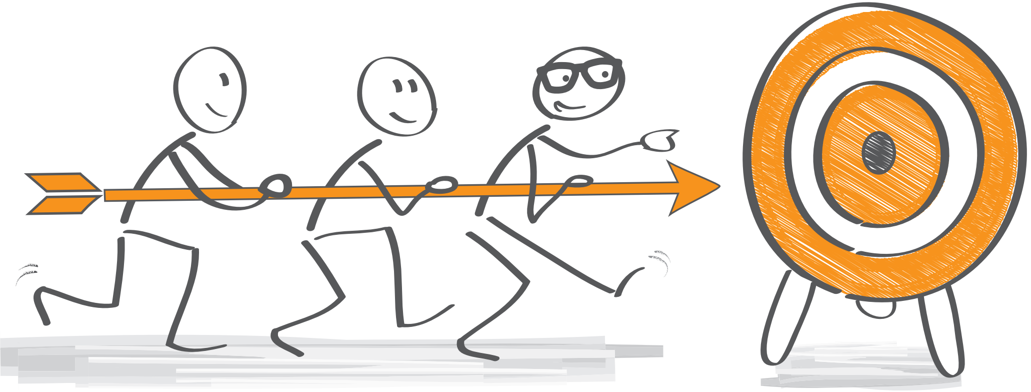
Regression Analysis
Regression Methods

Nicoleta Serban, Ph.D.
Associate Professor
Stewart School of Industrial and Systems Engineering

Regression Analysis: Overview

About this lesson



Simple Linear Regression

Data: $\{(x_1,Y_1),....,(x_n,Y_n)\}$

Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1,...,n$

Assumptions:

- *Linearity/Mean Zero Assumption :* $E(\varepsilon_i) = 0$
- *Constant Variance Assumption*: $Var(\varepsilon_i) = \sigma^2$
- *Independence Assumption* $\{\varepsilon_1,..., \varepsilon_n\}$ are independent random variables
- *Normality Assumption*: $\varepsilon_i \sim$ Normal

ANOVA

Data: Y_{ij} for $j = 1, \ldots, n_i; i = 1, \ldots, k$

Model: $Y_{ij} = \mu_i + \varepsilon_{ij}$ where ε_{ij} is an error term

Assumptions:

- *Constant Variance Assumption*: $\mathrm{Var}(\varepsilon_{ij}) = \sigma^2$
- *Independence Assumption*: $\{\varepsilon_{1j}, \ldots, \varepsilon_{nj}\}$ are independent random variables
- *Normality Assumption*: $\varepsilon_{ij} \sim \mathrm{Normal}(0, \sigma^2)$

Multiple Linear Regression

Data: $\{(x_{11},.., x_{1p}),Y_1\},....,\{(x_{n1},.., x_{np}),Y_n\}$

Model: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i$, $i = 1,...,n$

Assumptions:

- *Linearity/Mean Zero Assumption*: $E(\varepsilon_i) = 0$
- *Constant Variance Assumption*: $Var(\varepsilon_i) = \sigma^2$
- *Independence Assumption*: $\{\varepsilon_1,..., \varepsilon_n\}$ are independent random variables
- *Normality Assumption*: $\varepsilon_i \sim$ Normal

Logistic Regression

Data: $\{(x_{11},.., x_{1p}),Y_1\},....,\{(x_{n1},.., x_{np}),Y_n\}$ where $Y_1,.., Y_n$ binary responses

Model: Probability of success given predictor(s)

$$p = p(x_1,...,x_p) = P\{Y=1|x_1, \dots, x_p\}$$

link *p* to the predicting variables through ***logit*** *link function*

$$g(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0+\beta_1 x_1 + ... + \beta_p x_p$$

Assumptions:

- *Linearity Assumption*: $g\{p(x_1,...,x_p)\} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$
- *Independence Assumption*: $Y_1,.., Y_n$ are independent random variables
- *Logit link function*: $g(p) = \ln\left(\frac{p}{1-p}\right)$

Poisson Regression

Data: $\{(x_{11},.., x_{1p}),Y_1\},....,\{(x_{n1},.., x_{np}),Y_n\}$ where $Y_1,.., Y_n$ count response variable

Model: Model the conditional expectation**:**

$$\log(E(Y|x_1, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Assumptions:

- *Linearity Assumption*: $\log(E(Y|x_1, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
- *Independence Assumption*: $Y_1,.., Y_n$ are independent random variables
- *Variance Assumption*: $E(Y|x_1, \ldots, x_p) = V(Y|x_1, \ldots, x_p)$

Generalized Linear Model

Data: $\{(x_{11},.., x_{1p}),Y_1\},....,\{(x_{n1},.., x_{np}),Y_n\}$ where $Y_1,.., Y_n$ response variable with **a distribution from the exponential family**

Model: Model the conditional expectation**:**

$$g(E(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

OR

$$E(Y|x_1, \dots, x_p) = g^{-1}(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)$$

where g( ) is a *link function* and g^{-1}( ) the *inverse link function* depending on the distribution of Y.

Weighted Least Regression (WLS)

Data: $\{(x_{11},.., x_{1p}),Y_1\},....,\{(x_{n1},.., x_{np}),Y_n\}$

Model: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i$, $i = 1,...,n$

Assumptions: For the vector of errors $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$

- *Linearity/Mean Zero Assumption*: $E(\varepsilon) = 0$
- *Covariance-Variance Assumption*: $V(\varepsilon) = \Sigma$
- *Independence Assumption*: $\{\varepsilon_1,..., \varepsilon_n\}$ are independent random variables
- *Normality Assumption*: $\varepsilon \sim$ Normal

Generalized Additive Model (GAM)

Model: $Y_i = f(x_{i1},\ldots,x_{ip}) + \varepsilon_i$, $i = 1,\ldots,n$ with

$f(x_1,\ldots,x_p) = \alpha + f_1(x_1) + \ldots + f_p(x_p)$ where $f_1,\ldots, f_p$ are unknown smooth functions

Model Estimation:

- Backfitting algorithm:
  - Initialize: $\hat{\alpha}$, $\hat{f}_1,\ldots, \hat{f}_p$
  - Iterate until convergence: For $j=1,..,p$

$\check{R}_i = Y_i - \hat{\alpha} - \sum_{k\neq j} \hat{f}_k(x_{ki})$ and estimate $\hat{f}_j$ from regressing $\check{R}_i \sim x_{ji}$

Summary