Lab 2

Objectives: Scatter Plots, Correlation, and Least-Squares Regression

Summary of Commands

See Chapter 2 Appendix for detailed instructions

- Graph → Scatterplot
- Stat → Basic Statistics → Correlation
- Stat → Regression → Regression
- Stat → Regression → Fitted Line Plot

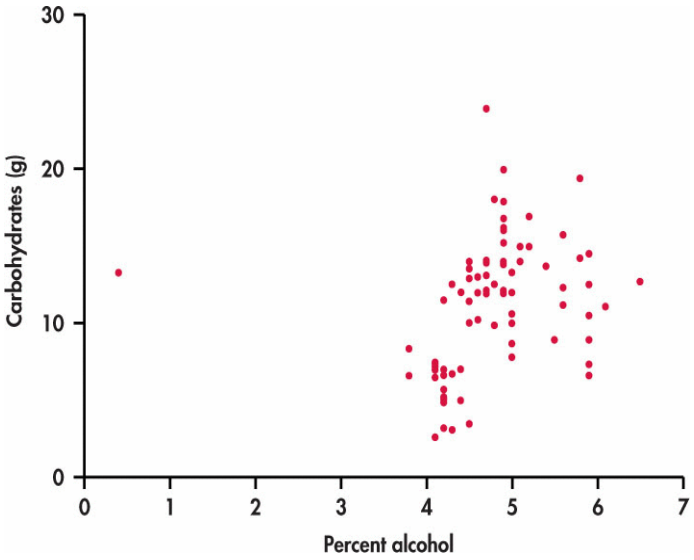


Figure 2.10

Problem 1 (**2.18, 2.19 & 2.35 Data Set: BEER**)

Beer100.com advertises itself as "Your Place for All Things Beer." One of their "things" is a list of 86 domestic beer brands with the percent alcohol, calories per 12 ounces, and carbohydrates (in grams). **Figure 2.10** gives a scatterplot of carbohydrates versus percent alcohol. One of the points is an outlier.

(a) Compute the correlation between the percent alcohol and the carbohydrates (**with the outlier**).

- Minitab: Stat → Basic Statistics → Correlation

(b) Use the data file to find the outlier brand of beer, remove the outlier from the data set and re-compute the correlation.

(c) Generate a scatterplot (with the least-squares regression line) of *carbohydrates* vs *percent alcohol* using the data **without the outlier**.

- Minitab: Graph → Scatterplot

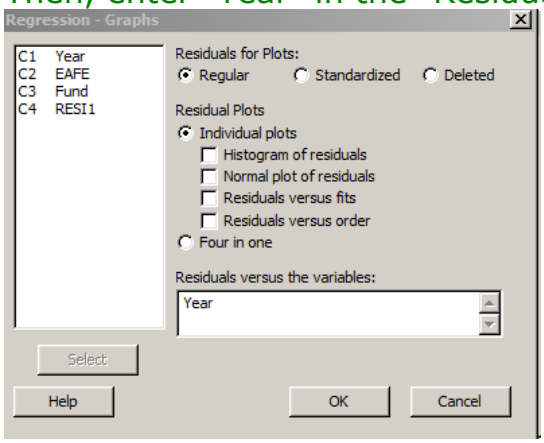
(d) Make a scatterplot (with the least-squares regression line) of *calories* vs *percent alcohol* using the data file **without the outlier**.

Problem 2 (2.46 & 2.71 Data Set: MUTUALFUNDS)
Many mutual funds compare their performance with that of a benchmark, an index of the returns on all securities of the kind that the fund buys. The Vanguard International Growth Fund, for example, takes as its benchmark the Morgan Stanley Europe, Australasia, Far East (EAFE) index of overseas stock market performance.
Here are the values of a $10,000 investment in the Vanguard Fund made in 1998 and the hypothetical value of a $10,000 investment in the benchmark (EAFE):

Year	EAFE	Fund	Year	EAFE	Fund
1998	10,000	10,000	2003	10,029	9,615
1999	12,110	11,787	2004	12,739	12,465
2000	11,440	10,939	2005	13,867	13,523
2001	9,365	8,846	2006	17,879	17,336
2002	8,378	7,740	2007	21,835	20,335

(a) Make a scatterplot (with the least-squares line) suitable for predicting fund returns from EAFE returns.
(b) Find the equation of the least-squares line.
- Minitab: Stat → Regression → Regression

(b) Plot the residuals versus year and observe the pattern.
In Stat → Regression → Regression, click on "Graphs" option. Then, enter "Year" in the "Residuals versus the variables:" box.



(c) What proportion of the variation in EAFE is explained by **year**? (Hint: You need to run "correlation" or "regression" with x=year, y=EAFE)

Problem 3 (**2.84 – 2.88 Move your business here.** **Data Set: OASIS**) City officials use a variety of tactics to encourage businesses to open offices in their city. One characteristic of a city that is viewed as desirable is open public space within the city limits. The New York City Open Accessible Space Information System Cooperative (OASIS) is an organization of public- and private-sector representatives that has developed an information system designed to enhance the stewardship of open space. Below are data from the OASIS Web site for 12 large U.S. cities. The variables are population in thousands and total park or open space within city limits in acres.

City	Population	Open space
Baltimore	651	5,091
Boston	589	4,865
Chicago	2,896	11,645
Long Beach	462	2,887
Los Angeles	3,695	29,801
Miami	362	1,329
Minneapolis	383	5,694
New York	8,008	49,854
Oakland	399	3,712
Philadelphia	1,518	10,685
San Francisco	777	5,916
Washington, D.C.	572	7,504

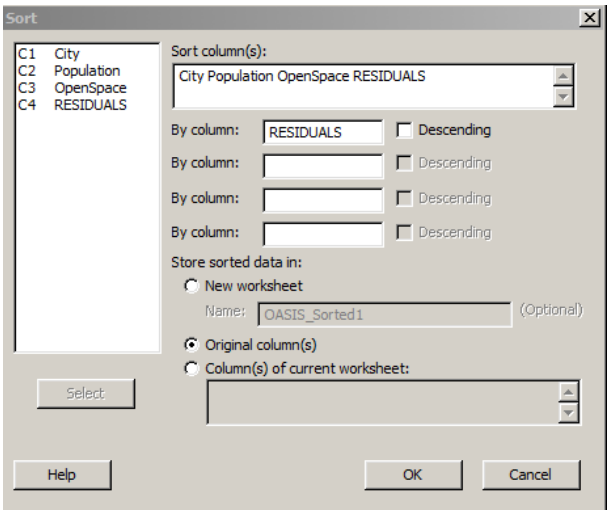
(a) Make a scatterplot of the data (with the least square regression line) using population as the explanatory variable and open space as the response variable.
(b) Find the least-squares regression line.
(c) What proportion of the variation in open space is explained by population?

Problem 4 (**2.85 & 2.86 Data Set: OASIS**) One way to compare cities with respect to the amount of open space that they have is to use the residuals from the regression analysis that you performed in the previous exercise. Cities with positive residuals are doing better than predicted, while those with negative residuals are doing worse.

(a) Find the residual for each city and make a table with the city name and the residual, ordered from worst to best by the size of the residual.

Minitab Tricks – Saving Residuals in the Table:
1) Click "Storage" when running regression
2) Select "Residuals"
3) A new column containing residuals will be added to your worksheet.

Minitab Tricks – Sorting the Data:
Use Data->Sort, then follow the directions (see picture below).



After Sorting the Data, use
Data->Display Data
The sorted data will be printed in the Minitab output window.
Then, right click the mouse and send the data to a word document.

(b) Find the data point corresponding to New York City. Is this point an outlier? Is it influential? (To see if New York City is influential, you need to re-run stat->regression without the data from New York City)

Problems 5 (**2.87 Open space per person. Data Set: OASIA**)
Open space in acres per person is an alternative way to report open space. ***Divide open space by population*** to compute the value of this variable for each city.

Minitab Tricks -- Creating a New Variable:

1) Create a new column named "OS per Person".
2) Choose "Editor->Formulas->Assign Formula to Column"
3) Enter appropriate formula for that new column.

(a) Make a scatterplot of the data (with the least square regression line) using **population** as the explanatory variable and **open space per person** as the response variable.
(b) Find the least-squares regression line.
(c) What proportion of the variation in **open space per person** is explained by **population**?