DSL 810 (Data Driven Design)

Instructor: Dr Jay Dhariwal, Dept of Design, IIT Delhi

In this tutorial, we would like to relate the simple linear regression output from MATLAB for an example with the math for better understanding.

The example that would be considered is what factors affect life expectancy? This is the Dataset for Life Expectancy predictors and response. Some details on this dataset.

Simple Linear Regression

If we consider the simple linear regression model to be:

$$y = \beta_0 + \beta_1 x + \in$$

This model would have one predictor and one response. For the above dataset on Life Expectancy data, we would consider Environmental Performance Index (EPI) as the predictor and Life Expectancy (in years) as the response.

Carrying out this computation in MATLAB:

```
% MATLAB code
load('dataset_life_expectancy.mat');

dataset_life_expectancy_onevar=dataset_life_expectancy(:,["EPI","LifeExpectancy"]
);
mdl = fitlm(dataset_life_expectancy_onevar)
```

MATLAB output:

```
mdl = Linear regression model: LifeExpectancy ~ 1 + EPI

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     42.941       1.8079    23.752    1.3006e-65
    EPI            0.50654     0.031506    16.078    3.0474e-40

Number of observations: 248, Error degrees of freedom: 246
Root Mean Squared Error: 5.42
R-squared: 0.512,  Adjusted R-Squared: 0.51
```

F-statistic vs. constant model: 258, p-value = 3.05e-40

The p-value of the intercept and the predictor “EPI” is significant. The simple linear regression model is:

$y = 42.941 + 0.507 \times EPI$

Working out the math for this computation:

The fitted simple linear regression model is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \; ; \quad S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Based on excel computation,

$\sum x_i = 13971$, $\sum y_i = 17726.2$,

$\sum x_i y_i = 1013558$, $\sum x_i^2 = 816596$

$$S_{xx} = 816596 - \frac{(13971)^2}{248} = 29585.2$$

$$S_{xy} = 1013558 - \frac{17726.2 \times 13971}{248} = 14986.10$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{14986.10}{29585.2} = 0.507$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{17726.2}{248} - 0.507 \times \frac{13971}{248}$$

$$\hat{\beta}_0 = 42.941$$

Hence, the least squares fit is:-

$$\hat{y} = 42.941 + 0.507x$$

Computation of R^2:

Coefficient of determination, $R^2 = \frac{SS_R}{SS_T}$

or $R^2 = 1 - \frac{SS_{Res}}{SS_T}$

where $SS_T = SS_R + SS_{Res}$ (from ANOVA)

$$SS_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 1281820 - 248 \times \left(\frac{17726.2}{248}\right)^2$$

$$SS_T = 14815.406$$

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy} = 14815.406 - 0.507 \times 14986.10$$

$SS_{Res} = 7224.331$ or $R^2 = 1 - \frac{7224.331}{14815.406} = 0.512$

Root mean squared error $(MS_{Res})^{1/2} = \sqrt{\frac{SS_{Res}}{n-2}}$

$$= \sqrt{\frac{7224.331}{248-2}} = 5.419$$