**When and why to standardize a variable**

* Variable Standardization is one of the most important concept of predictive modeling. It is a preprocessing step in building a predictive model.
* Standardization is also called Normalization and Scaling.

**Standardization / Scaling:**

* The concept of standardization comes into picture when continuous independent variables are measured at different scales.
* It means these variables do not give equal contribution to the analysis.
* **The idea is to rescale an original variable to have equal range and/or variance.**

**Methods of Standardization / Normalization:**

* Z Score
* Min-Max Scaling
* Standard Deviation Method
* Range Method

**Z score**

Z score standardization is one of the most popular method to normalize data. In this case, we rescale an original variable to have a **mean of zero** and **standard deviation of one**.

|  |
| --- |
|  |
| Z score |

Mathematically, scaled variable would be calculated by subtracting mean of the original variable from raw vale and then divide it by standard deviation of the original variable.

**Min-Max Scaling**

It is also called **0-1 scaling** because the standardized value using this method lies between 0 and 1.

**Standard Deviation Method**  
  
In this method, we divide each value by the standard deviation. The idea is to have **equal variance**, but different means and ranges.

**Range Method**  
  
In this method, we dividing each value by the range. **Formula :** x /(max(x) - min(x)). In this case, the means, variances, and ranges of the variables are still different, but at least the ranges are likely to be more similar.

**What is Centering?**  
Centering means subtracting a constant value from every value of a variable. The constant value can be average, min or max. Most of the times we use average value to subtract it from every value.

By default, scale() function with center=TRUE subtract mean value from values of a variable.

**When it is important to standardize variables?**  
1. It is important to standardize variables before running **Cluster Analysis**. It is because cluster analysis techniques depend on the concept of measuring the distance between the different observations we're trying to cluster. If a variable is measured at a higher scale than the other variables, then whatever measure we use will be overly influenced by that variable.  
  
2. Prior to **Principal Component Analysis**, it is critical to standardize variables. It is because PCA gives more weightage to those variables that have higher variances than to those variables that have very low variances. In effect the results of the analysis will depend on what units of measurement are used to measure each variable. Standardizing raw values makes equal variance so high weight is not assigned to variables having higher variances.  
  
3. It is required to standardize variable before using **k-nearest neighbors** with an Euclidean distance measure. Standardization makes all variables to contribute equally.  
  
4. All SVM kernel methods are based on distance so it is required to scale variables prior to running final **Support Vector Machine** (**SVM)**model.  
  
5. It is necessary to standardize variables before using **Lasso and Ridge Regression.**Lasso regression puts constraints on the size of the coefficients associated to each variable. However, this value will depend on the magnitude of each variable. The result of centering the variables means that there is no longer an intercept. This applies equally to ridge regression.  
  
6. In regression analysis, we can calculate **importance of variables** by ranking independent variables based on the descending order of absolute value of standardized coefficient.  
  
7. In regression analysis, when an interaction is created from two variables that are not centered on 0, some amount of collinearity will be induced. **Centering first addresses this potential problem**. In simple terms, having non-standardized variables interact simply means that when X1 is big, then X1X2 is also going to be bigger on an absolute scale irrespective of X2, and so X1 and X1X2 will end up correlated.  
  
8. In regression analysis, it is also helpful to standardize a variable when you include power terms X². Standardization removes collinearity.

**When it is not required to standardize variables**  
1. If you think model performance of linear regression model would improve if you standardize variables, **it is absolutely incorrect!** It does not change RMSE, R-squared value, Adjusted R-squared value, p-value of coefficients. It shows standardization does not affect model performance at all.

**Compare Coefficients, R-Squared and Adjusted R-Squared**

The value of coefficients are not same when we run regression analysis with and without standardizing independent variables. **It does not mean they are affected by scaling / standardization.** The values are different because of these are the slopes - how much the target variable changes if you change independent variable by 1 unit. In other words, standardization can be interpreted as scaling the corresponding slopes. **The adjusted r-squared and multiple r-squared value is exactly same.**

https://www.listendata.com/2017/04/how-to-standardize-variable-in-regression.html