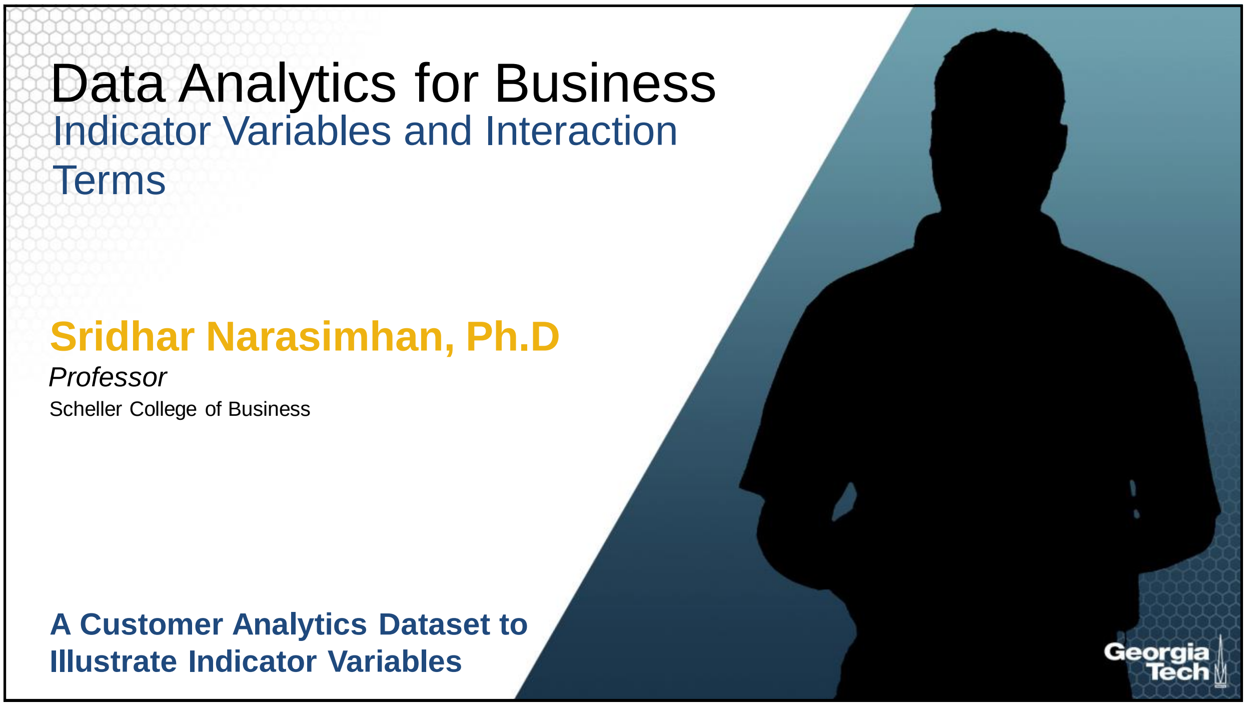


Data Analytics for Business

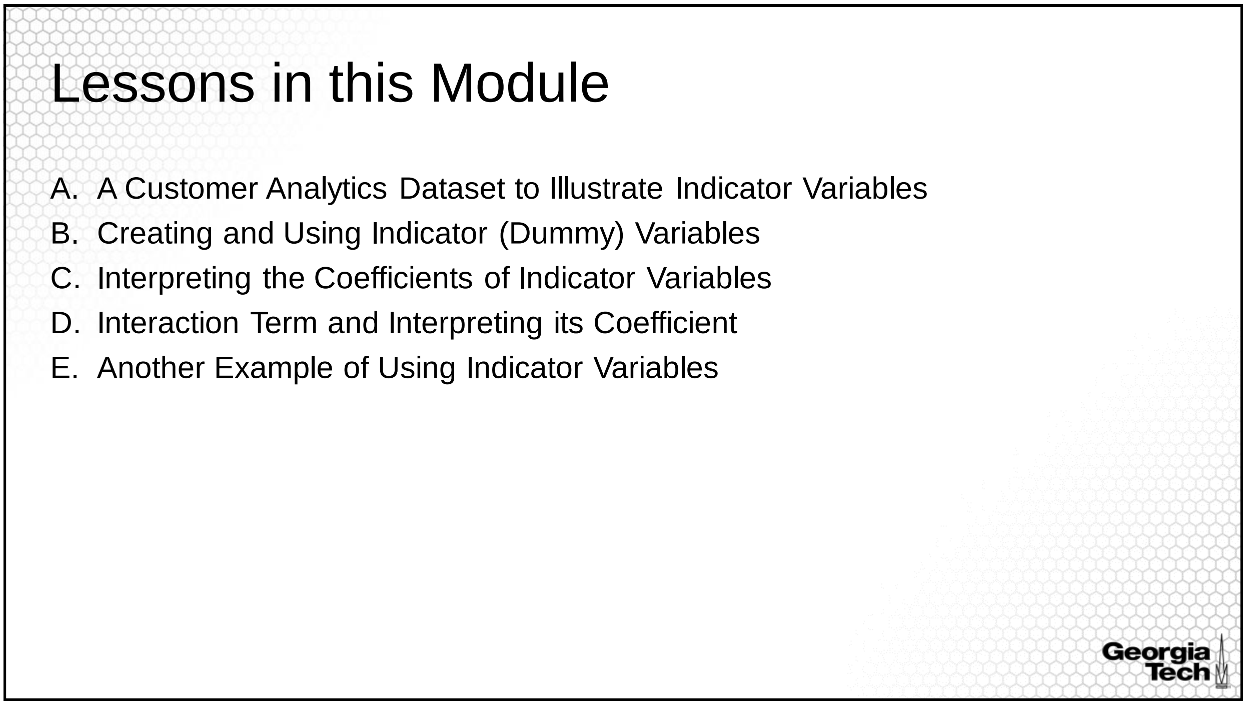
Indicator Variables and Interaction Terms

Sridhar Narasimhan, Ph.D

Professor

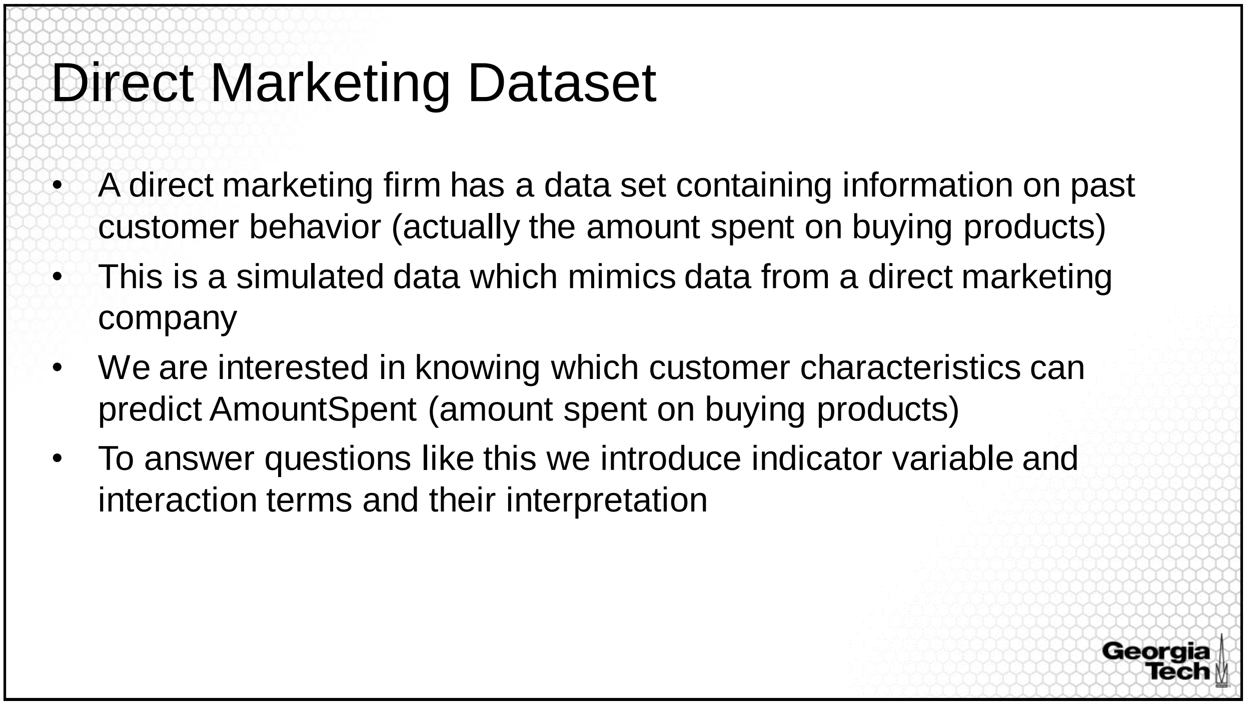
Scheller College of Business

A Customer Analytics Dataset to Illustrate Indicator Variables



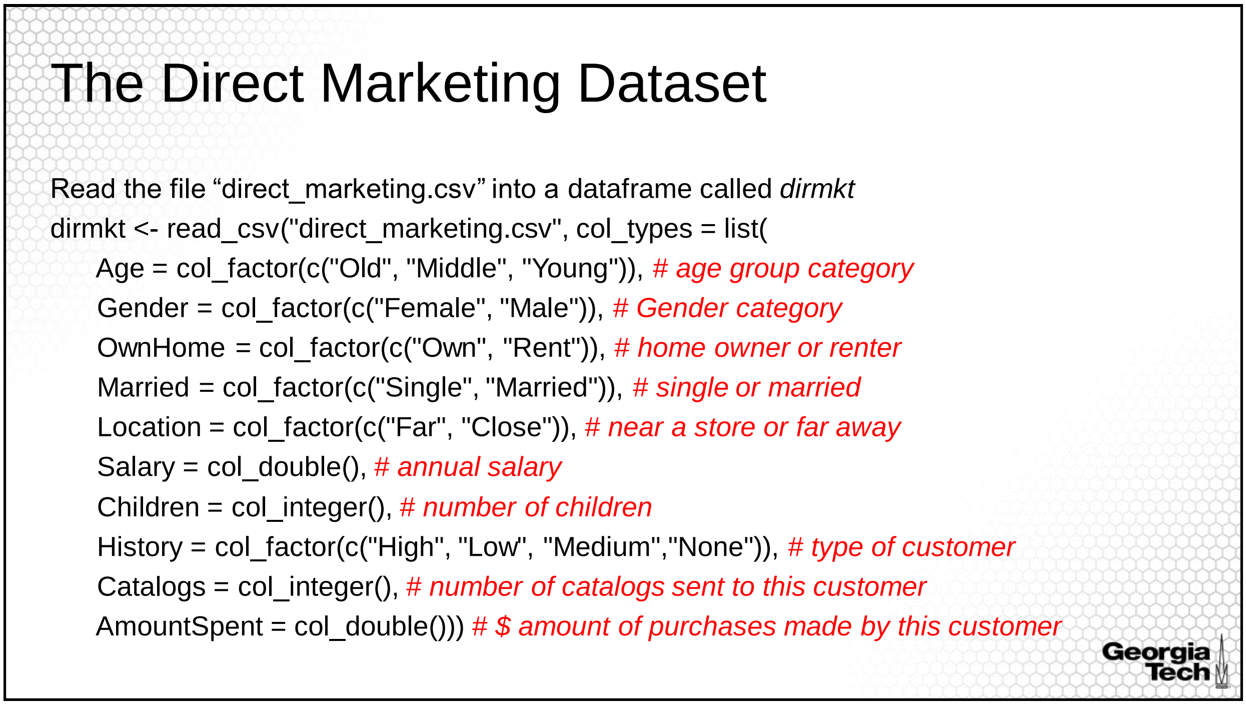
Lessons in this Module

A. A Customer Analytics Dataset to Illustrate Indicator Variables
B. Creating and Using Indicator (Dummy) Variables
C. Interpreting the Coefficients of Indicator Variables
D. Interaction Term and Interpreting its Coefficient
E. Another Example of Using Indicator Variables



Direct Marketing Dataset

- A direct marketing firm has a data set containing information on past customer behavior (actually the amount spent on buying products)
- This is a simulated data which mimics data from a direct marketing company
- We are interested in knowing which customer characteristics can predict AmountSpent (amount spent on buying products)
- To answer questions like this we introduce indicator variable and interaction terms and their interpretation



The Direct Marketing Dataset

Read the file “direct_marketing.csv” into a dataframe called *dirmkt*

```
dirmkt <- read_csv("direct_marketing.csv", col_types = list(
    Age = col_factor(c("Old", "Middle", "Young")), # age group category
    Gender = col_factor(c("Female", "Male")), # Gender category
    OwnHome = col_factor(c("Own", "Rent")), # home owner or renter
    Married = col_factor(c("Single", "Married")), # single or married
    Location = col_factor(c("Far", "Close")), # near a store or far away
    Salary = col_double(), # annual salary
    Children = col_integer(), # number of children
    History = col_factor(c("High", "Low", "Medium","None")), # type of customer
    Catalogs = col_integer(), # number of catalogs sent to this customer
    AmountSpent = col_double())) # $ amount of purchases made by this customer
```

The Direct Marketing Dataset…

```
str(dirmkt)  # what happened to the first row of the csv file?
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':      1000 obs. of  10 variables:
     $ Age        : Factor w/ 3 levels "Old","Middle",..: 1 2 3 2 2 3 2 2 2 1 ...
     $ Gender     : Factor w/ 2 levels "Female","Male": 1 2 1 2 1 2 1 2 1 2 ...
     $ OwnHome    : Factor w/ 2 levels "Own","Rent": 1 2 2 1 1 1 2 1 1 1 ...
     $ Married    : Factor w/ 2 levels "Single","Married": 1 1 1 2 1 2 1 1 2 2 ...
     $ Location   : Factor w/ 2 levels "Far","Close": 1 2 2 2 2 2 2 2 2 1 ...
     $ Salary     : num  47500 63600 13500 85600 68400 30400 48100 68400 51900 80700 ...
     $ Children   : int  0 0 0 1 0 0 0 0 3 0 ...
     $ History    : Factor w/ 4 levels "High","Low","Medium",..: 1 1 2 1 1 2 3 1 2 4 ...
     $ Catalogs   : int  6 6 18 18 12 6 12 18 6 18 ...
     $ AmountSpent: num  75.5 131.8 29.6 243.6 130.4 ...
```

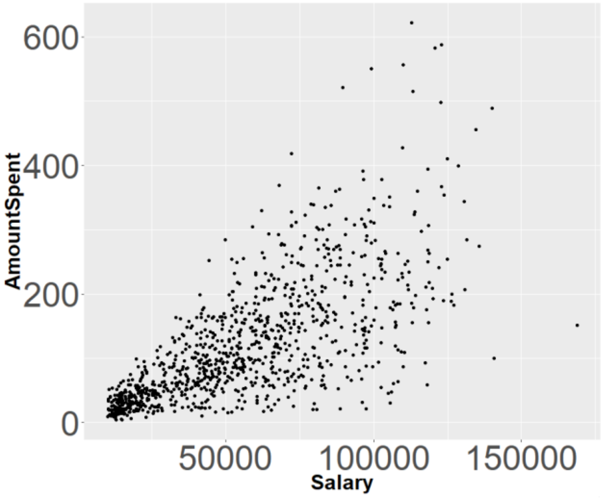
First 10 Rows of the *dirmkt* Dataframe

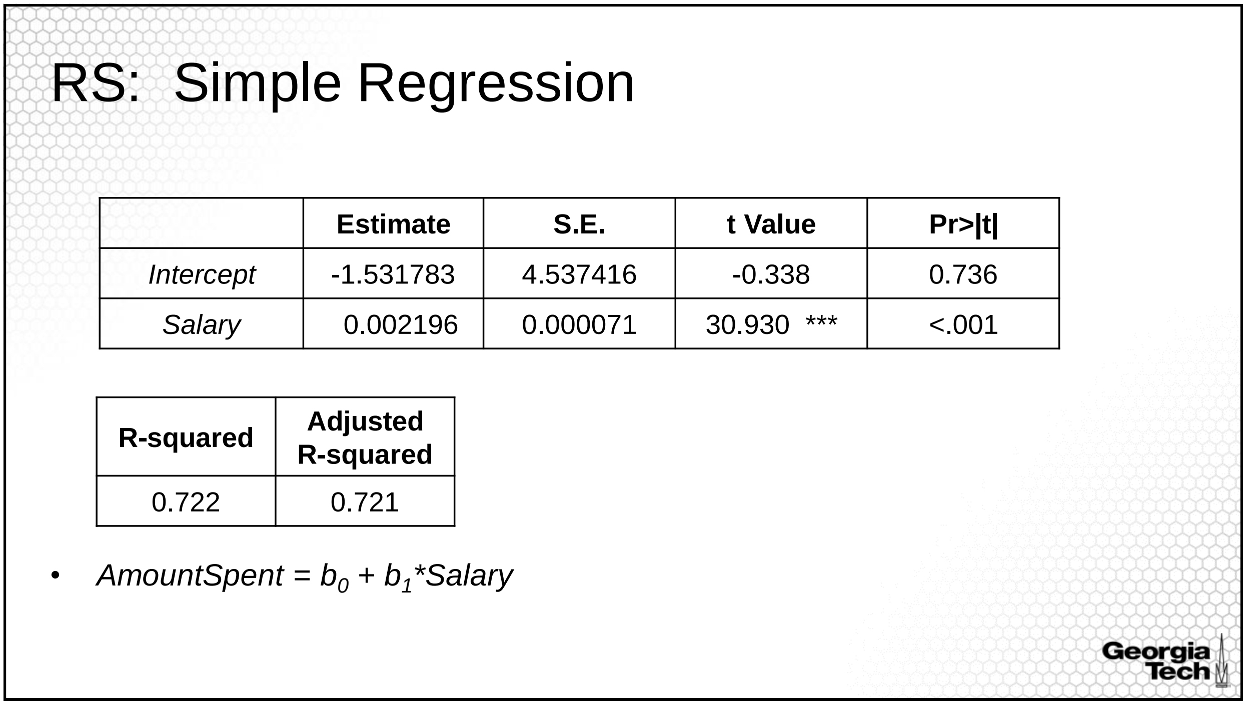
Age	Gender	OwnHome	Married	Location	Salary	Children	History	Catalogs	AmountSpent
Old	Female	Own	Single	Far	47500	0	High	6	75.5
Middle	Male	Rent	Single	Close	63600	0	High	6	131.8
Young	Female	Rent	Single	Close	13500	0	Low	18	29.6
Middle	Male	Own	Married	Close	85600	1	High	18	243.6
Middle	Female	Own	Single	Close	68400	0	High	12	130.4
Young	Male	Own	Married	Close	30400	0	Low	6	49.5
Middle	Female	Rent	Single	Close	48100	0	Medium	12	78.2
Middle	Male	Own	Single	Close	68400	0	High	18	115.5
Middle	Female	Own	Married	Close	51900	3	Low	6	15.8
Old	Male	Own	Married	Far	80700	0	None	18	303.4

Exploring the *dirmkt* Dataframe

- We would like to understand better the reasons why some individuals spend more than others
- In particular, we would like to investigate whether salary has an influence on AmountSpent
- So, how do we get started?
- To keep matters clear, we start with
  AmountSpent, Salary

Scatterplot



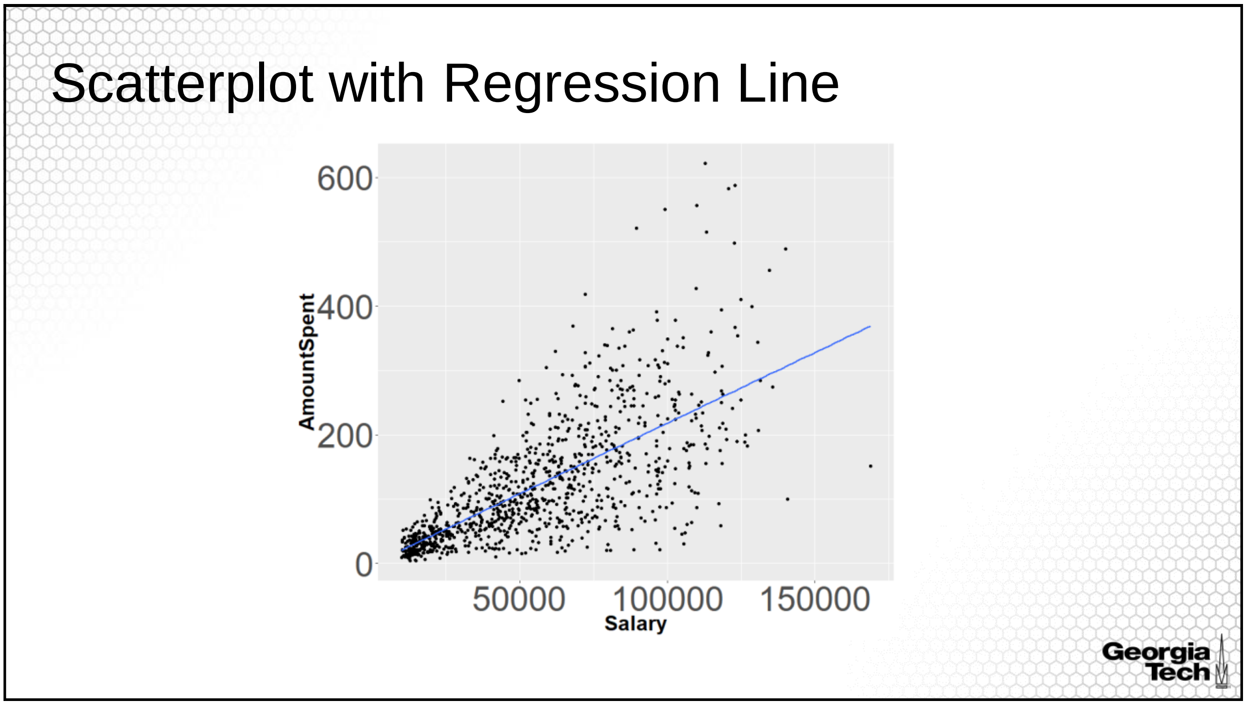


RS: Simple Regression

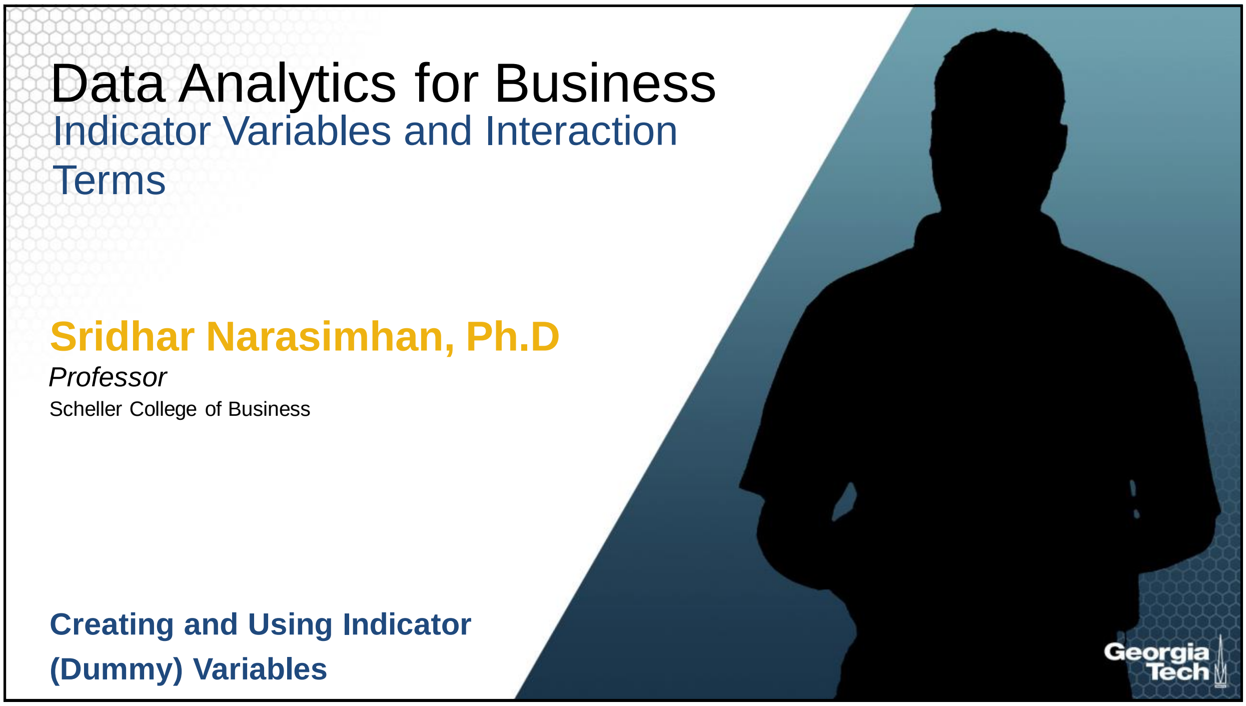
	Estimate	S.E.	t Value	Pr>|t|
Intercept	-1.531783	4.537416	-0.338	0.736
Salary	0.002196	0.000071	30.930 ***	<.001

R-squared	Adjusted R-squared
0.722	0.721

- *AmountSpent* = $b_0 + b_1$*Salary*



Scatterplot with Regression Line

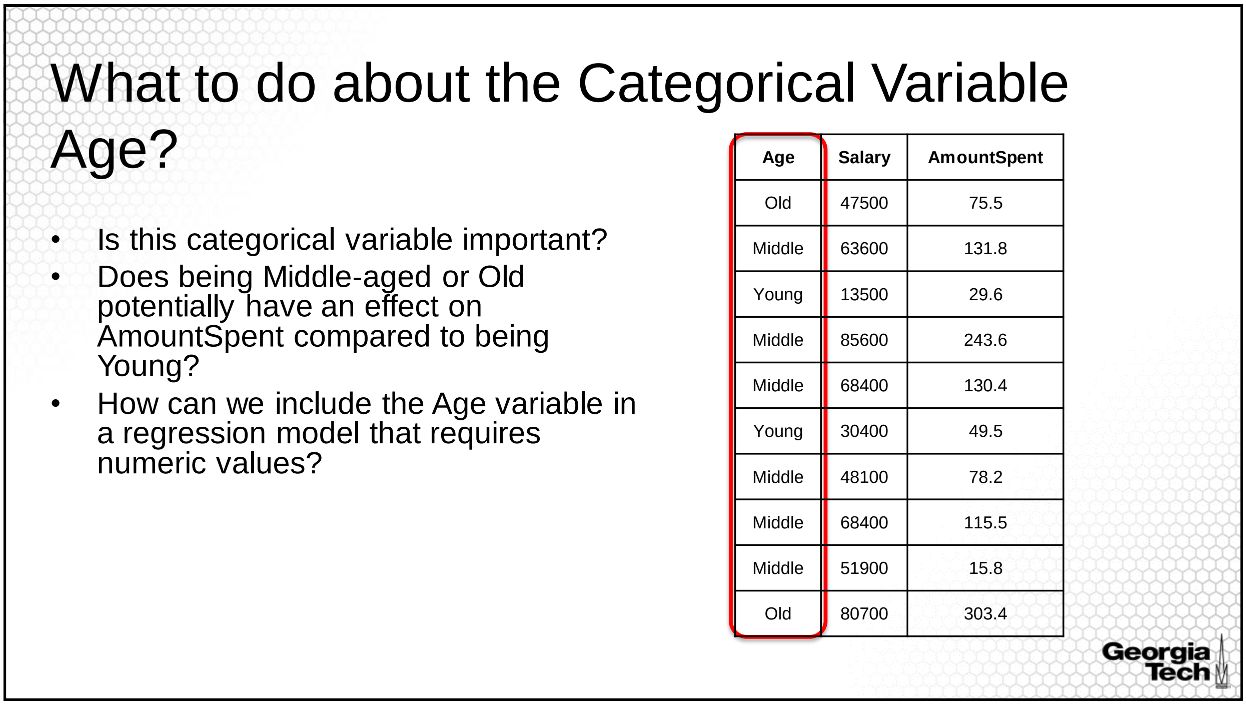


Data Analytics for Business

Indicator Variables and Interaction Terms

Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business

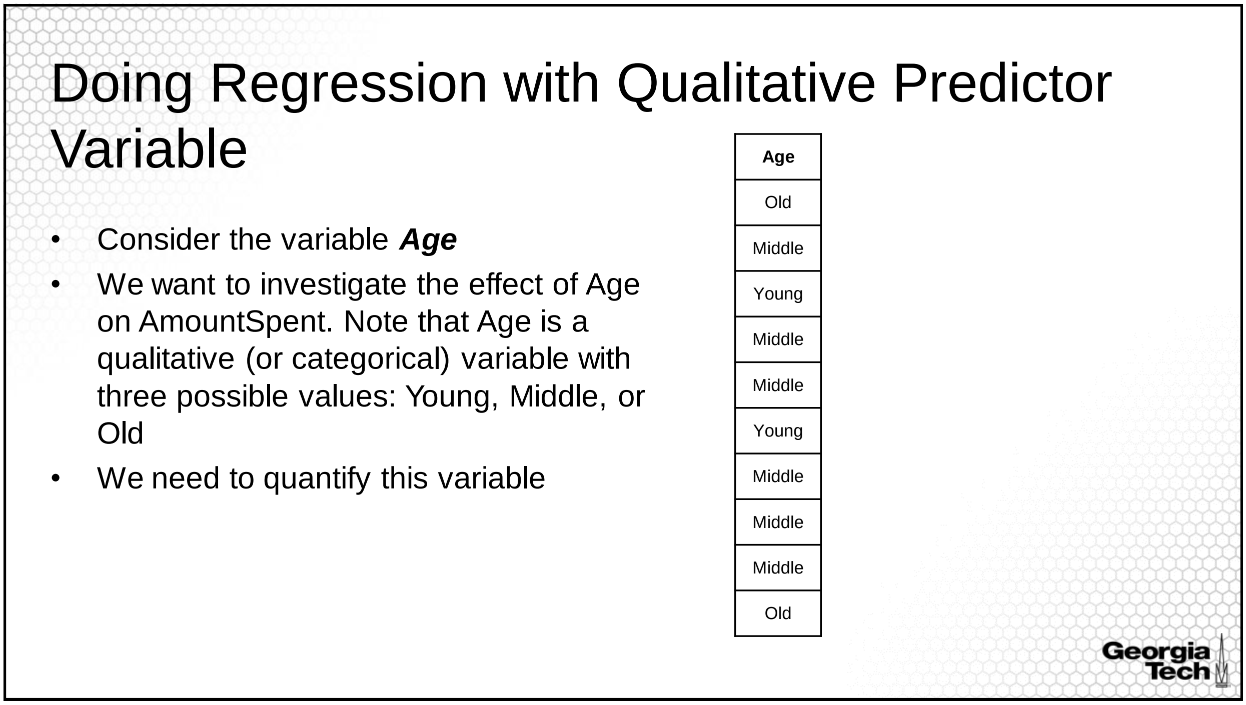
Creating and Using Indicator (Dummy) Variables



What to do about the Categorical Variable Age?

- Is this categorical variable important?
- Does being Middle-aged or Old potentially have an effect on AmountSpent compared to being Young?
- How can we include the Age variable in a regression model that requires numeric values?

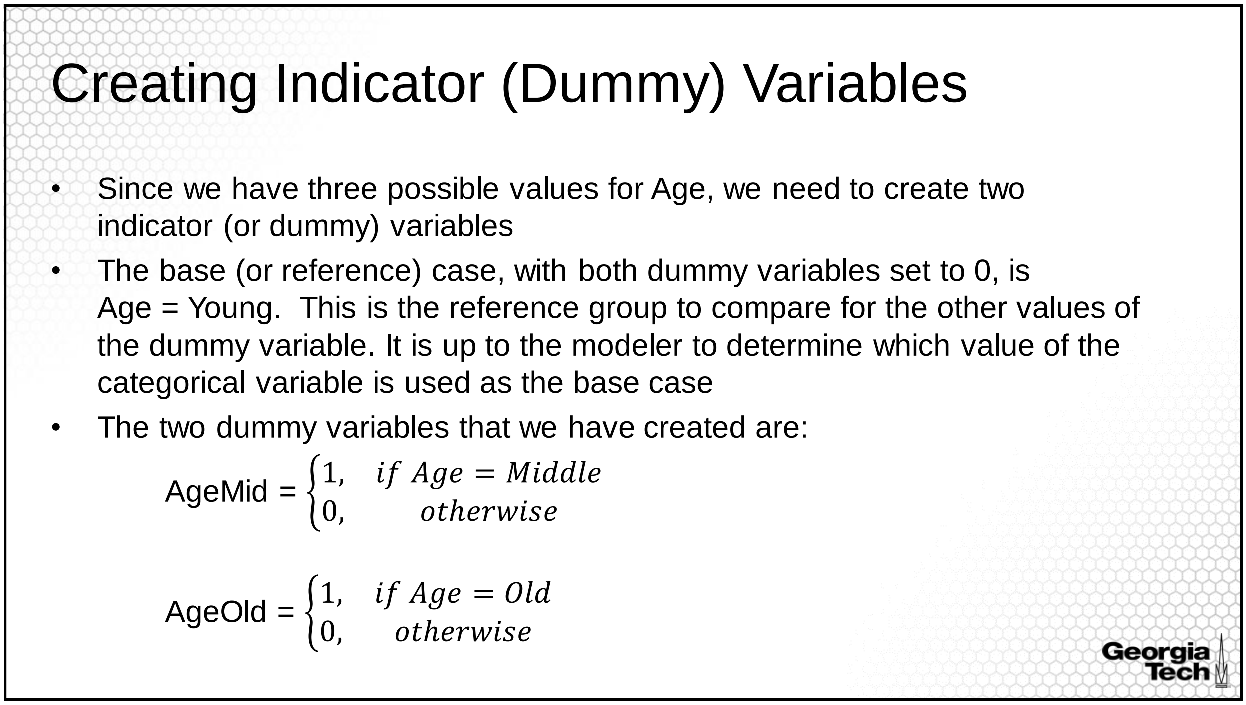
Age	Salary	AmountSpent
Old	47500	75.5
Middle	63600	131.8
Young	13500	29.6
Middle	85600	243.6
Middle	68400	130.4
Young	30400	49.5
Middle	48100	78.2
Middle	68400	115.5
Middle	51900	15.8
Old	80700	303.4



Doing Regression with Qualitative Predictor Variable

- Consider the variable ***Age***
- We want to investigate the effect of Age on AmountSpent. Note that Age is a qualitative (or categorical) variable with three possible values: Young, Middle, or Old
- We need to quantify this variable

Age
Old
Middle
Young
Middle
Middle
Young
Middle
Middle
Middle
Old



Creating Indicator (Dummy) Variables

- Since we have three possible values for Age, we need to create two indicator (or dummy) variables
- The base (or reference) case, with both dummy variables set to 0, is Age = Young. This is the reference group to compare for the other values of the dummy variable. It is up to the modeler to determine which value of the categorical variable is used as the base case
- The two dummy variables that we have created are:

$$\text{AgeMid} = \begin{cases} 1, & if\ Age = Middle \\ 0, & otherwise \end{cases}$$

$$\text{AgeOld} = \begin{cases} 1, & if\ Age = Old \\ 0, & otherwise \end{cases}$$

Assigning Values (0 or 1) to the New Indicator (Dummy) Variables

AgeMid = $\begin{cases} 1, & if\ Age = Middle \\ 0, & otherwise \end{cases}$

AgeOld = $\begin{cases} 1, & if\ Age = Old \\ 0, & otherwise \end{cases}$

We then run the regression,
AmountSpent = $b_0 + b_1$**AgeMid* + b_2**AgeOld*

Age	AgeMid	AgeOld
Old	0	1
Middle	1	0
Young	0	0
Middle	1	0
Middle	1	0
Young	0	0
Middle	1	0
Middle	1	0
Middle	1	0
Old	0	1

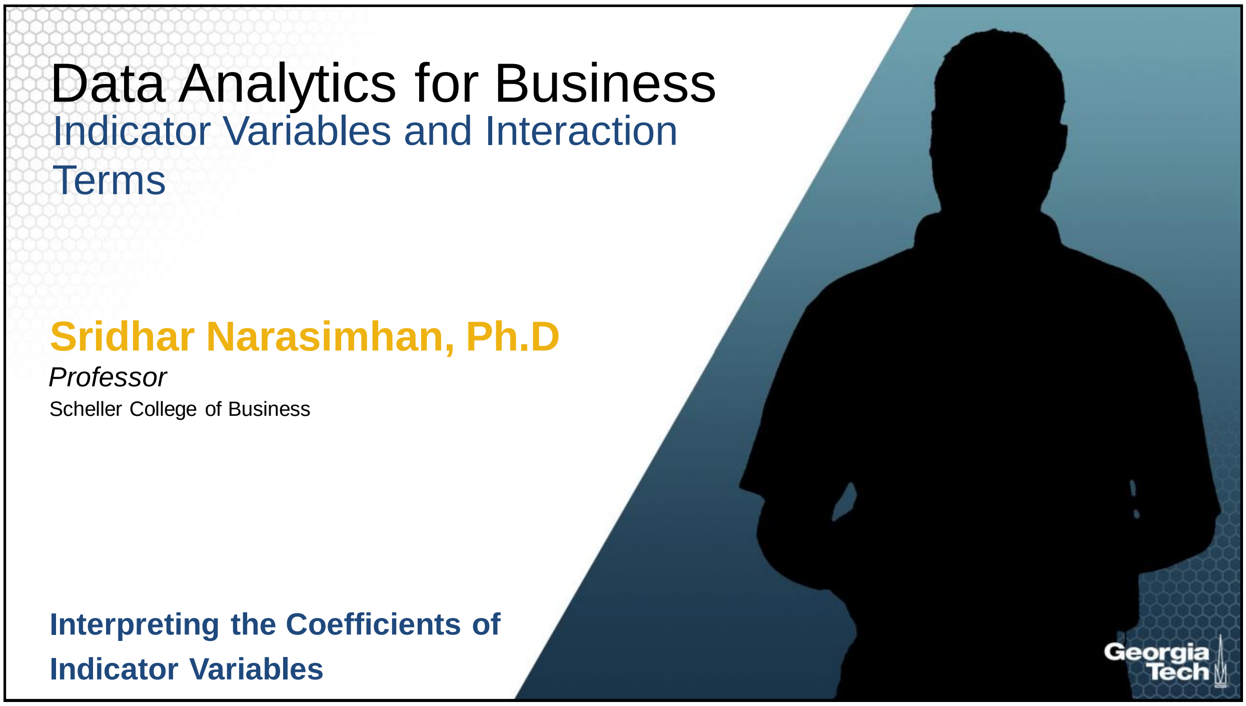
Quiz

With this Indicator variables coding scheme,

AgeMid = $\begin{cases} 1, & if\ Age = Middle \\ 0, & otherwise \end{cases}$ AgeOld = $\begin{cases} 1, & if\ Age = Old \\ 0, & otherwise \end{cases}$

- Can a record in the *dirmkt* dataframe have this value (AgeMid = 0, AgeOld = 0)? Answer: **YES**, because this record is for someone who is Young, i.e., the base case.

- Can a record in the *dirmkt* dataframe have this value (AgeMid = 1, AgeOld = 1)? Answer: **NO**, every individual has to be in exactly one age category



Data Analytics for Business

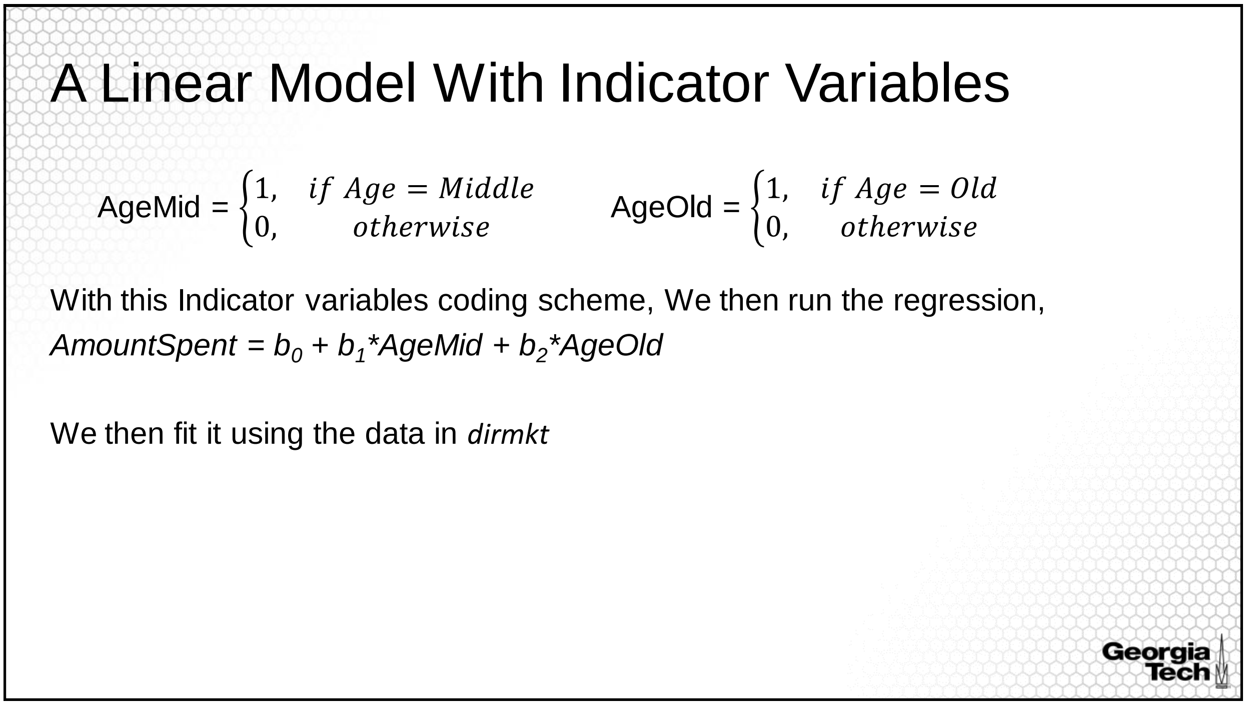
Indicator Variables and Interaction Terms

Sridhar Narasimhan, Ph.D

Professor

Scheller College of Business

Interpreting the Coefficients of Indicator Variables



A Linear Model With Indicator Variables

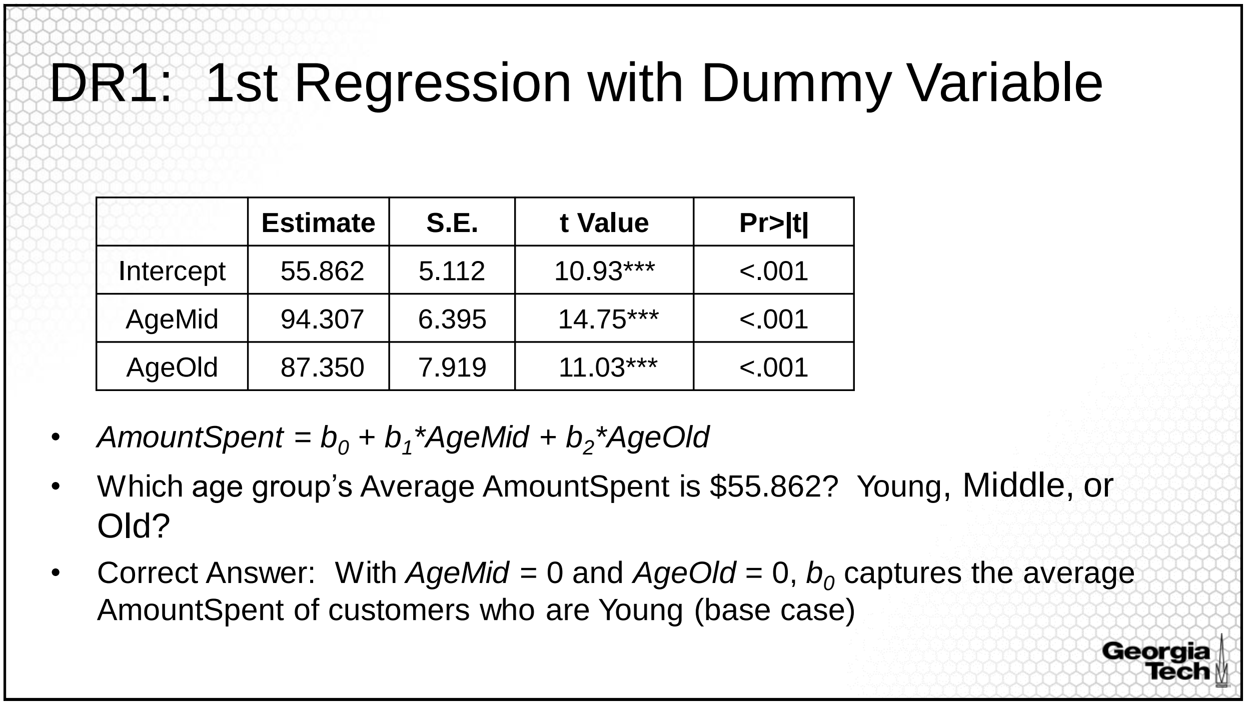
AgeMid = $\begin{cases} 1, & if\ Age = Middle \\ 0, & otherwise \end{cases}$ AgeOld = $\begin{cases} 1, & if\ Age = Old \\ 0, & otherwise \end{cases}$

With this Indicator variables coding scheme, We then run the regression,

$AmountSpent = b_0 + b_1{*}AgeMid + b_2{*}AgeOld$

We then fit it using the data in *dirmkt*

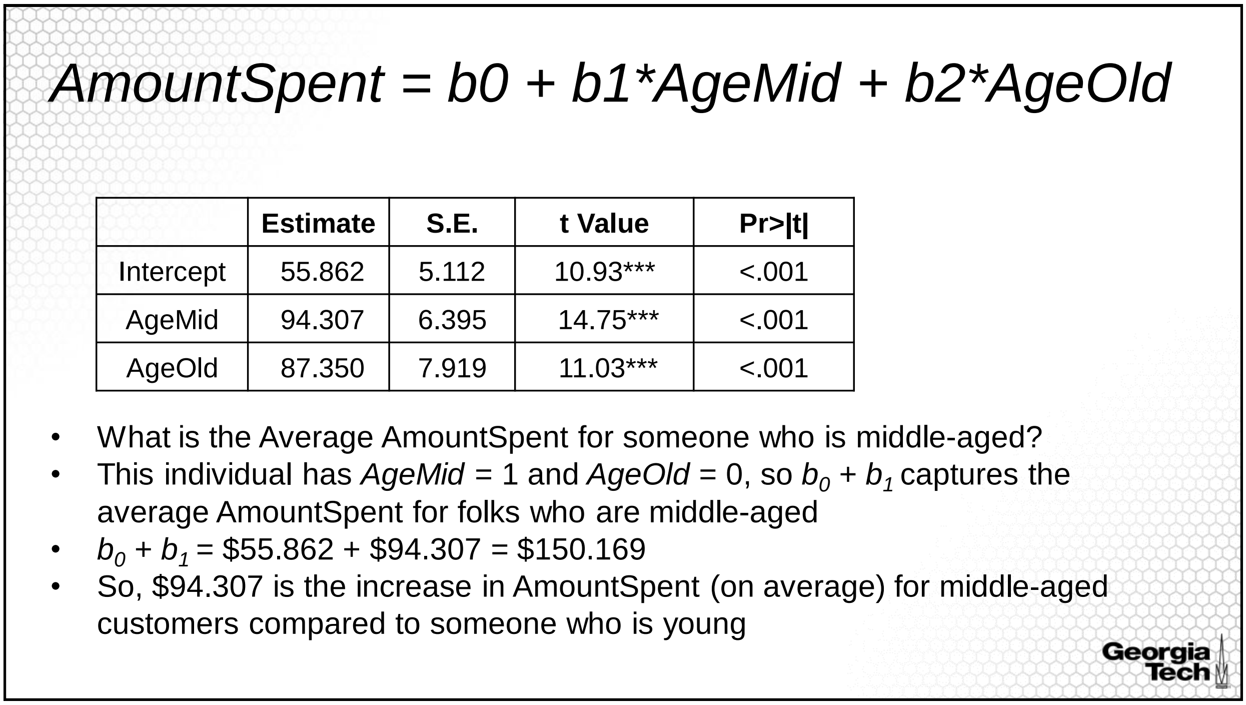
Georgia Tech



DR1: 1st Regression with Dummy Variable

	Estimate	S.E.	t Value	Pr>|t|
Intercept	55.862	5.112	10.93***	<.001
AgeMid	94.307	6.395	14.75***	<.001
AgeOld	87.350	7.919	11.03***	<.001

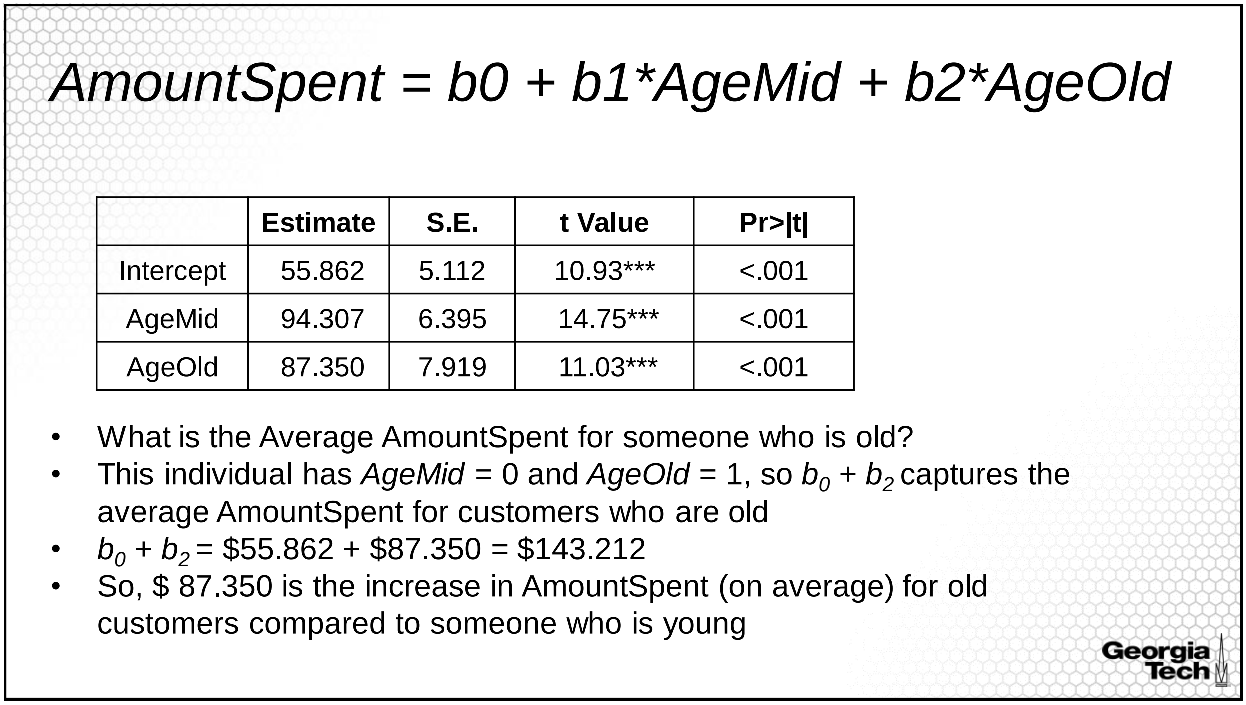
- *AmountSpent* = $b_0 + b_1$**AgeMid* + b_2**AgeOld*
- Which age group's Average AmountSpent is $55.862? Young, Middle, or Old?
- Correct Answer: With *AgeMid* = 0 and *AgeOld* = 0, b_0 captures the average AmountSpent of customers who are Young (base case)



*AmountSpent = b0 + b1*AgeMid + b2*AgeOld*

	Estimate	S.E.	t Value	Pr>|t|
Intercept	55.862	5.112	10.93***	<.001
AgeMid	94.307	6.395	14.75***	<.001
AgeOld	87.350	7.919	11.03***	<.001

- What is the Average AmountSpent for someone who is middle-aged?
- This individual has *AgeMid* = 1 and *AgeOld* = 0, so $b_0 + b_1$ captures the average AmountSpent for folks who are middle-aged
- $b_0 + b_1$ = $55.862 + $94.307 = $150.169
- So, $94.307 is the increase in AmountSpent (on average) for middle-aged customers compared to someone who is young

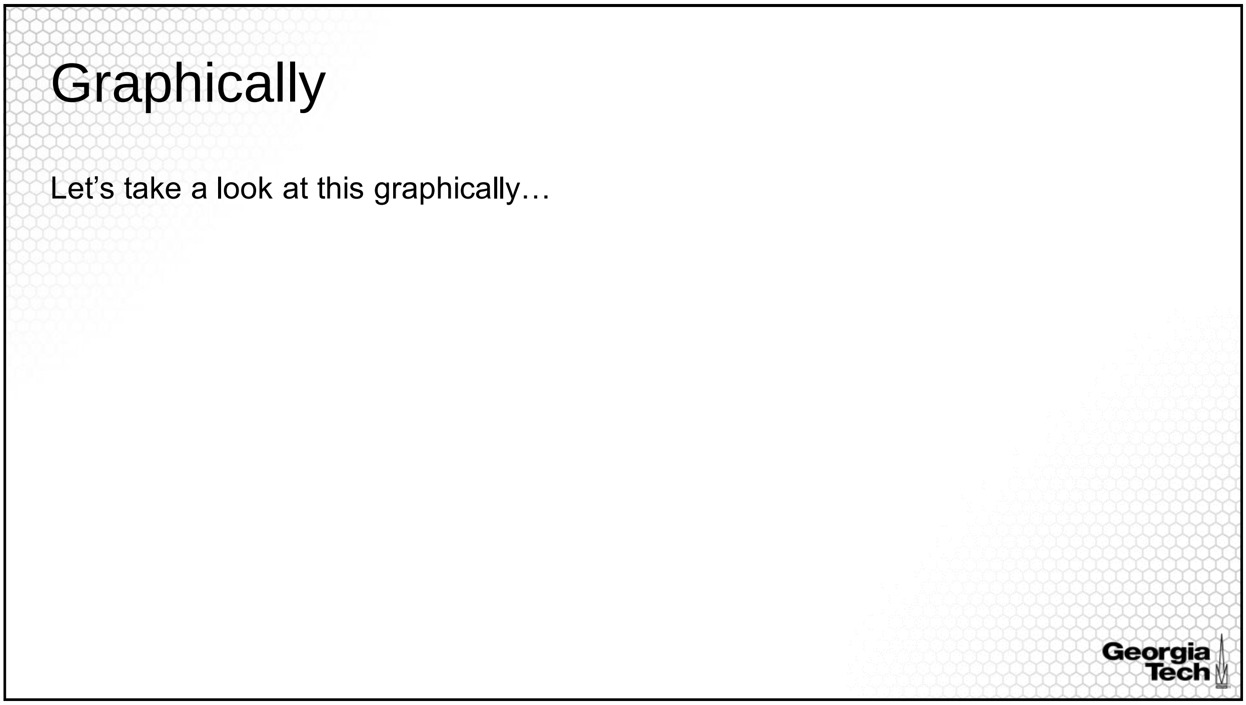


*AmountSpent = b0 + b1*AgeMid + b2*AgeOld*

	Estimate	**S.E.**	**t Value**	**Pr>|t|**
Intercept	55.862	5.112	10.93***	<.001
AgeMid	94.307	6.395	14.75***	<.001
AgeOld	87.350	7.919	11.03***	<.001

- What is the Average AmountSpent for someone who is old?
- This individual has *AgeMid* = 0 and *AgeOld* = 1, so $b_0 + b_2$ captures the average AmountSpent for customers who are old
- $b_0 + b_2$ = $55.862 + $87.350 = $143.212
- So, $ 87.350 is the increase in AmountSpent (on average) for old customers compared to someone who is young

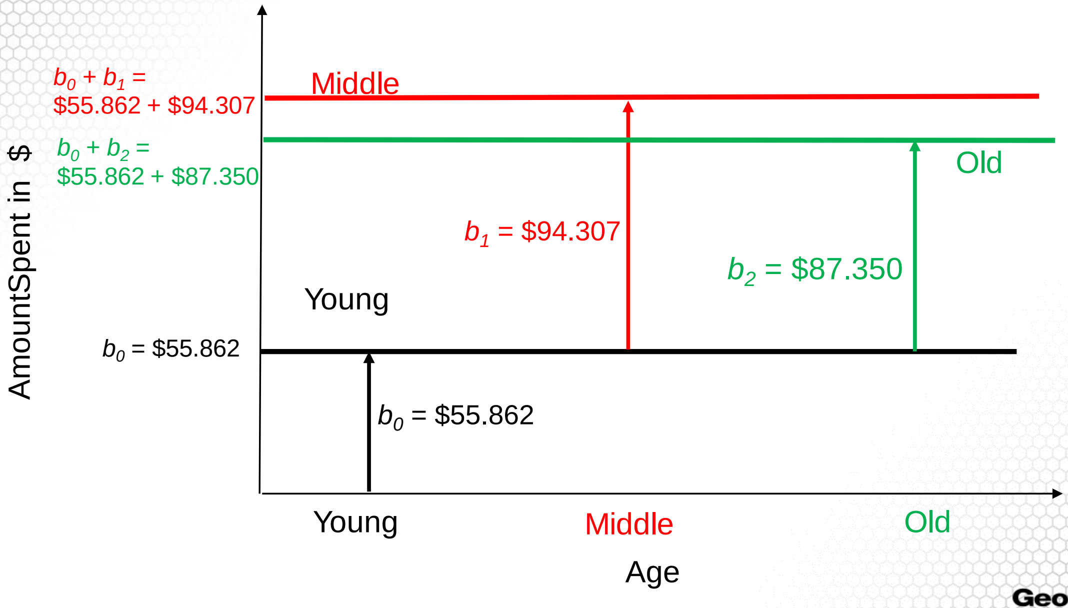
Georgia Tech



Graphically

Let’s take a look at this graphically…

Georgia Tech



Important Note

You can directly use a Factor Variable in regression in R instead of creating & using Dummy variables

lm(AmountSpent ~ Age, data=dirmkt)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	143.213	6.048	23.678	<2e-16 ***
AgeMiddle	6.956	7.165	0.971	0.332
AgeYoung	-87.350	7.919	-11.030	<2e-16 ***

- What is the base case? What is the average AmountSpent for the base case? The Base Case is Old with Average AmountSpent = $143.213
- What is the Average AmountSpent of Young? $143.213 - $87.350 = $55.863
- What is the Average AmountSpent of Middle? $143.213 + $6.956 = $150.169
- All three groups have the same answers as our coding scheme where Young was the base case!!!

R’s Indicator variable coding

R’s indicator variable coding scheme can be found by using:

contrasts(dirmkt$Age)

```
             Middle    Young
Old             0         0     (Old is the base case in this coding scheme)
Middle          1         0
Young           0         1
```

- In this case R uses a different coding scheme for dummy variables
- I find it more useful to use my own coding scheme!

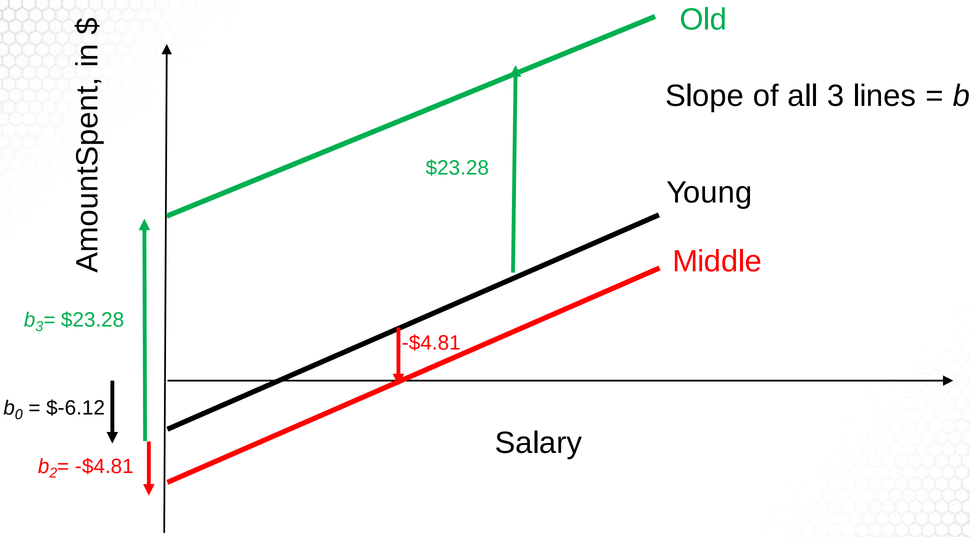
DR2: 2nd Regression with Salary and Dummy Variables

	Estimate	S.E.	t Value	Pr>|t|
Intercept	-6.12	4.72	-1.30	0.20
Salary	.002	.00009	25	<.001
AgeMid	-4.81	6.39	-0.75	0.45
AgeOld	23.28	6.72	3.46	<.001

- $AmountSpent = b_0 + b_1{*}Salary + b_2{*}AgeMid + b_3{*}AgeOld$
- What is the (average) increase in AmountSpent for a one unit increase in Salary?
- Answer: $.002

Graphically

Graphically, *AmountSpent = b0 + b1*Salary + b2*AgeMid + b3*AgeOld* would look like this…



Slope of all 3 lines = b_1 **=** 0.002

Quiz

	Estimate	S.E.	t Value	Pr>|t|
Intercept	-6.12	4.72	-1.30	0.20
Salary	.002	.00009	25	<.001
AgeMid	-4.81	6.39	-0.75	0.45
AgeOld	23.28	6.72	3.46	<.001

*AmountSpent = b0 + b1*Salary + b2*AgeMid + b3*AgeOld*

- What does this result mean?

A. Middle-aged customers spend the most

B. Old customers spend the least

C. Old customers spend more that young customers

D. At the same salary level, old customers spend more than young customers

What is the current answer?

D. At the same salary level, old customers spend more than young customers

Data Analytics for Business

Indicator Variables and Interaction Terms

Sridhar Narasimhan, Ph.D

Professor

Scheller College of Business

Interaction Term and Interpreting its Coefficient

Next Regression with Dummy Variables

- In the same dataset, Location is a categorical variable with a value equal to “Close” if the customer lives close to a store that sells similar merchandise, and has a value equal to “Far” otherwise

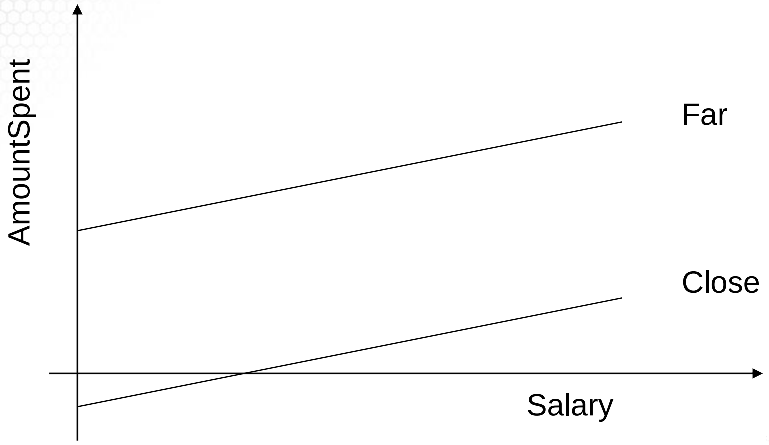
$$\text{Far} = \begin{cases} 1, & if\ Location = Far \\ 0, & otherwise \end{cases}$$

- We want to study the impact of Location on AmountSpent
- *AmountSpent* = $b_0 + b_1 Salary + b_2 Far$

	Estimate	**S.E.**	**t Value**	**Pr>|t|**
Intercept	-20.480	4.413	-4.64	<.0001
Salary	0.002	0.00007	34.05	<.0001
Far	59.060	4.414	13.38	<.0001

Multiple R-Squared: 0.5672,   Adjusted R-squared: 0.5663

Or Graphically…



	Estimate	**S.E.**	**t Value**	**Pr>|t|**
Intercept	-20.480	4.413	-4.64	<.0001
Salary	0.002	0.00007	34.05	<.0001
Far	59.060	4.414	13.38	<.0001

AmountSpent = $b_0 + b_1$*Salary* + b_2*Far*

What is the estimated AmountSpent for a customer who lives far?

A. 59.06

B. -20.48

C. -20.48 + 0.002 * Salary

D. 38.58 + 0.002 * Salary

For this customer since Far =1, the correct answer is
D. because AmountSpent = -20.48 + .002* Salary +59.06 * 1
= 38.58 + .002*Salary

But…

- *AmountSpent* = $b_0 + b_1$*Salary* + b_2*Far*
- In the above model, we assume that customers who live far away from a store that sells similar products will spend (at our direct market firm) at the same rate as customers who live close to a store
- Is this assumption realistic?
- So, can we investigate another scenario that the spending rate may be different? So how should we change the model?

Interaction Term

- Is spending rate higher for customers how live far away?
- To answer this question we need to construct a new variable SalaryFar
- SalaryFar = Salary * Far, is an Interaction Term

- $AmountSpent = b_0 + b_1 Salary + b_2 Far + b_3 SalaryFar$

Regression with Dummy Variable and Interaction Term

$AmountSpent = b_0 + b_1 Salary + b_2 Far + b_3 SalaryFar$

	Estimate	**S.E.**	**t Value**	**Pr>|t|**
Intercept	1.448	4.808	0.30	0.76
Salary	0.002	0.000	24.72	<.0001
Far	-13.460	8.680	-1.55	0.12
SalaryFar	0.001	0.000	9.57	<.0001

Multiple R-Squared: 0.6036, Adjusted R-squared: 0.6024

What does this result mean?

Regression with Experience, Dummy Variables, and Interaction Term

	Estimate	**S.E.**	**t Value**	**Pr>|t|**
Intercept	1.448	4.808	0.30	0.76
Salary	0.002	0.000	24.72	<.0001
Far	-13.460	8.680	-1.55	0.12
SalaryFar	0.001	0.000	9.57	<.0001

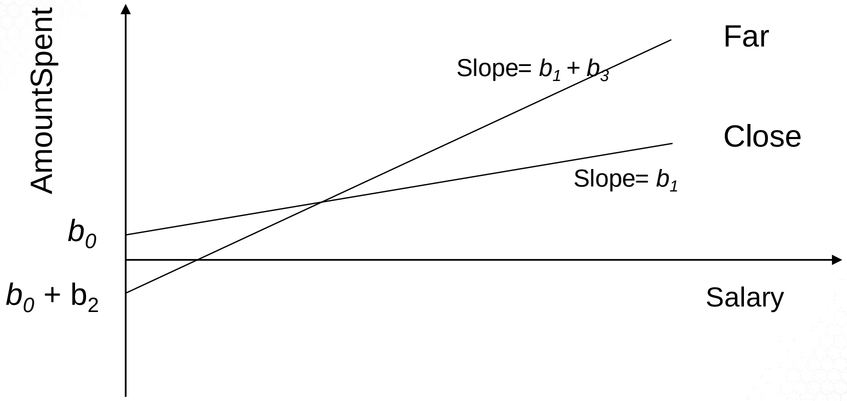
$AmountSpent = b_0 + b_1 Salary + b_2 Far + b_3 SalaryFar$

- How would you interpret b_3 the coefficient of *SalaryFar*?
- b_3 is the amount to add to b_1 to get the slope for individuals who live far away

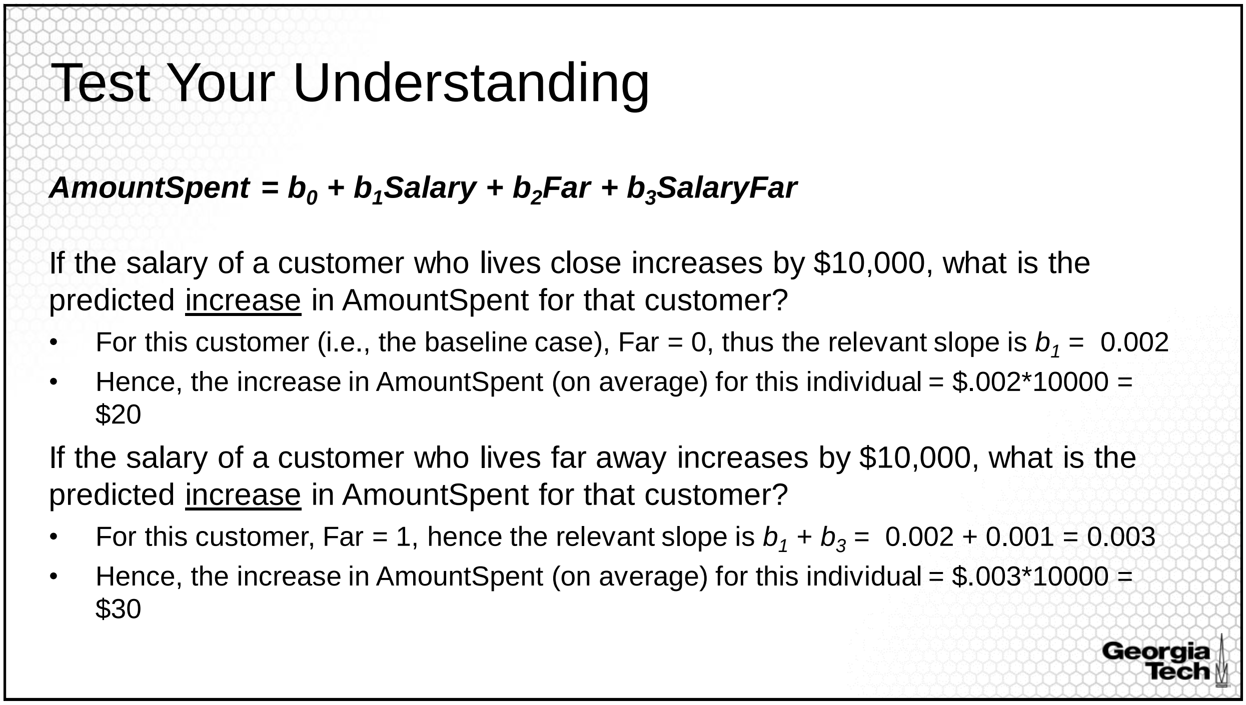
Georgia Tech

Graphically

$AmountSpent = b_0 + b_1 Salary + b_2 Far + b_3 SalaryFar$



Georgia Tech



Test Your Understanding

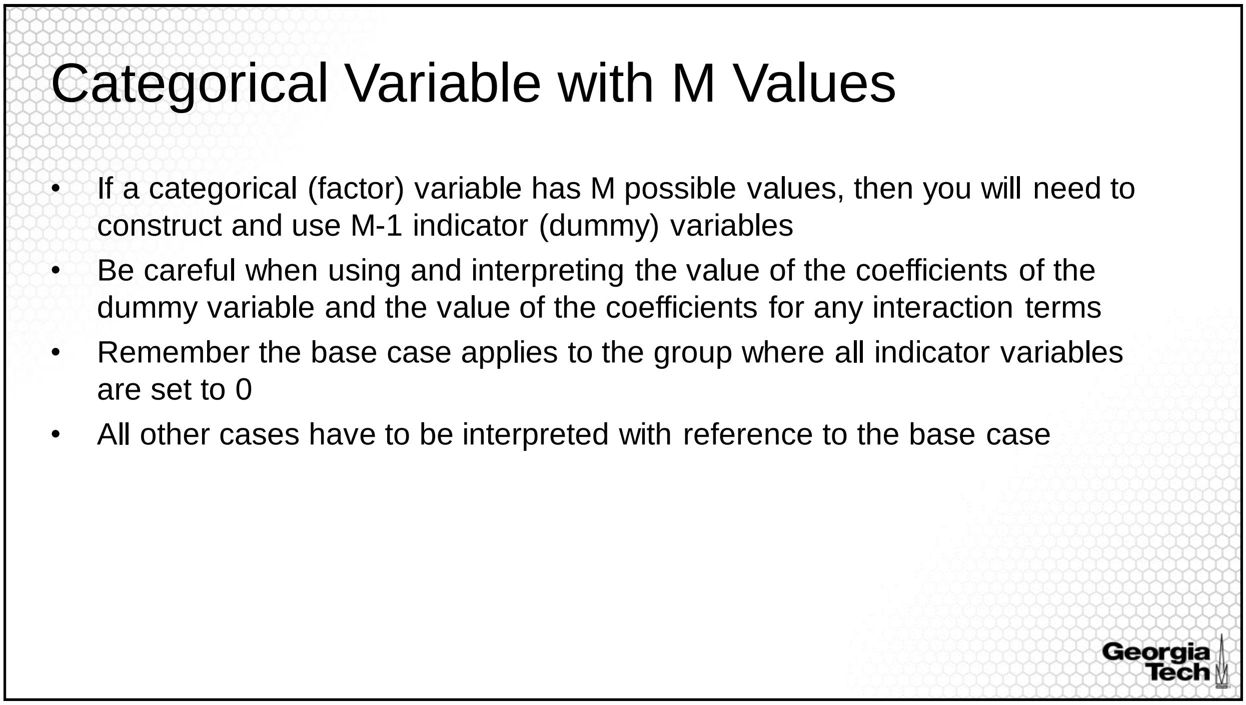
AmountSpent = b_0 + b_1Salary + b_2Far + b_3SalaryFar

If the salary of a customer who lives close increases by $10,000, what is the predicted increase in AmountSpent for that customer?

- For this customer (i.e., the baseline case), Far = 0, thus the relevant slope is b_1 = 0.002
- Hence, the increase in AmountSpent (on average) for this individual = $.002*10000 = $20

If the salary of a customer who lives far away increases by $10,000, what is the predicted increase in AmountSpent for that customer?

- For this customer, Far = 1, hence the relevant slope is b_1 + b_3 = 0.002 + 0.001 = 0.003
- Hence, the increase in AmountSpent (on average) for this individual = $.003*10000 = $30



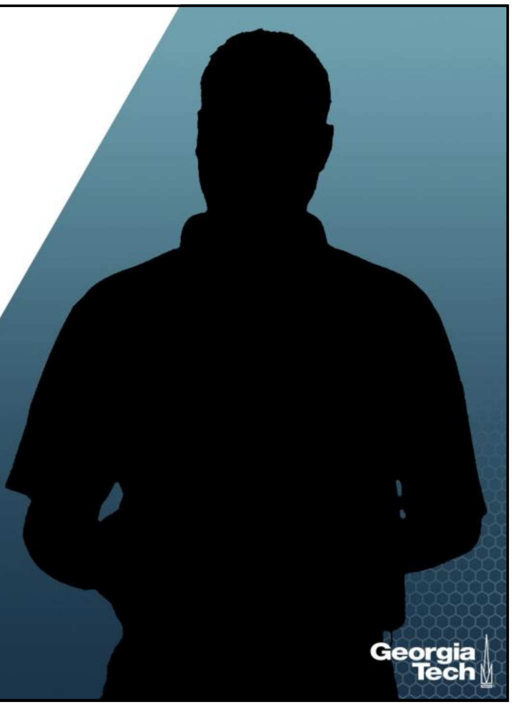
Categorical Variable with M Values

- If a categorical (factor) variable has M possible values, then you will need to construct and use M-1 indicator (dummy) variables
- Be careful when using and interpreting the value of the coefficients of the dummy variable and the value of the coefficients for any interaction terms
- Remember the base case applies to the group where all indicator variables are set to 0
- All other cases have to be interpreted with reference to the base case

Data Analytics for Business

Indicator Variables and Interaction Terms

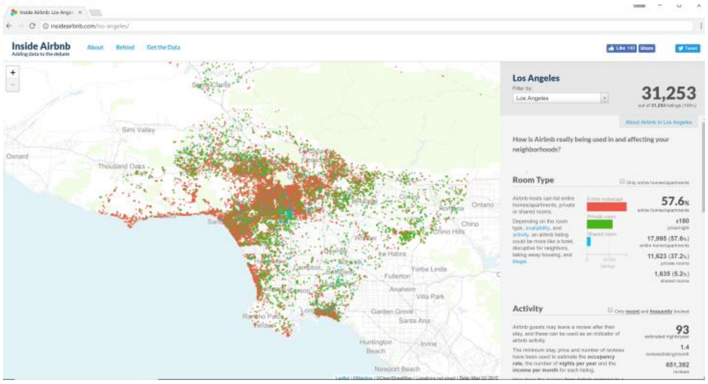
Sridhar Narasimhan, Ph.D
Professor
Scheller College of Business



Another Example of Using Indicator Variables

AirBnB – Los Angeles Rental Market

- Listing data on AirBnB is publicly available at http://insideairbnb.com/los-angeles/ and http://insideairbnb.com/get-the-data.html .



Los Angeles

Filter by: Los Angeles

31,253 out of 31,253 listings (100%)

About Airbnb in Los Angeles

How is Airbnb really being used in and affecting your neighborhoods?

Room Type

□ Only entire homes/apartments

Airbnb hosts can list entire homes/apartments, private or shared rooms.

Depending on the room type, availability, and activity, an airbnb listing could be more like a hotel, disruptive for neighbors, taking away housing, and illegal.

57.6% entire homes/apartments

$180 price/night

17,995 (57.6%) entire homes/apartments

11,623 (37.2%) private rooms

1,635 (5.2%) shared rooms

Activity

□ Only recent and frequently booked

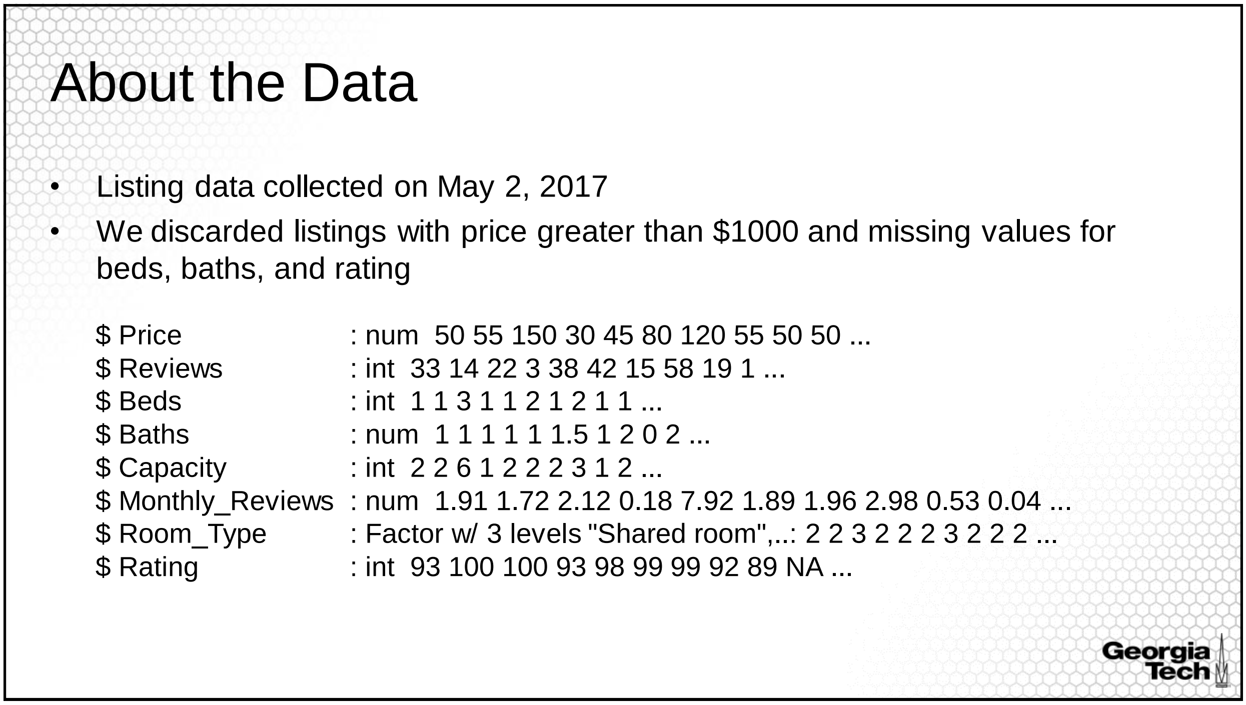
Airbnb guests may leave a review after their stay, and these can be used as an indicator of airbnb activity.

The minimum stay, price and number of reviews have been used to estimate the **occupancy rate**, the number of **nights per year** and the **income per month** for each listing.

93 estimated nights/year

1.4 reviews/listing/month

651,392 reviews

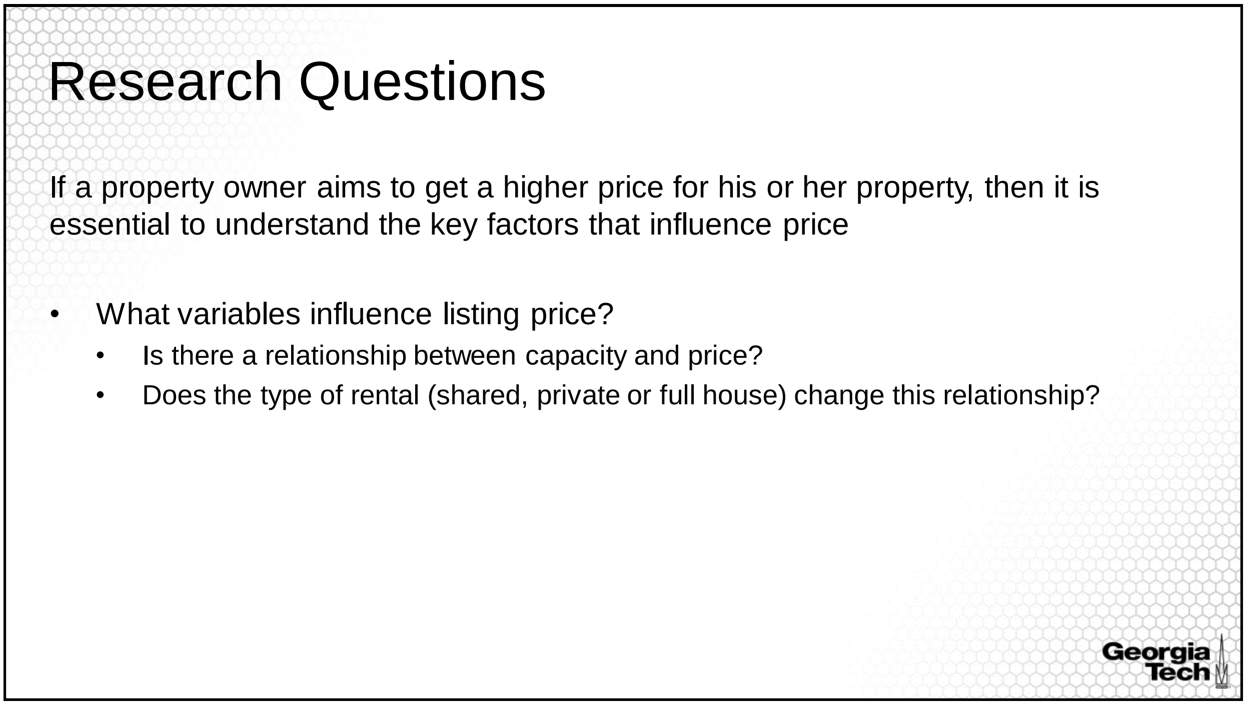


About the Data

- Listing data collected on May 2, 2017
- We discarded listings with price greater than $1000 and missing values for beds, baths, and rating

```
$ Price           : num  50 55 150 30 45 80 120 55 50 50 ...
$ Reviews         : int  33 14 22 3 38 42 15 58 19 1 ...
$ Beds            : int  1 1 3 1 1 2 1 2 1 1 ...
$ Baths           : num  1 1 1 1 1 1.5 1 2 0 2 ...
$ Capacity        : int  2 2 6 1 2 2 2 3 1 2 ...
$ Monthly_Reviews : num  1.91 1.72 2.12 0.18 7.92 1.89 1.96 2.98 0.53 0.04 ...
$ Room_Type       : Factor w/ 3 levels "Shared room",..: 2 2 3 2 2 2 3 2 2 2 ...
$ Rating          : int  93 100 100 93 98 99 99 92 89 NA ...
```

Georgia Tech



Research Questions

If a property owner aims to get a higher price for his or her property, then it is essential to understand the key factors that influence price

- What variables influence listing price?
  - Is there a relationship between capacity and price?
  - Does the type of rental (shared, private or full house) change this relationship?

Georgia Tech

Data Wrangling

```
la_listing <-  la_listing %>%
        mutate(Price = str_replace(Price, "[$]", "")) %>%
        mutate(Price = str_replace(Price, "[,]", "")) %>%
        mutate(Price = as.numeric(Price)) %>%
        mutate(Room_Type = factor(Room_Type, levels = c("Shared room", "Private room", "Entire home/apt"))) %>%
        mutate(Capacity_Sqr = Capacity * Capacity) %>%
        mutate(Beds_Sqr = Beds * Beds) %>%
        mutate(Baths_Sqr = Baths * Baths) %>%
        mutate(ln_Reviews = log(1+Reviews)) %>%
        mutate(ln_Monthly_Reviews = log(1+Monthly_Reviews))
        mutate(ln_Price = log(1+Price)) %>%
        mutate(ln_Beds = log(1+Beds)) %>%
        mutate(ln_Baths = log(1+Baths)) %>%
        mutate(ln_Capacity = log(1+Capacity)) %>%
        mutate(ln_Rating = log(1+Rating)) %>%
        mutate(Shared_ind = ifelse(Room_Type == "Shared room",1,0)) %>%
        mutate(House_ind = ifelse(Room_Type == "Entire home/apt",1,0)) %>%
        mutate(Private_ind = ifelse(Room_Type == "Private room",1,0)) %>%
        mutate(Capacity_x_Shared_ind = Shared_ind * Capacity) %>%
        mutate(Capacity_x_House_ind = House_ind * Capacity) %>%
        mutate(Capacity_x_Private_ind = Private_ind * Capacity) %>%
        mutate(ln_Capacity_x_Shared_ind = Shared_ind * ln_Capacity) %>%
        mutate(ln_Capacity_x_House_ind = House_ind * ln_Capacity) %>%
        mutate(ln_Capacity_x_Private_ind = Private_ind * ln_Capacity)
    filter(Price < 1000 , !is.na(Beds), !is.na(Baths), !is.na(Price), !is.na(Rating))
```

Convert price to numeric and room_type to factor

Create squared terms for testing non-linear relations

Create log terms for testing non-linear relations

Create dummy variables for room_type

Create interaction terms

Filter unwanted data

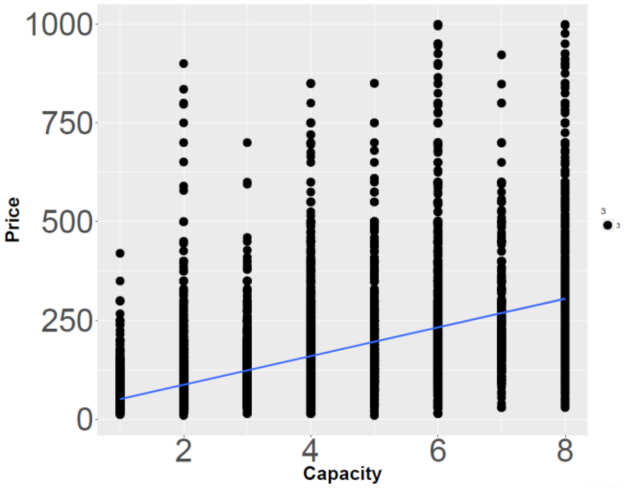
2RS: Simple Regression – How Does Price Vary by Room Capacity?

- *Price* = $b_0 + b_1$***Capacity*

	Estimate	S.E.	t Value	Pr>|t|
Intercept	15.039	1.141	13.19***	<.001
Capacity	38.272	0.316	114.72***	<.001

R-squared	Adjusted R-squared
0.367	0.367

Scatterplot with Regression Line



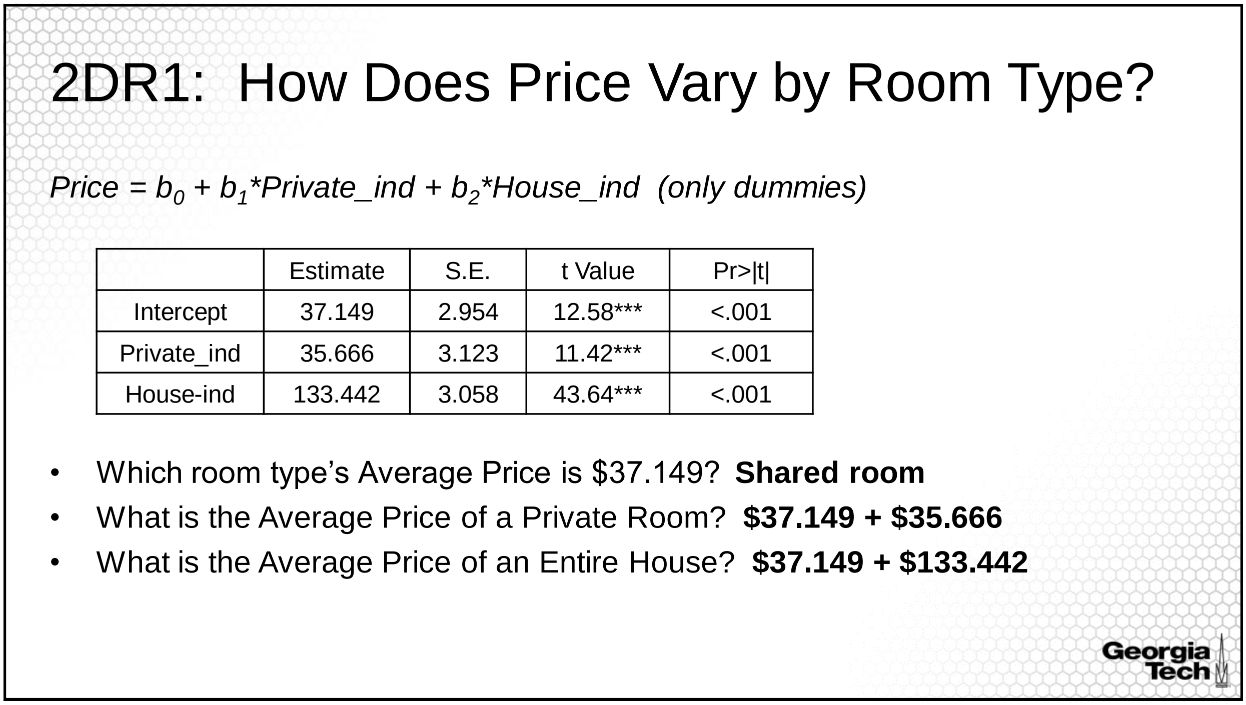
Creating Indicator (Dummy) variables

- We define two dummy variables:

Private_ind = $\begin{cases} 1, & if\ Room\ type = "Private\ room" \\ 0, & otherwise \end{cases}$

House_ind = $\begin{cases} 1, & if\ Room\ type = "Entire\ home/apt" \\ 0, & otherwise \end{cases}$

- The base (or reference) case, with both dummy variables set to 0, is Room type = “Shared.” This is the reference group to compare for the other values of the dummy variable

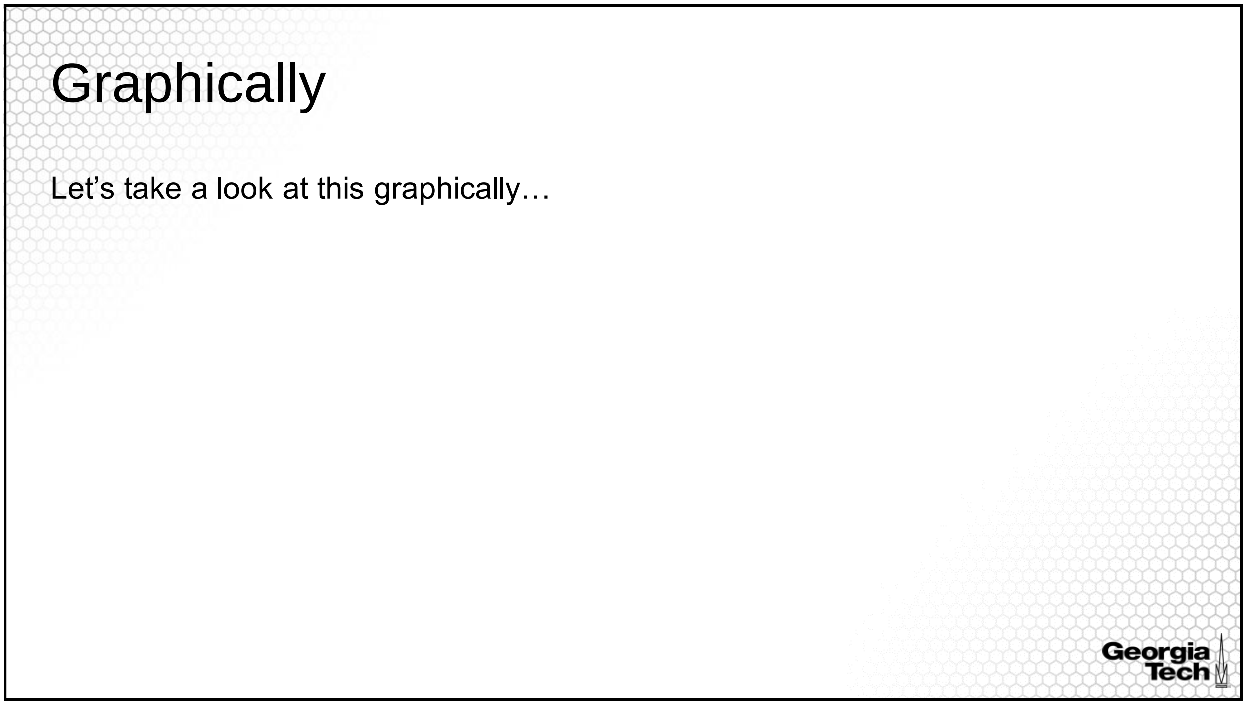


2DR1: How Does Price Vary by Room Type?

*Price = b_0 + b_1*Private_ind + b_2*House_ind (only dummies)*

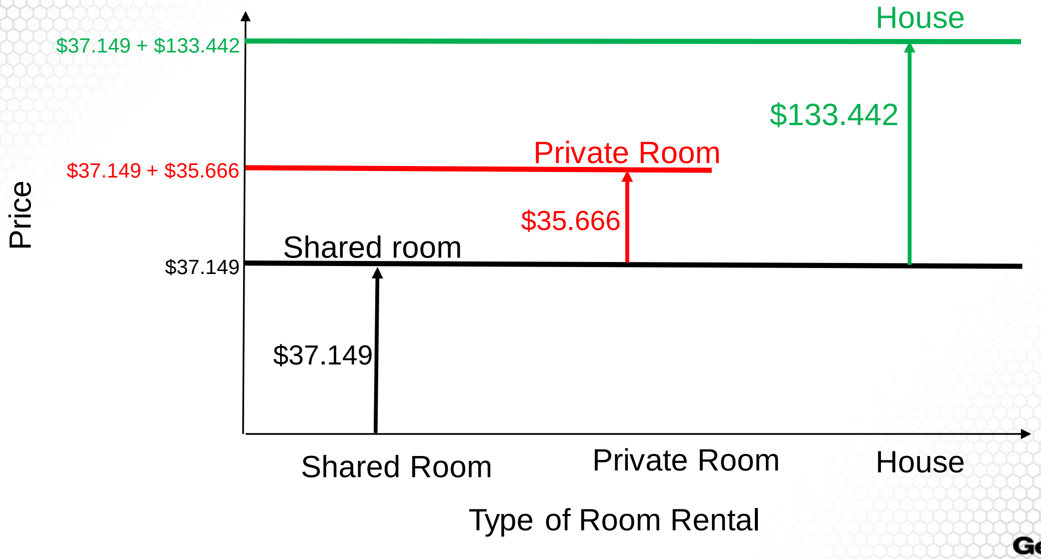
	Estimate	S.E.	t Value	Pr>|t|
Intercept	37.149	2.954	12.58***	<.001
Private_ind	35.666	3.123	11.42***	<.001
House-ind	133.442	3.058	43.64***	<.001

- Which room type's Average Price is $37.149? **Shared room**
- What is the Average Price of a Private Room? **$37.149 + $35.666**
- What is the Average Price of an Entire House? **$37.149 + $133.442**



Graphically

Let's take a look at this graphically…



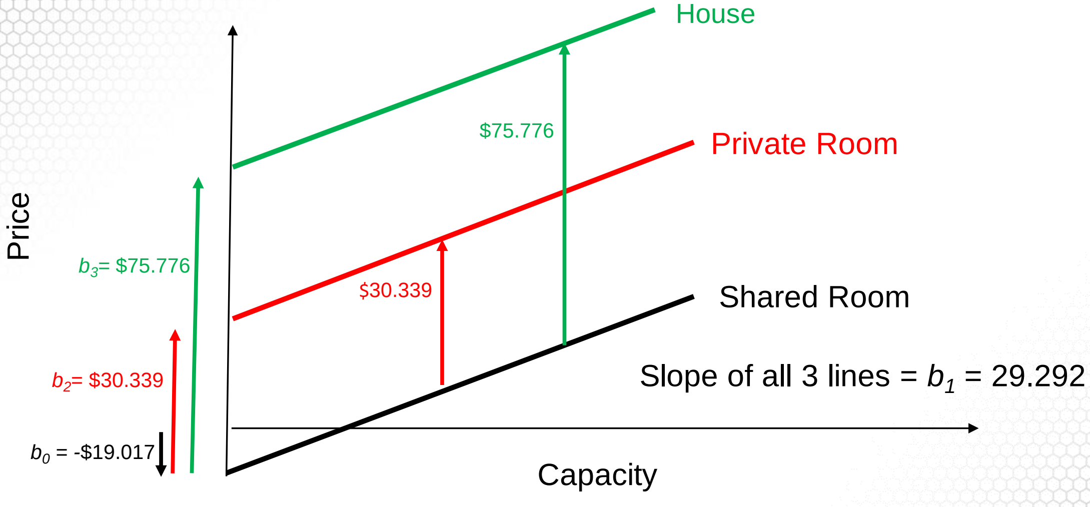
2DR2: 2nd Regression with Capacity and Dummy Variables

	Estimate	S.E.	t Value	Pr>|t|
Intercept	-19.017	2.678	-7.101	<.001
Capacity	29.292	0.355	82.605	<.001
Private_ind	30.339	2.739	11.076	<.001
House-ind	75.776	2.771	27.346	<.001

- *Price* = $b_0 + b_1$*Capacity + b_2*Private_ind + b_3*House_ind
- What is the (average) increase in Price for each additional individual? **Answer: $29.292**

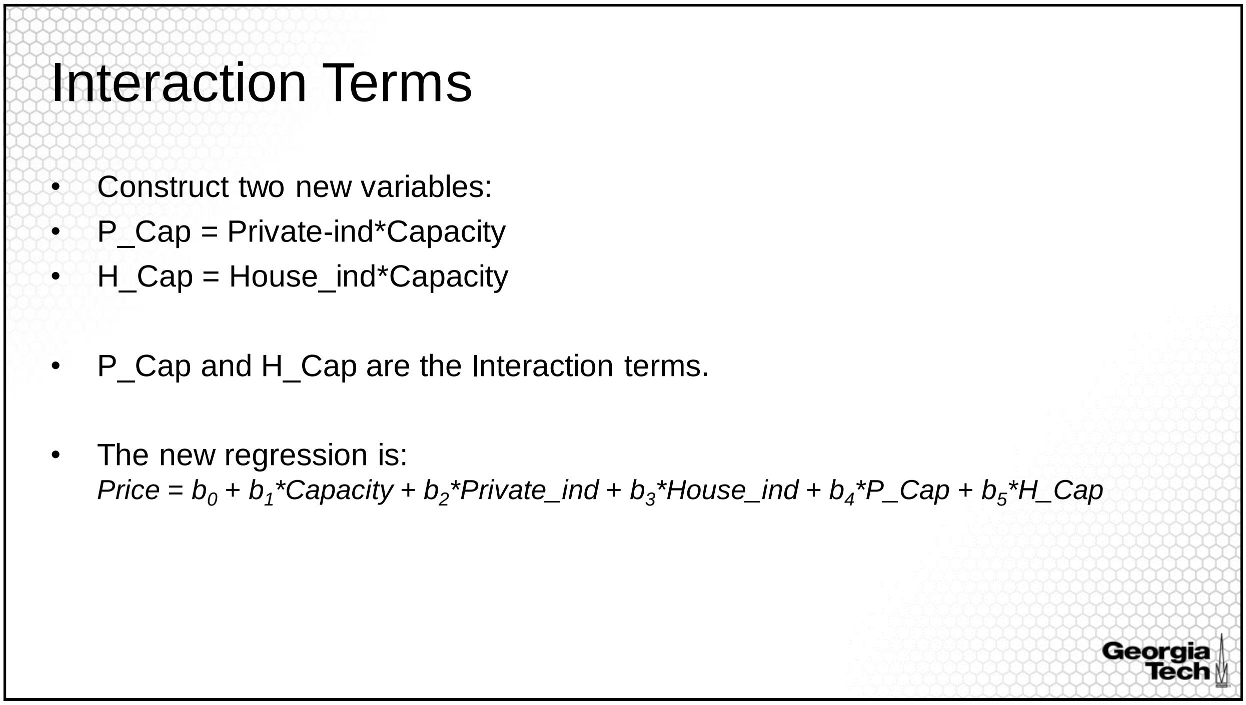
Graphically

Let's take a look at *Price* = b_0 + b_1**Capacity* + b_2**Private_ind* + b_3**House_ind* graphically…



Slope of all 3 lines = b_1 = 29.292

Price* = b_0 + b_1**Capacity* + b_2**Private_ind* + b_3**House_ind

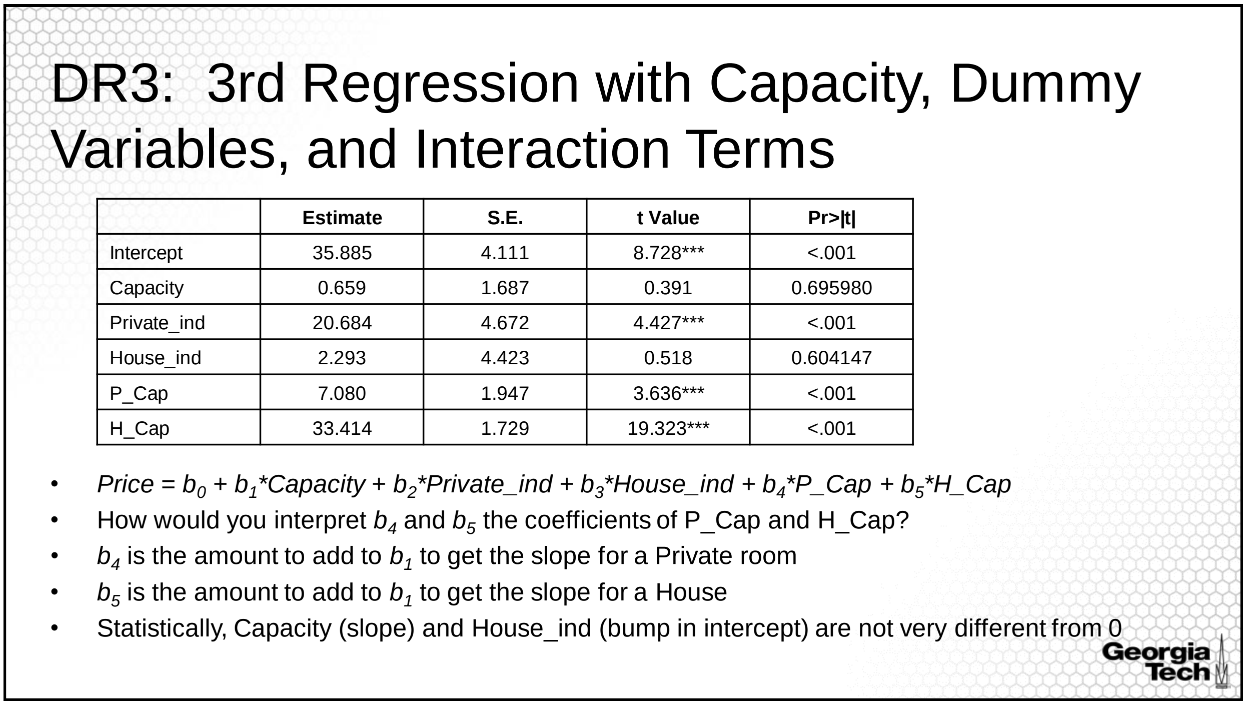


Interaction Terms

- Construct two new variables:
- P_Cap = Private-ind*Capacity
- H_Cap = House_ind*Capacity

- P_Cap and H_Cap are the Interaction terms.

- The new regression is:
  Price = b_0 + b_1*Capacity + b_2*Private_ind + b_3*House_ind + b_4*P_Cap + b_5*H_Cap



DR3: 3rd Regression with Capacity, Dummy Variables, and Interaction Terms

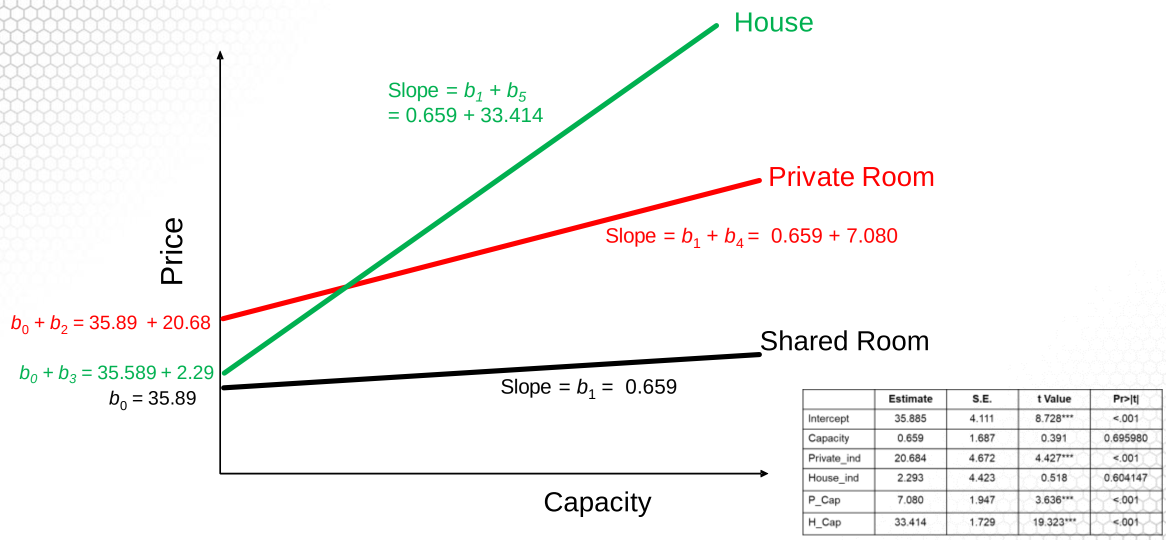
	Estimate	S.E.	t Value	Pr>|t|
Intercept	35.885	4.111	8.728***	<.001
Capacity	0.659	1.687	0.391	0.695980
Private_ind	20.684	4.672	4.427***	<.001
House_ind	2.293	4.423	0.518	0.604147
P_Cap	7.080	1.947	3.636***	<.001
H_Cap	33.414	1.729	19.323***	<.001

- *Price* = b_0 + b_1*Capacity + b_2*Private_ind + b_3*House_ind + b_4*P_Cap + b_5*H_Cap
- How would you interpret b_4 and b_5 the coefficients of P_Cap and H_Cap?
- b_4 is the amount to add to b_1 to get the slope for a Private room
- b_5 is the amount to add to b_1 to get the slope for a House
- Statistically, Capacity (slope) and House_ind (bump in intercept) are not very different from 0

Graphically

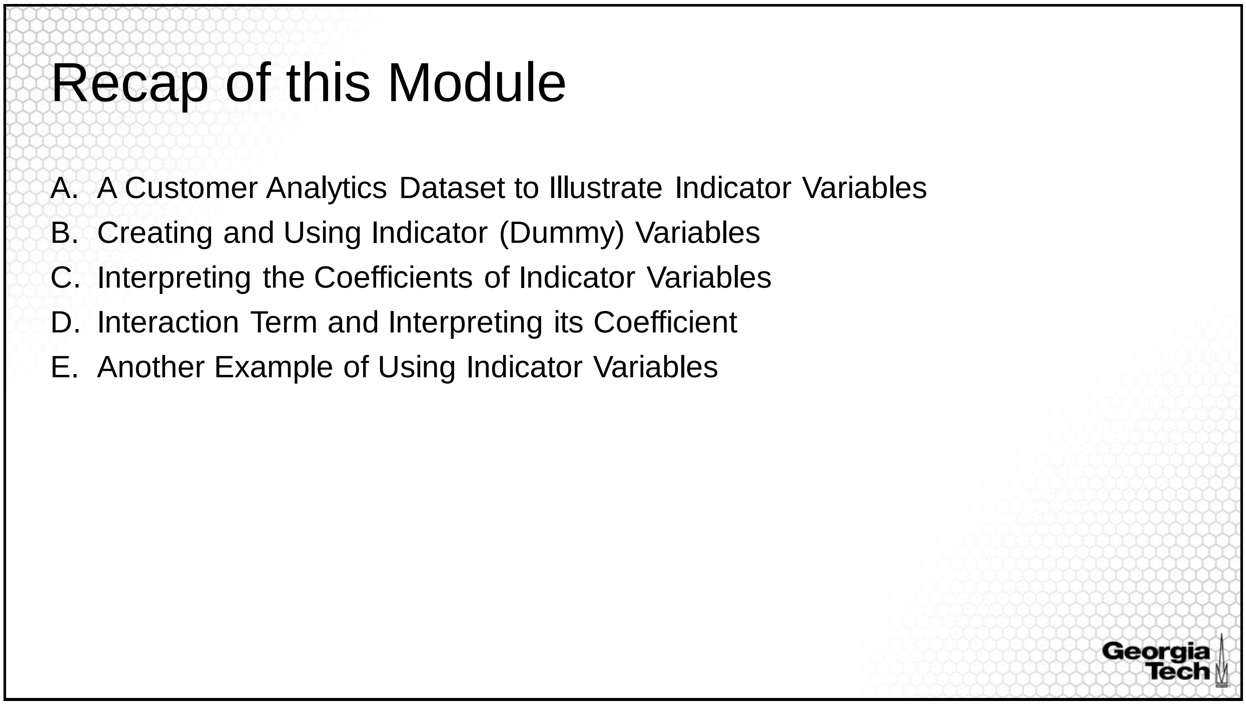
Let’s take a look at

Price = b_0 + b_1**Capacity* + b_2**Private_ind* + b_3**House_ind* + b_4**P_Cap* + b_5**H_Cap*
graphically…



	Estimate	S.E.	t Value	Pr>|t|
Intercept	35.885	4.111	8.728***	<.001
Capacity	0.659	1.687	0.391	0.695980
Private_ind	20.684	4.672	4.427***	<.001
House_ind	2.293	4.423	0.518	0.604147
P_Cap	7.080	1.947	3.636***	<.001
H_Cap	33.414	1.729	19.323***	<.001

Price = b_0 + b_1*Capacity + b_2*Private_ind + b_3*House_ind + b_4*P_Cap + b_5*H_Cap



Recap of this Module

A. A Customer Analytics Dataset to Illustrate Indicator Variables
B. Creating and Using Indicator (Dummy) Variables
C. Interpreting the Coefficients of Indicator Variables
D. Interaction Term and Interpreting its Coefficient
E. Another Example of Using Indicator Variables