1. Temperature and air pollution are known to be correlated. We collect data from two laboratories, in Boston and Montreal. Boston makes their measurements of temperature in Fahrenheit, and Montreal in degrees centigrade. Boston measures pollution in particles per cubic yard of air; Montreal uses cubic meters. Both report a correlation of exactly 0.58 between temperature and pollution. Which of the following is true:
   A. Boston really has the higher correlation, because Fahrenheit temperatures are higher than Centigrade.
   B. Montreal really has the higher correlation, because cubic meters are bigger than cubic yards.
   C. Both cities have the same correlation, because correlation is independent of the units of measurement.
   D. We do not know which city has the really higher correlation.

Answer: C. The strength of a correlation does not change if units change by a linear transformation such as:
   Fahrenheit = 32 + (5/9) * Centigrade

2. Which of the following is NOT a possible value of the correlation coefficient?
   A. negative 0.9
   B. zero
   C. positive 0.15
   D. positive 1.5
   E. negative .05

Answer: D. Correlations are always between -1 (perfect negative) and +1 (perfect positive)

3. We measure heights and weights of 100 twenty-year old male college students. Which will have the higher correlation:
   A. corr(height, weight) will be much greater than corr(weight, height)
   B. corr(weight, height) will be much greater than corr(height, weight)
   C. Both will have the same correlation.
   D. Both will be about the same, but corr(weight, height) will be a little higher.
   E. Both will be about the same, but corr(height, weight) will be a little higher.

Answer: C. Correlation is independent of the order in which the variables enter.

4. Two lists of numbers, X and Y, have a correlation of 0.3; X and Z have a correlation of -0.7
   We know that:
   A. the stronger correlation is the correlation of X and Y, since it is positive.
   B. the stronger correlation is the correlation of X and Z.
   C. the two correlations are equally strong, since 1.0 - 0.7 = 0.3
   D. We cannot tell which is stronger without more information.

Answer: B. The stronger correlation is determined by the absolute value, since it measures the scatter of points about a line. Whether the line has a positive or negative slope, the less the scatter, the greater the absolute value of the correlation.

5. Suppose men always married women who were 10 percent shorter than they were. The correlation coefficient of the heights of married couples would be:
   A. 0.10 if the correlation were computed with corr(male.height, female.height)
   B. -0.10 if the correlation were computed with corr(female.height, male.height)
   C. 0.10 no matter which way the correlation were computed.
   D. 1.0 since the height of the man is always predictable from the height of the woman.

Answer: D. All points in a graph of the correlation would line on the straight line \( H_f = 0.9 \times H_m \)
where \( H_f \) = female height and \( H_m \) = male height.
6. In one class, the correlation between the final and the midterm was 0.5, whereas the correlation between the final and the homework grades was 0.25. This means that:
   A. the relation of the final and the midterm was twice as linear as the relation between the final and the homework.
   B. the relation of the final and the homework was twice as linear as the relation between the final and the midterm.
   C. More of the variation of the final was explained by the homework grades than by the midterm grades.
   D. More of the variation of the final was explained by the midterm grades than by the homework grades.

   **Answer:** D. The phrase "more linear" is meaningless; a line is either linear or curved; and correlation is always correlation around a straight line. It does indicate (when squared) the percentage of the variance explained, and the midterm grade explained \( \text{square}(.5) = .25 \) or 25 percent of the variance, where the homework explained only \( \text{square}(.25) = .0625 \) or 6.25 percent of the variance.

7. The unemployment rate is related to inflation by the Phillips curve, which is typically a negative sloped curve looking like a hyperbola -- inflation is very high at very low rates of unemployment, and it takes very high rates of unemployment to bring inflation down to zero. We compute a correlation coefficient between unemployment rates and inflation, and find it is negative 0.5. The true relation between the two is most probably:
   A. exactly as reported by the correlation coefficient.
   B. stronger than reported by the correlation coefficient, due to the non-linearity.
   C. weaker than reported by the correlation coefficient, due to the great scatter of points around the line.

   **Answer:** B. The correlation coefficient reports the cluster of points around a straight line; if the true relation is curvilinear, the correlation will understate the strength of the relation.

8. An investigator is studying the relation between the physical and intellectual growth of primary schoolchildren (grades 1-6). At each grade level, she notes that the correlation between the height of children and the size of their vocabulary is zero. For all students in the school, the correlation is likely to be:
   A. Positive
   B. Negative
   C. About zero
   D. Cannot tell.

   **Answer:** A. While height and vocabulary size have nothing to do with each other, both have a lot to do with age, especially for ages 6 to 12. The common cause will lead to a strong correlation of the two, which of course could not be said of (say) ages 60 to 72.

9. A study is done of students commuting to a large university by bicycle. The correlation between the time spent waiting at traffic lights and total cycling time was 0.50. This means:
   A. The average rider spent half his cycling time waiting at traffic lights.
   B. The more time a rider spends waiting at traffic lights, the higher is total time is likely to be.
   C. If the rider's time at traffic lights increases by 5 minutes, he will spend an additional 10 minutes commuting, on the average.
   D. If the rider's time at traffic lights increases by 10 minutes, he will spend an additional 5 minutes commuting, on the average.

   **Answer:** B. You would need a regression equation to see whether or not time at lights predicted an exact time of commute. If the distances of the commute varied, using time at lights alone would lead to an omitted variable problem for a regression of total time on time at lights.
10. The correlation between the ages of the husbands and wives in the United States was which of the following?

A. +1.0  B. +0.85  C. zero  D. -0.85  E. -1.0

**Answer:** B. Men usually marry women of about their own age, but of course there are enough exceptions that the correlation is not perfect.

11. A study is done of the impact of a drug on body temperature and blood pressure. We have three observations:

<table>
<thead>
<tr>
<th>Temperature (F)</th>
<th>Pressure (mg)</th>
<th>Temp - 97</th>
<th>Pressure - 80</th>
</tr>
</thead>
<tbody>
<tr>
<td>99</td>
<td>100</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>97</td>
<td>80</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>104</td>
<td>150</td>
<td>7</td>
<td>70</td>
</tr>
</tbody>
</table>

The correlation coefficient is closest to

A. +1.0  B. +0.8  C. +0.5  D. zero  E. -0.5

**Hint:** A good guess is possible by plotting the data carefully. A better one is possible by subtracting the lowest number in each list from all the other numbers in that list.

**Answer:** A. Transform the first column of data by subtracting 97 from each number; transform the other column by subtracting 80 from each. The result is given above; the line Pressure = 10*Temp holds perfectly for all points.

12. The correlation between the average midterm score of each of 10 classes of statistics and the average final exam score was found to be 0.85. A statistics instructor concludes that if a student has a B on the midterm, he has an 85 percent chance of a B on the final. This conclusion is:

A. correct  
B. incorrect, because it is an ecological correlation, which means there is usually much more variance of individuals than of averages  
C. incorrect, because it is an average correlation, and there is usually much more variance of averages than of individuals.

**Answer:** B.
13. A regression tries to predict student GPA percentile (on a zero-100 scale) from their SAT-Math score. You are to make the relevant prediction for a student who has a SAT math score of 600 (note that SAT scores go from 200 to 800). He would be expected to be closest to the -- percentile of GPA on the basis of the regression output below:

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>20</td>
</tr>
<tr>
<td>SAT-Math</td>
<td>0.1</td>
</tr>
</tbody>
</table>

R-squared: 0.36

A. 10th percentile  B. 20th percentile  C. 60th percentile  D. 80th percentile  E. 100th percentile

Answer: D. Write the regression equation as GPA percentile = 20 + 0.1 * SAT math, so GPA percentile = 20 + 0.1 * 600 = 80

14. We know, on the basis of the above regression, that the strongest part of the explanation of the score is due to the coefficient of the:

A. intercept term, because the coefficient is larger.
B. intercept term, because the t-stat is smaller
C. SAT-math term, because the coefficient is smaller.
D. SAT-math term, because the t-stat is larger.

Answer: D. The size of the t-statistic, not the size of the coefficient, indicates the explanatory strength of a variable.

15. We know that the correlation coefficient between SAT-math and the GPA percentile is closest to:

A. zero  B. 0.30  C. 0.60  D. 0.80  E. 1.00

Answer: C. The R-squared is the square of the correlation coefficient r, so r = sqrt(.36) = 0.6

16. If we knew that the SD of the SAT math score was 100 and the SD of the GPA percentile was 25, we could calculate the standard error of the regression (the RMSE) to be closest to (look at the next question for a hint at the formula):

A. 20  B. 40  C. 60  D. 80  E. 100

Answer: C. The formula for the standard error of the regression (RMSE) is

A. (1 - R2) * SD(x)
B. Covariance (x,y) / SD(x) * SD(y)
C. sqrt(1- R2) * SD(y)
D. sqrt(1-R2) * SD(x)
E. r * SD(y) / SD(x)

Answer: C. The standard error of the regression gets smaller as the R-squared gets larger; if R-sq = 1, error = 0

18. The formula for the slope of the regression line is (same options as the previous question)

Answer: E

19. The formula for the correlation coefficient is (same options as the previous question)

Answer: B
20. We plot the data lying behind the question 13 regression with the R command

```r
>>> Plot(SAT-Math, GPA)
```

To draw a regression line on the plot, use the R command:

A. >>> abline(0.36)
B. >>> abline(20, 0.1)
C. >>> abline (0.3, 5.0)
D. >>> abline (0.1, 5.0)
E. We cannot draw a regression line because we do not know the correlation coefficient.

**Answer: B**

21. The regression line is drawn so that:

A. The line goes through more points than any other possible line, straight or curved
B. The line goes through more points than any other possible straight line.
C. The same number of points are below and above the regression line.
D. The sum of the absolute errors is as small as possible.
E. The sum of the squared errors is as small as possible.

**Answer: E**

22. In a regression, the --- that the standard error of the regression is, the greater the accuracy of the prediction will be.

A. smaller.
B. larger
C. we do not know unless we know whether the slope of the regression is positive or negative.

**Answer: A.**

23. In order for the regression technique to give the best and minimum variance prediction, all the following conditions must be met, EXCEPT for:

A. The relation is linear.
B. We have not omitted any significant variable.
C. Both the X and Y variables (the predictors and the response) are normally distributed.
D. The residuals (errors) are normally distributed.
E. The variance around the regression line is about the same for all values of the predictor.

**Answer: C**

24. Note that the last question said that all the conditions are needed for a “best and minimum variance prediction”. Very few real regressions can meet all the criteria, but their predictions may still be quite good -- at least unbiased, though not perhaps minimum variance. The cases in which the predictions will almost certainly be biased are:

A. conditions A and B are not met.
B. In case C, there is no problem, so regression will give the best and minimum variance prediction.
C. In cases D and E, regression will not give the minimum variance prediction.

25. Ecological correlations are weaker than other types of correlation because:

A. They are based on averages rather than individuals

26. If a regression has the problem of heteroscedasticity,

A. The predictions it makes will be wrong on average.
B. The predictions it makes will be correct on average, but we will not be certain of the RMSE
C. It will also have the problem of an omitted variable or variables.
D. It will also be based on a non-linear equation.

**Answer: B.** Heteroscedasticity implies that the variance will differ for different values of the regressor.
The following regression is based on a randomly chosen subset of the ecgrow data set, with 30 countries.

Call: ols(grate ~ invest + edu + gdp60 + openness + pop)

Coefficients:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>t-stat</th>
<th>Conf.interval (0.95)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.16</td>
<td>0.3</td>
</tr>
<tr>
<td>invest</td>
<td>0.09</td>
<td>4.6</td>
</tr>
<tr>
<td>edu</td>
<td>0.05</td>
<td>0.6</td>
</tr>
<tr>
<td>gdp60</td>
<td>-0.04</td>
<td>-5.0</td>
</tr>
<tr>
<td>openness</td>
<td>1.91</td>
<td>5.3</td>
</tr>
<tr>
<td>pop</td>
<td>0.06</td>
<td>0.4</td>
</tr>
</tbody>
</table>

-----------------------------------------------------

SE of regression(or RMSE) = 0.9155
R-squared = 0.6493

27. If we had to drop two variables from the regression, we would pick:
   A. Edu and pop, because they have the lowest t-stats.
   B. GDP60 and edu, because they have the lowest coefficients.
   C. GDP60 and pop, because they have the lowest t-stats.
   D. GDP60 and openness, because they have the highest t-stats in absolute value.

Answer: A.

28. If we wanted to make a 95 percent confidence interval prediction, we would make a point prediction with the equation, but place around that a margin of error of about:
   A. +/- 0.6493
   B. +/- 0.9155
   C. +/- 2 * .6493
   D. +/- 2 * .9155
   E. +/- 0.16

Answer: D

Suppose we had a much simpler regression, obtained with the command >>> regress(gdp85 ~ gdp60)

Coefficients:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>t-stat</th>
<th>Conf.interval (0.95)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.83</td>
<td>2.4</td>
</tr>
<tr>
<td>gdp60</td>
<td>1.04</td>
<td>21.9</td>
</tr>
</tbody>
</table>

SE of regression(or RMSE) = 10.7684
R-squared = 0.8267

29. If GDP in 1960 were 20 percent of US GDP (so gdp60 were 20), our prediction for gdp85 would be closest to:
   A. 20   B. 25   C. 30   D. 40   E. 50

Answer: B, since gdp85 = 3.83 + 1.04 * (20) = 3.83 + 20.8 = 24.63