STATS 501 Project Report

Akhila Reddy

2022-06-29

Introduction

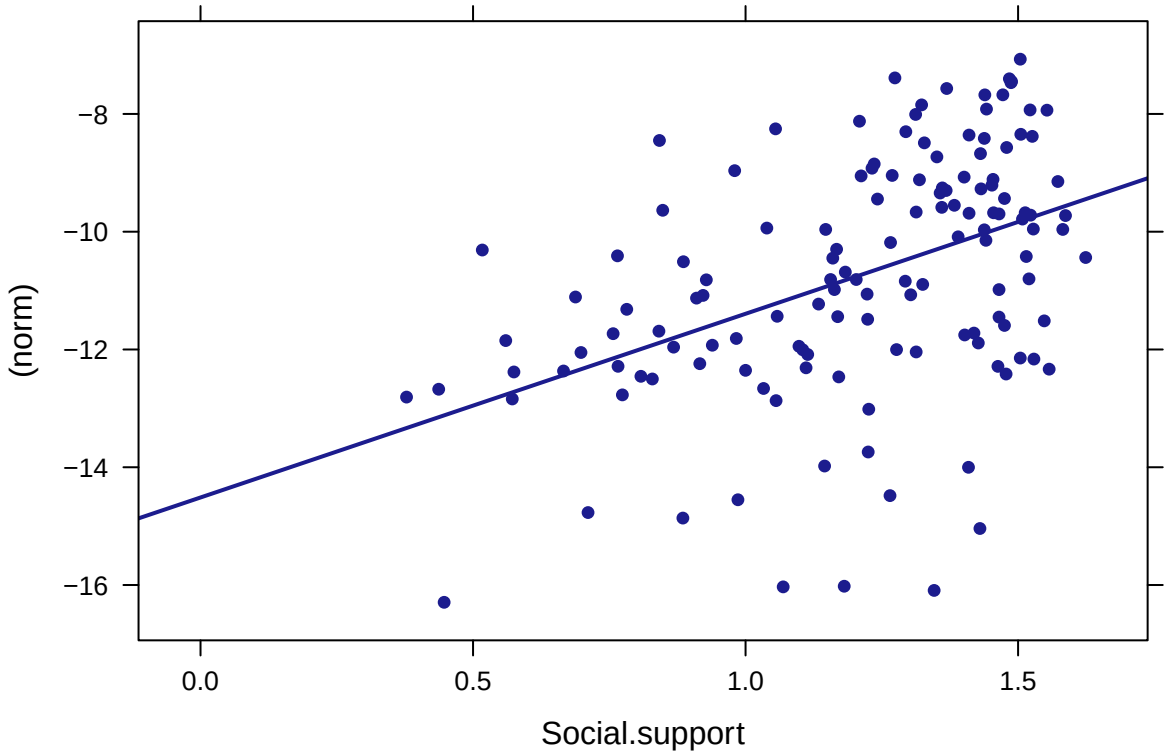
We have used data set "COVID-19 Dataset" and "World Happiness Report". These datasets are available on Kaggle. We have combined both the data sets. The main objective of this project is to analyze whether there is a linear relationship between Total Deaths and GDP per capita and between Total Deaths and social support. Our motivation to conduct this analysis is to understand if GDP per capita and social Support have any effect on covid-19 Total Deaths. Such an understanding would be important while devising health policies. Since we are analysis total deaths from different countries, normalized total deaths(total deaths/ population)would give us better understanding as- more the population, more active cases and more number of deaths. There is no mathematical formula to show relation between Total Deaths and GDP per capita or between Total Deaths and social support. Hence, we check for a linear relationship between the variables and see if the assumptions of the linear relationship satisfy.

DATA

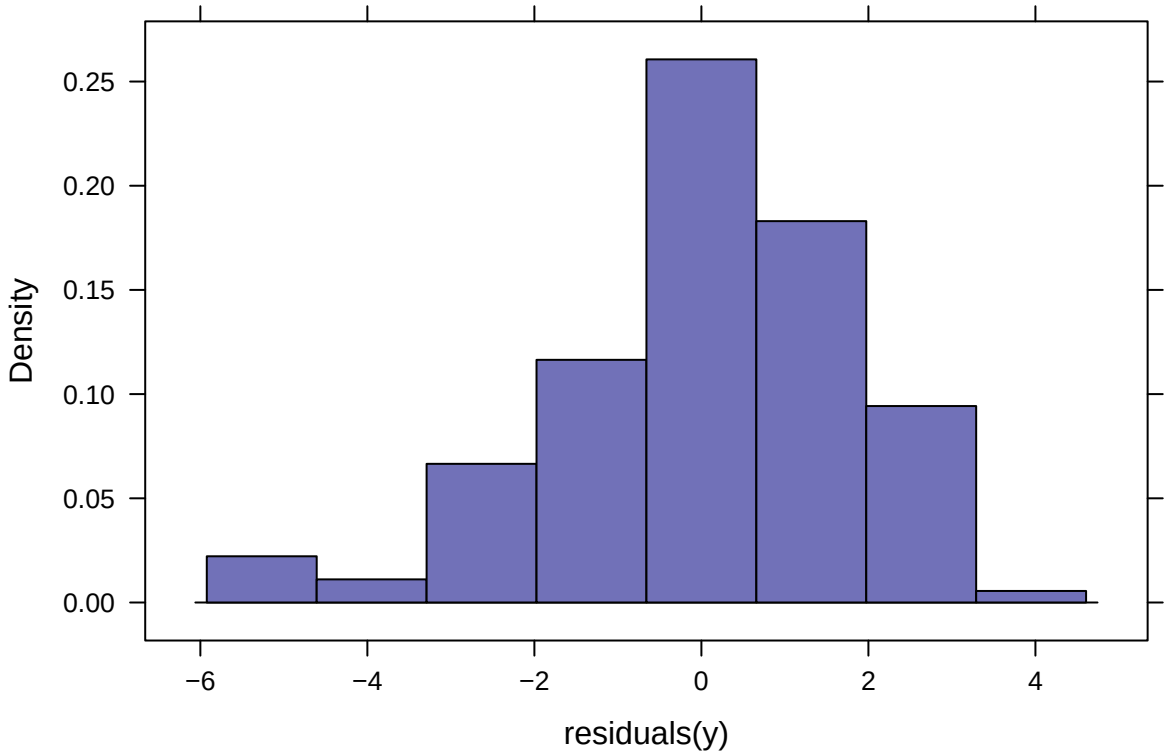
```
##   Overall.rank    Country.or.region      Score       GDP.per.capita
##  Min.   :  1.00   Length:156         Min.   :2.853   Min.   :0.0000
##  1st Qu.: 39.75   Class :character   1st Qu.:4.545   1st Qu.:0.6028
##  Median : 78.50   Mode  :character   Median :5.380   Median :0.9600
##  Mean   : 78.50                      Mean   :5.407   Mean   :0.9051
##  3rd Qu.:117.25                      3rd Qu.:6.184   3rd Qu.:1.2325
##  Max.   :156.00                      Max.   :7.769   Max.   :1.6840
##  Social.support  Healthy.life.expectancy Freedom.to.make.life.choices
##  Min.   :0.000   Min.   :0.0000          Min.   :0.0000
##  1st Qu.:1.056   1st Qu.:0.5477          1st Qu.:0.3080
##  Median :1.272   Median :0.7890          Median :0.4170
##  Mean   :1.209   Mean   :0.7252          Mean   :0.3926
##  3rd Qu.:1.452   3rd Qu.:0.8818          3rd Qu.:0.5072
##  Max.   :1.624   Max.   :1.1410          Max.   :0.6310
##    Generosity     Perceptions.of.corruption
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.1087   1st Qu.:0.0470
##  Median :0.1775   Median :0.0855
##  Mean   :0.1848   Mean   :0.1106
##  3rd Qu.:0.2482   3rd Qu.:0.1412
##  Max.   :0.5660   Max.   :0.4530
```

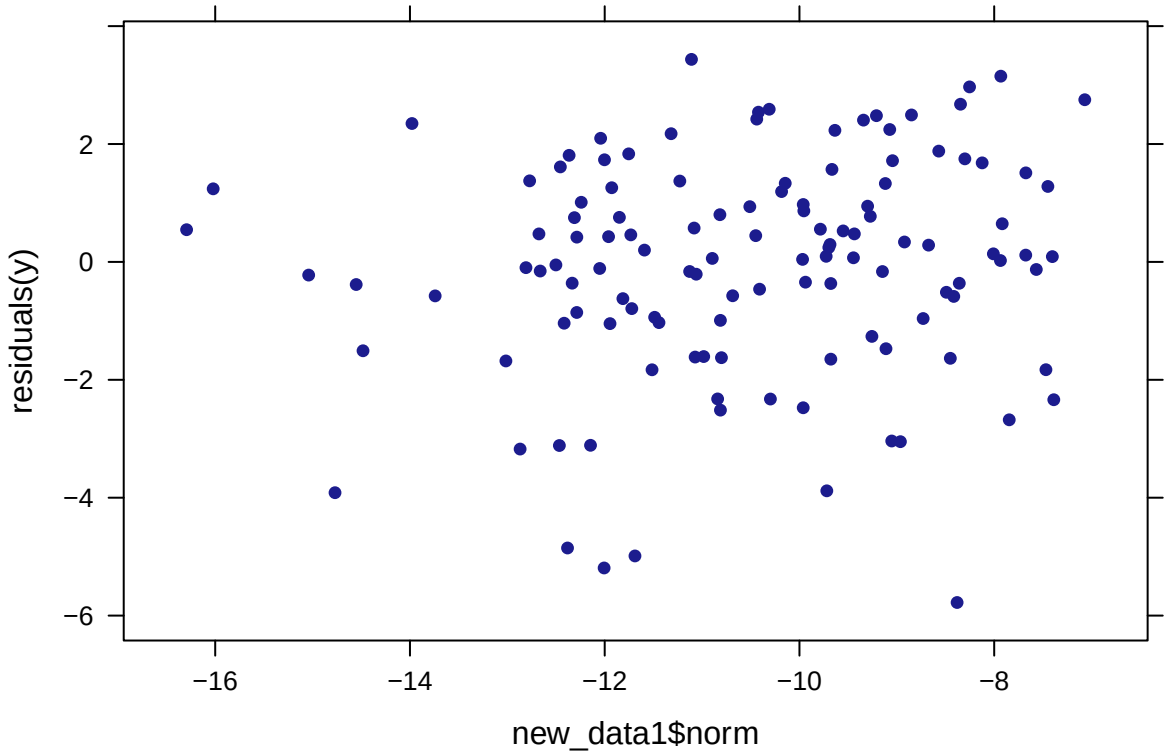
Including Plots

Total Deaths vs Social Support




```
##
## Call:
## lm(formula = (norm) ~ Social.support, data = new_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7778 -0.9607  0.1368  1.2804  3.4360
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -14.5137     0.6671 -21.757  < 2e-16 ***
## Social.support   3.1200     0.5365   5.816 4.16e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.83 on 135 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.2003, Adjusted R-squared:  0.1944
## F-statistic: 33.82 on 1 and 135 DF,  p-value: 4.16e-08
```


Density
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
residuals(y)

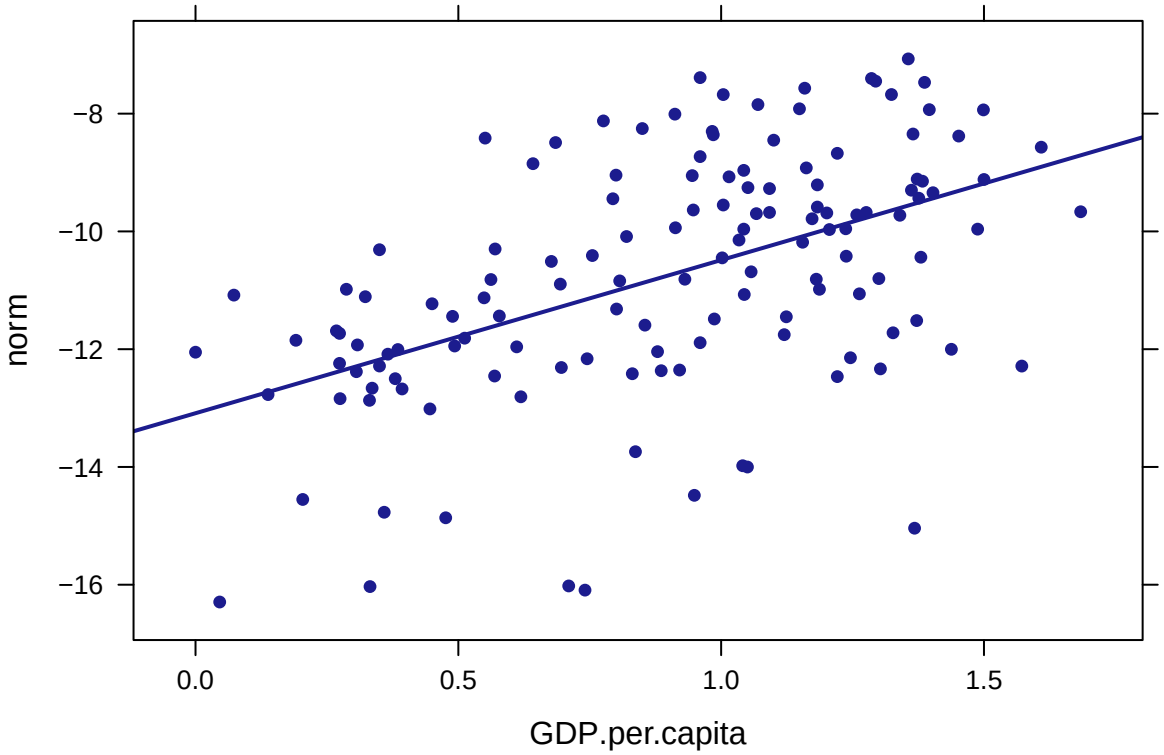


The slope is 3.12 This means that with every unit increase in Temperature, Humidity decreases by 1.293e-02 units. The value of R2 is 0.2 (from the model summary). This means that 20% of the variation in Total deaths can be explained Social Support. The correlation coefficient is sqrt(R^2), which is equal to 0.4472 which shows that there is weak linear relationship.

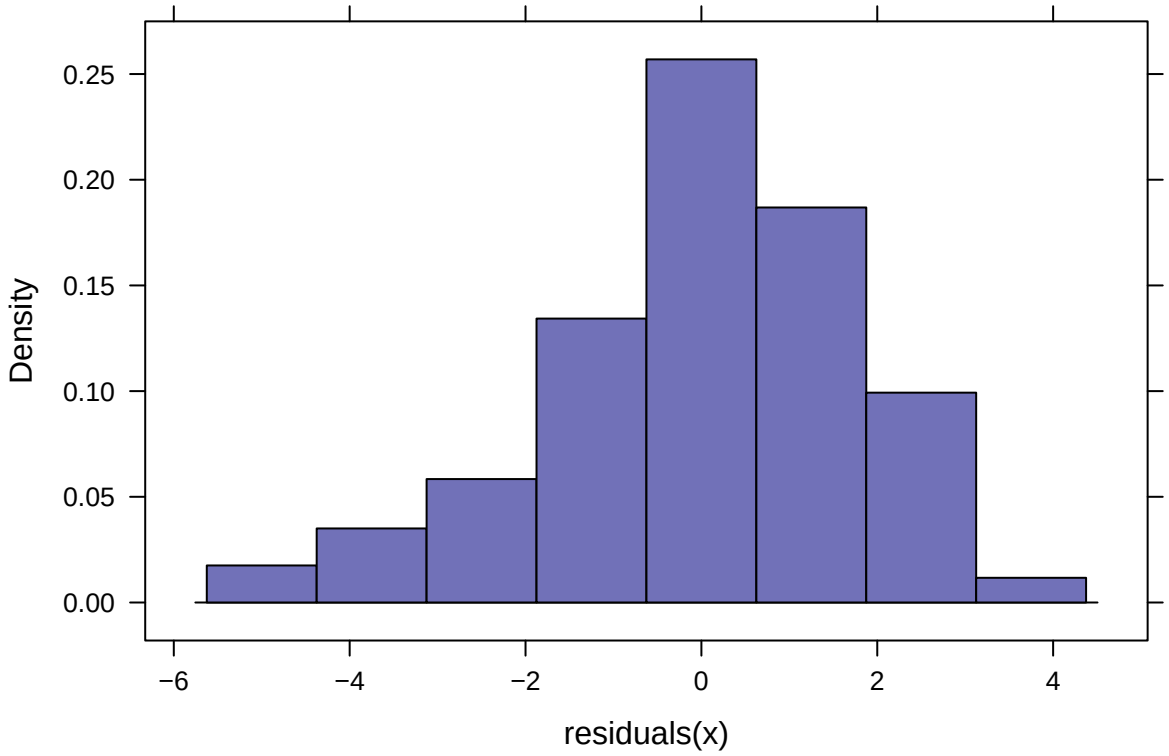
Assumptions for linear regression. 1) Linearity: This condition holds as seen from the scatterplot of Humidity vs Temperature. 2) Residuals are normally distributed: The residuals' distribution is approximately normal as seen form the histogram above. 3) Independence of residuals: As seen from the scatter plot above, there seems to be no relation between various residuals. 4) Residual variance is constant: The variance of residuals is constant as seen from the scatter plot below. 5) Expected value of residuals has to be zero: From the scatter plot it seems that the number of points above and below 0 are equal, so the mean of residuals is close to zero.

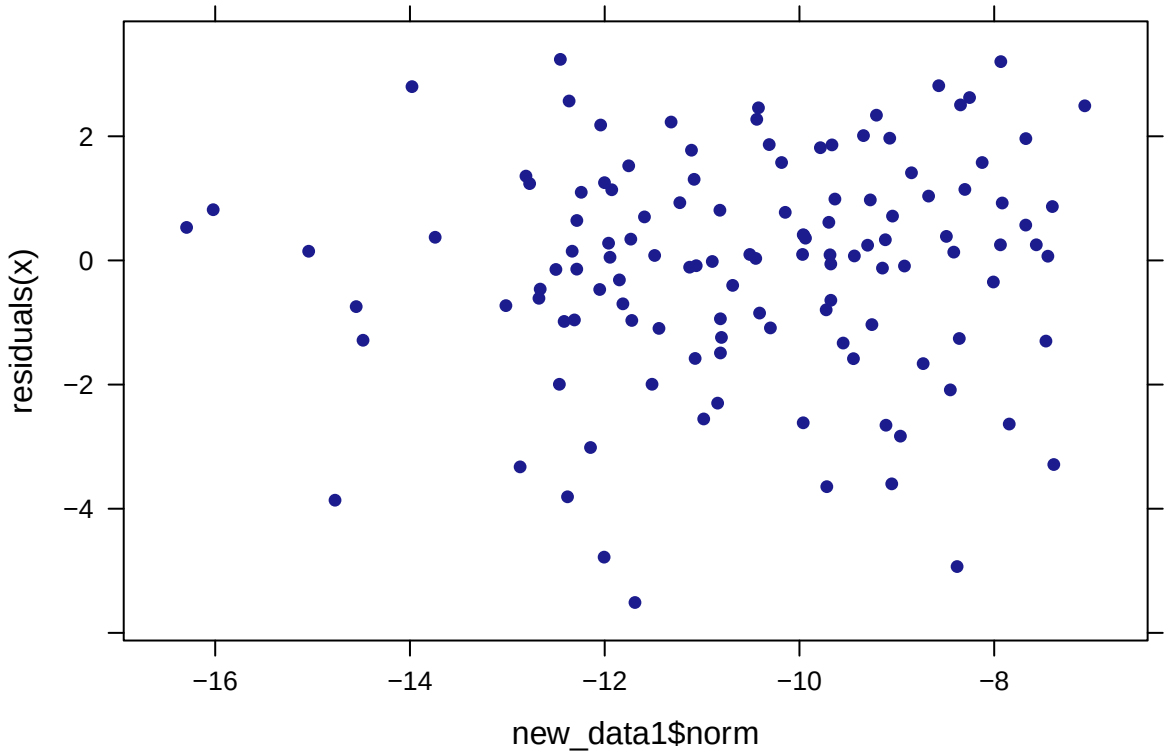
In conclusion, there are no strong violations of any assumptions. Hence, the linear model is appropriate.

Total Deaths vs GDP per capita




```
##
## Call:
## lm(formula = norm ~ GDP.per.capita, data = new_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5107 -0.9586  0.1338  1.1386  3.2391
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -13.0876     0.3743 -34.968  < 2e-16 ***
## GDP.per.capita   2.6010     0.3799   6.846 2.44e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.763 on 135 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.2577, Adjusted R-squared:  0.2522
## F-statistic: 46.86 on 1 and 135 DF,  p-value: 2.435e-10
```


Density
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
residuals(x)



The slope is 2.601. This means that with every unit increase in GDP , the Total Deaths increase by 2.601 units. The value of R2 is 0.258 (from the model summary). This means that 25.8% of the variation in Total Deaths can explained by GDP. The correlation coefficient is sqrt(R^2), which is equal to 0.507 which shows that there is moderate linear relationship. Assumptions for linear regression. 1) Linearity: This condition holds as seen from the scatterplot of Total Deaths vs GDP per capita. 2) Residuals are normally distributed: The residuals' distribution is approximately normal as seen form the histogram above. 3) Independence of residuals: As seen from the scatterplot below, there seems to be no visible pattern between the residual data points. Hence we can say that this condition is satisfied. 4) Residual variance is constant: The variance of residuals is constant as seen from the scatterplot below. 5) Expected value of residuals has to be zero: From the scatterplot it seems that the number of points above and below 0 are equal, so the mean of residuals is close to zero.

In conclusion, there are no strong violations of any assumptions. Hence, the linear model is appropriate.

Conclusion:

From the above analysis on the "Weather in Szeged 2006 - 2016 dataset", we can conclude that Humidity is moderately linearly related to Temperature and ApparentTemperature. Also, comparing the r values for each model(as shown in the above table), we can see that Humidity can be best predicted considering all the numerical variables and building a linear model than when only done with Temperature or Apparent-Temperature.