Title: Analyzing Overall Access to at Least Basic Water Source Using Time and Region through Additive and Interaction Models.

Abstract:

This is an observational study of the percentage of population using basic water service in individual countries over time from 2000 to 2017, given the data collected by the United Nations. To see if there is some progress made by the UN and the countries over time, I want to test the overall model using predictors region and time and then follow-up tests for individual predictors. I also want to test for interaction between the three variables since for regions where most third world countries are located, there are less overall access and in need for more progress. Through multiple regression and Anova tables, associations exist for individual predictors and overall access to basic water service (%), and interaction also exists. We can see that the UN is focusing on countries such as in South Asia and Sub-Saharan Africa's overall access more as time increases from the interaction.

Background and Significance:

In a first-world country, it may be hard to experience unexpected water shortages in one's own house. But according to the United Nations Sustainable Development Goals in 2017, "two billion people live in countries experiencing high water stress, and about 4 billion people experience severe water scarcity at least one month a year"; 785 million people also remain without even basic drinking water services. Basic water service, defined by the World Health Organization, is drinking water, given the collection time is 30 minutes or less, from an improved source. Improved water sources according to Gapminder include "piped water, boreholes or tube wells, protected dug wells and springs, and packaged of delivered water."

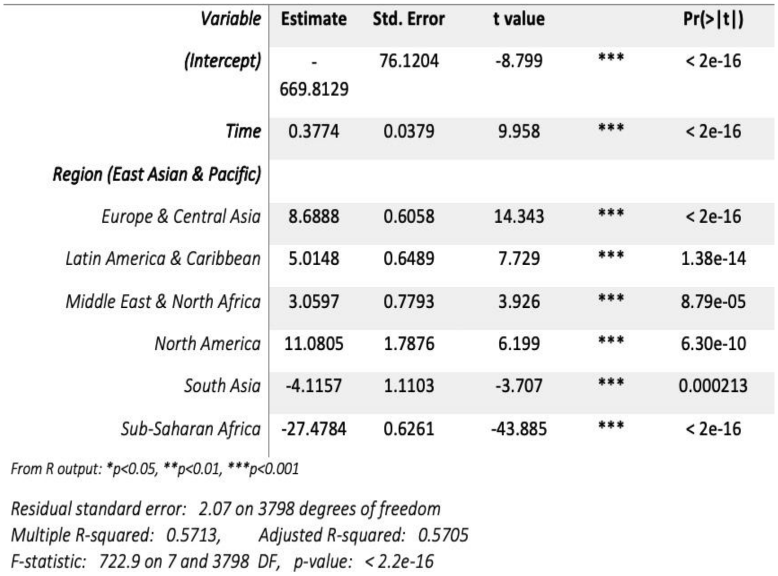
As the UNESCO World Water Assessment Programme claimed in 2019, "'Leaving no one behind' is at the heart of the commitment of the 2030 Agenda," and the Programme aimed to "achieve full realization of human rights" that included universal access to water. To see whether or not the access to basic water service increased over the years through actions taken by the United Nations and individual countries, I selected data on the percentage of population in individual countries using at least basic drinking water services. I want to test if there is an association or interaction between regions, time and the percentage of population with access to basic drinking water service.

Methods:

a. *Data Collection*: The data of "People using at least basic drinking water services (% of population)" from 2000 to 2017 is collected by WHO/UNICEF JMP. It is compiled by "country missions and regional workshops", "survey questions", and "nationally-led monitoring." The data includes 217 individual countries with defined regions and their corresponding percentages or NA for each year (18 years total). After reorganizing the data for R, it is made up of 3906 elements (rows) from 18 repeated 217 countries. None-response rate out of the 3906 responses is 2.56%.
b. *Variable Creation*: Variables include countries, region, SH.H2O.BASW.ZS, and time. In the reorganized data, countries variable is factor with 217 levels of individual countries; region, a factor with 7 levels: East Asia & Pacific, Europe & Central Asia, Latin America & Caribbean, Middle East & North Africa, North America, South Asia, and Sub-Saharan Africa; SH.H2O.BASW.ZS (response variable), a numerical percentages of population using at least basic water service, encompassing both people using basic water services as well as those using safely managed water service in the population; time, integer values spanning from 2000 to 2017.
c. *Analytic Methods*: Use Multiple Regression with an F-test to test the statistical significant of the overall model using Region and Time; if the overall model is useful (meaning there is at least one useful predictor), there will be a follow-up t-test or Anova table on association between SH.H2O.BASW.ZS and individual predictors. Whether or not there is an interaction between Region, Time, and SH.H2O.BASW.ZS is tested through the Anova table. Anova table is used because this is an observational study - there exists shared variations between explanatory variables.

Results:

Table 1 Multiple Regression for Percentage of People with Basic Water Services (Overall Access)



Variable	Estimate	Std. Error	t value		Pr(>|t|)
(Intercept)	-669.8129	76.1204	-8.799	***	< 2e-16
Time	0.3774	0.0379	9.958	***	< 2e-16
Region (East Asian & Pacific)					
Europe & Central Asia	8.6888	0.6058	14.343	***	< 2e-16
Latin America & Caribbean	5.0148	0.6489	7.729	***	1.38e-14
Middle East & North Africa	3.0597	0.7793	3.926	***	8.79e-05
North America	11.0805	1.7876	6.199	***	6.30e-10
South Asia	-4.1157	1.1103	-3.707	***	0.000213
Sub-Saharan Africa	-27.4784	0.6261	-43.885	***	< 2e-16

From R output: *p<0.05, **p<0.01, ***p<0.001

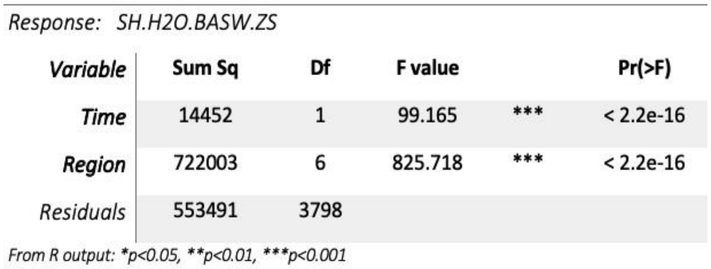
Residual standard error: 2.07 on 3798 degrees of freedom
Multiple R-squared: 0.5713, Adjusted R-squared: 0.5705
F-statistic: 722.9 on 7 and 3798 DF, p-value: < 2.2e-16

Table 1 shows two predictors: time (quantitative) and region (categorical with 7 categories and 6 indicators; all 0s is East Asian & Pacific.

57.13% of the variation in basic water service access (%) is explained by the additive model.

The F-statistic for the overall model is 722.9 with 7 and 3798 df, and the probability of getting that F-statistic or more extreme by random chance alone is very low, considering the p-value is $2.2 * 10^{-16}$ So there is strong evidence or statistical significance of at least one predictor (region and time) is useful. Following-up t-test for time gives a t-value of 9.958 and p-value of less than $2.2 * 10^{-16}$; there is strong evidence that time is associated with overall basic water service access (%). Given the intercept for time in Table 1, as time increases by 1 year, predicted basic water service access for a country increases by 0.3774%, holding region constant. The following-up test for region in Table 1, however, compares the six regions to East Asian & Pacific (baseline group), and shows the individual six regions are all statistically different from the baseline region.

Table 2 Anova Table (Type II tests) for Additive Model of Time and Region

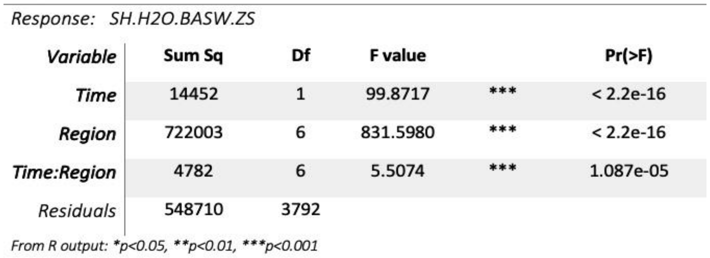


Response: SH.H2O.BASW.ZS

Variable	Sum Sq	Df	F value		Pr(>F)
Time	14452	1	99.165	***	< 2.2e-16
Region	722003	6	825.718	***	< 2.2e-16
Residuals	553491	3798			

From R output: *p<0.05, **p<0.01, ***p<0.001

Table 3 Anova Table (Type II tests) for Model of Time and Region with Interaction



Response: SH.H2O.BASW.ZS

Variable	Sum Sq	Df	F value		Pr(>F)
Time	14452	1	99.8717	***	< 2.2e-16
Region	722003	6	831.5980	***	< 2.2e-16
Time:Region	4782	6	5.5074	***	1.087e-05
Residuals	548710	3792			

From R output: *p<0.05, **p<0.01, ***p<0.001

To test the usefulness of the predictor region, the Anova table gives the unique contribution of time and region in the sum of squares. For variable region F-value is 825.718, and the p-value is less than $2.2 * 10^{-16}$; therefore, there is strong evidence of an association between region and the overall access (%) adjusting for time.

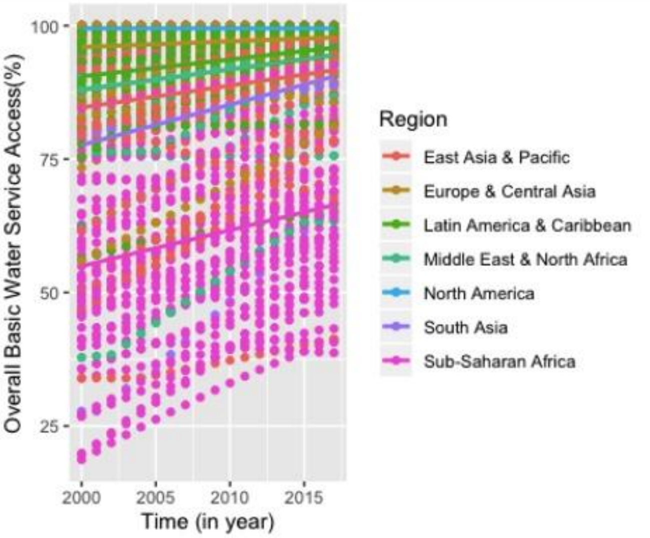
1.12% (SSTime/SSTotal) of variation in overall access (%) is explained by time; 55.97% (SSRegion/SSTotal) of variation is explained by region.

From Table 3, the interaction model is tested by Anova Table. The F-value for the interaction (Time:Region) is 5.5074, and the p-value is 1.087*10^(-5); since the p-value is very small, there is strong evidence of an interaction between time, region, and overall water access(%).

Discussions/Conclusions:

My objective was to see if there is an association between each predictor (time and region) and the overall access to basic water service %, and if there is an interaction between the three variables. The results from multiple regression using R support of my hypotheses of there exist an association and able to conclude by the very small p-values for the overall F-statistics 722.9 (Table 1) and individual t-statistic 9.958 (Table 1) for time and F-value from Anova 825.718 (Table 2) for region. I also wanted to find out if there is an interaction since in some regions there may be more focus from the UN to increase basic water service access. Since after testing using Anova there is a very small p-value, there is strong evidence of an interaction; and from Graph 1, we can see regardless of region, as time increases, countries without 100% basic water service tend to increase in access, but for countries in region Sub-Saharan Africa and South Asia, the change in percentage is larger on average, which shows that the UN had taken action to improve the percent of population that access basic water service in countries specially with low percentages.

Graph 1 Interaction Model Between Time, Region, and Overall Basic Water Service Access (%)



The limitation of this project is that since it is a time series, independence and stationarity is affected - although each country is independent of another country - because each country has multiple data, and numerical data increases as time increases. Autocorrelation may occur, which means residuals can predict the next residuals and gives explanation to variation that predictors do not explain. From Residual vs Fitted graph, linear condition is concerned by the curve. And there is non constant variance as fitted values increase - more negative residuals showing more overestimating actual percentage. The residuals seem normally distributed but a little bit skewed left, however the sample size is fairly large.

In conclusion, there is strong evidence of a positive association - but weak shown from slope 0.3774 and small R^2 0.0112 due to the large sample size - between predicted basic water service access (%) in each country and time; although regions do have an effect from the large R^2 (0.5597) on the countries' access, from the interaction graph, we can see regions with less access tend to increase greater in access - compared to those with higher overall access - on average as time increases. This shows that some progress has been achieved by the UN and individual countries from 2000 - 2017, although weak, we hope to continue expanding the effort to meet universal access to water in 2030.

References:

Gapminder. (n.d.). At least basic water source, overall access (%). Retrieved from https://www.gapminder.org/data/

JMP. (n.d.). Country and regional engagement. Retrieved from https://washdata.org/how-we-work/country-and-regional-engagement

UNESCO World Water Assessment Programme. (2019). The United Nations world water development report 2019: leaving no one behind. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000367306

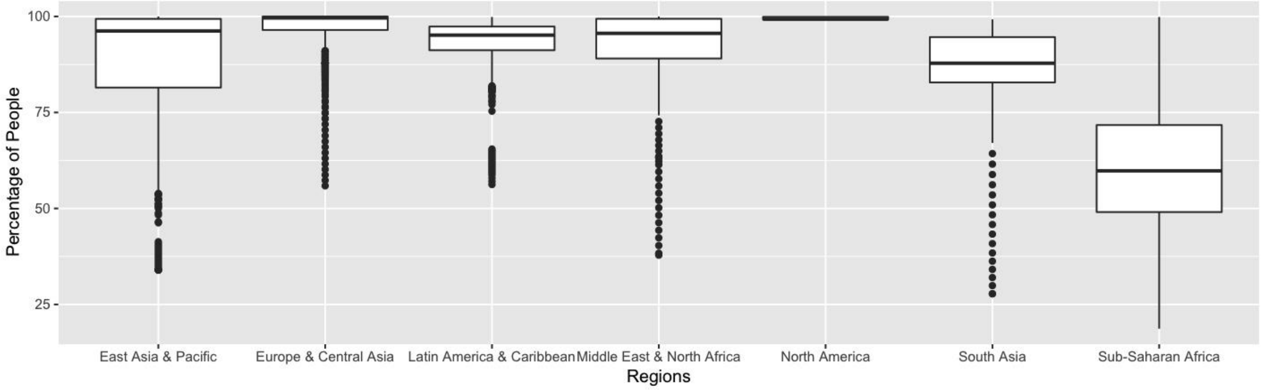
United Nations Statistical Division. (2019). Sustainable development goals overview. Retrieved from https://unstats.un.org/sdgs/report/2019/Overview/

WHO/UNICEF Joint Monitoring Programme (JMP) for Water Supply, Sanitation and Hygiene. (2019). People using at least basic drinking water services (% of population) using at least basic drinking water services (% of population). *The World Bank.* Retrieved from https://data.worldbank.org/indicator/SH.H2O.BASW.ZS

World Health Organization. (n.d.). Monitoring drinking-water. Retrieved from https://www.who.int/water_sanitation_health/monitoring/coverage/monitoring-dwater/en/

Appendix:

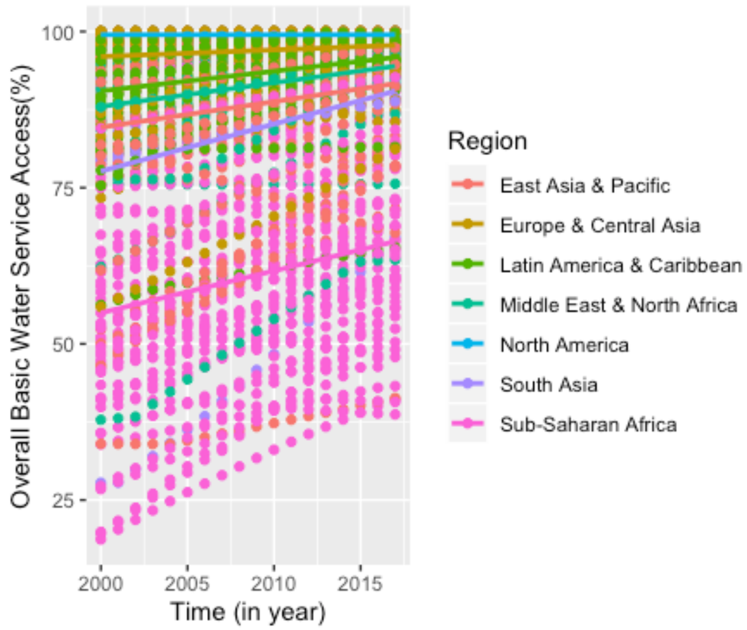
```
##STAT 4110H Project R code
BWA = read.csv(file= "OverallBasicWaterAccessinPercentage.csv",row.names = NULL)
dim(BWA)
[1] 3906    5
BWA2 = na.omit(BWA)
dim(BWA2)
[1] 3806    5

#descriptive statistic
tapply(BWA2$SH.H2O.BASW.ZS,BWA2$Region,mean)
 East Asia & Pacific          Europe & Central Asia
            96.22599                  99.62494
 Latin America & Caribbean  Middle East & North Africa
            95.14522                  95.61297
          North America              South Asia
            99.29436                  87.83213
        Sub-Saharan Africa
            59.79475
ggplot(data = BWA2, aes(x = Region, y = SH.H2O.BASW.ZS)) +
  geom_boxplot()+
  xlab("Percentage of Population") +
  ylab("Region")
```



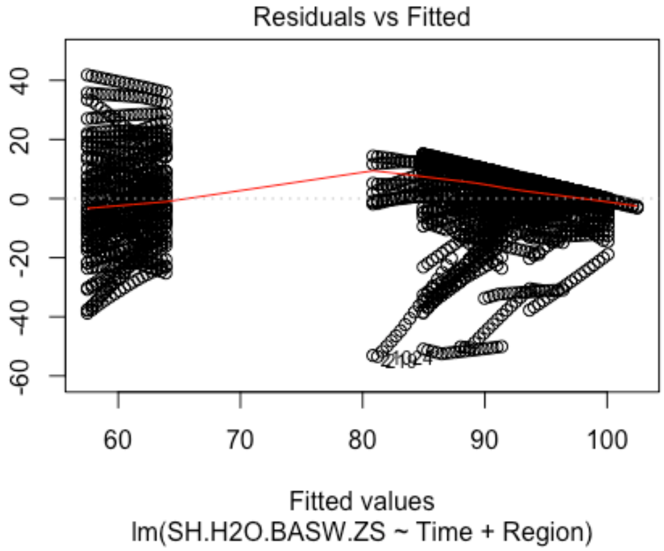
```
#inferential statistic
model1 = lm(data=BWA2,SH.H2O.BASW.ZS ~ Time + Region) #additive model
summary(model1)
Call:
lm(formula = SH.H2O.BASW.ZS ~ Time + Region, data = BWA2)

Residuals:
    Min      1Q  Median      3Q     Max
-53.422  -3.550   1.663   5.890  41.800

Coefficients:
```

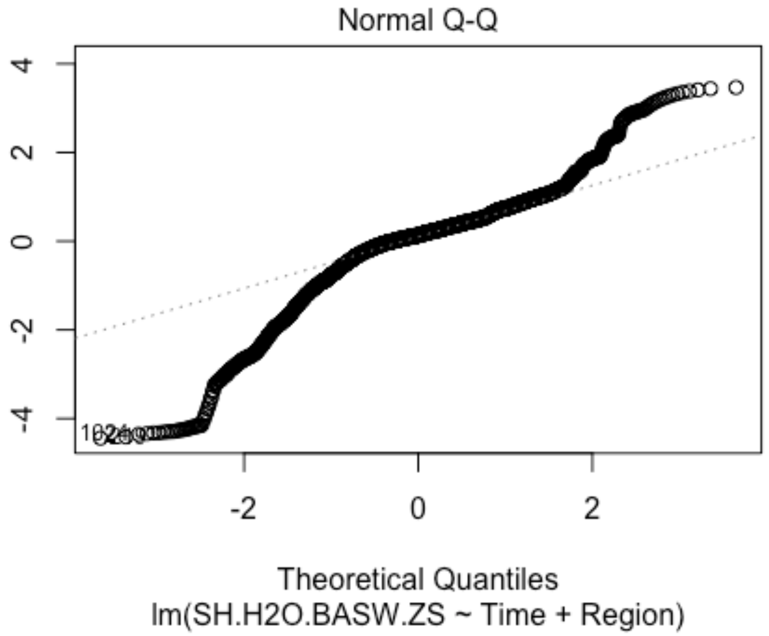
```
                                          Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)                               -669.8129    76.1204     -8.799    < 2e-16   ***
Time                                         0.3774     0.0379      9.958    < 2e-16   ***
RegionEurope & Central Asia                  8.6888     0.6058     14.343    < 2e-16   ***
RegionLatin America & Caribbean              5.0148     0.6489      7.729   1.38e-14   ***
RegionMiddle East & North Africa             3.0597     0.7793      3.926   8.79e-05   ***
RegionNorth America                         11.0805     1.7876      6.199   6.30e-10   ***
RegionSouth Asia                            -4.1157     1.1103     -3.707   0.000213   ***
RegionSub-Saharan Africa                   -27.4784     0.6261    -43.885   < 2e-16    ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.07 on 3798 degrees of freedom
Multiple R-squared:  0.5713,  Adjusted R-squared:  0.5705
F-statistic: 722.9 on 7 and 3798 DF,  p-value: < 2.2e-16
Anova(model1)
Anova Table (Type II tests)

Response: SH.H2O.BASW.ZS
             Sum Sq   Df      F value     Pr(>F)
Time         14452     1       99.165    < 2.2e-16 ***
Region       722003    6      825.718    < 2.2e-16 ***
Residuals    553491 3798
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

model2 = lm(data=BWA2,SH.H2O.BASW.ZS ~ Time*Region) #interaction model
summary(model2)
Call:
lm(formula = SH.H2O.BASW.ZS ~ Time * Region, data = BWA2)

Residuals:
    Min      1Q  Median      3Q     Max
-52.352  -3.436   2.204   5.477  44.321

Coefficients:
                                               Estimate   Std. Error     t value
(Intercept)                                  -739.45437  182.53346    -4.051
Time                                            0.41206    0.09088     4.534
RegionEurope & Central Asia          623.54209  234.99312    2.653
RegionLatin America & Caribbean               196.75084  252.29851     0.780
RegionMiddle East & North Africa               67.08778  301.83472     0.222
RegionNorth America                           839.22092  718.18620    1.169
RegionSouth Asia                             -694.38593  428.86203   -1.619
```

```
RegionSub-Saharan Africa                  -551.57921  242.58542  -2.274
Time:RegionEurope & Central Asia            -0.30610    0.11699    -2.616
Time:RegionLatin America & Caribbean        -0.09546    0.12562   -0.760
Time:RegionMiddle East & North Africa         -0.03188    0.15028  -0.212
Time:RegionNorth America              -0.41219    0.35746  -1.153
Time:RegionSouth Asia                         0.34368    0.21352   1.610
Time:RegionSub-Saharan Africa                 0.26094    0.12078   2.161
                                               Pr(>|t|)
(Intercept)                                   5.20e-05 ***
Time                                          5.96e-06 ***
RegionEurope & Central Asia                   0.00800 **
RegionLatin America & Caribbean               0.43554
RegionMiddle East & North Africa              0.82412
RegionNorth America                           0.24267
RegionSouth Asia                              0.10550
RegionSub-Saharan Africa                      0.02304 *
Time:RegionEurope & Central Asia              0.00892 **
Time:RegionLatin America & Caribbean          0.44732
Time:RegionMiddle East & North Africa         0.83202
Time:RegionNorth America                      0.24894
Time:RegionSouth Asia                         0.10758
Time:RegionSub-Saharan Africa                 0.03079 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.03 on 3792 degrees of freedom
Multiple R-squared:  0.575,    Adjusted R-squared:  0.5735
F-statistic: 394.6 on 13 and 3792 DF,  p-value: < 2.2e-16
Anova(model2)
Anova Table (Type II tests)

Response: SH.H2O.BASW.ZS
              Sum Sq   Df    F value    Pr(>F)
Time           14452    1    99.8717  < 2.2e-16 ***
Region        722003    6   831.5980  < 2.2e-16 ***
Time:Region     4782    6     5.5074  1.087e-05 ***
Residuals   548710 3792
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ggplot(model2,aes(x=Time,y=SH.H2O.BASW.ZS,col=Region)) +
  geom_point() +
  geom_smooth(method=lm,se=F) +
  labs(y= "Overall Basic Water Service Access(%)", x = "Time (in year)")
```

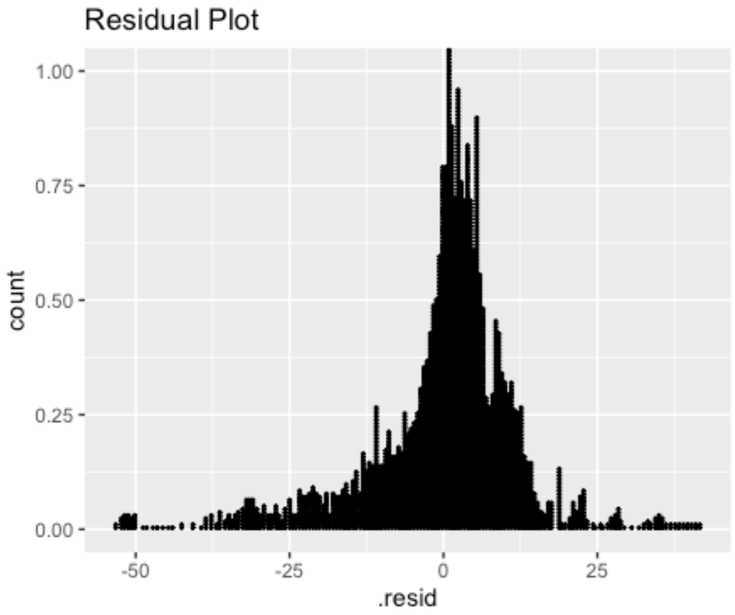


```
#validity conditions
plot(model1)
```





```
ggplot(model1,aes(x=.resid))+
  geom_dotplot(binwidth = 0.5)+
  ggtitle("Residual Plot")
```


Residual Plot
count
.resid