Econometric Data Analysis

Topic: Impact of Gender Inequality in Education on Economic Growth

INTRODUCTION

Any strategy towards the comprehensive socio-economic progress is incomplete without the incorporation of the role of women, as it is an important determinant of progress, because apart from constituting for half of population, women also bear the brunt of the daily survival, especially in a developing country. Moreover, "the role of women in development has been conceived as an integral process of economic growth and social progress in the contemporary world". (Imran Sharif Chaudhry, 2007)

On a contrary, Indian women have often been considered as the weaker and vulnerable section of the society in terms of education, health, workforce opportunities, livelihood conditions, etc. According to the UNDP's Human Development Report, India's Gender Development Index is 0.841, with a rank of 130 out of 189 countries. The report also elucidates the gender inequality prevailing in the country, as the Human Development index is 0.575 for females, while for males the index is 0.683. Similarly, the estimated GDP per women is just $2,772, while for men it is $9,729. It also suggests that the share in employment by Indian women is also very low. These figures indicate that India is not close in achieving the United Nations Sustainable Development goal of providing gender equality in the nation. This gender inequality adverse effects a large number of valuable development goals. Firstly, gender equality in education may prevent child mortality, increase in fertility, family planning and an expansion for the education to the upcoming generations. Secondly, these inequalities often lower the average level of human capital, which consequently hinder economic growth.

ECONOMIC GROWTH AND GENDER INEQUALITY IN EDUCATION

With the emergence of globalisation and increasing global competition, Education has become crucial for the in human capital formation in the 21st century. Education not only raises productivity and efficiency, but also opens the path for achieving socio-economic goals.

In India, in addition to the lower levels of literacy rate (69.3%), education in India also faces a great deal of gender inequality. According to the UNDP's report, mean years of schooling for female (4.8 years) is lower than that for male (8.2 years), population with at least some secondary education for female (39%) is lower than that for male (63.5) and literacy rate for female (65.46%) is lower than that for male (82.14%). The figures suggest that there is a great level of gender disparity in education in India, which not only affects efficiency of human capital but also reduces the economic growth.

According to the economic theory and numerous research papers, evidence suggests that as the gender inequality in education increases the economic growth reduces due to the poor quality of human capital, as women constitute almost half of the population. Similarly, with the reduction in education inequality, more women are educated which enhances the human capital and consequently increases the level of economic growth. The analysis is relevant because economic growth and gender equality in education, both plays a vital role in achieving socio-economic goals and paves the way for development for India.

LITERATURE REVIEW

The relationship between the gender inequality in education and economic growth has been explored extensively and provides a numerous amount of literature. Some frameworks are dependent on particular demographics while other differs from one another on the basis of the independent variables.

Imran Sharif Chaudhary (2007) explains the relationship in his paper **"Gender Inequality in Education and Economic growth: Case Study of Pakistan",** where he ran the regression analysis for the states in Pakistan for the years 1970-2005. The study illustrated that "gender inequality in education has a strong and significant impact on economic growth in Pakistan". Additionally, Klasen (1999) in his paper **"Does Gender Inequality Reduce Growth and Development?"**, argues that "there is a negative relation between gender inequalities in education on economic growth, but not vice versa", on the basis of the results of panel regressions, and the prediction of female-male ratio of growth in average education based on government spending on education and changes in fertility rates between 1960 and 1990.

Another paper, **"The Impact of Gender Inequality in Education and Employment on Economic Growth"** by Klasen and Lamanna (2009), which explains the relationship for the South Asian countries for the period 1960-2000. According to their findings, "gender inequality in education reduced economic growth till the 1990s, while the impact reduced afterwards with the decrease in the level of gender inequality". They further explained that gender inequality in labour force also reduces the level of economic growth for these countries, elucidating the importance of women in the process of development.

However, Nayef Al-Shammari and Monira Al Rakhis (2017) in their paper **"Impact of Gender Inequality on Economic Growth in the Arab Region",** a panel regression of 19 Arab countries over the period from 1990-2014, explains that there is a minimal effect of gender inequality on economic growth in Arab countries. The evidence suggests that, "economic growth depends more on capital accumulation and population growth and as long as these two factors are growing, gender inequality will not hinder the economic growth of these countries".

OBJECTIVE

The **objective** of this paper is to capture this impact of gender inequality in education on the economic growth across the Indian states and union territories for the year 2011-2012, using econometric analysis. After running the regression, we can conclude whether our hypothesis, gender inequality in education has a negative impact on economic growth, holds true or not.

DATA AND METHODOLOGY

According to the hypothesis, we test whether gender inequality in education has a negative effect on economic growth. For capturing the effect on economic growth, we take Gross State Domestic Product (GSDP) as an indicator of economic growth, which is the dependent variable. The data for the same has been collected from the **Statistics Handbook provided by Reserve bank of India** for the year 2011-12. The data is at Factor cost with 2011 as the base year and is normalized in terms of Rs. 1 crore. The report prepared by RBI has data for all States and Union Territories except Daman and Dui, Dadar and Nagar Haveli and Lakshwadeep therefore there are three missing values in the data set. The problem with this data set is that almost two-third of female economic activities go unrecorded in developing economies because of the nature of the activities, which mostly include household work. Thus, the analysis will suffer from this shortcoming.

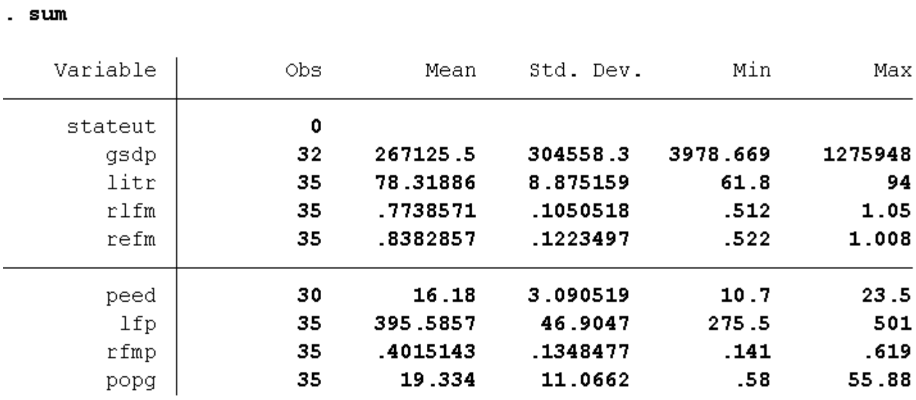
The primary independent variable in our model is the gender inequality in education. For capturing the inequalities, we use two sets of data as proximate variables: female to male Literacy rate (RLFM) and female to male enrolment ratios (REFM). We anticipate that with the increase in these variables i.e. gender equality rises, the level of economic growth also increases. The data has been collected from **Census of India** held in 2011.

Additionally, we take overall literacy rate (LITR), public expenditure on education as a proportion of total GSDP (PEED), labour force participation rate (LFP), female to male labour force participate ratio (RFMP) and population growth (POPG) as other explanatory variables in our model, which can possibly have an effect on economic growth. We anticipate that LITR, PEED, LFP and RFMP have a positive relation with the dependent variable, while POPG has a negative relationship with economic growth. The data for LITR and POPG has been collected from **Census** 2011, PEED has been collected from **RBI** where the data is missing for the union territories, LFP and RFMP has been collected from **NITI Aayog.**

The following table gives the variables and summarise them according to their number of observation, mean, standard deviation and minimum and maximum values.

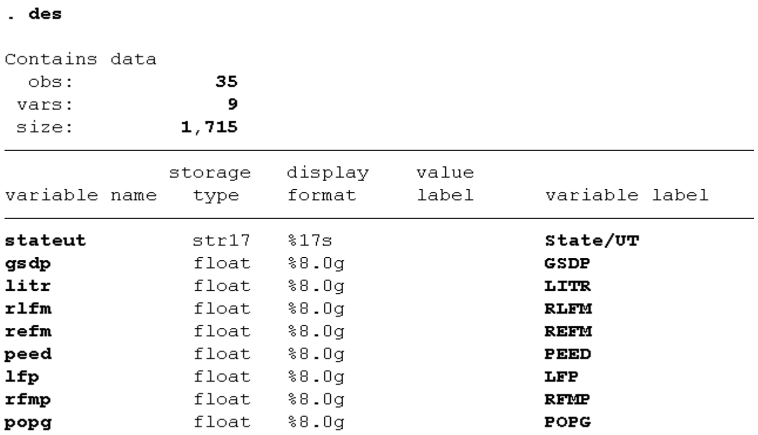
Variables	Description of the variables	Data Source
GSDP	Gross State Domestic Product	RBI-Handbook
LITR	Overall Literacy Rate	Census 2011
RLFM	Female to Male Literacy Ratio of 15 years and above	Census 2011
REFM	Female to Male Enrolment Ratio of 15 years and above	Census 2011
PEED	Public Expenditure on Education	RBI-Handbook
LFP	Labour Force Participation Rate	NITI Aayog
RFMP	Female to Male Labour Force Participation Ratio	NITI Aayog
POPG	Population Growth Rate	Census 2011

Summary of the data has been given below, which gives the number of observations, mean, standard deviation and minimum and maximum values.



```
. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     stateut |         0
        gsdp |        32    267125.5    304558.3   3978.669    1275948
        litr |        35    78.31886    8.875159       61.8         94
        rlfm |        35    .7738571    .1050518       .512       1.05
        refm |        35    .8382857    .1223497       .522      1.008
-------------+--------------------------------------------------------
        peed |        30       16.18    3.090519       10.7       23.5
         lfp |        35    395.5857     46.9047      275.5        501
        rfmp |        35    .4015143    .1348477       .141       .619
        popg |        35      19.334     11.0662        .58      55.88
```

Description of data shows that there are 35 observations on 9 variables. The first variable is stored as str17 and has a display format of %17s, while all the other variables are of storage type float and has a display format of %8.0g. None of the variable has any value label.



```
. des

Contains data
  obs:            35
 vars:             9
 size:         1,715
--------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
--------------------------------------------------------------
stateut         str17  %17s                  State/UT
gsdp            float  %8.0g                 GSDP
litr            float  %8.0g                 LITR
rlfm            float  %8.0g                 RLFM
refm            float  %8.0g                 REFM
peed            float  %8.0g                 PEED
lfp             float  %8.0g                 LFP
rfmp            float  %8.0g                 RFMP
popg            float  %8.0g                 POPG
```

METHODOLOGY

Keeping in view these variables, we plan the methodology of the findings. Following sets of equations are estimated to show the relationship between gender inequality in education and economic growth. Thus, in the basic specification the following two equations are estimated:

Model 1: Effect of gender disparity in education alone on the economic growth

$$GSDP = \beta_0 + \beta_1\, LITR + \beta_2\, RLFM + \beta_3\, REFM + \beta_4\, PEED + \varepsilon$$

Model 2: Effect of gender disparity in education along with other explanatory variables on the economic growth

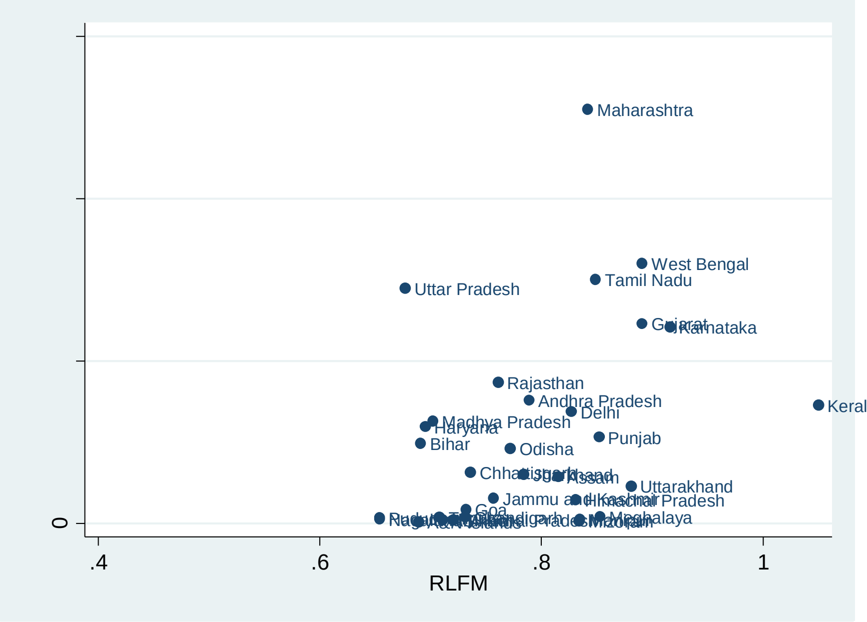
$$GSDP = \beta_0 + \beta_1\, LITR + \beta_2\, RLFM + \beta_3\, REFM + \beta_4\, PEED + \beta_5\, LEP + \beta_6\, REMP + \beta_7\, POPG + \varepsilon$$

We proceed by describing the statistic followed by regression analysis along with diagnostic tests. One of the limitations of the model is that sample size is low which might give fallacious results. In this model, we take the ratio of female to male for variables like literacy, enrolment and labour force participation instead of taking it separately for male and female, because we want to avoid the problem of multicollinearity in our regression.

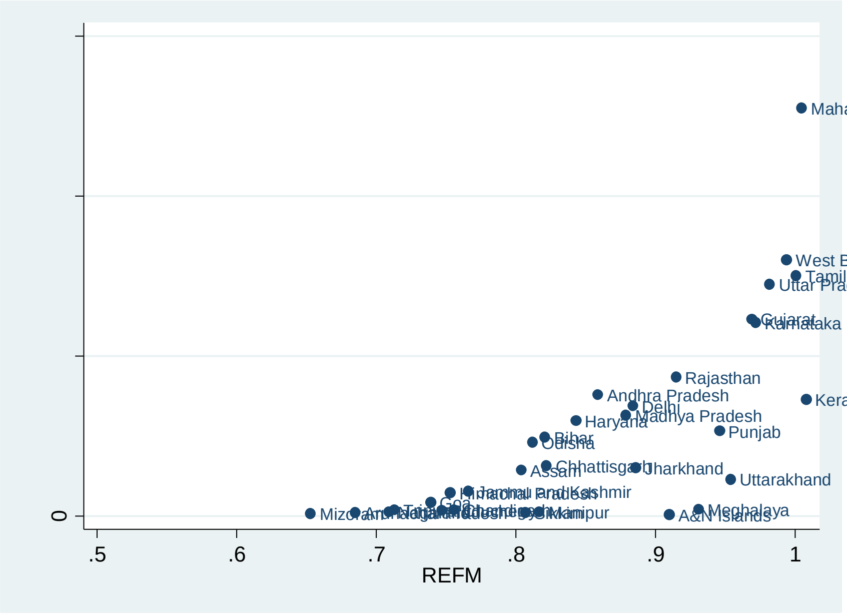
DESCRIPTIVE STATISTICS

We conduct the preliminary analysis by checking the relationship between the dependent variable and independent variable using the scatter plots and calculating the correlation between the two.

For our analysis, we plot GSDP with RLFM and GSDP with REFM. We also calculate the correlation between the two pairs which gives a correlation of **0.3665** between GSDP and RLFM, and a correlation of **0.7382** between GSDP and REFM. These figures indicate that with the increase in the level of RLFM and REFM, i.e. reduction in the inequality, the level of economic growth increases.



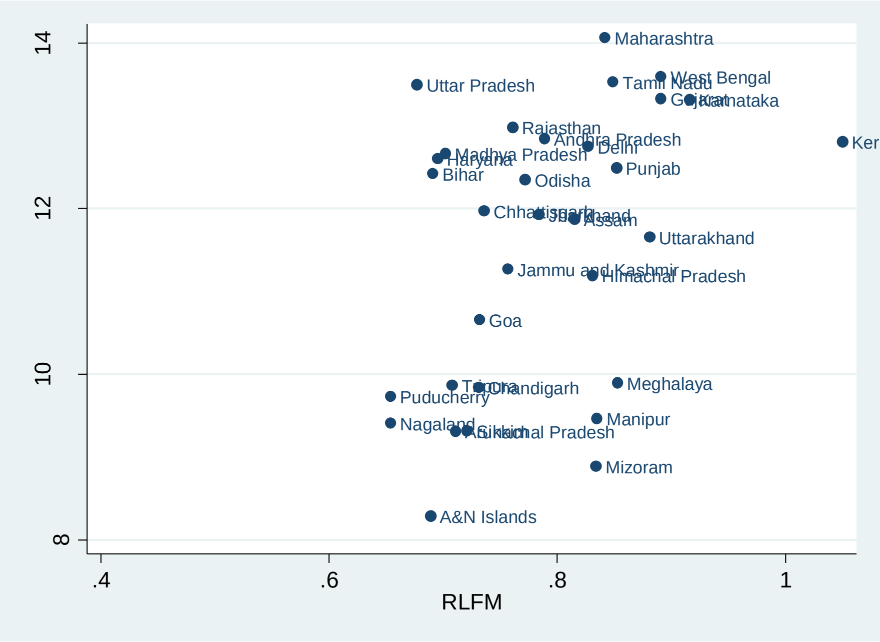
Correlation between GSDP and RLFM



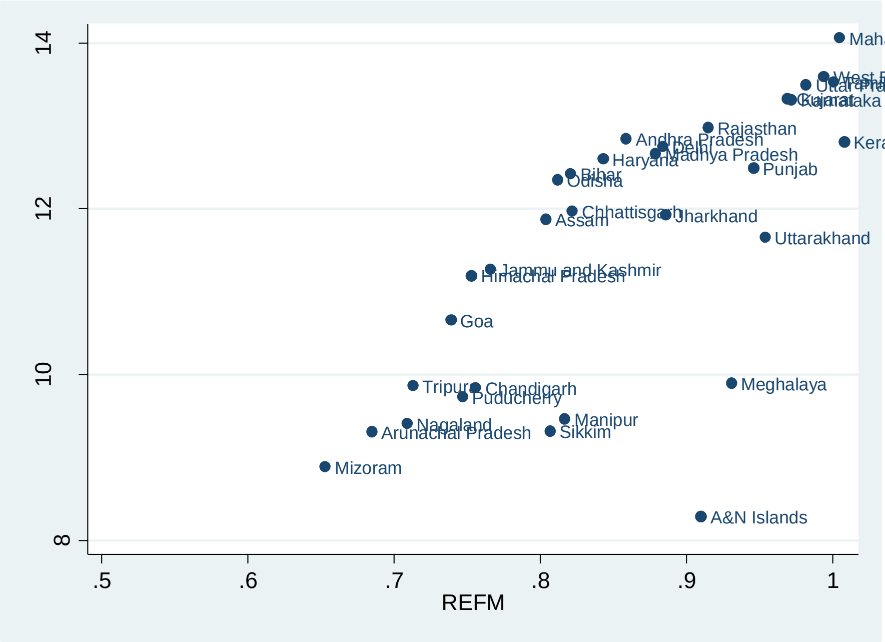
Correlation between GSDP and REFM

The summary of the statistics suggests that the data is highly skewed which can give misleading results. Further for running the linear regression, the variables must satisfy the Classical Liner Regression Model assumptions in order to obtain correct results from OLS method. We therefore perform transformations to make the variables normally distributed by using the command GLADDER. Therefore, we transform GSPD to its logarithmic form for getting better results. (We also check if other explanatory variables are normally distributed or not, and therefore we transform PEED to it logarithmic form and POPG to it square-root).

We again plot the log of GSDP with RLFM and REFM and yield the following scatter plots. Correlation between log GSDP and RLFM is **0.4084**, and between the log of GSDP and REFM is **0.7132**, which again suggests that there is a positive relationship between gender equality and economic growth or a negative relationship between gender inequality and economic growth.



Correlation between log of GSDP and RLFM



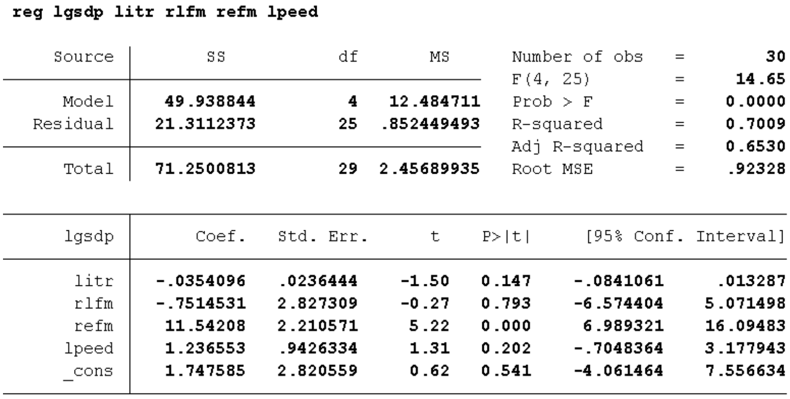
Correlation between log of GSDP and REFM

REGRESSION ANALYSIS

The hypothesis formed is confirmed by using confirmatory analysis. We establish the relationship between the concerned variables through the Ordinary Least Squares method by running a multiple linear regression model.

REGRESSION RESULTS

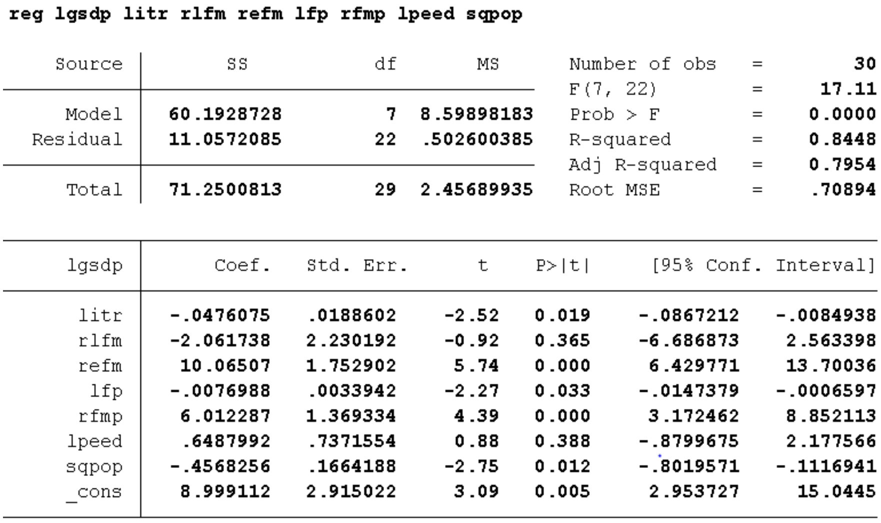
Model 1



```
reg lgsdp litr rlfm refm lpeed

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(4, 25)        =     14.65
       Model |   49.938844         4   12.484711   Prob > F        =    0.0000
    Residual |  21.3112373        25  .852449493   R-squared       =    0.7009
-------------+----------------------------------   Adj R-squared   =    0.6530
       Total |  71.2500813        29  2.45689935   Root MSE        =    .92328

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0354096   .0236444    -1.50   0.147    -.0841061     .013287
        rlfm |  -.7514531   2.827309    -0.27   0.793    -6.574404    5.071498
        refm |   11.54208   2.210571     5.22   0.000     6.989321    16.09483
       lpeed |   1.236553   .9426334     1.31   0.202    -.7048364    3.177943
       _cons |   1.747585   2.820559     0.62   0.541    -4.061464    7.556634
------------------------------------------------------------------------------
```

According to the economic theory, the expected signs of LITR and RLFM should be positive, but the regression results suggests that signs are negative. Additionally, the R-squared value of the regression is 0.7009 which indicates that the model has a good fit and has successfully explained 70% of the variation in dependent variables due to the variation in explanatory variables. The model is also jointly significant as we can reject the null hypothesis (slope coefficients are insignificant) at 95% confidence. The value of Root MSE (standard error) is 0.92328 and that of F statistic is 14.65 with a degree of freedom (4, 25)

Model 2



```
reg lgsdp litr rlfm refm lfp rfmp lpeed sqpop

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(7, 22)        =     17.11
       Model |  60.1928728         7  8.59898183   Prob > F        =    0.0000
    Residual |  11.0572085        22  .502600385   R-squared       =    0.8448
-------------+----------------------------------   Adj R-squared   =    0.7954
       Total |  71.2500813        29  2.45689935   Root MSE        =    .70894

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0476075   .0188602    -2.52   0.019    -.0867212   -.0084938
        rlfm |  -2.061738   2.230192    -0.92   0.365    -6.686873    2.563398
        refm |   10.06507   1.752902     5.74   0.000     6.429771    13.70036
         lfp |  -.0076988   .0033942    -2.27   0.033    -.0147379   -.0006597
        rfmp |   6.012287   1.369334     4.39   0.000     3.172462    8.852113
       lpeed |   .6487992   .7371554     0.88   0.388    -.8799675    2.177566
       sqpop |  -.4568256   .1664188    -2.75   0.012    -.8019571   -.1116941
       _cons |   8.999112   2.915022     3.09   0.005     2.953727     15.0445
------------------------------------------------------------------------------
```

According to the economic theory, the expected signs of LITR, RLFM and LFP should be positive, but the regression results suggests that signs are negative. Additionally, the R-squared value of the regression is 0.8448 which indicates that the model has a good fit and has successfully explained 84.48% of the variation in dependent variables due to the variation in explanatory variables. The model is also jointly significant as we can reject the null hypothesis (slope coefficient are insignificant) at 95% confidence. The value of Root MSE (standard error) is 0.70894 and that of F statistic is 17.11 with a degree if freedom (7, 22).

Before interpreting the regression results, we have to confirm that the regression models satisfy all the CLRM assumptions.

ASSUMPTIONS

In CLRM, there are a set of assumptions that the model must satisfy in order to obtain the best linear unbiased estimator of the coefficients given by the ordinary least squares estimator. These assumptions are as follows:

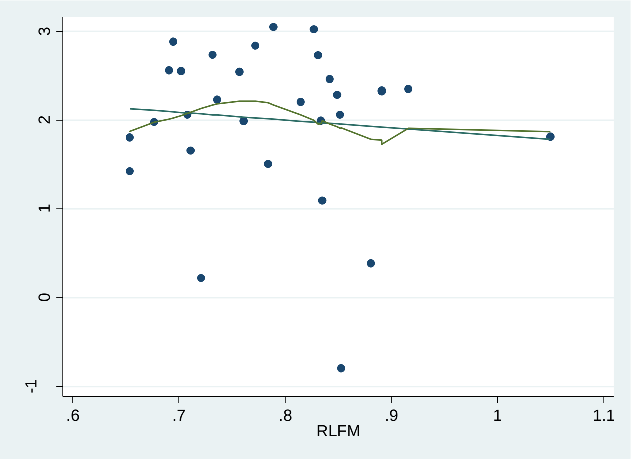
1. The relationship between dependent and independent variables are linear in the parameters.
2. Error terms has normal distribution with mean zero and constant variance. The assumption of constant variance across observation is termed as homoscedasticity assumption.
3. Error terms are not correlated with each other, which implies that covariance is zero. This is essentially the assumption of no autocorrelation.
4. Independent variables are not perfectly correlated with each other, i.e. there is no multicollinearity.
5. There is no specification error.
6. There is no measurement error.

We test both the models to see whether the assumptions are satisfied. Since our analysis is based on cross sectional analysis rather than time series analysis, therefore we can conclude that both the models are free from autocorrelation.

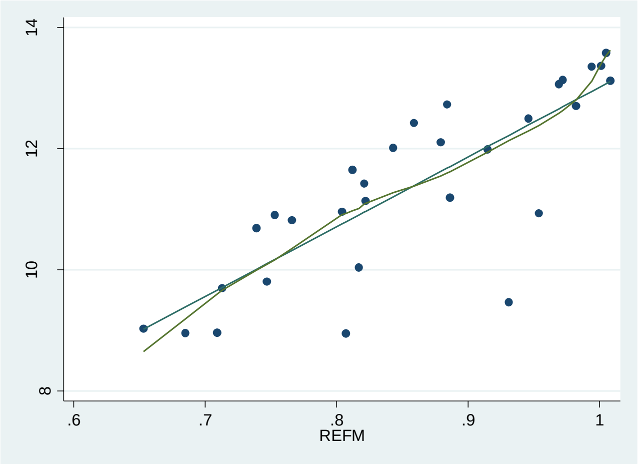
Model 1:

Testing Linearity

For testing linearity we plot acprplots.



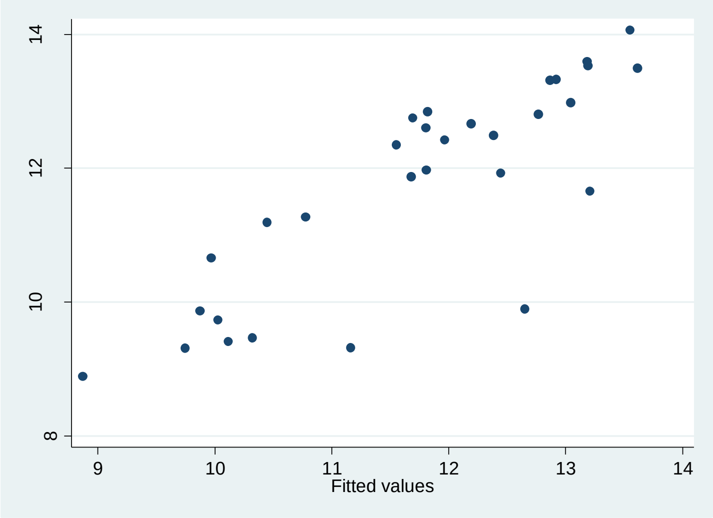
ACPR Plot for RLFM



ACPR Plot for REFM

The plots shows that the relationship between the log of GSDP and the primary explanatory variables, RLFM has a polynomial relation while REFM have an almost linear relationship with some non-linearity. However, LPEED is not linear. (Graph in Appendix)

The goodness of the model depends on how well it predicts dependent variables. We check this by predicting the dependent variable values and then plotting it against the dependent variable. Here, if the model is good then we expect the scatter plot to have 45 degrees pattern in the data.

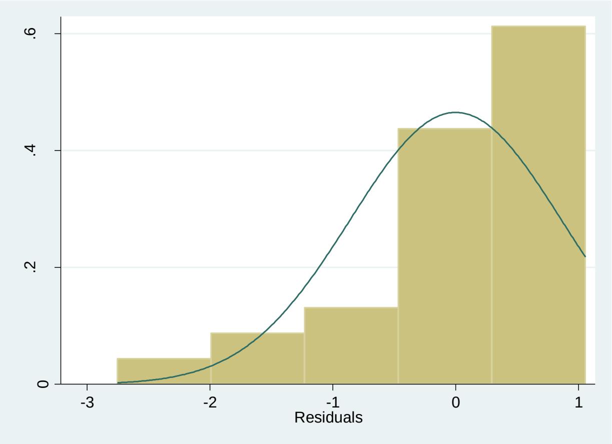


Correlation between dependent variable and fitted values

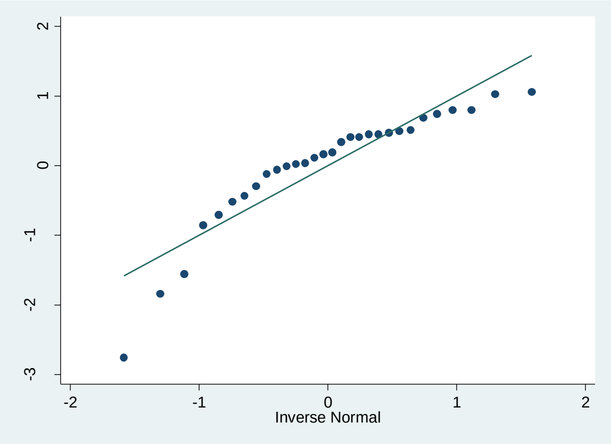
The graph between lgsdp and lgsdp_hat (predicted values) shows a 45 degrees pattern, which elucidates that the good fit of the model. Therefore, we can take that the linearity assumption is somewhat satisfied.

Testing normality of the residuals

After predicting the residual values, we can use either graphical method or formal test for checking the normality of the residuals. For graphical method, we either plot the residuals on a histogram or we can plot the Q-Norm plots. If the normality assumption is not satisfied, then the results might be misleading.

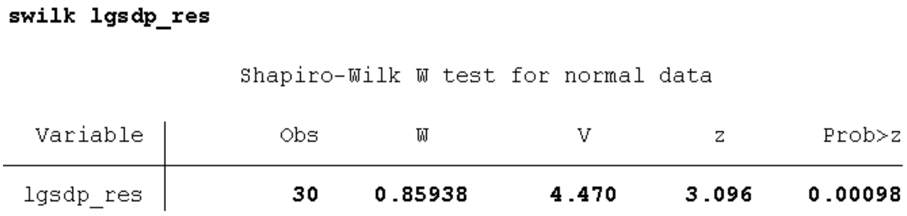


Histogram-Normal Plot for Residuals



Q-Norm Plot

These two graphs suggest that the residuals are not normally distributed. We can also run a formal test named Shapiro-Wilk test, where the null hypothesis is that the distribution of the residuals is normal. The result from the test also suggests that we can reject the null hypothesis for 5% significance level because the p value is 0.00098.



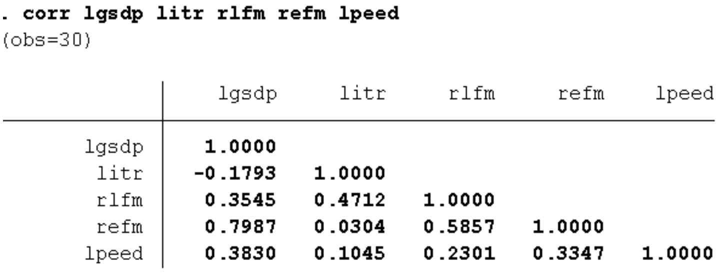
```
swilk lgsdp_res

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
   lgsdp_res |         30    0.85938      4.470     3.096    0.00098
```

Since the residuals do not follow a normal pattern, therefore we have to check for omitted variables and model specification. Although we assume the normality of residuals so that the t values and hypothesis testing is accurate, but normality assumption is not assumed to be a very strict assumption, therefore this non-normality will not affect our results significantly. The reason for the non-normality might be the low number of observations in our data set.

Testing for Multicollinearity

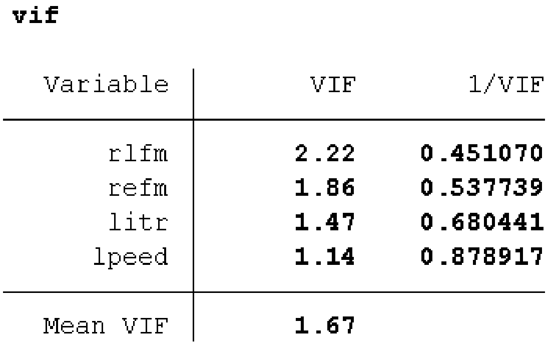
Multicollinearity essentially means that explanatory variables are highly correlated to one other. Although in the presence of multicollinearity the OLS estimators remain unbiased, but the variance is no longer minimum. Also due to this large variance, t values are incorrect due to which the hypothesis testing is often fallacious too, the R-squared value is also higher and the confidence intervals are wider. Additionally, the signs of beta values are also incorrect and the OLS estimators become sensitive towards small changes in the data. If the model suffers from multicollinearity then we can remove the variables with high correlation as a remedial measure of multicollinearity.

For testing multicollinearity, we can check the correlation between dependent and independent variables, or we can check the variance inflation factor (VIF). If the VIF is larger than 10, then it indicates that there is the presence of multicollinearity.



```
. corr lgsdp litr rlfm refm lpeed
(obs=30)

             |   lgsdp     litr     rlfm     refm    lpeed
-------------+---------------------------------------------
       lgsdp |  1.0000
        litr | -0.1793   1.0000
        rlfm |  0.3545   0.4712   1.0000
        refm |  0.7987   0.0304   0.5857   1.0000
       lpeed |  0.3830   0.1045   0.2301   0.3347   1.0000
```

The table shows that there is no indication of perfect correlation between the variables. This is also confirmed by the VIF, as it is less than 10 for every variable. Therefore, we conclude that our model does not suffer from multicollinearity.



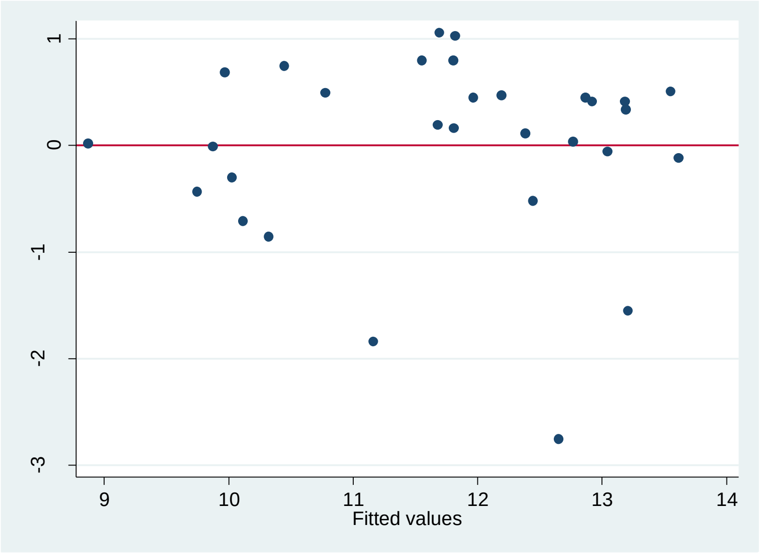
```
vif

    Variable |       VIF       1/VIF  
-------------+----------------------
        rlfm |      2.22    0.451070
        refm |      1.86    0.537739
        litr |      1.47    0.680441
       lpeed |      1.14    0.878917
-------------+----------------------
    Mean VIF |      1.67
```

Testing for Heteroscedasticity

When the error terms do not have constant variance, then the model suffers from the problem of heteroscedasticity. Due to heteroscedasticity, the estimators remain linear but are no longer unbiased where we cannot tell the direction of bias, also they do not have minimum variance. In addition to this, t values are incorrect and therefore the results from hypothesis testing are fallacious.

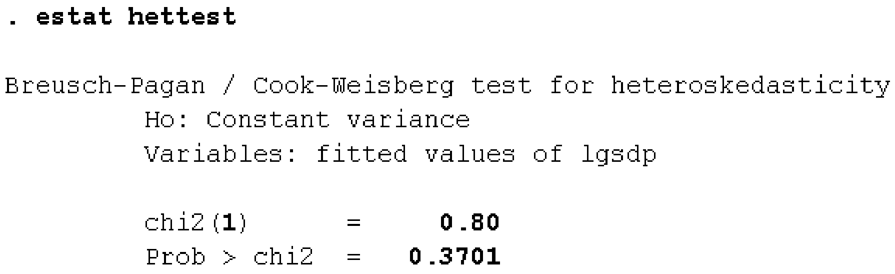
For testing for the presence of heteroscedasticity we can use either the graphical method or we can run formal tests. In the graphical method, we plot the residuals against the predicted values to get the pattern. If the scatter plot does not show any pattern, then our model is free from heteroscedasticity. These plots are obtained simply by using the command ‘rvfplot’ in Stata.

The rvfplot scatter plot shows that there is no pattern and therefore we can say that the model is free from heteroscedasticity.



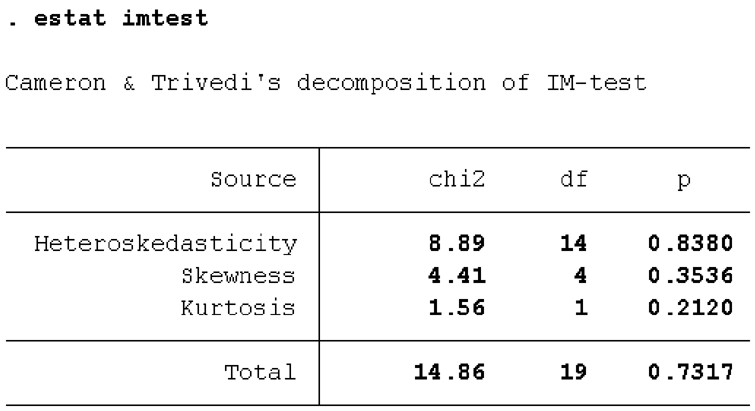
RVF Scatter Plot

The more formal methods are the Breusch-Pagan test and Cameron and Trivedi’s decomposition of IM-test, which are obtained by the commands ‘estat hettest’ and ‘estat imtest’ respectively, where the null hypothesis is that the variance of the constant variable is constant. The result from Breush-Pagan test suggests that we cannot reject the null hypothesis for the significance level 5% as the p value is 0.3701, with the chi-square value for 1 degree of freedom as 0.80.



```
. estat hettest

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

         chi2(1)      =     0.80
         Prob > chi2  =   0.3701
```

Similarly, in the Cameron and Trivedi’s decomposition test, we cannot reject the null hypothesis for 5% significance level as p value is 0.8380. We can thus conclude that the model is free from heteroscedasticity. Heteroscedasticity is treated by using Weight least squares or using robust regression.

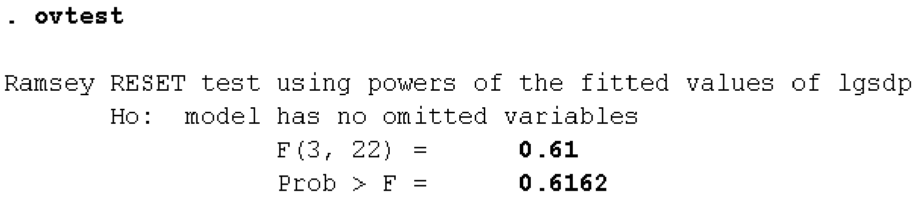


```
. estat imtest

Cameron & Trivedi's decomposition of IM-test

                   Source |       chi2     df      p
--------------------------+-----------------------------
       Heteroskedasticity |       8.89     14    0.8380
                 Skewness |       4.41      4    0.3536
                 Kurtosis |       1.56      1    0.2120
--------------------------+-----------------------------
                    Total |      14.86     19    0.7317
```

Testing for Omitted variable bias

When few variables are not added in the regression that must have been added, then the model suffers from omitted variable bias. In such bias, the estimator is biased while its variance is lower than the model where the other variables were also added. This trade-off between biased estimator and lower variance often confuses the researchers regarding adding the omitted variable in the regression or not. Testing for omitted variable bias is important since it is related to the assumption that the error term and the independent variables in the model are not correlated (E(e|X) = 0).

For testing omitted variable bias, we use the Ramsey RESET test, where the null hypothesis is that there is no omitted variable bias. The result suggests that at 5% significance level, we cannot reject the null hypothesis because the p value is 0.6162. Thus, we conclude that our model does not suffer from omitted variable bias.

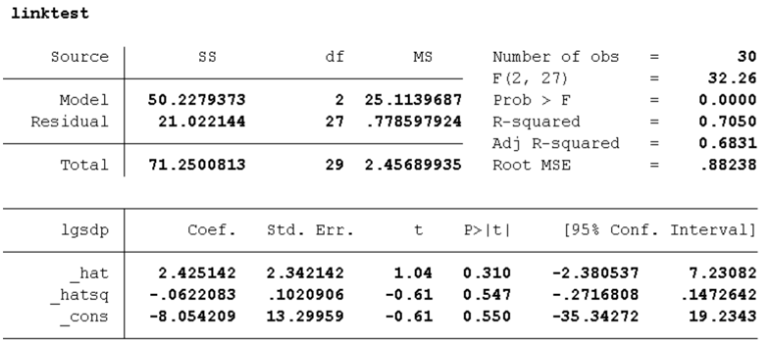


```
. ovtest

Ramsey RESET test using powers of the fitted values of lgsdp
       Ho:  model has no omitted variables
                  F(3, 22) =      0.61
                  Prob > F =      0.6162
```

Testing for specification error

When we use wrong forms of functional forms, omission of a relevant variable, inclusion of irrelevant variable and non-linearity then the model suffers from specification error. As a consequence of specification error, the estimators are not BLUE (best linear unbiased estimator), t values are insignificant and hypothesis testing is wrong.

For testing the specification in the model, we regress dependent variable (Y) against predicted values (Y-hat) and the square of predicted variables and then we look for the signification of the square of predicted values. If the Y-hat square is significant given the null hypothesis is that Y-hat square is insignificant, then we say that the model suffers from specification error. In Stata we simply use the command ‘linktest’ for running the above regression with a null hypothesis of no specification error in the model. The result form linktest shows that the Y-hat square value is insignificant at 5%significance level because we cannot reject the null hypothesis (Y-hat square=0) as p value is 0.547. We conclude that the model does not suffer from specification error.



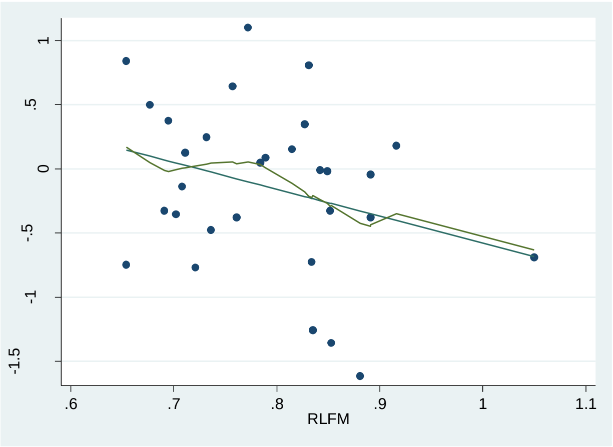
```
linktest

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(2, 27)        =     32.26
       Model |  50.2279373         2  25.1139687   Prob > F        =    0.0000
    Residual |   21.022144        27  .778597924   R-squared       =    0.7050
-------------+----------------------------------   Adj R-squared   =    0.6831
       Total |  71.2500813        29  2.45689935   Root MSE        =    .88238

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   2.425142   2.342142     1.04   0.310    -2.380537     7.23082
      _hatsq |  -.0622083   .1020906    -0.61   0.547    -.2716808    .1472642
       _cons |  -8.054209   13.29959    -0.61   0.550    -35.34272     19.2343
------------------------------------------------------------------------------
```

In conclusion, Model 1 satisfies all the assumptions except the normality assumption of the residuals.

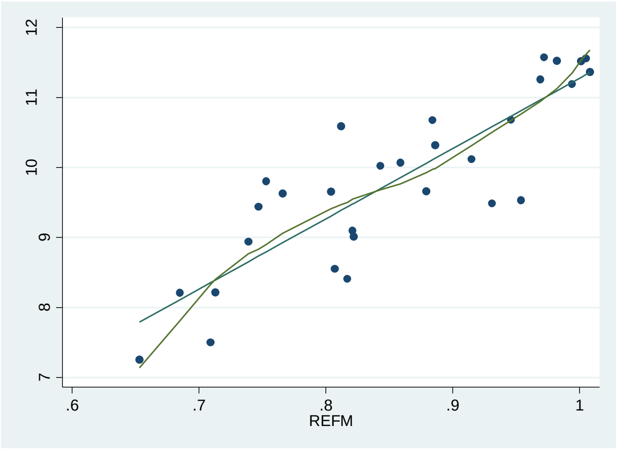
Model 2:

Testing Linearity

The plots shows that the relationship between the log of GSDP and the primary explanatory variables, RLFM and REFM have an almost linear relationship with slight non linearity.



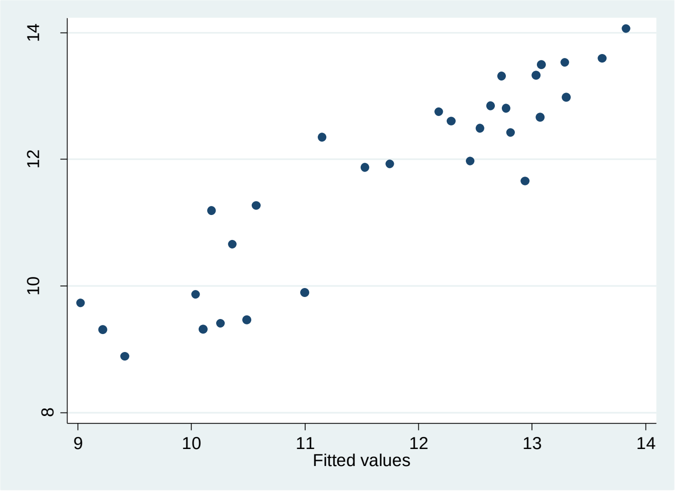
ACPR Plot for RLFM



ACPR Plot for REFM

However, some of the variables are not linear in terms of the parameters even after best possible transformation, which can be due to the low number of observations in our model.

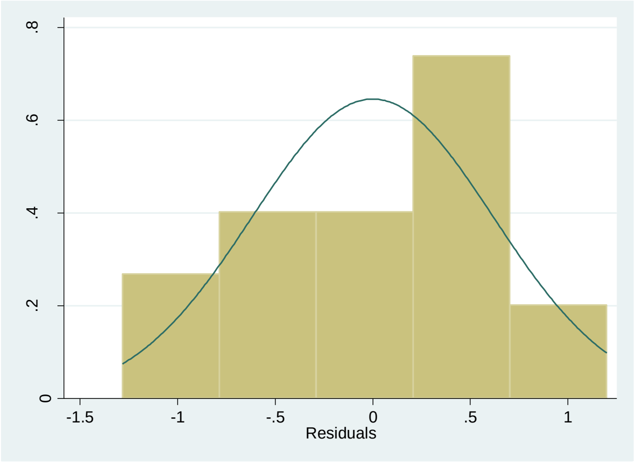
The graph between lgsdp and lgsdp_hat (predicted values) shows a 45 degrees pattern, which elucidates that the good fit of the model.

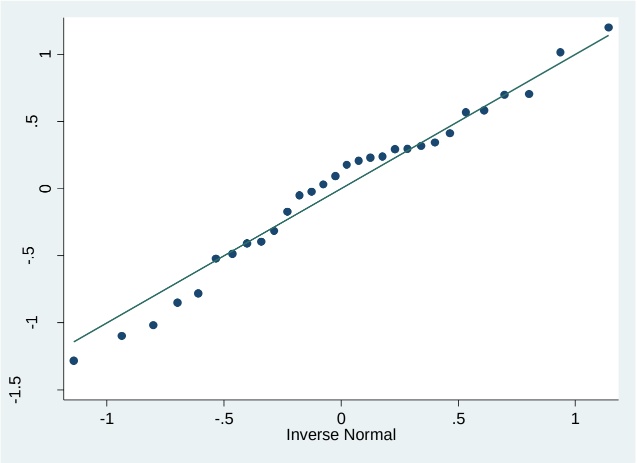


Correlation between dependent variable and fitted values

Testing normality of the residuals

Following are the graphs for histogram and the Q-Norm plots, which depicts that the residuals are normally distributed.

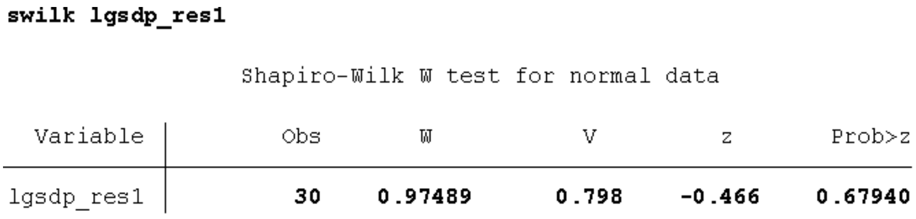




Histogram-Normal Plot for Residuals

Q-Norm Plot

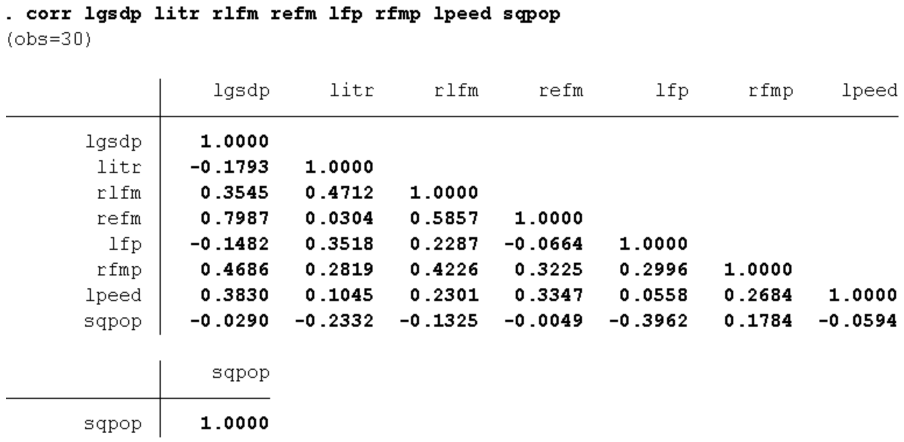
Similarly, the result from the Shapiro-Wilk test also suggests that we cannot reject the null hypothesis (normality of the residuals) for 5% significance level because the p value is 0.67940. Therefore, we can conclude that the normality assumption is satisfied.



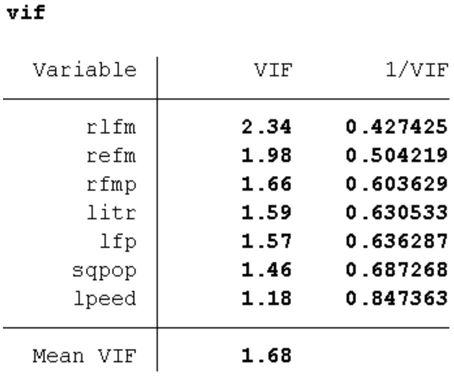
```
swilk lgsdp_res1

                  Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
  lgsdp_res1 |         30    0.97489      0.798    -0.466    0.67940
```

Testing for Multicollinearity

The table showing correlation between variables that there is no indication of perfect correlation. This is also confirmed by the VIF, as it is less than 10 for every variable. Therefore, we conclude that our model does not suffer from multicollinearity.



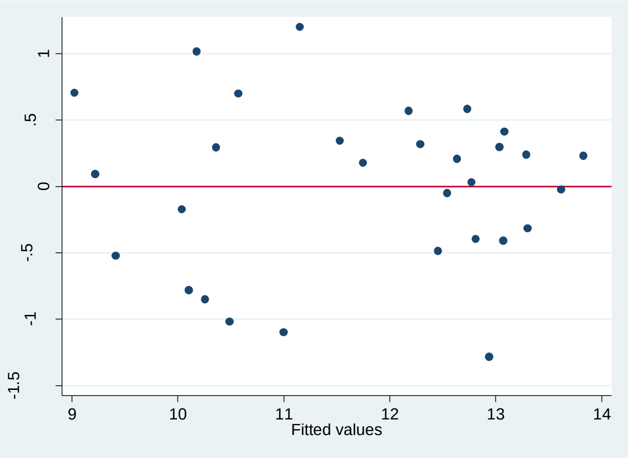
```
. corr lgsdp litr rlfm refm lfp rfmp lpeed sqpop
(obs=30)

             |    lgsdp     litr     rlfm     refm      lfp     rfmp    lpeed
-------------+---------------------------------------------------------------
       lgsdp |   1.0000
        litr |  -0.1793   1.0000
        rlfm |   0.3545   0.4712   1.0000
        refm |   0.7987   0.0304   0.5857   1.0000
         lfp |  -0.1482   0.3518   0.2287  -0.0664   1.0000
        rfmp |   0.4686   0.2819   0.4226   0.3225   0.2996   1.0000
       lpeed |   0.3830   0.1045   0.2301   0.3347   0.0558   0.2684   1.0000
       sqpop |  -0.0290  -0.2332  -0.1325  -0.0049  -0.3962   0.1784  -0.0594

             |    sqpop
-------------+---------
       sqpop |   1.0000
```



```
vif

    Variable |       VIF       1/VIF
-------------+----------------------
        rlfm |      2.34    0.427425
        refm |      1.98    0.504219
        rfmp |      1.66    0.603629
        litr |      1.59    0.630533
         lfp |      1.57    0.636287
       sqpop |      1.46    0.687268
       lpeed |      1.18    0.847363
-------------+----------------------
    Mean VIF |      1.68
```

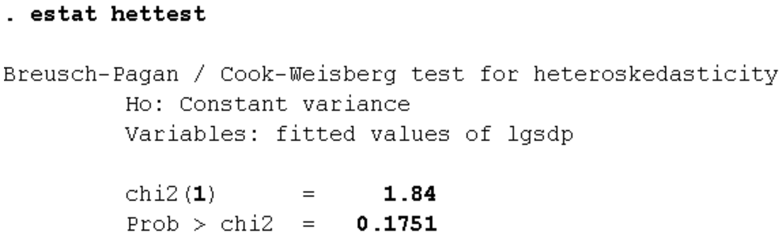
Testing for Heteroscedasticity

The rvfplot scatter plot shows that there is no pattern and therefore we can say that the model is free from heteroscedasticity.



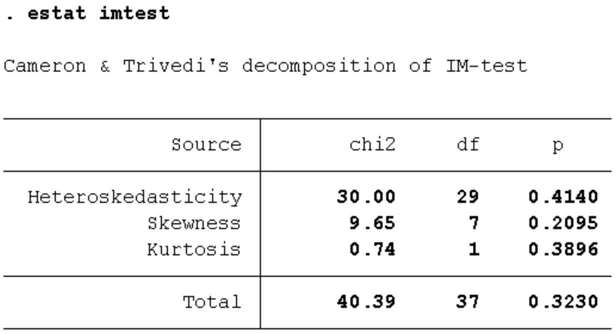
RVF Scatter Plot

The result from Breush-Pagan test suggests that we cannot reject the null hypothesis for the significance level 5% as the p value is 0.1751, with the chi-square value for 1 degree of freedom as 1.84.



```
. estat hettest

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

         chi2(1)      =     1.84
         Prob > chi2  =   0.1751
```

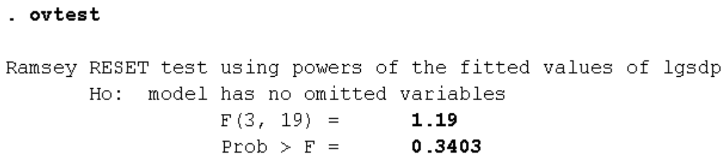
Similarly, in the Cameron and Trivedi's decomposition test, we cannot reject the null hypothesis for 5% significance level as p value is 0.4140. We can thus conclude that the model is free from heteroscedasticity.



```
. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      30.00     29    0.4140
            Skewness |       9.65      7    0.2095
            Kurtosis |       0.74      1    0.3896
---------------------+-----------------------------
               Total |      40.39     37    0.3230
---------------------------------------------------
```

Testing for Omitted variable bias

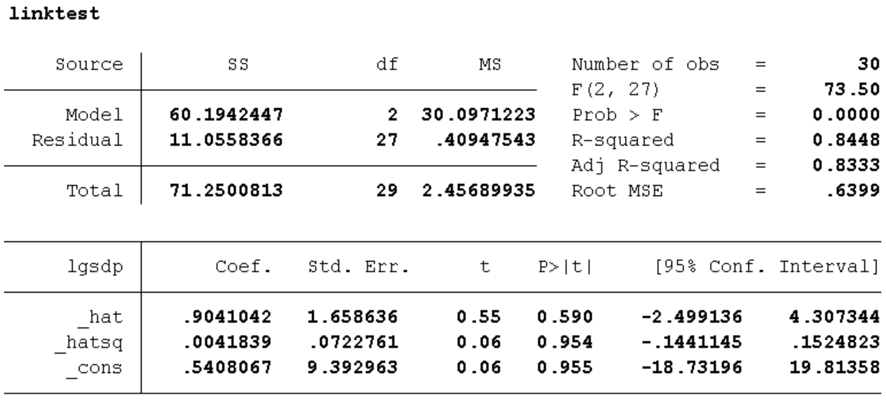
The result suggests that at 5% significance level, we cannot reject the null hypothesis because the p value is 0.3403. Thus, we conclude that our model does not suffer from omitted variable bias.



```
. ovtest

Ramsey RESET test using powers of the fitted values of lgsdp
       Ho:  model has no omitted variables
                 F(3, 19) =      1.19
                 Prob > F =      0.3403
```

Testing for specification error

The result form linktest shows that the Y-hat square value is insignificant at 5%significance level because we cannot reject the null hypothesis (Y-hat square=0) as p value is 0.954. We conclude that the model does not suffer from specification error.



```
linktest

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(2, 27)        =     73.50
       Model |  60.1942447         2  30.0971223   Prob > F        =    0.0000
    Residual |  11.0558366        27   .40947543   R-squared       =    0.8448
-------------+----------------------------------   Adj R-squared   =    0.8333
       Total |  71.2500813        29  2.45689935   Root MSE        =     .6399

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   .9041042   1.658636     0.55   0.590    -2.499136    4.307344
      _hatsq |   .0041839   .0722761     0.06   0.954    -.1441145    .1524823
       _cons |   .5408067   9.392963     0.06   0.955    -18.73196    19.81358
------------------------------------------------------------------------------
```

In conclusion Model2 satisfies all the CLRM assumptions. Since, model doesn’t suffer from any of the above mentioned problems, therefore there is no need for taking up the remedial measures. If the model was suffering from any of these problems then we could have used the robust form of regression for removing these problems. The robust regression form gives better results than the normal regression, but the basic result that there is a negative impact of gender inequality in education on the economic growth remains same. (Robust Regression in Appendix)

INTERPRETATION OF REGRESSION RESULTS

Model 1

In our first model we explain the relationship between the gender inequalities in education on the economic growth, without considering other possible variables that can affect the economic growth. For the multiple regression analysis, we first set up the hypothesis:

Null Hypothesis: β_0, β_1, β_2, β_3, β_4 =0

Alternate Hypothesis: β_0, β_1, β_2, β_3, β_4 $\neq$0

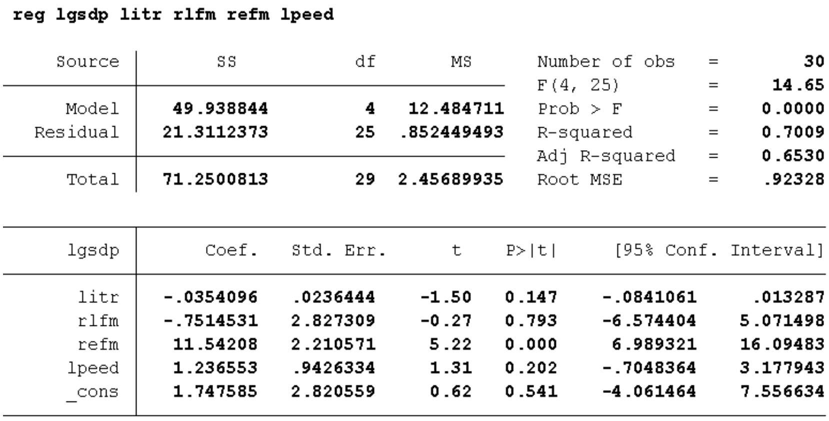
Here β_0 is the intercept term and β_1, β_2, β_3 and β_4 are slope terms for LITR, RLFM, REFM and log of PEED respectively.

It essentially means that, according to the null hypothesis the slope coefficients are insignificant i.e. there is no relationship between the independent variables and the dependent variables, while the alternative hypothesis states that there is some relationship between the dependent and the independent variables.

After running the regression, following results are obtained:

$$\text{PRF: } LGSDP = \beta_0 + \beta_1 LITR + \beta_2 RLFM + \beta_3 REFM + \beta_4 LPEED + \varepsilon$$

$$\text{SRF: } \widehat{LGSDP} = \beta_0 + \beta_1 \widehat{LITR} + \beta_2 \widehat{RLFM} + \beta_3 \widehat{REFM} + \beta_4 \widehat{LPEED} + \varepsilon$$



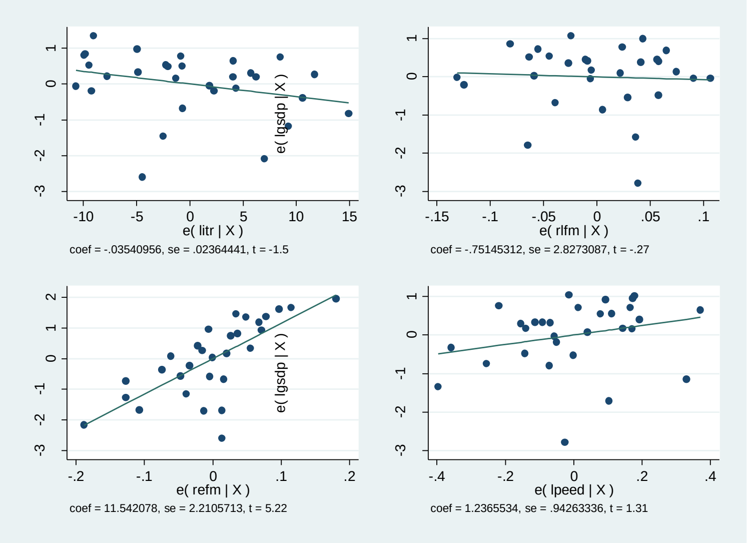
```
reg lgsdp litr rlfm refm lpeed

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(4, 25)        =     14.65
       Model |  49.938844          4  12.484711    Prob > F        =    0.0000
    Residual |  21.3112373        25  .852449493   R-squared       =    0.7009
-------------+----------------------------------   Adj R-squared   =    0.6530
       Total |  71.2500813        29  2.45689935   Root MSE        =    .92328

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0354096   .0236444    -1.50   0.147    -.0841061     .013287
        rlfm |  -.7514531   2.827309    -0.27   0.793    -6.574404    5.071498
        refm |   11.54208   2.210571     5.22   0.000     6.989321    16.09483
       lpeed |   1.236553   .9426334     1.31   0.202    -.7048364    3.177943
       _cons |   1.747585   2.820559     0.62   0.541    -4.061464    7.556634
------------------------------------------------------------------------------
```

The coefficients are interpreted as follows:

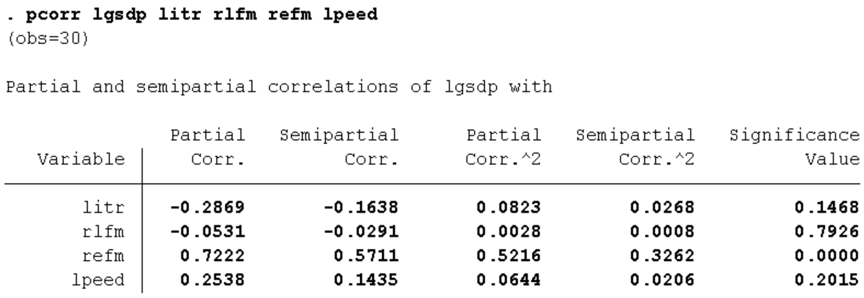
- β_0: When the explanatory variables are 0 then, on an average the log of GSDP is Rs. 1.7476.
- β_1: When the literacy rate increases by 1%, on an average the log of GSDP decreases by Rs. 0.354, keeping all the other variables as constant. This result is wrong because economic theory suggests that with an increase in the literacy rate, the economic growth increases rather than decrease because the level of human capital increases.
- β_2: When the RLFM increases by 1%, on an average the log of GSDP decreases by Rs. 0.7515, keeping all the other variables as constant. This result is wrong because economic theory suggests that with an increase in the gender equality in literacy rate, the economic growth increases rather than decrease because the level of human capital increases.
- β_3: When the REFM increases by 1%, on an average the log of GSDP increases by Rs. 11.542, keeping all the other variables as constant. This result essentially means that with the increase in the gender equality, the economic growth increases.

- β_4: When the log of PEED increases by 1%, on an average the log of GSDP increases by 1.237%, keeping all the other variables as constant. β_4 can also be interpreted as the partial elasticity coefficient of GSDP with respect to PEED, keeping all other variables as constant.

The following result shows that the coefficient β_3 is significant for 5% (α=0.05) significance level. Apart from this, the confidence interval at 95% confidence level lies above 0 indicating that β_3 is significant. The slope coefficient β_1, β_2 and β_4 are insignificant at 5% level and the confidence interval includes the value 0 for 95% confidence level which indicates that β_1, β_2 and β_4 takes the value 0 too. Here the intercept term is also insignificant. Similar results are also given by AVR Plots (which captures the partial influence of the variables) and the partial correlation, where the only significant value is of REFM.



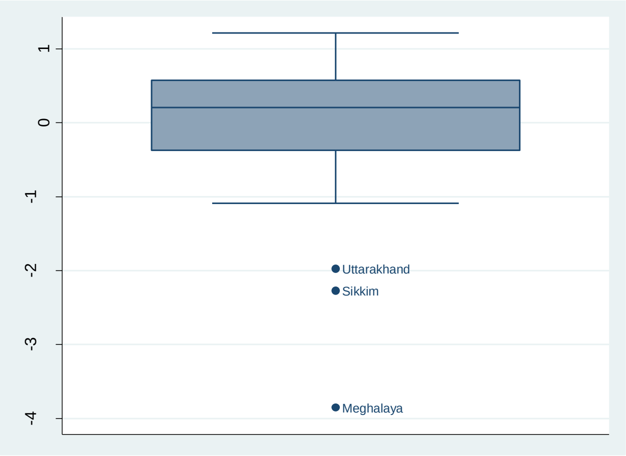
AVR Plots


```
. pcorr lgsdp litr rlfm refm lpeed
(obs=30)

Partial and semipartial correlations of lgsdp with

               Partial   Semipartial      Partial   Semipartial   Significance
   Variable |    Corr.         Corr.      Corr.^2       Corr.^2          Value
------------+-----------------------------------------------------------------
       litr |  -0.2869       -0.1638       0.0823        0.0268         0.1468
       rlfm |  -0.0531       -0.0291       0.0028        0.0008         0.7926
       refm |   0.7222        0.5711       0.5216        0.3262         0.0000
      lpeed |   0.2538        0.1435       0.0644        0.0206         0.2015
```

We conclude that although the economic growth does not get affected by the female to male literacy rates, but it gets affected by the female to male enrolment ratios, indicating that our assertion that gender inequality in education affects the economic growth negatively is partially true.

We further undertake the diagnostic tests to show whether the regression line obtained is influenced by outliers, leverage points and influential points.

Outliers: An outlier in a regression is a data point which has a large residual. Large in this context does not refer to the absolute size of a residual but to its size relative to most of the other residuals in the regression. Outliers are obtained by the studentized residual. We fit this residual and graph box-plot for the outliers. We observe that there are 3 outlier in the box-plot i.e. Uttrakhand, Sikkim and Mizoram.



Studentized residual Box Plot

Influential point: A data point is influential if removing it from the sample would markedly change the position of the least squares regression line (Moore and McCabe, 1989: 185).

We first fit dfbeta and then use rule of thumb for calculating the influential point. For points where dfbeta value is greater than 2, are considered to have high influence. While for points where dfbeta value is greater than $2/\sqrt{n}$, are considered to have low influence. There are no points having very high influence on the regression line while there is the presence of low influential points. We observe that Meghalya is a low influential point for the dfbeta values of LITR, Sikkim is a low influential point for the dfbeta values of RLFM and Manipur is a low influential point for the dfbeta values of Log of PEED.

Leverage: A data point is influential if removing it from the sample would markedly change the position of the least squares regression line (Moore and McCabe, 1989: 185). We predict the dfits value and compare it with the threshold value, which is $\sqrt{\frac{K}{N}}$. The comparison shows that there are no points with substantial Leverage.

Model 2

In our second model we explain the relationship between the gender inequalities in education on the economic growth, considering other possible variables (Employment) that can affect the economic growth. For the multiple regression analysis, we first set up the hypothesis:

Null Hypothesis: β_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7 =0

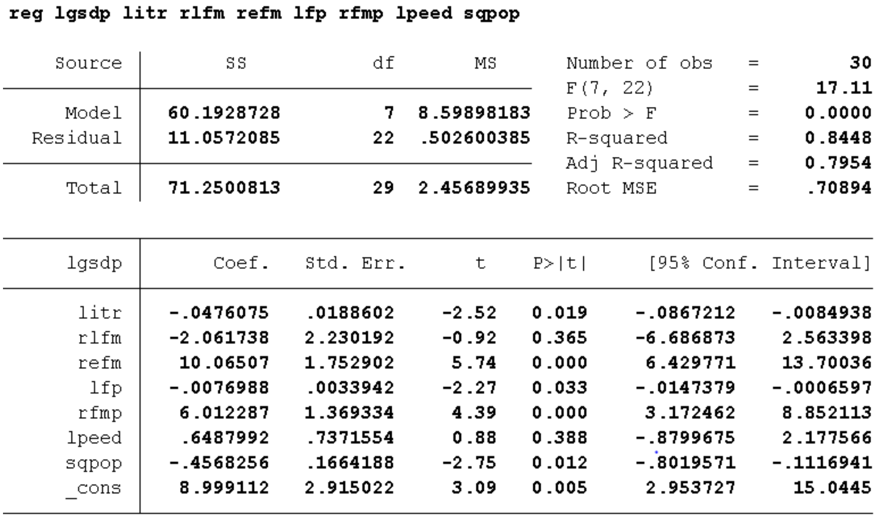
Alternate Hypothesis: β_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7 ≠0

Here β_0 is the intercept term and β_1, β_2, β_3, β_4, β_5, β_6, β_7 are slope terms for LITR, RLFM, REFM, LFP, RFMP, log of PEED and square-root of POPG respectively.

After running the regression, following results are obtained:

PRF: $LGSDP = \beta_0 + \beta_1 LITR + \beta_2 RLFM + \beta_3 REFM + \beta_4 LPF + \beta_5 RFMP + \beta_6 LPEED + \beta_7 SQPOP + \varepsilon$

SRF:$\widehat{LGSDP} = \beta_0 + \beta_1\widehat{LITR} + \beta_2\widehat{RLFM} + \beta_3\widehat{REFM} + \beta_4\widehat{LPF} + \beta_5\widehat{RFMP} + \beta_6\widehat{LPEED} + \beta_7\widehat{SQPOP} + \varepsilon$



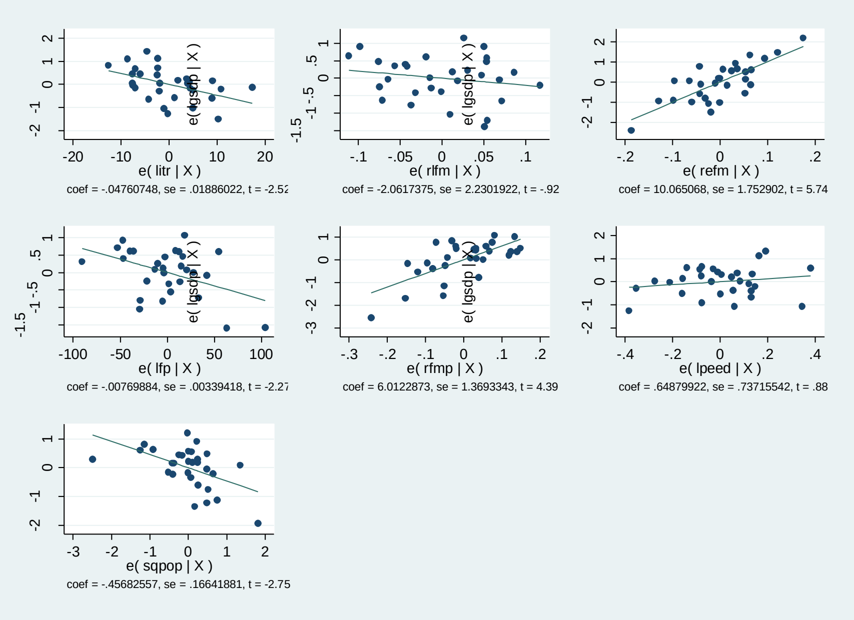
```
reg lgsdp litr rlfm refm lfp rfmp lpeed sqpop

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(7, 22)        =     17.11
       Model |  60.1928728         7  8.59898183   Prob > F        =    0.0000
    Residual |  11.0572085        22  .502600385   R-squared       =    0.8448
-------------+----------------------------------   Adj R-squared   =    0.7954
       Total |  71.2500813        29  2.45689935   Root MSE        =    .70894

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0476075   .0188602    -2.52   0.019    -.0867212   -.0084938
        rlfm |  -2.061738   2.230192    -0.92   0.365    -6.686873    2.563398
        refm |   10.06507   1.752902     5.74   0.000     6.429771    13.70036
         lfp |  -.0076988   .0033942    -2.27   0.033    -.0147379   -.0006597
        rfmp |   6.012287   1.369334     4.39   0.000     3.172462    8.852113
       lpeed |   .6487992   .7371554     0.88   0.388    -.8799675    2.177566
       sqpop |  -.4568256   .1664188    -2.75   0.012    -.8019571   -.1116941
       _cons |   8.999112   2.915022     3.09   0.005     2.953727     15.0445
------------------------------------------------------------------------------
```

The coefficients are interpreted as follows:

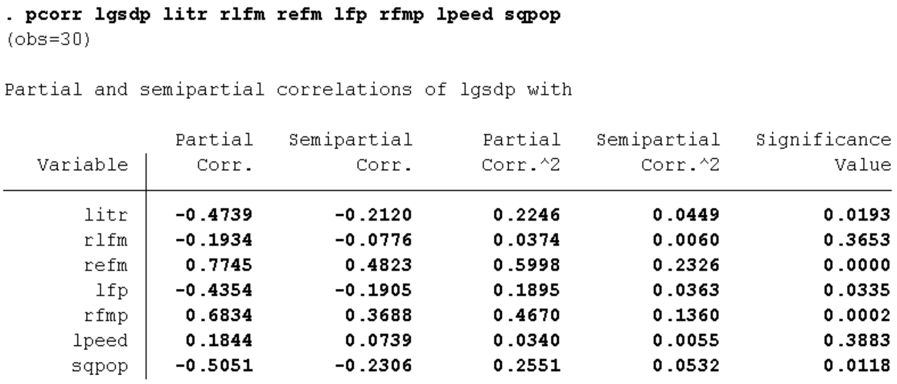
- β_0: When the explanatory variables are 0 then, on an average the log of GSDP is Rs. 8.999.
- β_1: When the literacy rate increases by 1%, on an average the log of GSDP decreases by Rs. 0.0476, keeping all the other variables as constant. This result is wrong because economic theory suggests that with an increase in the literacy rate, the economic growth increases rather than decrease because the level of human capital increases.
- β_2: When the RLFM increases by 1%, on an average the log of GSDP decreases by Rs. 2.0617, keeping all the other variables as constant. This result is wrong because economic theory suggests that with an increase in the gender equality in literacy rate, the economic growth increases rather than decrease because the level of human capital increases.

- β_3: When the REFM increases by 1%, on an average the log of GSDP increases by Rs. 10.065, keeping all the other variables as constant. This result essentially means that with the increase in the gender equality, the economic growth increases.
- β_4: When the LFP increases by 1%, on an average the log of GSDP decreases by Rs. 0.0077, keeping all the other variables as constant. This result is wrong because economic theory suggests that with an increase in the labour force participation, the economic growth should increase as more number of people are working which generates more income.
- β_5: When the REFM increases by 1%, on an average the log of GSDP increases by Rs. 6.0123, keeping all the other variables as constant. This result essentially means that with the increase in the gender equality in labour force participation, the economic growth increases.
- β_6: When the log of PEED increases by 1%, on an average the log of GSDP increases by 0.6488%, keeping all the other variables as constant. β_6 can also be interpreted as the partial elasticity coefficient of GSDP with respect to PEED, keeping all other variables as constant.
- β_7: When the square-root of POPG increases by 1%, on an average the log of GSDP decreases by Rs. 0.4568 keeping all the other variables as constant.

The following result shows that the coefficient β_0, β_1, β_3, β_4, β_5, and β_7 is significant for 5% (α=0.05) significance level. Apart from this, the confidence interval at 95% confidence level lies above 0 indicating that β_3 is significant. The slope coefficient β_2, and β_6 are insignificant at 5% level and the confidence interval includes the value 0 for 95% confidence level which indicates that β_2 and β_6 takes the value 0 too. Here the intercept term is also insignificant. Similar results are also given by AVR Plots and the partial correlation.



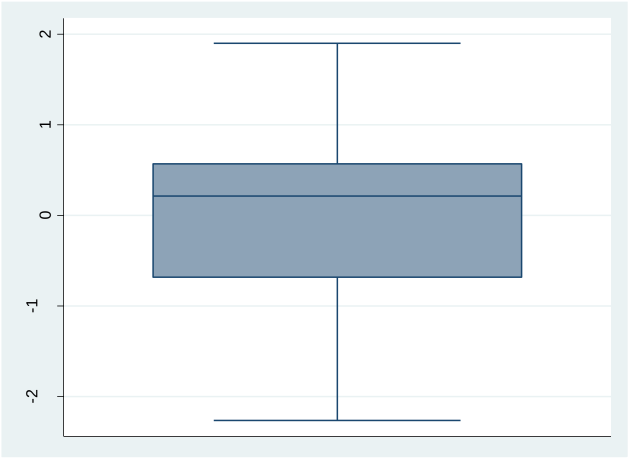
AVR Plots



```
. pcorr lgsdp litr rlfm refm lfp rfmp lpeed sqpop
(obs=30)

Partial and semipartial correlations of lgsdp with

               Partial   Semipartial      Partial   Semipartial   Significance
   Variable      Corr.         Corr.      Corr.^2       Corr.^2          Value

       litr    -0.4739       -0.2120       0.2246        0.0449         0.0193
       rlfm    -0.1934       -0.0776       0.0374        0.0060         0.3653
       refm     0.7745        0.4823       0.5998        0.2326         0.0000
        lfp    -0.4354       -0.1905       0.1895        0.0363         0.0335
       rfmp     0.6834        0.3688       0.4670        0.1360         0.0002
      lpeed     0.1844        0.0739       0.0340        0.0055         0.3883
      sqpop    -0.5051       -0.2306       0.2551        0.0532         0.0118
```

We further undertake the diagnostic tests to show whether the regression line obtained is influenced by outliers, leverage points and influential points.

Outliers: We fit this residual and graph box-plot for the outliers and observe that there are no outliers.



Studentized residual Box Plot

Influential Points: Dfbeta values show that Nagaland has a high influence for the dfbeta values of square-root of POPG, while there is the presence of low influential points. We observe that Nagaland and Puducherry are low influential points for the dfbeta values of LITR, Sikkim and Nagaland are low influential points for the dfbeta values of RLFM, Mizoram and Uttar Pradesh are low influential points for the dfbeta values of REFM, Himachal Pradesh and Uttrakhand are influential points for dfbeta values of LFP, Meghalaya and Sikkim are the low influential points for dfbeta values of RFMP Manipur and Odisha are low influential points for the dfbeta values of Log of PEED and Nagaland and Puducherry are the low influential points for dfbeta values of square-root of POPG.

Leverage: By predicting dfits value, we see that there are no points with substantial influence.

In conclusion, the models explains the relationship between the gender inequality in education and economic growth, however there is presence of outliers and influential points.

CONCLUSION

Gender Inequality in Education plays an important role in the development process as higher is the inequalities in a country, lower is the human capital and hence the realisation of development is not possible. In India, gender inequality in education is quite high which not only affect the human capital but also hinders economic growth.

After the exploratory and confirmatory analysis, the study has found that gender inequality in education significantly reduces economic growth, especially when the female to male ratio of enrolment in schools reduces, economic growth also reduces which clearly depicts the negative impact of gender inequality. Furthermore, models also depicts that there is a negative impact of gender inequality in labour force participation on economic growth, which further highlights the importance of women in the development process of any country. Therefore, of gender inequality reduces not only in education but also in other aspects like labour force participation, health etc. then the goals of sustainable development can be achieved.

In totality, both the models have explained the negative impact of gender inequality in education on economic growth. We also observed that Model 2 has given better results than Model 1, which is due to the increase in the number of variables in the second model. Furthermore, the residuals are non-normal for Model 1 but the problem is restored when we increased the number of variables.

Despite explaining the negative relation, the models has the presence of few influential points and outliers, some of the variables are insignificant and has opposite signs than expected, and all the variables do not have a linear relation in terms of parameters. This is due to the small number of observations in our analysis as the study is based on state level analysis.

Additionally, we have checked the regression results by lowering the number of observations, which illustrated that there is in-fact a very minimal change in the regression outputs and the basic result that there is a negative impact of gender inequality in education and labour force participation on the level of economic growth hold true. (Results in Appendix)

REFERENCES

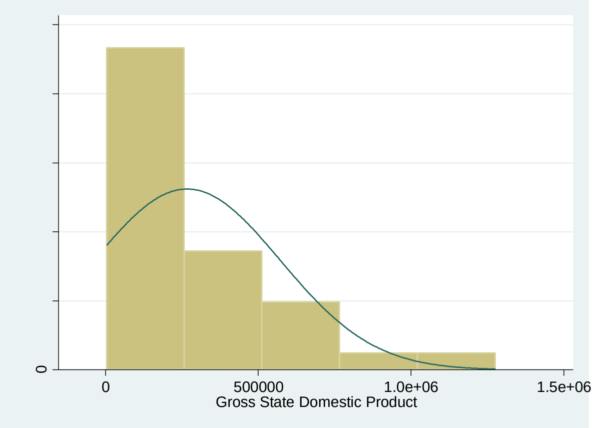
1. Reserve Bank of India: Statistical Handbook
   https://www.rbi.org.in/scripts/Publications.aspx
2. Census 2011
   https://www.census2011.co.in/
3. NITI Aayog
   http://niti.gov.in/state-statistics
4. Imran Sharif Chaudhry (2007) "

5. Gender Inequality in Education and Economic Growth: Case Study of Pakistan", Pakistan Horizon, Vol. 60, No. 4, Women's Concerns in International Relations (October 2007), pp. 81-91, Pakistan Institute of National Affairs.
6. Stephan Kalsen (1999) "Does Gender Inequality Reduce Growth and Development? Evidence from Cross-Country Regressions" Policy Research Report on Gender and Development Working Paper Series, No. 7. Development Research Group/PREMN, the World Bank.
7. Nayef Al-Shammari1, Monira Al Rakhis (2017) "Impact of Gender Inequality on Economic Growth in the Arab Region" Department of Economics, College of Business Administration, Kuwait University, Kuwait City, Kuwait.
8. Stephan Klasen and Francesca Lamanna (2009) "The Impact of Gender Inequality in Education and Employment on Economic Growth" Feminist Economics 15(3), July 2009, 91–132.

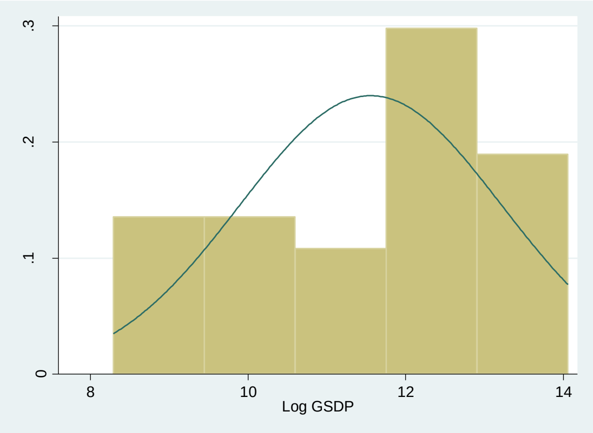
APPENDIX

1. We transform some of the variables to make them normal in order to get correct regression results.

   Gross State Domestic Product:

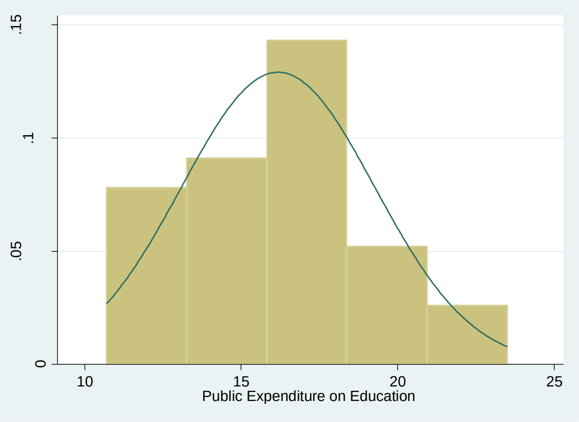


Before Transformation

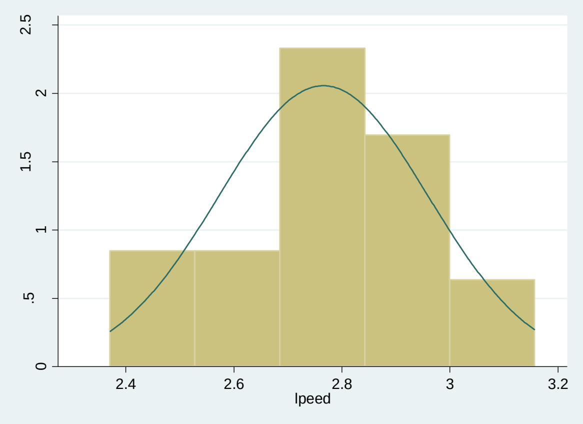


After Transformation

Public Expenditure on Education

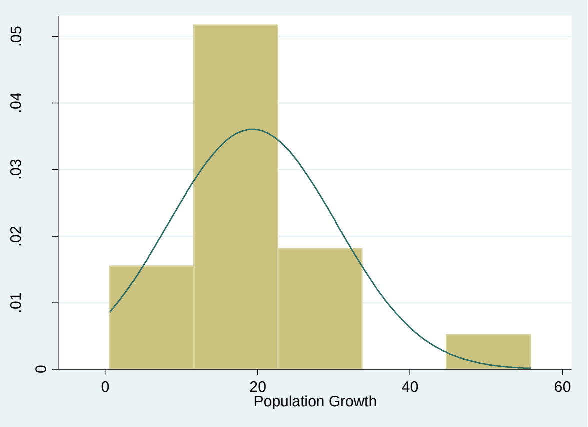


Before Transformation

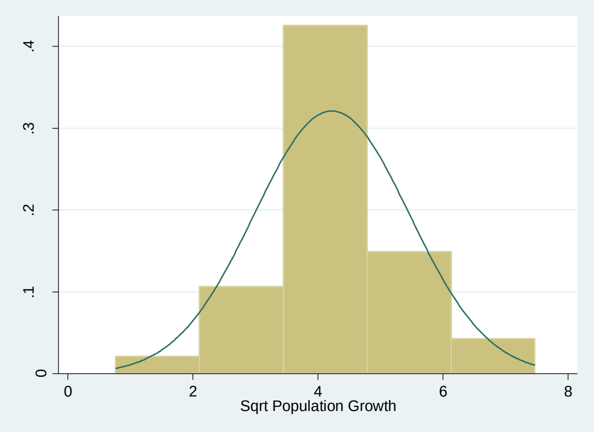


After Transformation

Population Growth

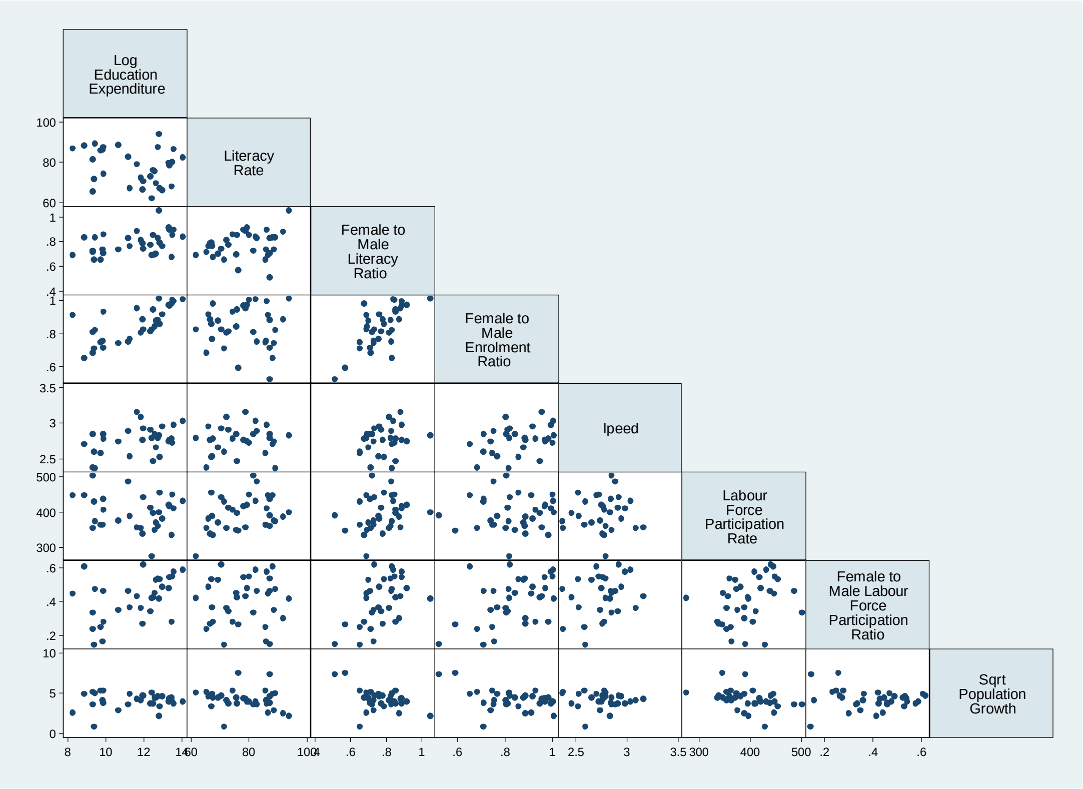


Before Transformation

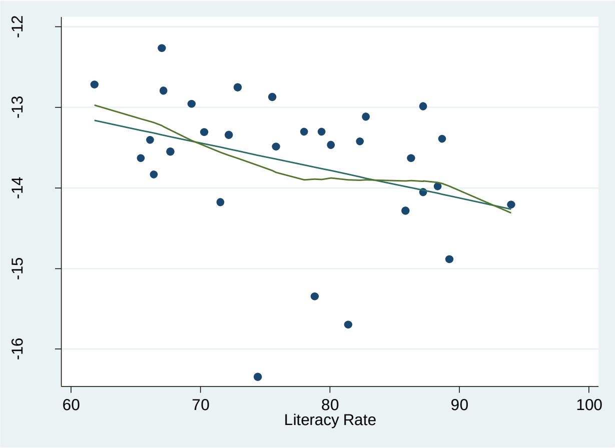


After Transformation

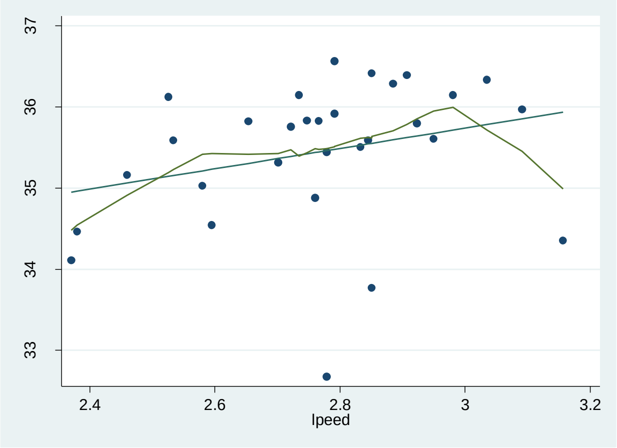
2. Correlation Graph Matrix



3. ACPR Plots for Model 1



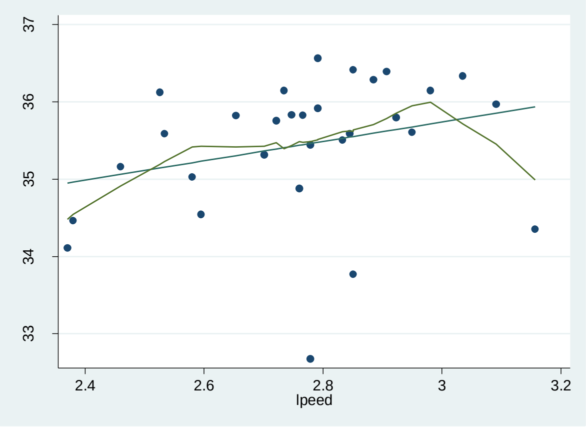
ACPR for LITR



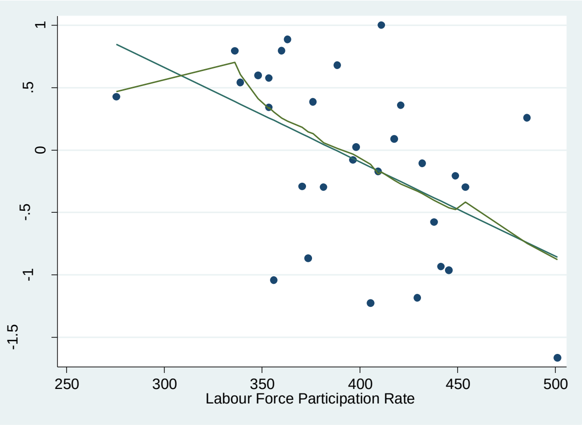
ACPR for LPEED

ACPR Plots suggest that while the relationship is linear with Literacy rate, log of GSDP does not hold a linear relationship with log of PEED. This might be due to the low number of observations in the data.

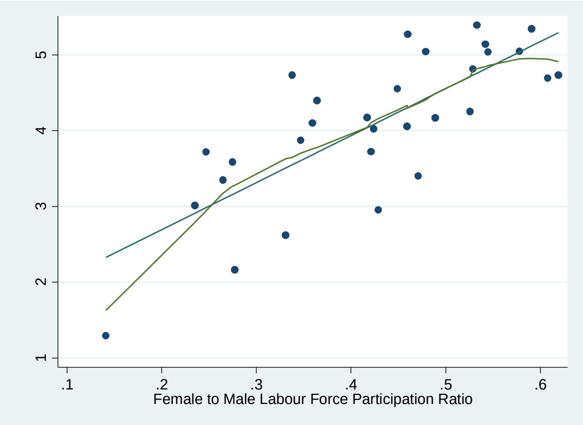
4. ACPR Plots for Model 2



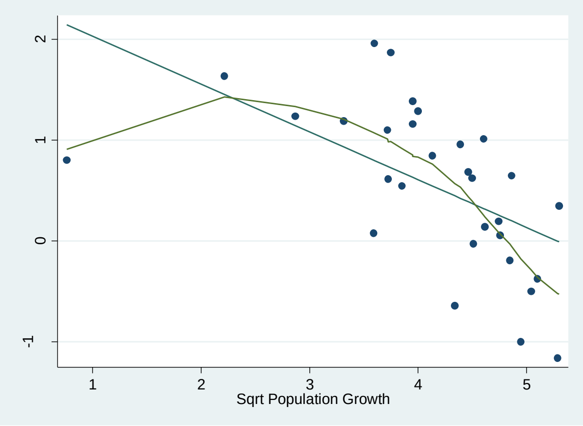
ACPR for LPEED



ACPR for LFP



ACPR for RFMP



ACPR for sqPOPG

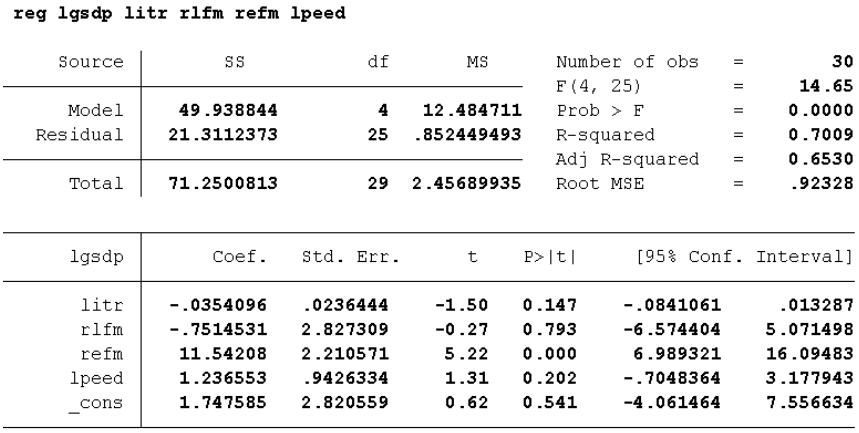
While few variables show an almost linear relationship with log of GSDP, variables like square-root of POPG and log of PEED show a quadratic relationship. This is due to the presence of outliers in the data set which pulls the ACPR into a quadratic relation. However, when we checked for the presence of outliers, influential points and leverage, there was no observation which had a very drastic influence on the regression line and there was no point which had high leverage. Additionally, this might be an effect of low number of observations in our data set.

5. Since the model suffers from the problem of low number of observations, we run regression with different number of observations and observe the changes that took place.

   In the current model, there are 30 observations, we further regress the variables with 26, 20, 16 and 11 observations.

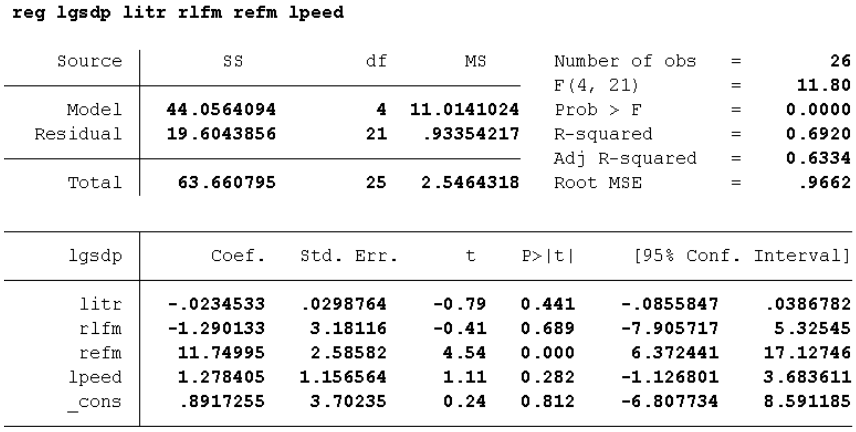
Model 1:

30 Observations:



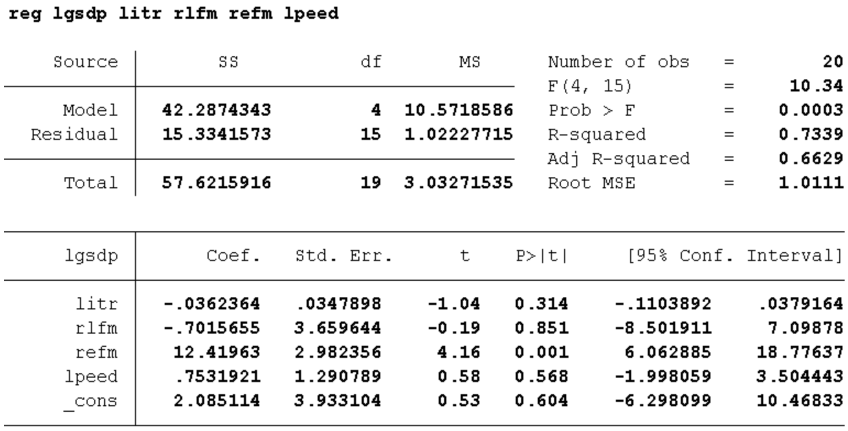
```
reg lgsdp litr rlfm refm lpeed

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(4, 25)        =     14.65
       Model |   49.938844         4   12.484711   Prob > F        =    0.0000
    Residual |  21.3112373        25  .852449493   R-squared       =    0.7009
-------------+----------------------------------   Adj R-squared   =    0.6530
       Total |  71.2500813        29  2.45689935   Root MSE        =    .92328

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0354096   .0236444    -1.50   0.147    -.0841061     .013287
        rlfm |  -.7514531   2.827309    -0.27   0.793    -6.574404    5.071498
        refm |   11.54208   2.210571     5.22   0.000     6.989321    16.09483
       lpeed |   1.236553   .9426334     1.31   0.202    -.7048364    3.177943
       _cons |   1.747585   2.820559     0.62   0.541    -4.061464    7.556634
------------------------------------------------------------------------------
```

26 Observations:



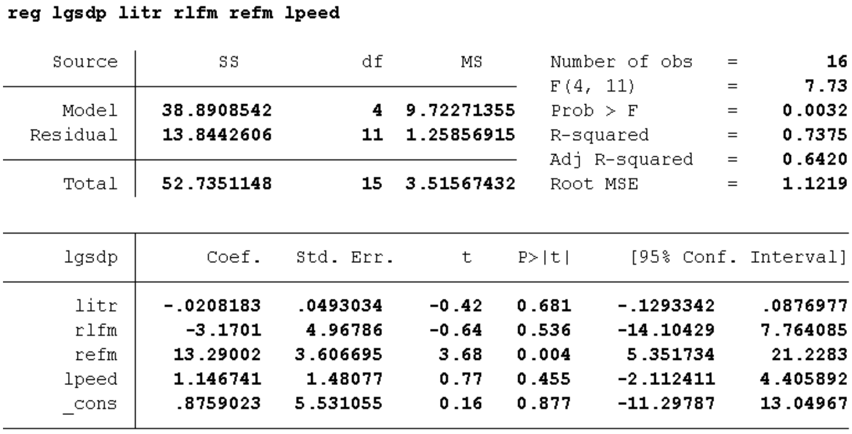
```
reg lgsdp litr rlfm refm lpeed

      Source |       SS           df       MS      Number of obs   =        26
-------------+----------------------------------   F(4, 21)        =     11.80
       Model |  44.0564094         4  11.0141024   Prob > F        =    0.0000
    Residual |  19.6043856        21   .93354217   R-squared       =    0.6920
-------------+----------------------------------   Adj R-squared   =    0.6334
       Total |   63.660795        25   2.5464318   Root MSE        =     .9662

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0234533   .0298764    -0.79   0.441    -.0855847    .0386782
        rlfm |  -1.290133    3.18116    -0.41   0.689    -7.905717     5.32545
        refm |   11.74995    2.58582     4.54   0.000     6.372441    17.12746
       lpeed |   1.278405   1.156564     1.11   0.282    -1.126801    3.683611
       _cons |   .8917255    3.70235     0.24   0.812    -6.807734    8.591185
------------------------------------------------------------------------------
```

20 Observations



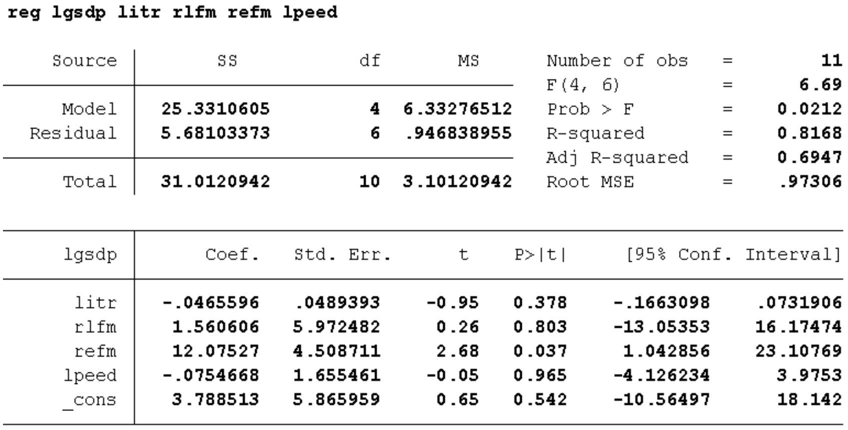
```
reg lgsdp litr rlfm refm lpeed

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(4, 15)        =     10.34
       Model |  42.2874343         4  10.5718586   Prob > F        =    0.0003
    Residual |  15.3341573        15  1.02227715   R-squared       =    0.7339
-------------+----------------------------------   Adj R-squared   =    0.6629
       Total |  57.6215916        19  3.03271535   Root MSE        =    1.0111

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0362364   .0347898    -1.04   0.314    -.1103892    .0379164
        rlfm |  -.7015655   3.659644    -0.19   0.851    -8.501911     7.09878
        refm |   12.41963   2.982356     4.16   0.001     6.062885    18.77637
       lpeed |   .7531921   1.290789     0.58   0.568    -1.998059    3.504443
       _cons |   2.085114   3.933104     0.53   0.604    -6.298099    10.46833
------------------------------------------------------------------------------
```

16 Observations



```
reg lgsdp litr rlfm refm lpeed

      Source |       SS           df       MS      Number of obs   =        16
-------------+----------------------------------   F(4, 11)        =      7.73
       Model |  38.8908542         4  9.72271355   Prob > F        =    0.0032
    Residual |  13.8442606        11  1.25856915   R-squared       =    0.7375
-------------+----------------------------------   Adj R-squared   =    0.6420
       Total |  52.7351148        15  3.51567432   Root MSE        =    1.1219

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0208183    .0493034    -0.42   0.681    -.1293342    .0876977
        rlfm |    -3.1701     4.96786    -0.64   0.536    -14.10429    7.764085
        refm |   13.29002    3.606695     3.68   0.004     5.351734     21.2283
       lpeed |   1.146741     1.48077     0.77   0.455    -2.112411    4.405892
       _cons |   .8759023    5.531055     0.16   0.877    -11.29787    13.04967
------------------------------------------------------------------------------
```

11 Observations

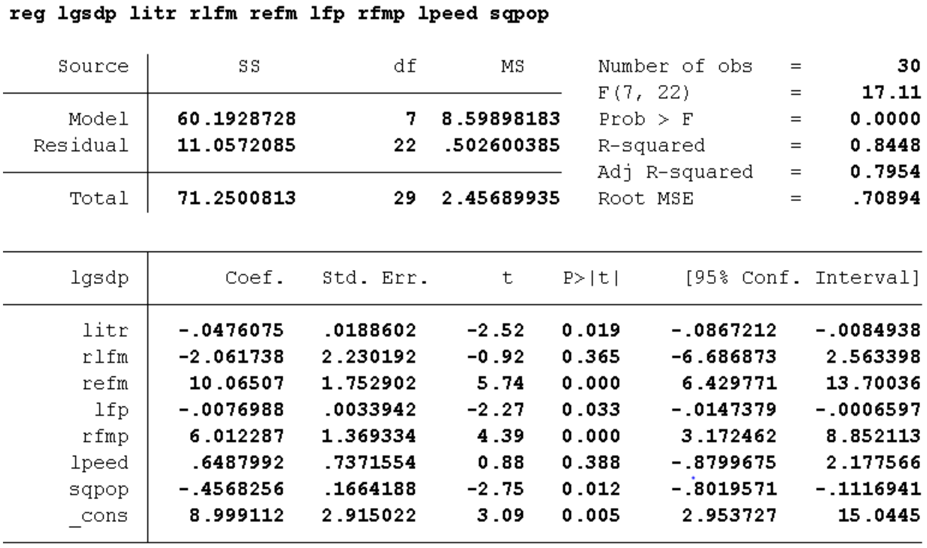


```
reg lgsdp litr rlfm refm lpeed

      Source |       SS           df       MS      Number of obs   =        11
-------------+----------------------------------   F(4, 6)         =      6.69
       Model |  25.3310605         4  6.33276512   Prob > F        =    0.0212
    Residual |  5.68103373         6  .946838955   R-squared       =    0.8168
-------------+----------------------------------   Adj R-squared   =    0.6947
       Total |  31.0120942        10  3.10120942   Root MSE        =    .97306

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0465596    .0489393    -0.95   0.378    -.1663098    .0731906
        rlfm |   1.560606    5.972482     0.26   0.803    -13.05353    16.17474
        refm |   12.07527    4.508711     2.68   0.037     1.042856    23.10769
       lpeed |  -.0754668    1.655461    -0.05   0.965    -4.126234      3.9753
       _cons |   3.788513    5.865959     0.65   0.542    -10.56497      18.142
------------------------------------------------------------------------------
```

The regression results for different observation suggests that with the decrease in the observation, the primary result that there is a negative impact of gender inequality in education on economic growth does not change. In all the regressions, the REFM value is significant, which essentially means that with the increase in the gender equality in terms of enrolment ratio, the economic growth also increase because the quality of human capital increases. Additionally, all the other slope coefficient in the regressions are insignificant. However, the R-squared value is fairly good and it increases with the decrease in the number of observations. Each regression model is jointly significant too as the p-value is lower than 0.05 (5% significance level), therefore we reject the null hypothesis, i.e. slope values are insignificant.

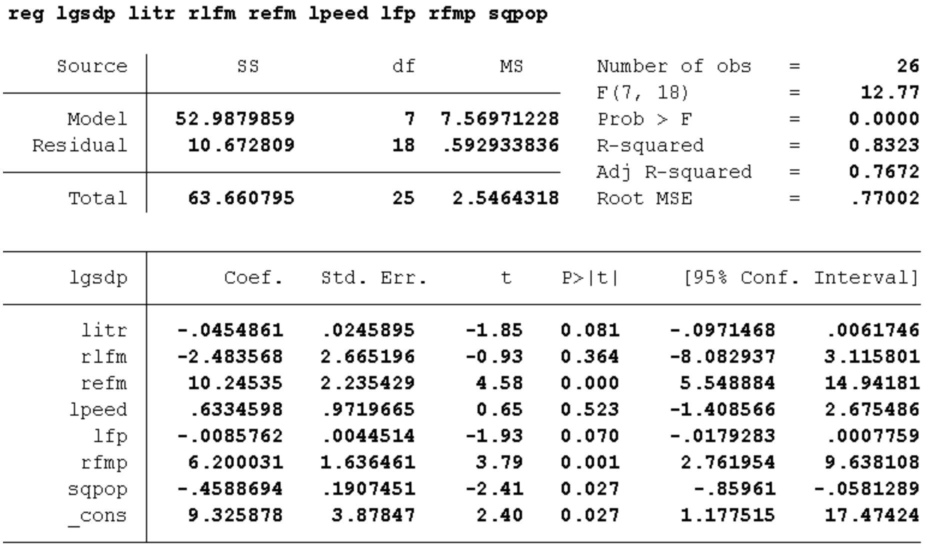
Model 2:

30 Observations



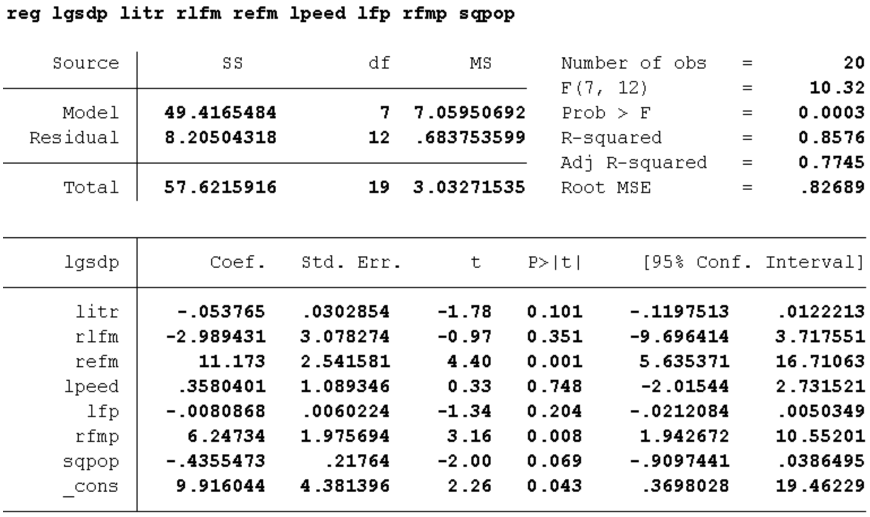
```
. reg lgsdp litr rlfm refm lfp rfmp lpeed sqpop

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(7, 22)        =     17.11
       Model |  60.1928728         7  8.59898183   Prob > F        =    0.0000
    Residual |  11.0572085        22  .502600385   R-squared       =    0.8448
-------------+----------------------------------   Adj R-squared   =    0.7954
       Total |  71.2500813        29  2.45689935   Root MSE        =    .70894

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0476075   .0188602    -2.52   0.019    -.0867212   -.0084938
        rlfm |  -2.061738   2.230192    -0.92   0.365    -6.686873    2.563398
        refm |   10.06507   1.752902     5.74   0.000     6.429771    13.70036
         lfp |  -.0076988   .0033942    -2.27   0.033    -.0147379   -.0006597
        rfmp |   6.012287   1.369334     4.39   0.000     3.172462    8.852113
       lpeed |   .6487992   .7371554     0.88   0.388    -.8799675    2.177566
       sqpop |  -.4568256   .1664188    -2.75   0.012    -.8019571   -.1116941
       _cons |   8.999112   2.915022     3.09   0.005     2.953727     15.0445
------------------------------------------------------------------------------
```

26 Observations



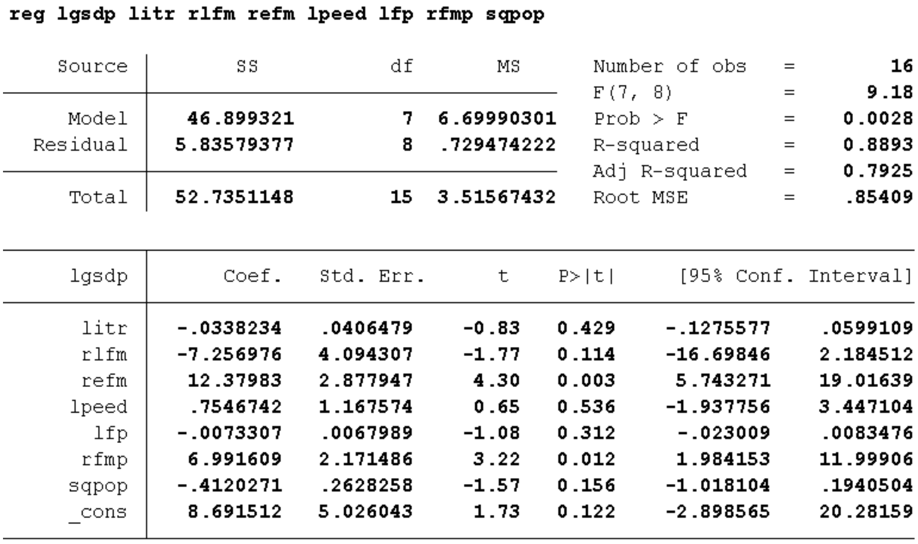
```
. reg lgsdp litr rlfm refm lpeed lfp rfmp sqpop

      Source |       SS           df       MS      Number of obs   =        26
-------------+----------------------------------   F(7, 18)        =     12.77
       Model |  52.9879859         7  7.56971228   Prob > F        =    0.0000
    Residual |   10.672809        18  .592933836   R-squared       =    0.8323
-------------+----------------------------------   Adj R-squared   =    0.7672
       Total |   63.660795        25   2.5464318   Root MSE        =    .77002

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0454861   .0245895    -1.85   0.081    -.0971468    .0061746
        rlfm |  -2.483568   2.665196    -0.93   0.364    -8.082937    3.115801
        refm |   10.24535   2.235429     4.58   0.000     5.548884    14.94181
       lpeed |   .6334598   .9719665     0.65   0.523    -1.408566    2.675486
         lfp |  -.0085762   .0044514    -1.93   0.070    -.0179283    .0007759
        rfmp |   6.200031   1.636461     3.79   0.001     2.761954    9.638108
       sqpop |  -.4588694   .1907451    -2.41   0.027      -.85961   -.0581289
       _cons |   9.325878    3.87847     2.40   0.027     1.177515    17.47424
------------------------------------------------------------------------------
```

20 Observations



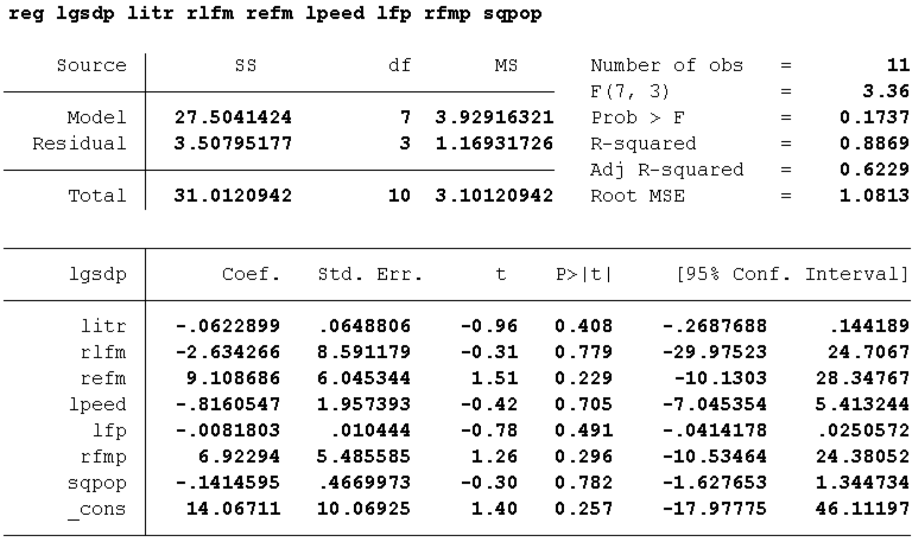
```
. reg lgsdp litr rlfm refm lpeed lfp rfmp sqpop

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(7, 12)        =     10.32
       Model |  49.4165484         7  7.05950692   Prob > F        =    0.0003
    Residual |  8.20504318        12  .683753599   R-squared       =    0.8576
-------------+----------------------------------   Adj R-squared   =    0.7745
       Total |  57.6215916        19  3.03271535   Root MSE        =    .82689

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |   -.053765   .0302854    -1.78   0.101    -.1197513    .0122213
        rlfm |  -2.989431   3.078274    -0.97   0.351    -9.696414    3.717551
        refm |     11.173   2.541581     4.40   0.001     5.635371    16.71063
       lpeed |   .3580401   1.089346     0.33   0.748     -2.01544    2.731521
         lfp |  -.0080868   .0060224    -1.34   0.204    -.0212084    .0050349
        rfmp |    6.24734   1.975694     3.16   0.008     1.942672    10.55201
       sqpop |  -.4355473     .21764    -2.00   0.069    -.9097441    .0386495
       _cons |   9.916044   4.381396     2.26   0.043     .3698028    19.46229
------------------------------------------------------------------------------
```

16 Observations



```
reg lgsdp litr rlfm refm lpeed lfp rfmp sqpop

      Source |       SS           df       MS      Number of obs   =        16
-------------+----------------------------------   F(7, 8)         =      9.18
       Model |   46.899321         7  6.69990301   Prob > F        =    0.0028
    Residual |  5.83579377         8  .729474222   R-squared       =    0.8893
-------------+----------------------------------   Adj R-squared   =    0.7925
       Total |  52.7351148        15  3.51567432   Root MSE        =    .85409

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0338234   .0406479    -0.83   0.429    -.1275577    .0599109
        rlfm |  -7.256976   4.094307    -1.77   0.114    -16.69846    2.184512
        refm |   12.37983   2.877947     4.30   0.003     5.743271    19.01639
       lpeed |   .7546742   1.167574     0.65   0.536    -1.937756    3.447104
         lfp |  -.0073307   .0067989    -1.08   0.312     -.023009    .0083476
        rfmp |   6.991609   2.171486     3.22   0.012     1.984153    11.99906
       sqpop |  -.4120271   .2628258    -1.57   0.156    -1.018104    .1940504
       _cons |   8.691512   5.026043     1.73   0.122    -2.898565    20.28159
------------------------------------------------------------------------------
```

11 Observations

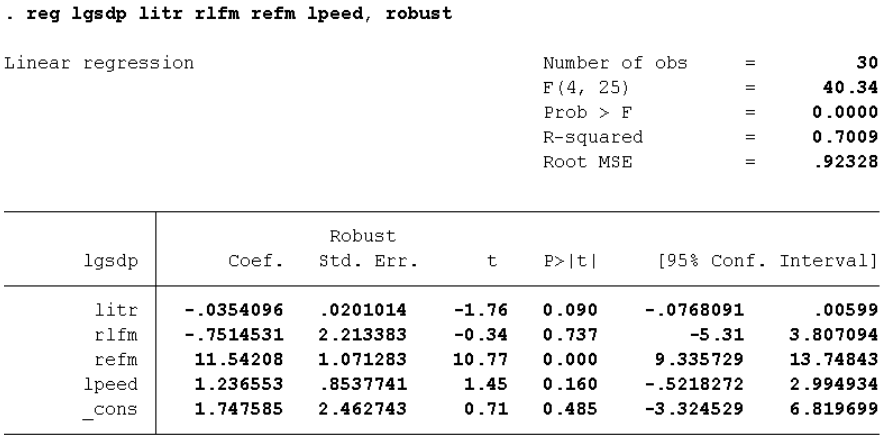


```
reg lgsdp litr rlfm refm lpeed lfp rfmp sqpop

      Source |       SS           df       MS      Number of obs   =        11
-------------+----------------------------------   F(7, 3)         =      3.36
       Model |  27.5041424         7  3.92916321   Prob > F        =    0.1737
    Residual |  3.50795177         3  1.16931726   R-squared       =    0.8869
-------------+----------------------------------   Adj R-squared   =    0.6229
       Total |  31.0120942        10  3.10120942   Root MSE        =    1.0813

------------------------------------------------------------------------------
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0622899   .0648806    -0.96   0.408    -.2687688     .144189
        rlfm |  -2.634266   8.591179    -0.31   0.779    -29.97523     24.7067
        refm |   9.108686   6.045344     1.51   0.229     -10.1303    28.34767
       lpeed |  -.8160547   1.957393    -0.42   0.705    -7.045354    5.413244
         lfp |  -.0081803    .010444    -0.78   0.491    -.0414178    .0250572
        rfmp |    6.92294   5.485585     1.26   0.296    -10.53464    24.38052
       sqpop |  -.1414595   .4669973    -0.30   0.782    -1.627653    1.344734
       _cons |   14.06711   10.06925     1.40   0.257    -17.97775    46.11197
------------------------------------------------------------------------------
```

The regression results illustrates that with the decrease in the number of observations, the results remain similar for all the regression analysis expect for the regression with 11 observations. Same variables are significant in all the regression analysis, while none of the slope coefficient is significant in the regression with 11 observations. R-squared value is fairly high and models are jointly significant for all except the regression with 11 observations. Hence the regression results clearly depicts that there is a negative impact of gender inequality in education on the economic growth, while this is not true for the regression with 11 observations. Additionally, the models also depicts that there is a negative relationship between the gender inequalities in labour force participation on the economic growth, except for the regression with 11 observations.

6. Robust regression results

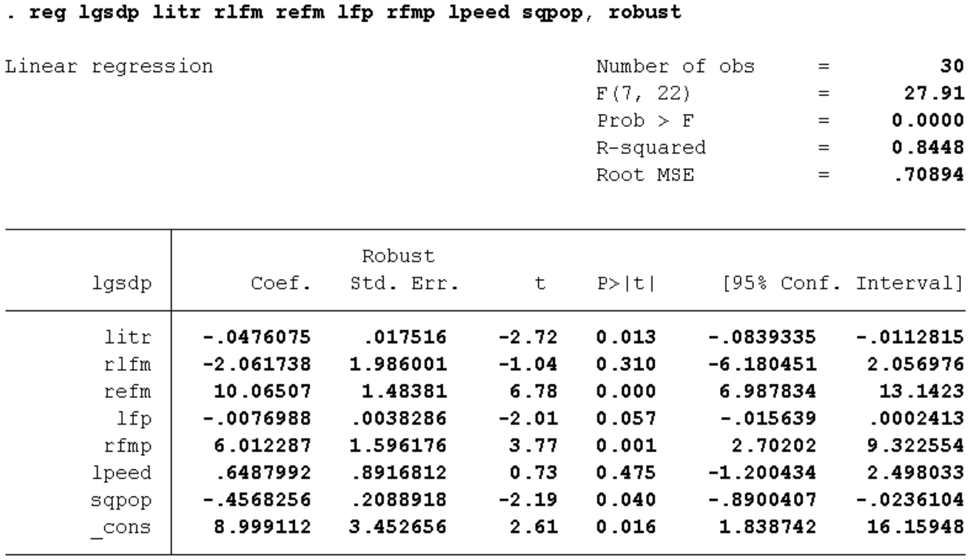
<u>Model 1:</u>



```
. reg lgsdp litr rlfm refm lpeed, robust

Linear regression                               Number of obs     =         30
                                                F(4, 25)          =      40.34
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7009
                                                Root MSE          =     .92328

------------------------------------------------------------------------------
             |               Robust
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0354096   .0201014    -1.76   0.090    -.0768091     .00599
        rlfm |  -.7514531   2.213383    -0.34   0.737        -5.31   3.807094
        refm |   11.54208   1.071283    10.77   0.000     9.335729   13.74843
       lpeed |   1.236553   .8537741     1.45   0.160    -.5218272   2.994934
       _cons |   1.747585   2.462743     0.71   0.485    -3.324529   6.819699
------------------------------------------------------------------------------
```

This model also explains that the slope coefficient of REFM is significant at 5% level of confidence, however the results from robust regression are better than the results from the normal regression.

<u>Model 2:</u>



```
. reg lgsdp litr rlfm refm lfp rfmp lpeed sqpop, robust

Linear regression                               Number of obs     =         30
                                                F(7, 22)          =      27.91
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8448
                                                Root MSE          =     .70894

------------------------------------------------------------------------------
             |               Robust
       lgsdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        litr |  -.0476075    .017516    -2.72   0.013    -.0839335   -.0112815
        rlfm |  -2.061738   1.986001    -1.04   0.310    -6.180451    2.056976
        refm |   10.06507    1.48381     6.78   0.000     6.987834     13.1423
         lfp |  -.0076988   .0038286    -2.01   0.057     -.015639    .0002413
        rfmp |   6.012287   1.596176     3.77   0.001      2.70202    9.322554
       lpeed |   .6487992   .8916812     0.73   0.475    -1.200434    2.498033
       sqpop |  -.4568256   .2088918    -2.19   0.040    -.8900407   -.0236104
       _cons |   8.999112   3.452656     2.61   0.016     1.838742    16.15948
------------------------------------------------------------------------------
```

This model also explains that the slope coefficient of REFM, LITR, LFP, RFMP, square-root of POPG are significant at 5% level of confidence, however the results from robust regression are better than the results from the normal regression.