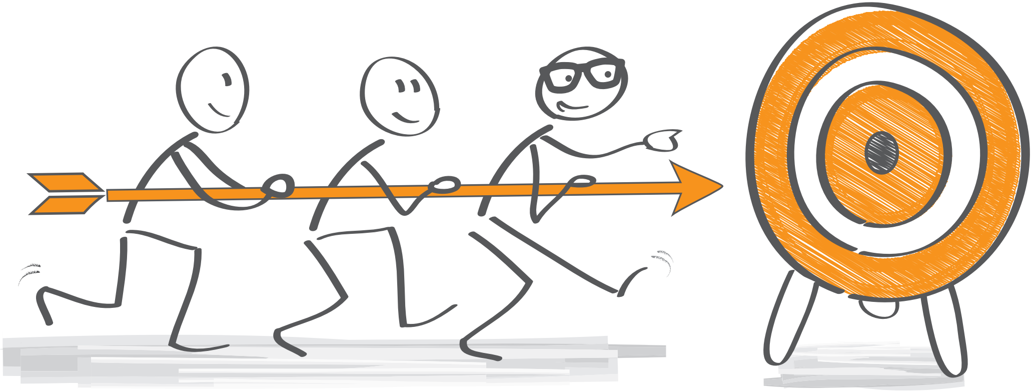
Regression Analysis
Other Regression Methods

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

Weighted Least Squares Regression

GTx

About this lesson



Multiple Linear Regression

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

What if the variance is not constant?

- **Transform the response variable using a variance-stabilizing transformation**
- **Weighted Least Squares Regression**

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

Example: Normal Approximation

- Assume D_i the number of diseased individuals in a population of size m_i; and $Y_i = \frac{D_i}{m_i}$
- Only observe Y_i but generally m_i is large, thus apply the normal approximation (CLT):
    - Use a regression analysis under the normality assumption instead of logistic regression
    - $\mathrm{V}(Y_i) = \frac{\sigma^2}{m_i}$ thus non-constant variance

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 <u>if Σ is a diagonal matrix</u>
- *Normality Assumption*: $\varepsilon \sim$ Normal

Parameter Estimation $(\beta_0,\beta_1,\beta_2,\ldots,\beta_p)$

To estimate $(\beta_0,\beta_1,\beta_2,\ldots,\beta_p)$, we find values that minimize squared error:

$$(Y - X\beta)^T \Sigma^{-1} (Y - X\beta)$$

$$\Rightarrow \quad \hat{\beta} = \left(\mathbf{X}^T \Sigma^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \Sigma^{-1} Y$$

Statistical Properties:

$$E(\hat{\beta}) = \beta$$

$$V(\hat{\beta}) = \sigma^2 \left(X^T \Sigma^{-1} X\right)^{-1}$$

Upshot: The covariance-variance matrix of the error terms is assumed known. However, it is needed for statistical inference. *How to get Σ?*

Simple WLS

The simplest WLS model: $\mathrm{V}(\varepsilon) = w_i \sigma^2$

How to implement WLS in R?

- **lm(y~x, weights = 1/w) where w is a vector of the weights**

How to estimate $w_i = \mathrm{w}(x_i)$ as a smooth function of x_i?

- **Several smoothing functions: simplest to use is 'lowess'**

  - Use external information: There are some cases where other information on the variance is available (e.g. measurement error) – <u>very rare</u>!
  - Use replications. If there are several Y's for each x_i, estimate $w_i \sigma^2 = \sigma_i{}^2$ as the sample variance of the replications
  - Estimate $w_i = \mathrm{w}(x_i)$ as a smooth function of x_i using nonparametric regression

Summary