Appendix B: Explanation of Less-Frequent Statistical Methods Used for Data Analysis

1. Bootstrapping

Bootstrapping is a non-parametric statistical method that involves randomly sampling a single observation from the set of collected data and creating a new data set with the same number of observations. A descriptive statistic, for example the mean, of this data set is calculated. This is done many times, creating a distribution of many values of the mean that could be expected from any data set from the target population with the same distribution as the collected data. Descriptive statistics of this distribution can then be used to determine a range of expected results that would be obtained with any similar sample of the target population.

For this study, bootstrapping was used to determine the expected number of UVD units in Thurston County with effluent levels exceeding 400 CFU/100 mL. In the study sample, one of the 22 effluent samples exceeded this value. By randomly sampling from the set of collected data, 10,000 new datasets were created, and the proportion of each dataset that was over 400 CFU/100 mL was calculated. The distribution of this value for each data set is shown in Figure B.1. The 95% confidence of this distribution was then calculated to determine which values would be expected in 95% of similar studies conducted in the same target population (Thurston County OSSs with UVD units). The result of this analysis is a 95% confidence interval of the expected proportion of OSSs with UVD units in Thurston County with fecal coliform effluent concentrations over 400 CFU/100 mL.

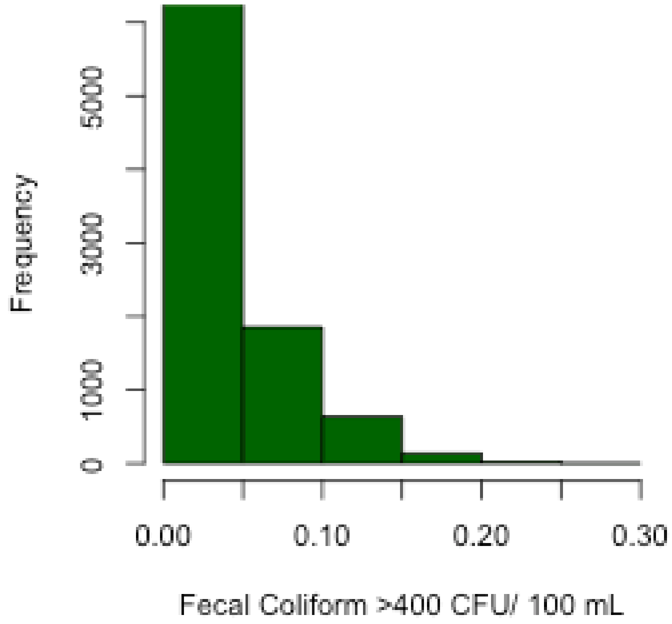

Figure B.1 Distribution of Bootstrapped Values for the Proportion of Fecal Coliform Measurements Exceeding 400 CFU/ 100 mL

2. L1 Regularized Regression Models

An L1 regularized regression model, also known as a lasso regression model, is useful for determining the associations between variables and an outcome of interest when the number of observations does not greatly exceed the number of variables included in the model. The method uses a penalization term to diminish the contribution of each variable to the model. As the penalization term increases, the number of variables that remain in the model decreases until only the variables that are most associated with the outcome are left. However, as variables are removed from the model, a smaller fraction of the variability in the outcome is explained by the model. An example is depicted in Figure B.2: as variables are added to the model (depicted by the colored lines diverging from the x-axis), a higher fraction of the deviance (or variation) is explained).

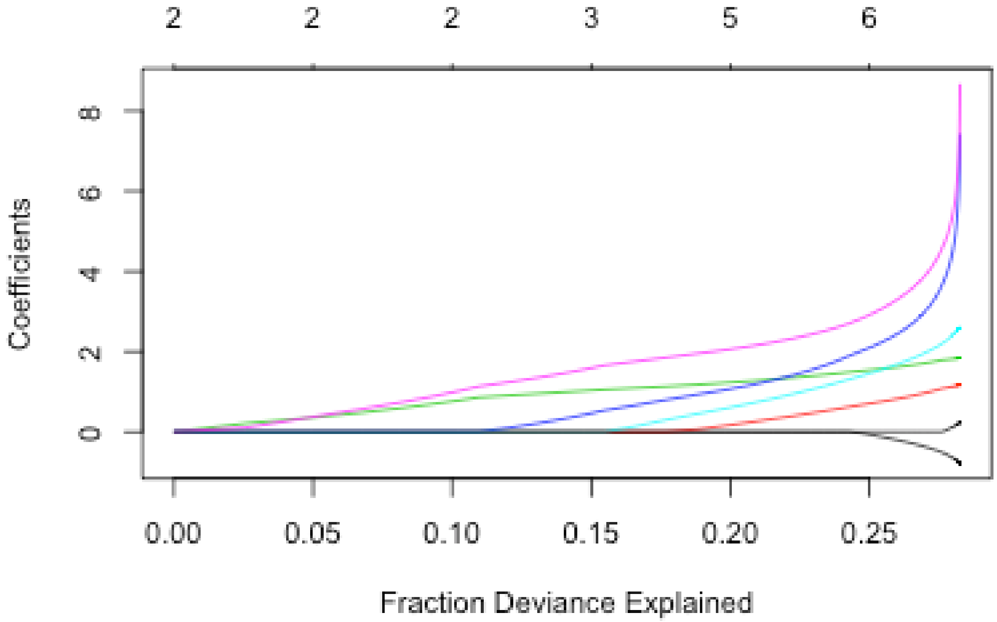

Figure B.2 Increase of Explained Deviance as Coefficients Are Added to L1 Regularized Regression Model

Cross-validation can be used to determine which lambda will provide a model with the smallest cross-validated (mean-squared) error, or that most accurately represents the data (Hastie & Qian, 2016). It is customary to use the model that includes the least number of predictors and has a cross-validated error that is within one standard error of the minimum, which ensures that the model is not over-fit. Thus, a model is created that is both simple and accurate (Krstajic, Buturovic, Leahy, & Thomas, 2014). The cross-validation method uses random number generation to calculate the one-standard-error lambda, so the method is usually run multiple times, and the mean of all results is used to calculate the final L1 regularized regression model coefficients.

For this study, an L1 regularized logistic regression model was used to evaluate the associations between installation, maintenance, and current functioning of OSSs and UV bulb malfunction. 18 observed indicators of installation, maintenance, and functioning were chosen as predictor variables. Indicators that were not directly connected to UVD unit function or that did not have significant variation in our sample were excluded. The cross-validation was run 100 times to determine the one-standard-error lambda. The result of this analysis provided coefficients for the variables that remained in the model, indicating the strength of their association with UV bulb malfunction. A similar method was performed to evaluation associations with log-transformed post-UV fecal coliform concentrations, the only difference being that an L1 regularized linear regression model was run instead of a logistic model.

Table B.1. Variables Used as Predictors in L1 Regularized Models

UVD unit on a non-dedicated circuit
Power switch for UVD unit inaccessible
Electrical connection to UVD unit unprotected
Electrical corrosion or damage
Inadequate cable slack
Cracks in UVD unit housing
UVD unit unprotected from flooding and debris
Leaking UV bulb protective sleeve
Level of biofilm deposit on UV bulb protective sleeve
Location of UVD unit
UVD unit make
ATU make
Age of OSS
Number of bedrooms served by OSS
Months since last OSS inspection
Months since UVD unit last cleaned
Months since UV bulb last replaced
Months since OSS last pumped

Appendix B from https://www.doh.wa.gov/Documents/Pubs/337-155.pdf.