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:

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

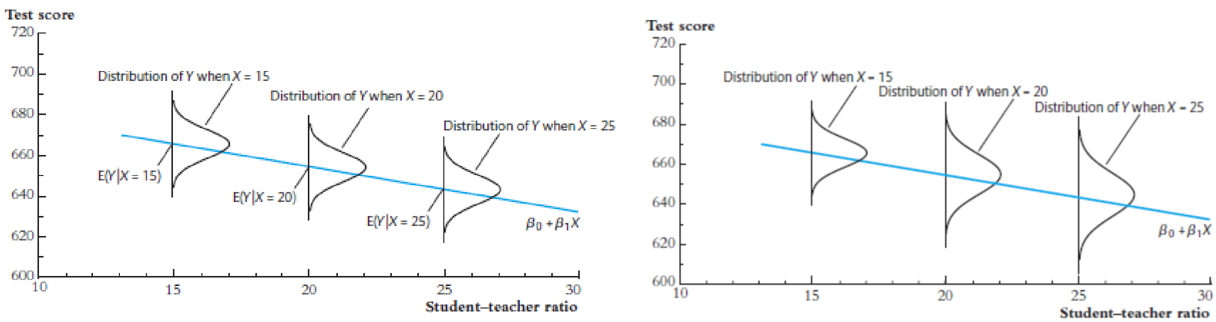
$$= Var(\beta_0) + Var(\beta_1 x_i | x_i) + Var(u_i | x_i) \qquad \text{(by i.i.d. data)}$$
$$= 0 + 0 + Var(u_i | x_i) = Var(u_i | x_i)$$

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:



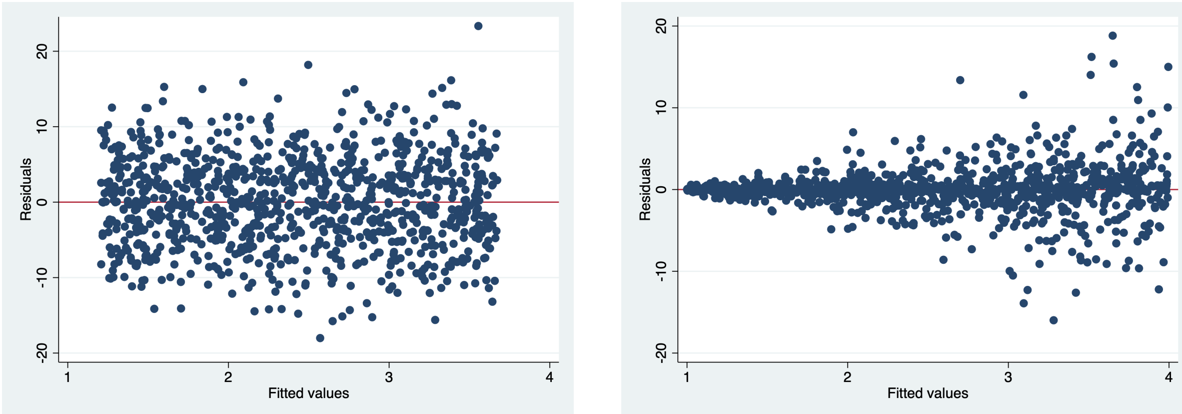
(a) Homoskedastic error

(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

Test	Stata Command and Output
Breusch-Pagan	`reg y x1 x2 x3 // run your regression first` `estat hettest, rhs fstat`
White	`reg y x1 x2 x3 // run your regression first` `estat imtest, white`

Breusch-Pagan output:

```
. 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
```

White output:

```
. 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
---------------------------------------------------
```

- 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.

- How can we test whether there's a difference between each group's population mean income?

$$H_0 : \mu_{\text{female}} = \mu_{\text{male}}$$
$$H_1 : \mu_{\text{female}} \neq \mu_{\text{male}}$$

  Recall from Econ 310 that such test can be performed using t-statistic:

  [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 mean income of male group

(Recall that $E[\text{income}|\text{female} = 0] = \beta_0$)

- β_1 records difference in mean income between male and female groups ($\mu_{\text{female}} - \mu_{\text{male}}$ to be exact)

  (Recall that $E[\text{income}|\text{female} = 1] = \beta_0 + \beta_1$)

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):

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

   Run the following commands in Stata:

```
use "http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.dta", clear
reg wage educ exper black urban married hours
```

   The output in Stata looks like the following:

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

. reg wage educ exper black urban married hours

      Source |       SS           df       MS      Number of obs   =       935
-------------+----------------------------------   F(6, 928)       =     42.94
       Model |  33186006.2         6  5531001.04   Prob > F        =    0.0000
    Residual |   119530162       928  128804.054   R-squared       =    0.2173
-------------+----------------------------------   Adj R-squared   =    0.2122
       Total |   152716168       934  163507.675   Root MSE        =    358.89

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   69.14725   6.122776    11.29   0.000     57.13116    81.16335
       exper |   16.01839   3.030074     5.29   0.000     10.07179    21.96498
       black |  -200.3573   36.03503    -5.56   0.000    -271.0769   -129.6377
       urban |   177.5962   26.25528     6.76   0.000     126.0696    229.1228
     married |   184.5965   38.31284     4.82   0.000     109.4067    259.7864
       hours |  -3.284626   1.641327    -2.00   0.046    -6.505769   -.0634829
       _cons |  -280.8906   129.3613    -2.17   0.030    -534.7651   -27.01605
------------------------------------------------------------------------------
```

   This means that our predicted model looks like the following:

   $$\widehat{\text{wage}}_i = \underset{(129.361)}{-280.891} + \underset{(6.123)}{69.147}\text{educ}_i + \underset{(3.030)}{16.018}\text{exper}_i - \underset{(36.035)}{200.357}\text{black}_i$$
   $$+ \underset{(26.255)}{177.596}\text{urban}_i + \underset{(38.313)}{184.597}\text{married}_i - \underset{(1.641)}{3.285}\text{hours}_i$$

   Side note: In academic journals and some other professional settings, regression estimates are also commonly reported in the following way:

	(1)
	wage
educ	69.15***
	(6.123)
exper	16.02***
	(3.030)
black	-200.4***
	(36.04)
urban	177.6***
	(26.26)
married	184.6***
	(38.31)
hours	-3.285*
	(1.641)
Constant	-280.9*
	(129.4)
Observations	935

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

An advantage of such table is that it reports not only all point estimates, but also the associated standard errors and information about p-value. If you're interested in producing such output, check out Stata's esttab command.

(To be very clear, this style of output is absolutely NOT required on your problem set. We're just pointing this out to you so that you're aware of such style of report.)

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

No. All p-value (recorded in the $P > |t|$ column in the regression Stata output) are less than .05, meaning that for any β_i, the null hypothesis $H_0 : \beta_i = 0$ can be rejected at significance level less than 5%.

If we can reject the null at significance level less than 5%, then we certainly can reject at 5% significance level. Thus, all slope coefficients are statistically significant.

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

Yes, OLS estimator is BLUE. This is because we run the default regression configuration in Stata, which assumes homoskedastic error. Under Gauss-Markov theorem, with error being homoskedastic and all four other assumptions satisfied, we have that OLS estimator is BLUE.

(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.

Heteroskedastic error states that $Var(u|\text{educ}, \text{exper}, \ldots, \text{hours}) \neq \sigma^2$ a constant, which is equivalent to say that $Var(\text{wage}|\text{educ}, \text{exper}, \ldots, \text{hours}) \neq \sigma^2$. If we want to argue that the error might be heteroskedastic, then we need to consider what makes wage have different level of variation at any specified level for all explanatory variables.

For example, holding all other variables constant, let's just think about the relationship between wage and exper. Although we control for all the other explanatory variables, we would still likely see small variation in wage for people with little experience, versus high variation in wage for people with many years of experience.

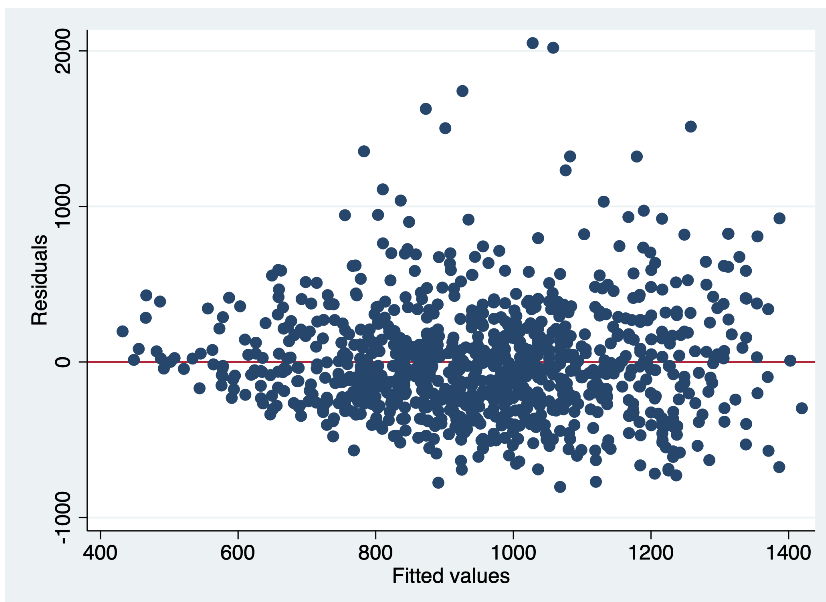
In a broader sense, we simply **cannot expect that the variation in wage is constant across all sorts of demographic groups**. This is why heteroskedastic error is the more appropriate way to model the error term in our linear regression model.

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

To perform a `rvfplot` visual test, make sure you run the regression first:

```
reg wage educ exper black urban married hours
rvfplot, yline(0)
```

The resulting residual-fitted value plot looks like the following:



With the plot having a "cone" shape (more specifically, variation of residuals seems small when fitted values are small, but the variation quickly becomes big when fitted values are big), we conclude that the error term follow a heteroskedastic pattern.

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

To perform a Breusch-Pagan test, make sure you already run the regrssion. Then in Stata, run

```
estat hettest, rhs fstat
```

The test output in Stata looks like the following:

```
. estat hettest, rhs fstat

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: educ exper black urban married hours

         F(6 , 928)    =      4.18
         Prob > F      =    0.0004
```

Since the p-value 0.0004 is less than 0.05, we can reject the null (null hypothesis here is homoskedastic error; alternative hypothesis is heteroskedastic error) and conclude that error term is heteroskedastic at 5% significance level.

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

To perform a White test, make sure you already run the regrssion. Then in Stata, run

`estat imtest, white`

The test output in Stata looks like the following:

```
. estat imtest, white

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

         chi2(24)     =     41.96
         Prob > chi2  =    0.0130

Cameron & Trivedi's decomposition of IM-test

-----------------------------------------------------
              Source |       chi2     df      p
---------------------+-------------------------------
  Heteroskedasticity |      41.96     24    0.0130
            Skewness |      25.68      6    0.0003
            Kurtosis |       6.80      1    0.0091
---------------------+-------------------------------
               Total |      74.44     31    0.0000
-----------------------------------------------------
```

Since the p-value 0.0130 is less than 0.05, we can reject the null (again, null hypothesis here is homoskedastic error; alternative hypothesis is heteroskedastic error) and conclude that error term is heteroskedastic at 5% significance level.

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

To reestimate the model in (a), attach the `robust` option to your `regress` command:

`reg wage educ exper black urban married hours, robust`

The output in Stata looks like the following:

```
. reg wage educ exper black urban married hours, robust

Linear regression                               Number of obs     =        935
                                                F(6, 928)         =      46.43
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2173
                                                Root MSE          =     358.89

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   69.14725    6.50851    10.62   0.000     56.37415    81.92036
       exper |   16.01839   3.042438     5.26   0.000     10.04753    21.98924
       black |  -200.3573   29.22863    -6.85   0.000    -257.7192   -142.9954
       urban |   177.5962   24.86559     7.14   0.000     128.7969    226.3955
     married |   184.5965   34.20315     5.40   0.000      117.472     251.721
       hours |  -3.284626   2.180172    -1.51   0.132    -7.563264    .9940122
       _cons |  -280.8906   146.7283    -1.91   0.056    -568.8484    7.067221
------------------------------------------------------------------------------
```

This means that our predicted model looks like the following:

$$\begin{aligned}\widehat{\text{wage}}_i = &\underset{(147.728)}{-280.891} + \underset{(6.509)}{69.147}\text{educ}_i + \underset{(3.042)}{16.018}\text{exper}_i - \underset{(29.229)}{200.357}\text{black}_i \\ &+ \underset{(24.866)}{177.596}\text{urban}_i + \underset{(34.203)}{184.597}\text{married}_i - \underset{(2.180)}{3.285}\text{hours}_i\end{aligned}$$

(i) Did any of the estimated coefficient change?

No, all the estimated coefficients remain the same. This is because changing the error structure only affects the standard error estimate of coefficients.

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

Yes. In general, robust standard errors (standard errors for beta estimates under heteroskedastic error) are bigger compared with standard errors in (a), but there are exceptions: specifically, robust standard error for coefficient on black is smaller than the original standard error.

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

No, OLS estimator is no longer BLUE under heteroskedasticity. Since robust standard errors are *generally* bigger, this tells us that the "best" (most efficient) aspect of BLUE is violated under heteroskedastic error.

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

Yes! Looking at the p-value for coefficient on hours. Now the p-value is 0.132, which is greater than 0.05. This tells us that we can no longer reject the null hypothesis that the true beta on hours is zero under 5% size, meaning that coefficient on hours is no longer statistically significant.

This change in interpretation highlights the importance of correctly specifying your model. When we incorrectly specified the error term to be homoskedastic, we thought coefficient on hours is statistically significant, meaning that it contributes to explaining wage. However, when we correctly adjust for the fact that the error term should actually be heteroskedastic, the statistic significance on coefficient for hours went away, telling us that hours actually don't contribute to explaining wage.