Statistik 2 – Tutorate
Sitzung 5: Inferenzstatistik

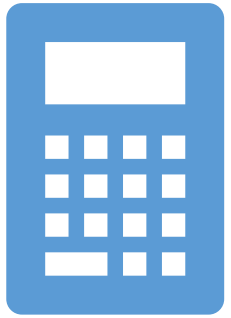
Lea Elina Hofer, Samuel Rauh, Sebastian Senn, Fynn Siefart, Marco Giesselmann

Lernziele dieser Einheit

Hypothesen

Formulierung von Null-

und Alternativhypothese



Inferenzstatistik

Standardfehler, t- und p-Wert

Konfidenzintervalle und Konfidenzband

R^2 und Vorhersageband

1 Hypothesen aufstellen

Uns beschäftigt der Zusammenhang zwischen Bildung und der Internetnutzung. Gibt es ihn und falls ja, welche Richtung hat er?

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

➡

Hypothese 1: Bildung hat einen positiven Einfluss auf den zeitlichen Umfang der Internetnutzung.

1 Hypothesen aufstellen

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

Der H1 (auch *Forschungshypothese* oder *Alternativhypothese* genannt) stellen wir immer (implizit) eine *Nullhypothese* gegenüber.

Wie lautet die Nullhypothese in diesem Fall?

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

1 Weiteres Vorgehen

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_ss <- select(ess8_CH, internet = netustm, eduyrs, idno)
summary(ess8_CH_ss)
sd(ess8_CH_ss$eduyrs, na.rm = TRUE)
sd(ess8_CH_ss$internet, na.rm = TRUE)

ess8_noNA <- na.omit(ess8_CH_ss)
```

mit dem **na.omit()**-Befehl können wir einen Datensatz ohne NAs für die finale Regressionsanalyse erstellen

eduyrs: Anzahl an abgeschlossenen Bildungsjahren.

netustm: Internetnutzung in Minuten pro Tag.

Durch Entfernung von Personen mit **NA**s in einer der relevanten Variablen reduziert sich der Datensatz von 1525 auf 1184 Merkmalsträger.

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.

Wir erhalten im Output zusätzliche Informationen, u.a.:

- einen Standardfehler[1] von 1.313
- einen t-Wert von 5.875
- einen p-Wert von 0.00000000548
- ein Signifikanzniveau von *** (0.001)

[1]Achtung: Sowohl der Standardfehler als auch Standardabweichung messen Variation, beziehen sich allerdings auf unterschiedliche Ebenen: Die Standardabweichung misst Variation ***innerhalb*** der Stichprobe, der Standardfehler misst Variation ***zwischen*** verschiedenen Stichproben.

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
```

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

Mit jedem zusätzlichen Bildungsjahr steigt die Internetnutzung um 0.047 Standardabweichungen.

Wir erhalten im Output zusätzliche Informationen, u.a.:

- einen Standardfehler[1] von 1.313
- einen t-Wert von 5.875
- einen p-Wert von 0.00000000548
- ein Signifikanzniveau von *** (0.001)

[1]Achtung: Sowohl der Standardfehler als auch Standardabweichung messen Variation, beziehen sich allerdings auf unterschiedliche Ebenen: Die Standardabweichung misst Variation ***innerhalb*** der Stichprobe, der Standardfehler misst Variation ***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 **Abweichung** eines Parameters der **Stichprobe** vom wahren Parameterwert in der **Grundgesamtheit** an.

Wie interpretieren wir den Standardfehler in diesem Fall?

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
```

Das Verhältnis von Koeffizient zum 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 bzw. seinen «Sicherheitsabstand» von der 0.

Er wird in Euren Auswertungen zumeist nicht berichtet.

In den Regressionsoutputs sind SE, t- und p-Wert standardmässig integriert und müssen nicht manuell berechnet werden.

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.

Bei sehr kleinem p-Wert wird dieser von R standardmässig in Exponentialschreibweise ausgegeben.

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 vorläufig gestützt.

[1]Auch wenn die von der R postulierte Forschungshypothese *einseitig* ist, können wir den abgeleiteten p-Wert einer Konvention folgend für die Prüfung unserer *gerichteten* Hypothese verwenden. Letztlich wird hierdurch die Schwelle für die Verwerfung der H0 höher gelegt (siehe Folien letzte Vorlesung)

2.6 Praktische Übung

Wie würdet ihr bei einem Hypothesentest für den Zusammenhang von **Bildung** und **Klimaverantwortung** vorgehen?

Formuliert Forschungs- und Nullhypothese.

Untersucht empirisch: wird die Forschungshypothese gestützt?

H1: *Mehr Bildung führt zu mehr Klimaverantwortung.*

Nullhypothese: *Mehr Bildung führt **nicht** zu mehr Klimaverantwortung.*

Konfidenz- und Vorhersageintervalle

Parameterkonfidenz

Haben wir den richtigen Regressionskoeffizienten gefunden?

→ **Konfidenzintervall des Regressionskoeffizienten**
Konfidenzband der Regressionsgerade

Vorhersagekonfidenz

Wie gross ist die Streuung um (mit der Regressionsgeraden) vorhergesagte Einzelwerte?

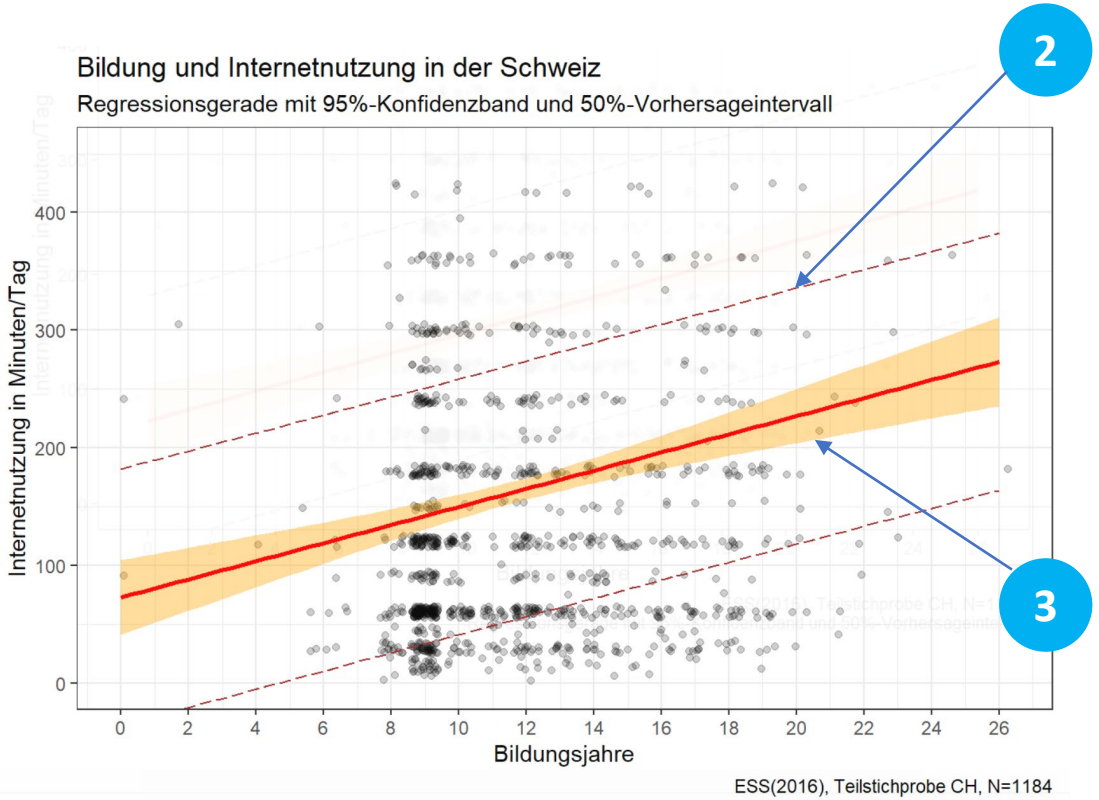
→ **Vorhersageband der Regressionsgerade**

3.1 Konfidenzintervall und- band, Vorhersageband

```
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
```

1



2

3

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 95%-Konfidenzband
3 50%-Vorhersageband

3.2 Konfidenzintervall

Das Konfidenzintervall des Koeffizienten zeigt an, zwischen welchen Werten der wahre Koeffizient in der Grundgesamtheit mit 95%-iger 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 näherungsweise mit der Daumenregel ermitteln (Koeffizient ± 2 * SE)*

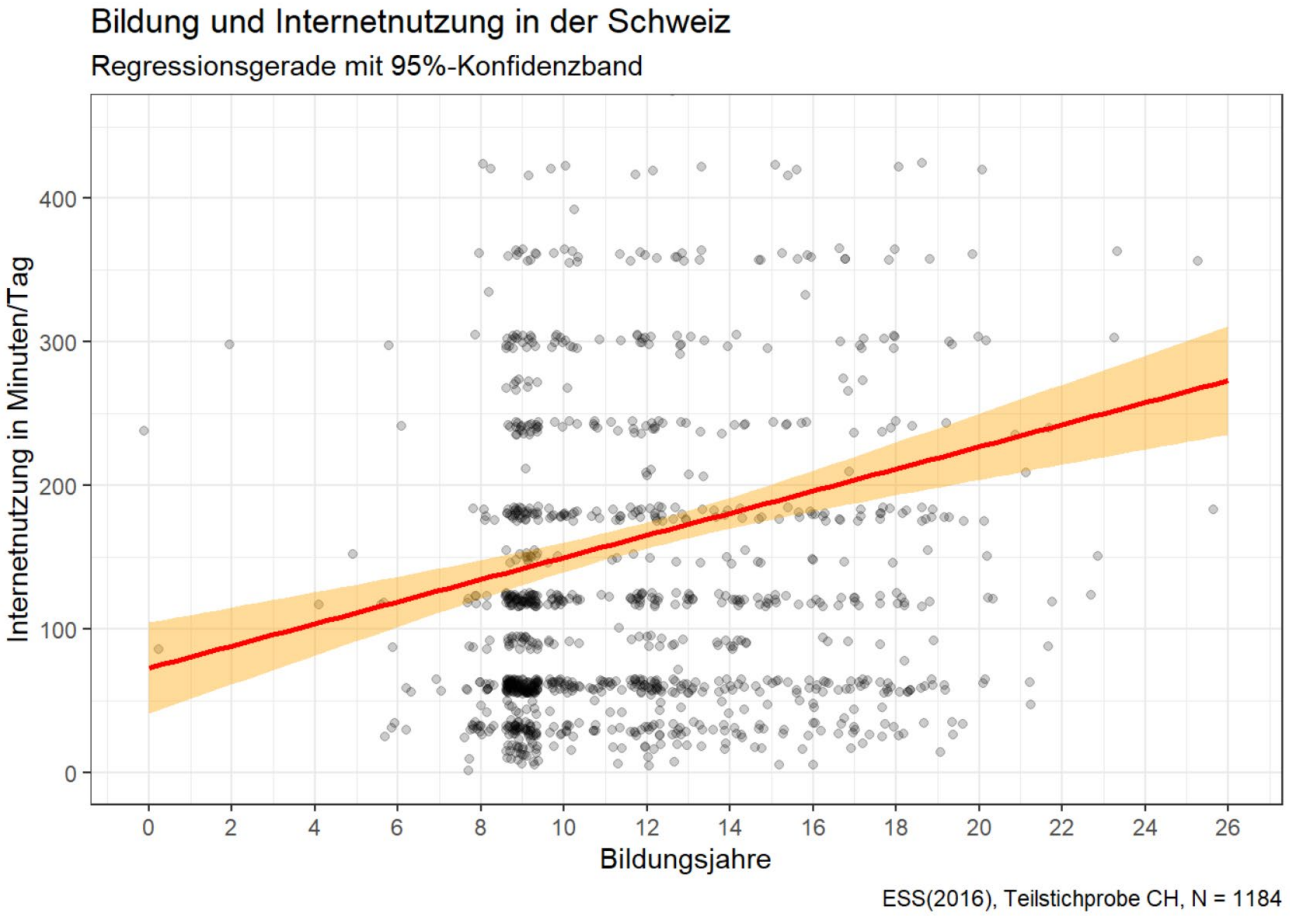
Frage: Wie könnt ihr den Output interpretieren?

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 5 min 08 und 10 min 17

3.3 Konfidenzband

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

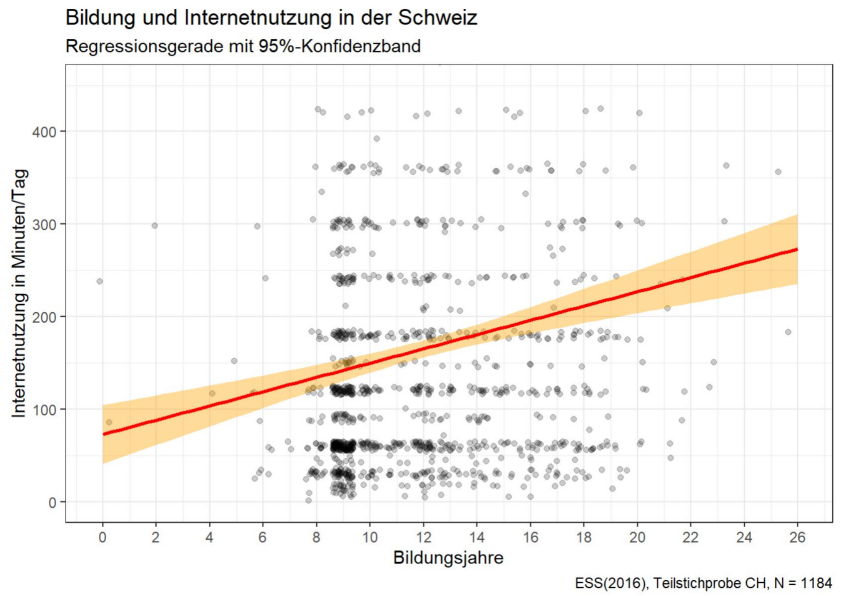


3.3 Konfidenzband

Das Konfidenzband können wir mit **ggplot()** einfach visualisieren (es ist sogar eine Default-Einstellung):

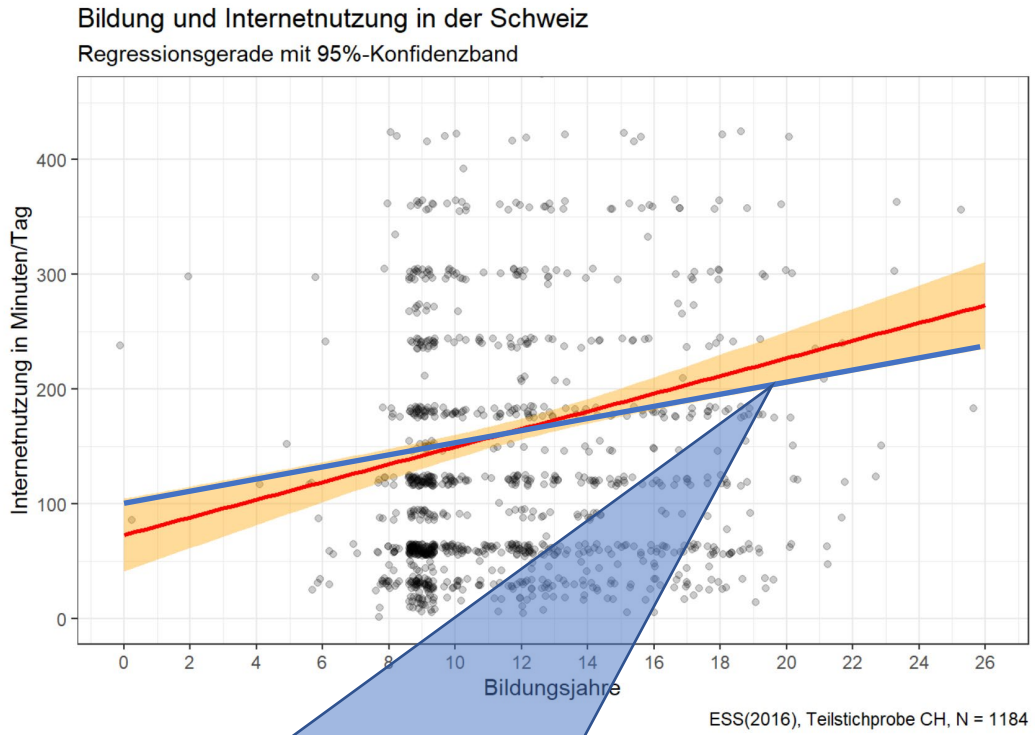
```
plot1 <- ggplot(ess8_noNA, aes(x = eduyrs, y = internet))+
  geom_jitter(alpha = 0.2, height = 5) +
  scale_x_continuous(breaks = seq(0,26,2))+
  coord_cartesian(ylim = c(0,450)) +
  geom_smooth(method = "lm", se = TRUE, color = "red", fill = "orange", level = 0.95)+
  theme_bw()+
  labs(title = "Bildung und Internetnutzung",
  y = "Internetnutzung in Minuten/Tag",
  x = "Bildungsjahre",
  caption = "ESS(2016), Teilstichprobe CH, N = 1184.\n Regressionsgerade mit
  95-Prozent-Konfidenzband.")
plot1
```

Mit **coord_cartesian()** *wählen wir einen Ausschnitt aus dem ggplot, ohne dass Werte ausserhalb dieses Bereichs für die Berechnung der Regressionsgerade ausgeschlossen werden. (Achtung: diese Eigenschaft haben andere Befehle zur Einschränkung des Ausschnittes (z.B. xlim) nicht!)*



3.3 Konfidenzband

Je schmaler das Konfidenzband, desto höher das Vertrauen in die ermittelte Regressionsgerade

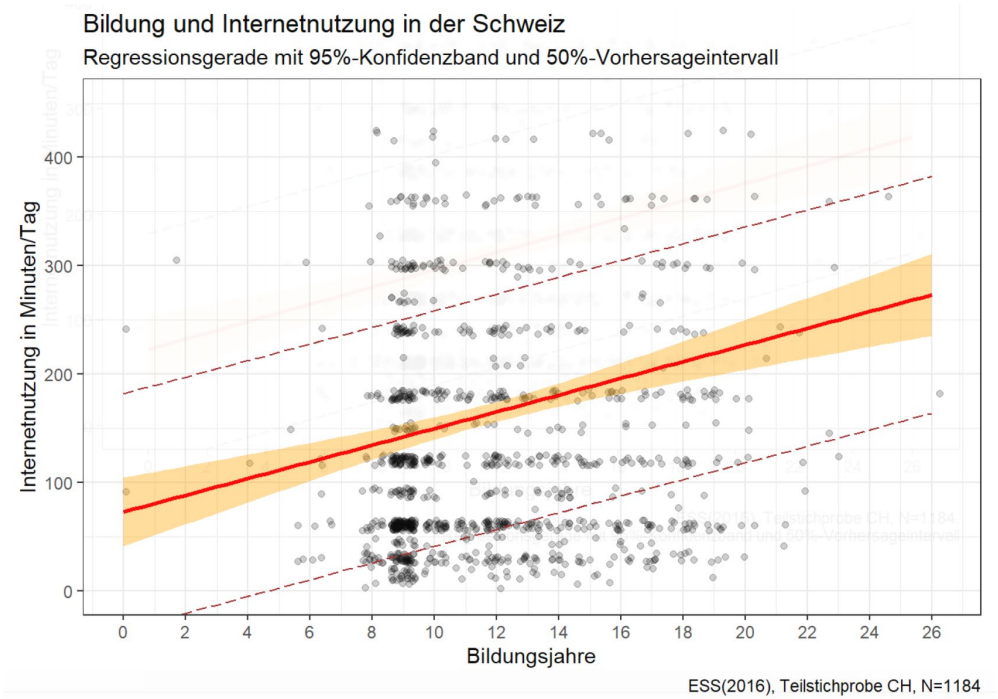


Hier hohes Vertrauen: Selbst wenn sich die tatsächliche Regressionsgerade am untere Rand des Sicherheitsbereichs befindet deutet sie noch immer auf einen substantiellen Zusammenhang hin.

3.4 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

Achtung: Für das Vorhersageband gibt es **keine Standardoption** im Rahmen des ggplot

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

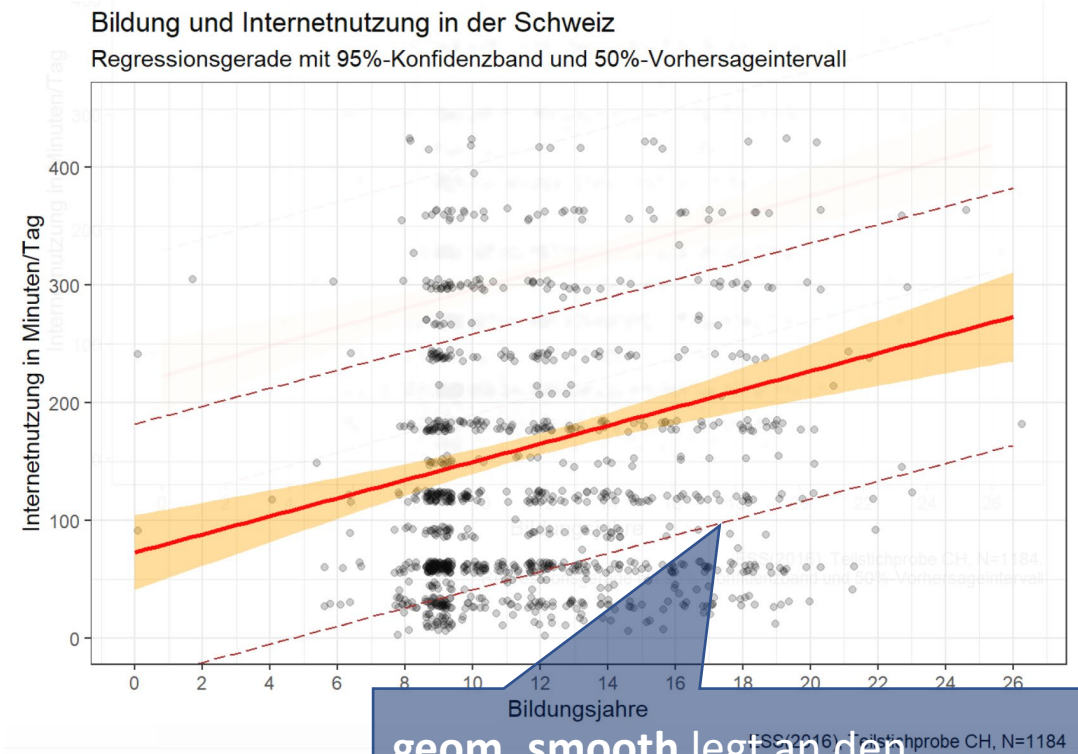
```
predictions <- ggpredict(fit, terms = "eduyrs", interval = "prediction", ci.lvl = 0.5)
```

	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

3.4 Vorhersageband

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

```
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
```



geom_smooth legt an den Grenzwerten des Vorhersagebandes orientierte, optimierte Kurven in den Plot.

3.4 Vorhersageband → Vorhersageintervall

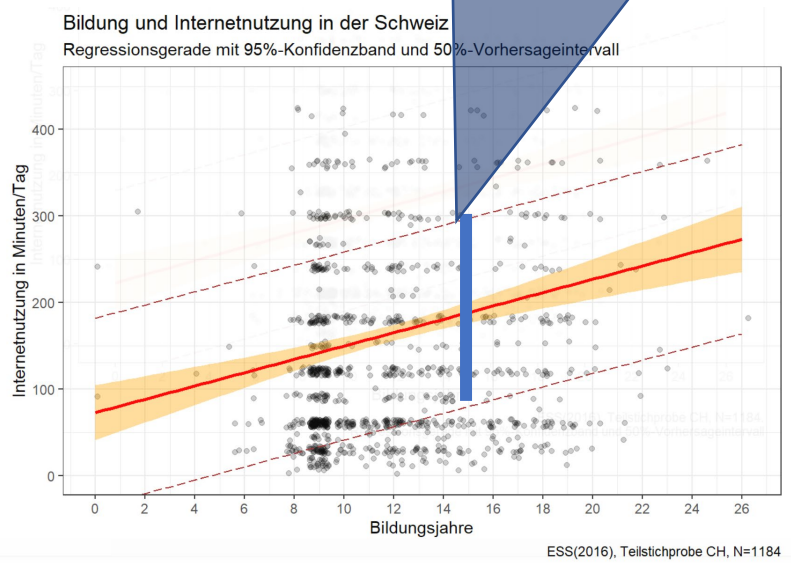
Die Breite des Vorhersagebandes an einer bestimmten Stelle kann ebenfalls mit dem **ggpredict()**-Befehl berechnet werden

```
> ggpredict(fit, terms = "eduyrs[13]", interval = "prediction", ci.lvl = 0.5)
# Predicted values of Internet use, how much time on typical day, in minutes

eduyrs | Predicted |          50% CI
-------------------------------------
    13 |    172.91 | [64.23, 281.60]
```

???

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



Frage: Wie könnt ihr das **abgeleitete 50% Vorhersageintervall** interpretieren?

Der tägliche Nutzungswert für eine Person mit 13 Bildungsjahren liegt mit 50% Sicherheit zwischen 64.23 und 281.60 Minuten.

Frage: Wie würde ein **90% Vorhersageband** aussehen?

3.5 Praktische Übung

Betrachtet das Konfidenzintervall, Konfidenzband und Vorhersageband für den Regressionsoutput von **Bildung** und **Klimaverantwortung** an.

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: Wie fällt der Wert der Klimaverantwortung für eine Person mit 9 Bildungsjahren mit 50% Sicherheit aus?

Hausaufgabe mit Selbstüberprüfung:

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