Making sense of Econometrics: Basics

Lecture 4: Qualitative influences and Heteroskedasticity

Hany Abdel-Latif & Anita Staneva

Egypt Scholars Economic Society

November 1, 2014

Assignment & feedback

enter classroom at

`http://b.socrative.com/login/student/`

room name c28efb78

Outline

nature of qualitative variables

- sometimes cannot obtain set of numerical values for all variables to use in a model
- because some variables cannot be quantified easily

examples

i. gender may play a role in determining salary levels
ii. different ethnic groups may follow different consumption patterns
iii. educational levels can affect earnings from employment

- in order to include factors like above, we define the so called **dummy variables**

nature of qualitative variables

- easier to have dummies for cross-sectional variables, but sometimes we have for time series as well

<u>examples</u>

i changes in political regime may affect production

ii war can impact on economic activities

iii certain days in week or certain months in year can have different effects on the fluctuation of stock prices

iv seasonal effects often observed in demand of various products

use of dummy variables

- consider following cross-sectional model

$$\underset{wage}{Y_i} = \beta_1 + \underset{work\ \ exp.}{\beta_2 X_i} + u_i$$

- what does the constant β_1 measure?
- this model assumes that the constant will be the same for all the observations in our data set
- what if we have two different subgroups
  - male and female, for example

أن كل عقل يفرق

use of dummy variables

- question is how to quantify the information that comes from the difference in the two groups
- we convert such qualitative information into a quantitative variable by creating a "dummy variable"

$$D = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

- note that
  - i the choice of which of the two different outcomes is to be assigned the value of 1 does not alter the results
  - ii the 0 classification is often referred to as the benchmark, or control category

use of dummy variables

- entering this dummy in the equation, we have the following model

$$\underset{wage}{Y_i} = \beta_1 + \underset{work\ \ exp.}{\beta_2 X_i} + \underset{dummy}{\beta_3 D_i} + u_i$$

- now we have two cases
  1. when $D = 0$ (female)

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

  2. when $D = 1$ (male)

$$Y_i = (\beta_1 + \beta_3) + \beta_2 X_i + u_i$$

Note: 2 groups (male & female) = 2 -1 dummies (D_i)

additive dummy variable

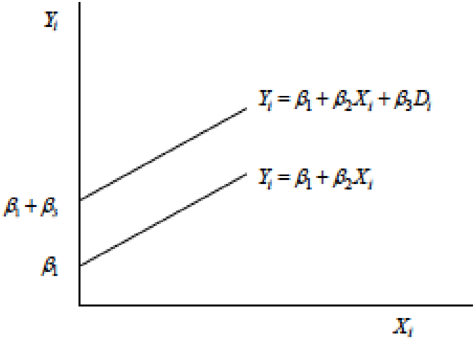
$$\underset{wage}{Y_i} = \beta_1 + \underset{work\ \ exp.}{\beta_2 X_i} + \underset{dummy}{\beta_3 D_i} + u_i$$

- dummy variable is included in "additive" form
  - as another regressor together with its corresponding coefficient
- the effect of the qualitative influence changes the intercept β_1
  - i.e., but not the 'slope' coefficient β_2
- β_3 is called the 'differential intercept coefficient'
- we can test the statistical significance of the qualitative influence (using t-test)
  - $H_0 : \beta_3 = 0$ vs. $H_1 : \beta_3 \neq 0$

additive dummy variable

$$\underbrace{Y_i}_{wage} = \beta_1 + \underbrace{\beta_2 X_i}_{work\ \ exp.} + \underbrace{\beta_3 D_i}_{dummy} + u_i$$

For $\beta_3 > 0$



multiplicative dummy variable

- situations where the non-quantitative factors affect one or more coefficients of the explanatory variables (slopes)
- hence, the dummy variables are included in "multiplicative" or "interactive" mode
  - i.e., multiplied by the corresponding regressor
- example
  - difference in the rate of consumption *MPC* between male and female

multiplicative dummy variable

- consider same case but now with the dummy affecting the slope

$$\underset{wage}{Y_i} = \beta_1 + \underset{work\ \ exp.}{\beta_2 X_i} + \underset{dummy}{\beta_3 D_i X_i} + u_i$$

- now we have two cases
  1. when $D = 0$ (female)

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

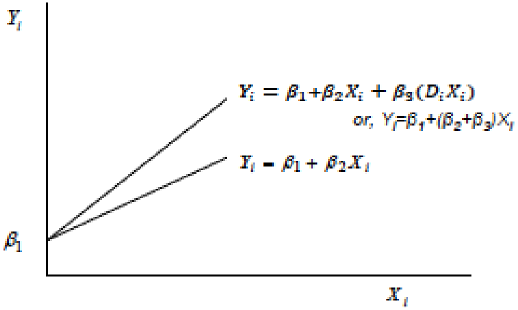
  2. when $D = 1$ (male)

$$Y_i = \beta_1 + (\beta_2 + \beta_3) X_i + u_i$$

multiplicative dummy variable

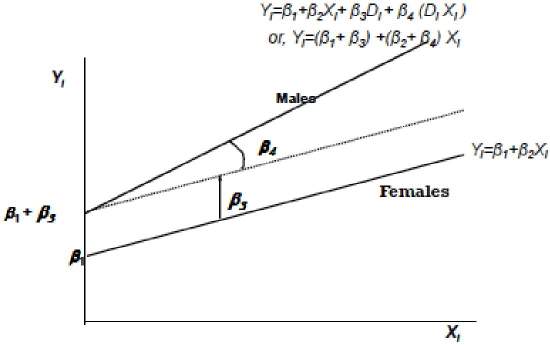
$$\underset{wage}{Y_i} = \beta_1 + \underset{work\ \ exp.}{\beta_2 X_i} + \underset{dummy}{\beta_3 D_i X_i} + u_i$$

For $\beta_3 > 0$



combined (additive & multiplicative) dummies

$$Y_i = \beta_1 + \beta_2 X_i + \underbrace{\beta_3 D_i}_{add.\ dummy} + \underbrace{\beta_4 D_i X_i}_{mult.\ dummy} + u_i$$



multiple dummies

- when you compare gender gap, there are only two groups (males & females)
- in some situations, there are more than 2 categories
- for example, you may want to examine the gender differences among the following four groups
  - married men
  - married women
  - single men
  - single women
- create dummy variables for all the categories except one category

multiple dummies

- we can define
  - $D_{1i} = 1$ if married man; 0 otherwise
  - $D_{2i} = 1$ if married woman; 0 otherwise
  - $D_{3i} = 1$ if single woman; 0 otherwise
- the excluded group is the single male
- the differences in wage among the four groups are estimated relative to single males
- we estimate

$$\underset{wage}{Y_i} = \beta_1 + \underset{work\ \ exp.}{\beta_2 X_i} + \underset{marr.\ man}{a_1 D_{1i}} + \underset{marr.\ wom.}{a_2 D_{2i}} + \underset{sing.\ man}{a_3 D_{3i}} + u_i$$

- consider various cases
  - i.e. $D_{1i} = 1,\ D_{2i} = D_{3i} = 0$ and so on

dummy variable trap

- using more than one dummy variable
  - gender (male; female)
  - education (primary; secondary; tertiary; BSc; MSc) and so on
- the interpretation (although it seems more complicated) is the same as before
- golden rule in applying dummy variables
  - use one less dummy variable than there are categories of classification for each qualitative influence (if an intercept is present)
  - including as many dummies as categories (and a constant) leads to the 'dummy variable trap'
  - this induces 'perfect multicollinearity', a situation where assumption **A6** of the CLRM is violated and coefficients cannot be estimated

seasonality

- consumer expenditure is usually
  - highest during the 4^{th} quarter of the year (i.e. Christmas)
  - lowest during the 1^{st} quarter (New Year)
  - then higher again in 2^{nd} quarter (summer)
  - lower in the 3^{rd} quarter (autumn)
- dummy variables provide one way of removing such systematic seasonal effects
  - e.g., quarterly; monthly
- again, we exclude one if constant term is present in the model

seasonality

- we can define

$$D_{2t} = \begin{cases} 1 & \text{if t is } 2^{nd} \text{ quarter} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{3t} = \begin{cases} 1 & \text{if t is } 3^{rd} \text{ quarter} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{4t} = \begin{cases} 1 & \text{if t is } 4^{th} \text{ quarter} \\ 0 & \text{otherwise} \end{cases}$$

seasonality

- and estimate

$$Y_t = \beta_1 + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + \beta_5 X_t + u_t$$

- note that
  - we have 4 groups and 3 dummy variables
  - the 1^{st} quarter effect will be captured by β_1
  - the dummy variable coefficients measure the intercept differences relative to β_1

hourly wage equation

Example 7.1 Wooldridge (1st & 2nd eds.)

TABLE 7.1 A Partial Listing of the Data in WAGE1.RAW

person	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

hourly wage equation

Example 7.1 Wooldridge (1st & 2nd eds.)

- we believe that wage depends on education, experience and tenure
- this can be shown as follows

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

- we want to test if there is wage discrimination?
  - male-female wage differential
- this can be done by using a dummy variable as follows

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

hourly wage equation

Example 7.1 Wooldridge (1st & 2nd eds.)

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

- *educ*, *exper*, and *tenure* are all relevant productivity characteristics
- *female* is a dummy variable as follows

$$female = \begin{cases} 1 & \text{if a female} \\ 0 & \text{if a male} \end{cases}$$

- we have 2 groups and 2-1 dummies
- male is our base (reference) group in this example

hourly wage equation

Example 7.1 Wooldridge (1st & 2nd eds.)

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

- we want to examine the wage discrimination between males and females
- we test the null hypothesis of no difference between men and women
  - $H_0 : \delta_0 = 0$
- against the alternative that there is discrimination against women
  - $H_1 : \delta_0 \neq 0$

hourly wage equation

Example 7.1 Wooldridge (1st & 2nd eds.)

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

- how can we actually test for wage discrimination?
  - estimate the model by OLS
  - use the usual t statistic
- nothing changes about the mechanics of OLS when some of the regressors dummy variables
- the only difference is in the interpretation of the coefficient on the dummy variable

hourly wage equation

Example 7.1 Wooldridge (1st & 2nd eds.)
import the data

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/wage1

.
. reg wage female educ exper tenure

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  4,   521) =   74.40
       Model |  2603.10658     4  650.776644           Prob > F      =  0.0000
    Residual |  4557.30771   521   8.7472317           R-squared     =  0.3635
-------------+------------------------------           Adj R-squared =  0.3587
       Total |  7160.41429   525  13.6388844           Root MSE      =  2.9576

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -1.810852   .2648252    -6.84   0.000    -2.331109   -1.290596
        educ |   .5715048   .0493373    11.58   0.000     .4745802    .6684293
       exper |   .0253959   .0115694     2.20   0.029     .0026674    .0481243
      tenure |   .1410051   .0211617     6.66   0.000     .0994323    .1825778
       _cons |  -1.567939   .7245511    -2.16   0.031    -2.991339    -.144538
------------------------------------------------------------------------------
```

hourly wage equation

$$\widehat{wage} = \underset{(0.031)}{-1.57} \underset{(0.000)}{-1.81} female \underset{(0.000)}{+0.57} educ + \underset{(0.000)}{0.025} exper + \underset{(0.000)}{0.14} tenure$$

p-value parentheses, $n = 526$, $R^2 = 0.3635$

- the negative intercept - the intercept for men, in this case - is not very meaningful, since no one has close to zero years of educ, exper, and tenure in the sample
- the coefficient on female measures the average difference in hourly wage between a woman and a man, given the same levels of educ, exper, and tenure

hourly wage equation

$$\widehat{wage} = \underset{(0.031)}{-1.57} \underset{(0.000)}{-1.81} female \underset{(0.000)}{+0.57} educ + \underset{(0.000)}{0.025} exper + \underset{(0.000)}{0.14} tenure$$

p-value parentheses, $n = 526$, $R^2 = 0.3635$

- if we take a woman and a man with the same levels of education, experience, and tenure, the woman earns, on average, \$1.81 less per hour than the man
- because we controlled for educ, exper, and tenure, the \$1.81 wage differential cannot be explained by different average levels of education, experience, or tenure between men and women

hourly wage equation

$$\hat{wage} = \underset{(0.031)}{-1.57} \underset{(0.000)}{-1.81}female \underset{(0.000)}{+0.57}educ + \underset{(0.000)}{0.025}exper + \underset{(0.000)}{0.14}tenure$$

p-value parentheses, $n = 526$, $R^2 = 0.3635$

- note that we reject the null hypothesis of $\delta_0 = 0$
- therefore we can conclude that the differential of 1.81 is due to gender

savings model

- consider the two-variable model

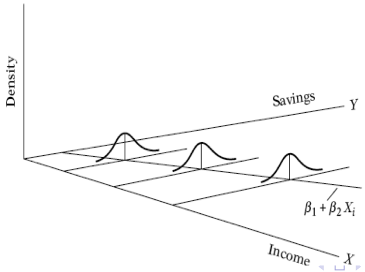
$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

- Y and X represent savings and income, respectively
- as income increases, savings on the average also increase
- **homoscedasticity**
    - the variance of savings remains the same at all levels of income

homoscedastic disturbance

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

- The CLRM assumes **homoscedasticity**
  - $var(u_i) = E(u_i^2|X_i) = \sigma^2$ i.e., constant variance
  - equal (homo) spread (scedasticity) or equal variance



Heteroscedastisity

- if the variances of u_i are not the same, there is **heteroscedasticity**
  - $var(u_i) = E(u_i^2|X_i) = \sigma_i^2$
  - the subscript of σ^2 refers to non-constant conditional variances of u_i

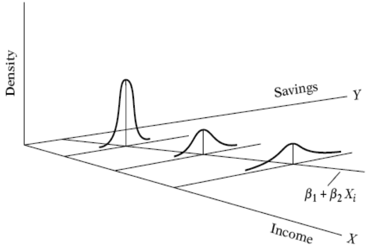


illustration of heteroscedasticity

- **why the variances of u_i may be variable**
  - following the error-learning models, as people learn, their errors of behaviour become smaller over time (σ_i^2 is expected to decrease)

example

- the number of typing errors made in a given time period on a test to the hours put in typing practice

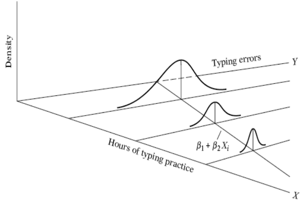


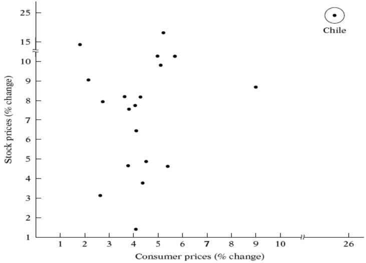
FIGURE 11.3 Illustration of heteroscedasticity.

illustration of heteroscedasticity

- **why the variances of u_i may be variable**
  - income growth
    - as income grows people have more discretionary income and hence more scope for choice about the disposition of their income. therefore, σ_i^2 is likely to increase with income
    - companies with larger profits are generally expected to show greater variability in their dividend policies than companies with lower profits
  - data collection
    - as data collecting techniques improve, σ_i^2 is likely to decrease
    - e.g., banks that have sophisticated data processing equipment are likely to commit fewer errors in the monthly or quarterly statements of their customers than banks without such facilities

illustration of heteroscedasticity

- **why the variances of u_i may be variable**
  - the presence of outliers
    - inclusion or exclusion of such an observation, especially if the sample size is small, can substantially alter the results of regression analysis



chile is an outlier because the given Y and X values are much larger than for the rest of the countries

illustration of heteroscedasticity

- **why the variances of u_i may be variable**
  - model misspecification
    - the CLRM assumes that the regression model is correctly specified
    - omitting some important variables from the model may lead to heterscedasticity
  - skewness in the distribution of one or more regressors included in the model
    - e.g., economic variables such as income, wealth, and education
    - the distribution of income and wealth in most societies is uneven, with the bulk of the income and wealth being owned by a few at the top

illustration of heteroscedasticity

- **why the variances of u_i may be variable**
  - incorrect data transformation
    - e.g., ratio or first difference transformations
  - incorrect functional form
    - e.g., linear versus log–linear models

note that

- heteroscedasticity is likely to be more common in cross-sectional than in time series data
- cross-sectional data:
  - members may be of different sizes, such as small or large firms
- time series data:
  - variables tend to be of similar orders of magnitude e.g., GNP, consumption expenditure, savings

consequences of heteroscedasticity

- OLS estimators are still unbiased
  - as long as mean of the error $E(u_i) = 0$
  - the sample and population estimates will be equal irrespective of whether there is heteroscedasticity or not
- OLS estimators are no longer efficient or have minimum variance
  - minimum variance can be achieved when variance across cross-section remains constant
  - when they vary significantly across cross-section, one would always expect higher variance of the estimators
  - hence, the estimators are no longer efficient

consequences of heteroscedasticity

- the formula used to estimate the coefficient standard errors are no longer correct
  - so the t-tests will be misleading
  - if the error variance is positively related to an independent variable then the estimated standard errors are biased downwards and hence the t-values will be inflated
  - confidence intervals based on these standard errors will be wrong

detection of heteroscedasticity

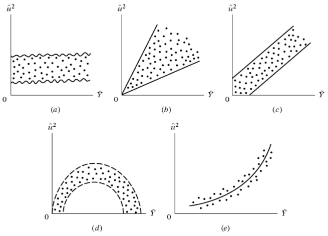
- two ways in general
  - informal (graphically)
  - formal
- examine the OLS residuals $\hat{u}_i$
  - since they are the ones we observe, and not the disturbances u_i
  - one hopes that they are good estimates of u_i
  - a hope that may be fulfilled if the sample size is fairly large

informal methods

- nature of the problem
  - very often the nature of the problem under consideration suggests whether heteroscedasticity is likely to be encountered
  - e.g., in cross-sectional data involving heterogeneous units, heteroscedasticity may be the rule rather than the exception
- in practice, one can examine the residual squared $\hat{u}_i^2$ to see if they exhibit any systematic pattern
  - although $\hat{u}_i^2$ are not the same as u_i^2, they can be used as proxies especially if the sample size is sufficiently large

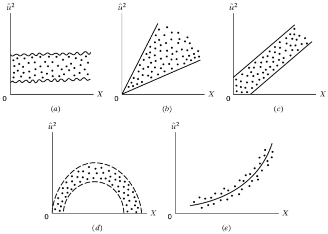
informal methods

- figures below represent hypothetical patterns of estimated squared residuals
- fig. (a) no systematic pattern between the two variables suggesting no heteroscedasticity is present in the data
- fig. (b) to (e) exhibit definite patterns
  - fig. (c) a linear relationship, fig. (d) and (e) a quadratic relationship



informal methods

- instead of plotting $\hat{u}_i^2$ against $\hat{Y}_i$, one may plot them against one of the explanatory variables X_i
- such pattern is shown in fig. (c)- the variance of the disturbance term is linearly related to the X variable



formal tests

- Goldfeld-Quandt test
- White's test
- Breusch-Pagan LM test
- Park LM test

Goldfeld-Quandt test

- assumes that the heteroscedastic variance σ_i^2 is a function of one of the explanatory variables
- known as 'classical heteroscedasticity' (e.g. as $X_i \uparrow \sigma_i^2 \uparrow$)
- consider the usual two-variable model:
  $Y_i = \beta_1 + \beta_2 X_i + u_i$
- suppose σ_i^2 is positively related to X_i as: $\sigma^2 X_i^2$, where σ^2 is a constant

Goldfeld-Quandt test

- step 1
  - identify one variable that is closely related to variance of the disturbances (e.g, X_i)
  - order (rank) the observations in ascending order (from lowest to highest)
- step 2
  - split the ordered sample into two equally sized sub-samples by omitting p central observations
  - so that the two samples will contain $1/2(n-p)$ observations
- step 3
  - run an OLS regression of Y on the X variable used in step 1 for each sub-sample and obtain RSS for each equation

Goldfeld-Quandt test

- compute the test statistic $GQ = \frac{RSS_2}{RSS_1}$
  - where RSS_2 is the RSS with the largest value and RSS_1 is the RSS with the smallest values
- under the null hypothesis of homoscedasticity and if ui are assumed to be normally distributed

$$GQ \sim F(\frac{n-p-2k}{2}, \frac{n-p-2k}{2})df$$

أن كل عقل يفرق

Goldfeld-Quandt test

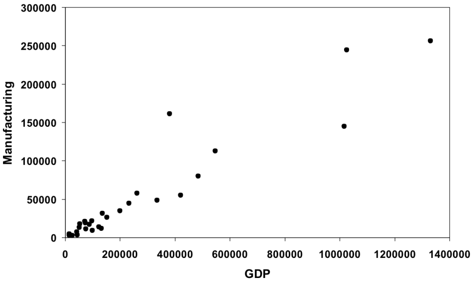
- step 4
  - reject the H_0 of homoscedastisity if $GQ - statistic > F^c$

note that

- the power of the test depends on the choice p
- choosing p too small is probably less problematic than choosing too big
- a simple rule for determining the choice of p
  - $p \cong \frac{4}{5} . \frac{n}{3}$ where n is sample size
  - e.g., for $n = 30$, $p = 8$; for $n = 60, p = 16$
- in multiple regression, repeat the GQ test for each explanatory variable

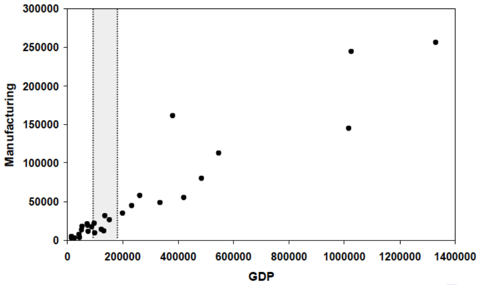
Goldfeld-Quandt example

- in the scatter diagram manufacturing output is plotted against GDP, both measured in US million dolars, for 28 countries for 1997
- the scatter diagram clearly points to the problem of heteroscedasticity



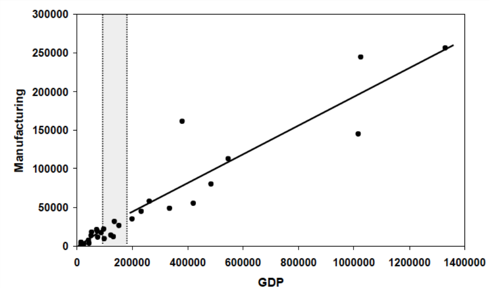
Goldfeld-Quandt example

- the sample of 28 observations is divided into three ranges
  - 11 of the observations with the smallest values of the X variable
  - 11 of the observations with the largest values
  - 6 in the middle



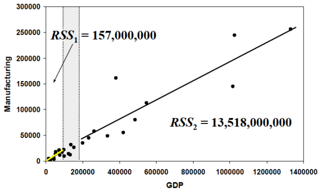
Goldfeld-Quandt example

- then fit regression lines to the lower and upper ranges of the observations
- the regression line for the lower range has been buried under the observations



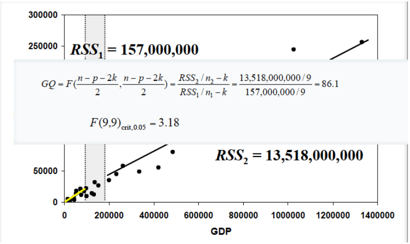
Goldfeld-Quandt example

- compare the residual sum of squares for the two regressions
- RSS_1 and RSS_2 denote the lower and upper ranges, respectively
  - if the disturbance term is homoscedastic, there should be no systematic difference between RSS_1 and RSS_2
  - if the standard deviation of the distribution of the disturbance term is proportional to the X variable, RSS_2 is likely to be greater than RSS_1



Goldfeld-Quandt example

- if it is greater, the question is whether it is significantly greater
- calculate GQ statistic
- since that $GQ > F^c$ (i.e., $86.1 > 3.18$), we reject the null hypothesis of homoscedasticity



$$GQ = F\left(\frac{n-p-2k}{2}, \frac{n-p-2k}{2}\right) = \frac{RSS_2/n_2-k}{RSS_1/n_1-k} = \frac{13{,}518{,}000{,}000/9}{157{,}000{,}000/9} = 86.1$$

$$F(9,9)_{\text{crit},0.05} = 3.18$$

Next Lecture

you should know

- assignment 2
  - available on BB tomorrow
  - due on November 15 (20:00 Cairo time)
  - to be emailed to `eses@egyptscholars.org`
- lecture 5
  - recorded
  - available on Saturday Nov. 8 20:00 (Cairo time)

next lecture

- regression in practice
  - autocorrelation
  - multicolinearity
  - specification bias