awm

angewandte **w**irtschafts- und **m**edienpsychologie

Thema 11:

Vertiefung zur Regressionsanalyse

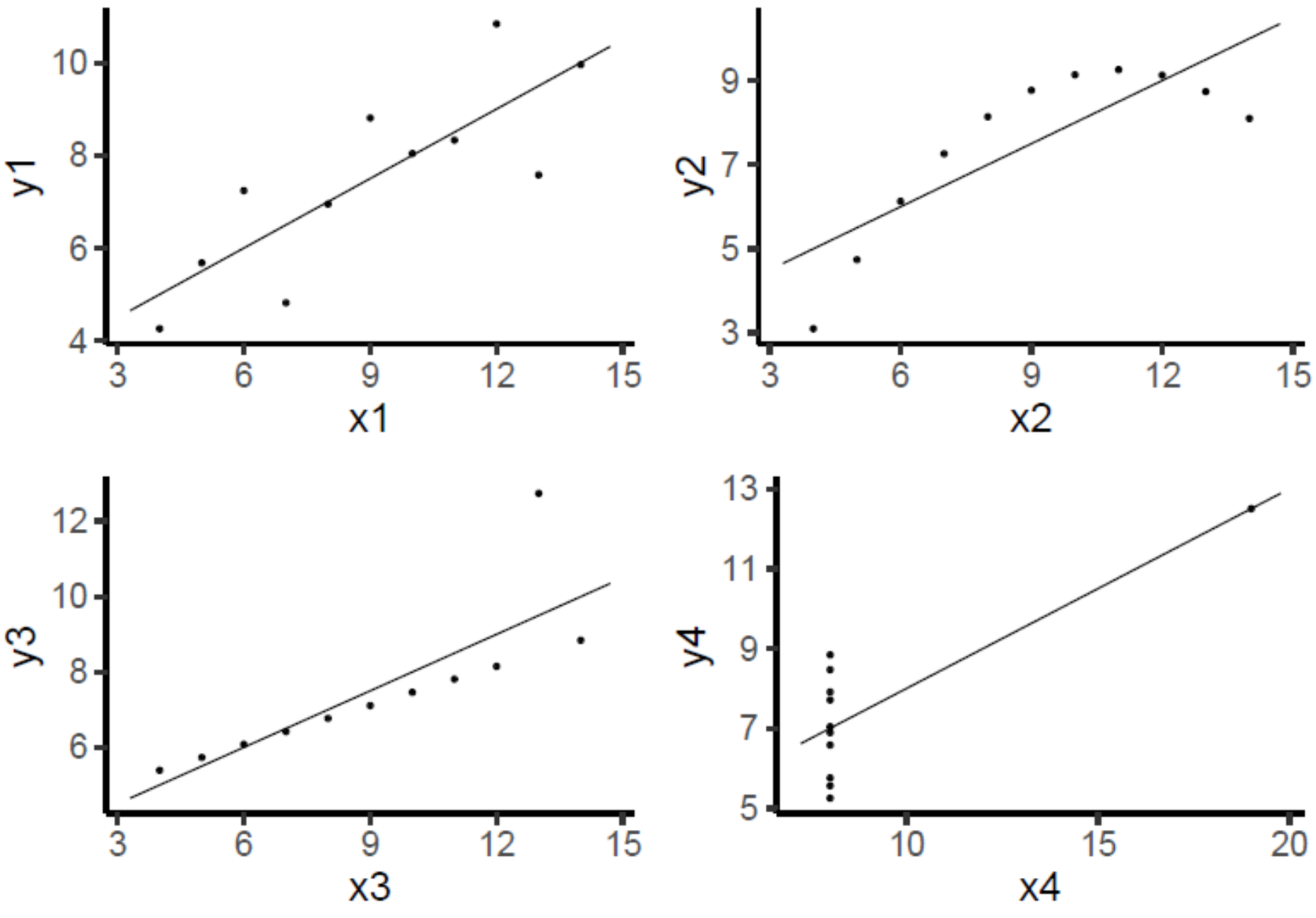
QM1, SoSe 22

HOCHSCHULE ANSBACH

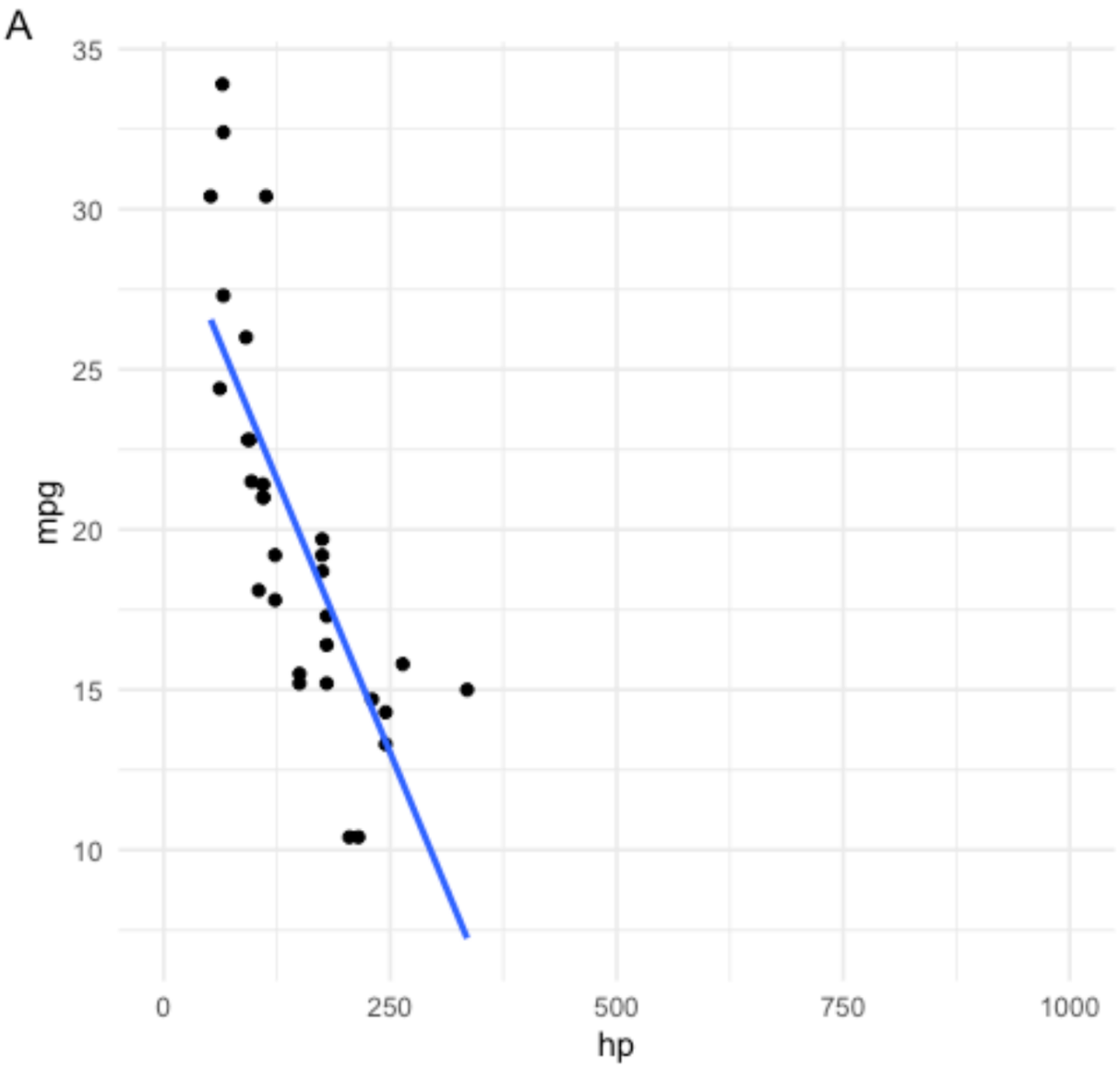
Grenzen der linearen Regression

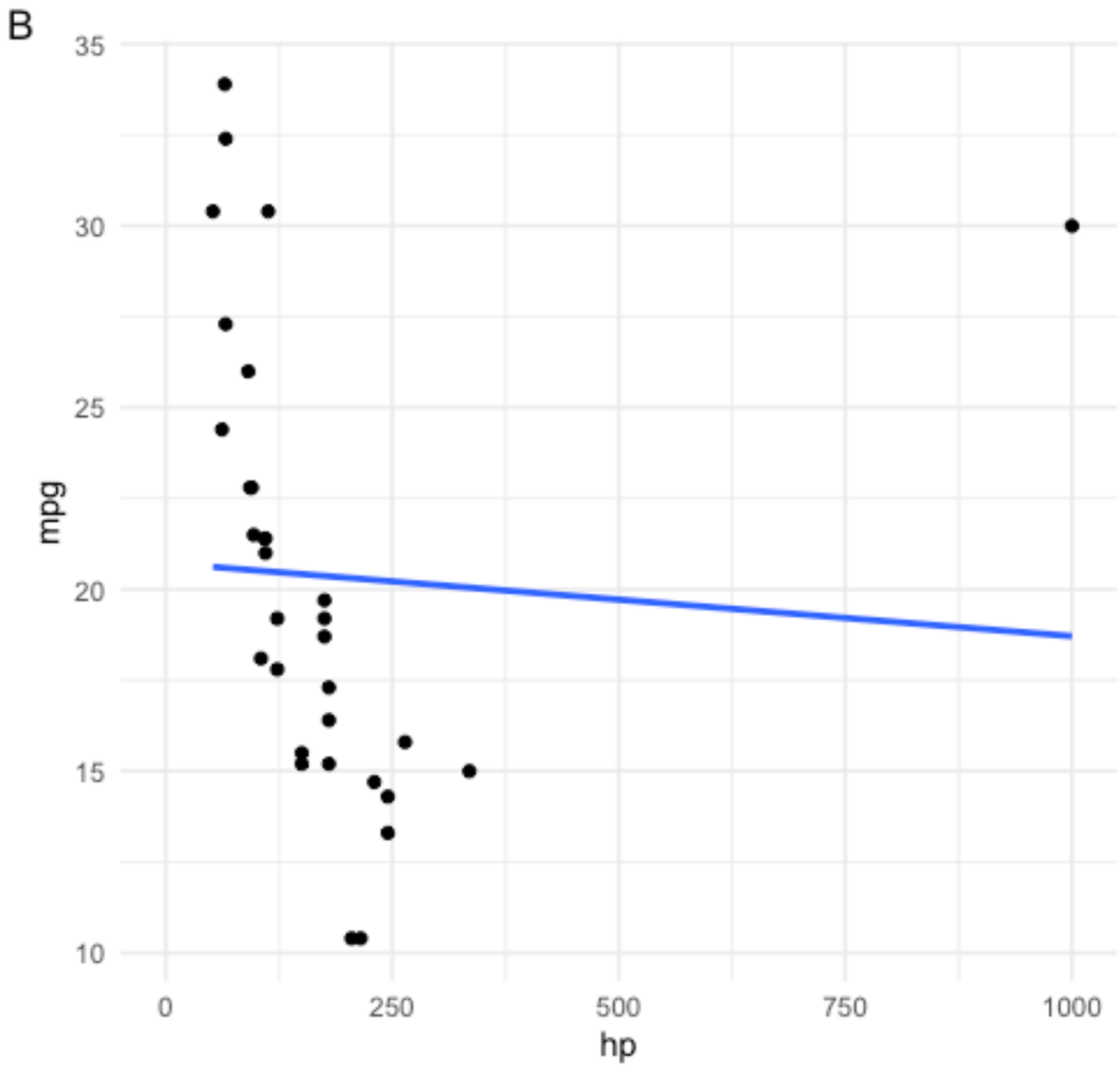
Anscombes Quartett revisited

Bei den vier Datensätzen im Anscombe-Quartett ist das R-Quadrat immer (in etwa) gleich. $R^2 \approx 0.67$



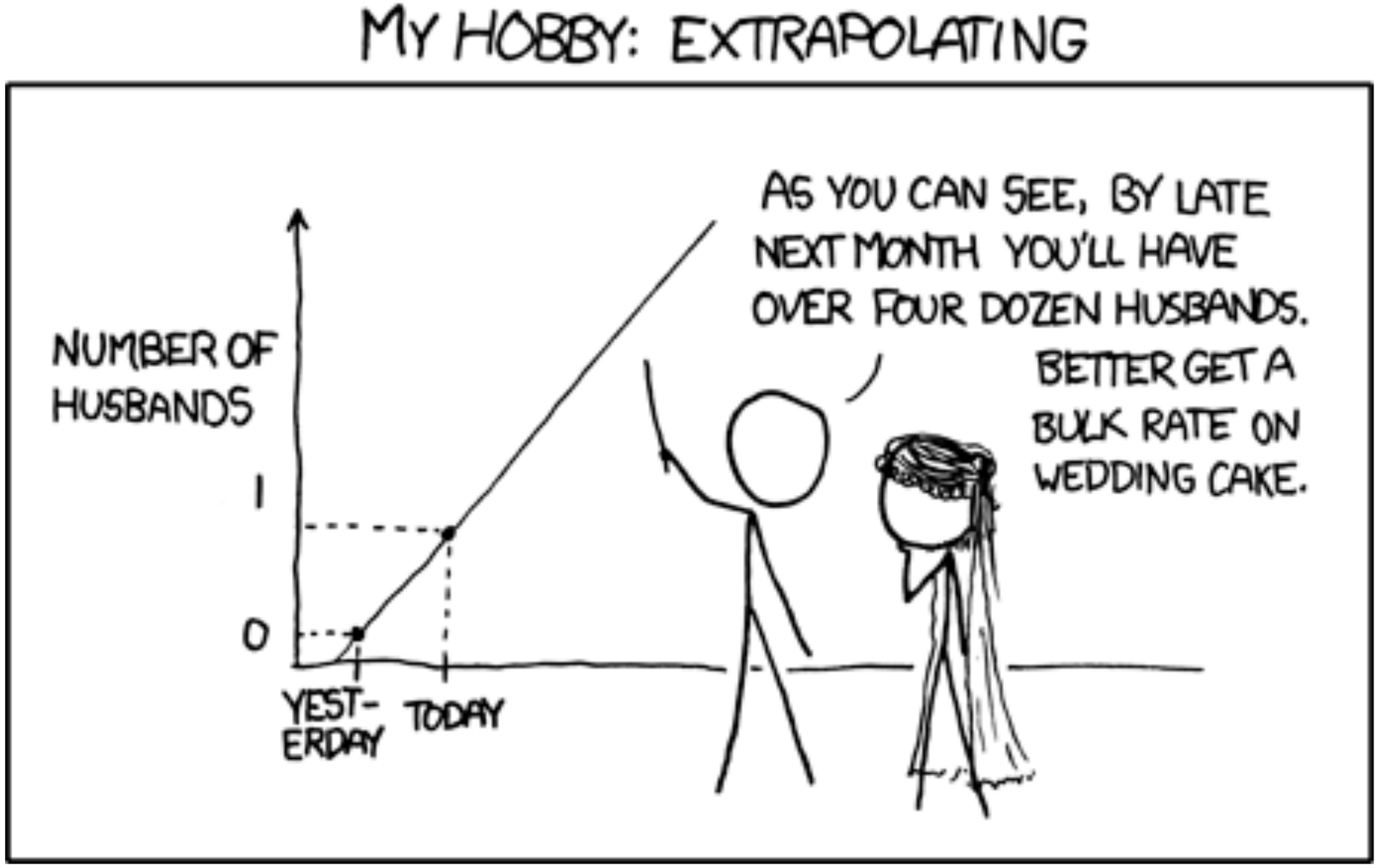
Ausreißer





- Beobachtungen mit extremen Ausprägungen im Prädiktor oder im Kriterium können einen starken Einfluss auf die Regressionsgerade – und damit auch auf die Modellgüte – haben.
- Man sollte prüfen, ob der Ausreißer ein korrekter Wert ist und ggf. korrigieren.

Extrapolation



[Quelle]

Nominalskalierte Prädiktoren

Zweistufig nominal:

Geschlecht	is_female
männlich	0
weiblich	1

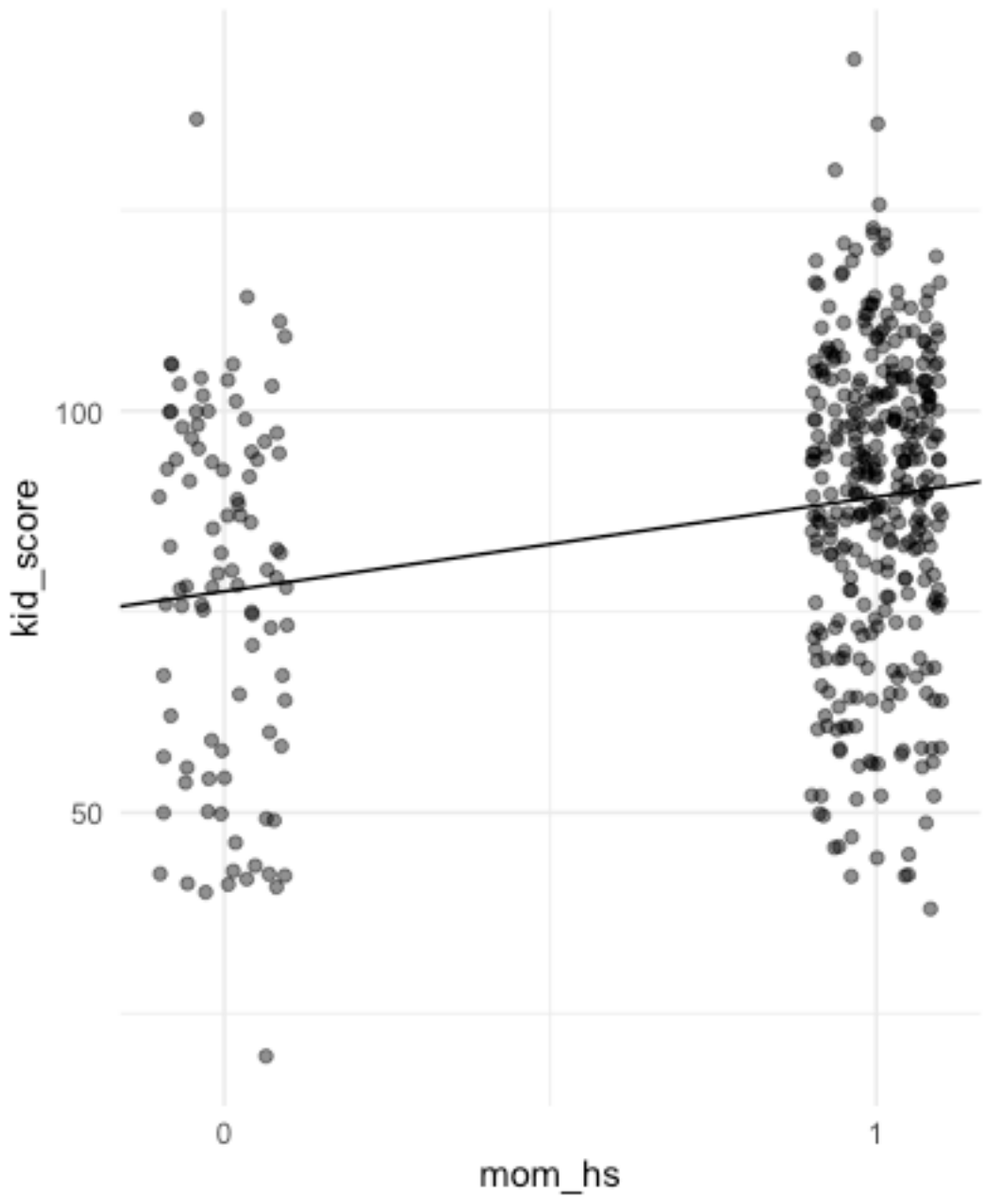
Mehrstufig nominal:

Geschlecht	is_female	is_male
männlich	0	1
weiblich	1	0
divers	0	0

- Die Regressionsanalyse benötigt numerische Variablen, um die Koeffizienten der Regressionsanalyse bestimmen zu können.
- Nominalskalierte Prädiktoren müssen daher in Zahlen umgewandelt werden, sofern möglich.
- Die AV (Kriterium) nehmen wir weiterhin als numerisch an.
- Bei einer zweistufig nominalskalierten Variablen wandeln wir in eine Indikatorvariable („Dummy-Variable“) um.
- Bei einer k-stufig nominalskalierten Variablen benötigen wir k-1 Indikatorvariablen.
- Allgemein wird immer eine Indikatorvariablen weniger benötigt, als der Prädiktor Ausprägungen hat.
- Grund dafür ist, dass sich die letzte Ausprägung ableiten lässt aus den übrigen Indikatorvariablen: Haben die übrigen Indikatorvariablen alle den Wert Null, so weist die betreffende Beobachtung die k-te Stufe der Prädiktors auf.

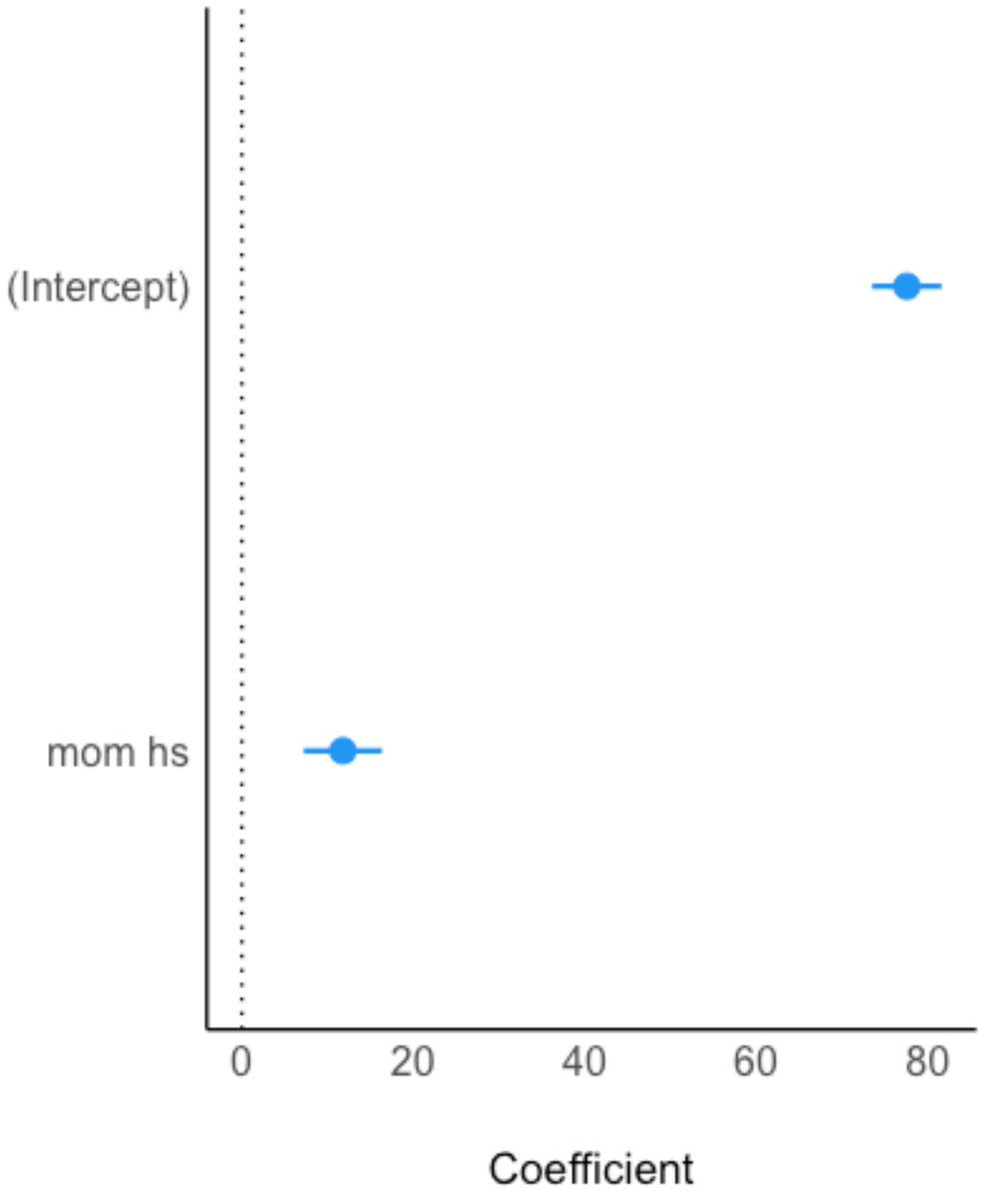
Einfache Regression mit einem nominalen, zweistufigen Prädiktor

IQ von Kindern, binärer Prädiktor



- Forschungsfrage: Unterscheidet sich der mittlere IQ-Wert (kid_score) von Kindern in Abhängigkeit davon, ob ihre jeweilige Mutter über einen Schlusabschluss (mom_hs) verfügt? (ceteris paribus)
- `m1: kid_score = 78 + 12*mom_hs + error`
- Der Achsensabschnitt (intercept, b0) ist der mittlere (bzw. vorhergesagte) IQ-Wert von Kindern, deren Mütter über *keinen* Schulabschluss verfügen:
- `kid_score = 78 + 0*12 + error`
- Das Regressionsgewicht (slope, b) ist der Unterschied im IQ-Wert von Kindern *von Mütter mit Schlulabschluss* (im Vergleich zum IQ-Wert von Kindern mit Mütter ohne Schlusabschluss). Dieser Unterschied entspricht der Steigung der Regressionsgerade:
- `78 + 1*12 + error = 90 + error`

Regressionskoeffizient als Mittelwertsdifferenz



- UV: binär (nominal zweistufig)
- AV: metrisch (quantitativ)
- Das Modell zeigt, dass der Unterschied im IQ auf ca. 12 Punkte geschätzt wird (im Vergleich von Kindern bei Müttern mit bzw. ohne Schulabschluss)
- Die Konfidenzintervalle (CI) geben Aufschluss über den Bereich plausibler Werte für die Schätzung
  - Der Wert 0.95 gibt die Genauigkeit des Schätzbereichs an
  - Je größer dieser Wert, desto ungenauer die Schätzung und desto größer der Schätzbereich

Parameter	Coefficient	CI	CI_low
(Intercept)	77.55	0.95	73.50
mom_hs	11.77	0.95	7.21

Eine metrische plus eine nominale UV

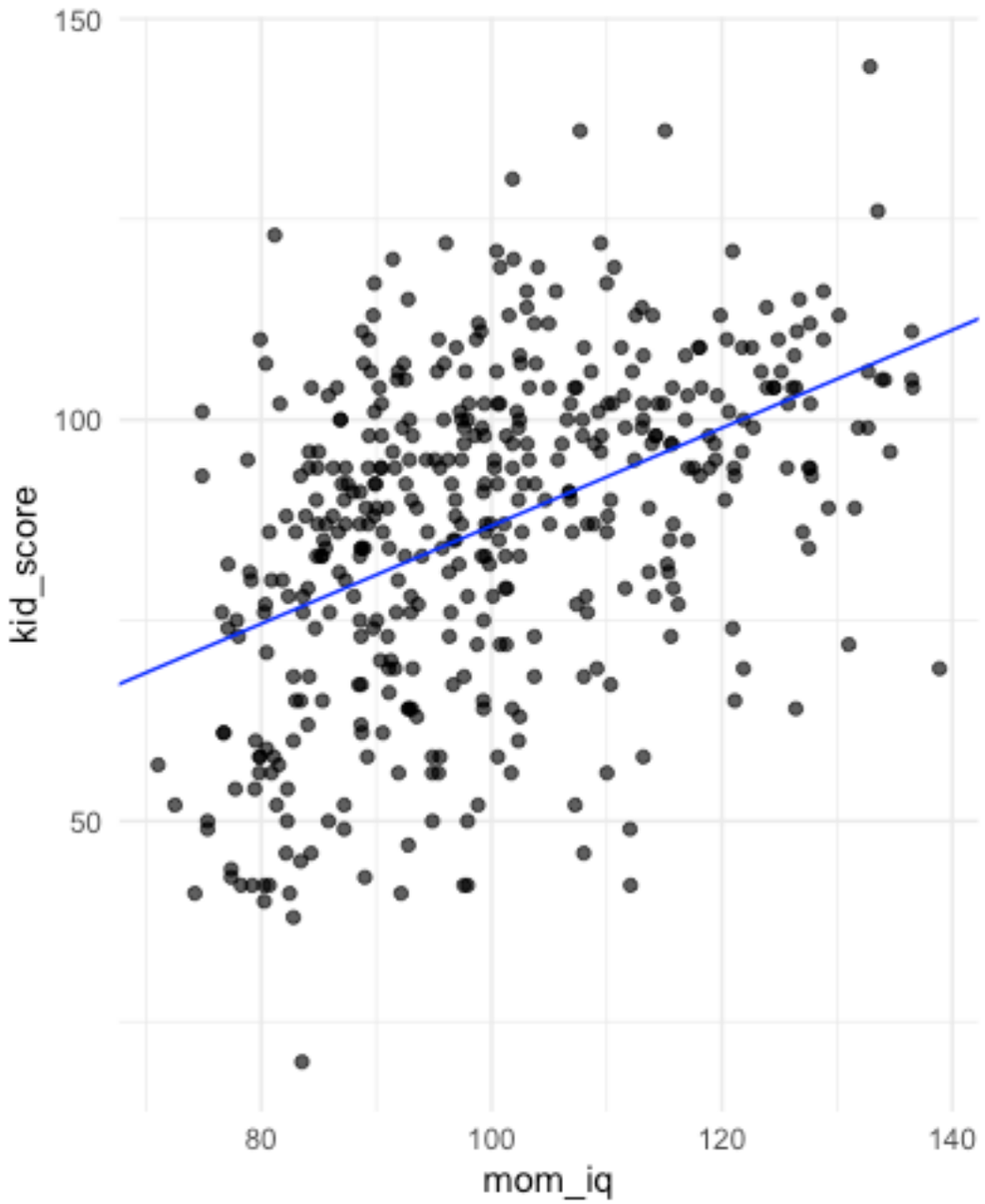
Vorhersage mit zwei Prädiktoren

- `kidiq ~ mom_iq + mom_hs`
- Wir sagen den IQ des Kindes vorher in Abhängigkeit von (als Funktion von) der IQ der Mutter sowie der Tatsache, ob die Mutter einen Schulabschluss besitzt.

variable	n	min	max	median	q1	q3	iqr	mad	mean	sd	se	ci
kid_score	434.00	20.00	144.00	90.00	74.00	102.00	28.00	19.27	86.80	20.41	0.98	1.93
mom_age	434.00	17.00	29.00	23.00	21.00	25.00	4.00	2.96	22.79	2.70	0.13	0.26
mom_hs	434.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00	0.79	0.41	0.02	0.04
mom_iq	434.00	71.04	138.89	97.92	88.66	110.27	21.61	15.89	100.00	15.00	0.72	1.42

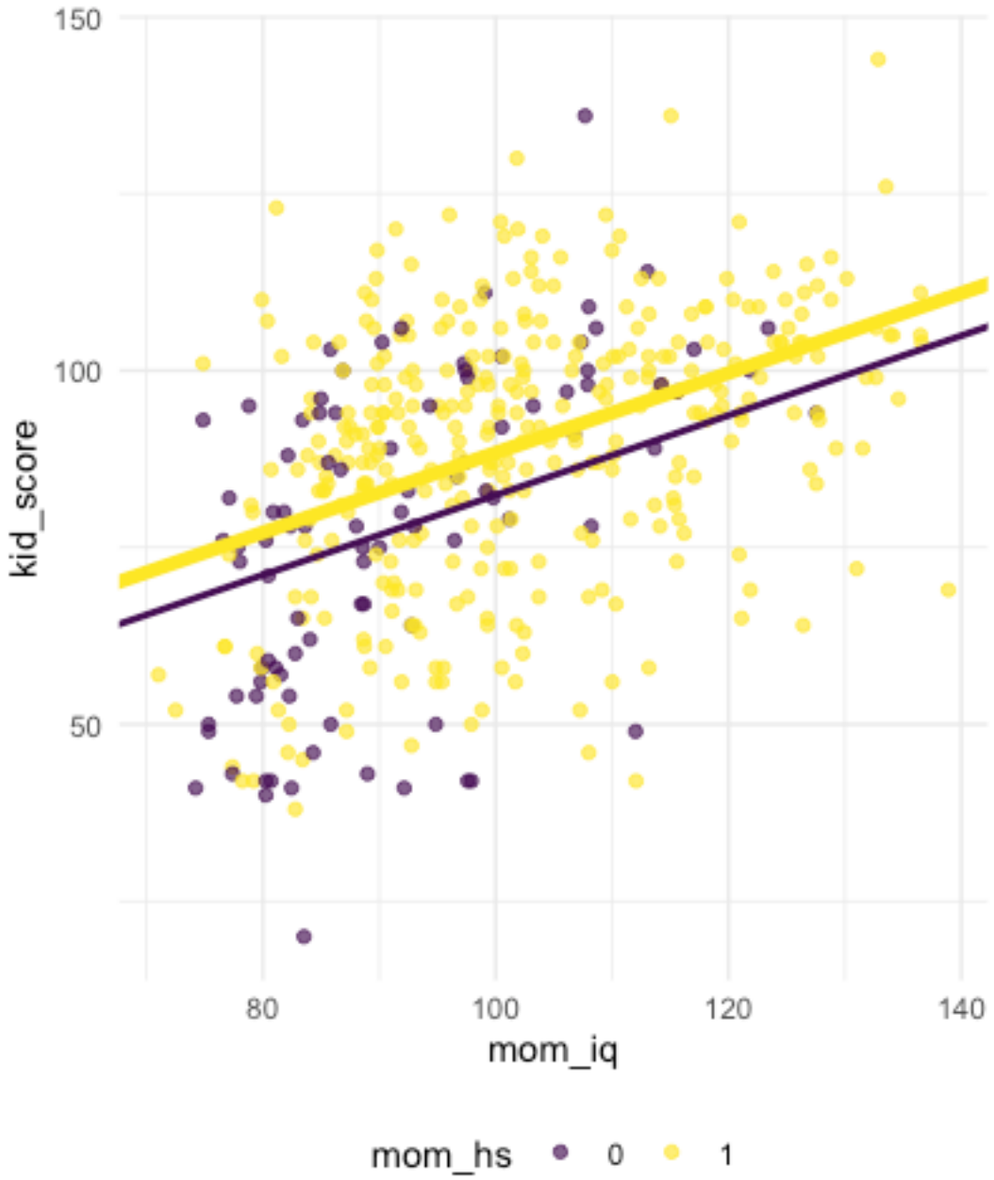
Ein metrischer Prädiktor

```
kid_score = 26 + 0.6 * mom_iq + error
```



- Die blaue Linie (Regressionsgerade) zeigt die vorhergesagten IQ-Werte der Kinder für verschiedene IQ-Werte der Mütter.
- Vergleicht man Teilpopulationen von Müttern mit mittleren Unterschied von einem IQ-Punkt, so findet man 0.6 IQ-Punkte Unterschied bei ihren Kindern im Durchschnitt.
- Der Achsenabschnitt hilft uns nicht weiter, da es keine Menschen mit einem IQ von 0 gibt.

Metrischer plus binärer Prädiktor



- **Achsenabschnitt**: Hat das Kind eine Mutter mit einem IQ von 0 und ohne Schulabschluss, dann schätzt das Modell den IQ-Wert des Kindes auf ca. 26.
- **Koeffizient zum mütterlichen Schulabschluss**: Vergleicht man Kinder von Müttern gleicher Intelligenz, aber mit Unterschied im Schulabschluss, so sagt das Modell einen Unterschied von ca. 6 Punkten im IQ voraus.
- **Koeffizient zur mütterlichen IQ**: Vergleicht man Kinder von Müttern mit gleichem Wert im Schulabschluss, aber mit 1 IQ-Punkt Unterschied, so sagt das Modell einen Unterschied von ca. 0.6 IQ-Punkten bei den Kindern voraus.

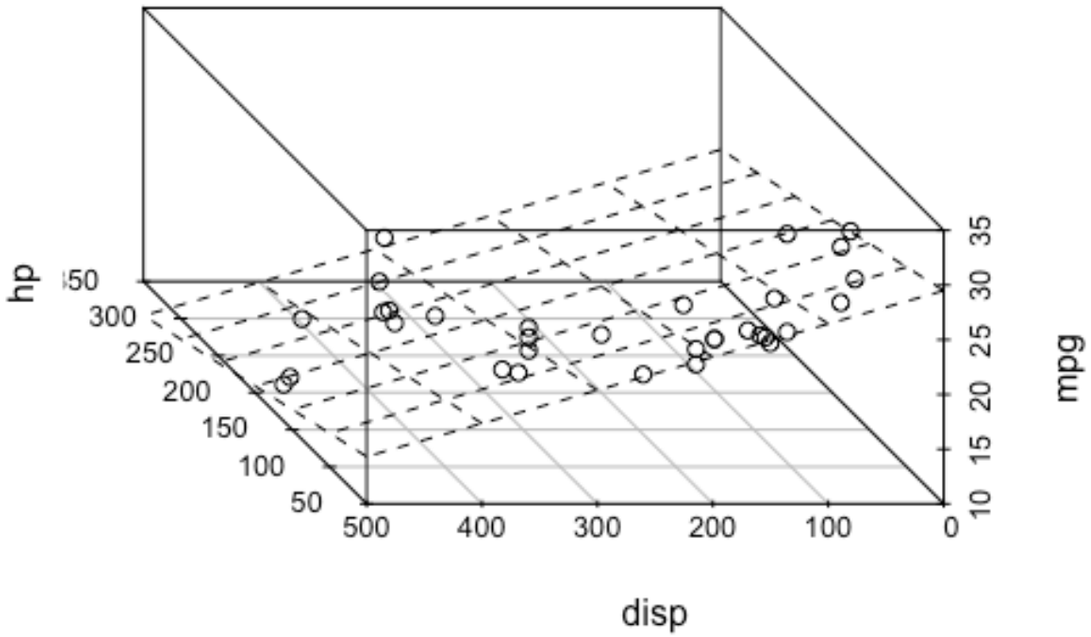
Parameter	Coefficient	CI	CI_low
(Intercept)	25.73	0.95	14.18
mom_iq	0.56	0.95	0.44
mom_hs	5.95	0.95	1.60

```
m3: kid_score = 26 + mom_hs + 0.6*mom_iq + error
```

Multiple Regression

Regression mit mehreren Prädiktoren

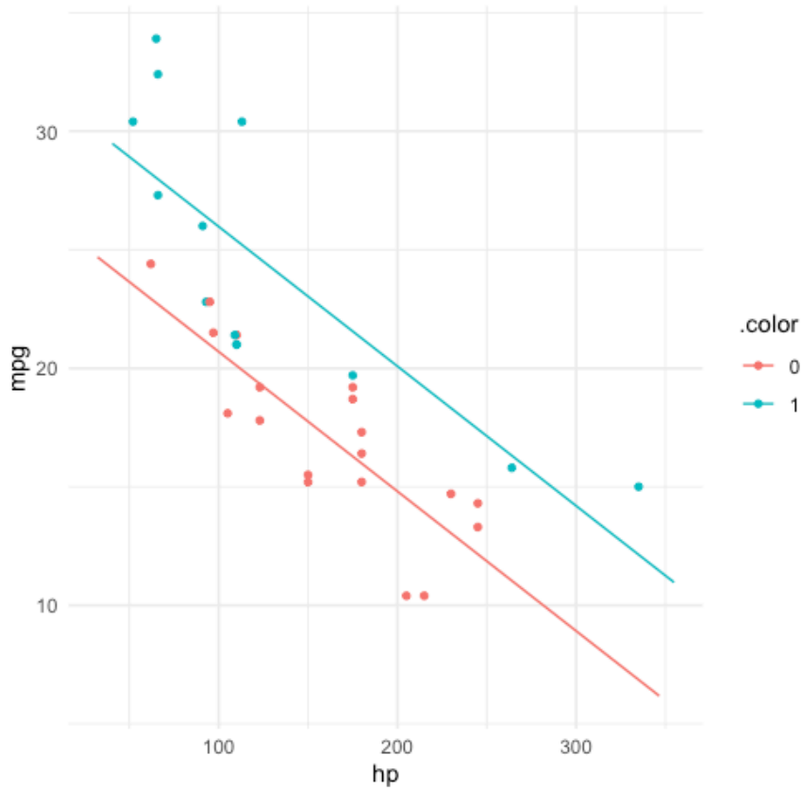
$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + e_i$$



- [3D-Diagramm] interaktiv

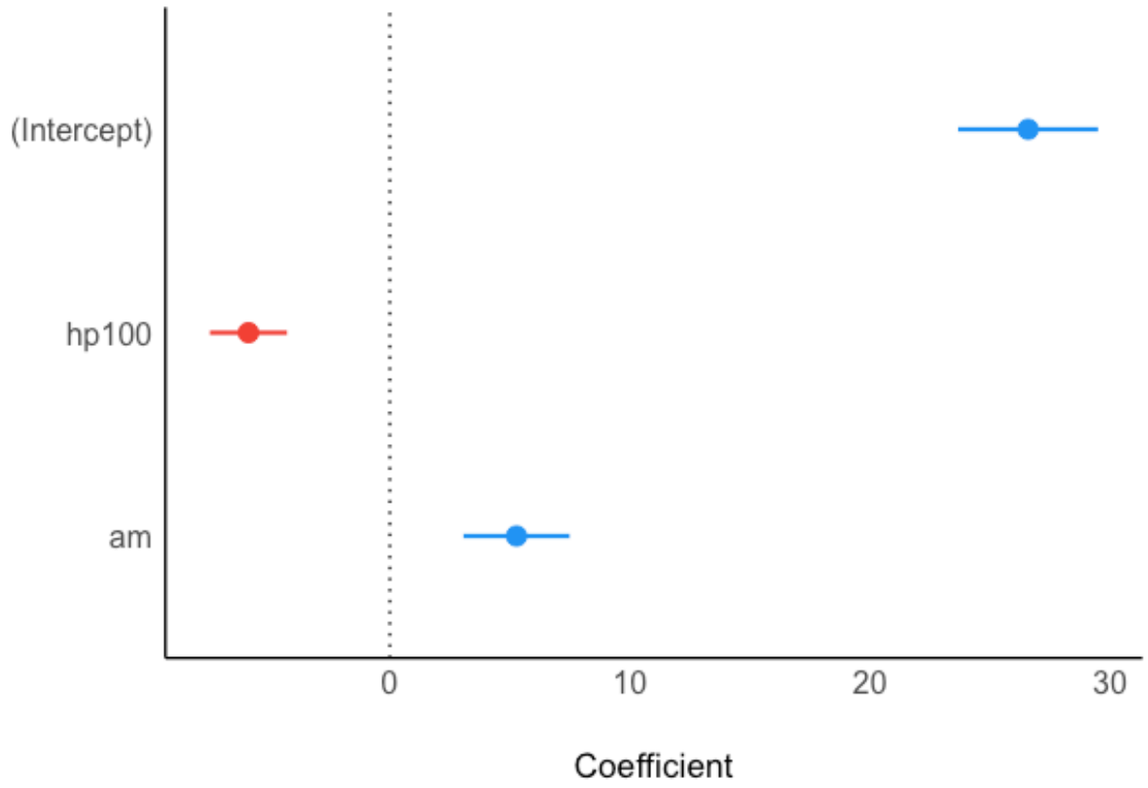
- Bei einer **multiplen Regressionsanalyse** wird versucht, den Wert einer Kriteriumsvariable (für eine Beobachtung) durch **mehrere (p) Prädiktorvariablen** vorherzusagen. Ansonsten ähnelt die multiple Regression der einfachen Regression.
- Multiple Regressionsanalysen sind dann sinnvoll, wenn die Vorhersagegüte steigt durch Hinzunahme weiterer Prädiktoren.
- Jeder Prädiktor hat dabei ein **Einflussgewicht** (auch: Regressionsgewicht; Geradensteigung; Koeffizient); diese Werte sind jeweils **bereinigt** von den Werten der anderen Prädiktoren. Eine multiple Regression ist mehreren einfachen Regressionen daher vorzuziehen.
- Das bedeutet, man betrachtet den den Zusammenhang eines Prädiktors mit der AV, wobei man gleichzeitig den anderen Prädiktor konstant hält.
- Bei zwei Prädiktorvariablen kann man sich das Modell als Ebene im Raum vorstellen (anstelle einer Geraden): stellen Sie sich ein Blatt Papier vor, das durch einen Bienenschwarm gelegt wird.

Modell mit zwei Prädiktoren



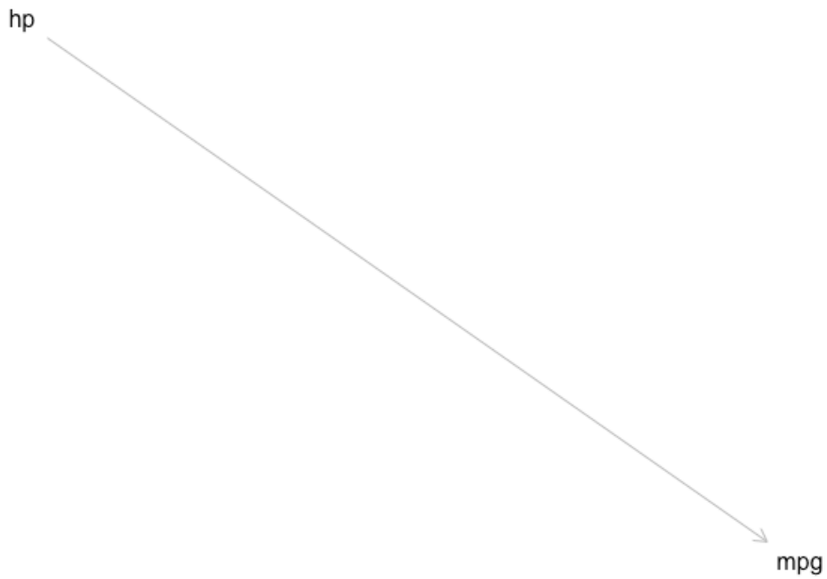
- Wir sagen Spritverbrauch vorher auf Basis zweier Prädiktoren: hp und am.
- Jeder (der drei) Modellkoeffizient (b0, b1, b2) gibt den Wert für mpg (Y) an, unter der Bedingung, dass die anderen Modellkoeffizienten Null sind.

$$\text{mpg} = b_0 + b_1\text{hp} + b_2\text{am} + e$$

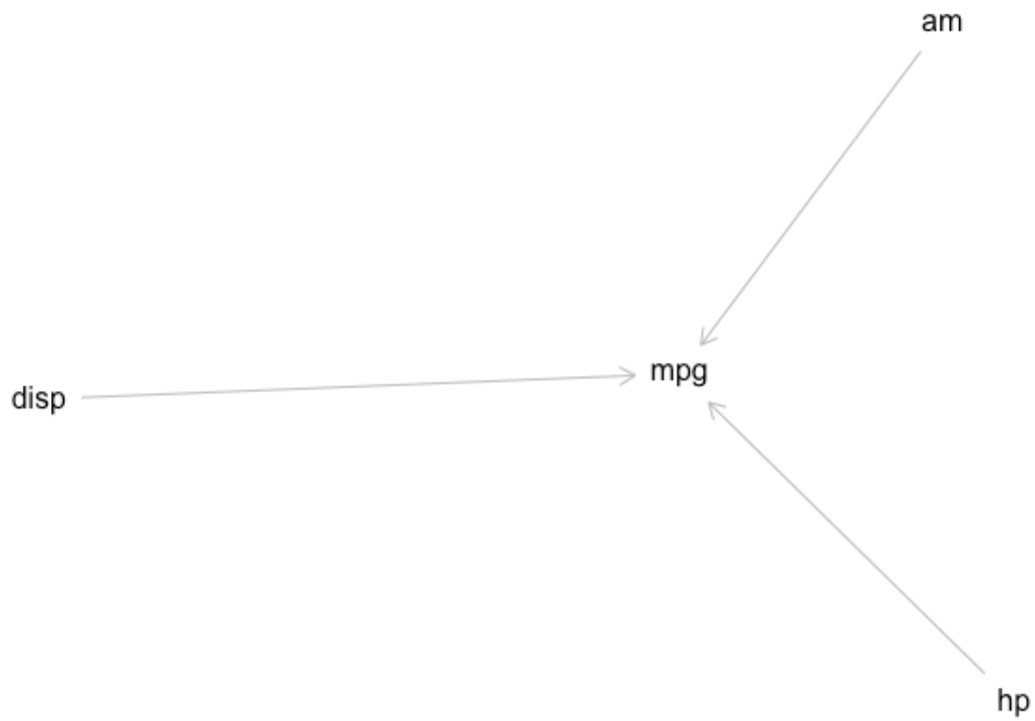


Welches Modell liefert wohl bessere Vorhersagen?

A



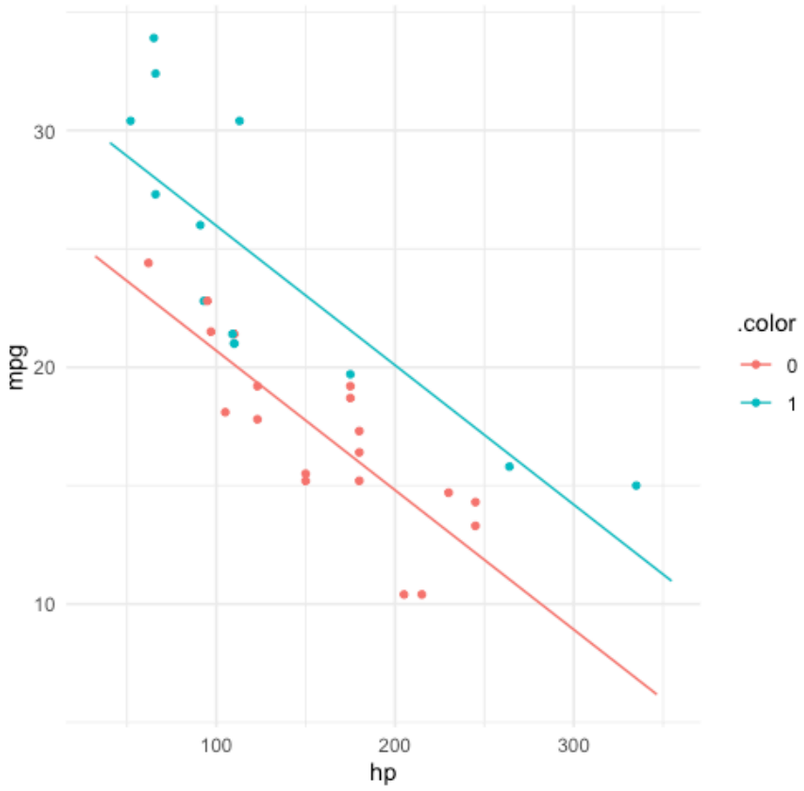
B



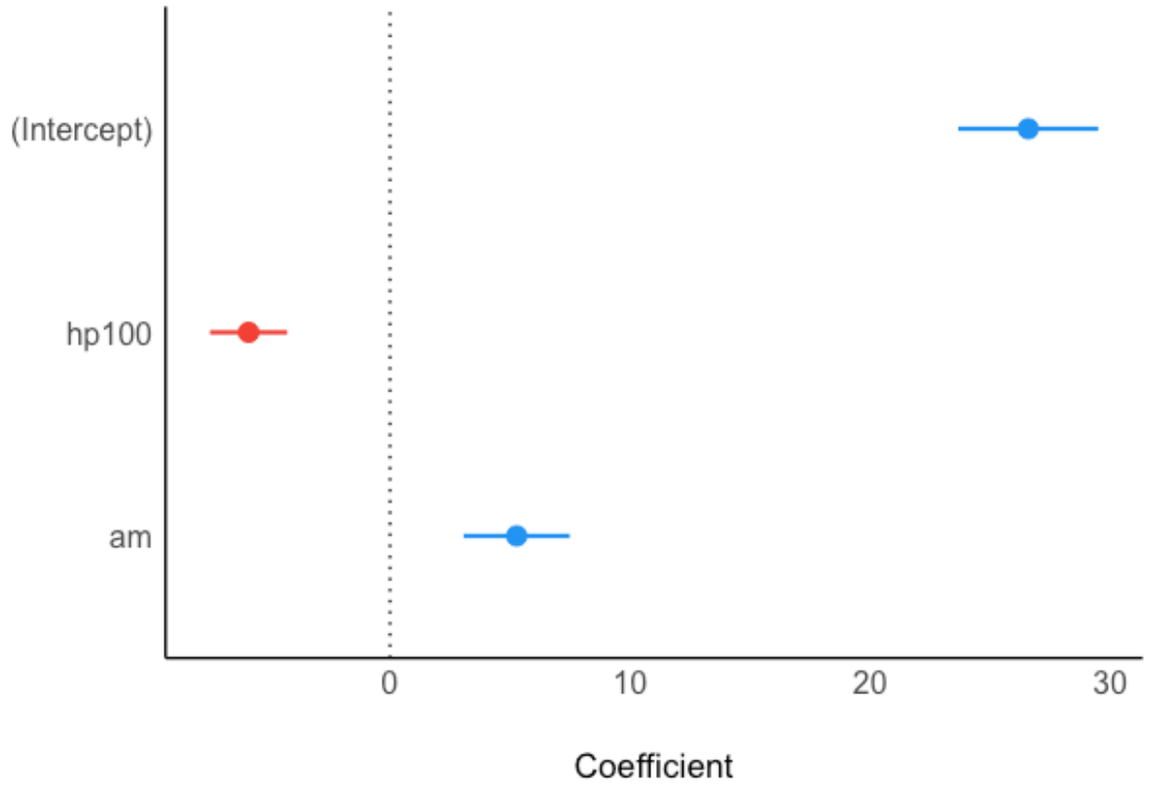
Sofern ein Modell mehr relevante Prädiktoren beinhaltet, wird es bessere Vorhersagen machen (als ein Modell mit weniger relevanten Variablen), unter sonst gleichen Umständen (ceteris paribus).

Ausgabe eines Regressionsmodells

$$\text{mpg} = b_0 + b_1\text{hp} + b_2\text{am} + e$$

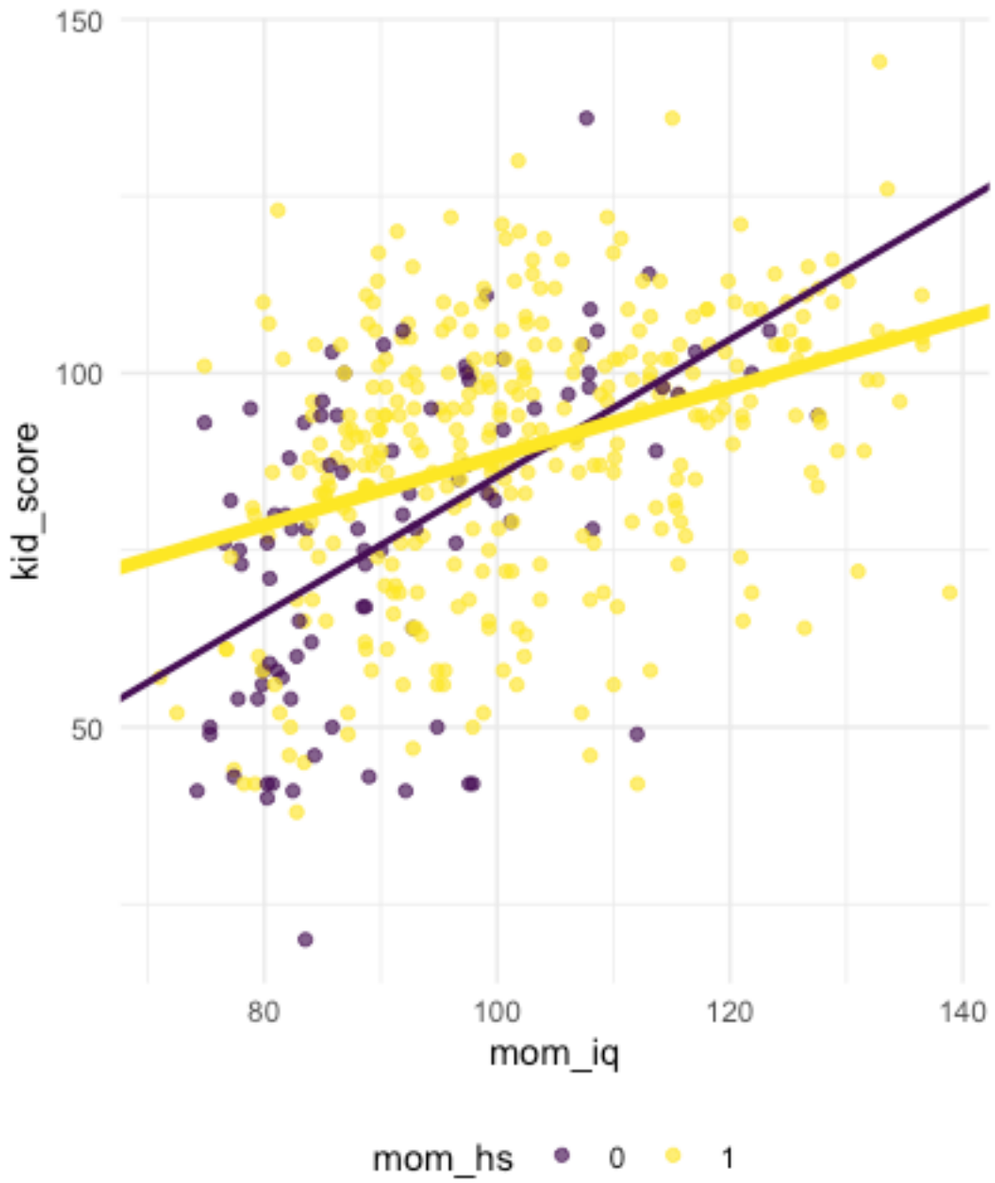


Parameter	Coefficient	CI	CI_low
(Intercept)	26.58	0.95	23.67
hp100	−5.89	0.95	−7.50
am	5.28	0.95	3.07



Interaktion

Interaktion von Mutter-IQ und Schulabschluss



- In Modell m3 haben wir die Regressionsgeraden gezwungen, **parallel** zu sein.
- Betrachtet man das Streudiagramm, so sieht man, das nicht-parallele Geraden besser passen.
- In m4 erlauben wir den Regressionsgeraden der Gruppen, nicht mehr parallel zu sein, was die Modellgüte erhöht
- Sind die Regressionsgeraden nicht parallel, so spricht man von einer **Interaktion** (synonym: Interaktionseffekt, Moderation).
- Liegt eine Interaktion vor, so unterscheidet sich also die Steigung in den Gruppen.

```
m4: kid_score ~ mom_iq + mom_hs + error
```

Parameter	Coefficient	CI	CI_low
(Intercept)	−11.48	0.95	−38.52
mom_iq	0.97	0.95	0.68
mom_hs	51.27	0.95	21.12
mom_iq:mom_hs	−0.48	0.95	−0.80

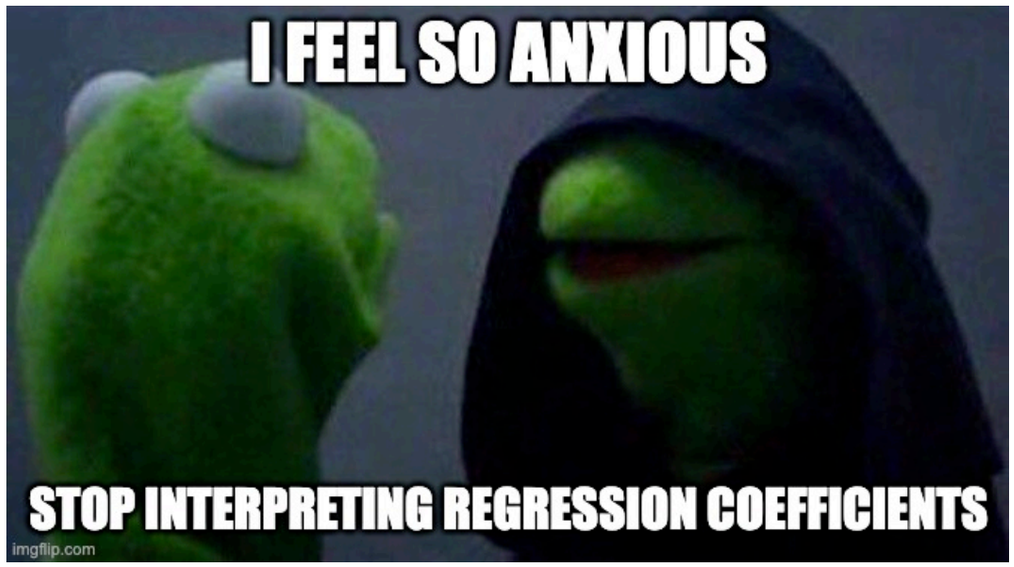
Interpretation einer Interaktion

- **Achsenabschnitt**: IQ-Schätzwerte für Kinder mit Mütter ohne Abschluss und mit einem IQ von 0. Kaum zu interpretieren.
- **mom_hs**: Unterschied der IQ-Schätzwerte zwischen Kindern mit Mutter ohne bzw. mit Schulabschluss und jeweils mit einem IQ von 0. Puh.
- **mom_iq**: Unterschied der IQ-Schätzwerte zwischen Kindern mit Müttern, die sich um einen IQ-Punkt unterscheiden aber jeweils ohne Schulabschluss (mom_hs = 0).
- **Interaktion**: Der Unterschied in den Steigungen der Regressionsgeraden, also der Unterschied des Koeffizienten für mom_iq zwischen Müttern mit bzw. ohne Schulabschluss.

```
mom_hs=0:
kid_score = -11 + 51*0 + 1.1* mom_iq + 0.5*0*mom_iq = -11 + 1.1*mom_iq

mom_hs=1:
kid_score = -11 + 51*1 + 1.1* mom_iq + 0.5*1*mom_iq = 40 + 0.6*mom_iq
```

Oft ist eine normale Regression schwer zu interpretieren



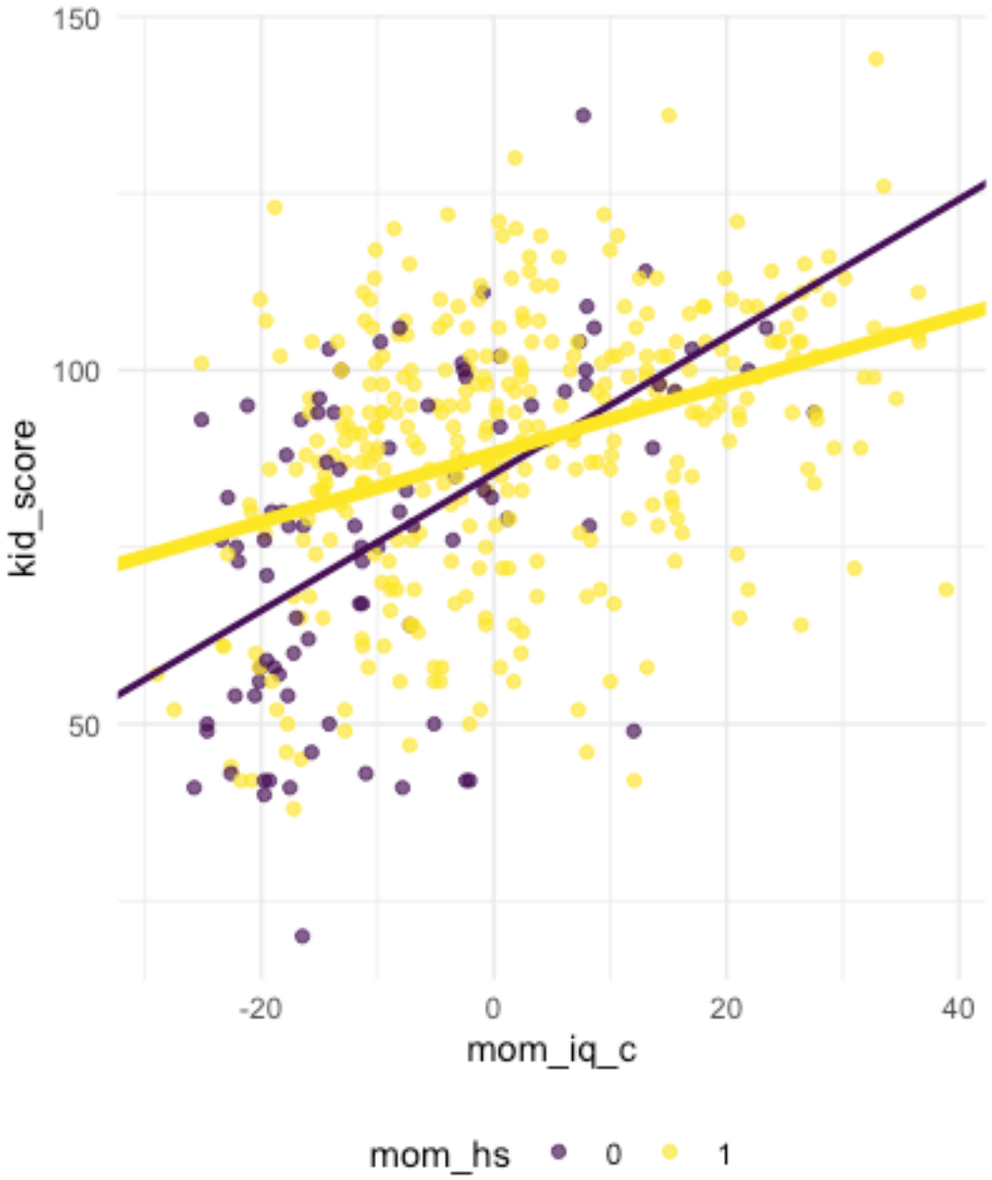
Quelle

- Da Regressionskoeffizienten sich darauf beziehen, dass die übrigen Koeffizienten den Wert Null haben, sind sie oft schwer (kaum) zu interpretieren.
- Daher ist es oft sinnvoll, die Rohvariablen zu zentrieren.
- Zentrierte Prädiktoren erlauben oft eine viel einfachere Interpretation der Koeffizienten.
- Unter Zentrieren (to center) versteht man das Bilden der Differenz eines Messwerts zu seinem Mittelwert.
- Zentrierte Werte (c wie centered) geben also an, wie weit ein Messwert vom mittleren (typischen) Messwert entfernt ist.

$$x_c = x - \bar{x}$$

Kermit kriegt zu viel

Modell mit zentrierten Prädiktoren

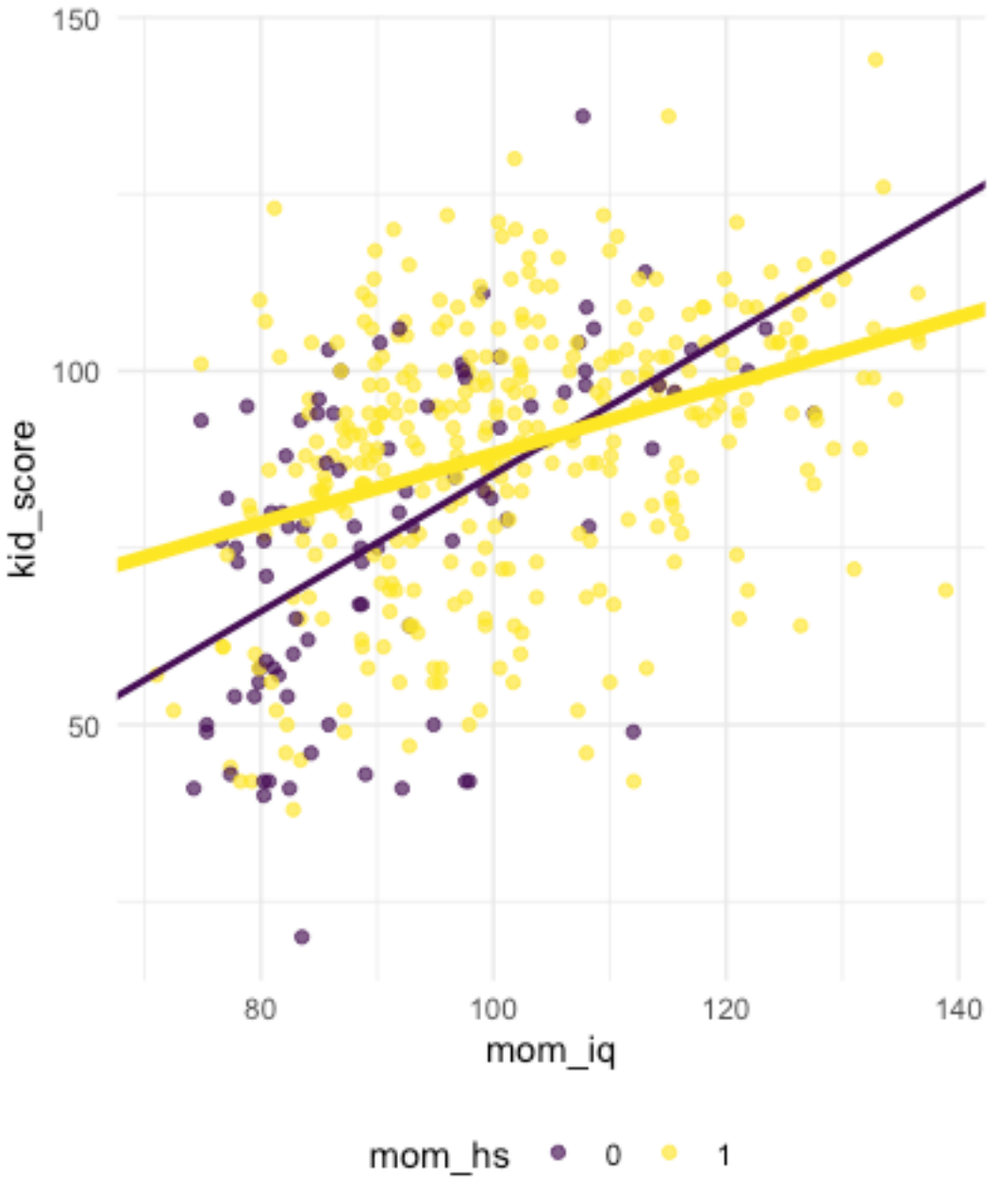


- Der Achsenabschnitt (Intercept) gibt den geschätzten IQ des Kindes an, wenn man eine Mutter mittlerer Intelligenz und ohne Schulabschluss betrachtet.
- mom_hs gibt den Unterschied im geschätzten IQ des Kindes an, wenn man Mütter mittlerer Intelligenz aber mit bzw. ohne Schlusabschluss vergleicht.
- mom_iq_c gibt den Unterschied im geschätzten IQ des Kindes an, wenn man Mütter ohne Schlusabschluss aber mit einem IQ-Punkt Unterschied vergleicht.
- mom_hs:mom_iq_c gibt den Unterschied in den Koeffizienten für mom_iq_c an zwischen den beiden Grupen von mom_hs.

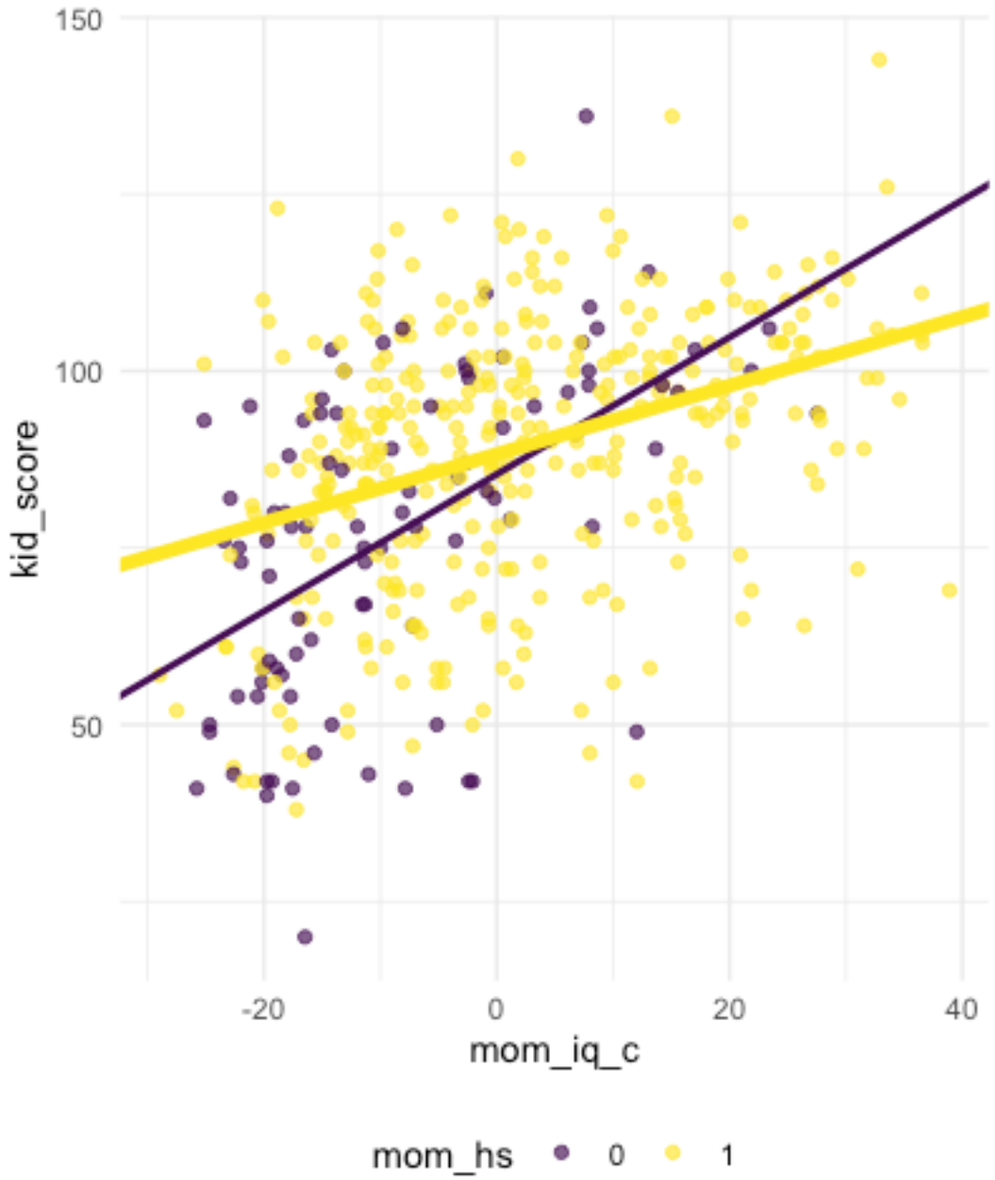
```
m5: kid_score ~ mom_iq_c + mom_iq_c + mom_iq_c:mom_hs
```

Zentrieren ändert nichts an den Vorhersagen

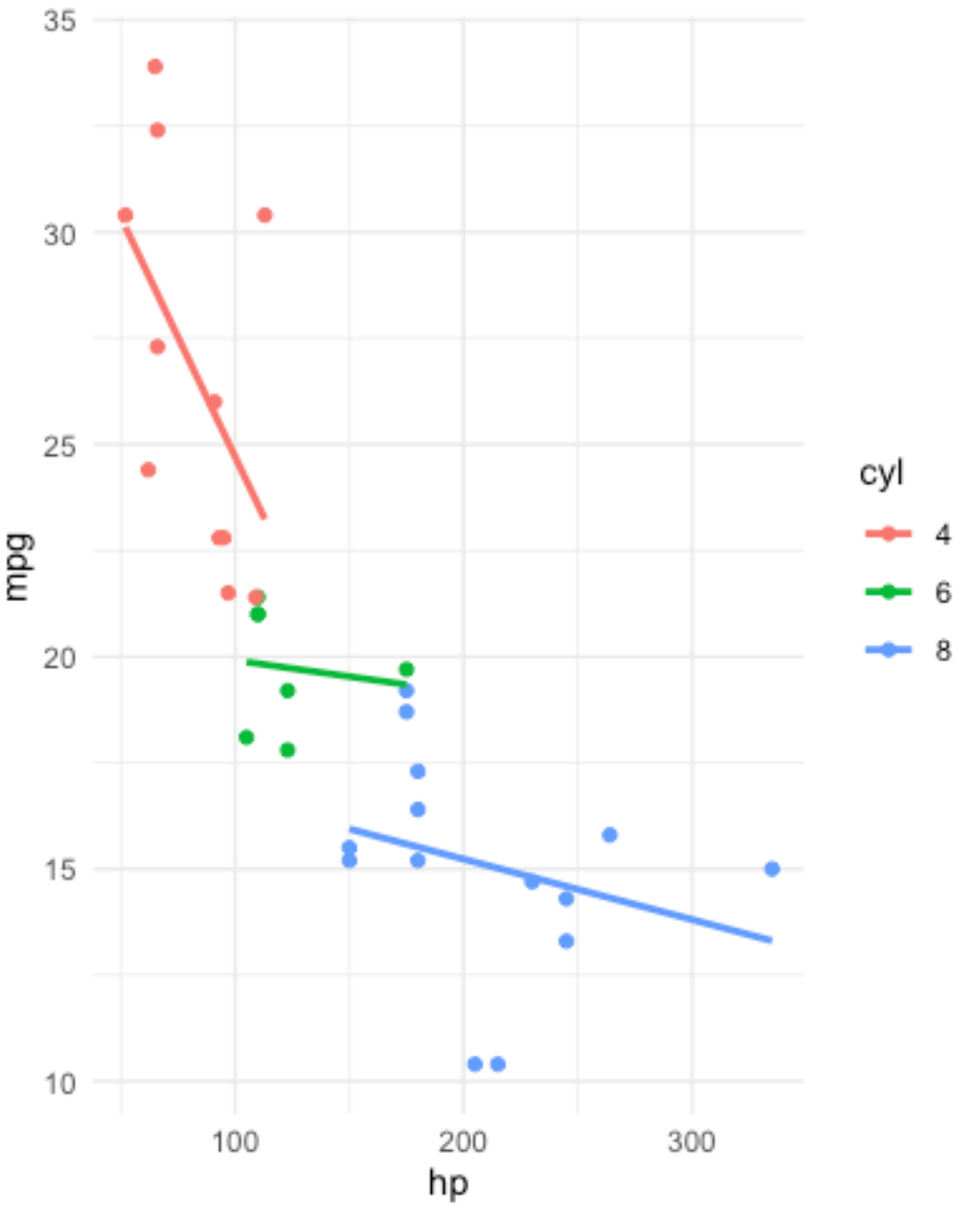
m4: unzentrierte Prädiktoren



m5: zentrierte Prädiktoren

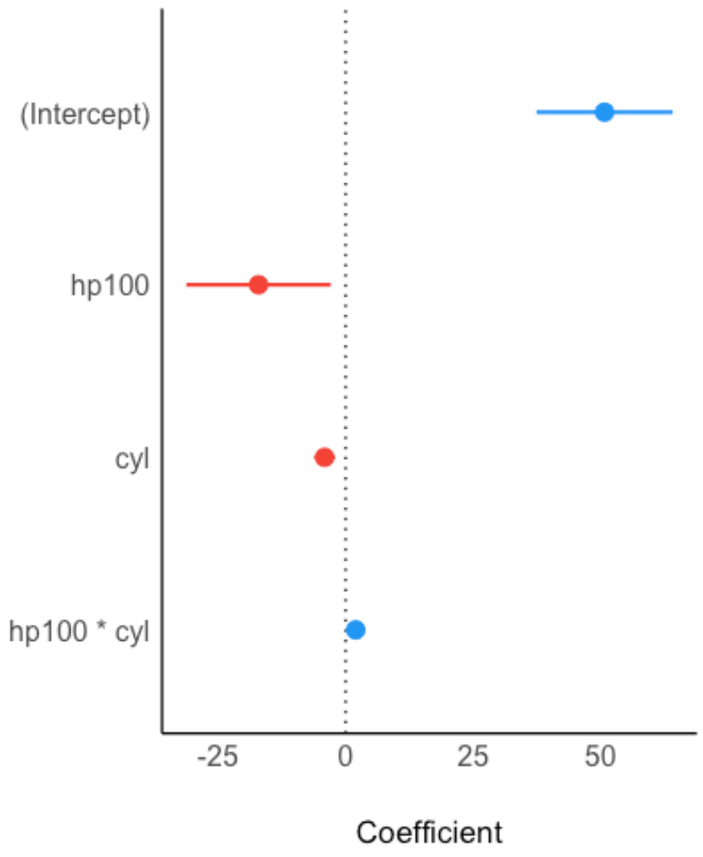


Interaktionseffekt bei mtcars



- Wirkt sich vielleicht die PS-Zahl unterschiedlich aus (auf den Spritverbrauch) je nach Anzahl der Zylinder des Autos?
- Hängt die Steigung einer Regressionsgeraden ab von der Ausprägung eines anderen Prädiktors, so liegt ein Interaktionseffekt (synonym: Wechselwirkung, Moderation) vor.
- Im Diagramm erkennt man einen **Interaktionseffekt** daran, dass die Regressionsgeraden **nicht parallel** sind.
- Der statistische Effekt eines Prädiktors ist dann abhängig von den Ausprägungen eines anderen Prädiktors.
- Die Steigung der Regressionsgeraden ist unterschiedlich je nach Gruppe (von cyl).

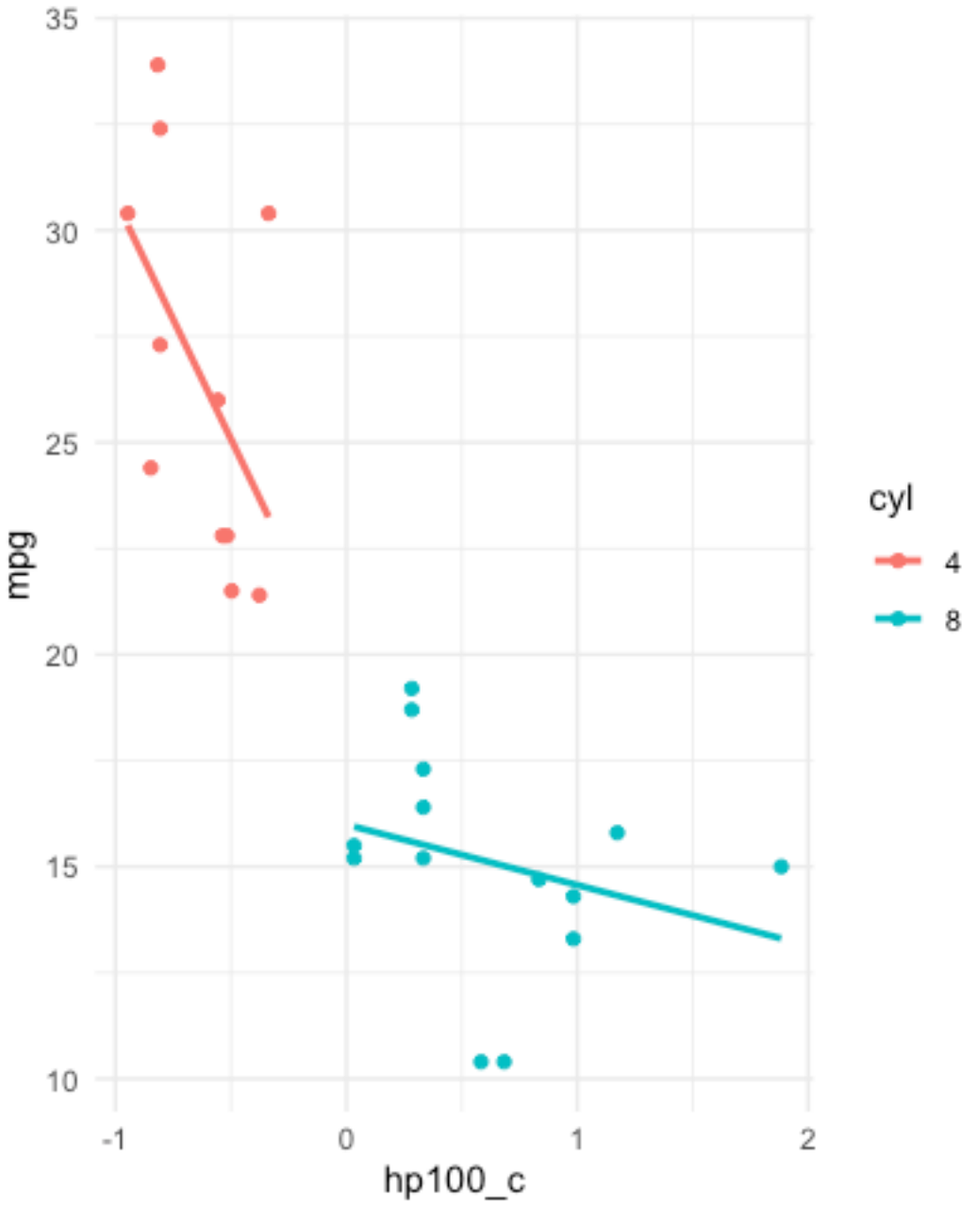
Koeffizienten bei Interaktionsmodell für mtcars



Parameter	Coefficient	CI	CI_low
(Intercept)	50.75	0.95	37.41
hp100	−17.07	0.95	−31.22
cyl	−4.12	0.95	−6.14
hp100:cyl	1.97	0.95	0.17

- Der Parameter **hp100**
  - gibt den Unterschied (im Verbrauch) zweier Autos an, die sich um 100 PS unterscheiden
  - unter der Annahme, dass die übrigen Prädiktoren gleich Null sind.
  - entspricht der Steigung der Regressionsgeraden.
- Der Parameter **cyl**
  - gibt den Unterschied im Verbrauch zweiter Autos an, die sich um einen Zylinder unterscheiden
  - unter der Annahme, dass die übrigen Prädiktoren gleich Null sind
- Der Parameter **hp100:cyl**
  - gibt den zusätzlichen Unterschied im Parameter hp100 an, wenn man Autos vergleicht, die sich in einem Zylinder unterscheiden.

mtcars-Interaktionsmodell vereinfacht



- **Intercept**: Autos mit mittlerer PS-Zahl und mit 4 Zylindern kommen mit einer Gallone Sprit im Schnitt ca. 19 Meilen weit
- **hp100_c**: Für je 100 PS mehr sinkt die Reichweite um im Schnitt ca. 11 Meilen, wenn das Auto 4 Zylinder hat
- **hp100_c:cyl8**: Verfügt es über 8 Zylinder, dann verringert sich die Reichweite im Schnitt hingegen nur um ca. 1 Meile (für je 100 PS mehr), da -11+10=-1
- **cyl8**: Autos mit mittlerer PS-Zahl und 8 Zylindern haben im Schnitt eine um ca. 4 Meilen (3.45) geringere Reichweite (als Autos mit 4 Zylindern und mittlerer PS-Zahl)

Parameter	Coefficient
(Intercept)	19.44
hp100_c	−11.28
cyl8	−3.45
hp100_c:cyl8	9.85

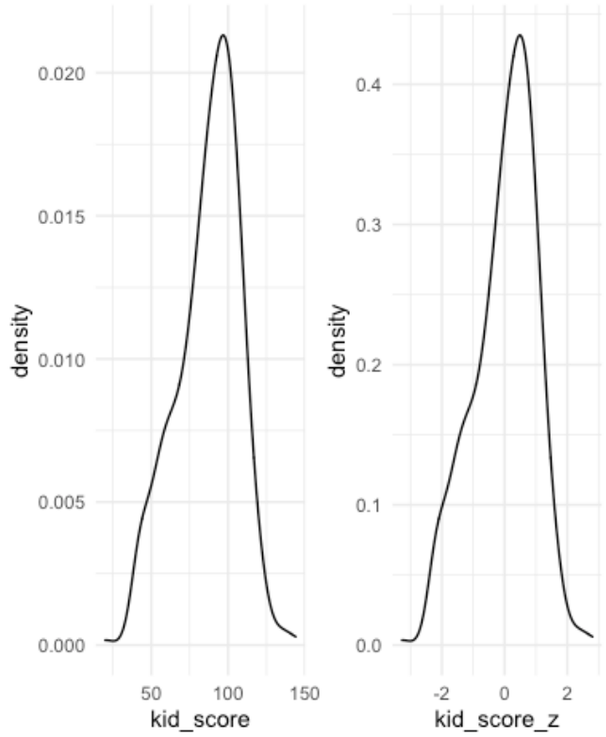
Prädiktorenrelevanz

Woher weiß man, welcher Prädiktor am wichtigsten ist?

- Welcher Prädiktor ist nun "wichtiger" oder "stärker" in Bezug auf den Zusammenhang mit der AV, mom_iq oder mom_age?
  - mom_iq hat den größeren Koeffizienten.
  - mom_age hat weniger Streuung.
- Um die Relevanz der Prädiktoren vergleichen zu können, müsste man vielleicht die Veränderung von kid_score betrachten, wenn man von kleinsten zum größten Prädiktorwert geht.
- Allerdings sind Extremwerte meist instabil (da sie von einer einzigen Beobachtung bestimmt werden).
- Sinnvoller ist es daher, die Veränderung in der AV zu betrachten, wenn man den Prädiktor von "unterdurchschittlich" auf "überdurchschnittlich" ändert.
- Das kann man mit z-Standardisierung erreichen.
- z-Standardisierung bedeutet, eine Variable so zu transformieren, dass sie über einen Mittelwert von 0 und eine SD von 1 verfügt.

$$z = \frac{x - \bar{x}}{sd(x)}$$

Modell mit z-standardisierten Prädiktoren



Parameter	Coefficient
(Intercept)	86.80
mom_iq_z	9.05
mom_age_z	1.05

- Der Achsenabschnitt gibt den Mittelwert der AV (kid_score) an, da kid_score_z = 0 identisch ist zum Mittelwert von kid_score.
- Der Koeffizient für mom_iq_z gibt an, um wie viele SD-Einheiten sich kid_score (die AV) ändert, wenn sich mom_iq um eine SD-Einheit ändert.
- Der Koeffizient für mom_age_z gibt an, um wie viele SD-Einheiten sich kid_score (die AV) ändert, wenn sich mom_age um eine SD-Einheit ändert.
- Jetzt sind die Prädiktoren in ihrer Relevanz (Zusammenhang mit der AV) vergleichbar.
- Man sieht, dass die Intelligenz der Mutter deutlich wichtiger ist das Alter der Mutter (im Hinblick auf die Vorhersage bzw. den Zusammenhang mit mit der AV).
- Der Wertebereich der Koeffizienten wird dadurch homogenisiert.

```
m6: kid_score ~ mom_iq_z + mom_age_z
```

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.