Relaxing some Assumptions

Part 6

EE325
Introductory Econometrics
Revision Aug 2020

List of the topics to cover

In this chapter, some assumptions imposed on either single or multiple linear regression will be relaxed. For the revision of all assumptions imposed on multiple linear regression, we can go back on slide page 67. Here are the assumption we are going to discuss.

- No multicollinearity: the assumption that each independent variables are not seriously correlated.
- Homoscedasticity: the error term across the range of independent variable are equal or $Var(u_i) = \sigma^2$.
- No autocorrelation: the error term are independently distributed or $cov(u_i, u_j) = 0$.

In each section, we will explore the follow topics

- Concept of relaxing the assumption.
- Effects on the estimated coefficients and variance, also standard error.
- How to detect the problem.
- What are remedial methods.

(1) Nature of multicollinearity

G. 321

This is a simplified version, compared to the book, of multicollinearity problem. Let λ_i be a constant, consider the following argument when X_{2i} and X_{3i} are linearly correlated perfectly, or **perfect multicollinearity**.

- $\lambda_2 X_{2i} + \lambda_3 X_{3i} = 0$

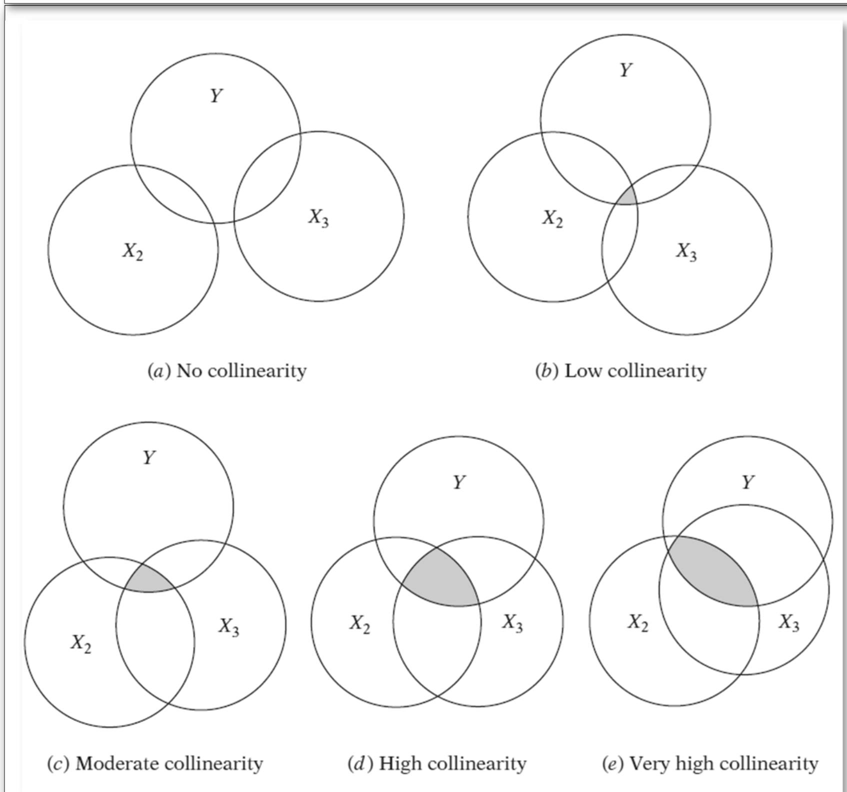
when $\lambda_i \neq 0$ for all i simultaneously. Another relation that describe a non-perfect collinearity, or **multicollinearity**, is

- $\lambda_2 X_{2i} + \lambda_3 X_{3i} + \nu_i = 0$

where υ_i is a stochastic error term. Now assumed that $\lambda_2 \neq 0$ then,

- $X_{2i} = -\frac{\lambda_3}{\lambda_2} X_{3i}$ for the first equation.
- $X_{2i} = -\frac{\lambda_3}{\lambda_2} X_{3i} - \nu_i$ for the second equation.

Levels of multicollinearity



(1) Nature of multicollinearity

G. 326

Precisely taken from the book, here are some causes of multicollinearity.

- *Data collection method* may limit range of values taken in the independent variables.
- *Constraints on the model or in the population being sampled.* E.g. income and house size tend to be correlated.
- *Model specification.* E.g. including a polynomial term especially when the range of X is small.
- *Overdetermined model* or when k is larger than n.
- *Common trend.* E.g. a time-series data consist of consumption expenditure, income, wealth, and number of population.

Looking from another perspective apart from stated above, multicollinearity is seen particularly as sampling problem, not a problem on a population, since when we postulate population regression, X variables included in a model have a separate or independent influence on Y.

Meaning that, as Goldberger coined the term, this may be considered as **micronumerosity** problem when our sampling may not be “rich” enough to capture X variability.

By the way, micronumerosity refers to the problem of small sample size.

Another important note is that multicollinearity is **much more common** in cross-sectional data compared to time-series data, in which autocorrelation is much more common.

(2) Effects on estimation

G. 324

(1) Perfect multicollinearity

Recall that the estimated coefficients are

- $\hat{\beta}_2 = \dfrac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i}x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}$

- $\hat{\beta}_3 = \dfrac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i}x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}$

If we assumed that $X_{3i} = \lambda X_{2i}$, replacing this into $\hat{\beta}_2$, we get,

- $\hat{\beta}_2 = \dfrac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2) - (\lambda \sum y_i x_{2i})(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - \lambda^2 (\sum x_{2i}^2)^2} = \dfrac{0}{0}$

Example: Consider the shoe size model with height (hei_i) as a regressor, now let's create a perfectly correlated variable of height * 2 (defined as $hei2_i$). The model will be

- $shoesize_i = \beta_1 + \beta_2 hei_i + \beta_3 hei2_i + u_i$

Throwing this model into STATA, we have the regression result as follows. We can see that STATA automatically rejects, or omits, one of these variables immediately because $\hat{\beta}_2$ cannot be estimated.

Regression results

```
      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(1, 22)        =     84.41
       Model |  115.781294         1  115.781294   Prob > F        =    0.0000
    Residual |  30.1770389        22  1.37168359   R-squared       =    0.7932
-------------+----------------------------------   Adj R-squared   =    0.7839
       Total |  145.958333        23  6.34601449   Root MSE        =    1.1712

------------------------------------------------------------------------------
          ss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hei |          0  (omitted)
        hei2 |   .1219928   .0132783     9.19   0.000     .0944553    .1495302
       _cons |  -.7057778   4.387244    -0.16   0.874    -9.804366     8.39281
------------------------------------------------------------------------------
```

Thus, if we use the program, perfect multicollinearity will not be much of a problem since this will be automatically detected and a variable will be dropped.

Remember that when we deal with a dummy variable, number of dummy must be $n-1$, here is the results of a model of show size with 2 gender dummy variables

Regression results

```
      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(1, 22)        =     42.71
       Model |  96.3333333         1  96.3333333   Prob > F        =    0.0000
    Residual |      49.625        22  2.25568182   R-squared       =    0.6600
-------------+----------------------------------   Adj R-squared   =    0.6446
       Total |  145.958333        23  6.34601449   Root MSE        =    1.5019

------------------------------------------------------------------------------
          ss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.sex |      -4.25   .6503386    -6.54   0.000     -5.59872    -2.90128
      1.sex2 |          0  (omitted)
       _cons |     42.375   .5309993    79.80   0.000     41.27377    43.47623
------------------------------------------------------------------------------
```

(2) Effects on estimation

(2) Multicollinearity

Given that $X_{3i} = \lambda X_{2i} + v_i$, replacing this into $\hat{\beta}_2$, we get,

- $\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum y_i x_{2i} + \sum y_i v_i)(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - \lambda^2 (\sum x_{2i}^2)^2}$

If v_i is approaching zero, the more this will be closer to perfect multicollinearity.

We are going to skip lots of proof to get to the conclusion what are affected as follows.

1) Coefficients estimated are still BLUE.

2) Large variances and covariances.

Recall that the variance of estimated coefficients can be written into this form.

- $Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$
- $Var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$

where r_{23}^2 is the coefficient of correlation between X_2 and X_3. The value is between 0 and 1, as an absolute value.

We can see clearly that when the correlation between X_2 and X_3 get higher, the denominator will be ………….. leading to ……………. variance.

If we separate a part of $Var(\hat{\beta}_2)$ like this,

- $Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2} \cdot \frac{1}{(1 - r_{23}^2)}$

we can define the latter part as **variance-inflating factor** or VIF

- $\text{VIF} = \frac{1}{(1 - r_{23}^2)}$

The higher r_{23}^2, the higher it is for VIF. We can also define the inverse of VIF as tolerance or TOL as

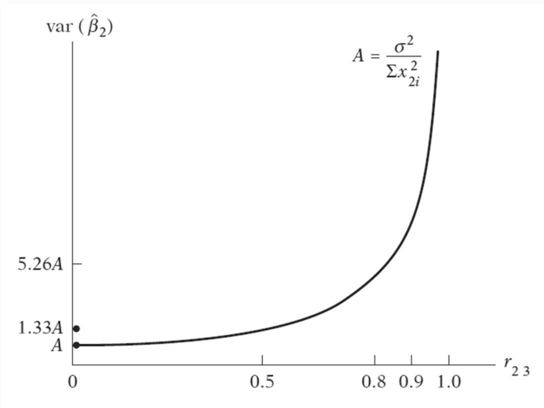
- $\text{TOL} = \frac{1}{\text{VIF}} = (1 - r_{23}^2)$

Note that these definition can be generalized for any pair of regressors.

(2) Effects on estimation

G. 330

$Var(\hat{\beta}_2)$ *and* r_{23}^2



3) Wider confidence interval

Since the variance is used to construct confidence interval, CI is stretched outward and the t-curve is flatter, which will later affect

4) Insignificant t ratios

The acceptance region also becomes larger when variance is high. It is more likely that we would accept the null hypothesis when we should reject, causing more-likely type-II error.

Example: Consider two shoe size models here, the first one take only height (hei_i) as a regressor while the second one include $heinorm_i$ which is height multiplied by a randomly generated number. The correlation between hei_i and $heinorm_i$ is 0.9956 to exaggerate the results as follows.

Regression results

```
      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(1, 22)        =     84.41
       Model |  115.781294         1  115.781294   Prob > F        =    0.0000
    Residual |  30.1770389        22  1.37168359   R-squared       =    0.7932
-------------+----------------------------------   Adj R-squared   =    0.7839
       Total |  145.958333        23  6.34601449   Root MSE        =    1.1712

------------------------------------------------------------------------------
          ss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hei |   .2439855   .0265566     9.19   0.000     .1889106    .2990605
       _cons |  -.7057778   4.387244    -0.16   0.874    -9.804366     8.39281
------------------------------------------------------------------------------
```

```
      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(2, 21)        =     44.31
       Model |  117.995532         2   58.997766   Prob > F        =    0.0000
    Residual |  27.9628013        21  1.33156197   R-squared       =    0.8084
-------------+----------------------------------   Adj R-squared   =    0.7902
       Total |  145.958333        23  6.34601449   Root MSE        =    1.1539

------------------------------------------------------------------------------
          ss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hei |   .6024787   .2792316     2.16   0.043     .0217848    1.183173
     heinorm |  -.1751522   .1358264    -1.29   0.211    -.4576187    .1073143
       _cons |  -2.065792   4.449407    -0.46   0.647    -11.31884    7.187255
------------------------------------------------------------------------------
```

(2) Effects on estimation

G. 331

5) High R^2 but few significant t ratios

The coefficient of determination or R^2 from a model with multicollinearity is likely to be high, also the F-stat of overall model test, since the estimation is 'tricked' to have similar regressors with more explanatory power.

We can also see from the previous example that when we intentionally add a colinear variable into the regression, R^2 is higher.

However, this is not due to more explanatory power of regressors, but multicollinearity. Therefore, each coefficient is not very likely to be significant.

6) Sensitivity of coefficient due to small changes in data.

You can read for this example in page 331. In conclusion, a very slight change in data will affect the value of coefficient tremendously. Some of the direction of coefficient can be different from theoretical speculation.

It would be beneficial to read an illustrative example from page 332 to 337.

(3) Detecting multicollinearity

G. 337

There are several ways to detect multicollinearity. Some of them are mentioned earlier. Some of them from the book are skipped.

Firstly, the phrase "**Rule of Thumb**" should be introduced. It refers to a specific level of criterion that is usually and mutually accepted as a threshold. More illustrative examples later here.

1) Conflicting test

This is already mentioned that when our estimation reports high R^2 or F value, but rarely coefficients are significant, we should suspect that there might be multicollinearity problem.

2) Pair-wise correlation among regressors

Another easy method to detect multicollinearity is to perform a pair-wise correlation on all regressors. (Easy when using STATA) The rule of thumb suggests that coefficient of correlation **exceeding 0.8** can be problematic and researcher may seek a remedial approach.

3) Auxiliary regressions

Regressing X_i on other X variables and obtain R_i^2 from the estimation then calculate

- $F_i = \frac{R^2_{xi \cdot x_2 x_3 \dots x_k}/(k-2)}{(1-R^2_{xi \cdot x_2 x_3 \dots x_k})/(n-k+1)}$

where k is the number of independent variables including intercept.

If F_i exceed critical value (d.f. of $k-2$ and $n-k+1$ obviously) from chosen level of significant, it means that X_i is collinear with other X.

Instead of testing all the auxiliary R_i^2, **Klein's rule of thumb** suggests that multicollinearity is troublesome if the R_i^2 is greater than R^2 from another model that we regress Y on these X_i and other X.

4) VIF and TOL

Again, we follow the rule of thumb that VIF should not exceed 10, which will happen when $r_{23}^2 = 0.9$, while TOL should be closer to 1 rather than 0.

(3) Detecting multicollinearity

G. 341

5) Scatter plot

Scatter plot is a good practice revealing linear relationship between two variables, see example below here.

Example: Setting up a ridiculous shoe size model once again as follows.

- $ss_i = \beta_1 + \beta_2 hei_i + \beta_3 heinorm_i + \beta_4 wei_i + \beta_5 ri_i + u_i$

where ss_i is shoe size,

hei_i is height in centimeters,

$heinorm_i$ is height multiplied by randomly generated number,

wei_i is weight in kilograms,

ri_i is received income per month.

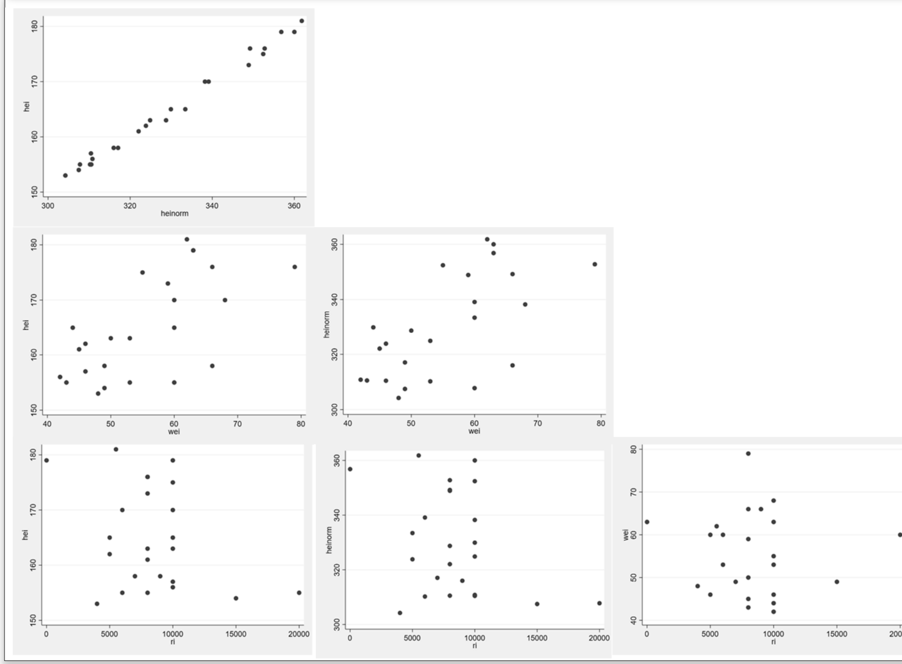
First, we should check coefficients of correlation of all these regressors.

Coefficients of correlation

	hei	heinorm	wei	ri
hei	1.0000			
heinorm	0.9956	1.0000		
wei	0.6521	0.6401	1.0000	
ri	-0.3255	-0.3342	-0.0523	1.0000

Then, we may check scatter plots, order corresponding to the coefficients of correlation.

Scatter plots



From both coefficients of correlation and scatter plot suggest that hei_i and $heinorm_i$ are seriously correlated.

Let's now figure out models of auxiliary regression.

(3) Detecting multicollinearity

Assumed that we acknowledge that all the coefficients and error terms in these models are all different, so that we do not have to write down all different notations.

- $hei_i = \beta_1 + \beta_2 heinorm_i + \beta_3 wei_i + \beta_4 ri_i + u_i$
- $heinorm_i = \beta_1 + \beta_2 hei_i + \beta_3 wei_i + \beta_4 ri_i + u_i$
- $wei_i = \beta_1 + \beta_2 hei_i + \beta_3 heinorm_i + \beta_4 ri_i + u_i$
- $ri_i = \beta_1 + \beta_2 hei_i + \beta_3 heinorm_i + \beta_4 wei_i + u_i$

Then we run all the regression, including the original model. If we consider **Klein's rule of thumb**, then we can see that the of R_i^2 of auxiliary model 1 and 2 exceeds the original model's.

Though it is obvious, let's perform F-test from the auxiliary model 1 and 4 to compare the result.

Aux model 1

- $F_i = \dfrac{R^2_{xi \cdot x_2 x_3 \ldots x_k}/(k-2)}{(1-R^2_{xi \cdot x_2 x_3 \ldots x_k})/(n-k+1)} =$
- $F_{cri} =$

Original model (shoe size)

```
      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(4, 19)        =     26.17
       Model |    123.5325         4  30.8831249   Prob > F        =    0.0000
    Residual |  22.4258337        19  1.18030704   R-squared       =    0.8464
-------------+----------------------------------   Adj R-squared   =    0.8140
       Total |  145.958333        23  6.34601449   Root MSE        =    1.0864

------------------------------------------------------------------------------
          ss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hei |   .5106412     .268842     1.90   0.073    -.0520515    1.073334
     heinorm |   -.156939    .1294631    -1.21   0.240    -.4279085    .1140305
         wei |   .0662845      .03225     2.06   0.054    -.0012155    .1337845
          ri |  -.0000716    .0000649    -1.10   0.284    -.0002075    .0000643
       _cons |   4.003158    5.136206     0.78   0.445    -6.747046    14.75336
------------------------------------------------------------------------------
```

Auxiliary models (1 and 2)

```
      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(3, 20)        =    787.33
       Model |  1928.62778         3  642.875926   Prob > F        =    0.0000
    Residual |  16.3305541        20  .816527704   R-squared       =    0.9916
-------------+----------------------------------   Adj R-squared   =    0.9903
       Total |  1944.95833        23  84.5634058   Root MSE        =    .90362

------------------------------------------------------------------------------
         hei |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     heinorm |   .4774118    .0141007    33.86   0.000     .4479982    .5068254
         wei |   .0231532    .0263193     0.88   0.389    -.0317479    .0780542
          ri |   9.17e-06     .000054     0.17   0.867    -.0001034    .0001217
       _cons |   6.118581    4.046982     1.51   0.146    -2.323276    14.56044
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(3, 20)        =    771.51
       Model |  8149.56423         3  2716.52141   Prob > F        =    0.0000
    Residual |  70.4210907        20  3.52105453   R-squared       =    0.9914
-------------+----------------------------------   Adj R-squared   =    0.9901
       Total |  8219.98532        23  357.390666   Root MSE        =    1.8764

------------------------------------------------------------------------------
     heinorm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hei |   2.058709    .0608056    33.86   0.000     1.931871    2.185547
         wei |  -.0264928    .0553858    -0.48   0.638    -.1420256    .0890399
          ri |  -.0000446    .0001117    -0.40   0.694    -.0002776    .0001884
       _cons |  -7.897395     8.69364    -0.91   0.374    -26.03201    10.23722
------------------------------------------------------------------------------
```

(3) Detecting multicollinearity

Aux model 4

- $F_i = \dfrac{R^2_{xi \cdot x_2 x_3 \dots x_k}/(k-2)}{(1-R^2_{xi \cdot x_2 x_3 \dots x_k})/(n-k+1)} =$

- $F_{cri} =$

Lastly, we can check both VIF and TOL, postestimation, the results are in the table on the right-hand side.

PS. We reject null hypothesis for auxiliary model 1, 2 and 3 but cannot reject for model 4. Note that F_i is tremendously different in model 1 and 2 compared to model 3 and 4, suggesting that wei_i and hei_i are correlated in some level, but not as serious as hei_i and $heinorm_i$.

Auxiliary models (3 and 4)

Source	SS	df	MS
Model	966.782514	3	322.260838
Residual	1134.84249	20	56.7421243
Total	2101.625	23	91.375

Number of obs	=	24
F(3, 20)	=	5.68
Prob > F	=	0.0056
R-squared	=	0.4600
Adj R-squared	=	0.3790
Root MSE	=	7.5327

wei	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hei	1.608961	1.828978	0.88	0.389	-2.206221	5.424142
heinorm	-.4269347	.8925475	-0.48	0.638	-2.288756	1.434887
ri	.0004259	.00044	0.97	0.345	-.0004919	.0013436
_cons	-72.76373	31.67798	-2.30	0.033	-138.8428	-6.684613

Source	SS	df	MS
Model	52222267.8	3	17407422.6
Residual	280017316	20	14000865.8
Total	332239583	23	14445199.3

Number of obs	=	24
F(3, 20)	=	1.24
Prob > F	=	0.3204
R-squared	=	0.1572
Adj R-squared	=	0.0308
Root MSE	=	3741.8

ri	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hei	157.3134	925.2591	0.17	0.867	-1772.743	2087.37
heinorm	-177.3861	444.1206	-0.40	0.694	-1103.805	749.0332
wei	105.0788	108.5597	0.97	0.345	-121.3727	331.5303
_cons	35098.58	15853.48	2.21	0.039	2028.8	68168.35

VIF and TOL

Variable	VIF	1/VIF
hei	119.10	0.008396
heinorm	116.73	0.008567
wei	1.85	0.539983
ri	1.19	0.842817
Mean VIF	59.72	

(4) Remedial measures

G. 342

1) Do nothing

If and only if when we do not any other choice than using deficient data set. Also, we may not be able to draw any meaningful insight from the regression.

2) Priori information

Supposed we have an income- consumption model as such,

- $Y_i = \beta_1 + \beta_2 income_i + \beta_3 wealth_i + u_i$

we know that income and wealth are highly colinear. If we know that from previous empirical work

- $\beta_3 = 0.1\beta_2$ then
- $Y_i = \beta_1 + \beta_2 income_i + 0.1\beta_2 wealth_i + u_i$ and
- $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$

where $X_{2i} = income_i + 0.1 wealth_i$, we can eliminate one of the variables.

3) Combining cross-sectional and time-series data

The most completed data would be panel data, repeated samples over time. If that is not possible to obtain, we may use '**pooled-data**', combining multiple waves of data into one large data set.

4) Dropping a variable(s) and specification bias

This is the easiest method. If we are sure which variable should be dropped, according to wrong specification of a model, dropping one of them is very easy and efficient. Consider dropping $heinorm_i$ from the show size model, our results would be as follows.

Original model (shoe size)

Source	SS	df	MS			
				Number of obs	=	24
				F(3, 20)	=	33.61
Model	121.798039	3	40.5993463	Prob > F	=	0.0000
Residual	24.1602944	20	1.20801472	R-squared	=	0.8345
				Adj R-squared	=	0.8096
Total	145.958333	23	6.34601449	Root MSE	=	1.0991

ss	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hei	.1875495	.0356159	5.27	0.000	.1132561	.2618429
wei	.0704423	.0324413	2.17	0.042	.0027709	.1381136
ri	-.0000646	.0000654	-0.99	0.335	-.000201	.0000719
_cons	5.242567	5.092152	1.03	0.316	-5.379477	15.86461

(4) Remedial measures

G. 345

5) Adding more observations

When possible, more observations lead to more variability in X and therefore, may lead to reduction of severity of multicollinearity problem.

6) Variable transformation

Transforming variables can be complicated, we can either perform

- **Ratio transformation** which will lead us to another problem of heteroscedasticity or using
- **First difference form** of variable which is popular in time-series data.

Therefore, transforming is not very much recommended.

(1) Nature of Heteroscedasticity

G. 365

Recall the assumption of homoscedasticity or

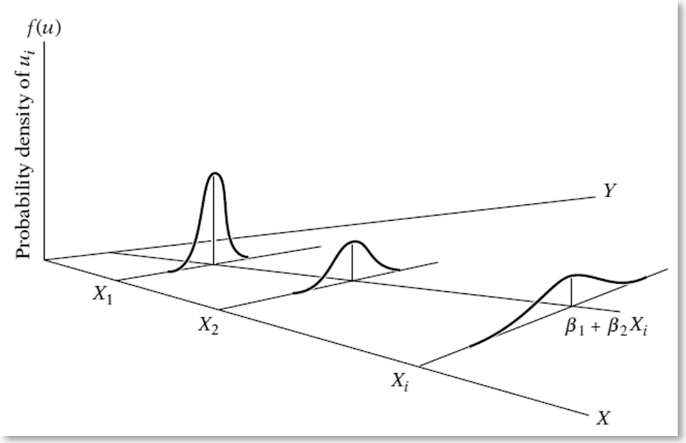
- $E(u_i^2|X_i) = \sigma^2$

when this assumption is relaxed, we have

- $E(u_i^2|X_i) = \sigma_i^2$

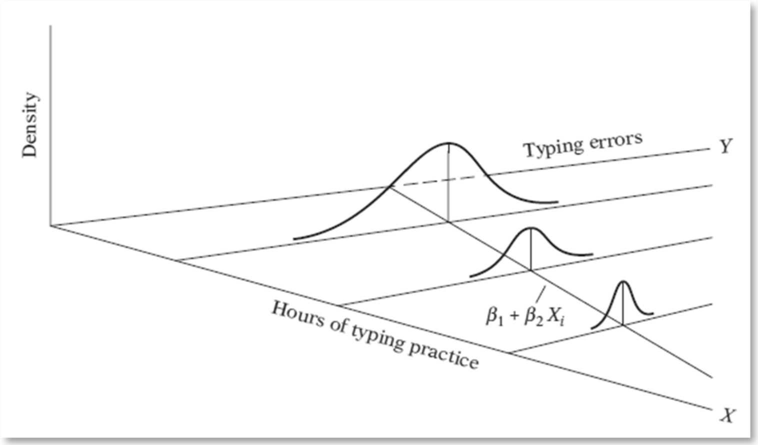
which means that we allow the error term scattered around each X_i to be different. Two classic examples are income-saving model and error-learning model as follows.

Income-saving model



As people get richer, they have more choice over their consumption-saving, leading to a larger variance of saving (Y_i) on larger income (X_i).

Error-learning model



Similarly, people practicing more hours on typing leads to lower typing errors. However, some people maybe pretty good at typing at first and some can be pretty bad due to their familiarity to keyboard layout or language. There are larger difference between people when start practicing but those difference will be minimized as they keep practicing.

Apart from the nature of data, there are some other causes for heteroscedasticity.

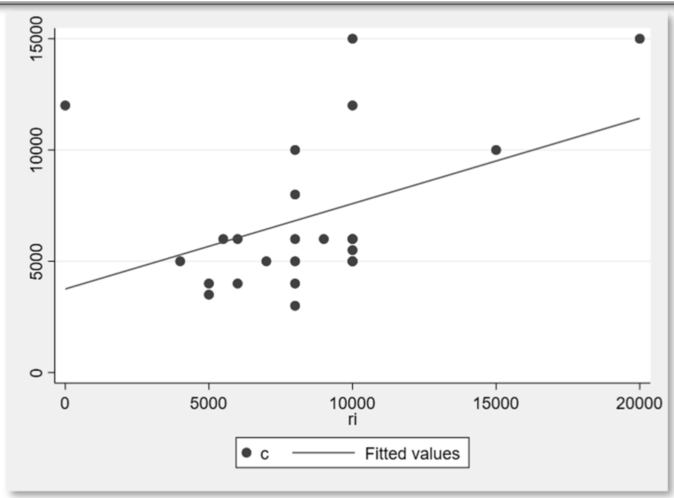
(1) Nature of Heteroscedasticity

G. 367

1) Presence of outliers

An outlier is an observation that is much different from the rest, either little or largely different. Inclusion and exclusion of an outlier can alter the result of a regression substantially.

A plot between income-consumption



2) Specification error

For example, a price demand model can be heteroscedastic if we do not include price of complementary or competing commodity. Other commodities' price may be the source of scattered quantity demanded at some level.

3) Skewness of distribution

For example, plotting wealth and education level can show this problem because there a fewer people with high wealth, leading to lower variance in education level, while larger groups of population with lower wealth.

4) Incorrect data transformation and functional form.

Details for this topic will be skipped on this point.

Heteroscedasticity is more likely to be found in cross-sectional data rather than time-series since they deal with members of a population at a given point of time.

(2) Effect on estimation

G. 371

We begin our examination on simple linear regression, recall that with homoscedasticity assumption yields the variance of the estimator as

- $Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}$

Relaxing the assumption, the variance becomes

- $Var(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{\left(\sum x_i^2\right)^2}$

$\hat{\beta}_2$ is not be BLUE, with heteroscedasticity. It is still linear and unbiased but it is not efficient anymore, comparing to deriving estimator from another method called **generalized least squares** (**GLS**). We setup a basic model as such.

(Note that an estimator not being efficient meaning that $Var(\hat{\beta}_i)$ is not the lowest.)

- $Y_i = \beta_1 + \beta_2 X_i + u_i$

Then we transform these variables by dividing by σ_i, if this standard error is known.

- $\frac{Y_i}{\sigma_i} = \frac{\beta_1}{\sigma_i} + \frac{\beta_2 X_i}{\sigma_i} + \frac{u_i}{\sigma_i}$

Then figure out the variance, we get

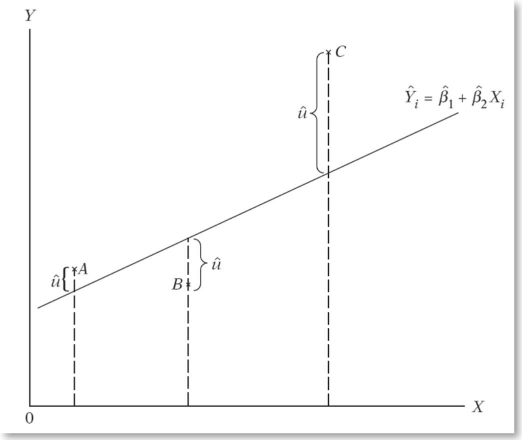
- $Var\left(\frac{u_i}{\sigma_i}\right) = E\left(\frac{u_i}{\sigma_i}\right)^2 = \frac{1}{\sigma_i^2} E(u_i^2)$ and if σ_i is known
- $Var\left(\frac{u_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2}(\sigma_i^2) = 1$

which is actually homoscedastic. To retrieve the estimators, we follow least squares method as usual

- $\min_{\hat{\beta}_1, \hat{\beta}_2} \sum \left(\frac{u_i}{\sigma_i}\right)^2 = \sum \left(\frac{Y_i}{\sigma_i} - \frac{\hat{\beta}_1}{\sigma_i} - \frac{\hat{\beta}_2 X_i}{\sigma_i}\right)^2$

This method is specifically called **weight least squares (WLS)**, weighing each term especially the error term with the error term itself. Estimators derived from this estimation are called **WLS estimators**. WLS is a class of GLS.

WLS



(2) Effect on estimation

G. 374

WLS estimators derived will have less variance compared to ordinary OLS. Therefore, estimators from OLS is not with the least variance anymore, losing the quality of being efficient.

Consequences of relying on OLS are as follows.

1) OLS estimation allowing heteroscedasticity

• Using OLS while assuming σ_i^2 are known, $Var(\hat{\beta}_2)$ is larger compared to the variance from WLS.

• Larger CI and t value is small, leading to insignificant conclusion.

2) OLS estimation disregarding heteroscedasticity

• Using OLS while assuming homoscedasticity (σ^2) when heteroscedasticity is present, $Var(\hat{\beta}_2)$ will be biased.

• We do not know whether the bias is positive (overestimate: actual variance is lower) or negative (underestimate: actual variance is higher), depending on the relationship between σ_i^2 and X_i.

• Conclusion or inference drawn from hypothesis tests may be misleading.

(3) Detecting heteroscedasticity

G. 377

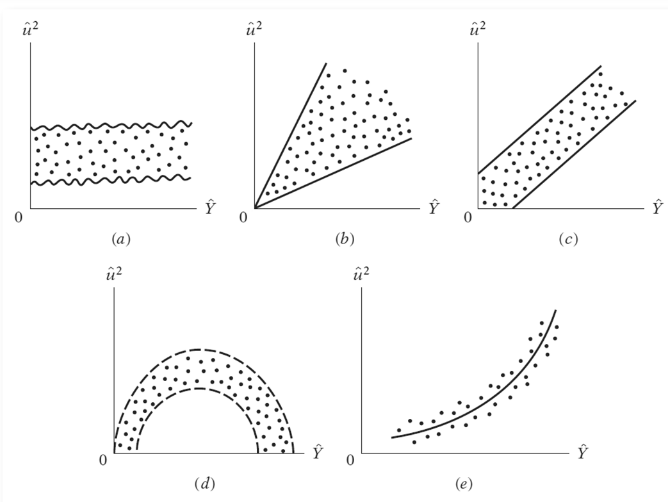
1) Graphical method

• **Step 1**: Estimate coefficients with OLS.

• **Step 2**: Retrieve $\hat{u}_i^2$. (In STATA, look for predicting residual)

• **Step 3**: Plot $\hat{u}_i^2$ with X_i or $\hat{Y}_i$. There might be multiple X_i in our regression function, so we can rely on $\hat{Y}_i$ as well.

Graphical method detecting heteroscedasticity



Which of these plots that heteroscedasticity is present?

2) Park Test

The intuition of Park test assumes that when heteroscedasticity is present, σ_i^2 is a kind of function of X_i. Steps are as follows.

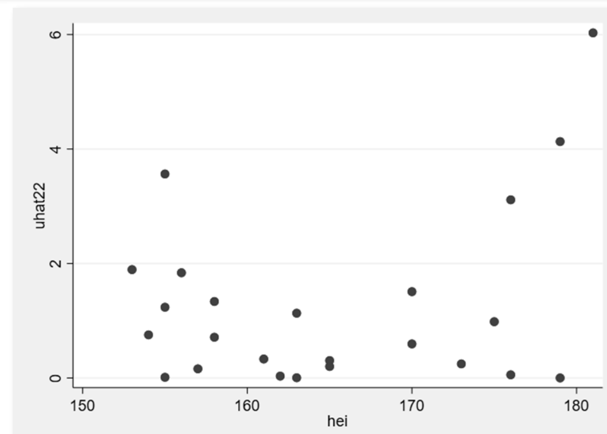
• **Step 1**: Estimate coefficients with OLS and retrieve $\ln \hat{u}_i^2$

• **Step 2**: Estimate $\ln \hat{u}_i^2 = \alpha + \beta \ln X_i + \nu_i$

• **Step 3**: Test the significance from zero of β. If it is, it would suggest that heteroscedasticity is present.

Example: Revisit the shoe size model with only one explanatory variable or height, when we plot $\hat{u}_i^2$ with height, the result is the picture below.

Graphical method of $\hat{u}_i^2$ and height



(3) Detecting heteroscedasticity

G. 387

The results of Park test also suggests that heteroscedasticity is not present in this model.

Park Test: shoe size model

```
    Source |       SS           df       MS      Number of obs   =        24
-----------+----------------------------------   F(1, 22)        =      0.05
     Model |  .256015921         1  .256015921   Prob > F        =    0.8259
  Residual |  113.705177        22  5.16841713   R-squared       =    0.0022
-----------+----------------------------------   Adj R-squared   =   -0.0431
     Total |  113.961193        23  4.95483447   Root MSE        =    2.2734

------------------------------------------------------------------------------
     uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
      hei2 |  -1.907324   8.569783    -0.22   0.826    -19.67997    15.86532
     _cons |     8.7586   43.74451     0.20   0.843    -81.96196    99.47916
------------------------------------------------------------------------------
```

Glejser Test, *Spearman's Rank Correlation Test*, and *Goldfeld-Quandt Test* will be skipped at this point.

3) Breusch-Pagan (BP) Test

This is a general case for other tests that follow Breusch and Pagan's idea (such as the Breusch-Pagan-Godfrey: BPG test in the textbook). This test is taken from Wooldridge page 270.

- **Step 1**: Estimate coefficients with OLS and retrieve $\hat{u}_i^2$.

- **Step 2**: Regress $\hat{u}_i^2 = \delta_1 + \delta_2 X_{2i} + \cdots + \delta_k X_{ki} + \nu_i$ and retrieve $R^2_{\hat{u}_i^2}$.

- **Step 3**: Calculate F-stat by

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

- **Step 4**: Test the null hypothesis of homoscedasticity. If we can reject the null hypothesis ($F_{cal} > F_{cri}$) at the selected significant level, heteroscedasticity is present in our model.

4) White's Test

White's test is also very similar to Breusch-Pagan. The differences are the residual model and test statistics. The steps here applies for 2 explanatory variables, but it is extendable. White's test has an advantage over BPG test as it is not sensitive to the assumption of normality.

- **Step 1**: Estimate coefficients with OLS and retrieve $\hat{u}_i^2$.

- **Step 2**: Regress

$$\hat{u}_i^2 = \delta_1 + \delta_2 X_{2i} + \delta_3 X_{3i} + \delta_4 X_{2i}^2 + \delta_5 X_{3i}^2 + +\delta_6 X_{2i} X_{3i} + \nu_i$$

and retrieve $R^2_{\hat{u}_i^2}$.

Additions of higher power and cross product imply that the error variance is functionally related to regressors, their squares, and their cross product.

(3) Detecting heteroscedasticity

• **Step 3**: Calculate the test stat, which in this case, we use Lagrange Multiplier (LM) stat

$$LM_{cal} = n \cdot R^2_{\hat{u}_i^2} \sim \chi^2_{k-1}.$$

LM is very useful in many cases due to its basic calculation and **asymptotically** distributed as Chi-square with k d.f. (number of coefficients minus 1 which in this case is 5).

• **Step 4**: Test the null hypothesis of homoscedasticity. If we can reject the null hypothesis ($LM_{cal} > \chi^2_{k-1.}$) at the selected significant level, heteroscedasticity is present in our model.

<u>**Example**</u>: Revisit the shoe size model now with 2 explanatory variables, height and weight.

• $ss_i = \beta_1 + \beta_2 hei_i + \beta_3 wei_i + u_i$

We will perform both BP and White's test. The models for the residual are respectively as follows.

• $\hat{u}_i^2 = \delta_1 + \delta_2 hei_i + \delta_3 wei_i + \nu_i$ (BP)

• $\hat{u}_i^2 = \delta_1 + \delta_2 hei_i + \delta_3 wei_i + \delta_4 hei_i^2 + \delta_5 wei_i^2 + \delta_6 hei_i wei_i + \nu_i$ (White's)

<u>**Regression results**</u>

```
      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(2, 21)        =     49.99
       Model |  120.620984         2  60.3104919   Prob > F        =    0.0000
    Residual |  25.3373496        21  1.20654045   R-squared       =    0.8264
-------------+----------------------------------   Adj R-squared   =    0.8099
       Total |  145.958333        23  6.34601449   Root MSE        =    1.0984

------------------------------------------------------------------------------
          ss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hei |   .2010805   .0328522     6.12   0.000     .1327606    .2694004
         wei |   .0632965    .031604     2.00   0.058    -.0024276    .1290206
       _cons |   2.866724   4.484679     0.64   0.530    -6.459676    12.19312
------------------------------------------------------------------------------
```

<u>**BP and White's test**</u>

```
      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(2, 21)        =      3.70
       Model |   10.364705         2  5.18235251   Prob > F        =    0.0422
    Residual |  29.4468707        21  1.40223194   R-squared       =    0.2603
-------------+----------------------------------   Adj R-squared   =    0.1899
       Total |  39.8115757        23  1.73093807   Root MSE        =    1.1842

------------------------------------------------------------------------------
       uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hei |   .0588498   .0354163     1.66   0.111    -.0148025    .1325021
         wei |   .0186662   .0340707     0.55   0.590    -.0521877      .08952
       _cons |  -9.685683   4.834709    -2.00   0.058    -19.74001    .3686444
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(5, 18)        =      4.01
       Model |  20.9803714         5  4.19607428   Prob > F        =    0.0128
    Residual |  18.8312043        18  1.04617802   R-squared       =    0.5270
-------------+----------------------------------   Adj R-squared   =    0.3956
       Total |  39.8115757        23  1.73093807   Root MSE        =    1.0228

------------------------------------------------------------------------------
       uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hei |  -2.913035    1.30776    -2.23   0.039    -5.660537   -.1655328
         wei |  -.4906248    .773815    -0.63   0.534     -2.11635      1.1351
        hei2 |   .0080558   .0046406     1.74   0.100    -.0016938    .0178054
        wei2 |  -.0027255   .0037403    -0.73   0.476    -.0105835    .0051324
      heiwei |   .0049706   .0066631     0.75   0.465     -.009028    .0189693
       _cons |   251.8105   99.79725     2.52   0.021     42.14427    461.4768
------------------------------------------------------------------------------
```

(3) Detecting heteroscedasticity

Calculate the test stats respectively

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

- $LM_{cal} = n \cdot R^2_{\hat{u}_i^2} =$

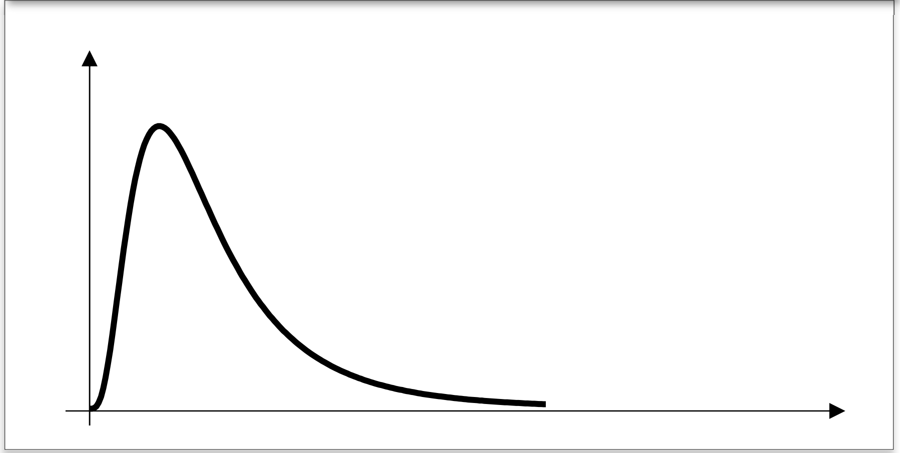
Then, find the critical value ($\alpha = 0.05$) and test the null hypothesis of homoscedasticity.

- $F_{0.05(k,n-k-1)} =$

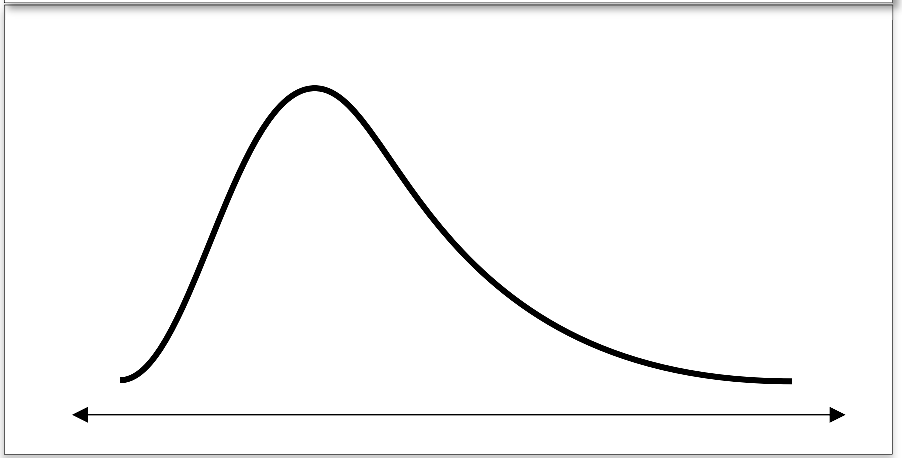
- $\chi^2_{\alpha,k-1.} =$

Conclude the tests below.

Simplified F-distribution



Simplified Chi-square



(3) Detecting heteroscedasticity

Well, actually, in STATA it is far simpler.

Note that the BP test package in STATA is not the general one, but it is the Breusch-Pagan / Cook-Weisberg test, which is a class of BP test. Hence, in the result, it relies on chi-square instead of F.

Additional note on the White's test is that

- The test that includes cross-product terms, it is a test of both heteroscedasticity and specification bias.
- If it is excluded, it is a pure heteroscedasticity test.

Regression results and tests

```
. reg ss hei wei

      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(2, 21)        =     49.99
       Model |  120.620984         2  60.3104919   Prob > F        =    0.0000
    Residual |  25.3373496        21  1.20654045   R-squared       =    0.8264
-------------+----------------------------------   Adj R-squared   =    0.8099
       Total |  145.958333        23  6.34601449   Root MSE        =    1.0984

------------------------------------------------------------------------------
          ss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hei |   .2010805   .0328522     6.12   0.000     .1327606    .2694004
         wei |   .0632965    .031604     2.00   0.058    -.0024276    .1290206
       _cons |   2.866724   4.484679     0.64   0.530    -6.459676    12.19312
------------------------------------------------------------------------------

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of ss

         chi2(1)      =     4.65
         Prob > chi2  =   0.0311

. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      12.65      5    0.0269
            Skewness |       0.24      2    0.8886
            Kurtosis |       0.78      1    0.3778
---------------------+-----------------------------
               Total |      13.66      8    0.0910
---------------------------------------------------
```

(4) Remedial measures

G. 389

(1) Weighted Least Squares (WLS)

It can be estimated when σ_i^2 is known.

(2) Data transformation: selecting a class of GLS

Advantage of this method is that we do not need asymptotic property. However, a major drawback is we need to speculate relationship between σ_i^2 and X_i.

For example, if we assume that σ_i^2 is proportional to X_i so that

- $E(u_i^2) = \sigma^2 X_i$ then $\sigma^2 = \frac{E(u_i^2)}{X_i} = \frac{E(u_i)}{\sqrt{X_i}}$

we can transform our model into

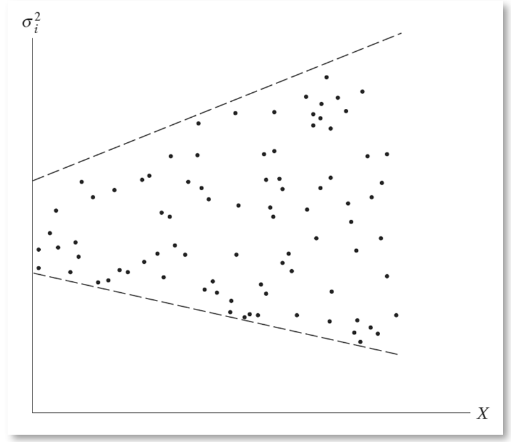
- $\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + \frac{u_i}{\sqrt{X_i}}$ where X_i must be > 0

$E\left(\frac{u_i}{\sqrt{X_i}}\right) = \sigma^2$ or homoscedastic.

However, we can see another drawback of this method is that the interpretation of coefficients are now different.

Moreover, there are multiple forms of transformation, due to relationship between σ_i^2 and X_i.

When $E(u_i^2) = \sigma^2 X_i$



(3) White's robust standard errors

This method assumes asymptotic property of our data, which is quite common in national cross-sectional data.

We do not cover how it is derived, even in the book Gujarati does not as well, but we compared the result of normal estimation with using White's robust standard errors.

(4) Remedial measures

Noted that White's robust standard errors do not change the estimated value of coefficients. The only differences here are standard errors.

In real-world scenario, both of these estimations are presented concurrently to check if there is any difference in drawing significance conclusion. Especially, when our data has large number of observations.

Disregarding heteroscedasticity versus White's robust S.E.

```
. reg ss hei wei

      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(2, 21)        =     49.99
       Model |  120.620984         2  60.3104919   Prob > F        =    0.0000
    Residual |  25.3373496        21  1.20654045   R-squared       =    0.8264
-------------+----------------------------------   Adj R-squared   =    0.8099
       Total |  145.958333        23  6.34601449   Root MSE        =    1.0984

------------------------------------------------------------------------------
          ss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hei |   .2010805   .0328522     6.12   0.000     .1327606    .2694004
         wei |   .0632965    .031604     2.00   0.058    -.0024276    .1290206
       _cons |   2.866724   4.484679     0.64   0.530    -6.459676    12.19312
------------------------------------------------------------------------------

. reg ss hei wei, rob

Linear regression                               Number of obs     =         24
                                                F(2, 21)          =      31.90
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8264
                                                Root MSE          =     1.0984

------------------------------------------------------------------------------
             |               Robust
          ss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hei |   .2010805   .0382944     5.25   0.000     .1214429     .280718
         wei |   .0632965   .0279869     2.26   0.034     .0050946    .1214984
       _cons |   2.866724   5.552371     0.52   0.611    -8.680064    14.41351
------------------------------------------------------------------------------
```