Final Project: Child Mortality

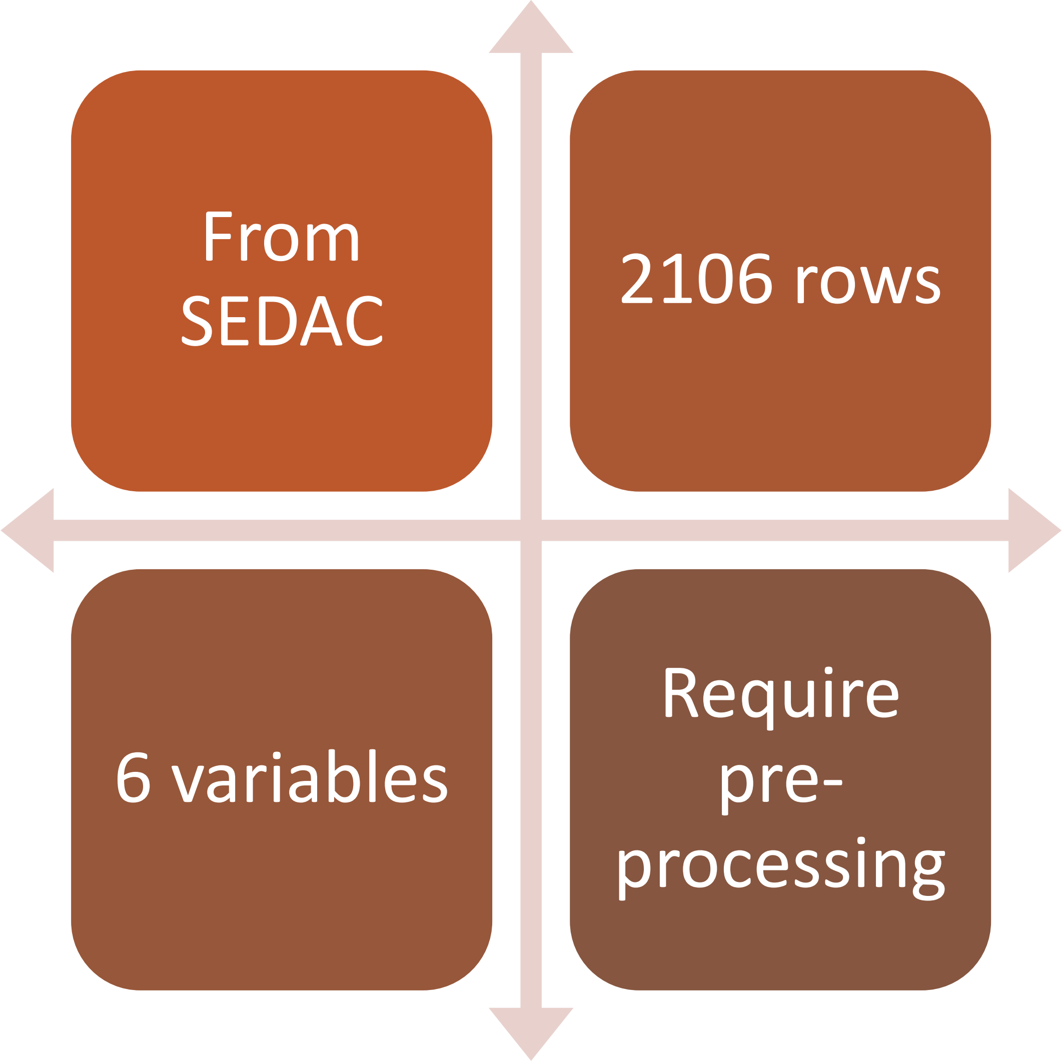
GRPUP-01

ADAM GERSOWITZ,
CHRISTOPHER BLOOME,
DAVID BLUMENSTIEL,
FORHAD AKBAR,
JEYARAMAN
RAMALINGAM,KEVIN
POTTER

Background and motivation

- Explore if 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 country.
- Examining whether access to “at least basic services” for both sanitation conditions and water access will impact child mortality rates
- 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.

Data



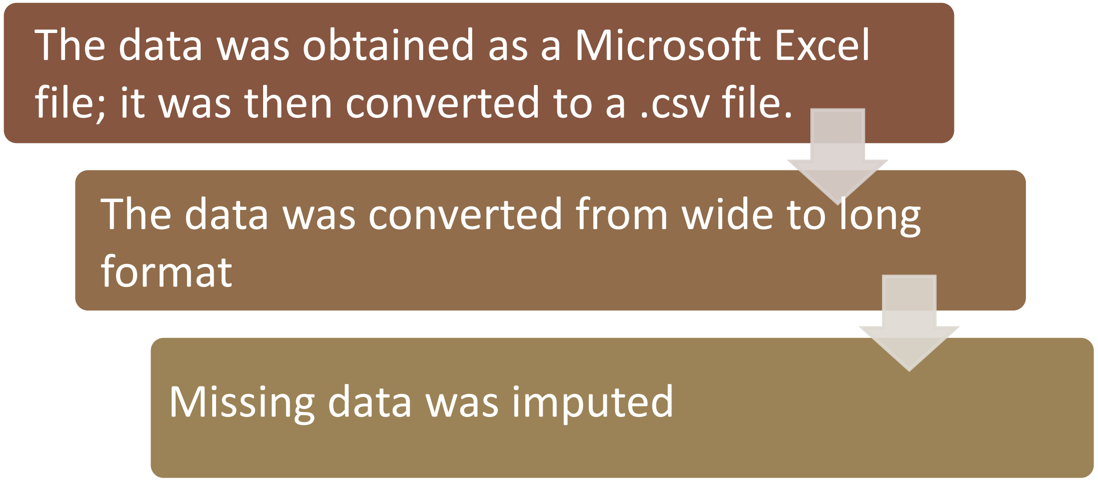
From SEDAC

2106 rows

6 variables

Require pre-processing

Data pre-processing

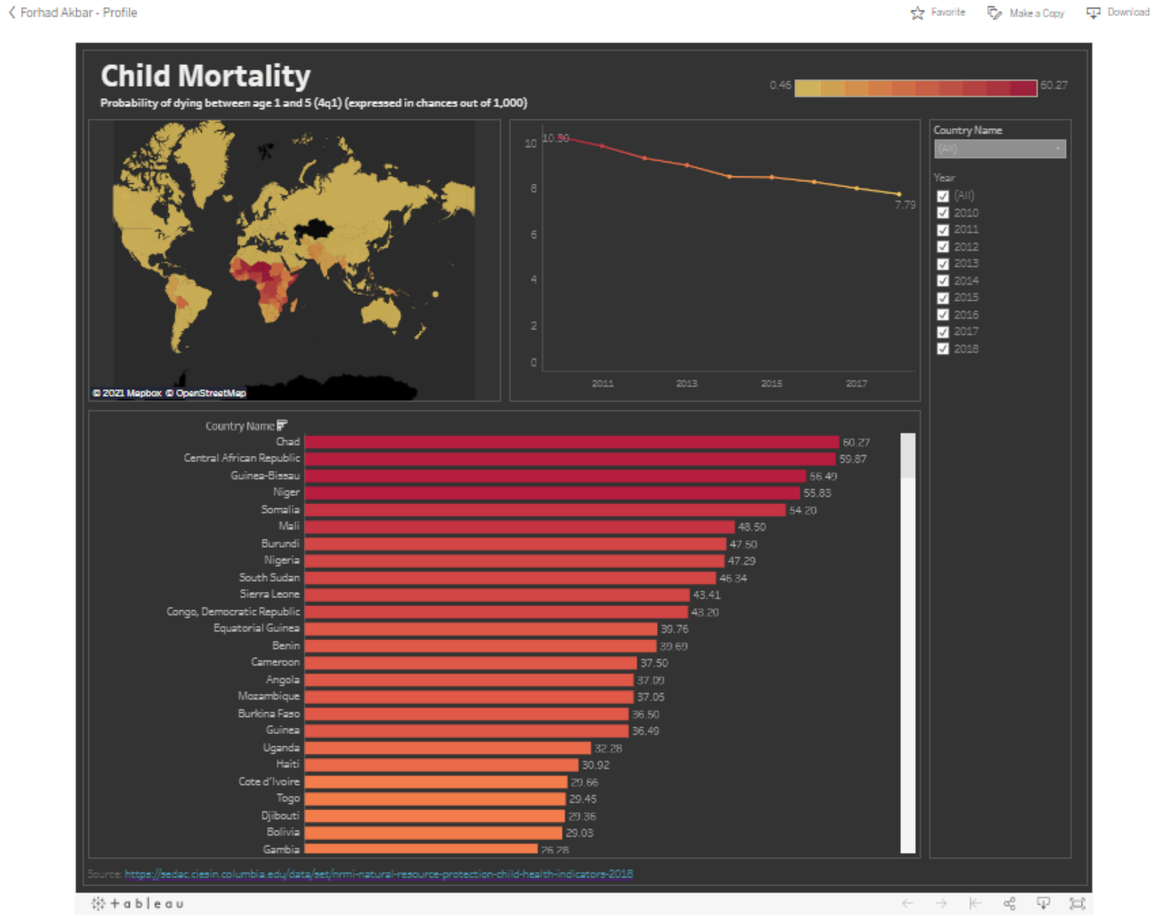


The data was obtained as a Microsoft Excel file; it was then converted to a .csv file.

The data was converted from wide to long format

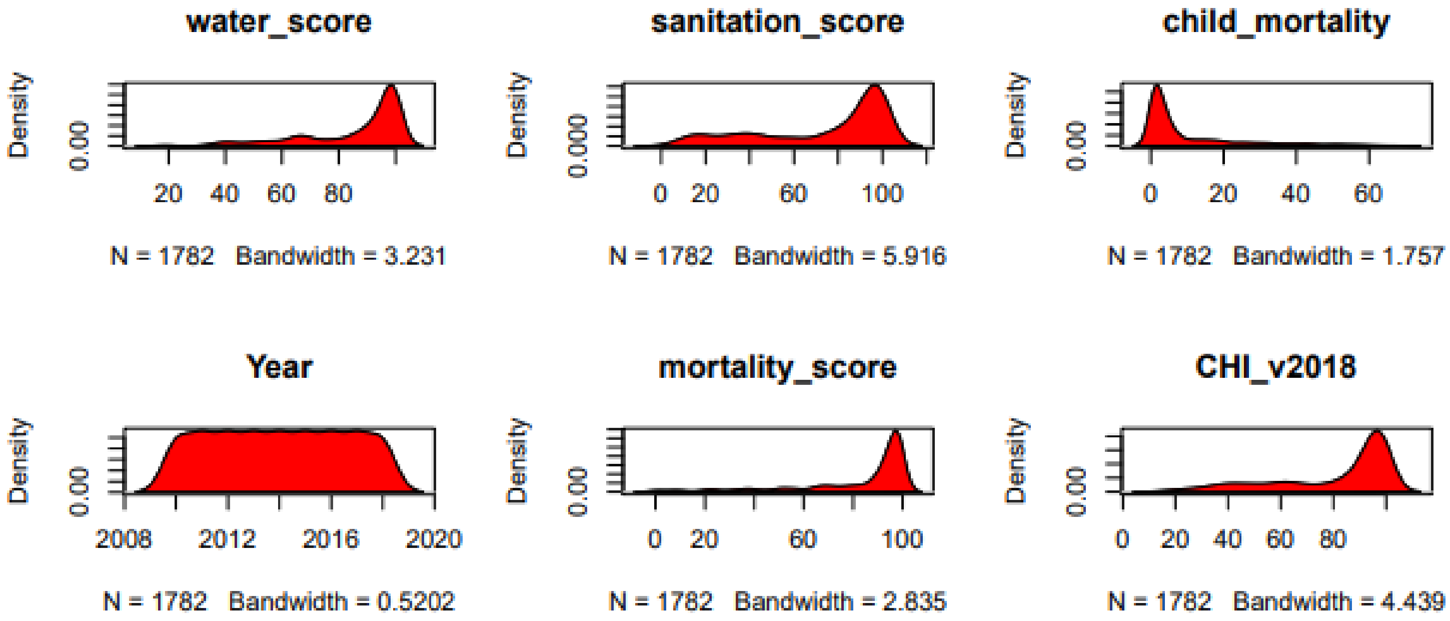
Missing data was imputed

Exploratory Data Analysis

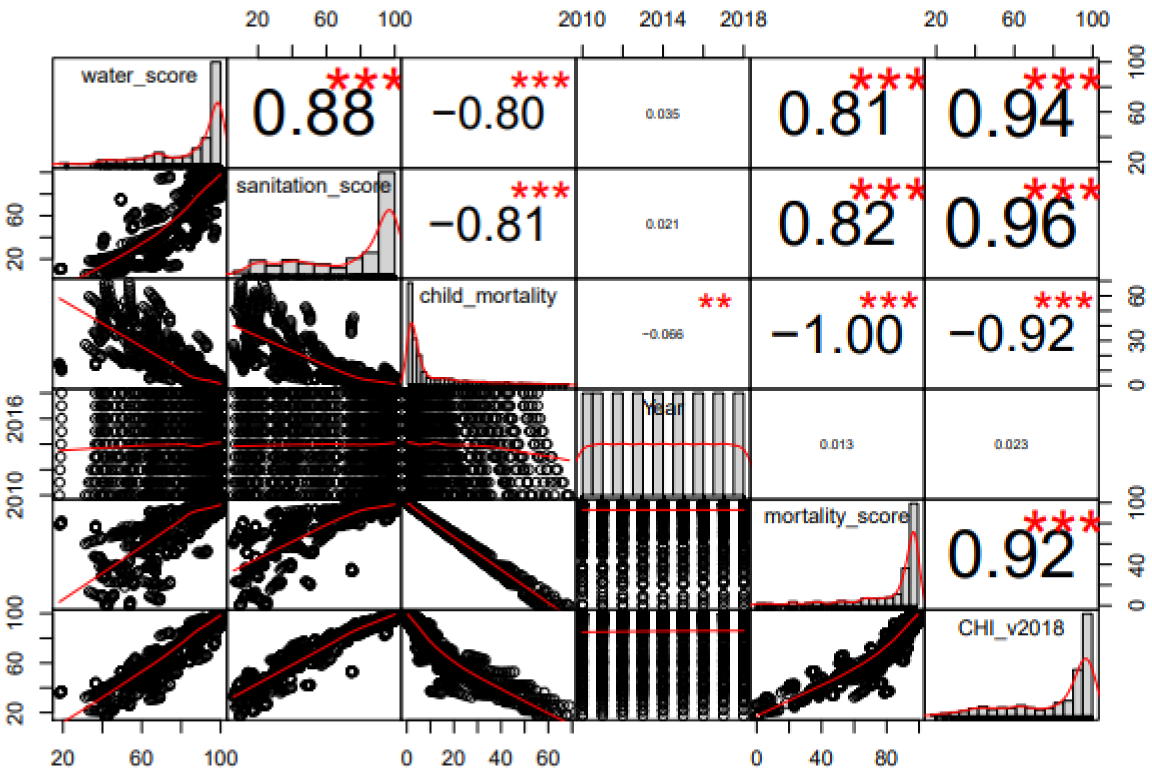


Interactive Tableau Dashboard:
https://public.tableau.com/views/ChildMortality_16211276461000/ChildMortality?:language=en&:display_count=y&:origin=viz_share_link

Exploratory Data Analysis



Correlation



Models

Model 1: Liner regression with water-score and sanitation-score as independent variables

Model 2: ridge regression with water-score and sanitation-score as independent variables

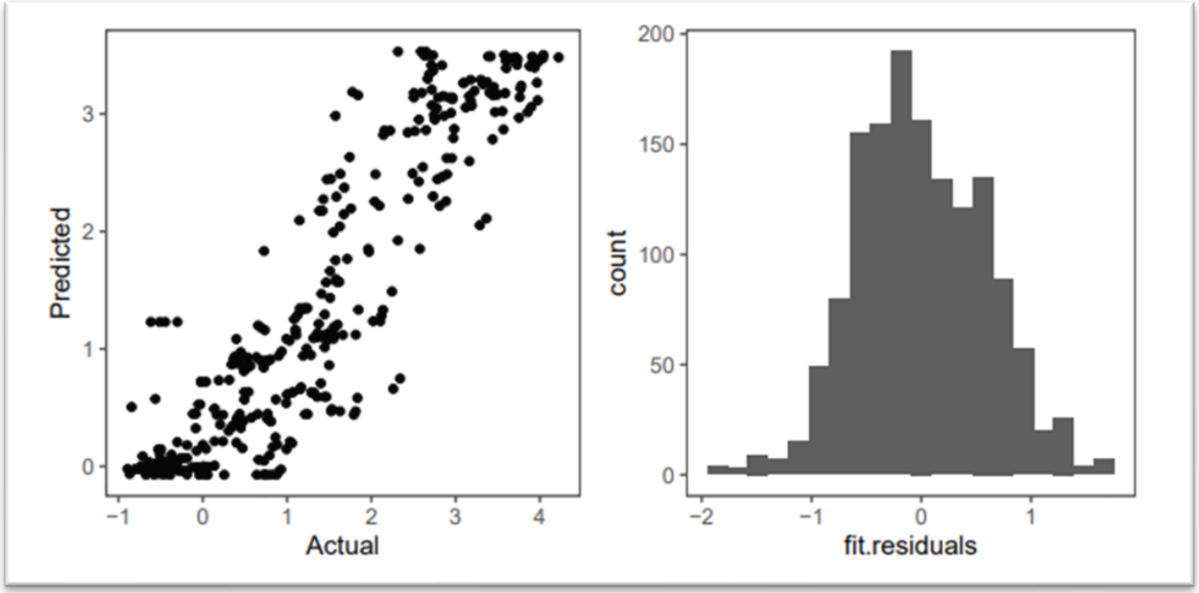
Model 3: Liner regression with water-score, sanitation-score and economy as independent variables

Model 4: Elastic-net regression with water-score, sanitation-score and economy as independent variables

Model 1: Liner regression with water-score and sanitation-score as independent variables

Key factors:

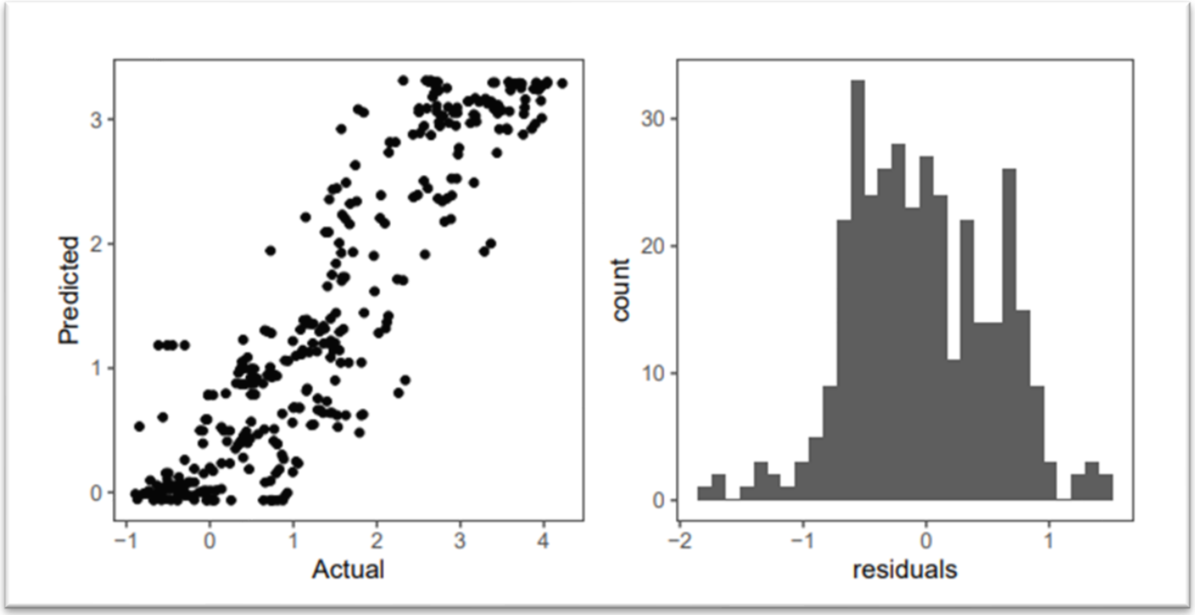
- The model achieved an r-squared value of 0.826, and when predictions were performed on the holdout set, the predictions fit the actual values with an r-squared of 0.821.
- Overall, the model meets the assumptions of linear regression
- It accurately predicts fewer deaths with higher water-score and sanitation-score, indicating that these variables are impactful towards child mortality



Model 2: ridge regression with water-score and sanitation-score as independent variables

Key factors:

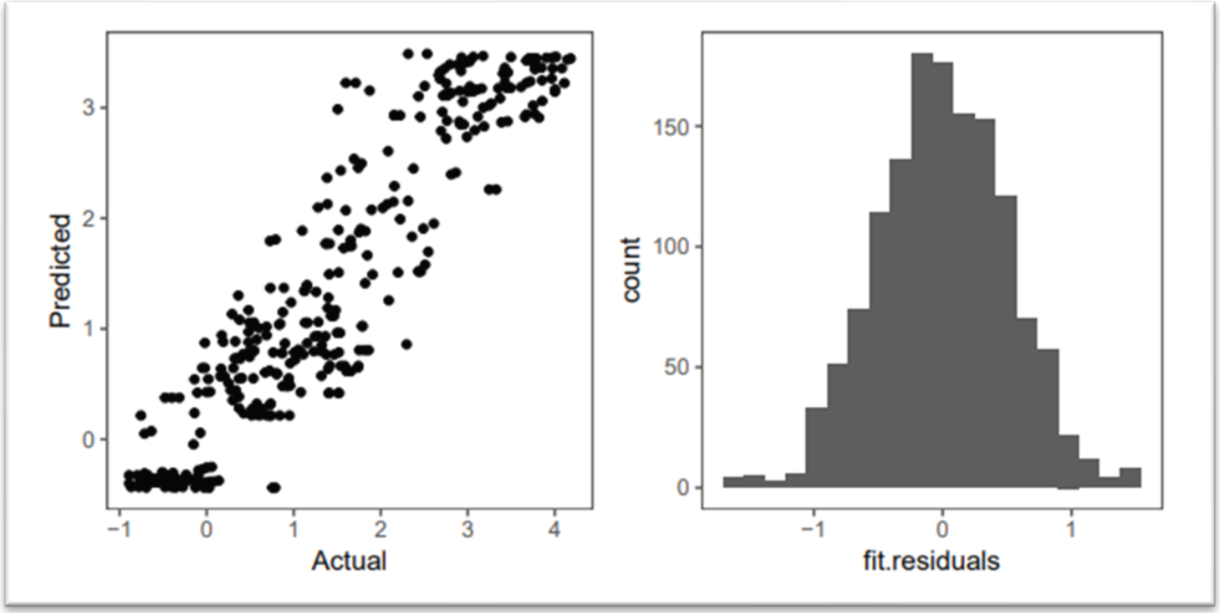
- The model achieved a fit of 0.824, and a validation fit of 0.822 , indicating a good fit without overfitting
- One interesting finding is that this model weighed water-score more than sanitation-score; the coefficient for water-score was about 1.29 times higher than for sanitation-score.
- This could indicate that access to "improved water sources" nearby one's dwelling might be more important for reducing child mortality than sanitation



Model 3: Liner regression with water-score, sanitation-score and economy as independent variables

Key factors:

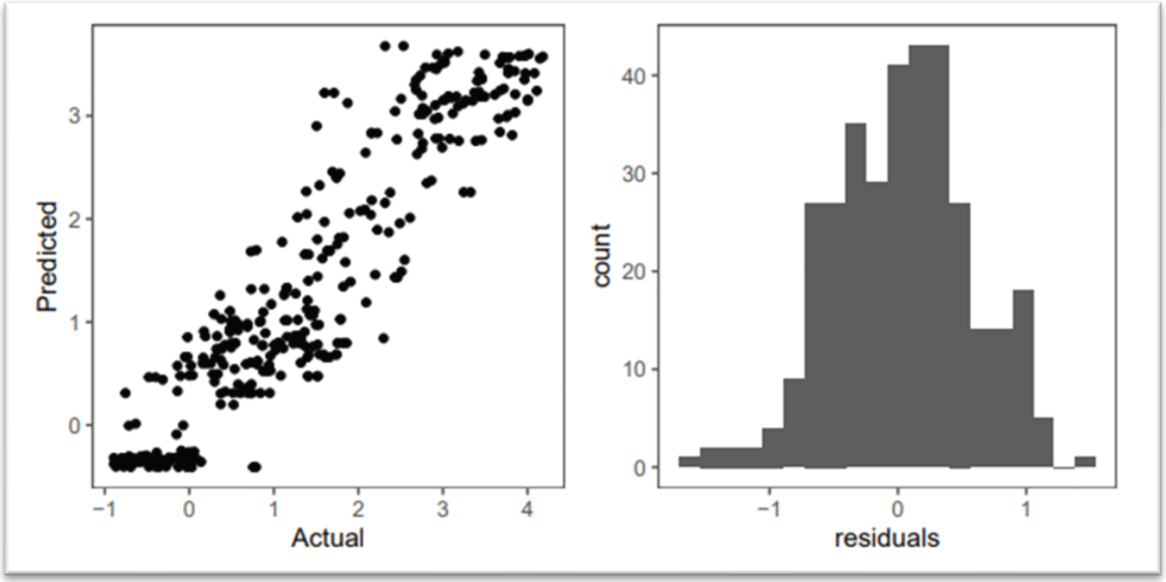
- This model achieved an r-squared value of 0.869, and a validation r-squared of 0.851, indicating some but very little overfitting, and a high overall fit
- This model expects more child mortality for the least developed
- Residuals are normally distributed, with slight heteroscedasticity



Model 4: Elastic-net regression with water-score, sanitation-score and economy as independent variables

Key factors:

- 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.



Model Selection

Key factors:

- 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 most of the 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

All models did something well

Best Model 4: Elastic-net regression

Conclusion

- A model can predict with some accuracy child mortality within a country

- Overall, access to improved water and sanitation can explain most of the 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