Thema 10:
Lineare Modelle

QM1, SoSe 22

HOCHSCHULE
ANSBACH

Grundlagen

Drei Arten von Zielen wissenschaftlicher Studien

Deskription

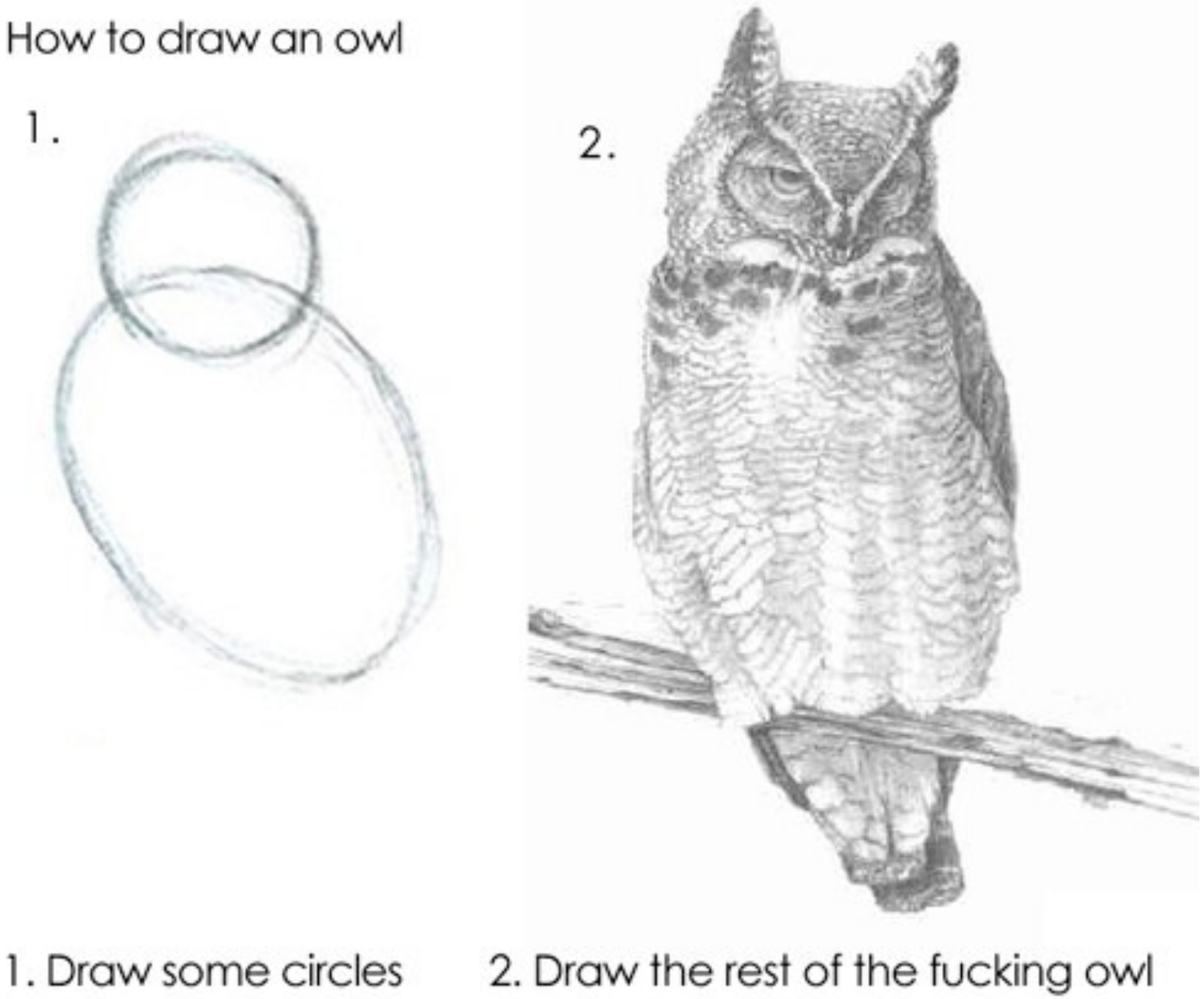
Explikation

Prognose

Was ist ein Modell?

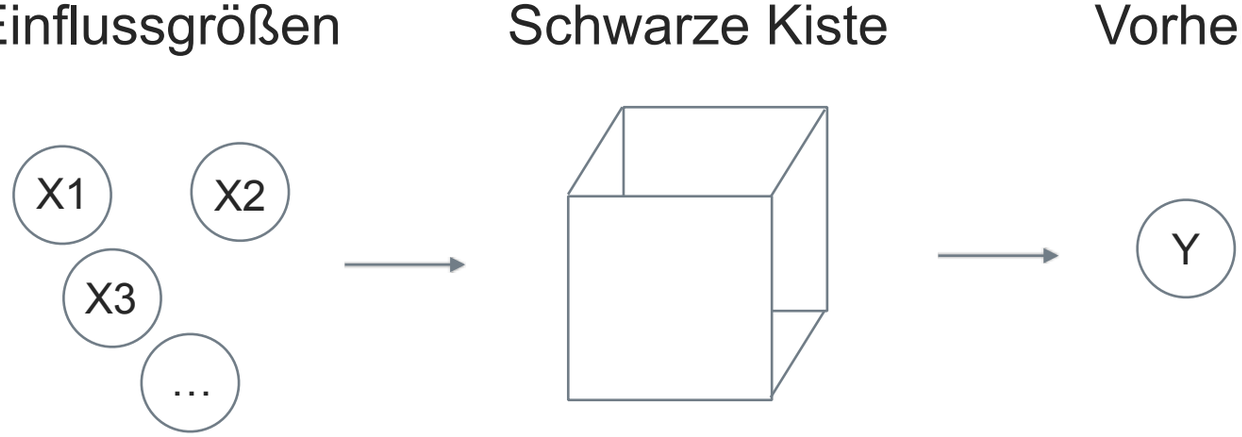
Quelle, CC-BY, unknown author, Wikimedia Commons

Modellieren als mirakulöser Zwischenschritt ?



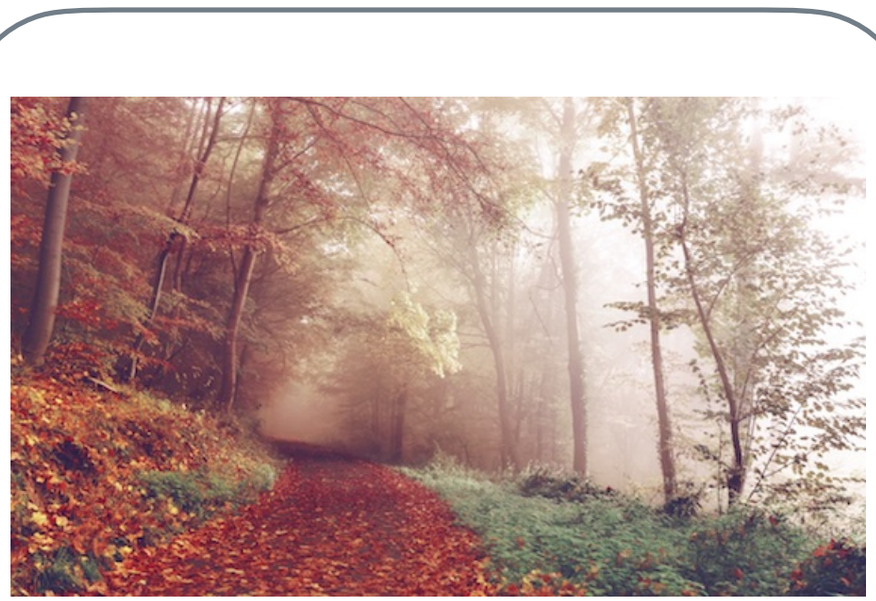
Quelle

Wie wichtig ist Transparenz im Modellieren?



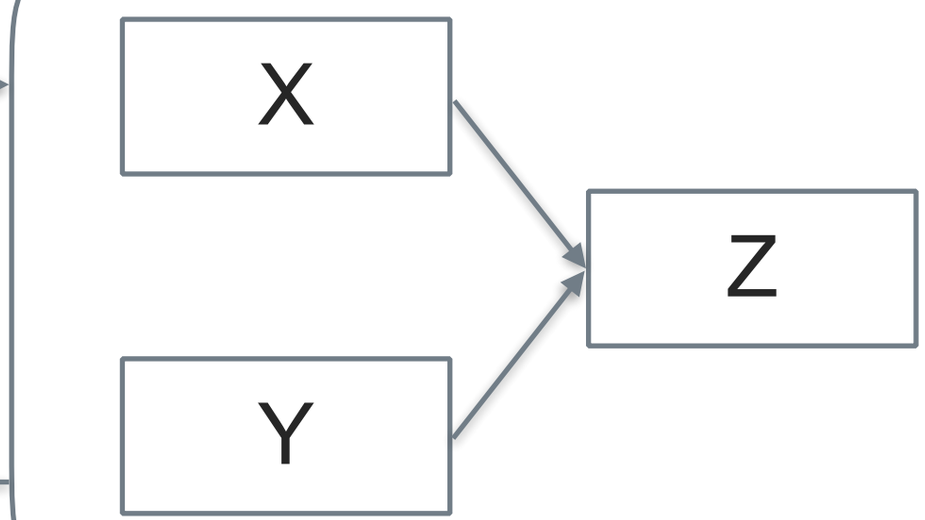
http://www.sciencecartoonsplus.com/pages/gallery.php

Modellieren



gestaltet

vereinfacht



Gegenstandsbereich

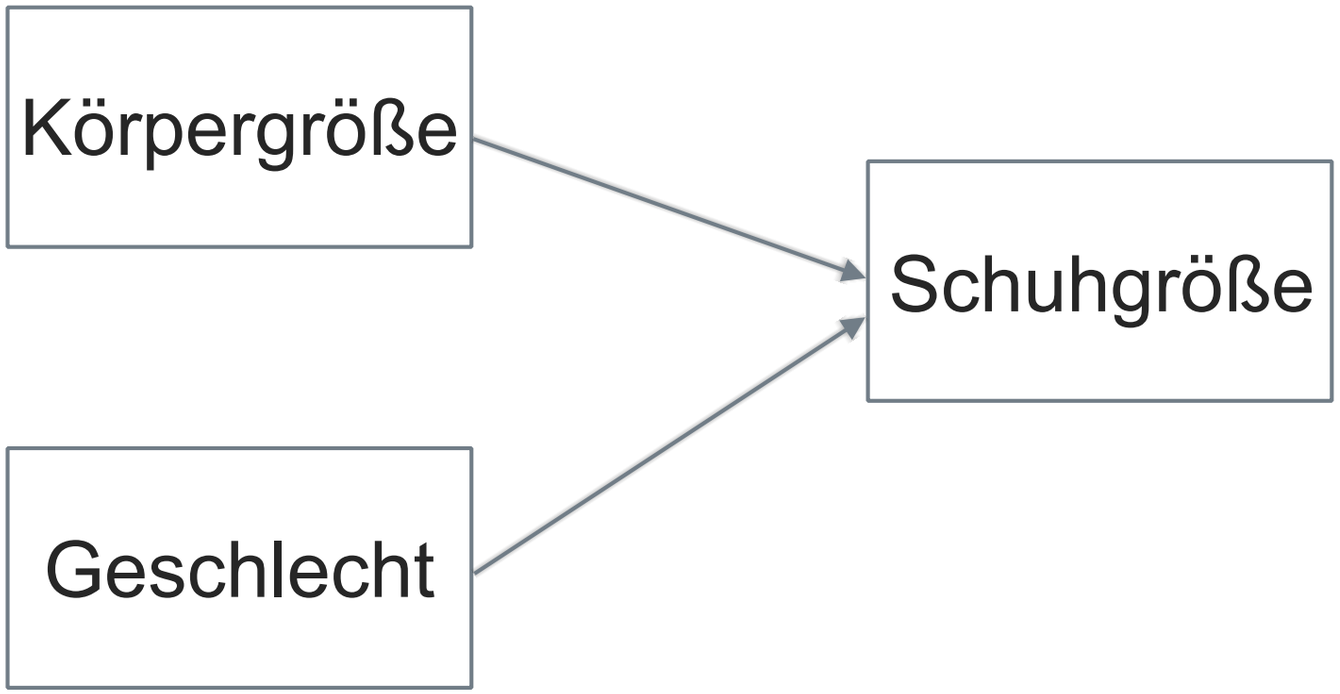
Modell

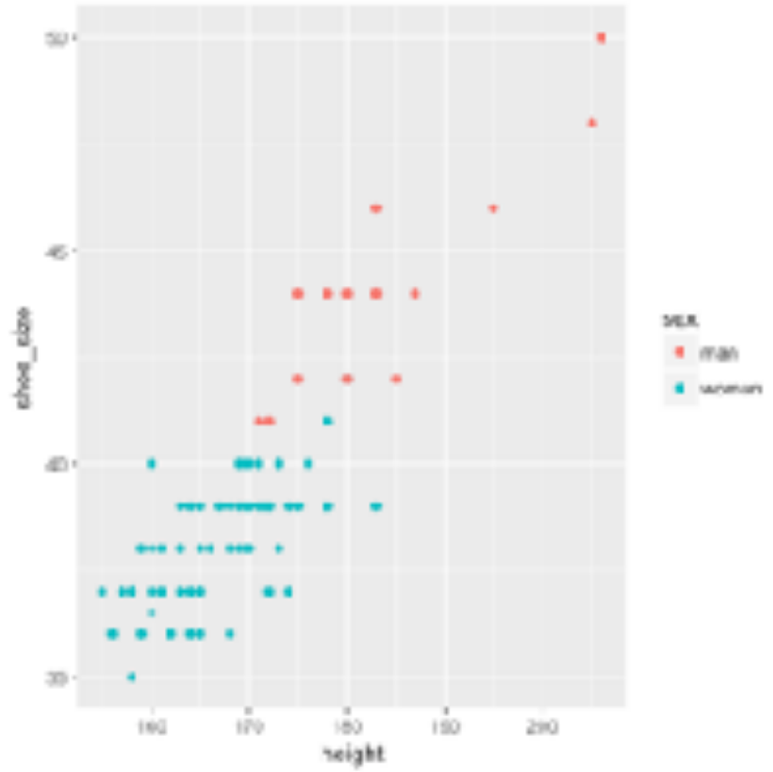
Beispiel zum Modellieren 1

Prädiktor, UV, X

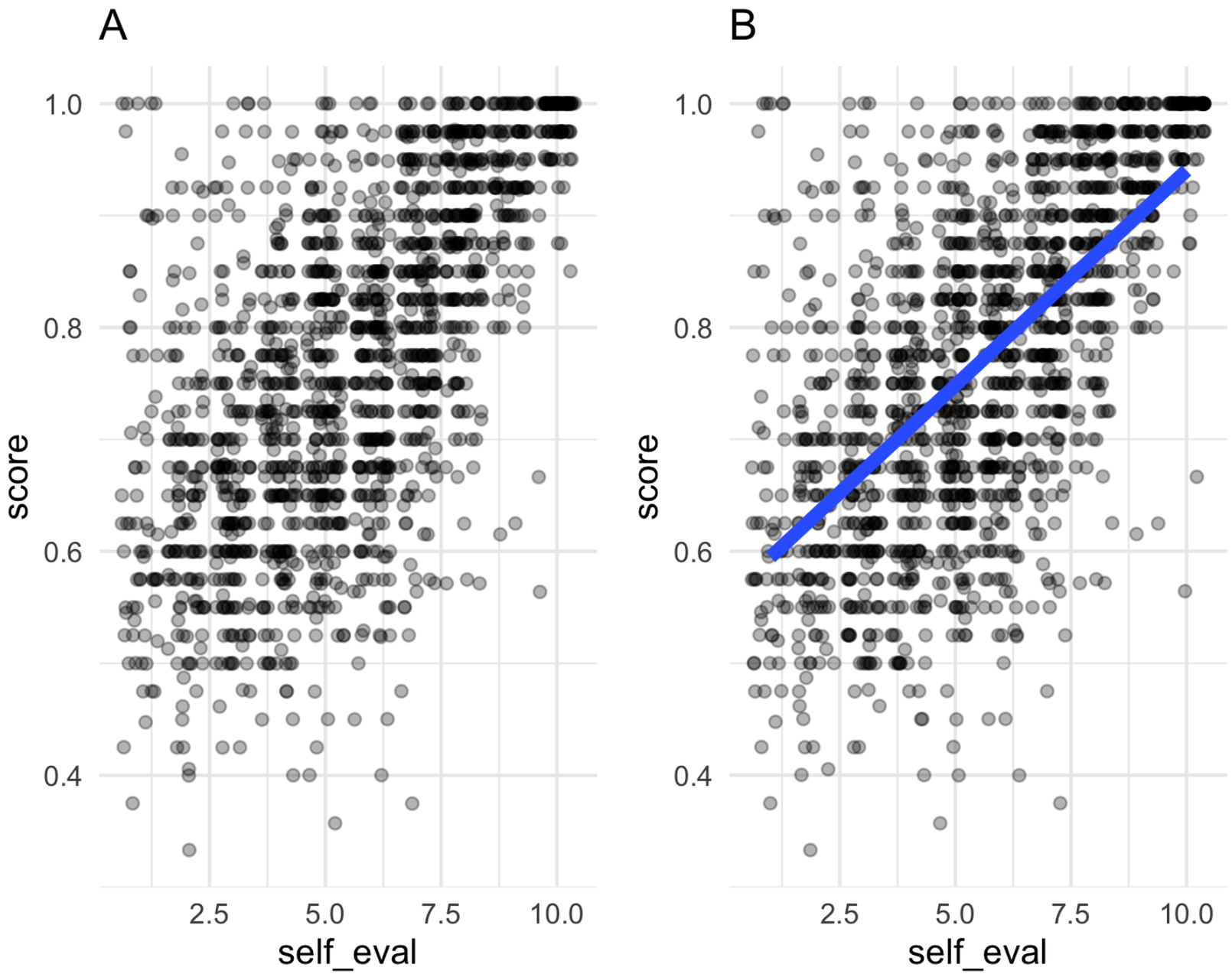
Kriterium, AV, Y

Beispiel zum Modellieren 2



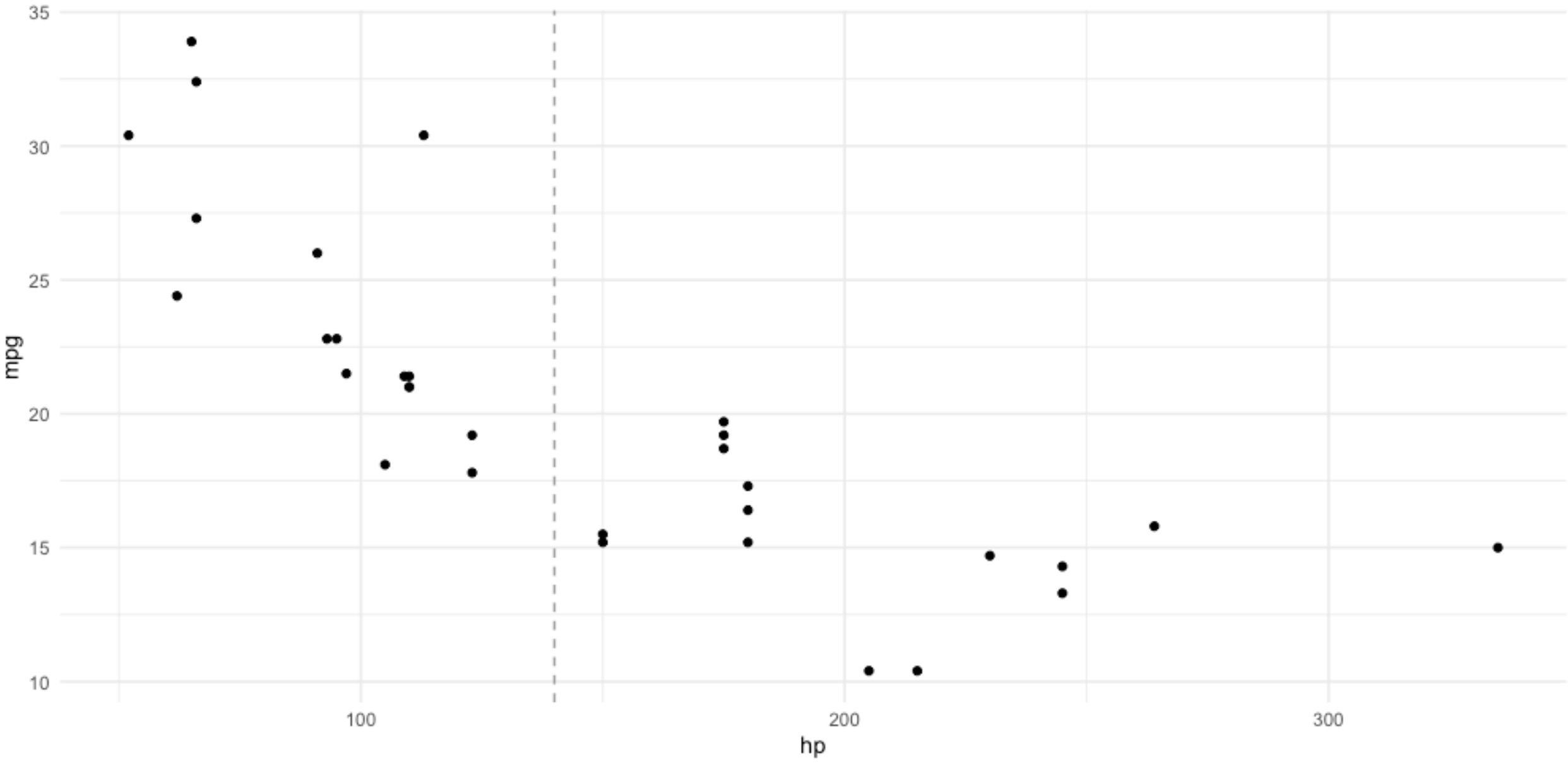


Beispiel zum Modellieren 2



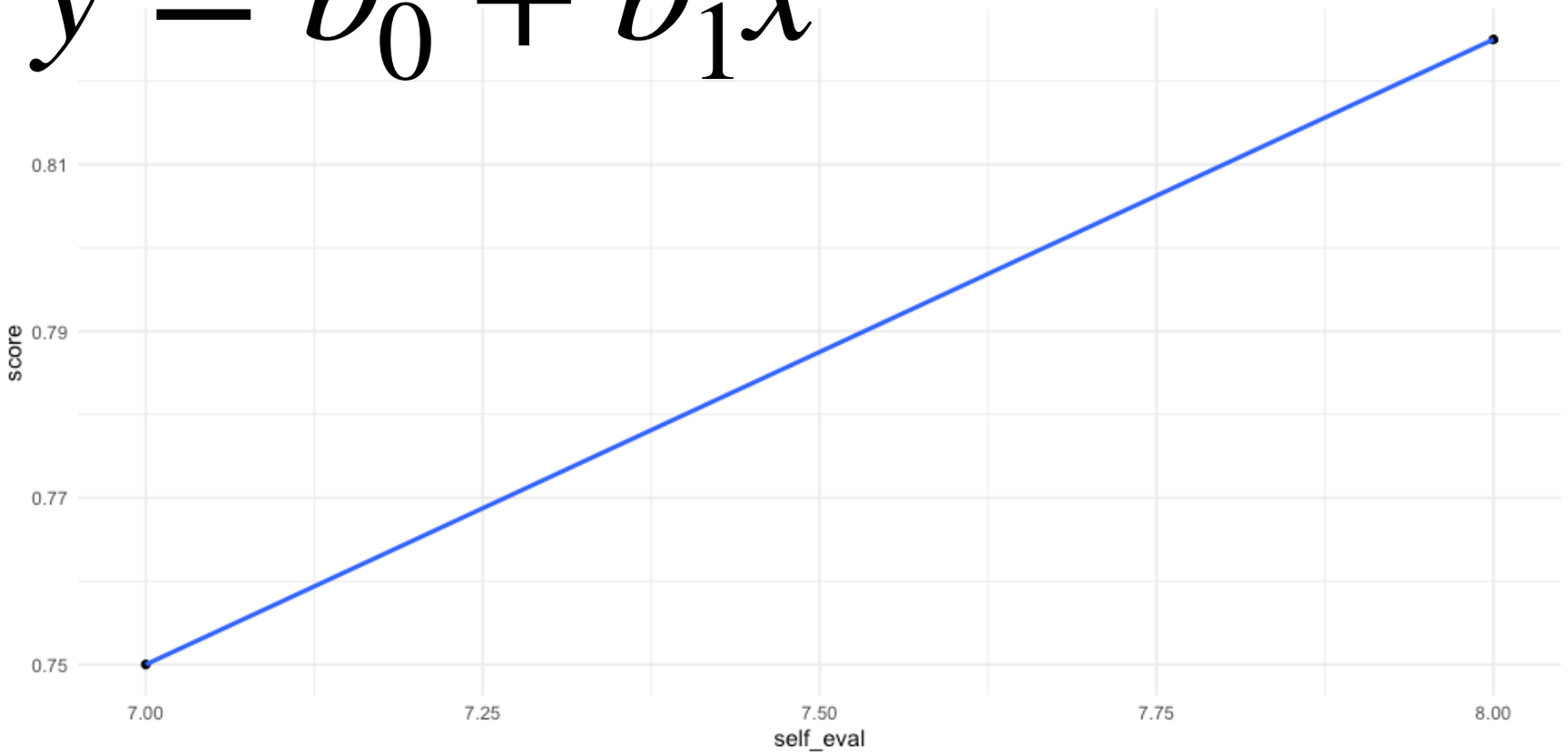
Geraden als Modelle

Wieviel Sprit braucht eine Karre mit 140 PS?



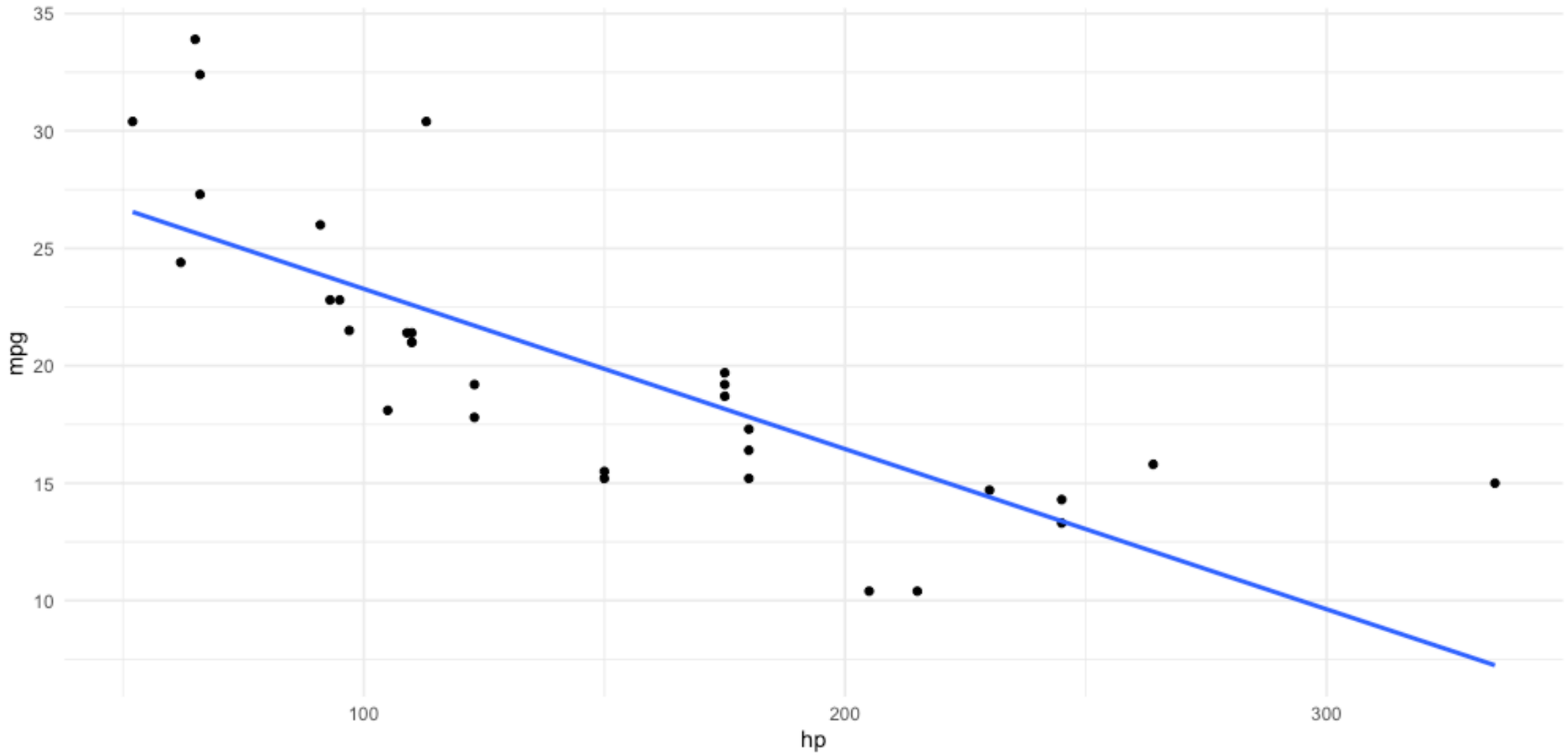
Eine Gerade als Modell

$$\hat{y} = b_0 + b_1 x$$



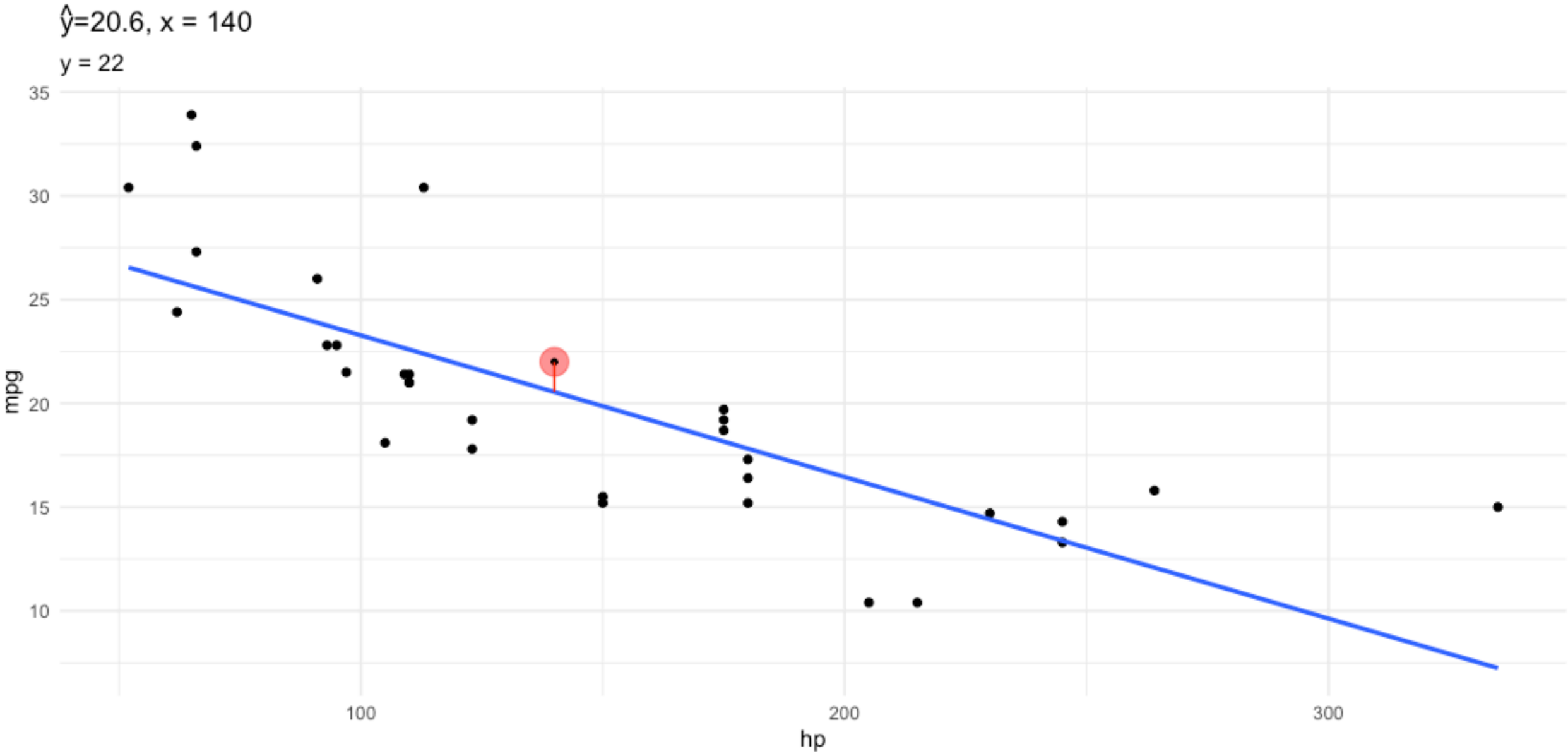
- Eine Gerade ist durch zwei Koeffizienten determiniert: Achsenabschnitt (b_0) und Steigungen (b_1).
- Kennt man die Koeffizienten, so kann man für jeden X-Wert den zugehörigen Y-Wert (Funktionswert) ausrechnen.

Spritverbrauch als Funktion von PS

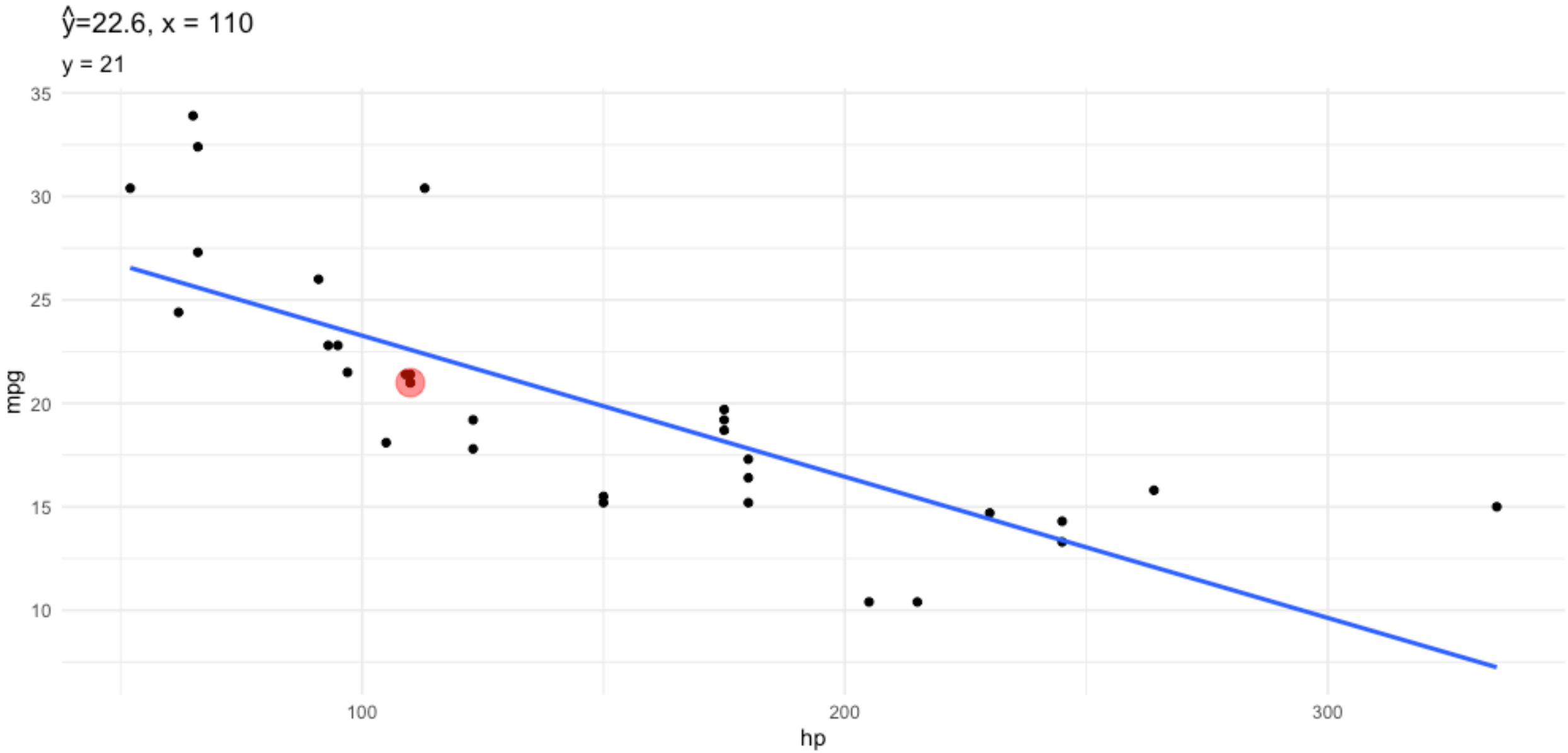


$$\hat{y} = 30 - 0.07 \cdot \text{PS}$$

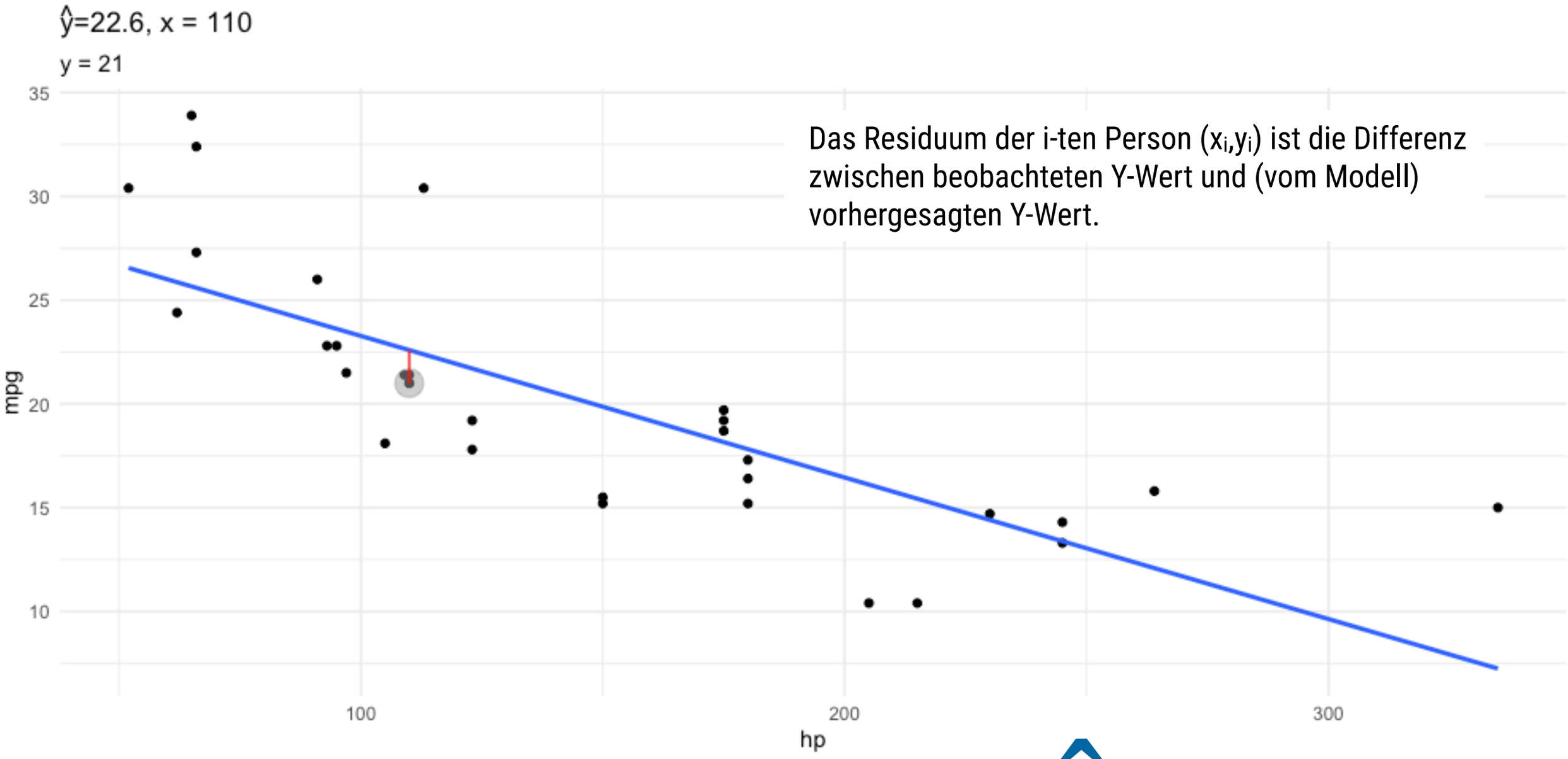
Gerade als Modell, nützlich zur Vorhersage



Wieviel Sprit braucht diese Karre?



Vorhersagefehler (Residuum, e)

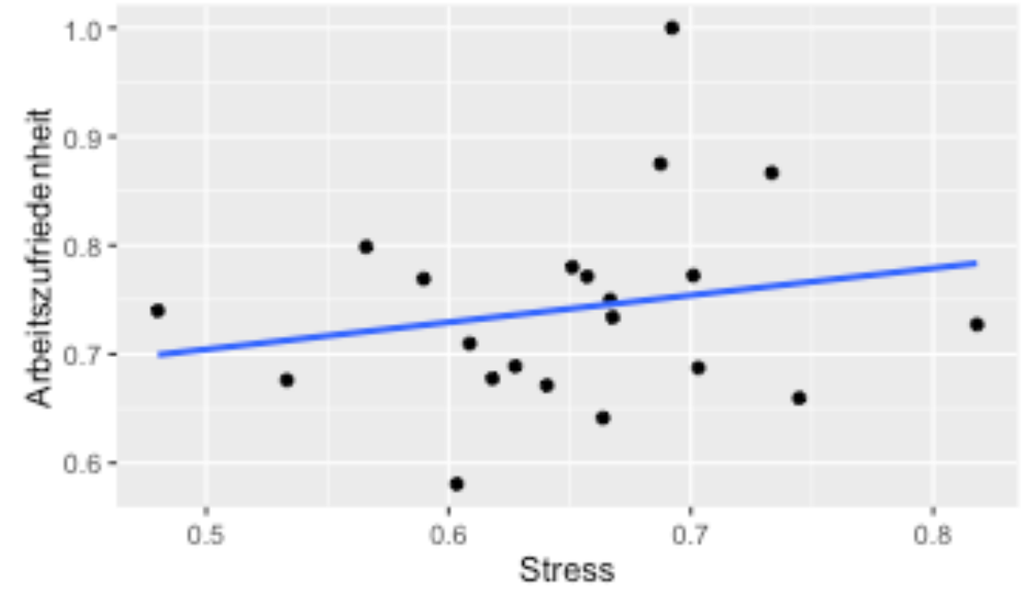


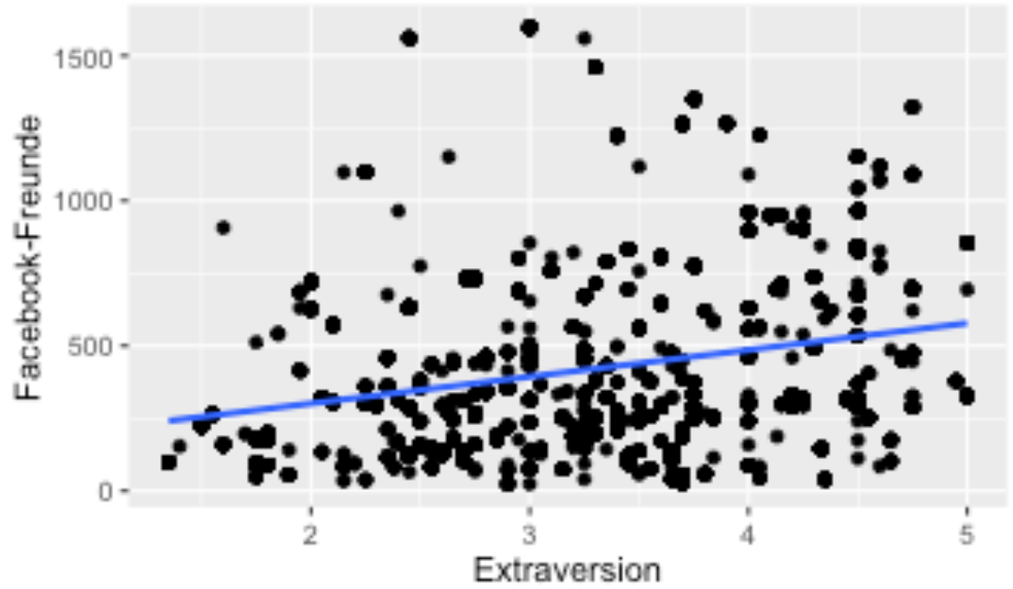
Das Residuum der i-ten Person (x_i,y_i) ist die Differenz zwischen beobachteten Y-Wert und (vom Modell) vorhergesagten Y-Wert.

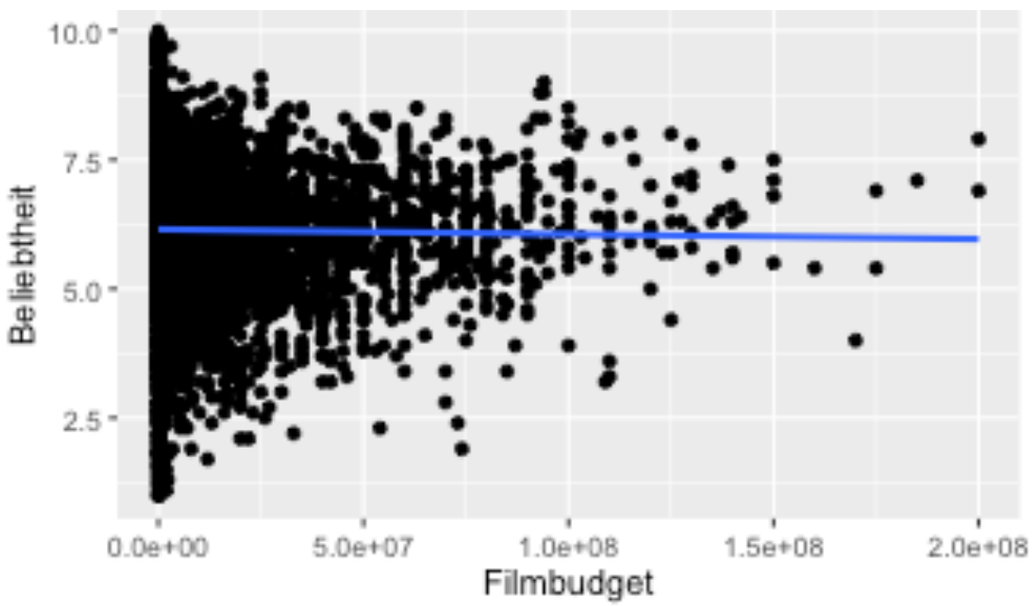
$$e_i = y_i - \hat{y}_i$$

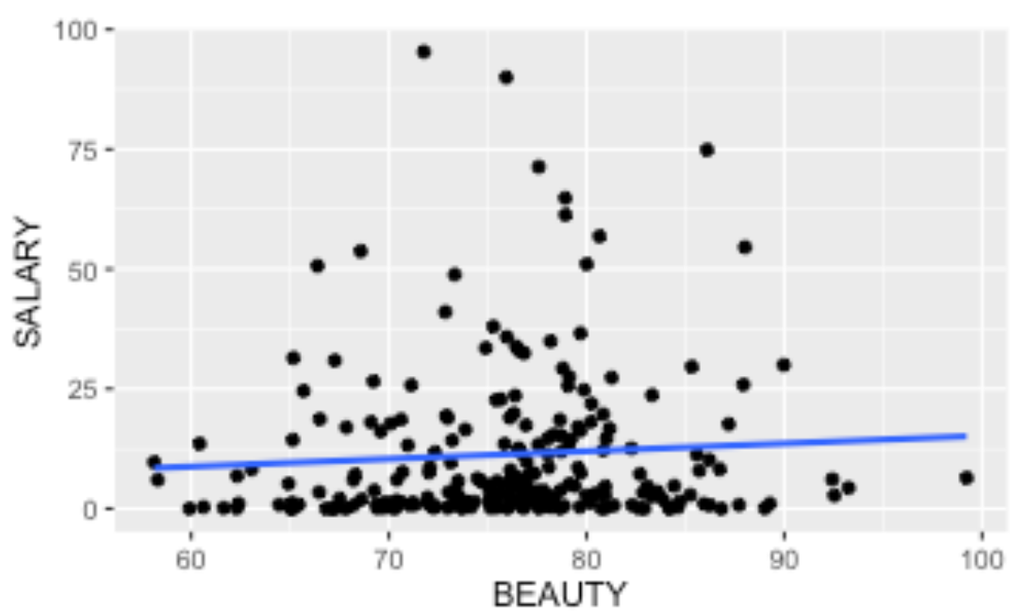
Regressionsgeraden können sich unterscheiden

- Regressionsgeraden können hinsichtlich Achsenabschnitt und Steigung unterscheiden
- Hat die Regressionsgerade eine Steigung von b=*0*, so leistet der Prädiktor keinen Beitrag zur Vorhersage (der Varianz) des Kriteriums.









Regressionsmodell

- Eine **Regressionsanalyse** ist ein Weg, den Wert einer metrischen **Kriteriumsvariable Y** (=abhängige Variable, AV) durch eine metrische **Prädiktorvariable X** (=unabhängige Variable, UV) zu **erklären** (damit ist *kein* kausaler Anspruch verbunden).
- Den *statistischen* Einfluss von X auf Y stellen wir anhand einer Gerade durch die Punktewolke dar; die Gerade soll die Punkte möglichst gut beschreiben.
- Dabei wird eine Gerade so in die Punktewolke hineingelegt, dass sie möglichst „mittig“ liegt – so, dass die (quadrierten) Abstände zwischen Geraden und Punkte möglichst **gering** sind.
- Anhand der Gerade können wir schätzen, welcher Y-Wert bei einem bestimmten X-Wert vorliegen sollte (man könnte sagen, wir führen einen Y-Wert auf seinen X-Wert zurück).
- Dabei werden wir Fehler machen, wenn unserer Vorhersage nicht perfekt ist.
- Eine Regressionsgerade ist – wie jede Gerade – durch folgende Gleichung gekennzeichnet:

$$\hat{Y} = b_1 x + b_0$$

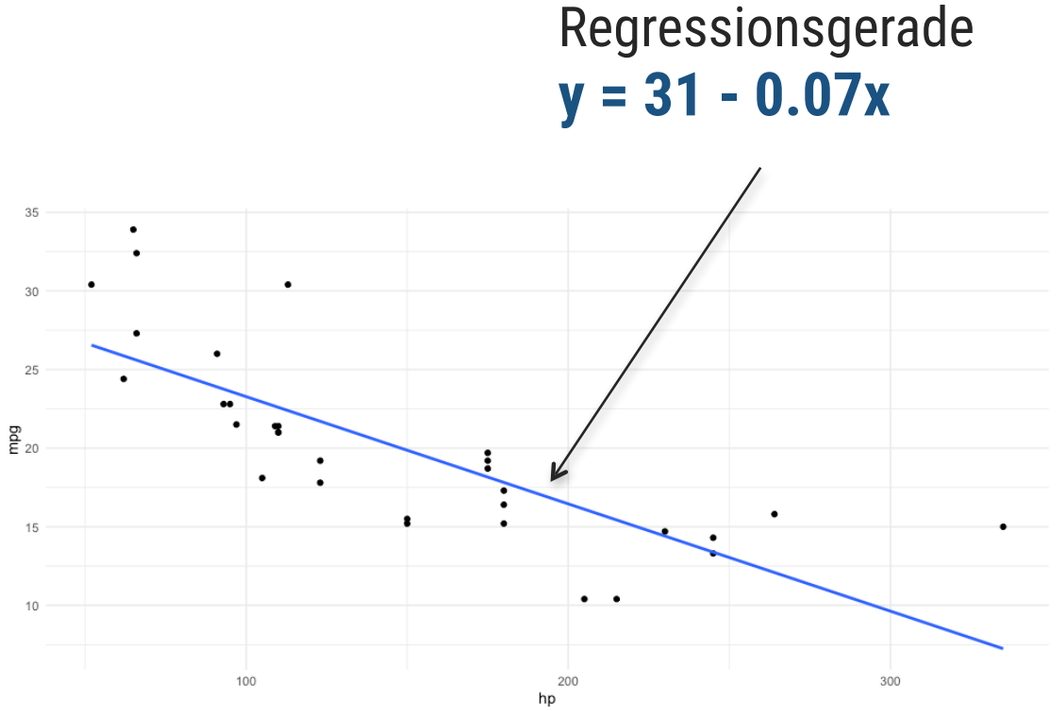
wobei *Ŷ* („Y-Dach“) für den *vorhergesagten* (geschätzten) *Y-Wert*, *b1* für die *Steigung* der Geraden, *x* für den *Prädiktor* und *b0* für den *Achsenabschnitt* (d.h. der Y-Wert wenn x = 0) steht

$$\hat{Y} = b_1 x + b_0 + \epsilon$$

Der *tatsächliche* (beobachtete) *Y-Wert* setzt sich zusammen aus dem geschätzten Y-Wert (Ŷ) plus einem Fehlerwert ε.

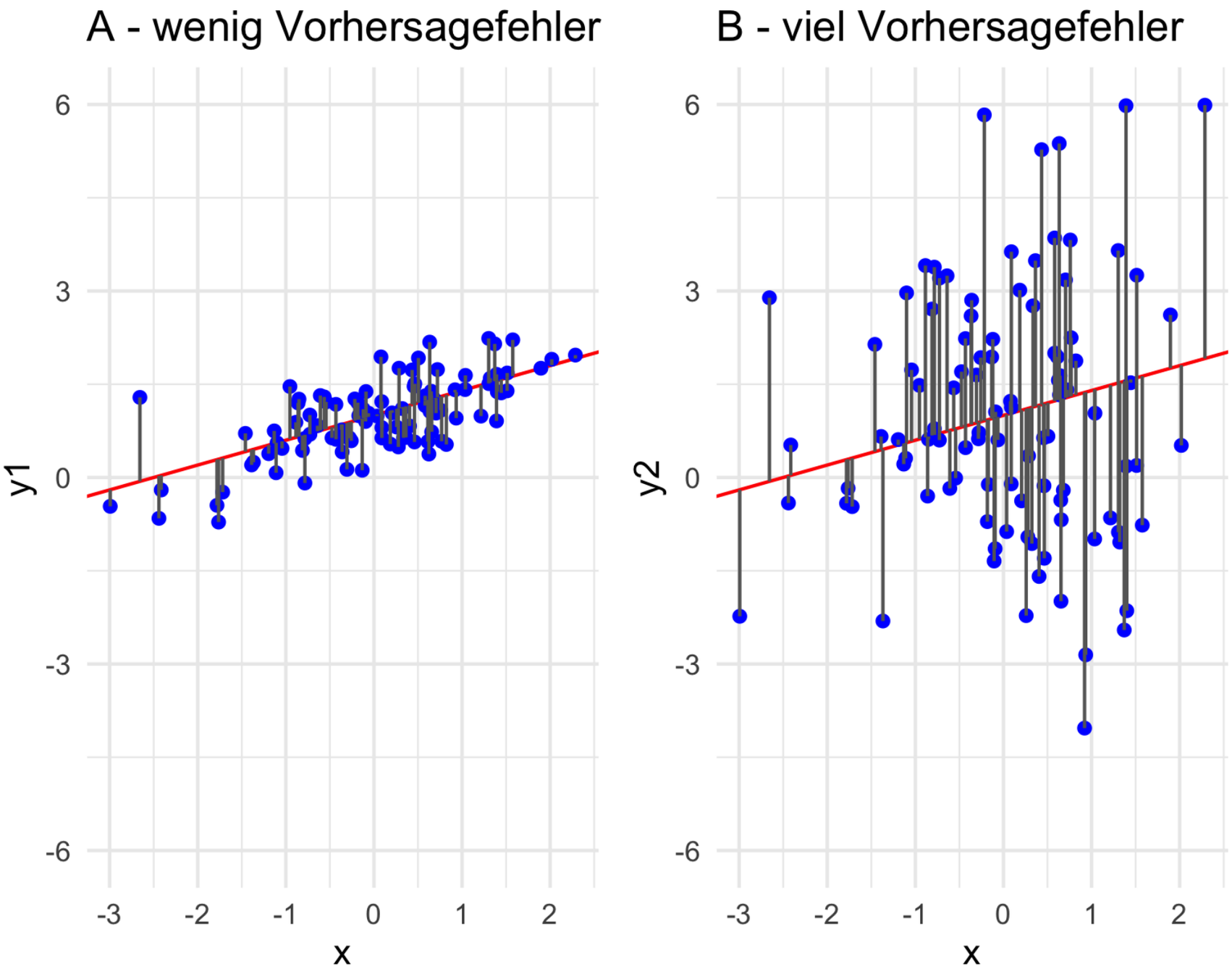
Die Steigung zeigt die Stärke des Zusammenhangs

- Die **Steigung *b1*** der Geraden quantifiziert die Stärke des Einflusses des Prädiktors X auf das Kriterium Y
- Steigung ist definiert als der Zuwachs in Y, wenn man X um eine Einheit erhöht (in R als *Estimate* bezeichnet)
- Abhängig von der Skalierung bei X und Y kann b1 alle möglichen Werte annehmen (positive und negative)
- Größere Werte von b sprechen tendenziell für einen größeren Einfluss von X auf Y
  - Beispiel: Zwei Autos, die sich um 1 PS unterscheiden, unterscheiden sich im Schnitt um ca. -0.07 MPG-Einheiten
- Der Achsenabschnitt b0 (engl. intercept) der Regressionsgeraden gibt den Y-Wert für X = 0 an
  - Beispiel: Bei 0 PS liegt der Spritverbrauch bei ca. 30 Meilen pro Gallone Sprit (theoretisch)



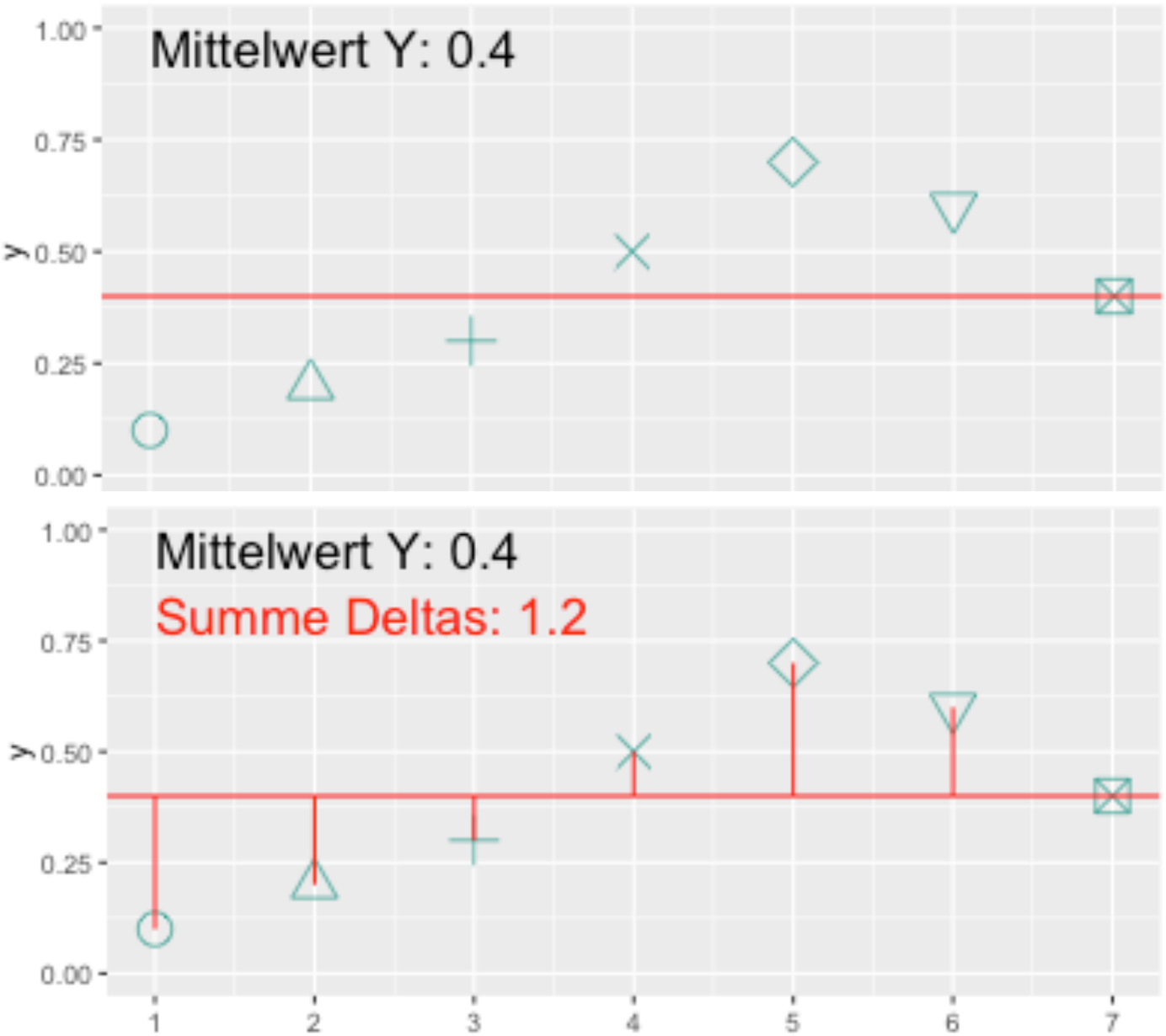
Modellgüte

Die Größe der Residuen zeigt die Modellgüte



Mittelwert als Referenziert

Eine Vorhersage hat nur dann Wert, wenn die Güte der Vorhersage (bzw. der Vorhersagefehler) bekannt ist/ bestimmt werden kann.

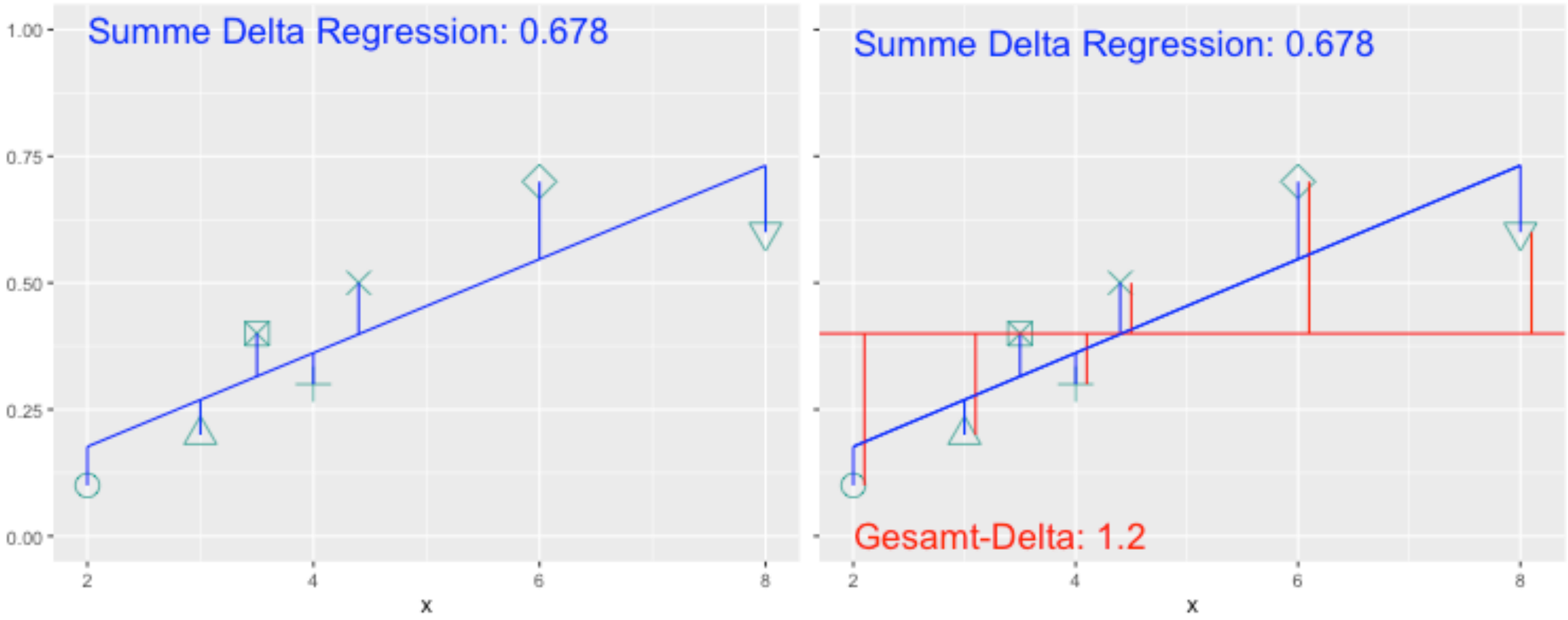


Ist Alberts Zufriedenheit (Y) *unbekannt*, so kann an den Mittelwert von Y ($\bar{Y}$) als Schätzer für ein bestimmtes Y_i (z.B. von Albert) nehmen, damit liegt man oft ganz gut. Dieses Verfahren besteht in einer Vorhersage von Y ohne Kenntnis eines Prädiktors (X).

Die roten „Stecken“ zeigen die Größe des Vorhersagefehlers an; der mittlere „Quadratstecken“ ist die Varianz. Die roten Stecken sind also ein Maß für die Güte der Vorhersage!

Wir legen eine „gut sitzende“ Gerade in die Daten

Und siehe da: die Summe der „Abweichungs-Stecken“ (Residuen, e) wird kürzer!



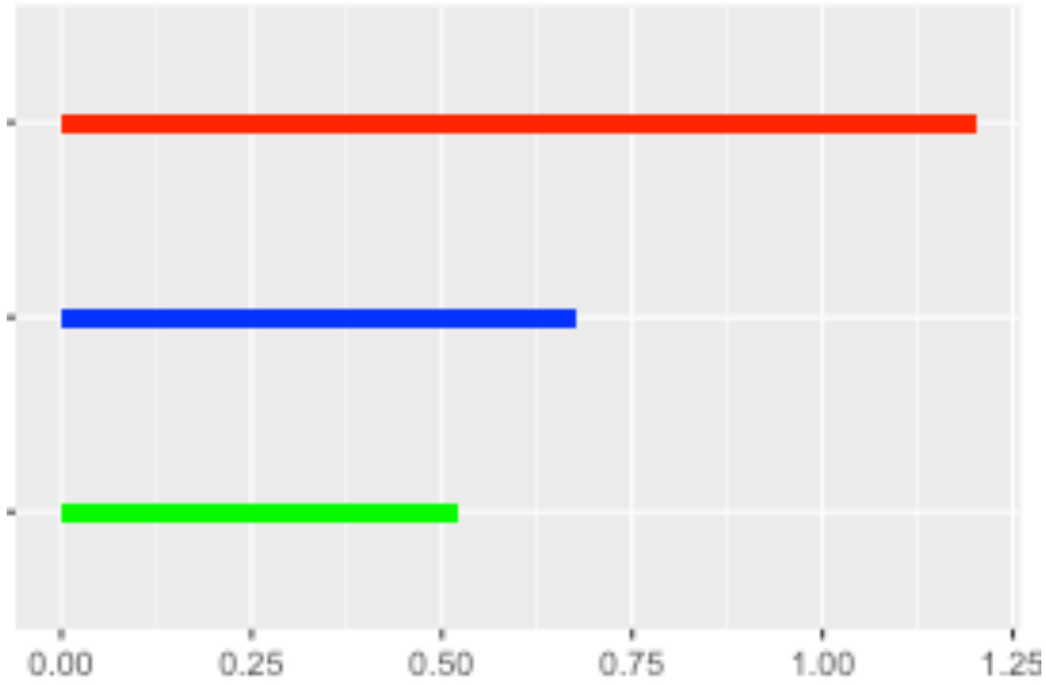
Die blauen Abweichungen (Deltas) sind in Summe kleiner als die roten (in Summe). Damit werden die Y_i-Werte durch die Regression insgesamt genauer geschätzt als bei Vorhersage duch $\bar{Y}$; der Vorhersagefehler wird kleiner.

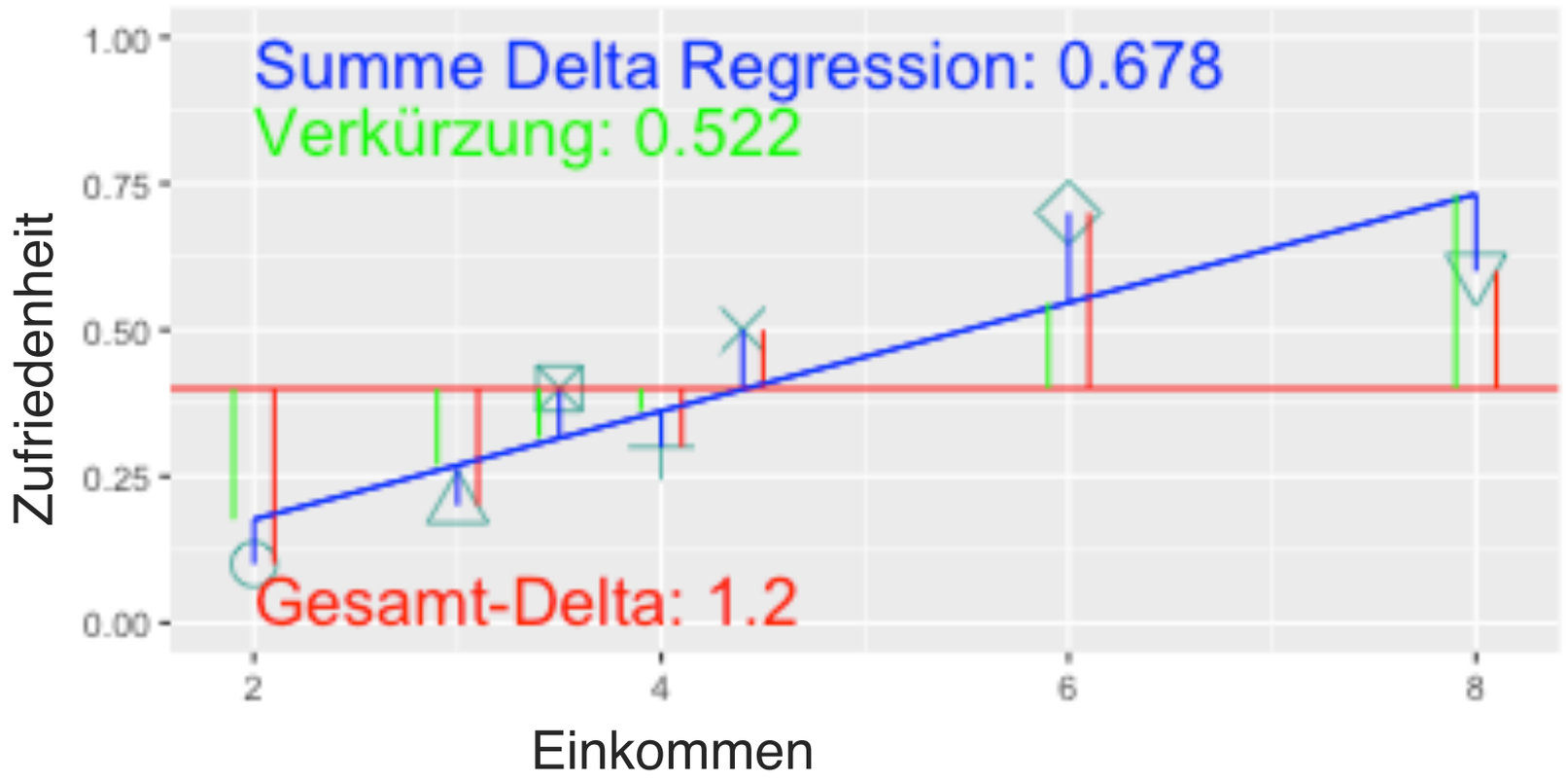
Unser Regressionsmodell verkürzt die Residuen

Abweichungen vom Mittelwert

Abweichungen von der Regressionsgeraden

Unterschied rot/blau





Die Abweichungen von der Regressionsgeraden (blau) sind in Summe kürzer als die Abweichungen von Mittelwert (rot); die Verbesserung lässt sich aus der Differenz dieser beiden Abweichungen bestimmen (grüner Balken).

Die Quadratsummen addieren sich

Die einzelnen „Gesamt-Abweichungsbalken“ bezeichnet man als *Quadratsummen* (engl. Sum of Squares, *SS*).

"Deutsche Übersetzung"

SS_T :

$$SS_t = \sum_{i=1}^{n} (\bar{y} - y_i)^2$$

die Summe der (quadrierten) Differenzen zwischen den erhobenen Daten und dem Mittelwert von Y (totale Varianz)

Gesamt-Varianz, maximale Streuung, totale Fehlerstreuung

SS_E :

$$SS_E = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Die Summe der der (quadrierten) Differenzen (Residuen, Error) zwischen den erhobenen Daten und der Regressionsgeraden

Gesamt-Vorhersage-Fehler, Summe der Abweichung von der Regressionsgeraden, Fehlerstreuung der Regression

SS_M :

$$SS_M = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Die Summe der der (quadrierten) Differenzen zwischen dem Mittelwert von Y und der Regressionsgeraden (dem Modell)

Verbesserung durch das Modell, erklärte Varianz, Verringerung des Vorhersage-fehlers durch die Regression

Wie gut ist mein Modell?

- Stets ist es das Ziel, die Residuen so klein wie möglich zu halten.
- Um zu prüfen, wie gut ein Regressionsmodell im Schnitt das Kriterium vorhersagt, werden die Abweichungen von vorhergesagten $\hat{Y}_i$ und tatsächlichen Kriteriumswerten Yi berechnet (die „blauen Abweichungstecken“).
- Ist der Wert des SSM größer als Null, so kann die Kriteriumsvariable Y mithilfe von SSM besser vorhergesagt werden als durch das arithmetische Mittel $\bar{Y}$ allein.

$$SS_T = SS_M + SS_E$$

Der mittlere Vorhersagefehler als Maß der Modellgüte

Root Mean Square Error (RMSE)

1. Bestimme das Residuum e für die 1. Beobachtung als Differenz von beobachteten und (vom Modell) vorhergesagten Wert
2. Quadriere das Residuum e: Voilà, das Quadrat-Residuum
3. Wiederhole das für alle Residuen
4. Teile durch die Anzahl der Beobachtung, um das mittlere Quadrat-Residuum zu erhalten
5. Ziehe die Wurzel daraus, um wieder zu einer Größenordnung zu gelangen, die den ursprünglichen Werten entspricht

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

$$RMSE = \sqrt{\frac{1}{n}SS_T}$$

Das Bestimmtheitsmaß R^2

- Das Bestimmtheitsmaß R^2 gibt den Anteil der im Modell erklärten Variation von Y an.
- Es ist ein Maß der Modellgüte: Je größer , desto besser erklärt das Modell die Daten.
- Allerdings ist es, wie die Korrelation (nach Pearson) ein Maß der Modellgüte nur für lineare Modelle.
- Bei einer Regression mit einem Prädiktor ist R^2 gleich dem Quadrat der Pearson'schen Korrelation (r). Es ist damit ein Maß für ein lineares Muster, nicht (zwangsläufig) für geringe Residuen.
- Zu wie viel Prozent die Variation in der Kriteriums variable durch die Variation der X-Werte linear erklärt wird, wird durch R^2 (Bestimmtheitsmaß) ausgedrückt.

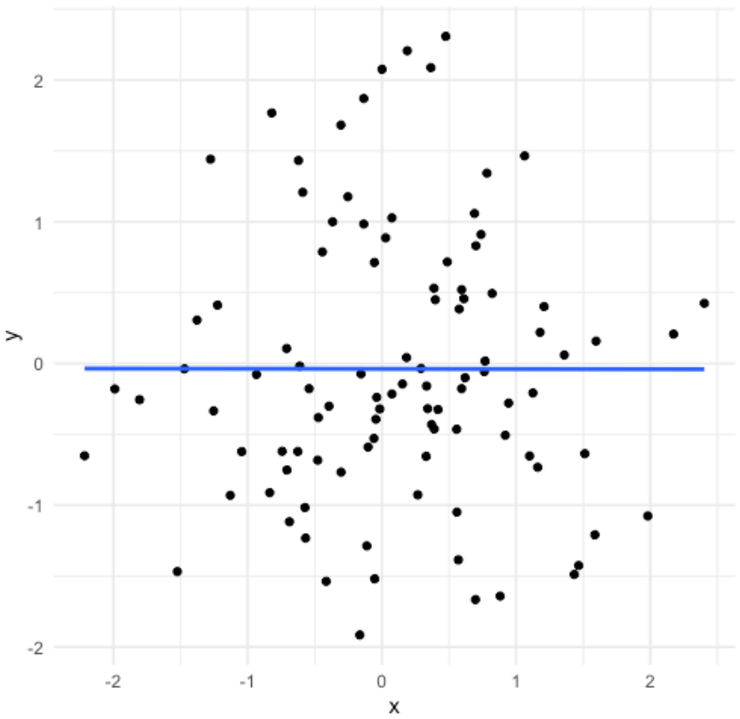
$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{SS_M}{SS_T}$$

$$R^2 = \frac{SS_M}{SS_T}$$

Einfachstes vs. bestes Modell

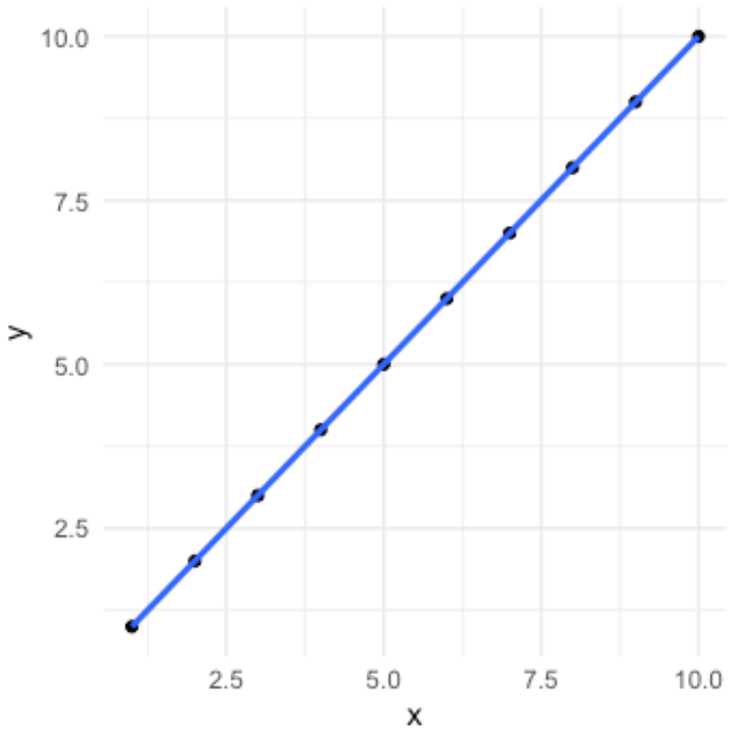
Einfachstes (oder einfaches) Modell:
Prognose durch Mittelwert.

$$\hat{y}_i = \bar{y} : R^2 = 0$$



Bestes Modell:
Prognose entspricht der Beobachtung

$$\hat{y}_i = y_i : R^2 = 1$$

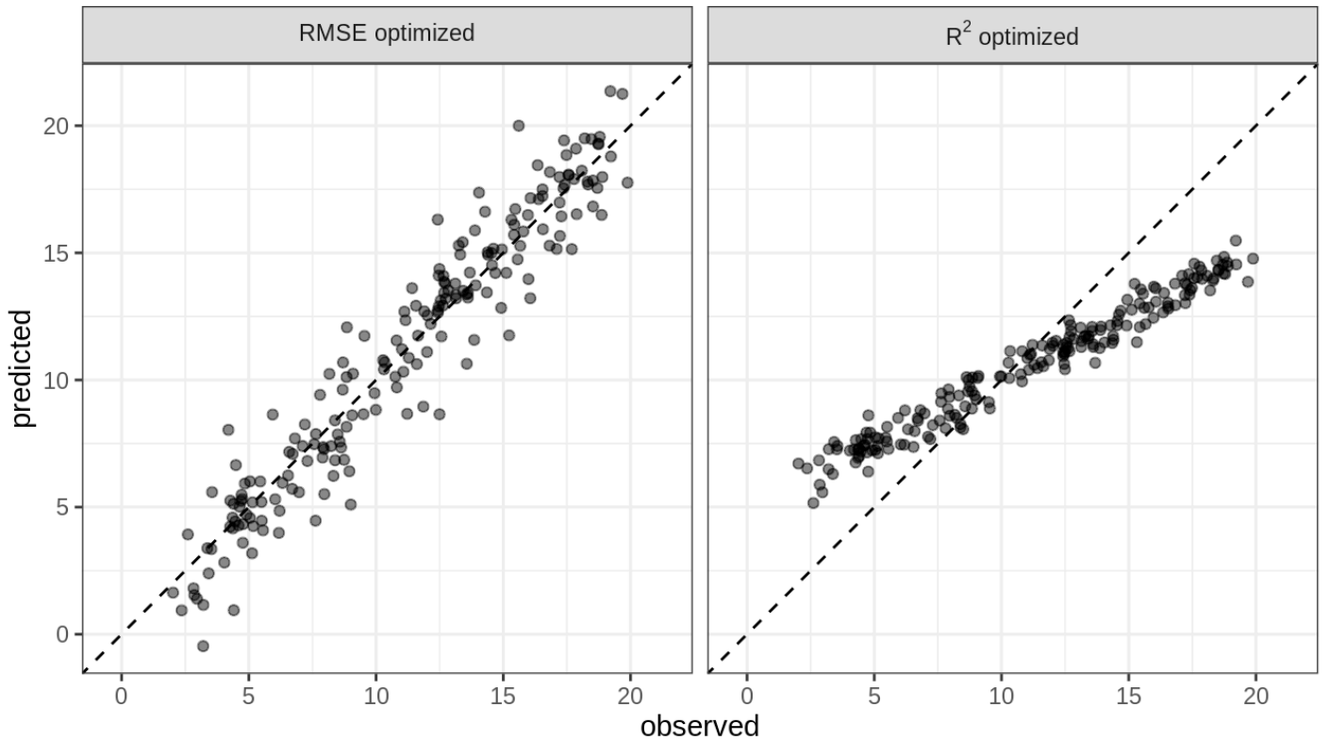


- Bei einer perfekten Korrelation ($R^2 = 1$) liegen die Punkte auf der Geraden; im schlimmsten Fall ($R^2 = 0$) ist die Vorhersage genauso gut wie wenn man für Y-Wert $\bar{Y}$ vorhersagen würde. R^2 ist also proportional zur Höhe der (linearen) Korrelation.

RMSE vs. R-Quadrat

RMSE misst die Kürze der Residuen; R-Quadrat misst die Korrelation

- RMSE und R2 werden oft ähnliche antworten, welches Modell gut ist (bzw. besser als ein anderes)
- RMSE und R2 können aber zu unterschiedlichen Antworten kommen, da sie nicht das gleiche messen
- Das linke Teilbild zeigt ein Modell mit
  - gutem Wert für RMSE
  - nicht so gutem Wert für R2
- Das rechte Teilbild zeigt ein Modell mit
  - gutem Wert für R2
  - nicht so gutem Wert für RMSE



Abschluss

Hinweise

- Dieses Dokument steht unter der Lizenz CC-BY 3.0.
- Autor: Sebastian Sauer
- Für externe Links kann keine Haftung übernommen werden.
- Dieses Dokument entstand mit reichlicher Unterstützung vieler Kolleginnen und Kollegen aus der FOM. Vielen Dank!
- Dieses Dokument baut in Teilen auf auf dem Skript zu quantitative Methoden des ifes-Instituts der FOM-Hochschule.