INDIAN INSTITUTE OF TECHNOLOGY

(INDIAN SCHOOL OF MINES) DHANBAD

MANAGEMENT STUDIES DEPARTMENT

"FINANCIAL ECONOMETRICS ASSIGNMENT"

Prepared By:

Jainendra Tripathy (21je0417) , Kriti Thawaria (21je0485)

Kashish (21je0457), Umang Sinha (21je1005), Saksham Jha (21je0806)

Course: Financial Econometrics

Course Instructor: Prof. Aparna Krishna

Financial Econometrics Assignment

Initially, we are considering a model where total expenditure is the dependent variable, while food and non-food expenditures act as explanatory variables. In this setup, the variation in total expenditure can be fully accounted for by the explanatory variables. This is logical since total expenditure is merely the aggregate of food and non-food expenditures.

```
. regress exptot expfd expnfd

      Source |       SS           df       MS      Number of obs   =     1,129
-------------+----------------------------------   F(2, 1126)      >  99999.00
       Model |  1.9336e+10         2  9.6678e+09   Prob > F        =    0.0000
    Residual |  .000475549     1,126  4.2234e-07   R-squared       =    1.0000
-------------+----------------------------------   Adj R-squared   =    1.0000
       Total |  1.9336e+10     1,128  17141429.8   Root MSE        =    .00065

------------------------------------------------------------------------------
      exptot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       expfd |          1   1.33e-08   7.5e+07   0.000            1           1
      expnfd |          1   6.25e-09   1.6e+08   0.000            1           1
       _cons |   .0000221   .0000498     0.44   0.657    -.0000755    .0001197
------------------------------------------------------------------------------
```

Given this observation, including all expenditure terms in subsequent models may not be meaningful. As a result, we are solely considering food expenditure as our dependent variable in further modelling endeavours, disregarding other forms of expenditure.

- ## Proposed Hypothesis:

Our models focus on per capita food expenditure as the primary variable we're interested in. We're leaving out other expenditure categories because they're closely related to food expenditure. In Bangladeshi culture, the head of the household usually earns the most. So, factors like the head's age and education play a significant role in how much the household spends on food. Older, more educated heads of households tend to earn more, which means they're likely to spend more on food. We're also looking at total household assets as a factor. Families with more assets are more likely to spend on higher-quality but pricier food. Surprisingly, we're not considering family size as a factor. This is because, oddly enough, we found a negative connection between family size and food spending. This goes against common sense, where you'd expect larger families to spend more on food.

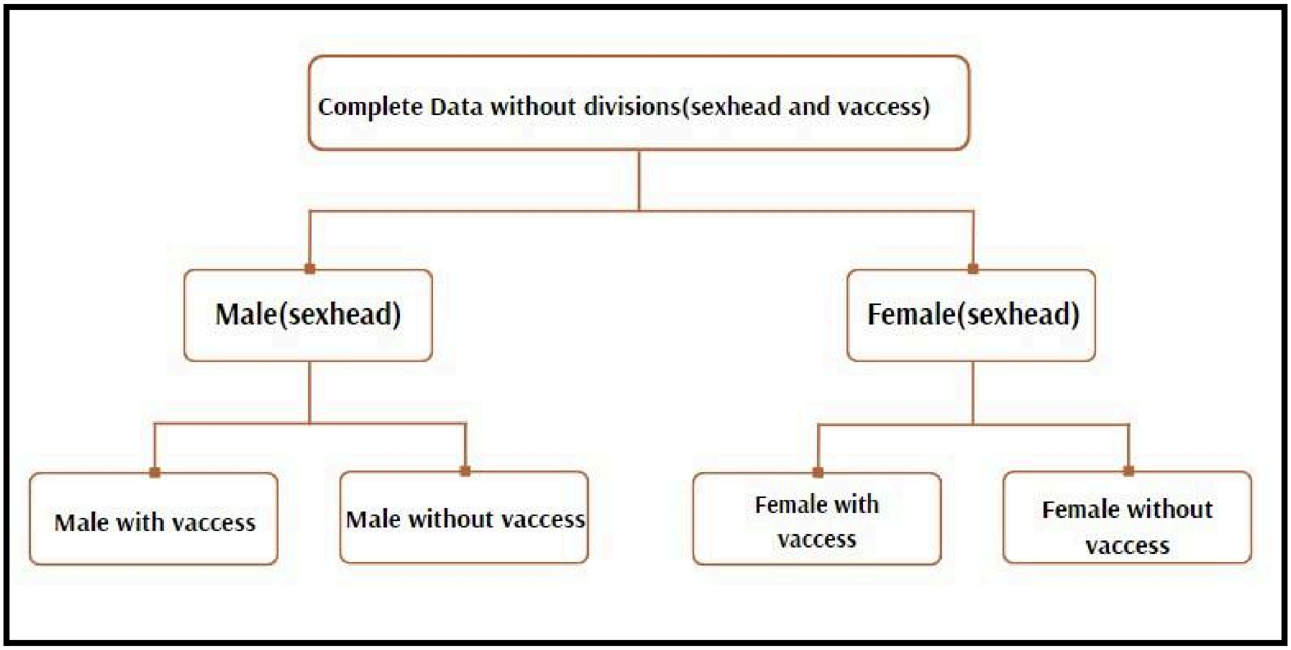
```
. corr famsize expfd
(obs=1,129)

             |  famsize    expfd
-------------+------------------
     famsize |   1.0000
       expfd |  -0.1327   1.0000
```

This is the case where we are considering the SRF to be,

$$expfd = \widehat{\beta o} + \widehat{\beta 1}\, hhasset + \widehat{\beta 2}\, agehead + \widehat{\beta 3}\, educhead + \widehat{Ui}$$

We propose different models while keeping the dependent and independent variables the same. Initially, we're testing our hypothesis using the entire dataset. Since we haven't learned about incorporating dummy variables in regression yet, we're dividing the data further based on the sex of the household head (male or female) and village access. We assume that these factors also play a significant role in per capita food consumption. Then, we apply our hypothesis to models based on these subsets of data.



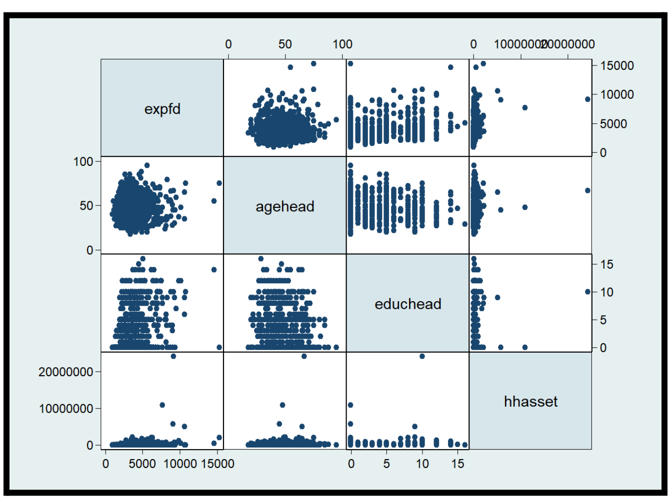
Model 01: Complete Data without divisions(sexhead and vaccess)

- **Outlier Detection and Removal:**

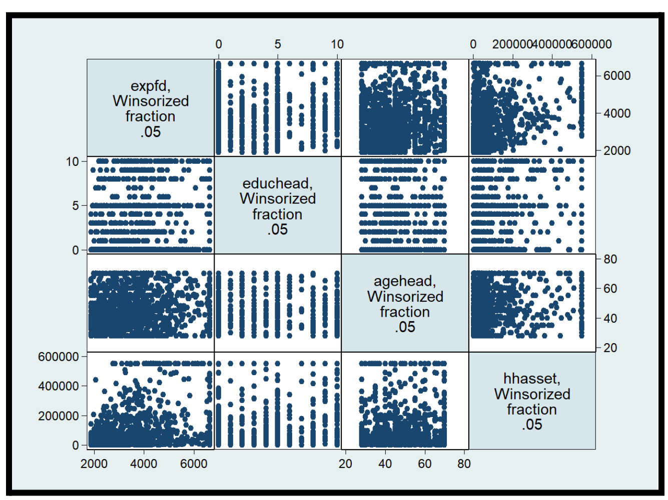
Code: graph matrix expfd agehead educhead hhasset **(Here we are checking for the outlier)**

Keypoint: Removing the outliers using winsorization: winsor variable_name, gen(winsorized_variable) p(#) and after this we have again checked the outlier.

Code: graph matrix winsorized_expfd winsorized_agehead winsorized_educhead winsorized_hhasse



Before (Removal of Outlier)



After (Removal of Outlier)

- ## Regression Analysis:

Now, we will make our linear regression based on the new variables(the removed outlier ones).

Code: regress winsorized_expfd winsorized_educhead winsorized_agehead winsorized_hhasset

Source	SS	df	MS
Model	334949859	3	111649953
Residual	1.5102e+09	1,125	1342405.89
Total	1.8452e+09	1,128	1635777.03

Number of obs	=	1,129
F(3, 1125)	=	83.17
Prob > F	=	0.0000
R-squared	=	0.1815
Adj R-squared	=	0.1793
Root MSE	=	1158.6

winsorized_expfd	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	91.34721	11.36524	8.04	0.000	69.04776	113.6467
winsorized_agehead	4.280413	2.977583	1.44	0.151	-1.561829	10.12265
winsorized_hhasset	.0024637	.0002701	9.12	0.000	.0019338	.0029936
_cons	2953.098	140.6958	20.99	0.000	2677.042	3229.153

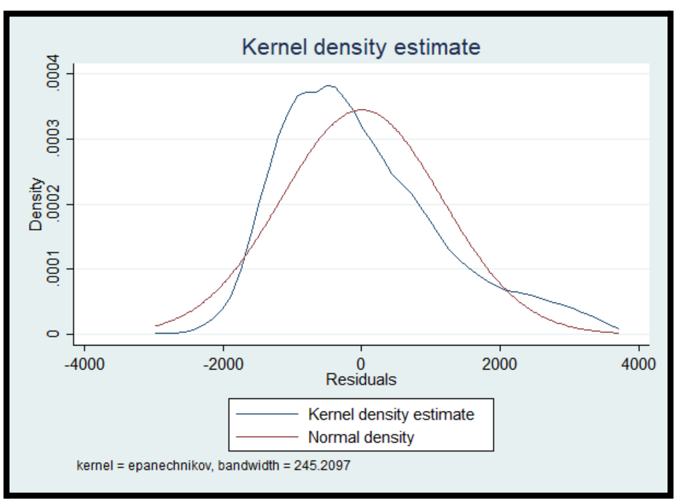
Interpretation: Here, we can see that the winsorized_agehead variable is not significant. And thus, will have no impact on expfd. While, others have a positive impact. And these variables can explain 17.93% of the dependent variable.

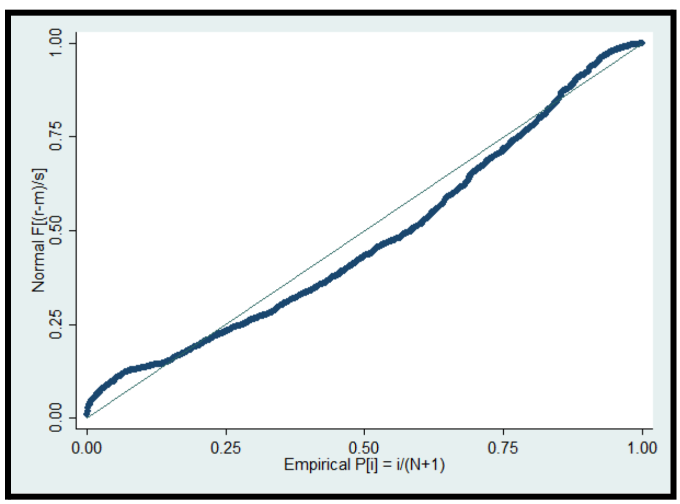
- ## Checking for the Normality of Residuals:

After this we will use predict r,residuals to store the residual in the variable r.

Code: predict r, residuals

Now we will Create plots to get a visual sense of whether or not normally distributed.





Code: kdensity r, normal

Code: pnorm r

Interpretation: Here we are getting an idea that the residuals may not be normally distributed.

Test for normality (Shapiro-Wilk W test):

Code: swilk r

The Shapiro-Wilk test is a statistical test used to assess the null hypothesis that a sample comes from a normally distributed population.

```
                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
           r |      1,129    0.95107     34.464     8.809    0.00000
```

Interpretation: So, it's possible that our residuals don't follow a normal distribution. We reject the null hypothesis, which assumes that the residuals are normally distributed, because the probability (Prob>Z) is less than 0.05.

- ## Checking Homoscedasticity of Residuals

Breusch-pagan test for heteroscedasticity

Code: hettest

The null hypothesis for the Breusch-Pagan test is that there is homoscedasticity, meaning that the variance of the residuals is constant across all levels of the independent variables.

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

         chi2(1)      =     5.78
         Prob > chi2  =   0.0162
```

Interpretation: But here, the p-value is less than 0.05, we would conclude that there is significant evidence of heteroskedasticity in the regression model.
Thus, we have heteroscedasticity. But, that won't be a problem if we are looking at the **central limit theorem.**

Removing Heteroscedasticity

Now we will focus on the **Robust Standard Errors**, this approach adjusts the standard errors of the regression coefficients to be robust to heteroskedasticity without directly modifying the estimation method.

Code: regress winsorized_expfd winsorized_educhead winsorized_agehead winsorized_hhasset, robust

Linear regression

Number of obs	= 1,129
F(3, 1125)	= 80.22
Prob > F	= 0.0000
R-squared	= 0.1815
Root MSE	= 1158.6

winsorized_expfd	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	91.34721	12.34543	7.40	0.000	67.12456	115.5699
winsorized_agehead	4.280413	2.981102	1.44	0.151	-1.568732	10.12956
winsorized_hhasset	.0024637	.0002856	8.63	0.000	.0019033	.0030241
_cons	2953.098	138.5844	21.31	0.000	2681.185	3225.011

- **Checking for Multicollinearity:**

Code: vif

Variable	VIF	1/VIF
winsorized~t	1.23	0.815326
winsor~chead	1.17	0.854796
winsor~ehead	1.06	0.939007
Mean VIF	1.15	

Interpretation: Our model does not have troublesome multicollinearity, as our VIF is less than 10.

- **Checking for Covariance between the independent variable and residual/error:**

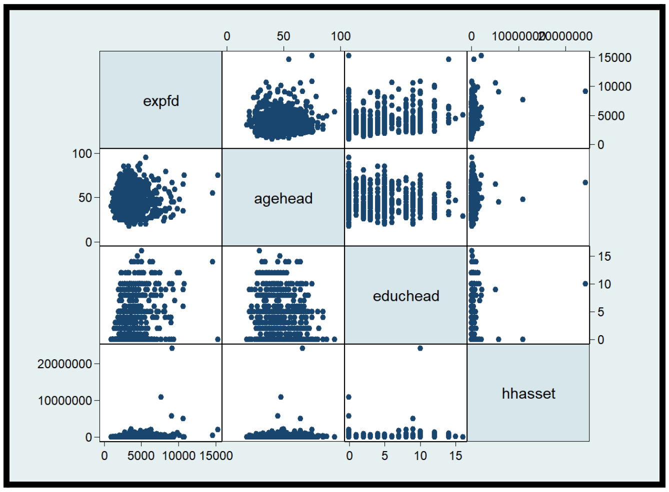
Code: corr r winsorized_educhead winsorized_agehead winsorized_hhasset

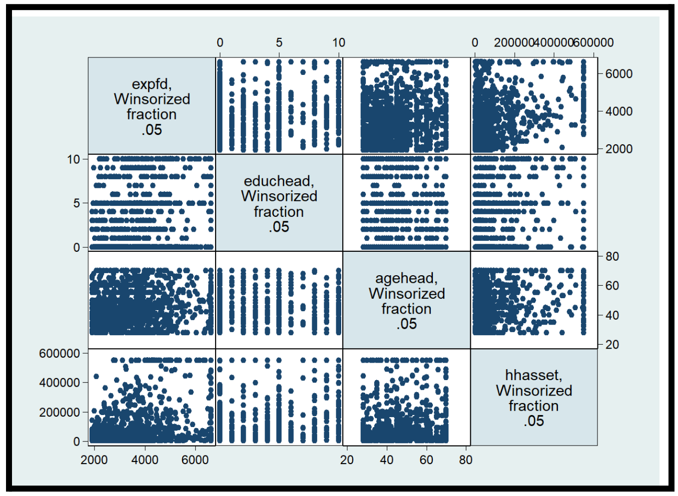
	r	wi~chead	wi~ehead	winsor~t
r	1.0000			
winsor~chead	0.0000	1.0000		
winsor~ehead	-0.0000	-0.0309	1.0000	
winsorized~t	0.0000	0.3641	0.2170	1.0000

Interpretation: Our model does not have multicollinearity, as our VIF is less than 10.

Model 02: Considering only the male and then taking the SRF of Model-01

- ## Outlier Detection and Removal:





Before (Removal of Outlier) **After (Removal of Outlier)**

- ## Regression Analysis:

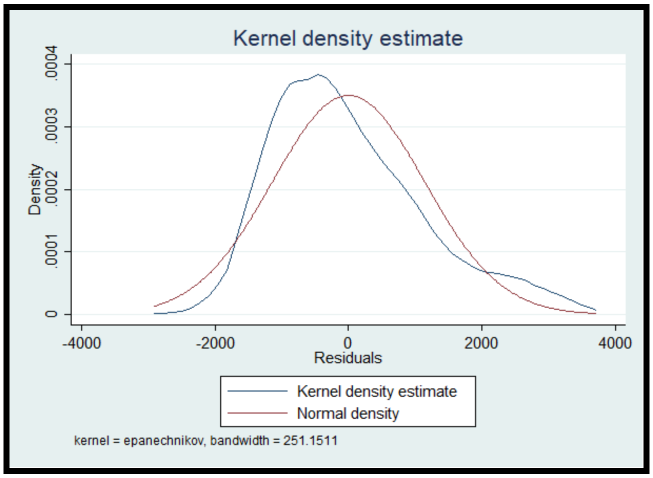
Now we will make our linear regression based on the new variables(t**he removed outlier one's**)

Source	SS	df	MS			
Model	298658855	3	99552951.8	Number of obs	=	1,025
Residual	1.3293e+09	1,021	1301994.71	F(3, 1021)	=	76.46
				Prob > F	=	0.0000
Total	1.6280e+09	1,024	1589839.31	R-squared	=	0.1835
				Adj R-squared	=	0.1811
				Root MSE	=	1141

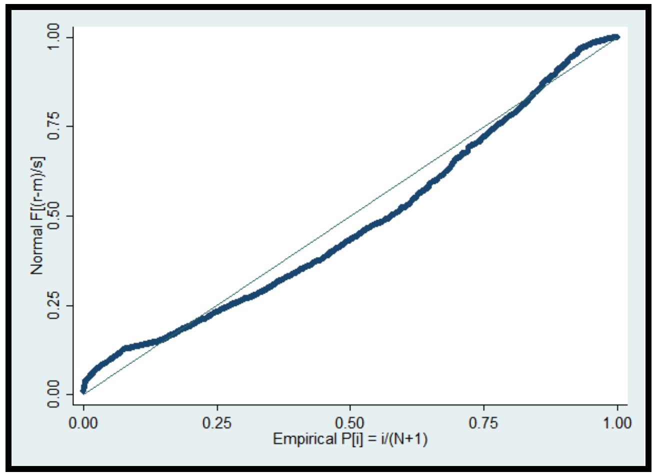
winsorized_expfd	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	92.92474	11.57992	8.02	0.000	70.20157	115.6479
winsorized_agehead	3.390813	3.086155	1.10	0.272	-2.665119	9.446744
winsorized_hhasset	.0023457	.0002799	8.38	0.000	.0017965	.0028949
_cons	2985.222	144.3872	20.68	0.000	2701.892	3268.551

Interpretation- For this case the **age of the head is insignificant** while the other **2 have a positive impact**, that we can see from the **T- test**. While the overall model is significant, the **explained variation is only 18.11%.**

- ## Checking for the normality of the residuals:



Code: kdensity r, normal



Code: pnorm r

Interpretation- The residuals slightly vary from having a normal distribution. Applying a more numerical test to check for the same.

<u>Test for normality (Shapiro-Wilk W test):</u>

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
r	**1,025**	**0.95456**	**29.309**	**8.374**	**0.00000**

Interpretation- The residuals are not normally distributed. But looking at the **central limit theorem,** as we have many data points it will not give us a problem.

- **<u>Checking Homoscedasticity of Residuals:</u>**

<u>Breusch-pagan test for heteroscedasticity</u>

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

         chi2(1)      =     7.70
         Prob > chi2  =   0.0055
```
:

Interpretation- We can **Reject the null hypothesis,** Thus our residuals are heteroscedastic. But looking at the **central limit theorem,** as we have many data points it will not give us a problem.

<u>Removing Heteroscedasticity</u>

Now we will focus on the **Robust Standard Errors:**

Linear regression

Number of obs	=	1,025
F(3, 1021)	=	72.17
Prob > F	=	0.0000
R-squared	=	0.1835
Root MSE	=	1141

winsorized_expfd	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	92.92474	12.66038	7.34	0.000	68.08139	117.7681
winsorized_agehead	3.390813	3.074245	1.10	0.270	-2.641749	9.423374
winsorized_hhasset	.0023457	.0002939	7.98	0.000	.001769	.0029225
_cons	2985.222	142.4931	20.95	0.000	2705.609	3264.835

Interpretation- For this case the age of the head is insignificant while the other 2 have a positive impact, that we can see from the **T- test**. While the overall model is significant from **F-test**, the **explained variation is only 18.35%**.

- <u>**Checking for Multicollinearity:**</u>

Variable	VIF	1/VIF
winsorized~t	1.24	0.807023
winsor~chead	1.18	0.850510
winsor~ehead	1.07	0.938335
Mean VIF	1.16	

Interpretation- Since our **VIF** is less than 10, we do not have a problem of multicollinearity.

- <u>**Checking for Covariance between the independent variable and residual/error:**</u>

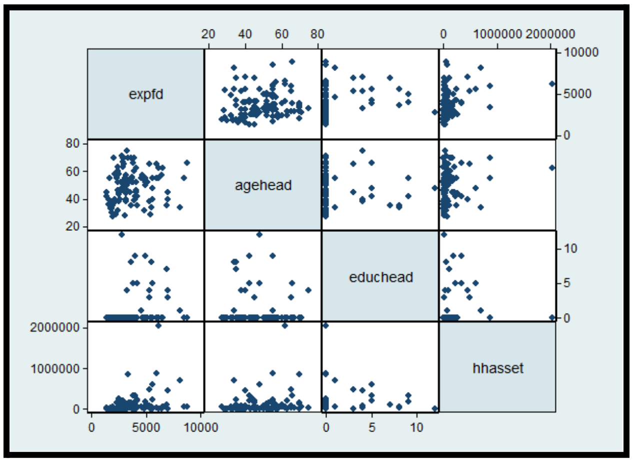
	r	wi~chead	wi~ehead	winsor~t
r	1.0000			
winsor~chead	-0.0000	1.0000		
winsor~ehead	0.0000	-0.0100	1.0000	
winsorized~t	0.0000	0.3742	0.2263	1.0000

Interpretation- So, we can see that our residual has **no correlation** with the independent variables.

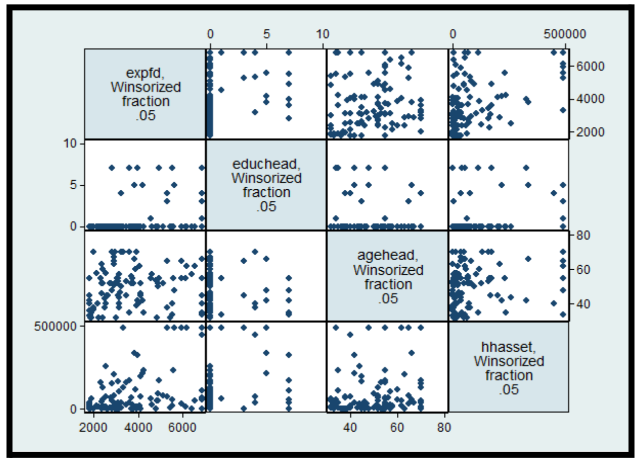
Conclusion- Thus, our regression model follows all the assumptions and is **verified.**

Model-03: Considering only the female and then taking the SRF of Model-01

Outlier Detection and Removal:



Before (Removal of Outlier)



After (Removal of Outlier)

- ## Regression Analysis:

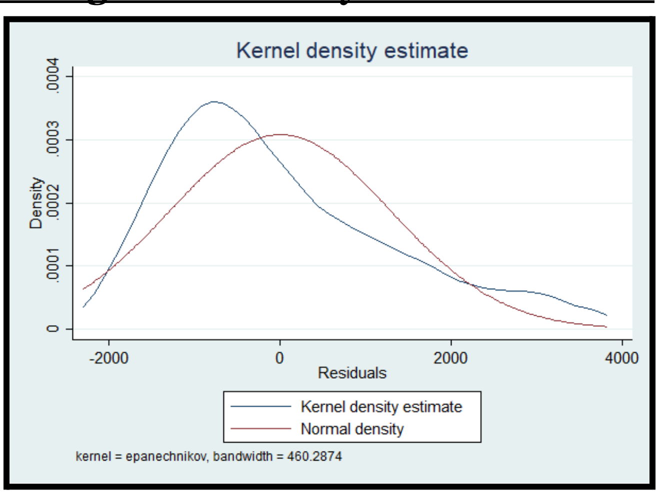
Now we will make our linear regression based on the new variables(t**he removed outlier one's**)

Source	SS	df	MS		
Model	45706294.6	3	15235431.5	Number of obs =	104
Residual	172672697	100	1726726.97	F(3, 100) =	8.82
				Prob > F =	0.0000
				R-squared =	0.2093
				Adj R-squared =	0.1856
Total	218378992	103	2120184.39	Root MSE =	1314

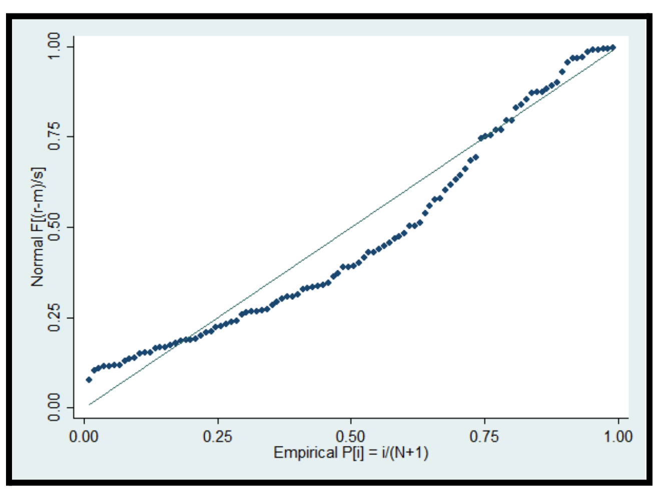
winsorized_expfd	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	175.2349	72.11932	2.43	0.017	32.15222	318.3176
winsorized_agehead	11.15372	11.76222	0.95	0.345	-12.1822	34.48964
winsorized_hhasset	.0035959	.0010793	3.33	0.001	.0014546	.0057373
_cons	2639.41	599.0597	4.41	0.000	1450.893	3827.928

Interpretation- For this case, the age of the head is insignificant, while the other 2 have a positive impact, as we can see from the T-test. While the overall model is significant, the explained variation is only 18.56%.

Checking the normality of the residuals:



Code: kdensity r, normal



Code: pnorm r

Interpretation- The residuals slightly vary from having a normal distribution. Applying a more numerical test to check for the same.

Test for normality (Shapiro-Wilk W test):

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
r	104	0.91185	7.520	4.485	0.00000

Interpretation- The residuals are not normally distributed. But looking at the **central limit theorem,** as we do not have many data points it will give us a problem.

- ## Checking Homoscedasticity of Residuals

Breusch-pagan test for heteroscedasticity

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

         chi2(1)      =     0.06
         Prob > chi2  =   0.8139
```

Interpretation- Here, we can infer that since **Prob>chi2** is greater than **81%** so our errors might be **homoscedastic**. That is our error's variance might be constant.
Our assumption of homoscedasticity is verified.

Here, we don't need to apply robust command to correct the violation of heteroskedasticity.
But, we will try to look at it.

```
Linear regression                               Number of obs     =        104
                                                F(3, 100)         =       9.93
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2093
                                                Root MSE          =       1314
```

winsorized_expfd	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	175.2349	76.59477	2.29	0.024	23.27306	327.1967
winsorized_agehead	11.15372	12.06031	0.92	0.357	-12.77358	35.08103
winsorized_hhasset	.0035959	.0011217	3.21	0.002	.0013706	.0058213
_cons	2639.41	589.323	4.48	0.000	1470.21	3808.61

Interpretation- From here, we can **reject Null Hypothesis** for **winsorized_educhead** and **winsorized_hhasset**. But, our model is not significant for **winsorized_agehead** and we can not **reject it on a 5% significance level**.

- **Checking for Covariance between the independent variable and residual/error:**

	r	wi~chead	wi~ehead	winsor~t
r	1.0000			
winsor~chead	0.0000	1.0000		
winsor~ehead	0.0000	-0.1368	1.0000	
winsorized~t	-0.0000	0.3037	0.1509	1.0000

Interpretation- We can see that the **covariance** between r and winsorized_educhead, winsorized_hhasset, winsorized_agehead is **0**. Thus, there is no correlation among them. Thus, **our assumption is satisfied**.

- **Checking for Multicollinearity:**

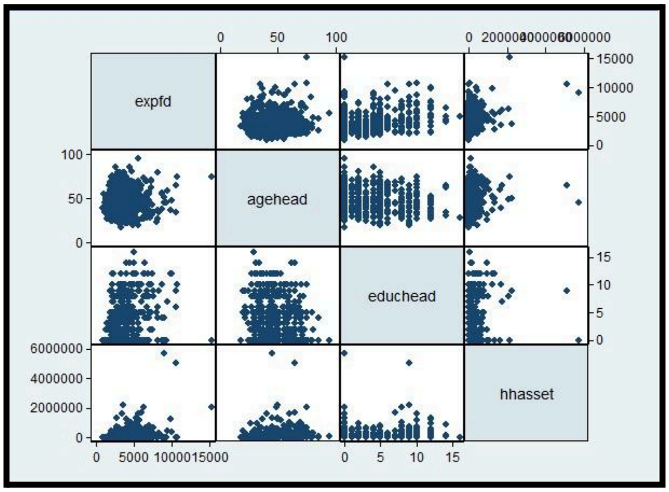
Variable	VIF	1/VIF
winsorized~t	1.15	0.870040
winsor~chead	1.14	0.873671
winsor~ehead	1.06	0.940489
Mean VIF	1.12	

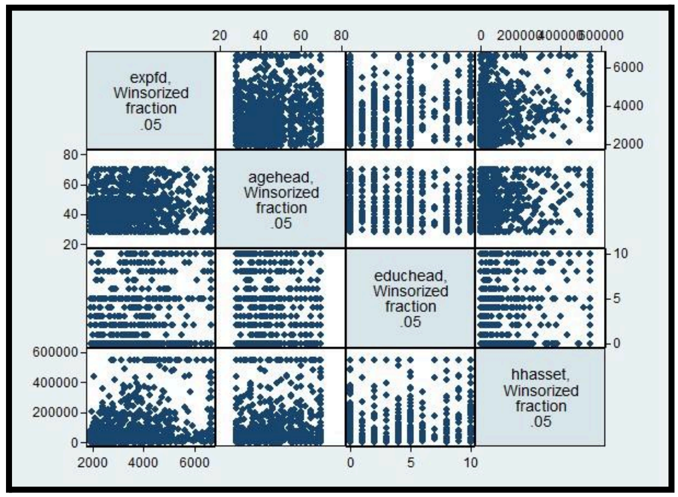
Interpretation- Since our **VIF** is less than 10, we do not have a problem of multicollinearity.

Conclusion- Thus, our regression model **does not follow all the assumptions**, we can say that our model is not fine**.** If proposed, the model will result in **biased analysis**.

Model 04: Considering only the male with vaccess and then taking the SRF of Model-01

- ## Outlier Detection and Removal:





Before (Removal of Outlier) **After (Removal of Outlier)**

- ## Regression Analysis:

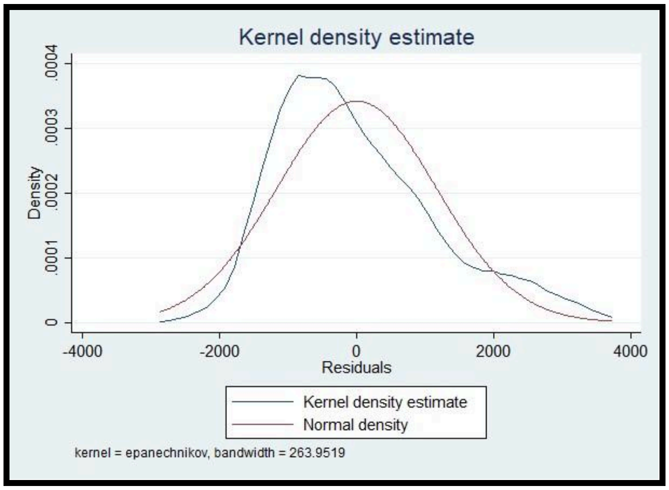
Now we will make our linear regression based on the new variables(t**he removed outlier one's**)

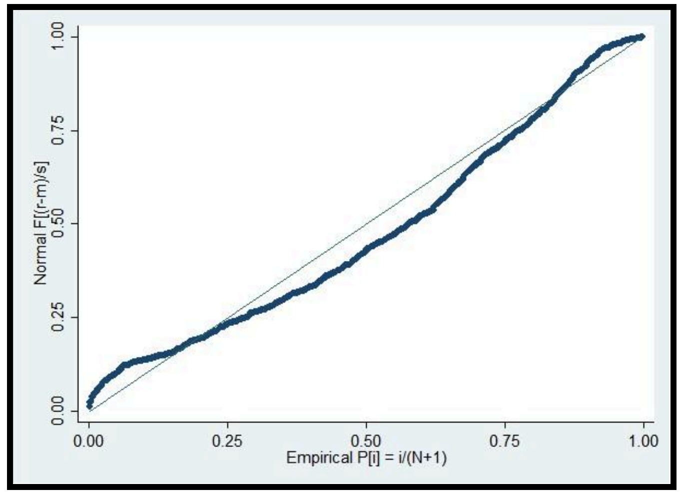
Source	SS	df	MS			
Model	240350145	3	80116715.2	Number of obs	=	860
Residual	1.1658e+09	856	1361941.26	F(3, 856)	=	58.83
				Prob > F	=	0.0000
Total	1.4062e+09	859	1636987.03	R-squared	=	0.1709
				Adj R-squared	=	0.1680
				Root MSE	=	1167

winsorized_expfd	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	107.3164	13.36998	8.03	0.000	81.07465	133.5582
winsorized_agehead	4.732908	3.48775	1.36	0.175	-2.112635	11.57845
winsorized_hhasset	.0019295	.0003207	6.02	0.000	.0013	.002559
_cons	2947.598	162.1943	18.17	0.000	2629.252	3265.943

Interpretation- For this case the **age of the head is insignificant** while the other **2 have a positive impact**, that we can see from the **T- test**. While the overall model is significant, the **explained variation is only 16.80%.**

- ## Checking for the normality of the residuals.





Interpretation- The residuals slightly vary from having a normal distribution. Applying a more numerical test to check for the same.

Test for normality (Shapiro-Wilk W test):

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
r	860	0.95085	27.023	8.113	0.00000

Interpretation- The residuals are not normally distributed. But looking at the **central limit theorem,** as we have many data points it will not give us a problem.

- ## Checking Homoscedasticity of Residuals

Breusch-pagan test for heteroscedasticity

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

         chi2(1)      =     6.67
         Prob > chi2  =   0.0098
```
:

Interpretation- We can **Reject the null hypothesis,** Thus our residuals are heteroscedastic. But looking at the **central limit theorem,** as we have many data points it will not give us a problem.

Removing Heteroscedasticity

To correct this,
Using **Robust stata command with regression we get-**

```
Linear regression                               Number of obs     =        860
                                                F(3, 856)         =      57.22
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1709
                                                Root MSE          =       1167
```

winsorized_expfd	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	107.3164	14.71345	7.29	0.000	78.43778	136.1951
winsorized_agehead	4.732908	3.471978	1.36	0.173	-2.08168	11.5475
winsorized_hhasset	.0019295	.0003248	5.94	0.000	.0012921	.002567
_cons	2947.598	159.4952	18.48	0.000	2634.55	3260.645

Interpretation- For this case the age of the head is insignificant while the other 2 have a positive impact, that we can see from the **T- test**. While the overall model is significant from **F-Test**, the explained variation is only **17.09%**.

- **Checking for the Covariance between the independent variables and residuals:**

	r	wi~chead	wi~ehead	winsor~t
r	1.0000			
winsor~chead	-0.0000	1.0000		
winsor~ehead	-0.0000	-0.0256	1.0000	
winsorized~t	-0.0000	0.3908	0.2434	1.0000

Interpretation- So, we can see that our residual has **no correlation** with the independent variables.

- **Checking for the Multicollinearity between the independent variables.**

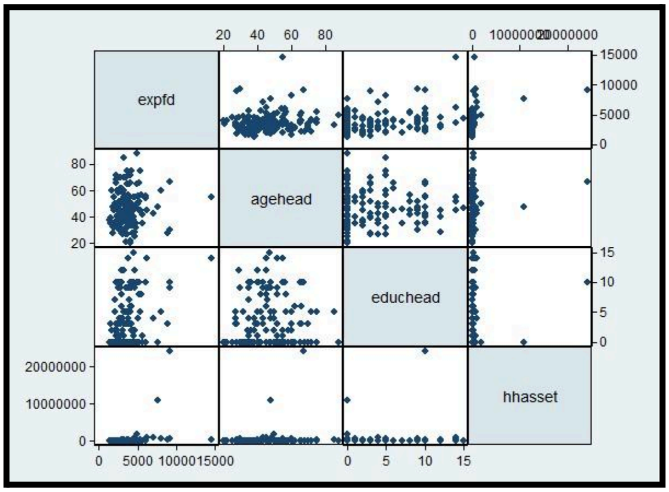
Variable	VIF	1/VIF
winsorized~t	1.28	0.783018
winsor~chead	1.20	0.831789
winsor~ehead	1.08	0.923545
Mean VIF	1.19	

Interpretation- Since our **VIF** is less than 10, we do not have a problem of multicollinearity.

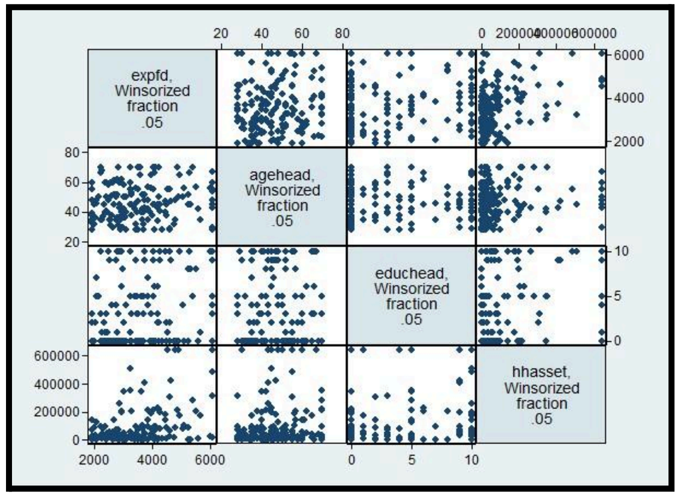
Conclusion- Thus, our regression model follows all the assumptions and is **verified.**

Model 05: Considering only the male without vaccess and then taking the SRF of Model-01

- ## Outlier Detection and Removal:



Before (Removal of Outlier)



After (Removal of Outlier)

- ## Regression Analysis:

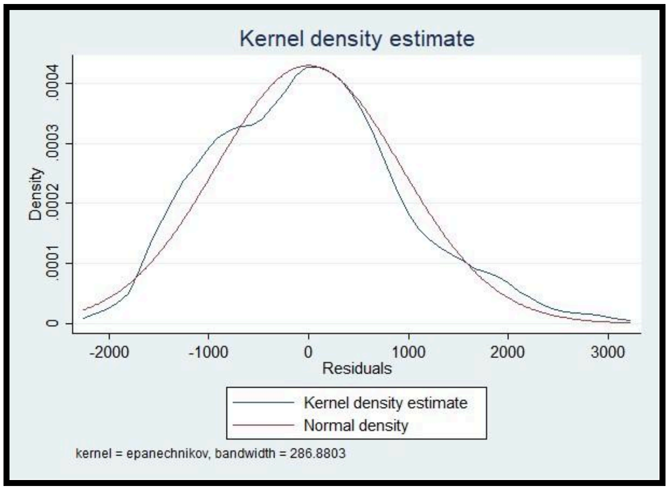
Now we will make our linear regression based on the new variables(t**he removed outlier one's**)

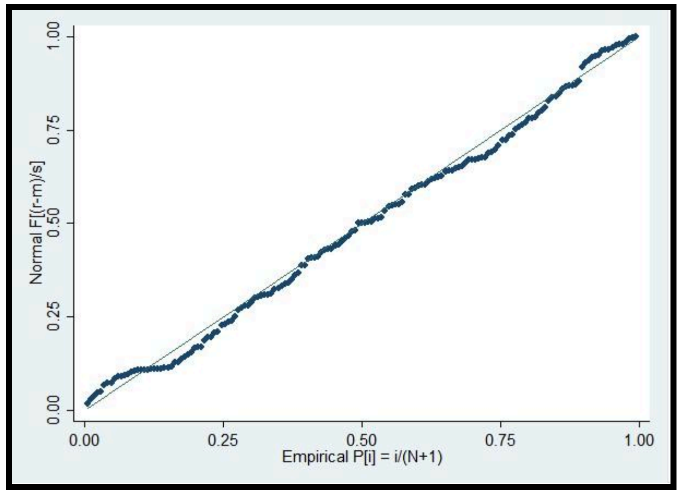
Source	SS	df	MS			
				Number of obs	=	165
				F(3, 161)	=	23.39
Model	61591075.6	3	20530358.5	Prob > F	=	0.0000
Residual	141292848	161	877595.327	R-squared	=	0.3036
				Adj R-squared	=	0.2906
Total	202883923	164	1237097.09	Root MSE	=	936.8

winsorized_expfd	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	49.59413	20.81326	2.38	0.018	8.491934	90.69633
winsorized_agehead	2.555757	6.071629	0.42	0.674	-9.434546	14.54606
winsorized_hhasset	.0032779	.000475	6.90	0.000	.0023398	.0042159
_cons	2940.176	292.833	10.04	0.000	2361.887	3518.464

Interpretation- For this case the **age of the head is insignificant** while the other **2 have a positive impact**, that we can see from the **T- test**. While the overall model is significant, the **explained variation is only 29.06%.**

- ## Checking for the Normality of Residuals:





Interpretation- The residuals slightly vary from having a normal distribution. Applying a more numerical test to check for the same.

<u>Test for normality (Shapiro-Wilk W test):</u>

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
r	**165**	**0.98042**	**2.473**	**2.063**	**0.01955**

Interpretation- The residuals are not normally distributed. But looking at the **central limit theorem,** as we have many data points it will not give us a problem.

- ## <u>Checking for homoscedasticity of residuals:</u>

<u>Breusch-pagan test for heteroscedasticity</u>

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

         chi2(1)      =     0.04
         Prob > chi2  =   0.8424
```
:

Interpretation- We can not **reject the null hypothesis,** Thus our residuals might be homoscedastic.That is our error's variance might be constant.
Our assumption of homoscedasticity is verified.

Here, we don't need to apply robust command to correct the violation of heteroskedasticity.
But, we will try to look at it.

Linear regression						
			Number of obs	=	165	
			F(3, 161)	=	28.34	
			Prob > F	=	0.0000	
			R-squared	=	0.3036	
			Root MSE	=	936.8	

winsorized_expfd	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	49.59413	22.67795	2.19	0.030	4.809534	94.37873
winsorized_agehead	2.555757	5.733985	0.45	0.656	-8.767763	13.87928
winsorized_hhasset	.0032779	.0004354	7.53	0.000	.0024181	.0041376
_cons	2940.176	282.9549	10.39	0.000	2381.394	3498.957

Interpretation- For this case the age of the head is insignificant while the other 2 have a positive impact, that we can see from the T- test. While the overall model is significant, the explained variation is only **30.36%**

- **Checking for the Covariance between the independent variables and residuals:**

	r	wi~chead	wi~ehead	winsor~t
r	1.0000			
winsor~chead	-0.0000	1.0000		
winsor~ehead	-0.0000	0.0465	1.0000	
winsorized~t	0.0000	0.2839	0.1355	1.0000

Interpretation- So, we can see that our residual has **no correlation** with the independent variables.

- **Checking for the Multicollinearity between the independent variables.**

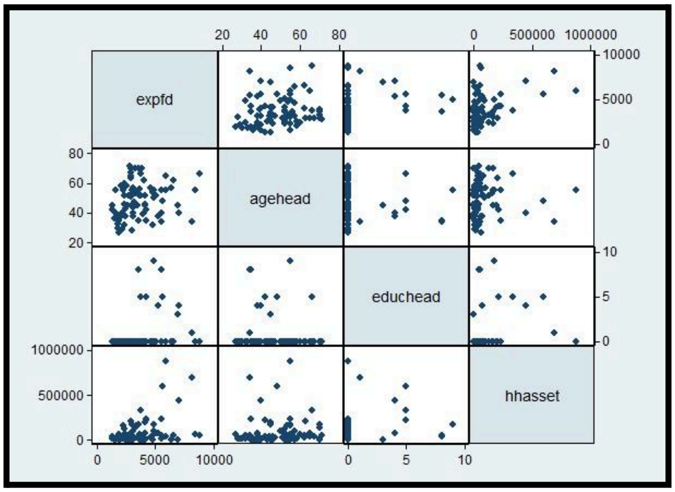
Variable	VIF	1/VIF
winsorized~t	1.11	0.904413
winsor~chead	1.09	0.919336
winsor~ehead	1.02	0.981567
Mean VIF	1.07	

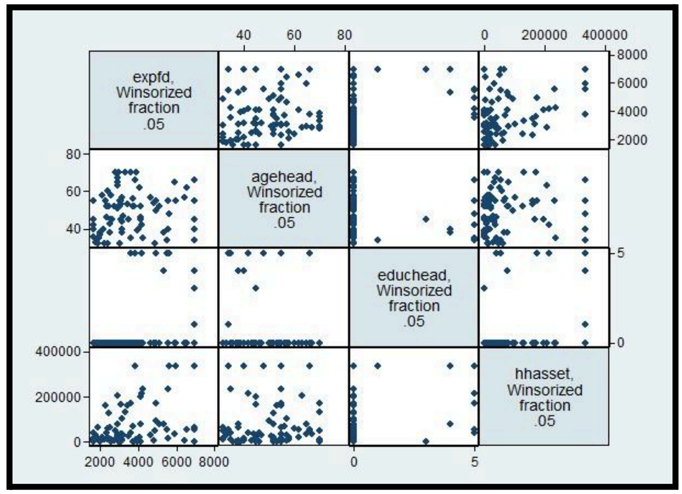
Interpretation- Since our **VIF** is less than 10, we do not have a problem of multicollinearity.

Conclusion- Thus, our regression model follows all the assumptions and is **verified.**

Model-06: Considering only the female with vaccess and then taking the SRF of Model-01

- ## Outlier Detection and Removal:





We have winsorized the data.

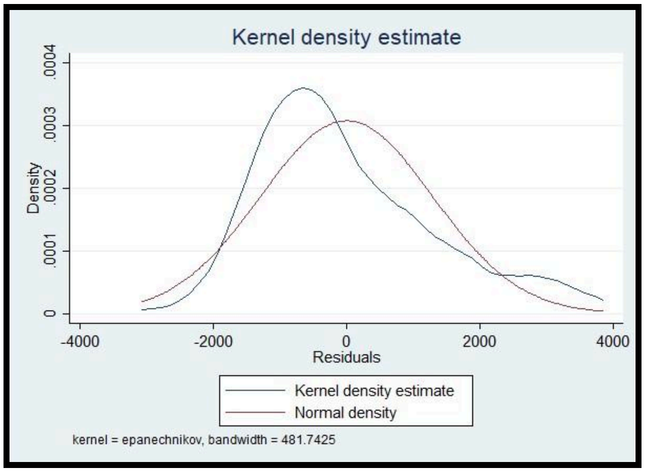
- ## Regression Analysis:

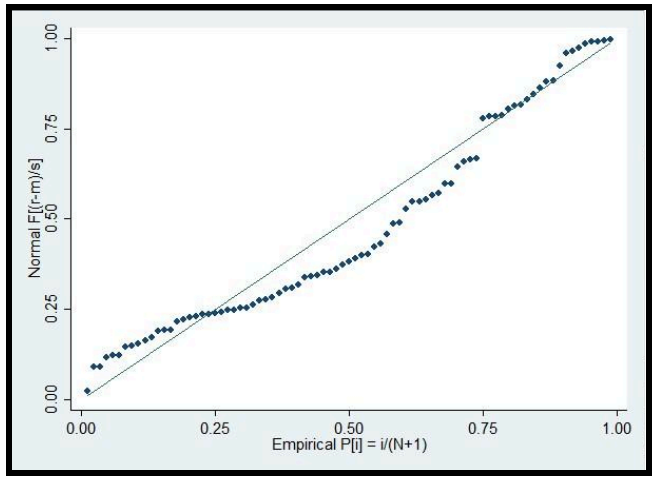
Source	SS	df	MS			
Model	43319485.1	3	14439828.4	Number of obs	=	83
Residual	137591002	79	1741658.25	F(3, 79)	=	8.29
				Prob > F	=	0.0001
				R-squared	=	0.2395
				Adj R-squared	=	0.2106
Total	180910487	82	2206225.45	Root MSE	=	1319.7

winsorized_expfd	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	271.4694	115.8654	2.34	0.022	40.84499	502.0937
winsorized_agehead	22.68626	13.68979	1.66	0.101	-4.562573	49.9351
winsorized_hhasset	.0045505	.0018337	2.48	0.015	.0009007	.0082003
_cons	2027.259	683.3939	2.97	0.004	666.9976	3387.521

Interpretation- For this case the age of the head is insignificant while the other 2 have a positive impact, that we can see from the T- test. While the overall model is significant, the explained variation is only **21.06%**.

- ## Checking for the normality of the residuals:





Interpretation- The residuals slightly vary from having a normal distribution. Applying a more numerical test to check for the same.

Test for normality (Shapiro-Wilk W test):

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
r	83	0.92376	5.394	3.700	0.00011

Interpretation- The residuals are not normally distributed. But looking at the **central limit theorem,** as we do not have many data points it will give us a problem.

- ## Checking Homoscedasticity of Residuals

Breusch-pagan test for heteroscedasticity

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

         chi2(1)      =     2.44
         Prob > chi2  =   0.1183
```

Interpretation- Here, we can infer that since **Prob>chi2** is greater than **0.05,** so our errors are **homoskedastic**. That is our error's variance is constant.
Our assumption of homoscedasticity is verified.

Here, we don't need to apply robust command to correct the violation of heteroskedasticity.
But, we will try to look at it.

Linear regression							
				Number of obs	=	83	
				F(3, 79)	=	6.44	
				Prob > F	=	0.0006	
				R-squared	=	0.2395	
				Root MSE	=	1319.7	

winsorized_expfd	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	271.4694	125.2832	2.17	0.033	22.09944	520.8393
winsorized_agehead	22.68626	14.75554	1.54	0.128	-6.683907	52.05644
winsorized_hhasset	.0045505	.0020961	2.17	0.033	.0003783	.0087227
_cons	2027.259	693.2941	2.92	0.005	647.2918	3407.227

Interpretation- From here, we can **reject Null Hypothesis** for **winsorized_educhead** and **winsorized_hhasset**. But, our model is not significant for **winsorized_agehead** and we can not **reject it on a 5% significance level**.While the overall model is significant, the explained variation is only **23.95%.**

- ## <u>Checking for Covariance between the independent variable and residual/error:</u>

	r	wi~chead	wi~ehead	winsor~t
r	1.0000			
winsor~chead	0.0000	1.0000		
winsor~ehead	0.0000	-0.1359	1.0000	
winsorized~t	0.0000	0.4538	0.1059	1.0000

Interpretation: We can see that the **covariance** between r and winsorized_educhead, winsorized_hhasset, winsorized_agehead is **0**. Thus, there is no correlation among them. Thus, **our assumption is satisfied**.

- ## <u>Checking for Multicollinearity:</u>

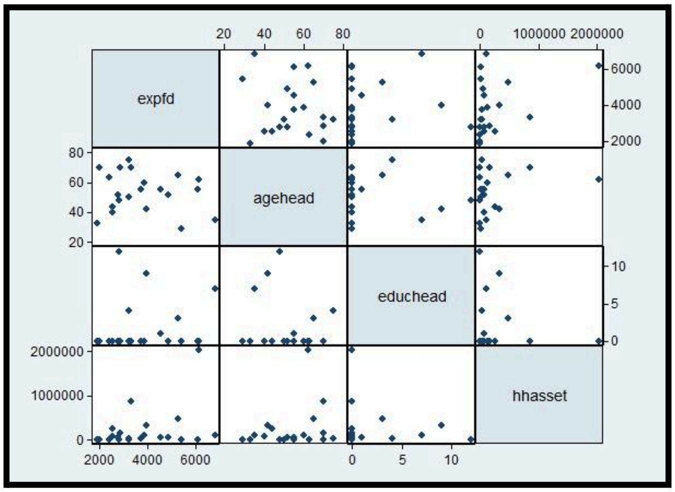
Variable	VIF	1/VIF
winsor~chead	1.32	0.759805
winsorized~t	1.31	0.765423
winsor~ehead	1.06	0.946176
Mean VIF	1.23	

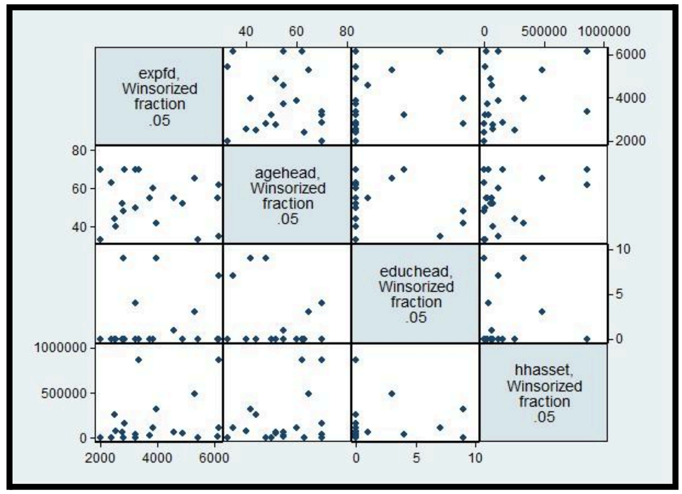
Interpretation: Since our **VIF** is less than 10, we do not have a problem of multicollinearity.

Conclusion- Thus, our regression model **does not follow all the assumptions**, we can say that our model is not fine**.** If proposed, the model will result in **biased analysis**.

Model-07: Considering only the female without vaccess and then taking the SRF of Model-01

- ## Outlier Detection and Removal:





We have winsorized the data.

- ## Regression Analysis:

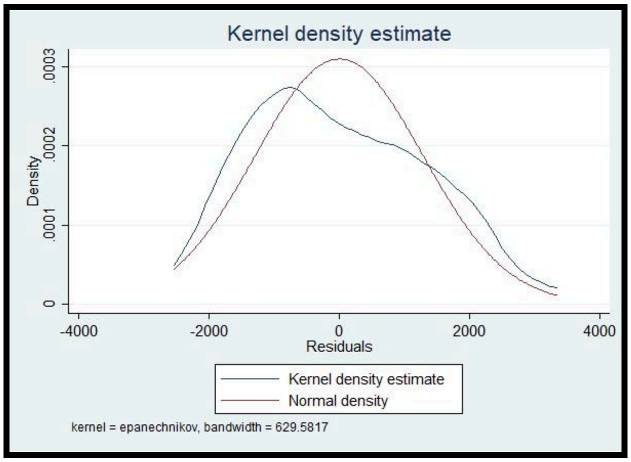
Source	SS	df	MS
Model	5726803.2	3	1908934.4
Residual	33077821.4	17	1945754.2
Total	38804624.6	20	1940231.23

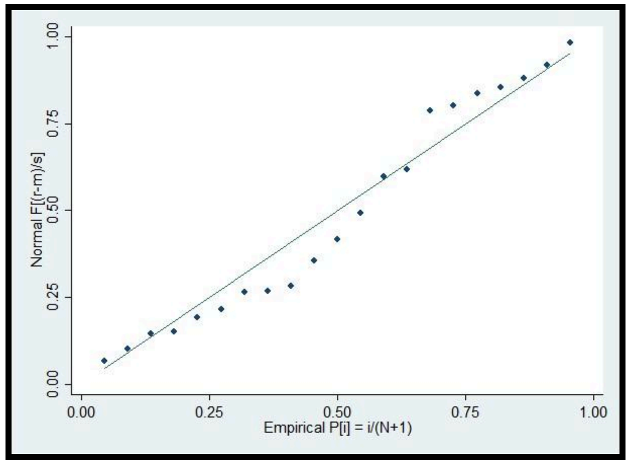
Number of obs	=	21
F(3, 17)	=	0.98
Prob > F	=	0.4249
R-squared	=	0.1476
Adj R-squared	=	-0.0028
Root MSE	=	1394.9

winsorized_expfd	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	42.35666	105.9325	0.40	0.694	-181.1414	265.8547
winsorized_agehead	-24.2779	27.58584	-0.88	0.391	-82.47893	33.92313
winsorized_hhasset	.0019567	.0012617	1.55	0.139	-.0007053	.0046187
_cons	4698.514	1500.238	3.13	0.006	1533.289	7863.739

Interpretation: Negative Adjusted R-Squared shows that our model doesn't fit the data means residual errors are larger than the total sum of squares and that can happen in no-intercept regression.

- ## Checking for the normality of the residuals:





Interpretation- The residuals slightly vary from having a normal distribution. Applying a more numerical test to check for the same.

Test for normality (Shapiro-Wilk W test):

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
r	21	0.95174	1.183	0.339	0.36723

Interpretation- The residuals are not normally distributed. But looking at the **central limit theorem,** as we have many data points it will not give us a problem.

- ## Checking for heteroscedasticity:

Breusch-pagan test for heteroscedasticity

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

         chi2(1)      =     0.26
         Prob > chi2  =   0.6088
```

Interpretation- Here, we can infer that since **Prob>chi2** is greater than **0.05** so our errors are **homoskedastic**. That is our error's variance is constant.
Our assumption of homoscedasticity is verified.

Here, we don't need to apply robust command to correct the violation of heteroskedasticity.
But, we will try to look at it.

Linear regression						
				Number of obs	=	21
				F(3, 17)	=	1.13
				Prob > F	=	0.3661
				R-squared	=	0.1476
				Root MSE	=	1394.9

winsorized_expfd	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
winsorized_educhead	42.35666	100.6141	0.42	0.679	-169.9205	254.6339
winsorized_agehead	-24.2779	28.03395	-0.87	0.399	-83.42437	34.86858
winsorized_hhasset	.0019567	.0012306	1.59	0.130	-.0006396	.004553
_cons	4698.514	1676.043	2.80	0.012	1162.372	8234.656

Interpretation- From here, we can **reject Null Hypothesis** for **winsorized_educhead** and **winsorized_hhasset**. But, our model is not significant for **winsorized_agehead** and we can not **reject it on a 5% significance level**. **Also now the r-squared is 14.76%.**

- **<u>Checking for the Covariance between the independent variables and residuals:</u>**

	r	wi~chead	wi~ehead	winsor~t
r	1.0000			
winsor~chead	0.0000	1.0000		
winsor~ehead	0.0000	-0.2406	1.0000	
winsorized~t	0.0000	-0.0092	0.3316	1.0000

Interpretation- We can see that the **covariance** between r and winsorized_educhead, winsorized_hhasset, winsorized_agehead is **0**. Thus, there is no correlation among them. Thus, **our assumption is satisfied**.

- **<u>Checking for the Multicollinearity:</u>**

Variable	VIF	1/VIF
winsor~ehead	1.20	0.833620
winsorized~t	1.13	0.884752
winsor~chead	1.07	0.936532
Mean VIF	1.13	

Interpretation- Since our **VIF** is less than 10, we do not have a problem of multicollinearity.

Conclusion- We will have to accept the null hypothesis for the entire model (F-test = 42%).