Lead Contamination Analysis in Flint City Based on Quasi-Poisson Regression

1 Summary

This paper is about the Flint water crisis. The main goal is to find the main factor that affect the lead level in water. By using the Quasi-Poisson regression I find out the main is the material of service line. Both copper, lead and galvanized line will affect the lead level.

2 Background

The Flint water crisis is a drinking water contamination issue in Flint, Michigan State, which started in April 2014. After Flint changed its water source form treated Detroit Water and Sewerage Department water (which was sourced from Lake Huron and Detroit River) to Flint River, its drinking water had a series of problems that culminated with lead contamination, creation a serious public health danger. From several researches, high level of lead in drinking water may be caused by the aging service pipe. Because of the government failed to apply corrosion inhibitors, the corrosive Flint River water caused lead from aging pipes, which may contain high level of lead, to leach into the water supply system. As result thousands of Flint citizens are suffer from the poisoned drinking water. And they may experience a range of serious health problems. According to the blood tests performed in 2013 and 2015, the percentage of Flint children with elevated blood lead levels risen from about 2.5% to 5%.

In this paper I build a Poisson Regression model to find out a reasonable way to explain the change in level of lead in Flint drinking water. By knowing the main cause factor(s) the government can deal with this crisis more efficiently and it is also helpful for avoiding such problem in future.

3 Data Analysis

The data set I used in my model is the combination of four rounds surveys about the lead concentration in the water supply, which are collected and posted by Michigan government through February to April in 2016
(source: https://www.michigan.gov/flintwater).

3.1 Raw Data instruction

Number of sample: 2533

Variables list:

Variable Name	Meaning	Variable type

Data	Date of sampling	date
Result_pb	Level of lead in water	numeric
Result_cu	Level of copper in water	Numeric
Add	Street of sample	Character
City	City of sample	Character
Zip	Zip code of sample	Character
Pipe_material	Service line material	character
Sample_loc	Sample location	Character
Group num	Group number	Character

Range of data collected:
Flint City (zip code: 48502 48503 48504 48505 48506 48507 48522 48532), shows as figure 1.

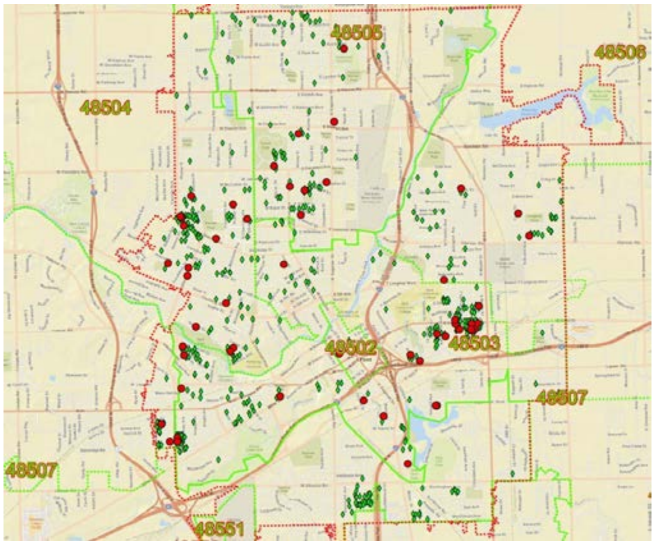

Figure 1

3.2 Variable selection

Since the data are collected in a relative short time interval (about 3 month) I don’t find any obvious tendency in lead level against time. From the *matplot* shows as figure2, except few point changed a lot with time, most of point don’t. So I decide to drop the date.

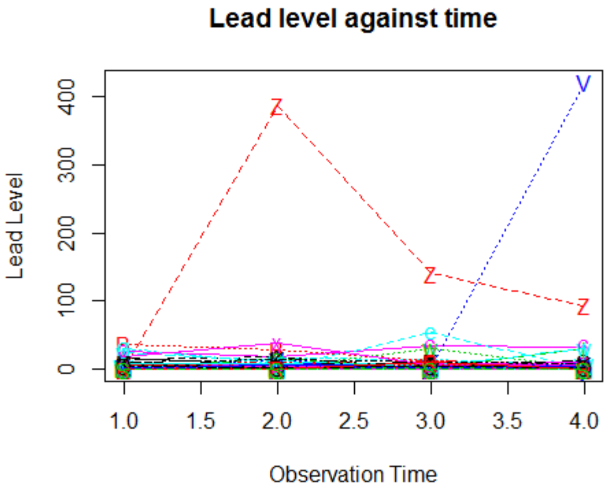

Figure 2

As for street, zip code and city, all of them represent the location information. Since all the data are collected in Flint city, so the city variable is useless here. And for "Address", the length of street a too vary. Some streets, like "Alexander Street", are very short, and others, like "Miller Road", are very long, it almost cross the whole Flint city. What's more, some of them are crossed with several streets. So it is not wise to choose "Address" as the location factor. So at last I chose the "zip code", which separate the Flint area basically equally, and each area is independent with others

And the group number, without further information I don't know the meaning of it. So I decide to drop it.

At last, the final variable list contains 5 variables, which are level of lead in water, level of copper in water, zip code, service line material, sample location.

4 Modeling

4.1 model selection and data transformation

According to Environmental Protection Agency (EPA), the standard for lead level in water is less than 15ppb, and the high level lead will be very harmful. This paper is more interested in whether the lead level is over the standard. So first I transfer the original data of lead level into binary response (Resp), the Resp equals to 1 if the lead level is larger than 15bbp, equals to 0 otherwise.

As we can see the raw data set is quite large (2533 observations) and only 210 samples are polluted by lead, which means the "success" probability is close to 8.29%, is quite small. So now the case has large number of trails and small "success" probability. And in this situation Poisson is a good approximation of Binomial distributed data. What's more if we try the linear model we have to consider the extreme values of the lead level, but without extra information we cannot decide whether the extreme values are input mistake, which we can treat them as outliers and delete them, or it may be the true value, which cannot be simply deleted. But for Poisson distribution all we have to consider is whether the value is larger than 15 or not, so we can avoid the effect of extreme value in linear regression. So I decide to use Poisson Regression in this paper.

Before performing the Poisson regression I need to modify the raw data. I grouping the data into 240 groups based on the three categorical data, “zip codes”, “service line material” and “sample locations”. And then calculate the number of sample point that lead level is larger than 15bbp of each group. I set this count number as response variable denoted as y.

So now we have the data set like this:

```
 y  Cu_leve   zip   material location

 01 45.00000 48502   Copper BATHROOM
18 50.90909 48503   Copper BATHROOM
28 58.57471 48503   Copper KITCHEN.
```

4.2 Fit the model

In this paper I let the “Number of sample point with lead level greater than 15 ppb” ($Y_i = \{y_{i1}, y_{i2}, \dots , y_{in}\}^T$) be response variable, and I use $\log(Y)$ as link function. Then the Poisson regression model is expressed as

$$\log(Y_i) = X_i^T \beta$$

In this case I denote copper level, zip code, service pipe line material and sample location as $X_{copper}, X_{zip}, X_{material}, X_{location}$ respectively, so the model becomes

$$\log(Y_i) = \beta_0 + \beta_1 * X_{copper} + \beta_2 * \ X_{zip} + \beta_3 * X_{material}, + \beta_4 * X_{location}$$

By using *glm* packages in R, I got the following results.

```
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)          -3.078e-01  1.297e+00  -0.237  0.81242    
result_cu            -2.416e-04  1.766e-04  -1.368  0.17138    
zip48503              3.003e+00  1.011e+00   2.969  0.00299 ** 
zip48504              2.571e+00  1.014e+00   2.535  0.01125 *  
zip48505              1.693e+00  1.026e+00   1.651  0.09882 .  
zip48506              8.174e-01  1.055e+00   0.775  0.43831    
zip48507              9.241e-01  1.054e+00   0.877  0.38038    
zip48522             -1.784e+01  3.468e+03  -0.005  0.99589    
zip48532              2.724e-01  1.101e+00   0.247  0.80462    
materialGalvanized   -9.978e-01  1.780e-01  -5.606 2.07e-08 ***
materialLead         -1.398e+00  2.070e-01  -6.755 1.43e-11 ***
materialOther        -3.092e+00  7.184e-01  -4.304 1.67e-05 ***
materialPlastic      -1.850e+01  1.340e+03  -0.014  0.98899    
materialUnknown      -2.387e+00  3.298e-01  -7.238 4.55e-13 ***
```

```
locationBATHROOM     -2.042e-01  8.277e-01  -0.247  0.80513
locationKITCHEN       8.486e-01  8.172e-01   1.038  0.29906
locationLAUNDRY SINK -2.926e-01  1.171e+00  -0.250  0.80274
locationUNKOWN       -5.908e-01  8.341e-01  -0.708  0.47879
---
```

4.2 Overdispersion test

Overdispersion is the presence of greater variability (statistical dispersion) in a data set than would be expected based on a given statistical model, it only happened in GLM. If overdispersion is present in the dataset, it may cause incorrect standard error and selection of overly complex models

The possibility for overdispersion exists because the commonly used distributions specify particular relationships between the variance and the mean. The most popular method for adjusting for overdispersion is quasilikelihood approach.

For Poisson regression the generalized model is

$$E(Y_i) = \mu_i \quad and \quad Var(Y_i) = \Phi\mu_i$$

The mean and variance of response variable should be equal, which means Φ should equal to 1. In this data set the mean of Y is 0.875 and the variance is 11.1057, obviously there are not equal, so it may have overdispersion, to confirm that I use the hypothesis test with $H_0: \Phi = 1 \;\; vs \;\; H_1: \Phi \neq 1$, if the p-value is less than 0.05, we can conclude to reject the H_0which means the model has overdispersion. By using the r function *dispersiontest()* I got the p-value equal to 0.0216, less than 0.05, which shows that I need to modify my previous model by using quasilikelihood approach.

By using *glm* packages in R with "quasipoisson" family, I got following result:

```
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -3.078e-01  1.718e+00  -0.179  0.85851    
result_cu            -2.416e-04  2.340e-04  -1.033  0.30649    
zip48503              3.003e+00  1.340e+00   2.242  0.02920 *  
zip48504              2.571e+00  1.344e+00   1.914  0.06108 .  
zip48505              1.693e+00  1.359e+00   1.246  0.21821    
zip48506              8.174e-01  1.397e+00   0.585  0.56096    
zip48507              9.241e-01  1.396e+00   0.662  0.51070    
zip48522             -1.784e+01  4.594e+03  -0.004  0.99692    
zip48532              2.724e-01  1.459e+00   0.187  0.85257    
materialGalvanized   -9.978e-01  2.358e-01  -4.232 9.23e-05 ***
materialLead         -1.398e+00  2.742e-01  -5.099 4.67e-06 ***
materialOther        -3.092e+00  9.516e-01  -3.250  0.00201 ** 
materialPlastic      -1.850e+01  1.775e+03  -0.010  0.99172    
materialUnknown      -2.387e+00  4.368e-01  -5.464 1.27e-06 ***
locationBATHROOM     -2.042e-01  1.096e+00  -0.186  0.85296    
locationKITCHEN       8.486e-01  1.082e+00   0.784  0.43656    
locationLAUNDRY SINK -2.926e-01  1.552e+00  -0.189  0.85114    
locationUNKOWN       -5.908e-01  1.105e+00  -0.535  0.59511    
---
```

By comparing the two outputs above, we can see that the significance of parameters don’t change much. Only two zip codes are significant, all others are not. Since the lead pollution is cover the whole Flint city, so there is no surprise that the lead level don’t change a lot between different zip code areas.

For the service line material, except plastic all other materials are highly significant. It indicates that the material of service line may be the main factor that affect the level of lead in water, which confirms the conclusion from the public research.

And all the “Sample locations” are not significant. It is reasonable, because the difference location in the same house or even in the same area will share the same service line. So the sample location shouldn’t be the main factor affect the lead level.

5 Diagnostic

5.1 Overall goodness-of-fit test

From the modeling part, I observed that only “Material” terms are highly significant with the lead level in water. But still we need to do the overall goodness-of-fit test.

$$H_{0:}\ \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

I build another model that only contains an intercept term, and compare it with Quasi-Poisson model by using likelihood ratio test. The results shows as follow:

```
Analysis of Deviance Table
Model_intercept : y~1
Model _Quasi-Poisson: y~ Copper+Zip+Material+Location
   Resid. Df  Resid. Dev Df   Deviance    Pr(>Chi)
1     239       957
2     53         93.826     186     863.175  < 2.2e-16
```

The results show that the likelihood ratio has a p-value < 2.2e-16. So we have a very strong evidence to reject H_0, which means the Quasi-Poisson model is good fitted.

5.2 Residual Checking

To further testify whether the model is appropriate, I also do the residual plot and Quantile-Quantile plot of the model.

The first is the residual plot, since $\log(Y_i) = X_i^T\beta$, so I do the plot of $\log(\hat{Y})$ with residuals. The result shows as figure 3.

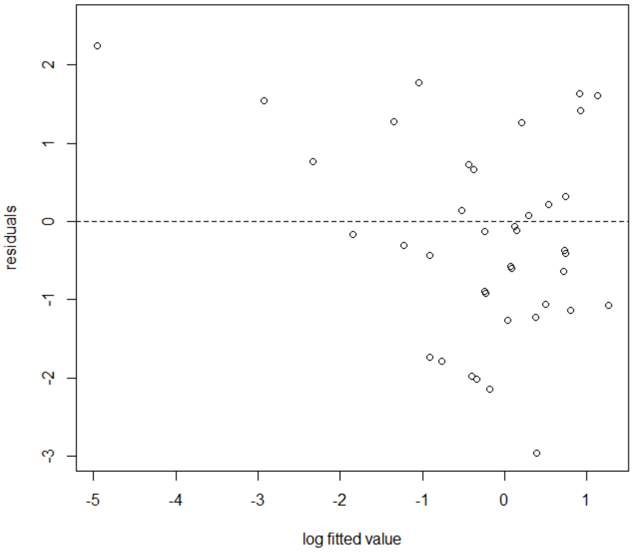

Figure 3

The result shows that all the points are located within the -2 to 2 around the line residuals=0. And it seems there is a non-linear negative tendency between residual and $\log(\hat{Y})$, but if we ignore the left 1 or 2 points, which is far away from the main part, then then tendency disappears. So I think the assumption for the linear is still hold.

For the QQ plot, shows as figure 4.

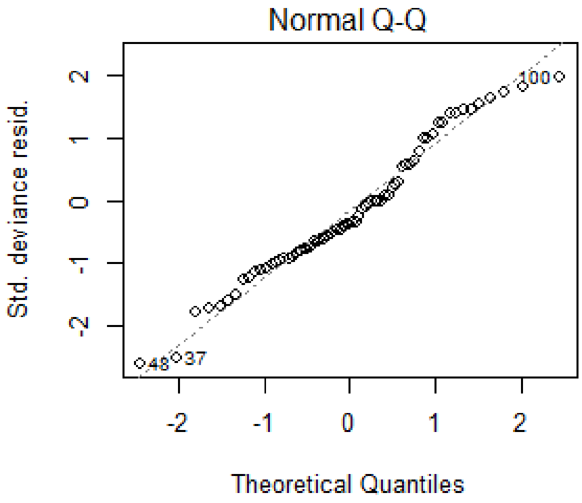

Figure 4

All the points are located around the line. So the model achieved the normality assumption.

So from the two tests mentioned above I can conclude that the whole model is well fitted.

6 Conclusion and further works

6.2 Conclusion

The Flint Water Crisis has happened almost 2 years, and after changing the water resource back the situation is much better. But it is still important to analysis the cause of this crisis, so that we can prevent such situation happen again.

Base on the analysis, I got the conclusion that the material of the service line is the main reason that affects lead level. So the government can avoid this crisis by choosing the pipe material more carefully.

6.2 Further works

Although the Quasi-Poisson model is well fitted, since the limitation of the data, there are still some further works need to be done. For example although the material of pipe is main factor compared with zip code and sample location, it maybe not the unique factor. For example, it is possible that the environment of plant near the water resource make the water itself contain some chemical element that react with the pipe, then cause the high level of lead. What's more, the data sets are collected in 2016, but the government has changed the water source back in 2015. So if the data before 2015 and the data beyond Flint City are available, I can do the comparison between those datasets and make the model more precise and generalized, I may can find the time tendency. So collecting more complete data is the main further work.