9. Heteroskedasticity
Cross-Sectional Analysis

Read Wooldridge (2013), Chapter 8

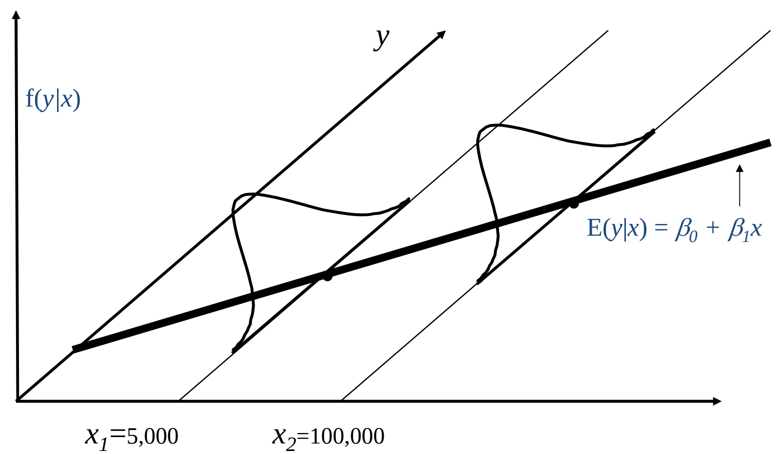
Outline

I. Consequences of Heteroskedasticity

- Motivation: Consider the model
  $sav_i = \beta_0 + \beta_1 inc_i + u_i$
  y = sav = *saving*
  x = *inc* = *income*

- Constant variances (MLR. 5)
  $\text{Var}(u_i | inc_i) = \sigma^2$,
  which implies that
  $\text{Var}(sav_i | inc_i) = \sigma^2$

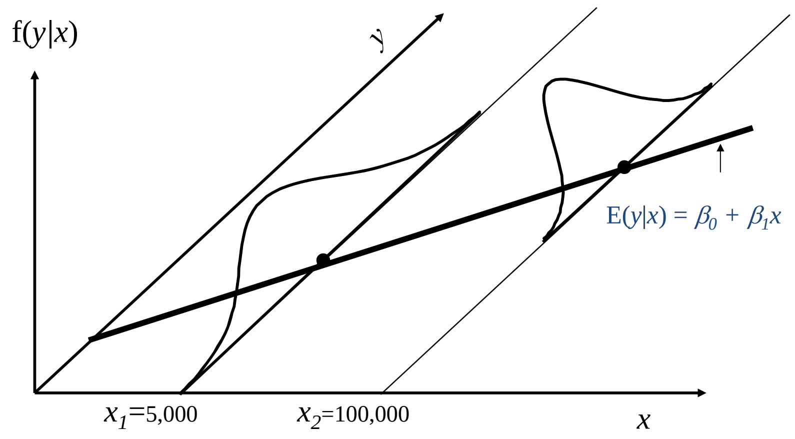
Homoskedastic Case



Heteroskedasticity

- Violation of homoskedasticity:
  What if the variability of savings of the rich is less than that of the lower income group?
- Here we say that the variance of savings – y (or unobserved factors – u) increases with income
  VAR(sav_i|inc_i = 5,000) = σ^2_5 (see x_1)
  VAR(sav_i|inc_i = 100,000) = σ^2_{100} (see x_2)
- When variances are unequal, this problem is called heteroskedasticity. (See Graph)

Example of Heteroskedasticity



MLR.5 violated: Heteroskedasticity

- Consider a model
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
  Heteroskedasticity
  VAR(u_i|$x_1,\ldots,x_k$) = σ_i^2

Properties unaffected by heteroskedasticity:
1) OLS estimators are still unbiased and consistent.
2) The interpretation is the same for goodness-of-fit measures, R^2 and R^2-bar.

Properties invalid under heteroskedasticity: unequal variances

1) The estimators of the variances, Var($\hat{\beta}_j$), are biased.
2) t, F and LM statistics no longer have t, F and LM distributions.
3) OLS is no longer best linear unbiased estimator (BLUE).

II. Testing for Heteroskedasticity

- There are many tests for heteroskedasticiy, but we will learn two modern tests:
  1) Breusch-Pagan Test for Heteroskedasticity
  2) White Test
     - use no cross terms.
     - use cross terms.
     - use fitted values of the LHS variable

- These modern tests assume that the variance of the error depends or does not depend upon the explanatory variables.

Breusch-Pagan Test

- Given MLR.1-MLR.4, consider the Model

  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$

- Want to test H_0: whether MLR.5 is true

  $H_0 : VAR(u_i | x_1, ..., x_3) = E(u_i^2) = \sigma^2$

  $u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + v$

- To test whether u_i^2 is related to x_i's

  $H_0 : \delta_1 = \delta_2 = ... = \delta_k = 0$

Use residuals for u_i^2

- Since we don't observe u_i, but we have estimates of $\hat{u}_i$ - residuals.

  $\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + v$

- Use F-test or LM test to test the overall significance

  $H_0 : \delta_1 = \delta_2 = ... = \delta_k = 0$

$$F = \frac{R^2_{\hat{u}^2} / k}{(1 - R^2_{\hat{u}^2}) / n - k - 1} \sim F_{k,(n-k-1)}$$

$$LM = n * R^2_{\hat{u}^2} \sim \chi_k^2$$

Example: Consider cigarette demand function

- *income*: annual income in dollars
- *cigpric*: state cigarette price, cents per pack
- *educ*: years of schooling
- *age*: age measured in years
- *restaurn* =1 if a state has restaurant smoking restrictions
  =0 if a state has no restaurant smoking restrictions

$$cigs = \beta_0 + \beta_1 \log(income) + \beta_2 \log(cigpric) + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 restaurn + u$$

BP Test for heteroskedasticity

Step 1: Estimate the above equation.

Step 2: obtain residuals from the *cigs* equation or $\hat{u}$

In Eviews, obtain residual series – *resid01*

Step 3: Regress $\hat{u}_i^2$ on all x_i's.

$resid01\text{^}2 = \beta_0 + \beta_1 \log(income) + \beta_2 \log(cigpric) + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 restaurn + u$

Step 4: Use F and LM Tests for heteroskedasticity. Compute F and LM statistics and compare to critical values of the $F_{k,n-k-1}$ and χ^2_k distributions

LM version of the Bruesch-Pagan Test

Step 4: Use F-test and LM-test

1) F-statistic = 5.55

   p-value = (0.000012)

2) LM statistic = Obs*R-squared

   = 807*.039973 = 32.26

   Chi square distribution with 6 DFs

   - c = 12.59 (5% significance level)
   - c = 16.81(1% significance level)

   – What can we say about heteroskedasticity?

Eviews: Step 1: Estimate the cigarette demand equation

Dependent Variable: CIGS
Sample: 1 807
Included observations: 807

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.639823	24.07866	-0.151164	0.8799
LOG(INCOME)	0.880268	0.727783	1.209519	0.2268
LOG(CIGPRIC)	-0.750862	5.773342	-0.130057	0.8966
EDUC	-0.501498	0.167077	-3.001596	0.0028
AGE	0.770694	0.160122	4.813155	0
AGE^2	-0.009023	0.001743	-5.176494	0
RESTAURN	-2.825085	1.111794	-2.541016	0.0112

R-squared	0.052737	Mean dependent var	8.686493
Adjusted R-squared	0.045632	S.D. dependent var	13.72152
S.E. of regression	13.40479	Akaike info criterion	8.037737
Sum squared resid	143750.7	Schwarz criterion	8.078448
Log likelihood	-3236.227	F-statistic	7.423062
Durbin-Watson stat	2.012825	Prob(F-statistic)	0

Proc/Make Residual Series. Step 2: name for residual series: *resid01*

Step 3: Regress resid01^2 ($\hat{u}_i^2$) on all x_i's

Dependent Variable: RESID01^2
Sample: 1 807

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-636.303	652.4945	-0.975186	0.3298
LOG(INCOME)	24.63847	19.7218	1.249302	0.2119
LOG(CIGPRIC)	60.97663	156.4487	0.389755	0.6968
EDUC	-2.38423	4.527535	-0.526606	0.5986
AGE	19.41748	4.339068	4.475034	0
AGE^2	-0.21479	0.047234	-4.547398	0
RESTAURN	-71.1814	30.12789	-2.362641	0.0184

R-squared	0.039973	Mean dependent var	178.1297
Adjusted R-squared	0.032773	S.D. dependent var	369.3519
S.E. of regression	363.2491	Akaike info criterion	14.63669
Sum squared resid	1.06E+08	Schwarz criterion	14.6774
Log likelihood	-5898.91	**F-statistic**	**5.551687**
Durbin-Watson stat	1.937302	**Prob(F-statistic)**	**0.000012**

White Test of Heteroskedasticity

- Consider the three-variable model

  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$

  $H_0 : Var(u_i | x_1, x_2, x_3) = \sigma^2$

- Weaker assumption by White (1980)

  u^2 is uncorrelated with $(x_1, x_2, x_3), (x_1^2, x_2^2, x_3^2), (x_1x_2, x_1x_3, x_2x_3)$

  $\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + v$

- $H_0 : \delta_1 = \delta_2 = ... = \delta_9 = 0$

  Use F and LM Test

  What are the rejection rules?

Intractable: more regressors

- Consider the model with 6 regressors

  $cigs = \beta_0 + \beta_1 \log(income) + \beta_2(cigpric) + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 restaurn + u$

- $\hat{u}_i^2 = \delta_0 +$ 25 regressors + v
  - LOG(INCOME), (LOG(INCOME))^2, (LOG(INCOME))*(LOG(CIGPRIC)), (LOG(INCOME))*EDUC, (LOG(INCOME))*AGE, (LOG(INCOME))*(AGE^2), (LOG(INCOME))*RESTAURN,
  - LOG(CIGPRIC), (LOG(CIGPRIC))^2, (LOG(CIGPRIC))*EDUC, (LOG(CIGPRIC))*AGE, (LOG(CIGPRIC))*(AGE^2), (LOG(CIGPRIC))*RESTAURN
  - EDUC, EDUC^2, EDUC*AGE, EDUC*(AGE^2), EDUC*RESTAURN
  - AGE, AGE^2, AGE*(AGE^2), AGE*RESTAURN, (AGE^2)^2, (AGE^2)*RESTAURN
  - RESTAURN

- $H_0 : \delta_1 = \delta_2 = \ldots = \delta_{25} = 0$

  k = 25

  n – k – 1 = n – 26

  What are the rejection rules?

Easier way to implement White Test

- Idea : use OLS fitted values in a test for heteroskedasticity

  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$

  When we square $\hat{y}$, we get a particular function of all the squares and cross products

- Simpler form of White Test

  $\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + v$

  – What are the null hypothesis and rejection rules?

Model:

$cigs = \beta_0 + \beta_1 \log(income) + \beta_2 \log(cigpric) + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 restaurn + u$

A Special case of the White Test for heteroskedasticity

Step 1: Estimate the above equation.

Step 2: obtain fitted value from the cigarette equation.

In Eviews, obtain residual series – *resid01* (or $\hat{u}_i$)

Note that $y_i = \hat{y}_i + \hat{u}_i$ or $cigs_i = \widehat{cigs}_i + \hat{u}_i$

Generating series for $\widehat{cigs}_i$ called *cigshat*

cigshat = *cigs*$_i$ – *resid01*

Step 3: Regress $\hat{u}_i^2$ on $\widehat{cigs}_i$ and $\widehat{cigs}_i^2$

$resid01\text{^}2 = \delta_0 + \delta_1\, cigshat + \delta_2\, cigshat^2 + v$

Step 4: Use F and LM Tests for heteroskedasticity. Compute F and LM statistics and compare to critical values of the $F_{2,n-3}$ and χ^2_2 distributions

Eviews: Step 1: Estimate the cigarette demand equation

Dependent Variable: CIGS
Sample: 1 807
Included observations: 807

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.639823	24.07866	-0.151164	0.8799
LOG(INCOME)	0.880268	0.727783	1.209519	0.2268
LOG(CIGPRIC)	-0.750862	5.773342	-0.130057	0.8966
EDUC	-0.501498	0.167077	-3.001596	0.0028
AGE	0.770694	0.160122	4.813155	0
AGE^2	-0.009023	0.001743	-5.176494	0
RESTAURN	-2.825085	1.111794	-2.541016	0.0112

R-squared	0.052737	Mean dependent var	8.686493
Adjusted R-squared	0.045632	S.D. dependent var	13.72152
S.E. of regression	13.40479	Akaike info criterion	8.037737
Sum squared resid	143750.7	Schwarz criterion	8.078448
Log likelihood	-3236.227	F-statistic	7.423062
Durbin-Watson stat	2.012825	Prob(F-statistic)	0

Step 2: Generating series for *cigs$_i$^* called *cigshat*

Step 3: Regress *resid01^2* on $\widehat{cigs}_i$ and $\widehat{cigs}_i$^2

Dependent Variable: RESID01^2				
Method: Least Squares				
Sample: 1 807				
Included observations: 807				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	14.05341	47.79852	0.294013	0.7688
CIGSHAT	14.05344	11.56743	1.214914	0.2248
CIGSHAT^2	0.491978	0.755633	0.65108	0.5152
R-squared	0.032928	Mean dependent var	178.1297	
Adjusted R-squared	0.030522	S.D. dependent var	369.3519	
S.E. of regression	363.6715	Akaike info criterion	14.63409	
Sum squared resid	1.06E+08	Schwarz criterion	14.65154	
Log likelihood	-5901.86	F-statistic	13.6876	
Durbin-Watson stat	1.93512	Prob(F-statistic)	0.000001	

Step 4. LM and F-test for White Test

- $resid01\text{^}2 = \delta_0 + \delta_1\, cigshat + \delta_2\, cigshat^2 + v$
  $H_0: \delta_1 = \delta_2 = 0$
- F-test
  F =13.6876
  p-value = .000001
  Could you find critical values to verify this?
- LM test
  LM = obs*R-squared = 807*.032928 = 26.57
  Chi square distribution with 2 DFs
  – c = 5.99 (5% significance level)
  – c = 9.21 (1% significance level)

Model:

$cigs = \beta_0 + \beta_1 \log(income) + \beta_2 \log(cigpric) + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 restaurn + u$

- We could follow similar steps to test for heteroskedasticity using White Tests.

3) White Test with no cross terms
   $\hat{u}_i^2 = \beta_0 + \beta_1 \log(income) + \beta_2 \log(cigpric) + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 restaurn + \beta_7[\log(income)]^2 + \beta_8[\log(cigpric)]^2 + \beta_9 educ^2 + \beta_{10}[age^2]^2 + v$
4) White Test with cross terms.
   $\hat{u}_i^2 = \beta_0 + 25 \text{ regressors} + v$

- Eviews:
  It has commands to find F and LM statistics using White Tests (methods 3-4).
  In the equation output window,
  - (3) Choose View/Residual Tests/White heteroskedasticity (with no cross terms)
  - (4) Choose View/Residual Tests/White heteroskedasticit (with cross terms) for (4)

In the "Equation" window, run the *following* regression.

Dependent Variable: CIGS				
Sample: 1 807				
Included observations: 807				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.639823	24.07866	-0.151164	0.8799
LOG(INCOME)	0.880268	0.727783	1.209519	0.2268
LOG(CIGPRIC)	-0.750862	5.773342	-0.130057	0.8966
EDUC	-0.501498	0.167077	-3.001596	0.0028
AGE	0.770694	0.160122	4.813155	0
AGE^2	-0.009023	0.001743	-5.176494	0
RESTAURN	-2.825085	1.111794	-2.541016	0.0112
R-squared	0.052737	Mean dependent var	8.686493	
Adjusted R-squared	0.045632	S.D. dependent var	13.72152	
S.E. of regression	13.40479	Akaike info criterion	8.037737	
Sum squared resid	143750.7	Schwarz criterion	8.078448	
Log likelihood	-3236.227	F-statistic	7.423062	
Durbin-Watson stat	2.012825	Prob(F-statistic)	0	

View/Residual Tests/White heteroskedasticity (no cross terms) for (3)

White Heteroskedasticity Test:			
F-statistic	3.732565	Probability	0.000065
Obs*R-squared	36.14649	Probability	0.000079

Dependent Variable: RESID^2
Included observations: 807

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	18981.92	18710.46	1.014509	0.3106
LOG(INCOME)	16.21011	277.0494	0.05851	0.9534
(LOG(INCOME))^2	0.413056	15.27585	0.02704	0.9784
LOG(CIGPRIC)	-9800.158	9251.297	-1.059328	0.2898
(LOG(CIGPRIC))^2	1219.708	1144.864	1.065374	0.287
EDUC	16.17731	27.71945	0.583609	0.5596
EDUC^2	-0.782004	1.09046	-0.717132	0.4735
AGE	35.95817	11.20129	3.210181	0.0014
AGE^2	-0.499351	0.185168	-2.69675	0.0071
(AGE^2)^2	1.92E-05	1.22E-05	1.572888	0.1161
RESTAURN	-64.69514	30.44613	-2.124905	0.0339

R-squared	0.044791	Mean dependent var	178.1297
Log likelihood	-5896.875	F-statistic	3.732565
Durbin-Watson stat	1.937056	Prob(F-statistic)	0.000065

View/Residual Tests/White heteroskedasticity (with cross terms) for (4)

White Heteroskedasticity Test:				
F-statistic	2.159258	Probability		0.000905
Obs*R-squared	52.17245	Probability		0.00114
Dependent Variable: RESID^2				
Included observations: 807				
Variable (25 regressors)	Coefficient	Std. Error	t-Statistic	Prob.
C	29374.77	20559.14	1.428794	0.1535
LOG(INCOME)	-1049.63	963.4359	-1.08947	0.2763
(LOG(INCOME))^2	-3.94118	17.07122	-0.23087	0.8175
(LOG(INCOME))*(LOG(CIGPRIC))	329.8896	239.2417	1.378897	0.1683
..	..	..	..	..
RESTAURN	-2868.2	2986.776	-0.9603	0.3372
R-squared Mean	0.06465	dependent var		178.1297
Log likelihood	-5888.4	F-statistic		2.159258
Durbin-Watson stat	1.933288	Prob(F-statistic)		0.000905

III. Heteroskedasticity – Robust Inference after OLS Estimation

- In the presence of heteroskedasticity, Eviews can adjust standard errors, t, F, and LM, statistics so that they are valid.
  - This method is called the heteroskedastic-robust procedure.
- Technically, this procedure is valid, at least in large samples, whether or not the errors have constant variances.

Sketch the procedure

- Consider the simple regression model
  $y_i = \beta_0 + \beta_1 x_i + u_i$
  $\mathrm{Var}(u_i|x_i) = \sigma_i^2$
- Steps to find robust standard errors:
  1) Find estimator of β_1
  2) Under the assumptions MLR.1 – MLR.4, the variance can be found.
  3) White(1980) suggests using $\hat{u}^2$ in place of σ_i^2. Thus, we can find the estimator of $\mathrm{VAR}(\hat{\beta}_j)$
  4) A heteroskedastic-robust standard error can be found.
- Sketch for the general case?

Variance with Heteroskedasticity

For the simple case,

$$\hat{\beta}_1 = \beta_1 + \frac{\sum (x_i - \bar{x}) u_i}{\sum (x_i - \bar{x})^2}, \text{ so}$$

$$Var(\hat{\beta}_1) = \frac{\sum (x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}, \text{ where } SST_x = \sum (x_i - \bar{x})^2$$

A valid estimator for this when $\sigma_i^2 \neq \sigma^2$ is

$$\frac{\sum (x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2},$$

where $\hat{u}_i$ are are the OLS residuals

Variance with Heteroskedasticity

For the general multiple regression model, a valid estimator of $Var(\hat{\beta}_j)$ with heteroskedasticity is

$$Va\hat{r}(\hat{\beta}_j) = \frac{\sum \hat{r}_{ij}\hat{u}_i^2}{SSR_j^2},$$

where $\hat{r}_{ij}$ is the i^{th} residual from regressing x_j on all other independent variables, and

SSR_j is the sum of squared residuals from this regression

Example: *cigs* equation

- Consider the model

  $cigs = \beta_0 + \beta_1 \log(income) + \beta_2 \log(cigpric) + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 restaurn + u$

- Find the *cigs* equation using heteroskedastic-robust procedure.

Steps in Eviews:

- Step 1: Estimate the log equation in usual OLS method.

  $cigs = \beta_0 + \beta_1 \log(income) + \beta_2 \log(cigpric) + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 restaurn + u$

- Step 2: Find the log equation with heteroskedasticity-robust standard errors.

  In the *"Equation"* window,
  - Choose Estimate.

  In the *"Equation Estimation"* box,
  - click *option* button. Then,
  - click "*heteroskedasticity consistent coefficient covariance*."
    - click "White"

Step 1: Estimate the *cigs* equation in usual way

Dependent Variable: CIGS
Included observations: 807

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.639823	24.07866	-0.151164	0.8799
LOG(INCOME)	0.880268	0.727783	1.209519	0.2268
LOG(CIGPRIC)	-0.750862	5.773342	-0.130057	0.8966
EDUC	-0.501498	0.167077	-3.001596	0.0028
AGE	0.770694	0.160122	4.813155	0
AGE^2	-0.009023	0.001743	-5.176494	0
RESTAURN	-2.825085	1.111794	-2.541016	0.0112

R-squared	0.052737	Mean dependent var	8.686493
Adjusted R-squared	0.045632	S.D. dependent var	13.72152
S.E. of regression	13.40479	Akaike info criterion	8.037737
Sum squared resid	143750.7	Schwarz criterion	8.078448
Log likelihood	-3236.227	F-statistic	7.423062
Durbin-Watson stat	2.012825	Prob(F-statistic)	0

Step 2: White Heteroskedasticity-Consistent s.e.'s

Dependent Variable: CIGS
Included observations: 807
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.63982	25.61646	-0.142089	0.887
LOG(INCOME)	0.880268	0.596011	1.476931	0.1401
LOG(CIGPRIC)	-0.75086	6.035401	-0.12441	0.901
EDUC	-0.5015	0.162394	-3.088167	0.0021
AGE	0.770694	0.138284	5.573262	0
AGE^2	-0.00902	0.001462	-6.170768	0
RESTAURN	-2.82509	1.008033	-2.802573	0.0052

R-squared	0.052737	Mean dependent var	8.686493
Adjusted R-squared	0.045632	S.D. dependent var	13.72152
S.E. of regression	13.40479	Akaike info criterion	8.037737
Sum squared resid	143750.7	Schwarz criterion	8.078448
Log likelihood	-3236.23	F-statistic	7.423062
Durbin-Watson stat	2.012825	Prob(F-statistic)	0

OLS and Robust estimates: compared

Dependent Variable: CIGS

Variable	Coefficient	s.e	s.e (robust)	Prob.	Prob. (robust)
C	-3.639823	24.07866	25.61646	0.8799	0.887
LOG(INCOME)	0.880268	0.727783	0.596011	0.2268	0.1401
LOG(CIGPRIC)	-0.750862	5.773342	6.035401	0.8966	0.901
EDUC	-0.501498	0.167077	0.162394	0.0028	0.0021
AGE	0.770694	0.160122	0.138284	0	0
AGE^2	-0.009023	0.001743	0.001462	0	0
RESTAURN	-2.825085	1.111794	1.008033	0.0112	0.0052

Interpretation: cigs demand equation

$$cigs = \beta_0 + \beta_1 \log(income) + \beta_2 \log(cigpric) + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 restaurn + u$$

1) In this application, any variable that is statistically significant at 1% level using the usual t test is still significant under the heteroskedasticity-robust procedure.
   variables: *educ, age,* age^2 and *restaurn.*
2) The robust standard errors are either larger or smaller than the usual standard error
3) The robust s.e. on log(*income*) becomes smaller, but that on log(*cigprice*) is larger.
4) How to interpret the coefficients of various variables in the model?

Robust F-test and Wald Test

$cigs = \beta_0 + \beta_1 \log(income) + \beta_2 \log(cigpric) + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 restaurn + u$

- Want to test the null hypothesis:

  $H_0: \beta_1 = \beta_2 = 0$

  In the equation with robust standard errors, we can easily obtain heteroskedastic robust F-statistic – also called heteroskedastic robust Wald statistic.

- Step 1. White Heteroskedasticity-Consistent s.e.'s
  Step 2. View/Coefficient Tests/Wald-Coefficient Restrictions
  Step 3. Type in c(2)=0, c(3)=0
- The F-statistic is

  F = 0.733326; p-value =.480631 (incorrect)

  robust F-statistic = 1.099; p-value=.3338 (correct)

  Since p-value > α, we do not reject H_0 and conclude that income and price together do not have an effect on cigarette demand.

Step 1: White Heteroskedasticity-Consistent s.e.'s

Dependent Variable: CIGS
Included observations: 807
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.63982	25.61646	-0.142089	0.887
LOG(INCOME)	0.880268	0.596011	1.476931	0.1401
LOG(CIGPRIC)	-0.75086	6.035401	-0.12441	0.901
EDUC	-0.5015	0.162394	-3.088167	0.0021
AGE	0.770694	0.138284	5.573262	0
AGE^2	-0.00902	0.001462	-6.170768	0
RESTAURN	-2.82509	1.008033	-2.802573	0.0052

R-squared	0.052737	Mean dependent var	8.686493
Adjusted R-squared	0.045632	S.D. dependent var	13.72152
S.E. of regression	13.40479	Akaike info criterion	8.037737
Sum squared resid	143750.7	Schwarz criterion	8.078448
Log likelihood	-3236.23	F-statistic	7.423062
Durbin-Watson stat	2.012825	Prob(F-statistic)	0

~~View/Coefficient Tests/Redundant Variables~~ …

~~Redundant Variables: LOG(INCOME) LOG(CIGPRIC)~~			
~~F-statistic~~	~~0.733326~~	~~Probability~~	~~0.480631~~
~~Log likelihood ratio~~	~~1.478131~~	~~Probability~~	~~0.47756~~

View/Coefficient Tests/Wald Coefficient Restrictions

Wald Test:			
Equation: Untitled			
Test Statistic	Value	df	Probability
F-statistic	1.098858	(2, 800)	0.3338
Chi-square	2.197717	2	0.3333
Null Hypothesis Summary:			
Normalized Restriction (= 0)		Value	Std. Err.
C(2)		0.880268	0.596011
C(3)		-0.75086	6.035401
Restrictions are linear in coefficients.			

IV. Weighted Least Squares

- Assume that MLR.1-MLR.4 hold, but MLR.5 does not.

  Let $\underline{x}$ denote $x_1, x_2, \ldots, x_k$

  $VAR(u_i|\underline{x}) = \sigma_i^2$ (heteroskedasticity)

  Let $\sigma_i^2 = \sigma^2 h(\underline{x})$

  $\therefore VAR(u_i|\underline{x}) = \sigma^2 h(\underline{x})$ $h(\underline{x}) > 0$ (since VAR > 0)

- Heteroskedastity can be corrected under two cases:
  1) $\sigma_i^2 = \sigma^2 h(\underline{x})$. $h(\underline{x})$ is known up to a multiplicative constant
  2) $h(\underline{x})$ has to be estimated – feasible GLS

Case 1: h(x) is known up to a multiplicative constant

- Consider the model
  Let $y_i = sav_i$; $x_i = inc_i$; $h_i = x_i$
  $y_i = \beta_0 + \beta_1 x_i + u_i$ (OLS)
  $\text{var}(u_i|x_i) = \sigma^2 h_i$ (heteroskedastic)
- Trick : Divide the equation by sqr(h_i)

$$\frac{y_i}{\sqrt{h_i}} = \beta_0 \frac{1}{\sqrt{h_i}} + \beta_1 \frac{x_i}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}} \quad \text{(IV.1)}$$

- Show that the errors in (IV.1) are homoskedastic!

Weighted Least Squares

- Weighted least squares – obtain the values of β_j^* that makes the weighted SSR as small as possible:

$$\sum_{i=1}^{n} \frac{\hat{u}_i^2}{h_i} = \sum_{i=1}^{n} \left[(y_i - \beta_0^* - \beta_1^* x_{i1} - \dots - \beta_k^* x_{ik})^2 \right] / h_i$$

  where each squared residual is weighted by $1/h_i$.
- Bring $1/h_i$ inside the squared residual:

$$\sum_{i=1}^{n} \frac{\hat{u}_i^2}{h_i} = \sum_{i=1}^{n} \left(\frac{y_i}{\sqrt{h_i}} - \beta_0^* \frac{1}{\sqrt{h_i}} - \beta_1^* \frac{x_{i1}}{\sqrt{h_i}} - \dots - \beta_k^* \frac{x_{ik}}{\sqrt{h_i}} \right)^2$$

- GLS is an efficient procedure.

Example: Saving equation

- OLS : $sav_i = \beta_0 + \beta_1 inc_i + u_i$
  $\text{VAR}(u_i|inc_i) = \sigma_i^2 = \sigma^2 inc_i$ is not constant.
  $\hat{\beta}_0$ and $\hat{\beta}_1$ are not BLUE.
- GLS : transformed equation
  $sav_i/(inc)^{1/2} = \beta_0/(inc)^{1/2} + \beta_1 inc_i/(inc)^{1/2} + u_i/(inc)^{1/2}$
  $\text{VAR}[u_i/inc_i] = \sigma^2$ is constant.
  β_0^* and β_1^* are BLUE
- The GLS estimators after correcting the error for heteroskedasticity is called weighted least squares (WLS) estimators.

OLS & WLS compared

	OLS	WLS
MLR.1-MLR-4	yes	yes
MLR.5	heteroskedasticity	homoskedasticity
error variance	not constant	constant
BLUE	no	yes
t and F Dist	invalid	valid
R-squared	meaningful	not meaningful

OLS and WLS Results: saving equation

OLS: $sav = 124.8 + 0.147inc$
prob {.8493} {0.0124}
n=100, R^2=.0621 R^2-bar=.0526

WLS: $sav/inc^{.5} = -125.0[1/inc^{.5}] + 0.172inc^{.5}$
prob {.7955} {.0032}
n=100, R^2=.0225 R^2-bar=.0125

- Eview Trick: WLS
  Step 1: Obtain Original output
  Step 2: In the [Equation:..] window,
  Choose Estimate and then options.
  click Weighted LS/TSLS
  Type in the weight, *1/inc^.5* or *1/sqr(inc)*

Step 1: OLS – Regress *sav* on *inc (with intercept)*

Dependent Variable: SAV
Method: Least Squares
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	124.8424	655.3931	0.190485	0.8493
INC	0.146628	0.057549	2.547897	0.0124

R-squared	0.062127	Mean dependent var	1582.51
Adjusted R-squared	0.052557	S.D. dependent var	3284.902
S.E. of regression	3197.415	Akaike info criterion	18.99787
Sum squared resid	1.00E+09	Schwarz criterion	19.04997
Log likelihood	-947.894	F-statistic	6.491778
Durbin-Watson stat	1.536387	Prob(F-statistic)	0.012391

WLS: Regress sav/(inc^.5) on 1/inc^.5 and inc^.5 (with no intercept)

Dependent Variable: SAV/(INC^.5)
Method: Least Squares
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
1/(INC^.5)	-124.953	480.8606	-0.259852	0.7955
INC^.5	0.171756	0.056813	3.023184	0.0032

R-squared	0.022487	Mean dependent var	15.25151
Adjusted R-squared	0.012512	S.D. dependent var	29.89943
S.E. of regression	29.71179	Akaike info criterion	9.640762
Sum squared resid	86513.48	Schwarz criterion	9.692866
Log likelihood	-480.038	Durbin-Watson stat	1.567035

Step 2: Compare to the regression of
$sav_i/(inc_i)1/2 = \beta_0/(inc_i)^{1/2} + \beta_1 inc_i/(inc_i)^{1/2} + u_i/(inc_i)^{1/2}$

Dependent Variable: SAV
Included observations: 100
Weighting series: 1/INC^.5

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-124.953	480.8606	-0.25985	0.7955
INC	0.171756	0.056813	3.023184	0.0032

Weighted Statistics			
R-squared	0.022487	Mean dependent var	1364.931
Adjusted R-squared	0.012512	S.D. dependent var	2675.843
S.E. of regression	2659.05	Akaike info criterion	18.62912
Sum squared resid	6.93E+08	Schwarz criterion	18.68123
Log likelihood	-929.456	F-statistic	9.139642
Durbin-Watson stat	1.567035	Prob(F-statistic)	0.003192
Unweighted Statistics			
R-squared	0.060303	Mean dependent var	1582.51
Adjusted R-squared	0.050714	S.D. dependent var	3284.902
S.E. of regression	3200.523	Sum squared resid	1.00E+09
Durbin-Watson stat	1.559211		

WLS: In practice, we rarely know h(*x*). Consider the model:
$sav_i = \beta_0 + \beta_1 inc_i + \beta_2 size_i + \beta_3 educ_i + \beta_4 age_i + \delta_0 black_i + u_i$
Does the variance depend on "age" or "education"?

Individual level data vs. Averages of data

- There is a case where weights needed arise naturally.

Example: the effect of earnings on the contribution to Private Provident Fund.

Individual data:

$contrib_{i,e} = \beta_0 + \beta_1 earn_{i,e} + u_{i,e}$

Assume MLR.1-MLR.4 and $Var(u_{i,e}) = \sigma^2$

i : denote a particular firm

e: an employee within the firm

$contrib_{i,e}$: annual contribution

$earn_{i,e}$: annual earnings

Average data at the firm level:

$$\overline{contrib}_i = \beta_0 + \beta_1 \overline{earn}_i + \bar{u}_i$$

VAR(*u*-bar) = σ^2/m_i (Heteroskedastic)

$h_i = 1/m_i$

weight = m_i

= $m_i^{1/2}$(in the transformed equation and Eviews)

Show the error in the transformed equation is homoskedastic!

WLS: to use with per-capita data

- A similar weighting arises when we use data at the city, province, or country level.
- In summary, WLS gives us an efficient way to treat averages of data.

Case 2: h(x) must be estimated

- Suppose the heteroskedasticity function is unknown
  i.e., $VAR(u_i|x_1, ..., x_k) = \sigma^2 h(\underline{x})$
  where *h(x)* is unknown and must be estimated; i.e., find $\hat{h}_i$.
- Using $\hat{h}_i$ in GLS transformation yields an estimator, called an FGLS estimator.
- FGLS is no longer unbiased but consistent in large samples.
  FGLS is no longer BLUE but asymptotically more efficient than OLS.

Feasible GLS Estimator, $\hat{h}_i$

- Assume that

  $VAR(u|\underline{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + ... + \delta_k x_k)$

  $u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + ... + \delta_k x_k) v$

  $\log(u^2) = \alpha_0 + \delta_1 x_1 + ... + \delta_k x_k + e$

- Estimate

  $\log(\hat{u}^2) = \alpha_0 + \delta_1 x_1 + ... + \delta_k x_k + e$

  $\hat{g}_i$ = fitted value of $\log(\hat{u}^2)$

  The estimates of h_i are simply

  $\hat{h}_i = \exp(\hat{g}_i)$

Properties: OLS, GLS and FGLS

	OLS	known h_i	$\hat{h}_i$: FGLS
Unbiasedness	yes	yes	**no longer unbiased but consistent**
BLUE	no	yes	**no longer BLUE but asymptotically more efficient**
t and F dist.	no	exact t and F distributions	***approximately* t and F distributions**

Example: FGLS and cigarette equation

$cigs = \beta_0 + \beta_1 \log(income) + \beta_2 \log(cigpric) + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 restaurn + u$

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + u$

- Using Breusch-Pagan Test or White Tests, we found that the variances are nonconstant.

Steps in running FGLS equation

Step 1: Run the regression of
y on $x_1, x_2, ..., x_6$ and obtain residuals $\hat{u}_i$ (called, *resid01*) in Eviews.

Step 2: Run the regression of $\log(\hat{u}_i^2)$ or
log(*resid01^2*) on $x_1, ..., x_k$
and obtain residuals $\hat{u}_i$ (called, *lresid02* in Eviews)

Step 3: From step 2, obtain $\hat{g}_i$ and $\hat{h}_i$
and the fitted values of $\log(\hat{u}_i^2)$ (called $\hat{g}_i$)
$\hat{g}_i$ = log(*resid01^2*)-*lresid02*
Since h_i^ = exp(g_i^), then in Eview
h01 = exp(log(*resid01^2*)-*lresid02*)

Step 4: Run the FGLS equation using 1/*h01* as weights. ($1/\sqrt{h01}$ *are weights in Eveiws)*

Eviews: Step 1: Estimate the cigarette demand equation

Dependent Variable: CIGS
Sample: 1 807
Included observations: 807

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.639823	24.07866	-0.151164	0.8799
LOG(INCOME)	0.880268	0.727783	1.209519	0.2268
LOG(CIGPRIC)	-0.750862	5.773342	-0.130057	0.8966
EDUC	-0.501498	0.167077	-3.001596	0.0028
AGE	0.770694	0.160122	4.813155	0
AGE^2	-0.009023	0.001743	-5.176494	0
RESTAURN	-2.825085	1.111794	-2.541016	0.0112

R-squared	0.052737	Mean dependent var	8.686493
Adjusted R-squared	0.045632	S.D. dependent var	13.72152
S.E. of regression	13.40479	Akaike info criterion	8.037737
Sum squared resid	143750.7	Schwarz criterion	8.078448
Log likelihood	-3236.227	F-statistic	7.423062
Durbin-Watson stat	2.012825	Prob(F-statistic)	0

Proc/Make Residual Series. Then, name for residual series: ***resid01***

Step 2: Run the regression of $\log(\hat{u}_i^2)$ or log(*resid01*^2) on $x_1, \ldots, x_k$

Dependent Variable: LOG(RESID01^2)
Included observations: 807

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.92069	2.563033	-0.749382	0.4538
LOG(INCOME)	0.29154	0.077468	3.763351	0.0002
LOG(CIGPRIC)	0.195418	0.614539	0.317992	0.7506
EDUC	-0.0797	0.017784	-4.481657	0
AGE	0.204005	0.017044	11.96928	0
AGE^2	-0.00239	0.000186	-12.89313	0
RESTAURN	-0.62701	0.118344	-5.298213	0

Log likelihood	-1428.44	F-statistic	43.82129
Durbin-Watson stat	2.024587	Prob(F-statistic)	0

Proc/Make Residual Series. Then, name for residual series: *lresid02*
Step 3: Generating Series: *h01* = exp(log(*resid01*^2)-*lresid02*)

Step 4: FGLS equation with weights, 1/sqr(h01)

Eview Trick: WLS; In the "Equation" window, Choose Estimate and then options. Click Weighted LS/TSLS. Type in the weight, *1/sqr(h01).*

Dependent Variable: CIGS
Included observations: 807
Weighting series: 1/SQR(H01)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.635471	17.80314	0.316544	0.7517
LOG(INCOME)	1.295239	0.437012	2.963855	0.0031
LOG(CIGPRIC)	-2.94031	4.460145	-0.659242	0.5099
EDUC	-0.46345	0.120159	-3.856953	0.0001
AGE	0.481948	0.096808	4.978378	0
AGE^2	-0.00563	0.000939	-5.989706	0
RESTAURN	-3.46106	0.795505	-4.350776	0

Weighted Statistics			
R-squared	0.002751	Mean dependent var	7.158227
Log likelihood	-3126.19	F-statistic	17.05549
Durbin-Watson stat	2.049719	Prob(F-statistic)	0
Unweighted Statistics			
R-squared	0.045739	Mean dependent var	8.686493

Alternatively, Step 4: Regress cigs/sqr(*h01*) on 1/sqr(*h01*), log(income)/sqr(*h01*), … with no intercept

Dependent Variable: CIGS/SQR(H01)
Included observations: 807

Variable	Coefficient	Std. Error	t-Statistic	Prob.
1/SQR(H01)	5.635471	17.80314	0.316544	0.7517
LOG(INCOME)/SQR(H01)	1.295239	0.437012	2.963855	0.0031
LOG(CIGPRIC)/SQR(H01)	-2.94031	4.460145	-0.659242	0.5099
EDUC/SQR(H01)	-0.46345	0.120159	-3.856953	0.0001
AGE/SQR(H01)	0.481948	0.096808	4.978378	0
AGE^2/SQR(H01)	-0.00563	0.000939	-5.989706	0
RESTAURN/SQR(H01)	-3.46106	0.795505	-4.350776	0

R-squared	0.002751	Mean dependent var	0.966192
Adjusted R-squared	-0.00473	S.D. dependent var	1.574979
S.E. of regression	1.578698	Akaike info criterion	3.759715
Sum squared resid	1993.831	Schwarz criterion	3.800425
Log likelihood	-1510.05	Durbin-Watson stat	2.049719

OLS & FGLS Results Compared

Dep. Var = *cigs*	OLS	FGLS	OLS	FGLS
Variable	Coefficient	Coefficient	Prob.	Prob.
C	-3.6398	5.635471	0.8799	0.7517
LOG(INCOME)	0.88027	1.295239	0.2268	0.0031
LOG(CIGPRIC)	-0.7509	-2.940314	0.8966	0.5099
EDUC	-0.5015	-0.463446	0.0028	0.0001
AGE	0.77069	0.481948	0	0
AGE^2	-0.009	-0.005627	0	0
RESTAURN	-2.8251	-3.461064	0.0112	0

Interpretation:

1. Income effect is now statistically significant and larger in magnitude.
2. Price effect is still statistically insignificant.
3. Cigarette smoking is negatively related to schooling.
4. Age has a diminishing marginal effect on smoking. Smoking increases with age up until 42.8 years old and then smoking decreases with age.
5. Cigarette smoking is negatively affected by restaurant smoking restrictions.

Recap of Heteroskedasticiy

- Consequences of Heteroskedasticity
- Testing for Heteroskedasticity
- Heteroskedasticity-Robust Inference
- Weighted Least Square Estimation