STA 250F, Mid-Term Test, 25/26 October 2000, 110 minutes

Write your answers on the answer sheet provided. Hand in the answer sheet **and this question sheet** at the end of the test. Answer sheets with no question sheet will not be marked.

For yes/no, true/false, and multiple choice questions, circle the correct answer. For questions with a numerical answer, write the number in the space provided. These questions will be marked as correct or incorrect, with no part marks, though for a few questions there is more than one correct answer (which are only slightly different).

The final discussion question, write your answer on the back of the answer sheet. Part marks will be possible for this question.

Use of pencil is recommended. You may use a calculator. No books or notes are allowed.

You should have a copy of Table B.2 from the textbook.

The questions are worth the numbers of marked indicated, out of 100 total.

Here is a stem-and-leaf plot of the weights in kilograms of 25 university students:

```
  The decimal point is 1 digit to the right of the |

   2 | 789
   3 | 033
   4 | 12347
   5 | 036
   6 | 389
   7 | 36
   8 | 9
   9 | 279
  10 |
  11 | 4
  12 |
  13 | 0
```

1) [2 marks] What is the median of this data?

2) [2 marks] What is the first quartile of this data? (There are several reasonable answers to this question, corresponding to slightly different definitions of the quartiles.)

3) [2 marks] Which of the following numbers is closest to the mean of this data? (You should be able to answer without actually calculating the mean.)

   (a) 53 (b) 97 (c) 61 (d) 48

4) [2 marks] Which of the following is the best description of the shape of the distribution of this data?

   (a) Close to normal (b) Skewed right (c) Skewed left (d) Uniform

Two six-sided dice (one red, one green) are rolled, and at the same time, a fair coin is flipped. The two dice and the coin do not affect each other. The following events are defined:

A is the event that the two dice show the same number
B is the event that the coin lands heads and both dice show the number 5
C is the event that the red die shows the number 1

5) [2 marks] How many outcomes are in the sample space describing this situation?

6) [1 mark] Are the events A and B mutually exclusive (disjoint)?

7) [1 mark] Are the events B and C mutually exclusive (disjoint)?

8) [1 mark] Are the events A and C mutually exclusive (disjoint)?

9) [1 mark] Are the events A and B independent?

10) [1 mark] Are the events B and C independent?

11) [1 mark] Are the events A and C independent?

12) [2 marks] What is the probability of event B?

Consider the following five measurements of the heights of corn plants, in inches:

62 57 63 63 65

13) [2 marks] What is the sample mean of this data set?

14) [2 marks] What is the sample variance of this data set?

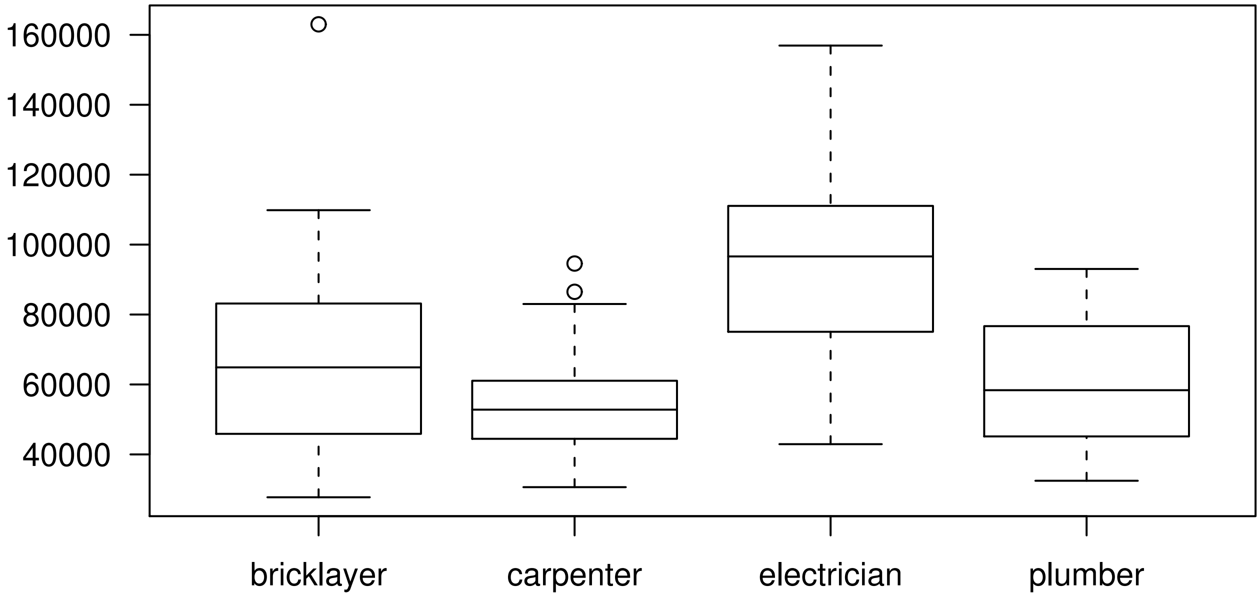
15) [2 marks] What is the sample standard deviation of this data set?

16) [1 mark] What would have been the sample mean of this data set if the heights had been measured in feet rather than inches? (Note: 1 foot = 12 inches.)

17) [1 mark] What would have been the sample variance of this data set if the heights had been measured in feet rather than inches?

18) [1 mark] What would have been the sample standard deviation of this data set if the heights had been measured in feet rather than inches?

A survey of incomes of workers in various trades was done by selecting a simple random sample of 500 people belonging to a trade association and sending each a questionnaire asking what their income was in the previous year. Of these 500 people, 157 returned the questionnaire — 12 bricklayers, 90 carpenters, 20 electricians, and 35 plumbers. Here are side-by-side boxplots of the incomes in dollars for the people who responded, arranged by trade:



19) [2 marks] For which trade was the median income (in this sample) the highest?

(a) bricklayer (b) carpenter (c) electrician (d) plumber

20) [2 marks] True or false: We can see from these boxplots that for each of the four trades, the distribution of income has only one mode.

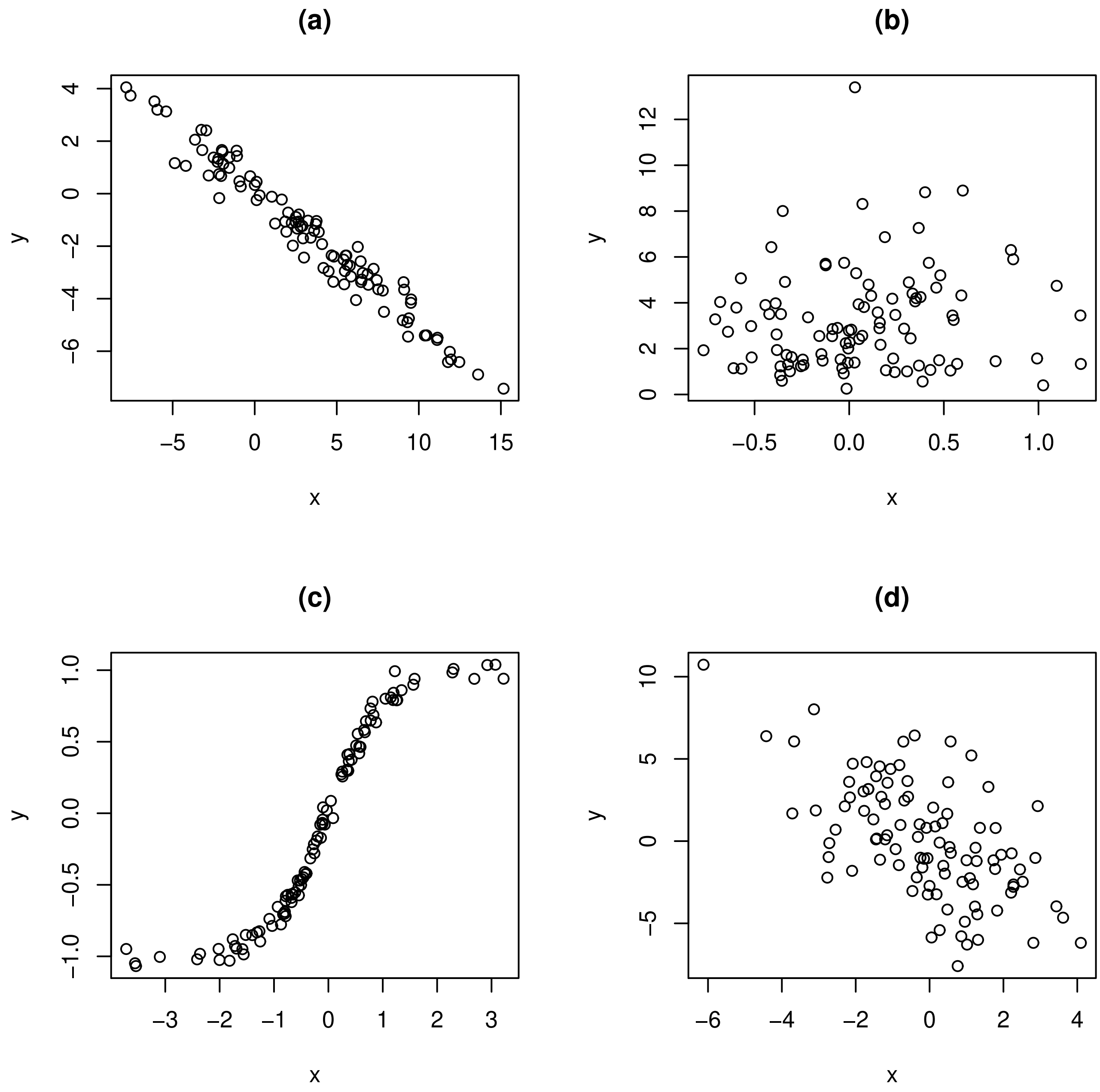
21) [2 marks] True or false: At least half of the electricians had incomes that were larger than the highest income earned by a plumber

22) [2 marks] True or false: The one bricklayer who earned over \$160000 dollars in income may be having a large effect on the median income for bricklayers in this sample. Before concluding that the median income for bricklayers is greater than the median income for carpenters, it is essential that the median income for bricklayers be recomputed with this outlier removed from the data set.

23) [2 marks] True or false: The inter-quartile range (IQR) for the bricklayers is much bigger than the IQR for the carpenters because the number of bricklayers (12) is much smaller than the number of carpenters (90), so it is expected that the answer found for the carpenters will be more precise.

24) [2 marks] True or false: Because we obtained responses from quite a few carpenters (90), we can be quite sure that the responses we have are a good indication of the incomes in the whole population of carpenters. For example, we can be quite sure that the median income for all carpenters is less than \$60000.

Answer the next four questions by giving the letter (a, b, c, or d) identifying one of the following scatterplots of variables x and y:



25) [2 marks] For which scatterplot is the Pearson's correlation close to zero?

26) [2 marks] For which scatterplot is the Pearson's correlation approximately -0.98?

27) [2 marks] For which scatterplot is the Pearson's correlation approximately -0.6?

28) [2 marks] For which scatterplot would you expect the Spearman's correlation to be larger (in absolute value) than the Pearson's correlation?

Six fair six-sided dice have been specially made so that they show the number 0 on two of their six sides, the number $+1$ on two other sides, and the number -1 on the remaining two sides. When one of these dice is thrown, it is equally likely to show 0, $+1$, or -1.

Let the random variable X be defined to be the sum of the numbers showing on all six of these dice after they are thrown. For example, if the six dice show the numbers

$$+1, \quad -1, \quad 0, \quad -1, \quad -1, \quad 0$$

the value of X will be -2.

29) [3 marks] What is the mean of the random variable X?

30) [3 marks] What is the variance of the random variable X?

31) [3 marks] What is the standard deviation of the random variable X?

A new computer software company has obtained five contracts to develop specialized programs for customers. The company hopes that it will be able to complete at least two of these contracts on time, so that it can use them as examples of its good work. However, the company thinks that each contract has only a 40% chance of being completed on time. They are interested in what the probability is that **at least two** contracts will be completed on time.

32) [2 marks] Which of the following would be a good reason to think that whether one contract is completed on time is **not** independent of whether one of the other contracts is completed on time? (Choose only one.)

   (a) The contracts are for different types of programs: eg, one is for an accounting program, and another is for a program to manage energy usage in buildings.

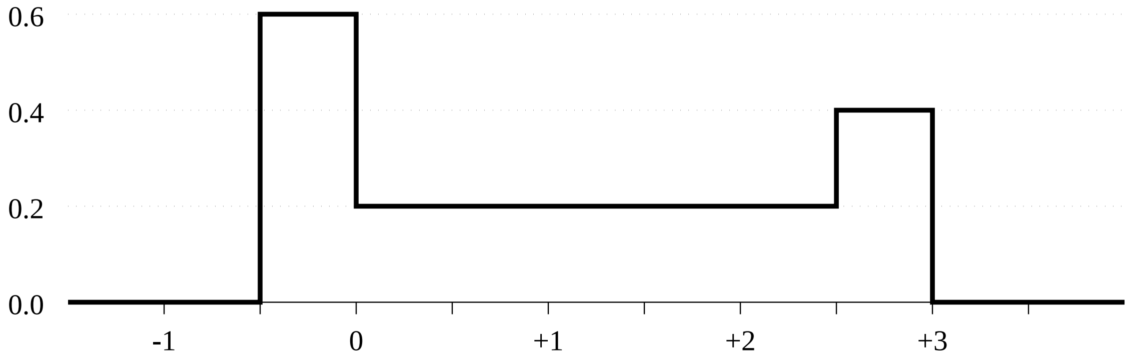
   (b) If the company completes one of the contracts early, the programmers who were working on that contract could help out on another contract.

   (c) Many of the things that need to be done for a contract depend on other things being done first. For example, the program design has to be completed before the user manual can be written.

   (d) Many programmers do stupid things if they aren't supervised closely, and they therefore cannot be trusted to work independently.

33) [5 marks] Suppose we do assume that whether one contract is completed on time is independent of whether another contract is completed on time. What is the probability that at least two of the five contracts will be completed on time?

Suppose the random variable X has the probability density function shown below:



34) [3 marks] What is the numerical value of $P(X \leq 0)$?

35) [3 marks] What is the numerical value of $P(1 \leq X \leq 3)$?

---

36) [2 marks] True or false: If a least-squares regression of a response variable y on a predictor variable x results in a value for r^2 that is close to zero, we can be sure that there is no relationship of any sort between x and y (or at least, we can be sure that any such relationship is very weak). In other words, a value for r^2 near zero indicates that finding out the value of x for a unit can't help you to predict the value of y for that unit.

37) [2 marks] Suppose we do a regression with the weight in kilograms of adults in Toronto as the response variable and the height of these people in feet as the predictor variable, and that the resulting value for r^2 is 0.7 and the standard deviation of the residuals is 1.1. If we convert the data so that weight is measured in pounds (1 kilogram = 2.2 pounds) and height in inches (1 foot = 12 inches), what will be the new value for the residual standard deviation?

38) [2 marks] Referring to the question above, what will be the value of r^2 for the regression with weight in pounds and height in inches?

39) [2 marks] True or false: If r^2 is greater than zero, the least squares regression line for y on x will pass **below** the point $(\bar{x}, \bar{y})$, where $\bar{x}$ and $\bar{y}$ are the sample means of x and y. The regression line will pass through the point $(\bar{x}, \bar{y})$ only if r^2 is zero.

40) [4 marks] Suppose we are interested in the question of whether or not regular exercise helps students learn better. Which of the following is an *experiment*, from which one could draw some conclusions about this question without worrying about the confounding influences of known or unknown lurking variables? (Choose only one.)

(a) We select a simple random sample of 100 U of T students, and ask them how many hours per week they exercise, and what their Grade Point Average was for the previous term. We then look at the relationship between these two variables.

(b) We select a simple random sample of 100 U of T students, and then randomly divide this sample into two groups of 50 students each. We carefully monitor how much the students in one of these groups exercise each week (looking at records for the athletic centre, for instance). We are careful to not let these students know that they are being monitored. We do nothing with the other group of students. After one school term, we find out the Grade Point Averages that term for all 100 students, and look at how the Grade Point Averages differ between students who exercised a lot, those who exercised only a litte, and those whose exercise was not monitored.

(c) We select a simple random sample of 100 U of T students, and then randomly divide this sample into two groups of 50 students each. We give the students in one group a booklet that explains the many benefits of exercise, and lists all the fun athletic activities that students at U of T can participate in. We do nothing with the other group of students. After one school term, we find out the Grade Point Averages that term for all 100 students, and look at how the Grade Point Averages differ between the students who were given the booklet and those who were not.

(d) We select a simple random sample of 100 U of T students, and ask each student whether they will agree to exercise at least 10 hours per week during the next term, in return for being paid \$100 per week. As it happens, 45 of the students agree to do this. After the term is over, we find out the Grade Point Averages that term for all 100 students, and compare the Grade Point Averages of the 45 students who agreed to exercise with the Grade Point Averages of the 55 students who did not agree.

---

41) [6 marks] Suppose that we know from past experience that the the total amount of gasoline sold in Toronto on a Saturday is normally distributed with mean of 300,000 litres and standard deviation of 40,000 litres. What is the probability that next Saturday more than 370,000 litres of gasoline will be sold?

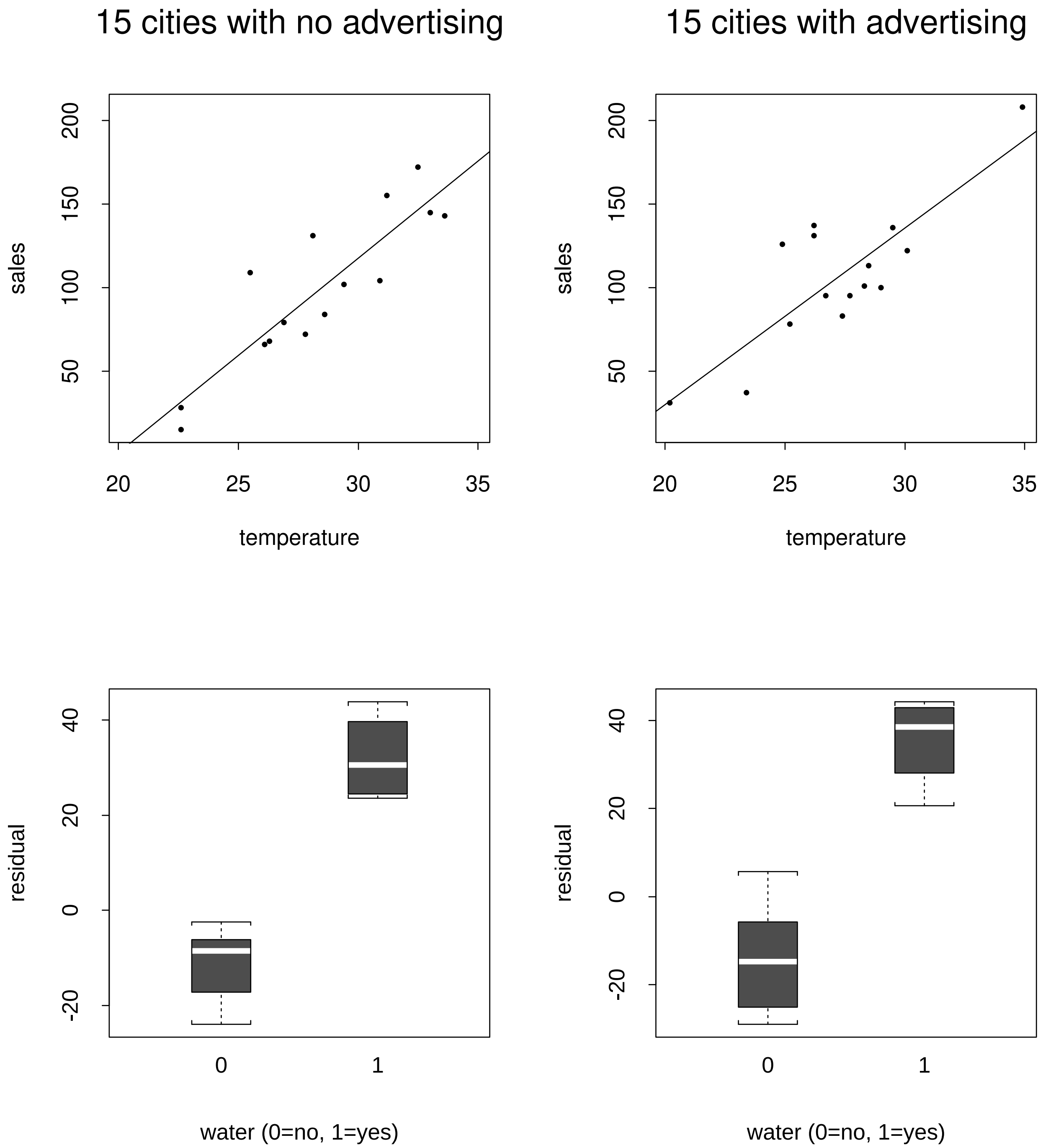
A swimwear company that has stores in 30 cities. One summer, they decide to conduct an advertising campaign in 15 of these cities, and do no advertising in the other 15 cities. They select the cities in which they will advertise randomly. They then see how many swim suits they sell in each city, and from that they calculate the number of swim suits sold per 100,000 people in each city. They also know the average summer maximum temperature for each city, and whether or not the city is near the ocean or a large lake. This gives a total of four variables that are available for each of the 30 cities:

ad	0 if no advertising was done, 1 if advertising was done
water	0 if no big lake or ocean near, 1 if one is near
temperature	average daily maximum temperature in summer (Celcius)
sales	number of swim suits sold per 100,000 people

Separate regressions were done with sales as the response and temperature as the predictor for the 15 cities with advertising and the 15 cities with no advertising. Here are the results from MINITAB:

```
REGRESSION USING DATA FOR THE 15 CITIES WHERE THERE WAS NO ADVERTISING

The regression equation is
sales = - 231 + 11.6 temperature

Predictor        Coef       StDev          T        P
Constant      -231.39       48.90      -4.73    0.000
temperature     11.630       1.714       6.79    0.000

S = 22.22       R-Sq = 78.0%     R-Sq(adj) = 76.3%


REGRESSION USING DATA FOR THE 15 CITIES WHERE ADVERTISING WAS DONE

The regression equation is
sales = - 181 + 10.6 temperature

Predictor        Coef       StDev          T        P
Constant      -181.12       56.01      -3.23    0.007
temperature     10.558       2.044       5.17    0.000

S = 25.40       R-Sq = 67.2%     R-Sq(adj) = 64.7%
```

Here are scatterplots for these two regressions, followed by side-by-side boxplots of the residuals for the two regressions according to whether or not the city is by the ocean or a large lake:



42) [13 marks] Discuss what you can and cannot conclude from the results of these regressions and from the plots above. Address the question of whether or not advertising seems to increase sales. Discuss whether or not possible lurking variables are a problem in trying to determine whether advertising increases sales. Make what comments you can on how well sales can be predicted based on the three other variables, and on anything else that seems interesting.