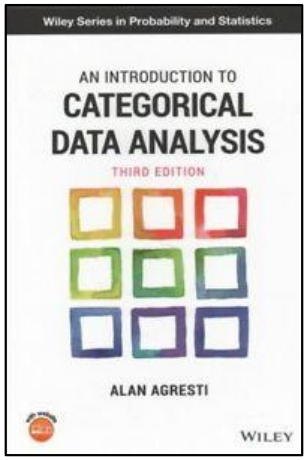
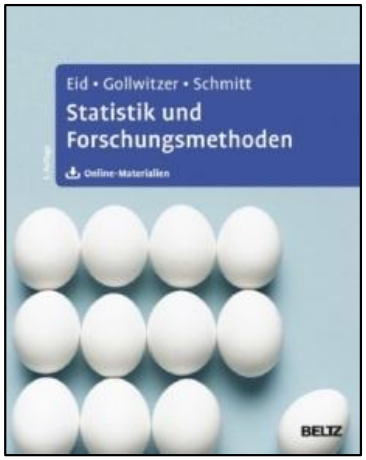
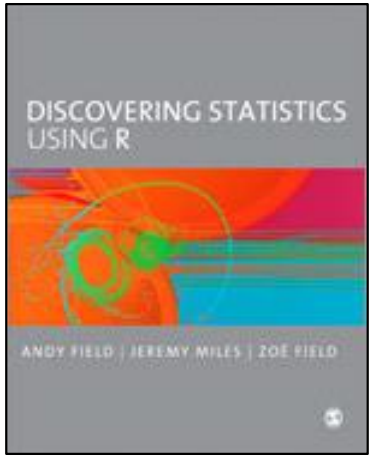
Literatur

- Agresti, A. (2019). *An Introduction into Categorical Data Analysis*. Wiley. [Kap. 3, 4]
- Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2018*). Multivariate Analysemethoden* (15. Aufl.). Berlin: Springer. [Kap. 5]
- Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5. Aufl.). Basel: Beltz. [Kap. 22]
- Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R.* Thousand Oaks: Sage. [Kap. 8]







Ausblick

1. **Verallgemeinertes lineares Modell**
2. Modell der logistischen Regression
3. Modellschätzung und Regressionsdiagnostik
4. Anwendungsbeispiel in R

Anhänge

- R-Funktionen

Nicht-metrische Kriterien in der Regression

- Bislang wurden ausschließlich Regressionsmodelle mit metrischen **abhängigen Variablen** betrachtet.
- Häufig beziehen sich Fragestellungen jedoch auf **nominal-** oder **ordinalskalierte** Kriterien:
  - Kann aufgrund der Arbeitszufriedenheit vorhergesagt werden, ob ein Mitarbeiter / eine Mitarbeitern eines Unternehmens innerhalb des kommenden Jahres kündigen wird?

    ⇨ dichotom: kündigt / kündigt nicht
  - Wie wahrscheinlich ist es an Lungenkrebs zu erkranken, wenn man die Dauer und Intensität des Rauchens berücksichtigt?

    ⇨ dichotom: raucht / raucht nicht
  - Führt bei Marathonläufern und - läuferinnen ein neuartiges Mentaltraining (im Vergleich zu einem klassischen Training) zu einer besseren Platzierung im Wettkampf?

    ⇨ Rangreihe: erster, zweiter, dritter,... Platz
  - Inwiefern bestimmen die motorischen Fähigkeiten einer Person, wie viele Aufgaben sie in einem Speedtest bearbeitet?

    ⇨ Häufigkeit: 0, 5, 7, ... Aufgaben

Verallgemeinertes lineares Modell

- Die bislang behandelte multiple lineare Regression setzt normalverteilte, metrische abhängige Variable voraus.
  - *Zur Erinnerung*: $Y = \mathrm{E}\left(Y \mid X_1, \ldots X_m\right) + E = b_0 + b_1 X_1 + \ldots + b_m X_m + E$

Das **Generalized Linear Model** (**GLM**) ist eine Erweiterung des ALM und umfasst verschiedene Regressionsmodelle für Kriterien mit unterschiedlichen Datentypen.

- GLMs bestehen aus **drei Komponenten**:
  1. Je nach Skalenniveau wird eine andere Wahrscheinlichkeitsverteilung der abhängigen Variable Y angenommen:
     - Metrische Y: Normalverteilung
     - Dichotome Y: Binomialverteilung
     - Häufigkeiten als Y (≥ 0): Poissionverteilung

     manifeste Verteilung

     diskrete Verteilungen
  2. Der Erwartungswert der abhängigen Variable wird durch lineare Prädiktoren $b_0 + b_1 x_1 + \ldots + b_m x_m$ erklärt.

     $$\mathrm{E}\left(Y \mid X_1, \ldots X_m\right) = b_0 + b_1 X_1 + \ldots + b_m X_m$$

Verallgemeinertes lineares Modell

- GLMs bestehen aus **drei Komponenten**:
  3. Die Linkfunktion g spezifiziert wie der Erwartungswert $E(Y \mid X_1, \ldots, X_m) = \mu$ der Wahrscheinlichkeitsverteilung von Y in Abhängigkeit der linearen Prädiktoren variiert: $g(\mu) = b_0 + b_1 x_1 + \ \ldots + b_m x_m$.
     - Es handelt sich also um eine Funktion, die den Erwartungswert in irgendeiner Form transformiert.
     - Dieser transformierte Erwartungswert hängt linear von den Prädiktoren ab.

     Es gibt verschiedene Linkfunktionen:

     - Die Identitäts-Linkfunktion $g(\mu) = \mu$ geht davon aus, dass der Erwartungswert direkt von den Prädiktoren abhängt:

       $$\mu = b_0 + b_1 x_1 + \ \ldots + b_m x_m$$ (lineares Modell)

       Diese wird für metrische Y (und somit die multiple lineare Regression) verwendet.

Linkfunktionen im GLM

Andere Linkfunktionen ermöglichen **nichtlineare Zusammenhänge** zwischen dem Erwartungswert und den Prädiktoren:

- Wenn der Erwartungswert nicht negativ werden kann (z.B. bei Häufigkeiten) wird die Log-Linkfunktion $g(\mu) = \log(\mu)$ verwendet.

$$\ln(\mu) = b_0 + b_1 x_1 + \ldots + b_m x_m \text{ (loglineares Modell)}$$

mit ln = natürlicher Logarithmus

- Bei dichotomen Y, wenn der Erwartungswert nur zwischen 0 und 1 liegen kann und somit einer Wahrscheinlichkeit entspricht, wird die Logit-Linkfunktion $g(\mu) = \log(\frac{\mu}{1-\mu})$ verwendet.

$$\ln(\frac{\mu}{1-\mu}) = b_0 + b_1 x_1 + \ldots + b_m x_m \text{ (Logit Modell)}$$

mit ln = natürlicher Logarithmus

Linkfunktionen im GLM

- Jede Wahrscheinlichkeitsverteilung von Y hat eine **natürliche Linkfunktion** für den Erwartungswert:
  - Für die Normalverteilung (metrische Y) ist es die Identitäts-Linkfunktion.
  - Für die Poissonverteilung (Häufigkeiten) ist es die Log-Linkfunktion.
  - Für die Binomialverteilung (dichotome Y) ist es die Logit-Linkfunktion.
- Die natürliche Linkfunktionen für jede Wahrscheinlichkeitsverteilung wird auch als kanonische Linkfunktion bezeichnet.
  - In vielen Fällen können für eine Wahrscheinlichkeitsverteilung zwar auch andere als die kanonische Linkfunktion verwendet werden, in der Praxis ist dies jedoch selten der Fall.
  - Die kanonische Linkfunktion ist der Standard in der meisten Software.

Wichtig

Die bisher behandelte **multiple lineare Regression** ist ein Spezialfall des GLM, das eine Normalverteilung von Y annimmt und den Erwartungswert von Y direkt über die Identitäts-Linkfunktion modelliert, $g(\mu) = \mu$.

Generalized linear models

Kriterium	Verteilung	Link-Funktion	Prädiktoren	Modell
Metrisch	Normal	Identität	Metrisch	Regression
Metrisch	Normal	Identität	Kategorial	Varianzanalyse (ALM)
Metrisch	Normal	Identität	Metrisch Kategorial	Kovarianzanalyse
Dichotom	Binomial	Logit	Metrisch Kategorial	Logistische Regression
Nominal	Multinomial	Logit	Metrisch Kategorial	Multinomiale Regression
Häufigkeit	Poisson	Log	Metrisch Kategorial	Loglineares Modell

wird nicht behandelt

bereits behandelt

Ausblick

1. Verallgemeinertes lineares Modell
2. **Modell der logistischen Regression**
3. Modellschätzung und Regressionsdiagnostik
4. Anwendungsbeispiel in R

Anhänge

- R-Funktionen

Binäre logistische Regression

- Viele kategoriale Variablen haben nur zwei Ausprägungen (z.B. ja / nein):
  - Arbeit versus arbeitslos
  - Raucht versus raucht nicht
  - Test bestanden versus Test nicht bestanden
  - Depression versus keine Depression

- Die binäre logistische Regression ist eine Variante der Regressionsanalyse, bei der die **abhängige Variable** Y **dichotom** in Form einer Dummy-Kodierung vorliegt:
  - 0 = nein / Misserfolg / trifft nicht zu / ...
  - 1 = ja / Erfolg / trifft zu / ...

Der Erwartungswert einer dichotomen Variable ist der Mittelwert

- Der Mittelwert einer dichotomen 0/1-Variable ist ein relativer Anteil bzw. die Wahrscheinlichkeit von Y = 1; daher: $E(Y) = \mu = P(Y = 1)$
- Die binäre logistische Regression schätzt nun die **bedingte Wahrscheinlichkeit**, dass das Kriterium den Wert 1 annimmt, bei gegebener Ausprägung auf dem Prädiktor X: $P(Y = 1 \mid X = x)$.

Probleme der linearen Regression

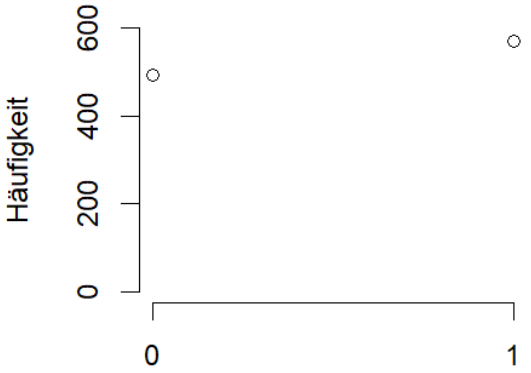
Zur Erinnerung: Modell der linearen Regression

$$Y = \mathrm{E}(Y \mid X) + E$$ E = Messfehler

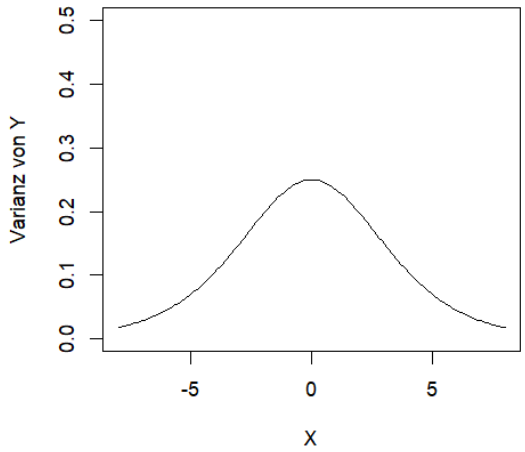
- Die Anwendung der linearen Regression auf dichotome Kriterien ist **problematisch**:
  1. Annahme der Normalverteilung der Residuen
     - Dichotome Variablen können nicht normalverteilt sein, da nur zwei Ausprägungen unterschieden werden: 0 oder 1.
  2. Annahme gleicher Varianzen der Residuen für jede Kombination von Prädiktorwerten
     - Die Annahme der Homoskedastizität ist bei dichotomen Daten immer falsch, da sich ihre Varianz aus der Multiplikation der Wahrscheinlichkeit für die beiden Kategorien ergibt:

       $$Var(Y = 1 \mid X) = P(Y = 1 \mid X) \cdot P(Y = 0 \mid X)$$

     - Die Varianz weist eine charakteristische Form auf, die symmetrisch um ein Maximum von 0.25 bei $X = 0$ ist.



Die Varianz wäre immer gleich wenn Homoskedastizität gegeben ist. Der Graf in Abb. unten rechts wäre gerade



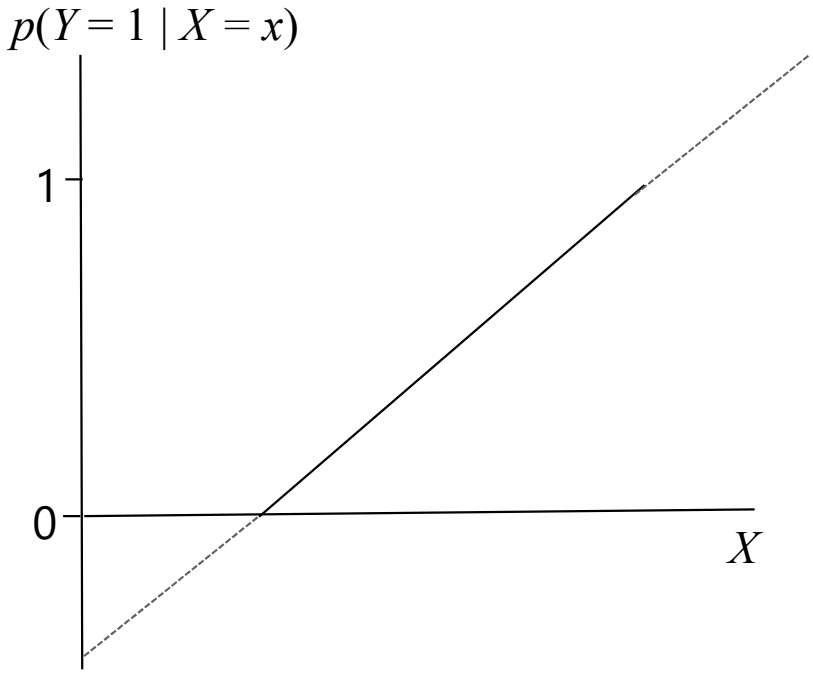
Probleme der linearen Regression

Zur Erinnerung: Modell der linearen Regression:

$$Y = \mathrm{E}\left(Y \mid X\right) + E$$

- Die Anwendung der linearen Regression auf dichotome Kriterien ist **problematisch**:
  1. Annahme der Normalverteilung der Residuen
  2. Annahme gleicher Varianzen der Residuen für jede Kombination von Prädiktorwerten
  3. Annahme einer linearen Beziehung zwischen $E(Y) = P(Y = 1)$ und den Prädiktoren

     - Eine lineare Regression mit Identitäts-Linkfunktion $P(Y = 1|X = x) = b_0 + b_1 x$ ist strukturell defekt, da die Wahrscheinlichkeit p zwischen 0 und 1 beschränkt ist, der Prädiktor X aber prinzipiell unbeschränkt ist.
     - Die Regression versucht daher $p < 0$ bzw. $p > 1$ vorherzusagen.

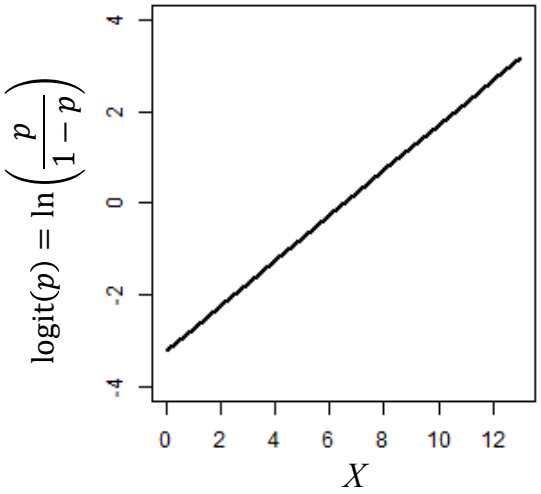


Logistische Regression

- Die logistische Regression ist ein Spezialfall des GLM mit ...
  - einer Binomialverteilung als Wahrscheinlichkeitsverteilung und
  - dem Logit als Linkfunktion.

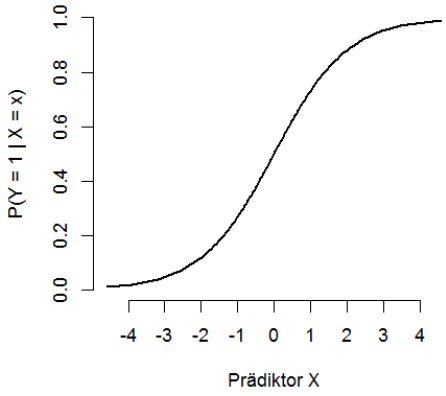
$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

mit $p = P(Y = 1|X = x) = \mu$



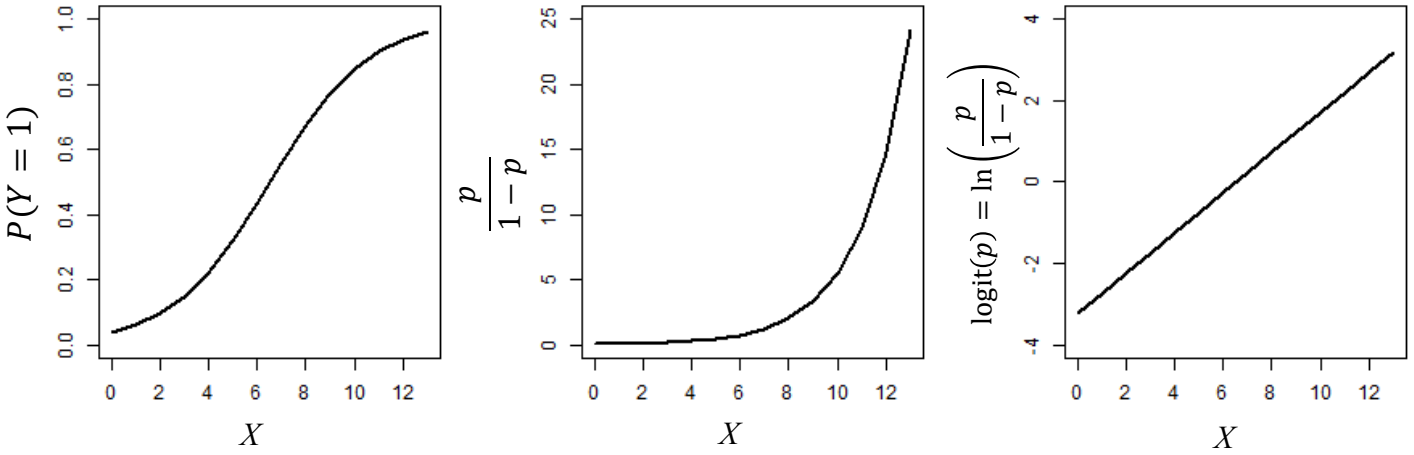
- Sie stellt den linearen Zusammenhang dem Prädiktor X mit dem Logit des Kriteriums Y dar.

- Es handelt sich um eine **nichtlineare** (logistische, S-förmige) **Regression** des dichotomen Kriteriums Y auf den Prädiktor X:
  - Die Regression ist im Mittelbereich annähernd linear.
  - Die vorhergesagten Wahrscheinlichkeiten näheren sich nur asymptotisch 0 bzw. 1 an.



Darstellungsformen

- Grundsätzlich gibt es unterschiedliche Möglichkeiten die Wahrscheinlichkeit für das Eintreten eines Ereignisses ($Y = 1$) darzustellen:
  - Wahrscheinlichkeit des Ereignisses: $p = P(Y = 1)$
  - Odds (Chance, Wettquotient) für ein Ereignis: $\frac{p}{1-p}$
  - Logit (log-Odds) für ein Ereignis: $\ln\left(\frac{p}{1-p}\right)$



Wichtig

Alle drei Darstellungsformen sind **äquivalent**, da man die anderen Varianten berechnen kann, wenn man eine Variante kennt.

Beispiel: Wenn die Odds für ein Ereignis bekannt sind, dann können sowohl das Logit für ein Ereignis wie auch die Wahrscheinlichkeit des Ereignisses berechnet werden.

Wahrscheinlichkeit für ein Ereignis

- Die **Wahrscheinlichkeit** für das Eintreten des Ereignisse $P(Y = 1) = p$ …
  - ist gleich 0, wenn das Ereignis sicher nicht eintritt.
  - ist gleich 1, wenn das Ereignis sicher eintritt.
  - liegt zwischen 0 und 1, wenn das Ereignis mit unterschiedlicher Sicherheit eintritt.

- Da es nur zwei Möglichkeiten gibt, das Ereignis tritt ein Y = 1 oder es tritt nicht ein Y = 0, aber nicht beides gleichzeitig der Fall sein kann, gilt: $P(Y = 0) = 1 - P(Y = 1)$

Odds für ein Ereignis

- Die **Odds** (Chance), dass ein Ereignis eintritt, stellen das Verhältnis aus der Wahrscheinlichkeit eines Ereignisses ($Y = 1$) und seiner Gegenwahrscheinlichkeit ($Y = 0$), dar (daher auch: **Wettquotient**):

$$Odds = \frac{P(Y = 1)}{P(Y = 0)} = \frac{P(Y = 1)}{1 - P(Y = 1)} = \frac{p}{1 - p}$$

  - … liegen zwischen 0 und ∞.
  - … ist gleich 0, wenn das Ereignis sicher nicht eintritt.
  - … ist gleich 1, wenn das Eintreten und Nicht-Eintreten des Ereignisses gleich wahrscheinlich sind (also 50%).
  - … steigt mit zunehmender Sicherheit, dass das Ereignis eintreten wird.

- *Beispiel*: In einer Stichprobe von insgesamt 179 Kindern sind 24 Kinder hyperaktiv.
  - Die Wahrscheinlichkeit für ein Kind hyperaktiv zu sein (Y = 1) ist also $P(Y = 1) = \frac{24}{179} = 0.13$.
  - Die Wahrscheinlichkeit für ein Kind nicht hyperaktiv zu sein (Y = 0) ist $P(Y = 0) = \frac{155}{179} = 0.87$ bzw. $P(Y = 0) = 1 - P(Y = 1) = 1 - 0.13 = 0.87$.
  - Der Wettquotient für ein Kind hyperaktiv zu sein (im Vergleich zu nicht hyperaktiv zu sein) beträgt also $\frac{0.13}{0.8\ 7} = \left(\frac{24}{179}\right) / \left(\frac{155}{179}\right) = \frac{25}{155} = 0.15$. bzw. liegt bei 0.15 : 1.

Logit

- Das **Logit** (log Odds) , dass ein Ereignis eintritt, ist der natürliche Logarithmus des Odds für das Ereignis:

$$\text{Logit} = \ln(\text{Odds}) = \ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \ln\left(\frac{p}{1-p}\right)$$

  - … liegt zwischen -∞ und∞.
  - … ist -∞, wenn das Ereignis sicher nicht eintritt.
  - … ist ∞, wenn das Ereignis sicher eintritt.
  - … ist gleich 0, wenn das Eintreten und Nicht-Eintreten des Ereignisses gleich wahrscheinlich sind (also 50%).
  - … steigt mit zunehmender Sicherheit, dass das Ereignis eintreten wird.

Umrechnungen

- Die Wahrscheinlichkeit p, die $Odds$ und der $Logit$ können ineinander umgerechnet werden:

	p	$Odds$	$Logit$
p		$\frac{p}{1-p}$	$\ln\left(\frac{p}{1-p}\right)$
$Odds$	$\frac{Odds}{1+Odds}$		$\ln(Odds)$
$Logit$	$\frac{e^{Logit}}{1+e^{Logit}}$	e^{Logit}	

mit $p = P(Y = 1)$ und $e = 2.718$ (Euler‘sche Zahl) = Exponentialfunktion

Rechenregel: $e^{\ln(x)} = x$

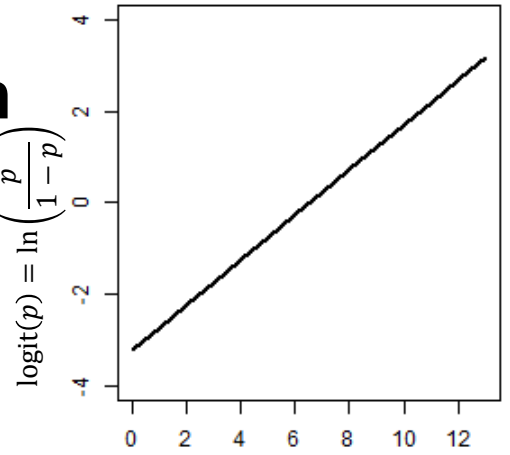
  - Wähle die Zeile für den Wert, der vorliegt, und die Spalte, in den umgerechnet werden soll.

Darstellungen der logistischen Regression

- Die logistische Regression ist auf Basis des **Logit** definiert:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

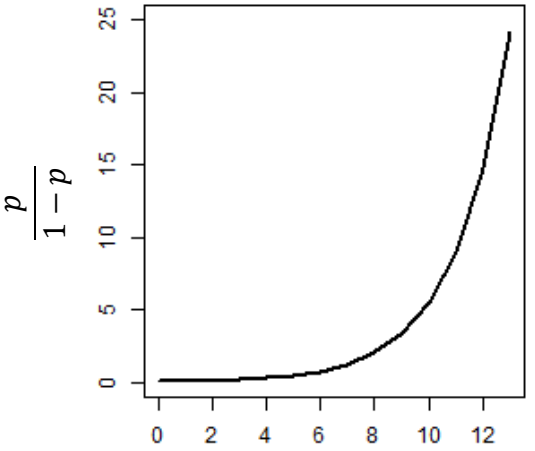
mit $p = P(Y = 1|X = x) = \mu$



- Die logistische Regression auf Basis des **Odds** ergibt sich als:

$$\frac{p}{1-p} = e^{b_0 + b_1 x}$$

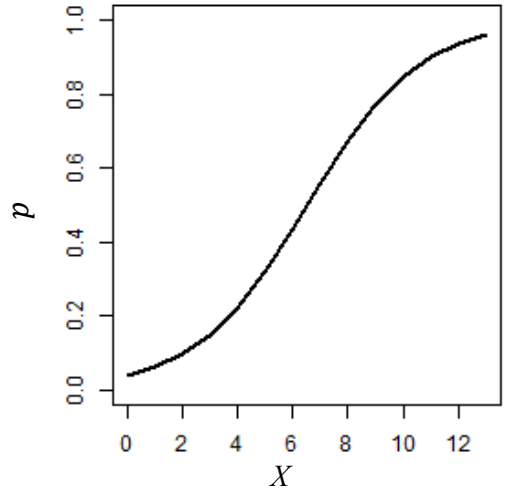
Rechenregel: $e^{\ln(x)} = x$



mit $p = P(Y = 1|X = x) = \mu$ und $e = 2.718$ (Euler'sche Zahl) = Exponentialfunktion

- Die logistische Regression auf Basis der **Wahrscheinlichkeit** ist:

$$p = P(Y = 1|X = x) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

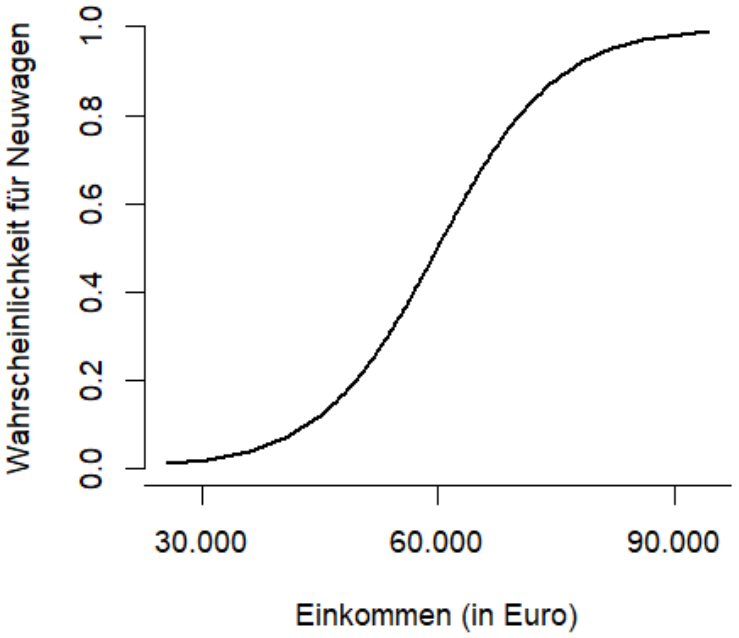


Logistische Regression

Wichtig

Die logistische Regression ist linear in Bezug auf das **Logit**, aber nicht-linear (logistisch, S-förmig) in Bezug auf die **Wahrscheinlichkeit** des Ereignisses.

- *Beispiel*: Wie hängt die Wahrscheinlichkeit ein neues ($Y = 1$) oder gebrauchtes ($Y = 0$) Auto zu kaufen vom Jahreseinkommen (X) ab?
  - Ein um 10,000 Euro höheres Einkommen X wird sich wahrscheinlich weniger stark auswirken, wenn das Jahreseinkommen 1,000,000 Euro beträgt und $P(X = 1)$ nahe 1 ist.
  - Ähnliche wird eine Einkommenserhöhung nur eine unwesentliche Rolle für die Kaufentscheidung spielen, wenn das Jahreseinkommen bei 30,000 Euro liegt und $P(X = 1)$ nahe 0 ist.
  - Am stärksten wird sich ein höheres Einkommen auswirken, wenn $P(X = 1)$ im mittleren Bereich liegt.



Interpretation des Intercepts

- Die Konstante b_0 entspricht dem Wert des Kriteriums an der Stelle $X = 0$ und ist somit analog zum Intercept der einfachen Regression zu interpretieren.

- b_0 ist das **Logit**, dass das Ereignis eintritt, wenn X = 0.

- e^{b_0} sind die **Odds**, dass das Ereignis eintritt, wenn X = 0.
  - Ist b_0 = 0, dann ist $e^{b_0} = 1$ und die Odds auch gleich 1; beide Kategorien haben somit die gleiche Wahrscheinlichkeit von 0.5.
  - Ist b_0 > 0, dann ist $e^{b_0} > 1$ und die Wahrscheinlichkeit der Kategorie $Y = 1$ ist größer als der Kategorie $Y = 0$.
  - Ist b_0 < 0, dann ist $e^{b_0} < 1$ und die Wahrscheinlichkeit der Kategorie $Y = 1$ ist kleiner als der Kategorie $Y = 0$.

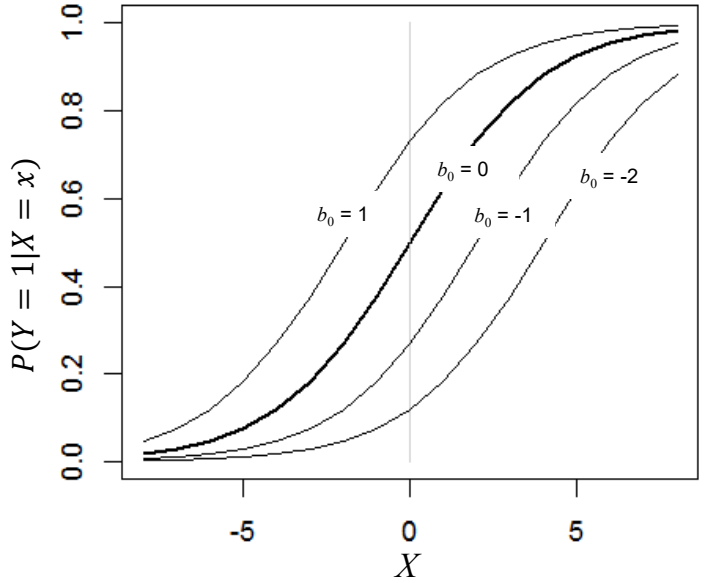
$$Logit = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

$$Odds = \frac{p}{1-p} = e^{b_0 + b_1 x}$$

$$p = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

Interpretation des Intercepts

- $\frac{e^{b_0}}{1+e^{b_0}}$ ist die **Wahrscheinlichkeit**, dass das Ereignis eintritt, wenn $X = 0$.
  - Die Konstante b_0 entspricht der bedingten Wahrscheinlichkeit an der Stelle $X = 0$.
  - Je größer b_0 ist, desto größer ist die bedingte Wahrscheinlichkeit.
  - Bei gegebenem X sind die Regressionsgeraden für unterschiedliche b_0 parallel und schneiden sich nicht.



$$Logit = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

$$Odds = \frac{p}{1-p} = e^{b_0+b_1 x}$$

$$p = \frac{e^{b_0+b_1 x}}{1+e^{b_0+b_1 x}}$$

Interpretation des Regressionsgewichts

- Das Regressionsgewicht b_1 gibt an, um welchen Wert der **Logit** sich ändert, wenn der Wert der Variablen X um eine Einheit erhöht wird.
  - Eine inhaltliche Interpretation des Logit ist schwierig.

- e^{b_1} gibt die Veränderung der **Odds** an, wenn der X um eine Einheit erhöht wird.
  - Er entspricht dem Verhältnis der Wettquotienten (Odds Ratio) für $X = x$ und $X = x + 1$:

$$\text{Odds Ratio} = \frac{P\,(Y = 1|\,X = x + 1)/(1\, - P(Y = 1|\,X = x + 1))}{P\,(Y = 1|X = x)/(1 - P(Y = 1|X = x))} = e^{b_1}$$

  - Ist b_1 = 0, dann ist $e^{b_1} = 1$ und die Chance hängt nicht vom Prädiktor ab.
  - Ist b_1 > 0, dann ist $e^{b_1} > 1$ und es besteht ein positiver Zusammenhang; d.h. die Chance die Kategorie $Y = 1$ zu wählen (im Vergleich zu $Y = 0$) steigt je größer X wird.
  - Ist b_1 < 0, dann ist $e^{b_1} < 1$ und es besteht ein negativer Zusammenhang; d.h. die Chance die Kategorie $Y = 1$ zu wählen (im Vergleich zu $Y = 0$) nimmt ab je größer X wird.

Das Odds Ratio

$$\text{Odds Ratio} = \frac{P\,(Y = 1|\,X = x + 1)/(1\, - P(Y = 1|\,X = x + 1))}{P\,(Y = 1|X = x)/(1 - P(Y = 1|X = x))} = e^{b_1}$$

Beispiel: Angenommen bei 20 Kindern sind auch andere Familienmitglieder hyperaktiv (X = 1) während bei 159 Kindern dies nicht der Fall ist (X = 0).

- Der Wettquotient für ein Kind mit Hyperaktivität in der Familiengeschichte (X = 1) beträgt **15 : 5 = 3 : 1**.
- Der Wettquotient für ein Kind ohne Hyperaktivität in der Familiengeschichte (X = 0) beträgt: **9 : 150 = 0.06 : 1**.
- Das Odds Ratio beträgt demnach (15 / 5) / (9 / 150) = **3 / 0.06 = 50**.

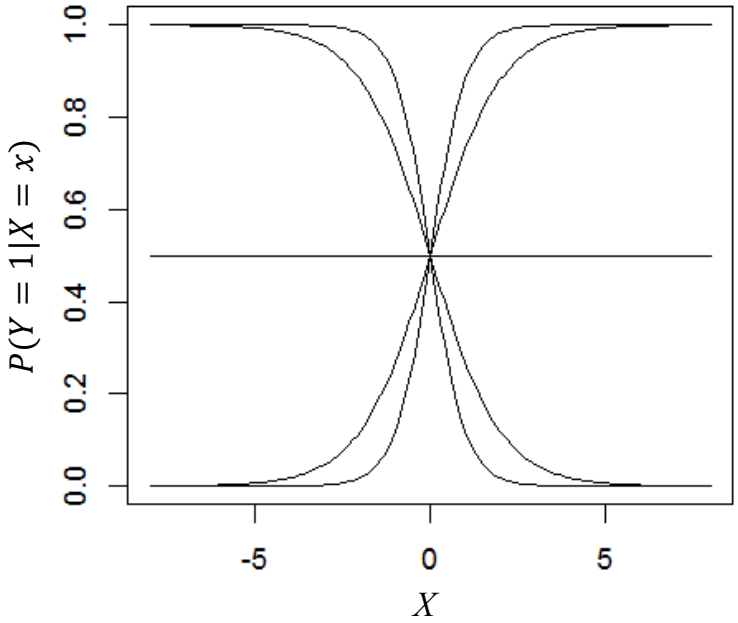
		Familiengeschichte Hyperaktivität: Ja	Familiengeschichte Hyperaktivität: Nein
Hyperaktivität	Ja	15	9
	Nein	5	150

- Die Chance hyperaktiv zu sein ist für ein Kind mit Hyperaktivität in der Familiengeschichte (X = 1) 50 mal höher als für ein Kind ohne familiäre Vorgeschichte (X = 0).

Interpretation des Regressionsgewichts

- $\frac{e^{b_0}}{1+e^{b_0}}$ gibt die Veränderung der **Wahrscheinlichkeit** für ein Ereignis an, wenn X um eine Einheit erhöht wird.

  - Ist $b_1 = 0$, dann sind beide Variablen unabhängig und die Wahrscheinlichkeitsfunktion ist konstant.
  - Ist $b_1 > 0$, dann ist steigt die bedingte Wahrscheinlichkeit die Kategorie $Y = 1$ zu wählen monoton an.
  - Ist $b_1 < 0$, dann ist nimmt die bedingte Wahrscheinlichkeit die Kategorie $Y = 1$ zu wählen monoton ab.



Wichtig

Unterschiede zwischen zwei Ausprägungen des Prädiktors X wirken sich unterschiedlich stark auf Unterschiede in den bedingten Wahrscheinlichkeiten $P(Y = 1 \mid X = x)$ aus. Sie wirken sich deutlich stärker um den Wendepunkte der Wahrscheinlichkeitsfunktion aus als an den Rändern.

Ausblick

1. Verallgemeinertes lineares Modell
2. Modell der logistischen Regression
3. **Modellschätzung und Regressionsdiagnostik**
4. Anwendungsbeispiel in R

Anhänge

- R-Funktionen

Modellschätzung und Inferenzstatistik

- Die binäre logistische Regression kann analog zur bereits behandelten linearen Regression erweitert werden:
  - M**ultiple** (metrische und kategoriale) **Prädiktoren**
  - **Interaktionen** (z.B. zur Untersuchung von Moderationseffekten)
  - Terme höherer Ordnung (z.B. zur Untersuchung von nicht-linearen Effekten)

- Die Modellschätzung und Inferenzstatistik erfolgt analog wie bei hierarchisch linearen Modellen (vgl. Foliensatz):
  - Die Parameterschätzung erfolgt iterativ über **Maximum Likelihood** Verfahren.
  - Für die Regressionsgewichte b_0 und b_1 kann die Nullhypothese $b = 0$ über **Wald-Tests** $z_b = b / SE_b$ geprüft werden.
  - Modellvergleiche können anhand des Likelihoods zweier hierarchisch geschachtelter Modelle mit **Devianztests** vorgenommen werden.

Effektstärkemaße

- Im Rahmen der linearen Regression der Determinationskoeffizient besprochen, welcher die Varianz des Kriteriums in die durch die Prädiktoren erklärte Varianz und die Residualvarianz zerlegt. Das Konzept der Residualvarianz ist in der logistischen Regression jedoch nicht definiert.
- Es wurden daher verschiedene **Pseudo-R^2** Indizes vorgeschlagen:
  - Der Cox-Snell-Index vergleicht die Likelihoods eines Modells ohne Prädiktoren L_0 mit dem geschätzten Modell L_1 und kann Werte zwischen 0 und 1 annehmen.

$$CS = 1 - \left(\frac{L_0}{L_1}\right)^{\frac{2}{n}}$$

    Bei kategorialen Kriterien ist CS aber immer < 1.
  - Das Nagelkerke-Index korrigiert CS, sodass die Obergrenze bei 1 liegt.

$$NK = \frac{CS}{CS_{\max}} = \frac{CS}{1 - (L_0)^{\frac{2}{n}}}$$

  - Diese Indizes sind trotzdem nur bedingt mit R^2 aus dem ALM vergleichbar und stellen keine Varianzaufklärung dar, sondern eher die Güte der Vorhersage.

Effektstärkemaße

- Als Effektstärkemaß für einen Prädiktor (analog zum standardisierten Regressionsgewicht) wird das **Odds Ratio** e^{b_1} und dessen Konfidenzintervall zurückgegriffen.

Wichtig

Das **Konfidenzintervall** für das Regressionsgewicht b_1 ist symmetrisch:

$$b_1 \pm z_{1-\left(\frac{\alpha}{2}\right)} \cdot SE_{b_1}.$$

Das Konfidenzintervall für das Odds Ratio e^{b_1} ist aber asymmetrisch, da e^{b_1} nach unten beschränkt ist und immer größer 0 ist.

$$e^{b_1 \pm z_{1-\left(\frac{\alpha}{2}\right)} \cdot SE_{b_1}}$$

Annahmen

- Für inferenzstatistischen Tests müssen in der Population folgende Annahmen / **Voraussetzungen** gelten:

  1. Die abhängige Variable hat genau zwei Ausprägungen (dichotom)
  2. Alle Variablen müssen eine Varianz größer als 0 aufweisen (keine Konstanten).
  3. Für alle Personen gilt die gleiche logistische Regressionsgleichung.
  4. Die bedingte Varianz (gegeben die Ausprägungen der Prädiktoren) der abhängigen Variable folgt einer Binomialverteilung.
  5. Die Kriteriumswerte sind statistisch unabhängig von einander
     $Cov\left(y_i, y_j\right) = 0$ für alle n Personen

3 Formen der logistischen Regression
Logit: Gerade
Auds: Exponentielle Steigung
Wahrscheinlichkeit: S-Form

Multikorinialität: Wenn die Prädikotren untereinander sehr stark korrelieren

Regressionsdiagnostik

- Verschiedene Ansätze zur **Regressionsdiagnostik** wurden bereits im Rahmen der linearen Regression besprochen und können vergleichbar bei der logistischen Regression angewandt werden.
- Eine korrekte Interpretation der Ergebnisse setzt voraus, dass das Modell richtig spezifiziert wurde:
  - Es liegen lineare Effekte der Prädiktoren auf das Logit vor.
  - Es wurden keine relevanten Prädiktoren (z.B. Interaktionen) ausgelassen (Underfitting) und keine irrelevanten Prädiktoren aufgenommen (Overfitting).
- Regressionskoeffizienten werden verzerrt geschätzt, wenn die Prädiktoren messfehlerbehaftet sind (Lösung: Strukturgleichungsmodelle).
- Ausreißer und einflussreiche Datenpunkte auf den Prädiktoren lassen sich über Hebel-Werte, Cooks Distanz, DfBETAS oder DfFITS identifizieren.
  - Ausreißeranalysen für das Kriterium sind bei dichotomen Daten jedoch wenig sinnvoll.
- Die Genauigkeit der Schätzungen der Regressionsgewichte nimmt bei Multikollinearität ab und kann anhand von VIF-Werten identifiziert werden.

Regressionsdiagnostik

- Vollständige Separierbarkeit liegt vor, wenn alle Personen aufgrund der Ausprägungen eines Prädiktors perfekt den beiden Kategorien des Kriteriums zugeordnet werden können.
  - Regressionsgewichte (und deren Standardfehler) können nicht geschätzt werden.
  - Identifikation über Kreuztabellen des Prädiktors mit dem Kriterium.

		X = Alter		
		15 Jahre	16 Jahre	17 Jahre
Y = Wahl-berechtigt	Ja	0	19	21
	Nein	15	0	0

Ausblick

1. Verallgemeinertes lineares Modell
2. Modell der logistischen Regression
3. Modellschätzung und Regressionsdiagnostik
4. **Anwendungsbeispiel in R**

Anhänge

- R-Funktionen

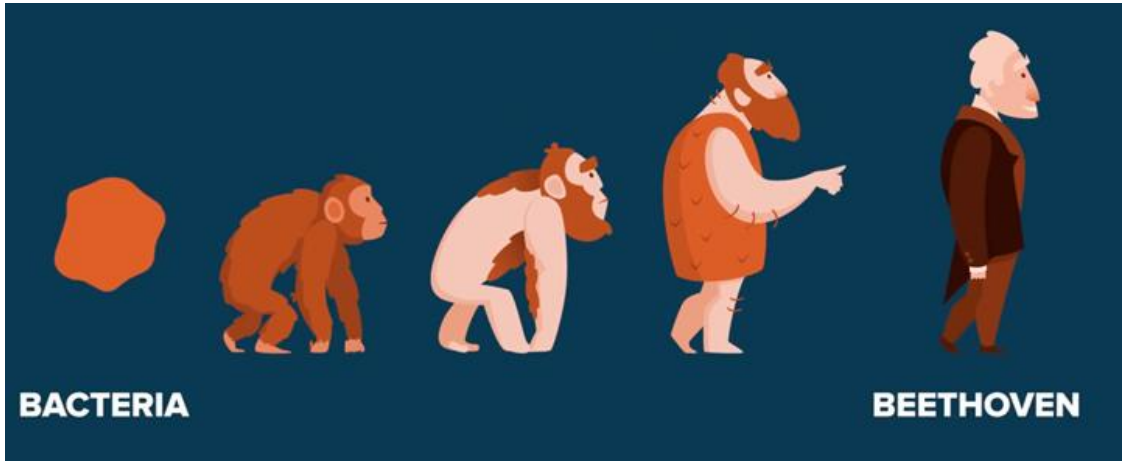
Anwendungsbeispiel

- Eine Stichprobe von 1,064 US-AmerikanerInnen wurden gefragt, ob sie an das Konzept der Evolution glauben: 571 Personen stimmten der Idee zu, während 493 Personen dies nicht taten. Hängt der Glaube an die Evolution von der politischen Orientierung (konservativ versus liberal) ab?

```
> str(dat)   # Beispieldaten
  'data.frame': 1064 obs. of 2 variables:
  $ ideology : num 0 3 5 6 5 6 1 1 2 5 ...
  $ evol     : num 0 1 1 0 0 1 1 1 1 0 ...
```

Prädiktor: Politische Orientierung (0-6)

Kriterium: Glaube an Evolution (0-1)



Logistische Regression in R

- Logistische Regressionen können in R über die `glm()`-Funktion geschätzt werden:

```
model <- glm(formula = Modellspezifikation, data = Daten,
             family = binomial("logit"))
```

Objekt, in der das Ergebnis gespeichert wird.

GLM mit Binomialverteilung und Logit-Link

Ein `data.frame` mit den Analysedaten.

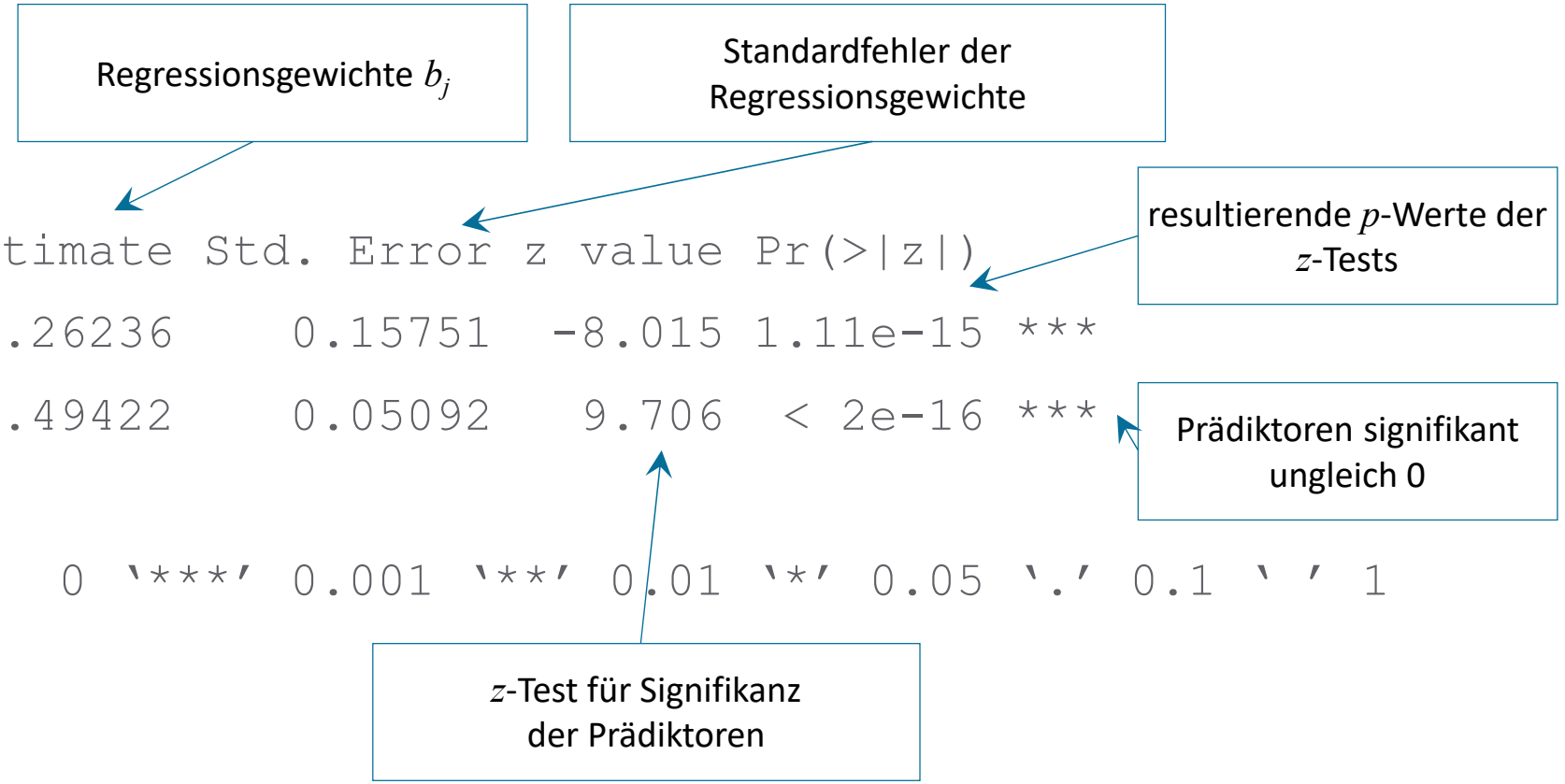
- Die `formula` besteht aus drei Teilen:
  - linke Seite: Variablenname der abhängigen Variable
  - Tilde: ~
  - rechte Seite: Variablennamen der unabhängigen Variablen (Prädiktoren), die mit `+` verbunden werden.

Die Zahl `1` steht für den Intercept, muss aber nicht explizit spezifiziert werden, da er automatisch in das Modell aufgenommen wird.

Regressionsgewichte

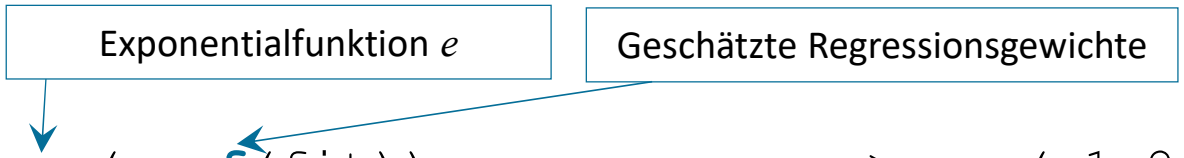
```
> fit <- glm(evol ~ ideology, data = dat,
             family = binomial(link = "logit"))
> summary(fit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.26236    0.15751  -8.015 1.11e-15 ***
ideology     0.49422    0.05092   9.706  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Odds Ratios

- Die Regressionsgewichte können in Odds Ratios, welche den Effekt auf den Wettquotienten darstellen, umgewandelt werden.




```
> exp(coef(fit))
(Intercept)    ideology
  0.2829853   1.6392175
```

```
> exp(-1.26236 + 0.49422 * 6)
5.49014
```

Odds für `ideology = 6`

- Die Chance (Odds) an Evolution zu glauben (im Vergleich zu nicht daran zu glauben) liegt bei politisch Konservativen (`ideology = 0`) bei 0.28 : 1.
- Die Chance (Odds) an Evolution zu glauben (im Vergleich zu nicht daran zu glauben) ist 1.64 mal höher, wenn die politische Orientierung um eine Einheit in Richtung liberal steigt.
- Die Chance (Odds) an Evolution zu glauben (im Vergleich zu nicht daran zu glauben) liegt bei politisch Liberalen (`ideology = 6`) bei 5.49 : 1.

Konfidenzintervalle für Regressionsgewichte

```
> round(confint(fit, level = .95), 2) # Regressionsgewicht
             2.5 % 97.5 %        Regressionsgewicht
(Intercept) -1.57  -0.95             -1.26
ideology     0.39   0.59              0.49
```

- Das Konfidenzintervall ist symmetrisch um das Regressionsgewicht.

```
> round(exp(confint(fit, level = .95)), 2) # Odds Ratio
             2.5 % 97.5 %        Odds Ratio
(Intercept)  0.21   0.39              0.28
ideology     1.48   1.81              1.64
```

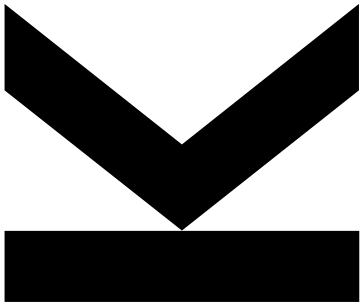
- Das Konfidenzintervall ist nicht symmetrisch um das Odds Ratio aufgrund der nicht-linearen Transformation der Exponentialfunktion, die nicht kleiner als 0 werden kann.

Pseudo-R^2

```
> # Zusatzpaket laden
> library(rsq)


> rsq(fit, type = "lr") # Cox & Snell Index
0.098
> rsq(fit, type = "n") # Nagelkerke Index
0.131
```

Anhang: R-Funktionen



Logistische Regression

R-Funktionen

Funktion	Paket	Beschreibung
`glm(formula, data, family = binomial( link = ("logit")))`		Berechnet für die Fälle in `data` die in `formula` spezifizierte logistische Regression.
`summary(object)`		Berichtet die Ergebnisse (inkl. Signifikanztests) des Regressionsmodells `object`.
`coef(object)`		Gibt die Regressionsgewichte aus dem Regressionsmodell `object` aus.
`exp(x)`		Exponentialfunktion von `x`
`rsq(model, type)`	`rsq`	Cox & Snell Index (`type = "lr"`) und Nagelkerke Index (`type = "n"`)
`confint(object, level)`		Konfidenzintervalle mit der Breite `level` für die Regressionsgewichte im Modell `object`.