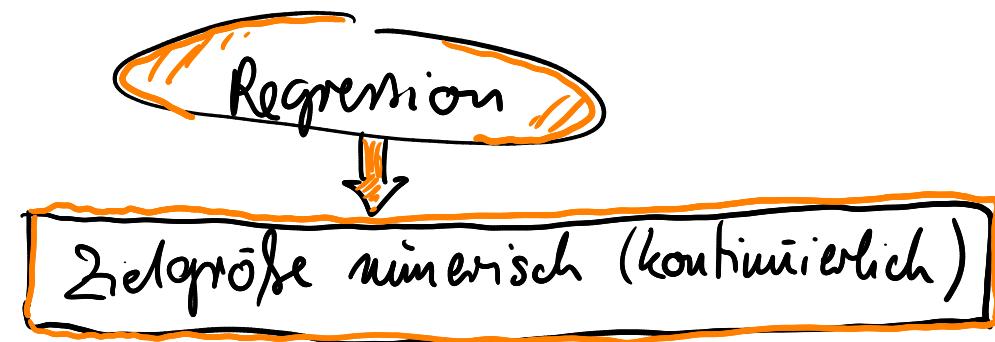


II. KLASISCHE VERFAHREN FÜR REGRESSION & KLASIFIKATION

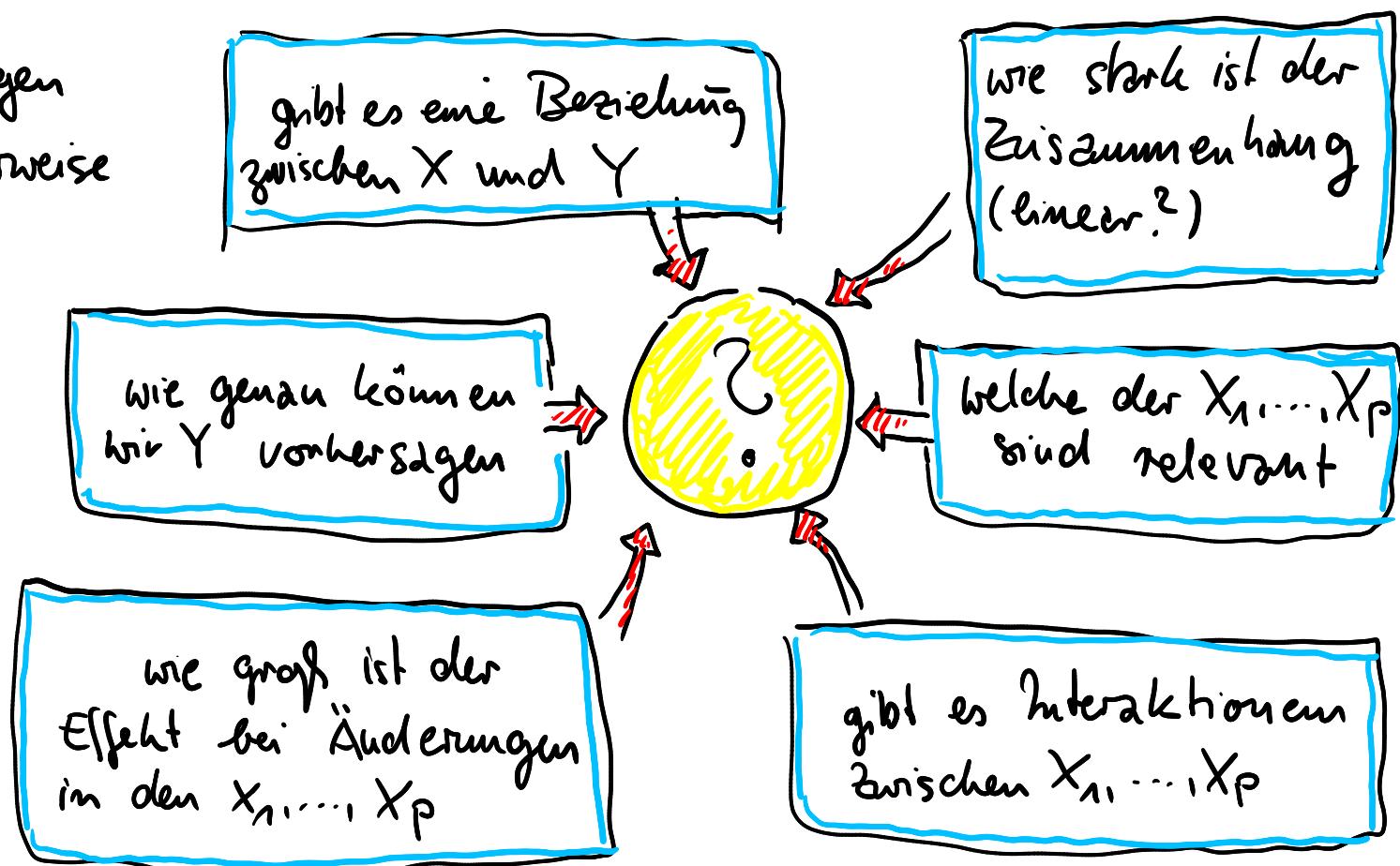
56



Bei typischen Anwendungen
stellt man sich üblicherweise
die folgenden Fragen:



Au Beispiel des
linearen Regressions-
modells lässt sich
gut über mögliche
Antworten nachdenken...



2.1) Lineare Regression aus Sicht des Machine Learning

Betrachte erklärende Größen X_1, \dots, X_p und die numerische Zielgröße Y mit n Realisierungen $(x_{i1}, \dots, x_{ip}, Y_i)$ ($i=1, \dots, n$) \leftarrow Datensamples

Annahme: die lineare Funktion

$$f(X_1, \dots, X_p) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

approximiert die Zielgröße Y

Linearität ist eine apriori Hypothese über die Struktur der Transferfunktion f (Modell)

zu schätzende Parameter $\beta_0, \beta_1, \dots, \beta_p$ ($p+1$ Stück)

→ Anzahl hängt von der Struktur von f ab

Im einfacheren Fall gibt es nur eine erklärende Größe X

$$f(X) = \beta_0 + \beta_1 X$$

2 Parameter: β_0, β_1

$$Y = f(X) + \varepsilon$$

2.1.1) Lineare Einfachregression

viele grundsätzliche Überlegungen zu (ellg.) linearen Modellen lassen sich am diesem einfachen Beispiel nach vollziehen.

MODELL

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \leftarrow \text{Störgröße}$$

$\underbrace{}_{f(X)}$

Sobald wir Schätzungen $\hat{\beta}_0, \hat{\beta}_1$ für β_0, β_1 haben können wir mit dem Modell Prognosen machen:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Prognose für Y
basierend auf X

und es gilt

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i$$

$\underbrace{\phantom{Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i}}_{\hat{Y}_i}$

empirische
lineare
Beziehung

also Residuum

$$\varepsilon_i = Y_i - \hat{Y}_i$$



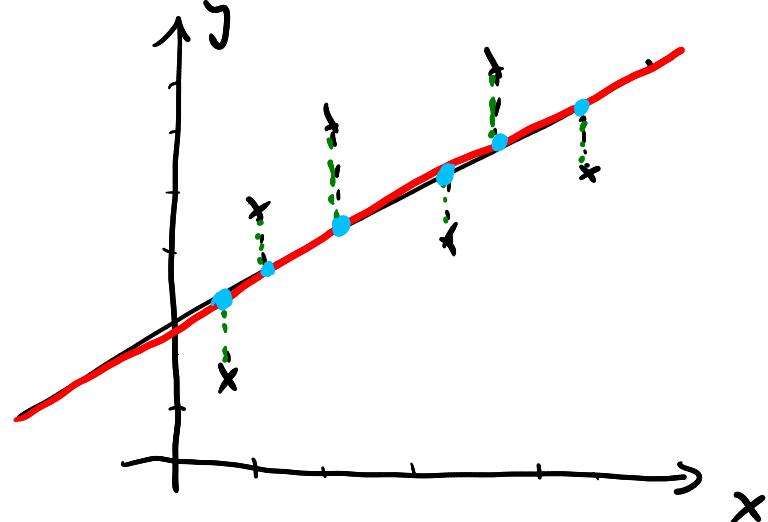
59

Schätze β_0, β_1 auf Basis der Trainingsdaten $(x_1, y_1), \dots, (x_n, y_n)$
so, dass die resultierende Schätzgerade optimal zu den Daten passt!

Wir brauchen also

① Fehlerfunktion, die passt wie gut es passt

② Verfahren um die Parameter mit Hilfe der Fehlerfunktion optimal zu belegen



HIER: Methode der kleinsten Quadrate

KQ - Verfahren

schon im letzten Semester gesehen \Rightarrow Wiederholung

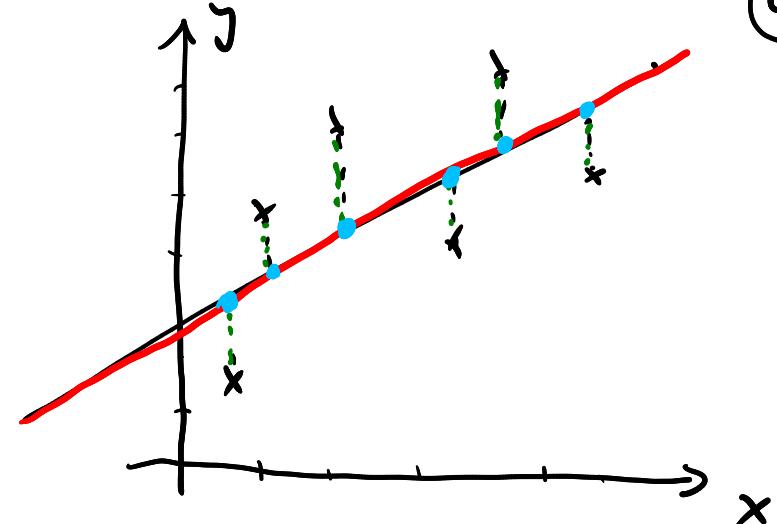
↪ eine Möglichkeit (es geht auch noch anders ...)

IDEE Sei $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

(60)

und $e_i = y_i - \hat{y}_i$ Residuum

$$\Rightarrow \text{Err}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$



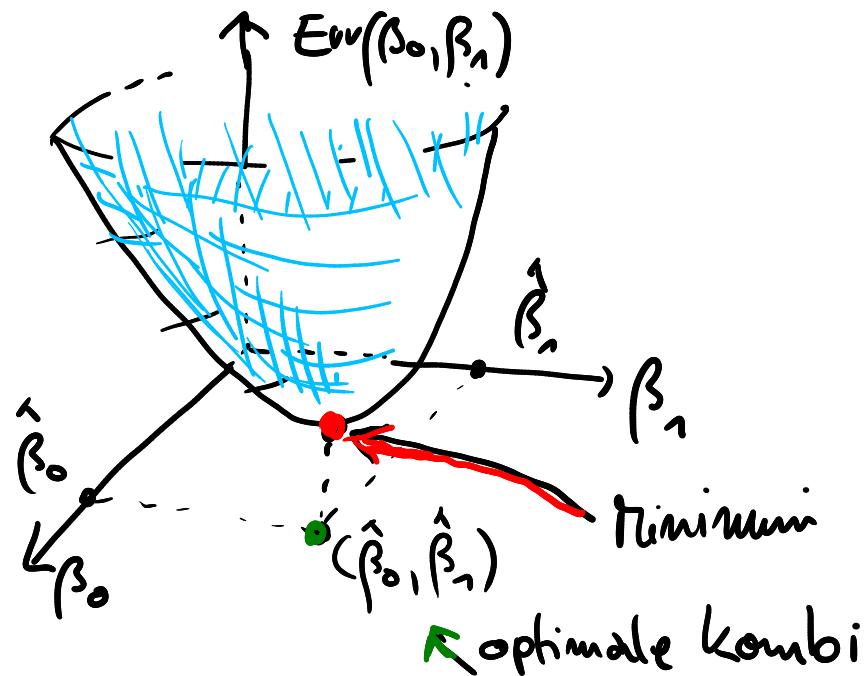
ist geeignete Fehlerfunktion (in $\hat{\beta}_0, \hat{\beta}_1$)

"Optimale Gerade" heißt dann: $\text{Err}(\hat{\beta}_0, \hat{\beta}_1)$ ist minimal!

$$\Rightarrow \text{grad}(\text{Err}(\hat{\beta}_0, \hat{\beta}_1)) = \vec{0}$$

optimale $\hat{\beta}_0, \hat{\beta}_1$

für



Zum Fall der Einfachregression ist $\min(\text{Err}(\hat{\beta}_0, \hat{\beta}_1))$ eindeutig und die Lösung lässt sich analytisch bestimmen!

Es ergeben sich Schätzungen

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= \frac{\text{cov}(x, y)}{s_x^2}\end{aligned}$$

wobei $s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ empirische Varianz von x

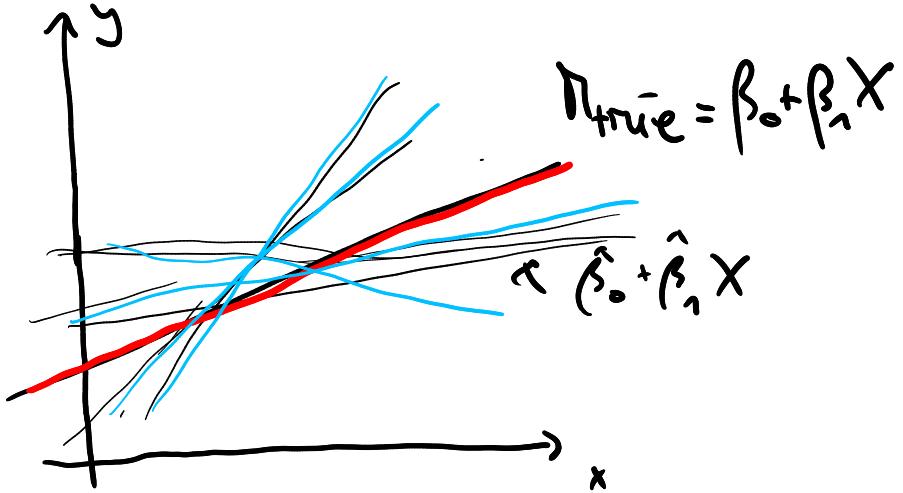
$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{kovarianz von } x \text{ und } y$$

Formal gesehen liefert uns das LQ-Verfahren Schätzfunktionen (zV) für β_0, β_1

$$\begin{aligned}T_{\beta_0}(x, Y) &= \bar{Y} - \beta_1 \bar{X} \\ T_{\beta_1}(x, Y) &= \frac{\text{cov}(X, Y)}{\text{Var}(X)}\end{aligned}$$

← sind sogar erwartungstreu – also Top 😊
(für ε unabh. von X)

⇒ Schätzen wir β_0, β_1 für unterschiedliche Trainingsdaten, dann kriegen wir auch verschiedene Realisierungen $\hat{\beta}_0, \hat{\beta}_1$ heraus!



ABER wegen der Erwartungstreue kommt der Mittelwert von α vielen dieser Schätzungen beliebig nahe an die echten, unbekannten β_0, β_1 heran ... ($\text{bias} = 0$)

... das können wir aber leider so nicht machen (zu wenige Daten, Rechenpower...)



Wie kann kommen wir mit einer konkreten Schätzung zu β_0, β_1 heran?

Der Standardfehler der Schätzung ist eine Möglichkeit das zu quantifizieren:

$$\sigma_{T_{\beta_i}} = \sqrt{\text{Var}(T_{\beta_i}(X, Y))} \quad (i=0,1)$$

Standardfehler der Schätzung für β_0, β_1

Sei ε Störgröße mit Realisierungen ε_i , die unkorreliert sind und Varianz σ^2 besitzen. Dann sind die Standardfehler der Schätzung $\hat{\beta}_0$ und $\hat{\beta}_1$ gegeben durch:

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

und

$$\text{Var}(\hat{\beta}_1) = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$n \cdot s_x^2$
emp. Varianz von x

Beweis: Skript

Bemerkung: σ^2 ist unbekannt, aber wir können $e_i = y_i - \hat{y}_i$ als Realisierungen von ε interpretieren und σ^2 (erwartungstreu) schätzen

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

möglich wegen Erwartungstreue!

hier Größe n :

$$\frac{1}{n-2} \approx \frac{1}{n}$$

Betrachten wir nochmal die Standardfehler:

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

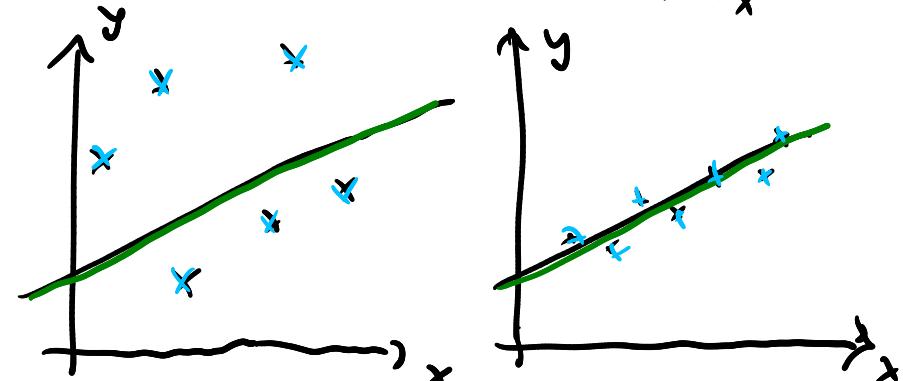
und

$$\text{Var}(\hat{\beta}_1) = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$n \cdot s_x^2$

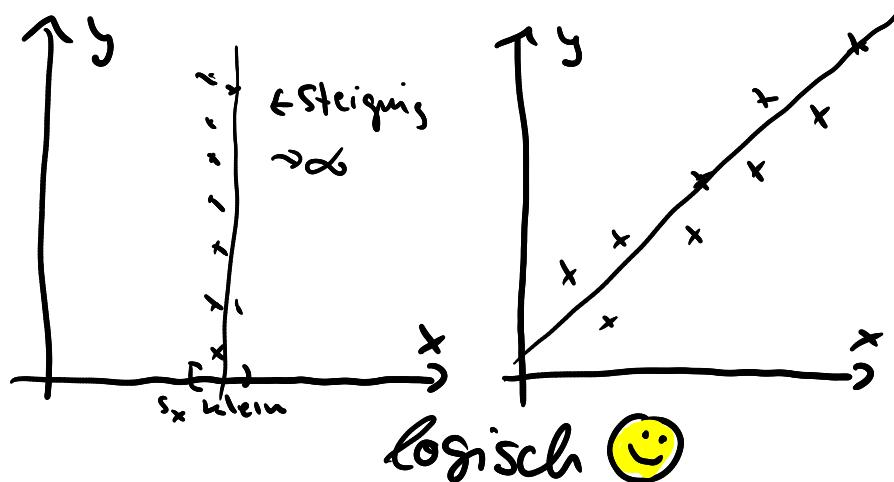
INTERPREATION

① Ist σ^2 gross \Rightarrow Standardfehler gross



logisch 😊

② Ist s_x^2 klein \Rightarrow Standardfehler gross



logisch 😊

... und der Einfluss der Daten? ?

$$\textcircled{3} \text{ wegen } \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 \sim \frac{1}{n-2}$$

\Rightarrow für großes n (vielen Daten) $\Rightarrow \hat{\sigma}^2$ klein \Rightarrow Standardfehler klein

Das ist auch keine große Überraschung und deckt sich mit unserer Erfahrung!



Welche Möglichkeiten gibt es noch um Qualität der Parameterschätzungen sowie Eignung des Modells zu bewerten? ?

1 Konfidenzintervalle

2 Hypothesentests

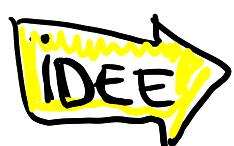
... noch mehr Statistik 

werden von Software automatisch ausgewertet ... \Rightarrow wichtig: Interpretation!

EXKURS : Konfidenzintervalle & Hypothesentests



besonders für kleine Stichproben sind Schätzungen häufig sehr ungenau! (gilt auch für unsere Modelle!)



Konstruiere ein Intervall $I = [I_u, I_o]$, so dass der unbekannte Parameter θ mit einer vorgegebenen Wahrscheinlichkeit von mind. $(1-\alpha)$ in I liegt.

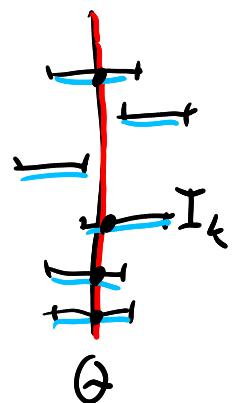
Also :

$$P(\theta \in I) \geq 1-\alpha$$

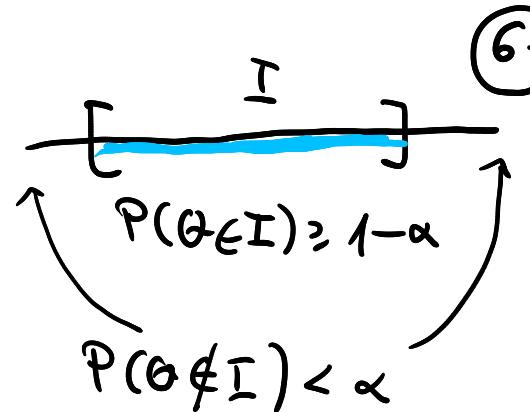
$1-\alpha$: Konfidenzniveau

α : Fehlerwahrscheinlichkeit

Interpretation : Zieht man k unabhängige Stichproben und konstruiert Intervalle I_k , dann enthalten $k(1-\alpha)$ davon das echte (unbekannte) θ

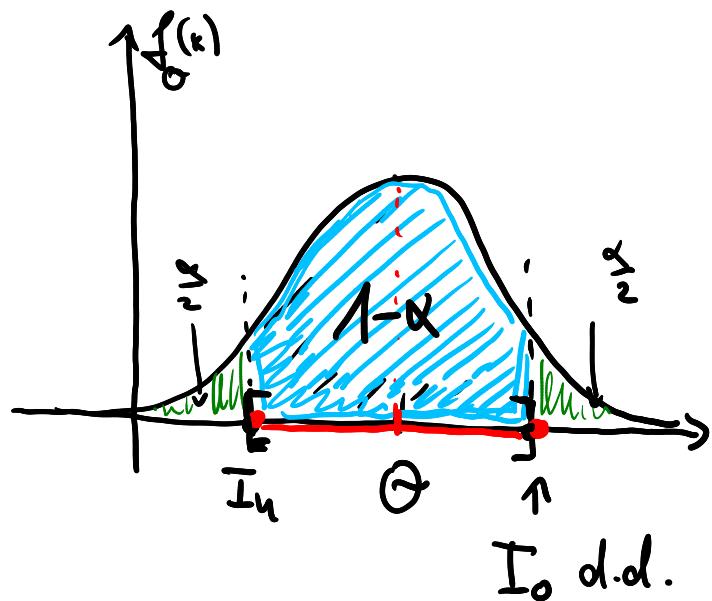


Hat man also $I = [I_u; I_o]$ konstruiert, so bleibt ein "Risiko" von α übrig, d.h. $\underline{\text{Q NICHT}}$ drin ist ...



Wie bestimmt man so ein Intervall?

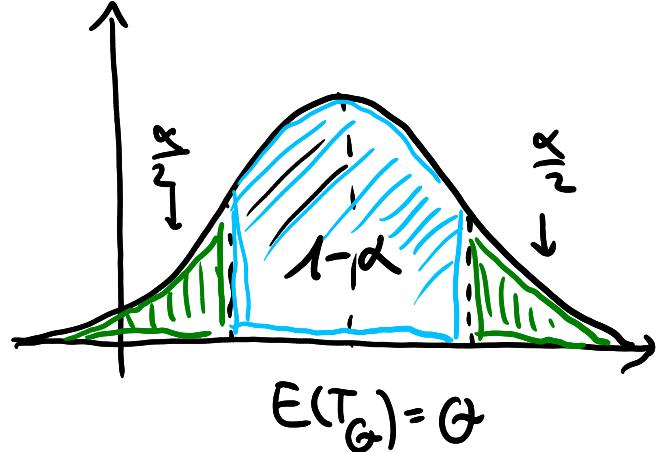
⇒ Die Verteilung der Schätzfunktion $T_Q(x)$ ist ausschlaggebend ...



$$P(T_Q < I_o) = 1 - \frac{\alpha}{2}$$

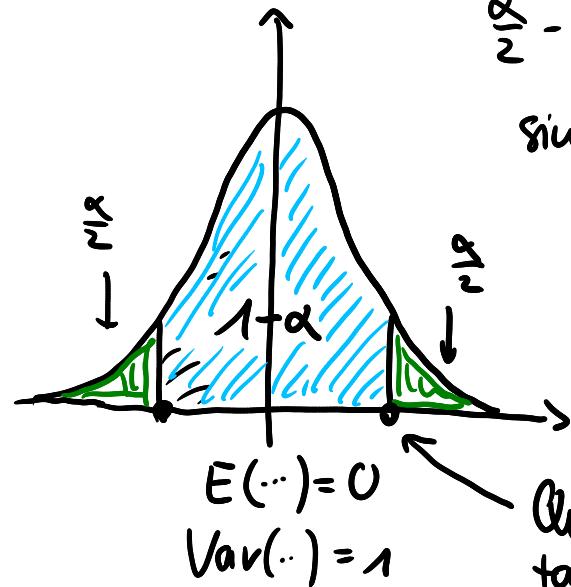
← aus der Verteilung der Schätzfunktion können wir Aussagen über die Wahrscheinlichkeit machen bestimmte Werte von $\underline{\text{Q}}$ zu beobachten!

Für konkrete Rechnungen benutzt man meistens die standardisierten Verteilungen ...



Transformation

$$\tilde{T}_0 = \frac{T_0 - \bar{\alpha}}{\sqrt{\text{Var}(T_0)}}$$



$\frac{\alpha}{2}$ -Quantile
sind Grenzen
von \tilde{T}

In Fall von $\hat{\beta}_0$ und $\hat{\beta}_1$ gilt:

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\text{Var}(\hat{\beta}_0)}} \sim t(n-2)$$

sowie

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}(\hat{\beta}_1)}} \sim t(n-2)$$

t -Verteilt mit
(n-2) Freiheitsgraden
↑
spezielle
Verteilung

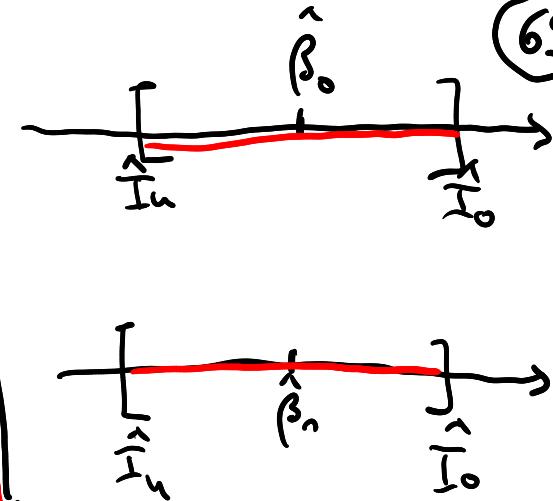
sh. Statistik-Vorlesung

aus den Daten ergeben sich dann die folgenden Intervalle:

rechnet der
Computer für
uns aus: :

$$\hat{I}_{\beta_0} = \left[\hat{\beta}_0 - t_{1-\frac{\alpha}{2}} \hat{\sigma}_{\beta_0}; \hat{\beta}_0 + t_{1-\frac{\alpha}{2}} \hat{\sigma}_{\beta_0} \right]$$

$$\hat{I}_{\beta_1} = \left[\hat{\beta}_1 - t_{1-\frac{\alpha}{2}} \hat{\sigma}_{\beta_1}; \hat{\beta}_1 + t_{1-\frac{\alpha}{2}} \hat{\sigma}_{\beta_1} \right]$$



~ Beobachtung: die Grenzen liegen symmetrisch um $\hat{\beta}_0, \hat{\beta}_1$ und die Intervallbreite hängt jeweils von

① $t_{1-\frac{\alpha}{2}}$ $\leftarrow 1-\frac{\alpha}{2}$ -Quantil der t -Verteilung mit 2 Freiheitsgraden

② Streuung von $\hat{\beta}_1, \hat{\beta}_2$ ($\hat{\sigma}_{\beta_0}$ und $\hat{\sigma}_{\beta_1}$) ab

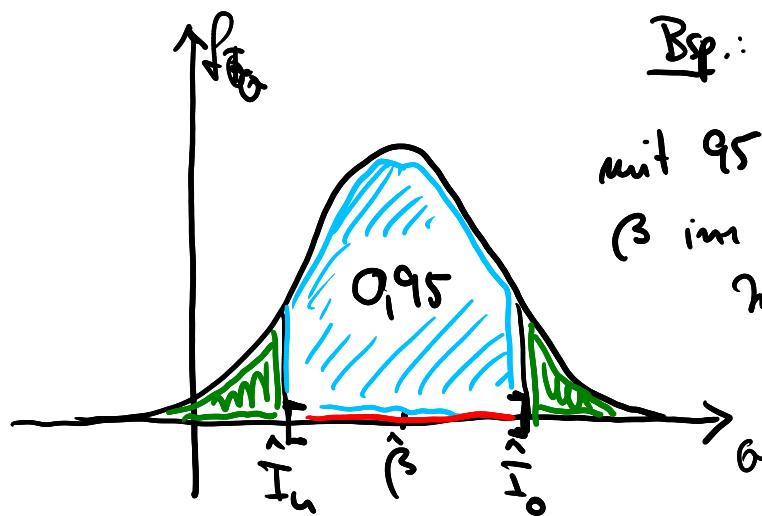
Was bedeutet "genaue" Schätzung??

↓ für festes $1-\alpha$ das Intervall klein \Rightarrow

(der Bereich in dem θ mit Sicherheit $(1-\alpha)$ liegt ist klein \Rightarrow genaue Schätzung!)

Welchen Einfluss hat die Auswahl von α auf das Intervall?

Wäre doch gut, wenn α möglichst klein ist. Oder ??

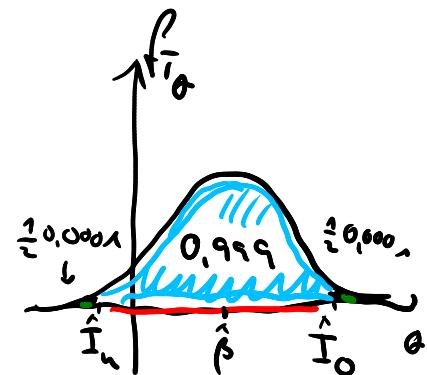


Bsp.: $\alpha = 0,05$ (5%) 70
mit 95% Sicherheit liegt β im konstruierten Intervall \hat{I}

Beobachtung: Bleiben alle anderen Größen fest, dann ergibt sich für kleineres α automatisch ein breiteres Intervall.

Für $\alpha \rightarrow 0 \Rightarrow I \rightarrow \mathbb{R} \Rightarrow P(Q \in \mathbb{R}) \approx 1$ "Q ist 100% sicher eine Zahl"

→ Das macht keinen Sinn!



Also

Die Irrtumswahrscheinlichkeit kann nicht beliebig klein gewählt werden.

Praxis: $\alpha = 0,05$ (default bei z.B. Python linregress, sm.OLS() etc.)

Was beeinflusst noch die Breite von I?

$\hat{\sigma}_{\beta_0}$ bzw. $\hat{\sigma}_{\beta_1}$ sind noch wichtig

Bsp: Varianz von β_1

$$\text{Var}(\hat{\beta}_1) = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

schon gesehen:

$\hat{\sigma}^2$ ist entscheidend

⇒ für großes n (viele Daten) wird \hat{I} automatisch kleiner!

⇒ Schätzungen werden zuverlässiger! ☺

Die Verteilung der Schätzfunktion kann auch benutzt werden um bestimmte Hypothesen bzgl. β_0, β_1 zu überprüfen.

Besonders wichtig: finde heraus, ob es einen (statistisch signifikanten) linearen Zusammenhang zwischen X und Y gibt!

Funktionsweise eines Tests:

① Aussage formulieren (z.B. über einen Parameter)

H_0 : Nullhypothese
 H_1 : Gegenhypothese



Gegenereignis zu H_0

hier wird das beweist, was eigentlich gezeigt werden soll

← die Nullhypothese kann durch den Test NICHT bewiesen werden.

Sie wird nur mit einer gewissen Wahrscheinlichkeit angenommen oder abgelehnt

② Bestimmen einer Prüfgröße + krit. Bereich K

$T(X)$ ZV Teststatistik ermitteln

$t = T(x_1, \dots, x_n)$ Realisierung

⇒ liegt $t \in K \Rightarrow H_0$ wird abgelehnt

$$P(H_0 \text{ ablehnen} \mid H_0 \text{ wahr}) \leq \alpha$$



Das Signifikanzniveau α bestimmt die Fehlerwahrscheinlichkeit beim Testen

Wird H_0 abgelehnt, so gilt H_1 mit Fehlerwahrscheinlichkeit α als bewiesen (Fehler 1. Art oder α -Fehler \leftarrow bekannt)

Allgemein gilt:

Wir brauchen

- ① sinnvolle Testgröße
- ② Verteilung des Schätzers im krit. Bereich zu ermitteln

Bsp. normalverteilte Daten

$X \sim N(0; 1)$ mit Realisierungen x_i ($i=1 \dots n$)

↪ unbekannt

Hypothese:

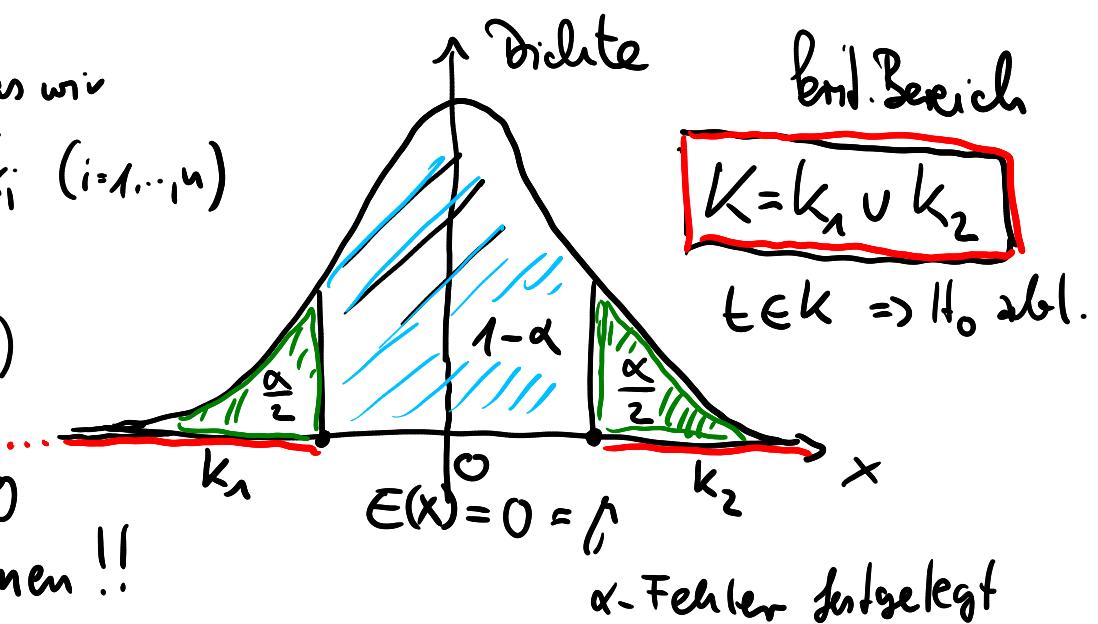
$$H_0: \mu = 0$$

$$H_1: \mu \neq 0$$

Testgröße: $t = \frac{1}{n} \sum_{i=1}^n x_i$

(NW)

⇒ liegt t "zu weit" von 0 entfernt müssen wir H_0 ablehnen !!

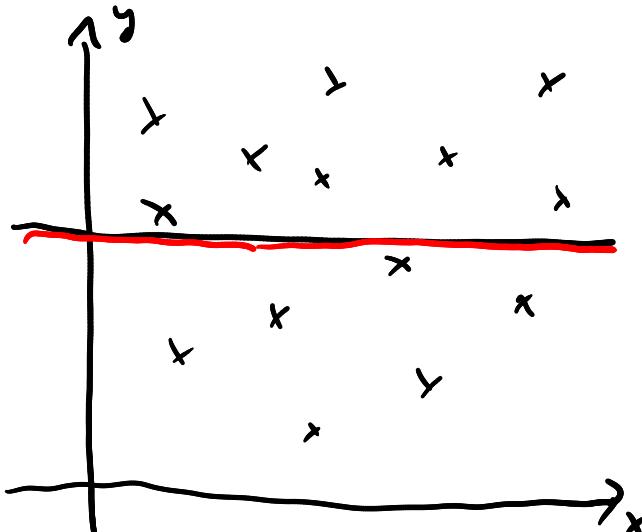




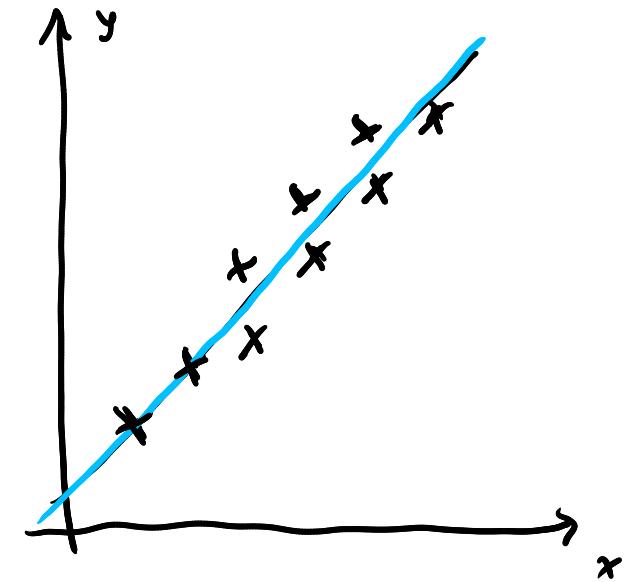
Wie formulieren wir geeignete Hypothesen um zu prüfen, daß ein lineares Modell Sinn macht?

Beobachtung:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



$$\begin{aligned} \text{cov}(x, y) &= 0 \\ \Rightarrow \hat{\beta}_1 &= 0 \end{aligned}$$

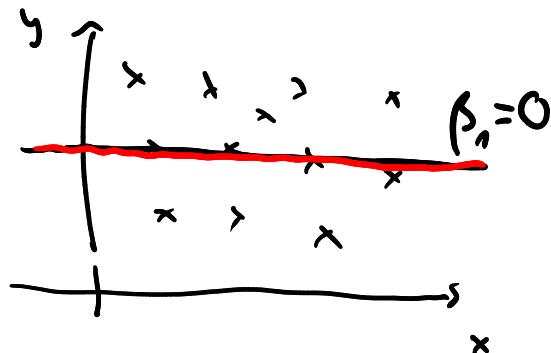


$$\begin{aligned} \text{cov}(x, y) &\neq 0 \\ \Rightarrow \hat{\beta}_1 &\neq 0 \end{aligned}$$

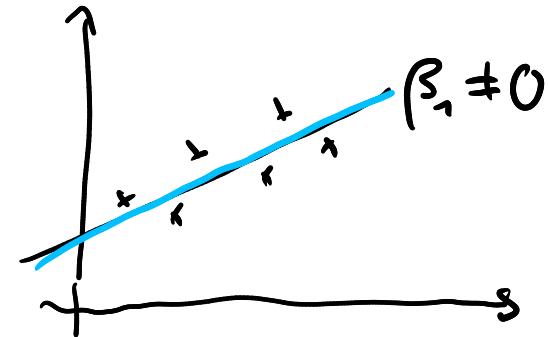
⇒ Ein (signifikant) von 0 verschiedener Steigungsparameter β_1 spricht für einen linearen Zusammenhang!



IDEE



75



Die geeigneten Hypothesen für einen statistischen Test sind also:

$$\begin{aligned} H_0: \beta_1 &= 0 \\ H_1: \beta_1 &\neq 0 \end{aligned}$$

Eine sinnvolle Testgröße beschreibt, ob (mit Fehler α) $\hat{\beta}_1$ "zu weit" von 0 weg liegt und die Nullhypothese verworfen werden muß $\Rightarrow H_1$ angenommen
 \Rightarrow lin. Modell sinnvoll

Testgröße:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\beta_1}}$$

\leftarrow Anzahl der Standardabweichungen, die $\hat{\beta}_1$ von 0 weg liegt
 \Rightarrow je größer $\hat{\sigma}_{\beta_1}$ desto größer muß $\hat{\beta}_1$ sein um H_0 zu verwirfen

Bem.: die Verteilung der Schätzfkt. bestimmt $K = (-\infty, -t^{(n-2), \frac{\alpha}{2}}] \cup [t^{(n-2), \frac{\alpha}{2}}, \infty)$

Interessant auch:

tatsächlich!

wie wahrscheinlich ist es, für $\beta_1 = 0 \downarrow$ eine Testgröße zu haben, die größer ist als die von uns aus den Daten berechnete?

⇒ solche Werte sorgen dafür, dass wir in K landen und folgern

H_0 abgelehnt $\Rightarrow H_1$ bewiesen \Rightarrow lin. Zusammenhang

obwohl tatsächlich
KEIN lin. Zshg.
besteht ↴ ↴

Def.: p-Wert

Der p-Wert ist die Wahrscheinlichkeit (unter der Bedingung, dass H_0 gilt) den beobachteten Wert der Teststatistik oder einen größeren zu erhalten. Der fiel dann eher in den krit. Bereich K.

Der p-Wert entspricht damit dem kleinsten Signifikanzniveau α bei dem H_0 gerade noch verworfen wird obwohl H_0 wahr ist.

Für unser lineares Modell heißt das:

Der p-Wert ruft, wie wahrscheinlich es ist einen linearen Zusammenhang zwischen X und Y rein zufällig zu bestätigen, obwohl es ihm in Wirklichkeit gar nicht gibt.

⇒ Ein kleiner p-Wert spricht für die Signifikanz des linearen Zusammenhangs !!

OLS Regression Results									
Dep. Variable:	sales	R-squared:	0.643						
Model:	OLS	Adj. R-squared:	0.641						
Method:	Least Squares	F-statistic:	284.6						
Date:	Thu, 19 Oct 2023	Prob (F-statistic):	3.58e-37						
Time:	07:54:07	Log-Likelihood:	-408.80						
No. Observations:	160	AIC:	821.6						
Df Residuals:	158	BIC:	827.7						
Df Model:	1								
Covariance Type:	nonrobust								
	coef	std err	t	P> t	[0.025	0.975]			
const	7.0688	0.483	14.634	0.000	6.115	8.023			
TV	0.0489	0.003	16.871	0.000	0.043	0.055			
Omnibus:	1.575	Durbin-Watson:		1.931					
Prob(Omnibus):	0.455	Jarque-Bera (JB):		1.414					
Skew:	-0.230	Prob(JB):		0.493					
Kurtosis:	3.002	Cond. No.		325.					

Python

Bsp.: Output statsmodels OLS



Guteweise für den linearen Fit → dazu später mehr

Koeffizienten $\hat{\beta}_0, \hat{\beta}_1$, Konfidenzintervall ($\alpha=5\%$), Standardfehler, T-Teststatistik, p-Wert

→ noch viel mehr interessante Infos 😊
... da braucht man aber noch mehr Statistik ...