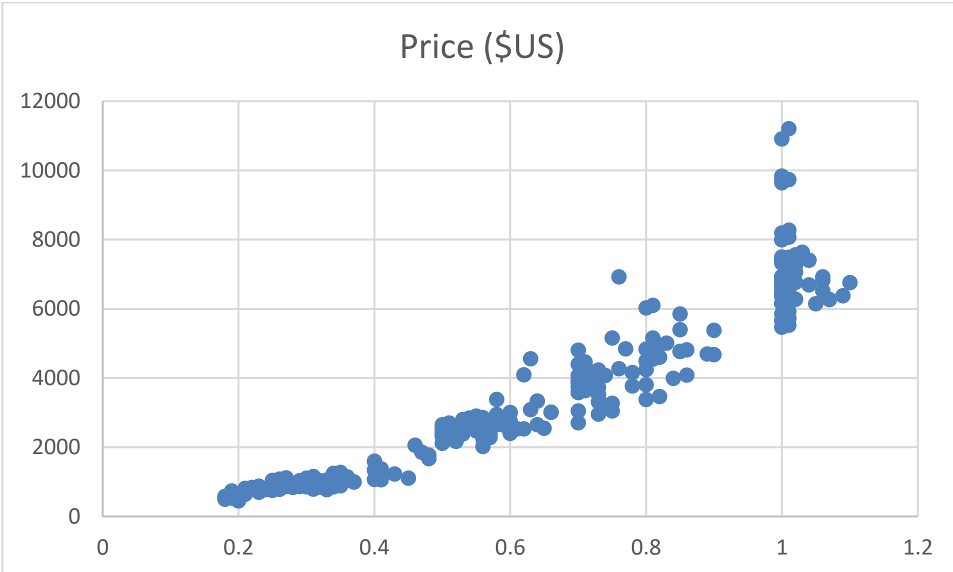
Intro to Correlation and Linear Regression

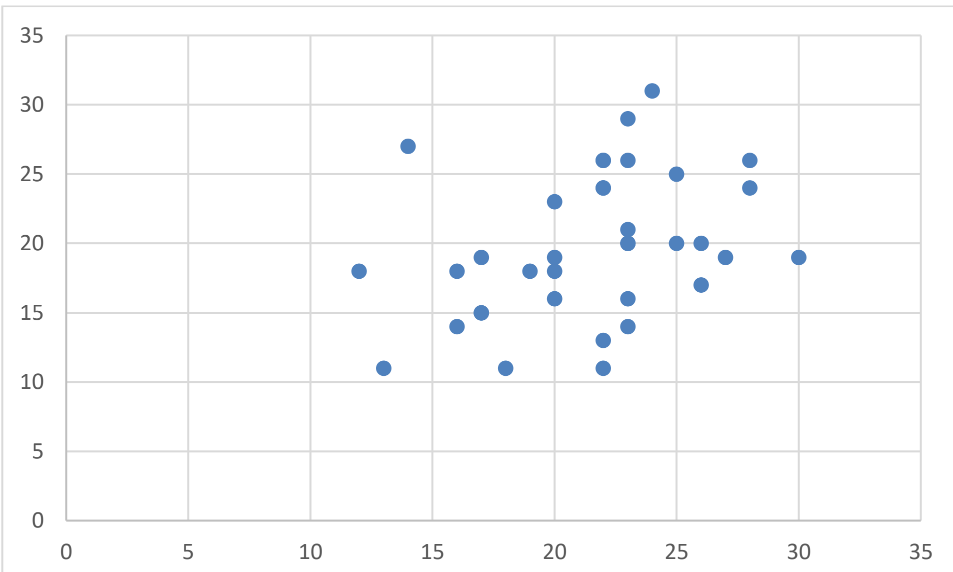
12.1 Bivariate Data and Correlation

To complete this section of homework watch Chapter Twelve, Lecture Example: 12.1.

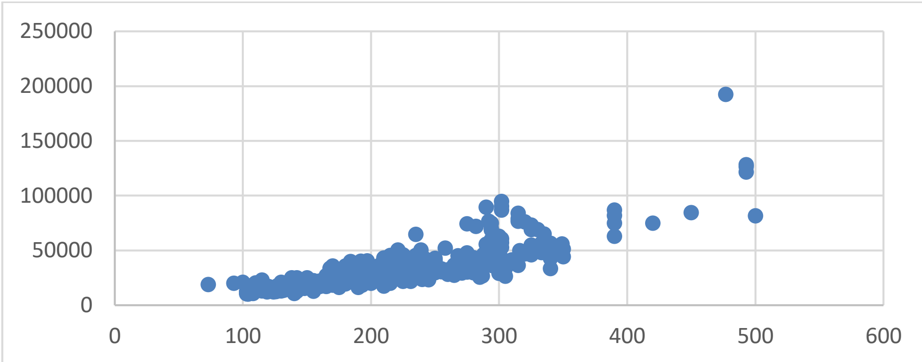
1. Determine whether the scatterplot shows little or no association, a negative association, a linear association, a moderately strong association, or a very strong association:



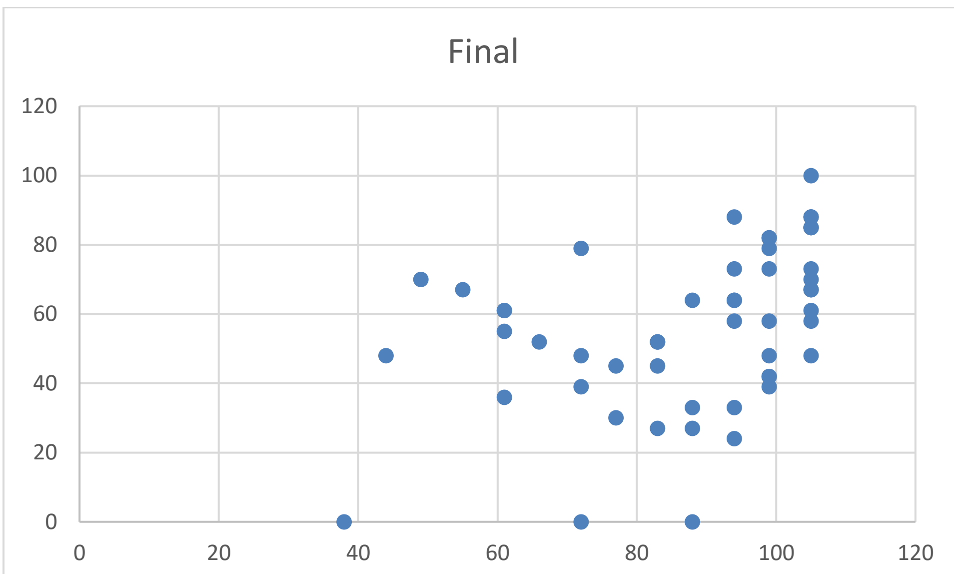
2. Determine whether the scatterplot shows little or no association, a negative association, a linear association, a moderately strong association, or a very strong association:



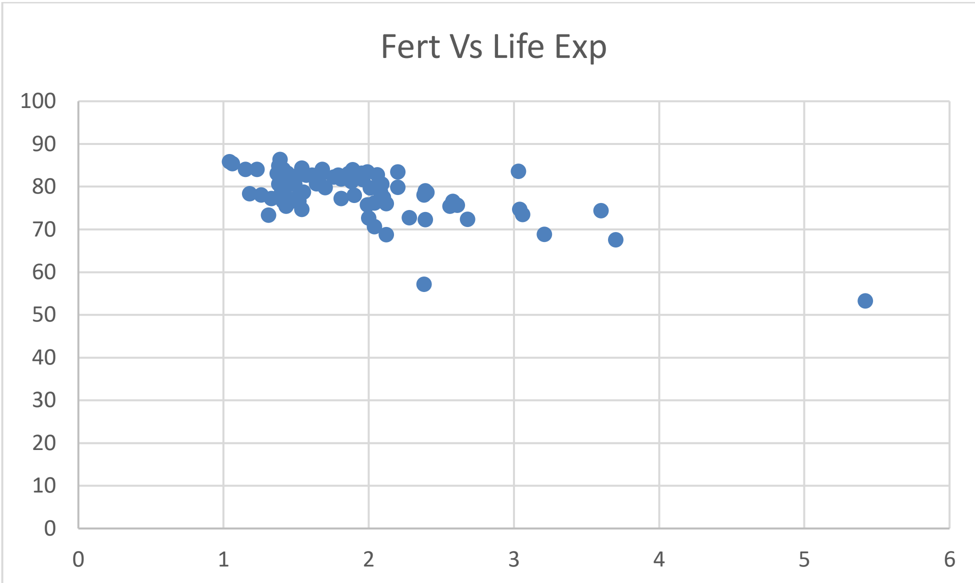
3. Determine whether the scatterplot shows little or no association, a negative association, a linear association, a moderately strong association, or a very strong association:



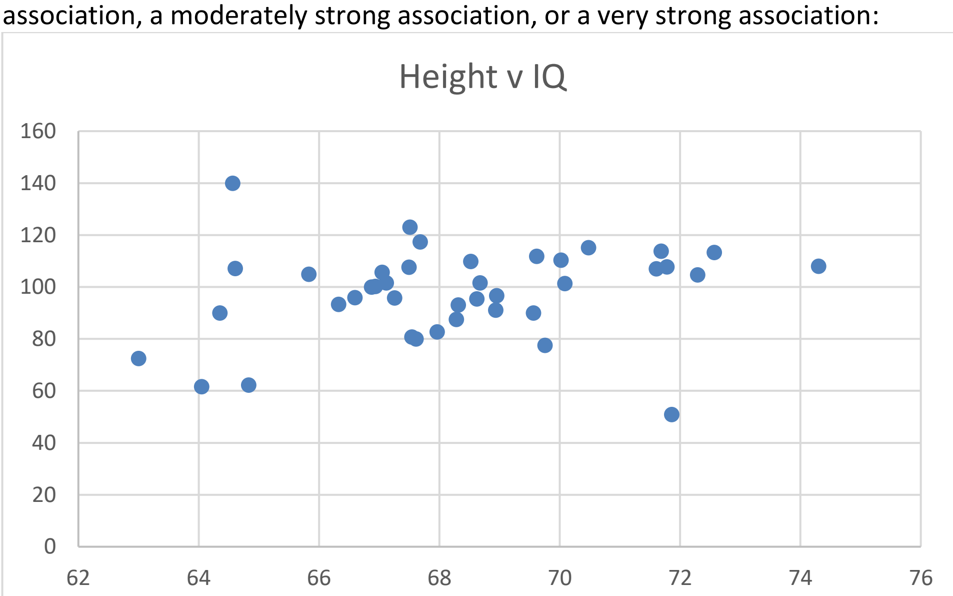
4. Determine whether the scatterplot shows little or no association, a negative association, a linear association, a moderately strong association, or a very strong association:



5. Determine whether the scatterplot shows little or no association, a negative association, a linear association, a moderately strong association, or a very strong association:



6. Determine whether the scatterplot shows little or no association, a negative association, a linear association, a moderately strong association, or a very strong association:



12.1 Answers

1. Very Strong Positive Linear Association/Correlation
2. Weak Positive Linear Association/Correlation
3. Strong Positive Linear Association/Correlation
4. Weak Positive Linear Association/Correlation
5. Moderate Negative Association/Correlation
6. Weak Positive Linear Association/Correlation

12.2 Pearson's Correlation Coefficient

To complete this section of homework watch Chapter Eleven, Lecture Example: 12.2 and 12.3.

7. Looking at GDP (Gross Domestic Product in trillions from 2010) and Female Life Expectancy (for 2010) for a selection of 93 countries, the following sum of squares were obtained. Use the sum of squares values to calculate Pearson's correlation coefficient (r) for the linear relationship between GDP and Female Life Expectancy.
$n = 93, SS_{xx}$ =28215.37461, SS_{xy} =2221.449175, SS_{yy} =5332.852258

8. Abdominal fat is dangerous for women (and men). It has been linked to heart disease. The data below lists waist measurements for women and their overall cholesterol level. Use the data below to find the correlation coefficient r for a linear relationship between waist circumference in women and their cholesterol level. $n = 14, SS_{xx} = 1860.943571, SS_{xy} = 8698.371429, SS_{yy} = 221713.4286$

Waist	67.2	82.5	66.7	93	82.6	75.4	73.6	81.4	99.4	67.7	100.7	99.3	85.7	85.7
Cholest.	264	181	267	384	98	62	126	89	531	130	175	280	149	112

9. The following list is from a used car lot. It includes the mileage and price for ten different 8-Series BMWs (Sept. 2025). Use the data (or the included sum of squares) to find Pearson's correlation coefficient (r) for the linear relationship between mileage and price for this series BMW.

Observation	Miles	Price
1	20	55990
2	37	56590
3	49	44590
4	51	43590
5	55	45990
6	6	71590
7	19	58990
8	20	53590
9	49	46990
10	10	54990
Sum	316	532900

$n = 10, SS_{xx} = 3{,}108.4, SS_{xy} = -1{,}214{,}000, SS_{yy} = 651{,}300{,}000$

10. Use Excel to find Pearson's correlation coefficient (r) for the linear relationship between number of years of experience as a chef and the time in minutes to prep (clean and chop vegetables) for a soup recipe.

Years	16	17	20	17	16	21	13	16	8	10	9	12	16
Prep Time	14.8	13.5	12.9	13	14.6	12.3	15.8	15.2	15.1	15	15.4	15.1	13.4

11. The calculated value of Pearson's correlation coefficient for the GDP (Gross Domestic Product in trillions from 2010)/Female Life Expectancy (for 2010) data turned out to be r = 0.181. Interpret this value of r. Do countries with higher GDPs have longer life expectancies for women?

12. The calculated value of Pearson's correlation coefficient for the waist measurement/cholesterol data turned out to be r = 0.428. Interpret this value of r. Does a smaller waist cause total cholesterol to drop lower?

13. The calculated value of Pearson's correlation coefficient for the used BMW mileage/price data turned out to be r = -0.853. Interpret this value of r. Is the price lower for BMWs with higher mileage than for BMWs with lower mileage?

14. The calculated value of Pearson's correlation coefficient for the years of experience/prep-time data turned out to be r = -0.791. Interpret this value of r. Do more experienced chefs finish prep work faster than chefs with less experience?

12.2 Answers

7. r = 0.181
8. r = 0.428
9. r = -0.853
10. r = -0.791
11. There is a weak positive linear relationship between these two variables, which implies that there is a slight trend for countries with higher GDPs to have longer life expectancies for woman. We cannot assume that higher GDP causes this increased life expectancy, but the two variables seem to move in the same direction (with the caveat that the relationship is weak). Perhaps the relationship would have been stronger had we used per capita GDP.
12. There is a moderate positive linear relationship between these two variables. A smaller waist size does not cause cholesterol to drop necessarily, but the association shows that smaller waist sizes tend to appear with less total cholesterol.
13. There is a strong negative linear relationship between these two variables. It seems that higher mileage BMWs tend to be cheaper than lower mileage ones. The negative relationship indicates that as miles go up on this type of BMW, the prices tend to move down.
14. There is a strong negative linear relationship between these two variables. It seems that more experienced chefs tend to finish the prep work faster. As experience goes up, the time to prep seems to trend down.

12.3 Linear Regression

To complete this section of homework watch Chapter Twelve, Lecture Examples 12.4, 12.4p2 12.5, and 12.6.

15. An educator wants to know if there is a relationship between the number of absences for a student and the student's final grade. Use the data below to find the least squares prediction line and to answer the questions below: VS

Absences	10	12	2	0	8	5
Final Grade	70	65	96	94	75	82

16. Based on this line, what would the average grade be for a student who missed 6 days of class?

: indicates the exercise has a video devoted to it in the corresponding section of STATSprofessor.com

17. What does x = 0 represent?

18. What average grade does the model predict for students who have perfect attendance?

19. The following set of data is randomly selected from a STA 2122 class from spring 2010. The list includes clicker points earned in class (clickers allow students to answer questions in class and to have their answers scored) and their final averages. Use the data to find the least squares prediction line: ($\sum x = 171, \sum x^2 = 4,737, \sum y = 648, \sum y^2 = 53,856, \sum xy = 15,013$) VS

Clicker points	32	11	34	41	16	15	7	15
Class Average	99	70	91	101	79	72	68	68

20. I plugged an entire class of 200 into SPSS to calculate the least squares line for the entire data set. The results were as follows: $\hat{y} = 0.512x + 70.196$. VS
   a. What does x = 0 represent here?
   b. What is the expected grade for students who do not earn any clicker points?
   c. What is the average grade for a student who has 20 clicker points?

21. The following table compares age at death and systolic blood pressure. Use the data to find the least squares prediction line and to answer questions 22 and 23:
($\sum x = 2678, \sum x^2 = 365,446, \sum y = 1166, \sum y^2 = 69,346, \sum xy = 153,860$)

BP	age at death		BP	age at death
158	46		134	59
157	46		157	59
157	49		150	62
160	49		117	62
131	50		126	64
138	51		109	65
160	53		120	68
122	54		111	69
123	57		107	71
122	58		119	74

22. What if any interpretation do we have for x = 0 in the model above?

23. What is the expected age at death for people with a systolic blood pressure of 140?

24. Use the data below to create the least squares prediction line and to predict the average weight for super models that are 69 inches tall.
$\sum x = 632, \sum x^2 = 44,399.5, \sum y = 1089, \sum y^2 = 132,223, \sum xy = 76,546$

Height	71	70.5	71	72	70	70	66.5	70	71
Weight	125	119	128	128	119	127	105	123	115

25. Use the data below to create the least squares prediction line for predicting the best finishing time of the New York City marathon given the temperature.
$\sum x = 478, \sum x^2 = 29,070, \sum y = 1,176.617, \sum y^2 = 173,068.7, \sum xy = 70,318.99$ VS

Temp	55	61	49	62	70	73	51	57
Time	145.283	148.717	148.3	148.1	147.617	146.4	144.667	147.533

26. Why do we say that the least squares line provides the "best fit" of any linear model?

27. What is the sum of all the errors ($\sum (y - \hat{y})$) made by any least squares line?

28. A researcher collected data involving the frequency of chirps made by a ground cricket, at various ground temperatures. He showed that there was a significant linear relationship between temperature and the frequency of chirps. The data he used is below. Would it be a good idea to use the model to estimate the average chirp frequency when the temperature was 51 degrees? Why or why not? VS

Temp	88.6	71.6	93.3	84.3	80.6	75.2	69.7	82	69.4	83.3	78.6	82.6	80.6	83.5	76.3
Chirps	20	16	19.8	18.4	17.1	15.5	14.7	17.1	15.4	16.2	15	17.2	16	17	14.1

12.3 Answers

15. $\hat{y} = -2.67x + 96.78$

16. $\hat{y} = -2.67(6) + 96.78 = 80.76$

17. It represents a student who missed zero class days = perfect attendance.

18. A 96.78%.

19. $\hat{y} = 1.07x + 58.04$

20. a. no clicker points b. 70.196 c. 80.436

21. $\hat{y} = -0.330x + 102.546$

22. There is no practical interpretation for x = 0, since that would imply the age at death for a person with zero blood pressure (which would basically mean your heart stopped working).

23. About 56 years old.

24. $\hat{y} = 3.88x - 152$; 116 lbs

25. $\hat{y} = 0.032x + 145.2$

26. Because the least square line has the minimum SSE of any linear model for a given set of data.

27. The SE is always equal to zero for the least squares line.

28. It is not a good idea since 51 degrees is far outside of the data range we used to create the model.