

Geodatenanalyse I:

Regressionsanalyse – Verallgemeinerte Lineare Modelle

Kathrin Menberg



Stundenplan

	08:30 – 12:30 Uhr	13:30 – 17:30 Uhr
Montag	Tag 1 / Block 1	Tag 1 / Block 2
Dienstag	Tag 2 / Block 1	Tag 2 / Block 2
Mittwoch	Tag 3 / Block 1	Tag 3 / Block 2
Donnerstag	Tag 4 / Block 1	Tag 4 / Block 2
Freitag	Tag 5 / Block 1	Tag 5 / Block 2

- ▶ 2.13 Regressionsanalyse – Lineare Regression
- ▶ **2.14 Regressionsanalyse – Verallgemeinerte lineare Modelle**
- ▶ 2.15 Fragestunde und Abschluss

Lernziele Block 2.14

Am Ende der Stunde werden die Teilnehmer:

- ▶ ... einen Überblick über verschiedene Arten von verallgemeinerten linearen Modellen und deren Einsatzgebiete haben.
- ▶ ... die mathematischen Grundlagen zur logistischen Regression kennen.
- ▶ ... eine logistische Regression in Python durchführen können.

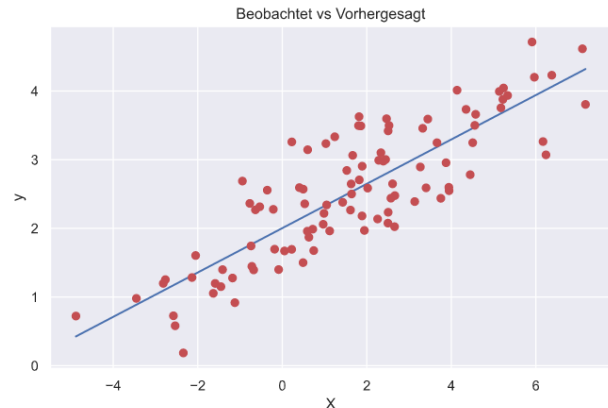
Anknüpfung an letzte Stunde...

► Lineare Regression

- Normalverteilte unabhängige Variablen und Residuen
- Kontinuierliche, abhängige Variable
- usw.

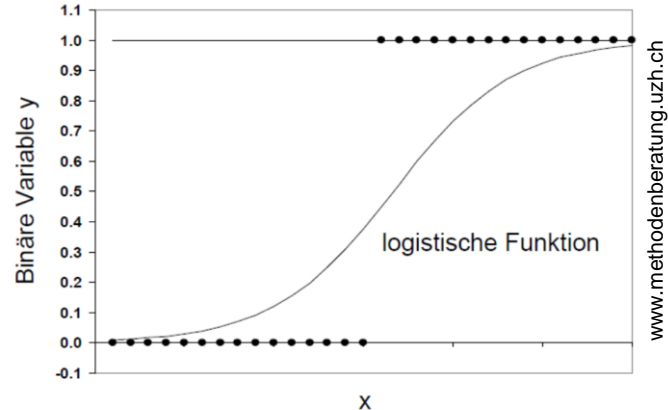
► Verallgemeinerte lineare Modelle (generalized linear models)

- Verteilungen aus der Exponentialfamilie (Poisson, Gamma, usw.)
- Diskrete abhängige Variablen



Logistische Regression




- ▶ engl. logistic (or logit) regression
- ▶ Binäre abhängige Variable: 1 („ja“, Erfolg, usw.) oder 0 („nein“, Misserfolg, usw.)
 - ▶ Wert „1“ sollte das bevorzugte Ergebnis darstellen
- ▶ Methode zur Klassifikation in zwei Kategorien
 - ▶ Funktion finden, die die Kategorien möglichst genau separiert
- ▶ Für nominale, bzw. ordinale abhängige Variablen
 - ▶ Multinominale logistische Regression
 - ▶ Ordinale logistische Regression



Grundlagen logistische Regression

► Logistische Regressionsfunktion

$$P(y = 1) = \frac{1}{1 + e^{-z}}$$

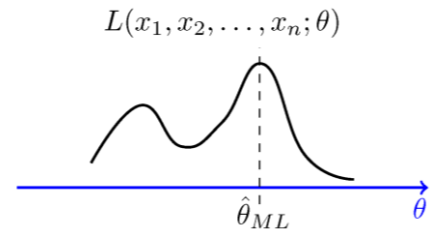
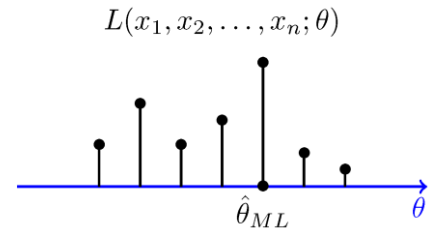
Eintrittswahrscheinlichkeit   abhängige Variable  Logit

mit $z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon_i$

- Ziel: Vorhersage der Eintrittswahrscheinlichkeit von y
- Werte nahe 0 → Eintreten unwahrscheinlich
- Werte nahe 1 → Eintreten wahrscheinlich
- Regressionskoeffizienten abschätzen mit Hilfe von Maximum Likelihood Methode

Maximum Likelihood Schätzung (MLS)

- ▶ Engl. Maximum Likelihood Estimation (MLE)
- ▶ Bestimmung der Regressionskoeffizienten so, dass für die beobachteten y-Werte möglichst hohe Wahrscheinlichkeiten vorausgesagt werden
- ▶ Maximierung der Likelihood-Funktion
 - ▶ $L(\theta) = \prod_{i=1}^n f_{\theta}(x_i)$
 - ▶ $\hat{\theta}_{ML} = \arg \max_{\theta \in \Theta} L(\theta)$
- ▶ Oft auch als Log-Likelihood, da effizienter zu bestimmen



Interpretation der Koeffizienten

$$P(y = 1) = \frac{1}{1 + e^{-\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon_i}}$$

- ▶ Vorzeicheninterpretation: $\beta > 0 \rightarrow$ Anstieg in β bewirkt Anstieg in $P(y = 1)$
- ▶ Interpretation mittels sog. „Odds Ratios“

$$Odd = \frac{P(y \text{ trifft ein})}{P(y \text{ trifft nicht ein})} = \frac{P(y \text{ trifft ein})}{1 - P(y \text{ trifft ein})}$$

$$Odds \text{ Ratio} = e^{\beta} = \frac{Odd \text{ nach dem Anstieg von } x \text{ um eine Einheit}}{Odd \text{ vor dem Anstieg von } x \text{ um eine Einheit}} = \frac{Odds_{nach}}{Odds_{vor}}$$

- ▶ Faktor, um den sich die Wahrscheinlichkeiten bei Veränderung von x ändern

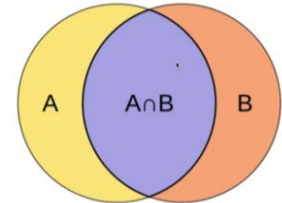
Überprüfung der Anpassungsgüte

- ▶ Jaccard Index
- ▶ Ähnlichkeitsmaß für Mengen und Vektoren
- ▶ Verhältnis von Schnittmenge und Vereinigungsmenge

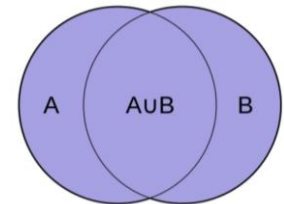
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- ▶ Logistische Regression: Vergleich der Vorhersagen aus Trainings- und Testdaten
- ▶ Bei vollständig korrekter Vorhersage $J = 1$
- ▶ Bei inkorrektter Vorhersage $J = 0$

Schnittmenge



Vereinigungsmenge



Überprüfung der Anpassungsgüte

- ▶ Wahrheitsmatrix (engl. confusion matrix)
- ▶ Anwenden des logistischen Modells auf den Testdatensatz
- ▶ Auswerten der korrekt und falsch vorhergesagten Werte für y
- ▶ Trefferquote (engl. precision)
- ▶ Maß für die Genauigkeit

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

$$P(\text{pos. Ergebnis} | \text{wirklich positiv}) = \frac{TP}{TP + FP}$$

Annahmen für logistische Regression

- ▶ Linearer Zusammenhang zwischen unabh. Variablen und $\log y$
- ▶ Aussagekräftige unabhängige Variablen, keine Multikollinearität
 - ▶ Korrelationsmatrix: Achtung bei binären Daten!
 - ▶ Extensive, sorgfältige explorative Datenanalyse
- ▶ Große Datensätze erforderlich!
 - ▶ mehrere Hundert Datenpunkte...

Übung 2.14: Logistische Regression

- ▶ Logistische Regression
 - ▶ Datensatz aus Gelman et al. (2020)
 - ▶ Arsenbelastungen in Brunnen in Bangladesch
 - ▶ Umfrage zu Wechseln von Brunnen (binäre Daten: ja/nein)
 - ▶ Weitere Daten: Distanz zu nächsten (unbelasteten) Brunnen, Arsenkonzentration, Bildungsniveau, Mitgliedschaft in lokalen Verbänden.
- ▶ Aufgaben in Jupyter Notebook: geodatenanalyse_1-2-14



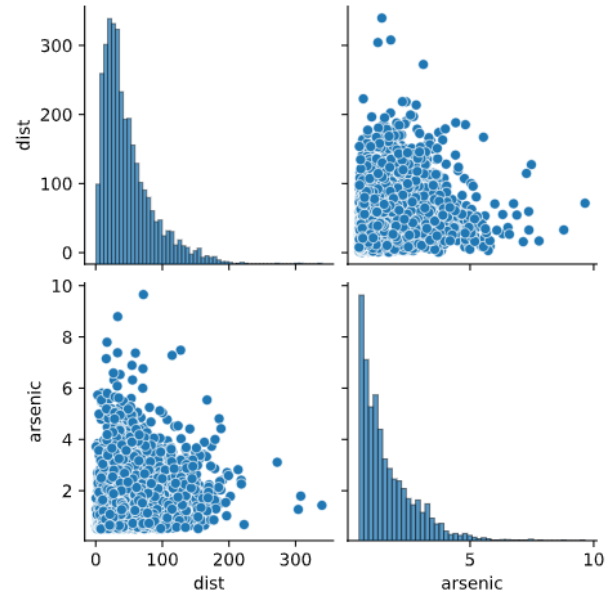
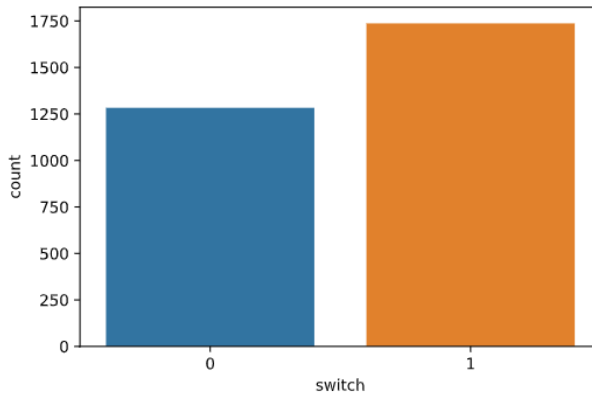
<https://practicalaction.org/>



<https://www.deutschlandfunkkultur.de/>

Aufgabenbesprechung

► Explorative Datenanalyse



Aufgabenbesprechung

► Logistische Regression mit *statsmodels.api*

Optimization terminated successfully.

Current function value: 0.650795

Iterations 5

Logit Regression Results

```

=====
Dep. Variable:          switch    No. Observations:          2416
Model:                Logit      Df Residuals:              2414
Method:                MLE       Df Model:                  1
Date:                  Mon, 22 Feb 2021    Pseudo R-squ.:            0.04470
Time:                  17:21:46           Log-Likelihood:           -1572.3
converged:              True           LL-Null:                  -1645.9
Covariance Type:        nonrobust        LLR p-value:              7.322e-34
=====

```

	coef	std err	z	P> z	[0.025	0.975]
dist100	-0.8925	0.105	-8.487	0.000	-1.099	-0.686
arsenic	0.4560	0.036	12.825	0.000	0.386	0.526

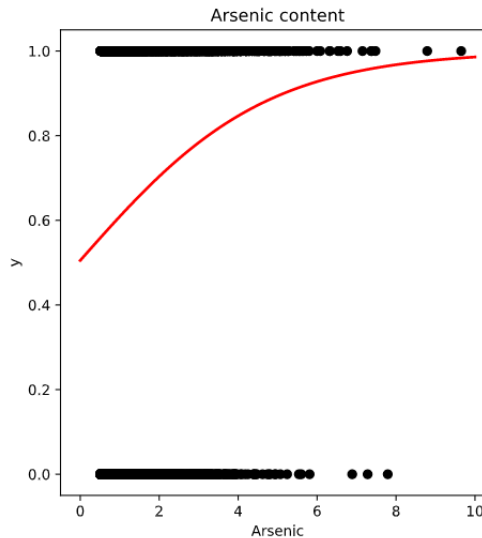
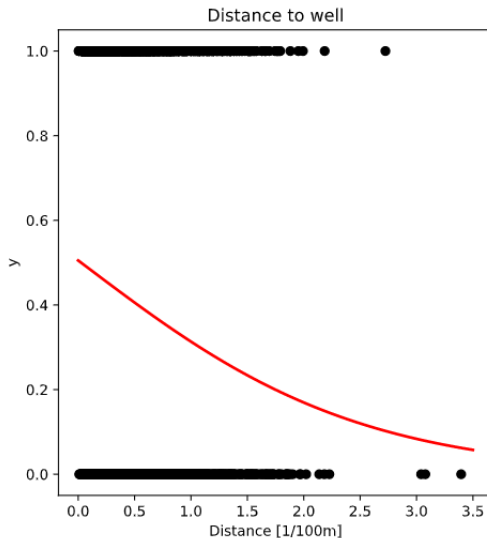
```

=====

```

Aufgabenbesprechung

► Logistische Regression mit *scikit-learn*



- Jacard = 0.55
- Confusion_matrix = $\begin{bmatrix} 71 & 190 \\ 49 & 294 \end{bmatrix}$
- Trefferquote = 0.59

Literatur

- ▶ Gelman et al. (2020) Regression and Other Stories, Cambridge University Press

Nützliche Weblinks:

- ▶ https://www.methodenberatung.uzh.ch/de/datenanalyse_spss.html
- ▶ https://www.probabilitycourse.com/chapter8/8_2_3_max_likelihood_estimation.php

