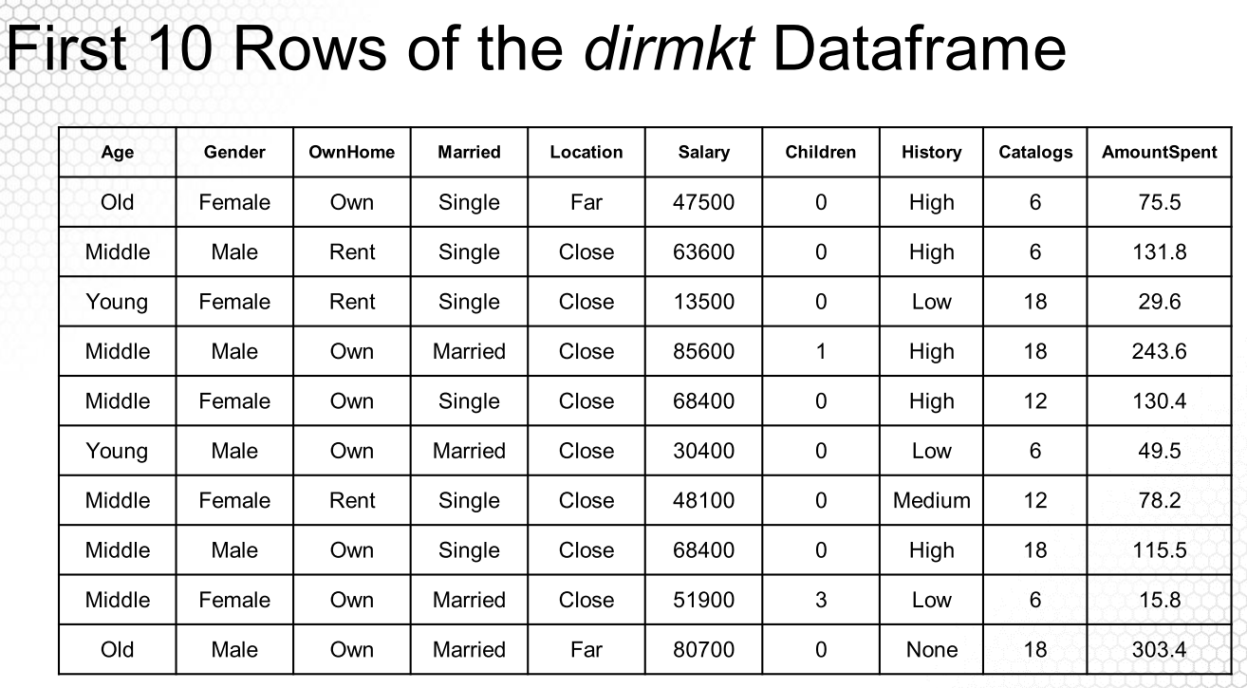
# Video 2.1 A Customer Analytics Dataset to Illustrate Indicator (Dummy) Variables

***\*\*My notes are in bullets whereas TA notes are in paragraph form\*\****

* The instructor used a simulated data set which mimics data from a direct marketing firm. The data set contains variables including indicator variables and numerical variables.
* **Indicator variable is also known as dummy variable**.
* We are trying to predict the amount customers spent on buying products using customer characteristics such as age (indicator variable), salary (numerical variable), location (indicator variable). A quick look of the dataset:
  + *Specifically, we want to understand why some individuals spend more than others*

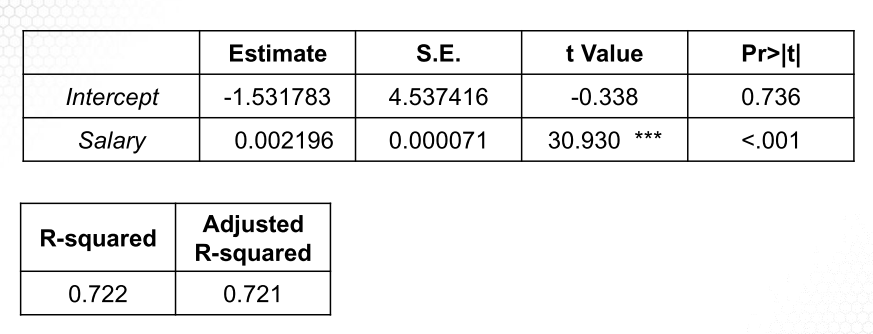


* As an example, we would like to investigate whether salary has an influence on AmountSpent. Shown below is the scatterplot of **Salary** to **AmountSpent** and the simple linear regression line.

A close up of text on a white background

Description automatically generated

* Note that the adjusted R-squared is the R-squared value adjusted for degree of freedom.



Video 2.2 Creating and Using Indicator (Dummy) Variables

* The question we are looking to solve in this segment is – How can we include categorical variables such as age in a model that requires numeric values?
* To setup **indicator** (aka dummy variables) you will need to create n-1 **dummy** **variables** and one **base** or r**eference case**.
* In our example we do the following:
  + We have three possible categories for age: {young, middle, old} for a total of 3. It’s our choice which becomes base but in this example, we select young and create two dummy variables. The formula is shown below:

A screenshot of a cell phone

Description automatically generated

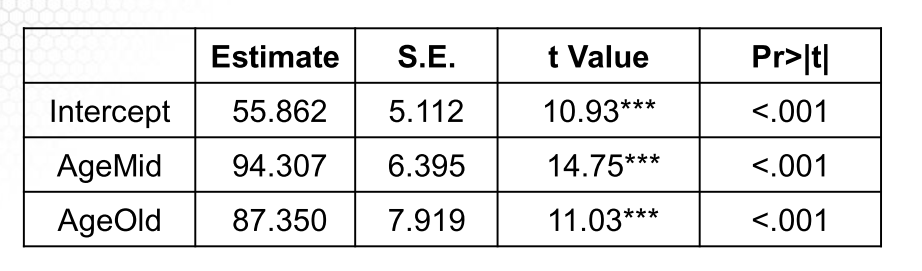
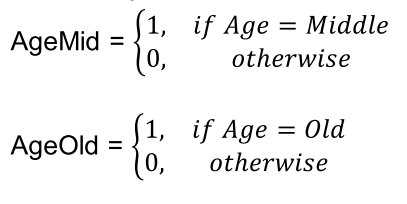
The instructor first tests whether categorical variable ‘Age’ has an effect on AmountSpent. ‘Age’ has three possible values: Young, Middle, or Old. To create indicator variables for Age, we need two indicator (dummy) variables. The base case, with both dummy variables set to 0, is Age = Young. It is up to modeler to determine which value of the categorical variable is used as the base case. Therefore, the two dummy variables we have are:

For example, when AgeMid = 0, AgeOld = 1, the record is for someone whose age is old. Note that AgeMid and AgeOld cannot be 1 at the same time since every individual has to be in exactly one age category.

Video 2.3 Interpreting the Coefficients of Indicator Variables

The instructor runs the regression model with results below:

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



* Note that with AgeMid and AgeOld are 0, 𝑏0 captures the average AmountSpent of customers who are Young. With AgeMid = 1 and AgeOld = 0, 𝑏0 + 𝑏1 capture the average AmountSpent of customer who are middle-aged. $94.307 is the increase in AmountSpent (on average) for middle- aged customers compared to young customers.
* In R, we can use a Factor Variable in regression to create dummy variables
  + Here’s the formula

**lm(AmountSpent ~ Age, data = dirmkt)**

A close up of text on a white background

Description automatically generated

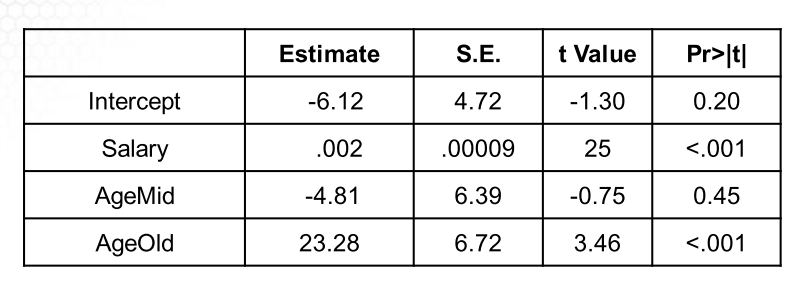
* where dirmkt is the dataset. R’s indicator variable coding scheme can be found by using - **contrasts(dirmkt$Age)** which is shown below:

A screenshot of a cell phone

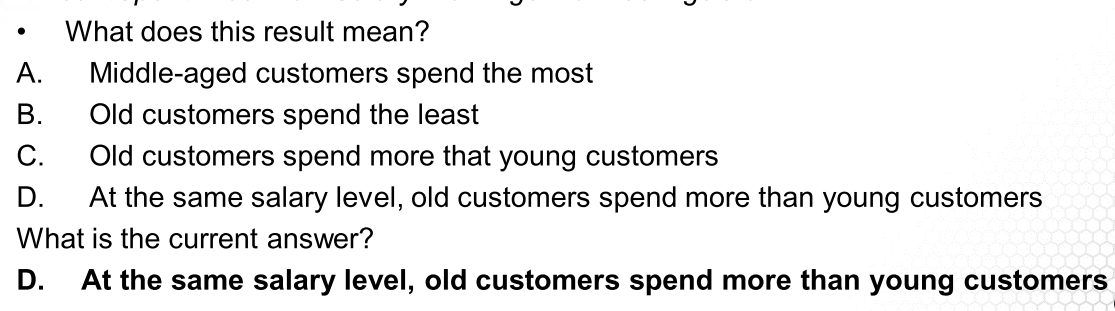
Description automatically generated

As a second example, here’s a regression equation with Salary and dummy variable Age. The output is shown further below:

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



For one unit increase in salary, the average AmountSpent increases by $0.002. Here is a quiz in the video, I think the answer is straightforward:

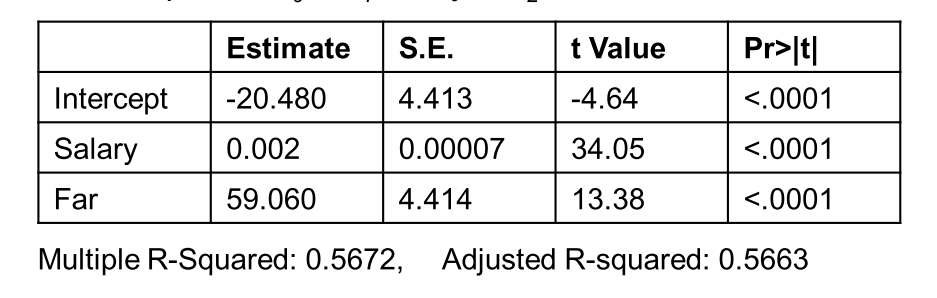


*D is the correct answer.*

Video 2.4 Interaction Term and Interpreting its Coefficient

The instructor first runs a regression using variables Salary and Location. Location is a categorical variable with ‘Close’ if the customer lives close to a store and ‘Far’ otherwise.

**𝐴𝑚𝑜𝑢𝑛𝑡𝑆𝑝𝑒𝑛𝑡 = 𝑏0 + 𝑏1 X 𝑆𝑎𝑙𝑎𝑟𝑦 + 𝑏2 X 𝐹𝑎𝑟**

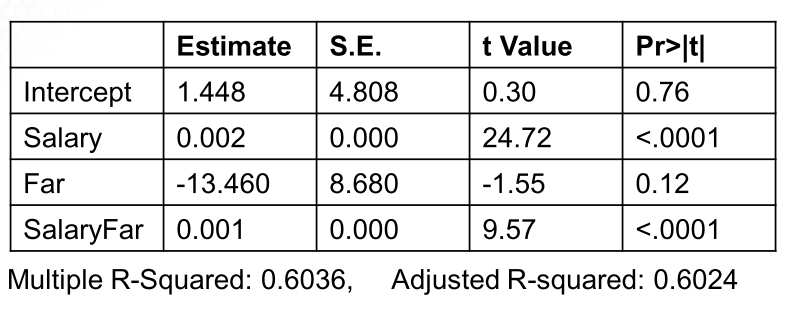


A picture containing bird, tree, flower

Description automatically generated

* However, in this above model, **we assume that customers who live far away from a store that sells similar products will spend at the same rate as customers who live close to a store.** 
  + For example, if the salary increases by $1000, no matter the customer lives far away or close, the average **AmountSpent** increase is $2.
  + One way of extending this model to allow for interaction effects is to include a third predictor, called an **interaction term**. In this case, we add a new variable **SalaryFar** which is Salary \* Far.

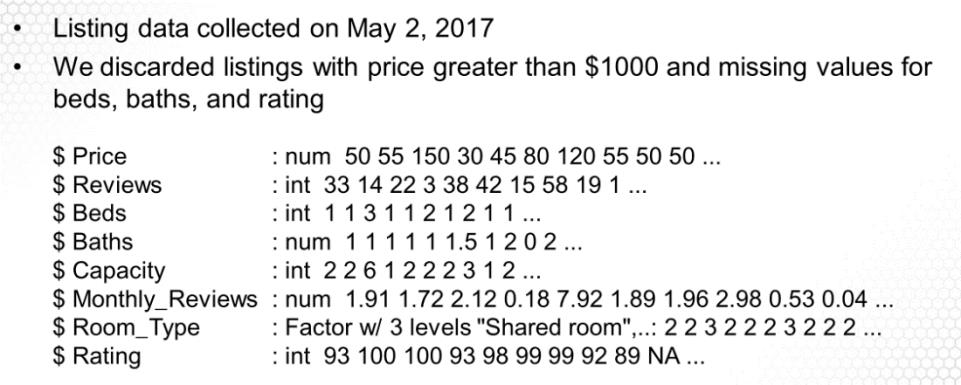
**𝐴𝑚𝑜𝑢𝑛𝑡𝑆𝑝𝑒𝑛𝑡 = 𝑏0 + 𝑏1 X 𝑆𝑎𝑙𝑎𝑟𝑦 + 𝑏2 X 𝐹𝑎𝑟 + 𝑏3 X 𝑆𝑎𝑙𝑎𝑟𝑦𝐹𝑎𝑟**



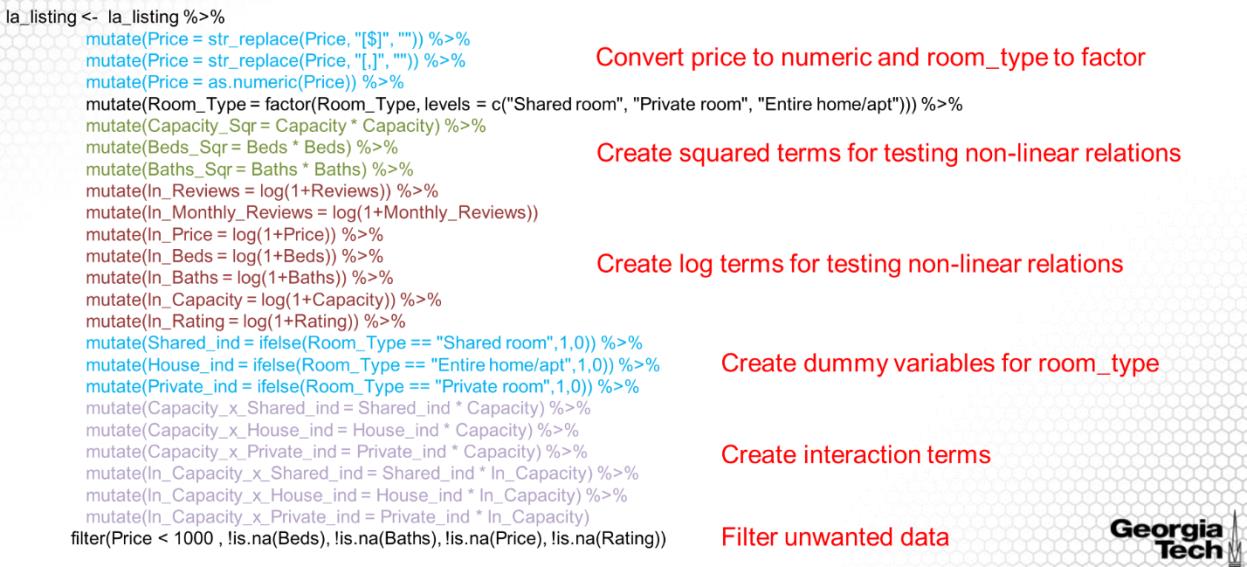
* The coefficient **𝑏3 (SalaryFar)**  is the amount to add to **𝑏1** **(Salary)** to get the slope for individuals who live far away.
  + If the salary of a customer who lives close increases by $10,000, the predicted increase in **AmountSpent** is 0.002 \* $10000 = $20.
  + If the salary of a customer who lives far away increases by $10,000, the predicted increase in **AmountSpent** is (0.002 + 0.001) \* $10000 = $30.

Video 2.5 Another Example of Using Indicator Variables

* The instructor uses a concrete example to summarize what we have learned. The dataset is an AirBnB for 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
* About the data used:



* We are going to find whether there is a relationship between capacity (capacity defined as can you add one more person) and price and whether Room\_Type (shared, private, or full house) changes this relationship.
* First, you need to clean the data. The visual below is the data wrangling that proceeded the analysis:



* First we run regression on Price against Capacity.

**𝑃𝑟𝑖𝑐𝑒 = 𝑏0 + 𝑏1 X 𝐶𝑎𝑝𝑎𝑐𝑖𝑡𝑦**

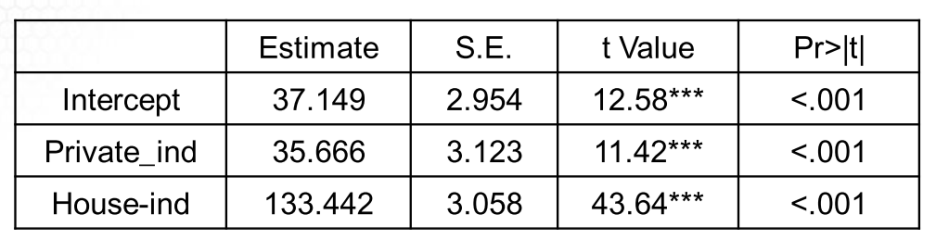
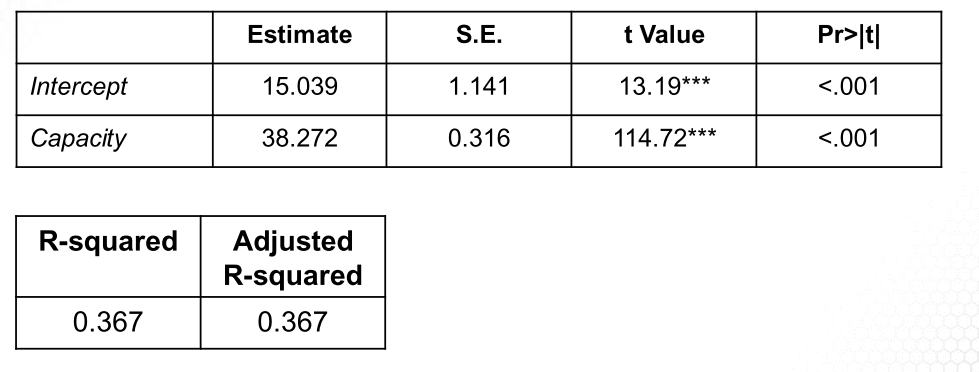
**A screenshot of a cell phone

Description automatically generated**

* Next, we use Room\_Type to run regression:

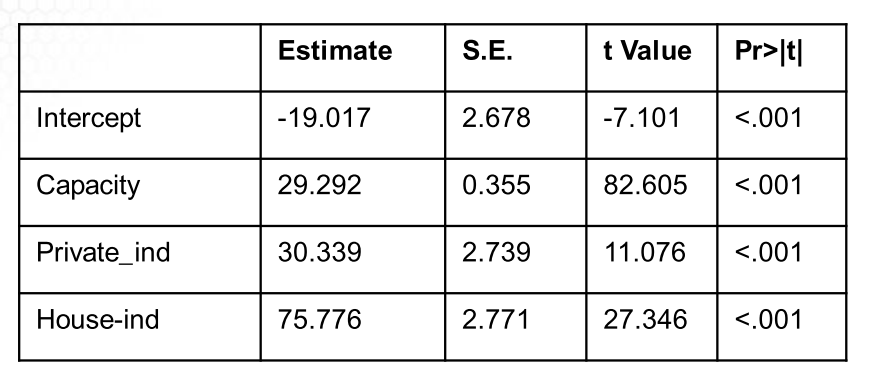
**𝑃𝑟𝑖𝑐𝑒 = 𝑏0 + 𝑏1 ∗ 𝑃𝑟𝑖𝑣𝑎𝑡𝑒\_𝑖𝑛𝑑 + 𝑏2 ∗ 𝐻𝑜𝑢𝑠𝑒\_𝑖𝑛𝑑**

* Note that the base case is ‘**Shared’**. The result is below:



* The shared room’s average price is $37.149; the private room’s average price is $37.149 + $ 35.666; the house’s average price is $37.149 + $133.442.
* Next we run regression using Capacity and Room\_Type,

**𝑃𝑟𝑖𝑐𝑒 = 𝑏0 + 𝑏1 X 𝐶𝑎𝑝𝑎𝑐𝑖𝑡𝑦 + 𝑏2 X 𝑃𝑟𝑖𝑣𝑎𝑡𝑒\_𝑖𝑛𝑑 + 𝑏3 X 𝐻𝑜𝑢𝑠𝑒\_𝑖𝑛𝑑**

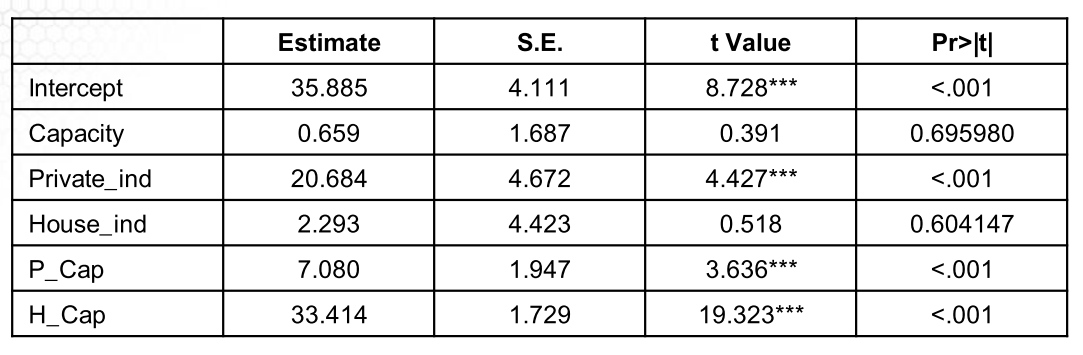


* Next we are adding interaction terms. Remember, **SharedRoom** is base case built into the intercept.

**𝑃𝑟𝑖𝑐𝑒=𝑏0 +𝑏1 X 𝐶𝑎𝑝𝑎𝑐𝑖𝑡𝑦+𝑏2 X 𝑃𝑟𝑖𝑣𝑎𝑡𝑒\_𝑖𝑛𝑑+𝑏3 X 𝐻𝑜𝑢𝑠𝑒\_𝑖𝑛𝑑+𝑏4 X 𝑃\_𝐶𝑎𝑝+𝑏5 X 𝐻\_𝐶𝑎𝑝**

Where:

* + 𝑃\_𝐶𝑎𝑝 = 𝑃𝑟𝑖𝑣𝑎𝑡𝑒\_𝑖𝑛𝑑 X𝐶𝑎𝑝𝑎𝑐𝑖𝑡𝑦,
  + 𝐻\_𝐶𝑎𝑝 = 𝐻𝑜𝑢𝑠𝑒\_𝑖𝑛𝑑 ∗ 𝐶𝑎𝑝𝑎𝑐𝑖𝑡𝑦
* The regression result is below where 𝑏4 is the amount to add to 𝑏1 to get the slope for a private room. 𝑏5 is the amount to add to 𝑏1 to get the slope for a house.



Video 3.1 Introduction to Non-linear Models

* Methods to detect non-linear relationships:
  1. **Q Q Plot:** scatterplot of quantiles graphed against each other
  2. **Fitted values vs. residuals:** plotting the fitted values on x-axis and residuals on the y.
  3. **Fitted values vs. sqrt residuals:**
  4. **Leverage vs standardized residuals:**
* In the following sections we will complete the following models to address non-linearity:

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

Video 3.2 Linear Log Model

* The purpose of transforming the predictor variable is to convert a non-linear relationship (i.e. exponential) into a linear relationship.

A screenshot of a cell phone

Description automatically generated

* Remember, a regression will have **unit changes** between x and y variables. Taking the log of one or both variables will effectively change the case from a ***unit***change to a ***percent***change.
  + That means a 0.1 unit change in log(x) is equivalent to a 10% in X.
* Below is a graph that shows the before and after of a transformation:

A screenshot of a map

Description automatically generated

A screenshot of a cell phone

Description automatically generated

# Video 3.3 Log-Linear Model

**A screenshot of a cell phone

Description automatically generated**

* Log of price is natural log of price hence the y = e ^ (bo + b1x) in the second bullet below in the interpretation.

**A screenshot of a cell phone

Description automatically generated**

# Video 3.3 Log-Log Model

A screenshot of a cell phone

Description automatically generated

**A screenshot of a cell phone

Description automatically generated**

* This model shows the best results when looking at diagnostics (below)

A close up of a map

Description automatically generated

# Video 3.4: Polynomial Model

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**1. Clarification About Homework 1**

Question 5: We want “percentage variation explained by speed” and “intercept of the regression model” and “coefficient of speed” separately but not “percentage variation explained by speed” and “percentage variation explained by intercept” and “percentage variation explained by speed coefficient”

Question 10: In option A “ error term = 1- confidence level”

Question 11: Choose the closet answer; set Facebook = 45 as a constant when calculating

Question 20: See below

**2. Diagnostic Plots**

* Residuals vs. Fitted  
  This plot shows if residuals have non-linear relationships. If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don’t have non- linear relationships or non-constant variances.
* Normal QQ plot  
  This plot shows whether the residuals are normally distributed https://stats.stackexchange.com/questions/101274/how-to- interpret-a-qq-plot

**3. Log Transformation and Interpretation**

In a regression model 𝑦 ~ 𝑥:  
If we say, 𝑥 increases by 1 unit, we simply mean 𝑥 → 𝑥 + 1

If we say, 𝑥 increases by 1%, we simply mean 𝑥 → (1 + 1%)𝑥 or  
𝑙𝑜𝑔𝑥 → 𝑙𝑜𝑔𝑥 + 0.01; The former is an approximation while the latter is the accurate formula in this course

The same terminology applies to 𝑦  
Notice: 𝑥 → (1 + 1%)𝑥 ⇔ 𝑙𝑜𝑔 𝑥 → 𝑙𝑜𝑔(1 + 1%)𝑥 = 𝑙𝑜𝑔 𝑥 +

𝑙𝑜𝑔1.01 ≈ 𝑙𝑜𝑔 𝑥 + 0.01

**In homework 1 from Question 17 to Question 20, please use** 𝒍𝒐𝒈𝒙 → 𝒍𝒐𝒈𝒙 + 𝟎. 𝟎𝟏 **to solve the problems**

**Linear-Linear Model:**

𝑦 = 𝑏0 + 𝑏1𝑥 and 𝑥 → 𝑥 + 1  
𝑦(𝑥+1)−𝑦(𝑥)=(𝑏0 +𝑏1(𝑥+1))−(𝑏0 +𝑏1𝑥)=𝑏1 As 𝑥 increases by 1 unit, 𝑦 increases by 𝑏1 units

**Linear-Log Model:**

𝑦=𝑏0+𝑏1𝑙𝑜𝑔𝑥and𝑥→(1+1%)𝑥or𝑙𝑜𝑔𝑥→𝑙𝑜𝑔𝑥 +0.01 𝑦(1.01𝑥) − 𝑦(𝑥) = (𝑏0 + 𝑏1𝑙𝑜𝑔(1.01𝑥)) − (𝑏0 + 𝑏1𝑙𝑜𝑔 𝑥)

= 𝑏1𝑙𝑜𝑔(1.01) ≈ 0.01𝑏1

𝑦(𝑙𝑜𝑔 𝑥 + 0.01) − 𝑦(𝑙𝑜𝑔 𝑥)  
=(𝑏0 +𝑏1(𝑙𝑜𝑔𝑥+0.01))−(𝑏0 +𝑏1𝑙𝑜𝑔𝑥)=0.01𝑏1

As 𝑥 increases by 1%, 𝑦 increases by 0.01𝑏1 units **Log-Linear Model:**

𝑙𝑜𝑔𝑦=𝑏0+𝑏1𝑥→𝑦=𝑒𝑏0+𝑏1𝑥 and𝑥→𝑥+1 𝑦(𝑥 + 1) 𝑒𝑏0+𝑏1(𝑥+1)

𝑦(𝑥) = 𝑒𝑏0+𝑏1𝑥 =𝑒𝑏1 ⇔𝑦→(1+(𝑒𝑏1 −1))𝑦 As 𝑥 increases by 1 unit, 𝑦 increases by 100(𝑒𝑏1 − 1)%

**Log-Log Model:**

𝑙𝑜𝑔𝑦=𝑏0+𝑏1𝑙𝑜𝑔𝑥→𝑦=𝑒𝑏0+𝑏1𝑙𝑜𝑔𝑥 =𝑒𝑏0𝑥𝑏1 and𝑥→(1+1%)𝑥 or𝑙𝑜𝑔𝑥 →𝑙𝑜𝑔𝑥+0.01

𝑦(𝑙𝑜𝑔 𝑥 + 0.01) 𝑒𝑏0+𝑏1(𝑙𝑜𝑔 𝑥+0.01)  
𝑦(𝑙𝑜𝑔 𝑥) = 𝑒𝑏0+𝑏1𝑙𝑜𝑔 𝑥 = 𝑒0.01𝑏1

⇔𝑦→(1+(𝑒0.01𝑏1 −1))𝑦  
As 𝑥 increases by 1%, 𝑦 increases by 100(𝑒0.01𝑏1 − 1)%

page3image28736320page3image28736128page3image28735936page3image28735744