# STAT 302 - Chapter 1 : Simple Linear Regression - Part 1

Harsha Perera

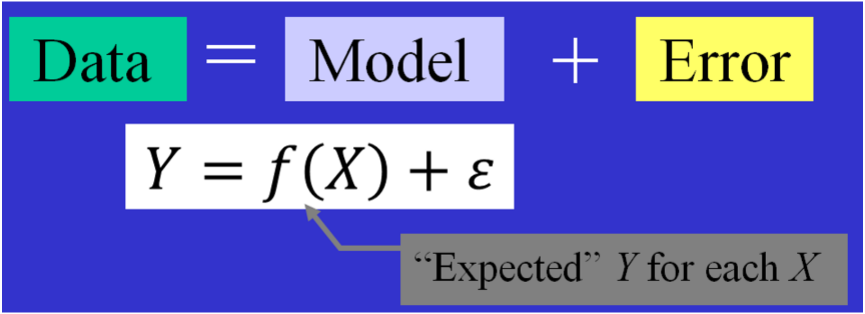
# Key Topics

- The Simple Linear Regression Model
- Conditions for a Simple Linear Regression Model
- Parameter Estimation

# The Simple Linear Regression Model

Notation: **Y** = Response variable **X** = Predictor variable

Assume (for now) that both Y and X are quantitative variables.



# Regression Line

A regression line is a straight line that describes how a response variable Y changes as an explanatory variable X changes. We often use a regression line to predict the value of y for a given value of x, when we believe the relationship between Y and X is linear.

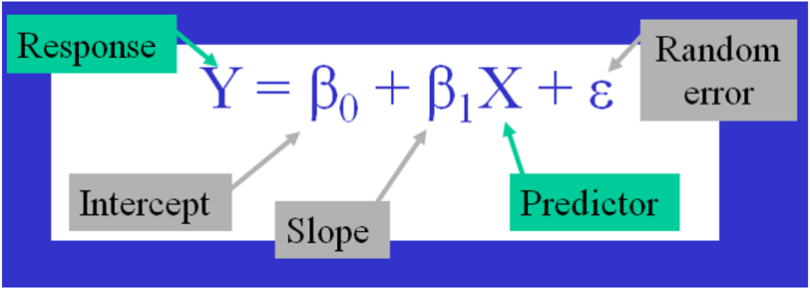
Some Examples for X and Y ;

- Y = final exam score
  X = midterm exam score

- Y = active pulse rate (after exercise)
  X = resting pulse rate

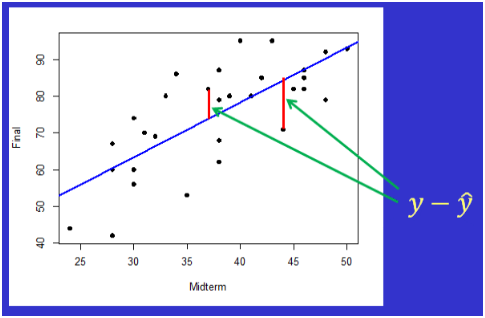
- Y = Price of an used car
  X = Mileage

# Simple Linear Regression



# The Least-Squares Regression Line

The **Least-Squares Regression Line** of Y on X is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.




$$Minimize\ SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

# The Least-Squares Regression Line

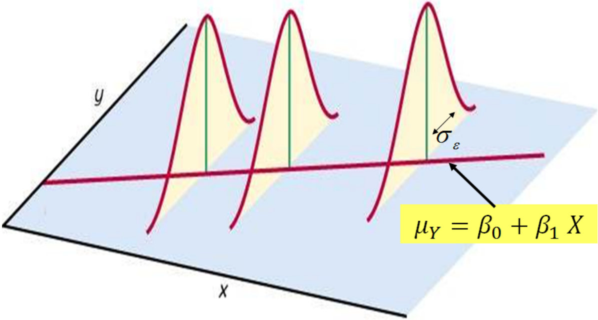
- Notation : Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the sample intercept and slope of the predicted line.
- The Least-Squares Regression Line : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- The residual at each point is (Actual – Predicted) $=$ $y$ - $\hat{y}$
- Choose the line to minimize :

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

# Conditions For a Simple Linear Regression Model

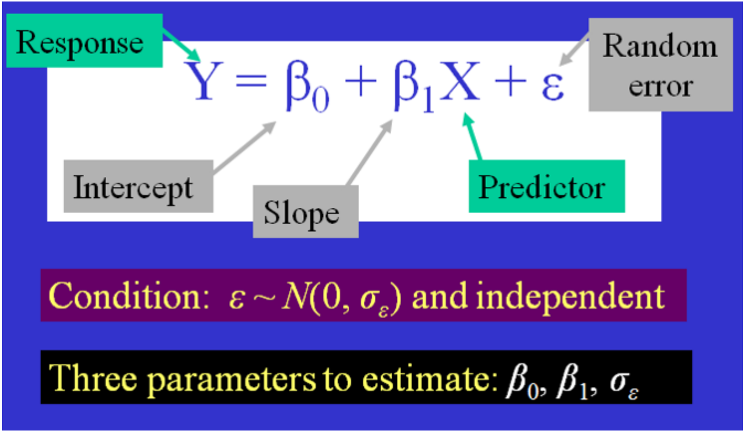
- **Model is Linear** : The overall relationship between variables has a linear pattern. The means for Y vary as a linear function of X. i.e $\mu_y = \beta_0 + \beta_1 X$
- **Errors have a Zero Mean** : The distribution of the errors is centered at zero.
- **Errors have a uniform spread/Constant Variance** : The variance of the response does not change as the predictor changes. Therefore the variance for Y is the same at each X (homoscedasticity).
- **Errors are Independent** : The errors are assumed to be independent from one another.
- **Errors are Normally Distributed** : For inference, the formulas for tests and intervals assume that the unseen errors in the model follow a normal probability distribution.

# Conditions for a Simple Linear Regression Model




- For each possible value of the explanatory variable $x$, the mean of the responses $\mu_y$ moves along this population regression line.
- The Normal curves show how $y$ will vary when $x$ is held fixed at different values. All of the normal curves have the same standard deviation $\sigma_\epsilon$, so the variability of $y$ is the same for all values of $x$.

# Simple Linear Regression Model



# Parameter Estimation

- $\beta_0 \rightarrow \hat{\beta}_0$
- $\beta_1 \rightarrow \hat{\beta}_1$
- $\sigma_\epsilon \rightarrow \hat{\sigma}_\epsilon = \sqrt{\frac{SSE}{n-2}}$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.6721     9.3311   2.001   0.0548 .
Midterm       1.4925     0.2413   6.186 9.58e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.651 on 29 degrees of freedom
Multiple R-squared: 0.5689,	Adjusted R-squared: 0.554
F-statistic: 38.26 on 1 and 29 DF,  p-value: 9.582e-07
```

$\hat{\sigma}_\varepsilon$

$n-2$

$$\widehat{Final} = 18.67 + 1.49 \cdot Midterm$$