Statistics 2000, Section 001, Quiz 2 (200 Points)

November 2, 2001, Dr. Jürgen Symanzik

Your Name: _____

First look at all 4 questions. Then start with the question that looks easiest to you. Continue with a more difficult question. Try to answer as many questions as possible in these 50 minutes.

Note that you will obtain at least partial credit if you indicate a correct formula but your final result is incorrect. If you just rely on your calculator without indicating the formula that should be used and your result is incorrect, you will obtain no credit at all for this part of a question.

Question 1: Linear Regression & Correlation (**50 Points**)

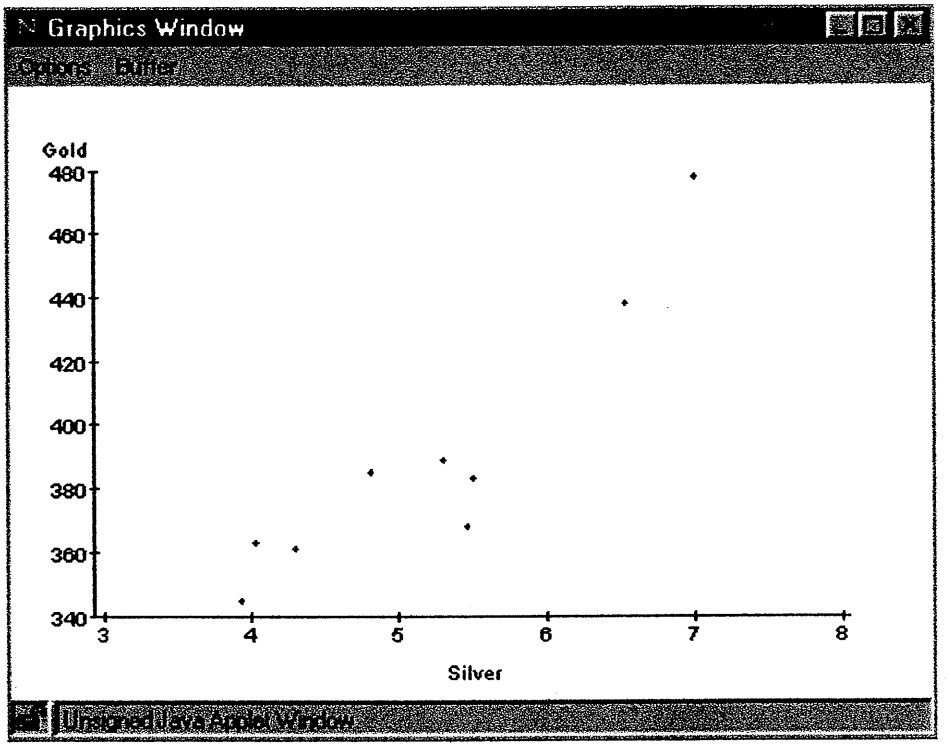
The average prices (in dollars) per ounce of gold and silver for the years 1986 through 1994 are given below as they appear in the WebStat main window. (Source: U.S. Bureau of Mines.) Note that var1 = Year, var2 = Price of Gold, var3 = Price of Silver, and var4 = Residual.

N WebStat 2.0

WebStat Data Stat Graphics Help

	var1	var2	var3	var4	var
1	1986	368	5.47	-31.372032	
2	1987	478	7.01	22.638151	
3	1988	438	6.53	0.089522004	
4	1989	383	5.5	-17.462744	
5	1990	385	4.82	9.260033	
6	1991	363	4.04	15.618511	
7	1992	345	3.94	1.2542136	
8	1993	361	4.3	4.165685	
9	1994	389	5.3	-4.1913385	
10					

The scatterplot of Gold vs Silver is shown below:



We used WebStat to fit a simple linear (least squares) regression line to the data. Here is the numerical output:

```
Results:

Simple linear regression results:

Independent variable: var3
Dependent variable: var2
Sample size: 9
Correlation coefficient: 0.9205
(See fitted line plot in Graphics Panel.)
Residuals stored in column var4

Estimate of sigma: 17.597992

Parameter   Estimate    Std. Err.   DF   Tstat       Pval
Intercept   200.49911   30.960272   7    6.4760127   2.0E-4
var3        36.357025   5.832346    7    6.233688    2.0E-4
```

1. Indicate the exact values for slope, y–intercept, the regression equation, and the correlation coefficient obtained from WebStat. Provide an interpretation of Pearson's correlation coefficient r between Gold and Silver for this given data set. (**15 Points**)

2. Based on your equation in (1.), what is the predicted price of an ounce of gold when the price of an ounce of silver is \$4.00?

   Do you think it is safe to predict that an ounce of gold costs about \$237 if the price of an ounce of silver drops to \$1.00? Explain. (**15 Points**)

3. Draw 2 different residual plots for this data set. One should contain the lurking variable. Is there any visible pattern in any of the your 2 residual plots? If there is a pattern, desribe the pattern. Finally, argue whether the least squares regression line describes the relationship between Silver and Gold reasonably well. (**20 Points**)

Question 2: Straight Lines & Correlation (**50 Points**)

1. Determine the slope and the y–intercept of the lines whose equations are given as: (**12 Points**)

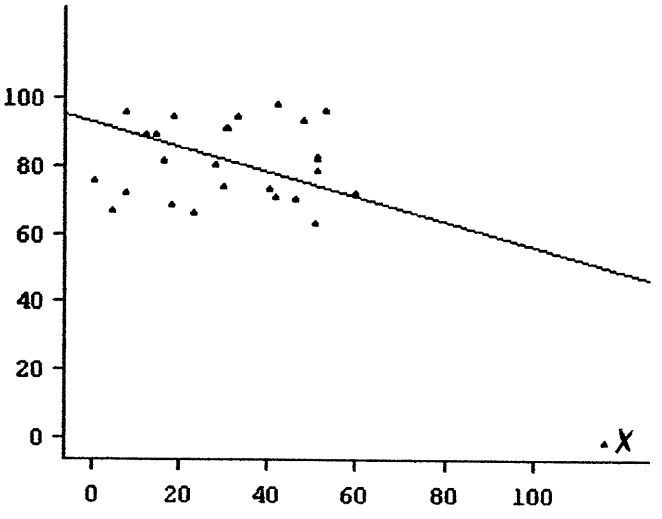
   (a) $10x - 5y = 25$

   (b) $x + 6y = 0$

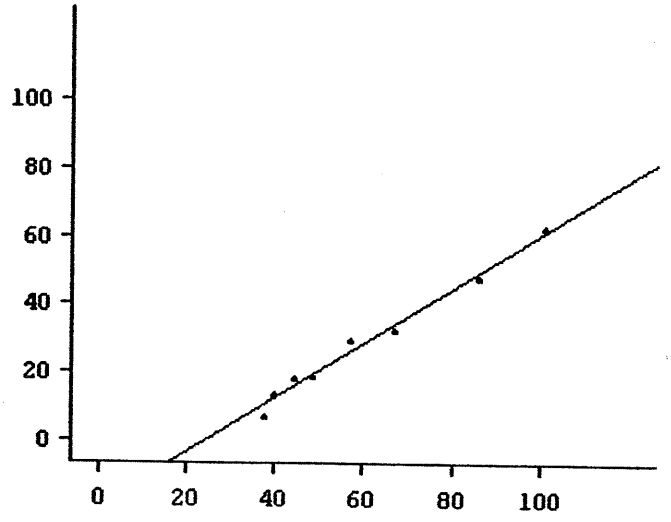
2. Indicate the slope and the y–intercept of the line that goes through the points $(x_1, y_1) = (1, 8)$ and $(x_2, y_2) = (3, 5)$. (**8 Points**)

3. The following 2 graphics show least squares regression lines that are fitted to 2 different data sets. **(18 Points)**

   (a) Describe the influence of the point that has been marked with an x in the scatterplot below using the appropriate statistical terminology. Sketch how the regression line will look when this point is being removed.



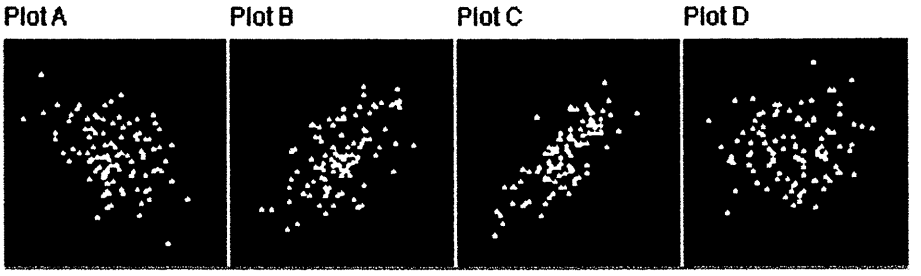
   (b) Describe what will happen to the regression line when we add an additional point at the approximate location $(x, y) = (70, 100)$ in the scatterplot below. Use the appropriate statistical terminology to describe this point. Sketch how the regression line will look when this point is being added.



4. The correlation coefficients for the data points displayed in these four scatterplots are 4 out of the following 9 values:

$$-0.97, \quad -0.78, \quad -0.37, \quad -0.12, \quad 0.12, \quad 0.37, \quad 0.58, \quad 0.78, \quad 0.99$$

For each plot below, indicate the corresponding correlation. **(12 Points)**



Correlation for Plot A:

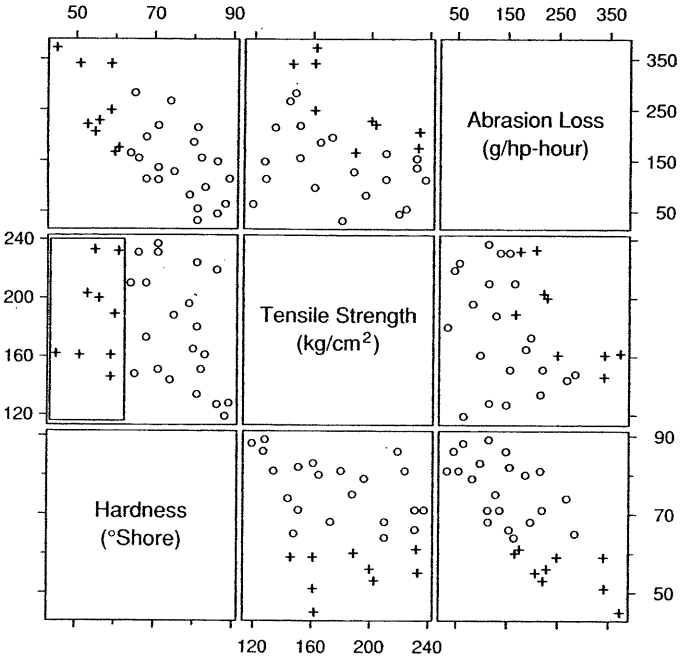
Correlation for Plot B:

Correlation for Plot C:

Correlation for Plot D:

Question 3: Scatterplot Matrix & Linked Brushing (**50 Points**)

The questions on the next page are based on the scatterplot matrix presented in William Cleveland's book "Visualizing Data". This scatterplot matrix has been reprinted below. It displays trivariate data that represents measurements of "Abrasion Loss", "Hardness", and "Tensile Strength" for 30 rubber specimens. "Abrasion Loss" is the dependent variable and "Hardness" and "Tensile Strength" are the independent variables. The goal of this study was to determine conditions that minimize the "Abrasion Loss".



1. Label the (individual) scatterplot that shows the "Tensile Strength" on the horizontal (x–)axis and the "Hardness" on the vertical (y–)axis with the letter "A". (**10 Points**)

2. What is the (approximate) range R of "Abrasion Loss"? (**10 Points**)

3. Describe the form, direction, and strength of the relationship in the scatterplots that show "Hardness" and "Abrasion Loss". (**10 Points**)

4. Which of these statements is correct/incorrect/undecidable?

   (a) The brush is located in the scatterplot that shows the "Tensile Strength" on the vertical (y–)axis and the "Hardness" on the horizontal (x–)axis. (**5 Points**)

   (b) The low values (i.e., values in the approximate range 40–60) of "Hardness", have been brushed. (**5 Points**)

5. Decide on **ONE** of the following options. The best way to minimize "Abrasion Loss" is a combination of (**10 Points**)

   (a) low "Hardness" and low "Tensile Strength"

   (b) low "Hardness" and high "Tensile Strength"

   (c) high "Hardness" and low "Tensile Strength"

   (d) high "Hardness" and high "Tensile Strength"

Question 4: Samples, Experiments and Studies (**50 Points**)

Answer the following questions:

1. A survey is done to estimate the average length of time that American college students sleep each night. Based on responses from 400 students, it is estimated that the average length of time that college students sleep is 7 hours per night. (**5 Points**)

   (a) What is the population of interest in this survey?

   (b) What is the sample in this survey?

   (c) What questions should be asked about this survey in order to know if the result is accurate? Indicate at least 3 different questions.

2. Suppose that your school's administration is doing a survey to see if students are in favor of building a new basketball arena. (**5 Points**)

   (a) Give an example of how they might choose a sample that does not represent all students.

   (b) Describe how they might choose a sample that is representative of all students.

3. Suppose a state has 10 universities, 25 four–year colleges, and 50 community colleges, each of which offer multiple sections of an introductory statistics class each year. Researchers want to conduct a survey of students taking introductory statistics in the state. Explain a method for collecting each of the following types of samples: (**5 Points**)

   (a) A stratified sample.

   (b) A cluster sample.

   (c) A simple random sample.

4. Suppose that 30 students will participate in an experiment in which the effectiveness using a web–based approach to teaching statistics is compared to the effectiveness of a textbook–based approach. (**5 Points**)

   (a) Describe how the researcher could assign participants to the two different approaches.

   (b) Why wouldn't the researcher want to let each student choose the method he or she wants to use?

5. The faculty senate at a large university wanted to know what proportion of the students thought a foreign language should be required for everyone. The statistics department offered to cooperate in conducting a survey, and a simple random sample of 500 students was selected from all students enrolled in statistics classes. A survey form was sent by email to these 500 students. (**10 Points**)

   (a) What is the population of interest to the faculty senate?

   (b) What is the sampling frame?

   (c) What is the sample?

   (d) Is the sample representative of the population of interest? Explain.

6. A psychiatrist compares two treatments for depression. She puts the names of 20 patients into a box and randomly draws the names of 10 people who will be assigned to treatment 1. The other 10 patients will be assigned to treatment 2. What type of data collection method is this? Mark the correct answer: (**5 Points**)

   - Census.
   - Comparative randomized experiment.
   - Observational study.
   - Sample survey.

7. The cholesterol levels of a sample of 200 heart attack patients in a hospital are compared to the cholesterol levels of 350 other patients. What type of study is this? Mark the correct answer: (**5 Points**)

   - Observational study.
   - Hawthone study.
   - Double–blind study
   - Randomized experiment.

8. A random sample of households in a city has been selected for a door–to–door survey asking questions about what recreational services the city should provide. You have been hired to conduct the survey. At the first house you contact, the person who answers the door tells you that her neighbor next door would be a better person to ask because she has more free time. What should you do and why? Mark the correct answer: (**5 Points**)

   - Go next door and ask the neighbor; the house is in the same neighborhood so it doesn't matter which household is included.
   - Try to get the person to respond anyway; if you substitute the neighbor the sample will be biased in favor of people with more time.
   - Ignore this household and reduce the sample size by one; people who don't have a lot for free time probably won't use the recreational facilities anyway.
   - Replace this household with another one chosen randomly from the city; one randomly chosen household is as good as another.

9. What is the relationship between a probability sampling plan and a simple random sampling plan? Mark the correct answer: (**5 Points**)

   - A simple random sampling plan is a special type of probability sampling plan.
   - A probability sampling plan is a special type of simple random sampling plan.
   - They are equivalent
   - They are completely different; neither is a special case of the other.