Data Science Workshop

Regression Analysis

Toulon, February 2019

Emmanuel Génard – *genard@fr.ibm.com*
Data Scientist & Cloud Developer Advocate Europe,
IBM Business Solution Center Nice, France

ISEN
ALL IS DIGITAL!
MÉDITERRANÉE

Objectives

- What is Regression Analysis?
- Why do we use Regression Analysis?
- What are the types of Regressions?
- How to select the right Regression Model?

What is Regression Analysis?

Basic Idea:
Use data to indetify **relationships** among variables and use these relationships to make **predictions**

- Form of predictive modelling technique
- Investigates relationship between **dependent** (target) and **independent variables** (predictor).
  - **Dependent Variable:** This is the main factor that you're trying to understand or predict.
  - **Independent Variables:** These are the factors that you hypothesize have an impact on your dependent variable.

Benefits of Regression Analysis?

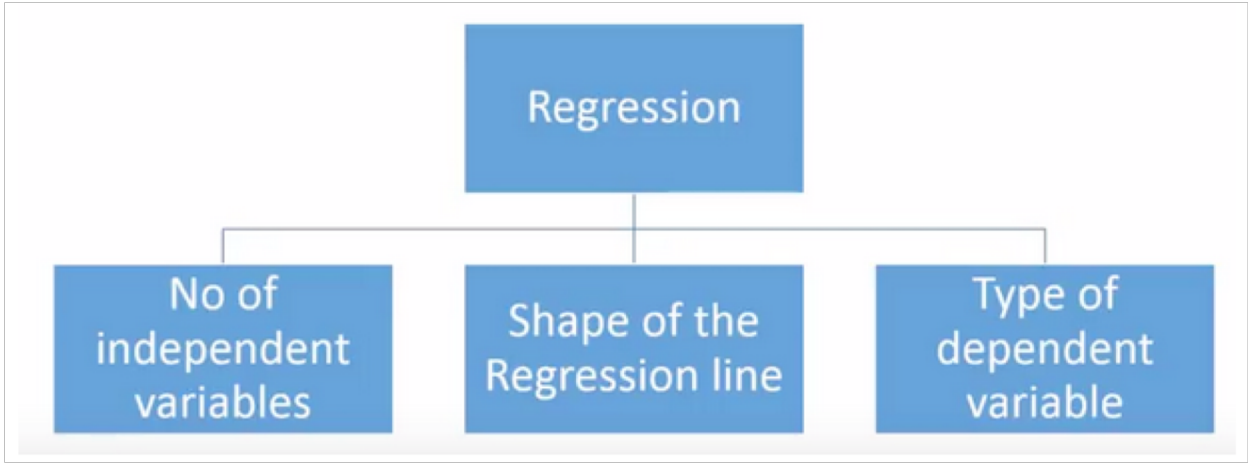
There are multiple benefits of using regression analysis. They are as follows:

- It indicates the **significant relationships** between dependent variable and independent variable.
- It indicates the **strength of impact** of multiple independent variables on a dependent variable.

Important Points

- There must be **linear relationship** between independent and dependent variables
- Multiple regression suffers from **multicollinearity, autocorrelation, heteroskedasticity**.
- Linear Regression is very sensitive to **Outliers**. It can terribly affect the regression line and eventually the forecasted values.

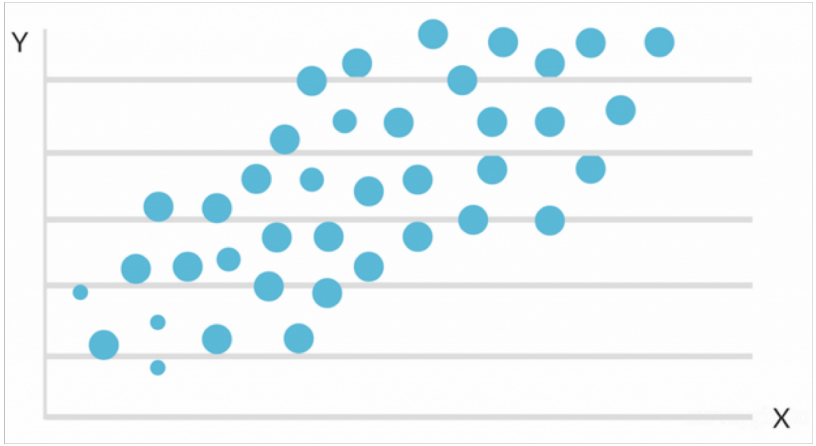
How many types of regression techniques do we have?

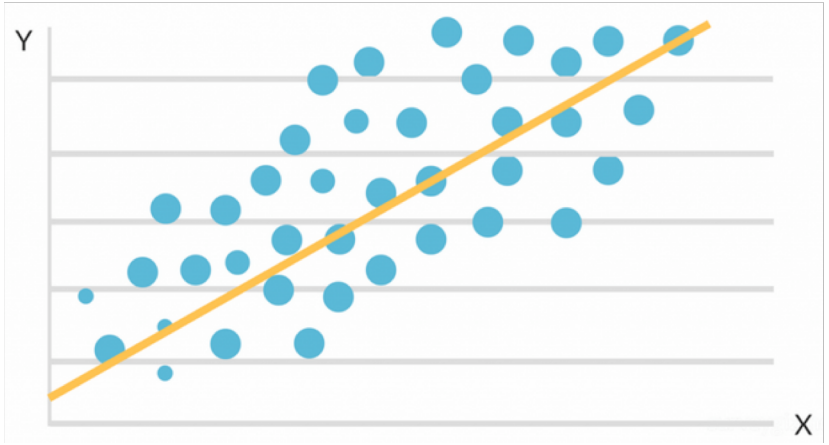


How does regression analysis work (Linear Regression) ?

- Start by plotting data points on a chart
- **Dependent variables (*Y*)** on **y-axis**
- **Independent variable (*X*)** on **x-axis**
- Draw a line in the middle of all the data points on the chart: **the regression line (curve/fit line)**

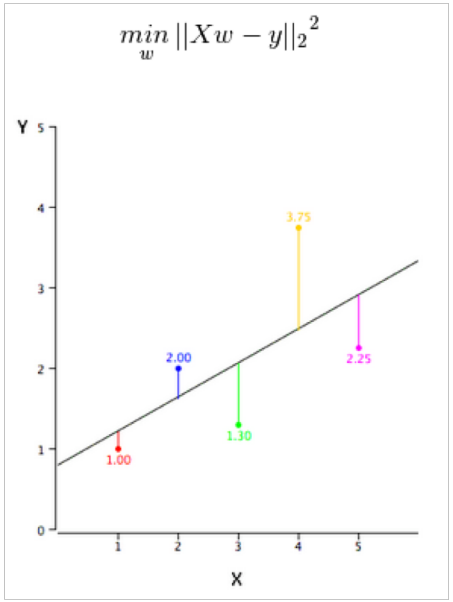
- A regression model relates *Y* to a function of *X* and
- $Y \approx f(X, \beta)$





How does regression analysis work (Linear Regression)?

- In linear regression
  - $Y \approx f(X, \beta)$
  - $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,,\ i = 1, \ldots, n$
- We obtain a model:
  - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- The **residual**, is the difference between the value of the dependent variable predicted by the model $\hat{y}_i$ and the true value of the dependent variable y_i
  - $e_i = y_i - \hat{y}_i$
- Ordinary Least Squares:
  - $SSR = \sum_{i=1}^{n} e_i^2$



How does regression analysis work (Linear Regression)?

Prediction and Forecasting

- Predict home sales for December given the interest rate for this month.
- Use time series data (e.g., sales vs. year) to forecast future performance (next year sales).
- Predict the selling price of houses in some area.
  - Collect data on several houses (# of BR, #BA, sq.ft, lot size, property tax) and their selling price.
  - Can we use this data to predict the selling price of a specific house?

Quantifying causality

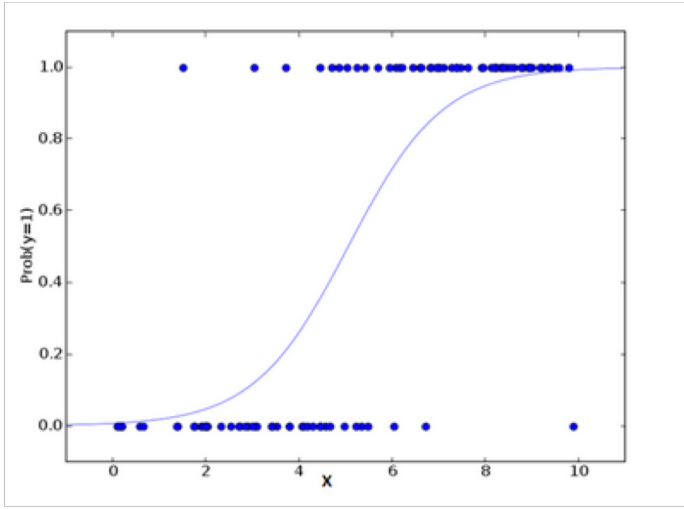
- Determine factors that relate to the variable to be predicted; e.g., predict growth for the economy in the next quarter: use past history on quarterly growth, index of leading economic indicators, and others.
- Want to determine advertising expenditure and promotion for the 1999 Ford Explorer.
  - Sales over a quarter might be influenced by: ads in print, ads in radio, ads in TV, and other promotions

Logistic Regression

- Used to find the probability of event=Success and event=Failure.
- **Dependent variable** is **binary** (0/1, True/False, Yes/No)

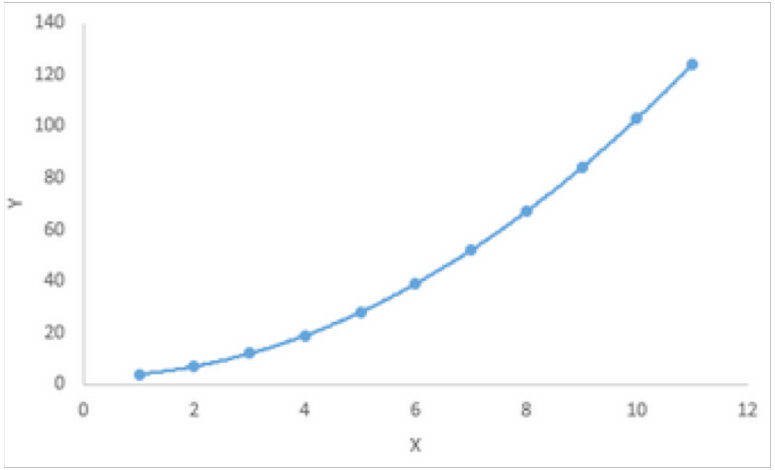
- $odds = \frac{p}{(1-p)} = \frac{Probability\ of\ event\ occurence}{Probability\ of\ not\ event\ occurence}$

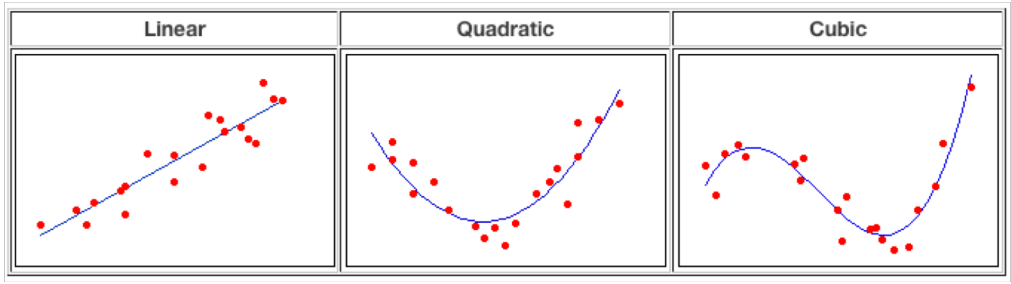
- It is widely used for **classification problems**
- Logistic regression **doesn't require linear relationship** between dependent and independent variables.
- To avoid over fitting and under fitting, we should include all significant variables.
- It requires **large sample sizes**



Other types of Regression

- **Polynomial Regression:**
  - $y = ax^2 + b$
- Helps in fitting curves in Linear Regression
- It can result in **over-fitting**





- Stepwise Regression
- Ridge Regression
- Lasso Regression
- ElasticNet Regression

How to select the right regression model?

1. **Data exploration** is an inevitable part of building predictive model. It should be you first step before selecting the right model like identify the relationship and impact of variables

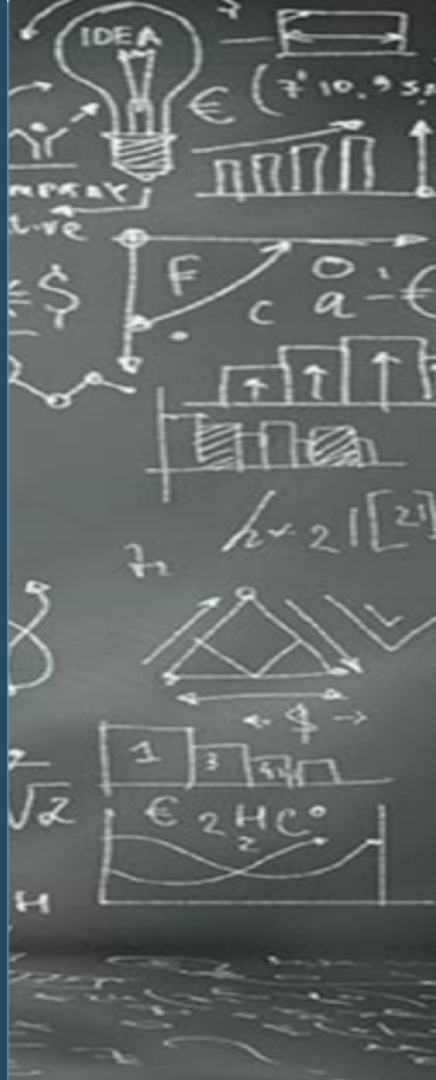
2. To compare the goodness of fit for different models, analyze different metrics like statistical significance of parameters, R-square, Adjusted r-square, AIC, BIC and error term.

3. Cross-validation is the best way to evaluate models used for prediction. Here you divide your data set into two group (train and validate). A simple mean squared difference between the observed and predicted values give you a measure for the prediction accuracy.

4. If your data set has multiple confounding variables, you should not choose automatic model selection method because you do not want to put these in a model at the same time.

5. It'll also depend on your objective. It can occur that a less powerful model is easy to implement as compared to a highly statistically significant model.

Thank You



Useful Links & Resources

External

Getting Started:
Service Homepage
Feature Requests / Suggestions

Case Studies:
OmniEarth
Aerialtronics
BlueChasm
iTrend

Tutorials & Best Practices:
Training models with Watson Studio
Getting started with Watson + Core ML
Stacking Multiple Custom Models
Create a Calorie Counting App
Watson Visual Recognition & Twilio
Best Practices for Custom Models

Code Patterns:
Classify vehicle damage
Analyze industrial equipment for defects
Create an Android calorie-counter app

External continued

Books:
Redbook: Building Cognitive Application using IBM Watson Services vol3 – Watson Visual Recognition

Blogs:
IBM Watson on Medium

Internal

Slack Channel: #ibmvisual-recognition
Service Roadmap
IBMer key limit increase request form
ZACS portal
Digital Sales Play
Content Request & Feedback Form