Statistik 2 – Tutorate
Sitzung 5: Inferenzstatistik

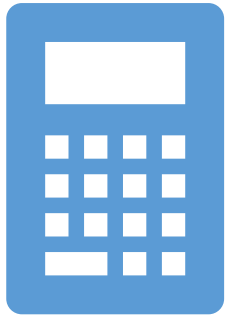
Marco Giesselmann, Rémy Blum, Federica Bruno, Simon Honegger, Nora Zumbühl

Lernziele dieser Einheit

Hypothesen

Formulierung von Null-

und Alternativhypothese



Inferenzstatistik

Standardfehler, t- und p-Wert

Konfidenzintervalle und Konfidenzband

R^2 und Vorhersageband

1 Hypothese formulieren

Wie hängen *Bildung* und *Dauer der täglichen Internetnutzung* zusammen?

Ad-hoc Vermutung (mit Brückenhypothese): Bildung führt oft zu computerbasierten und netzwerkintensiven Erwerbsprofilen, die wiederum mit vielfältiger und intensiver Internetnutzung verbunden sind.

Wie lauten entsprechend Forschungshypothese und Nullhypothese?

Hypothese (H1) / Forschungshypothese:
Bildung hat einen positiven Einfluss auf den zeitlichen Umfang der Internetnutzung.

Nullhypothese (H0):
*Bildung hat **keinen** positiven Einfluss auf den zeitlichen Umfang der Internetnutzung*

1 Hypothese empirisch prüfen

Hypothese (H1) / Forschungshypothese:
Bildung hat einen positiven Einfluss auf den zeitlichen Umfang der Internetnutzung.

Regressionsanalyse:

1. Berechnung des Regressionskoeffizienten und weiterer relevanter Kennwerte (p-Wert, Standardfehler)
2. Checks (machen wir heute aufgrund alternativen didaktischen Fokus nicht):
   - **Zusammenhangsform** beschreiben und einordnen: näherungsweise linear, unproblematisch linearitätsabweichend oder problematisch linearitätsabweichend?
   - **Ausreisserdiagnose:** Einflussreiche Werte vorhanden? Regressionsergebnisse ggf. robust?
3. Durch geeignete Interpretation und Visualisierung versuchen wir zu klären, ob der Koeffizient auf einen ***inhaltlich bedeutsamen Einfluss verweist***.
4. Dann prüfen wir, ob der Koeffizient ***statistische bedeutsam bzw. statistisch signifikant von 0 verschieden*** ist (über den Grad an Überzufälligkeit, ausgedrückt im p-Wert).
5. Liegt ein überzufälliges Ergebnis bzw. ***hinreichend niedriger p-Wert*** vor, wird unsere ***Forschungshypothese gestützt bzw. «die Nullhypothese abgelehnt».***

2.1 Datenmanagement – Inspektion und Selektion

Wir verwenden die Variablen **eduyrs** und **netustm aus dem ESS** und beschränken zudem die Stichprobe auf den Teildatensatz der Schweiz. Wir verschaffen uns einen Überblick über die beiden Variablen und reduzieren unseren Datensatz für die Regressionsanalyse.

```
ess8_CH <- filter(ess8, cntry == "CH")
look_for(ess8_CH, "eduyrs")
look_for(ess8_CH, "netustm")
ess8_CH <- select(ess8_CH, internet = netustm, eduyrs, idno)
summary(ess8_CH)
sd(ess8_CH$eduyrs, na.rm = TRUE)
sd(ess8_CH$internet, na.rm = TRUE)
```

- **eduyrs:** Anzahl an abgeschlossenen Bildungsjahren.
- **netustm:** Internetnutzung in Minuten pro Tag.

```
> sd(ess8_CH$internet, na.rm = TRUE)
[1] 163.1974
```

Die typische Abweichung vom Mittelwert der Internetnutzung beträgt 163 Minuten

2.2 Regressionsanalyse: Regressionskoeffizient

```
fit <- lm(internet ~ eduyrs, data = ess8_CH)
summary(fit)
```

```
Call:
lm(formula = internet ~ eduyrs, data = ess8_CH)

Residuals:
    Min      1Q  Median      3Q     Max
-204.19 -105.20  -52.06   44.16 1011.66

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

Residual standard error: 161.1 on 1182 degrees of freedom
  (341 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.02838,   Adjusted R-squared:  0.02755
F-statistic: 34.52 on 1 and 1182 DF,  p-value: 5.48e-09
```

Der Koeffizient zeigt einen positiven Zusammenhang an. Mit jedem Bildungsjahr steigt die tägliche Nutzungszeit des Internets im Schnitt um 7 Minuten und 43 Sekunden an.

- Interpretation Regressionskoeffizient?
- Grösseneinordnung Regressionskoeffizient?
- Bedeutung Standardabweichung vs. Standardfehler?

```
> sd(ess8_CH_ss$internet, na.rm = TRUE)
[1] 163.1974
```

Standardfehler und **Standardabweichung** sind unterschiedliche statistische Kennwerte:

- Die Standardabweichung misst **die Variation einer Variable** innerhalb der realisierten Stichprobe: Was ist die typische, erwartbare Abweichung eines realisierten Wertes von Mittelwert?
- Der Standardfehler misst die **Variation des Regressionskoeffizienten** zwischen verschiedenen hypothetischen Stichproben: Was ist die typische, erwartbare Abweichung des erzielten Koeffizienten in der Stichprobe vom wahren Koeffizienten der Population ?

2.3 Standardfehler

```
Call:
lm(formula = internet ~ eduyrs, data = ess8_CH)

Residuals:
    Min      1Q  Median      3Q     Max
-204.19 -105.20  -52.06   44.16 1011.66

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

Residual standard error: 161.1 on 1182 degrees of freedom
  (341 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.02838,   Adjusted R-squared:  0.02755
F-statistic: 34.52 on 1 and 1182 DF,  p-value: 5.48e-09
```

Der Standardfehler (SE) gibt die durchschnittliche bzw. erwartbare **Abweichung** eines Stichprobenkennwertes vom wahren Parameterwert in der **Grundgesamtheit** an.

Wie lautet die konkrete Interpretation *dieses* Standardfehlers?

Wir müssen erwarten, dass der «wahre» Anstieg der Nutzungsdauer (in der Population) pro Bildungsjahr um 1.3 Minuten grösser oder kleiner ausfällt als 7.7. Es handelt sich also um eine relativ vertrauenswürdige Schätzung des Regressionskoeffizienten.

2.4 t-Wert

```
Call:
lm(formula = internet ~ eduyrs, data = ess8_CH)

Residuals:
    Min      1Q  Median      3Q     Max
-204.19 -105.20  -52.06   44.16 1011.66

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

Residual standard error: 161.1 on 1182 degrees of freedom
  (341 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.02838,   Adjusted R-squared:  0.02755
F-statistic: 34.52 on 1 and 1182 DF,  p-value: 5.48e-09
```

Was misst der t-Wert?

Das Verhältnis von Koeffizient und SE wird durch den t-Wert angegeben.

$$t = \frac{b}{SE} = \frac{7.713}{1.313}$$

Dieser misst also die Grösse des Koeffizienten in Einheiten des Standardfehlers

2.5 p-Wert

```
Call:
lm(formula = internet ~ eduyrs, data = ess8_CH)

Residuals:
    Min      1Q  Median      3Q     Max
-204.19 -105.20  -52.06   44.16 1011.66

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

Residual standard error: 161.1 on 1182 degrees of freedom
  (341 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.02838,    Adjusted R-squared:  0.02755
F-statistic: 34.52 on 1 and 1182 DF,  p-value: 5.48e-09
```

Der p-Wert zeigt uns, wie wahrscheinlich der vorgefundene Stichprobenkoeffizient (oder ein grösserer) ist, wenn es in Wirklichkeit keinen Zusammenhang zwischen UV und AV gibt bzw. die (beidseitige Variante) der Nullhypothese richtig wäre.

Wie lautet die konkrete Interpretation dieses p-Wertes?

Unter der Bedingung, dass es in der Population keinen Zusammenhang zwischen Bildung und Webnutzung gibt, tritt das vorliegende (oder ein extremeres) Stichprobenergebnis mit einer Wahrscheinlichkeit von 0.00000000548 auf.

Bei sehr kleinem p-Wert wird dieser von R standardmässig in Exponentialschreibweise ausgegeben. Dezimaldarstellung hier: 0.00000000548

2.5 p-Wert

```
Call:
lm(formula = internet ~ eduyrs, data = ess8_CH)

Residuals:
    Min      1Q  Median      3Q     Max
-204.19 -105.20  -52.06   44.16 1011.66

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

Residual standard error: 161.1 on 1182 degrees of freedom
  (341 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.02838,   Adjusted R-squared:  0.02755
F-statistic: 34.52 on 1 and 1182 DF,  p-value: 5.48e-09
```

Als Konvention[1] gelten die Schwellenwerte $p < 0.05$ und $p < 0.01$ für die Feststellung statistischer Signifikanz bzw. Ablehnung der Nullhypothese.

Dem Signifikanzniveau entsprechend werden Sterne verteilt.

[1]Diese Konventionen sind nicht disziplinübergreifend. Diskussionen um den p-Wert als Kriterium und den angemessenen H0-Ablhenungsschwellenwert auf wissenschaftlicher Ebene dauern an.

2.6 Hypothesenevaluation

```
Call:
lm(formula = internet ~ eduyrs, data = ess8_CH)

Residuals:
    Min      1Q  Median      3Q     Max
-204.19 -105.20  -52.06   44.16 1011.66

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

Residual standard error: 161.1 on 1182 degrees of freedom
  (341 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.02838,   Adjusted R-squared:  0.02755
F-statistic: 34.52 on 1 and 1182 DF,  p-value: 5.48e-09
```

Bewerte die Nullhypothese auf Basis des p-Wertes

Wir können die Nullhypothese, dass es keinen [positiven] Zusammenhang zwischen den beiden Variablen gibt[1], ablehnen.

Unsere Forschungshypothese wird gestützt.

Warum die eckige Klammer in der Antwort? Eigentlich bezieht sich der von R::lm() berichtete p-Wert auf eine *zweiseitige*, also *tendenzlose* Hypothese. Wir können aber den abgeleiteten zweiseitig ermittelten p-Wert einer Konvention folgend für die Prüfung unserer *einseitigen* Hypothese verwenden. Letztlich wird hierdurch die Schwelle für die Verwerfung der H0 höher gelegt (siehe Folien letzte Vorlesung).

2.6 Übung

Führt einen Hypothesentest zum Zusammenhang von **Bildungsniveau** und **Migrationswertschätzung** durch (Siehe Einheit „III.Basics“)

- Formuliert Forschungs- und Nullhypothese.
- Interpretiert den Koeffizienten (Einfache technische Basisinterpretation).
- Interpretiert SE und p-Wert.
- Zusatzaufgabe: Leitet das Konfidenzintervall des Regressionskoeffizienten ab
- wird die Forschungshypothese gestützt?

```
##
## 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
```

In den letzten beiden Vorlesungen behandelt:

Konfidenz- und Vorhersageintervalle bzw. –bänder

Parameterkonfidenz

Haben wir den richtigen, „wahren“ Regressionskoeffizienten der Population gefunden?

→

- **Standardfehler des Regressionskoeffizienten**
- **p-Wert des Regressionskoeffizienten**
- **Konfidenzintervall des Regressionskoeffizienten**
- **Konfidenzband der Regressionsgerade**

Vorhersagekonfidenz

Wie gross ist die Streuung um (durch die Regressionsgerade) vorhergesagte Einzelwerte?

→

- **R^2**
- **Vorhersageband der Regressionsgerade**
- **Regressionsbasiertes Vorhersageintervall**

Tutorat heute: Einübung der Konzepte und Umsetzung mit R

3.1 Beispiel: Bildungsjahre → Internetnutzung (Minuten/Tag)

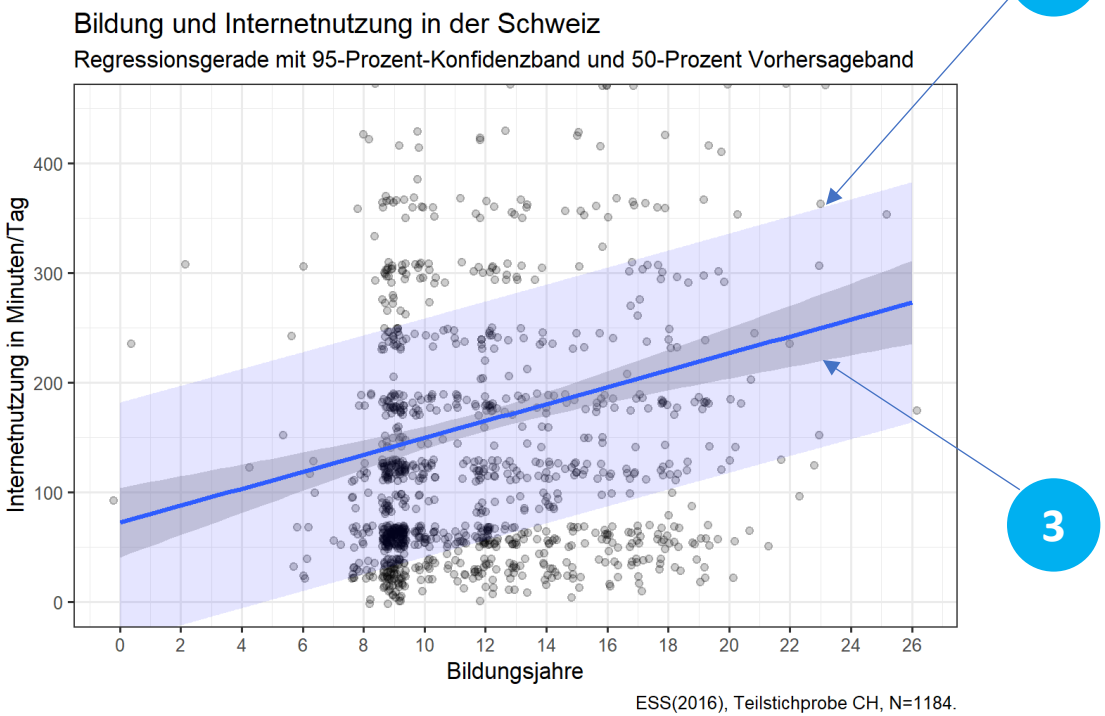
```
fit <- lm(internet ~ eduyrs, data = ess8_CH)
```

```
> confint(fit, level = 0.95)
                2.5 %    97.5 %
(Intercept) 40.918306 104.37282
eduyrs       5.137214  10.28828
```

1

2

3



Was sind *1*, *2* und *3*?

Antwort:

1 95%-Konfidenzintervall des Koeffizienten

2 50%-Vorhersageband

3 95%-Konfidenzband

Das Konfidenzintervall des Regressionskoeffizienten

```
> confint(fit, level = 0.95)
                2.5 %    97.5 %
(Intercept) 40.918306 104.37282
eduyrs       5.137214  10.28828
```

3.2 Konfidenzintervall des Koeffizienten

Das Konfidenzintervall des Koeffizienten zeigt an, zwischen welchen Werten der wahre Koeffizient in der Grundgesamtheit mit 95%-Sicherheit liegt.

```
> confint(fit, level = 0.95)
                2.5 %    97.5 %
(Intercept) 40.918306 104.37282
eduyrs       5.137214  10.28828
```

*Wir können die 95%-KI-Grenzen auch „per Hand“ ermitteln (Koeffizient ± 1,96 * SE)*

```
Coefficients:
             Estimate Std. Error
(Intercept)   72.646     16.171
eduyrs         7.713      1.313
---
Signif. codes:  0 ‘***’ 0.001 ‘*
```

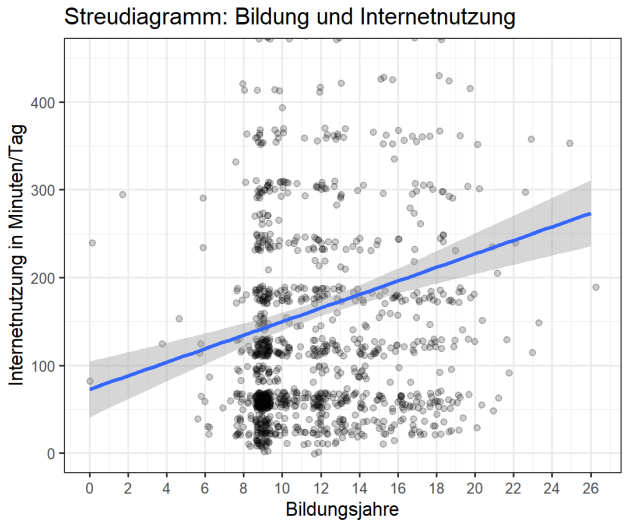
Interpretation im konkreten empirischen Fall?

Der wahre Steigungskoeffizient der Grundgesamtheit liegt mit 95% Sicherheit zwischen 5.14 und 10.29.

Eleganter , aber oft auch etwas sperrig: Mit Einbindung der Variablen:

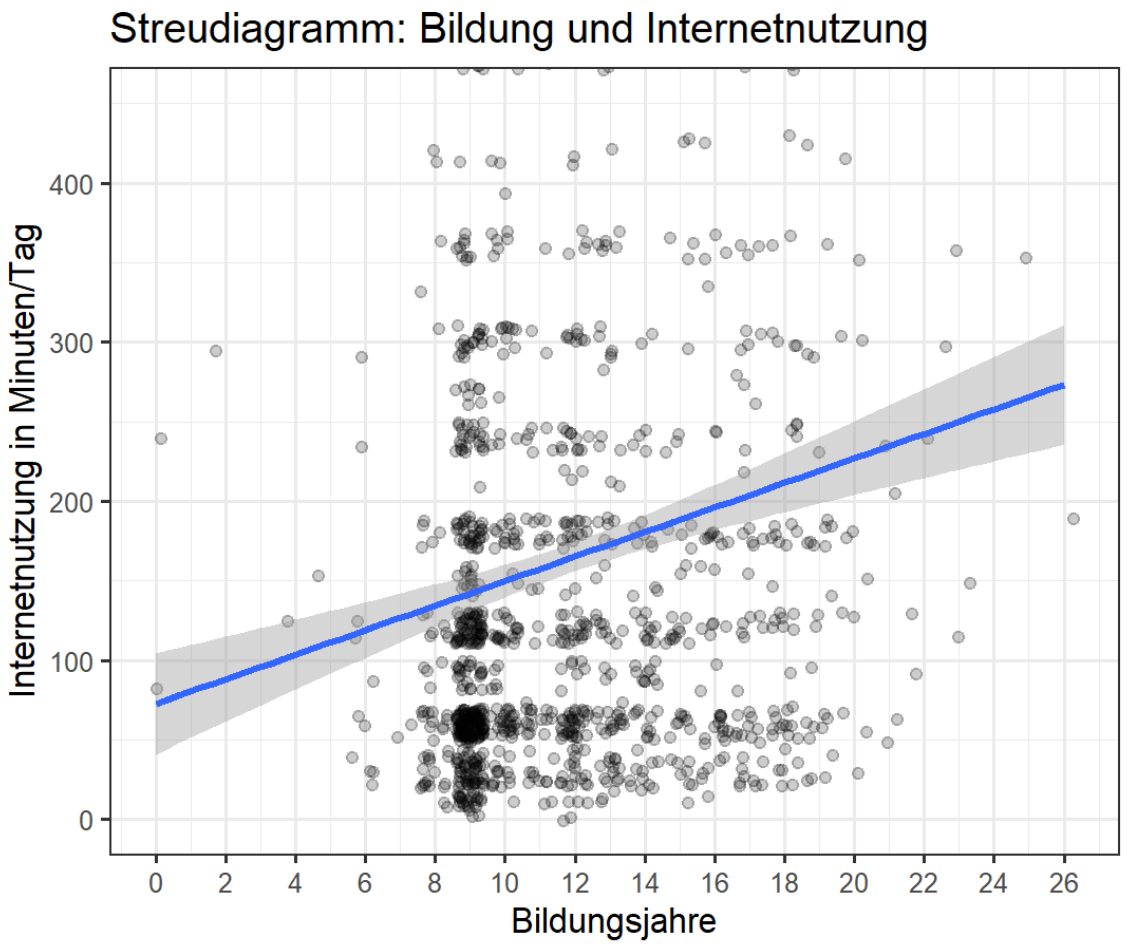
Mit 95% Sicherheit steigt mit jedem zusätzlichen Bildungsjahr die Dauer der durchschnittlichen Internetnutzung in der Population zwischen etwa 5 und 10 min

Das Konfidenzband der Regressionsgerade



3.3 Konfidenzband der Regressionsgerade

Bedeutung des (hier) grauen Bereichs?

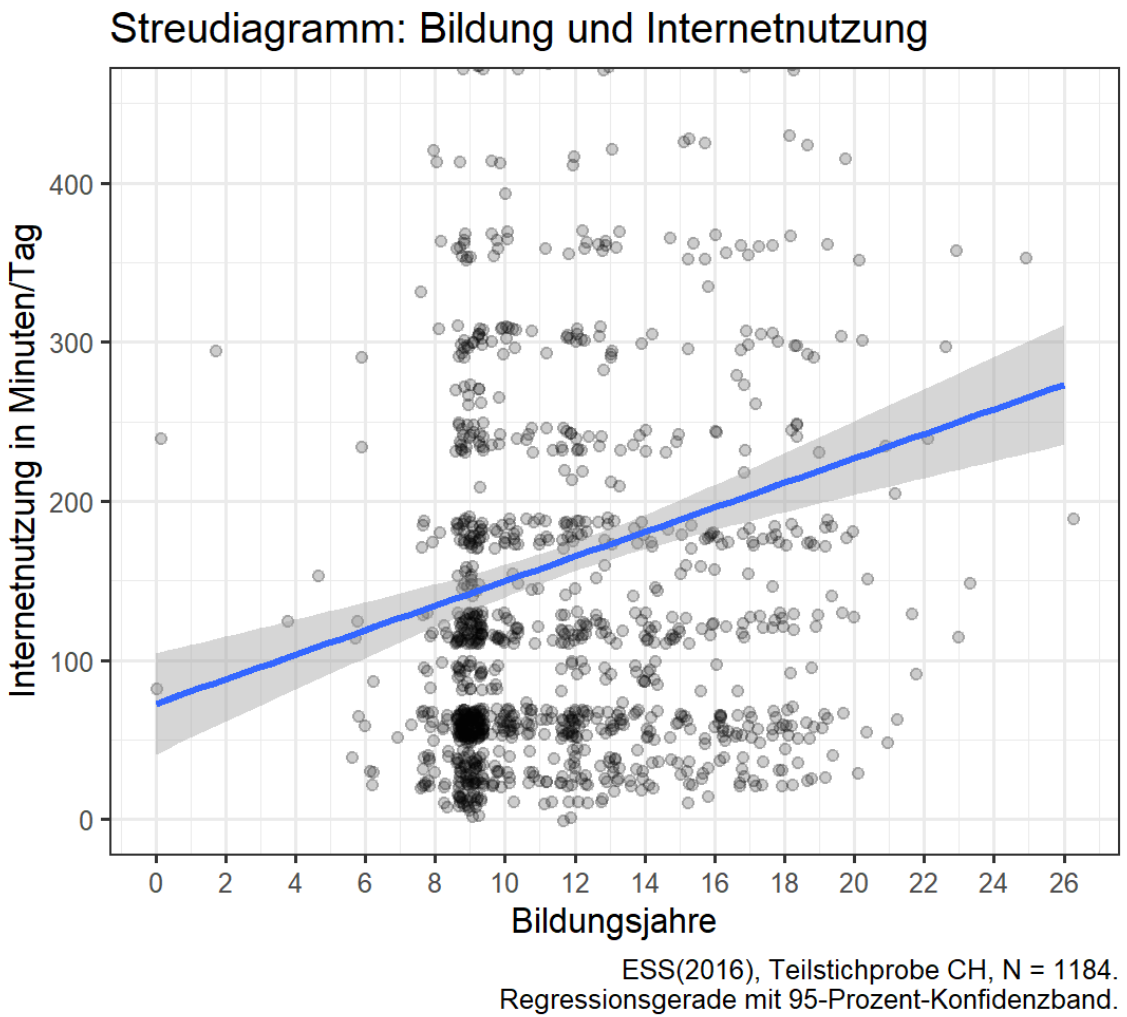


ESS(2016), Teilstichprobe CH, N = 1184.
Regressionsgerade mit 95-Prozent-Konfidenzband.

Das Konfidenzband zeigt den Bereich an, in dem die wahre Regressionsgerade der Population mit 95%-Sicherheit verläuft.

3.3 Konfidenzband der Regressionsgerade

Darstellung des Konfidenzbandes im ggplot-Scatterplot

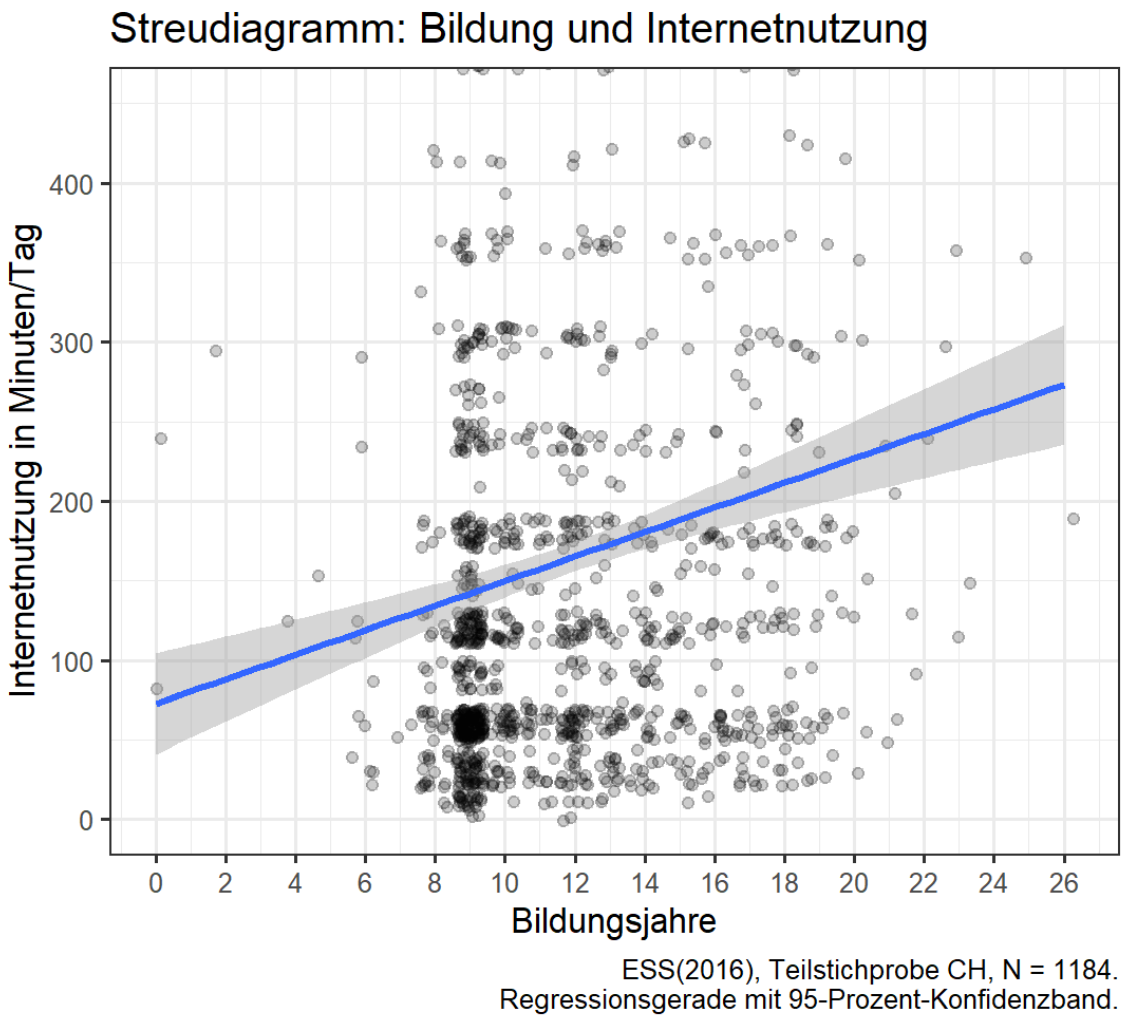



```
ggplot(ess8_CH, aes(x = eduyrs, y = internet))+
  geom_jitter(alpha = 0.2, height = 10) +
  scale_x_continuous(breaks = seq(0,26,2))+
  coord_cartesian(ylim = c(0,450)) +
  geom_smooth(method = "lm",
              se = TRUE,
              level = 0.95)+
  theme_bw()+
  labs(title = "Streudiagramm: Bildung und Internetnutzung",
       y = "Internetnutzung in Minuten/Tag",
       x = "Bildungsjahre",
       caption = "ESS(2016), Teilstichprobe CH, N = 1184.\n
```

Variiert die Spezifikation des Konfidenzbandes!

3.3 Konfidenzband der Regressionsgerade

Darstellung des Konfidenzbandes im ggplot-Scatterplot

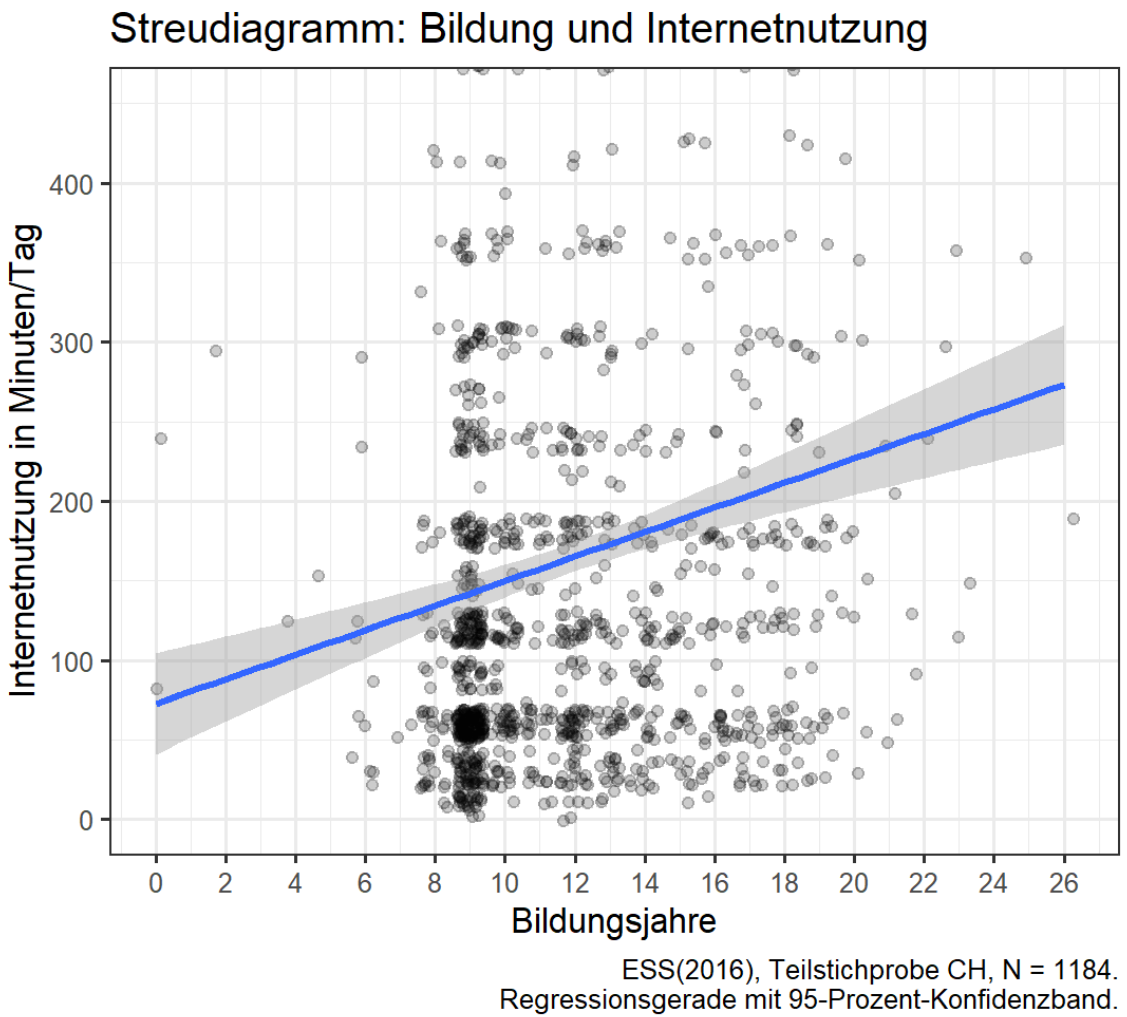



```
ggplot(ess8_CH, aes(x = eduyrs, y = internet))+
  geom_jitter(alpha = 0.2, height = 10) +
  scale_x_continuous(breaks = seq(0,26,2))+
  coord_cartesian(ylim = c(0,450)) +
  geom_smooth(method = "lm",
              se = TRUE,
              level = 0.95)+
  theme_bw()+
  labs(title = "Streudiagramm: Bildung und Internetnutzung",
       y = "Internetnutzung in Minuten/Tag",
       x = "Bildungsjahre",
       caption = "ESS(2016), Teilstichprobe CH, N = 1184.\n
```

Weshalb wird das Band bei «level=0.99» breiter?

3.3 Konfidenzband der Regressionsgerade

Darstellung des Konfidenzbandes im ggplot-Scatterplot

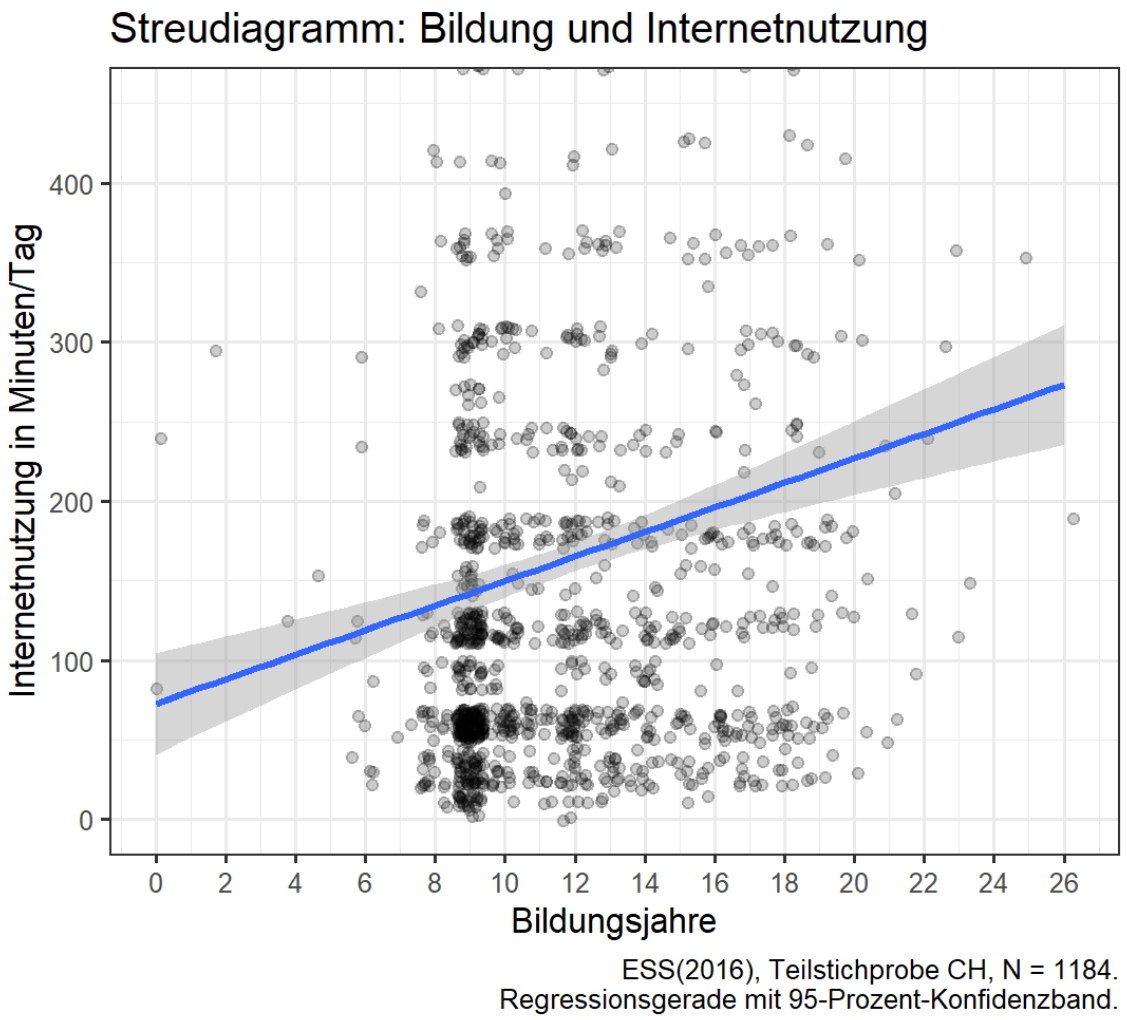



```
ggplot(ess8_CH, aes(x = eduyrs, y = internet))+
  geom_jitter(alpha = 0.2, height = 10) +
  scale_x_continuous(breaks = seq(0,26,2))+
  coord_cartesian(ylim = c(0,450)) +
  geom_smooth(method = "lm",
              se = TRUE,
              level = 0.95)+
  theme_bw()+
  labs(title = "Streudiagramm: Bildung und Internetnutzung",
       y = "Internetnutzung in Minuten/Tag",
       x = "Bildungsjahre",
       caption = "ESS(2016), Teilstichprobe CH, N = 1184.\n
```

Was passiert, wenn beide blau umrandeten Befehlselemente weggelassen werden?

3.3 Konfidenzband der Regressionsgerade

Darstellung des Konfidenzbandes im ggplot-Scatterplot

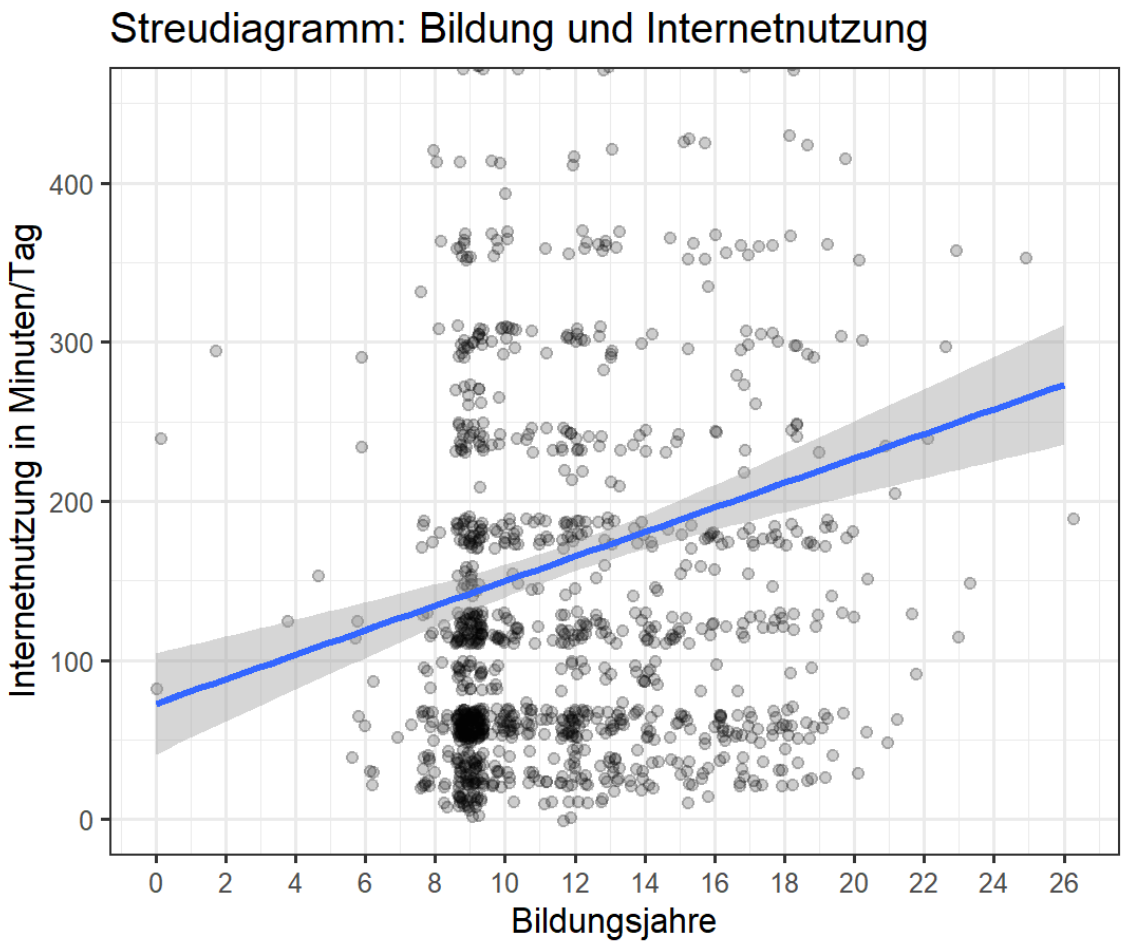



```
ggplot(ess8_CH, aes(x = eduyrs, y = internet))+
  geom_jitter(alpha = 0.2, height = 10) +
  scale_x_continuous(breaks = seq(0,26,2))+
  coord_cartesian(ylim = c(0,450)) +
  geom_smooth(method = "lm",
              se = TRUE,
              level = 0.95)+
  theme_bw()+
  labs(title = "Streudiagramm: Bildung und Internetnutzung",
       y = "Internetnutzung in Minuten/Tag",
       x = "Bildungsjahre",
       caption = "ESS(2016), Teilstichprobe CH, N = 1184.\n
```

PS: **coord_cartesian()** *wählt einen Ausschnitt aus dem Plot, nutzt aber Werte ausserhalb weiterhin für die Regressionsgerade. (Achtung: diese Eigenschaft haben andere Befehle zur Einschränkung des Ausschnittes (z.B. xlim) nicht!)*

3.3 Konfidenzband der Regressionsgerade

Aussagekraft des 95% Konfidenzbandes

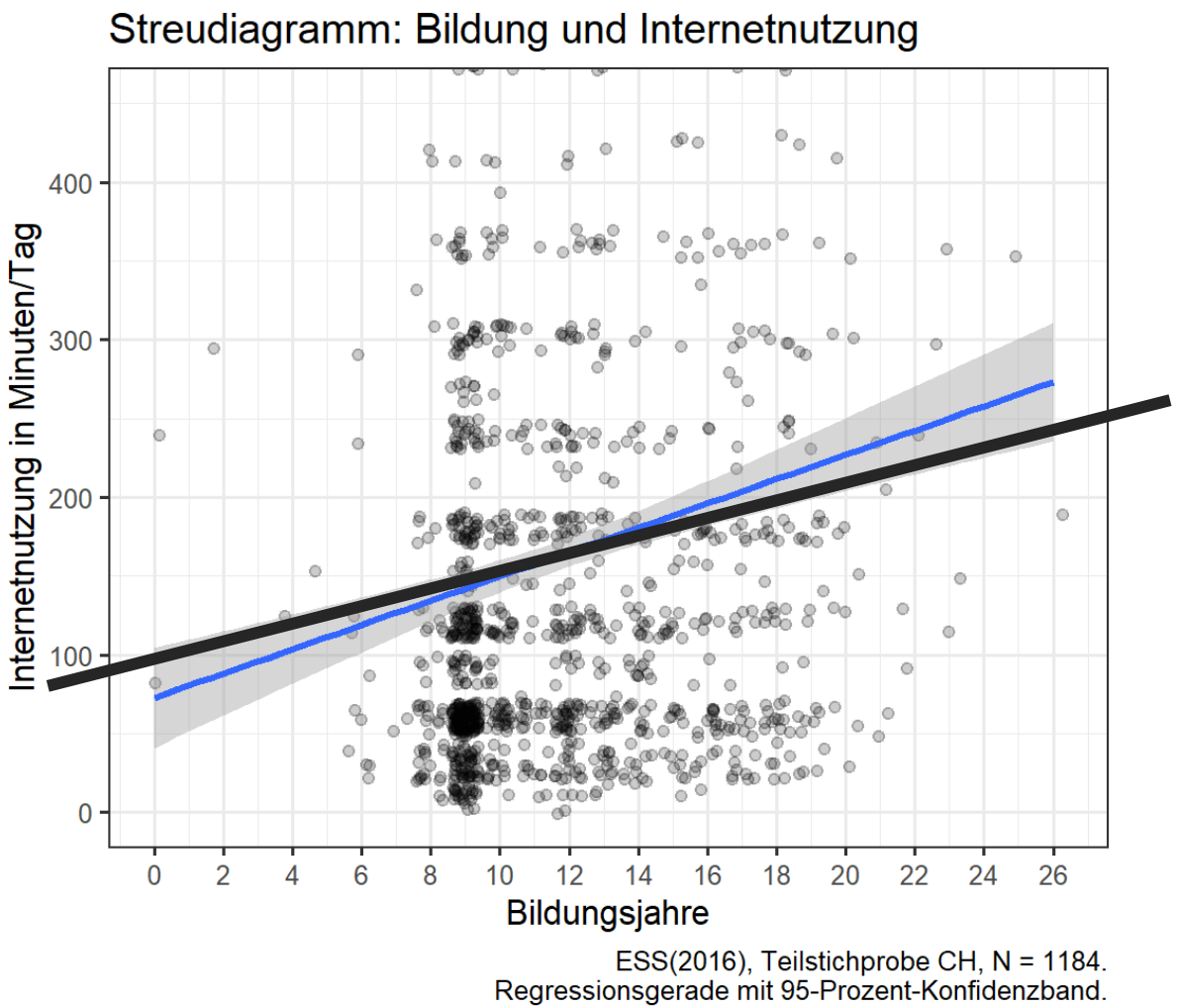


ESS(2016), Teilstichprobe CH, N = 1184.
Regressionsgerade mit 95-Prozent-Konfidenzband.

Welche konkreten Rückschlüsse auf die Sicherheit bzw. Vertrauenswürdigkeit des Koeffizienten lässt das Konfidenzband hier zu?

3.3 Konfidenzband der Regressionsgerade

Aussagekraft des 95% Konfidenzbandes



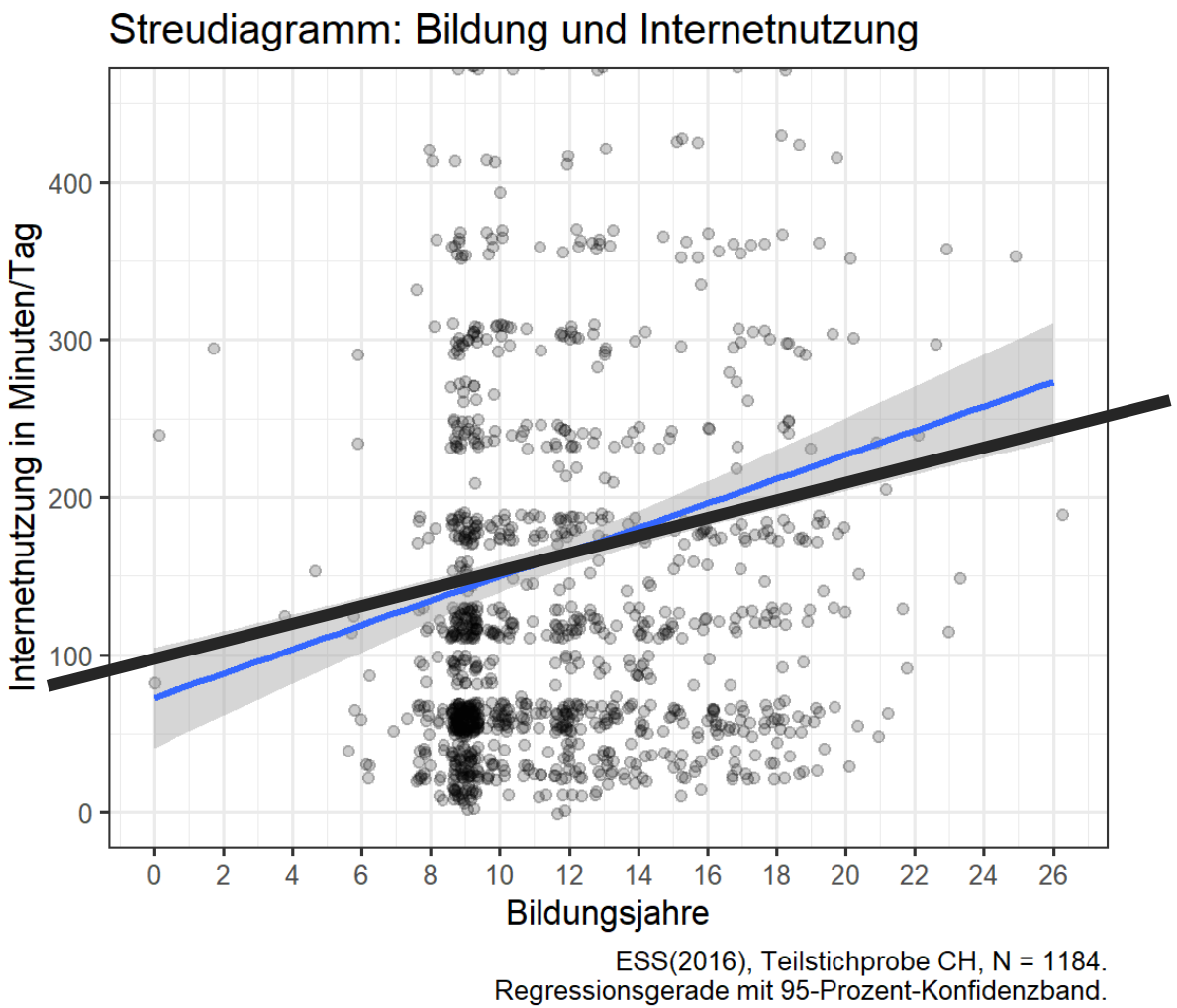
ESS(2016), Teilstichprobe CH, N = 1184.
Regressionsgerade mit 95-Prozent-Konfidenzband.

Welche konkreten Rückschlüsse auf die Sicherheit bzw. Vertrauenswürdigkeit des Koeffizienten lässt das Konfidenzband hier zu?

- *Selbst eine Gerade mit negativem Extremverlauf innerhalb des Konfidenzbandes weist eine deutlich positive Steigung auf.*
- *Der grosse Sicherheitsabstand zur Steigung 0 (und somit zum in der Nullhypothese aufgegriffenen Sachverhalt) unserer Regressionsgerade wird folglich deutlich.*
- *Dieses Konfidenzband drückt also ein **hohes Vertrauen** darin aus, dass die **wahre Regressionsgerade** einen **positiven Steigungskoeffizienten** hat!*

3.3 Konfidenzband der Regressionsgerade

Aussagekraft des 95% Konfidenzbandes



ESS(2016), Teilstichprobe CH, N = 1184.
Regressionsgerade mit 95-Prozent-Konfidenzband.

Welche konkreten Rückschlüsse auf die Sicherheit bzw. Vertrauenswürdigkeit des Koeffizienten lässt das Konfidenzband hier zu?

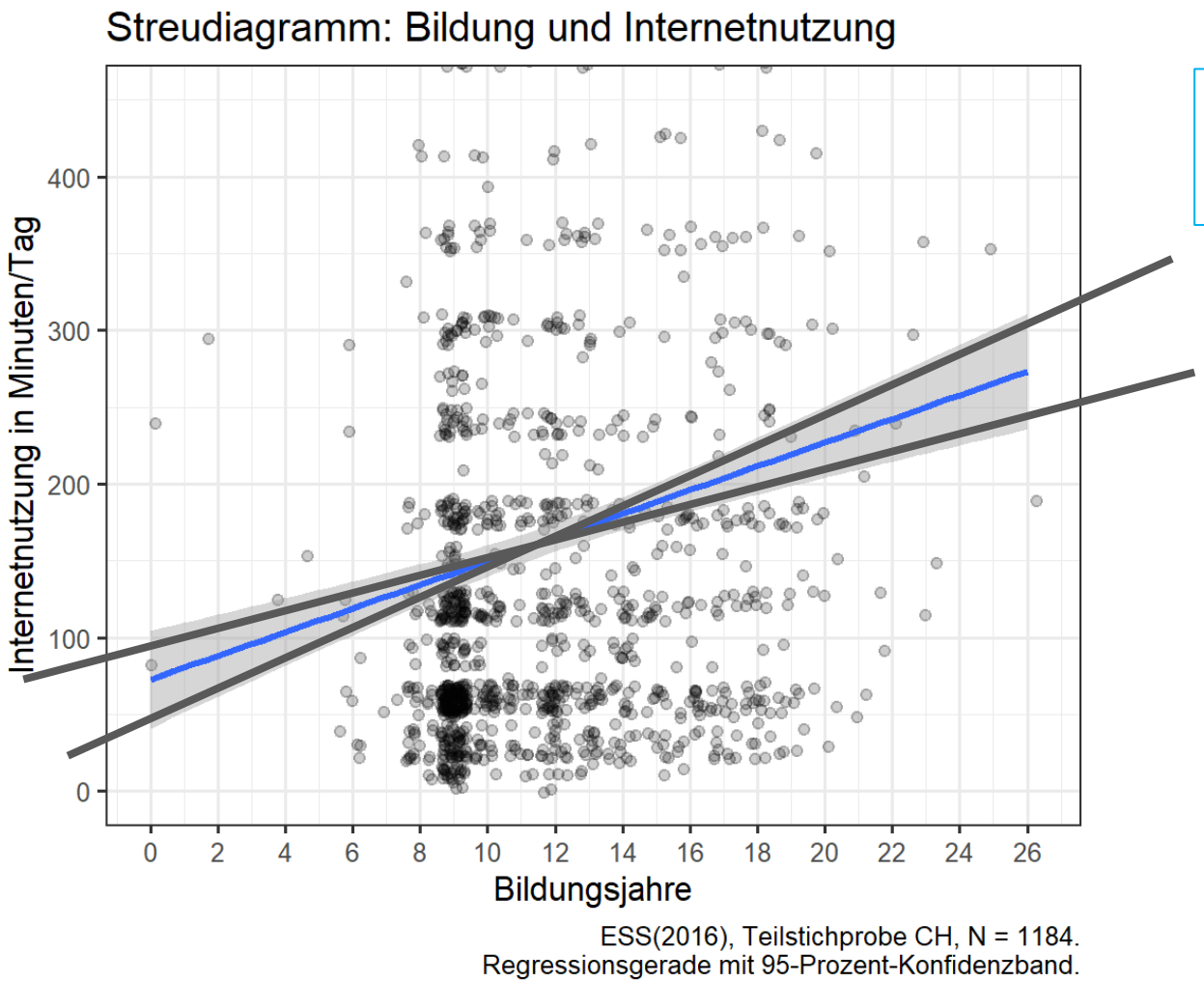
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    72.646     16.171   4.492 7.74e-06 ***
eduyrs          7.713      1.313   5.875 5.48e-09 ***
```

Die gleiche Information transportiert der p-Wert

- *Dieses Konfidenzband drückt also ein **hohes Vertrauen** darin aus, dass die **wahre Regressionsgerade** einen **positiven Steigungskoeffizienten** hat!*

3.3 Konfidenzband der Regressionsgerade

Aussagekraft des 95% Konfidenzbandes

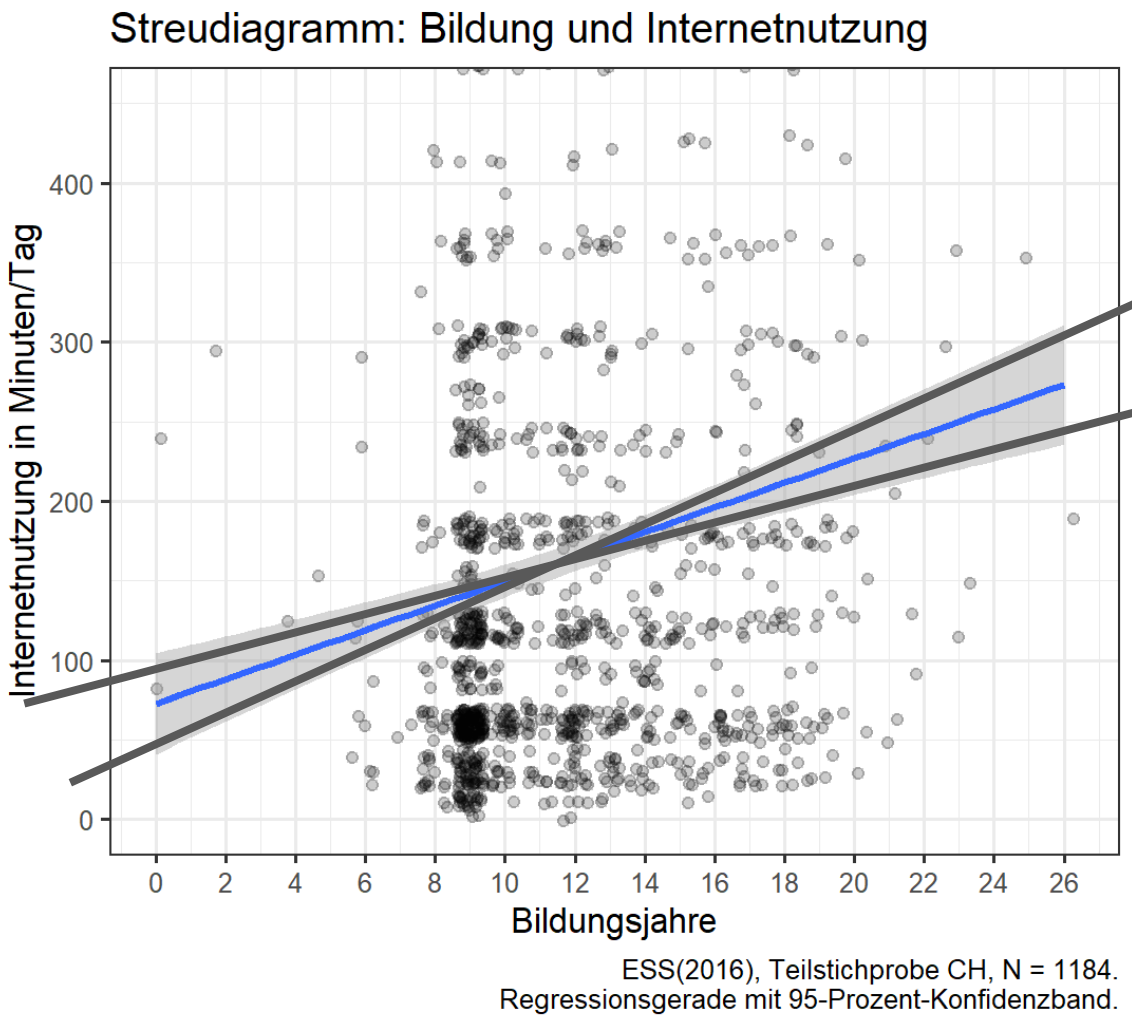


Welche inferenzstatischen Sachverhalte werden durch das Konfidenzband visualisiert?

- *Selbst eine Gerade mit Extremverlauf innerhalb des Konfidenzbandes weist eine deutlich positive Steigung auf.*
- ***Andererseits**: Die Steigungen der beiden Extremgraden innerhalb des Konfidenzbandes unterschieden sich deutlich voneinander: die Fächerform des Bandes ist ausgeprägt.*
- *Dieses Konfidenzband drückt also ein **nicht so hohes Vertrauen** darin aus, dass die **wahre Regressionsgerade** der ermittelten Regressionsgerade sehr ähnlich ist!*

3.3 Konfidenzband der Regressionsgerade

Aussagekraft des 95% Konfidenzbandes



ESS(2016), Teilstichprobe CH, N = 1184.
Regressionsgerade mit 95-Prozent-Konfidenzband.

Welche inferenzstatischen Sachverhalte werden durch das Konfidenzband visualisiert?

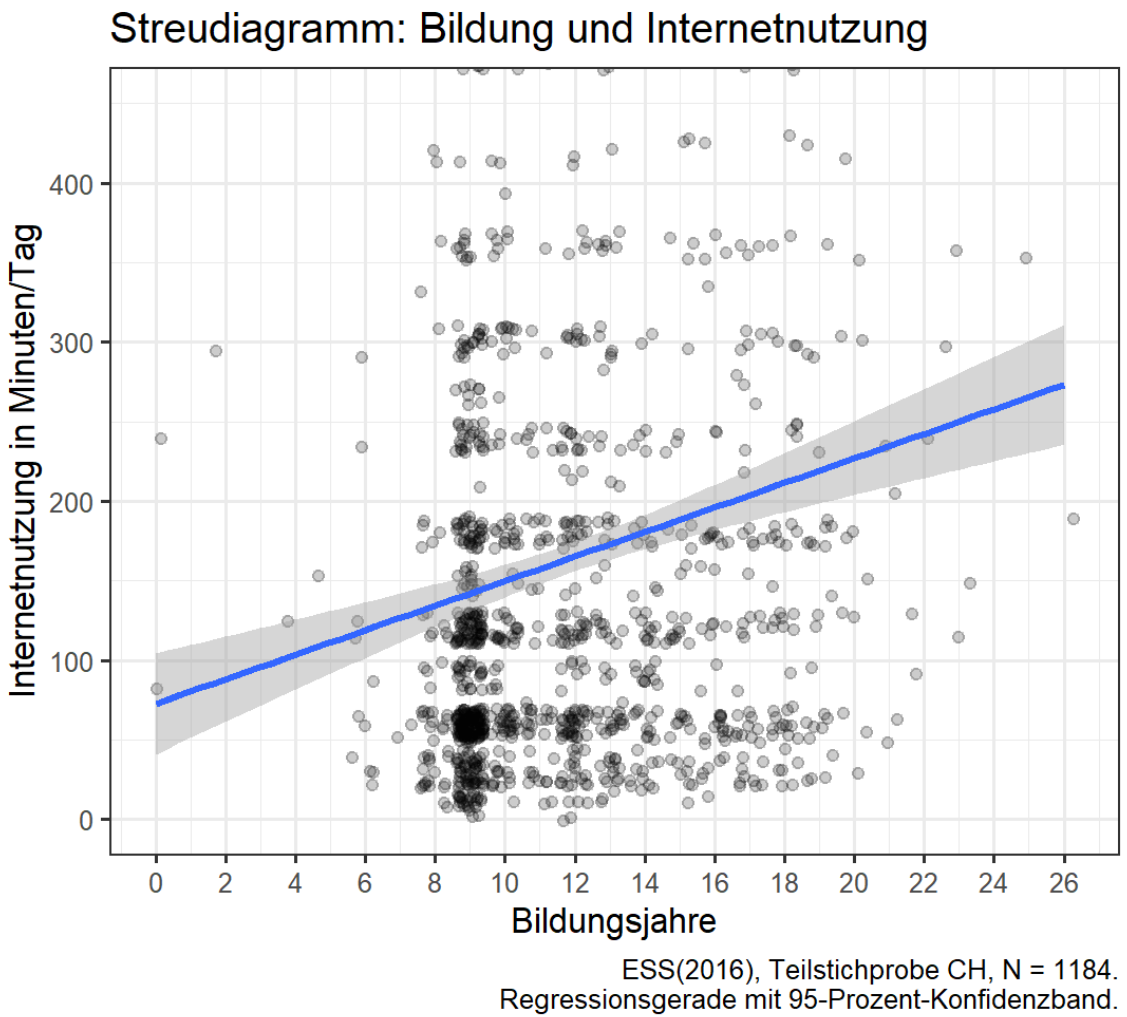
```
> confint(fit, level = 0.95)
                2.5 %    97.5 %
(Intercept) 40.918306 104.37282
eduyrs       5.137214  10.28828
```

Exakt die*selbe* Information transportiert das Konfidenzintervall des Koeffizienten

- *Dieses Konfidenzband drückt also ein **nicht so hohes Vertrauen** darin aus, dass die **wahre Regressionsgerade** der ermittelten Regressionsgerade sehr ähnlich ist!*

3.3 Konfidenzband der Regressionsgerade

Aussagekraft des 95% Konfidenzbandes



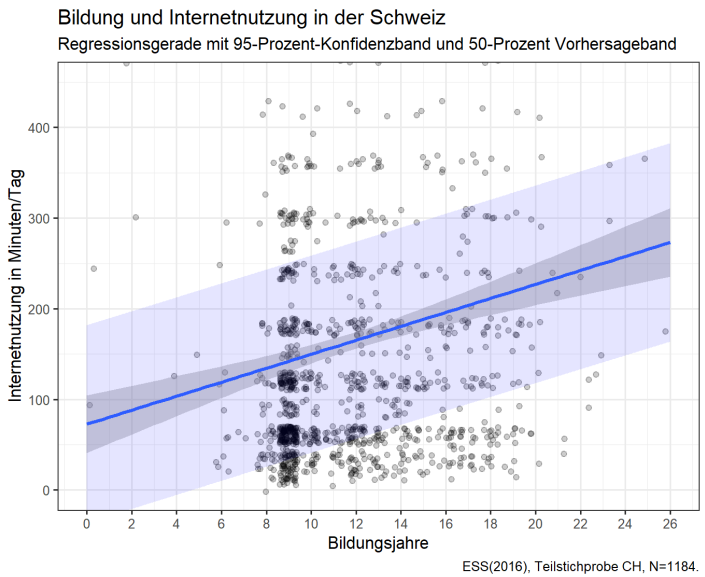
Welche inferenzstatischen Sachverhalte werden durch das Konfidenzband visualisiert?

Ein Konfidenzband stellt sowohl unser Vertrauen darin dar, dass bzw. ob

- **(a) sich der wahre Regressionskoeffizient in der Nähe des ermittelten findet (*hier eher mässig ausgeprägt*), und**
- **(b) der wahre Regressionskoeffizient von 0 verschieden ist (*hier stark ausgeprägt*).**

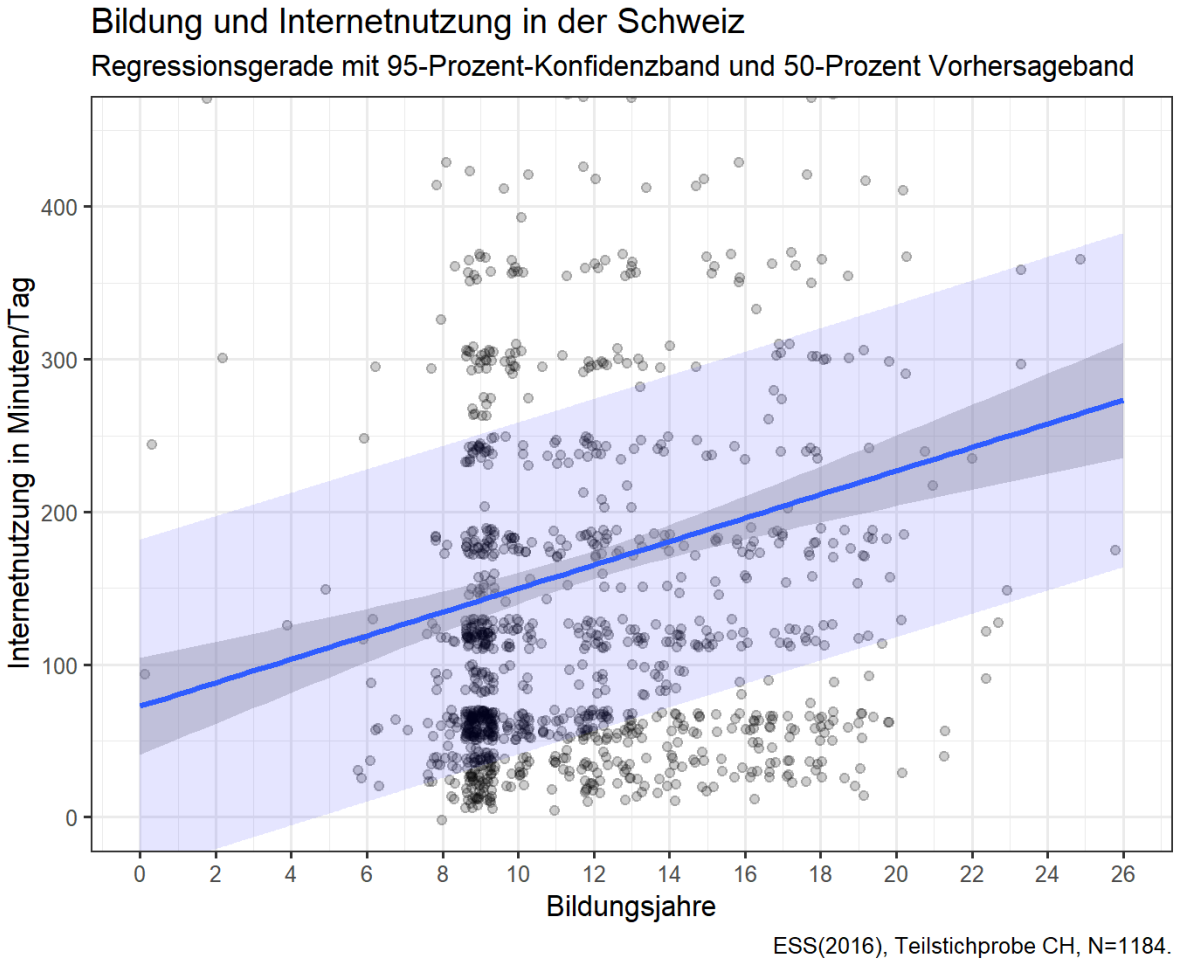
Es visualisiert somit die zentralen inferenzstatistischen Parameter der Regressionsanalyse.

3.4 Das Vorhersageband der Regressionsgerade



3.4 Vorhersageband der Regressionsgerade

Bedeutung des (hier) blau dargestellten Bereichs?

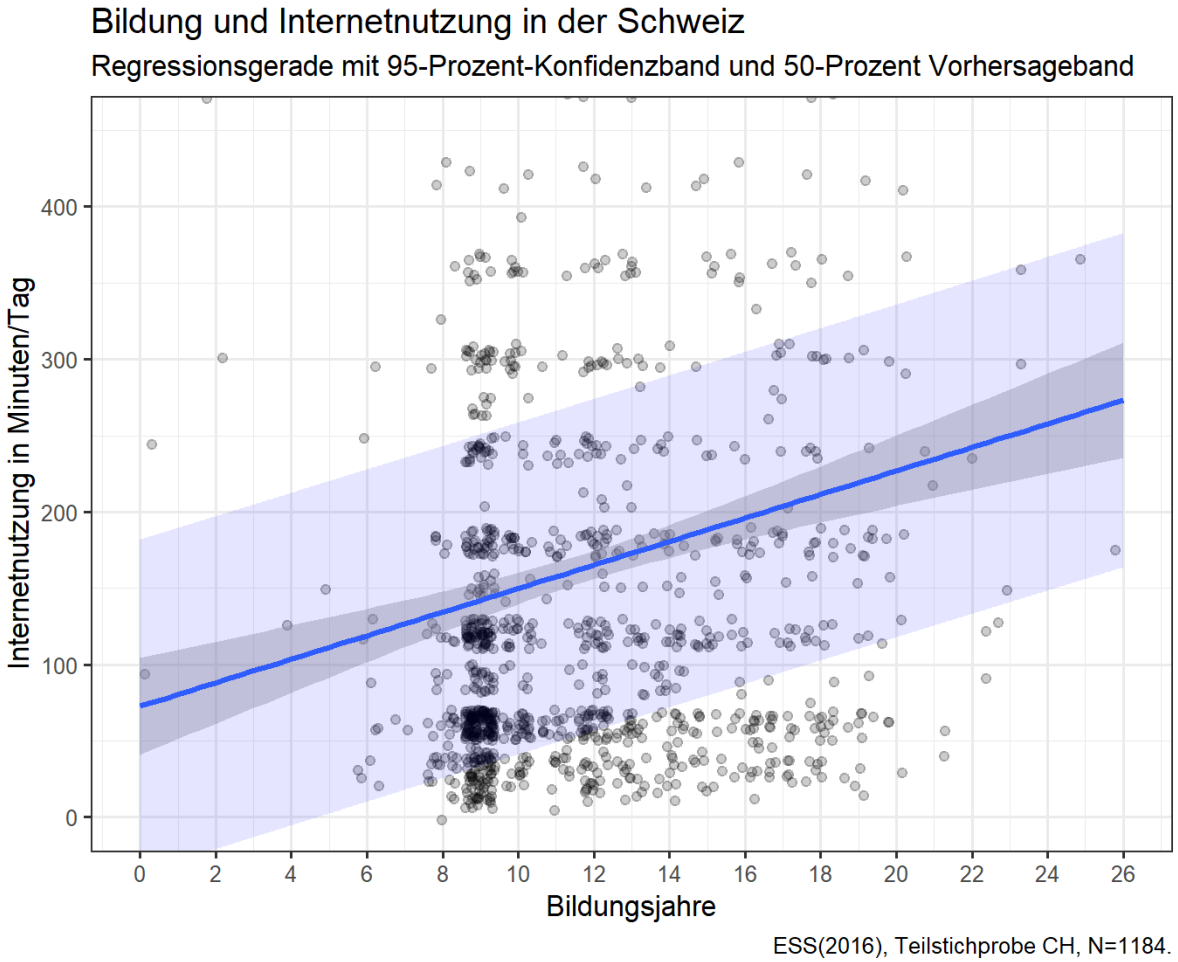


Das 50%-Vorhersageband zeigt den Bereich an, der 50% aller Werte enthält (in dem also ein weiterer Fall aus der Population mit 50% Sicherheit liegt).

Bei gleicher Sicherheitsstufe ist es immer deutlich breiter als das Konfidenzband und geht sogar oft in den irrealen Wertebereich. Daher, aber auch aus inhaltlichen Gründen und z.B. der konzeptionellen Nähe zum Interquartilabstand, ist es häufig sinnvoll, den Sicherheitswert hier niedriger (z.B. 50% oder 75%) anzusetzen.

3.4 Vorhersageband der Regressionsgerade

Das Vorhersageband verknüpft Vorhersageintervalle!




```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "prediction",
          ci_level = 0.50)
```

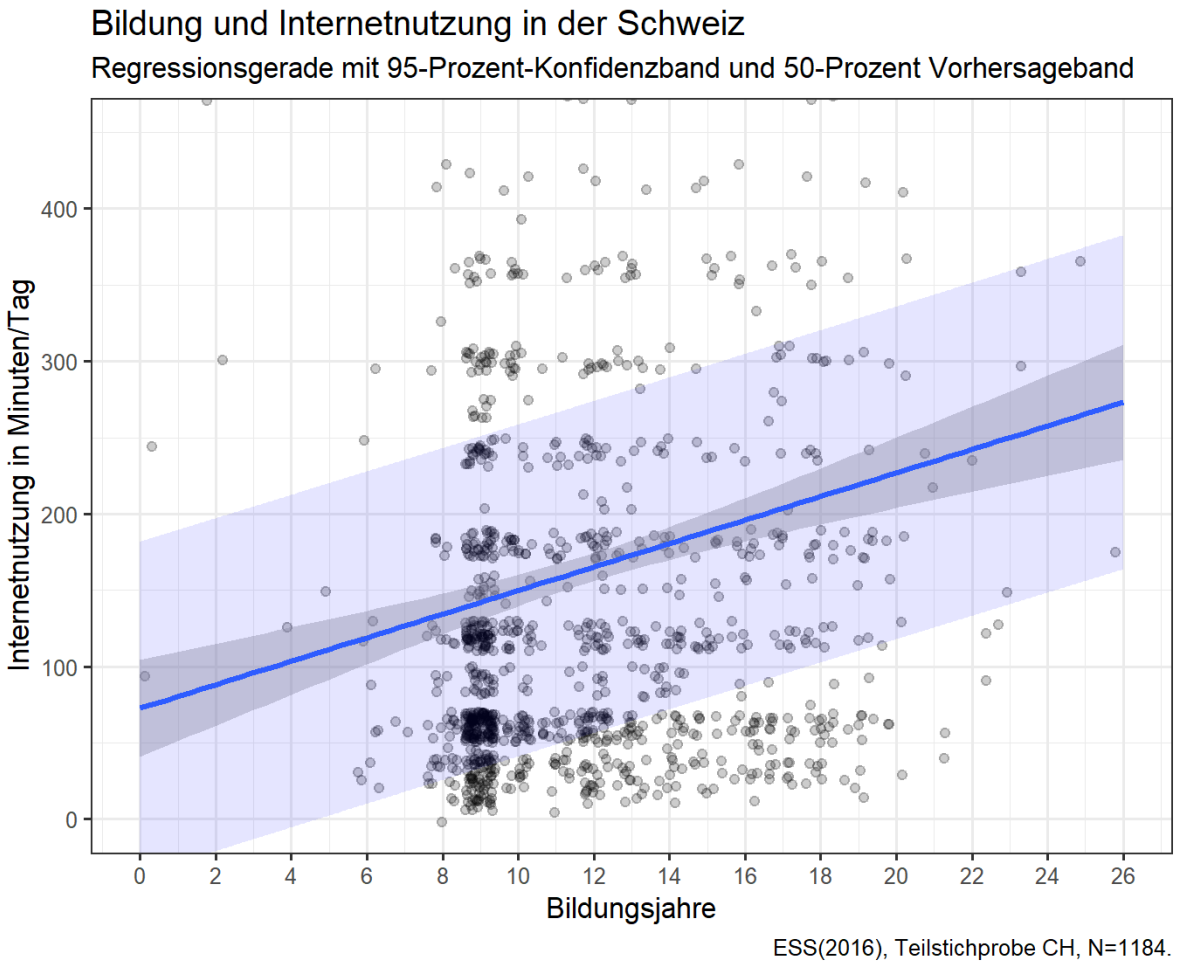
Ganz, ganz wichtig!

(Sonst werden Querschnitte/Intervalle des ***Konfidenz****bandes* ausgegeben)

Aktiviere den Befehl und erläutere die einzelnen Argumente

3.4 Vorhersageband der Regressionsgerade

Das Vorhersageband verknüpft Vorhersageintervalle!



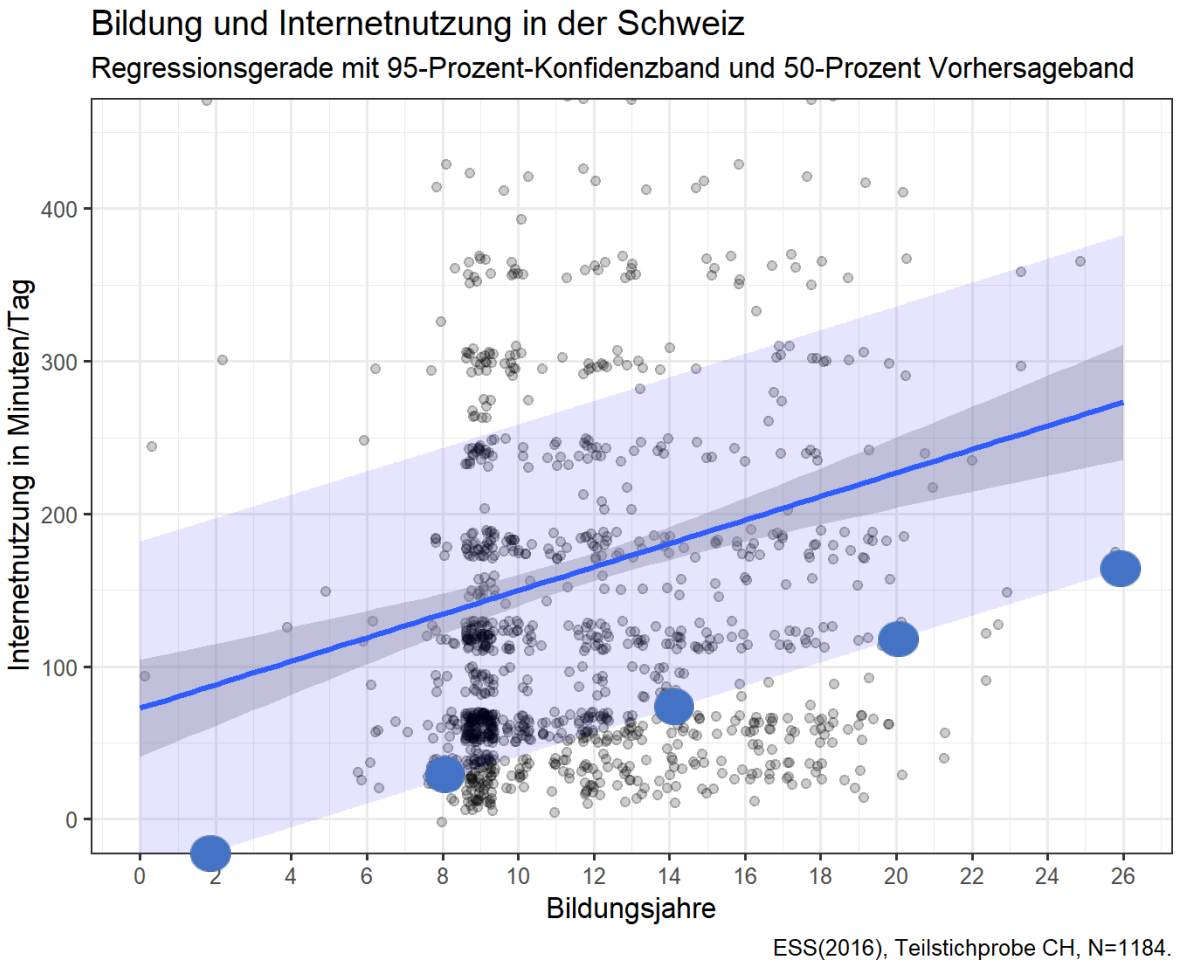

```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "prediction",
          ci_level = 0.50)
```

eduyrs	Predicted	50% CI
2	88.07	[-20.99, 197.13]
8	134.35	[25.58, 243.11]
14	180.62	[71.89, 289.36]
20	226.90	[117.94, 335.86]
26	273.18	[163.74, 382.62]

Was stellt hier die dritte Spalte des Outputs dar?

3.4 Vorhersageband der Regressionsgerade

Das Vorhersageband verknüpft Vorhersageintervalle!



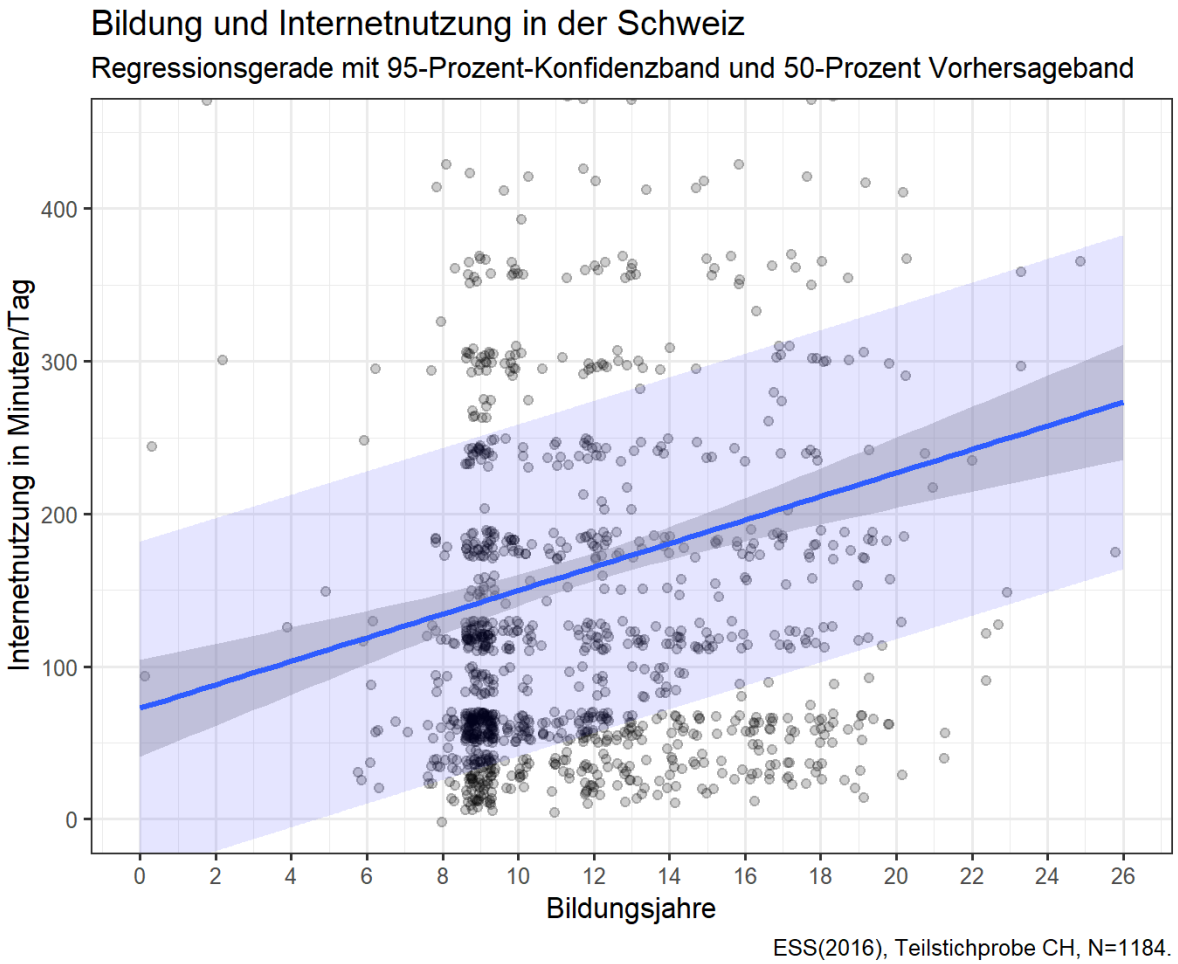

```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "prediction",
          ci_level = 0.50)
```

eduyrs	Predicted	50% CI
2	88.07	[-20.99, 197.13]
8	134.35	[25.58, 243.11]
14	180.62	[71.89, 289.36]
20	226.90	[117.94, 335.86]
26	273.18	[163.74, 382.62]

Unterer Grenzverlauf des 50%-Vorhersagebandes: Bei x=20 liegt die untere Grenze des 50%-Vorhersagebandes bei y=118

3.4 Vorhersageband der Regressionsgerade

Das Vorhersageband verknüpft Vorhersageintervalle!



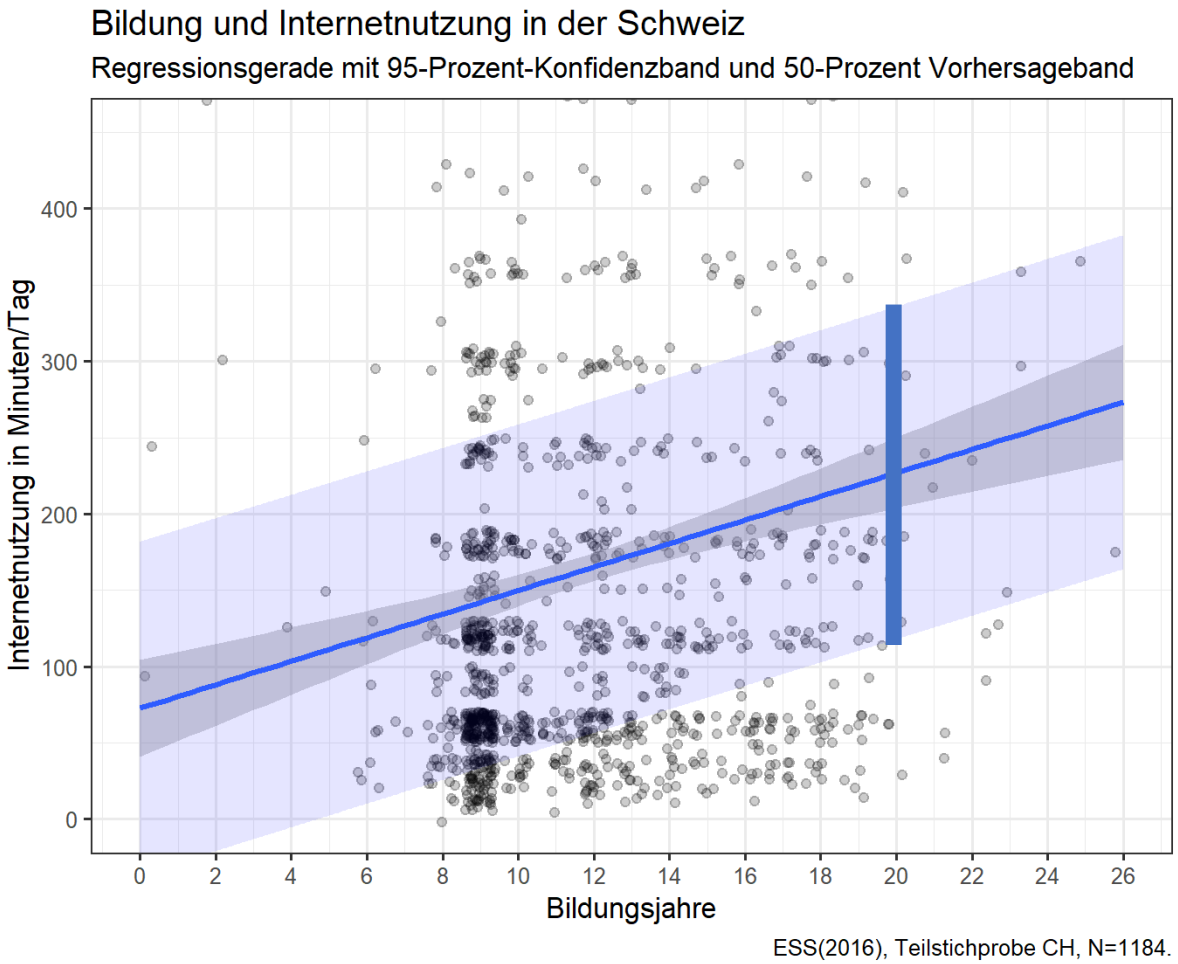

```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "prediction",
          ci_level = 0.50)
```

eduyrs	Predicted	50% CI
2	88.07	[-20.99, 197.13]
8	134.35	[25.58, 243.11]
14	180.62	[71.89, 289.36]
20	226.90	[117.94, 335.86]
26	273.18	[163.74, 382.62]

Interpretation einer Zeile des Outputs?

3.4 Vorhersageband der Regressionsgerade

Das Vorhersageband verknüpft Vorhersageintervalle!



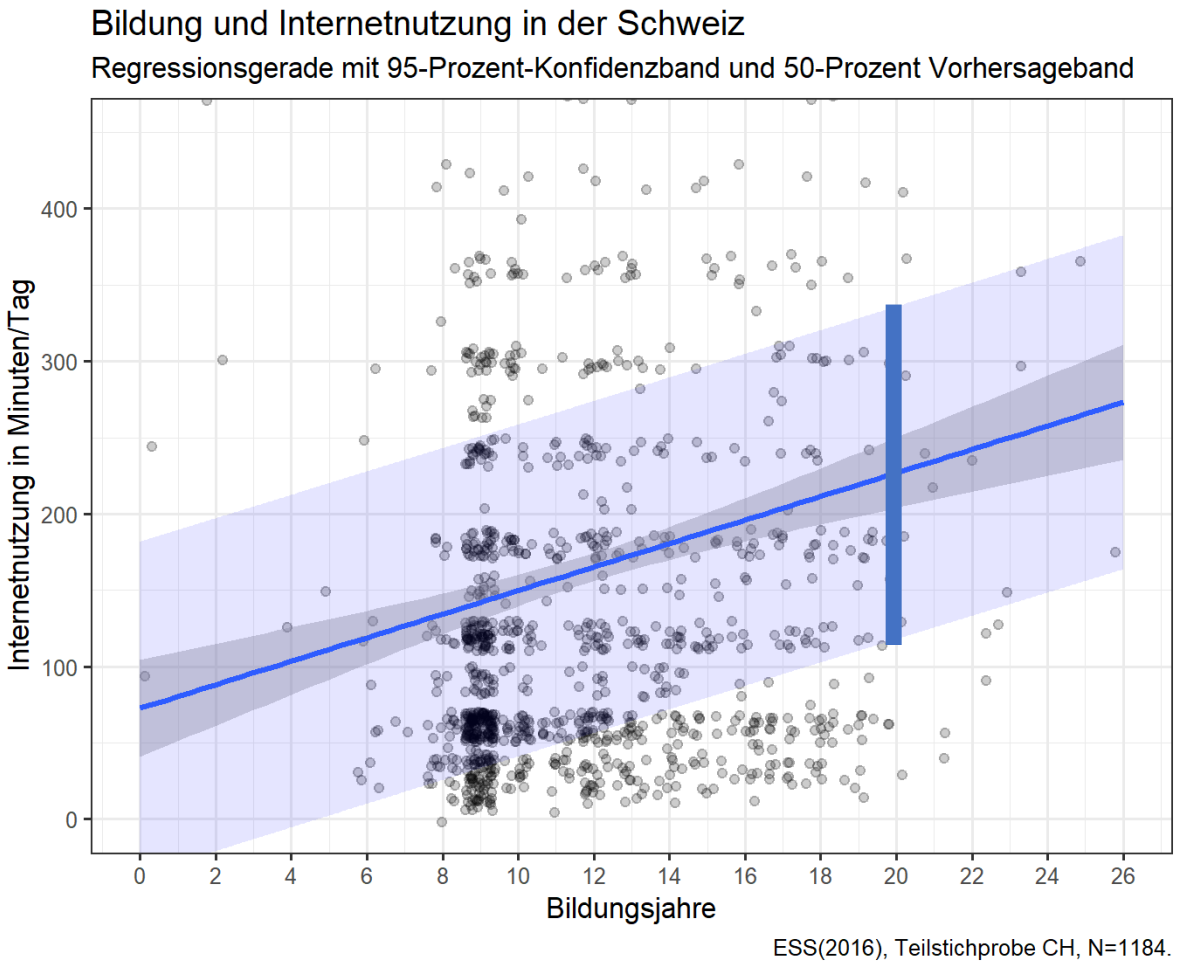

```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "prediction",
          ci_level = 0.50)
```

eduyrs	Predicted	50% CI
2	88.07	[-20.99, 197.13]
8	134.35	[25.58, 243.11]
14	180.62	[71.89, 289.36]
20	226.90	[117.94, 335.86]
26	273.18	[163.74, 382.62]

- ***Die Hälfte der Personen mit 20 Bildungsjahren verbringt täglich zwischen 118 und 336 Minuten im Internet. Oder:***
- ***Der tägliche Nutzungswert für eine Person mit 20 Bildungsjahren liegt mit 50% Sicherheit zwischen 118 und 336 Minuten.***

3.4 Vorhersageband der Regressionsgerade

Das Vorhersageband verknüpft Vorhersageintervalle!



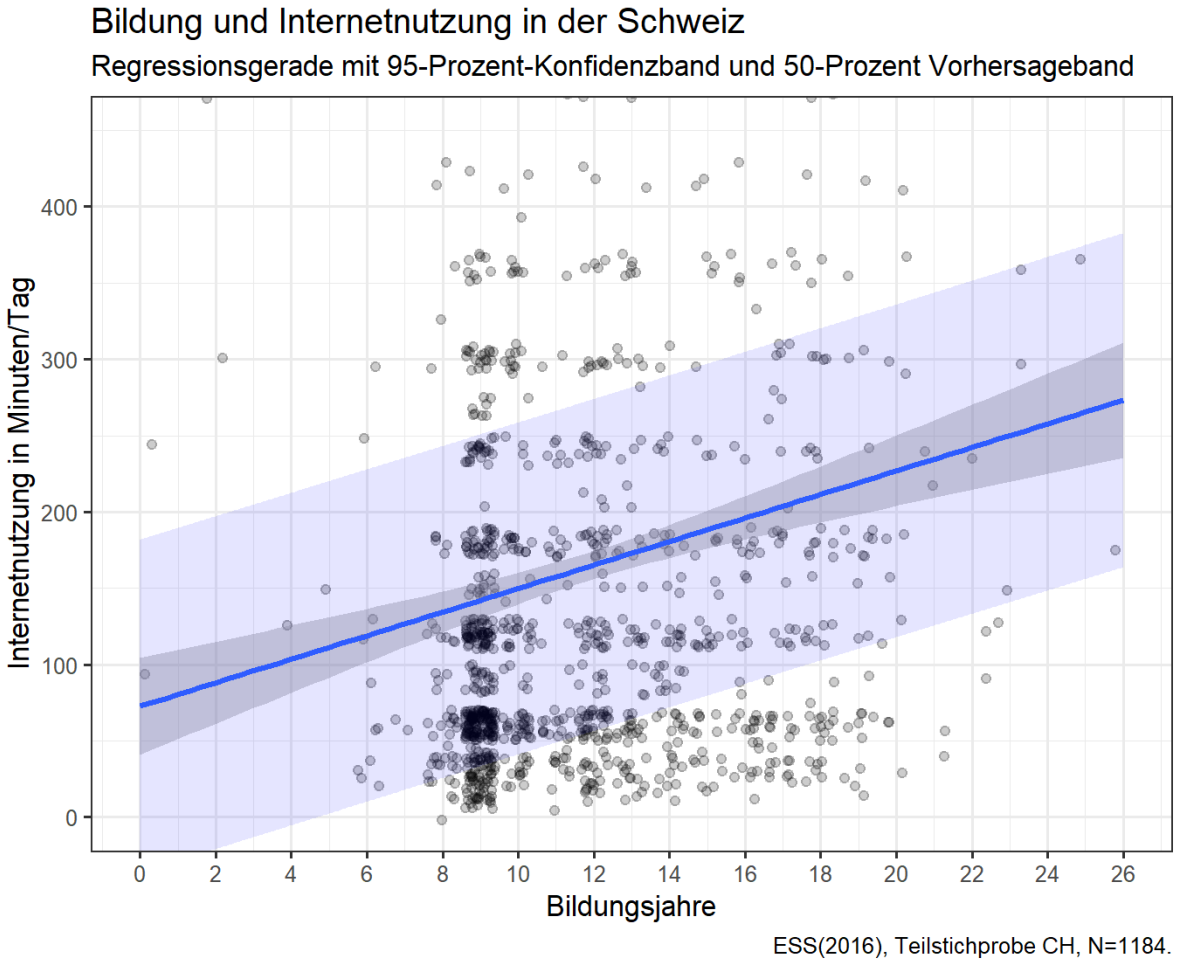

```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "prediction",
          ci_level = 0.50)
```

eduyrs	Predicted	50% CI
2	88.07	[-20.99, 197.13]
8	134.35	[25.58, 243.11]
14	180.62	[71.89, 289.36]
20	226.90	[117.94, 335.86]
26	273.18	[163.74, 382.62]

Frage: Wie würden sich die Grenzverläufe des **90% Vorhersagebandes** unterscheiden ?

3.4 Vorhersageband der Regressionsgerade

Darstellung des Vorhersagebandes im ggplot-Scatterplot

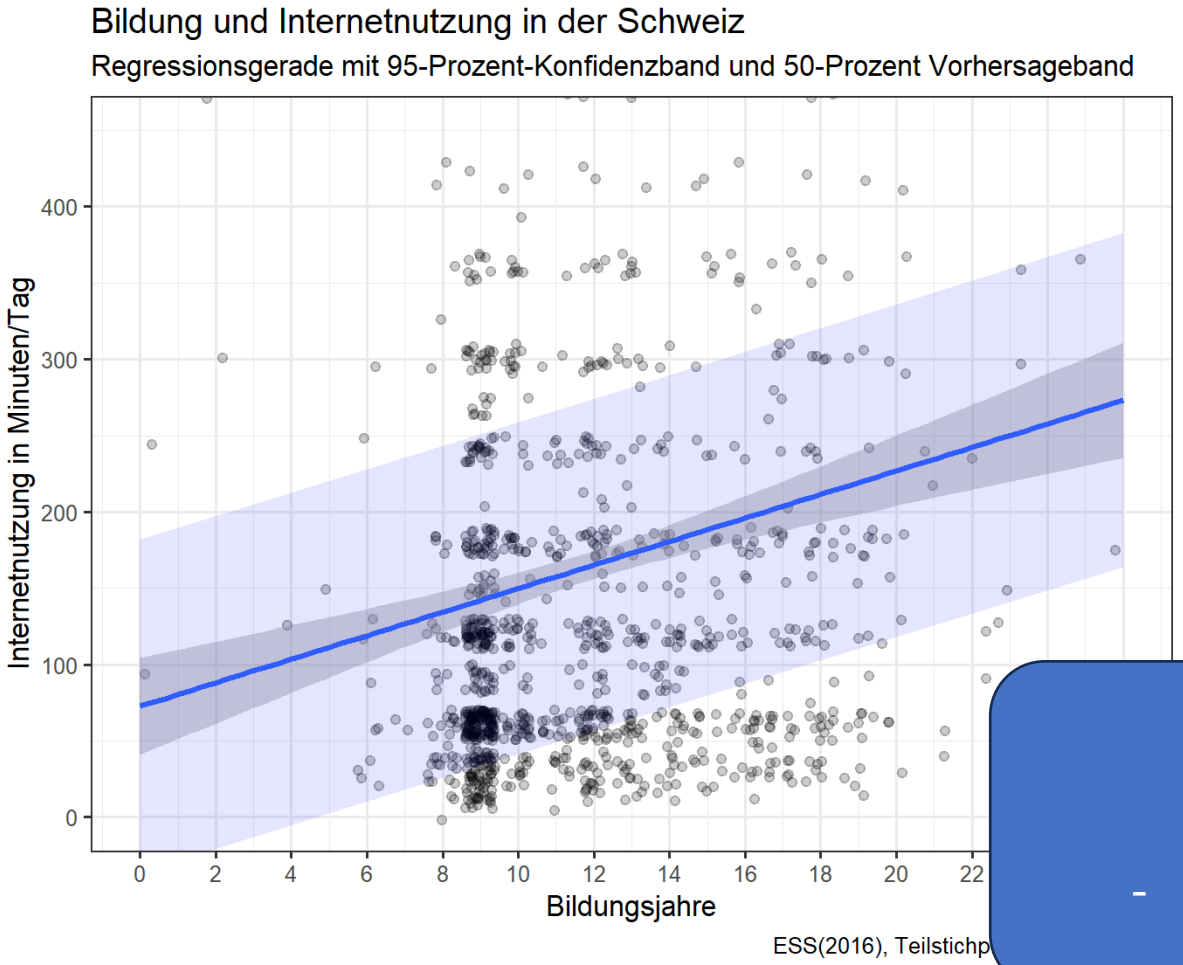


Bei Hypothesentests wird das Vorhersageband nur selten integriert. Integriert es auch in eure Prüfungspräsentation nur dann, wenn es für eure Argumentation wichtig ist! Entsprechend gibt es für das Vorhersageband auch **keine Standardoption** im Rahmen des ggplot.

…wir müssen also zunächst eine Datentabelle erstellen, welche Informationen zum Grenzverlauf des Vorhersagebandes enthält. Diese Aufgabe übernimmt wiederum **ggpredict()**

3.4 Vorhersageband der Regressionsgerade

Darstellung des Vorhersagebandes im ggplot-Scatterplot



…wir müssen also zunächst eine Datentabelle erstellen, welche Informationen zum Grenzverlauf des Vorhersagebandes enthält. Diese Aufgabe übernimmt wiederum **ggpredict()**

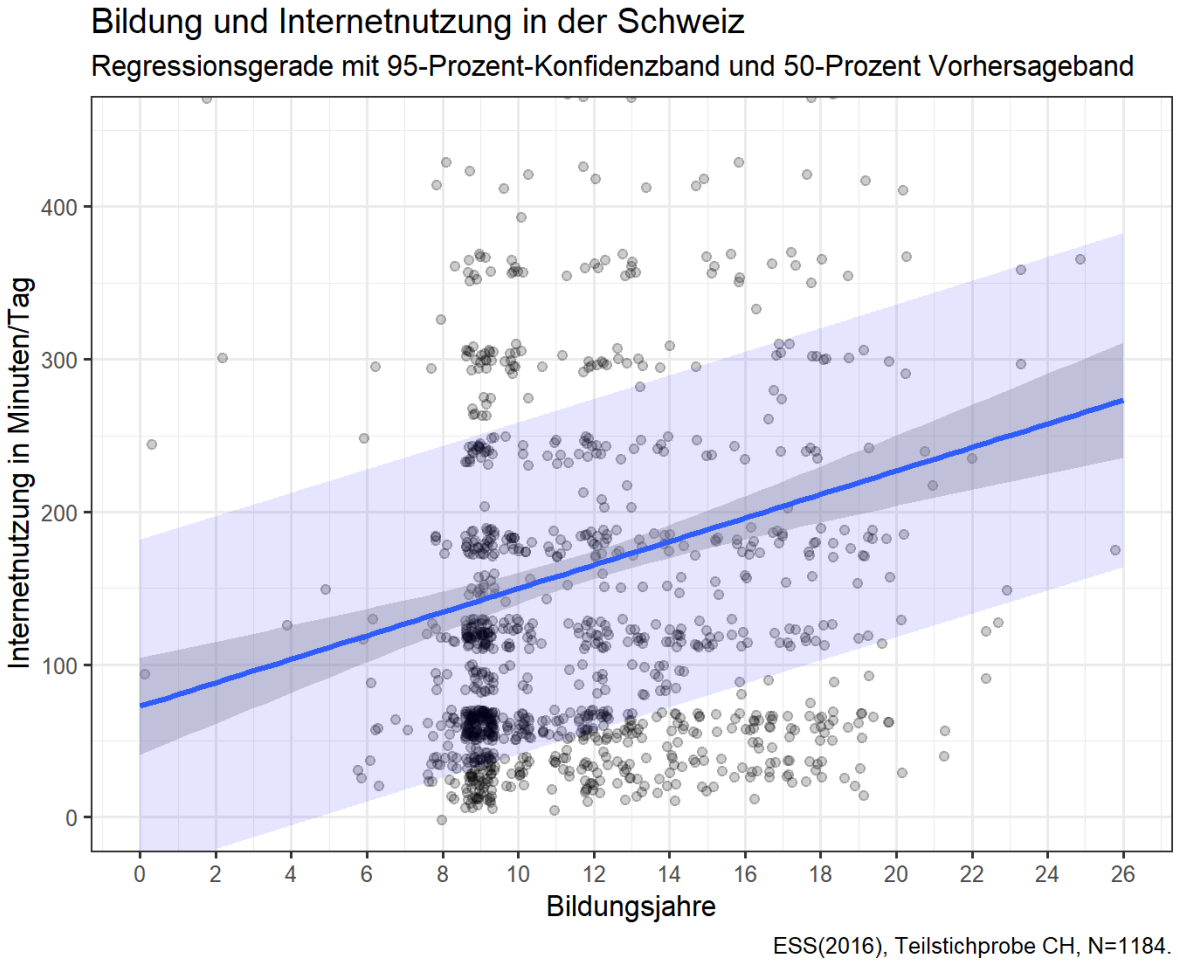
```
predictions <- ggpredict(fit,
                         terms = "eduyrs[0:26 by=0.05]",
                         interval = "prediction",
                         ci_level = 0.50)
```

Ähnlicher «ggpredict»-Befehl wie zuvor.

- Was aber ist der Unterschied?
- Eure Vermutung: Warum dieser Unterschied?
- Beschreibe das Objekt «predictions», dass der Befehl produziert

3.4 Vorhersageband der Regressionsgerade

Darstellung des Konfidenzbandes im ggplot-Scatterplot




```
predictions <- ggpredict(fit,
                         terms = "eduyrs[0:26 by=0.05]",
                         interval = "prediction",
                         ci_level = 0.50)
```

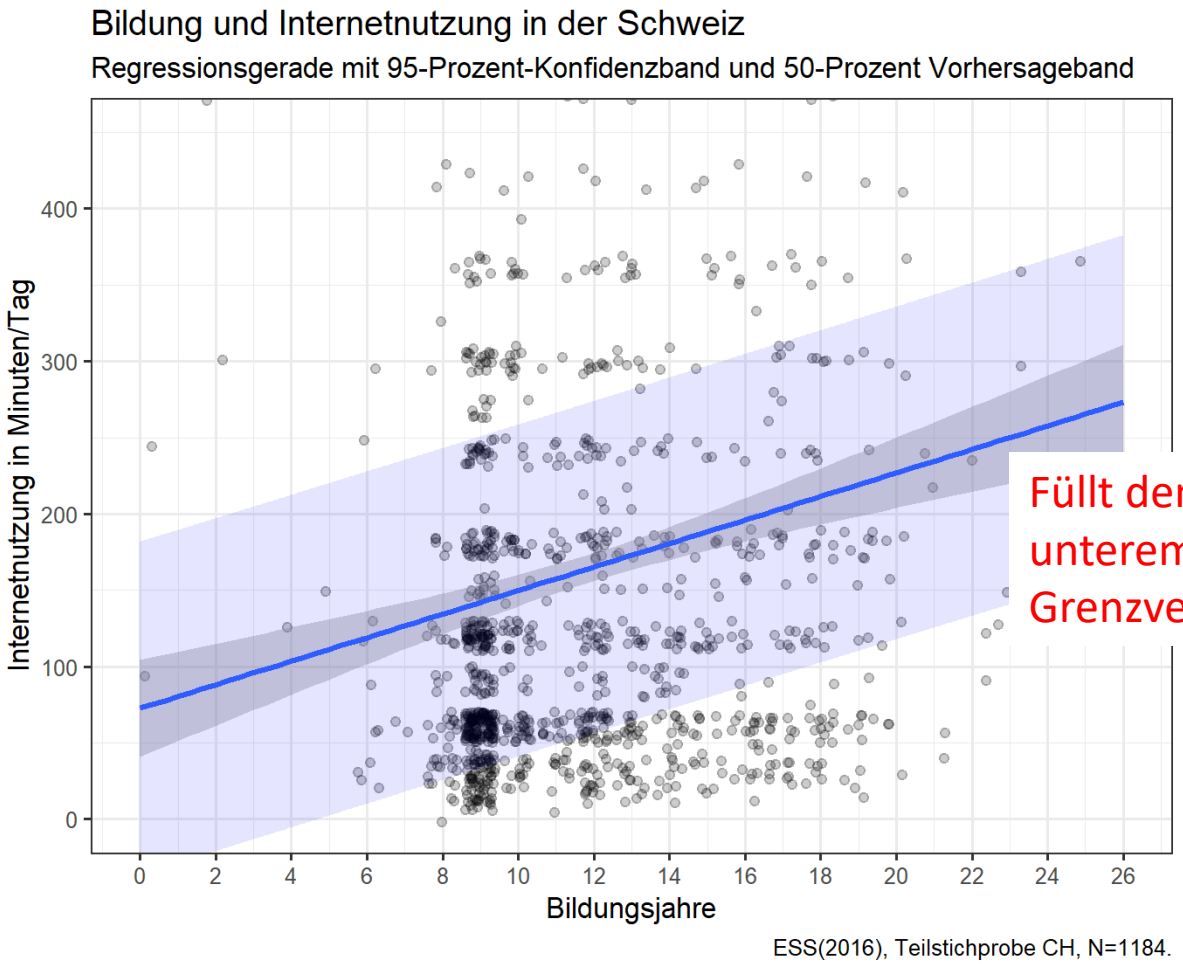
	x	predicted	std.error	conf.low	conf.high
1	0.00	72.64556	161.8715	-3.656870e+01	181.8598
2					2412
3					6227
4					0041
5					3856
6	0.25	74.57375	161.8404	-3.461956e+01	183.7671
7					86
8					01
9					16
10					32
11					47
12					53
13					79
14					96

Datentabelle mit Vorhersageintervallen

… im nächsten Schritt werden die Grenzwerte der Vorhersageintervalle im ggplot miteinander verknüpft. Daraus ergibt sich direkt der Grenzverlauf des Vorhersagebandes

3.4 Vorhersageband der Regressionsgerade

Darstellung des Konfidenzbandes im ggplot-Scatterplot




```
predictions <- ggpredict(fit,
                         terms = "eduyrs[0:26 by=0.05]",
                         interval = "prediction",
                         ci_level = 0.50)
```

Code entweder an den Basis-ggplot anfügen oder Basis-ggplot als Objekt «plot1» anlegen.

Füllt den Bereich zischen unterem und oberen Grenzverlauf aus

```
plot2 <- plot1 +
  geom_ribbon(data = predictions,
              aes(x = x,
                  ymin = conf.low,
                  ymax = conf.high),
              fill = "blue",
              alpha = 0.1,
              inherit.aes = FALSE) +
  labs(title = "Bildung und Internetnutzung ",
       subtitle = "Regressionsgerade mit...",
       y = "Internetnutzung in Minuten/Tag",
       x = "Bildungsjahre",
       caption = "ESS(2016),CH, N=1184.")
```

Hausaufgabe mit Selbstüberprüfung:

https://www.suz.uzh.ch/dataforstat/statistik2/infueb.html