# Lab #2 –Correlation and Regression

In this lab we will apply what we’ve learned about graphics, correlation, and regression. The data that we’ll be looking at was collected by a faculty member at the University of Iowa, at the time he was looking to buy a house in Iowa City and wanted to use data to make sure he was getting a good deal.

Write-up questions are colored in red. I expect to receive exactly one write-up per group (with everyone’s names on it), but I encourage you to keep your own copy.

**PLEASE ADDRESS THESE QUESTIONS IN A SEPARATE DOCUMENT**

The “Iowa City Homes Data” contains 19 different variables:

* Sale.amount – How much (in dollars) the house sold for
* Sale.date – The date when the house was sold
* Occupancy – A categorical variable describing how the property is designated
* Style – A categorical variable describing the home’s style
* Built – The year in which the home was constructed
* Bedrooms – The number of bedrooms
* Bsmt – The type of basement (full, none, 1/2, 3/4, or Crawl)
* Ac – Whether or not the home has central air conditioning
* Attic – Does the home have an attic and how is it set-up/accessed
* Area.base – Area of the ground floor (in square feet) of the home
* Area.add – Additional area (in square feet), not including the basement or garage
* Area.garage1 – Area (in square feet) of the attached garage
* Area.garage2 – Area (in square feet) of the detached garage
* Area.living - Total area (in square feet) of the home’s livable space
* Area.lot – Area of the lot (in square feet) that the house is built on
* Lon – The longitudinal coordinate of the home’s location
* Lat – The latitude of the home’s location
* Assessed – The home’s assessed value (in dollars) by the Iowa City Assessor

In this lab we will focus on a few different research questions:

1. **“How are various home attributes related to the home’s sale price?”**
2. **“Can we predict a home’s sale price based upon its attributes?”**
3. **“How can we tell if a home’s sale price is a good deal?”**

## Part 1 – Exploring the Variables

The first step in any statistical analysis is to explore the variables in your dataset. When doing this exploration it is important to consider:

* + Is the data a sample? What are the cases? What populations do they represent/not represent?
  + Which variables are categorical? Which variables are quantitative?
  + What are the categories of each categorical variable and are what their frequencies?
  + How are the quantitative variables distributed? Do they have any outliers?

### Question 1

I expect that you will have discussed with your group each of questions posed above. For your write-up I’d like you to include:

1. A sentence addressing the first bullet on the previous page
2. A one-way frequency table of the variable “style”
3. A five number summary of “area.living”
4. Histograms for the variables “assessed” and “sale.amount”

## Part 2 – Exploring the Variable Relationships

After getting a feel for the variables and cases in a data set, the next step in a statistical analysis is to get a preliminary understanding of how these variables might relate with each other. Usually there are too many possibilities to exhaustively look at all two variable combinations, instead we tend to focus on the variables that we think will be most important in our analysis. Importance is usually based upon your background knowledge of the research question.

Determining which variables are of interest before beginning a statistical analysis helps protect against “data dredging” or “p-hacking” (<https://en.wikipedia.org/wiki/Data_dredging>), the more variables we look at, the more likely it is for at least one variable to appear important just by random chance.

### Question 2

Using your knowledge of home pricing, identify 5 explanatory variables that you believe might have important or interesting relationships with home sale prices. Make sure that you choose at least 2 categorical variables and at least 2 quantitative variables. List these variables in your lab write-up.

### Question 3

Explore how the variables you identified in Question 2 relate with each other, as well as how they relate with the outcome variable, using the appropriate method (ie: scatterplots, boxplots, two-way frequency tables, etc.) Include in your lab write-up the 3 figures or tables that you found most interesting, at least 1 of these should involve the outcome variable “sale.amount”. Each should be accompanied a caption containing 1-2 sentences summarizing what the figure/table tells you.

## Part 3 – Analyzing the Data

Now that you’ve gotten a feel for the variables in the dataset and how they relate with each other, it is time to address our research questions, starting with the first question: “How do various home attributes relate to a home’s sale price?”

### Question 4

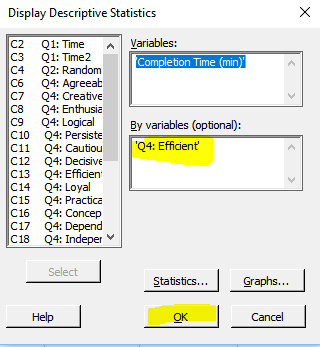
Describe how you could numerically determine (ie: using one or more *summary statistics*) which of your chosen *quantitative* variables is most related with sale price?

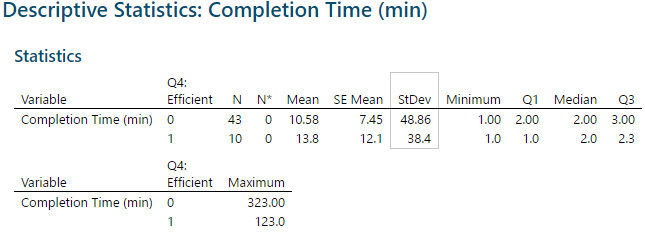
### Question 5

Describe how you could numerically determine (ie: using one or more *summary statistics*) which of your *categorical* variables is most related with sale price? (Hint: it might be beneficial to read the top of the next page, which talks about the “by” option in Minitab when thinking about question 5)

**The “by” option in Minitab:**

When implementing your answers to Question 5 you might need to make use of what is called a “by variable”. The images below show an example of how to calculate a 5-number summary of the variable “completion time” in the Class Survey Data, grouped by whether the student described themselves as "efficient”:





### Question 6

For the 5 variables you identified in Question 2, use the approach you described in Question 4 to identify the *quantitative* variable that is most related with sale price. Write 1-2 sentences describing your findings, make sure you include the relevant summary statistic(s) that you used to make your assessment.

### Question 7

For the 5 variables you identified in Question 2, use the approach you described in Question 5 to identify the *categorical* variable that is most related with sale price. Write 1-2 sentences describing your findings, make sure you include the relevant summary statistic(s) that you used to make your assessment.

**Addressing the second research question:**

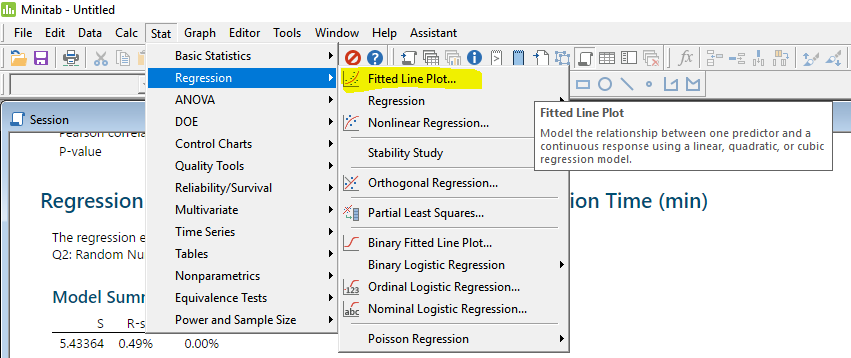
Now let’s move on the second research question: “Can we predict a home’s sale price based upon its attributes?” We’ve already seen that predictions can be made using the correlation coefficient or using the regression line, now we will practice both approaches.

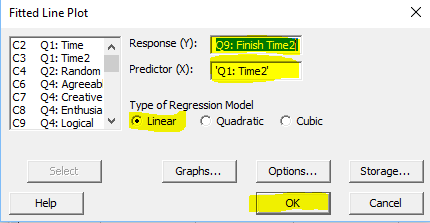
### Question 8

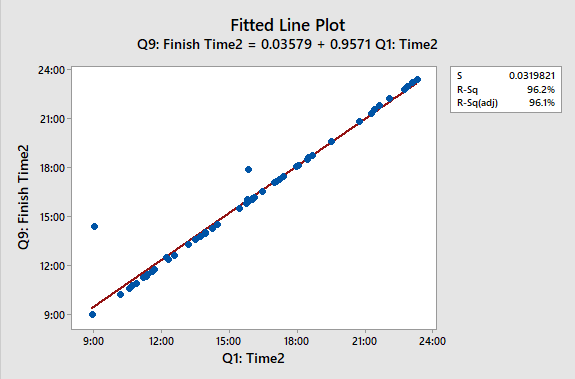
For the quantitative variable you identified in Question 6, use its correlation coefficient with sale price to predict the sale price of a home that is 1.5 standard deviations above average for the variable you identified. In your write-up list the relevant quantities and show your work. Just to be clear, I’m asking you to predict the sale price for a home that is 1.5 standard deviations above average for the quantitative explanatory variable you identified.

**Regression in Minitab:**

The next question will require you to fit a regression line. While there are many ways to perform regression in Minitab, the images below show what I would advise you do for this lab. Here I use start time to predict finish time for the class survey data:







As you can see, the slope of this line is 0.957 and the intercept is 0.036. The other piece of information that we might be interested in is “R-Sq”, the *coefficient of determination* or *R-squared*. is a summary statistic that conveys the proportion of variability in the outcome variable, Y, that can be explained by the explanatory variable, X:

In this example, where the outcome variable is “Finish Time”, and the explanatory variable is “Start Time”, we see that “Start Time” explains 96.2% of the variability in “Finish Time”.

### Question 9 A

Do you see any similarities in how is calculated and how the standard deviation of Y is calculated? What might be expressing (with regards to the standard deviation)? Explain in 1-2 sentences.

### Question 9 B

For your regression line, take the square root of “R-Sq”? Have you seen this number before when answering earlier questions?

### Question 10

For the quantitative variable you identified in Question 6, use your fitted regression line to predict the sale price of a home that is 1.5 standard deviations above average for the explanatory variable that you identified. In your write-up, include the fitted line plot, and describe how you used the regression line equation to arrive at your answer.

**Addressing the third research question:**

Now let’s turn our attention to the third research question: “How can we tell if a home’s sale price is a good deal?”

### Question 11

Consider the explanatory variable you identified in Question 6 and explain, in non-statistical terms, what a “good deal” means in regards to this variable and “sale.amount”, your answer should be 1-2 complete sentences.

### Question 12

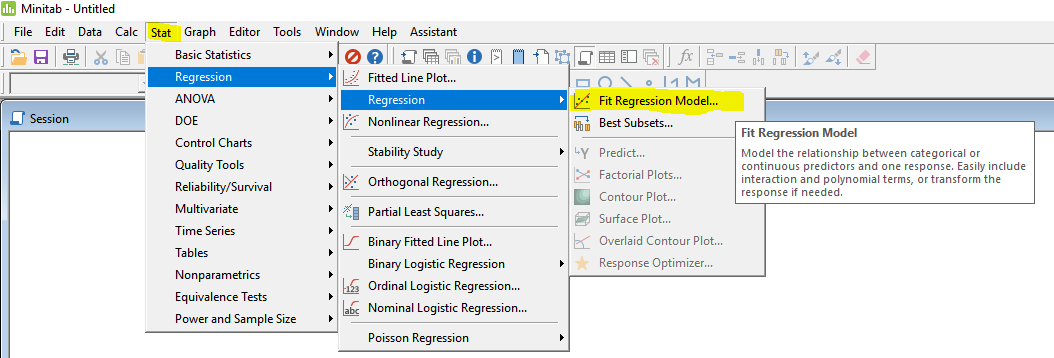
Now, explain using **residuals** and the regression line you constructed in Question 9, what a “good deal” means, your answer should be 1-3 complete sentences.

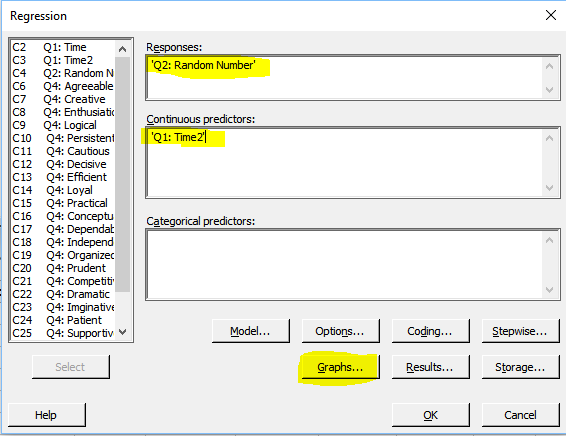
**Residuals in Minitab:**

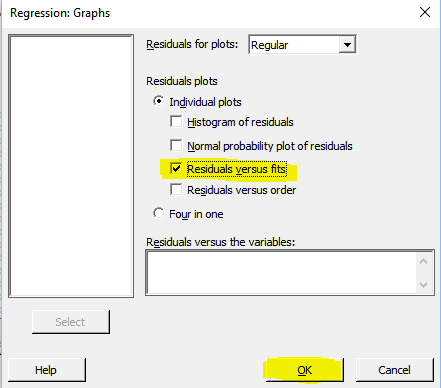
### Question 13

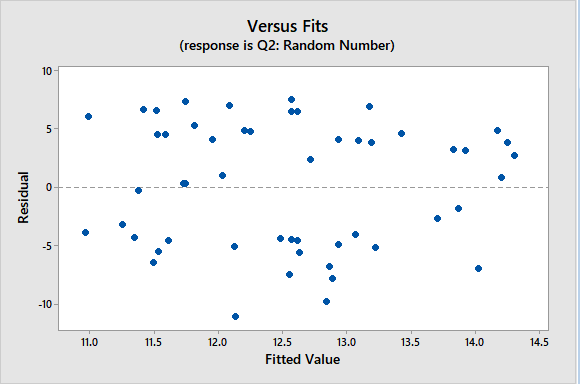
Use your fitted regression line to create a new variable called “sale\_residuals”, this variable should contain the residual for each predicted sale price from your regression model in Question 10, it should be calculated using a formula in Minitab.

Now let’s see a few different ways to view residuals in Minitab, the sequence of images below illustrates the process on the class survey data using a model with “Q1: Time2” its explanatory variable and “Q9: Finish Time” as its response:

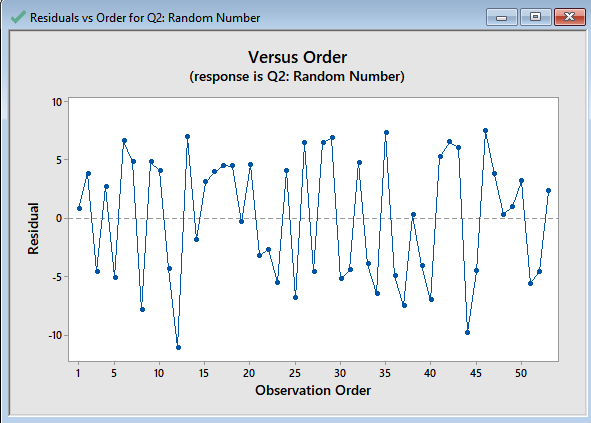








We can see that the biggest under prediction occurred for an individual whose predicted (fitted) random number was 12.2. We can use the plot option “Residuals versus order” to construct the following:



From this plot it is apparent that 12th case in our data has the largest negative residual.

### Question 14

Using residuals, identify 3 cases that you consider to be “good deals”, include information on your explanatory variable, the predicted sale prices, the observed sale prices, and a short explanation why these sales were “good deals”.

## Part 4 – Your Own Analysis

### Question 15

Devise your own research question for these data. In your write-up include the following:

1. A clearly stated research question
2. 1-2 sentences describing the statistical approach you can use to answer it
3. 1-2 sentences describing your results after implementing your approach (including 1-2 figures or tables)

### Question 16 (Optional)

Regression can also be done using categorical explanatory variables. Try this out using the variable “ac”. Looking at the fitted model, what does the intercept represent? What does the slope represent? What happens when you use a categorical variable with more than two categories?