Lab 2: Non-linear regression models

The flint water crisis has received much attention in the past few months. In today's lab, we will apply the non-linear regression to analyze the lead testing results data.

Starting in April 2014, Flint changed its water source from the treated Detroit and Sewerage department to the Flint River. Since then the drinking water in Flint city had a series of problems including the lead contamination. According to the laboratory testing done by a research group from Virginal Tech, the Flint River water was highly corrosive for lead. The introduction of the Flint water into the aging pipes caused lead to leach into water supply.

Here are a few details about the discovery of the lead contamination in Flint supply water. On Feb. 18, 2015, Ms. Walters detected 104 parts per billion (ppb) of lead in the drinking water at her home. On March 3, 2015, a second testing at Ms. Walters's home detected 397 ppb of lead in drinking water. On Sept. 2, 2015, Dr. Edwards reported that corrosiveness of water is causing lead to leach into water supply. On Sept. 24-25, 2015, a group of doctors found high levels of lead in the blood of children. On Oct. 16, 2015, Flint reconnected to Detroit's water. (Source: http://www.nytimes.com/interactive/2016/01/21/us/flint-lead-water-timeline.html?_r=0).

(Source of photos: http://flintwaterstudy.org).

Sampling time:

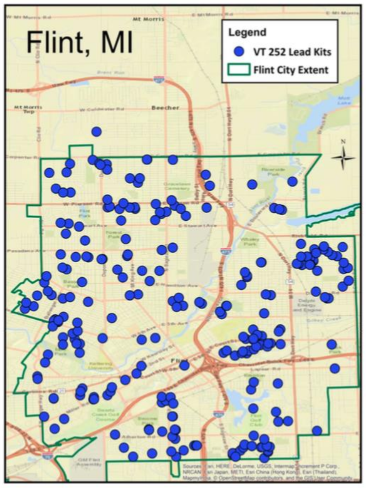
The data set was collected between Aug. 20 and Sept. 8, 2015 by a research team at Virginal Tech (VT). Part of the samples may be collected some time after Sept. 8, 2015. Since I did not find the exact information on the website (http://flintwaterstudy.org), I am not able to provide more accurate sampling time frame.

Sampling scheme:

The research team sent out 300 sample kits to Flint residents. They received 271 water samples for lab testing. According to the website (http://flintwaterstudy.org), we have the following information about the sampling scheme:

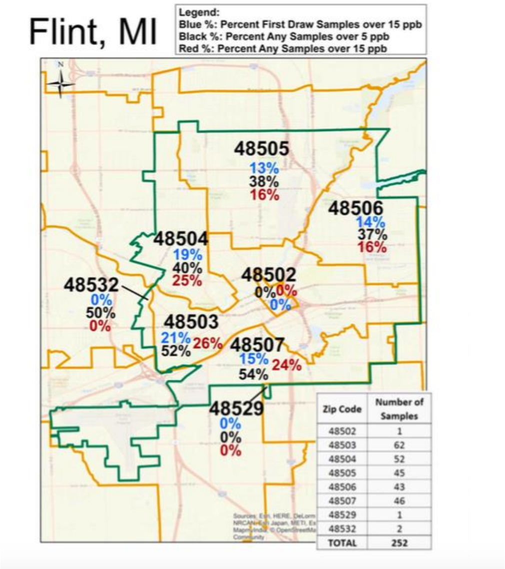
"We do not know who has lead pipes in Flint, and who does not. We did not recruit people who only have lead pipe. In that regard, the sampling approach is random relative to the key criteria of having (or not having) lead in the plumbing material. The actual volunteer citizens taking the samples are a self-selected group of people, but that self-selection has nothing to do with the likelihood of having a lead in water risk in their house, and is more likely related to citizens being concerned for their well-being and the well-being of their families."

To see where the lead kits were distributed, see the following map provided by the VT research team (source: http://flintwaterstudy.org/wp-content/uploads/2015/09/Town-Hall-Meeting-Marc-Edwards-and-Sid-Roy-FlintWaterStudy-NSF-RAPID.pdf).



Sampling areas:

The data set contains samples from areas with zip codes 48502-48507, 48529 and 48532. Below is a map obtained from the presentation slides of the VT research team (source: http://flintwaterstudy.org/wp-content/uploads/2015/09/Town-Hall-Meeting-Marc-Edwards-and-Sid-Roy-FlintWaterStudy-NSF-RAPID.pdf).



Flint, MI

Legend:
Blue %: Percent First Draw Samples over 15 ppb
Black %: Percent Any Samples over 5 ppb
Red %: Percent Any Samples over 15 ppb

Zip Code	Number of Samples
48502	1
48503	62
48504	52
48505	45
48506	43
48507	46
48529	1
48532	2
TOTAL	252

Testing results:

The data set contains the lead level measurements obtained at three times: the first draw, after 45 seconds flushing and after 2 minutes flushing.

Criteria:

According to Environmental Protection Agency (EPA), if the “worst case” homes are tested and more than 10% are over 15ppb lead, the city exceeds the federal standards for lead in water. This means that if the 90% quantile of the “worst case” is less than 15ppb, EPA does not require action. In the sampling conducted by the VT research team, they do not know “worse case” homes. The sampling they conducted was a random sampling. If the 90% quantile of the random sample exceeds 15ppb, then Flint water had a serious problem with lead.

The purpose of this lab is to determine that if the Flint water had a serious lead problem. Below are a few questions we would like to address:

Q1. Read the data into R. Then perform some initial data analysis. a) How many samples are distributed in each zip code? Are they evenly sampled? b) Check the histograms of the lead level at three repeated measurements. What are the shapes of the histograms? c) Plot the lead level versus the sample collecting time for all the households? Do you see any extreme cases? If you see extreme cases, remove the extreme cases and plot the curves again. d) Repeating part (c) for the data with zip codes 48503 to 48507. What patterns do you observe for each zip code? Do you see any nonlinearity pattern?

Answer: We first read the data set into R and rename the variables using the following R code

```
 setwd("…") ## put your working directory here
 flintlead<-read.csv(file="Flint-water-lead-dataset.csv",
 header=FALSE)
colnames(flintlead)<-c("SampleID","Zip Code","Ward", "Pb Bottle
1(ppb)-First Draw", "Pb Bottle 2(ppb)- 45 secs flushing", "Pb Bottle
3(ppb)- 2 mins flusing")
```

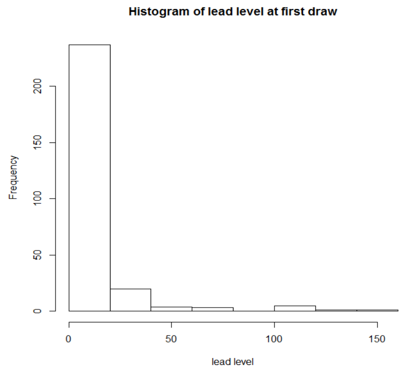
(a) To check the samples within each zip code, we use the following R code

```
table(flintlead[,2])
48502 48503 48504 48505 48506 48507 48529 48532
    1    69    55    48    44    51     1     2
```

Based on the above results, we see that there are 8 distinct zip codes and the samples are not evenly distributed among zip codes. In particular, in zip codes “48502”,”48529” and “48532”, there are only one or two samples. In rest of zip codes, the samples are more evenly distributed.

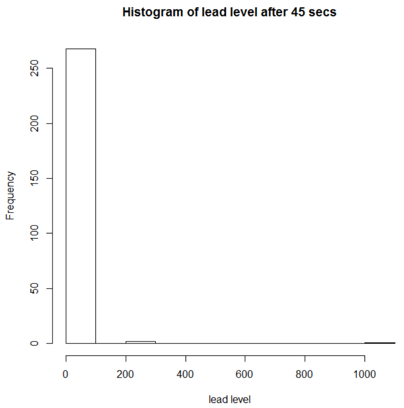
(b) The histogram of the lead levels at first sampling time is

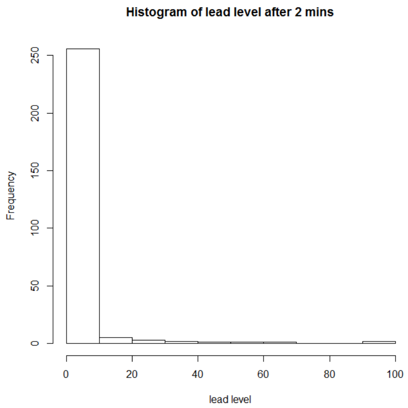
```
hist(flintlead[,4], main="Histogram of lead level at first draw",
xlab="lead level")
```



The histograms of the lead levels after 45 seconds and 2 minutes are, respectively,

```
hist(flintlead[,5], main="Histogram of lead level after 45 secs",
     xlab="lead level")
hist(flintlead[,6], main="Histogram of lead level after 2 mins",
     xlab="lead level")
```



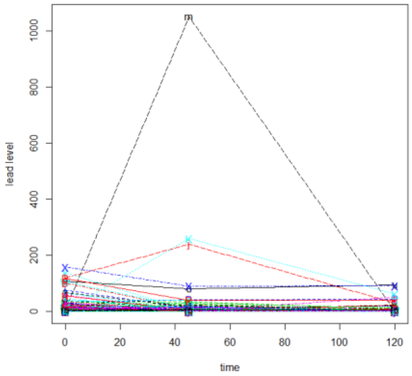


The above three histograms are heavily right skew due to some extreme values in each cases. For the histogram of the first draw, the majority of data is between 0-50 ppb. For the second histogram, the majority is between 0 and 100ppb. For the third histogram, the majority is between 0-10ppb.

(c) We first plot the lead levels versus the sampling time using the following code

```
time<-c(0, 45, 120)
matplot(time, t(flintlead[,4:6]),type="o", ylab="lead level")
```

The plot is shown below:



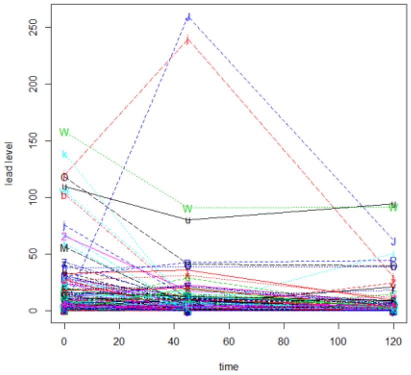
By examining the above plot, we see that most data points are within the range of 0 and 300ppb. But there exists a data point which is over 1000ppb. This point is clearly an extreme case compared to others. Moreover, we observed that, for most households, the lead level decreases as the sampling time increases. But this is not the case for the extreme case. In summary, the data point with extremely large value at the second sampling time is considered as an extreme case. To identify the ID and corresponding values, we use the following code

```
 extremeID<-which(flintlead[,5]>1000)
extremesvalues<-flintlead[extremeID,]
```

Based on the above code, we found that the extreme case is the #85 observation with sample ID 97. Then we remove the extreme case and plot the curves again.

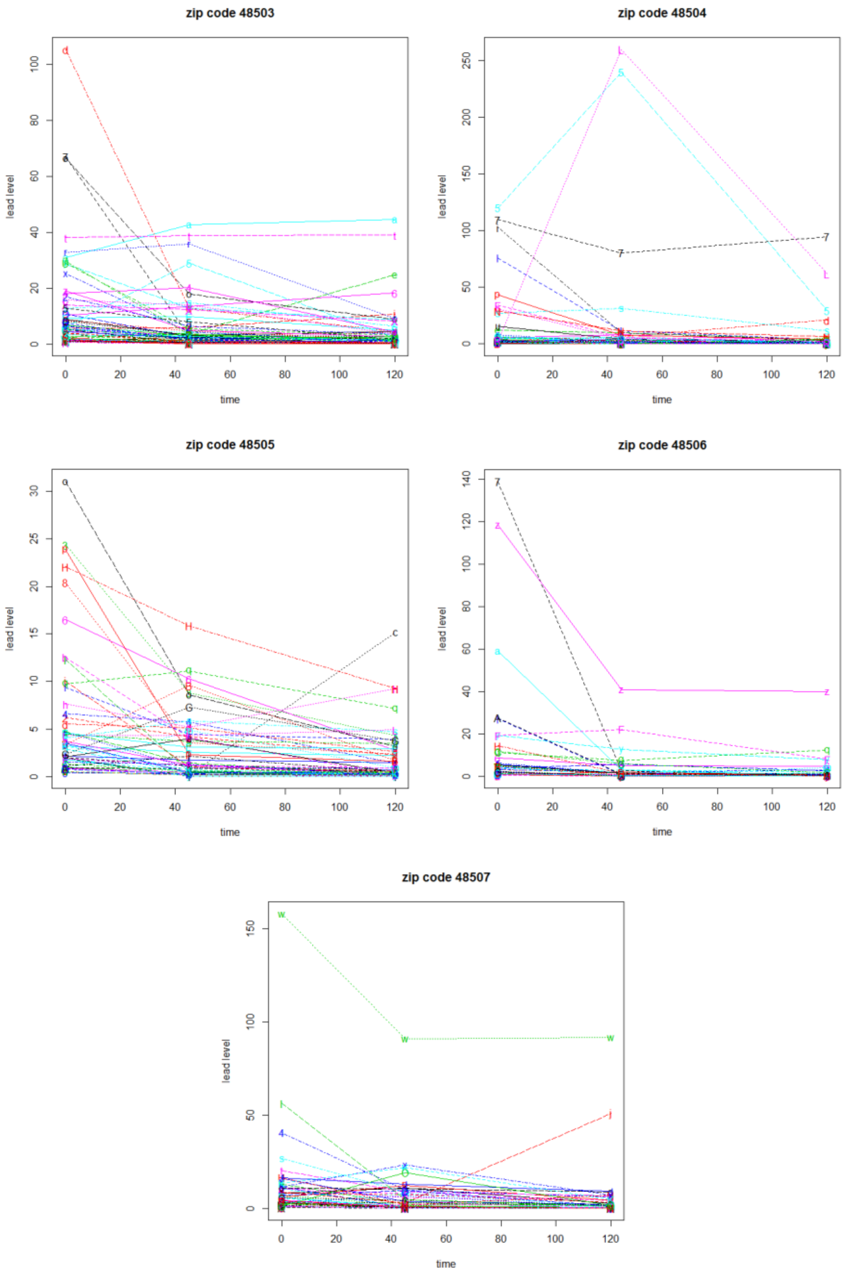
```
flintlead2<-flintlead[-extremeID,]
matplot(time,t(flintlead2[,4:6]),type="o",ylab="lead level")
```

The new plot is given below



(d) We now plot the lead level versus the sampling time for each zip code below:

```
zipcode1<-which(flintlead2[,2]==48503)
matplot(time,t(flintlead2[zipcode1,4:6]),type="o",ylab="lead
level",main="zip code 48503")
zipcode2<-which(flintlead2[,2]==48504)
matplot(time,t(flintlead2[zipcode2,4:6]),type="o",ylab="lead
level",main="zip code 48504")
zipcode3<-which(flintlead2[,2]==48505)
matplot(time,t(flintlead2[zipcode3,4:6]),type="o",ylab="lead
level",main="zip code 48505")
zipcode4<-which(flintlead2[,2]==48506)
matplot(time,t(flintlead2[zipcode4,4:6]),type="o",ylab="lead
level",main="zip code 48506")
zipcode5<-which(flintlead2[,2]==48507)
matplot(time,t(flintlead2[zipcode5,4:6]),type="o",ylab="lead
level",main="zip code 48507")
```



Overall, we observed from the above plots that, for most households, the lead levels decrease as the sampling time increases. For two households in zip code area 48504 and one household in zip code 48505 and 48507, we can also clearly see that the lead level decreases from the first sampling time to the second sampling time, and then increasing from the second sampling time to the third sampling time. For zip code 48503 to 48506, the non-linearity pattern is easy to observe. Because, for most curves, the slops are very sharp between the first two sampling times, and then the slops become smaller between the second and third sampling times. For the zip code 48507, the pattern may not be as clear as the first four zip codes, but one can still see that the difference slopes between the first two sampling times and the last two sampling times. Assuming a non-linear relationship between the sampling time and lead levels is reasonable for this data set.

Q2. Based on the initial data analysis in Q1, for each zip code, build a non-linear regression model using lead measurements as responses and sampling time as covariates. The commonly used exponential decay model is

$$f(x,\theta) = \theta_1 \exp(-\theta_2 x).$$

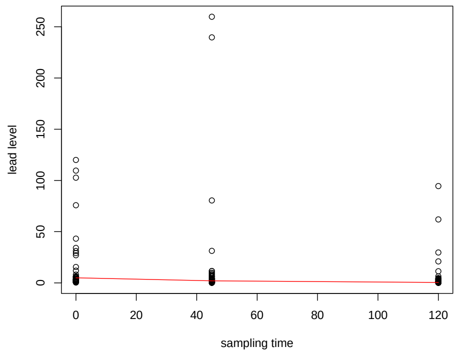
A generalized model could be

$$f(x,\theta) = \theta_1 / \{1 + \theta_2 \exp(\theta_3 x)\}.$$

Answer: In this question, we consider the first exponential decay model. We first consider the zip code 48503. The following R code is used to check if the initial values of θ_1 and θ_2 are appropriate. Note that θ_1 means the lead level at time 0 and θ_2 is the decay rate. In our analysis, θ_1 is chosen to close to the median of the lead level at time 0 and θ_2 is chosen so that it is close to the decay rate of the medians at three sampling times.

```
subset1<-which(flintlead3[,2]==48503)
subsetflintlead<-flintlead3[subset1,]
responses1<-unlist(subsetflintlead[,4:6])
sampletime1<-rep(time,each=dim(subsetflintlead)[1])
plot(sampletime1, responses1,xlab="sampling time",ylab="lead level")
theta1<-5
theta2<-0.02
lines(sampletime1, theta1*exp(-sampletime1*theta2),col=2)
```

The following plot provides the mean functions generated by initial values and the original data. By checking the plot, one might see if the initial values for θ_1 and θ_2 are appropriate.



Then we fit the nonlinear models as following

```
   nlsreg1<-nls(responses1~theta1*exp(-sampletime1*theta2),
start=list(theta1=5,theta2=0.02))
   summary(nlsreg1)
```

The output of the nonlinear regression model fitting is

```
Formula: responses1 ~ theta1 * exp(-sampletime1 * theta2)
Parameters:
        Estimate Std. Error t value Pr(>|t|)
theta1 10.592500   1.417368   7.473 2.23e-12 ***
theta2  0.010531   0.003685   2.858  0.00471 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 12.29 on 205 degrees of freedom
Number of iterations to convergence: 5
Achieved convergence tolerance: 3.866e-06
```

For zip code 48504 to 48507, we fit the non-linear regression model similarly to that for zip code 48503. For conciseness, we will not copy the plot for each case here.

For zip code 48504:

```
   subset2<-which(flintlead3[,2]==48504)
   subsetflintlead<-flintlead3[subset2,]
   responses2<-unlist(subsetflintlead[,4:6])
   sampletime2<-rep(time,each=dim(subsetflintlead)[1])
   plot(sampletime2, responses2,xlab="sampling time",ylab="lead level")
   theta1<-5
   theta2<-0.02
   lines(sampletime2, theta1*exp(-sampletime2*theta2),col=2)
   nlsreg2<-nls(responses2~theta1*exp(-sampletime2*theta2),
start=list(theta1=5,theta2=0.02))
   summary(nlsreg2)

   Formula: responses2 ~ theta1 * exp(-sampletime2 * theta2)
   Parameters:
           Estimate Std. Error t value Pr(>|t|)
   theta1 14.582360   4.188181   3.482 0.000642 ***
   theta2  0.006289   0.005836   1.078 0.282818
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
   Residual standard error: 33.19 on 160 degrees of freedom
   Number of iterations to convergence: 8
   Achieved convergence tolerance: 2.245e-06
```

For zip code 48505:

```
   subset3<-which(flintlead3[,2]==48505)
   subsetflintlead<-flintlead3[subset3,]
   responses3<-unlist(subsetflintlead[,4:6])
   sampletime3<-rep(time,each=dim(subsetflintlead)[1])
   plot(sampletime3, responses3,xlab="sampling time",ylab="lead level")
   theta1<-5
   theta2<-0.02
   lines(sampletime3, theta1*exp(-sampletime3*theta2),col=2)
   nlsreg3<-nls(responses3~theta1*exp(-sampletime3*theta2),
start=list(theta1=5,theta2=0.02))
   summary(nlsreg3)

   Formula: responses3 ~ theta1 * exp(-sampletime3 * theta2)
   Parameters:
          Estimate Std. Error t value Pr(>|t|)
   theta1 5.891855   0.689998   8.539 1.89e-14 ***
   theta2 0.010384   0.003191   3.254  0.00142 **
   ---
   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   Residual standard error: 4.996 on 142 degrees of freedom
   Number of iterations to convergence: 6
   Achieved convergence tolerance: 8.061e-06
```

For zip code 48506:

```
   subset4<-which(flintlead3[,2]==48506)
   subsetflintlead<-flintlead3[subset4,]
   responses4<-unlist(subsetflintlead[,4:6])
   sampletime4<-rep(time,each=dim(subsetflintlead)[1])
   plot(sampletime4, responses4,xlab="sampling time",ylab="lead level")
   theta1<-5
   theta2<-0.02
   lines(sampletime4, theta1*exp(-sampletime4*theta2),col=2)
   nlsreg4<-nls(responses4~theta1*exp(-sampletime4*theta2),
start=list(theta1=5,theta2=0.02))
   summary(nlsreg4)
   Formula: responses4 ~ theta1 * exp(-sampletime4 * theta2)
   Parameters:
          Estimate Std. Error t value Pr(>|t|)
   theta1 12.04647    2.54722   4.729 5.78e-06 ***
   theta2  0.02632    0.01507   1.746   0.0832 .
   ---
   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   Residual standard error: 16.93 on 130 degrees of freedom
   Number of iterations to convergence: 7
   Achieved convergence tolerance: 4.33e-06
```

For zip code 48507:

```
  subset5<-which(flintlead3[,2]==48507)
  subsetflintlead<-flintlead3[subset5,]
  responses5<-unlist(subsetflintlead[,4:6])
  sampletime5<-rep(time,each=dim(subsetflintlead)[1])
  plot(sampletime5, responses5,xlab="sampling time",ylab="lead level")
  theta1<-5
  theta2<-0.02
  lines(sampletime5, theta1*exp(-sampletime5*theta2),col=2)
  nlsreg5<-nls(responses5~theta1*exp(-sampletime5*theta2),
start=list(theta1=5,theta2=0.02))
  summary(nlsreg5)

  Formula: responses5 ~ theta1 * exp(-sampletime5 * theta2)
  Parameters:
          Estimate Std. Error t value Pr(>|t|)
  theta1 10.604467   2.314557   4.582 9.59e-06 ***
  theta2  0.008140   0.005063   1.608     0.11
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Residual standard error: 17.55 on 151 degrees of freedom
  Number of iterations to convergence: 5
  Achieved convergence tolerance: 7.246e-06
```

In summary, the nonlinear models fitted above are not very sensitive to the initial values. It typically takes only a few steps to reach convergence.

Q3. How the lead levels changing over the flushing time? Is flushing an efficient way to reduce the water lead level?
Answer: In the exponential decay model, $\theta_1$represents the mean of the lead level at time 0 and θ_2 represents the decay rate of the lead level as the flushing time increases. Based on the output in Q2, we see that θ_1 is significantly different from 0 for all the zip codes. This means that there exists significant amount of lead in water at initial time.

For the parameter θ_2, we found that the estimation of θ_2 is positive in all zip codes. This indicates that the lead level in water decreases as the flushing time increases. Moreover, we found that θ_2 is significant different from 0 for zip codes 48503 to 48505. But it is not significant at the nominal level 0.05 for zip codes 48506 and 48507. This could suggest that flushing is an efficient way to reduce lead level in zip code areas 48503 to 48505, but it is not that efficient for zip codes 48506 and 48507. This could imply some differences in the pluming systems among these zip codes.

Q4. Are there significant differences among areas with different zip codes in the lead contaminations?

Answer: To know if the differences among zip codes are significant or not, we perform hypothesis testing to examine if the parameter values among zip codes are the same or not.

More specifically, assume that the parameter values θ_1 and θ_2 for zip code 48503 is $\theta^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)})'$ and the parameter values for zip code 48504 is $\theta^{(2)} = (\theta_1^{(2)}, \theta_2^{(2)})$. The hypothesis we want to test is $H_0: \theta^{(1)} = \theta^{(2)}$ versus $H_1: \theta^{(1)} \neq \theta^{(2)}$. Let $\widehat{\theta^{(1)}}$ and $\widehat{\theta^{(2)}}$ be least squares estimators for $\theta^{(1)}$ and $\theta^{(2)}$. Note that the estimators are obtained using two different independent parts of data. Therefore, the estimators $\widehat{\theta^{(1)}}$ and $\widehat{\theta^{(2)}}$ are independent. We use a wald type of inference. The test statistic is defined as

$$Q_{n,12}=\left(\widehat{\theta^{(1)}} - \widehat{\theta^{(2)}}\right)' \left[\mathrm{Var}\left(\widehat{\theta^{(1)}}\right) + \mathrm{Var}(\widehat{\theta^{(2)}})\right]^{-1}(\widehat{\theta^{(1)}} - \widehat{\theta^{(2)}}).$$

We reject the null hypothesis if $Q_{n,12}$ is larger than the upper 5% quantile of chi-square distribution with one degree of freedom. The R code for comparing five zip codes is given below

```
   zipcodes<-c(48503,48504, 48505, 48506, 48507)
   Qnvec<-NULL
   for (i in 1:4)
    for (j in (i+1):5)
   {
    subset1<-which(flintlead3[,2]==zipcodes[i])
    subsetflintlead<-flintlead3[subset1,]
    responses1<-unlist(subsetflintlead[,4:6])
    sampletime1<-rep(time,each=dim(subsetflintlead)[1])
    nlsreg1<-nls(responses1~theta1*exp(-sampletime1*theta2),
start=list(theta1=5,theta2=0.02))
    subset2<-which(flintlead3[,2]==zipcodes[j])
    subsetflintlead<-flintlead3[subset2,]
    responses2<-unlist(subsetflintlead[,4:6])
    sampletime2<-rep(time,each=dim(subsetflintlead)[1])
    nlsreg2<-nls(responses2~theta1*exp(-sampletime2*theta2),
start=list(theta1=5,theta2=0.02))
```

```
    Qn<-t(coef(nlsreg1)-
coef(nlsreg2))%*%solve(vcov(nlsreg1)+vcov(nlsreg2))%*%(coef(nlsreg1)-
coef(nlsreg2))
    Qnvec<-rbind(Qnvec,c(i,j,Qn, 1-pchisq(Qn,1)))
  }
```

The output of the above R code is

```
      Zp1 Zp2         Qn         p-value
 [1,]   1   2  2.6453336 0.1038543564
 [2,]   1   3 11.5975428 0.0006603902
 [3,]   1   4  1.0750572 0.2998059332
 [4,]   1   5  0.2073597 0.6488447502
 [5,]   2   3  8.5729760 0.0034118901
 [6,]   2   4  2.4728557 0.1158273118
 [7,]   2   5  1.4336859 0.2311638427
 [8,]   3   4  5.5624010 0.0183502302
 [9,]   3   5  6.5168008 0.0106860057
[10,]   4   5  1.3100686 0.2523822431
```

In the above output, zip code 1 to 5 represents zip code areas 48503 to 48507 respectively. According to the p-values, we found that zip code 48505 are significantly different from rest of the zip codes, since all the corresponding p-values are smaller than 0.05. As we can see from the estimates of θ_1, the estimated value of θ_1 in zip code 48505 is much smaller than the other zip codes. This suggests that the zip code 48505 might be less severe than the other zip codes in lead contamination.

Q5. Estimate the 90% quantile curve of the lead level. Does it exceed the federal level 15 ppb?

Answer: By assumption, we know the response has a normal distribution with mean $f(t, \theta)$ and variance σ^2 at sampling time t. As a result, we can estimate the 90% quantile as $f\left(t, \hat{\theta}\right) + \hat{\sigma}\Phi^{-1}(0.9)$. Applying this formula, we can obtain the estimation of the 90% quantile at three sampling time points for each zip code. The R code and out put are given below:

```
Q90mat<-matrix(0,5,3)
zipcodes<-c(48503,48504, 48505, 48506, 48507)
timevec<-c(0,45,120)
```

```
   for (i in 1:5)
   {
    subset<-which(flintlead3[,2]==zipcodes[i])
    subsetflintlead<-flintlead3[subset,]
    responses<-unlist(subsetflintlead[,4:6])
    sampletime<-rep(time,each=dim(subsetflintlead)[1])
    nlsreg1<-nls(responses~theta1*exp(-sampletime*theta2),
start=list(theta1=5,theta2=0.02))
    for (j in 1:3)
    {
      theta1hat<-coef(nlsreg1)[1]
      theta2hat<-coef(nlsreg1)[2]
      meany<-theta1hat*exp(-timevec[j]*theta2hat)
      sigmahat<-summary(nlsreg1)$sigma
      y90quantile<-qnorm(0.9, meany, sigmahat)
      Q90mat[i,j]<-y90quantile
     }
   }
```

The output of the above R code is given below

```
Zip code        Flushing time
                0         45        120
48503    26.34515 22.34716 18.745996
48504    57.11656 53.52245 49.390605
48505    12.29387 10.09452   8.096755
48506    33.74771 25.38650 22.213103
48507    33.09380 29.84127 26.481946
```

Based on the above results, we observed that, except zip code 48505, the 90% quantiles of all the zip codes exceed the federal level 15 ppb. Even after 2 minutes flushing, the 90% quantiles are still above the federal level. The zip code area 48505 has lower 90% quantiles than the other zip codes.

Q6. Constructing 95% confidence intervals for the 90% quantiles of lead level at different sampling times. What is your conclusion? Do you think the inference is accurate enough?

Answer: The method for building 95% confidence intervals for the 90% quantiles of lead level was provided in a note send in the email. Please refer to the details given in that note. Here I will provide the R code and the output of the R code.

```
   Q90CImat<-matrix(0,5,6)
   zipcodes<-c(48503,48504, 48505, 48506, 48507)
   timevec<-c(0,45,120)
   for (i in 1:5)
   {
    subset<-which(flintlead3[,2]==zipcodes[i])
    subsetflintlead<-flintlead3[subset,]
    responses<-unlist(subsetflintlead[,4:6])
    sampletime<-rep(time,each=dim(subsetflintlead)[1])
    nlsreg1<-nls(responses~theta1*exp(-sampletime*theta2),
start=list(theta1=5,theta2=0.02))
    for (j in 1:3)
     {
      theta1hat<-coef(nlsreg1)[1]
      theta2hat<-coef(nlsreg1)[2]
      meany<-theta1hat*exp(-timevec[j]*theta2hat)
      sigmahat<-summary(nlsreg1)$sigma
      y90quantile<-qnorm(0.9, meany, sigmahat)
      Dtheta<-cbind(exp(-theta2hat*sampletime),-
sampletime*theta1hat*exp(-theta2hat*sampletime))
      Vartheta<-solve(t(Dtheta)%*%Dtheta)
      Gbeta<-c(exp(-theta2hat*timevec[j]),(-timevec[j])*theta1hat*exp(-
theta2hat*timevec[j]))
      Vary90<-t(Gbeta)%*%Vartheta%*%Gbeta
      CI4y90quantile<-c(y90quantile-
qnorm(0.975)*sigmahat*sqrt(Vary90),y90quantile+qnorm(0.975)*sigmahat*s
qrt(Vary90))
      Q90CImat[i,(j-1)*2+c(1:2)]<-CI4y90quantile
      }
     }
```

Using the above code, we obtain the following confidence intervals for the 90% quantiles:

Zip code		Flushing time	
	0	45	120
48503	(23.56716, 29.12314)	(20.391416, 24.30291)	(16.447920, 21.044071)
48504	(48.90788, 65.32525)	(48.150363, 58.89454)	(41.613546, 57.167664)
48505	(10.94150, 13.64624)	(9.145612, 11.04343)	(6.972026, 9.221485)
48506	(28.75525, 38.74018)	(20.693120, 30.07989)	(20.447694, 23.978513)
48507	(28.55736, 37.63025)	(26.798127, 32.88442)	(22.413622, 30.550269)

For zip codes 48503, 48504, 48506 and 48507, the confidence intervals for 90% quantiles are all above the federal level 15ppb, which further confirmed that the flint water did not meet the federal requirement. For the zip code 48505, the confidence intervals are below the federal level 15ppb, which suggests that this area has less serve lead problem when compared with other zip codes.

However, we also need to note that the statistical inference is based on asymptotic theory. We have two concerns: (1) the sample sizes are not big enough because most zip code areas have sample size around 50; (2) we use normality assumption in finding the 90% quantiles. This assumption might not be entirely appropriate based on our initial data analysis given in Q1.

Questions after lab:

Q7. Fit a non-linear regression model with using zip codes and sampling time as covariates, where we can treat zip codes as categorical variable.

Answer: If we use both zip codes and sampling time as covariates, we could use the following model

$$f(t,\theta) = (\theta_{11} + \sum_{j=1}^{4} \theta_{1(j+1)}\mathrm{Z}_j)\exp\{-t(\theta_{21} + \sum_{j=1}^{4} \theta_{2(j+1)}\mathrm{Z}_j)\}.$$

In the above model, Z_j are dummy variables representing different zip code areas. The zip code 48507 is used as the baseline. Specifically, Z_1=1 if data point is in the zip code 48503 otherwise Z_1=0. Z_2=1 if data point is in the zip code 48504 otherwise Z_2=0. Z_3=1 if data point is in the zip code 48505 otherwise Z_3=0. Z_4=1 if data point is in the zip code 48506 otherwise Z_4=0. The initial values for parameters could be chosen according to the results obtained in Q2.

The R code for implementing the above model is given as following:

```
   dummies<-
matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0),5,4,byrow=TRUE)
   zipcodes<-c(48503,48504, 48505, 48506, 48507)
   timevec<-c(0,45,120)
   responsevec<-NULL
   sampletimevec<-NULL
   zpdummy<-NULL
   for (i in 1:5)
   {
    subset1<-which(flintlead3[,2]==zipcodes[i])
    subsetflintlead<-flintlead3[subset1,]
    responses1<-unlist(subsetflintlead[,4:6])
    sampletime1<-rep(timevec,each=dim(subsetflintlead)[1])
    zpdummy1<-
matrix(rep(dummies[i,],3*dim(subsetflintlead)[1]),3*dim(subsetflintlea
d)[1], 4, byrow=TRUE)
    responsevec<-c(responsevec,responses1)
    sampletimevec<-c(sampletimevec,sampletime1)
    zpdummy<-rbind(zpdummy,zpdummy1)
   }
   dummy1<-zpdummy[,1]
   dummy2<-zpdummy[,2]
   dummy3<-zpdummy[,3]
   dummy4<-zpdummy[,4]
   nlsreg2<-
nls(responsevec~(theta11+theta12*dummy1+theta13*dummy2+theta14*dummy3+
theta15*dummy4)*exp(-
sampletimevec*(theta21+theta22*dummy1+theta23*dummy2+theta24*dummy3+th
eta25*dummy4)),
start=list(theta11=10,theta12=0.5,theta13=4.5,theta14=-
4,theta15=2,theta21=0.01,theta22=0.02,theta23=0.01,theta24=0.01,theta2
5=0.01))
   summary(nlsreg2)
```

The output of the R code is given below:

```
Parameters:
        Estimate Std. Error t value Pr(>|t|)
theta11 10.604596   2.549999   4.159 3.55e-05 ***
theta12 -0.012144   3.387101  -0.004    0.997
theta13  3.977813   3.529105   1.127    0.260
theta14 -4.712703   3.692373  -1.276    0.202
theta15  1.441635   3.867851   0.373    0.709
theta21  0.008141   0.005579   1.459    0.145
theta22  0.002390   0.008045   0.297    0.766
theta23 -0.001852   0.006533  -0.284    0.777
theta24  0.002243   0.013553   0.166    0.869
theta25  0.018175   0.018089   1.005    0.315
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 19.33 on 788 degrees of freedom
Number of iterations to convergence: 7
Achieved convergence tolerance: 5.605e-06
```

Q8. Bootstrap is another inference method that is more appropriate for small sample inference. Can you use Bootstrap method to construct confidence intervals for the 90% quantile of the lead level?

Answer: Bootstrap is a resampling method that contains three main steps (1) resample the data from the original data with replacement; (2) estimate the 90% quantiles for each resampling data set; (3) Repeat step (1) and (2) for B times. B should be big enough. Then use the lower and upper 2.5% quantiles to construct 95% confidence intervals for the 90% quantiles. The R code for implementing the Bootstrap is given below:

```
   BootStrapQ90CImat<-matrix(0,5,6)
   zipcodes<-c(48503,48504, 48505, 48506, 48507)
   timevec<-c(0,45,120)
   Brep<-1000
   for (i in 1:5)
   {
    subset<-which(flintlead3[,2]==zipcodes[i])
    subsetsize<-length(subset)
    y90quantilesmat<-matrix(0,Brep,3)
    for (j in 1:3)
    {
      for (b in 1:Brep)
      {
      subsetboot<-sample(subset, subsetsize, replace=TRUE)
      subsetflintlead<-flintlead3[subsetboot,]
      responses<-unlist(subsetflintlead[,4:6])
      sampletime<-rep(time,each=dim(subsetflintlead)[1])
      res<- try(nlsreg1<-nls(responses~theta1*exp(-sampletime*theta2),
start=list(theta1=median(subsetflintlead[,4]),theta2=0.01)))
      if(inherits(res, "try-error"))
      {
        y90quantilesmat[b,j]<-NA
      }
      theta1hat<-coef(nlsreg1)[1]
      theta2hat<-coef(nlsreg1)[2]
      meany<-theta1hat*exp(-timevec[j]*theta2hat)
      sigmahat<-summary(nlsreg1)$sigma;
      y90quantile<-qnorm(0.9, meany, sigmahat)
      y90quantilesmat[b,j]<-y90quantile
      }
     BootStrapQ90CImat[i,(j-1)*2+c(1:2)]<-quantile(y90quantilesmat[,j],
c(0.025, 0.975),na.rm=TRUE)
    }
   }
```

The Bootstrap confidence intervals for 90% quantiles are given in the following table.

Zip code	Flushing time		
	0	45	120
48503	(17.417249, 35.50380)	(15.300766, 28.97907)	(12.642311, 24.180339)
48504	(25.345554, 86.47634)	(17.627689, 81.11499)	(14.838799, 74.017574)

48505	(8.911765, 15.59846)	(7.610928, 12.22040)	(5.960811, 9.847349)
48506	(12.678214, 51.92893)	(8.827947, 36.84242)	(6.864102, 32.881601)
48507	(13.525934, 54.23085)	(11.385182, 48.48622)	(9.304525, 43.842985)

Based on the above table, we can see that the confidence intervals are mostly much wider than the confidence intervals based on asymptotic normality (given in question Q6). The conclusions based on the Bootstrap inference are quite different from those in Q6 because many intervals (except zip code 48503, time 0 and 45 and zip code 48504, time 0 and time 45) in the above table include 15ppb. Therefore, the 90% quantiles of the lead level might not be significantly different from 15ppb.