Annexure 1: COVID-19 early days regression analysis

#### Lindo Khoza

# Background

This exploratory data analysis notebook continues from the [Novel Coronavirus (COVID-19) Cases from a South African perspective](https://github.com/willkhoza/COVID_SA) repository. The purpose of this analysis is to understand the influence of demographic settings on the early progression of the COVID-19 within different countries.

## Data Dictionary

Albeit the demographic variables being intuitively named, a detailed data dictionary is provided in the [worldometer](https://github.com/willkhoza/COVID_SA/tree/master/worldometer) folder in the repository.

The following variables are defined:

: The amount of days it takes for a country to report it’s **first confirmed case** with a base date of 17 November 2019. This base date has been unofficiially reported to be the earliest detected case of the disease.

: The amount of days it takes for a country to report it’s **first death** which relates COVID-19.

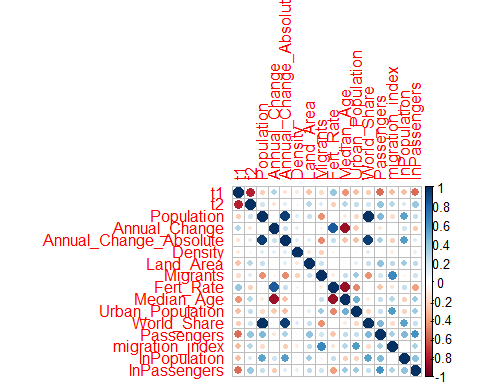
## Data Summary

# reading data into enviroment  
library(tidyverse)  
dat <- read\_csv("https://raw.githubusercontent.com/willkhoza/COVID\_SA/master/machinery/timeline.csv")  
dat <- dat %>%  
 mutate(Median\_Age = as.numeric(Median\_Age),   
 Urban\_Population = as.numeric(Urban\_Population),  
 migration\_index = sign(Migrants)\*log(abs(Migrants)),  
 country = as.factor(country),  
 lnPopulation = log(Population),  
 lnPassengers = log(Passengers))  
# remove countries with confirmed cases outside of the investigation period  
dat <- dat %>%  
 filter(t1 > 66)  
   
dat %>%  
 summary()

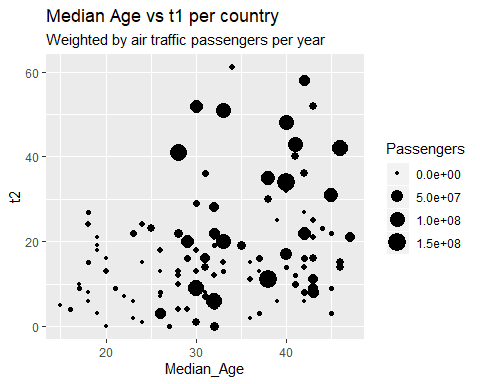
## country t1 t2 Population   
## Afghanistan : 1 Min. : 67.0 Min. : 0.00 Min. :8.010e+02   
## Albania : 1 1st Qu.:102.0 1st Qu.: 8.00 1st Qu.:2.586e+06   
## Algeria : 1 Median :110.0 Median :15.00 Median :9.228e+06   
## Andorra : 1 Mean :108.4 Mean :17.31 Mean :3.309e+07   
## Angola : 1 3rd Qu.:119.0 3rd Qu.:22.00 3rd Qu.:2.896e+07   
## Antigua and Barbuda: 1 Max. :135.0 Max. :61.00 Max. :1.380e+09   
## (Other) :168 NA's :53 NA's :4   
## Annual\_Change Annual\_Change\_Absolute Density Land\_Area   
## Min. :-1.3500 Min. : -259876 Min. : 2.0 Min. : 0   
## 1st Qu.: 0.3275 1st Qu.: 2316 1st Qu.: 32.5 1st Qu.: 2028   
## Median : 1.0750 Median : 53139 Median : 83.5 Median : 10680   
## Mean : 1.1961 Mean : 400224 Mean : 367.3 Mean : 492103   
## 3rd Qu.: 1.9300 3rd Qu.: 350403 3rd Qu.: 194.0 3rd Qu.: 62719   
## Max. : 3.8400 Max. :13586631 Max. :26337.0 Max. :16376870   
## NA's :4 NA's :4 NA's :4 NA's :4   
## Migrants Fert\_Rate Median\_Age Urban\_Population  
## Min. :-653249.0 Min. :1.200 Min. :15.00 Min. :13.00   
## 1st Qu.: -826.0 1st Qu.:1.700 1st Qu.:22.50 1st Qu.:43.00   
## Median : -4.0 Median :2.200 Median :30.00 Median :60.00   
## Mean : -3967.2 Mean :2.702 Mean :30.36 Mean :59.84   
## 3rd Qu.: 88.5 3rd Qu.:3.450 3rd Qu.:38.00 3rd Qu.:78.00   
## Max. : 543822.0 Max. :7.000 Max. :47.00 Max. :98.00   
## NA's :11 NA's :11 NA's :11 NA's :9   
## World\_Share Passengers migration\_index lnPopulation   
## Min. : 0.0000 Min. : 0 Min. :-13.3897 Min. : 6.686   
## 1st Qu.: 0.0300 1st Qu.: 49385 1st Qu.: -6.9748 1st Qu.:14.765   
## Median : 0.1200 Median : 1840720 Median : -1.6094 Median :16.037   
## Mean : 0.4245 Mean : 14140375 Mean : -0.7228 Mean :15.692   
## 3rd Qu.: 0.3675 3rd Qu.: 9654728 3rd Qu.: 4.8712 3rd Qu.:17.181   
## Max. :17.7000 Max. :153537550 Max. : 13.2064 Max. :21.045   
## NA's :4 NA's :10 NA's :15 NA's :4   
## lnPassengers   
## Min. : -Inf   
## 1st Qu.:10.81   
## Median :14.43   
## Mean : -Inf   
## 3rd Qu.:16.08   
## Max. :18.85   
## NA's :10

## Correlogram and Regression Plots

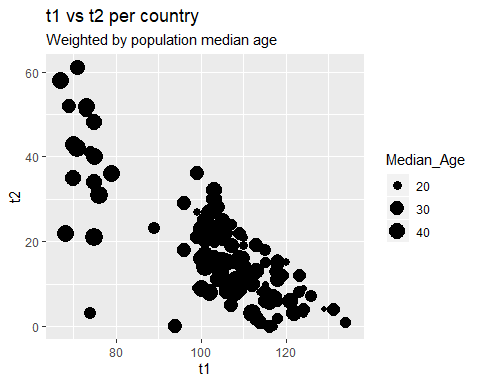
# exploratory data analysis  
library(Hmisc)  
library(corrplot)  
  
dat.corr <- dat %>%  
 na.omit() %>%  
 select(t1,  
 t2,  
 Population,   
 Annual\_Change,   
 Annual\_Change\_Absolute,  
 Density,  
 Land\_Area,  
 Migrants,  
 Fert\_Rate,  
 Median\_Age,  
 Urban\_Population,  
 World\_Share,  
 Passengers,  
 migration\_index,  
 lnPopulation,  
 lnPassengers) %>%  
 as.matrix() %>%  
 rcorr()  
  
dat.corr$r %>%  
 corrplot()



dat %>%  
 ggplot()+  
 aes(x = Median\_Age, y = t2, size = Passengers)+  
 geom\_point() +  
 ggtitle("Median Age vs t1 per country", "Weighted by air traffic passengers per year")



dat %>%  
 ggplot()+  
 aes(x = t1, y = t2, size = Median\_Age)+  
 geom\_point() +  
 ggtitle("t1 vs t2 per country", "Weighted by population median age")



## : Regression Analysis

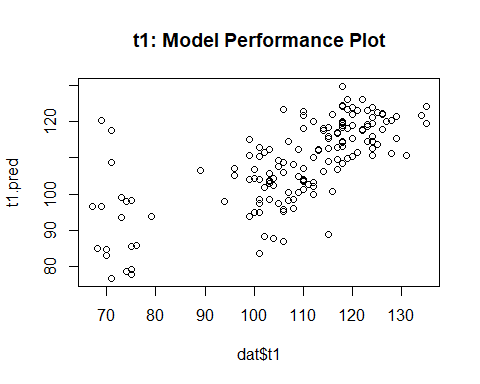
# fit regression models  
library(MASS, pos = .Machine$integer.max)  
library(ggpubr)  
  
# saturated linear model  
t1.mod0 <- lm(t1~ Population +  
 Annual\_Change +  
 Annual\_Change\_Absolute +  
 Density +  
 Land\_Area +  
 Migrants +  
 Fert\_Rate +  
 Median\_Age +  
 Urban\_Population +  
 World\_Share +  
 Passengers +  
 migration\_index +  
 lnPopulation +  
 lnPassengers,  
 data = na.omit(dat))  
# Stepwise regression model  
t1.mod1 <- stepAIC(t1.mod0, direction = "both",   
 trace = FALSE)  
  
summary(t1.mod0)

##   
## Call:  
## lm(formula = t1 ~ Population + Annual\_Change + Annual\_Change\_Absolute +   
## Density + Land\_Area + Migrants + Fert\_Rate + Median\_Age +   
## Urban\_Population + World\_Share + Passengers + migration\_index +   
## lnPopulation + lnPassengers, data = na.omit(dat))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -37.719 -5.094 0.303 7.689 26.109   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.898e+02 2.387e+01 7.954 3.98e-12 \*\*\*  
## Population 7.482e-06 5.300e-06 1.412 0.16136   
## Annual\_Change -6.823e-01 2.804e+00 -0.243 0.80831   
## Annual\_Change\_Absolute -2.015e-06 2.908e-06 -0.693 0.49013   
## Density -7.511e-03 4.963e-03 -1.513 0.13357   
## Land\_Area -5.299e-07 6.265e-07 -0.846 0.39979   
## Migrants -5.181e-06 1.583e-05 -0.327 0.74415   
## Fert\_Rate -1.800e+00 2.387e+00 -0.754 0.45282   
## Median\_Age -9.522e-01 3.412e-01 -2.791 0.00637 \*\*   
## Urban\_Population -3.821e-02 8.501e-02 -0.450 0.65409   
## World\_Share -5.822e+02 4.128e+02 -1.410 0.16180   
## Passengers -1.063e-07 5.000e-08 -2.126 0.03616 \*   
## migration\_index -2.780e-02 2.468e-01 -0.113 0.91055   
## lnPopulation -1.909e+00 1.126e+00 -1.695 0.09333 .   
## lnPassengers -7.827e-01 5.177e-01 -1.512 0.13391   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.47 on 94 degrees of freedom  
## Multiple R-squared: 0.5247, Adjusted R-squared: 0.4539   
## F-statistic: 7.412 on 14 and 94 DF, p-value: 3.541e-10

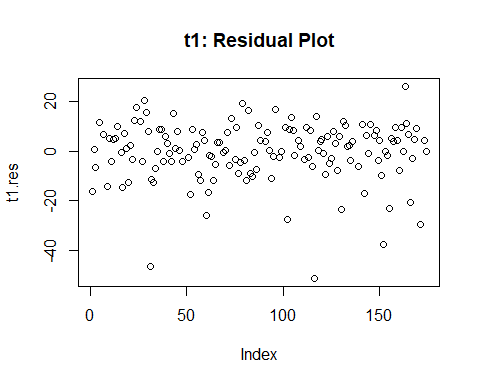
summary(t1.mod1)

##   
## Call:  
## lm(formula = t1 ~ Population + Density + Fert\_Rate + Median\_Age +   
## World\_Share + Passengers + lnPopulation + lnPassengers, data = na.omit(dat))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -34.237 -5.379 0.316 7.258 26.133   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.912e+02 1.952e+01 9.799 2.74e-16 \*\*\*  
## Population 7.391e-06 4.993e-06 1.480 0.141926   
## Density -7.283e-03 4.354e-03 -1.673 0.097510 .   
## Fert\_Rate -2.784e+00 1.883e+00 -1.478 0.142445   
## Median\_Age -9.569e-01 2.533e-01 -3.777 0.000269 \*\*\*  
## World\_Share -5.764e+02 3.893e+02 -1.481 0.141855   
## Passengers -1.271e-07 4.331e-08 -2.934 0.004148 \*\*   
## lnPopulation -1.909e+00 1.040e+00 -1.835 0.069469 .   
## lnPassengers -9.436e-01 4.841e-01 -1.949 0.054058 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.25 on 100 degrees of freedom  
## Multiple R-squared: 0.5136, Adjusted R-squared: 0.4747   
## F-statistic: 13.2 on 8 and 100 DF, p-value: 7.479e-13

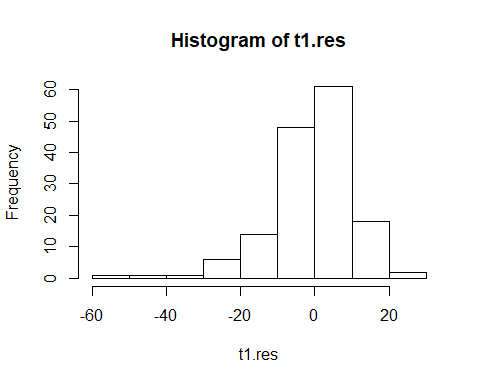
#Model Evaluation  
  
t1.pred = predict(t1.mod0, dat)  
plot(dat$t1, t1.pred, main = "t1: Model Performance Plot")



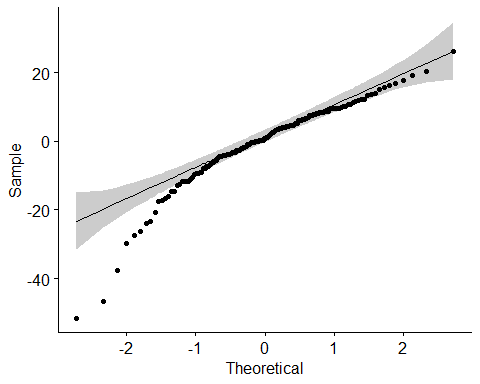
t1.res <- dat$t1-t1.pred  
plot(t1.res, main = "t1: Residual Plot")



hist(t1.res)



# test for normality  
ggqqplot(t1.res)



summary(t1.res)

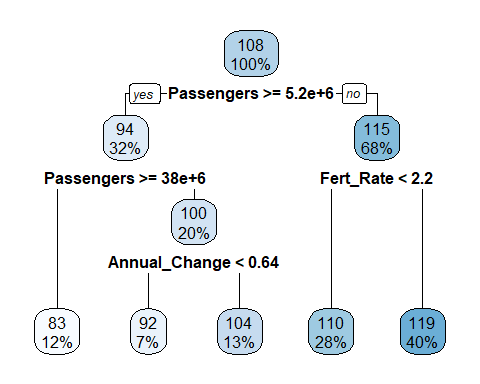
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## -51.4644 -4.6742 0.7389 -0.3796 7.5906 26.1091 22

var(t1.res, na.rm = TRUE)

## [1] 136.2532

## : Decision Tree Analysis

# regression tree to suppliment the t1 regression model  
require(rpart, warn.conflicts = FALSE)  
require(rpart.plot)  
  
t1.mod2 <- rpart(t1~ Population +  
 Annual\_Change +  
 Annual\_Change\_Absolute +  
 Density +  
 Land\_Area +  
 Migrants +  
 Fert\_Rate +  
 Median\_Age +  
 Urban\_Population +  
 World\_Share +  
 Passengers +  
 migration\_index +  
 lnPopulation,   
 data = dat,  
 control = rpart.control(cp = 0.025))  
  
rpart.plot(t1.mod2)



## : Regression Analysis

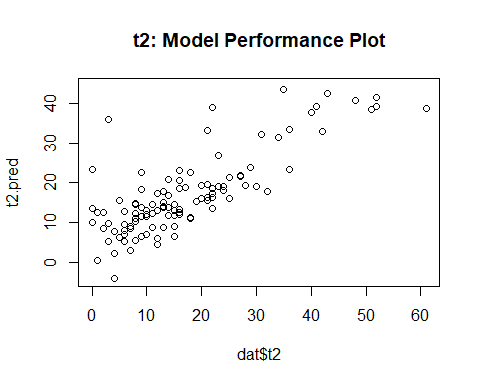
# regression model for t2  
# saturated linear model  
t2.mod0 <- lm(t2~ t1+   
 Annual\_Change +  
 Annual\_Change\_Absolute +  
 Density +  
 Land\_Area +  
 Migrants +  
 Fert\_Rate +  
 Median\_Age +  
 Urban\_Population +  
 World\_Share +  
 Passengers +  
 migration\_index +  
 lnPopulation+  
 lnPassengers,   
 data = na.omit(dat))  
# Stepwise regression model  
t2.mod1 <- stepAIC(t2.mod0, direction = "both",   
 trace = FALSE)  
  
summary(t2.mod0)

##   
## Call:  
## lm(formula = t2 ~ t1 + Annual\_Change + Annual\_Change\_Absolute +   
## Density + Land\_Area + Migrants + Fert\_Rate + Median\_Age +   
## Urban\_Population + World\_Share + Passengers + migration\_index +   
## lnPopulation + lnPassengers, data = na.omit(dat))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.967 -3.603 0.579 4.620 22.228   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.236e+02 2.170e+01 5.695 1.4e-07 \*\*\*  
## t1 -7.260e-01 7.245e-02 -10.021 < 2e-16 \*\*\*  
## Annual\_Change 1.011e-01 1.978e+00 0.051 0.9594   
## Annual\_Change\_Absolute 5.184e-07 2.042e-06 0.254 0.8001   
## Density -2.485e-03 3.530e-03 -0.704 0.4832   
## Land\_Area 5.078e-07 4.439e-07 1.144 0.2556   
## Migrants -1.908e-06 1.123e-05 -0.170 0.8654   
## Fert\_Rate -2.768e-01 1.679e+00 -0.165 0.8694   
## Median\_Age -2.152e-01 2.513e-01 -0.856 0.3940   
## Urban\_Population 9.095e-03 6.019e-02 0.151 0.8802   
## World\_Share 9.167e-02 1.779e+00 0.052 0.9590   
## Passengers -2.517e-08 3.609e-08 -0.698 0.4872   
## migration\_index 7.132e-02 1.735e-01 0.411 0.6820   
## lnPopulation -1.376e+00 8.113e-01 -1.696 0.0932 .   
## lnPassengers -4.924e-02 3.709e-01 -0.133 0.8947   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.144 on 94 degrees of freedom  
## Multiple R-squared: 0.6393, Adjusted R-squared: 0.5855   
## F-statistic: 11.9 on 14 and 94 DF, p-value: 2.594e-15

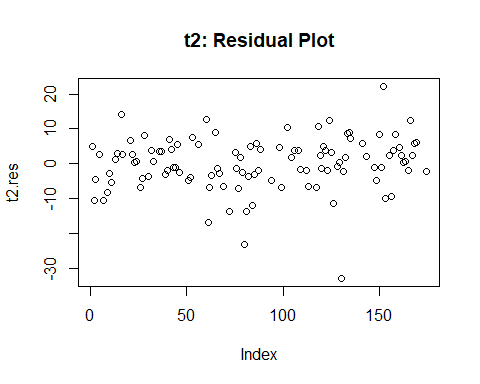
summary(t2.mod1)

##   
## Call:  
## lm(formula = t2 ~ t1 + Median\_Age + lnPopulation, data = na.omit(dat))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -35.287 -4.217 0.135 4.534 20.379   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 118.61624 16.31407 7.271 6.65e-11 \*\*\*  
## t1 -0.72402 0.06224 -11.632 < 2e-16 \*\*\*  
## Median\_Age -0.21348 0.10723 -1.991 0.0491 \*   
## lnPopulation -1.14529 0.58849 -1.946 0.0543 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.845 on 105 degrees of freedom  
## Multiple R-squared: 0.626, Adjusted R-squared: 0.6154   
## F-statistic: 58.59 on 3 and 105 DF, p-value: < 2.2e-16

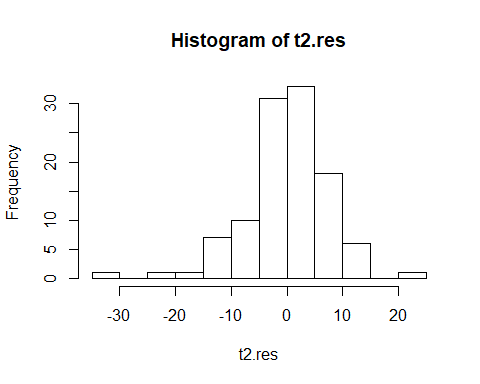
#Model Evaluation  
  
t2.pred = predict(t2.mod0, dat)  
plot(dat$t2, t2.pred, main = "t2: Model Performance Plot")



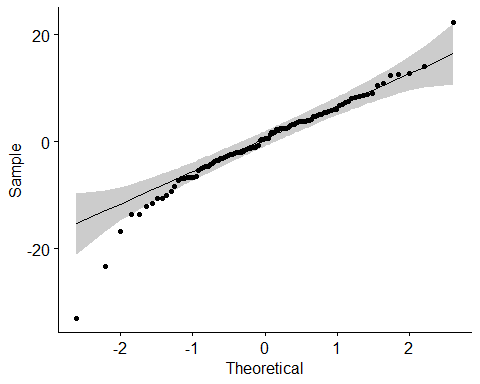
t2.res <- dat$t2-t2.pred  
plot(t2.res, main = "t2: Residual Plot")



hist(t2.res)



# test for normality  
ggqqplot(t2.res)



summary(t2.res)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## -32.9670 -3.6034 0.5791 0.0000 4.6197 22.2282 65

var(t2.res, na.rm = TRUE)

## [1] 57.72255