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 (x^2)

- 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 ($\sqrt{5.27} = 2.3$), 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:

In statistics, ordinary least squares (OLS) is a type of **linear least square**s method for estimating the **unknown parameters** in a **linear regression mode**l.
OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of **least squares**: **minimizing the sum of the squares** of the differences between the **observed dependent variable** (values of the variable being predicted) in the given dataset and **those predicted** by the linear function.

1. Normal equation

https://eli.thegreenplace.net/2014/derivation-of-the-normal-equation-for-linear-regression/
Complexity of the computation will increase as the number of features increase. It gets very slow when number of features grow large.

$$\text{Theta} = (X^T X)^{-1} X^T Y$$

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 / **coefficient of determination / goodness of fit**

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)

Loss function: least-squares cost

- Given the hypothesis function:

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n \qquad h_\theta(x) = \theta^T x \quad \begin{pmatrix} \theta_0 \\ \theta_1 \\ \cdots \\ \theta_n \end{pmatrix} \in \mathbb{R}^{n+1}$$

(1 * (n+1))

- Minimize the least-squares cost/mean square error:

$$J(\theta_{0\ldots n}) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \qquad J(\theta) = \frac{1}{2m}(X\theta - y)^T(X\theta - y)$$

Where $x^{(i)}$ is the i-th sample (from a set of *m* samples) and $y^{(i)}$ is the i-th expected result. For matrix notation. The X(uppercase X) is a matrix of m rows/ n+1 columns, in which each row is the i-th sample (the vector $x^{(i)}$).

Normal equation

https://eli.thegreenplace.net/2015/the-normal-equation-and-matrix-calculus/

- Throw the $\frac{1}{2m}$ part

$$J(\theta) = (X\theta)^T X\theta - (X\theta)^T y - y^T(X\theta) + y^T y$$
$$J(\theta) = \theta^T X^T X\theta - 2(X\theta)^T y + y^T y$$

- To find where the above function has a minimum, we will derive by θ and compare to 0.

$$\frac{\partial J}{\partial \theta} = 2X^T X\theta - 2X^T y = 0$$
$$\theta = (X^T X)^{-1} X^T y$$

Gradient descent

repeat until convergence {

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

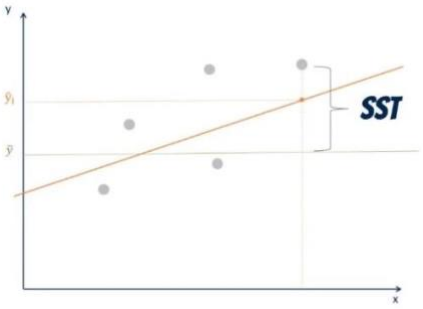
$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) \cdot x^{(i)}$$

}

The **sum of squares total**, the **sum of squares regression**, and the **sum of squares error**. https://365datascience.com/sum-squares/

SST (sum of squares total) or **TSS** (**total sum of squares)** is the squared differences between the observed dependent variable and its mean.

How much the data point move around the mean.



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

SSR (sum of squares regression) or **ESS** (**explained sum of squares)** is the sum of the differences between the *predicted* value and the **mean** of the *dependent variable*.

This gives information about how far estimated regression line is from the horizontal 'no relationship' line (average of actual output).

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

SSE (sum of squares error) or **RSS (residual sum of squares)**. The error is the difference between the *observed* value and the *predicted* value.

How much the target value varies around the regression line (predicted value).

$$\sum_{i=1}^{n} e_i^2$$

SST = **SSR** + **SSE**

$$r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

How to Interpret Regression Analysis Results: **P-values** and **Coefficients**
https://www2.isye.gatech.edu/~yxie77/isye2028/lecture12.pdf

- The **p-value** for each term tests the null hypothesis that the coefficient is equal to zero (no effect).
- Low P-value (< 0.05): Rejects null hypothesis indicating that the predictor value is related to the response
- High P-value: Changes in predictor are not associated with change in target.
- In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.

➔ Regression **coefficients** represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.

➔ This statistical control that regression provides is important because it isolates the role of one variable from all of the others in the model.