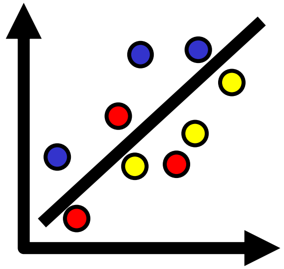
Continuous Improvement Toolkit

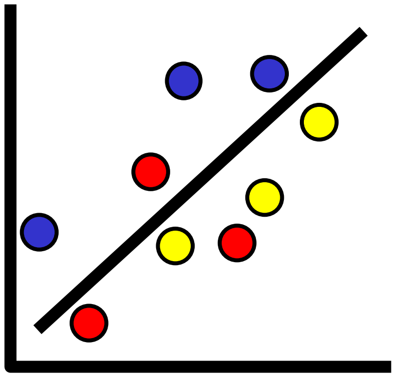
Regression

(Introduction)



Managing Risk
PDPC
FMEA
RAID Logs
Fault Tree Analysis
Risk Assessment*
Traffic Light Assessment
Deciding & Selecting
Pros and Cons
Importance-Urgency Mapping
Break-even Analysis
Cost -Benefit Analysis
Force Field Analysis
Voting
Pugh Matrix
Decision Tree
TPN Analysis
QFD
Matrix Diagram
Prioritization Matrix
Kano Analysis
Paired Comparison
Critical-to Tree
Planning & Project Management*
RACI Matrix
Stakeholders Analysis
PEST
PERT/CPM
Activity Diagram
SWOT
Roadmaps
Project Charter
Gantt Chart
PDCA
Control Planning
Gap Analysis
Hoshin Kanri
Kaizen
How-How Diagram
Tree Diagram**
Standard work
Identifying & Implementing Solutions***
Simulation
TPM
Mistake Proofing
Pull Systems
JIT
Ergonomics
Work Balancing
Automation
Bottleneck Analysis
Visual Management
Flow
Value Analysis
5S
Wastes Analysis
SMED
Understanding Cause & Effect
Cause & Effect Matrix
Pareto Analysis
Confidence Intervals
Probability Distributions
ANOVA
Design of Experiments
Hypothesis Testing
Regression
Scatter Plot
Correlation
5 Whys
Chi-Square Test
Multi-Vari Charts
Relations Mapping*
Fishbone Diagram
Understanding Performance
Lean Measures
KPIs
OEE
Capability Indices
MSA
RTY
Descriptive Statistics
Cost of Quality
Reliability Analysis
Graphical Analysis
Run Charts
Control Charts
Benchmarking
Sampling
Data Collection
Focus groups
Interviews
Photography
Check Sheets
Measles Charts
Surveys
Critical Incident Technique
Observations
Creating Ideas**
TRIZ***
Brainstorming
Analogy
SCAMPER***
Nominal Group Technique
Mind Mapping*
Affinity Diagram
Attribute Analysis
Lateral Thinking
Visioning
Designing & Analyzing Processes
Time Value Map
Process Redesign
IDEF0
Value Stream Mapping
SIPOC
Flow Process Chart
Process Mapping
Flowcharting
Service Blueprints

- Introduction to Regression



- **Regression** (& Correlation) is used when we have data inputs and we wish to explore if there is a relationship between the inputs and the output.
  - What is the strength of the relationship?
  - Does the output increase or decrease as we increase the input value?
  - What is the mathematical model that defines the relationship?
- Given multiple inputs, we can determine which inputs have the biggest impact on the output.
- Once we have a model (regression equation) we can **predict** what the output will be if we set our input(s) at specific values.

- Introduction to Regression

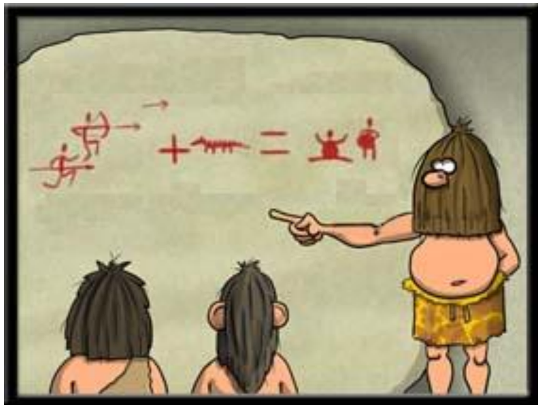
$$Y=f(x)$$

- Regression is a statistical forecasting model that is concerned with describing and evaluating the relationship between variables.
- It is the process of developing a mathematical **model** that represents the data.
- It provides an equation or model to describe the relationship between two (or more) variables.
- This regression equation can be used to predict future events.

- Introduction to Regression

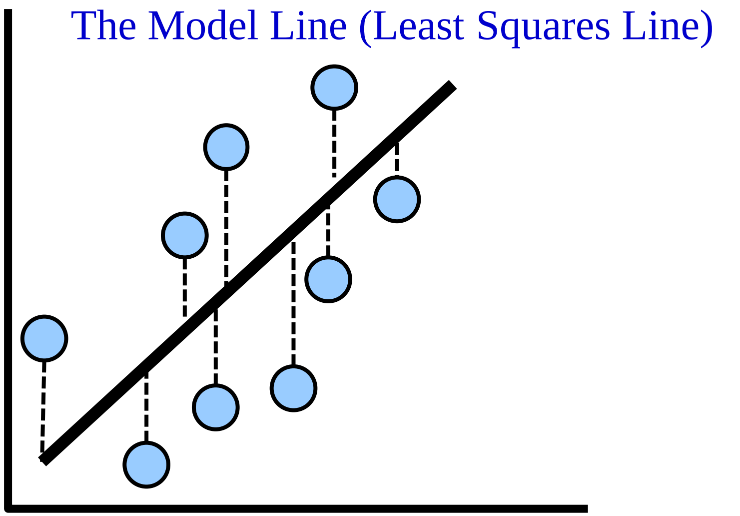
Two Types:

- **Simple Regression:**
  - We have only one explanatory variable.
  - The regression process can fit several shapes of line:
    - Linear.
    - Quadratic.
    - Cubic.
- **Multiple Regression:**
  - We may be interested in tow or more explanatory variables.



- Introduction to Regression

- It mathematically defines the relationship between the **explanatory** variable (X) and the **response** variable (Y).
- The regression process creates a line that best resembles the relationship between the process input and output.
- The best line is found by ensuring the errors between the data points and the line are minimized.



- Introduction to Regression

- All straight lines can be expressed as:

$$Y = \beta_0 + \beta_1 x$$

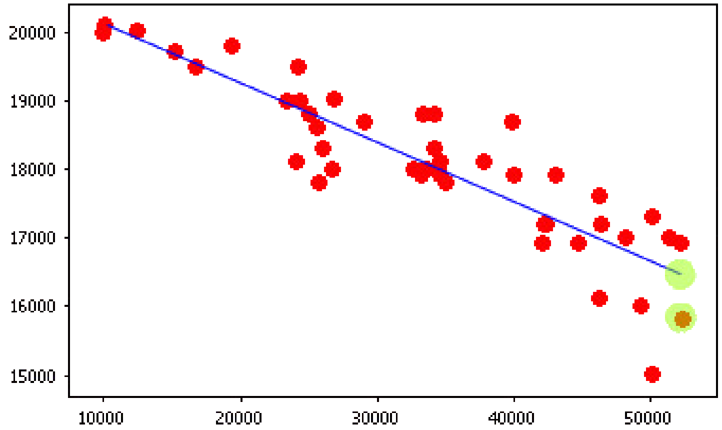
Y → The response variable.

X → The explanatory variable.

β0 → The intercept (The value of Y when x=0).

β1 → The slope (The impact of the explanatory variable on the response variable).

- Introduction to Regression



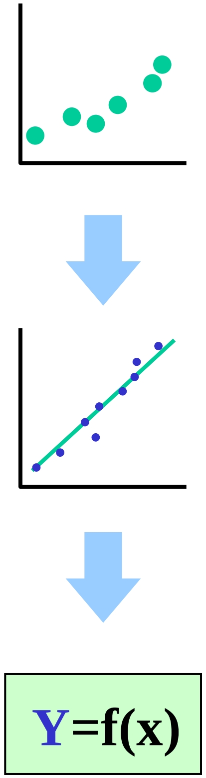
- The distances between the points and the regression line are called **residuals**.
- They represent the portion of the response that is not explained by the regression equation.
- Residuals (which are also referred as errors) must be encountered in the regression equation:

$$\mathbf{Y = \beta_0 + \beta_1 x + \varepsilon}$$

- Introduction to Regression

Approach:

- Collect random data.
- Create a scatter plot to check the relationship between the variables.
- Use correlation to quantify the strength and direction of the relationship.
- Use regression to develop an equation to describe the relationship.



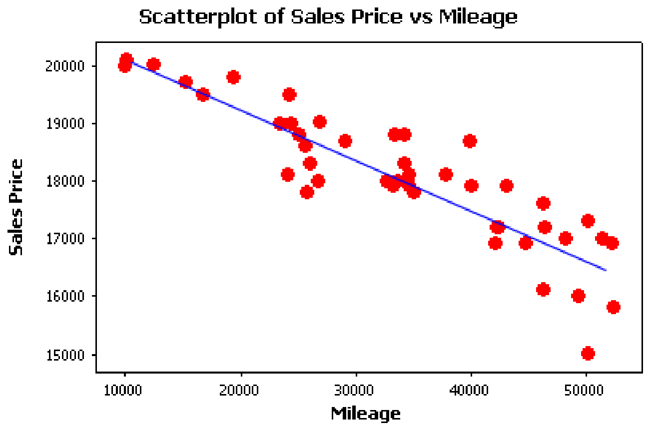
- Introduction to Regression

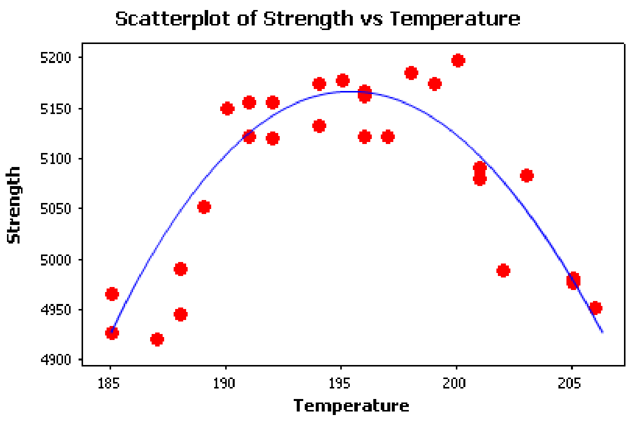
The Process:

Graph the Data	Scatter plot
Check the Correlations	Use Pearson Coefficient
1st Regression	Linear / Multiple regression
Evaluate Regression	R-squared & analyze residuals
Re-run Regression (If necessary)	Simple: With different model (Cubic) Multiple: Remove unnecessary items
Use the Results	Control critical process inputs & select best operating levels.

- Introduction to Regression

- With a linear relationship, we can use correlation and regression to evaluate the data.
- Sometimes the pattern is nonlinear.
- We need to use other advanced tools to evaluate the data.
- Such analysis tools are beyond the scope of this training.

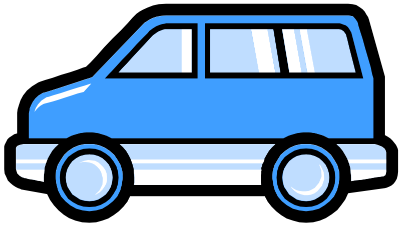




- Introduction to Regression

Example:

- Suppose that we conduct an experiment to examine the relationship between the vehicles sales price and the mileage.
- After we collected random data, we want to know how car mileage influence sales price.
- Which is the explanatory variable?



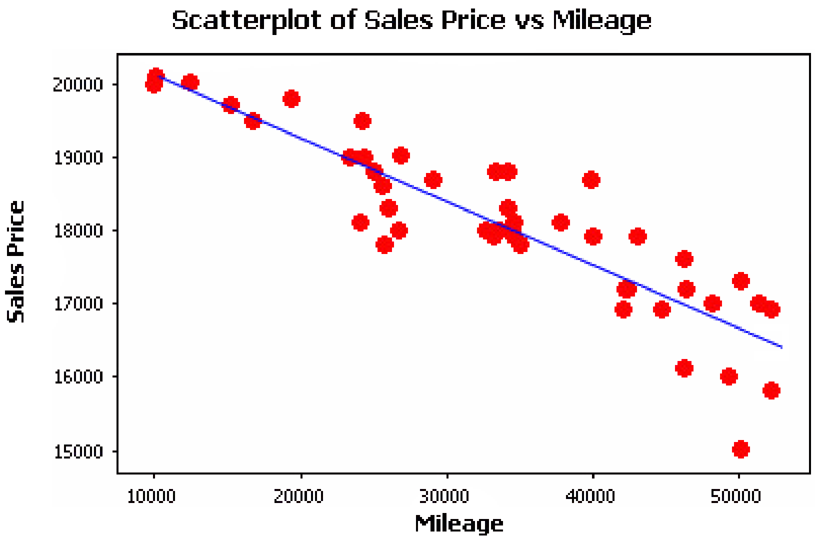
The mileage is the explanatory variable and sales price is the response variable.

Mileage	Sales Price
9980	19999
35000	17799
26870	19009
42100	16899
34200	18799
25056	18799
34212	17999
43070	17899
12431	20019
46221	16099
29000	18699
[illegible]	[illegible]

- Introduction to Regression

Example:

- We can see from the scatter plot that the variables are related.
- The Correlation between the variables is moderate to high negative (r = -0.79).
- As mileage increases, sales price of the car decreases.



- Using a statistical analysis, we can determine the regression model:

$$\text{Sales Price} = 21.015 - 0.0874 \times \text{Mileage} + \varepsilon$$

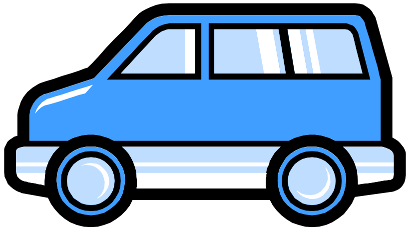
- Introduction to Regression

Example:

Sales Price = 21,015 – 0.0874 x Mileage + ε

- Use the regression equation above to predict what is the price of a vehicle when the mileage equals to 20,000?

- **Answer:** It will sell for about $19,267.



Example:

- We will use **R-Sq** to measure how much variability in the response is explained by the explanatory variable.
- As the points get closer to the regression line, R-Sq increases.

Regression Analysis: Sales Price versus Mileage

The regression equation is
Sales Price = 21015 - 0.0874 Mileage

Predictor	Coef	SE Coef	T	P
Constant	21014.6	246.3	85.31	0.000
Mileage	-0.087354	0.006860	-12.73	0.000

S = 537.537   R-Sq = 79.0%   R-Sq(adj) = 78.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	46855610	46855610	162.16	0.000
Residual Error	43	12424661	288946		
Total	44	59280271			

- The moderately high R-Sq value indicates that mileage greatly affect the sales price.
- However, other factors such as the condition of the car or its color may also influence the sales price.

The R2 Value:

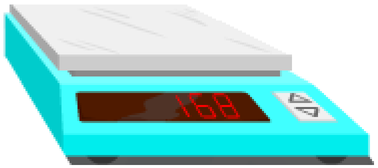
$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

$$0 \leq R^2 \leq 1$$

- R2 > 0.9 — Model can be used with full confidence.
- 0.7 < R2 < 0.9 — Model can be used carefully.
- R2 < 0.7 — Do not use the model.

Other Examples:

- The relationship between the height and the width of the man.
- The relation of the number of years of education someone has and that person's income.
- The relationship between the downtime of a machine and its cost of maintenance.



What About Attribute Data?

		Response (Y)	
		Variable	Attribute
Explanatory (Xs)	Variable	Regression	Logistic Regression
	Attribute	ANOVA	Contingency Table

Examples:

- Regression (Hardness of an alloy vs. its temperature).
- ANOVA (Shooting distance and ball material).
- Logistic reg. (% of discolored welds vs. current in welding process).
- Contingency Table (Process yield vs. Tool type).

Furthers Considerations:

- The Null and Alternative hypotheses must be clearly stated before the data is examined (or even collected).
- This hypotheses tests whether X can be considered a meaningful predictor of Y.

The Null Hypothesis → There is no relationship between X & Y.

Predictor	Coef	SE Coef	T	P
Constant	19959.0	582.4	34.27	0.000
Goals 2 previous weeks	1148.9	156.6	7.34	0.000

S = 2071.48   R-Sq = 62.0%   R-Sq(adj) = 60.8%

Analysis of Variance

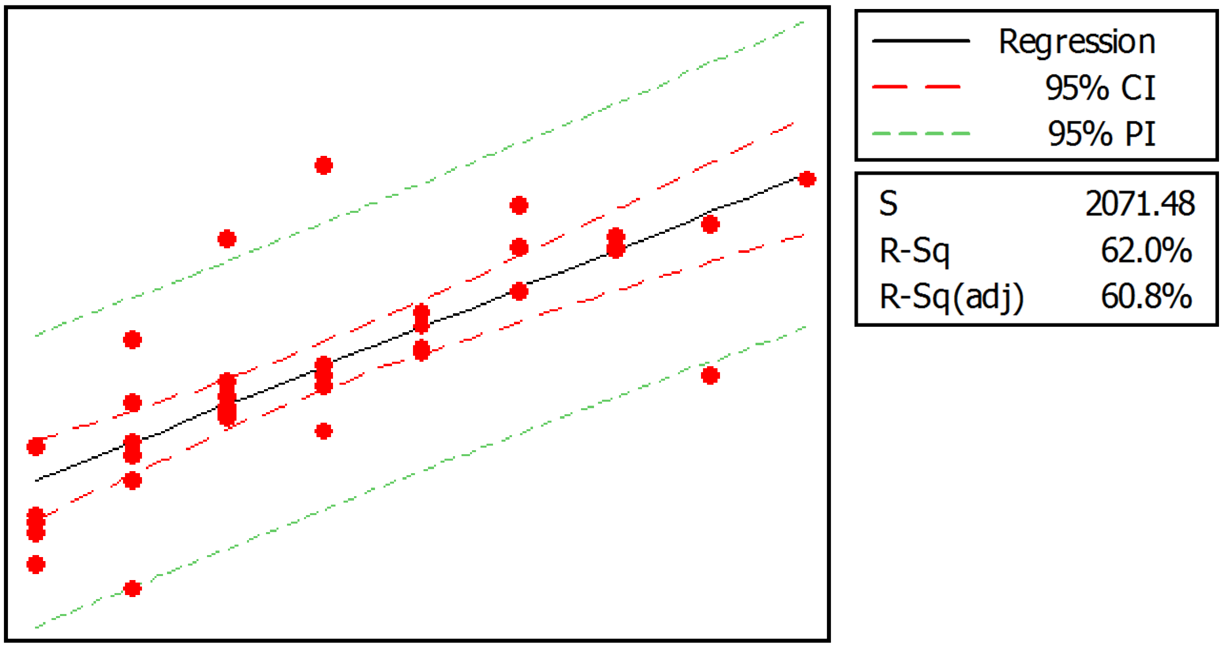
Source	DF	SS	MS	F	P
Regression	1	230953073	230953073	53.82	0.000
Residual Error	33	141603641	4291019		
Total	34	372556714			

As p-value<0.05, are confident there is a relationship between the two variables?

- Introduction to Regression

Furthers Considerations:

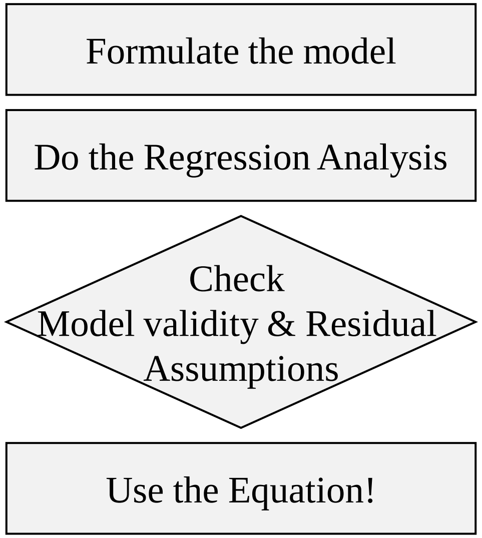
- Prediction and confidence intervals.



- Introduction to Regression

Further Information:

- For our regression model to be valid, we must be sure that the residuals can be explained by random error in the process.
- We must test the following assumptions:
  - The errors are random (each error is independent of each other error).
  - The errors are normally distributed with mean zero.
  - The errors variance does not change for different levels of x.



Further Information:

- Always perform a MSA before you do a regression because the measurement error will affect your R-Sq and the quality of your model.
- You should not use the model beyond the bounds of the data used to create it.
- In reality, the result of a process is rarely relationship with one input variable but instead more complex results of several factors.
- Forecasts must always be constantly compared with actual outcomes, and the effectiveness of the forecast reviewed.
- Only do the regression if it adds value.