Proof of SST=RSS+SSE

For a multivariate regression, suppose we have n observed variables $y_1, y_2, \cdots y_n$ predicted by n observations of k-tuple explanatory variables. Let $x_{i,j}, i \in \{1, \cdots, n\}, j \in \{1, \cdots, k\}$ be the i-th observation of the j-th explanatory variable.

The predicting equation for y_i is given by

$$y_i = x_{i,1} \cdot \beta_1 + x_{i,2} \cdot \beta_2 \cdots + x_{i,k} \cdot \beta_k + 1 \cdot \beta_0 + \varepsilon_i, i \in \{1, \cdots, n\}$$

where ε_i is the i-th error term.

If we put everything in a matrix form, i.e., let $\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$ and $\boldsymbol{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,k} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,k} \end{bmatrix}$ and $\boldsymbol{\beta}_0 = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_0 \end{bmatrix}$ and $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ (vector/matrix will be written in bold form), then we can get the predicting equation by

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$$

For the ordinary least squares estimation, we want to minimize sum of squared errors SSE, that is, the objective function is $\boldsymbol{\varepsilon^T \varepsilon}$. If we substitute the above equation to the SSE formula, we get the target optimization problem represented by

$$\min_{\boldsymbol{\beta}, \boldsymbol{\beta}_0} \{\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} = \boldsymbol{Y} - (\boldsymbol{X\beta} + \boldsymbol{\beta}_0)\}$$

$$= \min_{\boldsymbol{\beta}, \boldsymbol{\beta}_0} (\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{\beta}_0)^T (\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{\beta}_0)$$

Okay, let's recall the first order partial derivative in a matrix form, you can expand and verify the rules below in its scalar form.

If $\boldsymbol{W}$ is symmetric,

Rule #1: $(\boldsymbol{\beta}^T \boldsymbol{X})' = \boldsymbol{\beta}$, $(\boldsymbol{WX})' = \boldsymbol{W}$

Rule #2: $(\boldsymbol{X}^T \boldsymbol{WX})' = 2\boldsymbol{WX}$

In the special case for Rule #2 when $\boldsymbol{W} = \boldsymbol{I}$, $(\boldsymbol{X}^T\boldsymbol{X})' = 2\boldsymbol{X}$

Therefore, for this continuous function of SSE, the first order necessary optimality condition is given by$(\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon})' = 0$, that is, by the chain rule,

$$2\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{\beta}_0) = \boldsymbol{0}$$

Actually we can combine β_0 with the rest of k betas as $\boldsymbol{\beta} = \begin{bmatrix}\beta_0 \\ \vdots \\ \beta_k\end{bmatrix}$ and $\boldsymbol{X}_{\boldsymbol{n}\times(\boldsymbol{k}+\boldsymbol{1})} =$ $\begin{bmatrix}1 \\ \vdots \\ 1\end{bmatrix} || \begin{bmatrix}x_{1,1} & \cdots & x_{1,k} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,k}\end{bmatrix} = \begin{bmatrix}1 \, x_{1,1} & \cdots & x_{1,k} \\ \vdots \; \vdots & \ddots & \vdots \\ 1 \, x_{n,1} & \cdots & x_{n,k}\end{bmatrix}$, then the objective function can be re-written as

$$\min_{\boldsymbol{\beta}}\{\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X\beta}\}$$

$$= \min_{\boldsymbol{\beta}}(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta})$$

The optimality condition now becomes

$$\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X\beta}) = \boldsymbol{0}$$

Hence, the optimal $\boldsymbol{\beta}$ satisfies $\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}$**,** thus we can get

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^T\boldsymbol{Y}$$

and

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

where $(\boldsymbol{X}^T\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^T$ is called the left pseudo inverse of $\boldsymbol{X}$.

Note that for a simple regression (one explanatory variable), above reduces to

$$\beta_1 = \frac{cov(x, y)}{var(x)}$$

To see this, we write out the variables in their explicit form.

$$\boldsymbol{X}_{n\times 2} = \begin{bmatrix}1 & x_1 \\ \vdots & \vdots \\ 1 & x_n\end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix}y_1 \\ \vdots \\ y_n\end{bmatrix}$$

We get

$$\widehat{\boldsymbol{\beta}}_{2\times 1} = \begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^T\boldsymbol{Y}$$

$$= \left(\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

Bear in mind that we have

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

We can get

$$\beta_1 = \frac{n \sum x_i y_i - \sum x_i \cdot \sum y_i}{n \sum x_i^2 - \sum x_i \cdot \sum x_i} = \frac{cov(x, y)}{var(x)}$$

We now focus on proving

$$SST = RSS + SSE$$

The total sum of squares (SST) is given by

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = (\boldsymbol{Y} - \overline{\boldsymbol{Y}})^T (\boldsymbol{Y} - \overline{\boldsymbol{Y}})$$

$$= \boldsymbol{Y}^T \boldsymbol{Y} + \overline{\boldsymbol{Y}}^T \overline{\boldsymbol{Y}} - \boldsymbol{2} \boldsymbol{Y}^T \overline{\boldsymbol{Y}}$$

The sum of squared errors (SSE), a.k.a. sum of squared residuals (SSR), is given by

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \left(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\right)^T \left(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\right)$$

$$= \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

$$= \boldsymbol{Y}^T \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) - \widehat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

$$= \boldsymbol{Y}^T \boldsymbol{Y} - \boldsymbol{Y}^T \boldsymbol{X} \widehat{\boldsymbol{\beta}}$$

The regression sum of squares (RSS), a.k.a. explained sum of squares (ESS), is given by

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \left(\widehat{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}\right)^T \left(\widehat{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}\right)$$

$$= \left(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{Y}}\right)^T \left(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{Y}}\right)$$

$$= \widehat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{X} \widehat{\boldsymbol{\beta}} + \overline{\boldsymbol{Y}}^T \overline{\boldsymbol{Y}} - 2 \widehat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \overline{\boldsymbol{Y}}$$

Therefore,

$$SST - RSS - SSE$$
$$= \boldsymbol{Y}^T\boldsymbol{Y} + \overline{\boldsymbol{Y}}^T\overline{\boldsymbol{Y}} - \mathbf{2}\boldsymbol{Y}^T\overline{\boldsymbol{Y}} - \boldsymbol{Y}^T\boldsymbol{Y} + \boldsymbol{Y}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{Y}}^T\overline{\boldsymbol{Y}} + 2\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\overline{\boldsymbol{Y}}$$
$$= 2\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\overline{\boldsymbol{Y}} - \mathbf{2}\boldsymbol{Y}^T\overline{\boldsymbol{Y}} + \boldsymbol{Y}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

where

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^T\boldsymbol{Y}$$

We see that

$$\boldsymbol{Y}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}$$
$$= \boldsymbol{Y}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\left(\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$
$$= \boldsymbol{Y}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y}$$
$$= \boldsymbol{Y}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Y}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = 0$$

It suffices to prove that

$$2\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\overline{\boldsymbol{Y}} - \mathbf{2}\boldsymbol{Y}^T\overline{\boldsymbol{Y}} = 0$$

to get $SST = RSS + SSE$.

We may ask is this true in general??? No! But we do have assumptions when we conduct OLS regression.

Remember the moment restriction for a simple linear OLS regression.

- $E(y - b_0 - b_1x) = 0$
- $E[x(y - b_0 - b_1x)] = 0$

The expected value of the error term should be zero and the error term should be uncorrelated with the explanatory variables.

$$\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\overline{\boldsymbol{Y}} - \boldsymbol{Y}^T\overline{\boldsymbol{Y}} = -\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\overline{\boldsymbol{Y}} = -\boldsymbol{\varepsilon}^T\overline{\boldsymbol{Y}} = -\bar{y}\boldsymbol{\varepsilon}^T\boldsymbol{e} = 0$$

where $\boldsymbol{e}_{n\times 1} = \begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}$.

If the assumption that the expected value of the residual term is zero is violated, then

SST≠RSS+SSE

Classical assumptions for regression analysis include:

- The sample is representative of the population for the inference prediction.
- The error is a random variable with a mean of zero conditional on the explanatory variables.
- The independent variables are measured with no error. (Note: If this is not so, modeling may be done instead using errors-in-variables model techniques).
- The predictors are linearly independent, i.e. it is not possible to express any predictor as a linear combination of the others.
- The errors are uncorrelated, that is, the variance–covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- The variance of the error is constant across observations (homoscedasticity). If not, weighted least squares or other methods might instead be used.

Reference

Matrix Calculus in Wikipedia @ http://en.wikipedia.org/wiki/Matrix_calculus

CFA print curriculum Level 2, 2014

ESS in Wikipedia@ http://en.wikipedia.org/wiki/Explained_sum_of_squares