# Data Analytics in Business

## Linear Regression

**Sridhar Narasimhan, Ph.D**
*Professor*
Scheller College of Business

**Steps in Regression Analysis**

Georgia Tech

# Lessons in this Module

A. Steps in Regression Analysis
B. A Real Estate example
C. Notation
D. $R^2$, Adjusted $R^2$
E. Simple Regression (One Predictor Variable) Using R
F. Multiple Regression
G. $R^2$, Adjusted $R^2$ from Multiple Regression
H. Common Problems and Fixes in Linear Regression

Georgia Tech

# Steps in Regression Analysis

1. Statement of the problem
2. Using regression for:
   - Diagnostic,
   - Predictive, or
   - Prescriptive analytics?
3. Selection of potentially relevant response and explanatory variables
4. Data collection
   - Internal data external data, purchased data, experiments, etc.

Adapted from Chatterjee, S., & Hadi, A. S. (2013). *Regression Analysis by Example* (5th ed.). Somerset: Wiley.

Georgia Tech

# Steps in Regression Analysis (cont'd)

5. Choice of fitting method:
   - Ordinary least squares (OLS),
   - Generalized least squares,
   - Maximum likelihood,
   - Etc.
6. Model fitting
7. Model validation (diagnostics)
8. Refine the model & iterate from step 3
9. Use of the model

Adapted from Chatterjee, S., & Hadi, A. S. (2013). *Regression Analysis by Example* (5th ed.). Somerset: Wiley.

Georgia Tech

# Business Examples

| Y - Dependent Variable | X - Independent Variable(s) |
|---|---|
| 1. Used car price | odometer reading, age of car, condition |
| 2. Sales | advertisement spending |
| 3. Time taken to repair a product | experience of technician in years |
| 4. Product added to shopping cart? | ratings, price |
| 5. Starting salary of new employee | work experience, years of education |
| 6. Sale price of house | square feet, # of bedrooms, location |
| 7. Will customer default? | credit balance, income, age |
| 8. Will customer churn? | length of contract, age of customer |

Georgia Tech

# Quiz (True/False)

- Could a variable, say price, be either a dependent or an independent variable?

  Answer: **TRUE**. Depends on the purpose of your model; see where **price** appears in examples #1 and #4 in the previous slide.

- A variable that takes binary values (pass/fail or true/false) cannot be a dependent variable.

  Answer: **FALSE**. We do use 0/1 dependent variables in logistic regression models; #7 in the previous slide is one example.

Georgia Tech

# Data Analytics in Business

## Linear Regression

**Sridhar Narasimhan, Ph.D**
*Professor*
Scheller College of Business


**Real Estate Pricing Example**

# Linear Regression: A Sample Problem

- Assume that you need to sell your house
- You want to predict the listing price based on how other houses are listed in the market
- How would you approach this task?
- A typical approach is to ask realtors:
  - Realtors often will use “comparables” (i.e., recent sales of houses in your neighborhood) and somehow come up with a suggested sale price
- However, you want to be more analytical in your approach
  - You have access to recent actual home sales in your city
  - You’d like to know what are the impacts of factors such as lotsize, # of bedrooms, # of bathrooms, etc., on the price
  - Could you use linear regression to help you get a “better” estimate of the listing price?

Georgia Tech

# Use Housing Dataframe in Ecdat Package in R

- This data set is a sample of the real estate transactions in one city
- It is a cross-section of 546 home prices (from 1987) in the city of Windsor in Canada
- Alternatively, you could collect house prices from websites or scrape them from the web

Georgia Tech

# str(Housing)

- 'data.frame': 546 obs. of 12 variables:
- $ price: num 42000 38500 49500 60500 61000 66000 66000 69000 83800 88500 ...
- $ lotsize: num 5850 4000 3060 6650 6360 4160 3880 4160 4800 5500 ...
- $ bedrooms: num 3 2 3 3 2 3 3 3 3 3 ...
- $ bathrms: num 1 1 1 1 1 1 2 1 1 2 ...
- $ stories: num 2 1 1 2 1 1 2 3 1 4 ...
- $ driveway: Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
- $ recroom: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 2 2 ...
- $ fullbase: Factor w/ 2 levels "no","yes": 2 1 1 1 1 2 2 1 2 1 ...
- $ gashw: Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
- $ airco: Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 1 1 1 2 ...
- $ garagepl: num 1 0 0 0 0 0 2 0 0 1 ...
- $ prefarea: Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...

Georgia Tech

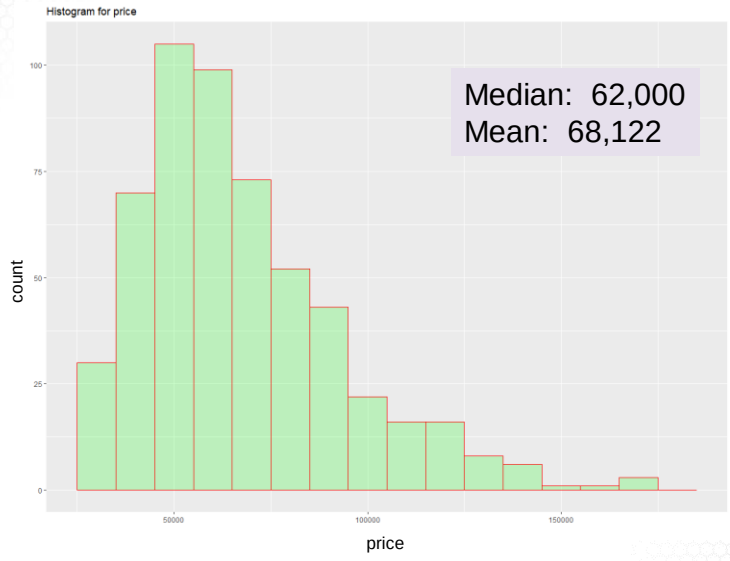
# The First 10 Records in Housing

*Housing* Dataset in the *Ecdat* package in R

| price | lotsize | bedrooms | bathrms | stories | driveway | recroom | fullbase | gashw | airco | garagepl | prefarea |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 42000 | 5850 | 3 | 1 | 2 | yes | no | yes | no | no | 1 | no |
| 38500 | 4000 | 2 | 1 | 1 | yes | no | no | no | no | 0 | no |
| 49500 | 3060 | 3 | 1 | 1 | yes | no | no | no | no | 0 | no |
| 60500 | 6650 | 3 | 1 | 2 | yes | yes | no | no | no | 0 | no |
| 61000 | 6360 | 2 | 1 | 1 | yes | no | no | no | no | 0 | no |
| 66000 | 4160 | 3 | 1 | 1 | yes | yes | yes | no | yes | 0 | no |
| 66000 | 3880 | 3 | 2 | 2 | yes | no | yes | no | no | 2 | no |
| 69000 | 4160 | 3 | 1 | 3 | yes | no | no | no | no | 0 | no |
| 83800 | 4800 | 3 | 1 | 1 | yes | yes | yes | no | no | 0 | no |
| 88500 | 5500 | 3 | 2 | 4 | yes | yes | no | no | yes | 1 | no |

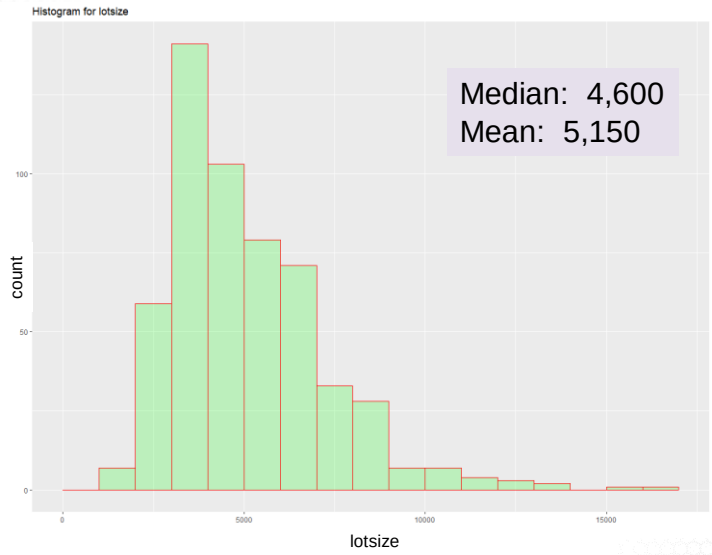
Georgia Tech

# Histogram of House Prices

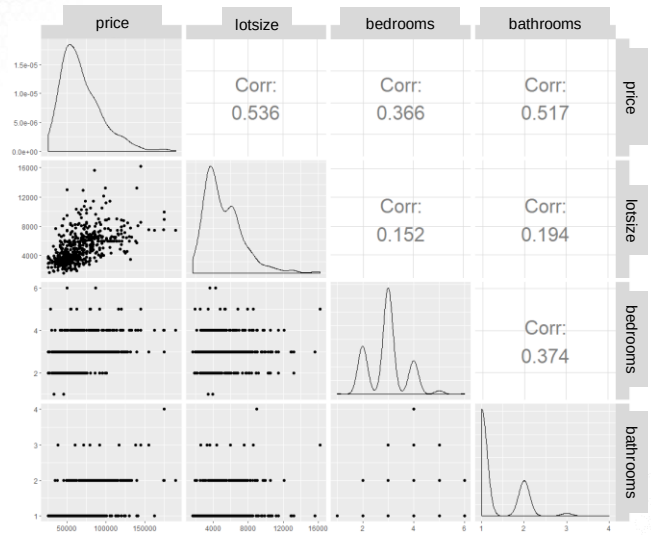



Georgia Tech

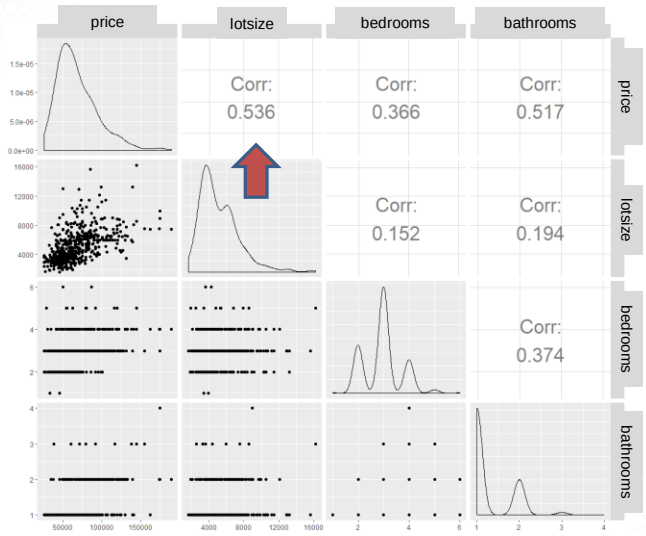
# Histogram of lotsize




# Correlation Matrix



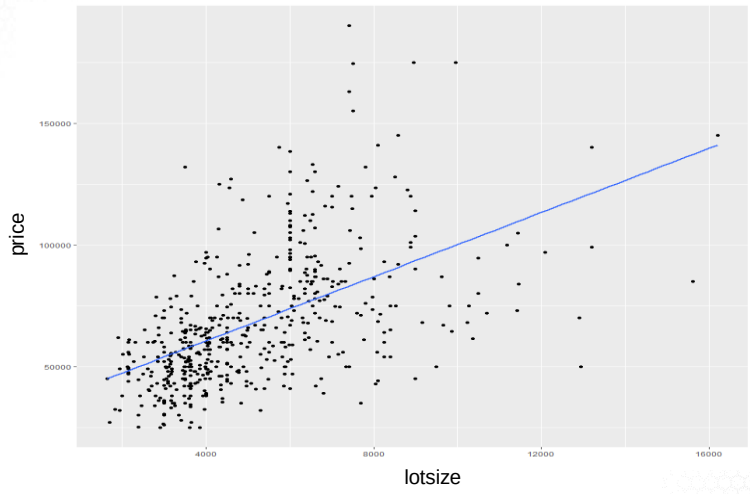
# Correlation Matrix




Georgia Tech

# Scatter Plot

Scatter Plot of price (*y*) against lotsize (*x*), including the linear regression line




Georgia Tech

# Quiz (True/False)

- The mean of a variable that has a right-skewed distribution is smaller than the median.

  Answer: **FALSE**.

- The correlation coefficient can capture the strength of both linear and non-linear relationships.

  Answer: **FALSE**.

Georgia Tech

# Data Analytics in Business

## Linear Regression

**Sridhar Narasimhan, Ph.D**

*Professor*

Scheller College of Business

**Notation**

Georgia Tech

# Linear Regression: Notation

| Notation | Meaning |
|---|---|
| $i = 1,2,\ldots,n$ | *i* refers to the *i*th observation or record in a data set of records (typically a sample of the population) |
| $(x_{11}, x_{21}, \ldots, x_{p1})$, $(x_{12}, x_{22}, \ldots, x_{p2})$, …, $(x_{1n}, x_{2n}, \ldots, x_{pn})$ | *n observations of the p explanatory variables* |
| $y_1, y_2, \ldots, y_n$ | *n* observations of the dependent variable |
| $\bar{y}$ | Mean value of the dependent (*y*) variable |
| $\bar{x}_k$ | Mean value of the $x_k$th explanatory (independent) variable |

Georgia Tech

# Linear Regression: Notation (cont'd)

| Notation | Meaning |
|---|---|
| $\beta_0, \beta_1, \ldots, \beta_p$ | Parameters of the regression line for the entire population |
| $b_0, b_1 \ldots, b_p$ | Estimates of the $\beta$ parameters obtained by fitting the regression to the sample data |
| $\varepsilon_i$ | Error term for the *i*th observation in the population |
| $e_i$ | Error term for the *i*th observation in the sample |
| $\hat{y}_i$ | Estimated value of *y* for the *i*th observation in a sample. This is obtained by evaluating the regression function at $x_i$ |

Georgia Tech

# Simple Linear Regression

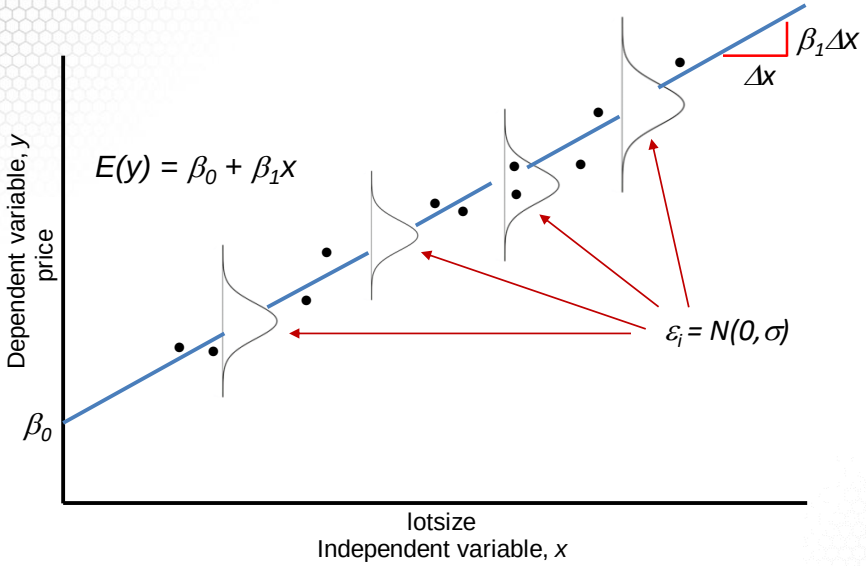
- We observe the data in the Housing dataset (which is a sample)
- We want to build a model for the **population**:

  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (which is the valid relation)
- $\varepsilon_i$ are independent and identically distributed (i.i.d.) random variables, which are normally distributed with mean 0 and standard deviation $\sigma$
- However, we do not know $\beta_0$, $\beta_1$, or $\sigma$; so we need to estimate them based on the **sample** in the Housing dataset
- Using this sample, we are going to build a model

  $Y_i = b_0 + b_1 X_i + e_i$

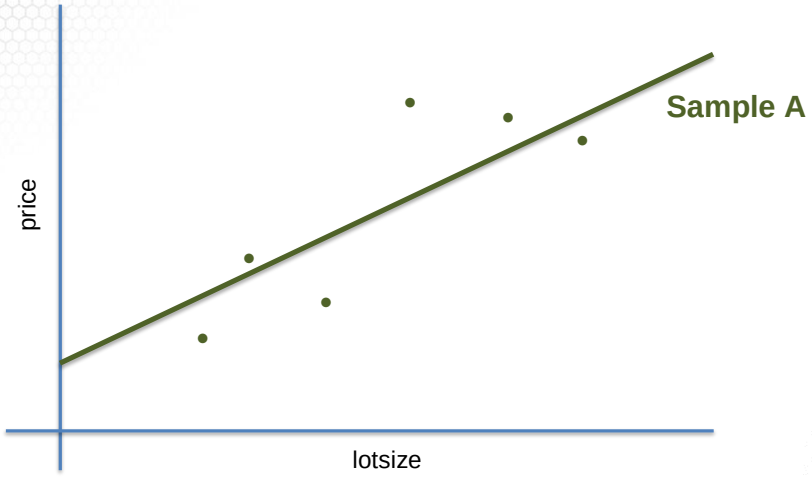
Georgia Tech

# Population model, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



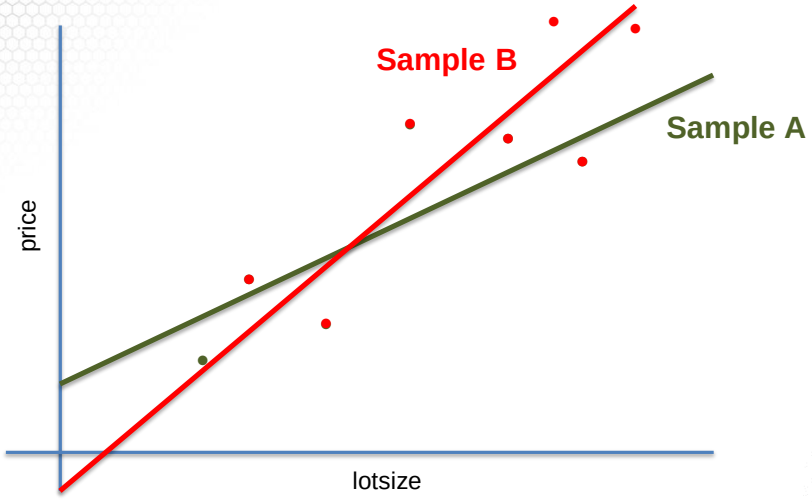

Georgia Tech

# Estimates of Slope and Intercept Depend on the Sample Being Used



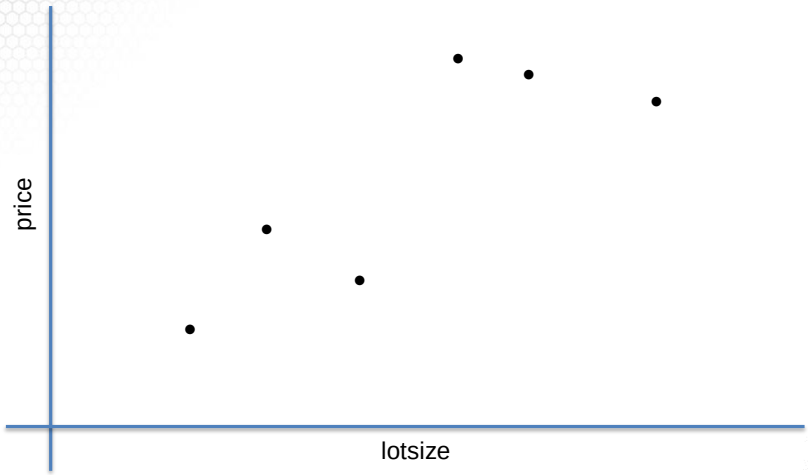

Georgia Tech

# Estimates of Slope and Intercept Depend on the Sample Being Used



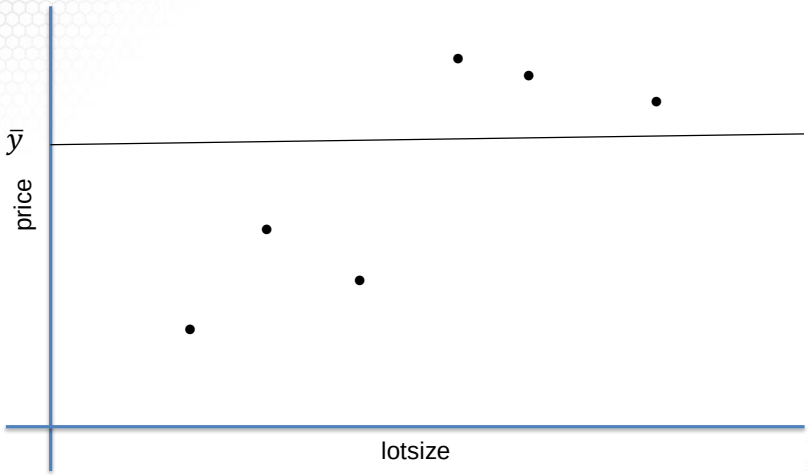

Georgia Tech

# Use Ordinary Least Squares (OLS) to Fit the Line $\hat{y} = b_0 + b_1 x$




Georgia Tech

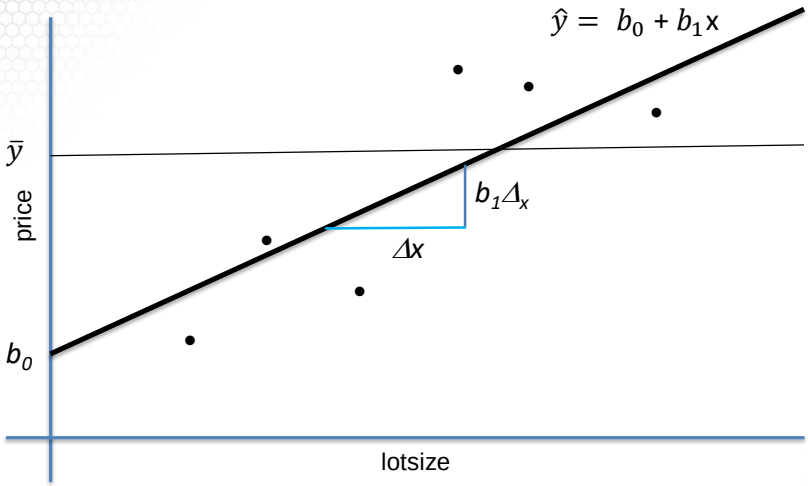
# Use Ordinary Least Squares (OLS) to Fit the Line $\hat{y} = b_0 + b_1 x$



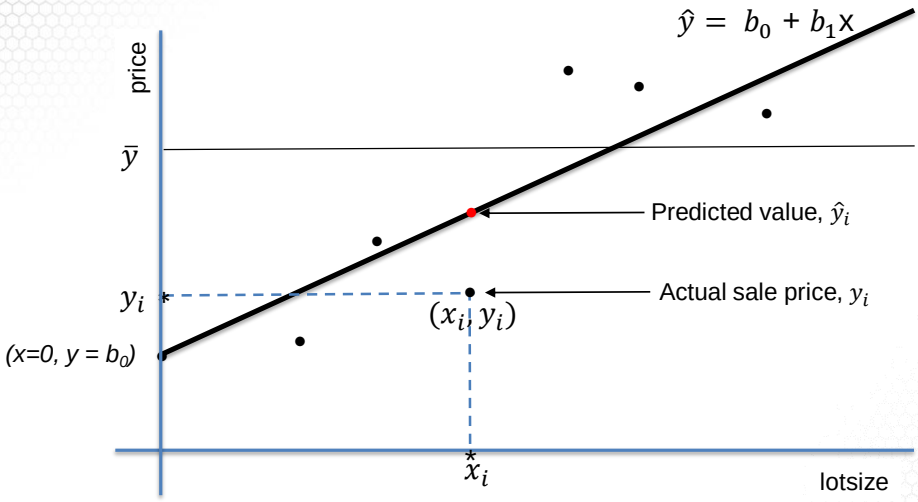

Georgia Tech

# Use Ordinary Least Squares (OLS) to Fit the Line $\hat{y} = b_0 + b_1 x$

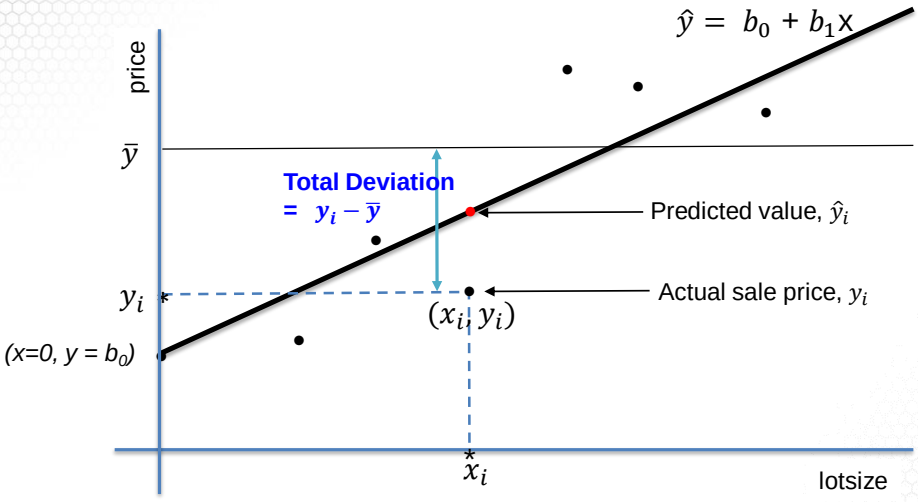

# Use Ordinary Least Squares (OLS) to Fit the Line $\hat{y} = b_0 + b_1 x$



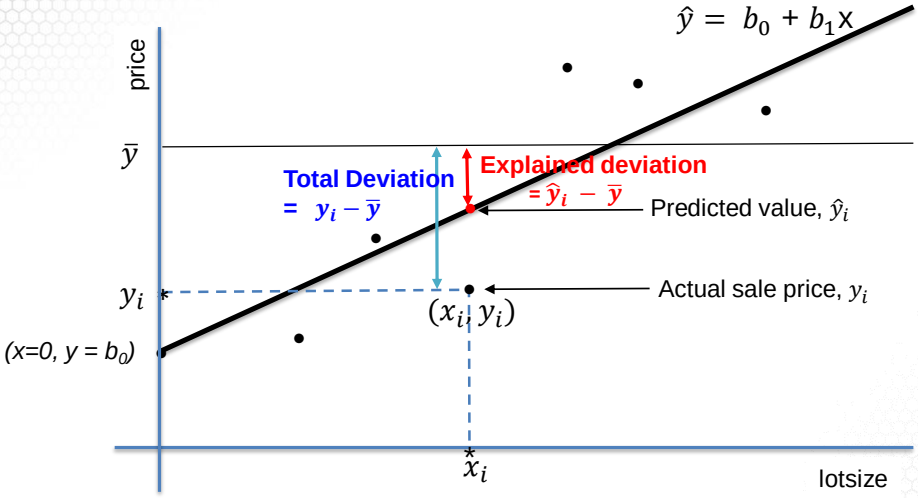
# Total Deviation = Explained Deviation + Unexplained Deviation



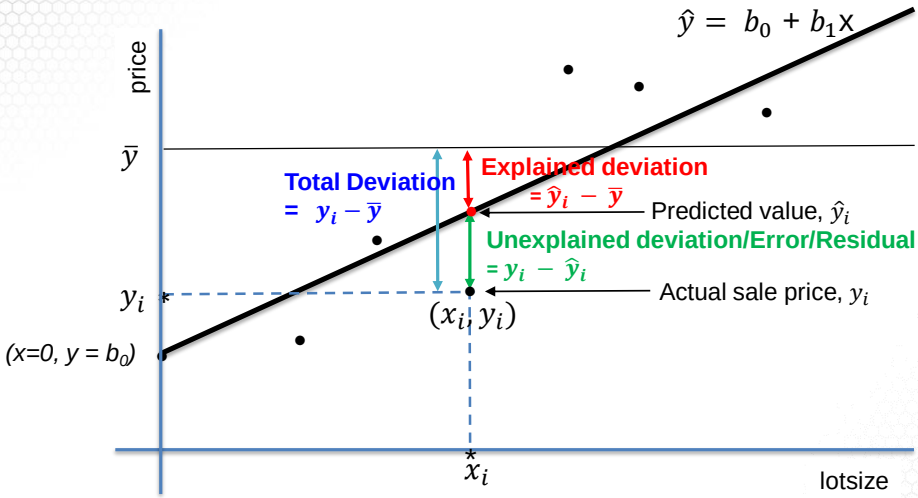

# Total Deviation = Explained Deviation + Unexplained Deviation



# Total Deviation = Explained Deviation + Unexplained Deviation




# Total Deviation = Explained Deviation + Unexplained Deviation



# Quiz (True/False)

- The total deviation at observation $(x_i , y_i)$ is $\boldsymbol{y_i - \bar{y}}$.
  Answer: **TRUE**

- In OLS, the estimates of slope and intercept do not depend on the sample being used.
  Answer: **FALSE**

Georgia Tech

# Data Analytics in Business

## Linear Regression

**Sridhar Narasimhan, Ph.D**
*Professor*
Scheller College of Business

**$R^2$, Adjusted $R^2$**

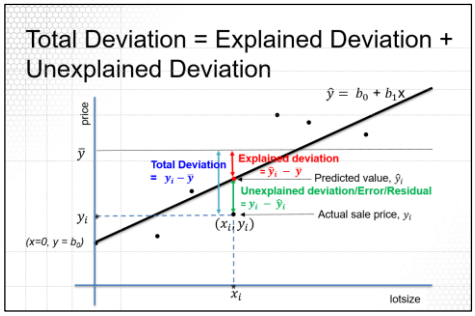
Georgia Tech

# Regression (Ordinary Least Squares): Sum of Squared Errors (SSE)

Regression (OLS) determines the line that minimizes the Sum of Squared Errors

- i.e., $b_0$ and $b_1$ are determined such that they minimize:

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (b_0 + b_1 x_i))^2$$




Georgia Tech

# Summing the Deviations

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\bar{y} - \hat{y}_i)^2$$

| | | | | |
|---|---|---|---|---|
| SST | = | SSE | + | SSR |
| Total Sum of Squares | = | Sum of Squared Errors | + | Sum of Squares Regression |

Georgia Tech

# Regression Output $R^2$ and Adjusted $R^2$

Coefficient of determination ($R^2$)

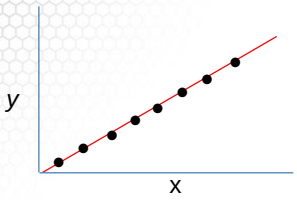
- A measure of the overall strength of the relationship between the dependent variable (Y) and independent variables (X)
- $R^2 = 1 - (SSE/SST) = SSR/SST$
  $= \text{Explained deviation (SSR)}/\text{Total Deviation (SST)}$
- $R^2$ → how much of the variation in Y (from the mean) has been explained

Adjusted $R^2$

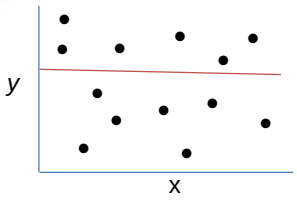
- Adding a penalty for the number of independent variables (p)
- Adjusted $R^2 = 1 - \{SSE/(n - p - 1)\}/\{SST/(n - 1)\}$

Georgia Tech

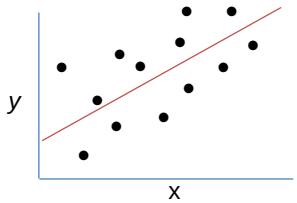
# $R^2$




$R^2 = 1$,
*X* accounts for all *Y* variation




$R^2 = 0$,
*X* accounts for none of the *Y* variation




$R^2 = 0.75$,
*X* accounts for most of the *Y* variation

Georgia Tech

# Quiz (True/False)

- $R^2 = 0$, implies that *X* values account for all of the variation in the *Y* values
  Answer: **FALSE**. $R^2 = 0$ implies that *X* values account for none of the variation in the *Y* values

- $R^2$ can take any value from – infinity to + infinity
  Answer: **FALSE**. It can take on values between 0 and 1

Georgia Tech

# Data Analytics in Business

## Linear Regression

**Sridhar Narasimhan, Ph.D**
*Professor*
Scheller College of Business

**Simple Regression (One Predictor Variable) Using R**

Georgia Tech

# Regression Output: Simple Linear Regression

lm(formula = price ~ lotsize, data = Housing)

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -69551 | -14626 | -2858 | 9752. | 106901 |

**Unexplained deviation/Error/Residual**
**= $y_i - \hat{y}_i$**

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.414e+04 | 2.491e+03 | 13.7 | <2e-16 *** |
| lotsize | 6.599e+00 | 4.458e-01 | 14.8 | <2e-16 *** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22570 on 544 degrees of freedom
Multiple R-squared: 0.2871, Adjusted R-squared: 0.2858
F-statistic: 219.1 on 1 and 544 DF, p-value: < 2.2e-16

Georgia Tech

# Regression Output: Coefficients

**$b_0$ and $b_1$ are estimates of the true parameters $\beta_0$ and $\beta_1$**
**$H_0$: the parameter is zero, $H_1$: The parameter is not zero**

lm(formula = price ~ lotsize, data = Housing)

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -69551 | -14626 | -2858 | 9752 | 106901 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.414e+04 | 2.491e+03 | 13.7 | <2e-16 *** |
| lotsize | 6.599e+00 | 4.458e-01 | 14.8 | <2e-16 *** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22570 on 544 degrees of freedom
Multiple R-squared: 0.2871, Adjusted R-squared: 0.2858
F-statistic: 219.1 on 1 and 544 DF, p-value: < 2.2e-16

Georgia Tech

# Regression Output: t-values for Coefficients

**p value: the probability of finding a t value of this size if the null hypothesis is true**

**$H_0$: the parameter is zero**

```
lm(formula = price ~ lotsize, data = Housing)

Residuals:
   Min     1Q Median     3Q    Max
-69551 -14626  -2858   9752 106901

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.414e+04  2.491e+03    13.7  <2e-16 ***
lotsize     6.599e+00  4.458e-01    14.8  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22570 on 544 degrees of freedom
Multiple R-squared:  0.2871, Adjusted R-squared:  0.2858
F-statistic: 219.1 on 1 and 544 DF,  p-value: < 2.2e-16
```

Georgia Tech

# Interpreting Coefficients

| | Estimate |
|---|---|
| (Intercept) | 3.414e+04 *** |
| lotsize | 6.599e+00 *** |

$b_0$ = 34,140

Intercept of the regression line with the y-axis (when lotsize is zero). Not useful

$b_1$ = 6.599

An increase of 1,000 square feet is associated with an increase of the sale price of a house by $6,599, keeping all else constant (*ceteris paribus*)

Georgia Tech

# Regression Output: Sum of Squares

**Analysis of Variance Table**

| | Df | Sum Sq |
|---|---|---|
| lotsize | 1 | 1.1156e+11 |
| Residuals | 544 | 2.7704e+11 |

SSR = 1.1156e+11

SSE = 2.7704e+11

SST = SSR + SSE = 3.886e+11

Georgia Tech

# Regression Output: $R^2$

```
lm(formula = price ~ lotsize, data = Housing)

Residuals:
   Min     1Q Median     3Q    Max
-69551 -14626  -2858   9752 106901

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) 3.414e+04  2.491e+03     13.7    <2e-16 ***
lotsize     6.599e+00  4.458e-01     14.8    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22570 on 544 degrees of freedom

Multiple R-squared:  0.2871, Adjusted R-squared:  0.2858

F-statistic: 219.1 on 1 and 544 DF,  p-value: < 2.2e-16
```

Georgia Tech

# Regression Output $R^2$ and Adjusted $R^2$

SSR = 1.1156e+11
SSE = 2.7704e+11
SST = SSR + SSE = 3.886e+11

Multiple R-squared: 0.2871, Adjusted R-squared: 0.2858
$R^2 = 1 - (SSE/SST) = SSR/SST = 1.1156e+11/3.886e+11$
$= 0.2871$
Note: $\sqrt{R^2} = \sqrt{0.2871} = 0.536$ (which is the correlation coefficient between price and lotsize)

Adjusted $R^2 = 1 - \{SSE/(n - p - 1)\}/\{SST/(n - 1)\}$
$= 1 - \{(2.7704e+11/(546 - 1 - 1)\} / \{3.886e+11/(546 - 1)\}$
$= 0.2858$

Georgia Tech

# F-test that the model is significant ($H_0$: $b_1 = 0$)

SSR = 1.1156e+11
SSE = 2.7704e+11
SST = SSR + SSE = 3.886e+11
If p is the number of independent variables, The F statistic
$= (SSR/p)/ (SSE/n - p - 1) = (R^2/p) / (1 - R^2)/(n - p - 1)$
The value of Prob(F) is the probability that $H_0$ is true (i.e., $b_1 = 0$).
For this model, p = 1,
$F = (0.2871/1) / (1 - 0.2871)/(546 - 1 - 1) = 219.1$
with (1,544) degrees of freedom.
F statistic: 219.1 with(1, 544) DF, p-value: < 2.2e-16. Hence $H_0$ is rejected.

Georgia Tech

# Data Analytics in Business
## Linear Regression

**Sridhar Narasimhan, Ph.D**
*Professor*
Scheller College of Business


**Multiple Regression**

# Multiple Linear Regression, with *p* Explanatory Variables

- **Regression coefficients**:
  $b_0, b_1, \ldots, b_p$ are estimates of $\beta_0, \beta_1, \ldots, \beta_p$
- **Prediction** for Y at $x_i$
  $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi}$
- **Residual**:
  $e_i = y_i - \hat{y}_i$

Goal: choose $b_0, b_1, \ldots, b_p$ to minimize the sum of squared errors

$$SSE = \sum_i (y_i - \hat{y}_i)^2 =$$

$$\sum_i (y_i - (b_0 + b_1 x_{1i} + \ldots b_p x_{pi}))^2$$

Georgia Tech

# Using R to Estimate a Linear Model

Using the *Housing* Dataset in the *Ecdat* package in R

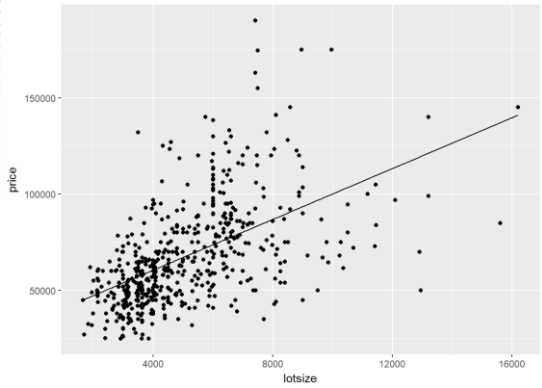
Georgia Tech

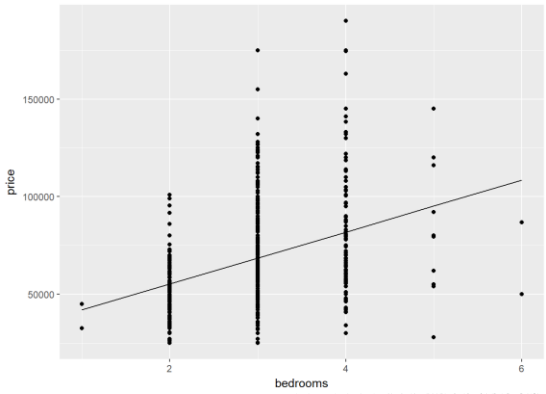
# Adding Bedrooms to the Analysis

| price | lotsize | bedrooms |
|---|---|---|
| 42,000 | 5,850 | 3 |
| 38,500 | 4,000 | 2 |
| 49,500 | 3,060 | 3 |
| 60,500 | 6,650 | 3 |
| 61,000 | 6,360 | 2 |
| 66,000 | 4,160 | 3 |
| 66,000 | 3,880 | 3 |
| 69,000 | 4,160 | 3 |
| 83,800 | 4,800 | 3 |
| 88,500 | 5,500 | 3 |
| 90,000 | 7,200 | 3 |
| 30,500 | 3,000 | 2 |
| 27,000 | 1,700 | 3 |
| 36,000 | 2,880 | 3 |
| 37,000 | 3,600 | 2 |

Georgia Tech

# Visualize (Plots)




Do the slopes make sense?

# Regression Output

lm(formula = price ~ lotsize + bedrooms, data = Housing)

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 5.613e+03 | 4.103e+03 | 1.368 | 0.172 |
| lotsize | 6.053e+00 | 4.243e-01 | 14.265 | < 2e-16 *** |
| bedrooms | 1.057e+04 | 1.248e+03 | 8.470 | 2.31e-16 *** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21230 on 543 degrees of freedom
Multiple R-squared: 0.3703, Adjusted R-squared: 0.3679
F-statistic: 159.6 on 2 and 543 DF, p-value: < 2.2e-16

# Regression Output: Coefficients

***$b_0$, $b_1$, …, $b_p$ are estimates of the true parameters $\beta_0$, $\beta_1$, …, $\beta_p$***

**$H_0$: the parameter is zero, $H_1$: The parameter is not zero**

```
lm(formula = price ~ lotsize + bedrooms, data = Housing)

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  5.613e+03  4.103e+03    1.368     0.172
lotsize      6.053e+00  4.243e-01   14.265   < 2e-16 ***
bedrooms     1.057e+04  1.248e+03    8.470   2.31e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21230 on 543 degrees of freedom
Multiple R-squared:  0.3703,    Adjusted R-squared:  0.3679
F-statistic: 159.6 on 2 and 543 DF,  p-value: < 2.2e-16
```


# Regression Output: Standard Error of the Coefficients

**Similar to Standard Deviation**

```
lm(formula = price ~ lotsize + bedrooms, data = Housing)

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  5.613e+03  4.103e+03    1.368     0.172
lotsize      6.053e+00  4.243e-01   14.265   < 2e-16 ***
bedrooms     1.057e+04  1.248e+03    8.470   2.31e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21230 on 543 degrees of freedom
Multiple R-squared:  0.3703,  Adjusted R-squared:  0.3679
F-statistic: 159.6 on 2 and 543 DF,  p-value: < 2.2e-16
```

# Regression Output: t-values for Coefficients

lm(formula = price ~ lotsize + bedrooms, data = Housing)

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 5.613e+03 | 4.103e+03 | 1.368 | 0.172 |
| lotsize | 6.053e+00 | 4.243e-01 | 14.265 | < 2e-16 *** |
| bedrooms | 1.057e+04 | 1.248e+03 | 8.470 | 2.31e-16 *** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21230 on 543 degrees of freedom
Multiple R-squared: 0.3703, Adjusted R-squared: 0.3679
F-statistic: 159.6 on 2 and 543 DF, p-value: < 2.2e-16


# Interpreting Coefficients

| | Estimate |
|---|---|
| (Intercept) | 5.613e+03 |
| lotsize | 6.053e+00 |
| Bedrooms | 1.057e+04 |

$b_0$ = 5613

Intercept of the regression line with the y-axis (when all x's are zero). Not useful

$b_1$ = 6.053

An increase of 1,000 square feet is associated with an increase of the sale price of a house by $6,053, keeping all else constant

$b_2$ = 10570

An additional bedroom is associated with an increase of the sale price of a house by $10,570, keeping all else constant

# Data Analytics in Business

## Linear Regression

**Sridhar Narasimhan, Ph.D**
*Professor*
Scheller College of Business

**$R^2$, Adjusted $R^2$ from Multiple Regression**


# Regression Output: Sum of Squares

**Analysis of Variance Table**

| | Df | Sum Sq |
|---|---|---|
| lotsize | 1 | 1.1156e+11 |
| bedrooms | 1 | 3.2329e+10 |
| Residuals | 543 | 2.4472e+11 |

SSR = (1.1156 + 0.32329) e+11

SSE = 2.4472e+11

SST = SSR + SSE = 3.88609e+11

Georgia Tech

# Regression Output: $R^2$

lm(formula = price ~ lotsize + bedrooms, data = Housing)

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 5.613e+03 | 4.103e+03 | 1.368 | 0.172 |
| lotsize | 6.053e+00 | 4.243e-01 | 14.265 | < 2e-16 *** |
| bedrooms | 1.057e+04 | 1.248e+03 | 8.470 | 2.31e-16 *** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21230 on 543 degrees of freedom

Multiple R-squared: 0.3703, Adjusted R-squared: 0.3679

F-statistic: 159.6 on 2 and 543 DF, p-value: < 2.2e-16

# Regression Output $R^2$ and Adjusted $R^2$

SSR = (1.1156 + 0.32329) e+11

SSE = 2.4472e+11

SST = SSR + SSE = 3.88609e+11

Multiple R-squared: 0.3703, Adjusted R-squared: 0.3679

$R^2$ = 1 – SSE/SST = SSR/SST = 1.43889e+11/3.88609e+11

= 0.3703

Adjusted $R^2$ = 1 – {SSE/(n – p – 1)}/{SST/(n – 1)}

= 1 – ((2.4472e+11/(546 – 2 – 1)} / {3.88609e+11/(546 – 1)}

= 0.3679

# F-test of the Overall Significance of the Model ($H_0$: $b_1 = b_2 = 0$)

SSR = (1.1156 + 0.32329) e+11
SSE = 2.4472e+11
SST = SSR + SSE = 3.88609e+11
The F statistic
= (SSR/p) / (SSE/(n – p – 1)) = ($R^2$/p) / ((1 – $R^2$)/(n – p – 1 ))

The value of Prob(F) is the probability that $H_0$ is true.

For this example, F = (0.3703/2)/(1 – 0.3703)/(546 – 2 – 1 ) = 159.6
F-statistic: 159.6 on 2 and 543 DF, p-value: < 2.2e-16. Hence $H_0$ is rejected.

Georgia Tech

# Simple vs. Multiple Regression

- For the Simple Regression we got:
  Multiple R-squared: 0.2871, Adjusted R-squared: 0.2858
- For the Multiple Regression we got:
  Multiple R-squared: 0.3703, Adjusted R-squared: 0.3679
- As you add variables, R-square will not decrease

Georgia Tech

# Comparing the Two Models

n: number of observations

p: number of variables (do not count the intercept)

Bigger model 2 which has (more) $p_2$ variables

Smaller model 1 which has (fewer) $p_1$ variables

We want to determine whether model 2 gives a *significantly* better fit to the data. Then use the F statistic shown below

- $F(p_2\text{-}p_1, n - p_2 - 1)$
- F test statistic is calculated as

$$F = \frac{(R_2^2 - R_1^2)/(p_2 - p_1)}{(1 - R_2^2)/(n - p_2 - 1)}$$

Georgia Tech

# Quiz (True/False)

- In general, adding more variables decreases the overall R-Square value of the multiple regression.

  Answer: **FALSE**.

- In the regression output shown below, a p-value of < 2e-16 *** means that there is not much evidence for the coefficient of lotsize to be different from zero.

  Answer: **FALSE**.

lm(formula = price ~ lotsize + bedrooms, data = Housing)

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 5.613e+03 | 4.103e+03 | 1.368 | 0.172 |
| lotsize | 6.053e+00 | 4.243e-01 | 14.265 | < 2e-16 *** |
| bedrooms | 1.057e+04 | 1.248e+03 | 8.470 | 2.31e-16 *** |

Georgia Tech

# Data Analytics in Business

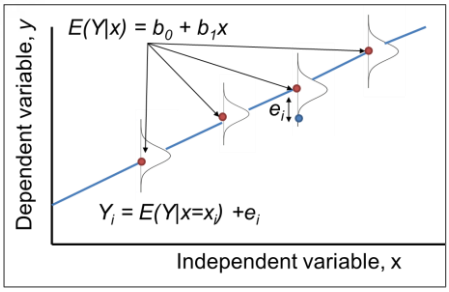
## Linear Regression

**Sridhar Narasimhan, Ph.D**
*Professor*
Scheller College of Business

**Common Problems and Fixes in Fitting Linear Regression, Part 1**

Georgia Tech

# Assumptions of Linear Regression




- **Linearity assumption**: E(y) = $b_0$ + $b_1$x, i.e., the expected value of Y at each value of X approximates to a straight line
- **Assumption about errors**: The error terms $e_i$ are independently and identically distributed (iid) normal random variables, each with mean zero and constant variance $\sigma^2$ (homoscedasticity)
- **Assumptions about predictors**: In multiple regression, the predictor variables are assumed to be linearly independent of one another

Georgia Tech

# Most Common Problems in Fitting Linear Regression

1. Non-linearity of the response-predictor relationships
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Collinearity

James, Gareth, et al. *An introduction to statistical learning: with applications in R* (Section 3.3.3). Springer, 2017.

Georgia Tech

# 1. Is the Relationship Nonlinear?

- Check the scatter plots of **Y** vs. each **X** variable. Linear?
- Another plot to use is the residuals plot vs. fitted values plot (especially useful in multiple regression)
  - We want to see no patterns

Georgia Tech

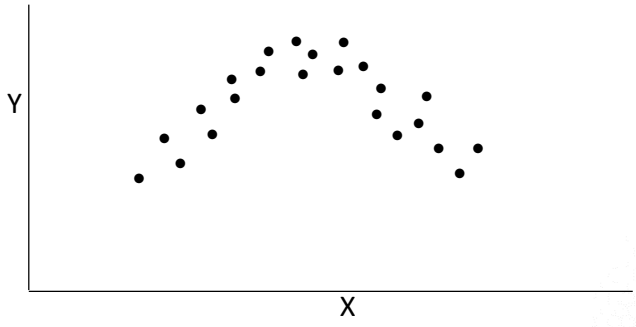
# 1. Is the Relationship Nonlinear?

- If there are concerns, then:
  - Can you model non-linear relationship with higher order terms (e.g., square)?
  - Use variance reducing transformation (such as log) that will give a better linear fit
  - Are there outliers or certain sections of the observations that seem to drive the non-linearity?
  - Is there any important variable that you left out from your model (e.g., age or gender)?
  - Or, maybe, was there systematic bias when collecting data, hence redesign data collection
- Checking residuals helps discover useful insights about your model and data

Georgia Tech

# Scatterplot of the Response Variable Versus the Explanatory Variable

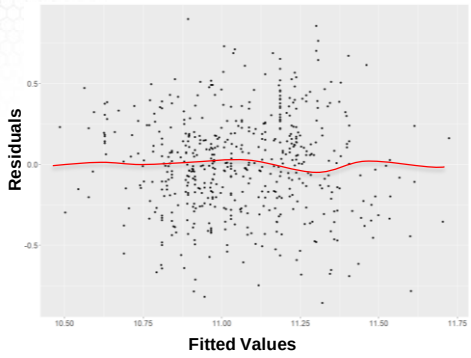
- This is useful to do before fitting a model
- You can plot Y vs. X to identify any patterns. For example, the scatter plot below suggests using $X^2$ rather than X as a predictor



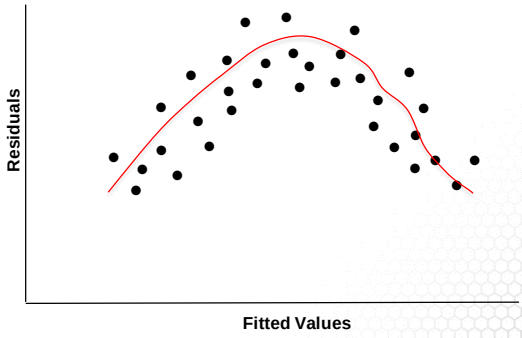

Georgia Tech

# Residuals vs. Fitted Values Plot

Case 1




Case 2




This plot should be examined after a model is fitted

Georgia Tech

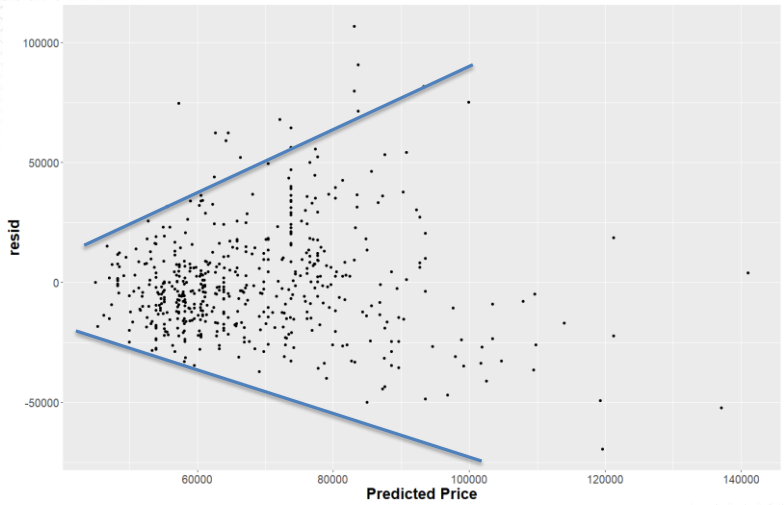
# 2. Correlation of Error Terms

- An important assumption is that error terms $e_1$, $e_2$, …,$e_n$ are uncorrelated. If they aren't, then we have autocorrelation
- So knowing the value of $e_i$ should not have any influence on the magnitude or size of $e_{i+1}$
- This property is used to estimate the standard errors of the parameters of the model
- If there is correlation in the error terms:
  - The estimated standard errors will underestimate the true standard errors
  - Confidence and prediction intervals will be narrower than they should be and p values will be lower than they should be
  - We may have sense of confidence in the model that is not warranted
- The Durbin-Watson test is used to detect autocorrelations in a linear model

Georgia Tech

# 3. Heteroskedasticity (Non-constant Error Variance)

- The assumption is that the spread of the responses around the straight line is the same at all levels of the explanatory variable (i.e., we have constant variance or homoscedasticity)
- You may have non-constant error present (e.g., the errors increase in size with the fitted values). You can detect this with the residuals vs. fitted values plot
- If non-constant error is present, then Hypotheses tests and Confidence Intervals can be misleading
- If there is Heteroscedasticity, then transformation of the Y variable may be called for
    - Example: ln(Y), or 1/Y, etc.



# Quiz (True/False)

- If the scatterplot of Y vs. X shows a nonlinear pattern, then we should not change our linear regression model.
  Answer: **FALSE**

- Autocorrelation is the correlation between each of the $e_i$ variables.
  Answer: **TRUE**

- Heteroskedasticity means having constant Error Variance.
  Answer: **FALSE**

Georgia Tech

# Data Analytics in Business

## Linear Regression

**Sridhar Narasimhan, Ph.D**
*Professor*
Scheller College of Business

**Common Problems and Fixes in Fitting Linear Regression, Part 2**

Georgia Tech

# 4. Outliers

- An outlier is a point that has a $y_i$ value that is far from its predicted value, $\hat{y}_i$
- One way to visualize outliers is to plot residuals (or, better yet, standardized residuals) against predicted values of *y*
- Outliers could occur because of incorrect data recording or because the phenomenon could very well be non-linear
- Do not assume that an outlier observation should be removed. It may signal a model deficiency (for e.g., a missing predictor)

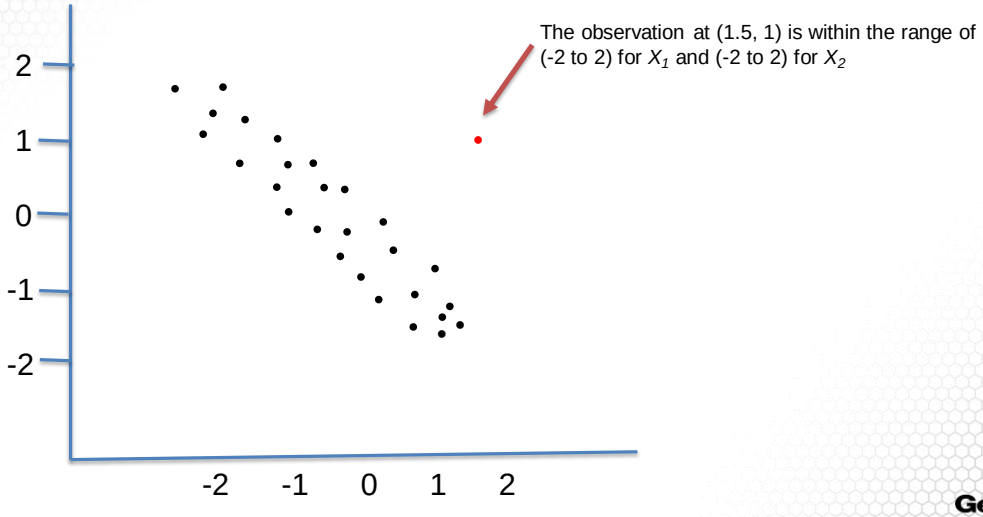
Georgia Tech

# 5. High Leverage Points

- In simple regression, look for observations that have a predictor value outside the normal range of observations
- A point has high leverage if its deletion (by itself or with 2 or 3 other points) causes noticeable changes in the model
- With many predictors in a model, one could have an observation that is within the range of each predictor's value but still be unusual

Georgia Tech

# 5. High Leverage Points - Example




# Cook's Distance

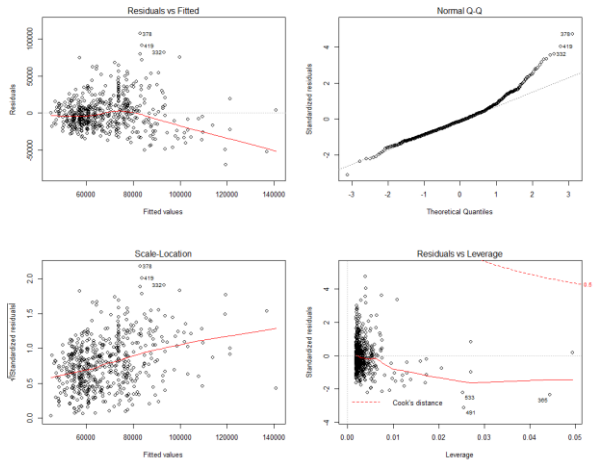
- One statistic to identify influential points is the ***Cook's Distance*** $C_i$ that measures the difference between the regression coefficients obtained
  (a) from the full data and
  (b) from deleting observation *i*
- A rule of thumb is to identify points with $C_i > 1$ as highly influential

# Plot(model) in R

- The plot function in R, when applied to a linear model, provides four plots that are:
  - Residual vs. Fitted (check if residuals have non-linear patterns)
  - Normal Q-Q (check if residuals are normally distributed)
  - Scale-Location (check if $\sqrt{}$ (standardized residuals) are spread equally along the range of fitted values)
  - Residuals vs Leverage (to find influential points, if any, with $C_i > 1$)

Georgia Tech

# a.lm <- lm(formula = price ~ lotsize , data = Housing) plot(a.lm)




Georgia Tech

# Outliers and Influential Points

- **Objective**: In fitting a model to a given body of data, we would like to ensure that the fit is not overly determined by one or a few observations, aka **outliers**
- There are two types of outliers:
  1. **Y (response) outlier**
  2. **X (predictor) outlier, Leverage Point**
- An outlier has the potential to be identified as an **influential** data point if it unduly influences the regression analysis
- The next few slides will define the two types of outliers and how to classify the outlier as an influential point

Reference: Chatterjee S, Hadi AS (2012) *Regression Analysis by Example* 5 edition. (Wiley, Hoboken, New Jersey).

Georgia Tech

# Outliers in the Response Y Variable

- An outlier is a point that has a $y_i$ value that is far from its predicted value $\widehat{y_i}$. It will have a large standardized residual
  - Since the standardized residuals are approximately normally distributed with mean = 0 and a standard deviation = 1, points with standardized residuals larger than 2 or 3 standard deviation away from the mean (zero) are called outliers
  - If removal of the outlier causes substantial change to the regression analysis, then it is an influential outlier (see influential point definition)
- **Detection**: One way to visualize/identify outliers is to plot residuals (or standardized residuals) against predicted values of y ($\widehat{y_i}$)
- Why do outliers occur?
  - Outliers could occur because of incorrect data recording, because of real anomalous events recorded correctly, or because the phenomenon could very well be non-linear
  - Do not assume that an outlier observation should automatically be removed. It may signal a model deficiency (i.e., a missing predictor)

Georgia Tech

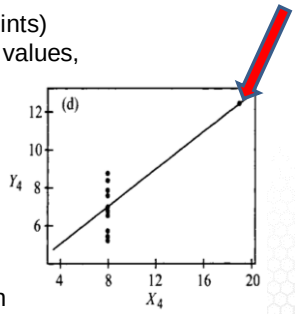
# Outliers in the Predictor X Variable (Leverage Points)

- Extreme *x* values (*x* is the predictor variable) are high leverage points
  - The data point $x_i$ will be unusually out of range of the other predictor X values
  - Does not have a large standardized residual
  - Can affect regression results
- **Detection**: Identify leverage points via index plot, dot plots, box plot, or Cook's Distance (next slide)
  - If the leverage point is flagged via Cook's distance, then it is also an influential point and thus has substantial influence on the fitted model (next slide)
- Why does the leverage point exist?
  - Requires a case-by-case data analysis
  - Often, it is best to analyze the leverage point by creating models with and without the data point to see the effect it has on the fitted line
  - This method of analysis applies to both *y* (response) outliers and influential points

Georgia Tech

# Influential Points

- An outlier is an *influential point* if its deletion (by itself or with 2 or 3 other points) causes substantial changes in the fitted model (estimated coefficients, fitted values, slope, t-tests, hypothesis tests, etc.)
  - Deletion must cause large changes, thus the data point has undue influence
  - See plot of (X,Y) least squares fitted line with an influential point




**Detection:**

- With several variables, we cannot detect influential points graphically
- Measure influential points via **Cook's Distance ($C_i$):** The difference between
  1. the regression coefficients obtained from the full data WITH '*i*th'data point
  2. and the regression coefficients obtained by DELETING the '*i*th' data point
     - Rule of thumb is to identify points with $C_i > 1$ as highly influential

Georgia Tech

# 6. (Multi)Collinearity

- Run these two linear regression models in R and then compare their respective coefficient of cylinders:
  - Reg1 <- lm(formula = mpg ~ cylinders, data = Auto)
  - Reg2 <- lm(formula = mpg ~ cylinders + displacement + weight, data = Auto)
- Reg1

  cylinders -3.5581 0.1457 -24.43 <2e-16 ***
- Reg2

  cylinders -0.2678 0.4131 -0.648 0.517
- In Reg2, cylinder's coefficient is no longer statistically significant!

Georgia Tech

# What Could Be the Reason?

- Such a change in a parameter estimate could indicate the presence of multicollinearity in Reg2
- Multicollinearity: two or more of the explanatory variables are more or less linearly related
- To detect multicollinearity, one approach is to use **Variance Inflation Factors (VIF)**
- Regress predictor variable $X_j$ against all other predictor (X) variables. Name the resulting $R^2$ as $R_j^2$
- Define $VIF_j = 1/(1-R_j^2)$ , $j = 1,2,..., p$
- If $X_j$ has a strong linear relationship to other X variables, then $R_j^2$ is close to 1, and $VIF_j$ will be large.
- Values of VIF > 5 signify presence of multicollinearity (rule of thumb)

Georgia Tech

# VIF

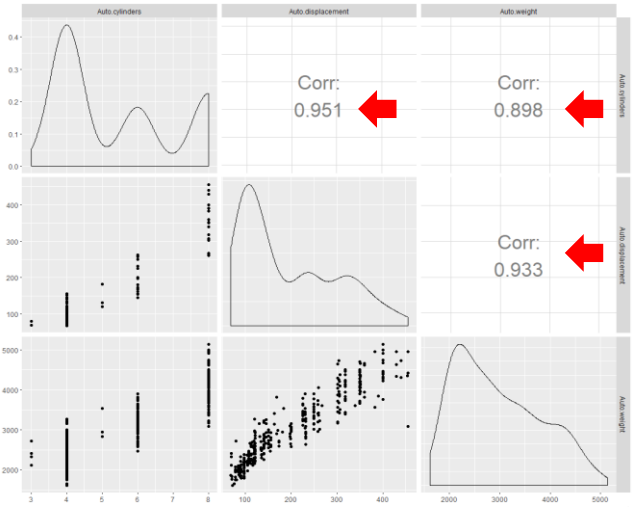
- We use the vif function on the predictors of the Reg2 model and obtain

```
vif(Reg2)
     cylinders      displacement      weight
     10.515508      15.786455       7.788716
```

- These high VIF values indicate Multicollinearity problems

Georgia Tech

# Correlation Matrix




Georgia Tech

# Consequences of Multicollinearity

- Multicollinearity: the explanatory variables are highly correlated
  - If $VIF_j = 1/(1-R_j^2) > 5$, multicollinearity present…
- Consequences of multicollinearity:
  - OLS estimated parameters may have large variances and covariances, thus making precise estimation difficult
  - The confidence intervals of the estimated parameters tend to be bigger, hence we may not be able to reject $H_0$ (the null hypothesis, $b_i = 0$)
  - Regression coefficients have the wrong sign, or
  - Regression coefficients are not significantly different from 0 although $R^2$ is high
  - Adding an explanatory variable changes other variables' coefficients

Georgia Tech

# Consequences of Multicollinearity

- Solution?
  - Pick one variable if two measure the same "thing"
  - Use Principal Components Analysis or Factor Analysis to create more useful variable(s)

Georgia Tech

# Recap of this Module

A. Steps in Regression Analysis
B. A Real Estate example
C. Notation
D. $R^2$, Adjusted $R^2$
E. Simple Regression (One Predictor Variable) Using R
F. Multiple Regression
G. $R^2$, Adjusted $R^2$ from Multiple Regression
H. Common Problems and Fixes in Linear Regression