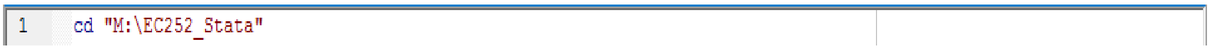
Daniel Howes 1604133

EC252 Introduction to Econometrics

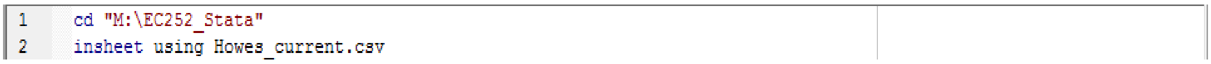
Introduction:

I imported my excel file to Stata by using the command ‘cd’ this changes the current working directory to the specified drive and directory.



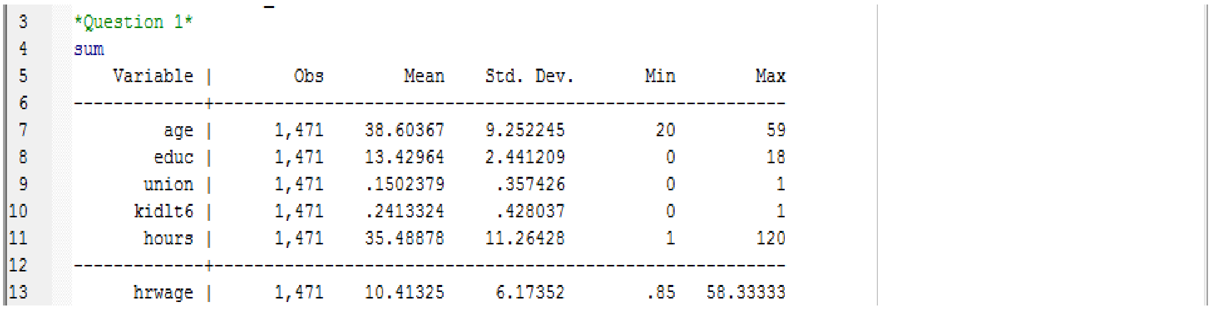
```
cd "M:\EC252_Stata"
```

I then used the command ‘insheet using Howes_current.csv’ to read the data created by my excel file.



```
cd "M:\EC252_Stata"
insheet using Howes_current.csv
```

Question 1: To find the summary statistics you can use the command ‘sum’. This creates a table of the summary statistics



```
*Question 1*
sum
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      1,471    38.60367    9.252245         20         59
        educ |      1,471    13.42964    2.441209          0         18
       union |      1,471    .1502379     .357426          0          1
      kidlt6 |      1,471    .2413324     .428037          0          1
       hours |      1,471    35.48878    11.26428          1        120
-------------+---------------------------------------------------------
      hrwage |      1,471    10.41325     6.17352        .85   58.33333
```

Here we find the values for mean, standard dev and minimum and maximum value.

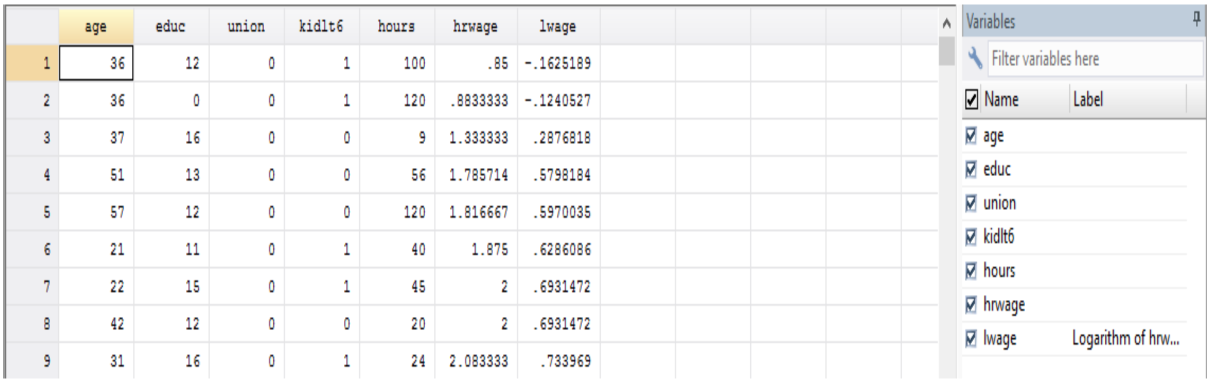
Question 2a: To generate a new variable we use the gen command, and we set this equal to the natural logarithm of hrwage.

```
*Question 2a*
gen lwage = ln(hrwage)
```

We then use the label command to give the new variable a name. “We call this the Logarithm of hrwage”

```
label variable lwage "Logarithm of hrwage"
```

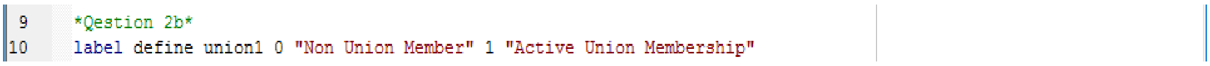
We can see this new variable when we enter the command ‘browse’ to bring up our dataset:



	age	educ	union	kidlt6	hours	hrwage	lwage
1	36	12	0	1	100	.85	-.1625189
2	36	0	0	1	120	.8833333	-.1240527
3	37	16	0	0	9	1.333333	.2876818
4	51	13	0	0	56	1.785714	.5798184
5	57	12	0	0	120	1.816667	.5970035
6	21	11	0	1	40	1.875	.6286086
7	22	15	0	1	45	2	.6931472
8	42	12	0	0	20	2	.6931472
9	31	16	0	1	24	2.083333	.733969

☑ Name	Label
☑ age	
☑ educ	
☑ union	
☑ kidlt6	
☑ hours	
☑ hrwage	
☑ lwage	Logarithm of hrw...

Question 2b: To label the indicator values I used the following command:



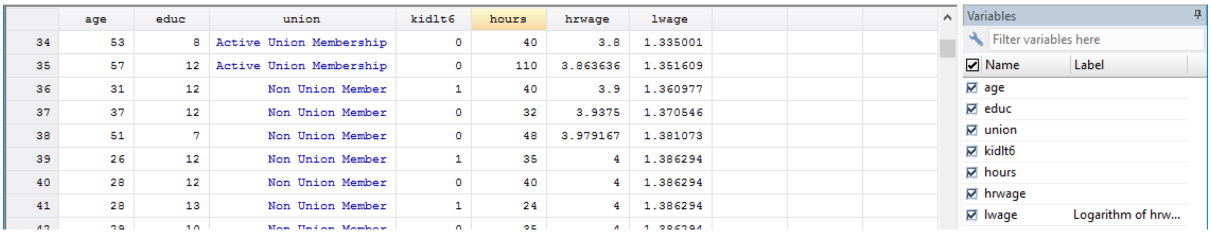
```
*Qestion 2b*
label define union1 0 "Non Union Member" 1 "Active Union Membership"
```

Under the variable union the string variables were labelled with 0 as “Non union Member” and 1 “Active union member”

Question 2c: This command will replace the values of zero and one for union membership with the variables stated above.

```
*Questiom 2c*
label values union union1
```

When we go onto browse after typing in this command we can see the new labels for union membership:



	age	educ	union	kidlt6	hours	hrwage	lwage
34	53	8	Active Union Membership	0	40	3.8	1.335001
35	57	12	Active Union Membership	0	110	3.863636	1.351609
36	31	12	Non Union Member	1	40	3.9	1.360977
37	37	12	Non Union Member	0	32	3.9375	1.370546
38	51	7	Non Union Member	0	48	3.979167	1.381073
39	26	12	Non Union Member	1	35	4	1.386294
40	28	12	Non Union Member	0	40	4	1.386294
41	28	13	Non Union Member	1	24	4	1.386294

Variables

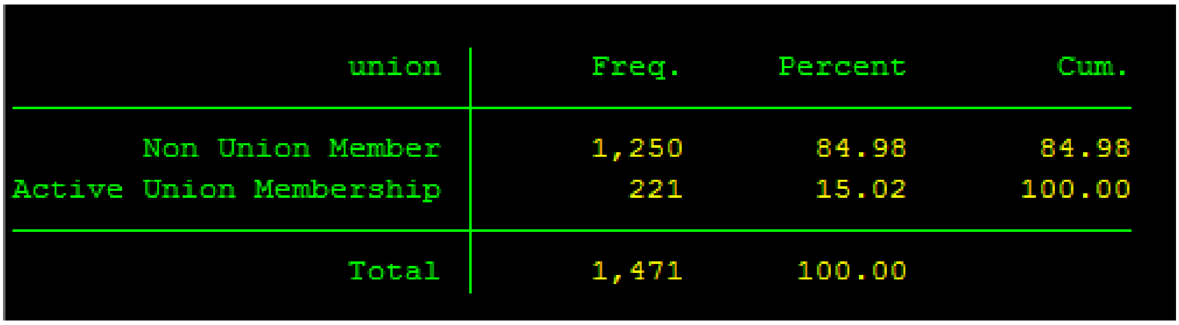
Filter variables here

☑ Name	Label
☑ age	
☑ educ	
☑ union	
☑ kidlt6	
☑ hours	
☑ hrwage	
☑ lwage	Logarithm of hrw...

Question 2d: For this we use the command tabulate to find the percentages and frequency, we can see from running this command:

```
*Question 2d*
tab union
```

This shows us this table:



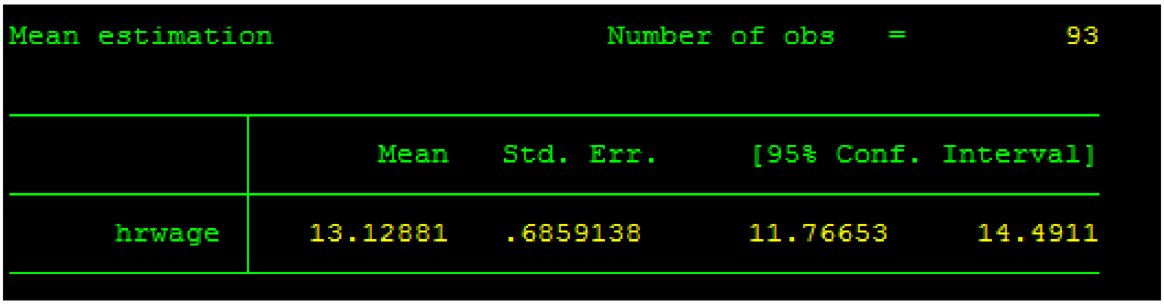
```
                  union |      Freq.     Percent        Cum.
------------------------+-----------------------------------
       Non Union Member |      1,250       84.98       84.98
Active Union Membership |        221       15.02      100.00
------------------------+-----------------------------------
                  Total |      1,471      100.00
```

So we have Non union members at 84.98% and Active union Members at 15.02%. For the next part to find the mean average hourly wage for union members we use the command:

```
mean hrwage if (age >=40 & age<=49 & union==1)
```

This will calculate the mean of hourly wage given the age is greater than or equal to 40 and less than or equal to 49 and they are a member of a union.

This produces the results:



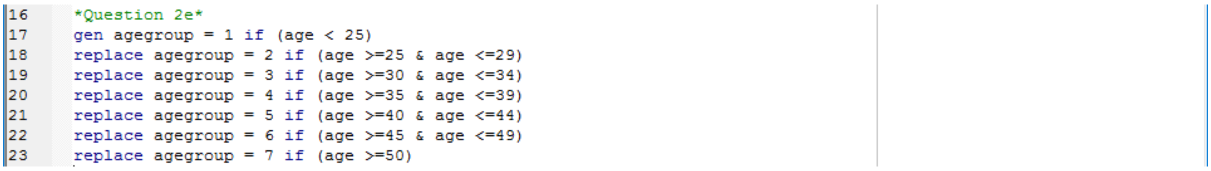
```
Mean estimation                   Number of obs   =         93

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
      hrwage |   13.12881   .6859138      11.76653     14.4911
--------------------------------------------------------------
```

So the mean hourly wage given the age parameters is 13.12881.

Question 2e: I generated a new variable called agegroup and set value = 1 if age was less than 25

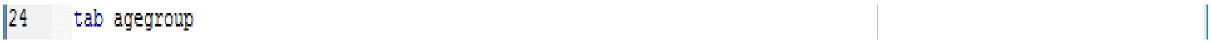
```
gen agegroup = 1 if (age < 25)
```

I the replaced the value of 1 in the new variable depending on their age using the replace command, which changes values in a variable. I used the if command to only let certain ages be replaced with a different value

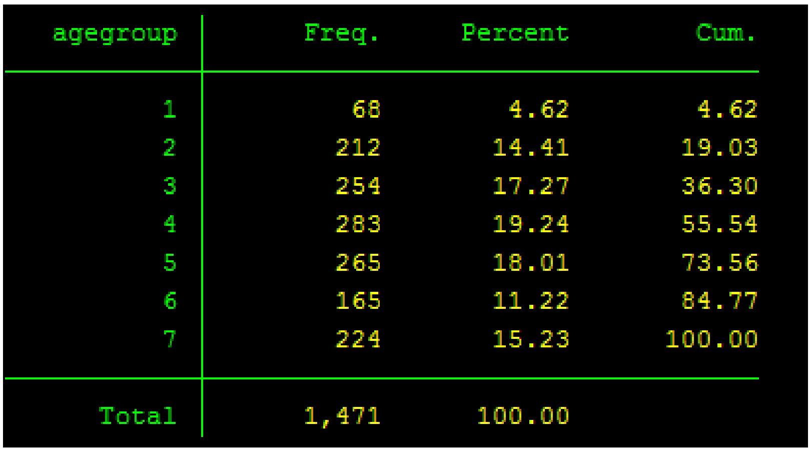


```
*Question 2e*
gen agegroup = 1 if (age < 25)
replace agegroup = 2 if (age >=25 & age <=29)
replace agegroup = 3 if (age >=30 & age <=34)
replace agegroup = 4 if (age >=35 & age <=39)
replace agegroup = 5 if (age >=40 & age <=44)
replace agegroup = 6 if (age >=45 & age <=49)
replace agegroup = 7 if (age >=50)
```

To tabulate the new variable we simply use the tabulate command:



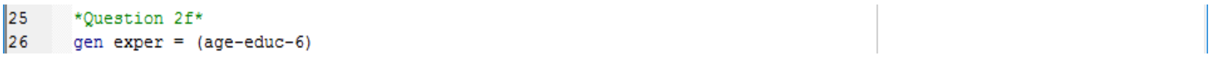
```
tab agegroup
```



```
   agegroup |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         68        4.62        4.62
          2 |        212       14.41       19.03
          3 |        254       17.27       36.30
          4 |        283       19.24       55.54
          5 |        265       18.01       73.56
          6 |        165       11.22       84.77
          7 |        224       15.23      100.00
------------+-----------------------------------
      Total |      1,471      100.00
```

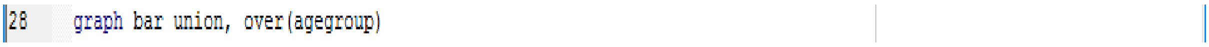
There are only women in the dataset as we are analysing married women, so the number of women in the 30-34 group is simply 254 as age group 3 corresponds to the 30-34 age group.

Question 2f: I generated the new variable using the command:

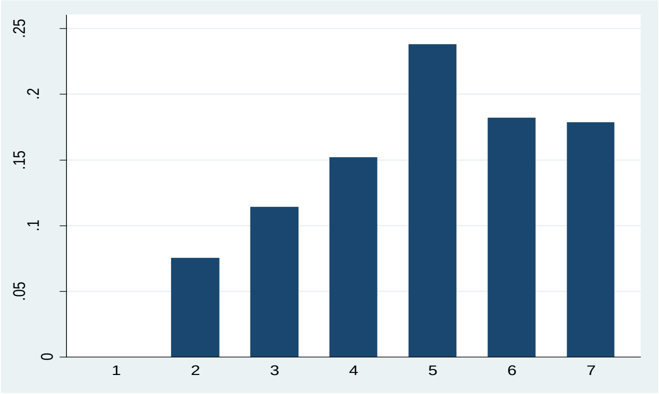


```
*Question 2f*
gen exper = (age-educ-6)
```

Question 3: To create a bar graph I use the command graph bar, this will create a bar graph when given further information. I then use the command mean to use the mean result of union. I used the information from age group 2 to create the chart.

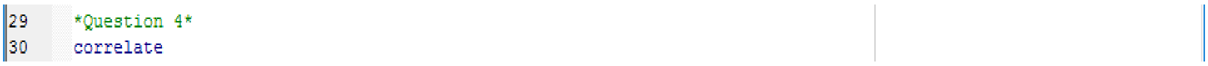


```
graph bar union, over(agegroup)
```

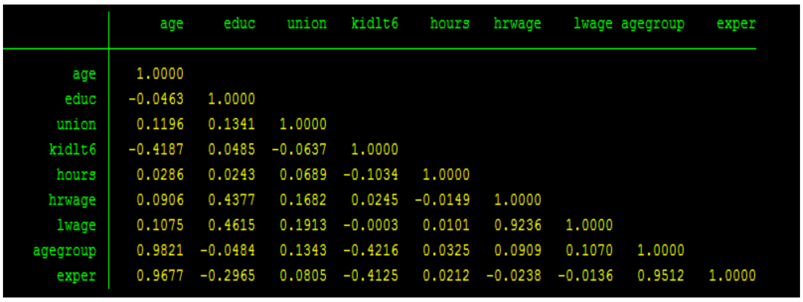


This creates this bar graph, which shows the older age group tends to have higher union membership, age group one for example does not have a single union member. While age group 5 has almost 25% of members in a union .

Question 4: To calculate the correlation coefficient we simply use the correlate command to find the correlation coefficient.



```
*Question 4*
correlate
```

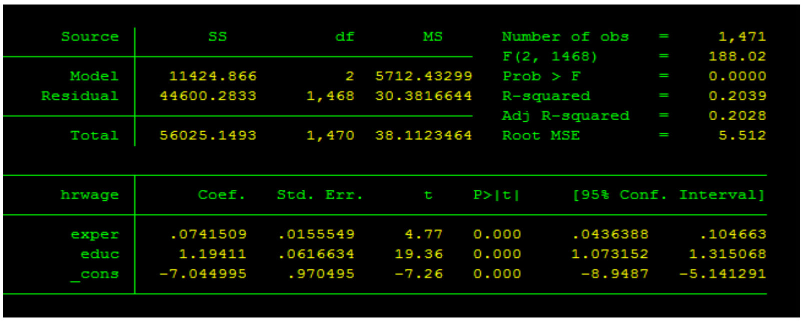


```
             |      age     educ    union   kidlt6    hours   hrwage    lwage agegroup    exper
-------------+---------------------------------------------------------------------------------
         age |   1.0000
        educ |  -0.0463   1.0000
       union |   0.1196   0.1341   1.0000
      kidlt6 |  -0.4187   0.0485  -0.0637   1.0000
       hours |   0.0286   0.0243   0.0689  -0.1034   1.0000
      hrwage |   0.0906   0.4377   0.1682   0.0245  -0.0149   1.0000
       lwage |   0.1075   0.4615   0.1913  -0.0003   0.0101   0.9236   1.0000
    agegroup |   0.9821  -0.0484   0.1343  -0.4216   0.0325   0.0909   0.1070   1.0000
       exper |   0.9677  -0.2965   0.0805  -0.4125   0.0212  -0.0238  -0.0136   0.9512   1.0000
```

As we can see from the table, the result for the correlation coefficient between hourly wage and education is 0.4377. This shows a fair positive correlation between the two variables.

Question 5a: For this question I used the regress command to regress hourly wage, experience and education.

```
*Question 5*
regress hrwage exper educ
```



```
      Source |       SS           df       MS      Number of obs   =     1,471
-------------+----------------------------------   F(2, 1468)      =    188.02
       Model |   11424.866         2  5712.43299   Prob > F        =    0.0000
    Residual |  44600.2833     1,468  30.3816644   R-squared       =    0.2039
-------------+----------------------------------   Adj R-squared   =    0.2028
       Total |  56025.1493     1,470  38.1123464   Root MSE        =     5.512

------------------------------------------------------------------------------
      hrwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0741509   .0155549     4.77   0.000     .0436388     .104663
        educ |    1.19411   .0616634    19.36   0.000     1.073152    1.315068
       _cons |  -7.044995    .970495    -7.26   0.000      -8.9487   -5.141291
------------------------------------------------------------------------------
```

As we can see from the bottom table we have a constant value, or alpha value, as -7.044995. Our two beta values, experience and education will be denoted as β1 and β2 respectively are 0.0741509 and 1.19411.

Our general equation:

$$\widehat{hrwage} = \alpha + \ \beta_1 exper + \beta_2 educ + \in$$

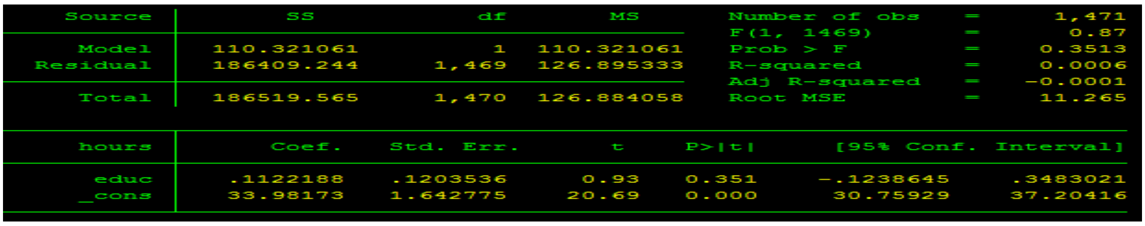
Substituting in my values

$$\widehat{hrwage} = \ -7.044995 + \ 0.0741509 exper + 1.19411 educ + \in$$

This regression does not make much sense; hourly wage would be negative if experience and education were zero, which is obviously impossible. Experience also has a very low effect on the hourly wage which in practice is probably incorrect. Education has an extremely high effect. A 1 unit change in experience causes a 0.0741509 change to hourly wage, and a one unit change in education causes a 1.19411 unit change in hourly wage.

Question 5b: For this question I ran a new regression on hours worked to education to find the new linear regression

```
regress hours educ
```



```
      Source |       SS           df       MS      Number of obs   =     1,471
-------------+----------------------------------   F(1, 1469)      =      0.87
       Model |  110.321061         1  110.321061   Prob > F        =    0.3513
    Residual |  186409.244     1,469  126.895333   R-squared       =    0.0006
-------------+----------------------------------   Adj R-squared   =   -0.0001
       Total |  186519.565     1,470  126.884058   Root MSE        =    11.265

------------------------------------------------------------------------------
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1122188   .1203536     0.93   0.351    -.1238645    .3483021
       _cons |   33.98173   1.642775    20.69   0.000     30.75929    37.20416
```

The coefficient for education is 0.1122188, so when we have a three unit change from 8 to 11 we have:

$\widehat{hours} = \alpha + \ \beta_1 educ{+}\in,\ \widehat{hours} = 33.98 + \ 0.11 educ{+}\in$

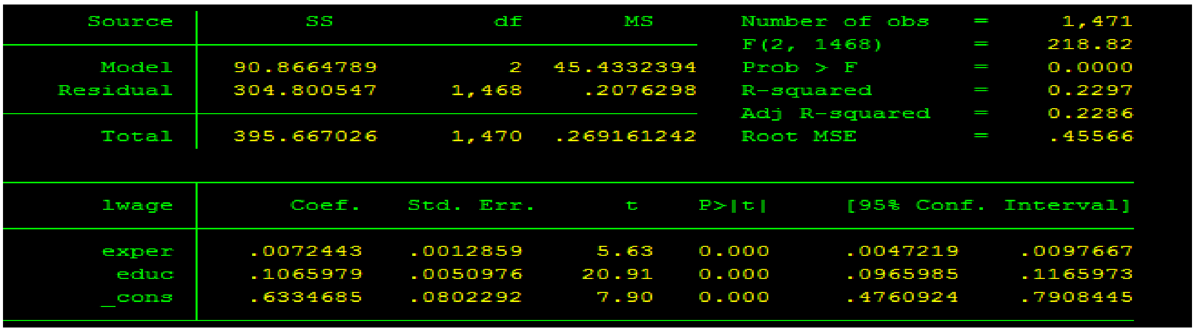
So hours changes by 0.337 per week from a 3 unit education change.

Question 5c: A three unit change to education will cause a change of the same amount as in Question 5b, so the change will be 0.337 per week.

Question 6a: To run this regression we use the regress command again

```
*Question 6*
regress lwage exper educ
```

Producing these results:



```
      Source |       SS           df       MS      Number of obs   =     1,471
-------------+----------------------------------   F(2, 1468)      =    218.82
       Model |  90.8664789         2  45.4332394   Prob > F        =    0.0000
    Residual |  304.800547     1,468   .2076298   R-squared       =    0.2297
-------------+----------------------------------   Adj R-squared   =    0.2286
       Total |  395.667026     1,470  .269161242   Root MSE        =    .45566

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0072443   .0012859     5.63   0.000     .0047219    .0097667
        educ |   .1065979   .0050976    20.91   0.000     .0965985    .1165973
       _cons |   .6334685   .0802292     7.90   0.000     .4760924    .7908445
------------------------------------------------------------------------------
```

We formulate the general equation:

$\widehat{lnwage} = 0.633 + \ 0.007 + 0.1066 educ{+}\in$

Question 6b: If the education increases by one point we would see a 10.66% Increase in hourly wage**.**

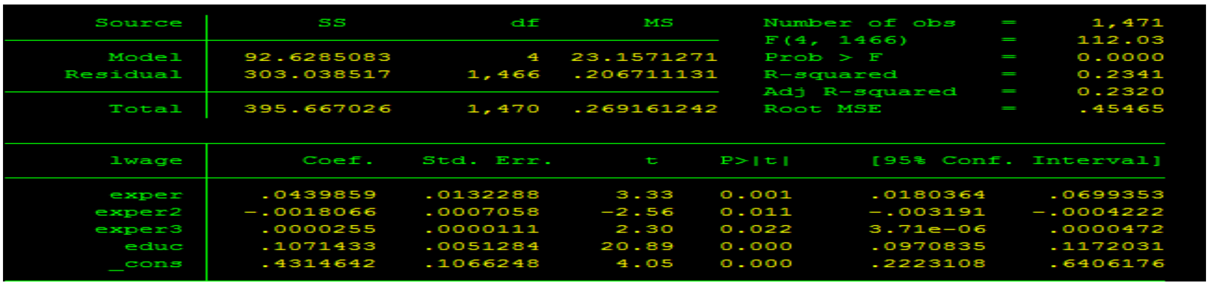
Question 6c: If education increases by one point we would also see a 10.66% increase in hourly wage.

Question 7a: To answer this question I first generated two new variables for exper^2 and exper^3 and labelled them as exper2 and exper3 respectively.

```
gen exper2 = (exper^2)
gen exper3 = (exper^3)
```

I then regressed lwage against the new variables and education

```
regress lwage exper exper2 exper3 educ
```



```
      Source |       SS           df       MS      Number of obs   =     1,471
-------------+----------------------------------   F(4, 1466)      =    112.03
       Model |  92.6285083         4  23.1571271   Prob > F        =    0.0000
    Residual |  303.038517     1,466  .206711131   R-squared       =    0.2341
-------------+----------------------------------   Adj R-squared   =    0.2320
       Total |  395.667026     1,470  .269161242   Root MSE        =    .45465

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0439859   .0132288     3.33   0.001     .0180364    .0699353
      exper2 |  -.0018066   .0007058    -2.56   0.011     -.003191   -.0004222
      exper3 |   .0000255   .0000111     2.30   0.022     3.71e-06    .0000472
        educ |   .1071433   .0051284    20.89   0.000     .0970835    .1172031
       _cons |   .4314642   .1066248     4.05   0.000     .2223108    .6406176
------------------------------------------------------------------------------
```

This creates the general regression:

$\widehat{lwage} = \alpha + \ \beta_1 exper + \beta_2 exper2 + \beta_3 exper3 + \beta_4 educ{+}\in$

Putting in our beta and alpha values:

$$\widehat{lwage} = 0.4315 + \ 0.0440exper - 0.002exper2 + 0.00003exper3 + 0.1071educ{+}\in$$

We see that our constant level is 0.4315; this is viable as when we take the exponential of this we get 1.54. This may seem as though this does not makes sense due to it being below the level of minimum wage, however we can see values in the dataset that are below this so it could make sense. Our beta values for exper, exper2 and exper3 show a weak positive correlation tending towards zero and education shows a relatively strong positive correlation between itself and lwage.

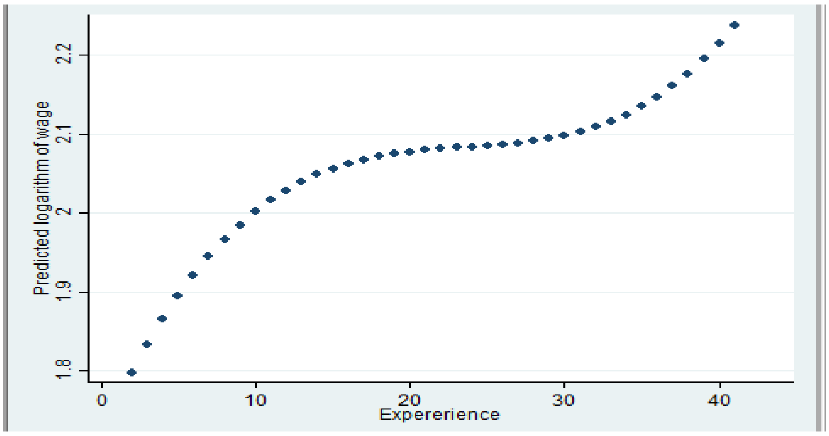
Question 7b: To plot the regression on experience with education held constant 12 years I used the command line:

```
predict lwage_hat if educ==12
```

This essentially creates a new variable for lwage using the linear regression model to find the predicted values given a fixed value of education equal to 12.

```
twoway scatter lwage_hat exper
```

This creates a scatter graph with predicted lwage values on the y axis and experience values on the x axis.

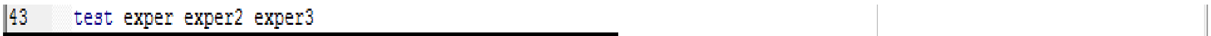


We can see from this graph that in general wage rises with experience especially in the extremes of the dataset.

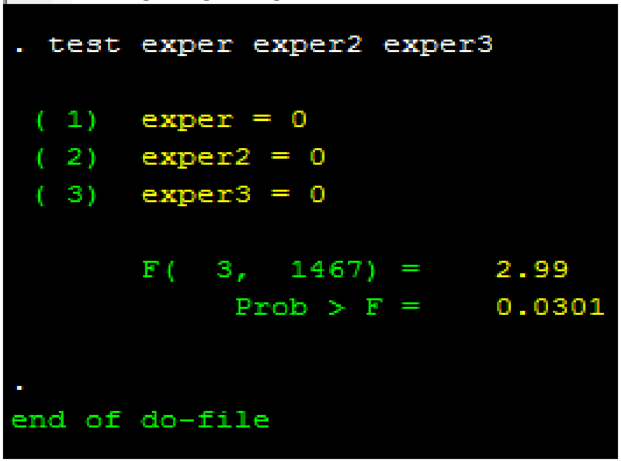
Question 7c: To test if there is more than one zero value in the linear regression we must assume that the samples are randomly selected in an independent manner from the k treatment populations. All k populations have distributions that are approximately normal and the k population variances are equal. We first regress exper exper2 and exper3 against lwage.

```
regress lwage exper exper2 exper3
```

We can then run a joint significance test by using the command line:



```
test exper exper2 exper3
```



```
. test exper exper2 exper3

 ( 1)  exper = 0
 ( 2)  exper2 = 0
 ( 3)  exper3 = 0

       F(  3,  1467) =    2.99
            Prob > F =    0.0301

.
end of do-file
```

This is an F test and from running this test we get the following data.

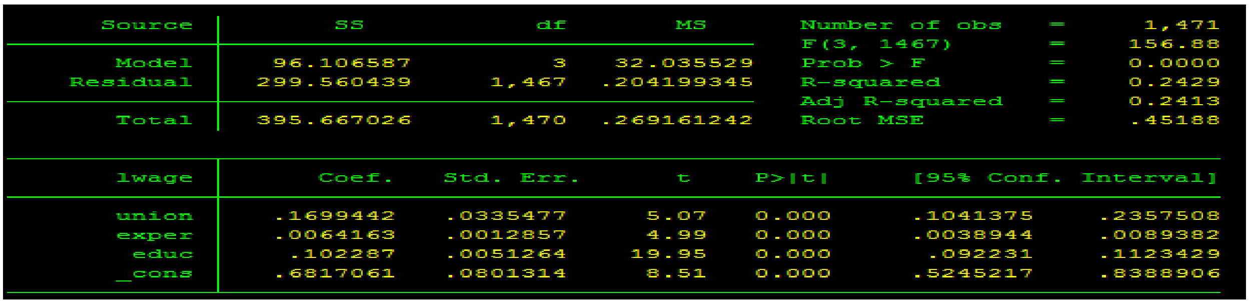
We can see from the table that the critical value of the F test is 2.99. And the corresponding probability that the values are zero is 3.01%. Regarding this data we can say that the test passes at the 5% significance level. So we can say that there is above a 95% probability that there is a linear relationship. *we reject the null hypothesis at the 5% level and accept the alternate to be true and the estimated coefficients are different from zero.*

Question 8: In general it is always true that adding more variables to the right hand side of the regression cannot decrease the explanatory power of the model, only increase or have no effect. Thusly it could be logical to say that the regression in 7(a) is the better model due to this. However the two variables that were added (exper2 and exper3) were both insignificantly different from zero. The change in explanatory power would likely be incredibly insignificant and use of either model would likely cede the same explanatory power, this is easy to see as the values of exper exper2 and exper3 are intrinsically linked due to the fact they are raised to the first, second and third power. However we also see that using the exper as a quadratic could better predict the effect on lwage than having exper as a linear variable. As higher values of exper may correspond to more than proportionally high lwage it may in fact be better to use the quadratic transformations instead of the linear variable.

Question 9a: As before we use the following command to create this regression:

```
regress lwage union exper educ
```

Following running this regression we get this regression table:



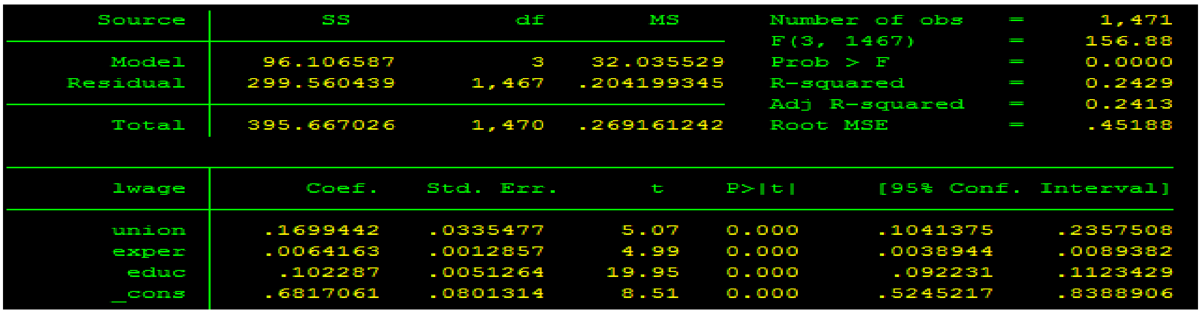
```
      Source |       SS           df       MS      Number of obs   =     1,471
-------------+----------------------------------   F(3, 1467)      =    156.88
       Model |   96.106587         3  32.035529   Prob > F        =    0.0000
    Residual |  299.560439     1,467  .204199345   R-squared       =    0.2429
-------------+----------------------------------   Adj R-squared   =    0.2413
       Total |  395.667026     1,470  .269161242   Root MSE        =    .45188

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   .1699442   .0335477     5.07   0.000     .1041375    .2357508
       exper |   .0064163   .0012857     4.99   0.000     .0038944    .0089382
        educ |    .102287   .0051264    19.95   0.000      .092231    .1123429
       _cons |   .6817061   .0801314     8.51   0.000     .5245217    .8388906
------------------------------------------------------------------------------
```

From this we get the general linear equation

$$\widehat{lwage} = 0.6817 + \ 0.1699union + 0.0064exper + 0.1022educ{+}\in$$

We can see from this that we have a constant value equal to 0.6817, this is the value lwage takes when union, experience and education equal zero. The beta values give the unit change in lwage given the same change to the beta value and show how strong the relationship between lwage and the other variable is.

Question 9b: We can produce a t-test from our previous table:



```
      Source |       SS           df       MS      Number of obs   =     1,471
-------------+----------------------------------   F(3, 1467)      =    156.88
       Model |   96.106587         3  32.035529   Prob > F        =    0.0000
    Residual |  299.560439     1,467  .204199345   R-squared       =    0.2429
-------------+----------------------------------   Adj R-squared   =    0.2413
       Total |  395.667026     1,470  .269161242   Root MSE        =    .45188

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   .1699442   .0335477     5.07   0.000     .1041375    .2357508
       exper |   .0064163   .0012857     4.99   0.000     .0038944    .0089382
        educ |    .102287   .0051264    19.95   0.000      .092231    .1123429
       _cons |   .6817061   .0801314     8.51   0.000     .5245217    .8388906
```

Our test will be $H_0 : \beta_{Union} = 0$, $H_1: \beta_{Union} \neq 0$. Our test statistic will be $\beta_{Union}/se(\beta_{Union})$

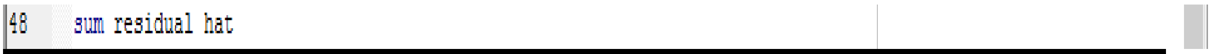
= 5.07, and the test statistic for the 1% confidence level at a normal distribution will be 2.33. As we can see our test statistic will easily pass the ttest and we can reject the H_0 in favour of $H_{A.}$

Question 9c: The old value of R^2 was 0.2297 whereas the new value of R^2 will be 0.2429 this is a difference of 0.0132 or a 5.75% increase on the explanatory power of the model.

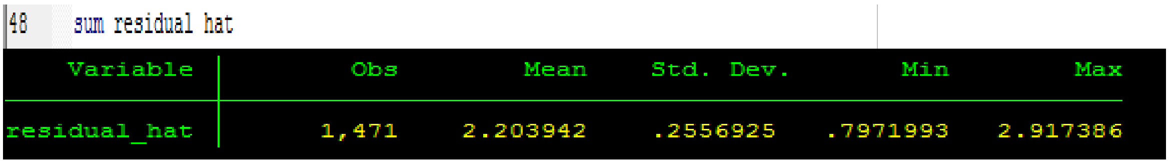
Question 9d: To create a prediction for the residuals we use the command:

```
predict residual_hat
```

This creates a new variable for the linear regression predictions of the residual values. We then simply use the sum command to find the descriptive statistics:



```
sum residual hat
```



```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
residual_hat |      1,471    2.203942    .2556925   .7971993   2.917386
```

We have a mean value of 2.2 and a standard deviation of 0.26.

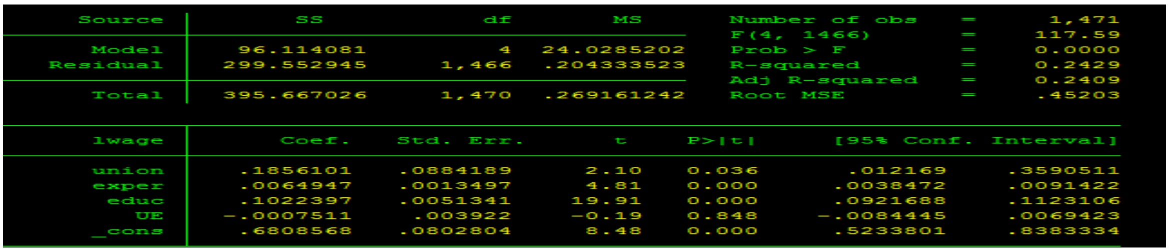
Question 9e: To find the amount of education needed to offset a one unit change from union to non-union we simply divide the union coefficient by the education coefficient: 0.1699442/0.102287 = 1.6614 years of education needed to offset someone joining a union.

Question 10a: To add the interaction term into the model I generate a new variable called UE and defined it as the union*exper

```
gen UE = union*exper
```

I then made a new regression with this variable.

```
reg lwage union exper educ UE
```

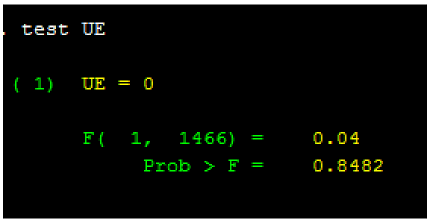


```
      Source |       SS           df       MS      Number of obs   =     1,471
-------------+----------------------------------   F(4, 1466)      =    117.59
       Model |   96.114081         4  24.0285202   Prob > F        =    0.0000
    Residual |  299.552945     1,466  .204333523   R-squared       =    0.2429
-------------+----------------------------------   Adj R-squared   =    0.2409
       Total |  395.667026     1,470  .269161242   Root MSE        =    .45203

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   .1856101   .0884189     2.10   0.036      .012169    .3590511
       exper |   .0064947   .0013497     4.81   0.000     .0038472    .0091422
        educ |   .1022397   .0051341    19.91   0.000     .0921688    .1123106
          UE |  -.0007511    .003922    -0.19   0.848    -.0084445    .0069423
       _cons |   .6808568   .0802804     8.48   0.000     .5233801    .8383334
```

The coefficient of the interaction term UE measures whether or not experience has an effect on lwage if they are a member of a union. The reason for this is so we can test whether or not experience has an effect on lwage depending on union membership.

Question 10b: To test this I ran the test command to find the probability it was zero.

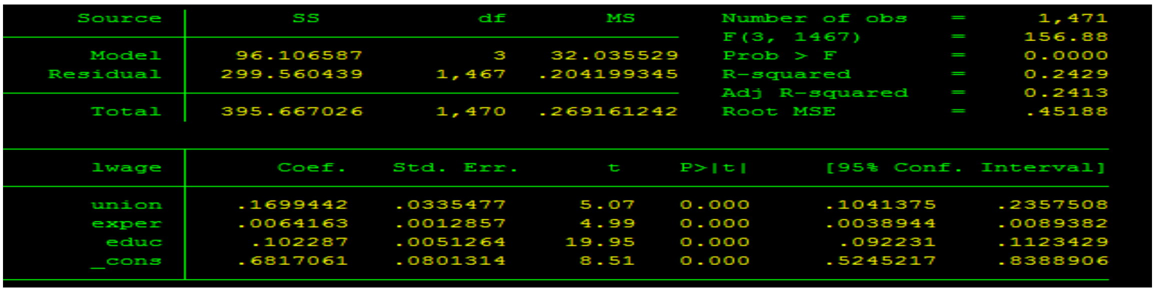
```
test UE
```



```
 test UE

 ( 1)   UE = 0

        F(  1,  1466) =    0.04
             Prob > F =    0.8482
```

As we can see there is an 84.8% probability that our value for the coefficient of UE is zero. This is obviously an insignificant result, we cannot put any faith in the result for the coefficient of UE and therefore the effect of experience on lwage does not change for different values of union.

Question 10c: If we run the regression: reg lwage educ exper union we get the table:



```
      Source |       SS           df       MS      Number of obs   =     1,471
-------------+----------------------------------   F(3, 1467)      =    156.88
       Model |   96.106587         3  32.035529   Prob > F        =    0.0000
    Residual |  299.560439     1,467  .204199345   R-squared       =    0.2429
-------------+----------------------------------   Adj R-squared   =    0.2413
       Total |  395.667026     1,470  .269161242   Root MSE        =    .45188

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   .1699442   .0335477     5.07   0.000     .1041375    .2357508
       exper |   .0064163   .0012857     4.99   0.000     .0038944    .0089382
        educ |    .102287   .0051264    19.95   0.000      .092231    .1123429
       _cons |   .6817061   .0801314     8.51   0.000     .5245217    .8388906
------------------------------------------------------------------------------
```

And we fix values of educ=12, exper=10 and union=1 we get the linear regression:

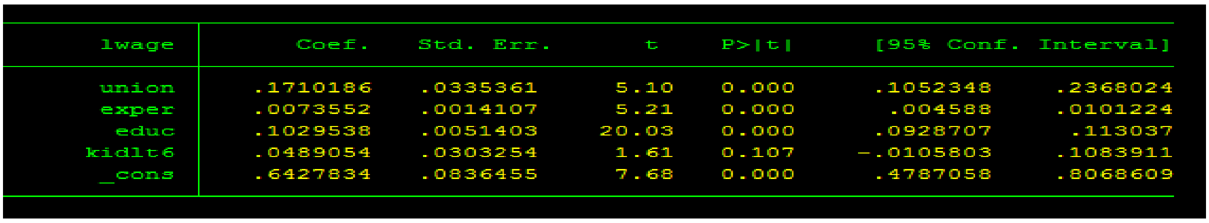
$$\widehat{lwage} = 0.6817 + \ 0.1699 * 1 + 0.0064 * 10 + 0.1022 * 12 {+} \epsilon \quad = 2.142$$

Question 10d:

$$\widehat{lwage} = 0.6817 + \ 0.1699 * 0 + 0.0064 * 15 + 0.1022 * 10 {+} \epsilon \quad = 1.7997$$

Question 11a: Running a new regression with kidlt6:

```
reg lwage union exper educ kidlt6
```

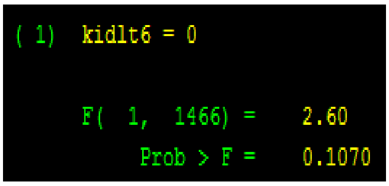


```
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   .1710186   .0335361     5.10   0.000     .1052348    .2368024
       exper |   .0073552   .0014107     5.21   0.000      .004588    .0101224
        educ |   .1029538   .0051403    20.03   0.000     .0928707     .113037
      kidlt6 |   .0489054   .0303254     1.61   0.107    -.0105803    .1083911
       _cons |   .6427834   .0836455     7.68   0.000     .4787058    .8068609
------------------------------------------------------------------------------
```

The estimations of the coefficients are union: 0.1710, experience: 0.007, education: 0.1029, kidlt6: 0.0489 and a constant value of 0.6429.

Question 11b: The presence of a small child will cause in increase to lwage by the beta value of kidlt6, which is 0.0489. No child will simply cause the value of increase to be zero. We produce a significance test using the test command in stata:

```
test kidlt6
```



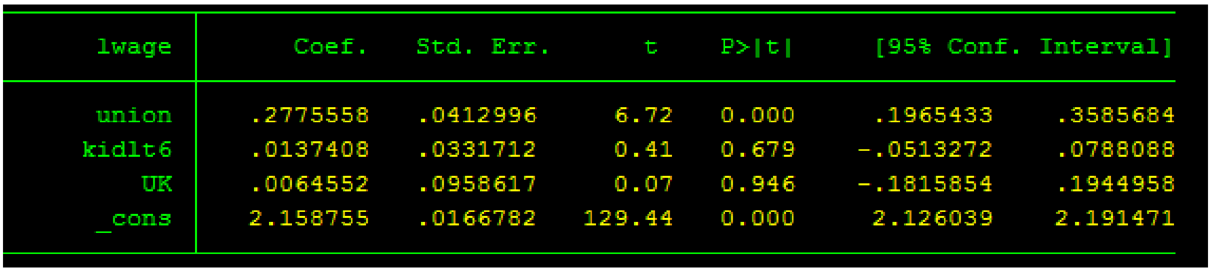
```
 ( 1)   kidlt6 = 0

        F(  1,  1466) =    2.60
             Prob > F =    0.1070
```

We can see that the critical value of our test was 2.60 and the probability that the F result will be greater than this is 89.3% So the probability that kidlt6 is = 0 is 10.7% so our test just fails to pass at the 10% level of significance.

Question 11c: We can use an interaction term between kidlt6 and union as we did before in question 10 to find whether or not parents with young children benefit more from union membership. To do this test we create a new interaction term UK.

```
gen UK = union*kidlt6
```

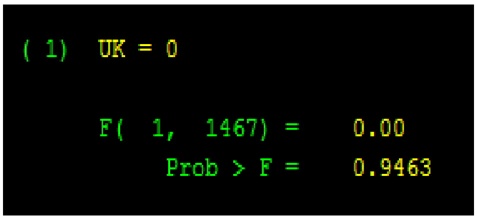
And regress this against lwage with union and kidlt6.



```
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   .2775558   .0412996     6.72   0.000     .1965433    .3585684
      kidlt6 |   .0137408   .0331712     0.41   0.679    -.0513272    .0788088
          UK |   .0064552   .0958617     0.07   0.946    -.1815854    .1944958
       _cons |   2.158755   .0166782   129.44   0.000     2.126039    2.191471
------------------------------------------------------------------------------
```

And then test the UK coefficient

```
test UK
```



```
 ( 1)  UK = 0

       F(  1,  1467) =    0.00
            Prob > F =    0.9463
```

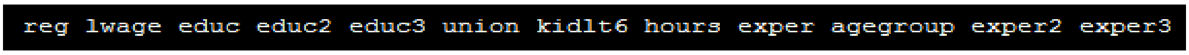
As we see, we cannot reject the H_0 that UK=0 as our probability is 0.9463 which is far greater that the 5% or 10% confidence level, therefore we can say with some certainty that there is no effect on lwage from families with a child under 6 and union membership.

Question 12:

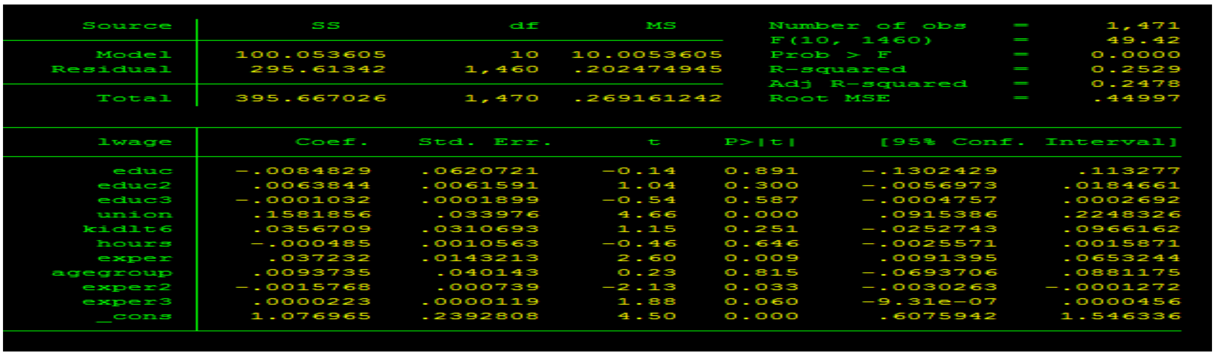
Dependant variable is logarithm of wage			
	6a	9a	11a
Experience (β)	0.0072443	0.0064163	0.0073552
(standard error)	0.0023859**	0.0012857**	0.0014107**
Education (β)	0.1065979	0.102287	0.1029538
(standard error)	0.0050976**	0.0051264**	0.0051403**
Union (β)		0.1699442	0.1710186
(standard error)		0.0335477**	0.0335361**
Kidlt6 (β)			0.0489054
(standard error)			0.0303254
Constant (α)	0.6334685	0.6817061	0.6427834
(standard error)	0.0802292**	0.0801314**	0.0836455**
Observations	1471	1471	1471
R^2 Value	0.2297	0.2429	0.2442
Adjusted R^2 Value	0.2286	0.2413	0.2422
Mean	2.203942	2.203942	2.203942
Standard Deviation	.2486241	.2556925	.2563972

Note: * indicates significance at the 5% level. ** indicates significance at the 1% level

Question 13: From these regressions I have found that the predictions for lwage are best explained when we increase the number of predictor variables. In my tests I found that the regressions in 11a had a higher R^2 value than that of 9a or 6a. This leads me to believe that I could increase the explanatory power of our regression by adding some more variables included in the dataset. Instead of the regression in 11a I ran a new regression with new variables of educ2 and educ3 corresponding to the square and cube of education.

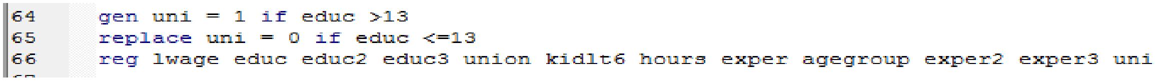


```
reg lwage educ educ2 educ3 union kidlt6 hours exper agegroup exper2 exper3
```



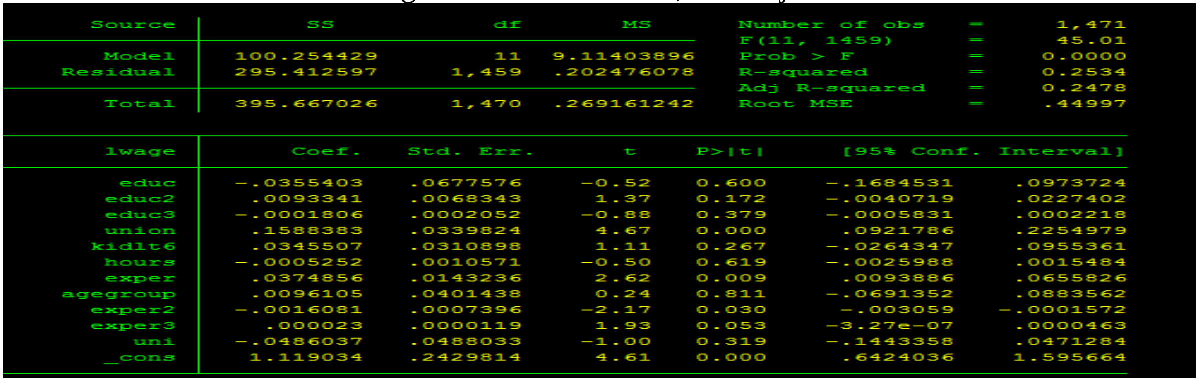
```
      Source |       SS           df       MS      Number of obs   =     1,471
-------------+----------------------------------   F(10, 1460)     =     49.42
       Model |  100.053605        10  10.0053605   Prob > F        =    0.0000
    Residual |   295.61342     1,460  .202474945   R-squared       =    0.2529
-------------+----------------------------------   Adj R-squared   =    0.2478
       Total |  395.667026     1,470  .269161242   Root MSE        =    .44997

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0084829   .0620721    -0.14   0.891    -.1302429     .113277
       educ2 |   .0063844   .0061591     1.04   0.300    -.0056973    .0184661
       educ3 |  -.0001032   .0001899    -0.54   0.587    -.0004757    .0002692
       union |   .1581856    .033976     4.66   0.000     .0915386    .2248326
      kidlt6 |   .0356709   .0310693     1.15   0.251    -.0252743    .0966162
       hours |   -.000485   .0010563    -0.46   0.646    -.0025571    .0015871
       exper |    .037232   .0143213     2.60   0.009     .0091395    .0653244
    agegroup |   .0093735    .040143     0.23   0.815    -.0693706    .0881175
      exper2 |  -.0015768    .000739    -2.13   0.033    -.0030263   -.0001272
      exper3 |   .0000223   .0000119     1.88   0.060    -9.31e-07    .0000456
       _cons |   1.076965   .2392808     4.50   0.000     .6075942    1.546336
------------------------------------------------------------------------------
```

This had the effect of increasing R^2 to 0.2529 and Adj R^2 to 0.2478 now 25.29% of the variation in lwage could be attributed to variation in the explanatory variables, which is quite a bit higher than the other regressions. To further increase the explanatory of the model I created a new dummy variable called uni, with the command:



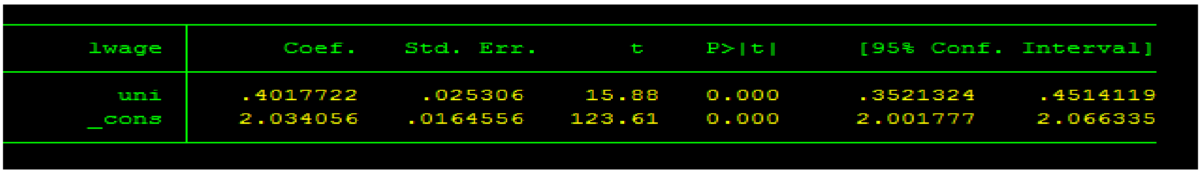
```
gen uni = 1 if educ >13
replace uni = 0 if educ <=13
reg lwage educ educ2 educ3 union kidlt6 hours exper agegroup exper2 exper3 uni
```

This had the effect of increasing the R^2 value further, but Adj R^2 remains constant.



```
      Source |       SS           df       MS      Number of obs   =     1,471
-------------+----------------------------------   F(11, 1459)     =     45.01
       Model |  100.254429        11  9.11403896   Prob > F        =    0.0000
    Residual |  295.412597     1,459  .202476078   R-squared       =    0.2534
-------------+----------------------------------   Adj R-squared   =    0.2478
       Total |  395.667026     1,470  .269161242   Root MSE        =    .44997

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0355403   .0677576    -0.52   0.600    -.1684531    .0973724
       educ2 |   .0093341   .0068343     1.37   0.172    -.0040719    .0227402
       educ3 |  -.0001806   .0002052    -0.88   0.379    -.0005831    .0002218
       union |   .1588383   .0339824     4.67   0.000     .0921786    .2254979
      kidlt6 |   .0345507   .0310898     1.11   0.267    -.0264347    .0955361
       hours |  -.0005252   .0010571    -0.50   0.619    -.0025988    .0015484
       exper |   .0374856   .0143236     2.62   0.009     .0093886    .0655826
    agegroup |   .0096105   .0401438     0.24   0.811    -.0691352    .0883562
      exper2 |  -.0016081   .0007396    -2.17   0.030     -.003059   -.0001572
      exper3 |    .000023   .0000119     1.93   0.053    -3.27e-07    .0000463
         uni |  -.0486037   .0488033    -1.00   0.319    -.1443358    .0471284
       _cons |   1.119034   .2429814     4.61   0.000     .6424036    1.595664
------------------------------------------------------------------------------
```

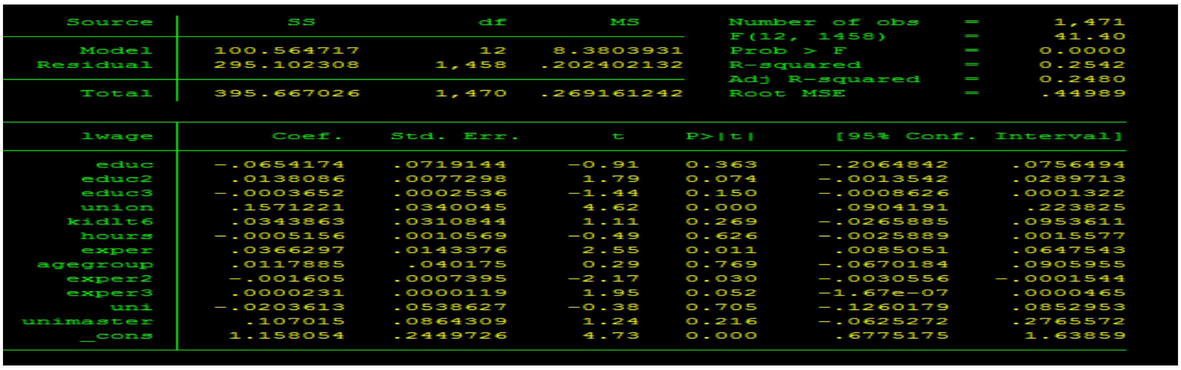
So now 25.34 % of the variation can be attributed to the explanatory variables. I also saw that when I regressed uni on lwage the coefficient for uni was much higher than that of education:



```
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         uni |   .4017722    .025306    15.88   0.000     .3521324    .4514119
       _cons |   2.034056   .0164556   123.61   0.000     2.001777    2.066335
------------------------------------------------------------------------------
```

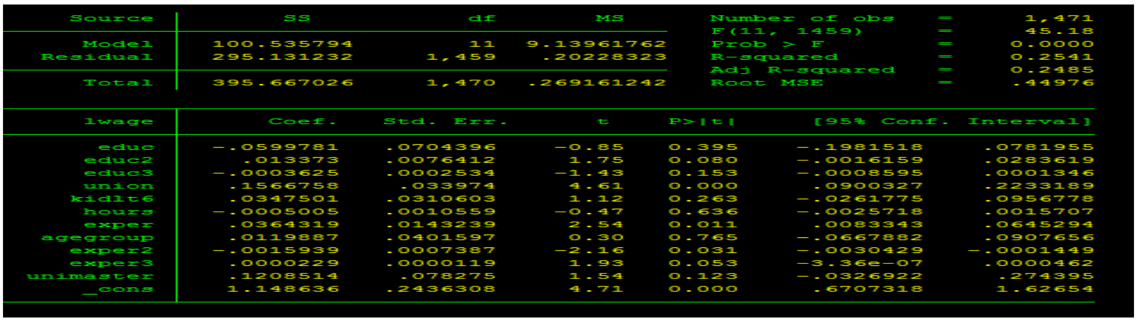
At 0.401 as opposed to 0.098

Following on from this I created a new dummy variable called unimaster if someone had completed enough years of education to complete a master’s course, at 17 years. Generating the variable as before and regressing against lwage gave us:



```
      Source |       SS           df       MS      Number of obs   =     1,471
-------------+----------------------------------   F(12, 1458)     =     41.40
       Model |  100.564717        12  8.3803931    Prob > F        =    0.0000
    Residual |  295.102308     1,458  .202402132   R-squared       =    0.2542
-------------+----------------------------------   Adj R-squared   =    0.2480
       Total |  395.667026     1,470  .269161242   Root MSE        =    .44989

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0654174   .0719144    -0.91   0.363    -.2064842    .0756494
       educ2 |   .0138086   .0077298     1.79   0.074    -.0013542    .0289713
       educ3 |  -.0003652   .0002536    -1.44   0.150    -.0008626    .0001322
       union |   .1571221   .0340045     4.62   0.000     .0904191     .223825
      kidlt6 |   .0343863   .0310844     1.11   0.269    -.0265885    .0953611
       hours |  -.0005156   .0010569    -0.49   0.626    -.0025889    .0015577
       exper |   .0366297   .0143376     2.55   0.011     .0085051    .0647543
    agegroup |   .0117885    .040175     0.29   0.769    -.0670184    .0905955
      exper2 |   -.001605   .0007395    -2.17   0.030    -.0030556   -.0001544
      exper3 |   .0000231   .0000119     1.95   0.052    -1.67e-07    .0000465
         uni |  -.0203613   .0538627    -0.38   0.705    -.1260179    .0852953
   unimaster |    .107015   .0864309     1.24   0.216    -.0625272    .2765572
       _cons |   1.158054   .2449726     4.73   0.000     .6775175     1.63859
------------------------------------------------------------------------------
```

This increased our R^2 value even further to 25.42%. However as we can see the adjusted R^2 remained constant with the introduction of uni, but rose with unimaster. So when we exclude uni but include unimaster our regression actually has more explanatory power under the adjusted R^2 measure!



```
      Source |       SS           df       MS      Number of obs   =     1,471
-------------+----------------------------------   F(11, 1459)     =     45.18
       Model |  100.535794        11  9.13961762   Prob > F        =    0.0000
    Residual |  295.131232     1,459  .20228323    R-squared       =    0.2541
-------------+----------------------------------   Adj R-squared   =    0.2485
       Total |  395.667026     1,470  .269161242   Root MSE        =    .44976

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0599781   .0704396    -0.85   0.395    -.1981518    .0781955
       educ2 |    .013373   .0076412     1.75   0.080    -.0016159    .0283619
       educ3 |  -.0003625   .0002534    -1.43   0.153    -.0008595    .0001346
       union |   .1566758    .033974     4.61   0.000     .0900327    .2233189
      kidlt6 |   .0347501   .0310603     1.12   0.263    -.0261775    .0956778
       hours |  -.0005005   .0010559    -0.47   0.636    -.0025718    .0015707
       exper |   .0364319   .0143239     2.54   0.011     .0083343    .0645294
    agegroup |   .0119887   .0401597     0.30   0.765    -.0667882    .0907656
      exper2 |  -.0015939   .0007387    -2.16   0.031    -.0030429   -.0001449
      exper3 |   .0000229   .0000119     1.93   0.053    -3.36e-07    .0000462
   unimaster |   .1208514    .078275     1.54   0.123    -.0326922     .274395
       _cons |   1.148636   .2436308     4.71   0.000     .6707318     1.62654
------------------------------------------------------------------------------
```

Our Adjusted R^2 value is now 24.85, the highest I have achieved, by adding educ2 educ3 and unimaster.

From running the regressions including interaction variables I found that the effect of experience on lwage depending on union membership was insignificant, in other words, people in a union did not have a higher effect of experience on lwage. We also found the same for kidlt6, there is no effect on lwage from families with a child under the age of 6 and a union membership.

Many variables could be added to the dataset to increase the explanatory power of the model, of those with the highest probable increase would be variables such as region and blue collar or white collar workers.

In conclusion these regressions showed us what variables had the greatest unit effect on lwage, such as education and union, and which variables had no little significance such as kidlt6. We also went deeper with interaction variables to find effect on lwage given x. By adding more explanatory variables such as unimaster we managed to achieve a higher adjusted R^2 value than what would otherwise have been possible.