

Vorhersage und Interpretation des Verkaufpreises von Immobilien mittels linearer Regression

Karin Lassnig, Phillip Grafendorfer, Martin Huf, Raphael Peer

Daten

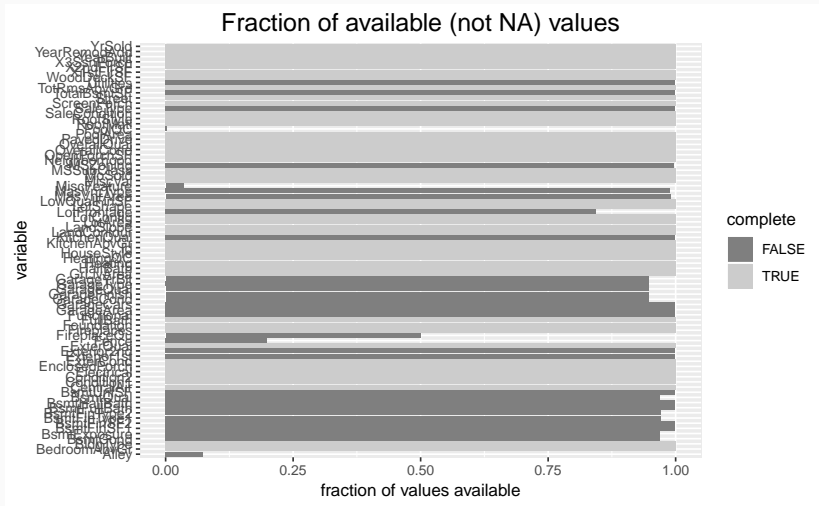
Ames House price Dataset

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



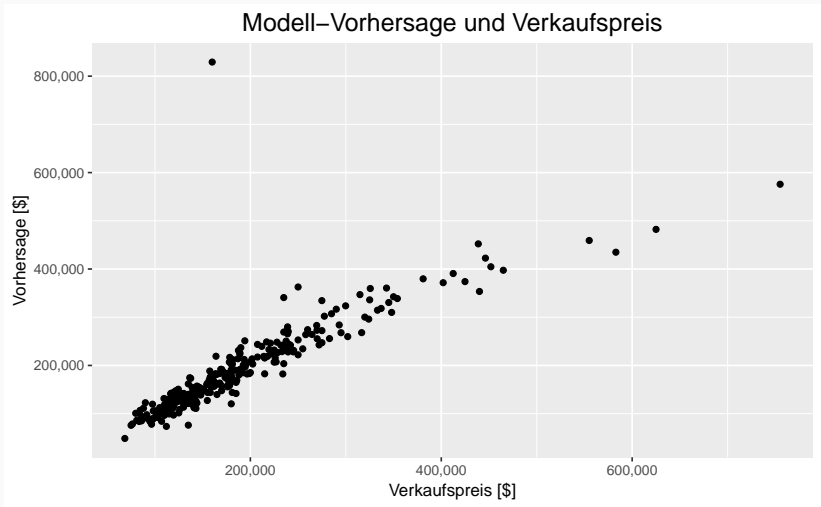
Fehlende Werte: Strategie

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')

Modelle

Einfaches Model mit allen Variablen



Standardisierte Koeffizienten

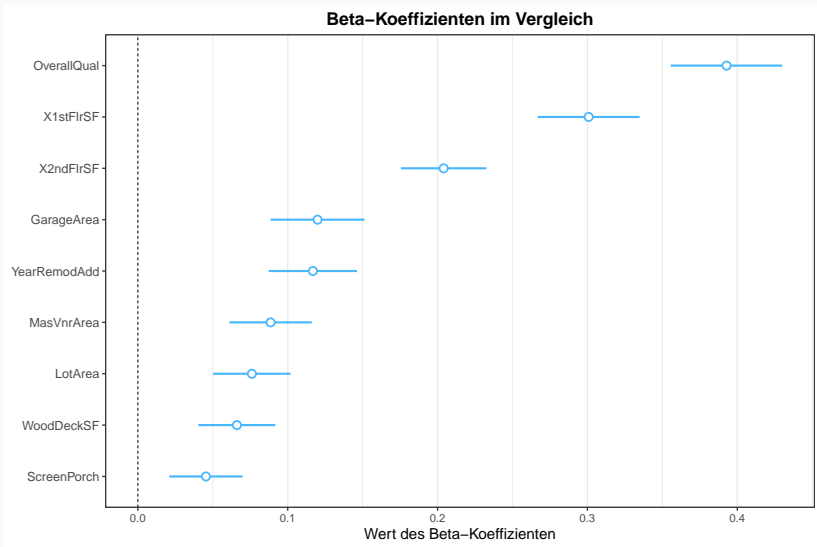
Problemstellung

Nachteil unstandardisierter Regressionskoeffizienten

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

Lösung: Standardisierte Koeffizienten

Beta-Koeffizienten im Vergleich



Regularisierung

Problemstellung I

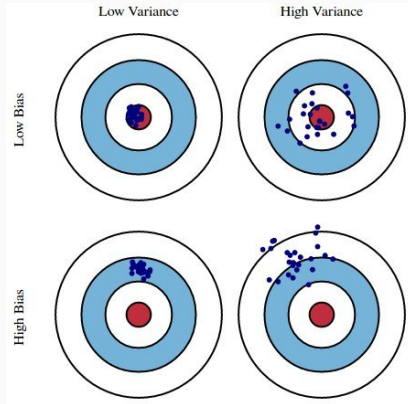


Figure 1: Quelle: kdnuggets.com

- 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

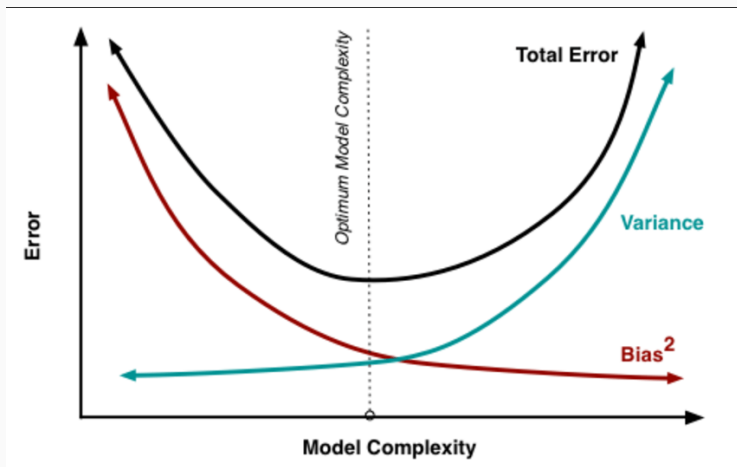


Figure 2: Quelle: researchgate.net

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, BIC etc.)

Crossvalidierung

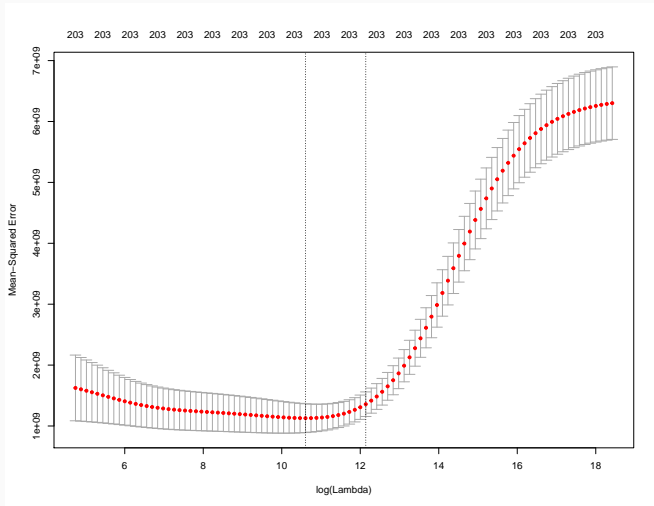


Figure 3: Lambda Tuning

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

Anwendung

myfile

Call:

```
lm(formula = scale(SalePrice) ~ 0 + scale(LotArea) + scale(YearRemodAdd) +  
    scale(MasVnrArea) + scale(X1stFlrSF) + scale(X2ndFlrSF) +  
    scale(GarageArea) + scale(WoodDeckSF) + scale(ScreenPorch) +  
    scale(OverallQual), data = train_processed)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.8579	-0.2341	-0.0194	0.1944	3.9476

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
scale(LotArea)	0.07608	0.01317	5.775	9.40e-09	***
scale(YearRemodAdd)	0.11680	0.01501	7.781	1.36e-14	***
scale(MasVnrArea)	0.08857	0.01401	6.322	3.44e-10	***
scale(X1stFlrSF)	0.30084	0.01734	17.352	< 2e-16	***
scale(X2ndFlrSF)	0.20406	0.01455	14.025	< 2e-16	***
scale(GarageArea)	0.11991	0.01598	7.504	1.07e-13	***
scale(WoodDeckSF)	0.06608	0.01308	5.051	4.96e-07	***
scale(ScreenPorch)	0.04543	0.01248	3.640	0.000282	***
scale(OverallQual)	0.39285	0.01896	20.718	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4697 on 1451 degrees of freedom

Multiple R-squared: 0.7806, Adjusted R-squared: 0.7792

F-statistic: 573.5 on 9 and 1451 DF, p-value: < 2.2e-16

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

Vielen Dank für
Ihre Aufmerksamkeit!