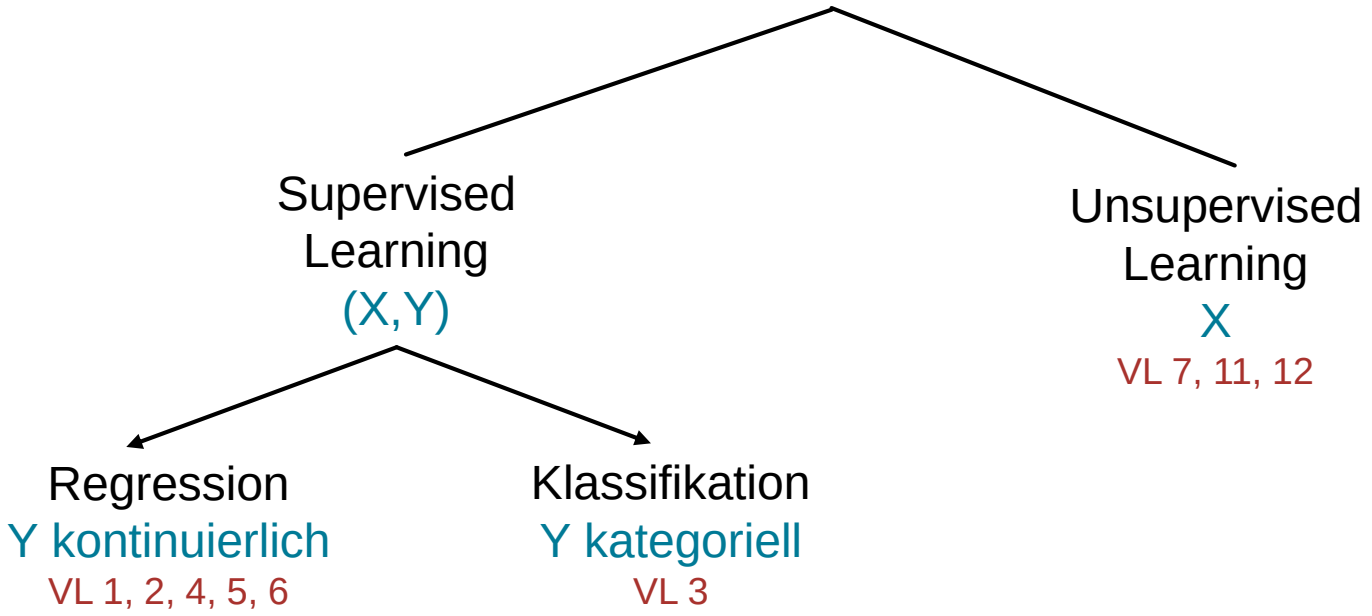
ETH zürich

Logistische Regression

Big Picture: Statistisches Lernen



Bsp. Klassifikation: Schuldenausfall (‘Default’) bei Kreditkartenfirma

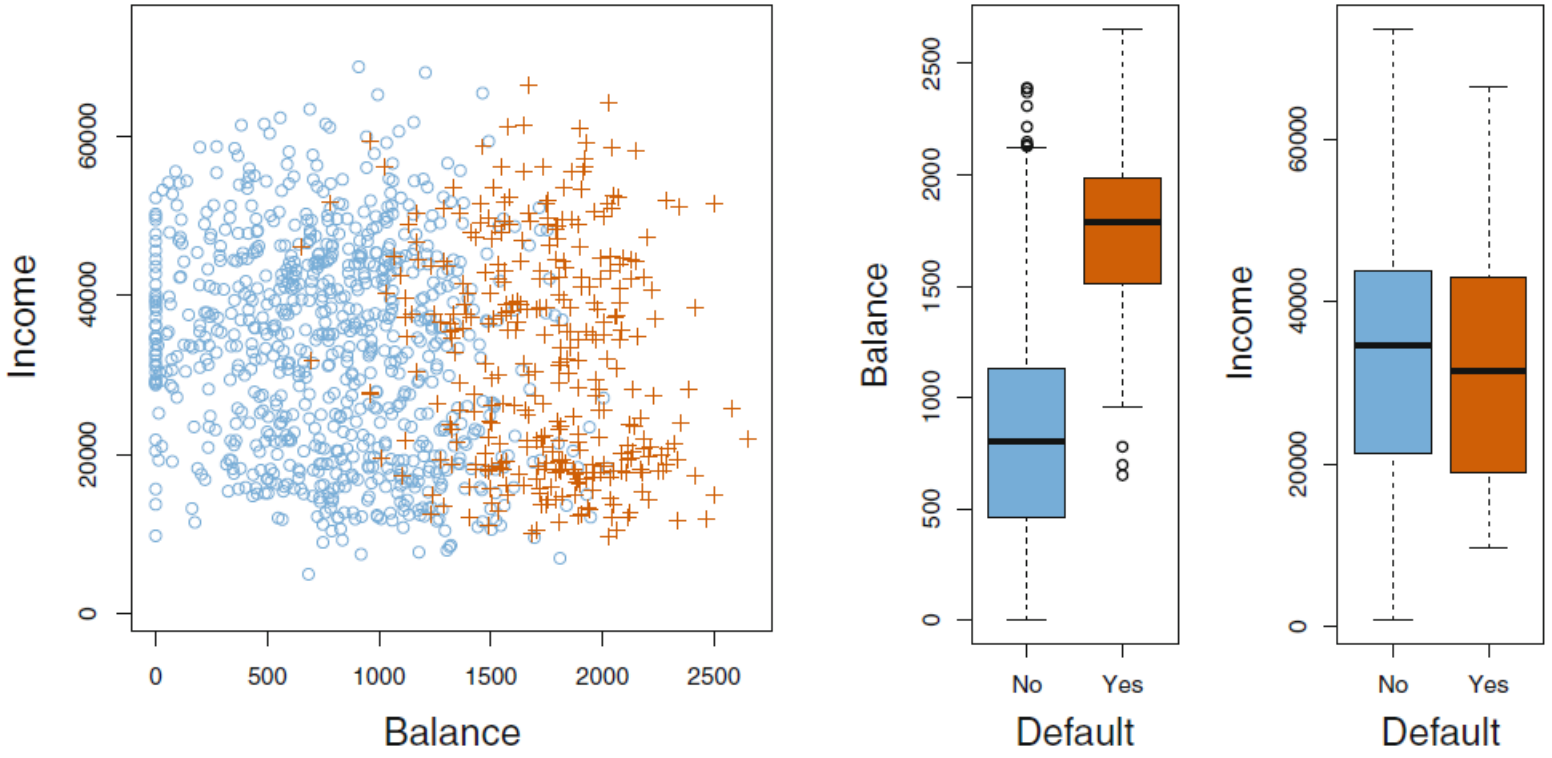
Bsp: Kreditausfall (‘Default’)

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 ?

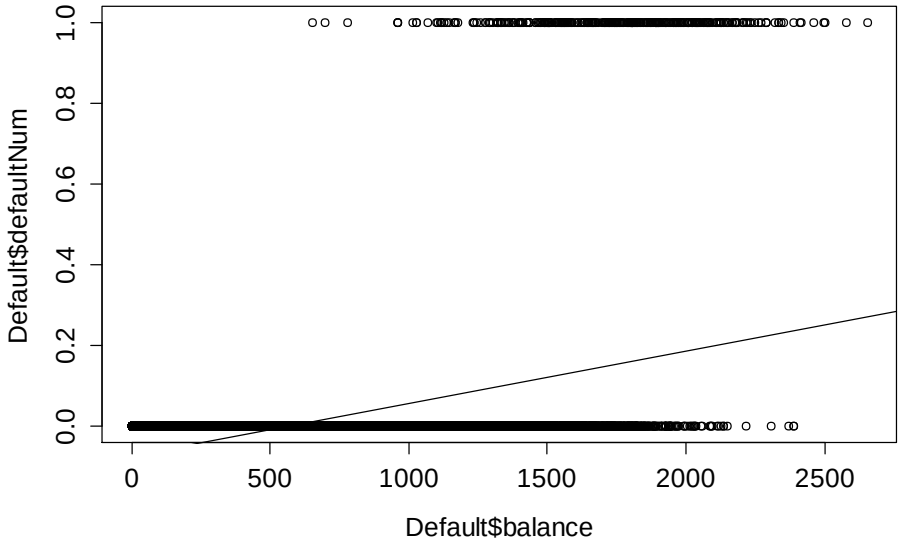
Kreditausfall: Überblick



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)

Wdh: Odds

- Alternative zu Wahrscheinlichkeit
- $odds(A) = \frac{P(A)}{1-P(A)} \rightarrow P(A) = \frac{odds(A)}{1+odds(A)}$
- Log-odds: $\log\left(\frac{P(A)}{1-P(A)}\right)$

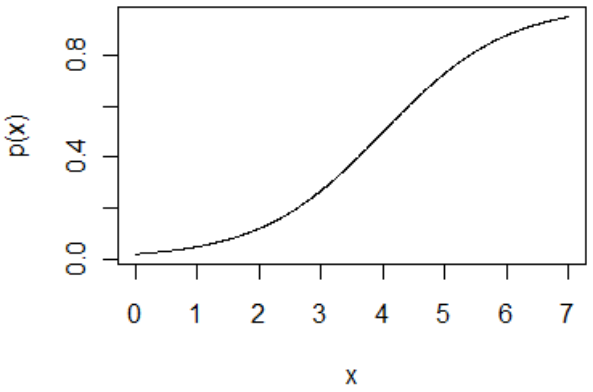
- Log-odds und odds wachsen monoton mit der Wahrscheinlichkeit:

  $P(A)$ grösser ~ $odds(A)$ grösser ~ $log - odds(A)$ grösser

Spezielles GLM: Logistische Regression

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

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

Logistische Funktion

Linear in β's

- "Logistische Regression", "Binomialregression"

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:

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

- 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

```
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 ***
```

- 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 !

Vorhergesagte Wahrscheinlichkeit

Richtig oder falsch?

Angenommen die *balance* ist 1000. Die Wahrscheinlichkeit für *Default* ist dann gemäss unserem Modell etwa 0.5.

Nehmen Sie folgenden R-Output an:

Parameter	Koeffizient
(Intercept)	-10
balance	0.01

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: $(1.19, 1.89)$)

- 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: $default \sim balance + income + student$

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.

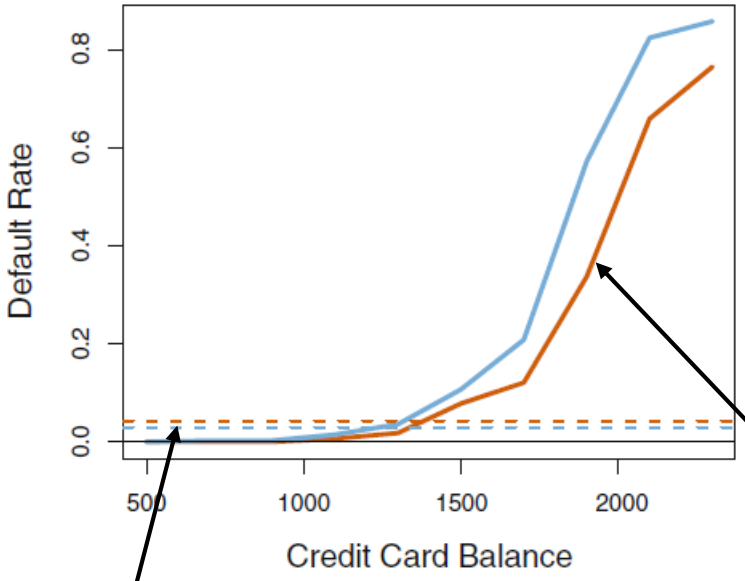
- Multiple Regression: “Bereinigter Effekt”

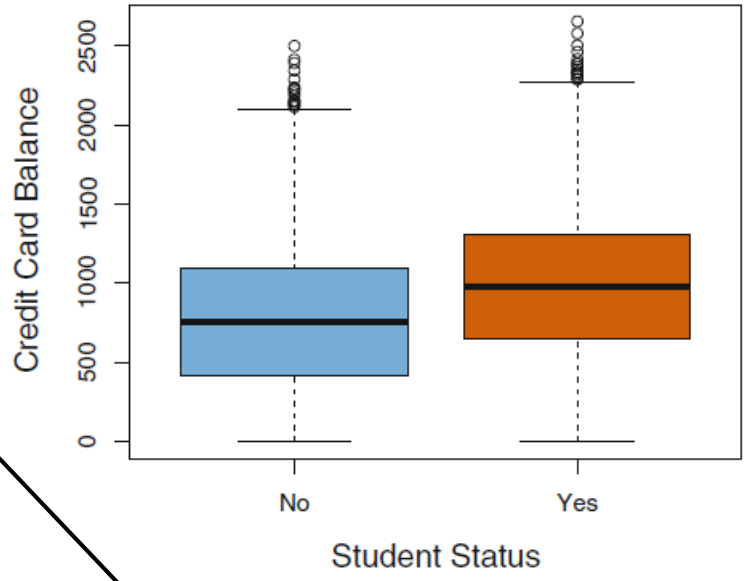
default ~ student + income + balance:
Studenten haben bei gleichem ‘income’ und ‘balance’ **kleinere** Default-Wa.

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)

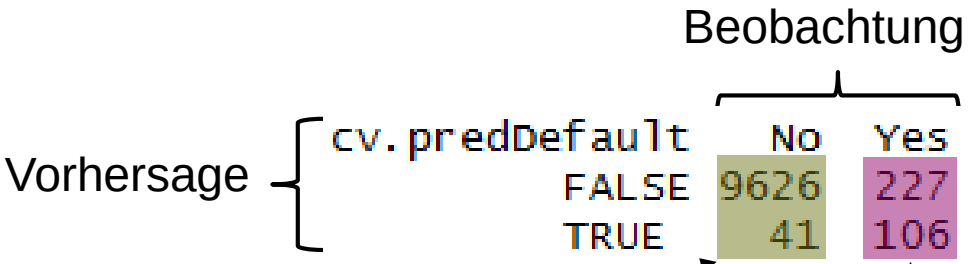
Klassifikation

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

- Einfache aber weit verbreitete Methode
  Bsp: Google Ads

Qualität der Klassifikation

- Beurteilung mit CV um Overfitting zu vermeiden
- “Confusion matrix”:



- Fehlerrate:
  (41 + 227) / 10000 = 0.00268

Personen ohne Kreditausfall werden gut vorhergesagt

Personen mit Kreditausfall werden schlecht vorhergesagt

- Fehlerrate alleine evtl. zu optimistisch; Confusion matrix hat mehr Aussagekraft

Feintuning

- Die Krankheit xyz tritt in der Bevölkerung bei etwa jeder tausendsten Person auf. Es wurde ein Screening für die breite Bevölkerung entwickelt. Die Fehlerrate in der breiten Bevölkerung ist 0.2%. Finden Sie ein Verfahren mit einer besseren Fehlerrate. Welche Fehlerrate kann sicher erreicht werden?

1. 0%
2. 0.001%
3. 0.1%
4. Ohne weitere Infos ist keine Antwort möglich

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)