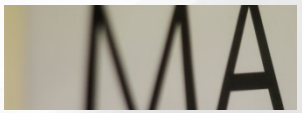
INSTITUT FÜR DEUTSCHE SPRACHE

Sandra Hansen-Morath
Sascha Wolfer



STATISTIK MIT R

Regression

Mitglied der Leibniz-Gemeinschaft

WIEDERHOLUNG KORRELATION

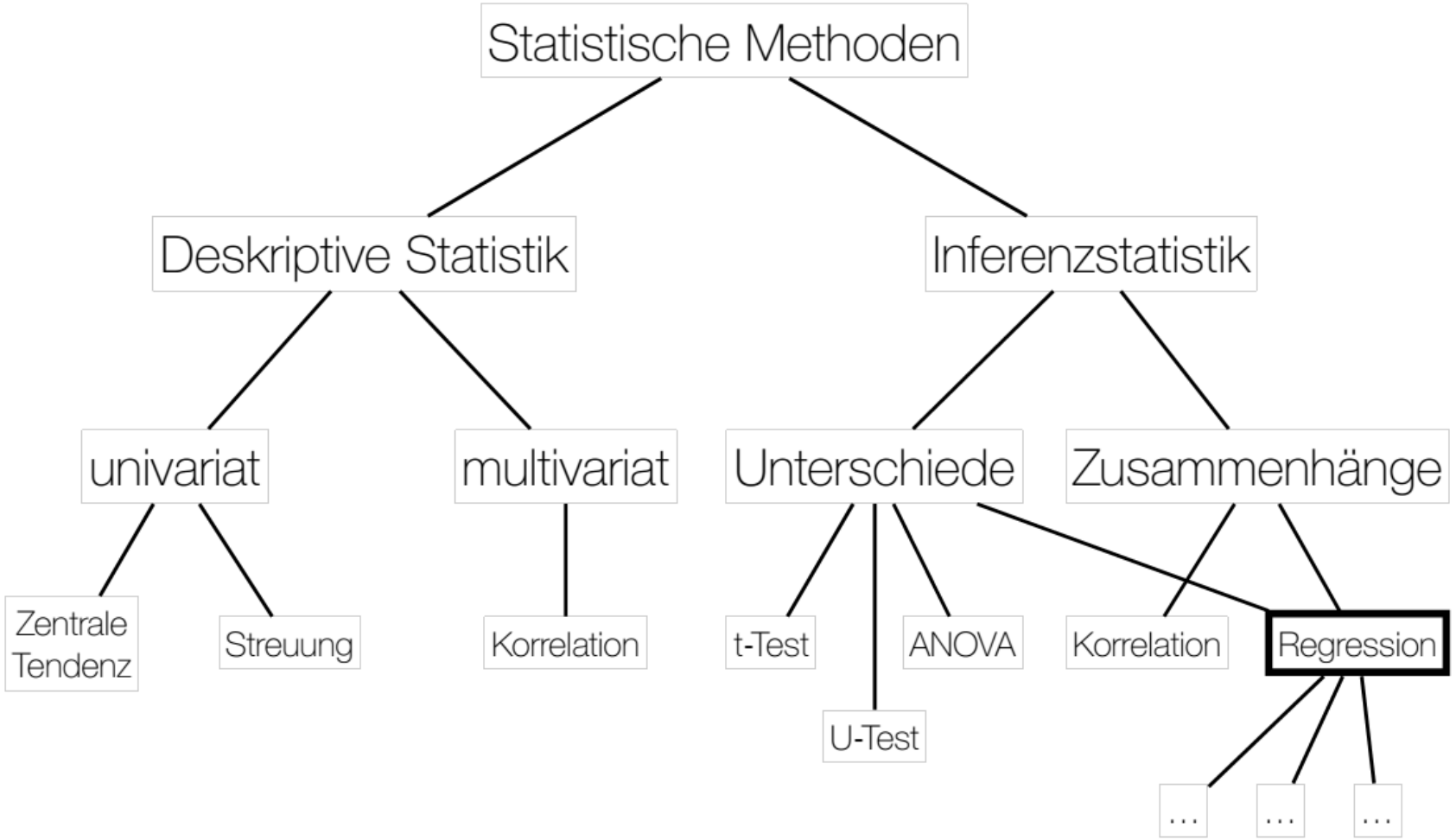
- Feststellen von Richtung und Stärke eines Zusammenhanges zwischen zwei (oder mehr) Variablen.

  → keine Aussage zu kausaler Beziehung oder Ursache-Wirkung!
- Die Stärke der linearen Assoziation der Variablen wird durch einen Korrelationskoeffizienten *r* festgestellt.

  → z.B. *r* nach Pearson bei intervallskalierten Variablen
- Grafisch kann man Zusammenhänge zwischen zwei Variablen gut in einem Scatterplot darstellen
- Korrelationen müssen auch immer auf Signifikanz geprüft werden!
- Korrelationen können von Ausreißerwerten künstlich vergrößert werden!

ÜBUNG ZUR KORRELATION

EINORDNUNG


Statistische Methoden
Deskriptive Statistik
Inferenzstatistik
univariat
multivariat
Unterschiede
Zusammenhänge
Zentrale Tendenz
Streuung
Korrelation
t-Test
ANOVA
U-Test
Korrelation
Regression
...
...
...

EINORDNUNG

- Regressionsmethoden sind grundsätzlich dem Feld der **Inferenzstatistik** zuzuordnen.
- Sie sind daher zuallererst einmal ein Werkzeug zur **gezielten Überprüfung von Hypothesen**.
- Es können auch aus einer Menge möglicher Einflussfaktoren die relevanten gefunden werden ("Modellanpassung").

TERMINOLOGIE

- **Unabhängige Variable**: Eigenschaft, die von ForscherIn manipuliert wird.
- **Abhängige Variable**: Eigenschaft, auf der der Einfluss der Manipulation der unabhängigen Variable gemessen werden soll.
- Im Kontext von Regressionsmodellen spricht man auch von **Kriteriums- und Prädiktorvariablen**.
  - Prädiktoren: Variablen, die zur Vorhersage des Kriteriums herangezogen werden.
  - Kriteriumsvariable: Variable, die vom Modell vorhergesagt werden soll.

TERMINOLOGIE

- Über ein Regressionsmodell finden wir den besten **Fit** (= die beste Modellanpassung) der Prädiktoren an die Kriteriumsvariable.
  - Der Fit wird über die **Estimates** (= Schätzwerte / Effektschätzer) bestimmt.
- Prädiktorvariablen können **binär**, **kategorial** und **intervallskaliert** sein.
- Dasselbe gilt für die Kriteriumsvariable.
  - Wir werden allerdings nur Regressionsmodelle mit binären und intervallskalierten Kriteriumsvariablen kennenlernen.

DAS BESTE MODELL???

Anzahl der anwesenden GPs

Geschlecht der GP

Frequenz des Lexems

Beruf der GP

Wird eine Variable dialektal od. standardnah artikuliert?
JA vs. NEIN

Alter der GP

Geografische Lage des Erhebungsortes

Anzahl Exploratoren

Morphologische Komplexität des Lexems

DAS BESTE MODELL???

Anzahl der anwesenden GPs

Geschlecht der GP

Frequenz des Lexems

Beruf der GP

Wird eine Variable dialektal od. standardnah artikuliert?
JA vs. NEIN

Alter der GP

Geografische Lage des Erhebungsortes

Anzahl Exploratoren

Morphologische Komplexität des Lexems

DAS BESTE MODELL???

Anzahl der anwesenden GPs

Geschlecht der GP

Frequenz des Lexems

Beruf der GP

Wird eine Variable dialektal od. standardnah artikuliert?
JA vs. NEIN

Alter der GP

Geografische Lage des Erhebungsortes

Anzahl Exploratoren

Morphologische Komplexität des Lexems

DAS BESTE MODELL???

Geschlecht der GP

Anzahl der anwesenden GPs

Frequenz des Lexems

Beruf der GP

Wird eine Variable dialektal od. standardnah artikuliert?
JA vs. NEIN

Alter der GP

Alter der GP x Beruf der GP

Geografische Lage des Erhebungsortes

Anzahl Exploratoren

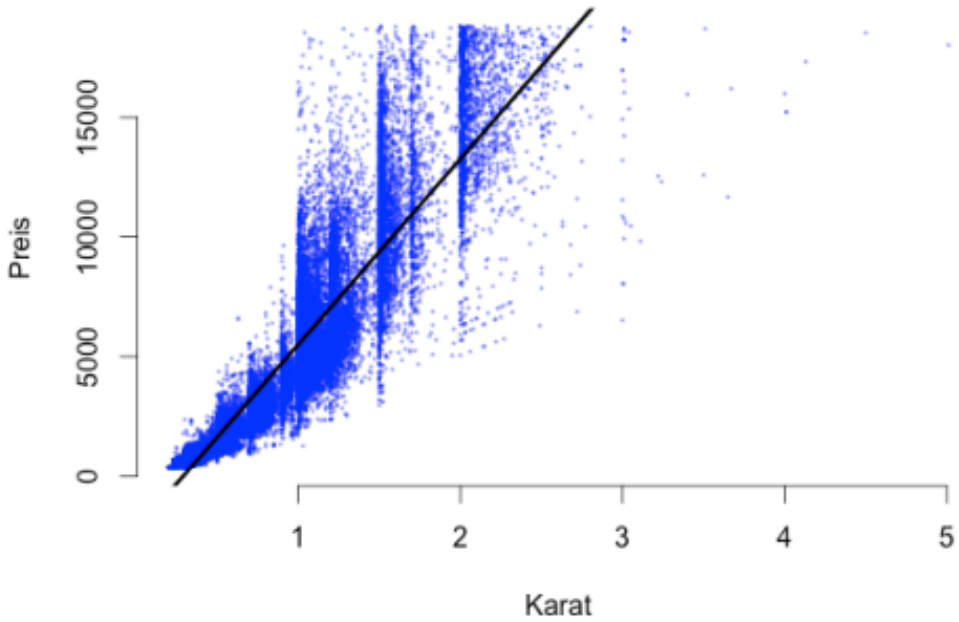
Morphologische Komplexität des Lexems

ZUSAMMENFASSUNG

- Prädiktor(variable) = unabhängige Variable
- Kriterium(svariable) = abhängige Variable
- Fit = Anpassung des Regressionsmodells an die Daten
- Estimates/Effektschätzer = Parameter des Regressionsmodells

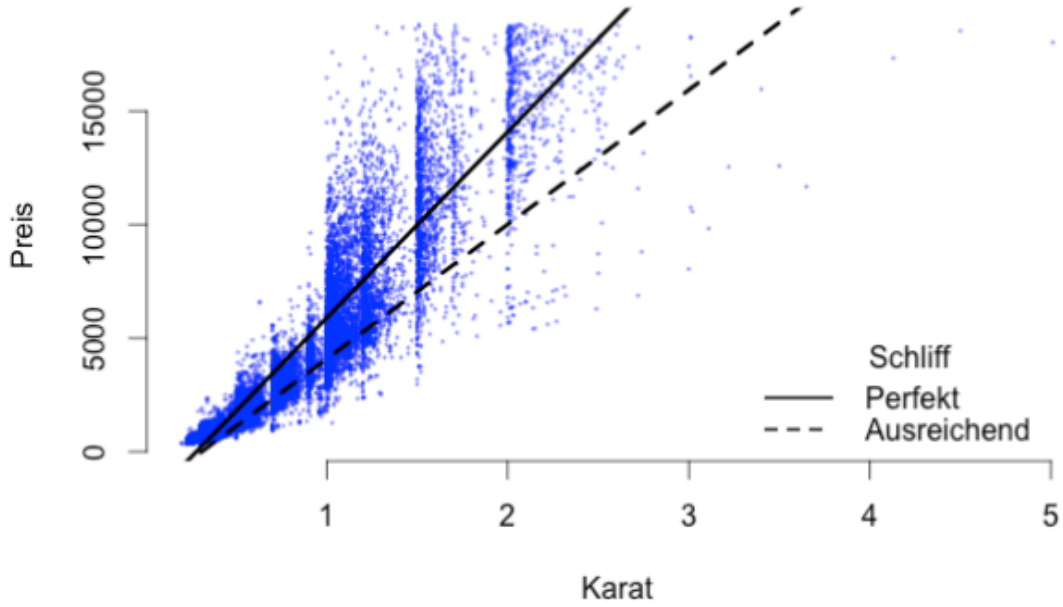
WOFÜR BRAUCHEN WIR REGRESSIONSMETHODEN?

- Die Regression ist eine Methode zur **Detektion von Zusammenhängen**.
- Einfachster Fall: Zusammenhang zweier intervallskalierter Variablen.
  - Diamantengröße und –preis
  - Wortfrequenz und Lesezeit
  - Wortfrequenz und Wortlänge
  - Weglänge und Laufzeit
  - Komplexität und Reaktionszeit



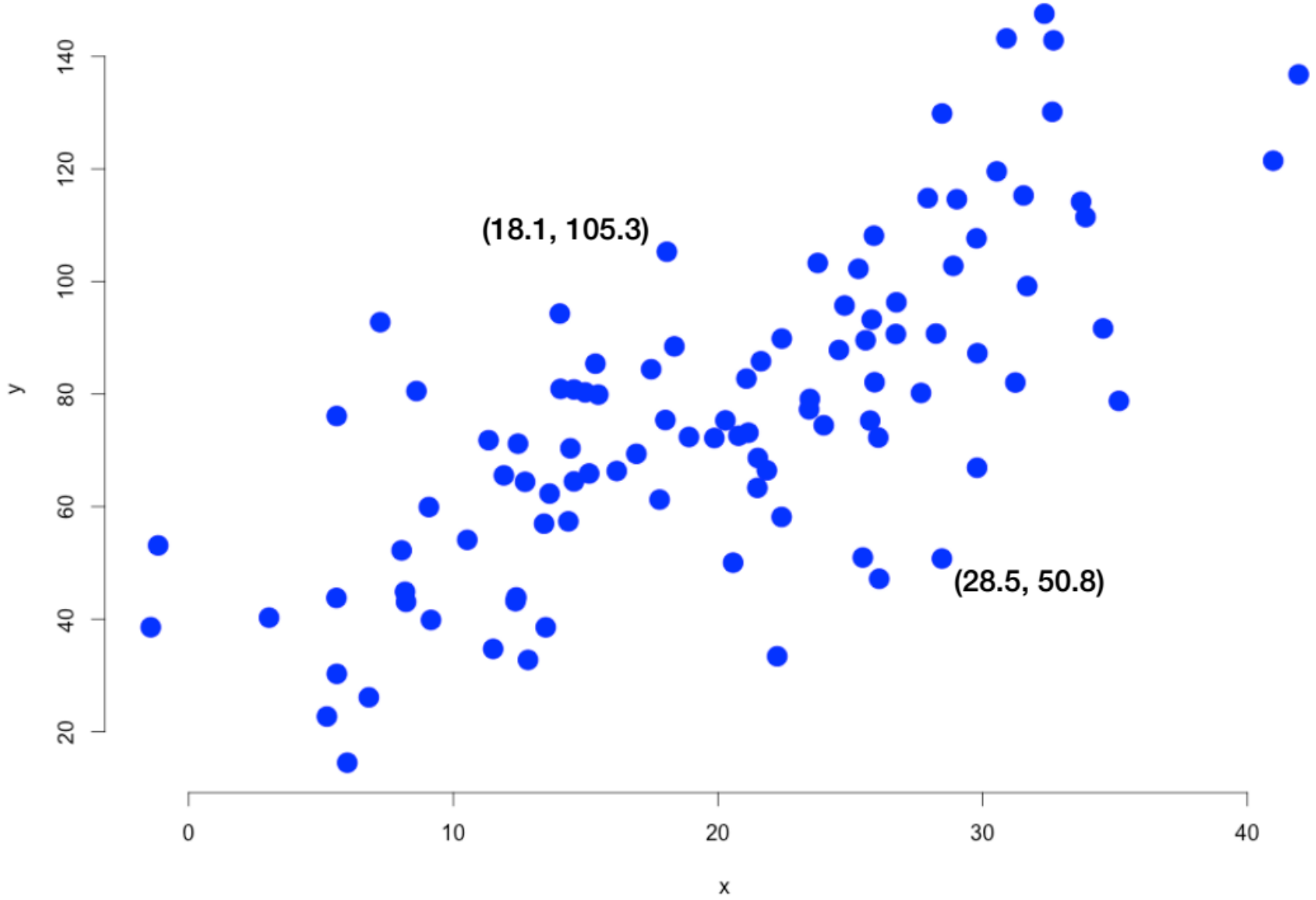
WOFÜR BRAUCHEN WIR REGRESSIONSMETHODEN?

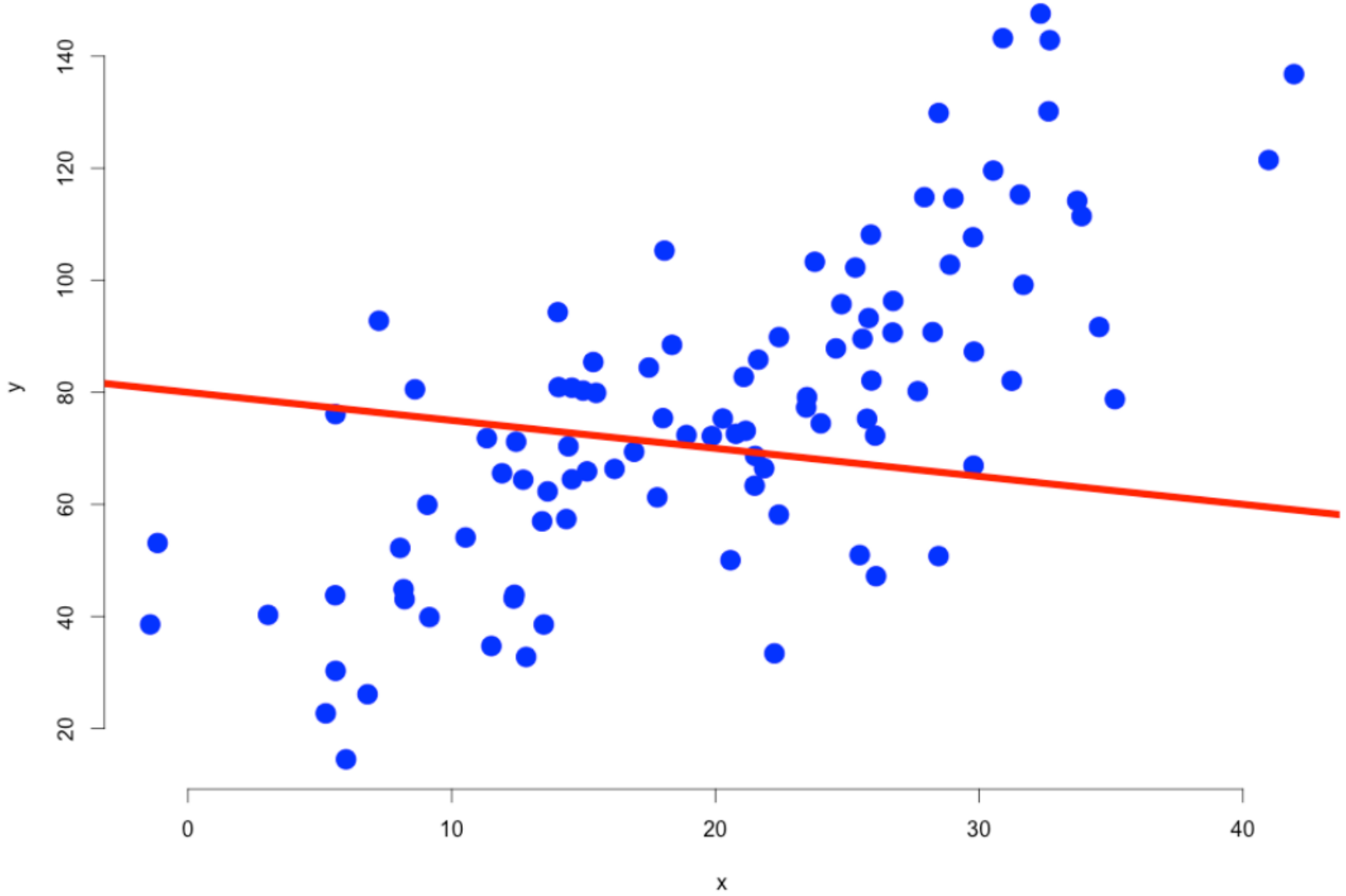
- Grundsätzlich können auch Einflüsse von nominalskalierten Prädiktoren analysiert werden.
- Außerdem ist auch eine Kombination von nominal- und intervallskalierten Prädiktoren möglich (bspw. in Form einer **Interaktion**).

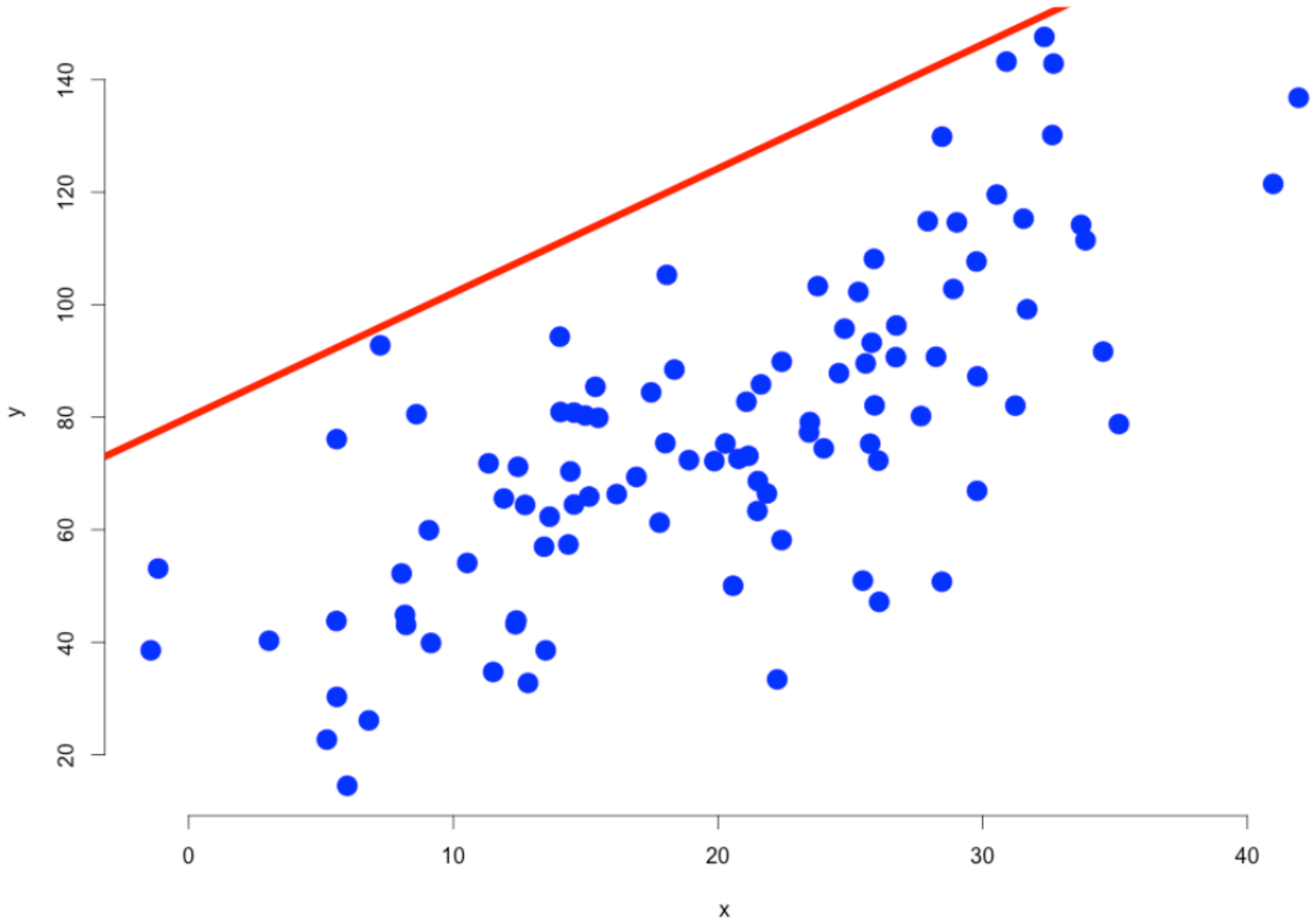


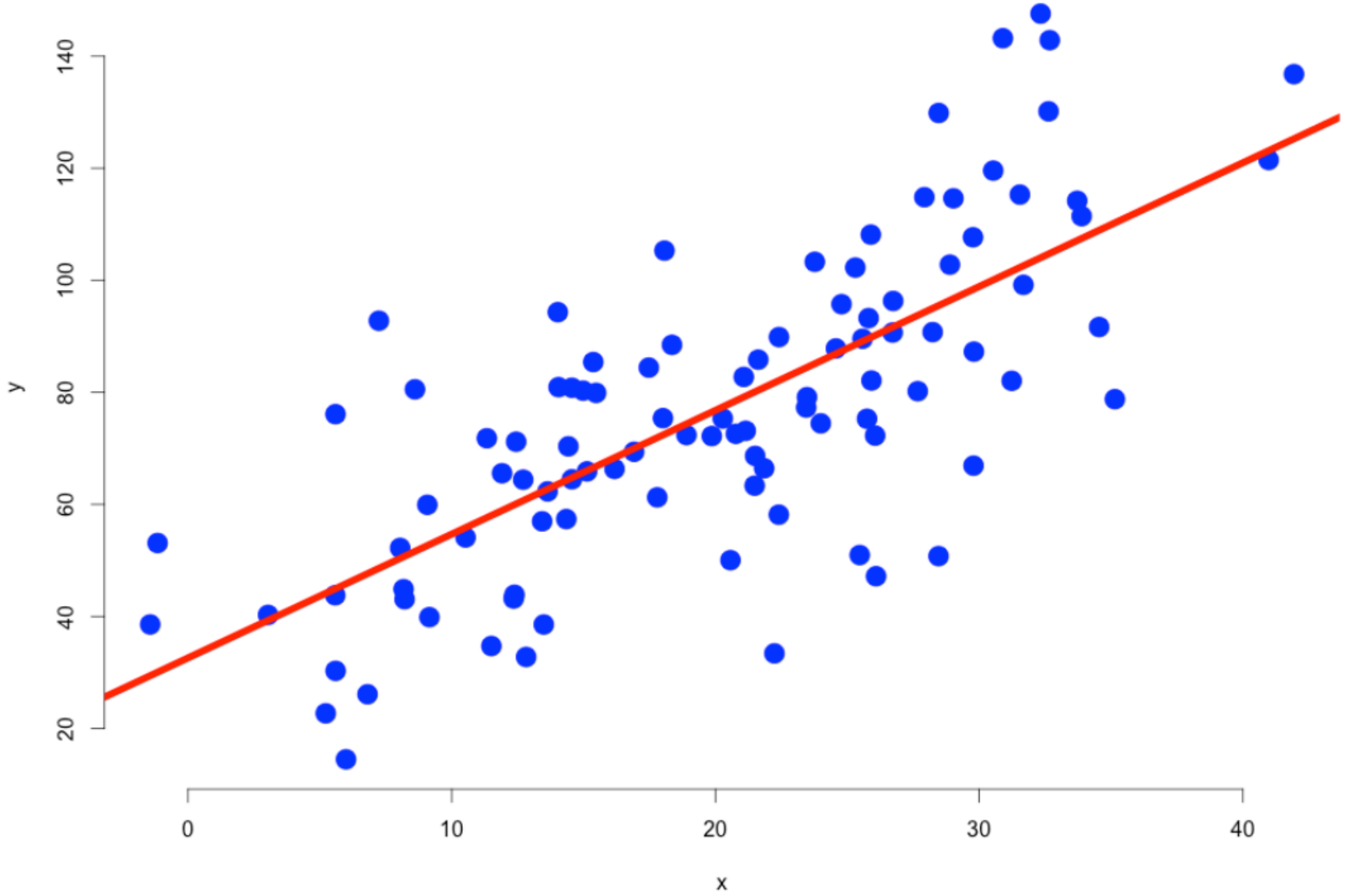
GRUNDIDEE DER REGRESSION

- Einfachster Fall: 1 intervallskalierte Prädiktorvariable *x*, 1 intervallskalierte Kriteriumsvariable *y*.
  - Lineare, univariate Regression.
- Anpassung einer **Gerade** an eine **Punktwolke**.
  - Koordinaten eines Punktes in der Punktwolke ergeben sich auf dem Wert für *x* und *y*.
- Regressionsrechnungen finden die **optimale Gerade** für eine Vorhersage von *y* aus *x*.
  - Optimal: Die Summe aller quadrierten Abweichungen von der Gerade sind möglichst klein.
  - Anders ausgedrückt: Es gibt keine andere Gerade, für die die quadrierten Abweichungen kleiner sind.


(18.1, 105.3)
(28.5, 50.8)
y
x
20
40
60
80
100
120
140
0
10
20
30
40


y
20
40
60
80
100
120
140
0
10
20
30
40
x

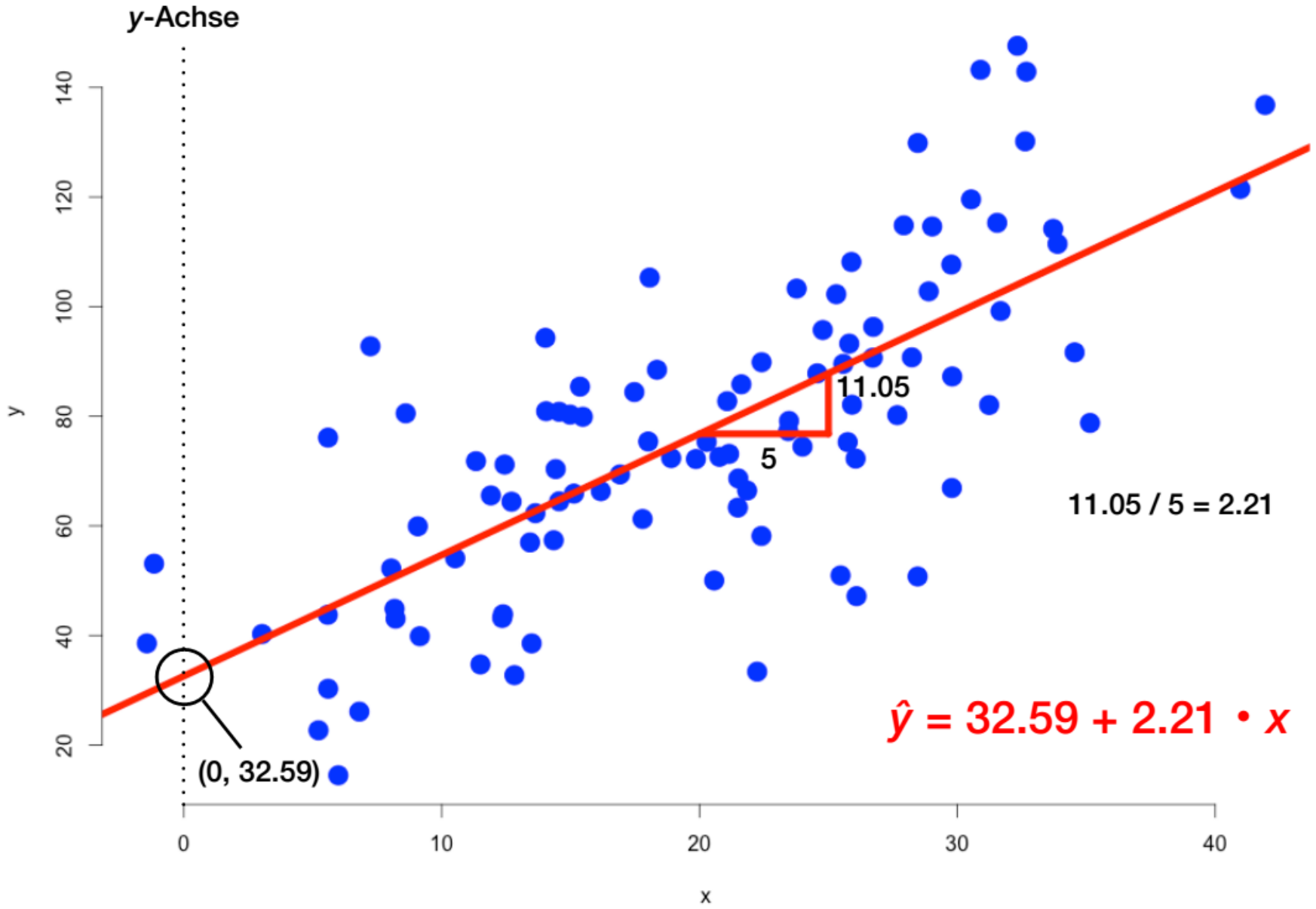

y
20
40
60
80
100
120
140
0
10
20
30
40
x


y
20
40
60
80
100
120
140
0
10
20
30
40
x

GERADEN

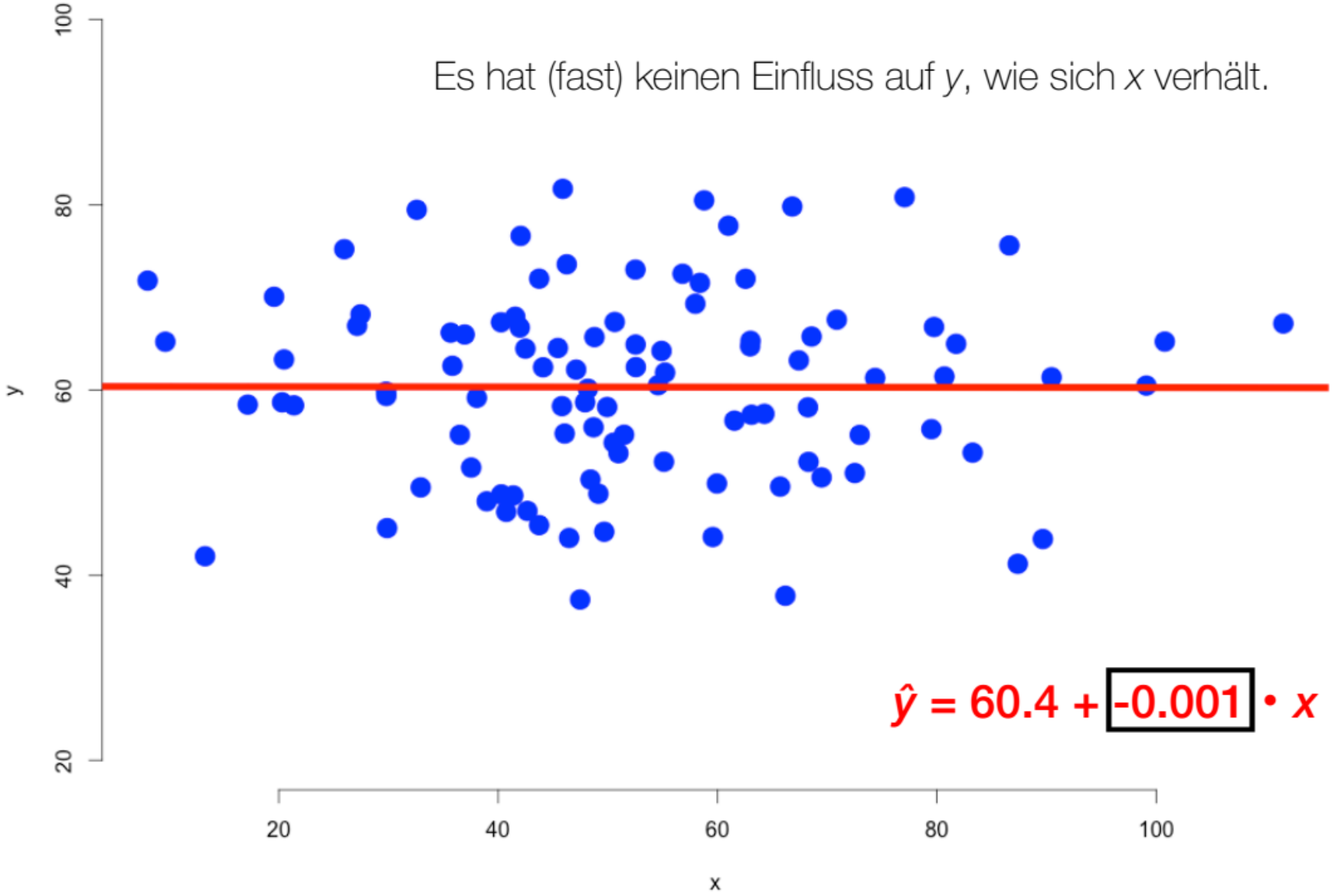
- Geraden definieren sich immer durch zwei Parameter.
  - Intercept / Schnittpunkt mit der *y*-Achse: *a*
  - Slope / Steigung: β

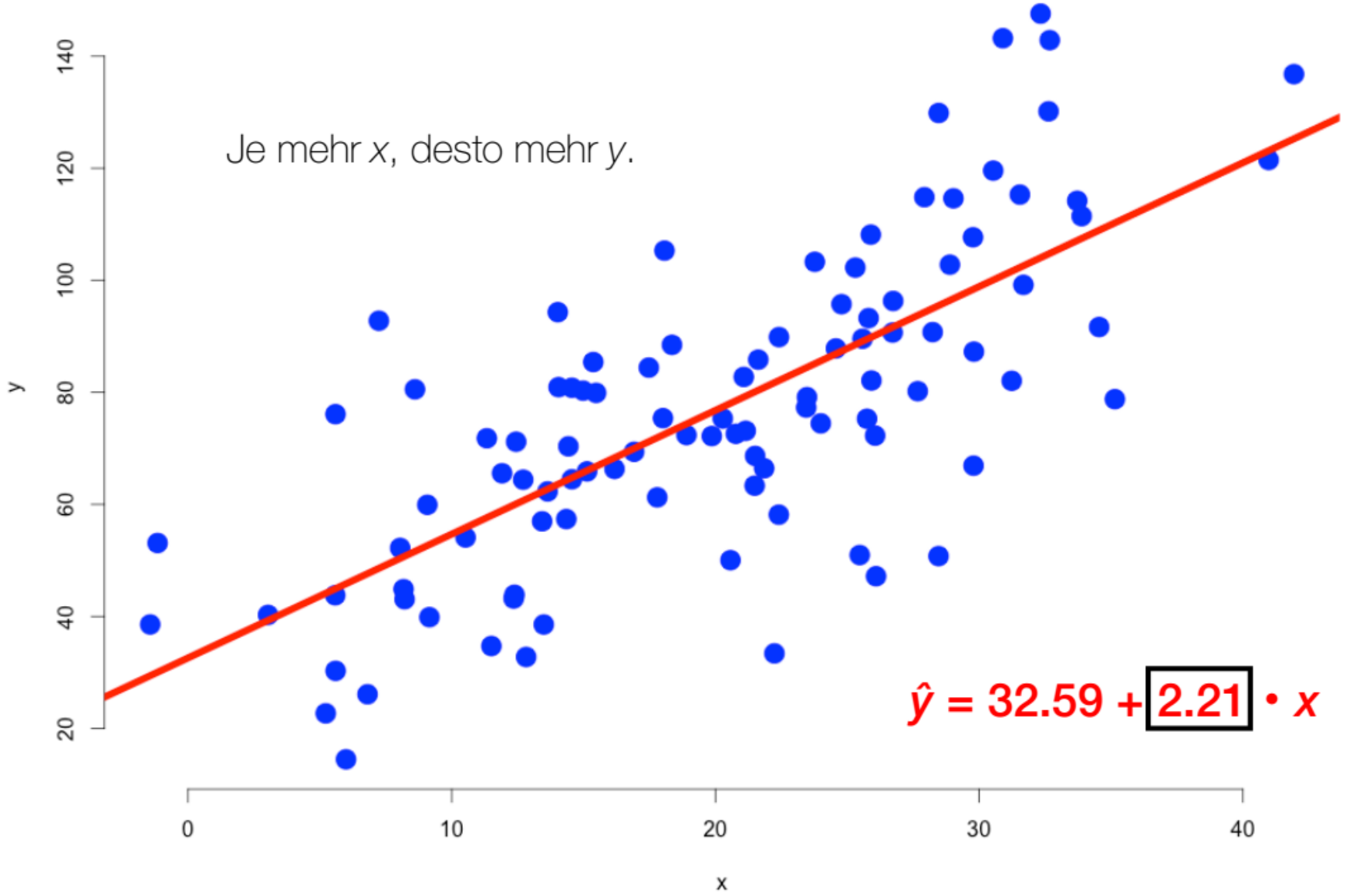
$$y = a + \beta \cdot x$$

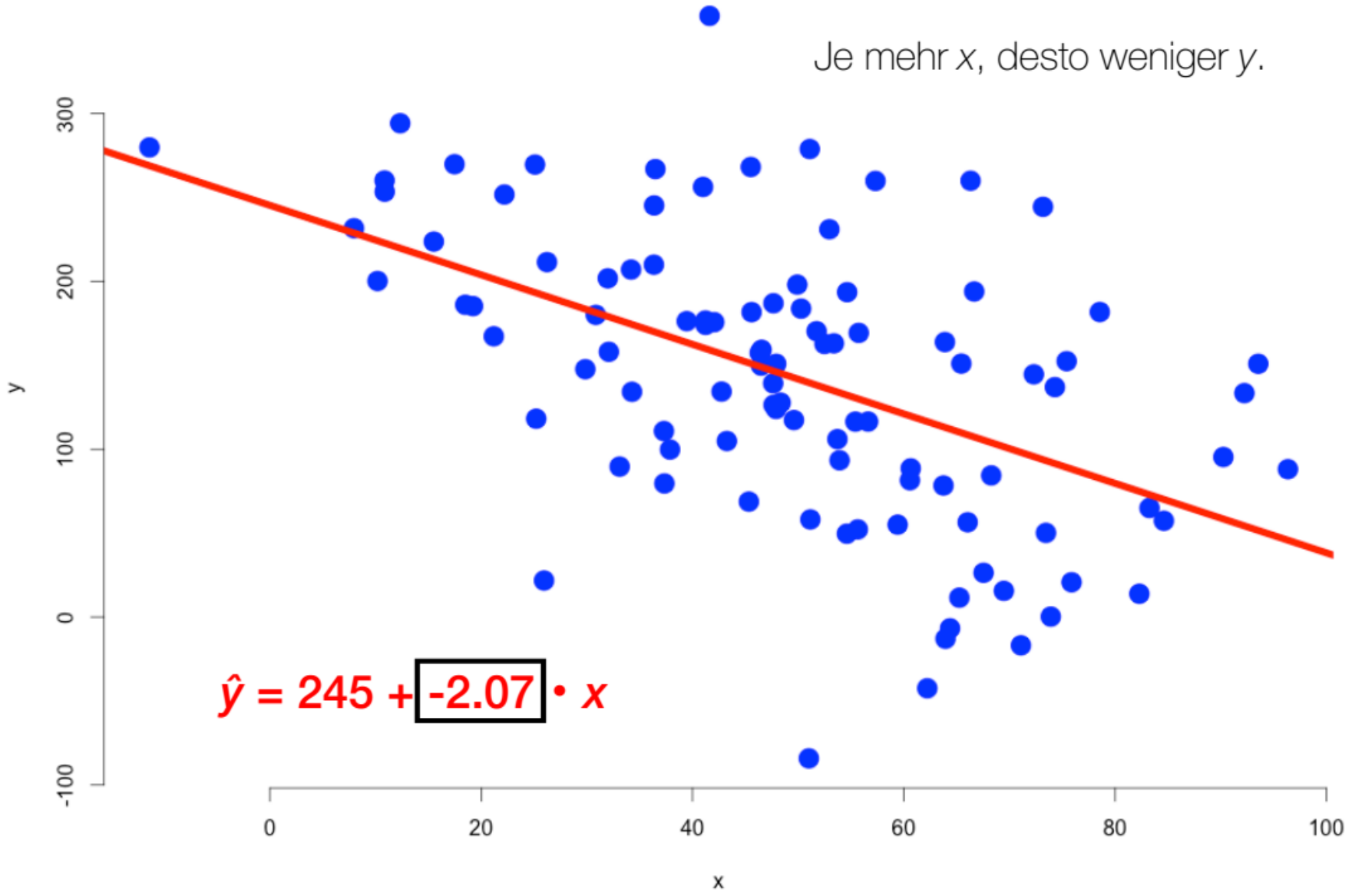

y-Achse
y
x
20
40
60
80
100
120
140
0
10
20
30
40
11.05
5
11.05 / 5 = 2.21
(0, 32.59)
ŷ = 32.59 + 2.21 • x

STEIGUNG

- *y* ist unsere Kriteriumsvariable, wird also aus *x* vorhergesagt.
- Kennen wir die Steigung der Regressionsgeraden, wissen wir, ob und wie *y* von *x* abhängt.
- β und der Korrelationskoeffizient *r* haben für dieselben Daten immer das gleiche Vorzeichen!
- Betrachten wir einmal ein Extrembeispiel mit einer Steigung β nahe 0.

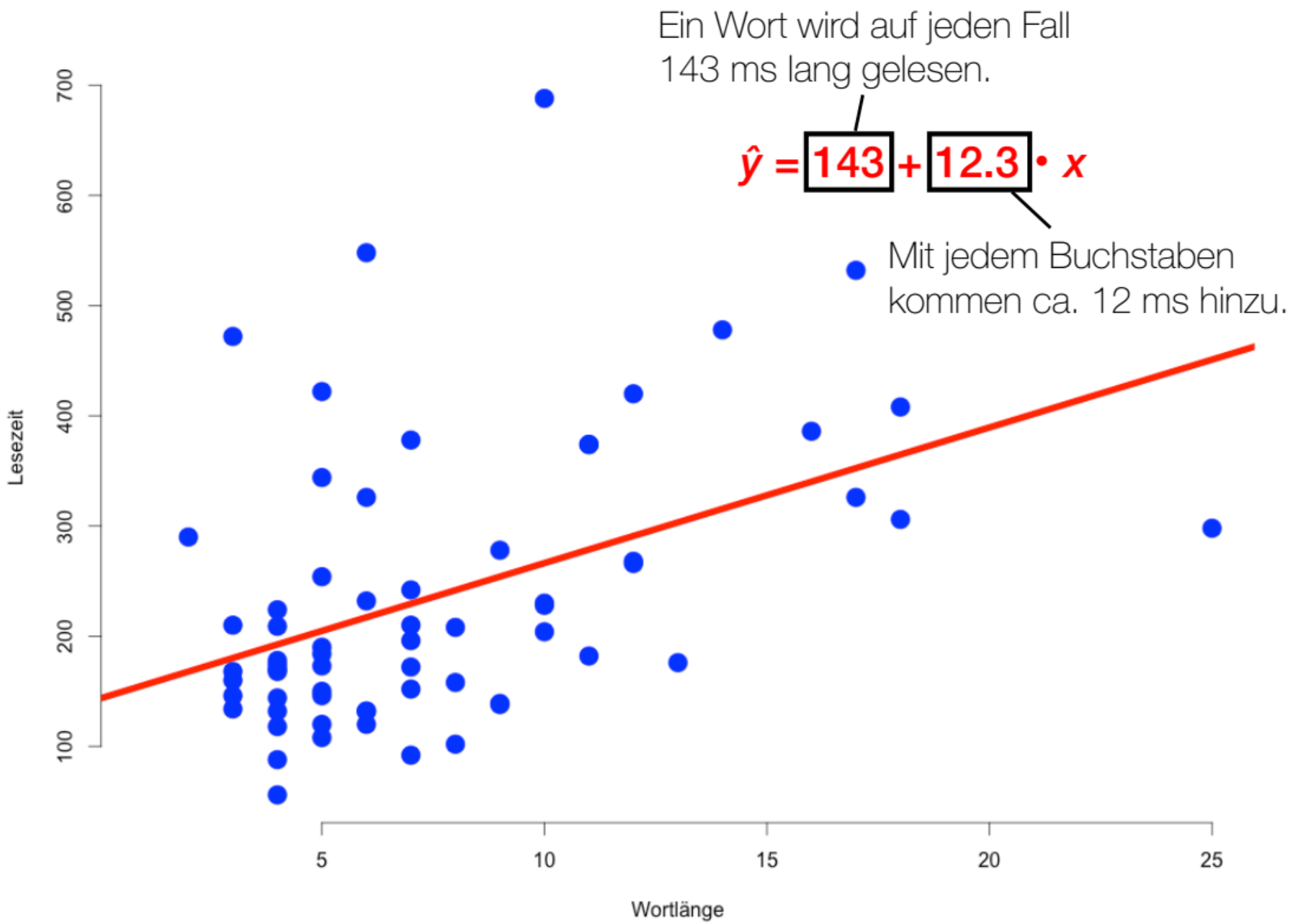

Es hat (fast) keinen Einfluss auf y, wie sich x verhält.
y
x
20
40
60
80
100
ŷ = 60.4 + -0.001 • x

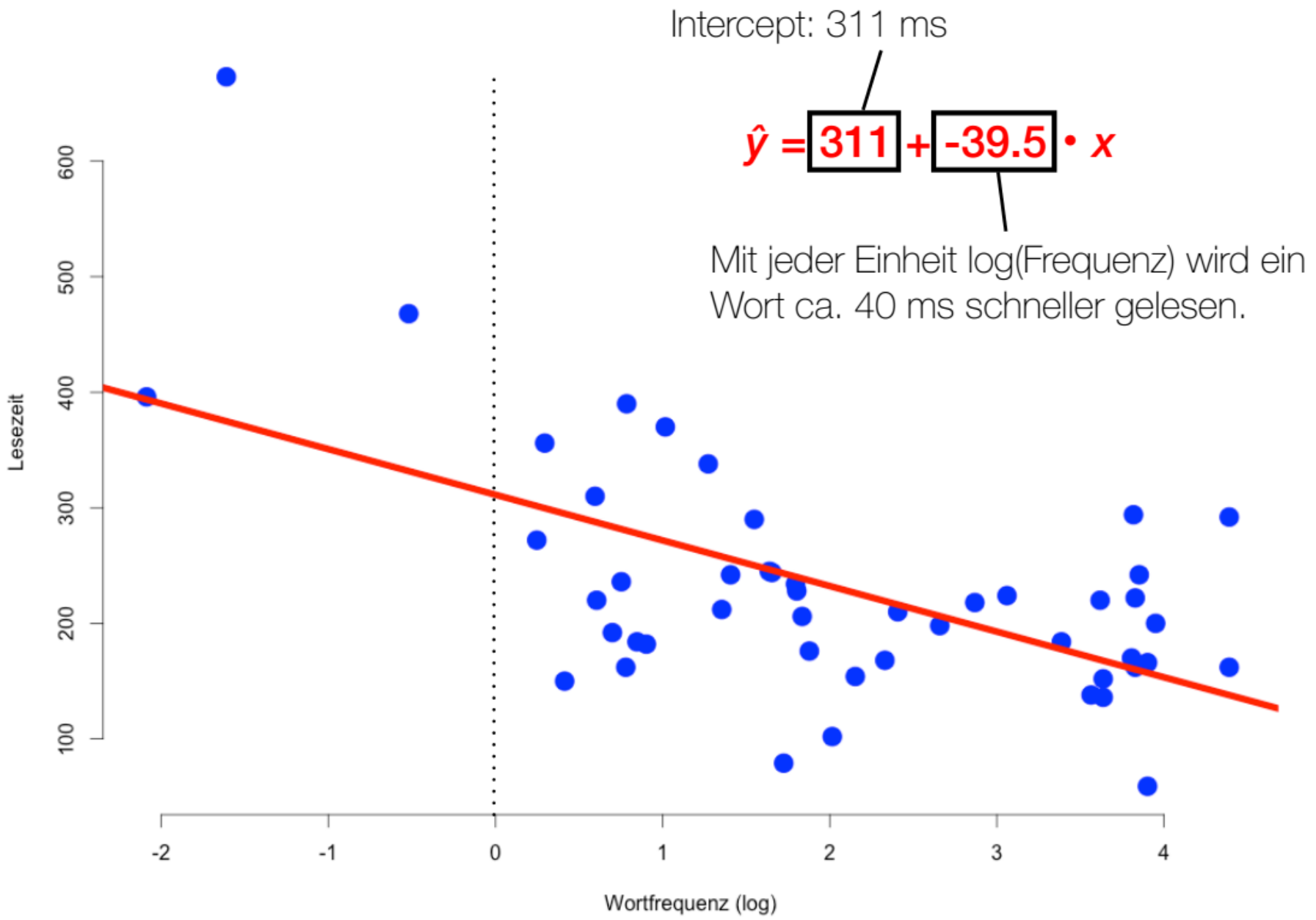

Je mehr x, desto mehr y.
$\hat{y} = 32.59 + \boxed{2.21} \cdot x$
y
x
20
40
60
80
100
120
140
0
10
20
30
40

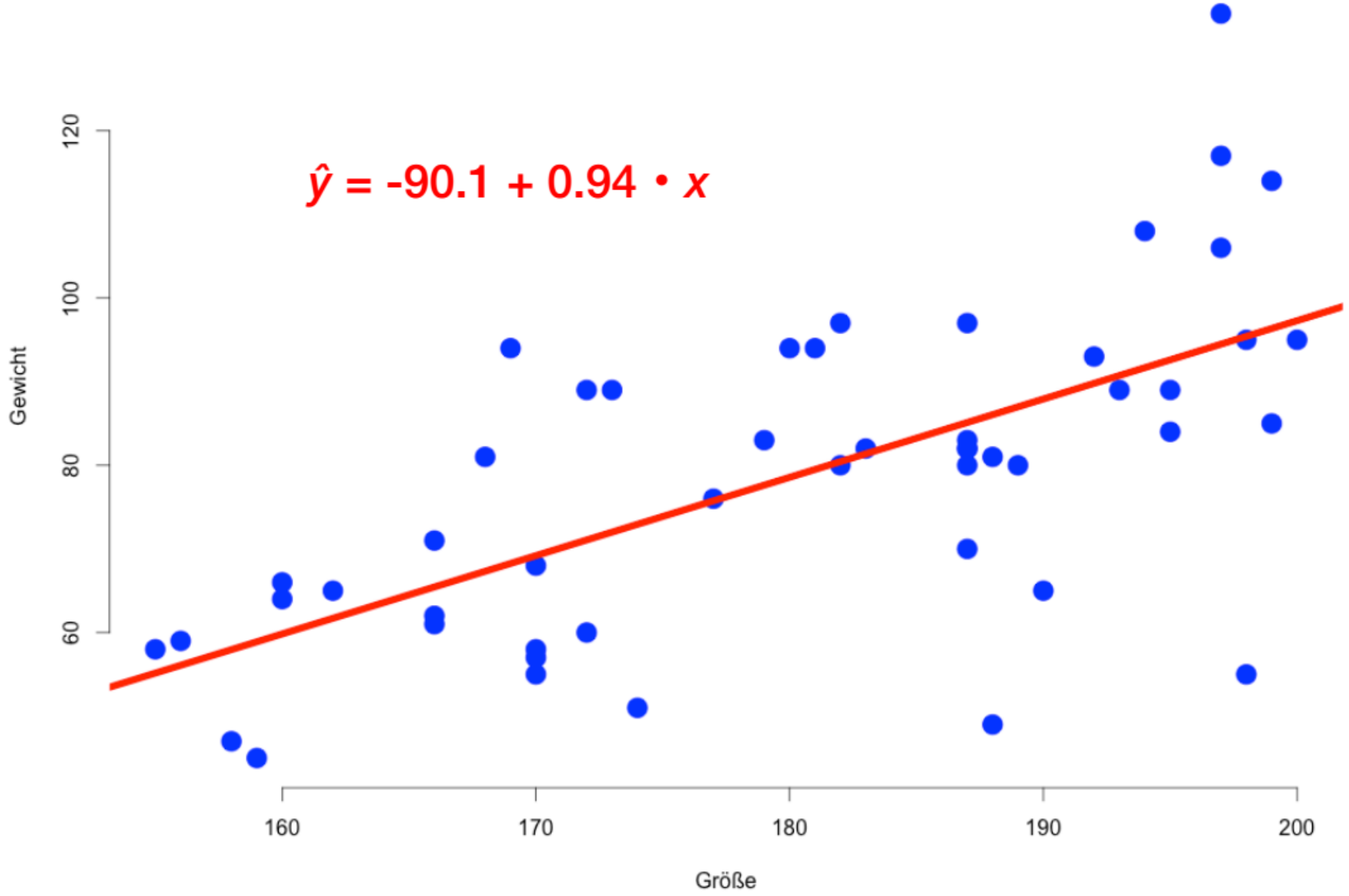

Je mehr x, desto weniger y.
ŷ = 245 + -2.07 • x
y
x
300
200
100
0
-100
0
20
40
60
80
100

REGRESSIONSPARAMETER

- Interpretation der Regressionsparameter.
  - Der Intercept (Schnittpunkt mit der *y*-Achse) ist selten sinnvoll interpretierbar.
  - Die Steigung ist immer so zu interpretieren, dass sich die Kriteriumsvariable *y* mit **einem** Schritt von *x* um den Wert β (Steigung) verändert.


Ein Wort wird auf jeden Fall
143 ms lang gelesen.
ŷ = 143 + 12.3 • x
Mit jedem Buchstaben
kommen ca. 12 ms hinzu.
Lesezeit
100
200
300
400
500
600
700
5
10
15
20
25
Wortlänge


Intercept: 311 ms
ŷ = 311 + -39.5 • x
Mit jeder Einheit log(Frequenz) wird ein Wort ca. 40 ms schneller gelesen.
Lesezeit
100
200
300
400
500
600
-2
-1
0
1
2
3
4
Wortfrequenz (log)


$\hat{y}$ = -90.1 + 0.94 • x
Gewicht
120
100
80
60
160
170
180
190
200
Größe

FIT / ANPASSUNG

- Der Regressionsalgorithmus findet jene Gerade, für die gilt: Die Summe aller quadrierten Abweichungen (*sum of squares*, Quadratsumme) von den Punkten zur Gerade ist möglichst klein.
- Die Summe aller Abweichungen (nicht quadriert) ist immer gleich 0!
- Wichtig: Im Gegensatz zur Korrelationsrechnung ist die Regression nicht bidirektional!
  - Die Vorhersage von *y* aus *x* ergibt eine andere Regressionsgerade als die Vorhersage von *x* aus *y*!

REGRESSION IN R

- Um eine lineare Regression mit intervallskalierten Prädiktorvariablen in R durchzuführen, benötigen wir die Funktion `lm()` für "linear model".
- Die Syntax lautet in allgemeiner Form:

```
lm(<Formel>, data = <Datensatz>)
```

?

R-FORMELN

- "Formeln" drücken in R das anzupassende Modell aus.
- Sie sind quasi eine Repräsentation dessen, was das Regressionsmodell für uns überprüfen soll (unsere Hypothese).
- Formeln enthalten immer ein **Tilde-Zeichen** (~), das auch als "*predicted by*" oder "*depends on*" gelesen werden kann.
- Vor und hinter der Tilde stehen immer Spalten in unserem Datensatz.

```
lm(gewicht ~ groesse, data = datensatz)
```

"Berechne ein lineares Modell, das das Gewicht aus der Größe vorhersagt. Benutze dazu den Datensatz "datensatz".

MULTIPLE REGRESSION

- Bisher gezeigte Fälle: Eine Prädiktorvariable.
- In der Praxis häufiger: Mehrere Prädiktoren!
- Sehr oft interessiert uns auch nicht nur der einzelne Effekt der Prädiktoren, sondern das **Zusammenwirken** zweier Prädiktoren.
  - **Interaktionen**: Der Effekt eines Prädiktors ist unterschiedlich für die Stufen des anderen Prädiktors.
  - Beispiel 1: Das Gewicht eines Diamanten wirkt sich stärker auf den Preis aus, wenn der Diamant perfekt geschliffen ist (im Vergleich zu einem ausreichend guten Schliff).
  - Beispiel 2: Je frequenter ein Wort ist, desto eher wird es reduziert. Das gilt aber insbesondere für konzeptionell mündliche Textgattungen.

MULTIPLE REGRESSION

- Sobald wir mehr als einen Prädiktor in unser Modell aufnehmen, sprechen wir von **multipler Regression**.
- Dabei wird jedem Prädiktor die "gleiche Chance" eingeräumt, einen signifikanten Einfluss auf die Kriteriumsvariable zu zeigen.
- Für jeden Prädiktor wird ein Parameter (Slope/Steigung, Estimate/Effektschätzer) im Regressionsmodell hinzugefügt.
- Die unabhängige Wirkung eines Prädiktors wird auch **Haupteffekt** genannt.
- Die R-Formeln müssen entsprechend erweitert werden…

R-FORMELN MULTIPLE REGRESSION

1 Prädiktor

```
lm(lesezeit ~ wortlänge, data = dat)
```

2 Prädiktoren, Haupteffekte

```
lm(lesezeit ~ wortlänge + wortfrequenz,
                data = dat)
```

2 Prädiktoren, Haupteffekte

```
lm(lesezeit ~ wortlänge + wortfrequenz +
  wortlänge:wortfrequenz, data = dat)
```

2 Prädiktoren, Interaktion

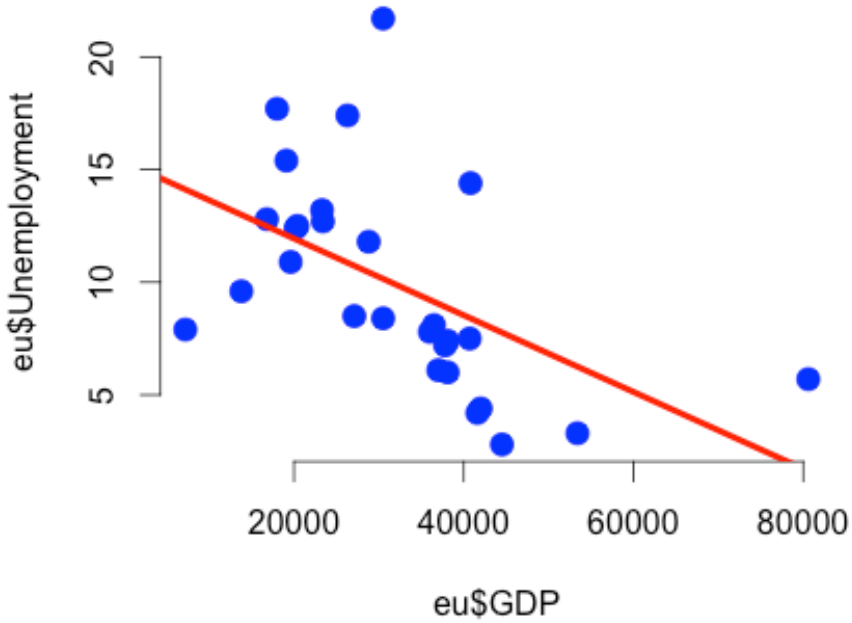
2 Prädiktoren, Haupteffekte & Interaktion

```
lm(lesezeit ~ wortlänge * wortfrequenz,
                data = dat)
```

REGRESSION IN R

- Typisches Vorgehen beim Berechnen eines Regressionsmodells:
  - `mod1 <- lm(<Formel>, <Daten>)`
  - `summary(mod1)`
- So haben wir das "Modell-Objekt" `mod1` immer verfügbar und können weitere Informationen daraus extrahieren.
  - Plotten
  - Effektschätzer
  - *p*-Werte
  - Rohdaten
  - …
  - ➔ alle im Modell-Objekt enthaltenen Informationen: `str(mod1)`

OUTPUT




```
> summary(mod1)
Call:
lm(formula = Unemployment ~ GDP, data = eu)
Residuals:
    Min      1Q  Median      3Q     Max
-6.220 -2.881 -1.070   1.474 11.547

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.535e+01  1.875e+00   8.183 1.15e-08 ***
GDP         -1.702e-04  5.373e-05  -3.168   0.0039 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.049 on 26 degrees of freedom
Multiple R-squared:  0.2785,       Adjusted R-squared:  0.2508
F-statistic: 10.04 on 1 and 26 DF,  p-value: 0.003898
```

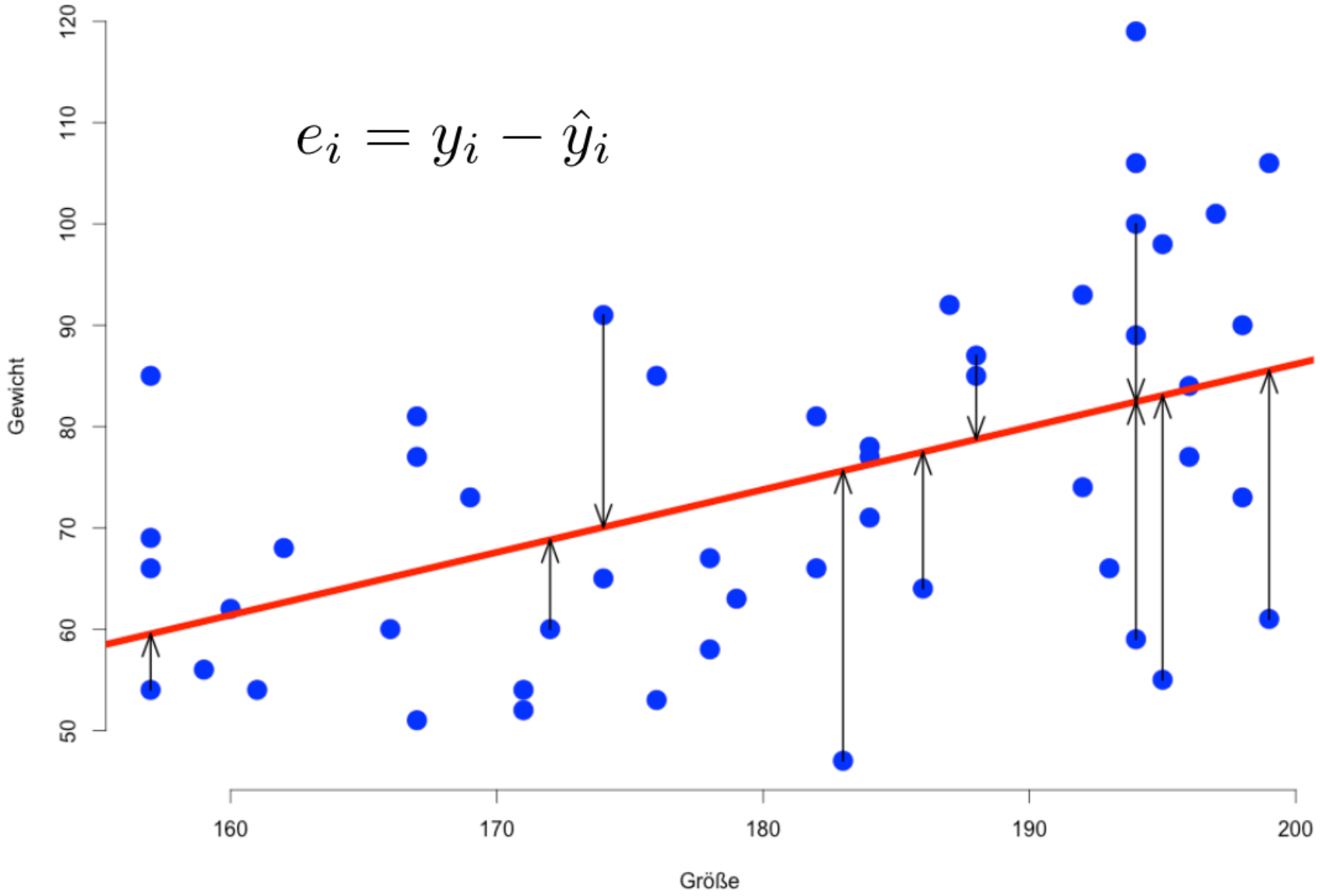
LOGISTISCHE REGRESSION

- Ist die Kriteriumsvariable nicht intervallskaliert, sondern **binär**, kann eine logistische Regression durchgeführt werden.
- Typische Kriteriumsvariablen:
  - Eintreten vs. Nicht-Eintreten
  - korrekt vs. falsch
- In R können logistische Regressionen mit **generalisierten linearen Modellen** berechnet werden.
  - Der Parameter `family` muss auf `"binomial"` gesetzt werden.
  - `glm(skipped ~ freq, data = dat, family = "binomial")`

R-ÜBUNG

VORHERSAGEFEHLER

- Eine fehlerfreie Vorhersage von *y* aus *x* wäre nur möglich, wenn alle Punkte auf **einer** Linie lägen.
  - Diese Linie wäre dann auch die Regressionsgerade und kein Punkt würde von ihr abweichen.
- Diesen Fall werden wir in der empirischen Realität aber niemals antreffen ➔ Jede Vorhersage ist mit Fehlern behaftet.
- Vorhersagefehler = "**Residuen**".
  - Das, was von der Varianz in *y* "übrigbleibt", wenn *x* beachtet ist.


$e_i = y_i - \hat{y}_i$
Gewicht
50
60
70
80
90
100
110
120
160
170
180
190
200
Größe

RESIDUEN

- Die Quadratsumme der Fehler wird bei der Regression minimiert, und die Summe der Fehler beträgt 0.
  - Das gilt **nur für die Regressionsgerade**.
- Das Residuum für einen Datenpunkt x_i wird berechnet aus der Abweichung des tatsächlichen und des vorhergesagten Werts.
  - $e_i = y_i - \hat{y}_i$

RESIDUEN

- Vorhersagefehler sind in vielen Fällen inhaltlich interpretierbar.
  - Ein Wort wird zu kurz/lang gelesen, gegeben dessen Länge.
  - Ein Wort wird zu häufig/selten nachgeschlagen, gegeben dessen Korpusfrequenz.
  - Jemand ist zu leicht/schwer, gegeben die Körpergröße.
  - Ein Fußballclub hat zu wenig/zu viel Punkte, gegeben den Wert des Kaders.
- In manchen Fällen kann es attraktiv sein, die Residuen einer Regression als Kriteriumsvariable in einer weiteren Analyse zu verwenden.

VORAUSSETZUNGEN

- Die Residuen sollten untereinander nicht korreliert sein.
  - Bei Zeitreihenanalysen könnten die Störeinflüsse einer Periode von den Störeinflüssen der vorherigen Periode beeinflusst sein.
- Die Residuen sollten annähernd normalverteilt sein.
  - Prüfung anhand von Histogrammen, Dichtekurven, ...
- Die Streuung der Residuen sollte über den ganzen Wertebereich der abhängigen Variablen in etwa konstant sein (Homoskedastizität).
  - Prüfung anhand eines Plots Prädiktor vs. Residuum.

ZUSAMMENFASSUNG

- Mit Regressionsmethoden wird eine Kriteriumsvariable aus einem oder mehreren Prädiktor(en) vorhergesagt.
- Damit möchte man den Zusammenhang von Variablen klären.
- Bei der linearen Regression mit einer Prädiktorvariable wird eine Gerade an eine Punktwolke angepasst.
  - Geraden definieren sich durch Intercept und Slope (Steigung).
  - Steigung repräsentiert den Effekt von *x* auf *y*.
- Die Abweichungen der beobachteten von den vorhergesagten Werten (Vorhersagefehler) nennt man "Residuen".
  - Die Summe der Residuen ist 0.

SELEKTION DER REGRESSIONSMETHODE

- Das zu verwendende Regressionsmodell ergibt sich aus
  - Prädiktorstruktur
  - Skalierung der Kriteriumsvariable

	Prädiktorstruktur	
	ohne Zufallseffekte	mit Zufallseffekten
binär	logistisches lineares Modell	logistisches lineares gemischtes Modell
kontinuierlich	lineares Modell	lineares gemischtes Modell