**PPOL502-01, Spring 2016**
**Course Notes # 18: Heteroskedasticity**

---

## I. HETEROSKEDASTICITY ("Hetero") DEFINITION AND DETECTION

- Remember the Gauss-Markov assumptions that, when all were met, indicate that OLS is BLUE? Well, we're about to throw another assumption out the window and see what happens.

- In particular, we're interested in assumption MLR.5:

  Var $(u \mid X_1, X_2, \ldots X_k) = \sigma^2$

  In other words, the variance of the unobservable error ($u$), conditional on the explanatory variables $X_1, X_2, \ldots X_k$, is the same for all values of the explanatory variables. (i.e., $\sigma^2$ is constant, or the same, for each value of the explanatory variables)

  This condition is known as the *homoskedasticity* assumption.

- The assumption of homoskedasticity fails when the **variance** of the unobservable/unmeasured factors **changes** across different levels of an explanatory variable of interest.

- *Heteroskedasticity* (or hetero for short) is present when homoskedasticity fails; i.e., heteroskedasticity indicates that the variances of the unobservables are different across different levels of an explanatory variable of interest.

  How about another way: what's left unexplained by the regression for a particular observation (i.e. the residual, $(Y_i - \hat{Y}_i)$ for each observation) is more clustered for some levels of a particular explanatory variable, and is more disperse for other levels of that explanatory variable (we'll look at an example in a bit).

- Key points about hetero:

  (1) The presence of heteroskedasticity does *not* affect the biasedness of the OLS estimators: OLS produces unbiased coefficient estimates, even if the errors are heteroskedastic (as long as assumptions MLR1 through MLR4 hold)

  (2) However, the coefficient estimates' standard errors, *t*-statistics, and *p*-values that Stata (or any computer program) spits out are no longer valid (even with large sample sizes).

- But don't worry! Many before you have thought of ways to address this problem. Wooldridge chapter 8 goes into great detail about some of these ways; even greater detail is available in other sources.

- For this course, the main goal is to (•) make you aware of the problem of hetero, (•) introduce some basic tests to detect presence of hetero, (•) describe a general solution to the problem of hetero, and (•) mention some alternative fixes.

Example: consumption and income (using data set SAVING.RAW, minus observation #100)

- Suppose you want to estimate the model:

  $CONS = \beta_0 + \beta_1 AGE + \beta_2 INC + \beta_3 SIZE + \eta$

where

```
1. cons                    annual consumption, $ (1970)
2. age                     age of household head
3. inc                     annual income, $  (1970)
4. size                    family size
```

- You estimate the model as usual in Stata:

Dependent Variable: cons
[first part of output omitted]

```
------------------------------------------------------------------------------
        cons |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   4.958145   30.93474     0.16   0.873    -56.45508    66.37137
         inc |   .8733883   .0402623    21.69   0.000     .7934574    .9533191
        size |   110.0096    149.668     0.74   0.464     -187.119    407.1381
       _cons |  -755.4562   1471.259    -0.51   0.609    -3676.275    2165.363
------------------------------------------------------------------------------
```

- Interpret the coefficient on INC:

- As you can see, everything *looks* fine.

- However, this is an example of a model where we may be worried about heteroskedasticity: conditional on the household head's age and the size of the family, it is likely that the *residual* $(Y_i - \hat{Y}_i)$ varies much more for higher levels of income than for lower levels of income: because consumption itself may vary much more for higher levels of income, our predictions of that consumption may be less precise than at lower levels.

- Income and consumption models (or income and savings models) are well-known for exhibiting heteroskedasticity. In your own work, as you read literature reviews and other material for your topic, you may find that researchers often correct for heteroskedasticity. This is a good indication that you should too: you can't ignore it.

- So, are you just left to theory or prior knowledge to understand whether hetero is a problem? Not entirely. We'll start out with some simple diagnostics:

- Diagnostic #1: A simple plot of the residuals against the offending variable(s)

    - Save the residuals after estimating the model:

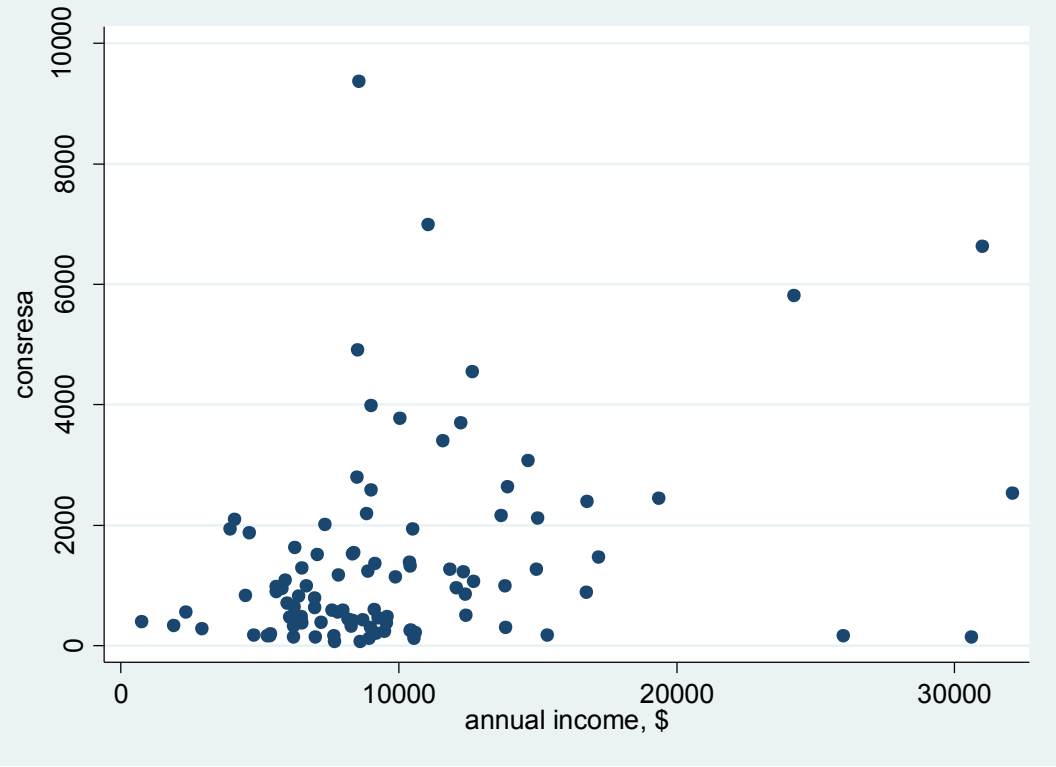
```
reg cons age inc size
predict consres , residuals

gen consresa  = abs(consres)
gen consressq = (consres*consres)
```

- You can either plot the absolute values of the residuals against the independent variable that you suspect is causing the heteroskedasticity; or you can plot the actual values of the residuals.

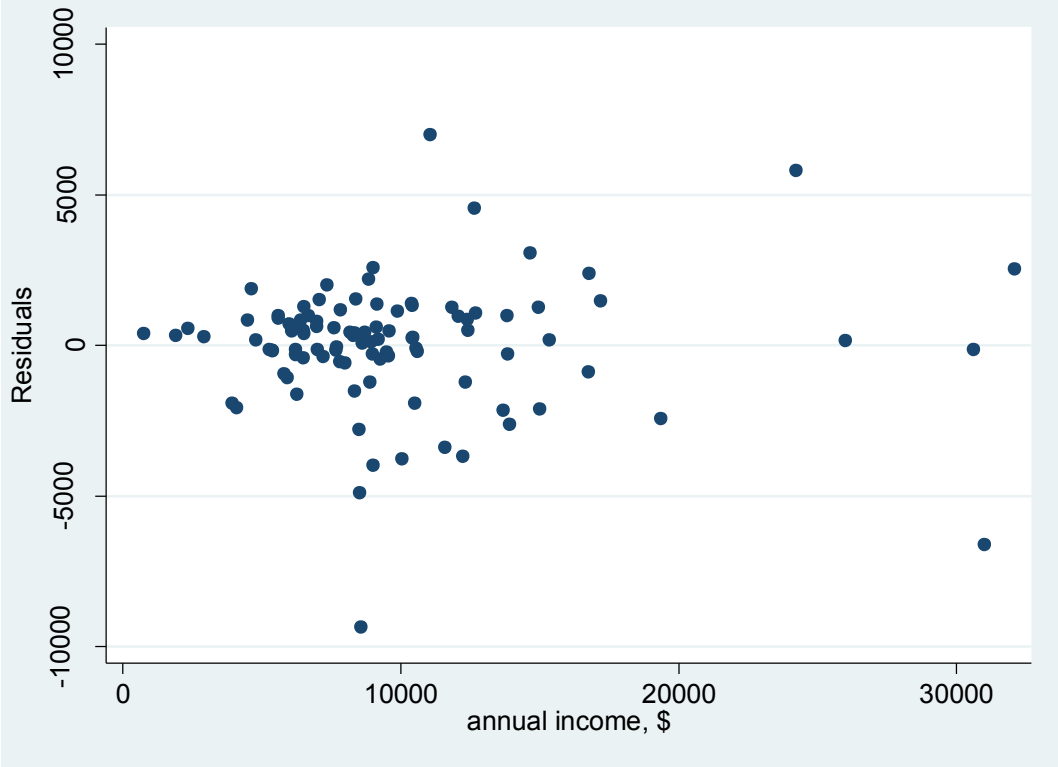
- If you plot the absolute values, look for an "opening up" in one end of the X-variable distribution.

```
twoway scatter consresa inc
```




- If you plot the actual values, now the baseline is zero, so look for a fanning out in both directions:

```
twoway scatter consres inc
```




- Q: what do you think a plot of the errors versus INCOME would look like if the homoskedasticity assumption held?

- You can plot the residuals against multiple independent variables separately (e.g., CONSRES*AGE, CONSRES*SIZE in this case), and examine each relationship for evidence of hetero.

  With lots of explanatory variables, this can become tedious and you want to save your energy for the variables that you think are really driving the hetero.

- The visual plot is not something that Wooldridge discusses, but it is a very simple, accepted way of getting a sense of whether hetero is a problem. As we know, though, our eyes often fail us and we need a more precise measure of whether hetero is a problem.

- The next two tests are intended to do this. There are many other tests for hetero that are used (the White test, which Wooldridge discusses; the Park test; the Glejser test; Loess plots, etc.). These various tests have advantages and drawbacks. Usually, you will run multiple tests and if you reach the same conclusion from all of them, you know you do or don't have a hetero problem.

- Diagnostic #2: The F-test from a regression of the squared residuals on the *X*s.

-- This test regresses the SQUARE of the residuals on all the *X*s in the original model. The idea is that if the assumption Var $(\mu \mid X_1, X_2, \ldots X_k) = \sigma^2$ is met, then the *X*s will not explain the variation in the squared residuals:

$$\hat{u}^2 = \delta_0 + \delta_1 \text{AGE} + \delta_2 \text{ INC} + \delta_3 \text{SIZE} + e$$

```
. reg consressq age inc size

      Source |       SS       df       MS              Number of obs =      99
-------------+------------------------------           F(  3,    95) =    3.07
       Model |  1.1940e+15     3  3.9802e+14           Prob > F      =  0.0317
    Residual |  1.2333e+16    95  1.2982e+14           R-squared     =  0.0883
-------------+------------------------------           Adj R-squared =  0.0595
       Total |  1.3527e+16    98  1.3803e+14           Root MSE      =  1.1e+07

------------------------------------------------------------------------------
   consressq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   204817.3   163414.3     1.25   0.213      -119601    529235.7
         inc |   444.4265   212.6878     2.09   0.039     22.18789    866.6652
        size |  -570481.2   790628.6    -0.72   0.472     -2140077      999115
       _cons |   -5421708    7771999    -0.70   0.487    -2.09e+07    1.00e+07
------------------------------------------------------------------------------
```

- This F-stat is statistically significant at the 0.04 level, indicating that together, *the Xs explain a statistically significant amount of the variation in the squared residuals from the initial regression.* This suggests that *heteroskedasticity* is present.

- Instead of including all the Xs in this auxiliary regression, you may just include the *X* variables that you suspect are causing the hetero (here, the INCOME variable):

```
. reg consressq inc

      Source |       SS       df       MS              Number of obs =      99
-------------+------------------------------           F(  1,    97) =    6.64
       Model |  8.6609e+14     1  8.6609e+14           Prob > F      =  0.0115
    Residual |  1.2661e+16    97  1.3052e+14           R-squared     =  0.0640
-------------+------------------------------           Adj R-squared =  0.0544
       Total |  1.3527e+16    98  1.3803e+14           Root MSE      =  1.1e+07

------------------------------------------------------------------------------
   consressq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   530.1882   205.8235     2.58   0.012     121.6856    938.6909
       _cons |  -793648.7    2341932    -0.34   0.735     -5441736     3854439
------------------------------------------------------------------------------
```

- <u>Diagnostic #3:</u> The Breusch-Pagan Test

  - This test uses the exact same model as the F-test for hetero, but uses a different test statistic: the Lagrange Multiplier (or LM) statistic, which has a $\chi^2_k$ distribution, where *k* is the number of *X* variables in the regression that explains the squared residuals (*not* the number of *X*s in the original equation).

- The LM test is calculated as: $LM = n * R^2_{\hat{u}^2}$
- So in this case (using the first auxiliary regression), LM = 99*0.0883 = 8.74

- $\chi^2_3$ at the 10% sig level = 6.25
  at the 5% sig level = 7.81
  at the 1% sig level = 11.34

- So we know the p-value for this statistic is $0.01 < p < 0.05$.

- In this case, the visual plots as well as both of these diagnostic statistics indicate that heteroskedasticity may be a problem in this model. Therefore, while the coefficient estimates in the model on p. 3 are unbiased (assuming MLR1 through MLR4 hold), we cannot make inferences about their stat sig based on the standard errors that Stata calculated: those standard errors are incorrect, because hetero is present.

- We need to "fix" the standard errors

## II. HETEROSKEDASTICITY CORRECTIONS: THE GENERAL FIX

- pp. 265--271of Wooldridge discuss a general correction for heteroskedasticity when "the form of heteroskedasticity is unknown" [in a little bit, we'll talk briefly about what it means to "know the form of hetero" and what you can do about it]. It is most often the case that you will *not* know, or will not be able to specify easily, the form of hetero. Therefore, this fix is often a standard, safe thing to implement (provided sufficiently large sample sizes).

- These standard errors corrected for heteroskedasticity are known by a variety of names: robust, Eicker-White, Huber-White, White. All these names indicate that this general correction has been implemented.

- The robust standard error is the square root of the coefficient's variance, calculated as follows:

$$\hat{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}, \text{ where:}$$

$\hat{u}_i$ = residual *i* from the original model of Y on the Xs

$\hat{r}_{ij}$ = residual *i* obtained by regressing the particular $x_j$ on all *other* Xs in the original model

$SSR_j$ = sum of squared residuals from the regression of a particular $x_j$ on all the other Xs in the original model

*NOTE:* Some versions of this statistic include a correction for *n/(n-k)* or *n/(n-k-1)* or other corrections. The precise formulas used by different statistical packages may vary somewhat, depending on what version of the statistic they use.

- Once you obtain the correct standard errors, you can calculate a *t*-statistic as usual using these new, improved standard errors, and carry on with inference.

- Using Stata, it is very easy to implement the general correction for heteroskedasticity: you just include the option "robust" in the code in this way:

```
Estimates with robust standard errors

. reg cons age inc size , robust

Linear regression                                      Number of obs =     100
                                                       F(  3,    96) =   73.24
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.6931
                                                       Root MSE      =  3223.5

------------------------------------------------------------------------------
             |               Robust
        cons |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   24.29426   39.48549     0.62   0.540    -54.08381    102.6723
         inc |   .8436222   .0673404    12.53   0.000     .7099526    .9772917
        size |  -59.47328   212.6565    -0.28   0.780     -481.593    362.6464
       _cons |  -711.1004   1247.055    -0.57   0.570    -3186.485    1764.284
------------------------------------------------------------------------------

Estimates with non-robust standard errors from above
------------------------------------------------------------------------------
        cons |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   24.29426   46.15682     0.53   0.600     -67.3263    115.9148
         inc |   .8436222   .0600349    14.05   0.000     .7244538    .9627905
        size |  -59.47328   222.4773    -0.27   0.790    -501.0873    382.1407
       _cons |  -711.1004   2198.814    -0.32   0.747    -5075.712    3653.511
------------------------------------------------------------------------------
```

- Robust standard errors are sometimes larger, and sometimes smaller, than the usual standard errors calculated by OLS. [Most often, they are larger]

- Even though the standard error is larger in this particular case, the inference we make will not change – the coefficient estimate on INC is still stat sig.

- Above, we considered a general fix for heteroskedasticity problems – the Eicker-Huber-White robust standard errors. So, whenever you read or hear that someone used robust standard errors, this fix is what they're talking about.

- Should you *always* calculate and use the robust standard errors, just to be safe? The short answer is that you should probably use them when you have large samples (greater than n=150 or so), but *should not use them with smaller samples*. This is because the robust standard errors were developed using asymptotic (i.e., large-sample) properties; in small samples they can be very inefficient. SO… robust s.e.'s SHOULD NOT have been used in the previous example (the sample size was only 100).

- One strategy is to either report both standard errors and their resulting hypothesis tests, or (this is preferred), to *use robust standard errors when you're dealing with large samples*. However, also calculate the usual standard errors, and report in a

footnote to a table, or in appendix whether your inferences are sensitive to the type of standard error you use: would your conclusions about stat sig be different if you didn't use the robust s.e.'s? Some researchers report the ratio of robust to ordinary standard errors, which allows a reader to determine whether inferences made using ordinary s.e's are sensitive to potential hetero correction.

**ANOTHER EXAMPLE:**

```
. regress lwage IQ exper expersq educ KWW momdad14 sinmom14 step14 libcrd14

      Source |       SS       df       MS              Number of obs =    2034
-------------+------------------------------           F(  9,  2024) =   55.01
       Model |  69.7397877     9   7.7488653           Prob > F      =  0.0000
    Residual |  285.121686  2024  .140870398           R-squared     =  0.1965
-------------+------------------------------           Adj R-squared =  0.1930
       Total |  354.861474  2033  .174550651           Root MSE      =  .37533

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          IQ |   .0030832   .0006817     4.52   0.000     .0017463    .0044201
       exper |    .083202   .0100787     8.26   0.000     .0634363    .1029676
     expersq |  -.0024638   .0005065    -4.86   0.000    -.0034571   -.0014704
        educ |   .0561251   .0055652    10.09   0.000     .0452111    .0670392
         KWW |   .0075454   .0014433     5.23   0.000     .0047149    .0103759
    momdad14 |   .0718106   .0389176     1.85   0.065    -.0045121    .1481333
    sinmom14 |   .0202218    .047688     0.42   0.672    -.0733008    .1137444
      step14 |   .0057517   .0573323     0.10   0.920    -.1066847    .1181881
    libcrd14 |   .0386801   .0204674     1.89   0.059    -.0014592    .0788193
       _cons |   4.384855   .1071474    40.92   0.000     4.174725    4.594986
------------------------------------------------------------------------------

. predict lwageres, residuals
. gen lwagesq=(lwageres*lwageres)

. regress lwagesq IQ exper expersq educ KWW momdad14 sinmom14 step14 libcrd14

      Source |       SS       df       MS              Number of obs =    2034
-------------+------------------------------           F(  9,  2024) =    0.90
       Model |  .418516484     9  .046501832           Prob > F      =  0.5241
    Residual |  104.547149  2024   .05165373           R-squared     =  0.0040
-------------+------------------------------           Adj R-squared = -0.0004
       Total |  104.965666  2033  .051630923           Root MSE      =  .22727

------------------------------------------------------------------------------
     lwagesq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          IQ |  -.0007643   .0004128    -1.85   0.064    -.0015739    .0000452
       exper |   -.000938    .006103    -0.15   0.878    -.0129068    .0110308
     expersq |   7.44e-06   .0003067     0.02   0.981    -.0005941    .0006089
        educ |   .0032478   .0033699     0.96   0.335     -.003361    .0098567
         KWW |  -.0005078    .000874    -0.58   0.561    -.0022217    .0012062
    momdad14 |   .0301473    .023566     1.28   0.201    -.0160689    .0763636
    sinmom14 |   .0182936   .0288768     0.63   0.526    -.0383379     .074925
      step14 |   .0243046   .0347168     0.70   0.484    -.0437799     .092389
    libcrd14 |    .012931   .0123937     1.04   0.297    -.0113748    .0372369
       _cons |   .1613479   .0648818     2.49   0.013     .0341059      .28859
------------------------------------------------------------------------------
```

```
. regress lwage IQ exper expersq educ KWW momdad14 sinmom14 step14 libcrd14

      Source |       SS       df       MS              Number of obs =    2034
-------------+------------------------------           F(  9,  2024) =   55.01
       Model |  69.7397877     9   7.7488653           Prob > F      =  0.0000
    Residual |  285.121686  2024  .140870398           R-squared     =  0.1965
-------------+------------------------------           Adj R-squared =  0.1930
       Total |  354.861474  2033  .174550651           Root MSE      =  .37533

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          IQ |   .0030832   .0006817     4.52   0.000     .0017463    .0044201
       exper |    .083202   .0100787     8.26   0.000     .0634363    .1029676
     expersq |  -.0024638   .0005065    -4.86   0.000    -.0034571   -.0014704
        educ |   .0561251   .0055652    10.09   0.000     .0452111    .0670392
         KWW |   .0075454   .0014433     5.23   0.000     .0047149    .0103759
    momdad14 |   .0718106   .0389176     1.85   0.065    -.0045121    .1481333
    sinmom14 |   .0202218    .047688     0.42   0.672    -.0733008    .1137444
      step14 |   .0057517   .0573323     0.10   0.920    -.1066847    .1181881
    libcrd14 |   .0386801   .0204674     1.89   0.059    -.0014592    .0788193
       _cons |   4.384855   .1071474    40.92   0.000     4.174725    4.594986


. regress lwage IQ exper expersq educ KWW momdad14 sinmom14 step14 libcrd14, robust

Linear regression                               Number of obs =    2034
                                                F(  9,  2024) =   54.91
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1965
                                                Root MSE      =  .37533

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          IQ |   .0030832   .0007361     4.19   0.000     .0016396    .0045268
       exper |    .083202   .0097367     8.55   0.000     .0641069     .102297
     expersq |  -.0024638   .0004816    -5.12   0.000    -.0034083   -.0015192
        educ |   .0561251   .0057837     9.70   0.000     .0447825    .0674678
         KWW |   .0075454   .0014732     5.12   0.000     .0046563    .0104346
    momdad14 |   .0718106   .0361468     1.99   0.047     .0009217    .1426994
    sinmom14 |   .0202218   .0447933     0.45   0.652    -.0676241    .1080676
      step14 |   .0057517   .0553293     0.10   0.917    -.1027567      .11426
    libcrd14 |   .0386801   .0200914     1.93   0.054    -.0007219     .078082
       _cons |   4.384855   .1068562    41.04   0.000     4.175296    4.594415
------------------------------------------------------------------------------
```

## III. HETEROSKEDASTICITY CORRECTIONS: MORE COMPLICATED VERSIONS

- The robust standard error we used above can be used when "the form of the heteroskedasticity is unknown." O.K., so what does it mean for the form to be *known*?

  $$\text{Var}(\mu \mid X_1, X_2, \ldots X_k) = \sigma^2 * h(\boldsymbol{X})$$

  That is, the conditional variance is some known function of the explanatory variables, and this is known up to some constant (here, $h(\boldsymbol{X})$).

- If this is known, then the original equation can be transformed to correct in advance (i.e., prior to estimation) for this known heteroskedasticity: Example:

  Original equation: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_k x_{ik} + u_i$

  Transformed equation:

  $$\frac{y_i}{\sqrt{h_i(\mathbf{X})}} = \frac{\beta_0}{\sqrt{h_i(\mathbf{X})}} + \beta_1\left(\frac{x_{ii1}}{\sqrt{h_i(\mathbf{X})}}\right) + \beta_2\left(\frac{x_{ii2}}{\sqrt{h_i(\mathbf{X})}}\right) + \ldots\beta_k\left(\frac{x_{iik}}{\sqrt{h_i(\mathbf{X})}}\right) + \frac{u_i}{\sqrt{h_i(\mathbf{X})}}$$

  Then, this new model is estimated by OLS as usual.

  Because the variables are transformed by (i.e., weighted by) by the $h(\boldsymbol{X})$ function, this method is known as *weighted least squares* (or WLS).

- WLS is one example of a more general method of estimation called *generalized least squares* (GLS). GLS methods can be used to solve a variety of problems (some of which you'll hear more about next semester). In the case of hetero, these estimators are more efficient (i.e. lower variance) than OLS in the presence of heteroskedasticity.