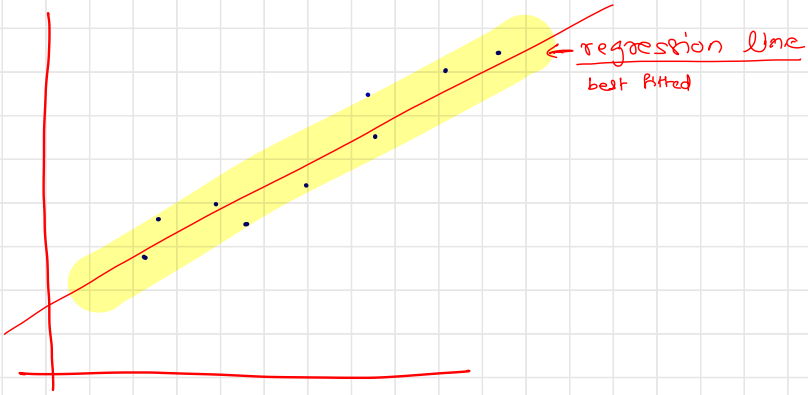
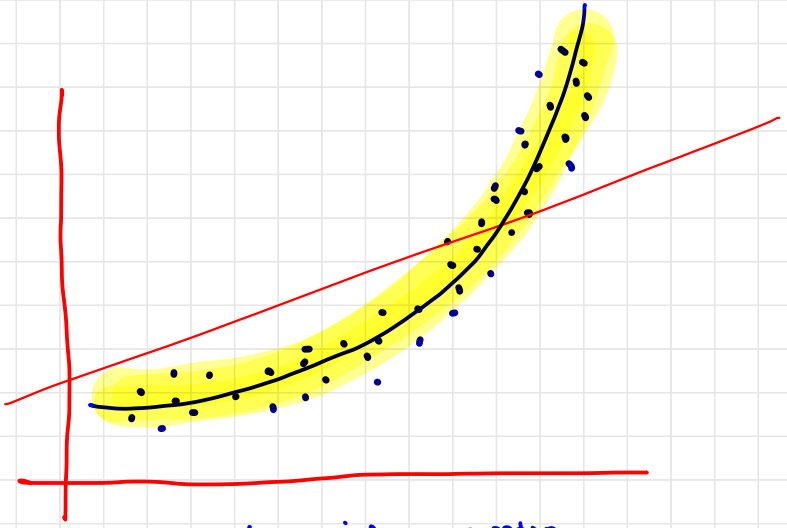
non-linear relationship

Polynomial Regression



Simple Regression



polynomial regression

Introduction

- Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression

exponential power of independent variable(s)

- Suppose we have X as Independent data and Y as dependent data. Before feeding data to a mode in preprocessing stage we convert the input variables into polynomial terms using some degree
- Consider an example my input value is 35 and the degree of a polynomial is 2 so I will find 35 power 0, 35 power 1, and 35 power 2 And this helps to interpret the non-linear relationship in data. The equation of polynomial becomes something like this.

polynomial ⇒ $y = a_0 + a_1x_1 + a_2x_1^2 + \ldots + a_nx_1^n$

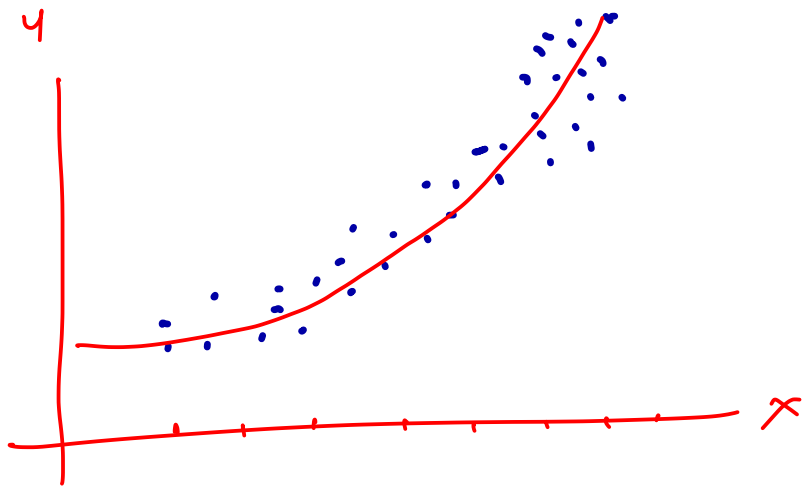
polynomial feature

simple linear ⇒ $y = a_0 + a_1x_1$ (a_0, a_1: constants)

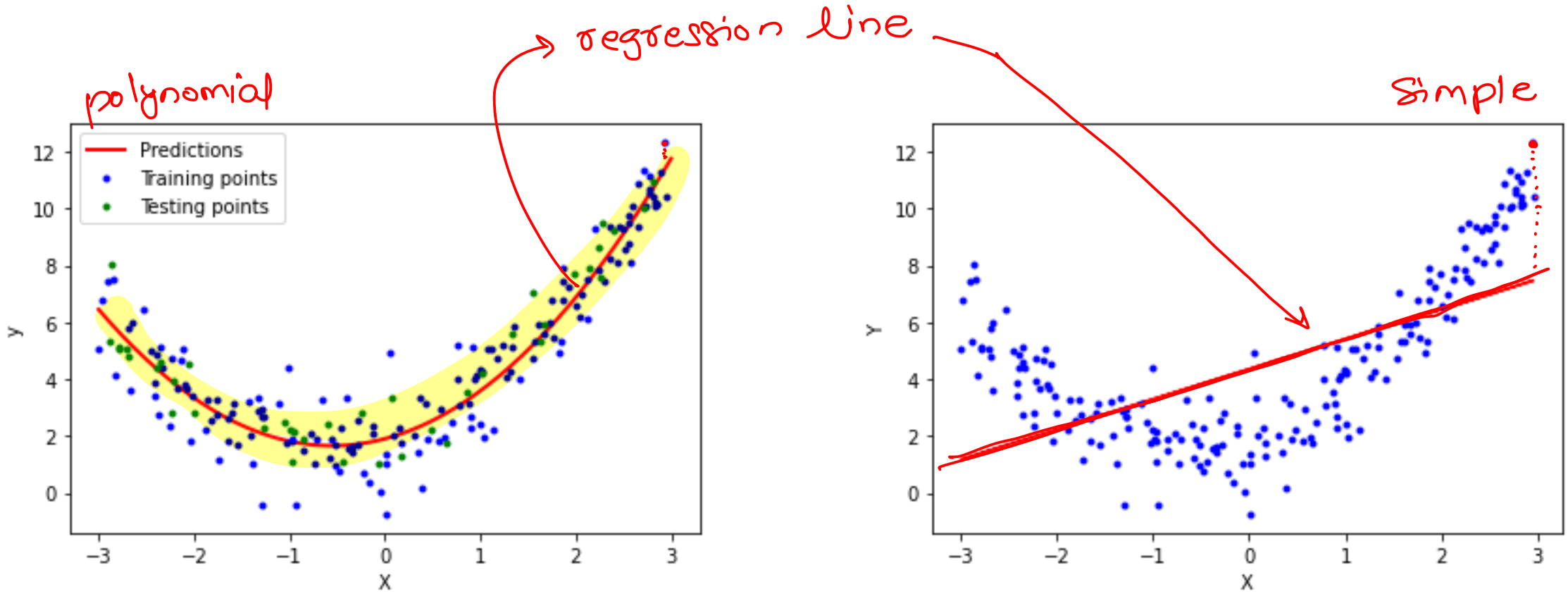
multi-linear ⇒ $y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n$ ($a_0, a_1, a_2, \ldots, a_n$: coefficients)

Introduction

- The degree of order which to use is a Hyperparameter, and we need to choose it wisely. But using a high degree of polynomial tries to overfit the data and for smaller values of degree, the model tries to underfit so we need to find the optimum value of a degree by using trial and error method
- If you see the equation of polynomial regression carefully, then we can see that we are trying to estimate the relationship between coefficients and y



Polynomial vs Simple Linear

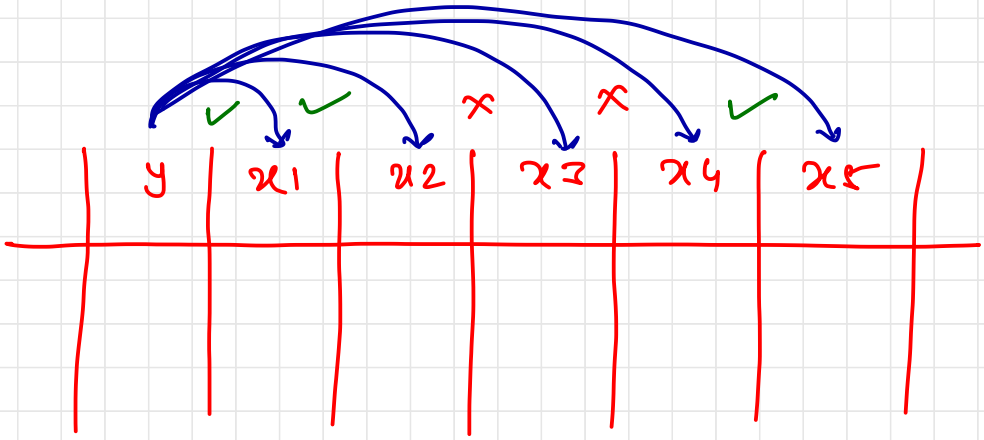


Stepwise Regression

Introduction

- Stepwise regression is a variable selection procedure for independent variables
- It consists of a series of steps designed to find the most useful x-variables to include in a regression model → correlation
- At each step of procedure, each variable is evaluated using a set criterion to see if should remain in the model or not
- Basis for selection could be
  1) Choosing variable that satisfies the stipulated criterion
  2) Removing the variable that least satisfies the criterion
- Example of such a criterion is t-statistic value

t-statistics
f-statistics

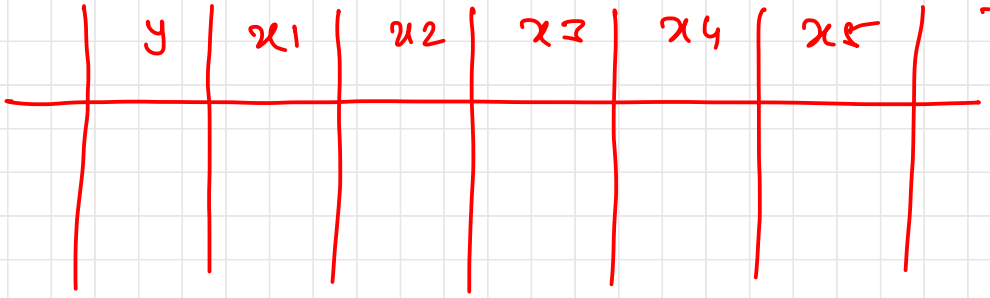


forward Selection

[x1, x2, x5]

t-statistics (y, x1) = score

t-statics (y, x2) = score



Backward Elimination

[x1, x2, ~~x3~~, ~~x4~~, x5]

Approaches

- **Forward selection**
  - Begins with no variables in the model
  - Tests each variable as it is added to the model
  - Then keeps those that are deemed most statistically significant
  - Repeating the process until the results are optimal
- **Backward elimination**
  - Starts with a set of independent variables
  - Deleting one at a time
  - Then testing to see if the removed variable is statistically significant
- **Bidirectional elimination**
  - It is a combination of the first two methods that test which variables should be included or excluded

Advantages

- The ability to manage large amounts of potential predictor variables, fine-tuning the model to choose the best predictor variables from the available options
- It's faster than other automatic model-selection methods
- Watching the order in which variables are removed or added can provide valuable information about the quality of the predictor variables

Disadvantages

- Stepwise regression often has many potential predictor variables but too little data to estimate coefficients meaningfully. Adding more data does not help much, if at all.
- If two predictor variables in the model are highly correlated, only one may make it into the model
- R-squared values are usually too high
- Adjusted r-squared values might be high, and then dip sharply as the model progresses. If this happens, identify the variables that were added or removed when this happens and adjust the model
- Predicted values and confidence intervals are too narrow
- P-values are given that do not have the correct meaning
- Regression coefficients are biased and coefficients for other variables are too high
- Collinearity is usually a major issue. Excessive collinearity may cause the program to dump predictor variables into the model.
- Some variables (especially dummy variables) may be removed from the model, when they are deemed important to be included. These can be manually added back in.

Elastic Net Regression

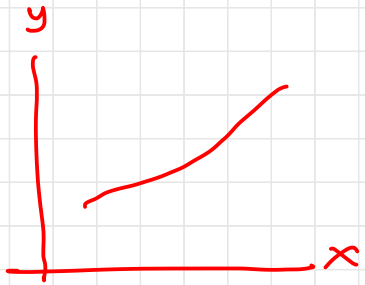
↳ L1 & L2 Regularization

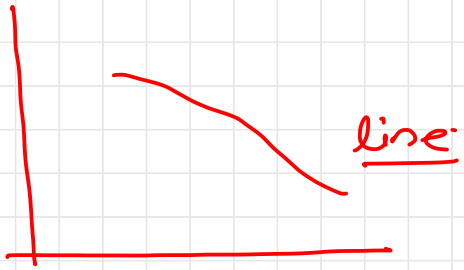
Introduction

- Linear regression refers to a model that assumes a linear relationship between input variables and the target variable
- With a single input variable, this relationship is a line, and with higher dimensions, this relationship can be thought of as a hyperplane that connects the input variables to the target variable
- The coefficients of the model are found via an optimization process that seeks to minimize the sum squared error between the predictions (*yhat*) and the expected target values (*y*)
- A problem with linear regression is that estimated coefficients of the model can become large, making the model sensitive to inputs and possibly unstable
- This is particularly true for problems with few observations (*samples*) or more samples (*n*) than input predictors (*p*) or variables
- Elastic net is a penalized linear regression model that includes both the L1 and L2 penalties during training

①

x	y





②

y	x1	x2