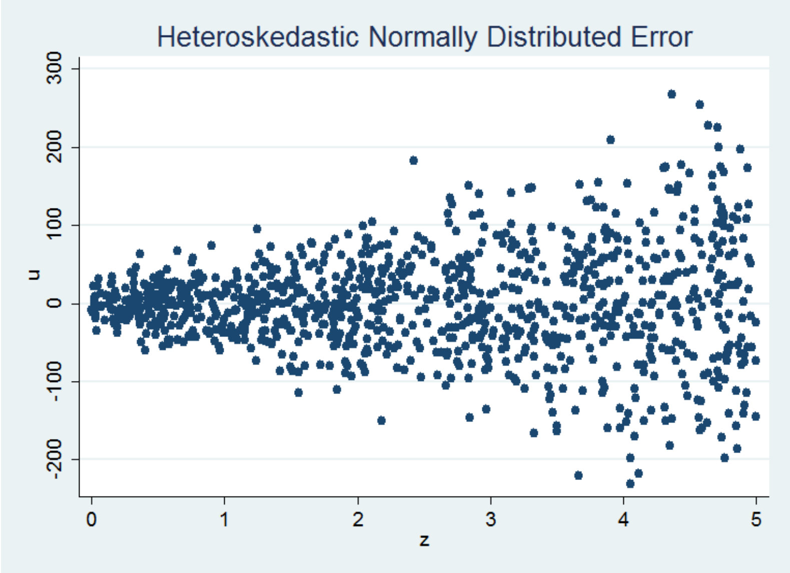
Heteroskedasticity & Robust Standard Errors

- *Recap I: Those SLR's, MLR's, LUEs and BLUE*
- *… II: MSE's, RMSE's, and Standard Errors*
- *… III: Homoskedasticity v. Heteroskedasticity*
- *Heteroskedasticity: Learning from the Sample Mean Estimator*
- *Bogus Standard Errors & Weighted Least Squares*
- *Robust standard errors to the rescue!*
- *… , robust in action! An example*



Recap I: *Those SLR's, MLR's, LUEs and BLUE*

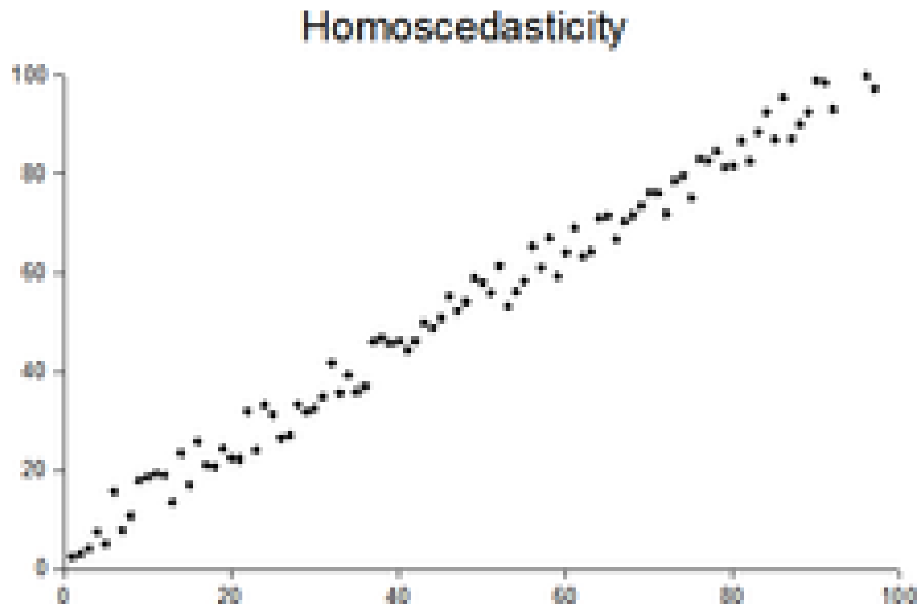
	SLR	MLR
Those SLR/MLR Assumptions	**SLR.1: Linear Model** $Y = \beta_0 + \beta_1 X + U$	**MLR.1: Linear Model** $Y = \beta_0 + \beta_x X_i + \beta_z Z_i + \ldots + U_i$
	SLR.2: Random sampling	**MLR.2**: Random sampling
	SLR.3: Sampling variation of the x's	**MLR.3**: No perfect collinearity
	SLR.4: Zero Conditional Mean $E(U \mid X = x) = 0$ for any x	**MLR.4**: Zero Conditional Mean $E(U \mid x, z, \ldots) = 0$ for any x, z, …
If all 4 hold:	**LUE: OLS is a LUE**	**LUE: OLS is an LUE**
Add one more assumption	**SLR.5**: Homoskedastic errors (conditional on x)	**MLR.5**: Homoskedastic errors (conditional on the x's, z's, …)
Gauss-Markov Thm	**SLR.1-SLR.5 $\rightarrow$ BLUE**	**MLR.1-MLR.5 $\rightarrow$ BLUE**

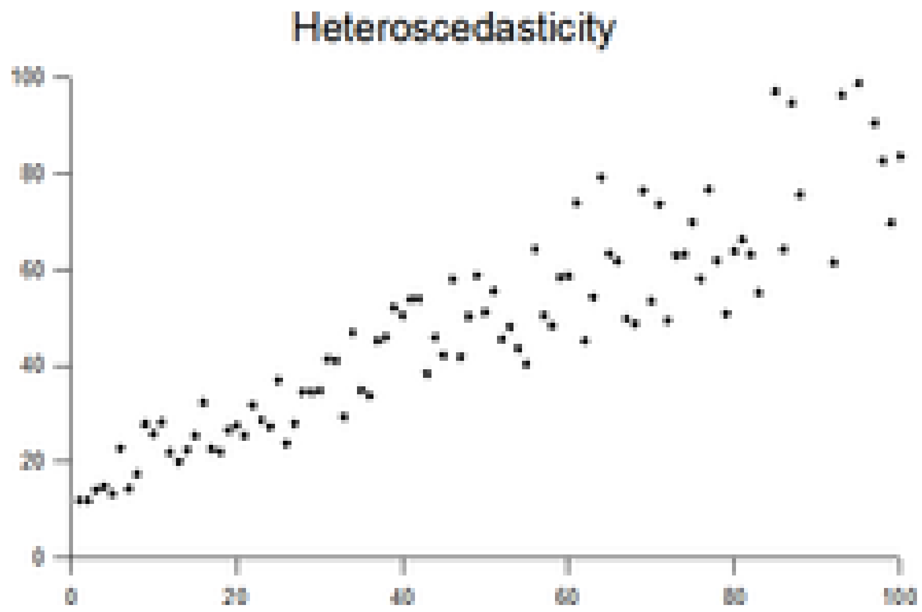
Recap II: *MSE's, RMSE's, and Standard Errors*

	SLR	**MLR**
Assume	SLR.1 – SLR.5	MLR.1 – MLR.5
Variance of the OLS slope estimator	$Var(B_1 \mid x's) = \frac{\sigma^2}{(n-1)S_{xx}}$	$Var(B_x \mid x, z\ldots) = \frac{\sigma^2}{(n-1)S_{xx}} VIF_x$
Standard Deviation of B_1	$sd(B_1) = \frac{\sigma}{S_x\sqrt{n-1}}$	$sd(B_x) = \frac{\sigma}{S_x\sqrt{n-1}}\sqrt{VIF_x}$
MSE unbiased estimator of σ^2	$MSE = \frac{SSR}{n-2} = \hat{\sigma}^2,\ E(MSE) = \sigma^2$	$MSE = \frac{SSR}{n-k-1} = \hat{\sigma}^2,\ E(MSE) = \sigma^2$
Standard Error of B_1 (estimator of $sd(B_1)$)	$se(B_1) = \frac{RMSE}{S_x\sqrt{n-1}}$	$se(B_x) = \frac{RMSE}{S_x\sqrt{n-1}}\sqrt{VIF_x}$
Add one last Assumption	SLR.6: $U_i \sim N(0, \sigma^2)$ and indept. of X	MLR.6: $U_i \sim N(0, \sigma^2)$ and indept. of the RHS variables
Distribution of *t statistic*	$\frac{B_1 - \beta_1}{se(B_1)} \sim t_{n-2}$	$\frac{B_x - \beta_x}{se(B_x)} \sim t_{n-k-1}$

Recap III: *Homoskedasticity v. Heteroskedasticity*

- **SLR.5: *Homoskedasticity*** (constant conditional variance of the error term, U)
  - To derive the variances of the estimators, we make one additional assumption:
    - SLR.5: $Var(U \mid X = x) = \sigma^2$ for all x
  - Note that SLR.5 holds if U is independent of X, so that $Var(U \mid X = x) = Var(U) = \sigma^2$.
  - *Heteroskedasticity*: the conditional variances are not all the same.





Heteroskedasticity: *Learning from the Sample Mean Estimator*

- Under homoskedasticity:
  - Linear Unbiased Estimator: $W = \beta_1 Y_1 + \beta_2 Y_2 + \ldots + \beta_n Y_n$, where $\sum_{i=1}^{n} \beta_i = 1$
  - BLUE: min $Var(W) = \sum \beta_i^{\,2} \sigma^2 = \sigma^2 \sum \beta_i^{\,2}$ subject to $\sum_{i=1}^{n} \beta_i = 1$ … which yields
  - The Sample Mean Estimator is BLUE: $W = \frac{1}{n} \sum Y_i = \bar{Y}$

- But given heteroskedasticity, we now have:
  - BLUE: min $Var(W) = \sum \beta_i^{\,2} \sigma_i^2$ subject to $\sum_{i=1}^{n} \beta_i = 1$, which yields

$$\frac{\partial Var(W)}{\partial \beta_i} = \frac{\partial Var(W)}{\partial \beta_j} \Rightarrow 2\sigma_i^{\,2}\beta_i = 2\sigma_j^{\,2}\beta_j \text{ and } \frac{\beta_i}{\beta_j} = \frac{\sigma_j^{\,2}}{\sigma_i^{\,2}}, \text{ or } \beta_i = \frac{(1/\sigma_i^{\,2})}{\sum_j 1/\sigma_j^{\,2}}$$

  - And so the BLUE will be a weighted average of the sampled values, where the weights are proportional to the inverse of the respective variances.

OLS and Heteroskedasticity:
Bogus Standard Errors & Weighted Least Squares

Given heteroskedasticity:

- **BLUE**: Now min $Var\left[\sum b_i Y_i\right] = \sum b_i^2 \sigma_i^2$ subject to $\sum b_i = 0$ and $\sum b_i x_i = 1$
  - Weighted Least Squares (WLS) is now **BLUE**

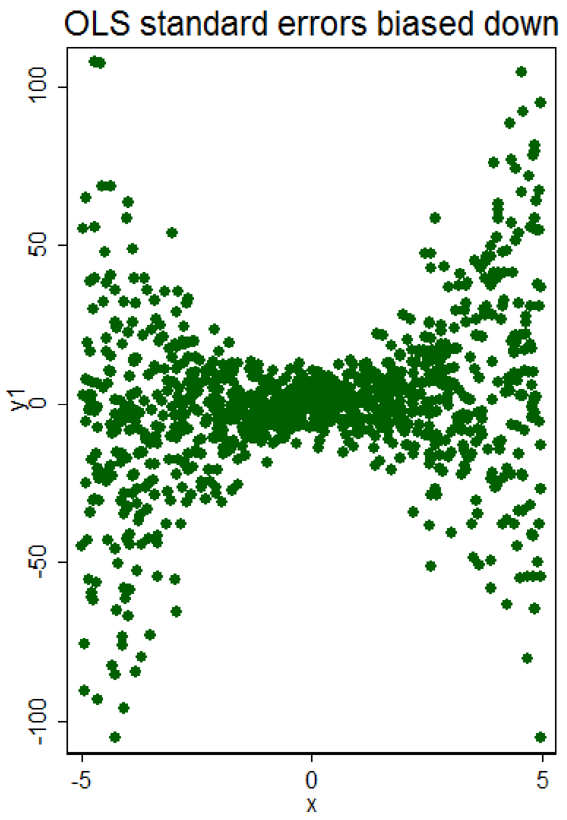
$$min\ wgtSSR = \sum \frac{1}{\sigma_i^2}(y_i - \beta_0 - \beta_1 x_i)^2$$

- Now, $E(MSE) = E\left(\frac{SSR}{n-2}\right) \neq \sigma^2$ and $Var(B_1) \neq \frac{\sigma^2}{\sum (x_i - \overline{x})^2}$

  (the σ_i^2 are no longer constant across observations… SLR.5 is violated!)

- The reported OLS standard errors are no longer relevant… ditto the t stats, p values and Confidence Intervals! ***Say goodbye to inference! … or maybe not?***

What's a researcher to do?
Robust standard errors to the rescue!



- **Don't despair!**
- Run weighted least squares if you can… and if you can't weight by $1/\sigma_i^2$ (who knows the σ_i^2's anyway?), maybe you can use a proxy.
- ***Robust standard errors***: Add **, robust** to your regression command and generate ***robust*** (*corrected*) standard errors… and as well, *corrected* t stats, p values and Confidence Intervals. *And move on!*
- It's OK to add ***, robust*** even if you may not have an issue with Heteroskedasticity. There's no harm in doing this… and besides, you can brag about having ***robust*** standard errors!

, robust in action!: An example

```
. reg Brozek hgt wgt abd

      Source |       SS           df       MS      Number of obs   =       252
-------------+----------------------------------   F(3, 248)       =    213.67
       Model |  10872.5504         3  3624.18347   Prob > F        =    0.0000
    Residual |  4206.46623       248  16.9615574   R-squared       =    0.7210
-------------+----------------------------------   Adj R-squared   =    0.7177
       Total |  15079.0166       251  60.0757635   Root MSE        =    4.1184

------------------------------------------------------------------------------
      Brozek |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hgt |  -.1181607   .0824192    -1.43   0.153    -.2804915    .0441701
         wgt |   -.120415   .0222516    -5.41   0.000    -.1642411   -.0765888
         abd |    .879846   .0579164    15.19   0.000     .7657751    .9939168
       _cons |  -32.66247    6.51936    -5.01   0.000    -45.50285    -19.8221
------------------------------------------------------------------------------

. reg Brozek hgt wgt abd, robust

Linear regression                               Number of obs     =        252
                                                F(3, 248)         =     183.63
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7210
                                                Root MSE          =     4.1184

------------------------------------------------------------------------------
             |               Robust
      Brozek |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hgt |  -.1181607   .0594908    -1.99   0.048    -.2353323   -.0009892
         wgt |   -.120415   .0249741    -4.82   0.000    -.1696034   -.0712265
         abd |    .879846   .0552738    15.92   0.000       .77098    .9887119
       _cons |  -32.66247   4.817425    -6.78   0.000    -42.15076   -23.17419
------------------------------------------------------------------------------
```