Übungsaufgabe 4

Fabio Votta

27.November 2018

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.

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

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

```
tidy(fit1) %>%
   select(term, estimate, std.error) %>%
   mutate_if(is.numeric, ~round(., digits = 3)) %>% # alle numeric Variablen runden
   mutate(low_se_95 = estimate - 1.96 * std.error) %>%
   mutate(high_se_95 = estimate + 1.96 * std.error) %>%
   kable()
```

term	estimate	std.error	low_se_95	high_se_95
(Intercept)	6.884	0.411	6.07844	7.68956
bildung_rec	1.062	0.161	0.74644	1.37756
geschl_rec	2.549	0.334	1.89436	3.20364
alter0	-0.002	0.008	-0.01768	0.01368
bild_tert	-1.836	0.468	-2.75328	-0.91872
geschl_rec:alter0	0.029	0.009	0.01136	0.04664
alter0:bild_tert	0.075	0.010	0.05540	0.09460

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

Mit 95%-iger Wahrscheinlichkeit liegt das β für tertiäre Bildung zwischen den Grenzen von -2.75 und -0.92.
Mit 95%-iger Wahrscheinlichkeit liegt das β für Alter zwischen den Grenzen von -0.03 und 0.01.

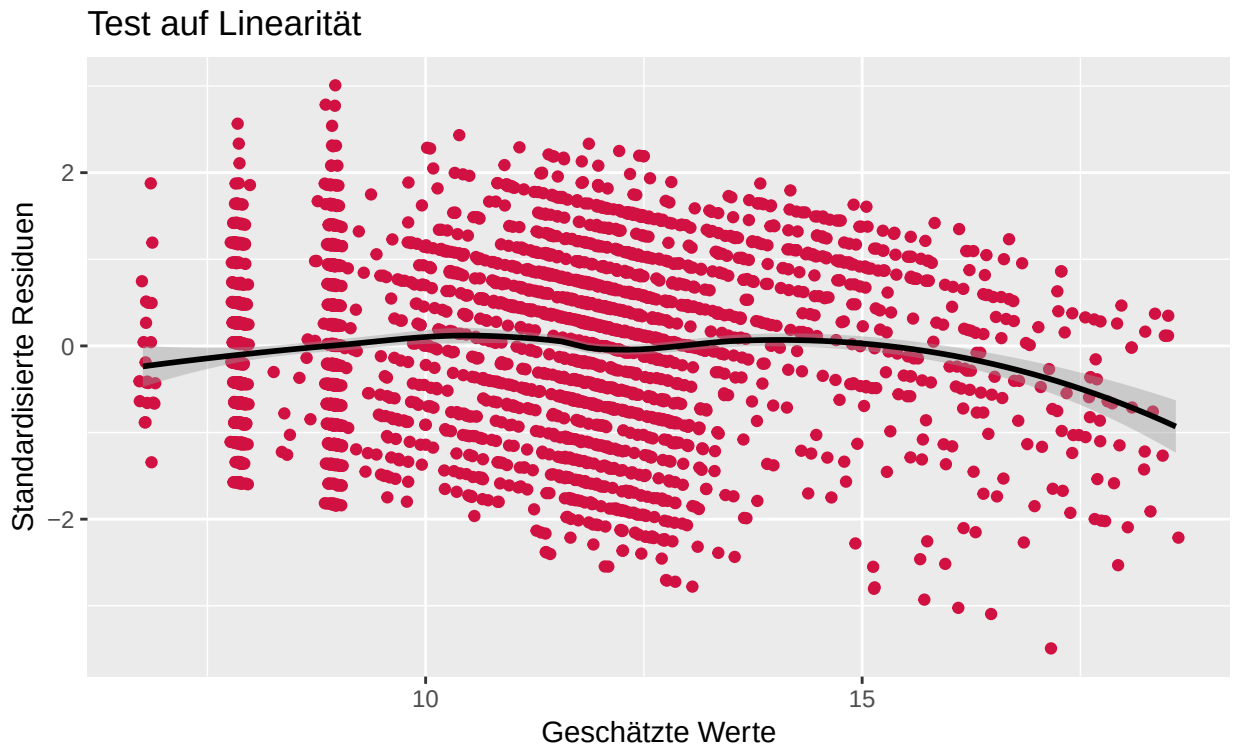
Da das Konfidenzintervall von Alter die 0 mit einschließt, lässt sich davon ausgehen dass der b-Koeffizient nicht statistisch signifikant ist (Signifikanzniveau = 95%).

Aufgabe 1b

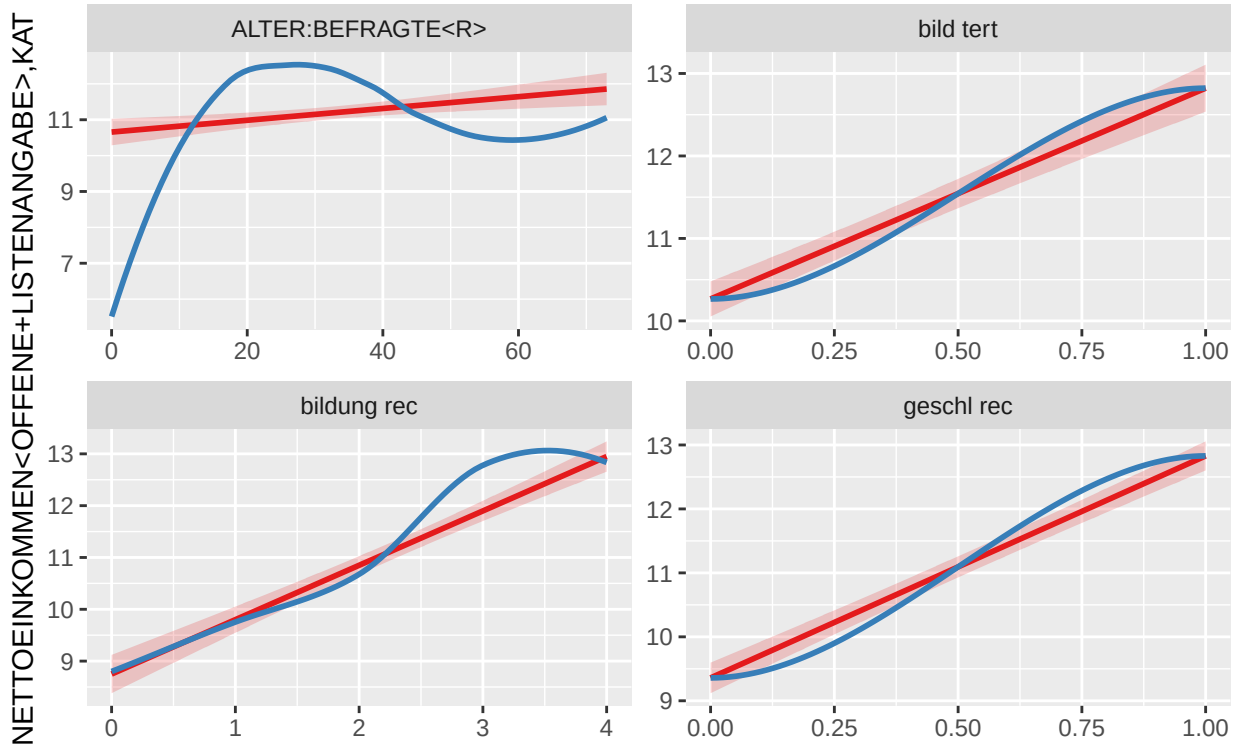
Testen Sie das Gesamtmodell auf Linearität.

```
diag_mod <- augment(fit1)


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"))
```




```
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?

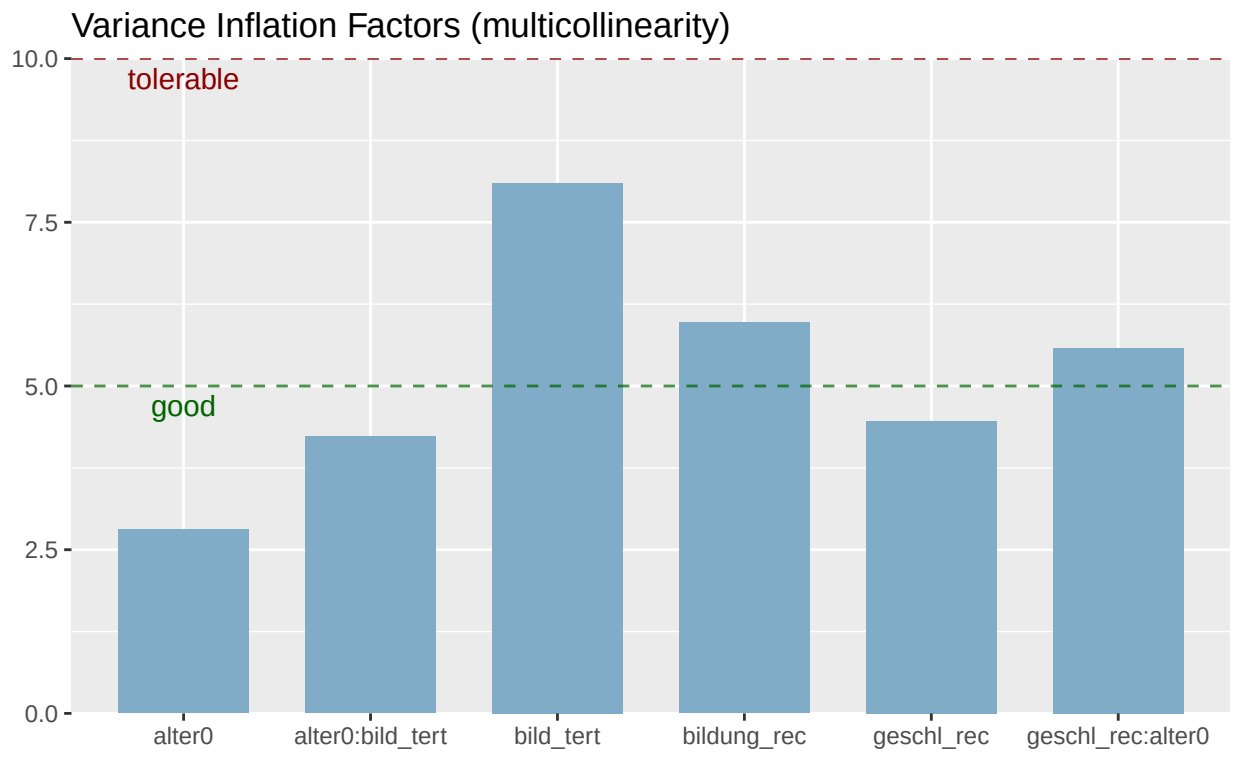
Eine Multikollinerarität liegt vor, wenn unabhängige Variablen untereinander korrelieren. Problematisch ist das, weil bei einer zu hohen Korrelation zwischen den unabhängigen Variablen die Effekte nicht mehr auf einzelne unabhängige Variablen zurückzuführen sind und die b-Koeffizienten sowie Standardfehler 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

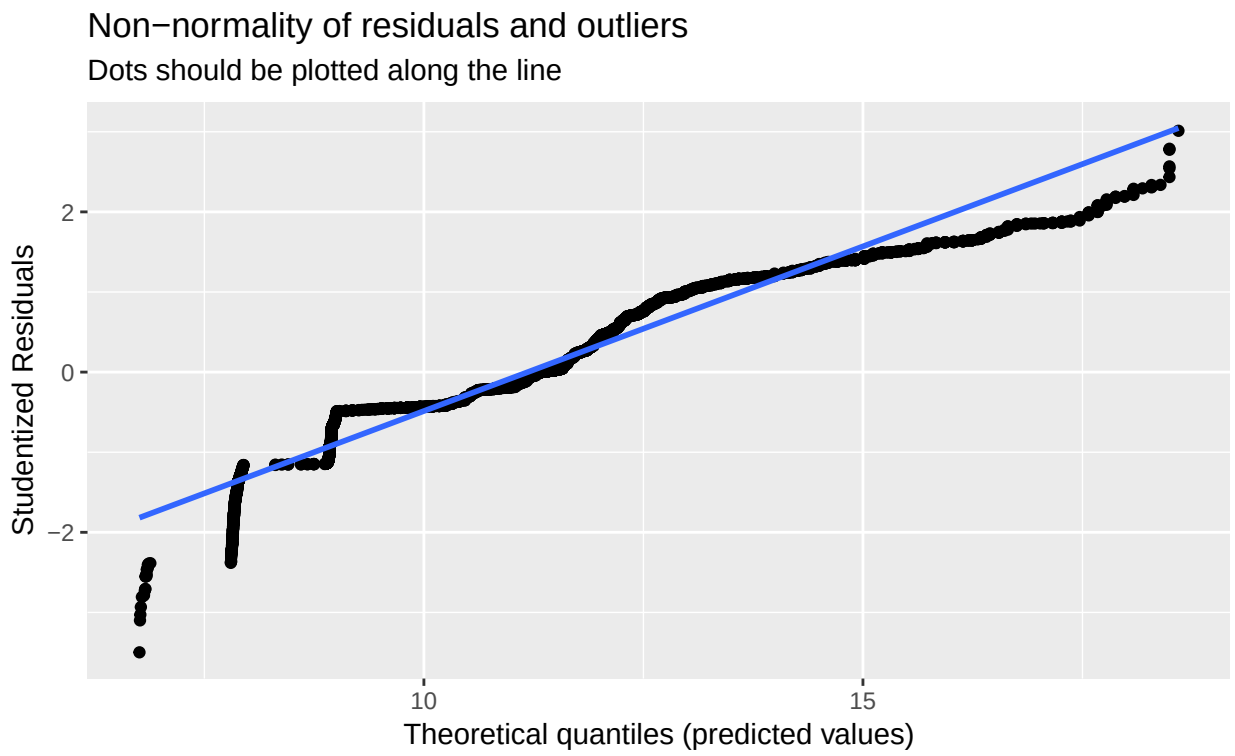
Wie ausgeprägt ist die Multikollinearität im Regressionsmodell von Aufgabe 1? Welche Gründe (inhaltliche) lassen sich für die Multikollinearität identifizieren?

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

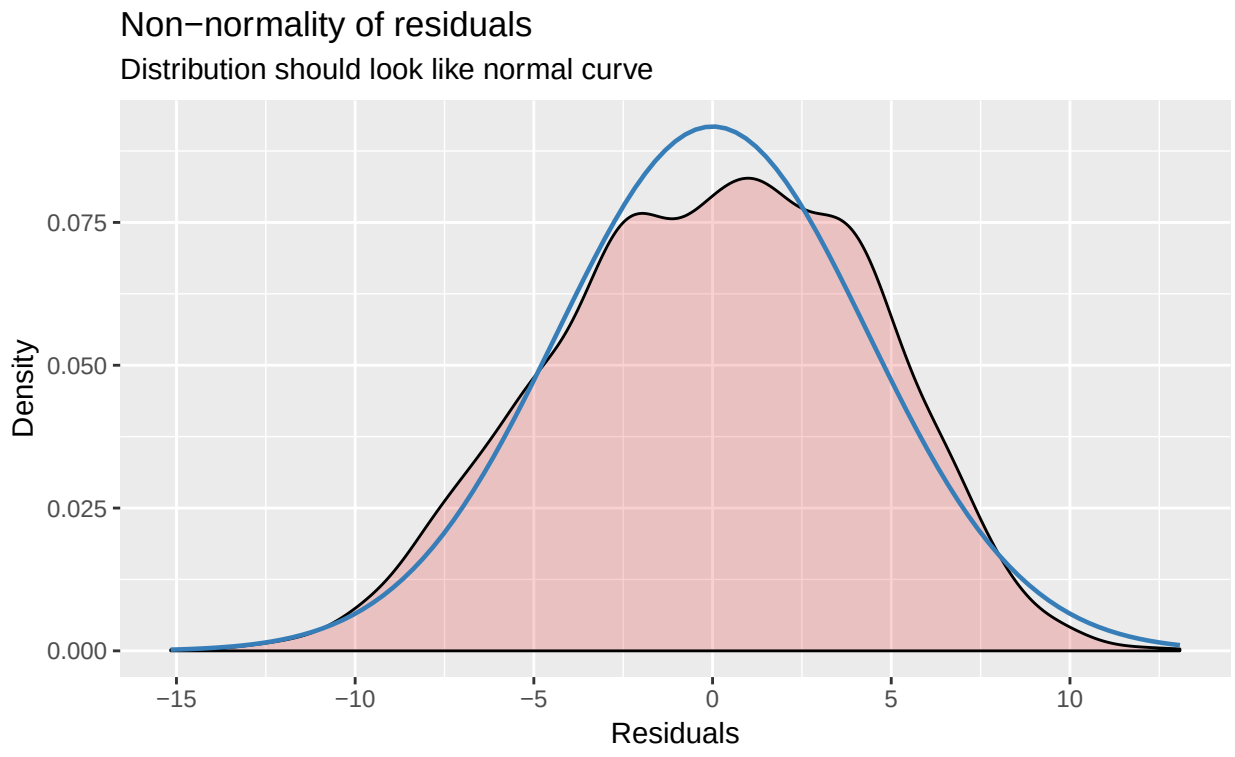
```
## [[1]]
```



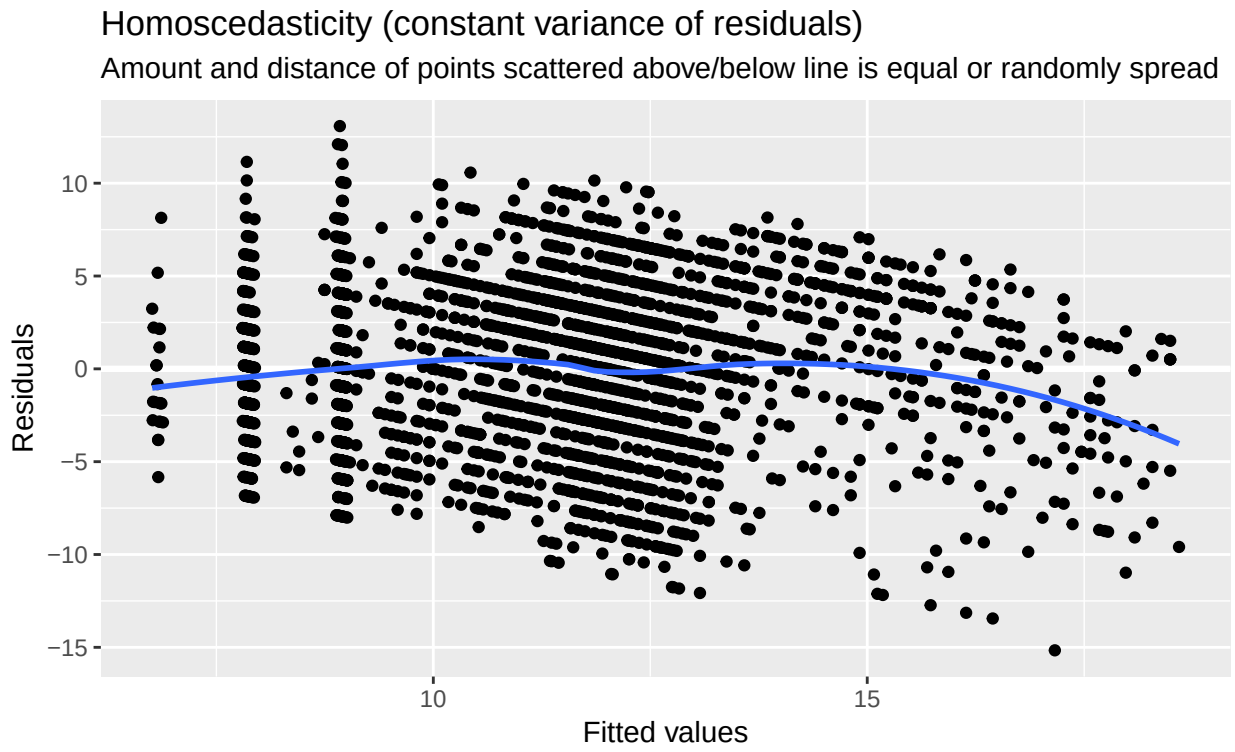

```
##
## [[2]]
```


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]]
```


Non−normality of residuals
Distribution should look like normal curve
Density
0.075
0.050
0.025
0.000
−15
−10
−5
0
5
10
Residuals

```
##
## [[4]]
```




```
data.frame(vif = vif(fit1),
           toleranz = (1/vif(fit1))) %>%
   kable()
```

	vif	toleranz
bildung_rec	5.971818	0.1674532
geschl_rec	4.456750	0.2243787
alter0	2.812942	0.3554997
bild_tert	8.103775	0.1233993
geschl_rec:alter0	5.582849	0.1791200
alter0:bild_tert	4.234264	0.2361686

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.

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.

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)

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).

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

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

Aufgabe 5

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

Der Fehler 1.Art wird von einem niedrigen Signifikanzniveau "begünstigt". denn dadurch wird die H1 angenommen, auch wenn in der Tat H0 beibehalten werden sollte. Es wird angenommen dass es ein Effekt besteht, auch wenn dieser nicht vorhanden ist.

Aufgabe 6

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

Das Signifikanzniveau höher als 95% ergibt Sinn, wenn die Fallzahl des untersuchten Samples sehr hoch ist, da eine hohe Fallzahl oftmals signifikante Ergbenisse erzeugen. Ein anderer Anwendungsfall wäre auch wenn die potentiellen Folgen eines begangenen Fehlers (*false positives* oder *false negatives*) zu kostspielig sind (z.B. medizinische Untersuchungen).