Logistische Regression

Prof. Dr. Karsten Lübke

SoSe 2017

Modellierung

$$y = f(x) + \epsilon$$

Hier nur für eine unabhängige Variable:

- Lineare Regression: abhängige Variable y numerisch:
  $y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$
- Logistische Regression: abhängige Variable y binär, d. h., kategorial mit zwei Merkmalsausprägungen $y_i \in \{0, 1\}$. p_i sei die Wahrscheinlichkeit, dass $y_i = 1$, dann:
  $logit(p_i) = \ln(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 \cdot x_i$
  - Kunde/ kein Kunde
  - Abwanderung Ja/ Nein
  - Raucher/ Nichtraucher
  - Kreditausfall Ja/ Nein

Analyse Extraversion

Extraversionstest nach Dr. Satow, angepasst von Prof. Dr. Sebastian Sauer.

Fragebogen: http://bit.ly/1HBhKWU

Analyse Extraversionsdaten: Alternative Einlesen

Menü RStudio:

`File -> Import Dataset-> From Excel ...`

Datei “Extraversion.xlsx” auswählen (`Browse`) `-> Import`

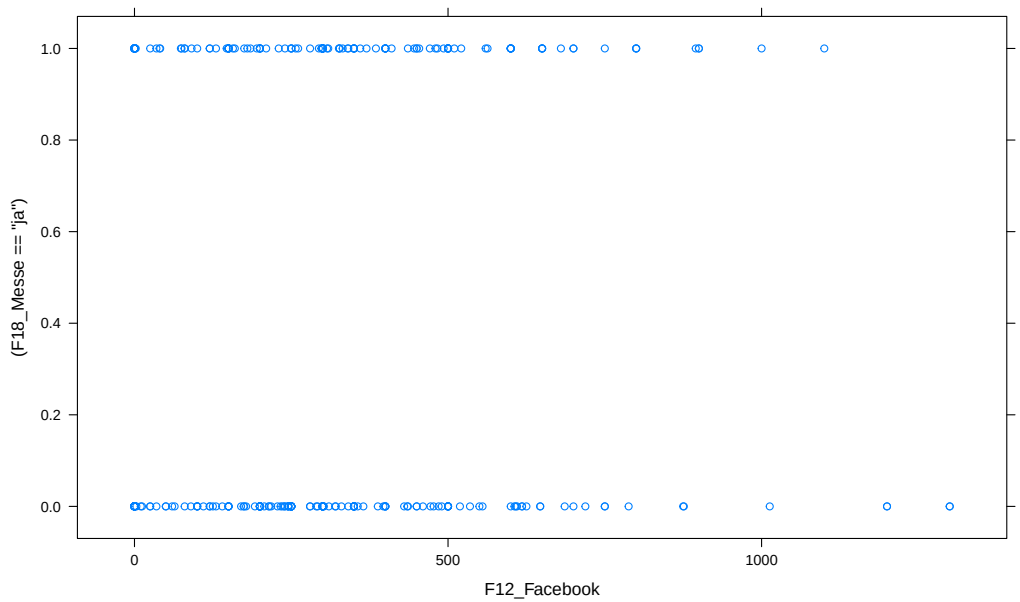
Analyse Extraversionsdaten: Einlesen

```r
# Gggfs. Einmalig vorab installieren
# install.packages("readxl")

# Paket zum Einlesen von Excel Dateien laden
library(readxl)

# Daten einlesen

# Daten "Extraversion.xls" einlesen
# und als Datensatz "Extraversion" in R speichern
# Achtung: Pfad zur Datei anpassen
Extraversion <- read_excel("Extraversion.xlsx")
```

Bereitschaft sich freiwillig zur Messe zu melden

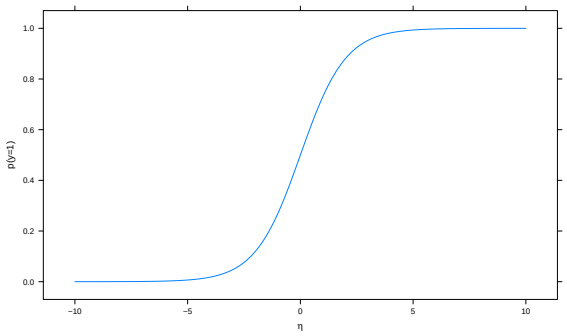
Modellierung durch die Anzahl Facebook-Freunde.

```
library(mosaic)

xyplot( (F18_Messe=="ja") ~ F12_Facebook,
        data = Extraversion)
```



Modellierung Logit

$$p(y=1) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{e^{\beta_0+\beta_1 \cdot x_i}}{1+e^{\beta_0+\beta_1 \cdot x_i}} = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x_i)}}$$



Schätze β anhand der Daten: $\hat{\beta}$:

- $\beta > 0$: Wahrscheinlichkeit steigt
- $\beta < 0$: Wahrscheinlichkeit fällt

Vorbereitung: Modellierung Messebesuch

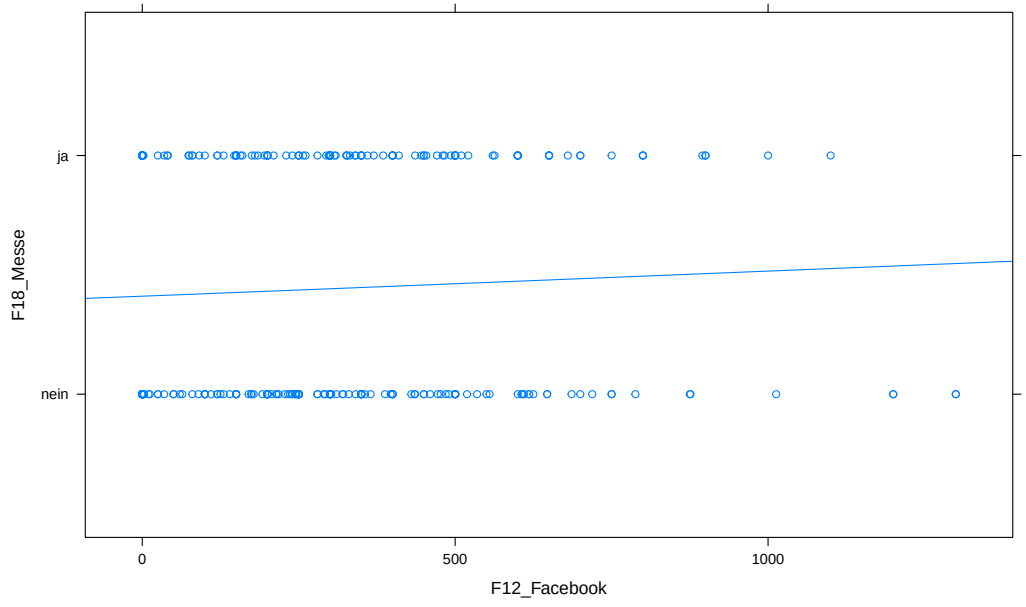
R modelliert y anhand der *Faktorstufen*: In der logistischen Regression ist die erste Ausprägung die 0, alle weiteren 1

```
# Als Faktor definieren
Extraversion$F18_Messe <- factor(Extraversion$F18_Messe)
# Referenzklasse festelgen
Extraversion$F18_Messe <- relevel(Extraversion$F18_Messe,
                                  ref="nein")
# Kontrolle
levels(Extraversion$F18_Messe)
```

```
## [1] "nein" "ja"
```

Logistische Regression: Messebesuch

```
ergglm <- glm(F18_Messe ~ F12_Facebook,
              data = Extraversion,
              family = binomial("logit"))

plotModel(ergglm)
```



Ergebnis Logistische Regression

```
summary(ergglm)
```

```
##
## Call:
## glm(formula = F18_Messe ~ F12_Facebook, family = binomia
##     data = Extraversion)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.259  -1.071  -1.028   1.274   1.334
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.3616264  0.1719854  -2.103   0.0355 *
## F12_Facebook  0.0004245  0.0004381   0.969   0.3326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
```

Regressionskoeffizienten

	Estimate	Std. Error	z value	Pr(>
(Intercept)	-0.3616264	0.1719854	-2.1026571	0.03549
F12_Facebook	0.0004245	0.0004381	0.9689802	0.33255

Übung 1: Regressionskoeffizienten

Welche der folgenden Aussagen stimmt?

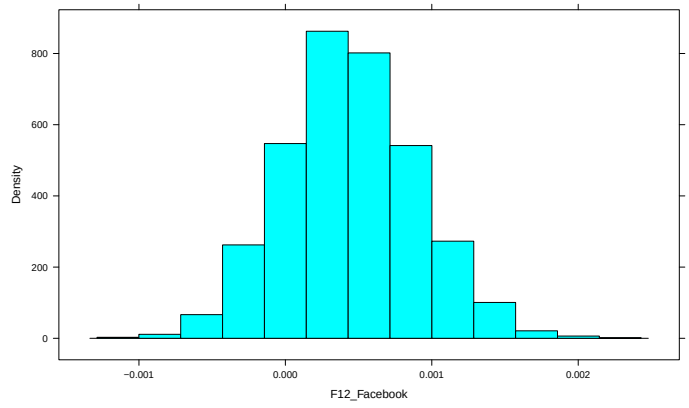
- A: In der Stichprobe steigt die Bereitschaft sich freiwillig zur Messe zu melden mit der Anzahl der Facebook-Freunde
- B: In der Stichprobe sinkt die Bereitschaft sich freiwillig zur Messe zu melden mit der Anzahl der Facebook-Freunde
- C: In der Stichprobe ist die Bereitschaft sich freiwillig zur Messe zu melden unverändert mit der Anzahl der Facebook-Freunde

Bootstrap Regressionkoeffizient (I/II)

```
set.seed(1896)

Bootvtlg <- do(5000) *
  coef(glm(F18_Messe ~ F12_Facebook,
           data = resample(Extraversion),
           family = binomial("logit")))[2]
```

Bootstrap Regressionkoeffizient (II/II)

```
histogram( ~ F12_Facebook, data = Bootvtlg)
```




```
quantile( ~ F12_Facebook, data = Bootvtlg,
          probs = c(0.025, 0.975))
```

```
##           2.5%          97.5%
## -0.0004170976  0.0013528616
```

Übung 2: Inferenz: Nullhypothese

Wie lautet die Nullhypothese, wenn die Variable x keinen Einfluss auf $p(y = 1)$ in der Population hat:

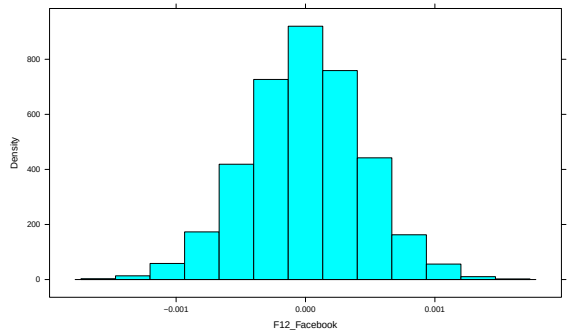
- A: $H_0 : \beta_1 = 0$
- B: $H_0 : \beta_1 = 1$
- C: $H_0 : \hat{\beta}_1 = 0$
- D: $H_0 : \hat{\beta}_1 = 1$

Permutationstest Regressionskoeffizient (I/II)

```
Nullvtlg <- do(5000) *
  coef(glm(F18_Messe ~ shuffle(F12_Facebook),
           data = Extraversion,
           family = binomial("logit")))[2]
```

Permutationstest Regressionskoeffizient (II/II)

```
histogram( ~ F12_Facebook, data = Nullvtlg)
```




```
prop( ~ abs(F12_Facebook) >= coef(ergglm)[2],
      data=Nullvtlg )
```

```
##    TRUE
## 0.3298
```

Übung 3: Inferenz Anzahl Facebook Freunde

Liefern die Daten Belege dafür, einen Zusammenhang zwischen der Anzahl Facebook-Freunde und der Bereitschaft sich freiwillig zur Messe zu melden in der Population zu zeigen (Forschungsthese)?

- Ja.
- Nein.

Modellierung Messebesuch durch Alter

```
ergglm2 <- glm(F18_Messe ~ F14_Alter,
               data = Extraversion,
               family = binomial("logit"))

summary(ergglm2)
```

```
##
## Call:
## glm(formula = F18_Messe ~ F14_Alter, family = binomial('
##     data = Extraversion)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.199  -1.071  -1.035   1.278   1.336
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.78278    0.67527  -1.159    0.246
```

Koeffizienten Modellierung Messebesuch durch Alter

	Estimate	Std. Error	z value	Pr(>|z|
(Intercept)	-0.7827803	0.6752712	-1.1592087	0.246371
F14_Alter	0.0219464	0.0266934	0.8221656	0.410982

Übung 4: Ergebnis Modellierung Messebesuch durch Alter

Wer hat im Modell die höchste Wahrscheinlichkeit sich freiwillig zur Messe zu melden?

- A: Max, 20 Jahre
- B: Tina, 24 Jahre
- C: Susi, 30 Jahre

Übung 5: Inferenz Modellierung Messebesuch durch Alter

Ist in dem Modell der Einfluss der Variable `F14_Alter` *signifikant*?

- Ja.
- Nein.

Vorhersagen Logistische Regression

Für Susi, 30 Jahre:

```
predict(ergglm2,
        newdata = data.frame(F14_Alter=30),
        type="response")
```

```
##         1
## 0.4689432
```

Modellierung Messebesuch durch Geschlecht

```
ergglm3 <- glm(F18_Messe ~ F15_Geschlecht,
               data = Extraversion,
               family = binomial("logit"))

summary(ergglm3)
```

```
##
## Call:
## glm(formula = F18_Messe ~ F15_Geschlecht, family = binor
##     data = Extraversion)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2392  -0.9809  -0.9809   1.1168   1.3875
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.4815     0.1459  -3.300 0.000968
```

Koeffizienten Modellierung Messebesuch durch Geschlecht

	Estimate	Std. Error	z value	
(Intercept)	-0.4814510	0.1459040	-3.299780	0
F15_GeschlechtMann	0.6257006	0.2312028	2.706285	0

Übung 6: Ergebnis Modellierung Messebesuch durch Geschlecht

Wer hat im Modell die höchste Wahrscheinlichkeit sich freiwillig zur Messe zu melden?

- A: Max
- B: Tina
- C: Beide gleich

Übung 7: Inferenz Modellierung Messebesuch durch Geschlecht

Kann im Modell die Nullhypothese $\beta_{\text{F15\_Geschlecht}} = 0$ verworfen werden?

- Ja.
- Nein.

Odds Ratio

$$OR = \frac{\frac{p_{\text{Mann}}}{1-p_{\text{Mann}}}}{\frac{p_{\text{Frau}}}{1-p_{\text{Frau}}}}$$

```
exp(coef(ergglm3))
```

```
##          (Intercept) F15_GeschlechtMann
##            0.6178862           1.8695554
```

Die Chance, dass sich ein Mann freiwillig meldet ist 1.87 mal so groß wie die einer Frau.

Multiple Logistische Regression

```r
ergglm4 <- glm(F18_Messe ~ F14_Alter
               + F15_Geschlecht
               + F12_Facebook,
               data = Extraversion,
               family = binomial("logit"))

summary(ergglm4)
```

```
##
## Call:
## glm(formula = F18_Messe ~ F14_Alter + F15_Geschlecht + F
##     family = binomial("logit"), data = Extraversion)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3645  -1.0245  -0.9356   1.1868   1.4687
##
## Coefficients:
```

Koeffizienten Multiple Logistische Regression

	Estimate	Std. Error	z value	
(Intercept)	-1.1901749	0.7730309	-1.5396214	(
F14_Alter	0.0234367	0.0287012	0.8165753	(
F15_GeschlechtMann	0.5879846	0.2340288	2.5124453	(
F12_Facebook	0.0004693	0.0004675	1.0040207	(

Übung 8: Ergebnis Multiple Logistische Regression

Welche Variablen erhöhen die Wahrscheinlichkeit im Modell, dass sich die Person freiwillig zur Messe meldet?

- A: Nur steigendes Alter
- B: Nur Geschlecht Mann
- C: Nur steigende Anzahl Facebook-Freunde
- D: Alle Variablen
- E: Keine der Variablen

Übung 9: Inferenz Multiple Logistische Regression

Welche Variablen sind *signifikant*?

- A: Nur Alter
- B: Nur Geschlecht
- C: Nur Anzahl Facebook-Freunde
- D: Alle Variablen
- E: Keine der Variablen

Offene Übung: Modellierung Geschlecht

Modellieren Sie die Wahrscheinlichkeit, dass es sich bei einer Person um eine Frau handelt als Funktion der Variablen `F12_Facebook`, `F13_Kater`, `F19_Partybesuche`.