Math 219 Final Project Spring 2021

Meherab Hossain

How Income per Worker is affected by Labor Force, Years Spent in Schooling, Average Age and Capital per Worker.

1. In this project I was tasked with tackling the following dataset. I was presented with a panel data that holds information about the population (pop0099), average age (avgage), time worked in hours by the labor force (labfor), average years spend in schooling (yrs) [this was specified by the professor], capital per worker (kpw) and output per worker (ypw) for each state in US ranging from years 1840 to 2000 with increments of 10 years. The state names (statealp) were given unique integer identifiers (state). I noticed another variable (id) which was either 0 or 1 and were used to show if all the information about the variables were present for a year with a 1. In order to tackle this panel data, I decided to use a Fixed Effects Model using the Cobbs-Douglas equation with indicator (ind) as a dummy variable.

$$\ln(ypw_{it}) - \overline{\ln(ypw_i)} = \beta_1(\ln(kpw_{it}) - \overline{\ln(kpw_i)}) + \beta_2(ln(labfor_{it}) - \overline{\ln(labfor_i)}) + \beta_3(yrs_{it} - yrs_i) + \beta_4(ind_i - \overline{ind}) + \varepsilon_{it}$$

This is similar to a simpler OLS regression model like:

$$\ln(ypw_{it}) = \beta_1(\ln(kpw_{it}) + \beta_2 ln(labfor_{it}) + \beta_3(yrs_{it}) + \varepsilon_{it}$$

I believe that these are excellent variables to explain how the output of a worker is affected nationally despite the variability of time and State. I hypothesize that as these explanatory variables increase, we will see a significant increase in output per worker since they will have more machinery, more hours worked and more educated. The expressions with dashed line over them are the average of those variables over time. The dummy variable should end up being zero.

Throughout the project, I will try my best to compare this two models and see how my results are.

2. Here are my summary statistics of the variables:

```
summarize logkpw loglabfor yrs logypw
```

Variable	Obs	Mean	Std. Dev.	Min	Max
logkpw	776	10.90629	.8306841	8.689911	12.59671
loglabfor	801	13.06626	1.432935	7.756196	16.65405
yrs	793	7.090316	3.831267	.2432265	14.13723
logypw	776	9.717204	.7763157	7.886107	11.3198

```
summarize kpw labfor yrs ypw
```

Variable	Obs	Mean	Std. Dev.	Min	Max
kpw	776	73905.92	54429.19	5942.655	295583.3
labfor	801	1085360	1641590	2336	1.71e+07
yrs	793	7.090316	3.831267	.2432265	14.13723
ypw	776	21964.06	15988.76	2660.069	82438.04

Table 1.

I represented my data in two ways. First shows the summary of the data of the variables I will implement. These have been log transformed to fit the Cobbs-Douglas function. The next summary shows these variables before transformation. We first notice that we are missing some values from some States but that should not be a huge problem. Iowa in 1840 had the lowest capital per worker with 5942.655 machines per worker while DC in 2000 had the most with 295583.3 machines per worker. The hours worked by labor force per year was highest in California in 2000 with 17090815 hours. Minnesota had the lowest in 1850 with only 2336 hours. DC had the highest fraction of education in terms of years in 2000 with 14.13723 years and while North Carolina had the lowest in 1840 with a mere 0.2432265 years. In terms of output per worker, DC had the highest in 2000 again with 82438 and South Carolina had the lowest with 2660.069 in 1840. There are high standard deviations but that is because the values of the variables have increased over time drastically as the US Economy progressed with more and more investment. One can already notice that since all these statistics are increasing and with our knowledge from economics, we can suggest that the explanatory variables have a strong effect on the output per worker. To do so, we must dive in a bit further.

3. Here is the correlation of my regressors:

	logkpw	loglab~r	yrs
logkpw	1.0000		
loglabfor	0.5031	1.0000	
yrs	0.9501	0.5469	1.0000

Table 2.

From correlating the regressors we see that the average fraction of years spent in school is highly correlated with the log of capital per worker. This value of 0.9501 is high perhaps due the increase of both education and more capital through investment throughout the years. It could also be that more education meant more people were working and thus more capital was required per person.

This is an issue of collinearity. High multicollinearity is when two or more regressors are highly correlated. This would lead to wide confidence intervals, large changes in estimates when we add a few more observations as well as high standard errors on the parameter estimates even if the regressors are jointly significant. Since *yrs* and *logkpw* are positively correlated, it might be hard to distinguish the effect of the individual variable on the output per worker as they increase with each other. In order to create a good model, one could substitute out *yrs* with perhaps average age of the labor force (*avgage*) assuming that it is not correlated with other regressors. One could also opt to remove *yrs* altogether. This breaks our OLS (1, s) assumption.

4. Here is a simple OLS regression summary:

$$\ln(ypw_{it}) = \beta_1(\ln(kpw_{it}) + \beta_2 ln(labfor_{it}) + \beta_3(yrs_{it}) + \varepsilon_{it}$$

```
. regress logypw logkpw loglabfor yrs
```

Source	SS	df	MS
Model	439.959279	3	146.653093
Residual	26.1806568	770	.034000853
Total	466.139936	773	.603027084

Number of obs	=	774
F(3, 770)	=	4313.22
Prob > F	=	0.0000
R-squared	=	0.9438
Adj R-squared	=	0.9436
Root MSE	=	.18439

logypw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
logkpw	.8428557	.0256508	32.86	0.000	.7925019	.8932095
loglabfor	-.0055685	.0059609	-0.93	0.351	-.0172702	.0061331
yrs	.0159099	.0057324	2.78	0.006	.0046569	.0271628
_cons	.4837948	.2566356	1.89	0.060	-.0199937	.9875833

Table 3.

We notice that log of capital per worker (*logkpw*) and average fraction of educations in years (*yrs*) has a positive effect on log of output per worker (*logypw*). These values are also significant sine their P-values are very small.

However, it seems as if the log of hours put in by the labour force per year (*loglabfor*) has a small negative effect on output per worker. This value is not significant since it has a high P-value of 0.351. The R-squared value is high indicating that a large proportion of the variability the data points can be explained by the model and that the data points are close to our fitted regression line.

For comparison, here is my fixed-effects regression summary (specified in the first page):

```
. xtreg logypw logkpw loglabfor yrs i.ind,fe
note: 1.ind omitted because of collinearity

Fixed-effects (within) regression               Number of obs     =        774
Group variable: state                           Number of groups  =         51

R-sq:                                           Obs per group:
     within  = 0.9574                                         min =          6
     between = 0.8208                                         avg =       15.2
     overall = 0.9223                                         max =         17

                                                F(3,720)          =    5397.90
corr(u_i, Xb)  = -0.1666                        Prob > F          =     0.0000

------------------------------------------------------------------------------
      logypw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logkpw |   .7581838   .0267195    28.38   0.000     .7057264    .8106412
   loglabfor |  -.1092327    .011403    -9.58   0.000    -.1316197   -.0868457
         yrs |   .0564468   .0065307     8.64   0.000     .0436254    .0692682
       1.ind |          0  (omitted)
       _cons |   2.481367   .2965943     8.37   0.000     1.899074    3.063659
-------------+----------------------------------------------------------------
     sigma_u |   .1630698
     sigma_e |  .15232223
         rho |  .53403733   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(50, 720) = 8.17                     Prob > F = 0.0000
```

Table 4.

This shows that the log of hours put in by the labor force has a negative effect on the output per worker and is significant since the P-value is almost zero. It also shows that the log of capital per worker as well as the fraction of years spent in schooling is also significant. However, the former has a significant effect on the output per worker.

```
e(r2_a) =  .9542973997280456
```

The adjusted R-Squared as well as the R-Squared values are both high and slightly higher than the simple OLS regression model. I got the adjusted R-Squared value through the command “ereturn list”.

5. Here is my OLS regression summary with an additional regressor.

$$\ln(ypw_{it}) = \beta_1(\ln(kpw_{it}) + \beta_2 ln(labfor_{it}) + \beta_3(yrs_{it}) + \beta_4(avg_{it}) + \varepsilon_{it}$$

```
. regress logypw logkpw loglabfor yrs avgage
```

Source	SS	df	MS			
				Number of obs	=	774
				F(4, 769)	=	3240.36
Model	440.032856	4	110.008214	Prob > F	=	0.0000
Residual	26.1070799	769	.033949389	R-squared	=	0.9440
				Adj R-squared	=	0.9437
Total	466.139936	773	.603027084	Root MSE	=	.18425

logypw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
logkpw	.83681	.0259583	32.24	0.000	.7858524	.8877675
loglabfor	-.0060584	.0059657	-1.02	0.310	-.0177694	.0056526
yrs	.0110742	.006603	1.68	0.094	-.0018879	.0240363
avgage	.0051652	.0035086	1.47	0.141	-.0017224	.0120528
_cons	.4247668	.2595571	1.64	0.102	-.0847577	.9342912

Table 5.

Here we see that the R-Sqaured value increased very slightly and that Adjusted R-Squared was more or less the same. This might mean that the additional regressor does not have a significant impact on our simple OLS regression. However, only *logkpw* is significant in this case since most of the P-values are not less than 0.05.

I repeated this with my fixed effects model adding the same additional regressor of average age of the labor force:

$$\ln(ypw_{it}) - \overline{\ln(ypw_i)} = \beta_1(\ln(kpw_{it}) - \overline{\ln(kpw_i)}) + \beta_2(ln(labfor_{it}) - \overline{\ln(labfor_i)}) + \beta_3(yrs_{it} - yrs_i) + \beta_4(ind_i - \overline{ind}) + \beta_5(avgage_{it} - \overline{avgage_i}) + \varepsilon_{it}$$

```
. xtreg logypw logkpw loglabfor yrs avgage i.ind,fe
note: 1.ind omitted because of collinearity

Fixed-effects (within) regression               Number of obs     =        774
Group variable: state                           Number of groups  =         51

R-sq:                                           Obs per group:
     within  = 0.9595                                         min =          6
     between = 0.8056                                         avg =       15.2
     overall = 0.9201                                         max =         17

                                                F(4,719)          =    4256.09
corr(u_i, Xb)  = -0.1647                        Prob > F          =     0.0000
```

logypw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
logkpw	.8132282	.0276401	29.42	0.000	.7589632	.8674931
loglabfor	-.1012117	.0112125	-9.03	0.000	-.1232247	-.0791986
yrs	.0720231	.0068801	10.47	0.000	.0585156	.0855306
avgage	-.0242686	.0040271	-6.03	0.000	-.0321748	-.0163624
1.ind	0	(omitted)				
_cons	2.444038	.2896434	8.44	0.000	1.875391	3.012686
sigma_u	.17175612					
sigma_e	.14871846					
rho	.57151651	(fraction of variance due to u_i)				

```
F test that all u_i=0: F(50, 719) = 9.23                     Prob > F = 0.0000
```

Table 6.

In this case, we see that the values are significant and that average age of the labour force (*avgage*) as well as log of the hours put in the by the labor force (*loglabfor*) have a negative effect on the log of output per worker.

```
e(r2_a) =  .9564343648921015
```

The R-Squared and the adjusted R-Squared value both increased slightly showing that average age helps explain a bit more of the variability of the data.

6. In my Fixed Effects model, all my variables are significant at 1%, 5% and 10% levels. However, in the simple OLS model this is different.

   Before adding an additional regressor, *logkpw* and *yrs* were significant at all levels. However, *loglabfor* was not significant since its P-value was 0.351.

   After adding the additional regressor, only *logkpw* was significant at all levels. However, *yrs* was only significant at the 10% level. Other variables like *avgage and loglabfor* were not.

7. Here is my OLS regression with an additional regressor where I scale all my variables up by 2.

```
. regress twologypw twologkpw twologlabfor twoyrs twoavgage

      Source |       SS           df       MS      Number of obs   =       774
-------------+----------------------------------   F(4, 769)       =   3240.36
       Model |  1760.13142         4  440.032856   Prob > F        =    0.0000
    Residual |   104.42832       769  .135797555   R-squared       =    0.9440
-------------+----------------------------------   Adj R-squared   =    0.9437
       Total |  1864.55974       773  2.41210834   Root MSE        =    .36851

------------------------------------------------------------------------------
   twologypw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   twologkpw |     .83681    .0259583    32.24   0.000     .7858524    .8877675
twologlabfor |  -.0060584    .0059657    -1.02   0.310    -.0177694    .0056526
      twoyrs |   .0110742     .006603     1.68   0.094    -.0018879    .0240363
   twoavgage |   .0051652    .0035086     1.47   0.141    -.0017224    .0120528
       _cons |   .8495336    .5191142     1.64   0.102    -.1695153    1.868583
------------------------------------------------------------------------------
```

Table 7.

   Here all the coefficients, standard errors, significance (P-values) as well as the R-Squared and adjusted R-Squared remains the same. However, the constant changes as well as the Root MSE to adjust for the scaling. The stand error and, confidence interval and coefficient of the constant double since we are scaling up. The Root MSE increases since the scaling means that the data are now further apart from the fitted regression line.

I repeated the same thing with my Fixed Effects regression with similar results.

```
. xtreg twologypw twologkpw twologlabfor twoyrs twoavgage i.ind,fe
note: 1.ind omitted because of collinearity

Fixed-effects (within) regression               Number of obs     =        774
Group variable: state                           Number of groups  =         51

R-sq:                                           Obs per group:
     within  = 0.9595                                         min =          6
     between = 0.8056                                         avg =       15.2
     overall = 0.9201                                         max =         17

                                                F(4,719)          =    4256.09
corr(u_i, Xb)  = -0.1647                        Prob > F          =     0.0000
```

twologypw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
twologkpw	.8132282	.0276401	29.42	0.000	.7589632	.8674932
twologlabfor	-.1012117	.0112125	-9.03	0.000	-.1232247	-.0791986
twoyrs	.0720231	.0068801	10.47	0.000	.0585156	.0855306
twoavgage	-.0242686	.0040271	-6.03	0.000	-.0321748	-.0163624
1.ind	0	(omitted)				
_cons	4.888077	.5792868	8.44	0.000	3.750781	6.025373
sigma_u	.34351224					
sigma_e	.29743691					
rho	.57151652	(fraction of variance due to u_i)				

```
F test that all u_i=0: F(50, 719) = 9.23                     Prob > F = 0.0000
```

Table 8.

Again, only the constant values change as well as its standard error and confidence interval double. The standard deviation of time invariant individual specific term (sigma_u) as well as the standard deviation of the error term (sigma_e) both double. However, this is expected.

In both cases, the R-Squared value remained constant as well as the P-values and t-values.

8. This is the summary of my simple OLS regression where I only multiply the y-values.

```
. regress twologypw logkpw loglabfor yrs avgage
```

Source	SS	df	MS			
				Number of obs	=	774
				F(4, 769)	=	3240.36
Model	1760.13142	4	440.032856	Prob > F	=	0.0000
Residual	104.42832	769	.135797555	R-squared	=	0.9440
				Adj R-squared	=	0.9437
Total	1864.55974	773	2.41210834	Root MSE	=	.36851

twologypw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
logkpw	1.67362	.0519166	32.24	0.000	1.571705	1.775535
loglabfor	-.0121168	.0119314	-1.02	0.310	-.0355388	.0113053
yrs	.0221484	.0132061	1.68	0.094	-.0037758	.0480726
avgage	.0103305	.0070172	1.47	0.141	-.0034447	.0241056
_cons	.8495335	.5191142	1.64	0.102	-.1695154	1.868582

Table 9.

In this case, all the coefficients their standard errors and confidence interval doubles to adjust for the scaling of data in the y-direction. The Root MSE also doubles to adjust for the scaling since the data are further apart. However, the significance and t values do not change.

Next, I did the same with my Fixed Effects regression:

```
. xtreg twologypw logkpw loglabfor yrs avgage i.ind,fe
note: 1.ind omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =       774
Group variable: state                           Number of groups   =        51

R-sq:                                           Obs per group:
     within  = 0.9595                                         min =          6
     between = 0.8056                                         avg =       15.2
     overall = 0.9201                                         max =         17

                                                F(4,719)           =   4256.09
corr(u_i, Xb)  = -0.1647                        Prob > F           =    0.0000
```

twologypw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
logkpw	1.626456	.0552802	29.42	0.000	1.517926	1.734986
loglabfor	-.2024234	.0224249	-9.03	0.000	-.2464495	-.1583972
yrs	.1440463	.0137602	10.47	0.000	.1170313	.1710613
avgage	-.0485372	.0080541	-6.03	0.000	-.0643496	-.0327248
1.ind	0	(omitted)				
_cons	4.888077	.5792868	8.44	0.000	3.750781	6.025373
sigma_u	.34351223					
sigma_e	.29743691					
rho	.57151651	(fraction of variance due to u_i)				

```
F test that all u_i=0: F(50, 719) = 9.23                     Prob > F = 0.0000
```

Table 10.

Just as before, we get similar results where the standard errors, confidence intervals and the coefficients double but the t and P values remain the same. The Root MSE also doubles but the R-Squared values remain the same.

9. For my Hypothesis, I wanted to see whether *loglabfor* had a greater effect on output per worker than *avgage* using both models.

USING SIMPLE OLS MODEL

$$\ln(ypw_{it}) = \beta_1(\ln(kpw_{it}) + \beta_2 ln(labfor_{it}) + \beta_3(yrs_{it}) + \beta_4(avg_{it}) + \varepsilon_{it}$$

This made my hypothesis as the following:

$$H_0: \beta_2 - \beta_4 = 0$$
$$H_1: \beta_2 - \beta_4 > 0$$

First, we need the covariance matrix.

```
. matrix list e(V)

symmetric e(V)[5,5]
                logkpw   loglabfor         yrs      avgage       _cons
   logkpw    .00067383
loglabfor    .00001101   .00003559
      yrs   -.00012347  -7.627e-06    .0000436
   avgage   -.00001441  -1.167e-06  -.00001152   .00001231
    _cons   -.00614728  -.00049638   .00150535  -.00014068   .06736988
```

Table 11.

Degrees of Freedom = 774 − 4 = 770

$$t_{stat} = \frac{\widehat{\beta_2} - \widehat{\beta_4}}{se(\widehat{\beta_2} - \widehat{\beta_4})}$$

$$t_{stat} = \frac{(-0.0060584 - 0.0051652)}{\sqrt{(0.00003559 + 0.00001231 - 2 \times (-1.167 \times 10^{-6}))}}$$

$$t_{stat} = -1.58356$$

$$t_{770}^{0.10} = 1.28265$$

$$t_{770}^{0.05} = 1.64683$$

$$t_{770}^{0.01} = 2.3312$$

$$|t_{stat}| > |t_{770}^{0.10}|$$

It is only significant for the 10% level, but not for 5% or 1% level. For the 10% level we can reject the null hypothesis and say that *loglabfor* has a bigger effect on *logypw* than *avgage*. However, we cannot reject the null hypothesis for more significant levels.

Using Fixed Effects Model

$$\ln(ypw_{it}) - \overline{\ln(ypw_i)} = \beta_1(\ln(kpw_{it}) - \overline{\ln(kpw_i)}) + \beta_2(ln(labfor_{it}) - \overline{\ln(labfor_i)}) + \beta_3(yrs_{it} - yrs_i) + \beta_4(ind_i - \overline{ind}) + \beta_5(avgage_{it} - \overline{avgage_i}) + \varepsilon_{it}$$

This made my hypothesis as the following:

$$H_0: \beta_2 - \beta_5 = 0$$
$$H_1: \beta_2 - \beta_5 > 0$$

First, we need the covariance matrix.

```
. matrix list e(V)

symmetric e(V)[6,6]
                                                                 1o.
                logkpw   loglabfor         yrs      avgage         ind       _cons
   logkpw    .00076398
loglabfor     .0000434   .00012572
      yrs   -.00012284  -.00003256   .00004734
   avgage   -.00003678  -5.360e-06  -.00001041   .00001622
   1o.ind            0           0           0           0           0
    _cons   -.00684205  -.00172251   .00176394   .00002494           0    .0838933
```

Table 12.

Degrees of Freedom = 774 − 5 = 769

$$t_{stat} = \frac{\widehat{\beta_2} - \widehat{\beta_5}}{se(\widehat{\beta_2} - \widehat{\beta_5})}$$

$$t_{stat} = \frac{(-0.1012117 - (-0.0242686))}{\sqrt{(0.00001622 + 0.00012572 - 2 \times (-5.360 \times 10^{-6}))}}$$

$$t_{stat} = -6.22789$$

$$t_{769}^{0.10} = 1.28265$$

$$t_{769}^{0.05} = 1.64683$$

$$t_{769}^{0.01} = 2.3312$$

$$|t_{stat}| > |t_{770}^{0.10}|, |t_{stat}| > \left|t_{770}^{0.05}\right|, |t_{stat}| > |t_{770}^{0.01}|$$

Since the t-stat value is greater that the critical t-value at all significance levels, we can safely reject the null hypothesis and say that log of hours put in by the labour force in a year has a more significant impact than average age of the population.

10. Testing for heteroscedasticity in our data using simple OLS model.

First with Breusch Pagan test

```
. estat hettest

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

         chi2(1)      =    33.87
         Prob > chi2  =   0.0000
```

Table 13.

Since the probability of the chi-square statistic being less than 0.05 is 0, we can reject the null hypothesis of constant variance. Thus, we have heteroscedasticity.

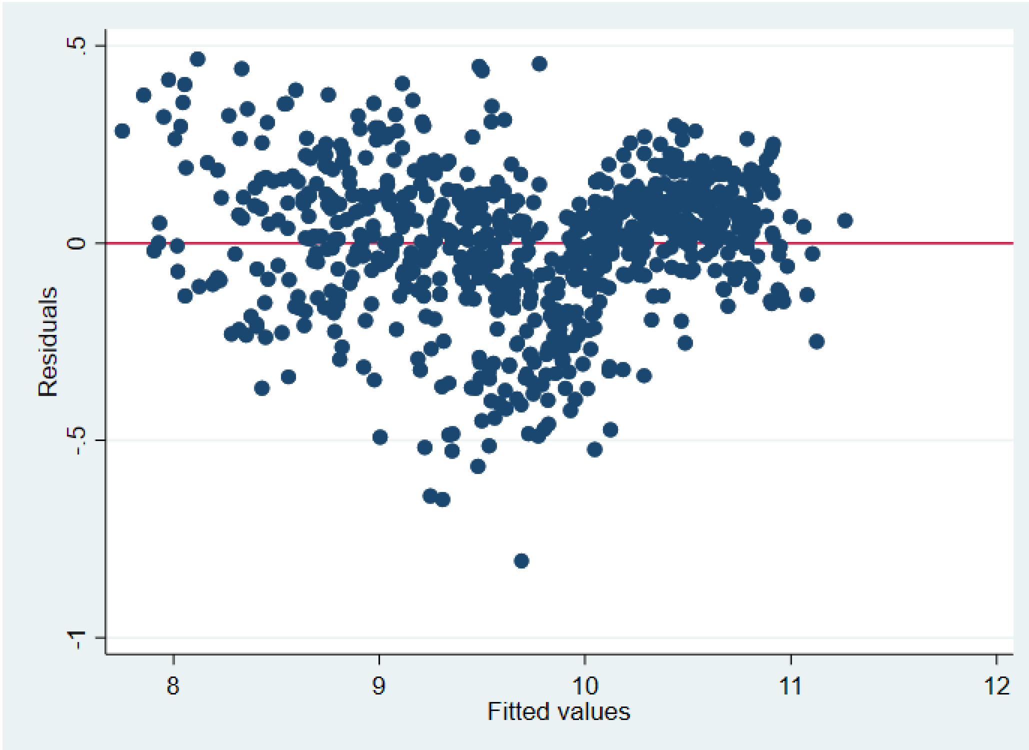
Using the White test.

```
. estat imtest, white

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

         chi2(14)     =     77.71
         Prob > chi2  =    0.0000

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

Source	chi2	df	p
Heteroskedasticity	77.71	14	0.0000
Skewness	28.90	4	0.0000
Kurtosis	4.92	1	0.0265
Total	111.53	19	0.0000

Table 14.

Again, we get similar results proving that we have heteroscedasticity. This shown by the following graph.



Graph 1.

Robust regression is used to identify outliers whilst minimizing their impact on the coefficient estimates. We make a robust regression with the same variables in our OLS model.

```
. regress logypw logkpw loglabfor yrs avgage, robust

Linear regression                               Number of obs     =        774
                                                F(4, 769)         =    3655.22
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9440
                                                Root MSE          =     .18425
```

logypw	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
logkpw	.83681	.0325818	25.68	0.000	.7728501	.9007698
loglabfor	-.0060584	.0060388	-1.00	0.316	-.0179128	.005796
yrs	.0110742	.0076961	1.44	0.151	-.0040338	.0261821
avgage	.0051652	.0034941	1.48	0.140	-.0016939	.0120244
_cons	.4247668	.3112374	1.36	0.173	-.1862089	1.035742

Table 16.

11. OLS (2, s) is a crucial assumption where we assume that the error term has a zero-conditional mean. It is likely that it is violated in both our models since it might have lagged dependent variables since output of the previous year decade might have an effect on the next. Next, we might have omitted an important variable such as productivity of the labor force. It is likely that the correlation between *yrs* and *logkpw* also led to this error (they might have been jointly determined).

12. An instrument variable is a third variable that is correlated the X variables but is not correlated error term. It will help find the true correlation between the explanatory variable and the response variable.

    I think the log of the total factor of productivity is a good IV in this case, which I believe is not correlated with the error term and can help alleviate the omitted variable bias.