MWSUG 2016 - Paper PO13

Regression Analysis of the Levels of Chlorine in the Public Water Supply in Orange County, FL

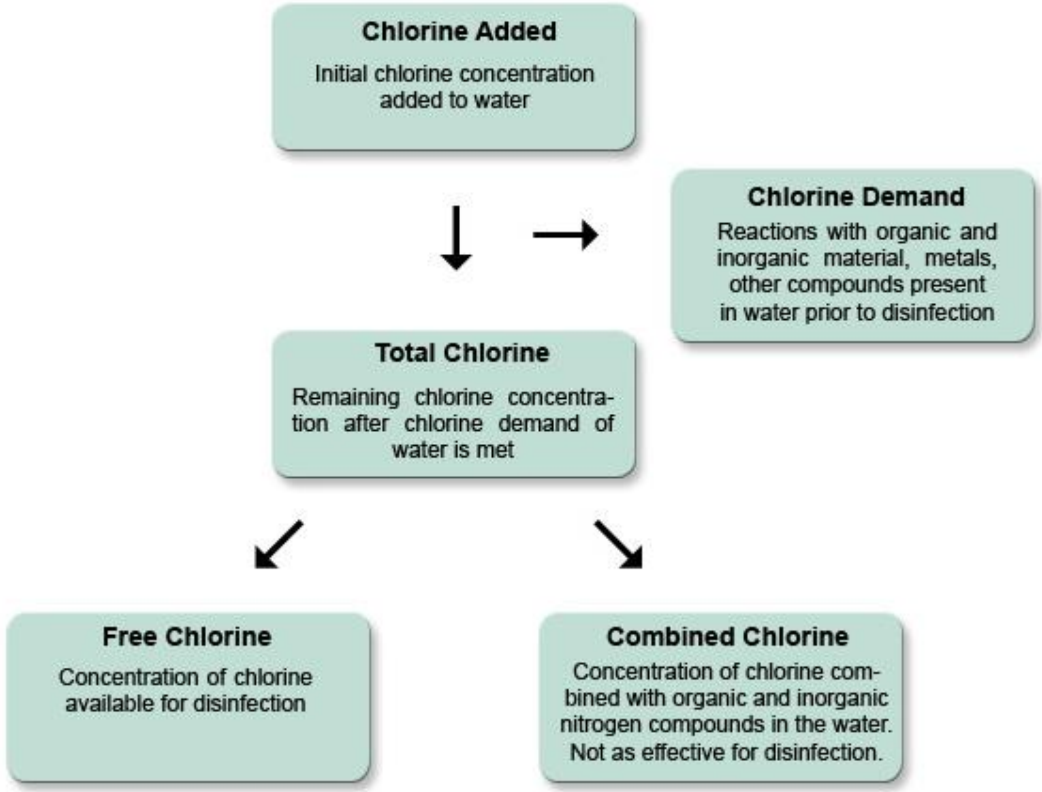
Drew Doyle, University of Central Florida, Orlando, FL

ABSTRACT

Public water supplies contain disease-causing microorganisms in the water or transport ducts. In order to kill off these pathogens, a disinfectant, such as chlorine, is added to the water. Chlorine is the most widely used disinfectant in all U.S. water treatment facilities. Chlorine is known to be one of the most powerful disinfectants to restrict harmful pathogens from reaching the consumer. In the interest of obtaining a better understanding of what variables affect the levels of chlorine in the water, this thesis will analyze a particular set of water samples randomly collected from locations in Orange County, Florida. Thirty water samples will be collected and have their chlorine level, temperature, and pH recorded. The chlorine levels will be read by a LaMotte Model DC1100 Colorimeter and will output the amount of chlorine in parts per million (ppm). This colorimeter will read the total chlorine of the sample, including both free and combined chlorine levels. A linear regression analysis will be performed on the data collected with several qualitative and quantitative variables. Water age, temperature, time of day, location, pH, and dissolved oxygen level will be the independent variables collected from each water sample. All data collected will be analyzed through various Statistical Analysis System (SAS®) procedures. Partial residual plots will be used to determine possible relationships between the chlorine level and the independent variables and stepwise selection to eliminate possible insignificant predictors. From there, several possible models for the data will be selected. F tests will be conducted to determine which of the models appears to be the most useful. All tests will include hypotheses, test statistics, p values, and conclusions. There will also be an analysis of the residual plot, jackknife residuals, leverage values, Cook's D, press statistic, and normal probability plot of the residuals. Possible outliers will be investigated and the critical values for flagged observations will be stated along with what problems the flagged values indicate. A nonparametric regression analysis can be performed for further research of the existing data.

INTRODUCTION

Public water supplies contain disease-causing microorganisms in the water or transport ducts. In order to kill off these pathogens, a disinfectant, such as chlorine, is added to the water. "Disinfection is the last treatment stage of a Drinking Water Treatment Plant (DWTP) and is carried out to maintain a residual concentration of disinfectant in the water distribution system" (Sorlini, 2014). The introduction of water disinfectants in the 20th century was considered to be one of the greatest progressions in health decreasing both typhoid and cholera outbreaks (Lyon, 2014). Chlorine is the most widely used disinfectant in all U.S. water treatment facilities. "Chlorine is still an indispensable disinfection agent because of the assurance of a high microbiological stability of water in the distribution subsystem…" (Zimoch, 2014). Chlorine is used as a disinfectant for a variety of reasons. "As a chemical disinfectant, chlorine has been applied to treat potable water widely because it is relatively cheap and effective." (Wang) Chlorine is known to be one of the most powerful disinfectants to restrict harmful pathogens from reaching the consumer. "While disinfectants have provided a novel method as a means to clean water, their usage leads to the formation of unwanted drinking water disinfection by-products (DBPs)" (Ali, 2014). These DBP's can form from the interaction between the disinfectant and the organic materials naturally within the water. By trying to eliminate harmful pathogens from our water supply, we are creating a new threat that our bodies must defend against. "Several epidemiological studies have shown that consumption or exposure to water above the maximum containment levels of DBPs in water have been associated with problems of liver, kidney, the central nervous system and increased risks of bladder, and colorectal cancers" (Ali, 2014). If someone has to choose, people are better off drinking elevated DBPs than they are drinking inadequately disinfected water. This method of cleansing the water is not perfect, but it is better than not disinfecting the water at all.

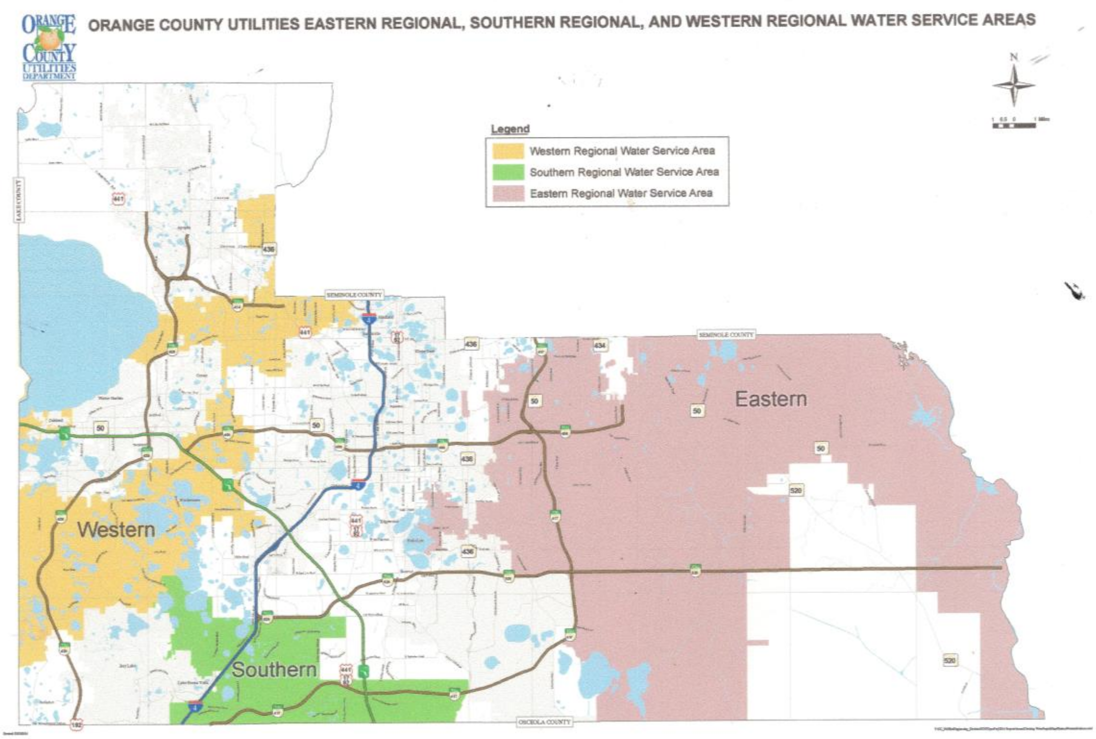


This image above, provided by the Centers for Disease Control and Prevention, summarizes what happens to chlorine when it is added to the water. When chlorine is added to the water it is broken into Chlorine Demand and Total Chlorine. The Total Chlorine is separated into two categories: Free Chlorine and Combined Chlorine. The Combined Chlorine is where the DBP's are formed when the chlorine reacts with the other compounds present in the water. The Combined Chlorine is not as effective for disinfecting the water, unlike the remaining Free Chlorine.

Several variables can affect the chlorine currently in the water, whether they increase or decrease the amount of chlorine. Ideally a consumer would like to decrease the amount of chlorine in their water before consuming or using it. "Chlorine decays in water because of its reactions with inorganic and organic solutes that impose chlorine demands." (Liu, 2014). The amount of chlorine in the water will decrease as it reacts with the microorganisms present in the water. "Chlorine loss in aged samples (samples left in open bottles) was greatest (approximately 40 mg/L free chlorine loss in 24 h) in low pH (approximately 2.5) and high chloride (Cl-) concentrations (greater than 150 mg/L)" (Waters, 2014). As water is left to sit, the amount of chlorine present should decrease. Chlorine levels should be lower when the pH level is more acidic.

METHODOLOGY

In the interest of obtaining a better understanding of what variables affect the levels of chlorine in the water, this paper will analyze a particular set of water samples randomly collected from locations in Orange County, Florida. Thirty water samples will be collected and have their chlorine level, temperature, pH, and dissolved oxygen level recorded. The chlorine levels will be read by a LaMotte Model DC1100 Colorimeter and will output the amount of chlorine in parts per million (ppm). This colorimeter will read the total chlorine of the sample, including both free and combined chlorine levels. The collected data "tells us about how one or more factors might influence the variable of interest" (Bowerman, 2014). In this research the variable of interest is the chlorine level of the water for Orange County, FL.



"Regression analysis answers questions about the dependence of a response variable on one or more predictors, including prediction of future values of a response, discovering which predictors are important, and estimating the impact of changing a predictor or a treatment on the value of the response." (Weisberg, 2005) A Simple Linear Regression model will be performed on the data collected with several qualitative and quantitative variables. Sample storage time, temperature, time of day, location, pH, and dissolved oxygen level will be the independent variables collected from each water sample. Water age refers to the amount time between when the water leaves the treatment plant and reaches its point of extraction. The sample storage time variable will be counted as the number of hours between water sample collection and chlorine level reading. For this particular analysis, water age will be ignored and sample storage time will be used instead. The time of day variable will be recorded as the number of minutes since noon. The location was recorded as the Eastern, Western, or Northern water treatment plant of Orange County, FL from which the water for sample came from. Two dummy variables will be created, E and W, to represent when the sample was taken from each of the treatment plants. All data collected will be analyzed through various Statistical Analysis System (SAS) procedures. Partial residual plots will be used to determine possible relationships between the chlorine level and the independent variables and stepwise selection to eliminate possible insignificant predictors. From there, several possible models for the data will be selected. F tests will be conducted to determine which of the models appears to be the most useful. There will also be an analysis of the residual plot, jackknife residuals, leverage values, Cook's D, press statistic, and normal probability plot of the residuals. Possible outliers will be investigated and the critical values for flagged observations will be stated along with what problems the flagged values indicate.

GETTING THE DATA INTO SAS

The first step is to correctly get your data into SAS. The first variable read in is Location for the treatment plant, which the water sample came from. A number one was used to represent water samples from the Eastern treatment plant of Orange County, a number two was used to represent water samples from the Western treatment plant of Orange County, and the number three was used to represent water samples from the Northern treatment plant of Orange County. The next variable read in is Time, for the time of day the sample was collected recorded as the number of minutes since noon. After that the storage time of the water sample, Storage, will be read in as the number of hours between collection and testing of the sample. The temperature of the water sample at time of sampling in degrees Celsius, Temp, is read in following Storage. The pH of the water sample is then read in with the typical 0-14 scale. The dissolved oxygen, in percent, of the water sample, DO, is read in preceding the pH variable. The last variable read in is the chlorine level, in ppm, under the variable name Chlor. An if-else statement is then used to create a dummy variable, E, for those samples from the Eastern water treatment plant. Another if-else statement is used to create a second dummy variable, W, for those samples from the Western water treatment plant.

```
DATA Chlorine;
INPUT Location Time Storage Temp pH DO Chlor;
 if Location=1 then E=1;
    else E=0;
 if Location=2 then W=1;
    else W=0;
DATALINES;
1   15     0      22.19 7.84  7.50  0.83
3   105    0      23.94 7.97  10.13 0.89
2   120    0      23.64 8.02  8.04  0.68
3   135    0      28.02 8.01  7.63  0.44
1   150    0      26.42 7.97  6.85  0.67
2   165    0      29.19 7.96  7.40  0.50
3   210    0      17.44 8.03  9.42  0.34
2   255    0      15.43 8.10  8.86  0.09
1   240    1      24.56 7.99  6.68  0.24
3   360    2      24.88 8.01  5.84  0.37
1   300    3      19.93 7.91  6.45  0.06
3   0      3      21.20 7.94  6.50  0.93
2   255    4      23.09 7.41  8.68  0.22
2   270    4      23.04 7.84  8.80  0.35
2   180    5      20.80 7.57  9.06  0.30
3   210    5      22.57 7.20  8.62  0.45
2   60     6      20.84 8.60  7.64  0.03
1   90     6      20.85 7.88  9.02  0.07
3   225    7      22.92 7.77  8.60  0.60
2   285    7      22.70 7.50  8.45  0.00
1   30     8      21.32 7.91  6.66  0.34
1   45     8      22.14 7.94  7.20  0.18
2   210    10     21.23 7.86  8.61  0.21
3   270    10     21.57 7.90  7.93  0.16
1   360    12     20.55 7.76  9.61  0.09
3   390    12     21.00 7.96  9.24  0.02
2   180    15     21.04 8.07  9.08  0.01
1   300    15     21.52 8.01  9.12  0.02
1   315    24     21.08 7.74  9.10  0.01
3   360    24     22.00 7.51  8.46  0.00
;
RUN;
```

Now that there was a successful creation of dummy variables for location, the Location variable will no longer be used in part of the analysis.

FINDING THE BEST MODEL

Through the stepwise selection method, the best model for this particular data will be chosen. Stepwise, backward, and forward selection will all be used to see if they all select the same model. In order to do so, PROC STEPWISE will be used. For this to work properly the model must have the dependent variable, Chlor, in this instance, set equal to each independent variable for which the user wants to include in the model. The model is followed by a forward slash and the options of the type of model selection the user would like. For this analysis, forward selection, backward elimination, and stepwise selection will be used, which means forward, backward, and stepwise must be included in the options. If these options are not included then the PROC will default to only running a stepwise selection. If the forward and backward options are included but the stepwise option is not, then the PROC will only run a forward selection and backward elimination. All three options should be included if the user wants all three selection methods to be used. This method can be a bit more challenging when working with dummy variables. Some users choose to run this PROC without incorporating the dummy variables and then adding them to the chosen models. Other users will run the PROC with the dummy variables and will add them to the model if all the dummy variables are not selected, or, they will create new dummy variables depending on the selection. In this case, the selection process is being run with the dummy variables and will be added to the model if only one is selected.

```
PROC STEPWISE;
    MODEL Chlor = Time Storage Temp pH DO E W / forward backward stepwise;
RUN;
```

After the PROC has run, then all of the steps of all of the selection methods will be shown. One must be careful when picking the selected method. Check the step number to be sure the last step of the selection is the chosen model, not the eliminated variables. There can be a lot of output depending on your data and variables, therefore only the summary of the selection tables will be shown below.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Storage	1	0.3743	0.3743	19.3482	16.75	0.0003
2	Time	2	0.1103	0.4846	13.3516	5.78	0.0233
3	Temp	3	0.0660	0.5506	10.5676	3.82	0.0615
4	W	4	0.0489	0.5995	9.0232	3.05	0.0929
5	E	5	0.0658	0.6653	6.2559	4.72	0.0400
6	pH	6	0.0254	0.6907	6.4167	1.89	0.1828

The forward selection chose the model containing the storage time, time of day, temperature of the sample, both dummy variables and pH. The variable DO was the only variable dropped from the complete model. From this table in the output, we can see the p-values for each one of the selected variables. Each has a p-value below an alpha of 0.10 except for the pH variable, this is because the forward selection uses an alpha of 0.50. Forward selection starts with no variables and adds variables one at a time. Most users do not use forward selection as their preferred method due to a low alpha level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	DO	6	0.0057	0.6907	6.4167	0.42	0.5253
2	pH	5	0.0254	0.6653	6.2559	1.89	0.1828

It appears that the backward elimination only selected a model with DO and pH. This is where one has to be careful. The summary shown above is telling the user what variables were eliminated from the model.

Therefore, the model that backward elimination chose contains time of day, storage time, temperature of the sample, and both dummy variables. Backward elimination starts with the full model and eliminates one variable at a time until the best model remains. Backward elimination compares each variable's p-value to an alpha of 0.10, which is why this time pH was eliminated from this model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Storage		1	0.3743	0.3743	19.3482	16.75	0.0003
2	Time		2	0.1103	0.4846	13.3516	5.78	0.0233
3	Temp		3	0.0660	0.5506	10.5676	3.82	0.0615
4	W		4	0.0489	0.5995	9.0232	3.05	0.0929
5	E		5	0.0658	0.6653	6.2559	4.72	0.0400

Through the stepwise selection the model containing the storage time, time of day, temperature, and location dummy variables were selected. This is the same model that was chosen by backward elimination. Stepwise selection compares each variable's p-value to an alpha of 0.15, which is why pH and DO were also eliminated from this model. Stepwise selection is the preferred method because it is similar to a combination of forward and backward selection. It starts with no variables in the model and adds one at a time, checking the new variable's p-value along with the variables already in the model.

Based on the selections listed above, the chosen model to analyze is the one containing the independent variables Time, Storage, Temp, E, and W.

ANALYZING THE BEST MODEL

In order to see if this model is useful we must check and analyze the conditions necessary for this to be true. A global F test will be done to see if the model is deemed useful. We will also investigate residual plots, jackknife residuals, leverage values, Cook's D, PRESS statistic, and normal probability plot of the residuals. Possible outliers will be flagged based on these findings. We will also look into any problems with collinearity between the variables. This will all be done using the code below.

```
PROC REG;
model Chlor = Time Storage Temp E W / partial influence VIF;
 output out=new cookd=cook rstudent=jack  h=lev r=resid;
RUN;
PROC PRINT data= new;
RUN;

PROC UNIVARIATE normal plot;
var resid;
RUN;

PROC CORR;
var Time Storage Temp E W;
RUN;
```

F TEST

Through PROC REG with the previously selected model one is able to perform a global F test on the model to test its significance.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1.54238	0.30848	9.54	<.0001
Error	24	0.77589	0.03233		
Corrected Total	29	2.31827			

This proposed model was deemed significant at an alpha of 0.01 with an F value of 9.54.

PREDICTION QUALITY

Through PROC REG with the previously selected model one is able to compute the mean square error and R-square values of the model to see how well the model predicts values.

Root MSE	0.17980	R-Square	0.6653
Dependent Mean	0.30333	Adj R-Sq	0.5956
Coeff Var	59.27542		

We expect about 95% of chlorine levels to fall within 2*0.17980 = 0.3596 ppm of the fitted regression equation. This model explains 66.5% of the observed variability in chlorine levels. This model also explains 59.6% of the observed variability in the chlorine levels after adjusting for the sample size of 30 and the 5 variables in the model.

PARAMETER ESTIMATES

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.21432	0.31375	0.68	0.5011	0
Time	1	-0.00108	0.00034548	-3.13	0.0045	1.30872
Storage	1	-0.01587	0.00590	-2.69	0.0128	1.37402
Temp	1	0.02442	0.01288	1.90	0.0700	1.07951
E	1	-0.18007	0.08291	-2.17	0.0400	1.41762
W	1	-0.21980	0.08128	-2.70	0.0124	1.36223

As the amount of minutes since noon increases, the estimated mean chlorine level decreases by 0.00108 ppm. As the number of hours between sample collection and testing increases, the estimated mean chlorine level decreases by 0.01587 ppm. As the temperature of the water increases, the estimated mean chlorine level increases by 0.024442 ppm. If a sample was from the eastern region, the estimated mean chlorine level is 0.180007 ppm less. If a sample was from the western region then the estimated mean chlorine level is 0.21980 ppm less.

PRESS STATISTIC

Sum of Residuals	0
Sum of Squared Residuals	0.77589
Predicted Residual SS (PRESS)	1.20403

It is ideal to have a small PRESS statistic value and in this particular case the PRESS statistic is 1.20. The PRESS statistic is similar to the R-square value in respect to saying how well the model explains the observed variability.

OUTLIERS

Using PROC REG we can also check for possible outliers. This code is using an output option to extract and rename the output of interest for analyzing residuals. These variables are saved into a new data set and printed out.

Obs	Location	Time	Storage	Temp	pH	DO	Chlor	E	W	resid	cook	lev	jack
1	1	15	0	22.19	7.84	7.50	0.83	1	0	0.27010	0.11341	0.19528	1.74439
2	3	105	0	23.94	7.97	10.13	0.89	0	0	0.20469	0.04639	0.15379	1.25208
3	2	120	0	23.64	8.02	8.04	0.68	0	1	0.23805	0.05082	0.13129	1.45296
4	3	135	0	28.02	8.01	7.63	0.44	0	0	-0.31248	0.23197	0.25545	-2.16296
5	1	150	0	26.42	7.97	6.85	0.67	1	0	0.15290	0.04216	0.21537	0.95837
6	2	165	0	29.19	7.96	7.40	0.50	0	1	-0.02879	0.00337	0.34162	-0.19333
7	3	210	0	17.44	8.03	9.42	0.34	0	0	-0.07295	0.01740	0.30570	-0.47904
8	2	255	0	15.43	8.10	8.86	0.09	0	1	-0.00537	0.00018	0.41668	-0.03825
9	1	240	1	24.56	7.99	6.68	0.24	1	0	-0.11841	0.02202	0.19659	-0.72750
10	3	360	2	24.88	8.01	5.84	0.37	0	0	-0.03057	0.00197	0.23779	-0.19078
11	1	300	3	19.93	7.91	6.45	0.06	1	0	-0.08868	0.01630	0.23520	-0.55576
12	3	0	3	21.20	7.94	6.50	0.93	0	0	0.24558	0.16298	0.27530	1.66230
13	2	255	4	23.09	7.41	8.68	0.22	0	1	0.00105	0.00000	0.12115	0.00611
14	2	270	4	23.04	7.84	8.80	0.35	0	1	0.14851	0.01931	0.12887	0.88079
15	2	180	5	20.80	7.57	9.06	0.30	0	1	0.07168	0.00368	0.11008	0.41525
16	3	210	5	22.57	7.20	8.62	0.45	0	0	-0.00888	0.00005	0.10159	-0.05100
17	2	60	6	20.84	8.60	7.64	0.03	0	1	-0.31329	0.14253	0.18643	-2.05785
18	1	90	6	20.85	7.88	9.02	0.07	1	0	-0.28080	0.07509	0.13744	-1.75259
19	3	225	7	22.92	7.77	8.60	0.60	0	0	0.18055	0.02114	0.10154	1.06220
20	2	285	7	22.70	7.50	8.45	0.00	0	1	-0.12935	0.01422	0.12593	-0.76274
21	1	30	8	21.32	7.91	6.66	0.34	1	0	-0.05547	0.00474	0.19422	-0.33725
22	1	45	8	22.14	7.94	7.20	0.18	1	0	-0.21926	0.06335	0.17427	-1.36596
23	2	210	10	21.23	7.86	8.61	0.21	0	1	0.08299	0.00565	0.12249	0.48483
24	3	270	10	21.57	7.90	7.93	0.16	0	0	-0.13018	0.01263	0.11358	-0.76224
25	1	360	12	20.55	7.76	9.61	0.09	1	0	0.13394	0.02766	0.19418	0.82430
26	3	390	12	21.00	7.96	9.24	0.02	0	0	-0.09465	0.01263	0.18263	-0.57409
27	2	180	15	21.04	8.07	9.08	0.01	0	1	-0.06548	0.00754	0.21183	-0.40301
28	1	300	15	21.52	8.01	9.12	0.02	1	0	0.02293	0.00059	0.15614	0.13597
29	1	315	24	21.08	7.74	9.10	0.01	1	0	0.18274	0.11675	0.31667	1.24337
30	3	360	24	22.00	7.51	8.46	0.00	0	0	0.01890	0.00163	0.36088	0.12873

There were no observations that were flagged as possible outliers with the dependent or independent variables.

VIF

Variance Inflation
0
1.30872
1.37402
1.07951
1.41762
1.36223

The variation inflation factor was attached to the previous table for the parameter estimates. Small Variance Inflation Factors for all variables in the model, which tells us that there are no problems with collinearity between the independent variables.

PEARSON CORRELATION COEFFICIENTS

Another method to check for any collinearity between the variables is by using PROC CORR to create a correlation matrix.

Pearson Correlation Coefficients, N = 30 Prob > |r| under H0: Rho=0					
	Time	Storage	Temp	E	W
Time	1.00000	0.44215 0.0144	-0.12659 0.5050	-0.12034 0.5264	-0.03253 0.8645
Storage	0.44215 0.0144	1.00000	-0.25474 0.1743	0.14451 0.4461	-0.13728 0.4694
Temp	-0.12659 0.5050	-0.25474 0.1743	1.00000	-0.04824 0.8002	-0.03649 0.8482
E	-0.12034 0.5264	0.14451 0.4461	-0.04824 0.8002	1.00000	-0.50000 0.0049
W	-0.03253 0.8645	-0.13728 0.4694	-0.03649 0.8482	-0.50000 0.0049	1.00000

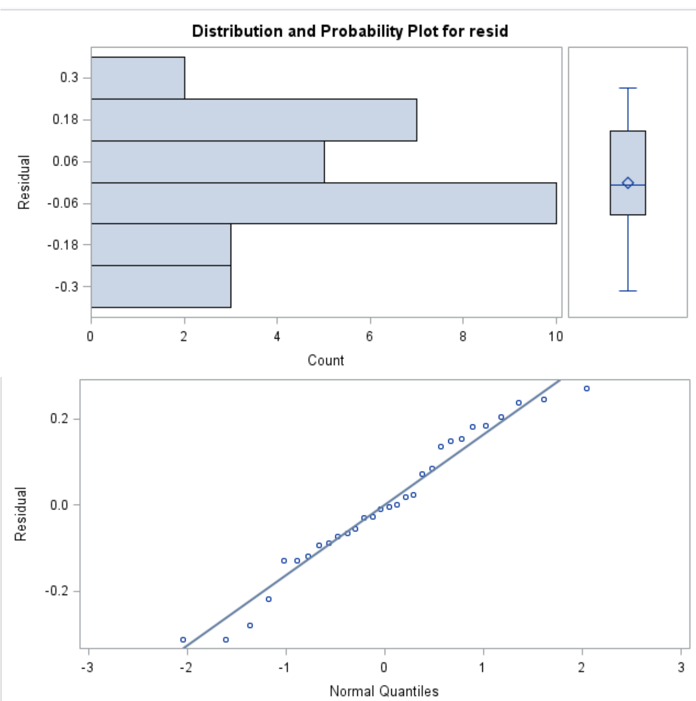
Each box gives the correlation coefficients between the two variables and below it the corresponding p-values. A small p-value tells us that the variables are correlated with one another. Ideally, we do not want them correlated with each other because this means they affect each other. The following variables are significantly correlated with one another: Time and Storage, East and West. Time and Storage could affect each other due to the fact that it was easier for a sample to have a long storage time when it was collected early in the day. This may be something to fix if further data collection is done. The two dummy variables cannot really affect each other because they cannot occur at the same time. A sample cannot be from both the eastern and western water treatment plant.

NORMAILTY

We want to test to see if the residuals are normally distributed. Using PROC UNIVARIATE we can look at the plots of the residuals and hypothesis tests for normality.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.963015	Pr < W	0.3690
Kolmogorov-Smirnov	D	0.093572	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.038275	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.301686	Pr > A-Sq	>0.2500

According to both the Shapiro-Wilk and Kolmogorov-Smirnov tests for normality, we can say the distribution of the residuals is normal. Both produce a test statistic with a p-value greater than an alpha of 0.15, which means we cannot reject the null hypothesis that the residuals are normally distributed.



We next look at the histogram and box plot of the residuals to check for normality. We can see that both are approximately normal.

The points on the normal quartiles chart should form a linear shape. The points do form roughly a linear shape in the graph above.

CONCLUSION

The assumptions for the regression analysis held for this chlorine model. When a water sample is collected later in the day, it will read a lower total chlorine level. Overall, a water sample with a lower temperature will have a lower chlorine level. The longer a water sample is left to sit out it will read a lower total chlorine level. The western region contains, on average, the least amount of chlorine in comparison to the eastern and northern regions. The northern region contains higher chlorine levels than the western and eastern regions.

FUTURE RESEARCH

A nonparametric regression analysis can be performed for further research of the existing data. A nonparametric analysis is appropriate if the data contains outlier that may be inaccurate, but there is insufficient evidence to remove the data points. The parametric and nonparametric regressions will be compared with each other to see which is a better predictor of the chlorine level. "…seasonal changes in temperature (as well seasonal changes in precipitation) can contribute to the variability in municipal drinking water quality" (Dyck, 2015). Data can be collected throughout the year, for a total of 12 months. By doing so, one can observe any seasonal relationship between the season and the chlorine level. Due to seasonal changes in temperature and precipitation the levels of chlorine in the water could also be affected. This change is worth investigating to see if it is significant in the regression model for predicting the chlorine levels. Water systems try to maintain an effect chlorine level throughout the entire water system. "This requires a much higher concentration of chlorine at entry than the concentration that is to be achieved at the extremities," (Fisher, 2010). There can be a measureable difference in chlorine levels between water samples collected near the water treatment plants and those further away. This could lead to the addition of a distance variable to account for a water sample's location in comparison to the water treatment plant. By contacting the water treatment plants the estimated water age of the samples can be collected and used to see if it is influential in predicting the levels of chlorine

REFERENCES

Ali, Aftab, Malgorzata Kurzawa-Zegota, Mojgan Najafzadeh, Rajendran C. Gopalan, Michael J. Plewa, and Diana Anderson. "Effect of Drinking Water Disinfection By-products in Human Peripheral Blood Lymphocytes and Sperm." *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 770 (2014): 136-43. Web. 15 Mar. 2015.

Dyck, Roberta, Geneviève Cool, Manuel Rodriguez, and Rehan Sadiq. "Treatment, Residual Chlorine and Season as Factors Affecting Variability of Trihalomethanes in Small Drinking Water Systems." Frontiers of Environmental Science & Engineering 9.1 (2015): 171-79. Print.

Fisher, Ian, George Kastl, and Arumugam Sathasivan. "A Suitable Model of Combined Effects of Temperature and Initial Condition on Chlorine Bulk Decay in Water Distribution Systems." *Water Research* 46.10 (2010): 3293-303. Web. 5 Mar. 2015.

"Free Chlorine Testing." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, 17 July 2014. Web. 20 Mar. 2015.

Liu, Boning, David A. Reckhow, and Yun Li. "A Two-site Chlorine Decay Model for the Combined Effects of PH, Water Distribution Temperature and In-home Heating Profiles Using Differential Evolution." *Water Research* 53 (2014): 47-57. Web. 10 Mar. 2015.

Lyon, Bonnie. "Integrated Chemical and Toxicological Investigation of UV-Chlorine/ Chloramine Drinking Water Treatment." Environmental Science & Technology 48.12 (2014): 6743-753. Print.

Sorlini, Sabrina, Francesca Gialdini, Michela Biasibetti, and Carlo Collivignarelli. "Influence of Drinking Water Treatments on Chlorine Dioxide Consumption and Chlorite/chlorate Formation."*Water Research* 54 (2014): 44-52. Web. 20 Mar. 2015.

Wang, Yifei, Aiyin Jia, Yue Wu, Chunde Wu, and Lijun Chen. "Disinfection of Bore Well Water with Chlorine Dioxide/sodium Hypochlorite and Hydrodynamic Cavitation." *Enivironmental Technology* 36.4 (2015): 479-86. Web. 20 Mar. 2015.

"Water Quality." *Water Quality*. N.p., n.d. Web. 29 Mar. 2015. <http://www.orangecountyfl.net/Water,GarbageRecycling/WaterQuality.aspx#.VUD5BK3BzGc>.

Waters, Brian W., and Yen-Con Hung. "The Effect of PH and Chloride Concentration on the Stability and Antimicrobial Activity of Chlorine-Based Sanitizers." *Journal of Food Science* 79 (2014): n. pag. *Biological Abstracts [EBSCO]*. Web. 13 Mar. 2015.

Weisberg, Sanford. Preface. *Applied Linear Regression*. 3rd ed. Hoboken: Wiley Series in Probability and Statistics, 2005. N. pag. Print.

Zimoch, Izabela. "The Optimization of Chlorine Dose in Water Treatment Process in Order to Reduce the Formation of Disinfection By-Products." Desalination and Water Treatment 52 (2014): 3719-724. Print.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Drew Doyle
Enterprise: University of Central Florida
Work Phone: (321) 230-4054
E-mail: drewdoyle@knights.ucf.edu