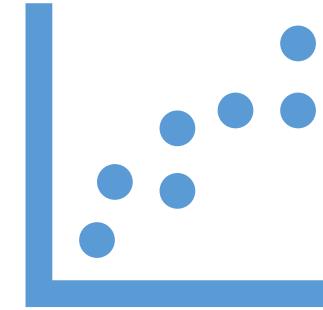


Statistik 2 – Tutorate

Sitzung 3: Regressionsanalyse Basics

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



0

Theoretische Vorüberlegungen

Fragestellung: Wie beeinflusst Bildung die Einstellungen zur Migration?



Formuliert und begründet eine prüfbare Hypothese

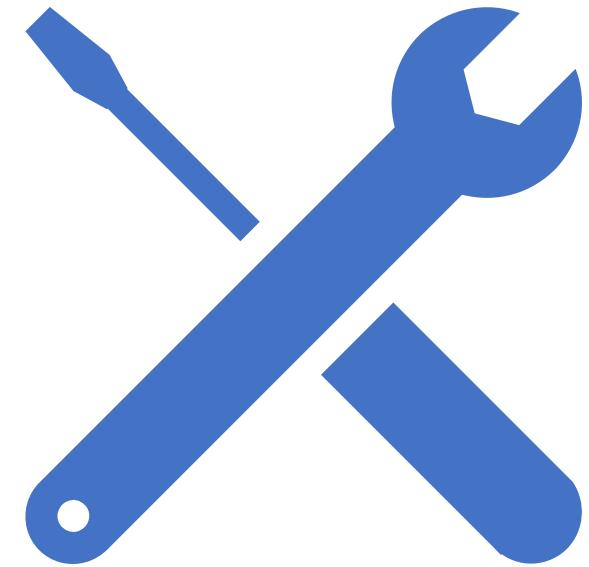
Je höher die Bildung desto positiver sind die Einstellungen zur Migration



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



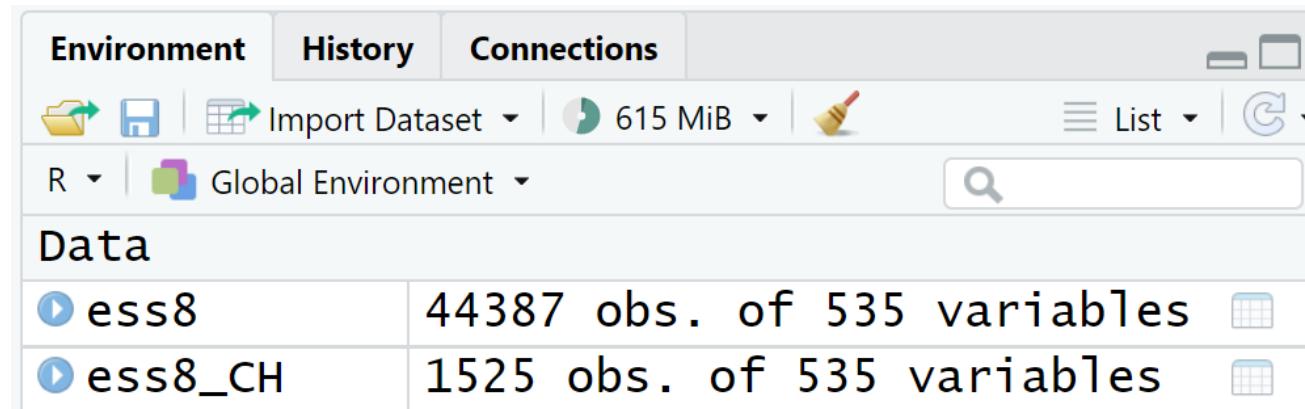
Teil 1: Datenmanagement



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	

The screenshot shows the RStudio interface. In the top-left, there's a code editor window with the following R code:

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

In the top-right, there's a viewer pane containing a table of datasets:

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

Below the viewer pane is a data grid. The top row of the grid has the following column headers: pos, variable, label, col_type, missing, levels, and value. The data rows below show various survey questions and their characteristics. The 'label' column contains descriptive text for each question, and the 'variable' column lists the corresponding R variable names. The 'value' column shows the possible values for each variable.

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:

- **edulvib**: Highest level of education
- **eisced**: Highest level of education, ES – ISCED
- **edlvdch**: Highest level of education, Switzerland
- **eduys**: Years of full-time education completed

EINSTELLUNGEN ZUR MIGRATION:

- **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 beiden Konzepte werden jeweils durch verschiedene Variablen abgebildet. In der Praxis hingt 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 **eduys** und **imueclt**.

1.2

Selektieren von Merkmalen

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

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

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

eduysr		imueclt
	Years of full-time education completed	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

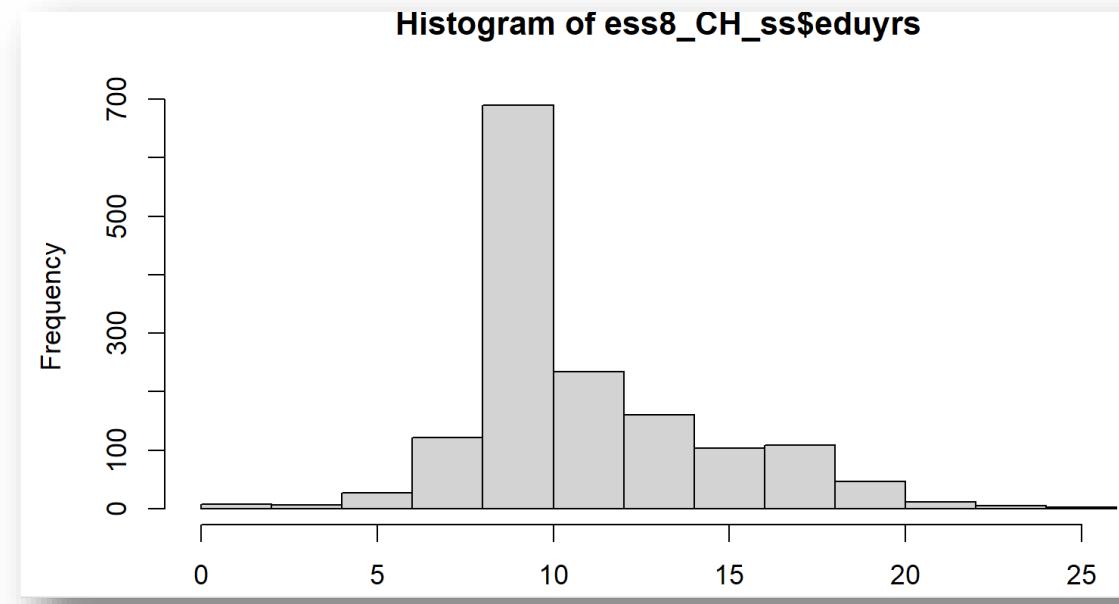
```
attributes(ess8_CH_ss$eduys)
summary(ess8_CH_ss$eduys)
hist(ess8_CH_ss$eduys)
sd(ess8_CH_ss$eduys, na.rm = TRUE)
```

```
$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

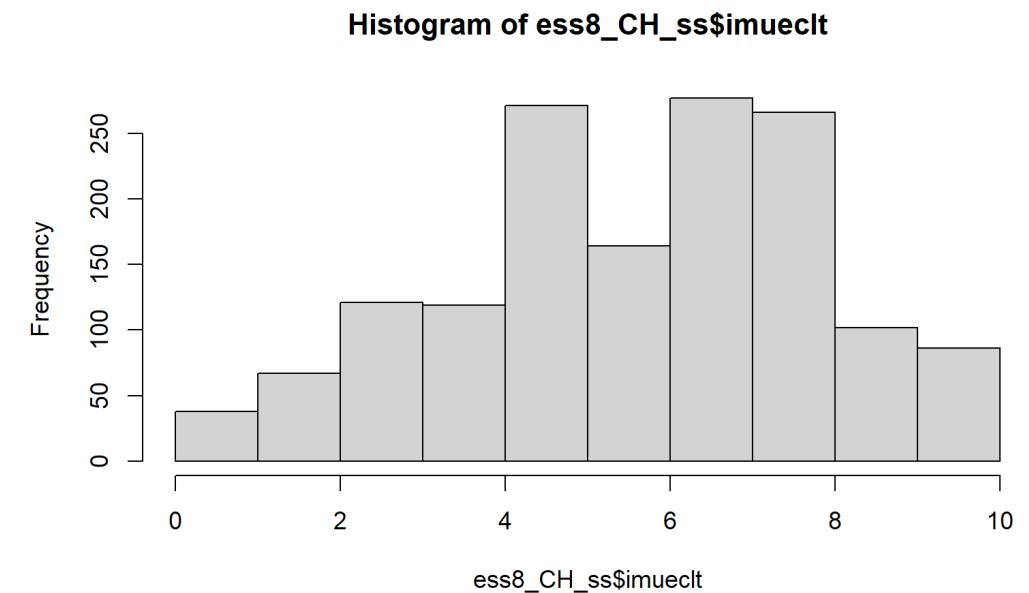
```
attributes(ess8_CH_ss$imueclt)
summary(ess8_CH_ss$imueclt)
hist(ess8_CH_ss$imueclt)
sd(ess8_CH_ss$imueclt, na.rm = TRUE)

$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

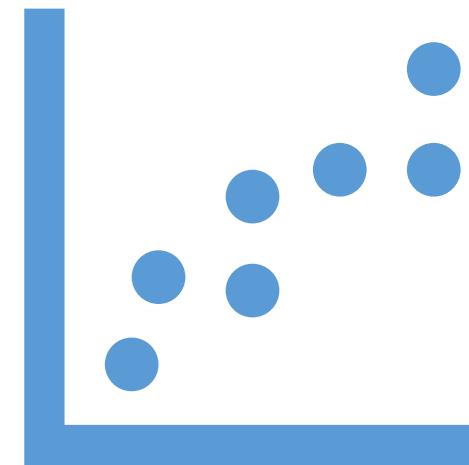
Festlegung des Analyseformates

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

Übersicht: Zusammenhangsmasse (bivariate Statistik)

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

Teil 2: Visualisierung – Streudiagramm per **ggplot**

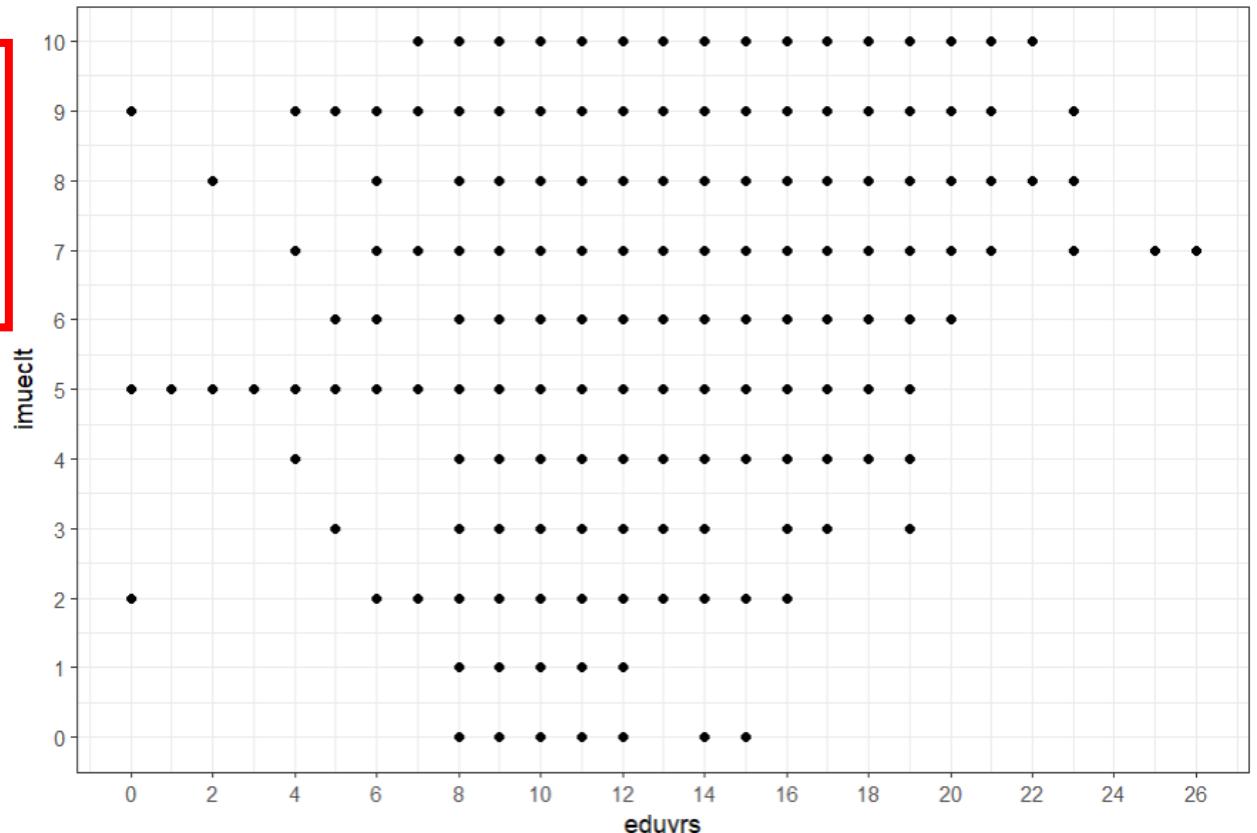


1.6 Visualisierung des Zusammenhangs: Scatterplot mit ggplot()

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?

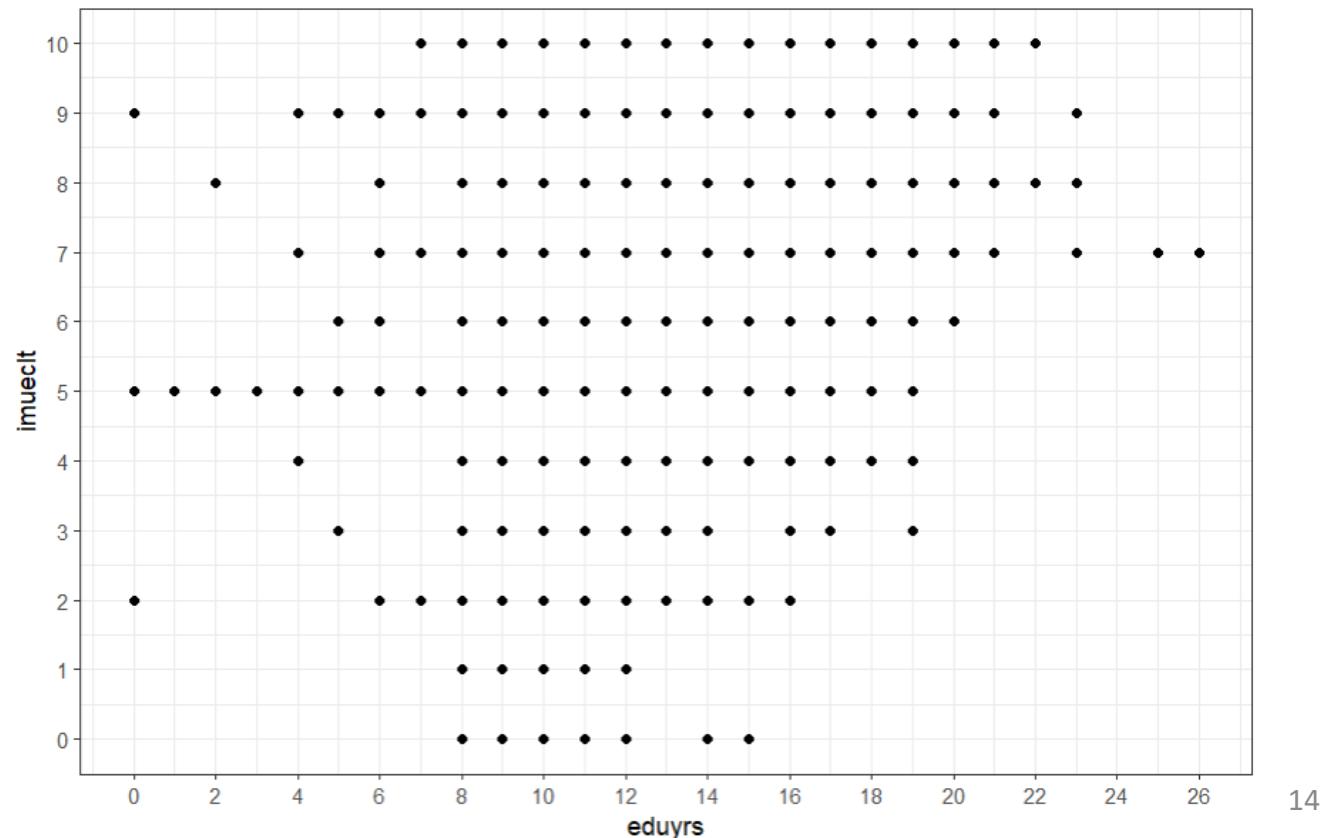
```
ggplot(ess8_CH_ss,  
       aes(x = eduys, y = imueclt)) +  
  scale_x_continuous(breaks = seq(0, 26, 2)) +  
  scale_y_continuous(breaks = seq(0, 10, 1)) +  
  geom_point() +  
  theme_bw()
```



1.6 Visualisierung des Zusammenhangs: Scatterplot

Verwende die Datenmatrix «ess8_CH_ss»

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



1.6

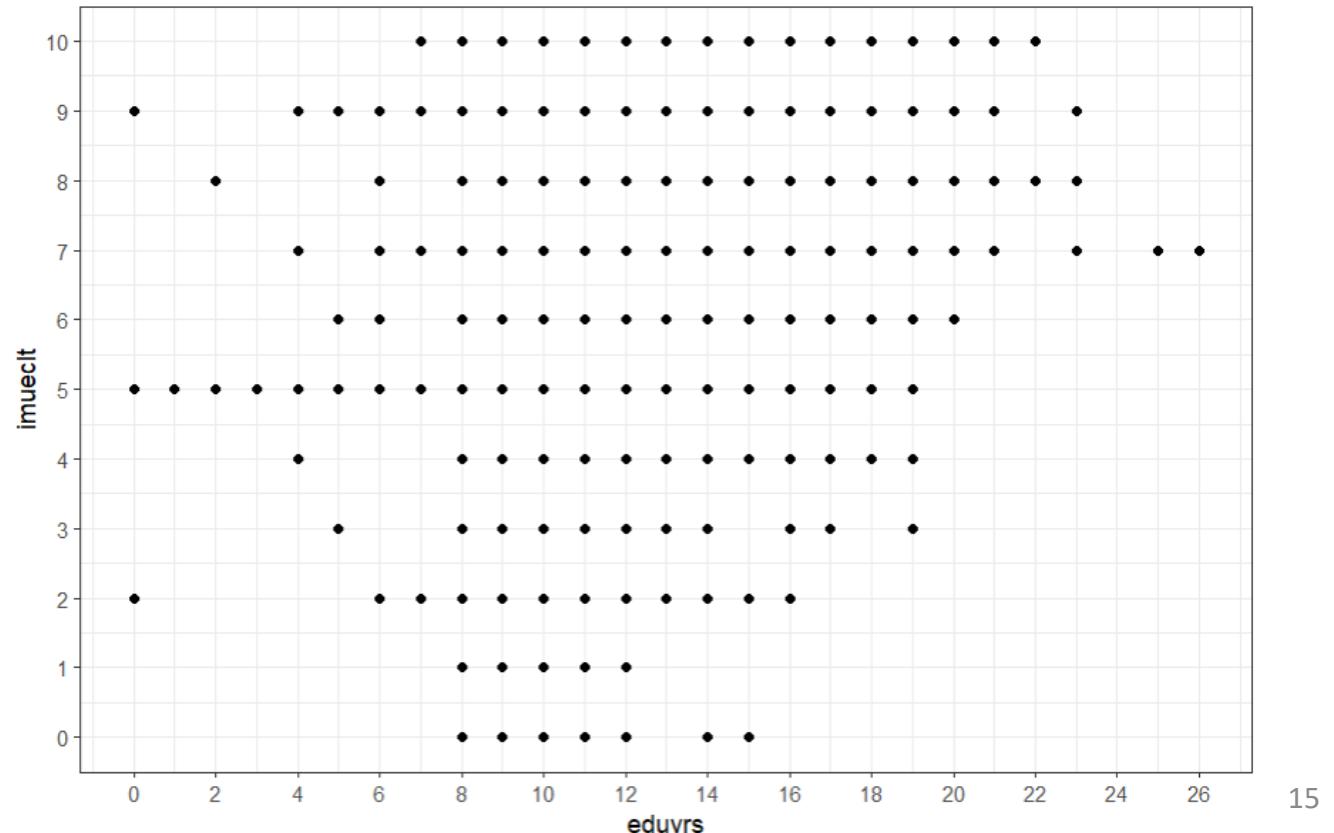
Visualisierung des Zusammenhangs: Scatterplot

Unterfunktion «aes»:

Wie soll die Abbildung aufgebaut sein?

```
ggplot(b8_CH_ss,  
       aes(x = eduhrs, y = imueclt)) +  
  scale_x_continuous(breaks = seq(0, 26, 2)) +  
  scale_y_continuous(breaks = seq(0, 10, 1)) +  
  geom_point() +  
  theme_bw()
```

... «eduhrs» definiert x-Achse, «imueclt»
definiert die y-Achse

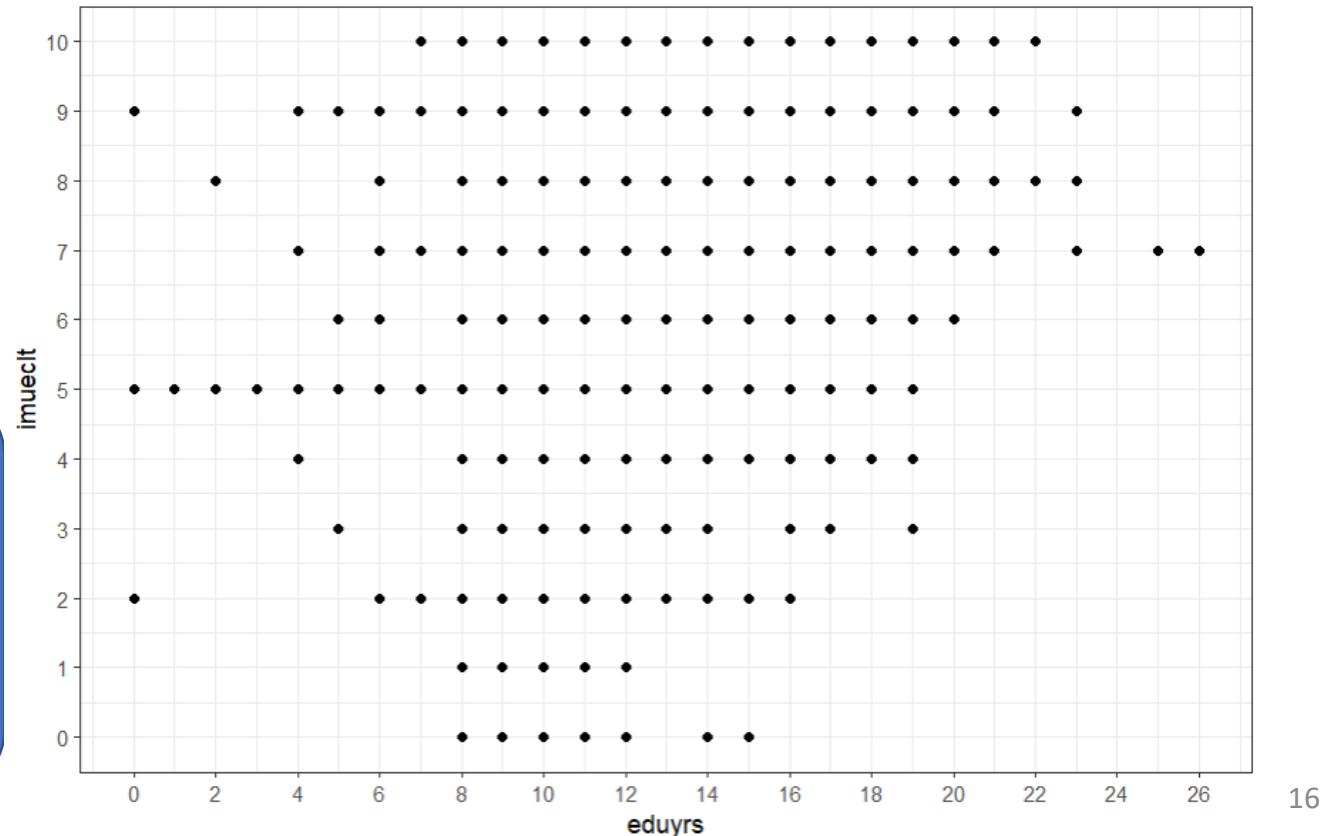


1.6 Visualisierung des Zusammenhangs: Scatterplot

```
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)

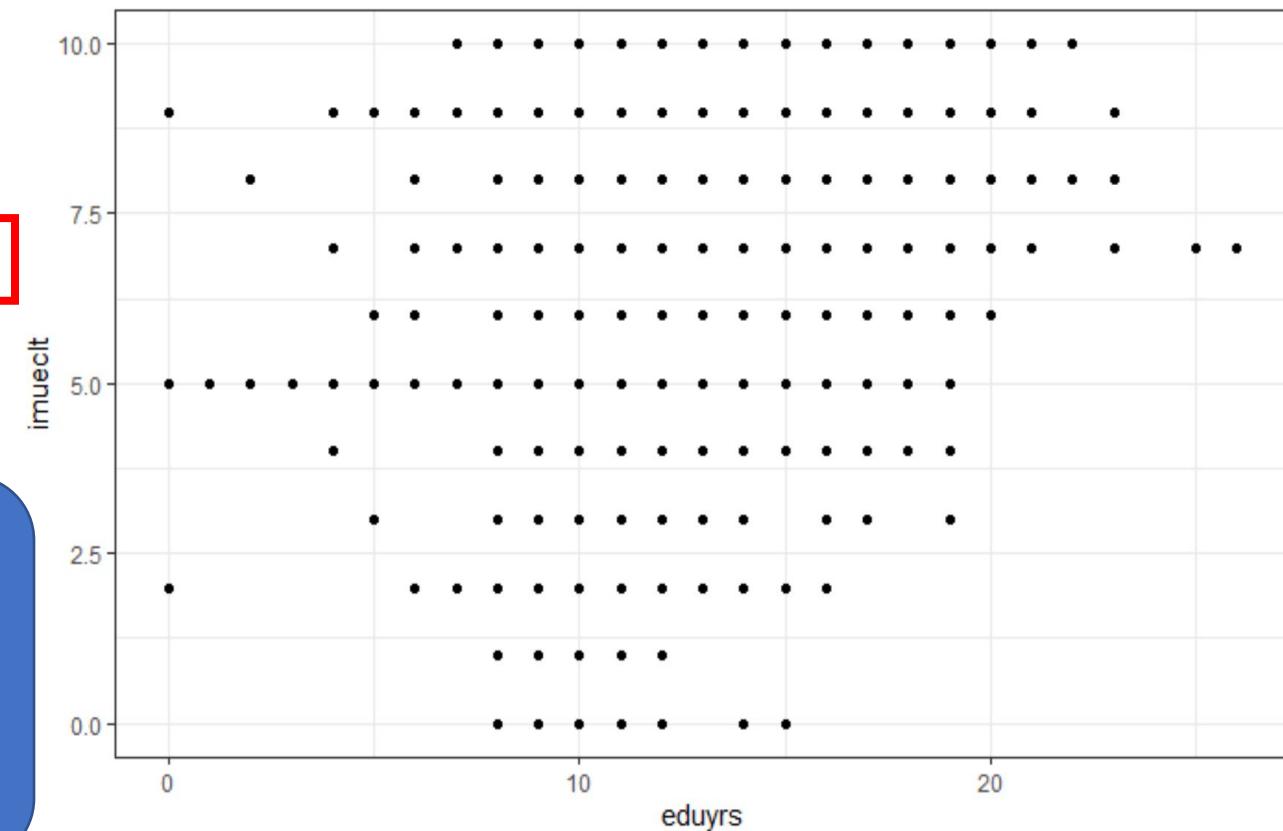


1.6

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 = eduys, y = imueclt)) +  
  [redacted]  
  geom_point() +  
  theme_bw()
```



Probleme bei Default:

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

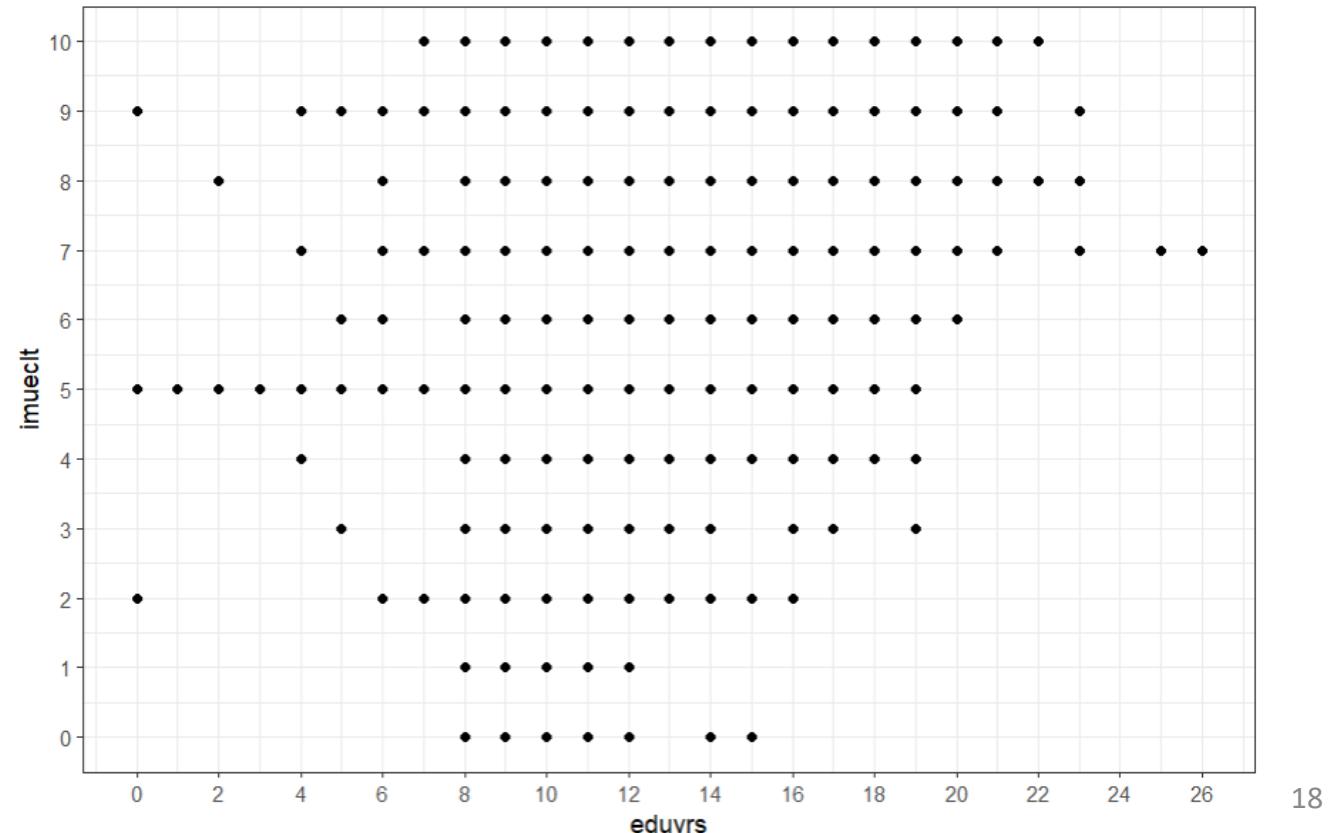
1.6

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 = eduys, 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)



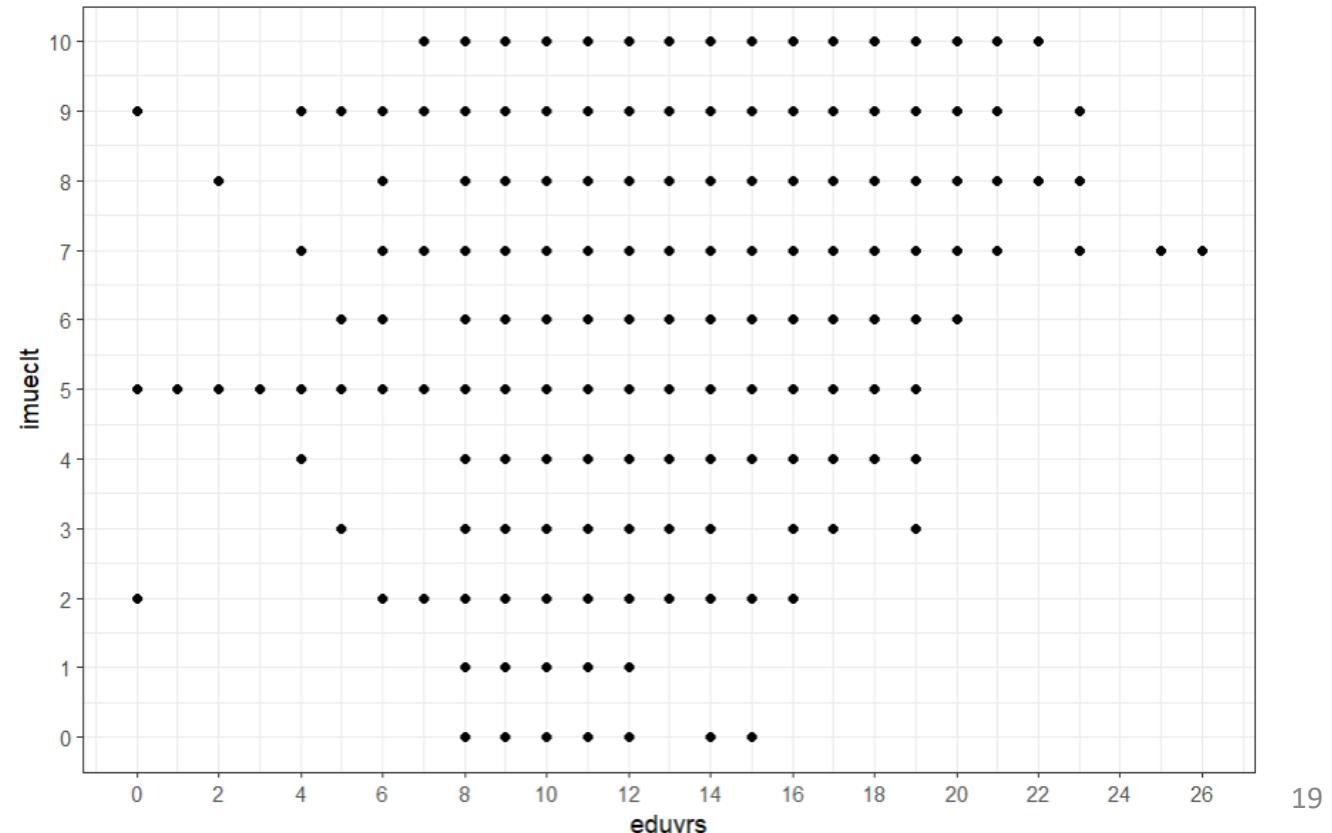
1.6

Visualisierung des Zusammenhangs: Scatterplot

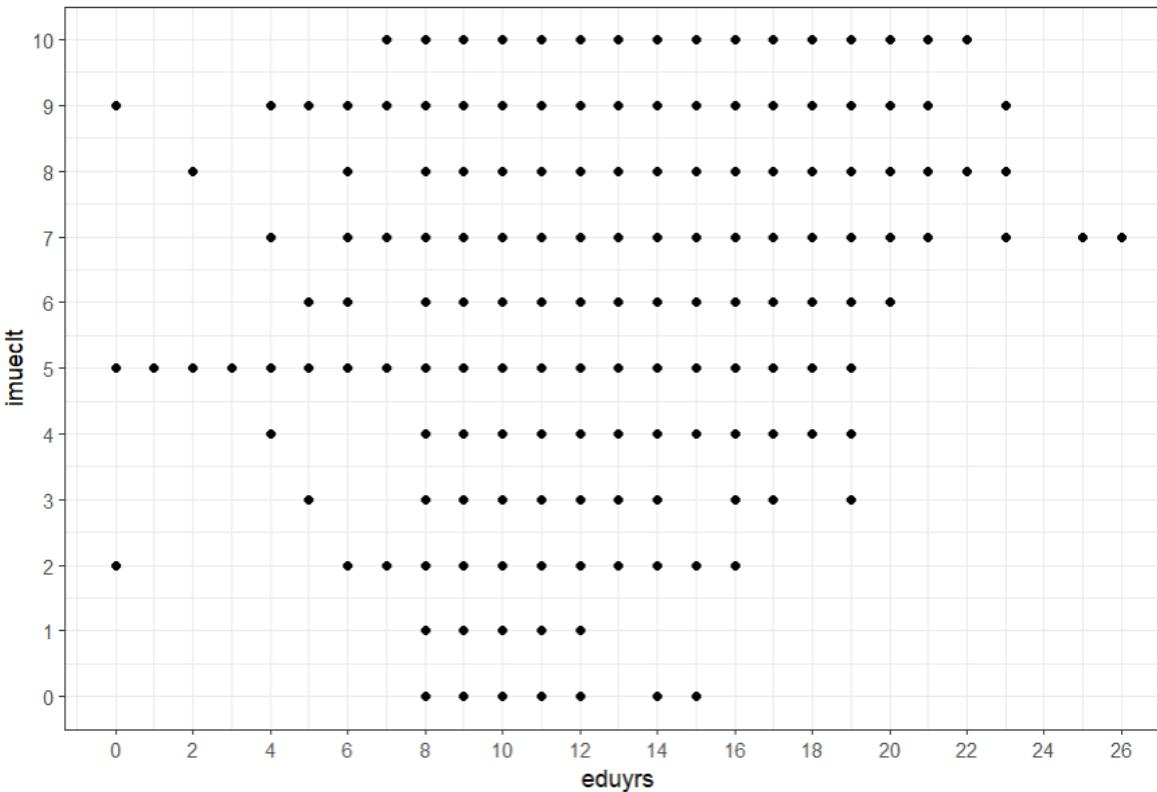
Grundlage der Regressionsanalyse ist die Visualisierung des Zusammenhangs auf Grundlage des **ggplot()** Befehls aus dem **ggplot2** Package.

```
ggplot(ess8_CH_ss,  
       aes(x = eduys, 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 Muster erkennen?

Lässt sich ein Zusammenhang erkennen?

Warum ist die Grafik nicht aussagekräftig?

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

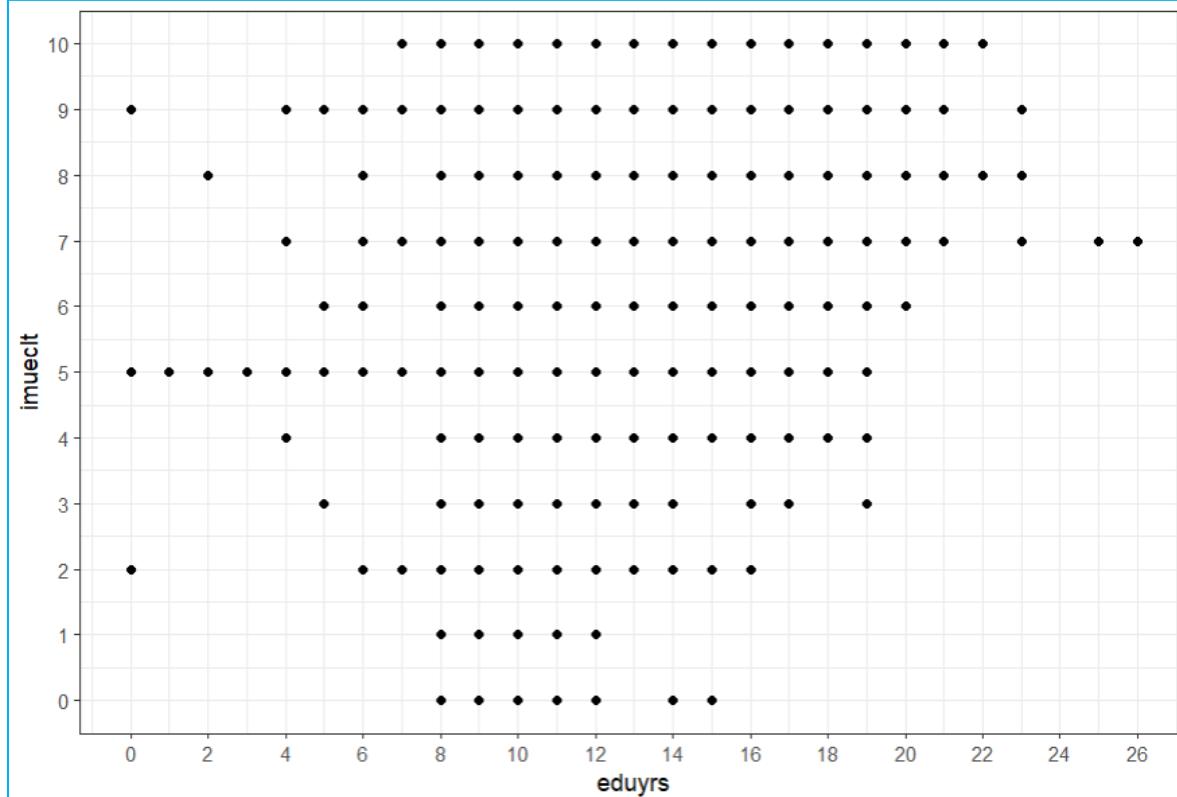
1.6 Scatterplot - Beschriftung

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 «eduys»)
- **Caption** mit Datenquelle und (Teil-) Stichprobengrösse

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



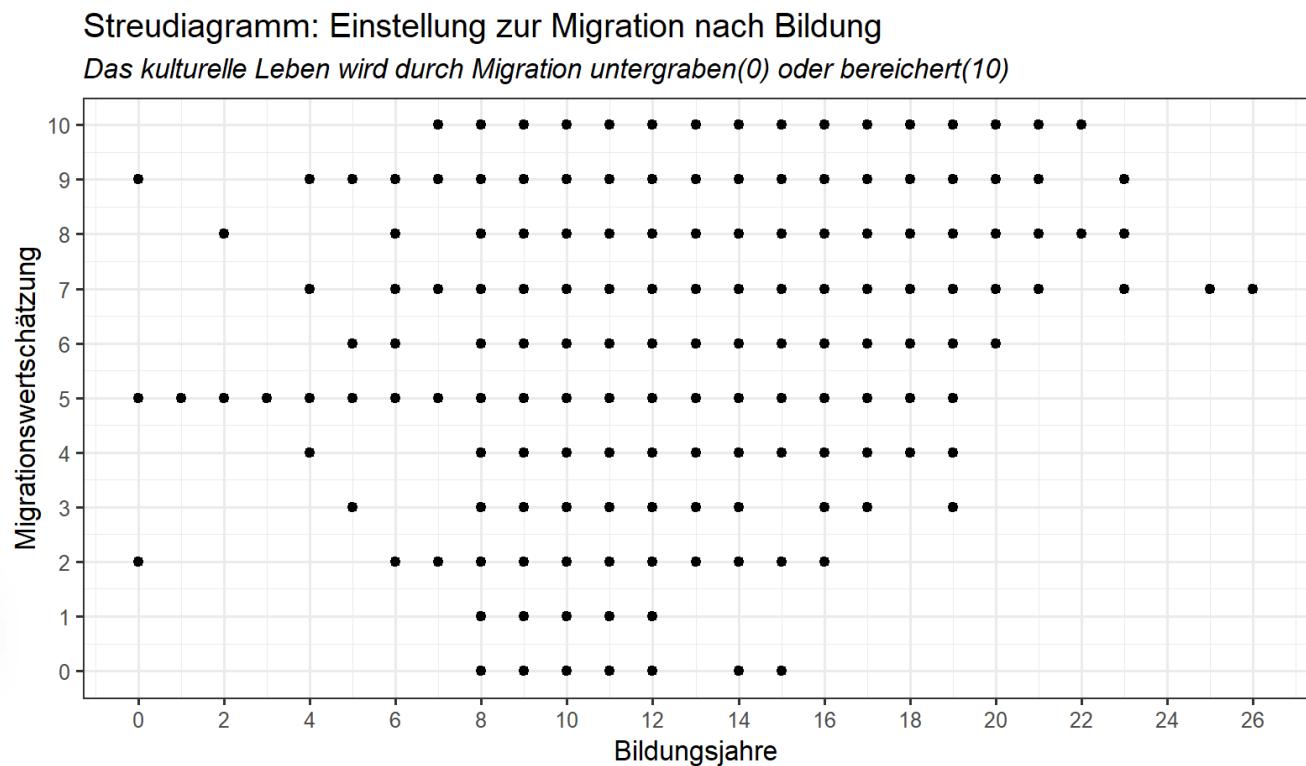
1.6 Scatterplot - Beschriftung

```
ggplot(ess8_CH_ss, aes(x = eduys, 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"))
```

Data	
ess8	44387 obs.
ess8_CH	1525 obs.
ess8_CH_ss	1525 obs.

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

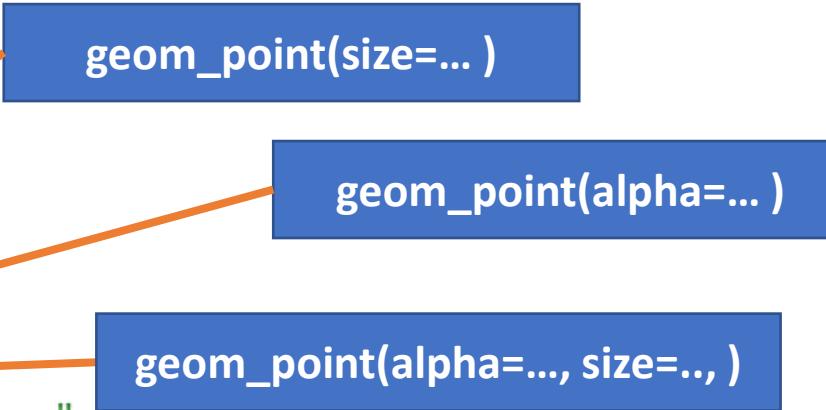
Info zu Fallzahl im Plot aus ggplot-Meldung ableiten



1.6 Scatterplot - Overplotting

Lösungsmöglichkeit 1: «Alpha»

```
ggplot(ess8_CH_ss,  
       aes(x = eduhrs, 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()
```



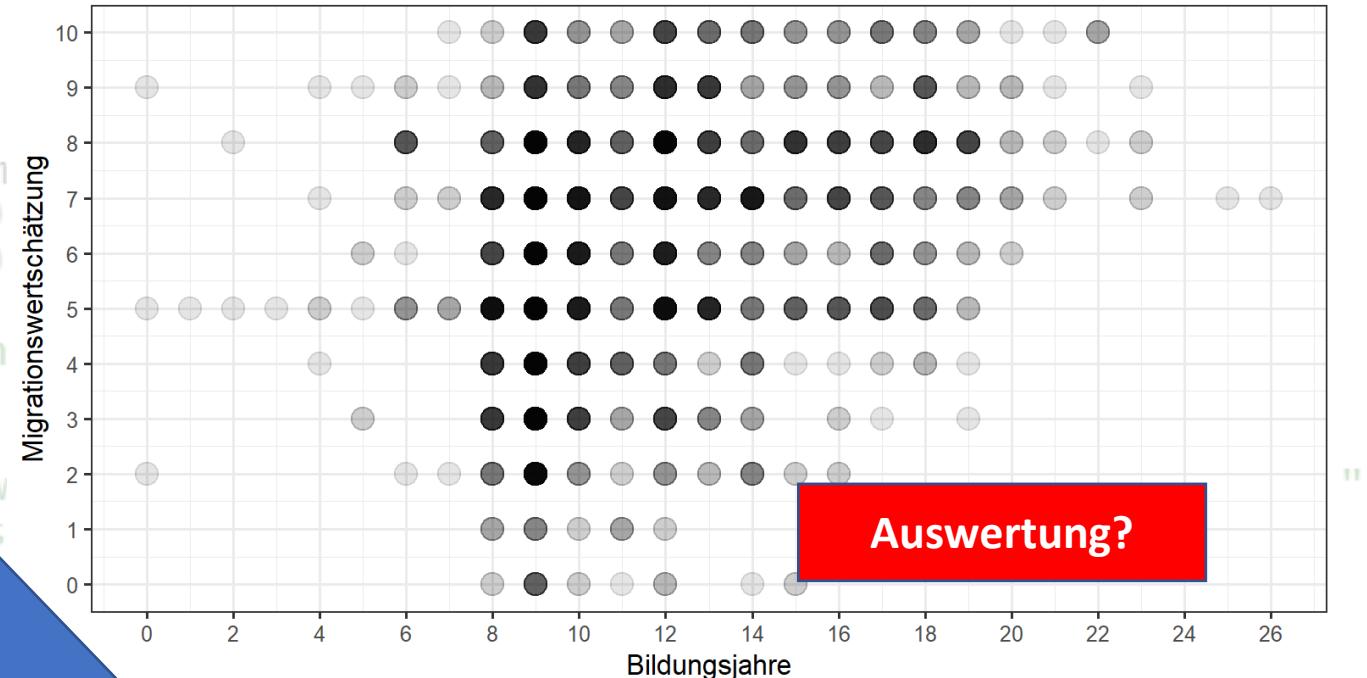
Ersetzt den `geom_point()` Befehl durch die hier vorgeschlagenen Alternativen.
Wie verändert sich der Plot? Was bewirken die einzelnen Teilarumente?

1.6 Scatterplot – Overplotting: “alpha”

```
ggplot(ess8_CH_ss, aes(x = eduyears, 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 wird durch Migration untergraben(0) oder bereichert(10)  
  caption = "Daten ESS8(2016), Teilstichprobe CH(N=1509)"  
  )
```

“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.

Streudiagramm: Einstellung zur Migration nach Bildung
Das kulturelle Leben wird durch Migration untergraben(0) oder bereichert(10)



“Size” bestimmt die Grösse der Datenpunkte.

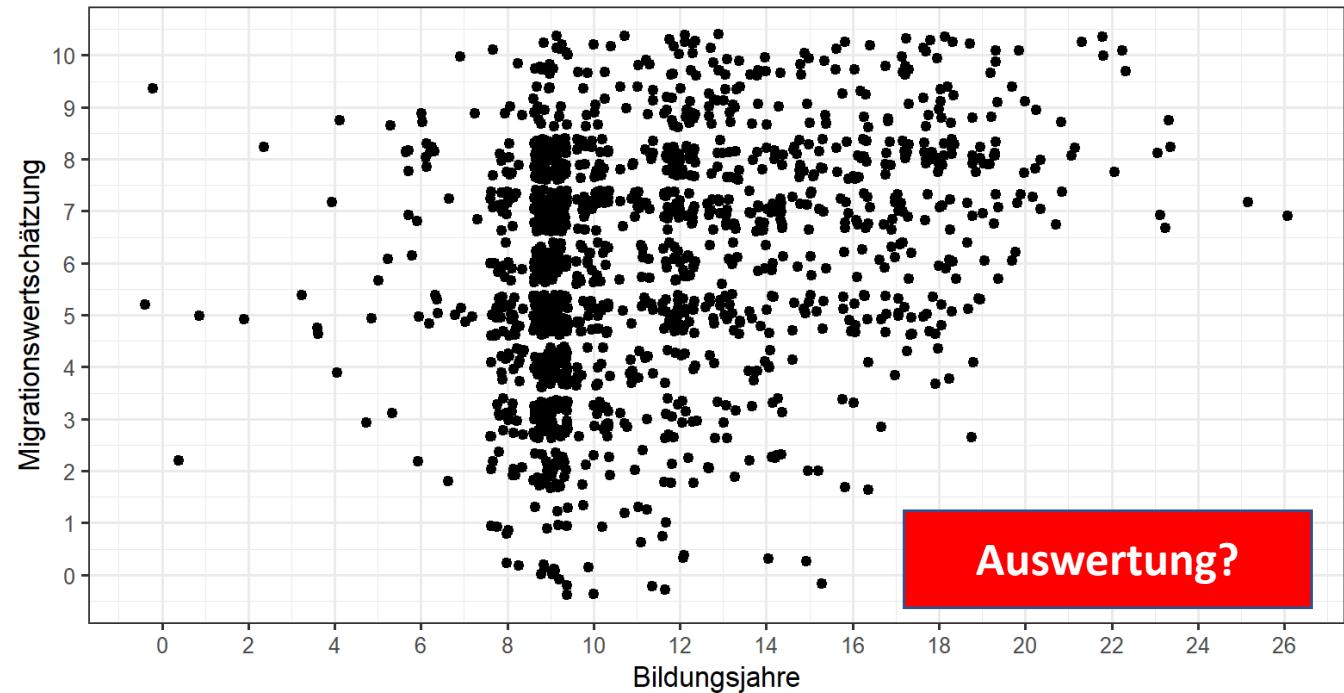
1.6 Scatterplot – Overplotting: “jitter”

geom_jitter() – Beschreibt die Funktion

```
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  
Data: ESS8(2016), Teilstic
```

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

Streudiagramm (Jitter): Einstellung zur Migration nach Bildung
Das kulturelle Leben wird durch Migration untergraben(0) oder bereichert(10)



Daten ESS8(2016), Teilstichprobe CH(N=1509)

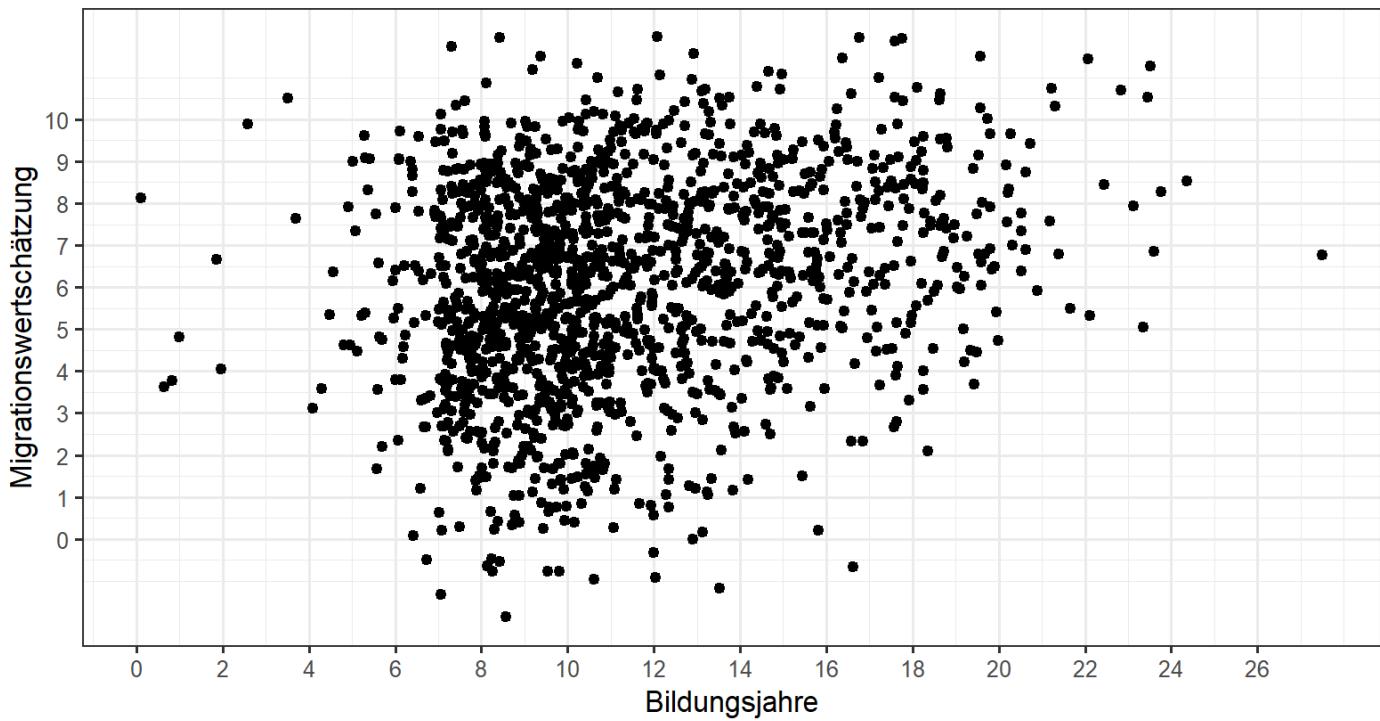
1.6 Scatterplot – Overplotting: “jitter”

`geom_jitter(height=, width=)`

```
ggplot(ess8_CH_ss, aes(x = eduhrs, 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 = "Einstellung zur Migratio  
x = "Bildungsjahre",  
y = "Migrationsestimation",  
subtitle = "Das kulturelle Leben wird durch Mi  
teil
```

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

Streudiagramm (Jitter): Einstellung zur Migration nach Bildung
Das kulturelle Leben wird durch Migration untergraben(0) oder bereichert(10)



Daten ESS8(2016), Teilstichprobe CH(N=1509)

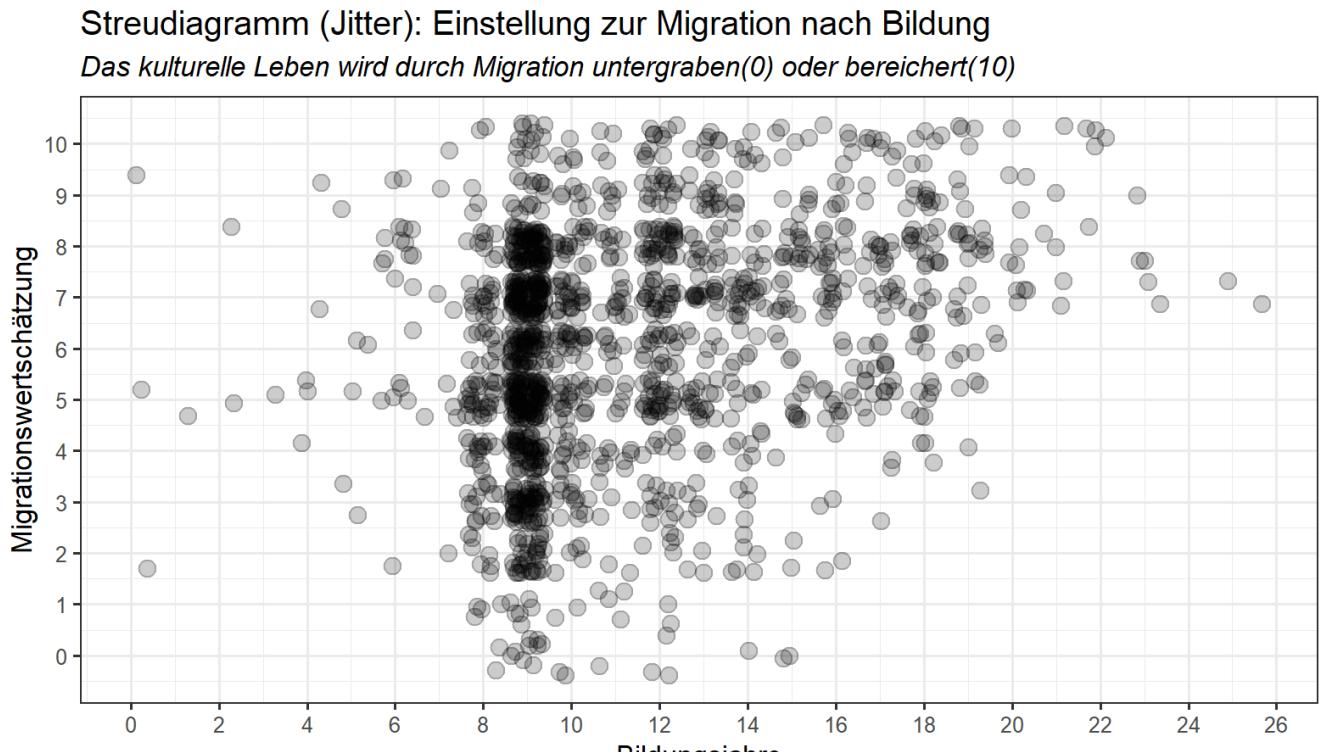
Achtung: Je grösser die Parameter im Jitter-Befehl, desto starker die Verfremdung der Daten.
Hier kommt es z.B. zu einer unschönen Kategorienüberlappung der Punkte!

1.6 Scatterplot – Overplotting: “jitter”

`geom_jitter(alpha =, size =)`

```
ggplot(ess8_CH_ss, aes(x = eduhrs, 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 = "Bildungsjahre",  
  y = "Migrationswertschätzung",  
  subtitle = "Das kulturelle Leben wird durch Migration untergraben(0) oder bereichert(10)")
```

Falls «defensives» Jittern nicht ausreicht:
Besser zusätzlich Alpha (Transparenz) anpassen
statt «aggressives» Jittern!



Daten ESS8(2016), Teilstichprobe CH(N=1509)

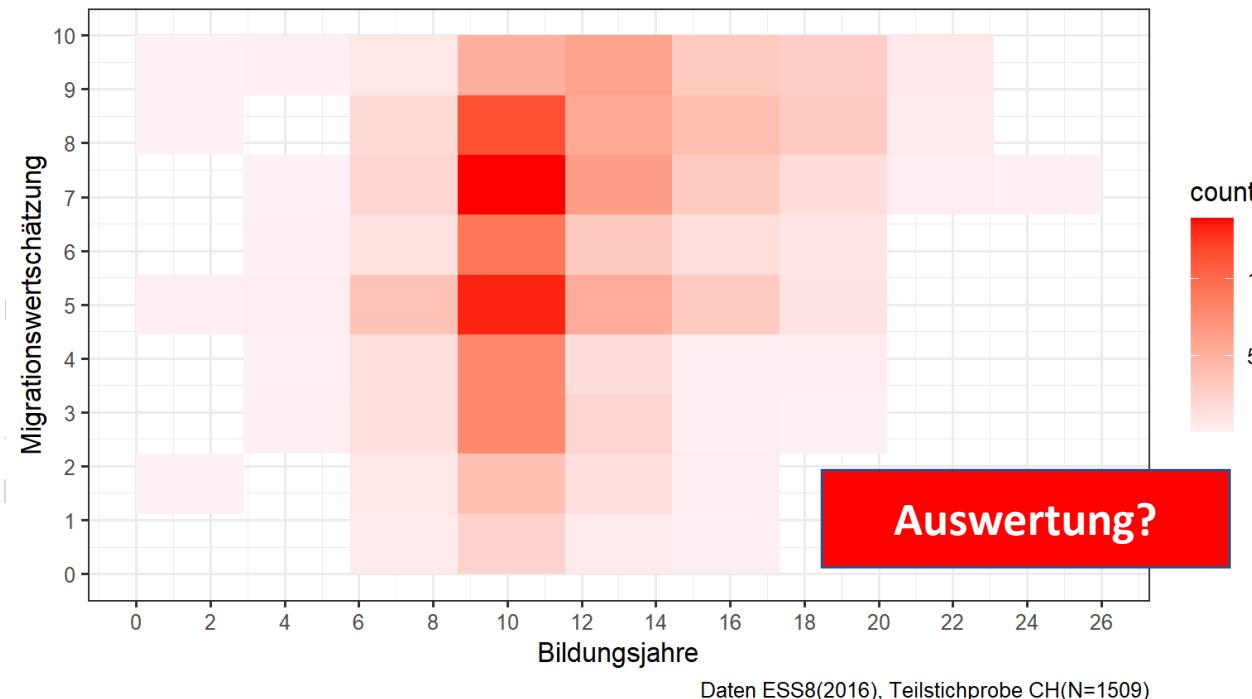
1.6 Scatterplot – Overplotting: “heatmap” (siehe HP!)

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

```
ggplot(ess8_CH_ss, aes(x = eduhrs, 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 Bildung",  
       x = "Bildungsjahre",  
       y = "Migrationswertschätzung",  
       subtitle = "Das kulturelle Leben wird durch Migration untergraben(0) oder bereichert(10)",  
       color = "count",  
       fill = "count")
```

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

Heatmap: Einstellung zur Migration nach Bildung
Das kulturelle Leben wird durch Migration untergraben(0) oder bereichert(10)



Punkte werden zu Clustern zusammengebündelt – eigentlich kein Streudiagramm mehr.
Daher: Nur bei sehr grossen Datensätzen (oder wenigen Merkmalskombinationen) verwenden

1.6

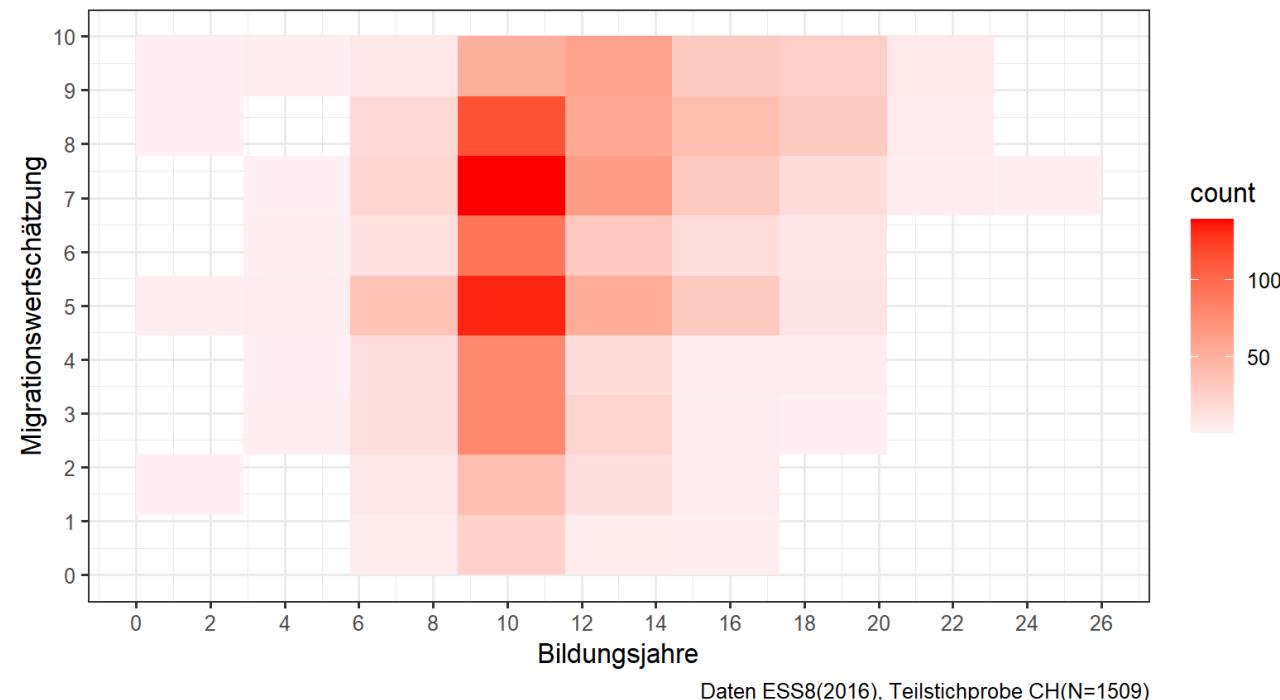
Scatterplot – Overplotting: “heatmap”

```
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.

Heatmap: Einstellung zur Migration nach Bildung
Das kulturelle Leben wird durch Migration untergraben(0) oder bereichert(10)

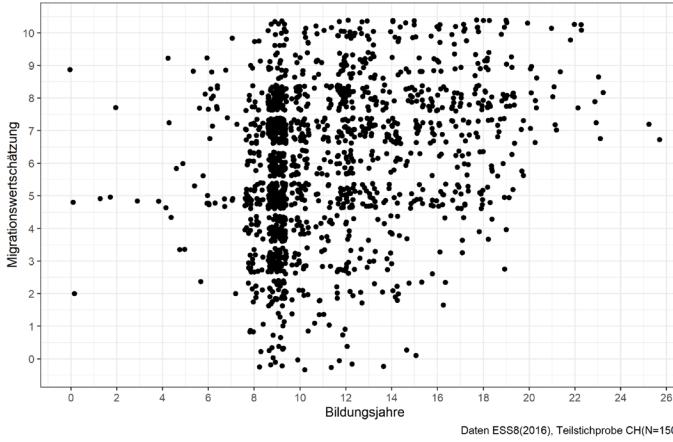


Daten ESS8(2016), Teilstichprobe CH(N=1509)

1.6

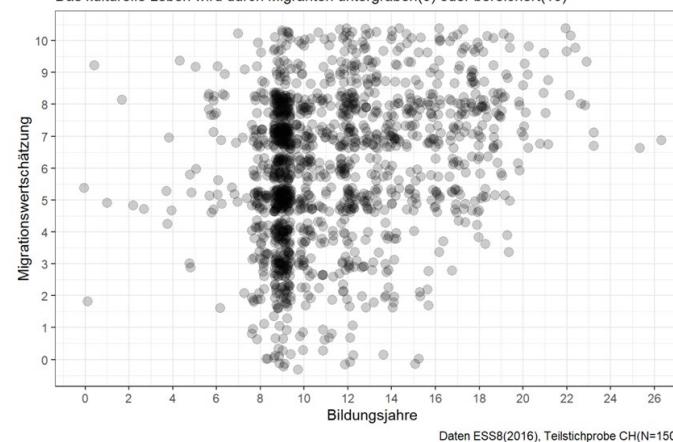
Welche Darstellungsform bei Overplotting?

Einstellung zur Migration nach Bildungsjahren
Das kulturelle Leben wird durch Migranten untergraben(0) oder bereichert(10)



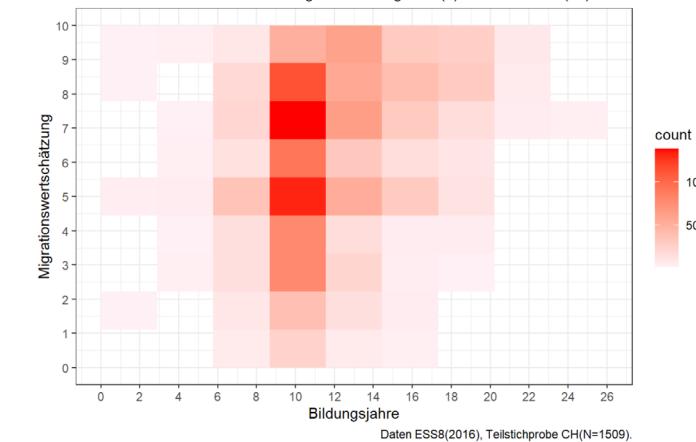
Daten ESS8(2016), Teilstichprobe CH(N=1509).

Einstellung zur Migration nach Bildungsjahren
Das kulturelle Leben wird durch Migranten untergraben(0) oder bereichert(10)



Daten ESS8(2016), Teilstichprobe CH(N=1509).

Einstellung zur Migration nach Bildungsjahren
Das kulturelle Leben wird durch Migranten untergraben(0) oder bereichert(10)



Daten ESS8(2016), Teilstichprobe CH(N=1509).

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

Generell gilt:

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

Teil 3: Regressionsgerade und Regressionskoeffizient



2.1

Ermittlung des Regressionskoeffizienten

Was misst b_1 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 ~ eduysr,  
    data = ess8_CH_ss)
```

Interpretation der Konstante?

```
call:  
lm(formula = imueclt ~ eduysr, data = ess8_CH_ss)
```

einfache technische
Interpretation b_1

```
Coefficients:  
(Intercept) eduysr
```

4.0520

0.1789

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

Die Regressionsgleichung mit berechneten Koeffizienten & Konstante, der «fit»:

$Migrationswertschätzung = 4.05 + 0.18 * Bildungsjahre$

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 zur 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

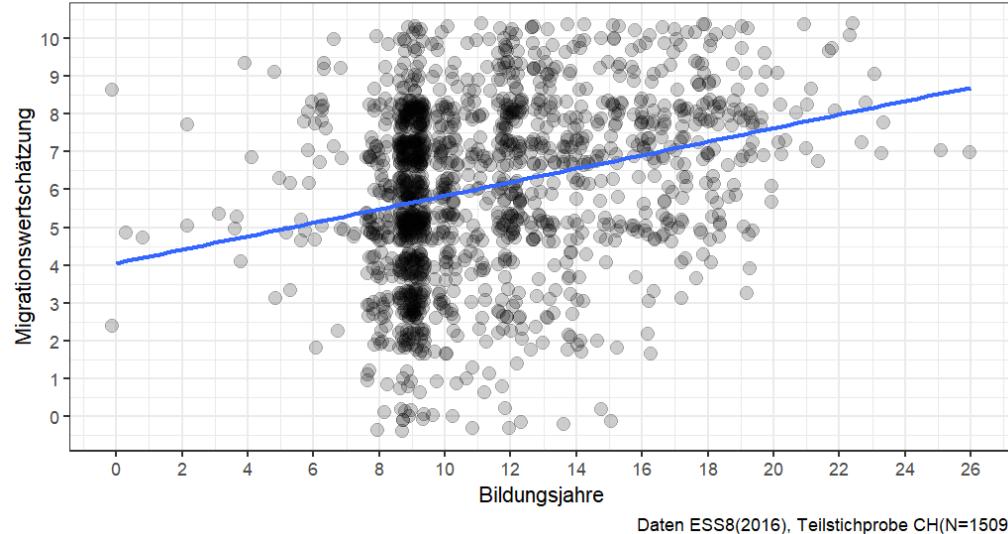
Ohne dabei auf den lm-Output zuzugreifen, bietet die Funktion **geom_smooth()** die Möglichkeit, die Regressionsgerade ins **ggplot**-Streudiagramm zu integrieren.

```
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

Streudiagramm (mit Regressionsgerade): Einstellung zur Migration nach Bildung
Das kulturelle Leben wird durch Migration untergraben(0) oder bereichert(10)

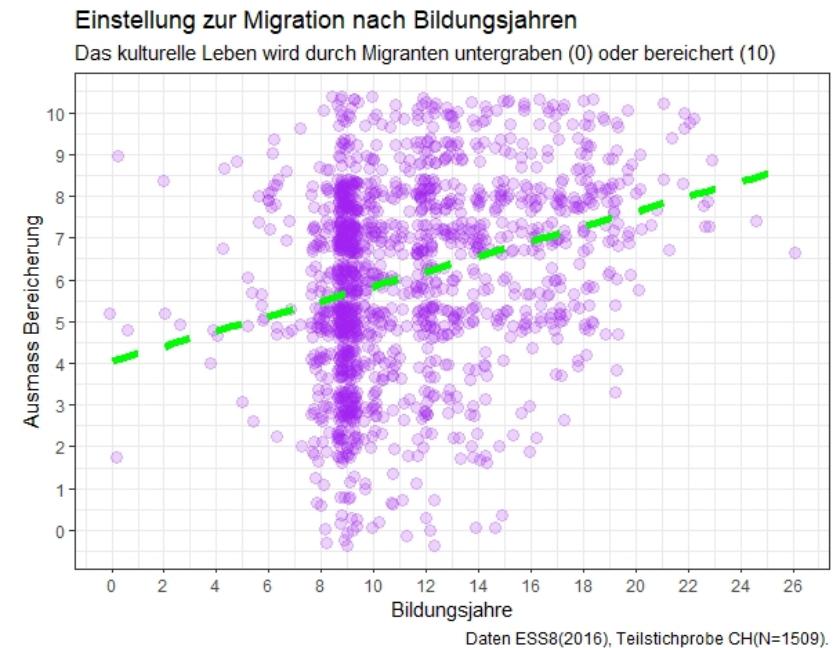
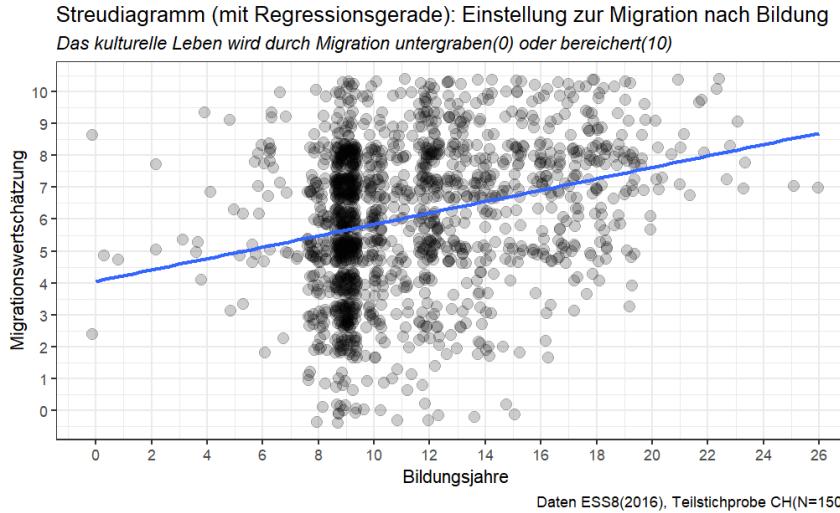


Aufgabe: Modifiziert den Code so, dass die Gerade grün und gestrichelt sowie die Punkte in pink erscheinen

```
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



Here's the rewritten code with the desired changes:

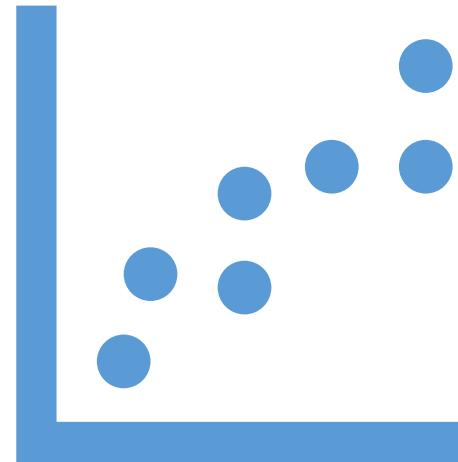


```
perl
```

```
ggplot(ess8_CH_ss, aes(x = eduhrs, 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 = "dashed") +  
  labs(title = "Einstellung zur Migration nach Bildungsjahren",  
       x = "Bildungsjahre",  
       y = "Ausmass Bereicherung",  
       subtitle = "Das kulturelle Leben wird durch Migranten untergraben (0) oder bereichert (10)",  
       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



```
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
```

2.3

Fortgeschrittene Interpretationen (Muster): Jeweils Sinnvoll oder nicht?

Einfache technische bzw. Basis-Interpretation:

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

Reskalierung von x, z.B.

6 zusätzliche Bildungsjahre – also in etwa der Umfang eines vollständigen Studiums (BA+MA) – erhöhen die Migrationswertschätzung im Mittel um etwas mehr als einen Skalenpunkt.

Reskalierung (in Einheiten) von y, z.B.

Um die Migrationswertschätzung um einen ganzen Skalenpunkt zu erhöhen, braucht es im Mittel etwa 5.5 Bildungsjahre.

Vorhersagepaar, z.B.

Variante: Für Personen mit 12 Bildungsjahren (Matura) wird eine Migrationswertschätzung von etwa 6.2 Skalenpunkten, für solche mit 18 Bildungsjahren (BA+MA) eine von etwa 7.3 vorhergesagt.

Standardabweichungs-Bezug?

Was war noch mal die Standardabweichung?

```
Call:  
lm(formula = imueclt ~ eduysrs, data = ess8_CH_ss)  
  
> sd(ess8_CH_ss$eduysrs, na.rm = TRUE)  
[1] 3.496909  
> sd(ess8_CH_ss$imueclt, na.rm = TRUE)  
[1] 2.258924
```

Exkurs: Standardabweichung

Standardabweichung ~ Wie stark weicht eine Person typischerweise vom Mittelwert ab?

Gedankenexperiment:

Sie wählen zufällig eine Person aus der Population aus. Die erwartbare Abweichung ihres Wertes zum Mittelwert ist die Standardabweichung.

Standardabweichung ~ Wie stark weicht eine Person typischerweise vom Mittelwert ab?

Standardabweichung...

... der Körpergrösse Schweizer Männer: **9 cm**

... der Körpergrösse Schweizer Frauen: **7 cm**

... des Monatslohns unter Schweizer Erwerbstätigen: **4500 CHF**

... der Wohnungsgrösse Schweizer Haushalte: **50 qm**

... der Note in der Stat 1 - Klausur unter allen Klausurteilnehmenden: **0.8**

Eine Differenz im Umfang einer Standardabweichung ist der statistische Prototyp eines **grossen Unterschiedes!**

2.3 Fortgeschrittene Interpretation (Muster): Sinnvoll oder nicht?

Einfache technische bzw. Basis-Interpretation:

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

Reskalierung von x, z.B.

6 zusätzliche Bildungsjahre – also in etwa der Umfang eines vollständigen Studiums (BA+MA) – erhöhen die Migrationswertschätzung im Mittel um etwas mehr als einen Skalenpunkt.

Reskalierung (in Einheiten) von y, z.B.

Um die Migrationswertschätzung um einen ganzen Skalenpunkt zu erhöhen, braucht es im Mittel etwa 5.5 Bildungsjahre.

Vorhersagepaar, z.B.

Variante: Für Personen mit 12 Bildungsjahren (Matura) wird eine Migrationswertschätzung von etwa 6.2 Skalenpunkten, für solche mit 18 Bildungsjahren (BA+MA) eine von etwa 7.3 vorhergesagt.

Standardabweichungs-Bezug?

Um die Migrationswertschätzung um eine ganze Standardabweichung zu erhöhen, braucht es im Mittel etwa 12,5 Bildungsjahre.

```
Call:  
lm(formula = imueclt ~ eduysr, data = ess8_CH_ss)  
  
Coefficients:  
(Intercept)      eduysr  
        4.0520       0.1789  
  
> sd(ess8_CH_ss$eduysr, na.rm = TRUE)  
[1] 3.496909  
> sd(ess8_CH_ss$imueclt, na.rm = TRUE)  
[1] 2.258924
```

2.3 Fortgeschrittene Interpretation (Muster): Sinnvoll oder nicht?

```
Call:  
lm(formula = imueclt ~ eduysrs, data = ess8_CH_ss)  
[1] 3.496909  
> sd(ess8_CH_ss$eduysrs, na.rm = TRUE)  
[1] 2.258924  
Coefficients:  
(Intercept)      eduysrs  
4.0520          0.1789
```

Um die Migrationswertschätzung um eine ganze Standardabweichung zu erhöhen, braucht es im Mittel etwa 12,5 Bildungsjahre.

Einfluss der Bildung hier gross oder klein?

2.3

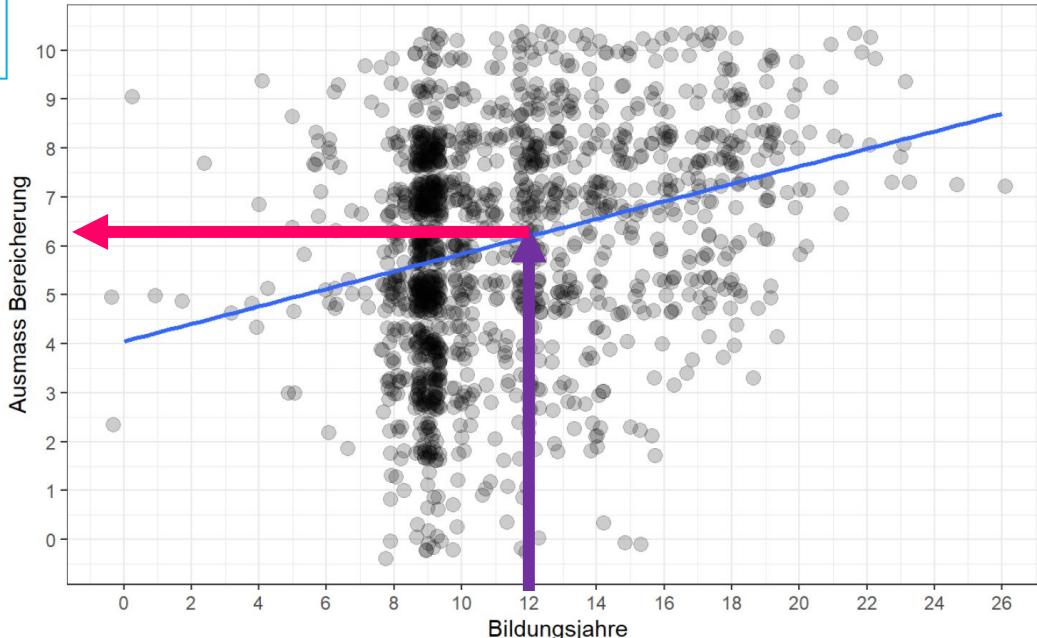
Interpretation – Vorhersagen einfach machen mit R

Welche Migrationswertschätzung wird für eine Person mit 12 bzw. 15 Bildungsjahren vorhergesagt?

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

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

Einstellung zur Migration nach Bildungsjahren
Das kulturelle Leben wird durch Migranten untergraben(0) oder bereichert(10)



`ggpredict()` übernimmt die Arbeit für uns!

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

Predicted values of country's cultural life undermined or enriched by immigrants

eduys	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 Interpretation: Überblick Standardisierungen

```
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 Standardabweichungen (also etwas weniger als 1/10 eines typischen Einstellungsunterschiedes) positiver.

Standardisierung von Bezugs- und Zieleinheit (UV & AV)

$(0.18*3.50/2.25 = 0.28)$ Ein typischer Bildungsunterschied erzeugt im Mittel eine Veränderung der Einstellung zur Migration um etwa 0.28 Standardabweichungen.

Übung

Untersucht nun selbstständig den Einfluss der *Internetnutzung in Minuten* (UV) auf die *Lebenszufriedenheit* (AV) in der Schweiz (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. Bietet ein sinnvolles Interpretationsangebot zur Verdeutlichung der Effektgrösse an

Weitere Übung: Siehe HP