Prediction of Covid-19 spread and identification of contributing factors using lasso and ridge models

Valerie Brouwers (461925)

Supervisor: Dr. Anastasija Tetereva

Second assessor: Dr. Wendun Wang

ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Quantitative Finance

July 5, 2020

Abstract

The worldwide spread of Covid-19 has caused an increasing demand on medical care. Hence it is of importance to identify which factors influence the spread, and find a model that can forecast the regional spread of the disease. The data used consist of 213 features, which are county-specific data on health, socioeconomics and weather. This paper aims to predict the weekly increase of Covid-19 cases per county using these features with multiple lasso and ridge regressions, and determines which regressions perform best using the mean squared error. It also identifies which factors are associated with the spread of Covid-19 and accordingly makes policy recommendations. It determines that the pooled lasso and a version of the joint lasso method are the best models for predicting the increase in cases. Two policy recommendations are made: 1. Covid-19 tests and related health-care should be financially accessible for everyone 2. Residents of multi-unit housing should be given priority in testing.

Contents

1 Introduction

On December 31, 2019, a cluster of pneumonia cases was reported in Wuhan, China (World Health Organization, 2020). Later, the World Health Organization (WHO) declared a novel coronavirus called Covid-19 as the cause of the infections. Though at first the outbreak of Covid-19 was mainly limited to China, it soon after started to infect more people in Western countries, including the United States (US). On March 11, 2020, the spread of the disease had become uncontrollable in an alarming amount of countries in different parts of the world, for the WHO to declare the virus a pandemic. With 2.37 million confirmed Covid-19 cases as per June 21, the US currently has the most reported cases in the world. Because of the increasing demand on medical care in the US, it is imperative to identify which populations are at risk of contracting Covid-19 and protect high risk communities.

As the disease is still relatively new, much is uncertain about which factors play a role in Covid-19 transmission and fatality rates. It is suspected that health, socioeconomics, and weather conditions are all factors that play a role in this. This paper tries to identify these factors, combining county-level data on confirmed Covid-19 cases, health, socioeconomics, and weather conditions. Together these factors comprise of 213 features, which are used to predict the weekly increase of Covid-19 cases per county in the US.

Because there are 213 possible explanatory variables, the data is of high dimensionality. High-dimensional data sets can undermine the predictive ability of the model and compromise its efficiency because of high variance. Multiple models that are often used when the data are of high-dimensionality will be reviewed in this paper.

A popular selection method that is often used when the data is of high-dimensionality is lasso regression, which was proposed by Tibshirani (1996). Lasso regression performs parameter estimation and variable selection simultaneously, by setting some of the coefficients to zero. This process is called regularization and induces sparsity of the model, which reduces the high-dimensionality of the model and its corresponding negative effects. Another similar method is ridge regression, which is a shrinkage method. Ridge differs from lasso in the way it sets the coefficients: whereas lasso sets certain coefficients to zero, thereby selecting certain variables, ridge will only shrink coefficients closer to zero.

Recently, Dondelinger & Mukherjee (2018) proposed a new variation of the lasso method, called joint lasso regression. This method assumes that the dependent variables can be divided into subgroups, each having a different underlying model. If this is indeed the case for counties, a single regression model used for all the data, such as a normal lasso-regression, could be misspecified. It might seem intuitive to simply split the data into different groups, and then build a seperate model for each group. However, the sample size per model would then decline, which would exacerbate the negative effects of high-dimensionality even more. Furthermore, although the underlying models of the subgroups might not be entirely the same, there can still be similarities among them. The joint lasso method takes advantage of this: it allows subgroups to have different sparsity patterns, but at the same time induces global sparsity and encourages similarity between subgroup-specific coefficients. Dondelinger & Mukherjee (2018) showed the joint lasso has promising results

on biomedical applications. Not much research has been done in other applications yet. As there are many features involved in the spread of COVID-19, it poses an interesting field to apply the joint lasso to.

For the implementation of the joint lasso method, it is necessary to divide the counties into subgroups in some logical way. Because there is much uncertainty about the way Covid-19 spreads, it is difficult to determine which types of subgroups have different underlying models. Consequently, subgroups can only be defined based on speculation and logical reasoning, not on empirical/experimental data. Therefore, multiple ways of splitting the counties into subgroups have been evaluated, of which two perform best: subgroups based on the amount of Covid-19 cases per head, and subgroups based on the population density. Hence, only the results from these subgroups will be reviewed in this paper.

This paper evaluates the performance of the aforementioned models: it reviews a lasso and ridge regression performed on all data pooled together, lasso and ridge regressions performed on each subgroup individually, and multiple variations of the joint lasso. These lasso-regressions use tuning parameters, which are set using 10-fold cross-validation. The mean squared error (MSE) is used to compare each method with another: the lower the MSE, the better the model. To test whether the MSE of one model is significantly lower than the MSE of another, a bootstrapping procedure is used to determine the distribution of the MSE's. For an out-of-sample model evaluation, the weekly increase of Covid-19 cases of a week later will be used. Next to evaluating these models, an analysis of the coefficients and policy recommendations are made.

After applying all models, it was found that the pooled lasso regression and one form of the joint lasso regression with cases per head subgroups perform best in terms of the out-of-sample MSE. The group-wise lasso and ridge, as well as some of the joint lassos perform worse than a regression with only an intercept in regards of the out-of-sample MSE. Therefore, these models do not add any value in the prediction of the spread of Covid-19. Additionally, none of the other methods that use subgroups significantly outperform the pooled lasso, which indicates that the defined subgroups might not have differing underlying regression models.

Analysis of the coefficients yields multiple contributing factors to the spread of the virus, which consist of bad weather conditions, socioeconomic factors, and health factors. Two important policy recommendations are made after analysis. Firstly, insurance status of people was found to be a strong contributing factor to the spread of the infection. The rationale behind this is that unsinsured residents are less likely to take a test, in spite of showing symptoms, and as they remain undetected are likely to infect others, hereby resulting in a faster spread. Therefore, the first recommendation is to make tests and related health-care financially accessible to all members of the public, regardless of whether one is insured. Secondly, it was found that the percentage of people living in multi-unit housing is a contributing factor to the spread of the virus. This is a sensible finding, as habitants of these buildings live very close to each other, which increases the chance of the infection being transmitted to another person. Hence, residents of multi-unit housing should be given priority in testing, so infected habitants can be placed quickly into lockdown, thereby mitigating the spread of the virus.

The paper is organized as follows. Section 2 presents the data used in this paper and the cleaning procedures of the data. In Section 3, the formulation of the lasso regressions is described and the methods to evaluate their performance. Section 4 presents the results of this paper. Finally, the main findings are highlighted in section 5.

2 Data

This section discusses the data used and the way subgroups are defined. It elaborates on cleaning the data and the motivation behind the data and the subgroups. All variables are standardized before implementing them in models.

2.1 Socioeconomics, weather conditions and health features

Data on health, socioeconomics, and weather conditions on county-level are provided by John Davis on Kaggle (Davis, 2020).

A lot of the features have missing data. Some of these have a large amount of missing data (>10% missing) and are therefore deleted from the dataset. Others only have a small amount of missing data (<10% missing). The imputation method "classification and regression trees" (CART) is used to fill in these missing spots, using R package "mice"(Breiman et al., 1984). CART are a popular class of machine learning algorithms, because they can deal with multicollinearity and skewed distributions, are robust against outliers, and can fit interactions and nonlinear relations. See Saar-Tsechansky & Provost (2007) for an introductory overview on the idea of using CART methods for imputation of missing data.

Another problem encountered with the data is that weather conditions vary daily, whereas socioeconomics and health data are constant over time. Because the time of infection with Covid-19 is not the same as the time that the infection is diagnosed and reported, weather conditions prior to diagnosis should be considered, instead of weather conditions at the day of diagnosis. Determining the exact time of infection is impossible, due to large variations in the incubation period and time between testing and official diagnosis. The incubation time of Covid-19 varies considerably between patients, with 95% of people developing symptoms between 2.2 and 11.5 days and the median being 5.1 days (Lauer et al., 2020). It should be noted that there is also time between the onset of symptoms and testing, of which the statistics are not known. Time between testing and official diagnosis also varies substantially. Patients who are very ill or high-risk typically receive their results within 24 hours, while others wait for multiple weeks (Cleveland Clinic, 2020), (Petri, 2020). It can also depend on the county, as some testing facilities are overwhelmed by the number of people getting tested, which protracts the waiting time. Therefore the time of infection that is linked to a positive test can have occured between approximately three to thirty days before diagnosis. For this reason, averages of weather conditions will be used in the regression and will be taken for 3 to 9 days before diagnosis, 10 to 16 days, 17 to 23 days, 24 to 30 days, and 3 to 30 days. The dataset includes 12 different weather conditions,

so with five different averages, there are now $12 \cdot 5 = 60$ weather conditions used in the regressions.

Next to these weather conditions, there are also 153 features about socioeconomics and health after cleaning the data. Thus, in total there are 213 features. An overview of these features can be found in Table 11 and Table 12 in the Appendix.

2.2 Covid-19 cases

This paper uses weekly data on Covid-19 cases in the US per county, provided by Johns Hopkins University (Johns Hopkins University, n.d.).

As spreading of the disease varies per county, the increase of cases will be measured in percentage of cases reported the week before, to make comparison between counties fair. Specifically, the weekly increase from May 7th until May 14th is used. This week is chosen for a specific reason: lockdown had then started more than a month ago in all counties, meaning almost all diagnosed people during those dates should have been infected during the lockdown. Hence, the spread of the disease is measured fully during lockdown, which is especially important, as the disease will spread very differently in a state of lockdown than in a state of no lockdown. Therefore, the underlying regression model will also be different.

The US comprises 3,142 counties in total, but not all counties have reported cases and some counties only very little. Counties with only a small amount of reported cases can show extremely large increases percentage-wise, which can erroneously and dramatically influence the model. Therefore only the counties having 100 cases or more on May 7th are used, which are 787 in total.

2.3 Subgroup division

Multiple definitions of subgroups have been evaluated. The two definitions that gerenate results with the highest predictive value are a subgroup division based on the ascending order of the population density of the county, and one based on the ascending order of the amount of reported Covid-19 cases divided by total population of the county, which will be called the cases per head subgroup division. This paper will therefore only review the results on these two definitions of subgroups. Dividing the data into three and five subgroups have both been evaluated. Using five subgroups gives the most promising results, which is why this paper will only review these results in this paper. Ideally more numbers of subgroups would have been evaluated, but because the computations take up a lot of cpu this was not possible due to time and computer restrictions.

The motivation behind using subgroups based on population density is that in densely populated areas the disease might spread differently than in less dense areas, as is suggested in Florida (2020). The motivation behind using subgroups based on cases per head, is that the degree of immunization can affect the spread of the disease (D'Souza & Dowdy, 2020). Table 1 and 2 give an overview of the defined subgroups.

Table 1: Overview of the subgroups based on cases per head (cph)

Group	# of obs	Condition on cph
1	164	cph ≤ 0.0015
2	212	$0.0015 <$ cph ≤ 0.003
3	148	$0.003 <$ cph ≤ 0.005
4	134	$0.005 <$ cph ≤ 0.01
5	129	cph > 0.01

Table 2: Overview of the subgroups based on population density (pd)

Group	# of obs	Condition on pd
1	162	pd ≤ 75
2	187	$75 <$ pd ≤ 200
3	222	$75 <$ pd ≤ 200
4	108	$500 <$ pd ≤ 1100
5	108	pd > 1100

3 Methodology

This section first discusses the models used and their formulations. Secondly, it discusses the evaluation methods. Next, it discusses an analysis of the coefficients. Lastly, it discusses the setting of the tuning parameters and the optimization methods.

3.1 Models used

This paper uses ridge regression and multiple forms of lasso regression, which will be discussed in this section. Ridge and Lasso regression are both simple machine learning techniques that reduce model complexity and prevent over-fitting, which could result from linear regression with high-dimensional data.

The ridge regression as proposed by Hoerl & Kennard (1970) is formulated as:

$$\hat{B} = \underset{B}{\operatorname{argmin}} \frac{1}{n} ||y - XB||_2^2 + \lambda ||B||_q^q, \tag{3.1}$$

where

X = the n x m matrix of observed features

y = the n x 1 vector of observed dependent variables

B = the m x 1 vector of coefficients to be estimated

λ = a prespecified tuning parameter that determines the amount of regularization

$|| \cdot ||_q$ = the ℓ_q norm of its argument

q = 2

n = the number of observations

m = the number of features.

(3.1) is the same as minimizing the sum of squares with constraint $\sum B_m^2 \leq c$, where c is a prespecified postive number. Hence the last term in (3.1) essentially forces the sum of the squared values of the coefficients to be less than c, thus shrinking the coefficients and helping to reduce the model complexity and multi-collinearity. Consequently, ridge regression reduces model variance at the cost of bias.

The lasso-method as proposed by Tibshirani (1996) can also be formulated as (3.1), but uses an ℓ_1 norm in the last term ($q = 1$) instead of an ℓ_2 norm. Just like ridge regression, lasso is the same as minimizing the sum of squares with a constraint added, only the constraint takes the sum of the absolute values of coefficients instead of the squared: $\sum |B_m| \leq c$, where c is again a prespecified postive integer. So the last term in (3.1) essentially forces the sum of the absolute value of the coefficients to be less than c. This affects the way the coefficients are set: whereas ridge shrinks coefficients, lasso forces certain coefficients to be set to zero and therefore effectively chooses a simpler model with less variables.

Equation (3.1) contains the tuning parameter λ. This tuning parameter controls the strength of the ℓ_q norm, thus controlling the amount of shrinkage/parameter selection. When λ is set to zero, (3.1) reduces to the sum of squared errors. Ergo, the coefficients obtained will be the same as the ones obtained with linear regression. When λ increases, more coefficients in (3.1) will be shrunk in the ℓ_2 norm case or set to zero in the ℓ_1 norm case. Consequently, the larger λ is, the more bias and the less variance the model contains. λ has to be set prior to optimization. This paper uses k-fold cross validation to do so, which will be elaborated on in Section 3.3.

The joint lasso method as proposed in Dondelinger & Mukherjee (2018) is an extension on the lasso-method as proposed by Tibshirani (1996). Two versions of it are formulated, which differ in the type of ℓ_q norm used. The first formulation is:

$$\hat{B} = \underset{B=[\beta_1 \ldots \beta_k]}{\operatorname{argmin}} \sum_{k=1}^{K} \left\{ \frac{1}{n_k} ||y_k - X_k \beta_k||_2^2 + \lambda ||\beta_k||_1 + \gamma \sum_{k'>k} \tau_{k,k'} ||\beta_k - \beta_{k'}||_2^2 \right\}, \tag{3.2}$$

and the second formulation is:

$$\hat{B} = \underset{B=[\beta_1 \ldots \beta_k]}{\operatorname{argmin}} \sum_{k=1}^{K} \left\{ \frac{1}{n_k} ||y_k - X_k \beta_k||_2^2 + \lambda ||\beta_k||_1 + \gamma \sum_{k'>k} \tau_{k,k'} ||\beta_k - \beta_{k'}||_1 \right\}, \tag{3.3}$$

where

k	= a subgroup and $k \in \{1, \ldots K\}$
n_k	= the sample size of subgroup k
y_k	= the n_k x 1 vector of dependent variables
X_k	= the n_k x m matrix of features
β_k	= the m x 1 vector of coefficients
B	= $[\beta_1 \ldots \beta_K]$ is an m x K matrix that contains all coefficients
γ	= a tuning parameter
$\tau_{k,k'}$	= a tuning parameter.

Whereas the lasso method uses all observations, the joint lasso uses predefined subgroups of observations that can have differing coefficients, while simultaneously encouraging similarity between subgroup-specific coefficients. The last term in (3.2) and (3.3) is a fusion-type penalty between subgroups. The tuning constant γ determines the amount of similarity encouraged for the model as a whole, whereas the constants

$\tau_{k,k'}$ determine the extent to which similarity is encouraged for specific pairs of subgroups. When γ is set to zero, (3.2) and (3.3) reduce to the classical lasso applied to all subgroups seperately. When the value of γ increases, more similarity is encouraged between subgroups.

The difference between models (3.2) and (3.3), is that (3.2) encourages similarity between subgroup-specific coefficients, whereas (3.3) allows for exact similarity. This is because (3.2) has an ℓ_2 norm in the last term, and (3.3) an ℓ_1 norm in the last term.

This paper evaluates a lasso and ridge regression as defined in (3.1) pooled on all data together, which is called the pooled lasso and pooled ridge. Next to that, it evaluates a lasso and ridge regression applied to every subgroup separately, called the groupwise lasso and groupwise ridge regression. This is done for both the population density subgroups, as well as the cases per head subgroups. It also evaluates the joint lasso as described in (3.3) and (3.2), which uses the same subgroups. Multiple variations on the joint lasso are made, which differ from each other in the way the tuning constants $\tau_{k,k'}$ are calculated. This will be elaborated on in Section 3.4.

3.2 Model evaluation

The general aim of lasso and ridge regressions is to induce sparsity of the model, which lowers variance at the cost of more bias, known as the bias-variance trade-off. Therefore a comparison metric is needed that takes both bias as well as variance into account. The mean squared error (MSE) is a suitable metric for this, as it is composed of two factors: the squared bias and the variance. The MSE is computed as follows:

$$\sum_{i=1}^{N} \frac{(y_i - \hat{y_i})^2}{N}, \tag{3.4}$$

where $\hat{y_i}$ is the fitted value of y_i with the obtained coefficients, y_i is the actual value and N is the total number of observations. The lower the MSE, the better the model is able to predict the increase in COVID-19 cases.

To evaluate the performance of the methods, it is compared to the MSE of a regression with only an intercept. As the lasso and ridge regressions are prone to overfitting, mainly the out-of-sample performance of the methods is reviewed. If a method performs worse than a simple regression with only an intercept, it does not add any value in the prediction of Covid-19 cases.

To determine whether the MSE of one model is significantly lower than the MSE of another, a bootstrapping method has to be performed as the distribution of the MSE is unknown. Technically, this could also be done by cross-validation, but since only 10-fold cross-validation is used due to computational reasons (which will be discussed later in Section 3.4), only ten observations are obtained which is too small for determining the distribution. After performing the bootstrapping procedure, it is necessary to determine whether the MSE's are normally distributed, which is done with the Jarque-Bera (JB) test:

$$JB = \frac{N}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right), \tag{3.5}$$

where N is the total number of observations, S the skewness and K the kurtosis. The JB test determines whether the data have skewness and kurtosis matching a normal distribution. Consequently, if skewness deviates much from zero and/or kurtosis deviates much from three, the null-hypothesis of normality will be rejected. Under the null-hypothesis, the JB statistic asymptotically has a chi-squared distribution with two degrees of freedom.

If the data are indeed normally distributed, the two-sample t statistic can be used to determine whether the MSE's differ significantly from eachother:

$$t = \frac{MSE_1 - MSE_2}{\sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}}, \tag{3.6}$$

where MSE_i is the MSE of model i, s_i the estimated standard devation of MSE_i, and n_i the number of observations of MSE_i.

If the data are not normally distributed, the Mann-Whitney U Test is used, which is a comparison test that does not assume normality. This test compares the number of times a score from one sample is ranked higher than a score from another sample. The test statistic is denoted U and is the smaller of U_1 and U_2:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, \tag{3.7}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2, \tag{3.8}$$

where R_i is the sum of the ranks for model i. The smaller the value of the U statistic, the more likely the null-hypothesis of equal MSE is rejected.

3.3 Interpretation of model

Next to finding the model that performs best in predicting the increase of COVID-19 cases per county, this paper also interprets the obtained coefficients of the best model, excluding the joint lasso regressions. The joint lasso regressions are excluded from this, because the coefficients cannot be analyzed on their P-value.

Generally, lasso regression tends to overfit the model by including coefficients that are not significant (Tibshirani, 1996). It is therefore necessary to check whether the coefficients obtained from the lasso are indeed significantly different from zero. To do so, a linear regression is made with the variables that are selected with the lasso regression. If there are coefficients with P-value > 0.05, the general-to-specific method will be applied. The general-to-specific method first deletes the variable with the largest P-value from the data, and then performs another linear regression with the remaining variables. This process is repeated until all obtained coefficients are significantly different from zero on a 5% significance level.

The coefficients of the remaining variables are then interpreted. Note that the variables have been standardized, as this alters the interpretation of the coefficients. For example, if a coefficient is -0.5, this means that if the corresponding variable is one standard deviation above its mean and all other variables are constant, then the increase in percentage of COVID-19 cases is 0.5 standard deviation lower than normal.

3.4 Setting the tuning parameters

Ideally all tuning parameters are set using some form of cross validation. However, the computation of the tuning parameter $\tau_{k,k'}$ using cross-validation is onerous. Hence, we pre-set the value of $\tau_{k,k'}$, and only set γ and λ according to cross-validation.

3.4.1 Setting $\tau_{k,k'}$

The tuning parameters $\tau_{k,k'}$ control the extent of fusion between specific subgroups. This paper uses both weighted and unweighted fusion. With unweighted fusion, the extent of fusion does not differ between specific subgroups, meaning all $\tau_{k,k'}$ are set to unity. Weighted fusion means the extent of fusion does differ between subgroups, and therefore $\tau_{k,k'}$ has to be specified before regressing the joint lasso. This paper uses two different methods to compute $\tau_{k,k'}$ which are discussed below.

3.4.2 Distance function based on the features

An alternative method as proposed by Dondelinger & Mukherjee (2018) that does not use cross validation, is setting $\tau_{k,k'}$ using some distance function $d(k,k')$ based on the features. One of the methods they use is setting the distance function using the means of the features, according to:

$$d(k,k') = ||\mu_k - \mu_{k'}||_2, \tag{3.9}$$

where μ_k and $\mu_{k'}$ are the sample means of the standardized features in the subgroups k and k' respectively. The parameters are then set according to the following formula:

$$\tau_{k,k'} = 1 - \frac{d(k,k')}{d_{max}}, \tag{3.10}$$

where d_{max} is the largest distance between any pair of subgroups k, k'. By using this formula, we have $0 \leq \tau_{k,k'} \leq 1$.

Note that the method above essentially means there will be more fusion between subgroups that are similar with respect to their features. Thus, this method assumes that similarity between the underlying regression coefficients reflects similarity in its features, which might not necessarily be the case in this COVID-19 application. Moreover, in the applications used in Dondelinger & Mukherjee (2018), joint lasso with weighted $\tau_{k,k'}$ never outperforms the unweighted joint lasso, which might indicate that their method is not proper for setting the tuning constants. Therefore this paper also investigates other methods of setting the tuning parameters that do not use a distance function based on the features, which possibly reflect more accurate values of $\tau_{k,k'}$.

3.4.3 Distance function based on regression coefficients from the subgroup wise lasso

An alternative method to set the tuning parameters is to use the regression coefficients from the subgroup wise lasso as a measure for the distance function. Using these instead of the underlying features, there will

be more fusion between subgroups that have similar coefficients, which is a more sensible way of determining the amount of similarity between the underlying regression model of subgroups. A straightforward method to determine the distance function is to use $d(k, k') = ||\beta_k - \beta_{k'}||_2$, where β_k and $\beta_{k'}$ are the regression coefficients of the features from the subgroup wise lasso in subgroups k and k' respectively. The parameters $\tau_{k,k'}$ are set according to (3.10).

3.4.4 Setting γ and λ

After $\tau_{k,k'}$ is set, the tuning paramaters γ and λ are set by a non-nested k-fold cross validation (CV). A non-nested k-fold cross validation is a re-sampling procedure that splits the data randomly into a training set and a test set. The training set is used to fit the model, and the test set is used to evaluate the model with some form of metric. This is done k times, so for k different training and test sets. The tuning parameters that on average perform best in the k evaluations are then chosen as tuning parameters. This form of cross validation poses the problem that the same data is used to tune model parameters and evaluate model performance. Information can therefore "leak" into the model and overfit the data. A nested CV solves this issue by using an inner loop CV nested in an outer CV. The inner loop is responsible for determining the tuning paramaters, while the outer loop is used for determining the MSE. Hereby, information cannot "leak" into the model anymore and the data will not be overfitted. Because of the benefits of NCV, it was initially chosen as method to evaluate the best values for γ and λ. A drawback of this method is that it is computationally strenuous, and given the small period of time and limited access to computationally fast computers, it was chosen to use non-nested CV instead.

This paper uses 10-fold cross-validation, and the MSE as metric to evaluate the model. It evaluates 47 values of γ and 47 values of λ between zero and one in the cross-validation, thus evaluating $47 \cdot 47 = 2209$ unique combinations of λ and γ. The exact values of λ and γ are found in Table 10 in the Appendix. If after evaluation the optimal value of either λ or γ is the lowest or highest value of the 47 values, then the range of values for λ and γ is broadened to make sure the optimal value is found.

3.5 Optimization

To optimize the pooled lasso and ridge and subgroup lasso and ridge, glmnet software is used in R (Friedman, 2010) . This software contains extremely efficient procedures for fitting the lasso regularization path for linear regression. The joint lasso is computationally much more complicated to optimize than the aforementioned models, and cannot directly be optimized with glmnet. Fortunately, Dondelinger & Mukherjee (2018) have published their R code used for the optimization of the joint lasso as R package fuser on CRAN (https://cran.r-project.org/web/packages/fuser), which is used in this paper as well to optimize the joint lasso.

4 Results

Firstly, the results of the regressions are presented and their performances are evaluated. Secondly, the coefficients are evaluated and policy recommendations are made.

4.1 Performance of models

The performance of the model with only an intercept is evaluated first. Second, the performance of the pooled regressions is analyzed. Then, the regressions that use subgroups based on cases per head are analyzed, and afterwards the regressions that use subgroups based on population density. Lastly, all regressions are compared with each other.

4.1.1 Linear regression with only an intercept

Because the data are standardized, the in-sample MSE of the model with only an intercept is approximately equal to one. The out-of-sample data are standardized with the mean and standard deviation of the in-sample data. Consequently, the out-of-sample MSE of the regression with only an intercept is not equal to one, but 0.504. This indicates that the out-of-sample variance of the cases per county is smaller than in-sample. These results, along with the standard error, can be found in Table 3. The Jarque-Bera test rejects the null-hypothesis of normality. Thus, all statistical comparisons with the MSE of the linear regression with only an intercept are made with the Mann-Whitney U test.

Table 3: Mean squared error of the regression with an intercept only

Model	MSE in-sample	MSE out-of-sample
Intercept only	0.995 (0.169)	0.502 (0.039)

In between brackets () is the standard error of the MSE

4.1.2 Pooled regressions

Table 4 presents the results of the lasso and ridge regressions on the data pooled together. The lasso and ridge regressions both perform significantly better than a regression with only an intercept term, both in- and out-of-sample. In-sample, the ridge seems to perform slightly better than the lasso regression, but out-of-sample it is the opposite: here the lasso performs better. Thus the problem of over-fitting seems to occur more in the ridge regression than in the lasso regression. Note that although the out-of-sample MSE is lower than in-sample, this does not imply that the out-of-sample results are better than the in-sample results. The reason the out-of-sample MSE is lower is because the standard deviation of the out-of-sample data is lower, which is why the out-of-sample MSE and in-sample MSE cannot directly be compared with each other.

Table 4: Tuning parameter and in-sample mean squared error for the pooled lasso and ridge regression

Method	λ_1	MSE in-sample	MSE out-of-sample
lasso	0.02	0.788* (0.144)	0.442* (0.033)
ridge	0.8	0.773* (0.137)	0.460* (0.034)

In between brackets () is the standard error of the MSE

* Model has lower MSE than the intercept model at the 5% significance level

4.1.3 Cases per head subgroups

Table 5 presents the results of the groupwise cases per head lasso and ridge regressions. The MSE of the groupwise lasso and ridge is much higher than the MSE of the pooled versions. This can be explained by the fact that the sample size per regression is much smaller, which results in higher variance and thus higher MSE. This is observed in the tuning constants as well: the λ's of the groupwise regressions are much higher than the λ's of the pooled regressions, which shows that in the groupwise regressions lasso and ridge set more coefficients to zero / shrink coefficients more in order to reduce the variance at the cost of more bias.

In-sample, the groupwise lasso does not perform significantly better than a simple regression with only an intercept, which shows the explanatory power of the groupwise lasso is very poor. The groupwise ridge, however, does perform significantly better in-sample than a simple regression with only an intercept, but not out-of-sample. In fact, both the groupwise ridge as well as the lasso perform significantly worse out-of-sample as compared to a regression with only an intercept.

Table 5: Tuning parameters for each subgroup and mean squared error with subgroups based on cases per head

Method	λ_1	λ_2	λ_3	λ_4	λ_5	MSE in-sample	MSE out-of-sample
lasso	0.2	0.4	0.1	0.1	0.2	0.989 (0.124)	0.638 (0.052)
ridge	1	4	4	3	30	0.916* (0.116)	0.604 (0.054)

In between brackets () is the standard error of the MSE

* Model has lower MSE than the intercept model at the 5% significance level

Table 6 presents the results of the joint lasso method. In-sample, both the joint lasso with ℓ_1 norm as well as ℓ_2 norm perform significantly better than a regression with only an intercept. The joint lasso with ℓ_1 norm performs significantly better than the ℓ_2 norm. However, out-of-sample only the ℓ_2 norm performs better than a regression with only an intercept, for all values of τ. The ℓ_1 is the worst model up until now in terms of out-of-sample performance for all values of τ.

The ℓ_2 norm with τ set with the coefficients method performs best out-of-sample as well as in-sample. This result can also be seen in the value of γ: The value of γ is relatively higher with the coefficients method

for τ as compared to the other method. This implies that the subgroups are being forced to have the same coefficients more than with the other methods, indicating that the inter-subgroup relations are better defined with the coefficients method.

Table 6: Tuning parameters and mean squared error for the joint lasso methods with subgroups based on cases per head

Method	γ	λ	τ	MSE in-sample	MSE out-of-sample
ℓ_1	6.00E-02	3.00E-05	Unweighted	0.622* (0.101)	0.796 (0.048)
ℓ_1	1.00E-01	6.00E-05	Coef method	0.617* (0.099)	0.810 (0.044)
ℓ_1	9.00E-02	8.00E-05	Mean method	0.633* (0.110)	0.791 (0.046)
ℓ_2	4.00E-06	3.00E-07	Unweighted	0.866* (0.165)	0.445* (0.041)
ℓ_2	4.00E-05	3.00E-07	Coef method	0.842* (0.160)	0.431* (0.037)
ℓ_2	6.00E-06	3.00E-07	Mean method	0.870* (0.163)	0.448* (0.039)

In between brackets () is the standard error of the MSE

* Model has lower MSE than the intercept model at the 5% significance level

4.1.4 Population density subgroups

Table 7 presents the results of the groupwise population density lasso and ridge regression. Just like in the cases per head groups, the MSE of the groupwise lasso and ridge is much higher than the MSE of the pooled versions. Again, this is explained by the increase in variance because of the smaller group sizes, which is also observed in the large tuning constants.

In-sample, both the groupwise lasso and ridge perform significantly better than a simple regression with only an intercept. Out-of-sample, however, both perform significantly worse than a simple regression with only an intercept, similar to the results of the groupwise lasso with cases per head subgroups.

Table 7: Tuning parameters for each subgroup and mean squared error with subgroups based on population density

Method	λ_1	λ_2	λ_3	λ_4	λ_5	MSE in-sample	MSE out-of-sample
lasso	0.1	0.09	0.05	0.05	0.05	0.924* (0.123)	0.566 (0.122)
ridge	20	20	3	3	3	0.931* (0.048)	0.558 (0.047)

In between brackets () is the standard error of the MSE

* Model has lower MSE than the intercept model at the 5% significance level

Table 8 presents the results of the joint lasso method with population density subgroups. In-sample, both the joint lasso with ℓ_1 norm as well as ℓ_2 norm perform significantly better than a regression with only an intercept. The joint lasso with ℓ_1 norm performs significantly better than the ℓ_2 norm. However, out-of-

sample both methods perform worse than a regression with only an intercept, for all values of τ. Just like in the cases per head subgroups, the ℓ_1 norm performs significantly better than the ℓ_2 norm out-of-sample. On average, the coefficients method for setting τ yields the lowest MSE and thus performs best. Again, this can also be seen by the higher value of γ when τ is set with the coefficients method.

Table 8: Tuning parameters and mean squared error for the joint lasso methods with subgroups based on population density

Method	γ	λ	τ	MSE in-sample	MSE out-of-sample
ℓ_1	8.00E-02	4.00E-05	Unweighted	0.634* (0.091)	0.825 (0.041)
ℓ_1	1.00E-01	6.00E-05	Coef method	0.634* (0.092)	0.815 (0.045)
ℓ_1	8.00E-02	4.00E-05	Mean method	0.634* (0.097)	0.825 (0.043)
ℓ_2	3.00E-06	4.00E-07	Unweighted	0.918* (0.159)	0.517 (0.038)
ℓ_2	5.00E-05	4.00E-07	Coef method	0.896* (0.152)	0.509 (0.034)
ℓ_2	9.00E-06	4.00E-07	Mean method	0.904* (0.150)	0.509 (0.037)

In between brackets () is the standard error of the MSE

* Model has lower MSE than the intercept model at the 5% significance level

4.1.5 Comparison of the models

As lasso and ridge regressions are prone to overfitting, only the out-of-sample results of the models are compared. Generally, both the pooled lasso, as well as the joint lasso with ℓ_2 norm with cases per head subgroups perform the best in terms of their MSE. Ultimately, the joint lasso with ℓ_2 norm, τ set with the coefficients method and cases per head subgroups yields the lowest MSE of all models. All groupwise regressions and joint lasso with ℓ_1 perform worse than a simple regression with intercept. Moreover, the joint lasso with ℓ_2 norm never *significantly* outperforms the pooled lasso. This indicates that the defined subgroups (both based on population density as well as cases per head) might not have a different underlying regression model, especially because the joint lassos do perform better than the pooled regressions in the biomedical applications in Dondelinger & Mukherjee (2018).

Although some models do perform significantly better than the regression with only an intercept, the difference in MSE is small. The lowest obtained out-of-sample MSE is only 14% lower than that of the regression with only an intercept. This means that even the best model cannot predict the increase in cases of the virus with a high accuracy.

4.2 Interpretation of coefficients

Excluding the joint lasso models, the regression with the lowest out-of-sample MSE is the pooled lasso. Thus, the coefficients of this model are examined in this section. Note that all data are standardized, which affects

the way the coefficients are interpreted. This has been discussed in Section 3.3.

The pooled lasso regression selects 53 variables to be in the model. After performing a linear regression on these 53 variables, 39 variables are insignificant at the 5% significance level. After performing the general-to-specific method, 14 variables remain that have a significant factor in predicting the spread. The coefficients and statistical values are found in Table 9.

Of the 14 coefficients seven are weather variables. Note that six of the seven variables are either from the period 17-23 days or 24-30 days prior diagnosis, which could indicate that infection takes place on average between 17 to 30 days before diagnosis. A contradictory finding in the weather coefficients is that snow, rain and thunder 17-23 days before diagnosis have positive coefficients, whereas snow and precipitation 24-30 days before diagnosis have negative coefficients. Thus, "bad" weather, eg. rain, snow or thunder, from 24-30 days prior generally increases the spread of Covid-19, whereas "bad" weather from 17-23 days prior generally decreases the spread of Covid-19. Adding to the confusion is the fact that rain from 10-16 days prior again increases the spread of Covid-19, as this coefficient is positive. This makes it difficult to determine the influence of the weather on the spread of Covid-19, and prohibits making a definite conclusion.

The other seven coefficients are socioeconomic and health variables. The percentage uninsured coefficient, which indicates the percentage of the population that have no health insurance, is -0.29. This means that the more people that are uninsured in a county, the less the number of confirmed Covid-19 cases increases. A possible explanation for this phenomenon is that uninsured citizens will be more reluctant to get tested, as it will cost them relatively more money. This means that the amount of people that are uninsured only lowers the amount of detected cases, not the real amount of cases. This coefficient has the largest absolute value of the 14 coefficients, demonstrating the significance of its influence.

The percentage multi-unit housing coefficient is 0.21, which means that the more people live in multi-unit housing, such as apartment buildings, the more the amount of Covid-19 cases increases. This can be explained by the fact that habitants of these buildings live very closely to each other, so when one of the habitants gets infected it is more likely they will infect someone else, thus worsening the spread of the virus. This is also supported by empirical evidence, as there are numerous counts of severe outbreaks in apartment buildings (Khan, 2020).

The percentage physically inactive coefficient is 0.16, indicating that the more people are physically inactive, the more the amount of Covid-19 cases increases. As a general rule, physical activity ameliorates the immune system (Shephard & Shek, 1994). Consequently, the immune system of physically inactive people is inferior at combatting infections such as Covid-19, which explains the relationship between the amount of Covid-19 cases and the percentage physically inactive people.

The percentage with annual mammogram coefficient is -0.094, meaning the more women have an annual mammogram, the less the amount of confirmed Covid-19 cases increases. This seems like a strange relation, and could simply be coincidental. After all, correlation does not imply causation. Furthermore, it is the smallest coefficient in absolute value of all 14 coefficients, which further bolsters suspicion that it is not a

true factor in the spread of the virus.

The income ratio coefficient is -0.16, meaning the higher the income ratio, the less Covid-19 spreads. The income ratio is the ratio between the amount of income at the 80th percentile, and the amount of income at the 20th percentile. Thus, the higher the income ratio, the bigger the difference is in income between household in a county, and the less the virus spreads. This is an interesting finding, but no straightforward explanation can be made for it. It could be coincidental, just like the mammogram coefficient seems to be.

The percentage no highschool diploma coefficient is 0.225, meaning the more people did not finish high school, the more the amount of confirmed Covid-19 cases increase. An explanation for this could be that people without a high school diploma generally have a lower income, and will often not have desk jobs. Therefore in a state of lockdown, these people cannot work as easily from home as other people can, and thus go to work since they need the money.

The average traffic volume per meter of major roadways coefficient is -0.114, meaning the more traffic there is on major roadways, the less the amount of confirmed Covid-19 cases increase. At first this seems like a strange relationship. However, it can be an indirect relationship: a lot of traffic on roadways could indicate a lot of people take the car, and subsequently not much people take public transportation. Then this would imply that the more people take public transportation, the more the amount of confirmed Covid-19 cases increase, which makes sense as there is more interaction between people when they take public transportation instead of their own car. This indirect relationship is possible, since there is no variable included about public transportation. It could however also be coincidental, just like it seems to be with the percentage with annual mammogram coefficient.

4.2.1 Policy recommendations

Based on the results found using the pooled lasso model and analysing the various possible contributing factors, two clear policy recommendations can be made. First of all, it was found that the percentage uninsured people has a strong negative influence on the spread of the disease. It is assumed that uninsured people are less likely to take a test as a result of the high medical costs. If these symptomatic people do not isolate themselves and are in fact infected with Covid-19, they can infect other people which will exacerbate the spread. Further research should be conducted to determine if this is indeed the case. If so, Covid-19 tests and related healthcare should be made financially more accessible to these people.

Second of all, there was a positive relationship found between the number of people living in multi-unit housing and the increase of the spread of Covid-19. This indicates that the virus can spread more easily between residents of such buildings. The second policy recommendation is therefore to give residents of multi-unit housing priority in testing for Covid-19, as an undetected infected habitant could cause a severe outbreak in such apartment buildings. When one person in such a residency has tested positive for the virus, it could also be intrumental to test everyone in the residency that shows Covid-19 symptoms. These steps could prevent an outbreak in such a building, thereby forestalling the need to take drastic measures such as

Table 9: Coefficients obtained from the pooled lasso after the general-to-specific method has been applied

Variable	Estimate	Std. Error	P-value
income ratio	-0.163	0.058	0.005
percentage multi-unit housing	0.214	0.068	0.002
percentage with annual mammogram	-0.094	0.045	0.039
percentage uninsured	-0.285	0.110	0.010
percentage physically inactive	0.155	0.062	0.013
percentage no highschool diploma	0.225	0.092	0.015
average traffic volume per meter of major roadways	-0.114	0.051	0.025
rain 10-16 days prior	0.146	0.067	0.030
snow 17-23 days prior	-0.180	0.060	0.003
rain 17-23 days prior	-0.115	0.057	0.044
thunder 17-23 days prior	-0.110	0.055	0.045
visibility 24-30 days prior	0.148	0.047	0.002
precipitation 24-30 days prior	0.115	0.039	0.003
snow 24-30 days prior	0.205	0.074	0.006

an entire appartment lockdown.

5 Conclusion

In this study it was investigated which lasso/ridge model performs best in predicting the spread of Covid-19 and which parameters are the strongest contributors. The lasso regression pooled on all data together is tough to beat in terms of out-of-sample mean squared error. Only one of the 12 investigated versions of the joint lasso (insignificantly) outperforms the pooled lasso regression. The groupwise lasso and ridge perform significantly worse than the pooled regression. Together with the fact that none of joint lasso methods significantly outperform the pooled lasso, this indicates that the defined subgroups might not have a different underlying regression model, especially since the joint lasso *does* outperform the pooled lasso in some biomedical applications in Dondelinger & Mukherjee (2018). Additionally, the lowest obtained out-of-sample MSE from our models is still only 14% lower than that of a regression with only an intercept, which is not much of an improvement. It is therefore interesting for future research to investigate whether there are models that can predict the spread of Covid-19 better, such as time series models.

The coefficients obtained from the pooled lasso have been analyzed. According to this analysis, weather conditions, socioeconomics and health all seem to have influence on the spread of the virus. From the analysis of the coefficients, two policy recommendations are made. Firstly, it seems that uninsured people are less likely

to take a Covid-19 test. It is necessary to investigate if the costs of the tests are an obstacle for people with no insurance to take tests. If this is indeed the case, it is necessary to make the test financially more accesible, as undetected infections could exacerbate the spread. Secondly, residents in multi-unit housing should be given priority in taking tests, as there is increased hazard of an outbreak in such buildings. Another finding was that bad weather conditions seem to influence the spread of the virus, but because of contradictory coefficients, no conclusion can be made on the specific relationship it has with regards to the spread.

References

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. (1984). *Classification and regression trees.* Taylor & Francis.

Cleveland Clinic. (2020). *Frequently asked questions about coronavirus disease 2019 (covid-19).* Retrieved from https://newsroom.clevelandclinic.org/2020/03/18/frequently-asked-questions-about-coronavirus-disease-2019-covid-19/

Davis, J. (2020). *Us counties: Covid19 + weather + socio/health data.* Retrieved from https://www.kaggle.com/johnjdavisiv/us-counties-covid19-weather-sociohealth-data?select=us_county_sociohealth_data.csv

Dondelinger, F., & Mukherjee, S. (2018). The joint lasso: high-dimensional regression for group structured data. *Biostatistics*, *21*(2), 219–235.

D'Souza, G., & Dowdy, D. (2020). *What is herd immunity and how can we achieve it with covid-19?* Retrieved from https://www.jhsph.edu/covid-19/articles/achieving-herd-immunity-with-covid19.html

Florida, R. (2020). *The geography of coronavirus.* Retrieved from https://www.bloomberg.com/news/articles/2020-04-03/what-we-know-about-density-and-covid-19-s-spread

Friedman, J. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1 – 22.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression:biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Johns Hopkins University. (n.d.). *Covid-19 united states cases by county.* Retrieved from https://coronavirus.jhu.edu/us-map

Khan, S. (2020). *Coronavirus: people in tall buildings may be more at risk – here's how to stay safe.* Retrieved from https://theconversation.com/coronavirus-people-in-tall-buildings-may-be-more-at-risk-heres-how-to-stay-safe-135845

Lauer, S. A., Grantz, K. H., Bi, Q., & Jones, F. K. (2020). The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*.

Petri, A. E. (2020). *The experience of getting tested for coronavirus.* Retrieved from https://www.nytimes.com/article/test-for-coronavirus.html

Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. *Journal of Machine Learning Research*, *8*, 1625–1657.

Shephard, R. J., & Shek, P. N. (1994). Potential impact of physical activity and sport on the immune system - a brief review. *British Journal of Sports Medicine*, *28*, 247–255.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*(1), 267–288.

World Health Organization. (2020). *Timeline of covid-19.* Retrieved from `https://www.who.int/news-room/detail/27-04-2020-who-timeline\OT1\textemdashcovid-19`

6 Appendix

Table 10: Values used for λ and γ in initial cross-validation.

Values of both λ and γ				
0	1.00E-05	2.00E-04	3.00E-03	4.00E-02
1.00E-06	2.00E-05	3.00E-04	4.00E-03	5.00E-02
2.00E-06	3.00E-05	4.00E-04	5.00E-03	6.00E-02
3.00E-06	4.00E-05	5.00E-04	6.00E-03	7.00E-02
4.00E-06	5.00E-05	6.00E-04	7.00E-03	8.00E-02
5.00E-06	6.00E-05	7.00E-04	8.00E-03	9.00E-02
6.00E-06	7.00E-05	8.00E-04	9.00E-03	1.00E-01
7.00E-06	8.00E-05	9.00E-04	1.00E-02	
8.00E-06	9.00E-05	1.00E-03	2.00E-02	
9.00E-06	1.00E-04	2.00E-03	3.00E-02	

If the optimal value found was 0 or 1.00E-06, the range was broadened with values between 0 and 1.00E-06.

If the optimal value found was 1.00E-01, the range was broadened with values above 1.00E-01

Table 11: Overview of used variables in the lasso and ridge regressions (part 1 of 2)

Part 1 of the variables used in the regressions		
mean_temp_3-9	percent_low_birthweight	num_households_with_severe_cost_burden
min_temp_3-9	percent_smokers	percent_severe_housing_cost_burden
max_temp_3-9	percent_adults_with_obesity	percent_less_than_18_years_of_age
dewpoint_3-9	food_environment_index	percent_65_and_over
station_pressure_3-9	percent_physically_inactive	num_black
visibility_3-9	percent_with_access_to_exercise_opportunities	percent_black
wind_speed_3-9	percent_excessive_drinking	num_american_indian_alaska_native
max_wind_speed_3-9	num_alcohol_impaired_driving_deaths	percent_american_indian_alaska_native
precipitation_3-9	num_driving_deaths	num_asian
fog_3-9	percent_driving_deaths_with_alcohol_involvement	percent_asian
rain_3-9	num_chlamydia_cases	num_native_hawaiian_other_pacific_islander
snow_3-9	chlamydia_rate	percent_native_hawaiian_other_pacific_islander
mean_temp_10-16	teen_birth_rate	num_hispanic
min_temp_10-16	num_uninsured	percent_hispanic
max_temp_10-16	percent_uninsured	num_non_hispanic_white
dewpoint_10-16	num_primary_care_physicians	percent_non_hispanic_white
station_pressure_10-16	primary_care_physicians_rate	num_not_proficient_in_english
visibility_10-16	num_dentists	percent_not_proficient_in_english
wind_speed_10-16	dentist_rate	percent_female
max_wind_speed_10-16	num_mental_health_providers	num_rural
precipitation_10-16	mental_health_provider_rate	percent_rural
fog_10-16	preventable_hospitalization_rate	num_housing_units
rain_10-16	percent_with_annual_mammogram	num_households_CDC
snow_10-16	percent_vaccinated	num_below_poverty
thunder_10-16	high_school_graduation_rate	num_unemployed_CDC
mean_temp_17-23	num_some_college	per_capita_income
min_temp_17-23	population	num_no_highschool_diploma
max_temp_17-23	percent_some_college	num_age_65_and_older
dewpoint_17-23	num_unemployed_CHR	num_age_17_and_younger
station_pressure_17-23	labor_force	num_disabled
visibility_17-23	percent_unemployed_CHR	num_single_parent_households_CDC
wind_speed_17-23	percent_children_in_poverty	num_minorities
max_wind_speed_17-23	eightieth_percentile_income	num_multi_unit_housing
precipitation_17-23	twentieth_percentile_income	num_mobile_homes
fog_17-23	income_ratio	num_overcrowding
rain_17-23	num_single_parent_households_CHR	num_households_with_no_vehicle
snow_17-23	percent_single_parent_households_CHR	num_institutionalized_in_group_quarters

Table 12: Overview of used variables in the lasso and ridge regressions (part 2 of 2)

Part 2 of the variables used in the regressions		
thunder_17-23	num_associations	percent_below_poverty
mean_temp_24-30	social_association_rate	percent_unemployed_CDC
min_temp_24-30	annual_average_violent_crimes	percent_no_highschool_diploma
max_temp_24-30	violent_crime_rate	percent_age_65_and_older
dewpoint_24-30	num_injury_deaths	percent_age_17_and_younger
station_pressure_24-30	injury_death_rate	percent_disabled
visibility_24-30	average_daily_pm2_5	percent_single_parent_households_CDC
wind_speed_24-30	presence_of_water_violation	percent_minorities
max_wind_speed_24-30	percent_severe_housing_problems	percent_limited_english_abilities
precipitation_24-30	severe_housing_cost_burden	percent_multi_unit_housing
fog_24-30	overcrowding	percent_mobile_homes
rain_24-30	inadequate_facilities	percent_overcrowding
snow_24-30	percent_drive_alone_to_work	percent_no_vehicle
thunder_24-30	percent_long_commute_drives_alone	percent_institutionalized_in_group_quarters
mean_temp_3-30	life_expectancy	percentile_rank_below_poverty
min_temp_3-30	num_deaths_2	percentile_rank_unemployed
max_temp_3-30	age_adjusted_death_rate	percentile_rank_per_capita_income
dewpoint_3-30	percent_frequent_physical_distress	percentile_rank_no_highschool_diploma
station_pressure_3-30	percent_frequent_mental_distress	percentile_rank_socioeconomic_theme
visibility_3-30	percent_adults_with_diabetes	percentile_rank_age_65_and_older
wind_speed_3-30	num_food_insecure	percentile_rank_age_17_and_younger
max_wind_speed_3-30	percent_food_insecure	percentile_rank_disabled
precipitation_3-30	num_limited_access	percentile_rank_single_parent_households
fog_3-30	percent_limited_access_to_healthy_foods	percentile_rank_household_comp_disability_theme
rain_3-30	percent_insufficient_sleep	percentile_rank_minorities
snow_3-30	percent_uninsured_2	percentile_rank_limited_english_abilities
thunder_3-30	num_uninsured_3	percentile_rank_minority_status_and_language_theme
total_population	percent_uninsured_3	percentile_rank_multi_unit_housing
area_sqmi	other_primary_care_provider_rate	percentile_rank_mobile_homes
population_density_per_sqmi	median_household_income	percentile_rank_overcrowding
years_of_potential_life_lost_rate	percent_enrolled_in_free_or_reduced_lunch	percentile_rank_no_vehicle
percent_fair_or_poor_health	average_traffic_volume_per_meter_of_major_roadways	percentile_rank_institutionalized_in_group_quarters
average_number_of_physically_unhealthy_days	num_homeowners	percentile_rank_housing_and_transportation
average_number_of_mentally_unhealthy_days	percent_homeowners	percentile_rank_social_vulnerability