

## Computer Exercise 5 Answers – Heteroskedasticity

. u cex5

Using the summary command you can inspect the means, maxima and minima of the variables

. su

Variable	Obs	Mean	Std. Dev.	Min	Max
years	807	12.47088	3.057161	6	18
cigprice	807	60.30041	4.738469	44.004	70.129
white	807	.8785626	.3268375	0	1
age	807	41.23792	17.02729	17	88
income	807	19304.83	9142.958	500	30000
numcigs	807	8.686493	13.72152	0	80
rest	807	.2465923	.4312946	0	1
lninc	807	9.687315	.7126952	6.214608	10.30895
age2	807	1990.135	1577.166	289	7744
lncigp	807	4.096032	.0829194	3.784281	4.250336

So that the mean number of cigarettes smoked each day is 8.6 with a minimum of zero and maximum of 80.

. reg numcigs age age2 yearsed lncigp lninc

Source	SS	df	MS	Number of obs =	807
Model	6842.82045	5	1368.56409	F( 5, 801) =	7.56
Residual	144910.862	801	180.912437	Prob > F =	0.0000
Total	151753.683	806	188.280003	R-squared =	0.0451
				Adj R-squared =	0.0391
				Root MSE =	13.45

numcigs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.7806352	.1606188	4.860	0.000	.4653517 1.095919
age2	-.0091056	.0017487	-5.207	0.000	-.0125381 -.0056731
years	-.5141422	.167571	-3.068	0.002	-.8430722 -.1852121
lncigp	-2.853151	5.733184	-0.498	0.619	-14.10699 8.400687
lninc	.7582931	.7286683	1.041	0.298	-.6720319 2.188618
_cons	5.368736	23.89722	0.225	0.822	-41.53983 52.2773

Since age is entered as a quadratic, an extra year of age influences the number of cigarettes smoked each day by around

$$d(\text{numcigs})/d\text{Age} = 0.78 - 2 * 0.009\text{Age}$$

so that cigarette consumption reaches a (theoretical) maximum at  $0.78 / .018 = 43$  (and then begins to fall back - the oldest person in the data set is 88)

Years of education is linear, so one extra year of education reduces cigarette consumption by around 0.5 a day, on average.

Price and income do not appear to be statistically significant in this formulation.

Breusch Pagan test for Heteroskedasticity

Save residuals from the OLS regression. Regress the square of these predicted residuals on the original set of regressors

```
. predict reshat, resid
. g reshat2=reshat^2
. reg reshat2 age age2 yearsed lncigp lninc
```

Source	SS	df	MS	Number of obs =	807
Model	3594658.95	5	718931.791	F( 5, 801) =	5.25
Residual	109586286	801	136811.843	Prob > F =	0.0001
				R-squared =	0.0318
				Adj R-squared =	0.0257
				Root MSE =	369.88
Total	113180945	806	140423.009		

reshat2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	19.2736	4.416967	4.364	0.000	10.6034	27.9438
age2	-.2124767	.0480874	-4.419	0.000	-.3068689	-.1180844
yearsded	-2.717985	4.608148	-0.590	0.555	-11.76346	6.327487
lncigp	5.919533	157.6607	0.038	0.970	-303.5574	315.3965
lninc	23.43158	20.03815	1.169	0.243	-15.9019	62.76506
_cons	-409.7185	657.1658	-0.623	0.533	-1699.689	880.252

Under the null hypothesis of homoskedasticity, the statistic

$$N \cdot R^2 = 807 \cdot 0.0318 = 25.7 \sim \chi^2(k)$$

Where k is the number of regressors in the auxiliary regression (**excluding** the constant, since if only the constant is significant this means the residual are also constant)

From tables the  $\chi^2(5)$  critical value at the 95% level is 11.07

Therefore  $\chi^2 > \chi^2(5)_{\text{critical}}$

So reject null of homoskedasticity (estimated residual variance does appear to vary with levels of the right hand side variables - notably age - in the auxiliary regression above).

(Note that this is not quite the same test as appears in the lecture notes, but Johnston & DiNardo chapter 6 shows that the above is asymptotically equivalent).

The **White Test** for heteroskedasticity involves the levels, squares and cross products of all the right hand side variables.

```
. g age4=age2^2
. g ed2=yearsded^2
. g lnc2=lncigp^2
. g lninc2=lninc^2
. g ageage2=age*age2
. g ageed=age*yearsded
. g agecp=age*lncigp
. g agein=age*lninc
```

```
. g age2ed=age2*yearsed
. g age2cp=age2*lnincigp
. g age2in=age2*lninc
. g edcp=yearsed*lnincigp
. g edin=yearsed*lninc
. g cpin=lnincigp*lninc
```

These terms (and the rhs variables in the original regression are then regressed on the squared residuals - the same squared residuals as in the Breush Pagan test).

```
. reg reshat2 age age2 yearsed lninc lnincigp age4 ed2 lnc2 lninc2 ageage2 aged
> agecp agein age2ed age2cp edcp edin cpin
```

Source	SS	df	MS	Number of obs =	807
Model	5866260.04	18	325903.335	F( 18, 788) =	2.39
Residual	107314685	788	136186.148	Prob > F =	0.0010
				R-squared =	0.0518
				Adj R-squared =	0.0302
Total	113180945	806	140423.009	Root MSE =	369.03

reshat2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-249.8941	235.6643	-1.060	0.289	-712.4981	212.7099
age2	3.554355	3.196707	1.112	0.267	-2.720713	9.829423
yearsed	-134.1134	252.0662	-0.532	0.595	-628.9141	360.6873
lninc	-850.0402	961.5075	-0.884	0.377	-2737.459	1037.379
lnincigp	-15042.66	9487.555	-1.586	0.113	-33666.53	3581.211
age4	.0001389	.0001443	0.963	0.336	-.0001443	.0004222
ed2	-.3943341	1.297562	-0.304	0.761	-2.941422	2.152753
lnc2	1330.966	1172.89	1.135	0.257	-971.393	3633.325
lninc2	-7.22581	16.7408	-0.432	0.666	-40.08765	25.63603
ageage2	-.0237558	.028469	-0.834	0.404	-.0796399	.0321283
aged	3.364368	1.683282	1.999	0.046	.0601207	6.668615
agecp	51.18759	55.37556	0.924	0.356	-57.51346	159.8886
agein	-.8802429	1.174691	-0.749	0.454	-3.186137	1.425651
age2ed	-.0329717	.0170998	-1.928	0.054	-.0665383	.0005949
age2cp	-.4976826	.5948478	-0.837	0.403	-1.665356	.6699911
edcp	38.38581	59.44379	0.646	0.519	-78.30111	155.0727
edin	-9.470243	8.091913	-1.170	0.242	-25.3545	6.414012
cpin	280.252	237.3359	1.181	0.238	-185.6334	746.1373
_cons	37935.54	19995.43	1.897	0.058	-1315.062	77186.14

The test is again an LM test that all the variables in this regression are zero, except the constant of the form

$$N \cdot R^2 = 807 \cdot 0.0518 = 41.8 \sim \chi^2(18)$$

Again  $\chi^2 > \chi^2(18)$ critical

So again reject null of homoskedasticity (estimated residual variance does appear to vary with levels of the right hand side variables - notably age - in the auxiliary regression above.

The **Goldfeld-Quandt** Test assumes knowledge of the variable potentially causing heteroskedasticity, in this case age.

The idea is to first split the data into 2 subsets of approximately the same size (most text books give examples of the same size but this is not necessary)

```
. reg numcigs age age2 yearsed lncigp lninc if age<=40
```

Source	SS	df	MS	Number of obs =	441
Model	5103.46252	5	1020.6925	F( 5, 435) =	6.16
Residual	72074.025	435	165.687414	Prob > F =	0.0000
				R-squared =	0.0661
				Adj R-squared =	0.0554
Total	77177.4875	440	175.403381	Root MSE =	12.872

numcigs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	1.896592	.8553263	2.217	0.027	.2155058	3.577678
age2	-.0268437	.0150699	-1.781	0.076	-.0564625	.0027751
yearsed	-1.106068	.2485324	-4.450	0.000	-1.594542	-.6175949
lncigp	-5.501084	7.689271	-0.715	0.475	-20.61383	9.611659
lninc	.8683389	.872849	0.995	0.320	-.8471868	2.583865
_cons	6.375362	33.90145	0.188	0.851	-60.25564	73.00636

```
. reg numcigs age age2 yearsed lncigp lninc if age>40
```

Source	SS	df	MS	Number of obs =	366
Model	3485.90428	5	697.180855	F( 5, 360) =	3.53
Residual	71073.3225	360	197.425896	Prob > F =	0.0039
				R-squared =	0.0468
				Adj R-squared =	0.0335
Total	74559.2268	365	204.271854	Root MSE =	14.051

numcigs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.3714549	.6914006	0.537	0.591	-.9882364	1.731146
age2	-.0053264	.0057564	-0.925	0.355	-.0166468	.005994
yearsed	-.1472604	.2426356	-0.607	0.544	-.6244216	.3299007
lncigp	-1.188689	8.580715	-0.139	0.890	-18.06331	15.68593
lninc	.672531	1.321469	0.509	0.611	-1.926237	3.271299
_cons	5.435032	42.50687	0.128	0.898	-78.15794	89.028

The idea is to split the sample into high variance (older) and low variance (younger) groups and test whether the estimated residual variances from the 2 subsets are statistically different.

From above output, the Root MSE is Stata's estimate of the model residual standard error, so just square it to obtain the residual variance

$$SO\ GQ = \frac{\sigma_{high\ variance}^2}{\sigma_{low\ variance}^2} \sim F(N_{high}-k_{high}, N_{low}-k_{low})$$

(Note that if the sample sizes in the 2 sub-groups are equal, this reduces to the ratio of the residual sum of squares in the 2 groups)

$$so\ GQ = (14.051)^2 / (12.872)^2 \sim F(366-6, 441-6)$$

$$= 1.19 \quad \sim F(361, 435)$$

Given 5% critical value  $F(\alpha, \alpha) = 1.99$ . Then cant reject null that age is not responsible for heteroskedasticity.

On the assumption that the residual variance can be modelled as

$$\text{Var}[u_i / X_i] = \sigma^2 \text{age}_i^2$$

then **Feasible GLS (Weighted Least Squares)** can be done by transforming all the data (dependent and RHS variables) in the original equation by the square root of the variable thought to be causing heteroskedasticity. In this case the residuals in this transformed equation satisfy the Gauss-Markov criteria and OLS estimation of this transformed equation can give efficient estimates (unlike the original OLS estimation)

```
. g numage=numcigs/age
. g oneage=1/age
. g age2oage=age2/age
. g edoage=yearsred/age
. g cpoage=lncigp/age
. g inoage=lninc/age

. reg numage oneage age2oage edoage cpoage inoage
```

Source	SS	df	MS			
Model	9.93000592	5	1.98600118	Number of obs =	807	
Residual	122.260849	801	.152635268	F( 5, 801) =	13.01	
				Prob > F =	0.0000	
				R-squared =	0.0751	
				Adj R-squared =	0.0693	
				Root MSE =	.39069	
Total	132.190855	806	.164008505			

  

	numage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	oneage	25.97943	22.52427	1.153	0.249	-18.23413	70.19299
	age2oage	-.0117926	.0022057	-5.346	0.000	-.0161223	-.007463
	edoage	-.7371059	.1784982	-4.129	0.000	-1.087485	-.3867264
	cpoage	-8.644436	5.434412	-1.591	0.112	-19.31181	2.022933
	inoage	.9344365	.5648385	1.654	0.098	-.174302	2.043175
	_cons	1.011789	.1766175	5.729	0.000	.6651009	1.358477

Note that the constant in this regression becomes the efficient estimate of  $b_1$  (the coefficient on age in the original equation) since  $\text{age}/\text{age} = 1 = \text{constant}$ . Similarly the coefficient of  $1/\text{age}$  in the FGLS equation is the efficient estimate of the constant in the original OLS equation

Note also that these estimates are only efficient **if** the assumption that heteroskedasticity can be modelled as  $\text{Var}[u_i / X_i] = \sigma^2 \text{age}_i^2$  is correct

Comparing the results with the original OLS equation, the coefficient on age appears to have risen (from 0.78 to 1.01) and the coefficient on years of education is now more negative and more significant.

In practice it is very difficult to obtain the precise form of heteroskedasticity (except in a few specialised cases, such as when working with grouped data – see lecture notes). Most researchers therefore usually adjust the OLS estimates – more specifically the OLS standard errors - to take account of the presence of heteroskedasticity using the **White adjustment** formula (see lecture

notes). These adjusted OLS standard errors, whilst not fully efficient are at least consistent, unlike the unadjusted OLS standard errors and may be preferred when making statistical inferences on the basis of t and F tests.

This can be done easily in stata using the command

```
. reg numcigs age age2 yearsed lncigp lninc, robust
```

Regression with robust standard errors

```
Number of obs =      807
F( 5, 801) =      11.36
Prob > F      =      0.0000
R-squared     =      0.0451
Root MSE     =      13.45
```

numcigs	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
age	.7806352	.1401858	5.569	0.000	.5054602	1.05581
age2	-.0091056	.0014817	-6.146	0.000	-.012014	-.0061972
years	-.5141422	.1627735	-3.159	0.002	-.8336552	-.1946291
lncigp	-2.853151	5.989163	-0.476	0.634	-14.60946	8.903157
lninc	.7582931	.5979058	1.268	0.205	-.4153542	1.93194
_cons	5.368736	25.37082	0.212	0.832	-44.43241	55.16988

Comparing the adjusted standard errors with the unadjusted ones in the original OLS regression, it appears that the standard errors on age are smaller and the t values correspondingly higher in the adjusted estimates.