

DATA 621 Final

Critical Thinking Group 1

5/20/2021

Abstract:

Kevin

Use 250 words or less to summarize your problem, methodology, and major outcomes.

Key words:

Adam

Select a few key words (up to five) related to your work.

Child Mortality, Sanitation, Drinking-Water, Community, Health, Natural Resource

Introduction:

Adam

Describe the background and motivation of your problem.

The quality of life of billions of humans across the world is directly tied to their environment. The World Health Organization (2008) has stated the following: “ensuring poor people’s access to safe drinking-water and adequate sanitation and encouraging personal, domestic and community hygiene will improve the quality of life of millions of individuals.”

In this paper we will explore whether the improvement in quality of community resources, in particular the access to improved water and access to improved sanitation, will impact child mortality rates in a given country. We will be examining whether access to “at least basic services” for both sanitation conditions and water access will impact child mortality rates. For water access, this means households have access to an adequate water source with water collection times of no longer than 30 minutes round trip. For sanitation, this means households that are using adequate sanitation facilities that are not shared with other households. Additionally, we will use a categorical grouping variable for the type of economic region the country is a part of ranging from 1, (Developed Region - G7) to 7, (Least Developed Region). These metrics were used to predict the probability of an individual dying between ages 1 and 5.

The death of a child is a horrific event regardless of its impact on a community, society, or country. However, beyond the mental and emotional toll, the death of a child will have a profound societal impact. Kirigia’s (2013) study found that “The discounted value of future non-health GDP loss due to the deaths of children under 5 years old in 2013 will be in the order of Int\$ 150.3 billion.” This is particularly impactful in Africa where there was a loss of “approximately 6 % of its non-health GDP from the future years of life lost among the 2,976,000 child deaths that occurred in 2013.” These findings make it clear that countries and policy makers should make providing resources to combat child mortality a top priority due to its economic impact in addition to its moral obligation.

Access to safe drinking water and sanitary conditions are some of the most impactful factors that lead to improved quality of life. So much so, the WHO made them key targets of their Millennium Development Goals. Eight Millennium Development Goals, which were set with a goal year of 2015, were established as the most important factors in meeting the needs of the world poorest communities and improving their quality of life. The World Health Organization (2008) stated that Millennium Development Goal 7, To Ensure environmental sustainability, had the following aspects:

- Target 10: Reduce by half the proportion of people without sustainable access to safe drinking water and basic sanitation
- Indicator 30: Proportion of the Population with Sustainable Access to an Improved Water Source
- Indicator 31: Proportion of the Population with Access to Improved Sanitation. (p. 5)

Relying on the World Health Organization's expertise we will examine these key indicators and their impact on a specific aspect of a nation's quality of life, child mortality.

Literature review:

Adam

Discuss how other researchers have addressed similar problems, what their achievements are, and what the advantage and drawbacks of each reviewed approach are. Explain how your investigation is similar or different to the state-of-the-art. Please cite the relevant papers where appropriate.

The fourth goal of the 2000 Millennium Summit is to reduce child mortality. More specifically, to reduce the under-five mortality rate by two thirds by 2015. By 2015 the global under-five child mortality rate dropped from 90 deaths per 1000 live births, to 43 deaths per 1000 live births. While a stark improvement, this improvement still fell short of their goal of improving to 30 deaths per 1000 live births. Some factors that can lead to higher child mortality rates are: children in rural areas have a higher mortality rate than those in urban areas, children of mothers with secondary education or higher are more likely to survive (United Nations, 2015), and children are less likely to die when they have access to improved water and sanitation infrastructure (Lavy et al. 1996).

The link between access to an adequate water source and child mortality has been found by many other studies. For example, the prevalence and duration of diarrhea among young children in India was found to be much lower for families that had access to piped water (Jalan & Ravallion 2003). Both of these studies use logit regression models to predict the likelihood that a child would die and using that number and a threshold to predict a dichotomous result.

The link between sanitation and child mortality is also well established. In some studies, it has been found that sanitation has a higher impact on child mortality rates than access to water (Abou-Ali, 2003). Other studies have found that while access to improved water was not associated with non-infant child mortality, but sanitation was (Fink, 2011). Fink found that:

Access to improved sanitation was associated with lower mortality (OR=0.77, 95% CI 0.68–0.86), a lower risk of child diarrhea (OR=0.87, 95% CI 0.85–0.90) and a lower risk of mild or severe stunting (OR=0.73, 95% CI 0.71–0.75). Access to improved water was associated with a lower risk of diarrhea (OR=0.91, 95% CI 0.88–0.94) and a lower risk of mild or severe stunting (OR=0.92, 95% CI 0.89–0.94), but did not show any association with non-infant child mortality (OR=0.97, 95% CI 0.88–1.04). (p.1)

Despite this discrepancy Fink still found that there are “large health consequences of lacking access to water and sanitation for children aged <5 years in low- and middle-income countries.” Similarly, to Lavy and Jalan, Fink used a logit logistic regression model to produce a dichotomous outcome. Abou-Ali used a probit logistic regression model to predict child mortality which is similar to logit models in its attempt to predict a dichotomous result and only differs in the distribution it uses to make the prediction.

We used a linear regression model to predict the correct child mortality rate of a country rather than using that rate to predict if a death will occur. We made this decision because we had child mortality rates at a national level as well as other national level variables. We felt to predict the likelihood of a child dying would be difficult to test with these inputs, so we opted to focus on evaluating the accuracy of the child mortality rate at a national level.

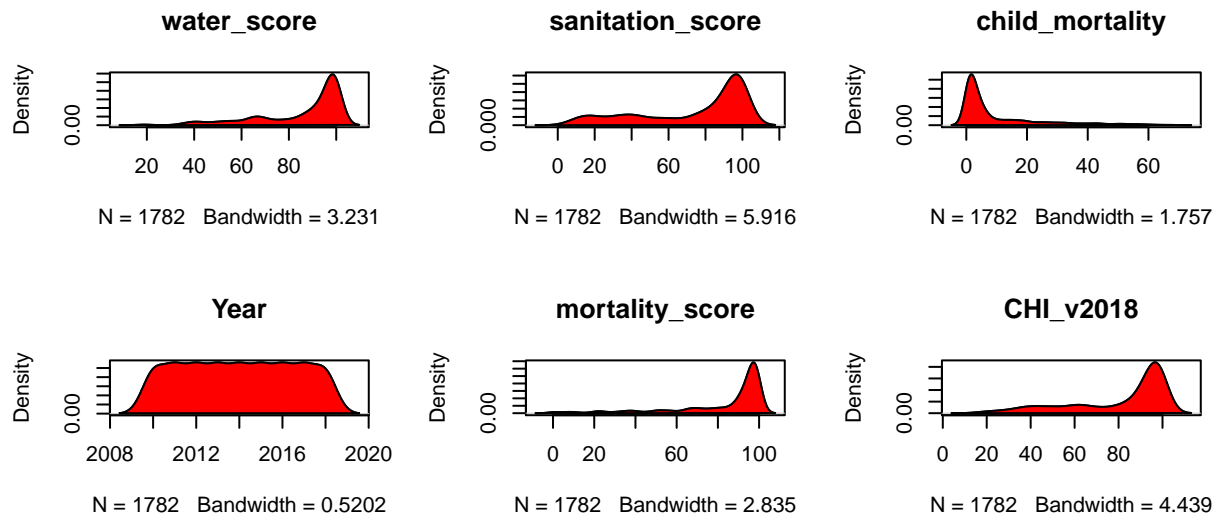
Methodology:

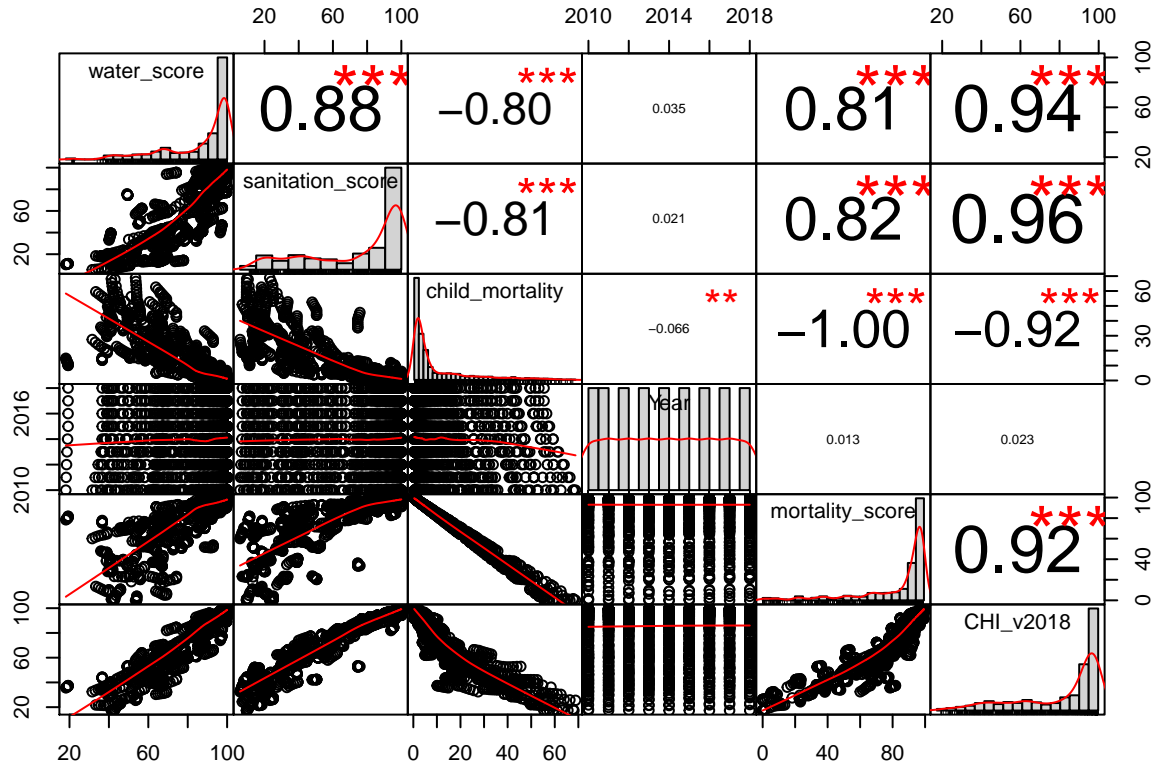
David

Discuss the key aspects of your problem, data set and regression model(s). Given that you are working on real-world data, explain at a high-level your exploratory data analysis, how you prepared the data for regression modeling, your process for building regression models, and your model selection.

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, pour-flush latrines and simple pit or ventilated improved pit latrines”, provided the latrines or pits are not public (Center, 2018). 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.





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:

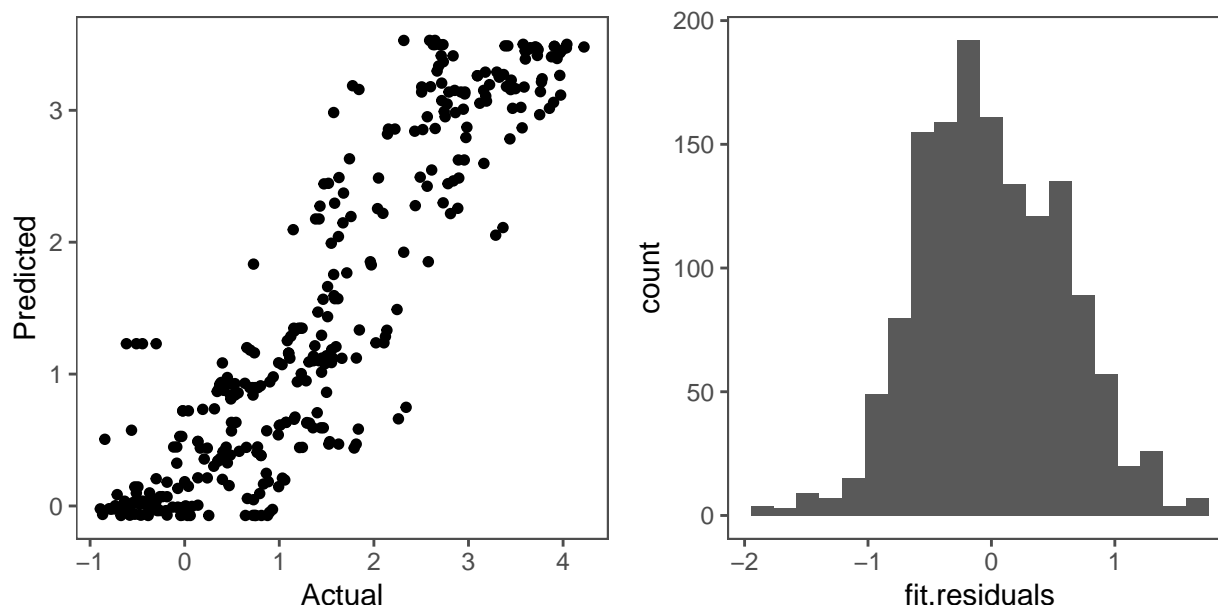
- Model 1: linear regression with water-score and sanitation-score as independent variables
- Model 2: ridge regression with water-score and sanitation-score as independent variables
- Model 3: linear regression with water-score, sanitation-score and economy as independent variables
- Model 4: linear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

Experimentation and Results:

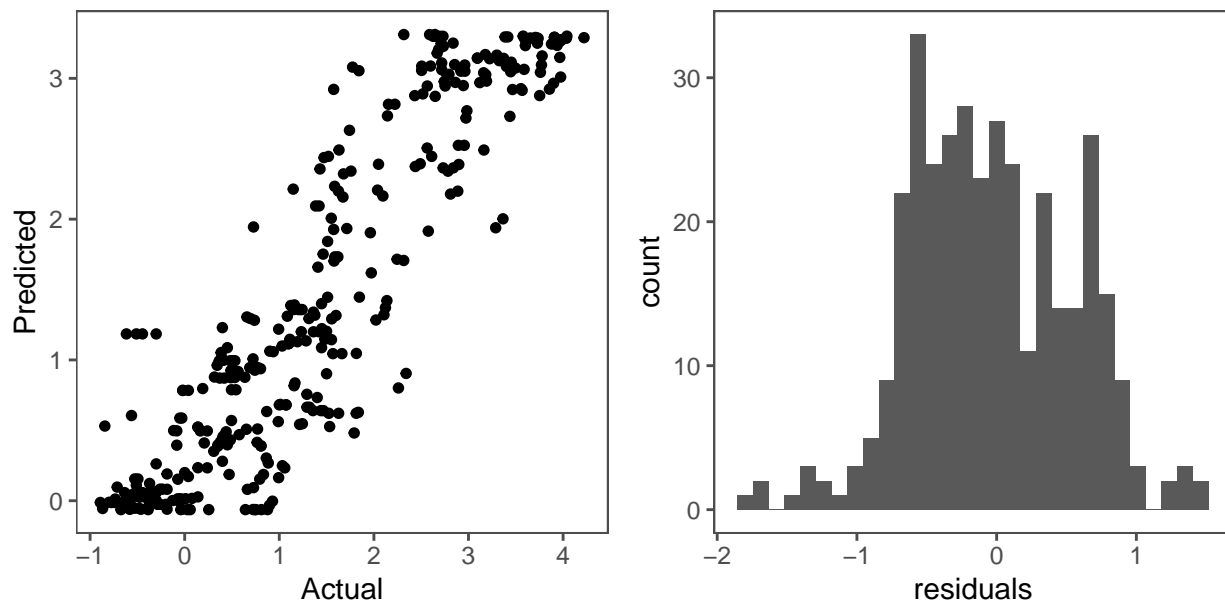
David

Describe the specifics of what you did (data exploration, data preparation, model building, model selection, model evaluation, etc.), and what you found out (statistical analyses, interpretation and discussion of the results, etc.).

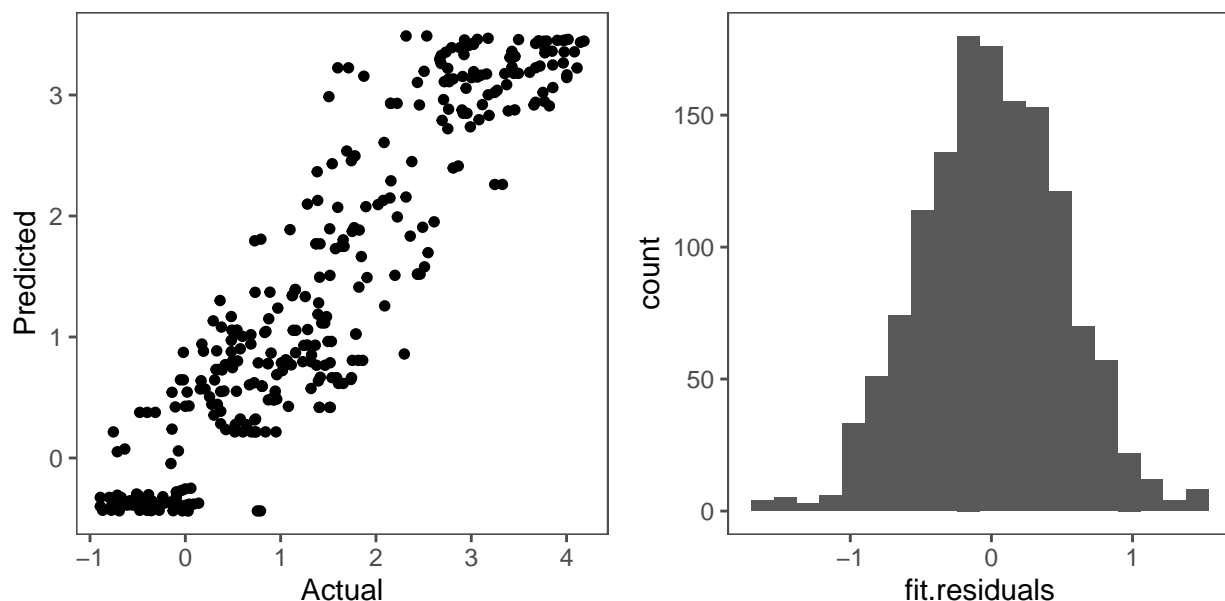
Model 1: This linear regression model used an interaction term between sanitation-score and water-score to get around collinearity assumptions; water and sanitation access were highly correlated (0.88 Pearson correlation coefficient). The fourth power of both water and sanitation were used in the interaction 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.



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.

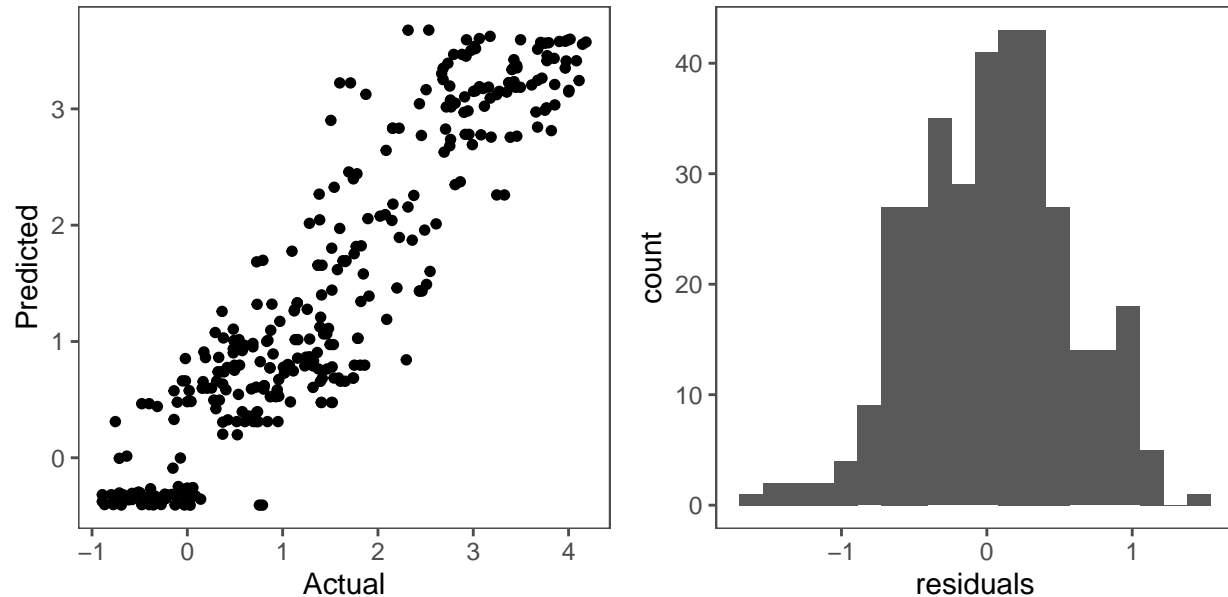


Model 3: this is a linear regression model that uses water-score, sanitation-score, and economy. A log transformation of child mortality along with exponential transformations (4th power) on water and sanitation scores were performed. Water-score and sanitation-score were wrapped in an interaction term 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.



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.



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.

Discussion and Conclusions:

Kevin

Conclude your findings, limitations, and suggest areas for future work.

References:

Be sure to cite all references used in the report (APA format).

Abou-Ali, H. (2003). The effect of water and sanitation on child mortality in Egypt. 1-29. Retrieved May 18, 2021, from <https://gupea.ub.gu.se/bitstream/2077/2828/1/gunwpe0112.pdf?origin=publication>.

Center for International Earth Science Information Network - CIESIN - Columbia University. 2018. Natural Resource Protection and Child Health Indicators, 2018 Release. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). <https://doi.org/10.7927/6t8a-es66>. Accessed DAY MONTH YEAR.

Fink, G., Günther, I., & Hill, K. (2011). The effect of water and sanitation on child health: Evidence from the demographic and health surveys 1986–2007. *International Journal of Epidemiology*, 40(5), 1196-1204. doi:10.1093/ije/dyr102

- Jalan, J., & Ravallion, M. (2003). Does piped water reduce diarrhea for children in rural India? *Journal of Econometrics*, 112(1), 153-173. doi:10.1016/s0304-4076(02)00158-6
- Kirigia, J. M., Muthuri, R. D., Nabyonga-Orem, J., & Kirigia, D. G. (2015). Counting the cost of child mortality in the World Health Organization African region. *BMC Public Health*, 15(1). doi:10.1186/s12889-015-2465-z
- Lavy, V., Strauss, J., Thomas, D., & Vreyer, P. D. (1996). Quality of health care, survival and health outcomes in Ghana. *Journal of Health Economics*, 15(3), 333-357. doi:10.1016/0167-6296(95)00021-6
- Prüss-Üstün A, Bos R, Gore F, Bartram J. Safer water, better health: costs, benefits and sustainability of interventions to protect and promote health. World Health Organization, Geneva, 2008.
- United Nations Millennium Development Goals. (n.d.). Retrieved from <https://www.un.org/millenniumgoals/childhealth.shtml>

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.

Appendices:

Supplemental tables and/or figures.

R statistical programming