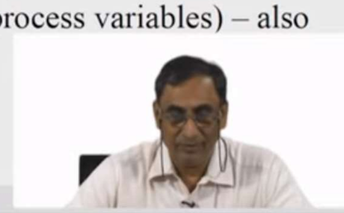
LINEAR REGRESSION

Motivation

- Purpose is to build a functional relationship (model) between *dependent variable(s)* and *independent variable(s)*
- Example
  - Business : What is the effect of price on sales? (Can be used to fix the selling price of an item)
  - Engineering : Can we infer difficult to measure properties of a product from other easily measured variables? (mechanical strength of a polymer from temperature, viscosity or other process variables) – also known as a soft sensor



Regression - Basics

- One of the widely used statistical techniques
- Dependent variables also known as *Response variable, Regressand, Predicted variable, output variable* - denoted as variable/s y
- Independent variable also known as *Predictor variable, Regressor, Exploratory variable, input variable* - denoted as variable/s x

Regression types

- Classification of Regression Analysis
  - Univariate vs Multivariate
    - *Univariate*: One dependent and one independent variable
    - *Multivariate*: Multiple independent and multiple dependent variables
  - Linear vs Nonlinear
    - *Linear*: Relationship is linear between dependent and independent variables
    - *Nonlinear:* Relationship is nonlinear between dependent and independent variables
  - *Simple vs Multiple*
    - Simple: One dependent and one independent variable (SISO)
    - Multiple: One dependent and many independent variables (MISO)

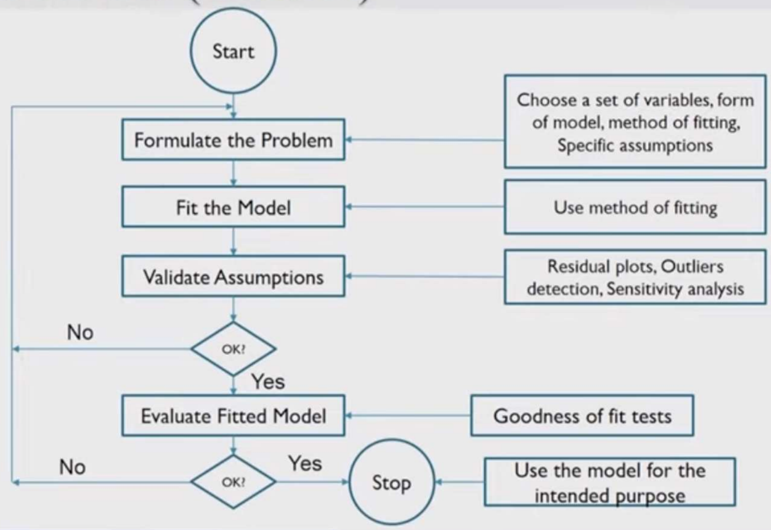
Regression analysis

- Is there a relationship between these variables?
- Is the relationship linear and how strong is the relationship?
- How accurately can we estimate the relationship?
- How good is the model for prediction purposes?

Regression methods

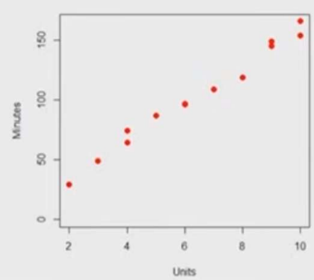
- Linear regression methods
  - Simple linear regression
  - Multiple linear regression
  - Ridge regression
  - Principal component regression
  - Lasso
  - Partial least squares
- Nonlinear regression methods
  - Polynomial regression
  - Spline regression
  - Neural networks

Regression Process (Iterative)



Ordinary Least Squares (OLS)

- Fourteen observations obtained on time taken in minutes for service calls and number of units repaired
- Objective is to find relationship between these variables (useful for judging service agent performance)



Ordinary Least Squares (OLS)

- Linear model between y_i and x_i, $i = 1,\dots, n$

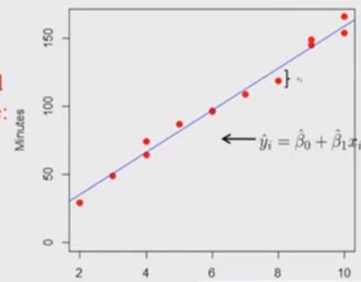
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Error in only dependent variable and no error in independent variable:

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$$

- The sum of squares of errors (SSE)

$$\sum_i \epsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$



- The minimization of SSE gives estimates of β_0 and β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

GyanData Private Limited

OLS: Testing Goodness of Fit

- Prediction using the regression equation: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Coefficient of determination - R^2 is a measure of variability in output variable explained by input variable

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

Variability explained by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Total variability in y

- R^2 values: Between 0 and 1
  - Values close to 0 indicates poor fit
  - Values close to 1 indicates a good fit (However, should not be used as sole criterion to judge that a linear model is adequate)
- Adjusted $\bar{R}^2$

$$\bar{R}^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2/(n-p-1)}{\sum (y_i - \bar{y})^2/(n-1)}$$

OLS: Example using R

```
Call:
lm(formula = Minutes ~ Units)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2318 -3.3415 -0.7143  4.7769  7.8033

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.162      3.355    1.24    0.239
Units         15.509      0.505   30.71 8.92e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.392 on 12 degrees of freedom
Multiple R-squared:  0.9874, Adjusted R-squared:  0.9864
F-statistic: 943.2 on 1 and 12 DF,  p-value: 8.916e-13
```

$\boxed{\hat{\beta}_0}$ $\boxed{\hat{\beta}_1}$