A Regression Study of Children Mortality Rate

Yiu Chung, WONG[1], Tsz Wing, WONG[1], Yiu Fung Frankie, CHAN[1], & Wai Lan, LI[1]

[1] The Chinese University Of Hong Kong

Author Note

Yiu Chung Wong, Department of Statistics, The Chinese University Of Hong Kong.

Tsz Wing Wong, Department of Statistics, The Chinese University Of Hong Kong.

Yiu Fung Frankie Chan, Department of Statistics, The Chinese University Of Hong Kong.

Wai Lan Li, Department of Statistics, The Chinese University Of Hong Kong.

This report was supported in part by class from the M.Sc. in Data Science and Business Statistics Program, Department of Statistics, The Chinese University Of Hong Kong.

Correspondence concerning this article should be addressed to Yiu Chung, WONG, 603A, Wong Foo Yuan Blg, Chung Chi RD, The Chinese University Of Hong Kong.
E-mail: s1155017920@link.cuhk.edu.hk

Abstract

According to Worldbank, “mortality indicators are important indicators of health status in a country.” Data on the incidence and prevalence of diseases are frequently unavailable. A prediction in mortality rate can help identify vulnerable populations. The present study investigates children mortality rates using the dataset obtained from The World Bank: World Development Indicators. A Linear Model is used to perform a prediction analysis on the response variable: Mortality rate, using 6 predictor variables. The goal is to identify a subset of features in the dataset which best predicts Mortality rate of children under the age of five. Here we show the mortality rate of children under the age of 5 can be resonibly predicted by 6 other variables. The result identifies previously unknown relationship between children mortality rate and other existing variables. This shall allow researcher to better understand the relationship between children mortality and other variables. For example, data on the incidence and prevalence of diseases are frequently unavailable. A prediction in mortality rate can help identify vulnerable populations.

Keywords: Mortality, Develope, Regression, Lasso, Best subset

Preliminary Data Manipulation

The World Bank data set contains relevant data from 214 countries and jurisdictions for the year 2010, covering 36 variables. Please refer to the appendix for details about the variables.

First remove a few variables: `Year`, `YearCode`, `Country Name`, `Country Code`

These variables are included for naming purpose and contribute no added value to our subsequent analysis.

The variable `Age dependency ratio (% of working-age population)`"`includes people who are below 15 or above than 64. Mean while, the variable`Age dependency ratio, young (% of working-age population)`only includes people below 15. To separate these two, we will subtract the send from the first. The original "Age dependency ratio (% of working-age population)" is renamed to`"`Age dependency ratio (% of working-age population)`.

Missing Data

The data set contains some obviously problematic data. For instance, some variables contains more then 90% missing data. Here, we remove variables with more than 5% missing data, and remove cases with more than 5 missing fields. These cut offs are set objectively deeming that any variables or cases with certain number missing data would undermine the its usefulness. After removal, 214 cases and `ncol(world_bank)` variables remains.

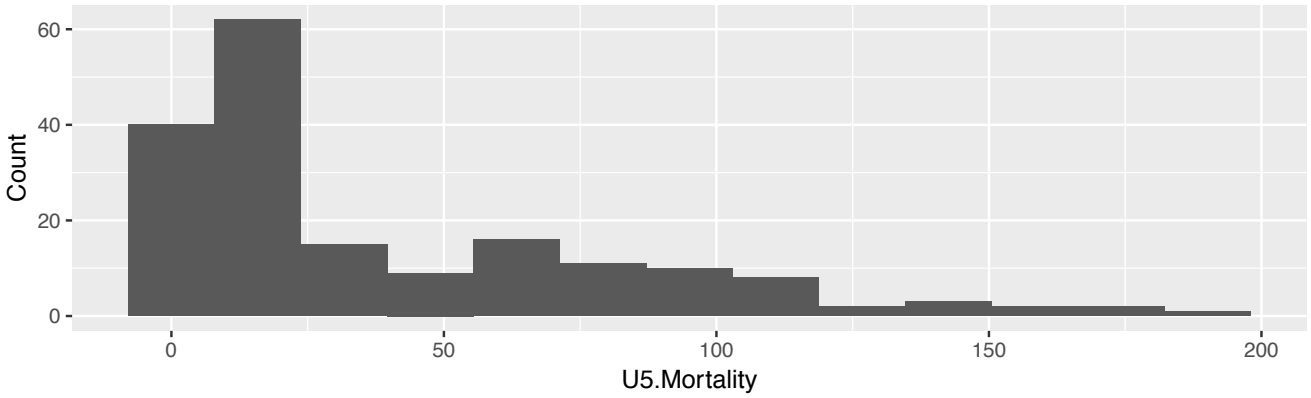
Please refer to the Appendix for an updated list of variables after variable removal.

Impute Data with Multiple Imputation by Chained Equations (MICE)

Missing data are imputed using Multiple (or also called Multivariate) Imputation by Chained Equations (MICE). Many variables in the original dataset are linear combinations of others. Hence parametric approach for estimating missing data is impossible because the dataset is singular. This report uses data imputed by means of Random Forest, which is a non-parametric method. Random Forest has an additional benefit of not requiring data to be missing at random (which is otherwise required for parametric MICE) (Shah, Bartlett, Carpenter, Nicholas, & Hemingway, 2014).

Exploratory Data Analysis

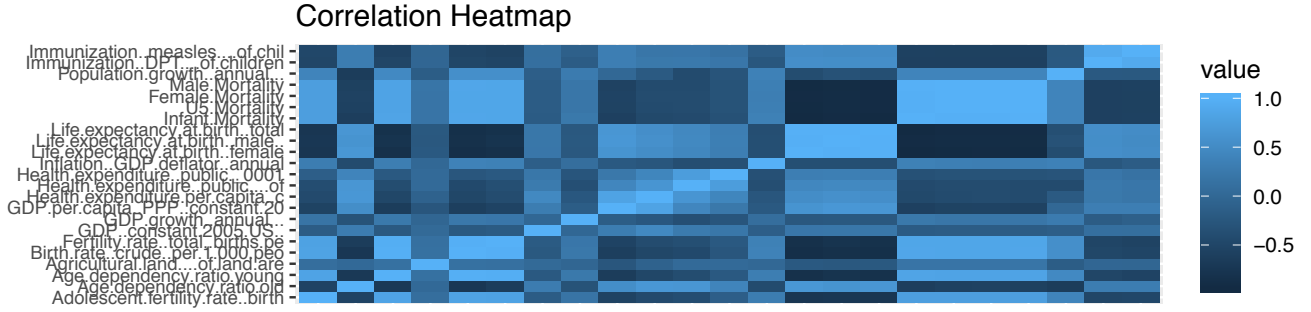
Distributions of Mortality Rate Under 5, Per 1000/Births



The response variable for this project is the quantitative variable, **Mortality Rate Under 5, Per 1000/Births**; let's call it `U5.Mortality` from here onward. `U5.Mortality` is measured in every 1000 deaths per births. While rate data is theoretically infinitely divisible, `U5.Mortality` is Poisson distributed because it is derived from count data. There are decimal points only because it is divided by 1,000 for easy comparison. The response variable having Poisson properties will violate many of the linear regression assumptions. This will be discussed in a later session. Exploring the relationship between `U5.Mortality` and each of the predictor variables would very much be beneficial. However, there are still 23 variables remaining in the dataset, even after deletion. Hence we will focus on the predictor variables instead.

Correlation Heat Map

A heatmap give a rough idea to how variables correlate to each other. The lighter the colour, the higher the correlation between two variables. Here, the bottom tick labels are omitted since variables are grouped by names anyway. Variables have high correlation with neighboring variables. This is expected as variables which are similar in nature have similar names. What's interesting is that instead of variables correlate individually, sets of variable often correlate highly with each other. This can be seen in the graph as there are many light blue boxes and dark blue boxes. For example, all Mortality related variables have high correlation with all Life Expectancy related variables. This can be a problem because there are many variables that are highly correlated to each other. Worst, there may be many variables that are linear combinations of others. One must be judicious when choosing variable in a regression analysis to avoid (multi)collinearity.



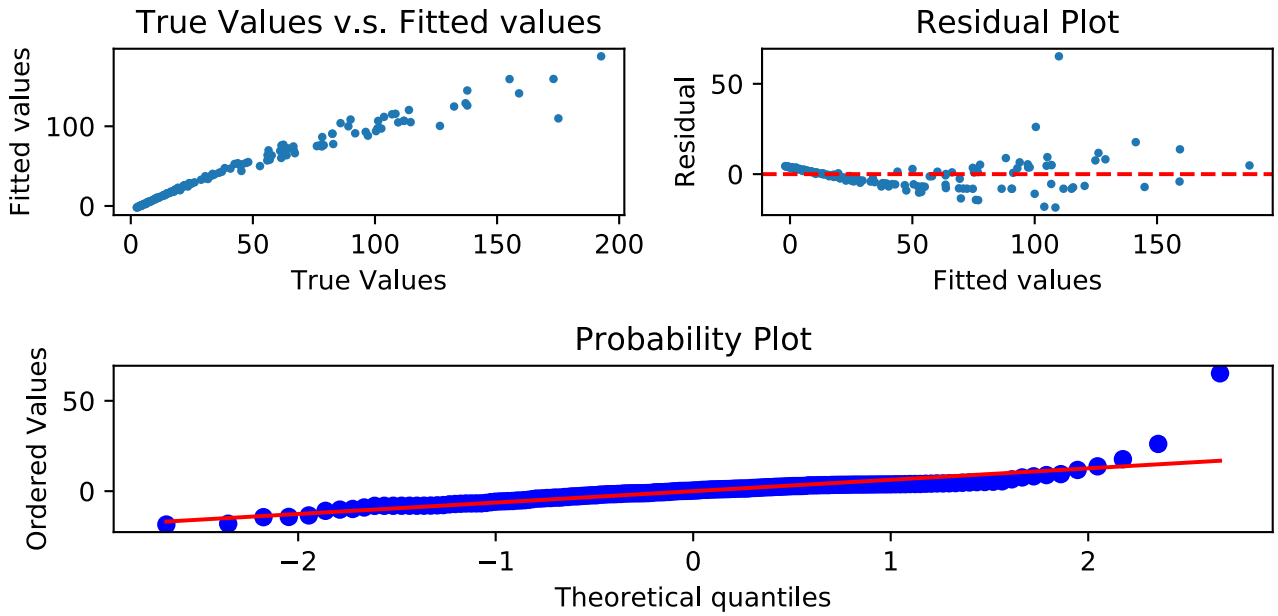
Exploring correlations with OLS regression

Let's start by looking at the response variable, i.e. the Mortality rates

From the correlation heatmap, all Mortality related variables are highly correlated. Highly correlated predictors can lead to instability in our estimator, as well as increased variance.

Not only do these variables co-vary, their value are also extremely close to each other. It is very tempting to just use a one of these variables to predict the response variable. Let's pick one of these predictors to predict `U5.Mortality`? Here `Mortality rate, infant (per 1,000 births)` is used.

```
## lm(formula = U5.Mortality ~ Infant.Mortality, data = world_bank_imputed)
```



Normality Test

Error in a regression model are expected to be i.i.d. Both Shapiro–Wilk (Shapiro & Wilk, 1965) and Anderson-Darling (Anderson & Darling, 1952) tests tests the null hypothesis that a statistical sample $x_1, \ldots, x_n$ came from a normally distributed population. Here, these tests are used to test for residuals normality. Normality tests tend to be not very statistically powerful, hence more than one tests are used.

H_0: Residual follow a normal distribution

H_1: Residual do not follow a normal distribution

```
## [1] "According to Shapiro-Wilk test of normality, and alpha at .01,"
## [1] "residuals do not look Gaussian (reject H0)"
```

```
## [1] "According to Anderson-Darling test for the composite hypothesis of normality, and alpha at .01,"
## [1] "residuals do not look Gaussian (reject H0)"
```

Heteroskedasticity test

Residuals are expected to be not correlated with the predictor variable. The **Breusch–Pagan test** (Breusch & Pagan, 1979) is used to test for heteroskedasticity in a linear regression model.

H_0: Variance of the residuals do not dependent on the values of the independent variables.

H_1: Variance of the residuals is dependent on the values of the independent variables.

```
## [1] "According to Breusch-Pagan test against heteroskedasticity, and alpha at .01,"
## [1] "residuals do not look constant (reject H0)"
```

The assumptions mentioned has to be confirmed before any testing conclusion can be drawn. None of the Normality test passed; the key assumptions of regression are valid in this model.

The bottom line is, we cannot use Simple Linear Regression (OLS) simply because high correlation between predictor and outcome. OLS will not be further discussed.

How about the other two mortality indicators? After some eyeball investigation, it is not difficult to see that the response variable is merely the average of Male/Female mortality rate. Hence these two variables will not be included in analysis; they should not be included in any formal analysis because one could argue that this is a form of data leakage. However, it might be interesting to see how the DIFFERENCE in Male/Female mortality rate affects the response variable. Therefore a new variable is added: DIFFERENCE in Male/Female mortality rate. Note that this is a linear combination of two variables which is originally highly correlated to the response. The outcome is still highly correlated to the response but having one fewer variable is likely to reduce variance.

Finally, because of the following:

- infant mortality rate is highly correlated to our response
- they are measured in the same unit
- almost no difference in absolute value
- domain knowledge tell us that they are indeed almost the same measure

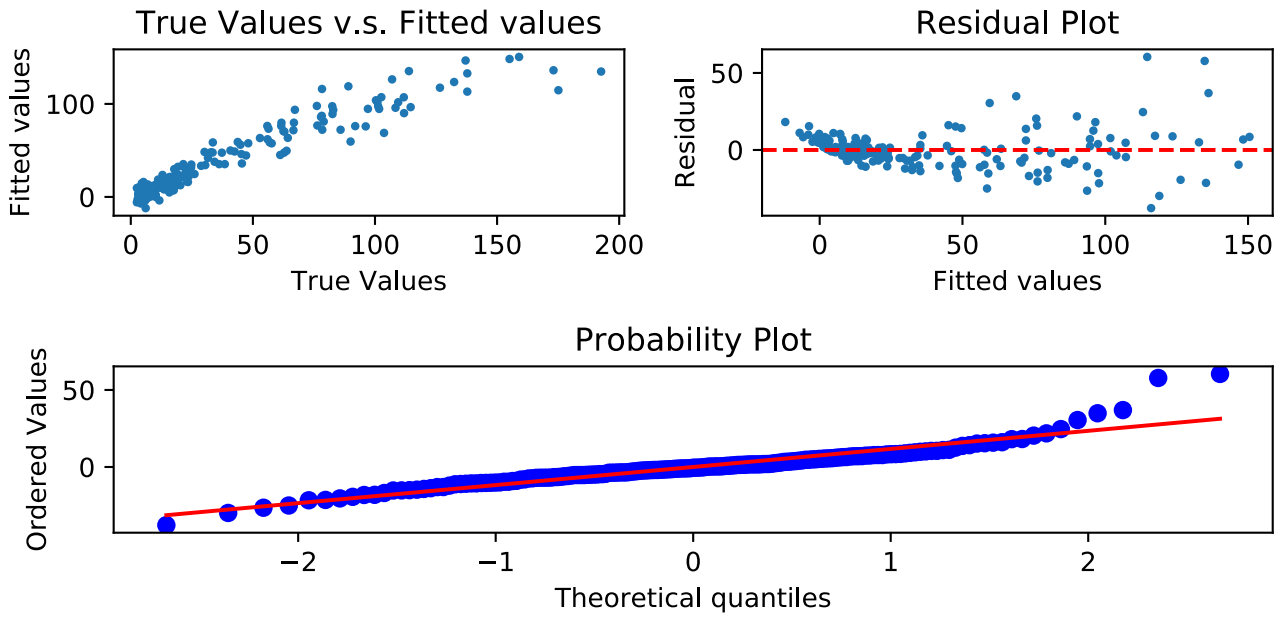
infant mortality rate is removed.

The same is done for Life Expectancy: "Life expectancy at birth", both male and female, are combined and is called `Life.exp.gender.diff`.

Multiple Regression

Perhaps a set of predictor variables can be better explain the variations in Mortality rate?

```
## lm(formula = U5.Mortality ~ ., data = world_bank_imputed)
```



Normality Test

```
## [1] "According to Shapiro-Wilk test of normality, and alpha at .01,"
## [1] "residuals do not look Gaussian (reject H0)"
```

```
## [1] "According to Anderson-Darling test for the composite hypothesis of normality, and alpha at .01,"
## [1] "residuals do not look Gaussian (reject H0)"
```

Heteroskedasticity test

```
## [1] "According to Breusch-Pagan test against heteroskedasticity, and alpha at .01,"
## [1] "residuals do not look constant (reject H0)"
```

Again, none of the assumption tests passed; the key assumptions of multiple regression are not valid in this model.

Multicollinearity and VIF

Multicollinearity exists when a predictor variable is a linear combination of other(s) predictor variables. Multicollinearity can result in instability of regression coefficient: the estimated regression coefficient of one variable sway hugely when new predictors are placed. Statistical power also suffers due to inflated standard error: proving regression coefficient to be useful become more difficult. The overall effect is reduced model precision.

Variance Inflation Factor (VIF) is a score used to measure the extend of which (multi)colinearity exist among predictor variables. This measure is easy to reproduce.

To calculate the VIF of the first predictor variable (X_1): perform OLS regression on X_1 as a function of all the other predictor variables:

$X_i = \alpha_0 + \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_k X_k + e$

Input R^2 of the model above into the formula below to calculate VIF:

$$\text{VIF}_\text{i} = \frac{1}{1 - R_i^2}$$

```
## Birth.rate..crude..per.1.000.peo Fertility.rate..total..births.pe
##                         88.37404                         62.12002
```

The first two VIF indicates enormous milticolinearity. For example, `Birth rate crude/1.000 people` has VIF well above 90. Many of these predictor variables are linear combinations of others. They provide no additional information; their presence increase variance in coefficient estimation and reduce statistical power in proving the coefficients are useful. According to Hair et al.(n.d.), a score of 10 and above is considered problematic and that particular variable should be removed.

Regression Model

The work above are of exploratory nature. We found that both OLS and multiple regression result in residuals that are both heteroskedastic and non Gaussian.

The following is our current regression model:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon \tag{1}$$

where

$$\epsilon \sim N(\mu, \sigma^2) \tag{2}$$

A continuous variable Y is being regressed on a set of Xs. The model assumes Y to be the sum of all Xs multiplied by their corresponding coefficients, plus random normal error. This model works well if numbers on the Right Hand Side(RHS) result in the value on the Left Hand Side (LHS). The regression model shown above relies heavily on certain assumptions:

- Homoskedastic Error
- Gaussian Error
- LHS of the model be strictly positive.

The first two assumptions have already been violated.

Note that the sum on the RHS can sometimes be negative depending on the input and coefficients. Unfortunately, the response variable `U5.Mortality` can not take negative values; the model is broken when ever the sum on of the RHS comes out negative. Worst, as mentioned in the exploratory data analysis section, `U5.Mortality` exhibits Poisson properties: the variance in `U5.Mortality` is somewhat correlated to its mean.

Because of this, a log transformation is performed on the response variable.

$$\log(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon \tag{3}$$

A log transformation on the response variable has the following benefits:

1. Values on both sides of the equation may take negative values
2. A log transformation on a Poisson variable decouples the its variance from the mean. This ensures each y_i to be independently distributed.

Simultaneous transformation on predictor variables may be required to maintain linearity and some other assumptions. This will be discussed later.

The confidence interval covers zero, hence the best simple transformation for the response is logarithm.

Feature Selection Using Lasso Regression

There are 181 observations (countries) remaining in the dataset, along with 19 predictors. One might overfit the data because there are few observations, which inflates variance. Feature selection is therefore helpful in reducing the number of variables.

The Lasso regression imposes a L1 penalty on the coefficients: every coefficient is being shrinked, and some even as low as zero. This is a form of feature selection since feature with zero coefficient has no effect in the model.

```
## lm(formula = log(U5.Mortality) ~ ., data = Lasso_train_set[c(lasso_terms,
##     "U5.Mortality")])
```

F test for overall significance

H_0: No linear relationship between Mortality rate and all variables

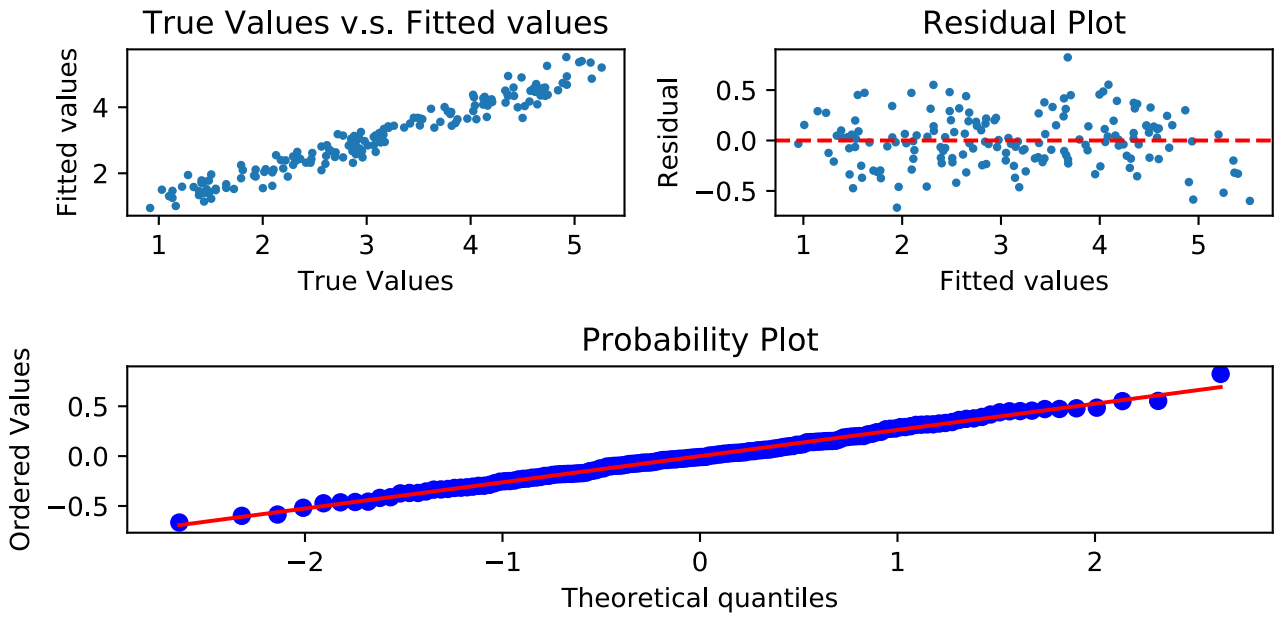
H_1: At least one independent variables affects Y

The p-value is 0, H0 is rejected; independent variables have an overall effect on Y. R^2_{adj} is 0.9466252. The model is able to explain over 94.7% of variance in `U5.Mortality`.

VIF

L1 norm penalisation is successful in removing predictor variables which are linear combinations of other terms. The largest VIFs are both below 10.

```
## Birth.rate..crude..per.1.000.peo         Age.dependency.ratio.old
##                         7.207851                         4.092055
```



Normality Test

```
## [1] "According to Shapiro-Wilk test of normality, and alpha at .01,"
## [1] "residuals look Gaussian (fail to reject H0)"

## [1] "According to Anderson-Darling test for the composite hypothesis of normality, and alpha at .01,"
## [1] "residuals look Gaussian (fail to reject H0)"
```

Heteroskedasticity test

```
## [1] "According to Breusch-Pagan test against heteroskedasticity, and alpha at .01,"
## [1] "residuals look constant (fail to reject H0)"
```

Through the above testing, we can prove the key assumptions of multiple regression are valid in this model.

Using the terms yielded by cross validation Lasso regression, the results looks promising.

- Residuals are Gaussian and constant.
- VIF shows there are no multicolinearity.
- T-test on most individual coefficient estimates, as well as F-test on the overall model are significant.

This model is worth keeping since results are significant and all model assumptions are met.

Best Subset Selection

The following procedure attempts to select the 'best' set of predictor variables used for predicting the response.

Algorithm

Let M_0 denote the null model which contains no predictors, this model simply predicts the sample mean of each observation

- for $k = 1, 2, ...n$

1. Fit all $\binom{n}{k}$ models that contain exactly k predictors
2. Pick the best among these $\binom{n}{k}$ models, and call it M_k. Here the best is defined as having the smallest RSS, or an equivalent measure.

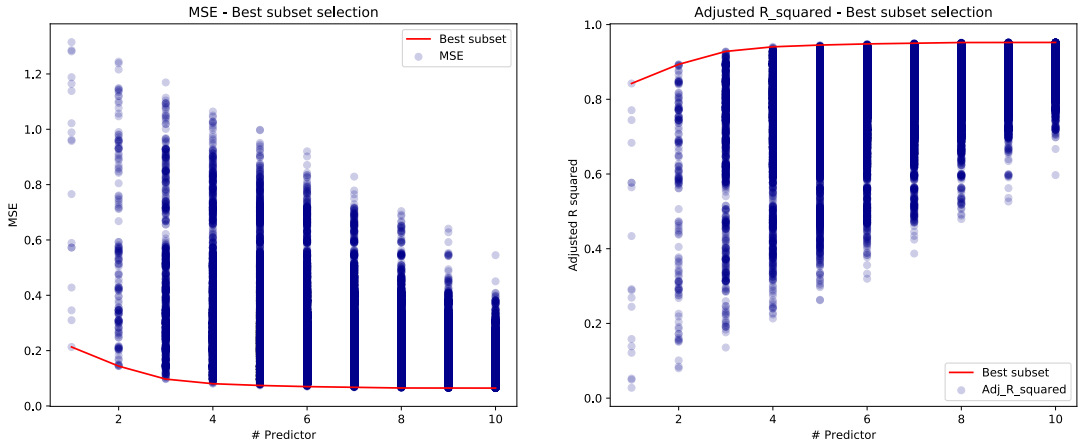
- Select the single best model among M_0, M_1,...,M_n mean squared error and R^2_{adj}.

There is a total of subsets.

Best Subsets

For every k_i (from 1 to 19), we pick out the subset with smallest test MSE, then store it in a dataframe. We also pick out the subset with highest test R^2_{adj}, and store them in another dataframe. We then sort the two data frames by R^2_{adj} and MSE.

Plotting the best subset selection process



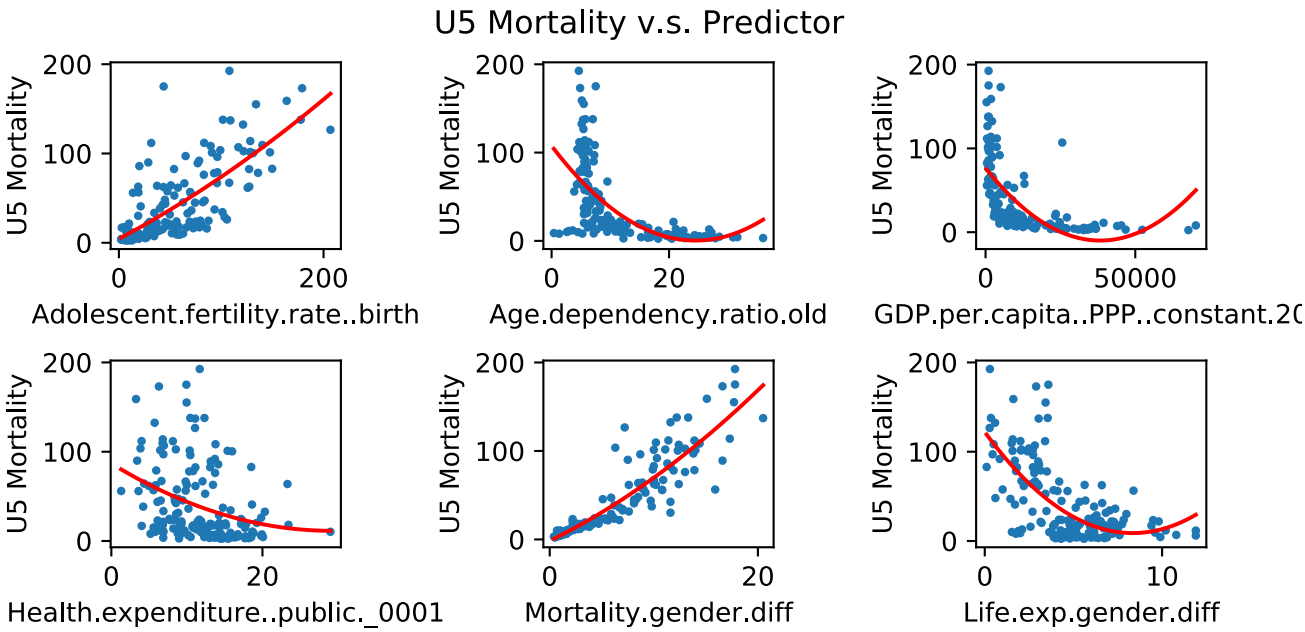
The graph on the left illustrates the mean square error (mse) for each model among each k number of predictor variables. The red line connects models with the lowest mse in each k. The graph on the right shows the same but in R^2_{adj}.

The graphs shows that the more predictor variables are included, the lower the mse and higher the R^2_{adj}. However, both red lines level off at 6 predictors. This suggests the difference in model performance between 6 and more predictor variables are minuscule. In the spirit of parsimony, one should opt for a model with fewer terms. This minimises colinearity and variance in estimates of predictor coefficients.

```
## ['Age.dependency.ratio.old', 'GDP.per.capita..PPP..constant.20']
```

```
## ['Health.expenditure..public._0001', 'Mortality.gender.diff', 'Life.exp.gender.diff']
```

Plot Response variable against Best subset predictors



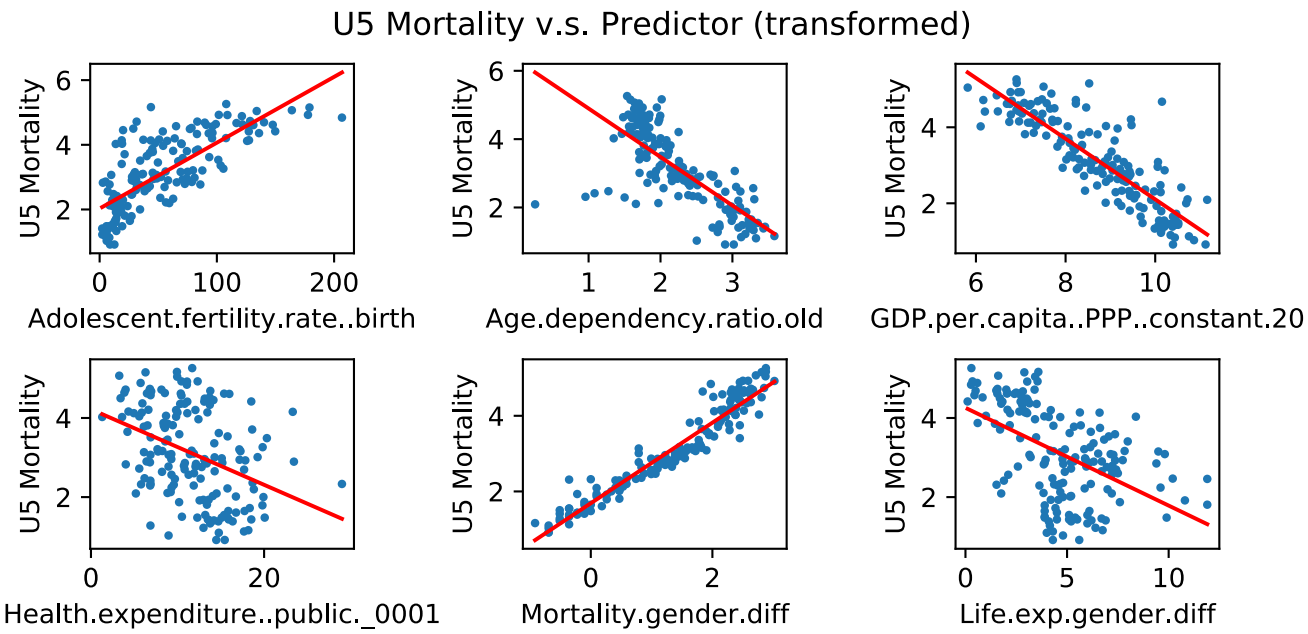
Residuals when Y is regressed on each individual predictor

The regression plot gives some insights:

- The relation between `U5.Mortality` and the predictor variables are highly non-linear. e.g. `GDP per Capita` has a L-shaped relationship with `U5.Mortality`. This suggest a logarithmic relationship.

The following procedures can improve the above issues:

- Log transformation on both the response and predictor to obtain or maintain linear regression relation.
- Remove outliners with high cook distance and high influence.



Much better! In particular, there are no more L-shaped relationship between the response and predictors such as `GDP per Capita`.

Best Subset Multiple Regression

```
## lm(formula = U5.Mortality ~ ., data = log_best_subset)
```

F test for overall significance

H_0: No linear relationship between Mortality rate and all variables

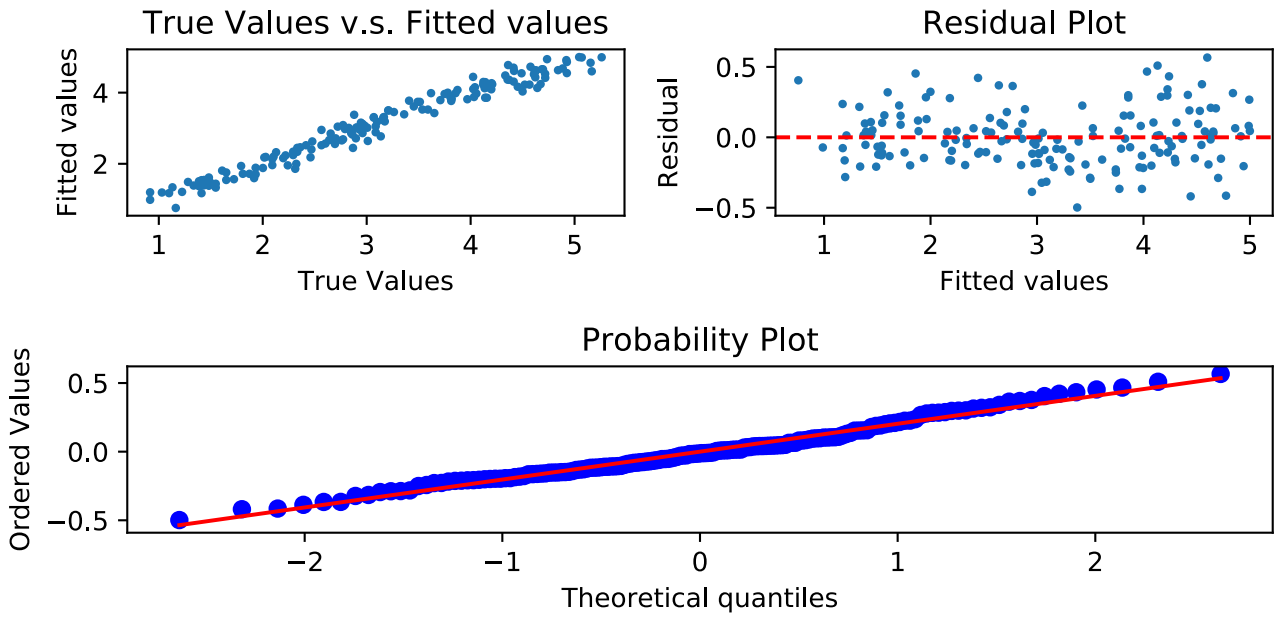
H_1: At least one independent variables affects Y

The p-value is 0, H0 is rejected; independent variables have an overall effect on Y. R^2_{adj} is 0.9685498. The model is able to explain over 96.9% of variance in `U5.Mortality`.

VIF

There don't seem to be any colinearity

```
##            Mortality.gender.diff GDP.per.capita..PPP..constant.20
##                         5.283221                         3.619694
```



Normality Test

```
## [1] "According to Shapiro-Wilk test of normality, and alpha at .01,"
## [1] "residuals look Gaussian (fail to reject H0)"

## [1] "According to Anderson-Darling test for the composite hypothesis of normality, and alpha at .01,"
## [1] "residuals look Gaussian (fail to reject H0)"
```

Heteroskedasticity test

```
## [1] "According to Breusch-Pagan test against heteroskedasticity, and alpha at .01,"
## [1] "residuals look constant (fail to reject H0)"
```

Through the above testing, we can prove the key assumptions of multiple regression are valid in this model.

Model Selection

According to validation, the Best subset model yields RMSE:0.2259276 and R^2: 0.9581884; Lasso model yields RMSE:0.3271431 and R^2: 0.923145 There are less error in Best subset's prediction, as well as more explained variance. Hence Best subset model is the model of choice.

Discussion

Three family member of the Generalised Linear Model are examined in this report: Ordinary Least Square, Multiple Regression, and Lasso Regression. Four models are generated and two are chosen for validation. Two models are

chosen because they consider the non-linear relationship between the response and the predictors. Results from these two models are both satisfactory as they both show statistical significance and agreeing to all assumptions imposed on linear regression. The Best subset model is chosen to be the model of choice because of its performance in validation test.

References

Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain“ goodness of fit” criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 193–212.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, 1287–1294.

Hair, J. (n.d.). JR., anderson, re, tatham, rl, & black, wc (1995). *Multivariate Data Analysis (3rd Ed). New York: Macmillan*.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study. *American Journal of Epidemiology*, *179*(6), 764–774.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

STAT5102-2018-Group-Project Appendix

Final project for 2018R1 Regression in Practice (STAT5102)

Variables

For each variable, you may click 'Details' on the top right corner of each graph for additional information.

- Year (***removed***)
- YearCode (***removed***)
- Country Name (***removed***)
- Country Code (***removed***)
- ARI treatment (% of children under 5 taken to a health provider) (***removed***)
  - Children with acute respiratory infection (ARI) who are taken to a health provider refers to the percentage of children under age five with ARI in the last two weeks who were taken to an appropriate health provider, including hospital, health center, dispensary, village health worker, clinic, and private physician.
  - Link
- Adjusted savings: education expenditure (% of GNI)
  - Education expenditure refers to the current operating expenditures in education, including wages and salaries and excluding capital investments in buildings and equipment.
  - Link
- Adolescent fertility rate (births per 1,000 women ages 15-19)
  - Adolescent fertility rate is the number of births per 1,000 women ages 15-19.
  - Link
- Age dependency ratio (% of working-age population) (***removed***)
  - Age dependency ratio is the ratio of dependents–people younger than 15 or older than 64–to the working-age population–those ages 15-64. Data are shown as the proportion of dependents per 100 working-age population.
  - Link
- Age dependency ratio, young (% of working-age population)
  - Age dependency ratio, young, is the ratio of younger dependents–people younger than 15–to the working-age population–those ages 15-64. Data are shown as the proportion of dependents per 100 working-age population.
  - Link
- Agricultural land (% of land area)
  - Agricultural land refers to the share of land area that is arable, under permanent crops, and under permanent pastures. Arable land includes land defined by the FAO as land under temporary crops (double-cropped areas are counted once), temporary meadows for mowing or for pasture, land under

market or kitchen gardens, and land temporarily fallow. Land abandoned as a result of shifting cultivation is excluded. Land under permanent crops is land cultivated with crops that occupy the land for long periods and need not be replanted after each harvest, such as cocoa, coffee, and rubber. This category includes land under flowering shrubs, fruit trees, nut trees, and vines, but excludes land under trees grown for wood or timber. Permanent pasture is land used for five or more years for forage, including natural and cultivated crops.
  - Link

- Birth rate, crude (per 1,000 people)
  - Crude birth rate indicates the number of live births occurring during the year, per 1,000 population estimated at midyear. Subtracting the crude death rate from the crude birth rate provides the rate of natural increase, which is equal to the rate of population change in the absence of migration.
  - Link

- CPIA gender equality rating (1=low to 6=high) (***removed***)
  - Gender equality assesses the extent to which the country has installed institutions and programs to enforce laws and policies that promote equal access for men and women in education, health, the economy, and protection under law.
  - Link

- Central government debt, total (% of GDP) (***removed***)
  - Debt is the entire stock of direct government fixed-term contractual obligations to others outstanding on a particular date. It includes domestic and foreign liabilities such as currency and money deposits, securities other than shares, and loans. It is the gross amount of government liabilities reduced by the amount of equity and financial derivatives held by the government. Because debt is a stock rather than a flow, it is measured as of a given date, usually the last day of the fiscal year.
  - Link

- Children with fever receiving antimalarial drugs (% of children under age 5 with fever) (***removed***)
  - Malaria treatment refers to the percentage of children under age five who were ill with fever in the last two weeks and received any appropriate (locally defined) anti-malarial drugs.
  - Link

- Fertility rate, total (births per woman)
  - Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.
  - Link

- GDP (constant 2005 US$)
  - GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2010 U.S. dollars. Dollar figures for GDP are converted from domestic currencies using 2010 official exchange rates. For a few countries where the official exchange

rate does not reflect the rate effectively applied to actual foreign exchange transactions, an alternative conversion factor is used.
  - Not available on WorldBank website

- GDP growth (annual %)
  - Annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2010 U.S. dollars. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources.
  - Link

- GDP per capita, PPP (constant 2011 international $)
  - GDP per capita based on purchasing power parity (PPP). PPP GDP is gross domestic product converted to international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as the U.S. dollar has in the United States. GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2011 international dollars.
  - Link

- GINI index (World Bank estimate) (***removed***)
  - Gini index measures the extent to which the distribution of income (or, in some cases, consumption expenditure) among individuals or households within an economy deviates from a perfectly equal distribution. A Lorenz curve plots the cumulative percentages of total income received against the cumulative number of recipients, starting with the poorest individual or household. The Gini index measures the area between the Lorenz curve and a hypothetical line of absolute equality, expressed as a percentage of the maximum area under the line. Thus a Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality.
  - Link

- Health expenditure per capita (current US$)
  - Current expenditures on health per capita in current US dollars. Estimates of current health expenditures include healthcare goods and services consumed during each year.
  - Link

- Health expenditure (% of GDP)
  - Level of current health expenditure expressed as a percentage of GDP. Estimates of current health expenditures include healthcare goods and services consumed during each year. This indicator does not include capital health expenditures such as buildings, machinery, IT and stocks of vaccines for emergency or outbreaks.
  - Link

- Health expenditure, public _ 0001

  - Not available

- Income share held by lowest 20% (***removed***)
  - Percentage share of income or consumption is the share that accrues to subgroups of population indicated by deciles or quintiles. Percentage shares by quintile may not sum to 100 because of rounding.
  - Link

- Inflation, GDP deflator (annual %)
  - Inflation as measured by the annual growth rate of the GDP implicit deflator shows the rate of price change in the economy as a whole. The GDP implicit deflator is the ratio of GDP in current local currency to GDP in constant local currency.
  - Link

- Life expectancy at birth, female (years) (***removed***)
  - Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
  - Link

- Life expectancy at birth, male (years) (***removed***)
  - Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
  - Link

- Life expectancy at birth, total (years)
  - Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
  - Link

- Mortality rate, infant (per 1,000 live births) (***removed***)
  - Infant mortality rate is the number of infants dying before reaching one year of age, per 1,000 live births in a given year.
  - Link

- Mortality rate, under-5 (per 1,000 live births)
  - Under-five mortality rate is the probability per 1,000 that a newborn baby will die before reaching age five, if subject to age-specific mortality rates of the specified year.
  - Link

- Mortality rate, under-5, female (per 1,000 live births) (***removed***)
  - Under-five mortality rate, female is the probability per 1,000 that a newborn female baby will die before reaching age five, if subject to female age-specific mortality rates of the specified year.
  - Link

- Mortality rate, under-5, male (per 1,000 live births) (***removed***)
  - Under-five mortality rate, male is the probability per 1,000 that a newborn male baby will die before reaching age five, if subject to male age-specific mortality rates of the specified year.

  - Link

- Population growth (annual %)
  - Annual population growth rate for year t is the exponential rate of growth of midyear population from year t-1 to t, expressed as a percentage . Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.
  - Link

- Immunization, DPT (% of children ages 12-23 months)
  - Child immunization, DPT, measures the percentage of children ages 12-23 months who received DPT vaccinations before 12 months or at any time before the survey. A child is considered adequately immunized against diphtheria, pertussis (or whooping cough), and tetanus (DPT) after receiving three doses of vaccine.
  - Link

- Immunization, measles (% of children ages 12-23 months)
  - Child immunization, measles, measures the percentage of children ages 12-23 months who received the measles vaccination before 12 months or at any time before the survey. A child is considered adequately immunized against measles after receiving one dose of vaccine.
  - Link

- Physicians (per 1,000 people) (***removed***)
  - Physicians include generalist and specialist medical practitioners.
  - Link

- Women's share of population ages 15+ living with HIV (%) (***removed***)
  - Prevalence of HIV is the percentage of people who are infected with HIV. Female rate is as a percentage of the total population ages 15+ who are living with HIV.
  - Link

Generated Variables

- Age dependency ratio, old (% of working-age population)
- Age dependency ratio minus Age dependency ratio (young).
- Mortality rate, under-5, Gender Difference
- Difference between female and male Mortality rate, under-5, (female minus male).
- Life expectancy at birth, Gender Difference
- Difference between female and male Life expectancy at birth, (female minus male).

Code Appendix

```
#some constants
STAT5102 <- 5102
ALPHA <- 0.010

library("knitr")
library("ggplot2")
library("mice")
library("caret")
library("reticulate")
library("papaja")

use_condaenv(condaenv = "stat5102_", required = TRUE)
opts_chunk$set(echo = FALSE,
               results = FALSE,
               render = TRUE,
               warning = FALSE,
               message = FALSE,
               cache =TRUE,
               prompt = FALSE,
               python = reticulate::eng_python)

set.seed(STAT5102)
#import scipy
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import scipy.stats as stats
from sklearn import linear_model
import itertools

shapiro_test <- function(residuals){
  shapiro <- shapiro.test(residuals)
  print("According to Shapiro-Wilk test of normality, and alpha at .01,")
  if (shapiro$p.value > ALPHA) {
    print("residuals look Gaussian (fail to reject H0)")
    } else {
      print("residuals do not look Gaussian (reject H0)")
    }
  }
```

```
anderson_test <- function(residuals){
  anderson <- nortest::ad.test(residuals)
  print("According to Anderson-Darling test for the composite hypothesis of normality, and alpha at .01,")
  if (anderson$p.value > ALPHA) {
    print("residuals look Gaussian (fail to reject H0)")
  } else {
    print("residuals do not look Gaussian (reject H0)")
    }
}

breusch_pagan_test <- function(model){
  bp <- lmtest::bptest(model)
  print("According to Breusch-Pagan test against heteroskedasticity, and alpha at .01,")
  if (bp$p.value > ALPHA) {
    print("residuals look constant (fail to reject H0)")
  } else {
    print("residuals do not look constant (reject H0)")
      }
}
#some self defined functions

def calculate_mse(y_true, y_pred, n, k):
  resid = (y_true - y_pred)
  rss = np.sum(resid**2)
  mse = rss / (n - (k + 1))
  return mse

def regression_graph(true_values, fitted_values, residuals):
  fig = plt.figure(figsize = (7,3.5))
  ax1 = plt.subplot(221)
  ax2 = plt.subplot(222)
  ax3 = plt.subplot(212)

  ax1.scatter(true_values, fitted_values, s = 5)
  ax1.set_xlabel("True Values")
  ax1.set_ylabel("Fitted values")
  ax1.set_title("True Values v.s. Fitted values")

  ax2.scatter(fitted_values, residuals, s = 5)
  ax2.axhline(y = 0, color='r', linestyle = '--')
  ax2.set_xlabel("Fitted values")
  ax2.set_ylabel("Residual")
  ax2.set_title("Residual Plot")
```

```
  stats.probplot(residuals, dist="norm", plot=ax3)
  plt.tight_layout()
  plt.show()

def response_vs_predictor_plot(X, Y, title, order = 1):
  fig = plt.figure(figsize = (7, 3.5))
  for i in range(0, X.shape[1]):
    ax = fig.add_subplot(2,3,i+1)

    ax.scatter(X.iloc[:,i], Y, s = 5)
    ax.set_xlabel(X.columns.tolist()[i])
    ax.set_ylabel("U5 Mortality")

    weights = np.polyfit(X.iloc[:,i], Y, order)
    model = np.poly1d(weights)
    pred = model(X.iloc[:,i])
    xp = np.linspace(X.iloc[:,i].min(),X.iloc[:,i].max(),100)
    pred_plot = model(xp)
    ax.plot(xp, pred_plot, "r")

  plt.suptitle(title)
  plt.tight_layout()
  fig.subplots_adjust(top=0.9)
  plt.show()

def fit_linear_reg(X,Y):
  n = X.shape[0]
  k = X.shape[1]

  #Fit linear regression model and return MSE and R squared values
  model_k = linear_model.LinearRegression(fit_intercept = True, n_jobs=-1)
  model_k.fit(X, Y)

  #get MSE
  fitted_values = model_k.predict(X)
  MSE = calculate_mse(Y, fitted_values, n, k)

  #get R^2
  R_squared = model_k.score(X,Y)
  adj_R_squared = 1 - ( (1-R_squared)*(n-1)/(n-(k+1)) )

  #returning the test RSS and test R^2
```

```
  return MSE, adj_R_squared



def find_best_subset(X, Y, up_to = 10):
  MSE_list, adj_R_squared_list, feature_list = [],[], []
  numb_features = []

  #Looping over k = 1 to k = 11 features in X
  for k in range(1, up_to + 1):
    #Looping over all possible combinations: from 11 choose k
    for combo in itertools.combinations(X.columns,k):
      tmp_result = fit_linear_reg(X[list(combo)],Y)   #Store temp result
      MSE_list.append(tmp_result[0])                  #Append lists
      adj_R_squared_list.append(tmp_result[1])

      feature_list.append(combo)
      numb_features.append(len(combo))

  #Store in DataFrame
  best_sub_features = pd.DataFrame({'numb_features': numb_features,
  'MSE': MSE_list,
  'Adj_R_squared':adj_R_squared_list,
  'features':feature_list})
  return best_sub_features
world_bank = sas7bdat::read.sas7bdat("project_data.sas7bdat", debug=FALSE)
#Drop the column Year and YearCode
world_bank = world_bank[!names(world_bank) %in% c("Year",
                                                  "YearCode",
                                                  "Country.Name",
                                                  "Country.Code")]

colnames(world_bank)[colnames(world_bank) ==
                       "Age.dependency.ratio....of.worki"] <- "Age.dependency.ratio.old"
colnames(world_bank)[colnames(world_bank) ==
                       "Age.dependency.ratio..young....o"] <- "Age.dependency.ratio.young"
colnames(world_bank)[colnames(world_bank) ==
                       "Mortality.rate..infant..per.1.00"] <- "Infant.Mortality"
colnames(world_bank)[colnames(world_bank) ==
                       "Mortality.rate..under.5..per.1.0"] <- "U5.Mortality"
colnames(world_bank)[colnames(world_bank) ==
                       "Mortality.rate..under.5..male..p"] <- "Male.Mortality"
colnames(world_bank)[colnames(world_bank) ==
```

```
                        "Mortality.rate..under.5..female"] <- "Female.Mortality"

world_bank["Age.dependency.ratio.old"] =
  world_bank["Age.dependency.ratio.old"] - world_bank["Age.dependency.ratio.young"]
head(world_bank, 5)
#remove case where Mortality rate, under-5 (per 1,000 live births) is Nan
world_bank = world_bank[!is.na(world_bank$U5.Mortality),]
rownames(world_bank) <- NULL

#remove columns with more than 5% missing
world_bank <- world_bank[,-which(colMeans(is.na(world_bank)) > 0.10)]
rownames(world_bank) <- NULL

#remove rows with more than 5 field missing
world_bank <- world_bank[-which(rowSums(is.na(world_bank)) > 5),]
rownames(world_bank) <- NULL
world_bank_imputed <- mice(data = world_bank,
                           m = 10,
                           method="rf",
                           seed = STAT5102,
                           printFlag = FALSE)
world_bank_imputed <- complete(world_bank_imputed)
qplot(world_bank_imputed$U5.Mortality,
      geom="histogram",
      bins = sqrt(nrow(world_bank_imputed)),
      xlab = "U5.Mortality", ylab = "Count")
cormat <- round(cor(world_bank_imputed),2)
melted_cormat <- reshape2::melt(cormat)
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  ggtitle("Correlation Heatmap")+
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.y = element_text(size=8))
OLS_model <- lm(data = world_bank_imputed, U5.Mortality ~ Infant.Mortality)
summary(OLS_model)$call
OLS_fitted_values <- predict(OLS_model, newdata = world_bank_imputed)
OLS_residuals <- resid(OLS_model)
regression_graph(r.world_bank_imputed["U5.Mortality"], r.OLS_fitted_values, r.OLS_residuals)
shapiro_test(OLS_residuals)
anderson_test(OLS_residuals)
```

```r
breusch_pagan_test(OLS_model)
#add gender difference
world_bank_imputed$Mortality.gender.diff <-
  abs(world_bank_imputed$Male.Mortality - world_bank_imputed$Female.Mortality)

#remove mortality rate for male and female
world_bank_imputed <-
  world_bank_imputed[,-which(names(world_bank_imputed) %in% c("Male.Mortality", "Female.Mortality"))]

#remove infant mortality rate
world_bank_imputed <-
  world_bank_imputed[,-which(names(world_bank_imputed) %in% c("Infant.Mortality"))]
#do the same thing for Life expectancy
world_bank_imputed["Life.exp.gender.diff"] <-
  abs(world_bank_imputed["Life.expectancy.at.birth..female"] -
        world_bank_imputed["Life.expectancy.at.birth..male.."])


#remove Life expectency for male and female
world_bank_imputed <-
  world_bank_imputed[,-which(names(world_bank_imputed) %in%
                               c("Life.expectancy.at.birth..female",
                                 "Life.expectancy.at.birth..male..",
                                 "Life.expectancy.at.birth..total"))]
Multiple_model <- lm(data = world_bank_imputed, U5.Mortality ~ .)
summary(Multiple_model)$call
Multiple_fitted_values <- predict(Multiple_model, newdata = world_bank_imputed)
Multiple_residuals <- resid(Multiple_model)
regression_graph(r.world_bank_imputed["U5.Mortality"],
r.Multiple_fitted_values,
r.Multiple_residuals)
shapiro_test(Multiple_residuals)
anderson_test(Multiple_residuals)
breusch_pagan_test(Multiple_model)
sort(car::vif(Multiple_model), decreasing = TRUE)[1:2]
bc <- MASS::boxcox(lm(world_bank_imputed$U5.Mortality~1), plotit = FALSE)
bcCI <- range(bc$x[bc$y > max(bc$y)-qchisq(0.95,1)/2])
print(paste0(
  "The 95% confidence interval for the lambda parameter that maximises the log-likelihood function is: ",
  round(bcCI[1], 4),
  ", ",round(bcCI[2], 4)))
Lasso_inTrain <- createDataPartition(world_bank_imputed$U5.Mortality, p=0.90, list=FALSE)
Lasso_validation_set <- world_bank_imputed[c(-Lasso_inTrain),]
```

```
Lasso_train_set <- world_bank_imputed[Lasso_inTrain,]
mod_cv <- hdi::lasso.cv(x =  as.matrix(dplyr::select(Lasso_train_set, -U5.Mortality)),
                        y = log(Lasso_train_set$U5.Mortality),
                        nfolds = 5)
lasso_terms <- names(Lasso_train_set)[mod_cv]
lasso_terms
Lasso_multiple_model <- lm(data = Lasso_train_set[c(lasso_terms, "U5.Mortality")],
                           log(U5.Mortality) ~ .)
summary(Lasso_multiple_model)$call
Lasso_multiple_fitted_values <- predict(Lasso_multiple_model, newdata = Lasso_train_set)
Lasso_multiple_residuals <- resid(Lasso_multiple_model)
Lasso_full_coef <- paste(round(coef(Lasso_multiple_model), 2))
Lasso_f_test <- anova(lm(data = Lasso_train_set, log(U5.Mortality)~1), Lasso_multiple_model)
Lasso_f_p <- round(Lasso_f_test$`Pr(>F)`[2], 2)
sort(car::vif(Lasso_multiple_model), decreasing = TRUE)[1:2]
regression_graph(r.Lasso_train_set["U5.Mortality"].apply(np.log),
r.Lasso_multiple_fitted_values,
r.Lasso_multiple_residuals)
shapiro_test(Lasso_multiple_residuals)
anderson_test(Lasso_multiple_residuals)
breusch_pagan_test(Lasso_multiple_model)
best_sub_inTrain <- createDataPartition(world_bank_imputed$U5.Mortality, p=0.90, list=FALSE)
best_sub_validation_set <- world_bank_imputed[c(-best_sub_inTrain),]
best_sub_train_set <- world_bank_imputed[best_sub_inTrain,]
#Initialization variables

X = r.best_sub_train_set.drop(["U5.Mortality"], axis = 1)
Y = r.best_sub_train_set["U5.Mortality"].apply(np.log)

#best_sub_features = find_best_subset(X, Y)
best_sub_features = pd.read_pickle('best_sub_features.pkl')
min_MSE = (best_sub_features[best_sub_features.groupby('numb_features')['MSE'].transform(min) ==
best_sub_features['MSE']])
max_adj_R2 = (best_sub_features[best_sub_features.groupby('numb_features')['Adj_R_squared'].transform(max)
best_sub_features['Adj_R_squared']])
#min_MSE <- dplyr::select(py$min_MSE, -Adj_R_squared)
#dplyr::arrange(min_MSE, by_group = MSE)
#Adding columns to the dataframe with MSE and R squared values of the best subset
best_sub_features['min_MSE'] = best_sub_features.groupby('numb_features')['MSE'].transform(min)
best_sub_features['max_R_squared'] = best_sub_features.groupby('numb_features')['Adj_R_squared'].transform

fig = plt.figure(figsize = (16,6))
```

```
ax = fig.add_subplot(1, 2, 1)
ax.scatter(best_sub_features.numb_features,best_sub_features.MSE, alpha = .2,
color = 'darkblue')
ax.set_xlabel('# Predictor')
ax.set_ylabel('MSE')
ax.set_title('MSE - Best subset selection')
ax.plot(best_sub_features.numb_features,
best_sub_features.min_MSE,color = 'r',
label = 'Best subset')
ax.legend()

ax = fig.add_subplot(1, 2, 2)
ax.scatter(best_sub_features.numb_features,best_sub_features.Adj_R_squared, alpha = .2, color = 'darkblue'
ax.plot(best_sub_features.numb_features,best_sub_features.max_R_squared,color = 'r', label = 'Best subset'
ax.set_xlabel('# Predictor')
ax.set_ylabel('Adjusted R squared')
ax.set_title('Adjusted R_squared - Best subset selection')
ax.legend()

plt.show()
best_subset_features = min_MSE[min_MSE["numb_features"] == 6]["features"].tolist()
best_subset_features = list(best_subset_features[0])
print(best_subset_features[1:3])
print(best_subset_features[3:6])

X = r.best_sub_train_set[best_subset_features]
Y = r.best_sub_train_set["U5.Mortality"]

response_vs_predictor_plot(X, Y, "U5 Mortality v.s. Predictor", order = 2)
process_best_subset_data <- function(data, best_subset_features){
  #Pick best subset variables
  log_best_subset = data[c(best_subset_features, "U5.Mortality")]

  #log Y
  log_best_subset$U5.Mortality <- log(log_best_subset$U5.Mortality)

  #log some variables
  log_best_subset[,"Age.dependency.ratio.old"] <- log(log_best_subset[,"Age.dependency.ratio.old"])
  log_best_subset[,"Mortality.gender.diff"] <- log(log_best_subset[,"Mortality.gender.diff"])
  log_best_subset[,"GDP.per.capita..PPP..constant.20"] <-
    log(log_best_subset[,"GDP.per.capita..PPP..constant.20"])

  #there is a outliner in Age.dependency.ratio.old
```

```
  log_best_subset <- log_best_subset[log_best_subset$Age.dependency.ratio.old >= 0,]
  #remove outliners
  # log_best_subset <- apply(log_best_subset, 2, remove_outliers)
  # log_best_subset <- log_best_subset[complete.cases(log_best_subset),]
  # log_best_subset <- as.data.frame(log_best_subset)
}

log_best_subset_validate <- process_best_subset_data(best_sub_validation_set,
                                                     py$best_subset_features)
log_best_subset <- process_best_subset_data(best_sub_train_set, py$best_subset_features)
X = r.log_best_subset.drop("U5.Mortality", axis = 1)
Y = r.log_best_subset["U5.Mortality"]

response_vs_predictor_plot(X, Y, "U5 Mortality v.s. Predictor (transformed)")
Best_sub_model <- lm(data = log_best_subset, U5.Mortality~.)
summary(Best_sub_model)$call
Best_sub_fitted_values <- predict(Best_sub_model, newdata = log_best_subset)
Best_sub_residuals <- resid(Best_sub_model)
Best_sub_full_coef <- paste(round(coef(Best_sub_model), 2))
Best_sub_f_test <- anova(lm(data = log_best_subset, U5.Mortality~1), Best_sub_model)
Best_sub_f_p <- round(Best_sub_f_test$`Pr(>F)`[2], 2)
sort(car::vif(Best_sub_model), decreasing = TRUE)[1:2]
regression_graph(r.log_best_subset["U5.Mortality"], r.Best_sub_fitted_values, r.Best_sub_residuals)
shapiro_test(Best_sub_residuals)
anderson_test(Best_sub_residuals)
breusch_pagan_test(Best_sub_model)
Best_Subset_Validation_Result <-
  postResample(pred = predict.lm(Best_sub_model,
                                 log_best_subset_validate),
               obs = log_best_subset_validate$U5.Mortality)
Lasso_Validation_Result <-
  postResample(pred = predict.lm(Lasso_multiple_model, Lasso_validation_set),
               obs = log(Lasso_validation_set$U5.Mortality))
papaja::r_refs(file = "project-references.bib")
```