30. Which of the following statements about the correlation r are true?

I. When $r = 0$, there is no relationship between the variables.
II. When $r = .5$, 50% of the variables are closely related.
III. When $r = 1$, there is a perfect cause-and-effect relationship between the variables.

(A) I only
(B) II only
(C) III only
(D) I, II, and III
(E) All the statements are false.

31. Consider n pairs of numbers. Suppose $\bar{x} = 2$, $s_x = 3$, $\bar{y} = 4$, and $s_y = 5$. Of the following, which could be the least squares line?

(A) $\hat{y} = -2 + x$
(B) $\hat{y} = 2x$
(C) $\hat{y} = -2 + 3x$
(D) $\hat{y} = \frac{5}{3} - x$
(E) $\hat{y} = 6 - x$

Answer Key

1. **E**	8. **C**	14. **D**	20. **A**	26. **E**
2. **A**	9. **E**	15. **C**	21. **A**	27. **E**
3. **B**	10. **B**	16. **A**	22. **D**	28. **A**
4. **D**	11. **A**	17. **E**	23. **B**	29. **B**
5. **E**	12. **A**	18. **B**	24. **B**	30. **E**
6. **A**	13. **C**	19. **E**	25. **E**	31. **E**
7. **E**				

Answers Explained

1. **(E)** The slope is 20/10 = 2; that is, a woman's risk of developing ovarian cancer rises 2% for every gram of fat consumed per day. The other statements may be true, but they do not answer the question.

2. **(A)** Slope $= .15(\frac{42{,}000}{1.3}) \approx 4850$ and intercept $= 208{,}000 - 4850(6.2) \approx 178{,}000$.

3. **(B)** log(revenue in \$1000), not revenue, goes up 0.82 per year.

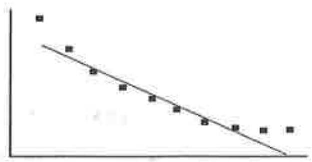
log(revenue in \$1000) = 0.82(5) + 0.67 = 4.77 gives revenue = $10^{4.77}$ thousand dollars ≈ 59 million dollars.

$r^2 = (.73)^2 = 53\%$ of the variation in log(revenue in \$1000), not revenue, can be explained by variation in time.

4. **(D)** The correlation coefficient is not changed by adding the same number to each value of one of the variables or by multiplying each value of one of the variables by the same positive number.

5. **(E)** A negative correlation shows a tendency for higher values of one variable to be associated with lower values of the other; however, given any two points, anything is possible.

6. **(A)** This is the only scatterplot in which the residuals go from positive to negative and back to positive.



7. **(E)** Since (2, 5) is on the line $y = 3x + b$, we have $5 = 6 + b$ and $b = -1$. Thus the regression line is $y = 3x - 1$. The point $(\bar{x}, \bar{y})$ is always on the regression line, and so we have $\bar{y} = 3\bar{x} - 1$.

8. **(C)** The correlation r measures association, not causation.

9. **(E)** The correlation r cannot take a value greater than 1.

10. **(B)** It can be shown that r^2, called the coefficient of determination, is the ratio of the variance of the predicted values $\hat{y}$ to the variance of the observed values y. Alternatively, we can say that there is a partition of the y-variance and that r^2 is the proportion of this variance that is predictable from a knowledge of x. In the case of perfect correlation, $r = \pm 1$.

11. **(A)** Correlation has the formula

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

We see that x and y are interchangeable, and so the correlation does not distinguish between which variable is called x and which is called y. The formula is also based on standardized scores (z-scores), and so changing units does not change the correlation. Finally, since means and standard deviations can be strongly influenced by outliers, the correlation is also strongly affected by extreme values.

12. **(A)** Removal of scores with large residuals but average x-values may not have a great effect on the regression line.

13. **(C)** An influential score may have a small residual but still have a greater effect on the regression line than scores with possibly larger residuals but average x-values.

14. **(D)** The sum and thus the mean of the residuals are always zero. In a good straight-line fit, the residuals show a random pattern.

15. **(C)** The coefficient of determination r^2 gives the proportion of the y-variance that is predictable from a knowledge of x. In this case $r^2 = (.632)^2 = .399$ or 39.9%.

16. **(A)** The regression equation is $\hat{y} = 7386.87 - 3.627x$, and so its slope is -3.627 and the y-value drops 3.627 for every unit increase in the x-value. Answer (C) is a true statement but doesn't pertain to the question.

17. **(E)** $-3.627(1983) + 7386.87 = 194.5$

18. **(B)** On each exam, two students had scores of 100. There is a general negative slope to the data showing a moderate negative correlation.

19. **(E)** On the scatterplot all the points lie perfectly on a line sloping up to the right, and so $r = 1$.

20. **(A)** The slope and the correlation are related by the formula

$$b_1 = r\frac{s_y}{s_x}$$

The standard deviations are always positive, and so b_1 and r have the same sign. Positive and negative correlations with the same absolute value indicate data having the same degree of clustering around their respective regression lines, one of which slopes up to the right and the other of which slopes down to the right. While $r = .75$ indicates a better fit with a linear model than $r = .25$ does, we cannot say that the linearity is threefold.

21. **(A)** The correlation is not changed by adding the same number to every value of one of the variables, by multiplying every value of one of the variables by the same positive number, or by interchanging the x- and y-variables.

22. **(D)** The slope is negative (–2.4348); that is, the regression line slopes down to the right, indicating that nonagenarians with greater strength have lower walk times and thus greater functional mobility.

23. **(B)** The slope and the correlation coefficient have the same sign. Multiplying every y-value by -1 changes this sign.

24. **(B)** Correlation shows association, not causation. In this example, older school children both weigh more and have faster reading speeds.

25. **(E)** A scatterplot readily shows that while the first three points lie on a straight line, the fourth point does not lie on this line. Thus no matter what the fifth point is, all the points cannot lie on a straight line, and so r cannot be 1.

26. **(E)** Only III is true. A positive correlation indicates that higher values of x tend to be associated with higher values of y.

27. **(E)** All three scatterplots show very strong nonlinear patterns; however, the correlation r measures the strength of only a linear association. Thus $r = 0$ in the first two scatterplots and is close to 1 in the third.

28. **(A)** Using your calculator, find the regression line to be $\hat{y} = 9x - 8$. The regression line, also called the least squares regression line, minimizes the sum of the squares of the vertical distances between the points and the line. In this case (2, 10), (3, 19), and (4, 28) are on the line, and so the minimum sum is $(10 - 11)^2 + (19 - 17)^2 + (28 - 29)^2 = 6$.

29. **(B)** When transforming the variables leads to a linear relationship, the original variables have a nonlinear relationship, their correlation (which measures linearity) is not close to 1, and the residuals do not show a random pattern.

30. **(E)** These are all misconceptions about correlation. Correlation measures only linearity, and so when $r = 0$, there still may be a nonlinear relationship. Correlation shows association, not causation.

31. **(E)** The least squares line passes through $(\bar{x}, \bar{y}) = (2, 4)$, and the slope b satisfies $b = r\frac{s_y}{s_x} = \frac{5r}{3}$. Since $-1 \le r \le 1$, we have $-\frac{5}{3} \le b \le \frac{5}{3}$.