Angewandte Statistik

Julian Hinz — Universität Bielefeld

Session 3

Parametrische Regression

Poisson Regression

Update Tutorien

- Online-Teilnahme scheint gut zu funktionieren
- aber: Donnerstagnachmittag nur 4 Personen vor Ort
- “neues” Problem: Feiertage

Lernziel

Letzte Woche
- *Parametrische* Regression - Wiederholung Lineare Regression

Heute
- Poisson Regression

Fußballergebnisse

Nächstes Spiel: Dortmund - Leipzig

- Mit welcher Wahrscheinlichkeit gewinnt Dortmund dieses Spiel?
- Bekommt Dortmund den Champions League Platz?
- Sollten wir auf Dortmund wetten?

Poisson Regression

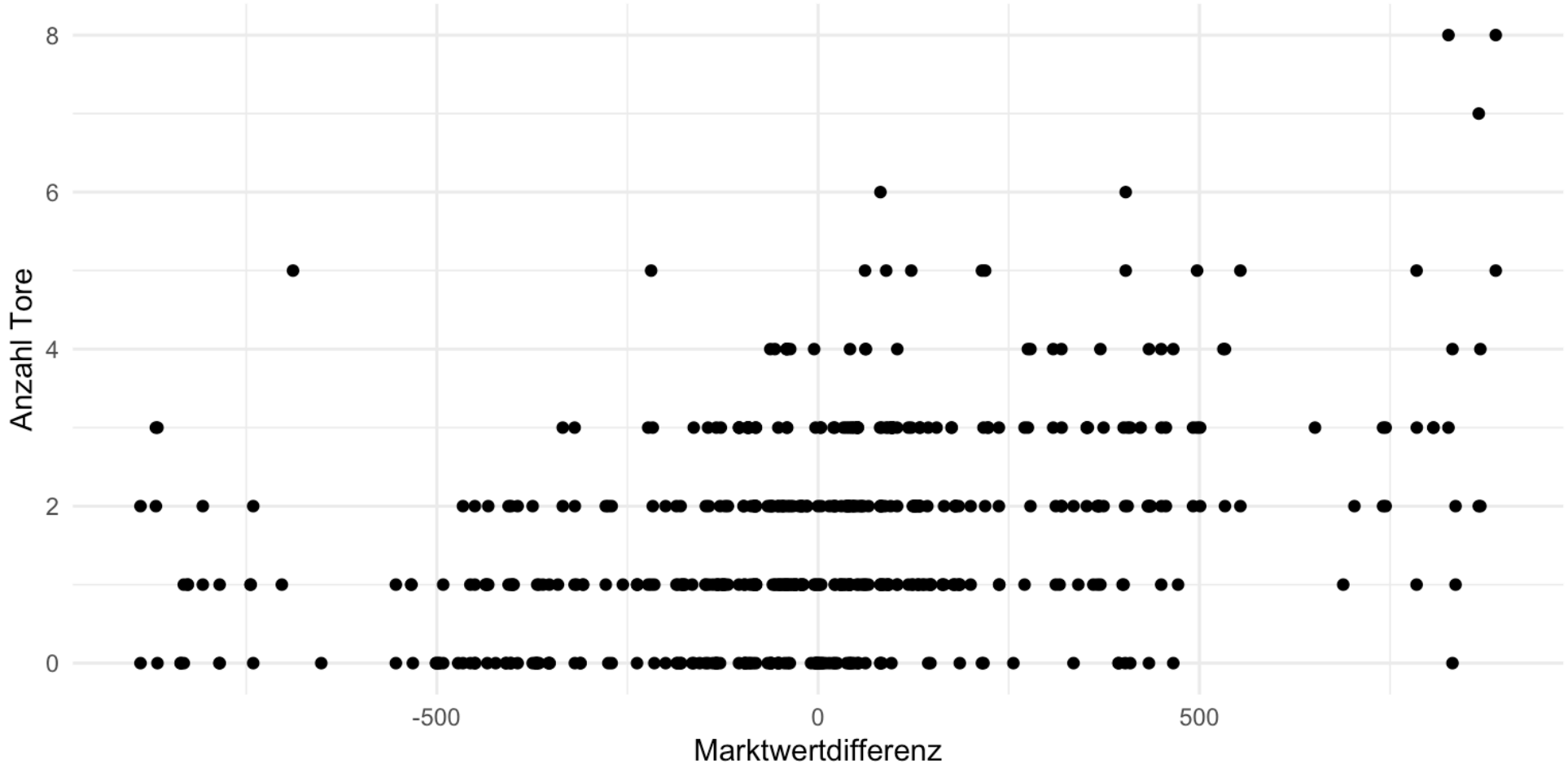
- **statistisches Modell** für **ganze Zahlen**
- hier: Anzahl geschossener Tore vorhersagen
- dann: daraus Gewinnwahrscheinlichkeit schätzen

Datenbasis

Team	Gegner	Spieltag	Tore	Heimspiel	Differenz Marktwert
1.FC Heidenheim	RB Leipzig	30	1	1	-435.85
1.FC Köln	Darmstadt 98	30	0	1	52.7
1.FSV Mainz 05	SC Freiburg	30	1	0	-85.8
B. Leverkusen	Bor. Dortmund	30	1	0	130.85
Bayern München	Union Berlin	30	5	0	784.8
Bor. Dortmund	B. Leverkusen	30	1	1	-130.85
…	…	…	…	…	…
Bor. M’gladbach	TSG Hoffenheim	30	3	0	40.88
Darmstadt 98	1.FC Köln	30	2	0	-52.7

- Bundesliga-Ergebnisse: zwei Datenpunkte pro Spiel
- Zuerst: Anzahl Tore als Funktion der Marktwertdifferenz

Anzahl Tore vs. Marktwertdifferenz



Statistische Modellierung durch Regression

Allgemeines Regressionsmodell (mit nur einer erklärenden Variable):

$$Y_i = f(x_i) + \epsilon_i, \qquad E(\epsilon_i) = 0, \qquad i = 1, \ldots$$

$$E(Y_i) = f(x_i), \qquad i = 1, \ldots$$

- Y_i: geschossene Tore (von einer Mannschaft in einem Spiel)
- x_i: Marktwertdifferenz (aus Sicht der betrachteten Mannschaft)

Ein solches Regressionsmodell könnte man dann benutzen, um die erwartete Anzahl geschossener Tore von Bielefeld bzw. Leipzig vorherzusagen.

Warum nicht einfach lineare Regression?

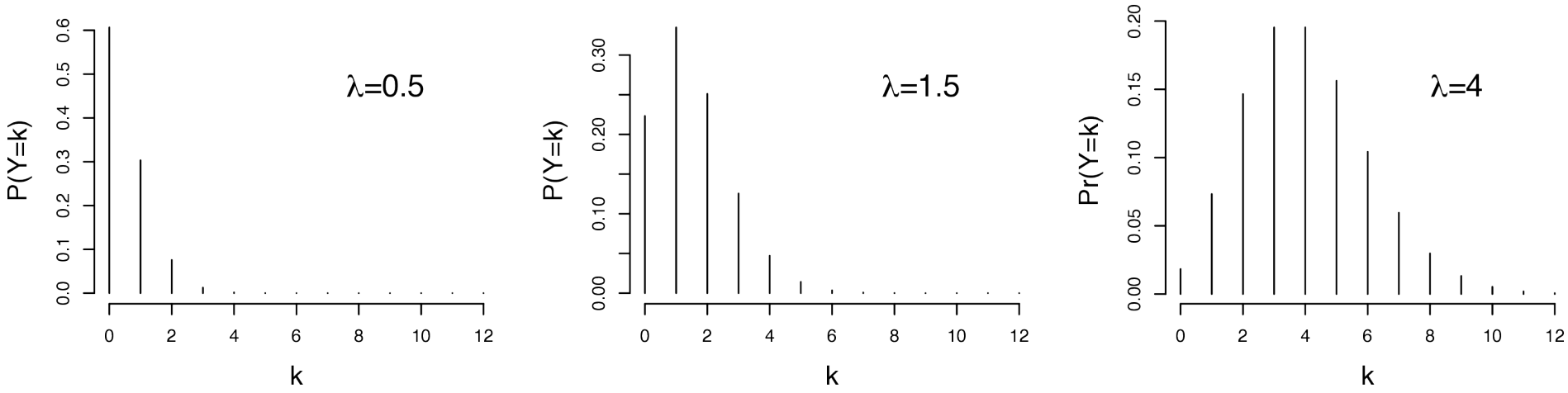
Wenig sinnvoll lineares Regressionsmodell zur Vorhersage:

- Unter der üblichen Annahme einer Normalverteilung der ϵ_i wäre die Vorhersageverteilung für die Anzahl geschossener Tore stetig
- Ein solches Modell könnte für bestimmte Werte von x_i sogar negative Werte für $E(Y_i)$ ergeben.

Gewünscht: ein Modell mit **sinnvoller Verteilungsannahme** und $E(Y_i) > 0$.

Verteilung der Zielgröße

- Zielvariable Y_i beschreibt Zähldaten: der Wertebereich ist $\{0, 1, 2, 3, \ldots\}$
- Zähldaten werden in der Regel mit der **Poisson–Verteilung** modelliert

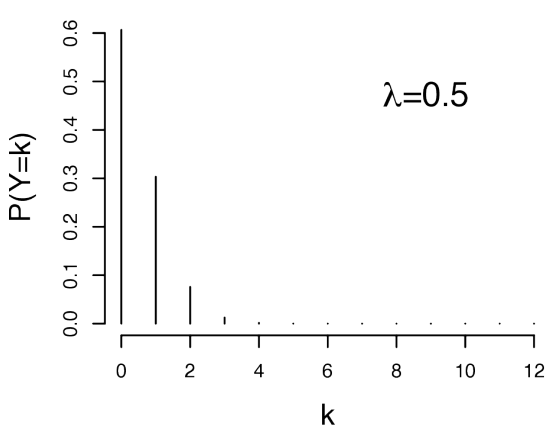


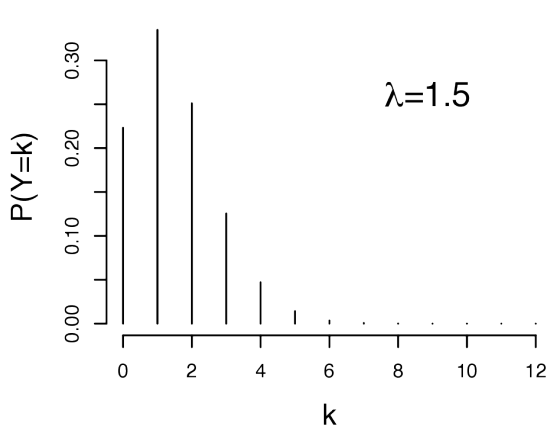
Poisson Verteilung

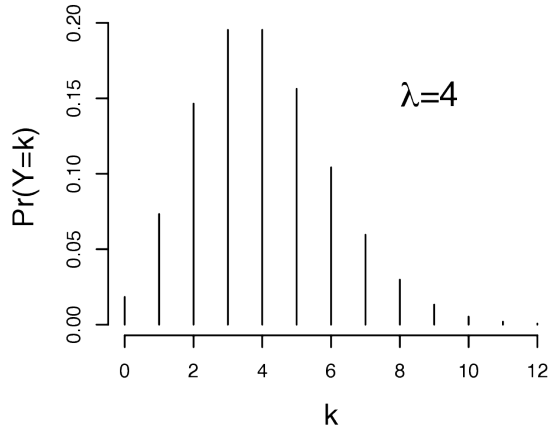
$Y \sim Po(\lambda)$, mit $\lambda > 0$, wenn

$$P(Y = k) = \frac{\lambda^k}{k!} \exp(-\lambda)$$

- Eigenschaften: $E(Y) = \lambda$ und $Var(Y) = \lambda$







Regression mit Poisson-verteilter Zielgröße

Folgendes Modell ist demnach zunächst einmal naheliegend:

$$Y_i \sim Po(\lambda_i), \qquad \text{wobei } \lambda_i = E(Y_i) = \beta_0 + \beta_1 x_i$$

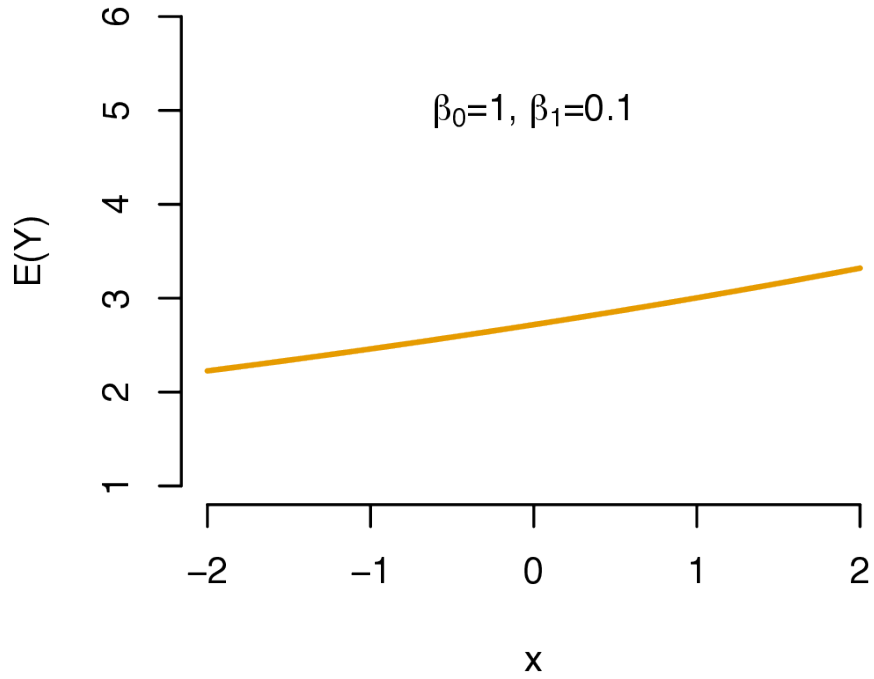
Problem: für manche x_i–Werte kann $\beta_0 + \beta_1 x_i$ negative Werte für λ_i ergeben

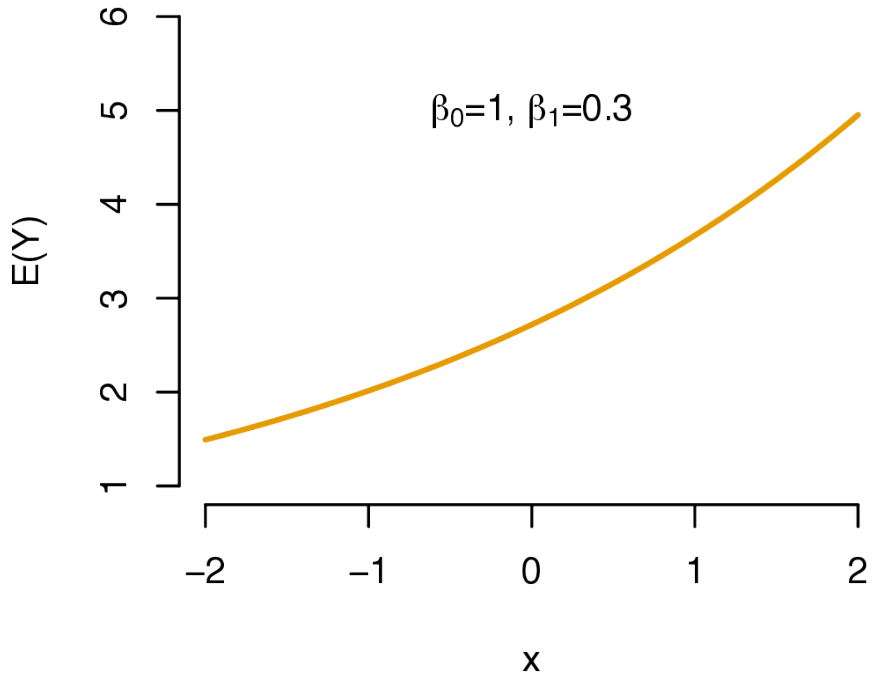
Regression mit Poisson-verteilter Zielgröße

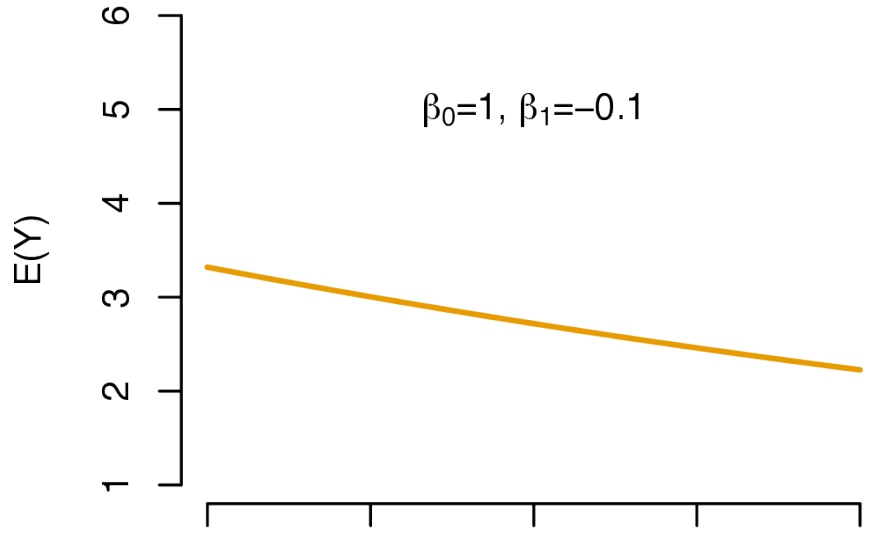
Betrachte stattdessen das Modell:

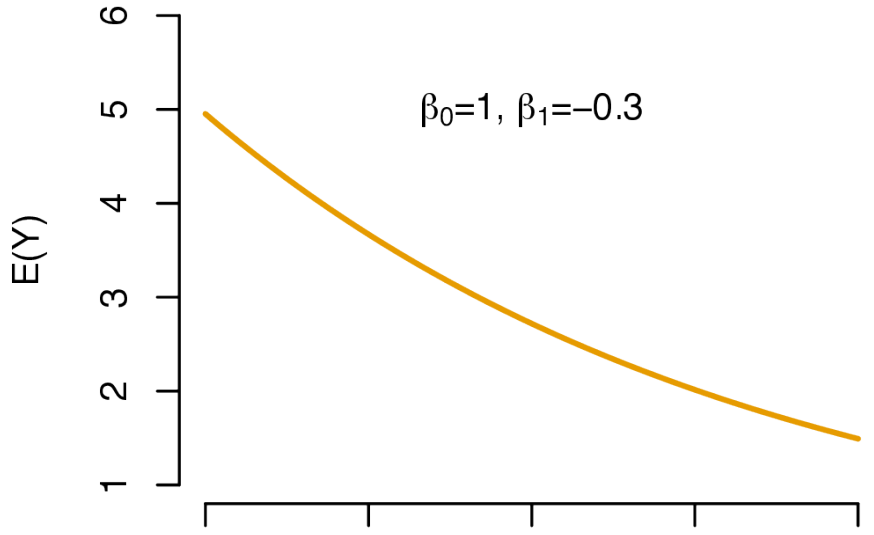
$$Y_i \sim Po(\lambda_i), \qquad \text{wobei } \lambda_i = E(Y_i) = e^{\beta_0 + \beta_1 x_i}$$

Lösung: Transformation durch Exponentialfunktion *garantiert*, dass λ_i positiv ist!


β_0=1, β_1=0.1
E(Y)
1
2
3
4
5
6
−2
−1
0
1
2
x


β_0=1, β_1=0.3
E(Y)
1
2
3
4
5
6
−2
−1
0
1
2
x


β_0=1, β_1=−0.1
E(Y)
1
2
3
4
5
6


β_0=1, β_1=−0.3
E(Y)
1
2
3
4
5
6

Poissonregression — Parameterschätzung

Wir wollen β_0 und β_1 schätzen. Ist Kleinste-Quadrate-Schätzung möglich?

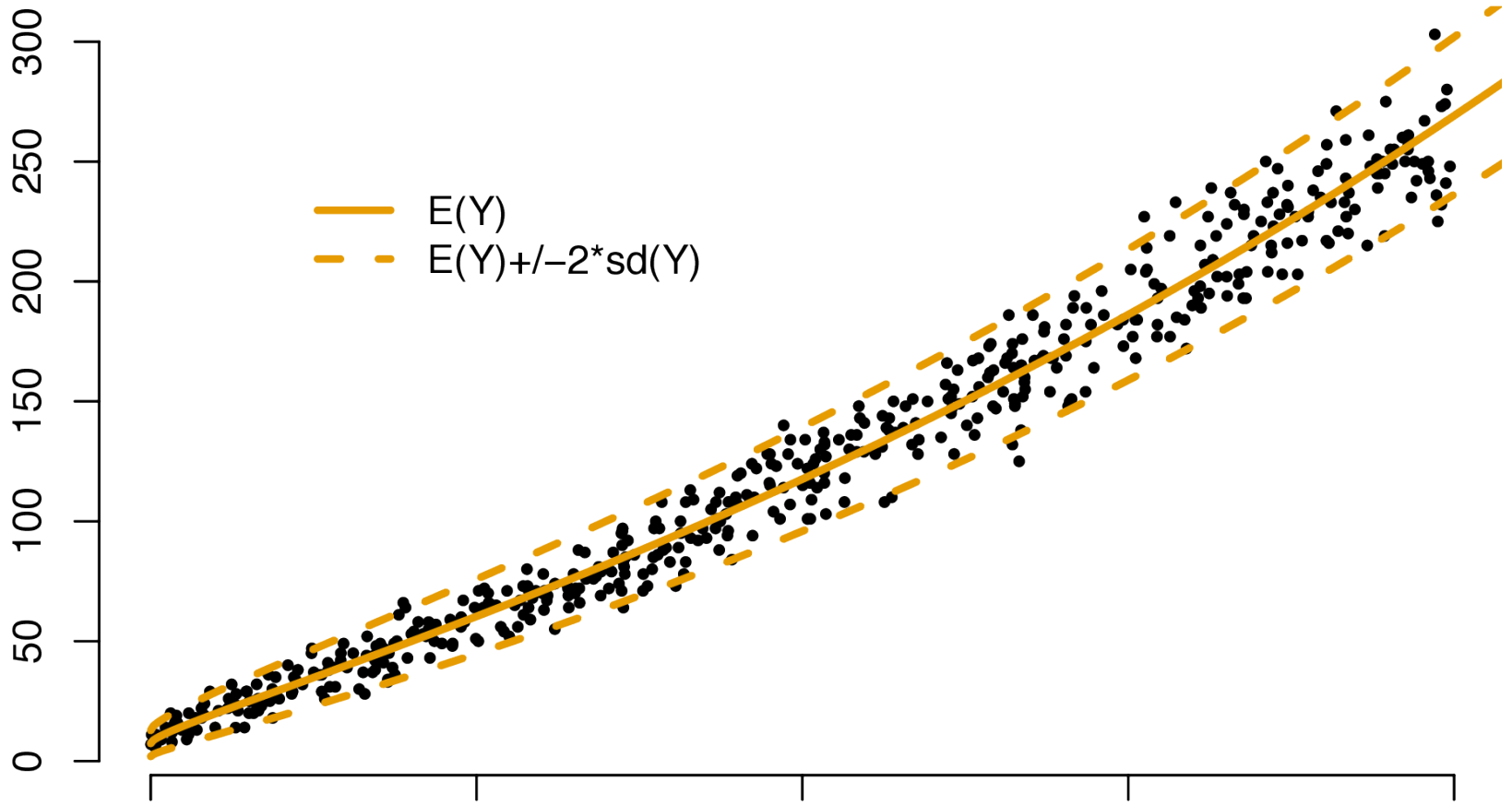
$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - e^{\beta_0+\beta_1 x_i} \right)^2$$

- Im Prinzip schon, aber: **Heteroskedastizität**
- bei Poissonregression:

$$\operatorname{Var}(Y_i) = E(Y_i) = e^{\beta_0+\beta_1 x_i}$$

und damit $\operatorname{Var}(Y_i) \neq \text{konstant}$

Poissonregression — Parameterschätzung



Poissonregression — Parameterschätzung

Wir minimieren daher die Summe der **gewichteten** quadrierten Fehler:

$$S_{\text{gewichtet}}(\beta_0, \beta_1) = \sum_{i=1}^{n} w_i \left(y_i - e^{\beta_0 + \beta_1 x_i} \right)^2.$$

- Gewichte sind $w_i = \frac{1}{\text{Var}(Y_i)}$
- Varianz der Y_i hängt jedoch von den unbekannten Parametern β_0 und β_1 ab — ein Teufelskreis

Poissonregression — Parameterschätzung

Methoden der iterativ gewichteten kleinsten Quadrate

1. Wähle/rate Startwerte für $\widehat{\beta}_0, \widehat{\beta}_1$
2. Bestimme die entsprechenden Gewichte, w_i
3. Minimiere $S_{\text{gewichtet}}(\beta_0, \beta_1)$, um verbesserte $\widehat{\beta}_0, \widehat{\beta}_1$ zu erhalten
4. Wiederhole 2. und 3. bis sich keine Änderung mehr ergibt

Good news: R erledigt das für uns, mit Hilfe der Funktion `glm()`.

Tore ~ MWdiff

```r
library(data.table)

daten = fread("../data/data3/bundesliga_vorhersage.csv")

regression = glm(Tore ~ Marktwert_Differenz,
                 data = daten,
                 family = poisson)
summary(regression)
```

```
Call:
glm(formula = Tore ~ Marktwert_Differenz, family = poisson, data = daten)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.4075213  0.0359620   11.33   <2e-16 ***
Marktwert_Differenz 0.0010162  0.0001011   10.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 671.82  on 539  degrees of freedom
Residual deviance: 572.93  on 538  degrees of freedom
  (72 observations deleted due to missingness)
AIC: 1638.2

Number of Fisher Scoring iterations: 5
```

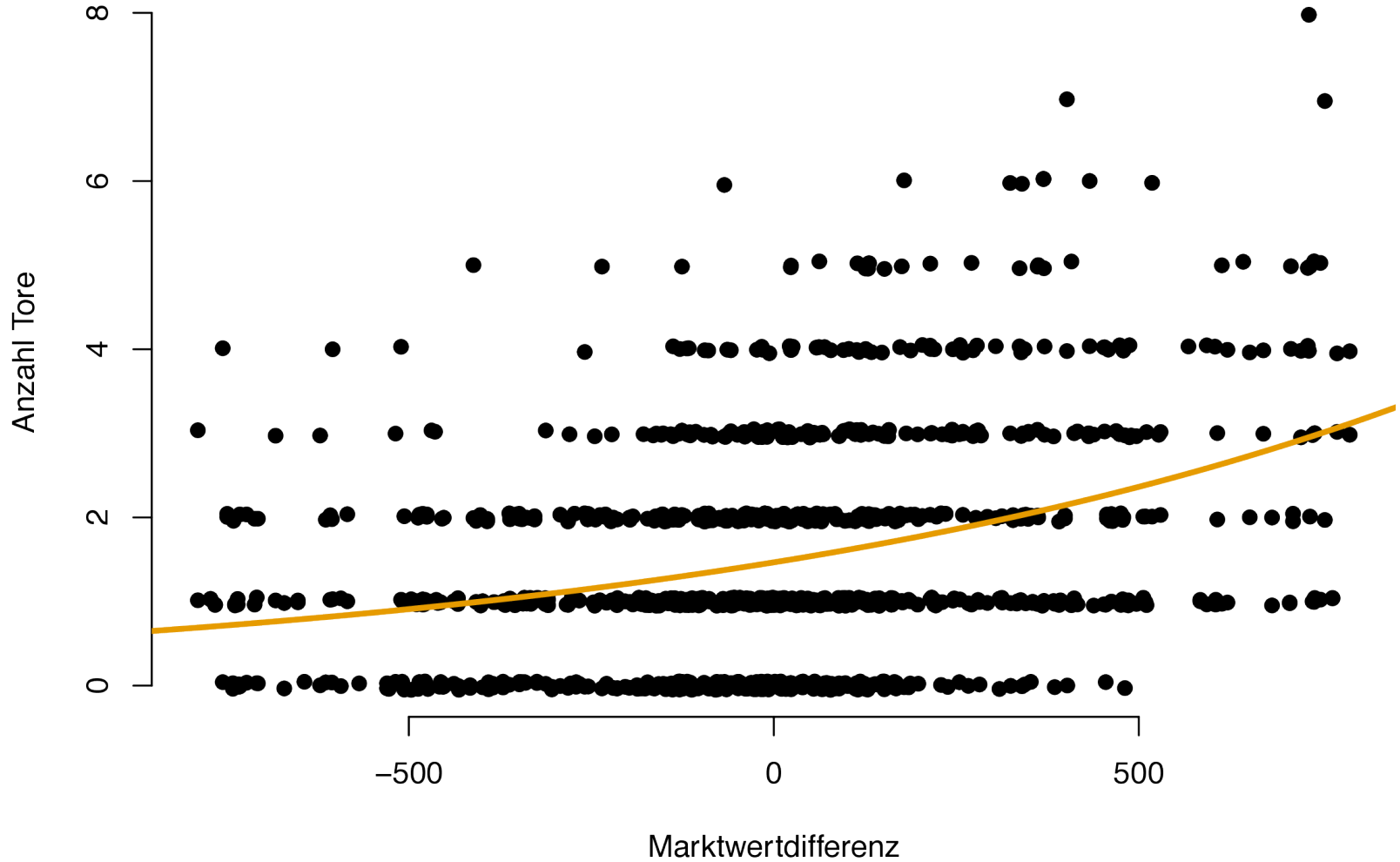
Tore ~ MWdiff

Folgendes Modell wurde hier geschätzt:

$$\text{Tore}_i \sim Po(\lambda_i), \quad \text{mit } \lambda_i = E(\text{Tore}_i) = e^{0.4075213+0.0010162 \cdot \text{Marktwert\_Differenz}_i}$$

bzw.:

$$\text{Tore}_i \sim Po(e^{0.4075213+0.0010162 \cdot \text{Marktwert\_Differenz}_i})$$


Anzahl Tore
0
2
4
6
8
−500
0
500
Marktwertdifferenz

Hypothesentest im Poissonregressionsmodell

Die im R–Output gegebenen p–Werte beziehen sich auf den Test von

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0.$$

Unter H_0, d.h. für $\beta_j = 0$, gilt

$$Z = \widehat{\beta}_j / \widehat{\sigma}_{\widehat{\beta}_j} \overset{\text{approx.}}{\sim} N(0, 1).$$

- handelt sich hier um einen **approximativen Gaußtest**
- also keinen t–Test wie beim linearen Modell

Allgemeine Poisson-Regression

Das **Poisson-Regressionsmodell** für unabhängige Zufallsvariablen Y_i mit dem Wertebereich $\{0, 1, 2, \ldots\}$ hat die Form

$$Y_i \sim Po(\lambda_i), \qquad \text{wo } \lambda_i = E(Y_i) = e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}, \qquad i = 1, \ldots, n.$$

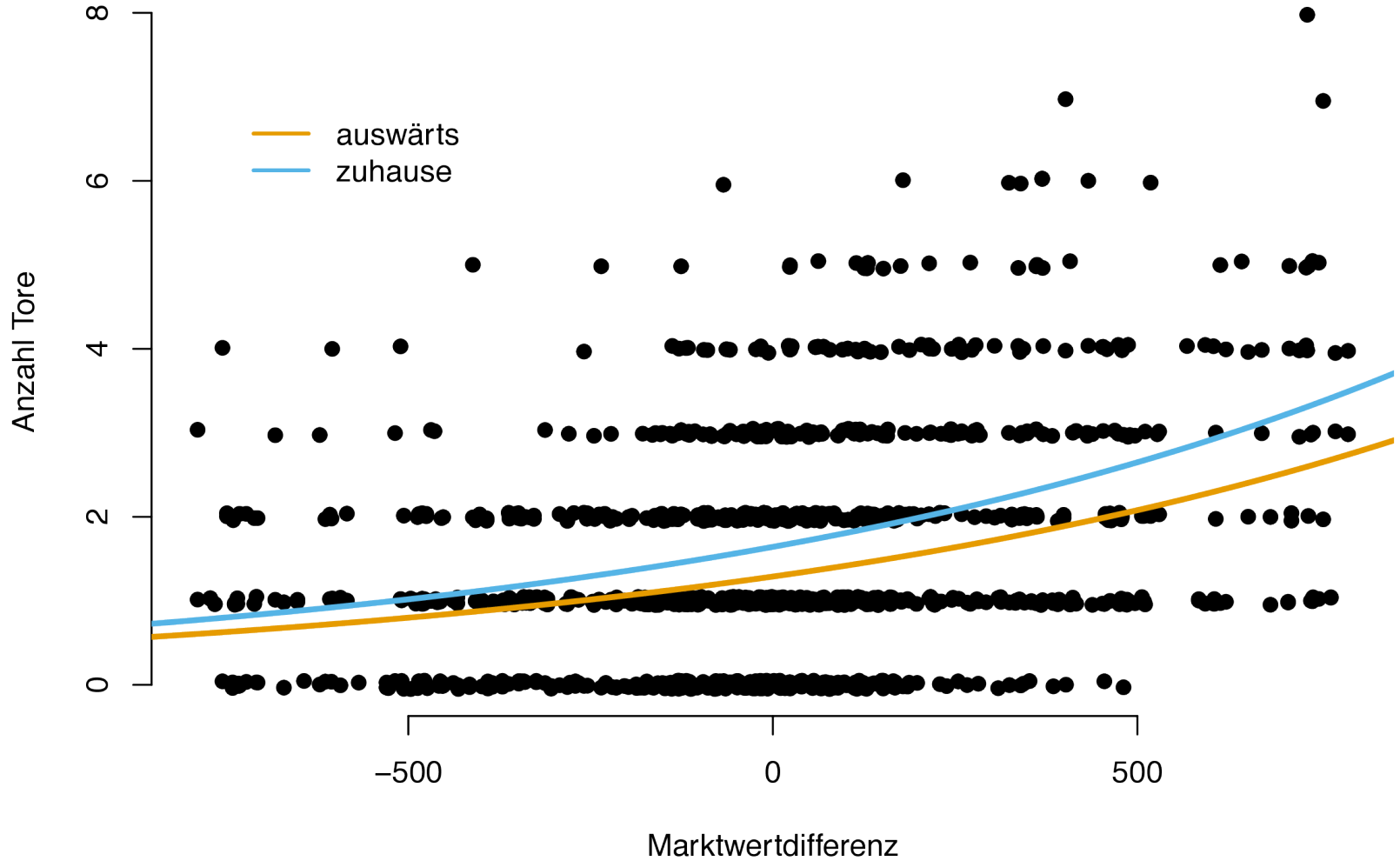
Bemerkungen:

- **linearer Prädiktor** $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ ebenso flexibel wie im linearen Modell in Bezug auf quadratische Terme, Interaktionen usw.
- Die Modellgleichung bestimmt nicht nur $E(Y_i)$, sondern auch $\mathrm{Var}(Y_i)$.
- Ein häufiges Problem bei der Analyse von echten Daten ist die **Überdispersion**, d.h., $\mathrm{Var}(Y_i) > E(Y_i)$, was das Poisson-Modell nicht zulässt.

Modell einschließlich Heimvorteil: Tore ~ MWdiff + Home

```r
library(data.table)

daten = fread("../data/data3/bundesliga_vorhersage.csv")

regression = glm(Tore ~ Marktwert_Differenz + Heimspiel,
                 data = daten,
                 family = poisson)
summary(regression)
```

```
Call:
glm(formula = Tore ~ Marktwert_Differenz + Heimspiel, family = poisson,
    data = daten)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.2745497  0.0528045   5.199    2e-07 ***
Marktwert_Differenz 0.0010161  0.0001011  10.054  < 2e-16 ***
Heimspiel           0.2503383  0.0687747   3.640 0.000273 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 671.82  on 539  degrees of freedom
Residual deviance: 559.57  on 537  degrees of freedom
  (72 observations deleted due to missingness)
AIC: 1626.9
```


auswärts
zuhause
Anzahl Tore
0
2
4
6
8
−500
0
500
Marktwertdifferenz

Interpretation der Modellparameter in der Poisson-Regression

Allgemeines Poisson-Regressionsmodell:

$$E(Y_i) = e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}$$

Angepasstes Modell im Fußballbeispiel:

$$E(\text{Tore}_i) = e^{0.254 + 0.001 \cdot \text{MWdiff}_i + 0.242 \cdot \text{Heim}_i}$$

Wie können wir zum Beispiel $\hat{\beta}_1 = 0.001$ interpretieren?

Zur Erinnerung: in der linearen Regression verändert das Modell

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip},$$

$E(Y_i)$ um "+β_1", wenn x_{i1} um 1 erhöht wird.

Erhöht man x_{i1} um 1 in Gleichung (2), ergibt sich:

$$e^{\beta_0+\beta_1(x_{i1}+1)+\ldots+\beta_p x_{ip}} = e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}+\beta_1} = e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}} \cdot e^{\beta_1}$$

also: Erhöhung von x_{i1} um 1 verändert $E(Y_i)$ um "e^{β_1}".

Im Fußballbeispiel:

- Erhöht sich MWdiff_i um 1, verändert sich $E(\mathrm{Tore}_i)$ um "$e^{0.001}$".
- $\mathrm{Heim}_i = 1$ verändert $E(\mathrm{Tore}_i)$ um "$e^{0.242}$" (verglichen mit $\mathrm{Heim}_i = 0$).

Anwendung des Modells auf Dortmund vs. Leipzig

Geschätztes Modell:

$$\text{Tore}_i \sim Po(\lambda_i), \qquad \lambda_i = E(\text{Tore}_i) = e^{0.254+0.001\cdot \text{MWdiff}_i+0.242\cdot \text{Heim}_i}$$

- Dortmund: Marktwert 63.55
- Leipzig: Marktwert 465.55

Daraus folgt:

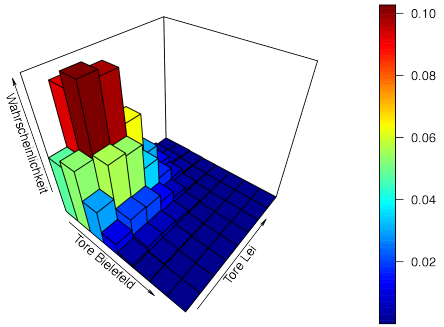
$$T_{Bie} = \text{Tore Dortmund} \sim Po\left(e^{0.254+0.001\cdot(-402)+0.242\cdot 1}\right) = Po(1.099)$$

$$T_{Lei} = \text{Tore Leipzig} \sim Po\left(e^{0.254+0.001\cdot 402+0.242\cdot 0}\right) = Po(1.927)$$

Prädiktive Verteilung unter dem Modell**

Unter der **Annahme der Unabhängigkeit** ist die Wahrscheinlichkeit eines spezifischen Spielausgangs wie folgt:

$$P(T_{BVB} = j, T_{Lei} = k) = P(T_{BVB} = j) \cdot P(T_{Leipzig} = k) = e^{-1.099} \frac{1.099^j}{j!} \cdot e^{-1.927} \frac{1.927^k}{k!}$$



Wahrscheinlichkeitsverteilung

In R: `dpois(j,1.099)*dpois(k,1.927)`

Further Applications of Poisson Regression

Poisson regression is useful when we want to model **count data** via regression.

Possible examples:

- Number of defects in a machine $\sim$ Machine running time.
- Number of cancer cases in a community $\sim$ Air pollution data.
- Number of sick days of an employee $\sim$ Overtime, salary, etc.
- Number of children of a woman $\sim$ Income.
- Number of traffic accidents on a day $\sim$ Weather conditions, day of the week.
- Number of doctor visits $\sim$ Financial incentives.