Regression Classification

🕐 Created	@Jan 20, 2021 12:13 AM
👤 Created By	🅚 Khanh Vương
👤 Last Edited By	🅚 Khanh Vương
🕐 Last Edited Time	@Jan 30, 2021 7:44 PM
☰ Module	Regression
⊙ Status	Archived
⊙ Type	Regression Problem

What is Overfitting ?

- Best describe the `training data`
- Cannot give good prediction

Nature of Linear Regression

- Is one of the most widely-used technique

- Nature of **Regression Line** is `Linear` .

What is the Least Squared Method ?

- This is the most common method used for fitting a `Regression Line`

$$min\|X_w - y\|_2^2$$

How can we evaluate a Linear Regression model ?

- We can evaluate the model performance using the metric r^2

What should be noticed about this type of model ?

- **Linear Regression** is very sensitive to outliers

- There must be **Linear Relationship** between `Independent` and `Dependent Variables`

- **Multiple Regression** suffers from **multicollinearity, autocorrelation, heteroskedasticity**.

Nature of Logistic Regression

What is the definition of Logistic Regression ?

- **Logistic Regression** is used to find the probability of the events that only have two outputs, meaning that the **Dependent Variable** is binary, which is follow the **Binomial Distribution**

When should we use Logistic Regression ?

- Used for **Classification Problem**

- **Logistic Regression** doesn't require `linear` relationship between `dependent` and `independent` variables. It can handle various types of relationships because it applies a non-linear `log` transformation to the predicted odds ratio

- To avoid `overfitting` and `underfitting`, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the **Logistic Regression**

- The `Independent Variables` should not be correlated with each other i.e. **no multi collinearity**. However, we have the options to include interaction effects of categorical variables in the analysis and in the model.

- If the values of `Dependent Variable` is ordinal, then it is called as **Ordinal Logistic Regression**

- If `Dependent Variable` is multi class then it is known as **Multinomial Logistic Regression**.

What is the different between evaluating a Linear Regression model and Logistic Regression model ?

- **Linear Regression**: Aim to use **Least Square Method**, where we will want to minimize the `Error`, means finding the `best-fit-line`

- **Logistic Regression**: Aim to maximize the likelihood between the `predict` and the actual value. Therefor, **Logistic Regression** will require **large sample sizes** because maximum likelihood estimates are less powerful at low sample sizes than ordinary **Least Square**
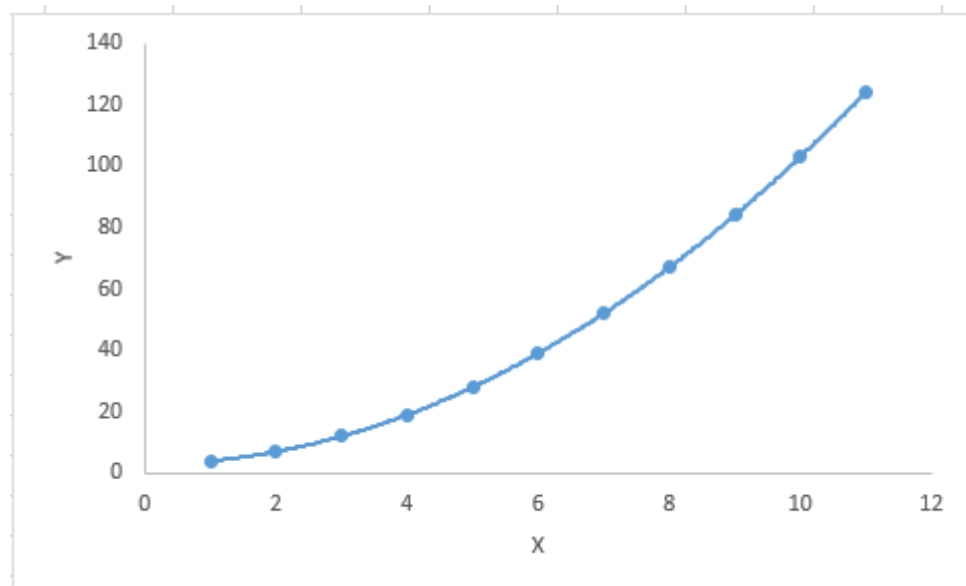
Polynomial Regression Definition

- The power of **Independent Variable** is more than 1, means increase the flexibility of the model

$$y = a + bx^2$$

What is the shape of the Regression Line in the Polynomial Regression ?

- In this technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

What should we notice about Polynomial Regression ?

- While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in `over-fitting`. Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem

- Especially look out for curve towards the ends and see whether those shapes and trends make sense. Higher **Polynomials** can end up producing weird results on extrapolation.

Stepwise Regression Definition

- Aim to use **Stepwise Regression** when we have multiple **Independence Variables**

- The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the method to handle higher dimensionality of data set.

What are the basic concepts of Stepwise Regression ?

- **Independence Variables** aims to automatically select the **Independence Variables**

- Basically, it will try to fit the **Regression Model** by adding/dropping co-variates one at a time based on a specified criterion like $R^2, t-stats$ and AIC

  - **Standard Stepwise Regression** does two things. It adds and removes predictors as needed for each step.

  - **Forward Selection** starts with most significant predictor in the model and adds variable for each step.

  - **Backward Elimination** starts with all predictors in the model and removes the least significant variable for each step.

Ridge Regression Definition

- When we have many **Independence Variables** and they are correlated to each others

Nature or Ridge Regression

- In multicollinearity, even though the **Least Squares** estimates (**OLS**) are unbiased, their variances are large which deviates the observed value far from the true value.

- By adding a degree of bias to the regression estimates, **Ridge Regression** reduces the standard errors.