

... letztes Mal gesehen:

Modell

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

mit empirischer Beziehung

Rendition

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$$

(= Realisierung
der Störgröße ε)

bisher möglich

① Bewertung der Parameter -

Schätzung: Standardfehler, Konfidenzintervalle

geschätzte
Parameter

② Antwort auf die Frage: "besteht ein linearer Zusammenhang zwischen X und Y"

\Rightarrow Hypothesentest mit

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

hat man eine gute Schätzung und ein brauchbares Modell \Rightarrow

Es gilt

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$$

für $i = 1, \dots, n$

also

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Mit Hilfe des Standardfehlers der Schätzung $\hat{\sigma}$

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

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

empirische Entsprechung
 $MSE = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2 \approx \hat{\sigma}^2$
für große n

lässt sich die Modellgüte dann (global) bewerten!



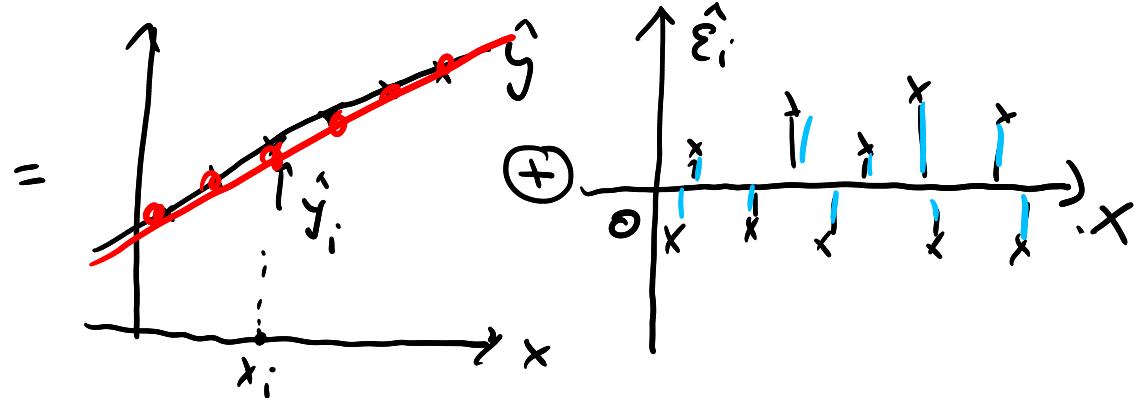
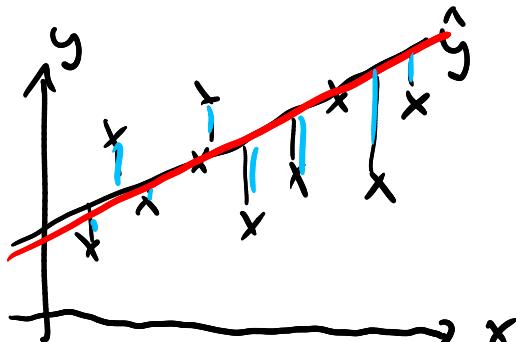
Welche Retriken gibt es noch?



Betrachte die Varianz der Zielgröße Y , das heißt
in der Realität die Struktur der y_i (als Schätzung)

Beobachtung:

$$y_i = \hat{y}_i + \hat{\varepsilon}_i$$



\Rightarrow Die Streuung der y_i setzt sich zusammen aus der Streuung der \hat{y}_i und der Streuung der "Prognosen" \hat{y}_i .

$\Leftrightarrow 0$ wegen $E(\varepsilon) = 0$!

also $\sum_{i=1}^n (y_i - \bar{y}_i)^2 = \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}}_i)^2 + \sum_{i=1}^n (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}_i)^2$

\uparrow

$= \bar{y}_i$ denn $E(f(x)) = Y$

\Rightarrow STREUUNGSZERLEGUNG

$$\boxed{\sum_{i=1}^n (y_i - \bar{y}_i)^2 = \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}}_i)^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$\hat{\varepsilon}_i^2$

\rightarrow SQT = SQE + SQR

SQT (total) = SQE (explained) + SQR (residual)

FRAGE Wann ist ein Modell gut?

... logischerweise dann, wenn der Anteil von SQE an SQT möglichst groß ist ☺

Def.: Bestimmtheitsmaß R^2 (R^2 -Statistik)

$$R^2 = \frac{\text{SQE}}{\text{SQT}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

↔

$$\sqrt{R^2} = R = g(y, \hat{y}) = \frac{\text{cov}(y, \hat{y})}{S_y \cdot S_{\hat{y}}}$$

PEARSON
KORRELATION

\uparrow
 $R^2 \in [0; 1]$ ist super, weil $\text{SQR} = 0$
 \uparrow also $\hat{e}_i = 0 \quad \forall i = 1 \dots n$

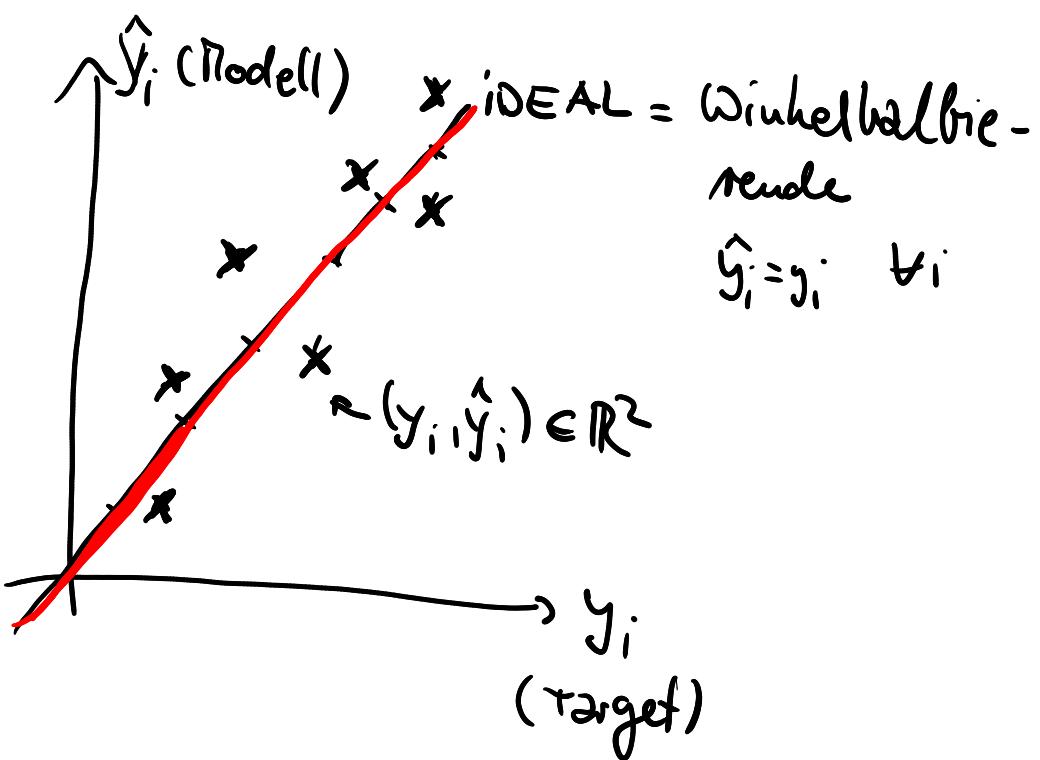
geht gar nicht (Modell $\hat{y}_i = y_i \quad \forall i = 1 \dots n$)

↑ Maß für die Stärke des linearen Zusammenhangs zwischen \hat{y} und y
 $g(x, y) \in [-1, 1]$ (idealerweise)

INTERPRETATION

$$R^2 = 1 \quad g(x_i, y_i) = 1$$

\Rightarrow je stärker die Prognosen \hat{y}_i von den echten y_i abweichen, desto mehr streuen die Punkte im Scatter plot um die Winkelhalbierende $\rightarrow R^2$ ist kleiner!



Bemerkung:

R^2 ist für alle Regressionsmodelle (auch nicht-lineare) geeignet! \Rightarrow keine Sache!

Es ist egal, wie die \hat{y}_i erzeugt würden. Es zählt nur die Stärke des linearen Zusammenhangs zwischen y_i und \hat{y}_i .

\Rightarrow Siehe Scatter Plot !!

! In der Anwendung gibt es Fälle (wenn auch selten), in denen die Software $R^2 < 0$ ausgibt !! ↴ ↴ STRANGE, ODER ?

Wie lässt sich das erklären ??

Bsp.: Sklearn 'LinearRegression' (und viele andere) geben den

R^2 -Score aus

$$R^2 = 1 - \frac{SQR}{SQT}$$

$$\overbrace{SQT = SQE + SQR} \text{ gilt nur, falls } E(f(x)) = Y !!$$

→ idealerweise wäre $SQR < SQT$, aber rein rechnerisch könnte auch was größeres rauskommen $SQR > SQT$ ↴

Gründe:

Das Modell ist so schlecht, dass die Prognose schlechter ist als $\hat{y}_i = \bar{Y}$ $\forall i$

Dann $E(f(x)) \neq Y$ und $SQE + SQR$ übersteigen $SQT \dots \text{YY}$



Falls man also so etwas beobachtet braucht man dringend ein besseres Modell !!

2.1.2) Multivariate lineare Regression

KLAR: In der Praxis hängt Y oft nicht nur von einer erklärenden Variable ab sondern von mehreren Größen X_1, \dots, X_p

⇒ **Modell**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip} + \varepsilon_i$$

und für $i=1, \dots, n$

Letztes Semester gesehen: günstige Darstellung

$$\vec{Y} = \vec{X} \cdot \vec{\beta} + \vec{\varepsilon}$$

Matrix-Form

wobei $\vec{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $\vec{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$ und $\vec{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$ ← Datenmatrix

Eine Schätzung für den Parametervektor $\hat{\beta}$ erhält man mit dem LQ - Verfahren

$$\hat{\beta} = (X^t \cdot X)^{-1} \cdot X^t \cdot \hat{y}$$

Voraussetzung
 $(X^t \cdot X)^{-1}$ existiert!
 $\Rightarrow X_i, X_j$ linear unabhängig für $i \neq j$

Für den Standardfehler der Schätzung ergibt sich:

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

$(\hat{\sigma}^2 \approx \text{MSE für große } n)$

Jetzt wird die Frage interessant, ob ein Modell mit allen Variablen X_1, \dots, X_p am besten abschneidet, oder ob eine Teilmenge der X_i ($i=1 \dots p$) gleich gut \ besser ist. \rightarrow Vergleich der $\hat{\sigma}^2$ wäre möglich (ist mit p skaliert)

Bemerkung: $R^2 = \frac{SQE}{SQT}$ ist nur bedingt geeignet, denn

für Modelle mit mehr Features bleibt zwar SQT gleich, aber
SQE wächst automatisch $\Rightarrow \underline{R^2 steigt}$

\Rightarrow Modelle mit vielen Input-Variablen haben automatisch ein
größeres R^2 als solche mit wenigen \Rightarrow Anzahl der X_j muss noch
mit berücksichtigt werden!

 Def: Angepasstes Bestimmtheitsmaß

Für X_1, \dots, X_p und $Y = f(X_1, \dots, X_p) + \varepsilon$ mit $i=1, \dots, n$ Realisierungen gilt:

$$\bar{R}^2 = 1 - (1 - R^2) \frac{(n-1)}{n - (p+1)}$$

\Leftarrow angepasstes Bestimmtheitsmaß!
(wird viel diskutiert ob Skalierung reicht...)

... damit lassen sich jetzt auch Modelle mit unterschiedlicher Anzahl von Einflußgrößen vergleichen.

⚠ Die Koeffizienten $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_P$ sind nicht die gleichen wie in

$$Y = \beta_0 + \beta_j X_j + \varepsilon$$

← das würde nur dann gelten, wenn es absolut keine Interaktionen zwischen den X_j ($j=1, \dots, P$) gäbe!

Das heißt, die X_j ($j=1, \dots, P$) müssten unabhängige ZV sein
(im Sinne der Statistik!)

Das gilt praktisch nie!!

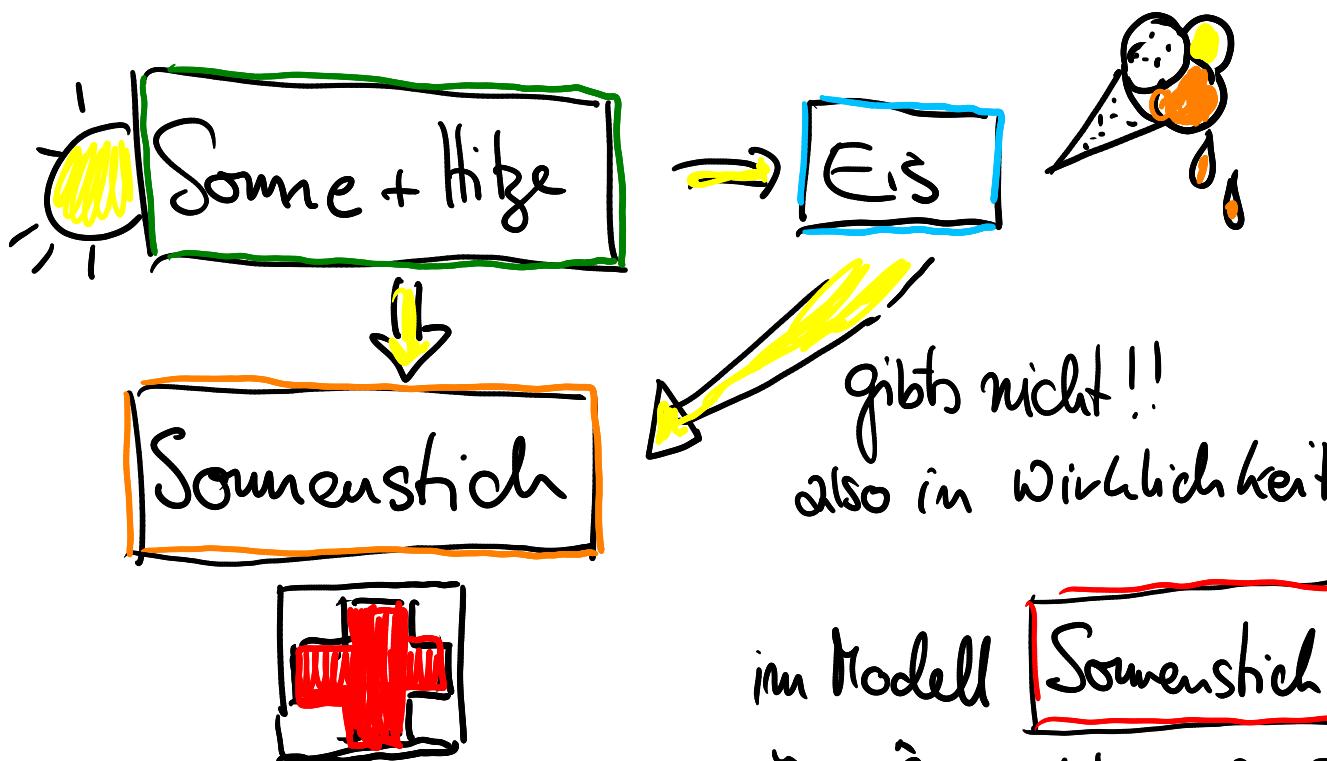
→ Stichwort: Scheinkorrelation!

Bsp.:

$Y \hat{=} \text{Anzahl Personen mit Sonnenstich}$
im Notaufnahme

$X \hat{=} \text{Verkauftes Eis am entsprechenden Tag}$

eigentlich:



im Modell

im Modell $\hat{\text{Sonnenstich}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Eis} + \hat{\beta}_2 \text{Hitze}$
wäre $\hat{\beta}_1$ wohl groß und signifikant

$\hat{\text{Sonnenstich}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Eis} + \hat{\beta}_2 \text{Hitze}$
wäre $\hat{\beta}_2$ kleiner, weniger signifikant
Hitze wäre "besser"



Wie entscheiden wir, ob bzw. welche der X_1, \dots, X_p gut fürs Modell sind?

Was bedeutet X_j relevant in

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_j X_j + \dots + \beta_p X_p + \epsilon$$

für X_1, \dots, X_p im üblichen Wertebereich lassen sich die β_j als Gewichtungsfaktoren interpretieren.

D.h.: $\beta_j \neq 0 \Rightarrow X_j$ relevant

⇒ Hypothesentest

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_1: \exists j \text{ mit } \beta_j \neq 0$$

← Teststatistik?

kritischer Bereich?

→ in der Literatur zu finden

Teststatistik: F-Statistik

$$F = \frac{\left(\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right) \cdot \frac{1}{p}}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 \cdot \frac{1}{m-(p+1)}}$$

↑ ganz schön kompliziert ⚡⚡

Was beschreibt F intuitiv?

↓
F vergleicht zwei Dinge

durch das Modell erklärte
Varianz von y (wie gut passt das Modell)
Varianz der Residuen (Nenner)

wird der Zähler sehr groß (top Modell) $\Rightarrow F$ groß \Rightarrow

dann kann $H_0: \beta_0 = \beta_1 = \dots = \beta_p$ wohl nicht stimmen $\Rightarrow H_0$ ablehnen

... das geht so auch für Teilmengen der X_1, \dots, X_p

alle
Vgr
↓

Dann

$$H_0: \beta_{p-q+1} = \dots = \beta_p = 0$$

H_1 : wenigstens eins davon $\neq 0$

$$SQE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad j=1, \dots, p$$

$$SQE_0 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad i=p-q+1, \dots, p$$

↑
nur die = 0

⇒

$$F = \frac{(SQE_0 - SQE) \cdot \frac{1}{q}}{SQE \cdot \frac{1}{n-(p+1)}}$$

⇒ Die F-Statistik und die zugehörigen p-Werte geben Auskunft über die Sinnhaftigkeit eines linearen Modells und bestimmen erklärende Größen X_j im Modell

Insgesamt: gute Idee F-Statistik & p-Werte zu berechnen.

Problem: man braucht vorher eine Idee was grundsätzlich sinnvoll sein könnte, denn

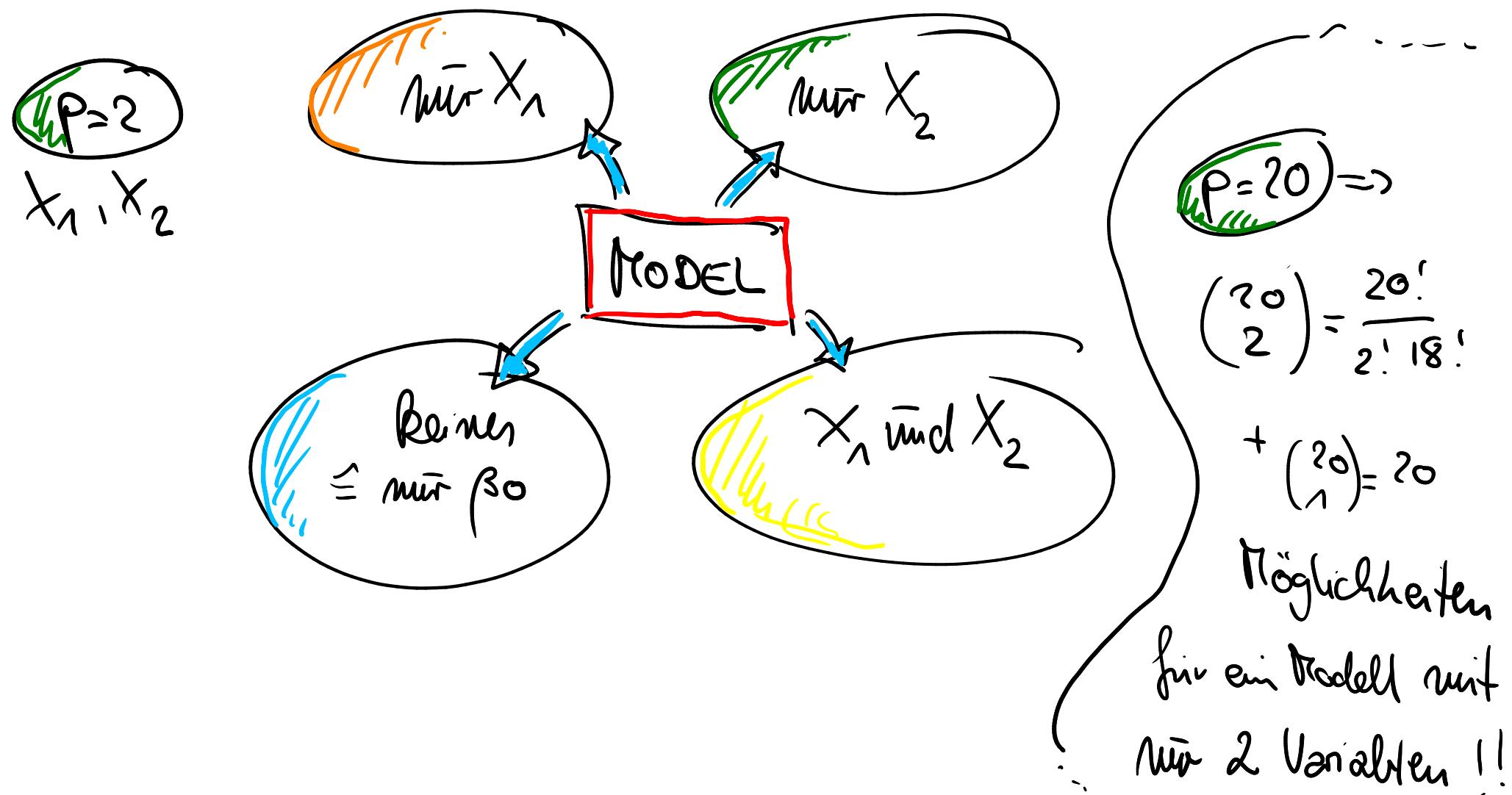
→ insbesondere bei sehr hochdimensionalen Feature-Räumen kann man sich nicht zu 100% auf das Ergebnis verlassen.



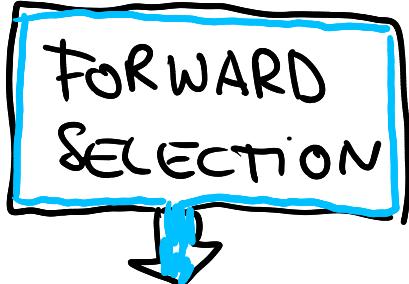
weitere Möglichkeiten der Variablen Selektion

"Variable Selection"

PROBLEM: Alle Kombis von X_1, \dots, X_p einfach ausprobieren
geht wahrscheinlich nicht.
Besonders dann nicht, wenn p groß ist ↴ ↴



3 gebräuchliche (machbare) Strategien:

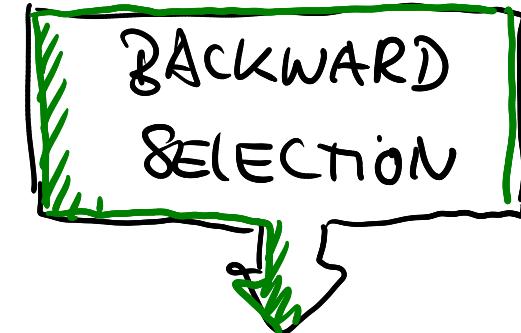


- starte mit $n=\text{const.}$
- filte p Einflachregressionen
- wähle X_j mit dem besten Ergebnis
- filte $(p-1)$ Modelle mit einem zus. X_j
- wähle bestes
- wiederhole Prozess bis Performance-kriterium erreicht ist

auch für $p > n$



- starte mit Forward-Selection
- steigt ab einem best. Punkt der p-Wert einer Größe zu sehr an
⇒ entfernen
- führe \leftarrow Schritte so lange aus bis alle X_j im Modell kleine und alle anderen große p-Werte haben



- starte mit allen X_j im Modell
- streiche X_j mit dem größten p-Wert
- filte neues Modell mit $(p-1)$ Größen
- wiederhole so lange bis alle p-Werte unter Threshold (z.B. 0,0001)

kür für $p < n$ möglich