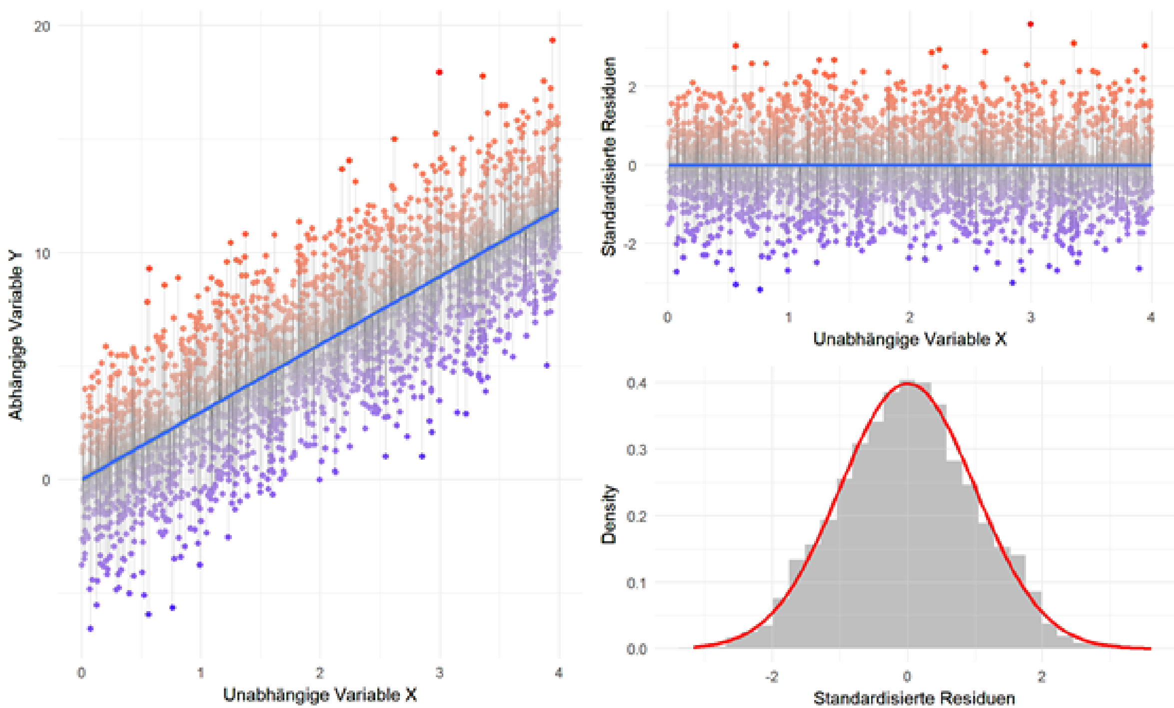
Universität Stuttgart
Abteilung IV (Prof. Urban)
Soziologie und empirische Sozialforschung

SM II: Tutorium - 5. Sitzung



Kontakt:

Fabio Votta

favstats
@favstats
www.favstats.eu
fabio.votta@gmail.com

2018-22-11

Übersicht

1. Übungsaufgabe 4 in SPSS
2. Übungsaufgabe 4 in R
3. Bestimmung des benötigten Stichprobenumfangs

Tabelle A4 zum F-Test (Urban/Mayerl, S. 358-359)

Multikollinearität Blogpost

Link zum Datensatz

Link zur Übungsaufgabe 4 - SPSS

Link zur Übungsaufgabe 4 - R

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).

Aufgabe 1a

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

$$KI_{95} = b \pm t_n \times SE_b$$

Für $n > 120$ und 95% Signifikanzniveau ist der kritische Wert $t_{krit} = 1.96$

Für $n > 120$ und 99% Signifikanzniveau ist der kritische Wert $t_{krit} = 2.58$

Aufgabe 1b

Testen Sie das Gesamtmodell auf Linearität.

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?

Tipp:

VIF Werte über 5 bzw. Toleranz unter 0.2 gelten als grenzwertig.

VIF Werte über 10 bzw. Toleranz unter 0.1 gelten als sehr problematisch.

Aufgabe 3

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

Aufgabe 4

Bestimmen Sie den minimalen Stichprobenumfang für eine Variablenbeziehung in der Höhe von ca. f2=0.1. Die Variablenbeziehung soll in einem Regressionsmodell mit 20 weiteren Kontrollvariablen mit einer Power von 0.8 und einem Signifikanzniveau von 95% (bzw. Irrtumswahrscheinlichkeit 0.05) getestet werden. Stellen Sie Ihren Denk- /Rechenvorgang dar.

Tipp: siehe Urban/Mayerl 2011: 159f.

$$N = \frac{\lambda}{f^2}$$

Aufgabe 5

Welche Form von Fehlschluss wird durch ein niedriges Signifikanzniveau "begünstigt"?

Aufgabe 6

In welchen Fällen ist es sinnvoll das Signifikanzniveau höher anzusetzen als 95%?

Übungsaufgabe SPSS

Erstellen einer Interaktionsvariablen:

```
COMPUTE int_alter_gender = alter * gender.
```

Testen des Gesamtmodells auf Linearität:

```
REGRESSION
  /DEPENDENT demzufriedenheit
  /METHOD=ENTER alter gender int_alter_gender
  /SCATTERPLOT=(*ZRESID , *ZPRED).
```

(Multi-)Kollinearitätskoeffezienten anzeigen lassen:

Fügen wir zu `/STATISTICS` noch `BCOV` und `TOL` hinzu, so bekommen wir eine Korrelationstabelle für die Koeffizienten und Toleranzwerte bzw. VIF Werte aus.

```
REGRESSION
  /STATISTICS COEFF OUTS CI(95) R BCOV TOL
  /DEPENDENT demzufriedenheit
  /METHOD=ENTER alter gender int_alter_gender
```

Übungsaufgabe R

Ausgeben der Koeffizienten eines Modells mit `tidy`

```
tidy(mod1)
```

term	estimate	std.error	statistic	p.value
(Intercept)	26.9660246	0.8317190	32.422036	0.00e+00
alter0	0.0848804	0.0208275	4.075392	4.79e-05
bildung_rec	6.3914253	0.2165824	29.510365	0.00e+00

Übungsaufgabe R

Berechnung von 95% Konfidenzintervallen:

```
tidy(mod1) %>%
  mutate(low_se_95 = estimate - 1.96 * std.error) %>%
  mutate(high_se_95 = estimate + 1.96 * std.error)
```

term	estimate	std.error	statistic	p.value	low_se_95	high_se_95
(Intercept)	26.9660246	0.8317190	32.422036	0.00e+00	25.3358553	28.5961939
alter0	0.0848804	0.0208275	4.075392	4.79e-05	0.0440584	0.1257024
bildung_rec	6.3914253	0.2165824	29.510365	0.00e+00	5.9669238	6.8159267

Übungsaufgabe R

Berechnung von 99% Konfidenzintervallen:

```
tidy(mod1)  %>%
  mutate(low_se_99 = estimate - 2.58 * std.error) %>%
  mutate(high_se_99 = estimate + 2.58 * std.error)
```

term	estimate	std.error	statistic	p.value	low_se_99	high_se_99
(Intercept)	26.9660246	0.8317190	32.422036	0.00e+00	24.8201895	29.1118597
alter0	0.0848804	0.0208275	4.075392	4.79e-05	0.0311453	0.1386154
bildung_rec	6.3914253	0.2165824	29.510365	0.00e+00	5.8326427	6.9502078

Übungsaufgabe R

Geschätzte Werte (fitted values) und standardisierte Residuen können wir uns mit `augment` ausgeben:

```
diag_mod <- augment(mod1)
```

Nur zum angucken :)

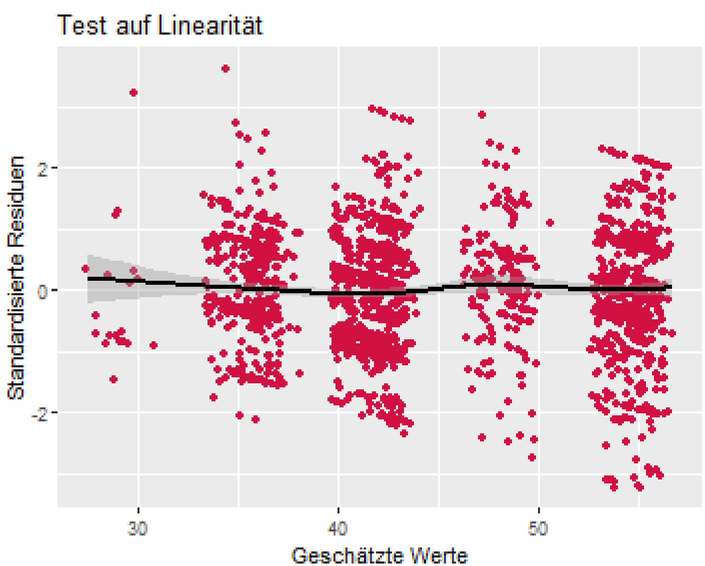
```
head(diag_mod) %>%
  select(.fitted, .std.resid)
```

.fitted	.std.resid
53.80493	0.8218703
42.46505	0.5886210
42.97433	0.0515600
43.22897	-1.8773068
42.97433	1.9982687
53.12589	-0.8649826

Übungsaufgabe R

Testen des Gesamtmodells auf Linearität:

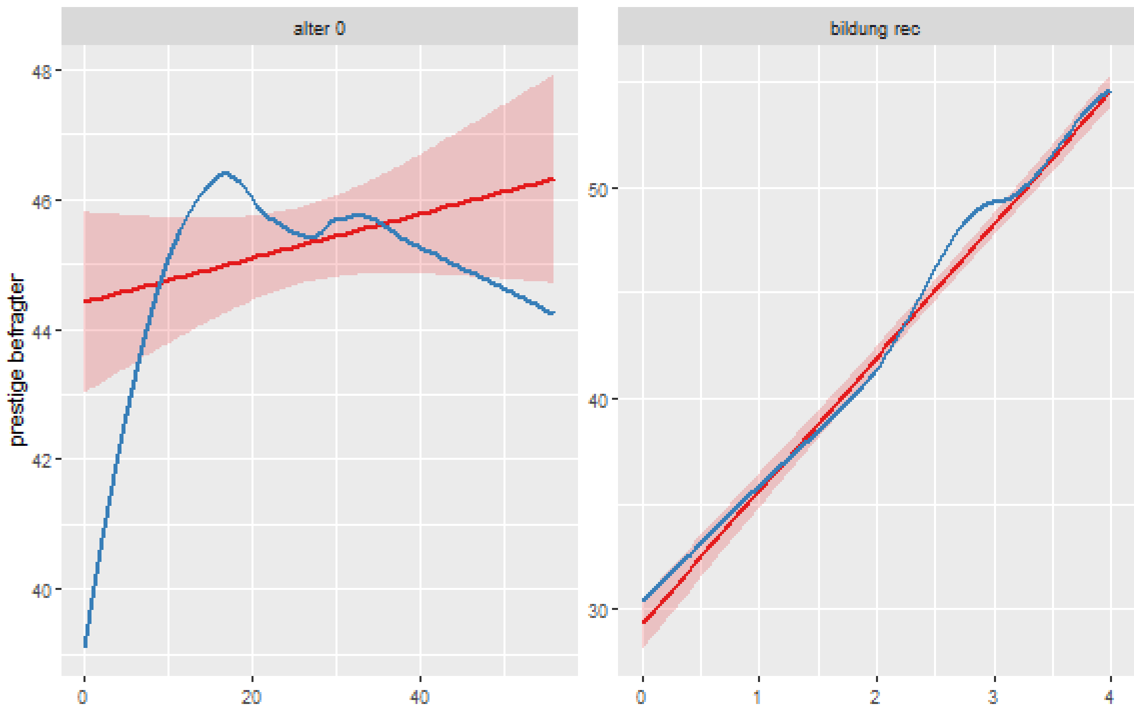
```
diag_mod %>% #Datensatz
  plot_scatter(.fitted, .std.resid, #x und y definieren
               fit.line = "loess",  #zeige eine loess Kurve
               show.ci = T,         #zeige das Konfidenzintervall
               title = "Test auf Linearität",    #Titel der Grafik
               axis.titles = c("Geschätzte Werte",
                               "Standardisierte Residuen"))
```



Übungsaufgabe R

Testen des aller UVs auf Linearität:

```
plot_model(mod1, type = "slope")
```



Übungsaufgabe R

Testen auf Multikollinearität

VIF - Werte:

```
vif(mod1)
```

```
##        alter0 bildung_rec
##      1.006896    1.006896
```

Toleranzwerte:

```
(1/vif(mod1))
```

```
##        alter0 bildung_rec
##     0.9931517   0.9931517
```

Übungsaufgabe R

Testen auf Multikollinearität

Gemeinsam:

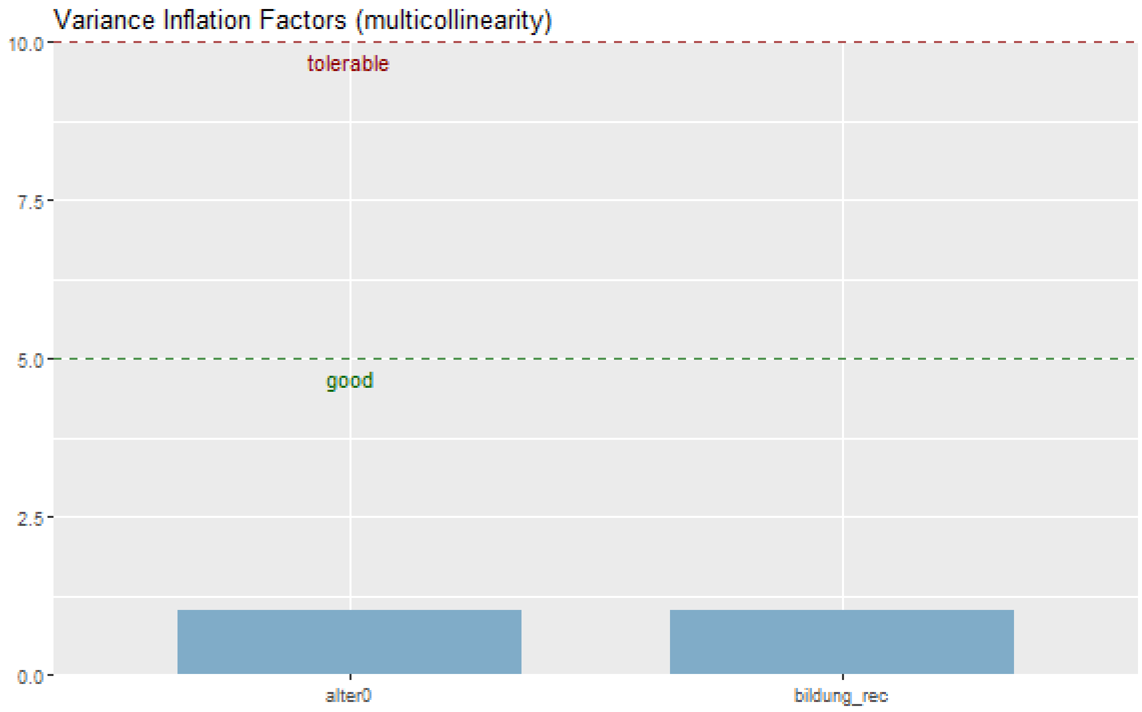
```
data.frame(vif = vif(mod1),
           toleranz = (1/vif(mod1)))
```

	vif	toleranz
alter0	1.006896	0.9931517
bildung_rec	1.006896	0.9931517

Übungsaufgabe R

Eine (nahezu) komplette Regressionsdiagnostik mit `plot_model`:

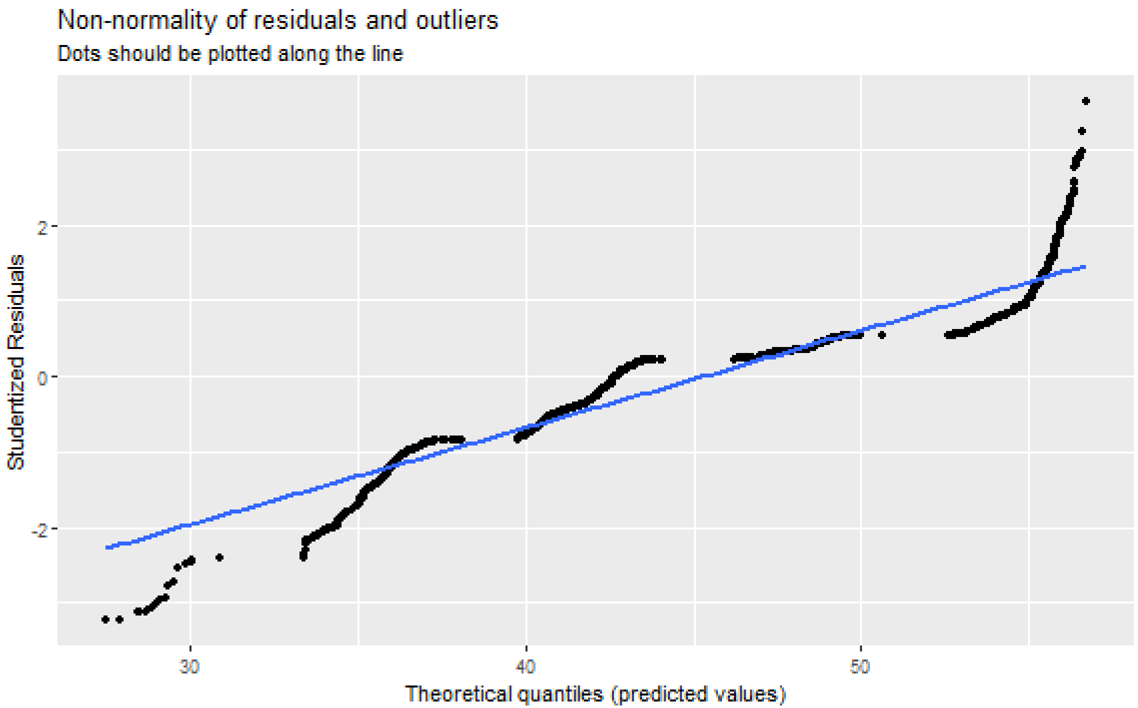
```
plot_model(mod1, type = "diag")
```



Übungsaufgabe R

Eine (nahezu) komplette Regressionsdiagnostik mit `plot_model`:

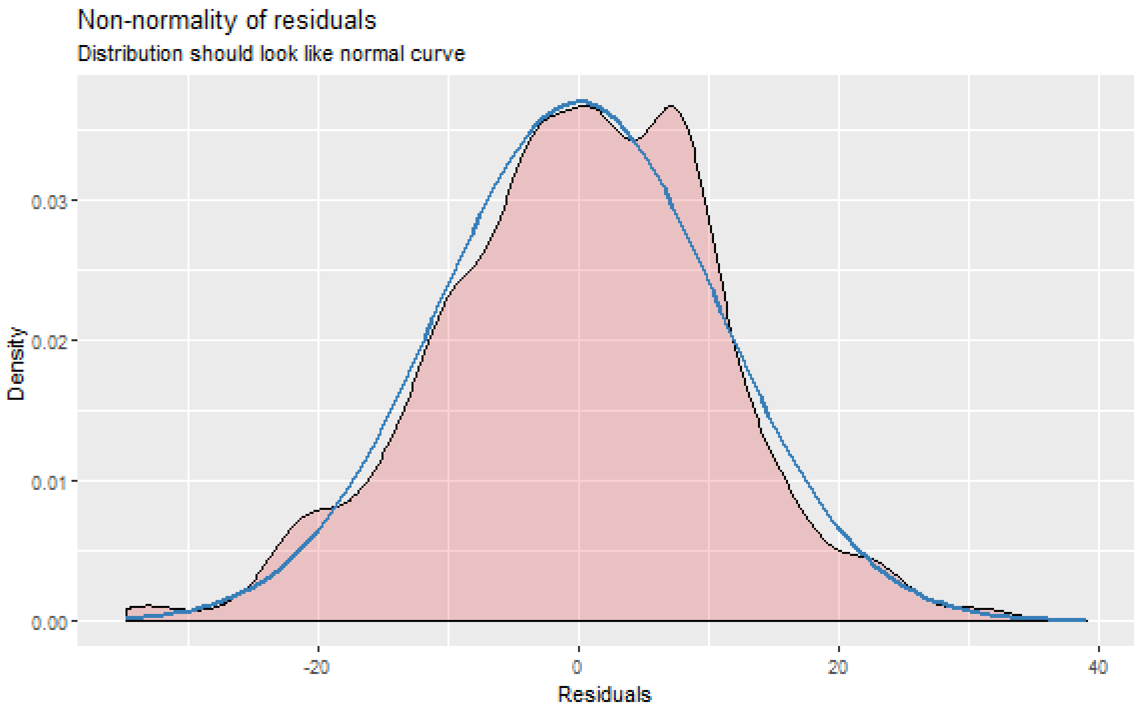
```
plot_model(mod1, type = "diag")
```



Übungsaufgabe R

Eine (nahezu) komplette Regressionsdiagnostik mit `plot_model`:

```
plot_model(mod1, type = "diag")
```



Übungsaufgabe R

Eine (nahezu) komplette Regressionsdiagnostik mit `plot_model`:

```
plot_model(mod1, type = "diag")
```

Bestimmung des benötigten Stichprobenumfangs

Die Teststärke ist eine Funktion der drei Faktoren Signifikanzniveau, geschätzte Effektstärke und Stichprobenumfang. Deswegen lässt sich der notwendige Stichprobenumfang aus einer vorab festgelegten Teststärke, einem bestimmten Signifikanzniveau und einer erwarteten Effektstärke ableiten (a-priori-Analyse).

Folgende Werte sind gegeben:

- Variablenbeziehung: $f^2 = 0.1$
- Anzahl der Kontrollvariablen: 20
- Teststärke: 80%
- Signifikanzniveau: 95% (Irrtumswahrscheinlichkeit $\alpha = 0.05$)

Gesucht:

- N (Stichprobenumfang)

Bestimmung des benötigten Stichprobenumfangs

Nun ist in einem ersten Schritt den Nonzentralitätsparameter λ zu berechnen. Ist dieser berechnet sind sämtliche Größen vorhanden um die umgeformte Gleichung zur Berechnung von N auflösen zu können.

Der Nonzentralitätsparameter λ ergibt sich aus einer Teststärkentabelle für die Analyse mit Alpha = 0.05, gemäß dem gewählten Signifikanzniveau. Unser **u**, also die Anzahl unabhängiger Modellvariablen, beträgt 21. Somit betrachten wir in der Tetstärkentabelle die Zeile mit u = 21 bzw. 20.

In der entsprechenden Zeile der Teststärkentabelle wird nun der erste Wert gesucht, bzw. der kleinste Wert, der die geforderte Teststärke von $80\% (= 0.8)$ erstmalig überschreitet. Ist dieser gefunden, kann die notwendige Stichprobengröße abgeleitet werden, aus der umgeformten Gleichung:

$$N = \frac{\lambda}{f^2}$$

Für den Nonzentralitätsparameter λ wird auf diese Weise ein Wert von 24 aus der Teststärkentabelle ermittelt (u = 20).

Bestimmung des benötigten Stichprobenumfangs

Somit ergibt sich unter den angeführten Randbedingungen eine optimale Stichprobengröße von 240, um mit einer Wahrscheinlichkeit von 80 Prozent einen signifikanten Effekt zu entdecken.

$$N = \frac{24}{0.1} = 240$$