Study several factors that affect the performance and interpretation of a simple model, such as Linear Regression Analysis. The factors that will be discussed include

* Model mismatch,
* The presence of outliers in the data,
* The presence of hidden confounders or of selection bias,
* The presence of correlations between the covariates (multicollinearity).

You will characterize how these factors

1. Affect performance and the interpretation of the parameters in the model.

2. You will examine different solutions to eliminate or mitigate these effects (e.g. normalization, or transformation of the covariates).

The presence of outliers in the data

*Data points that are noticeably diverge from the rest of the data are called ‘Outliers’. Graphically, there might be only either an extreme x value or an extreme y value, or both. A point that is too far from the rest of the data points without having an extreme x or an extreme y value could also be an Outlier.*

Affect on the performance of the model and the interpretation of the parameters in the model:

The presence of outliers specially the influential points can dramatically impact the magnitude as well as the direction of regression coefficients.

Consequently, the effect of the presence of Outliers on the model’s performance is straight forward, that is, it can completely change the model’s equation resulting in a bad prediction.

The effect of outliers on the regression model may be easily checked by dropping these points and refitting the regression equation.

Generally, the impacts of the Outliers on the model’s performance can be following:

* Increase in error variance
* Decrease the normality of the data
* Outliers can impact against the basic assumptions a model has, including Linear Regression.

Solutions to mitigate or eliminate these effects:

*Even with a thorough understanding of the data, outliers can be hard to define. [...]Great care should be taken not to hastily remove or change values, especially if the sample size is small.*

*Page 33, Applied Predictive Modeling, 2013.*

Outliers should always be carefully investigated for the reasons of their presence. If there is some fault behind them, it should be corrected or as a last resort, should be deleted. Deleting or altering an outlier for a sake for better model performance could give user a false sense of precision.

In a nutshell, the following approaches could be taken to handle the outliers once their adverse effects have been confirmed.

Using Standardization techniques:

Certain standardization techniques such as Robust Scaling. Robust Scaling removes the median and scales the data according to the quantile and retains some effect of outliers.

Removal of Outliers:

Removal of the Outliers using Standard Deviation or Inter Quartile Range. Usually a suitable threshold is choosen and used to define the bounds of 'non-outliers' or acceptable values.

Binning:

In cases where a feature could be binned and the outliers could be justified to be a part of a bin, binning technique can be a remedy to lessen the impact of outliers.

The presence of correlations between the covariates (multicollinearity).

*Multi Collinearity means independent variables are highly correlated to each other. It is an assumption in Regression analysis that the multi collinearity doesn’t exist in the model*

Effect on the performance of the model and the interpretation of the parameters in the model:

This phenomenon can have effects on the extra sums of squares, fitted values and predictions, regression coefficients, and many other parts of multiple linear regression. The coefficients become very sensitive to small changes in the model. Also, Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of your regression model

Multicollinearity may not affect the accuracy of the model as much but we might lose reliability in determining the effects of individual independent features on the dependent feature in model and that can be a problem when we want to interpret the model.

Solutions to mitigate or eliminate these effects:

Removal of one of the highly correlated independent variables from the data. This can be achieved by manually checking the Correlation Coefficients or with calculating Variance Inflation Factor (VIF) and removing the one with higher VIF.

Dimensionality reduction using Principal Components Analysis (PCA).

Ridge and Lasso Regression– This is an alternative estimation procedure to ordinary least squares. Penalizes for the duplicate information and shrinks or drops to zero the parameters of a regression model.

By standardizing the variables i.e. reducing the multi collinearity by ‘centering’ the variables.