Logistische Regression

Big Picture: Statistisches Lernen

- Supervised Learning
  (X,Y)
  - Regression
    Y kontinuierlich
    VL 2, 3, 4, 6, 7, 8
  - Klassifikation
    Y kategoriell
    VL 5
- Unsupervised Learning
  X
  VL 9, 13

Bsp: Ausfall («Default») bei Kreditkartenfirma kaggle

Bsp: Kreditausfall

limited
Risky Business
Predict the risk of customer credit default

Private	44	14 months ago

Predict the risk of customer credit default

Improve credit risk models by predicting the probability of default on a consumer credit product in the next 18 months. More accurate credit risk evaluations allow issuers of credit to be able to responsibly extend and manage credit lines for their customers. The goal of this contest is to make the most accurate ranking of customers' credit risk given key data related to customer behavior.

How well can you detect the early signs of credit distress?

Enter Now!

This competition is only open to Masters-level participants who meet the eligibility criteria. Visit the Enter the Competition page to view the eligibility criteria and request entrance.

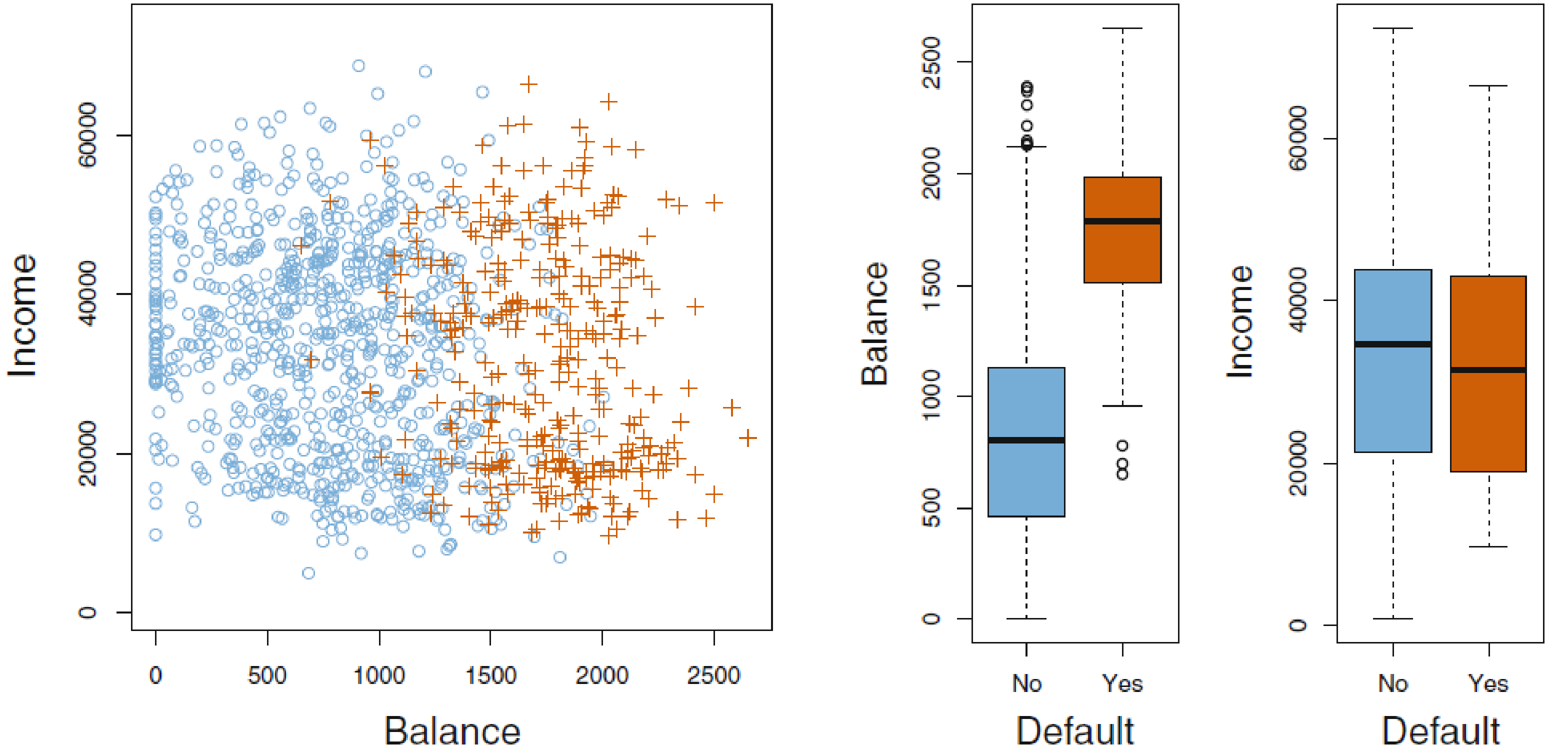
Bsp: Kreditausfall mit simulierten Daten

Simulierter Datensatz von 10000 Kunden

- ‘default’: Kam es zu Kreditausfall ? (Yes / No)
- ‘student’: Ist Kunde Student ? (Yes / No)
- ‘balance’: Monatliche Schulden ? (in USD)
- ‘income’: Jährl. Einkommen des Kunden (in USD)

- Wie hängt Kreditausfall mit erklärenden Variablen zusammen ?
- Kann man Kreditausfall gut vorhersagen ?

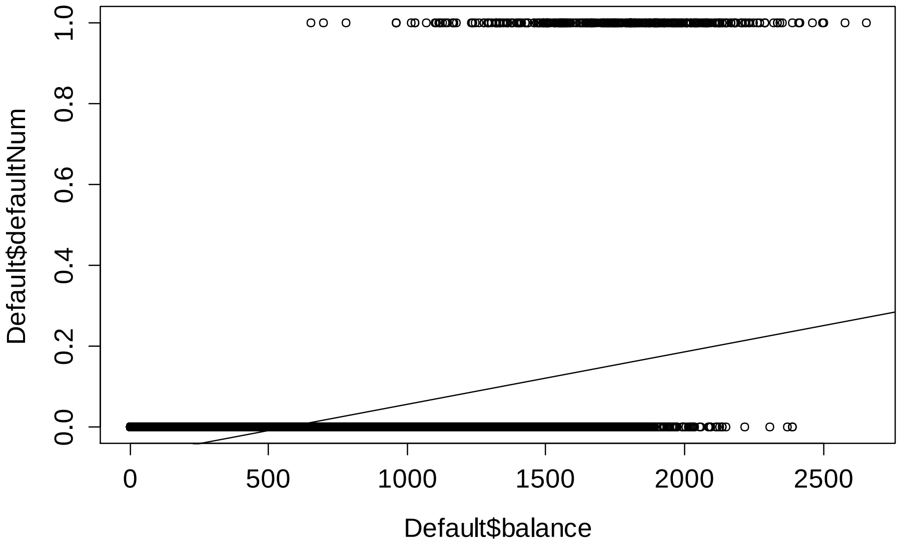
Kreditausfall: Überblick



Orange: Student, blau: Kein Student

Erster Versuch: Einfache Lineare Regression

- Dummy coding: 0 – kein Kreditausfall, 1 – Kreditausfall
- lm: DefaultNumeric ~ balance



Problem mit einfacher linearer Regression

- Was bedeutet kontinuierliche Zielgrösse ?
  Evtl. eine Art “Wahrscheinlichkeit”?
- Was bedeuten Werte grösser als 1 und kleiner als 0 ?

- → Schwierig zu interpretieren

- Besseres Werkzeug: Logistische Regression
  - Modelliere Wa. für Kreditausfall gegeben erklärende Var.
  - Vorhersage “Kreditausfall”, z.B. falls Wa. grösser 50%

Big picture: Generalized Linear Models (GLMs)

- **Bisher:** Population wird mit einer Verteilung beschrieben
  Bsp: Medikament wirkt mit 30% Wa. Wie wa. ist es, dass bei 10 Patienten mindestens 5 gesund werden?
  $$X \sim Bin(10, \pi = 0.3)$$
- **Neu:** Parameter dieser Verteilung hängt von erklärenden Variablen ab.
  Bsp: Wirkwa. hängt von Dosis D ab. Bei welcher Dosis werden im Mittel 90% der Patienten gesund?
  $X \sim Bin(10, \pi)$ *und* $\pi = f(D)$
- **Generalized Linear Models:** Zshg zw. erklärenden Variablen (z.B. Dosis) und Parametern einer Verteilung (z.B. Erfolgswa. in Binomialverteilung)

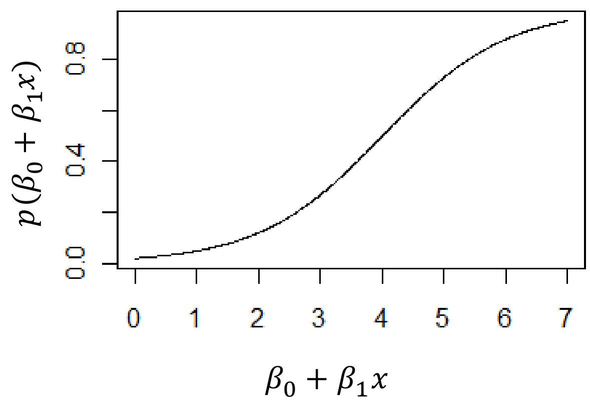
GLM ist ein zweistufiges Modell

- Bsp: Lineare Regression
  D: $\mu(x) = \beta_0 + \beta_1 x$ ← deterministisch
  S: $Y \sim N(\mu(x), \sigma^2)$ ← stochastisch

Weiteres Bsp: Logistische Regression

- X: Dosis des Wirkstoffs; n: Patienten, p: Genesungswa.
  Y: Anz. gesunder Patienten nach Behandlung
- S: $Y \sim Bin(n, p(x))$
- Zshg. zwischen p und x z.B.:

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$



- Kann man umformen zu:
  D: $\underbrace{\log\left(\frac{p(x)}{1-p(x)}\right)}_{\text{Logistische Funktion}} = \beta_0 + \beta_1 x$
- “Logistische Regression”, “Binomialregression”

Wdh: Odds, log-odds, odds ratio

- Alternative zu Wahrscheinlichkeit
- $odds(A) = \frac{P(A)}{1-P(A)} \rightarrow P(A) = \frac{odds(A)}{1+odds(A)}$
- Log-odds: $\log\left(\frac{P(A)}{1-P(A)}\right)$
- Log-odds und odds wachsen monoton mit der Wahrscheinlichkeit:
  $P(A)$ grösser ~ $odds(A)$ grösser ~ $log - odds(A)$ grosser
- Odds-ratio: $\frac{odds(A|B)}{odds(A|B^c)}$

Aus: “Statistical Sleuth”, Kap 18.1.3

Wdh: Bsp odds-ratio

	Lungenkrebs (L)	Kontrolle	Total
Raucher (R)	83	72	155
Nichtraucher	3	14	17
Total	86	86	172

- Alternative zu Wa.: odds, log-odds, odds-Ratio (OR)
- $P(L|R) \approx \frac{83}{155} = 0.54;\ P(L|R^C) \approx \frac{3}{17} = 0.18$

$$Risk - ratio = \frac{P(L|R)}{P(L|R^C)} = 3$$

- $odds(L|R) = \frac{P(L|R)}{P(L^C|R)} = \frac{P(L|R)}{1-P(L|R)} \approx \frac{0.54}{0.46} = 1.17;\ odds(L|R^C) \approx 0.22$
- $log - odds(L|R) = \log\big(odds(L|R)\big) = \log(1.17) \approx 0.157$ (zur Basis e)
- $OR = \frac{odds(L|R)}{odds(L|R^C)} = \frac{1.17}{0.22} = 5.33$

“Odds ratio”

Wie genau?

“Die odds an Lungenkrebs zu erkranken sind für Raucher ca. 5 mal grösser.”

GLMs

- Lineare Regression:
  D: $\mu(x) = \beta_0 + \beta_1 x$
  S: $Y \sim N(\mu(x), \sigma^2)$

Link Funktion: Identität $g(x) = x$
Linearer (in β) Prädiktor

- Logistische Regression:
  D: $\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$
  S: $Y \sim Bin(1, p(x))$

Link Funktion: log-odds
Linearer (in β) Prädiktor

- Allgemein:
  D: $g(\theta(x)) = \beta_0 + \beta_1 x$
  S: $Y \sim F(\theta)$

Link Funktion: g()
Linearer (in β) Prädiktor

- In dieser VL beschränken wir uns auf Lineare und Logistische Regression als Bsp von GLMs

Parameterschätzung: Maximum Likelihood

- Gegeben: x_i kontinuierlich, y_i binär (z.B. 0/1)
- Gesucht: $\widehat{\beta_0}, \widehat{\beta_1}$
- Prinzip: Suche $\widehat{\beta_0}, \widehat{\beta_1}$, sodass *likelihood function* l maximal ist:

$$l(\widehat{\beta_0}, \widehat{\beta_1}) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} \left(1 - p(x_j)\right)$$

- In R: Funktion 'glm()' mit der Option 'family = binomial'

```
fm1 <- glm(default ~ balance, data = Default, family = binomial)
```

```
fm2 <- glm(default ~ student, data = Default, family = binomial)
```

Default ~ balance: Interpretation

β_0

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
```

β_1

- Wenn man 'balance' um eine Einheit erhöht, erhöhen sich die ***log-odds*** um 0.0055 (95%-VI: $0.0055 \pm 2 * 0.00022$)
- Wenn man 'balance' um eine Einheit erhöht, erhöhen sich die ***odds*** um den **Faktor** $\exp(0.0055)$ (95%-VI: $\exp(0.0055 \pm 2 * 0.00022)$) d.h., 1.0055 (95%-VI: $(1.00507, 1.00596)$)

- Eine einfache Aussage über die Änderung der Wahrscheinlichkeit ist nicht möglich !

Vorhersage

- Vorhersage für gegebenes x möglich für:
  - log-odds
  - odds
  - Wahrscheinlichkeit

- Nur Vertrauensintervall (kein Vorhersageintervall) möglich: Warum?

Default ~ student: Interpretation

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
```

- Wenn man ‘student’ vom Status ‘No’ zum Status ‘Yes’ ändert, *erhöhen* sich die ***log-odds*** um 0.405 (95%-VI: $0.405 \pm 2 * 0.115$)
- Wenn man ‘student’ vom Status ‘No’ zum Status ‘Yes’ ändert, *erhöhen* sich die ***odds*** um den **Faktor** $\exp(0.405)$ (95%-VI: $\exp(0.405 \pm 2 * 0.115)$)
  d.h., 1.50 (95%-VI: $(1.19, 1.89)$)
- Dieser Faktor (1.50) ist gerade das **odds-ratio** (inkl. VI);
  d.h., bei binären Faktoren als erklärende Variablen, kann man das odds-ratio mit $\exp()$ aus dem summary output ablesen
- Eine einfache Aussage über die Änderung der Wahrscheinlichkeit ist nicht möglich !

Multiple Logistische Regression

- $p(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + etc.)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 x_2 + etc.)}$
- Erklärende Variablen können kontinuierlich oder diskret (Faktoren) sein
- Interaktion ist möglich genau wie bei Linearer Regression

- Bsp: $\mathrm{d}efault \sim balance + income + student$

Default ~ balance + income + student Interpretation 1

Wenn man 'balance' um eine Einheit erhöht
und 'income' und 'student' gleich lässt,
erhöhen sich die ***log-odds*** um
0.0057 (95%-VI: $0.0055 \pm 2 * 0.00023$)

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
```

Widerspruch zu Bsp
Default ~ student ?

Wenn man 'student' vom Status 'No' zum Status 'Yes' ändert,
und 'income' und 'balance' gleich lässt,
erhöhen sich die ***log-odds*** um
-0.647 (95%-VI: $-0.647 \pm 2 * 0.236$)

Einfache vs. multiple Regression: Paradox ?

- Einfache Regression: “Totaler Effekt”

default ~ student: Studenten haben **grössere** Default-Wa.

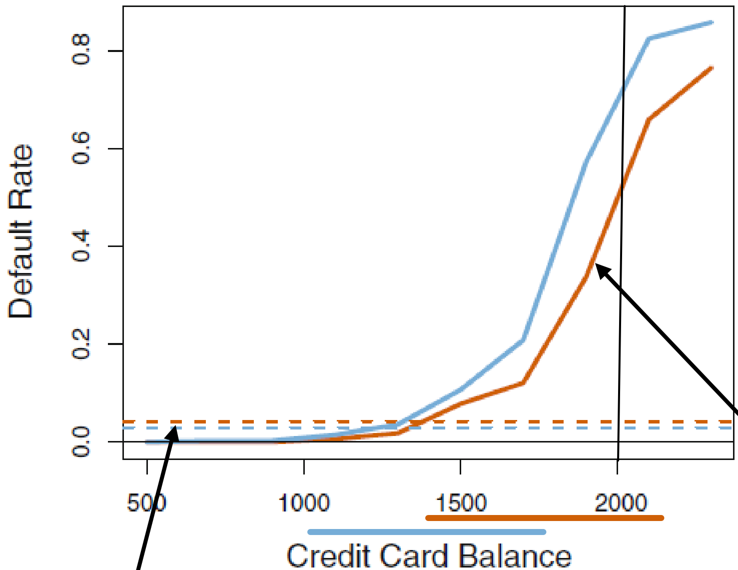
- Multiple Regression: “Bereinigter Effekt”

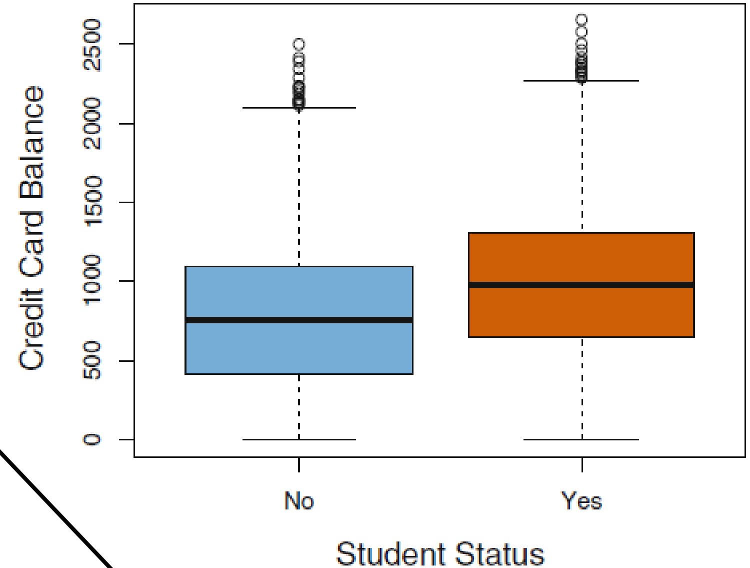
default ~ student + income + balance:
Studenten haben bei gleichem ‘income’ und ‘balance’ **kleinere** Default-Wa.

Erklärung (vgl. Simpson-Paradox)

Wer viele Schulden hat, fällt eher aus

Studenten haben viele Schulden





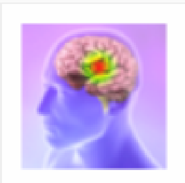
Ein Student hat eine grössere Ausfalls-Wa. als ein Nicht-Student (Einfache Regression)

Ein Student hat eine kleinere Ausfalls-Wa. als ein Nicht-Student *mit gleichem Schuldenbetrag* (Multiple Regression)

Klassifikation

- Vorhersage von Logistischer Regression liefert Wa., dass Beobachtung zu Klasse “Default” gehört
- Falls Wa. > 50%: Klassifiziere zu Klasse “Default”, sonst Klasse “Kein Default”

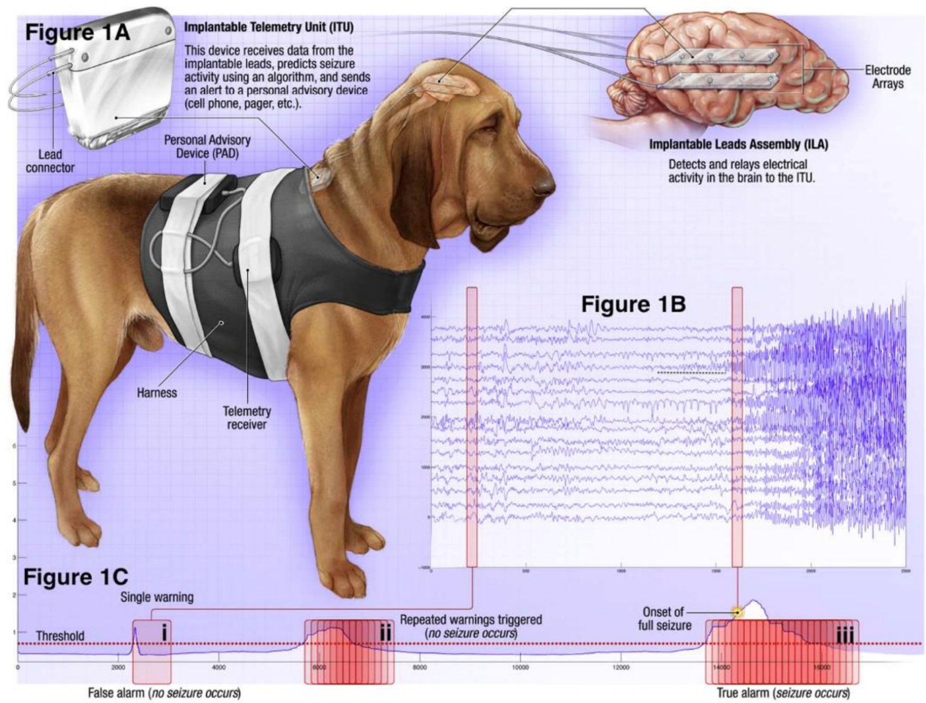
Reales Beispiel: Epilepsie



American Epilepsy Society Seizure Prediction Challenge Predict seizures in intracranial EEG recordings	$25,000	504	8 months ago

EEG Daten → Anfall steht bevor: Ja / Nein

Logistische Regression ↑



Qualität der Klassifikation

- Beurteilung mit CV um Overfitting zu vermeiden
- "Confusion matrix":

Beobachtung

Vorhersage

```
cv.predDefault   No  Yes
         FALSE 9626  227
         TRUE    41  106
```

Personen ohne Kreditausfall werden gut vorhergesagt

Personen mit Kreditausfall werden schlecht vorhergesagt

- Fehlerrate:
  (41 + 227) / 10000 = 0.00268
- Fehlerrate alleine evtl. zu optimistisch; Confusion matrix hat mehr Aussagekraft

Klassifikation bei mehr als 2 Klassen

- Logistische Regression: Zielgrösse binär (z.B. Mann / Frau)
- Falls Zielgrösse mehr als zwei Stufen hat
  (z.B. Augenfarbe blau / braun / grün):
  - Erweiterung von Logistischer Regression möglich (‘one vs. all’)
  - einfacher: Verwende andere, massgeschneiderte Methode (nicht Teil von diesem Kurs)
- Z.B.: Lineare Diskriminanz Analyse (ISLR Kap. 4.4), Nearest-Neighbor Methoden (ISLR Kap. 4.6.5), Random Forest (ISLR Kap. 8)