Memorandum

To: Carly Lesne
From: Lesley Arenas, Ryan Destefano, Lena Joseph, AJ Motter
Date: April 26, 2023
Re: CEC Wastewater Treatment Project: Statistical Analysis and Findings

---

The purpose of this memo is to describe the statistical methods and key findings from your CEC Pond 13-16 Weekly Average data. We hope that this information helps you address the following research questions for your thesis:

1. *"Is the overall average TAN removal for each group significantly different from one another?"*
2. *"What is the variation within each group?"*
3. *"What share of the variation in each group can be attributed to each of the explanatory variables?"*
4. *"Are there any issues with auto-correlation?"*
5. *"How to solve issues with auto-correlation?"*

This memo is organized into five sections:

- The first section, "**Abstract of Key Findings"**, includes an overview of key results in our analysis. (page 2)
- The second section, "**Background and Data"**, includes a summary of our understanding of your research questions and data. (page 3)
- The third section, "**Statistical Methods"**, provides a description of the models and methods for your consideration. (page 4)
- The fourth section, "**Results and Discussion"**, includes the interpretation of results from our analysis, limitations, and numerical and graphical summaries. (page 5)
- The fifth section, "**Technical Output"**, provides computer output for reference. (page 10)

  * We have presented you with statistical output from both the JMP and R statistical software, as we understand that these are the programs you are most familiar with.

For future meetings or questions regarding our analysis, please contact Professor Smith and Professor Glanz at hsmith@calpoly.edu and hglanz@calpoly.edu. Due to the nature of this course, we unfortunately are not able to meet for further follow up meetings.

Ⅰ. Abstract of Key Findings

We found a statistically significant difference in the TAN removal for the two levels of HRT (t-ratio = 4.78, p-value < .001). The average TAN removal for HRT=4 days is, on average, .902 greater than the average TAN removal for HRT=8 days, after accounting for all other explanatory variables in the model (AvgVSS, MinDailyDO, etc). We found that the average VSS explained the most variation in the TAN removal, at 15%, followed by HRT, at 13%. We did find that there were significant issues of autocorrelation present in the data, and the method of addressing it was to include a new variable, Month, of when the observation was collected. This allowed us to incorporate the effect of time into the analysis, and begin to address the autocorrelation in the data.

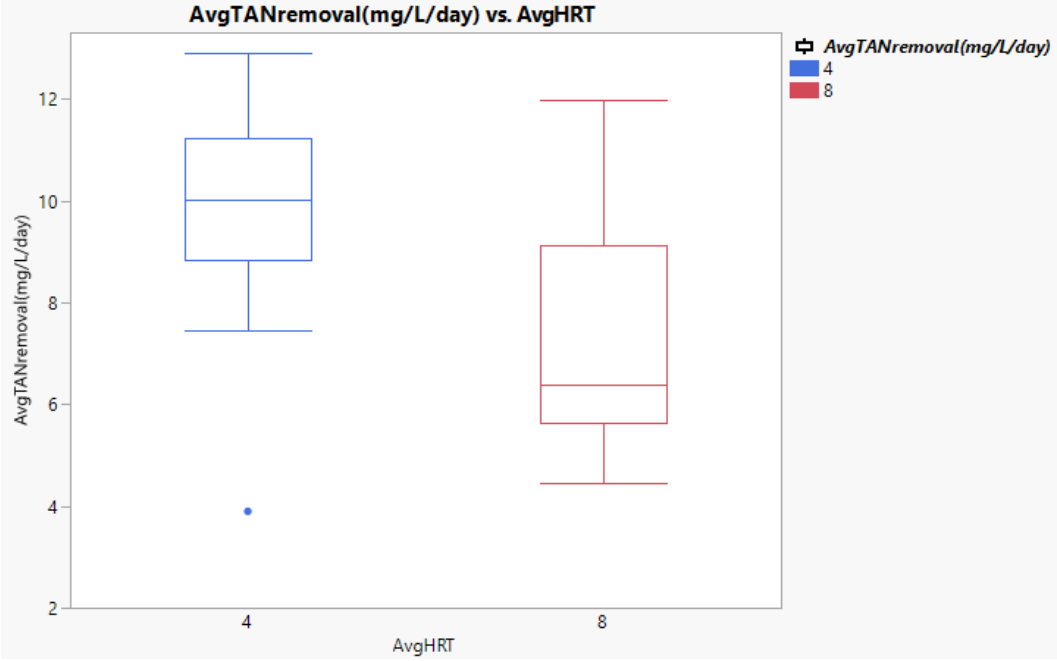


Figure 1: Comparing Average TAN removal distribution for the two levels of HRT.

Ⅱ. Background and Data

From our meeting on April 19th, we understand your project to be on the topic of improving the energy efficiency of wastewater treatment in California. You are working for the California Energy Commission, and your work concerns the removal of nutrients from wastewater that bacteria can use to grow, specifically the removal of nitrogen. Your project is investigating an alternate way of aerating the wastewater while it is being treated, through the use of algae. Your primary research goal was to investigate two different levels of the Hydraulic Residence Time (HRT), 4 and 8 days, to see if there was a significant difference in the Total Ammonia Nitrogen (TAN) removal for two different times to replace all the water in a tank with new influent water. You accomplished this by creating several different ponds of wastewater, assigning a different level of HRT to each set of 2 ponds. You also wanted to understand the variation in TAN removal within each of these two groups, as well as how much of this variation you can explain using several variables. These explanatory variables include:

- **VSS**: volatile suspended solids, a measure of the concentration of biomass in the ponds, with concentrations ranging from 150 to 1050 mg/L.
- **AvgMinDailyDO**: the minimum amount of dissolved oxygen in the ponds for each week, a method to measure the activity of the algae, with values ranging from 1-6 mg/L.
- **Temperature**: the average temperature in the ponds for each week, with temperatures ranging from 5 to 20 ℃..
- **pH**: the average pH of the ponds for each week, with values ranging from 3-9.

You also expressed in your request for assistance that you were concerned with issues of autocorrelation in your data, due to the nature of how you collected the data, week after week, for 11 months. If there were issues with autocorrelation, you wanted to know possible strategies for solving these issues. When we began examining the data, we learned that there were indeed issues with autocorrelation. We tried multiple methods for accounting for this dependence of observations in our analysis, and this led us to explore the creation of a couple new explanatory variables, Month and Season. Month simply takes on the value of the month the observation was taken in, and Season takes the value of the season that the observation was taken in. You mentioned during our initial meeting that you might be interested in investigating the seasonal nature of the data. Additionally, this allows us to introduce the effect of time into the analysis, to account for the autocorrelation present on the data.

Ⅲ. Statistical Methods

We recommend using the statistical method Multiple Linear Regression. We believe you mentioned this as the method you attempted to complete the analysis previously and hope that will lead to a seamless transition into understanding these models. With this multiple regression model you will be able to **discern the difference in average TAN removal between the two HRT groups, explain the amount of variation accounted for by given explanatory variables, and assess any issues with autocorrelation.**

The model explained above is of the following form:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13} + \beta_{14} x_{14} + \beta_{15} x_{15} + \beta_{16} x_{16} + \varepsilon$$

Where $x_1, \ldots, x_{11}$ = Month (Jan = x_1 to Nov = x_{11}, Dec being the baseline)
x_{12} = HRT (taking a value of 0 for HRT=8 and 1 for HRT=4)
x_{13} = VSS (taking its value from the data)
x_{14} = Minimum Average Daily DO (also taking its value from the data)
x_{15} = Temperature, and
x_{16} = pH

This model best allows us to investigate the primary research question, if there is a significant difference in average TAN removal for the two levels of HRT. The associated coefficient, β_{12}, will be our main interest for this research question. This coefficient effectively creates two regression lines, one line for HRT=4, and another for HRT=8, with a different intercept for each line. Testing if this coefficient is significantly different from 0 will allow us to tell if the effect of HRT on TAN differs for the two groups.

An important note to consider when using this method is that any observations with missing values for the explanatory variables will automatically be excluded by JMP when fitting the regression model. This removes a significant amount of the data from the analysis. There are methods for filling in these missing values, but they all introduce bias into the analysis. Additionally, when we tried one of these methods, the resulting regression model had even worse issues with autocorrelation than the models that excluded the observations with missing data. Thus, we thought it was appropriate to continue with the analysis on the data as it was given to us, without editing the missing values, and allowing them to be excluded.

Ⅳ. Results and Discussions

We have conducted 3 different analyses using this method of Multiple Linear Regression, all of which build upon each other until we reach a final analysis that we propose for your consideration. When considering which of these models was best, we turned to the Durbin-Watson test, which is a test that allows you to check for autocorrelation in the residuals (the differences between the actual, observed values from the data, and the predicted values created by the model equation) from a regression model. A small p-value on this test means that there is a significant issue of autocorrelation in the model. We evaluated these tests at the 5% significance level.

We also considered adding interaction terms for a variety of variables as we thought that this could explain more of the variation in average TAN removal. However, upon adding these interaction terms we found that these added terms were not significant and they did not contribute to the model for the better, these terms often led to worsening the issue of autocorrelation within the models.

Analysis #1: (With all variables of interest)

The first model we considered using the following explanatory variables: HRT, VSS, minAvgDailyDO, Temp, pH, and Season. We tried this first because you had expressed interest in the effect of season on the analysis, and because it could solve the issue of the autocorrelation in the data.

Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**
Model	7	250.68344	35.8119	15.4312
Error	69	160.13112	2.3207	**Prob > F**
C.Total	76	410.81456		< .0001

Table 1: Analysis of variance output for model 1

Durbin-Watson Test

Durbin-Watson	**Number of Obs.**	**Autocorrelation**	**Prob<DW**
1.6069456	77	0.1882	0.0102

Table 2: Durbin Watson output for model 1

The p-value for the Durbin-Watson test for this model is small, meaning there is a significant issue of autocorrelation present with this model. For this reason, we **do not** recommend using this model.

Analysis #2: (With Month)

The next model we fit used the explanatory variables: HRT, VSS, minAvgDailyDO, Temp, pH, and Month. We theorized that because Month has more possible values than Season, it might account for more of the variation created by time.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	139.96880	23.3281	9.1848
Error	48	121.78078	2.5371	**Prob > F**
C.Total	54	261.74958		< .0001

Table 3: Analysis of variance output for model 2

Durbin-Watson Test

Durbin-Watson	Number of Obs.	Autocorrelation	Prob<DW
1.4499105	55	0.2258	0.0080

Table 4: Durbin Watson output for model 2

The p-value for the Durbin-Watson test for this model is also small, meaning there is a significant issue of autocorrelation present with this model. For this reason, we **do not** recommend using this model.

Analysis #3: (Without Temp)

When both of the prior models returned low p-values for the Durbin-Watson test, we started looking into variables we could remove, ones that weren't as important to your research goals. During our initial meeting, you expressed that you weren't as interested in the effects of Temperature and pH on the average TAN removal. We also realized that Temperature and Month are likely to have a relationship with each other, so we tried removing it from the model.

Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**
Model	15	221.80189	14.7868	8.4463
Error	43	75.27894	1.7507	**Prob > F**
C.Total	58	297.08083		< .0001

Table 5: Analysis of variance output for model 3

Durbin-Watson

Durbin-Watson	**Number of Obs.**	**Autocorrelation**	**Prob<DW**
1.9037226	59	0.0352	0.0603

Table 6: Durbin Watson output for model 3

This time, the p-value for the Durbin-Watson test was greater than 0.05, meaning that this model did not have a significant issue of autocorrelation. While the result of this test is still of borderline significance, this was the best model we found that included all of the important variables you highlighted in our initial meeting, that had the least issue of autocorrelation. It is worth noting that some autocorrelation being present in regression analysis is common.

Parameter Estimates

Term	**Estimate**	**P > |t|**
Intercept	12.749	.0001
AvgHRT[4]	.906	.0001
AvgVSS	-.006	.0001
AvgMinDailyDO	-.817	.0001
AvgpH	-.022	.9163

Table 7: Final model parameter estimates

Table 7 illustrates the parameter estimates for the final model (the β estimates), note that the month variable is not included in this table to shorten the length of the table (since there are 11 coefficients for month), however this does not mean that these coefficients are not part of the model.

Research Q1:

We have found that the overall average TAN removal for the two groups to be significantly different from each other. We know this because looking at the p-value for the HRT[4] term in table 7, we see that this p-value is very small (p-value: 0.0001). Looking at the HRT[4] estimate (.906), this tells us that that on average the average TAN removal for ponds **with HRT value of 4** is **.906 mg/L/day higher** than the average TAN removal for ponds **with HRT value of 8**, after adjusting for all other variables in the model. This adjustment is essentially saying that this difference found in TAN removal between the two groups is true in the presence of the other explanatory variables. This difference is also apparent in Figure 2.

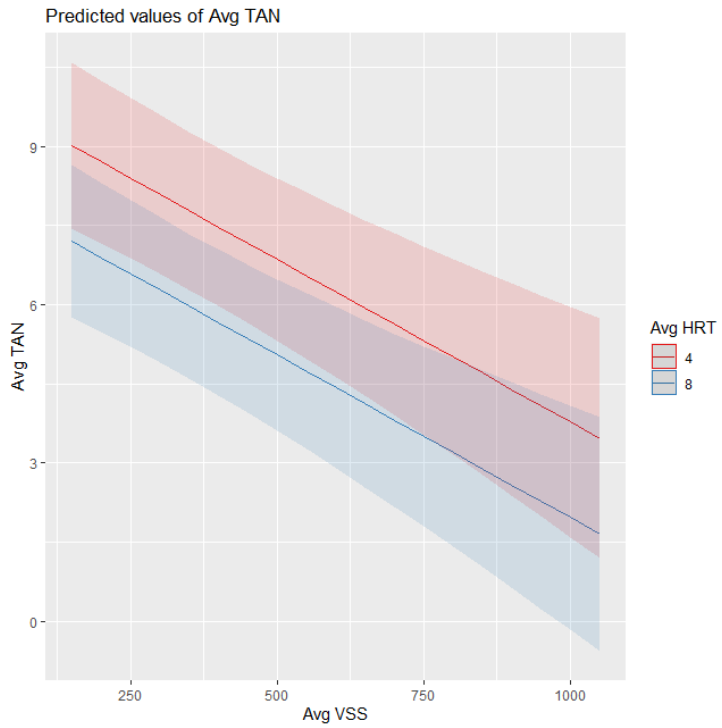


Figure 2: Predicted values of average TAN removal

Research Q2 and Q3:

Table 8 includes the share of the variation in average TAN Removal that can be attributed to the following explanatory variables: AvgHRT, AvgVSS, AvgMinDailyDO, Month, and pH. The variation attributed to each of the explanatory variables was calculated by dividing every predictors regression Sum of Squares value by the Total Sum of Squares value. The Sum of Squares Value measures the variation between observed Average TAN Removal and predicted Average TAN Removal that is being explained by the model proposed, while the Total Sum of Squares measures the amount of variation there is in the Average TAN Removal. By comparing the sum of squares for each predictor to the total sum of squares, we were able to determine the proportion of the variation that is explained by every predictor in our regression model.

Explanatory Variable	**% of Variation Attributed**
AvgHRT	0.13468
AvgVSS	0.15241
AvgMinDailyDO	0.1097
Month	0.2345
pH	0.0000659

Table 8: Variation Attributed to Explanatory Variables

V. Technical Output

Model 1 JMP Output:

Summary of Fit

RSquare	0.610211
RSquare Adj	0.570667
Root Mean Square Error	1.523398
Mean of Response	8.64017
Observations (or Sum Wgts)	77

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	250.68344	35.8119	15.4312
Error	69	160.13112	2.3207	Prob > F
C. Total	76	410.81456		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.687892	1.162267	10.06	<.0001*
AvgTemp ©	0.0250608	0.057595	0.44	0.6648
AvgHRT[4]	1.052416	0.182128	5.78	<.0001*
AvgMinDailyDo (mg/L)	-0.817644	0.222877	-3.67	0.0005*
AvgVSS (mg/L)	-0.004957	0.001001	-4.95	<.0001*
Season[Autumn]	-0.025493	0.304388	-0.08	0.9335
Season[Spring]	1.139353	0.345966	3.29	0.0016*
Season[Summer]	-0.157204	0.324394	-0.48	0.6295

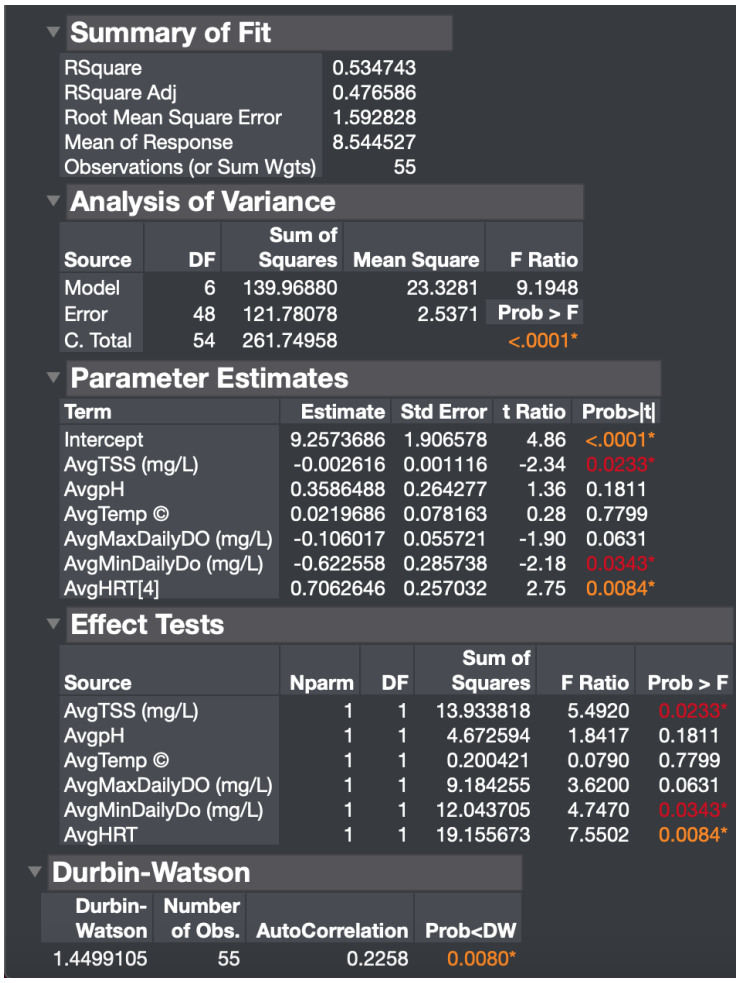
Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
AvgTemp ©	1	1	0.439383	0.1893	0.6648
AvgHRT	1	1	77.490107	33.3902	<.0001*
AvgMinDailyDo (mg/L)	1	1	31.233804	13.4585	0.0005*
AvgVSS (mg/L)	1	1	56.960406	24.5441	<.0001*
Season	3	3	26.689162	3.8334	0.0134*

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.6069456	77	0.1882	0.0102*

Model 2 JMP Output:



Summary of Fit

RSquare	0.534743
RSquare Adj	0.476586
Root Mean Square Error	1.592828
Mean of Response	8.544527
Observations (or Sum Wgts)	55

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	139.96880	23.3281	9.1948
Error	48	121.78078	2.5371	Prob > F
C. Total	54	261.74958		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	9.2573686	1.906578	4.86	<.0001*
AvgTSS (mg/L)	-0.002616	0.001116	-2.34	0.0233*
AvgpH	0.3586488	0.264277	1.36	0.1811
AvgTemp ©	0.0219686	0.078163	0.28	0.7799
AvgMaxDailyDO (mg/L)	-0.106017	0.055721	-1.90	0.0631
AvgMinDailyDo (mg/L)	-0.622558	0.285738	-2.18	0.0343*
AvgHRT[4]	0.7062646	0.257032	2.75	0.0084*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
AvgTSS (mg/L)	1	1	13.933818	5.4920	0.0233*
AvgpH	1	1	4.672594	1.8417	0.1811
AvgTemp ©	1	1	0.200421	0.0790	0.7799
AvgMaxDailyDO (mg/L)	1	1	9.184255	3.6200	0.0631
AvgMinDailyDo (mg/L)	1	1	12.043705	4.7470	0.0343*
AvgHRT	1	1	19.155673	7.5502	0.0084*

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.4499105	55	0.2258	0.0080*

Model 3 JMP Output:

Summary of Fit

RSquare	0.746605
RSquare Adj	0.658211
Root Mean Square Error	1.32313
Mean of Response	8.477572
Observations (or Sum Wgts)	59

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	15	221.80189	14.7868	8.4463
Error	43	75.27894	1.7507	Prob > F
C. Total	58	297.08083		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	12.749274	1.506245	8.46	<.0001*
Month[1]	-2.134527	0.674764	-3.16	0.0029*
Month[2]	-0.98226	1.245269	-0.79	0.4346
Month[3]	1.7263811	0.644989	2.68	0.0105*
Month[4]	1.6919894	0.705913	2.40	0.0210*
Month[5]	1.1007909	0.513663	2.14	0.0378*
Month[6]	1.1474928	0.5805	1.98	0.0545
Month[7]	-1.223058	0.508163	-2.41	0.0205*
Month[8]	-0.459277	0.531257	-0.86	0.3921
Month[9]	-0.193655	0.897884	-0.22	0.8303
Month[10]	1.2619804	0.544606	2.32	0.0253*
Month[11]	-1.179354	0.5417	-2.18	0.0350*
AvgpH	-0.021858	0.206651	-0.11	0.9163
AvgMinDailyDo (mg/L)	-0.816505	0.189234	-4.31	<.0001*
AvgVSS (mg/L)	-0.006156	0.001211	-5.09	<.0001*
AvgHRT[4]	0.9056626	0.189438	4.78	<.0001*

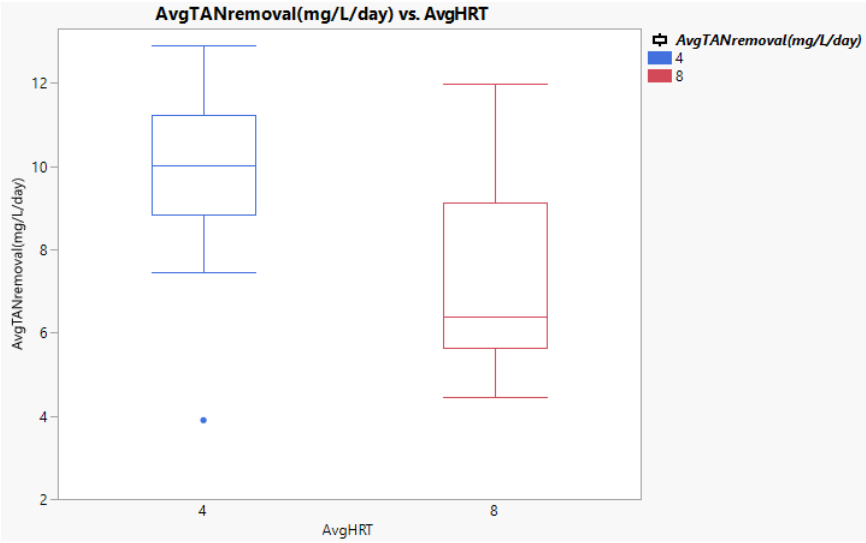
Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Month	11	11	69.665791	3.6176	0.0011*
AvgpH	1	1	0.019587	0.0112	0.9163
AvgMinDailyDo (mg/L)	1	1	32.593087	18.6175	<.0001*
AvgVSS (mg/L)	1	1	45.279025	25.8638	<.0001*
AvgHRT	1	1	40.013258	22.8559	<.0001*

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.9037226	59	0.0352	0.0603

Boxplot comparing Average TAN removal by HRT value:



Predicted values of average TAN removal