TECHNISCHE UNIVERSITÄT
IN DER KULTURHAUPTSTADT EUROPAS
CHEMNITZ

Professur Psychologie digitaler Lernmedien

Institut für Medienforschung

Philosophische Fakultät

Statistik II

Logistische Regression

Divergent (2014). Concorde Filmverleih.

Prof. Dr. Günter Daniel Rey

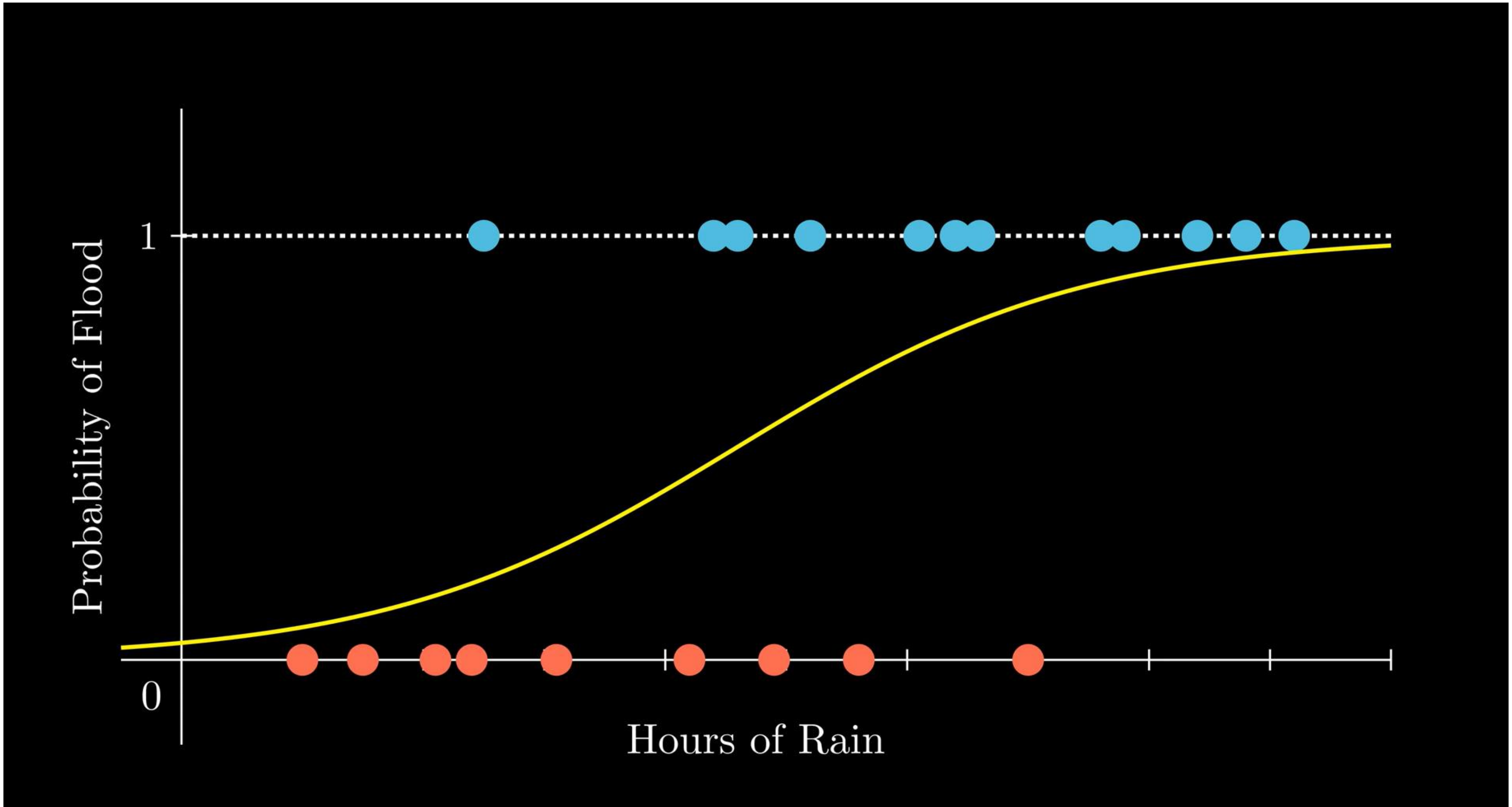
Überblick

- Einleitung
- Drei Darstellungsformen
- Multiple logistische Regression
- Parameterschätzung
- Hypothesenprüfung
- Zerlegung der Likelihood-Ratio-Teststatistik
- Klassifikation
- Voraussetzungen der Maximum-Likelihood-Schätzung und Hypothesentestung

Einleitung (Eid, Gollwitzer & Schmitt, 2017)

- Logistische Regression: Regression, bei dem die Prädiktorvariablen zumindest teilweise metrisch und die Kriteriumsvariable kategorial ist
- Einfache logistische Regressionsanalyse: Logistische Regression mit einer einzelnen Prädiktorvariable

Kurzes Erklärvideo zur logistischen Regressionsanalyse



https://www.youtube.com/watch?v=EKm0spFxFG4

Drei Darstellungsformen der logistischen Regressionsanalyse (Eid, Gollwitzer & Schmitt, 2017)

- In Form bedingter Wahrscheinlichkeiten: Wahrscheinlichkeit einer Kategorie der Kriteriumsvariablen als Funktion der Prädiktorvariablen
- In Form bedingter Wettquotienten: Modellierung des Wettquotienten als Funktion der Prädiktorvariablen
- In Form bedingter Logits: Zerlegung des Logits in Linearkombination der Prädiktorvariablen

Darstellung in Form bedingter Wahrscheinlichkeiten (Eid, Gollwitzer & Schmitt, 2017)

- Exponentialfunktion mit zwei Parametern
  - β_0: Angabe, dass generell die Wahrscheinlichkeit, eine bestimmte Kategorie zu wählen, bei X = 0 auf einem höheren oder geringeren Niveau liegen kann
  - β_1: Angabe, wie stark die Wahrscheinlichkeit, die Kategorie zu wählen, mit Zunahme der Werte auf der Prädiktorvariable ansteigt
- Formel der Exponentialfunktion: $\boxed{P(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 \cdot X}}{1 + e^{\beta_0 + \beta_1 \cdot X}}}$
- Vergleich mit der linearen Regression: $\boxed{\hat{y} = b_0 + b_1 \cdot x}$

Darstellung in Form bedingter Wahrscheinlichkeiten (Eid, Gollwitzer & Schmitt, 2017)

- Auswirkungen verschiedener Parameter / Regressionsgewichte

β_0

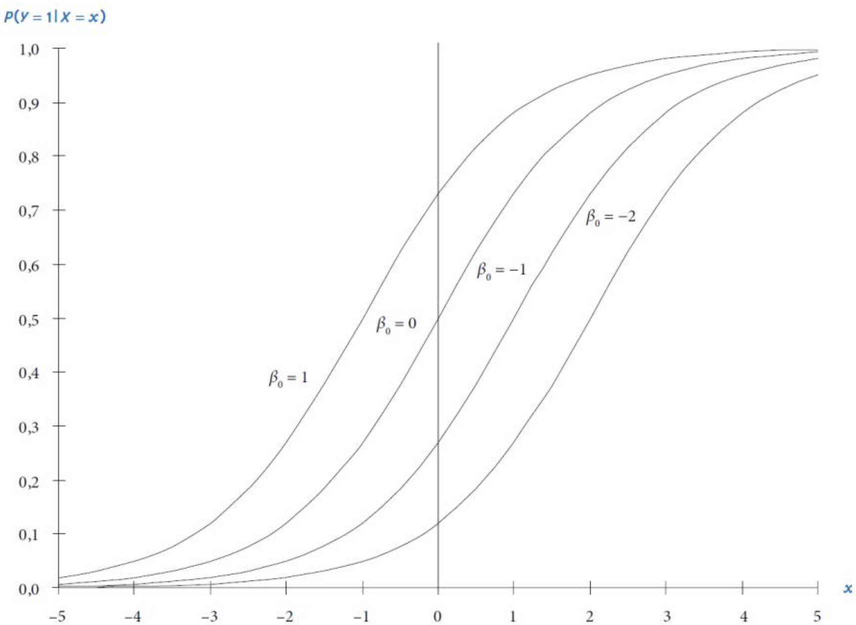


Abbildung 22.2 Abhängigkeit der bedingten Wahrscheinlichkeit von einer metrischen unabhängigen Variablen im logistischen Regressionsmodell: Auswirkungen verschiedener Regressionskonstanten β_0

β_1

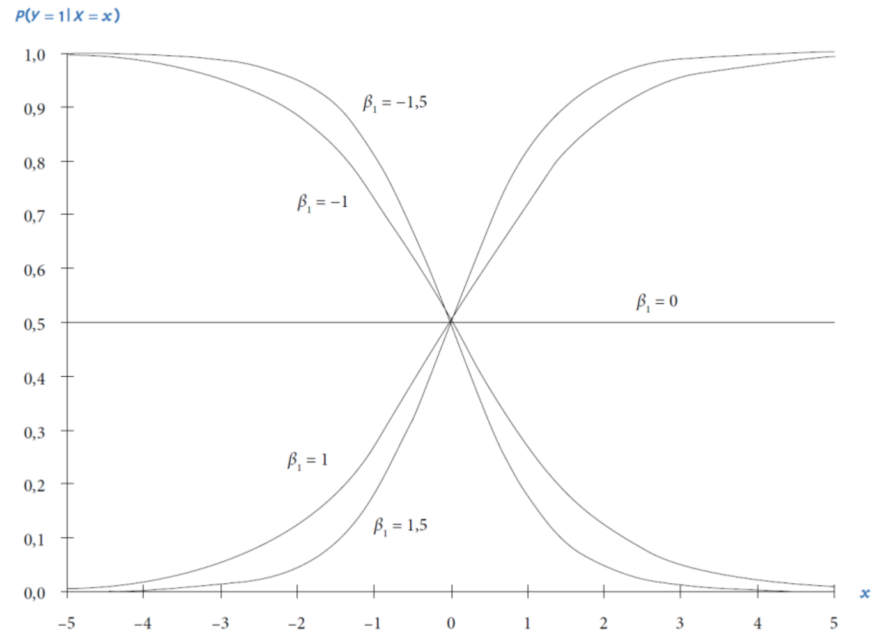


Abbildung 22.3 Abhängigkeit der bedingten Wahrscheinlichkeit von einer metrischen unabhängigen Variablen im logistischen Regressionsmodell: Auswirkungen verschiedener Regressionsgewichte β_1

Quellen: Eid, Gollwitzer und Schmitt (2017)

Darstellung in Form bedingter Wettquotienten (Eid, Gollwitzer & Schmitt, 2017)

- Wettquotienten (Odds): Verhältnis aus der Wahrscheinlichkeit eines Ereignisses und seiner Gegenwahrscheinlichkeit
- Bedingter Wettquotient im logistischen Regressionsmodell: Verhältnis aus der bedingten Wahrscheinlichkeit $P\,(\,Y = 1 \mid X = x\,)$ und der Gegenwahrscheinlichkeit $[1 - P\,(\,Y = 1 \mid X = x\,)]$
- Formel: $\boxed{\frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\beta_0+\beta_1 \cdot X} = e^{\beta_0} \cdot e^{\beta_1 \cdot X} = e^{\beta_0} \cdot (e^{\beta_1})^X}$
- Bedeutung der Parameter:
  - e^{β_0}: Entspricht dem Wettquotienten an der Stelle $X = 0$; Wenn $\beta_0 = 0 \rightarrow e^{\beta_0} = 1$ und Wettquotient = 1; Wenn $\beta_0 > 0 \rightarrow e^{\beta_0} > 1$; Wenn $\beta_0 < 0 \rightarrow e^{\beta_0} < 1$
  - e^{β_1}: Odds-Ratio; gibt die Veränderung des Wettquotienten an, wenn die Prädiktorvariable um eine Einheit erhöht wird

Darstellung in Form des Logits (Eid, Gollwitzer & Schmitt, 2017)

- Logit: Logarithmierte Wettquotient
- Formel: $\boxed{\ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 \cdot X}$
- Rechte Seite der Gleichung: Entspricht exakt der einfachen linearen Regression
- Logit strebt gegen +∞, wenn der Wettquotient gegen ∞ strebt
- Bedeutung der Parameter:
  - β_0: Konstante entspricht dem Wert des Logits an der Stelle X = 0 → Achsenabschnitt in der einfachen linearen Regression
  - β_1: Regressionsgewicht zeigt an, um welchen Wert der Logit sich ändert, wenn der Wert der Variablen X um eine Einheit erhöht wird; $\beta_1 = 0$ → Kein Zusammenhang zwischen beiden Variablen

Multiple logistische Regression
(Eid, Gollwitzer & Schmitt, 2017)

- Multiple logistische Regression: Logistische Regression mit mehreren Prädiktorvariablen
- Indizierung der Prädiktorvariablen: $j = 1, \ldots, k$
- Nachfolgend die Formeln zu den drei Darstellungsformen:
- Exponentialfunktion: $P(Y = 1|X_1, \ldots, X_k) = \frac{e^{\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2+\cdots+\beta_k\cdot X_k}}{1+e^{\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2+\cdots+\beta_k\cdot X_k}}$
- Bedingter Wettquotient: $\frac{P(Y=1|X_1,\ldots,X_k)}{1-P(Y=1|X_1,\ldots,X_k)} = e^{\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2+\cdots+\beta_k\cdot X_k}$
- Logit: $\ln\left(\frac{P(Y=1|X_1,\ldots,X_k)}{1-P(Y=1|X_1,\ldots,X_k)}\right) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$

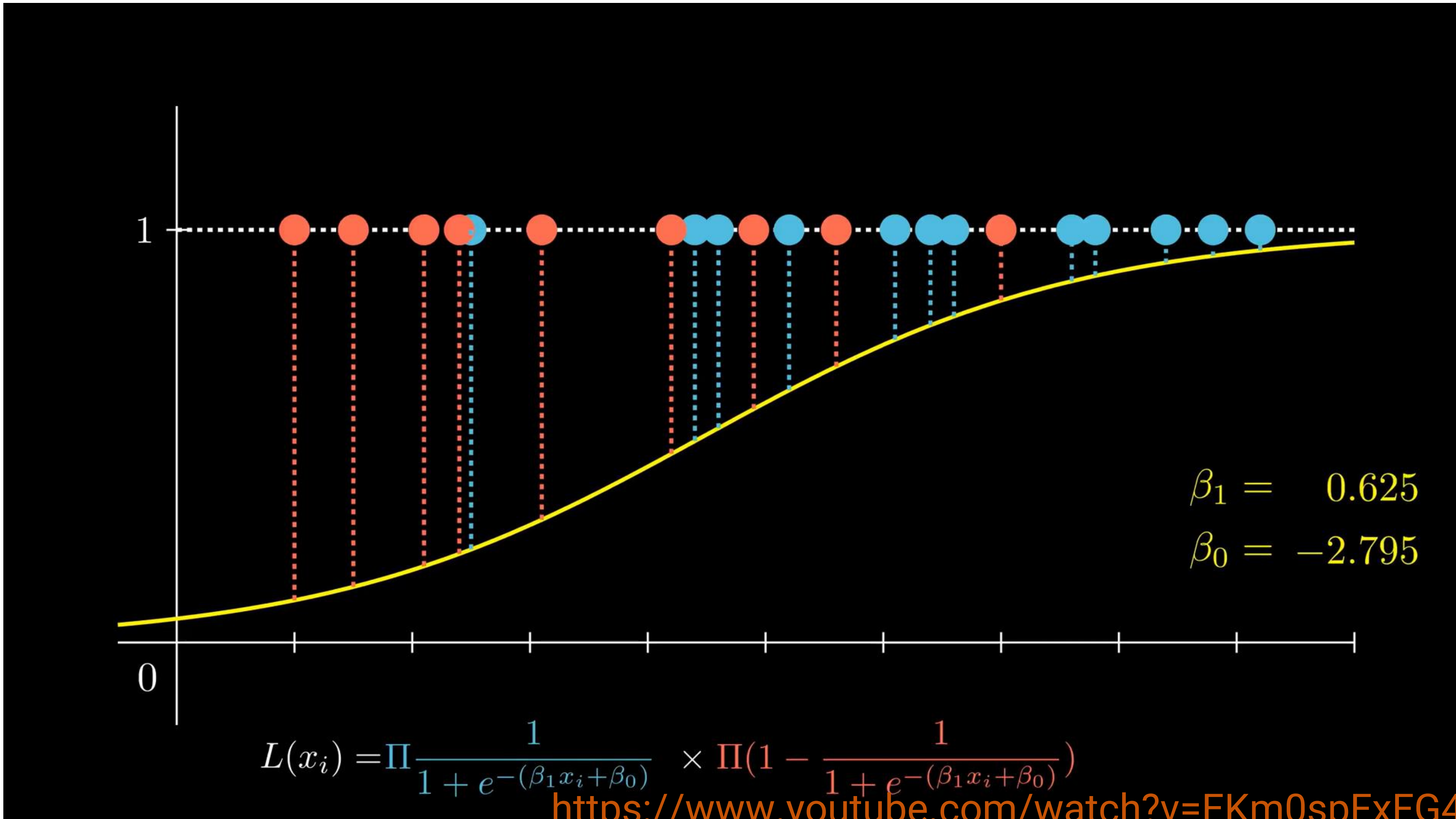
Multiple logistische Regression
(Eid, Gollwitzer & Schmitt, 2017)

- Skalenniveau: Prädiktorvariablen können metrischer als auch qualitativer Natur sein
- Wechselwirkungen: Interaktionseffekte zwischen den Prädiktorvariablen ebenfalls modellierbar
- Indikatorcodierung (Dummy-Codierung): Berücksichtigung nominalskalierter und ordinalskalierter Prädiktorvariablen mittels Indikatorcodierung
- Nonlineare Abhängigkeiten: Mittels Quadrierung oder Potenzen höherer Ordnung der Prädiktorvariablen ebenfalls modellierbar

Parameterschätzung (Eid, Gollwitzer & Schmitt, 2017)

- Parameterschätzung: Mittels Maximum-Likelihood-Verfahren
- Likelihood-Funktion: beschreibt die Wahrscheinlichkeit der Daten, die man in einer Untersuchung erhalten hat, als Funktion der Modellparameter unter der Voraussetzung, dass das Modell gilt
- Bedingte Wahrscheinlichkeiten: Ausschließlich abhängig von den Regressionsparametern
- Maximierung der Likelihood-Funktion L: Ziel bei der Schätzung der Regressionsparameter
- Schätzung der Regressionsparameter: Mittels iterativer statistischer Verfahren

Kurzes Erklärvideo zur Parameterschätzung in der logistischen Regressionsanalyse



https://www.youtube.com/watch?v=EKm0spFxFG4

Hypothesenprüfung (Eid, Gollwitzer & Schmitt, 2017)

- Hypothesenprüfung analog zur multiplen Regressionsanalyse
  1. H_0: Einzelner Parameter (β_0 oder β_j) = 0
  2. H_0: Alle Parameter $\beta_1 = \ldots = \beta_j = \ldots = \beta_k = 0$
  3. H_0: Satz von UVs keinen Einfluss über andere UVs hinaus

- Zu 1. H_0: Einzelner Parameter (β_0 oder β_j) = 0
- Drei Testverfahren
  - *z*-Test
  - Wald-Test
  - Likelihood-Ratio-Test (Likelihood-Quotienten-Test)
- Verfahren werden nachfolgend kurz skizziert

Hypothesenprüfung (Eid, Gollwitzer & Schmitt, 2017)

- *z*-Test: Geschätzter Parameter wir durch seinen geschätzten Standardfehler geteilt
- Formeln: $z = \frac{b_0}{\hat{\sigma}_{B_0}}$ bzw. $z = \frac{b_j}{\hat{\sigma}_{B_j}}$
- Prüfgröße: Asymptotisch standardnormalverteilt
- Bestimmung des kritischen Wertes: Für ein a priori festgelegtes α-Niveau bzw. die *p*-Werte anhand der Quantile der Standardnormalverteilung
- Signifikanzprüfung: Vergleich des empirischen *z*-Wertes mit dem kritischen *z*-Wert

Hypothesenprüfung (Eid, Gollwitzer & Schmitt, 2017)

- Wald-Test: Basiert auf dem quadrierten *z*-Wert
- Formeln: $z^2 = \frac{b_0^2}{\hat{\sigma}_{B_0}^2}$ bzw. $z^2 = \frac{b_j^2}{\hat{\sigma}_{B_j}^2}$
- Prüfgröße: Wald-Statistik genannt; asymptotisch χ^2-verteilt
- Bestimmung des kritischen Wertes: Für ein a priori festgelegtes α-Niveau bzw. die *p*-Werte anhand der Quantile der χ^2-Verteilung
- Signifikanzprüfung: Vergleich des empirischen z^2-Wertes mit dem kritischen χ^2-Wert

Hypothesenprüfung (Eid, Gollwitzer & Schmitt, 2017)

- Likelihood-Ratio-Test: Vergleich der Likelihoods zweier logistischer Regressionsmodelle
- Modellvergleich: Vergleich eines uneingeschränkten Modells, in dem alle Modellparameter frei geschätzt werden können mit einem eingeschränkten Modell, in dem ein Parameter (β_j) nicht mehr frei geschätzt, sondern auf 0 fixiert wird
- Prüfung durch Likelihood-Ratio-Test: Sinkt die Likelihood durch die Fixierung auf 0 signifikant?
- Formel: $\mathrm{LR} = -2 \cdot \ln\left(\frac{L_e}{Lu}\right) = -2 \cdot [\ln(L_e) - \ln(L_u)]$

L_e	= Likelihood des eingeschränkten Modells
L_u	= Likelihood des uneingeschränkten Modells

- Signifikanzprüfung: Vergleich des LR-Wertes mit einem kritischen χ^2-Wert

Hypothesenprüfung (Eid, Gollwitzer & Schmitt, 2017)

- Vergleich der drei Tests (*z*-Test, Wald-Test, Likelihood-Ratio-Test)
- Selbes Ergebnis: *z*-Test und Wald-Test
- Teststärke: Geringere Teststärke des Wald-Tests im Vergleich zum LR-Test, wenn die Regressionsgewichte einen großen (positiven oder negativen) Wert annehmen oder bei kleineren Stichproben
- Fazit aufgrund der Teststärke: Likelihood-Ratio-Test ist dem Wald-Test und dem *z*-Test vorzuziehen

Zerlegung der Likelihood-Ratio-Teststatistik (Eid, Gollwitzer & Schmitt, 2017)

- Additive Zerlegung von Likelihood-Ratio-Teststatistiken in Teilmodelle mit unterschiedlich vielen UVs
- Beispiel: Teilmodelle mit vier UVs

$$M_u\text{:}\ \ln\left(\frac{P(Y=1|X_1,X_2,X_3,X_4)}{1-P(Y=1|X_1,X_2,X_3,X_4)}\right)$$
$$=\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2+\beta_3\cdot X_3+\beta_4\cdot X_4$$

$$M_{e1}\text{:}\ \ln\left(\frac{P(Y=1|X_1,X_2,X_3)}{1-P(Y=1|X_1,X_2,X_3)}\right)$$
$$=\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2+\beta_3\cdot X_3$$

$$M_{e2}\text{:}\ \ln\left(\frac{P(Y=1|X_1,X_2)}{1-P(Y=1|X_1,X_2)}\right)$$
$$=\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2$$

$$M_{e3}\text{:}\ \ln\left(\frac{P(Y=1|X_1)}{1-P(Y=1|X_1)}\right)=\beta_0+\beta_1\cdot X_1$$

$$M_{e4}\text{:}\ \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=\beta_0$$

Zerlegung der Likelihood-Ratio-Teststatistik (Eid, Gollwitzer & Schmitt, 2017)

- Zwei generelle Strategien zur Auswahl unabhängiger Variablen (= Prädiktorvariablen) analog zum Vorgehen in der multiplen, linearen Regressionsanalyse
  - Auswahl aufgrund theoretischer Überlegungen
  - Datengesteuerte Auswahl
- Strategien bei der datengesteuerten Auswahl
  - Vorwärtsselektion
  - Rückwärtselimination

Effektgrößen (Eid, Gollwitzer & Schmitt, 2017)

- Determinationskoeffizient: Allgemein akzeptierte Effektgröße in der multiplen Regressionsanalyse; gibt den Varianzanteil der AV an, der durch die UVs aufgeklärt wird
- Weitere Effektgrößen: Siehe Sitzung zur Stichprobenumfangsplanung
- Effektgrößen in der logistischen Regression: Kein generell anerkanntes Maß, sondern nur verschiedene Vorschläge
- Vorschläge: Nachfolgend drei Indizes, die typischerweise in Statistikprogrammen berichtet werden: Koeffizienten nach…
    - McFadden (1974)
    - Cox und Snell (1989)
    - Nagelkerke (1991)

Effektgrößen (Eid, Gollwitzer & Schmitt, 2017)

- McFadden-Index: Index (*MF*) von McFadden (1974), der auf einem Vergleich von folgenden logarithmierten Likelihoods beruht
  - L_M: Likelihood des Modells mit allen UVs
  - L_0: Likelihood des Modells ohne UVs (nur mit Regressionskonstante)
  - Differenz $\ln(L_M) - \ln(L_0)$: Gibt den (nicht normierten) Erklärungsgewinn an, der aus der Hinzunahme der UVs in das Modell resultiert
  - L_S: Likelihood des saturierten Modells, welches keine Restriktionen enthält und die Daten somit perfekt reproduziert ($L_S = 1$)
- Formel: $MF = \frac{\ln(L_M) - \ln(L_0)}{\ln(L_S) - ln(L_0)} = \frac{ln(L_0) - \ln(L_M)}{ln(L_0)}$
- Wertebereich: Zwischen 0 (keine Erklärung durch die UVs) und 1 (perfektes Modell)

Effektgrößen (Eid, Gollwitzer & Schmitt, 2017)

- Cox-Snell-Index: Index (*CS*) nach Cox und Snell (1989) vergleicht die Likelihood des Modells, das nur den Parameter β_0, aber keine UVs enthält (L_0), mit dem Modell, das die *k* UVs enthält (L_M)
- Formel: $CS = 1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}}$
- Funktionsweise: Je größer L_M im Vergleich zu L_0 wird, d. h., je erklärungsstärker die UVs sind, umso kleiner wird der Wert L_0/L_M und umso größer wird der Index *CS*
- Cox-Snell-Index = Determinationskoeffizient (bei Anwendung auf eine multiple Regressionsanalyse)
- Wertebereich: Zwischen 0 (keine Erklärung durch die UVs) und…(siehe nächste Folie)

Effektgrößen (Eid, Gollwitzer & Schmitt, 2017)

- Wertebereich des Cox-Snell-Index: Nagelkerke (1991) hat gezeigt, dass der maximal mögliche Wert des *CS*-Index ist:
- $CS_{max} = 1 - (L_0)^{\frac{2}{n}}$
- Nagelkerke-Index: Standardisierung des CS-Index an CS_{max}:
- Formel: $NK = \frac{CS}{CS_{max}}$
- Wertebereich: Zwischen 0 und 1

Klassifikation (Eid, Gollwitzer & Schmitt, 2017)

- Klassifikation von Personen: Zuordnung von Personen zu einer Klasse von Personen (Kategorie der AV) anhand der Regressionsgleichung mittels logistischer Regressionsanalyse
- Wahrscheinlichkeiten: Zuordnung der Person mittels der (anhand der Regressionsgleichung) geschätzten Wahrscheinlichkeiten für die Kategorien der AV
- Beispiel: Prognose für spätere Alzheimer-Erkrankung

Voraussetzungen der Maximum-Likelihood-Schätzung & Hypothesentestung (Eid, Gollwitzer & Schmitt, 2017)

- Korrekte Modellspezifikation: Annahme erfüllt, wenn das Modell die relevanten UVs enthält und die bedingte Wahrscheinlichkeitsfunktion der postulierten Funktion entspricht
- Bedingte Binomialverteilung: Annahme impliziert, dass die bedingte Varianz der Varianz der Binomialverteilung folgt
- Unabhängigkeit der Beobachtungen: Annahme, dass Beobachtungen unabhängig voneinander sind (z. B. bei einer echten Zufallsstichprobe gegeben)

Zusammenfassung I

- Logistische Regression: Regression, bei dem die Prädiktorvariablen zumindest teilweise metrisch und die Kriteriumsvariable kategorial ist
- Darstellungsformen der logistischen Regressionsanalyse: In Form bedingter Wahrscheinlichkeiten, bedingter Wettquotienten oder bedingter Logits
- Multiple logistische Regression: Logistische Regression mit mehreren Prädiktorvariablen
- Parameterschätzung: Mittels Maximum-Likelihood-Verfahren
- Hypothesenprüfung: Mittels der Testverfahren *z*-Test, Wald-Test oder Likelihood-Ratio-Test
- Zerlegung der Likelihood-Ratio-Teststatistik analog zur multiplen Regressionsanalyse

Zusammenfassung II

- Klassifikation Zuordnung von Personen zu einer Klasse von Personen (Kategorie der AV) anhand der Regressionsgleichung
- Voraussetzungen der Maximum-Likelihood-Schätzung und Hypothesentestung: Korrekte Modellspezifikation, bedingte Binomialverteilung und Unabhängigkeit der Beobachtungen

Prüfungsliteratur

- Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5. Aufl.). Weinheim: Beltz.
  - Logistische Regression (S. 799–839)

Weiterführende Literatur

- Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2025). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung* (18. Auflage). Wiesbaden: Springer Gabler.
  - Kapitel 5: Logistische Regression (S. 293–387)
- Kalisch, M., & Meier, L. (2021). *Logistische Regression: eine anwendungsorientierte Einführung mit R.* Springer Nature.