Statistik 2 – Tutorate
Sitzung 5: Inferenzstatistik

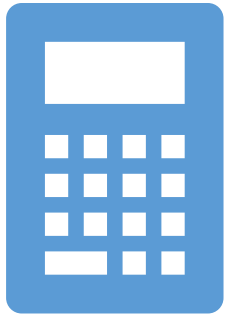
Marco Giesselmann, Rémy Blum, Federica Bruno, Rebecca Hobel, Kristina Trajkovic

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 Internetnutzung zusammen?

Vermutung: Personen mit mehr Bildung nutzen das Internet zu vielfältigeren Zwecken und somit insgesamt länger.

Wie lauten prüfbare Hypothese 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 überprüfen

Hypothese 1 (H1): *Bildung hat einen positiven Einfluss auf den zeitlichen Umfang der Internetnutzung*

Regressionsanalyse:

1. Berechnung des Regressionskoeffizienten und weiterer relevanter Kennwerte (p-Wert, Standardfehler)
2. Durch eine geeignete Interpretation und Visualisierung des Koeffizienten versuchen wir zu klären, ob dieser eine ***inhaltlich substanzielle Bedeutung*** aufweist.
3. Zudem prüfen wir, ob der Koeffizient von ***statistischer Bedeutsamkeit*** ist (über den Grad an Überzufälligkeit, ausgedrückt im p-Wert).
4. 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** 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_ss$internet, na.rm = TRUE)
[1] 163.1974
```

?

Die typische Abweichung von Mittelwert der Internetnutzung beträgt 163 Minuten

2.2 Regressionsanalyse

```
fit <- lm(internet ~ eduyrs, data = ess8_noNA)
summary(fit)
```

```
Call:
lm(formula = internet ~ eduyrs, data = ess8_noNA)

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
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.

p-Wert=0.00000000548

- Interpretation Regressionskoeffizient?
- Bedeutung Standardabweichung vs. Standardfehler?

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

Sowohl der Standardfehler als auch Standardabweichung messen Variation, beziehen sich allerdings auf unterschiedliche Ebenen und unterschiedliche statistische Grössen: Die Standardabweichung misst Variation der **abhängigen Variable *innerhalb*** der Stichprobe, der Standardfehler misst Variation des **Regressionskoeffizienten *zwischen*** verschiedenen Stichproben.

2.3 Standardfehler

```
Call:
lm(formula = internet ~ eduyrs, data = ess8_noNA)

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
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.

2.4 t-Wert

```
Call:
lm(formula = internet ~ eduyrs, data = ess8_noNA)

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
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.

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

Er misst die Grösse des Koeffizienten in Einheiten des Standardfehlers

2.5 p-Wert

```
Call:
lm(formula = internet ~ eduyrs, data = ess8_noNA)

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
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_noNA)

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
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_noNA)

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
Multiple R-squared:  0.02838,   Adjusted R-squared:  0.02755
F-statistic: 34.52 on 1 and 1182 DF,  p-value: 5.48e-09
```

Frage:
Was können wir hier ausgehend vom p-Wert zur Nullhypothese sagen?

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

Unsere Forschungshypothese wird gestützt.

Auch wenn die von der R postulierte Forschungshypothese *zweiseitig* ist, können wir den abgeleiteten 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 Praktische Übung

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

- Formuliert Forschungs- und Nullhypothese.
- Interpretiert SE und p-Wert.
- 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
```

Konfidenz- und Vorhersageintervalle (bzw –bänder)

Parameterkonfidenz

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

→

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

Vorhersagekonfidenz

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

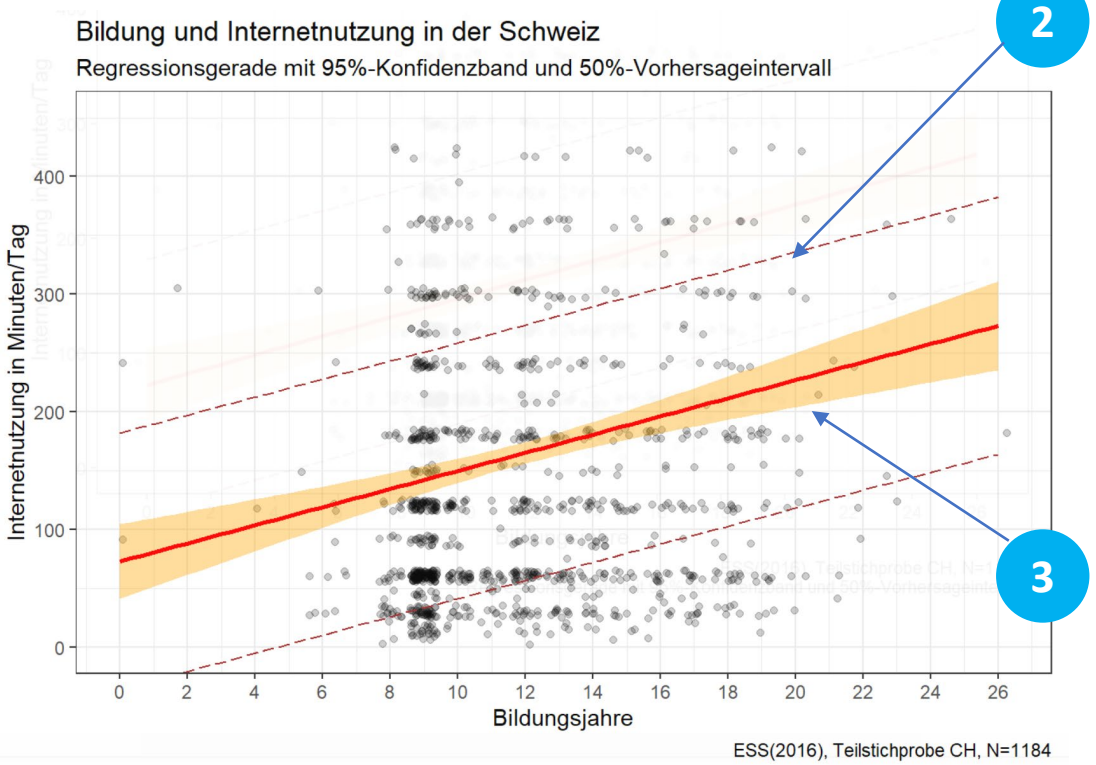
→

- **Vorhersageband der Regressionsgerade**
- **Regressionsbasiertes Vorhersageintervall**

3.1 Beispiel: Bildungsjahre → Internetnutzung (Minuten/Tag)

```
fit <- lm(internet ~ eduyrs, data = ess8_noNA)

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



Ergänzt wurden hier das **Konfidenzband**, das **Konfidenzintervall des Koeffizienten** und das **Vorhersageband**.

Frage: Wo sind diese drei Werte im Output zu finden?

Antwort:

1 95%-Konfidenzintervall des Koeffizienten
2 50%-Vorhersageband
3 95%-Konfidenzband

3.2 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)*

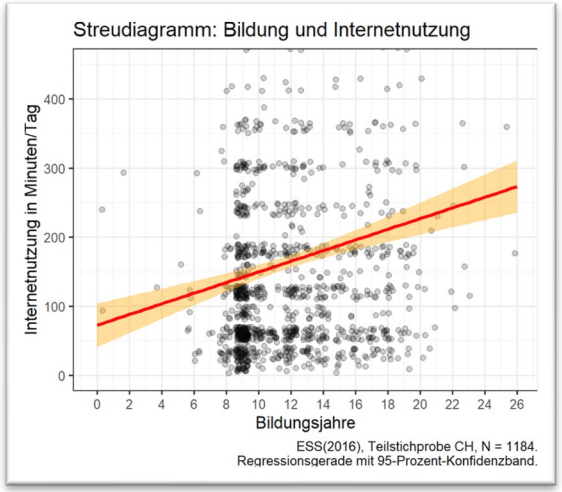
```
Coefficients:
             Estimate Std. Error
(Intercept)   72.646     16.171
eduyrs         7.713      1.313
---
Signif. codes:  0 ‘***’ 0.001 ‘*
```

Konkrete Interpretation?

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

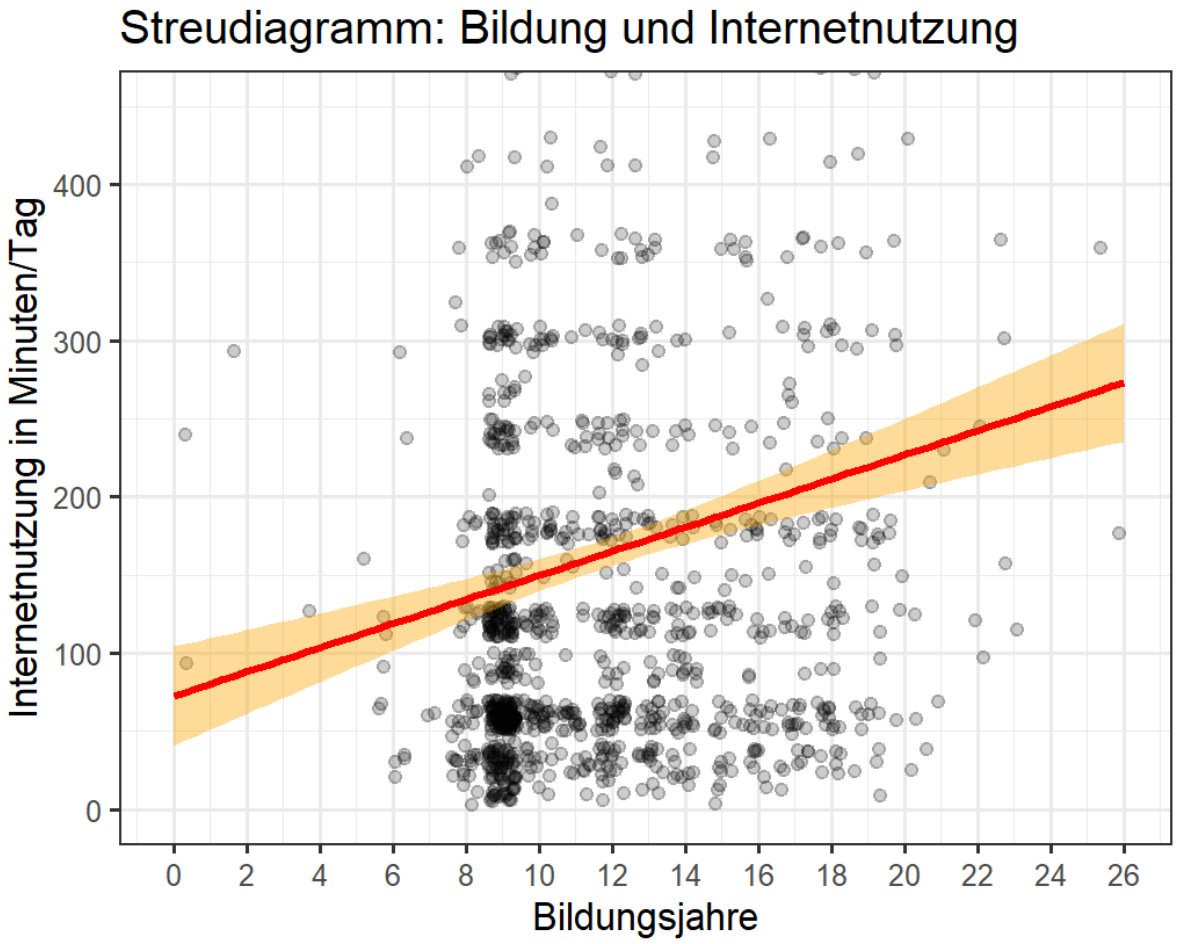
Mit 95% Sicherheit steigt mit jedem zusätzlichen Bildungsjahr die Dauer der durchschnittlichen Internetnutzung in der GG zwischen etwa 5 und 10 min

3.3 Das Konfidenzband der Regressionsgerade



3.3 Konfidenzband der Regressionsgerade

Bedeutung des (hier) orangenen Bereichs?

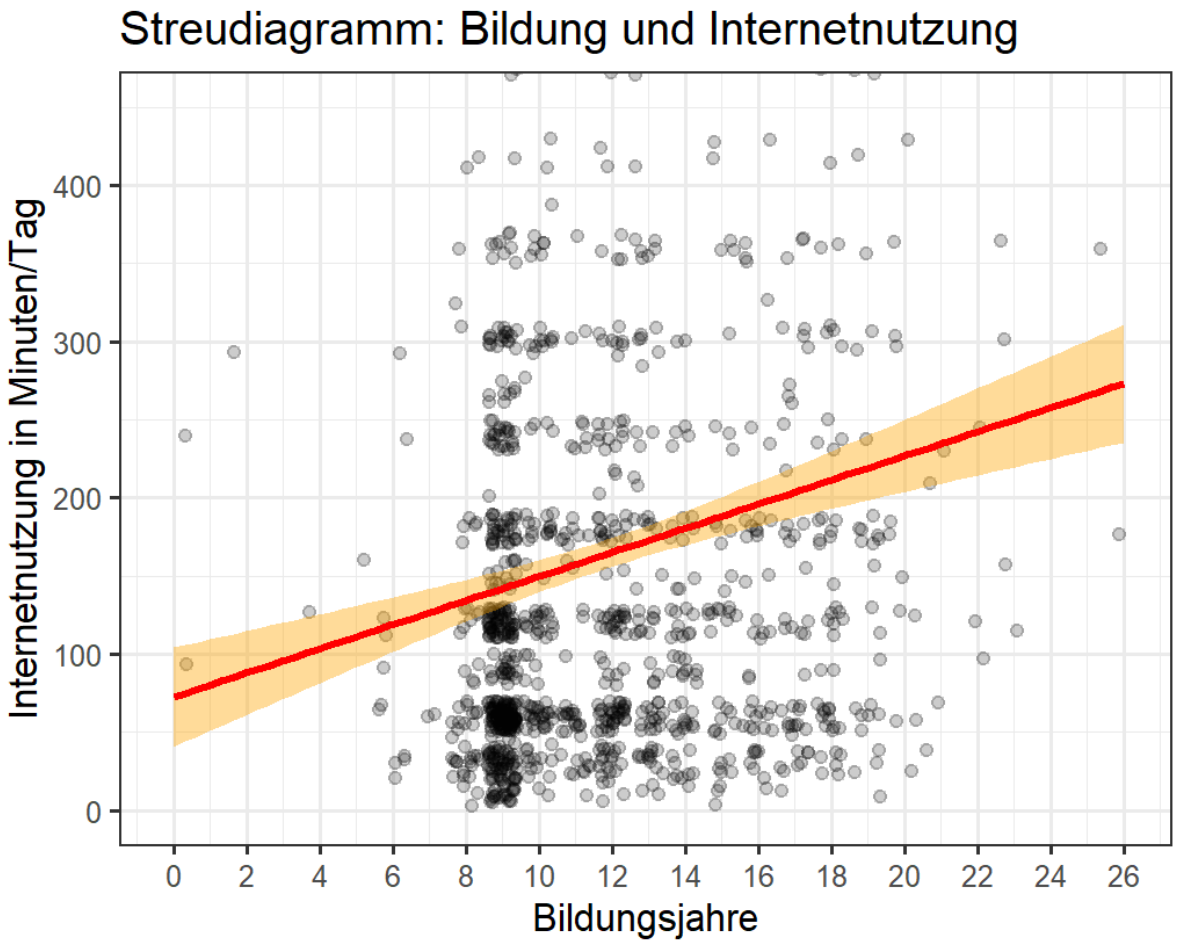


ESS(2016), Teilstichprobe CH, N = 1184.
Regressionsgerade mit 95-Prozent-Konfidenzband.

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

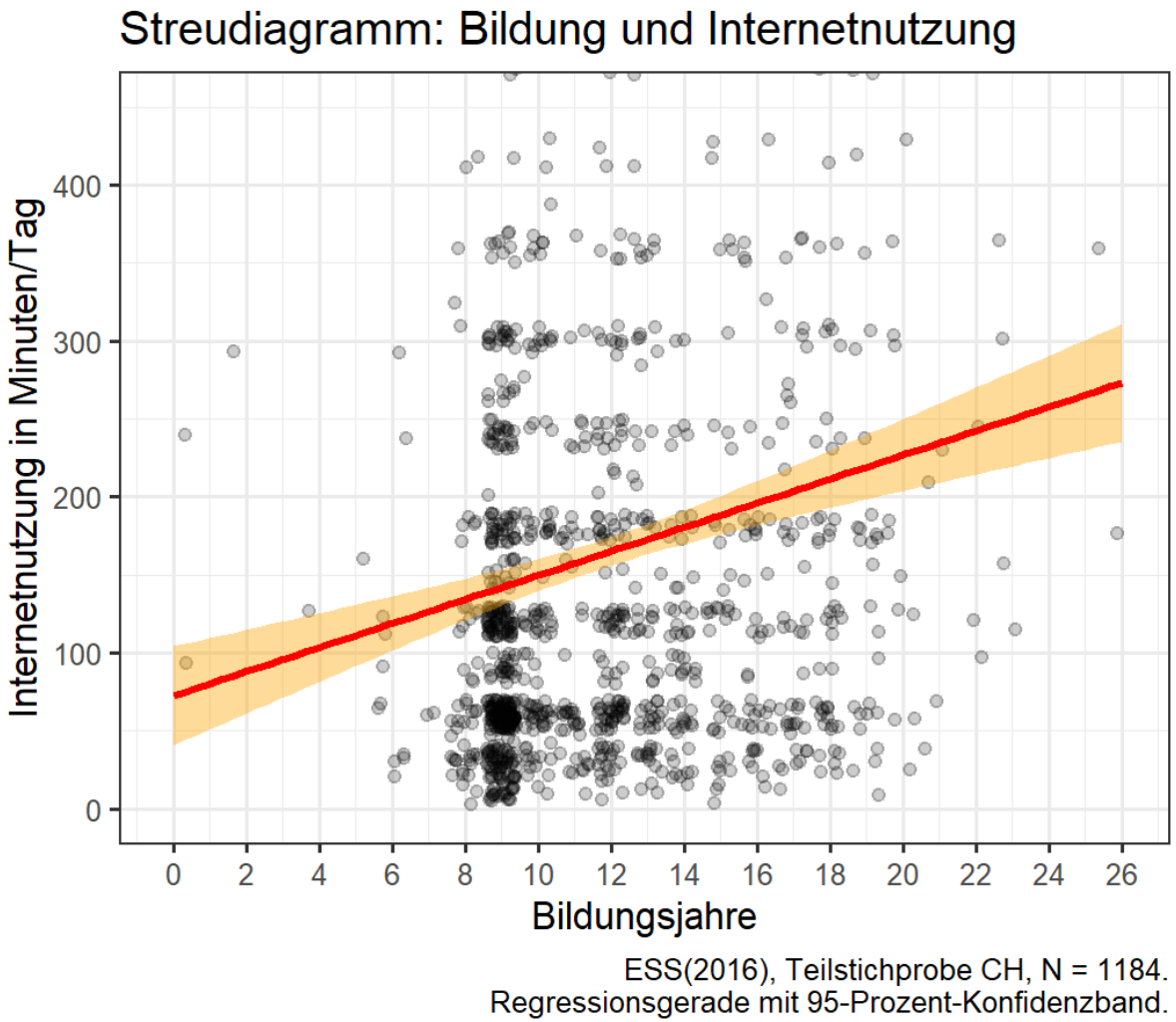



```
ggplot(ess8_CH, aes(x = eduyrs, y = internet))+
geom_jitter(alpha = 0.2, height = 10) +
scale_x_continuous(breaks = seq(0,26,2))+
coord_cartesian(ylim = c(0,450)) +
geom_smooth(method = "lm",
            se = TRUE,
            color = "red",
            fill = "orange",
            level = 0.95)+
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

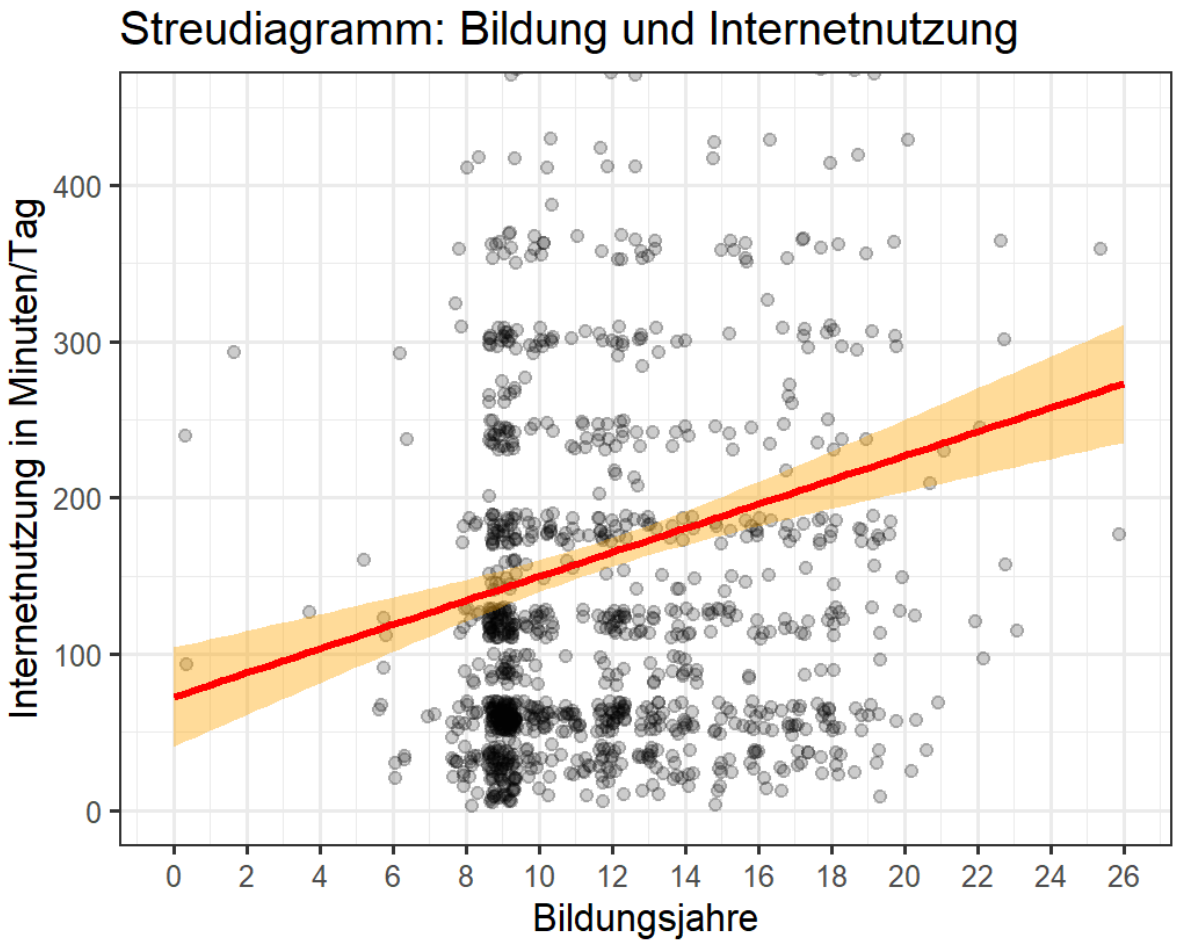



```
ggplot(ess8_CH, aes(x = eduyrs, y = internet))+
geom_jitter(alpha = 0.2, height = 10) +
scale_x_continuous(breaks = seq(0,26,2))+
coord_cartesian(ylim = c(0,450)) +
geom_smooth(method = "lm",
            se = TRUE,
            color = "red",
            fill = "orange",
            level = 0.95)+
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

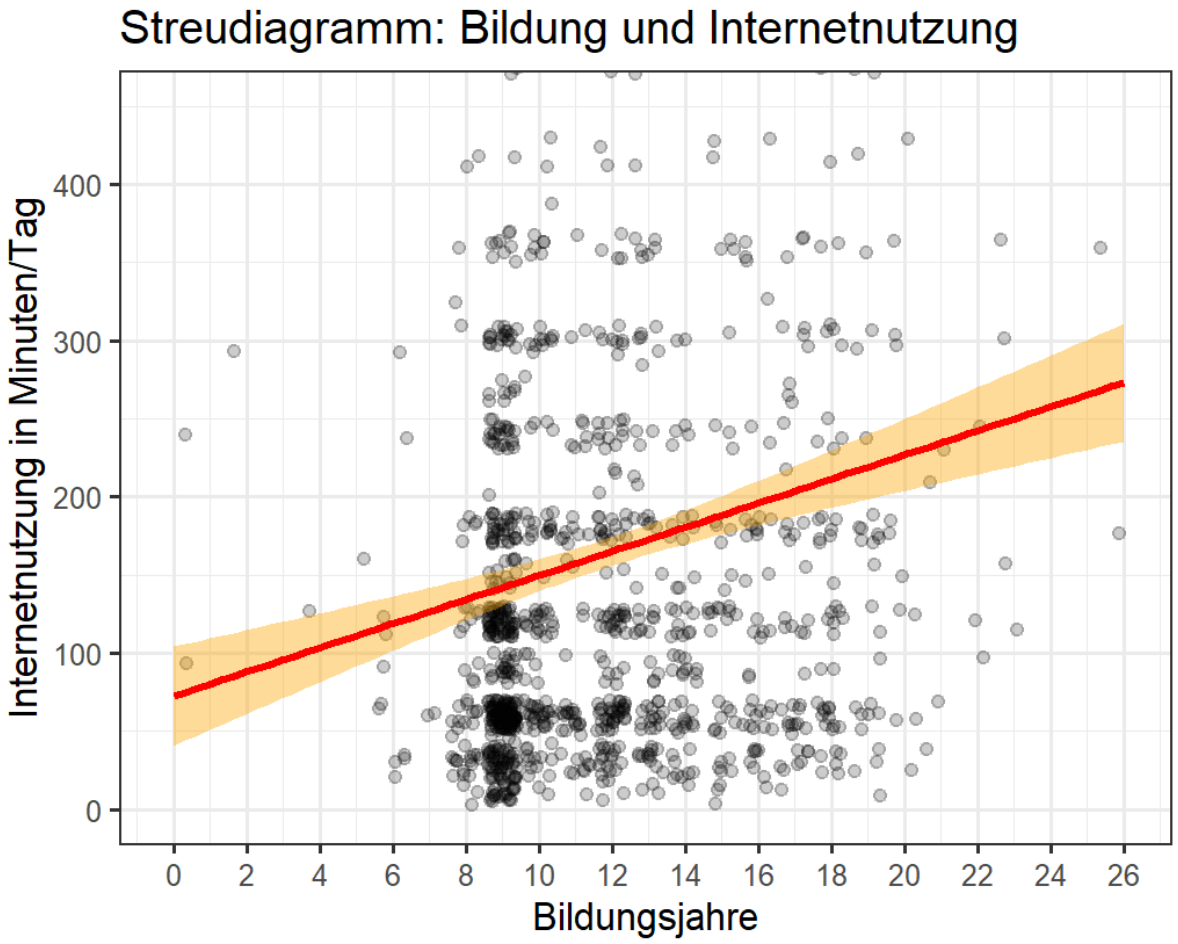



```
ggplot(ess8_CH, aes(x = eduyrs, y = internet))+
geom_jitter(alpha = 0.2, height = 10) +
scale_x_continuous(breaks = seq(0,26,2))+
coord_cartesian(ylim = c(0,450)) +
geom_smooth(method = "lm",
            se = TRUE,
            color = "red",
            fill = "orange",
            level = 0.95)+
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 alle blau umrandeten Befehlselemente weggelassen werden?

3.3 Konfidenzband der Regressionsgerade

Darstellung des Konfidenzbandes im ggplot-Scatterplot

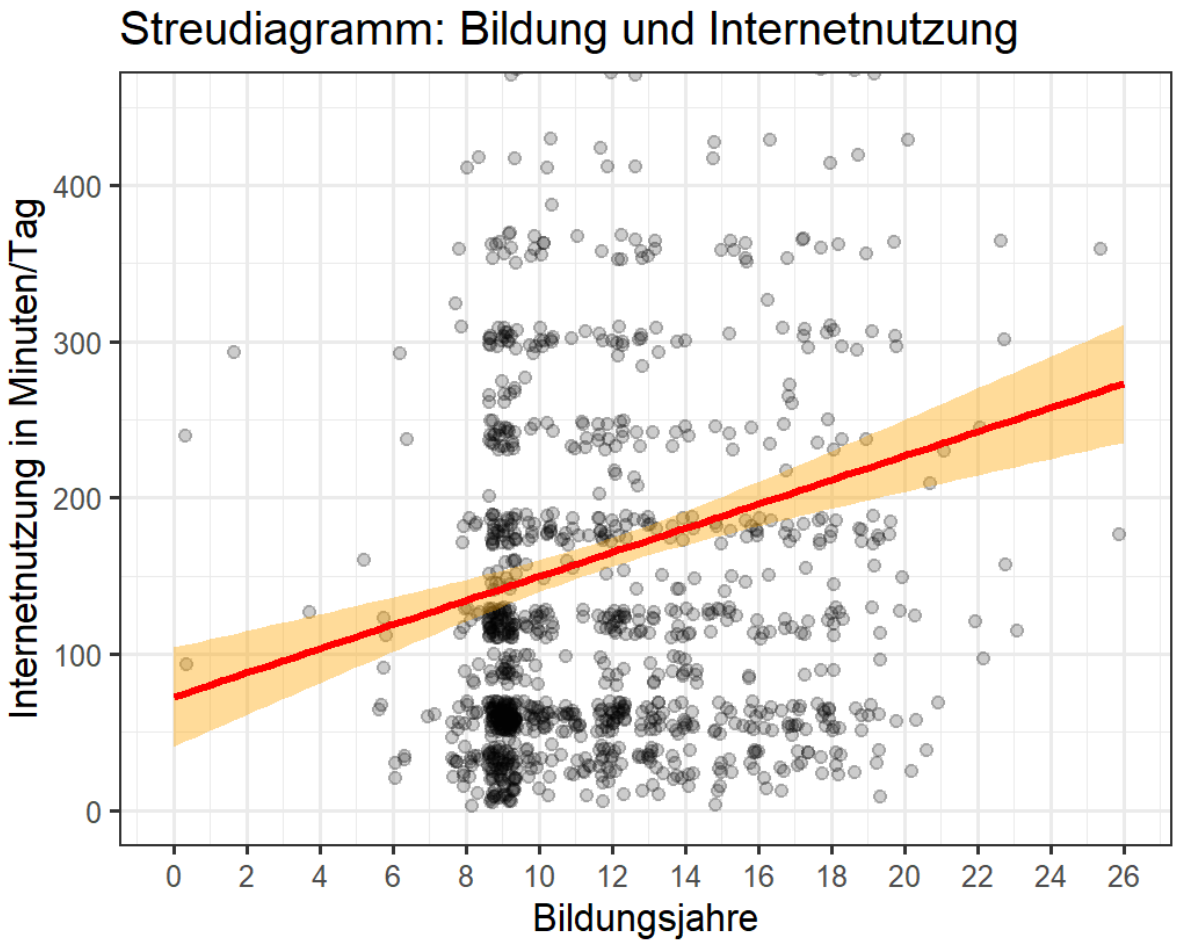



```
ggplot(ess8_CH, aes(x = eduyrs, y = internet))+
geom_jitter(alpha = 0.2, height = 10) +
scale_x_continuous(breaks = seq(0,26,2))+
coord_cartesian(ylim = c(0,450)) +
geom_smooth(method = "lm",
            se = TRUE,
            color = "red",
            fill = "orange",
            level = 0.95)+
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

Berechnung der Grenzen des Konfidenzbandes mit R



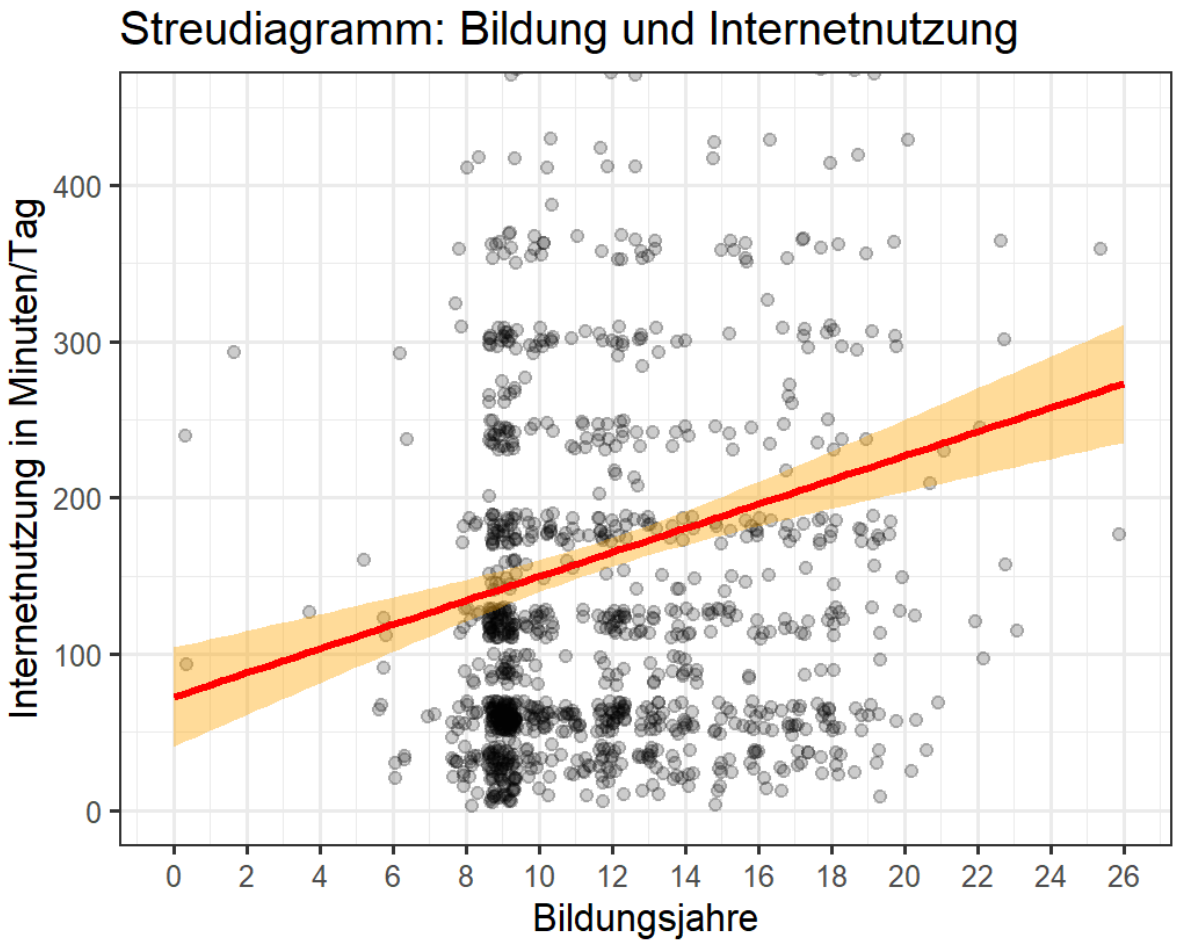
ESS(2016), Teilstichprobe CH, N = 1184.
Regressionsgerade mit 95-Prozent-Konfidenzband.

```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "confidence",
          ci.lvl = 0.95)
```

Aktiviere den Befehl und erläutere die einzelnen Argumente

3.3 Konfidenzband der Regressionsgerade

Berechnung der Grenzen des Konfidenzbandes mit R



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

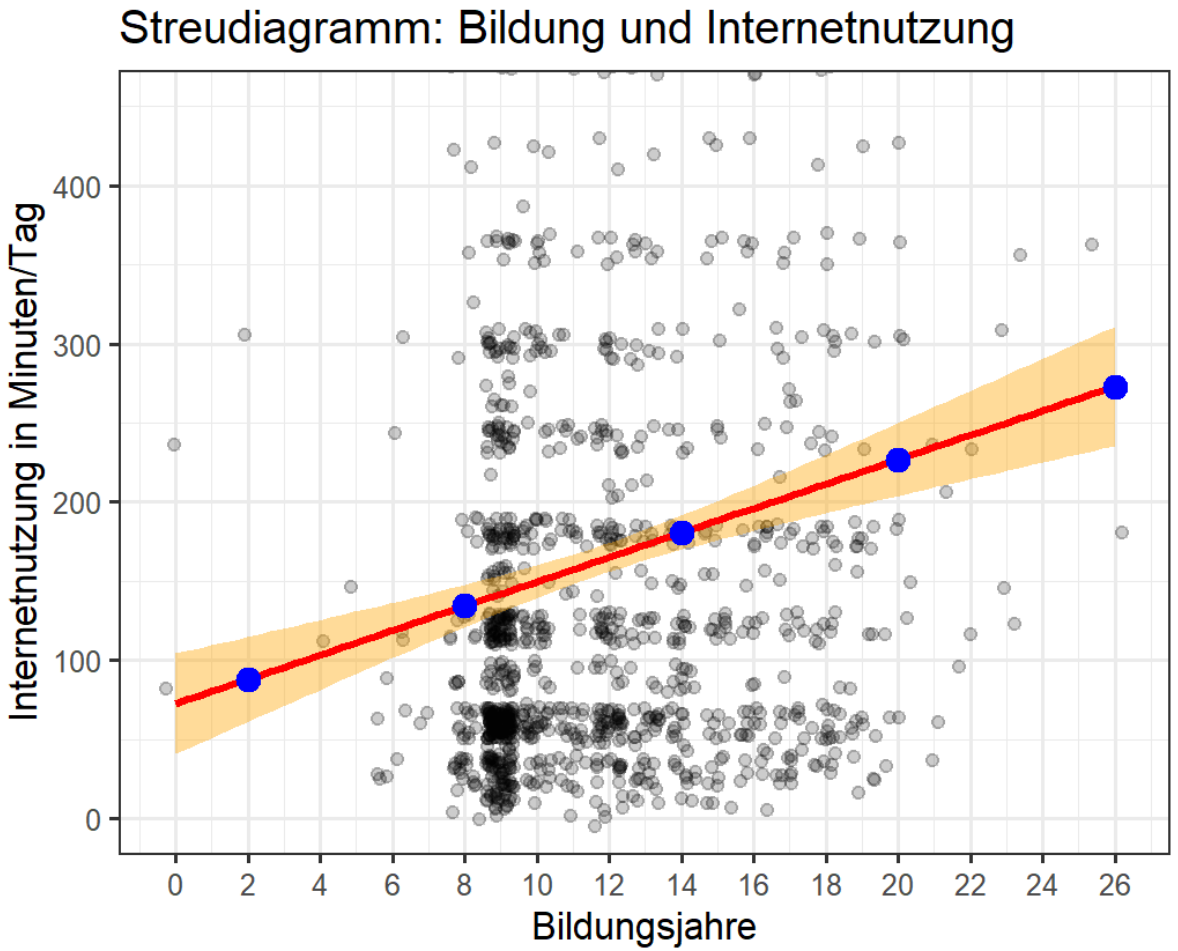
```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "confidence",
          ci.lvl = 0.95)
```

eduyrs	Predicted	95% CI
2	88.07	[61.23, 114.91]
8	134.35	[120.94, 147.75]
14	180.62	[169.82, 191.43]
20	226.90	[203.85, 249.95]
26	273.18	[235.45, 310.91]

Was stellt die zweite, was die dritte Spalte des Outputs dar?

3.3 Konfidenzband der Regressionsgerade

Berechnung der Grenzen des Konfidenzbandes mit R



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

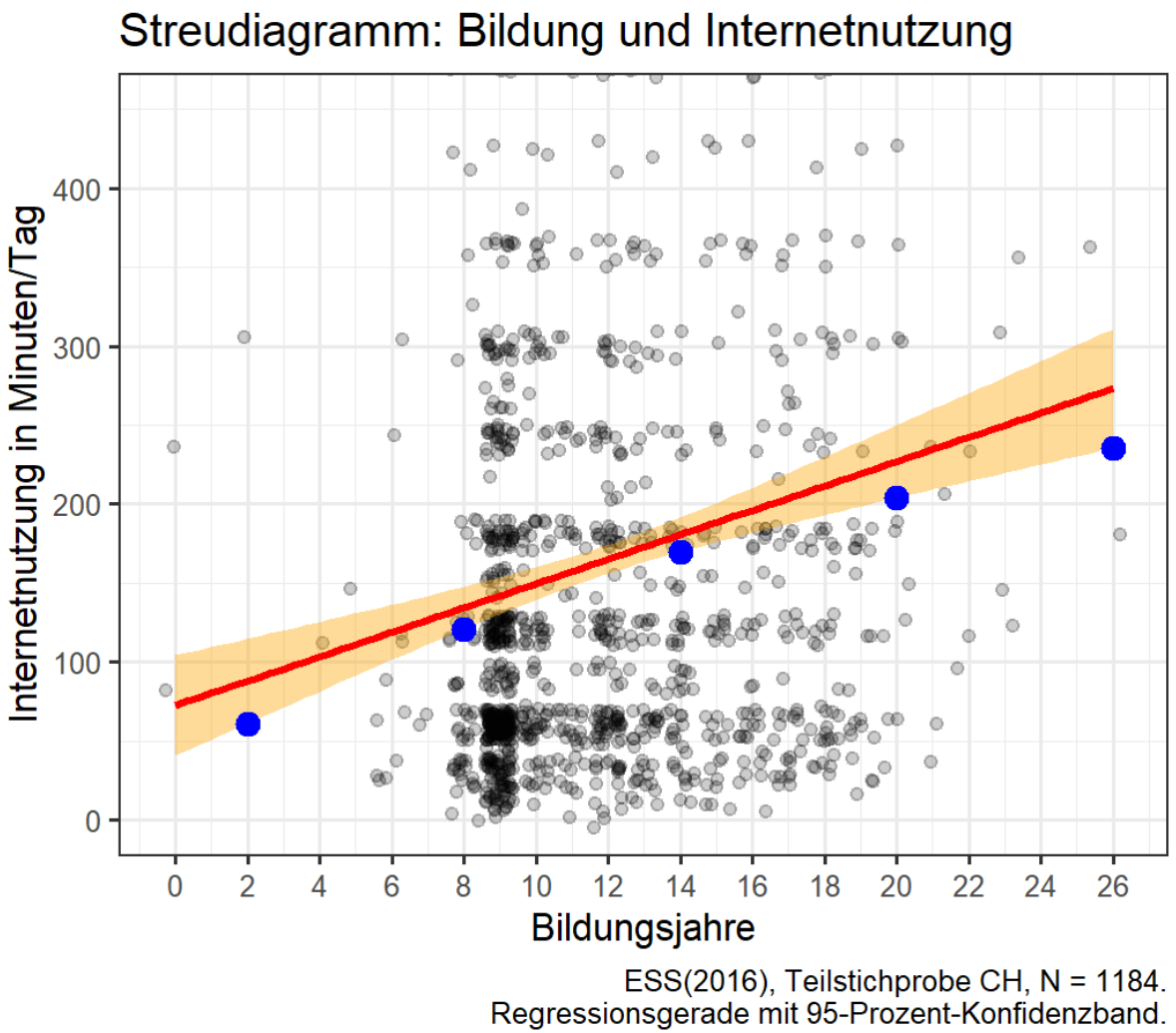
```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "confidence",
          ci.lvl = 0.95)
```

eduyrs	Predicted	95% CI
2	88.07	[61.23, 114.91]
8	134.35	[120.94, 147.75]
14	180.62	[169.82, 191.43]
20	226.90	[203.85, 249.95]
26	273.18	[235.45, 310.91]

Verlauf der Regressionsgerade bzw. vorhergesagte Werte: Für eine Person mit 2 Bildungsjahren werden 88 Internetminuten vorhergesagt

3.3 Konfidenzband der Regressionsgerade

Berechnung der Grenzen des Konfidenzbandes mit R



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

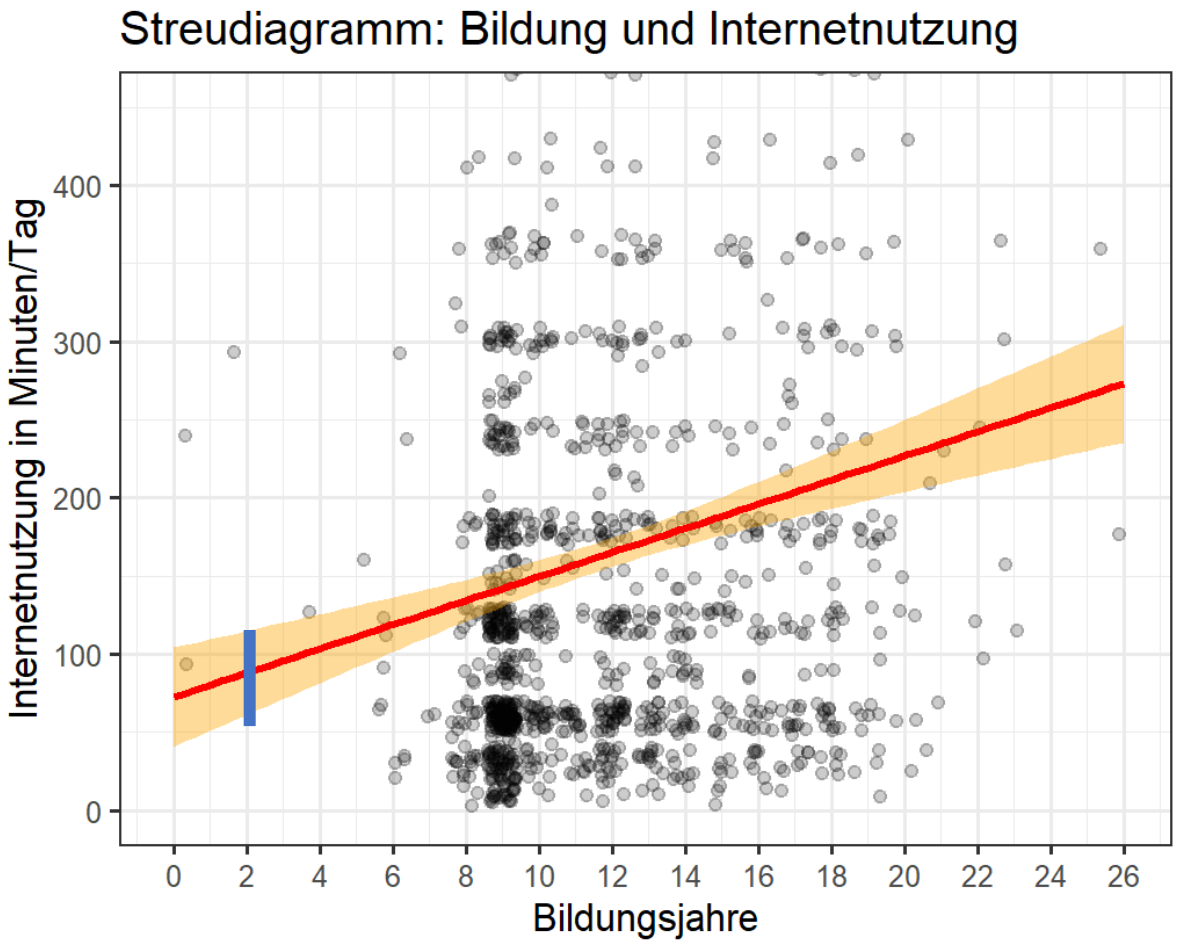
```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "confidence",
          ci.lvl = 0.95)
```

eduyrs	Predicted	95% CI
2	88.07	[61.23, 114.91]
8	134.35	[120.94, 147.75]
14	180.62	[169.82, 191.43]
20	226.90	[203.85, 249.95]
26	273.18	[235.45, 310.91]

Unterer Grenzverlauf des 95%-Konfidenzbandes: Bei x=2 liegt die untere Grenze des 95%-Konfidenzbandes bei y=61

3.3 Konfidenzband der Regressionsgerade

Berechnung der Grenzen des Konfidenzbandes mit R



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

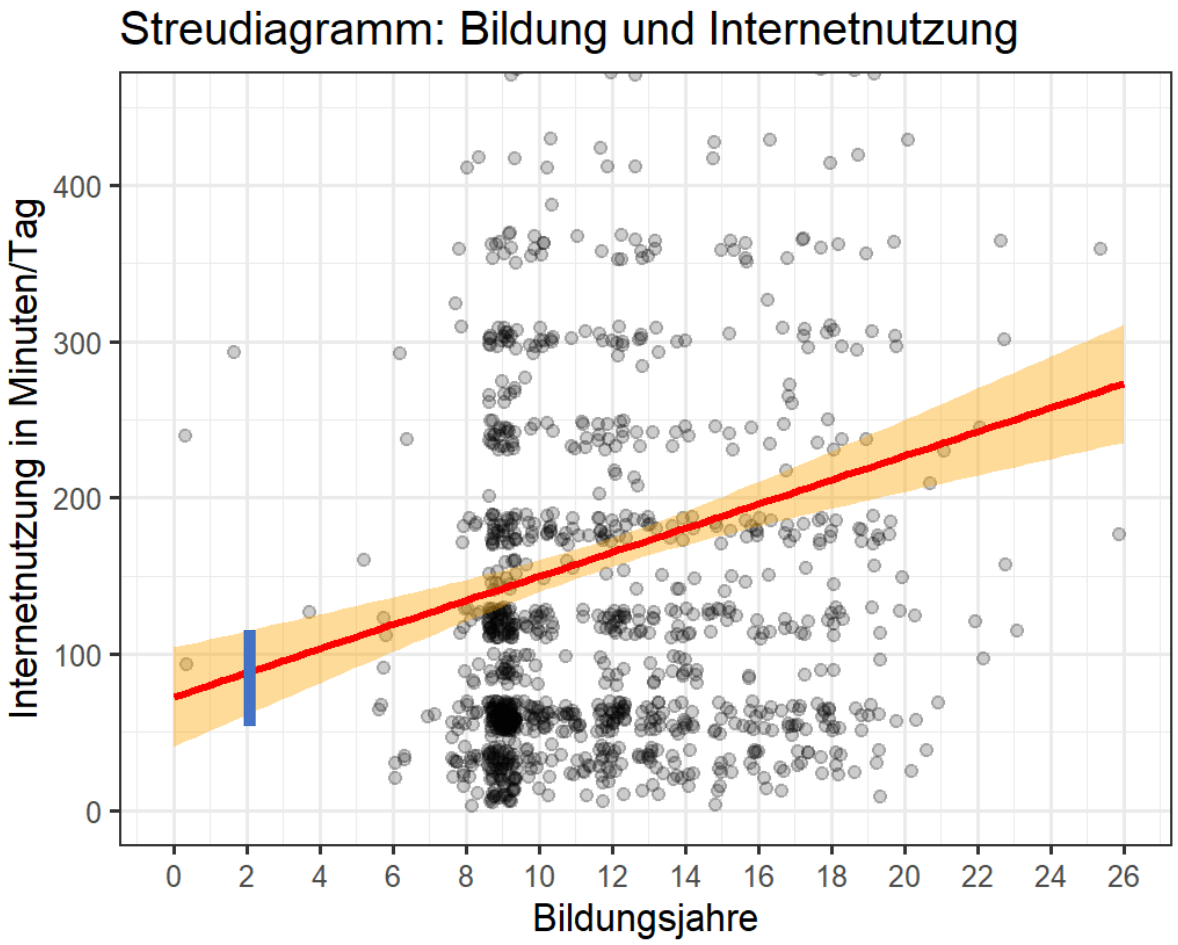
```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "confidence",
          ci.lvl = 0.95)
```

eduyrs	Predicted	95% CI
2	88.07	[61.23, 114.91]
8	134.35	[120.94, 147.75]
14	180.62	[169.82, 191.43]
20	226.90	[203.85, 249.95]
26	273.18	[235.45, 310.91]

Interpretation einer Zeile des Outputs?

3.3 Konfidenzband der Regressionsgerade

Berechnung der Grenzen des Konfidenzbandes mit R



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

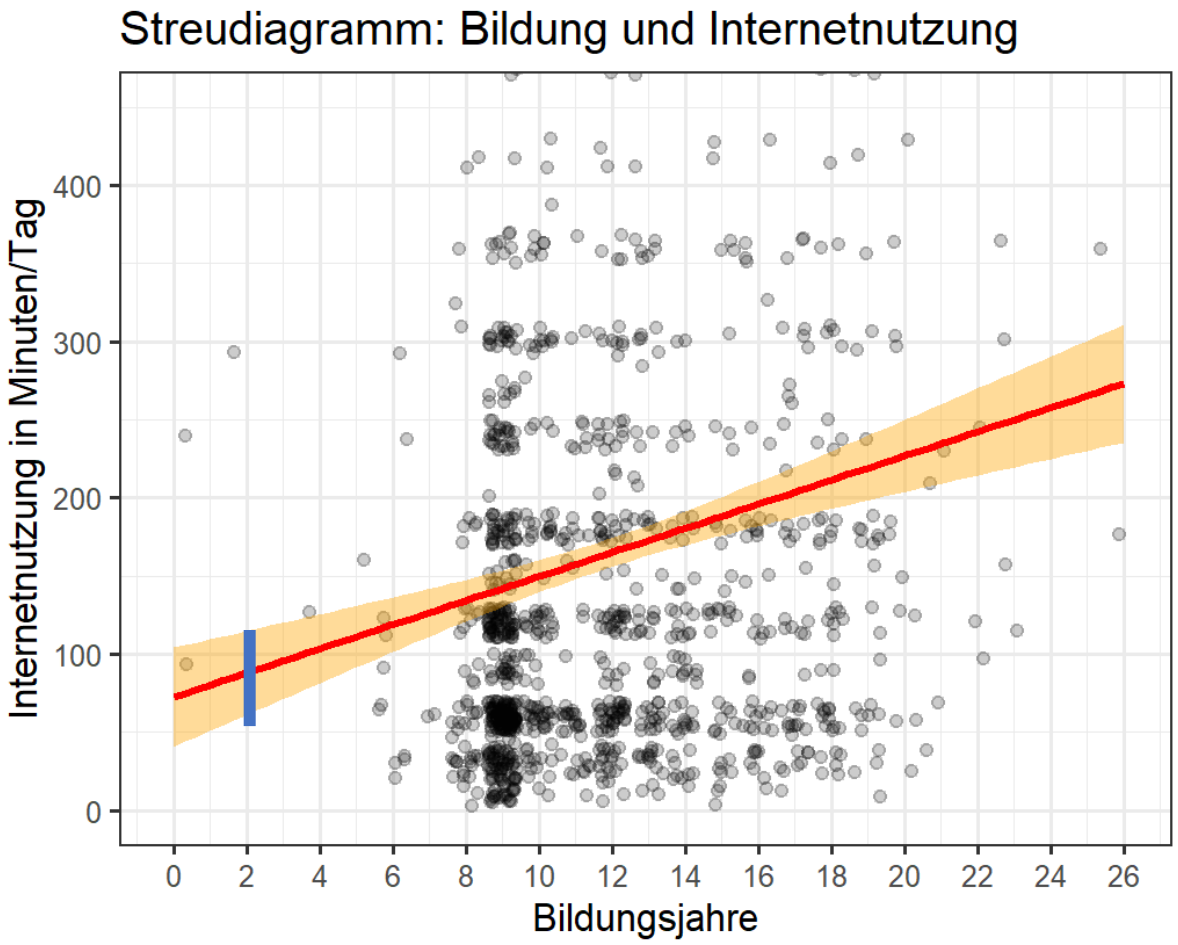
```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "confidence",
          ci.lvl = 0.95)
```

eduyrs	Predicted	95% CI
2	88.07	[61.23, 114.91]
8	134.35	[120.94, 147.75]
14	180.62	[169.82, 191.43]
20	226.90	[203.85, 249.95]
26	273.18	[235.45, 310.91]

- ***Regressionsbasierte Vorhersage für x=2:*** *siehe vorletzte Folie*
- ***Grenzwerte des Konfidenzbandes für x=2:*** *Nur technische, keine inhaltlich gehaltvolle Interpretation möglich…*

3.3 Konfidenzband der Regressionsgerade

Berechnung der Grenzen des Konfidenzbandes mit R



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

```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "confidence",
          ci.lvl = 0.95)
```

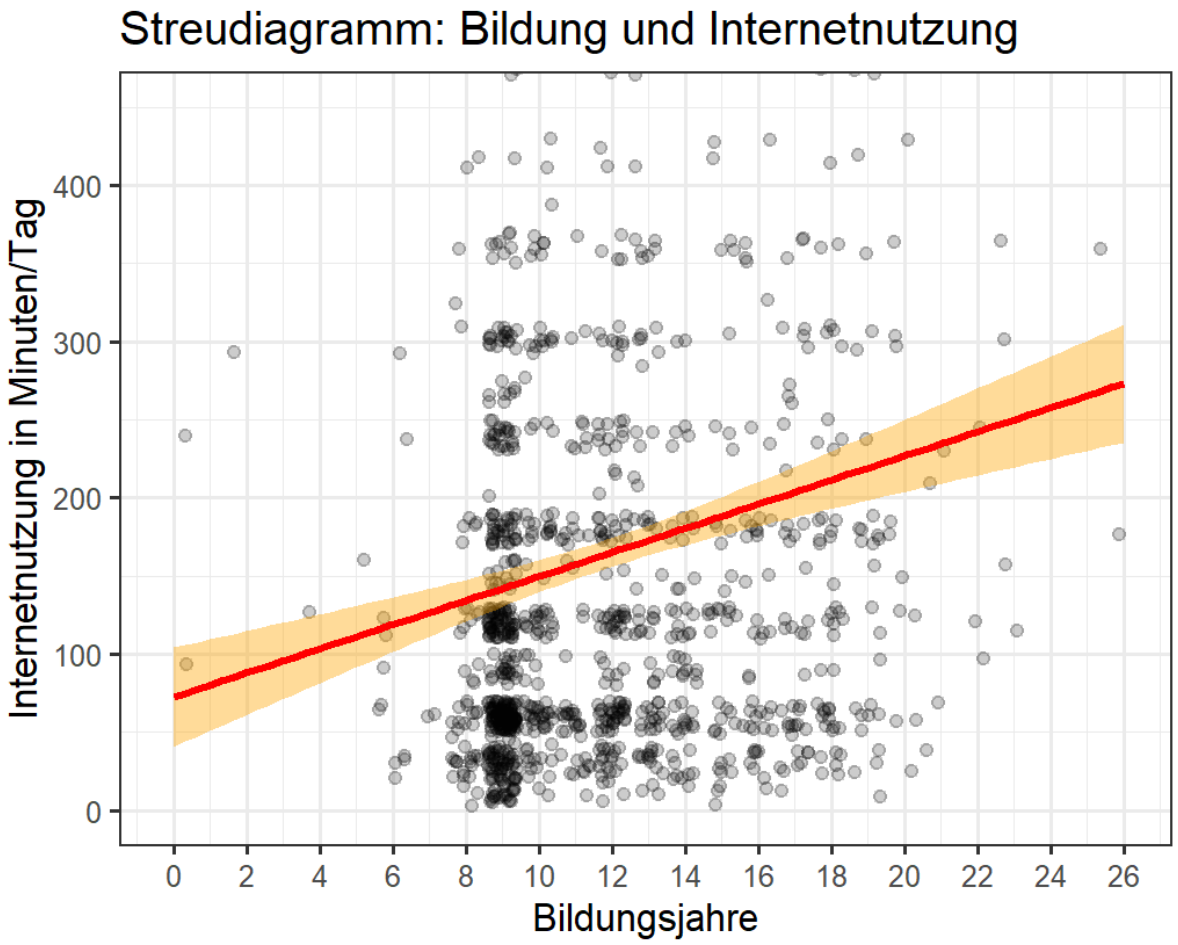
eduyrs	Predicted	95% CI
2	88.07	[61.23, 114.91]
8	134.35	[120.94, 147.75]
14	180.62	[169.82, 191.43]
20	226.90	[203.85, 249.95]
26	273.18	[235.45, 310.91]

- *„Das 95%-Konfidenzband verläuft bei 2 Bildungsjahren zwischen 61 und 114 Minuten“, oder*
- *„Mit 95%-Sicherheit verläuft die wahre Regressionsgerade der Population bei x=2 oberhalb von y=61 und unterhalb von y=115“*

Trotz der wenig anschaulichen Bedeutung seiner Grenzwerte hat das Konfidenzband hohe inferenzstatistische Darstellungskraft…

3.3 Konfidenzband der Regressionsgerade

Darstellungskraft des 95% Konfidenzbandes

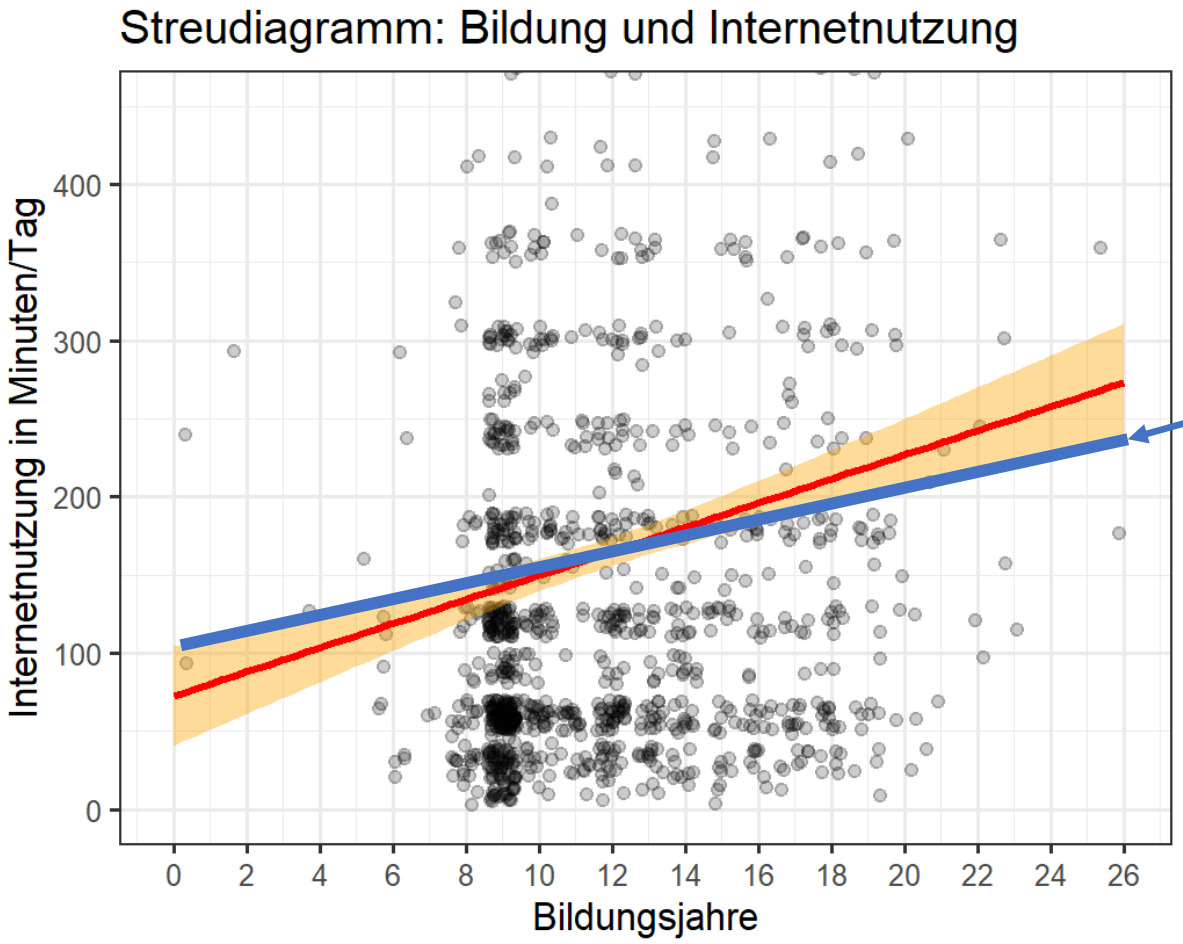


ESS(2016), Teilstichprobe CH, N = 1184.
Regressionsgerade mit 95-Prozent-Konfidenzband.

Welche inferenzstatischen Sachverhalte werden durch das Konfidenzband visualisiert?

3.3 Konfidenzband der Regressionsgerade

Darstellungskraft des 95% Konfidenzbandes



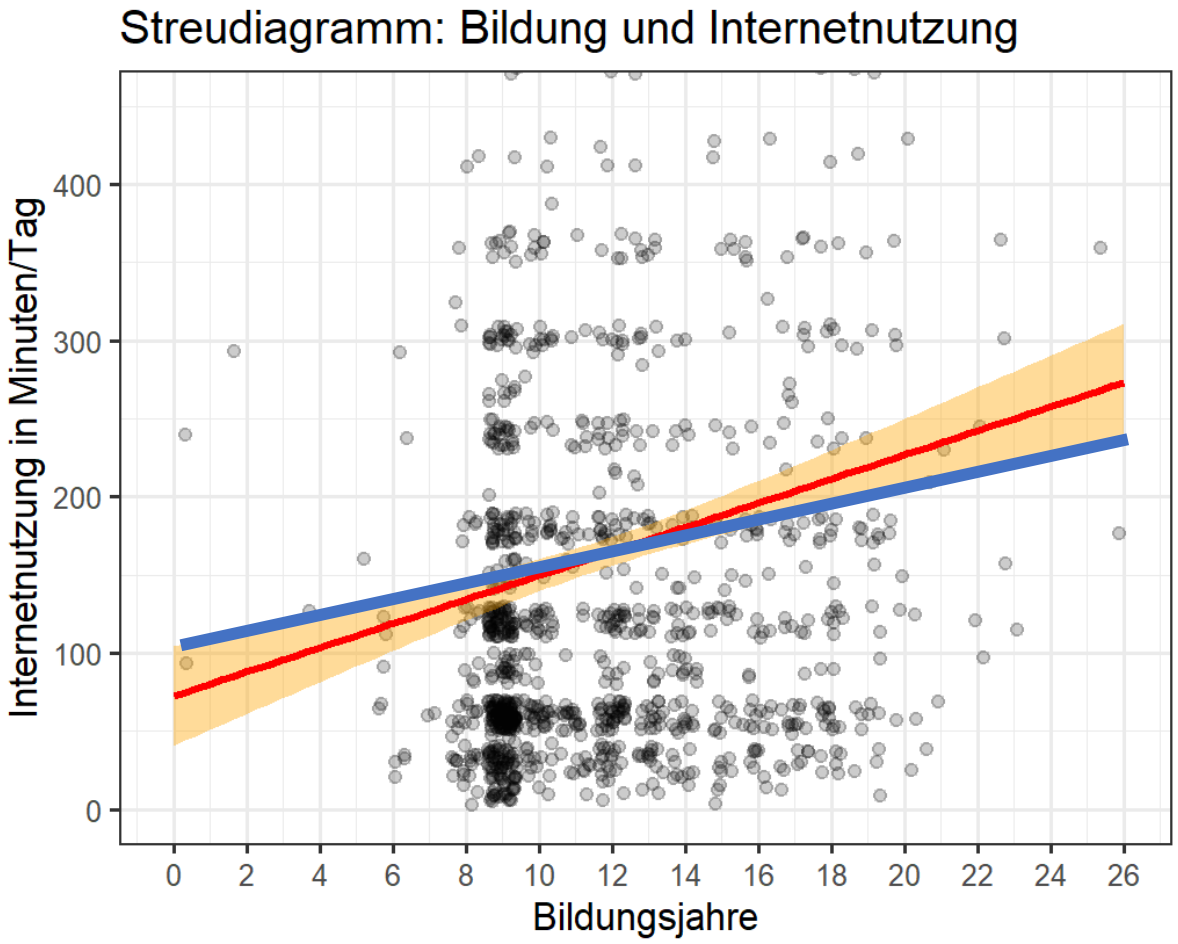
ESS(2016), Teilstichprobe CH, N = 1184.
Regressionsgerade mit 95-Prozent-Konfidenzband.

Welche inferenzstatischen Sachverhalte werden durch das Konfidenzband visualisiert?

- *Selbst eine Gerade mit 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

Darstellungskraft des 95% Konfidenzbandes



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

Welche inferenzstatischen Sachverhalte werden durch das Konfidenzband visualisiert?

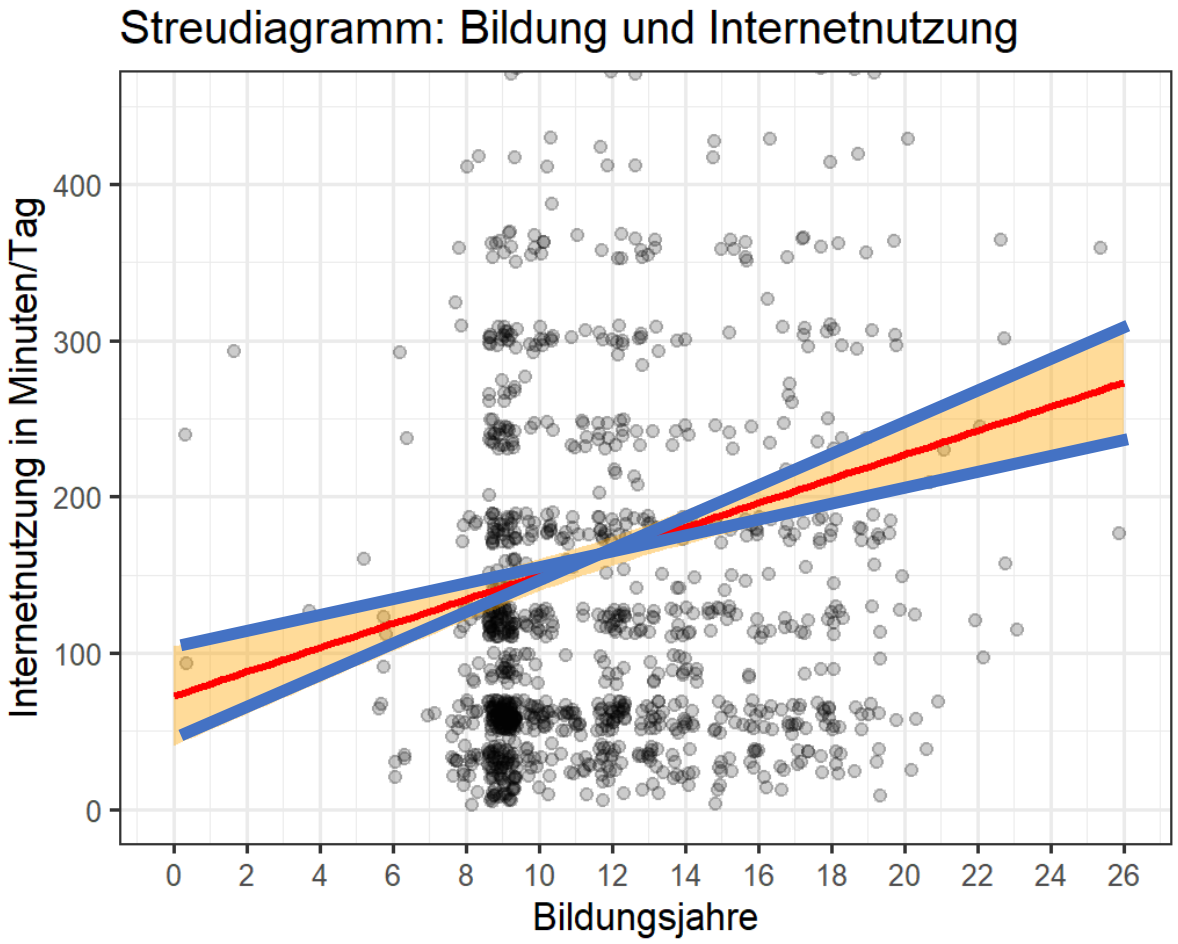
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    72.646     16.171   4.492 7.74e-06 ***
eduyrs          7.713      1.313   5.875 5.48e-09 ***
```

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

Darstellungskraft des 95% Konfidenzbandes



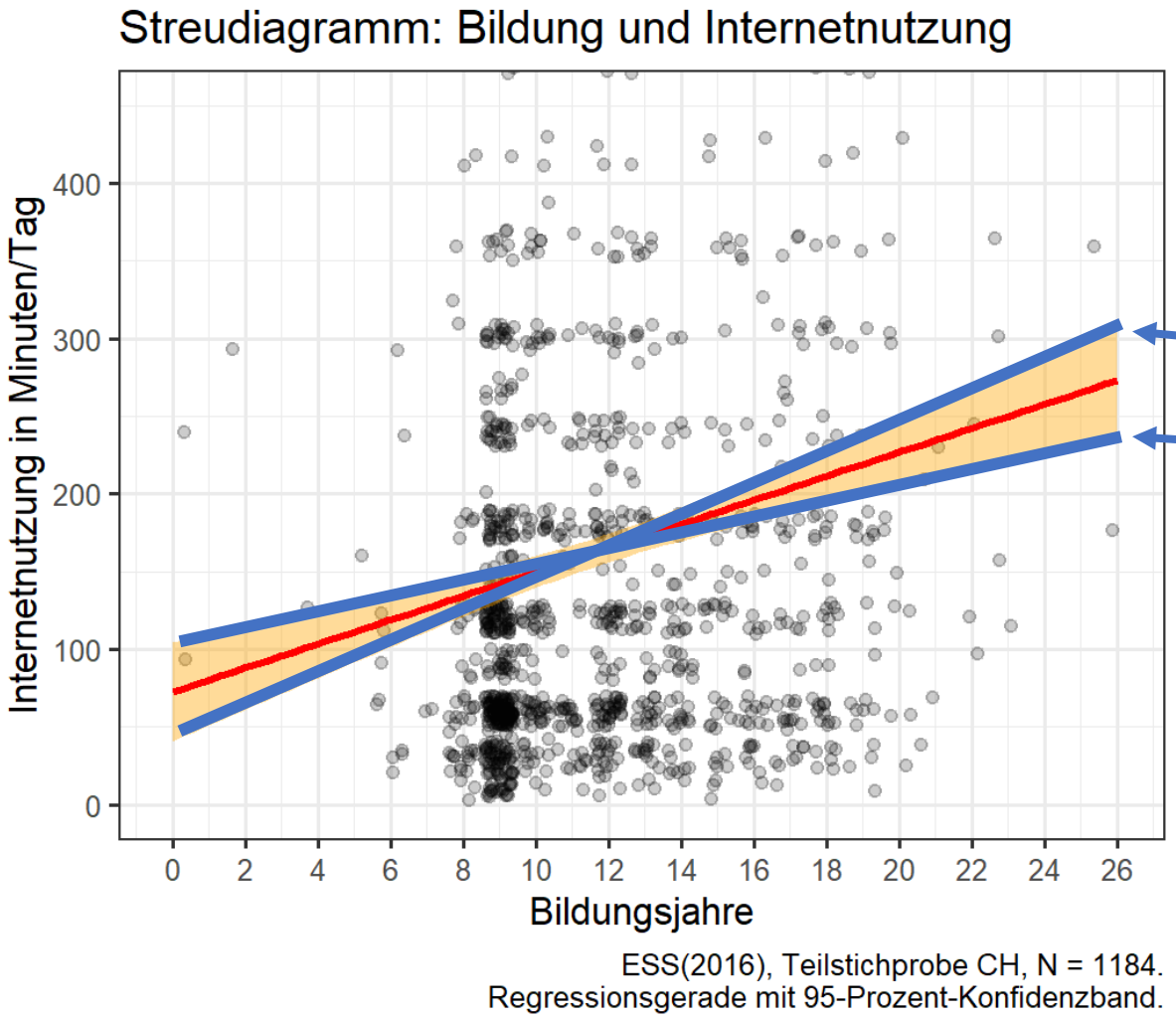
ESS(2016), Teilstichprobe CH, N = 1184.
Regressionsgerade mit 95-Prozent-Konfidenzband.

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

Darstellungskraft des 95% Konfidenzbandes



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

Welche inferenzstatischen Sachverhalte werden durch das Konfidenzband visualisiert?

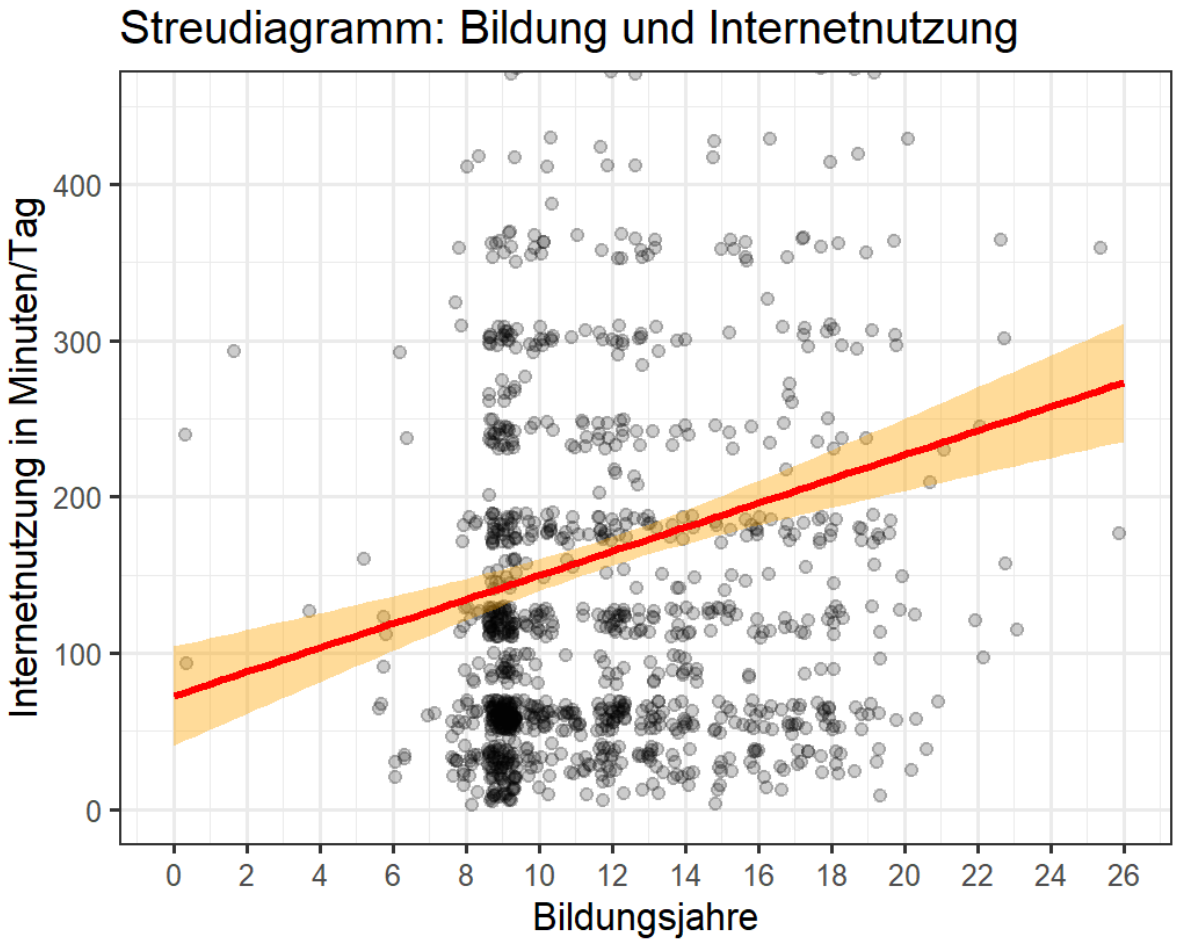
```
> confint(fit, level = 0.95)
                2.5 %    97.5 %
(Intercept) 40.918306 104.37282
eduyrs       5.137214  10.28828
```

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

Darstellungskraft des 95% Konfidenzbandes



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

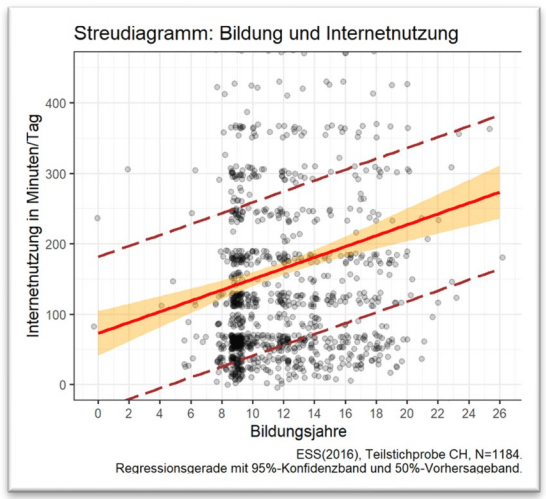
Welche inferenzstatischen Sachverhalte werden durch das Konfidenzband visualisiert?

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*).

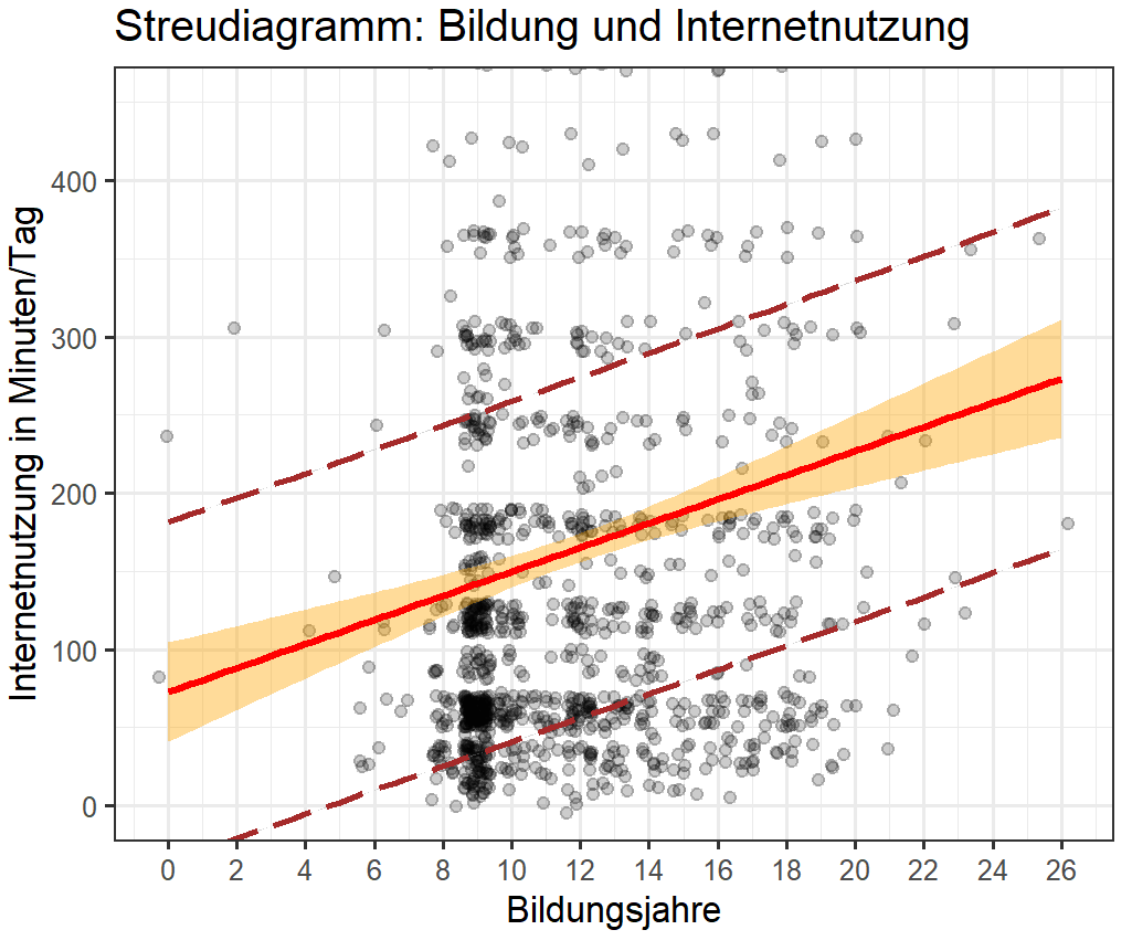
Es visualisiert somit die zentralen inferenzstatistischen Parameter der Regressionsanalyse.

3.4 Das Vorhersageband der Regressionsgerade



3.4 Vorhersageband der Regressionsgerade

Bedeutung des (hier) dunkelrot gestrichelten Bereichs?



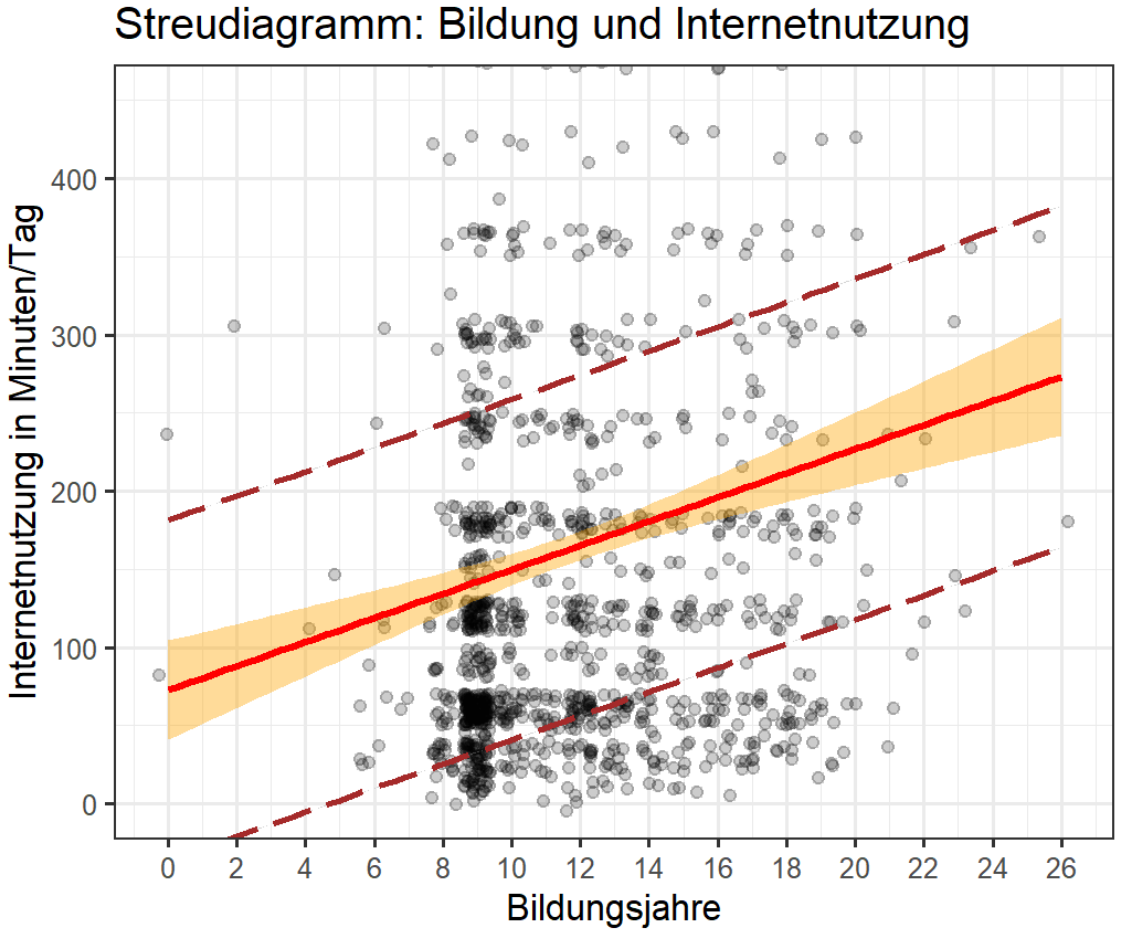
ESS(2016), Teilstichprobe CH, N=1184.
Regressionsgerade mit 95%-Konfidenzband und 50%-Vorhersageband.

Das 50%-Vorhersageband zeigt den Bereich an, in dem 50% aller Werte der Grundgesamtheit liegen bzw. in dem ein einzelner Wert der Grundgesamtheit 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, ist es häufig sinnvoll, den Sicherheitswert hier niedriger (z.B. 50% oder 75%) anzusetzen.

3.4 Vorhersageband der Regressionsgerade

Berechnung der Grenzen des Vorhersagbandes mit R



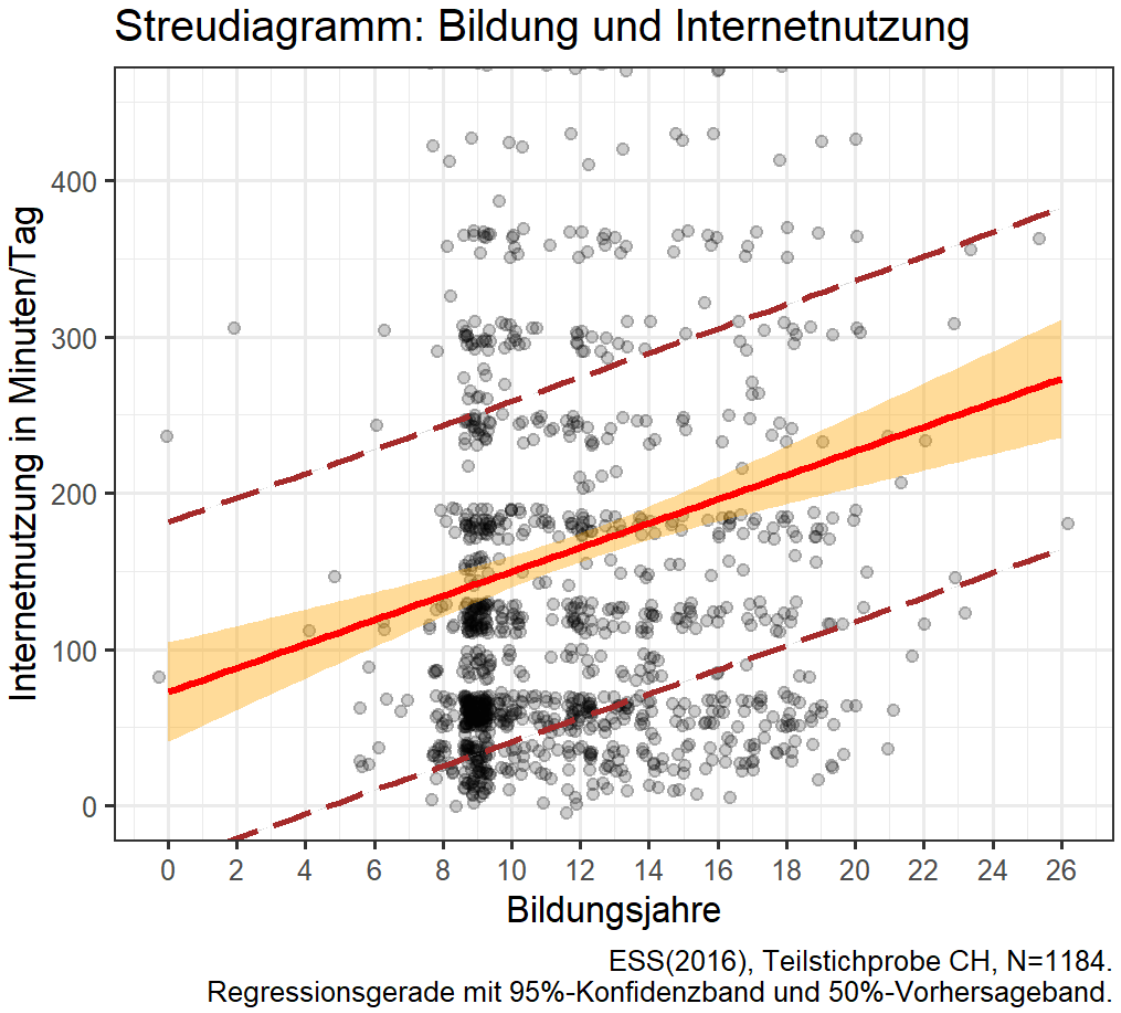
ESS(2016), Teilstichprobe CH, N=1184.
Regressionsgerade mit 95%-Konfidenzband und 50%-Vorhersageband.

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

Aktiviere den Befehl und erläutere die einzelnen Argumente

3.4 Vorhersageband der Regressionsgerade

Berechnung der Grenzen des Vorhersagbandes mit R



ESS(2016), Teilstichprobe CH, N=1184.
Regressionsgerade mit 95%-Konfidenzband und 50%-Vorhersageband.

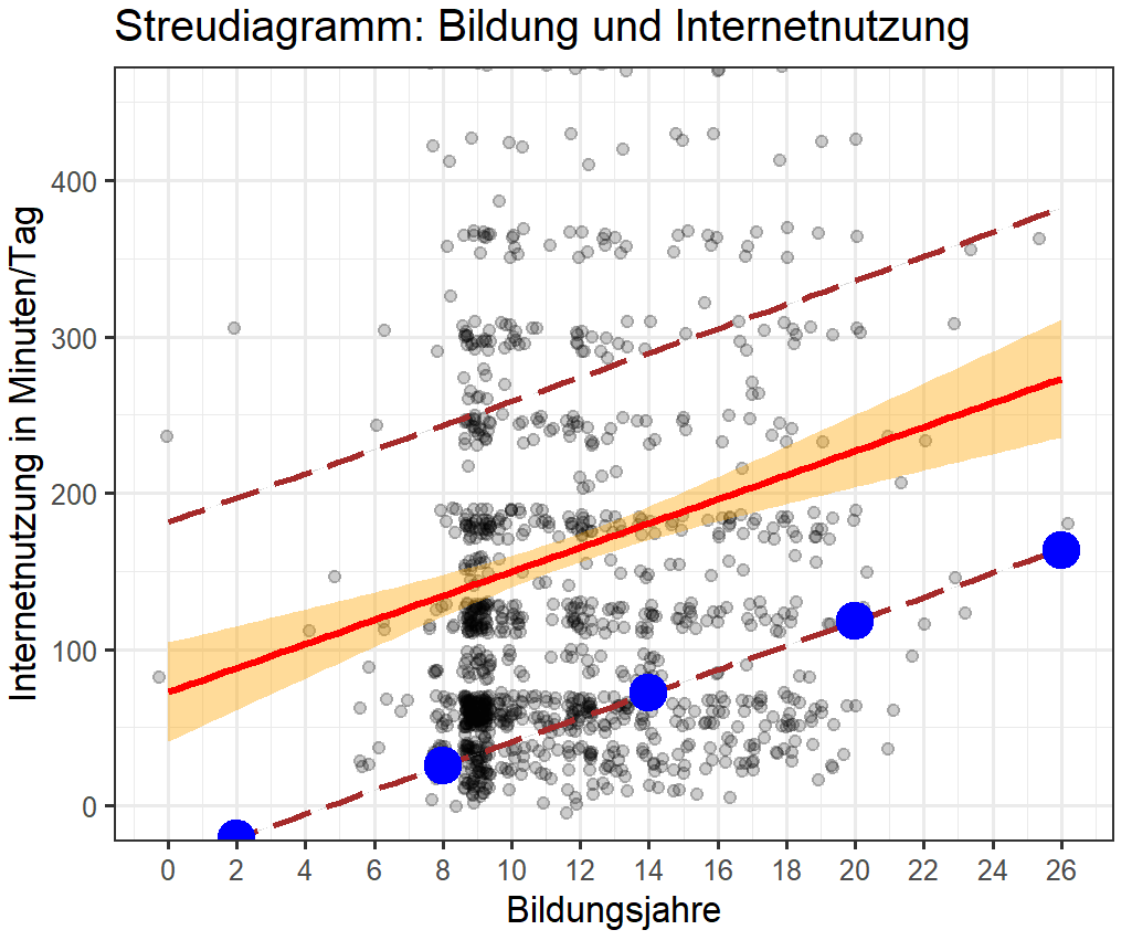
```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "prediction",
          ci.lvl = 0.50)
```

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

Berechnung der Grenzen des Vorhersagbandes mit R



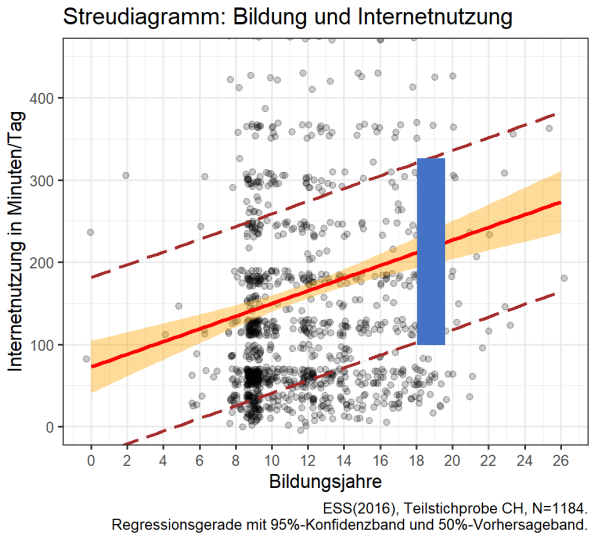
ESS(2016), Teilstichprobe CH, N=1184.
Regressionsgerade mit 95%-Konfidenzband und 50%-Vorhersageband.

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

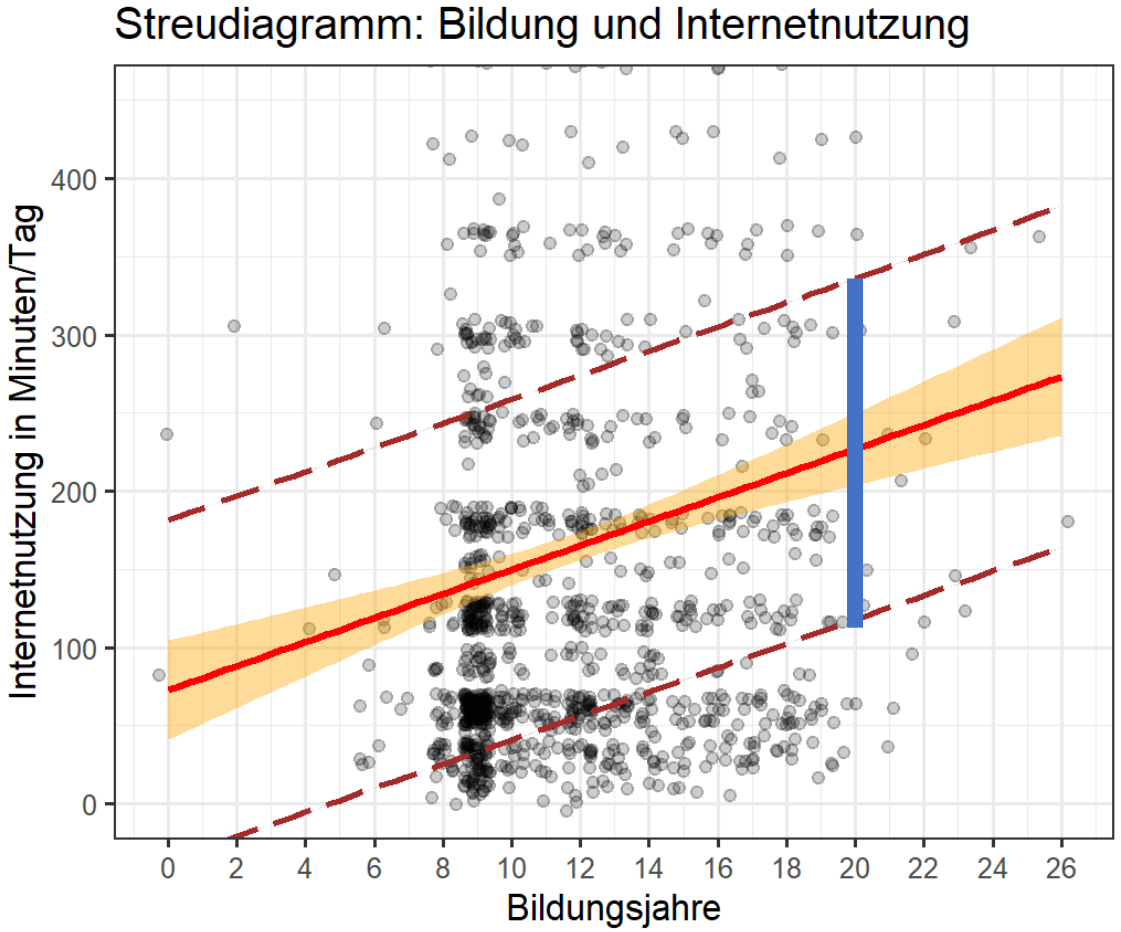
Unterer Grenzverlauf des 95%-Vorhersagebandes: Bei x=20 liegt die untere Grenze des 95%-Vorhersagebandes bei y=118

3.4b Das Vorhersageintervall der Regressionsvorhersage



3.4 Vorhersageband der Regressionsgerade

Berechnung der Grenzen des Vorhersagbandes mit R: Das Vorhersageintervall



ESS(2016), Teilstichprobe CH, N=1184.
Regressionsgerade mit 95%-Konfidenzband und 50%-Vorhersageband.

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

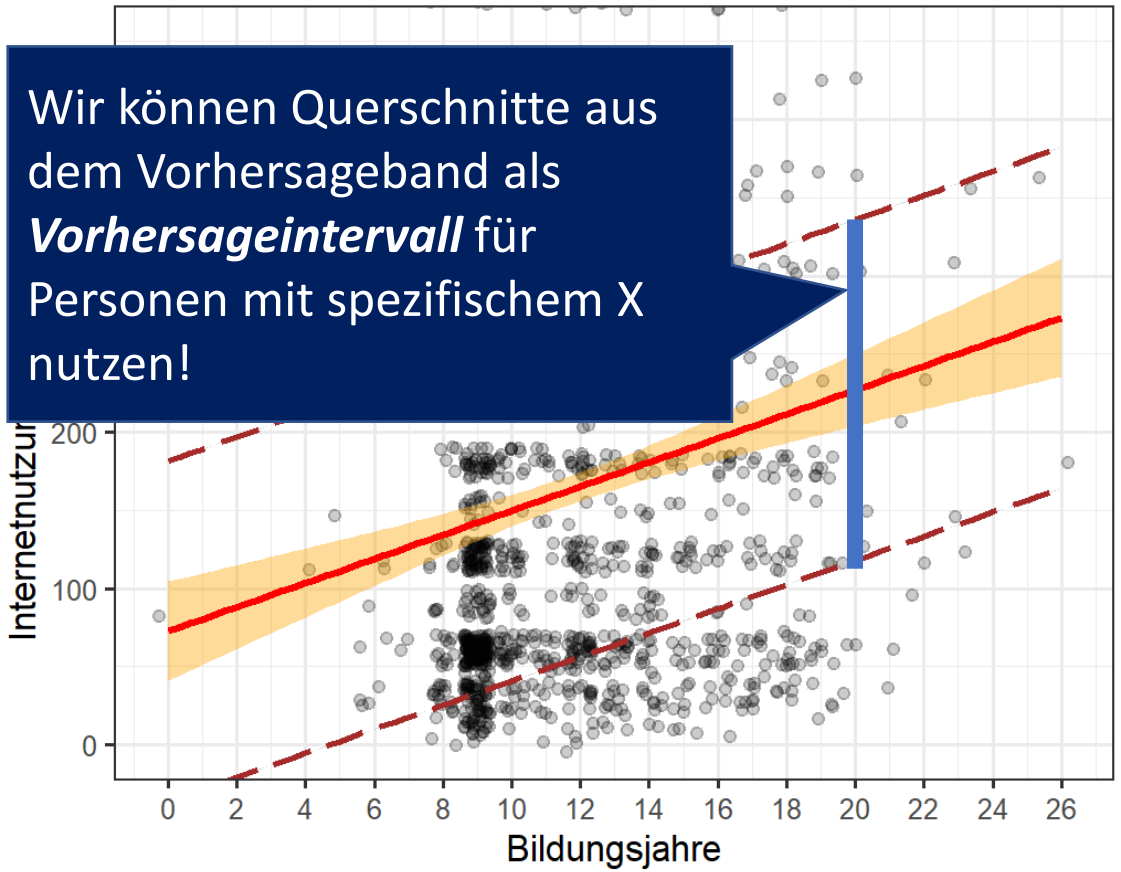
eduyrs	Predicted	50% CI
2	88.07	[-20.99, 197.13]
8	134.35	[25.58, 243.11]
14	180.62	[71.89, 289.36]
20	226.90	[117.94, 335.86]
26	273.18	[163.74, 382.62]

Interpretation einer Zeile des Outputs?

3.4 Vorhersageband der Regressionsgerade

Berechnung der Grenzen des Vorhersagbandes mit R: Das Vorhersageintervall

Streudiagramm: Bildung und Internetnutzung



ESS(2016), Teilstichprobe CH, N=1184.
Regressionsgerade mit 95%-Konfidenzband und 50%-Vorhersageband.

Wir können Querschnitte aus dem Vorhersageband als ***Vorhersageintervall*** für Personen mit spezifischem X nutzen!

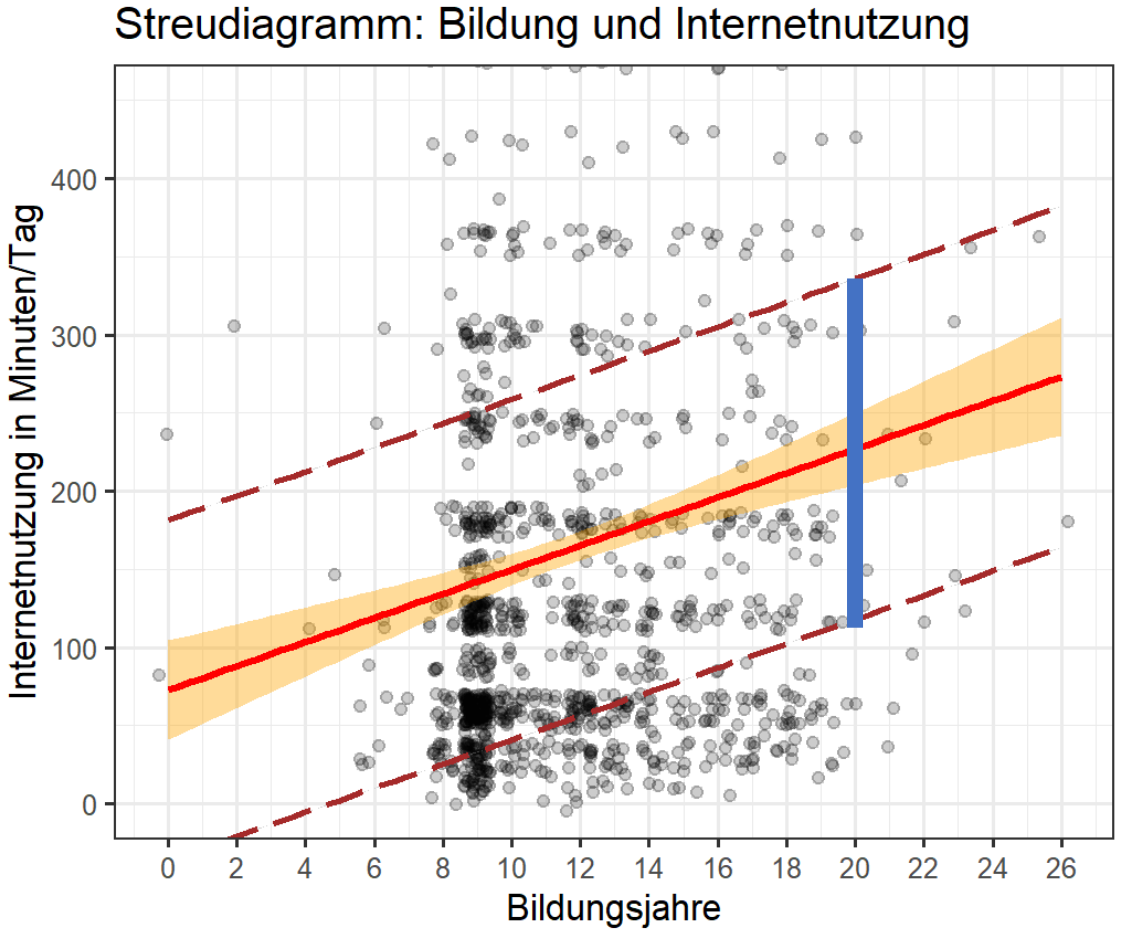
```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "prediction",
          ci.lvl = 0.50)
```

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

Berechnung der Grenzen des Vorhersagbandes mit R: Das Vorhersageintervall



ESS(2016), Teilstichprobe CH, N=1184.
Regressionsgerade mit 95%-Konfidenzband und 50%-Vorhersageband.

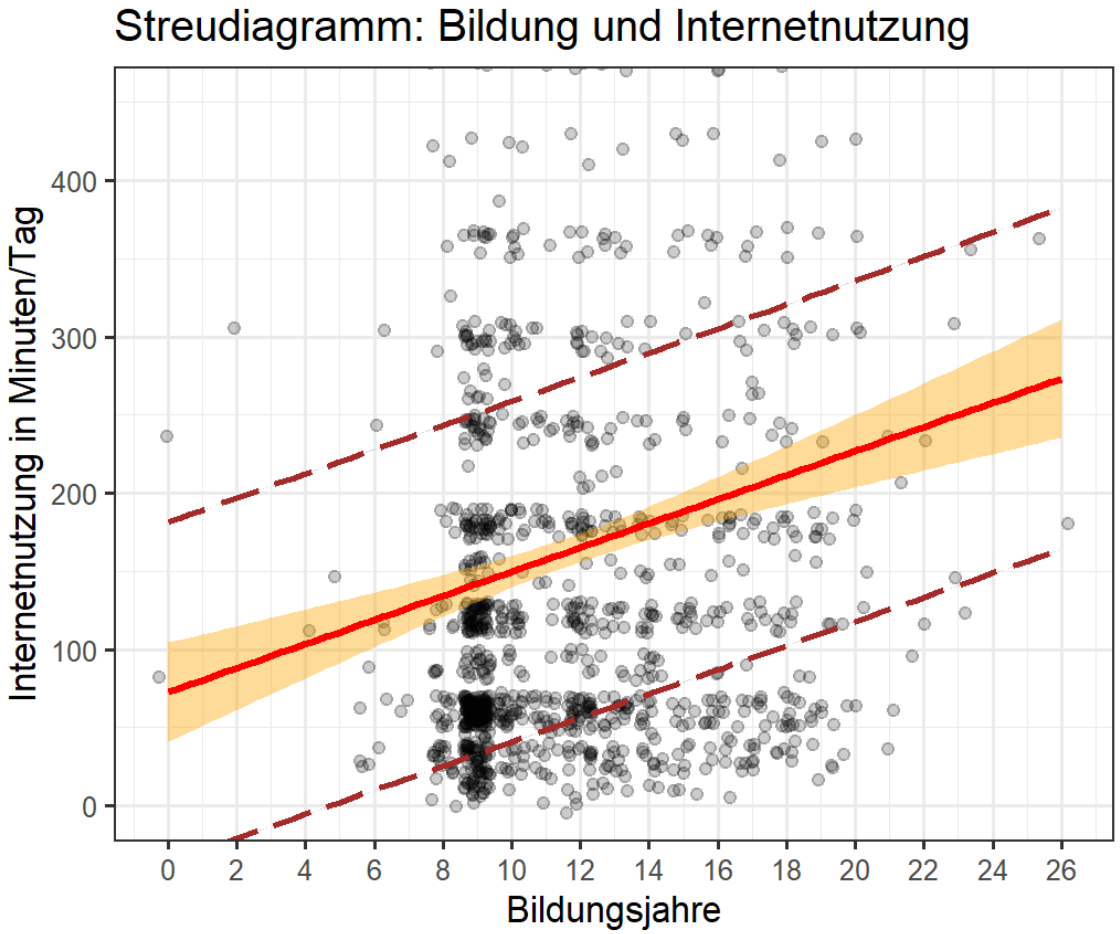
```
library (ggeffects)
ggpredict(fit,
          terms = "eduyrs[2, 8, 14, 20, 26]",
          interval = "prediction",
          ci.lvl = 0.50)
```

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 Konfidenzbandes im ggplot-Scatterplot



ESS(2016), Teilstichprobe CH, N=1184.
Regressionsgerade mit 95%-Konfidenzband und 50%-Vorhersageband.

Das Vorhersageband wird nur selten integriert. Entsprechend gibt es für das Vorhersageband gibt es **keine Standardoption** im Rahmen des ggplot.

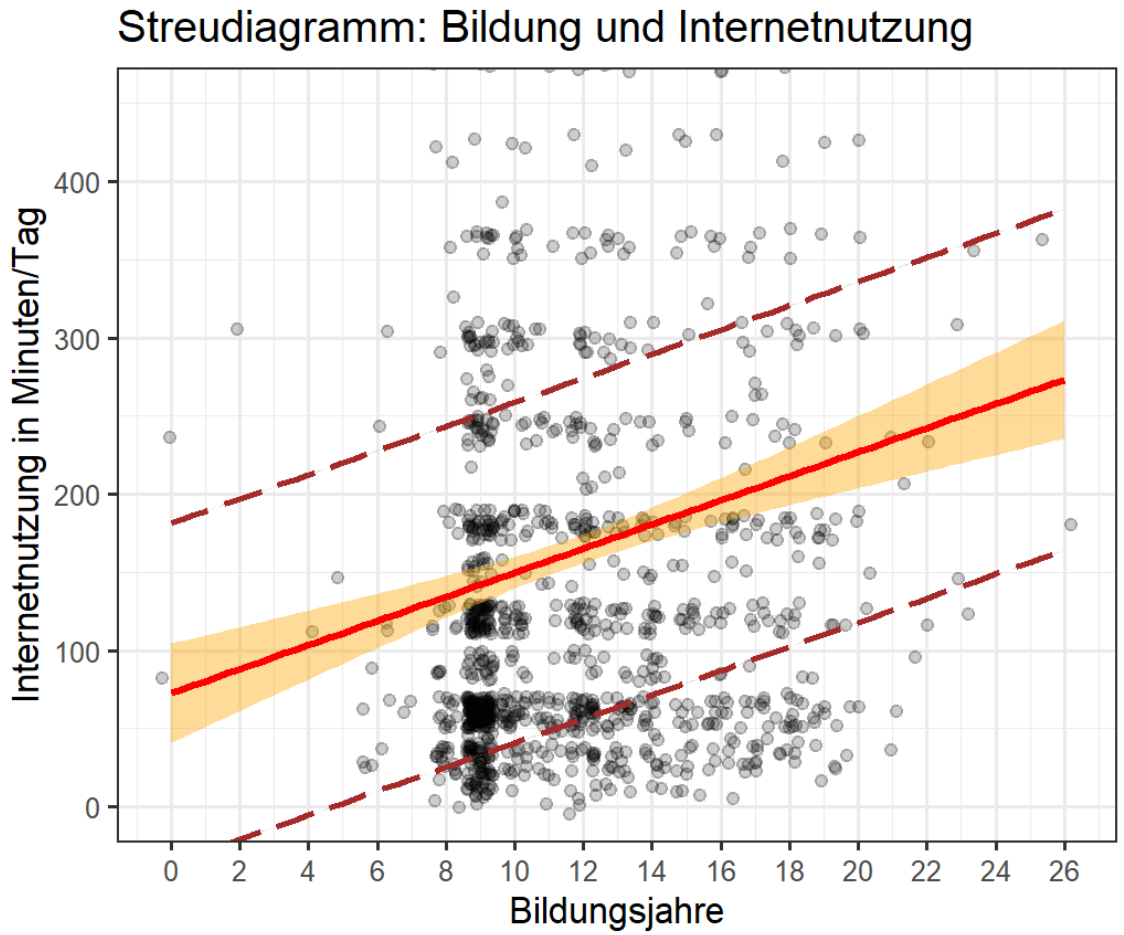
...wir müssen zunächst eine Datentabelle erstellen, welcher Informationen zum Grenzverlauf des Vorhersagebandes enthält. Diese Aufgabe übernimmt wiederum **ggpredict()**

```
predictions <- ggpredict(fit,
                         terms = "eduyrs",
                         interval = "prediction",
                         ci.lvl = 0.50)
```

Derselbe Befehl wie zuvor, definiert jetzt aber ein Objekt. Beschreibe dieses Objekt!

3.4 Vorhersageband der Regressionsgerade

Darstellung des Konfidenzbandes im ggplot-Scatterplot



ESS(2016), Teilstichprobe CH, N=1184.
Regressionsgerade mit 95%-Konfidenzband und 50%-Vorhersageband.

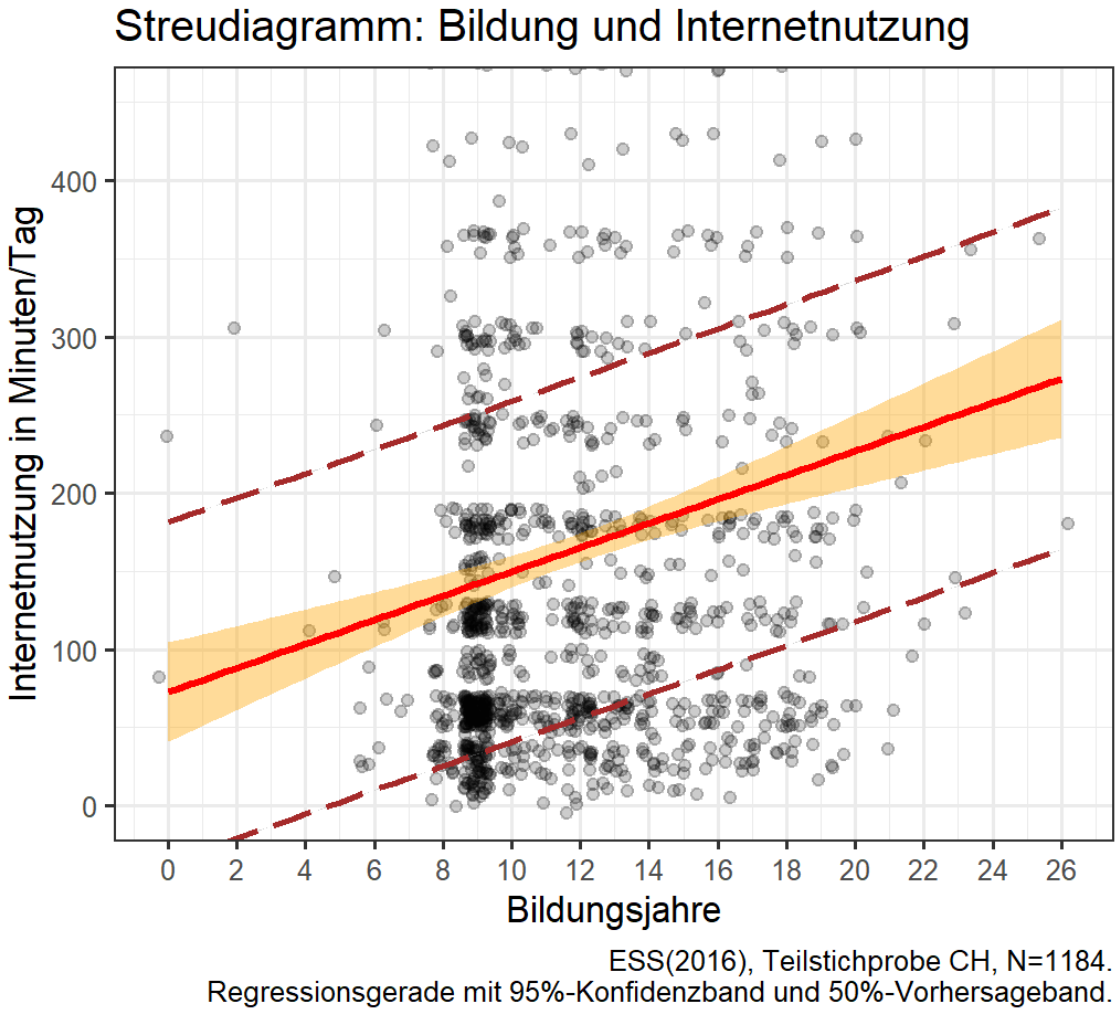
```
predictions <- ggpredict(fit,
                         terms = "eduyrs",
                         interval = "prediction",
                         ci.lvl = 0.50)
```

	x	predicted	std.error	conf.low	conf.high	group
1	0	72.64556	161.8715	-36.535091	181.8262	1
2	2	88.07106	161.6416	-20.954514	197.0966	1
3	4	103.49655	161.4540	-5.402516	212.3956	1
4	6	118.92204	161.3090	10.120804	227.7233	1
5	8	134.34753	161.2066	25.615368	243.0797	1
6	10	149.77303	161.1469	41.081120	258.4649	1
7	12	165.19852	161.1299	56.518030	273.8790	1
8	14	180.62401	161.1558	71.926087	289.3219	1
9	16	196.04951	161.2244	87.305306	304.7937	1
10	18	211.47500	161.3357	102.655724	320.2943	1
11	20	226.90049	161.4896	117.977400	335.8236	1
12	22	242.32598	161.6860	133.270417	351.3816	1
13	24	257.75148	161.9248	148.534878	366.9681	1
14	26	273.17697	162.2057	163.770909	382.5830	1

… und dann im zweiten Schritt diese Tabelle (bzw. den dort dargestellten Grenzverlauf) in den ggplot einbinden:

3.4 Vorhersageband der Regressionsgerade

Darstellung des Konfidenzbandes im ggplot-Scatterplot




```
predictions <- ggpredict(fit,
                         terms = "eduyrs",
                         interval = "prediction",
                         ci.lvl = 0.50)


plot2 <- plot1 +
  geom_smooth(data = predictions,
              aes(x = x, y = conf.high),
              size = 0.5,
              color = "brown",
              linetype = "longdash")+
  geom_smooth(data = predictions,
              size = 0.5,
              color = "brown",
              aes(x = x, y = conf.low),
              linetype = "longdash")+
  labs(title = "Bildung und Internetnutzung",
       y = "Internetnutzung in Minuten/Tag",
       x = "Bildungsjahre",
       caption = "ESS(2016), Teilstichprobe CH, N=1184.
       \n Regressionsgerade mit 95-Prozent-Konfidenzband
       und \n 50-Prozent Vorhersageintervall.")
plot2
```

3.5 Praktische Übung

Berichtet das Konfidenzintervall, Konfidenzband und Vorhersageband für den Regressionsanalyse von **Bildung** und **Migrationswertschätzung** (Lerneinheit 3).

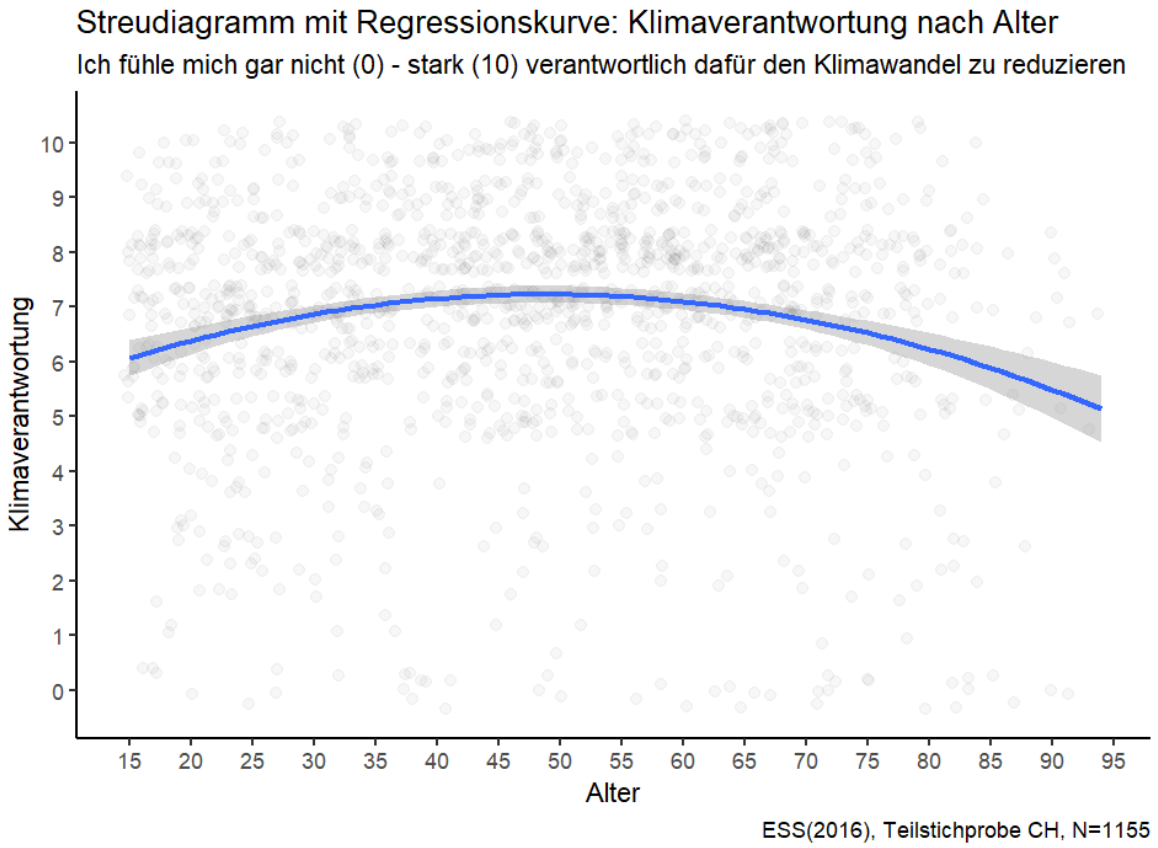
Aufgabe 0: Wie gross ist der Regressionskoeffizient in der Stichprobe, mit welcher Abweichung zum Regressionskoeffizienten in der Population muss gerechnet werden?

Aufgabe 1: In welchem Intervall liegt der wahre Koeffizient der Grundgesamtheit mit 95% Sicherheit?

Aufgabe 2: Visualisiert das Konfidenzband im Scatterplot: Wie schätzt ihr die Vertrauenswürdigkeit der Regressionsgerade ein?

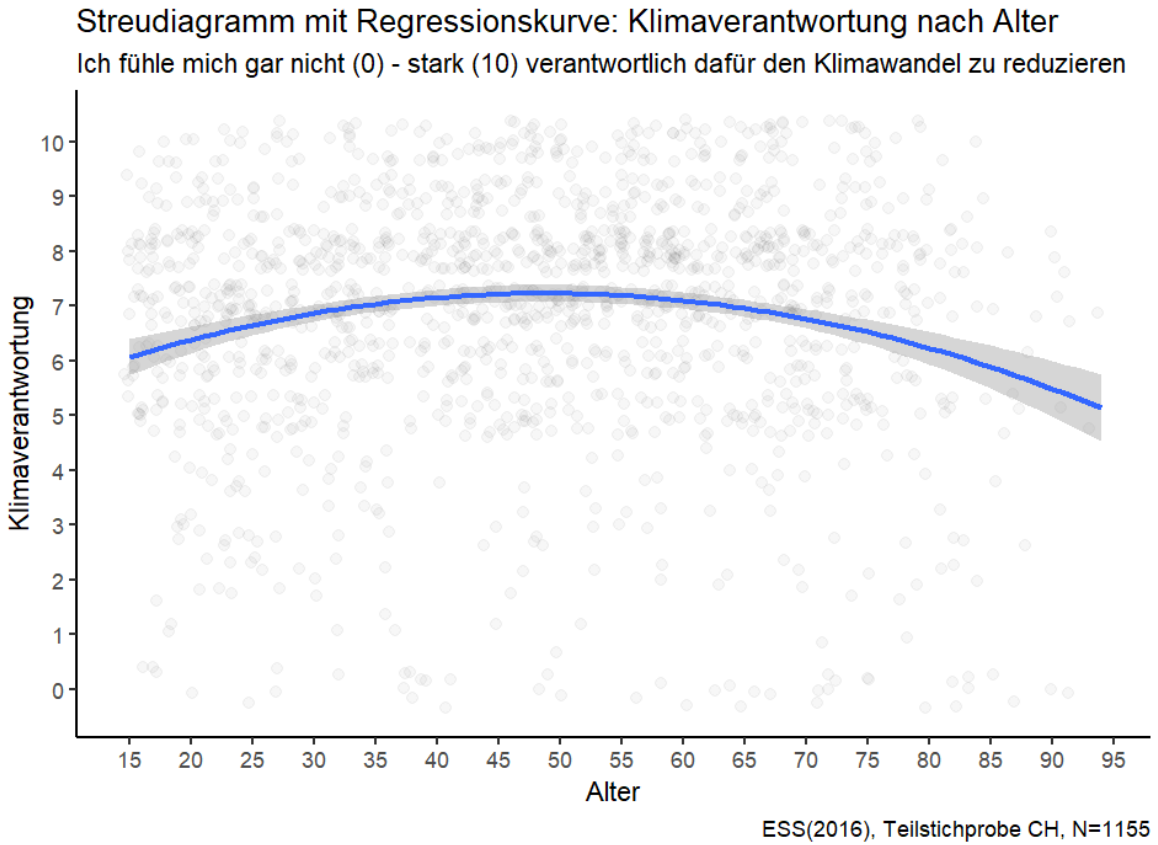
Aufgabe 3: In welchem Bereich liegt der Wert der Klimaverantwortung für eine Person mit 9 Bildungsjahren mit 50% Sicherheit?

3.6 Konfidenzband und Vorhersageband bei nicht-Linearen Regressionen




```
ggplot(ess8_ch_ss_1, aes(alter, klima_ver))+
  geom_jitter(alpha = 0.03, size =2)+
  scale_x_continuous(breaks = seq(15,100,5))+
  scale_y_continuous(breaks = seq(0,10,1))+
  geom_smooth(method = "lm"  se = T, aes(show.legend=FALSE),
              formula = y ~ poly(x,2), data = ess8_ch_ss_1)+
  scale_color_manual(values = c("blue", "red"),
                     labels =c ("Regressionskurve"),
                     name= "Legende")+
  theme_classic()+
```

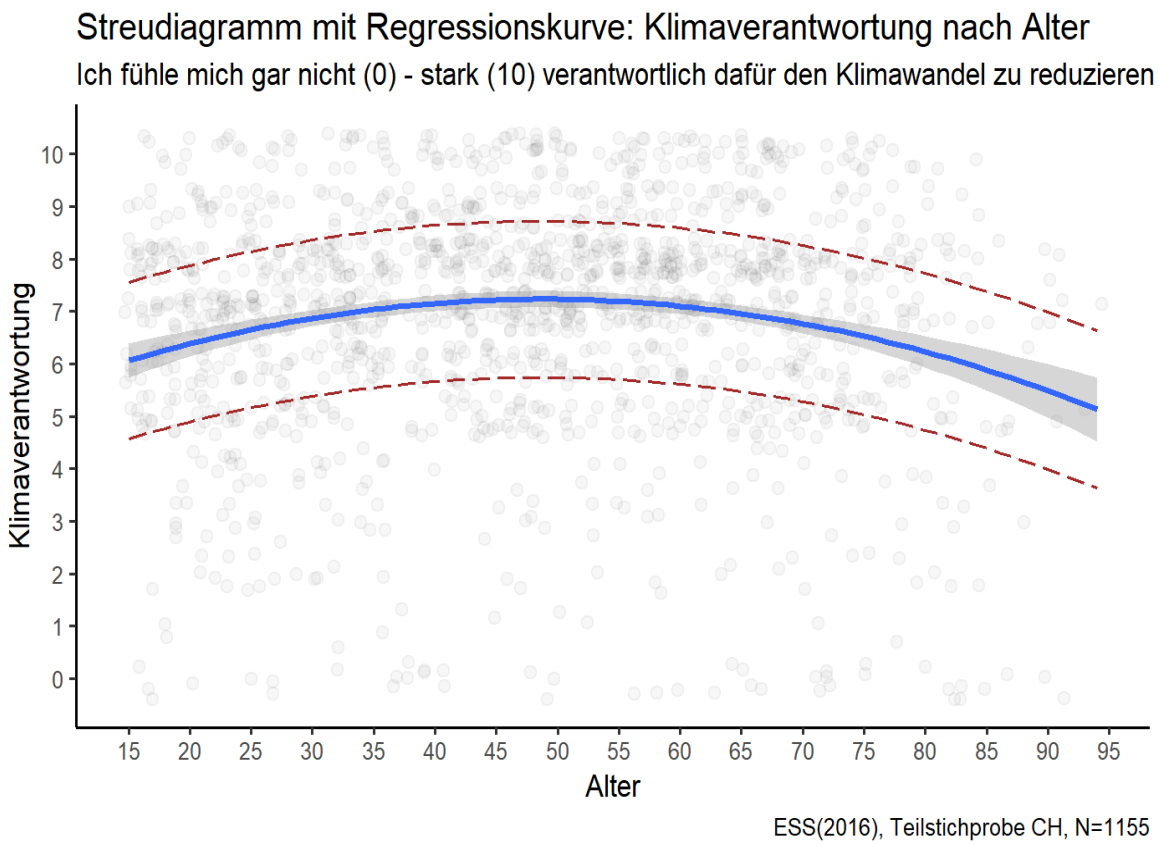
3.6 Konfidenzband und Vorhersageband bei nicht-Linearen Regressionen




```
ggpredict(model_sqr1,
          terms = "alter[20, 35, 50, 65, 80]")
```

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

3.6 Konfidenzband und Vorhersageband bei nicht-Linearen Regressionen




```
ggpredict(model_sqr1,
          terms = "alter[20, 35, 50, 65, 80]")
```

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

```
ggpredict(model_sqr1,
          terms = "alter[20, 35, 50, 65, 80]",
          interval="prediction",
          ci.lvl=0.50)
```

alter	Predicted	50% CI
20	6.38	[4.89, 7.87]
35	7.04	[5.55, 8.53]
50	7.23	[5.74, 8.72]
65	6.96	[5.47, 8.45]
80	6.23	[4.74, 7.72]

Hausaufgabe mit Selbstüberprüfung:

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