Data Analytics for Business

Indicator Variables and Interaction Terms

Sridhar Narasimhan, Ph.D
Professor
Co-Director, Business Analytics Center
Scheller College of Business

A Dataset to Illustrate Indicator Variables

GTx

Lessons in this Module

A. A Dataset to Illustrate Indicator Variables
B. Creating and Using Indicator (Dummy) Variables
C. Interpreting the Coefficients of Indicator Variables
D. Interaction Terms and Interpreting Their Coefficients
E. Another Example of Using Indicator Variables

GTx

A Synthetic Dataset

Read the file “EDSAL.csv” into a dataframe called *edsal*

```
edsal <- read_csv("EDSAL.csv", col_types = list(
    Education = col_factor(c("HS", "UG", "GRAD")),
    Experience = col_integer(),
    Salary = col_double()))
```

```
> str(edsal)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':     300 obs. of 6 variables:
```

$ Education: Factor w/ 3 levels "HS","UG","GRAD“ … (*highest educational level*)
$ Experience: int 2 14 36 16 36 33 36 8 3 21 ... (*in years of work experience*)
$ Salary: num 34.4 59.2 113.3 69.1 106 ... (*in thousands of $)*

GTx

First 10 Rows of the *edsal* Dataframe

Education	Experience	Salary
HS	2	34.432
HS	14	59.157
HS	36	113.270
HS	16	69.147
HS	36	106.016
HS	33	58.634
HS	36	101.081
HS	8	51.329
HS	3	58.197
HS	21	77.964

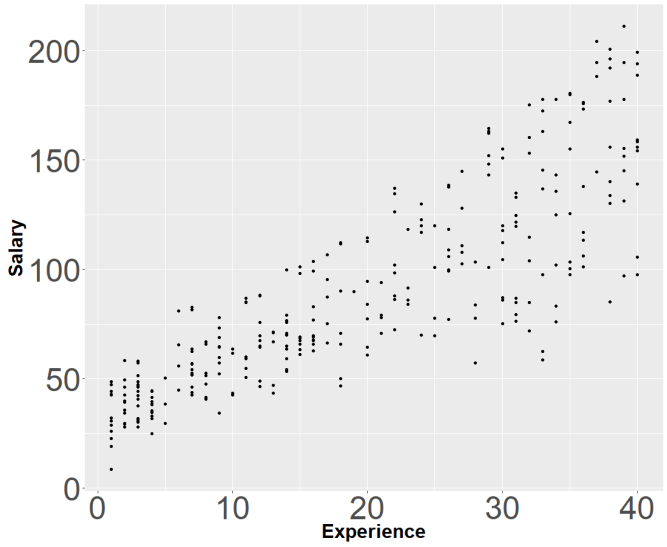
GTx

Exploring the *edsal* Dataset

- We would like to understand better the reasons why some individuals have larger salaries than others.
- In particular, we would like to investigate whether work experience has an influence on salaries.
- There is the *education* variable in this dataset.
- So, how do we get started?
- To keep matters clear, we start with
  Salary, Experience

GTx

Scatterplot



GTx

RS: Simple Regression

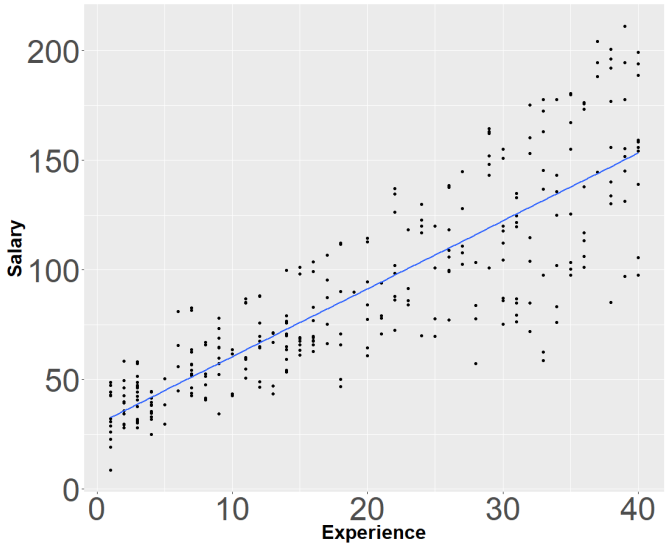
	Estimate	S.E.	t Value	Pr>ltl
Intercept	29.4679	2.5673	11.48 ***	<.001
Experience	3.0959	0.1113	27.81 ***	<.001

R-squared	Adjusted R-squared
0.722	0.721

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

GTx

Scatterplot with Regression Line



GTx

Data Analytics for Business

Indicator Variables and Interaction Terms

Sridhar Narasimhan, Ph.D
Professor
Co-Director, Business Analytics Center
Scheller College of Business

Creating and Using Indicator (Dummy) Variables

GTx

What to do About the Categorical Variable Education?

- Is this categorical variable important?
- Does having an undergraduate or graduate degree potentially have an effect on Salary compared to only studying through high school?
- How do we include this variable in a regression model that requires numeric values?

Education	Experience	Salary
HS	2	34.432
HS	14	59.157
HS	36	113.270
HS	16	69.147
HS	36	106.016
HS	33	58.634
HS	36	101.081
HS	8	51.329
HS	3	58.197
HS	21	77.964

GTx

Doing Regression with Qualitative Predictor Variable

- Consider the variable ***Education***
- We want to investigate the effect of Education on Salary. Note that Education is a qualitative (or categorical) variable with three possible values: HS, UG, or GRAD
- We need to quantify this variable

Education
HS
UG
GRAD
UG
UG
GRAD
GRAD
HS
HS
GRAD

GTx

Creating Indicator (Dummy) variables

- Since we have three possible value for Education, we need to create two indicator variables.
- The base (or reference) case, with both dummy variables set to 0, is Education = UG. 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:

$$\text{HS} = \begin{cases} 1, & if\ Education = HS \\ 0, & otherwise \end{cases}$$

$$\text{Graduate} = \begin{cases} 1, & if\ Education = GRAD \\ 0, & otherwise \end{cases}$$

GTx

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

HS = $\begin{cases} 1, & if\ Education = HS \\ 0, & otherwise \end{cases}$

Graduate = $\begin{cases} 1, & if\ Education = GRAD \\ 0, & otherwise \end{cases}$

We, then run the regression,
Salary = $b_0 + b_1$*HS + b_2*Graduate

Education	HS	Graduate
HS	1	0
UG	0	0
GRAD	0	1
UG	0	0
UG	0	0
GRAD	0	1
GRAD	0	1
HS	1	0
HS	1	0
GRAD	0	1

GTx

Quiz

With this Indicator variables coding scheme,

HS = $\begin{cases} 1, & if\ Education = HS \\ 0, & otherwise \end{cases}$ Graduate = $\begin{cases} 1, & if\ Education = GRAD \\ 0, & otherwise \end{cases}$

- Can a record in the edsal dataframe have this value (HS = 0, Graduate = 0)?
  Answer: **YES**, because this record is for someone with an UG degree, i.e., the base case.
- Can a record in the *edsal* dataframe have this value (HS = 1, Graduate = 1)?
  Answer: **NO**, no individual can have their highest level of education be both HS and graduate degree!

GTx

Data Analytics for Business

Indicator Variables and Interaction Terms

Sridhar Narasimhan, Ph.D
Professor
Co-Director, Business Analytics Center
Scheller College of Business

Interpreting the Coefficients of Indicator Variables

GTx

A Linear Model With Indicator Variables

$$\text{HS} = \begin{cases} 1, & if\ Education = HS \\ 0, & otherwise \end{cases} \qquad \text{Graduate} = \begin{cases} 1, & if\ Education = GRAD \\ 0, & otherwise \end{cases}$$

With this Indicator variables coding scheme, we build an initial linear regression model:

Salary = $b_0 + b_1$**HS* + b_2**Graduate*

We then fit it using the data in *edsal*.

GTx

DR1: 1st Regression with Dummy Variable

	Estimate	S.E.	t Value	Pr>ltl
Intercept	87.923	4.062	21.645***	<.001
HS	-22.487	5.759	-3.905 ***	<.001
Graduate	27.518	5.774	4.766 ***	<.001

- $Salary = b_0 + b_1*HS + b_2*Graduate$
- Which group's Average Salary is $87,923: High School, Undergraduate, or Graduate?
- Correct Answer: With *HS* = 0 and *Graduate* = 0, b_0 captures the average salary of UG (base case)

GTx

$Salary = b_0 + b_1*HS + b_2*Graduate$

	Estimate	S.E.	t Value	Pr>ltl
Intercept	87.923	4.062	21.645***	<.001
HS	-22.487	5.759	-3.905 ***	<.001
Graduate	27.518	5.774	4.766 ***	<.001

- What is the Average Salary for someone with only a HS diploma?
- This individual has *HS* = 1 and *Graduate* = 0, so $b_0 + b_1$ captures the average Salary for folks with a HS diploma.
- $87,923 - $22,487 = $65,436
- So, $22,487 is the decrease in salary (on average) for HS compared to UG.

GTx

$Salary = b_0 + b_1{*}HS + b_2{*}Graduate$

	Estimate	S.E.	t Value	Pr>ltl
Intercept	87.923	4.062	21.645***	<.001
HS	-22.487	5.759	-3.905 ***	<.001
Graduate	27.518	5.774	4.766 ***	<.001

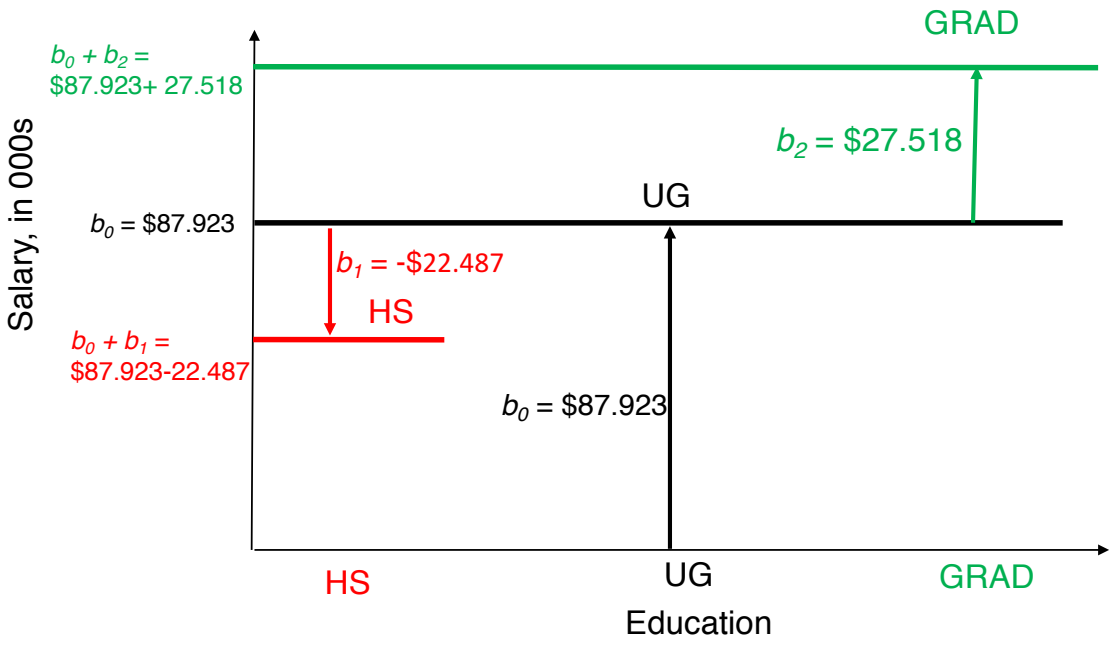
- What is the Average Salary for someone with a Graduate degree?
- This individual has *HS* = 0 and Graduate = 1, so $b_0 + b_2$ captures the average Graduate Salary.
- $87,923 + $27,518 = $115,441
- So, $27,518 is the additional salary (on average) for GRAD over UG education.

GTx

Graphically

Let's take a look at this graphically…

GTx



GTx

Important Note

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

lm(Salary ~ Education, data=edsal)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**65.436**	**4.082**	**16.029**	**< 2e-16 *****
EducationUG	**22.487**	**5.759**	**3.905**	**0.000117 *****
EducationGRAD	**50.006**	**5.788**	**8.640**	**3.54e-16 *****

- What is the base case? What is the average salary for the base case? The Base Case is HS with Average Salary = $65,436
- What is the Average Salary of UG? $65,436 + 22,487 = $87,923
- What is the Average Salary of GRAD? $65,436 + 50,006 = $115,442
- All three groups have the same answers as our coding scheme where UG was the base case!!!

GTx

R's Indicator variable coding

R's indicator variable coding scheme can be found by using

contrasts(edsal$Education)

	UG	GRAD	
HS	0	0	(HS is the base case in this coding scheme).
UG	1	0	
GRAD	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!

GTx

DR2: 2nd Regression with EXPERIENCE and Dummy Variables

	Estimate	S.E.	t Value	Pr>|t|
Intercept	28.359	2.348	12.079	<.001
Experience	2.971	0.080	37.084	<.001
HS	-15.062	2.436	-6.183	<.001
Graduate	25.925	2.434	10.650	<.001

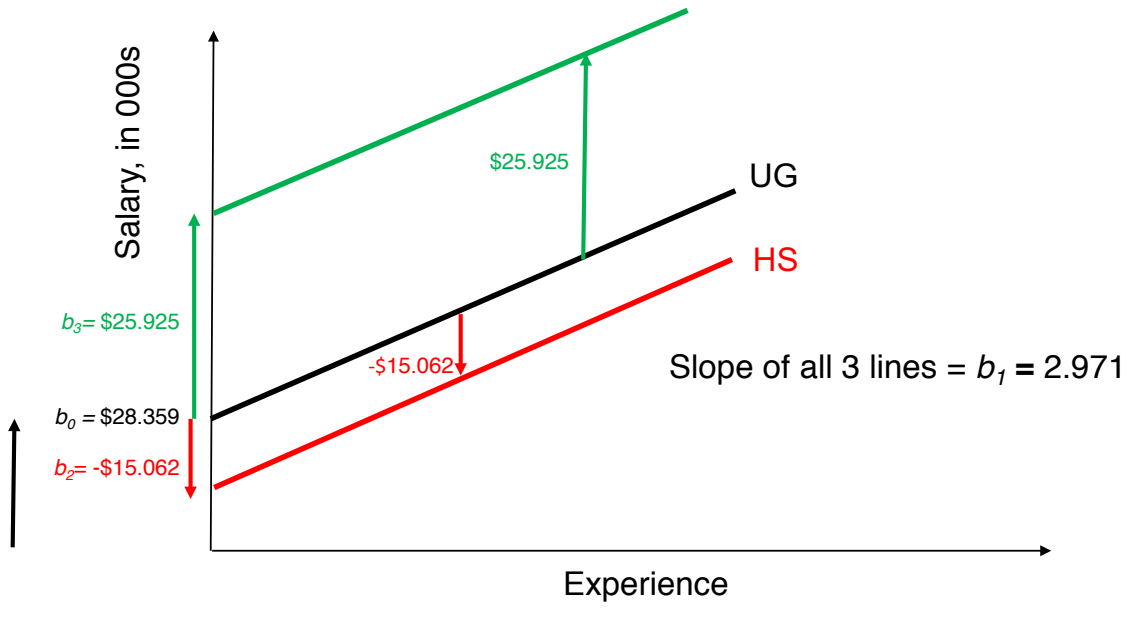
- $Salary = b_0 + b_1*Experience + b_2*HS + b_3*Graduate$
- What is the (average) increase in Salary for a one year increase in Experience?
- Answer: $2,971

GTx

Graphically

Graphically, *SALARY = b_0 + b_1*EXPERIENCE + b_2*HS + b_3*Graduate would look like this…*

GTx



GTx

Quiz

	Estimate	S.E.	t Value	Pr>|t|
Intercept	28.359	2.348	12.079	<.001
Experience	2.971	0.080	37.084	<.001
HS	-15.062	2.436	-6.183	<.001
Graduate	25.925	2.434	10.650	<.001

Salary = $b_0 + b_1$**Experience* + b_2**HS* + b_3**Graduate*

- What is the Base case in this Model?

  Answer: The base case has HS = 0, Graduate = 0, i.e., individuals with an Undergrad degree
- What is the salary (on average) for an individual with an Undergraduate Degree and has NO work experience (Experience = 0, HS = 0, Graduate = 0)

  Answer: $28,359

GTx

Data Analytics for Business

Indicator Variables and Interaction Terms

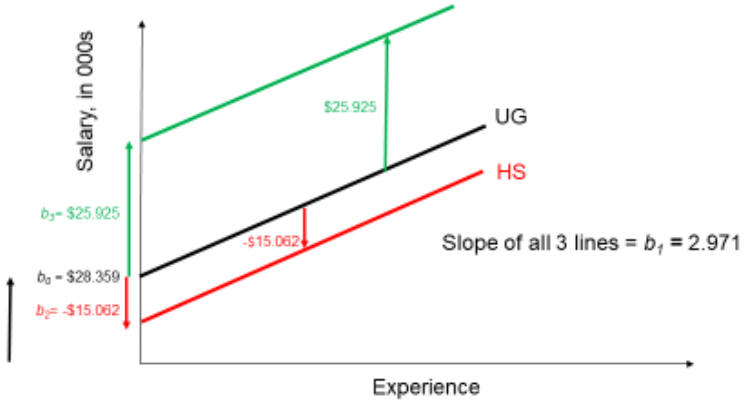
Sridhar Narasimhan, Ph.D
Professor
Co-Director, Business Analytics Center
Scheller College of Business

Interaction Terms and Interpreting Their Coefficients

GTx

*Salary = b_0 + b_1*Experience + b_2*HS + b_3*Graduate*

In this model, we had the same slope for all three values of Education, the Categorical Variable,

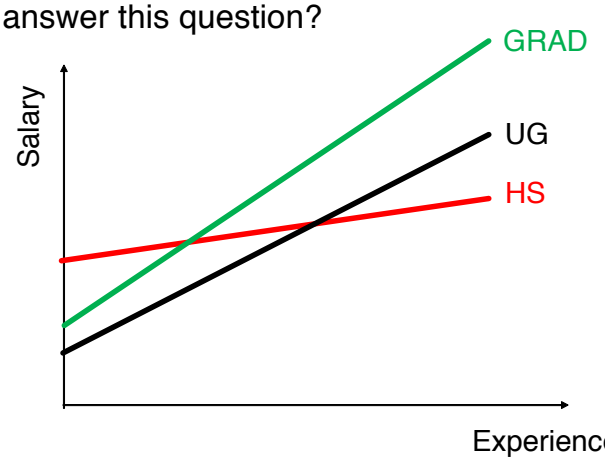


GTx

GTx

Interaction Terms

- But, do we really know that the salary for HS and GRAD increase with experience at the same rate as for UG (the base case)?
- How can regression help answer this question?



GTx

Interaction Terms

- Construct two new variables:
  - H_Exp = HS*Experience
  - G_Exp = Graduate*Experience
- H_Exp and G_Exp are the Interaction terms.
- The new regression is:
  $Salary = b_0 + b_1*Experience + b_2*HS + b_3*Graduate + b_4*H\_Exp + b_5*G\_Exp$

GTx

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

	Estimate	S.E.	t Value	Pr>|t|
Intercept	27.34	1.92	14.27 ***	<.001
Experience	3.022	0.08	37.38 ***	<.001
HS	11.48	2.64	4.35***	<.001
Graduate	0.30	2.76	0.11	0.912
H_Exp	-1.51	0.12	-12.80 ***	<.001
G_Exp	1.24	0.12	10.78 ***	<.001

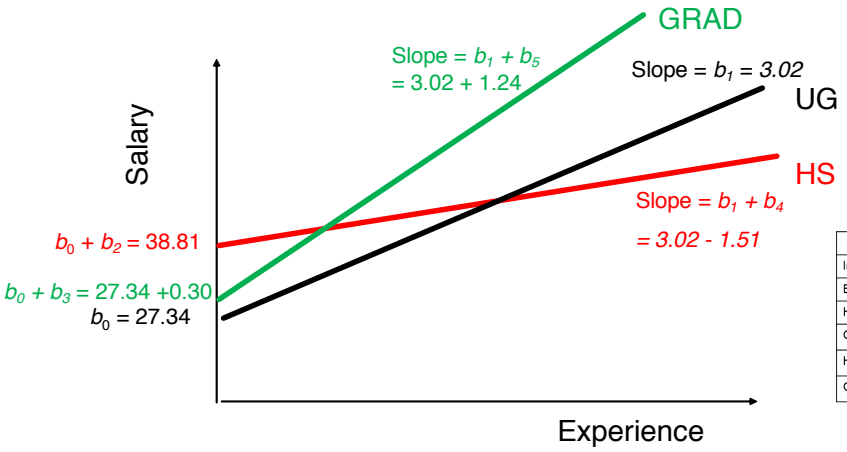
$Salary = b_0 + b_1*Experience + b_2*HS + b_3*Graduate + b_4*H\_Exp + b_5*G\_Exp$

- How would you interpret b_4 and b_5, the coefficients of H_Exp and G_Exp?
- b_4 is the amount to add to b_1 to get the slope for individuals with a HS diploma
- b_5 is the amount to add to b_1 to get the slope for individuals with a Graduate degree

GTx

Graphically

$$Salary = b_0 + b_1*Experience + b_2*HS + b_3*Graduate + b_4*H\_Exp + b_5*G\_Exp$$



	Estimate	S.E.	t Value	Pr>|t|
Intercept	27.34	1.92	14.27 ***	<.001
Experience	3.022	0.08	37.38 ***	<.001
HS	11.48	2.64	4.35***	<.001
Graduate	0.30	2.76	0.11	0.912
H_Exp	-1.51	0.12	-12.80 ***	<.001
G_Exp	1.24	0.12	10.78 ***	<.001

GTx

Test Your Understanding

$$\textbf{Salary} = b_0 + b_1*\textbf{Experience} + b_2*\textbf{HS} + b_3*\textbf{Graduate} + b_4*\textbf{H\_Exp} + b_5*\textbf{G\_Exp}$$

If the experience of an individual with an UG education increases by 10 years, what is the predicted <u>increase</u> in Salary for that individual?

- For this individual (i.e., the baseline case) HS=0, Graduate = 0, thus the relevant slope is b_1 = 3.02.
- Hence, the increase in Salary (on average) for this individual = \$3.02K*10 = \$30.2K

If the experience of an individual with a HS education increases by 10 years, what is the predicted <u>increase</u> in Salary for that individual?

- For the HS case, HS = 1, Graduate = 0, hence the relevant slope is $b_1 + b_4$ = 3.02 - 1.51 = \$1.51
- Hence, the increase in Salary (on average) for this individual = \$1.51K*10 = \$15.1K

GTx

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.

GTx

Quiz

$$Salary = b_0 + b_1*Experience + b_2*HS + b_3*Graduate + b_4*H\_Exp + b_5*G\_Exp$$

- If the experience of an individual with an Graduate education increases by 10 years, what is the predicted increase in Salary for that individual?

  Answer:
  - For this case, Graduate = 1, HS= 0, hence the relevant slope is $b_1 + b_5 =$ 3.02 +1.24 = \$4.26
  - Hence, the increase in Salary (on average) for this individual = \$4.26K*10 = \$42.6K

GTx

Data Analytics for Business

Indicator Variables and Interaction Terms

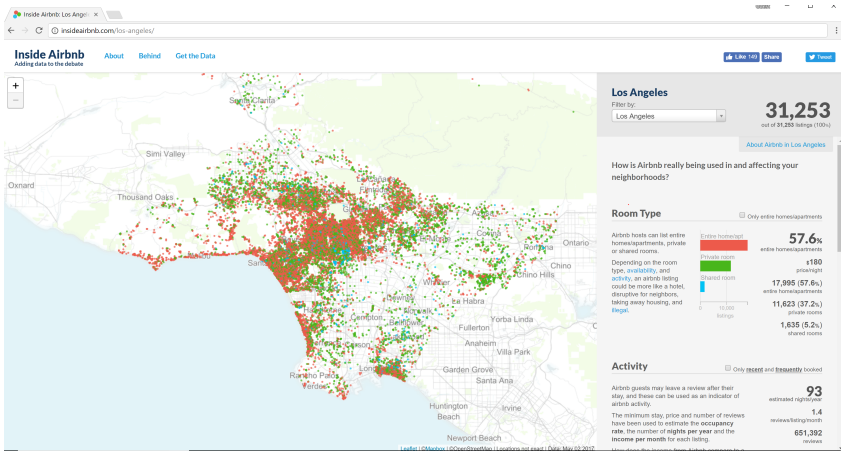
Sridhar Narasimhan, Ph.D
Professor
Co-Director, Business Analytics Center
Scheller College of Business

Another Example of Using Indicator Variables

GTx

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 .



GTx

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 ...
```

GTx

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?

GTx

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

GTx

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

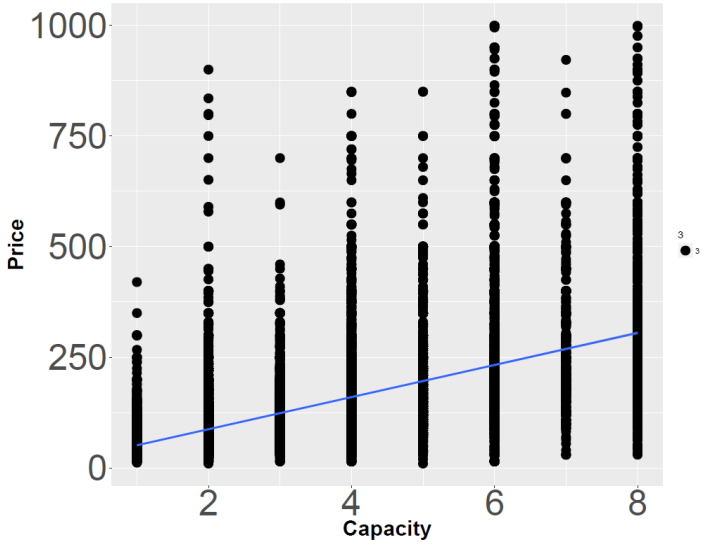
- $Price = b_0 + b_1 * Capacity$

	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

GTx

Scatterplot with Regression Line



GTx

Creating Indicator (Dummy) variables

- We define two dummy variables:

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

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

- 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.

GTx

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>ltl
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

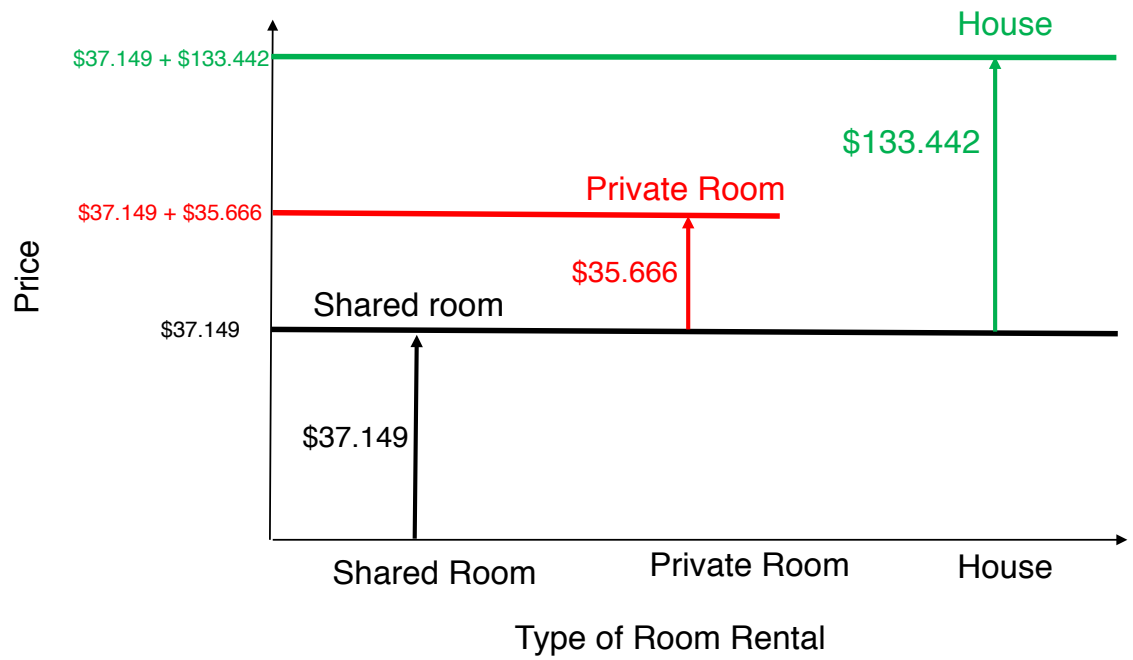
- Which room type's Average Price is $37.149? **Shared room**
- What is the Average Price of a Private Room? **$37.149 + $35.666**
- What is the Average Price of an Entire House? **$37.149 + $133.442**

GTx

Graphically

Let's take a look at this graphically…

GTx



GTx

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

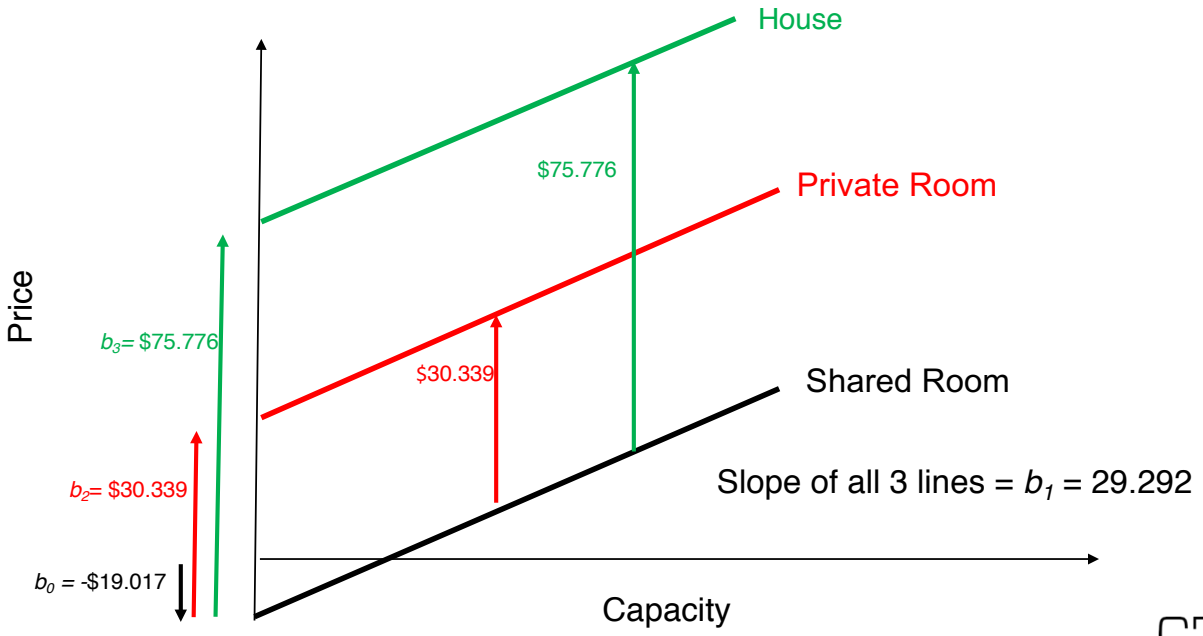
GTx

Graphically

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

GTx

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



GTx

Interaction Term

- 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:
  $Price = b_0 + b_1*Capacity + b_2*Private\_ind + b_3*House\_ind + b_4*P\_Cap + b_5*H\_Cap$

GTx

DR3: 3rd Regression with Experience, 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 room
- Statistically, Capacity (slope) and House_ind (bump in intercept) are not very different from 0

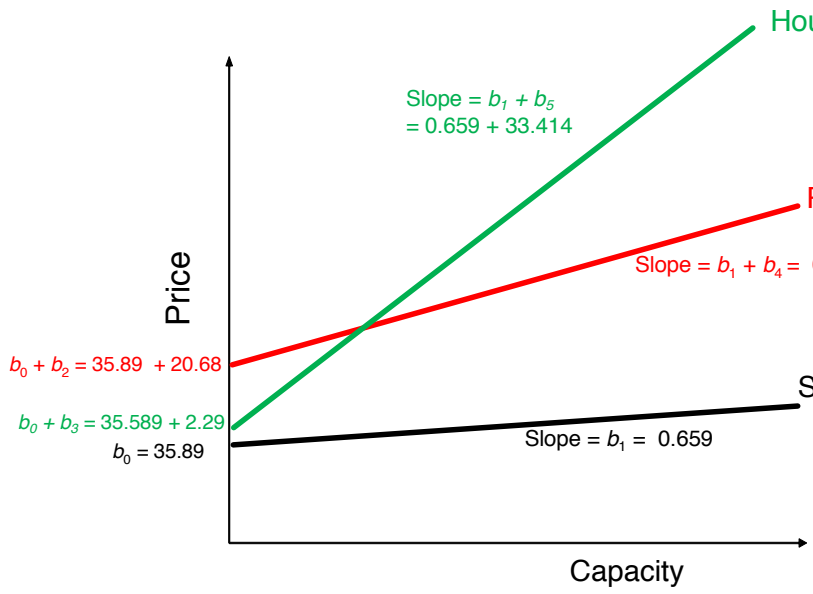
GTx

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…

GTx

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



	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

GTx

Recap of this Module

A. A Dataset to Illustrate Indicator Variables
B. Creating and Using Indicator (Dummy) Variables
C. Interpreting the Coefficients of Indicator Variables
D. Interaction Terms and Interpreting Their Coefficients
E. Another Example of Using Indicator Variables

GTx

Data Analytics for Business

Indicator Variables and Interaction Terms

Sridhar Narasimhan, Ph.D
Professor
Co-Director, Business Analytics Center
Scheller College of Business

End

GTx