Dis 6: BLUE; Homoskedasticity vs. Heteroskedasticity; Comparing Two Populations

1 BLUE (Gauss-Markov theorem)

- BLUE is a feature of OLS (ordinary least squares) regression technique.

Theorem 1 (Gauss-Markov). OLS estimator of β_i, $\hat{\beta}_i$, is the **B**est (most efficient) **L**inear conditionally **U**nbiased **E**stimator (**BLUE**), as long as the following OLS assumptions holds:

1. **Zero conditional mean**: $E[u_i|x_1, x_2, \ldots, x_k] = 0$
2. **I.I.D. Data**: $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ are i.i.d. (independent and identically distributed).
3. **Large outliers are unlikely**: There doesn't exist some $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ that live in a dramatically different region. Could be measured as the fourth moment of each variable is finite (i.e. $0 < E[x_{1i}^4] < \infty$, $0 < E[x_{2i}^4] < \infty$, ..., $0 < E[x_{ki}^4] < \infty$, and $0 < E[y_i^4] < \infty$)
4. **No perfect multicollinearity**: One of the regressors cannot be a perfect linear function of some other regressors.
5. **Homoskedasticity**: $Var(u_i|x_1, x_2, \ldots, x_k) = \sigma^2$ is a constant.

- **Best** and **unbiased** are the two most important features:
  - **Best** is in the sense that the estimators achieved under OLS have the smallest standard errors compared with all other potential estimators. When the estimator has the smallest standard error possible, the estimator is called **the most efficient**.
  - **Unbiased** means that on average, estimators yield the true value: $E(\hat{\beta}_i) = \beta_i$.

2 Homoskedasticity vs. Heteroskedasticity

- Homoskedasticity and heteroskedasticity are features of **conditional variance of error term u_i given information on x_i**
  - When $Var(u_i|x_{1i}, x_{2i}, \ldots, x_{ki}) = \sigma^2$ (a constant), the error term u_i is **homoskedastic**
  - When $Var(u_i|x_{1i}, x_{2i}, \ldots, x_{ki}) \neq \sigma^2$, the error term u_i is **heteroskedastic**
    * **Impure heteroskedasticity**: $Var(u_i|x_{1i}, x_{2i}, \ldots, x_{ki}) \neq \sigma^2$ due to model misspecification (ex. omitted variable bias)
    * **Pure heteroskedasticity**: $Var(u_i|x_{1i}, x_{2i}, \ldots, x_{ki}) \neq \sigma^2$ arises even when the model is correctly specified.

  For the rest of this handout, we only consider pure heteroskedasticity.

- How do they reflect in data?
  - Consider linear model $y_i = \beta_0 + \beta_1 x_i + u_i$
  - Conditional variance of y_i given x_i looks like the following:

$$Var(y_i|x_i) = Var(\beta_0 + \beta_1 x_i + u_i|x_i)$$

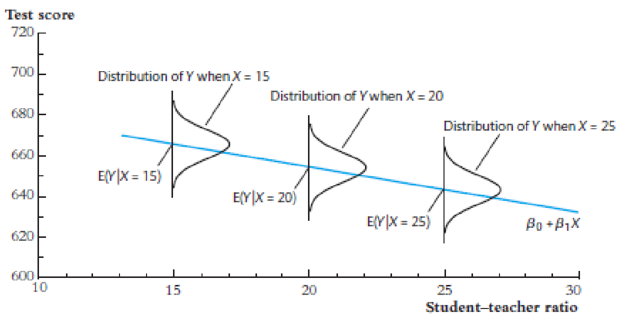
$$= Var(\beta_0) + Var(\beta_1 x_i | x_i) + Var(u_i | x_i) \qquad \text{(by i.i.d. data)}$$
$$= 0 + 0 + Var(u_i | x_i) = Var(u_i | x_i)$$

So $Var(u_i|x_i)$ is directly reflected onto $Var(y_i|x_i)$!

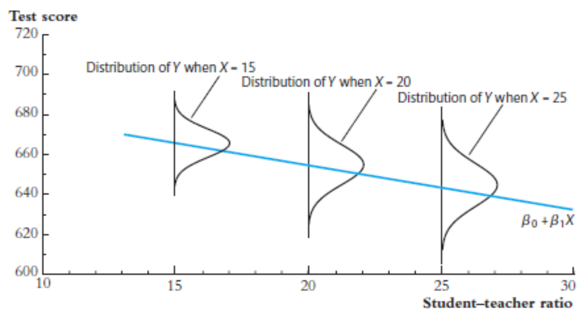
Ex. Say that we are interested in studying the relationship between test score and student-teacher ratio. A simple (univariate) linear regression model is proposed:

$$\text{test score}_i = \beta_0 + \beta_1 \text{student-teacher ratio}_i + u_i$$

Each type of error term reflects in data in the following way:



(a) Homoskedastic error

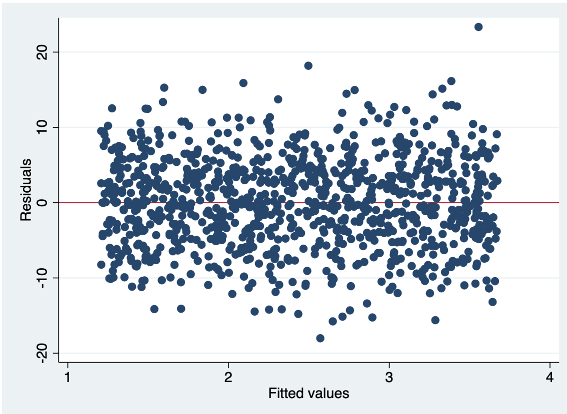


(b) Heteroskedastic error

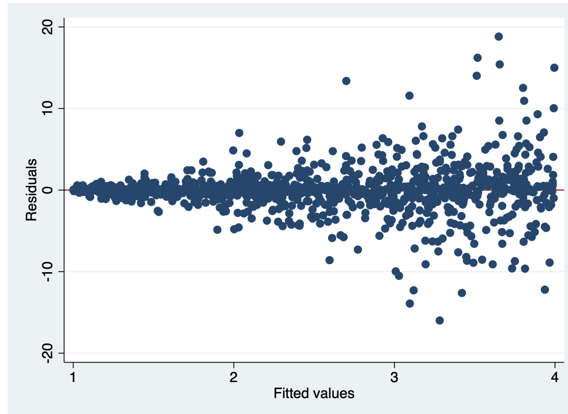
- What difference does these two types of error specification make?
  - OLS estimate of $\hat{\beta}_i$ is BLUE only under homoskedasticity; it's not under heteroskedasticity.
    Homoskedasticity simplifies the error structure $\rightarrow$ *in general*, standard error of the $\hat{\beta}$ estimator is smaller under homoskedasticity.
    This means that if we specify the error term to be heteroskedastic, then the **best** part of BLUE is violated under heteroskedasticity.
  - As long as model is correctly specified (either error is homoskedastic or pure heteroskedastic), then $\hat{\beta}_i$ yields unbiased estimate of the true β_i.
    Neither error specification affects the $\hat{\beta}_i$ point estimate though – the unbiased part of BLUE has nothing to do with how standard error of $\hat{\beta}_i$ looks like.
  - In many contexts, heteroskedasticity is the more accurate way to model the error structure.
- How to test whether heteroskedasticity is the way we should model the error term?
  - Visually: Plot residual $\hat{u}$ (sample analog of error) against predicted value $\hat{y}$.
    Doing this in Stata:

```
reg y x1 x2 x3 // run your regression first
rvfplot, yline(0) // the yline(0) option adds a horizontal line at y = 0
```

    Each type of error term has the following regress postestimation diagnostic plot:



(a) Homoskedastic error



(b) Heteroskedastic error

– Analytically: Breusch-Pagan or White test

<table>
<tr><th>Test</th><th>Stata Command and Output</th></tr>
<tr><td>Breusch-Pagan</td><td><pre>
reg y x1 x2 x3 // run your regression first
estat hettest, rhs fstat

. quietly reg y x

. estat hettest, rhs fstat

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: x

         F(1 , 998)   =   2.26
         Prob > F   =   0.1333
</pre></td></tr>
<tr><td>White</td><td><pre>
reg y x1 x2 x3 // run your regression first
estat imtest, white

. quietly reg y x

. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(2)   =   2.33
         Prob > chi2 =   0.3121

Cameron & Trivedi's decomposition of IM-test

              Source | chi2   df   p
---------------------+-----------------------------
  Heteroskedasticity | 2.33   2   0.3121
            Skewness | 0.77   1   0.3788
            Kurtosis | 0.33   1   0.5649
---------------------+-----------------------------
               Total | 3.44   4   0.4878

</pre></td></tr>
</table>

- How to incorporate homoskedasticity and heteroskedasticity in regression model estimation?
  - Homoskedasticity and heteroskedasticity only affect the standard error of $\hat{\beta}$ estimators
  - Calculating by hand:
    - Under homoskedasticity:

$$Var(\hat{\beta}_1|x_1, x_2, \ldots, x_k) = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}$$

    - Under heteroskedasticity:

$$Var(\hat{\beta}_1|x_1, x_2, \ldots, x_k) = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2\hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2\right]^2}$$

    Homoskedasticity can be viewed as a special case of heteroskedasticity: $\hat{u}_i$ varies constantly across all i under homoskedasticity.

    Squared root of the variance under heteroskedasticity is called **robust standard error**.
  - Using Stata:
    - Under homoskedasticity: same as what we've been doing (homoskedasticity is the default for linear regression model)

      `reg y x1 x2 x3 x4`

```
. use "http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.dta", clear

. reg wage educ exper hours

      Source |       SS           df       MS      Number of obs   =       935
-------------+----------------------------------   F(3, 931)       =     49.31
       Model |  20939351.2         3  6979783.75   Prob > F        =    0.0000
    Residual |   131776817       931  141543.305   R-squared       =    0.1371
-------------+----------------------------------   Adj R-squared   =    0.1343
       Total |   152716168       934  163507.675   Root MSE        =    376.22

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    76.7356    6.31113    12.16   0.000     64.34991    89.12129
       exper |   17.55185   3.162026     5.55   0.000     11.34632    23.75737
       hours |  -1.995166     1.7116    -1.17   0.244    -5.354206    1.363875
       _cons |  -190.8808   128.0893    -1.49   0.137     -442.258    60.49628
------------------------------------------------------------------------------
```

    - Under heteroskedasticity: add `robust` option

      `reg y x1 x2 x3 x4, robust`

```
. reg wage educ exper hours, robust

Linear regression                               Number of obs     =        935
                                                F(3, 931)         =      43.29
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1371
                                                Root MSE          =     376.22

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    76.7356    6.74783    11.37   0.000     63.49288    89.97832
       exper |   17.55185   3.121662     5.62   0.000     11.42554    23.67816
       hours |  -1.995166   2.275691    -0.88   0.381    -6.461243    2.470912
       _cons |  -190.8808   146.1022    -1.31   0.192    -477.6087    95.84704
------------------------------------------------------------------------------
```

3 Comparing two populations: Econ 310 vs. Econ 400 regression technique

- Consider two different populations. In the context of PS 5 Question 1, let the two populations be female and male.

  We'd love to know something about the female and male population, so suppose that we constructed a representative sample, which contains two variables, x_{female} and x_{male}, that record income of female and male sampled from the respective population.

  [Correction begins here]

  - If the two populations have equal variances:

    $$t = \frac{(\overline{x_{\text{female}}} - \overline{x_{\text{male}}}) - (\mu_{\text{female},H_0} - \mu_{\text{male},H_0})}{\sqrt{\frac{s^2_{\text{pooled}}}{n_{\text{female}}} + \frac{s^2_{\text{pooled}}}{n_{\text{male}}}}} \quad \sim \quad t_{n_{\text{female}}+n_{\text{male}}-2}$$

    where

    $$s^2_{\text{pooled}} = \frac{\sum_i^{n_{\text{female}}}(x_{\text{female},i} - \overline{x_{\text{female}}})^2 + \sum_i^{n_{\text{male}}}(x_{\text{male},i} - \overline{x_{\text{male}}})^2}{n_{\text{female}} + n_{\text{male}} - 2}$$

  - If the two populations have unequal variances:

    $$t = \frac{(\overline{x_{\text{female}}} - \overline{x_{\text{male}}}) - (\mu_{\text{female},H_0} - \mu_{\text{male},H_0})}{\sqrt{\frac{s^2_{\text{female}}}{n_{\text{female}}} + \frac{s^2_{\text{male}}}{n_{\text{male}}}}} \quad \sim \quad t_{DOF}$$

    where

    $$DOF = \frac{(s^2_{\text{female}}/n_{\text{female}} + s^2_{\text{male}}/n_{\text{male}})^2}{\frac{(s^2_{\text{female}}/n_{\text{female}})^2}{n_{\text{female}}-1} + \frac{(s^2_{\text{male}}/n_{\text{male}})^2}{n_{\text{male}}-1}}$$

    (DOF stands for degree of freedom)

  [Correction ends here]

- How do we perform the same test using regression technique?

  $$\text{income}_i = \beta_0 + \beta_1 \text{female}_i + u_i$$

  where female is a dummy variable.

  Here,

  - β_0 records ____

  - β_1 records ____

So to test whether there's a difference between female and male population income, we can equivalently test whether β_1 is nonzero:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- To test whether $\beta_1 = 0$, we can use what we learned from Dis 4, and construct t-statistic for test. Recall that t-statistic looks like the following:

$$t = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{se(\hat{\beta}_1)} \quad \sim \quad t_{n-k-1}$$

- This tells us that how $se(\hat{\beta}_1)$ looks like matters:
  * When homoskedastic error based $se(\hat{\beta}_1)$ is used, this is equivalent to Econ 310's test under equal population variance.
  * When heteroskedastic error based $se(\hat{\beta}_1)$ (i.e. robust standard errors) is used, this is equivalent to Econ 310's test under unequal population variance.

4 Problems

1. Load the dataset from `http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.dta` into Stata (don't forget to first change your working directory).

   Dataset codebook is available at `http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.des`

   (a) Estimate the following multivariate linear model using Stata's default regression setting (i.e. assuming homoskedastic error):

   $$\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{black}_i + \beta_4 \text{urban}_i + \beta_5 \text{married}_i + \beta_6 \text{hours}_i + u_i$$

   (b) At 5% significance level, is there any slope coefficient that is not statistically significant?

   (c) Is the OLS estimator BLUE under the current configuration?

   (d) The default linear regression assumes homoskedastic error. One worries that heteroskedastic error is more appropriate here. Without performing any test, give a reason on why you'd think that heteroskedasticity might hold here.

(e) Perform a visual test on heteroskedasticity by creating the regression postestimation diagnostic plot (`rvfplot`).

(f) Perform Breusch-Pagan test on heteroskedasticity at 5% significance level.

(g) Perform White test on heteroskedasticity at 5% significance level.

(h) With the correct error specification, reestimate the linear regression model in (a).

(i) Did any of the estimated coefficient change?

(j) Did any of the standard error for coefficient change?

(k) Is the OLS estimator BLUE under the new configuration?

(l) At 5% significance level, is there any slope coefficient that is not statistically significant now?