Middle East Technical University **Fall 2013**
Department of Economics
ECON302

Instructor: Dr. H. Ozan Eruygur
Research Assistant: Fatma Taşdemir

PROBLEM SET 1 – HETEROSCEDASTICITY

Problem 1

Consider the model $y_i = \beta_1 + \beta_2 x_i + e_i$ with heteroscedastic variance $var(e_i) = \sigma_i^2$ and its transformed homoscedastic version $y_i^* = \beta_1 \sigma_i^{-1} + \beta_2 x_i^* + e_i^*$ where $y_i^* = \sigma_i^{-1} y_i$, $x_i^* = \sigma_i^{-1} x_i$, and $e_i^* = \sigma_i^{-1} e_i$. The normal equations whose solution yield the generalized least squares estimators $\widehat{\beta_1}$ and $\widehat{\beta_2}$ are

$$\left(\sum \sigma_i^{-2}\right)\widehat{\beta_1} + \left(\sum \sigma_i^{-1}\, x_i^*\right)\widehat{\beta_2} = \sum \sigma_i^{-1} y_i^*$$
$$\left(\sum \sigma_i^{-1}\, x_i^*\right)\widehat{\beta_1} + \left(\sum x_i^{*^2}\right)\widehat{\beta_2} = \sum x_i^* y_i^*$$

a) Show that $\widehat{\beta_1}$ and $\widehat{\beta_2}$ can be written as

$$\widehat{\beta_2} = \frac{\dfrac{\sum \sigma_i^{-2} y_i x_i}{\sum \sigma_i^{-2}} - \left(\dfrac{\sum \sigma_i^{-2} y_i}{\sum \sigma_i^{-2}}\right)\left(\dfrac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right)}{\dfrac{\sum \sigma_i^{-2} x_i^2}{\sum \sigma_i^{-2}} - \left(\dfrac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right)}$$

$$\widehat{\beta_1} = \frac{\sum \sigma_i^{-2} y_i}{\sum \sigma_i^{-2}} - \left(\frac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right)\widehat{\beta_2}$$

b) Show that $\widehat{\beta_1}$ and $\widehat{\beta_2}$ are equal to the least squares estimators b_1 and b_2 when $\sigma_i^2 = \sigma^2$ for all i. That is the error variances are constant.

c) Does a comparison of the formulas for $\widehat{\beta_1}$ and $\widehat{\beta_2}$ with those for b_1 and b_2 suggest an interpretation for $\widehat{\beta_1}$ and $\widehat{\beta_2}$?

Problem 2

Consider the simple regression model

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

where the e_i are independent errors with $E(e_i\,) = 0$ and $var(e_i) = \sigma_i^2 x_i^2$. Suppose that you have the following five observations

$$y = (4,3,1,0,2) \quad x = (1,2,1,3,4)$$

Use a hand calculator to find the *generalized least squares* estimates of β_1 and β_2.

Problem 3

A sample of 200 Chicago households was taken to investigate how far American households tend to travel when they take vacation. Measuring distance in miles per year, the following model was estimated

$$MILES_i = \beta_1 + \beta_2 INCOME_i + \beta_3 AGE_i + \beta_4 KIDS_i + e_i$$

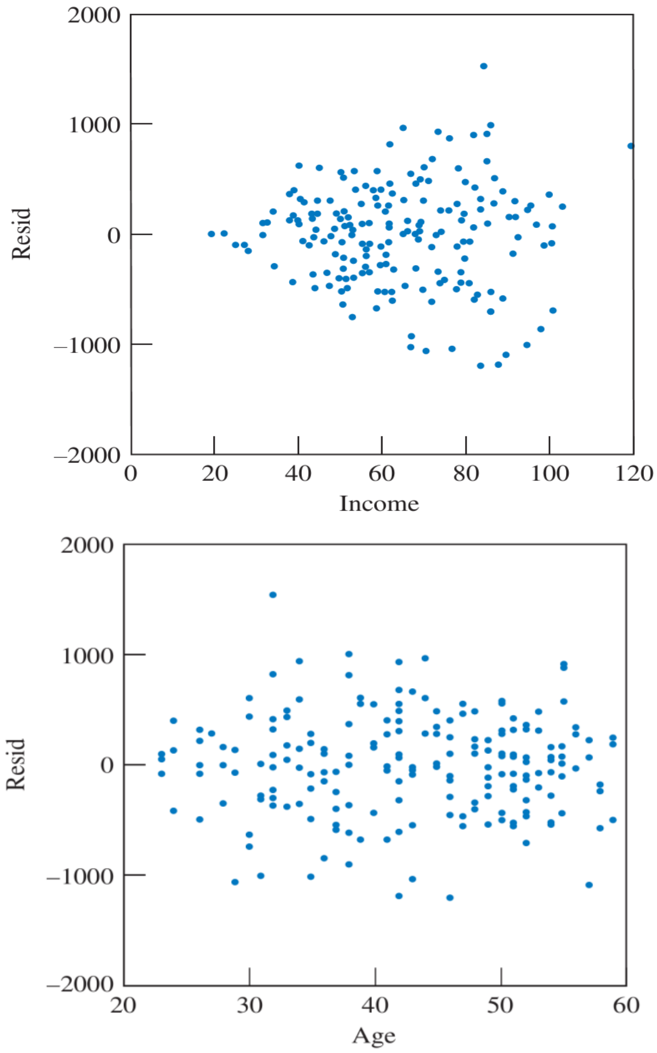

Figure 1: Residual plots for Problem 3: vacation data

The variables are self-explanatory except perhaps for AGE, the average of the adult members of the household. The data are in the vacation.dat.

a) The equation was estimated by least squares and the residuals are plotted against age and income in Figure 1. What do these graphs suggest to you?

b) Ordering the observations according to descending values of $INCOME$, and applying least squares to the first 100 observations, and again to the second 100 observations, yields the sums of squared errors

$$SSE_1 = 2.9471 \times 10^7 \qquad SSE_2 = 1.0479 \times 10^7$$

Use the Goldfeld-Quandt test to test for heteroscedastic errors. Include specification of the null and alternative hypotheses.

c) Table 1 contains three sets of estimates: those from least squares, those from least squares with White's standard errors, and those from generalized least squares under the assumption $\sigma_i^2 = \sigma^2 \times INCOME^2$.

  i. How do vacation miles traveled depend on income, age, and the number of the kids in the household?
  ii. How do White's standard errors compare with the least squares standard errors? Do they change your assessment of the precision of estimation?
  iii. Is there evidence to suggest the generalized least squares estimates are better estimates?

Table 1: Output for Problem 3

Variable	Coefficient	Std Error	t-value	p-value
Least squares estimates				
C	-391.55	169.78	-2.31	0.022
INCOME	14.20	1.80	7.89	0.000
AGE	15.74	3.76	4.19	0.000
KIDS	-81.83	27.13	-3.02	0.003
Least squares estimates with White standard errors				
C	-391.55	142.65	-2.74	0.007
INCOME	14.20	1.94	7.32	0.000
AGE	15.74	3.97	3.97	0.000
KIDS	-81.83	29.15	-2.81	0.006
Generalized least squares estimates				
C	-425.00	121.44	-3.50	0.001
INCOME	13.95	1.48	9.42	0.000
AGE	16.72	3.02	5.53	0.000
KIDS	-76.81	21.85	-3.52	0.001

Problem 4

Consider the model

$$y_i = \beta_1 + \beta_2 x_i + e_i \qquad E(e_i\,) = 0 \quad var(e_i) = \sigma_i^2 = \exp(\alpha z_i)$$

You have the following eight observations on $y_i, x_{i,}$ and z_i:

Y	1.1	-0.5	18.9	-0.9	6.4	1.8	4.5	-0.2
X	-0.5	-3	3.2	-1.8	3.4	-3.5	2.4	-0.2
Z	3.3	0.3	7.0	4.7	1.9	6.8	2.3	6.4

Use a hand calculator to

a) Find least squares estimeates of β_1 and β_2.
b) Find the least squares residuals.
c) Estimate α.
d) Find variance estimates $\hat{\sigma}_i^2$.
e) Find generalized least squares estimates of β_1 and β_2. (*Hint:* Use the results in Problem 1).

Problem 5

State with brief reason whether the following statements are true, false, or uncertain:

a) In the presence of heteroscedasticity OLS estimators are biased as well as inefficient.
b) If heteroscedasticity is present, the conventional t and F tests are invalid.
c) In the presence of heteroscedasticity, the usual OLS method always overestimates the standard errors of estimators.
d) If residuals estimated from an OLS regression exhibit a systematic pattern, it means heteroscedasticity is present in the data.
e) There is no general test of heteroscedasticity that is free of any assumption about which variable the error term is correlated with.
f) If a regression model is misspecified (e.g., an important varible is omitted), the OLS residuals will show a distinct pattern.
g) If a regressor that has nonconstant variance is (incorrectly) omitted from a model, the (OLS) residuals will be heteroscedastic.

Problem 6

In a regression of average wages (W, \$) on the number of employees (N) for a random sample of 30 firms, the following regression results were obtained:

$$\widehat{W} = 7.5 + 0.009N \qquad (1)$$
$$t \; = n.a \quad (16.10) \qquad R^2 = 0.90$$

$$\widehat{W}/N = 0.008 + 7.8(\tfrac{1}{N}) \qquad (2)$$
$$t \quad = (14.43) \quad (76.58) \qquad R^2 = 0.99$$

a) How do you interpret the two regressions?
b) What is the author assuming in going from Eq. (1) to (2)? Was he worried about heteroscedasticity? How do you know?
c) Can you relate the slopes and intercepts of the two models?
d) Can you compare the R^2 values of the two models? Why or why not?

Problem 7

For pedagogic purposes Hanushek and Jackson estimate the following model:

$$C_t = \beta_1 + \beta_2 GNP_t + \beta_3 D_t + u_i \qquad (1)$$

where C_t=aggregate private consumption expenditure in year t, GNP_t=gross national product in year t, and D_t=national defense expenditures in year t, the objective of the analysis being to study the effect of defense expenditures on toher expenditures in the economy.

Postulating that $\sigma_t^2 = \sigma^2 GNP_t^2$, they transform (1) and estimate

$$C_t/GNP_t = \beta_1(1/GNP_t) + \beta_2 + \beta_3(D_t/GNP_t) + u_t/GNP_t \qquad (2)$$

The empirical results based on the data for 1946-1975 were as follows (standard errors in the parentheses):

$$\widehat{C_t} = 26.19 + 0.6248GNP_t - 0.4398D_t$$
$$(2.73) \quad (0.0060) \qquad\quad (0.0736) \qquad\quad R^2 = 0.999$$

$$\widehat{C_t/GNP_t} = 25.92(1/GNP_t) + 0.6246 - 0.4315(D_t/GNP_t)$$
$$(2.22) \qquad\qquad\quad (0.0068) \;\; (0.0736) \qquad\quad R^2 = 0.875$$

a) What assumption is made by the authors about the nature of heteroscedasticity? Can you justify it?
b) Compare the results of the two regressions. Has the transformation of the original model improved the results, that is, reduced the estimation standard errors? Why or why not?
c) Can you compare the two R^2 values? Why or why not? (*Hint:* Examine the dependent variables.)

Problem 8

You are given the following data:

- SSR_1 based on the first 30 observations=55, df=25
- SSR_2 based on the last 30 observations=140, df=25

Carry out the Goldfeld-Quandt test of heteroscedasticity at the 5 percent level of significance.

Problem 9

You are provided with the following information:

Equation (1)

(1a) $\hat{u}^2 = 0.114\text{-}0.051 \ln W + 0.006 (\ln W)^2$ $\quad R^2=0.2572, \quad T=36$

(1b) $\ln q=-8.714+0.700 \ln W$ $\quad SSR=0.00515, \quad t=1947\text{-}1960, \quad T=14$

(1c) $\ln q=-6.693+0.314 \ln W,$ $\quad SSR=0.01250, \quad t=1969\text{-}1982, \quad T=14$

Equation (2)

(2a) $\hat{u}^2 = -0.087+0.042 \ln W+0.565D_1+0.017 D_2-0.235(D_1 \ln W)-0.053(D_2 \ln W)$
$-0.005(\ln W)^2+0.025(D_1 \ln W)^2+0.006(D_2 \ln W)^2$

$R^2=0.5245, \quad T=36$

a) Carry out all the tests of heteroscedasticity you can for equations (1) and (2).
b) Do you conclude that taking into account of structural change has solved the problem of heteroscedasticity? Why or why not?

Problem 10

The following estimation results are based on cross-sectional data pertaining to consumpion expenditures (C) and disposable income (Y), of 30 families. The data has been put in ascending order with respect to the values of Y.

(1) $\hat{C}_t = 1480.0000+0.7885\, Y_t,$ $\quad t=1.........30$
$\quad (449.5059) \quad (0.0268)$ $\quad SSR=4994182$

(2) $\hat{C}_t = 845.6667+0.8367\, Y_t$ $\quad t=1..........12$
$\quad (1143.5661) \quad (0.0844)$ $\quad SSR=106900$

(3) $\hat{C}_t = 2306.6667 + 0.7467\,Y_t$ t=19.........30

$(2916.3173)\quad(0.1493)$ SSR=3344000

(4) $\hat{U}_t^2 = 527507.10 - 76.4307\,Y_t + 0.0032\,Y_t^2$ t=1.........30

$(1302030.30)\quad(161.1689)\quad(0.0049)$ $R^2=0.185593$

Carry out tests of heteroscedasticity at the $\alpha=0.05$ level of significance. Do all test results agree with each other? What conclusion would you reach as a result of these tests?

Problem 11

In a survey the following data are obtained:

MEAN AGE(A_t)	*MEDIAN SALARY(S_t)*
20	850
25	1150
30	2250
35	2870
40	3700
45	4500
50	4500
55	4500
60	3250
65	2500
70	2000
75	1750

a) Develop a suitable regression model explaining median salary in relation to mean age.
b) Assuming that the variance of disturbance term is proportional to the square of the mean age transform the data so as to make the resulting disturbance term homoscedastic.
c) Test for heteroscedasticity using GQ test.
d) Find GLS and EGLS estimates of B_0 and B_1 respectively, assuming :
   i. $\sigma^2_t=\sigma^2_t\,A^2_t$
   ii. $\sigma^2_t=\sigma^2\,(E(S_t))^2$

Problem 12

The following estimation results are based on 50 observations. The data has been put in ascending order with respect to the values of X_1.

$\hat{Y}_t = 3+2.1X_{t1}+1.9X_{t2}$ t=1,2,.......,50 SSR=55.2

$\hat{Y}_t = 2.5+2.19X_{t1}+1.1X_{t2}$ t=1,2,.......,20 SSR=9.3

$\hat{Y}_t = 1.9+2.2X_{t1}+0.8X_{t2}$ t=31,32,...,50 SSR=75.7

$\hat{u}_t^2 = 5.1+1.7X_{t1}+2.3X_{t2}+1.1X_{t1}^2+0.8X_{t2}^2+0.6X_{t1}\,X_{t2}+e_t$ $R^2=0.28$

Test if the model is heteroscedastic using two different tests and name the tests you have used.

Problem 13

Assume a regression model is specified as Y_t=bX_t+u $_t$ and we have 10 observations on X and Y.

X_t	Y_t
6	25
7	23
9	20
11	17
13	15
14	14
12	19
16	18
20	22
21	25

Estimate b using the most efficient technique assuming:

a) $E(u_t)=0$ $\qquad$ $E(u_t^2)=2$

b) $E(u_t)=0$ $\qquad$ $E(u_t^2)=\sigma^2 X_t^2$

c) $E(u_t)=0$ $\qquad$ $E(u_t^2)=\sigma^2 X_t$

d) $E(u_t)=0$ $\qquad$ $E(u_t^2)=\sigma^2 (E(Y_t))^2$

Problem 14

Given the model:

$$Y_t=B_0+B_1 X_t+u_t \qquad t=1......18$$

where;

$$E(u_t^2)=\begin{cases} \sigma_1^2=1/4 & t=1....6 \\ \sigma_2^2=1/9 & t=7....12 \\ \sigma_3^2=1/16 & t=13...18 \end{cases}$$

a) Transform the model into a homoscedastic one writing the weights clearly for each observation.

b) Derive the GLS estimates of B_1.

c) Assume that the data have been ranked in ascending order according to X and two separate regression are estimated for the first seven and the last seven observations and the following sum of the squared residuals are obtained:

$$SSR_1=\sum_{t=1}^{7}\hat{u}_{t1}^2 = 45.8 \qquad SSR_2=\sum_{t=12}^{18}\hat{u}_{t2}^2 = 675.9 \ .$$

Test for heteroscedasticity and name the test you used.

Problem 15

Consider the following estimated equations:

Eq.1 $\hat{FE}_t = 8.20+0.17\ YD_t$

(3.00) (0.03)

[2.80] [0.02] $\sum_{t=1}^{10} \hat{u}_t^2 = 90$, R^2: 0.61, $\hat{\sigma}^2 = 18$, T= 1,2,…,24

Eq.2 $\hat{FE}_t = 12.7+0.08\ YD_t$

(1.10) (0.01)

R^2: 0.87, SSR=8, T= 1,…,10

Eq.3 $\hat{FE}_t = 12.0+0.18\ YD_t$

(3.80) (0.01)

R^2: 0.78, SSR=72, T= 15,…,24

Eq.4 $\hat{u}_t^2 = -41.8+0.80\ YD_t - 0.002\ YD_t^2$

(54.5) (1.00) (0.004)

R^2: 0.17, SSR=11000, T=1,…,24

Eq.5 $\hat{u}_t^2 = -21.9+0.40\ YD_t - 0.001\ YD_t^2+2.70\ N_t^2$;

(39.0) (0.80) (0.003) (0.60)

R^2: 0.60, SSR=5400, T=1,…,24

Eq.6 $\hat{u}_t^2 = 3.30+2.90\ N_t^2$;

(1.10) (0.50)

R^2: 0.58, SSR=5500, T=1,…,24

where, FE=food expenditure, YD=disposable income, N=number of persons in each household, $\hat{u} =$ OLS residuals from eq.1.
The values in parenthesis (.) are standard errors. HCSE's are presented in brackets [.].

A)

i. Consider equation 1. Test for heteroscedasticity using all the test statistics computable by the information below.

ii. Test the hypothesis that the disposable income coefficient is unity in equation 1.

iii. Briefly interpret and compare equations 4, 5, and 6.

B) The following 2-step EGLS equation is estimated under the maintained hypothesis that $\sigma_t^2 = \gamma_0 + \gamma_1 N_t^2$, where γ_0 and γ_1 are unknown parameters.

Eq. 7. $\hat{FE}_t^* = 8.00\ \hat{w}_t + 0.150\ YD_t^*$;

(2.00) (0.02)

R^2: 0.860, SSR=23, T= 1,…,24, Q_{white}=0.40

i. Is there any theoretical reason and/or empirical result to support the maintained hypothesis on the form of heteroscedasticity

ii. Which of the following expressions is correct for equation 7?

a) FE_t *=$FE_t\ w_t$ and $w_t=1/(\lambda_t)^{1/2}$

b) FE_t *=$FE_t\ \hat{w}_t$ and $\hat{w}_t = 1/(\hat{\lambda}_t)^{1/2}$

c) FE_t *=FE_t / $\hat{FE}_t$ and $\hat{w}_t = 3.30+2.90\ N_t^2$

d) FE_t *=FE_t / $(\alpha_0+\alpha_1\ N_t^2)^{1/2}$ and $w_t=1/(\lambda_t)^{1/2}$

e) None of the above. Since;

FE_t *=_____________, and $\hat{w}_t$ = ____________. (*Please complete*).

C) The following equation is also estimated by using the same sample of observations:

Eq. 8: $\hat{FE}_t = 9.09+0.12\ YD_t+2.50\ N_t$;

(3.00) (0.03) (0.50)

[2.80] [0.02] [0.40]

R^2: 0.860, SSR=145, T= 1,…,24, Q_{white}=0.60

i. Describe the regression to be run to conduct the White test (Q_{white}) for equation 8.

ii. Compare equations 1, 7 and 8.

Problem 16

Given,

$Y_t=b_0+u_t$ and

i. $E(u_t)=0$ for all t.

ii. $E(u_t\ u_s)=0,\ t \neq s$

iii. $E(u^2_t)=\sigma^2 / X^2_t$

Obtain the GLS estimator of b_0 and its variance.

Problem 17

You wish to estimate the model, $Y_t=b_0+b_{t1}\ X_{t1}+b_{t2}\ X_{t2}+u_t$ based on 110 observations. The data arranged in ascending order with respect to the values of X_{t1} and the following regressions are estimated using OLS.

(1)	$\hat{Y}_{t1} = 4.0+3.6\ X_{t1}+1.2\ X_{t2}$	t=1,2,…,110	SSR=95,	R^2=0.90
(2)	$\hat{Y}_{t2} = 3.1+2.8\ X_{t1}+0.9\ X_{t2}$	t=1,2,…,44	SSR =5,	R^2=0.94
(3)	$\hat{Y}_{t3} = 4.2+3.9\ X_{t1}+1.5\ X_{t2}$	t=67,…,110	SSR =70,	R^2=0.85
(4)	$\hat{Y}_{t4} = 4.8+3.0\ X_{t1}+1.3\ X_{t2}$	t=77,…,110	SSR =60,	R^2=0.88
(5)	$\hat{u}_t^2 = 2.1+2.3\ X_{t1}+1.9\ X_{t2}+1.6\ X^2_{t1}+1.2\ X^2_{t2}+0.6\ X_{t1}\ X_{t2}$	t=1,2,…,110	SSR =1.5,	R^2=0.75

Test if the model is heteroscedastic and name the tests you have used (clearly state your null and alternative hypothesis, degrees of freedom, test statistic you use).

Problem 18

What is meant by heteroscedasticity? What are effects on

a) OLS estimators and their variances?
b) Confidence intervals?
c) The use of “t” and “F” test of significance?

Problem 19

Consider the following model

$$Y_t=\alpha+\beta X_t+u_t \qquad \text{where } E(u_t^2)=\sigma^2 X_t^4$$

a) How would you transform the model in order to achieve homoscedastic error variance?
b) How would you estimate the transformed model? List the necessary steps.

Problem 20

Suppose you are given the equation

$$Y_t=b_0+u_t$$

and

i. $E(u_t)=0$ for all t.
ii. $E(u_t u_s)=0$, $t \neq s$
iii. $E(u^2{}_t)=\sigma^2 X^4{}_t$

Show that the OLS estimator of b_0 is not efficient.