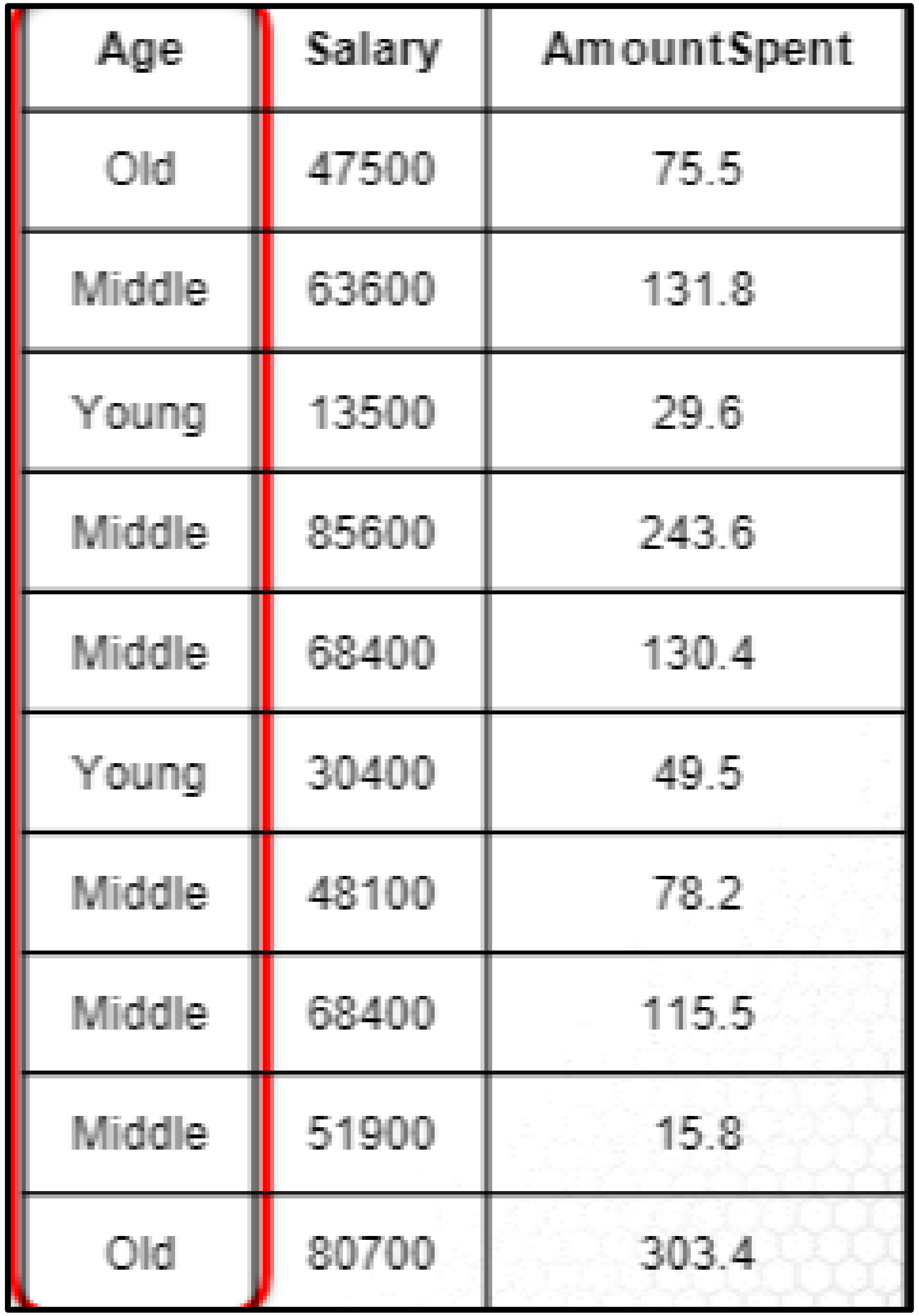
Week 2 TA OH-Indicator and Interaction Terms

MGT 6203 Spring 2021

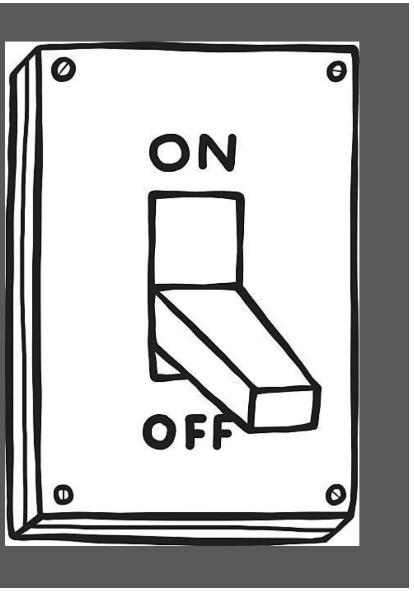
Indicator/Dummy Variables Definition

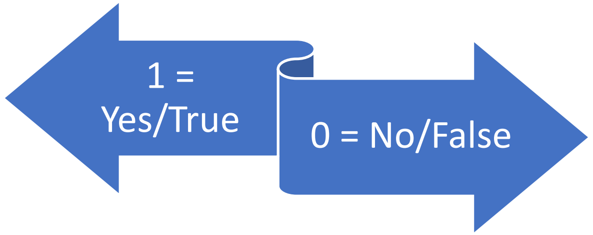
- Used in regression when we have a categorical variable (qualitative not quantitative) that we want to measure (e.g., Age w/ 3 categories)
- Create 1 less than the number of categories
- R will automatically create dummy variables but then you have no decision over the base case



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

Indicator/Dummy Variables Assignment





1 = Yes/True

0 = No/False

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

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

- Variables can't have more than one True (1) dummy variable
- The base case/intercept is when every other category = 0/False
- In the example, the base case is young when AgeMid = 0 and AgeOld = 0, leaving only the young category remaining

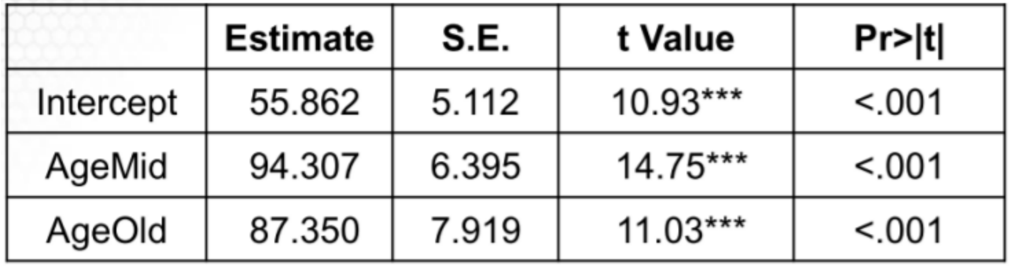
	AgeMid	**AgeOld**
Intercept/Base/Young	0	0
Middle	1	0
Old	0	1

Simple Indicator/Dummy Variables Equation

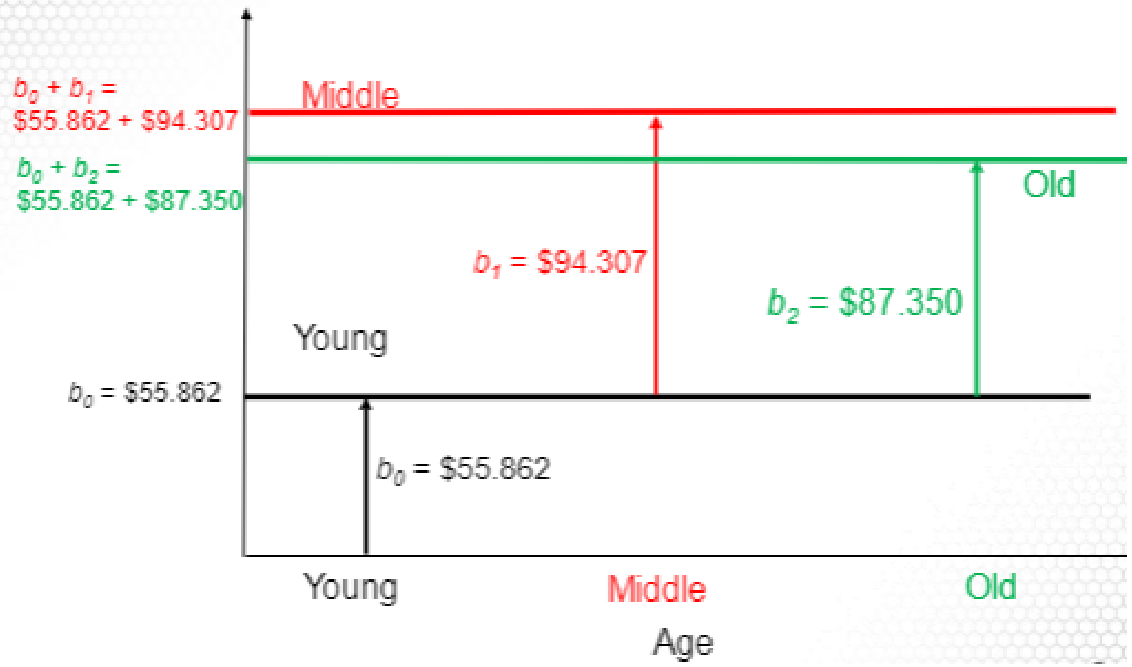
$$AmountSpent = b_0 + b_1 * AgeMid + b_2 * AgeOld$$

- Young AmountSpent = $b_0 + (b_1 * 0) + (b_2 * 0)$ = b_0 = 55.862
- Middle AmountSpent= $b_0 + (b_1 * 1) + (b_2 * 0) = b_0 + b_1$= 55.862 + 94.307 = 150.169
- Old AmountSpent = $b_0 + (b_1 * 0) + (b_2 * 1) = b_0 + b_2$ = 55.862+87.350 = 143.212

- On avg., Middle-aged spend 94.307 more than the Young, and Old spend 87.350 more than the Young

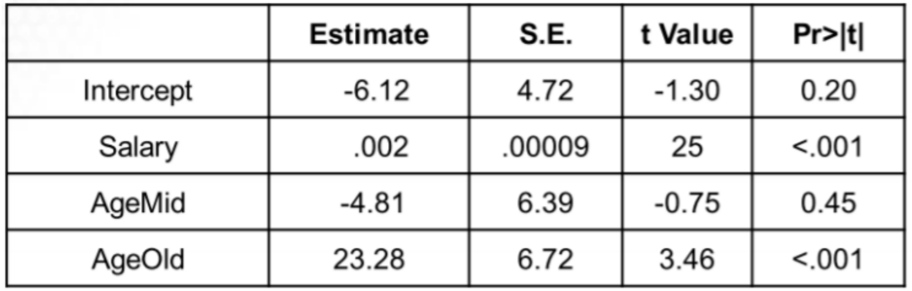


	Estimate	S.E.	t Value	Pr>|t|
Intercept	55.862	5.112	10.93***	<.001
AgeMid	94.307	6.395	14.75***	<.001
AgeOld	87.350	7.919	11.03***	<.001



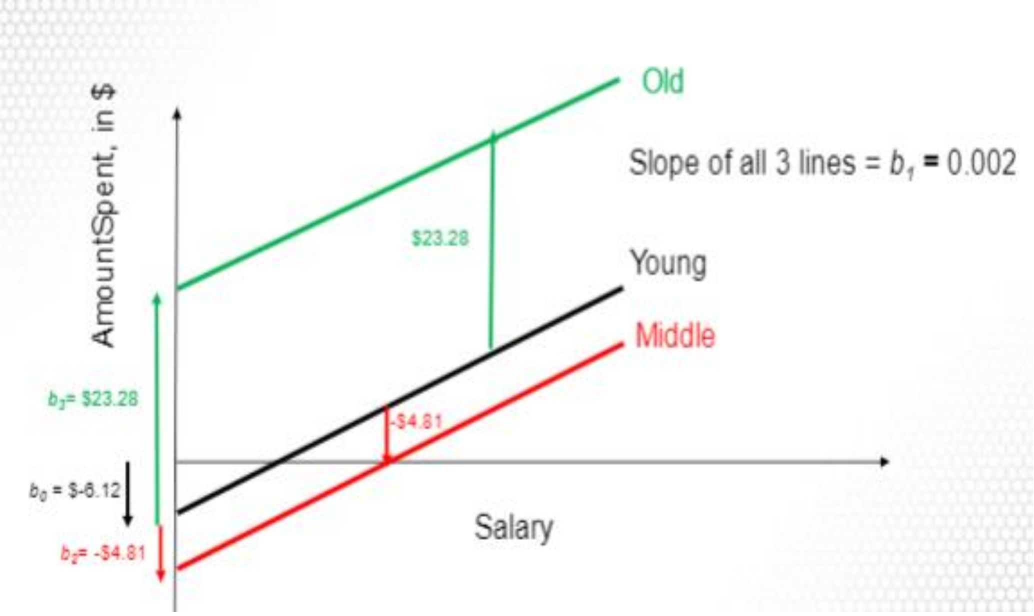
Indicator/Dummy Variables Equation with Quantitative Variable

$$AmountSpent = b_0 + b_1 * Salary + b_2 * AgeMid + b_3 * AgeOld$$



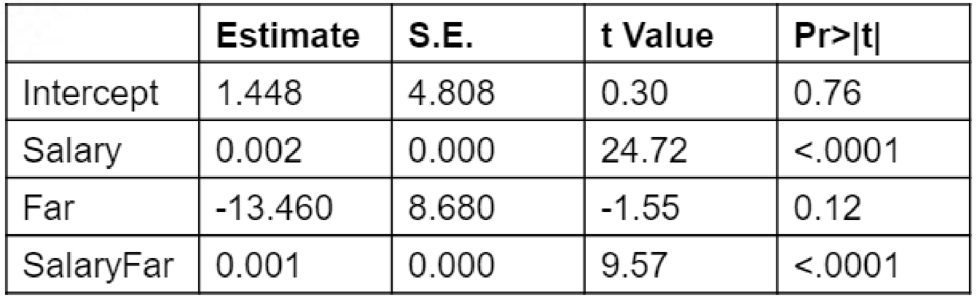
	Estimate	S.E.	t Value	Pr>|t|
Intercept	-6.12	4.72	-1.30	0.20
Salary	.002	.00009	25	<.001
AgeMid	-4.81	6.39	-0.75	0.45
AgeOld	23.28	6.72	3.46	<.001

- Interpretation of Salary: for one unit increase in salary, the average AmountSpent increases by $.002, all else constant
- If Salary = $100,000
- Young AmountSpent = $b_0 + (b_1 * Salary) + (b_2 * 0) + (b_3 * 0) = b_0 + (b_1 * Salary)$ = (-6.12) + (.002 * Salary).
  - (-6.12) + 200 = 193.88
- Middle AmountSpent = $b_0 + (b_1 * Salary) + (b_2 * 1) + (b_3 * 0) = b_0 + (b_1 * Salary) + b_2$= (-6.12) + (.002 * Salary) + (-4.81).
  - (-6.12) + 200 + (-4.81) = 189.07
- Old AmountSpent = $b_0 + (b_1 * Salary) + (b_2 * 0) + (b_3 * 1) = b_0 + (b_1 * Salary) + b_3$= (-6.12) + (.002 * Salary) + 23.28.
  - (-6.12)+200 = 217.16
- Interpretation: On avg, Middle-aged spend 4.81 less and Old spend 23.28 more than Young, holding Salary constant



Interaction Term Definition

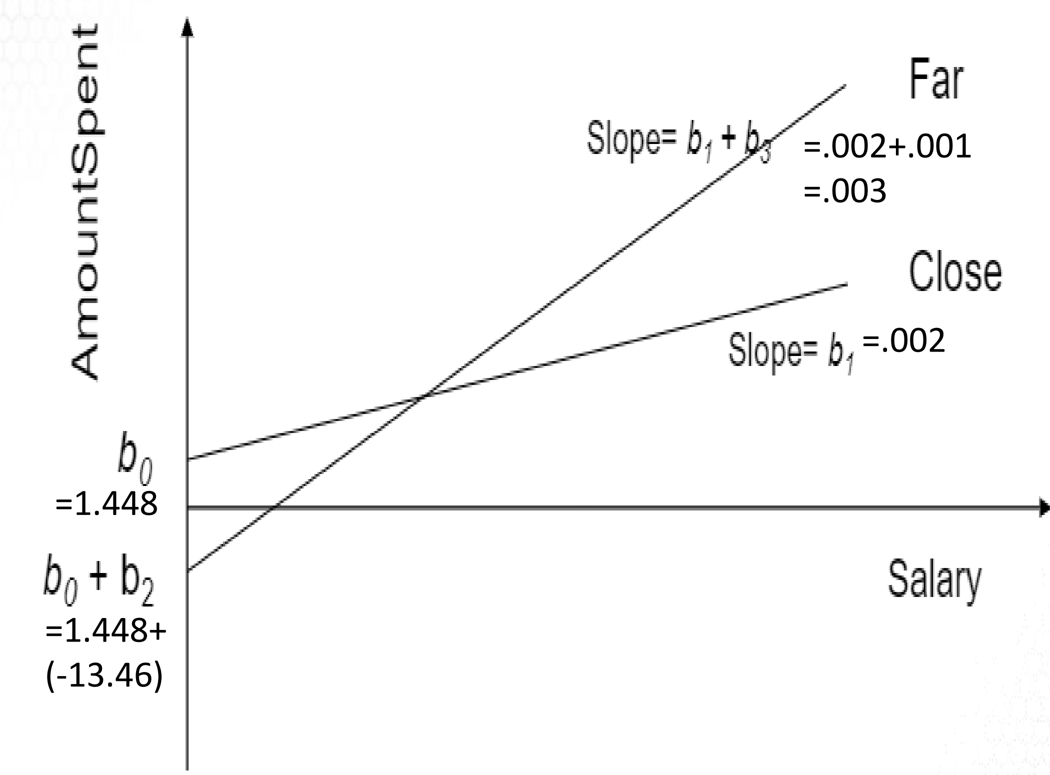
- Using a factor of multiple columns/variables in your dataset to gather further insight into your data
- Interpretation: b_3 the coefficient of the interaction term SalaryFar is the amount you add to b_1 to get the slope for people who are far away
- An increase of $10,000 in Salary for someone who lives Close = .002*10,000 = 20 increase in AmountSpent
- An increase of $10,000 in Salary for someone who lives Far = (.002+.001)*10,000 = 30 increase in AmountSpent



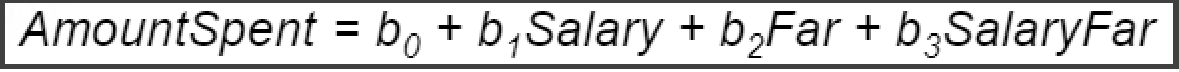
	Estimate	S.E.	t Value	Pr>|t|
Intercept	1.448	4.808	0.30	0.76
Salary	0.002	0.000	24.72	<.0001
Far	-13.460	8.680	-1.55	0.12
SalaryFar	0.001	0.000	9.57	<.0001

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

$$AmountSpent = b_0 + b_1Salary + b_2Far + b_3SalaryFar$$

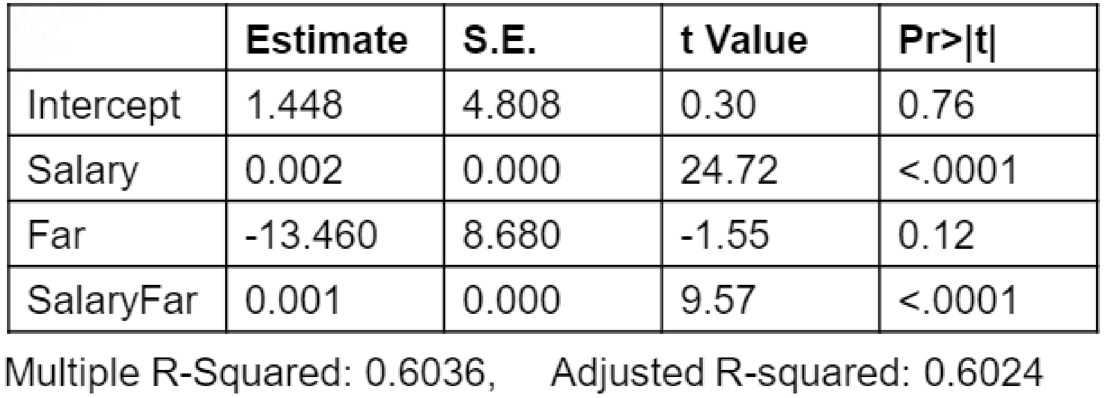


Interaction Term Equation



$AmountSpent = b_0 + b_1Salary + b_2Far + b_3SalaryFar$

- If Salary = $100,000
- Close AmountSpent = $b_0 + (b_1 * Salary) + (b_2 * 0) + (b_3 * 0 * Salary)$= $b_0 + (b_1 * Salary)$= 1.448 + 200 = 201.448
- Far AmountSpent = $b_0 + (b_1 * Salary) + (b_2 * 1) + (b_3 * 1 * Salary)$= $b_0 + b_2 + (b_1 + b_3) * Salary$= 1.448 + (-13.46)+(.002 + .001)*100,000 = 287.988



	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

Useful steps:

1. Read in the data
2. Inspect the data with head(), str(), and/or cor()
   i. If categorical variables aren’t factors: change them to factors
3. Create your own dummy variables or leave as is
   i. If left as is use contrasts() to see which one is the base case
4. Create any interaction terms
5. Fit model and summarize
6. Use plot(model) to check model assumption plots

Now-Let’s code. Useful functions to remember:

- **str**() – structure of dataframe. Will tell you which variables are factors/categories
- **as.factor**() – if categorical variable isn’t already assigned to factor data type
- model <- **lm**(Response~Explanatory, **data** = data) – R automatically creates dummy variables for factor variables if you don’t assign yourself
- **summary**(model) – get the results of your regression
- **contrasts**(data$Indicator) – check the dummy variables coding scheme
- data$new <- **ifelse**(data$old == “Yes”, 1, 0) – assign dummy variables
- **mutate**(data, data$new = data$old1 * data$old2) or just data$new = data$old1 * data$old2