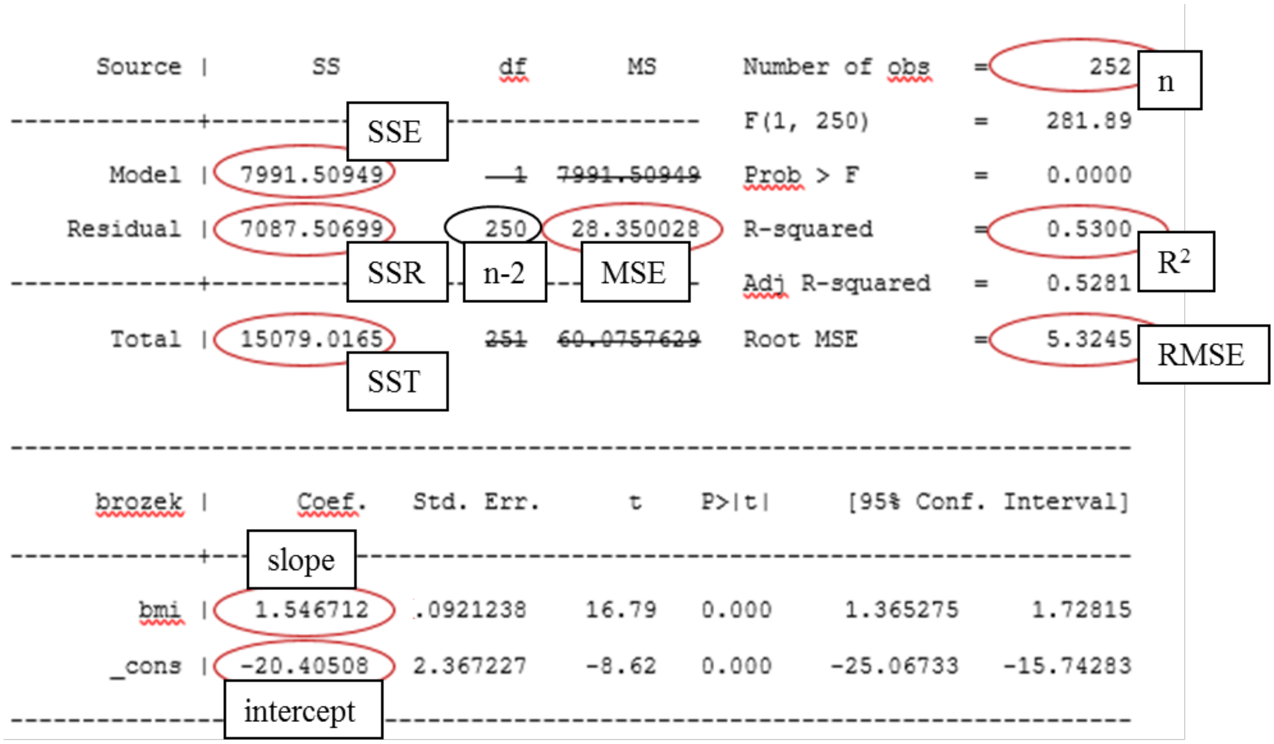
A Quick Review of OLS\SLR Analytics and Assessment

Here are some SLR regression results using the *bodyfat* dataset:

```
. reg brozek bmi
```



```
      Source |       SS           df       MS      Number of obs   =       252
-------------+----------------------------------   F(1, 250)       =    281.89
       Model |  7991.50949       ~~1~~  ~~7991.50949~~   Prob > F        =    0.0000
    Residual |  7087.50699       250  28.350028    R-squared       =    0.5300
-------------+----------------------------------   Adj R-squared   =    0.5281
       Total |  15079.0165       ~~251~~  ~~60.0757629~~   Root MSE        =    5.3245

------------------------------------------------------------------------------
      brozek |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   1.546712   .0921238    16.79   0.000     1.365275     1.72815
       _cons |  -20.40508   2.367227    -8.62   0.000    -25.06733   -15.74283
------------------------------------------------------------------------------
```

(*brozek* is the y variable and *bmi* is the x variable)

1) The estimated OLS/SLR coefficients

   a) OLS: Minimize $SSR = \sum (u_i)^2 = \sum \left(y_i - \left(b_0 + b_1 x_i \right) \right)^2$ $wrt\ b_0\ and\ b_1$ (FOCs and SOCs)

   b) Slope coefficient (*bmi*):

      i) $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = \rho_{xy} \frac{S_y}{S_x} = 1.546712$,

      ii) Also a weighted average of the slopes of lines connecting the different data points to the sample means $\left(\bar{x}, \bar{y} \right)$, so $slope_i = \left[\frac{(y_i - \bar{y})}{(x_i - \bar{x})} \right]$:

      $$(1)\ \ \hat{\beta}_1 = \sum_i w_i \left[\frac{(y_i - \bar{y})}{(x_i - \bar{x})} \right] \text{, where } w_i = \frac{(x_i - \bar{x})^2}{(n-1)S_{xx}} = \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2} \geq 0\ and \sum w_i = 1$$

   c) Intercept coefficient (*_cons*): $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -20.40508$

2) SRF (Sample Regression Function; *predicteds*): $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -20.41 + 1.55x$

   a) average marginal effect: $\frac{d\hat{y}}{dx} = \hat{\beta}_1 = 1.55$

   b) elasticity@means: $\varepsilon = \frac{d\hat{y}}{dx}\frac{\overline{x}}{\overline{y}} = \hat{\beta}_1 \frac{\overline{x}}{\overline{y}} = 1.546712 \frac{25.4369}{18.93849} = 2.08$, since

```
. summ brozek bmi

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      brozek |        252    18.93849    7.750856          0       45.1
         bmi |        252     25.4369    3.648111       18.1       48.9
```

3) The ANOVA table (w/ a constant (intercept) term in the model)

   a) SST: Total Sum of Squares … $\sum (y_i - \overline{y})^2 = (n-1)S_{yy} = 15,079.$

   b) SSE: Explained Sum of Squares … $\sum (\hat{y}_i - \overline{y})^2 = (n-1)S_{\hat{y}\hat{y}} = 7,992$, since $\frac{1}{n}\sum \hat{y}_i = \overline{y}$

   c) SSR: Residual Sum of Squares… $\sum \hat{u}_i^2 = \sum (y_i - \hat{y}_i)^2 = (n-1)S_{\hat{u}\hat{u}} = 7,088$, since $\frac{1}{n}\sum \hat{u}_i = 0$

   d) Result: SST = SSE + SSR … or $S_{yy} = S_{\hat{y}\hat{y}} + S_{\hat{u}\hat{u}}$

      i) The variance of the y's is the sum of the variances of the predicteds and of the residuals.

      ii) Also recall that the residuals will be uncorrelated with the predicteds.

4) *Goodness-of-Fit* metrics: MSE/RMSE and R^2

   a) (Root) Mean Squared Error:

      i) $MSE = \frac{SSR}{n-2} = \frac{7087.50699}{250} = 28.350028$

         (1) sort of an average squared residual… sort of, but not quite since dividing by n-2, and not n

         (2) also sort of the sample variance of the residuals… again, sort of, but not quite since dividing by n-2, and not n-1

ii) $RMSE = \sqrt{MSE} = \sqrt{\dfrac{SSR}{n-2}} = \sqrt{28.350028} = 5.3245$

(1) sort of an average residual… sort of, but not quite… more like the square root of sort of an average squared residual… sort of, but not quite …

b) R^2: Coefficient of Determination:

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{7087.50699}{15079.0165} = 0.5300$$

$$R^2 = \frac{SSE}{SST} = \frac{7991.50949}{15079.0165} = 0.5300$$

i) The percentage of the variance of the dependent variable *explained* by the predicteds… which is to say, explained by the model

(1) $R^2 = \dfrac{S_{\hat{y}\hat{y}}}{S_{yy}} = \dfrac{31.8387}{60.0768} = 0.5300$, since …

```
. corr Brozek yhat, covar
(obs=252)

             |   Brozek     yhat
-------------+------------------
      Brozek |  60.0758
        yhat |  31.8387  31.8387
```

ii) R^2 also effectively reflects the correlation between the dependent variable y and the independent variable x

(1) $R^2 = \rho_{xy}^2 = \rho_{\hat{y}y}^2 = .7280^2 = 0.5300$, since …

```
. corr Brozek BMI yhat
(obs=252)

             |   Brozek      BMI     yhat
-------------+---------------------------
      Brozek |   1.0000
         BMI |   0.7280   1.0000
        yhat |   0.7280   1.0000   1.0000

. di .7280^2
.529984
```