Research Question 1:
Q2. To what extent do demographic or community factors correlate with the rate of improving instruction

Datasets: NCES, CWIS

CWIS

Rows = 80267 & Columns = 106

- Survey responses
- Also, if a district met with a coaching team, but didn't fill out the CWIS survey, you'll find the district in the Coaching Logs but not in the CWIS survey data.
- Building/School level data i.e identifier State.School.ID
- 867 unique school IDs

NCES

Rows = 2456 & Columns = 26

- National Common Core data that includes descriptions of the buildings (**e.g., Free/Reduced lunch rate, student:teacher ratio, rural/urban, etc.**)
- Building/School level data i.e identifier State.School.ID
- 2456 nces unique school IDs

CWIS + NCES

Rows = 78558 & Columns = 123

- Merging CWIS,NCES based on State.School.ID
- **1709 missing rows are because :**
- There are only 839 common school IDs between CWIS and NCES.
- NCES data doesn’t have all the school IDs

Research Question 2:
Q3. What attributes of external support (externally provide training, coaching, DESE support) influence the rate of improving instruction? What are the conditions that cause the contribution of these variables to vary?

Datasets: Coaching, CWIS

Coaching Logs ⟶ Districts Aggregate

Rows = 5924 & Columns = 58 **Rows = 192 & Columns = 14**

- Each line of the coaching logs is an interaction between a school and a Coach.
- Identifier for merging tables - **State.District.ID**
- **MO-001090(adair co. r-i) -> (multiple schools, building in CWIS,NCES) -> we need an aggregate**

Need for Aggregation?

- Doesn't have building level or school level data unlike NCES, CWIS.
- To Merge NCES,CWIS data with Coaching logs we need to generalize the identifier over the 3 files.
- **MO-001090(adair co. r-i) -> (multiple schools, building in CWIS,NCES) -> we need an aggregate**

District Level Data - Coaching Logs

- Each Row represent the coaching log for a school in that district. But we don't have any school ID to help us understand which school represents which row in the given district
- 0 - represents the school didn't receive the coaching of that particular type
- 1 - represents the school did receive the coaching of that particular type

Input

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	State.District.ID	Districts	Collabor	Comm	Data.b	Effect	Instructic	School.ba	Collect	Practice.	Self.asses	Learning	DESE.virtu	CWIS
2	MO-001090	adair co. r-i	0	0	0	1	0	0	0	0	1	0	1	0
3	MO-001090	adair co. r-i	0	0	0	0	0	0	0	1	0	0	1	1
4	MO-001090	adair co. r-i	0	0	1	1	0	0	0	1	1	0	1	1
5	MO-001090	adair co. r-i	0	0	0	1	0	0	0	1	1	0	0	1
6	MO-001090	adair co. r-i	0	0	0	0	0	1	0	1	1	1	1	0
7	MO-001090	adair co. r-i	0	0	0	1	0	0	0	1	1	1	1	0
8	MO-001090	adair co. r-i	1	0	1	1	0	0	0	1	0	1	0	0

Output

MO-001090	**adair co. r-i**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**1**	**1**	**1**	**1**

CWIS → CWIS Aggregation

Rows = 80267 & Columns = 97

Rows = 230 & Columns = 90

```
cwis.dt[, cwis_mean_cols] <- cwis.dt[, lapply(.SD, mean), .SDcols = cwis_mean_cols]
cwis.aggregates <- unique(cwis.dt, by = "State.District.ID")
```

- cwis_required_cols include all the columns(except the 7 common practices, state, school Identifiers).
- ETL Average for district will be the average of all the schools in the given district.

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	State.District.	experien	member_g	admin_receive_c	district_accept_c	ETL.AV	CFA.A	DBDM	PD.AV	common_practic	common_practic	common_practic	common_practic	collab_
2	MO-002097	8.099499	0.6206286	-0.9800789864	0.002217598764	2.9458	3.626	2.7429	3.4299	3.640269849	3.506700014	3.676486227	3.552602938	2.6
3	MO-004110	8.099499	0.6206286	-0.9800789864	0.002217598764	2.9458	3.626	2.7429	3.4299	3.640269849	3.506700014	3.676486227	3.552602938	2.6
4	MO-005121	8.099499	0.6206286	-0.9800789864	0.002217598764	2.9458	3.626	2.7429	3.4299	3.640269849	3.506700014	3.676486227	3.552602938	2.6
5	MO-005123	8.099499	0.6206286	-0.9800789864	0.002217598764	2.9458	3.626	2.7429	3.4299	3.640269849	3.506700014	3.676486227	3.552602938	2.6
6	MO-007129	8.099499	0.6206286	-0.9800789864	0.002217598764	2.9458	3.626	2.7429	3.4299	3.640269849	3.506700014	3.676486227	3.552602938	2.6

CWIS Aggregation

Rows = 230 & Columns = 90

Districts Aggregate

Rows = 192 & Columns = 14

cwis.aggregrate.districts <- districts.aggregate.dt[cwis.aggregate,on=.(State.District.ID),nomatch = NULL]

CWIS Districts Aggregation

Rows = 183 & Columns = 103

- Each row represents combination of cwis survey per district and coaching level data.
- **38 missing rows are because :**
- Indicating few of the districts weren't involved in Coaching has no CWIS survey record.

Problems

- Need building/School level data in Coaching Logs to get Improving Instruction school wise. If not can we continue with the district level data for now?
- What should be the Consistency metric per month? Is it reasonable to consider the Fall(Aug-Mar)(To be effective?), Spring(April -July)
- When combining CWIS + Coaching for ETL average as we are using districts as identifiers?

1. Do we join for dates based on Closest Survey Data(nearest neighbors approach)?
2. Do we join based on months of coaching and survey?

i.e Coaching (Aug 12 - 15)

Consider surveys from Aug(16- until next coaching)? This is because we don't have date/duration of coaching for a given school in Coaching logs.

Baseline/Featureless Models

- Simple Models that provide reasonable results and requires less expertise or time to build.
- Baselines predictions are independent of the inputs.
- Why Baseline?

Hypothesis: If there is any relation between inputs and outputs

- If the baselines do better than our models(regression or classification) it's mean we can assume the inputs has no relation with the outputs.

<u>Baselines Used:</u>

- **L0 - Classification Baseline -> most frequent values**
- **L1 - Regression Baseline -> median**
- **L2 - Regression Baseline -> mean**

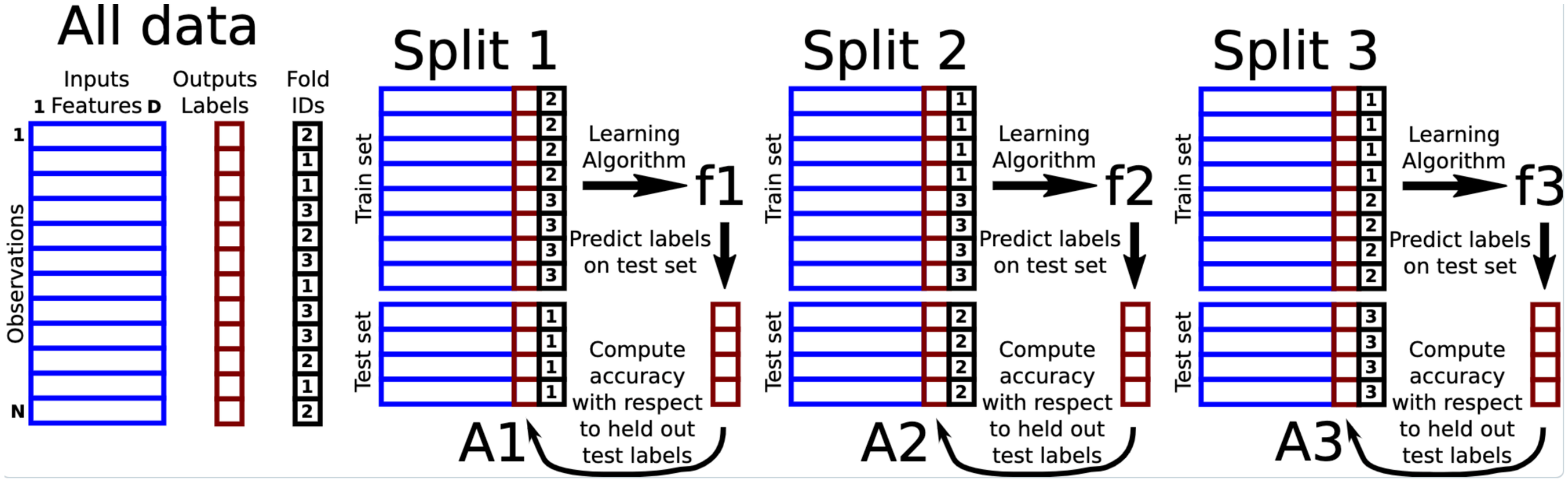
Linear Regression

LASSO

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

- Sum of Squared Residuls + lambda|slope| <- lasso regression penalty
- Regularization works by penalizing the magnitude of coefficients of the features.
- L1 tends to shrink coefficients to zero.
- L1 is therefore useful for feature selection, as we can drop any variables associated with coefficients that go to zero

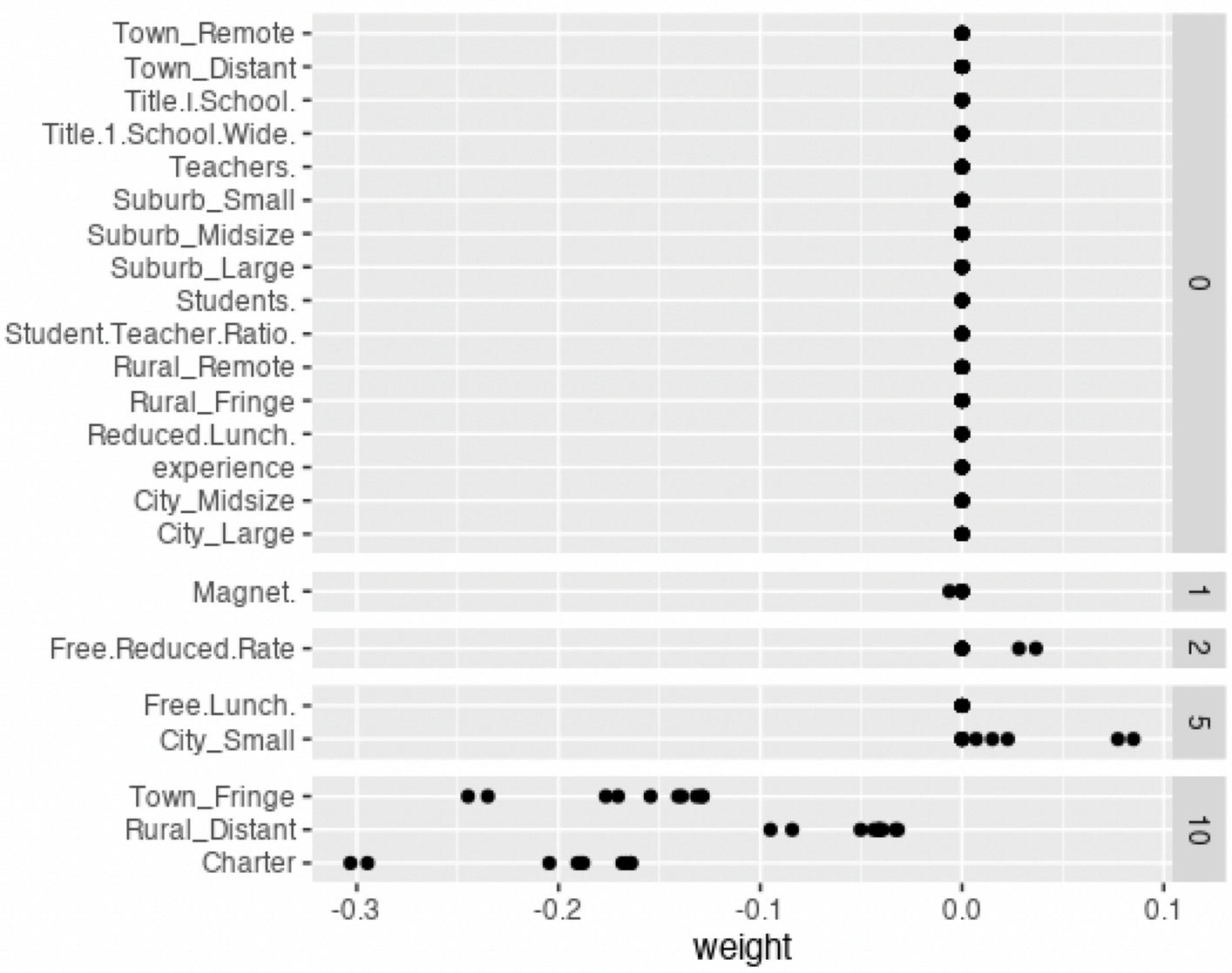
cvglmnet - cross validation



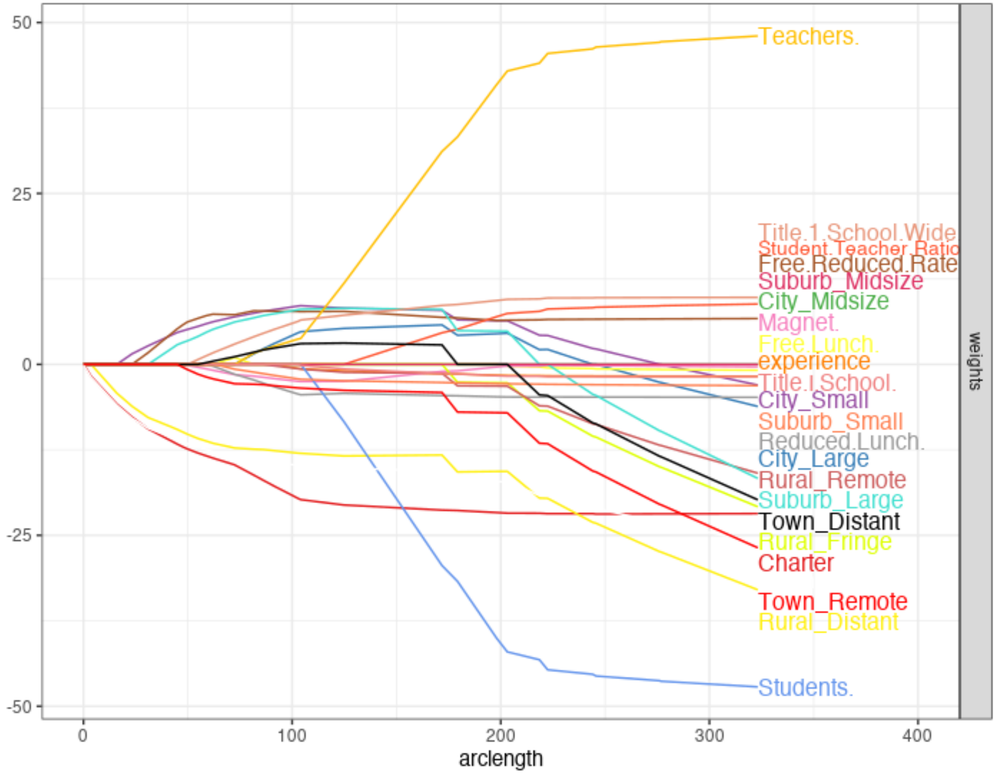
Source – https://github.com/tdhock/2020-yiqi-summer-school

Results

cv.glmnet - 10 folds



Lasso - Complete DataSet

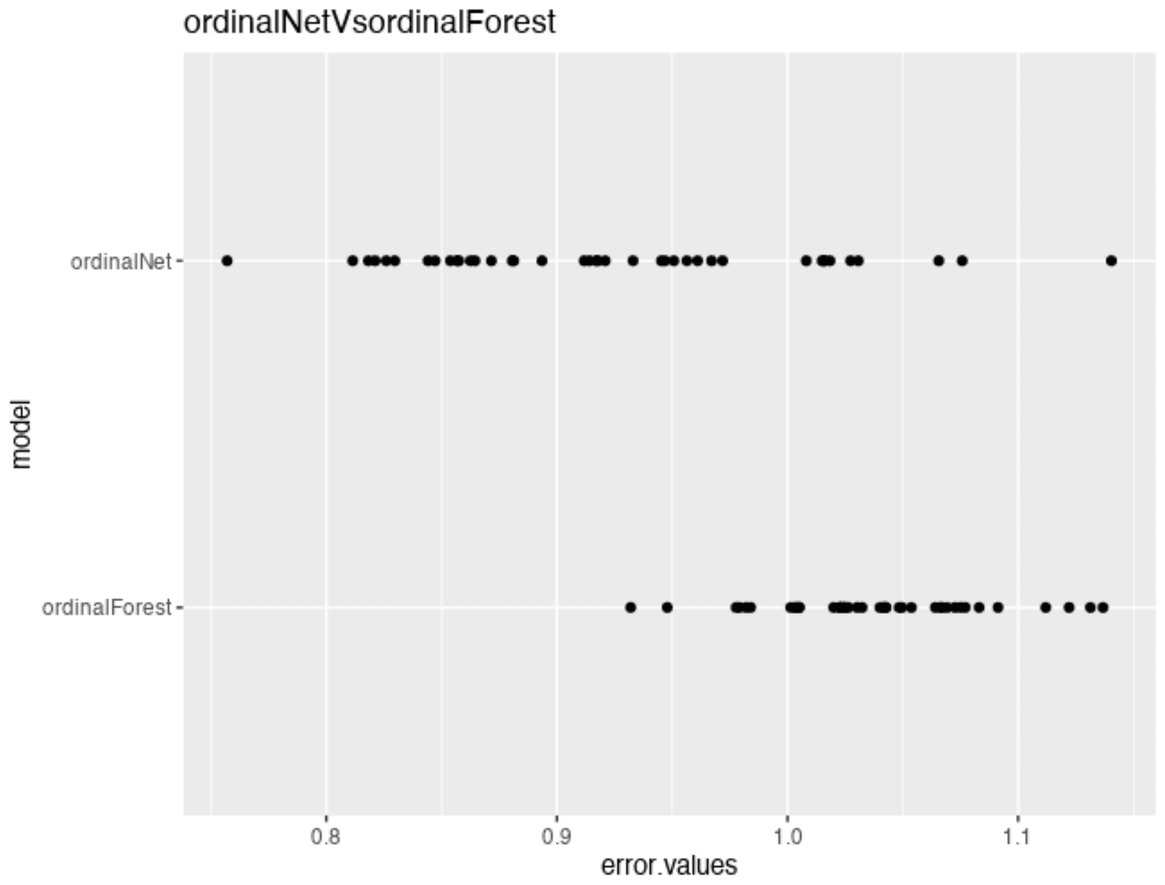


OrdinalNet

- In statistics, ordinal regression (also called "ordinal classification") is a type of regression analysis used for predicting an ordinal variable, with 'ordered' multiple categories and independent variables.
- Ordinal Net fits ordinal regression models with elastic net penalty.

Ordinal Forest

- The ordinal forest (OF) method allows ordinal regression with high-dimensional and low-dimensional data.
- Moreover, by means of the (permutation-based) variable importance measure of OF, it is also possible to rank the covariates with respect to their importance in the prediction of the values of the ordinal target variable.



OrdinalNet Vs glmnet Vs Baselines