SLR Models: *Inference*

- ***SLR Assessment II: Precision/Inference***
- ***Sample Means and Inference: Conceptual Review***
- ***Onwards to SLR Inference***
- ***… SLR.6: U has a Normal Distribution***
- ***Distribution of the OLS Estimators (given SLR.1-SLR.6)***
- ***… Standard Errors and t Stats***
- ***t Statistics and Inference***
- ***… Confidence Intervals***
- ***… Hypothesis Testing***
- ***… p values and Statistical Significance***
- ***SLR Assessment Metrics Converge: t Stats and R^2***
- ***Example: Bodyfat***

SLR Assessment II: *Precision/Inference*

- When we initially considered the topic of SLR Assessment, we started with:
  - *After we have derived the OLS parameter estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, the question always arises: How well did we do? How close are the estimated coefficients to the true parameters, β_0 and β_1? We'll have several answers. None will be entirely satisfactory… though they will be informative, nonetheless.*
- We then discussed two approaches to SLR Assessment:
  - ***Goodness-of-Fit*** metrics (MSE/RMSE and R^2), which measured the extent to which our model explained the variation in the dependent variable, and
  - ***Precision/Inference*** metrics, which measured the precision with which we had estimated the unknown parameter values, β_0 and β_1.
- At that time there was extensive discussion of Goodness-of-Fit metrics (SLR Assessment I)…. but we totally punted on precision/inference.
- But we punt no more!
  - … and now turn to the second approach to SLR Assessment: ***Precision/Inference***

Samples Means and Inference: *Review*

- Recall from the *Review of Inference* and the case of estimating the mean of the distribution:
  - Under certain assumptions (including homoskedasticity) we found that the Sample Mean was a ***BLUE*** estimator of the unknown mean.
  - To generate confidence intervals or perform hypothesis testing, we made distributional assumptions, and assumed a Normal distribution.
  - Under those assumptions:
    - ***Confidence Intervals***: Interval estimators… *Sample Mean* +/- *c Standard Errors* (the critical value c comes from a t distribution with n-1 degrees of freedom)
    - ***Hypothesis Testing***: We reject the Null hypothesis ($H_0 : \mu = 0$) at significance level α only if the reported *p value* is less than α (or if the $|t\ stat| > c$, the critical value)
- These results carry over to the SLR models, virtually unchanged … just replace $(n-1)$ with $(n-2)$.

Recall those SLR Assumptions/Conditions

- **SLR.1 – *Linear model*** (the true model/DGM is in fact linear): $Y = \beta_0 + \beta_1 X + U$
- **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 all the same
- **SLR.4 – *Zero conditional mean of the error term***: $E(U \mid X = x) = 0$ for all x
- **SLR.5 – *Homoskedasticity*** (constant conditional variance of the error term):
  $Var(U \mid X = x) = \sigma^2$ for all x

SLR.1-.4: OLS = LUE

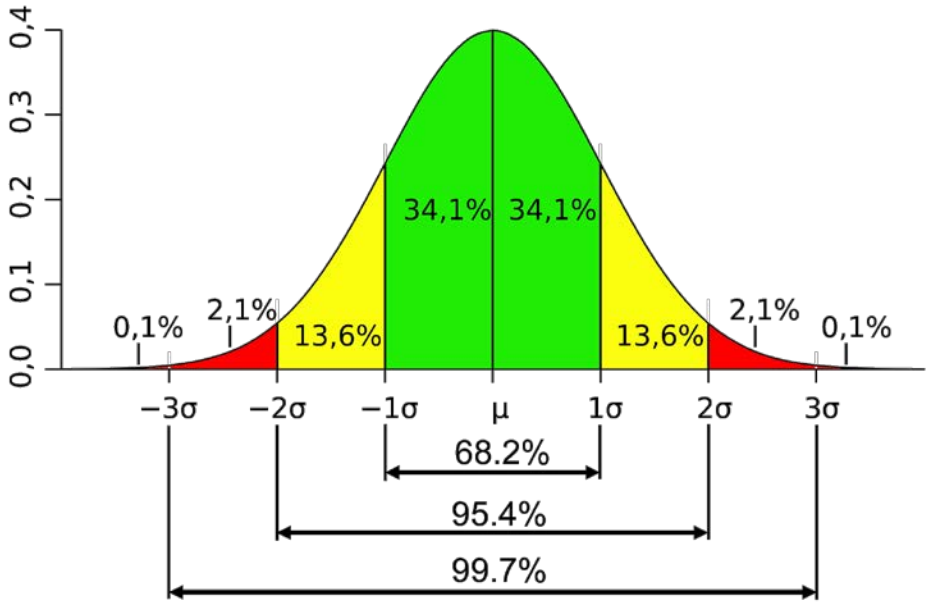
+ SLR.5: OLS = *BLUE*

Under those Assumptions/Conditions…

- ***LUEs***. Given SLR.1 – SLR.4, the OLS estimators are *LUE's* of the true parameters of the DGM, β_0 and β_1, so that $E\left(B_0\right)=\beta_0$ and $E\left(B_1\right)=\beta_1$, where:
  - $B_1=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_j-\bar{X})^2}=\frac{S_{XY}}{S_{XX}}=\frac{\sum(X_i-\bar{X})Y_i}{\sum(X_j-\bar{X})^2}$ and,
  - $B_0=\bar{Y}-B_1\bar{X}$.
- ***MSE** and **BLUE***. Adding in SLR.5 we have:
  - $\hat{\sigma}^2=MSE=\frac{SSR}{n-2}$ is an unbiased estimator of σ^2 , the conditional variance of U,
  - $\frac{MSE}{\sum(x_i-\bar{x})^2}$ is an unbiased estimator of $Var(B_1)$, and most importantly,
  - OLS estimators are ***BLUE*** estimators (***Best Linear Unbiased Estimators*** of β_0 and β_1). This last result is the **Gauss-Markov Theorem**.

SLR.6: U has a Normal Distribution

- Inference requires that we make one additional SLR assumption: ***Normal Distribution***
- **SLR.6 – *Normality***: U is independent of the RHS variable X and is Normally distributed with mean 0 and variance σ^2.
- Note that SLR.6 requires more than SLR.4 (U has conditional mean 0) and SLR.5 (homoskedasticity)… since it now specifies the actual distribution of U, not just its mean and variance.



- Recall that the Population Regression Function (PRF) is defined by: $E(Y \mid X = x) = \beta_0 + \beta_1 x$.
- SLR.6 implies that we know the actual the conditional <u>distribution</u> of Y (given $X = x$): $Y \mid X = x \sim Normal\left(\beta_0 + \beta_1 x, \sigma^2\right)$

Distribution of the OLS Estimators (given SLR.1-SLR.6)

- Given SLR.1-SLR.6, and conditional on the sample values of the x's, the OLS estimators will be Normally distributed:

$$B_1 \sim Normal\left(\beta_1, Var(B_1)\right), \text{ where } Var(B_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}.$$

- We can standardize B_1, so that: $\frac{B_1 - \beta_1}{sd(B_1)} \sim Normal\left(0,1\right)$, where $sd(B_1) = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}$.

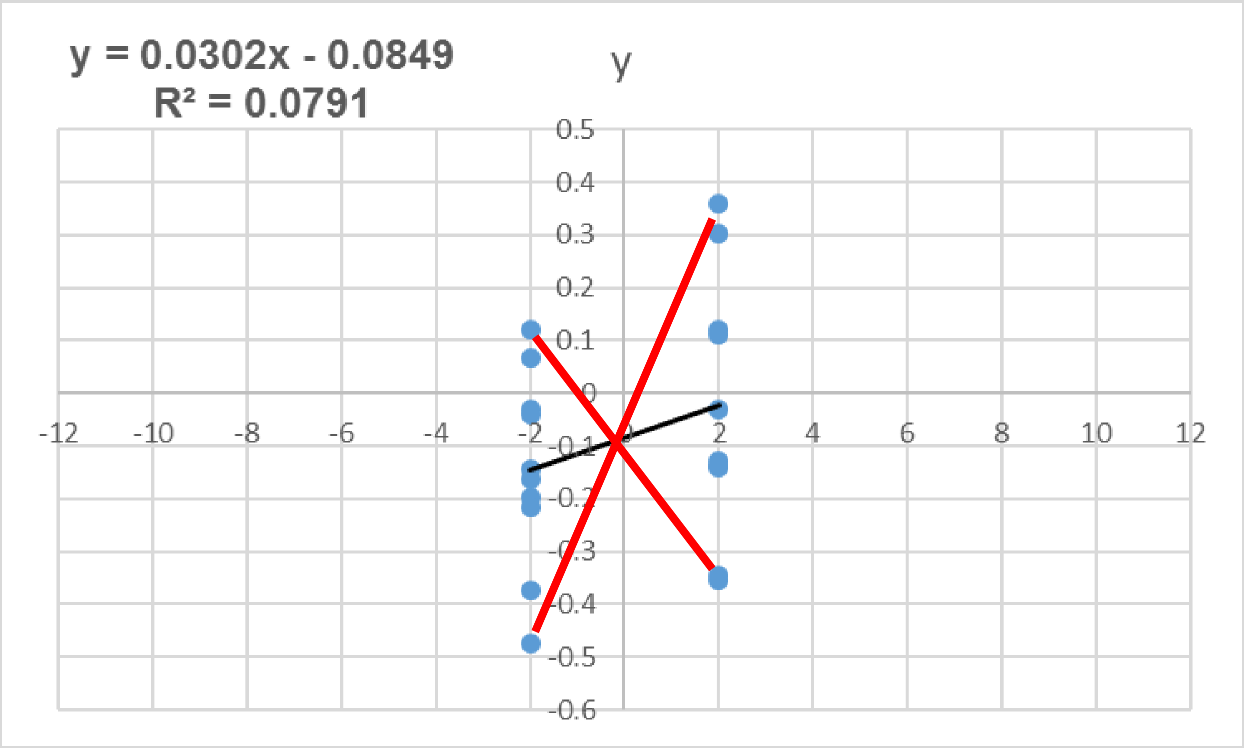
- Given SLR.1-SLR.5, and conditional on the x's, we have unbiased estimators of variances:

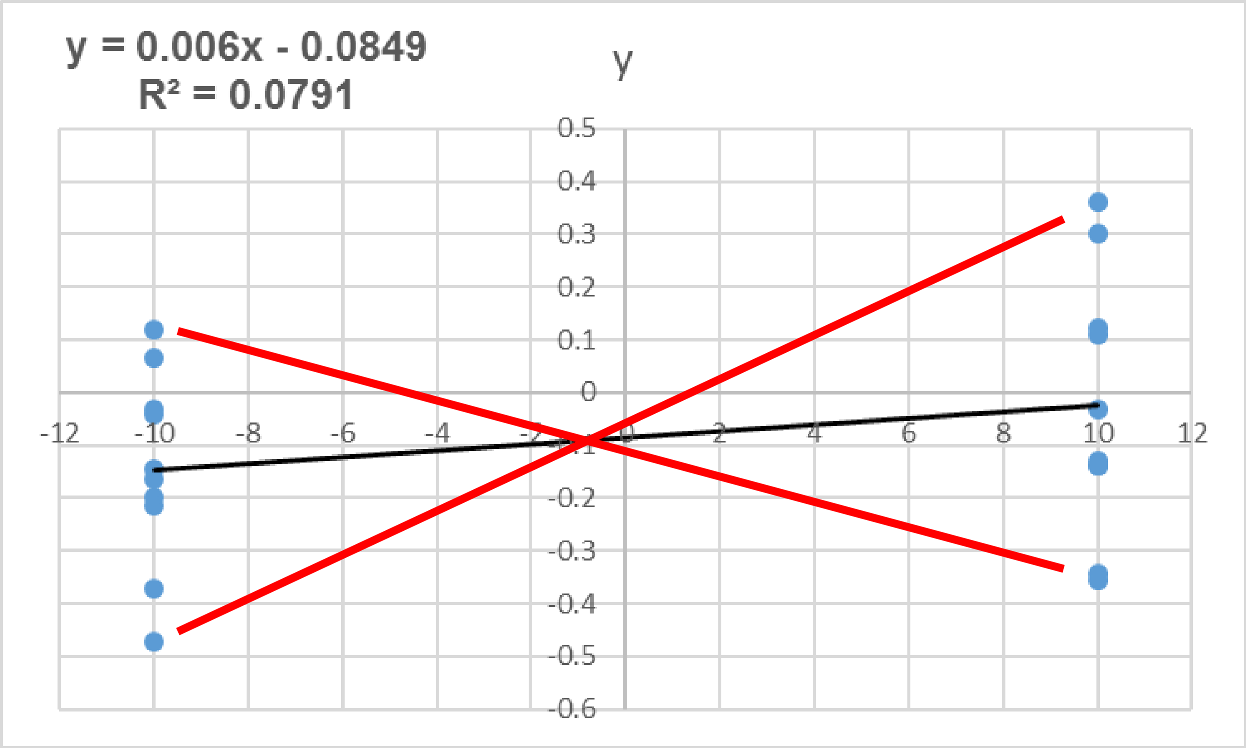
$$E(MSE) = \sigma^2, \text{ and } E\left(\frac{MSE}{\sum (x_i - \bar{x})^2}\right) = Var\left(B_1\right)$$

- … and so we use the standard error of B_1, $se(B_1)$ to estimate $sd(B_1)$:

$$se(B_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}} = \frac{RMSE}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{RMSE}{S_x \sqrt{(n-1)}}$$

Some Intuition? Why variance in the x's matters for std errs

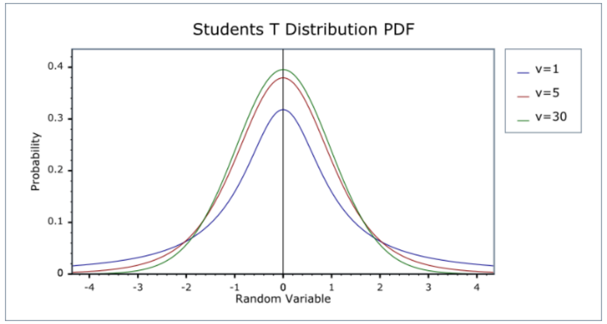




- ***Some intuition, maybe***: SLR.5 keeps the conditional variances constant. And so, as the x's are more spread out, there's less of a possible variation in the slopes.
  - On the left: close x's and lots of variation in the possible slopes (std err = .0243)
  - On the right: x's farther apart, and less variation in the possible slopes (std err = .0049)

The t Statistic, t Distribution and Confidence Intervals

- Recall the ***t statistic*** $\frac{B_1 - \beta_1}{se(B_1)}$ … the ***Cornerstone of Inference***… which enables us to:
  - to develop confidence intervals for β_1, and
  - to test hypotheses about β_1.



- Under the SLR.1 - SLR.6, the ***t statistic*** $\frac{B_1 - \beta_1}{se(B_1)}$ will have a t distribution with n-2 dofs.

… and Confidence Intervals

- Since $\frac{B_1 - \beta_1}{se(B_1)} \sim t_{n-2}$, the interval estimator, $\left[B_1 - c \cdot se(B_1),\ \ B_1 + c \cdot se(B_1)\right]$

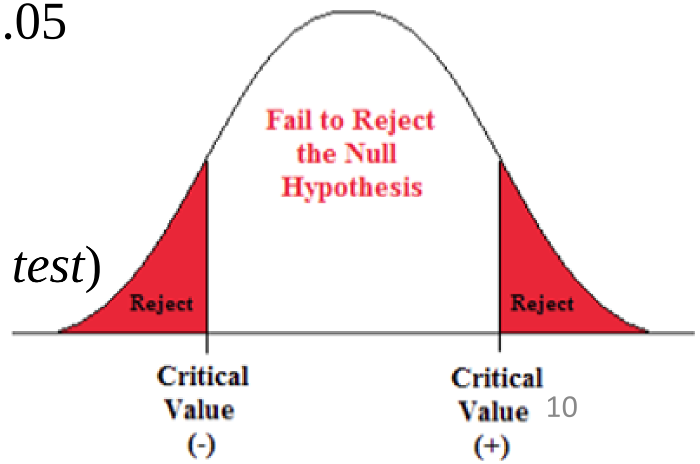
  will form, say, a 95% confidence interval for β_1 if c is defined by: $P\left(\left|t_{n-2}\right| \leq c\right) = .95$. (where t_{n-2} has a t distribution with (n-2) degrees of freedom).

SLR Inference: Hypothesis Testing

- ***The Null Hypothesis***: $H_0 : \beta_1 = 0$ (the most common Null Hypothesis in econometrics)
  - the t statistic (or ***t stat***) under H_0: $t\ stat = \dfrac{B_1 - 0}{se(B_1)} = \dfrac{B_1}{se(B_1)}$

    (the slope estimator divided by its standard error)
  - t stats can be positive or negative, and will always have the same sign as the $\hat{\beta}_1$ (since standard errors are always positive)
- ***The Hypothesis Test***: To conduct the test at, say, the 5% significance level:
  - ***Critical Value***: determine the critical value c defined by $P(|t_{n-2}| > c) = .05$

    (the *two-tailed* probability will be 5%)
  - ***Critical Region***: Reject $H_0 : \beta_1 = 0$ if $|t\ stat| = \left|\dfrac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right| > c$ (*two-tailed test*)



p values: Hypothesis tests the easy way

- ***The Null Hypothesis***: $H_0 : \beta_1 = 0$
- **The Test I**: Critical Value, c, defined by the significance level, α , and t_{n-2}

   a) Reject $H_0 : \beta_1 = 0$ if $|t\ stat| = \left|\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right| > c$; c is defined by $P(|t_{n-2}| > c) = \alpha$

- ***The p value***: *p value* = $P\left(|t_{n-2}| > |t\ stat|\right)$, where t_{n-2} is a random variable with a t distribution with (n-2) degrees of freedom

   a) The p value is just the probability in the tails (of the t_{n-2} distribution) outside $\pm tstat$.

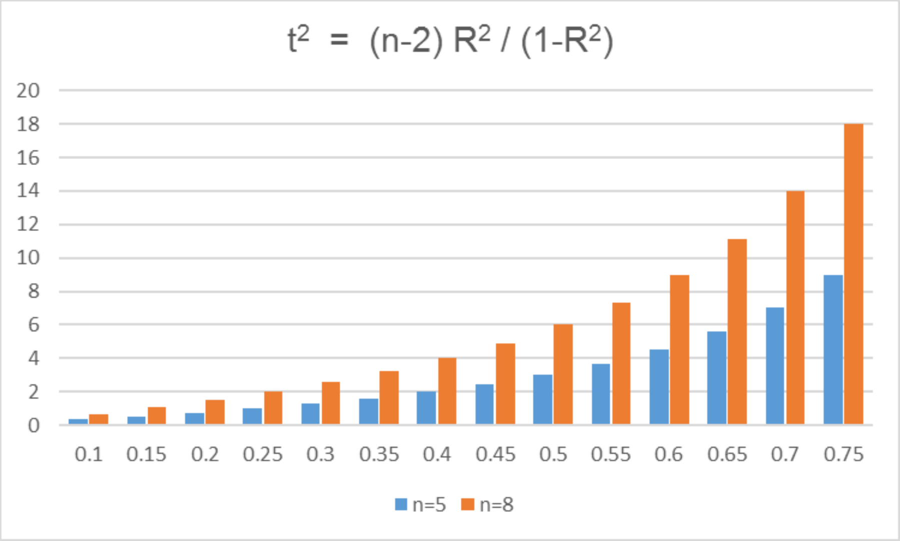
- **The Test II**: p Value

   a) Reject $H_0 : \beta_1 = 0$ if $P\left(|t_{n-2}| > |t\ stat|\right) = p < \alpha$, if the p-value is smaller than the significance level, α

   b) As in the case of the inference and the Sample Mean, you can reject the Null Hypothesis at all significance levels above the *p value*, but not at significance levels below the *p value*.

Convergence: SLR Assessment I & II
Who saw this coming?

- Goodness-of-Fit and Precision/Inference metrics converge in SLR models:
  - $t^2_{\hat{\beta}_1} = (n-2)\dfrac{R^2}{1-R^2} = (n-2)\dfrac{SSE}{SSR}$
- This expression is increasing in n and R^2, and so you hope that both n and R^2 are large.
- Since $SSE + SSR = SST$, the t stat reflects the division of SSTs between SSEs and SSRs… since $t^2_{\hat{\beta}_1}$ is proportional to $\dfrac{SSE}{SSR}$, for given n.
- The higher the SSE/SSR ratio, the greater the magnitude of the t stat.



An Example: Bodyfat

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      Brozek |        252    18.93849    7.750856          0       45.1
         hgt |        252    70.14881    3.662856       29.5      77.75

. corr Brozek hgt

             |   Brozek      hgt
-------------+------------------
      Brozek |   1.0000
         hgt |  -0.0891   1.0000

. corr Brozek hgt, covar

             |   Brozek      hgt
-------------+------------------
      Brozek |  60.0758
         hgt | -2.52975  13.4165

. reg Brozek hgt

      Source |       SS           df       MS      Number of obs   =       252
-------------+----------------------------------   F(1, 250)       =      2.00
       Model |  119.726679         1  119.726679   Prob > F        =    0.1585
    Residual |  14959.2899       250  59.8371598   R-squared       =    0.0079
-------------+----------------------------------   Adj R-squared   =    0.0040
       Total |  15079.0166       251  60.0757635   Root MSE        =    7.7354

------------------------------------------------------------------------------
      Brozek |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hgt |  -.1885553   .1332996    -1.41   0.158    -.4510886     .073978
       _cons |   32.16542   9.363495     3.44   0.001     13.72403    50.60681
------------------------------------------------------------------------------
```

- $Coef. = \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-2.53}{13.42} = \rho_{xy}\frac{S_y}{S_x} = -.0891\left(\frac{7.75}{3.66}\right) - .1885553$
- $Std.Err. = se(\hat{\beta}_1) = \frac{RMSE}{\sqrt{\sum(x_i - \bar{x})^2}} = \frac{RMSE}{S_x\sqrt{n-1}} = \frac{7.7354}{3.66\sqrt{251}} = .1332996$
- $t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{Coef.}{Std.\ Err.} = \frac{-.1886}{.1333} - 1.41$
- $P > |t|\,(p\ value) : P\left(\left|t_{250}\right| > \left|t\ stat\right|\right) = 0.158$
- [95% Conf. Interval]: $\left[Coef. \pm c \cdot Std.\ Err.\right] = [-.1886 \pm 1.97(.1333)\ ]$
  $= [-.4511, .0740]$ where $c = 1.97$ and $P\left(\left|t_{250}\right| \le c\right) = P\left(\left|t_{250}\right| \le 1.97\right) = .95$
- The *hgt* coefficient is statistically significant at the 15.9% level, but not at the 15% level, or any smaller level of statistical significance.
- Connecting t stats and R^2: The reported t stat for the *hgt* variable is -1.41.
  - $t_{\hat{\beta}_1}^2 = (n-2)\frac{R^2}{1-R^2} = 250\frac{.0079}{.9921} = 1.99 \ldots$ and so $\left|t_{\hat{\beta}_1}\right| = \sqrt{1.99} = 1.41$
  - $t_{\hat{\beta}_1}^2 = (n-2)\frac{SSE}{SSR} = 250\frac{119.727}{14,959} = 2.00 \ldots$ and so $\left|t_{\hat{\beta}_1}\right| = \sqrt{2.00} = 1.41$

Onwards to *MLR Estimation and Inference*