Machine Learning for Big Data: Simple Linear Regression

Lionel Fillatre

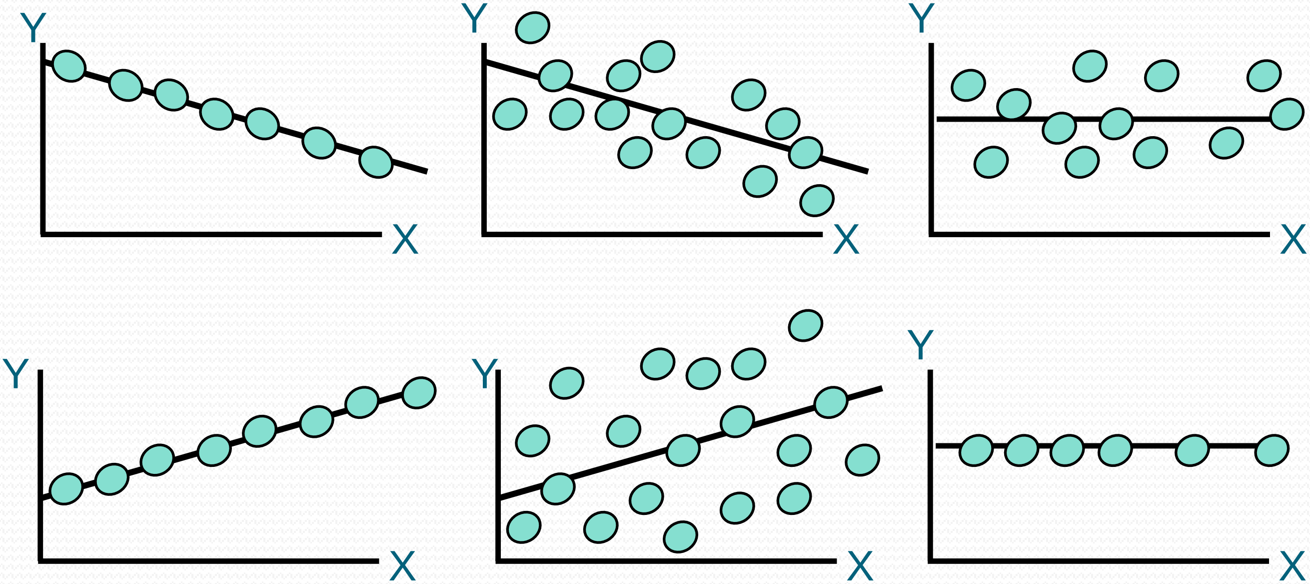
fillatre@unice.fr

Topics

- Introduction
- Compute the best linear model
- Testing significance
- Residuals analysis
- Conclusion

1 Introduction

Scatter Plots of Two Variables



Simple Linear Regression Model

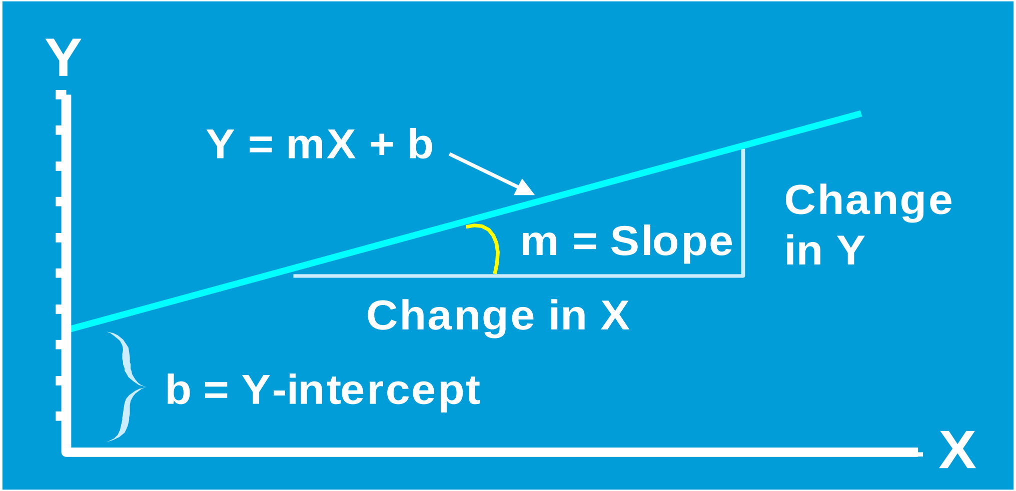
- The equation that describes how y is related to x and an error term is called the **regression model**.
- The **simple linear regression model** is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

  - β_0 and β_1 are called **parameters of the model**.
  - ε is a random variable called the **error term**. Generally, this error is centered:

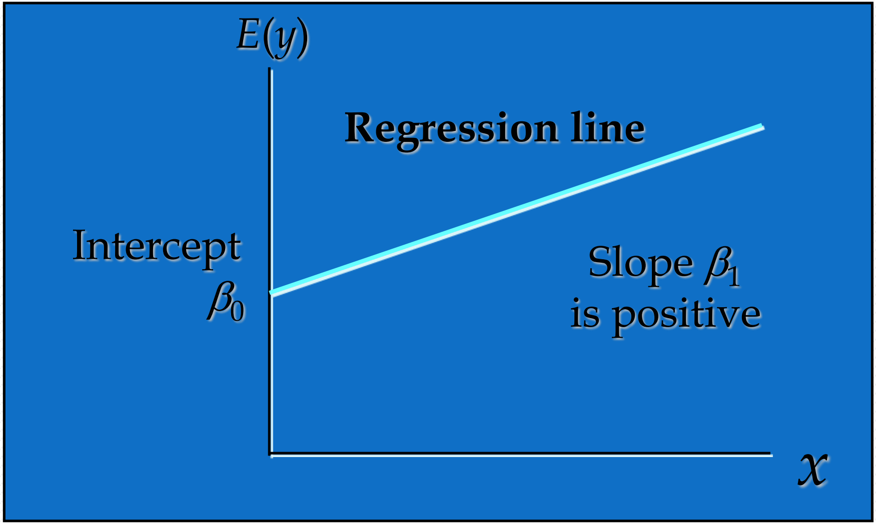
$$E(\varepsilon) = 0$$

Linear Model $E(y) = \beta_0 + \beta_1 x$



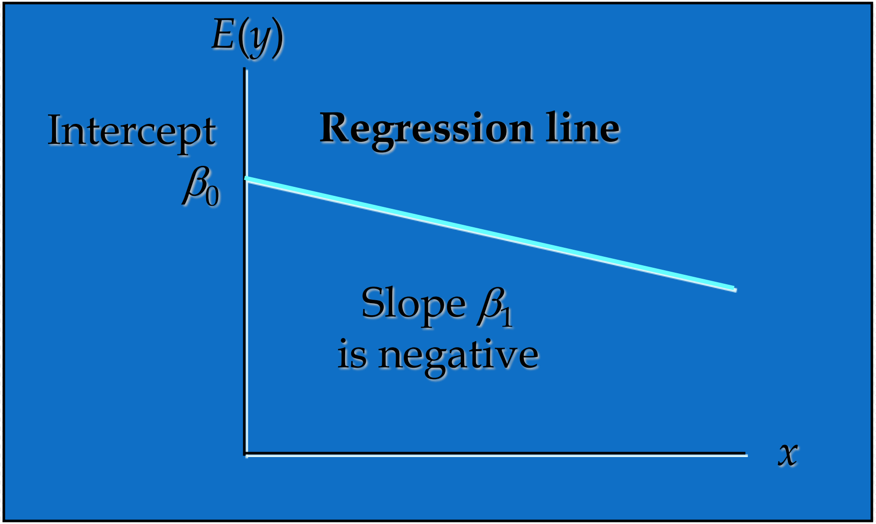
Simple Linear Regression Equation

Positive Linear Relationship



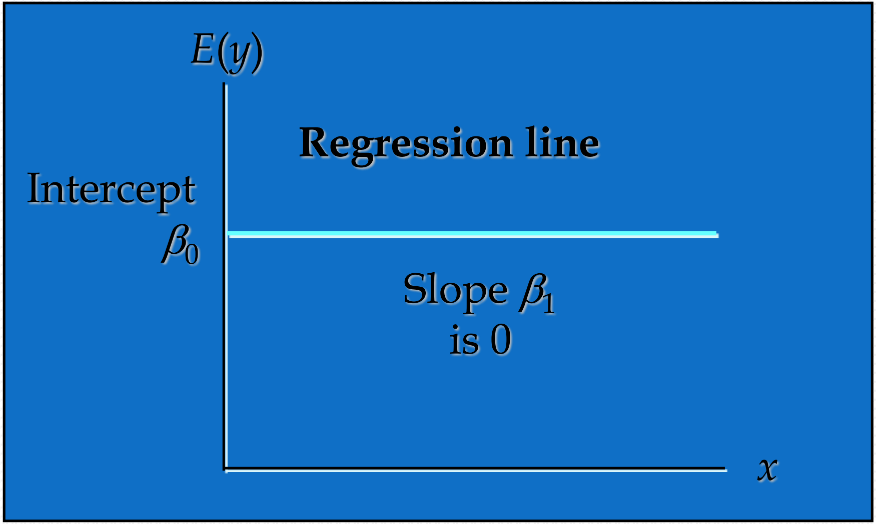
Simple Linear Regression Equation

Negative Linear Relationship



Simple Linear Regression Equation

No Relationship



2 Compute the best linear model

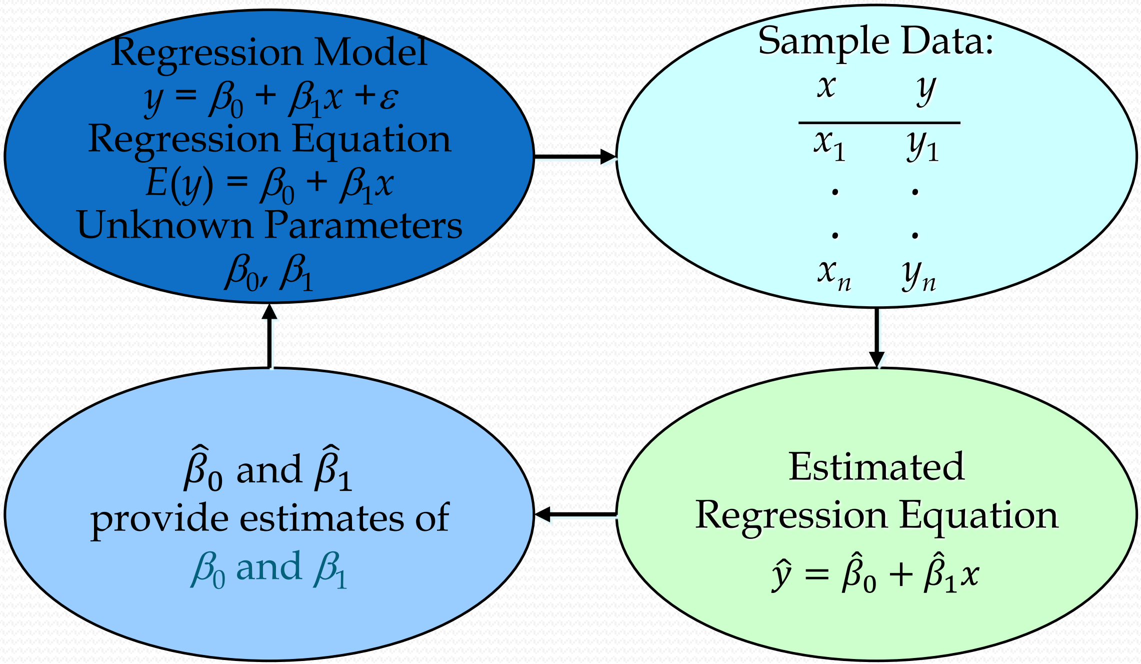
Estimated Regression Equation

The estimated equation is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The graph is called the estimated regression line.
- $\hat{\beta}_0$ is the y intercept of the line.
- $\hat{\beta}_1$ is the slope of the line.
- $\hat{y}$ is the estimated value of y for a given x value.

Estimation Process



Least Squares Method

- **Least Squares Criterion (SSE = Sum of Squared Errors)**

$$SSE = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where:

- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ = observed value of the dependent variable for the ith observation
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ = estimated value of the dependent variable for the ith observation

Coefficient Equations

- Prediction equation

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

- Sample slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

- Sample Y-intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{with} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Derivation of Estimates (1)

- Least Squares (L-S): Minimize squared error

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i \right)^2}{\partial \beta_0}$$

$$= -2\left(n\overline{y} - n\beta_0 - n\beta_1 \overline{x} \right)$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Derivation of Estimates (2)

- Least Squares (L-S): Minimize squared error

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i \right)^2}{\partial \beta_1}$$
$$= -2 \sum x_i \left(y_i - \beta_0 - \beta_1 x_i \right)$$
$$= -2 \sum x_i \left(y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i \right)$$

$$\beta_1 \sum x_i \left(x_i - \bar{x} \right) = \sum x_i \left(y_i - \bar{y} \right)$$
$$\beta_1 \sum \left(x_i - \bar{x} \right)\left(x_i - \bar{x} \right) = \sum \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right)$$
$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

3 Testing significance

The Coefficient of Determination

- We compare our fit to a null model $y_i = \alpha + \varepsilon_i$, in which we don't use the independent variable x

- Analysis of Variance = relationship among SST, SSR, SSE

$$\text{SST} = \text{SSR} + \text{SSE}$$

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

where:

- SST = total sum of squares
- SSR = sum of squares due to regression
- SSE = sum of squares due to error

The Coefficient of Determination

- The coefficient of determination is:

$$r^2 = \text{SSR}/\text{SST}$$

  where:
  - SST = total sum of squares
  - SSR = sum of squares due to regression

- r^2 is the proportional reduction in squared error due to the linear regression.

- "Good" values of r^2 vary widely in different fields of application.

The Correlation Coefficient

- The correlation coefficient gives the strength and direction of the relationship.

- Sample Correlation Coefficient

$$r_{xy} = (\text{sign of } \hat{\beta}_1)\sqrt{r^2}$$

where:

- $\hat{\beta}_1$= the slope of the estimated regression equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Model Assumptions

- Conventional assumptions about the error term ε
  1. The error ε is a random variable with mean of zero.
  2. The variance of ε, denoted by σ^2, is the same for all values of the independent variable.
  3. The values of ε are independent.
  4. The error ε is a normally distributed random variable.

Testing for Significance

- To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero.
- Two tests are commonly used
  - t Test
  - F Test (not presented here)
- Both tests require an estimate of σ^2, the variance of ε in the regression model.

Testing for Significance

- An estimate of σ^2 :
  - The mean square error (MSE) provides the estimate of σ^2 :

$$\widehat{\sigma^2} = \text{MSE} = \text{SSE}/(\text{n-2})$$

where

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- An estimate $\hat{\sigma}$ of σ :
  - To compute $\hat{\sigma}$ we take the square root of $\widehat{\sigma^2}$.
  - The resulting $\hat{\sigma}$ is called the standard error of the estimate.

Testing for Significance: t Test

- Hypotheses:

$$H_o: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

- Test Statistic: $t = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$

where $s_{\hat{\beta}_1} = \dfrac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}$

Testing for Significance: *t* Test

- Rejection Rule:
  - Reject H_0 if $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
- where:
- $t_{\alpha/2}$ is based on a t distribution with n - 2 degrees of freedom
- $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of the t distribution.
- Meaning of α:
  - the significance level α is the probability of rejecting the null hypothesis when it is true.
  - For example, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

Some Cautions about the Interpretation of Significance Tests

- Rejecting H_0: $\beta_1 = 0$ and concluding that the relationship between *x* and *y* is significant does not enable us to conclude that a **cause-and-effect relationship** is present between *x* and *y*.

- Just because we are able to reject H_0: $\beta_1 = 0$ and demonstrate statistical significance does not enable us to conclude that there is a **linear relationship** between *x* and *y*.

4 Residual Analysis

Residual Analysis

- Residual for observation *i*

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

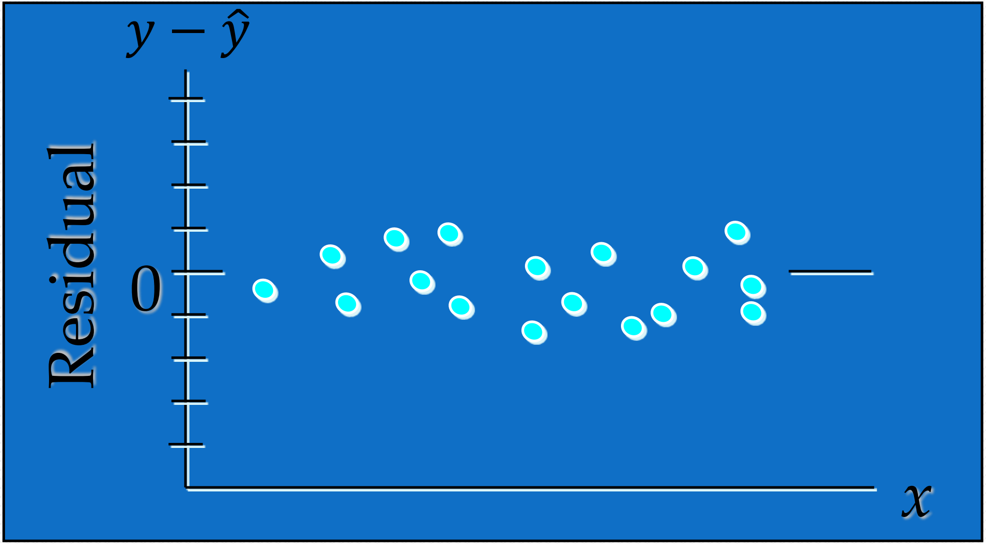
- Standardized Residual for observation *i*

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

where: $$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

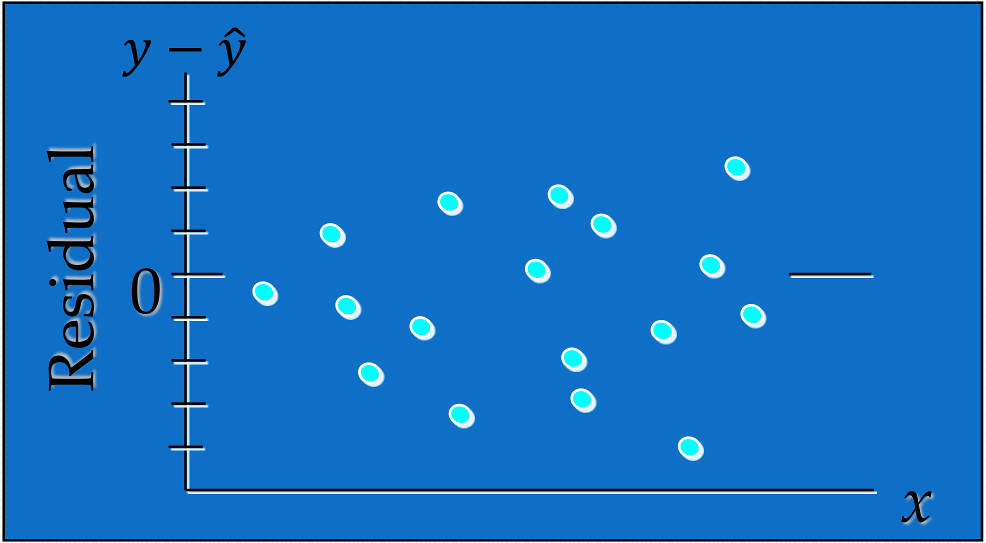
Residual Analysis

- Residual Plot: Good pattern

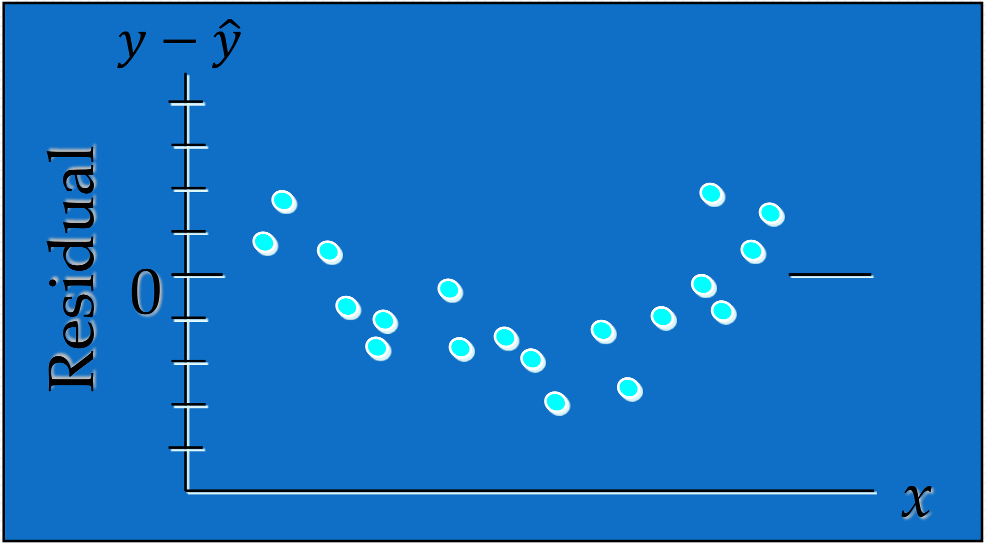


Residual Analysis

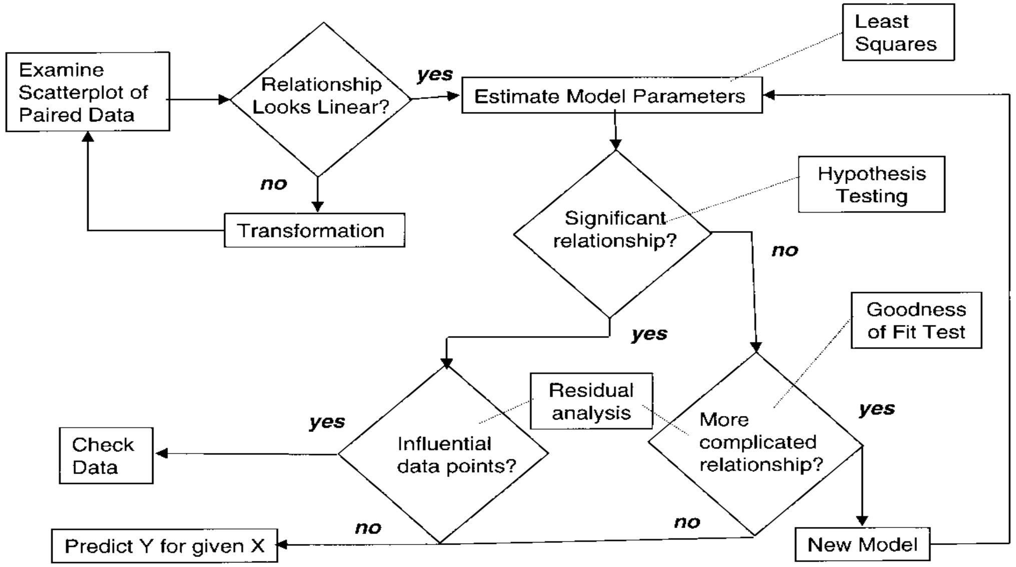
- Residual Plot: Nonconstant variance



Residual Analysis

- Residual Plot: Model form not adequate
- 

How is a Simple Linear Regression Analysis done? A Protocol



5 Conclusion

Conclusion

- Linear regression: a very famous parametric method!
- Many tools for interpreting the results
- Interpretation should be done carefully
- Extension to multiple linear regression