Chapter 11 Heteroscedasticity

남준우·허인 (2018), 제 9장

Gujarati/Porter (2018), Chapter 10

1. Sources, Nature (and Estimation)
2. Problems of Least Squares Estimator
3. Detecting Heteroscedasticity
4. Estimation
5. White's Heteroscedasticity-Autocorrelation Consistent Standard Errors

   (White's Robust Standard Errors)
6. Logarithms and Heteroscedasticity
7. Example

- Homoscedasticity vs. Heteroscedasticity

$V(\varepsilon_i) = \sigma^2 = V(\varepsilon_j)$ for $i \neq j$ vs. $V(\varepsilon_i) = \sigma_i^2 \neq \sigma_j^2 = V(\varepsilon_j)$ for $i \neq j$

1. <u>Sources, Nature</u>

- Conditional density function

► In general, as $x\uparrow$, $y\uparrow$ with variability of $y\uparrow$.

► $V(Y|X)$ increases as X increases.

☞(graph)

► Frequently encountered in cross-sectional models.

(Example)

① Income-Saving in low and high income group. (Discretion)

② Unit problem in long time series.

③ Small and large firms.

④ Learning by doing

⑤ Grouped data

⑥ Outlier

(Example) Data: UNIDO

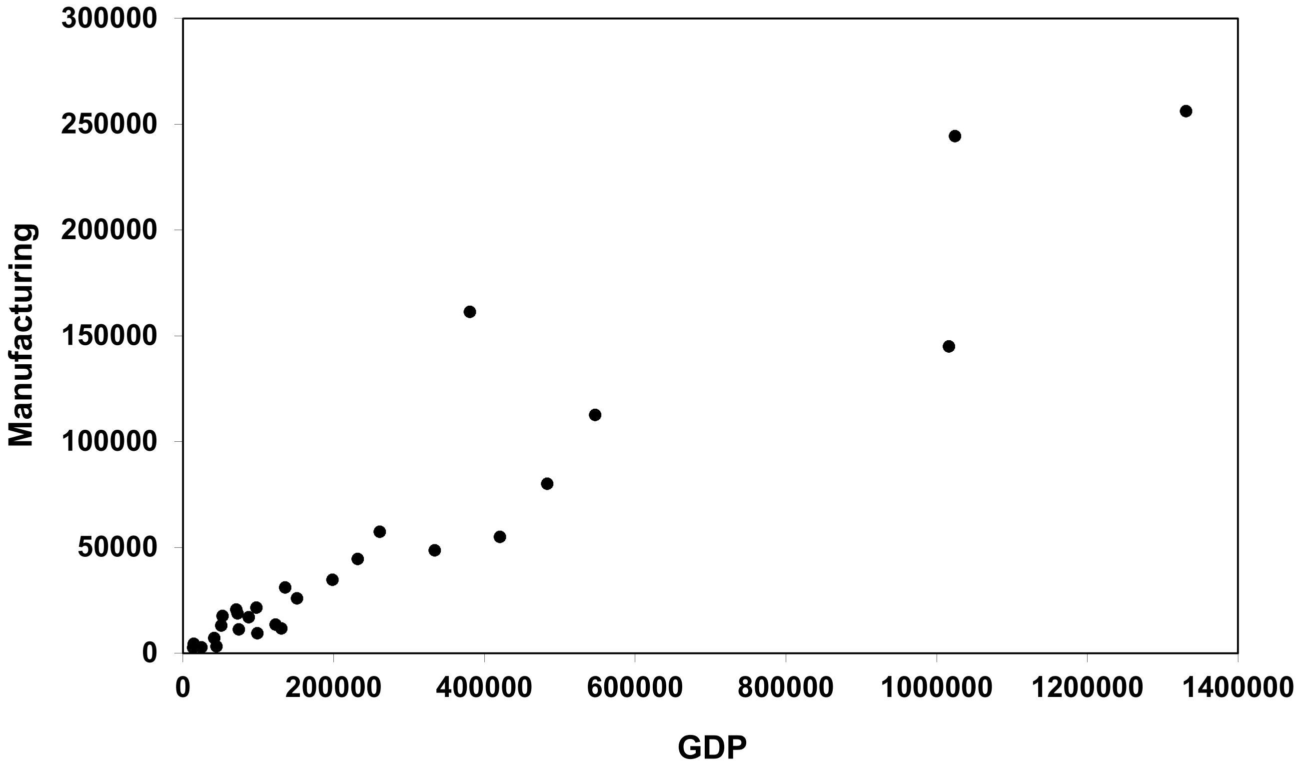
- Relationship between value added in manufacturing(MANU) and GDP in cross-country data.

- Manufacturing output and GDP(millions of $) for 28 countries in 1994.

$$MANU_i = \beta_1 + \beta_2 GDP_i + \varepsilon_i$$

- Likely to have heteroscedasticity.

► Rich countries will have larger errors than poor countries.



2. Consequences of Heteroscedasticity

(1) Assumptions of classical regression model is violated.

► Gauss-Markov theorem does not hold.

► LS estimator is not BLUE.

► There may be better(more efficient) estimator ⇒ Generalized Least Squares Estimator.

- OLS is inefficient: There are other estimators that have lower variances.

(Note) Ordinary LS estimator vs. Generalized LS estimator

(2) Problem of OLS estimator in inferences

• OLS estimator: $(b_2, \quad S_{b_2})$

► $E(b_2) = \beta_2$.

► $$V(b_2)\Big|_{\text{heteroscedasticity}} = \frac{\sum_{i=1}^{n} \sigma_i^{\,2} (X_i - \bar{X})^2}{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2 \right)^2} \quad \neq \quad \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = V(b_2)\Big|_{\text{homoscedasticity}}.$$

► OLS estimator $$s_{b_2}^{\;\;2} = \frac{s^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2},$$ $$E(s_{b_2}^{\;\;2}) \neq V(b_2)\Big|_{\text{heteroscedasticity}}.$$

• So, inferences based on OLS is invalid.

► White’s heteroscedasticity-(autocorrelaton)-corrected (HAC) estimator.

(<u>Summary</u>)

- OLS estimators still unbiased.
- However, standard errors are wrong $\Rightarrow$ use White's robust standard errors.
- Use $\left(b_2, \quad \text{White's robust standard error}\right)$ rather than $(b_2, \quad S_{b_2})$.

3. Detecting Heteroscedasticity

(0) Graphical Method: residual plot (e_i^2 on X_i or $\hat{y}_{\mathrm{i}}$).

(1) Goldfeld-Quandt Test

• Applicable under the assumption that the heteroscedastic variance σ_i^2 is monotonically related to one of the explanatory variables, $\sigma_i^2 = \sigma^2 X_i^2$.

① Order the sample by value of X.

$$X_i^{(1)} \le X_i^{(2)} \le \cdots\cdots\cdots \le X_i^{(n)}$$

② Drop the middle 10-15%(c) observations and take first n' and last n' observations(n'=(n-c)/2).

③ Run separate OLS on the two subsamples and get RSS1(from first n') and RSS2(from last n').

$$F = \frac{RSS_2}{RSS_1} \sim F(n'-k, n'-k) \text{ under } H_0.$$

If $F \ge F(n'-k, n'-k; \alpha)$, reject H_0 ⇒ heteroscedasticity.

If $F < F(n'-k, n'-k; \alpha)$, do not reject H_0 ⇒ homoscedasticity.

(2) <u>Breusch –Pagan Test</u>

► More general test

- Looks for any kind of association between σ_i^2 and independent variables, not just proportionality.

- Specified variables cause variation in the disturbances across observations.

- Model: $y_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$,

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{i2} + \cdots + \alpha_p Z_{ip} + v_i$$

$H_0 : \alpha_2 = \cdots = \alpha_p = 0$ (homoscedasticity). $H_1 :$ not H_0 (heteroscedasticity).

① LS regression of y on $1,\ X_2,\ \cdots,\ X_k$ and get residual e_i.

② Run the following auxiliary regression and get R^2: $e_i^2 = a_1 + a_2 Z_{i2} + \cdots + a_p Z_{ip} + \hat{v}_i$

③ Compute $n \cdot R^2$.

If $n \cdot R^2 \geq \chi^2(p-1;\alpha)$, reject H_0. $\Rightarrow$ heteroscedasticity.

If $n \cdot R^2 < \chi^2(p-1;\alpha)$, do not reject H_0. $\Rightarrow$ homoscedasticity.

(3) White test

- Model: $y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i$

$H_0 : \sigma_i^2 = \sigma^2$ for all i, H_1 : not H_0.

$$\sigma_i^2 = f(X_i, Z_i, X_i^2, Z_i^2, X_i Z_i) + v_i$$

$$\Rightarrow \quad \sigma_i^2 = a_1 + a_2 X_i + a_3 Z_i + a_4 X_i^2 + a_5 Z_i^2 + a_6 X_i Z_i + v_i$$

① LS regression of y on 1, X, Z and get residual e_i.

② Run the following auxiliary regression and get R^2:

$$e_i^2 = a_1 + a_2 X_i + a_3 Z_i + a_4 X_i^2 + a_5 Z_i^2 + a_6 X_i Z_i + \hat{v}_i$$

③ Compute $n \cdot R^2$.

If $n \cdot R^2 \geq \chi^2(5;\alpha)$, reject H_0. $\Rightarrow$ heteroscedasticity.

If $n \cdot R^2 < \chi^2(5;\alpha)$, do not reject H_0. $\Rightarrow$ homoscedasticity.

► Can extend to k-independent variables.

4. Estimation

(1) Generalized Least Squares Estimator: $V(\varepsilon_i) = \sigma_i^2$ is known

$$y_i = \beta_1 + \beta_2 X_i + \varepsilon_i, \quad V(\varepsilon_i) = \sigma_i^2$$

$$\Rightarrow \quad \frac{y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{X_i}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i} .$$

$$\Rightarrow \quad y_i^* = \beta_1 Z_i^* + \beta_2 X_i^* + \varepsilon_i^* \qquad \text{with } V(\varepsilon_i^*) = 1$$

- Since this new model is homoscedastic, OLS estimators will be efficient.

So, OLS of $\dfrac{y_i}{\sigma_i}$ on $\dfrac{1}{\sigma_i}, \dfrac{X_i}{\sigma_i}$ $\Rightarrow$ GLS estimation.

(Note)

- OLS of $\dfrac{y_i}{\sigma_i}$ on $\dfrac{1}{\sigma_i}, \dfrac{X_i}{\sigma_i}$ == GLS of y_i on $(1, X_i)$ with weight $\dfrac{1}{\sigma_i}$; Weighted LS (WLS)

$$
\begin{aligned}
&\min \sum_{i=1}^{n} \left(y_i * - \beta_1 Z_i * - \beta_2 X_i * \right)^2 \\
&= \min \sum_{i=1}^{n} \left(\frac{y_i}{\sigma_i} - \beta_1 \frac{1}{\sigma_i} - \beta_2 \frac{X_i}{\sigma_i} \right)^2 \quad . \\
&= \min \sum_{i=1}^{n} \left\{ \frac{1}{\sigma_i} \left(y_i - \beta_1 - \beta_2 X_i \right) \right\}^2
\end{aligned}
$$

- By weighting the observations by a factor $1/\sigma_i$, we are attaching greater importance to the observations with low σ_i.

$\Rightarrow$ Weighted Least Squares(WLS) estimator.

(Note) No constant term in the weighted regression.

(2) Feasible Generalized Least Squares Estimator(FGLS)

① Proportional Heteroscedasticity:

$y_i = \beta_1 + \beta_2 X_i + \varepsilon_i,$ Assume $V(\varepsilon_i) = \sigma_i^2 = \sigma^2 X_i$

- To make model with homoscedastic errors,

$$\frac{y_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + \frac{\varepsilon_i}{\sqrt{X_i}} \quad \text{with } V\left(\frac{\varepsilon_i}{\sqrt{X_i}}\right) = \sigma^2$$

- Since this new model is homoscedastic, OLS estimators will be efficient.

So, LS of $\frac{y_i}{\sqrt{X_i}}$ on $\frac{1}{\sqrt{X_i}}, \sqrt{X_i}$. $\Rightarrow$ FGLS estimation.

(Example) What if we assume $V(\varepsilon_i) = \sigma_i^2 = \sigma^2 X_i^2$?

② <u>White's Estimator</u>

• Model: $y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i$, $\sigma_i^{\ 2} = \alpha_1 + \alpha_2 X_i + \alpha_3 Z_i + \alpha_4 X_i^{\ 2} + \alpha_5 Z_i^{\ 2} + \alpha_6 X_i Z_i + v_i$

(a) LS regression of y on $1,\ X,\ Z$ and get residual e_i.

(b) Run the following auxiliary regression $e_i^2 = a_1 + a_2 X_i + a_3 Z_i + a_4 X_i^{\ 2} + a_5 Z_i^{\ 2} + a_6 X_i Z_i + \hat{v}_i$

and get $\hat{\sigma}_i^{\ 2} \equiv \widehat{e_i^2} = a_1 + a_2 X_i + \cdots + a_6 X_i Z_i$.

(c) To make model with homoscedastic errors, take $\hat{\sigma}_i \equiv \sqrt{\widehat{e_i^2}}$

$$\frac{y_i}{\hat{\sigma}_i} = \beta_1 \frac{1}{\hat{\sigma}_i} + \beta_2 \frac{X_i}{\hat{\sigma}_i} + \beta_3 \frac{Z_i}{\hat{\sigma}_i} + \varepsilon_i^{\ *}.$$

• Since this new model is homoscedastic, OLS estimators will be efficient.

So, OLS of model (c) $\Rightarrow$ FGLS estimation.

(Note)

① In getting $\hat{\sigma}_i$ in step (b), (c),

$\hat{\sigma}_i = \sqrt{\hat{\sigma}_i^{\,2}} = \sqrt{\widehat{e_i^2}}$,   NOT $\hat{\sigma}_i = \hat{e}_i$

② Sometimes, in step (b),

$\hat{\sigma}_i^{\,2} < 0$ is possible, so $\sigma_i = \sqrt{\hat{\sigma}_i^{\,2}}$ cannot be defined.

• What if we assume $\sigma_i^{\,2} = \sigma^2 \cdot \exp\left(a_1 X_i + a_2 Z_i + a_3 X_i^{\,2} + a_4 Z_i^{\,2} + a_5 X_i Z_i + v_i \right)$?

Then, $\widehat{\ln \sigma_i^{\,2}} = a_0 + a_1 X_i + a_2 Z_i + a_3 X_i^{\,2} + a_4 Z_i^{\,2} + a_5 X_i Z_i$,

get $\hat{\sigma}_i^{\,2} = \exp\left(\widehat{\ln \sigma_i^{\,2}} \right)$.

Then, go to step (c).

(3) White's Heteroscedasticity-Corrected Standard Errors (White's Robust Standard Errors)

- FGLS is efficient when we are confident about the form of heteroscedasticity.

  What if we do not know the form of heteroscedasticity?

- One method: Use OLS after correcting the variance term $\Rightarrow$ White's Robust Standard Errors

  Some econometricians say that OLS with White's robust standard error is superior to FGLS.

- For simple regression, since $V(b_2) = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2 V(\varepsilon_i)}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sigma_i^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}$,

  use $\hat{V}(b_2) = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2 e_i^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}$ as estimated variance or standard errors.

$\Rightarrow$ White's HAC (Heteroscedasticity and autocorrelation consistent (or corrected)) estimator.

5. Logarithms and Heteroscedasticity

- Sometimes, taking logarithm of dependent variable reduce the degree of heteroscedasticity.
- However, semilog specification would impose a particular shape; thus misspecification.

(Example 1)

File name: UNIDO

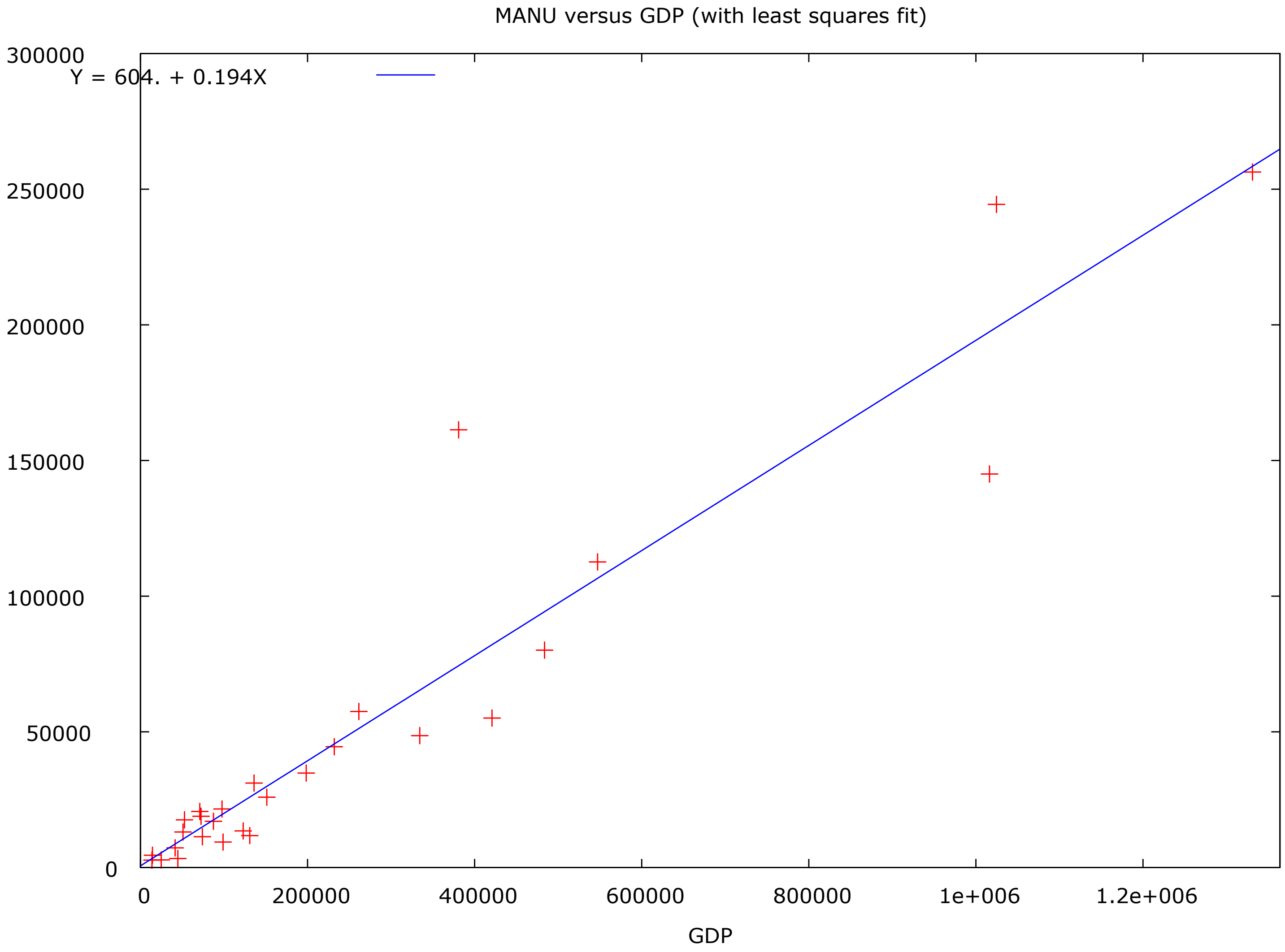
Source: UNIDO Yearbook 1997

Note:

MANU: value added in manufacturing (US$ million)

GDP: gross domestic product (US$ million)

POP: population(million)


MANU versus GDP (with least squares fit)
Y = 604. + 0.194X
300000
250000
200000
150000
100000
50000
0
0
200000
400000
600000
800000
1e+006
1.2e+006
GDP

(1) OLS

Model 1: OLS, using observations 1-28
Dependent variable: MANU

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	603.953	5699.67	0.1060	0.9164	
GDP	0.193693	0.0133428	14.52	<0.0001	***

Mean dependent var	52595.05	S.D. dependent var	69472.47
Sum squared resid	1.43e+10	S.E. of regression	23461.93
R-squared	0.890172	Adjusted R-squared	0.885948
F(1, 26)	210.7346	P-value(F)	5.53e-14
Log-likelihood	−320.4605	Akaike criterion	644.9211
Schwarz criterion	647.5855	Hannan-Quinn	645.7356

(2) White’s test for heteroscedasticity

► (1)의 결과에서 $\boxed{\text{Tests}}$ ⇒ $\boxed{\text{Heteroscedasticity Tests}}$ ⇒ $\boxed{\text{White’s Test}}$ 를 선택

```
White's test for heteroskedasticity
OLS, using observations 1-28
Dependent variable: uhat^2

             coefficient      std. error     t-ratio   p-value
  ------------------------------------------------------------
  const      -4.21382e+08     4.51255e+08   -0.9338   0.3593
  GDP       6271.89         2758.25          2.274    0.0318  **
  sq_GDP     -0.00411546      0.00226259    -1.819    0.0809  *

  Unadjusted R-squared = 0.211391

Test statistic: TR^2 = 5.918939,
with p-value = P(Chi-square(2) > 5.918939) = 0.051846
```

(3) White's robust standard error: HAC estimator

Model / OLS Estimation 창 아래 왼쪽 Robust standard errors 항을 체크한다.

Model 3: OLS, using observations 1-28
Dependent variable: MANU
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	603.953	3542.39	0.1705	0.8659	
GDP	0.193693	0.0179541	10.79	<0.0001	***

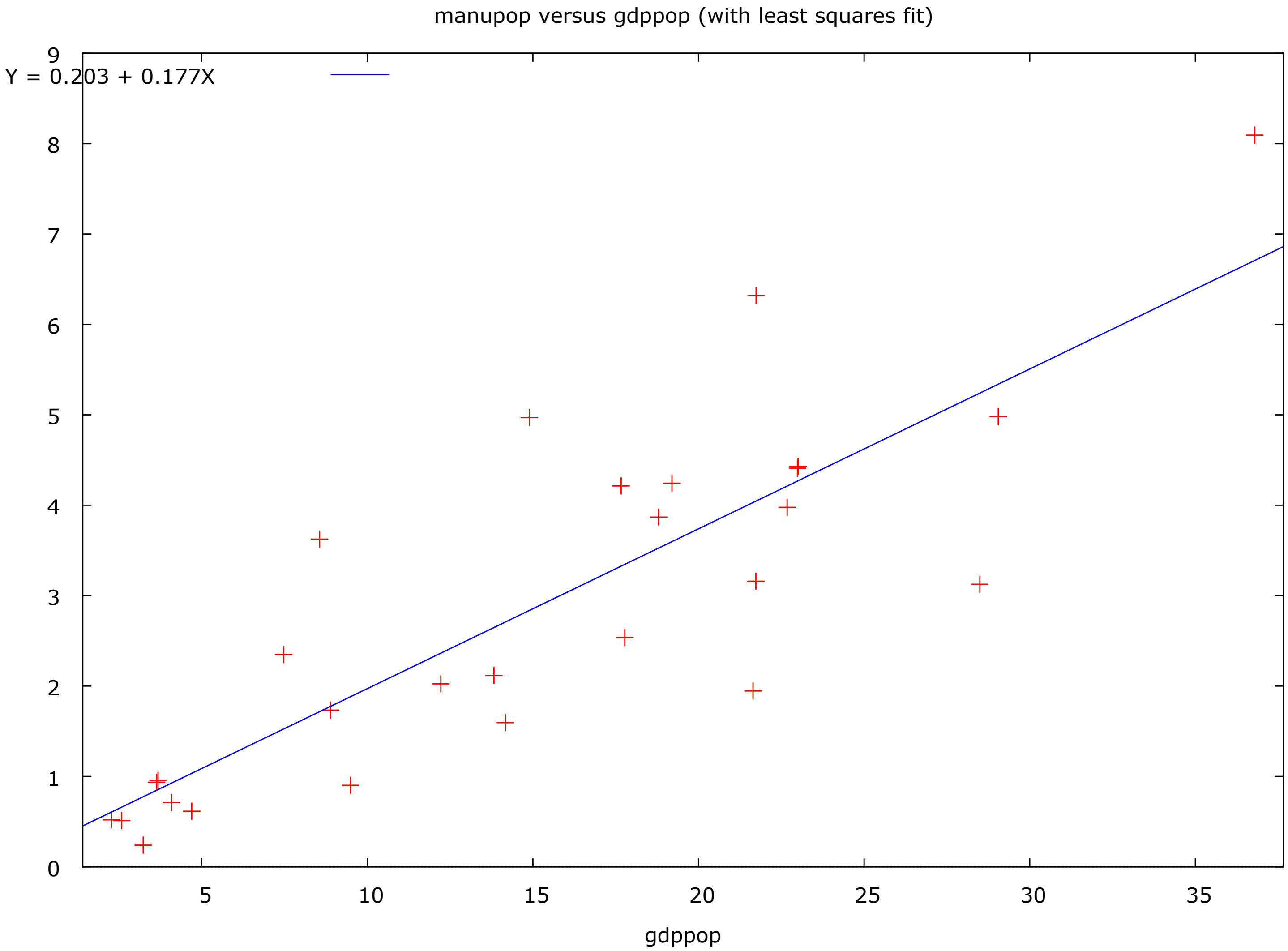
Mean dependent var	52595.05	S.D. dependent var	69472.47
Sum squared resid	1.43e+10	S.E. of regression	23461.93
R-squared	0.890172	Adjusted R-squared	0.885948
F(1, 26)	116.3853	P-value(F)	4.27e-11
Log-likelihood	−320.4605	Akaike criterion	644.9211
Schwarz criterion	647.5855	Hannan-Quinn	645.7356

(4)변수의 변환

Manupop=manu/pop

Gdppop=gdp/pop

Invconst=1/pop


manupop versus gdppop (with least squares fit)
Y = 0.203 + 0.177X
0
1
2
3
4
5
6
7
8
9
5
10
15
20
25
30
35
gdppop

WLS: Per capita data

► Manupop를 Invconst, Gdppop에 회귀분석

Model 1: OLS, using observations 1-28
Dependent variable: manupop

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
invconst	612.275	1370.51	0.4468	0.6588	
gdppop	0.182253	0.0155108	11.75	<0.0001	***

Mean dependent var	2.825215	S.D. dependent var	1.961627
Sum squared resid	31.00287	S.E. of regression	1.091979
Uncentered R-squared	0.905302	Centered R-squared	0.701596
F(2, 26)	124.2786	P-value(F)	4.93e-14
Log-likelihood	−41.15653	Akaike criterion	86.31306
Schwarz criterion	88.97747	Hannan-Quinn	87.12760

(Example 2) data: USEDCAR

변 수 명	변 수 설 명	비 고
ABSEQ	ABS장치의 유무 여부	ABS 장치가 없으면=0
		ABS 장치가 있으면=1
AIRBAG	에어백 개수	에어백이 없는 경우=0
		운전석에만 있는 경우=1
		운전석과 조수석에 모두 있는 경우=2
AUTO	자동변속기 유무 여부	자동변속기=1, 기타=0
CC	배기량	단위: cc
CVT	무단변속기 유무 여부	무단변속기=1, 기타=0
DIESEL	디젤 엔진 여부	디젤엔진의 경우=1, 기타=0
GAS	휘발유 사용 여부	휘발유 엔진의 경우=1, 기타=0
LPG	LPG연료 사용 여부	LPG엔진의 경우=1, 기타=0
MILEAGE	주행거리	단위: km
PERIOD	출고후 경과기간	단위: 개월
PRICE	중고차 (매매 완료) 가격	단위: 만원
DWOO	제조회사	대우차=1, 기타=0
HYUN	제조회사	현대차=1, 기타=0
KIA	제조회사	기아차=1, 기타=0
SSANG	제조회사	쌍용차=1, 기타=0

2002년 말 현재 여러 인터넷 중고차 매매 사이트의 정보를 통해 실제 판매된 145개 중고 차량을 대상으로 조사한 자료.

Model: $\log(price_i) = \beta_1 + \beta_2 CC_i + \beta_3 Period_i + \beta_4 Airbag_i + \varepsilon_i$

(1) OLS estimation

Model 1: OLS, using observations 1-145
Dependent variable: l_PRICE

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	6.09645	0.0763532	79.85	<0.0001	***
AIRBAG	0.186334	0.0309484	6.021	<0.0001	***
CC	0.000518743	3.38428e-05	15.33	<0.0001	***
PERIOD	−%s	0.000838089	−%#.4g	<0.0001	***

Mean dependent var	6.856559	S.D. dependent var	0.576598
Sum squared resid	9.108667	S.E. of regression	0.254166
R-squared	0.809741	Adjusted R-squared	0.805693
F(3, 141)	200.0312	P-value(F)	1.34e-50
Log-likelihood	−5.101803	Akaike criterion	18.20361
Schwarz criterion	30.11054	Hannan-Quinn	23.04180

(2) White's test

```
White's test for heteroskedasticity
OLS, using observations 1-145
Dependent variable: uhat^2

              coefficient    std. error    t-ratio   p-value
  ----------------------------------------------------------
  const       -0.0736790     0.0780381     -0.9441   0.3468
  AIRBAG      -0.0781653     0.0780761     -1.001    0.3185
  CC           0.000126545   7.60720e-05    1.663    0.0985   *
  PERIOD       0.000955978   0.00164686     0.5805   0.5626
  sq_AIRBAG    0.0110297     0.0175151      0.6297   0.5299
  X2_X3        1.45754e-05   2.84971e-05    0.5115   0.6099
  X2_X4        0.000339351   0.000641647    0.5289   0.5978
  sq_CC       -2.25646e-08   1.98339e-08   -1.138    0.2573
  X3_X4       -1.68336e-06   6.64106e-07   -2.535    0.0124   **
  sq_PERIOD    4.36323e-05   9.00784e-06    4.844    3.44e-06 ***

  Unadjusted R-squared = 0.423644

Test statistic: TR^2 = 61.428373,
with p-value = P(Chi-square(9) > 61.428373) = 0.000000
```

(3) White Robust Standard Error: HAC estimator

Model 2: OLS, using observations 1-145
Dependent variable: l_PRICE
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	6.09645	0.0813171	74.97	<0.0001	***
AIRBAG	0.186334	0.0264546	7.044	<0.0001	***
CC	0.000518743	3.07805e-05	16.85	<0.0001	***
PERIOD	−%s	0.00149739	−%#.4g	<0.0001	***

Mean dependent var	6.856559	S.D. dependent var	0.576598
Sum squared resid	9.108667	S.E. of regression	0.254166
R-squared	0.809741	Adjusted R-squared	0.805693
F(3, 141)	168.0053	P-value(F)	2.35e-46
Log-likelihood	−5.101803	Akaike criterion	18.20361
Schwarz criterion	30.11054	Hannan-Quinn	23.04180

(4) Weighted Least Squares

① OLS estimation에서 구한 e_i^2을 White's test에 사용된 변수에 회귀분석하여 $\widehat{e_i^2}$을 구하여 save한다.

② Model / other linear model / WLS에서 $\widehat{e_i^2}$을 weight로 지정함.

Model 8: WLS, using observations 1-145
Dependent variable: l_PRICE
Variable used as weight: weight

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	6.01642	0.0633770	94.93	<0.0001	***
AIRBAG	0.210118	0.0257631	8.156	<0.0001	***
CC	0.000508055	2.82205e-05	18.00	<0.0001	***
PERIOD	−%s	0.00111600	−%#.4g	<0.0001	***

Statistics based on the weighted data:

Sum squared resid	177.3833	S.E. of regression	1.121623
R-squared	0.789884	Adjusted R-squared	0.785413
F(3, 141)	176.6860	P-value(F)	1.45e-47
Log-likelihood	−220.3606	Akaike criterion	448.7212
Schwarz criterion	460.6281	Hannan-Quinn	453.5593

Statistics based on the original data:

Mean dependent var	6.856559	S.D. dependent var	0.576598
Sum squared resid	9.571608	S.E. of regression	0.260545