# DATA 621 Final

Critical Thinking Group 1

Spring 2021

## Abstract:

This document and its appendices use methods and packages in the programming language R to investigate the potential linear relationship between a countries access to modern sanitation and child mortality rate. The data used in our study was part of a global initiative to help measure the impact of sanitation and quality of drinking water in underdeveloped nations. The research used in our findings come from a number of national organizations including but not limited to Center for International Earth Science Information Network, NASA Socioeconomic Data and Applications Center, United Nations, and Wold Health Organization.

Throughout the document we explore the human impact of this crisis, explain how the data was used, methods for creating the model, and metrics to measure the quality. Also included is a summary of our findings, discussion of how we believe our research can be used, and potential next steps.

## Key words:

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

## Introduction:

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:

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:

The data was obtained through NASA Socioeconomic Data and Applications Center (SEDAC) and The Center for International Earth Science Information Network's (CIESIN) public online data portal at: https://sedac.ciesin.columbia.edu/data/sets/browse. It was converted from it's original Microsoft Excel format to .csv. From there it was loaded into an R-markdown file for analysis. 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 (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, correlation analyses, and a map depicting the scale of child mortality.

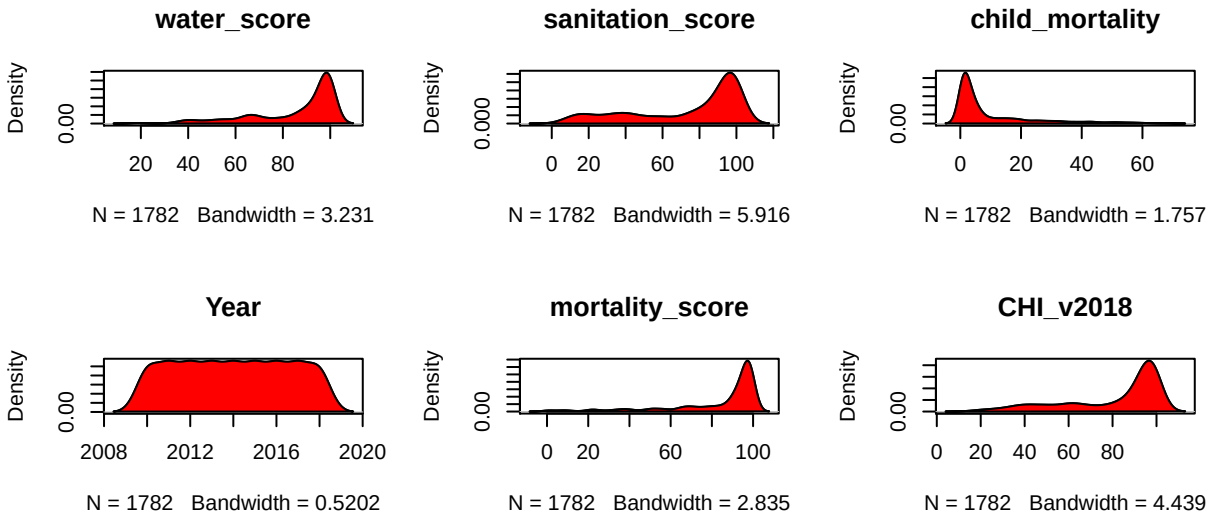
Models were compiled using linear regression techniques to predict child mortality. Two models used basic multiple linear regression, one used ridge regression, and one used elastic-net regression. All models used water-score and sanitation-score as independent variables, but only two of them used economy. 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.

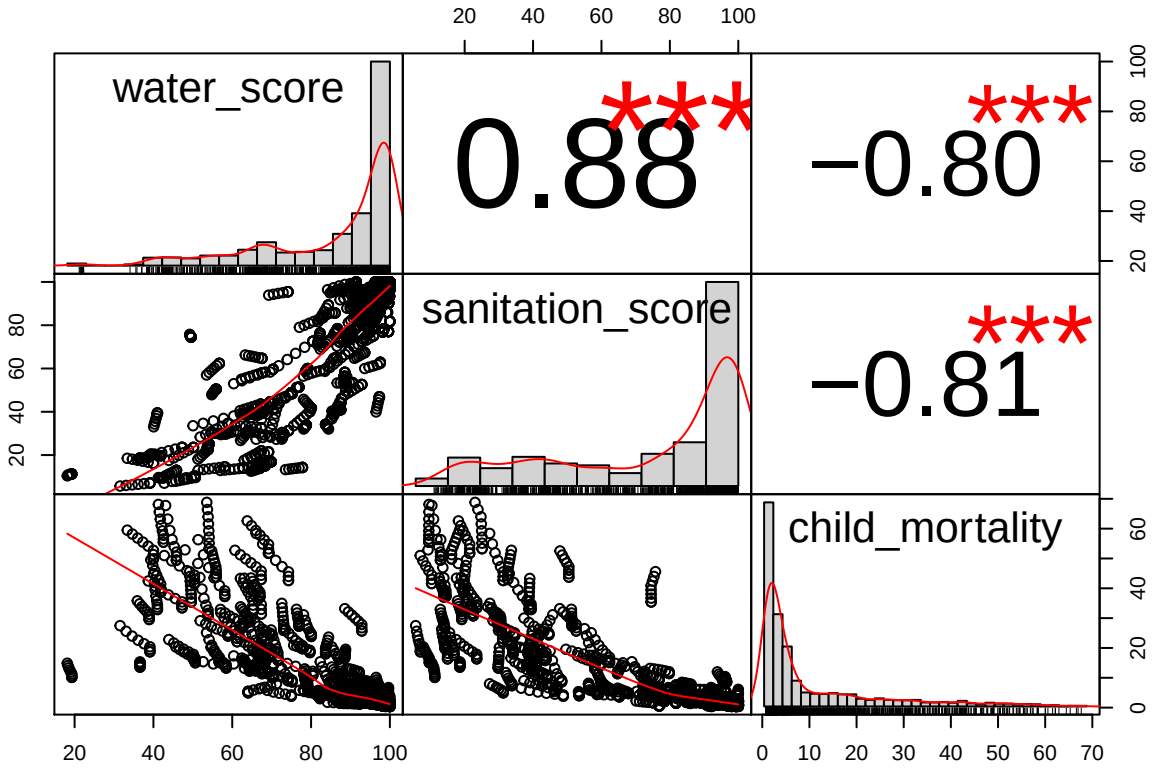
A brief description of each final model:

- 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

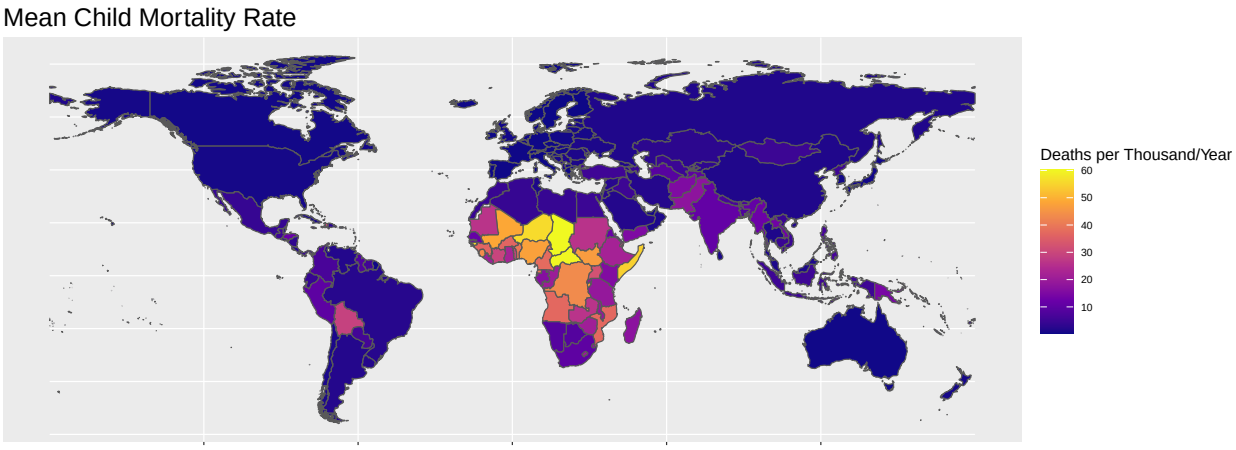
## Experimentation and Results:

Our analysis of the variables used revealed high degrees of correlation. Of particular concern, water-score and sanitation-score were strongly correlated, with a correlation coefficient of 0.88. This raised concerns about multicollinearity in the models, and influenced our choice of models. On the other hand, water and sanitation scores were both strongly inversely correlated to child mortality, with correlation coefficients of -0.80 and -0.81 respectively; this indicates that both variables may be highly predictive of child mortality.

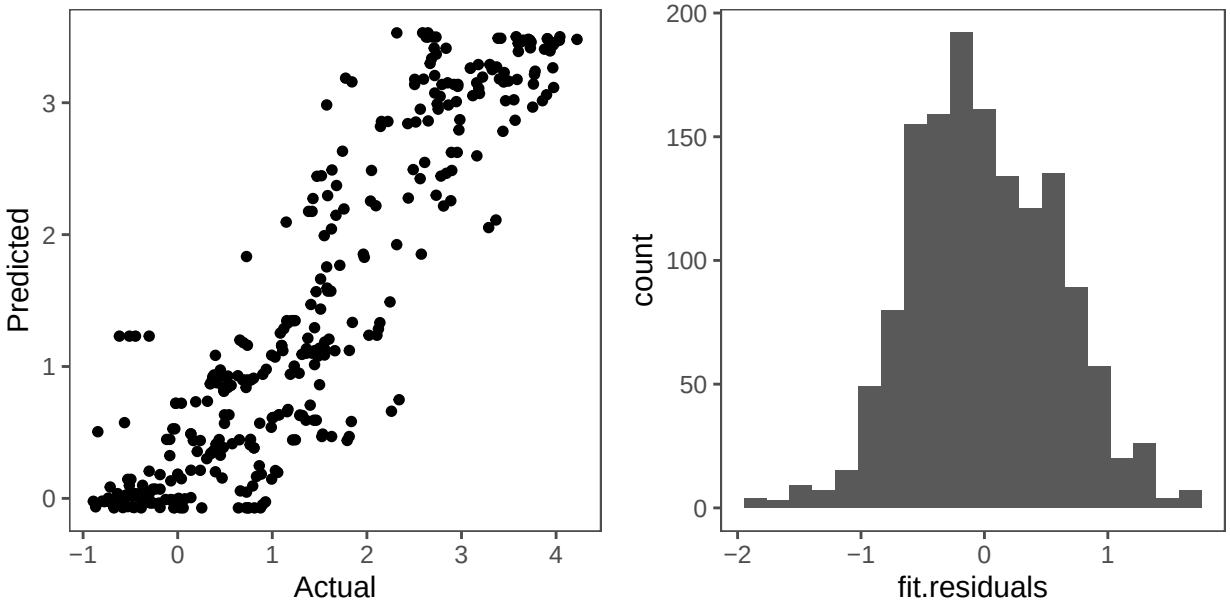




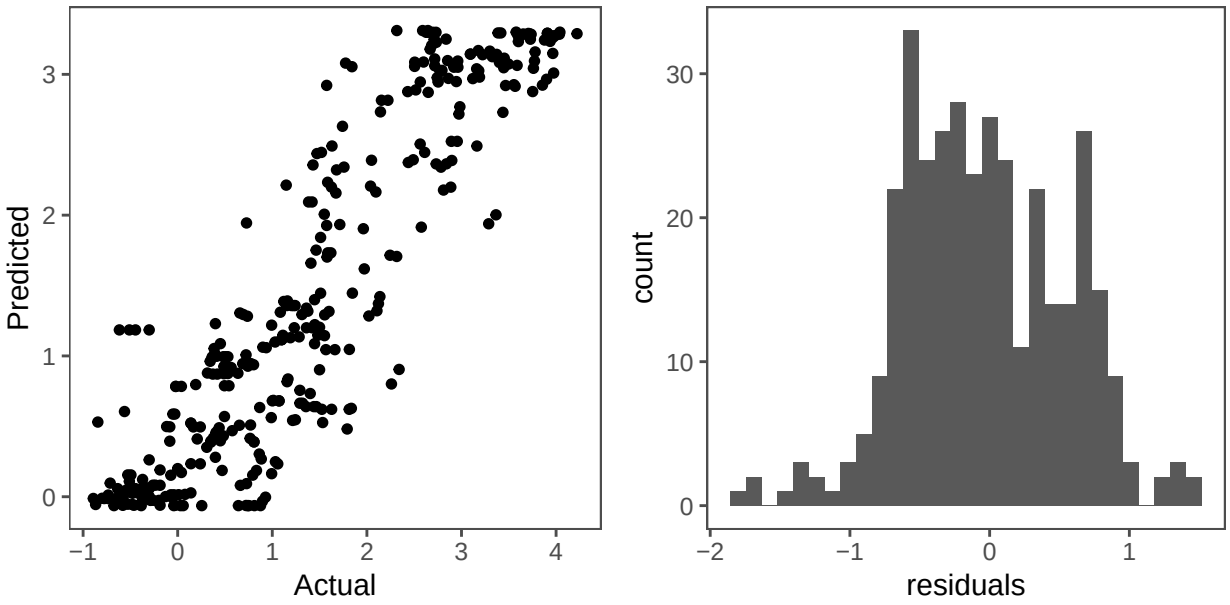

Child mortality rates tend to be highest in sub-Saharan Africa, which is also where access to improved drinking water sources and adequate sanitation is the lowest. Outside of Africa, child mortality is somewhat higher than average in south Asia, and south America (especially Bolivia).



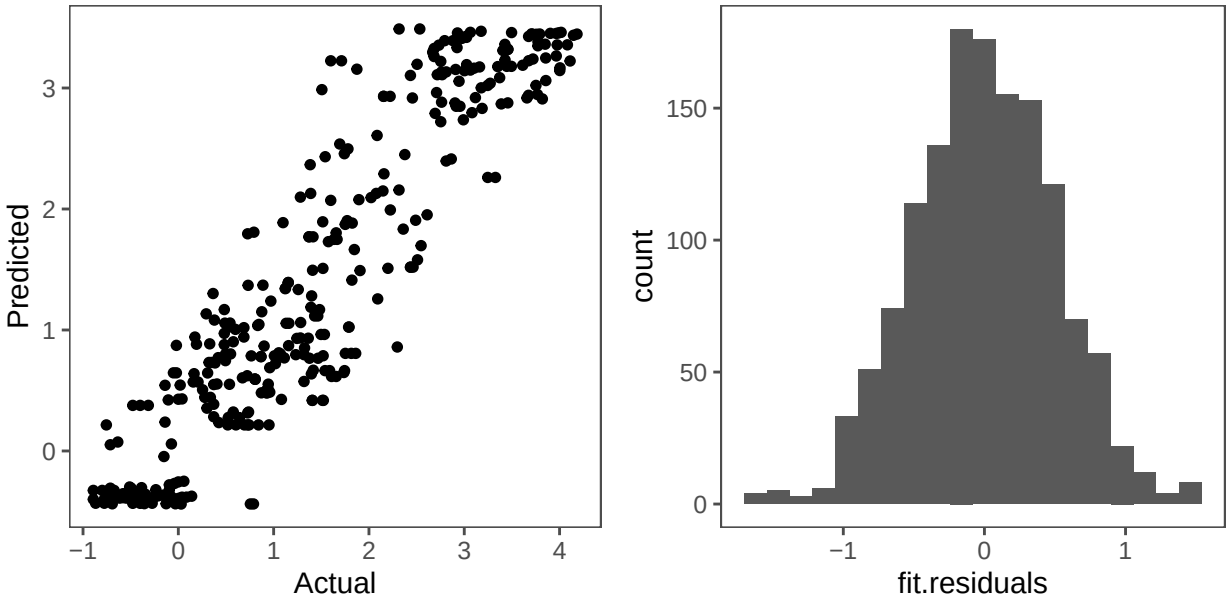
Model 1: This linear regression model used an interaction term between sanitation-score and water-score to get around collinearity assumptions; water and sanitation access were highly correlated (0.88 Pearson correlation coefficient). The fourth power of both water and sanitation were used in the interaction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

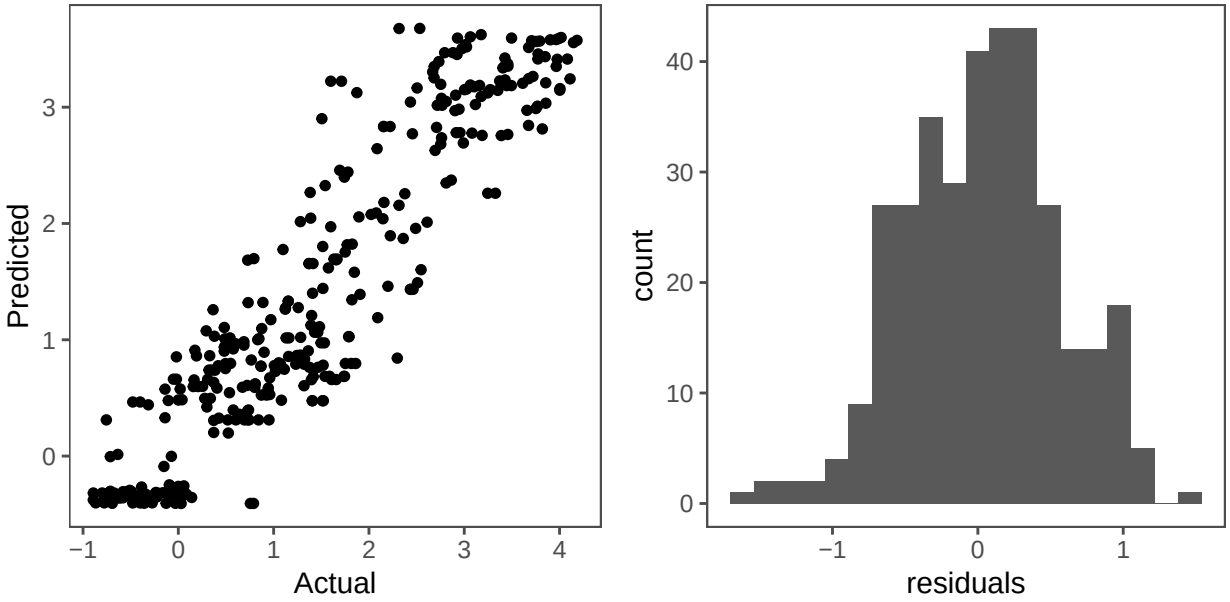



Model 3: this is a linear regression model that uses water-score, sanitation-score, and economy. A log transformation of child mortality along with exponential transformations (4th power) on water and sanitation scores were performed. Water-score and sanitation-score were wrapped in an interaction term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

Through the course of our modeling and external research it's evident that nations who provide their citizens with access to modern sanitation and clean drinking water will result in lower child mortality rate. Virtually all the models we ran resulted in a high rate of predictability. The explained variance and residual analysis show there is a definite linear relationship between child mortality and poor conditions of drinking water and sanitation. The factors that lead to these predictions indicate that these conditions are more prevalent in emerging MIKT countries. This model could be used for many applications including how changes to the contributing factors can positively impact child death rates. A country or non-profit organization could use this model to help decide where to allocate spending if the goal is to mitigate this risk.

This issue impacts a large potion of the world and although most of these studies were done more than five years ago it's still an issue today. It's important to note there are a number of other factors that also correlate to countries that have high child mortality like access to healthcare, quality of healthcare, political instability, and other societal factors. One of the major limitations of the data is child mortality rate does not differentiate between cause of death which makes it difficult to say that the rate is directly attributed to hygiene related illness.

This is an important issue and something often times overlooked or taken for granted by those in developed nations. Although the problem has been recognized by many the solution is not a simple one. Most cities in developed nations were built around modern sanitation strategically placed pump stations and underground

sewers that run throughout the city. Because many cities in underdeveloped nations suffer from over crowding and poor living situations its not feasible to uproot entire cities to build this infrastructure. To solve the issue of sanitation in much of the world it will require new and innovative solutions.

## References:


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

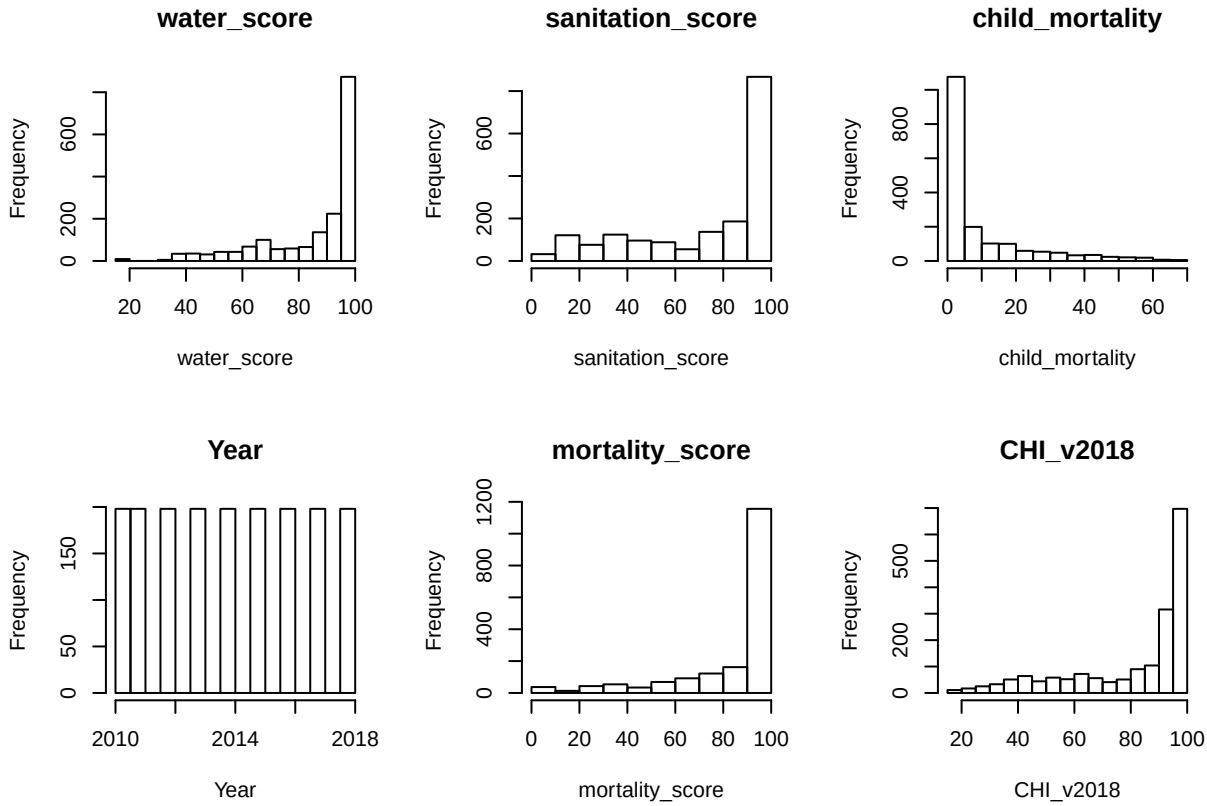
Lavy, V., Strauss, J., Thomas, D., & Vreyer, P. D. (1996). Quality of health care, survival and health outcomes in Ghana. Journal of Health Economics, 15(3), 333-357. doi:10.1016/0167-6296(95)00021-6

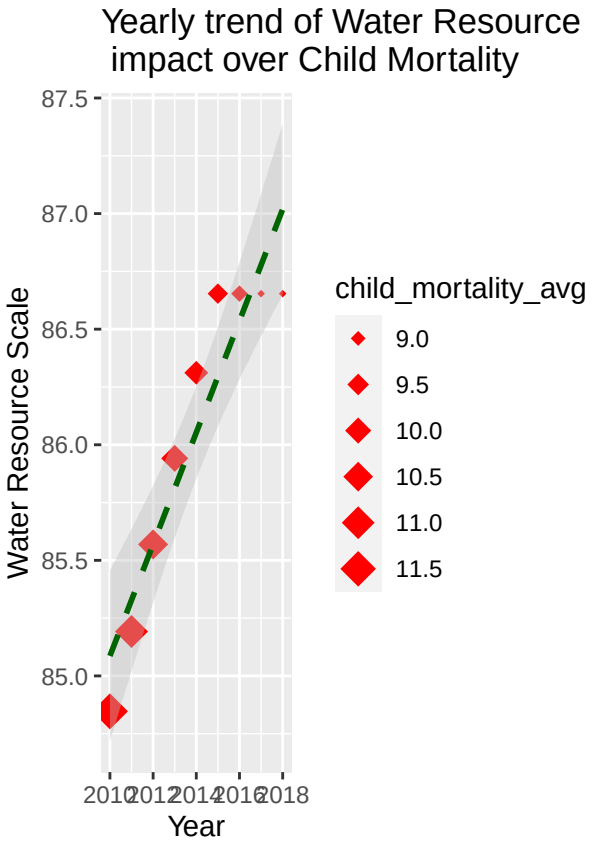
Prüss-Üstün A, Bos R, Gore F, Bartram J. Safer water, better health: costs, benefits and sustainability of interventions to protect and promote health. World Health Organization, Geneva, 2008.

United Nations Millennium Development Goals. (n.d.). Retrieved from https://www.un.org/millenniumgoals/childhealth.shtml

## Appendices:

**Supplemental tables and/or figures.**



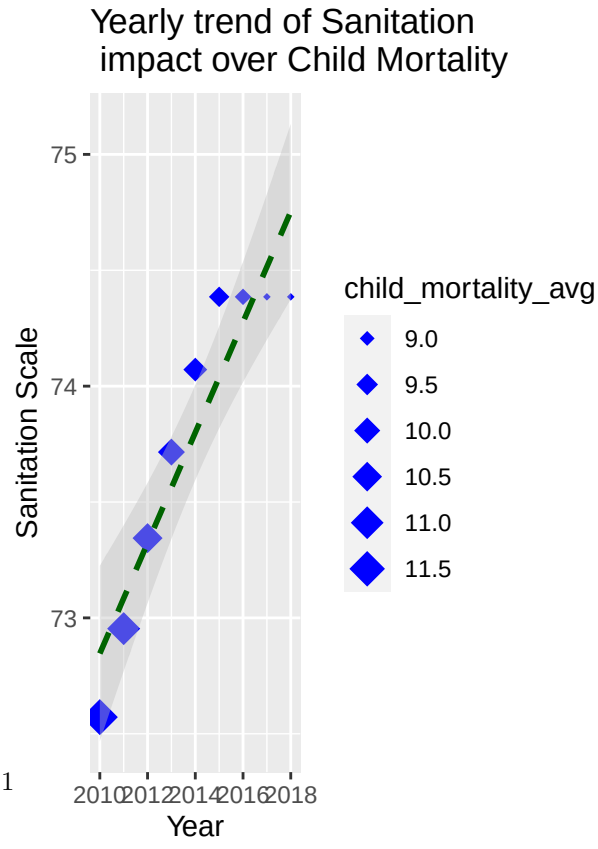
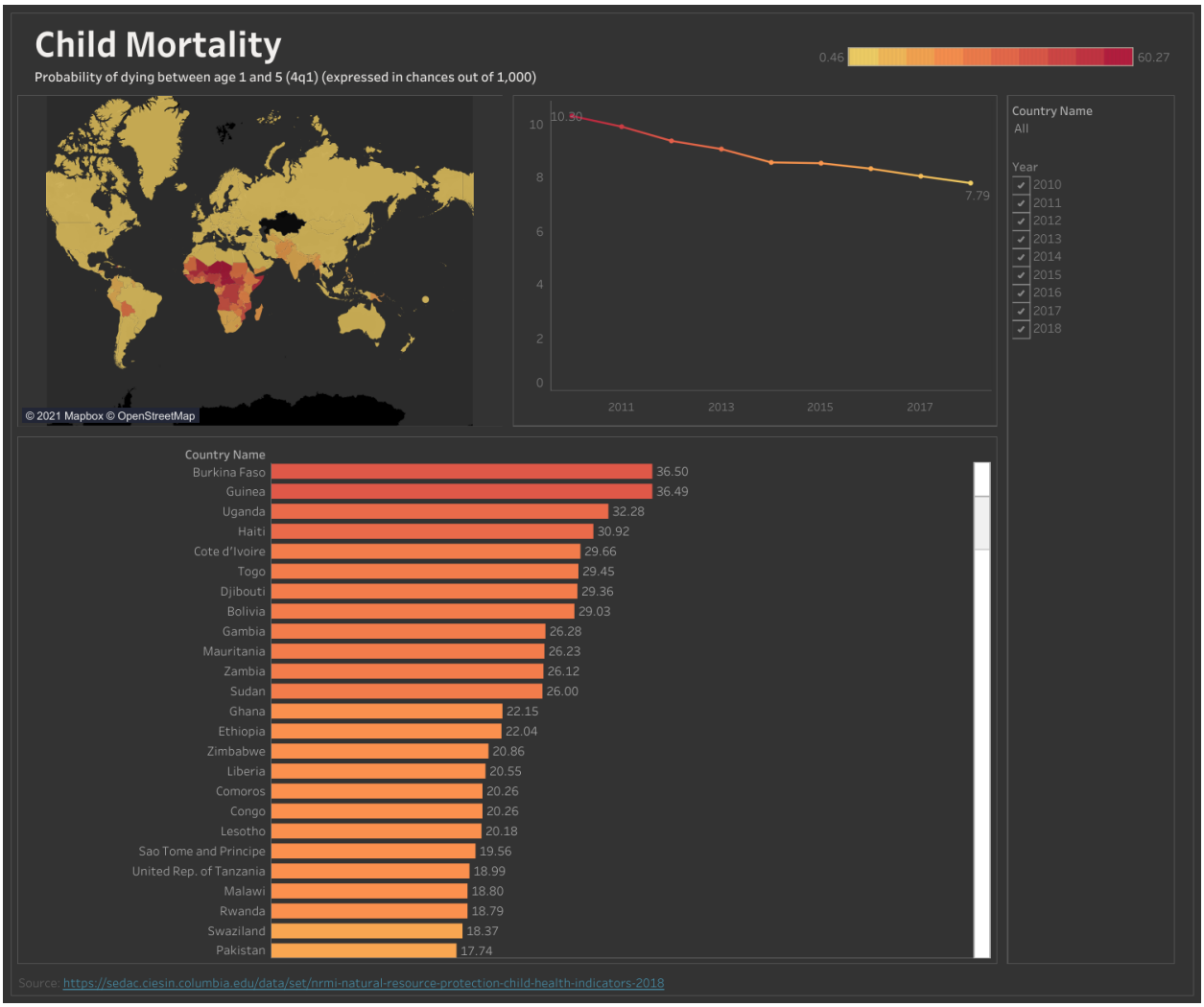



Tableau Notebook:

https://public.tableau.com/profile/forhad.akbar#!/vizhome/ChildMortality_16211276461000/ChildMortality




### R statistical programming

```
# Data Prep

load_chi2018 <- function(var_select = c("water_score", "sanitation_score", "child_mortality",
    "Year", "ISO3", "CountryName")) {


    library(tidyr)
    library(dplyr)
    library(data.table)
    library(caret)
    library(rpms)
```

```r
set.seed(1234567890)

df = read.csv("https://raw.githubusercontent.com/davidblumenstiel/
                    CUNY-MSDS-DATA-621/main/Final_Project/chi-2018.csv")

# Took some code from:
# https://stackoverflow.com/questions/50010196/replacing-na-values-from-another-dataframe-by-id
# and https://stackoverflow.com/questions/25908772/r-column-mean-by-factor

x1 = df %>% pivot_longer(cols = starts_with("wat_"), names_to = "Year", names_prefix = "wat_",
    values_to = "water_score")
x1 = x1 %>% left_join(setDT(x1)[, mean(water_score, na.rm = TRUE), by = CountryName],
    by = "CountryName") %>% mutate(water_score = ifelse(is.na(water_score), V1,
    water_score)) %>% # Replace NA with mean of values if available
dplyr::select("Year", "water_score", "CountryName")

x2 = df %>% pivot_longer(cols = starts_with("san_"), names_to = "Year", names_prefix = "san_",
    values_to = "sanitation_score")
x2 = x2 %>% left_join(setDT(x2)[, mean(sanitation_score, na.rm = TRUE), by = CountryName],
    by = "CountryName") %>% mutate(sanitation_score = ifelse(is.na(sanitation_score),
    V1, sanitation_score)) %>% # Replace NA with mean of values if available
select("Year", "sanitation_score", "CountryName")



x3 = df %>% pivot_longer(cols = starts_with("chmort_"), names_to = "Year", names_prefix = "chmort_"
    values_to = "child_mortality")
x3 = x3 %>% left_join(setDT(x3)[, mean(child_mortality, na.rm = TRUE), by = CountryName],
    by = "CountryName") %>% mutate(child_mortality = ifelse(is.na(child_mortality),
    V1, child_mortality)) %>% # Replace NA with mean of values if available
select("Year", "child_mortality", "CountryName")



x4 = df %>% pivot_longer(cols = starts_with("mortality_"), names_to = "Year",
    names_prefix = "mortality_", values_to = "mortality_score")
x4 = x4 %>% left_join(setDT(x4)[, mean(mortality_score, na.rm = TRUE), by = CountryName],
    by = "CountryName") %>% mutate(mortality_score = ifelse(is.na(mortality_score),
    V1, mortality_score)) %>% # Replace NA with mean of values if available
select("Year", "mortality_score", "CountryName")

x5 = df %>% pivot_longer(cols = starts_with("CHI_v2018_"), names_to = "Year",
    names_prefix = "CHI_v2018_", values_to = "CHI_v2018")
x5 = x5 %>% left_join(setDT(x5)[, mean(CHI_v2018, na.rm = TRUE), by = CountryName],
    by = "CountryName") %>% mutate(CHI_v2018 = ifelse(is.na(CHI_v2018), V1, CHI_v2018)) %>%
    # Replace NA with mean of values if available
select("Year", "CHI_v2018", "CountryName")

out = x1 %>% merge(x2, by = c("CountryName", "Year")) %>% merge(x3, by = c("CountryName",
    "Year")) %>% merge(x4, by = c("CountryName", "Year")) %>% merge(x5, by = c("CountryName",
    "Year"))

out = as.data.frame(out)
```

```
    # Adds back ISO3 abbreviations
    out <- out %>% merge(x = out, y = df[, 1:2], by.x = "CountryName", by.y = "CountryName")
    colnames(out)[8] <- "ISO3"

    # NA dropping
    out <- data.frame(out[, var_select]) %>% drop_na()


    return(out)


}

# Basic EDA

df <- load_chi2018(var_select = c("water_score", "sanitation_score", "child_mortality",
    "Year", "ISO3", "CountryName", "mortality_score", "CHI_v2018"))

summary(df)
hist(df$water_score)
hist(df$sanitation_score)
hist(df$child_mortality)
hist(df$mortality_score)
hist(df$CHI_v2018)
df <- df[complete.cases(df), ]

# All very exponentialy distributed.

library(corrplot)
corrplot(cor(df[, c(1, 2, 3, 7, 8)], use = "pairwise.complete.obs"), type = "upper")

# Also alot of collinearity.

plot(df$child_mortality ~ df$water_score)
plot(df$child_mortality ~ df$sanitation_score)

# Going to need transformations to avoid problems with residuals and the
# assumptions of linear regression

# Modeling

# Let's first explore the individual relationships between water and sanitation
# with child mortality.


df <- load_chi2018()
df$child_mortality <- log(df$child_mortality)
df$water_score <- df$water_score^6
splitdex <- createDataPartition(df$child_mortality, p = 0.8, list = FALSE)
train <- df[splitdex, ]
val <- df[-splitdex, ]
fit <- lm(child_mortality ~ water_score, data = train)
```

```r
summary(fit)
plot(fit)
plot(predict(fit, val) ~ val$child_mortality)
hist(fit$residuals, breaks = 20)
print(paste("Validation R^2: ", cor(predict(fit, val), val$child_mortality)^2))


# after variable transformation, water-score alone will account for about 0.78 of
# the variation in child mortality


df <- load_chi2018()
df$child_mortality <- log(df$child_mortality)
df$sanitation_score <- df$sanitation_score^2.5
splitdex <- createDataPartition(df$child_mortality, p = 0.8, list = FALSE)
train <- df[splitdex, ]
val <- df[-splitdex, ]
fit <- lm(child_mortality ~ sanitation_score, data = train)
summary(fit)
plot(fit)
plot(predict(fit, val) ~ val$child_mortality)
hist(fit$residuals, breaks = 20)
print(paste("Validation R^2: ", cor(predict(fit, val), val$child_mortality)^2))


# after variable transformation, sanitation-score alone can also account for
# about 0.77-0.78 of the variation in child mortality


df <- load_chi2018()
splitdex <- createDataPartition(df$child_mortality, p = 0.8, list = FALSE)
train <- df[splitdex, ]
val <- df[-splitdex, ]
fit <- lm(log(child_mortality)^1.1 ~ I(water_score * sanitation_score), data = train)
summary(fit)
plot(fit)
plot((exp(predict(fit, val)))^(1/1.1) ~ val$child_mortality)
hist(fit$residuals, breaks = 20)

# Basic linear regression with log and slight exponential transformaton.  Uses an
# interaction term to get around collinearity issues.  Likely meets the criteria
# for linear regression althought there is slight heteroskedascity; residuals are
# still fairly normal.


# Claims to have a fit of r^2=0.8, but the predictions vs fitted plot insicates
# this may not hold.  Also looks heteroskedastic.


## Final model 1

# Also linear regression using an interactin term to avoid collinearity issues,
# but with exponenial transformations of the response and predictor variables.
```

```r
# It's a somewhat better fit, with holds true when predictions are plotted
# against residuals.  Still has somewhat heteroskedastic residuals.


df <- load_chi2018()
df$child_mortality <- log(df$child_mortality)
df$sanitation_score <- df$sanitation_score^4
df$water_score <- df$water_score^4
splitdex <- createDataPartition(df$child_mortality, p = 0.8, list = FALSE)
train <- df[splitdex, ]
val <- df[-splitdex, ]
fit <- lm(child_mortality ~ I(water_score + sanitation_score), data = train)
summary(fit)
plot(fit)
plot(predict(fit, val) ~ val$child_mortality)
hist(fit$residuals, breaks = 20)
print(paste("Validation R^2: ", cor(predict(fit, val), val$child_mortality)^2))


## Final Model 2

# A ridge regression model (handles collinearity) with exponential
# transformations on the variables.  Fits about as well as the previous model,
# but maybe less heteroskedascity?  Residuals are kinda normally distributed, but
# maybe a little bimodal.


df <- load_chi2018()
df$child_mortality <- log(df$child_mortality)
df$sanitation_score <- df$sanitation_score^5
df$water_score <- df$water_score^5
splitdex <- createDataPartition(df$child_mortality, p = 0.8, list = FALSE)
train <- df[splitdex, ]
val <- df[-splitdex, ]
library(glmnet)
train_X <- model.matrix(~water_score + sanitation_score, data = train)
train_Y <- train$child_mortality
val_X = model.matrix(~water_score + sanitation_score, data = val)
# Makes a series of crossvalidated glmnet models for 100 lambda values (default)
# lamba values are constants that define coefficient shrinkage.
ridge_model <- cv.glmnet(x = train_X, y = train_Y, family = "gaussian", nfolds = 10,
    type.measure = "mse", alpha = 0)
# setting lambda.min uses the lambda value with the minimum mean cv error (picks
# the best model)
predictions <- predict(ridge_model, newx = val_X, type = "response", s = ridge_model$lambda.min)
# Print's the coefficients the model uses
print(coef.glmnet(ridge_model, s = ridge_model$lambda.min))
# r^2 :
# https://stats.stackexchange.com/questions/266592/how-to-calculate-r2-for-lasso-glmnet
r2 <- ridge_model$glmnet.fit$dev.ratio[which(ridge_model$glmnet.fit$lambda == ridge_model$lambda.min)]
print(paste("R^2: ", r2))
# Correct for transformation
predictions <- predictions
```

```r
residuals <- val$child_mortality - predictions
plot(predictions ~ val$child_mortality)
plot(residuals ~ predictions)
hist(residuals, breaks = 30)
print(paste("Validation R^2: ", cor(predictions, val$child_mortality)^2))


## Adding 'world' data


# Checking for collinearity
library(rnaturalearth)
# Adds new data
world <- ne_countries(scale = "medium", returnclass = "sf")
df <- load_chi2018()
df <- merge(df, world, by.x = "ISO3", by.y = "iso_a3")
# Preserved all of the countries in the origional set
corrplot(cor(df[c("water_score", "sanitation_score", "child_mortality", "pop_est",
    "gdp_md_est")], use = "pairwise.complete.obs"), type = "upper")


# population and and the other new variable seem uncorrelated to child mortality,
# but somehwat correlated to eachother; probably best to just use one if either.
# We can also an economy variable (categorical), of which there are two which are
# presumably fairly correlated



world <- ne_countries(scale = "medium", returnclass = "sf")
df <- load_chi2018()
df <- merge(df, world, by.x = "ISO3", by.y = "iso_a3")
df$child_mortality <- log10(df$child_mortality)
# Still needs a transformation
splitdex <- createDataPartition(df$child_mortality, p = 0.8, list = FALSE)
train <- df[splitdex, ]
val <- df[-splitdex, ]
fit <- lm(child_mortality ~ economy, data = train)
summary(fit)
plot(fit)
plot(predict(fit, val) ~ val$child_mortality)
hist(fit$residuals, breaks = 20)

# Economy seems to be fairly well correlated with child_mortality. Pop_est was
# also tested and proved to be highly unpredictive.

# Let's see what economy + the previous variables can acomplish

## Final model 3


world <- ne_countries(scale = "medium", returnclass = "sf")
df <- load_chi2018()
df <- merge(df, world, by.x = "ISO3", by.y = "iso_a3")
```

```r
df$child_mortality <- log(df$child_mortality)
df$sanitation_score <- df$sanitation_score^4
df$water_score <- df$water_score^4
splitdex <- createDataPartition(df$child_mortality, p = 0.8, list = FALSE)
train <- df[splitdex, ]
val <- df[-splitdex, ]
fit <- lm(child_mortality ~ economy + I(water_score + sanitation_score), data = train)
summary(fit)
plot(fit)
plot(predict(fit, val) ~ val$child_mortality)
hist(fit$residuals, breaks = 20)
print(paste("Validation R^2: ", cor(predict(fit, val), val$child_mortality)^2))


# Very well fitting; adding economy helps.  Maybe slighltly heteroscedastic
# residuals, although they are still normally distributed.

## Final model 4


world <- ne_countries(scale = "medium", returnclass = "sf")
df <- load_chi2018()
df <- merge(df, world, by.x = "ISO3", by.y = "iso_a3")
df$child_mortality <- log(df$child_mortality)
df$sanitation_score <- df$sanitation_score^3
df$water_score <- df$water_score^3
splitdex <- createDataPartition(df$child_mortality, p = 0.8, list = FALSE)
train <- df[splitdex, ]
val <- df[-splitdex, ]
library(glmnet)
train_X <- model.matrix(~water_score + sanitation_score + economy, data = train)
train_Y <- train$child_mortality
val_X = model.matrix(~water_score + sanitation_score + economy, data = val)
# Makes a series of crossvalidated glmnet models for 100 lambda values (default)
# lamba values are constants that define coefficient shrinkage.
ridge_model <- cv.glmnet(x = train_X, y = train_Y, family = "gaussian", nfolds = 10,
    type.measure = "mse", alpha = 0.5)
# Alpha = 0 is ridge. setting lambda.min uses the lambda value with the minimum
# mean cv error (picks the best model)
predictions <- predict(ridge_model, newx = val_X, type = "response", s = ridge_model$lambda.min)
# Print's the coefficients the model uses
print(coef.glmnet(ridge_model, s = ridge_model$lambda.min))
# r^2 :
# https://stats.stackexchange.com/questions/266592/how-to-calculate-r2-for-lasso-glmnet
r2 <- ridge_model$glmnet.fit$dev.ratio[which(ridge_model$glmnet.fit$lambda == ridge_model$lambda.min)]
print(r2)
# Correct for transformation
predictions <- predictions
residuals <- val$child_mortality - predictions
mean(residuals)
plot(predictions ~ val$child_mortality)
plot(residuals ~ predictions)
# ridge_model$glmnet.fit
```

```r
hist(residuals, breaks = 30)
print(paste("Validation R^2: ", cor(predictions, val$child_mortality)^2))
```