08 Regressionen

Einführung in die quantitativen Forschungsmethoden

Heute

- Regressionen
- Wie gut ist unsere Vorhersage?
- Annahmen bei Regressionen

Regressionen

Regressionen

- *statistische Methode um Beziehung zwischen einer abhängigen und (einer oder mehreren) unabhängigen Variablen zu modellieren*

Zwei **Nutzen** von Regressionen:

- um Vorhersagen über eine Variable aus anderen Variablen zu treffen (heute)
  - z.B. Licht bei Nacht & GDP - siehe Llaudet & Imai
- um den Einfluss mehrerer Variablen auf eine andere Variable zu quantifizieren (Schwerpunkt nächste Woche)
  - z.B. Bildungsgrad & Einkommen

Zwei Schritte: **Schätzung eines Modells** & **Vorhersage**

Regressionsanalyse

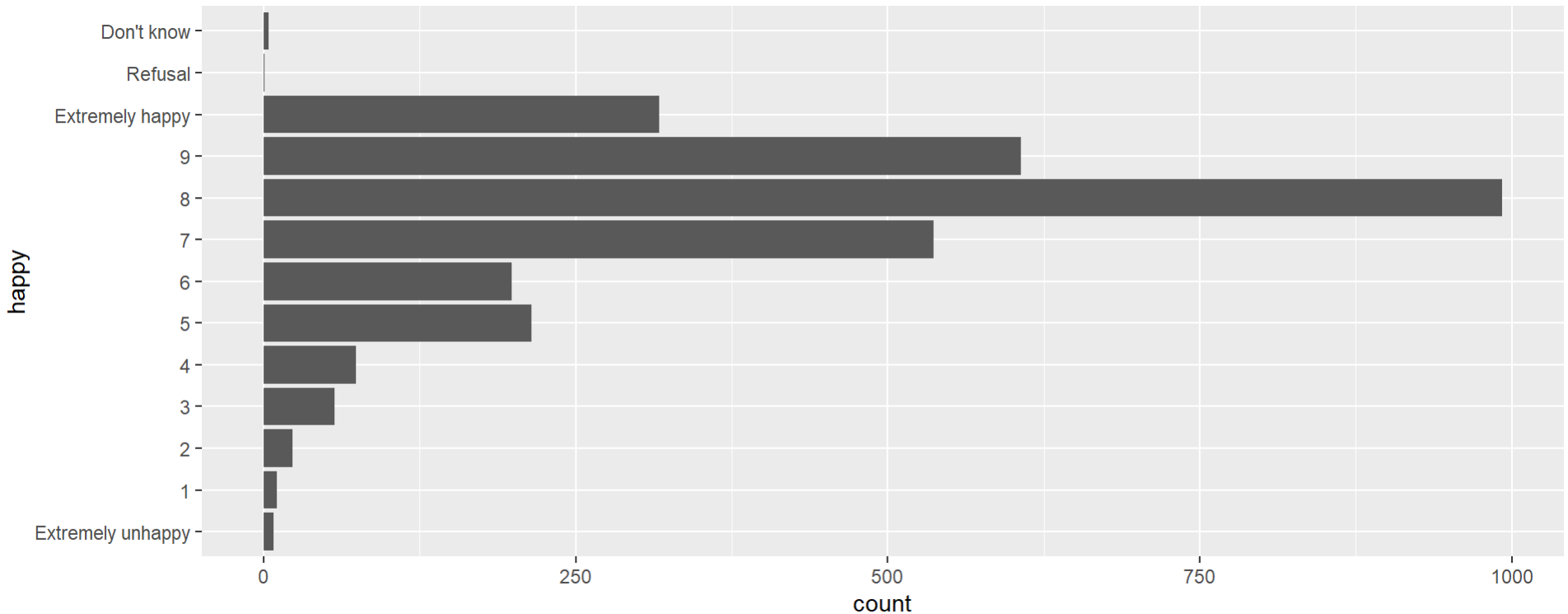
- **abhängige Variable** wird auf **unabhängige Variable(n)** zurückgeführt (‘regrediert’)
  - Einfluss von Variable x auf y
  - Vorhersage von y durch x

Beispiel Forschungsfrage

Macht Geld glücklich? (Und wieviel Geld braucht man um ‘extremely happy’ zu sein?)

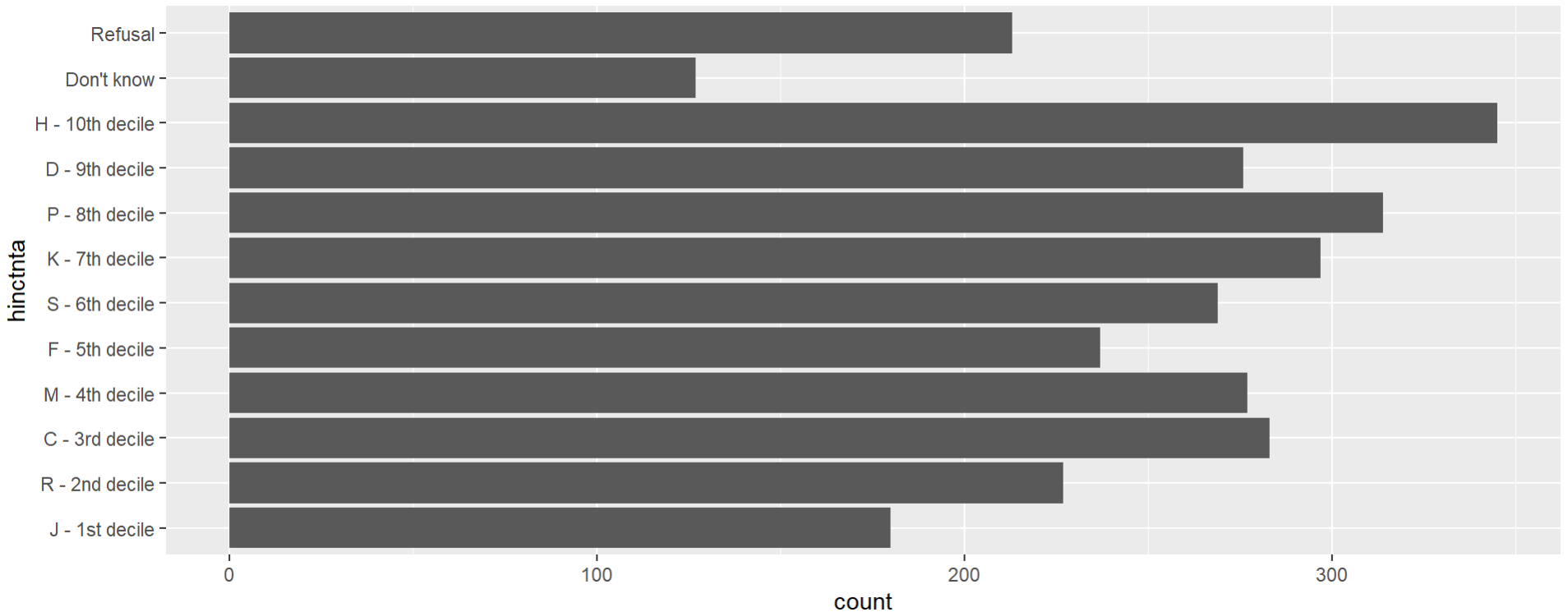
happy

Alles in allem betrachtet, was würden Sie sagen, wie glücklich sind Sie? (äußerst unglücklich … äußerst glücklich)



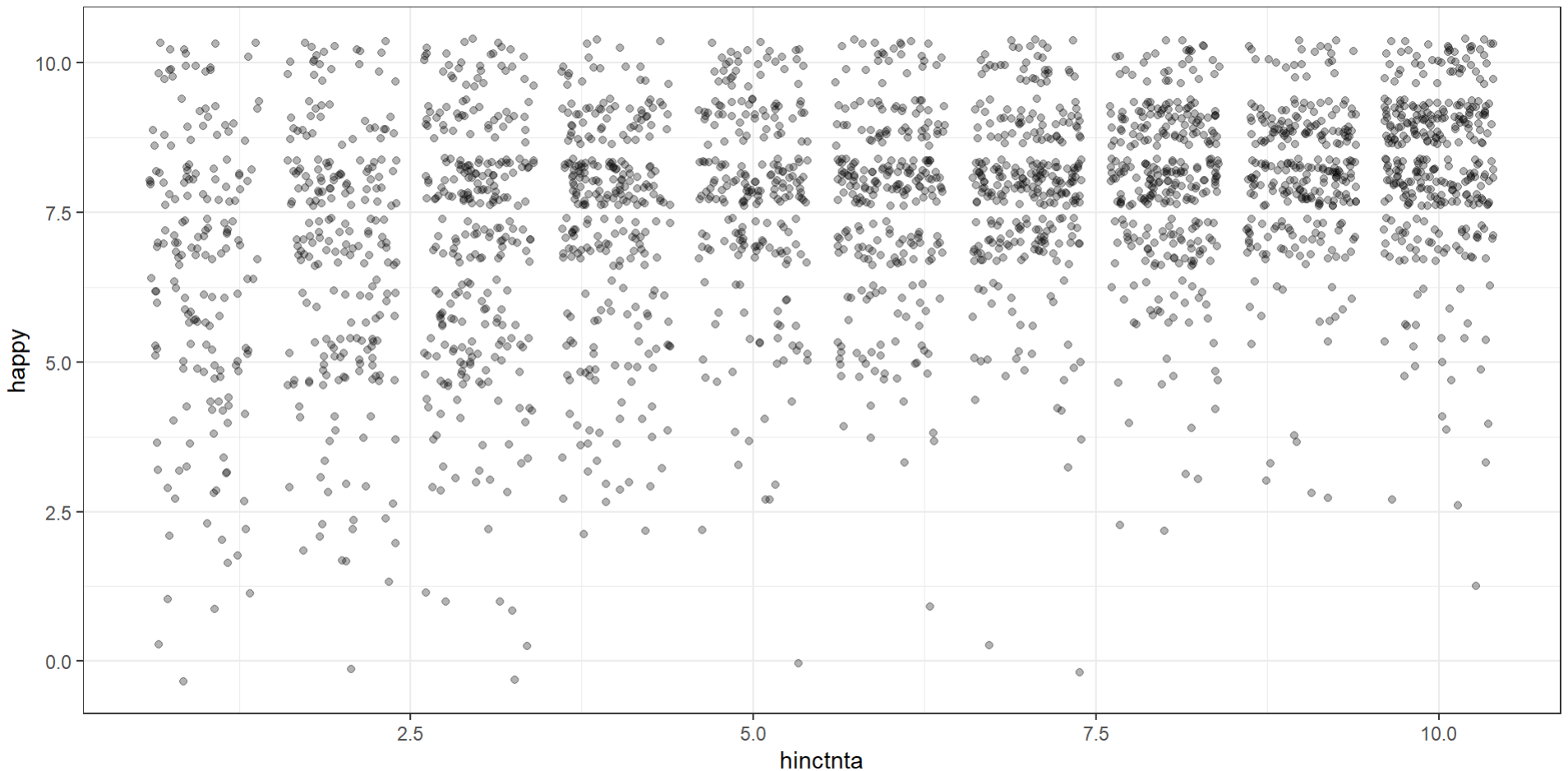
hinctnta

Wenn Sie die Einkommen aus allen Quellen zusammenzählen: Welcher Buchstabe auf Liste 58 trifft für das gesamte Nettoeinkommen Ihres Haushalts zu? (Zehn Bereiche)



Macht Geld glücklich?

Scatterplot (zur Ermittlung der Beziehung)

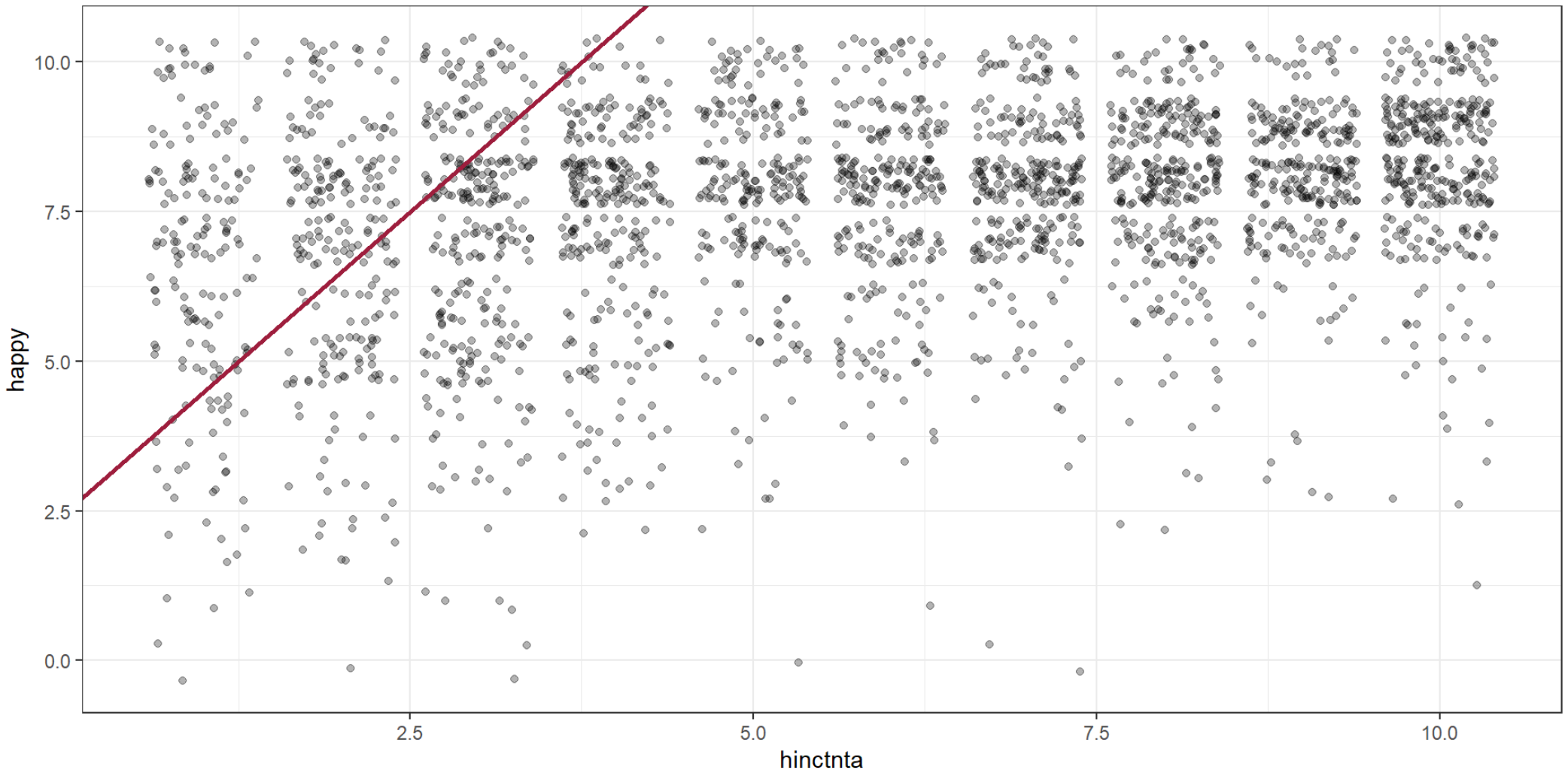


Macht Geld glücklich?

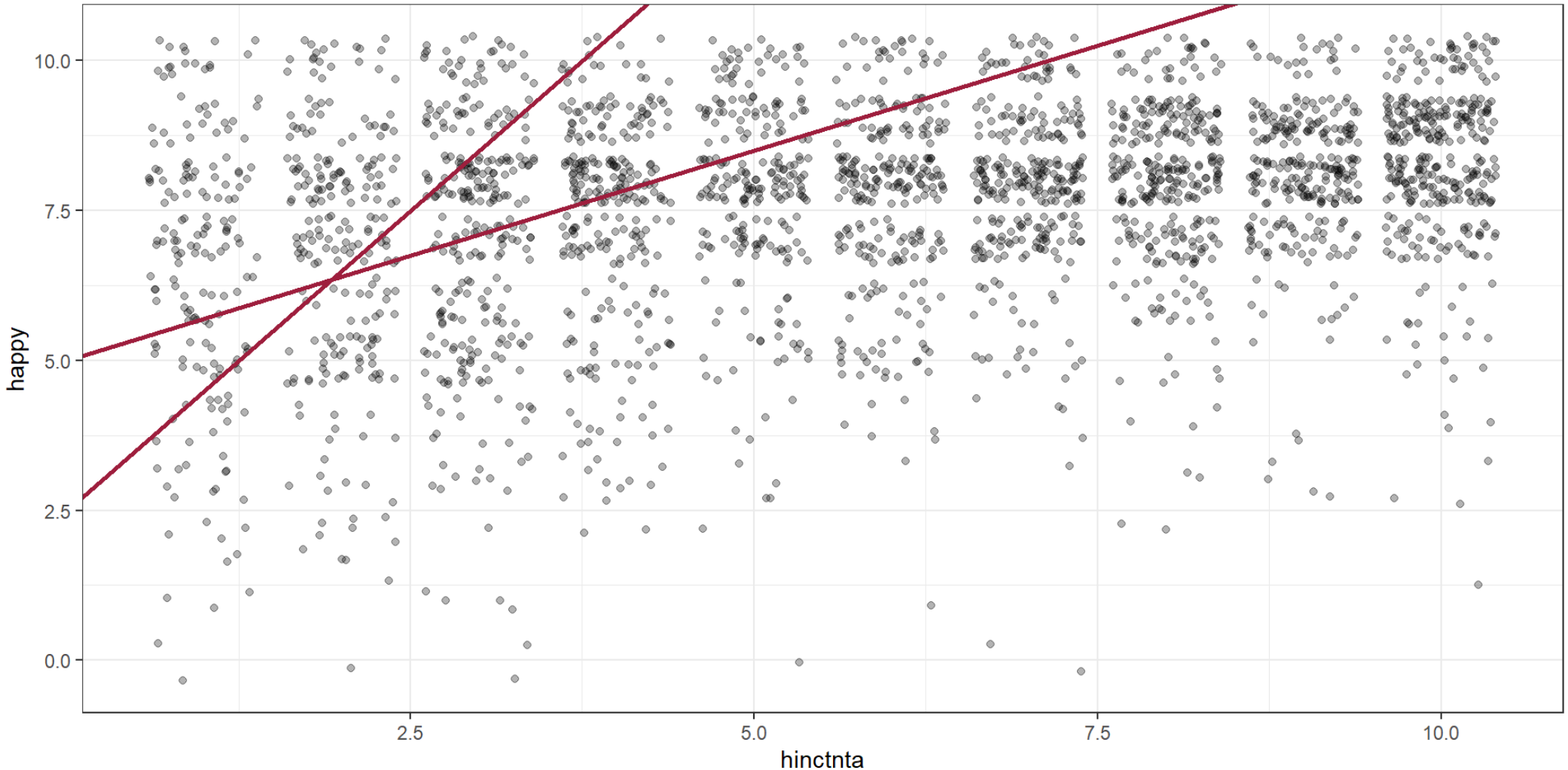
→ Welche (Regressions-)Gerade verknüpft X- und Y-Werte?

→ Wieviel Geld sagt ein bestimmtes Maß an Glück voraus?

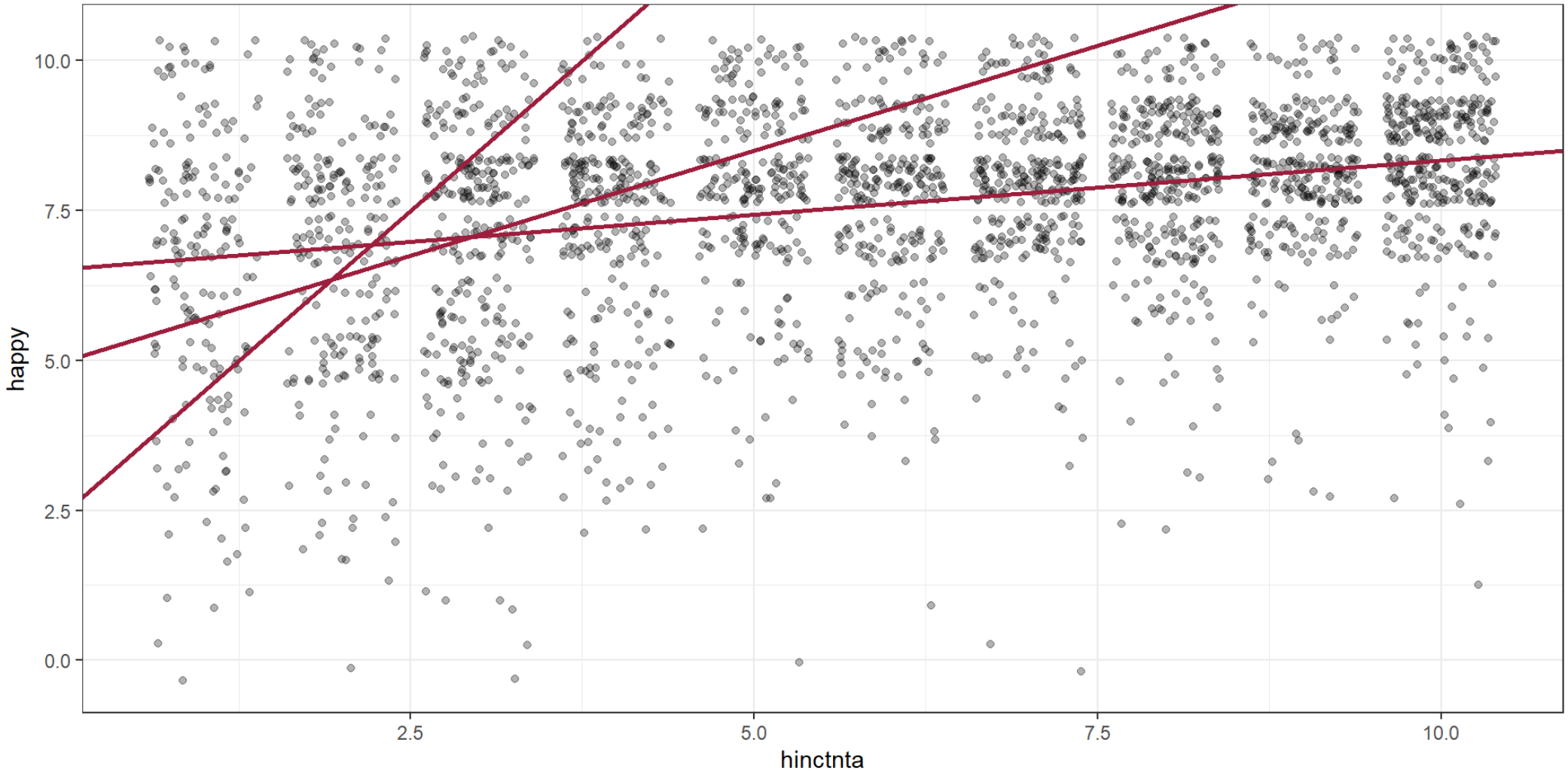
Macht Geld glücklich?



Macht Geld glücklich?



Macht Geld glücklich?

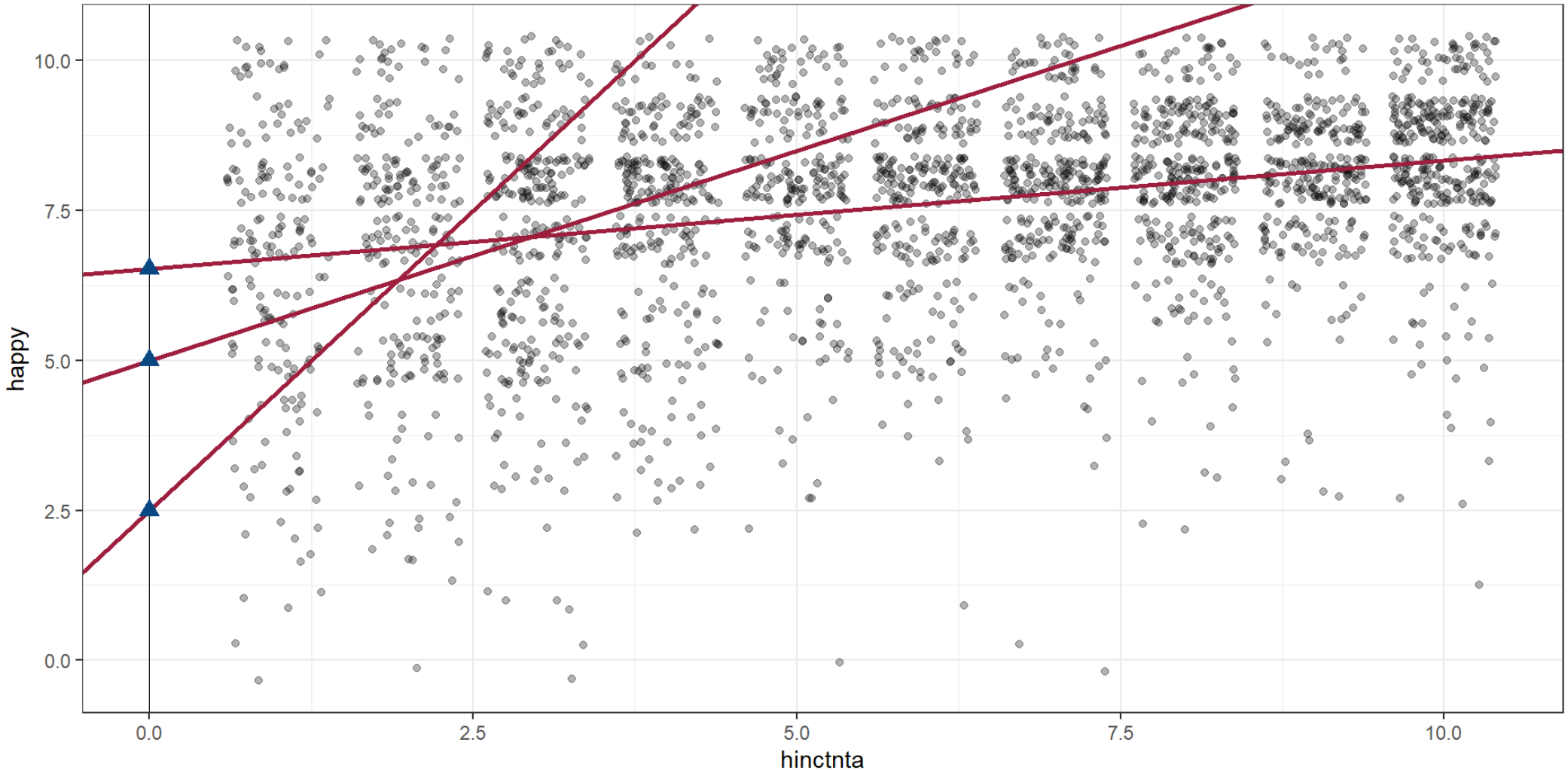


(Regressions)gerade

Konstante / Intercept

- vertikale Spezifizierung der Linie
- Wert, wenn X null ist
  - typischerweise: Y-Wert, wenn die Linie die X-Achse schneidet

Konstante / Intercept



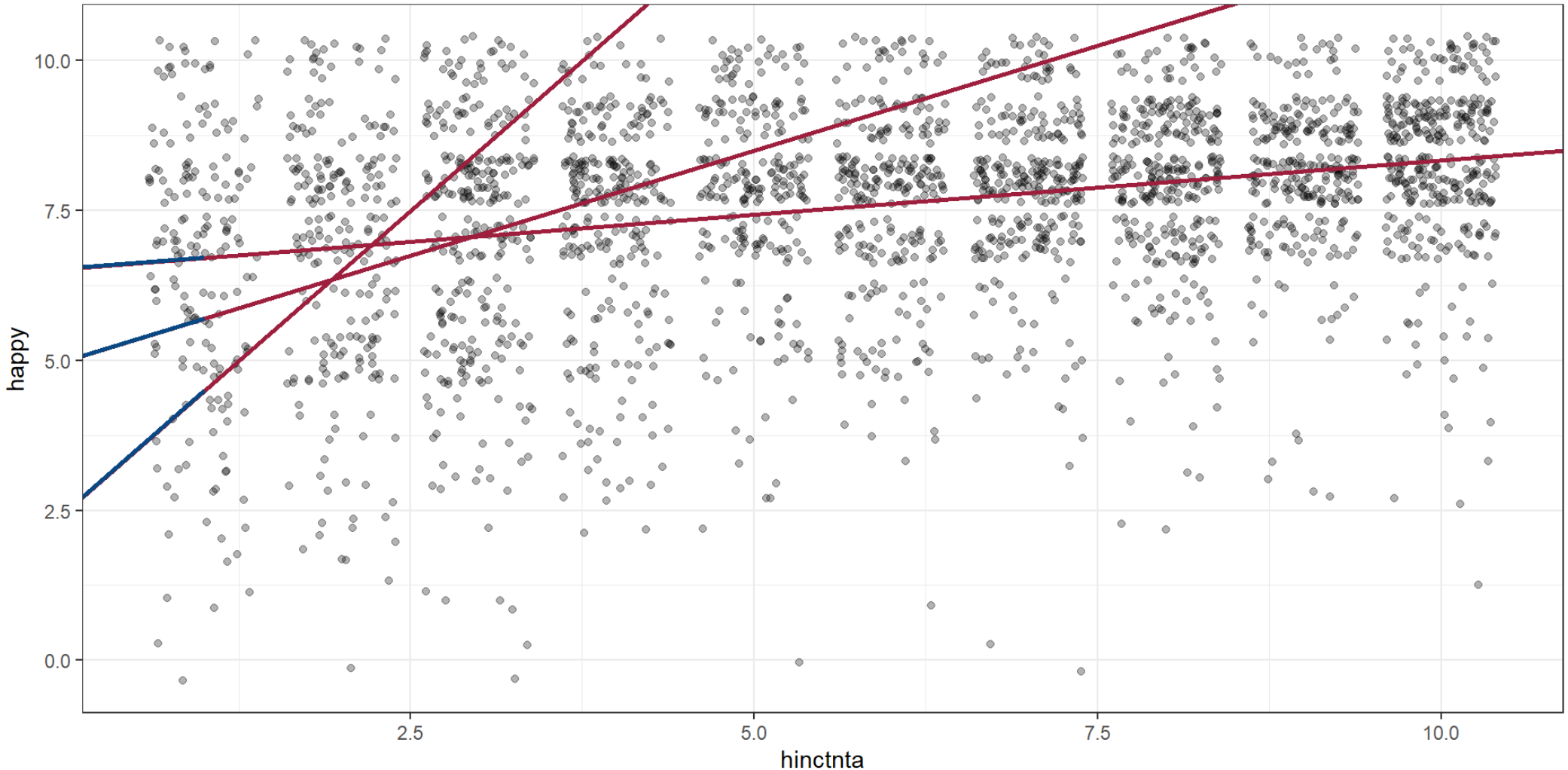
(Regressions)gerade

Steigung / Slope

- die Steigung charakterisiert den Winkel der Regressionslinie
- Änderung des Y-Werts über (1 Einheit) Änderung des X-Werts

$$\hat{\beta} = \frac{rise}{run} = \frac{\Delta \hat{Y}}{\Delta X}$$

Steigung / Slope



Mathematisches Modell

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- Y_i: abhängige Variable für Observation i
- α: Konstante
- $beta$: Steigung ('slope')
- X_i: unabhängige Variable für Observation i
- ϵ_i: Fehler

$$\mathrm{happy_i} = \alpha + \beta(\mathrm{hinctnta_i}) + \epsilon$$

Mathematisches Modell

Wir suchen die Linie, die die Relation von X & Y zusammenfasst (‘*Modell schätzen*’)

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

- $\hat{Y}_i$: vorhergesagter Wert von Y an Position i
- $\hat{\alpha}$: vorhergesagte Konstante
- $\hat{\beta}$: vorhergesagte Steigung in Relation zur unabhängigen Variable X an Position i

Regressionen in R

In R können wir Regressionen mit `lm()` berechnen:

```
1  lm1 <- lm(happy~hinctnta,data=ess_sample)
2  lm1
```

```
Call:
lm(formula = happy ~ hinctnta, data = ess_sample)

Coefficients:
(Intercept)     hinctnta
     6.5290       0.1806
```

- intercept / Konstante
- Steigung / Koeffizient für hinctnta

→ Aber was steht hinter diesen Werten?

Wie erhalten wir eine gute Vorhersage?

Es gibt verschiedene Methoden, die beste Linie zu schätzen

Idee: Minimierung der Abweichung der Vorhersagen von den vorliegenden Beobachtungen

→ *…Aber was heißt schon minimieren?*

Mathematisches Modell: Residuen

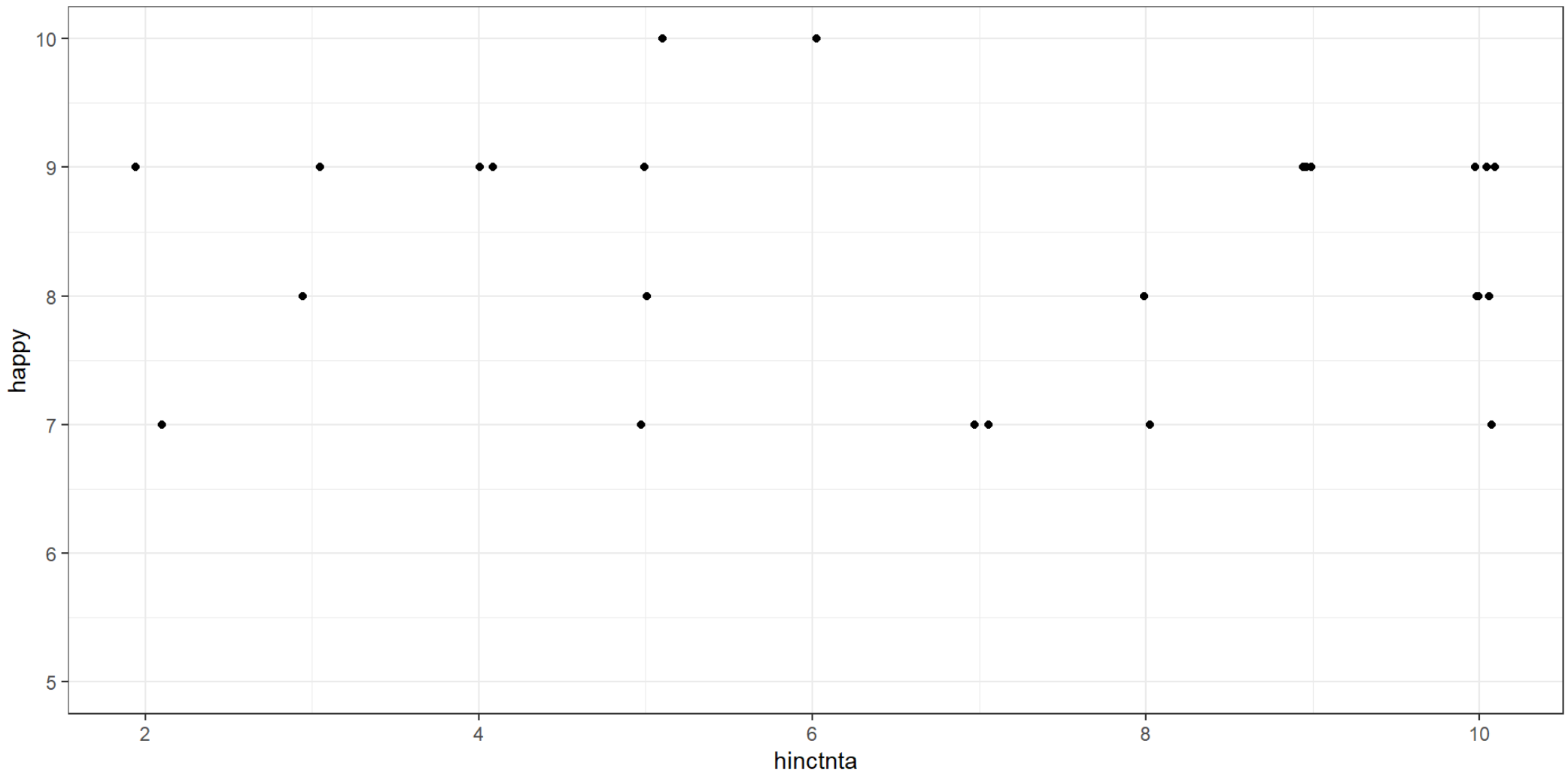
Residuen: Abweichungen der realen Werte von den vorhergesagten Werten ('geschätzter Fehler')

→ Differenz zwischen geschätztem & beobachtetem Y für Observation i:

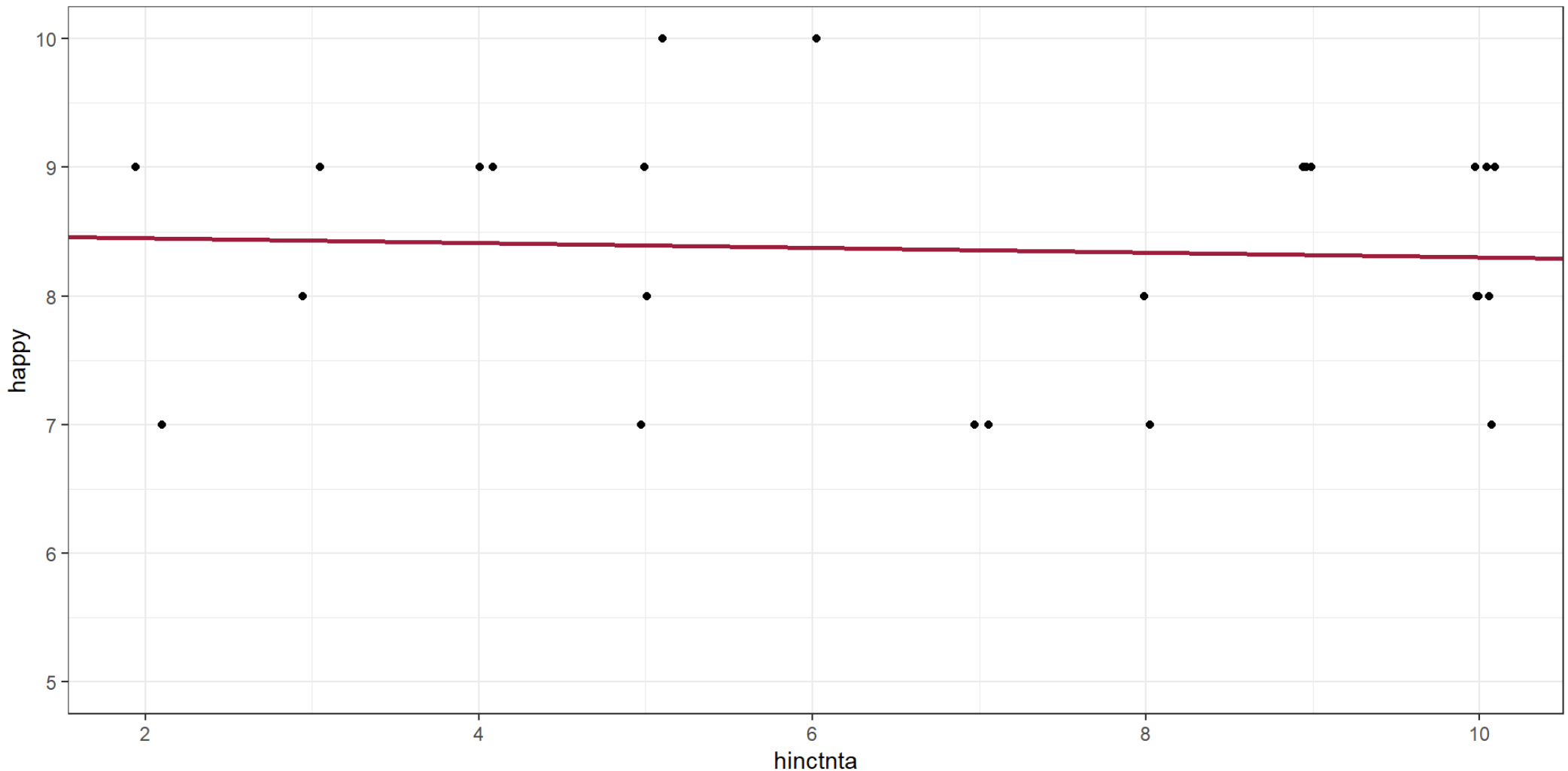
$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

→ die geschätzten Werte liegen auf der Regressionsgeraden

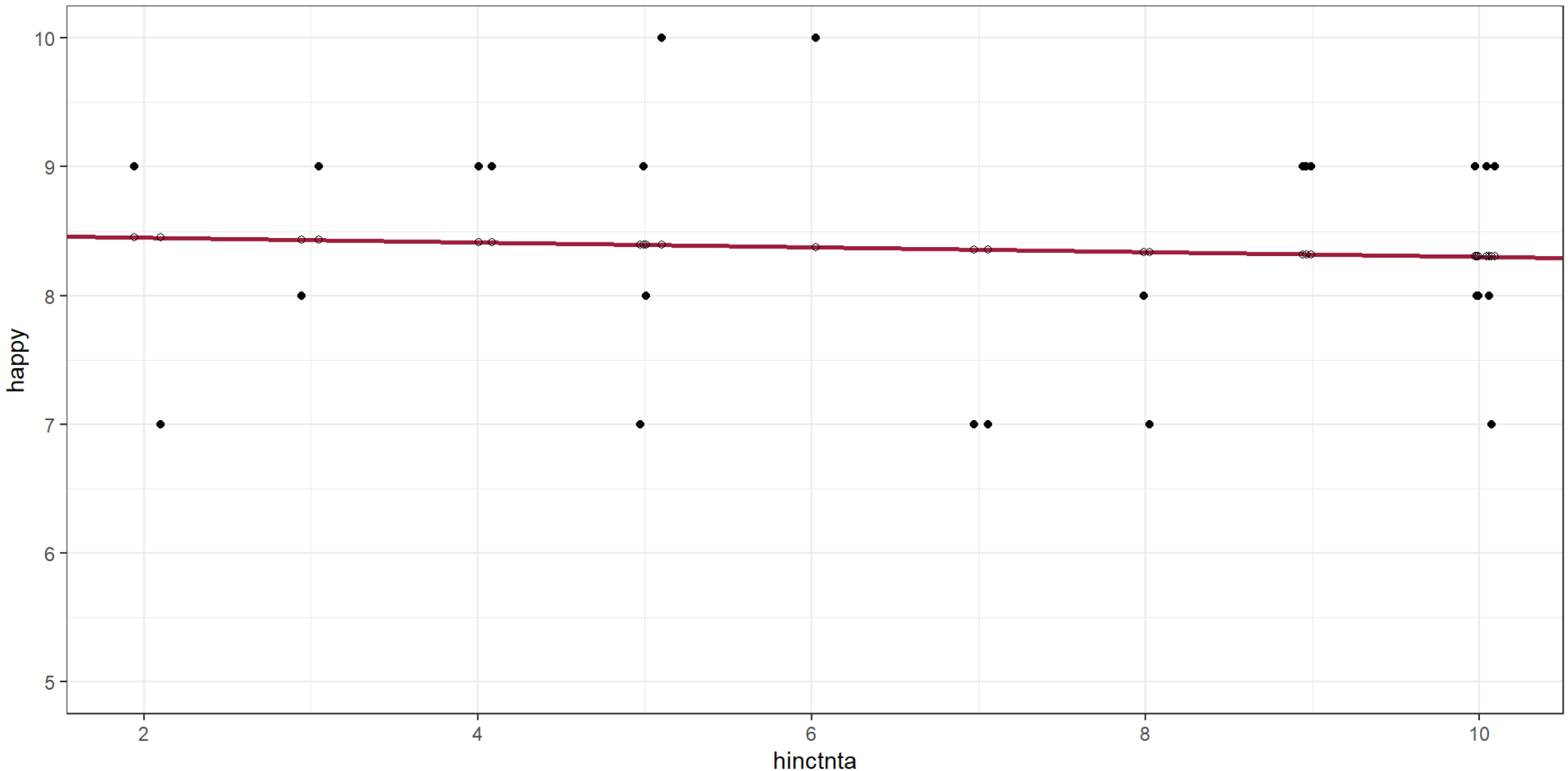
Beispiel



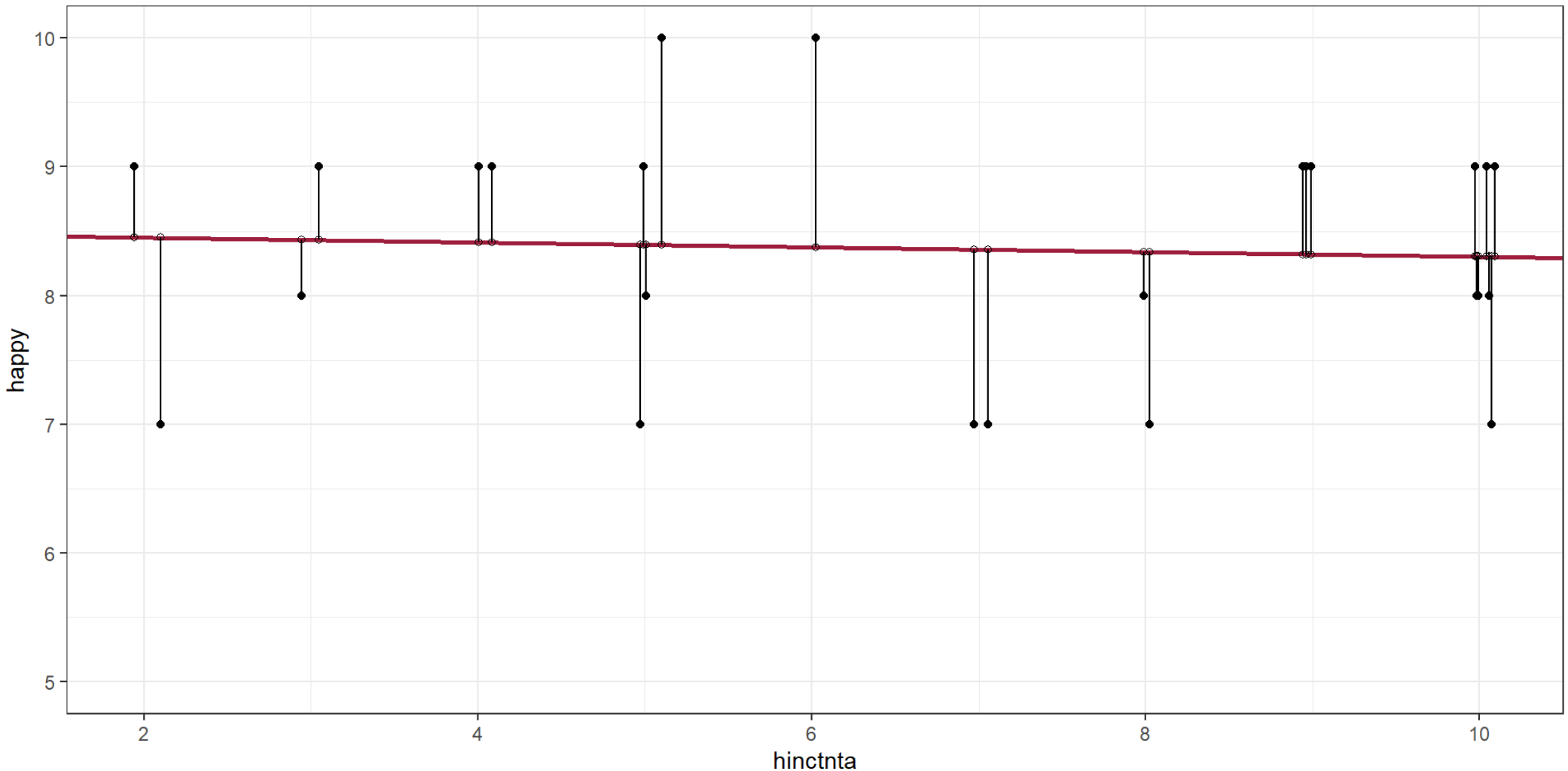
Beispiel



Beispiel



Beispiel



Mathematisches Modell

Ordinary Least Squares (OLS) / Kleinste-Quadrate-Methode: Gleichung, die die Summe der quadrierten Residuale minimiert (ausführliches Berechnungsvideo)

→ **Mathematische Berechnung der Steigung und der Konstante der Regressionsgeraden**

Detail: Vorteile der Quadrierung

- symmetrische Werte für positive und negative Abweichungen
- stärkere Bestrafung starker Abweichungen

. . .

→ häufigste Definition von Minimierung der Abstände

Wie gut ist unsere Vorhersage?

Wie gut ist unsere Vorhersage?

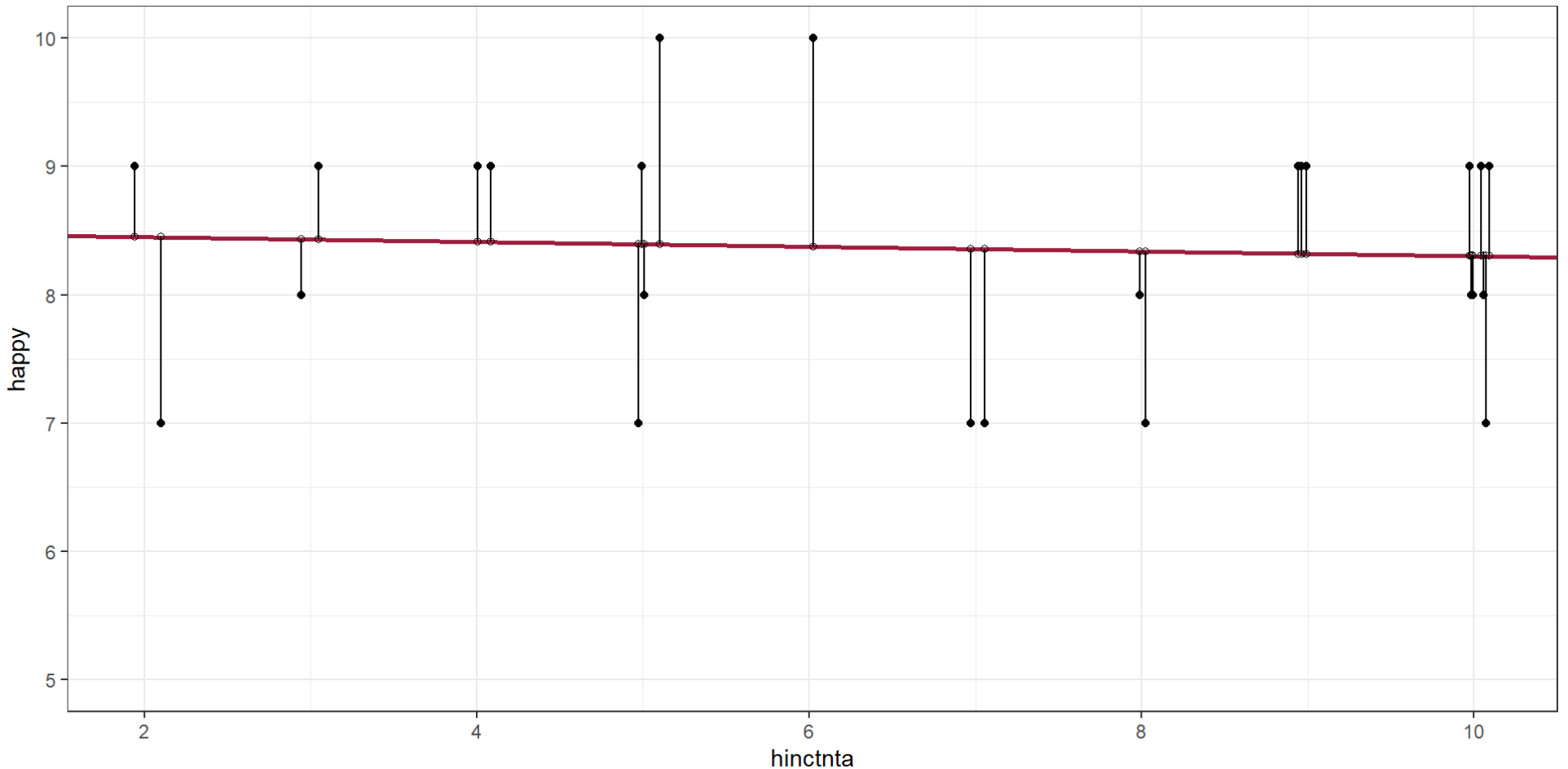
Auch zur Güte einer Regression gibt es ein standardisiertes Maß: Den **Determinationskoeffizienten** R^2

$$R^2 = \frac{erklärte\ Variation}{gesamte\ Variation} = 1 - \frac{unerklärte\ Variation}{gesamte\ Variation}$$

→ Berechnungen über Summe der quadrierten Residuen . . .

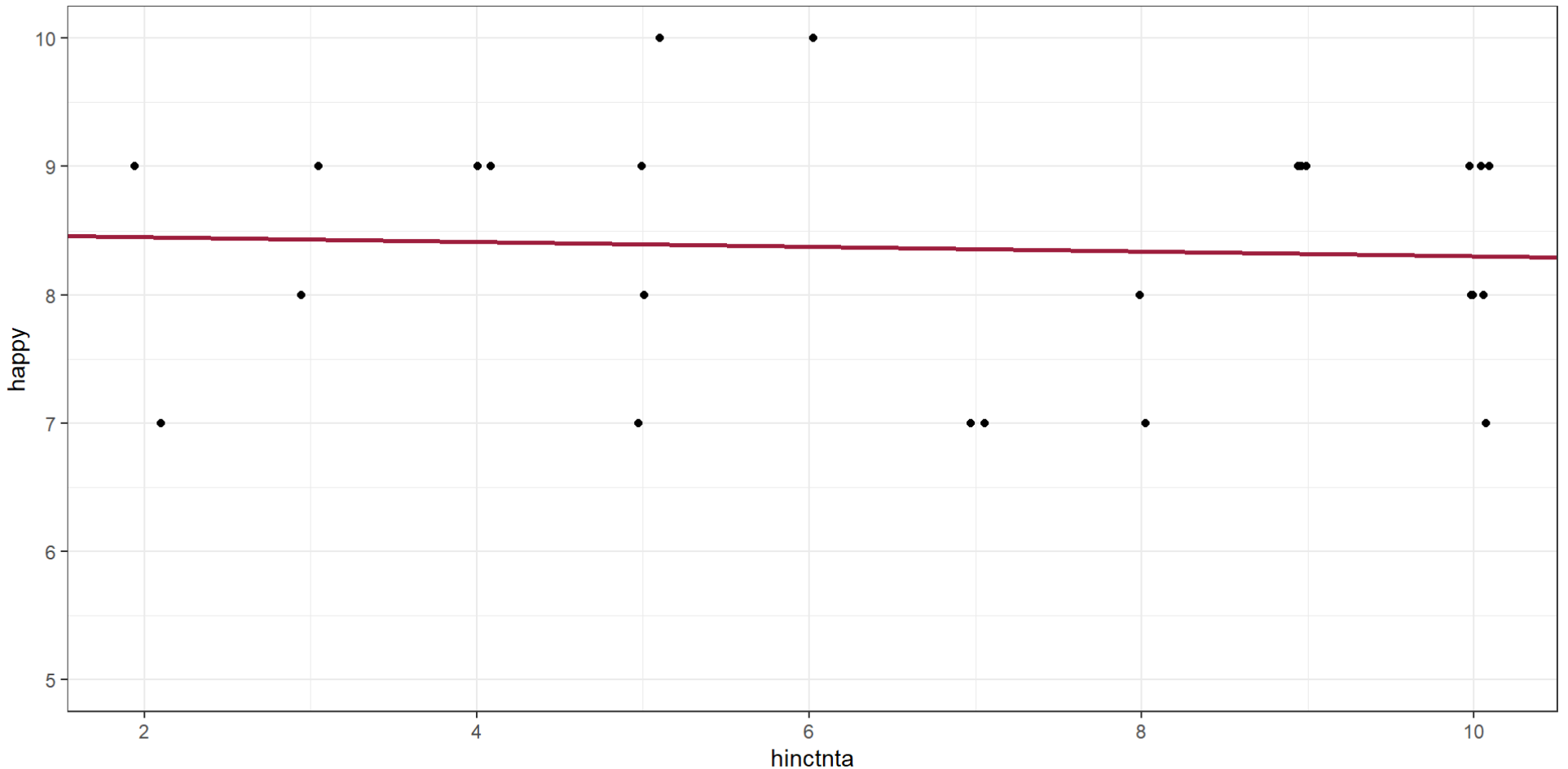
$$R^2 = 1 - \frac{\sum i(Y_i - \hat{Y}_i)^2}{\sum i(Y_i - \bar{Y}_i)^2}$$

Beispiel

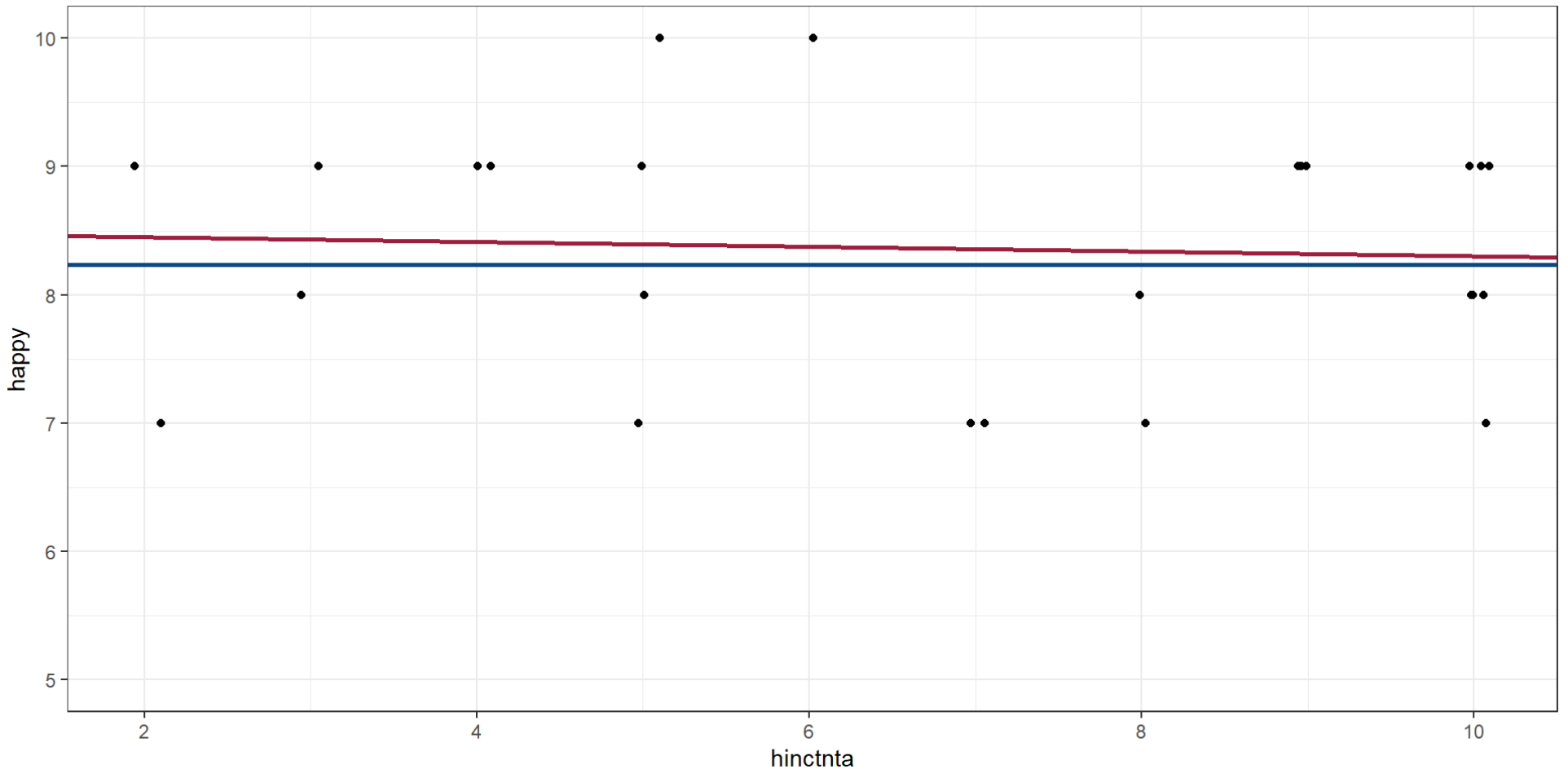


Unerklärte Variation: Abweichungen von der Regressionsgeraden

Beispiel

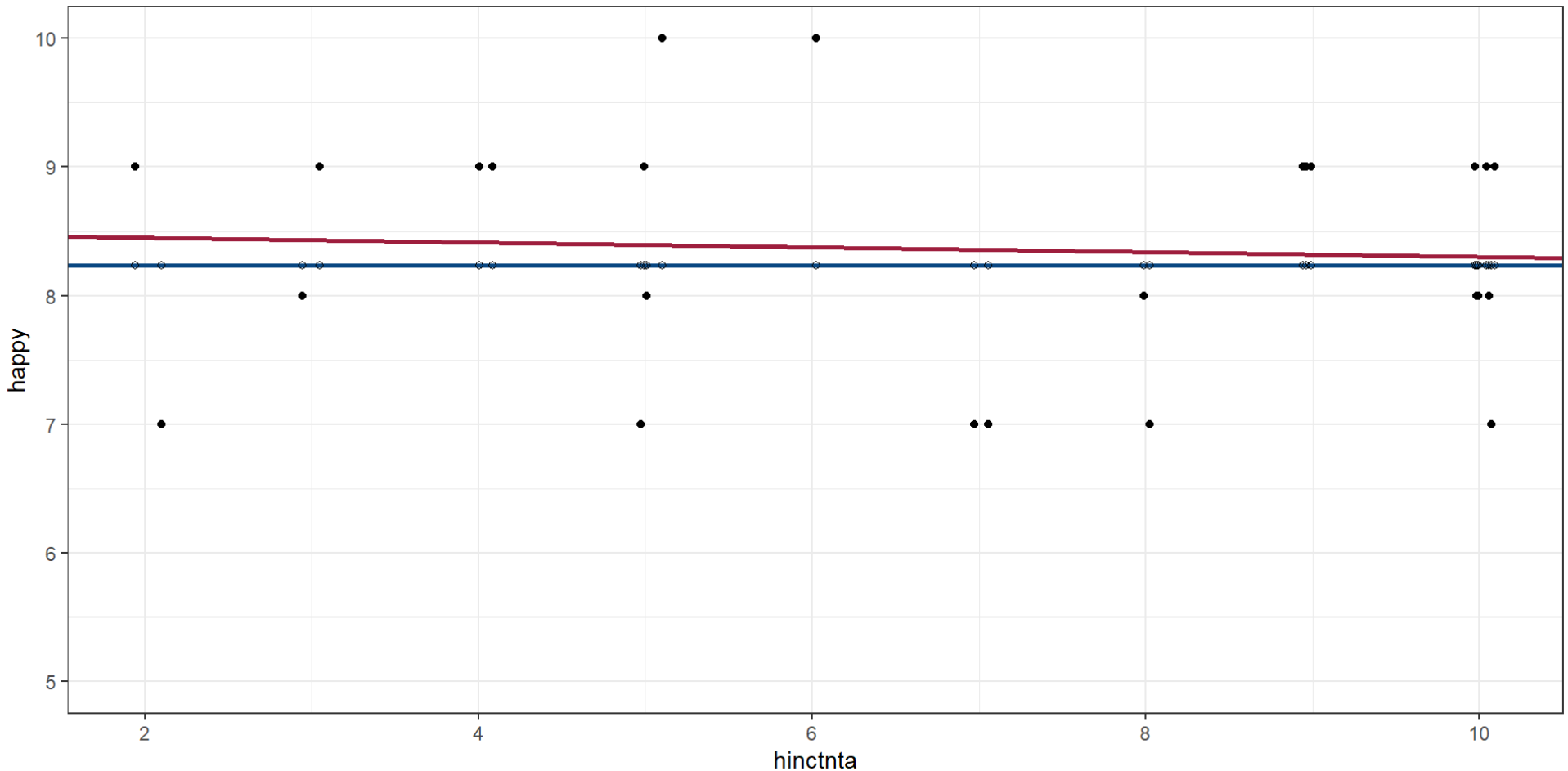


Beispiel

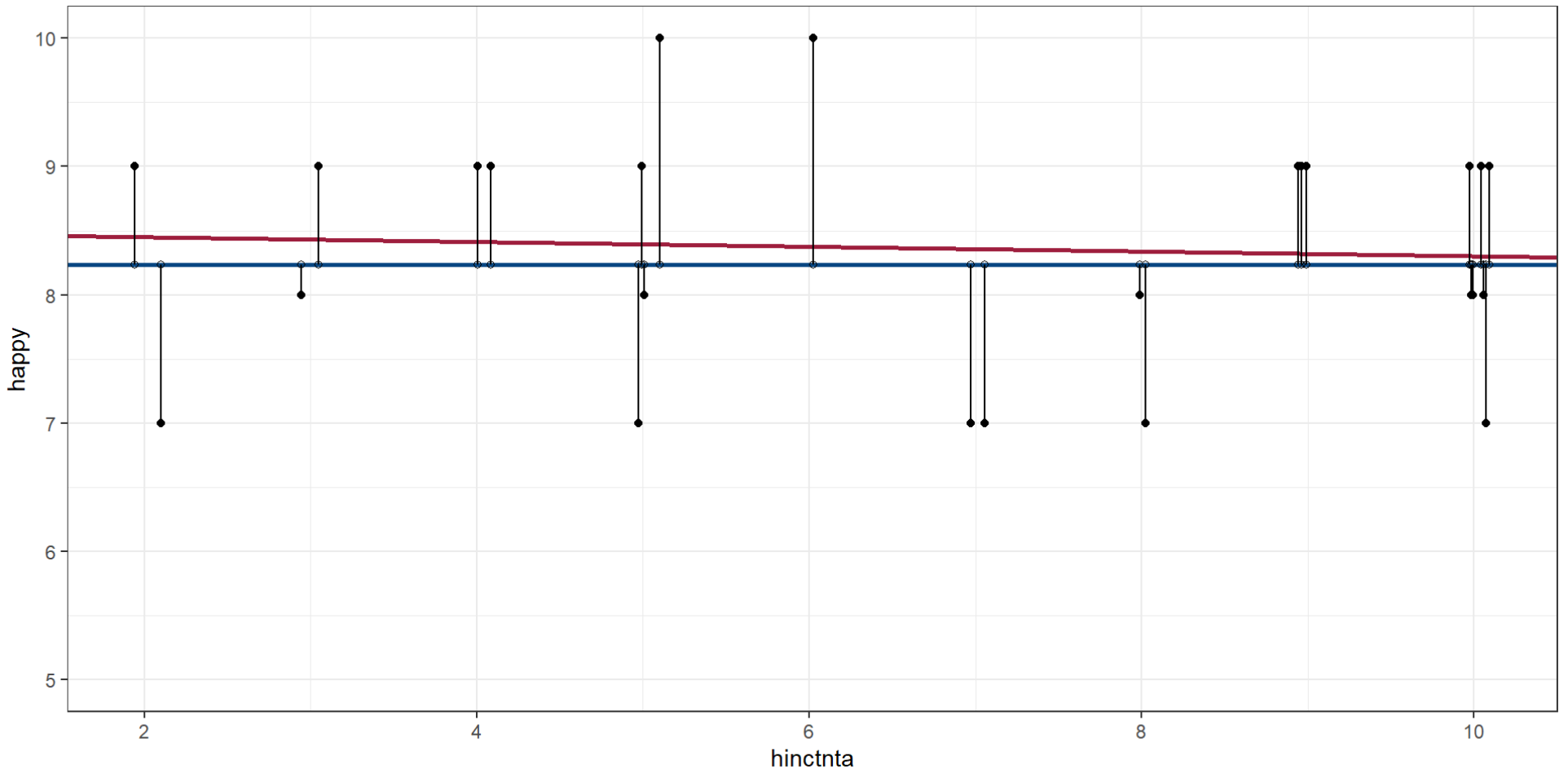


Mittelwert-Gerade

Beispiel



Beispiel



Abweichungen vom Mittelwert (‘gesamte Variation’)

Interpretation

Wir interpretieren R^2 als **den Anteil der Variation, die durch unsere Regression erklärt wird**

Regressionen in R

```r
lm1 <- lm(happy~hinctnta,data=ess_sample)
lm1
```

```
Call:
lm(formula = happy ~ hinctnta, data = ess_sample)

Coefficients:
(Intercept)     hinctnta
     6.5290       0.1806
```

Regressionen in R

```
1  summary(lm1)
```

```
Call:
lm(formula = happy ~ hinctnta, data = ess_sample)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7931 -0.7931  0.2069  1.0263  3.2904

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.52900    0.07467   87.44   <2e-16 ***
hinctnta     0.18058    0.01144   15.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.677 on 2700 degrees of freedom
  (343 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.08446,   Adjusted R-squared:  0.08412
F-statistic: 249.1 on 1 and 2700 DF,  p-value: < 2.2e-16
```

→ ‘R-squared’

Regressionen in R

Aus der `summary()` Funktion können wir noch weitere Informationen entnehmen - wir werden uns damit in den nächsten Wochen noch weiter beschäftigen!

Übung 1

Bearbeiten Sie den ersten (und zweiten?) Teil des Übungsskripts.

Annahmen bei Regressionen

Annahmen bei Regressionen

Annahmen bei Regressionen

- Format der abhängigen Variable: kontinuierlich
  - → unsere Vorhersage von Glück durch Einkommen ist mathematisch zweifelhaft, weil Glück nur wenige Werte hat / nicht kontinuierlich ist
- in der Praxis werden Variablen mit vielen Leveln (ab 7-10) und Normalverteilung (→ Sitzung 9) aber oft als (beinahe-)kontinuierlich behandelt
- Alternative: spezielle Regressionsmodelle
- Alternative: Umrechnung auf 0/1 (binär / ‘dummy’) und berechnen der Wahrscheinlichkeit für Wert 1

Annahmen bei Regressionen

- Format der unabhängigen Variable(n): verschiedene Formate möglich
  - z.B. kontinuierlich, diskret, …
- Art des Zusammenhangs: geradlinig / linear
  - alle Vorhersagewerte liegen auf einer Regressionsgeraden

Regression vs. Korrelation

- Regression und Korrelation sind verwandt im Bezug auf die Beziehung von Variablen
  - Korrelation: **Zusammenhang**
  - Regression: **Vorhersage**, typischerweise mit Ursache-Wirkungs-Annahme im Hintergrund
- nächste Woche: Einbeziehung verschiedener Variablen in die Regression (‘multiple Regression’)

Nächste Woche: Multivariate Regressionen

Thema: Wie können wir Zusammenhänge zwischen Variablen beschreiben, wenn es mehrere Einflussfaktoren gibt?

- Llaudet and Imai (2023), 129-161
- optional zur Vertiefung: Urban and Mayerl (2011)
- optional zur Vertiefung: Regressionsanalyse in R (Uni Zürich)

Referenzen

Llaudet, Elena, and Kosuke Imai. 2023. *Data Analysis for Social Science: A Friendly and Practical Introduction*. Princeton: Princeton University Press.

Urban, Dieter, and Jochen Mayerl. 2011. *Regressionsanalyse: Theorie, Technik und Anwendung: Lehrbuch ; Neu: jetzt auch mit logistischer Regression*. 4., überarb. und erw. Aufl. Studienskripten zur Soziologie. Wiesbaden: VS, Verl. für Sozialwiss.