Name: Birru Lavanya **Roll number: 21HS10018**

Objective:

The main objective of the project is to explain the infant mortality rate (per 1000 live births).

The dependent variable here is infant mortality rate (per 1000 live births) of different countries in the world. The independent variables are Poverty headcount ratio at $2.15 a day (2017 PPP) (% of population), current health exp (% of GDP), Births attended by skilled health staff (% of total). The objective is to find the variables that influence infant mortality rate of a country and their relative impact on the dependent variable, to test if the estimated model suffers from heteroscedasticity and finding the remedial measures to solve the heteroscedasticity.

Specification of the econometric model:

First of all, I collected data of the infant mortality rate of each country during 2017-2018 from world bank data. Then I gave some thought to what factors might be influencing it. I ended up with these variables:

% of skilled health staff - I could find Births attended by skilled health staff (% of total)

Govt. Health expenditure (%of GDP)- I could find data for this.

Teenage mothers (% of total pregnant women) - I couldn't find the data for most of the countries. So, I had to drop this.

People below poverty line (% of population) - I could find Poverty headcount ratio at $2.15 a day (2017 PPP) (% of population).

% of women receiving prenatal care -I couldn't find data for % of women receiving prenatal care. Then I found this is closely related to govt health expenditure which I already included in my model.

$$Y = A_0 + A_1 X_1 + A_2 X_2 + A_3 X_3$$

Y = infant mortality rate (per 1000 live births)

X_1 = Poverty headcount ratio at $2.15 a day (2017 PPP) (% of population)

X_2 = current health exp (% of GDP)

X_3 = Births attended by skilled health staff (% of total).

Justification of the model:

Intuitively, infant children are most likely to not survive if delivery wasn't done under the surveillance of skilled health staff. The health of infants also depends on the health of the mother. Poverty determines the level of nutrition in the mothers. If the household is poor, then it is most likely that members of the family are undernourished. Infant mortality rate also depends on health facilities in one's country. Even if the children are born weak, medical facilities like incubators can be helpful in improving the health of infants. So, the health expenditures of the country can be used as one of the independent variables.

Regression:

```
             |     mort  poverty      exp  skilled
-------------+------------------------------------
        mort |   1.0000
             |
             |
     poverty |   0.8680   1.0000
             |   0.0000
             |
         exp |  -0.5364  -0.4121   1.0000
             |   0.0001   0.0029
             |
     skilled |  -0.9068  -0.8556   0.4254   1.0000
             |   0.0000   0.0000   0.0021
             |

. pcorr mort poverty exp skilled
(obs=50)

Partial and semipartial correlations of mort with

               Partial   Semipartial      Partial   Semipartial   Signi
> ficance
   Variable |    Corr.         Corr.      Corr.^2       Corr.^2
>   Value
------------+-----------------------------------------------------------
> -------
    poverty |   0.4178        0.1618       0.1746        0.0262
>  0.0031
        exp |  -0.3899       -0.1490       0.1520        0.0222
>  0.0062
    skilled |  -0.6363       -0.2902       0.4048        0.0842
>  0.0000
```

Statistically significant pair-wise correlation coefficients.

Statistically significant partial correlation coefficients.

I regressed the model in Stata. The results can be checked below:

```
. reg mort poverty exp skilled

      Source |       SS           df       MS      Number of obs   =        50
-------------+----------------------------------   F(3, 46)        =    108.48
       Model |  14676.1963         3  4892.06543   Prob > F        =    0.0000
    Residual |  2074.38019        46  45.0952215   R-squared       =    0.8762
-------------+----------------------------------   Adj R-squared   =    0.8681
       Total |  16750.5765        49  341.8485     Root MSE        =    6.7153

------------------------------------------------------------------------------
        mort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     poverty |   1.011945   .3244232     3.12   0.003     .3589152    1.664975
         exp |  -1.085441   .3780167    -2.87   0.006    -1.846349    -.324533
     skilled |  -.8443069   .1509356    -5.59   0.000    -1.148125   -.5404892
       _cons |   98.21013   14.86557     6.61   0.000     68.28729     128.133
------------------------------------------------------------------------------
```

Interpretation of the results:

The estimated model is statically significant at 1% significance level. R-squared and adjusted R-squared values are significantly high. Regression shows that about 87% of variations in the dependent variable are explained by the model. All the independent variables are statistically significant at 1% significance level and neither of them has wrong sign. When poverty is high, mortality rate is also expected to be high and hence the coefficient is positive. The more skilled the health staff, the less are the chances of infant mortality. Hence the coefficient is negative. The more the government spends on health, the better are the health facilities, the less are the chances of infant mortality. Hence the coefficient is negative.

Tests for heteroskedasticity:

Since model is regressed on cross sectional data, it is more likely to have heteroscedasticity. So, I carried out the following tests in Stata.

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
     skilled |      3.82    0.261596
     poverty |      3.77    0.265169
         exp |      1.23    0.810379
-------------+----------------------
    Mean VIF |      2.94

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of mort

         chi2(1)      =    19.05
         Prob > chi2  =   0.0000

. estat hettest poverty exp skilled

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: poverty exp skilled

         chi2(3)      =    20.96
         Prob > chi2  =   0.0001

. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(9)      =     16.87
         Prob > chi2  =    0.0508

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      16.87      9    0.0508
            Skewness |       6.94      3    0.0738
            Kurtosis |       1.90      1    0.1676
---------------------+-----------------------------
               Total |      25.71     13    0.0186
---------------------------------------------------
```

The vif value for skilled variable is considerably high. So, there is a possibility of severe multicollinearity problem. The statistical tests performed to test if there's heteroscedasticity have rejected the null hypothesis. So, the model suffers from heteroscedasticity.

To solve for multicollinearity, I dropped the skilled variable. Upon regressing the model on remaining independent variables, the results are as follows:

```
. reg mort poverty exp

      Source |       SS           df       MS      Number of obs   =        50
-------------+----------------------------------   F(2, 47)        =     89.44
       Model |  13265.1295         2  6632.56473   Prob > F        =    0.0000
    Residual |  3485.44701        47  74.1584471   R-squared       =    0.7919
-------------+----------------------------------   Adj R-squared   =    0.7831
       Total |  16750.5765        49  341.8485     Root MSE        =    8.6115

------------------------------------------------------------------------------
        mort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     poverty |   2.509093   .2351248    10.67   0.000     2.036083    2.982103
         exp |   -1.41212   .4789392    -2.95   0.005    -2.375621   -.4486184
       _cons |   16.80246   3.887792     4.32   0.000     8.981233    24.62369
------------------------------------------------------------------------------

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
         exp |      1.20    0.830193
     poverty |      1.20    0.830193
-------------+----------------------
    Mean VIF |      1.20

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of mort

         chi2(1)      =    17.36
         Prob > chi2  =   0.0000

. estat hettest poverty exp

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: poverty exp

         chi2(2)      =    19.76
         Prob > chi2  =   0.0001

. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(5)      =    11.84
         Prob > chi2  =   0.0370

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      11.84      5    0.0370
            Skewness |       7.77      2    0.0206
            Kurtosis |       1.78      1    0.1818
---------------------+-----------------------------
               Total |      21.39      8    0.0062
---------------------------------------------------
```

Vif values are low. So, there is no severe multicollinearity. But this didn't solve the problem of heteroscedasticity, since all the statistical tests have rejected the null hypothesis.

To solve for heteroscedasticity, I took logarithmic transformation on all the variables. For poverty variable some of the observations are 0. So, Stata removed such observations. Upon regressing,

```
. reg mort_ pov_ exp_ skl

      Source |       SS           df       MS      Number of obs   =        40
-------------+----------------------------------   F(3, 36)        =     36.18
       Model |  33.5967694         3  11.1989231   Prob > F        =    0.0000
    Residual |  11.1418163        36  .309494898   R-squared       =    0.7510
-------------+----------------------------------   Adj R-squared   =    0.7302
       Total |  44.7385858        39  1.14714322   Root MSE        =    .55632

------------------------------------------------------------------------------
       mort_ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pov_ |   .2525873   .0807035     3.13   0.003     .0889129    .4162616
        exp_ |  -.8963993   .2969214    -3.02   0.005    -1.498584   -.2942149
         skl |  -1.634498    .675527    -2.42   0.021    -3.004531   -.2644662
       _cons |   11.22611   2.930443     3.83   0.000     5.282892    17.16932
------------------------------------------------------------------------------

. vif

    Variable |       VIF       1/VIF  
-------------+----------------------
        pov_ |      2.07    0.482644
         skl |      1.92    0.521696
        exp_ |      1.88    0.533176
-------------+----------------------
    Mean VIF |      1.95

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity 
         Ho: Constant variance
```

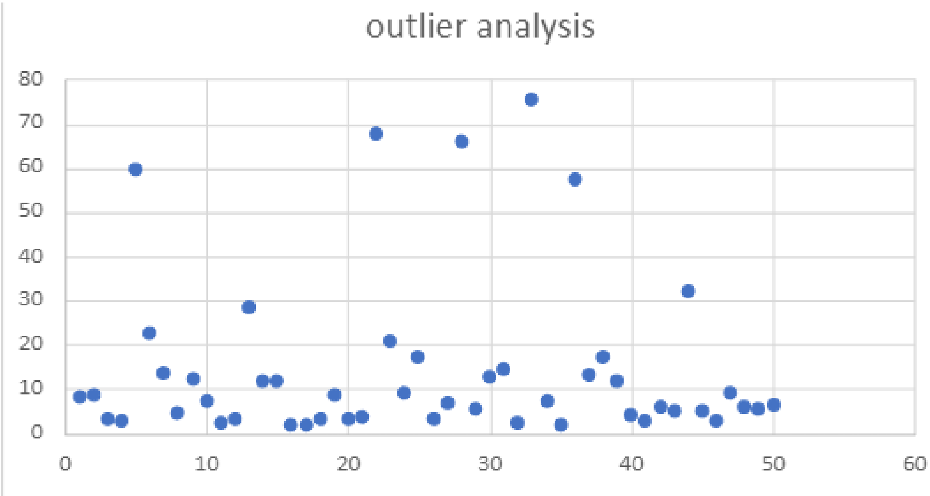
```
        Variables: fitted values of mort_

        chi2(1)      =     1.34
        Prob > chi2  =   0.2473

. estat hettest pov_ exp_ skl

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
        Ho: Constant variance
        Variables: pov_ exp_ skl

        chi2(3)      =     1.98
        Prob > chi2  =   0.5773

. estat imtest,white

White's test for Ho: homoskedasticity
        against Ha: unrestricted heteroskedasticity

        chi2(9)      =     8.57
        Prob > chi2  =   0.4781

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	8.57	9	0.4781
Skewness	1.56	3	0.6674
Kurtosis	0.36	1	0.5512
Total	10.49	13	0.6537

The estimated model is statistically significant at a 1% significance level. R-squared and adjusted R-squared values are considerably high. All the coefficients are statistically significant at 5% significance level. Constant term of the model is highly positive and statistically significant. These results are consistent with our expectations. Vif values are low. So, there is no problem of multicollinearity. All the computed statistical tests for heteroscedasticity don't reject the null hypothesis. So, there is no problem of heteroscedasticity.

Alternate procedure to solve for heteroscedasticity and multicollinearity:

Later I removed the outliers from the original data to check if they are causing the problem of heteroscedasticity. So, I plotted the dependent variable a scatter plot to check for outliers in excel.



I removed those 5 outliers and regressed the data in Stata. Surprisingly, the outlier removal has solved both multicollinearity and heteroscedasticity.

Source	SS	df	MS		
				Number of obs =	45
				F(3, 41) =	15.69
Model	1153.13921	3	384.379737	Prob > F =	0.0000
Residual	1004.23277	41	24.4934821	R-squared =	0.5345
				Adj R-squared =	0.5005
Total	2157.37198	44	49.0311813	Root MSE =	4.9491

morta	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pover	1.201831	.4552266	2.64	0.012	.2824821	2.12118
expe	-.7717225	.2869192	-2.69	0.010	-1.351168	-.1922774
skill	-.4084158	.1608219	-2.54	0.015	-.7332022	-.0836295
_cons	52.46637	16.03691	3.27	0.002	20.07918	84.85357

```
. vif
```

Variable	VIF	1/VIF
pover	1.53	0.654368
skill	1.50	0.664564
expe	1.06	0.943797
Mean VIF	1.36	

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance

. estat hettest pover expe skill

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: pover expe skill

         chi2(3)      =     4.34
         Prob > chi2  =   0.2268

. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(9)      =      3.63
         Prob > chi2  =    0.9341

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	3.63	9	0.9341
Skewness	2.62	3	0.4538
Kurtosis	1.17	1	0.2793
Total	7.42	13	0.8792

As it can be seen, the estimated model is statistically significant at a 1% significance level. R-squared and adjusted R- squared values are considerably high. All the coefficients are statistically significant at 5% significance level. Constant term of the model is highly positive and statistically significant. These results are consistent with our expectations. Vif values are low. So, there is no problem of multicollinearity. All the computed statistical tests for heteroscedasticity don't reject the null hypothesis. So, there is no problem of heteroscedasticity.

Conclusions:

Hence it can be concluded that the model is better due to robustness in explaining the variations in data. It doesn't suffer from the problems of heteroscedasticity and multicollinearity.

Further to our analysis, we can try to include more variables in the model like environmental pollution, % of teenage mothers, genetics to better my model. We can also check if we can change the functional forms of some more variables. We can try to find better proxies for the variables which we have already used. We can also try to change the years of some of our variables or use lagged data. For example, if we use lagged health expenditure on my model, it might be giving better conclusions. Since we know that implementation of government policies takes time to show effects.