

# LAB: INTRO TO STAT ANALYSIS

---

Zeyi Qian

[zeqian@clarku.edu](mailto:zeqian@clarku.edu)

Office Hours: JC 201, Tuesday 4-5 PM & Thursday 3-4 PM

September 20, 2024

## Question 1

- Fast food is often considered unhealthy because much of it is high in fat and calories. Are **fat** and **calories** related? Analyze the association between fat content and calories by calculating the correlation coefficient and drawing a scatterplot (Use a computer)

Fat (g)	19	31	34	35	39	39	41
Calories	410	560	585	570	640	680	660

# Question 1

Q1	Fat (g)	19	31	34	35	39	39	41
	Calories	410	560	585	570	640	680	660

=CORREL(C1:I1,C2:I2)

Function Arguments ? X

CORREL

<b>Array1</b>	C1:I1	↑	= {19,31,34,35,39,39,41}
<b>Array2</b>	C2:I2	↑	= {410,560,585,570,640,680,660}

= 0.977514742

Returns the correlation coefficient between two data sets.

**Array1** is a cell range of values. The values should be numbers, names, arrays, or references that contain numbers.

---

Formula result = 0.977514742

[Help on this function](#) OK Cancel

# Question 1

Page Layout Formulas Data Review View Automate Developer Help Power Pivot Chart Design Format

Table Pictures 3D Models Shapes SmartArt Checkboxes Recommended Charts Maps PivotChart Line Column Win/Loss Slicer Timeline Link Comment Text Box Header & Footer

Illustrations Controls

Change Chart Type

Recommended Charts All Charts

Recent Templates Column Line Pie Bar Area XY (Scatter) Map Stock Surface Radar Treemap Sunburst Histogram Box & Whisker Waterfall Funnel Combo

Calories

	D	E	F	G	H	I	J
9	31	34	35	39	39	41	
10	560	585	570	640	680	660	

Calories

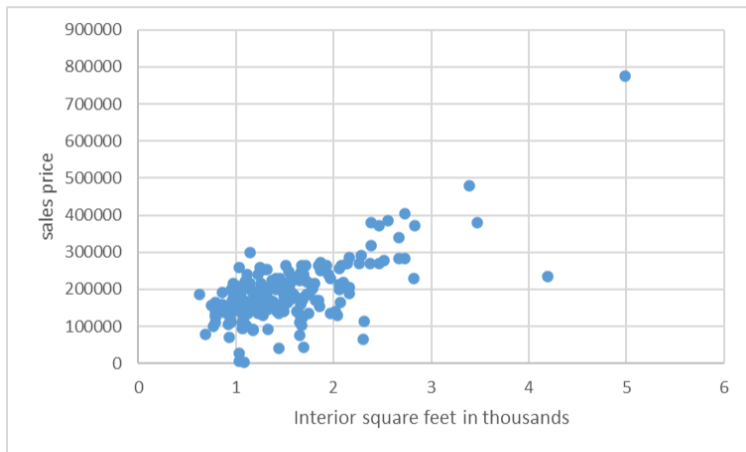
## Question 2

This question asks you to work with housing sales data from the 01602 zip code of Worcester, which is across Park Ave from the Clark neighborhood. All of the houses are single-family houses. Please use the Excel Spreadsheet that is on the page for this assignment.

- a. Please use the “Single Family” Tab first. Please find the scatterplot for **floor area of the house (interior square feet in thousands)** and **sales price**. Why would it seem to make the most sense to put the floor area on the “x-axis and the sales price on the “y-axis.”?

It is logical to treat sales price as something that responds to changes in floor area

## Question 2: a



- It is logical to treat sales price as something that responds to changes in floor area

## Question 2: b

- b. Please describe the scatterplot you have created according to these criteria:
  - i. direction of association (positive or negative?) **positive**
  - ii. is the form curved, straight, or “exotic?” **straight (linear)**
  - iii. is the strength of association apparently strong or weak? **strong**

## Question 2: c

- c. Using the data from the “Single Family” tab, calculate the correlation coefficient by hand and then check your work using the excel command. **(0.688743)**

$$Z_{x_i} = \frac{x_i - \bar{x}}{Stdev_x}, \quad Z_{y_i} = \frac{y_i - \bar{y}}{Stdev_y}$$

$$r = \sum_{i=1}^n \frac{Z_{x_i} Z_{y_i}}{n - 1}$$

- Order of calculation by hand
  - Mean of x & y
  - Deviation from x & deviation from y
  - Stdev of x & y
  - Z score of x & y
  - Correlation

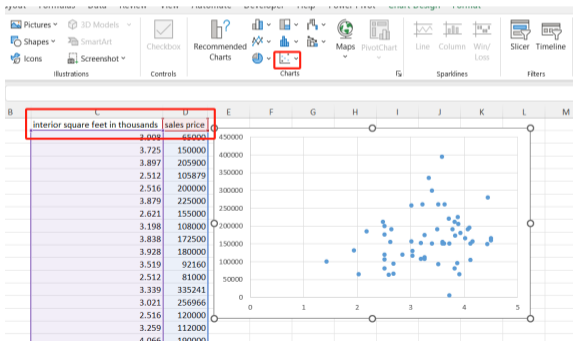


## Question 2: d

- d. Using the data on the “Three Family” tab, you will find three-family houses that were sold in the adjacent 01603 zip code that is in South Worcester further down Main Street away from downtown. Please find the same kind of scatter plot that you did in part a., but now it will be for three-family houses and describe it using the criteria in part b.:
  - i. direction of association (positive or negative?)
  - ii. is the form curved, straight, or “exotic?”
  - iii. is the strength of association apparently strong or weak?

## Question 2: d

- Positive
- Straight, though there is a considerable spread in the data points
- (Moderate to) Strong
- Other factors may also play a significant role, leading to the observed variability



## Question 2: e & f

- e. Using the data on the “Three Family” tab, test the regression assumptions
  - Linearity Assumption
  - Independence Assumption
  - Equal Variance Assumption
  - Normal Population Assumption
- f. Estimate a regression model using Excel. What is the interpretation of the coefficients (i.e., put the coefficients in a sentence)?

## Question 2: e &amp; f

The screenshot shows the Microsoft Excel interface. In the top right corner, there is a 'Comments' button and a green icon. The ribbon contains several tool icons: 'Filter', 'Clear', 'Reapply', 'Text to Columns', 'What-If', 'Forecast', and 'Outline'. The 'Data Analysis' button is highlighted with a red box. Below the ribbon, the 'Data Analysis' dialog box is open. The 'Analysis Tools' list includes: Descriptive Statistics, Exponential Smoothing, F-Test Two-Sample for Variances, Fourier Analysis, Histogram, Moving Average, Random Number Generation, Rank and Percentile, Regression, and Sampling. The 'Regression' option is highlighted with a blue background and a red box. To the right of the list are three buttons: 'OK', 'Cancel', and 'Help'. The background shows a spreadsheet with columns labeled 'Q', 'R', and 'S'.

## Question 2: e &amp; f

Regression

Input

Input Y Range:

Input X Range:

Labels  Constant is Zero

Confidence Level:  %

Output options

Output Range:

New Worksheet Ply:

New Workbook

Residuals

Residuals  Residual Plots

Standardized Residuals  Line Fit Plots

Normal Probability

Normal Probability Plots

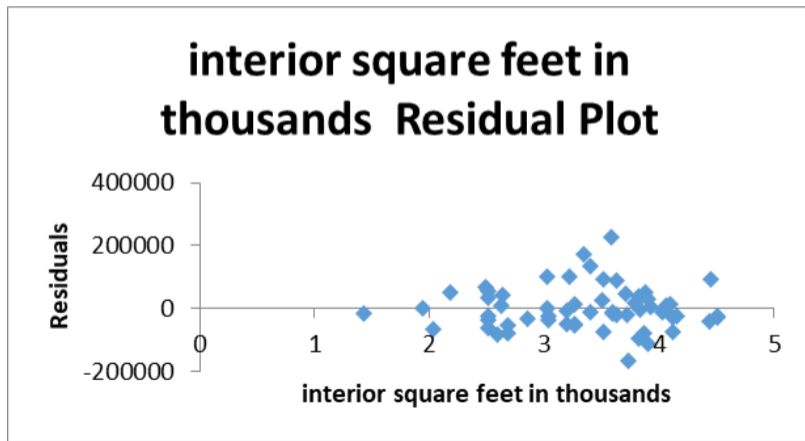
OK

Cancel

Help

## Question 2: e & f

- Independence Assumption



# Question 2: e & f

- Linearity Assumption **Put y into new sheet**

Microsoft Excel ribbon showing the **Insert** tab with the **Charts** group highlighted in red. The **Recommended Charts** button is also highlighted.

The **Insert Chart** task pane is open, showing the **Recommended Charts** tab. The **Scatter** chart type is selected, and a preview of a scatter plot titled "Residuals" is displayed.

The background shows a spreadsheet with the following data:

	SS	MS	F	Signif
1	15165459692	15165459692	3.116139	0.
7	2.77405E+11	4866747481		
8	2.9257E+11			

	Standard Error	t Stat	P-value	Lower
4	45103.04486	1.875713688	0.065819	-5
4	13209.92873	1.765258792	0.082877	-3

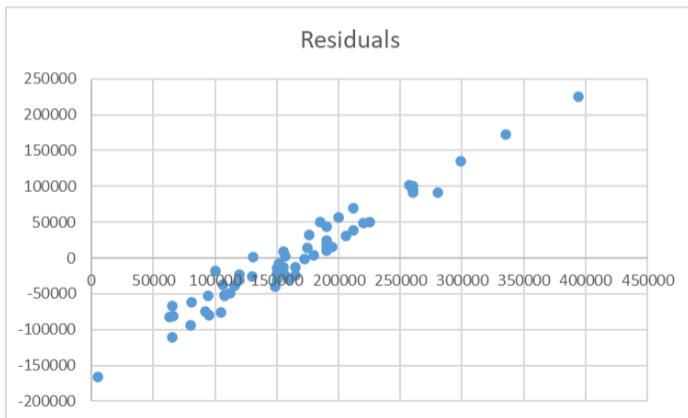
At the bottom of the spreadsheet, the text "PROBABILITY" is visible.

The **Scatter** chart preview shows a scatter plot titled "Residuals" with a y-axis ranging from -200,000 to 250,000 and an x-axis ranging from 0 to 450,000. The data points show a positive linear relationship.

A scatter chart is used to compare at least two sets of values or pairs of data. Use it to show relationships between sets of values.

## Question 2: e & f

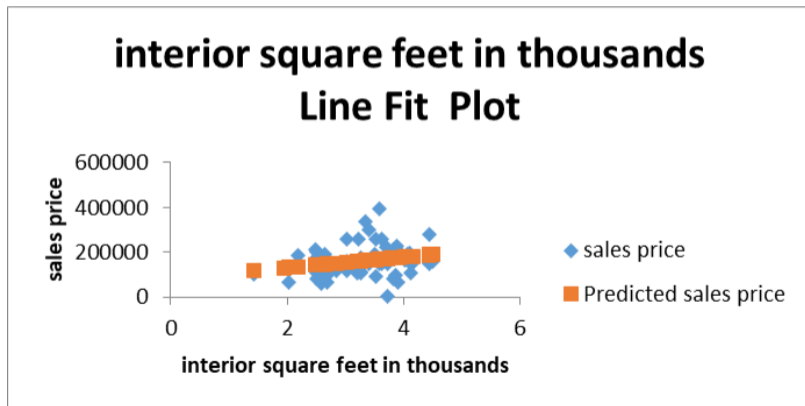
- Linearity Assumption





## Question 2: e & f

- Equal Variance Assumption



## Question 2: e & f

- Normal Population Assumption

The screenshot shows an Excel spreadsheet with a regression analysis table. The table has three columns: 'Predicted sales price', 'Residuals', and 'Standard Error'. The 'Residuals' column is highlighted with a red box. A 'Data Analysis' dialog box is open over the spreadsheet, with the 'Histogram' option selected and highlighted with a red box. The dialog box also shows other analysis tools like 'F-Test Two-Sample for Variances', 'Fourier Analysis', 'Moving Average', 'Random Number Generation', 'Rank and Percentile', 'Regression', 'Sampling', 't-Test: Paired Two Sample for Means', and 't-Test: Two-Sample Assuming Equal Variances'. The 'OK', 'Cancel', and 'Help' buttons are visible on the right side of the dialog box.

	Predicted sales price	Residuals	Standard Error
1	175730.8272	-110730.8272	-1.60
2	171463.4607	-21463.46068	-0.3
3	175474.3188	30425.68115	0.4
4	143177.583	-37298.58303	-0.5
5	143270.8588	56729.1412	0.8
6	175054.5779	49945.42212	0.7
7	145719.3478	9280.652205	0.1
8	159174.3778	-51174.37781	-0.7
9	174098.5012	-1598.501222	-0.0
0	176197.2061	3802.793923	0.054986991
1	166659.7585	-74499.75846	-1.077238899

## Question 2: e &amp; f

Histogram ? ×

**Input**

Input Range:

Bin Range:

Labels

**Output options**

Output Range:

New Worksheet Ply:

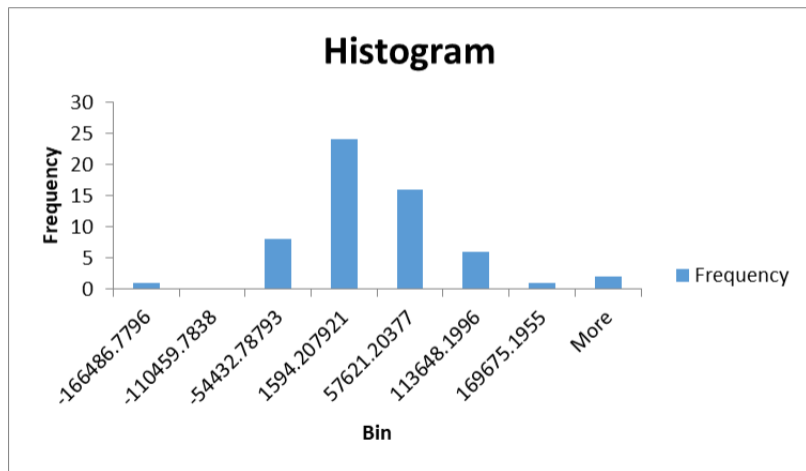
New Workbook

Pareto (sorted histogram)

Cumulative Percentage

Chart Output

## Question 2: e &amp; f



## Question 2: e & f

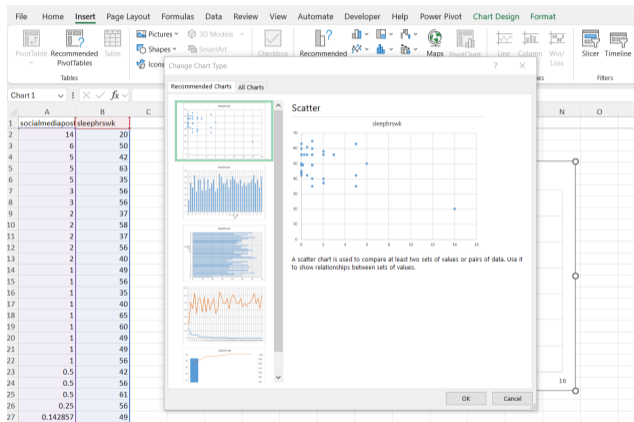
- All else being equal, for each additional 1,000 square feet of house area, the expected increase in the sales price is \$23,318.94

## Question 3

For this question, please use the data from the student survey on social media use (“socialmedia&sleep” Tab).

- a. Find the scatter diagram for social media posts per day and hours of sleep per week.
- b. Find the correlation coefficient using excel.
- c. Suppose that we multiplied the “posts per day” variable by 7 so that it was measured in terms of weeks (rather than days). Would that influence the correlation coefficient ( $r$ )? How?
- d. Calculate the mean and standard deviation of the two variables. Then calculate the regression slope and regression intercept by hand. Write out the regression model.
- e. Estimate the regression model using Excel. Do the answers in d. and e. match?

# Question 3: a & b & c



- $r = \text{CORREL}(A2:A40, B2:B40) = -0.477231403$
- If we multiplied the “posts per day” variable by 7,  $r = -0.477231403$  still holds

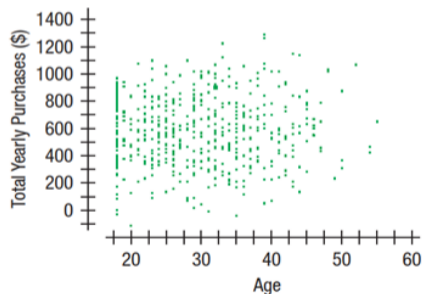
## Question 3: d & e

- hours of sleep per week =  $b_0 + b_1$  social media posts per day
- mean of  $x$  =  $\text{AVERAGE}(A2:A40)=1.59$
- mean of  $y$  =  $\text{AVERAGE}(B2:B40)=49.76$
- Stdev of  $x$  =  $\text{STDEV}(A2:A40)=2.59$
- Stdev of  $y$  =  $\text{STDEV}(B2:B40)=9.53$
- $b_1=r S_y/S_x = \text{SLOPE}(B2:B40,A2:A40)=1.75$
- $b_0=\text{mean of } y - b_1 \text{ mean of } x = \text{INTERCEPT}(B2:B40,A2:A40)=52.55$
- **Data Analysis** will give you same answer



## Question 4

**Online clothes** An online clothing retailer keeps track of its customers' purchases. For those customers who signed up for the company's credit card, the company also has information on the customer's *Age* and *Income*. A random sample of 500 of these customers shows the following scatterplot of *Total Yearly Purchases* by *Age*:



The correlation between *Total Yearly Purchases* and *Age* is  $r = 0.037$ . Summary statistics for the two variables are:

## Question 4

	Mean	SD
Age	29.67 yr	8.51 yr
Total Yearly Purchase	\$572.52	\$253.62

- What is the linear regression equation for predicting *Total Yearly Purchase* from *Age*?
- Do the assumptions and conditions for regression appear to be met?
- What is the predicted *Total Yearly Purchase* for an 18-year-old? For a 50-year-old?
- What percent of the variability in *Total Yearly Purchases* is accounted for by this model?
- Do you think the regression might be a useful one for the company? Explain.

## Question 4

- Total Yearly Purchases =  $b_0 + b_1 \text{ Age}$
- $b_1 = r S_y / S_x = 11.37$
- $b_0 = \text{mean of } y - b_1 \text{ mean of } x = 268.56$
- Total Yearly Purchases =  $268.56 + 11.37 \times \text{Age}$
- $R^2 = 0.001369$
- If age = 18, predicted  $y = 268.56 + 11.37 \times 18 = 473.22$
- If age = 50, predicted  $y = 268.56 + 11.37 \times 50 = 837.06$