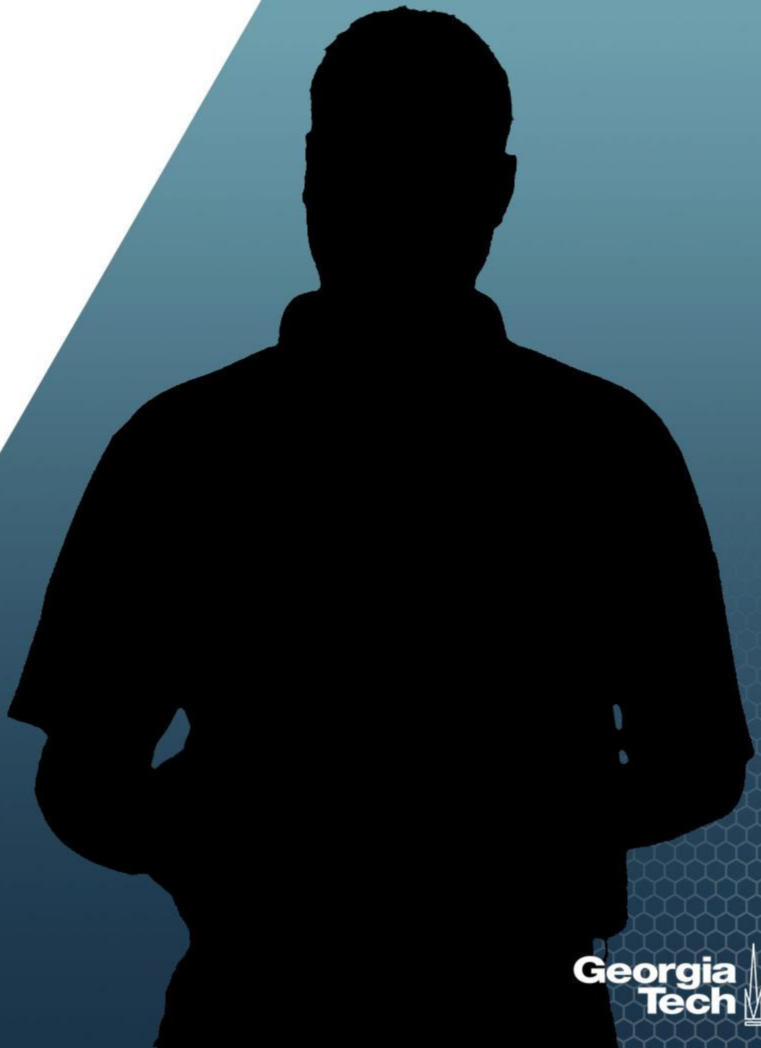
Regression Analysis

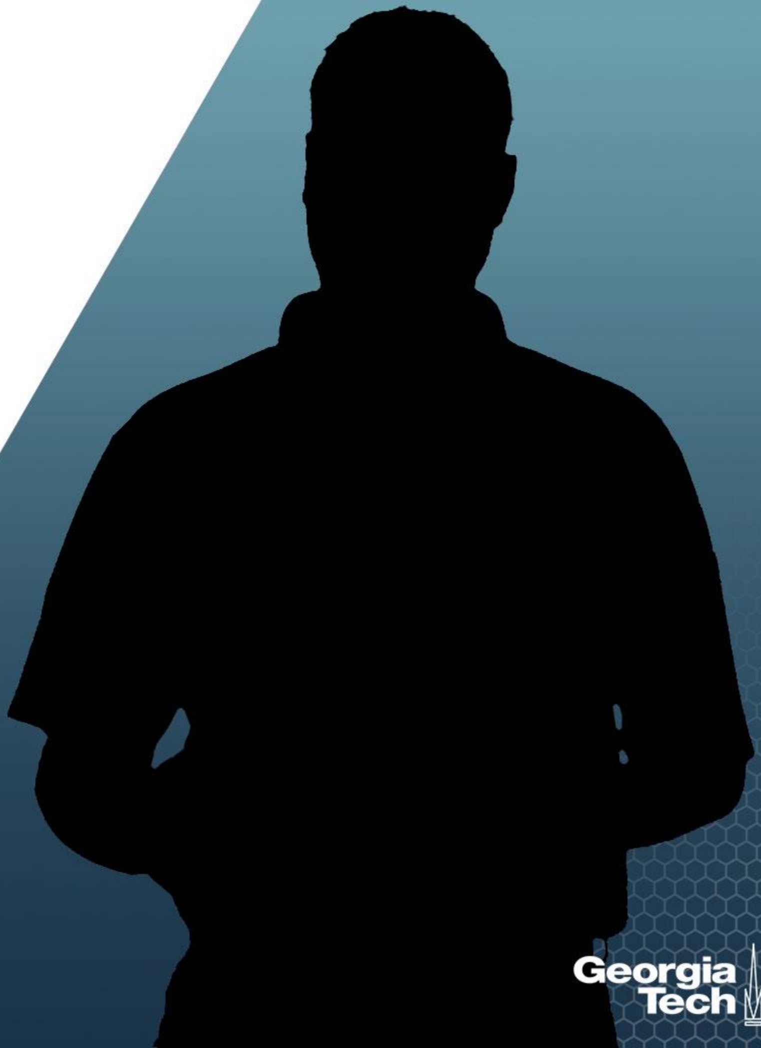
Multiple Linear Regression



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

Basics of Multiple Regression

About This Lesson



Georgia Tech

Multiple Linear Regression: Model

Data: $\{(x_{1,1},\dots,x_{1,p}),y_1\},\dots,\{(x_{n,1},\dots,x_{n,p}),y_n\}$
Model: $Y_i=\beta_0+\beta_1 x_{i,1}+\beta_2 x_{i,2}+\cdots+\beta_p x_{i,p}+\varepsilon_i,\ \ i=1,\dots,n$

Assumptions:

- *Linearity/Mean Zero Assumption*: $\mathrm{E}(\varepsilon_i)=0$
- *Constant Variance Assumption*: $\mathrm{Var}(\varepsilon_i)=\sigma^2$
- *Independence Assumption*: $\{\varepsilon_1,\dots,\varepsilon_n\}$ are independent random variables
- ε_i ~ Normally distributed *for confidence/prediction intervals, hypothesis testing*

Multiple Linear Regression: Model

Data: $\{(x_{1,1}, \dots, x_{1,p}), y_1\}, \dots, \{(x_{n,1}, \dots, x_{n,p}), y_n\}$
Model: $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i, \ \ i = 1, \dots, n$

Assumptions:

- *Linearity/Mean Zero Assumption*: $\mathrm{E}(\varepsilon_i) = 0$
- *Constant Variance Assumption*: $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption*: $\{\varepsilon_1, \dots, \varepsilon_n\}$ are independent random variables
- $\varepsilon_i \sim$ Normally distributed *for confidence/prediction intervals, hypothesis testing*

The model parameters are: $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1, \dots, \beta_p, \sigma^2}$

- **Unknown regardless how much data are observed**
- **Estimated given the model assumptions**
- **Estimated based on data**

Georgia Tech

Multiple Linear Regression: Model

Data: $\{(x_{1,1}, \ldots, x_{1,p}), y_1\}, \ldots, \{(x_{n,1}, \ldots, x_{n,p}), y_n\}$
Model: $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i,\ \ i = 1, \ldots, n$

Model in Matrix Form: $Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Response

$$\text{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

Design Matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

Coefficients

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Error

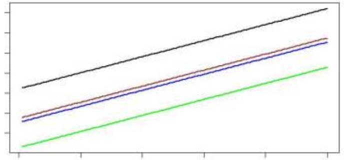
$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Georgia Tech

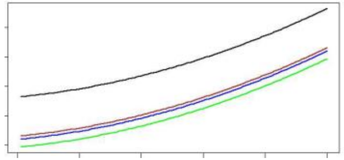
Model Flexibility: Main Effects & Interactions

For *k* = 2 predicting variables, four useful regressions:

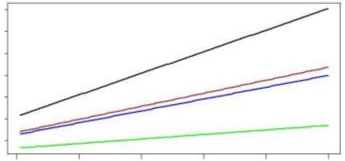
- ***1st* Order Model:**
  $$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$



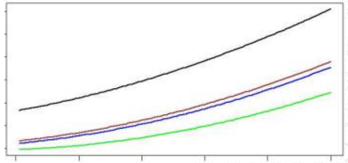
- ***2nd* Order Model:**
  $$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \varepsilon$$



- ***1st* Order Interaction Model:**
  $$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$



- ***2nd* Order Interaction Model:**
  $$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$$



Georgia Tech

Quantitative and Qualitative Variables

Simple Linear Regression: Linear regression with one quantitative predicting variable

ANOVA: Linear regression with one or more qualitative predicting variables

Multiple Linear Regression: Multiple quantitative and qualitative predicting variables

Quantitative and Qualitative Variables

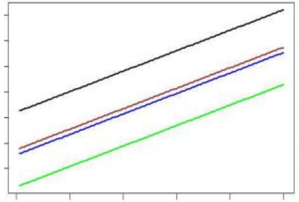
Multiple Linear Regression: Multiple quantitative/qualitative predicting variables

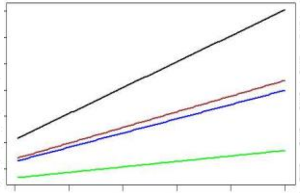
$\boldsymbol{x_1}$ quantitative
$\boldsymbol{x_2}$ qualitative with three levels: $\mathbf{D_1}$, $\mathbf{D_2}$, and $\mathbf{D_3}$ dummy variables

Model: $\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 d_1} + \boldsymbol{\beta_3 d_2} + \boldsymbol{\varepsilon}$ ➡ **Intercept varies**

If $d_1{=}0, d_2{=}0$: $\boldsymbol{\beta_0 + \beta_1 x_1}$
If $d_1{=}1, d_2{=}0$: $(\boldsymbol{\beta_0}{+}\boldsymbol{\beta_2}) + \boldsymbol{\beta_1 x_1}$
If $d_1{=}0, d_2{=}1$: $(\boldsymbol{\beta_0}{+}\boldsymbol{\beta_3}) + \boldsymbol{\beta_1 x_1}$

Parallel regression lines





If $\boldsymbol{x_1}$ $\boldsymbol{x_2}$ interaction: Nonparallel regression lines

Georgia Tech

Linear Regression: Example 1

Linear Regression: Example 1

Quantitative Predicting Variables:

$\boldsymbol{X_1}$ = The amount (in hundreds of dollars) spent on advertising in 1999
$\boldsymbol{X_2}$ = The total amount of bonuses paid in 1999
$\boldsymbol{X_3}$ = The market share in each territory
$\boldsymbol{X_4}$ = The largest competitor's sales

Qualitative Predicting Variable:

$\boldsymbol{X_5}$ = Indicates the region of the office (1 = south, 2 = west, 3 = midwest)

Linear Regression: Example 3

Bike sharing systems are of great interest due to their important role in traffic management.

Dataset: Historical data for years 2011-2012 for the bike sharing system in Washington D.C.

Georgia Tech

Linear Regression: Example 3

Qualitative predicting variables:

$\boldsymbol{X_1}$ = Day of the week
$\boldsymbol{X_2}$ = Month of the year
$\boldsymbol{X_3}$ = Hour of the day (ranging 0-23)
$\boldsymbol{X_4}$ = Year (2011, 2012)
$\boldsymbol{X_5}$ = Holiday Indicator
$\boldsymbol{X_6}$ = Weather condition (with four levels from good weather for level 1 to severe condition for level 4)

Quantitative predicting variables:

$\boldsymbol{X_7}$ = Normalized temperature
$\boldsymbol{X_8}$ = Normalized humidity
$\boldsymbol{X_9}$ = Wind speed

Linear Regression: Example 3

Qualitative predicting variables:

$\boldsymbol{X}_1$ = Day of the week
$\boldsymbol{X}_2$ = Month of the year
$\boldsymbol{X}_3$ = Hour of the day (ranging 0-23)
$\boldsymbol{X}_4$ = Year (2011, 2012)
$\boldsymbol{X}_5$ = Holiday Indicator
$\boldsymbol{X}_6$ = Weather condition (with four levels from good weather for level 1 to severe condition for level 4)

Quantitative predicting variables:

$\boldsymbol{X}_7$ = Normalized temperature
$\boldsymbol{X}_8$ = Normalized humidity
$\boldsymbol{X}_9$ = Wind speed

Year: A *quantitative* or a *qualitative* predicting variable?

- If observations are made over many years, then consider it to be *quantitative*
- If observations are made over only a few years, then consider it to be *qualitative*

Summary