Regularisierung von linearer Regression

Phillip Grafendorfer, Michael Kastner, Raphael Peer

Daten

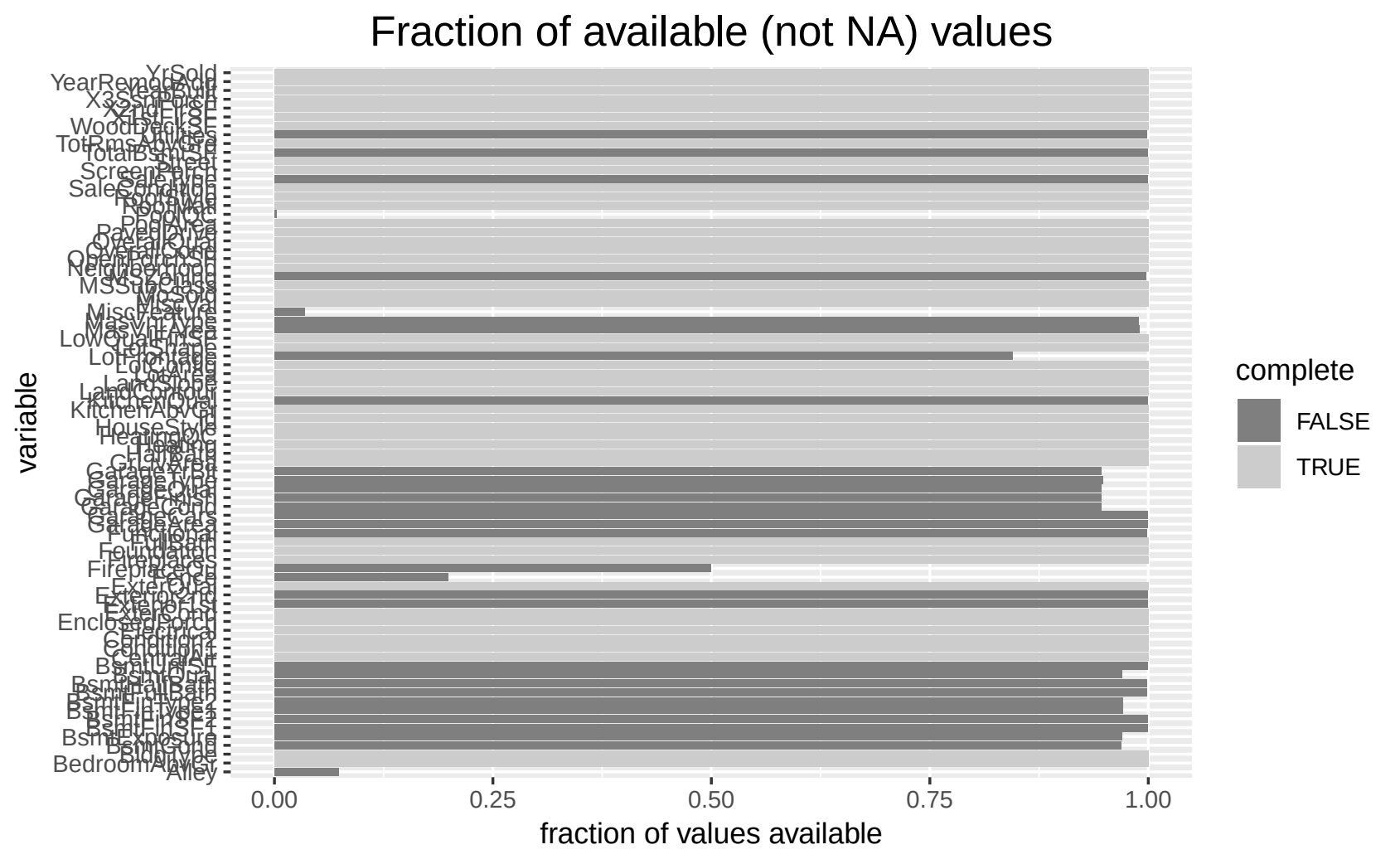
Ames House price Dataset

Datensatz:

- 1460 Häuser
- 79 erklärende Variablen (numerisch und kategorisch)
- bekannter Übungsdatensatz

Quelle: `https://www.kaggle.com/c/house-prices-advanced-regression-techniques`

Fehlende Werte: Übersicht



Fehlende Werte: Strategie

Umgang mit fehlenden Werten:

- Bei mehr als 10% Fehlenden Werten: Variable verworfen
- Bei numerischen Variablen: NA durch Median der Variable ersetzt
- Bei kategorischen Variablen: NA als eigene Kategorie (Kategorie 'unbekannt')

Problem mit validation-set: seltene Factor-levels

data-frame

einige levels im validation-set aber nicht im trainings-set

$\Longrightarrow$ unbekannte dummy Variablen im validation-set

$\Longrightarrow$ error

design-matrix

einige levels in validation-set aber nicht im trainings-set

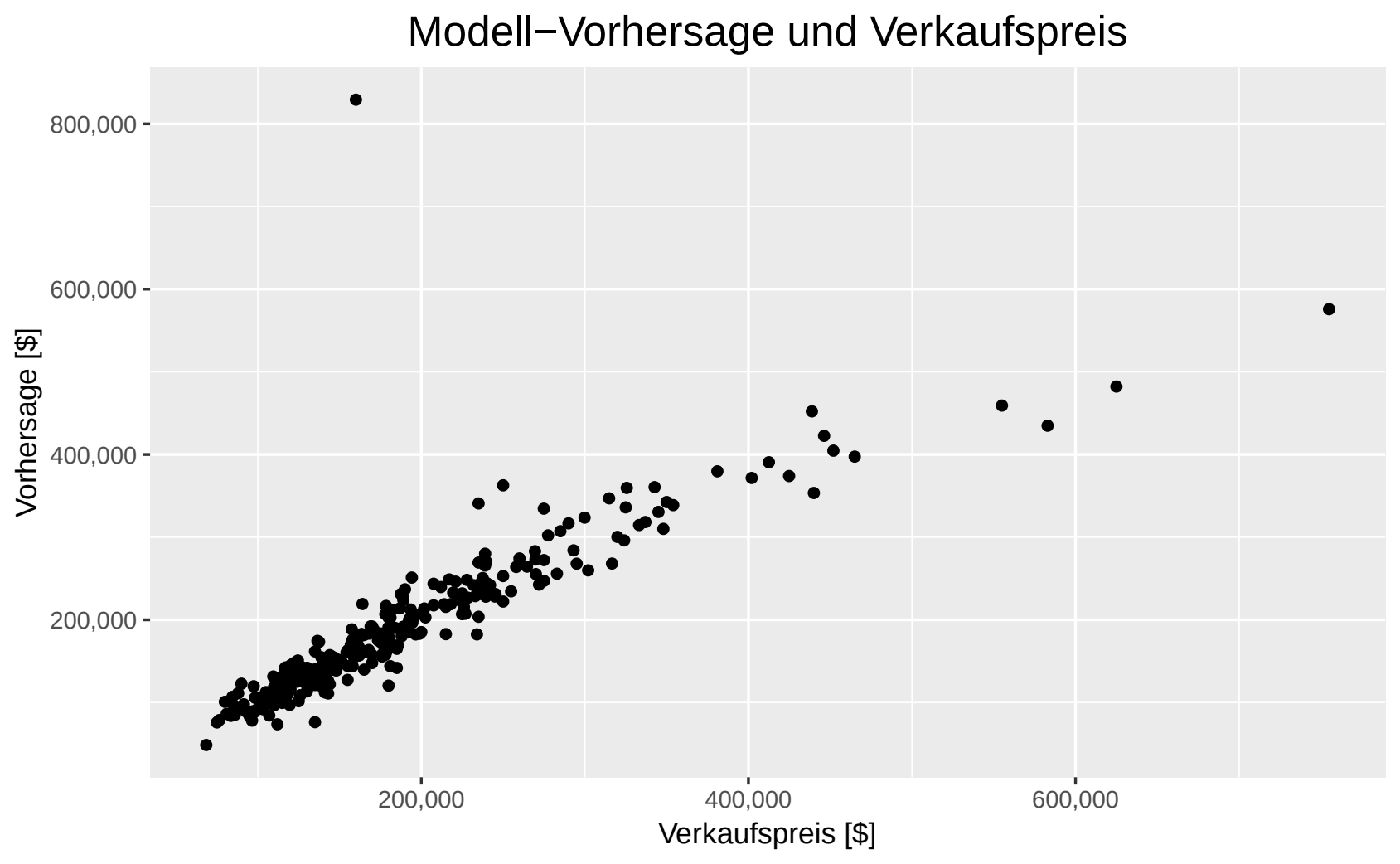
$\Longrightarrow$ dummy variable immer null im trainings set

$\Longrightarrow$ Koeffizeint ≈ 0

$\Longrightarrow$ kein Einfluss

Standard lineare Regression

Einfaches Model mit allen Variablen



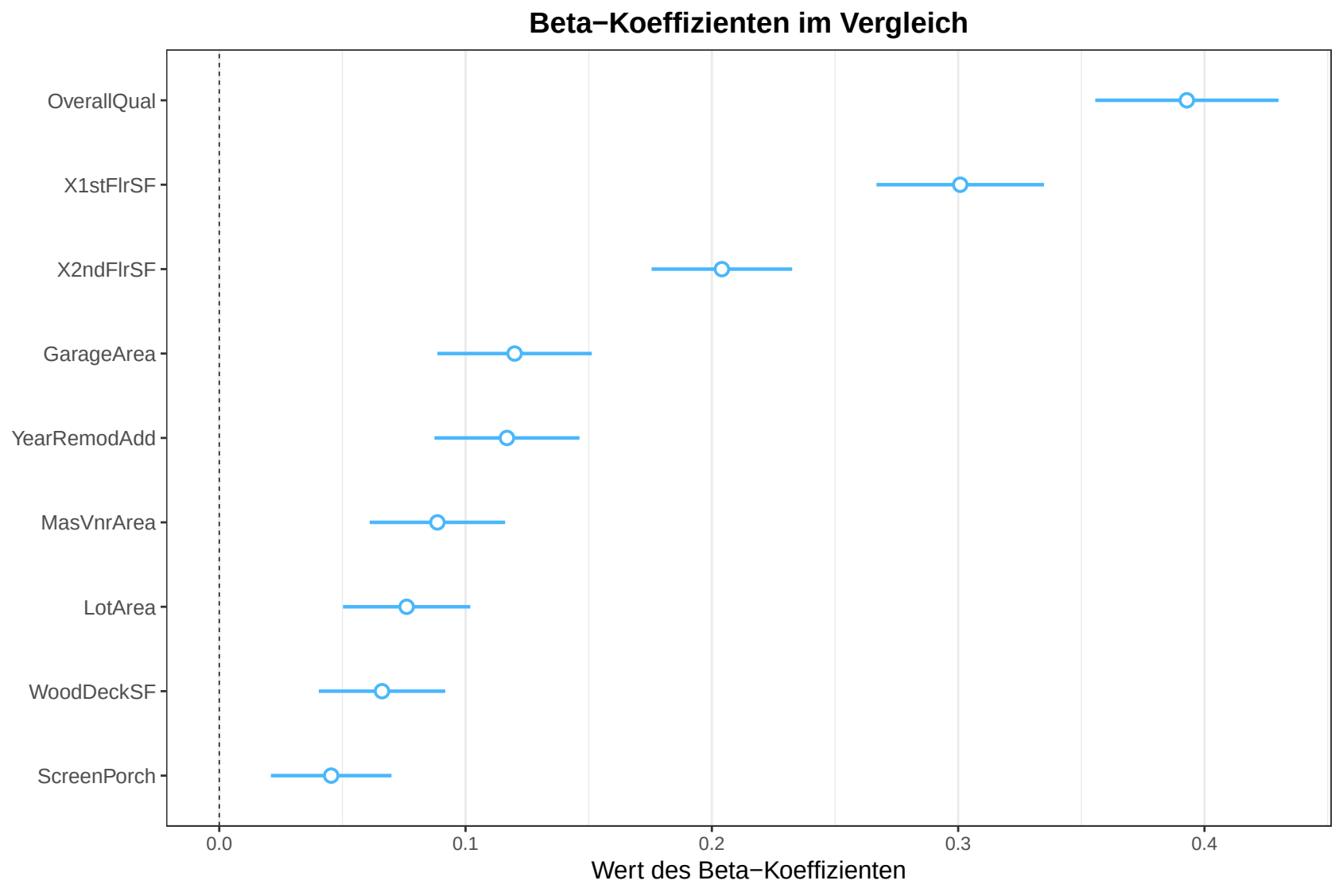
Interpretierbare Koeffizienten

Nachteil unstandardisierter Regressionskoeffizienten

- Von den Maßeinheiten für X und Y abhängig
- Daher schlechtere Vergleichbarkeit

Lösung: Standardisierte Koeffizienten

Beta-Koeffizienten im Vergleich



Regularisierung

Problemstellung I

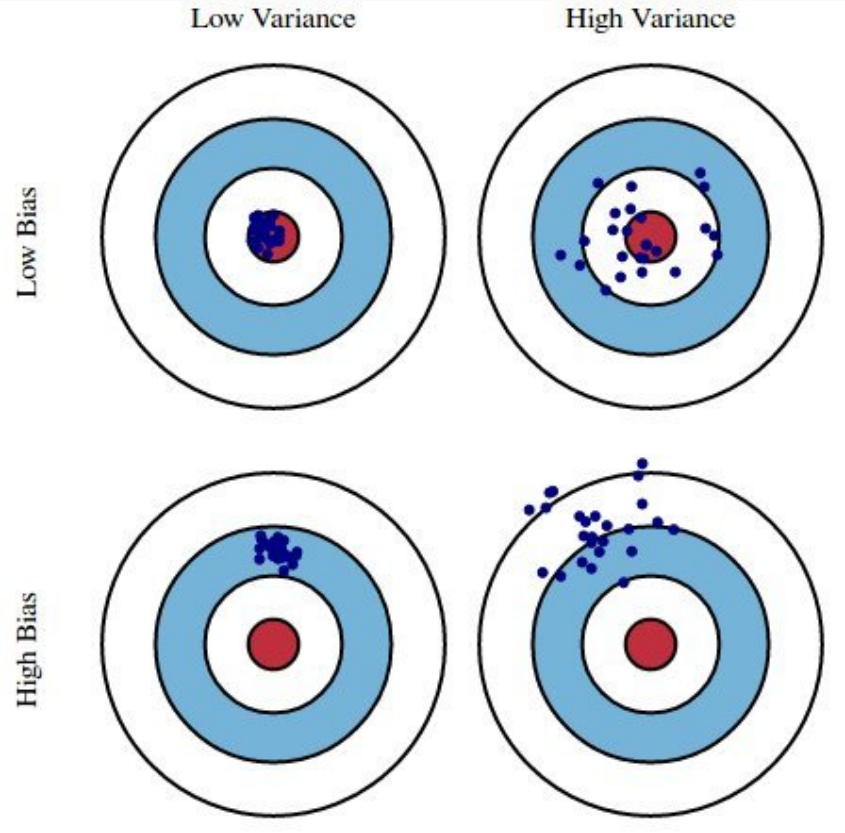


Figure 1: Quelle: kdnuggets.com

- Bias- Variance Tradeoff
- OLS Schätzer ist "unbiased" aber kann große Varianz haben

Problemstellung II

Wann tritt große Varianz auf?

- Wenn die Prediktoren hohe Korrelation aufweisen
- Bei vielen Prediktoren. Wenn die Anzahl Prediktoren nahe bei Anzahl der Beobachtungen geht die Varianz gegen unendlich.

Lösung I

Verringerung der Varianz auf Kosten des Bias

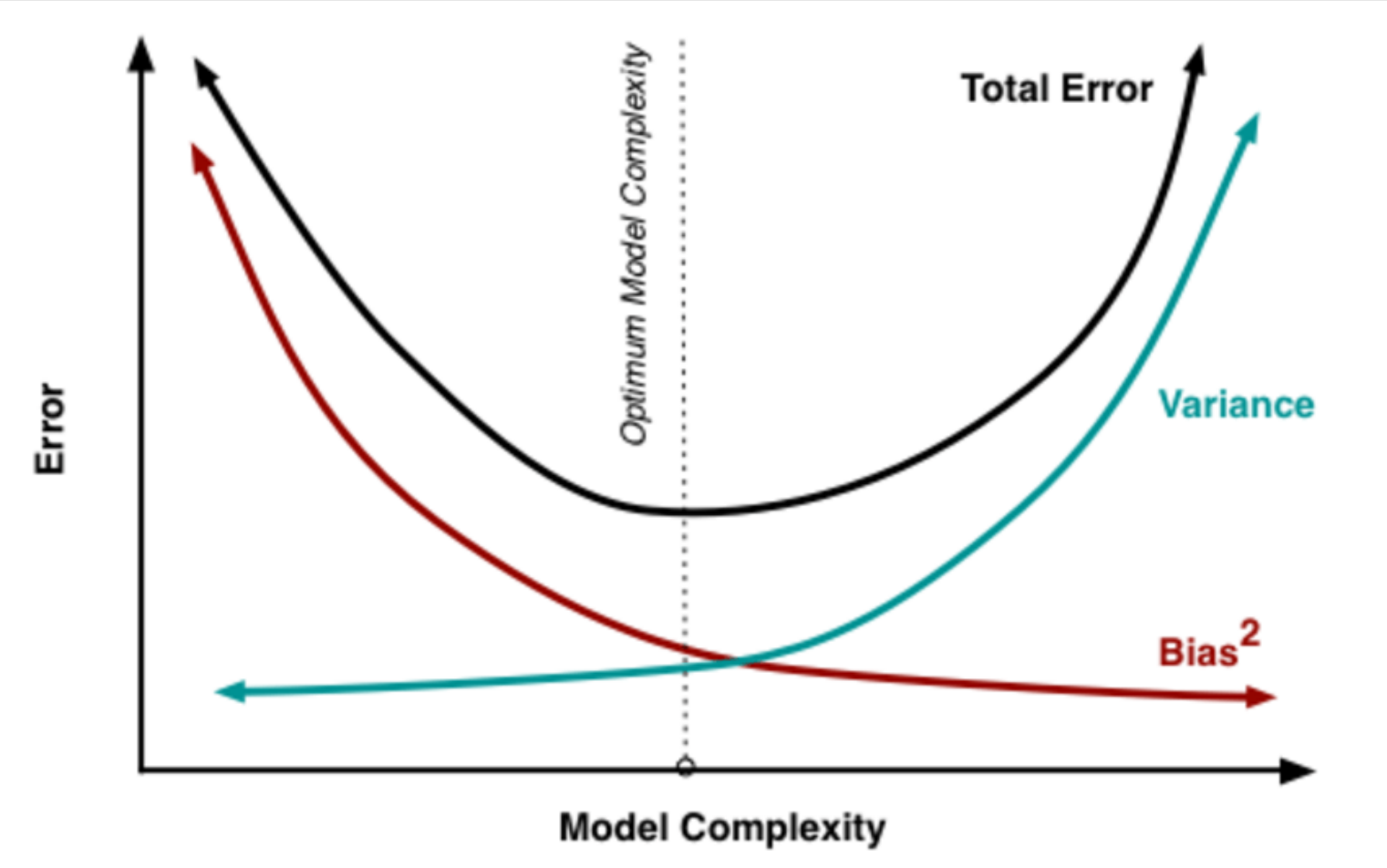


Figure 2: Quelle: researchgate.net

Lösung II

$$L_{ridge}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{m} \hat{\beta}_j^2 = ||y - X\hat{\beta}||^2 + \lambda||\hat{\beta}||^2$$

Die Diskussion dieser Likelihood Funktion liefert für jeden Parameter λ ein set von Schätzern $\hat{\beta}$. Falls $\lambda \Rightarrow 0$, dann $\hat{\beta}_{ridge} \Rightarrow \hat{\beta}_{OLS}$ Frage: wie wird der Regularisierungs- Parameter gewählt?

- Crossvalidierung (hier benutzt)
- Minimierung eines weiteren Informationskriteriums (AIC, BCI etc.)

Crossvalidierung

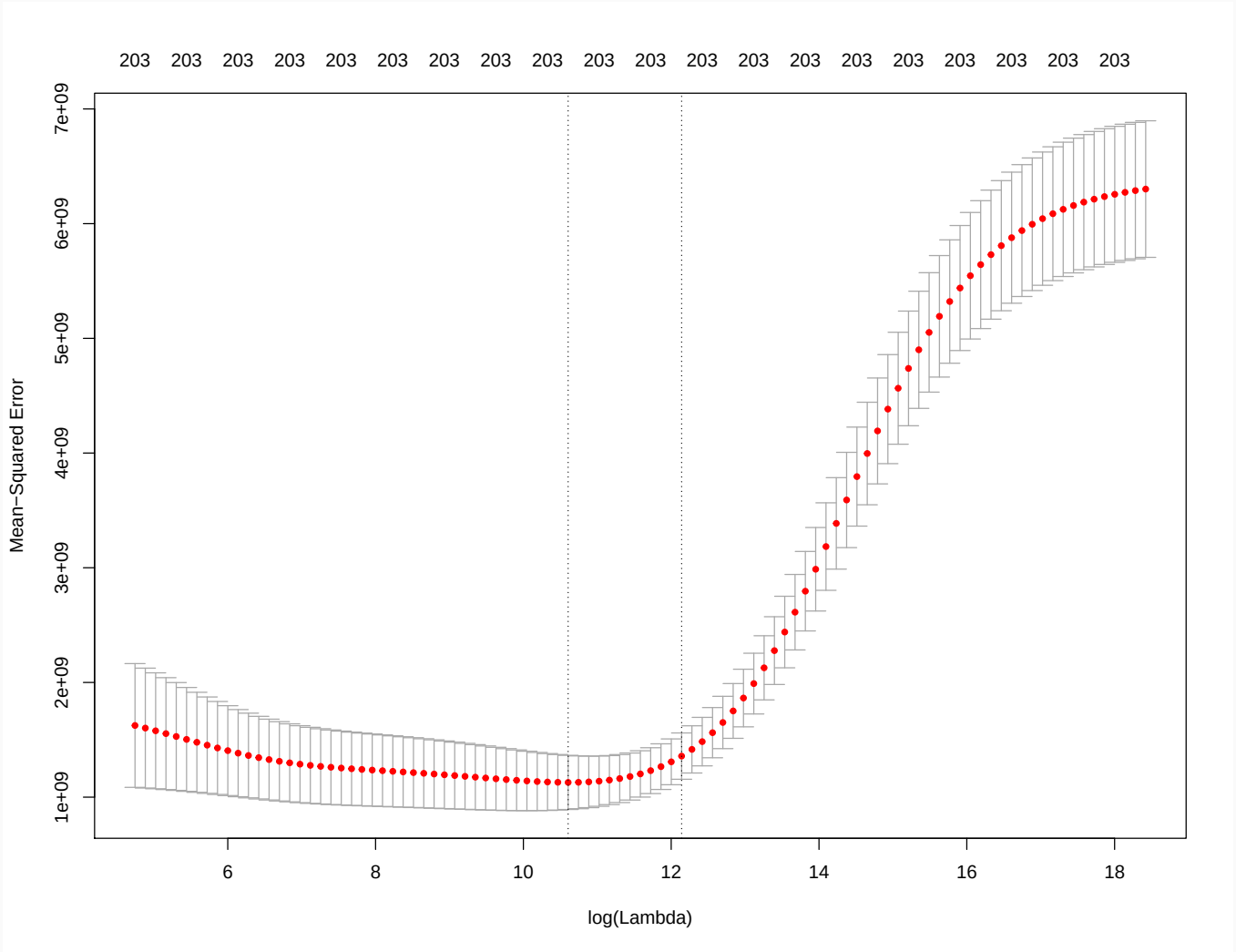


Figure 3: Lambda Tuning

Umsetzung

Abhängige und unabhängige Variablen werden standardisiert

- Mittelwert $= 0$
- Varianz $= 1$

Regressionskoeffizienten:

- $\hat{\beta}_i = \beta_i * \frac{s_{x_i}}{s_y}$
- $\hat{\beta}_i$ sollte im Intervall [-1, 1] liegen (sonst Hinweis auf Multikollinearität)

Vor- und Nachteile

Vorteile

- Operiert mit Änderungen von Standardabweichungen
- $\Rightarrow$ Stärke und Richtung eines Effektes können besser interpretiert und verglichen werden

Nachteile

- Nur für Variablen anwendbar, bei denen Heranziehen einer Standardabweichung sinnvoll ist (zB nicht Dummyvariablen)
- Abhängigkeit von Stichprobe
- Kann zu Missverständnissen führen

Vergleich der Modelle

Modelle	R^2	MAD
Top5	0.762	25410
Top9	0.779	24407
Naives Modell	0.933	20117
RReg (Vset)	0.903	19624
RReg (Ges)	0.896	
RReg (Spez)	0.952	23251

Fragen und Diskussion

Vielen Dank für
Ihre Aufmerksamkeit!