SLR Models: *Estimation*

- ***Those OLS Estimates***
- ***Estimators (ex ante) v. estimates (ex post)***
- ***The Simple Linear Regression (SLR) Conditions SLR.1-SLR.4***
- ***An Aside: The Population Regression Function (PRF)***
- ***B_0 and B_1 are Linear Estimators (conditional on the x's)***
- ***OLS estimators are unbiased! (under SLR.1-SLR.4)***
- ***… but B_1 is not alone***
- ***OLS estimators have a variance***
- ***SLR.5 – Homoskedasticity***
- **Variance of the OLS Estimators (assuming SLR.1-SLR.5)**
- ***MSE/RMSE (Goodness-of-Fit) and Standard Errors***
- ***OLS estimators are BLUE! (under SLR.1-SLR.5)***

SLR.1: Linear (DGM) Model
SLR.2: Random Sample
SLR.3: Sample variation in the RHS variable
SLR.4: U has zero mean | RHS variable

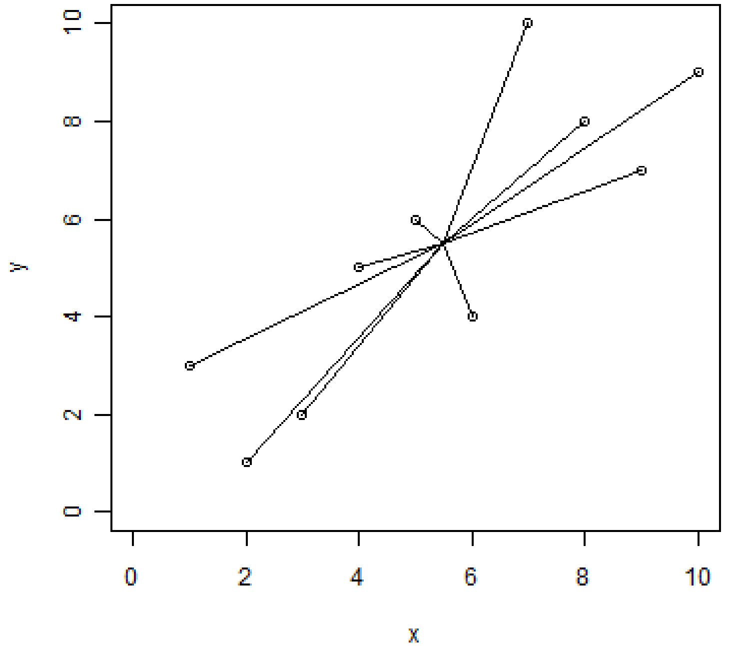
SLR.5: Homoskedasticity | RHS variable

SLR Models Estimation: *Those OLS estimates*

- Your data: (x, y): $\{x_i, y_i\}$ $\; i = 1, 2, \dots n$.
- You fit a straight line to the data: $y_i \sim \beta_0 + \beta_1 x_i$
  - OLS: estimate β_0 (intercept parameter) and β_1 (slope parameter) found by
    min $SSR = \sum \left(y_i - \left(b_0 + b_1 x_i \right) \right)^2$ wrt b_0 and b_1.
- OLS estimates (for your dataset):
  - ***Slope***: $\hat{\beta}_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

    $$= \sum \frac{(x_i - \bar{x})^2}{(n-1)S_{xx}} \left(\frac{y_i - \bar{y}}{x_i - \bar{x}} \right) = \sum w_i \left(\frac{y_i - \bar{y}}{x_i - \bar{x}} \right),$$

    where $w_i = \dfrac{(x_i - \bar{x})^2}{(n-1)S_{xx}}$ and $\sum w_i = 1 \dots$
  - ***Intercept***: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.



ex Post estimates v. ex Ante estimators

- **Estimates: *exPost*** *(actual; after the event)*:
  - Numbers driven by the specific sample
  - ***Slope estim<u>ate</u>***: $\hat{\beta}_1 = \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$, and
  - ***Intercept estim<u>ate</u>***: $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$.
- **Estimators: *exAnte*** *(before the event)*
  - Random variables… will take on different values depending on the actual sample
  - ***Slope estim<u>ator</u>***: $B_1 = \dfrac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_j - \overline{X})^2} = \dfrac{\sum (X_i - \overline{X}) Y_i}{\sum (X_j - \overline{X})^2}$
  - ***Intercept estim<u>ator</u>***: $B_0 = \overline{Y} - B_1 \overline{X}$

SLR Models Estimation: *Let’s review notation!*

- **Random variables** (upper case letters): X's and Y's
- **Data** (lower case letters): x's and y's
- **True parameters** (to be estimated): $\beta_0 \; and \; \beta_1$
- **Parameter estima<u>tors</u>** (random variables; upper case letters): $B_0 \; and \; B_1$
- **Parameter estima<u>tes</u>** (estimated coefficients; (true) parameter estimates; denoted with *hats*): $\hat{\beta}_0 \; and \; \hat{\beta}_1$

Those SLR Conditions: SLR.1-SLR.4

- ***SLR.1 – Linear model (DGM)***: $Y_i = \beta_0 + \beta_1 X_i + U_i \ldots \; i = 1, \ldots, n$
    - $X's, Y's$ and $U's$ are random variables
    - β_0 *and* β_1 are (true) parameters to be estimated.
    - ***DGM***: *Data Generation Mechanism*
- **SLR.2 – *Random sampling***: the sample $\{(x_i, y_i)\}$ is a random sample
- **SLR.3 – *Sample variation in the independent variable***: the $x_i's$ are not identical
- ***SLR.4 – U has zero conditional mean:*** $E(U \mid X = x) = 0$ for all x. This implies:
    - $E(U) = 0$ (U has mean zero)
    - $Cov(X, U) = 0$ (X and U are uncorrelated)

SLR.1: Linear (DGM) Model
SLR.2: Random Sample
SLR.3: Sample variation in the RHS variable
SLR.4: U has zero mean | RHS variable

PRFs and Linear Estimators

- ***Population Regression Function (PRF)***: $E(Y \mid X = x)$
  - PRF: The conditional means of the dependent variable Y (conditional on the x's)
  - $E(Y \mid X = x) = \beta_0 + \beta_1 x$ given SLR.1 and SLR.4
- **B_0 and B_1 are Linear Estimators** (conditional on the x's)
  - B_1 is linear in the Y_i's (conditional on the x's):

    $$B_1 = \sum b_i Y_i \text{ , where } b_i = \frac{(x_i - \overline{x})}{\sum (x_j - \overline{x})^2} = \frac{(x_i - \overline{x})}{(n-1)S_{xx}}$$

  - Note that *conditional on the x's* means that we are taking the x values as given, and not as random variables with values to be determined.
  - B_0 is also linear in the Y_i's (conditional on the x's):

    $$B_0 = \sum \frac{1}{n} Y_i - \overline{x} \sum b_i Y_i = \sum \left[\frac{1}{n} - b_i \overline{x} \right] Y_i$$

OLS Estimators are *Unbiased!* *Who saw this coming?*

- Recall ***SLR.1***: $Y = \beta_0 + \beta_1 X + U$ (β_0 and β_1 to be estimated).
- Recall the OLS slope and intercept estimators (conditional on the x's)
  - OLS slope estimator: $B_1 = \sum w_i \dfrac{(Y_i - \overline{Y})}{(x_i - \overline{x})}$

$$\text{OLS} \triangleq \text{LUE}$$

  where $w_i = \dfrac{(x_i - \overline{x})^2}{(n-1)S_{xx}}$ are non-negative weights that sum to 1, $\sum w_i = 1$

  - OLS intercept estimator: $B_0 = \overline{Y} - B_1 \overline{x}$
- **Given SLR.1-SLR.4: *B_0* and *B_1* are unbiased estimators!**
  - **B_1 is unbiased! -** $E\left(B_1 \mid x's\right) = \beta_1$ all x's for all x's implies $E\left(B_1\right) = \beta_1$
  - **B_0 is also unbiased! -** $E\left(B_0 \mid x's\right) = \beta_0$ all x's for all x's implies $E\left(B_0\right) = \beta_0$
- So: **OLS = LUE!**

But B_1 is not Alone!

- Given SLR.1-SLR.4: There are an ***infinite*** number of linear unbiased slope estimators.
- Any weighted average of the slopes of the lines connecting the data points to the samples means will also be a LUE (conditional on the x's) of the slope parameter:
  - Here's a LUE: $\sum \alpha_i \left(\frac{Y_i - \bar{Y}}{X_i - \bar{X}} \right)$, where $\sum \alpha_i = 1$.
  - Then conditional on the x's:

    $$E\left(\sum \alpha_i \left(\frac{Y_i - \bar{Y}}{x_i - \bar{x}} \right) \right) = \sum \alpha_i E\left(\frac{Y_i - \bar{Y}}{x_i - \bar{x}} \right) = \sum \alpha_i \beta_1 = \beta_1$$

    , since $\sum \alpha_i = 1$.
  - And since this is the case for all x's, we have an unbiased estimator of β_1.
  - Since we only require $\sum \alpha_i = 1$, we have an infinite number of unbiased slope estimators (as we vary the α_i's).
- So the fact that OLS gives you LUE's does not make OLS so special!

So many LUEs! Test your understanding!

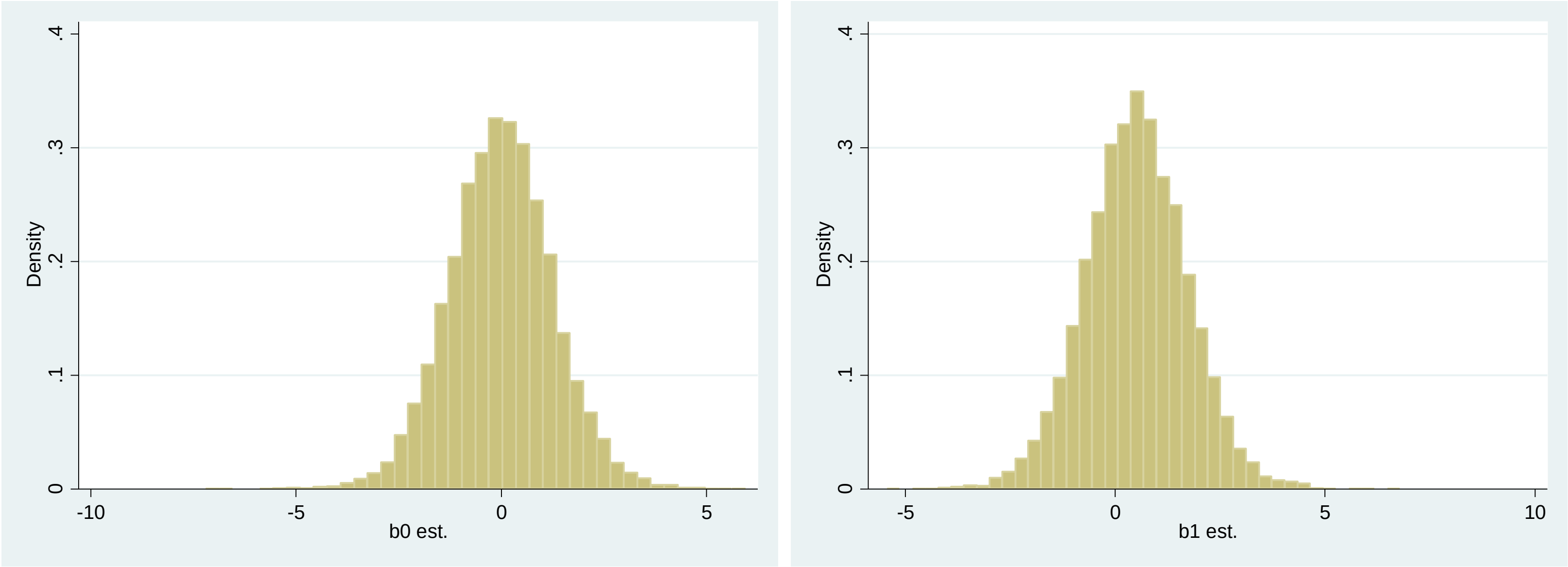
- From before,
  - Given SLR.1-SLR.4, $E\left(Y_i \mid x_i\right) = \beta_0 + \beta_1 x_i$ and $E\left(\bar{Y} \mid x's\right) = \beta_0 + \beta_1 \bar{x}$.
  - Then $E\left(\frac{Y_i - \bar{Y}}{X_i - \bar{X}} \middle| x's\right) = \beta_1$, since $E\left(\frac{Y_i - \bar{Y}}{X_i - \bar{X}} \middle| x's\right) = \beta_1 = \frac{E\left(Y_i - \bar{Y} \mid x's\right)}{(x_i - \bar{x})}$

$$= \frac{E\left(Y_i \mid x's\right) - E\left(\bar{Y} \mid x's\right)}{(x_i - \bar{x})} = \frac{\left(\beta_0 + \beta_1 x_i\right) - \left(\beta_0 + \beta_1 \bar{x}\right)}{(x_i - \bar{x})} = \frac{\beta_1 (x_i - \bar{x})}{(x_i - \bar{x})} = \beta_1$$

- So each of the following is an unbiased slope estimator, conditional on the x's:
  - $B_1 = \left(\frac{Y_1 - \bar{Y}}{X_1 - \bar{X}}\right)$ ***Answer***: $E\left(B_1 \mid x's\right) = \frac{\left(\beta_0 + \beta_1 x_1\right) - \left(\beta_0 + \beta_1 \bar{x}\right)}{(x_1 - \bar{x})} = \frac{\beta_1 (x_1 - \bar{x})}{(x_1 - \bar{x})} = \beta_1$
  - $B_1 = .5\left(\frac{Y_1 - \bar{Y}}{X_1 - \bar{X}}\right) + .5\left(\frac{Y_5 - \bar{Y}}{X_5 - \bar{X}}\right)$
  - $B_1 = \left(\frac{Y_1 - Y_5}{X_1 - X_5}\right)$
  - $B_1 = .9\left(\frac{Y_1 - \bar{Y}}{X_1 - \bar{X}}\right) + .1\left(\frac{Y_5 - \bar{Y}}{X_5 - \bar{X}}\right)$
  - $B_1 = .5\left(\frac{Y_1 - Y_5}{X_1 - X_5}\right) + .5\left(\frac{Y_3 - Y_7}{X_3 - X_7}\right)$
  - $B_1 = 1.5\left(\frac{Y_1 - \bar{Y}}{X_1 - \bar{X}}\right) - .5\left(\frac{Y_5 - \bar{Y}}{X_5 - \bar{X}}\right)$
  - $B_1 = 1.5\left(\frac{Y_1 - Y_5}{X_1 - X_5}\right) - .5\left(\frac{Y_3 - Y_7}{X_3 - X_7}\right)$

OLS Estimators (B_0 and B_1) have Variances I

- Variances of the OLS estimators: The OLS estimators, B_0 and B_1, are random variables, with a joint distribution, means, variances and a covariance. The sample you are working with is just one of many possible samples.
- **An example**.
  - A random sample: $Y_i = 0 + .5X_i + U_i$, $X_i \sim Uniform[0,1]$; $U_i \sim N(0,1)$; nObs=10
  - Here are distributions of those 10,000 estimated intercepts and slopes:



The means of the 10,000 estimates are quite close to the true parameter values… but notice the large variation driven by the random nature of the DGM.

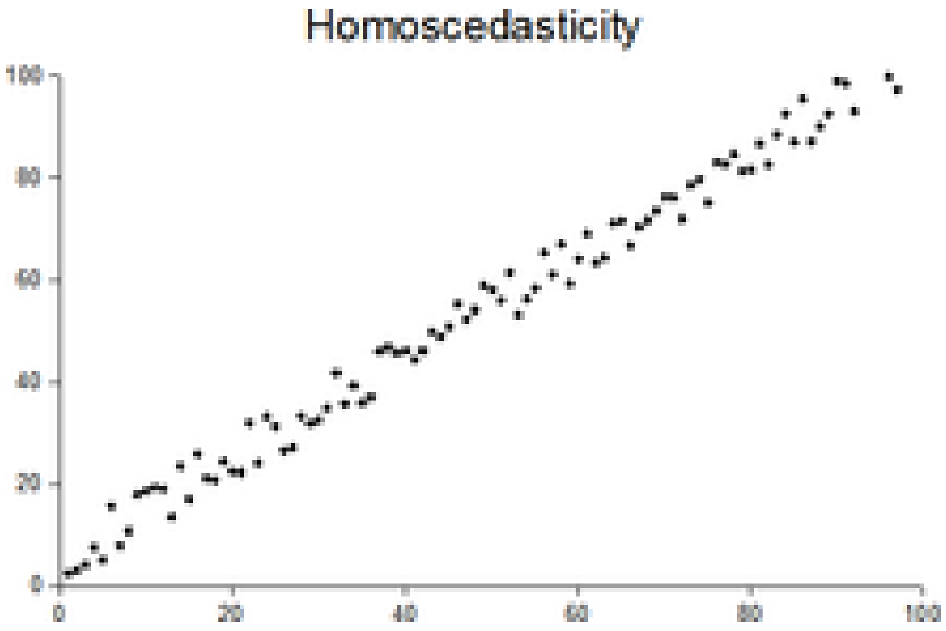
*The **BLUE** Challenge: Which LUE has the smallest variance?*

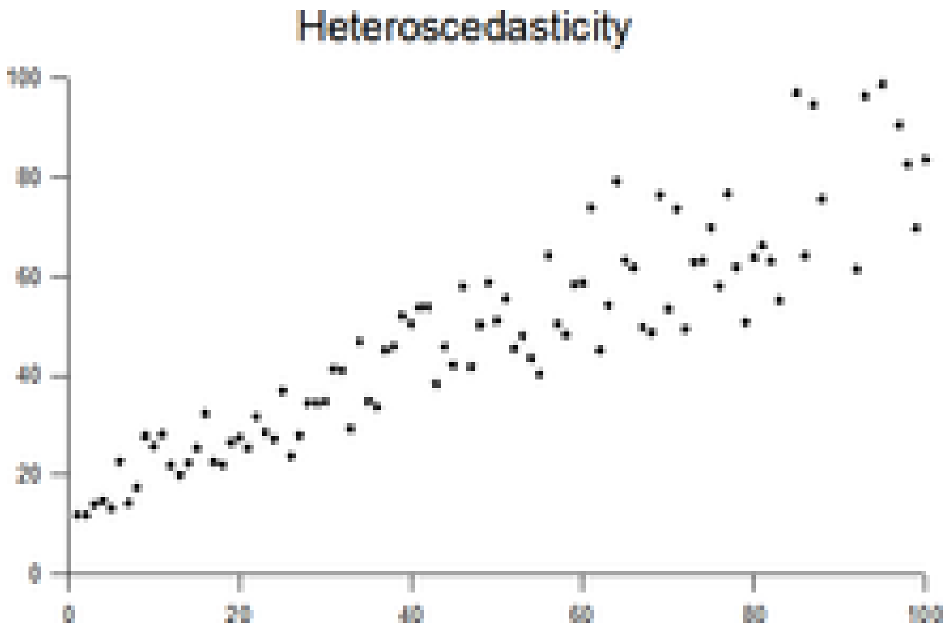
- ***I say again!***: Because they are random variables, estimators have means, variances, covariances, correlations, etc etc etc.
- ***In particular***: The OLS estimators B_0 and B_1 are random variables, with a joint distribution, means, variances, and a covariance. Different samples will generate different intercept and slope estimates. Who knows if your sample is representative? … your estimates could in fact be not at all close to the true parameter values. It all depends on your sample!
- Getting to ***BLUE*** (***Best Linear Unbiased Estimators***):

  This will be all about finding the LUE(s?) (amongst the many) with the minimum variance.

SLR.5: *Homoskedasticity*

- **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.





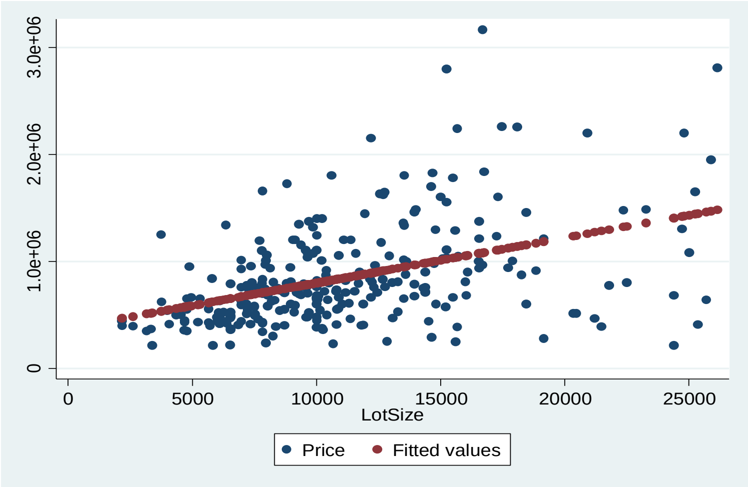
Heteroskedasticity Example: *Real Estate valuation*

- Newton real estate sales prices and lot sizes (heteroskedasticity)

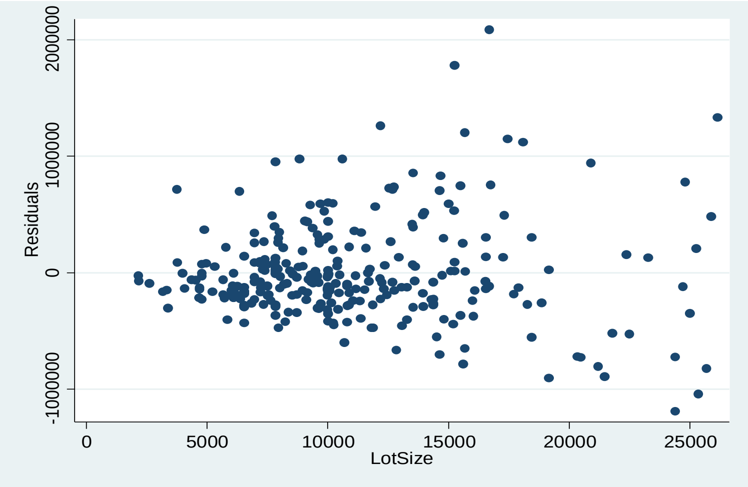
```
      Source |       SS           df       MS      Number of obs   =       284
-------------+----------------------------------   F(1, 282)       =     69.24
       Model |  1.2374e+13         1  1.2374e+13   Prob > F        =    0.0000
    Residual |  5.0402e+13       282  1.7873e+11   R-squared       =    0.1971
-------------+----------------------------------   Adj R-squared   =    0.1943
       Total |  6.2776e+13       283  2.2182e+11   Root MSE        =   4.2e+05

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lotsize |   42.22929   5.075149     8.32   0.000     32.23931    52.21928
       _cons |   374248.4   61384.29     6.10   0.000     253418.8      495078
------------------------------------------------------------------------------
```

SLR.1: Linear (DGM) Model
SLR.2: Random Sample
SLR.3: Sample variation in the RHS variable
SLR.4: U has zero mean | RHS variable

SLR.5: Homoskedasticity | RHS variable



predicteds v. actuals



residuals v. lot size

OLS Estimators (B_0 and B_1) have Variances II

- If SLR.5 holds, in addition to SLR.1-SLR.4, then we have the following variances of the OLS estimators, conditional on the particular sample of $\{x_i\}$:

  - $Var(B_1 \mid x's) = \dfrac{\sigma^2}{\sum (x_j - \overline{x})^2}$ and $StdDev(B_1 \mid x's) = sd(B_1 \mid x's) = \dfrac{\sigma}{\sqrt{\sum (x_i - \overline{x})^2}}$

  - $Var(B_0 \mid x's) = \dfrac{\sigma^2}{n} \dfrac{\sum x_i^2}{\sum (x_j - \overline{x})^2}$

- Comments:

  - $Var(B_1)$ decreases with *decreases* in the error variance, σ^2, and with *increases* in the variation of the independent variable. Makes sense?

  - Where does this variance come from? The estimator is always just the OLS estimator, so the variation is coming from the DGM.

MSE/RMSE and the *Standard Error of the Regression*

- ***Mean Squared Error*** (**MSE**): Typically, we don't know the actual value of the variance σ^2. But we can estimate it with the: $\hat{\sigma}^2 = \frac{SSR}{n-2} = MSE$.
  - Recall: MSE is one of our *Goodness-of-Fit* metrics in OLS/SLR Assessment.
- ***RMSE:*** The *standard error of the regression*, sometimes called the Root MSE (or RMSE), is the square root of this: $\hat{\sigma} = \sqrt{\frac{SSR}{n-2}} = \sqrt{MSE} = RMSE$.

MSE is an *Unbiased Estimator* of *var(U|x)*

Unbiasedness I: $E\left(MSE\right) = \sigma^2 = \text{var}(U \mid X = x)$, given SLR.1-SLR.5

- $MSE = \hat{\sigma}^2$ is an *unbiased estimator of the variance* , σ^2 (the homoscedastic error) , given SLR.1-SLR.5 and conditional on the x's

Unbiasedness II: $E\left(\dfrac{MSE}{(n-1)S_{xx}}\right) = Var\left(B_1\right)$, given SLR.1-SLR.5

- Given the above, and since $Var(B_1) = \dfrac{\sigma^2}{\sum (x_i - \overline{x})^2}$, we have:

$\dfrac{MSE}{\sum (x_i - \overline{x})^2} = \dfrac{MSE}{(n-1)S_{xx}}$ is an unbiased estimator of $Var(B_1)$.

Standard Errors of B_1: *Estimates of sd(B_1)*

- We don't typically know the actual value of σ, and so we usually can't derive $sd(B_1) = \dfrac{\sigma}{\sqrt{\sum (x_i - \overline{x})^2}}$.

- But we can approximate $sd(B_1)$, with the *standard error* of B_1, $se(B_1)$, by approximating σ with $RMSE = \hat{\sigma}$:

  - $StdErr(B_1) = se(B_1) = \dfrac{\hat{\sigma}}{\sqrt{\sum (x_i - \overline{x})^2}} = \dfrac{RMSE}{\sqrt{\sum (x_i - \overline{x})^2}} = \dfrac{RMSE}{S_x \sqrt{n-1}}$.

  - $se(B_1)$ is useful in statistical inference… for constructing confidence intervals for, and testing hypotheses about, β_1, the true slope parameter in the DGM.

Onwards to Gauss, Markov, BLUE… and Inference!