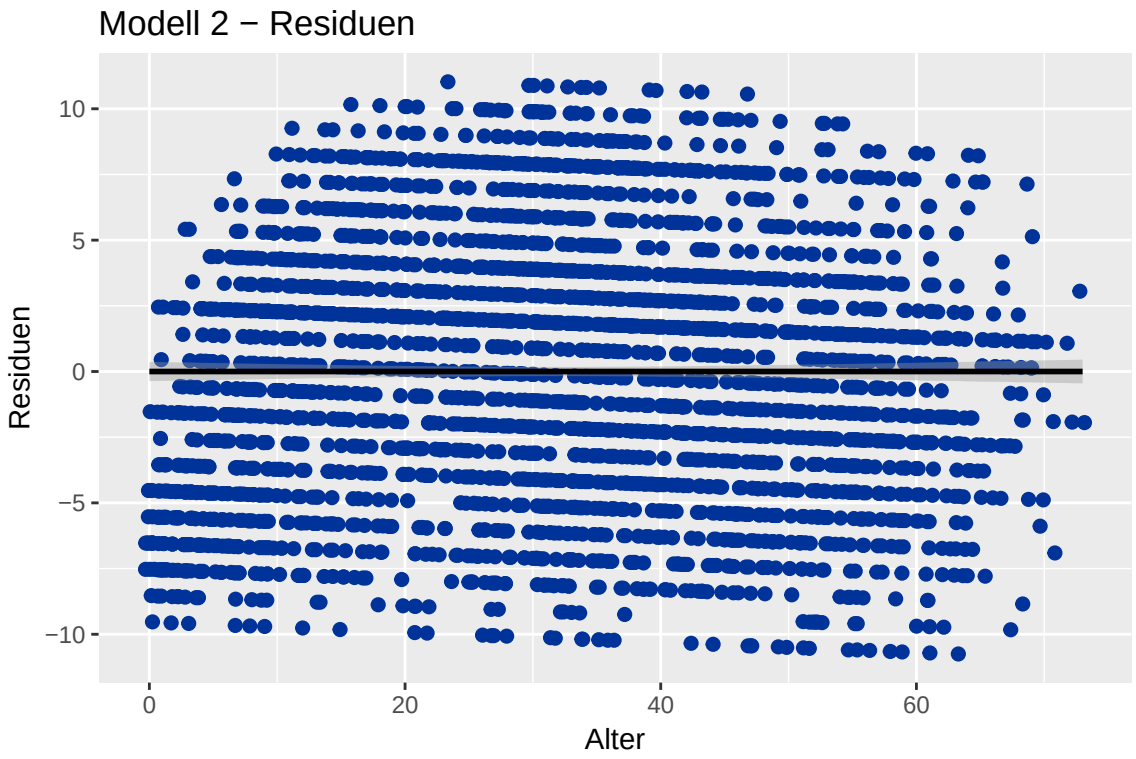
```
diag_mod2 <- augment(mod2)

diag_mod2 %>% #Datensatz einladen
  select(alter0, .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 = "lm",   #"loess" statt "lm"
         show.ci = T,        #zeige zusätzliche das Konfidenzintervall
         title = "Modell 2 - Residuen",    #Titel der Grafik
         axis.titles = c("Alter",
                         "Residuen"))
```

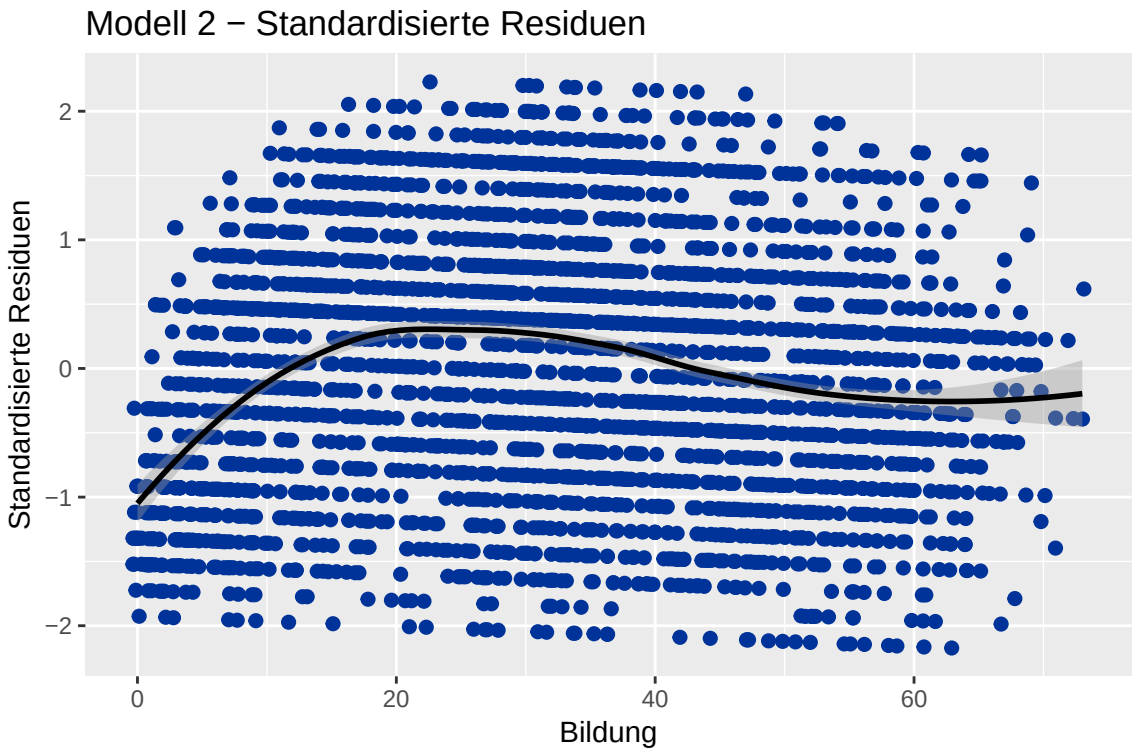
	Model 1
(Intercept)	10.53***
	(0.19)
alter0	0.02***
	(0.01)
R^2	0.00
Adj. R^2	0.00
Num. obs.	3064
RMSE	4.95

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

Table 2: Statistical models




```
diag_mod2 %>% #Datensatz einladen
  select(alter0, .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 2 - Standardisierte Residuen",    #Titel der Grafik
         axis.titles = c("Bildung",
                         "Standardisierte Residuen"))
```



Aufgabe 2a

Was ist der Unterschied zwischen den beiden Residuenarten und wie werden die standardisierten Residuen gebildet?

Die standardisierten Residuen sind z-standardisierte Residuen, zeigen also die standardisierten Abstände zur Regressionsgeraden, während die unstandardisierten Residuen absolute Fehlerwerte sind.

Aufgabe 2b

Beschreiben Sie die beiden Streudiagramme. Gibt es Unterschiede, falls nein, hängt dies mit der Art der Erstellung der Variablen zusammen? Beziehen Sie sich bei Ihrer Antwort auf die statistischen Kenngrößen, die mittels FREQ-Befehl ausgegeben werden.

```
diag_mod2 %>%
  select(.resid, .std.resid) %>%
  describe() %>%
  select(n, mean, sd, skew , kurtosis) %>%
  kable()
```

	n	mean	sd	skew	kurtosis
.resid	3064	0.000000	4.949657	0.0433585	-0.8645087
.std.resid	3064	-0.000042	1.000148	0.0432203	-0.8645970

Visuell gibt es keine nennenswerten unterschiede zwischen den beiden Streudiagrammen. Aufgrund der Standardisierung kann die Varianzstreuung in der Regel besser eingeschätzt werden, weil die Abstände nicht absolut sind, sondern standardisiert. Beide Residuenarten zeigen einen Mittelwert von 0, weil Residuen die

Abstände zur Regressionsgeraden aufzeigen. Auch bei unstandardisierten Residuen ist der Mittelwert 0, allerdings unterscheidet sich die Standardabweichung.

Aufgabe 3

Anhand der Scatterplots (Streudiagramme) lässt sich vermuten, dass zwischen verschiedenen Altersgruppen Varianzinhomogenität (Verletzung A1) besteht.

Aufgabe 3a

Erläutern Sie das Prinzip nach welchem der Test funktioniert, der angewendet werden muss, um zu prüfen, ob die identifizierten Gruppen Varianzhomogenität aufweisen.

Zunächst ist eine visuelle Überprüfung der Varianz der Residuen zu empfehlen. Nachdem man bei der ÜerprÜfung des Streudiagramms eine Verletzung der A1 Annahme vermutet, sollte man den sogenannten *Levene-Test* anwenden.

Für einen Levene-Test werden zunÄchst Gruppen gebildet, welche die Bereiche der unterschiedlichen Streuungen möglichst genau abgrenzen sollen. Mit diesen Variablen wird der Levene-Test durchgeführt. Es wird untersucht, ob die Varianz innerhalb der Gruppen signifikant voneinander abweicht. Wird das Testergebnis also signifikant, so kann von einer signifikanten Abweichung der Varianzen in den verschiedenen Gruppen ausgegangen werden (Heteroskedastizität) und so kann die Annahme 1 als verletzt gesehen werden.

Aufgabe 3b

Testen Sie die Annahme A3 und A5 (Test der Gesamtheit der Residuen) für die Residuen von Aufgabe 3. Sind die Annahmen erfüllt? Welche Einschränkungen müssen bzgl. des Tests für A5 beachtet werden?

Test von Annahme 3

Zuerst eine absolute Residuenvariable erstellen:

```
diag_mod2 <- diag_mod2 %>%
  mutate(.std.resid.abs = abs(.std.resid))
```

Als nächstes eine Korrelationstabelle erstellen:

```
diag_mod2 %>%
  select(alter0, .std.resid.abs) %>%
  corstar() %>%
  kable()
```

	alter0	.std.resid.abs
alter0	1.000***	-0.045*
	(0.000)	(0.012)
.std.resid.abs	-0.045*	1.000***
	(0.012)	(0.000)

Alternativ:

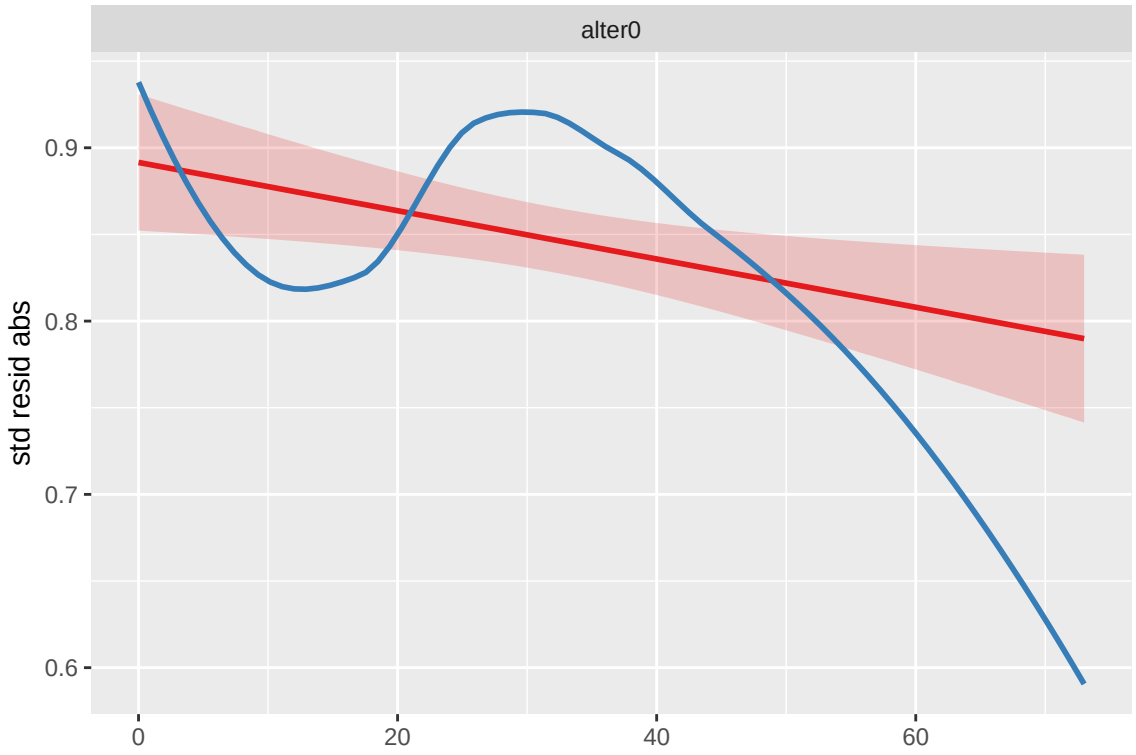
```
resid_mod <- lm(.std.resid.abs ~ alter0, data = diag_mod2)
texreg(resid_mod)
```

	Model 1
(Intercept)	0.89***
	(0.02)
alter0	−0.00*
	(0.00)
R^2	0.00
Adj. R^2	0.00
Num. obs.	3064
RMSE	0.53

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

Table 5: Statistical models

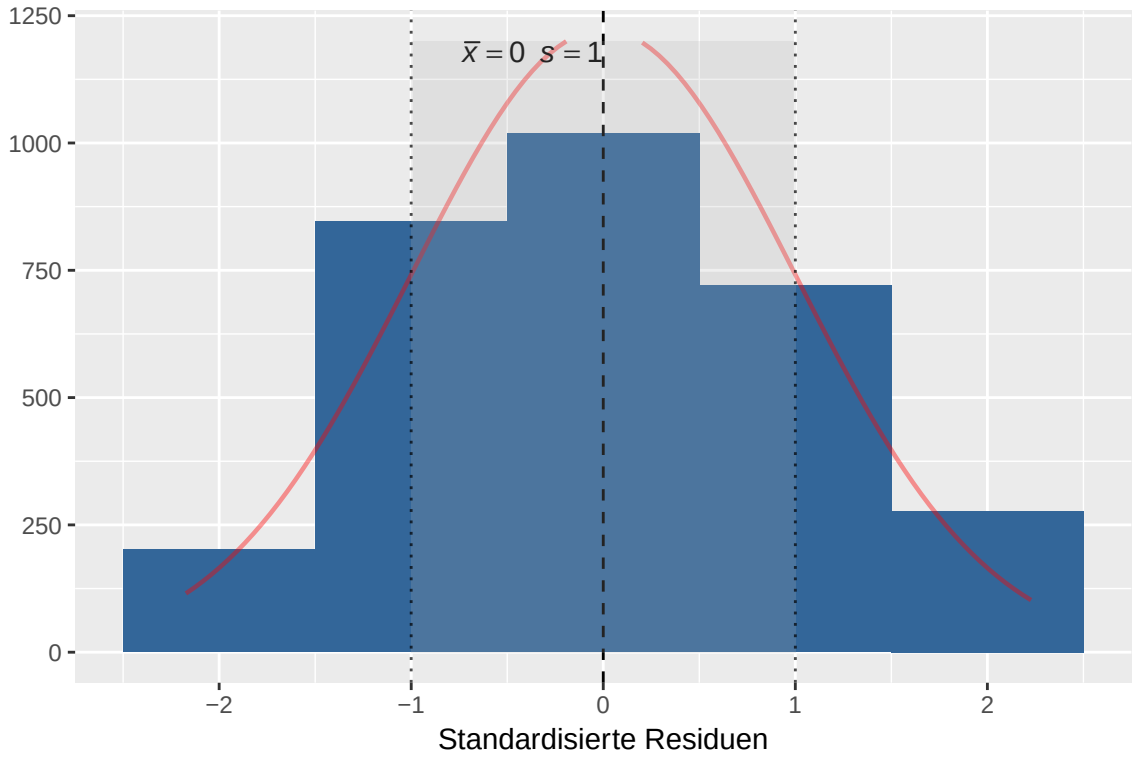
```
plot_model(resid_mod)
```



Die Standardisierten Residuen der Regression von Einkommen auf alter0 sind signifikant mit alter0 korreliert. Allerdings ist diese Korrelation relativ gering, daher kann die A3 als nicht verletzt gelten. Eine genauere Inspizion des Sachverhalts wäre angebracht.

Test von Annahme 5

```
diag_mod2 %>%
  select(.std.resid) %>%
  sjplot(fun = "frq",         #univariate Zähldaten
         type = "hist",       #Histogramm
         normal.curve = TRUE, #Normalverteilung
         show.mean = TRUE,    #Mittelwert darstellen
         axis.title = "Standardisierte Residuen") #Namen für x-Achse
```




```
diag_mod2 %>%
  select(.resid, .std.resid) %>%
  describe() %>%
  select(n, mean, sd, skew , kurtosis) %>%
  kable()
```

	n	mean	sd	skew	kurtosis
.resid	3064	0.000000	4.949657	0.0433585	-0.8645087
.std.resid	3064	-0.000042	1.000148	0.0432203	-0.8645970

```
shapiro.test(diag_mod2$.std.resid)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diag_mod2$.std.resid
## W = 0.98205, p-value < 0.00000000000000022
```

Die Sichtung des Histogramms lässt erkennen, dass die Residuenverteilung annäherungsweise einer Normalverteilung ähnelt. Betrachtet man Schiefe und Steilheit liegen diese innerhalb des Intervalls von +/- 1. Auf dieser Grund wird A5 nicht verworfen. Ein Shapiro-Wilk Normality Test wird signifikant ($p < 0.05$), was bedeutet, dass die Residuenverteilung signifikant von einer Normalverteilung abweicht, weswegen A5 als nicht erfüllt angesehen werden müsste. Da die visuelle Inspektion und Skewness/Kurtosis nicht auf eine Verletzung der A5 hinweisen, wird allerdings kein Verstoß angenommen.