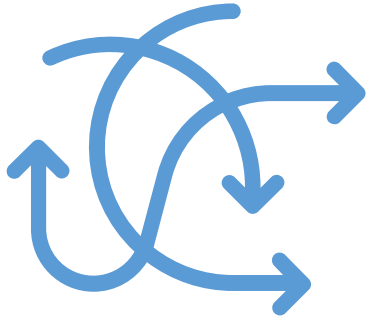
Statistik 2 – Tutorate
Thema 4: Probleme der Regressionsanalyse

Marco Giesselmann, Rémy Blum, Federica Bruno, Simon Honegger, Nora Zumbühl

Lernziele dieser Sitzung

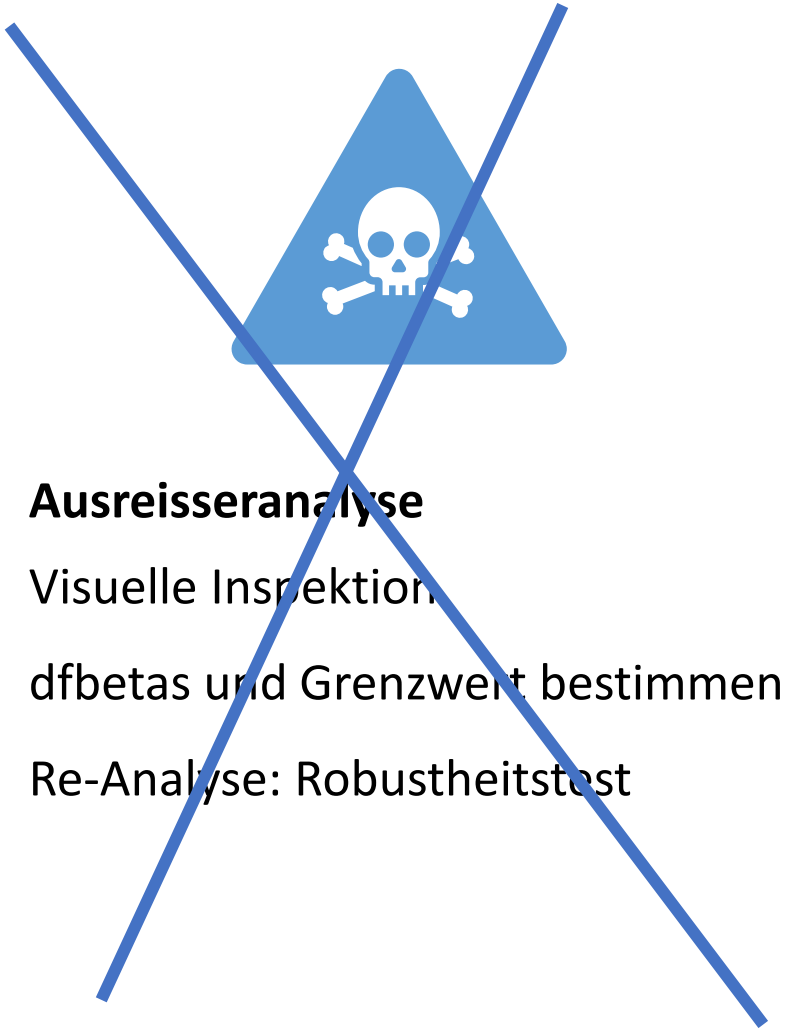


Linearitätsprüfung

Visuelle Inspektion

Statistische Inspektion: Multigruppenanalyse

Quadratische Regression



Ausreisseranalyse

Visuelle Inspektion

dfbetas und Grenzwert bestimmen

Re-Analyse: Robustheitstest

Linearitätsdiagnose: Vorüberlegungen

Zur Linearitätsdiagnose gehören:

- **Theoretische Überlegungen**: Gründe für die Annahme einer Linearitätsabweichung
- **Visuelle Inspektion**: Überprüfung anhand Streudiagramm
- **Analyse**: Identifizierung von Linearitätsabweichungen auf Basis statistischer Parameter

Wir betrachten den Zusammenhang zwischen dem *Lebensalter* und der *individuell wahrgenommenen Verantwortung dafür, den Klimawandel einzudämmen.*

(a) Welche *Einflusstendenz* und (b) welche *Zusammenhangsform* zwischen diesen beiden Merkmalen vermutet ihr?

Vorbereitungen zur Inspektion und Analyse

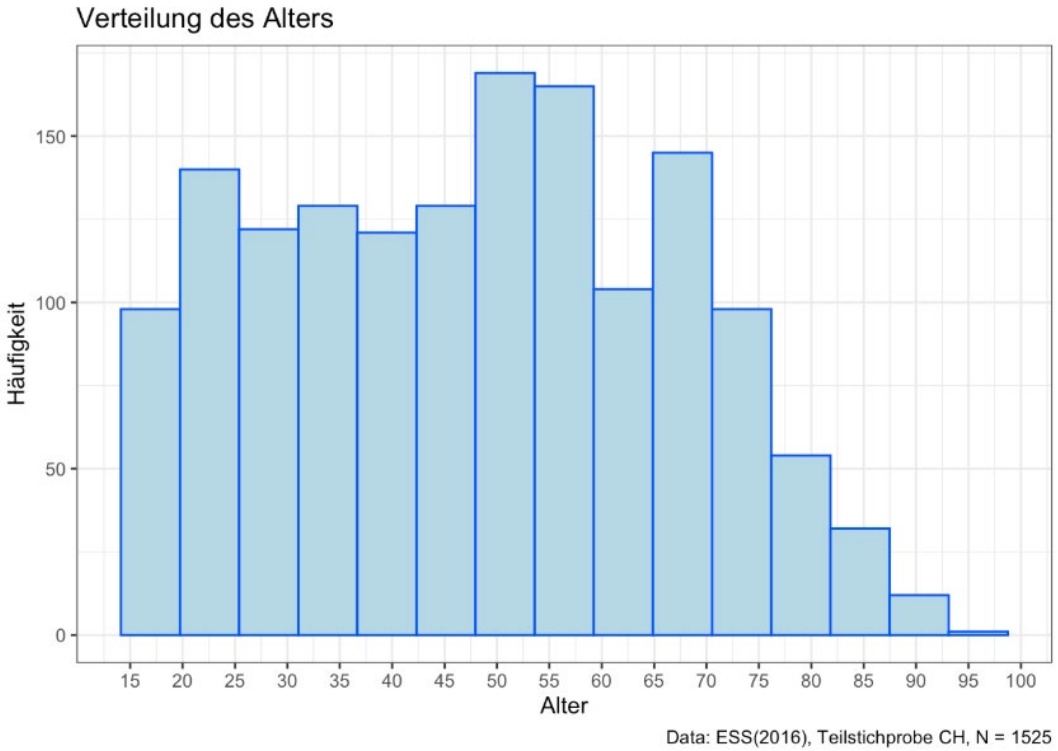
Analyse: **Lebensalter -> Klimaverantwortung**
Benötigte Variablen: **idno, agea, ccrdprs**
Sample: **ESS8, Schweizer Subsample**

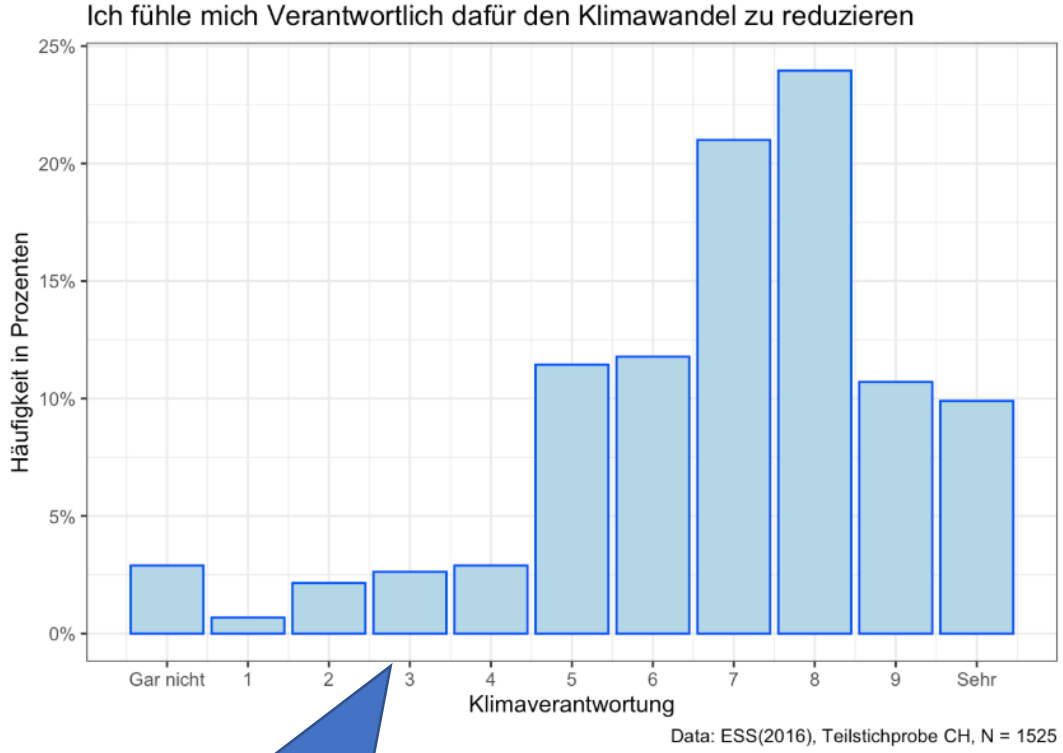
```
ess8_ch <- filter(ess8, cntry == "CH")


ess8_ch_ss_1 <- select(ess8_ch, identifyer = idno,
                       alter = agea,
                       klima_ver = ccrdprs)
```

Finde mit **attributes()** und **summary()** heraus, was die Variablen messen und ob es Werte gibt, welche eine Rekodierung erfordern (Missings!).

- **agea** misst das Alter in Jahren
- Die Variable **ccrdprs** misst die Klimaverantwortung auf einer 10er-Skala von 0 (= kein Verantwortungsgefühl) bis 10 (= starkes Verantwortungsgefühl)
- Es sind keine Rekodierungen erforderlich

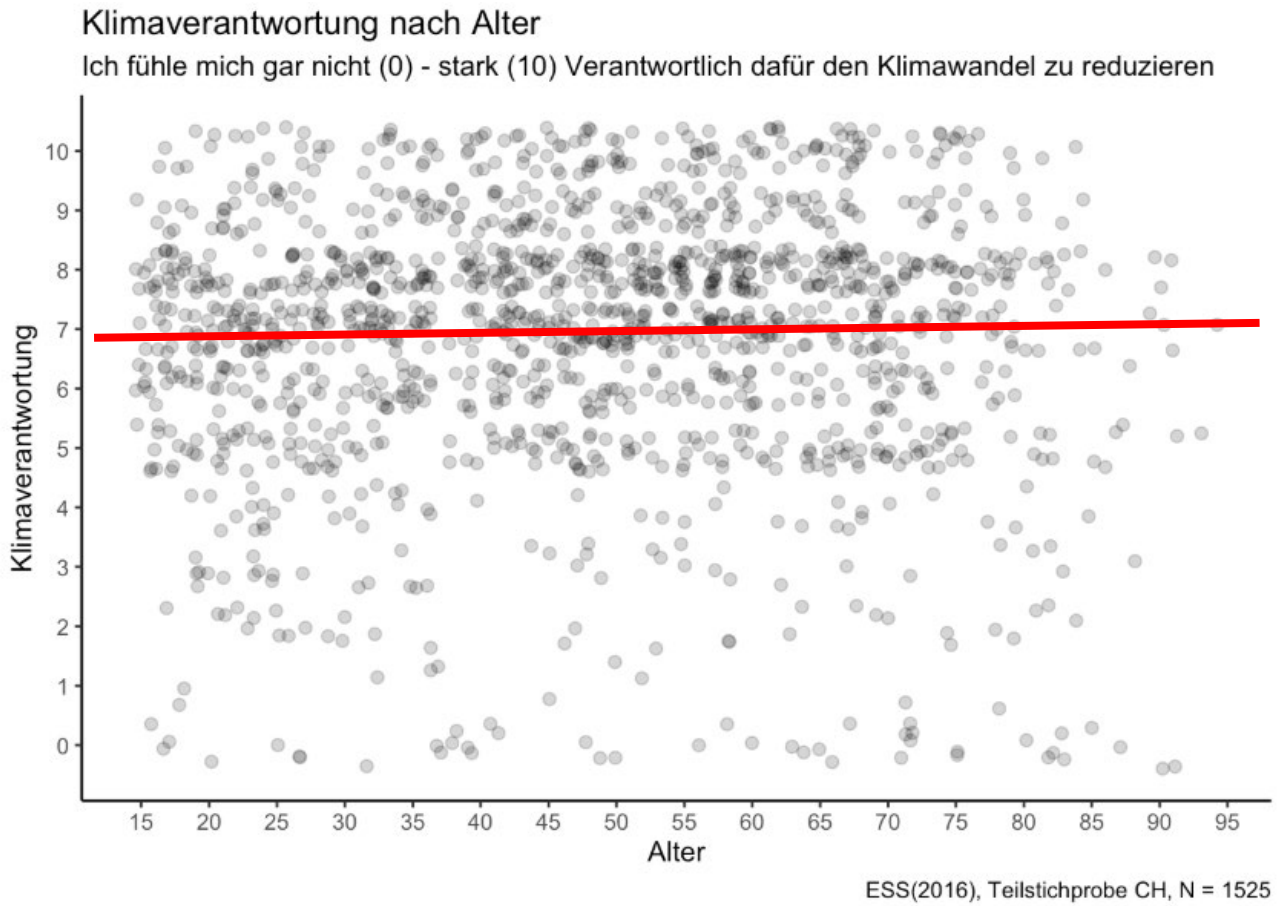
Vorbereitungen zur Inspektion und Analyse





Den R Code zu den Diagrammen findet ihr auf der HP

Linearitätsdiagnose: Visuelle Inspektion Streudiagramm




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.8426286  0.1586060  43.142   <2e-16 ***
alter       0.0004828  0.0030855   0.156    0.876
```

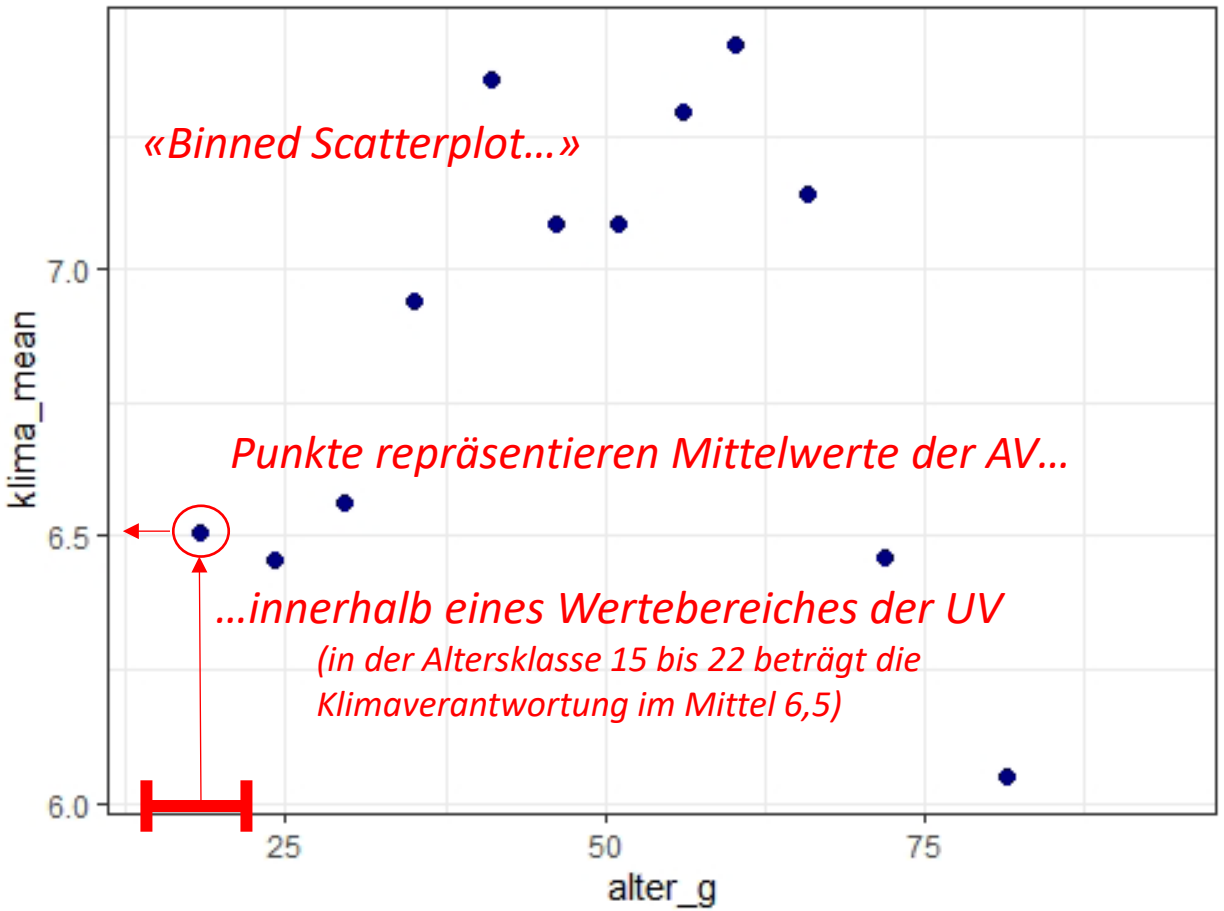
Ist hier ein Zusammenhang (linear, nicht -linear) erkennbar?

Die visuelle Inspektion liefert weder einen Hinweis auf einen linearen (s. auch Regressionsgerade), noch auf einen nicht-linearen Zusammenhang.

Aufgrund der vielen Datenpunkte und des hier unübersichtlichen Verlaufs der Punktwolke können wir jedoch einen nicht-linearen Zusammenhang noch nicht sicher ausschliessen.

Linearitätsdiagnose: Visuelle Inspektion «Binplot»

```
1  library(binsreg)
2  ess_sample_bins<-as.data.frame(ess8_ch_ss_1)
3  ess_sample_bins$klima_mean <- as.numeric(ess_sample_bins$klima_ver)
4  ess_sample_bins$alter_g <- as.numeric(ess_sample_bins$alter)
5  binsreg(data = ess_sample_bins, x = alter_g, y = klima_mean
```



Was ist hier abgebildet?
Welche Zusammenhangsform zeigt sich?

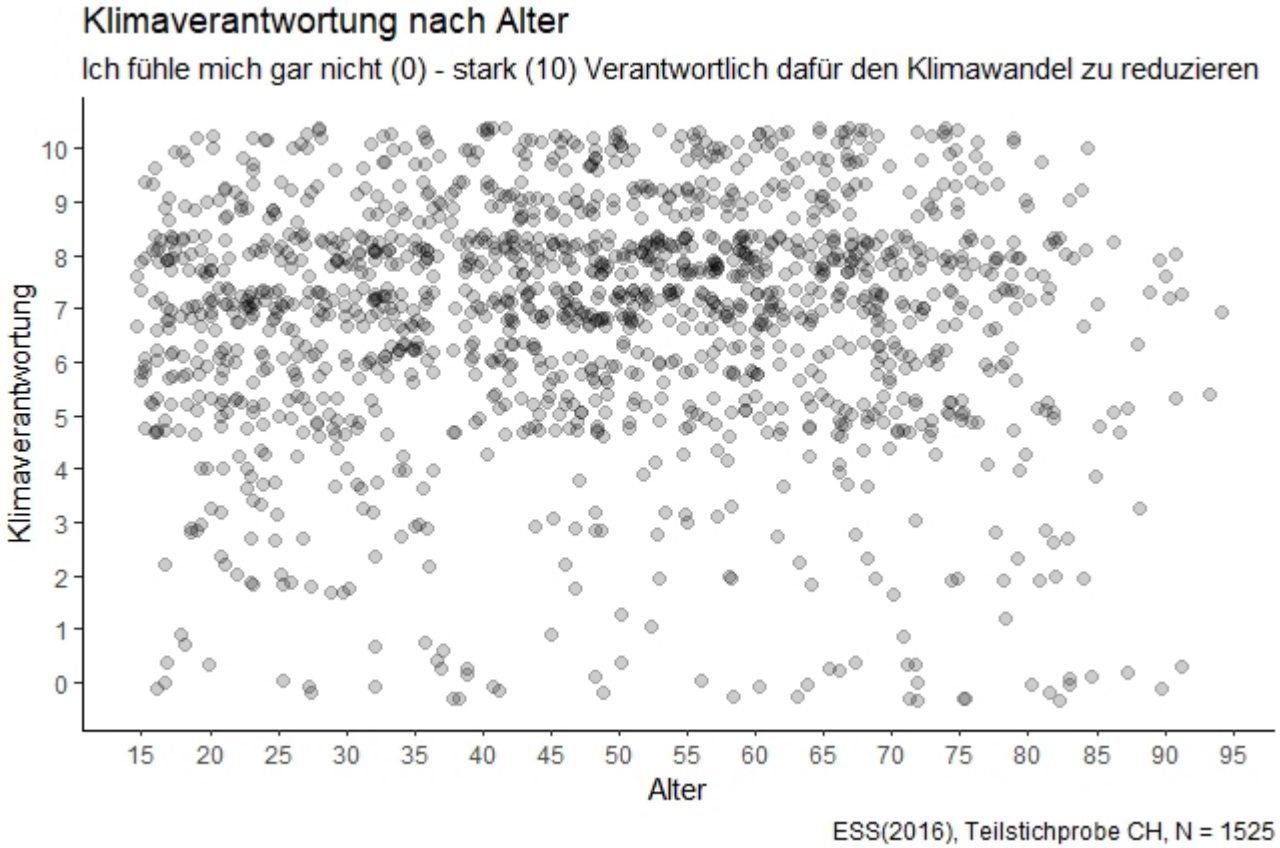
Ein «Binned Scatterplot» illustriert den Verlauf klassenspezifischer Mittelwerte der AV

Die Klassenbreiten variieren entsprechend der Verteilungsdichte, aber umfassen alle gleich viele Personen. *Weitere Infos: HP*

Der Binplot entlarvt einen umgekehrt U-förmiger Zusammenhang.
Bestätigt sich dieser visuelle Eindruck in einer Multigruppenanalyse?

Linearitätsdiagnose: Multigruppenanalyse

```
summary(ess8_ch_ss_1$alter)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
15.00   32.00   48.00   47.83   62.00   94.00       6
```



Wie viele Splits sollten wir machen und bei welchen Werten der UV?

Optionen:

(a) Theoretische Begründung

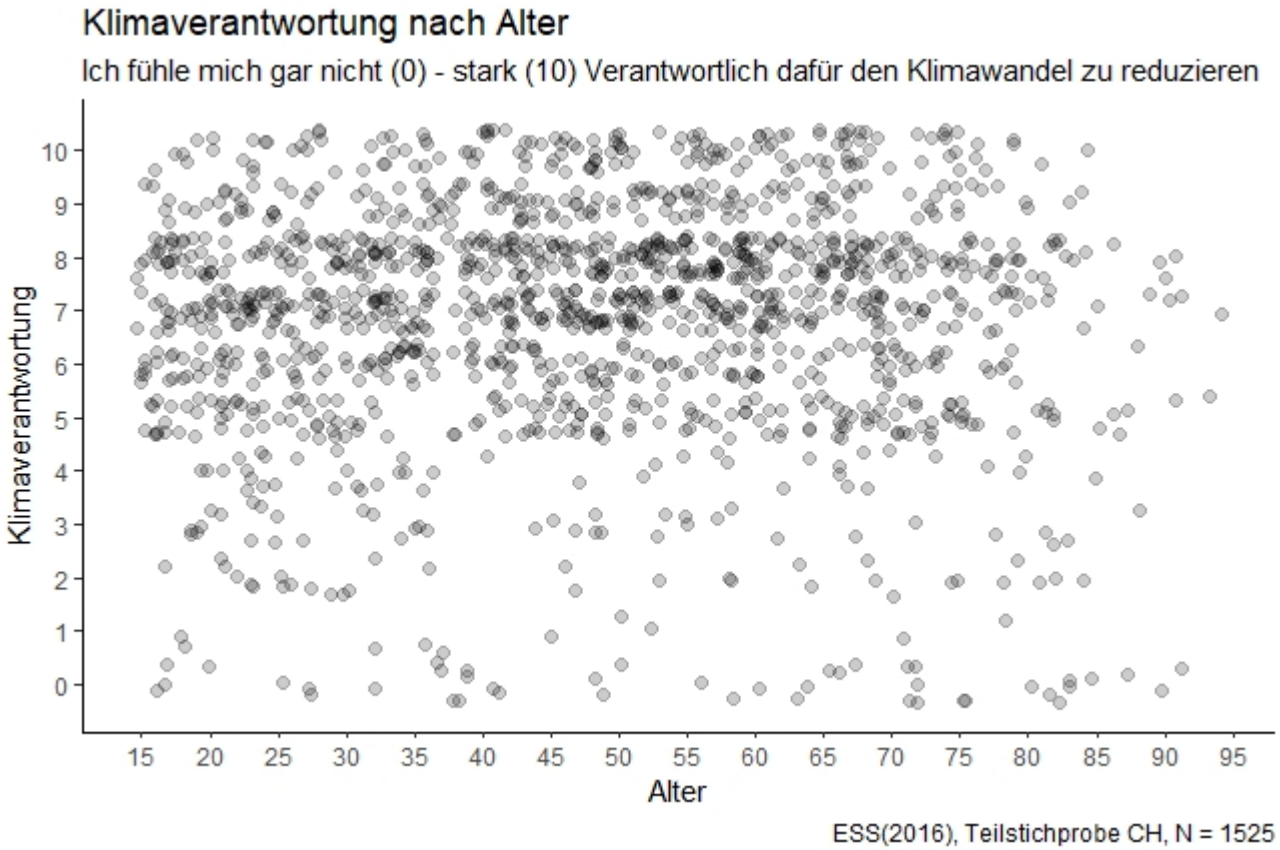
(b) Datengetriebene Identifikation: Scheitelpunkt

(c) Konvention: 1 Split, Median

(d) Alternativ: 3 Splits, Quartile

Linearitätsdiagnose: Multigruppenanalyse

```
summary(ess8_ch_ss_1$alter)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
15.00   32.00   48.00   47.83   62.00   94.00       6
```



Aufgabe: Time to split

1. Erstellt einen neuen Teildatensatz **ess8_split1**, in dem nur Merkmalsträger mit Alter bis zum Median vorhanden sind.
2. Erstellt einen neuen Teildatensatz **ess8_split2**, in dem nur Merkmalsträger mit Alter ab dem Median vorhanden sind.
3. Berechnet für die zwei Teildatensätze den jeweiligen Koeffizienten.
4. Vergleicht die beiden Koeffizienten. Bestätigt sich der visuelle Befund einer Linearitätsabweichung?

Linearitätsdiagnose: Multigruppenanalyse

1. Erstellt einen neuen Teildatensatz **ess8_split1**, in dem nur Merkmalsträger mit Alter bis zum Median vorhanden sind.
2. Erstellt einen neuen Teildatensatz **ess8_split2**, in dem nur Merkmalsträger mit Alter ab dem Median vorhanden sind.

```
ess8_split1 <- filter(ess8_ch_ss_1, alter <= 48)
ess8_split2 <- filter(ess8_ch_ss_1, alter > 48)
```

3. Berechnet für die zwei Teildatensätze das jeweilige lineare Regressionsmodell.

```
model_split1 <- lm(klima_ver~alter, data = ess8_split1)
model_split2 <- lm(klima_ver~alter, data = ess8_split2)
```

4. Vergleicht die beiden Modelle, liegt eine *inhaltlich* substantielle Linearitätsabweichung vor? Interpretiert zudem den Zusammenhang auf Basis der beiden Koeffizienten.

```
Coefficients:
(Intercept)        alter
     5.9076       0.0281
```

```
Coefficients:
(Intercept)        alter
    9.58318     -0.04157
```

Ja, denn:
- Umkehrung
- Beide Steigungen inhaltlich bedeutsam

Linearitätsdiagnose: Multigruppenanalyse

1. Erstellt einen neuen Teildatensatz **ess8_split1**, in dem nur Merkmalsträger mit Alter bis zum Median vorhanden sind.
2. Erstellt einen neuen Teildatensatz **ess8_split2**, in dem nur Merkmalsträger mit Alter ab dem Median vorhanden sind.

```
ess8_split1 <- filter(ess8_ch_ss_1, alter <= 48)
ess8_split2 <- filter(ess8_ch_ss_1, alter > 48)
```

3. Berechnet für die zwei Teildatensätze das jeweilige lineare Regressionsmodell.

```
model_split1 <- lm(klima_ver~alter, data = ess8_split1)
model_split2 <- lm(klima_ver~alter, data = ess8_split2)
```

4. Vergleicht die beiden Modelle, liegt eine *statistisch* substantielle Linearitätsabweichung vor?

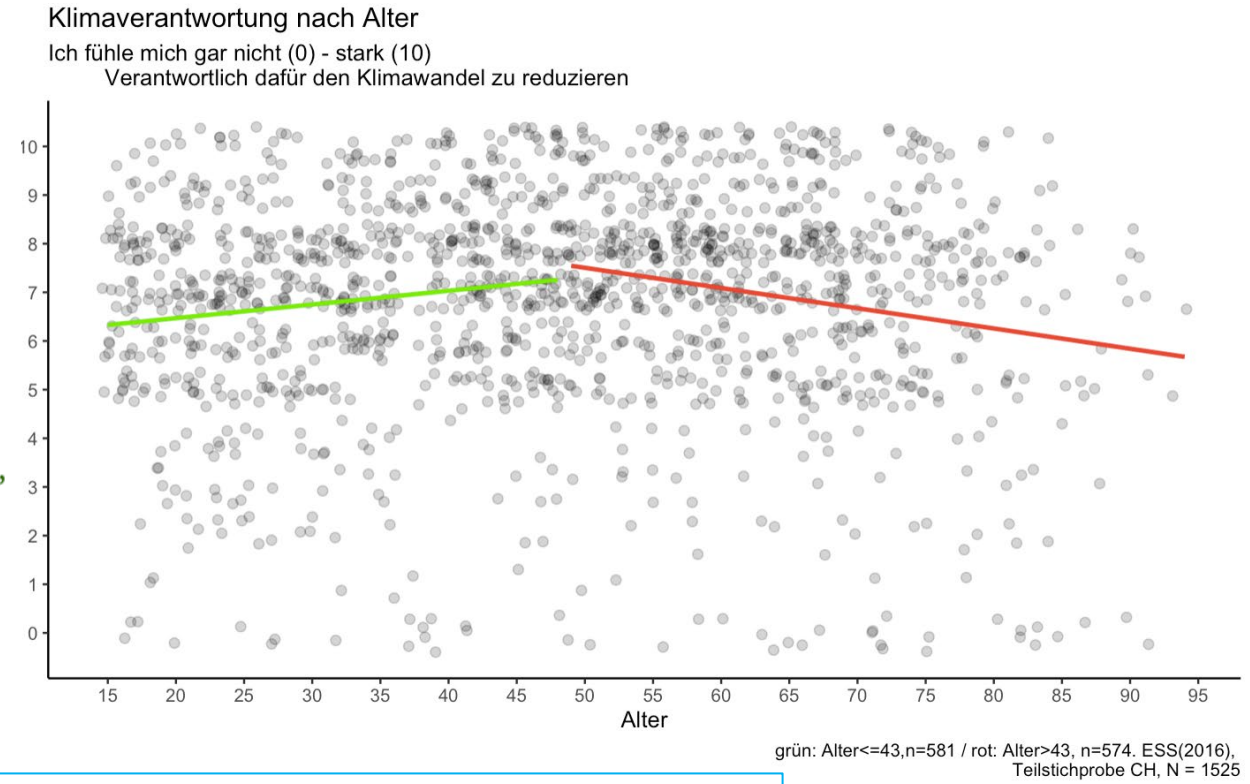
```
Call:
lm(formula = klima_ver ~ alter, data = ess8_split1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.90764    0.26513  22.282  < 2e-16 ***
alter        0.02810    0.00784   3.584 0.000359 ***
```

```
Call:
lm(formula = klima_ver ~ alter, data = ess8_split2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.583184   0.535634  17.891  < 2e-16 ***
alter       -0.041573   0.008254  -5.037 5.99e-07 ***
```

Selbst nach sehr strengem inferenzstatistischem Kriterium – *beide Koeffizienten sind jeweils signifikant in unterschiedliche Richtungen* – ist die Linearitätsabweichung indiziert.

Linearitätsdiagnose: Multigruppenanalyse im Plot

```
ggplot(ess8_ch_ss_1, aes(alter, klima_ver))+
  geom_jitter(alpha = 0.2, size =2)+
  scale_x_continuous(breaks = seq(15,100,5))+
  scale_y_continuous(breaks = seq(0,10,1))+
  geom_smooth(method = "lm", se = F, color = "green", data = ess8_split1)+
  geom_smooth(method = "lm", se = F, color = "red", data = ess8_split2)+
  theme_classic() +
  labs(title = "Klimaverantwortung nach Alter",
       y = "Klimaverantwortung", x = "Alter",
       subtitle= "Ich fühle mich gar nicht (0) - stark (10)
       Verantwortlich dafür den Klimawandel zu reduzieren",
       caption = "grün: Alter<=43,n=581 / rot: Alter>43, n=574. ESS(2016),
       Teilstichprobe CH, N = 1525")
```



Warum hier keine einfache lineare Regression?

Wie weiter: Quadrierung oder Logarithmierung?

Regression mit quadriertem Term: Analyse

Aufgrund unserer theoretischen Vorüberlegungen und gestützt auf die visuelle und analytische Inspektion nehmen wir eine Quadrierung der UV vor.

Zur Umsetzung bilden wir eine quadrierte Altersvariable und fügen sie dem lm-Befehl hinzu.

```
ess8_ch_ss_1$alter_sqr <- ess8_ch_ss_1$alter^2
model_sqr <- lm(klima_ver~alter_sqr+alter, data = ess8_ch_ss_1)
summary(model_sqr)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.7966541  0.3530216  13.587  < 2e-16 ***
alter_sqr   -0.0010240  0.0001584  -6.465 1.37e-10 ***
alter        0.0998233  0.0156638   6.373 2.47e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.206 on 1479 degrees of freedom
  (43 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.0275,    Adjusted R-squared:  0.02619
```

	alter Age of respondent, calculated	alter_sqr
1	56	3136
2	29	841
3	67	4489
4	53	2809
5	68	4624
6	55	3025
7	36	1296

Regression mit quadriertem Term: Analyse

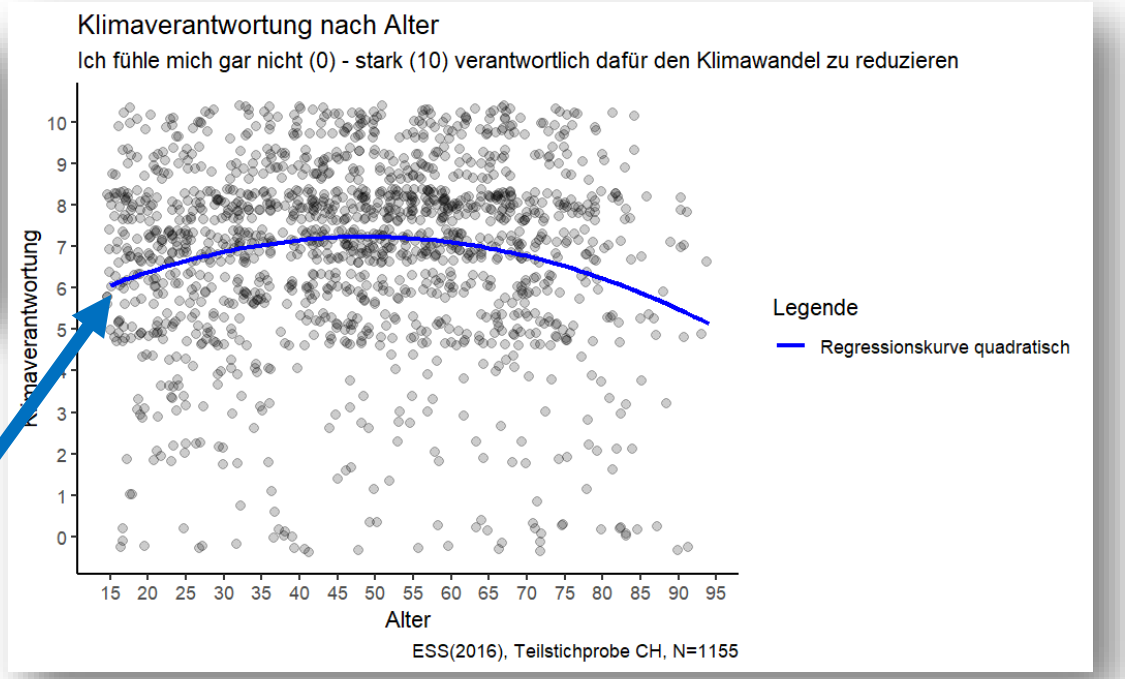
Wir können im lm-Befehl auch automatisch eine entsprechende Variable anlegen und integrieren lassen:

```
model_sqr <- lm(klima_ver ~ alter + I(alter^2), data = ess8_ch_ss_1)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.7966541  0.3530216  13.587  < 2e-16 ***
alter        0.0998233  0.0156638   6.373 2.47e-10 ***
I(alter^2)  -0.0010240  0.0001584  -6.465 1.37e-10 ***
```

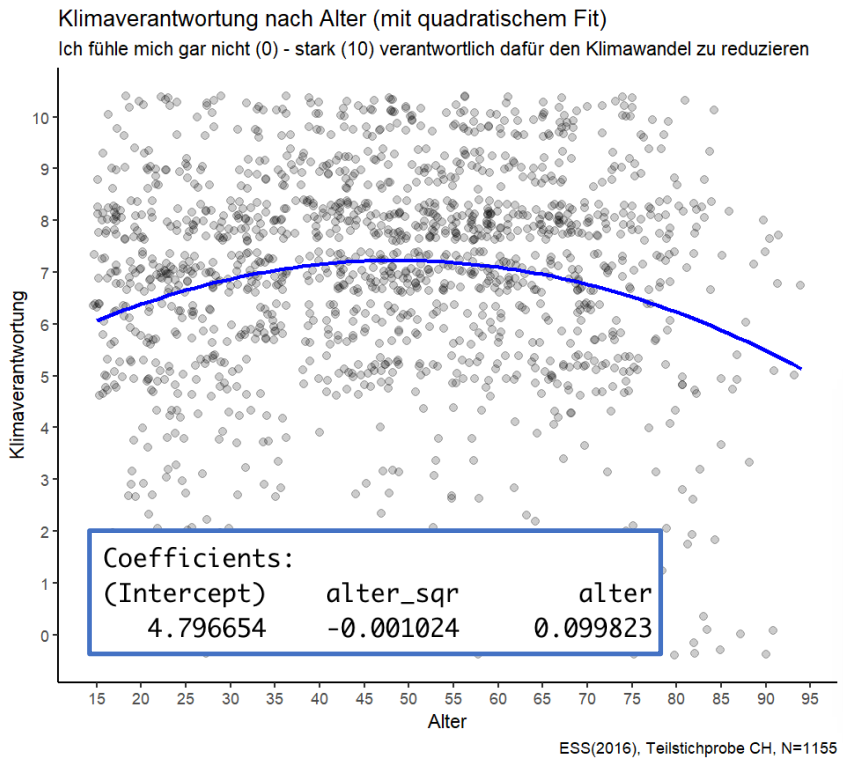
Auch hier stellen die Koeffizienten in der Tabelle die Parameter einer Gleichung dar! Wie sieht diese aus?

$$\widehat{klima\_ver} = 4.8 + 0.1 * Alter - 0.001 * Alter^2$$



Was können wir auf Grundlage der Koeffizienten über den Zusammenhang aussagen?

Regression mit quadriertem Term: Vorhersagebasierte Interpretation



Welche Vorhersagen ergibt der (quadratische) Regressionsfit für Personen mit 20, 35, 50, 65 und 80 Lebensjahren?

```
library(ggeffects)
sd (ess8_ch_ss_1$klima_ver, na.rm=TRUE)
ggpredict(model_sqr1, terms = "alter[20, 35, 50, 65, 80]")
```

alter	Predicted	95% CI
20	6.38	[6.14, 6.63]
35	7.04	[6.89, 7.18]
50	7.23	[7.07, 7.39]
65	6.96	[6.80, 7.11]
80	6.23	[5.93, 6.53]

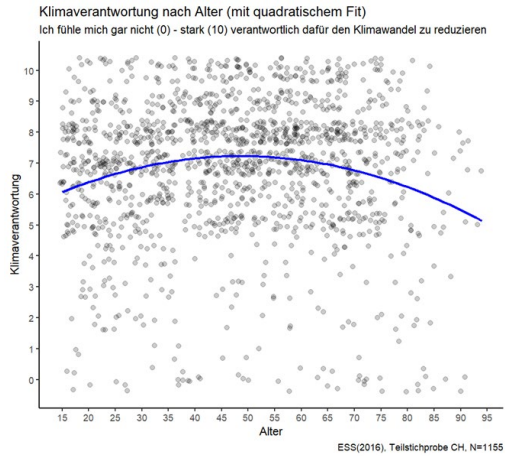
Auswertung?

Beispielhaftes Auswertungsmuster:

Der quadratische fit sagt das Maximum der Klimaverantwortung für 50-jährige voraus und eine um fast einen Skalenpunkt (bzw. mehr als 0,3 Standardabweichungen) geringere Klimaverantwortung an den Rändern der Altersverteilung Erwachsener. Obgleich der Verantwortungsanstieg also bis zum 50. Lebensjahr im Mittel stetig positiv ist, ist er im jungen Erwachsenenleben (Anstieg um zwei Drittel Skalenpunkte zwischen 20 und 35) deutlich stärker ausgebildet als im mittleren Alter (Anstieg um weniger als einen Fünftel Skalenpunkt zwischen 35 und 50). Ausweislich des quadratischen Regressionsfits zeigt sich also ein relativ hohes und stabiles Plateau der Klimaverantwortung in der Lebensmitte (35-65), gerahmt von einem fulminanten Auf- und Abstieg in den rahmenden, frühen und späten Lebensphasen.

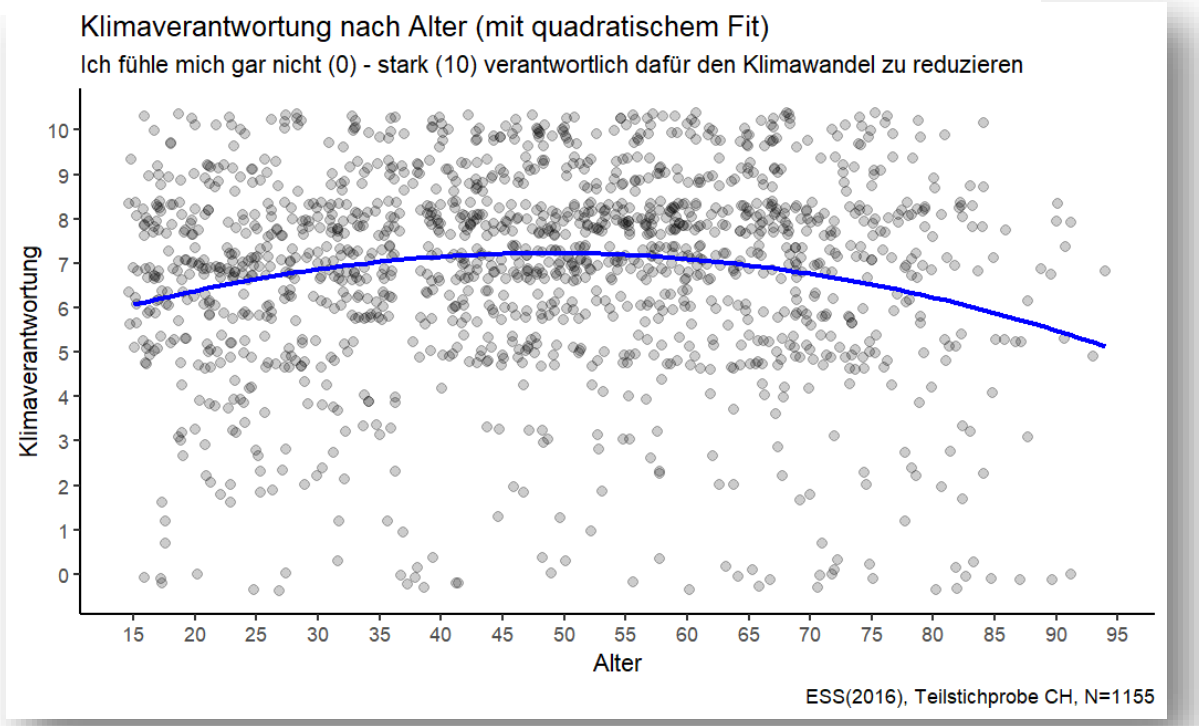
Regression mit quadriertem Term: Visualisierung

Aufgabe: Stellt die quadratische Regressionskurve im Scatterplot dar.
(Tip: geom_smooth muss modifiziert werden)



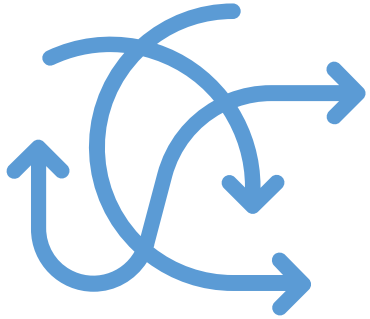
Regression mit quadriertem Term: Visualisierung

```
ggplot(ess8_ch_ss_1, aes(alter, klima_ver))+
  geom_jitter(alpha = 0.2, size =2)+
  scale_x_continuous(breaks = seq(15,100,5))+
  scale_y_continuous(breaks = seq(0,10,1))+
  geom_smooth(method = "lm", se = F, color = "blue", formula = y ~ poly(x,2))+
  theme_classic()+
  labs(title = "Klimaverantwortung nach Alter (mit quadratischem Fit)",
       y = "Klimaverantwortung",
       x = "Alter",
       subtitle = "Ich fühle mich gar nicht (0) - stark (10) verantwortlich dafür den Klimawandel zu reduzieren",
       caption = "ESS(2016), Teilstichprobe CH, N=1155")
```



- Hinweis auf Übung HP
- Hinweis auf Folgefolien und Übung zu Ausreisser (wird im Tutorat in diesem Semester nicht behandelt)

Lernziele dieser Sitzung



Linearitätsprüfung

Visuelle Inspektion

Statistische Inspektion: Multigruppenanalyse

Polynomiale Regression

Ausreisseranalyse

Visuelle & theoretische Diagnose

Statistische Diagnose: dfbetas

Re-Analyse: Robustheitstest

Ausreisserdiagnose

Zur Ausreisserdiagnose im Kontext der Regression gehören:

- **Nachdenken**: Ist es plausibel von Ausreissern auszugehen?
- **Visuelle Inspektion**: Überprüfung anhand des Streudiagramms bzw. Residualplots
- **Analyse**: Identifizierung von Ausreissern auf Basis statistischer Parameter

Wir betrachten den Zusammenhang zwischen den **Bildungsjahren** (eduyrs) und dem **zeitlichen Umfang der täglichen Internetnutzung** (netustm).

- Welchen Zusammenhang zwischen diesen Variablen erwartet ihr?
- Ist dieser Zusammenhang grundsätzlich anfällig für Ausreisser?

Vorbereitungen zur Ausreisserdiagnose

- Benötigte Variablen: **eduyrs** und **netustm**
- Reduktion des Datensatzes auf die Schweiz
- Inspektion des Datensatzes / der Variablen

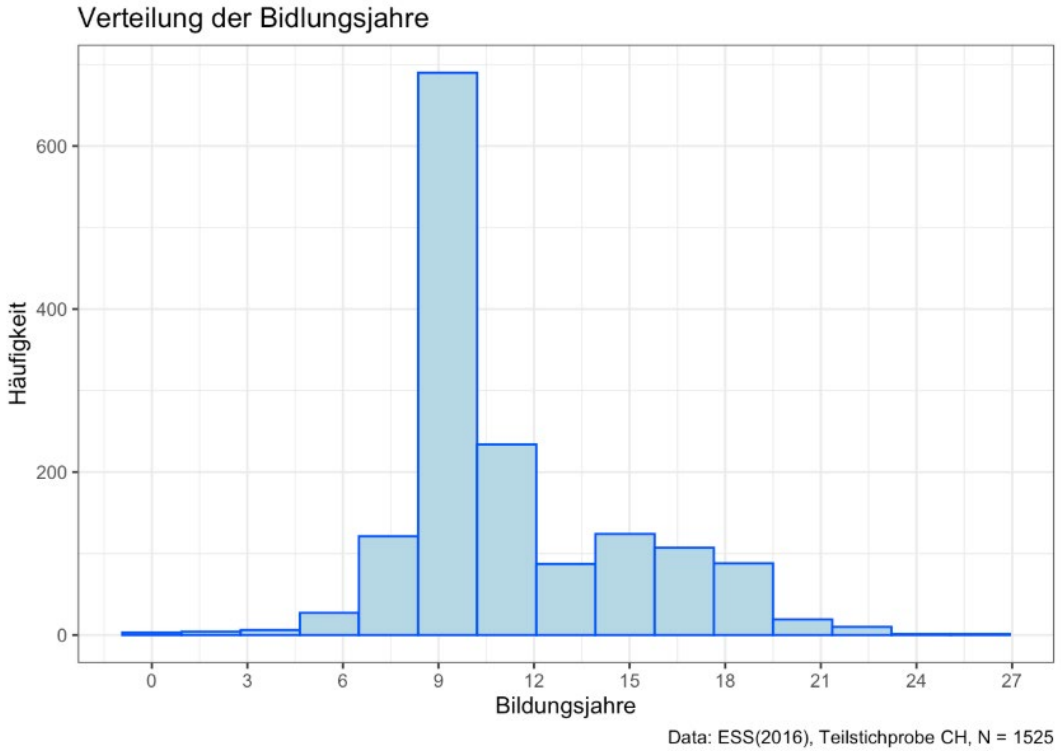
```
ess8 <- read_dta("ESS8e02_1.dta")

ess8_ch <- filter(ess8, cntry == "CH")


ess8_ch_ss_2 <- select(ess8_ch, identifyer = idno,
                       eduyrs, internet = netustm)
```

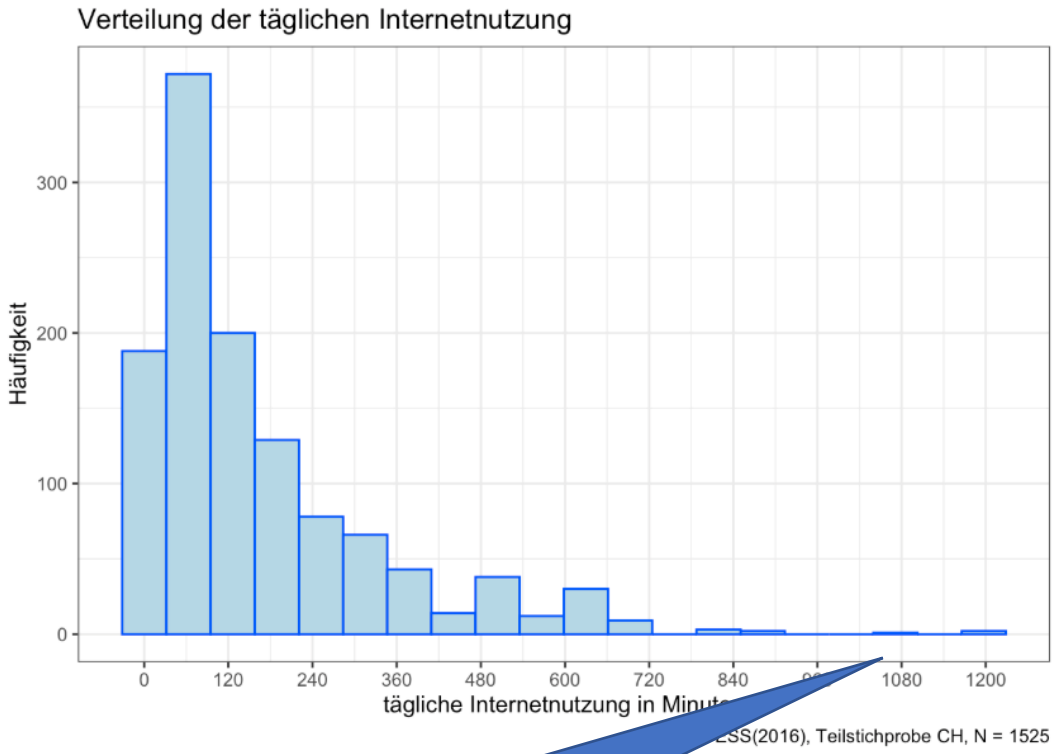
Inspiziere mit **attributes(), summary()** und ggf. **hist()** die beiden Variablen.

- **eduyrs** misst die Anzahl Bildungsjahre einer Person
- **netustm** misst die Zeit (in Minuten), die eine Person täglich im Internet verbringt.
- Es sind keine Rekodierungen erforderlich

➢ netustm ist ausreisserbehaftet (siehe Abbildung!)

Vorbereitungen zur Ausreisserdiagnose





Offene Skala, abseitige Werte: Starke Indizien, aber nicht Belege für Zusammenhangs-Ausreisser

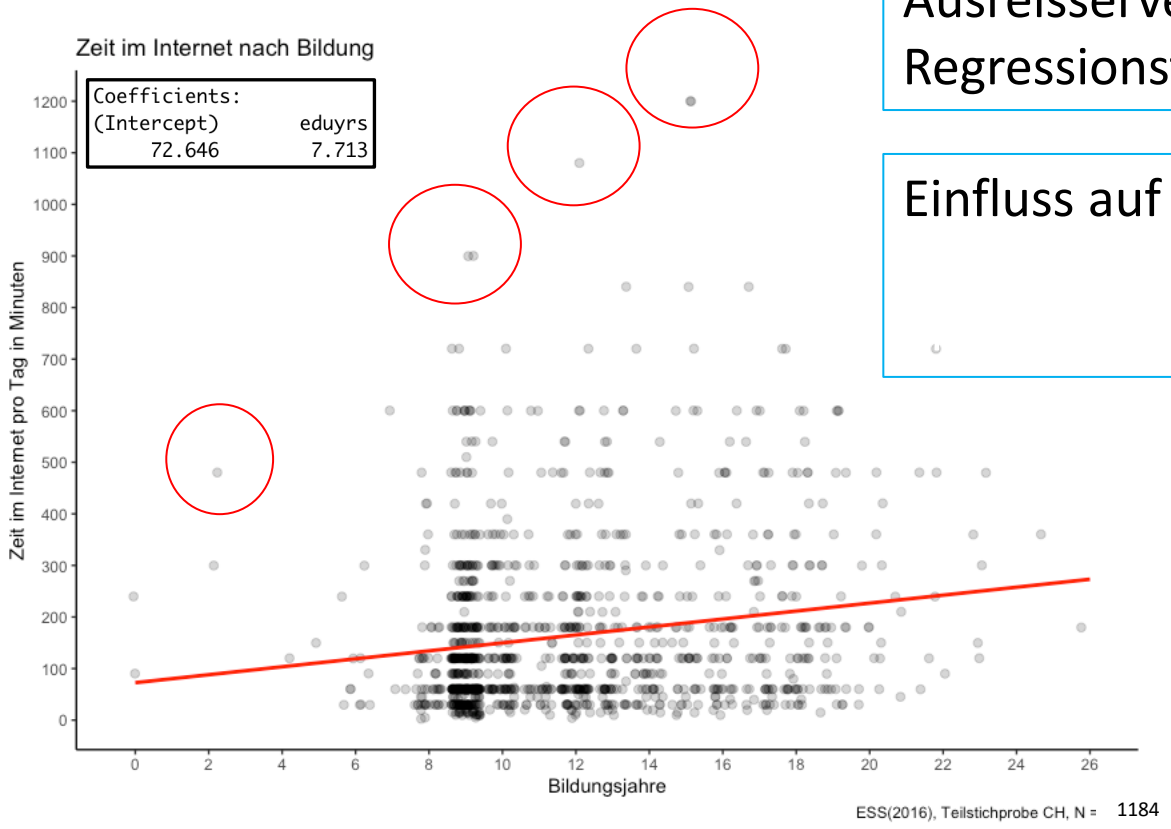
Regressionsanalyse

```
net_model <- lm(internet ~ eduyrs, data = ess8_ch_ss_2)
net_model
```

```
Coefficients:
(Intercept)       eduyrs
     72.646        7.713
```

Ausreisserdiagnose – Visuelle Inspektion

Sind anhand der visuellen Inspektion des Scatterplots potentielle Ausreisser zu erkennen?
Falls ja: Welchen Einfluss haben diese wohl auf das Regressionsergebnis?



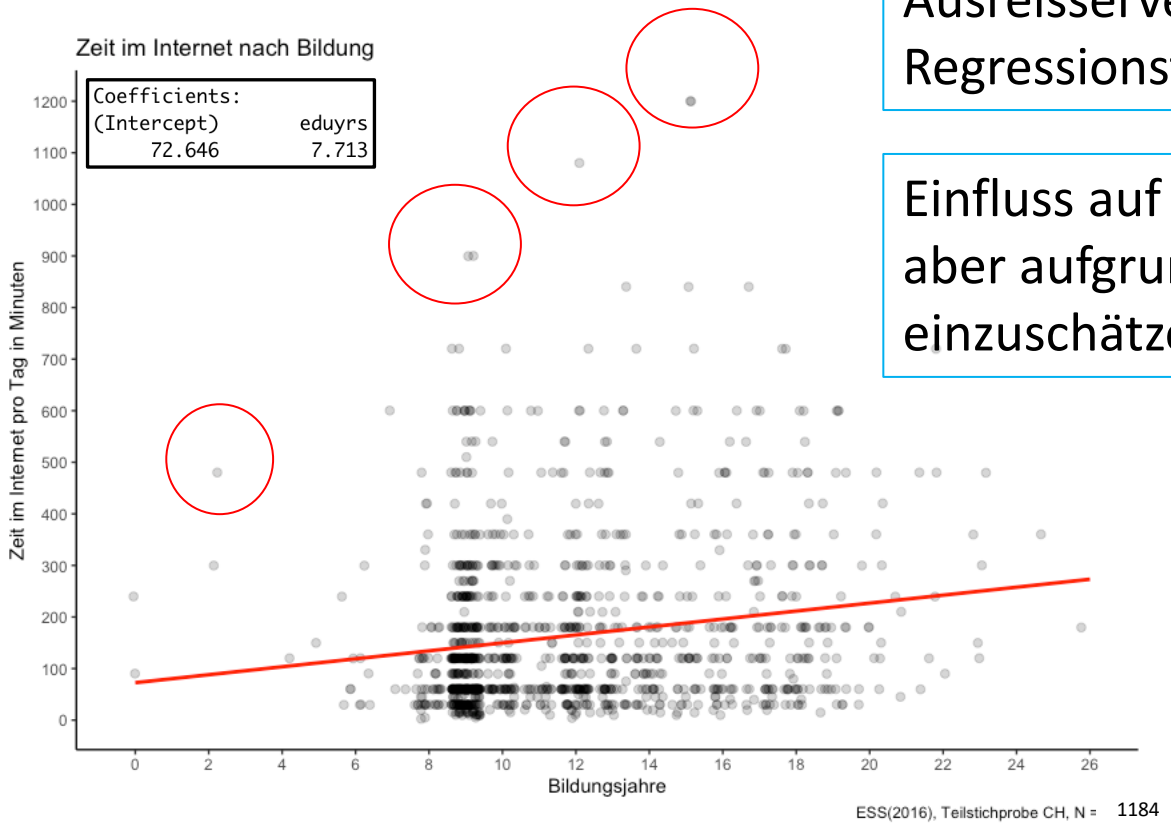
Ausreisserverdacht! Einige Werte liegen sehr weit vom Regressionsfit entfernt bzw. deutlich abseits der Punktwolke.

Einfluss auf den Steigungskoeffizienten:

```
ggplot(ess8_ch_ss_2,
       aes(x = eduyrs, y = internet))+
  geom_jitter(alpha = 0.2, size = 2)+
  scale_x_continuous(breaks = seq(0,26,2))+
  scale_y_continuous(breaks = seq(0,1200,100))+
  geom_smooth(method = "lm", se =F, color = "red")+
  theme_classic()+
  labs(title = "Zeit im Internet nach Bildung",
       caption = "ESS(2016), Teilstichprobe CH, N = 1525",
       y = "Zeit im Internet pro Tag in Minuten",
       x = "Bildungsjahre")
```

Ausreisserdiagnose – Visuelle Inspektion

Sind anhand der visuellen Inspektion des Scatterplots potentielle Ausreisser zu erkennen?
Falls ja: Welchen Einfluss haben diese wohl auf das Regressionsergebnis?



Ausreisserverdacht! Einige Werte liegen sehr weit vom Regressionsfit entfernt bzw. deutlich abseits der Punktwolke.

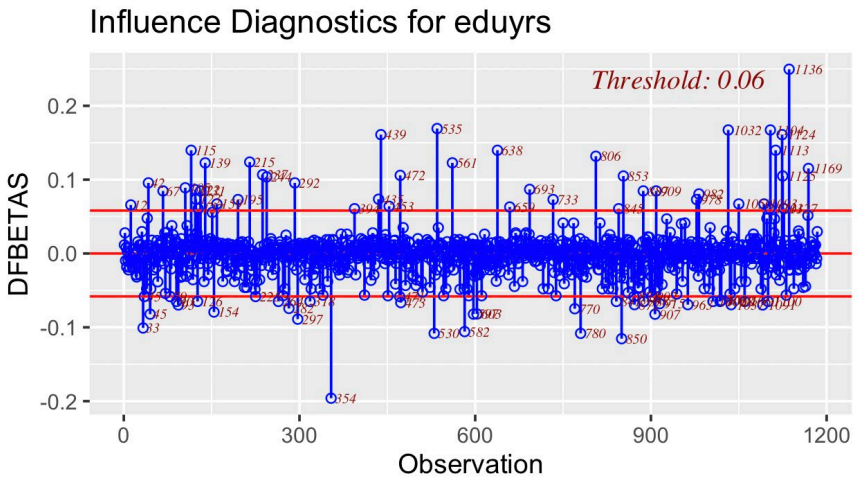
Einfluss auf den Steigungskoeffizienten: Wahrscheinlich positiv, aber aufgrund zentraler UV-Lage der Ausreisser schwierig einzuschätzen!

```
ggplot(ess8_ch_ss_2,
       aes(x = eduyrs, y = internet))+
  geom_jitter(alpha = 0.2, size = 2)+
  scale_x_continuous(breaks = seq(0,26,2))+
  scale_y_continuous(breaks = seq(0,1200,100))+
  geom_smooth(method = "lm", se =F, color = "red")+
  theme_classic()+
  labs(title = "Zeit im Internet nach Bildung",
       caption = "ESS(2016), Teilstichprobe CH, N = 1525",
       y = "Zeit im Internet pro Tag in Minuten",
       x = "Bildungsjahre")
```

Ausreisserdiagnose - Einflussanalyse

- Ausreisser = Extrem einflussreiche Messungen
- Einflussbestimmung über dfbetas (siehe Vorlesung "Ausreisser")
- **dfbetas**: Indikator für messungsspezifische Einflüsse
- Abgleich mit Grenzwert nach Formel: $\frac{2}{\sqrt{n}}$ (Belsley et al.). ***Konkreter Grenzwert in diesem Fall?***

Mit n=1184 erhalten wir über die Formel $\frac{2}{\sqrt{n}}$ einen Grenzwert von 0.06. Folglich sind alle Merkmalsträger mit einem dfbetas unter -0.06 und über 0.06 kritisch.



Abkürzung: Mit dem Befehl **ols_plot_dfbetas** aus dem Package **olsrr** können wir auf einen Blick erkennen, ob mögliche kritische Messungen vorliegen.

```
library(olsrr)
ols_plot_dfbetas(net_model)
```

Erläutert und Interpretiert die Abbildung!

Re-Analyse ohne Ausreisser - Robustheitstest

Was machen wir nun mit den Ausreissern?

- **Falsch:** Ausreisser unreflektiert eliminieren und ausschliesslich ausreisserbefreite Ergebnisse berichten
- **Richtig:** Erstmal prüfen: Messfehler? Datenmanipulation? Invalide Antwort?
- **Richtig:** Checken: Steckt Linearitätsabweichung hinter dem Ausreisserbefund?
- **Richtig:** Ergänzend zur Hauptanalyse mit Ausreissern führen wir als Robustheitstest eine Analyse ohne Ausreisser durch. So können wir abschätzen (und ggf. berichten), wie stabil die Ergebnisse sind (bzw. wie gravierend der Einfluss der Ausreisser ist).

Re-Analyse ohne Ausreisser - Robustheitstest

Was machen wir nun mit den Ausreissern?

- **Richtig:** Ergänzend zur Hauptanalyse mit Ausreissern führen wir als Robustheitstest eine Analyse ohne Ausreisser durch. So können wir abschätzen (und ggf. berichten), wie stabil die Ergebnisse sind (bzw. wie gravierend der Einfluss der Ausreisser ist).
- Kritischer dfbetas: +/-0.06

Füge eine Variable an, die für jede Beobachtung des Datensatzes den dfbetas-Wert enthält...

```
ess8_ch_mitdfb <- cbind(filter(ess8_ch_ss_2, !is.na(eduyrs),!is.na(internet)), data.frame(betas = dfbetas(net_model)))
```

Befehl aus baseR, der Wertelisten mit dfbetas zum Regressionsfit in der Klammer erstellt.

Re-Analyse ohne Ausreisser - Robustheitstest

Was machen wir nun mit den Ausreissern?

- **Richtig:** Ergänzend zur Hauptanalyse mit Ausreissern führen wir als Robustheitstest eine Analyse ohne Ausreisser durch. So können wir abschätzen (und ggf. berichten), wie stabil die Ergebnisse sind (bzw. wie gravierend der Einfluss der Ausreisser ist).
- Kritischer dfbetas: +/-0.06

Füge eine Variable an, die für jede Beobachtung des Datensatzes den dfbetas-Wert enthält...

```
ess8_ch_mitdfb <- cbind(filter(ess8_ch_ss_2, !is.na(eduyrs),!is.na(internet)), data.frame(betas = dfbetas(net_model)))
```

Organisiere die Werteliste der dfbetas als Datenmatrix; betitel die Wertelisten mit den Präfix «betas»

Re-Analyse ohne Ausreisser - Robustheitstest

Was machen wir nun mit den Ausreissern?

- **Richtig:** Ergänzend zur Hauptanalyse mit Ausreissern führen wir als Robustheitstest eine Analyse ohne Ausreisser durch. So können wir abschätzen (und ggf. berichten), wie stabil die Ergebnisse sind (bzw. wie gravierend der Einfluss der Ausreisser ist).
- Kritischer dfbetas: +/-0.06

Füge eine Variable an, die für jede Beobachtung des Datensatzes den dfbetas-Wert enthält...

```
ess8_ch_mitdfb <- cbind(filter(ess8_ch_ss_2, !is.na(eduyrs),!is.na(internet)), data.frame(betas = dfbetas(net_model)))
```

Verknüpfe die um NA bereinigte Ausgangsdatenmatrix mit den dfbetas

Re-Analyse ohne Ausreisser - Robustheitstest

Was machen wir nun mit den Ausreissern?

- **Richtig:** Ergänzend zur Hauptanalyse mit Ausreissern führen wir als Robustheitstest eine Analyse ohne Ausreisser durch. So können wir abschätzen (und ggf. berichten), wie stabil die Ergebnisse sind (bzw. wie gravierend der Einfluss der Ausreisser ist).
- Kritischer dfbetas: +/-0.06

Füge eine Variable an, die für jede Beobachtung des Datensatzes den dfbetas-Wert enthält...

```
ess8_ch_mitdfb <- cbind(filter(ess8_ch_ss_2, !is.na(eduyrs),!is.na(internet)), data.frame(betas = dfbetas(net_model)))
```

eduyrs Years of full-time education completed	**internet** Internet use, how much time on typical day, in minutes	**betas..Intercept.**	**betas.eduyrs**
19	600	-1.158197e-01	0.1421898015
8	30	-2.491944e-02	0.0203165338
19	180	1.212432e-02	-0.0148848093
10	540	5.620380e-02	-0.0369078056
8	30	-2.491944e-02	0.0203165338
13	300	-7.030375e-04	0.0078234735

Kritischer Wert, weil grösser als 0.06

Re-Analyse ohne Ausreisser - Robustheitstest

Nun können wir alle Messungen ausschliessen, die als Ausreisser identifiziert wurden…

```
ess8_noOut <- filter(ess8_ch_mitdfb, betas.eduyrs > -0.06 & betas.eduyrs < 0.06)
```

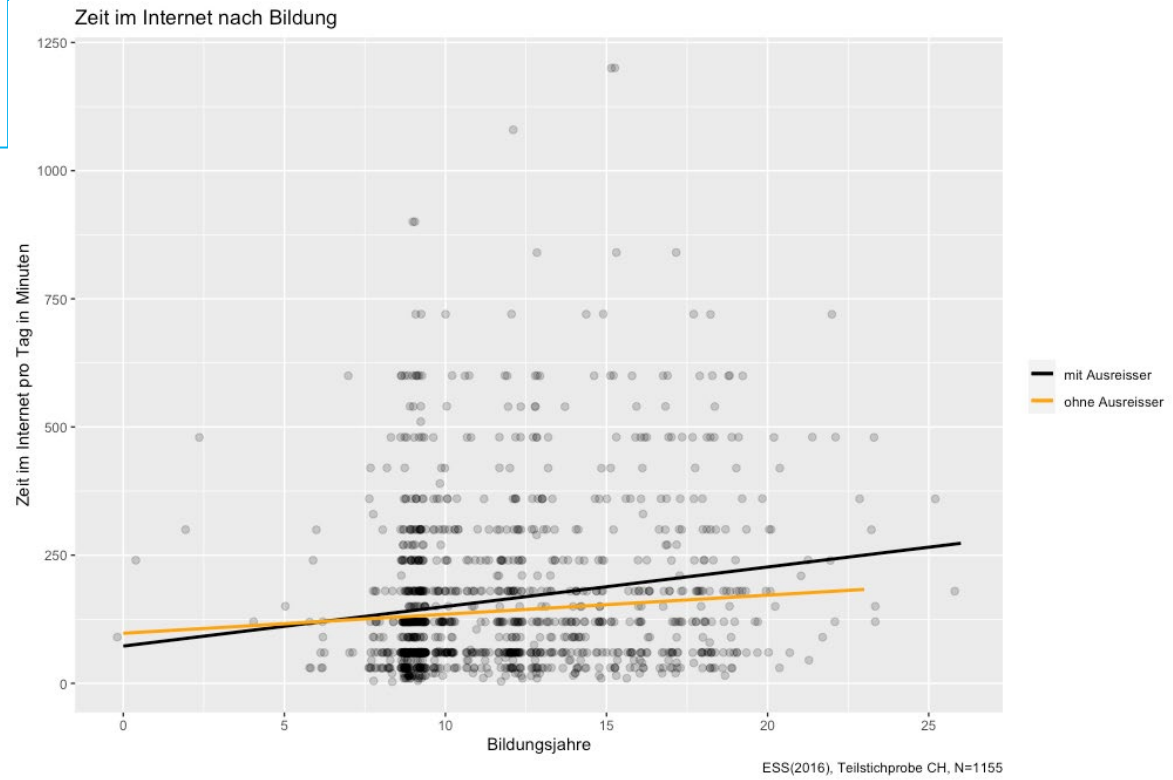
… und den Robustheitstest (=Regression ohne Ausreisser) durchführen

Ausgangsergebnisse

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   72.646     16.171   4.492 7.74e-06 ***
eduyrs         7.713      1.313   5.875 5.48e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ohne Ausreisser

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   97.627     14.065   6.941 6.62e-12 ***
eduyrs         3.718      1.177   3.159  0.00163 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Kommt ihr zu einer ähnlichen Einschätzung der Effektgrösse in den Analysen mit und ohne Ausreisser?

Nein, der Koeffizient verringert sich um etwas mehr als die Hälfte. Das Regressionsergebnis ist aus inhaltlicher Perspektive ***nur bedingt robust*** gegenüber einem Ausreisserausschluss.

Re-Analyse ohne Ausreisser - Robustheitstest

Nun können wir alle Messungen ausschliessen, die als Ausreisser identifiziert wurden…

```
ess8_noOut <- filter(ess8_ch_mitdfb, betas.eduyrs > -0.06 & betas.eduyrs < 0.06)
```

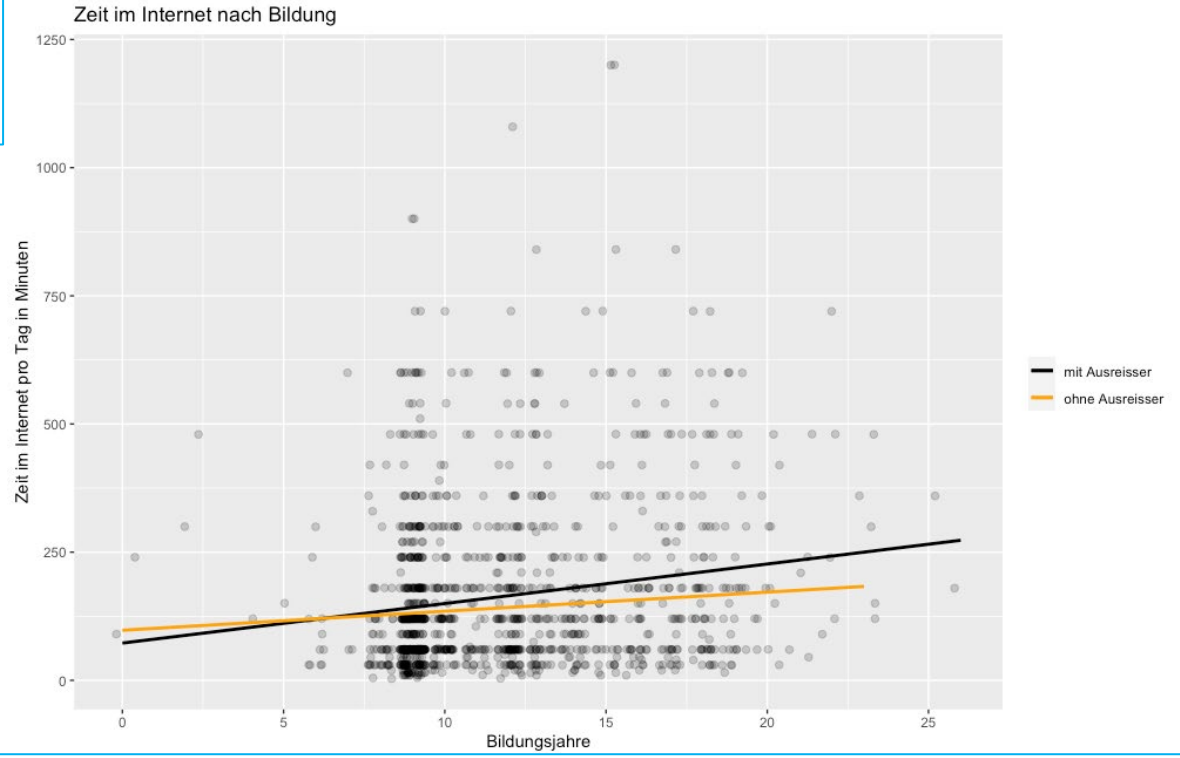
… und den Robustheitstest (=Regression ohne Ausreisser) durchführen

Ausgangsergebnisse

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   72.646     16.171   4.492 7.74e-06 ***
eduyrs         7.713      1.313   5.875 5.48e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ohne Ausreisser

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   97.627     14.065   6.941 6.62e-12 ***
eduyrs         3.718      1.177   3.159  0.00163 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Konklusion: Wir würden in der Darstellung unserer Analyse den Hinweis ergänzen, dass die Ergebnisse aus inhaltlicher Perspektive nur bedingt robust gegenüber Ausreisserausschluss sind. Ausserdem sinnvoll: Auseinandersetzung mit Messung und Zusammenhangsform als mögliche Ursache für das Ausreisserproblem.

Hausaufgabe mit Selbstüberprüfung:

- Übung zur Sitzung IV Linearität und Ausreisser auf der Tutoratswebseite.