Statistik zur Datenanalyse

Dr. Meike Wocken

Klausurvorbereitung

HS Bielefeld
Digitale Technologien (M.Sc.)
WiSe 2023/24

Meike.Wocken@codecentric.de

Lösung

Probeklausur

Probeklausur – Aufgabe 1

```
Call:
lm(formula = Volume ~ Diameter + Height, data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.182027 -0.074603 -0.004944  0.061530  0.240610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.64166    0.24524  -6.694 2.89e-07 ***
Diameter     5.25283    0.29572  17.763  < 2e-16 ***
Height       0.03144    0.01212   2.593   0.0149 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1102 on 28 degrees of freedom
Multiple R-squared:  0.9477,    Adjusted R-squared:  0.9439
F-statistic: 253.5 on 2 and 28 DF,  p-value: < 2.2e-16
```

1. (20 Punkte) Gegeben ist ein Datensatz mit 31 Beobachtungen von Obstbäumen (Schwarzkirsche). Es sind die drei Variablen gegeben:

   - Volumen (Volume) des Baumes, gemessen in Kubikmeter (m^3).
   - Durchmesser (Diameter) des Baumes, gemessen in Meter (m).
   - Höhe (Height) des Baumes, gemessen in Meter (m).

   Für das multiple lineare Regressionsmodell

   $$Volume_i = \beta_0 + \beta_1 Diameter_i + \beta_2 Height_i + u_i$$

   erhalten Sie das Schätzergebnis in der Statistik-Software **R**, das in Abbildung 1 zu sehen ist.

   (a) Machen Sie bitte eine Aussage zur Güte des Modells.

   (b) Interpretieren Sie den quantitativen Effekt von *Diameter*

   (c) Stellen Sie die Nullhypothese und Alternativhypothese auf für den t-Test auf statistische Signifikanz von *Height*. Geben Sie die Prüfgröße an und interpretieren Sie das Ergebnis.

Lösung – Aufgabe 1

LÖSUNG:

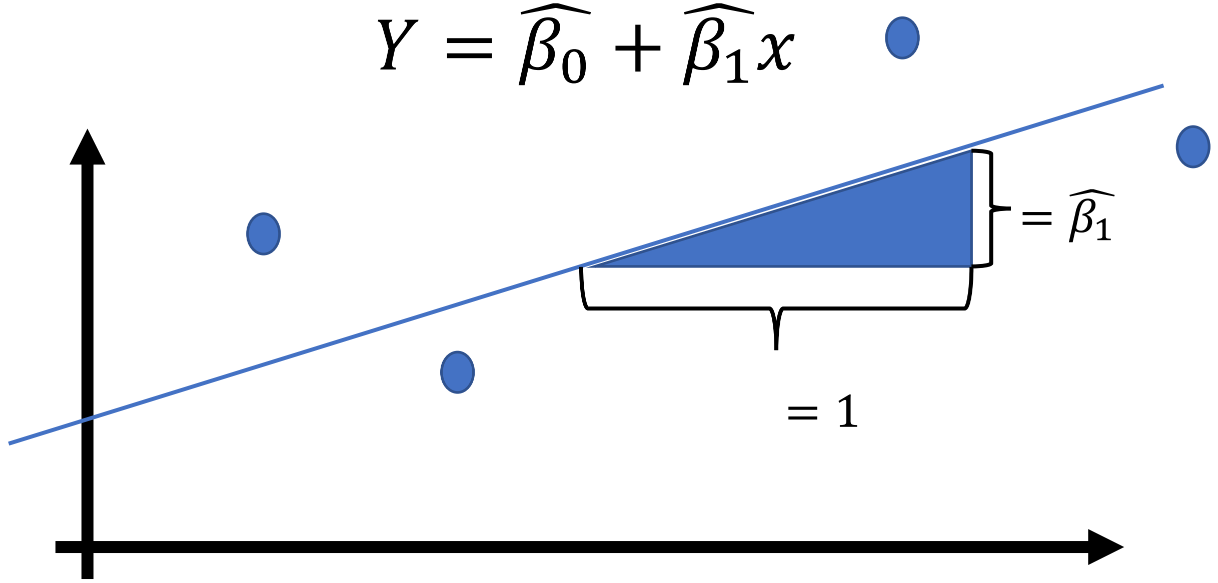
(a) (5 Punkte) Residuen als Schätzer der Abweichungen zwischen Beobachtungen und Modell sind für die Aussage zur Güte des Modells verwendbar Der $RSE/n-p-1 = RSE/28$ ist hier $0,1102$. Allerdings ist der RSE ein absolutes Maß. Das

R^2 hingegen gibt den Anteil der Gesamtstreuung von $Volume$ an, der durch das Modell erklärt werden kann. Hier ist der Anteil mit $94,77\%$ sehr hoch. Damit hat das Modell eine sehr hohe Güte.

(b) (5 Punkte) *Volume* ändert sich im Durchschnitt um 5, 25283 Einheiten, wenn *Diameter* um eine Einheit, ceteris paribus, zunimmt.

Änderung von x um eine Einheit bewirkt ceteris paribus (c.p.) eine durchschnittliche Änderung von Y um $\hat{\beta}_1$ Einheiten

$$Y = \widehat{\beta_0} + \widehat{\beta_1} x$$



Exkurs

Kategoriale Variablen

Exkurs: Kategoriale Variablen

	Dependent variable:	
	y	
	(1)	(2)
x1	−4.059***	−4.055***
	(0.039)	(0.039)
x2		−0.127
		(0.135)
gender	1.999***	2.039***
	(0.276)	(0.280)
Constant	70.112***	74.343***
	(1.981)	(4.931)
Observations	40	40
R^2	0.997	0.997
Adjusted R^2	0.997	0.996
Residual Std. Error	0.840 (df = 37)	0.842 (df = 36)
F Statistic	5,569.701*** (df = 2; 37)	3,701.273*** (df = 3; 36)
Note:		*p<0.1; **p<0.05; ***p<0.01

Zwei Modelle (1) und (2) sind geschätzt worden, um die Größe y zu erklären.

Die Variable $gender$ ist eine Dummy-Variable, die nur zwei Werte annehmen kann:

$gender$ = 0, falls weibliche Person, sonst $gender$ = 1.

(1) $y = \beta_0 + \beta_1 x_1 + \beta_2\ gender + \epsilon$

Erklärende Variablen müssen nicht metrisch sein!

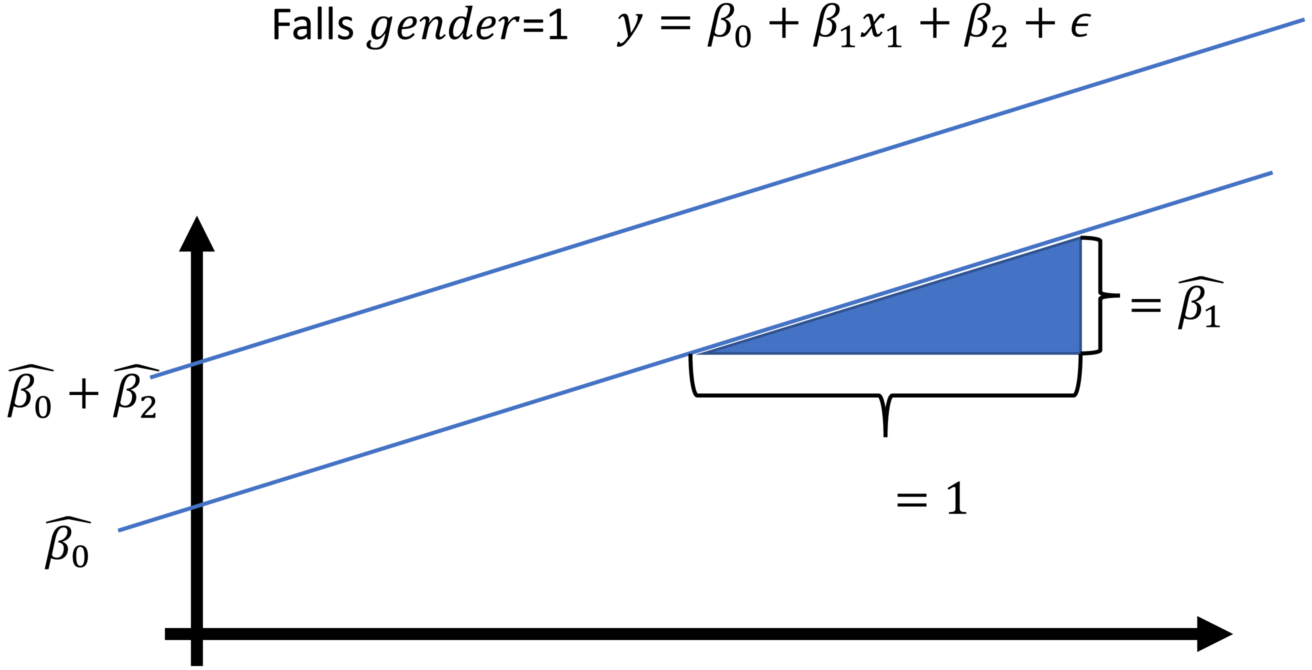
Falls $gender$=0 $y = \beta_0 + \beta_1 x_1 + \epsilon$

Falls $gender$=1 $y = \beta_0 + \beta_1 x_1 + \beta_2 + \epsilon$

Interpretation: Im Vergleich zu weiblichen Personen, haben nicht weibliche Personen im Durchschnitt c.p. ein $\widehat{\beta_2}$ Einheiten höheres y.

Falls $gender$=0 $y = \beta_0 + \beta_1 x_1 + \epsilon$

Falls $gender$=1 $y = \beta_0 + \beta_1 x_1 + \beta_2 + \epsilon$

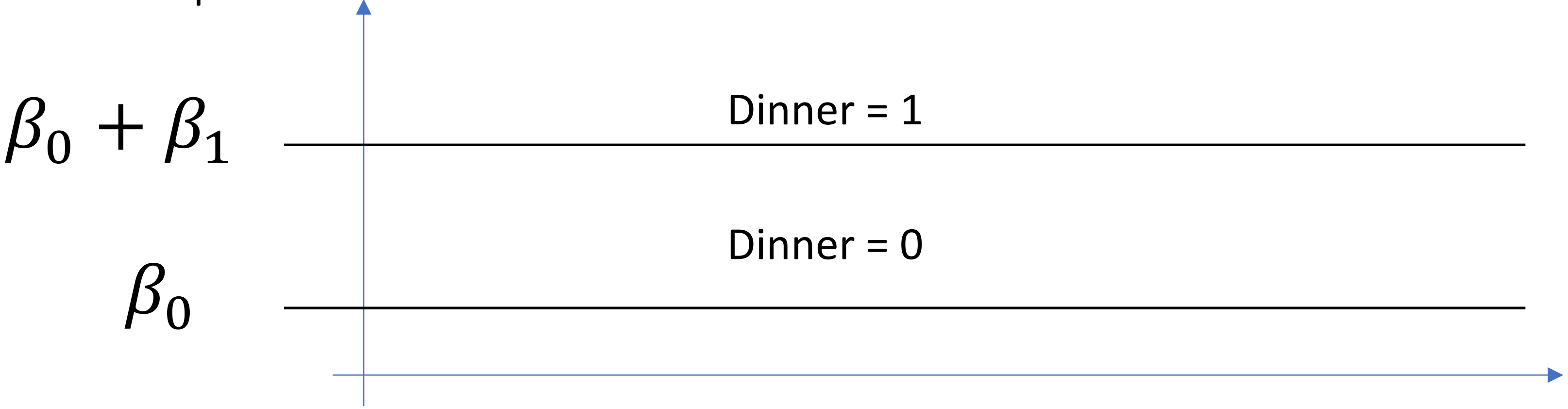


In dem Modell muss auch gar keine weitere metrische Variable zur Erklärung genutzt werden.

- Einfaches Beispiel: zwei Klassen miteinander vergleichen.

Dinner= 1 für Aktivität „Dinner“ und 0 für Aktivität „Breakfast“

$$Temperatur = \beta_0 + \beta_1 \cdot Dinner$$

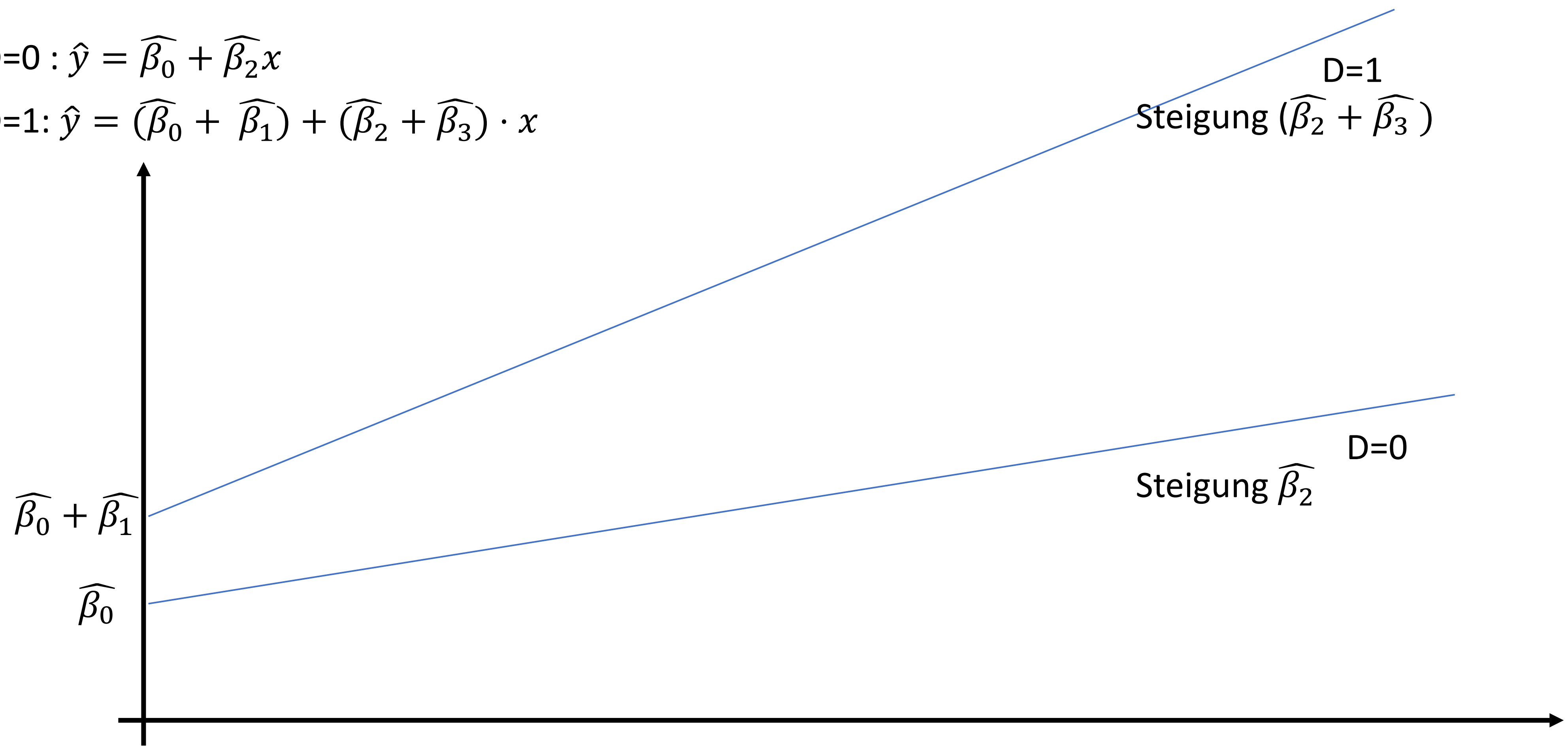


„Advanced“: Modellierung unterschiedlicher Steigungen mit Dummy Variablen

$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} \cdot D + \widehat{\beta_2}x + \widehat{\beta_3} \cdot D \cdot x$

D=0 : $\hat{y} = \widehat{\beta_0} + \widehat{\beta_2}x$

D=1: $\hat{y} = (\widehat{\beta_0} + \widehat{\beta_1}) + (\widehat{\beta_2} + \widehat{\beta_3}) \cdot x$



Mehr als zwei Klassen

$$beste.lage = \begin{cases} 1 & Wohnlage = 1 \text{ (Beste Wohnlage)} \\ 0 & \text{sonst} \end{cases}$$

$$gute.lage = \begin{cases} 1 & Wohnlage = 2 \text{ (Gute Wohnlage)} \\ 0 & \text{sonst} \end{cases}$$

$$normale.lage = \begin{cases} 1 & Wohnlage = 3 \text{ (Normale Wohnlage)} \\ 0 & \text{sonst} \end{cases}$$

- Bodenrichtwert.csv
- 1=beste Lage, 2= gute Lage, 3= normale Lage
- Für jede Kategorie wird ein separater Effekt geschätzt, dafür wird aber immer eine Dummy Variable weniger ins Modell aufgenommen, als Kategorien insgesamt da sind (sonst perfekte Kollinearität). Eine Kategorie ist Referenzkategorie (im ersten Beispiel Breakfast). Hier also 2 Dummy Variablen im Modell (gute und normale Wohnlage im Vergleich zu beste Wohnlage.
- Schätzergebnisse im Vergleich zur Referenzkategorie zu interpretieren
- R kümmert sich automatisch um das erstellen der Dummy-Variablen

Lösung – Aufgabe 1

(c) (5 Punkte) Zweiseitiger Test $H_0 : \beta_{Height} = 0$ und $H_1 : \beta_{Height} \neq 0$.

$$t = 0,03144/0,01212 = 2,593$$

Die Prüfgröße könnte nun mit dem $1 - \alpha/2$-Quantil der t-Verteilung verglichen werden oder es wird der p-Wert ermittelt, d.h. die Wahrscheinlichkeit einen extremeren Wert als die berechnete Prüfgröße unter Annahme von H_0 zu erhalten. Im R Output ist gibt die Angabe von einem Sternchen an dass der p-Wert größer als $0,01$ und kleiner als $0,05$ sein wird. Daher ist H_0 mit einem Signifikanzniveau von $/alpha = 0,05$ abzulehnen. $Height$ besitzt zum Signifikanzniveau $/alpha = 0,05$ einen statistisch signifikanten Einfluß.

Der t-Test prüft, ob die einzelnen Variablen einen signifikanten Einfluss haben. Hätten sie keinen Einfluss, wäre der Schätzer des Koeffizienten 0. Somit wird die Hypothese getestet, ob der Schätzer gleich Null ist.

Wiederholung

Hypothesentests

Neue Stichprobe

→ neue Schätzung

→ Werte der geschätzten Parameter ändern sich.

① $H_0: \beta_{Height} = 0$ $\quad \beta_{Height} \neq 0$

Zweiseitiger Test

$\beta_{Height} > 0$

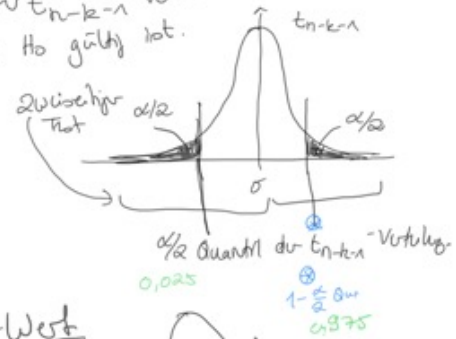
$\beta_{Height} < 0$

② <u>Prüfgröße berechnen</u>

$$PG = \frac{\hat{\beta}_{Height} - 0}{se(\hat{\beta}_{Height})} = \frac{0{,}03144}{0{,}01212} = 2{,}593$$

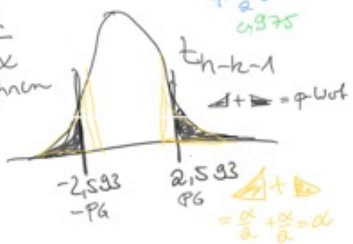
③ überprüfen, wo PG in der Verteilung liegt:

$PG \sim t_{n-k-1}$ verteilt unter der Annahme, dass H_0 gültig ist.



P-Wert

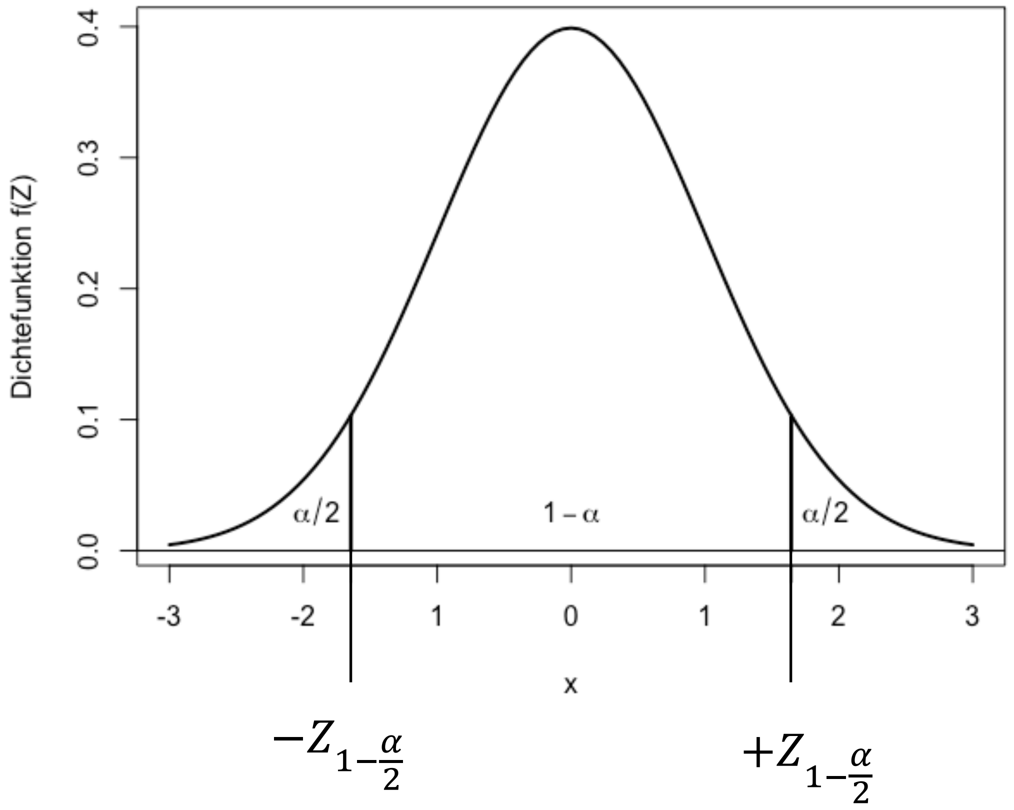
P-Wert < α
↳ H_0 ablehnen



Beurteilung der Abweichung $\bar{X} - \mu_0$

- Es wird immer eine Abweichung zwischen $\bar{X}$ und μ_0 geben, auch wenn $H_0\colon \mu = \mu_0$ wahr ist (Produktionsstück hat im Durchschnitt Sollwert-Länge).

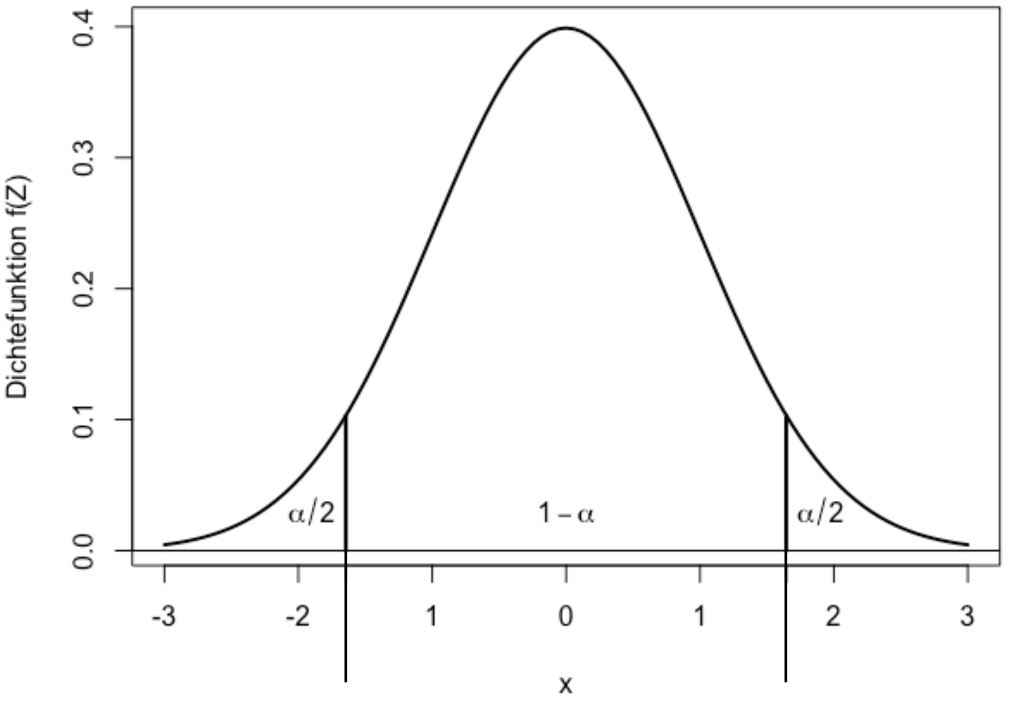
Die Größe Z folgt der Standard-Normalverteilung. Unter der Annahme von H_0 beträgt die Wahrscheinlichkeit $(1-\alpha)\cdot 100\%$, das Z im Bereich zwischen $\left[-Z_{1-\frac{\alpha}{2}}, +Z_{1-\frac{\alpha}{2}}\right]$ liegt. D.h., solange Z im mittleren Bereich liegt, spricht alles für die Gültigkeit von H_0.



Wie ist es zu beweten, wenn Z Werte außerhalb des mittleren Bereiches liegen?

Die Wahrscheinlichkeit ist gering $(\alpha \cdot 100\%)$, dass ich Z unter der Annahmen, dass H_0gültig ist, außen zu beobachten.

D.h. entweder beobachten wir damit ein sehr unwahrscheinliches Ereignis oder es ist ein Hinweis, dass die Annahme H_0 ungültig ist.



$-Z_{1-\frac{\alpha}{2}}$ $+Z_{1-\frac{\alpha}{2}}$

Fehler 1. und 2. Art

- Fehler 1. Art:

H_0 wird verworfen, obwohl H_0 wahr ist

- Fehler 2. Art:

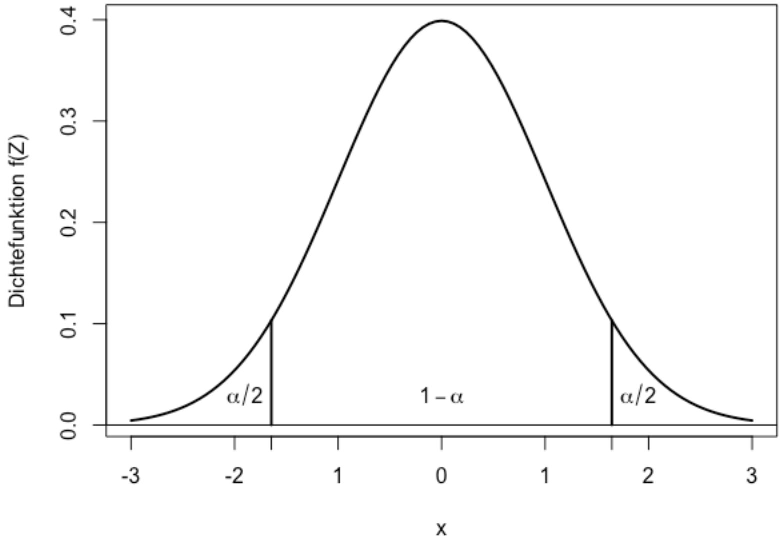
H_0 wird beibehalten, obwohl H_1 wahr ist.

Statistische Tests können nur den Fehler 1. Art kontrollieren:

Test zum Signifikanzniveau $\alpha, 0 < \alpha < 1$ mit

$$P(H_1 annehmen\ |H_0 wahr) \leq \alpha$$

Über das Signifikanzniveau α wird gesteuert, wie viel Wahrscheinlichkeit „rechts" und „links" bleibt.



Zusammenfassung

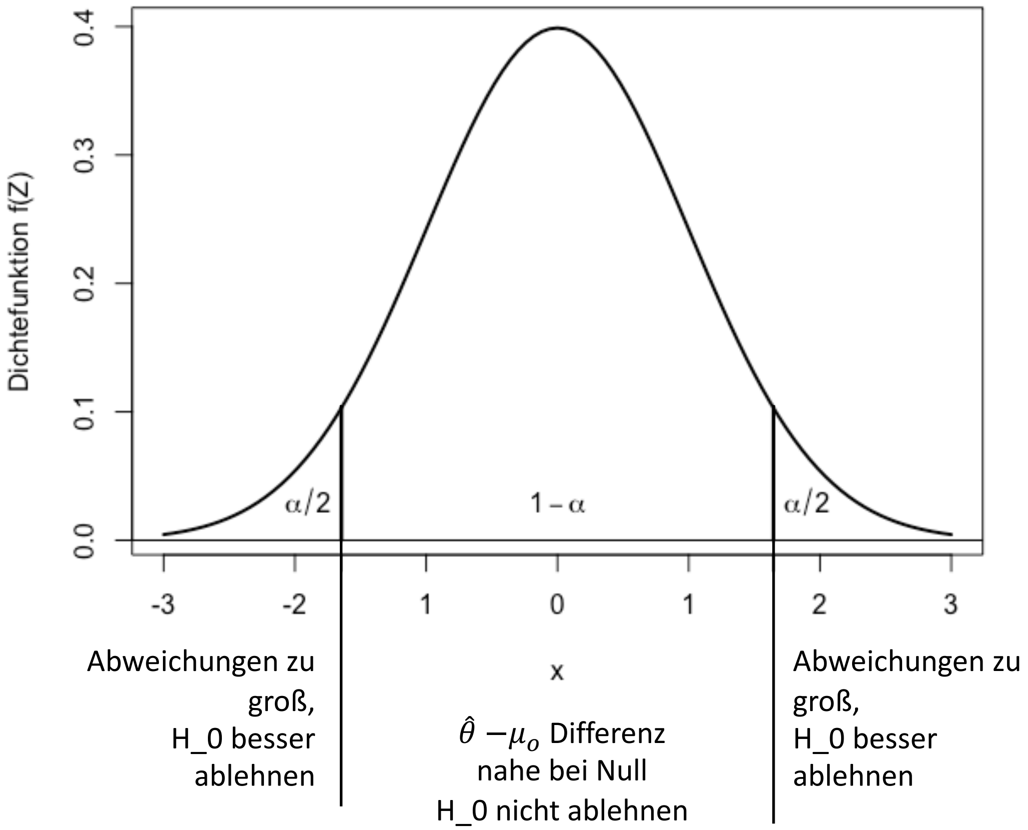
- Aussage zu einem Parameter $\theta = \mu$ gewünscht
- Schätzfkt. $\hat{\theta} = t(x_1, \dots, x_n) = \frac{\sum x_i}{n}$
- Zu prüfende Hypothesen aufstellen

$$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0$$

- Wenn H_0 wahr ist, dann gilt

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

Wir berechnen die Prüfgröße Z



Kapitel 10 - Aufgabe 1

1. Lebenshaltungskosten:

Ein Marktforschungsinstitut führt jährliche Untersuchungen zu den Lebenshaltungskosten durch. Die Kosten für einen bestimmten Warenkorb beliefen sich in den letzten Jahren auf durchschnittlich 600 EUR. Im Beispieljahr wurde in einer Stichprobe von

40 zufällig ausgewählten Kaufhäusern jeweils der aktuelle Preis des Warenkorbs bestimmt. Als Schätzer für den aktuellen Preis des Warenkorbs ergab sich ein mittlerer Preis von 605 EUR. Die Varianz $\sigma^2 = 225$ sei aufgrund langjähriger Erfahrung bekannt. Gehen Sie von einer Normalverteilung des Preises für den Warenkorb aus.

- Hat sich der Preis des Warenkorbs im Vergleich zu den Vorjahren signifikant zum Niveau $\alpha = 0,01$ erhöht? Wie lautet das zugehörige statistische Testproblem?

1) Zu prüfende Hypothesen aufstellen

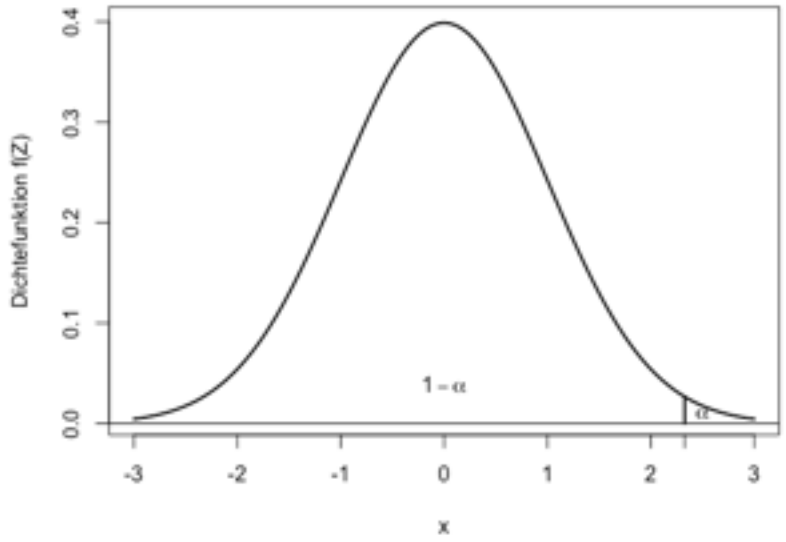
$$H_0: \mu = 600 \quad H_1: \mu > 600$$

Einseitiger Test!!

2) Prüfgröße berechnen:

$$Z = \frac{605 - 600}{\sqrt{225/40}} = \frac{5}{15/\sqrt{40}} = 2{,}108$$

Unter H_0 gilt $Z \sim N(0{,}1)$



Aufgabe 1

Da gilt Z $= 2{,}108 < 2{,}326$ kann Ho nicht abgelehnt werden.

Da gilt $\mathrm{Z} = 2{,}108 < 2{,}326$ kann Ho nicht abgelehnt werden.

H0 nicht ablehnen | H0 ablehnen

`qnorm(0.99, 0, 1)=2.326`

Kapitel 10 - Aufgabe 2

2. Abfüllanlage:

Der Output einer Abfüllanlage lag in allen Test seit der Inbetriebnahme im Mittel bei $26.60ml$ je abgefüllter Einheit mit einer Varianz von $0.025ml^2$. Der letzte Test lag einige Monate zurück. Nun wird eine Zufallststichprobe von 40 Einheiten entnommen und die Füllmenge überprüft. Der Mittelwert der Stichprobe liegt bei $26.65ml$ je abgefüllter Einheit. Die bisherigen Testergebnisse werden als Parameter der Grundgesamtheit angenommen, d.h. der Mittelwert der Grundgesamtheit beträgt $\mu_0 = 26.60ml$, die Varianz beträgt $\sigma_0^2 = 0.025ml^2$. Es kann eine Normalverteilung der Abfüllmenge unterstellt werden.

- Weicht die Abfüllmenge statistisch signifikant zum Niveau $\alpha = 0,05$ von den vorherigen Werten ab? Wie lautet das zugehörige statistische Testproblem?

```
qnorm(0.025,0,1) = -1.9599

qnorm(0.975,0,1) = 1.9599
```

Aufgabe 2 - Lösung

- 1) Zu prüfende Hypothesen aufstellen

$$H_0: \mu = 26{,}60 \quad H_1: \mu \neq 26{,}60$$

Zweiseitiger Test!!

2) Prüfgröße berechnen:

$$Z = \frac{26{,}65 - 26{,}60}{\sqrt{0{,}025/40}} = \frac{26{,}65 - 26{,}60}{\sqrt{0{,}025}}\sqrt{40} = 2$$

Unter H_0 gilt $Z \sim N(0{,}1)$

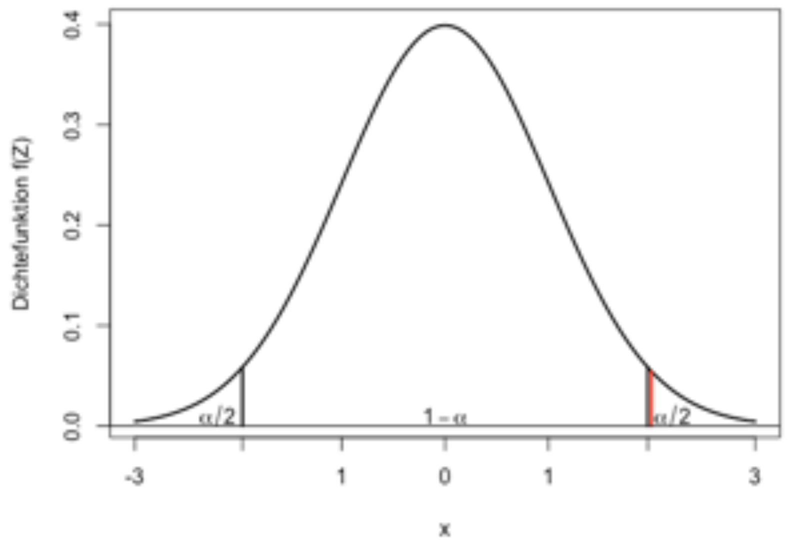
Schwarze Linie:

`qnorm(0.025, 0, 1) = -1.9599` (2,5%-Quantil)

`qnorm(0.975, 0, 1) = 1.9599` (97,5%-Quantil)

Rote Linie: Z=2

H0 ist abzulehnen!



Exkurs: Berechnung t-Test

	Dependent variable:	
	y	
	(1)	(2)
x1	−4.059***	−4.055***
	(0.039)	(0.039)
x2		−0.127
		(0.135)
gender	1.999***	2.039***
	(0.276)	(0.280)
Constant	70.112***	74.343***
	(1.981)	(4.931)
Observations	40	40
R^2	0.997	0.997
Adjusted R^2	0.997	0.996
Residual Std. Error	0.840 (df = 37)	0.842 (df = 36)
F Statistic	5,569.701*** (df = 2; 37)	3,701.273*** (df = 3; 36)
Note:		*p<0.1; **p<0.05; ***p<0.01

In den Klammern steht der geschätzte Standardfehler.

Test in Modell (1):
Ist der Effekt von x1 statistisch signifikant zum Signifikanzniveau alpha = 0,05?

1) Hypothese aufstellen:

$$H_0: \beta_{X1} = 0 \quad H_1: \beta_{X1} \neq 0$$

Zweiseitiger Test!!

2) Prüfgröße berechnen

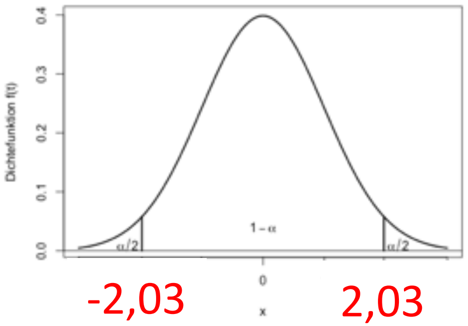
	Dependent variable:	
	y	
	(1)	(2)
x1	−4.059***	−4.055***
	(0.039)	(0.039)
x2		−0.127
		(0.135)
gender	1.999***	2.039***
	(0.276)	(0.280)
Constant	70.112***	74.343***
	(1.981)	(4.931)
Observations	40	40
R^2	0.997	0.997
Adjusted R^2	0.997	0.996
Residual Std. Error	0.840 (df = 37)	0.842 (df = 36)
F Statistic	5,569.701*** (df = 2; 37)	3,701.273*** (df = 3; 36)
Note:		*p<0.1; **p<0.05; ***p<0.01

Parameterschätzer $\widehat{\beta}$ ist selber Zufallsvariable, abhängig von der Stichprobe. Unter der Annahme, das Homoskedastizität, Erwartungstreue $E(\widehat{\theta}) = \theta$ und Konsistenz ($n \uparrow \Rightarrow \widehat{\theta} \to \theta$, wenn die Anzahl an Beobachtungen groß wird, konvergiert der Schätzer gegen den wahren Wert) vorliegen (siehe Annahmen für lineare Modelle), können wir Aussagen zur Verteilung machen.

Prüfgröße:

$$t = \frac{\widehat{\beta_{X1}} - 0}{se(\widehat{\beta_{X1}})} = \frac{-4{,}059 - 0}{0{,}039} = -104{,}08$$

Die Prüfgröße folgt einer t-Verteilung mit n-2-1 Freiheitsgraden.
Gegeben:
0,025-Quantil der t-Verteilung mit 27 Freiheitsgraden = - 2,03
0,05-Quantil der t-Verteilung mit 27 Freiheitsgraden = -1,69



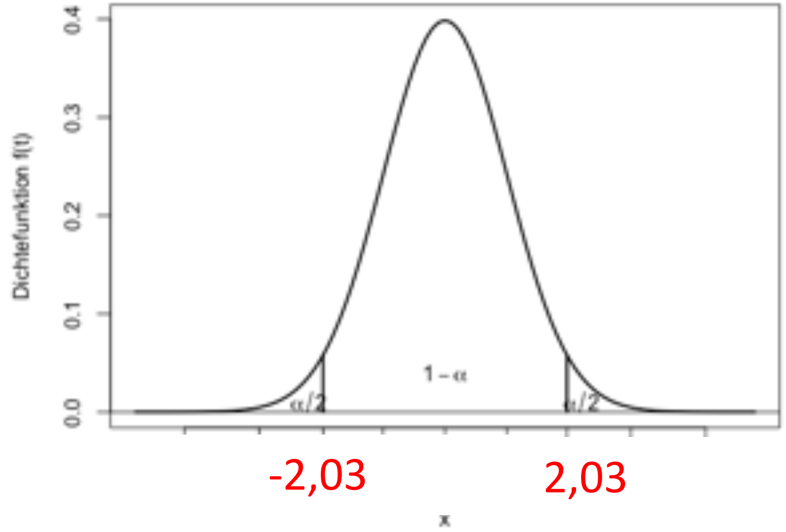
R Output

	Dependent variable:	
	y	
	(1)	(2)
x1	-4.059^{***}	-4.055^{***}
	(0.039)	(0.039)
x2		-0.127
		(0.135)
gender	1.999^{***}	2.039^{***}
	(0.276)	(0.280)
Constant	70.112^{***}	74.343^{***}
	(1.981)	(4.931)
Observations	40	40
R^2	0.997	0.997
Adjusted R^2	0.997	0.996
Residual Std. Error	0.840 (df = 37)	0.842 (df = 36)
F Statistic	$5{,}569.701^{***}$ (df = 2; 37)	$3{,}701.273^{***}$ (df = 3; 36)
Note:		*p<0.1; **p<0.05; ***p<0.01

Prüfgröße:

$$t = \frac{\widehat{\beta_{X1}} - 0}{se(\widehat{\beta_{X1}})} = \frac{-4{,}059 - 0}{0{,}039} = -104{,}08$$

Für eine Entscheidung müssen wir entweder die Prüfgröße -104,08 mit dem $\left(\frac{\alpha}{2}\right)$-Quantil der t-Verteilung vergleichen oder die Fläche, die rechts und links von der Prüfgröße liegt auswerten (p-Werte)



p-Wert

Der p-Wert ist die Wahrscheinlichkeit, einen noch extremeren Wert als die berechnete Prüfgröße unter H0 zu erhalten. Ist p-Wert$< \alpha$, dann kann H0 zum Signifikanzniveau α abgelehnt werden.

Bei Heteroskedastizität und/oder Kollinearität im Modell sind Hypothesentests nicht verlässlich, aufgrund einer vergrößerten Standardabweichung der Schätzer.

Test auf statistische Signifikanz von kategorialer Variable

```
Call:
lm(formula = Salary ~ AtBat + Hits + Walks + CAtBat + CRuns +
    CRBI + CWalks + League + Division + PutOuts + Assists, data = Hitters)

Residuals:
   Min     1Q Median     3Q    Max
-932.2 -175.4  -29.2  130.4 1897.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  135.75122   71.34623   1.903 0.058223 .
AtBat         -2.12775    0.53746  -3.959 9.81e-05 ***
Hits           6.92370    1.64612   4.206 3.62e-05 ***
Walks          5.62028    1.59064   3.533 0.000488 ***
CAtBat        -0.13899    0.05609  -2.478 0.013870 *
CRuns          1.45533    0.39270   3.706 0.000259 ***
CRBI           0.78525    0.20978   3.743 0.000225 ***
CWalks        -0.82286    0.26361  -3.121 0.002010 **
LeagueN       43.11162   39.96612   1.079 0.281755
DivisionW   -111.14603   39.21835  -2.834 0.004970 **
PutOuts        0.28941    0.07478   3.870 0.000139 ***
Assists        0.26883    0.15816   1.700 0.090430 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 311.7 on 251 degrees of freedom
Multiple R-squared:  0.5426,    Adjusted R-squared:  0.5226
F-statistic: 27.07 on 11 and 251 DF,  p-value: < 2.2e-16
```

- LeagueN ist 1, wenn League =N ist, sonst 0 (League = A).
- $H_0: \beta_{League} = 0$ kann nicht abgelehnt werden. Was bedeutet das?

Der **Unterschied** zwischen den Leagues ist statistisch nicht signifikant!!

F-Test auf Gesamtsignifikanz (Overall F test)

```
Call:
lm(formula = Salary ~ AtBat + Hits + Walks + CAtBat + CRuns +
    CRBI + CWalks + League + Division + PutOuts + Assists, data = Hitters)

Residuals:
   Min     1Q Median     3Q    Max 
-932.2 -175.4  -29.2  130.4 1897.2 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  135.75122   71.34623   1.903 0.058223 .  
AtBat         -2.12775    0.53746  -3.959 9.81e-05 ***
Hits           6.92370    1.64612   4.206 3.62e-05 ***
Walks          5.62028    1.59064   3.533 0.000488 ***
CAtBat        -0.13899    0.05609  -2.478 0.013870 *  
CRuns          1.45533    0.39270   3.706 0.000259 ***
CRBI           0.78525    0.20978   3.743 0.000225 ***
CWalks        -0.82286    0.26361  -3.121 0.002010 ** 
LeagueN       43.11162   39.96612   1.079 0.281755    
DivisionW   -111.14603   39.21835  -2.834 0.004970 ** 
PutOuts        0.28941    0.07478   3.870 0.000139 ***
Assists        0.26883    0.15816   1.700 0.090430 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 311.7 on 251 degrees of freedom
Multiple R-squared:  0.5426,	Adjusted R-squared:  0.5226 
F-statistic: 27.07 on 11 and 251 DF,  p-value: < 2.2e-16
```

- Im R Output gibt es F-Test auf Gesamtsignifikanz für lineare Regressionen

- $H_0: \beta_1 = \cdots = \beta_p = 0$

(die **erklärenden** Variablen können nichts erklären)

- H_1: mind. ein $\beta_i \neq 0, i \in 1, \ldots, p$
- Einseitiger Test

- Prüfgröße ist F-verteilt mit p und n-p-1 Freiheitsgraden

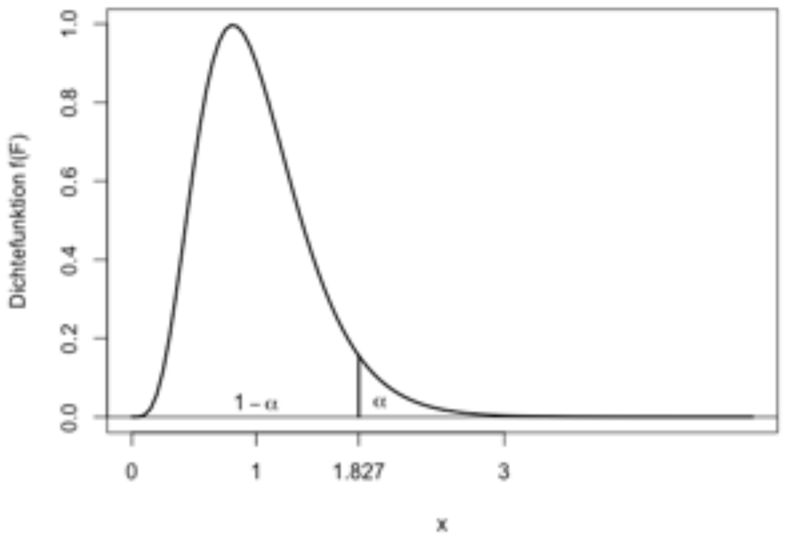
F-Test auf Gesamtsignifikanz (Overall F test)

```
Call:
lm(formula = Salary ~ AtBat + Hits + Walks + CAtBat + CRuns +
    CRBI + CWalks + League + Division + PutOuts + Assists, data = Hitters)

Residuals:
   Min     1Q Median     3Q    Max
-932.2 -175.4  -29.2  130.4 1897.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  135.75122   71.34623   1.903 0.058223 .
AtBat         -2.12775    0.53746  -3.959 9.81e-05 ***
Hits           6.92370    1.64612   4.206 3.62e-05 ***
Walks          5.62028    1.59064   3.533 0.000488 ***
CAtBat        -0.13899    0.05609  -2.478 0.013870 *
CRuns          1.45533    0.39270   3.706 0.000259 ***
CRBI           0.78525    0.20978   3.743 0.000225 ***
CWalks        -0.82286    0.26361  -3.121 0.002010 **
LeagueN       43.11162   39.96612   1.079 0.281755
DivisionW   -111.14603   39.21835  -2.834 0.004970 **
PutOuts        0.28941    0.07478   3.870 0.000139 ***
Assists        0.26883    0.15816   1.700 0.090430 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 311.7 on 251 degrees of freedom
Multiple R-squared:  0.5426,    Adjusted R-squared:  0.5226
F-statistic: 27.07 on 11 and 251 DF,  p-value: < 2.2e-16
```



Prüfgröße 27,07 ist größer als das .95-Quantil der F-Verteilung (=1,827).
P-Wert ist sehr klein. H0 ist zum Signifikanzniveau $\alpha = 0.05$ abzulehnen!

```
Call:
lm(formula = log(Volume) ~ log(Diameter) + log(Height), data = df)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.169537 -0.048572  0.004428  0.063542  0.129237 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.57663    0.69137   -2.28   0.0304 *  
log(Diameter)  1.98371    0.07522   26.37  < 2e-16 ***
log(Height)    1.11439    0.20485    5.44 8.34e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08154 on 28 degrees of freedom
Multiple R-squared:  0.9776,	Adjusted R-squared:  0.976 
F-statistic: 610.4 on 2 and 28 DF,  p-value: < 2.2e-16
```

Abbildung 2: 2. Aufgabe: Ausgabe Schätzergebnis aus der Statsitik-Software **R**.

2. (10 Punkte) Mit den Daten aus Aufgabe 1 ist ein weiteres Modell geschätzt worden.

$$\log(Volume_i) = \beta_0 + \beta_1 \log(Diameter_i) + \beta_2 \log(Height_i) + u_i$$

Die Ergebnisse sind in Abbildung 2 zu sehen.

(a) In der Praxis werden häufig Variablentransformationen durchgeführt (z.B. durch den Einsatz von log()), bevor lineare Modelle geschätzt werden. Warum wird das gemacht?

(b) Machen Sie bitte eine Aussage, ob und wie die beiden Modelle aus Aufgabe 1 und Aufgabe 2 miteinander vergleichbar sind.

LÖSUNG:

(a) (5 Punkte) Die lineare Regression schätzt einen linearen Zusammenhang in den Daten. Eine notwendige Annahme ist, dass der Zusammenhang der Variablen linear ist. Ist das nicht gegeben, helfen Variablentransformationen bei der Überführung in lineare Zusammenhänge.

(b) (5 Punkte) Die Modelle aus Aufgabe 1 und Aufgabe 2 sind nicht miteinander vergleichbar, da die Modelle unterschiedliche zu erklärende Variablen besitzen.

3. (20 Punkte) In der multiplen Regression ist der Forward-Algorithmus eine Möglichkeit, um systematisch, insbesondere bei einer hohen Anzahl an erklärenden Variablen, ein Modell mit hoher Güter aus der Vielzahl an möglichen Modellen zu bestimmen. Erklären Sie die dafür notwendigen Schritte. Gehen Sie insbesondere auf die Verwendung und Interpretation des korrigierten R^2 ein.

(20 Punkte) Die Variable y ist mit den Variablen $x_1, \ldots, x_p$ zu erklären. Die Forward Selektion ist ein systematisches Verfahren, um aus der Vielzahl an Modellen ein Modell mit hoher Güte zu bestimmen.

- 4 Punkt: p einfache lineare Regressionen schätzen für jede einzelne erklärende x-Variable

$$y = \beta_0 + \beta_1 x_j$$

  für $j = 1, ..., p$
- 4 Punkt: Die erklärende Variable x_j mit dem Modell der geringsten Residuenquadratesumme auswählen
- 4 Punkt: $p - 1$ lineare Regressionen schätzen zur Erklärung von y

$$y = \beta_0 + \beta_1 x_j + \beta_2 x_k$$

  füer $k = 1, ..., p - 1$ mit $k/neqj$
- 4 Punkt: Die erklärende Variable x_k mit dem Modell mit der geringsten Residuenquadratesumme auswählen
- 4 Punkt: u.s.w. bis eine Stopp-Regel erreich ist, z.B. wenn ein bestimmter Wert des korrigierten R^2 erreicht ist.

Korrigiertes R^2 wird für Modellvergleich verwendet, wenn die Modelle die gleiche zu erklärende Variable besitzen, aber mit unterschiedlicher Anzahl an erklärenden Variablen. Hintergrund: das R^2 neigt dazu, größer zu werden bei zunehmender Anzahl an Variablen, auch wenn die weiteren dazukommenden Variablen nicht weiter zu Erklärung beiträgt.

Exkurs 2: Modellvergleich

	Dependent variable:	
	y	
	(1)	(2)
x1	-4.059^{***}	-4.055^{***}
	(0.039)	(0.039)
x2		-0.127
		(0.135)
gender	1.999^{***}	2.039^{***}
	(0.276)	(0.280)
Constant	70.112^{***}	74.343^{***}
	(1.981)	(4.931)
Observations	40	40
R^2	0.997	0.997
Adjusted R^2	0.997	0.996
Residual Std. Error	0.840 (df = 37)	0.842 (df = 36)
F Statistic	$5{,}569.701^{***}$ (df = 2; 37)	$3{,}701.273^{***}$ (df = 3; 36)
Note:		*p<0.1; **p<0.05; ***p<0.01

Hier ist Modellvergleich möglich, da gleiche zu erklärende Variable. Modell (1) und Modell (2) haben unterschiedliche Anzahl erklärender Variablen, daher wird hier das korr. R^2 verwendet.

4. (20 Punkte) Empirische Korrelation

(a) Was misst die empirische Korrelation?

(b) Wie ist die empirische Korrelation zu interpretieren?

(c) Wie wird die empirische Korrelation von zwei metrischen Variablen X und Y berechnet?

(d) Die Variablen X und Y besitzen eine metrische Skala. Was ist die Besonderheit einer Ordinalskala?

(e) Wie kann eine Korrelation berechnet werden, wenn zwei ordinal-skalierte Variablen Z und V vorliegen? Was misst diese berechnete Korrelation?

LÖSUNG:

(a) (2 Punkte) Linearen Zusammenhang

(b) (4 Punkte) Der Korrelationskoeffizient ist begrenzt zwischen -1 und 1. Die Extremwerte treten auf, wenn alle Beobachtungen auf einer Gerade liegen. Es gilt $r_{XY} = 1$ auf Gerade mit positiver Steigung und $r_{XY} = -1$, auf Gerade mit negativer Steigung. Je näher die Beobachtungspunkte an einer Geraden liegen, umso näher ist r bei 1 bzw. bei -1.
$r_{XY} \approx 0$ gibt an, das kein linearer Zusammenhang in den Daten vorliegt, es kann aber dennoch ein nicht-linearer Zusammenhang in den Daten vorkommen.

(c) (5 Punkte)

$$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

(d) (2 Punkte) Ordinalskala: die Datenpunkte besitzen eine Ordnung, aber die Abstände sind nicht interpretierbar.

(e) (7 Punkte) Spearman-Korrelationskoeffizient wird bestimmt. Es handelt sich dabei um die Messung des monotonen (nicht linearen) Zusammenhangs. Die Werte X und Y werden geordnet und die Ränge der Beobachtung werden genutzt. Für die Rangpaare $rg(X_i), rg(Y_i)$ wird dann der Korrelationskoeffizient berechnet.

5. (20 Punkte) In einer Qualitätskontrolle wird überprüft, ob ein Produktionsprozess eines Bauteils noch gemäß der Annahme einer Normalverteilung mit einem Sollwert von $\mu_x = 15mm$ und einer Standardabweichung von $\sigma_x = 0,5mm$ produziert. Dafuür wird in regelmäßigen Abständen eine Stichprobe der Größe $n = 25$ von den Bauteilen aus der Produktion gezogen. Der Mittelwert der Stichprobe liegt bei $15,25mm$.

   (a) Weicht die Länge des Bauteils statistisch signifikant zum Niveau $\alpha = 0,05$ vom Sollwert ab? Es gilt für die Verteilungsfunktion $\Phi()$ der Standardnormalverteilung:

$$\Phi(-1,96) = 0,025, \Phi(0,015) = 0,506, \Phi(1,65) = 0,95$$

   (b) Was sind die Fehler 1. Art und 2. Art bei statistischen Hypothesentests?

LÖSUNG:

   (a) (15 Punkte) $H_0 : \mu = 15mm \quad H_1 : \mu \neq 15mm$ (3 Punkte)
   Prüfgröße (Berechnung 6 Punkte):

$$z = \frac{15,25 - 15}{0,5}\sqrt{25} = 0,5 \cdot 5 = 2,5$$

   Vergleich der Prüfgröße mit kritischem Wert der Normalverteilung $1,96$. Rechts von $1,96$ liegt $0,025$ und links von $-1,96$ liegt $0,025$ an Fläche, das aufsummiert $0,05$ ergibt. (3 Punkte)
   Die Prüfgröße $z = 2,5$ ist größer als $1,96$, damit ist H_0 abzulehnen. (3 Punkte)

   (b) (5 Punkte) Fehler 1. Art: H_0 wird verworfen, obwohl H_0 wahr ist
   Fehler 2. Art: H_0 wird behalten, obwohl H_1 wahr ist
   Der Fehler 1. Art wird durch das Signifikanzniveau α kontrolliert.

Wiederholung

Schätzungen von linearen Modellen basieren auf Annahmen bzgl. f und Störgröße epsilon. Auswirkungen der Verletzungen der Annahmen.

Wiederholung

Bedingte WSK, UND-WSK, Unabhängigkeit von Ereignissen

Bayessche Netze

Kurze Wiederholung Wahrscheinlichkeiten

- UND Verknüpfung von Ereignissen $p(E1 \cdot E2) = p(E2|E1)p(E1)$
- Unabhängigkeit $p(E1 \cdot E2) = p(E2)p(E1)$
- Totale Wahrscheinlichkeit

$$P(A) = \sum_{j=1}^{n} P\left(E_j\right)p(A|E_j)$$

Mit $\sum_{j=1}^{n} P\left(E_j\right) = 1$ und für alle $j \neq k$ gilt $P\left(E_j \cdot E_k\right) = 0$ (gegenseitiges Ausschließen der Ereignisse)

- Satz von Bayes $P\left(E_j \middle| A\right) = \frac{P\left(E_j\right)P(A|E_j)}{P(A)}$
- Grenzen von Naive Bayes Klassifikatoren: bedingte Unabhängigkeit sehr restriktiv
- Unsupervised Learning:
    - Vielzahl an Zielgrößen (z.B. Fehler im Drucker, Diagnosen)
    - Unsichere Informationen
    - Fehlende Informationen

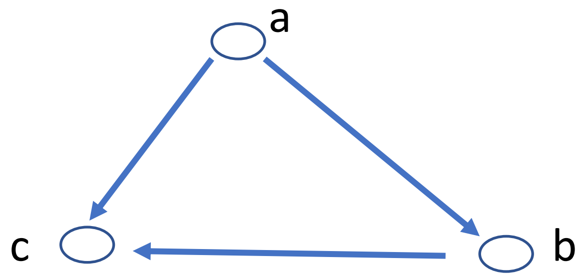
Bayessche Netze: Modelle zur Umsetzung von lernenden Systemen

- Was ist ein Bayessches Netz?
  - Gerichteter azyklischer Graph
  - Annotation dieses Graphen mit Wahrscheinlichkeiten, deren Kombination eine Verteilung definiert
  - Knoten sind Variablen
  - Kanten sind Wahrscheinlichkeitsbeziehungen

Gemeinsame Wahrscheinlichkeitsverteilung

$$p(a, b, c) = p(c|a, b) \cdot p(a \cdot b) = p(c|a, b) \cdot p(b|a) \cdot p(a)$$

(Kettenregel)



Zerlegung hätte auch anders aussehen können (z.B. $p(a|b, c) \cdot p(b|c) \cdot p(c)$), dann wäre es aber eine andere grafische Darstellung.

Gemeinsame Wahrscheinlichkeitsverteilung

Die gemeinsame Verteilung definiert durch einen Graphen (Bayessches Netz). Gemeinsame Wahrscheinlichkeitsverteilung ist gegeben durch Produkt über alle Knoten vom Graphen und den bedingten Verteilungen für jeden Knoten in Abhängigkeiten zu den Eltern des jeweiligen Knotens $X = \{x_1, \dots, x_K\}$

$$p(X) = \prod_{k=1}^{K} p(x_k | pa_k)$$

mit pa_k ist Set an Eltern (parents) für x_k. Der Graph hat K Knoten.

Anwendung Bayessche Netze

- Abbildung Expertenwissen
- Regelbasierte Systeme mit Unsicherheiten erweitern

Lernen mit Bayesschen Netzen

- Netztopologie/Struktur
- Parameter für jeden Knoten (bedingte Wahrscheinlichkeitstabelle für jede Variable)

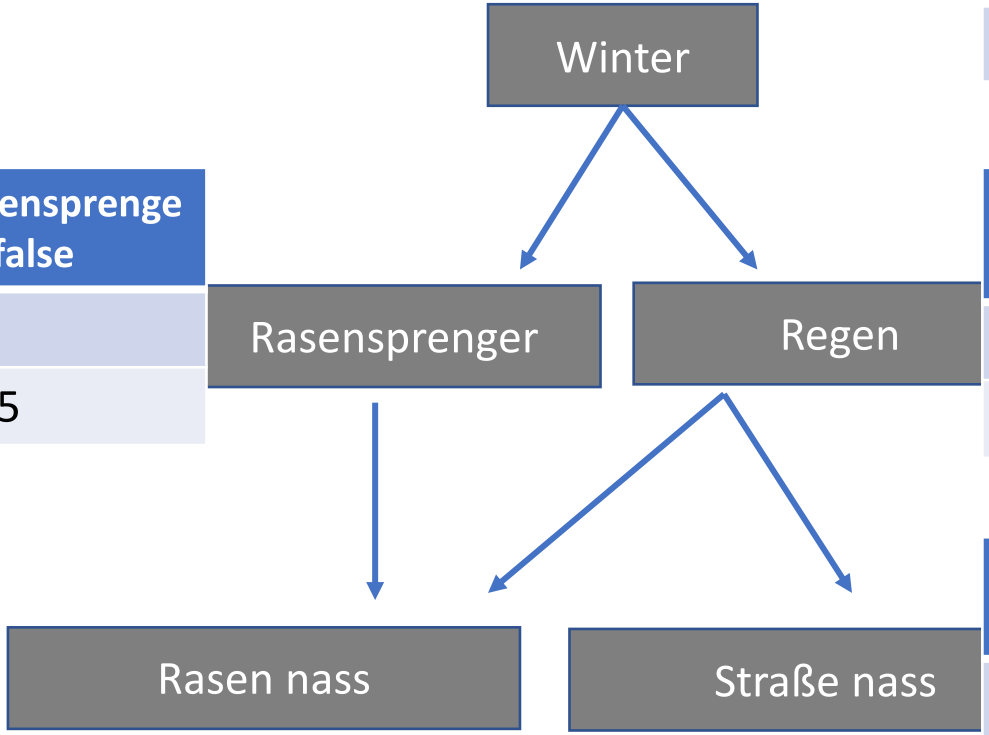
Hier:

- Annahme diskrete Variablen
- Experte gibt Netzstruktur vor (sonst z.B. über Greedy Search)
- Trainingsdaten vorhanden

Dann Maximum-Likelihood Schätzung möglich:

$$\theta_{i,j,k} = p(x_i = k | pa_i = j) = \frac{N_{ijk}}{\sum_k N_{ijk}}$$

Beispiel



Winter = true	Winter = false
0,6	0,4

Winter	Rasensprenger = true	Rasensprenger = false
true	0,2	0,8
false	0,75	0,25

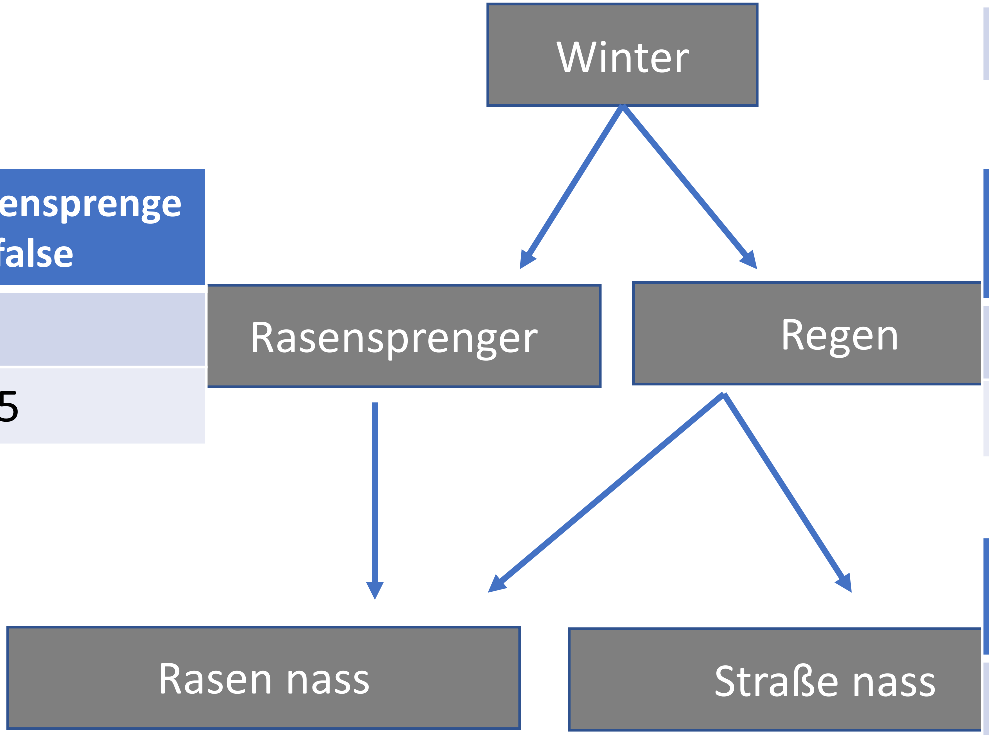
Winter	Regen = true	Regen = false
true	0,8	0,2
false	0,1	0,9

Regen	Straße nass = true	Straße nass = false
true	0,7	0,3
false	0	1

Rasensprenger	Regen	Rasen nass = true	Rasen nass = false
true	true	0,95	0,05
true	false	0,9	0,1
false	true	0,8	0,2
false	false	0	1

Beispiel

Winter = true	Winter = false
0,6	0,4



Winter	Rasensprenger = true	Rasensprenger = false
true	0,2	0,8
false	0,75	0,25

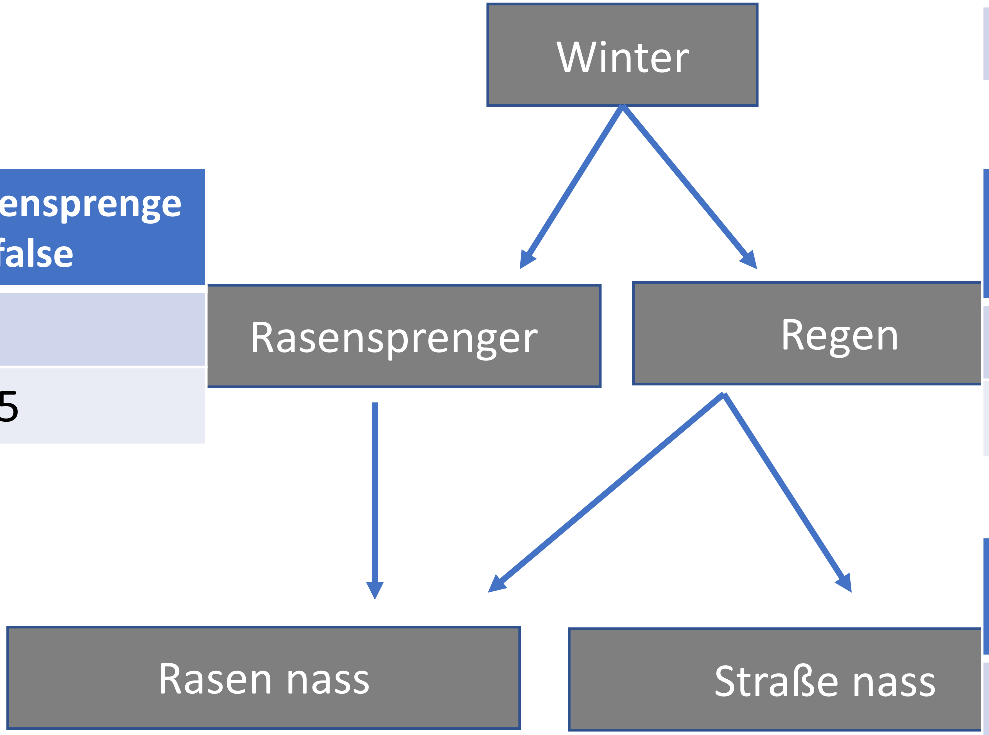
Winter	Regen = true	Regen = false
true	0,8	0,8
false	0,1	0,9

Regen	Straße nass = true	Straße nass = false
true	0,7	0,3
false	0	1

Rasensprenger	Regen	Rasen nass = true	Rasen nass = false
true	true	0,95	0,05
true	false	0,9	0,1
false	true	0,8	0,2
false	false	0	1

Gemeinsame Verteilung:
P(**Winter** und **kein Rasensprenger** und **Regen** und **Rasen nass** und **Straße nass**)?

Beispiel



Winter = true	Winter = false
0,6	0,4

Winter	Rasensprenger = true	Rasensprenger = false
true	0,2	0,8
false	0,75	0,25

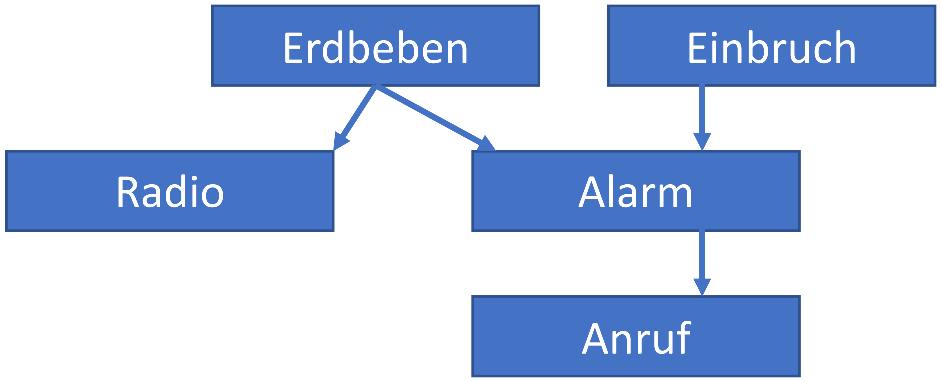
Winter	Regen = true	Regen = false
true	0,8	0,8
false	0,1	0,9

Regen	Straße nass = true	Straße nass = false
true	0,7	0,3
false	0	1

Rasensprenger	Regen	Rasen nass = true	Rasen nass = false
true	true	0,95	0,05
true	false	0,9	0,1
false	true	0,8	0,2
false	false	0	1

Gemeinsame Verteilung:
P(**Winter** und **kein Rasensprenger** und **Regen** und **Rasen nass** und **Straße nass**) =
0,6 * 0,8 * 0,8 * 0,7 * 0,8 = 0,21504

Explaining Away



- Erdbeben ist unbekannt
- Radio und Anruf sind abhängig (Tail-to-Tail)
- Es gilt dann:

$$P(\text{Einbruch}|\text{Anruf}) > P(\text{Einbruch}|\text{Anruf, Radio})$$

- Radiomeldung über Erdbeben erklärt einen Alarm, daher wird Einbruch unwahrscheinlicher. (Zahlenbeispiel dazu im Skript)

Zahlenbeispiel aus Skript

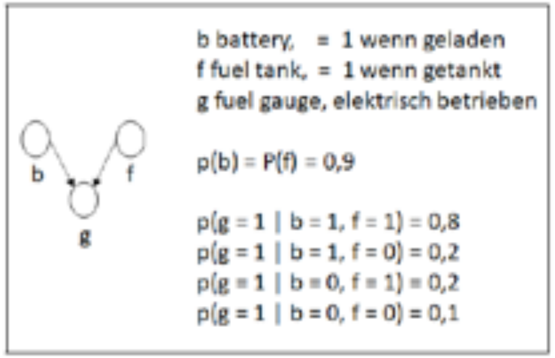
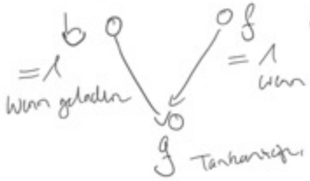


Abbildung 49: Beispiel Head to Head.



$p(b) = p(f) = 0{,}9$

Explaining away:

die Erklärung eines Ereignisses reduziert die WSK möglicher anderer Ursachen:

wenn $g = 0$, wie hoch ist WSK, $f = 0$?

$$p(g=0) = \sum_b \sum_f p(g=0 \mid b, f) \cdot p(b) p(f)$$

$$= \underbrace{0{,}1}_{p(b=0)} \cdot \left(\underbrace{0{,}1}_{p(f=0)} \cdot 0{,}9 + \underbrace{0{,}9}_{p(f=1)} \cdot 0{,}8 \right)$$

$$+ \underbrace{0{,}9}_{p(b=1)} \left(\underbrace{0{,}1}_{p(f=0)} \cdot \underbrace{0{,}8}_{p(g=0 \mid b=1, f=0)} + \underbrace{0{,}9}_{p(f=1)} \cdot 0{,}2 \right)$$

$$= 0{,}315$$

$$p(g=0 \mid f=0) = \sum_b p(g=0 \mid b, f=0)\, p(b)$$

$$= p(b=0) \cdot p(g=0 \mid b=0, f=0)$$

$$+ p(b=1) \cdot p(g=0 \mid b=1, f=0)$$

$$= 0{,}81$$

$$p(f=0 \mid g=0) = \frac{p(g=0 \mid f=0) \cdot p(f=0)}{p(g=0)}$$

$$\approx 0{,}257 > \underbrace{0{,}1}_{p(f=0)}$$

$g=0$ beobachten, dann liefert das Evidenz, dass Tank leer ist, daher steigt WSK für leeren Tank.

Wenn aber $b=0$ beobachtet:

$$p(f=0 \mid g=0, b=0) = \frac{p(g=0 \mid b=0, f=0)\, p(f=0)}{\sum_f p(g=0 \mid b=0, f)\, p(f)}$$

$$\approx 0{,}111$$

Inferenz

- Alarm, wie wahrscheinlich ist ein Einbruch?
  Einbruch E → Alarm A:

$$P(E|A) = \frac{P(E \cdot A)}{P(A)}$$

- Anruf, wie wahrscheinlich ist ein Einbruch?
  Einbruch E→ Alarm A → (Telefon-)Anruf T:

$$P(T) = P(E \cdot T) + P(kein\ E \cdot T)$$

$$P(E|T) = \frac{P(E \cdot T)}{P(T)}$$