

Homework 8_9

Heteroskedasticity and Autocorrelation.

Question 1: Heteroskedasticity

Using the (houthak.dta) dataset, which was used for an analysis in a famous paper by Houthakker (1951), we can analyze the demand for electricity in a cross section of towns in Great Britain. After renaming the data, the original data and description are:

Name	new label	description
income:	inc	Income of electricity consumers (pounds per year)
p36:	p1	Price of electricity in 1935-36 (pence per kwh)
p38:	p2	Price of electricity in 1937-38 (pence per kwh)
gas36:	pg	Price of gas in 1935-36 (pence per kwh)
equip:	he	Average holding of heavy equipment by consumers (kwh)
consump:	con	domestic consumption per customer
expend:	exp	average total expenditure on electricity by consumer
num:	num	Number of customers in each town

1. Estimate a model of consumption with the following variables: Income, inverse of the electricity price (P36), price of gas, and heavy equipment.

```
. gen invp1=1/p1
. reg con inc invp1 pg he
```

Source	SS	df	MS	Number of obs =	42
-----+-----				F(4, 37) =	31.96
Model	13126642.7	4	3281660.68	Prob > F	= 0.0000
Residual	3799648.37	37	102693.199	R-squared	= 0.7755
-----+-----				Adj R-squared =	0.7512
Total	16926291.1	41	412836.368	Root MSE	= 320.46

con	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					
inc	1.917337	.1819347	10.54	0.000	1.548702 2.285972
invp1	752.7082	164.8723	4.57	0.000	418.6453 1086.771
pg	1.750985	34.29271	0.05	0.960	-67.73265 71.23462
he	286.524	98.69427	2.90	0.006	86.55046 486.4976

_cons | -1507.593 498.015 -3.03 0.004 -2516.667 -498.5189

2. Then check for multicollinearity and specification errors:

. estat vif

Variable	VIF	1/VIF
-----+-----		
pg	1.16	0.862853
invp1	1.15	0.873294
inc	1.13	0.882435
he	1.11	0.900714
-----+-----		
Mean VIF	1.14	

Suggests multicollinearity will not inflate the standard errors.

. estat ovtest

Ramsey RESET test using powers of the fitted values of con

Ho: model has no omitted variables

F(3, 34) = 1.77

Prob > F = 0.1723

This suggests model is ok.

Unfortunately,

. estat ovtest, rhs

Ramsey RESET test using powers of the independent variables

Ho: model has no omitted variables

F(12, 25) = 2.79

Prob > F = 0.0147

This indicates that entering our variables in linear fashion is not sufficient to explain the data generating process. We should probably investigate a better theoretical functional form before proceeding.

3. Now, check for normality of the error structure:

```
. predict uhat, resid  
. sfrancia uhat
```

Shapiro-Francia W' test for normal data

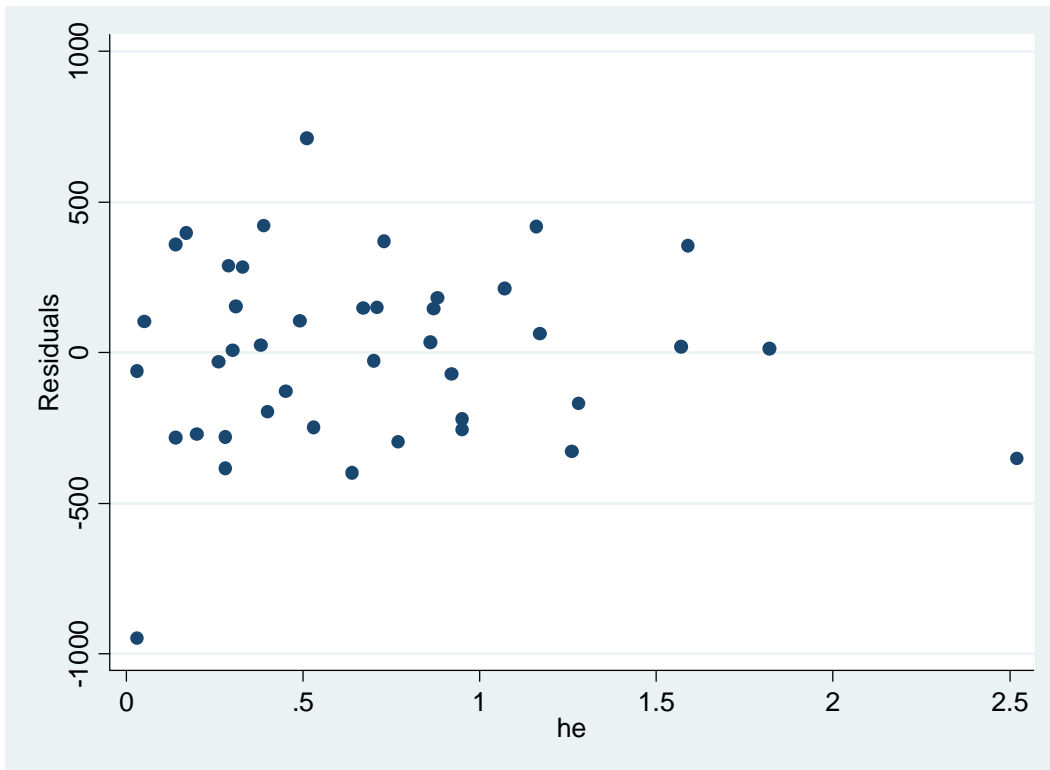
Variable	Obs	W'	V'	z	Prob>z
uhat	42	0.95933	1.842	1.152	0.12472

Errors appear normal.

4. Evaluate for Heteroscedasticity.

4A. print the errors in relation to 'he'

```
. rvpplot he
```



The above Graph shows a possible heteroskedastic pattern in the errors with respect to the 'he' variable.

4B Conduct the G-Q test for the (pg) variable and discuss your findings. (note: I set c=4)

```
. sort pg
. quietly reg con inc invp1 pg he in 1/19
. scalar pp1=e(rss)
. quietly reg con inc invp1 pg he in 24/42
. scalar pp2=e(rss)
. gen pp3=pp1 if pp1>=pp2
. replace pp3=pp2 if pp2>=pp1
. gen pp4=pp1 if pp1<=pp2
. replace pp4=pp2 if pp2<=pp1
. gen pp=pp3/pp4
. di Ftail(14,14,pp)
.069263
```

Result: At the 5% level of significance, the pg variable does not appear to be a source of heteroscedasticity.

4C. Conduct the (second) Glejser test for the (1/p1) variable and discuss your findings.

```
. quietly reg con inc invp1 pg he
. predict uhat, resid
. gen abs_uhat=abs(uhat)
. regress abs_uhat p1
```

Source	SS	df	MS	Number of obs =	42
-----+-----				F(1, 40) =	0.01
Model	371.756795	1	371.756795	Prob > F =	0.9201
Residual	1457731.69	40	36443.2923	R-squared =	0.0003
-----+-----				Adj R-squared =	-0.0247
Total	1458103.45	41	35563.4987	Root MSE =	190.9

abs_uhat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]

```

-----+-----
      p1 | -31.58213  312.6949  -0.10  0.920  -663.5621  600.3978
      _cons | 253.2161  171.8454   1.47  0.148  -94.09643  600.5287
-----+-----

```

As shown, the p1 variable is not a source of heteroscedasticity.

4D. Conduct the White test but drop all the interaction terms. Discuss your findings.

```

. gen uhatsq=uhat*uhat
. gen inc2=inc*inc
. gen invp12=invp1*invp1
. gen pg2=pg*pg
. gen he2=he*he
. quietly regress uhatsq inc inc2 invp1 invp12 pg pg2 he he2
. scalar pp3= e(N)*e(r2)
. di chi2tail(8, pp3)
.44273082
. quietly regress con inc invp1 pg he
. hettest inc2 invp12 pg2 he2, rhs iid

```

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: inc2 invp12 pg2 he2 inc invp1 pg he

chi2(8) = 7.91

Prob > chi2 = 0.4427

4E. Look for a weighting scheme. Notice that the data are collected by averaging across very different number of consumers.

```
. summarize inc num
```

Variable	Obs	Mean	Std. Dev.	Min	Max
inc	42	592.7143	292.8343	279	1422
num	42	16.5881	18.63596	1.3	88.7

```
. summarize num, detail
      num
```

Percentiles		Smallest		
1%	1.3	1.3		
10%	2.2	2.2	Obs	42
25%	5.5	2.2	Sum of Wgt.	42
50%	8.9		Mean	16.5881
		Largest	Std. Dev.	18.63596
75%	20.4	41.7		
99%	88.7	88.7		

This suggest that NUM could be used as a weight! In other words, we might think that:

$$\sigma_i^2 = \frac{\sigma^2}{n_i}$$

Stata knows what you are planning to do with $\sigma_i^2 = \sigma^2 / n_i$ in terms of weighting. All you need to do is append [aweight=num] to the end of the regression.

```
. regress con inc invp1 pg he [aweight=num]
```

```
(sum of wgt is 6.9670e+02)
```

Source	SS	df	MS	Number of obs =	42
-----+-----				F(4, 37) =	48.69
Model	7899684.02	4	1974921.01	Prob > F =	0.0000
Residual	1500864.06	37	40563.8934	R-squared =	0.8403

```
-----+-----
Total | 9400548.08  41 229281.661      Adj R-squared = 0.8231
      |                               Root MSE   = 201.4
```

```
-----+-----
con |   Coef.  Std. Err.   t  P>|t|  [95% Conf. Interval]
-----+-----
inc |  2.341007  .2006208  11.67  0.000   1.93451  2.747503
invp1 | 604.0278  124.9037   4.84  0.000  350.9488  857.1068
pg |  40.88995  21.16741   1.93  0.061  -1.999305  83.7792
he |  267.7428  61.91254   4.32  0.000  142.296  393.1895
_cons | -1666.649  310.4055  -5.37  0.000  -2295.59 -1037.708
-----+-----
```

Doing this manually:

Generate transformed variables by multiplying each term (and the constant) by \sqrt{num} (generate commands not shown).

```
. reg wcon snum winc winvp1 wpg whe, noconstant
```

```
Source |   SS   df   MS       Number of obs =   42
-----+-----
Model | 972432411   5 194486482       F( 5, 37) = 289.04
Residual | 24896475.9 37 672877.727       Prob > F   = 0.0000
-----+-----
R-squared   = 0.9750
Adj R-squared = 0.9717
Total | 997328887  42 23745925.9       Root MSE   = 820.29
```

```
-----+-----
wcon |   Coef.  Std. Err.   t  P>|t|  [95% Conf. Interval]
-----+-----
snum | -1666.649  310.4055  -5.37  0.000  -2295.59 -1037.708
winc |  2.341007  .2006208  11.67  0.000   1.93451  2.747503
winvp1 | 604.0278  124.9037   4.84  0.000  350.9488  857.1068
wpg |  40.88995  21.16741   1.93  0.061  -1.999302  83.7792
```

whe | 267.7428 61.91254 4.32 0.000 142.296 393.1895

5. Dealing with Outliers:

An automated way to deal with outliers is to run the `rreg` in Stata. This procedure provides an innovative way to weight the errors such that impact of outliers are mitigated to the point that they are possibly completely ignored. You should investigate the literature developing this procedure before running it. For our data:

```
. rreg con inc invp1 pg he, graph genwt(w)
Huber iteration 1: maximum difference in weights = .52750159
Huber iteration 2: maximum difference in weights = .07683591
Huber iteration 3: maximum difference in weights = .03866779
Biweight iteration 4: maximum difference in weights = .15488038
Biweight iteration 5: maximum difference in weights = .03837982
Biweight iteration 6: maximum difference in weights = .0222037
Biweight iteration 7: maximum difference in weights = .01307324
Biweight iteration 8: maximum difference in weights = .00792082
Robust regression                Number of obs = 42
                                F( 4, 37) = 39.01
                                Prob > F   = 0.0000
```

con	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
inc	2.057845	.172599	11.92	0.000	1.708126	2.407564
invp1	646.8404	156.412	4.14	0.000	329.9195	963.7613
pg	-9.970968	32.53302	-0.31	0.761	-75.88913	55.9472
he	258.8521	93.62989	2.76	0.009	69.13995	448.5643
_cons	-1262.863	472.46	-2.67	0.011	-2220.158	-305.5686

Notice that this produces a very different set of results compared to the `reg` command. Make sure this is the correct procedure.

Question 2. Autocorrelation

In a trade context, dumping is an action in which a nation or firm exports below the cost of production. Theoretically, dumping can force hardship on domestic industries that, if severe enough, could force them out of business. Antidumping laws are used to protect nations from firms looking to unload excess quantities on the world market. Data from a study by Krupp and Pollard (1996) were used to analyze antidumping legislation designed to protect barium chloride producers in the U.S. against actions in China. These data are on the web (bariumch.dta):

The descriptors are:

1. chnimp	Chinese imports, bar. chl.	17. lrtwex	log(rtwex)
2. bchlimp	total imports bar. chl.	18. lchempi	log(chempi)
3. befile6	=1 for all 6 mos before filing	19. t	time trend
4. affile6	=1 for all 6 mos after filing	20. feb	=1 if month is feb
5. afdec6	=1 for all 6 mos after decision	21. mar	=1 if month is march
6. befile12	=1 all 12 mos before filing	22. apr	etc.
7. affile12	=1 all 12 mos after filing	23. may	.
8. afdec12	=1 all 12 mos after decision	24. jun	
9. chempi	chemical production index	25. jul	
10. gas	gasoline production	26. aug	
11. rtwex	exchange rate index	27. sep	
12. spr	=1 for spring months	28. oct	
13. sum	=1 for summer months	29. nov	
14. fall	=1 for fall months	30. dec	
15. lchnimp	log(chnimp)	31. percchn	% imp. from china
16. lgas	log(gas)		

1. Estimate a linear regression model: $lchnimp=f(lchempi, lgas, lrtwex, befile6, affile6, afdec6)$

2. Get the Durbin Watson statistic and interpret it.

. use <http://fmwww.bc.edu/ec-p/data/wooldridge/BARIUM>

. quietly regress lchnimp lchempi lgas lrtwex befile6 affile6 afdec6

. tsset t

time variable: t, 1 to 131

delta: 1 unit

. estat dwatson

Durbin-Watson d-statistic(7, 131) = 1.458417

Using any textbook table, and noting that $k'=6$, the table for $n=100$ indicates that $d_L=1.55$.

Our statistic is below this critical level, thus, there is statistical evidence of autocorrelation in the error structure.

3. Run and interpret the Breusch-Godfrey Test

```
. estat bgodfrey
```

Breusch-Godfrey LM test for autocorrelation

lags(p)	chi2	df	Prob > chi2
1	9.829	1	0.0017

H0: no serial correlation

This statistic reveals considerable evidence of first order autocorrelation.

Note: To get the Breusch-Godfrey result in longhand:

```
. predict uhat, resid
. gen uhat_lag1=uhat[_n-1]
. quietly regress uhat lchempi lgas lrtwex befile6 affile6 afdec6 uhat_lag1
. di e(N)*e(r2)
9.7534721
. di chi2tail(1,9.75347)
.00178985
```

4. For $p=3$, calculate the Box Pierce and Ljung-Box statistics and interpret.

```
. predict uhat, resid
. gen uhatsq=uhat*uhat
. egen r_denom=total(uhatsq)
. gen z1=(uhat*uhat[_n-1])
. gen z2=(uhat*uhat[_n-2])
. gen z3=(uhat*uhat[_n-3])
. egen r1_num=total(z1)
. egen r2_num=total(z2)
. egen r3_num=total(z3)
. gen r1=r1_num/r_denom
. gen r2=r2_num/r_denom
```

```
. gen r3=r3_num/r_denom
```

Box Pierce Test Statistic=

```
. di e(N)*(r1^2+r2^2+r3^2)
```

```
20.725208
```

```
. di chi2tail(3,20.725)
```

```
.00012007
```

Reject the null of no autocorrelation using the Box Pierce test.

Ljung-Box Test Statistic=

```
. di e(N)*(e(N)+2)*((r1^2/(e(N)-1))+r2^2/(e(N)-2))+r3^2/(e(N)-3))
```

```
21.332487
```

```
. di chi2tail(3,21.332)
```

```
.00008982
```

Reject the null of no autocorrelation using the Ljung-Box test.

5. Run the Prais-Winsten procedure

```
. prais lchnimp lchempi lgas lrtwex befile6 affile6 afdec6
```

```
Iteration 0: rho = 0.0000
```

```
Iteration 7: rho = 0.2932
```

Prais-Winsten AR(1) regression -- iterated estimates

Source	SS	df	MS	Number of obs =	131
-----+-----				F(6, 124) =	5.24
Model	10.3252189	6	1.72086981	Prob > F	= 0.0001
Residual	40.7594601	124	.328705323	R-squared	= 0.2021
-----+-----				Adj R-squared =	0.1635
Total	51.084679	130	.392959069	Root MSE	= .57333

	lchnimp	lchempi	lgas	lrtwex	befile6	affile6	afdec6
Coef.	2.940963	2.940963	1.046299	1.132774	-.0164782		
Std. Err.	.6328381	.6328381	.9773411	.5066564	.3193797		
t	4.65	4.65	1.07	2.24	-0.05		
P> t	0.000	0.000	0.286	0.027	0.959		
[95% Conf. Interval]	1.6884	1.6884	-.8881332	.1299595	-.64862		
	4.193527	4.193527	2.980731	2.135589	.6156637		

affile6	-.0331578	.3218095	-0.10	0.918	-.6701089	.6037933
afdec6	-.5768112	.341986	-1.69	0.094	-1.253697	.1000749
_cons	-37.07582	22.77843	-1.63	0.106	-82.16072	8.009074

rho	.2932139
-----	----------

Durbin-Watson statistic (original) 1.458417
Durbin-Watson statistic (transformed) 2.087180

- Describe what the iterative process is doing The model stops when the optimal value for rho is identified. If first identifies the rho variable from the initial regression. Then the data is transformed. The model is rerun and a new rho is calculated. If the new rho is different from the original rho, the new rho is used to transform the data. This process continues until the rho is stabilized.