Predicting Number of Publications by Biochemistry Ph.D. Students (1950-1970)

Shuchi Goyal, Suoyi Yang, Heather Zhou

STATS 202B
Instructor: Dr. Chad Hazlett
University of California, Los Angeles

October 10, 2019

Overview

Introduction

- **915 observations** of biochemistry Ph.D. students (1950-1970, data from Long (1990))
  - Training: 615 observations
  - Testing: 300 observations

Introduction

- **915 observations** of biochemistry Ph.D. students (1950-1970, data from Long (1990))
  - Training: 615 observations
  - Testing: 300 observations
- **Count outcome**: number of articles produced in last 3 years of Ph.D.

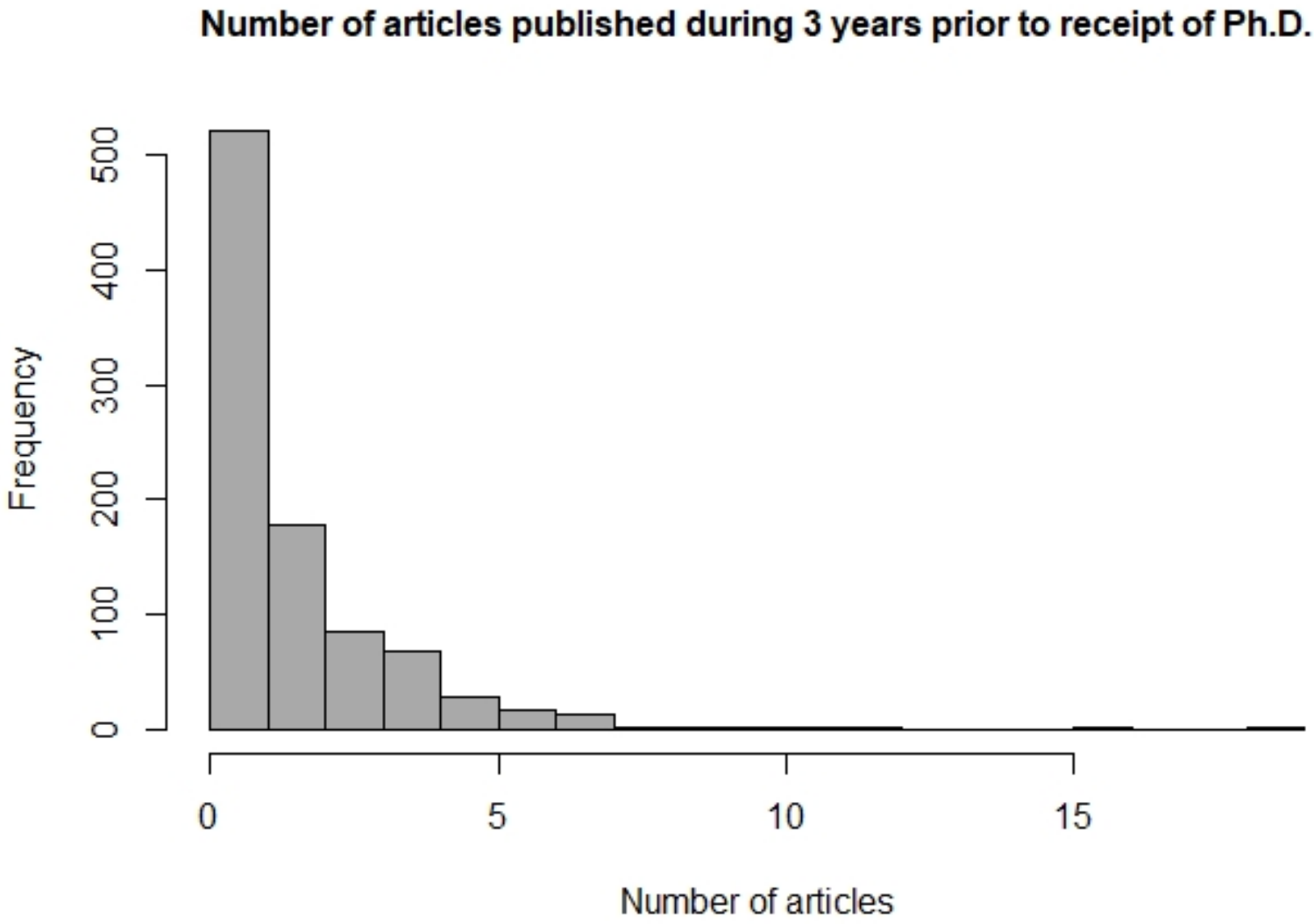
Introduction

- **915 observations** of biochemistry Ph.D. students (1950-1970, data from Long (1990))
  - Training: 615 observations
  - Testing: 300 observations
- **Count outcome**: number of articles produced in last 3 years of Ph.D.
- **5 covariates**:
  - Gender (binary)
  - Marital status (binary)
  - Number of children aged 5 or younger (count)
  - Prestige of department (ranging between 0.755 and 4.62)
  - Productivity of mentor (number of articles produced by the Ph.D.s mentor during the same 3 years)

Introduction

- **915 observations** of biochemistry Ph.D. students (1950-1970, data from Long (1990))
  - Training: 615 observations
  - Testing: 300 observations
- **Count outcome**: number of articles produced in last 3 years of Ph.D.
- **5 covariates**:
  - Gender (binary)
  - Marital status (binary)
  - Number of children aged 5 or younger (count)
  - Prestige of department (ranging between 0.755 and 4.62)
  - Productivity of mentor (number of articles produced by the Ph.D.s mentor during the same 3 years)
- **Goal**:
  - Build and select models
  - Interpret the effect of the covariates
  - Compare the models using test data

Histogram


Number of articles published during 3 years prior to receipt of Ph.D.
Frequency
0
100
200
300
400
500
0
5
10
15
Number of articles

GLM Models

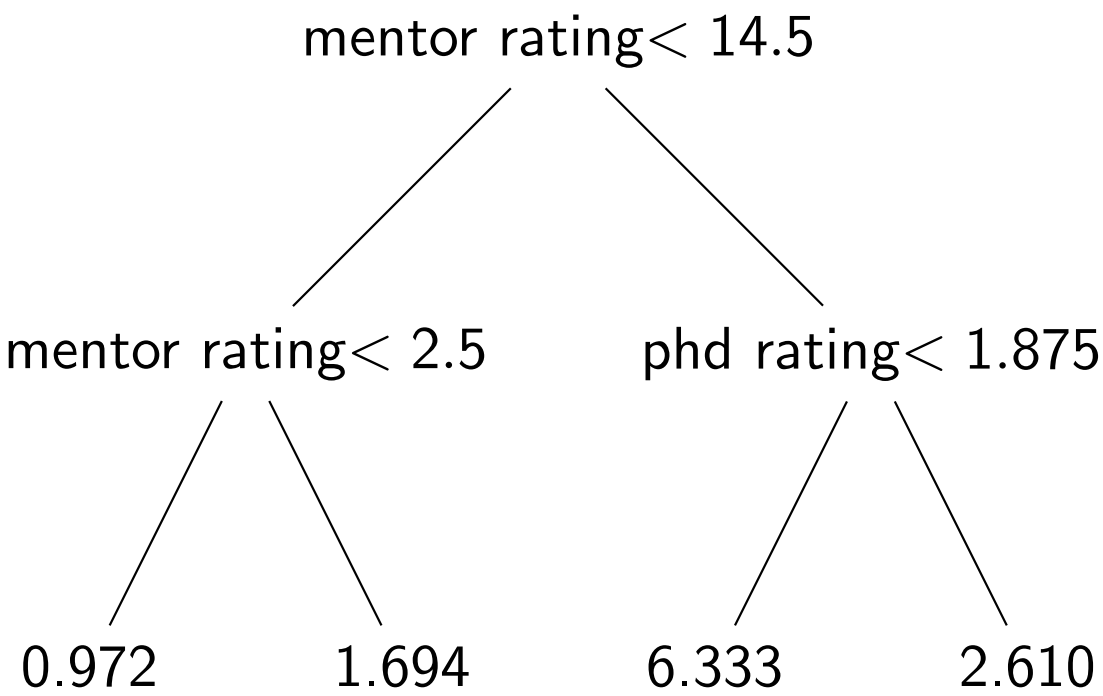
Motivation		**Baseline** Poisson	**Overdispersion** NegBin	**Excess 0's** Hurdle Poisson	Hurdle NB	ZI Poisson	ZI NB
Coef. Estimate	int	0.114	0.049	0.497 **	0.139	0.354 *	0.049
	female	-0.219 **	-0.208 *	-0.252 **	-0.264 *	-0.237 **	-0.208
	married	0.249 **	0.243 *	0.172 *	0.189	0.218 **	0.243
	numKid$\leq$ 5	-0.250 ***	-0.240 ***	-0.250 ***	-0.272 **	0.251 ***	-0.240 ***
	phdRating	0.050	0.058	0.026	0.049	0.042	0.058
	mentRating	0.025 ***	0.029 ***	0.020	0.025 ***	0.022 ***	0.029 ***
Model Eval.	numParameters	6	7	7	8	7	8
	Loglikeli	-1080.441	-1025.808	-1077.468	-1043.847	-1063.258	-1025.808
	RMSE (Training)	2.184	2.182	1.821	1.820	1.801	1.827

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

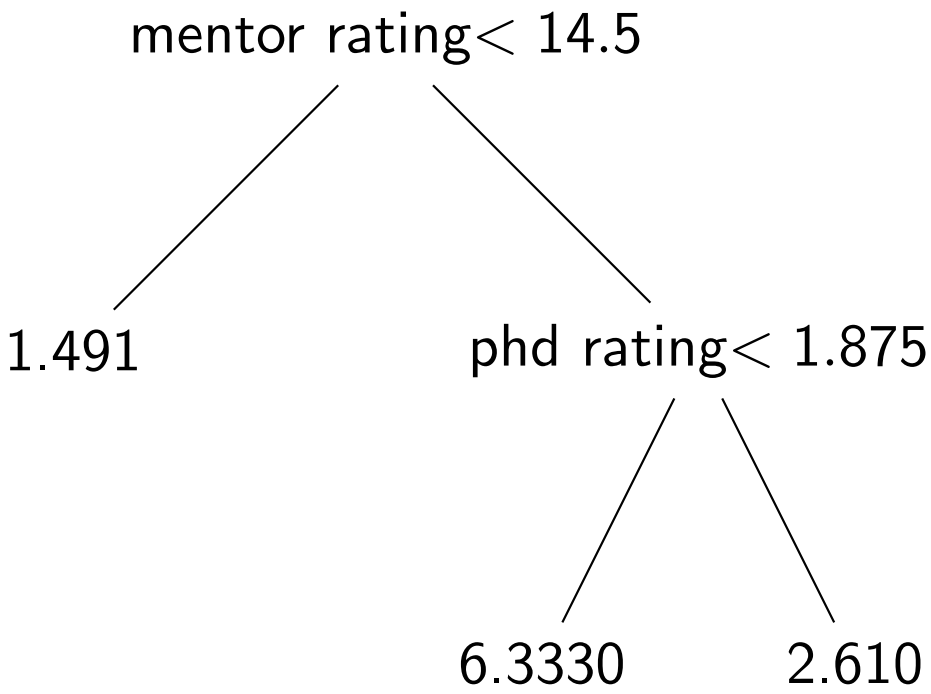
Regularization and Kernel Models

Motivation				**Regularization**		**Kernelization**
		NegBin	Hurdle NB	Ridge	Lasso	KRLS
Coef. Estimate	int	0.049	0.139	1.077	1.099	na
	female	-0.208	-0.264	0.333	-0.340	-0.278
	married	0.243	0.189	0.358	0.372	0.181
	numKid$\leq$ 5	-0.240	-0.272	-0.321	-0.341	-0.149
	phdRating	0.058	0.049	0.051	0.034	0.059
	mentRating	0.029	0.025	0.054	0.058	0.045
Model Eval.	numParameters	7	8	6	6	5
	RMSE (Training)	2.182	1.820	1.787	1.787	2.376

CART - full tree



CART - pruned tree (n=3)



Recap and (Training) RMSEs

- **GLM**
  - Poisson, Negative Binomial
  - Hurdle Poisson, Hurdle NB
  - Zero-Inflated Poisson, Zero-Inflated NB
- **Regularization/KRLS**: Ridge, Lasso, KRLS
- **Tree-based methods**: CART, Random Forest, XG Boost

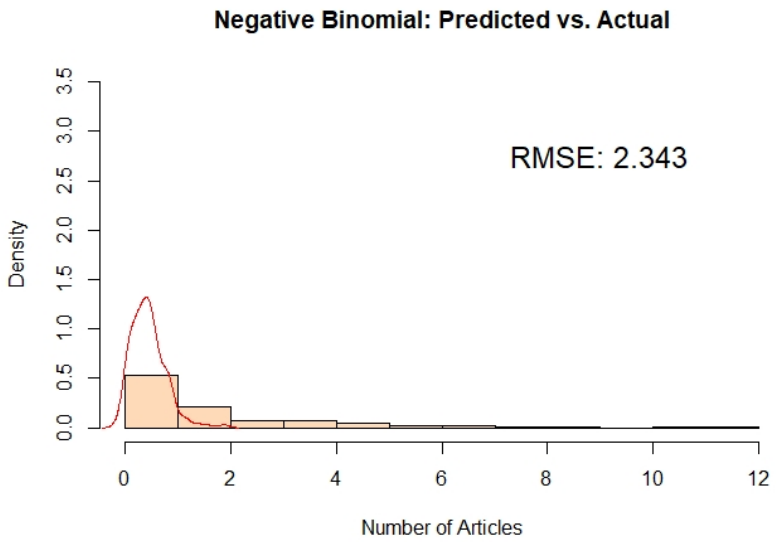
Method	**GLM**		**Regularization/KRLS**			**Tree-based Methods**		
	NegBin	Hurdle NB	Ridge	Lasso	KRLS	CART (n=3)	Random Forest	XG Boost
RMSE	2.182	1.820	1.787	1.787	2.376	1.796	1.829	1.562

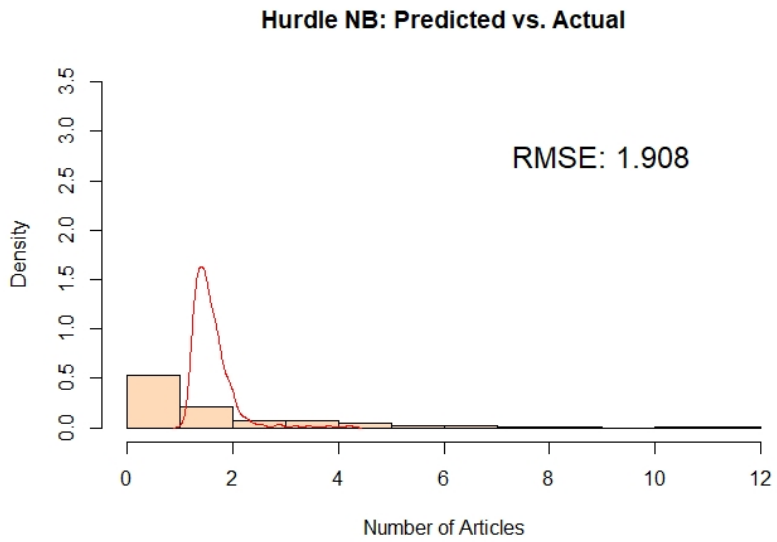
Interpretation

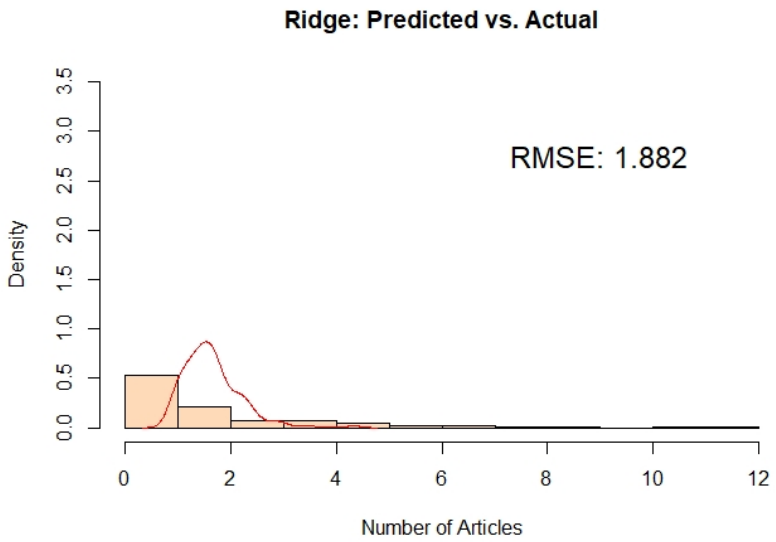
- **GLM, Ridge, and Lasso**: interpretable coefficients
- **KRLS**: average marginal effects
- **CART**: the pruned tree chose mentor rating and Ph.D. program rating
- Conceptually, which models are the most helpful for the question we want to answer?

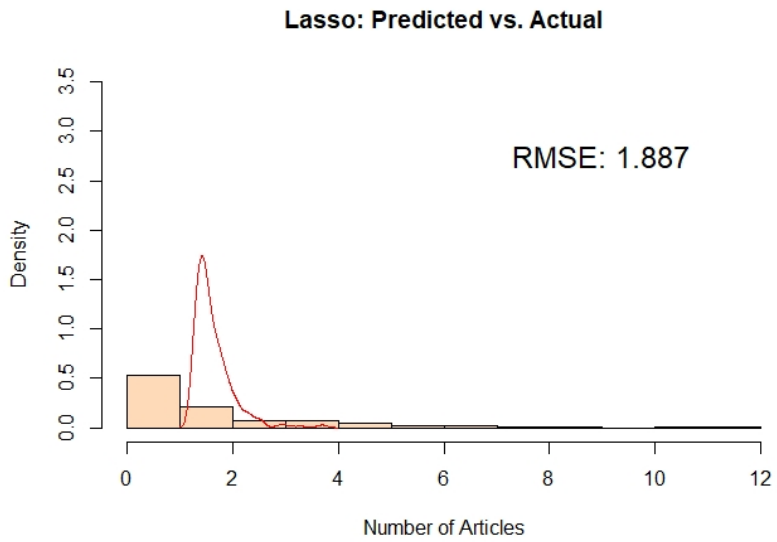
Negative Binomial	
Intercept	0.049
1fem	-0.208 *
1(married)	0.243 *
kid$\leq$ 5	-0.240 ***
phdrating	0.058
mentorrating	0.029 ***
Number of parameters	7
Log likelihood	-1025.808
RMSE (Training)	2.182

Application to Test Data

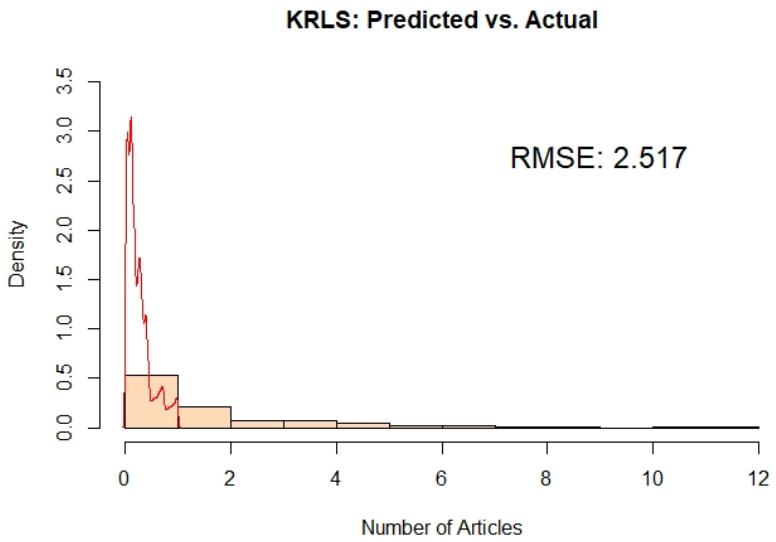


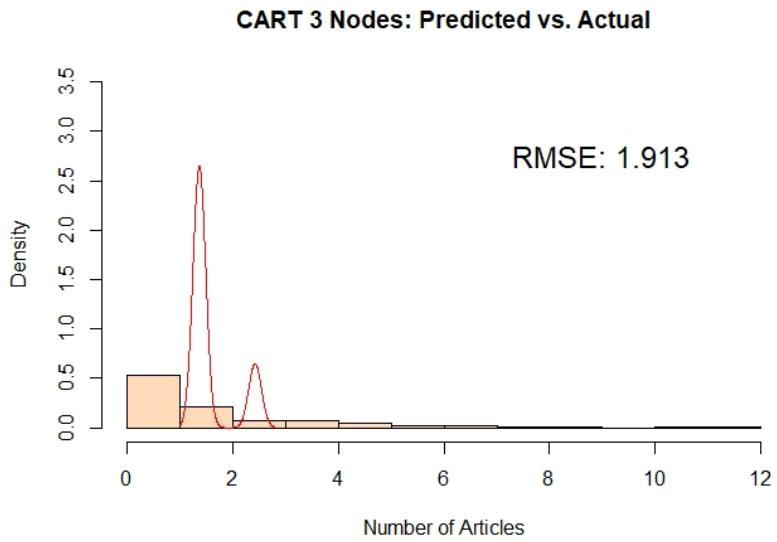


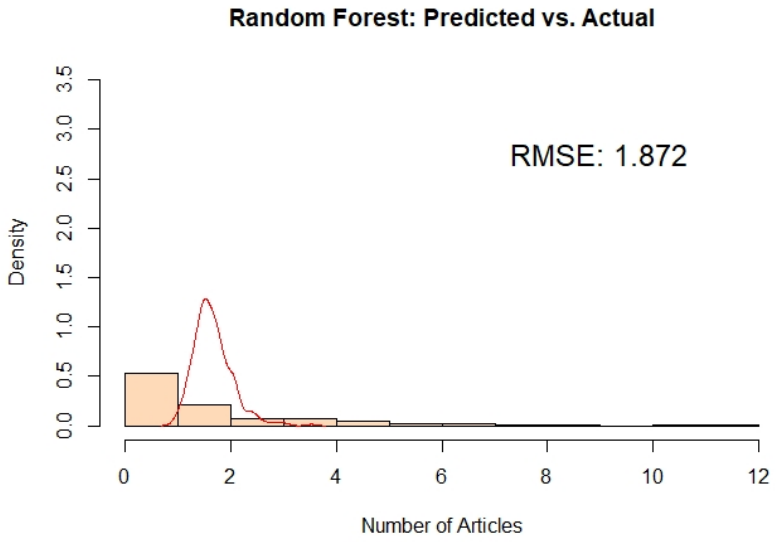


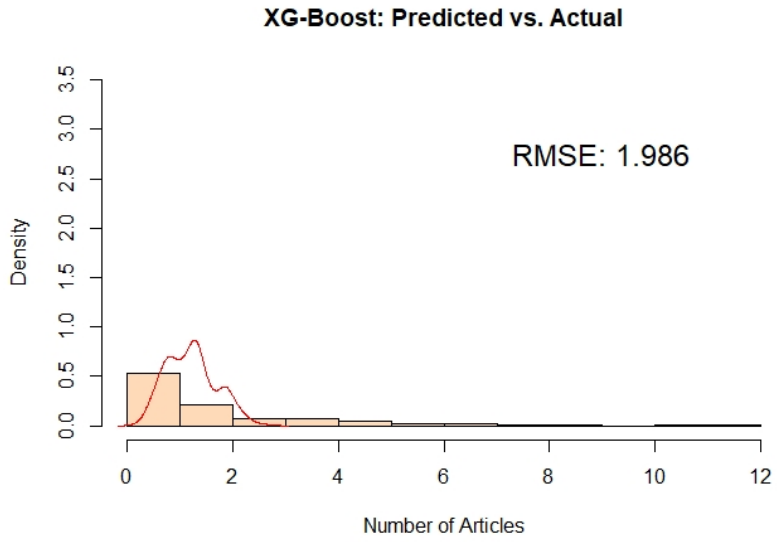


Application to Test Data (Cont'd)









Conclusion

- GLM methods are more interpretable, while tree-based methods give more accurate predictions.
- Regularization methods strike a good balance between interpretability and accuracy
  - But are they the way to go?
- Chosen method should align with goals of researcher (prediction of count variable versus estimation of effect size)
- We can try tuning XG boost further in the future.

References

Thanks. B)