Mini-Lecture 4.1

Scatter Diagrams and Correlation

Objectives

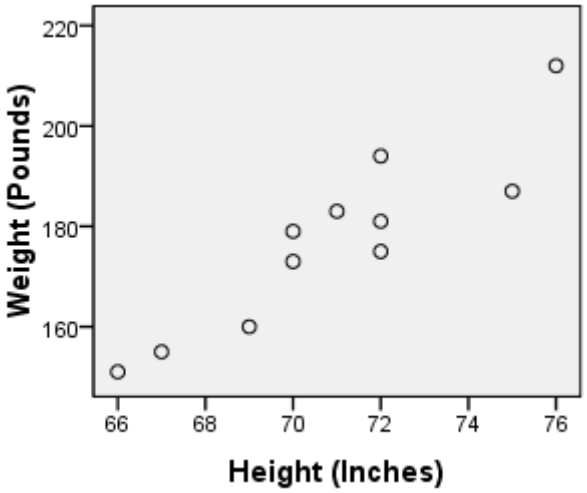
1. Draw and interpret scatter diagrams
2. Describe the properties of the linear correlation coefficient
3. Compute and interpret the linear correlation coefficient
4. Determine whether there is a linear relation between two variables
5. Explain the difference between correlation and causation

Examples

1. The heights and weights of 11 men between the ages of 21 and 26 were measured. The data are presented in the table below.

Height (Inches), x	75	66	71	67	70	72	72	70	72	76	69
Weight (Pounds), y	187	151	183	155	179	175	181	173	194	212	160

   a. Draw a scatter diagram of the data, treating the height as the explanatory variable.

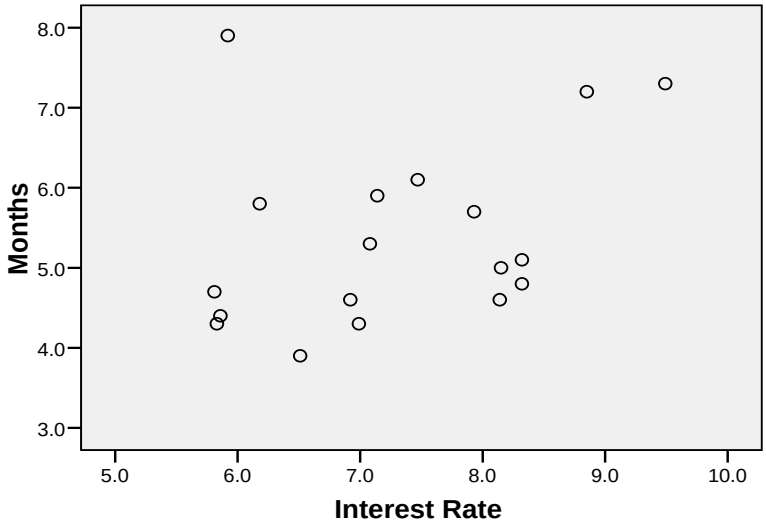


   b. Compute the linear correlation coefficient between the height and the weight of the men in the sample. (0.914)
   c. Comment on the type of relation that appears to exist between the height and the weight of the men based on the scatter diagram and the linear correlation coefficient. (There is a fairly strong linear association between the height and the weight.)

2. The interest rate for a 30-year conventional mortgage and the median number of months homes were listed for sale are given in the table below. The data represent the values for April of each year. (Source: economagic.com.)

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999
Interest, x	9.49	8.85	7.47	8.32	8.32	7.93	8.14	7.14	6.92
Months, y	7.3	7.2	6.1	5.1	4.8	5.7	4.6	5.9	4.6
Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Interest, x	8.15	7.08	6.99	5.81	5.83	5.86	6.51	6.18	5.92
Months, y	5	5.3	4.3	4.7	4.3	4.4	3.9	5.8	7.9

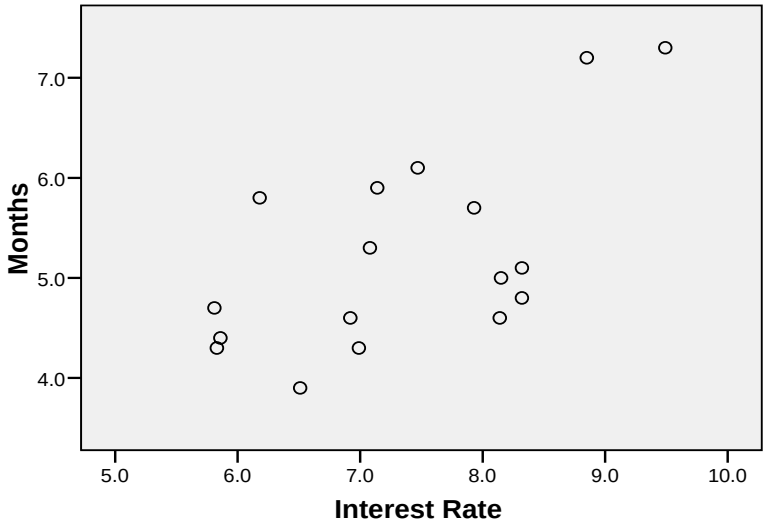
a. Draw a scatter diagram of the data, treating the interest rate as the explanatory variable.



b. Compute the linear correlation coefficient between the interest rate and the median number of months homes were listed for sale. (The correlation coefficient is 0.333)

c. Is there a linear relation between the interest rate and the median number of months homes are listed for sale? (There does not appear to be a linear relationship between these two variables. If there is a linear relationship, it is weak.)

d. Delete the observation for the year 2008. Repeat parts (a) through (c).



(With the data for the year 2008 omitted,the correlation coefficient is 0.624; there appears to be a linear relationship between the interest rate and the median number of months homes are listed for sale after omitting the data for 2008)

e. Does it seem that the data for the year 2008 is inconsistent with the data for the other years? If so, can you speculate on the cause of this irregularity? (Yes; in 2008, the United States was in a "mortgage crisis," and many investors were leery of real estate. The Federal Reserve lowered interest rates to stimulate growth in the economy, but many people were still reluctant or unable to purchase homes.)

3. A team of researchers studied the eating habits of n=502 adults in their mid-30s. (Source: Kvaavik E, Lien N, Tell GS and Klepp K-I (2005) Psychosocial explanatorys of eating habits among adults in their mid-30s: The Oslo Youth Study follow-up 1991–1999. International Journal of Behavioral Nutrition and Physical Activity (2)9.) For each of the subjects in their study, they observed: their fruit and vegetable intake (g/day), their whole grain intake (g/day), and their added sugar intake (recorded as a percent of the total energy intake.) The correlation coefficient for each pair of these variables was computed, and the results are given below. Interpret the strength and direction of these correlations.
   a. The correlation between the fruit and vegetable intake and the whole grain intake was 0.20. (There is a weak positive association between these variables, suggesting that people who eat a lot of fruits and vegetables also tend to eat a lot of whole grain foods.)
   b. The correlation between the fruit and vegetable intake and the added sugar intake was –0.14. (There is a very weak negative association between these variables, suggesting that people who eat a lot of fruits and vegetables tend to eat a lower amount of added sugar.)
   c. The correlation between the whole grain intake and the added sugar intake was –0.22. (There is a weak negative association between these variables, suggesting that people who eat a lot of whole grain foods tend to eat a lower amount of added sugar.)

d. Sketch a scatter diagram that could represent the general relationship between the fruit and vegetable intake (x-axis) and the whole grain intake (y-axis). (Answers will vary, but a gentle positive slope should be evident.)
e. Sketch a scatter diagram that could represent the general relationship between the whole grain intake (x-axis) and the added sugar intake (y-axis). (Answers will vary, but a gentle negative slope should be evident.)

Mini-Lecture 4.2

Least-Squares Regression

Objectives

1. Find the least-squares regression line and use the line to make predictions
2. Interpret the slope and the *y*-intercept of the least-squares regression line
3. Compute the sum of squared residuals

Examples

1. The heights and weights of 11 men between the ages of 21 and 26 were measured. The data are presented in the table below.

Height (Inches), x	75	66	71	67	70	72	72	70	72	76	69
Weight (Pounds), y	187	151	183	155	179	175	181	173	194	212	160

   a. Create at a scatter diagram to confirm that an approximately linear relationship exists between x and y. (See Example 1, Section 4.1)
   b. Find the least squares regression line, treating height, x, as the explanatory variable and weight, y, as the response variable. ($\hat{Y} = 5.371 - 203.580\ X$)
   c. Interpret the slope and intercept, if appropriate. (The slope is the estimated average increase in the weight of men in pounds, for each additional inch of height. It is not appropriate to interpret the Y-intercept.)
   d. Use the regression line to predict the weight of a man who is 73 inches tall? (188.5 pounds)

2. The interest rate for a 30-year conventional mortgage and the median number of months homes were listed for sale are given in the table below. The data represent the values for April of each year. In Mini-Lecture 4.1, Problem 2, we discovered that 2008 was a peculiar year. We will omit that observation in this analysis. (Source: economagic.com.)

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999
Interest, x	9.49	8.85	7.47	8.32	8.32	7.93	8.14	7.14	6.92
Months, y	7.3	7.2	6.1	5.1	4.8	5.7	4.6	5.9	4.6
Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Interest, x	8.15	7.08	6.99	5.81	5.83	5.86	6.51	6.18	--
Months, y	5	5.3	4.3	4.7	4.3	4.4	3.9	5.8	--

   a. Find the least squares regression line treating the interest rate as the explanatory variable and the median number of months as the response variable. ($\hat{Y} = 1.174 + 0.552\ X$)
   b. Interpret the slope and intercept, if appropriate. (The slope is the additional increase in the median number of months required to sell a home for each percentage point the interest rate increases. So, if the interest rate increases by 1%, the median duration required to sell a home

is expected to increase by about 0.552 months. It is not appropriate to interpret the intercept, since the interest rate cannot be 0% on a 30-year mortgage.)

c. Predict the median number of months required to sell a home if the interest rate is 5.92%. (This is the interest rate in 2008.) (4.4 months; you may want to compare this to the actual value given in Example 2, Section 4.1.)

Mini-Lecture 4.3

Diagnostics on the Least-Squares Regression Line

Objectives

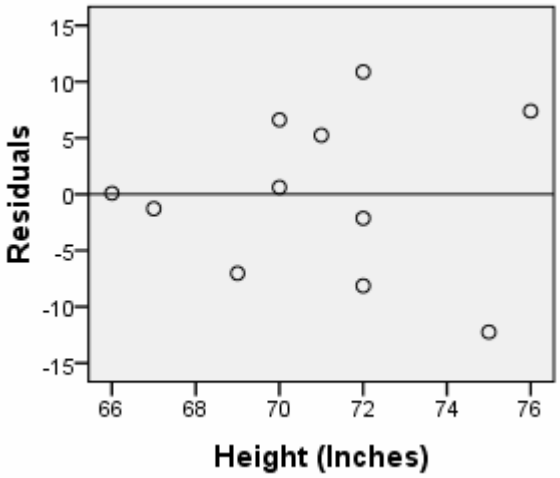
1. Compute and interpret the coefficient of determination
2. Perform residual analysis on a regression model
3. Identify influential observations

Examples

1. The heights and weights of 11 men between the ages of 21 and 26 were measured. The data are presented in the table below.

Height (Inches), x	75	66	71	67	70	72	72	70	72	76	69
Weight (Pounds), y	187	151	183	155	179	175	181	173	194	212	160

   a. Find the coefficient of determination, R^2. (0.836)
   b. Interpret the coefficient of determination. (83.6% of the variation in the weights is explained by the least-squares regression line.)
   c. Plot the residuals against height.

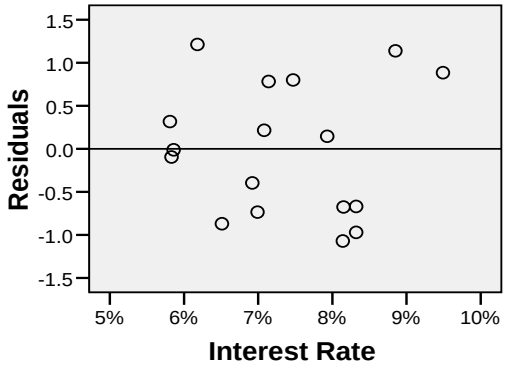


   d. Do you think the least-squares regression line is a good model? Why? (Answers may vary. A least-squares regression line is not appropriate when a pattern is apparent.)

2. The interest rate for a 30-year conventional mortgage and the median number of months homes were listed for sale are given in the table below. The data represent the values for April of each year. In Mini-Lecture 4.1, Problem 2, we discovered that 2008 was a peculiar year. We will omit that observation in this analysis. (Source: economagic.com.)

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999
Interest, x	9.49	8.85	7.47	8.32	8.32	7.93	8.14	7.14	6.92
Months, y	7.3	7.2	6.1	5.1	4.8	5.7	4.6	5.9	4.6
Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Interest, x	8.15	7.08	6.99	5.81	5.83	5.86	6.51	6.18	--
Months, y	5	5.3	4.3	4.7	4.3	4.4	3.9	5.8	--

a. Find the coefficient of determination, R^2. (0.390)
b. Interpret the coefficient of determination. (39% of the variation in the number of months to sell a home is explained by the least-squares regression line.)
c. Plot the residuals against the interest rate.



d. Do you think the least-squares regression line is a good model? Why? (The residual plot does not show any pattern, so the least-squares regression line seems to be appropriate.)

Mini-Lecture 4.4

Contingency Tables and Association

Objectives

1. Compute the marginal distribution of a variable
2. Use the conditional distribution to identify association among categorical data
3. Explain Simpson's Paradox

Examples

1. A professor at a community college in New Mexico conducted a study to assess the effectiveness of delivering an introductory statistics course via traditional lecture-based method, online delivery (no classroom instruction), and hybrid instruction (online course with weekly meetings) methods, the grades students received in each of the courses were tallied.

	Traditional	Online	Hybrid
A	36	39	24
B	52	55	66
C	57	68	90
D	46	38	41
F	46	54	31

a. Construct a frequency marginal distribution.

	Traditional	Online	Hybrid	Total
A	36	39	24	99
B	52	55	66	173
C	57	68	90	215
D	46	38	41	125
F	46	54	31	131
Total	237	254	252	743

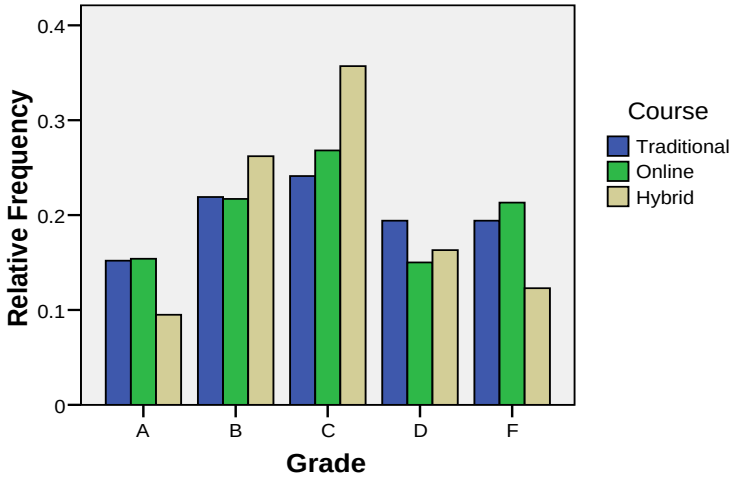
b. Construct a relative frequency marginal distribution.

	Traditional	Online	Hybrid	
A	36	39	24	0.133
B	52	55	66	0.233
C	57	68	90	0.289
D	46	38	41	0.168
F	46	54	31	0.176
	0.319	0.342	0.339	1.000

c. What percentage of students in the study received a grade of C? (28.9%)
d. What percentage of students in the study was in the online course? (34.2%)
e. Construct a conditional distribution by type of course traditional, online or hybrid.

	Traditional	Online	Hybrid
A	0.152	0.154	0.095
B	0.219	0.217	0.262
C	0.241	0.268	0.357
D	0.194	0.150	0.163
F	0.194	0.213	0.123

f. Draw a bar graph of the conditional distribution you found in Part e.



g. Is the type of course associated with students' grades? (Answers may vary, but there does appear to be a relationship between the type of course and the grades earned.)

2. The following table summarizes the percentage of the jury pool in New Zealand who belong to the Maori ethnic group. (Source: Westbrooke, Ian (1998). Simpson's Paradox: An example in a New Zealand survey of jury composition. *Chance, New Directions for Statistics and Computers*, 11, 40-42.)

District	Percentage Maori ethnic group Eligible Population (Age 20-64)	Jury Pool
Whangarei	17.0%	16.8%
Auckland	9.2%	9.0%
Hamilton	13.5%	11.5%
Rotorua	27.0%	23.4%
Gisborne	32.2%	29.5%
Napier	15.5%	12.4%
New Plymouth	8.9%	4.1%
Palmerston North	8.9%	4.3%
Wellington	8.7%	7.5%
Nelson	3.9%	1.7%
Christchurch	4.5%	3.3%
Dunedin	3.3%	2.4%
Invercargill	8.4%	4.8%
All Districts	9.5%	10.1%

Notice that in every district, the Maori people appear to be underrepresented in the jury pool. However, when all the districts are combined, the Maori ethic group appears to be overrepresented in the jury pool. How is this possible? To address this issue, consider the cities Rotorua and Nelson. The following table gives the counts of the total number of people of Maori and non-Maori descent in the eligible population and the jury pool for each city.

	Eligible Population		Jury Pool	
City	Maori	Non-Maori	Maori	Non-Maori
Rotorua	8,889	24,009	79	258
Nelson	1,329	32,658	1	56
Combined	10,218	56,667	80	314

a. Compute the percentage of the eligible population in Rotorua that are of the Maori ethnic group. (8889 / 32898 = 27.02%; compare this value to the value observed in the first table.)
b. Compute the percentage of the jury pool in Rotorua that are of the Maori ethnic group. (79 / 337 = 23.44%; compare this value to the value observed in the first table.)
c. Compare your answers to parts a and b. (The jury pool contains a lower proportion of Maori people than the general population.)

d. Compute the percentage of the eligible population in Nelson that are of the Maori ethnic group. (3.91%; compare this value to the value observed in the first table.)
e. Compute the percentage of the jury pool in Nelson that are of the Maori ethnic group. (1.75%; compare this value to the value observed in the first table. Note the value in the table should have been rounded to 1.8%, but this table was reproduced directly from the source.)
f. Compare your answers to parts d and e. (The jury pool contains a lower proportion of Maori people than the general population.)
g. Compute the percentage of the eligible population for both cities combined that are of the Maori ethnic group. (15.28%)
h. Compute the percentage of the jury pool for both cities combined that are of the Maori ethnic group. (20.30%)
i. Compare your answers to parts g and h. (The jury pool contains a higher proportion of Maori people than the general population.)
j. Discuss the results from parts c, f, and i. (This is an example of Simpson's Paradox. Because the jury pool in Rotorua is so large, it contributes a large number of Maori individuals to the combined jury pool. This leads to a larger overall percentage of Maori individuals in the jury pool than in the general population.)

Mini-Lecture 4.5 (on CD)

Nonlinear Regression: Transformations

Objectives

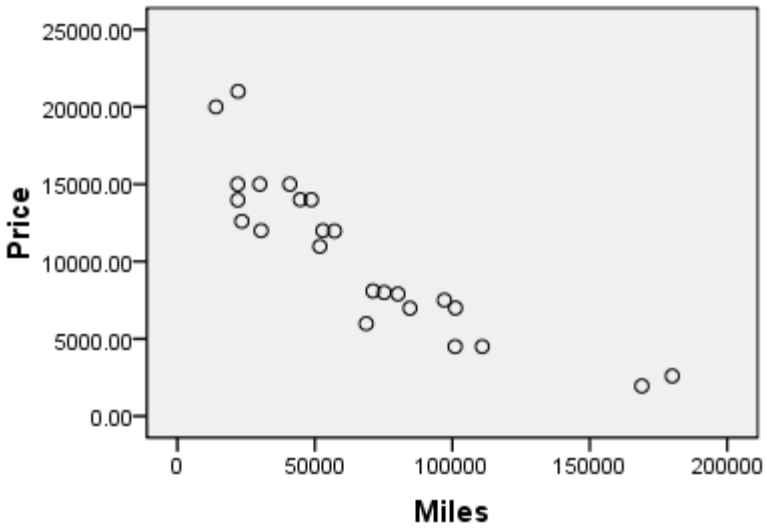
1. Change exponential expressions to logarithmic expressions and logarithmic expressions to exponential expressions
2. Simplify expressions containing logarithms
3. Use logarithmic transformations to linearize exponential relations
4. Use logarithmic transformations to linearize power relations

Examples

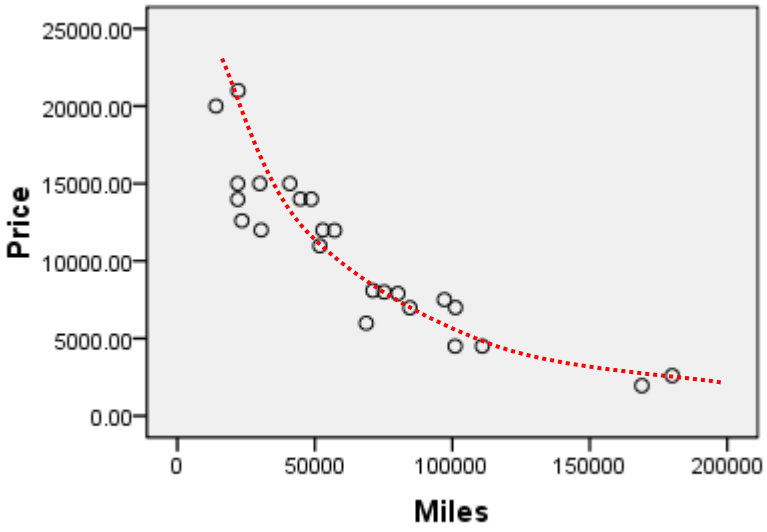
1. The asking price (in dollars) and the number of miles on the odometer were recorded for 24 randomly selected used Chrysler Sebring convertibles listed for sale in Atlanta, Georgia. The data are given below. Use this information to answer the following questions. (Source: www.cars.com)

Price	Miles	Price	Miles
12599	23431	4498	110887
1950	168954	13980	21930
10980	51771	13998	44680
14998	29965	11998	30436
2600	180000	13995	48656
19996	14017	6977	84566
5980	68677	6995	101126
7499	97101	11975	57125
7998	75189	7900	80200
8095	71085	11998	52880
14995	21907	14998	40877
20999	22065	4500	101000

   a. Draw a scatter diagram, treating the number of miles on the odometer as the explanatory variable.

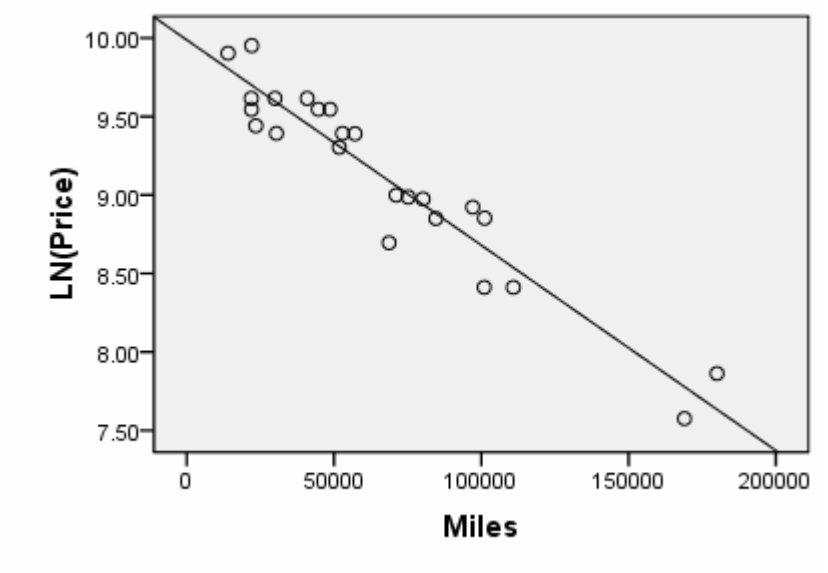


b. Does simple linear regression seem appropriate for these data? Justify your answer.



(No; there is distinct curvature in the scatter diagram)

c. Determine the logarithm of the y-values so that $Y = \log y$. Draw a scatter diagram of the transformed data. What do you observe?



(There is a linear relationship in the data)

d. Find the equation of the least-squares regression line for the transformed data. ($\hat{Y} = 9.988 - 0.00001308\,X$)

e. Find the exponential equation of best fit. ($\hat{y} = 21764\, e^{-0.00001308\,X}$)

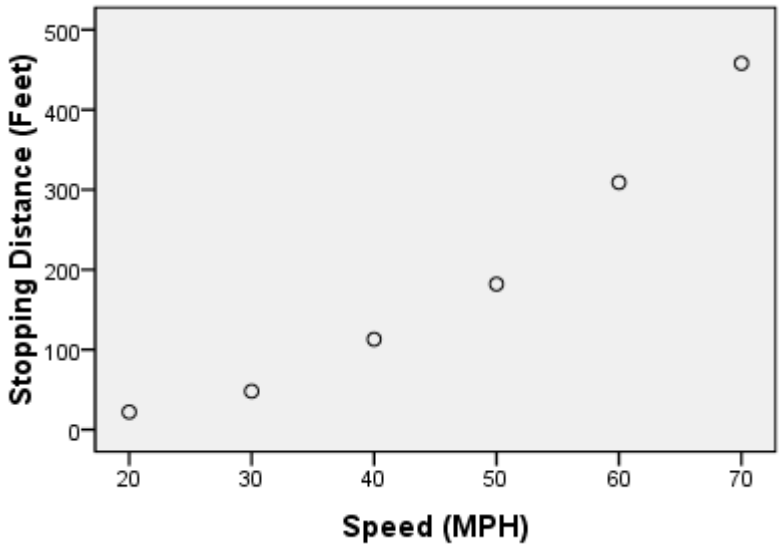
f. Use the exponential equation of best fit to predict the asking price for a Sebring convertible that has 23431 miles on the odometer. ($16019)

g. Compare this to the asking price for the first car listed in the data set. Interpret the result. (This value is higher than the asking price of $12599, suggesting that this car is priced lower than we would have expected. This car has the potential to be a bargain.)

2. The distance required to stop a car depends on the speed at which it is traveling. A group of students measured the stopping distance (in feet) required for a car traveling at speeds between 20 and 70 miles per hour (MPH). Their results are given below.

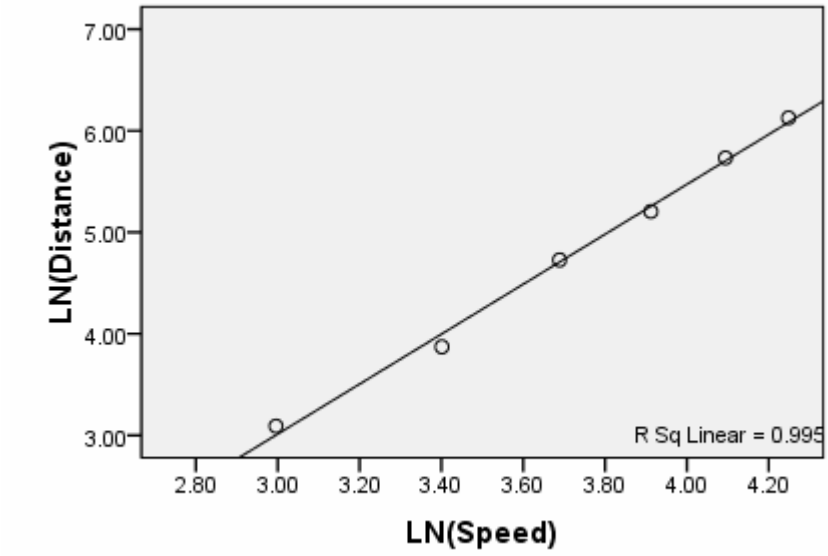
Speed (MPH)	Stopping Distance (Ft)
20	22
30	48
40	113
50	182
60	309
70	458

a. Draw a scatter diagram treating speed as the explanatory variable.



b. Does simple linear regression seem appropriate for these data? Justify your answer. (No; there is distinct curvature in the scatter diagram)

c. Determine the logarithm of both the x- and y-values so that $X = \log x$ and $Y = \log y$. Draw a scatter diagram of the transformed data. What do you observe?



(There is a linear relationship in the data)

d. Find the equation of the least-squares regression line for the transformed data. ($\hat{Y} = -4.365 + 2.459\,X$)

e. Find the power equation of best fit.
($\hat{y} = 0.0127\,x^{2.459}$)

f. Use the power equation of best fit to predict the stopping distance required for a car traveling at 65 miles per hour. (364.6 ft)

g. Does your result to part (f) seem reasonable, given the data for 60 and 70 mph? (yes)