**CE 269 Applied Geostatistics**

Homework 1: Normality and Correlation

Thomas Adler

August 30, 2019

# Test of Normality

With the hydrochemcial dataset provided, the sample parameters within the dataset are evaluated for normality. This is a critical attribute of the data to be aware of as many statistical tests rest upon the assumption of normality. Deviations from normality render those statistical tests inaccurate so it is important to know if the data is normally distributed. Provided in part 1 of this assessment are the summary statistics, bar plots, quantile plots and normal quantile plots associated with each parameter.

## **Parameter: Reduction-oxidation potential (Eh)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 595 | | 99.5% |  | 595 | | 97.5% |  | 512.125 | | 90.0% |  | 190 | | 75.0% | quartile | 147.5 | | 50.0% | median | 62.5 | | 25.0% | quartile | -200 | | 10.0% |  | -200 | | 2.5% |  | -200 | | 0.5% |  | -200 | | 0.0% | minimum | -200 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 1.25 | | Std Dev | 191.376 | | Std Err Mean | 25.573693 | | Upper 95% Mean | 52.500827 | | Lower 95% Mean | -50.00083 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.833397 |  | <.0001\* |  **Normally Distributed?** **No** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter: pH** | Quantiles  |  |  | | --- | --- | | 100.0% | 7.9 | | 99.5% | 7.9 | | 97.5% | 7.9 | | 90.0% | 7.6 | | 75.0% | 7.1 | | 50.0% | 6.75 | | 25.0% | 6.4 | | 10.0% | 6.2 | | 2.5% | 5.785 | | 0.5% | 5.7 | | 0.0% | 5.7 | |
| Summary Statistics Mean 6.8196429  Std Dev 0.5261938  Std Err Mean 0.0703156  Upper 95% Mean 6.9605585  Lower 95% Mean 6.6787272  N 56 | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.974760 |  | 0.2875 |  **Normally Distributed?** **Yes** |

## **Parameter: SO4**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 107 | | 99.5% |  | 107 | | 97.5% |  | 100.625 | | 90.0% |  | 40.6 | | 75.0% | quartile | 24 | | 50.0% | median | 10 | | 25.0% | quartile | 6 | | 10.0% |  | 5 | | 2.5% |  | 5 | | 0.5% |  | 5 | | 0.0% | minimum | 5 | |
| Summary Statistics Mean 18.785714  Std Dev 23.090111  Std Err Mean 3.0855459  Upper 95% Mean 24.969287  Lower 95% Mean 12.602142  N 56 | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.615901 |  | <.0001\* |  **Normally Distributed?** **No** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter: NO3** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 2.5 | | 99.5% |  | 2.5 | | 97.5% |  | 2.3725 | | 90.0% |  | 1.002 | | 75.0% | quartile | 0.2 | | 50.0% | median | 0.2 | | 25.0% | quartile | 0.2 | | 10.0% |  | 0.1 | | 2.5% |  | 0.1 | | 0.5% |  | 0.1 | | 0.0% | minimum | 0.1 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 0.3785714 | | Std Dev | 0.5149066 | | Std Err Mean | 0.0688073 | | Upper 95% Mean | 0.5164643 | | Lower 95% Mean | 0.2406785 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.515160 |  | <.0001\* |  **Normally Distributed?** **No** |
| **Parameter: NH3** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 230 | | 99.5% |  | 230 | | 97.5% |  | 225.75 | | 90.0% |  | 145 | | 75.0% | quartile | 20.5 | | 50.0% | median | 0.5 | | 25.0% | quartile | 0.044 | | 10.0% |  | 0.04 | | 2.5% |  | 0.04 | | 0.5% |  | 0.04 | | 0.0% | minimum | 0.04 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 29.571107 | | Std Dev | 62.475177 | | Std Err Mean | 8.3485967 | | Upper 95% Mean | 46.302069 | | Lower 95% Mean | 12.840145 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.535427 |  | <.0001\* |  **Normally Distributed?** **No** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter: TKN** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 260 | | 99.5% |  | 260 | | 97.5% |  | 251.5 | | 90.0% |  | 151 | | 75.0% | quartile | 19.25 | | 50.0% | median | 1 | | 25.0% | quartile | 0.5 | | 10.0% |  | 0.5 | | 2.5% |  | 0.21675 | | 0.5% |  | 0.17 | | 0.0% | minimum | 0.17 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 32.674107 | | Std Dev | 68.959723 | | Std Err Mean | 9.2151307 | | Upper 95% Mean | 51.141642 | | Lower 95% Mean | 14.206573 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.529541 |  | <.0001\* |  **Normally Distributed?** **No** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter: COD** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 18000 | | 99.5% |  | 18000 | | 97.5% |  | 15875 | | 90.0% |  | 333 | | 75.0% | quartile | 132.5 | | 50.0% | median | 22.5 | | 25.0% | quartile | 11.75 | | 10.0% |  | 10 | | 2.5% |  | 10 | | 0.5% |  | 10 | | 0.0% | minimum | 10 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 648.23214 | | Std Dev | 2929.76 | | Std Err Mean | 391.50565 | | Upper 95% Mean | 1432.827 | | Lower 95% Mean | -136.3627 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.218704 |  | <.0001\* |  **Normally Distributed?** **No** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter: BOD5** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 19000 | | 99.5% |  | 19000 | | 97.5% |  | 18150 | | 90.0% |  | 214 | | 75.0% | quartile | 33.25 | | 50.0% | median | 4 | | 25.0% | quartile | 3.25 | | 10.0% |  | 2 | | 2.5% |  | 2 | | 0.5% |  | 2 | | 0.0% | minimum | 2 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 690.10714 | | Std Dev | 3369.4152 | | Std Err Mean | 450.25705 | | Upper 95% Mean | 1592.4424 | | Lower 95% Mean | -212.2281 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.202475 |  | <.0001\* |  **Normally Distributed?** **No** |
| **Parameter: TOC** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 7300 | | 99.5% |  | 7300 | | 97.5% |  | 4401.5 | | 90.0% |  | 123 | | 75.0% | quartile | 38.5 | | 50.0% | median | 6.6 | | 25.0% | quartile | 1.25 | | 10.0% |  | 1 | | 2.5% |  | 1 | | 0.5% |  | 1 | | 0.0% | minimum | 1 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 165.91143 | | Std Dev | 973.7857 | | Std Err Mean | 130.12759 | | Upper 95% Mean | 426.69294 | | Lower 95% Mean | -94.87009 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.150310 |  | <.0001\* |  **Normally Distributed?** **No** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter: Fe** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 246 | | 99.5% |  | 246 | | 97.5% |  | 232.4 | | 90.0% |  | 122.7 | | 75.0% | quartile | 56.075 | | 50.0% | median | 10.15 | | 25.0% | quartile | 2.01 | | 10.0% |  | 0.425 | | 2.5% |  | 0.04935 | | 0.5% |  | 0.04 | | 0.0% | minimum | 0.04 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 40.47425 | | Std Dev | 58.880259 | | Std Err Mean | 7.8682056 | | Upper 95% Mean | 56.242486 | | Lower 95% Mean | 24.706014 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.714881 |  | <.0001\* |  **Normally Distributed?** **No** |
| **Parameter: Mn** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 25.9 | | 99.5% |  | 25.9 | | 97.5% |  | 22.1175 | | 90.0% |  | 9.079 | | 75.0% | quartile | 4.89 | | 50.0% | median | 1.08 | | 25.0% | quartile | 0.225 | | 10.0% |  | 0.09268 | | 2.5% |  | 0.0356875 | | 0.5% |  | 0.0325 | | 0.0% | minimum | 0.0325 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 3.2998589 | | Std Dev | 5.2563695 | | Std Err Mean | 0.7024119 | | Upper 95% Mean | 4.7075239 | | Lower 95% Mean | 1.892194 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.651848 |  | <.0001\* |  **Normally Distributed?** **No** |

# Transformed Data

The log-transformation is widely used in environmental research to deal with skewed data that is not normally distributed. As noted from part 1, all of the parameters in the given data set with the exception of pH is not normally distributed. To make the data conform to the assumption of normality, we employ the natural log transformation to all the parameters. Provided in part 2 of this assessment are the summary statistics, bar plots, quantile plots and normal quantile plots associated with each parameter after this transformation has been used.

The issue with missing data in the context of oxidation-reduction potential (Eh) is addressed by simply adding 201 to each of the known values before taking the natural log. Be it that -200 is the lowest reading of Eh, this allows to include all the data in the analysis. While adding this arbitrary constant to the data is harmless in the scheme of normality analysis, it has been shown to alter the p-value in other statistical tests (Feng, 2014).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter: ln(Eh+201)** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 6.67959918 | | 99.5% |  | 6.67959918 | | 97.5% |  | 6.56017237 | | 90.0% |  | 5.96870756 | | 75.0% | quartile | 5.85356028 | | 50.0% | median | 5.57364813 | | 25.0% | quartile | 0 | | 10.0% |  | 0 | | 2.5% |  | 0 | | 0.5% |  | 0 | | 0.0% | minimum | 0 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 3.4108722 | | Std Dev | 2.8831751 | | Std Err Mean | 0.3852805 | | Upper 95% Mean | 4.1829915 | | Lower 95% Mean | 2.6387529 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.680909 |  | <.0001\* |  **Normally Distributed?** **No** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter: ln(SO4)** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 4.67282883 | | 99.5% |  | 4.67282883 | | 97.5% |  | 4.60863672 | | 90.0% |  | 3.70351650 | | 75.0% | quartile | 3.17805383 | | 50.0% | median | 2.30258509 | | 25.0% | quartile | 1.79175946 | | 10.0% |  | 1.60943791 | | 2.5% |  | 1.60943791 | | 0.5% |  | 1.60943791 | | 0.0% | minimum | 1.60943791 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 2.4864134 | | Std Dev | 0.8607378 | | Std Err Mean | 0.1150209 | | Upper 95% Mean | 2.7169205 | | Lower 95% Mean | 2.2559063 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.870219 |  | <.0001\* |  **Normally Distributed?** **No** |
| **Parameter: ln(NO3)** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 0.91629073 | | 99.5% |  | 0.91629073 | | 97.5% |  | 0.86196154 | | 90.0% |  | 0.00001765 | | 75.0% | quartile | -1.609437 | | 50.0% | median | -1.609437 | | 25.0% | quartile | -1.609437 | | 10.0% |  | -2.302585 | | 2.5% |  | -2.302585 | | 0.5% |  | -2.302585 | | 0.0% | minimum | -2.302585 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | -1.412798 | | Std Dev | 0.8025612 | | Std Err Mean | 0.1072468 | | Upper 95% Mean | -1.19787 | | Lower 95% Mean | -1.627725 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.761112 |  | <.0001\* |  **Normally Distributed?** **No** |
| **Parameter: ln(NH3)** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 5.43807930 | | 99.5% |  | 5.43807930 | | 97.5% |  | 5.41918730 | | 90.0% |  | 4.96516117 | | 75.0% | quartile | 3.01950157 | | 50.0% | median | -0.693147 | | 25.0% | quartile | -3.126497 | | 10.0% |  | -3.218875 | | 2.5% |  | -3.218875 | | 0.5% |  | -3.218875 | | 0.0% | minimum | -3.218875 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 0.1139411 | | Std Dev | 3.0227609 | | Std Err Mean | 0.4039334 | | Upper 95% Mean | 0.9234418 | | Lower 95% Mean | -0.69556 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.873692 |  | <.0001\* |  **Normally Distributed?** **No** |
| **Parameter: ln(COD)** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 9.79812703 | | 99.5% |  | 9.79812703 | | 97.5% |  | 9.65982251 | | 90.0% |  | 5.80804854 | | 75.0% | quartile | 4.88135190 | | 50.0% | median | 3.11128813 | | 25.0% | quartile | 2.45818578 | | 10.0% |  | 2.30258509 | | 2.5% |  | 2.30258509 | | 0.5% |  | 2.30258509 | | 0.0% | minimum | 2.30258509 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 3.8309337 | | Std Dev | 1.6867998 | | Std Err Mean | 0.2254081 | | Upper 95% Mean | 4.2826616 | | Lower 95% Mean | 3.3792058 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.812804 |  | <.0001\* |  **Normally Distributed?** **No** |
| **Parameter: ln(TKN)** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 5.5606816 | | 99.5% |  | 5.5606816 | | 97.5% |  | 5.5266634 | | 90.0% |  | 4.9967693 | | 75.0% | quartile | 2.9551025 | | 50.0% | median | 0 | | 25.0% | quartile | -0.693147 | | 10.0% |  | -0.693147 | | 2.5% |  | -1.559885 | | 0.5% |  | -1.771956 | | 0.0% | minimum | -1.771956 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 1.1716572 | | Std Dev | 2.1777399 | | Std Err Mean | 0.2910127 | | Upper 95% Mean | 1.7548598 | | Lower 95% Mean | 0.5884547 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.824357 |  | <.0001\* |  **Normally Distributed?** **No** |
| **Parameter: ln(BOD5)** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 9.85219425 | | 99.5% |  | 9.85219425 | | 97.5% |  | 9.80492336 | | 90.0% |  | 5.30130535 | | 75.0% | quartile | 3.50326719 | | 50.0% | median | 1.38629436 | | 25.0% | quartile | 1.17053280 | | 10.0% |  | 0.69314718 | | 2.5% |  | 0.69314718 | | 0.5% |  | 0.69314718 | | 0.0% | minimum | 0.69314718 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 2.5027435 | | Std Dev | 2.1184269 | | Std Err Mean | 0.2830867 | | Upper 95% Mean | 3.0700619 | | Lower 95% Mean | 1.935425 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.785978 |  | <.0001\* |  **Normally Distributed?** **No** |
| **Parameter: ln(TOC)** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 8.89562962 | | 99.5% |  | 8.89562962 | | 97.5% |  | 7.73884612 | | 90.0% |  | 4.81150455 | | 75.0% | quartile | 3.64824972 | | 50.0% | median | 1.88522972 | | 25.0% | quartile | 0.17328679 | | 10.0% |  | 0 | | 2.5% |  | 0 | | 0.5% |  | 0 | | 0.0% | minimum | 0 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 2.2041903 | | Std Dev | 1.9716299 | | Std Err Mean | 0.2634701 | | Upper 95% Mean | 2.7321962 | | Lower 95% Mean | 1.6761843 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.907883 |  | 0.0004\* |  **Normally Distributed?** **No** |
| **Parameter: ln(Fe)** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 5.5053315 | | 99.5% |  | 5.5053315 | | 97.5% |  | 5.4461054 | | 90.0% |  | 4.7998617 | | 75.0% | quartile | 4.0234419 | | 50.0% | median | 2.3173644 | | 25.0% | quartile | 0.6978069 | | 10.0% |  | -0.870346 | | 2.5% |  | -3.032617 | | 0.5% |  | -3.218875 | | 0.0% | minimum | -3.218875 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | 2.1903102 | | Std Dev | 2.2270284 | | Std Err Mean | 0.2975992 | | Upper 95% Mean | 2.7867123 | | Lower 95% Mean | 1.5939081 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.951545 |  | 0.0249\* |  **Normally Distributed?** **No** |
| **Parameter: ln(Mn)** | Quantiles  |  |  |  | | --- | --- | --- | | 100.0% | maximum | 3.2542429 | | 99.5% |  | 3.2542429 | | 97.5% |  | 3.0753053 | | 90.0% |  | 2.1951023 | | 75.0% | quartile | 1.5869626 | | 50.0% | median | 0.0769181 | | 25.0% | quartile | -1.492374 | | 10.0% |  | -2.386499 | | 2.5% |  | -3.338268 | | 0.5% |  | -3.426515 | | 0.0% | minimum | -3.426515 | |
| Summary Statistics  |  |  | | --- | --- | | Mean | -0.060748 | | Std Dev | 1.7560441 | | Std Err Mean | 0.2346613 | | Upper 95% Mean | 0.4095238 | | Lower 95% Mean | -0.53102 | | N | 56 | | Shapiro-Wilk W Results  | **W** |  | **Prob<W** | | --- | --- | --- | | 0.968792 |  | 0.1547 |  **Normally Distributed?** **Yes** |

# Remaining Parameters

As we saw in part 2, many of the parameters that were initially non-normally distributed stayed that way even after the log transformation was taken. This exemplifies one of the limitations associated with log transformations. Additionally, studies have shown that log-transformations can alter the results of hypothesis testing thus making them useless in facilitating inferences concerning the original data (Feng, 2014). With this dataset it is a better to abandon transformations and instead move forward with non-parametric statistical tools that do not rely on the assumption of normally distributed data.

# Regression Models

Linear regression models are generated for a number of combinations as displayed in Figures 4.1 through 4.5. The fit of the models is determined with the parameters R2 and R2 adjusted. While they both provide information on the quality of the fit, R2 assumes that every single variable explains the variation in the dependent variable while R2 adjusted defines the percentage of variation explained only by the independent variables that actually affect the dependent variable. In other words, the R2 adjusted accounts for the number of terms in the model so if more useless variables are added, the R2 adjusted will decrease.

Of the models displayed in Figures 4.1 through 4.5, the one with the best fit defines the relationship between ln(COD) and ln(BOD5). This model shows that these two parameters maintain a positive linear relationship (R2adjust = 0.81). Chemically this makes logical sense considering that COD is a measure of the oxygen equivalent of the organic matter content that is susceptible to oxidation while BOD is the amount of oxygen that bacteria take-up. They are intrinsically linked and both dependent on the same two variables, oxygen and organic matter.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Summary of Fit  |  |  | | --- | --- | | RSquare | 0.22309 | | RSquare Adj | 0.208702 | | Root Mean Square Error | 52.37688 | | Mean of Response | 40.47425 | | Observations | 56 | |

**Figure 4.1: Fe vs. pH Regression**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Summary of Fit  |  |  | | --- | --- | | RSquare | 0.192812 | | RSquare Adj | 0.177864 | | Root Mean Square Error | 4.76604 | | Mean of Response | 3.299859 | | Observations | 56 | |

**Figure 4.2: Mn vs. pH Regression**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Summary of Fit  |  |  | | --- | --- | | RSquare | 0.700039 | | RSquare Adj | 0.694484 | | Root Mean Square Error | 2.905378 | | Mean of Response | 3.299859 | | Observations (or Sum Wgts) | 56 | |

**Figure 4.3: Mn vs. Fe Regression**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Summary of Fit  |  |  | | --- | --- | | RSquare | 0.644343 | | RSquare Adj | 0.637503 | | Root Mean Square Error | 80.73323 | | Mean of Response | 49 | | Observations (or Sum Wgts) | 54 |   Outlier Wells Removed: MW 308 (Samples 13 and 34) |

**Figure 4.4: TOC vs. BOD5 Regression**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Summary of Fit  |  |  | | --- | --- | | RSquare | 0.818993 | | RSquare Adj | 0.815512 | | Root Mean Square Error | 0.548909 | | Mean of Response | 3.615953 | | Observations (or Sum Wgts) | 54 | |

**Figure 4.5: ln(COD) vs. ln(BOD5) Regression**

# TOC by BOD5 Regression

In Figure 4.4, the relationship between BOD5 and TOC is displayed. The model to define this relationship is generated with the exclusion of the samples collected from monitoring well MW-308. This well produced significantly higher values than the rest of the wells and as a result would have had a lot more weight in the model if they were included. There is a clear positive relationship between these two variables, however the quality of the model to define this relationship is not great (R2adj = 0.64). Potential confounding variables include the spatial and temporal distributions of the data. Depending on the question you are trying to address, having samples collected at different points in time or space will likely drive the part of the variability. In a further analysis it might be beneficial to group the data by sampling date or to plot the spatial span of the monitoring wells.

# References

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*, *26*(2), 105–109. doi:10.3969/j.issn.1002-0829.2014.02.009