36-617 Midterm Report I (Statistician)

Mengran He

15 October, 2016

Exploratory Data Analysis

This research is to analyze the relationship between arsenic concentration in toenails and arsenic levels in the well water. The samples were collected in the area of New Hampshire. The dataset includes 21 observations and five variables age (Age), gender (Sex), the measurement of how often they use well water to cook (Cook) and drink (Drink), arsenic concentration in the well (Water) and arsenic concentration in individuals' toenails (Toes). Among the 21 well water samples, 15 well water samples were detected to have trace concentration in arsenic. By looking at the histograms and boxplots of Water and Toenails, we can see that both variables are right skewed, and there are three outliers in both boxplots that are far away from upper whiskers. Those outliers may pull the regression line towards them and will appear in residuals' QQplot, so further investigation on those outliers is necessary.

```
arsenic <- read.csv("arsenic.csv", header = TRUE)
par(mfrow = c(1, 4))
hist(arsenic$Water, xlab = "Water", main = "Histogram of Water", ylim = c(0,
    20), col = "grey")
hist(arsenic$Toes, xlab = "Toes", main = "Histogram of Toes", col = "grey",
    ylim = c(0, 20))
boxplot(arsenic$Water, xlab = "Water", main = "Boxplot of Water")
boxplot(arsenic$Toes, xlab = "Toes", main = "Boxplot of Toes")
```

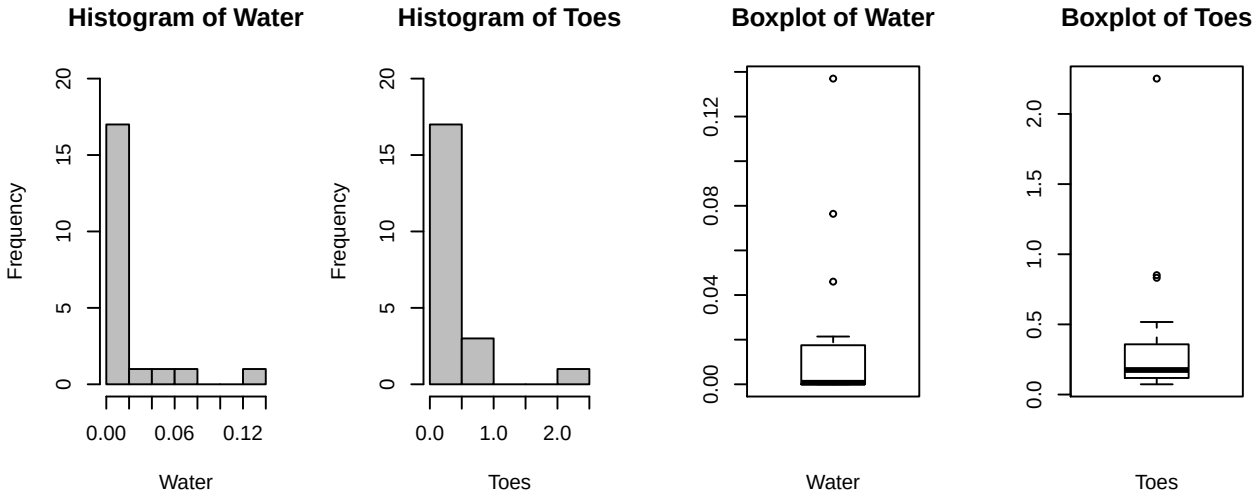

Figure 1: Histogram and Boxplot of Water and Toenails

Moveroever, table 1 summarizes the mean, range and variance of Water and Toes. By calculation, the mean of Water (0.016) and Toes (0.366) are both greater than their median, 0.001 and 0.175 respectively, another representation of right-skewness. By comparing the variance of these two variables, Toes (0.237) has greater variablity than Water (0.001).

```r
# summary(arsenic$Water) var(arsenic$Water) summary(arsenic$Toes)
# var(arsenic$Toes) use summary() and var() to extract values of mean,
# median, range and variance
out <- data.frame(Minimum = c(0, 0.07), Median = c(0.001, 0.18), Mean = c(0.016,
    0.37), Maximum = c(0.137, 2.25), Variance = c(0.001, 0.24))
row.names(out) <- c("Water", "Toes")
pandoc.table(out, style = "rmarkdown", "Summary of Predictor (Toes) and Response (Water)",
    split.table = Inf)
```

Table 1: Summary of Predictor (Toes) and Response (Water)

	Minimum	Median	Mean	Maximum	Variance
Water	0	0.001	0.016	0.137	0.001
Toes	0.07	0.18	0.37	2.25	0.24

Initial Modeling

By assigning Water as predictor variable and Toes as response variable, we have the scatterplot as shown in figure 2. The majority of data cluster together where arsenic concentration in well water is below 0.05, except one data point on the upper right cornor where arsenic concentration in well water is above 0.1. By looking at the scatterplot, we assume that Toes and Water might have positive relationship. So, we fit our initial linear model such that $Toes = 0.16 + 12.99 \times Water$, which means for one unit change in arsenic concentration in well water, the arsenic concentation in toenails is expected to increase by 12.99. If arsenic concentration in well water is 0, the arsenic concentration in toesnails, on average, is expected to be 0.16. Accordingly, we have the regression line as shown in figure 2. Because the sample size is small (n = 21), we can't see a clear linear pattern from the scatterplot, even though it appears to be linear. So, we should check the validity of this linear model.

```r
model.1 <- lm(Toes ~ Water, data = arsenic)
summary(model.1)$coefficient
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.16      0.054     2.9  9.7e-03
Water          12.99      1.473     8.8  3.8e-08
```

```r
ggplot(arsenic, aes(x = Water, y = Toes)) + geom_point() + geom_smooth(method = "lm",
    se = FALSE) + labs(x = "Arsenic in Water", y = "Arsenic in Toes",
    title = "Scatterplot of Toes vs Water")
```

Diagnostics and Transformations

The diagnostics plot for initial model is shown in figure 3. Residuals vs fitted values plot shows some pattern unlike residual plot under homoscedasticity assumption. Also, the residuals' QQplot clearly doesn't follow the normality assumption line, especially on two tails, point 15, 17 and 14 largely deviate from the straight line, so the residual's normality assumption is violated. The scale-location plot also indicates that the variance is not constant as pattern exists. Furthermore, the residuals vs leverage plot suggests that point 14 is a potential influential point with large leverage score (0.7) and high residual score (>2), and point 17 is an outlier as it

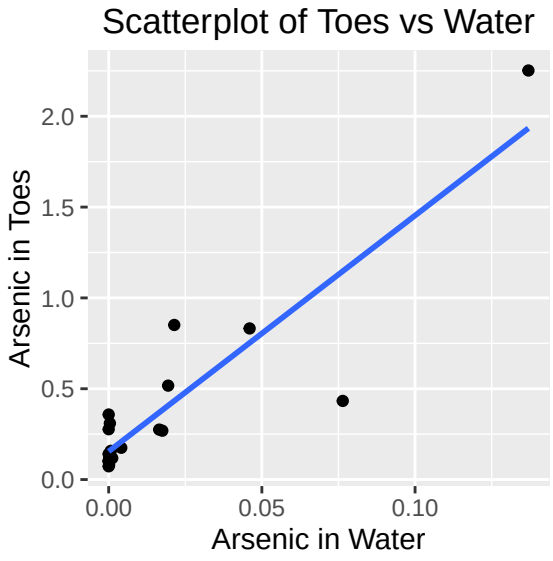

Figure 2: Scatterplot of Toes vs Water

has large residual score. According to the diagnostics plot, this model doesn't fit well as assumptions are violated. So, taking logarithms transformation on response variable or predictor variable or both might fix the problems of skewness, hetermoscedasticity and non-normality.

```
par(mfrow = c(1, 4))
plot(model.1, cex.caption = 0.7)
```

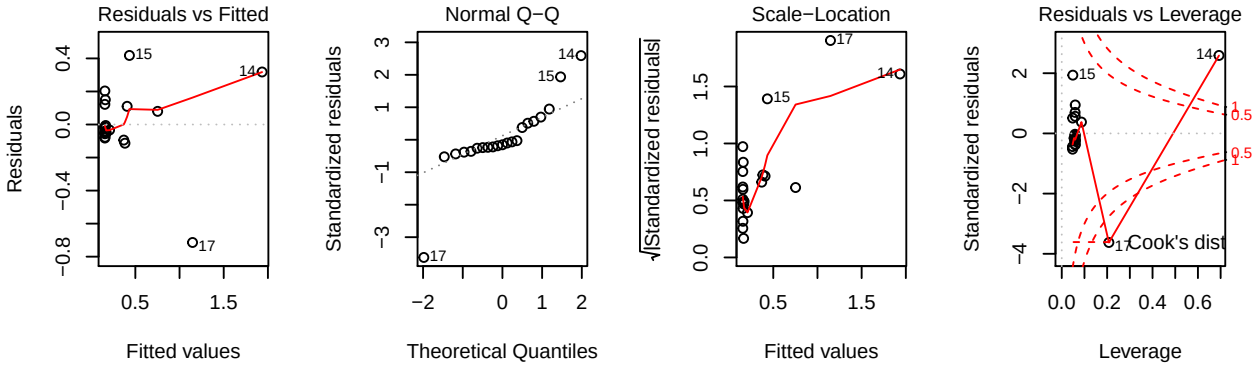

Figure 3: a) Residuals vs Fitted values, b) Residuals QQplot, c)Scale-Location plot, d)Residual vs Leverage

```
arsenic15 <- arsenic[-c(3, 5, 6, 10, 11, 18), ]  #15 samples with detectable arsenic concentrations
fit.1 <- lm(log(Toes) ~ Water, data = arsenic15)
fit.2 <- lm(Toes ~ log(Water), data = arsenic15)
fit.3 <- lm(log(Toes) ~ log(Water), data = arsenic15)
stargazer(fit.1, fit.2, fit.3, type = "latex", title = "Summaries of regression models",
    header = FALSE)
```

Then, we took logarithms transformation on the 15 samples with detectable arsenic concentration, the summaries of three linear models are shown in table 2. As we can see from the summary, R^2 is the largest in the third model ($R^2 = 0.71$), such that 71% of variability in log of arsenic concentration in Toenails can be

Table 2: Summaries of regression models

	Dependent variable:		
	log(Toes)	Toes	log(Toes)
	(1)	(2)	(3)
Water	20.000***		
	(4.000)		
log(Water)		0.150***	0.330***
		(0.048)	(0.058)
Constant	−1.700***	1.300***	0.550
	(0.170)	(0.290)	(0.350)
Observations	15	15	15
R^2	0.660	0.450	0.710
Adjusted R^2	0.630	0.400	0.690
Residual Std. Error (df = 13)	0.570	0.430	0.520
F Statistic (df = 1; 13)	25.000***	10.000***	32.000***
Note:			*p<0.1; **p<0.05; ***p<0.01

explained by log of arsenic concentration in well water under this regression model. In addition, by comparing at scatterplot of response variable and predictor variable from three log-transformed models, shown in figure 4, we can see that the third model depicts linearilty best, because the points evenly spread out across the regression line. As comparison, the first log-transformed model shows points cluster together where arsenic concentration in toenails is below 1 except one point at upper right cornor. The second log-transformed model shows that points are lined up vertically on the left hand side where arsenic concentration in water is below 0.04. So, we decided to build linear model by log-transforming both response and predictor variables.

```
par(mfrow = c(1, 3))
plot(log(arsenic15$Water), arsenic15$Toes, xlab = "log(Water)", ylab = "Toes",
    main = "Toes vs log(Water)")
abline(1.3, 0.15, col = "red")
plot(arsenic15$Water, log(arsenic15$Toes), xlab = "Water", ylab = "log(Toes)",
    main = "log(Toes) vs Water")
abline(-1.7, 20, col = "red")
plot(log(arsenic15$Water), log(arsenic15$Toes), xlab = "log(Water)", ylab = "log(Toes)",
    main = "log(Toes) vs log(Water)")
abline(0.55, 0.33, col = "red")
```

After transformation, we checked the validity of this model. Based on the diagnostics plot in fiture 5, we can see that the residuals' QQplot in fiture 5b, follows the normality assumption line except points on left tail deviate a little from the straight line. The residuals plot in figure 5a and 5c seem to show some patterns, but after conducting studentized Breusch-Pagan test, p-value is 0.2 which is greater than $\alpha = 0.05$, so the equal variance assumption holds and the pattern shown in residuals plot is acceptable. In addition, the residuals vs leverage plot suggests that point 14 is potential outlier with high residual.

```
par(mfrow = c(1, 4))
plot(fit.3, cex.caption = 0.7)
```

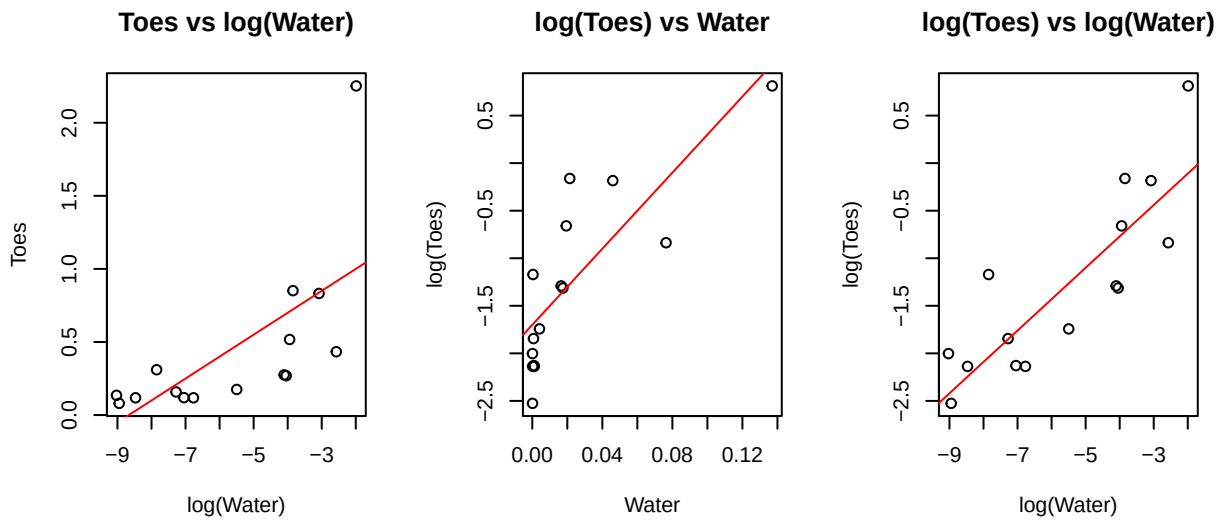

Figure 4: a) Toes vs log(Water), b) log(Toes) vs Water, c) log(Toes) vs log(Water)

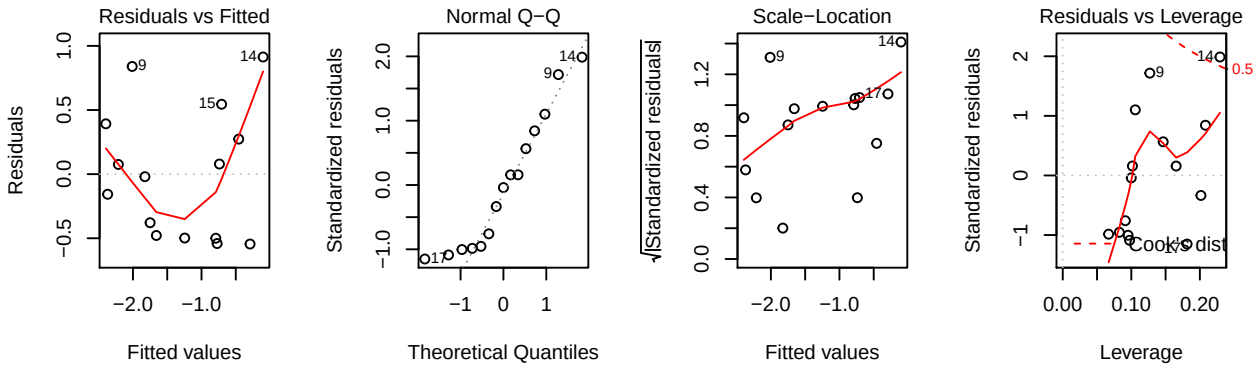

Figure 5: a) Residuals vs Fitted values, b) Residuals QQplot, c)Scale-Location plot, d)Residual vs Leverage

```
bptest(fit.3)
```

```
    studentized Breusch-Pagan test

data:  fit.3
BP = 2, df = 1, p-value = 0.2
```

Model Interfence and Results

Based on the diaganostics analysis of 15 samples who had detectable water levels of arsenic, the ideal linear model for investigating the relationship between arsenic concentration in well water and arsenic concentration in toenails is $log(Toes) = 0.55 + 0.33 \times log(Water)$, shown in table 3. Namely, given 10% increase in arsenic concentration in well water, the arsenic concentration in toenails is expected to increase by 3.2%, $(1.1^{\beta_1}) - 1 = (1.1^{0.33}) - 1 = 0.032$. Also, We are 95% confident that the true proportional increase in arsenic concentration in toenails is between 1.9% and 4.4%, given 10% increase in arsenic concentration in well water that household consumes. In addition, our intial hypothesis is whether the arsenic concentration in toenails can be reflected by arseninc concentration in well water that hourseholds use, because the p-value for estimated coefficient of log(Water) is less than $\alpha = 0.05$, we rejected the null hypothesis and concluded that the log of arsenic concentration in well water has positive linear relationship with log of arsenic concentration in toenails.

```
pandoc.table(summary(fit.3)$coefficient, "Summary of log-transformed model",
    split.cells = c("50%", "12.5%", "12.5%", "12.5%", "12.5%"))
```

Table 3: Summary of log-transformed model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5458	0.3533	1.545	0.1463
log(Water)	0.3257	0.05796	5.62	8.344e-05

According to the research, the arsenic concentration in toenails may be associated with participants' characteristics like age, gender and frequency of using well water for cooking or drinking. Next, we focused on whether these confounding variables will change the relationship between arsenic concentration in well water and in toenails. Fiture 6 generates the scatterplot of log(Toes) vs log(Water) given by four categories and we can see that, for each category, the categorical points seem to spread out across the range of log(Water), no cluster in a particular level is identified. We assumed that the confounding variables wouldn't change the positive linear relationship between log(Water) and log(Toes) as we analyzed previously.

```
ageCategory <- function(x) {
    # category of age
    if (x < 40)
        return("<40") else if (x >= 40 && x <= 59)
        return("40-59") else if (x > 59)
        return(">=60")
}
arsenic15$age <- apply(arsenic15[1], 1, ageCategory)  #create a new column for category of age
a <- ggplot(arsenic15, aes(x = log(Water), y = log(Toes), fill = factor(age),
    colour = factor(age))) + geom_point() + labs(title = "age")
```

```
b <- ggplot(arsenic15, aes(x = log(Water), y = log(Toes), fill = factor(Drink),
    colour = factor(Drink))) + geom_point() + labs(title = "Frequency of drinking water")
c <- ggplot(arsenic15, aes(x = log(Water), y = log(Toes), fill = factor(Cook),
    colour = factor(Cook))) + geom_point() + labs(title = "Frequency of cooking")
d <- ggplot(arsenic15, aes(x = log(Water), y = log(Toes), fill = factor(Sex),
    colour = factor(Sex))) + geom_point() + labs(title = "Gender")
grid.arrange(a, b, c, d, nrow = 2, ncol = 2, newpage = FALSE)
```

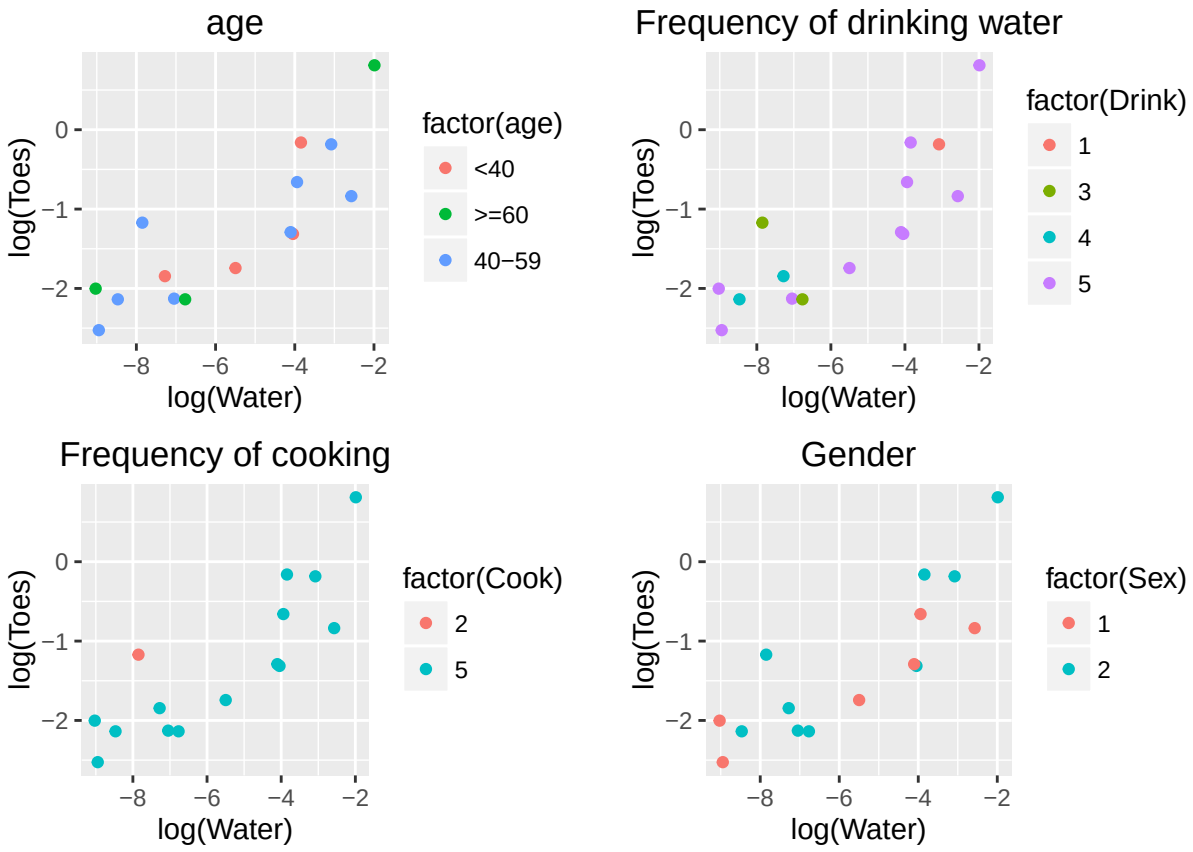

Figure 6: Scatter Plot by Category

Furthermore, we took a closer look at the confounding variables by including them selectively into the linear model, as shown in table 4. The table representing multiple linear regression suggests that p-values of log(Water), in all models, are less than $\alpha = 0.05$ and the estimated coefficients of log(Water) are all statistically significant and positive, which indicates that by adding up confounding variables in the model won't change the relationship between log(Water) and log(Toes). In addition, table 5 shows ANOVA summary by comparing with model with and without confounding variables. As F statistics is small and p-value is greater than $\alpha = 0.05$, we didn't reject the null. Therefore, we can assume that the model including these confounding variables doesn't appear more valid than the model that omits them based on the 15 samples that have detectable arsenic concentration.

```
try.1 <- lm(log(Toes) ~ log(Water) + age, data = arsenic15)
try.2 <- lm(log(Toes) ~ log(Water) + factor(Drink), data = arsenic15)
try.3 <- lm(log(Toes) ~ log(Water) + factor(Cook), data = arsenic15)
try.4 <- lm(log(Toes) ~ log(Water) + factor(Sex), data = arsenic15)
try.5 <- lm(log(Toes) ~ log(Water) + age + factor(Drink) + factor(Cook),
```

```
    data = arsenic15)
try.6 <- lm(log(Toes) ~ log(Water) + age + factor(Drink) + factor(Cook) +
    factor(Sex), data = arsenic15)
stargazer(try.1, try.2, try.3, try.4, try.5, try.6, type = "latex",
    title = "Summaries of multiple linear regression models",
    header = FALSE, font.size = "tiny", column.sep.width = "0.2pt")
```

Table 4: Summaries of multiple linear regression models

	Dependent variable:					
	log(Toes)					
	(1)	(2)	(3)	(4)	(5)	(6)
log(Water)	0.330^{***}	0.330^{***}	0.350^{***}	0.320^{***}	0.360^{***}	0.340^{***}
	(0.061)	(0.077)	(0.055)	(0.057)	(0.055)	(0.058)
age>=60	0.410				0.880^{**}	0.890^{*}
	(0.420)				(0.360)	(0.370)
age40-59	0.093				−0.058	0.021
	(0.340)				(0.270)	(0.290)
factor(Drink)3		−0.063			-1.600^{*}	-1.600^{*}
		(0.780)			(0.690)	(0.700)
factor(Drink)4		−0.210			−0.130	−0.150
		(0.800)			(0.590)	(0.600)
factor(Drink)5		−0.330			−0.480	−0.330
		(0.630)			(0.460)	(0.500)
factor(Cook)5			-0.960^{*}		-2.300^{**}	-2.200^{**}
			(0.510)		(0.660)	(0.690)
factor(Sex)2				0.340		0.250
				(0.270)		(0.300)
Constant	0.450	0.840	1.600^{**}	0.330	3.300^{***}	2.800^{**}
	(0.410)	(0.630)	(0.640)	(0.380)	(0.870)	(1.000)
Observations	15	15	15	15	15	15
R^2	0.730	0.720	0.780	0.740	0.910	0.920
Adjusted R^2	0.660	0.610	0.740	0.700	0.810	0.810
Residual Std. Error	0.540 (df = 11)	0.580 (df = 10)	0.480 (df = 12)	0.510 (df = 12)	0.400 (df = 7)	0.410 (df = 6)
F Statistic	10.000^{***} (df = 3; 11)	6.600^{***} (df = 4; 10)	21.000^{***} (df = 2; 12)	17.000^{***} (df = 2; 12)	9.800^{***} (df = 7; 7)	8.300^{***} (df = 8; 6)
Note:						*p<0.1; **p<0.05; ***p<0.01

```
pandoc.table(anova(fit.3, try.6), style = "rmarkdown", "ANOVA Table", split.table = Inf,
    justify = "left")  #fit.3 is reduced model, try.6 is hypothesized full model
```

Table 5: ANOVA Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
13	3.563	NA	NA	NA	NA
6	1.016	7	2.547	2.15	0.185

To sum up, we collected 21 observations which include households' age and gender and frequency of cooking or drinking well water, and then we fitted the initial simple linear regression model by arsenic concentration in toenails (Toes) versus arsenic concentration in well water (Water). However, due to the equal variance assumption and normality assumption were both violated in this model, we decided to take logarithms transformation on both response and predictor variables, by using 15 samples with detectable arsenic concentraiton. The transformed linear model turns out to be linear and it supports both equal variance and normality assumptions, so we concluded that the arsenic concentration in toenails can be reflected by the arsenic concentration in well water that households consumes, and their relationship appears to be positively linear. In addition, after analyzing the potential lurking variables, age, gender and frequency of cooking and drinking in the multiple regression models, we concluded that by adding up these confounding variables won't change the positive linear relationship between arsenic concentration in toenails and arsenic concentration in well water, and it is acceptable to omit them from the model.

On the other hand, this research only includes 21 samples from households and corresponding private wells, which limits the objectiveness of the linear model. For example, observations from age less than 40 and age greater than 60 are too few to compare with observation from age 40-59. Same problem applies to frequency of drinking and cooking, the observations from different levels are not balanced, which leads to skewness and bias in the model. Furthermore, the reason why we used 15 samples with detectable arsenic concentration instead of total samples is because log-transformation can't apply to value of 0 in well water's arsenic concentration. We built our linear model based on detectable samples but didn't infer the situation when arsenic in well water is 0 ppm while arsenic concentration exists in toenails. Moreover, the study of research didn't collect the information regarding individuals' diet, tobacco use and water consumption, which may contribute to the arsenic concentration in their toenails as well. They may also downplay the effect of arsenic concentration in well water at the same time. Therefore, for future study, the research should analyze participants' diet, tobacco use, water comsumption and other lurking variables, and collect the dataset regardless of geological constraint.