Angewandte Statistik II

Dr. Uli Wannek

Skript erstellt von Alina Renz

Sommersemester 2018

Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Inhaltsverzeichnis

1 Lineare Modelle

1.1 Zufallsvariable Y

- Verteilung, Erwartungswert, Varianz, Form (Schiefe, Kurtosis,...)
- Parameter der Verteilung (μ, σ), (λ),...
  - Punktschätzer $(\widehat{\mu})$, $(\widehat{\theta})$, ...
  - Konfidenzintervall
- Zusätzlich abhängig von einer Variablen X:

$$\begin{aligned} \mathcal{E}(Y_i) &= \mu_i \\ Y_i &\sim \mathcal{N}(\mu_i, \sigma^2) \end{aligned}$$

  - mit der linearen Abhängigkeit

$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

- Ausprägungen
  - nominal, z.B. rot/grün/blau; f/m; Städte
  - ordinal, z.B. kein/etwas/viel; Schulabschluss
  - kardinal/metrisch, z.B. Dosis, Stimulusintensität, -Abstand, -Anzahl
  - speziell dichotom, z.B. ja/nein; klein/groß; 0/1

1.2 Einfaches Lineares Modell

$$Y = \beta_0 + \beta_1 X$$

- abhängige Variable: Zufallsvariable Y
  - (mehrfache) Messung/Realisierung, ergibt Wert y_i
  - *response*
  - fehlerbehaftet
- unabhängige Variable X
  - mit Wert x_i, vom Experimentator vorgegeben, '*control*'
  - mit Wert x_i, fest, mitgemessen, '*covariate*'
  - Vorhersageparameter '*predictor*'

- Linearer Zusammenhang
  - kausale Abhängigkeit Y von X
  - Proportionalitätsfaktor β_1
  - y-Achsenabschnitt β_0 '*intercept*'
- Streuung zulassen

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Konventionen

Schrift	Bedeutung	Beispiel
Großbuchstaben	Zufallsvariable	Y
Kleinbuchstaben	Realisierung einer Zufallsvariale, Messwert	x_i, y
fett	Vektor oder Matrix	$\mathbf{X}, \mathbf{y}, \boldsymbol{\epsilon}$
Griechisch	Parameter	β, μ
Hut $\hat{}$	Schätzer	$\widehat{\beta}_0$
Index $_i$	Index für Werte	x_i
Index $_j$	Index für Parameter	β_j
Index $^{(m)}$	Index für Iteration	$b^{(m+1)}$

- Lineares Modell - Matrix Schreibweise
  - Seien Y_i *i.i.d.* Zufallsvariablen mit normalverteilter Streuung ϵ

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i & \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \\ \mathcal{E}(Y_i) = \mu_i &= \beta_0 + \beta_1 X_i & Y_i &\sim \mathcal{N}(\mu_i, \sigma^2) \end{aligned}$$

  - n-malige *unabhängige, identische* Wiederholung des Versuchs
    - Messtupel (X_i, Y_i) mit $i \in [1 \dots n]$
    - Erlaubte Streuung in Y_i
  - Abhängige Variable Y

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

  - Parametervektor β

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

    - bestimmt die Modell-Abhängigkeit $y_i \sim x_i$
  - unabhängige Variable X
    - Vektor $\mathbf{X} = [X_1, X_2, \dots, X_n]^T$
    - erweitert um den y-Achsenabschnitt *intercept*

  - Vektor $\mathbf{1} = [1, 1, \ldots, 1]^T$

- $\Rightarrow$ Designmatrix $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

  * unabhängige Variablen in Spalten
  * Indikator- (Pseudo-) Variable für unabhängige Kategorien

- $\Rightarrow$ Lineares Modell

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\mathcal{E}(\mathbf{Y}) = \mathbf{X} \quad \boldsymbol{\beta}$$

$$\mathcal{E}(Y_i) = 1 \cdot \beta_0 + X_i \cdot \beta_1$$

- ϵ Streuungen in y
  * Messfehler
  * Ungenauigkeiten
  * Residuen: Abweichungen vom Modell

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \epsilon$$

$$\mathcal{E}(\mathbf{y}) = \mathbf{X}\,\boldsymbol{\beta}$$

- Gesucht: Parameter des Modells $\boldsymbol{\beta}$
- Lösung dieser Aufgabe:

  mittels Anpassen der Parameter durch iterative Anwendung von Matrixinversion aus Maximum-Likelihood-Prinzip / Kleinste-Quadrate-Schätzung

- Ergebnis: Parametervektor β
  * Punktschätzer $\widehat{\beta}$ mit Konfidenzintervall
  * *Signifikanz*

1.3 Additives und Interaktives Lineares Modell

- Additives Lineares Modell
  - k unabhängige Variablen X_j als Spalten der Länge n in die **Designmatrix** einfügen

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & & \ddots & & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}$$

  - den **Parametervektor** erweitern zu

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

  - ergibt das additive Lineare Modell

$$\mathcal{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

- Interaktives Lineares Modell
  - sind die unabhängigen Variablen X_l und X_m untereinander unabhängig, dann ist

$$x_{io} = x_{il} \cdot x_{im}$$

    eine *neue* unabhängige Variable und kann als Spalte der Designmatrix hinzugefügt werden
  - Interaktion: Beeinflussung von X_l auf X_m
  - Designmatrix mit zusätzlichem Interaktions-Term

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,k} & x_{1,k+1} = x_{1,l} \cdot x_{1,m} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,k} & x_{n,k+1} = x_{n,l} \cdot x_{n,m} \end{bmatrix}$$

  - Schätzung der Parameter analog

$$\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_0,\ \widehat{\beta}_1, \dots \widehat{\beta}_k,\ \widehat{\beta}_{lm}]^T$$

  - Formelbeschreibung in `patsy` beispielweise
    - `'y ~ x1: x2'`: beinhaltet eine Spalte mit Term $x1 * x2$ in Designmatrix
    - `'y ~ x1 * x2 + x3'`: Abkürzung für: Spalte mit Termen 1, x1, x2, x1*x2 und x3

1.4 Kennwerte Linearer Modelle

- Einzelne Messwerte
$$Y_i = 1\beta_0 + X_{i1}\beta_1 + \cdots + X_{ik}\beta_k + \epsilon_i$$
  - mit Zufall/Streuung/Rauschen ("Homoskedastizitätsannahme", (Residuen-) Varianzhomogenität)
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$
  - Dann
$$\begin{aligned} \mathcal{E}(Y_i) &= \beta_0 + X_{i1}\beta_1 + \cdots + X_{ik}\beta_k \\ Var(Y_i) &= \sigma^2 \end{aligned}$$

- vektoriell
  - Erwartungswert
$$\mathcal{E}(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$
  - Varianz-Kovarianz-Matrix
$$\mathbf{V}_{jk} = \mathcal{E}\Big((Y_j - \mu_j) \cdot (Y_k - \mu_k)\Big)$$
    - im unabhängigen Fall
$$\begin{aligned} Var(Y_j) &= V_{jj} = \sigma_j^2 \\ Cov(Y_j, Y_k) &= V_{jk} = 0 \quad \text{für } k \neq j \end{aligned}$$
    - im i.i.d.-Fall
$$Var(Y_j) = V_{jj} = \sigma^2$$
    - Definition:
$$Cov(Y_j, Y_k) = \mathcal{E}\Big((Y_j - \mathcal{E}(Y_j)) \cdot (Y_k - \mathcal{E}(Y_k))\Big)$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y) \cdot (x - \mathcal{E}(X)) \cdot (y - \mathcal{E}(Y))\, \mathrm{d}y\, \mathrm{d}x$$
    - daraus folgt im unabhängigen Fall (siehe oben):
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_X(x) \cdot f_Y(y) \cdot (x - \mathcal{E}(X)) \cdot (y - \mathcal{E}(Y))\, \mathrm{d}y\, \mathrm{d}x = 0 \quad q.e.d.$$

1.5 Generalisierte Lineare Modelle

- Lineares Modell

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\,\boldsymbol{\beta} + \epsilon \\ \mathcal{E}(\mathbf{y}) &= \mathbf{X}\,\boldsymbol{\beta} \end{aligned}$$

- Generalisiertes Lineares Modell mit Link-Funktion g

$$\begin{aligned} \mathcal{E}(\mathbf{Y}) &= \boldsymbol{\mu} \\ g(\boldsymbol{\mu}) &= \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \end{aligned}$$

  - insbesondere hilfreich mit
    - kategorialer *abhängiger* Variable
    - dichotomer *abhängiger* Variable
- Spezialfall
  - Link-Funktion **Identität**

$$\eta = g(\mu) = \mu$$

  - Streuung **Normalverteilung**

$$f(\mathbf{Y}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^2}{2\sigma^2}\right)$$

  - Dann ergibt sich

$$\begin{aligned} \mathcal{E}(\mathbf{Y}) &= \mathbf{X}\boldsymbol{\beta} \\ \mathrm{Var}(\mathbf{Y}) &= \sigma^2 \end{aligned}$$

    ... das (einfache) **Lineare Modell**
- Fragestellungen
  - Das Modell ist festgelegt
    - Theorie
    - Erfahrung
    - Vorversuch
  - Die Modell-Parameter
    - sind unbekannt
    - oder dienen der Überprüfung einer Theorie
    - gilt es, aus Messungen von X_i und Y_i zu bestimmen
  - Schlussfolgerung
    - Ist Y von X abhängig? (Signifikanz)
    - Ist die Abhängigkeit stärker unter Versuchsbedingung A als unter B? (Vergleich)

1.6 Fragestellung

- Ziel: Parameter $\boldsymbol{\beta}$
- Anpassung (fit) des Linearen Modells, so dass die Residuen minimiert werden.
- Erinnerung: Homoskedastizitätsannahme der Normalverteilten Residuen.
  - Summe der Abweichungsbeträge L_1

$$S_1(\mathbf{y}, \widehat{\mathbf{y}}) = \sum_{i=1}^{n} |y_i - \widehat{y}_i|$$

  - Element der maximalen Aweichung L_∞

$$S_\infty(\mathbf{y}, \widehat{\mathbf{y}}) = max_i(|y_i - \widehat{y}_i|)$$

  - Euklidische Abstandsquadratsumme L_2

$$S_2(\mathbf{y}, \widehat{\mathbf{y}}) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

  - Euklidische Norm: $||\mathbf{z}|| = \sqrt{S_2(\mathbf{z}, \mathbf{0})} = \sqrt{\mathbf{z}^T\mathbf{z}} = \sqrt{\sum_{i=1}^{n} z_i^2}$
  - Quadratfehlersumme

$$\text{RSS} = S_2(\mathbf{y}, \widehat{\mathbf{y}}) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

    * Wird verwendet, wenn Gauß'sche Fehler vorhanden sind
- Gauß-Markov-Theorem
  - L_2 liefert die kleinste Varianz zu einem erwartungstreuen (*unbiased*) linearen Schätzer
  - Voraussetzung:
    * unabhängige Parameter
    * Fehler *i.i.d.* (*independently identically distributed*)
  - Nicht zwingend hier:
    * Normalverteilung

1.7 Lösung der Aufgabe

Lösung 1: Kleinste Quadrate Schätzer

- Für das Lineare Modell
$$\widehat{\mathbf{y}} = \mathcal{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$
- Speziell: Ausgleichsgerade
$$\widehat{\mathbf{y}} = \mathcal{E}(\mathbf{Y}) = \beta_0 + \beta_1\mathbf{x}$$
  - Ansatz
$$S_2(\mathbf{y}, \widehat{\mathbf{y}}) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i))^2 \stackrel{!}{=} min_{\beta_0,\beta_1}$$
  - führt dank einfacher Rechnung zu
$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$$
  - Residuenvarianz (bereits zwei Werte geschätzt, reduziert Freiheitsgrade)
$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{\epsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

Lösung 2: Matrix-Ansatz

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Minimieren der Fehlerquadratsumme
$$S_2(\mathbf{y}, \widehat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \stackrel{!}{=} min_{\boldsymbol{\beta}}$$
  - führt zu
$$\mathbf{X}^T\,\mathbf{y} = \mathbf{X}^T\,\mathbf{X}\,\boldsymbol{\beta}$$
  - mit Lösung
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\,\mathbf{X})^{-1}\mathbf{X}^T\,\mathbf{y}$$
- Mit Gewichtung
  - Minimieren der Fehlerquadratsumme mit reziprok gewichteten Varianzen
$$S_2(\mathbf{y}, \widehat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \stackrel{!}{=} min_{\boldsymbol{\beta}}$$
  - (Varianz-Kovarianz-Matrix $\mathbf{V}$; $\mathbf{V}_{jk} = Cov(Y_j, Y_k)$) führt zu
$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}$$

  - mit Lösung
$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

- Gilt für beliebige Dimensionen
  - hier mit 2x2 Matrix einfach
- Höherdimensional möglich, nur technisch schwer.
  - Dann iterativ zu bestimmen
- Numerisch instabil mit Kovarianzen
- Unlösbar oder stark fehlerbehaftet durch Gleitkommafehler
  - wenn unterbestimmt durch unglückliche Verteilung der Fehler
  - zu wenig Freiheitsgrade
- Implementiert in Python `statsmodels.ols`:
  - `pinv`: Moore-Penrose pseudoinverse
  - `qr`: Q-R-Zerlegung

Lösung 3: Maximum Likelihood Schätzer

- Ansatz über Verbund-Wahrscheinlichkeitsverteilung $f_{\boldsymbol{\theta}}(\mathbf{y})$ = Likelihood $L_{\mathbf{y}}(\boldsymbol{\theta})$
$$L(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} f(y_i|\boldsymbol{\theta})$$
  - Daraus Log-Likelihood
$$l(\boldsymbol{\theta}|\mathbf{y}) := \log L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{N} \log f(y_i|\boldsymbol{\theta})$$
  - zu maximieren
$$l(\widehat{\boldsymbol{\theta}}) \stackrel{!}{=} max_{\boldsymbol{\theta}}$$

- Beispiel Normalverteilung
  - Lineares Modell $\quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\mu} = \mathcal{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$
  - Normalverteilung $\quad f(y_i|\,\mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right)$
  - Parametervektor $\quad \boldsymbol{\theta} = [\beta_0, \beta_1, \sigma]^T$
  - Log-Likelihood:
$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log f(y_i|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right) \\ &= -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\Big(y_i - (\beta_0 + \beta_1 x_i)\Big)^2 \end{aligned}$$

  - Maximieren der Log-Likelihood führt zum Parametervektor-Schätzer $\widehat{\boldsymbol{\theta}} = [\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}]^T$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$
$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

Vergleich der Lösungen

- Kleinste-Quadrate-Methode
  - Minimieren S_2 der Residuen
  - Findet *Kleinste-Quadrate-Schätzer (least square, LSE)* für Parameter
- Max-Likelihood-Methode
  - Maximiert Log-Likelihood
  - Findet *Max-Likelihood-Schätzer (MLE)* für Parameter
- Meist das selbe Ergenis
  - Bei Normalverteilung identisch

Anwendungsbeispiel: log(Gehirnmasse) $\sim$ log(Körpermasse)

- Designmatrix
  - Zeilen:
    - Daten der einzelnen Tiere (i)
  - Spalten:
    - unabhängige Variable 'Körpergewicht'
    - Konstante für den y-Achsenabschnitt (*intercept*) β_0
- Designmatrix mit `numpy`: `np.vstack((np.ones(len(x1)), x1)).T`
- Berechne den Punktschätzer des Parametervektors aus Designmatrix und Datenvektor

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\,\mathbf{X})^{-1}\mathbf{X}^T\,\mathbf{y}$$

1.8 Praktische Lösung mittels Python `statsmodels`

- Homepage: `http://www.statsmodels.org/stable/`
- Beschreibung
  - `GLS` = Generalized least squares regression
  - `OLS` = Ordinary least square regression
  - `GLM` = Generalized linear models
    - fit = `smf.glm(formula='log_BrainWt ~ log_BodyWt', data=animalsdata).fit()`
    - Ergebnis/Ausgabe:
      - Parametervektorschätzer
      - Standardabweichung
      - z-Wert der Gauß-Statistik
      - p-Wert dazu
      - 95%-Konfidenzintervall
- Daten interpolieren, extrapolieren
  - Modell an die Daten anpassen (fit) ergibt den Parameter-Schätzer

$$\widehat{\boldsymbol{\beta}}$$

  - Der vorhergesagte Wert $\widehat{\mathbf{y}}$ ist

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

$$\widehat{y}_i = (\mathbf{X}\widehat{\boldsymbol{\beta}})_i = \sum_{j=0}^{m} x_{ij}\beta_j = 1\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{im}\beta_m$$

1.8.1 Python `statsmodels`

- `statsmodels.formula.glm.fit()` beschreibt ein **lineares Datenmodell**
  - Eingabe Datensatz `data =`
    - `pandas.DataFrame` mit *Variablennamen*
    - unabhängige Variablen bzw. Designmatrix
    - abhängigen Variablen
  - Eingabe Modell `formula=`
    - `patsy`-Formel mit abhängiger Variable ~ unabhängiger Variablen
    - `'y ~ x1 + x2 + x3'`
    - berücksichtigt bereits die Konstantenspalte der Designmatrix *intercept*

            - ⇒ explizit ausschließen `'~ -1'`

- `statsmodels.GLM.fit()`
  - Eingabe Daten
    - `exog`: unabhängige Variablen in Spalten der Designmatrix X
      - zusätzlich Konstante *intercept* anfügen `sm.add_constant(X)`
      - Bei *Interaktion* sind zusätzliche Spalten zu berechnen
    - `endog`: abhängige Variable, gemessene Daten y
- `statsmodels.___.fit()`
  - Ausgabe Parametervektor
    - Punktschätzer
      - Standardabweichung
      - Vertrauensintervall
      - Z-Wert der Gauß-Statistik
      - p-Wert
  - Ausgabe Statistiken und Kennzahlen
    - ...
  - Ausgabe Fit-Werte
    - `fittedvalues`: (als pandas-Daten-`Series`)
    - `resid_response`: verbleibende Fehler `(Series)`
    - `predict`(x): Zwischenwerte vorhersagen/extrapolieren
      - x als `DataFrame` mit passend benannten Spalten

1.8.2 Python `Pandas`

- Python `Pandas` für Umgang mit Daten
  - Homepage: `http://pandas.pydata.org/pandas-docs/stable/overview.html`
  - Daten aus Datei einlesen `read_csv()`
  - Variable vom Typ `DataFrame`
    - Auswahl der in Spalten enthaltenen Variablen durch Namensstring
    - Auswahl nach Kriterien, Index, Eigenschaften, ...
    - Umfangreiche Methoden
      - sortieren `sort()`
  - Beispiel: Abhängigkeit von Körpergewicht und Hirngewicht

    - Lösung? Zufällige Abweichungen zwischen Messung y_i und Modell-Vorhersage $\widehat{y}_i$
    - Residuen
$$r_i = y_i - \widehat{y}_i$$

1.8.3 Python `Patsy`

- Designmatrix mit `patsy`
  - Homepage: `http://patsy.readthedocs.io/en/latest/overview.html`
  - Patsy erlaubt Formulierung
    - des Modells
    - der zu benutzenden Daten
  - Eingabe:
    - `y, X = patsy.dmatrices('yvar ~ xvar1 + xvar2', df)`
    - verwendet *pandas* `DataFrame df`
  - Ausgabe
    - Designmatrix x als `patsy.design_info.DesignMatrix,` $N * K$ Array, mit y-Achsenabschnittskonstante
    - Gemessene Daten y als `patsy.design_info.DesignMatrix,` $N * 1$ Array
  - Generelle Form: Innerhalb eines Strings $y \sim x$
    - links der Tilde die abhängige Variable
    - rechts die unabhängige Variablen
  - Anschaulich lassen sich die Namen der Datenfelder aus dem `DataFrame` benutzen

1.9 Ergebnis lineare Modelle in Python

- Daten lassen sich in `DataFrames` komfortabel bearbeiten
- lassen sich durch `Patsy`-Formel beschreiben
- Schätzer für Parameter lassen sich durch `statsmodels.glm` berechnen
- Rückgabewerte:
  - Kennzahlen
  - Statistik
  - Punktschätzer für Parameter (Steigung und Achsenabschnitt) und deren
  - Intervallschätzer
  - ...

1.10 Bestes Modell?

- Ein perfekt passendes Modell muss nicht das beste sein
- Gleiche Versuchsbedingung, identische Zeile in Designmatrix:
  Streuung in $\mu_{i_1} = \mu_{i_2} = \ldots$
- $\Rightarrow$ Fehler zulassen

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

- Theorie
- Ockham's razor

Verdichtung der Information

- Nicht von Interesse: alle einzelnen μ_i der abhängigen Variablen Y
- Von Interesse:
  - Einfluss der unabhängigen Variablen (*erklärende* Variablen, Pediktoren) X
    - kategorial
    - kontinuierlich
    - Versuchsbedingungen $i \quad i \in [1 \ldots n]$
  - zugehörige Parameter
    - modellieren X, *Gewichtung* der Einflüsse
    - Parameter $\beta_j \quad j \in [1 \ldots k] \quad k \ll n$

= das Modell

Ergebnis

- Modell = Entscheidung für Vereinfachung
- Es verbleiben Residuen

Residuen

- Verteilung der Residuen

$$\begin{aligned} Y_i &= \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i & \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \\ \mathcal{E}(Y_i) = \mu_i &= \mathbf{X}_i^T \boldsymbol{\beta} & Y_i &\sim \mathcal{N}(\mu_i, \sigma^2) \end{aligned}$$

- Anforderung an Residuen
  - Modell soll gut abbilden, 'in der Mitte' $\Rightarrow$ $\mathcal{E}(R) = 0$
  - Streuung in Verteilung hat dieselben Ursachen
    - *Lineares Modell*, *Gauß*- Verteilung: $\Rightarrow$ $Var(R) = const.$
    - Gemäß Verteilung
  - Gutes Modell erklärt Messdaten
    - Keine (wenig) Information in den Residuen:
      $\Rightarrow$ **unabhängig**, homoskedastisch
- Homoskedastizität und Unabhängigkeit
  - Systematische Abweichungen? $\Rightarrow$ Auf den Grund gehen!

1.11 Modell-Vergleich

- Quadratfehlersumme, *sum of squared residua, RSS*

$$RSS = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - (\mathbf{X}\widehat{\boldsymbol{\beta}})_i)^2$$

  - Ist eine charakteristische Kennzahl
    - Für Gauß-Verteilungen: standardisierte Quadratfehlersumme $\tilde{S} = \frac{RSS}{\sigma^2}$
    - $\tilde{S} \sim \chi^2(n-p)$
  - Abhängigkeit nur von
    - n Werten der *abhängigen* Variablen
    - n Werten der *unabhängigen* Variablen
    - p geschätzte Parameterwerte
  - je kleiner RSS, desto näher liegt das Modell an den Daten
- Schätzer für $\boldsymbol{\beta}$
  - $\widehat{\boldsymbol{\beta}}$ aus Max-Likelihood oder Kleinste-Quadtrate (k Komponenten)
- Schätzer für $\boldsymbol{\mu}$
  - $\widehat{\mu}_i = \mathbf{X}_i^T \widehat{\boldsymbol{\beta}}$ aus dem linearen Modell
- Schätzer für *Störparameter* σ^2
  - Seien y_i Normalverteilt (mindestens näherungsweise; Zentraler Grenzwertsatz) dann ist mit

$$RSS = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} (y_i - (\mathbf{X}\widehat{\boldsymbol{\beta}})_i)^2$$

$$\widehat{\sigma}^2 = \frac{1}{N-p} RSS$$

  - ein erwartungstreuer Schätzer der Varianz σ^2 für das Lineare Modell

$$\mathcal{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \qquad Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$\widehat{\sigma}^2 = \frac{1}{N-p} \sum_{i=1}^{N} r_i^2 = \frac{1}{N-p} \sum_{i=1}^{N} (y_i - (\mathbf{X}\widehat{\boldsymbol{\beta}})_i)^2$$

- Verteilung der standardisierten Fehlerquadratsumme

$$\frac{RSS}{\sigma^2} \sim \chi^2(N-p)$$

  - Die Verteilung der Zufallsvariable *Schätzer der Residuen-Varianz* $\widehat{\sigma}^2$ ist dann skaliert:

$$\widehat{\sigma}^2 \sim \chi^2(\texttt{df} = N-p,\ \texttt{scale} = \frac{\sigma^2}{N})$$

  - ... unter der Nullhypothese, dass das Modell korrekt ist!
- Problem 1: Woher kennen wir das wahre σ^2?
- Problem 2: Was ergibt die Berechnung mit dem Schätzer?
- Vergleich der beiden Modelle
  - Voraussetzung: Modelle bauen aufeinander auf, Modell B ist eine Erweiterung/Verallgemeinerung des einfacheren Modells A
  - Ist Modell B (hier $p_B = 3$ Parameter) angemessen?
    * Nein $\Rightarrow$ beide Modelle verwerfen
    * Ja $\Rightarrow$ vergleiche mit Modell A
  - Ist Modell A (hier $p_A = 2$ Parameter) angemessen?
    * Nein $\Rightarrow$ wähle Modell B
    * Ja $\Rightarrow$ Vergleich mit Modell B ergibt ...

Wiederholung Tests

1. Formulierung des Problems
2. Modellannahme
   - Welcher Art sind die Daten
   - Welche Verteilung wird erwartet
3. Aufstellen der Nullhypothese und der Alternativhypothese
   - Ziel soll es sein, die Nullhypothese ablehnen zu können
   - einseitiger Test
   - zweiseitiger Test
4. Festlegen des Signifikanzniveaus
   - zulässige Irrtumswahrscheinlichkeit α
5. Teststatistik / Prüfgröße aussuchen
   - verdichtet Information aus der Stichprobe
   - Verteilung unter H_A sollte sich deutlich von der unter H_0 unterscheiden
6. Verteilungsfunktion F bestimmen
   - theoretisch bestimmbar
   - asymptotisch bestimmbar
   - Simulation
7. Verwerfungsbereich
   - Statistik: Verteilung der Prüfgröße

- Hypothese: Richtung einseitig/zweiseitig
- Signifikanzniveau: Irrtumswahrscheinlichkeit α

a) Verwerfungsbereich bestimmen
   - Wert für t der Teststatistik T aus Daten bestimmen
   - Tabelle oder berechnen oder

b) p-Wert bestimmen
   - Tabelle oder berechnen

8. Entscheidung fällen
   - t im Verwerfungsbereich: Verwerfen der Nullhypothese
   - p außerhalb α: Verwerfen der Nullhypothese
   - sonst: H_0 nicht verwerfbar

Gauß-Test / t-Test

- Neue Differenz in Kategorien = Zusätzlicher Parameter
  - Modellannahme
  - Nullhypothese: Parameter `IsMonkey` ist nicht nötig, Einfluss $\beta_1 = 0$
  - Alternativhypotehse: Parameter `IsMonkey` ist relevant, Einfluss $\beta_1 \neq 0$
  - Teststatistik standardisierte Differenz - *Gauß-Test* für $\beta_{IsMonkey}$

$$Z = \frac{\overline{X_a} - \overline{X_b}}{\sqrt{S_a^2/n_a + S_b^2/n_b}} \sim \mathcal{N}(0,1) = \varphi$$

  - Verwerfungsbereich festlegen und bestimmen
    - Zur Irrtumswahrscheinlichkeit $\alpha = 0.1\%$
  - Wert der Statistik berechnen, p-Wert
  - Ergebnis und Entscheidung
- Problem: kumulierter α-Fehler

F-Tests

- F-Test: Vergleich des Varianzenverhältnisses

$$F = \frac{SQE/(n_c - 1)}{SQR/(n - n_c)} \sim \mathcal{F}(n_c - 1, n - n_c)$$

- Siehe Varianzanalyse (ANOVA)

Vergleich der Likelihood

- Verhältnis der Likelihood $= \frac{L_A}{L_B}$
- Differenz der Log-Likelihood $log(L_A) - log(L_B) = l_A - l_B$
- Maximal mögliche Likelihood?
  - *Vollständiges* Modell $\widehat{y}_i \equiv y_i$ mit Likelihood L_V
- Deviance
  - (Doppelter) Unterschied zur Log-Likelihood des vollständigen Modells

$$D := 2(l_V - l_A)$$

1.12 Deviance

Verallgemeinert die Quadratfehlersumme von Normalverteilten Modellen.

- Anwendung: Modellvergleich
  - Voraussetzung: Modelle bauen aufeinander auf (*nested models*)
- Definition

$$D(\widehat{\boldsymbol{\theta}}; \mathbf{y}) := 2\Big(l(\tilde{\boldsymbol{\theta}}; \mathbf{y}) - l(\widehat{\boldsymbol{\theta}}; \mathbf{y})\Big)$$

  - $\mathbf{y}$ Werte der abhängigen Variable
  - $\widehat{\boldsymbol{\theta}}$ Schätzer der Parameter
  - $\tilde{\boldsymbol{\theta}}$ Schätzer der Parameter eines *vollständigen* Modells $\widehat{y}_i \equiv y_i$
- Beispiel Lineares Modell mit Normalverteilung(en)

$$l(\boldsymbol{\mu}; \mathbf{y}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2 - n \log\left(\sigma\sqrt{2\pi}\right)$$

$$\begin{aligned} D &= 2\Big(l(\tilde{\boldsymbol{\mu}}; \mathbf{y}) - l(\widehat{\boldsymbol{\mu}}; \mathbf{y})\Big) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2 \end{aligned}$$

  - entspricht damit Pearsons standardisierter Quadratfehlersumme, also

$$D \sim \chi^2(n - k)$$

  - Begründung: Abhängigkeiten der $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, es verbleiben k Komponenten, Freiheitsgrade in $\boldsymbol{\beta}$
  - Verteilung $\sim \chi^2(k)$ mit Anzahl der *zusätzlichen* Parameter k zum erweiterten Modell
- auch für andere Verteilungen
  - näherungsweise χ^2-verteilt

Scaled Deviance

Streuung σ ist unbekannt

- Die angegebene *scaled Deviance* ist aus den Daten berechenbar

$$D' = \sigma^2 D = \sum_{i=1}^{n} (y_i - \mu_i)^2$$

Unterscheidung

Unterscheiden sich die beiden Modelle?

- Unterschied in Deviance ΔD:

$$\Delta D(\widehat{\boldsymbol{\theta}}_A, \widehat{\boldsymbol{\theta}}_B; \mathbf{y}) = D(\widehat{\boldsymbol{\theta}}_A; \mathbf{y}) - D(\widehat{\boldsymbol{\theta}}_B; \mathbf{y}) = 2l(\widehat{\boldsymbol{\theta}}_B; \mathbf{y}) - 2l(\widehat{\boldsymbol{\theta}}_A; \mathbf{y}) > 0$$

  - $\mathbf{y}$ Werte der abhängigen Variable
  - $\widehat{\boldsymbol{\theta}}_A$ Schätzer der Parameter (k_A Stk.) des einfachen Modells
  - $\widehat{\boldsymbol{\theta}}_B$ Schätzer der Parameter (k_B Stk.) des erweiterten Modells
  - $\Delta D \geq 0$
  - Verteilung

$$\Delta D \sim \chi^2(k_B - k_A)$$

- Fisher $\mathcal{F}$-Test für Deviance
  - Betrachte das Verhältnis

$$F = \frac{D_0 - D_1}{k - q} \Big/ \frac{D_1}{n - k} \sim \mathcal{F}(k - q, n - k)$$

  - Unterschied?
    * Nullhypothese: Modell A (alle Säugetiere) ist ebenso gut wie das bessere Modell B(Affen getrennt)
    * Alternativhypothese: Modell B beschreibt den linearen Zusammenhang besser

Ergebnis

- Im Beispiel ist der Unterschied höchst signifikant ($\alpha = 0.1\%$)
  - t-Test/Gauß-Test für Parameter β_{IsMonkey}
  - Varianzanalyse für Residuen zwischen beiden Modellen
  - F-Test der Deviance zwischen beiden Modellen
- Unterschied in Deviance
  - in guter Näherung χ^2-verteilt
- Die Deviance ist eine sinnvolle Erweiterung der Pearson Quadratfehlersumme
- Konzept der Deviance gilt auch für andere Verteilungen der Exponentialfamilie

2 Generalisierte Lineare Modelle - GLM

2.1 Motivation Generalisiertes Lineares Modell

- Problemstellung
  - Jet-Piloten erfahren unter besonderes hohen Beschleunigungskräften (bezogen auf die Erdbeschleunigung g) Blackouts
- Versuch
  - Glaister und Miller (1990) erzeugten ähnliche Symptome, indem sie den Körper der Versuchspersonen einem Luftunterdruck aussetzten
- Fragestellung
  - Hängt die Ohnmacht vom Alter ab?
- Ansatz
  - Linearer fit `'symptoms ~ age'`
  - Problem: Linearer fit nicht aussagekräftig hier
- Lösung: Logit-Link
  - Wahrscheinlichkeit des Bernoulli-Ereignisses $\pi \in [0...1]$
  - Linearer Term $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$
  - Link-Funktion **logit**

$$\mathcal{E}(\mathbf{Y}) = \boldsymbol{\pi} \qquad g(\boldsymbol{\pi}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$
$$\mathcal{E}(\mathbf{Y}) = \boldsymbol{\pi} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$$

    * logit-Funktion

$$g^{-1}(\eta) = \text{logit}(\eta) = \frac{1}{1 + e^{-\eta}}$$

    * Umkehrfunktion: logarithmisches Chancenverhältnis *log-odds-ratio*

$$\eta = g(\pi) = \ln \frac{\pi}{1 - \pi}$$

- Bernoulliverteilung
  - Wahrscheinlichkeitsverteilung des Ereignisses $y \in [0, 1]$

$$f(y|\pi) = \pi^y (1 - \pi)^{1-y}$$

$$\mathcal{E}(y) = \pi$$

- Binomialverteilung
  - Wahrscheinlichkeitsverteilung der y = *Anzahl der Erfolge* mehrerer Bernoulli-Ereignisse

$$P(y|N,\pi) = \binom{N}{y}\pi^y(1-\pi)^{(N-y)} \qquad y \in \{0\dots n\}$$

$$\mathcal{E}(y) = N\pi$$

- Ergebnis Link-Funktion: Eine Link Funktion $g(\mu)$
  - kann Anforderungen an Randbedingungen von Zufallsvariablen erfüllen
    - ∞-Problem ✓
    - Verteilung der Streuung berücksichtigen ✓
  - erweitert das Lineare Modell
    - verbindet lineare Vorhersage $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$
    - und zentralen Parameter der Wahrscheinlichkeitsverteilung μ_i
- Ergebnis 'Generalisiertes Lineares Modell'

$$g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$$

$$\mathcal{E}(Y_i) = \mu_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$$
$$Y_i \sim f(\mu_i, \sigma^2, \dots)$$

2.1.1 Kategoriale Variable und Residuen

- Beispieldaten: Allison, Cicchetti (1976) *Sleep in mammals: ecological and constitutional correlates.* Science **194:** 732-734
  - Lineares Modell des Gehirn-Gewichts gegen das Körpergewicht
  - Interessant: Abweichungen vom Modell
    - systematisch?
    - Zufall (wie im Modell vorgesehen)?
- Ergebnis Residuen-Analyse
  - **Systematische** Abweichungen
    - Ausreißer, Auffälligkeit
    - Affen haben positive Residuen: eher *kein Zufall*
  - **Zufällige** Abweichungen
    - Verteilung gemäß Modell: Streuung
- Erweitertes Modell
  - Affen als eigene Kategorie

    * Kategoriale Variable `['IsMonkey']`
    * Anpassen der Designmatrix
    * Indikatorvariable c für Kategorie Affe `['IsMonkey']='no'` $= 0$ und `['IsMonkey']='yes'` $= 1 \Rightarrow \beta_1$

$$
\begin{aligned}
\mathcal{E}(\mathbf{Y}) &= \mathbf{X} \quad \boldsymbol{\beta} \\
\mathcal{E}(Y_i) &= 1 \cdot \beta_0 + c_i \cdot \beta_1 + X_i \cdot \beta_2 \\
\begin{bmatrix} Y_1 \\ \vdots \\ Y_a \\ Y_{a+1} \\ \vdots \\ Y_n \end{bmatrix} &= \begin{bmatrix} 1 & 0 & X_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & X_a \\ 1 & 1 & X_{a+1} \\ \vdots & \vdots & \vdots \\ 1 & 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}
\end{aligned}
$$

- Ergebnis Kategoriale Variable
  - wirkt als Schalter
    * Wert $X_{ij} \in [0, 1]$
    * für Parameter β_j
  - Kategorien werden von `Patsy` automatisch erkannt (z.B. wenn *String*)
    * erzwingen mit `'C(variable)'`
  - fügt sich formal in Lineares Modell ein
  - erweiterbar auf mehrere Ausprägungen
    * mehrere Spalten
    * *nicht Zahlen!*

2.1.2 Modellvergleich

- Residuen der beiden Modelle
  - Modell A: $r_{Ai} = y_i - \widehat{\mu}_{Ai} = y_i - (\mathbf{X}_A \widehat{\boldsymbol{\beta}}_A)_i$
  - Modell B: $r_{Bi} = y_i - \widehat{\mu}_{Bi} = y_i - (\mathbf{X}_B \widehat{\boldsymbol{\beta}}_B)_i$
- Residuen gehören zu einem Modell
- Minimieren
  - Kleinste-Quadrate
  - Matrix Zerlegung
  - Maximum-Log-Likelihood
- Überprüfen, ob Modellvoraussetzungen erfüllt sind

  - Scatter-Plot
  - Histogramm

2.1.3 Verdichtung der Information

- Nicht von Interesse: alle einzelnen μ_i
- Von Interesse:
  - Einfluss der unabhängigen Variablen (*erklärende* Variablen, Pediktoren) X
    - kategorial
    - kontinuierlich
    - Versuchsbedingungen $i \quad i \in [1 \dots n]$
  - zugehörige Parameter
    - modellieren X, *Gewichtung* der Einflüsse
    - Parameter $\beta_j \quad j \in [1 \dots k] \quad k \ll n$

2.2 *Generalisierte* Lineare Modelle

Link-Funktion g

verbindet additiven Einfluss (η_i) der unabhängigen Variablen $\mathbf{x}_i$ auf die (erwünschte) Verteilung der abhängigen Y_i um (μ_i)

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

Beispiel Bernoulli-Verteilung

- Exponentiell abfallende Abhängigkeit

$$P(Y_i{=}1) = e^{-\lambda t} = \pi$$
$$P(Y_i{=}0) = 1 - e^{-\lambda t} = 1 - \pi$$

- führt unter Verwendung der Link-Funktion

$$g(\pi) = \log(\pi) = -\lambda t$$

- auf eine lineare Abhängigkeit

$$g(E(Y)) = -\lambda t$$

- mit

$$\mathbf{x}_i = [t] \quad \boldsymbol{\beta} = [-\lambda]$$

- zum Generalisierten Linearen Modell

$$E(Y) = g^{-1}(x\beta)$$

Anwendung

- Biologie: Genetischer Stammbaum
- Linguistik: Abspaltung von Sprachen zum Zeitpunkt t mit gemeinsamem Wortschatz (=1) in unterschiedliche Entwicklung von Worten (=0)
- Physik: Spannung bei Kondensatorentladung über konstanten Widerstand

Modell und Fragestellung

- **Gesucht** sind die Parameter des Modells β
  - Verdichtung der Information
  - Signifikanz einer Teil-Abhängigkeit, Parameter β_j
  - Unterschiedliche Abhängigkeit bei anderen Daten
  - Unterschiedliche Modelle

2.3 Exponentialfamilie

Exponentialfamilie für Wahrscheinlichkeitsdichteverteilungen

$$f(y;\theta) = \exp\Big(a(y)b(\theta) + c(\theta) + d(y)\Big)$$

Einige wichtige bekannte Verteilungen sind Mitglied der Exponentialfamilie

- Normalverteilung
  - Parameter θ ist μ
- Binomialverteilung
  - Der einzige interessierende Parameter bei gegebenem n ist π
  - $y \in \{0 \dots n\}$
- Poissonverteilung
  - Der einzige interessierende Parameter ist λ.
  - $y \in \mathbb{N}$

Sie haben

- Gemeinsame Eigenschaften
- Gemeinsame Methoden
- und lassen sich mittels GLM-Formalismus lösen

Implementiert in `statsmodels glm`

- `Binomial ()`
- `Gamma ()`
- `Gaussian ()`
- `InverseGaussian ()`
- `NegativeBinomial ()`
- `Poisson ()`

2.3.1 Allgemeine Eigenschaften der Exponentialfamilie

- Erwartungswert
$$\mathcal{E}\left(a(Y)\right) = -\frac{c'(\theta)}{b'(\theta)}$$
- Varianz
$$Var\left(a(Y)\right) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

2.3.2 Log-Likelihood-Funktion

- Exponentialfamilie

$$l(\theta; y) = \log(f_Y) = a(y) \cdot b(\theta) + c(\theta) + d(y)$$

Score Statistik U

- Ableiten der Log-Likelihood-Funktion nach θ ergibt die *score statistic* U, als Funktion von Y eine Zufallsvariable

$$U(\theta; y) := \frac{\mathrm{d}l(\theta; y)}{\mathrm{d}\theta} = a(y) \cdot b'(\theta) + c'(\theta)$$

- mit Erwartungswert

$$\mathcal{E}(U) = 0$$

Information $\mathcal{I}$

- Varianz von U oder *Information* $\mathcal{I}$

$$\mathcal{I} := \mathrm{Var}(U) = (b'(\theta))^2 \cdot \mathrm{Var}\left(a(y)\right) = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta)$$

- Aus dem **Verschiebungssatz** folgt mit $\mathcal{E}(U) = 0$

$$\mathrm{Var}(U) = \mathcal{E}(U^2)$$

- Des Weiteren gilt

$$\mathcal{E}(U') = -\mathrm{Var}(U)$$

- $\Rightarrow$ Information

$$\mathcal{I} := \mathrm{Var}(U) = -\mathcal{E}(U')$$

2.3.3 Kanonische Verteilung

Verteilungen mit

$$a(Y) = Y$$

nennt man **kanonisch**

- Normalverteilung, Poissonverteilung, Binomialverteilung sind kanonisch
- Erwartungswert und Varianz für y haben eine einfache Form
- Der Parameter im zugehörigen Term $b(\theta)$ heißt **natürlicher Parameter**

Verteilung	natürlicher Parameter $b(\theta)$	Funktion $c(\theta)$	Funktion $d(y)$
Normal	$\frac{\mu}{\sigma^2}$	$-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$	$-\frac{y^2}{2\sigma^2}$
Binomial	$\ln(\frac{\pi}{1-\pi})$	$n\ln(1-\pi)$	$\ln\binom{n}{y}$
Poisson	$\ln\lambda$	$-\lambda$	$-\ln y!$

Natürlicher Parameter

$$f(Y;\theta) = \exp\Big(Y \cdot b(\theta) + c(\theta) + d(Y)\Big)$$

- Wählt man $b(\theta) = \theta$, dann heißt θ selbst der natürliche Parameter der Verteilung

$$f(Y;\theta) = \exp\Big(Y\theta + c(\theta) + d(Y)\Big)$$

- Möchte man diesen natürlichen Parameter selbst linear vorhersagen

$$\theta = \mathbf{X}\boldsymbol{\beta}$$

- so wird aus der allgemeinen Link-Funktion g:

$$g(\mu) = \mathbf{X}\boldsymbol{\beta}$$

- die *natürliche Link-Funktion*

$$\theta = g(\mu)$$

Verteilung	natürlicher Param. $\theta = b(\theta)$	Erwartungswert	oder $\mu = g^{-1}(\theta)$
Normal	$\theta = \frac{\mu}{\sigma^2}$	$\mu = \mu$	$\mu = \sigma^2\theta$
Binomial	$\theta = \ln(\frac{\pi}{1-\pi})$	$\mu = n\pi$	$\pi = \frac{e^\theta}{1+e^\theta}$
Poisson	$\theta = \ln\lambda$	$\mu = \lambda$	$\lambda = e^\theta$

Vereinfachungen

- Für kanonische Verteilung $a(Y) = Y$ und natürlichen Parameter $b(\theta) = \theta$ ergibt sich

$$f(Y;\theta) = \exp\Big(Y\theta + c(\theta) + d(Y)\Big)$$

- Erwartungswert

$$\mathcal{E}\left(a(Y)\right) = -\frac{c'(\theta)}{b'(\theta)}$$
$$\mathcal{E}\left(Y\right) = -c'(\theta)$$

- Varianz

$$\mathrm{Var}\left(a(Y)\right) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$
$$\mathrm{Var}\left(Y\right) = -c''(\theta)$$

Verteilung	natürlicher Param. $b(\theta)$	c	c'	c"
Normal	$\theta = \frac{\mu}{\sigma^2}$	$-\frac{\sigma^2\theta^2}{2} - \frac{1}{2}\ln(2\pi\sigma^2)$	$-\sigma^2\theta$	$-\sigma^2$
Binomial	$\theta = \ln(\frac{\pi}{1-\pi})$	$-n\ln(1+e^\theta)$	$-\frac{e^\theta}{1+e^\theta}$	$-n\frac{e^\theta}{(1+e^\theta)^2}$
Poisson	$\theta = \ln\lambda$	$-e^\theta$	$-e^\theta$	$-e^\theta$

Natürlicher Parameter und kanonischer Link

- ... ist in GLM immer für die passende Verteilung implementiert

$$\mathcal{E}(Y) = -c'(\theta)$$

- **Normal-, Poisson-** und **Binomialverteilung** haben passende Parameter
- Andere Link-Funktionen sind ebenso gut möglich

2.3.4 Zusammengesetzte Wahrscheinlichkeitsverteilung - Skalarer Parameter θ

- Ein Satz *unabhängiger, identisch verteilter* (i.i.d.) Zufallsvariabler $\mathbf{Y} = [Y_1 \dots Y_N]^T$
- mit Wahrscheinlichkeitsverteilung $f(y_i, \theta)$ aus der kanonischen Exponentialfamilie
- hat eine gemeinsame Wahrscheinlichkeitsverteilung

$$\begin{aligned} f(\mathbf{Y}, \theta) &= \prod_{i=0}^{n} \exp\Big(y_i b(\theta) + c(\theta) + d(y_i)\Big) \\ &= \exp\Big(\sum_{i=0}^{n} y_i b(\theta) + \sum_{i=0}^{n} c(\theta) + \sum_{i=0}^{n} d(y_i)\Big) \end{aligned}$$

- mit

$$\mathcal{E}(Y_i) = (\dots) = \mu$$

- wobei

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

- als auch

$$\theta_i = fkt(\mathbf{x}_i^T \boldsymbol{\beta})$$

- mit unabhängigen $\beta_j;\ j \in [1 \dots k];\ k \ll n$

Maximum-Likelihood-Schätzung

- Für kanonische Verteilungen mit $a(y) = y$ gilt

$$\mathcal{E}(Y_i) = \mu_i \qquad g(\mu_i) = \eta_i$$

- **Gesucht:** Parameter θ
- **Ansatz: Max-Log-Likelihood**

$$l_i(\theta, y_i) = y_i \cdot b(\theta) + c(\theta) + d(y_i)$$

$$l(\theta, \mathbf{y}) = \sum_{i=0}^{n} l_i = \sum y_i b(\theta) + \sum c(\theta) + \sum d(y_i)$$

$$U = \frac{\mathrm{d}l}{\mathrm{d}\theta} \stackrel{!}{=} 0$$

- **Ziel:**
  - Parameter $\widehat{\theta}$
  - Maximum der Log-Likelihood $l_{max} = l(\widehat{\theta})$
- Numerische Lösung mittels Iteration nach Newton-Raphson
  (siehe Folien)
  - Für Mitglieder der Exponentialfamilie wird eine gute Näherung U' durch dessen Erwartungswert ersetzt

$$U' \leftarrow \mathcal{E}(U') = -\mathcal{I} = -\mathrm{Var}(U)$$

  - Damit iterative Lösung nach Newton-Raphson

$$\alpha^{(m)} = \alpha^{(m-1)} + \frac{U(\alpha^{(m-1)})}{\mathcal{I}(\alpha^{(m-1)})}$$

- Beispiel Ausfallwahrscheinlichkeit
  - Weibull-Verteilung

$$f(y, \lambda, \theta) = \frac{\lambda y^{\lambda-1}}{\theta^{\lambda}} \exp\Big(-\big(\frac{y}{\theta}\big)^{\lambda}\Big)$$

  - mit
    - $y > 0$ Zeit bis zum Ausfall
    - Parameter λ Form der Verteilung, hier $\lambda = 2$
      - $\lambda = 1$ wäre Exponentialverteilung mit konstanter Ausfallrate
      - *Rayleigh*-Verteilung; für gedächtnisbehaftete Lebensdauerverteilung
    - Parameter θ Skalierung. $\Rightarrow$ Diesen gilt es zu schätzen.
  - Darstellung als Exponentialfamilienmitglied:
    - $a(y) = y^{\lambda}$ (nicht kanonisch für $\lambda \neq 1$; wir benutzen $\lambda = 2$)

    - $b(\theta) = -\theta^{-\lambda}$
    - $c(\theta) = \log\lambda - \lambda\log\theta$
    - $d(y) = (\lambda - 1)\log y$
    - mit einem *Störparameter* λ
  - Log-Likelihood
    - damit kann U berechnet werden
    - $\mathcal{I}$ als Näherung $U' \leftarrow \mathcal{E}(U')$
      - im Falle der Weibull-Verteilung geschlossen lösbar
    - Damit **Scoring Methode**
- Ergebnis der *Score Methode*
  - Für die Verteilung aus der Exponentialfamilie

$$f_Y(y|\theta) = \exp\Big(a(y)b(\theta) + c(\theta) + d(y)\Big)$$

  - führt die iterative Anpassung des Verteilungsparameters θ durch die scoring Methode

$$\theta^{(m)} = \theta^{(m-1)} + \frac{U^{(m-1)}}{\mathcal{I}^{(m-1)}}$$

  - mit der *Score Statistik* U (erste Ableitung des Log-Likelihood)

$$U(\theta, y) := \frac{\mathrm{d}l}{\mathrm{d}\theta} = a(y) \cdot b'(\theta) + c'(\theta)$$

  - und der *Information* *Information* $\mathcal{I}$ (genäherte zweite Ableitung)

$$\mathcal{I} := \mathrm{Var}(U) = \mathcal{E}(U') = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta)$$

  - in wenigen Schritten zum Ergebnis
  - Die Methode lässt sich auf mehrdimensionale Parametervektoren $\boldsymbol{\theta}$ erweitern.

2.3.5 Zusammengesetzte Wahrscheinlichkeitsverteilung - Parametervektor β

- Mehrdimensional: Scoring Methode iterative Lösung

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} + \Big(\mathcal{I}(\boldsymbol{\beta}^{(m-1)})\Big)^{-1}\mathbf{U}(\boldsymbol{\beta}^{(m-1)})$$

  - Parameter $\alpha \Rightarrow$ Parametervektor $\boldsymbol{\beta}$
  - Score-Funktion $U \Rightarrow$ Score-Vektor $\mathbf{U}$
    - Gradientenvektor der Log-Likelihood $\mathbf{U} := \nabla l$

    - mit $U_j = \frac{\partial l}{\partial \beta_j}$
  - Information $\mathcal{I} \Rightarrow$ Informations-Matrix $\mathcal{I}$
- Modell-Parameter
  - Datentupel y_i, X_{ij}, Erwartungswerte μ_i und Verteilungs-Parameter θ_i mit $i \in [1 \dots n]$
  - Verdichtete Information in Parametervektor $\boldsymbol{\beta}$
  - Komponenten β_j mit $j \in [1 \dots p]$ mit i.A. $p \ll n$
- Ableitung für Max-Log-Likelihood-Schätzer
  - Berechnung unter Verwendung des Erwartungswerts
  - Umkehrfunktion
  - Kettenregel
  - $\Rightarrow$ 1. Teilergebnis:
    - Damit ergibt sich die vektorielle score-Funktion

$$U_j = \sum_{i=1}^{n} \Big(\frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \; x_{ij} \; \frac{\partial \mu_i}{\partial \eta_i} \Big)$$

      ausgedrückt durch zugängliche Größen
- Information

$$\mathcal{I} := \mathrm{Var}(U) = -\mathcal{E}(U')$$

  - Im mehrdimensionalen Fall ist die die Information $\mathcal{I}$ die Varianz-Kovarianz-Matrix der Score-Funktion U

$$\mathcal{I}_{jk} = \mathcal{E}(U_j \; U_k)$$

  - $\Rightarrow$ 2. Teilergebnis:
    - Damit ergibt sich die Informationsmatrix

$$\mathcal{I}_{jk} = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathrm{Var}(Y_i)} \Big(\frac{\partial \mu_i}{\partial \eta_i} \Big)^2$$

- Zwischenergebnis
  - Für die **Scoring Methode** ergibt sich

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} + \left(\mathcal{I}^{(m-1)} \right)^{-1} \mathbf{U}^{(m-1)}$$

  - mit dem Schätzer für den Parametervektor

$$\mathbf{b} = [\beta_1, \dots, \beta_k]^T$$

  - der Inversen Informationsmatrix

$$\mathcal{I}^{-1}$$

  - und dem *score*-Vektor

$$\mathbf{U}$$

- Erweiterung

$$\mathcal{I}^{(m-1)} \mathbf{b}^{(m)} = \mathcal{I}^{(m-1)} \mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)}$$

2.4 IRLS

Zu lösendes Gleichungssystem

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}^T\mathbf{W}\mathbf{z}$$

hat die selbe Form, wie die Normalengleichungen für ein lineares Modell

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

- Vergleiche: Kleinste Quadrate Methode
- Designmatrix $\mathbf{X}$
- Gewichtungsmatrix $\mathbf{W}^{(m-1)}$
- Zielvektor $\mathbf{z}^{(m-1)}$
- Lösung muss iterativ gewonnen werden
  - Sowohl $\mathbf{z}$
  - als auch $\mathbf{W}$
  - hängen über $\boldsymbol{\mu}$ und $Var(Y_i)$ von $\mathbf{b}^{(m-1)}$ ab

2.4.1 iterative reweighted least squares, *IRLS*

- wird in `GLM` der Python `statsmodels` verwendet

Algorithmus

1. Finde einen Startwert $\mathbf{b}^{(0)}$
2. Berechne damit $\mathbf{z}$ und $\mathbf{W}$
3. Löse $\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}^T\mathbf{W}\mathbf{z}$

$$\mathbf{b}^{(m)} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}$$

   - und wiederhole 2. und 3. bis
4. Abbruch bei Konvergenz

Ergebnis IRLS

$$\mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{z}^{(m-1)}$$

- mit mehrdimensionaler *Iterative Reweighted Least Squares*-Methode lösbar

$$\mathbf{b}^{(m)} = \left(\mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{z}^{(m-1)}$$

- konvergiert in wenigen Schritten zum Schätzer $\mathbf{b} = \widehat{\boldsymbol{\beta}}$

2.4.2 Implementierung Python `statsmodels GLM`

- kann *Generalized Linear Models* mit verschiedenen Verteilungsfamilien aus der Exponentialfamilie
- benutzt IRLS um den Parametervektor β des Modells zu bestimmen
- liefert Ergebnis
  - `.predict`
  - `.fittedvalues`
  - `.params`
- Verwendung der Likelihood
  - Wahrscheinlichkeitsverteilung der Daten aus Sicht der Parameter
- Log-Likelihood
  - für Punkt-Schätzung von Parametern mittels Maximierung
  - für Intervall-Schätzung bei genäherter Verteilungsstatistik
  - Score Statistik $\mathbf{U}$ und
  - Informationsmatrix $\mathcal{I}$
    - IRLS