# Regression Analysis
## Simple Linear Regression

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

Regression Concepts: Estimation

Georgia Tech

1

# About This Lesson

Georgia Tech

2

# Simple Linear Regression: Model

Our goal is to find the best line that describes a linear relationship; that is, find **$(\beta_0, \beta_1)$** where

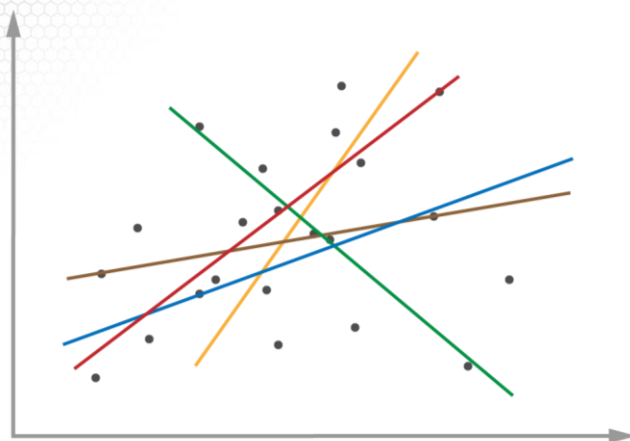$$Y = \beta_0 + \beta_1 x + \varepsilon$$

**Equivalently, estimating:**

1. $\beta_0$     *Intercept*
2. $\beta_1$     *Slope*

**$\varepsilon$** is the deviance of the data from the linear model

Georgia Tech

3

# Simple Linear Regression: Model



Our goal is to find the line that describes a linear relationship; that is, find **$(\beta_0, \beta_1)$** where

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

How to find the best line?

Georgia Tech

4

2

# Simple Linear Regression: Model

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

**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*

- (*Later we assume* $\varepsilon_i \sim$ *Normal*)

**Georgia Tech**

5

# Simple Linear Regression: Model

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

*The model parameters are:*
*$\beta_0$, $\beta_1$, $\sigma^2$*
- *Unknown regardless how much data are observed*
- *Estimated given the model assumptions*
- *Estimated based on data*

**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*

- (*Later we assume* $\varepsilon_i \sim$ *Normal*)
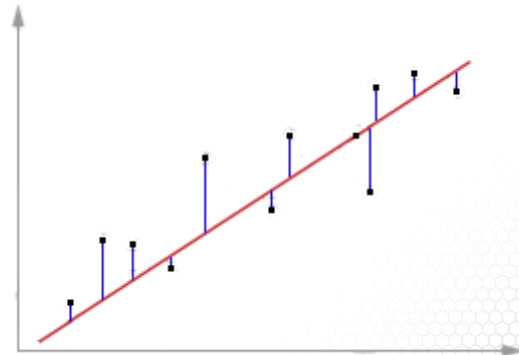
**Georgia Tech**

6

3

# Model Estimation: Approach

To estimate $(\beta_0, \beta_1)$ , we find values that minimize sum of squared errors:

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

⟹



**Georgia Tech**

7

# Model Estimation: Approach

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

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

⟹

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

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \equiv \frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

**Georgia Tech**

8

# Model Estimation: Approach

Begin with the minimization problem:

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

To solve, take the first order derivatives of the function to be minimized and equate to 0:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 = 0$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 = 0$$

➢ Result into a system of linear equation in $\beta_0$ and $\beta_1$
➢ Solve using linear algebra
➢ Solutions to the system are $\hat{\beta}_0$ and $\hat{\beta}_1$

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

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \equiv \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Georgia Tech

9

# Fitted Values and Residuals

Given the estimates of $b_0$ and $b_1$, we define:

- *Fitted values*: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- *Residuals*: $r_i = \hat{\varepsilon}_i = y_i - \hat{y}_i$

- *Mean squared error*: Estimator for $\sigma^2$

$$MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-2} = \frac{SSE}{n-2}$$

Georgia Tech

10

# Variance Sampling Distribution

$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n-2} \sim \chi_{n-2}^2$$

(chi-squared distribution with n-2 degrees of freedom)

Assuming $\hat{\epsilon}_i \sim \epsilon_i \sim N(0, \sigma^2)$

$\uparrow$

Estimating $\sigma^2$ $\longleftarrow$ Sample variance

**Georgia Tech**

11

# Variance Sampling Distribution (cont'd)

**What is the sample variance estimation?**

**Basic statistic concept:**

**Consider** $Z_1, \ldots, Z_n \sim N(\mu, \sigma^2)$ with $\mu$ and $\sigma^2$ unknown

The sample variance estimator: $S^2 = \frac{\sum (Z_i - \bar{Z})^2}{n-1} \rightarrow \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$

**Why n-1?**

We lose a degree of freedom because we replace $\mu \leftarrow \bar{Z}$

Now, going back to $\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n-2} \sim \chi_{n-2}^2$

This looks like the sample variance estimates except we use n-2 degrees of freedom.
**Why?**

**Georgia Tech**

12

# Variance Sampling Distribution (cont'd)

Recall that $\epsilon_i = \left( y_i - \left( \beta_0 + \beta_1 x_i \right) \right)$

$\uparrow$ Replaced by $\hat{\epsilon}_i = \left( y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right)$

We lose two degrees of freedom because
$$\beta_0 \leftarrow \hat{\beta}_0$$
$$\beta_1 \leftarrow \hat{\beta}_1$$

Thus, assuming that $\epsilon_i \sim N\left( 0, \sigma^2 \right)$

$\rightarrow$ $\hat{\sigma}^2 = MSE \sim \chi^2_{n-2}$

**(This is called the sampling distribution of $\hat{\sigma}^2$)**

Georgia Tech

13

# Model Parameter Interpretation

Commonly interested in the behavior of $\beta_1$

➢ A positive value of $\beta_1$ is consistent with a direct relationship between x and y; **e.g.,** higher values of height are associated with higher values of weight, or lower values of revenue are associated with lower values of profit;

➢ A negative value of $\beta_1$ is consistent with an inverse relationship between x and y; **e.g.,** higher price of a product is associated with lower demand, or a lower inflation rate is associated with a higher savings rate;

➢ A close-to-zero value of $\beta_1$ means that there is not a significant association between x and y.

Georgia Tech

14

# Model Estimate Interpretation

The Least Squares estimated coefficients have specific interpretations:

➢ $\hat{\beta}_1$ is the estimated expected change in the response variable associated with one unit of change in the predicting variable;

➢ $\hat{\beta}_0$ is the estimated expected value of the response variable when the predicting variable equals zero.

**Georgia Tech**

15

# Summary

**Georgia Tech**

16