Machine Learning: Lab 3

Jingpeng Hong

March 1, 2022

1 Descriptive Statistics

The descriptive statistics for the subset of PM COVID and Opportunity Insights variables filtered in question 1 are shown in Table 1 and Table 2, respectively.

Table 1: Descriptive Statistics (PM COVID Variables)

Statistic	N	Mean	St. Dev.	Min	Max
$PM_{2.5}$ Air Pollution	3,107	8.37	2.57	0.00	15.79
Missing $PM_{2.5}$ Variable Recoded to 0	3,107	0.004	0.06	0	1
Summer Max Temperature	3,107	303.13	3.17	290.46	313.87
Summer Max Relative Humidity	3,107	88.97	9.69	31.64	99.78
Winter Max Temperature	3,107	280.40	6.60	264.69	298.34
Winter Max Relative Humidity	3,107	87.47	4.81	58.16	97.67
Baseline Mortality Crude Rate	3,107	1,029.16	248.38	189.30	1,978.60

Table 2: Descriptive Statistics (Opportunity Insights Variables)

Statistic	N	Mean	St. Dev.	Min	Max
Urban Area	3,107	0.60	0.49	0	1
BRFSS: Fraction Current Smokers in Q1	3,107	0.21	0.15	0	1
BRFSS: Fraction Current Smokers in Q2	3,107	0.17	0.13	0	1
BRFSS: Fraction Current Smokers in Q3	3,107	0.13	0.13	0	1
BRFSS: Fraction Current Smokers in Q4	3,107	0.10	0.11	0	1
BRFSS: Fraction Obese in Q1	3,107	0.24	0.17	0	1
BRFSS: Fraction Obese in Q2	3,107	0.21	0.15	0	1
BRFSS: Fraction Obese in Q3	3,107	0.21	0.18	0	1
BRFSS: Fraction Obese in Q4	3,107	0.19	0.17	0	1
BRFSS: Fraction Exercised in Past 30 Days in Q1	3,107	0.46	0.27	0	1

BRFSS: Fraction Exercised in Past 30 Days in Q2	3,107	0.56	0.32	0	1
BRFSS: Fraction Exercised in Past 30 Days in Q3	3,107	0.60	0.36	0	1
BRFSS: Fraction Exercised in Past 30 Days in Q4	3,107	0.64	0.38	0	1
Missing BRFSS Variable(s) Recoded to 0	3,107	0.25	0.43	0	1
Percent Uninsured	3,107	18.47	5.54	3.63	41.37
Medicare $ Per Enrollee	3,103	9,302.74	1,590.93	3,663.53	18,443.22
30-day Hospital Mortality Rate Index	3,106	0.46	1.21	−7.78	8.47
30-day Mortality for Heart Attacks	3,106	0.17	0.04	0.00	0.44
30-day Mortality for Heart Failure	3,107	0.11	0.02	0.00	0.34
Mean of Z-Scores for Dartmouth Atlas Ambulatory Care Measures	3,012	−0.15	0.86	−4.85	3.48
Percent of Medicare Enrollees with at Least One Primary Care Visit	3,098	80.87	7.40	18.33	95.67
Percent Diabetic with Annual Hemoglobin Test	3,069	83.71	6.59	16.91	100.00
Percent Diabetic with Annual Eye Test	3,054	66.08	7.60	31.37	90.00
Percent Diabetic with Annual Lipids Test	3,057	78.31	7.85	19.66	94.48
Percent Female Aged 67-69 with Mammogram	3,029	63.11	8.40	30.00	95.24
Income Segregation	3,107	0.03	0.03	−0.01	0.44
Segregation of Poverty (<p25)	3,107	0.02	0.03	−0.02	0.75
Segregation of Affluence (>p75)	3,107	0.03	0.03	−0.002	0.20
Racial Segregation	3,107	0.08	0.08	0.00	0.71
Gini Index Within Bottom 99%	3,008	0.38	0.09	0.16	1.09
Poverty Rate	3,107	0.14	0.07	0.00	0.57
Top 1% Income Share	3,008	0.09	0.05	0.02	0.73
Fraction Middle Class (p25-p75)	3,106	0.55	0.09	0.22	0.88
Social Capital Index	3,107	0.0002	1.35	−4.26	9.91
Percent Religious	3,106	53.22	18.50	1.82	164.53
Percent Black	3,107	8.74	14.48	0.00	85.97
Percent Hispanic	3,107	6.21	12.05	0.08	97.54
Unemployment Rate in 2000	3,107	0.05	0.02	0.02	0.18
Labor Force Participation	3,107	0.61	0.07	0.32	0.86
Share Working in Manufacturing	3,107	0.16	0.09	0.00	0.49
Percent Foreign Born	3,107	3.44	4.84	0.00	50.94
Migration Inflow Rate	3,017	0.03	0.02	0.00	0.17
Migration Outflow Rate	3,017	0.03	0.01	0.00	0.15
Population Density	3,107	244.33	1,676.10	0.10	66,940.08

Fraction with Commute < 15 Min	3,107	0.40	0.14	0.10	0.82
Mean Household Income	3,107	32,853.50	6,975.84	10,511.81	77,942.65
Median House Value	3,107	112,180	63,189	0	1,333,001
School Expenditure per Student	3,080	6.09	2.10	3.03	53.26
Test Score Percentile (Income Adjusted)	3,069	0.08	9.01	−38.69	32.99
Fraction of Children with Single Mother	3,107	0.19	0.07	0.02	0.54
Local Government Expenditures	3,107	2,119.41	999.83	0.00	20,541.92
Local Tax Rate	3,107	0.02	0.01	0.00	0.21
Tax Progressivity	3,106	0.78	1.47	0.00	7.22

2 OLS Estimation

2.1 a

Table 3: MSE & R^2

	MSE	R^2
Training Set	1594.57	0.483
Test Set	1923.50	0.420

2.2 b

From the results in Table 3, we find that the MSE of the test set is much larger than that of the training set. Hence, we suspect that overfitting has occurred and we expect to see a less flexible model with a smaller test MSE.

3 Ridge & Lasso

3.1 Ridge

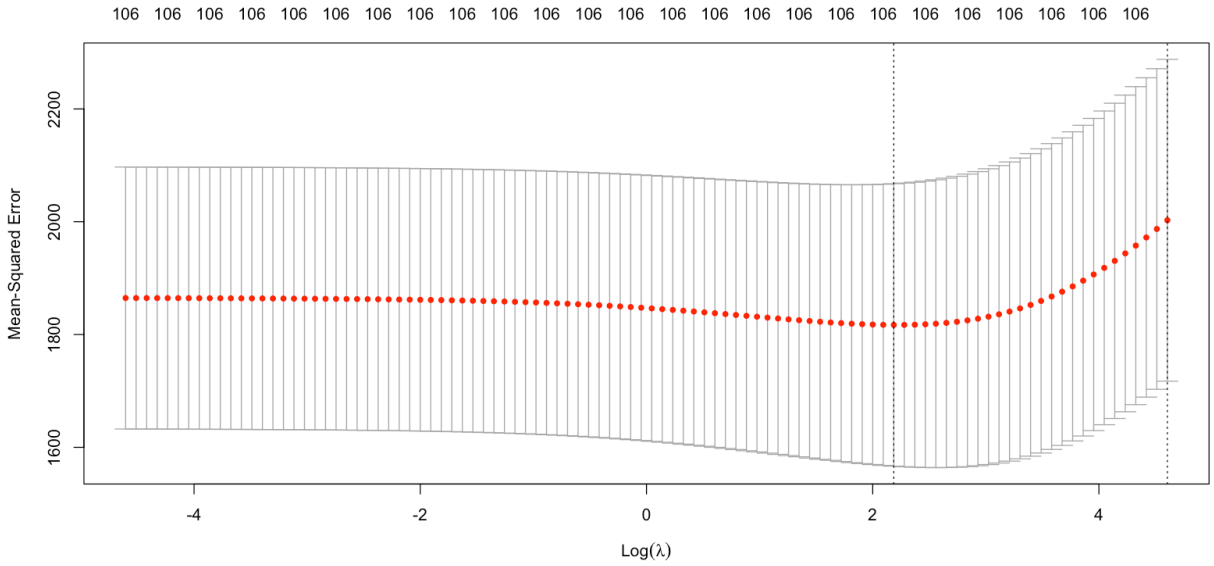

Figure 1: Ridge: Test error & λ

3.2 Lasso

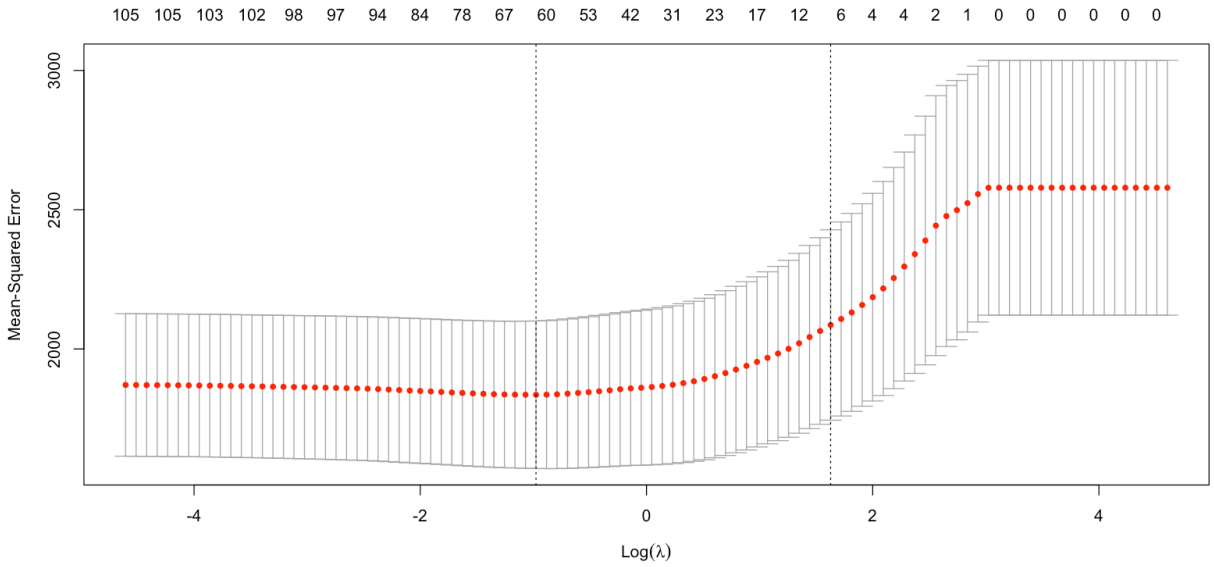

Figure 2: Lasso: Test error & λ

3.3 Estimation

	Lasso	Ridge
Optimal λ	0.376	8.902
Training MSE	1629.58	1636.39
Test MSE	1879.84	1749.64
Training R^2	0.367	0.365
Test R^2	0.433	0.472

Table 4: Lasso & Ridge Estimation Results

From Table 4, we find that Lasso and Ridge indeed mitigate the overfitting. Compared with OLS, Lasso and Ridge both have better out-of-sample performance, i.e., lower test set MSE and higher test set R^2.

3.4 Bonus Question

We find that the test-set prediction errors from the Ridge and the Lasso are smaller than that from OLS. We can consider this from the perspective of bias-variance tradeoff. When we increase the λ in Lasso and Ridge, we decrease the flexibility of the Ridge and Lasso regression fit, leading to decreased variance but increased bias. In this data set, we have over 100 variables, and some of them are highly correlated. Hence, the shrinkage of the Ridge and Lasso coefficient estimates will only slightly increase the bias, with a substantial reduction in the variance. Then, the test MSE, which is a function of the variance plus the squared bias, will drop considerably.