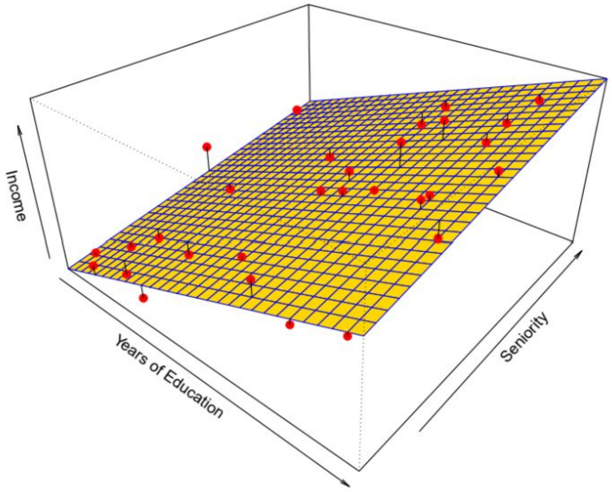
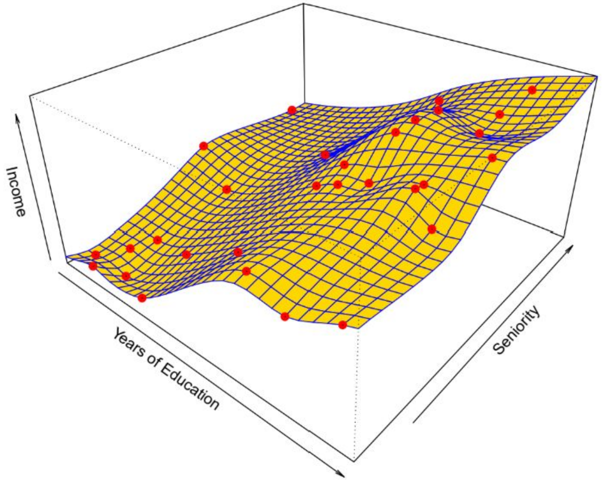
The Importance of Model Specification in Statistical Conclusions

Sophia Mao
Mentor: Abhijeet Mulgund

Why do different statistical models applied to the same data lead to different conclusions?





True model: $Y = f(X) + \epsilon$

What we learn: $\hat{Y} = \hat{f}(X)$

Different models → different $\hat{f}(X)$

Linear regression

d value for
n of Y

slope of X term
representing the
change in Y for a
unit change in X

error term
representing
$y_i - \bar{y}$

$$\hat{Y} = \beta_0 + \beta_1 x_i + \varepsilon_i$$

constant term
where y = 0

observed value for
the ith term of X

Logistic regression

$$\hat{Y} = \frac{1}{1+e^{-\sum_{j=0}^{M} \beta_j x_{ij}}}$$

j = 1...M (number of ind. variables)

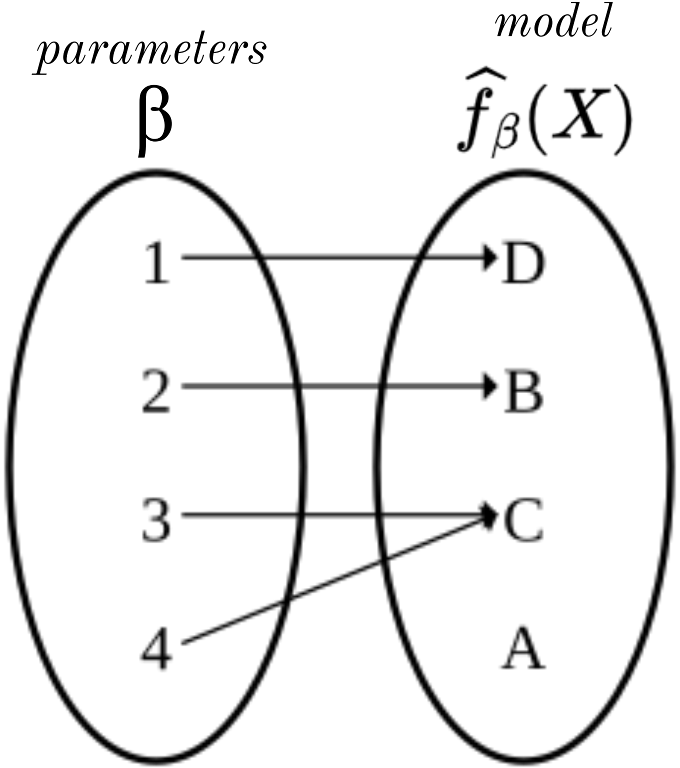
the j'th variable at
observation *i*

natural log

regression coefficients

Final result depends on...

1. the model class (what is allowed), and
2. the fitting process (what is chosen within it)


parameters
β
model
$\widehat{f}_\beta(X)$
1
2
3
4
D
B
C
A

We commonly use models for...

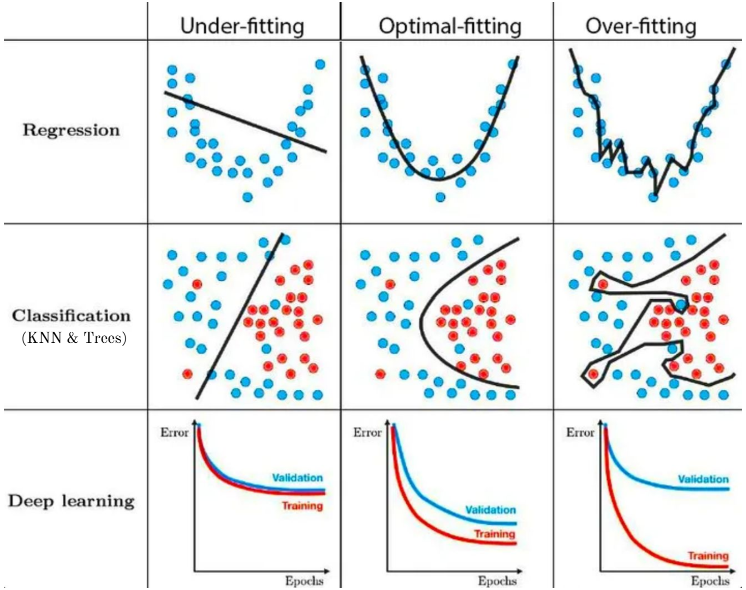
1. Prediction

Will a patient be diagnosed with diabetes?
What is the risk of hospital readmission?

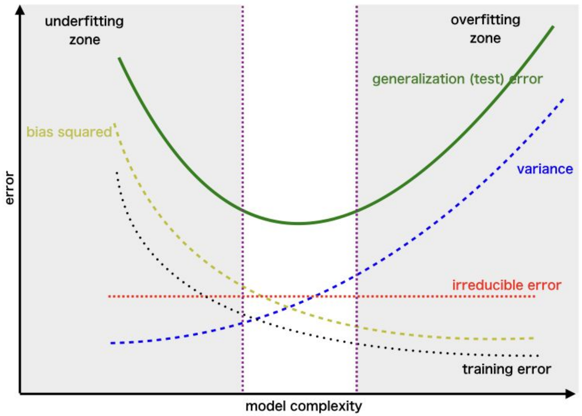
2. Inference

Does obesity increase disease risk?
Does housing status affect health outcomes?

1. More rigid model → underfits → misses real patterns (high bias)
2. More flexible model → overfits → fits noise (high variance)



Bias-Variance Tradeoff



$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

Regularization

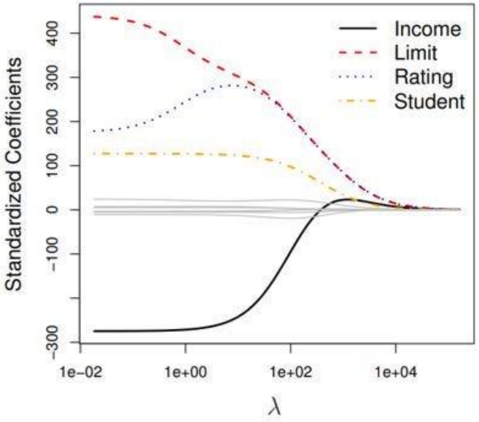
Ridge regression

- Adds a penalty that keeps coefficients ssmall

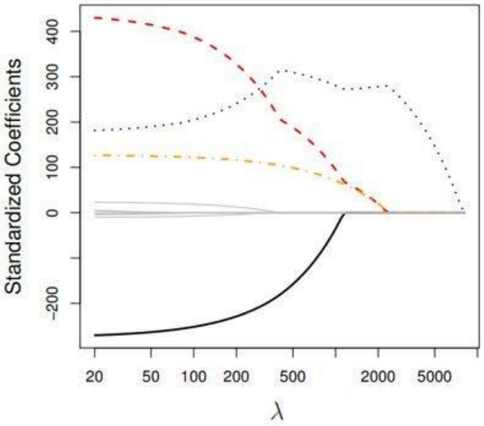
Lasso regression

- Makes some coefficients exactly zero

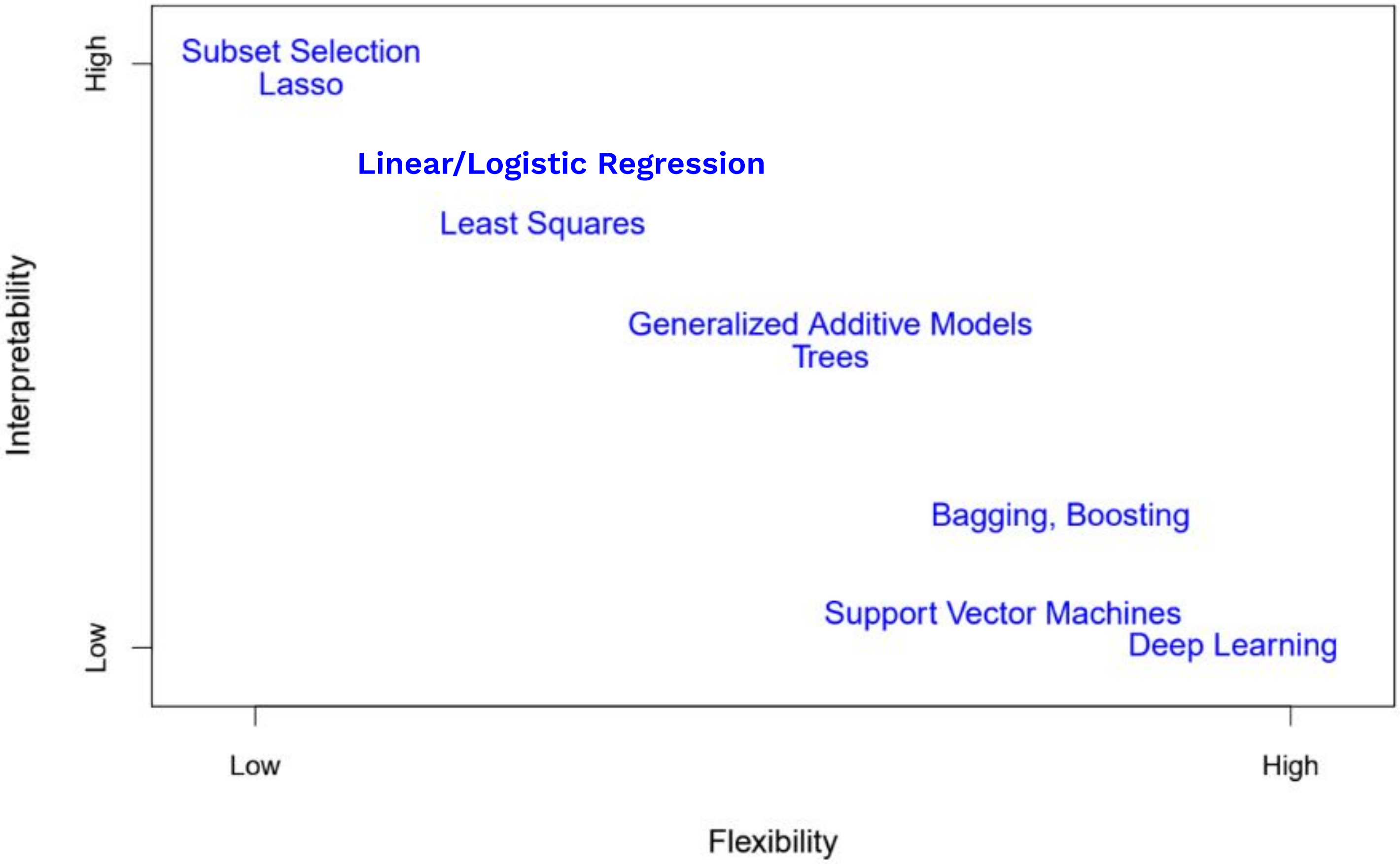
Both control variances at the cost of introducing some bias



Ridge Regression



Lasso Regression


Subset Selection
Lasso
Linear/Logistic Regression
Least Squares
Generalized Additive Models
Trees
Bagging, Boosting
Support Vector Machines
Deep Learning
Interpretability
High
Low
Flexibility
Low
High

```
model = smf.logit("HCV_active ~ C(Homeless) + Recept_sharing + Public_prop + Partner_HCV_prop",
    data=df).fit()
```

	coef	std err	z	P>|z|
Intercept	-6.7835	0.886	-7.660	0.000
C(Homeless)[T.1]	1.1375	1.147	0.992	0.321
Recept_sharing	0.2621	0.037	7.112	0.000
Public_prop	2.8854	2.346	1.230	0.219
Partner_HCV_prop	11.2222	1.573	7.134	0.000

```
model2 = smf.logit("HCV_active ~ C(Homeless) + Recept_sharing + Drug_out + Partners",
    data=df).fit()
```

	coef	std err	z	P>|z|
Intercept	-4.1533	0.401	-10.366	0.000
C(Homeless)[T.1]	1.4640	0.305	4.801	0.000
Recept_sharing	0.0768	0.027	2.850	0.004
Drug_out	0.0821	0.032	2.561	0.010
Partners	0.9752	0.162	6.033	0.000

> Science. 2019 Oct 25;366(6464):447-453. doi: 10.1126/science.aax2342.

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer [1] [2], Brian Powers [3], Christine Vogeli [4], Sendhil Mullainathan [5]

Affiliations + expand

PMID: 31649194 DOI: 10.1126/science.aax2342

“Essentially, all models are wrong, but some are useful”

— George E. P. Box