Statistik 2 – Tutorate
Sitzung 3: Regressionsanalyse Basics

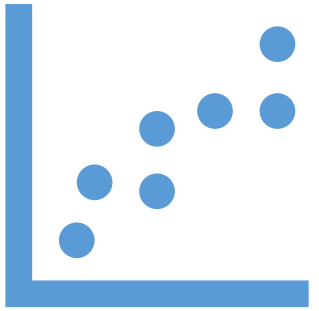
Marco Giesselmann, Rémy Blum, Federica Bruno, Rebecca Hobel, Kristina Trajkovic

Lernziele dieser Sitzung

Vorbereitung

Fragestellung entwickeln

Datenmanagement



Regressionsanalyse

Regressionskoeffizient *b* bestimmten

Visualisierung

Interpretation

0 Theoretische Vorüberlegungen

Fragestellung: Wie beeinflusst Bildung die Einstellungen zur Migration?

Formuliert und begründet eine prüfbare Hypothese

Je höher die Bildung desto positiver sind die Einstellungen zur Migration ✓

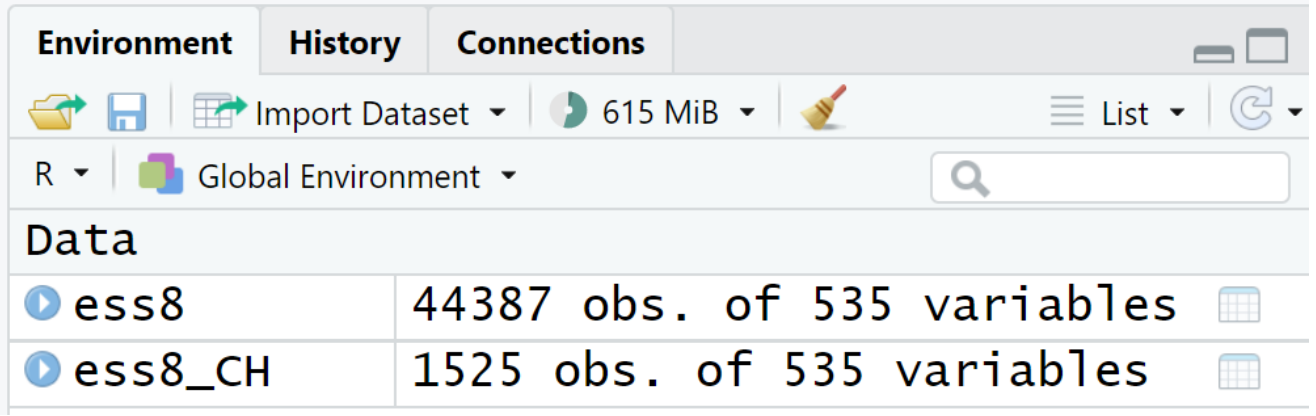
*Höhere Bildung **führt** zu positiveren Einstellungen zur Migration* ✓✓

Teil 1: Datenmanagement

1.1 Filtern von Merkmalsträgern

- Aktiviere tidyverse
- Lade/Importiere den ESS8
- Begrenze ihn auf Messungen aus der Schweiz

```
library(tidyverse)
ess8 <- read_dta("C:/Daten/ESS/ESS8e02_2.dta")
ess8_CH <- filter(ess8, cntry == "CH")
```



1.2 Selektieren von Merkmalen

Fragestellung: Wie beeinflusst Bildung die Einstellung zur Migration?

→ Welche Variablen zur Messung der beiden Konzepte?

→ Sucht mittels **labelled::look_for()**, **hmisc::contents()**, oder im Codebook nach passenden Variablen zur Operationalisierung.

Reminder: Variablensuche per look_for…

```
library(labelled)
codebook<-look_for(ess8_CH)
```

codebook	535 obs. of 7 variables
ess8	44387 obs. of 535 variables
ess8_CH	1525 obs. of 535 variables

m_m_data | reg.Rmd* | ess8 | Untitled1* | codebook | preset_linsen_v6.R

Filter | migra

	pos	variable	label	col_type	missing	levels	value
1	98	imsmetn	Allow many/few immigrants of same race/ethnic group as majority	dbl+lbl	45	NULL	c(`Allc
2	99	imdfetn	Allow many/few immigrants of different race/ethnic group from majority	dbl+lbl	44	NULL	c(`Allc
3	100	impcntr	Allow many/few immigrants from poorer countries outside Europe	dbl+lbl	47	NULL	c(`Allc
4	101	imbgeco	Immigration bad or good for country's economy	dbl+lbl	29	NULL	c(`Bad
5	102	imueclt	Country's cultural life undermined or enriched by immigrants	dbl+lbl	14	NULL	c(`Cul
6	103	imwbcnt	Immigrants make country worse or better place to live	dbl+lbl	36	NULL	c(`Wo
7	225	imsclbn	When should immigrants obtain rights to social benefits/services	dbl+lbl	59	NULL	c(`Imr

1.2 Selektieren von Merkmalen

Fragestellung: Wie beeinflusst Bildung die Einstellung zur Migration?
→ Welche Variablen zur Messung der beiden Konzepte?
→ Sucht mittels **labelled::look_for()**, **hmisc::contents()**, oder im Codebook nach passenden Variablen zur Operationalisierung.

BILDUNG:	EINSTELLUNGEN ZUR MIGRATION:
• **edulvlb**: Highest level of education • **eisced**: Highest level of education, ES – ISCED • **edlvdch**: Highest level of education, Switzerland • **eduyrs**: Years of full-time education completed	• **imsmetn**: Allow many/few immigrants of same race/ethnic group as majority • **Imdfetn:** Allow many/few immigrants of different race/ethnic group as majority • **imbgeco**: Immigration bad or good for country's economy • **imueclt**: Country's cultural life undermined or enriched by immigrants

Die Konzepte werden jeweils durch unterschiedliche Indikatoren abgebildet. In der Praxis hinge es von unseren theoretischen Überlegungen und vom Forschungsstand ab, mit welchen dieser Variablen wir arbeiten. In diesem didaktisch orientierten Beispiel arbeiten wir aus pragmatischen Gründen mit **eduyrs** und **imueclt**.

1.2 Selektieren von Merkmalen

Reduziere auf die analyserelevanten Variablen: **eduyrs, imueclt** und den/die **Identifier**

```
ess8_CH_ss <- select(ess8_CH, idno, eduyrs, imueclt)
```

Global Environment

Name	Type	Length	Size	Value
ess8	tbl_df	534	182.4 MB	44387 obs. of 534 varia…
ess8_CH	tbl_df	534	7.7 MB	1525 obs. of 534 variab…
ess8_CH_ss	tbl_df	3	41 KB	1525 obs. of 3 variables

	eduyrs Years of full-time education completed	**imueclt** Country's cultural life undermined or enriched by immigrants
1	9	6
2	12	7
3	18	10
4	9	3

1.3 Inspektion der Daten

Analysiere und inspiziere die beiden Variablen (univariate deskriptive Analyse): Variablenklasse, Mittelwerte, fehlende Werte, Verteilungseigenschaften

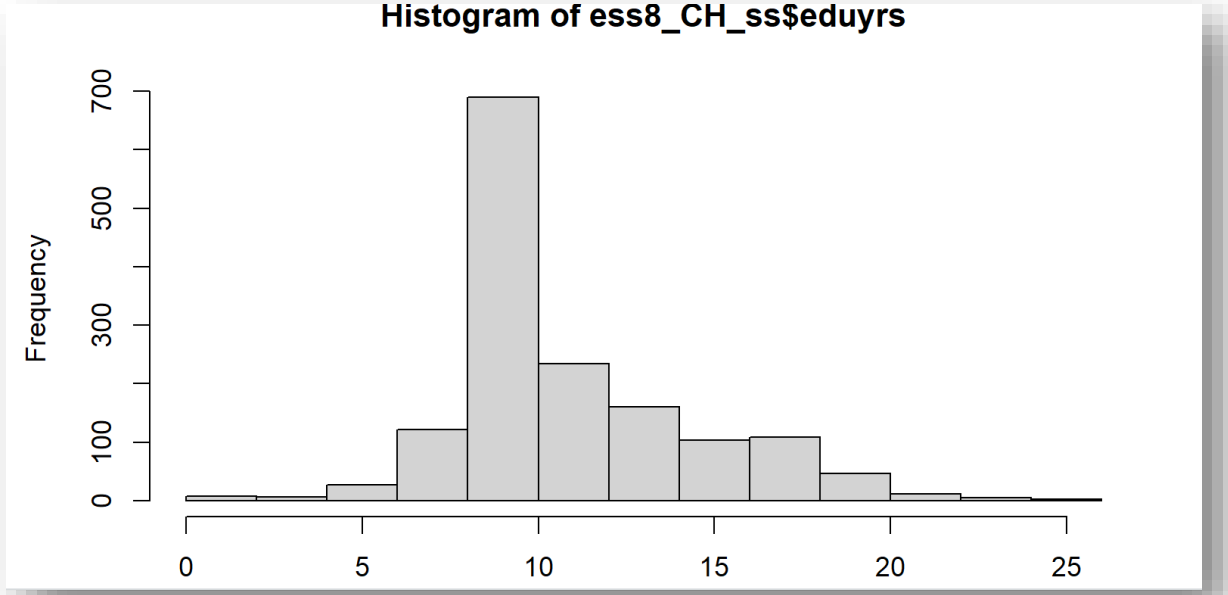
```
attributes(ess8_CH_ss$eduyrs)
summary(ess8_CH_ss$eduyrs)
hist (ess8_CH_ss$eduyrs)
sd(ess8_CH_ss$eduyrs, na.rm = TRUE)
```

```
$label
[1] "Years of full-time education completed"

$class
[1] "haven_labelled" "vctrs_vctr"     "double"

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0     9.0    10.0    11.3    13.0    26.0       3
[1] 3.496909
```

Checke immer auch die Anzahl NAs einer Variable. Sind es viele (Daumenregel >10%), muss dies im Rahmen der Auswertung explizit **berichtet oder begründet** werden. Gelegentlich wird dann auch auf eine alternative Variable ausgewichen.



1.3 Inspektion der Daten

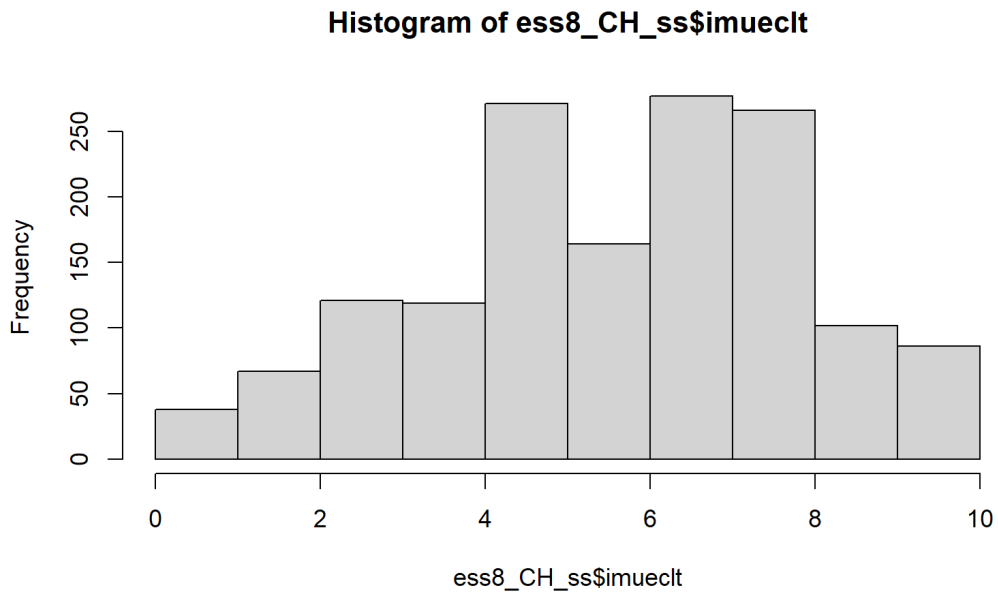
Analysiere und inspiziere die beiden Variablen (univariate deskriptive Analyse): Variablenklasse, Mittelwerte, fehlende Werte, Verteilungseigenschaften

```
attributes(ess8_CH_ss$imueclt)
summary(ess8_CH_ss$imueclt)
hist (ess8_CH_ss$imueclt)
sd(ess8_CH_ss$imueclt, na.rm = TRUE)
```

```
$labels
Cultural life undermined

$class
[1] "haven_labelled" "vctrs_vctr"     "double"

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   5.000   6.000   6.072   8.000  10.000      14
[1] 2.258924
```



Checke immer auch die Anzahl NAs einer Variable. Sind es viele (Daumenregel >10%), muss dies im Rahmen der Auswertung explizit **berichtet oder begründet** werden. Gelegentlich wird dann auch auf eine alternative Variable ausgewichen.

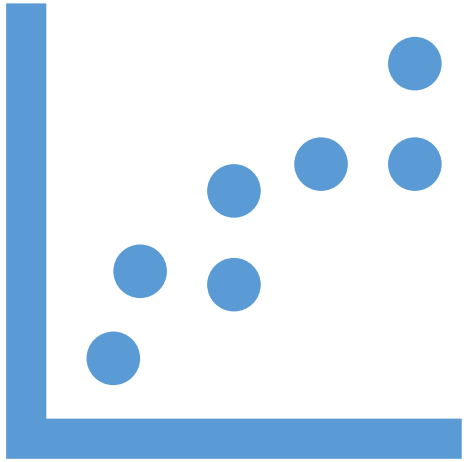
1.5 Festlegung des Analyseformates

Beide Variablen haben **metrisches Messniveau**. Zusammenhangsanalyse und Hypothesentest erfolgen somit durch die **Regressionsanalyse**, vgl. Vorlesung:

Übersicht: Zusammenhangsmasse (bivariate Statistik)

Unabhängige Variable \ Abhängige Variable	Nicht-Metrisch	Metrisch
Nicht-Metrisch	*Tabellenanalyse* *Prozentsatzdifferenz* [1] *Lambda* [2] *Chi-quadrat / Cramer's V* [3] *p-Wert («Chi²-test»)* [4]	*Varianzanalyse* *Mittelwertdifferenz* [1] *Eta-Quadrat* [2] *p-Wert («t-test»)* [4]
Metrisch		*Regressionsanalyse* *Regressionskoeffizient* [1] ***Korrelationskoeffizient r*** [3] *(=beta)*

Teil 2: Visualisierung - Streudiagramm

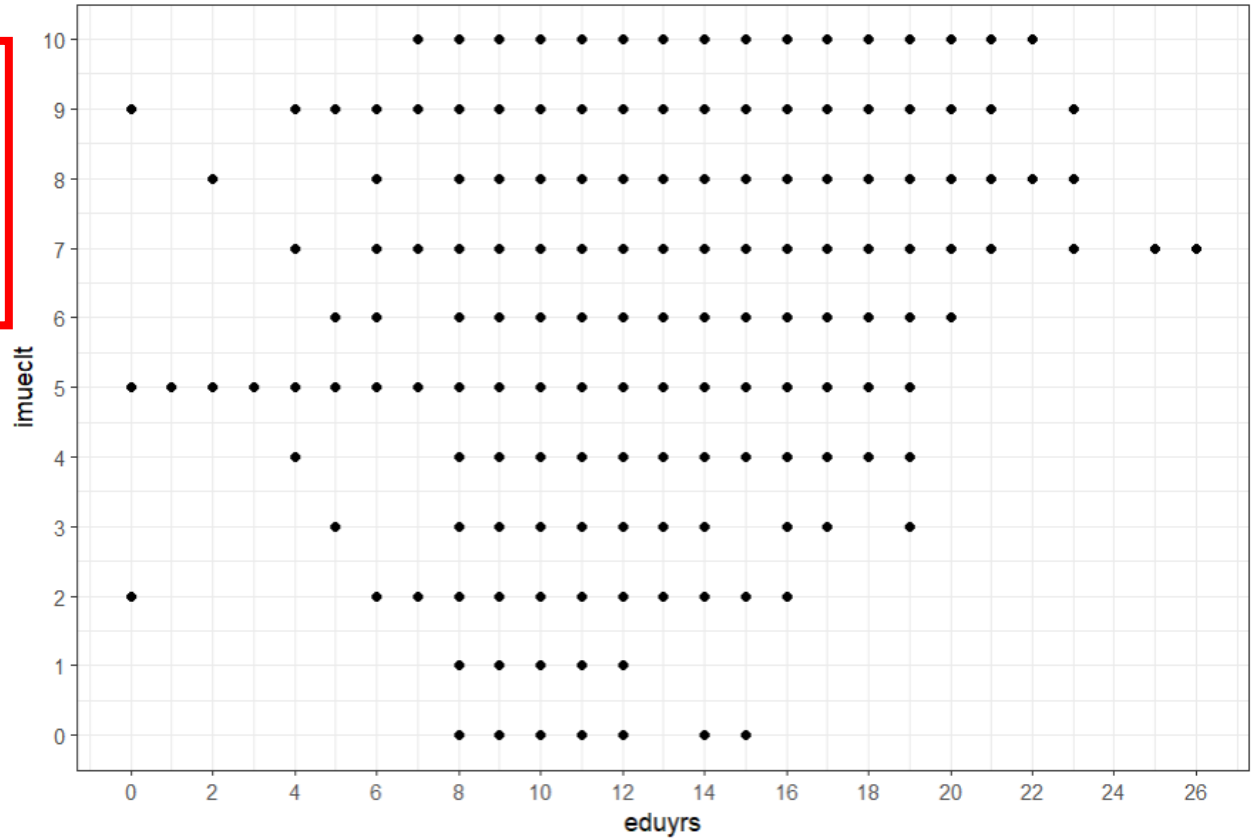


1.6 Visualisierung des Zusammenhangs: Scatterplot mit ggplot()

Beschreibt die einzelnen Funktionen des Befehls:

(a) Welches Element der Abbildung wird mit den Befehlen jeweils modifiziert?
(b) Was passiert, wenn in die «scale»-Befehle andere Werte eingesetzt werden?
(c) Was passiert, wenn die «scale»—Befehle ganz weggelassen werden?

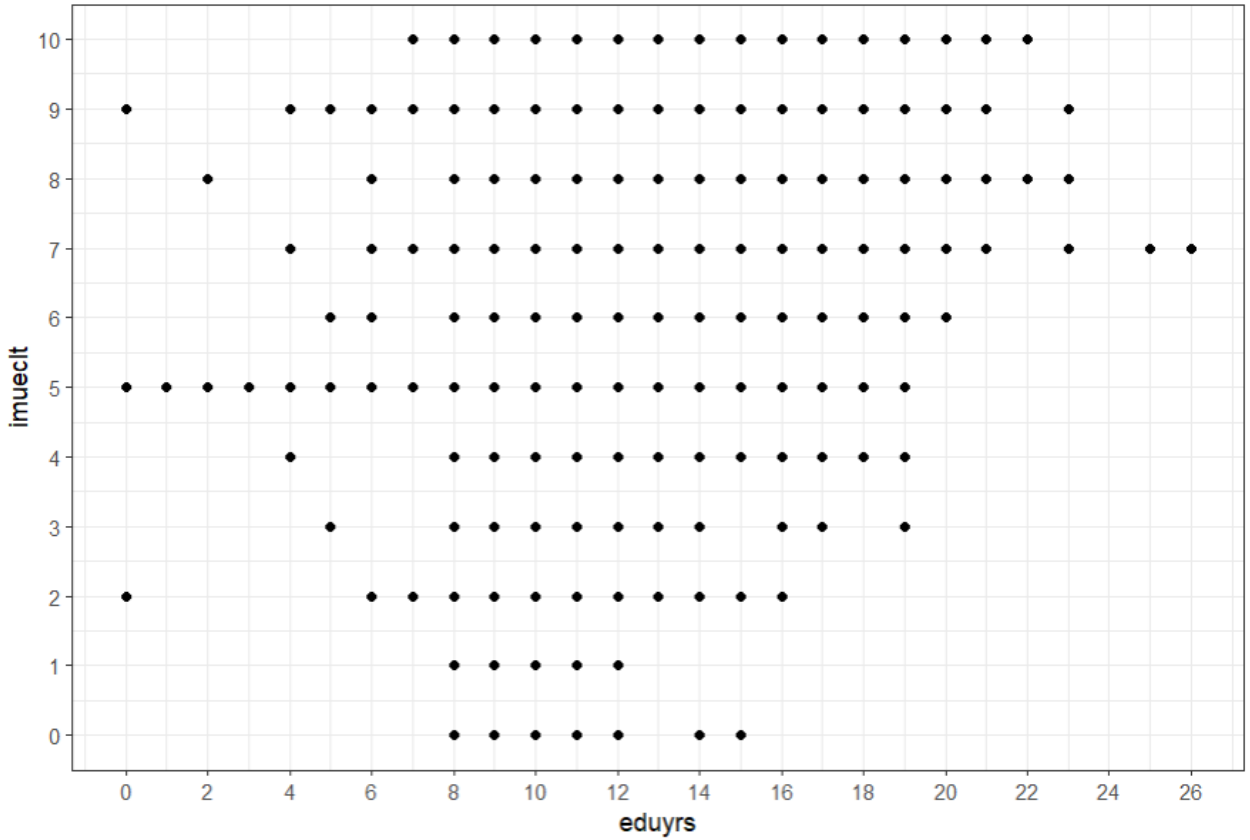
```
ggplot(ess8_CH_ss,
       aes(x = eduyrs, y = imueclt)) +
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_point()+
  theme_bw()
```



1.6 Visualisierung des Zusammenhangs: Scatterplot

Verwende die Datenmatrix «ess8_CH_ss»

```
ggplot(ess8_CH_ss,
       aes(x = eduyrs, y = imueclt)) +
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_point()+
  theme_bw()
```

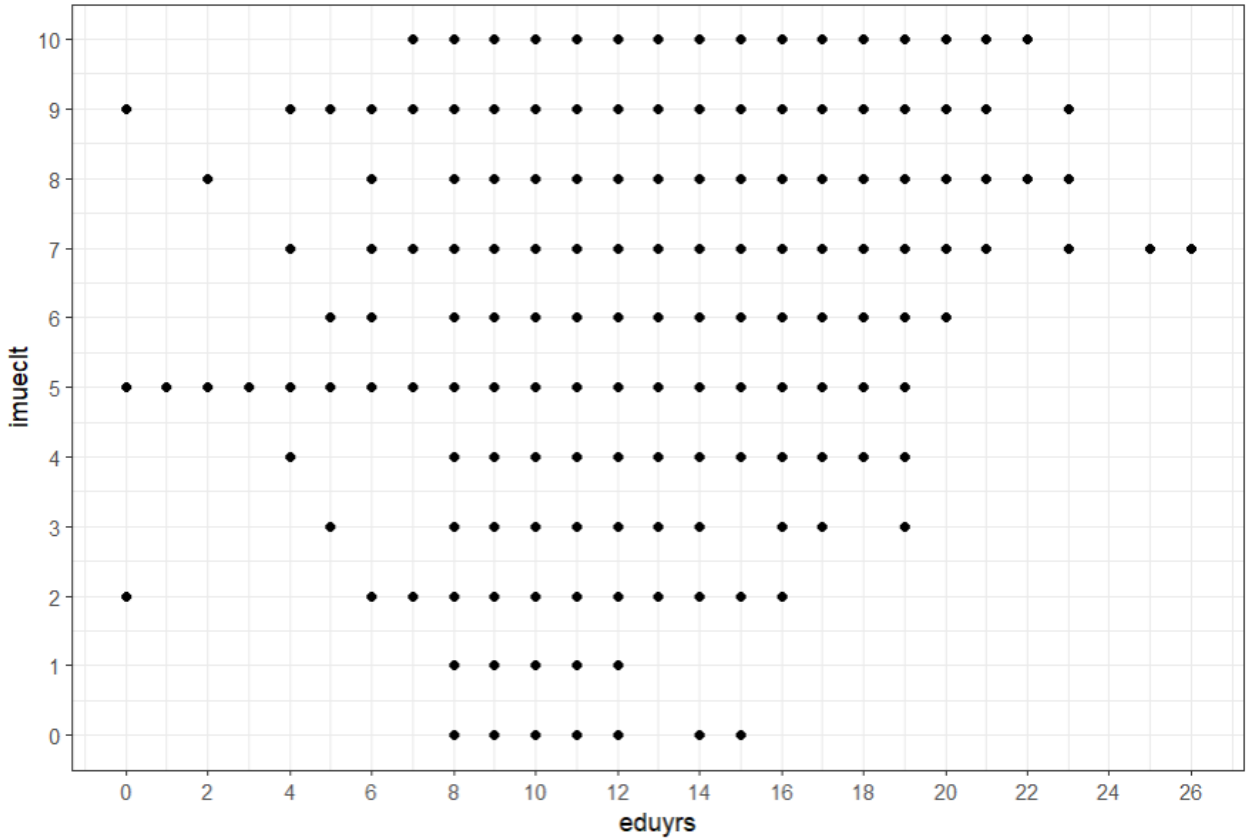


1.6 Visualisierung des Zusammenhangs: Scatterplot

Unterfunktion «aes»:
Wie soll die Abbildung aufgebaut sein?

```
ggplot(ESS8_CH_ss,
       aes(x = eduyrs, y = imueclt)) +
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_point() +
  theme_bw()
```

… «eduyrs» definiert x-Achse, «imueclt» definiert die y-Achse

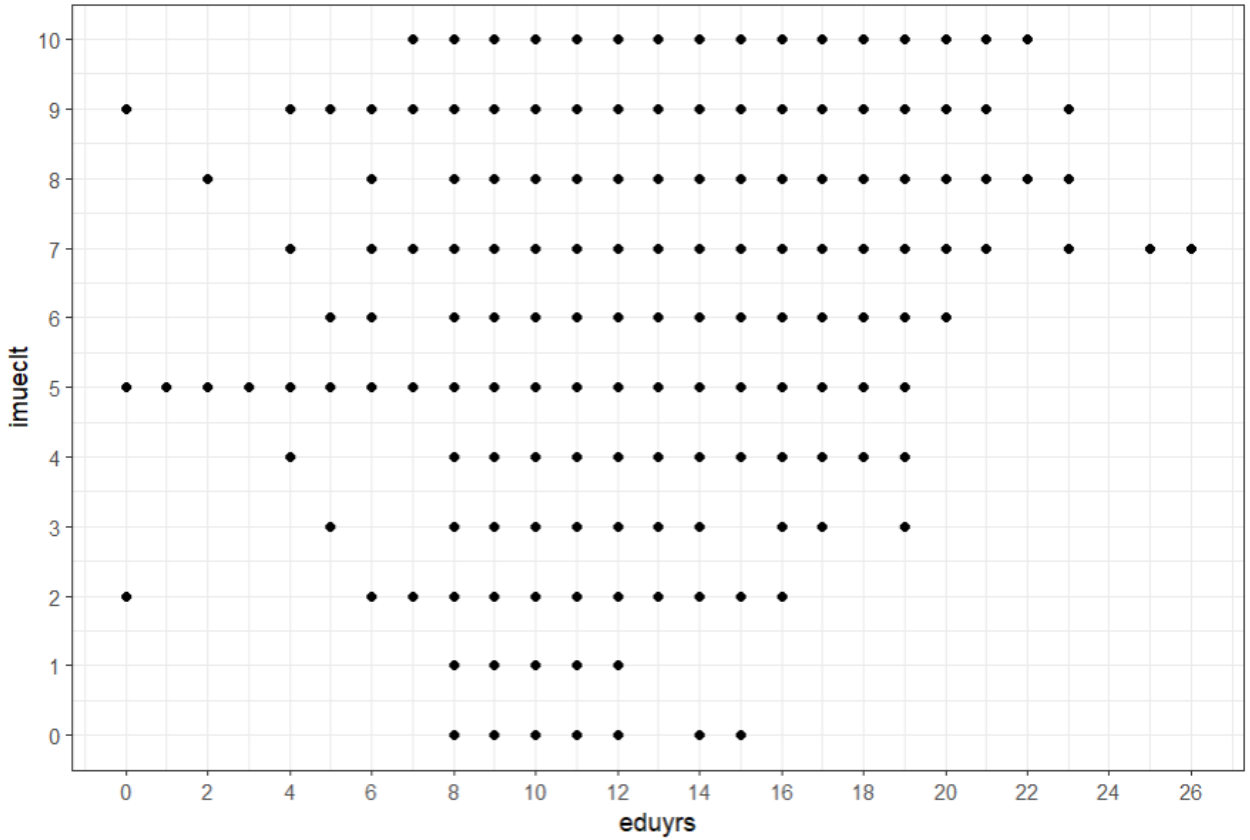


1.6 Visualisierung des Zusammenhangs: Scatterplot

```
ggplot(ess8_CH_ss,
       aes(x = eduyrs, y = imueclt)) +
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_point()+
  theme_bw()
```

Spezifikation der Achsengestaltung

(Achtet immer darauf, dass hinreichend viele Ticks & Label gesetzt sind, aber keine unsinnigen Werte dargestellt)

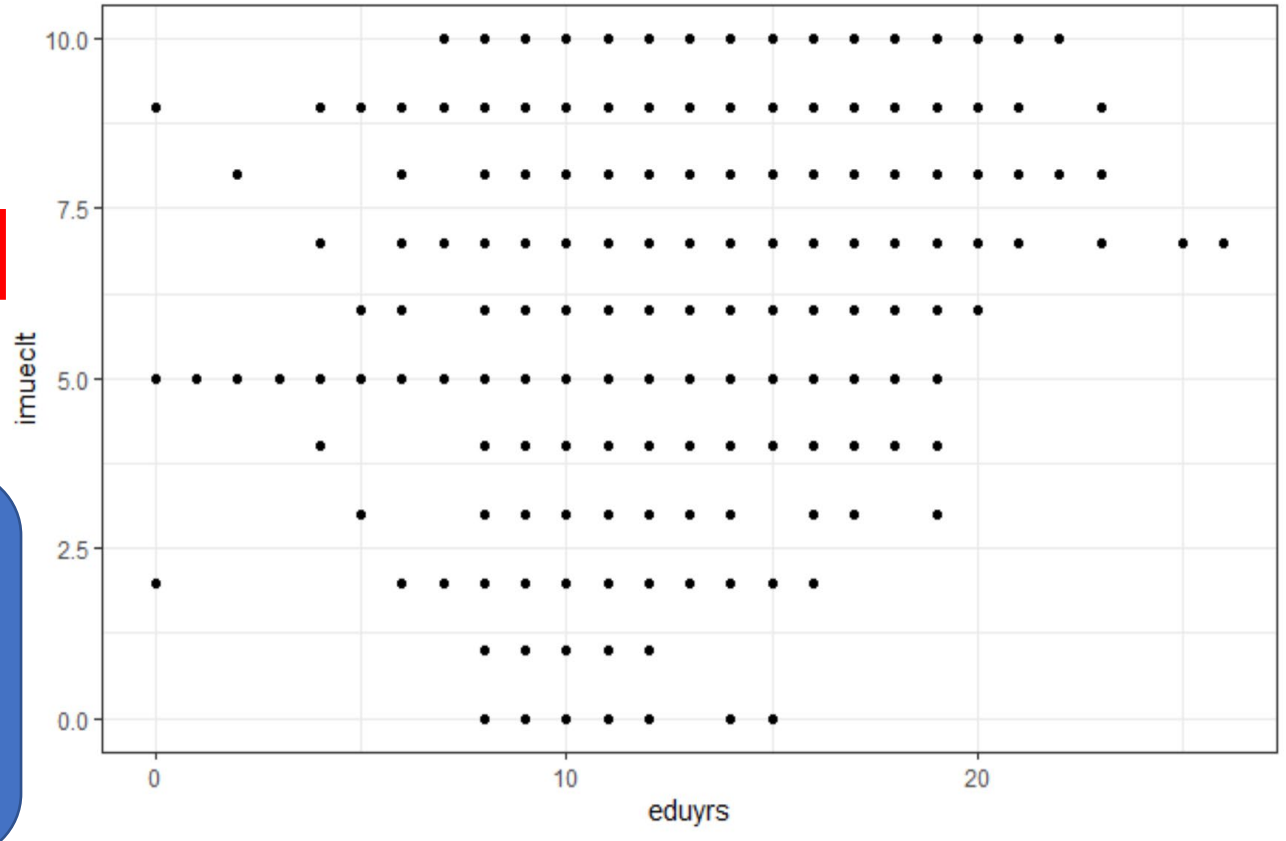


1.6 Visualisierung des Zusammenhangs: Scatterplot

Grundlage der Regressionsanalyse ist die Visualisierung des Zusammenhangs auf Grundlage des **ggplot()** Befehls aus dem **ggplot2** Package.

```
ggplot(ess8_CH_ss,
       aes(x = eduyrs, y = imueclt)) +

  geom_point()+
  theme_bw()
```



Probleme bei Default:

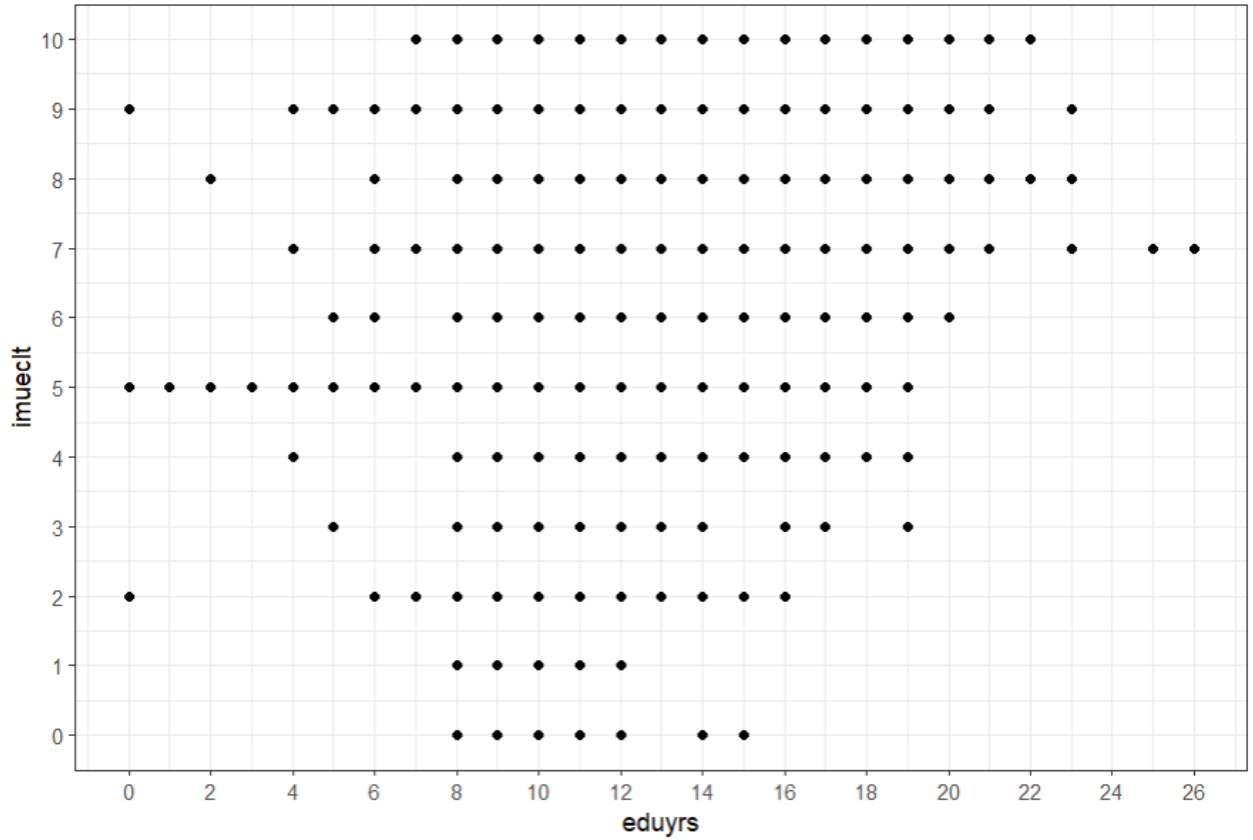
- zu wenig gelabelte Ticks auf der x-Achse
- Dezimalwerte auf der y-Achse: unsinnig bei diskreter Skalenvariable

1.6 Visualisierung des Zusammenhangs: Scatterplot

Grundlage der Regressionsanalyse ist die Visualisierung des Zusammenhangs auf Grundlage des **ggplot()** Befehls aus dem **ggplot2** Package.

```
ggplot(ess8_CH_ss,
       aes(x = eduyrs, y = imueclt)) +
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_point()+
  theme_bw()
```

Stelle die Datenträger bzw. Einheiten des Datensatzes als Punkte dar (scatterplot)

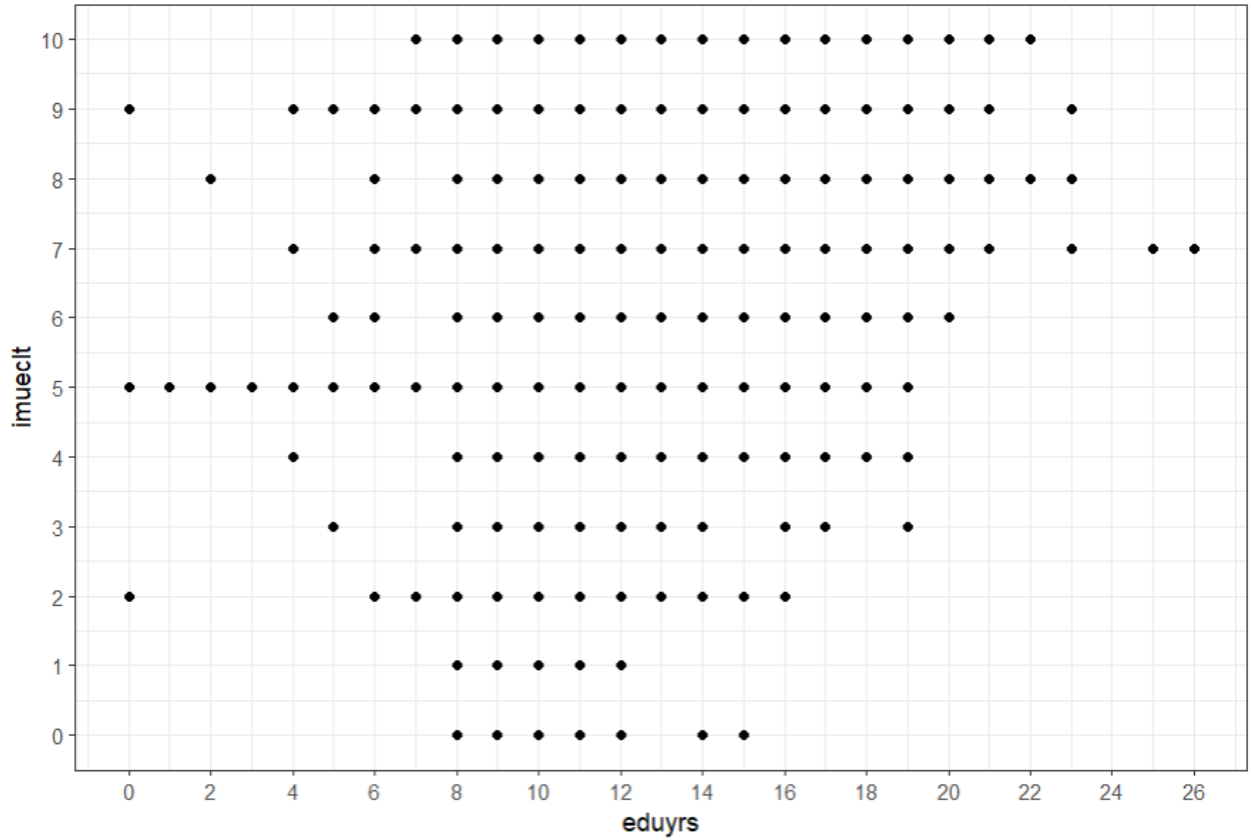


1.6 Visualisierung des Zusammenhangs: Scatterplot

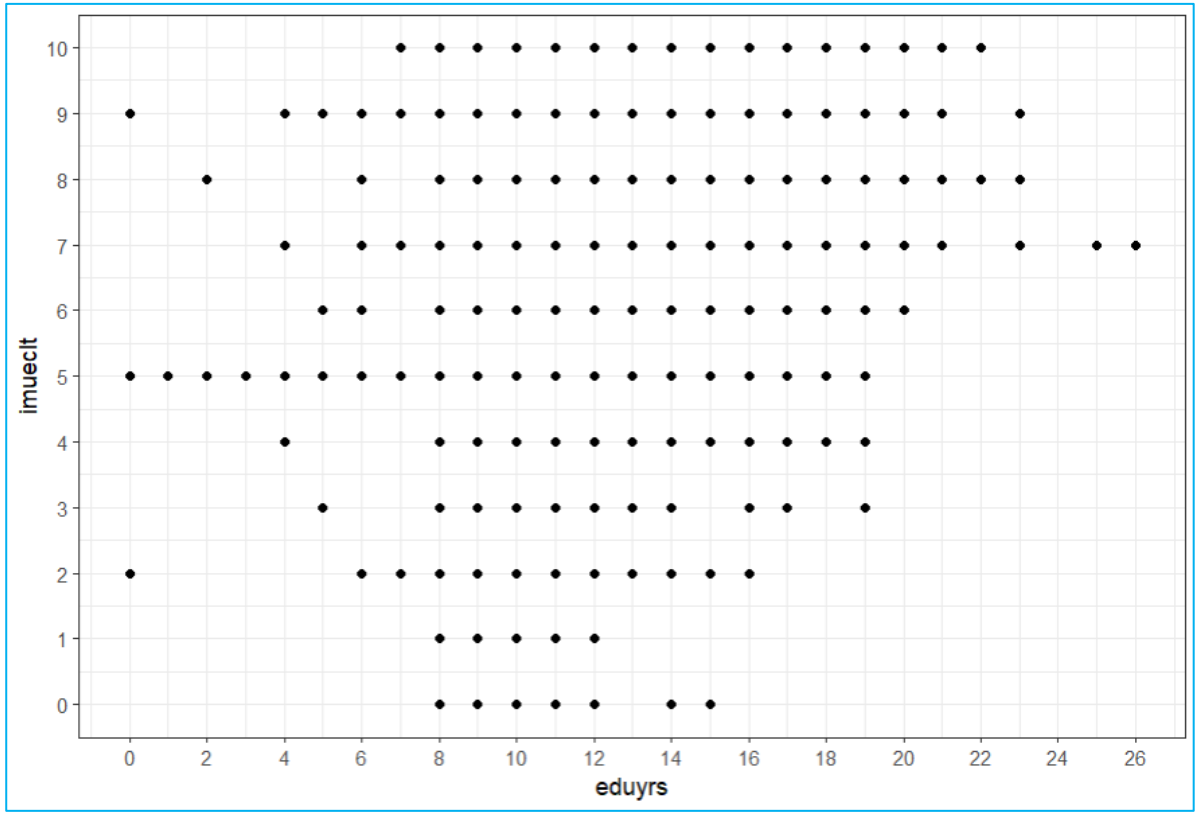
Grundlage der Regressionsanalyse ist die Visualisierung des Zusammenhangs auf Grundlage des **ggplot()** Befehls aus dem **ggplot2** Package.

```
ggplot(ess8_CH_ss,
       aes(x = eduyrs, y = imueclt)) +
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_point()+
  theme_bw()
```

Nutze den «Standardhintergrund»



1.6 Scatterplot



Fragen:
Lässt sich ein Trend erkennen?
Warum ist die Grafik nicht aussagekräftig?

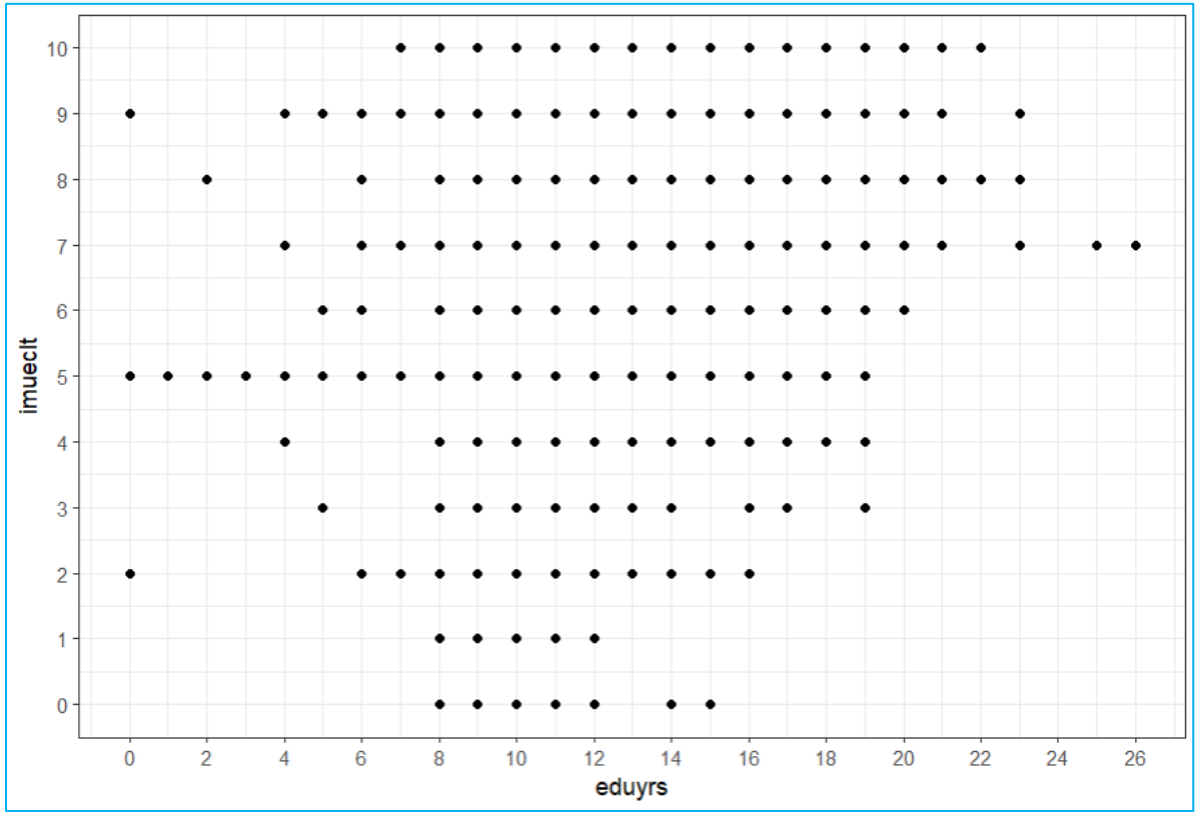
Overplotting sowie eine **unzureichende Beschriftung** schmälern hier die **Aussagekraft**.
Für beide Probleme gibt es Standardlösungen, die auch von ggplot bzw. R unterstützt werden.

1.6 Scatterplot - Beschriftung

Beschriftung:
Standards für die Annotation von Abbildungen insb. Streudiagrammen:

- **Titel:** Merkmale nennen, evtl. Verfahren referenzieren
- **Untertitel:** *Ggf.* Fragetext der AV
- **Achsenbeschriftung:** Messung: verständliche Variablennamen (*Bildungsjahre* statt «eduyrs»)
- **Caption** mit Datenquelle und (Teil-) Stichprobengrösse

Im ggplot können wir all dies mit **labs()** anpassen. Findet heraus, wie! (oder direkt Chatgpt machen lassen)



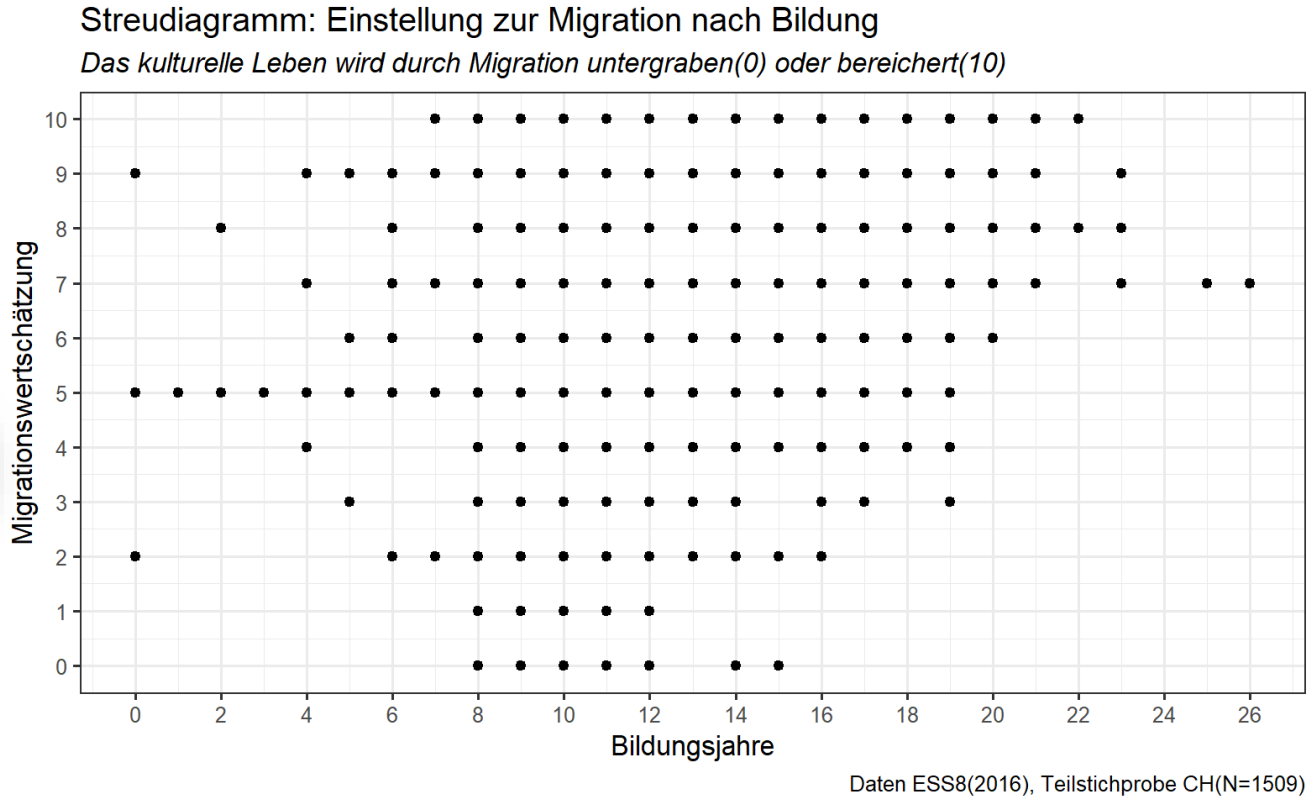
1.6 Scatterplot

```
ggplot(ess8_CH_ss, aes(x = eduyrs, y = imueclt)) +
  geom_point() +
  scale_x_continuous(breaks = seq(0, 26, 2)) +
  scale_y_continuous(breaks = seq(0, 10, 1)) +
  labs(
    title = "Streudiagramm: Einstellung zur Migration nach Bildung",
    subtitle = "Das kulturelle Leben wird durch Migration untergraben(0) oder bereichert(10)",
    x = "Bildungsjahre",
    y = "Migrationswertschätzung",
    caption = "Daten ESS8(2016), Teilstichprobe CH(N=1509)"
  ) +
  theme_bw() +
  theme(plot.subtitle = element_text(face = "italic"))
```

Schon besser. Aber das Overplotting-Problem ist noch nicht gelöst!

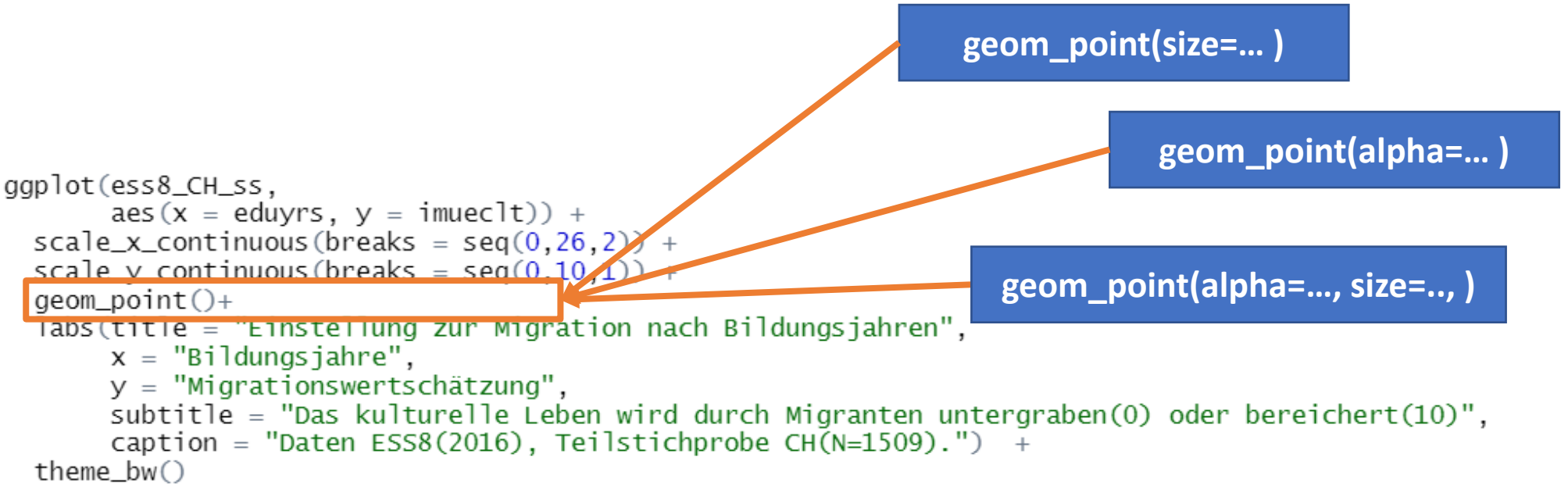
```
Warning message:
Removed 16 rows containing missing values (`geom_point()`).
```

Info zu fehlenden Werten aus ggplot-Meldung



1.6 Scatterplott - Overplotting

Es gibt verschiedene Wege das Overplotting zu lösen

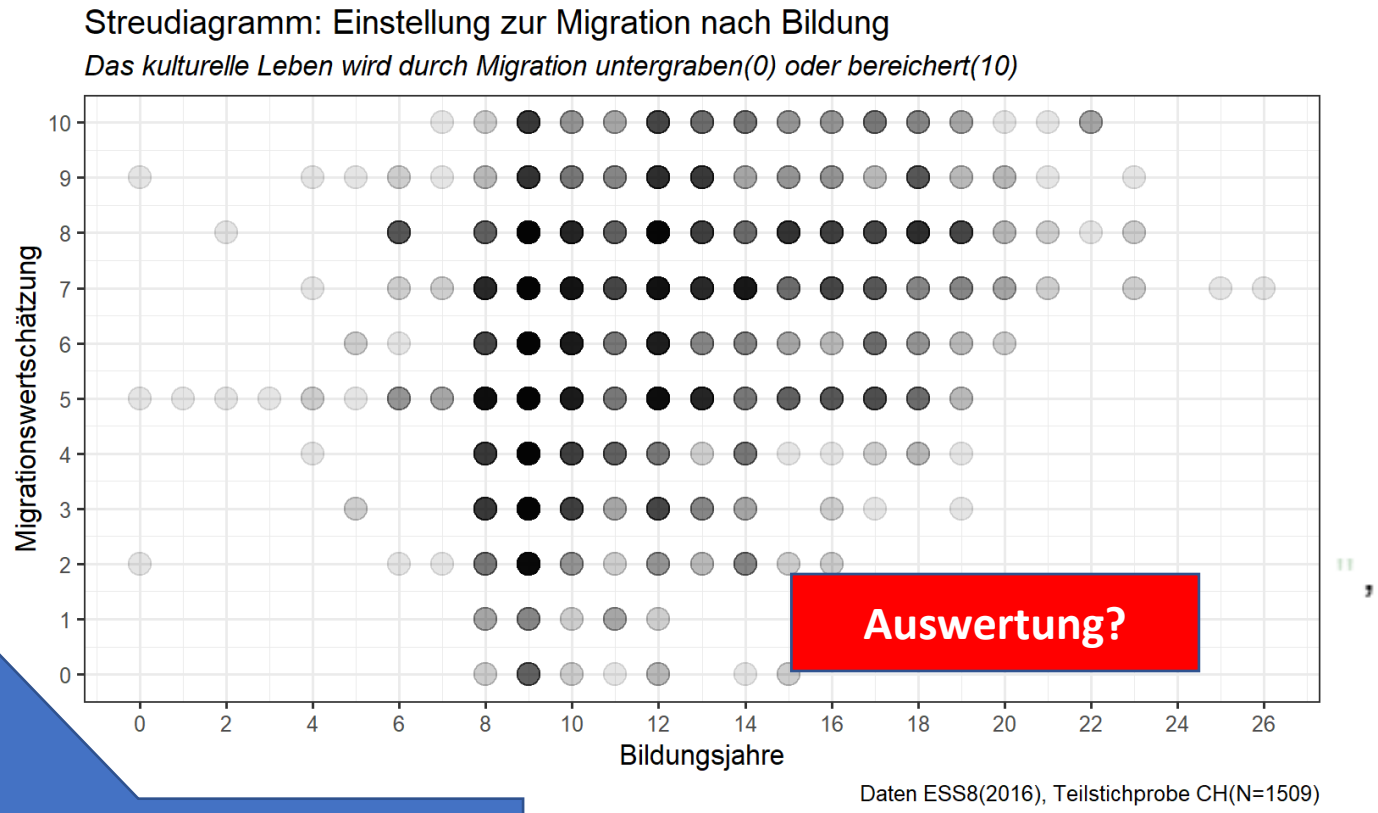



```
ggplot(ess8_CH_ss,
       aes(x = eduyrs, y = imueclt)) +
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_point()+
  labs(title = "Einstellung zur Migration nach Bildungsjahren",
       x = "Bildungsjahre",
       y = "Migrationswertschätzung",
       subtitle = "Das kulturelle Leben wird durch Migranten untergraben(0) oder bereichert(10)",
       caption = "Daten ESS8(2016), Teilstichprobe CH(N=1509).")  +
  theme_bw()
```

Ersetzt den **geom_point()** Befehl durch die hier vorgeschlagenen Alternativen.
Wie verändert sich der Plot? Was bewirken die einzelnen Teilargumente?

1.6 Scatterplott – Overplotting: “alpha”

```
ggplot(ess8_CH_ss, aes(x = eduyrs, y = im
  scale_x_continuous(breaks = seq(0,26,2)
  scale_y_continuous(breaks = seq(0,10,1)
  geom_point(alpha = 0.1, size = 4)+
  labs(title = "Einstellung zur Migration
       x = "Bildungsjahre",
       y = "Migrationswertschätzung",
       subtitle = "Das kulturelle Leben w
       caption = "Daten ESS8(2016), Teils
```



“Alpha” bestimmt die Transparenz der Datenpunkte (0=völlig durchsichtig, 1=voll deckend). Die Farbtiefe der Punkte lässt so auf die Belegungsdichte der einzelnen Merkmalskombinationen schliessen.

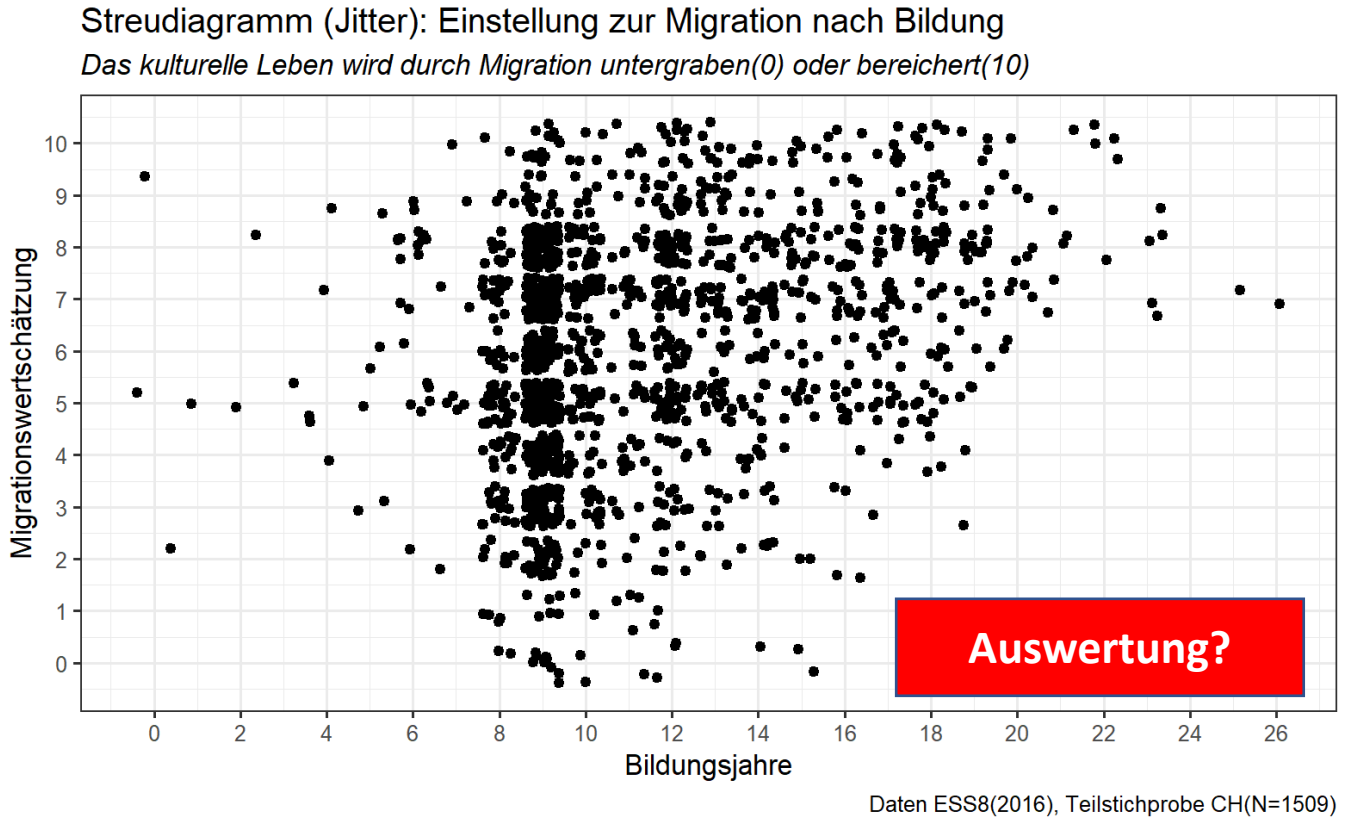
“Size” bestimmt die Grösse der Datenpunkte.

1.6 Scatterplott – Overplotting: “jitter”

geom_jitter()

```
ggplot(ess8_CH_ss, aes(x = eduyrs, y = imuec
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_jitter()+
  labs(title = "Einstellung zur Migration na
       x = "Bildungsjahre",
       y = "Migrationswertschätzung",
       s = "Das kulturelle Leben wird
```

geom_jitter() plottet wie auch **geom_point()** Punkte, fügt diesen aber Zufallsstreuung zu. Übereinanderliegende Punkte werden damit auseinander gezogen und so sichtbar gemacht.

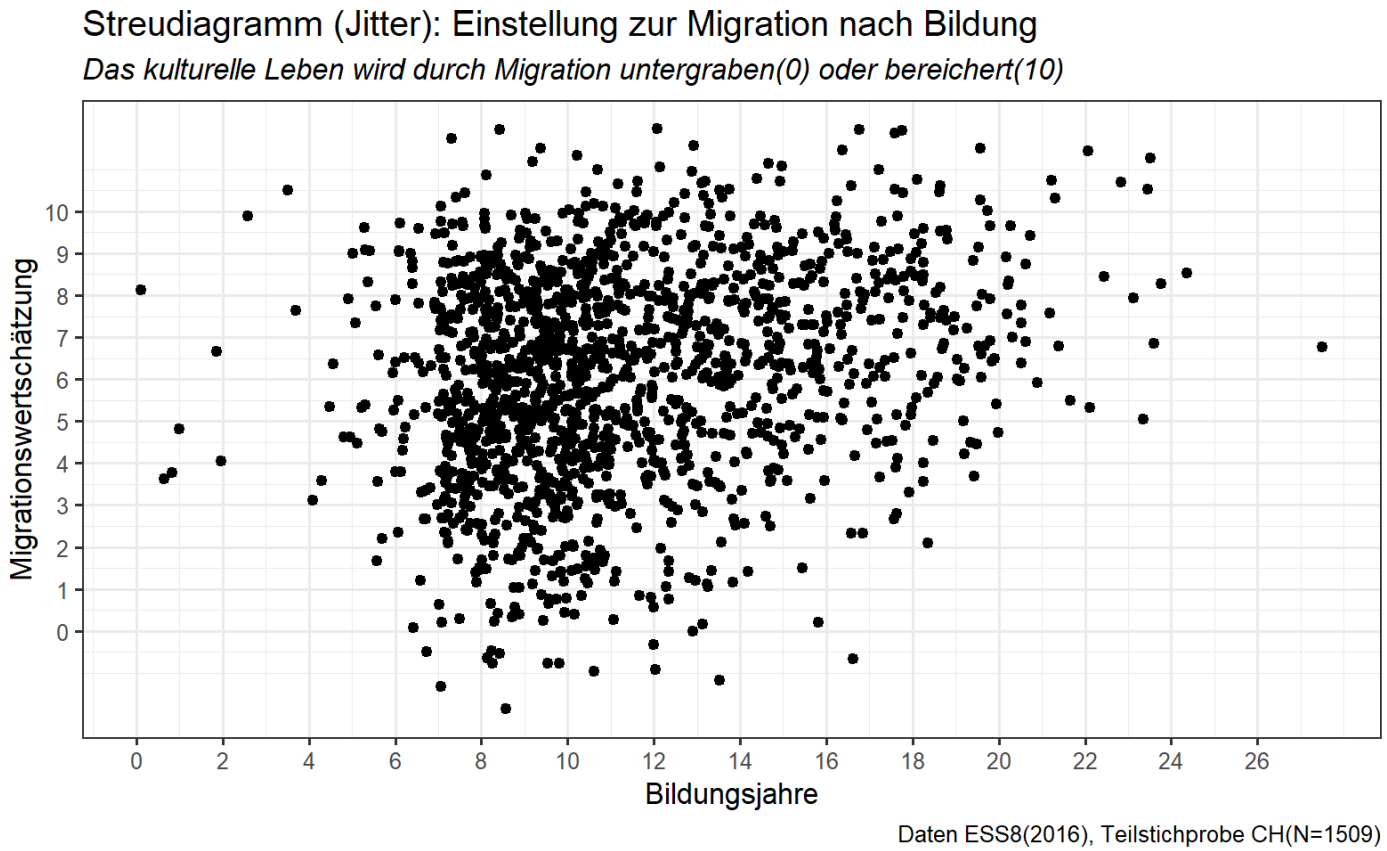


1.6 Scatterplott – Overplotting: “jitter”

geom_jitter(height=, width=)

```
ggplot(ess8_CH_ss, aes(x = eduyrs, y = i
  scale_x_continuous(breaks = seq(0,26,2
  scale_y_continuous(breaks = seq(0,10,1
  geom_jitter(height= 2, width = 2)+
  labs(title = "Ein ellung zur Migratic
       x = "Bildungs re",
       y = "Migratio schätzung",
       subtitle = "D elle Leben
                                    eil
```

Mit den Optionen **height** und **width** können wir bestimmen wie stark die Punkte in eine Richtung «jittern»

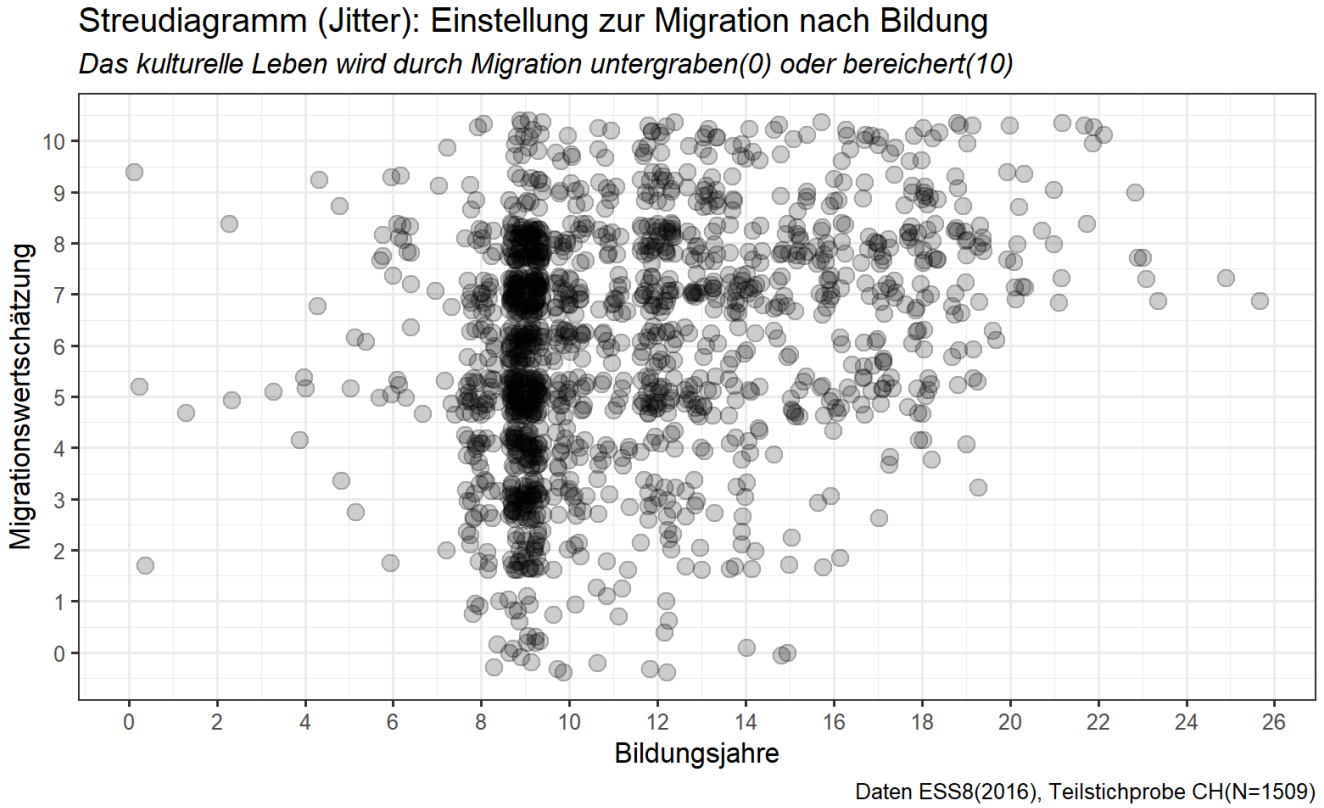


1.6 Scatterplott – Overplotting: “jitter”

geom_jitter(alpha =, size =)

```
ggplot(ess8_CH_ss, aes(x = eduyrs, y = imu
  scale_x_continuous(breaks = seq(0,26,2))
  scale_y_continuous(breaks = seq(0,10,1))
  geom_jitter(alpha = 0.2, size = 3)+
  labs(title = "Einstellung zur Migration
       x = "dungsjahre",
       y = ationswertschätzung",
       su "Das kulturelle Leben wi
```

Wir können hier auch das Alpha (Transparenz) und die Grösse der Punkte anpassen

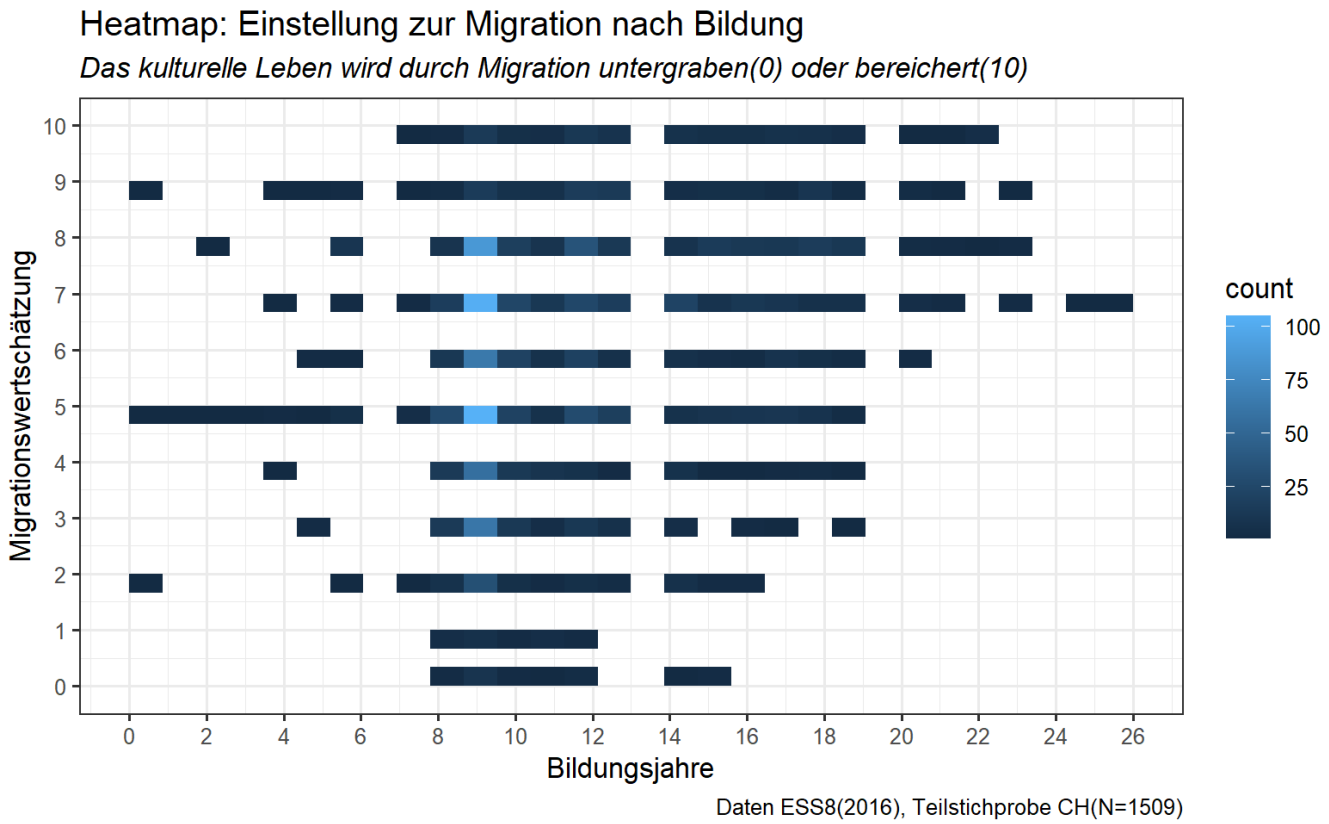


1.6 Scatterplott – Overplotting: “heatmap”

geom_bin2d()

```
ggplot(ess8_CH_ss, aes(x = eduyrs, y = imueclt))
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_bin2d() +
  labs(title = "Einstellung zur Migration nach B
       x = "Bildungsjahre",
       y = "Migrationswertschätzung",
       su = "Das kulturelle Leben wird dur
       c = "Daten ESS8(2016), Teilstichpro
```

geom_bin2d() stellt eine *Heatmap* her



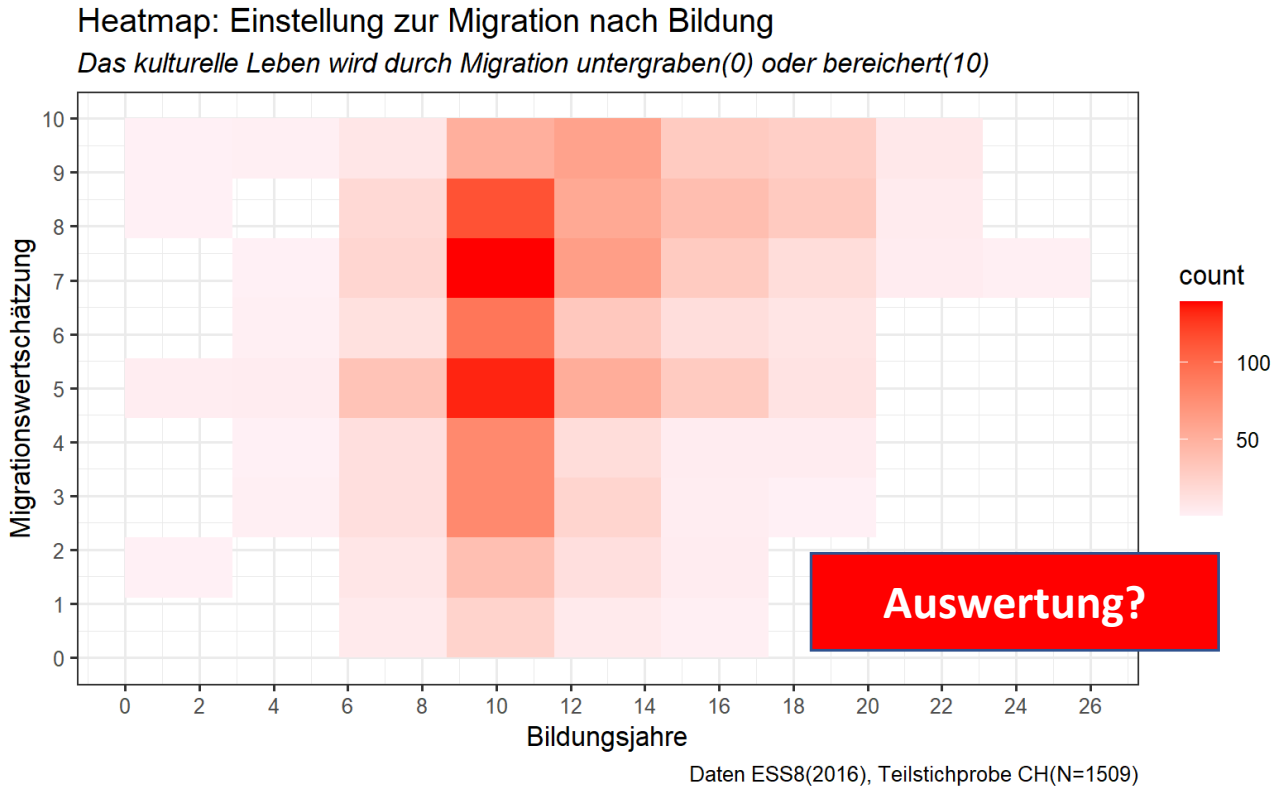
Was sind Schwächen dieser Visualisierung?

1.6 Scatterplott – Overplotting: “heatmap”

geom_bin2d(bins=)+
scale_fill_gradient(low=, high=)

```
ggplot(ess8_CH_ss, aes(x = eduyrs, y = imueclt)) +
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_bin2d(bins =9) +
  scale_fill_gradient(low = "lavenderblush",
                      high = "red") +
  labs(title = "Einstellung zur Migration nach Bildu
       x = "Bildungsjahre",
       y = "Migrationswertschätzung",
```

Über *bins* bestimmen wir die Anzahl Felder, während scale_fill_gradient() die Farbauswahl steuert.



1.6 Scatterplott – Overplotting: “heatmap”

geom_bin2d(bins=)+
scale_fill_gradient(low=, high=)

Unsere Leitlinien für Heatmaps im Streudiagramm:

(a) *Beide Dimensionen sollten geschlossen sein; die Felder also aneinander anliegen. Lücken im Plot verweisen dann eindeutig auf nicht besetzte Felder.*

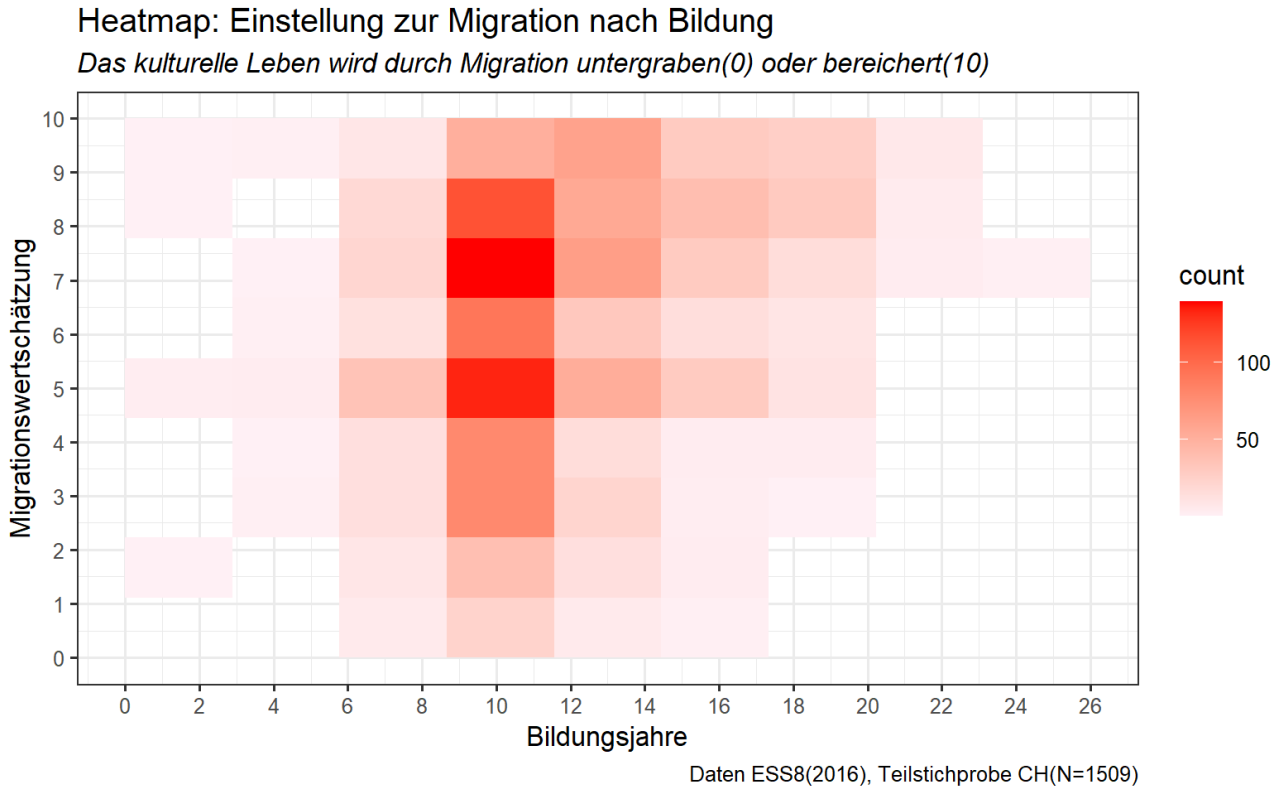
(b) *Je **dunkler** das Feld, desto **höher** die Belegung.*

(c) *Farb- und Bins-Schema werden so gewählt, dass die Charakteristika der Verteilung deutlich werden. Ggf. mit Vektoren statt Zahlen (bins) und/oder zusätzlicher «mid»-Kategorie (scale_fill_gradient) operieren.*

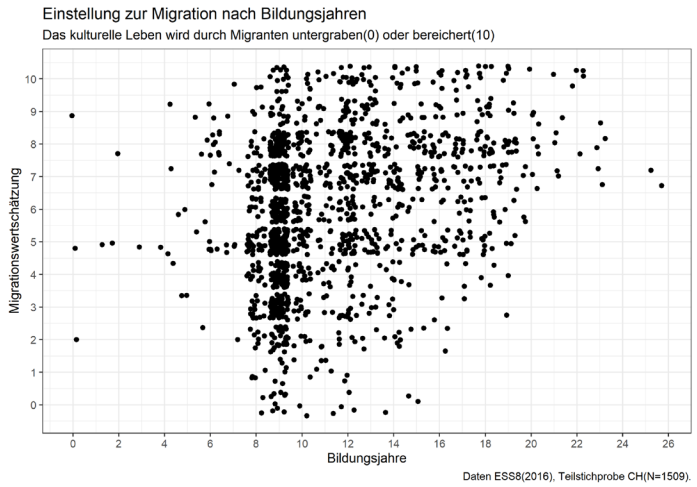
(d) Heatmap nur dann verwenden, wenn ein einfaches Streudiagramm zur Datenkommunikation ungeeignet ist.

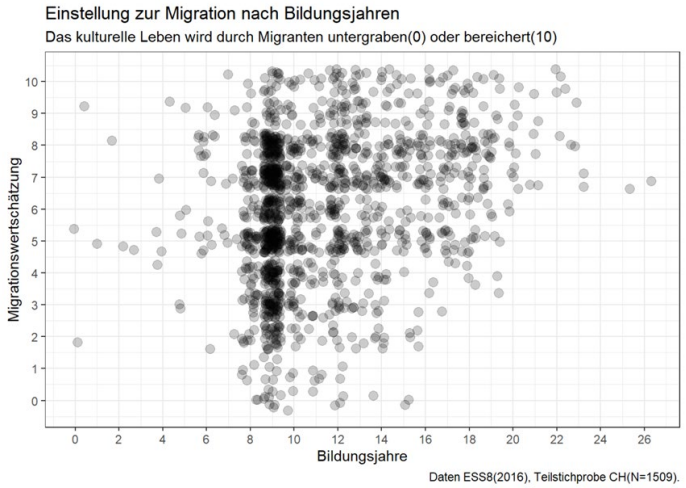
(e) *Heatmap **nur dann** verwenden, wenn der Zusammenhang damit gegenüber den Alternativen **an Anschaulichkeit gewinnt.***

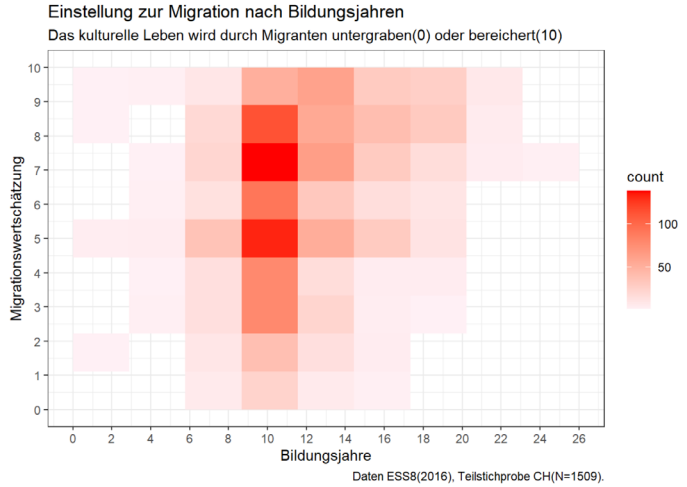
(f) *Immer wichtig (auch für einfaches Streudiagramm): Auf Beschriftung (Titel, Achsen, Note) achten.*



1.6 Welche Darstellungsform bei Overplotting?







Welche Darstellung ist die Beste? *Kriterium:* ***Aussagekraft****!*

Potential zur Veranschaulichung des Zusammenhangs im Vergleich zu anderen Optionen ohne dabei Ausgangsdaten zu verfremden.

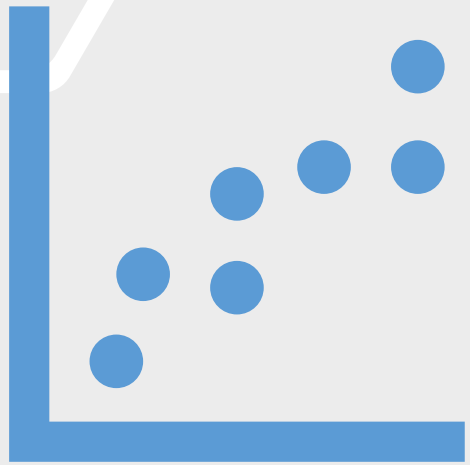
Generell gilt:
Wir wählen jeweils die Darstellungsform, auf der wir den Zusammenhang am besten erkennen und kommunizieren können.

1.6 Vor- und Nachteile: Anhaltspunkte

Achtung: Rezepthafte Entscheidungsvorgaben zur Auswahl des Plots sind nicht möglich: Die Aussagekraft hängt oft von detaillierten Spezifika der Datenlage ab. Zudem sind die Auswahlmöglichkeiten oft durch generelle Layout- oder Satzvorgaben limitiert (viele Journals akzeptieren z.B. nur S/W-Abbildungen). Es gibt aber allgemeine Leitbilder zur Datenvisualisierung, aus denen sich ein paar Empfehlungen und Erwägungen furs Streudiagramm ableiten lassen.

- **Less is more (!)** - Wenn ein **einfaches Streudiagramm** funktioniert, sollte dies auch verwendet werden.
  - Die **Transparenzoption** (alpha) ist minimal-invasiv und daher erste Alternative.
- **Jitterplots** sind insbesondere hilfreich, wenn *(a) diskrete Variablen* abgebildet werden (z.B. Likert-Skalen), die keine generische nuancierte Streuung aufweisen und *(b) mittelgrosse Stichproben* vorliegen (bei z.B. n > 10.000 hilft auch kein Jitter mehr).
  - **Jitterplots** halten das Repräsentationsniveau des einfachen Streudiagramms (*ein Punkt=eine Observation*), liefern aber keine datengetreue Abbildung – besonders problematisch ist dies in den Randbereichen und/oder wenn die jitter-Parameter zu hoch gesetzt werden.
- **Heatmaps** liefern datengetreue Abbildungen, halten aber nicht das Repräsentationsniveau des einfachen Streudiagramms (da *eine Fläche = mehrere Observationen*).
  - Das Erscheinungsbild von **Heatmaps** im Streudiagramm ist zuschnittsabhängig (bzw. abhängig davon, wie die Bins spezifiziert wurden) und somit manipulativ einsetzbar.
  - **Heatmaps** haben bei sehr schwerpunktlastigen Verteilungen nur eine begrenzte Aussagekraft.

Teil 3: Regressionsgerade und Regressionskoeffizient



2.1 Ermittlung des Regressionskoeffizienten

Was misst b1 hier?

Regressionsgleichung bzw. «Modell»: $Migrationswertschätzung = b_0 + b_1 * Bildungsjahre$

Ermittlung von b_0 und b_1 (Parameter der bestmöglichen Gerade): ***lm(AV~UV, data = DATENSATZ)***

lm = «linear Model», alternativer Ausdruck für «Regressionsanalyse»

```
lm(imueclt ~ eduyrs,
   data = ess8_CH_ss)
```

Interpretation der Konstante?

```
Call:
lm(formula = imueclt ~ eduyrs, dat        ess8_CH_ss)

Coefficients:
(Intercept)      eduyrs
     4.0520      0.1789
```

einfache technische Interpretation b1

Ad-hoc Einordnung zur Grösse des Koeffizienten bzw. des Einflusses (später mehr …)

Die Regressionsgleichung mit berechneten Koeffizienten & Konstante, der «fit»:
$Migrationswertschätzung = 4.05 + 0.18 * Bildung$

Der Basisoutput des lm-Befehls weist *Konstante und Regressionskoeffizienten* aus. Im Hintergrund legt der Befehl weitere Parameter der Regressionsanalyse (z.B. R^2, Teststatistik…) an. Um diese sichtbar zu machen, müssen wir einen kleinen Umweg gehen….

2.1 Ermittlung des Regressionskoeffizienten

(a) Regressionsergebnisse als Objekt erstellt
(b) z.B. mit **summary()**, an die Oberfläche holen.

```
fit <- lm(imueclt ~ eduyrs,
          data = ess8_CH_ss)

summary(fit)
```

```
Call:
lm(formula = imueclt ~ eduyrs, data = ess8_CH_ss)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7355 -1.5566  0.3379  1.5168  4.9480

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0520     0.1893   21.41   <2e-16 ***
eduyrs        0.1789     0.0160   11.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.172 on 1507 degrees of freedom
  (16 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.07661,   Adjusted R-squared:  0.07599
F-statistic:   125 on 1 and 1507 DF,  p-value: < 2.2e-16
```

- «Mit jedem Bildungsjahr steigt die Migrationswertschätzung um 0.18 Skalenpunkte»
- «Die vorhergesagte Migrationswertschätzung einer Person mit 0 Bildungsjahren liegt bei etwa 4 Skalenpunkten» [Unsinnig]

2.1 Ermittlung des Regressionskoeffizienten

(a) Regressionsergebnisse als Objekt erstellt
(b) z.B. mit **summary()**, an die Oberfläche holen.

```
fit <- lm(imueclt ~ eduyrs,
          data = ess8_CH_ss)

summary(fit)
```

```
Call:
lm(formula = imueclt ~ eduyrs, data = ess8_CH_ss)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7355 -1.5566  0.3379  1.5168  4.9480

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0520     0.1893   21.41   <2e-16 ***
eduyrs        0.1789     0.0160   11.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.172 on 1507 degrees of freedom
  (16 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.07661,   Adjusted R-squared:  0.07599
F-statistic:   125 on 1 and 1507 DF,  p-value: < 2.2e-16
```

- Beide Koeffizienten sind signifikant von 0 verschieden (p-Wert<0,0000000000000002)

2.1 Ermittlung des Regressionskoeffizienten

(a) Regressionsergebnisse als Objekt erstellt
(b) z.B. mit **summary()**, an die Oberfläche holen.

```
fit <- lm(imueclt ~ eduyrs,
          data = ess8_CH_ss)

summary(fit)
```

```
Call:
lm(formula = imueclt ~ eduyrs, data = ess8_CH_ss)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7355 -1.5566  0.3379  1.5168  4.9480

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0520     0.1893   21.41   <2e-16 ***
eduyrs        0.1789     0.0160   11.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.172 on 1507 degrees of freedom
  (16 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.07661,   Adjusted R-squared:  0.07599
F-statistic:   125 on 1 and 1507 DF,  p-value: < 2.2e-16
```

- «7,7% der Variation des Einstellungswertes können durch die Bildungsjahre (bzw. die Regressionsgerade) erklärt werden»
- «7,7% der Vorhersagefehler des Einstellungswertes können durch Berücksichtigung der Bildungsjahre reduziert werden»

2.1 Ermittlung des Regressionskoeffizienten

(a) Regressionsergebnisse als Objekt erstellt
(b) z.B. mit **summary()**, an die Oberfläche holen.

```
fit <- lm(imueclt ~ eduyrs,
          data = ess8_CH_ss)

summary(fit)
```

```
Call:
lm(formula = imueclt ~ eduyrs, data = ess8_CH_ss)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7355 -1.5566  0.3379  1.5168  4.9480

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0520     0.1893   21.41   <2e-16 ***
eduyrs        0.1789     0.0160   11.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.172 on 1507 degrees of freedom
  (16 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.07661,   Adjusted R-squared:  0.07599
F-statistic:   125 on 1 and 1507 DF,  p-value: < 2.2e-16
```

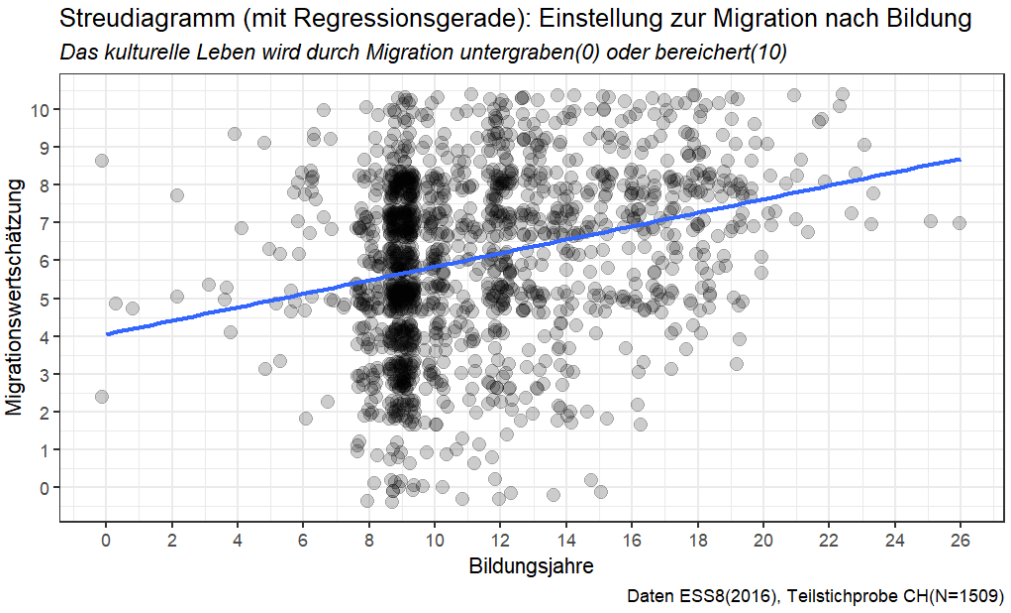
- Die grösste negative Abweichung einer realen Migrationseinstellung zum vorhergesagten Migrationseinstellung beträgt 6,7 Skalenpunkte
- Die grösste positive Abweichung einer realen Migrationseinstellung zur vorhergesagten Migrationseinstellung beträgt 4.9 Skalenpunkte

2.2 Visualisierung des Regressionskoeffizienten

Ohne dabei auf den lm-Output zuzugreifen, bietet die Funktion **geom_smooth()** die Möglichkeit, die Regressionsgerade ins **ggplot**-Streudiagramm zu integrieren.

```
ggplot(ess8_CH_ss,
       aes(x = eduyrs, y = imueclt)) +
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_jitter(alpha=0.2, size=3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Streudiagramm (mit Regressionsgerade): Einstellung zur Migration nach Bildung",
    subtitle = "Das kulturelle Leben wird durch Migration untergraben(0) oder bereichert(10)",
    x = "Bildungsjahre",
    y = "Migrationswertschätzung",
    caption = "Daten ESS8(2016), Teilstichprobe CH(N=1509)"
  ) +
  theme_bw() +
  theme(plot.subtitle = element_text(face = "italic"))
```

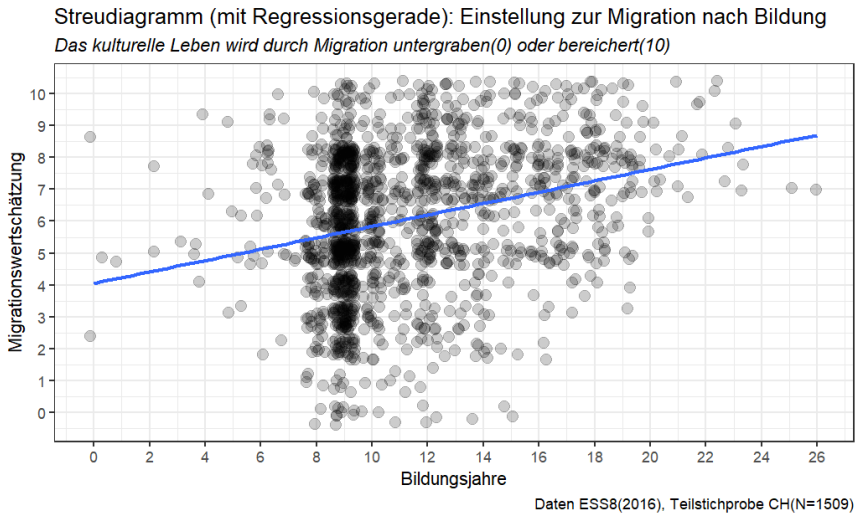
2.2 Visualisierung des Regressionskoeffizienten

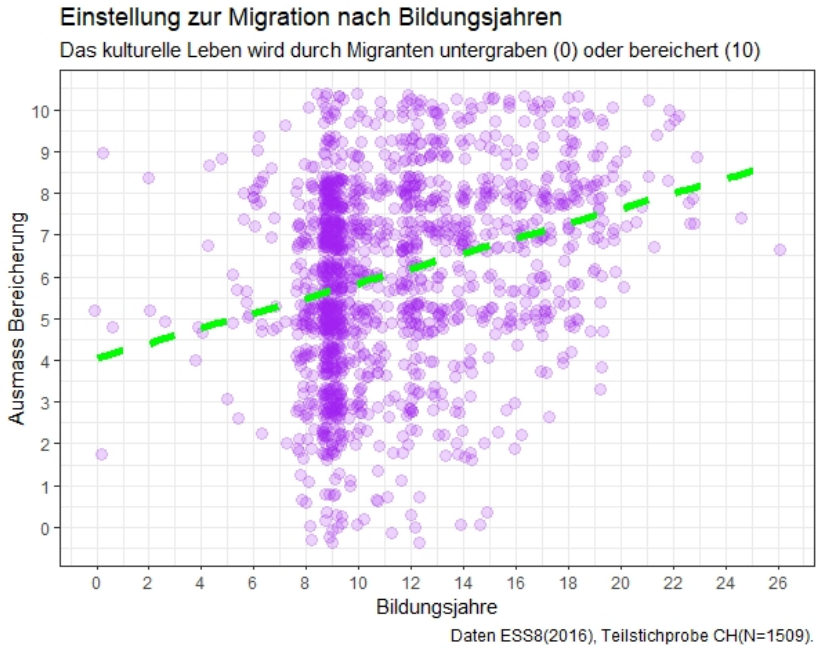


Aufgabe: Modifiziert den Code so, dass die Gerade grün und gestrichelt sowie die Punkte in pink erscheinen

```
ggplot(ess8_CH_ss,
       aes(x = eduyrs, y = imueclt)) +
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_jitter(alpha=0.2, size=3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Streudiagramm (mit Regressionsgerade): Einstellung zur Migration nach Bildung",
    subtitle = "Das kulturelle Leben wird durch Migration untergraben(0) oder bereichert(10)",
    x = "Bildungsjahre",
    y = "Migrationswertschätzung",
    caption = "Daten ESS8(2016), Teilstichprobe CH(N=1509)"
  ) +
  theme_bw() +
  theme(plot.subtitle = element_text(face = "italic"))
```

2.2 Visualisierung des Regressionskoeffizienten





Aufgabe: Modifiziert den Code so, dass die Gerade grün und gestrichelt sowie die Punkte in pink erscheinen

MA

rewrite the code so that the slope appears green, thick and dashed, and the points shall be purple

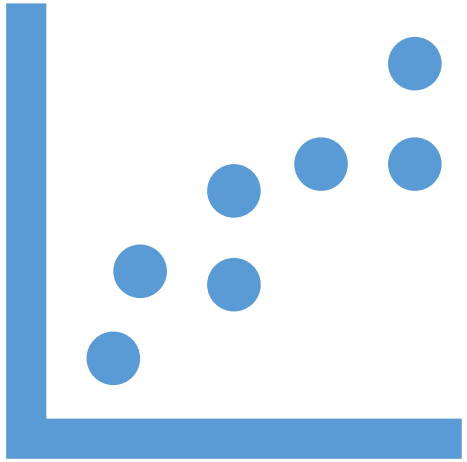
Here's the rewritten code with the desired changes:

perl — Copy code

```
ggplot(ess8_CH_ss, aes(x = eduyrs, y = imueclt)) +
  scale_x_continuous(breaks = seq(0,26,2)) +
  scale_y_continuous(breaks = seq(0,10,1)) +
  geom_jitter(alpha=0.2, size=3, color = "purple") + # Change point color to purple
  geom_smooth(method = "lm", se = FALSE, color = "green", size = 2, linetype = "das
  labs(title = "Einstellung zur Migration nach Bildungsjahren",
       x = "Bildungsjahre",
       y = "Ausmass Bereicherung",
       subtitle = "Das kulturelle Leben wird durch Migranten untergraben (0) oder b
       caption = "Daten ESS8(2016), Teilstichprobe CH(N=1509).") +
  theme_bw()
```

I've added the `color` argument to the `geom_jitter` function to change the color of the points to purple. For the `geom_smooth` function, I've added `color = "green"` to change the color of the line to green, `size = 2` to increase the thickness of the line, and `linetype = "dashed"` to change the line type to dashed.

Teil 4: Fortgeschrittene Interpretation



2.3 Fortgeschrittene Interpretation

Erinnert euch an die Interpretationsstrategien aus der Vorlesung 2.3 (ca. ab Folie 25):

1. **Einfache technische Interpretation (Basisinterpretation)**
2. **Interpretation mit Reskalierung**
3. **Interpretation mit Standardisierung**

```
Call:
lm(formula = imueclt ~ eduyrs, data = ess8_CH_ss)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7355 -1.5566  0.3379  1.5168  4.9480

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0520     0.1893   21.41   <2e-16 ***
eduyrs        0.1789     0.0160   11.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.172 on 1507 degrees of freedom
  (16 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.07661,   Adjusted R-squared:  0.07599
F-statistic:   125 on 1 and 1507 DF,  p-value: < 2.2e-16
```

```
> sd(ess8_CH_ss$eduyrs, na.rm = TRUE)
[1] 3.496909
> sd(ess8_CH_ss$imueclt, na.rm = TRUE)
[1] 2.258924
```

Aufgabe: Wendet verschiedene Interpretationsstrategien auf diesen empirischen Fall an. Welchen helfen beim Verständnis der Einflussgrösse, welche nicht? Woran liegt das jeweils?

2.3 Fortgeschrittene Interpretation

```
Call:
lm(formula = imueclt ~ eduyrs, data = ess8_CH_ss)

Coefficients:
(Intercept)       eduyrs
     4.0520       0.1789
```

Einfache technische Interpretation:

Mit jedem zusätzlichen Bildungsjahr erhöht sich die Migrationswertschätzung im Mittel um 0.18 Skalenpunkte.

Reskalierung / Vorhersage, z.B.

SchweizerInnen mit 15 Bildungsjahren (BA) wird eine um etwa einen halben Skalenpunkte positivere Einstellung zur Migration vorhergesagt als SchweizerInnen mit 12 Bildungsjahren (Matura).

Variante: Für SchweizerInnen mit 12 Bildungsjahren (Matura) wird eine Migrationswertschätzung von etwa 6.2 Skalenpunkten, für solche mit 15 Bildungsjahren (BA) eine von etwa 6.7 vorhergesagt.

Reskalierung / Vorhersage, z.B.

Die vorhergesagte Migrationswertschätzung wird mit 6 zusätzlichen Bildungsjahren um einen ganzen Skalenpunkt positiver.

2.3 Interpretation – Vorhersagen einfach machen mit R

Welche Migrationswertschätzung wird für eine Person mit 12 Bildungsjahren vorhergesagt?

```
Call:
lm(formula = imueclt ~ eduyrs, data = ess8_CH_ss)

Coefficients:
(Intercept)       eduyrs
     4.0520       0.1789
```

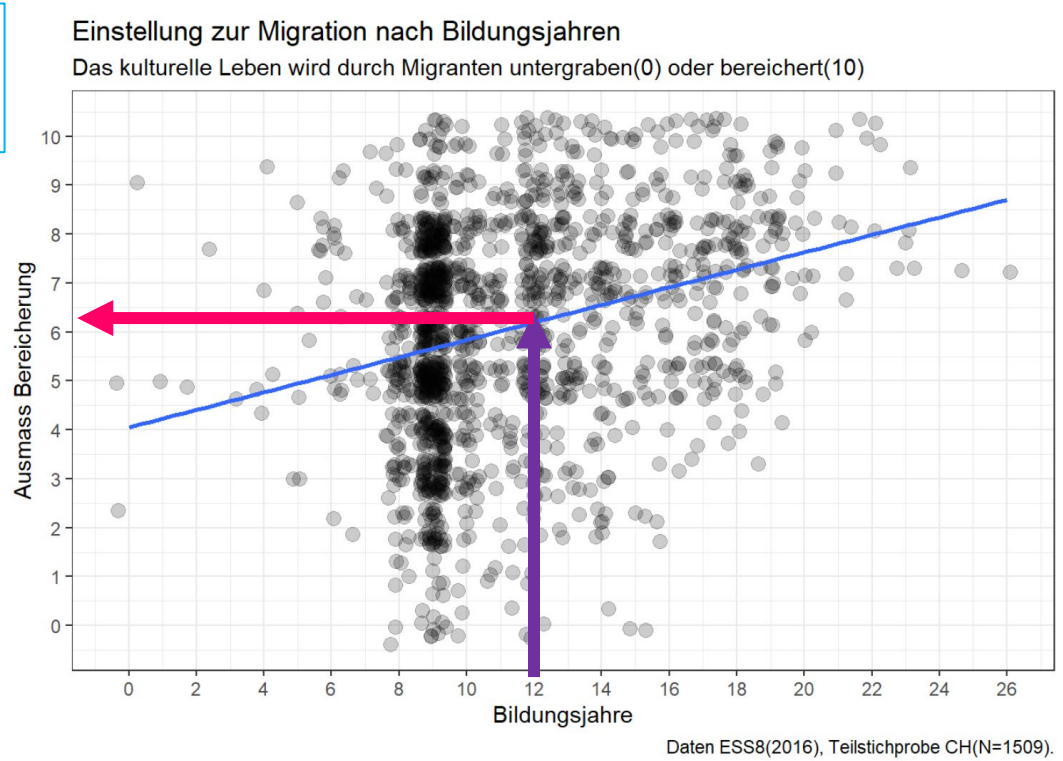
$$\hat{y} = 4,05 + 0,18 * x$$

$$\hat{y} = 4,05 + 0,18 * 12$$

$$6,21 = 4,05 + 0,18 * 12$$

ggpredict() übernimmt die Arbeit für uns!

```
library(ggeffects)
ggpredict(fit, terms = "eduyrs[9, 12, 15, 17]")
```




```
# Predicted values of Country's cultural life
undermined or enriched by immigrants

eduyrs | Predicted |       95% CI
-------------------------------------
     9 |      5.66 | [5.53, 5.79]
    12 |      6.20 | [6.09, 6.31]
    15 |      6.74 | [6.58, 6.90]
    17 |      7.09 | [6.88, 7.30]
```

2.3 Fortgeschrittene Interpretati

```
Call:
lm(formula = imueclt ~ eduyrs, data = ess8_CH_ss)

Coefficients:
(Intercept)       eduyrs
     4.0520       0.1789
```

```
> sd(ess8_CH_ss$eduyrs, na.rm = TRUE)
[1] 3.496909
> sd(ess8_CH_ss$imueclt, na.rm = TRUE)
[1] 2.258924
```

Standardisierung der Bezugseinheit (UV)

*(0.18 * 3.50 = 0.62)* Ein typischer Bildungsunterschied erzeugt im Mittel eine Veränderung der Einstellung zur Migration um 0.62 Skalenpunkte.

Standardisierung der Zieleinheit (AV)

(0.18/2.25= 0.08) Mit jedem zusätzlichen Bildungsjahr wird die Migrationseinstellungen im Schnitt um 0.08 Standardeinheiten (also etwas weniger als 1/10 eines typischen Einstellungsunterschiedes) positiver.

Standardisierung von Bezugs- und Zieleinheit (UV & AV)

*(0.18*3.50/2.25 = 0.14)* Ein typischer Bildungsunterschied erzeugt im Mittel eine Veränderung der Einstellung zur Migration um etwa 0.14 Standardabweichungen zu. *(Interpretation eher nicht sinnvoll, aber Verwendung von beta (bzw. r=beta) kann zu Vergleichszwecken praktisch sein)*

2.3 Fortgeschrittene Interpretati

```
Call:
lm(formula = imueclt ~ eduyrs, data = ess8_CH_ss)

Coefficients:
(Intercept)       eduyrs
     4.0520       0.1789
```

```
> sd(ess8_CH_ss$eduyrs, na.rm = TRUE)
[1] 3.496909
> sd(ess8_CH_ss$imueclt, na.rm = TRUE)
[1] 2.258924
```

Standardisierung der Bezugseinheit (UV)

*(0.18 * 3.50 = 0.62)* Ein typischer Bildungsunterschied erzeugt im Mittel eine Veränderung der Einstellung zur Migration um 0.62 Skalenpunkte.

Standardisierung der Zieleinheit (AV)

(0.18/2.25=
Standardein

Ordnet den Zusammenhang zwischen Bildung und Migrationseinstellungen vergleichend mit anderen, im Rahmen der Veranstaltung analysierten, Zusammenhängen ein.

Standardisierung von Bezugs- u eit (UV & AV)

*(0.18*3.50/2.25 = 0.14)* Ein typischer Bildungsunterschied erzeugt im Mittel eine Veränderung der Einstellung zur Migration um etwa 0.14 Standardabweichungen zu. *(Interpretation eher nicht sinnvoll, aber Verwendung von beta (bzw. r=beta) kann zu Vergleichszwecken praktisch sein)*

Übung

Untersucht nun selbständig den Zusammenhang zwischen Lebenszufriedenheit und Positionierung auf der Links-Rechts-Skala für die GG aller SchweizerInnen (auf Basis des ESS!)

1. Findet passende Variablen
2. Inspiziert sie und kodiert um falls nötig
3. Visualisiert den Zusammenhang (probiert verschiedene Varianten aus)
4. Berechnet den Regressionskoeffizienten
5. Trefft Vorhersagen für ausgewählte Werte der UV

Aufgabe für nächste Woche

- Löst die Übung zur Sitzung 3 auf der Tutoratswebseite.