11. Vorlesung Statistik II
Logistische Regression

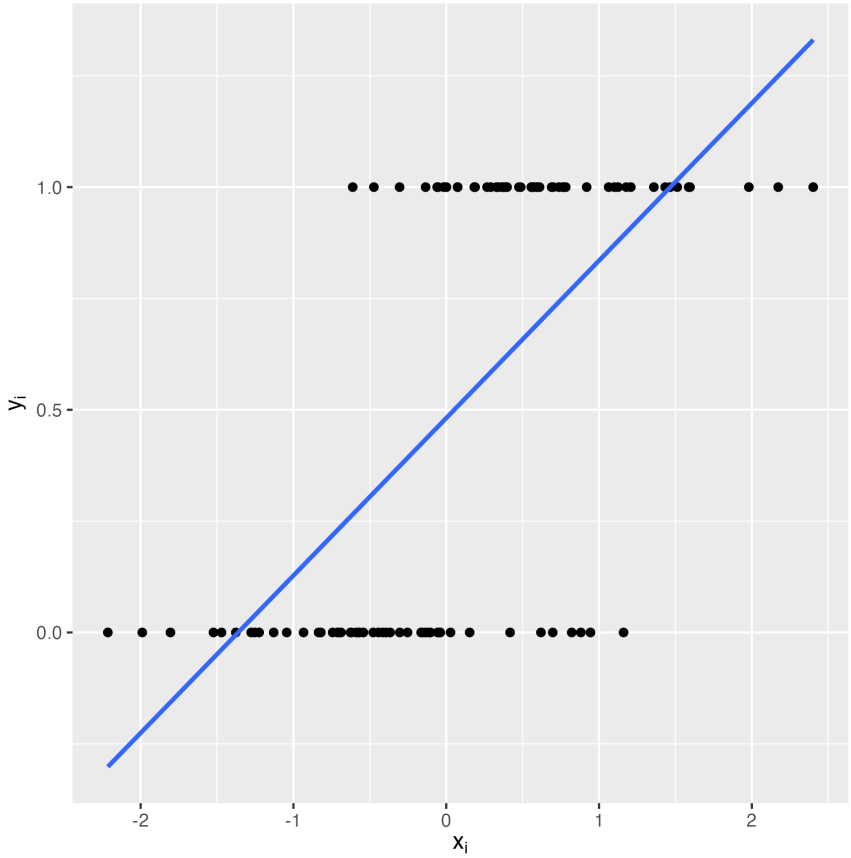
Motivation

- In allen bisher besprochenen Regressionsmodellen war die abhängige Variable (das Kriterium) immer eine stetige Variable.

- In der heutigen Vorlesung werden wir **Regressionsmodelle mit diskreter AV** kennenlernen.

- Wir beschränken uns auf eine diskrete AV mit zwei Ausprägungen in Dummy-Kodierung, d.h. auf binäre Variablen.

Motivation II



- Wir betrachten zunächst ein Beispiel mit einer binären AV und einer stetigen UV.
- Beispiel 1:
  - AV: Diagnose Depression nach ICD-10 ($Y_i = 0$ falls nein, $Y_i = 1$ falls ja)
  - UV: (stetige) negative Selbstbewertung
- Frage: Wie lässt sich der Zusammenhang zwischen den Variablen beschreiben?
- Problem: Ein einfaches lineares Regressionsmodell der Form

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

erscheint nicht sinnvoll, da die meisten der mithilfe der Gerade vorhergesagten Werte gar nicht tatsächlich beobachtet werden könnten.

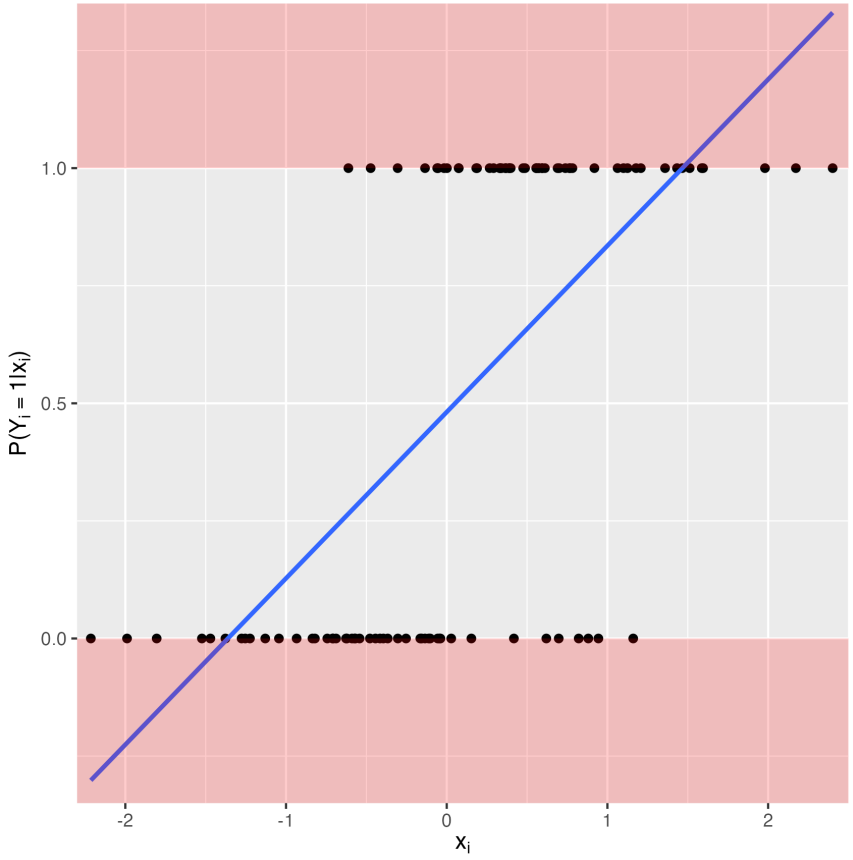
Modell für die bedingten Wahrscheinlichkeiten

- In statistischen Modellen für binäre Kriteriumsvariablen wird in der Regel (anstatt der Variable Y_i selbst) die **bedingte Wahrscheinlichkeit** modelliert, dass die Variable Y_i einen der beiden möglichen Werte (0 und 1) annimmt, bei gegebenen Werten einer Reihe von Prädiktoren $x_{i1}, \dots, x_{ik}$.

- Dabei ist es ausreichend nur $P(Y_i = 1|x_{i1}, \dots, x_{ik})$ zu modellieren, da

$$P(Y_i = \mathbf{1}|x_{i1}, \dots, x_{ik}) = \mathbf{1} - P(Y_i = \mathbf{0}|x_{i1}, \dots, x_{ik})$$

Naive Idee: Lineares Wahrscheinlichkeitsmodell

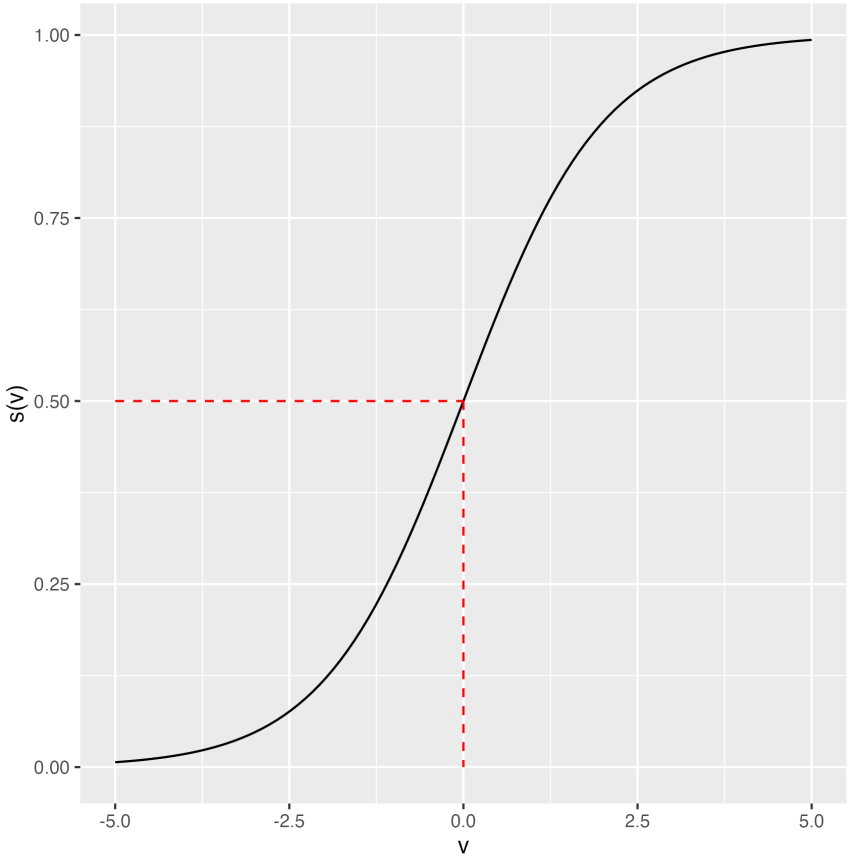


- Wir betrachten erneut das Beispiel mit einem stetigen Prädiktor.
- Beispiel 1:
  - AV: Diagnose Depression nach ICD-10 ($Y_i = 0$ falls nein, $Y_i = 1$ falls ja)
  - UV: (stetige) negative Selbstbewertung
- Ein naiver Ansatz wäre ein lineares Modell für die bedingten Wahrscheinlichkeiten:

$$P(Y_i = 1|x_i) = \alpha + \beta x_i$$

- **Ähnliches Problem:** Eine Wahrscheinlichkeit liegt zwischen 0 und 1, der Wertebereich von $\alpha + \beta x_i$ liegt jedoch zwischen $-\infty$ und $+\infty$

Die logistische Funktion



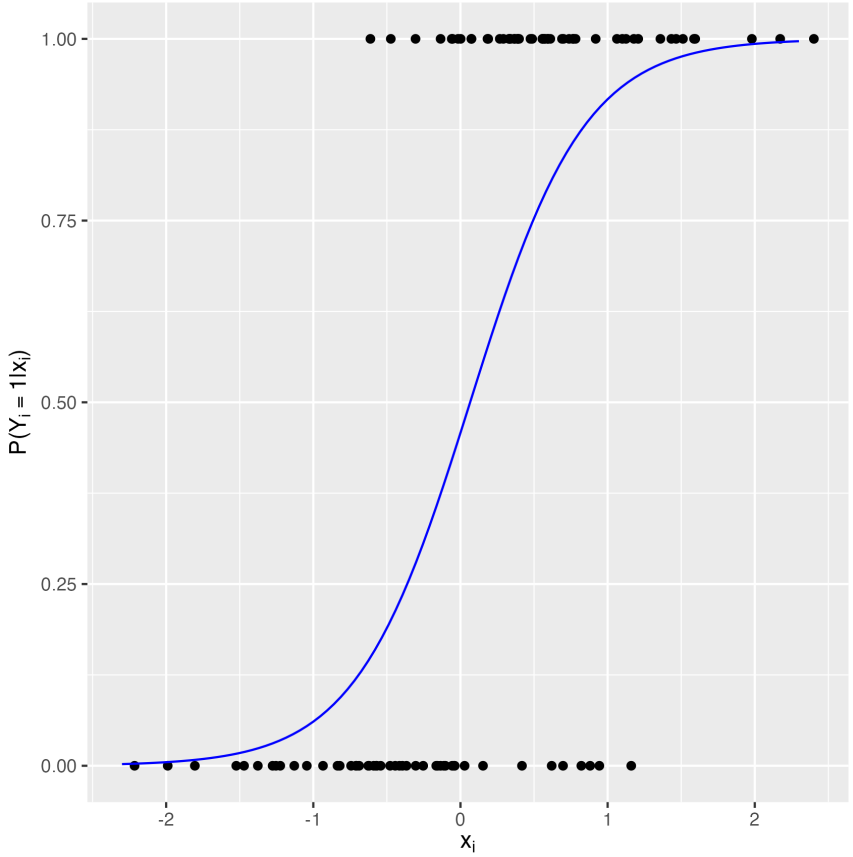
Lösung:

- Transformiere den Wertebereich $]-\infty;+\infty[$ des linearen Wahrscheinlichkeitsmodells in Werte zwischen 0 und 1 (Wertebereich von Wahrscheinlichkeiten)
- Eine Möglichkeit ist die logistische Funktion:

$$s(v) = \frac{e^v}{1 + e^v}$$

- Für $v = 0$ nimmt die logistische Funktion den Wert 0.5 an
- Die Ausprägung von v kann nun in Abhängigkeit von Prädiktoren modelliert werden.

Einfache logistische Regression: Modellgleichung



- Modellgleichung der **einfachen logistischen Regression**:

$$P(Y_i = 1|x_i) = s(\alpha + \beta x_i) = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$$

- Da wir nicht Y_i selbst, sondern die bedingte Wahrscheinlichkeit für $Y_i = 1$ modellieren, benötigen wir keine Fehlervariable, um Zufallseinflüsse zu berücksichtigen.
- Zwei Personen mit gleichen Werten auf der UV haben unter dem Modell zwar die gleiche Wahrscheinlichkeit, einen Wert von 1 auf der AV aufzuweisen. Dies bedeutet jedoch nicht, dass sie auch tatsächlich den gleichen Wert aufweisen.

Interpretation der Parameter

- Modellgleichung der **einfachen logistischen Regression**:

$$P(Y_i = 1|x_i) = s(\alpha + \beta x_i) = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$$

- Die Parameter des Modells sind α und β.
- Wir betrachten zunächst den Parameter β:
  - Falls $\beta = 0$ ist, ist die Wahrscheinlichkeit für $Y_i = 1$ unabhängig von der UV.
  - Falls $\beta > 0$ ist, gilt: Je höher der Wert auf der UV, desto höher die Wahrscheinlichkeit für $Y_i = 1$.
  - Falls $\beta < 0$ ist, gilt: Je höher der Wert auf der UV, desto niedriger die Wahrscheinlichkeit für $Y_i = 1$.
- Für eine Interpretation des genauen Wertes von β müssen wir das Modell umstellen.

Odds Schreibweise I

- Der Bruch aus der Wahrscheinlichkeit für ein bestimmtes Ereignis geteilt durch die Wahrscheinlichkeit für das Gegenereignis, wird in der Statistik als „Odds“ bezeichnet
- In der logistischen Regression interessiert man sich für folgende Odds:

$$\frac{P(Y_i = 1|x_i)}{P(Y_i = 0|x_i)}$$

- Die Odds sind größer als 1, falls das Ereignis $Y_i = 1$ wahrscheinlicher ist als das Ereignis $Y_i = 0$ (bei gleichen Werten für x_i)
- Die Odds sind kleiner als 1, falls das Ereignis $Y_i = 0$ wahrscheinlicher ist als das Ereignis $Y_i = 1$ (bei gleichen Werten für x_i)
- Setzt man die Modellgleichung in die Formel der Odds ein, ergibt sich:

$$\frac{P(Y_i = 1|x_i)}{P(Y_i = 0|x_i)} = \frac{P(Y_i = 1|x_i)}{1 - P(Y_i = 1|x_i)} = \cdots = e^{\alpha+\beta x_i} = e^{\alpha} \cdot e^{\beta x_i}$$

Rechenregeln für Potenzen:

$$e^{a+b} = e^{a} \cdot e^{b}$$

Odds Schreibweise II

- Für ein gegebenes x_i gilt für die Odds:

$$\frac{P(Y_i = 1|x_i)}{P(Y_i = 0|x_i)} = e^{\alpha+\beta x_i} = e^{\alpha} \cdot e^{\beta x_i}$$

- Für $x_i + 1$, also falls sich die UV um eine Einheit erhöht, gilt:

$$\frac{P(Y_i = 1|x_i + 1)}{P(Y_i = 0|x_i + 1)} = e^{\alpha} \cdot e^{\beta(x_i+1)} = e^{\alpha} \cdot e^{\beta x_i+\beta} = e^{\alpha} \cdot e^{\beta x_i} \cdot e^{\beta} = \frac{P(Y_i = 1|x_i)}{P(Y_i = 0|x_i)} \cdot e^{\beta}$$

- Das heißt: Erhöht sich die UV um eine Einheit, erhöhen sich die Odds **um den Faktor** e^{β}.
  - $e^{\beta} = 1$: kein Einfluss der UV auf die Odds
  - $e^{\beta} > 1$: je höher die UV, desto höher die Odds
  - $e^{\beta} < 1$: je höher die UV, desto niedriger die Odds

Hinweis:
$e^0 = 1$
$e^1 \approx 2.718$
$e^{-1} \approx 0.368$

Odds Schreibweise III

- Falls $x_i = 0$ ist, gilt für die Odds

$$\frac{P(Y_i = 1|x_i = 0)}{P(Y_i = 0|x_i = 0)} = e^{\alpha} \cdot e^{\beta \cdot 0} = e^{\alpha} \cdot e^{0} = e^{\alpha} \cdot 1 = e^{\alpha}$$

- Das heißt:
  Für Personen mit einem Wert von 0 auf der UV, sind die Odds gleich e^{α}.

- Ob diese Interpretation sinnvoll ist, hängt genau wie in den linearen Regressionsmodellen davon ab, ob ein UV-Wert von 0 sinnvoll interpretierbar ist.

Odds Schreibweise IV

- Wir haben also eine einigermaßen intuitive Interpretation von e^{α} und e^{β}.
- Wie sieht es mit α und β selbst aus?
- Hierfür müssen wir das Modell noch einmal umstellen.

Log – Odds Schreibweise I

- Nimmt man von den Odds den natürlichen Logarithmus, so erhält man die „Log – Odds“ bzw. „Logits“:

$$ln\left(\frac{P(Y_i = 1|x_i)}{P(Y_i = 0|x_i)}\right)$$

Hinweis:
$\ln(1) = 0$

- Die Log-Odds sind genau dann gleich 0, wenn $P(Y_i = 1|x_i) = 0.5$ ist.
- Sie sind **negativ**, wenn $P(Y_i = 1|x_i) < P(Y_i = 0|x_i)$ ist,
- und sind **positiv**, wenn $P(Y_i = 1|x_i) > P(Y_i = 0|x_i)$ ist.

Log – Odds Schreibweise II

- Einsetzen der Modellgleichung für die Odds in die Definition der Log-Odds ergibt:

$$ln\left(\frac{P(Y_i = 1|x_i)}{P(Y_i = 0|x_i)}\right) = ln\left(e^{\alpha+\beta x_i}\right) = \alpha + \beta x_i$$

- Das logistische Regressionsmodell ist also ein lineares Modell für die Log – Odds.

Log – Odds Schreibweise III

- Die Parameter α und β können damit wie folgt interpretiert werden:
  - α entspricht den Log-Odds für Personen mit einem UV-Wert von 0.
  - Falls sich die UV um eine Einheit erhöht, erhöhen sich die Log-Odds um β.
- Da die Log-Odds weniger intuitiv als die Odds sind, werden jedoch meist e^{α} und e^{β} statt α und β interpretiert.

Einfache logistische Regression: Beispiel 1

- Beispiel 1:
  - AV: Diagnose Depression nach ICD-10 ($Y_i = 0$ falls nein, $Y_i = 1$ falls ja)
  - UV: Negative Selbstbewertung

- Interpretation von e^{α} und e^{β}:
  - Die Odds für eine Depressionsdiagnose bei einer Person mit negativer Selbstbewertung von 0 sind e^{α}.
  - Falls sich die negative Selbstbewertung um eine Einheit erhöht, erhöhen sich die Odds für eine Depressionsdiagnose um den Faktor e^{β}.

- Interpretation von α und β:
  - Die Log-Odds für eine Depressionsdiagnose bei einer Person mit negativer Selbstbewertung von 0 sind α.
  - Falls sich die negative Selbstbewertung um eine Einheit erhöht, erhöhen sich die Log-Odds für eine Depressionsdiagnose um β.

Multiple logistische Regression

- Wie bei der linearen Regression lässt sich auch die einfache logistische Regression zur multiplen logistischen Regression erweitern:

$$P(Y_i = 1|x_{i1}, \dots, x_{ik}) = \frac{e^{\alpha+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}}}{1 + e^{\alpha+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}}}$$

- Odds-Schreibweise:

$$\frac{P(Y_i = 1|x_{i1}, \dots, x_{ik})}{P(Y_i = 0|x_{i1}, \dots, x_{ik})} = e^{\alpha+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}}$$

- Log – Odds Schreibweise:

$$ln\left(\frac{P(Y_i = 1|x_{i1}, \dots, x_{ik})}{P(Y_i = 0|x_{i1}, \dots, x_{ik})}\right) = \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

Multiple logistische Regression: Interpretation der Parameter

- Aus der Odds-Schreibweise ergibt sich:
  - e^{α} entspricht den Odds von Personen, die auf **allen UVs** den Wert 0 haben.
  - e^{β_j} entspricht dem Faktor, um den sich die Odds erhöhen, falls sich die UV *j* um eine Einheit erhöht und **alle anderen UVs konstant** bleiben.

- Aus der Log-Odds-Schreibweise ergibt sich:
  - α entspricht den Log-Odds von Personen, die auf **allen UVs** den Wert 0 haben.
  - Falls sich die UV *j* um eine Einheit erhöht und **alle anderen UVs konstant** bleiben, erhöhen sich die Log-Odds um β_j.

Multiple logistische Regression: Beispiel 2 I

- Beispiel 2:
  - AV: Diagnose Depression nach ICD-10 ($Y_i = 0$ falls nein, $Y_i = 1$ falls ja)
  - UV 1: Negative Selbstbewertung
  - UV 2: Abhängigkeitskognitionen

- Interpretation von e^{α}, e^{β_1} und e^{β_2} :
  - Die Odds für eine Depressionsdiagnose bei einer Person mit negativer Selbstbewertung von 0 und Abhängigkeitskognitionen von 0 sind e^{α}.
  - Falls sich die negative Selbstbewertung um eine Einheit erhöht und die Abhängigkeitskognitionen konstant bleiben, erhöhen sich die Odds für eine Depressionsdiagnose um den Faktor e^{β_1}.
  - Falls sich die Abhängigkeitskognitionen um eine Einheit erhöhen und die negative Selbstbewertung konstant bleibt, erhöhen sich die Odds für eine Depressionsdiagnose um den Faktore^{β_2}.

Multiple logistische Regression: Beispiel 2 II

- Beispiel 2:
  - AV: Diagnose Depression nach ICD-10 ($Y_i = 0$ falls nein, $Y_i = 1$ falls ja)
  - UV 1: Negative Selbstbewertung
  - UV 2: Abhängigkeitskognitionen

- Interpretation von α, β_1 und β_2:
  - Die Log-Odds für eine Depressionsdiagnose bei einer Person mit negativer Selbstbewertung von 0 und Abhängigkeitskognitionen von 0 sind α.
  - Falls sich die negative Selbstbewertung um eine Einheit erhöht und die Abhängigkeitskognitionen konstant bleiben, erhöhen sich die Log-Odds für eine Depressionsdiagnose um β_1.
  - Falls sich die Abhängigkeitskognitionen um eine Einheit erhöhen und die negative Selbstbewertung konstant bleibt, erhöhen sich die Log-Odds für eine Depressionsdiagnose um β_2.

Modelle mit diskreten Prädiktoren

- In logistische Regressionsmodelle können genau wie in lineare Regressionsmodelle diskrete Prädiktoren in Dummy-Kodierung aufgenommen werden.

- Beispiel für einen diskreten Prädiktor mit zwei Ausprägungen:

$$D_i = \begin{cases} 1, & \text{falls Person i nicht zur Referenzkategorie gehört} \\ 0, & \text{falls Person i zur Referenzkategorie gehört} \end{cases}$$

$$P(Y_i = 1|d_i) = \frac{e^{\alpha+\beta d_i}}{1 + e^{\alpha+\beta d_i}}$$

Modelle mit diskreten Prädiktoren: Interpretation der Parameter

- Die Interpretation der Parameter dieser Modelle ergibt sich wieder aus den Modellgleichungen in Odds- und Log-Odds-Schreibweise.

- Beispiel für einen diskreten Prädiktor mit zwei Ausprägungen:

- Odds-Schreibweise:
  - e^{α} entspricht den Odds von Personen aus der Referenzkategorie.
  - e^{β} entspricht dem Faktor, um den die Odds in der Nicht-Referenzkategorie höher sind als in der Referenzkategorie.

- Log-Odds-Schreibweise:
  - α entspricht den Log-Odds von Personen aus der Referenzkategorie.
  - β entspricht der Differenz der Log-Odds zwischen Personen aus der Nicht-Referenzkategorie und Personen aus der Referenzkategorie.

Modelle mit diskreten Prädiktoren: Odds Ratio

- Im Beispiel für einen diskreten Prädiktor mit zwei Ausprägungen gilt

$$\frac{P(Y_i = 1|d_i)}{P(Y_i = 0|d_i)} = e^{\alpha} \cdot e^{\beta d_i}$$

- Für eine Person aus der Nicht-Referenzkategorie gilt damit

$$\frac{P(Y_i = 1|1)}{P(Y_i = 0|1)} = e^{\alpha} \cdot e^{\beta \cdot 1} = \frac{P(Y_i = 1|0)}{P(Y_i = 0|0)} \cdot e^{\beta}$$

- Alternative Interpretation: e^{β} entspricht dem sogenannten Odds Ratio

$$\frac{\frac{P(Y_i = 1|1)}{P(Y_i = 0|1)}}{\frac{P(Y_i = 1|0)}{P(Y_i = 0|0)}} = e^{\beta}$$

Modelle mit diskreten Prädiktoren: Risk Ratio

- In Anwendungen liest man manchmal folgende **falsche Interpretation** für das Odds Ratio e^{β}: „Um welchen Faktor ist das Auftreten von $Y_i = 1$ in der Nicht-Referenzkategorie wahrscheinlicher ist als in der Referenzkategorie?“

- Diese Interpretation gilt jedoch nicht für das Odds Ratio, sondern für das sogenannte Risk Ratio:

$$\frac{P(Y_i = 1|1)}{P(Y_i = 1|0)} = \frac{\frac{e^{\alpha+\beta\cdot 1}}{1+e^{\alpha+\beta\cdot 1}}}{\frac{e^{\alpha+\beta\cdot 0}}{1+e^{\alpha+\beta\cdot 0}}} = \frac{\frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}}}{\frac{e^{\alpha}}{1+e^{\alpha}}} = \cdots = \frac{1+e^{-\alpha}}{1+e^{-\alpha-\beta}}$$

- Das Risk Ratio entspricht also dem Faktor, um den die Wahrscheinlichkeit für $Y_i = 1$ in der Nicht-Referenzkategorie höher ist als in der Referenzkategorie.

Modelle mit diskreten Prädiktoren: Beispiel 3

- Beispiel 3:
  - AV: Diagnose Depression nach ICD-10 ($Y_i = 0$ falls nein, $Y_i = 1$ falls ja)
  - UV: Bildung mit Referenzkategorie „kein Abitur“

- Interpretation von e^{α} und e^{β}:
  - Die Odds für eine Depressionsdiagnose bei einer Person ohne Abitur sind e^{α}.
  - Die Odds für eine Depressionsdiagnose sind bei Person mit Abitur um den Faktor e^{β} höher als bei einer Person ohne Abitur.

- Interpretation von α und β:
  - Die Log-Odds für eine Depressionsdiagnose bei einer Person **ohne** Abitur sind α.
  - Die Log-Odds für eine Depressionsdiagnose sind bei einer Person **mit** Abitur um β höher als bei einer Person ohne Abitur.

Statistische Inferenz

- Wir werden die inferenzstatistischen Verfahren für die logistischen Modelle nicht im Detail besprechen, sondern uns auf die Durchführung und Interpretation in R beschränken.

- Wichtig ist in diesem Kontext jedoch, dass Konfidenzintervalle und Hypothesentests in logistischen Regressionsmodellen auch bei erfüllten Annahmen nicht exakt sind. Sie haben nur für großes *n* approximativ die gewünschten Eigenschaften. Alternativ können auch so genannte Bootstrapverfahren (die keine Verteilungsannahme, jedoch große, repräsentative Stichproben benötigen) verwendet werden.

- Also: Große Stichproben!

- Bemerkung: Weitere statistische Verfahren (gerichtete Hypothesentests, Konfidenzintervalle für Kombinationen von Parametern, etc.) können wie bisher auch mit dem multcomp Paket in R berechnet werden.

Vorhersage von AV-Werten I

- Genau wie bei den linearen Regressionsmodellen können wir mithilfe von logistischen Regressionsmodellen unbekannte AV-Werte von Personen mit gegebenen UV-Werten vorhersagen.

- Eine Möglichkeit: Wir sagen für eine Beobachtung i den Wert $\hat{y}_i = 1$ vorher, sofern die mithilfe der Schätzwerte a, b_1, …, b_k geschätzte Wahrscheinlichkeit

$$\hat{P}(Y_i = 1 | x_{i1}, \dots, x_{ik}) = \frac{e^{a+b_1 x_{i1}+\cdots+b_k x_{ik}}}{1 + e^{a+b_1 x_{i1}+\cdots+b_k x_{ik}}}$$

größer oder gleich 0.5 ist.

Vorhersage von AV-Werten II

- Formal ergibt sich also für die Vorhersagewerte:

$$\hat{y}_i = \begin{cases} 1, & \text{falls } \frac{e^{a+b_1 x_{i1}+\cdots+b_k x_{ik}}}{1+e^{a+b_1 x_{i1}+\cdots+b_k x_{ik}}} \geq 0.5 \\ 0, & \text{sonst} \end{cases}$$

- Die logistische Funktion $s(v)$ ist genau dann größer als 0.5, wenn v größer als 0 ist. Damit vereinfacht sich die obige Darstellung zu:

$$\hat{y}_i = \begin{cases} 1, & falls\ a + b_1 x_{i1} + \cdots + b_k x_{ik} \geq 0 \\ 0, & sonst \end{cases}$$

Effektstärken, Stichprobenplanung und Regressionsdiagnostik

- Auch für die logistische Regression spielen in der Praxis Effektstärken, Stichprobenplanung und Regressionsdiagnostik eine wichtige Rolle. Bei allen drei Themen existiert aber häufig keine eindeutig beste Variante, weshalb wir hier auf eine Beschreibung der Methoden verzichten.

- Bemerkung 1: In der logistischen Regression gibt es mehrere Möglichkeiten, eine globale Effektstärke in Anlehnung an ρ^2 in der linearen Regression zu definieren.

- Bemerkung 2: Für die Stichprobenplanung in der logistischen Regression müssen meist individuell Daten simuliert werden. Einfache Methoden wie wie die Verwendung des pwr Pakets in R sind nicht möglich.

- Bemerkung 3: Für die Regressionsdiagnostik in der logistischen Regression gibt es mehrere Möglichkeiten, Residuen mit sinnvollen Eigenschaften zu definieren.

Anwendung 1 I

Beispiel 1:

- AV: Diagnose Depression nach ICD-10 ($Y_i = 0$ falls nein, $Y_i = 1$ falls ja)

- UV: Negative Selbstbewertung z-standardisiert (d.h. ein UV-Wert von 0 entspricht einer durchschnittlichen negativen Selbstbewertung, ein Wert von 1 entspricht einer negativen Selbstbewertung die eine Standardabweichung über dem Durchschnitt liegt)

Anwendung 1 II

Punktschätzwerte und Hypothesentests:

```
Call:
glm(formula = icd_dep ~ scale(fie_nsb), family = "binomial",
       data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6424  -1.0651  -0.7456  1.1020   1.7119

Coefficients:
                Estimate  Std. Error   z value      Pr(>|z|)
(Intercept)      -0.03158     0.15161    -0.208     0.835
scale(fie_nsb)    0.63843     0.16061     3.975     0.0000703 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Schätzwerte für α und β

p-Werte der Hypothesentests für

$H_0: \alpha = 0$
$H_1: \alpha \neq 0$

und

$H_0: \beta = 0$
$H_1: \beta \neq 0$

- Wir gehen davon aus, dass es einen Zusammenhang zwischen negativer Selbstbewertung und einer Depressionsdiagnose gibt.

Anwendung 1 III

Konfidenzintervalle für e^{α} und e^{β}:

```
                   2.5 %     97.5 %
(Intercept)     0.7192067   1.304806
scale(fie_nsb)  1.3940932   2.622550
```

- Wir gehen davon aus, dass die Odds für eine Depressionsdiagnose bei einer Person mit durchschnittlicher negativer Selbstbewertung zwischen 0.72 und 1.30 liegen.

- Wir gehen davon aus, dass die Odds für eine Depressionsdiagnose um den Faktor 1.39 bis 2.62 steigen, falls die negative Selbstbewertung um eine Standardabweichung steigt.

Anwendung 2 I

Beispiel 2:

- AV: Diagnose Depression nach ICD-10 ($Y_i = 0$ falls nein, $Y_i = 1$ falls ja)
- UV 1: Negative Selbstbewertung z-standardisiert
- UV 2: Abhängigkeitskognitionen z-standardisiert

Anwendung 2 II

Ergebnis der Punktschätzung und Hypothesentests:

```
Call:
glm(formula = icd_dep ~ scale(fie_nsb) + scale(fie_abk), family = "binomial",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6201  -1.0546  -0.7459   1.0990   1.7272

Coefficients:
                Estimate  Std. Error   z value   Pr(>|z|)
(Intercept)     -0.03006     0.15184    -0.198   0.843072
scale(fie_nsb)   0.69224     0.17868     3.874   0.000107 ***
scale(fie_abk)  -0.12328     0.17137    -0.719   0.471902
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Interpretation der Hypothesentests?

Anwendung 2 III

Konfidenzintervalle für e^{α}, e^{β_1} und e^{β_2}:

```
                   2.5 %     97.5 %
(Intercept)     0.7200147   1.307463
scale(fie_nsb)  1.4229813   2.874758
scale(fie_abk)  0.6283923   1.234167
```

- Wir gehen davon aus, dass die Odds für eine Depressionsdiagnose bei Personen mit durchschnittlicher negativer Selbstbewertung und durchschnittlichen Abhängigkeitskognitionen zwischen 0.72 und 1.31 liegen.
- Wir gehen davon aus, dass sich die Odds für eine Depressionsdiagnose um den Faktor 1.42 bis 2.87 erhöhen, falls sich die negative Selbstbewertung um eine Standardabweichung erhöht und die Abhängigkeitskognitionen konstant bleiben.
- Wir gehen davon aus, dass sich die Odds für eine Depressionsdiagnose um den Faktor 0.63 bis 1.23 erhöhen, falls sich die Abhängigkeitskognitionen um eine Standardabweichung erhöhen und die negative Selbstbewertung konstant bleibt.

Anwendung 3 I

Beispiel 3:

- AV: Diagnose Depression nach ICD-10 ($Y_i = 0$ falls nein, $Y_i = 1$ falls ja)
- UV: Bildung mit Referenzkategorie „kein Abitur“

Anwendung 3 II

Punktschätzwerte und Hypothesentests:

```
Call:
glm(formula = icd_dep ~ bildung, family = "binomial", data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.330  -1.078  -1.078   1.280   1.280

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -0.2373      0.1787   -1.328    0.1842
bildung        0.5901      0.3121    1.891    0.0587 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Interpretation der Hypothesentests?

Anwendung 3 III

Konfidenzintervalle für e^{α} und e^{β}:

```
                2.5 %     97.5 %
(Intercept)  0.5538018   1.117885
bildung      0.9826203   3.351317
```

- Wir gehen davon aus, dass die Odds für eine Depressionsdiagnose bei Personen ohne Abitur zwischen 0.55 und 1.12 liegen.

- Wir gehen davon aus, dass die Odds für eine Depressionsdiagnose bei Personen mit Abitur um den Faktor 0.98 bis 3.35 höher sind als bei Personen ohne Abitur.

Anwendung 3 IV

- Interessiert man sich für das Risk Ratio, erhält man den Punktschätzwert, indem man die Schätzwerte für α und β aus der logistischen Regression in die Formel für das Risk Ratio einsetzt:

$$\frac{\hat{P}(Y_i = 1|1)}{\hat{P}(Y_i = 1|0)} = \frac{1 + e^{-a}}{1 + e^{-a-b}} = \frac{1 + e^{-(-0.24)}}{1 + e^{-(-0.24)-0.59}} = 1.33$$

- Bemerkung: Ein Konfidenzintervall für das Risk Ratio lässt sich nicht ohne weiteres berechnen. Hierfür können jedoch „händisch“ konstruierte Bootstrap Percentilintervalle verwendet werden (siehe folgender Exkurs).

Exkurs: Anwendung 3 V


```
risk_ratio <- function(a, b) {(1 + exp(-a)) / (1 + exp(-a - b))}
stat <- function(dat, i) {
          fit <- glm(icd_dep ~ bildung, data = dat[i,], family = "binomial")
          risk_ratio(coef(fit)[1], coef(fit)[2])
          }
set.seed(1)
bootsamples <- boot(Daten, statistic = stat, R = 10000)
boot.ci(bootsamples, conf = 0.95, type = "perc")
```

```
> BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
> Based on 10000 bootstrap replicates
>
> CALL :
> boot.ci(boot.out = bootsamples, conf = 0.95, type = "perc")
>
> Intervals :
> Level     Percentile
> 95%   ( 0.989,  1.773 )
> Calculations and Intervals on Original Scale
```

- Wie gehen davon aus, dass eine Depressionsdiagnose bei Personen mit Abitur um den Faktor 0.989 bis 1.773 wahrscheinlicher ist als bei Personen ohne Abitur.

Anwendung 4 I

Beispiel 4:

- AV: Diagnose Depression nach ICD-10 ($Y_i = 0$ falls nein, $Y_i = 1$ falls ja)
- UV1: Bildung mit Referenzkategorie „kein Abitur“
- UV 2: Negative Selbstbewertung z-standardisiert
- UV3: Interaktion zwischen Bildung und Negativer Selbstbewertung

Anwendung 4 II

Punktschätzwerte und Hypothesentests:

```
Call:
glm(formula = icd_dep ~ bildung * scale(fie_nsb), family = "binomial",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8042  -1.0474  -0.7464   1.1075   1.7313

Coefficients:
                      Estimate  Std. Error   z value  Pr(>|z|)
(Intercept)            -0.2473      0.1845    -1.340   0.18020
bildung                 0.6797      0.3374     2.014   0.04397 *
scale(fie_nsb)          0.5444      0.2016     2.700   0.00694 **
bildung:scale(fie_nsb)  0.2779      0.3478     0.799   0.42432
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Interpretation der Hypothesentests?

Anwendung 4 III

Konfidenzintervalle für e^{α}, e^{β_1}, e^{β_2} und e^{β_3}:

```
                           2.5 %     97.5 %
(Intercept)             0.5416184   1.119126
bildung                 1.0277969   3.885598
scale(fie_nsb)          1.1733143   2.598514
bildung:scale(fie_nsb)  0.6783329   2.681683
```

- Wir gehen davon aus, dass die Odds für eine Depressionsdiagnose bei Personen ohne Abitur mit einer durchschnittlichen negativen Selbstbewertung zwischen 0.54 und 1.12 liegen.

- Wir gehen davon aus, dass die Odds für eine Depressionsdiagnose bei einer durchschnittlichen negativen Selbstbewertung für Personen mit Abitur um den Faktor 1.03 bis 3.89 höher sind als bei Personen ohne Abitur.

Anwendung 4 IV

Konfidenzintervalle für e^{α}, e^{β_1}, e^{β_2} und e^{β_3}:

```
                          2.5 %     97.5 %
(Intercept)            0.5416184   1.119126
bildung                1.0277969   3.885598
scale(fie_nsb)         1.1733143   2.598514
bildung:scale(fie_nsb) 0.6783329   2.681683
```

- Wir gehen davon aus, dass sich bei Personen ohne Abitur die Odds für eine Depressionsdiagnose um den Faktor 1.17 bis 2.60 erhöhen, falls sich die negative Selbstbewertung um eine Standardabweichung erhöht.

- Wir gehen davon aus, dass sich bei Personen mit Abitur die Odds für eine Depressionsdiagnose um den Faktor 0.68 bis 2.68 stärker erhöhen als bei Personen ohne Abitur, falls sich die negative Selbstbewertung um eine Standardabweichung erhöht.

Exkurs: Anwendung 4 V

Berechnung eines 95% Bootstrap Percentilintervalls für das Risk Ratio:

```
risk_ratio <- function(a, b) {(1 + exp(-a)) / (1 + exp(-a - b))}
stat <- function(dat, i) {
          fit <- glm(icd_dep ~ bildung*scale(fie_nsb), data = dat[i,], family = "binomial")
          risk_ratio(coef(fit)[1], coef(fit)[2])
          }
set.seed(1)
bootsamples <- boot(Daten, statistic = stat, R = 10000)
boot.ci(bootsamples, conf = 0.95, type = "perc")

> BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
> Based on 10000 bootstrap replicates
>
> CALL :
> boot.ci(boot.out = bootsamples, conf = 0.95, type = "perc")
>
> Intervals :
> Level     Percentile
> 95%   ( 1.014,  1.901 )
> Calculations and Intervals on Original Scale
```

Exkurs: Anwendung 4 VI

- Da die negative Selbstbewertung z-standardisiert wurde, entspricht das Risk Ratio dem Faktor, um den die Wahrscheinlichkeit für eine Depressionsdiagnose bei einer Person mit Abitur höher ist als bei einer Person ohne Abitur, wenn beide Personen eine durchschnittliche negative Selbstbewertung aufweisen.

Interpretation konkret:

- Wir gehen davon aus, dass eine Depressionsdiagnose bei Personen mit Abitur und einer durchschnittlichen negativen Selbstbewertung um den Faktor 1.01 bis 1.90 wahrscheinlicher ist als bei Personen ohne Abitur und einer durchschnittlichen negativen Selbstbewertung.