Moderne Kausalanalyse

Rubin Causal Model und Directed Acyclic Graphs

Simon Ress

Ruhr-Universität Bochum

31.10.2018

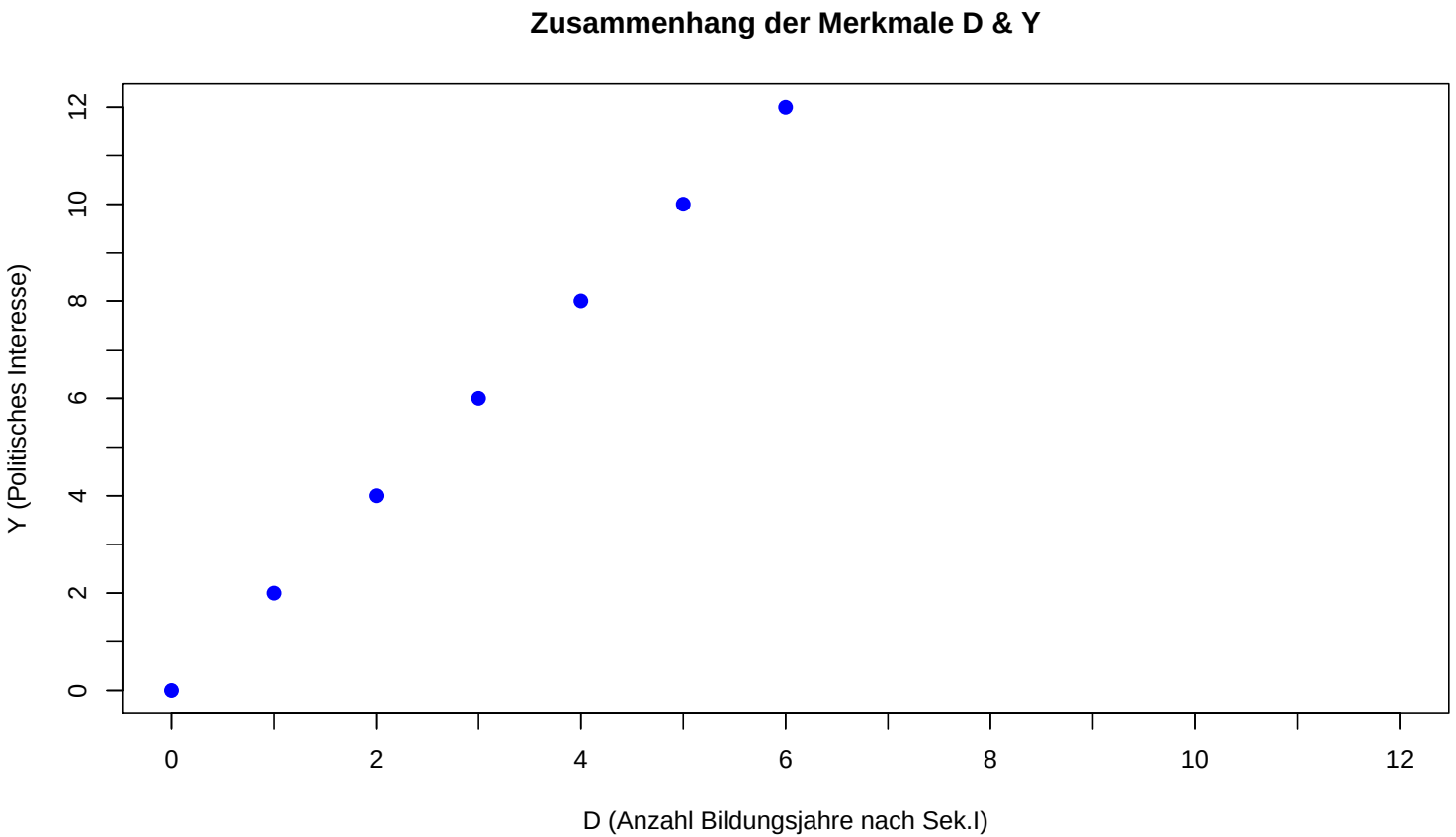
Inhalt

Warum eine Regressionsanalyse durchführen?

- Die Ausprägungen der unabhängigen Variable [D] sollen auf die Ausprägungen der abhängigen Variable [Y] zurückgeführt werden
- Zusammenhang beider Variablen beschreiben (mit Hilfe von Formeln)
- Postulierte Hypothese(n) testen

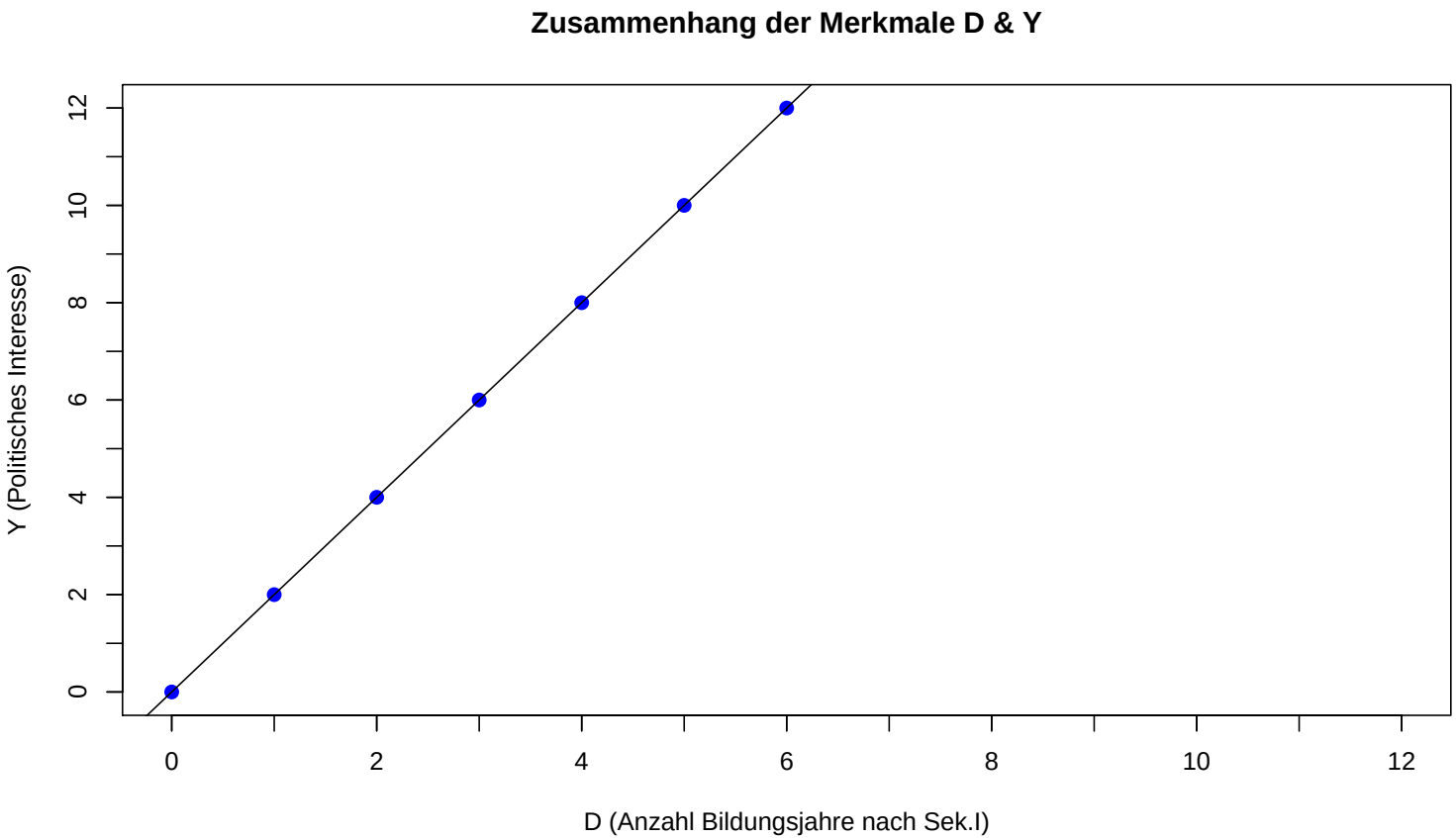
Warum eine Regressionsanalyse durchführen

- Die Ausprägungen von zwei Variablen (Y & D) sind für sechs Einheiten dargestellt
- Mit welchem mathematischen Ausdruck können die Werte von Y auf D zurück geführt werden?



Linearer (perfekter) Zusammenhang

- Die Variable Y ist immer doppelt so groß wie die Variable D
- Mathematische Formulierung: y = d * 2 (Kleinschreibung bei realisierten Werten einer Variable)



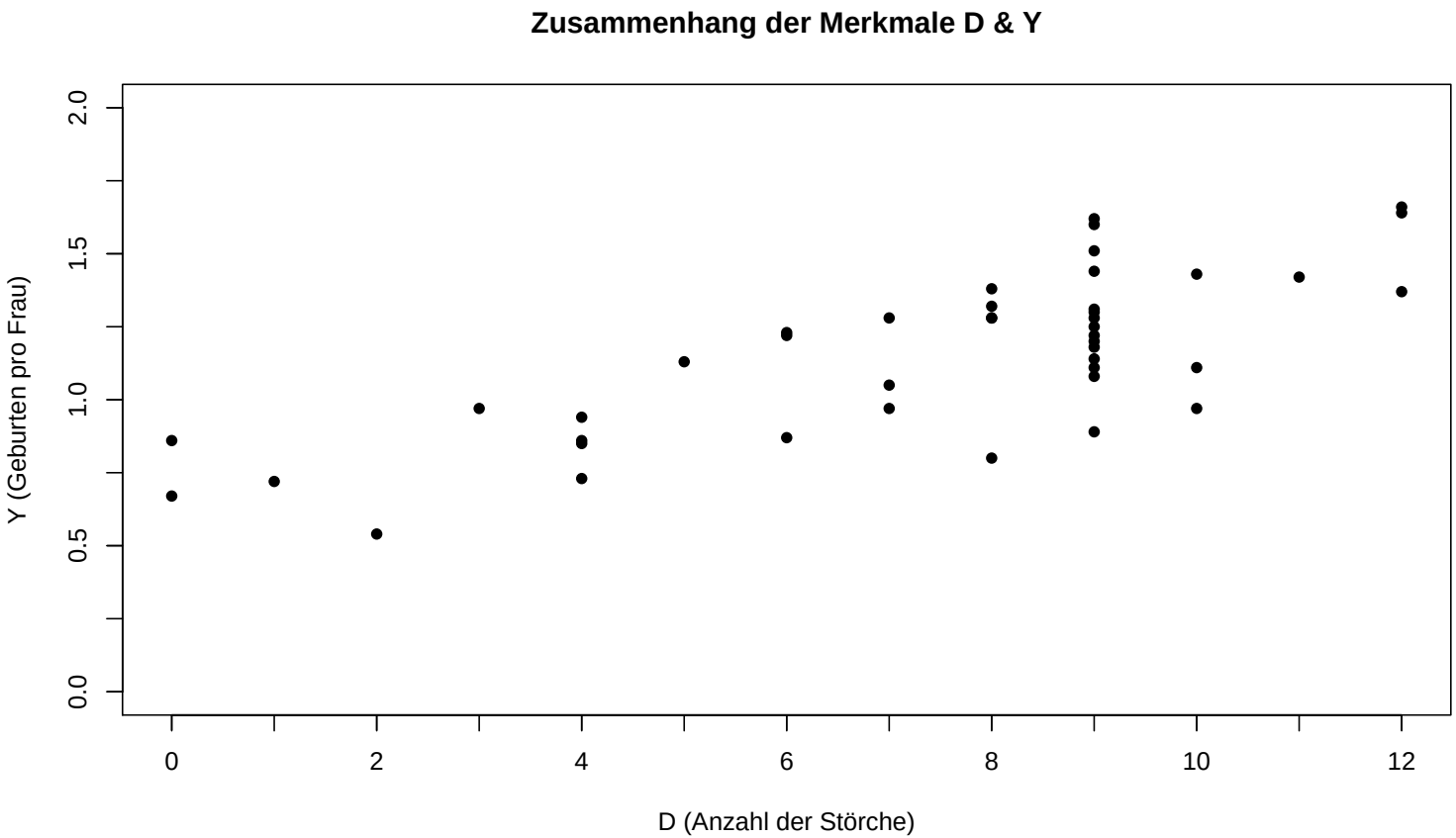
Datensatz

	Geburten	Störche	Einwohnerdichte
1	1.8	17.0	185.0
2	1.6	14.0	241.0
3	1.7	17.0	170.0
4	1.1	9.0	291.0
5	1.8	15.0	187.0
6	1.1	13.0	252.0
7	0.7	4.0	358.0
8	0.9	4.0	349.0
9	0.8	8.0	314.0
10	2.2	25.0	89.0
11	0.5	2.0	367.0
12	1.2	9.0	281.0
13	1.3	8.0	301.0
14	1.3	8.0	297.0
15	1.4	9.0	290.0
16	1.1	5.0	326.0
17	2.1	18.0	176.0
18	0.7	1.0	405.0
19	2.2	19.0	151.0
20	1.3	9.0	280.0
21	2.8	33.0	3.0
22	1.7	12.0	259.0
23	1.9	14.0	212.0
24	1.4	10.0	271.0
25	1.1	9.0	284.0

Tabelle: Geburten, Störche und Einwohnerdichte (25 von 100 Beobachtungen)

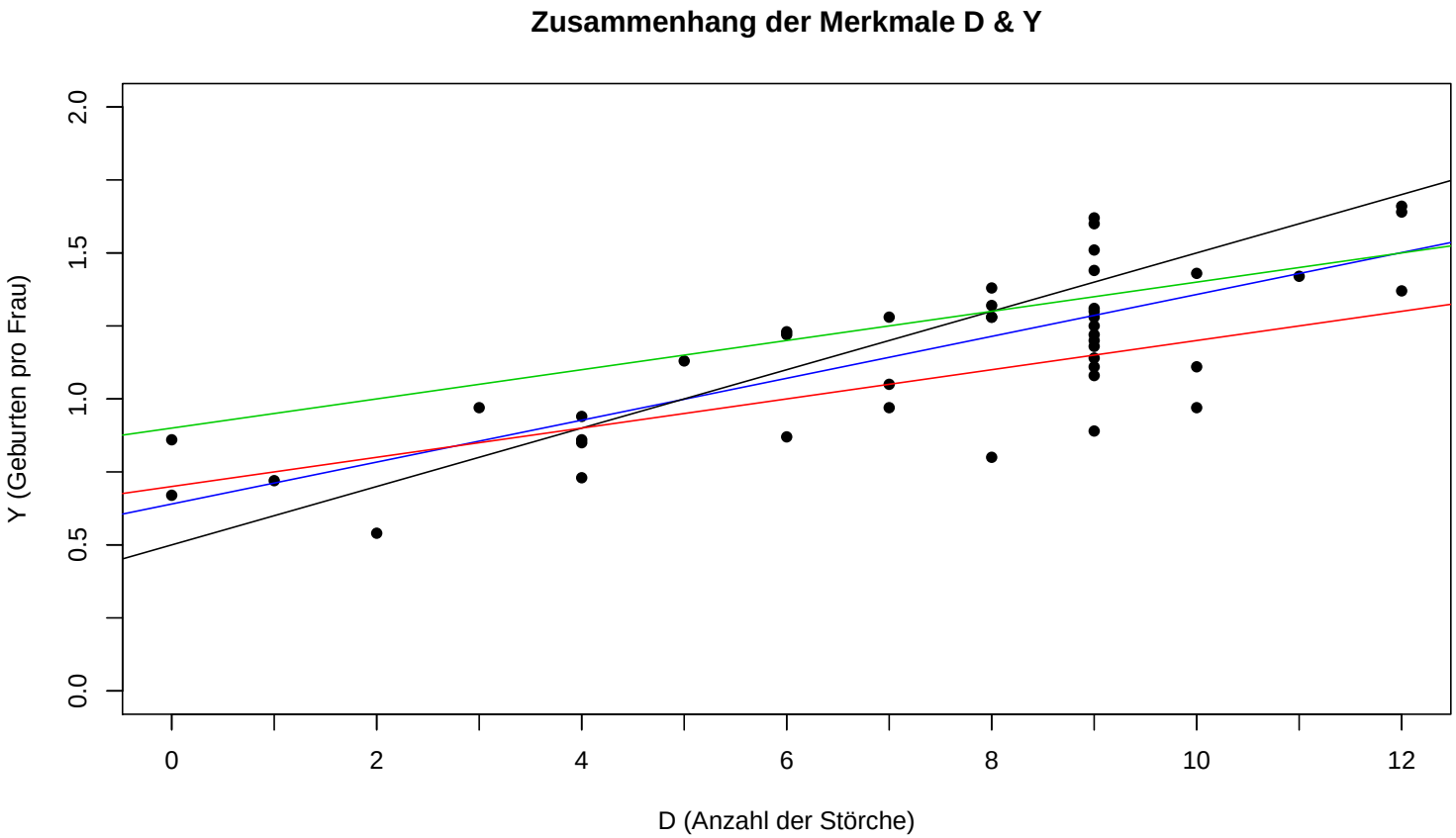
Linearer Zusammenhang

- Der Zusammenhang zwischen D (Störche) und Y (Geburten) ist nicht perfekt
- Wie kann dieser Zusammenhang beschreiben werden?



Linearer Zusammenhang

- Welche (lineare) mathematische Funktion beschreibt den Zusammenhang ab besten?
- Welches Kriterium kann zur Entscheidung herangezogen werden?



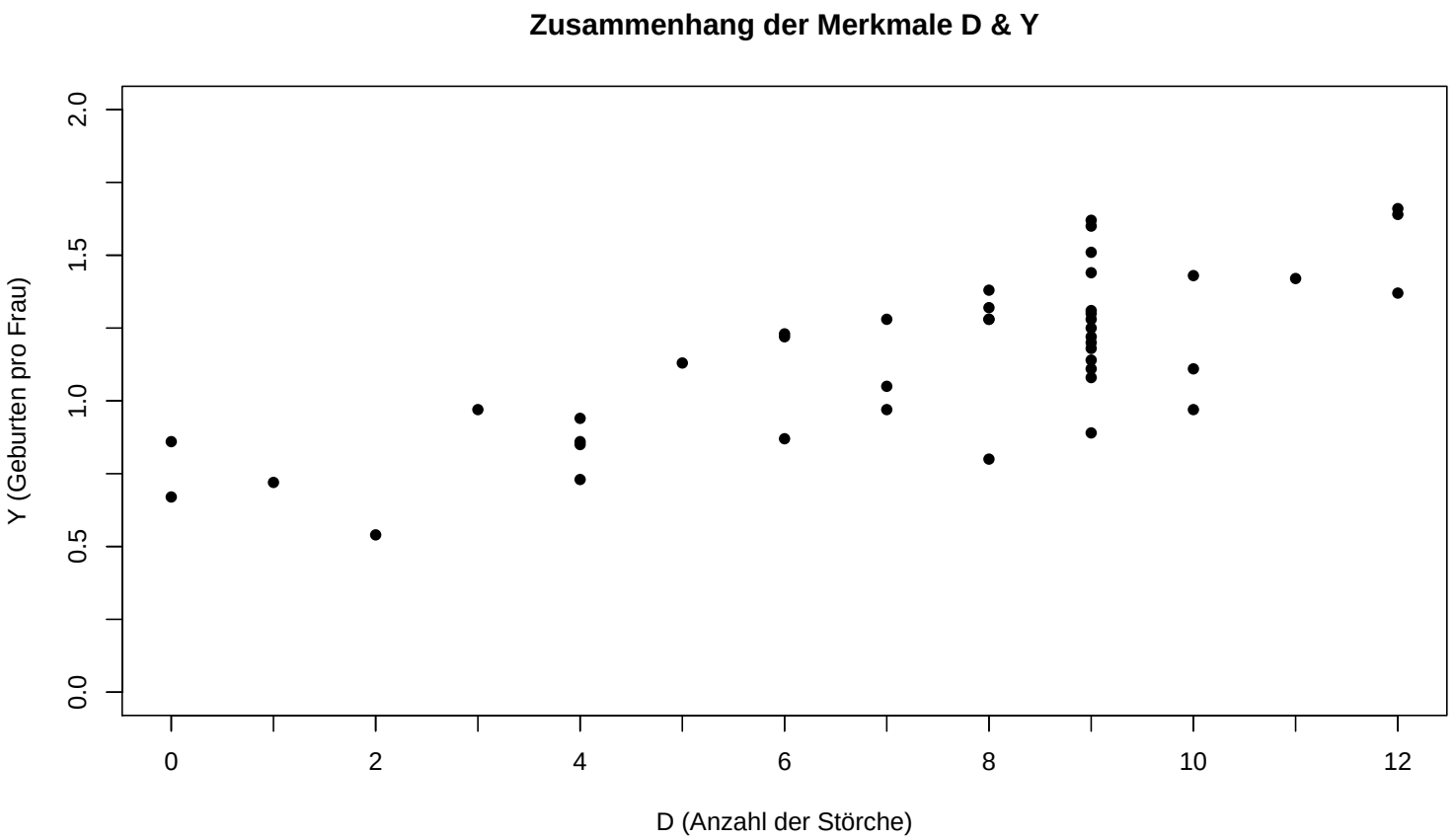
Ordinary Least Squares (OLS-) Methode

- (dt.) Methode der kleinsten Quadrate
- Idee: Mathematische Funktion bestimmen, welche möglichst gut den Zusammenhang der Daten beschreibt
- Abweichung der Datenpunkte von dem geschätzten Ergebniss: Residuen
- Auswahl der Funktion: Minimale Summe der quadrierten Abweichungen
- Für dieses Minimierungsproblem existieren verschiedene Algorithmen

Voraussetzungen

- Variable Y hat ein metrisches Skalenniveau
- Auch kategoriale Skalenniveaus möglich, die dann als metrisch angenommen werden
- Mindestvoraussetzungen: mindestens fünf Kategorien, ordinales Messniveau, Abstände zwischen Ausprägungen sind gleich groß, Kategorien als Wertintervalle

Beispiel Störche und Geburten



Variablen

Für die Berechnung einer Regression kann folgender Befehl in R genutzt werden

```
lm(data$Geburten~data$Störche)
```

Lineare Regression

```
##
## Call:
## lm(formula = data$Geburten ~ data$Störche)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49298 -0.14636  0.01052  0.13789  0.49702
##
## Coefficients:
##              Estimate Std. Error t value
## (Intercept)  0.640069   0.045626   14.03
## data$Störche 0.071763   0.003065   23.41
##              Pr(>|t|)
## (Intercept)    <2e-16 ***
## data$Störche   <2e-16 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 0.1 ' ' 1
```

Lineare Regression

```
##
## Call:
## lm(formula = data$Geburten ~ data$Störche + data$Einwohner
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5057 -0.1191  0.0097  0.1210  0.4438
##
## Coefficients:
##                          Estimate Std. Error
## (Intercept)              3.304096   0.591504
## data$Störche            -0.010859   0.018512
## data$Einwohnerdichte    -0.006723   0.001489
##                         t value Pr(>|t|)
## (Intercept)               5.586 2.12e-07 ***
## data$Störche             -0.587    0.559
## data$Einwohnerdichte     -4.515 1.78e-05 ***
## ---
```

Datensatz

	ID	Land	GesAnteilGDP	leftseat	leftcab	GDP	chronic
26	1	Australia	8	53	100	1409795	462
57	2	Austria	10	42	53	294628	592
88	3	Belgium	10	32	21	365101	511
119	4	Canada	11	28	0	1662130	461
150	6	Czech Republic	7			3953651	823
181	7	Denmark	10	41	0	1798649	568
212	8	Estonia	6			14718	868
243	9	Finland	9	38	12	187100	534
274	10	France	11	40	0	1998481	408

Tabelle: Gesundheitsausgaben und weitere Merkmale in OECD-Ländern

Datensatz

Die Werte einzelner Variablen können durch die Eingabe des Datensatzes, gefolgt von einem "$" und dem Variablennamen abgerufen werden

```
data$GesAnteilGDP
```

```
##  [1]  8.4717 10.1123  9.9392 10.5875  6.9435
##  [6] 10.4464  6.3192  8.8801 10.7189 10.9956
## [11]  9.8523  7.5679  8.8203  7.0087  8.9536
## [16]  9.4921  6.4051  6.1541  7.1467  6.0200
## [21] 10.4324  9.6589  8.9103  6.4201  9.8195
## [26]  7.8231  8.5616  9.0136  8.4864 10.4590
## [31] 16.3918
```

Variablen

Für die Zuordnung der Werte zu dem Länder kann folgender Befehl genutzt werden

```
data[c("Land","GesAnteilGDP"), digits = 1]
```

	Land	GesAnteilGDP
26	Australia	8.5
57	Austria	10.1
88	Belgium	9.9
119	Canada	10.6
150	Czech Republic	6.9
181	Denmark	10.4
212	Estonia	6.3
243	Finland	8.9
274	France	10.7

Lineare Regression

Mit dem Befehl "lm()"kann eine lineare Regression berechnet werden.

```
lm(data$GesAnteilGDP~data$chronic +data$GDP)
```

Lineare Regression

```
##
## Call:
## lm(formula = data$GesAnteilGDP ~ data$chronic + data$GDP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0178 -0.8572  0.0051  0.7145  6.9916
##
## Coefficients:
##                Estimate Std. Error t value
## (Intercept)   1.275e+01  1.274e+00  10.002
## data$chronic -6.199e-03  2.049e-03  -3.026
## data$GDP     -2.736e-09  1.448e-09  -1.890
##                Pr(>|t|)
## (Intercept)  1.41e-10 ***
## data$chronic  0.00539 **
## data$GDP      0.06952 .
## ---
```

Logistische Regression

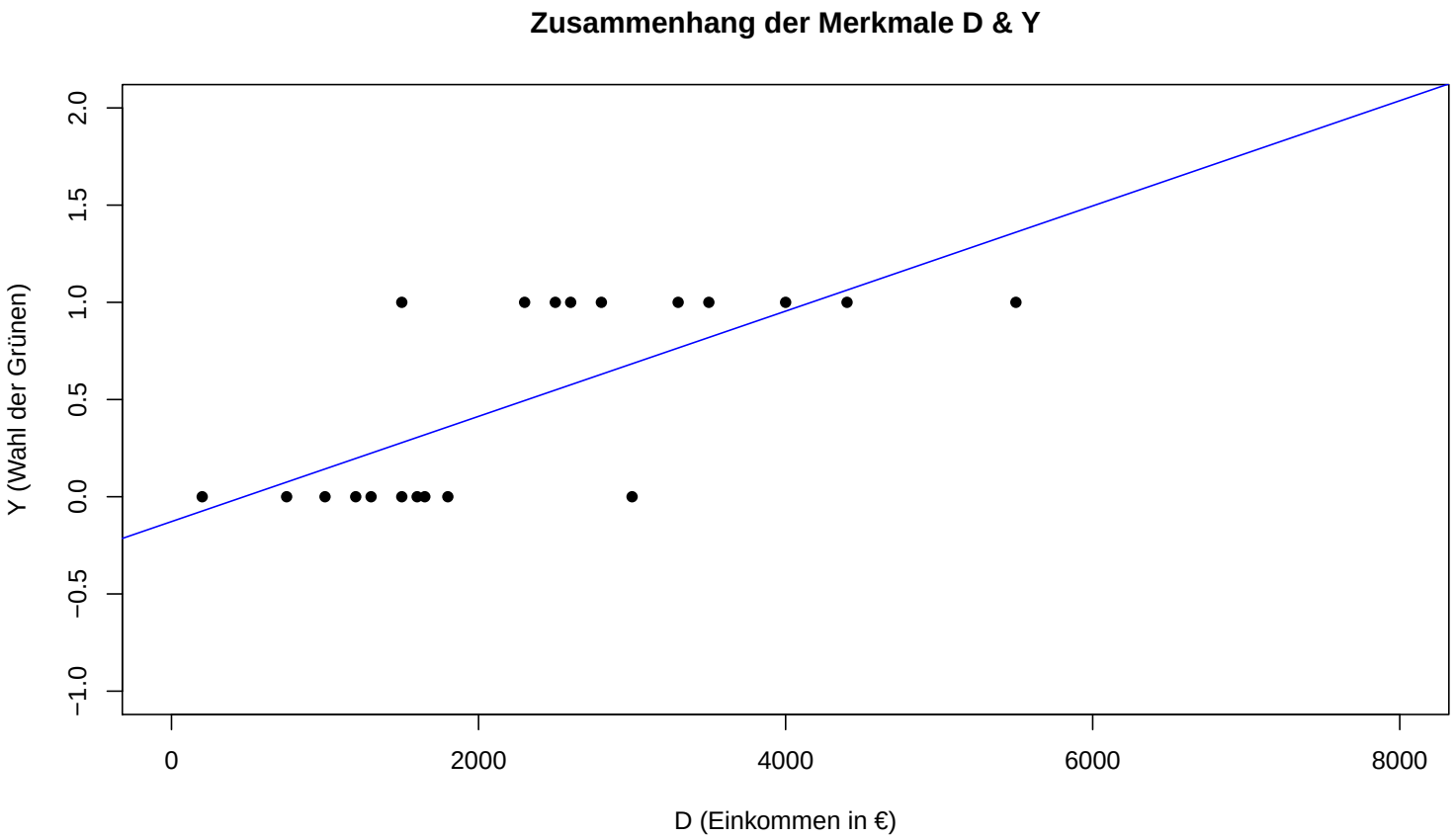
- andere Bezeichnung: Logit-Modell
- zumeist ist die binomiale logistische Regression gemeint
- hierbei ist abhängige Variable dichotom (meist Ausprägungen 0 und 1)
- eine weitere Form ist die multinominale logistische Regression (abhängige Variable: multinominal)

Warum keine lineare Regression (wenn Y binär)?

- Interpretation der Ergebnisse nicht sinnvoll
- Beispiel: Einfluss Einkommen(D) auf Wahl der Partei Die Grünen (Y)
- mögliches Ergebniss: y=0,001*D + 0,2
- Interpretation: Jeder zusätzliche Euro an Einkommen erhöht die Wahl der Partei Die Grünen um 0,001 (nicht sinnvoll!)

Weitere Probleme einer linearen Regression

- für dichotome Merkmale wie die Wahl der Grünen sind nur die Ausprägungen 0 & 1 sinnvoll
- in diesem Beispiel also Wahl der Gründen ja (y=1) oder nein (y=0)



Ziel der logistischen Regression

Schätzung der Eintrittswahrscheinlichkeit eines Ereignisses in Abhänigkeit verschiedener möglicher Einflussgrößen.

Beispielhafte Fragestellungen für logistische Regressionen

- erhöht das Einkommen die Wahrscheinlichkeit Die Grünen zu wählen?
- senkt der Schulabbruch ohne Abschluss die Wahrscheinlichkeit wählen zu gehen?
- haben Alleinerziehende eine höhere Wahrscheinlichkeit dauerhaft in ALGII zu verbleiben als Familien mit zwei Elternteilen?

Grundzüge der logistischen Regression

- abhängige Variable (Y): Wahl der Grünen
- Ausprägungen: Nein (y=0) & Ja (y=1)
- Wahrscheinlichkeiten: p(y=0) + p(y=1) = 1
- Abhängige Variable ist dichotom (meist Ausprägungen 0 und 1)

Logistische Funktion

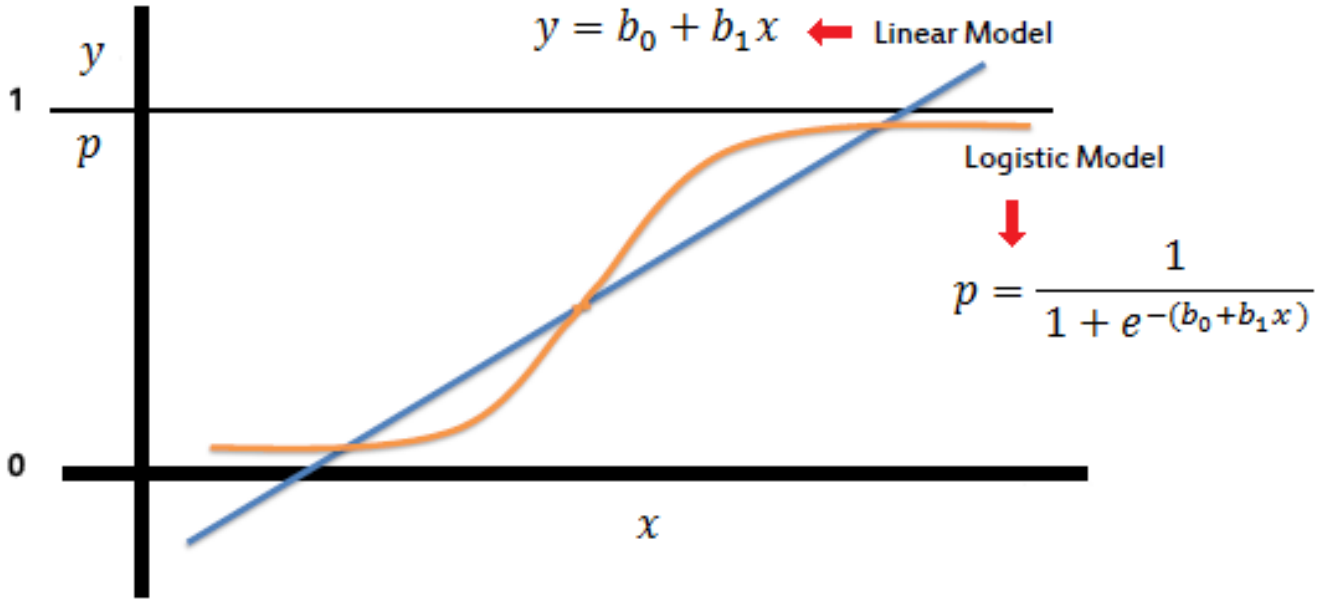


Abbildung: Vergleich logitischer und linearer Funktion

Odds Ratios

- in Statistikprogrammen können verschiede Ausgabeformen für Beta-Koeffizienten eingestellt werden
- die umfangreichsten Interpretationsmöglichkeiten bieten Odds Ratios
- Odds Ratios(Chancenverhältnis): Verhältnis von Eintrittswahrscheinlichkeit zur Wahrscheinlichkeit das Ereignis nicht Eintritt für verschiedene Ausprägungen der unabhängigen Variable

Odds Ratios

- Interpretation: OR=1,2; für ein um einen Euro höheres Einkommen wird eine um durchschittlich (1,2-1=0,2) 20% höhere Wahrscheinlichkeit der Wahl der Grünen geschätzt
- Interpretation: OR=0,8; für ein um einen Euro höheres Einkommen wird eine durchschittliche Wahrscheinlichkeit der Wahl der Grünen geschätzt die 80% der vorherigen Wahrscheinlichkeit entspricht

Annahmen & Voraussetzungen

- keine Multikollinearität (zwei oder mehr unabhängige Variablen korrelieren sehr stark miteinander)
- pro Ausprägung der abhängigen Variable mindestens 25 Beobachtungen
- aussagekräftige Schätzungen ab 100 Beobachtungen pro Gruppe