Regression Analysis

MD Arshad Ahmad
15 Years+ Experience in Data Science
Mentored 100+ people

Decoding Data Science

Agenda

- Introduction to Regression Analysis
  - What is Regression Analysis
  - Why do we need Regression Analysis in Business – Introduction to Modeling
- Introduction to OLS Regression
- Introduction to Modeling Process

What is Regression Analysis?

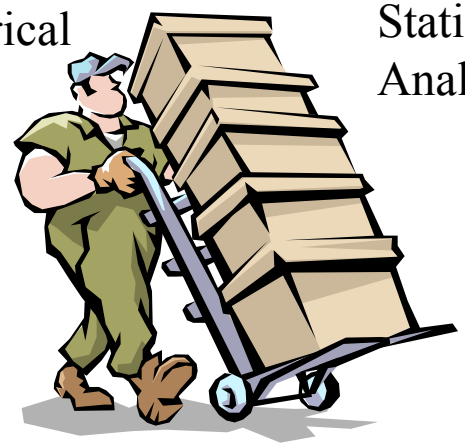
Regression Analysis captures the relationship between one or more response variables (dependent/predicted variable – denoted by Y) and the its predictor variables (independent/explanatory variables – denoted by X) using historical observations of both.

Hence its estimates the functional relationship between a set of independent variables $X_1, X_2, \ldots, X_p$ with the response variable Y which estimate of the functional form best fits the historical data.

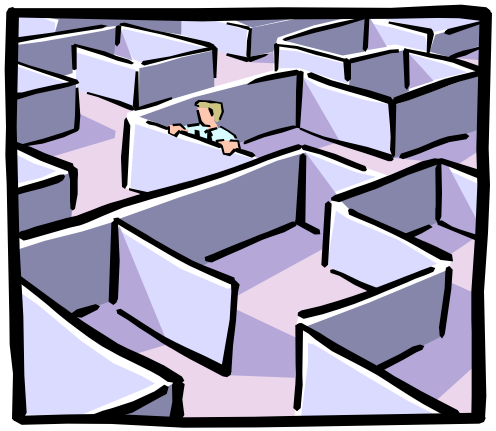
$$Y = f(X_1, X_2 .., X_p) + \epsilon$$

where ϵ denotes the "Residual" or unexplained part of Y

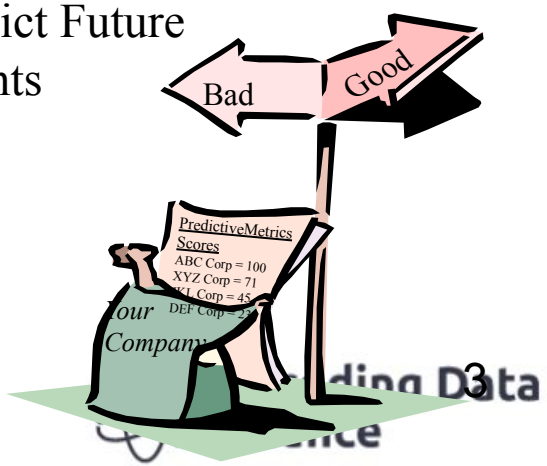
Historical Data



Statistical Analyses



Predict Future Events



Types of Regression Analysis

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$

There are various kinds of Regressions based on the nature of : -

- the functional form of the relationship
- the residual
- the dependent variable
- the independent variables

Functional Form	Residual	Dependent Var	Independent Var

▪ **Linear**
▪ Non-Linear – *Out of scope for this presentation*
	▪ Based on the distribution of the residual – normal, binomial, poisson, exponential	▪ **Single**
		  ▪ **Continuous**
		  ▪ **Discrete**
		  ▪ **Binary**
		▪ Multiple – *Out of scope for this presentation*
			▪ Numerical
			  ▪ Discrete
			  ▪ Continuous
			▪ Categorical
			  ▪ Ordinal
			  ▪ Nominal

Types of Linear Regression

Dependent Variable Type	Residual Distribution	Types of Regression
Continuous	Normal (with constant variance)	Ordinary Least Squares (OLS)
Continuous	Normal (without constant variance)	Generalized Least Square
Binary	Binomial	Logistic Regression
Discrete	Poisson	Poisson Regression
Rational	Exponential Family of Distributions	Generalized Least Squares

Other Types of Regression Related Techniques

- Simultaneous Equation Models
  - When both X & Y are dependent on each other
- Structural Equation Modeling / Pathways
  - Captures the inter-relations between Xs i.e. captures how Xs affect each other before affecting Y
- Survival Analysis
  - Predicts a decay curve for a probability of an event
- Hierarchal Bayesian
  - Estimates a non-linear equation

Agenda

- Introduction to Regression Analysis
  - What is Regression Analysis
  - Why do we need Regression Analysis in Business – Introduction to Modeling
- Introduction to OLS Regression
- Introduction to Modeling Process

What is Modeling?

- ✔ Is based on Regression Analysis
- ✔ It can be used for the following two distinct but related purposes
  - ✔ Predict certain events
  - ✔ Identify the drivers of certain events based on some explanatory variables
- ✔ Isolates individual effects and then quantifies the magnitude of that driver to its impact on the dependent variable
- ✔ It is required because
  - ✔ Knowledge of Y is crucial for decision making but is not deterministic
  - ✔ X is available at the time of decision making and is related to Y

$$Volume = Base\ Sales + b_2(GRPs) + b_3(Dist) \ldots + b_n(Price)$$

- **Predict the sales that a customer would contribute, given a certain set of attributes like demographic information, credit history, prior purchase behavior, etc.**
- **Predict the probability of response from a direct mail thus saving cost and acquire potential customers.**
- **Identify high responsive and high profit segments and targeting only these segments for direct mail campaigns**
- **Identify the most effective marketing levers & quantify their impact**
- **To find out what differentiates between buyers and non buyers based on their past 3 months usage of the product and the age group**

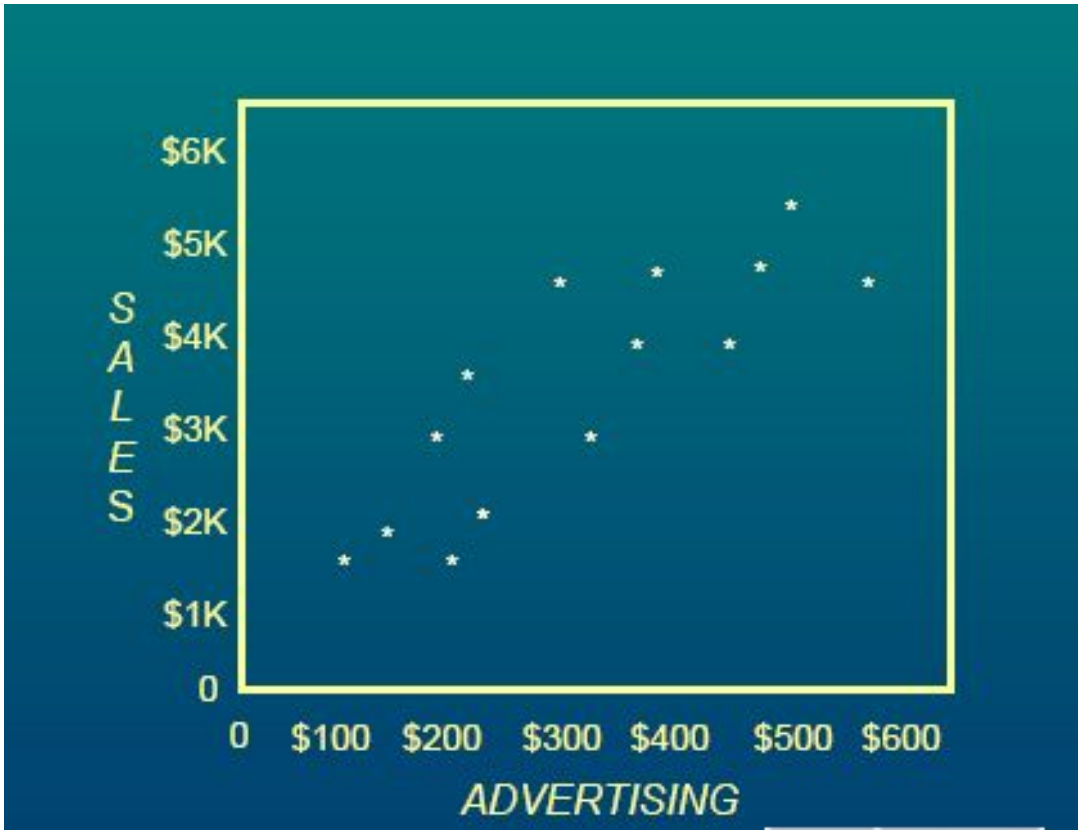
- Introduction to Regression Analysis
- Introduction to OLS Regression
- Introduction to Modeling Process

Introduction to Ordinary Least Squares

Dependent Variable Type	Residual Distribution	Types of Regression
Continuous	**Normal (with constant variance)**	**Ordinary Least Squares (OLS)**
Continuous	Normal (without constant variance)	Generalized Least Square
Binary	Binomial	Logistic Regression
Discrete	Poisson	Poisson Regression
Rational	Exponential Family of Distributions	Generalized Least Squares

Introduction to Ordinary Least Squares – Simple Regression

Advertising	Sales
$120	$1,503
$160	$1,755
$205	$2,971
$210	$1,682
$225	$3,497
$230	$1,998
$290	$4,528
$315	$2,937
$375	$3,622
$390	$4,402
$440	$3,844
$475	$4,470
$490	$5,492
$550	$4,398

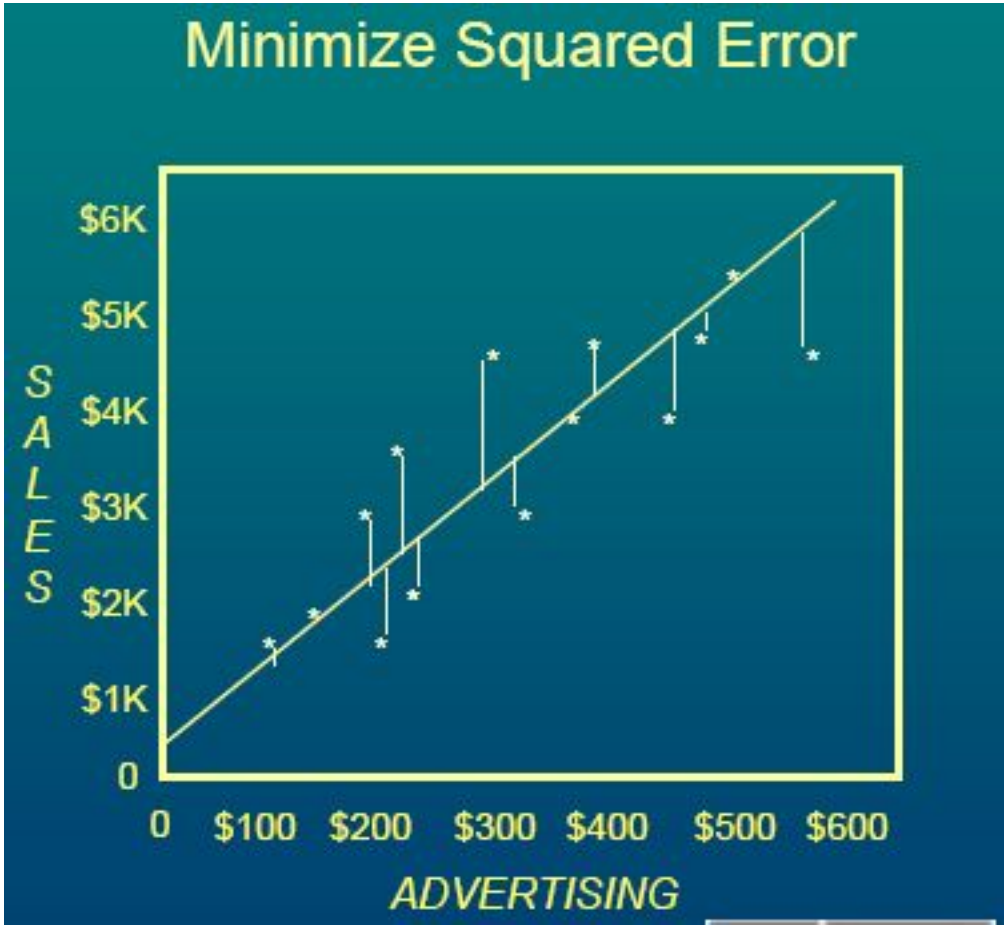


Goal: characterize relationship between advertising and sales

Introduction to Ordinary Least Squares – Simple Regression

Result: equation that predicts sales dollars based on advertising dollars spent

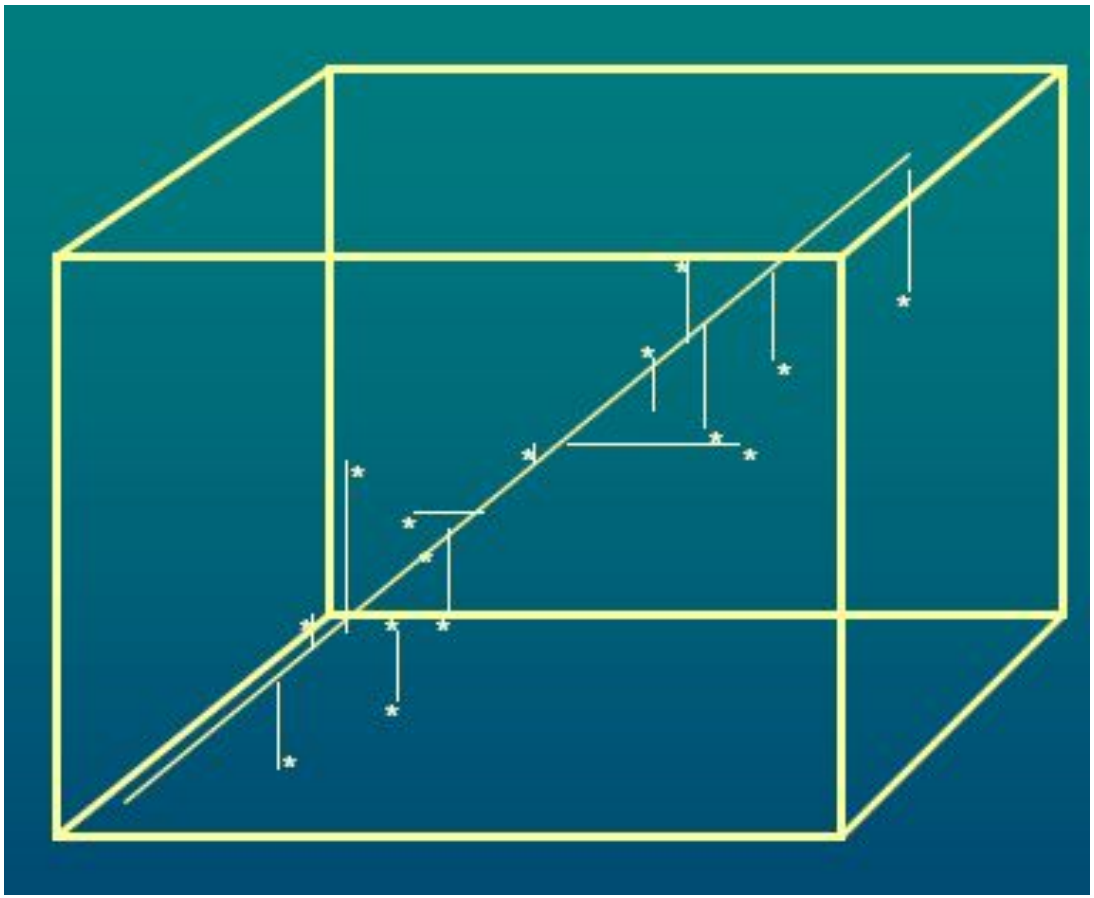
$$Sales = B_0 + B_1 * Adv.$$



Minimizes Error sum of squares ,Hence the name "Ordinary Least Square Regression"

Introduction to Ordinary Least Squares – Multiple Regression

- Credit card balances
  - payment amount
  - years
  - gender (0/1)
- Minimizes squared error in N-dimensional space



Balances = 2.1774 +.0966*Payment + 1.2494*Months + .4412*Gender

OLS Model Assumptions

1. Linearity

   Model is linear in parameters

   $$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_p X_{pi} + e_i$$

2. Spherical Errors

   Error distribution is Normal with mean 0 & constant variance

   $$e_i \sim \text{Normal}(0, \sigma^2)$$

3. Zero Expected Error

   The expected value (or mean) of the errors is always zero

   $$E(e_i) = 0 \text{ for all } i$$

4. Homoskedasticity

   The errors have constant variance

   $$\text{Variance}(e_i) = \text{constant for all } i$$

5. Non-Autocorrelation

   The errors are statistically independent from one another. This implies the data is a random sample of the population

   $$\text{corr}(e_i, e_j) = 0 \text{ for all } i \neq j$$

6. Non-Multicollinearity

   The independent variables are not collinear

   $$\text{Covariance}(X_i, X_j) = 0$$

Steps in OLS Regression

Assume all OLS assumptions hold

Run regression in software (R/Python)

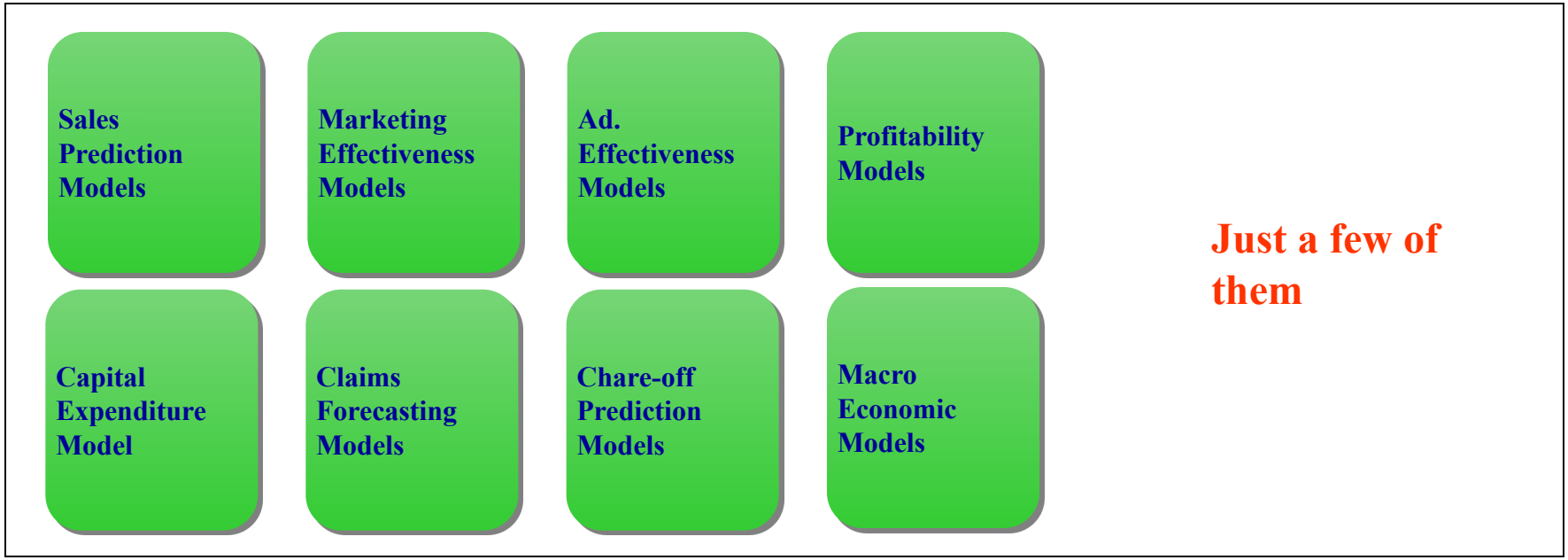
Check if assumptions really hold

Check if Fit is good

Check Hypothesis testing results
i.e. variable significance

Iterate to make “BEST” model

Applications of OLS Regression in Business



Thank You!
To know more Get In Touch!

Kick start your Data Science Career

Book Mentoring Session

www.decodingdatascience.com

Decoding Data Science