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**)

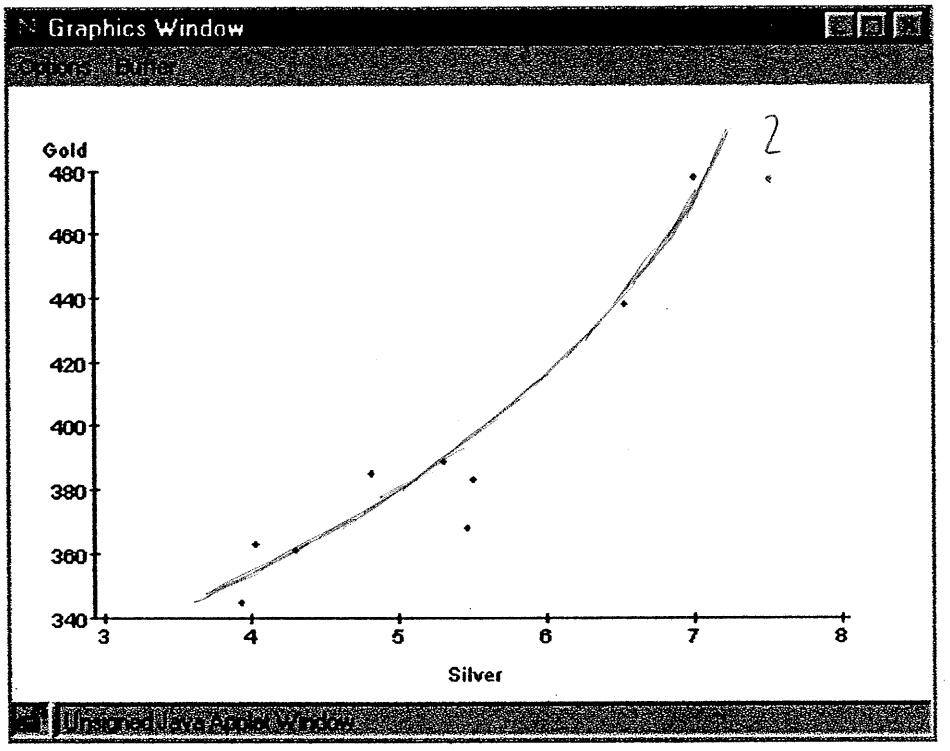
The average prices (in dollars) per ounce of gold and silver for the years 1986 through 1994 are given below as they appear in the WebStat main window. (Source: U.S. Bureau of Mines.) Note that var1 = Year, var2 = Price of Gold, var3 = Price of Silver, and var4 = Residual.

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	

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**)

slope: $b_1 = 36.357025$ (3)

y-intercept: $b_0 = 200.49911$ (3)

regression equation: $\hat{y} = b_0 + b_1 \cdot x = 200.49911 + 36.357025 \cdot x$ (2)

correlation: $r = 0.9205$ (3)

Interpretation: There is a strong positive linear relationship between gold price and silver price since the correlation is close to +1. In a scatterplot the data points fall close to a straight rising line. (4)

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**)

   • $x = 4.00 \Rightarrow \hat{y} = 200.49911 + 36.357025 \cdot 4.00 = 345.92721$ (6)
   predicted gold price if silver costs \$ 4.00

   • $x = 1.00 \Rightarrow \hat{y} = 200.49911 + 36.357025 \cdot 1.00 = 236.856135$ (9)
   Since the minimum price of silver in the years 1986 to 1994 was \$ 3.94, a price of \$1.00 is considerably less (i.e., only 25% of the minimum). Something has changed significantly – perhaps an overall economic crisis, new massive source of silver have been detected, or other unpredictable events may have occurred that also may or may not affect the gold price. A gold price of \$237 in the case that silver drops to \$1.00 is very uncertain, i.e., this value is not very reliable due to extreme extrapolation of the x-value.

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**)

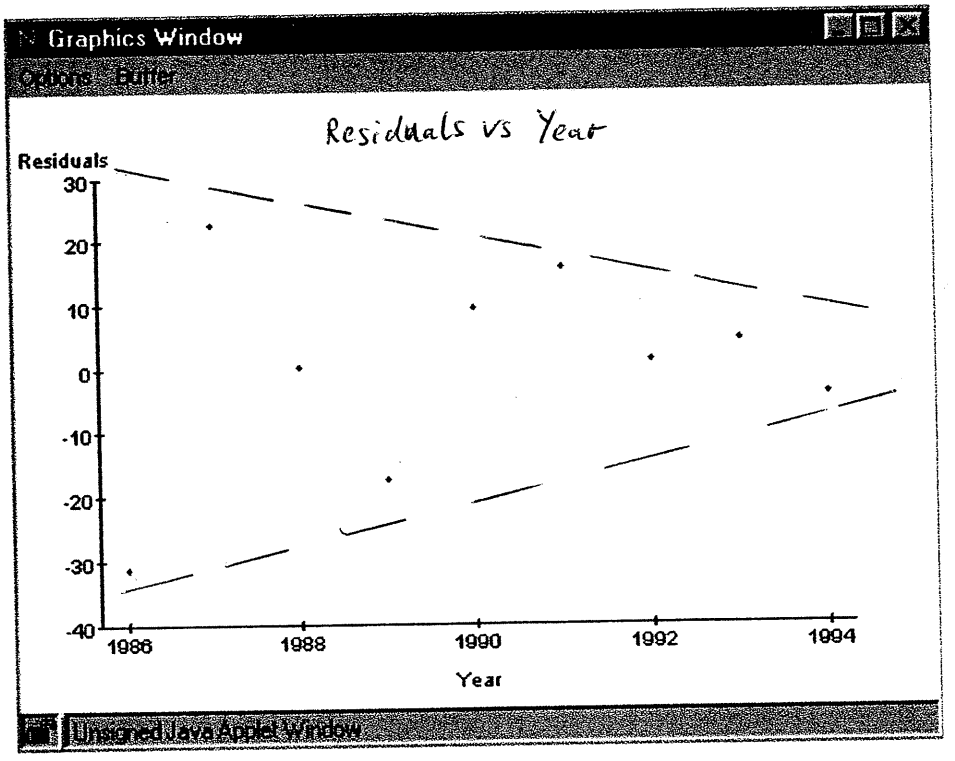
see next page

3. Since we know one additional (lurking) variable (Year), we should definitely plot the residuals e_i vs the Year.

The second residual plot should be a plot of e_i vs x_i, $\hat{y}_i$, or y_i.

Here are the 2 residual plots: (−3) if Residuals vs Year not plotted

(6)



(3)

The plot Residuals vs. Year shows a slight $\cup$-shaped pattern, indicating that responses for the first 2 years (1986 & 1987) are fitted less consistently than for future years.

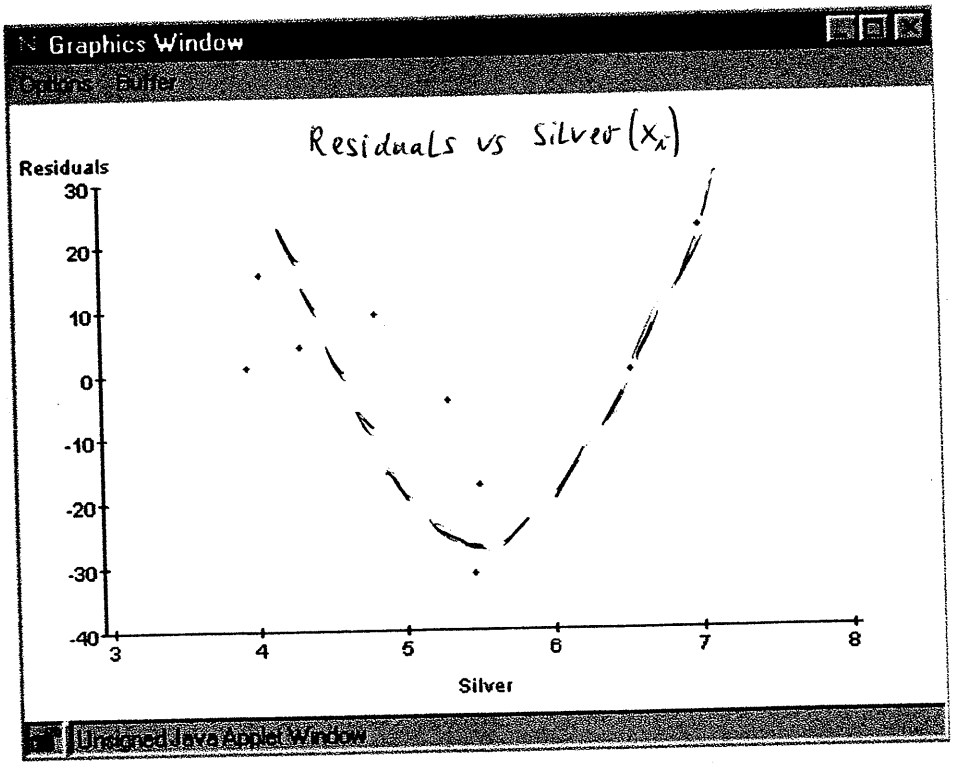
(3)

The plot Residuals vs. Silver shows smallest (negative) residuals for Silver values in the middle and highest (positive) residuals for low and for high Silver values.

(2)

Based on this 2nd residual plot, one might ask whether a straight line describes the relationship between Silver and Gold correctly. Perhaps a curved line (quadratic or exponential) might be better for this data (see sketch on page 2).

(6)



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**)

(-2) if not labeled

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

$\Leftrightarrow$ $y = -5 + 2x$

$\Rightarrow$ y-intercept: -5 (3)

slope: 2 (3)

(b) $x + 6y = 0$ $\Leftrightarrow$ $6y = 0 - x$

$\Leftrightarrow$ $y = 0 - \frac{1}{6}x$

$\Rightarrow$ y-intercept: 0 (3)

slope: $-\frac{1}{6}$ (3)

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**)

slope $b_1 = \frac{\Delta y}{\Delta x} = \frac{(5-8)}{(3-1)} = \frac{-3}{2} = -1.5$ // (4)

use one point, e.g. $(x_1, y_1) = (1, 8)$ to solve for b_0:

$8 = b_0 - \frac{3}{2} \cdot 1$

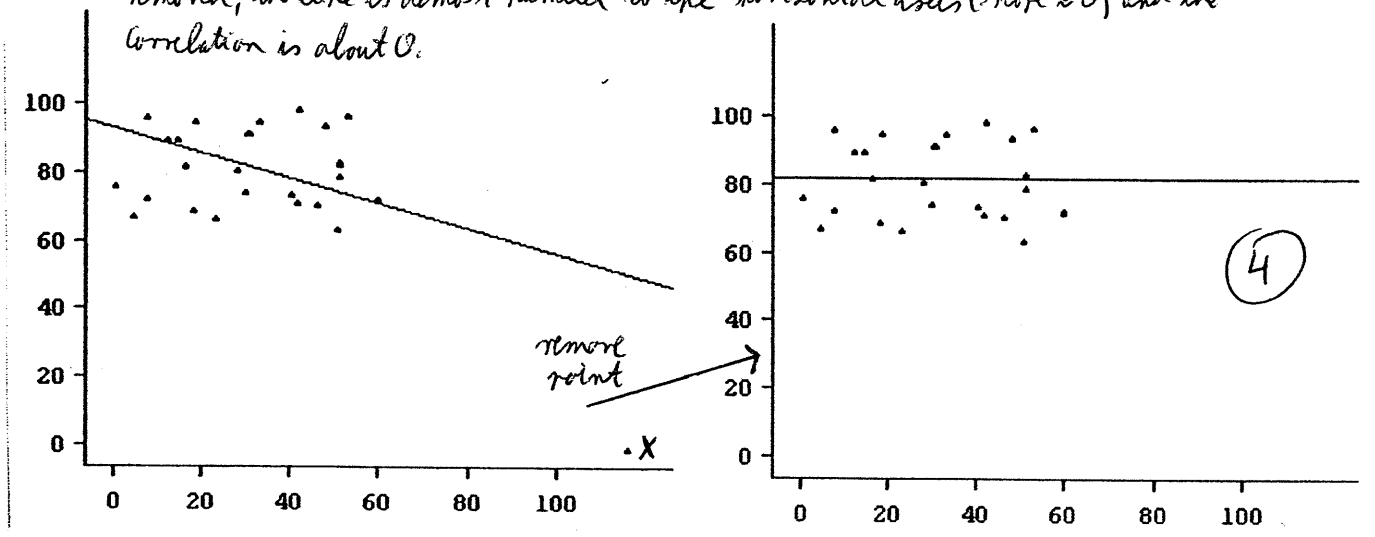
$\Rightarrow$ y-intercept $b_0 = 8 + \frac{3}{2} = 9.5$ // (4)

(-1) if not labeled

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.

The point marked "x" is an influential point, here an outlier in x direction & y direction. When removed, the regression line changes considerably. With this point "x" removed, the line is almost parallel to the horizontal axis (slope ≈ 0) and the correlation is about 0. (5)

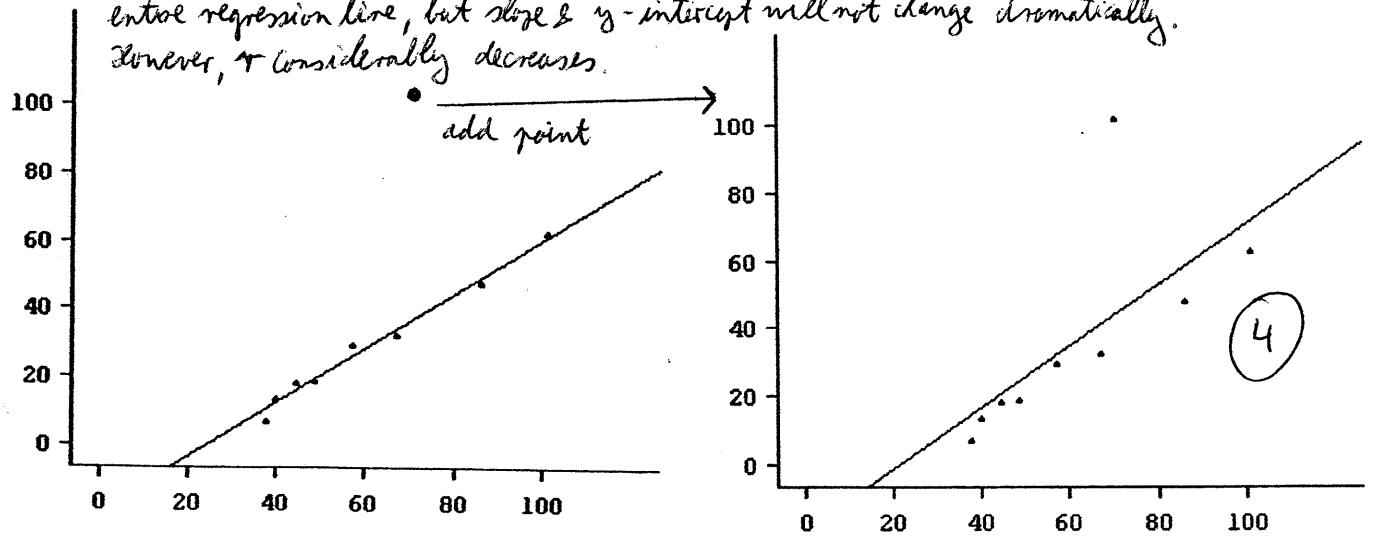


Y = 92.77 - 0.3672 (X), SD of resids = 14.69, r = -0.4724

Y = 81.65 - 8.205E-4 (X), SD of resids = 10.57, r = -0.001343

(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.

When a point "•" is added, it will be an outlier in y direction. It will raise the entire regression line, but slope & y-intercept will not change dramatically. However, r considerably decreases. (5)



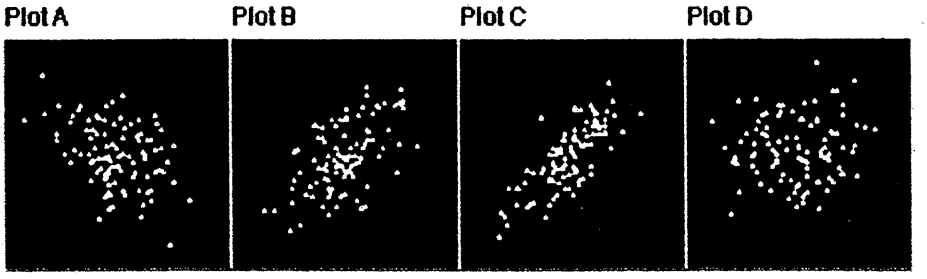
Y = -18.63 + 0.7951 (X), SD of resids = 2.091, r = 0.9928

Y = -18.62 + 0.8955 (X), SD of resids = 19.15, r = 0.6967

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:

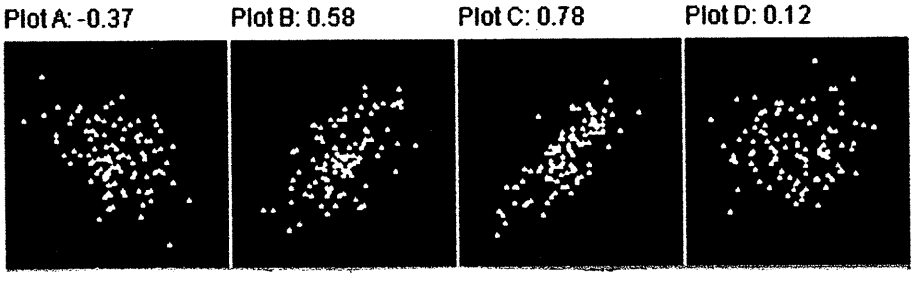
Correlation for Plot B:

Correlation for Plot C:

Correlation for Plot D:

(-3) if far off
(-2) if slightly off
(3) each

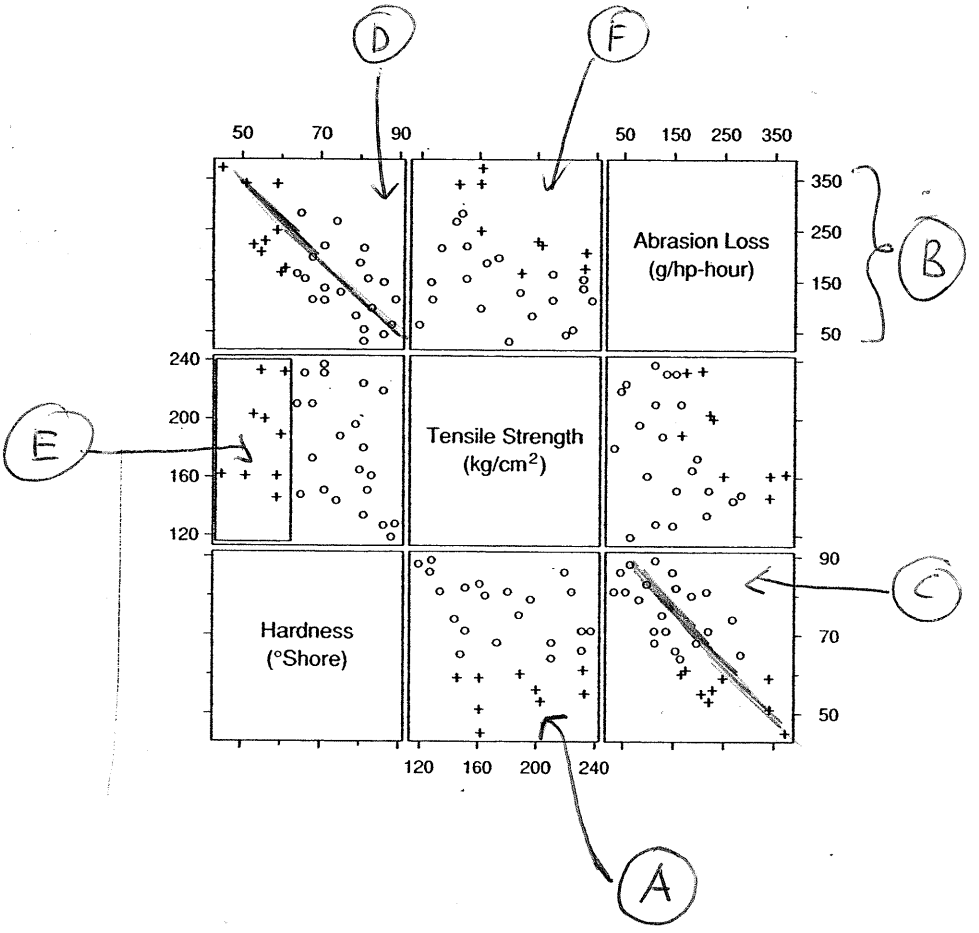
Here are the answers:



In none of the plots are the points close to a straight line – this eliminates −0.97 and 0.99.
In plot C, points are overall closest to an (increasing) straight line – this must be 0.78
There is no plot with a similar decreasing straight line – this eliminates −0.78
In plot B, points also are related to an increasing straight line but are farther apart from the line than in plot C and not as far apart as in plot A – this must be 0.58
In plot A, points are related to a decreasing straight line – the only reasonable option left is −0.37
In plot D, we see a very weak linear relationship that is certainly increasing – the only reasonable option is 0.12.

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)**

(-5) if axes flipped

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

look at (B): $R \approx 350 - 50 = 300$ (in g/hp-hour)

(-5) if maxes min listed

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

look at (C) or (D).
The relationship between Hardness and Abrasion Loss can be described by a line that is linear (4) and has negative (3) slope. Overall, the relationship is medium to strong (3)

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)**

   look at (E): this is correct ⇒ correct

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

   look at (E): this is correct ⇒ correct

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"

look at (F): low values of Hardness have been brushed with a "+" symbol; in average, for identical Tensile Strength, low values of Hardness relate to a higher Abrasion Loss than high values of Hardness; overall, Abrasion Loss decreases with an increase of Tensile Strength; so (d)

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?

   The population is all American college students.

   (b) What is the sample in this survey?

   The sample is the 400 students.

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

   Possible questions include:
   - How was the sample selected?
   - Is the sample representative of all American college students?
   - How was the question asked? Is it an unbiased question?

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.

   Ask students as they leave a basketball game.

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

   Randomly select and survey a sample of students from the list of all students who attend the school.

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.

   a. Stratified sample: use the 3 types of schools as strata. Create a list of all students for each of the 3 strata. Draw a simple random sample from each of the 3 lists.
   b. Cluster sample: Use individual schools or individual classes as clusters. Take a random sample of clusters, measure all students in those clusters.
   c. Simple random sample: Obtain a list of all students in the classes at all schools; take a simple random sample from that combined list.

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.

   Assign the numbers 1 to 30 to the thirty students. In some way, randomly select 15 numbers between 1 and 30. The students with those numbers are assigned to the web-based approach. The other students are assigned to the textbook approach.

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

   If each student chooses the method he or she will use, there is a risk of an unfair comparison. Possibly, a particular type of learner may be more likely to choose the web-based approach. For instance, students with generally better quantitative skills might be inclined choose the web-based approach. These students would be likely to fare better than the other students regardless of the approach used.

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?

   The population of interest is the population of all students at the university.

   (b) What is the sampling frame?

   The sampling frame is the collection of all students enrolled in statistics classes at the time of the survey.

   (c) What is the sample?

   The sample is the 500 students to whom the survey was mailed.

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

   The extent to which the sample represents the population of interest depends on what types of students are enrolled in statistics classes that term. Depending on the university requirements and the term in question, there may be more freshman or seniors than in the general student body, and certain majors are likely to be over- or under-represented.

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.