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

1. You may recall that 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.

2. We then discussed two approaches to SLR Assessment:

   a. ***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

   b. ***Precision/Inference*** metrics, which measured the precision with which we had estimated the unknown parameter values, β_0 and β_1.

3. At that time there was extensive discussion of Goodness-of-Fit metrics (SLR Assessment I)…. but we totally punted on precision/inference, saying:

   Later on, we will have lots to say about precision of estimation… but that discussion awaits the development of the tools of inference, including Confidence Intervals and Hypothesis Tests.

   While those inferential tools won't with certainty answer the question How Close?, they will give us probabilistic assessments as to how close our estimated coefficients are to the true unknown parameter values: levels of confidence for confidence intervals and significance levels for hypothesis testing.

4. ***Well that time has arrived!*** We will shortly consider precision/inference in the context of SLR models. But first, it is useful to review the case of the Sample Mean estimator.

Sample Means and Inference: Conceptual Review

5. Recall from the *Review of Inference* and the case of estimating the mean of the distribution:
   a. Under certain assumptions (including homoskedasticity) we found that the Sample Mean was a ***BLUE*** estimator of the unknown mean.
   b. To create confidence intervals or do hypothesis testing, we had to make an additional assumption about the distribution of the population. We assumed a Normal distribution.
   c. Under those assumptions:
      i. Confidence intervals were the *Sample Mean* plus or minus *c Standard Errors*, where the critical value c came from the t distribution with n-1 degrees of freedom.
      ii. We rejected the Null hypothesis ($H_0 : \mu = 0$) at some significance level α only if the reported *p value* was less than α … or if the t stat was larger in magnitude than the critical value.
6. These results carry over to the SLR models, virtually unchanged … just replace $(n-1)$ with $(n-2)$.

Onwards to SLR Inference

7. Recall the five SLR assumptions:
   a. **SLR.1 – *Linear model*** (the true model/DGM is in fact linear): $Y = \beta_0 + \beta_1 X + U$
   b. **SLR.2 – *Random sampling***: the sample $\{(x_i, y_i)\}$ is a random sample
   c. **SLR.3 – *Sample variation in the independent variable***: the x_i's are not all the same
   d. **SLR.4 – *Zero conditional mean of the error term***: $E(U \mid X = x) = 0$ for all x
   e. **SLR.5 – *Homoskedasticity*** (constant conditional variance of the error term): $Var(U \mid X = x) = \sigma^2$ for all x
8. Previously we showed:
   a. **LUEs**. Given SLR.1 – SLR.4, the OLS estimators are linear unbiased estimators of the true parameters of the DGM, β_0 and β_1, so that $E(B_0) = \beta_0$ and $E(B_1) = \beta_1$, where:
      i. $B_1 = \dfrac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_j - \bar{X})^2} = \dfrac{S_{XY}}{S_{XX}} = \dfrac{\sum (X_i - \bar{X}) Y_i}{\sum (X_j - \bar{X})^2}$ and,
      ii. $B_0 = \bar{Y} - B_1 \bar{X}$.

b. ***MSE* and *BLUE***. Adding in SLR.5 we have:

i. $\hat{\sigma}^2 = MSE = \dfrac{SSR}{n-2}$ is an unbiased estimator of σ^2, the conditional variance of *U*,

ii. $\dfrac{MSE}{\sum (x_i - \bar{x})^2}$ is an unbiased estimator of $Var(B_1)$, and most importantly,

iii. OLS estimators are BLUE estimators (the Best Linear Unbiased Estimators of β_0 and β_1). This last result is the **Gauss-Markov Theorem**.

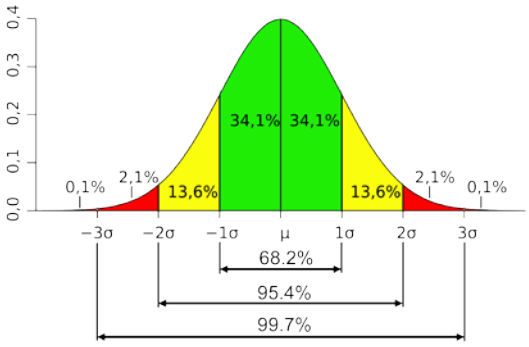
… SLR.6: U has a Normal Distribution

9. To create confidence intervals for the estimated parameters, or do hypothesis tests, we need to make one additional assumption (as we did previously for inference with the sample mean estimator above):

a. **SLR.6 – *Normality***: U is independent of the RHS variable X and is Normally distributed with mean 0 and variance σ^2.

i. 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.

ii. This may or may not be a good assumption… but it does simplify computations!



b. 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)

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

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

a. We'll skip the proof, but it follows from the fact that sums of independent normally distributed random variables are themselves normally distributed.

b. And as we did with the Sample Mean estimator, we can standardize B_1, so that:

$$\frac{B_1 - \beta_1}{sd(B_1)} \sim Normal(0,1),$$

where the standard deviation of B_1 is $sd(B_1) = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}$.

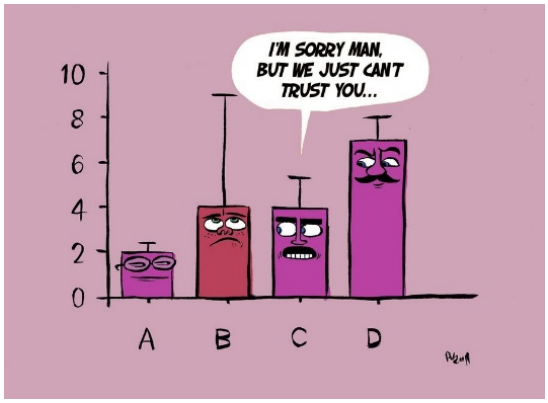
c. Since σ^2 is unknown, we don't know $sd(B_1)$. But as with the Sample Mean, we can use the standard error of B_1, $se(B_1)$ to estimate $sd(B_1)$.

d. Given SLR.1-SLR.5, MSE is an unbiased estimator can estimate σ^2, and $\frac{MSE}{\sum (x_i - \bar{x})^2}$ is an unbiased estimator of $Var(B_1)$ (conditional on the x's).

… *Standard Errors and t Stats*

Standard Errors

11. Standard errors (*se*'s) provide us with a measure of precision in the estimation of the unknown parameters. Knowing the *se* alone however is typically not very helpful, since it is often difficult to know whether a particular standard error is small or large. As with the Sample Mean estimator, we will circumvent this shortcoming by focusing on *t stats*, which effectively standardize the standard error, and gives us a metric that is more readily interpretable.



12. Recall that the ***standard error*** of B_1, $se(B_1)$, provides an estimate of $sd(B_1)$, is defined by: $se(B_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}}$

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

This is the **Std. Err.** that is reported in the SLR regression results.

13. Perhaps not surprisingly, the standard error is:

a. increasing in RMSE (reported standard errors will be smaller with models that do a better job of fitting the data),

b. decreasing in n (more observations will lead to smaller reported standard errors), and

c. decreasing in the variance of x (this is perhaps less intuitive, but increased variance in your RHS variable is a good thing and will lead to a smaller reported standard errors).

t-Stats

14. Comparing the standard error to the estimated coefficient, $\hat{\beta}_1$, often tells us something about how reliably we've estimated the unknown slope parameter, β_1. Before assessing reliability, though, we'll need to define one more term, the *t stat*:

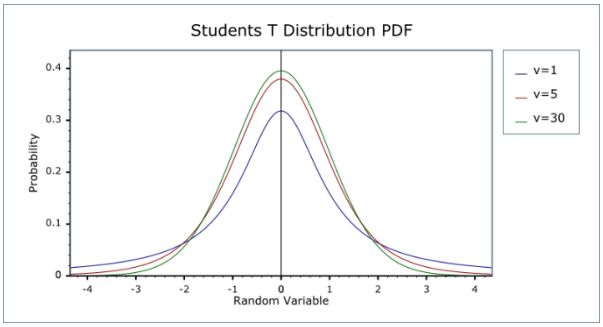
$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{se_{\hat{\beta}_1}} \ .$$

15. The absolute value of *t stat* tells you the magnitude of the estimated slope coefficient, $\hat{\beta}_1$, measured in units of standard errors. Once you know the *t stat*, you can apply some general rules of thumb to assess precision of estimation.
16. We'll be more precise below, but in general, the larger the t stat, the greater the likely precision (as you'll see later, n also matters in assessing precision)... so you should take comfort seeing high t stats, and fret over low ones. In terms of ranges and emotions, and assuming a sizable n:
    a. if $|t| > 2$ or so… then you have likely done a pretty good job of estimating the unknown slope parameter, β_1,

    b. if $|t| < 1$ ish … then you have likely done a not so good job of estimating β_1,
    c. and for in-between magnitudes of t… while the results aren't as strong as you might like, there's hope and reason to believe that with further work your model will be something to brag about. So definitely no reason to lose hope!

t Statistics and Inference

17. 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 degrees of freedom. Sometimes we write this as: $\frac{B_1 - \beta_1}{se(B_1)} \sim t_{n-2}$.



    a. This looks very similar to what we saw in the sample mean example, except in that case, we had the t distribution with n-1 degrees of freedom.
    b. Knowing the distribution of $\frac{B_1 - \beta_1}{se(B_1)}$ enables us to develop confidence intervals for β_1 and test hypotheses about β_1.
    c. And remember, as with the Sample Mean, ***the t statistic is the cornerstone of inference.***

… Confidence Intervals

18. 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: $P\left(\left|t_{n-2}\right| \le c\right) = .95$. (where t_{n-2} has a t distribution with (n-2) degrees of freedom).

    a. Notice that the confidence interval is centered around B_1, which will vary with the sample. Additionally, although c is fixed, the width of the interval, $2c \cdot se(B_1)$, will also vary with the sample, since $se(B_1)$ varies with the drawn sample.

    b. Many regression packages automatically report (95%?) confidence intervals for the different parameters.

… Hypothesis Testing

19. Testing $H_0 : \beta_1 = 0$**:** This is far and away the most common hypothesis test in econometrics. If the true slope parameter is 0 then changes in Y do not in general relate to changes in X, and so any apparent covariance is being driven solely by noise (U).

    a. From above, we know that the t statistic, $\frac{B_1 - \beta_1}{se(B_1)}$, has a t distribution with n-2 *dofs.*

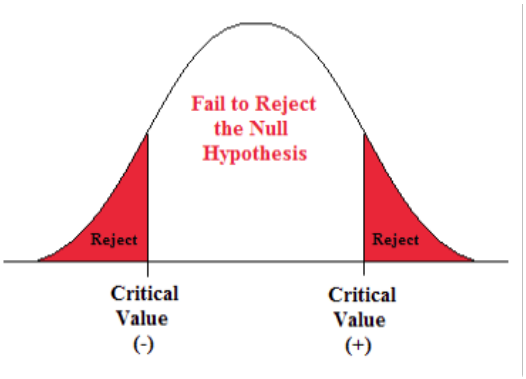
    b. Under the Null hypothesis, $H_0 : \beta_1 = 0$, the t statistic (or ***t stat***), $t\ stat = \frac{B_1 - 0}{se(B_1)} = \frac{B_1}{se(B_1)}$.

        i. t stats (assuming $H_0 : \beta_1 = 0$) are normally reported in regression output (as are parameter estimates and standard errors). They can be positive or negative, and will have the same sign as the $\hat{\beta}_1$ since standard errors are always positive.

    c. To conduct the test at the 5% significance level, determine the critical value c defined by: $P(\left|t_{n-2}\right| > c) = .05$.

        i. As before, unless n-2 is fairly small, this critical value will typically be about 2.

        ii. Critical Region: Reject $H_0 : \beta_1 = 0$ if $\left|t\ stat\right| = \left|\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right| > c$ (so the t statistic value is larger in magnitude than c).



        iii. If we reject we reject the null hypothesis, we say that *the coefficient is statistically significant at the 5% level.*

        iv. If we cannot reject the null hypothesis, we say that *the coefficient is not statistically significant at the 5% level*… or that it's *statistically insignificant at that level.*

… p values and Statistical Significance

20. As before, the *p value* is the smallest significance level at which the null hypothesis can be rejected:

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

    $t\ stat = \dfrac{\hat{\beta}_1}{se(\hat{\beta}_1)}$ for the given sample. (This is just the probability in the tails outside $\pm tstat$.)

    a. 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*. Small *p values* are evidence against the null hypothesis; large values not so much.

21. So, for significance level α, we Reject the Null Hypothesis, $H_0 : \beta_1 = 0$, if:

a. The t stat is larger in magnitude than the critical value: $\left|t\ stat\right| > c$, where the critical value c is defined by , $P(\left|t_{n-2}\right| > c) = \alpha$, or if

b. The p-value is smaller than the significance level: $P\left(\left|t_{n-2}\right| > \left|t\ stat\right|\right) = p < \alpha$

c. And yes, all of this is virtually identical to what we saw with the Sample Mean.

SLR Assessment Metrics *Converge*: t Stats and R^2

22. There's a connection between the measure of precision, $t_{\hat{\beta}_1}$, and the R^2 measure of goodness of fit, as well as SSE and SSR:

$$t^2_{\hat{\beta}_1} = (n-2)\frac{R^2}{1-R^2} = (n-2)\frac{SSE}{SSR}.$$

Who knew? …Goodness-of-Fit and Precision/Inference metrics are connected in SLR models!

23. These equations make it clear that precision in estimation is a function of both R^2, how well the model fits the data, as well and the number of observations, n. It may not be so obvious, but this expression is increasing in n and R^2. And so ideally, both n and R^2 are large.

24. Importance of *n and* R^2: If you have high R^2 but low n, or high n (lots of observations) but poor fit (low R^2), then it's likely that your slope estimate is not so precise. But a healthy R^2 together with lots of observations means that that you have likely done a nice job estimating the unknown parameter, β_1. So:

    a. low n and low R^2: bad news… *get back to work!*

    b. (low n and high R^2) or (high n and low R^2)… *still not so great!*

    c. high n and high R^2: *well done!*

25. Note that since $SSE + SSR = SST$, the t stat will depend on how the SSTs are divided between SSEs and SSRs, since $t^2_{\hat{\beta}_1}$ will be proportional to $\frac{SSE}{SSR}$, for given n. The higher the SSE/SSR ratio, the greater the magnitude of the t stat.

Example: Bodyfat

26. Here's an example using the bodyfat dataset.

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

a. $Coef. = \hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \rho_{xy}\dfrac{S_y}{S_x} = -.1885553$

b. $Std.Err. = se(\hat{\beta}_1) = \dfrac{RMSE}{\sqrt{\sum (x_i - \bar{x})^2}} = \dfrac{RMSE}{S_x\sqrt{n-1}} = .1332996$

c. $t = \dfrac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \dfrac{Coef.}{Std.\,Err.} = -1.41$

d. $P > |t|\,(p\ value) : P\left(|t_{250}| > |t\ stat|\right) = 0.158$

e. [95% Conf. Interval]: $\left[Coef. \pm c \cdot Std.\,Err.\right] = [-.4510886\ ,\ .073978]$ where $P\left(|t_{250}| \le c\right) = .95$

f. The *hgt* coefficient is statistically significant at the 15.9% level, but not at the 15% level, or any smaller level of statistical significance.

g. Connecting t stats and R^2: The reported t stat for the *hgt* variable is -1.41. Applying the formulas above, we have:

i. $t^2_{\hat{\beta}_1} = (n-2)\dfrac{R^2}{1-R^2} = 250\dfrac{.0079}{.9921} = 1.99\ldots$ and so $\left|t_{\hat{\beta}_1}\right| = \sqrt{1.99} = 1.41$

ii. $t^2_{\hat{\beta}_1} = (n-2)\dfrac{SSE}{SSR} = 250\dfrac{119.727}{14,959} = 2.00\ldots$ and so $\left|t_{\hat{\beta}_1}\right| = \sqrt{2.00} = 1.41$

Appendix

27. Proof of the relationship between Goodness-of-Fit and precision/Inference metrics:

$$t^2_{\hat{\beta}_1} = (n-2)\frac{R^2}{1-R^2}$$

a. By definition, $t^2_{\hat{\beta}_1} = \dfrac{\hat{\beta}_1^2}{se^2_{\hat{\beta}_1}} = \dfrac{\hat{\beta}_1^2 \sum (x_i - \overline{x})^2}{MSE} = \dfrac{\hat{\beta}_1^2 \sum (x_i - \overline{x})^2}{SSR/(n-2)}$.

b. We know from the proof of $\rho^2_{xy} = R^2$ that $SSE = \hat{\beta}_1^2 \sum (x_i - \overline{x})^2$.

c. And so $t^2_{\hat{\beta}_1} = (n-2)\dfrac{SSE}{SSR} = (n-2)\dfrac{SSE/SST}{SSR/SST} = (n-2)\dfrac{R^2}{1-R^2}$.