Exercise 6

Problem 1
The file *stockton96.gdt* contains 940 observations on home sales in Stockton, CA in 1996.

a) Use least squares to estimate a linear equation that relates house price *PRICE* to the size of the house in square feet *SQFT* and the age of the house in years *AGE*. Interpret all the estimates.
ols price const age sqft

```
Model 1: OLS, using observations 1-940
Dependent variable: price

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       5193.15      3586.64        1.448    0.1480
  age         -217.843       35.0976     -6.207    8.11e-010 ***
  sqft          68.3907       2.16868    31.54     2.39e-149 ***

Mean dependent var   97937.83   S.D. dependent var   34179.37
Sum squared resid    4.76e+11   S.E. of regression   22539.63
R-squared            0.566050   Adjusted R-squared   0.565124
F(2, 937)            611.1178   P-value(F)           1.4e-170
Log-likelihood      -10753.95   Akaike criterion     21513.90
Schwarz criterion    21528.43   Hannan-Quinn         21519.44
```

b) Suppose that you own two houses. One has 1400 square feet; the other has 1800 square feet. Both are 20 years old. What price do you estimate you will get for each house?

$$\widehat{p_1} = 5193 + 20 * (-217) + 68.39 * 1400$$
$$\widehat{p_2} = 5193 + 20 * (-217) + 68.39 * 1800$$

c) Test the hypothesis that the size and the age of the house are important determinants of its price (separately as well as jointly). **Both have three stars. Also jointly significant according to above output**

d) Using the Breusch-Pagan test for heteroscedasticity, test whether the model satisfies the homoscedasticity assumption by using the command for the BP test in Gretl.

You could certainly use software to do the test for you, which will be
modtest --breusch-pagan

according to the test, LM test statistic is very large 148 as well as the P-value is extremely small, therefore, you are rejecting the H_0 hypothesis that there is no heteroskedasticity:

```
? modtest --breusch-pagan

Breusch-Pagan test for heteroskedasticity
OLS, using observations 1-940
Dependent variable: scaled uhat^2

             coefficient   std. error    t-ratio   p-value
  ---------------------------------------------------------
  const      -1,54079      0,480089      -3,209    0,0014    ***
  sqft        0,00165699   0,000290289    5,708    1,53e-08  ***
  age         0,00542698   0,00469798     1,155    0,2483

  Explained sum of squares = 297,816

Test statistic: LM = 148,907753,
with p-value = P(Chi-square(2) > 148,907753) = 0,000000
```

You could do the test also with more manual way and it is important to be able to do so, because BP test in the software tests heteroskedasticity for all the variables at the same time in your regression. If you are asked to test for heteroskedasticity by just one variable for example in your multivariate regression, then standard BP test will not do it (at least I could not find appropriate command in Gretl, Stata has it). For that matter we need to do several steps:

Step 1. Run original regression

ols price const age sqft

Step 2. Generate residuals and its squares

series resid=$uhat

genr sq_resid=resid^2

Step 3. Run regression of squared residuals on the explanatory variable(s) of interest

ols sq_resid sqft age const

```
Model 5: OLS, using observations 1-940
Dependent variable: sq_resid

             coefficient        std. error     t-ratio  p-value
  --------------------------------------------------------------
  const      -7,80277e+08      2,43124e+08    -3,209    0,0014    ***
  sqft       839121            147006          5,708    1,53e-08  ***
  age         2,74830e+06      2,37912e+06     1,155    0,2483

Mean dependent var   5,06e+08   S.D. dependent var   1,55e+09
Sum squared resid    2,19e+21   S.E. of regression   1,53e+09
R-squared            0,033740   Adjusted R-squared   0,031677
F(2, 937)            16,35898   P-value(F)           1,04e-07
Log-likelihood      -21210,61   Akaike criterion     42427,22
Schwarz criterion    42441,76   Hannan-Quinn         42432,76
```

Step 4. Derive LM test statistic by taking R^2 from the regression in step 3 and multiplying it by the number of observations

In this case, LM=0.0337*940=31.68

Step 5. Find critical value in the $\chi^2(2)$ distribution table which at 1% significance level will be 9.21 and we can again reject the H_0

e) Use the White test to test for heteroskedasticity.

You could certainly use software to do the test for you, which will be

modtest –white

Don’t forget to re-run the original regression before doing the test

```
White's test for heteroskedasticity
OLS, using observations 1-940
Dependent variable: uhat^2

              coefficient       std. error     t-ratio   p-value
  ---------------------------------------------------------------
  const          4,47842e+08      9,36332e+08   0,4783   0,6326
  sqft      -878575               1,19718e+06  -0,7339   0,4632
  age        144565               1,19978e+07   0,01205  0,9904
  sq_sqft       598,200         381,040         1,570    0,1168
  X2_X3       -2063,04         6764,87         -0,3050   0,7605
  sq_age      65307,8         86504,5           0,7550   0,4505

  Unadjusted R-squared = 0,037499

Test statistic: TR^2 = 35,248808,
with p-value = P(Chi-square(5) > 35,248808) = 0,000001
```

according to the test, LM test statistic is very large 35.25 as well as the P-value is extremely small, therefore, you are rejecting the H_0 hypothesis that there is no heteroskedasticity.

Manual version:

Step 1. Run original regression
ols price const age sqft
Step 2. Generate residuals and its squares
series resid=$uhat
genr sq_resid=resid^2
Step 3. Generate squares and interaction terms of the explanatory variables
genr sq_sqft=sqft^2
genr sq_age=age^2
genr agesqft=sqft*age
Step 4. Run regression of squared residuals on the explanatory variable(s), their squared terms and the interaction terms

```
? ols sq_resid const sq_sqft sqft age sq_age agesqft

Model 9: OLS, using observations 1-940
Dependent variable: sq_resid

             coefficient       std. error     t-ratio   p-value
  ---------------------------------------------------------------
  const          4,47842e+08      9,36332e+08   0,4783    0,6326
  sq_sqft      598,200          381,040         1,570     0,1168
  sqft     -878575                1,19718e+06  -0,7339    0,4632
  age       144565                1,19978e+07   0,01205   0,9904
  sq_age     65307,8          86504,5           0,7550    0,4505
  agesqft    -2063,04          6764,87         -0,3050    0,7605

Mean dependent var   5,06e+08   S.D. dependent var   1,55e+09
Sum squared resid    2,18e+21   S.E. of regression   1,53e+09
R-squared            0,037499   Adjusted R-squared   0,032346
F(5, 934)            7,277666   P-value(F)           1,06e-06
Log-likelihood      -21208,78   Akaike criterion     42429,56
Schwarz criterion    42458,63   Hannan-Quinn         42440,64

Excluding the constant, p-value was highest for variable 3 (age)
```

Step 5. Derive LM test statistic by taking R^2 from the regression in step 4 and multiplying it by the number of observations

In this case, LM=0.03749*940=35.24 (just like in the software version, yey)

Step 6. Find critical value in the $\chi^2(5)$ distribution table which at 1% significance level will be 15.09 and we can again reject the H_0

f) What do you conclude regarding the heteroskedasticity? Does your conclusion depend on the choosing a specific test? Discuss also drawbacks of the BP and White tests.
There is heteroskedasticity
A weakness of the BP test is that it assumes the heteroskedasticity is a linear function of the independent variables. Failing to find evidence of heteroskedasticity with the BP doesn't rule out a nonlinear relationship between the independent variable(s) and the error variance.
The weakness of white test is that if you have many variables, the number of possible interactions plus the squared variables plus the original variables can be quite high.

g) Test the hypothesis that the size and the age of the house are important determinants of its price (separately as well as jointly). Hint: choose appropriate standard errors. Does your conclusion differ from part (c)?
ols price const age sqft –robust
compare the robust and non-robust standard errors and parameters. You can see that the parameters did not change, while standard errors increased. Still, conclusions have not changed, based on the F-statistic

```
? ols price const sqft age --robust

Model 10: OLS, using observations 1-940
Dependent variable: price
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       5193.15      3648.56       1.423    0.1550
  sqft          68.3907       2.46807   27.71     6.35e-124 ***
  age         -217.843       36.3142    -5.999    2.84e-09  ***

Mean dependent var   97937.83   S.D. dependent var   34179.37
Sum squared resid    4.76e+11   S.E. of regression   22539.63
R-squared            0.566050   Adjusted R-squared   0.565124
F(2, 937)            476.5571   P-value(F)           1.7e-143
Log-likelihood      -10753.95   Akaike criterion     21513.90
Schwarz criterion    21528.43   Hannan-Quinn         21519.44

? ols price const sqft age

Model 11: OLS, using observations 1-940
Dependent variable: price

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       5193.15      3586.64       1.448    0.1480
  sqft          68.3907       2.16868   31.54     2.39e-149 ***
  age         -217.843       35.0976    -6.207    8.11e-010 ***

Mean dependent var   97937.83   S.D. dependent var   34179.37
Sum squared resid    4.76e+11   S.E. of regression   22539.63
R-squared            0.566050   Adjusted R-squared   0.565124
F(2, 937)            611.1178   P-value(F)           1.4e-170
Log-likelihood      -10753.95   Akaike criterion     21513.90
Schwarz criterion    21528.43   Hannan-Quinn         21519.44
```

Problem 2

Using the data in *cps4_small.gdt* estimate the following wage equation with least squares and heteroskedasticity-robust standard errors:

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + \beta_5 (EXPER x EDUC) + e$$

(a) Report the results.

genr exper2=exper^2
genr experedu=exper*educ
genr lnwage=ln(wage)
ols lnwage educ exper exper2 experedu const --robust

```
? ols lnwage educ exper exper2 experedu const --robust

Model 4: OLS, using observations 1-1000
Dependent variable: lnwage
Heteroskedasticity-robust standard errors, variant HC1

             coefficient    std. error    t-ratio    p-value
  ---------------------------------------------------------------
  const       0.529677      0.252825       2.095    0.0364    **
  educ        0.127195      0.0169597      7.500    1.41e-013 ***
  exper       0.0629807     0.0113775      5.536    3.97e-08  ***
  exper2     -0.000713939   9.20134e-05   -7.759    2.11e-014 ***
  experedu   -0.00132239    0.000636794   -2.077    0.0381    **

Mean dependent var   2.856988   S.D. dependent var   0.580619
Sum squared resid    254.4216   S.E. of regression   0.505668
R-squared            0.244548   Adjusted R-squared   0.241511
F(4, 995)            85.06746   P-value(F)           3.57e-62
Log-likelihood      -734.5572   Akaike criterion     1479.114
Schwarz criterion    1503.653   Hannan-Quinn         1488.441
```

(b) Add MARRIED to the equation and re-estimate. Holding education and experience constant, do married workers get higher wages? Using a 5% significance level, test a null hypothesis that wages of married workers are less than or equal to those of unmarried workers against the alternative that wages of married workers are higher.

```
? ols lnwage educ exper exper2 experedu married const --robust

Model 5: OLS, using observations 1-1000
Dependent variable: lnwage
Heteroskedasticity-robust standard errors, variant HC1

             coefficient    std. error    t-ratio    p-value
  ---------------------------------------------------------------
  const       0.541061      0.254209       2.128    0.0335    **
  educ        0.126120      0.0170564      7.394    3.02e-013 ***
  exper       0.0613731     0.0115877      5.296    1.45e-07  ***
  exper2     -0.000693346   9.55671e-05   -7.255    8.07e-013 ***
  experedu   -0.00130912    0.000638420   -2.051    0.0406    **
  married     0.0402895     0.0339231      1.188    0.2352

Mean dependent var   2.856988   S.D. dependent var   0.580619
Sum squared resid    254.0582   S.E. of regression   0.505561
R-squared            0.245627   Adjusted R-squared   0.241833
F(5, 994)            69.11228   P-value(F)           4.41e-62
Log-likelihood      -733.8426   Akaike criterion     1479.685
Schwarz criterion    1509.132   Hannan-Quinn         1490.877
```

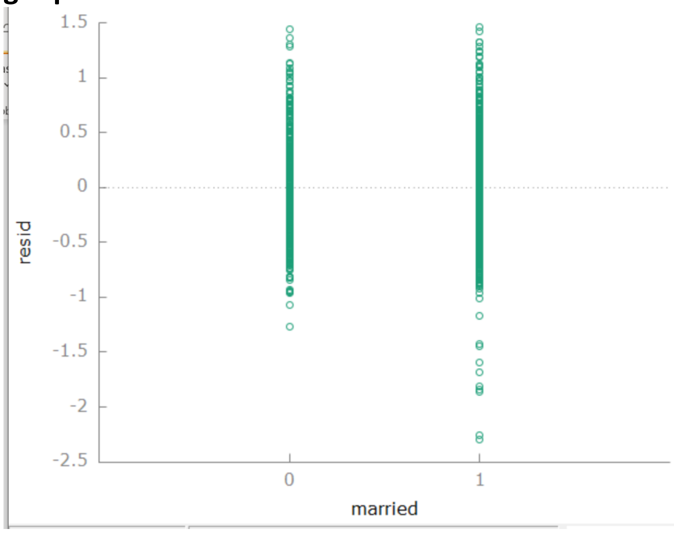
The null and alternative hypotheses for testing whether married workers get higher wages are given by

$$H_0: \beta_6 \leq 0$$
$$H_1: \beta_6 > 0$$

The test value is: 1.188, the critical value at the 5% level of significance is 1.646. Since the test value is less than the critical value, we do not reject the null hypothesis at the 5% level. We conclude that there is insufficient evidence to show that wages of married workers are greater than those of unmarried workers.

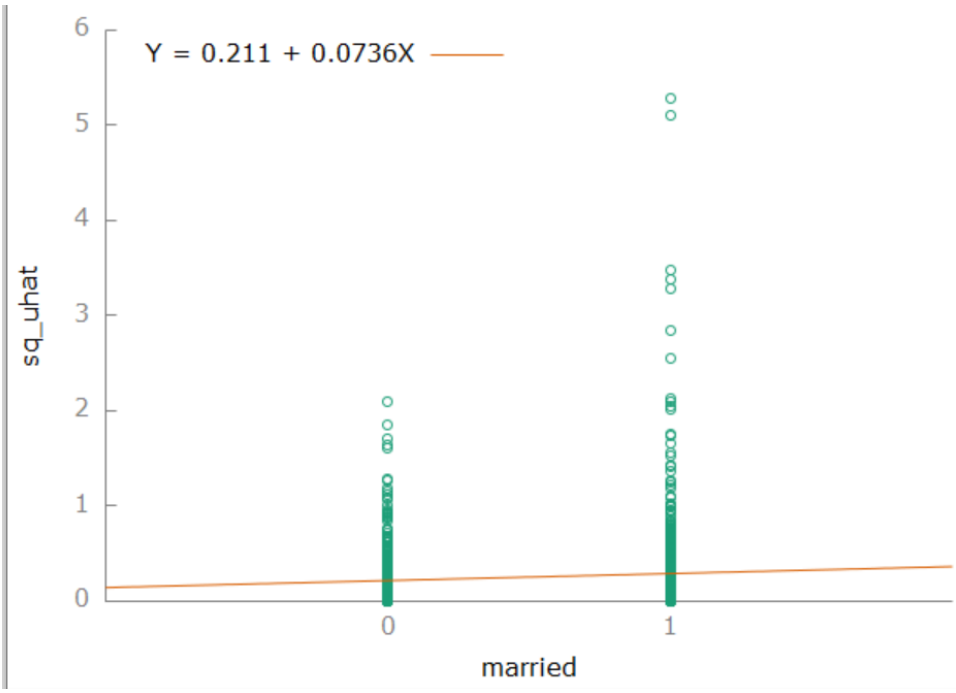
(c) Plot the residuals from part (a) against the two values of MARRIED. Is there evidence of heteroskedasticity?

series uhat=$uhat
genr sq_uhat=uhat^2
gnuplot uhat married

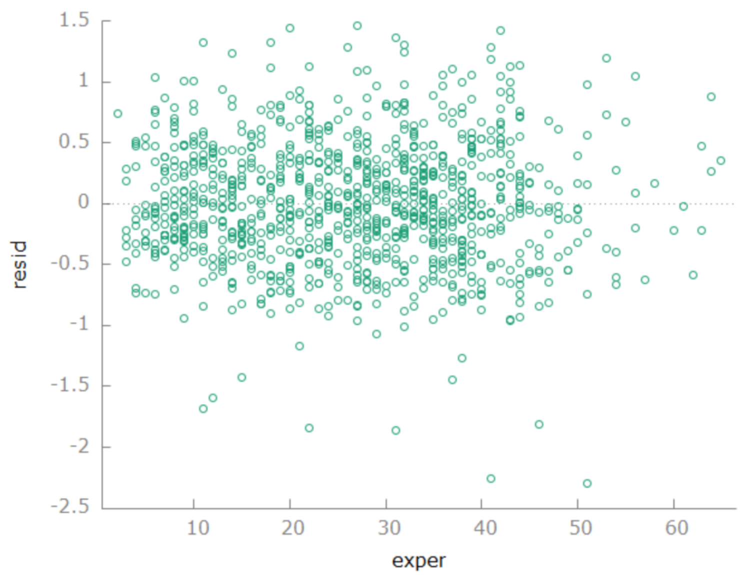


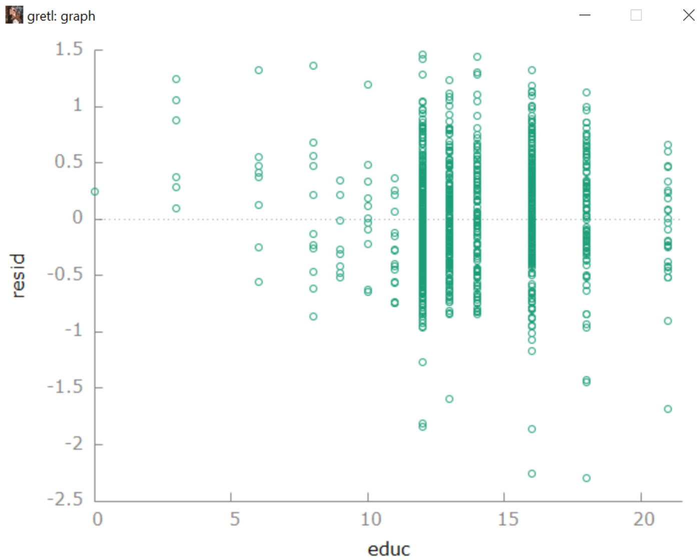
The residual plot suggests the variance of wages for married workers is greater than that for unmarried workers. Thus, there is the evidence of heteroskedasticity.
It probably makes better sense to plot squared residuals against the married variable because in reality, variance is a squared term. However, above figure still shows the change in the dispersion of the data-cloud given the explanatory variable. As we can see, the slope of the fitted line is not horizontal, meaning that there is a heteroskedasticity issue

gnuplot sq_uhat married



(d) Plot the least squares residuals against EDUC and against EXPER. What do they suggest?





Both residual plots exhibit a pattern in which the absolute magnitudes of the residuals tend to increase as the values of *EDUC* and *EXPER* increase, although for *EXPER* the increase is not very pronounced. Thus, the plots suggest there is heteroskedasticity with the variance dependent on *EDUC* and possibly *EXPER*. Again, we should better plot squared residuals against the explanatory variables

(e) Test for heteroskedasticity using a Breusch-Pagan test where the variance depends on EDUC, EXPER and MARRIED. What do you conclude at a 5% significance level?

Since this question asks to use all the variables from the original regression (and not the subset of it (well interaction terms and squares still involve these variables, although they are independent variables derived from the original variables, but it is up to you how you understand the question) , we can just use the software to calculate automatically

modtest --breusch-pagan

```
? modtest --breusch-pagan

Breusch-Pagan test for heteroskedasticity
OLS, using observations 1-1000
Dependent variable: scaled uhat^2

             coefficient    std. error    t-ratio   p-value
  ---------------------------------------------------------
  const       1.44427       0.767360       1.882    0.0601  *
  educ       -0.0482079     0.0498622     -0.9668   0.3339
  exper      -0.0456217     0.0325651     -1.401    0.1615
  exper2      0.000390635   0.000303371    1.288    0.1982
  experedu    0.00262156    0.00167371     1.566    0.1176
  married     0.247908      0.114282       2.169    0.0303  **

  Explained sum of squares = 52.2061

Test statistic: LM = 26.103073,
with p-value = P(Chi-square(5) > 26.103073) = 0.000085
```

The null and alternative hypotheses are

$$\boldsymbol{H_0}\text{: errors are homoskedastic}$$

$$\boldsymbol{H_1}\text{: errors are heteroskedastic}$$

With H_1 implying the error variance depends on one or more of *EXPER*, *EDUC* or *MARRIED*. The value of the test statistic is 26.1, with P value 0.000085, therefore, we reject the null hypothesis and conclude that heteroskedasticity exists.

Feel free to use the manual method by yourself as well as try the white test (manually it will be hard to put all the squares and interactions…)