Modelldiagnose
Lineare Regression

Signifikanz, Modellannahmen, Residuenanalyse, Multikolinearität, Ausreißer

Treffen **30.11.2017**
Fabio & Simon

Aufbau

01 **Recap: Lineare Regression**

Lineare Parameterschätzung

02 **Die Logik von Signifikanz**

Mechanismen der Inferenz (Standardfehler, Konfidenzintervalle und NHTS

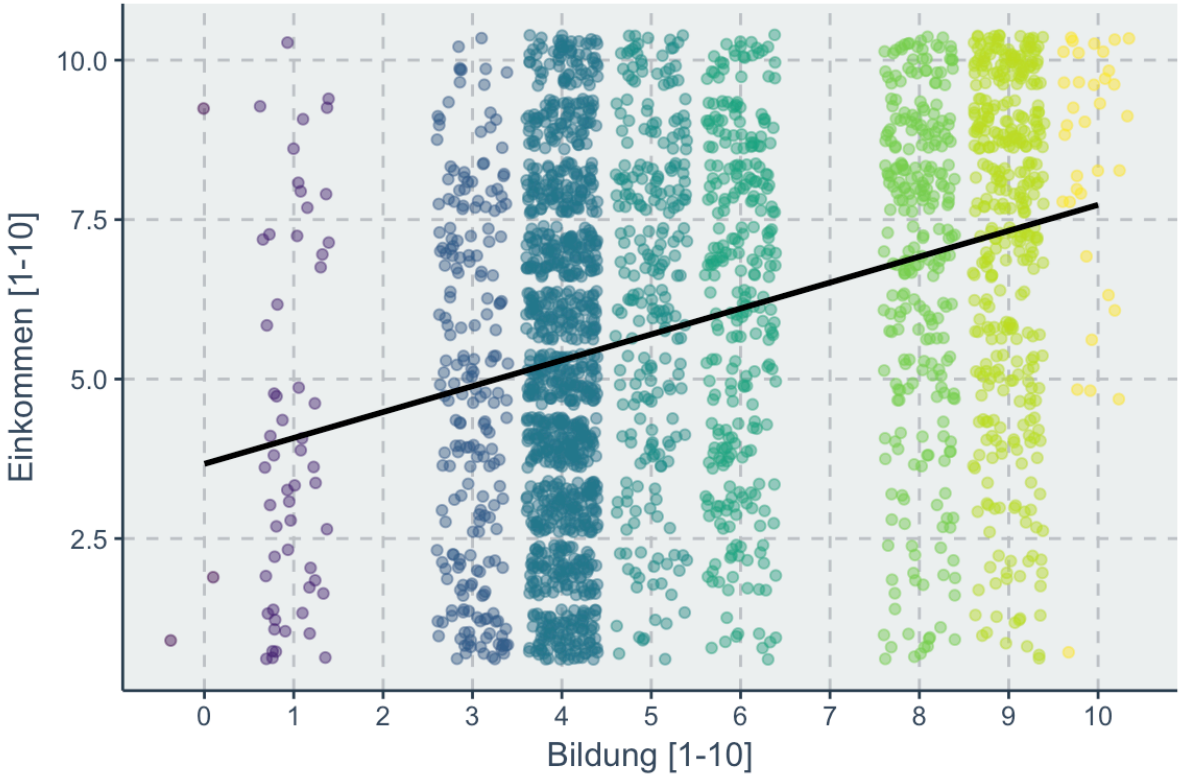
03 **Modellannahmen**

NEW

`learnr`

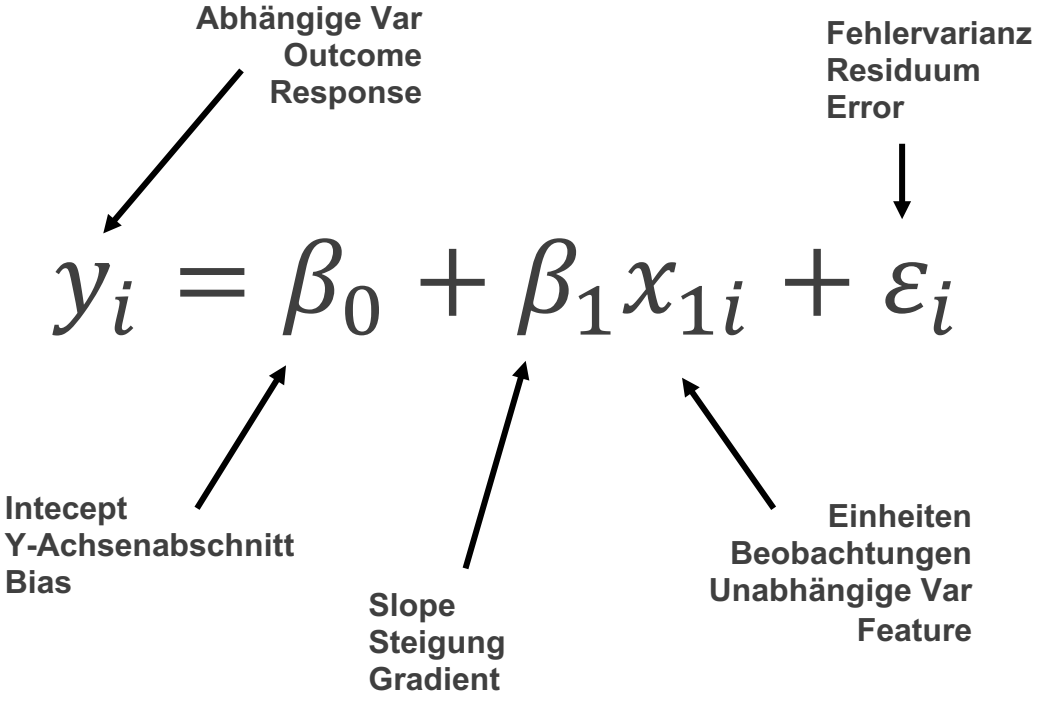
Problemstellungen und Annahmeverletzungen

Lineare Regression
Recap



R User Group Stuttgart

Lineare Regression
Notation



R User Group Stuttgart

Lineare Regression
Parameter

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

$$Beobachtung_i = \underbrace{Parameter * Daten}_{\text{Modell}} + \text{Fehler}$$

$$Beobachtung_i = Lineare\ Fun\ von\ x + \text{Fehler}$$

$$Beobachtung_i = Vorhersage + \text{Fehler}$$

Parameter?!

Input Parameter

Vor Allem beim programmieren

```
x_quer <- mean(x, na.rm = T)
```

Die Funktionsargumente `x` und `na.rm` können auch als Eingabeparameter betrachtet werden, die das Verhalten und den Output der Funktion beeinflusst.

Estimated Parameter

Statistische Modellierung zielt darauf ab, unbekannte Maßzahlen aus Daten zu schätzen, um diese komprimiert zusammenzufassen. Wird auch Koeffizient genannt.

$$\bar{x}$$

$$\beta_0 + \beta_1$$

R User Group Stuttgart

Daten > Schätzung > Params

Unsicherheit

Schätzunsicherheit

```
ess_ger %>%
    lm(imm_econ ~ edu + income + age + I(age^2), data = .) %>%
    broom::tidy()
```

	Dependent variable:
	imm_econ
edu	0.192***
	(0.023)
income	0.074***
	(0.018)
age	-0.070***
	(0.014)
I(age2)	0.001***
	(0.0001)
Constant	5.935***
	(0.308)
Observations	2,511
R^2	0.048
Adjusted R^2	0.047
Residual Std. Error	2.270 (df = 2506)
F Statistic	31.903*** (df = 4; 2506)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
##          term      estimate  std.error statistic      p.value
## 1 (Intercept)  5.9349814831 0.3082548254 19.253491 3.759595e-77
## 2         edu  0.1915797340 0.0227268945  8.429649 5.763811e-17
## 3      income  0.0741948714 0.0175057650  4.238311 2.333559e-05
## 4         age -0.0702992536 0.0136998941 -5.131372 3.096792e-07
## 5    I(age^2)  0.0006791562 0.0001380618  4.919217 9.252048e-07
```

Steigungsparameter

Signifikanztests

Inferenzparameter

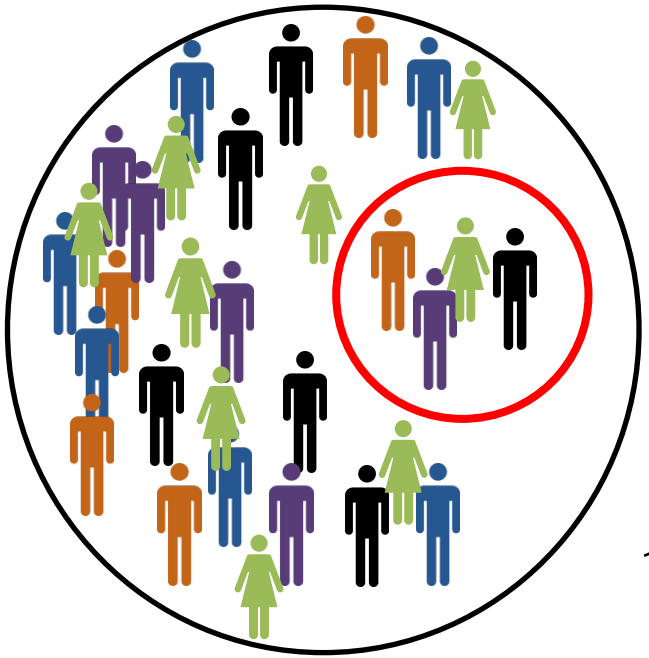
- Standardfehler
- t-Statistik
- p-Werte/ NHTS

R User Group Stuttgart

Inferenz

Regression mit **Stichproben**-Daten

$$y_i = b_0 + b_1 x_{1i} + u_i$$



Inferenz

$\hat{\beta}$ oder b

Zufallsauswahl
(randomness)

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

Regression mit Daten der **Grundgesamtheit**

R User Group Stuttgart

Standardfehler
Standard Error

```
##          term      estimate    std.error statistic      p.value
## 1 (Intercept)  5.9349814831 0.3082548254 19.253491 3.759595e-77
## 2         edu  0.1915797340 0.0227268945  8.429649 5.763811e-17
## 3      income  0.0741948714 0.0175057650  4.238311 2.333559e-05
## 4         age -0.0702992536 0.0136998941 -5.131372 3.096792e-07
## 5    I(age^2)  0.0006791562 0.0001380618  4.919217 9.252048e-07
```

	Dependent variable:
	imm_econ
edu	0.192***
	(0.023)
income	0.074***
	(0.018)
age	-0.070***
	(0.014)
I(age2)	0.001***
	(0.0001)
Constant	5.935***
	(0.308)
Observations	2,511
R^2	0.048
Adjusted R^2	0.047
Residual Std. Error	2.270 (df = 2506)
F Statistic	31.903*** (df = 4; 2506)
Note:	*p<0.1; **p<0.05; ***p<0.01

Streuung um den Intercept

$$SE(\beta_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

← uninteressant

Streuung um einen Slope

$$SE(\beta_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{\sigma}{s_x\sqrt{n}}$$

← **Sehr wichtig, da der SE die Streuung/ Abweichungen um die lineare Vorhersage beschreibt.**

R User Group Stuttgart

t-Test
Signifikanztests

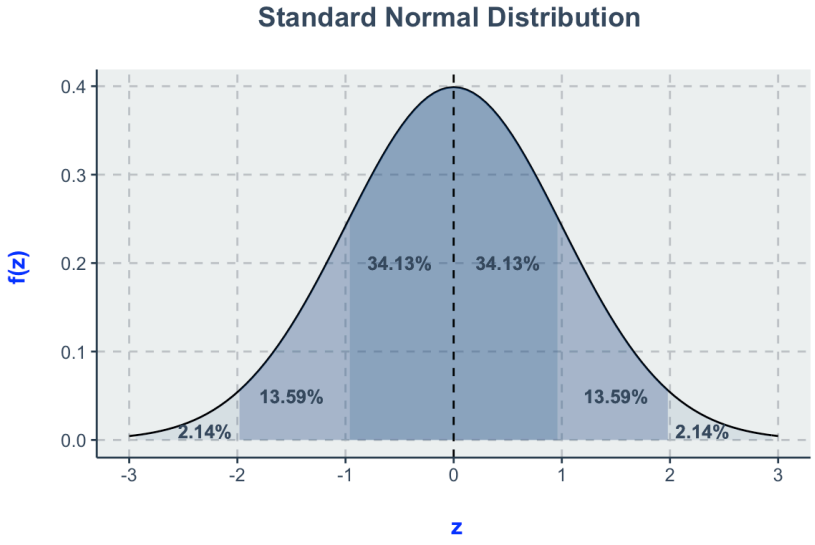
	term	estimate	std.error	statistic	p.value
1	(Intercept)	5.9349814831	0.3082548254	19.253491	3.759595e-77
2	edu	0.1915797340	0.0227268945	8.429649	5.763811e-17
3	income	0.0741948714	0.0175057650	4.238311	2.333559e-05
4	age	-0.0702992536	0.0136998941	-5.131372	3.096792e-07
5	I(age^2)	0.0006791562	0.0001380618	4.919217	9.252048e-07

Vergleich des empirischen **t-Wertes** und dem theoretischen/ kritischen t-Wert.

$$t = \frac{\beta_1 - 0}{SE(\beta_1)} \sim N(0,1) \text{ genauer } \sim t_{n-2}$$

Welche **Verteilung**?
- t-Verteilung
- z-Verteilung

Beide Standardnormalverteilt. Ab t > 30 konvergiert die t- zur z-Verteilung



Konfidenzintervalle

$$CI_\beta = \hat{\beta}_1 \pm t_{\frac{\alpha}{2}} \, SE(\beta_1)$$

$$[\hat{\beta}_1 - t_{\frac{\alpha}{2}} \, SE(\beta_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2}} \, SE(\beta_1)]$$

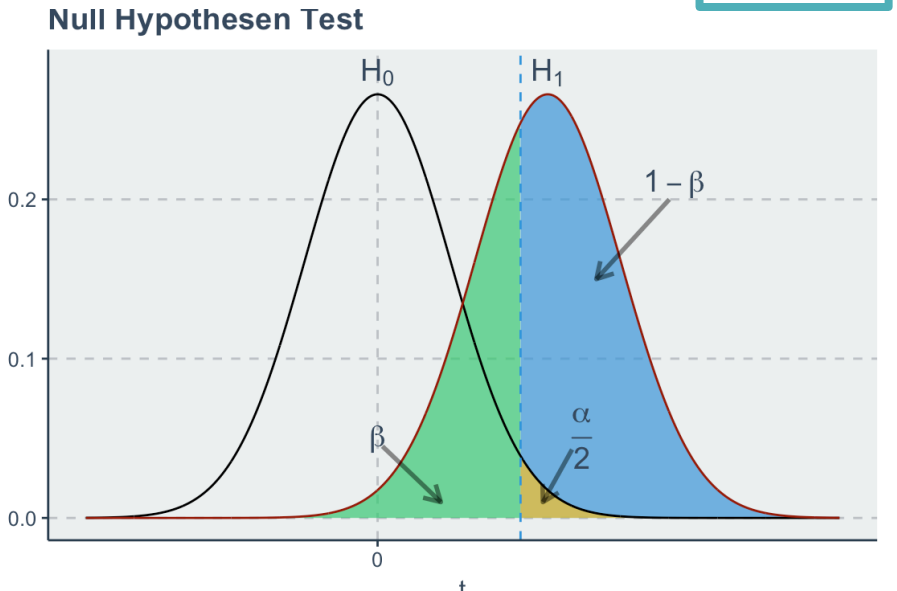
confint(fit1)

```
##                 2.5 %    97.5 %
## (Intercept) 3.3950307 3.9478453
## edu         0.3585679 0.4532925
```

Nullhypothesentests

p-Werte

	term	estimate	std.error	statistic	p.value
1	(Intercept)	5.9349814831	0.3082548254	19.253491	3.759595e-77
2	edu	0.1915797340	0.0227268945	8.429649	5.763811e-17
3	income	0.0741948714	0.0175057650	4.238311	2.333559e-05
4	age	-0.0702992536	0.0136998941	-5.131372	3.096792e-07
5	I(age^2)	0.0006791562	0.0001380618	4.919217	9.252048e-07



$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Modellannahmen

Der linearen Regression

01 **Linearität der Parameter**

02 **Unabhängigkeit der Residuen**

03 **Homoskedastizität**

04 **Normalverteilung der Residuen (IID)**

05 **Multikolinearität**

06 **Ausreißer**

Und jetzt ... `learnr`

https://github.com/favstats/rgroup_diagnostik