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)*** **(separate handout)**

Those OLS Estimates

1) Recall the birth of OLS estimates:

   a) You have a dataset consisting of n observations of (x, y): $\{x_i, y_i\}$ $i = 1, 2, \ldots n$.

   b) You believe that except for random noise in the data, there is a linear relationship between the x's and the y's: $y_i \sim \beta_0 + \beta_1 x_i$ … and want to estimate the unknown parameters β_0 (intercept parameter) and β_1 (slope parameter).

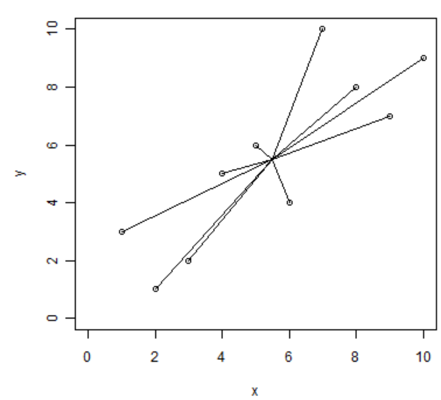
   c) Adopting OLS, you estimate β_0 and β_1 by minimizing sum squared *residuals* (*SSR*s):

   $$\min SSR = \sum \left(y_i - \left(\beta_0 + \beta_1 x_i \right) \right)^2 \text{ wrt } \beta_0 \textit{ and } \beta_1$$

   d) For the given sample, the OLS estimates of the unknown intercept and slope parameters are:

      i) ***Slope***: $$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \sum \frac{(x_i - \bar{x})^2}{(n-1)S_{xx}} \left(\frac{y_i - \bar{y}}{x_i - \bar{x}} \right) = \sum w_i \left(\frac{y_i - \bar{y}}{x_i - \bar{x}} \right),$$ where

      $w_i = \dfrac{(x_i - \bar{x})^2}{(n-1)S_{xx}}$ and $\sum w_i = 1$… so the w_i's sum to 1 and are proportional to the square of the x-distance from $\bar{x}$.



      ii) Recall that $\left(\dfrac{y_i - \bar{y}}{x_i - \bar{x}} \right)$ is the slope of the line connecting (x_i, y_i) to the sample means point, $(\bar{x}, \bar{y})$.

      iii) ***Intercept***: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Estimates (ex post) v. Estimators (ex ante)

2) ***exPost*** *(actual; after the event)*: After the data set is generated, OLS provides numeric ***estimates*** of the slope and intercept parameters, for the given dataset (estimates are numbers, not random variables).

   a) After we have drawn the sample $\{x_i, y_i\}$, we have the OLS *estimates*:

      i) ***Slope estimate***: $\hat{\beta}_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$, and

      ii) ***Intercept estimate***: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

exAnte-exPost

3) ***exAnte*** *(before the event)*: But prior to the generation of the data, the x's and y's are random variables, X's and Y's, and OLS provides a **rule** for estimating the OLS coefficients (for any realized data set).

   a) We call these rules ***estimators***. Estimators will take on different values depending on the actual drawn sample, the actual x's and y's. Accordingly, estimators are random variables.

   b) The X_i's and Y_i's are random variables, and OLS provides us with slope and intercept *estimators*:

      (1) ***Slope estimator***: $B_1 = \dfrac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_j - \bar{X})^2} = \dfrac{\sum (X_i - \bar{X})Y_i}{\sum (X_j - \bar{X})^2}$

      (2) ***Intercept estimator***: $B_0 = \bar{Y} - B_1 \bar{X}$

4) To review notation, we have:

   a) Random variables (upper case letters): X's and Y's

   b) Data (lower case letters): x's and y's

   c) True parameters: β_0 *and* β_1

   d) Parameter estimators (random variables; upper case letters): B_0 *and* B_1

   e) Parameter estimates (estimated coefficients; denoted with *hats*): $\hat{\beta}_0$ *and* $\hat{\beta}_1$

SLR Models – *Estimation*

The *Simple Linear Regression* (*SLR*) Conditions SLR.1-SLR.4

5) ***SLR.1 – Linear model (DGM)***: $Y = \beta_0 + \beta_1 X + U$, where X, Y *and* U are random variables and β_0 *and* β_1 are unknown parameters to be estimated.[1]
   a) This is sometimes referred to as the ***Data Generation Mechanism*** (***DGM***), as it describes the process by which the data are assumed to have been generated.
   b) Notice that X, U and Y are now random variables, reflecting the random nature of the *DGM*.
   c) U is the ***unexplained*** (or *unobserved*) *error* term (or *disturbance*), and captures other factors (excluded from the model) that explain Y:
      i) $U = Y - \beta_0 - \beta_1 X$
      ii) Put differently: U is the part of Y not explained/generated by the linear function of X.
6) **SLR.2 – *Random sampling***: the sample $\{(x_i, y_i)\}$ is a random sample; the i[th] elements in the sample, (x_i, y_i), is the realization of two random variables (X_i, Y_i), where $Y_i = \beta_0 + \beta_1 X_i + U_i$.
7) **SLR.3 – *Sample variation in the independent variable***: the x_i's are not all the same value
8) ***SLR.4 – U has zero conditional mean*:** $E(U \mid X = x) = 0$ for all x
   a) $E(U \mid X = x) = 0$ for any x implies that:
      i) $E(U) = 0$ (U has mean zero)
         (1) If the expected value of U is 0 conditional on any value of x, then the overall expected value of U, which will be a weighted average of the conditional expectations of U, will also be 0.
         (2) Put differently: $E(U) = \sum_x prob(X = x)E(U \mid X = x) = \sum_x prob(X = x) \cdot 0 = 0$
      ii) $Cov(X, U) = 0$ (X and U are uncorrelated)[2]
         (1) $Cov(X,U) = E\left((X - \mu_X)(U - \mu_U)\right) = E\left((X - \mu_X)U\right)$ since $\mu_U = 0$
         $= E\left(XU\right) - \mu_X E\left(U\right) = E\left(XU\right) = \sum_x prob(X = x)E(xU \mid X = x)$
         $= \sum_x prob(X = x)xE(U \mid X = x) = \sum_x prob(X = x)x \cdot 0 = 0$

---

[1] At times we will consider the X values to be exogenously given, in which case the explanatory variable is not random.

[2] Recall that if X and U are independent then $Cov(X,U) = 0$. A covariance of 0, however, does not imply independence, but rather than X and U do not move together in much of a linear way.

(2) Notice the connection to Omitted Variable Bias, which is driven by correlation between U and X.

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

An Aside: The Population Regression Function (PRF)

9) Under these assumptions (specifically SLR.1 and SLR.4):

   a) $E(Y \mid X = x) = \beta_0 + \beta_1 x + E(U \mid x) = \beta_0 + \beta_1 x + 0$, so the expected value of Y given x is $E(Y \mid X = x) = \beta_0 + \beta_1 x$.

10) ***Population Regression Function (PRF)***: $E(Y \mid X = x) = \beta_0 + \beta_1 x$

   a) This traces out the conditional means (conditional on the values of X, the particular values of the RHS variable) of the dependent variable Y.

B_0 and B_1 are Linear Estimators (conditional on the x's)

11) The OLS slope estimator, B_1, is linear in the Y_i's (conditional on the x's), since we can express the estimator as:

   a) $B_1 = \sum b_i Y_i$, where $b_i = \dfrac{(x_i - \bar{x})}{\sum (x_j - \bar{x})^2} = \dfrac{(x_i - \bar{x})}{(n-1)S_{xx}}$

   b) 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.

12) And the OLS intercept estimator is also linear in the Y_i's (conditional on the x's) since:

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

13) So conditional on the x's, B_0 and B_1 are linear estimators.

OLS estimators are unbiased! (under SLR.1-SLR.4) ***Who saw this coming***?

14) Given assumptions/conditions SLR.1-SLR.4, B_0 *and* B_1, the OLS estimators of the intercept and slope parameters, are unbiased estimators (so for each, the expected value of the estimator is the true parameter value).

  a) We'll be proving this below.

  b) The trick in the proof is to assume a particular set of sample x_i's, and to show that for any such sample, the OLS estimators will be unbiased. And since this is true for any sample of x_i's, it must be true in *expectation*.

    i) This is sometimes referred to as the *Law of Iterated Expectation*. We will skip the proof of this *Law*… but the intuition is pretty straightforward, yes? … and you saw it in action above in the proof that $E(U) = 0$ given SLR.4.

15) Conditional on the x's, the OLS estimators (random variables… note the capital B's below) are defined by:

  a) $B_1 = \dfrac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} = \dfrac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{(n-1)S_{xx}} = \sum \left[\dfrac{(x_i - \bar{x})^2}{(n-1)S_{xx}} \right] \left(\dfrac{Y_i - \bar{Y}}{x_i - \bar{x}} \right)$,

    and so $B_1 = \sum w_i \dfrac{(Y_i - \bar{Y})}{(x_i - \bar{x})}$,

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

  b) $B_0 = \bar{Y} - B_1 \bar{x}$

16) An interesting result:

  a) Assume SLR.1-SLR.4.

    i) Then $Y = \beta_0 + \beta_1 X + U$, and $\mu_Y = \beta_0 + \beta_1 \mu_X$ (since $E(U) = 0$)

    ii) $Cov(X,Y) = Cov(X, \beta_0 + \beta_1 X + U) = \beta_1 Cov(X,X) + Cov(X,U) = \beta_1 Cov(X,X)$ since $Cov(X,U) = 0$.

  b) But then:

    i) $\beta_1 = \dfrac{Cov(X,Y)}{Cov(X,X)}$, and

    ii) $\beta_0 = \mu_Y - \beta_1 \mu_X = \mu_Y - \dfrac{Cov(X,Y)}{Cov(X,X)} \mu_X$.

  c) Notice the resemblance to the OLS estimators!

17) **The OLS slope estimator B1 is unbiased!** - $E\left(B_1\right)=\beta_1$

a) Step 1 - $E\left(\frac{Y_i-\bar{Y}}{x_i-\bar{x}}\right)=\beta_1$: To evaluate the expected value of B_1 conditional on the x's, we need to determine $E\left(\frac{Y_i-\bar{Y}}{x_i-\bar{x}}\right)$, for each *i*. But $E\left(Y_i \mid x_i\right)=\beta_0+\beta_1 x_i+E(U_i \mid x_i)$ $=\beta_0+\beta_1 x_i$ (given SLR.4). And since $E\left(\bar{Y} \mid x's\right)=\frac{1}{n}\sum E\left(Y_i \mid x_i\right)=\frac{1}{n}\sum\left(\beta_0+\beta_1 x_i\right)$ $=\beta_0+\beta_1\bar{x}$, we have: $E\left(\frac{Y_i-\bar{Y}}{x_i-\bar{x}}\right)=\frac{\left(\beta_0+\beta_1 x_i\right)-\left(\beta_0+\beta_1\bar{x}\right)}{(x_i-\bar{x})}=\frac{\beta_1(x_i-\bar{x})}{(x_i-\bar{x})}=\beta_1$.

b) Step 2 - $E\left(B_1 \mid x's\right)=\beta_1$ (B_1 in an unbiased estimator of β_1, conditional on the x's):

i) Since $E\left(B_1 \mid x's\right)=E\left(\sum w_i\left(\frac{Y_i-\bar{Y}}{x_i-\bar{x}}\right)\right)=\sum w_i E\left(\frac{Y_i-\bar{Y}}{x_i-\bar{x}}\right)=\sum w_i\beta_1=\beta_1$ (since the weights sum to 1… $\sum w_i=1$).

c) Step 3 - $E\left(B_1\right)=\beta_1$: And since $E\left(B_1 \mid x's\right)=\beta_1$ for all x's, $E\left(B_1\right)=\beta_1$.[3]

18) **The OLS intercept estimator B0 is also unbiased!** - $E\left(B_0\right)=\beta_0$

a) Step 1 - $E\left(B_0 \mid x's\right)=\beta_0$ (B_0 in an unbiased estimator of β_0, conditional on the x's):
$E\left(B_0 \mid x's\right)=E\left(\bar{Y}\right)-E\left(B_1\right)\bar{x}=\beta_0+\beta_1\bar{x}-\beta_1\bar{x}=\beta_0$.

b) Step2 - $E\left(B_0\right)=\beta_0$: And since $E\left(B_0 \mid x's\right)=\beta_0$ for all x's, $E\left(B_0\right)=\beta_0$.

19) $OLS \equiv LUE$ (given SLR.1-SLR.4): So given the SLR condition 1-4, the OLS slope and intercept estimators are ***linear unbiased estimators*** (***LUEs***) of the unknown parameter values![4]

OLS ≙ LUE

20) ***Who saw this coming***? Who ever thought that process of minimizing SSRs would lead to unbiased estimators?

[3] This last step is an application of the *Law of Iterated Expectations*.
[4] But remember that they are only linear conditional on the x's.

But B_1 is not alone!!!

… there are in fact an ***infinite*** number of linear unbiased slope estimators (given SLR.1-SLR.4)

21) 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:

   a) Consider the estimator defined by $\sum \alpha_i \left(\frac{Y_i - \bar{Y}}{X_i - \bar{X}} \right)$, where $\sum \alpha_i = 1$.

   b) Then conditional on the x's:

   $$E\left(\sum \alpha_i \left(\frac{Y_i - \bar{Y}}{x_i - \bar{x}} \right) \right) = \sum \alpha_i E\left(\frac{Y_i - \bar{Y}}{x_i - \bar{x}} \right) = \sum \alpha_i \beta_1 = \beta_1, \text{ since } \sum \alpha_i = 1.$$

   c) And since this is the case for all x's, we have an unbiased estimator of β_1.

   d) Since we only require that the α_i's sum to one, we have an infinite number of unbiased slope estimators (as we vary the α_i's).

22) ***Test your understanding!*** From before, you know that given SLR.1-SLR.4, $E\left(Y_i \mid x_i\right) = \beta_0 + \beta_1 x_i$ and $E\left(\bar{Y} \mid x's\right) = \beta_0 + \beta_1 \bar{x}$. Use these results to show that each of the following will be unbiased slope estimators, conditional on the x's. And accordingly, by the Law of Iterated Expectations, unbiased slope estimators overall:

   a) $B_1 = \left(\frac{Y_1 - \bar{Y}}{X_1 - \bar{X}} \right)$ Answer: $E\left(B_1 \mid x's\right) = \frac{(\beta_0 + \beta_1 x_1) - (\beta_0 + \beta_1 \bar{x})}{(x_1 - \bar{x})} = \frac{\beta_1 (x_1 - \bar{x})}{(x_1 - \bar{x})} = \beta_1$, all x's.

   b) $B_1 = .5\left(\frac{Y_1 - \bar{Y}}{X_1 - \bar{X}} \right) + .5\left(\frac{Y_5 - \bar{Y}}{X_5 - \bar{X}} \right)$

   c) $B_1 = .9\left(\frac{Y_1 - \bar{Y}}{X_1 - \bar{X}} \right) + .1\left(\frac{Y_5 - \bar{Y}}{X_5 - \bar{X}} \right)$

   d) $B_1 = 1.5\left(\frac{Y_1 - \bar{Y}}{X_1 - \bar{X}} \right) - .5\left(\frac{Y_5 - \bar{Y}}{X_5 - \bar{X}} \right)$

   e) $B_1 = \left(\frac{Y_1 - Y_5}{X_1 - X_5} \right)$

   f) $B_1 = .5\left(\frac{Y_1 - Y_5}{X_1 - X_5} \right) + .5\left(\frac{Y_3 - Y_7}{X_3 - X_7} \right)$

23) And so getting to ***BLUE*** (***Best Linear Unbiased Estimators***) will be all about finding the LUE(s?) (amongst the many) with the minimum variance.

OLS estimators (B_0 and B_1) have variances

24) Variances of the OLS estimators: Yes, because they are 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 that you could have drawn.

25) **An example**.

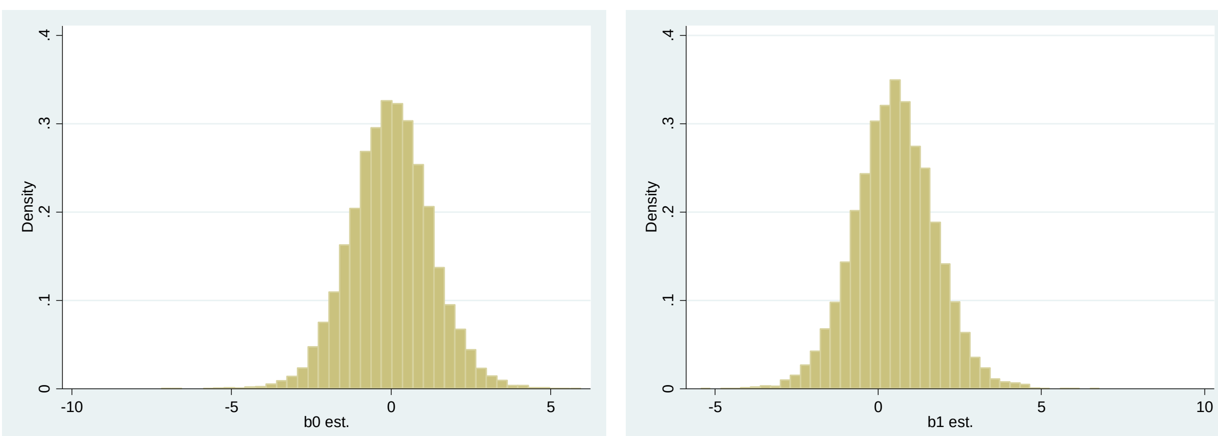
a) Consider $Y_i = 0 + .5X_i + U_i$, $X_i \sim Uniform[0,1]$ and $U_i \sim N(0,1)$.

b) The following show the results from 10,000 samples, each with 10 observations generated by the random process above… with one slope and intercept estimate per sample.

c) Distribution of OLS intercept and slope estimates:

```
. summ b0est b1est

Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   b0est |    10,000   -.0077699    1.277106  -7.205879   5.944109
   b1est |    10,000    .5024118    1.229428  -5.437995   6.791059
```



26) The means of the 10,000 estimates are quite close to the true parameters values… but notice the large variation in slope and intercept estimates, driven by the random nature of the DGM and depending on the particular sample that you are working with.

27) **Worth repeating!!!:** Because they are estimators, 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!

SLR Models – *Estimation*

SLR.5 – Homoskedasticity

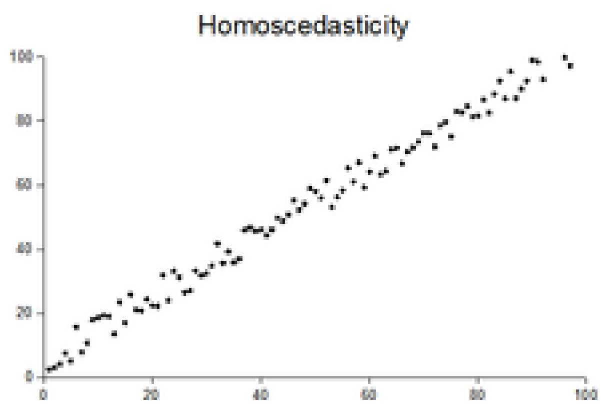
28) SLR.5: *Homoskedasticity* (constant conditional variance of the error term)

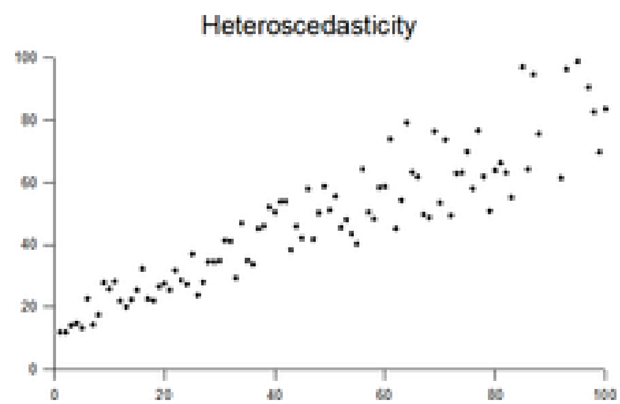
   a) To derive the variances of the estimators, we make one additional assumption:

      SLR.5: $Var(U \mid X = x) = \sigma^2$ for all x

   b) Note that SLR.5 holds if U is independent of X, so that $Var(U \mid X = x) = Var(U) = \sigma^2$.

   c) *Heteroskedasticity*: the conditional variances are not all the same.

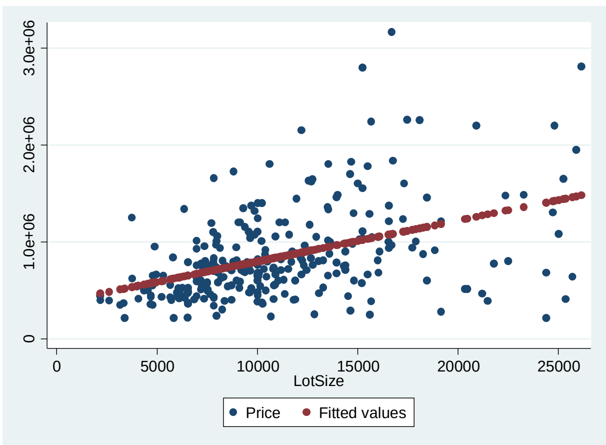




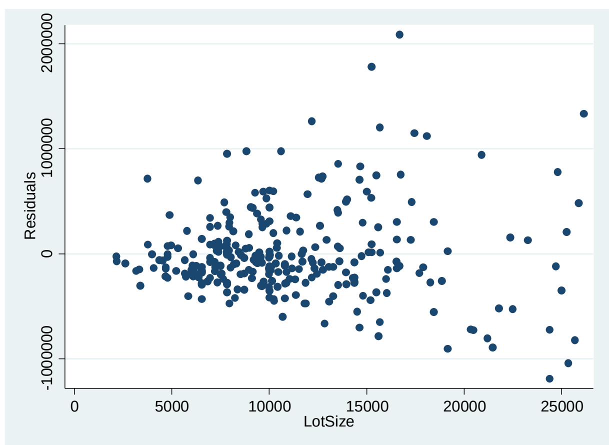
Example: Newton real estate sales prices and lot sizes (heteroskedasticity)

```
      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
------------------------------------------------------------------------------
```



predicteds v. actuals



residuals v. lot size

29) **A simple test for heteroskedasticity:** Regress the squared residuals on the RHS variable. Are there more complicated test? *Of course! But simpler is always better!*

```
. predict resid, res
. gen resid2=resid^2

. reg resid2 lotsize

      Source |       SS           df       MS      Number of obs   =       284
-------------+----------------------------------   F(1, 282)       =     43.03
       Model |  6.3833e+24         1  6.3833e+24   Prob > F        =    0.0000
    Residual |  4.1829e+25       282  1.4833e+23   R-squared       =    0.1324
-------------+----------------------------------   Adj R-squared   =    0.1293
       Total |  4.8212e+25       283  1.7036e+23   Root MSE        =    3.9e+11

------------------------------------------------------------------------------
      resid2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lotsize |   3.03e+07    4623424     6.56   0.000     2.12e+07    3.94e+07
       _cons |  -1.57e+11   5.59e+10    -2.81   0.005    -2.67e+11   -4.73e+10
------------------------------------------------------------------------------
```

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

Variance of the OLS Estimators (assuming SLR.1-SLR.5)

30) 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 $\{x_i\}$:[5]

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

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

31) Comments:

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

[5] See the Appendix for the proof for the variance of the slope estimator.

b) Where does this variance come from? The estimator is always just the OLS estimator, so all of the variation is coming from the different possible data samples generated by the DGM. (See the Excel simulations.)

MSE/RMSE (Goodness-of-Fit) and Standard Errors

32) ***Mean Squared Error*** (**MSE**): Typically, we don't know the actual value of the variance σ^2. But we can estimate it with the: $\hat{\sigma}^2 = \frac{SSR}{n-2} = MSE$.

a) Recall that MSE was one of our *Goodness-of-Fit* metrics in OLS/SLR Assessment.

33) ***Unbiasedness I:*** $MSE = \hat{\sigma}^2$ is an *unbiased estimator of the variance* , σ^2

a) Under SLR.1-SLR.5 and conditional on the x's, $MSE = \hat{\sigma}^2$ will be an *unbiased estimator of the variance* , σ^2, of the error term U (the homoskedastic error).[6]

34) ***Unbiasedness II:*** $\frac{MSE}{(n-1)S_{xx}}$ is an unbiased estimator of $Var(B_1)$

a) Since $\hat{\sigma}^2 = MSE$ is an unbiased estimator of σ^2 (under SLR.1-SLR.5) and since $Var(B_1) = \frac{\sigma^2}{\sum (x_i - \overline{x})^2}$, $\frac{MSE}{\sum (x_i - \overline{x})^2} = \frac{MSE}{(n-1)S_{xx}}$ is an unbiased estimator of $Var(B_1)$.

35) ***RMSE:*** The *standard error of the regression*, sometimes called the Root MSE (or RMSE), is the square root of this: $\hat{\sigma} = \sqrt{\frac{SSR}{n-2}} = \sqrt{MSE} = RMSE$.

36) ***Standard Errors of B₁: Estimates of sd(B₁)***

a) As mentioned above, we don't typically know the actual value of σ, and so we can't usually derive $sd(B_1) = \frac{\sigma}{\sqrt{\sum (x_i - \overline{x})^2}}$. However, we can approximate $sd(B_1)$, with the *standard error* of B_1, $se(B_1)$.

b) If we approximate σ with $RMSE = \hat{\sigma}$, then we can derive the standard error of B_1:

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

c) This estimate of $sd(B_1)$ will prove to be useful in statistical inference… for constructing confidence intervals for, and testing hypotheses about, β_1, the true slope parameter in the DGM.

---

[6] For the proof, see the text.

Appendix:

37) Here's the derivation of the variance of the slope estimator.

a) From above we have $B_1 = \sum\left(\frac{(x_i - \overline{x})}{(n-1)S_{xx}}\right)Y_i \ = \sum \alpha_i Y_i$, where $\alpha_i = \frac{(x_i - \overline{x})}{(n-1)S_{xx}}$.

b) Since the $Y_i\,'s$ are independent, the variance of the sum is the sum of the variances, and $Var(B_1) = Var(\sum \alpha_i Y_i) = \sum \alpha_i^2 Var(Y_i) = \sum \alpha_i^2 \sigma^2 \ = \sigma^2 \sum \alpha_i^2$.

c) But $\alpha_i = \frac{(x_i - \overline{x})}{(n-1)S_{xx}}$ and so

$$\sum \alpha_i^2 = \sum \left[\frac{(x_i - \overline{x})}{(n-1)S_{xx}}\right]^2 = \frac{\sum (x_j - \overline{x})^2}{\left[(n-1)S_{xx}\right]^2} = \frac{(n-1)S_{xx}}{\left[(n-1)S_{xx}\right]^2} = \frac{1}{\sum (x_j - \overline{x})^2}$$

d) Therefore: $Var(B_1) = \frac{\sigma^2}{\sum (x_j - \overline{x})^2} = \frac{\sigma^2}{(n-1)S_{xx}}$