SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

To Hydrate or Chlorinate: A Regression Analysis of the Levels od Chlorine in the Public Water Supply

USERS PROGRAM

To Hydrate or Chlorinate:
A Regression Analysis of the Levels of Chlorine in the Public Water Supply

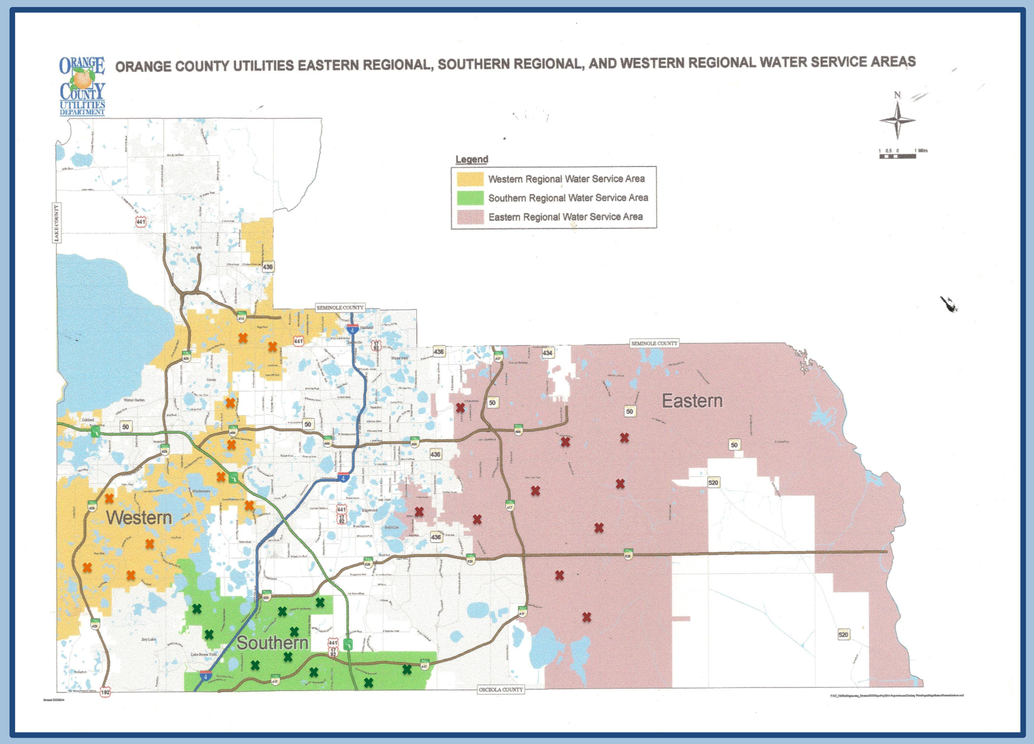
Drew A. Doyle
The University of Central Florida

ABSTRACT

This poster will analyze a particular set of water samples randomly collected from locations in Orange County, Florida. Thirty water samples were collected and had their chlorine level, temperature, and pH recorded. A linear regression analysis was performed on the data collected with several qualitative and quantitative variables. Water storage time, temperature, time of day, location, pH, and dissolved oxygen level were designated as the independent variables collected from each water sample. All data collected was analyzed through various Statistical Analysis System (SAS®) procedures. A partial residual plot was used for each variable to determine possible relationships between the chlorine level and the independent variables. Stepwise selection was used to eliminate possible insignificant predictors. From there, several possible models for the data were selected. F tests were conducted to determine which of the models appears to be the most useful. There was an analysis of the residual plot, jackknife residuals, leverage values, Cook's D, PRESS statistic, and normal probability plot of the residuals. Possible outliers were investigated and the critical values for flagged observations were stated along with what problems the flagged values indicate.

METHODS

A 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. 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.



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. In this research the variable of interest is the chlorine level of the water for Orange County, FL.

To Hydrate or Chlorinate:
A Regression Analysis of the Levels of Chlorine in the Public Water Supply

Drew A. Doyle
The University of Central Florida

GETTING THE DATA INTO SAS

The first step is to correctly get the 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 after 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
...(More data in here)*
3 360 24 22.00 7.51 8.46 0.00
;
RUN;
```

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.

$$\hat{y} = \beta_0 + \beta_1 Storage + \beta_2 Time + \beta_3 Temp + \beta_4 W + \beta_5 E$$

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

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.

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

This summary 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.

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.

To Hydrate or Chlorinate:
A Regression Analysis of the Levels of Chlorine in the Public Water Supply

Drew A. Doyle
The University of Central Florida

ANALYZING THE CHOSEN 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 following code.

```
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 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

Below are the parameter estimates of the chosen model.

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.

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.

To Hydrate or Chlorinate:
A Regression Analysis of the Levels of Chlorine in the Public Water Supply

Drew A. Doyle
The University of Central Florida

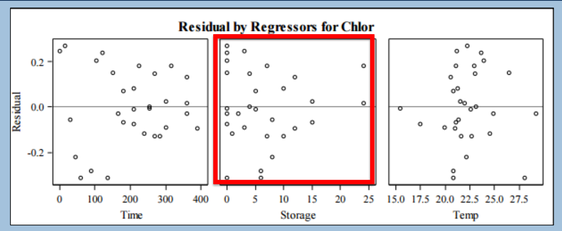
OUTLIERS

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.

Obs	resid	cook	lev	jack
1	0.27010	0.11341	0.19528	1.74439
2	0.20469	0.04639	0.15379	1.25208
3	0.23805	0.05082	0.13129	1.45296
4	-0.31248	0.23197	0.25545	-2.16296
.....				
29	0.18274	0.11675	0.31667	1.24337
30	0.01890	0.00163	0.36088	0.12873

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

RESIDUAL PLOTS



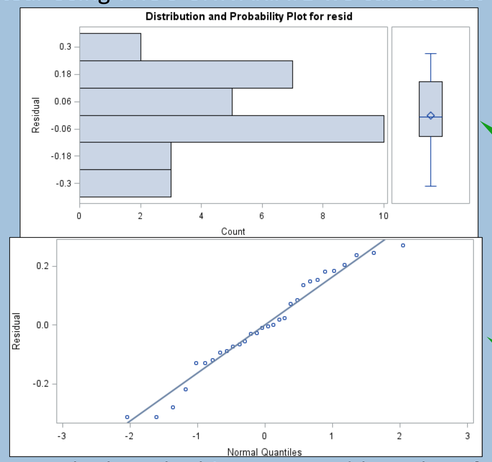
The normal plot of the residuals has a straight-line appearance. The plot of the residuals versus chlorine level has a vertical band appearance, as do the plots of the residuals versus the independent variables. We conclude that the regression assumptions approximately hold for the most part with the chlorine model.

NORMALITY

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 a mostly linear shape in the graph above.

PEARSON CORRELATION COEFFICIENTS

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

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. 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.

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

CONCLUSION

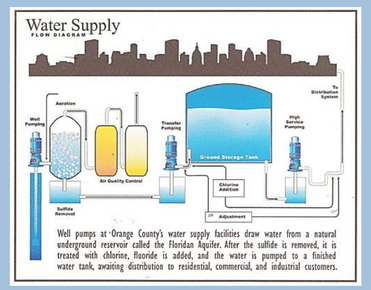
Most of the assumptions for the regression analysis held for this chlorine model. Based on the data and analysis, there was a negative correlation between when a water sample is collected later in the day and the total chlorine level. Overall, there is a positive correlation between a water sample's temperature and the total chlorine level. There is a negative correlation between a water sample's storage time and the 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. Further analysis on the data must be done in order to establish a possible cause and effect relationship between the independent and dependent variables. There was no testing of the interaction of the independent variables, which could explain some of the results.

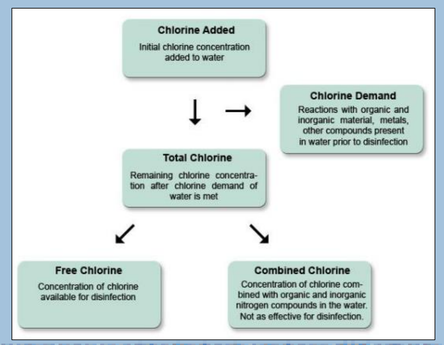
To Hydrate or Chlorinate:
A Regression Analysis of the Levels of Chlorine in the Public Water Supply

Drew A. Doyle
The University of Central Florida

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, 2015). 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. The interaction between the different independent variables should be investigated in order to see if these interactions lead to a better understanding of how they affect the chlorine levels. From the correlation matrix, one can see that adding an interaction between the storage time and the time of day or possibly of storage time and the temperature of the water sample. One could also test to see if there is a significant difference between the three different treatment areas. If there is a significant difference, one can look at each treatment area separately and see if this changes how the independent variables are affecting the total 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.

SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL