Kapitel 2

Parametrische Regression

2.1 Lineare Regression

Fragen? https://www.menti.com/alneo5om9sq5.

Analyse von (gerichteten) Abhängigkeiten zwischen Variablen

Wir wollen aus Daten ein Modell schätzen, welches den Einfluss der erklärenden Variablen $x_1, \ldots, x_p$ auf die Zielgröße Y gut beschreibt:

$$Y = f(x_1, \ldots, x_p) + \text{ zufälliger Fehler}$$

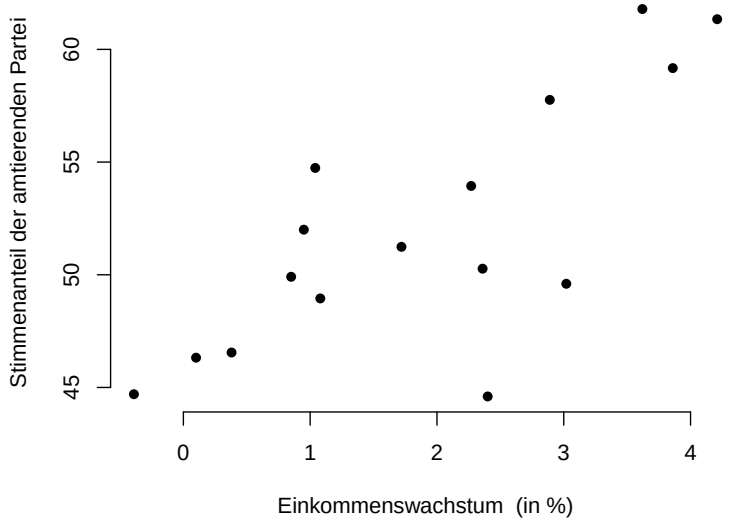
Beispiele:

- Verkaufszahlen eines Produkts hängen ab von Preis, Wochentag, Werbung, etc.
- Wettquote eines Fußballteams hängt ab von Stärke, Form, etc.
- Stimmanteil bei Wahl hängt ab von wirtschaftlicher Situation im Land, Partei, etc.
- Mietpreis hängt ab von Wohnfläche, Baujahr, Ausstattung, Lage, etc.

Im Idealfall erklärt f einen Großteil der Variabilität von Y, so dass der Fehlerterm klein ist $\rightsquigarrow$ dann sind z.B. ziemlich genaue Vorhersagen möglich.

Fragen? https://www.menti.com/alneo5om9sq5.

Beispiel: Einkommenswachstum und Wahlerfolg in den USA



Fragen? https://www.menti.com/alneo5om9sq5.

Modellformulierung

Das **lineare Regressionsmodell** hat die Form

$$\underbrace{Y_i}_{\text{Zielgröße}} = \underbrace{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}_{\text{systematischer Effekt}} + \underbrace{\epsilon_i}_{\text{zufälliger Fehler}}, \quad E(\epsilon_i) = 0, \; i = 1, \ldots, n$$

Hier wird eine spezielle lineare Form des systematischen Teils $f(x_{i1}, \ldots, x_{ip})$ gewählt. Warum ist diese Formulierung so wichtig?

- überraschend oft realistisch
- relativ robust gegenüber einzelnen Ausreißern
- sehr einfach umzusetzen und zu interpretieren
- viel flexibler als es auf den ersten Blick scheint

Fragen? https://www.menti.com/alneo5om9sq5.

Schätzung der Regressionskoeffizienten

Als “Gesamtabstand” zwischen Modell und Datenpunkten definieren wir die Summe der quadrierten Fehler:

$$S(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \Big)^2$$

Der **Kleinste-Quadrate-Schätzer** (KQS),

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \underset{\beta_0, \beta_1, \ldots, \beta_p}{\operatorname{argmin}} \; S(\beta_0, \beta_1, \ldots, \beta_p),$$

minimiert diesen Abstand.

Fragen? https://www.menti.com/alneo5om9sq5.

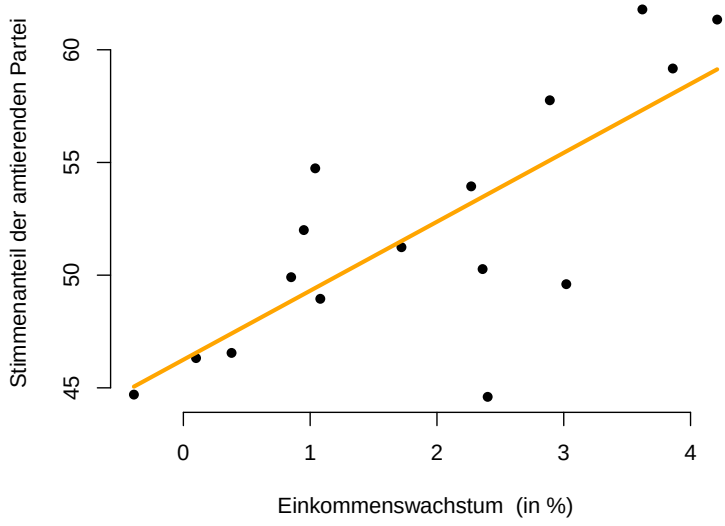
Lineare Regression in R

```r
USwahl_daten = read.table("https://tinyurl.com/ycx2kfmj", header = TRUE)
mod = lm(vote ~ growth, data = USwahl_daten)
summary(mod)
```

```
Call:
lm(formula = vote ~ growth, data = USwahl_daten)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9929 -0.6674  0.2556  2.3225  5.3094

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.2476     1.6219  28.514 8.41e-14 ***
growth        3.0605     0.6963   4.396  0.00061 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.763 on 14 degrees of freedom
Multiple R-squared:  0.5798,    Adjusted R-squared:  0.5498
F-statistic: 19.32 on 1 and 14 DF,  p-value: 0.00061
```

Fragen? https://www.menti.com/alneo5om9sq5.

Angepasstes Modell: $E(\text{Stimmenanteil}_i) = 42.248 + 3.061 \cdot \text{Einkommenswachstum}_i$



Fragen? https://www.menti.com/alneo5om9sq5.

Kapitel 2

Parametrische Regression

2.2 Poissonregression

Fragen? https://www.menti.com/alneo5om9sq5.

Motivation — Bielefeld vs. Dresden

- mit welcher Wahrscheinlichkeit gewinnt Bielefeld dieses Spiel?
- lohnt es sich, auf ein bestimmtes Ergebnis zu wetten?

Wir wollen ein **statistisches Modell** nutzen, um die Anzahl geschossener Tore einer Mannschaft vorherzusagen und daraus dann die Gewinnwahrscheinlichkeit zu schätzen.

Fragen? https://www.menti.com/alneo5om9sq5.

Daten

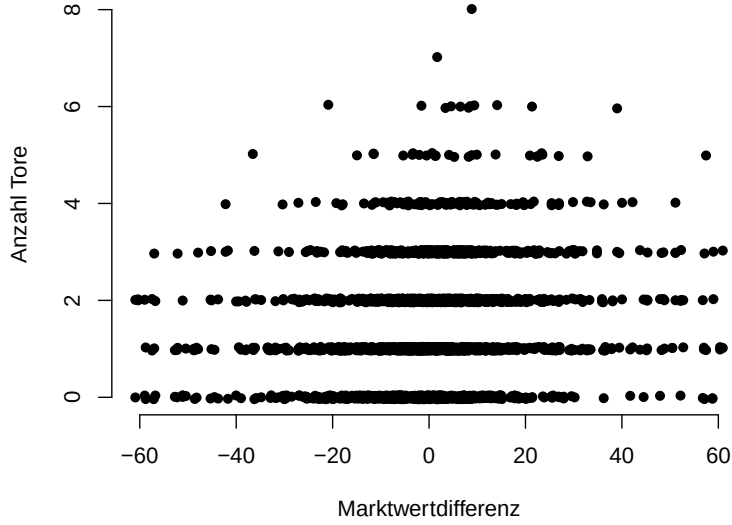
- Daten aus den Saisons 19/20, 20/21, 21/22, 22/23
- **zwei** Datenpunkte pro Spiel

Datum	Team	Gegner	MWDiff	Tore
26.07.2019	Stuttgart	Hannover	31.87	2
26.07.2019	Hannover	Stuttgart	-31.87	1
27.07.2019	Dresden	Nurnberg	-13.27	0
27.07.2019	Nurnberg	Dresden	13.27	1
27.07.2019	...	...	...	...
27.07.2019	...	...	...	...
14.05.2023	Paderborn	Heidenheim	-1.17	3
14.05.2023	Heidenheim	Paderborn	1.17	2
14.05.2023	Regensburg	Hamburg	-22.47	1
14.05.2023	Hamburg	Regensburg	22.47	5

Wir wollen zunächst die Anzahl Tore einer Mannschaft als Funktion der Marktwertdifferenz (“MWdiff”) aus Sicht der betrachteten Mannschaft modellieren.

Fragen? https://www.menti.com/alneo5om9sq5.

Anzahl Tore vs. Marktwertdifferenz



Fragen? https://www.menti.com/alneo5om9sq5.

Statistische Modellierung durch Regression

Allgemeines Regressionsmodell (mit nur einer erklärenden Variable):

$$Y_i = f(x_{i1}) + \epsilon_i, \quad E(\epsilon_i) = 0, \quad i = 1, \ldots, 2412$$

bzw. (äquivalente Schreibweise)

$$E(Y_i) = f(x_{i1}), \quad i = 1, \ldots, 2412$$

wobei hier

- Y_i: geschossene Tore (von einer Mannschaft in einem Spiel)
- x_{i1} : Marktwertdifferenz (aus Sicht der betrachteten Mannschaft)

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

Fragen? https://www.menti.com/alneo5om9sq5.

Warum nicht einfache lineare Regression?

Im Beispiel ist es wenig sinnvoll, ein lineares Regressionsmodell, d.h.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i \quad \text{bzw.} \quad E(Y_i) = \beta_0 + \beta_1 x_{i1}$$

zur Vorhersage zu nutzen, denn:

1. unter der üblichen Annahme einer Normalverteilung der ϵ_i wäre die Vorhersageverteilung für die Anzahl geschossener Tore stetig: $Y_i \sim N(\beta_0 + \beta_1 \cdot x_{i1}, \sigma^2)$
2. ein solches Modell könnte für bestimmte Werte von x_{i1} sogar negative Werte für $E(Y_i)$, d.h. für die erwartete Anzahl geschossener Tore, ergeben.

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

Fragen? https://www.menti.com/alneo5om9sq5.

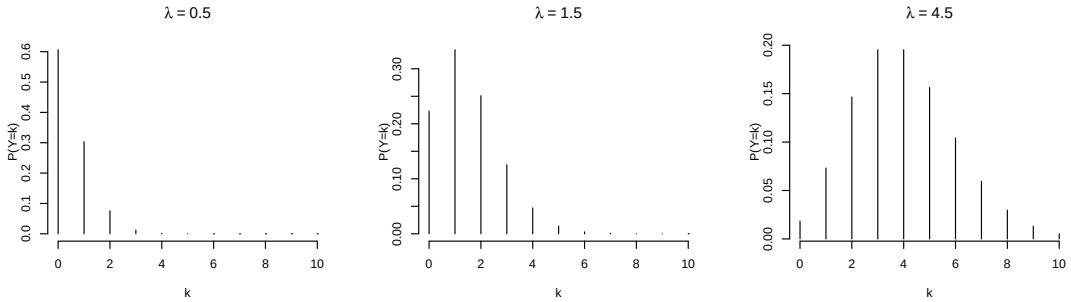
Verteilung der Zielgröße

Die Zielvariable Y_i beschreibt Zähldaten: der Wertebereich ist $\{0, 1, 2, 3, ...\}$.

Zähldaten können mit der **Poisson-Verteilung** modelliert werden.

Zur Erinnerung: Für $Y \sim Po(\lambda)$, mit $\lambda > 0$, ist die Wahrscheinlichkeitsfunktion

$$P(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad \big(E(Y) = \lambda, \text{Var}(Y) = \lambda\big)$$



Fragen? https://www.menti.com/alneo5om9sq5.

Regression mit Poisson-verteilter Zielgröße

Folgendes Modell ist demnach zunächst einmal naheliegend:

$$Y_i \sim Po(\lambda_i), \quad \text{wobei } \lambda_i = E(Y_i) = \beta_0 + \beta_1 x_{i1}$$

Problem: für manche x_{i1}-Werte kann $\beta_0 + \beta_1 x_{i1}$ negative Werte für λ_i ergeben.

Betrachte stattdessen das Modell

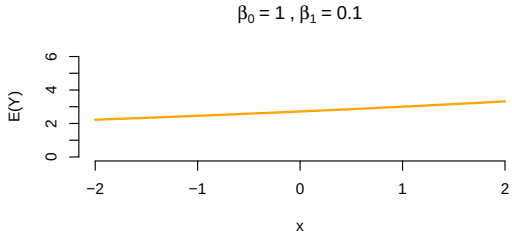
$$Y_i \sim Po(\lambda_i), \quad \text{wobei } \lambda_i = E(Y_i) = e^{\beta_0 + \beta_1 x_{i1}}$$

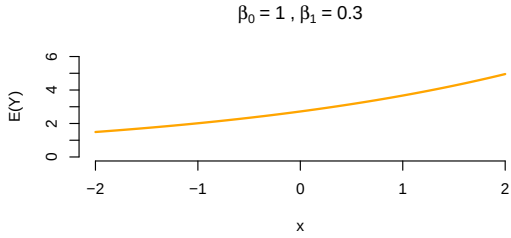
Die Transformation durch die Exponentialfunktion **garantiert**, dass λ_i **positiv ist**!

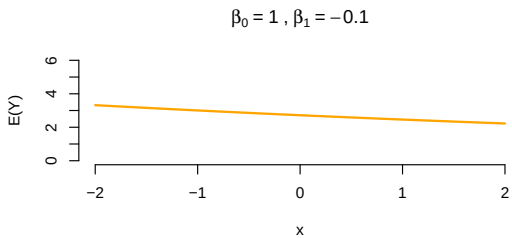
Fragen? https://www.menti.com/alneo5om9sq5.

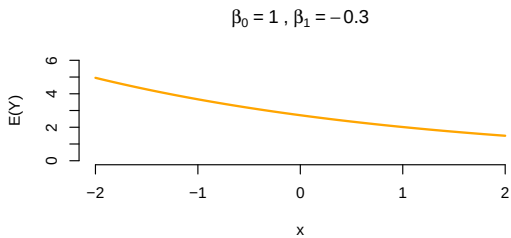
Verläufe der Regressionsfunktion

Mögliche Verläufe der Regressionsfunktion $e^{\beta_0+\beta_1 x}$ für verschiedene β_1:









Fragen? https://www.menti.com/alneo5om9sq5.

Poissonregression — Parameterschätzung

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

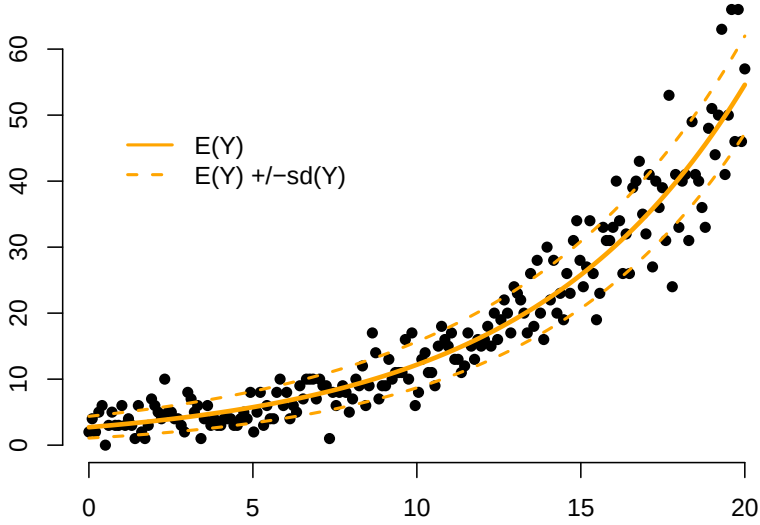
$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - e^{\beta_0 + \beta_1 x_{i1}} \right)^2$$

Im Prinzip schon, dieser Schätzer wäre wegen **Heteroskedastizität** aber nicht optimal. Bei der Poissonverteilung gilt $E(Y) = \text{Var}(Y)$, also:

$$\text{Var}(Y_i) = E(Y_i) = e^{\beta_0 + \beta_1 x_{i1}}$$

und damit ist $\text{Var}(Y_i)$ nicht konstant. Wenn $\text{Var}(\epsilon_i) > \text{Var}(\epsilon_j)$, dann sollte i-te Beobachtung weniger Gewicht erhalten.

Fragen? https://www.menti.com/alneo5om9sq5.



Fragen? https://www.menti.com/alneo5om9sq5.

Wir minimieren daher die Summe der gewichteten quadrierten Fehler,

$$S_{\text{gewichtet}}(\beta_0, \beta_1) = \sum_{i=1}^{n} w_i (y_i - e^{\beta_0 + \beta_1 x_{i1}})^2$$

Die Gewichte sind hierbei $w_i = \frac{1}{\text{Var}(Y_i)}$, die Varianz der Y_i hängt jedoch von den unbekannten Parametern β_0 und β_1 ab $\rightsquigarrow$ ein Teufelskreis...[1]

Methode der **iterativ gewichteten kleinsten Quadrate**:

1. wähle/rate Startwerte für $\hat{\beta}_0, \hat{\beta}_1$
2. bestimmte die entsprechenden Gewichte w_i
3. minimiere $S_{\text{gewichtet}}(\beta_0, \beta_1)$, um verbesserte $\hat{\beta}_0, \hat{\beta}_1$ zu erhalten
4. wiederhole Schritt 2. und 3. bis sich keine Änderung mehr ergibt.

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

[1] Bestimmung der β_j ist nur möglich, wenn die Varianz bekannt ist — und umgekehrt!

Fragen? https://www.menti.com/alneo5om9sq5.

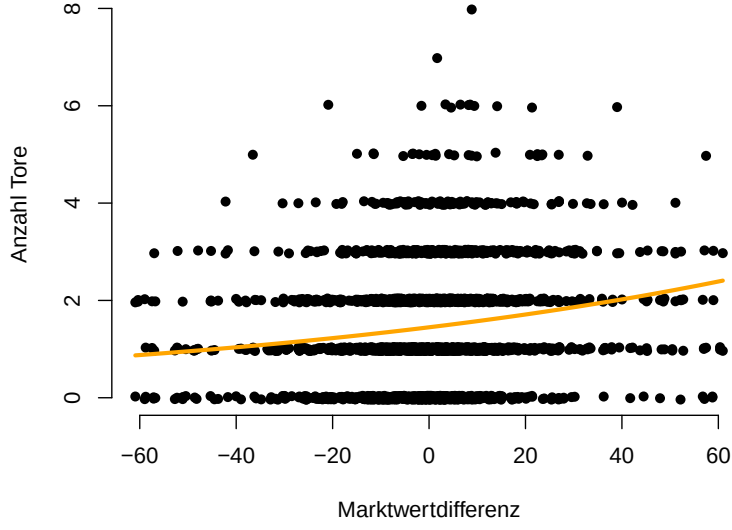
Poissonregression in R

```r
buli_daten = read.csv("https://tinyurl.com/5y2ss6sn")
mod = glm(Tore ~ MWDiff, family = "poisson", data = buli_daten)
summary(mod)
```

```
Call:
glm(formula = Tore ~ MWDiff, family = "poisson", data = buli_daten)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.370144   0.016999  21.774   <2e-16 ***
MWDiff      0.008342   0.001006   8.292   <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: 2878.1  on 2411  degrees of freedom
Residual deviance: 2810.2  on 2410  degrees of freedom
AIC: 7334

Number of Fisher Scoring iterations: 5
```

Fragen? https://www.menti.com/alneo5om9sq5.

Angepasstes Modell: $\text{Tore}_i \sim Po(e^{0.370+0.008 \cdot \text{MWdiff}_i})$



Fragen? https://www.menti.com/alneo5om9sq5.

Poissonregression allgemein

Das **Poissonregressionsmodell**, für unabhängige Zufallsvariablen Y_i jeweils mit Wertebereich $\{0, 1, 2, \ldots\}$, hat die Form

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

Bemerkungen:

- der **lineare Prädiktor** $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ ist ebenso flexibel wie im linearen Modell was quadratische Terme, Interaktionen etc. angeht
- die Modellgleichung legt nicht nur $E(Y_i)$, sondern auch $\text{Var}(Y_i)$ fest
- ein häufiges Problem bei der Analyse von Daten ist **Überdispersion**, d.h. $\text{Var}(Y_i) > E(Y_i)$, was das Poissonregressionsmodell nicht zulässt

Fragen? https://www.menti.com/alneo5om9sq5.

Modell inkl. Heimvorteil

```r
mod = glm(Tore ~ MWDiff + Heim, family = "poisson", data = buli_daten)
summary(mod)
```

```
Call:
glm(formula = Tore ~ MWDiff + Heim, family = "poisson", data = buli_daten)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.270002   0.025146  10.737  < 2e-16 ***
MWDiff      0.008342   0.001006   8.292  < 2e-16 ***
Heim        0.191163   0.033835   5.650 1.61e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2878.1  on 2411  degrees of freedom
Residual deviance: 2778.1  on 2409  degrees of freedom
AIC: 7303.9

Number of Fisher Scoring iterations: 5
```

Fragen? https://www.menti.com/alneo5om9sq5.

Poissonregression — Interpretation der Modellparameter

Allgemeines Poissonregressionsmodell:

$$Y_i \sim Po(\lambda_i), \quad \lambda_i = E(Y_i) = e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}$$

Im Fußballbeispiel angepasstes Modell:

$$\text{Tore}_i \sim Po(e^{0.270+0.008\cdot \text{MWDiff}_i+0.191\cdot \text{Heim}_i})$$

Wie kann man z.B. $\hat{\beta}_1 = 0.008$ interpretieren?

Fragen? https://www.menti.com/alneo5om9sq5.

Zur Erinnerung: bei linearer Regression, d.h. im Modell

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

ändert sich $E(Y_i)$ um β_1 bei Erhöhung von x_{i1} um 1.

Erhöhung von x_{i1} um 1 bei Poissonregression ergibt:

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

$\rightsquigarrow$ Erhöhung von x_{i1} um 1 ändert $E(Y_i)$ um "$\cdot e^{\beta_1}$".

Im Fußballbeispiel:

- Erhöhung von MWdiff_i um 1 ändert $E(\text{Tore}_i)$ um $\cdot e^{0.008}$ (also um etwa 0.8%)
- bei Heimspielen ($\text{Heim}_i = 1$) ändert sich $E(\text{Tore}_i)$ um $\cdot e^{0.191}$ (also um etwa 21.0%)

Fragen? https://www.menti.com/alneo5om9sq5.

Anwendung des Modells auf Bielefeld vs. Dresden

Geschätztes Modell:

$$\text{Tore}_i \sim Po(e^{0.270+0.008\cdot \text{MWDiff}_i+0.191\cdot \text{Heim}_i})$$

- Marktwert Bielefeld: 8.23
- Marktwert Dresden: 9.75

Somit:

$$T_{\text{BI}} \sim Po(\underbrace{e^{0.270+0.008\cdot(-1.52)+0.191\cdot 1}}_{=1.566})$$

$$T_{\text{DRE}} \sim Po(\underbrace{e^{0.270+0.008\cdot 1.52+0.191\cdot 0}}_{=1.326})$$

Fragen? https://www.menti.com/alneo5om9sq5.

Vorhersageverteilung unter dem Modell

Unter der **Annahme der Unabhängigkeit** ergibt sich die Wahrscheinlichkeit eines bestimmten Spielergebnisses wie folgt:

$$P(T_{\text{BI}} = j,\ T_{\text{DRE}} = k) = P(T_{\text{BI}} = j) \cdot P(T_{\text{DRE}} = k) = e^{-1.566}\frac{1.566^j}{j!} \cdot e^{-1.326}\frac{1.326^k}{k!}$$

$T_{\text{BI}} / T_{\text{DRE}}$	0	1	2	3	4	5
0	0.055	0.074	0.049	0.022	0.007	0.002
1	0.087	0.115	0.076	0.034	0.011	0.003
2	0.068	0.090	0.060	0.026	0.009	0.002
3	0.036	0.047	0.031	0.014	0.005	0.001
4	0.014	0.018	0.012	0.005	0.002	0.000
5	0.004	0.006	0.004	0.002	0.001	0.000

Fragen? https://www.menti.com/alneo5om9sq5.

Vorhersage — analytisch

$$
\begin{aligned}
\text{Wahrscheinlichkeit, dass Bielefeld gewinnt} &= \text{Summe aller Einträge unterhalb der Diagonalen} \\
&= 0.087 + 0.068 + \ldots + 0.001 \\
&= 0.425
\end{aligned}
$$

$$
\begin{aligned}
\text{Wahrscheinlichkeit für ein Unentschieden} &= \text{Summe aller Einträge auf der Diagonalen} \\
&= 0.055 + 0.115 + 0.060 + 0.014 + 0.002 \\
&= 0.246
\end{aligned}
$$

$$
\begin{aligned}
\text{Wahrscheinlichkeit, dass Dresden gewinnt} &= 1 - 0.425 - 0.246 \\
&= 0.329
\end{aligned}
$$

Fragen? https://www.menti.com/alneo5om9sq5.

Vorhersage — per Monte-Carlo-Simulation

- simuliere den Spielausgang sehr oft aus den beiden Poisson-Verteilungen
- dann entspricht die relative Häufigkeit des Eintretens eines Ereignisses approximativ der entsprechenden Wahrscheinlichkeit unter dem Modell

In R:

```
set.seed(123)
lambda_BI = 1.566
lambda_DRE = 1.326
mean(rpois(10^6, lambda_BI) > rpois(10^6, lambda_DRE))
```

```
[1] 0.42919
```

```
mean(rpois(10^6, lambda_BI) == rpois(10^6, lambda_DRE))
```

```
[1] 0.245963
```

```
mean(rpois(10^6, lambda_BI) < rpois(10^6, lambda_DRE))
```

```
[1] 0.323337
```

Fragen? https://www.menti.com/alneo5om9sq5.

Wetten

Würden Sie eine der folgenden Wetten eingehen?

Arminia Bielefeld
Dynamo Dresden

2,05	3,60	3,20

Fragen? https://www.menti.com/alneo5om9sq5.

Weitere Anwendungen von Poissonregression

Wir haben uns in diesem Abschnitt ausschließlich mit Fußballdaten beschäftigt.

Tatsächlich ist Poissonregression immer dann sinnvoll, wenn wir **Zähldaten modellieren** wollen.

Mögliche Beispiele:

- Anzahl Defekte einer Maschine $\sim$ Laufzeit der Maschine
- Anzahl Krebsfälle in einer Kommune $\sim$ Luftverschmutzungsdaten
- Anzahl Krankheitstage eines Arbeitnehmers $\sim$ Überstunden, Gehalt, etc.
- Anzahl Kinder $\sim$ Einkommen
- Anzahl Verkehrsunfälle an einem Tag $\sim$ Blitzermarathon, Wetterlage, Wochentag

Nächste Woche: logistische Regression für binäre Daten.

Fragen? https://www.menti.com/alneo5om9sq5.