Methodology

The data was obtained as a Microsoft Excel file; it was then converted to a .csv file. Some pre-processing took place before analysis. During pre-processing the data was converted from wide to long format. Some missing data was imputed, when possible, as the mean of a variable by country for all non-missing years for that variable; approximately 15% of the observations were removed due to missing, unimputed data. The variables included in the final analyses were access to clear water (water-core), access to sanitation (sanitation-score), child mortality, year, country, and level of economic development (economy). Water-score and sanitation score were continuous variables from 0 to 100, with 100 indicating everyone in the country has access to water or sanitation sources and 0 meaning no access. Access to water was defined in the data description as “20 liters of water per person per day from an “improved” source (household connections, public standpipes, boreholes, protected dug wells, protected springs, and rainwater collection) within one kilometer of the user's dwelling.” Access to sanitation was defined in the data description as “Facilities such as sewers or septic tanks, poor-flush latrines and simple pit or ventilated improved pit latrines”, provided the latrines or pits are not public. **(CITE THE SOURCE HERE)**.Child mortality was calculated as the number of deaths per 1000 children (ages 1-5) per year.

Exploratory data analysis included examinations of the distributions of each variable, and a correlation analysis. **(talk more about data exploration methods here)**.

Models were compiled using linear regression techniques to predict child mortality. Two models used basic multiple linear regression, one used ridge regression, and one used elastic-net regression. All models used water-score and sanitation-score as independent variables, but only two of them used economy. Year and country were excluded from the models, as the primary focus was on water, sanitation, and economic development. In addition, the individual relationships between water and sanitation scores and child mortality were examined using linear regression. Models were validated using a holdout set containing 20% of the data (not used in training the models). Log transformations of the response variable were performed for each model in attempt to normalize the response variable (child mortality). Exponential transformations were performed on the water-score and sanitation-score to improve the fit of the models and to correct for non-linearity. Analyses of the residuals and model fit were performed for each model using histograms and scatter-plots, along with measurements of fit including r-squared.

A brief description of each final model:

* Model 1: linear regression with water-score and sanitation-score as independent variables
* Model 2: ridge regression with water-score and sanitation-score as independent variables
* Model 3: linear regression with water-score, sanitation-score and economy as independent variables
* Model 4: linear regression with water-score, sanitation-score and economy as independent variables

Results and Discussion

**(Talk about exploratory data analysis here)**

Model 1: This linear regression model used an interaction term between sanitation-score and water-score to get around collinearity assumptions; water and sanitation access were highly correlated (0.88 Pearson correlation coefficient). The fourth power of both water and sanitation were used in the interaction term, and the log (base e) of child-mortality was used to adhere to the normality assumption of linear regression. The model achieved an r-squared value of 0.826, and when predictions were performed on the holdout set, the predictions fit the actual values with an r-squared of 0.821. It’s residuals were distributed approximately normally, although there is some slight heteroscedasticity when residuals are plotted against fit. Overall, the model meets the assumptions of linear regression and fits the data well, including on the validation set (does not overfit). It accurately predicts fewer deaths with higher water-score and sanitation-score, indicating that these variables are impactful towards child mortality.

**(Would include residual figures and predicted vs actual figure below)**

Model 2: This ridge regression model used both water and sanitation scores. The fifth power of both water and sanitation were used in the interaction term, and the log (base e) of child-mortality was used to attain a better fit. Ridge regression does well even though variables are collinear, which is why water and sanitation scores were included as separate variables. Residuals are distributed somewhat normally, with little heteroscedasticity. The model achieved a fit of 0.824, and a validation fit of 0.822 , indicating a good fit without overfitting. One interesting finding is that this model weighed water-score more than sanitation-score; the coefficient for water-score was about 1.29 times higher than for sanitation-score. This could indicate that access to “improved water sources nearby one’s dwelling might be more important for reducing child mortality than sanitation.

**(Would include residual figures and predicted vs actual figure below)**

Model 3: this is a linear regression model that uses water-score, sanitation-score, and economy. A log transformation of child mortality along with exponential transformations (4th power) on water and sanitation scores were performed. Water-score and sanitation-score were wrapped in an interaction term to avoid collinearity concerns. This model achieved an r-squared value of 0.869, and a validation r-squared of 0.851, indicating some but very little overfitting, and a high overall fit. Residuals are normally distributed, with slight heteroscedasticity. This model expects more child mortality for the least developed > emerging G20 member > developing > and emerging MIKT countries (as defined in the Natural Earth Countries dataset); it does not regard emerging BRIC countries or developed regions as significant. It’s interesting that it predicts more child mortality in emerging G20 member states than developing countries. Adding the economic data made for more accurate predictions, although it did create some overfitting. I suspect that the economic data is somewhat correlated to sanitation and water, which could explain overfitting.

**(Would include residual figures and predicted vs actual figure below)**

Model 4: this is an elastic-net regression model (L1 and L2 penalty mix) which includes water and sanitation scores along with the economic data used in model 3. Water and sanitation scores are included as separate terms. This model has an r-squared of 0.871, with a validation r-squared of 0.855; a very good fit with a little overfitting. Residuals are mostly normally distributed, with slight heteroscedasticity. It finds similar trends in the coefficients to model 3 for the economic data. However, unlike model 2, the coefficients for water-score and sanitation-score in this model are very similar. It is possible that adding in the economic data may account for some of the difference between the two.

**(Would include residual figures and predicted vs actual figure below)**

In addition, the individual relationships water and sanitation scores have to child mortality were examined. It was determined via linear regression that both water and sanitation scores can individually explain approximately 78% of the variation in child mortality after variable transformation. Overall, access to improved water and sanitation can explain the majority of excess child mortality within a country, with the best valid model used here (model 4) able to predict transformed responses with a fit of 0.855 (r-squared) on a holdout set.