kNN ok

$n >> p$ or $n << p$

kNN ☹

can fit Linear Regression

Linear regresson blows up

Statistical Machine Learning

Linear, Ridge and LASSO Regression and kNN

A comparison when n ≈ p

Day 18

n = # examples

p = # predictors

AI in the news…

“As part of an effort to combat the US’s growing prison population, the US attorney-general is required to develop an ‘evidence-based’ risk assessment system by July 2019 to help decide how long inmates remain incarcerated.”

FT, 27 April/28 April 2019

Let’s build a model to predict “Violent Crime”

About Citation Policy Donate a Data Set Contact

Search

Repository Web Google™

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

View ALL Data Sets

Communities and Crime Unnormalized Data Set

Download: Data Folder, Data Set Description

Abstract: Communities in the US. Data combines socio-economic data from the '90 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 FBI UCR

n

Data Set Characteristics:	Multivariate	**Number of Instances:**	2215	**Area:**	Social
Attribute Characteristics:	Real	**Number of Attributes:**	147	**Date Donated**	2011-03-02
Associated Tasks:	Regression	**Missing Values?**	Yes	**Number of Web Hits:**	121688

Source:

-- Creator: Michael Redmond (redmond 'at' lasalle.edu); Computer Science; La Salle University; Philadelphia, PA, 19141, USA
-- culled from 1990 US Census, 1995 US FBI Uniform Crime Report, 1990 US Law Enforcement Management and Administrative Statistics Survey, available from ICPSR at U of Michigan.
-- Donor: Michael Redmond (redmond 'at' lasalle.edu); Computer Science; La Salle University; Philadelphia, PA, 19141, USA

Available predictors…

1 -- population: population for community: (numeric - expected to be integer)

2 -- householdsize: mean people per household (numeric - decimal)

3 -- racepctblack: percentage of population that is african american (numeric - decimal)
4 -- racePctWhite: percentage of population that is caucasian (numeric - decimal)
5 -- racePctAsian: percentage of population that is of asian heritage (numeric - decimal)
6 -- racePctHisp: percentage of population that is of hispanic heritage (numeric - decimal)
7 -- agePct12t21: percentage of population that is 12-21 in age (numeric - decimal)
8 -- agePct12t29: percentage of population that is 12-29 in age (numeric - decimal)
9 -- agePct16t24: percentage of population that is 16-24 in age (numeric - decimal)
10 -- agePct65up: percentage of population that is 65 and over in age (numeric - decimal)
11 -- numbUrban: number of people living in areas classified as urban (numeric - expected to be integer)
12 -- pctUrban: percentage of people living in areas classified as urban (numeric - decimal)
13 -- medIncome: median household income (numeric - may be integer)
14 -- pctWWage: percentage of households with wage or salary income in 1989 (numeric - decimal)
15 -- pctWFarmSelf: percentage of households with farm or self employment income in 1989 (numeric - decimal)
16 -- pctWInvInc: percentage of households with investment / rent income in 1989 (numeric - decimal)
17 -- pctWSocSec: percentage of households with social security income in 1989 (numeric - decimal)
18 -- pctWPubAsst: percentage of households with public assistance income in 1989 (numeric - decimal)

…

106 -- PolicReqPerOffic: total requests for police per police officer (numeric - decimal)
107 -- PolicPerPop: police officers per 100K population (numeric - decimal)
108 -- RacialMatchCommPol: a measure of the racial match between the community and the police force.
109 -- PctPolicWhite: percent of police that are caucasian (numeric - decimal)
110 -- PctPolicBlack: percent of police that are african american (numeric - decimal)
111 -- PctPolicHisp: percent of police that are hispanic (numeric - decimal)
112 -- PctPolicAsian: percent of police that are asian (numeric - decimal)
113 -- PctPolicMinor: percent of police that are minority of any kind (numeric - decimal)
114 -- OfficAssgnDrugUnits: number of officers assigned to special drug units (numeric - expected to be integer)
115 -- NumKindsDrugsSeiz: number of different kinds of drugs seized (numeric - expected to be integer)
116 -- PolicAveOTWorked: police average overtime worked (numeric - decimal)
117 -- LandArea: land area in square miles (numeric - decimal)
118 -- PopDens: population density in persons per square mile (numeric - decimal)
119 -- PctUsePubTrans: percent of people using public transit for commuting (numeric - decimal)
120 -- PolicCars: number of police cars (numeric - expected to be integer)
121 -- PolicOperBudg: police operating budget (numeric - may be integer)
122 -- LemasPctPolicOnPatr: percent of sworn full time police officers on patrol (numeric - decimal)
123 -- LemasGangUnitDeploy: gang unit deployed (numeric - integer - but really nominal - 0 means NO, 10 means YES,
124 -- LemasPctOfficDrugUn: percent of officers assigned to drug units (numeric - decimal)
125 -- PolicBudgPerPop: police operating budget per population (numeric - decimal)

The Crime dataset

There are p=125 quantitative pieces of information available for n = 2215 communities to predict the number of violent crimes per 100,000.

Let

Y = number of violent crimes per 100,000 people

$\mathbf{X} = (1, X_1, \ldots X_{125})$

True Relationship: $Y = f(\mathbf{X}) + \varepsilon$

Which methods are appropriate to try?

- Linear regression ✓
- kNN regression ✓
- kNN classification ✗
- Logistic regression ✗
- LDA
- QDA
- Ridge regression ✓

There are p=125 quantitative pieces of information available for n = 2215 communities to predict the number of violent crimes per 100,000.

Let

Y = number of violent crimes per 100,000 people

$\mathbf{X} = (1, X_1, \ldots X_{125})$

Training/selection/assessment

Unfortunately, there are a lot of “?” in the dataset…

When I remove communities with at least one piece of missing information, there are only n=319 communities left.

I’ll omit the assessment step and just train and select a model to suggest to the Attorney-general.

I’ll use 3-fold cross-validation, means that models will be trained on about 200 communities. So $(n = 200) \approx (p=125)$

Selection by MAE

I'll select the model with the best cross-validated

Mean Absolute Error (MAE)

$$MAE = \frac{1}{m}\sum_{i=1}^{m} |y_i - \hat{y}_i|$$

all m vales in test set

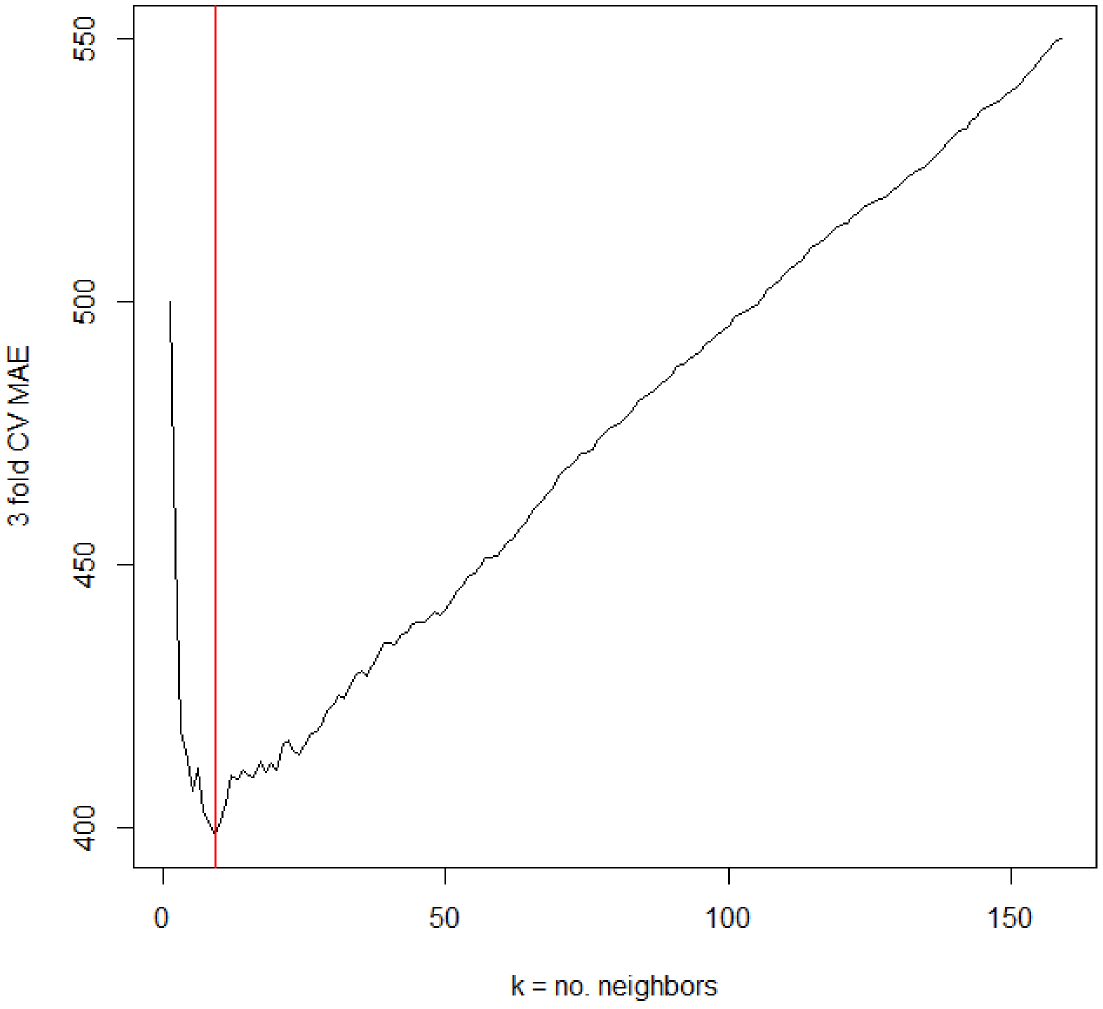
This is an easy measure to interpret – it's just the average error of the model

repeat

kNN Regression

3-fold cross-validated MAE: 398.7 at k=9

The model is off, on average, by 399 violent crimes per 100,000 people.

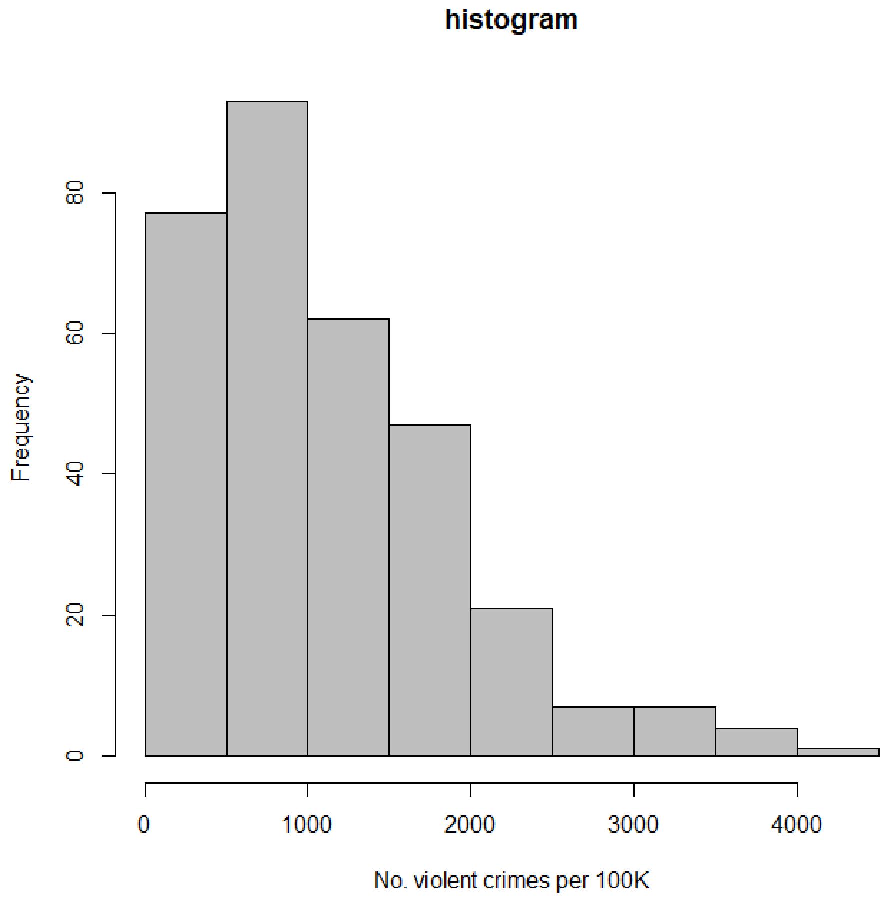


Linear Regression

3-fold cross-validated MAE: 666.9

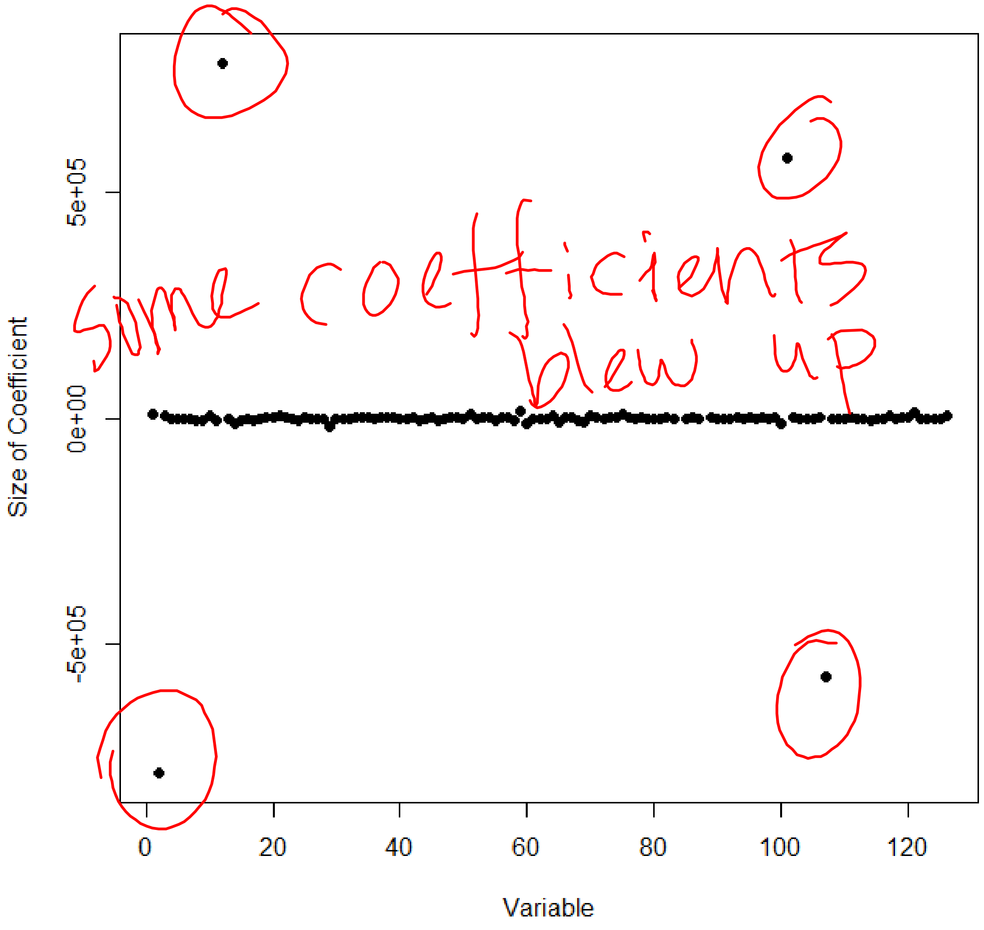
The model is off, on average, by 667 violent crimes per 100,000 people.

Some perspective:



What’s happening with Linear Regression?

```
Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9793.18    4855.33   2.017   0.0451 *
V6         -786386.38  351548.85  -2.237   0.0264 *
V7            6531.90    3870.36   1.688   0.0931 .
V8            -511.28    1611.55  -0.317   0.7514
V9            -319.70    1535.42  -0.208   0.8353
V10             84.62    1358.78   0.062   0.9504
V11           -360.91    1261.88  -0.286   0.7752
V12          -3608.42    4618.87  -0.781   0.4356
V13          -4675.59    4633.36  -1.009   0.3142
V14           4622.65    7321.53   0.631   0.5285
V15          -3919.81    3979.34  -0.985   0.3258
V16         787758.12  351498.75   2.241   0.0261 *
V17          -2117.90     952.00  -2.225   0.0272 *
V18         -12056.30    5431.52  -2.220   0.0276 *
V19          -5666.41    2627.07  -2.157   0.0322 *
V20          -2678.91    1010.73  -2.650   0.0087 **
V21          -3220.86    1288.51  -2.500   0.0133 *
V22          -1497.63    2752.04  -0.544   0.5869
V23            660.73    1421.50   0.465   0.6426
V24            850.01     948.38   0.896   0.3712
V25           5069.53    5104.68   0.993   0.3219
V26           2284.08    3741.31   0.611   0.5422
V27           -898.81    2791.56  -0.322   0.7478
V28          -4018.97    4053.86  -0.991   0.3227
V29           1350.29    1695.56   0.796   0.4268
V30            304.12     921.56   0.330   0.7418
V31           -716.43    1066.91  -0.671   0.5027
V32            -26.39     694.67  -0.038   0.9697
V33         -19731.01    7314.75  -2.697   0.0076 **
V34          -1229.27    1730.22  -0.710   0.4783
V35          -1504.30    1779.27  -0.845   0.3989
V36          -1448.93    1897.00  -0.764   0.4459
V37           2504.51    1602.29   1.563   0.1197
```



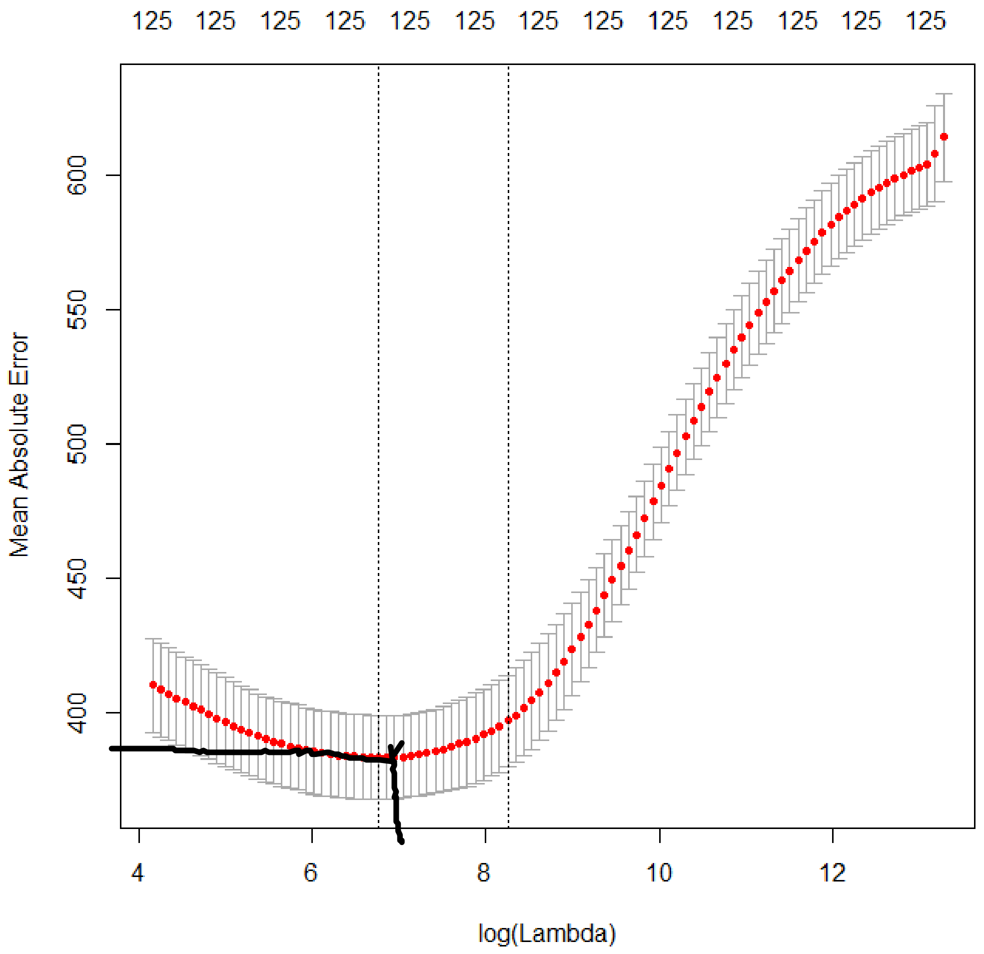
Ridge Regression

Best cross-validated MAE:

383.5 at λ = 865.3445

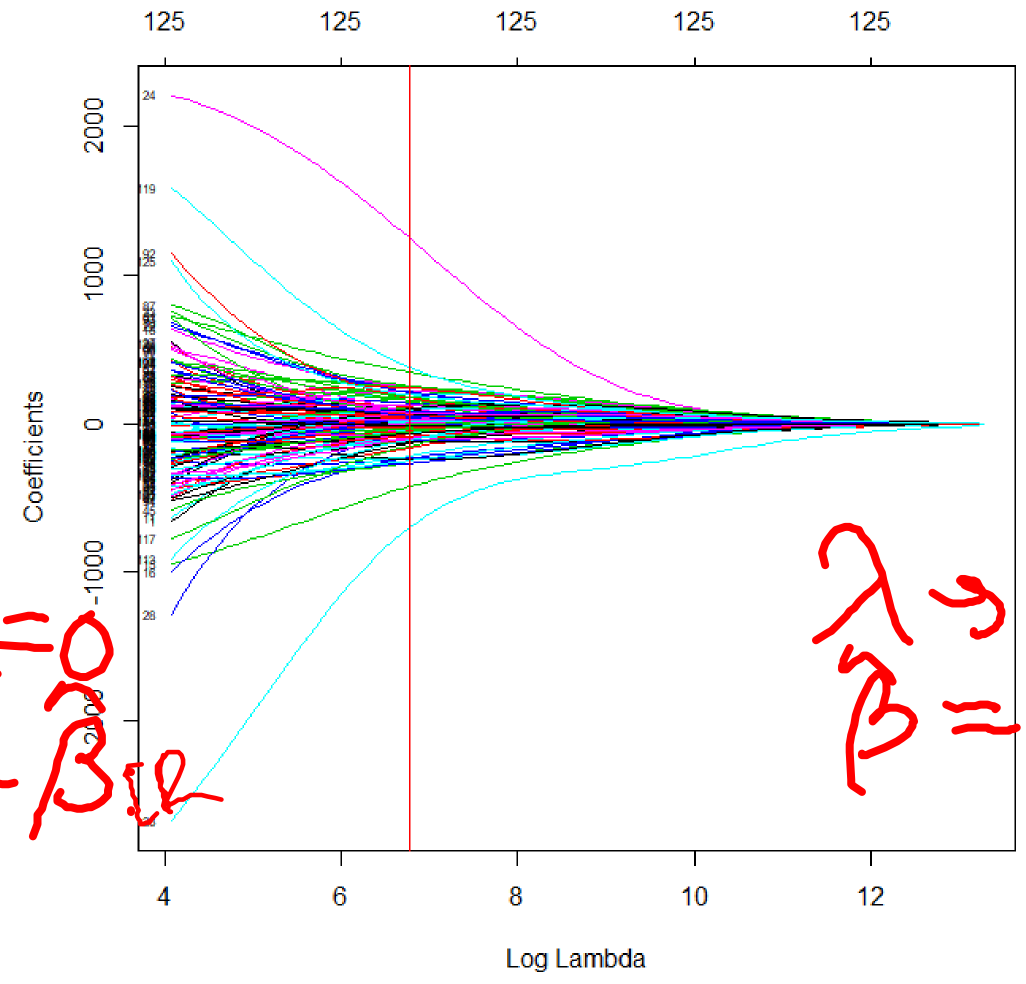
The model is off, on average, by

384 violent crimes per 100,000 people



Coefficients of “Best” Ridge Regression Model

var	coef	var	coef	var	coef
(Intercept)	1655.5299886	X43	68.1816928	X86	-21.7420117
X1	-25.9406933	X44	-268.6734087	X87	153.9386967
X2	31.6771839	X45	-270.0369352	X88	-12.2074142
X3	253.6215204	X46	-266.7784123	X89	94.9898535
X4	-241.1060424	X47	-241.9013078	X90	0.5753564
X5	-7.7106474	X48	79.1263784	X91	-135.9476513
X6	-0.8404995	X49	48.6251480	X92	171.7000971
X7	13.0517883	X50	37.7958524	X93	67.6431714
X8	-163.1090232	X51	345.1184421	X94	59.1452610
X9	-117.4990281	X52	58.7145850	X95	-101.4125248
X10	14.7159633	X53	-41.1922829	X96	61.3908462
X11	-25.4783550	X54	-32.2638457	X97	70.3107946
X12	26.9740744	X55	33.8815962	X98	60.0060485
X13	-117.7290150	X56	109.9933423	X99	-26.4912937
X14	-114.4387752	X57	-7.8824384	X100	-54.5273248
X15	-425.5980356	X58	13.1565185	X101	-63.8771521
X16	-234.0855010	X59	20.7717562	X102	-100.6357076
X17	-26.2195114	X60	38.4569656	X103	52.7116951
X18	244.8169960	X61	5.8848302	X104	42.9524209
X19	-126.2979337	X62	24.5868176	X105	157.5064347
X20	-85.0308688	X63	100.8602217	X106	-54.5175791
X21	25.4030863	X64	39.9275534	X107	-249.4361533
X22	212.6749419	X65	-15.2593080	X108	-92.3897942
X23	-703.3025639	X66	82.7231759	X109	100.6166765
X24	1251.5794359	X67	-58.8026489	X110	-87.8502968
X25	79.5174122	X68	-24.1535080	X111	180.1292554
X26	56.9404134	X69	159.0554848	X112	37.1604905
X27	141.0275606	X70	214.9933593	X113	-109.4079050
X28	-54.1535938	X71	-59.6524941	X114	31.1973694
X29	97.1378066	X72	69.7405702	X115	-68.5743070
X30	-56.4716972	X73	-239.6287457	X116	240.2279076
X31	22.8392993	X74	-38.3487533	X117	-152.2577582
X32	-9.4608207	X75	242.5791949	X118	83.0179093
X33	255.2446486	X76	-21.5011064	X119	382.8650393
X34	-113.7162632	X77	24.1582147	X120	74.2974657
X35	-216.4232295	X78	119.8836865	X121	-13.4444366
X36	96.9022541	X79	91.1979114	X122	48.0421912
X37	11.1448310	X80	-51.8435693	X123	-54.1244668
X38	-3.0859963	X81	-34.7838020	X124	52.0602389
X39	217.1387325	X82	-20.0464827	X125	193.4160188
X40	139.2293310	X83	8.6270720		
X41	201.2565753	X84	-96.3995678		
X42	218.5047427	X85	-19.6537283		



$\lambda = 0$

$\hat{\beta} = \hat{\beta}_{LR}$

$\lambda \to \infty$

$\hat{\beta} = 0$

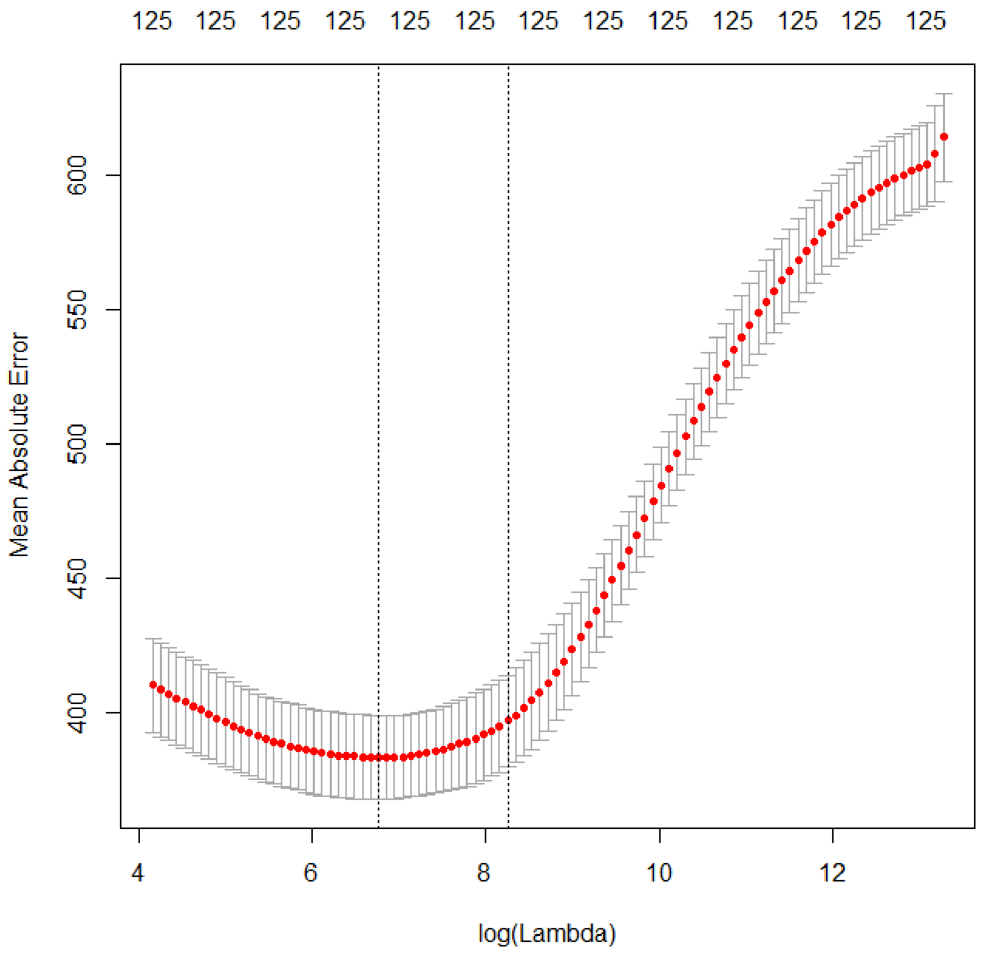
LASSO

Best cross-validated MAE:

381.5 at λ = 75.26325

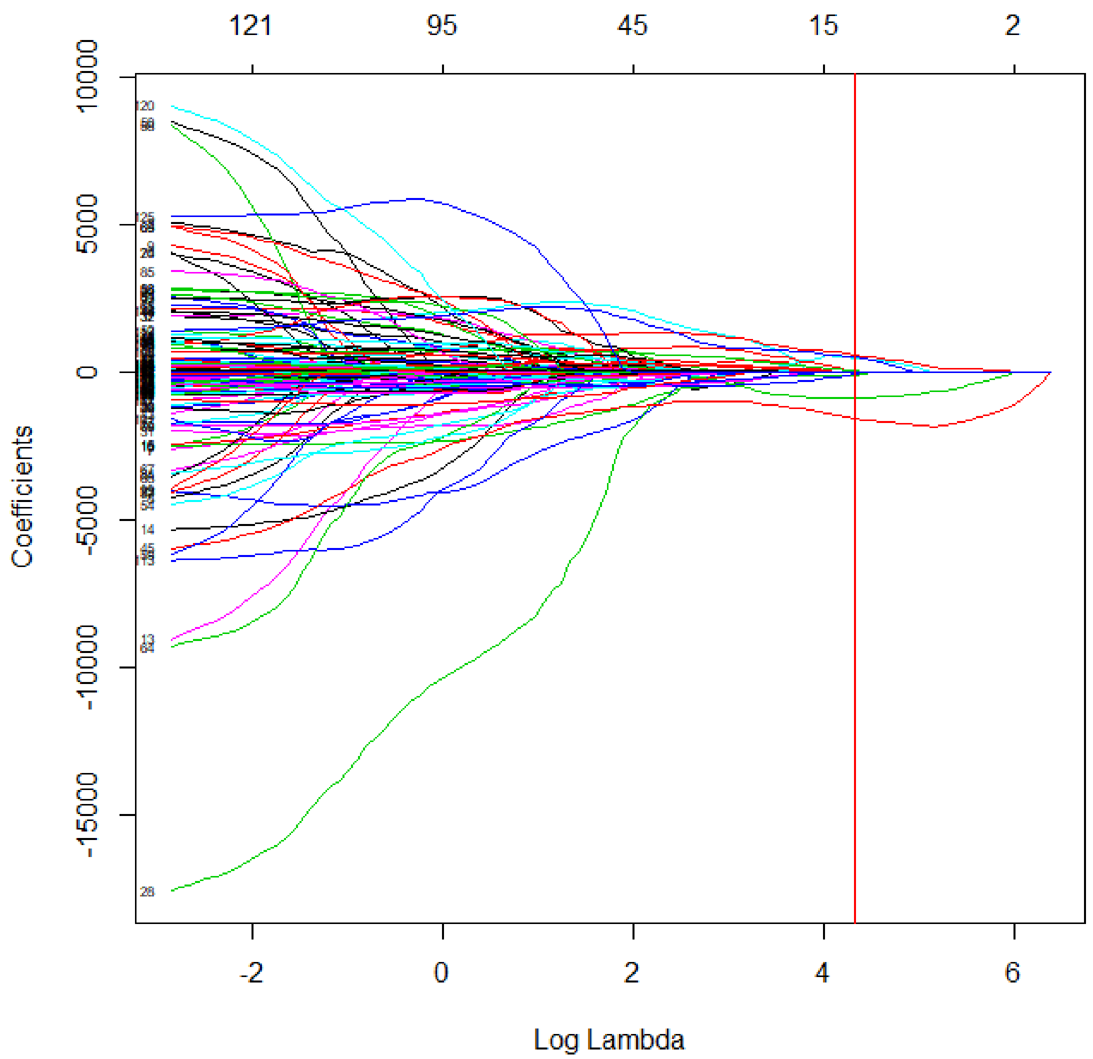
The model is off, on average, by

382 violent crimes per 100,000 people



Coefficients of “Best” LASSO Model

```
(Intercept)  2297.9344232
X1           .
X2           .
X3           .
X4           -864.6335292
X5           .
X6           .
X7           .
X8           .
X9           .
X10          .
X11          .
X12          .
X13          .
X14          .
X15          .
X16          -87.1812780
X17          .
X18          .
X19          .
X20          .
X21          .
X22          .
X23          .
X24          .
X25          .
X26          .
X27          .
X28          .
X29          .
X30          .
X31          .
X32          .
X33          0.3007172
X34          .
X35          -67.9043955
X36          .
X37          .
X38          .
X39          .
X40          .
X41          .
X42          494.3948635
X43          .
X44          .
X45          -1569.8071172
X46          .
X47          .
X48          .
X49          .
X50          .
X51          592.8828027
X52          .
X53          .
X54          .
X55          .
X56          .
X57          .
X58          .
X59          .
X60          .
X61          .
X62          .
X63          .
X64          .
X65          .
X66          .
X67          .
X68          .
X69          .
X70          35.9197947
X71          .
X72          .
X73          .
X74          .
X75          .
X76          .
X77          .
X78          .
X79          .
X80          .
X81          .
X82          .
X83          .
X84          .
X85          .
X86          .
X87          .
X88          .
X89          .
X90          .
X91          .
X92          .
X93          .
X94          .
X95          .
X96          .
X97          .
X98          .
X99          .
X100         .
X101         .
X102         .
X103         .
X104         .
X105         .
X106         .
X107         .
X108         .
X109         .
X110         .
X111         .
X112         .
X113         .
X114         .
X115         .
X116         .
X117         .
X118         .
X119         513.2171350
X120         .
X121         .
X122         .
X123         .
X124         .
X125         .
```



The “best” model out of kNN, Linear, Ridge and LASSO is…

LASSO, both in terms of performance (lowest MAE) and easy of interpretability:

Predicted number of Violent Crimes = 2298
-865 * standardized percentage of population that is Caucasian
-87 * standardized percentage of households with investment / rent income in 1989
0.3 * standardized percentage of people 16 and over, in the labor force, and unemployed
-68 * standardized percentage of people 16 and over who are employed in manufacturing
+492 * standardized percentage of females who are divorced
-1570 * standardized percentage of families (with kids) that are headed by two parents
+593 * standardized number of kids born to never married
+36 * standardized percent of persons in dense housing
+513 * standardized percent of people using public transit for commuting

What could we do to improve LASSO's performance?

Try

- Using communities with missing data
- Interactions or transformations of predictors
- Making minimal assumptions about the form of the relationship between Y and X…tomorrow!

What could we really do to improve the model's performance?

Artificial intelligence

US justice system's predictive tools under fire

CAMILLA HODGSON — SAN FRANCISCO

A research group founded by some of the world's most influential tech companies has found "serious shortcomings" in predictive policing tools being used across the US to make decisions about pretrial detention, probation and sentencing.

The Partnership on AI, set up in 2016 by companies including Google, Microsoft, Amazon and Facebook, said in an inaugural report yesterday that algorithmic risk assessment tools — which use statistical models to determine the probability of a future outcome — were not sufficiently accurate or transparent.

Law enforcement agencies are using such tools to predict, for example, whether someone will fail to appear in court based on their arrest history, demographic and how others have behaved in the past. But the report found "serious and unresolved problems with accuracy, validity and bias in both the data sets and statistical models that drive these tools".

A growing number of agencies in the US and overseas have begun experimenting with technologies such as predictive models, GPS tracking and facial recognition. But the technology has been criticised by opponents, who argue the tools reinforce racial biases and threaten human and civil rights.

Yesterday's paper was prompted by proposed legislation in California that would mandate the use of risk assessment tools in pretrial detention decision-making. The report said the use of such systems in the US criminal justice system was "expanding rapidly" despite "numerous, deeply concerning problems and limitations".

> Opponents argue the tools reinforce racial biases and threaten human and civil rights

As part of an effort to combat the US's growing prison population, the US attorney-general is required to develop an "evidence-based" risk assessment system by July 2019 to help decide how long inmates remain incarcerated.

But Peter Eckersley, the partnership's director of research, said that the tools currently available were "not suitable for deciding to detain or continue to detain individuals" and that in cases where the technology was required, defendants should also be granted in-person hearings.

The use of artificial intelligence has become increasingly controversial in recent years. Amazon has come under heavy criticism for selling its facial recognition tool to law enforcement, Google has disbanded its AI ethics board and Microsoft was revealed to have worked with a Chinese military-run university on AI that could be used for censorship and surveillance.

Nevertheless, many policymakers have endorsed the technology; since 2009, the US Department of Justice has given millions of dollars in grants to researchers and police forces for the development of "smart" policing tools.

However, just this month, the Los Angeles Police Department scrapped its "chronic offender" database, which was used to monitor people considered at high risk of committing violent crimes, following widespread concerns about inaccuracy and a damning audit from the department's inspector-general.