



# Thema 11: Vertiefung zur Regressionsanalyse

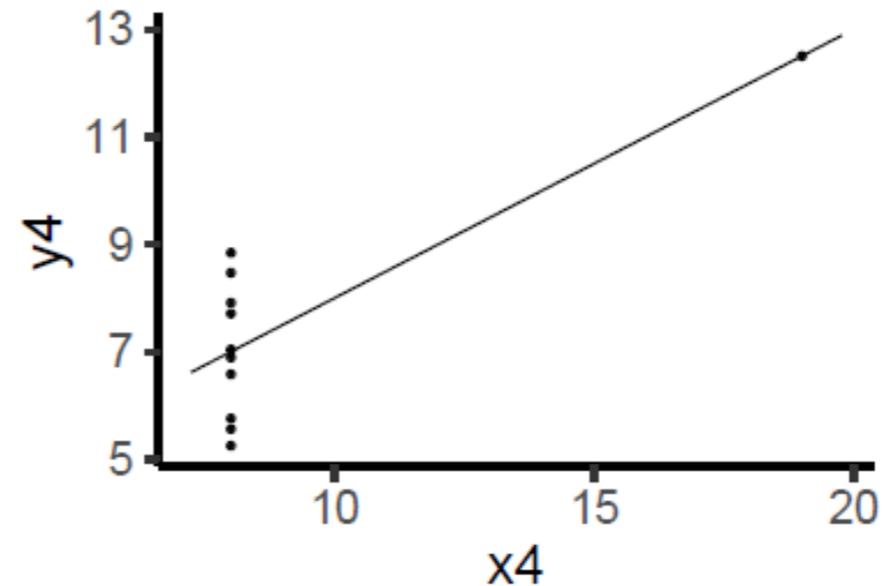
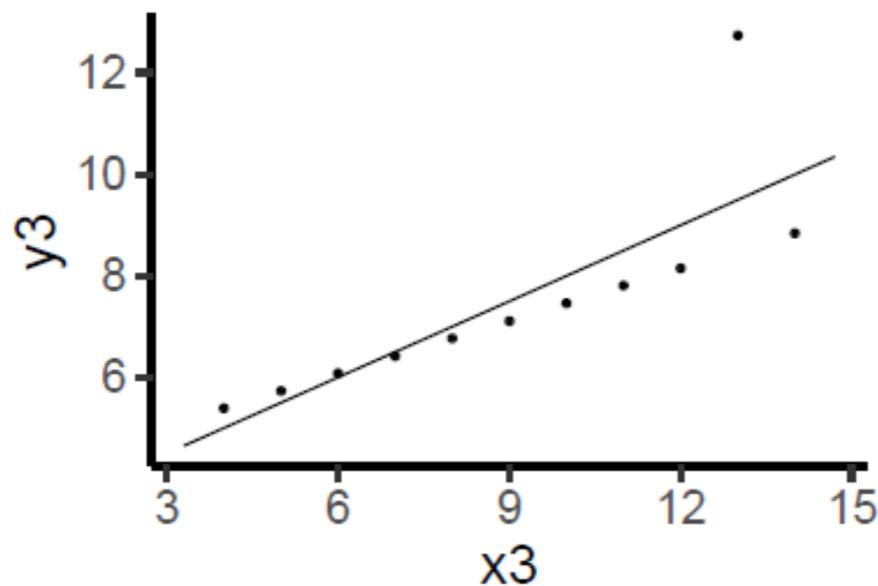
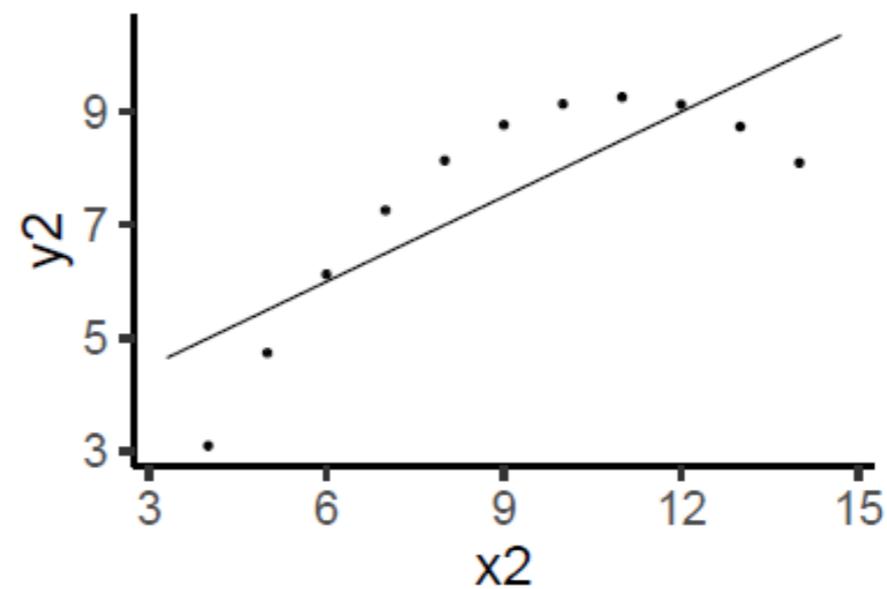
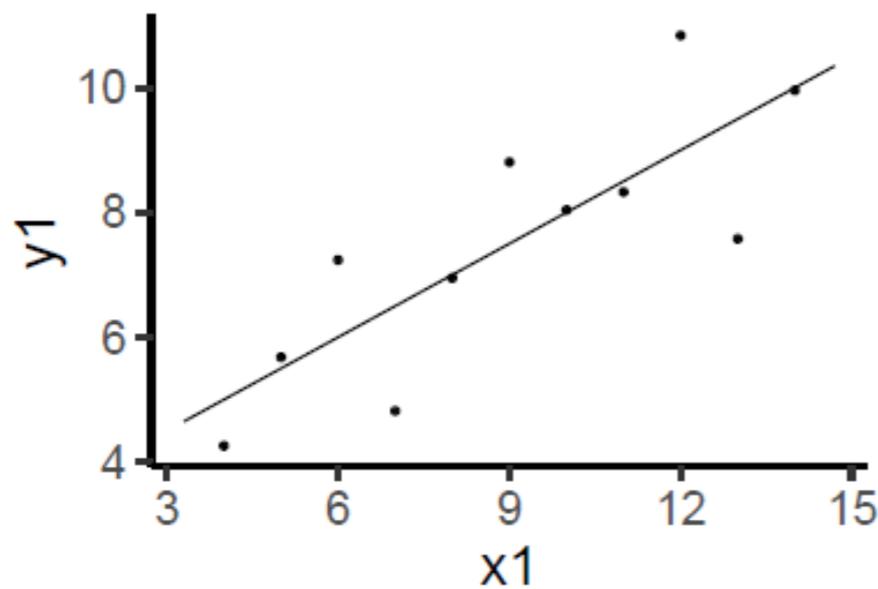
QM1, SoSe 22

# Grenzen der linearen Regression

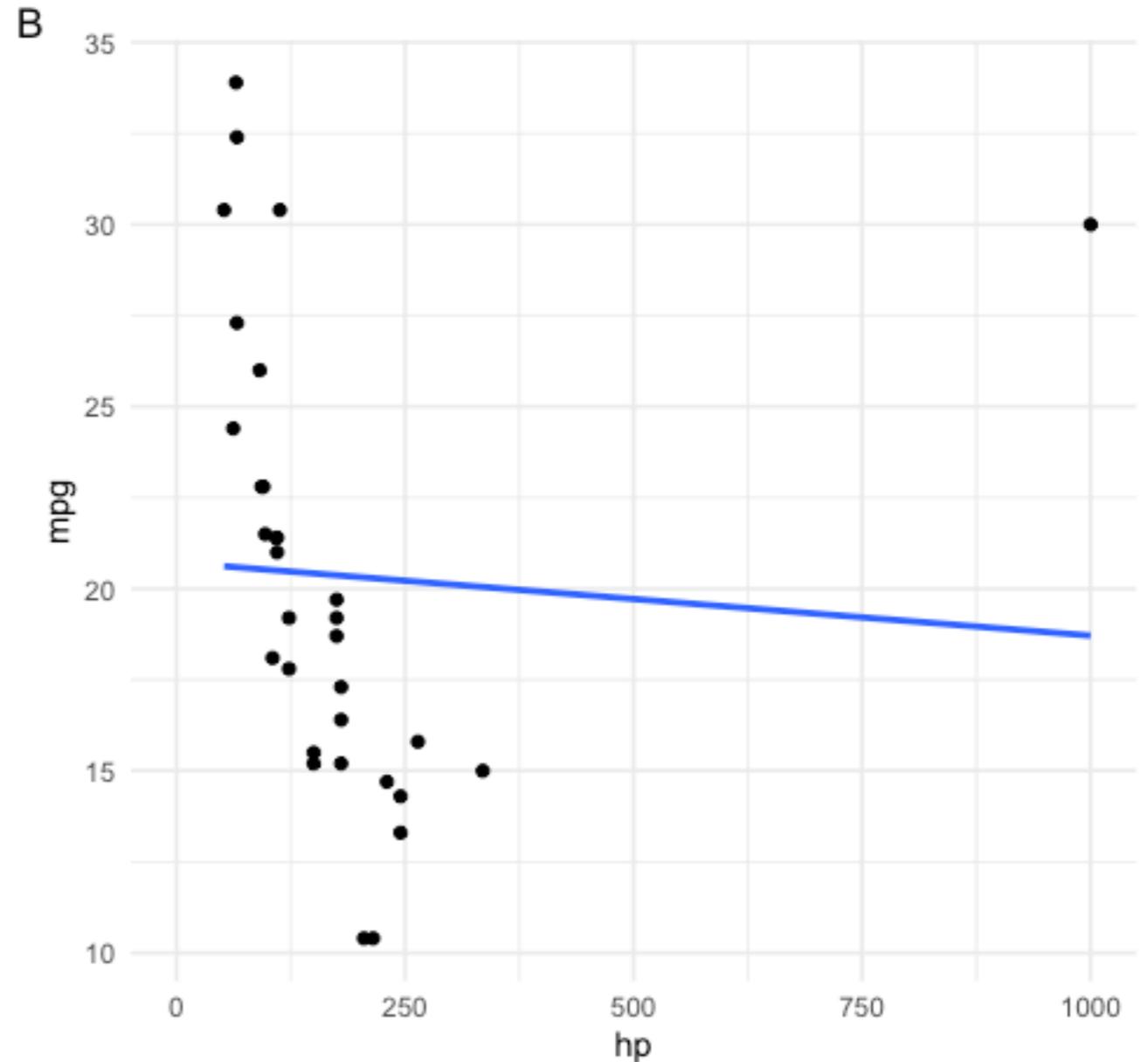
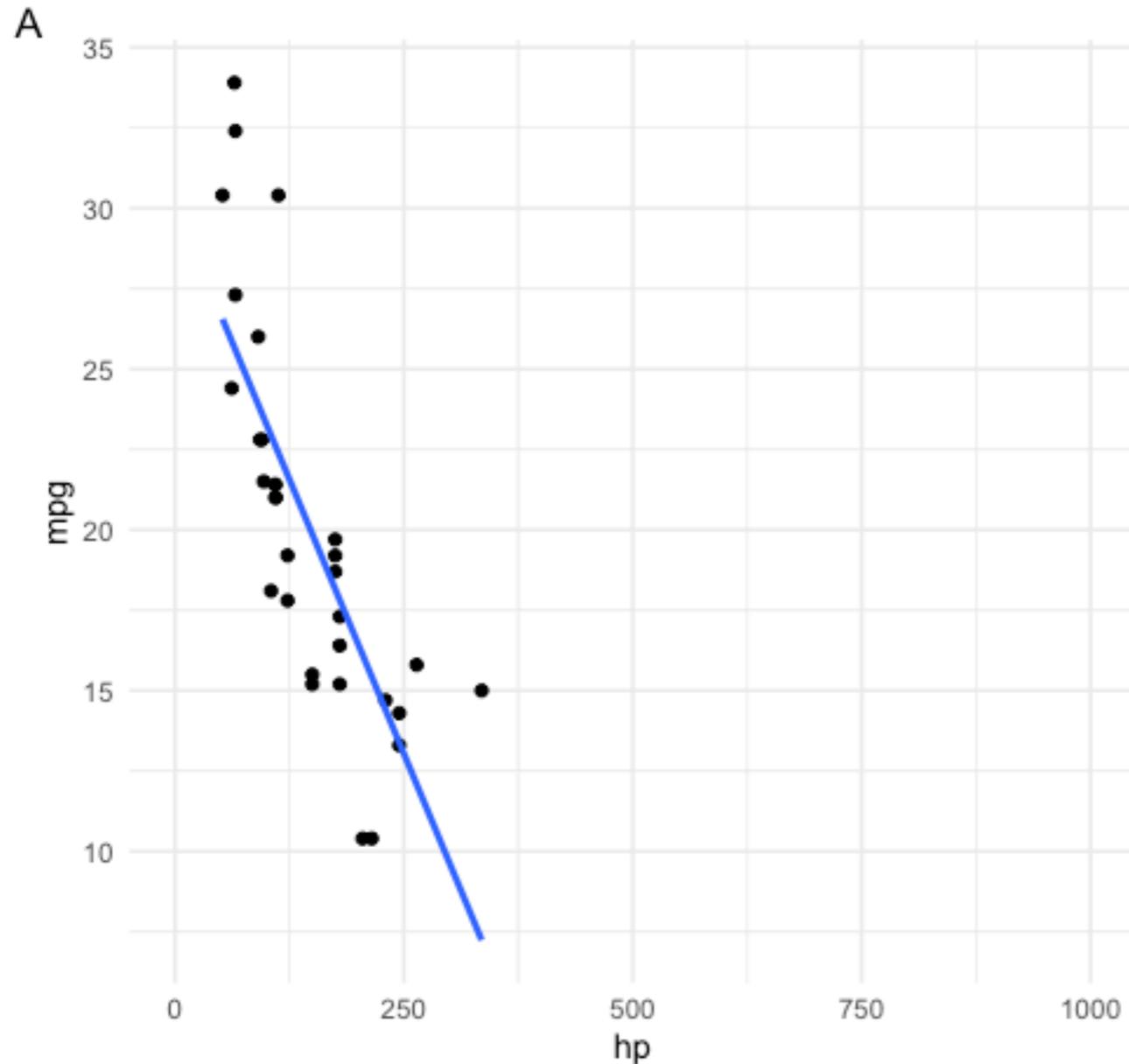
# Anscombes Quartett revisited: Nur lineare Trends

Bei den vier Datensätzen im Anscombe-Quartett ist das R-Quadrat immer (in etwa) gleich.

$$R^2 \approx 0.67$$



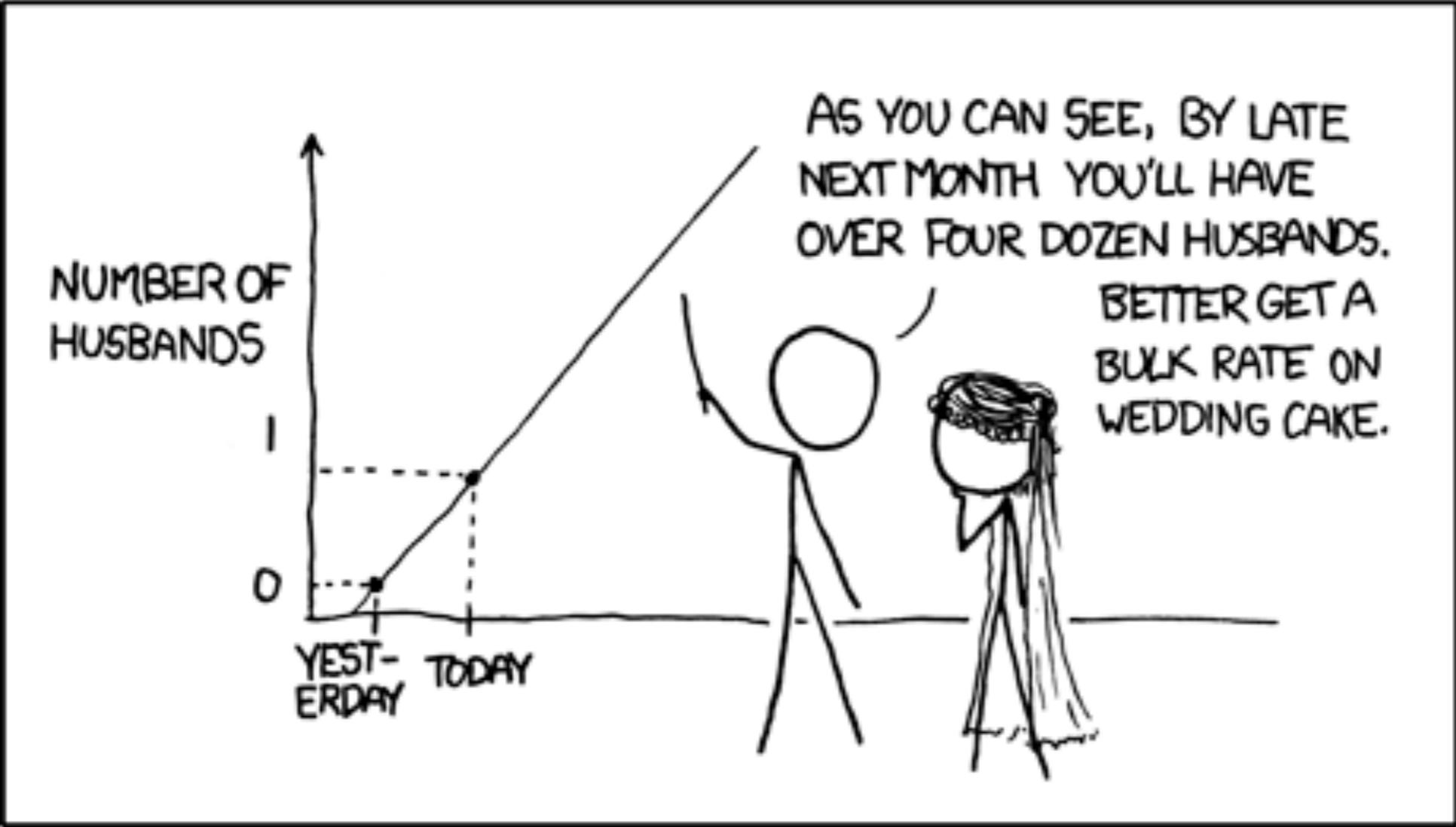
# Ausreißer



- ▶ 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

## MY HOBBY: EXTRAPOLATING



# Nominalskalierte Prädiktoren

Zweistufig nominal:

Geschlecht	is_female
männlich	0
weiblich	1

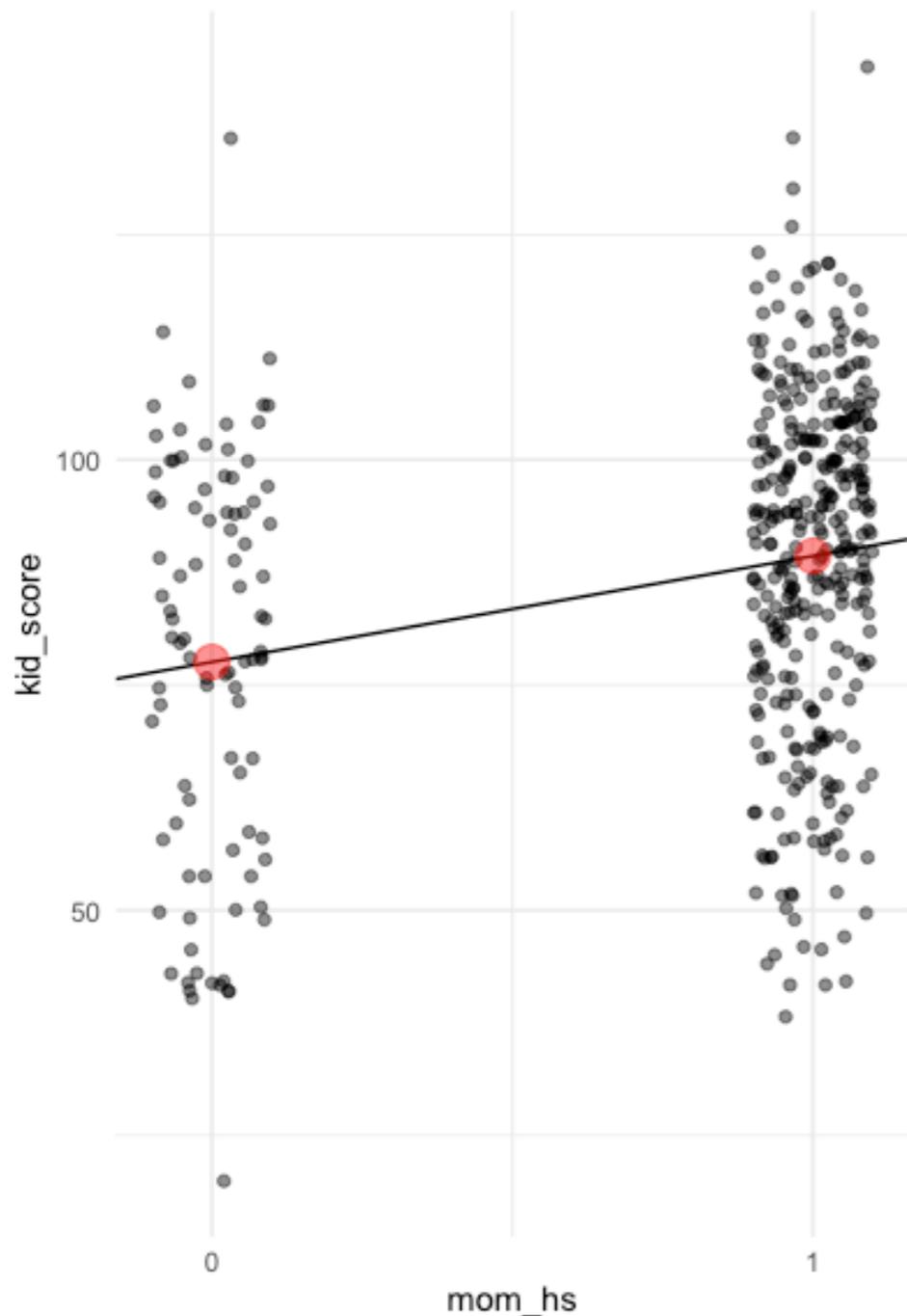
Mehrstufig nominal:

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.
- ▶ Praktischerweise wird die Umwandlung in Dummy-Variablen zumeist von der Software automatisch übernommen.

# Einfache Regression mit einem nominalen, zweistufigen Prädiktor

# IQ von Kindern, binärer Prädiktor (Modell m1)

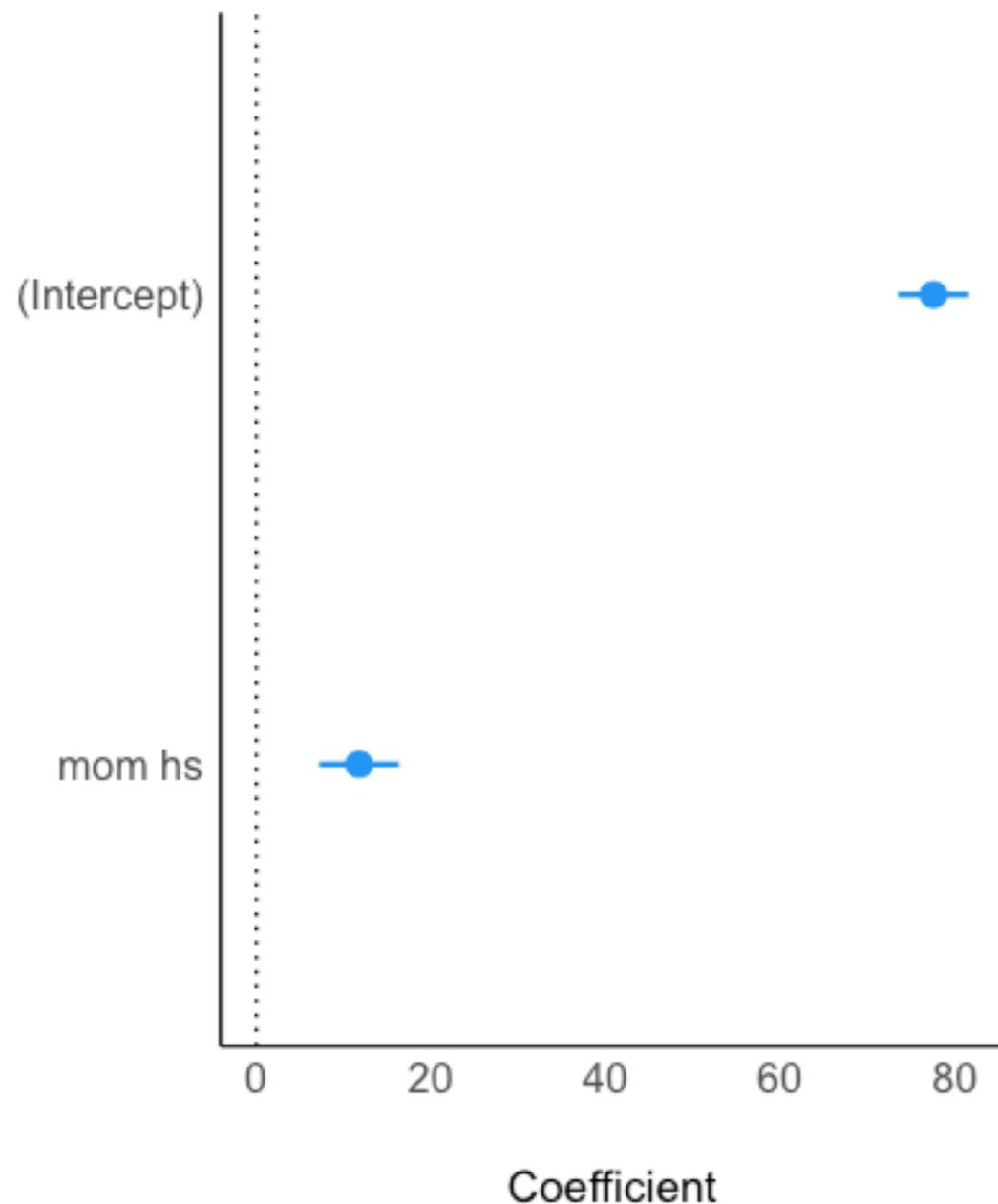


Rote Punkte symbolisieren die Mittelwerte der Gruppen

- ▶ *Forschungsfrage*: Unterscheidet sich der mittlere IQ-Wert (`kid_score`) von Kindern in Abhängigkeit davon, ob ihre jeweilige Mutter über einen Schulabschluss (`mom_hs`) verfügt? (*ceteris paribus*)
- ▶ `m1: kid_score = 78 + 12*mom_hs + error`
- ▶ Der Achsenschnitt (intercept,  $b_0$ ) ist der mittlere (bzw. vorhergesagte) IQ-Wert von Kindern, deren Mütter über *keinen* Schulabschluss verfügen:
- ▶ `kid_score = 78 + 0*12 + error`
- ▶ Das Regressionsgewicht (slope,  $b_1$ ) ist der Unterschied im mittleren IQ-Wert von Kindern *von Mütter mit Schulabschluss* (`mom_hs`; im Vergleich zum IQ-Wert von Kindern mit Mütter ohne Schulabschluss).
- ▶ Dieser Unterschied entspricht der Steigung der Regressionsgerade:
- ▶ `78 + 1*12 + error = 90 + error`

$$\hat{m}_1 : \text{kid score} \sim \text{mom hs}$$

# 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)
- ▶ Dieser Schätzwert ist identisch zum Mittelwertsunterschied zwischen den beiden Gruppen (mit vs. 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: Mit 95%-Wahrscheinlichkeit liegt der gesuchte Wert in der Population im angegebenen Intervall\*, laut Modell
  - ▶ Je größer dieser Wert, desto ungenauer die Schätzung und desto größer der Schätzbereich

Parameter	Coefficient	CI	CI_low	CI_high
(Intercept)	77.55	0.95	73.50	81.59
mom_hs	11.77	0.95	7.21	16.34

\* Bayesianische Interpretation

**Eine metrische plus eine nominale UV**

# Vorhersage mit zwei Prädiktoren

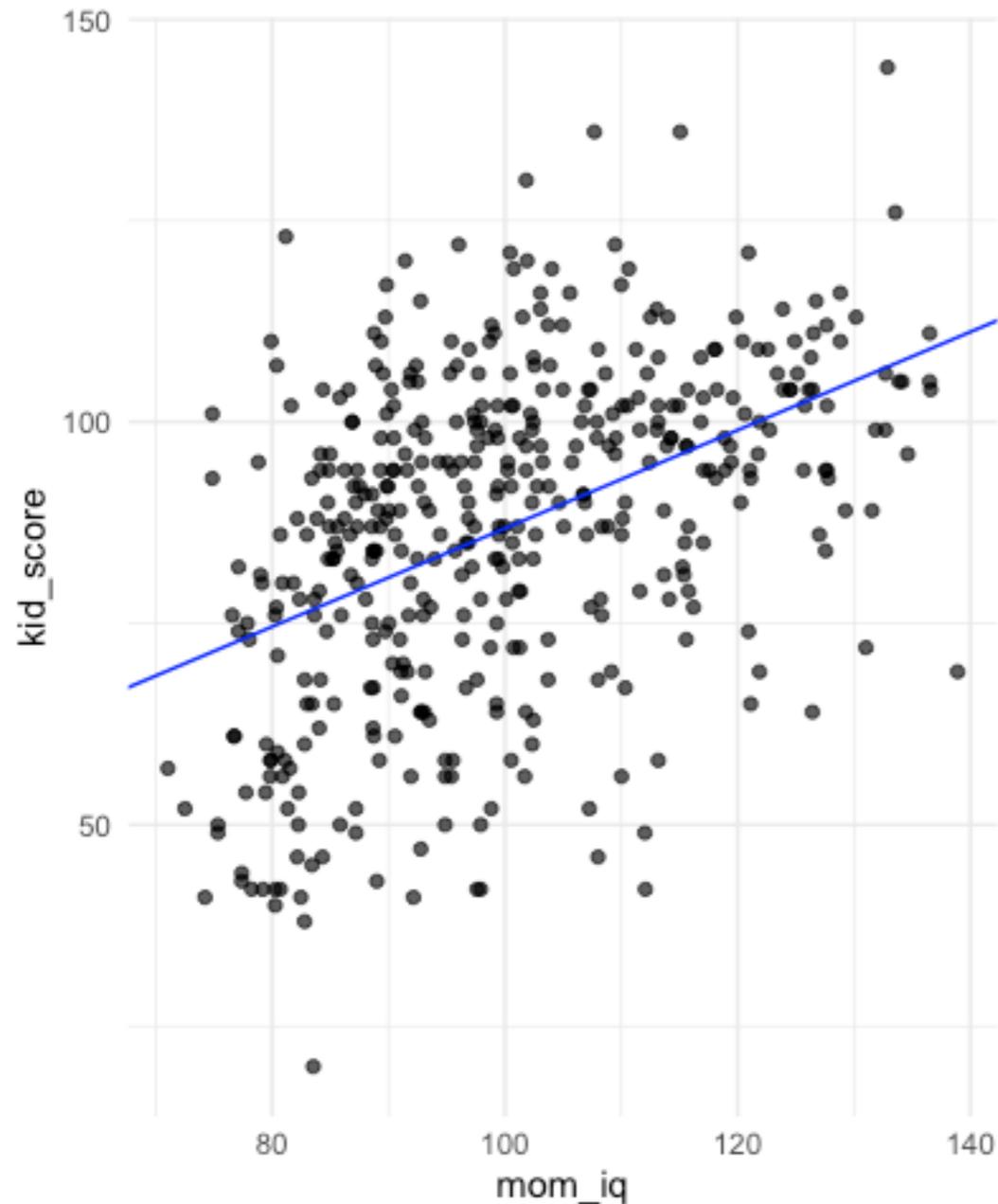
▶ `m3: kidiq ~ mom_iq + mom_hs`

- ▶ 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

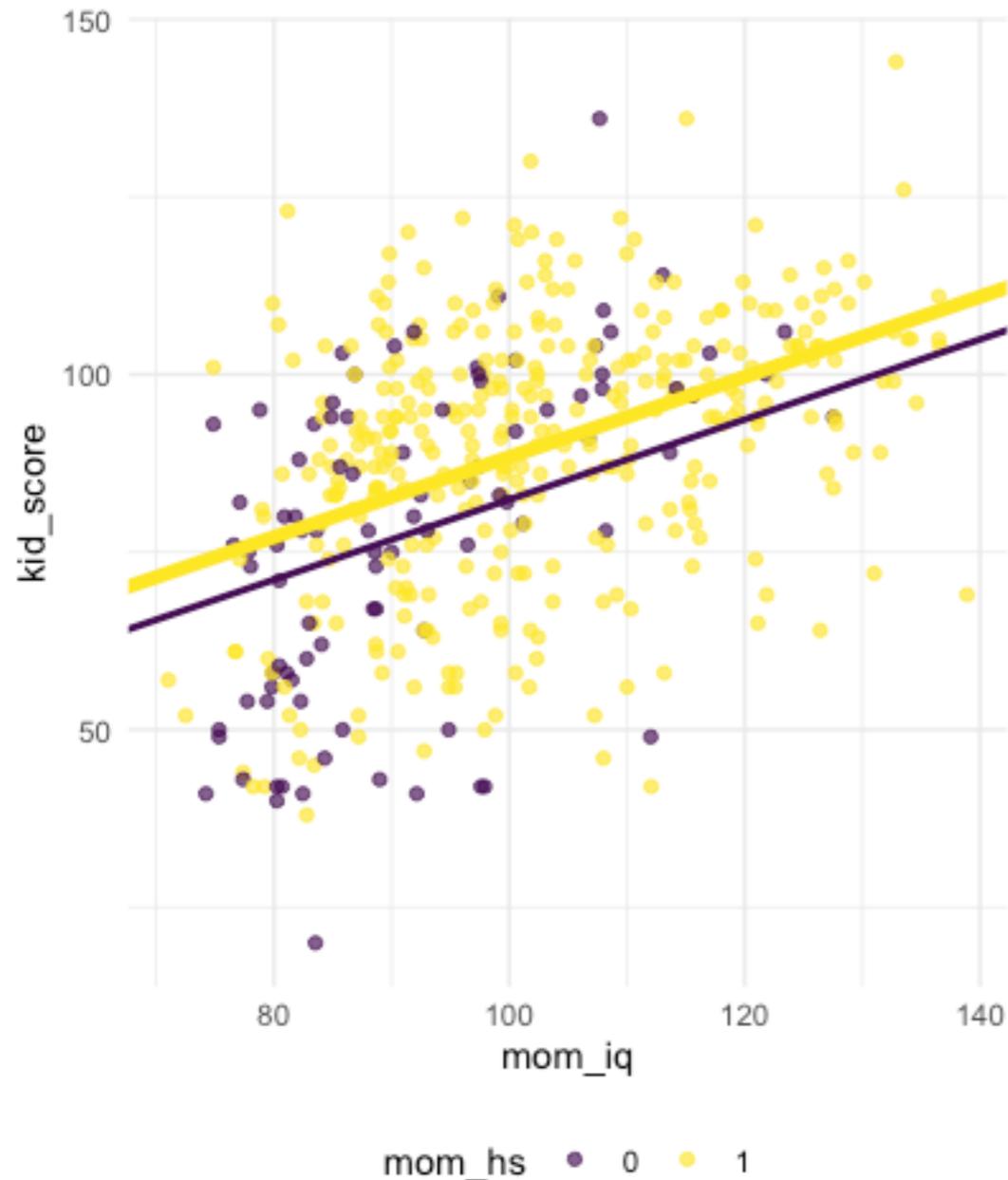
# Ein metrischer Prädiktor

$$\text{kid\_score} = 26 + 0.6 * \text{mom\_iq} + \text{error}$$



- ▶ 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.
- ▶ Dieser Unterschied („statistischer Effekt“) spiegelt sich in der Steigung der Regressionsgeraden wider ( $b_1 = 0.6$ ).
- ▶ Der Achsenabschnitt hilft uns nicht weiter, da es keine Menschen mit einem IQ von 0 gibt.

# Metrischer plus binärer Prädiktor – Interpretation



- ▶ **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.

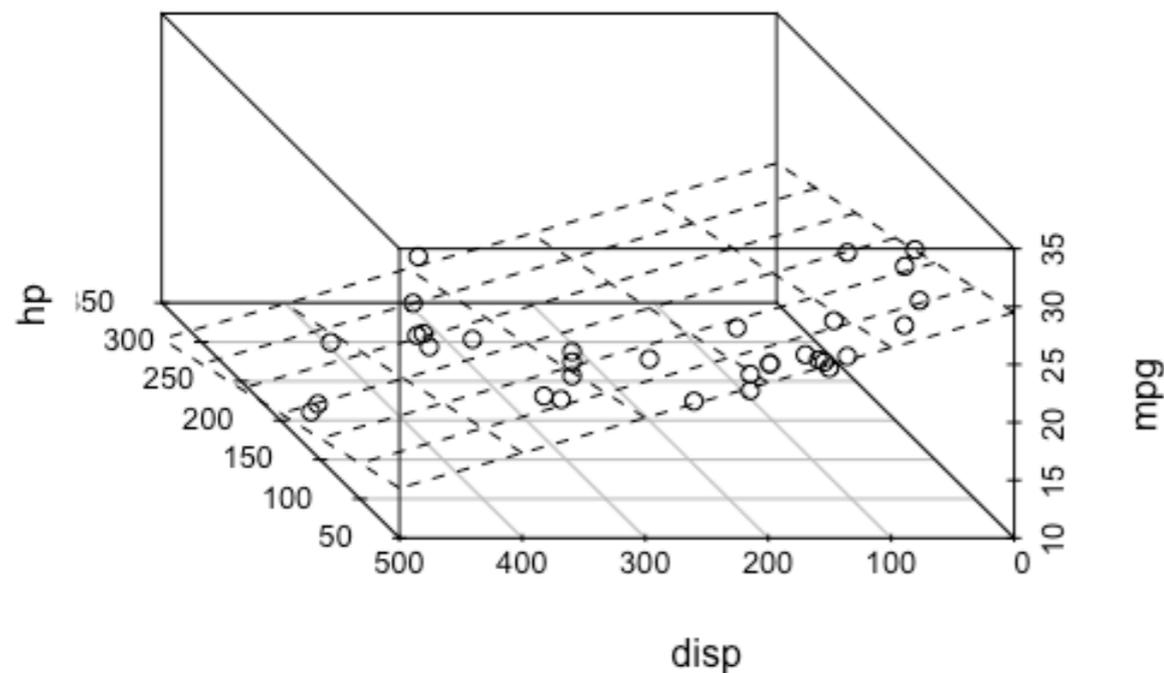
Parameter	Coefficient	CI	CI_low
(Intercept)	25.73	0.95	14.18
mom_iq	0.56	0.95	0.44
mom_hs	5.95	0.95	1.60

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

# Multiple Regression

# Regression mit mehreren Prädiktoren

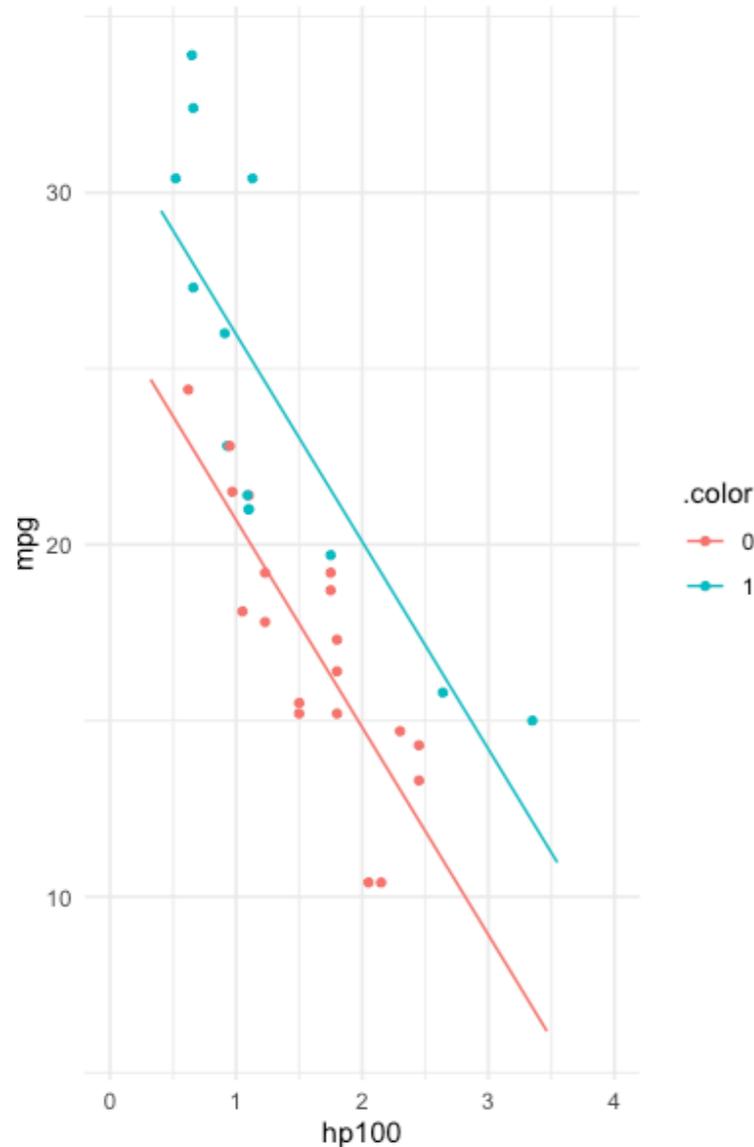
$$y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \dots + b_px_{ip} + e_i$$



- ▶ [3D-Diagramm](#) interaktiv

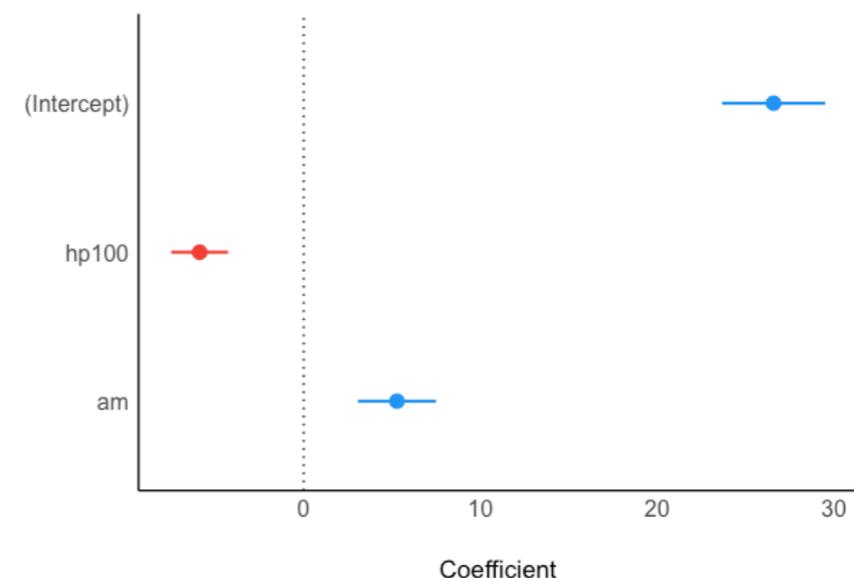
- ▶ Bei einer **multiplen Regressionsanalyse** wird versucht, den Wert einer Kriteriumsvariable (für eine Beobachtung) durch **mehrere (p) Prädiktorvariablen** vorherzusagen. Ansonsten ähnelt die multiple Regression der einfachen Regression.
- ▶ Multiple Regressionsanalysen sind dann sinnvoll, wenn die Vorhersagegüte steigt durch Hinzunahme weiterer Prädiktoren.
- ▶ Jeder Prädiktor hat dabei ein eigenes **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, sofern die Prädiktoren mit der UV korreliert sind.
- ▶ „Bereinigt“ bedeutet, dass das Regressionsgewicht  $b_x$  den Y-Wert für  $x=1$  angibt, wobei die übrigen Prädiktoren mit Wert 0 angenommen werden.
- ▶ Bei zwei Prädiktorvariablen kann man sich das Modell als Ebene im Raum vorstellen (anstelle einer Geraden): stellen Sie sich ein Blatt Papier vor, das „mittig“ durch einen Bienenschwarm gelegt wird.

# Modell mit zwei Prädiktoren (metrisch, binär)



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

- ▶ Wir sagen Spritverbrauch (mpg) vorher auf Basis zweier Prädiktoren: hp (metrisch) und am (binär).
- ▶ Jeder (der drei) Modellkoeffizienten ( $b_0$ ,  $b_1$ ,  $b_2$ ) gibt den Wert für mpg (Y) an, unter der Bedingung, dass die anderen Modellkoeffizienten Null sind.
- ▶ Die Geraden sind parallel.
- ▶  $b_{\text{hp}}$ : Der mpg-Wert unterscheidet um  $b_{\text{hp}}$  mpg-Einheiten, bei Autos, die sich um eine hp-Einheit unterscheiden.
- ▶  $b_{\text{am}}$ : Der mpg-Wert unterscheidet um  $b_{\text{am}}$  mpg-Einheiten, bei Autos, die sich um eine am-Einheit unterscheiden.
- ▶  $b_0$ : Der mpg-Wert von Autos mit  $\text{hp} = 0$  und  $\text{am} = 0$ .

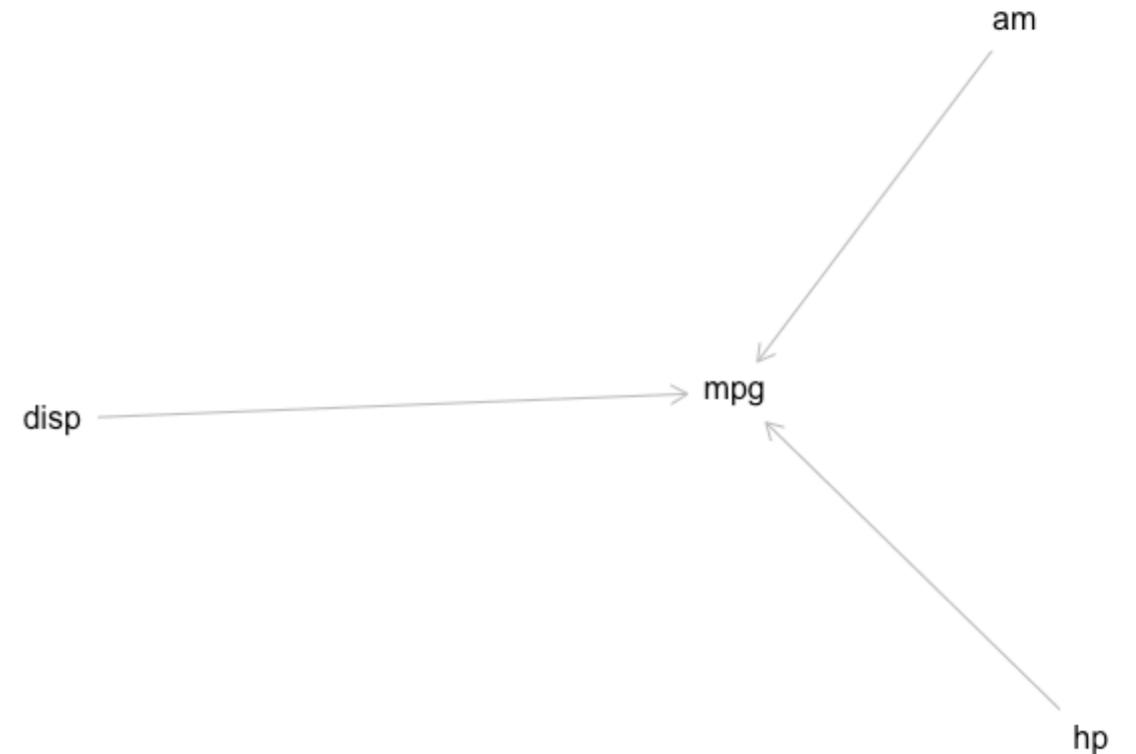


# Welches Modell liefert wohl bessere Vorhersagen?

A



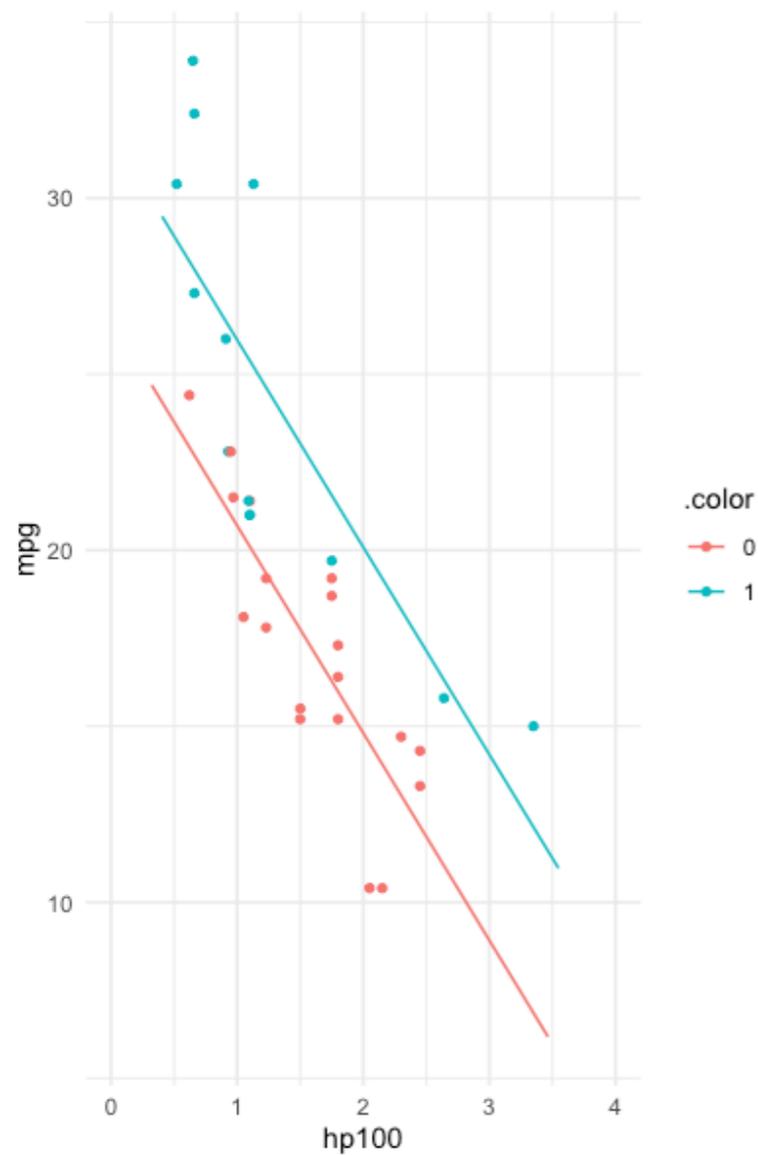
B



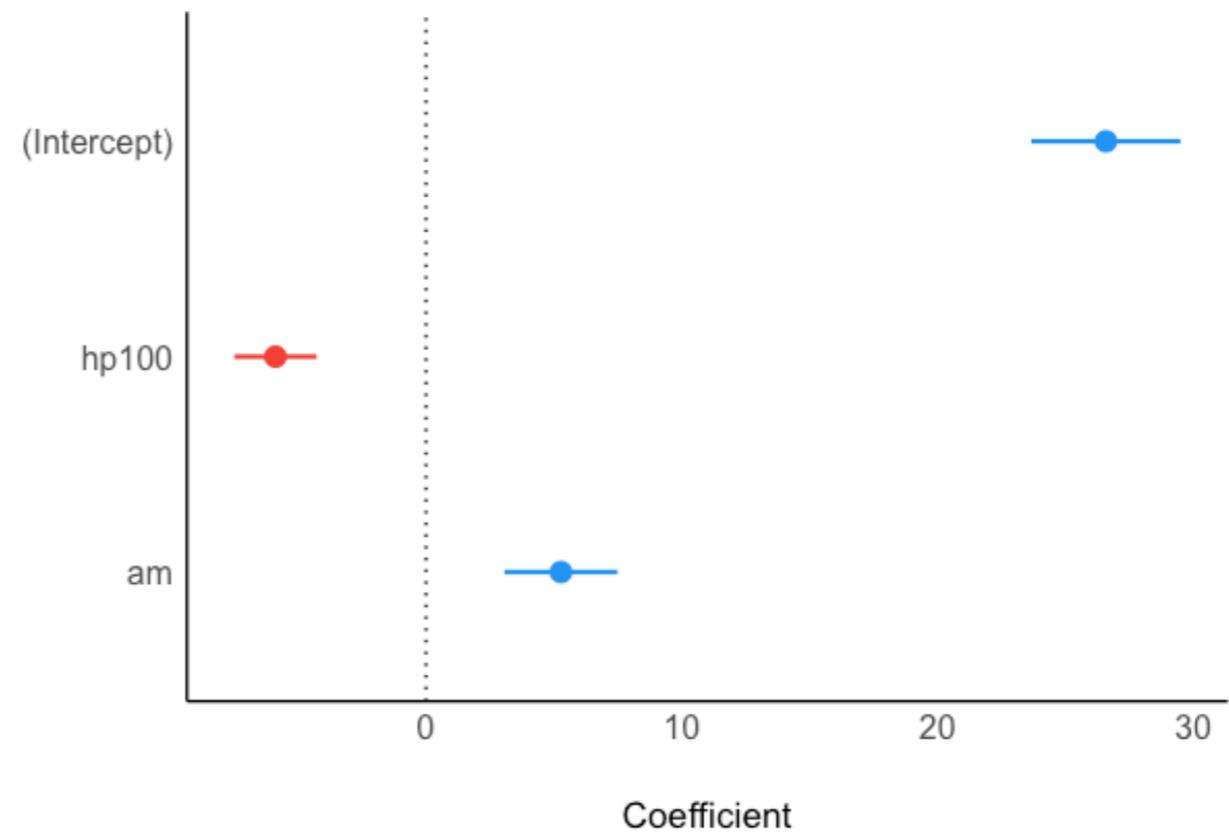
Sofern ein Modell mehr relevante Prädiktoren beinhaltet, wird es bessere Vorhersagen machen (als ein Modell mit weniger relevanten Variablen), vorausgesetzt die die zusätzlichen Prädiktoren sind mit der AV korreliert (unter sonst gleichen Umständen, d.h. ceteris paribus).

# Ausgabe eines Regressionsmodells

$$\text{mpg} = b_0 + b_1 \text{hp}_{100} + b_2 \text{am} + e$$

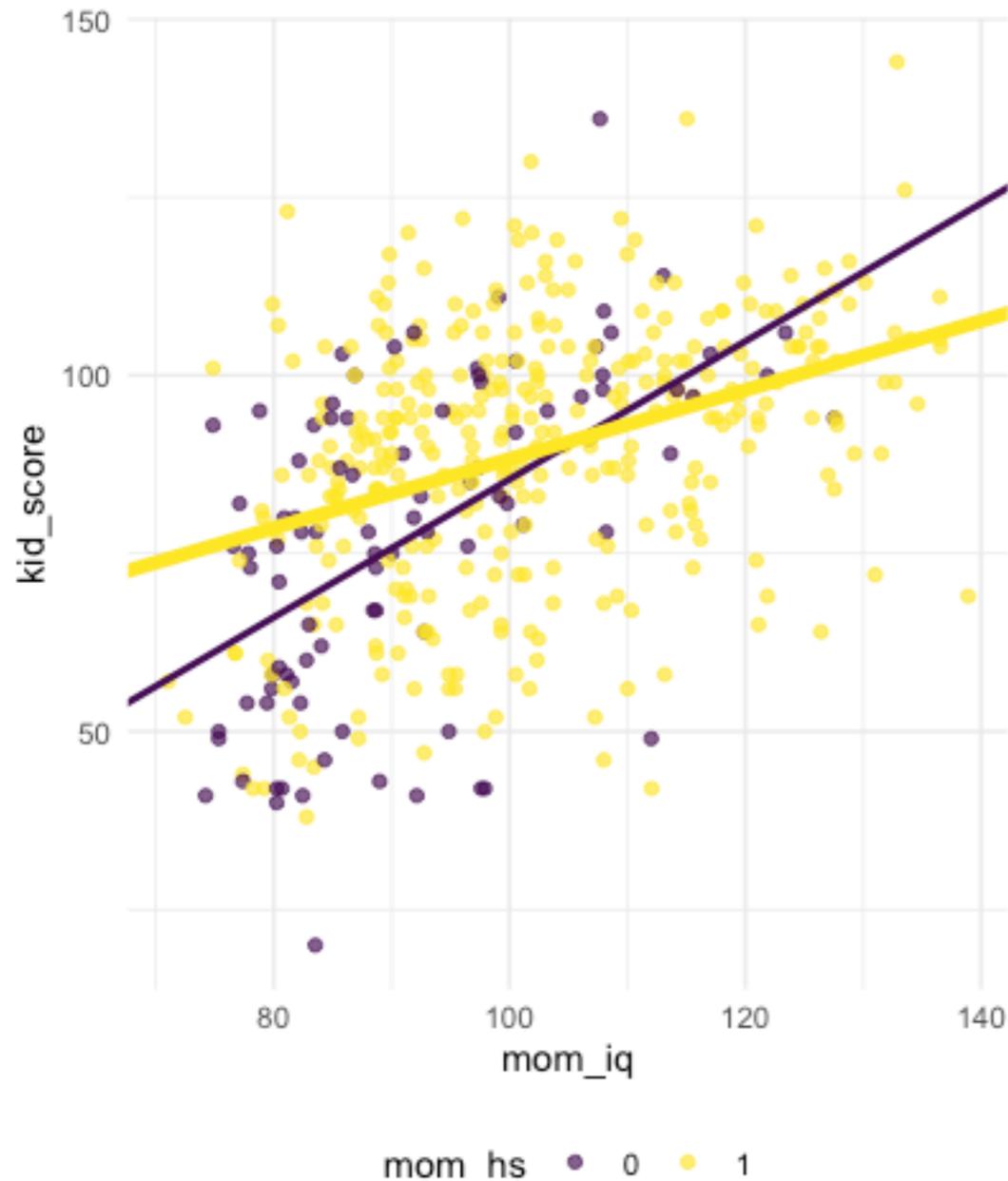


Parameter	Coefficient	CI	CI_low
(Intercept)	26.58	0.95	23.67
hp100	-5.89	0.95	-7.50
am	5.28	0.95	3.07



# Interaktion

# Interaktion von Mutter-IQ und Schulabschluss



- ▶ In Modell m3 hatten wir die Regressionsgeraden gezwungen, **parallel** zu sein.
- ▶ Betrachtet man dieses Streudiagramm, so sieht man, dass *nicht-parallel* 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*.

Parameter	Coefficient
(Intercept)	-11.48
mom_iq	0.97
mom_hs	51.27
mom_iq:mom_hs	-0.48

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

# 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 ( $\text{mom\_hs} = 0$ ).
- ▶ **Interaktion:** Der Unterschied in den Steigungen der Regressionsgeraden, also der Unterschied des Koeffizienten für  $\text{mom\_iq}$  zwischen Müttern mit bzw. ohne Schulabschluss. Die Interaktion wird als Produkt beider Prädiktoren dargestellt.
- ▶ Der Wert von  $\text{kid\_score}$  ist die Summe obiger Terme.

$\text{mom\_hs}=0$ :

$$\text{kid\_score} = -11 + 51*0 + 1.1*\text{mom\_iq} + 0.5*0*\text{mom\_iq} = -11 + 1.1*\text{mom\_iq}$$

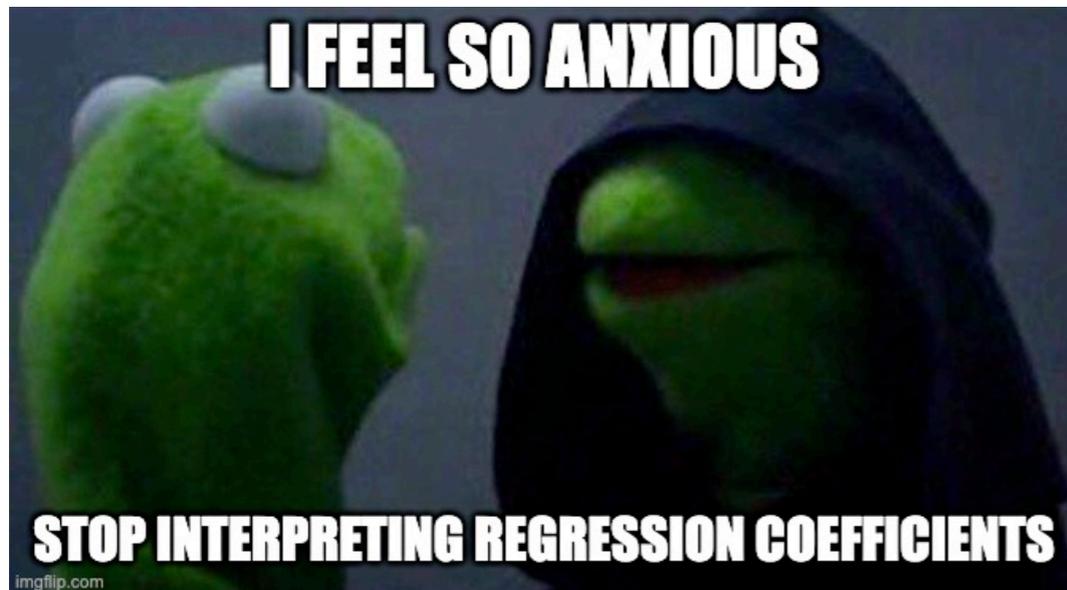
$\text{mom\_hs}=1$ :

$$\text{kid\_score} = -11 + 51*1 + 1.1*\text{mom\_iq} + 0.5*1*\text{mom\_iq} = 40 + 0.6*\text{mom\_iq}$$

allgemein:

$$\text{kid\_score} = b_0 + b_1*\text{hs} + b_2*\text{mom\_iq} + b_3*\text{mom\_hs}*\text{mom\_iq}$$

# Oft ist eine normale Regression schwer zu interpretieren

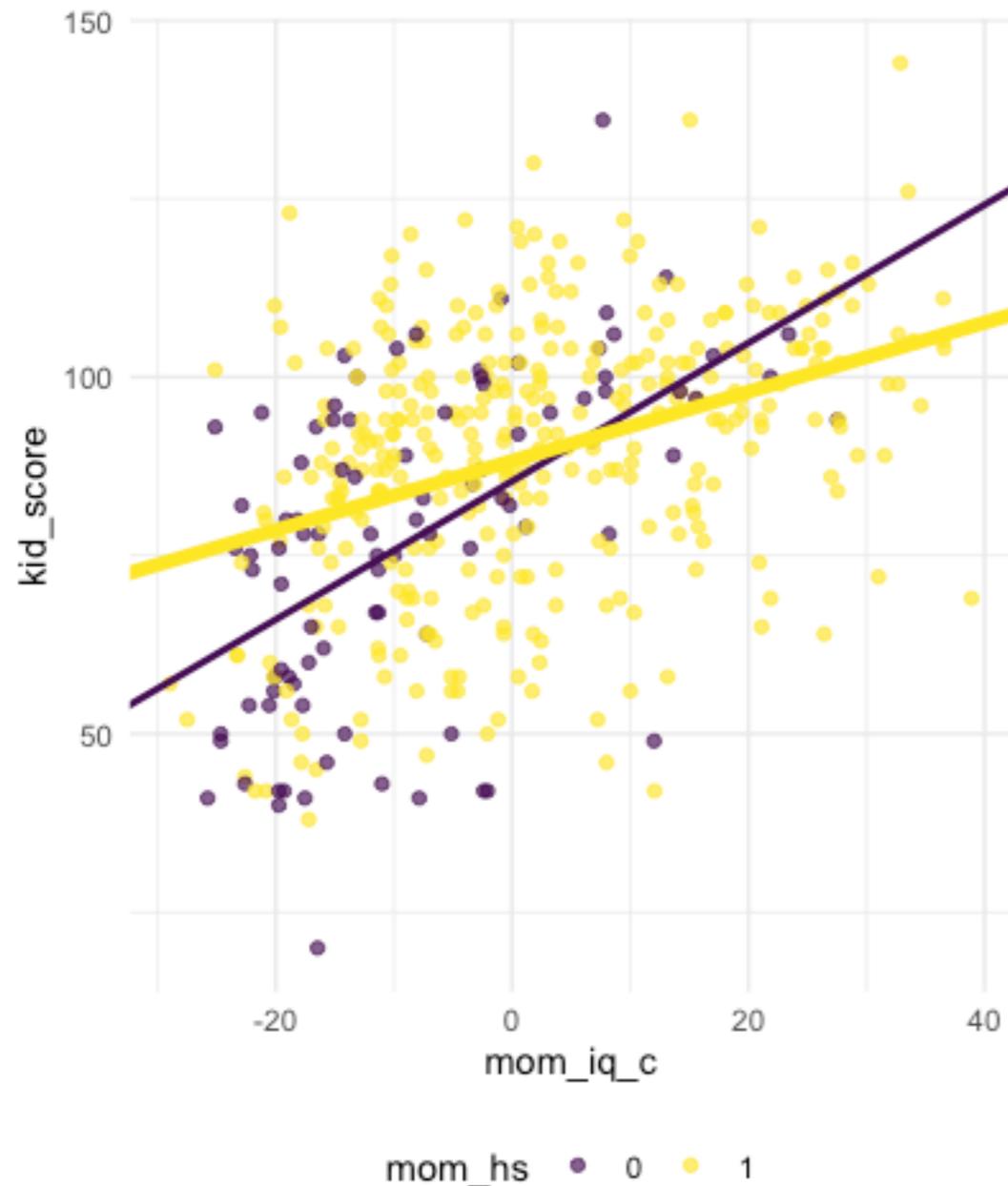


[Quelle](#)

- ▶ 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 **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.

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

# Modell mit zentrierten Prädiktoren interpretieren

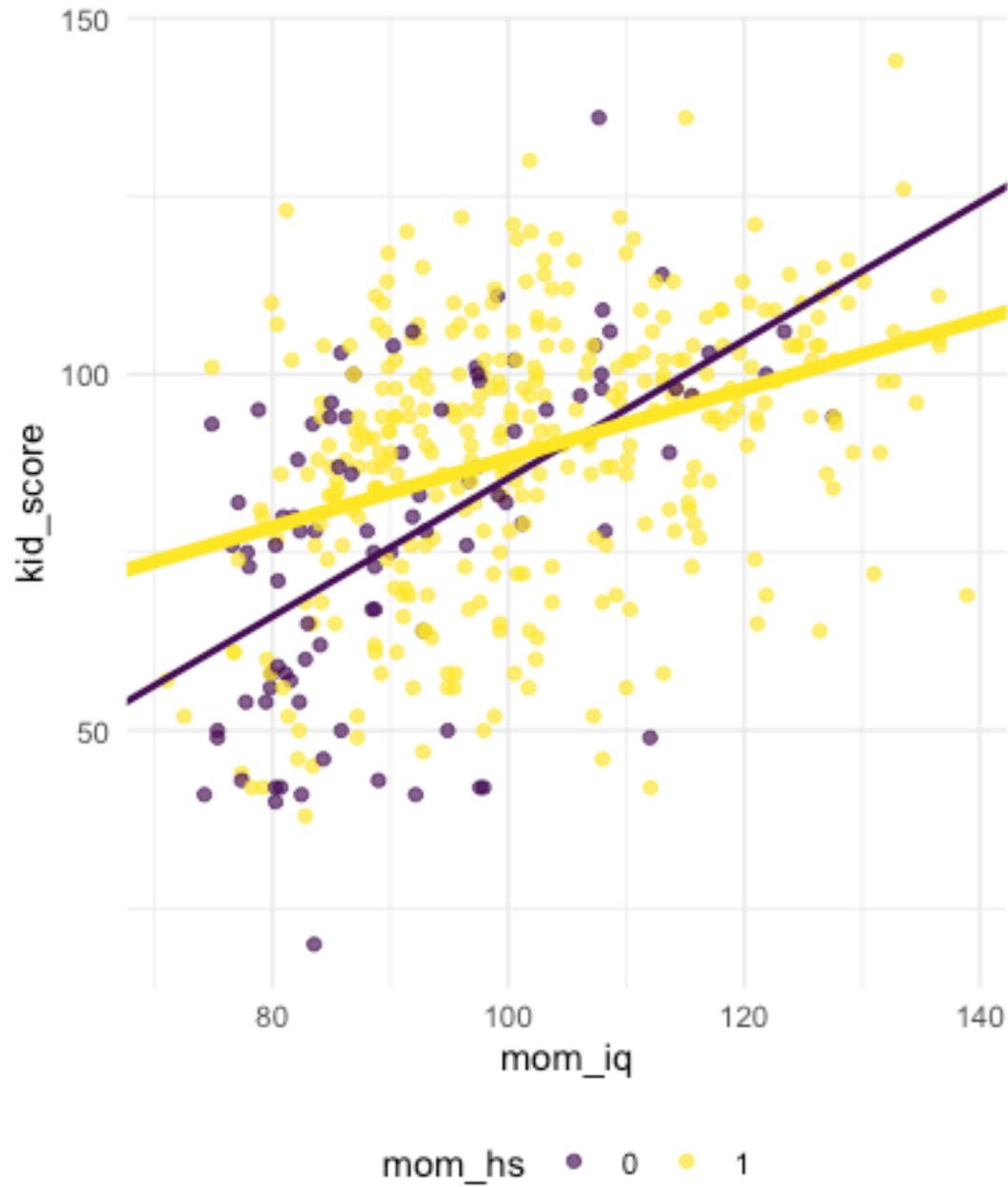


- ▶ Wir gehen im Folgenden von *zentrierten* Prädiktoren aus:
- ▶ Der Achsenabschnitt (`Intercept`) gibt den geschätzten IQ des Kindes an, wenn man eine Mutter mittlerer Intelligenz und ohne Schulabschluss betrachtet.
- ▶ `mom_hs` gibt den Unterschied im geschätzten IQ des Kindes an, wenn man Mütter mittlerer Intelligenz mit bzw. ohne Schulabschluss vergleicht.
- ▶ `mom_iq_c` gibt den Unterschied im geschätzten IQ des Kindes an, wenn man Mütter *ohne Schulabschluss 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 Gruppen von `mom_hs`. Der Doppelpunkt wird in R verwendet, um die Interaktion anzuzeigen. Das hat nichts mit Division zu tun, es sieht nur zufällig so aus.

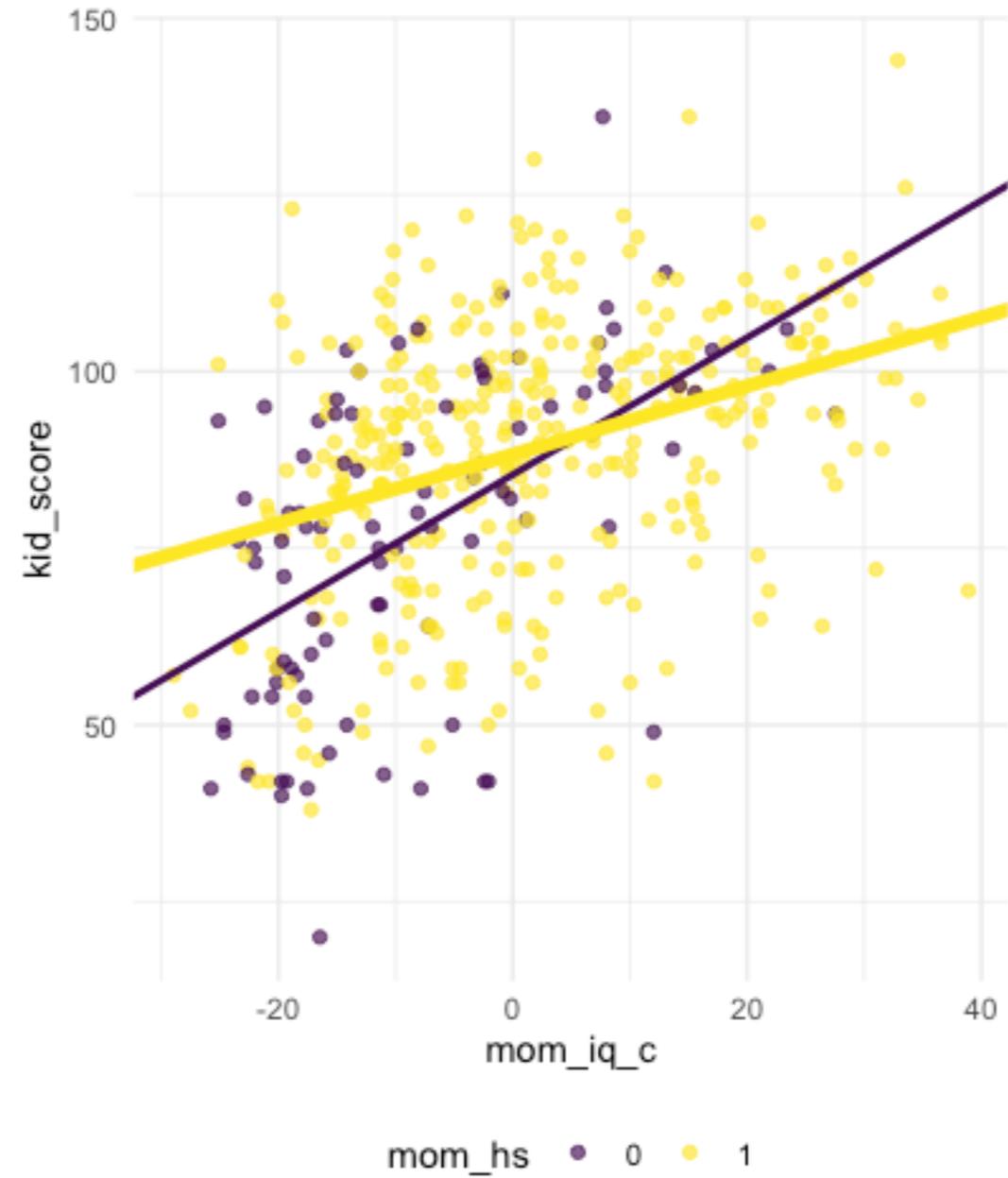
```
m5: kid_score ~ mom_iq_c + mom_iq_c + mom_iq_c:mom_hs
```

# Zentrieren ändert nichts an den Vorhersagen

m4: unzentrierte Prädiktoren

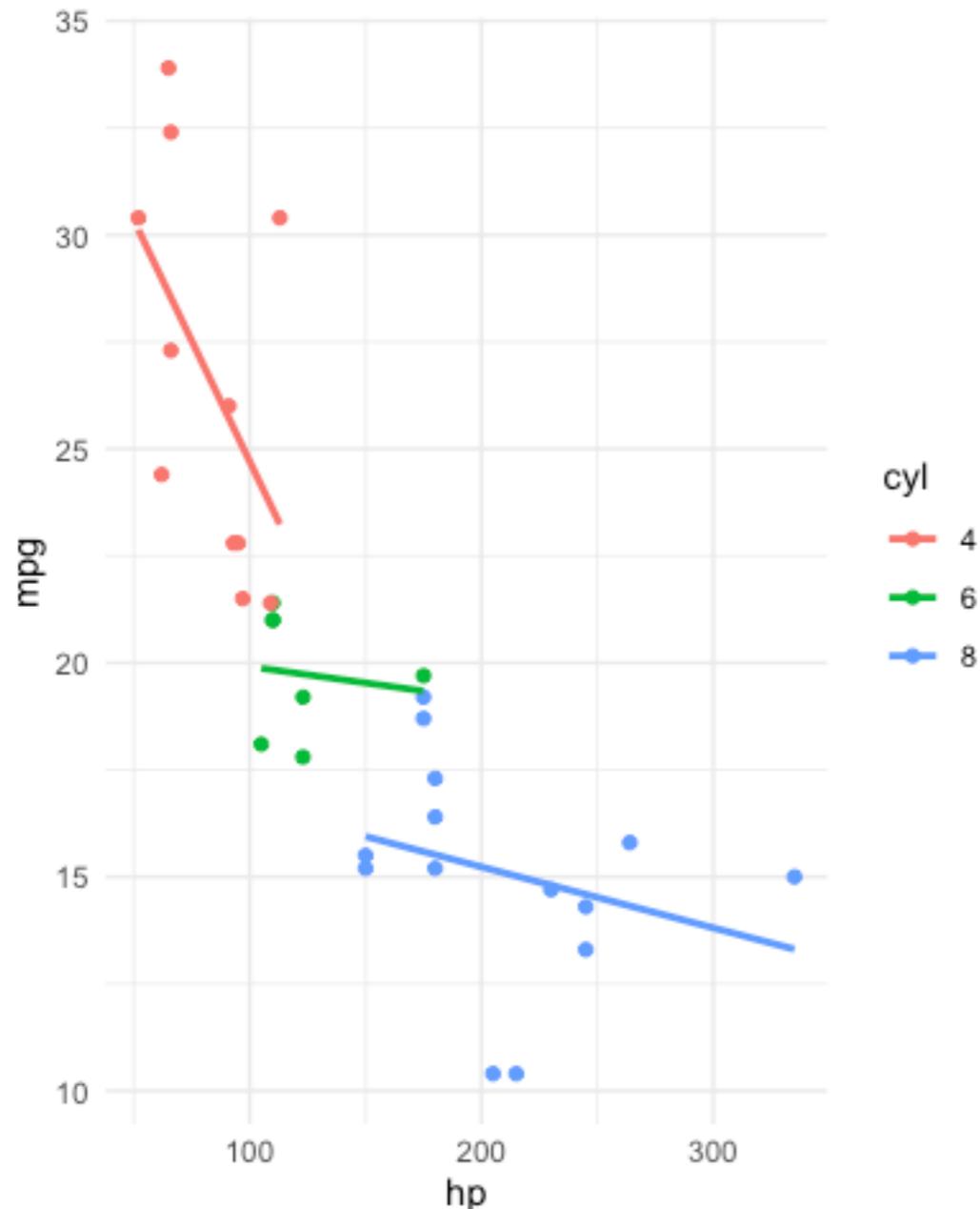


m5: zentrierte Prädiktoren



Nur die Prädiktoren sind zentriert: Ein Wert von Null entspricht dem Mittelwert in den untransformierten Daten.

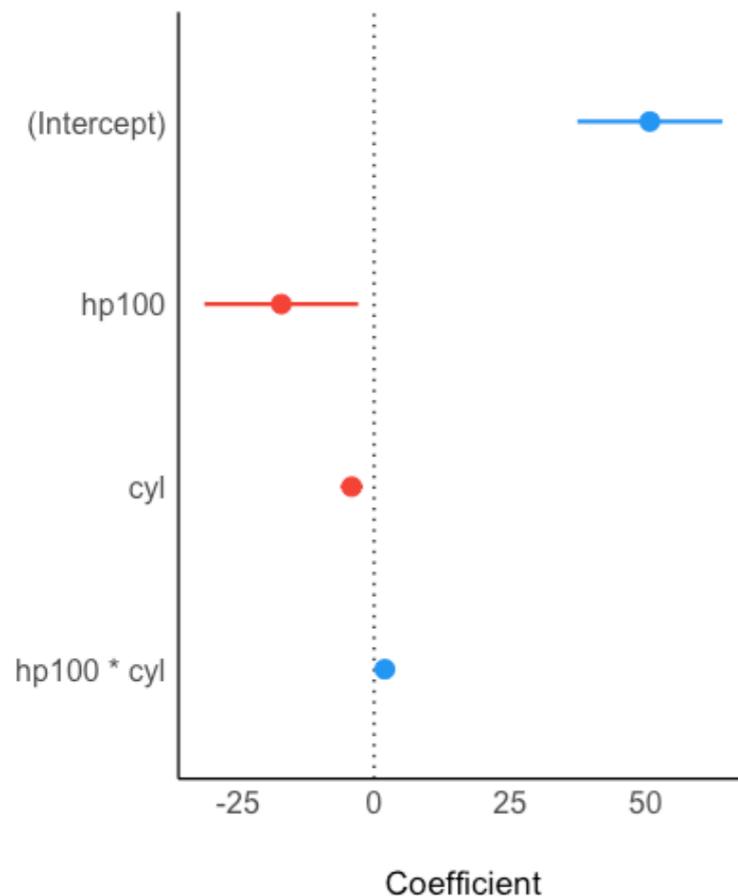
# Interaktionseffekt bei mtcars



- ▶ Wirkt sich vielleicht die PS-Zahl **unterschiedlich** aus (auf den Spritverbrauch, Y) je nach Anzahl der Zylinder des Autos?
- ▶ Wir fragen also noch einer **Interaktion** der beiden Prädiktoren.
- ▶ Hängt die Steigung einer Regressionsgeraden ab von der Ausprägung eines anderen Prädiktors, so liegt ein Interaktionseffekt (synonym: Wechselwirkung, Moderation) vor.
  - ▶ Mit „abhängen“ ist gemeint, dass die Veränderung in Y nicht gleich ist für alle Werte des Prädiktors X1, sondern sich je nach Wert eines anderen Prädiktors X2 unterscheidet.
- ▶ 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).

\* Wir machen hier keine Kausalaussagen, sondern nur Vorhersagen, die nur auf statistischen Abhängigkeiten (Korrelationen) aufbauen.

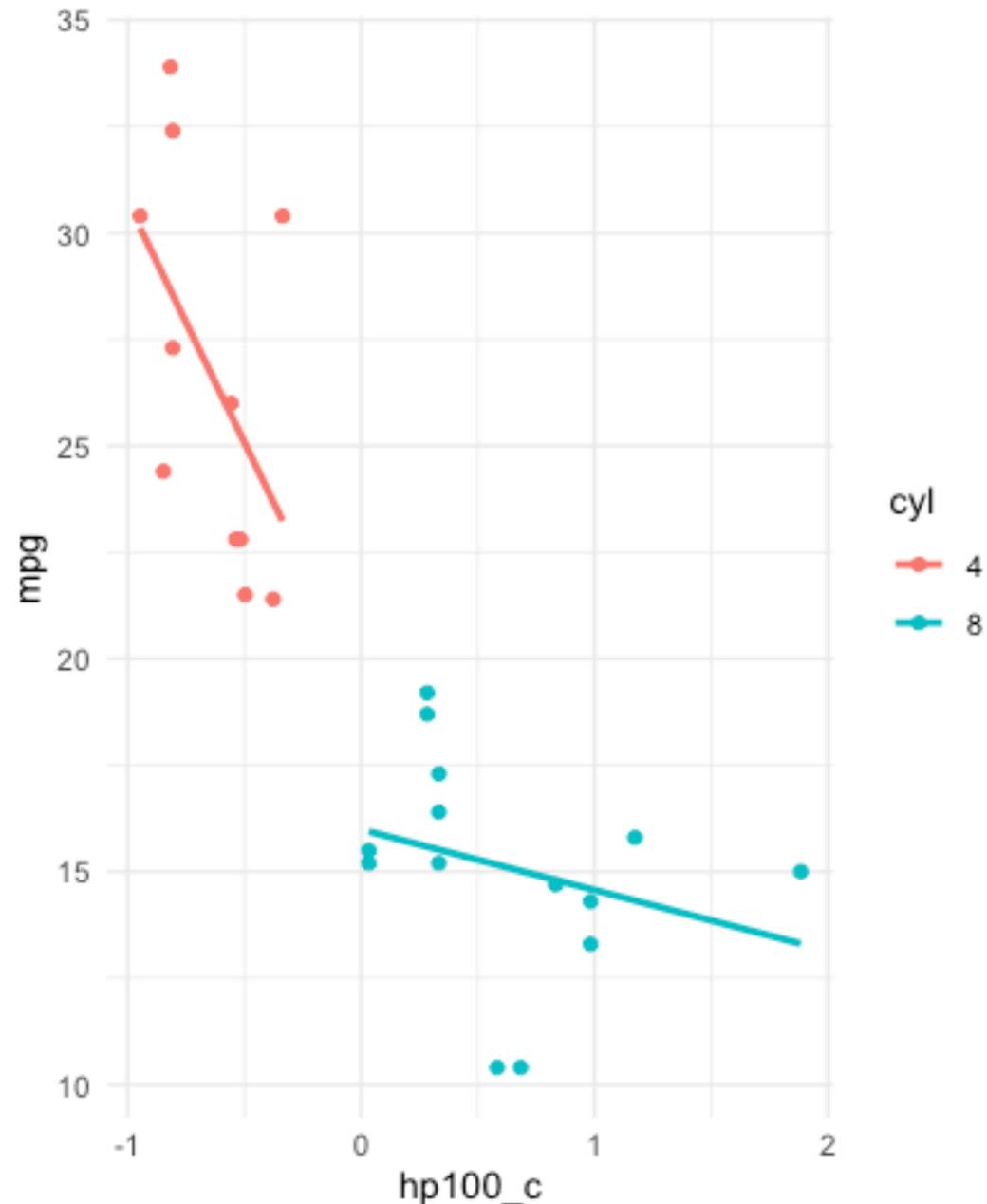
# Koeffizienten bei Interaktionsmodell für mtcars



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, Y) zweier Autos an, die sich um 100 PS (eine Einheit von hp100) unterscheiden
  - ▶ unter der Annahme, dass die übrigen Prädiktoren gleich Null sind.
  - ▶ entspricht der Steigung der Regressionsgeraden.
- ▶ Der Parameter **cyl**
  - ▶ gibt den Unterschied im Verbrauch (Y) zweier 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.
- ▶ Der **Achsenabschnitt**
  - ▶ gibt den mpg-Wert an für Autos mit 0 PS und 0 Zylindern.
- ▶ Der **Gesamtwert** (der Vorhersage) von mpg ist die **Summe** der obigen Terme.

# mtcars-Interaktionsmodell vereinfacht



- ▶ Der Einfachheit halber betrachten wir hier nur Autos mit 4 oder mit 8 Zylindern (also nur 2 Gruppen, nicht 3).
- ▶ **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

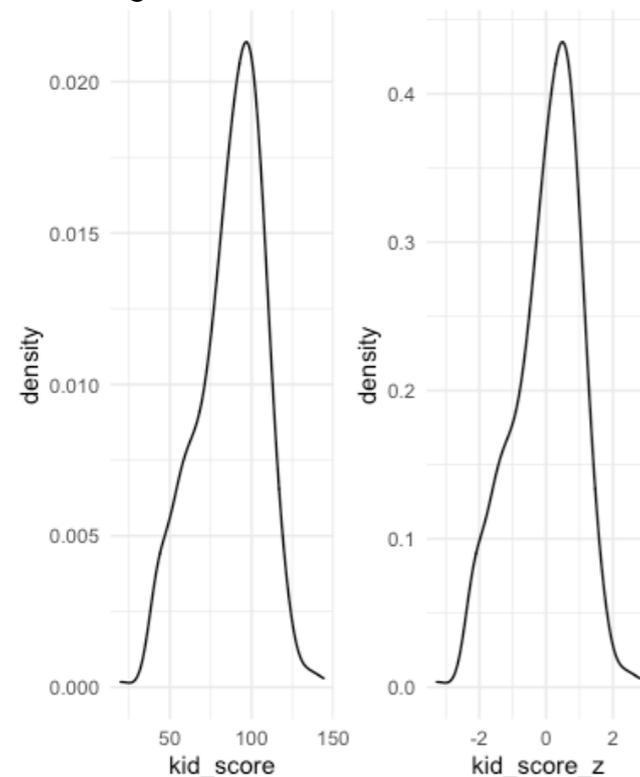
# Woher weiß man, welcher Prädiktor am wichtigsten ist?

- ▶ 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 "unterdurchschnittlich" 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

Verteilung der z-transformierten 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  $\text{kid\_score\_z} = 0$  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, da ihre Streuung gleich ist.
- ▶ Man sieht, dass die Intelligenz der Mutter deutlich wichtiger ist als das Alter der Mutter (im Hinblick auf die Vorhersage bzw. den Zusammenhang mit der AV).
- ▶ Der Wertebereich der Koeffizienten wird durch die z-Transformation homogenisiert.
- ▶ Eine z-Standardisierung der Prädiktoren ist daher **oft nützlich**.

m6:  $\text{kid\_score} \sim \text{mom\_iq\_z} + \text{mom\_age\_z}$

# 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.
- ▶ Die in diesem Skript angewandte R-Syntax findet sich in [diesem Ordner](#).