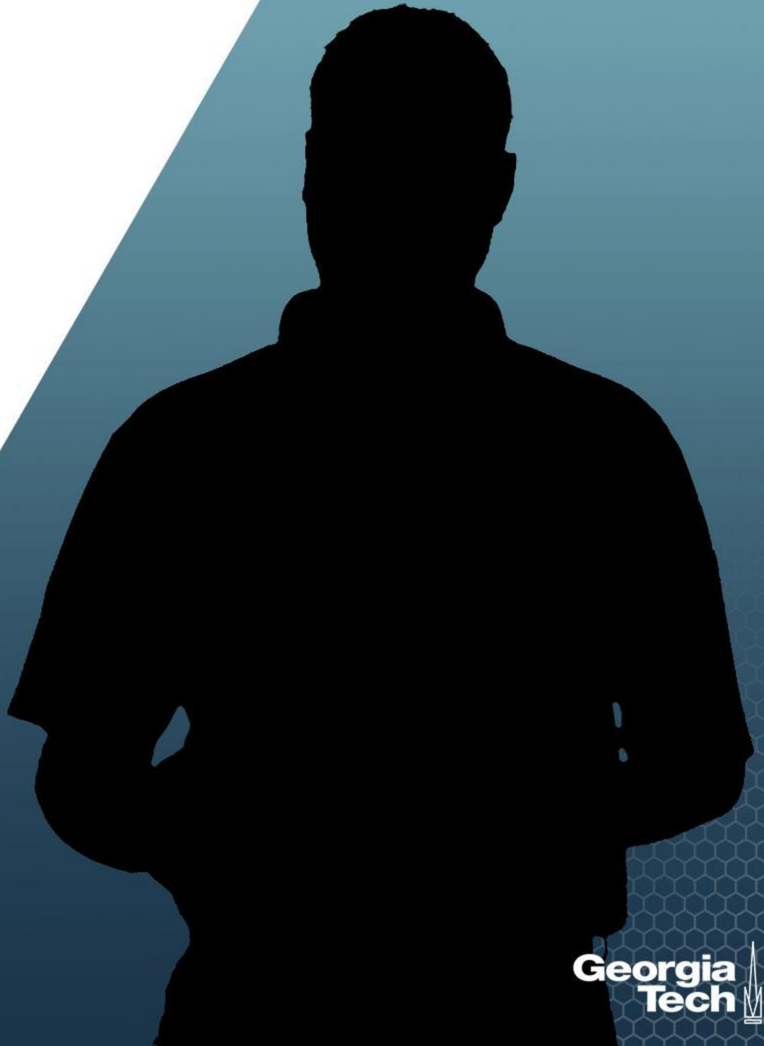
Regression Analysis

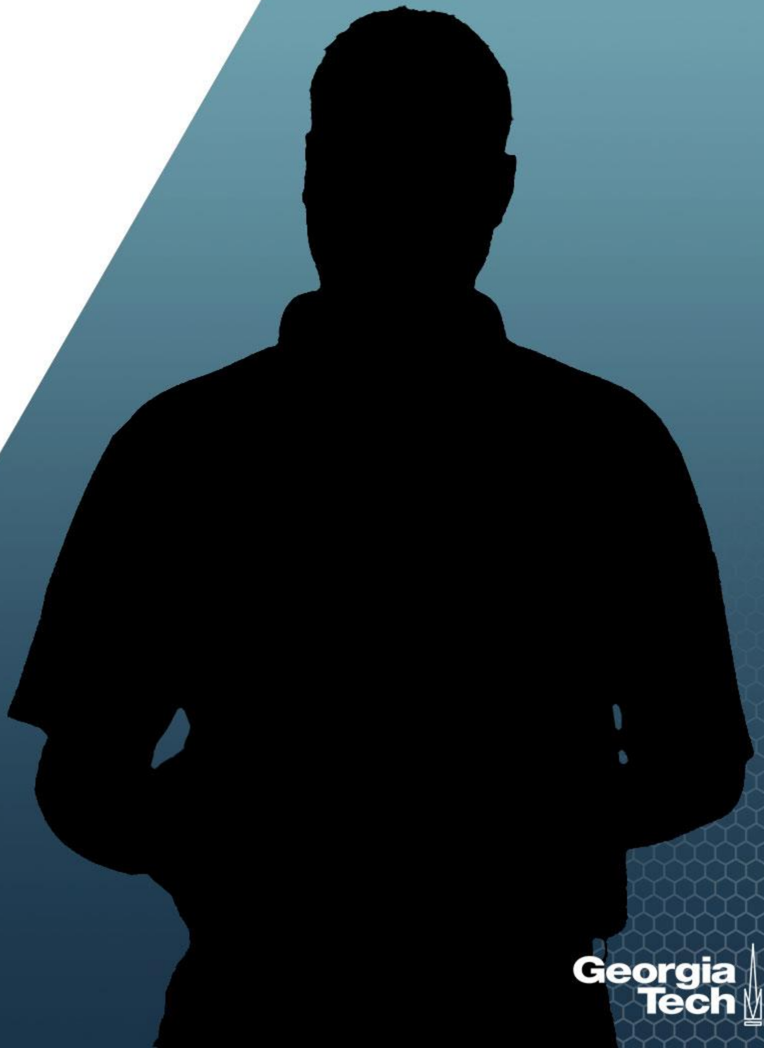
Simple Linear Regression

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

Regression Concepts:
Assumptions and Diagnostics



About This Lesson



Simple Linear Regression: Model

Data: $\{(x_1,y_1),\ldots,(x_n,y_n)\}$
Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \; i = 1,\ldots,n$

Assumptions:

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

Residual Analysis

Residual Values: $\varepsilon_{\mathrm{i}} \rightarrow \hat{\varepsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Graphical display: **Plot of the residuals $\boldsymbol{\varepsilon}_{\mathbf{i}}$**

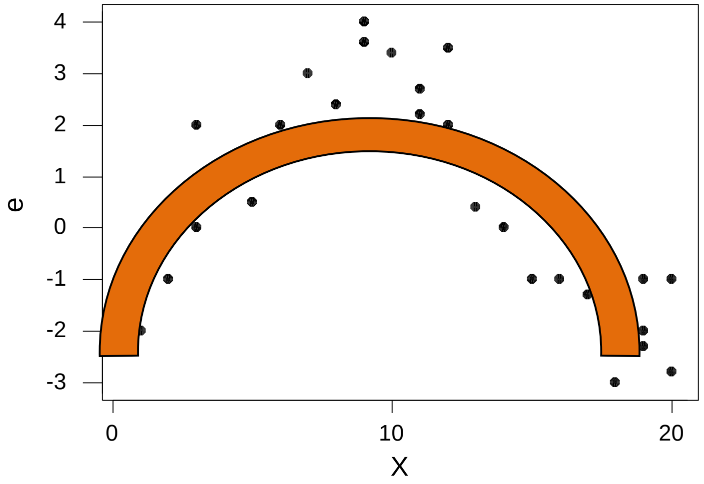
If the scatter of ε_{i} is **not random around zero line**, it could be that

- The relationship between X and Y is not linear
- Variances of error terms are not equal
- Response data are not independent

Checking Assumptions: Residual Analysis

Linearity Assumption:

This shows that there may be a non-linear relationship between X and Y.

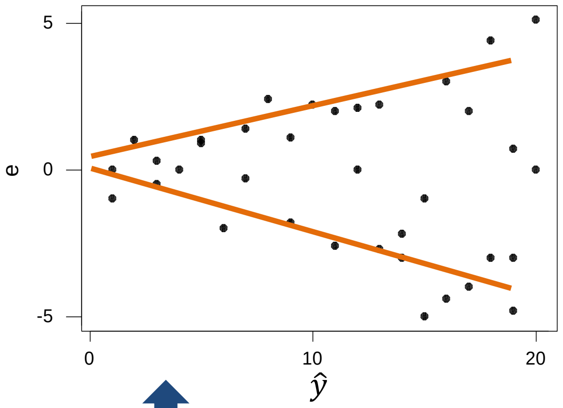


Georgia Tech

Checking Assumptions: Residual Analysis

Constant Variance Assumption:

The residuals show larger variance as the fitted values increase.

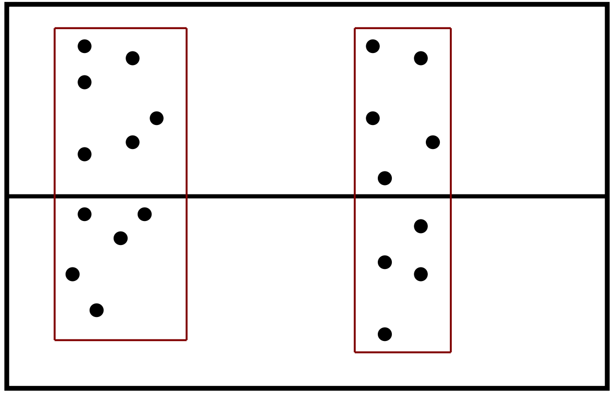


Here, it could be that σ^2 is not constant.

Checking Assumptions: Residual Analysis

Independence Assumption:

There are clusters of residuals: the independence assumption does not hold.



- **Using residual analysis, we check for uncorrelated errors but not independence.**
- **Independence is a more complicated matter. If the data are from a randomized trial, then independence is established, but most data are from observational studies.**

Checking the Assumption of Normality

One way to check this assumption in a regression is using a
Normal Probability Plot

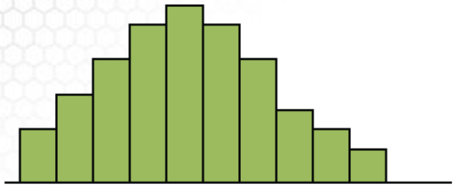
x-axis:	$\Phi^{-1}\left(\dfrac{r_i - 3/8}{n + 1/4}\right)$
y-axis:	e_i

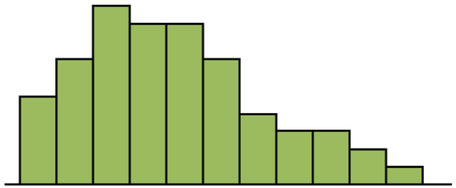
r_i = rank of e_i (between 1, n)

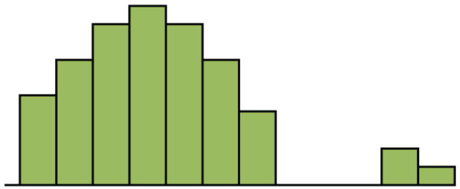
Φ = CDF of Normal Distribution

- Let the R statistical software do this for you!
- A straight line in normal probability plot implies assumption of normality is valid
- **Curvature** (especially at the ends) shows non-normality

Checking the Assumption of Normality







A complementary approach to check for the normality assumption is by plotting the **histogram** of the residuals

Normality Assumption:
The residuals should have an approximately symmetric distribution, unimodal, and with no gaps in the data.

Georgia Tech

Variable Transformation

- If the model fit is inadequate, it does not mean that a regression is not useful.

- One problem might be that the relationship between **X** and **Y** is ***not exactly linear***.

- To model the nonlinear relationship, we can transform **X** by some nonlinear function such as:

$$f(x) = x^a \quad \text{or} \quad f(x) = \log(x)$$

Normality Transformations

Problem: Normality or constant variance assumption does not hold.
Solution: Transform the response variable from y to y* via

$$y^* = y^{\lambda}$$

where the value of λ depends on how Var(Y) changes as X changes.

$\sigma_y(x) \propto const$	$\lambda = 1$	*(don't transform)*
$\sigma_y(x) \propto \sqrt{\mu_x}$	$\lambda = 1/2$	
$\sigma_y(x) \propto \mu_x$	$\lambda = 0$	$y^* = \ln(y)$
$\sigma_y(x) \propto 1/\mu_x$	$\lambda = -1$	

This is called Box-Cox Transformation: The parameter λ can be determined using R statistical software.

Georgia Tech

Summary