**Linear regression:**

The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

**Assumptions:**

* Linear relationship;

Can best be tested with scatter plots;

Solve: other model/high order terms (x2)

* Multivariate normality/all variables normal distribution;

Test: Q-Q-plot, Kolmogorov-Smirnov test;

Solve: non-linear transformation (log-transformation)

* No or little multicollinearity;

Test: 1) Correlation matrix: the matrix of Pearson’s Bivariate Correlation (correlation coefficients) among all independent variables; 2) Variance Inflation Factor (VIF): With VIF > 10 there is an indication that multicollinearity may be present; The square root of the VIF indicates how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the model. If VIF= 5.27 (√5.27 = 2.3), standard error for the coefficient of that predictor variable is 2.3 times as large as it would be if that predictor variable were uncorrelated with the other predictor variables.

Solve: remove independent variables with high VIF/Pearson’s values; add regularization(L1/Lasso can do feature selection work)

* No auto-correlation

Test: scatterplot; Durbin-Watson’s d tests: d between 0 and 4, around 2 indicate no autocorrelation. As a rule of thumb values of 1.5 < d < 2.5 show that there is no auto-correlation in the data.

* Homoscedasticity: the residuals are equal across the regression line

Test: scatterplot; Goldfeld-Quandt Test

**Co-efficient:**

1. Normal equation

Complexity of the computation will increase as the number of features increase. It gets very slow when number of features grow large.

2. Optimizing using gradient descent

The partial derivative of the cost function with respect to the parameter can give optimal coefficient value.

Iterating through different values of slope and intercept can yield different error values. Out of all values, there will be one point where error value will be minimum and parameters corresponding to this value will yield the optimal solution.

**Metrics for model evaluation:**

1. R-Squared value

This value ranges from 0 to 1. Value ‘1’ indicates predictor perfectly accounts for all the variation in Y. Value ‘0’ indicates that predictor ‘x’ accounts for no variation in ‘y’.

2. Mean squared error (MSE, loss function)/ Root Mean Squared Error (RMSE)

The average of the squares of the errors/the square root of the mean square error, same units as the quantity plotted on the vertical axis.

**Null-Hypothesis and P-value:**

The null hypothesis is that the coefficient is equal to zero (no effect).

Low P-value (<0.05): Rejects null hypothesis- predictor is related to the response

High P-value: Changes in predictor are not associated with change in target

**Categorical feature:**

One-hot-encoder: uniform the distance between different categories; too many features, clustering first or tree-based model (no need One-hot-encoder, based on entropy)

‘Other’ category (dump feature)