GAMES Practice Problem Solutions – Correlation and Simple Linear Regression

1. (a) Perfectly correlated. Either variable would be equally effective as the independent or dependent variable

   (b) The correlation between shoe size and grades is likely weak or nonexistent. It would be interesting to predict grades based on shoe size. It takes many months and sometimes a full year to earn a course grade. While measuring shoe size is quick and easy.

   (c) Knowing how much fuel is in the fuel tank would be a useful piece of information, so it would make a good dependent variable. For example, a driver would want to know how much fuel is left, so they can plan their journey. The correlation between kilometers driven and fuel remaining would be strong but not perfect as it omits other useful pieces of information like speed and car characteristics.

2. (a) See "GAMES-DescriptiveStats-Excel-PracticeProblemSolutions.xlsx"

   (b) Upward sloping, positive relationship

   (c) The underlying relationship appears linear

   (d) The relationship is quite strong. Adding a regression line reveals the observations remain close to the line.

   (e) There are no outliers.

3. (a) Altitude is fairly easy to measure with little variance, So it would make a good independent variable. Altitude is useful in its ability to predict temperature and temperature is going to have a large effect on human beings during any climb. The scatterplot which show a negative relationship – as the altitude increases the temperature falls. The underlying relationship could be linear and the strength would be weak or moderate (0.2-0.6 correlation). This is because the temperature has a lot of variability and factors other than altitude affect temperature.

   (b) Both would make equally good independent or dependent variables. Ice cream sales and air conditioner sales are both affected by daily temperatures - when it's hot sales of both increase. The scatterplot would be positive, straight and the correlation would be moderate – Ice cream is relatively inexpensive while air-conditioners are fairly expensive, so some divergence would be expected.

   (c) Predicting strength based on age would be of interest. The underlying relationship would be quadratic (curved) with children and younger people not being as strong than a rapid drop in strength in older adult. The scatterplot would be curved in the relationship between strength and age would be moderate since other variables like health and fitness are important, too.

   (d) It would be more interesting to predict a driver's ability to react based on their blood alcohol level although the other way works, too. The underlying relationship would be nonlinear, perhaps even exponential, negative and strong.

4. A) 0.70, B) -0.09, C) 0.06, D) -0.94. Note that C and B have a similarly weak correlation coefficient but for different reasons.

5. A) -0.71, B) 0.95, C) -0.94, D) -0.01.

6. Correlation is not causation. Even if these two phenomenon are correlated (to the best of the author's knowledge, these two phenomenon are not correlated), it could be that committing violent crimes causes people to play violent video games. Or it could be that a hidden third factor explains both behaviours.

7. Correlation is not causation. Children who are taller might also be more mature with more developed brains.

8. (a) There is a negative relationship that is moderately strong between GDP and clean drinking water.

   (b) The correlation coefficient only works between two quantitative variables. The hemisphere variable is categorical.

   (c) Correlation cannot be greater than 1.0 which is a perfect correlation. The student made a calculation error.

   (d) Correlation is not causation. An increase in GDP could cause higher literacy rates. Or both GDP and literacy depend on a hidden third factor.

9. (a) i. 0.5
   ii. 0.25
   iii. 7.5

   (b) the mean of v and w.

   (c) 0.25

   (d) $+15$

10. (a) $\sigma_{s,m} = r_{m,s}\sigma_m\sigma_s = \sqrt{625}\sqrt{3025}\,(0.7) = 962.5$

    (b) $b_1 = 0.318, b_0 = 24.1$

    (c) \$62.27

    (d) Observations would be moderately close to the regression line which would be upward sloping.

    (e) 49 percent

11. (a) See "GAMES_DescriptiveStats_Excel_PracticeProblemSolutions.xlsx"

    (b) See "GAMES_DescriptiveStats_Excel_PracticeProblemSolutions.xlsx" – the relationship looks a little curved which would violate the assumption that the underlying reltationship is linear.

    (c) i. See "GAMES_DescriptiveStats_Excel_PracticeProblemSolutions.xlsx" – The beta coefficients changed. In part b, b_1 is calculated using the ratio of $\frac{\sigma_e}{\sigma_{na}}$ while in part c, the ratio is $\frac{\sigma_{na}}{\sigma_e}$.

    ii. See "GAMES_DescriptiveStats_Excel_PracticeProblemSolutions.xlsx" – No, the correlation of determination r^2 is the same whichever variables are the dependant variable and independent variable. A simple linear regression cannot differentiate correlation from causation.

12. See "GAMES_DescriptiveStats_Excel_PracticeProblemSolutions.xlsx"

    When the price is lower, the observations are clustered closely together in a line, but when the price is higher, the observations form more of a cloud.

(b) As the price of bitcoin increases, the observations become more cloudy and appear further away from the line.

(c) See "GAMES_DescriptiveStats_Excel_PracticeProblemSolutions.xlsx" – The conditions for an accurate simple linear regression do not seem to be satisfied. The residuals are lower when the predictive theory in price is either higher or lower. And the residuals are higher when the predictive price is near the center of the distribution. This verifies what we saw in the previous scatterplot that the variability of the residuals is increasing when bitcoin and ethereal prices are increasing.

(d) When we use the percentage changes instead of the actual prices, our analysis changes. The scatterplot of bitcoin and ethereum price percentage changes reveals a positive linear relationship. The variance appears constant. When we complete the simple linear regression and look at the scatterplot of the residuals against the predicted ethereum prices, they have a cloud shape which suggests that the assumptions of the simple linear regression were satisfied: the underlying relationship is linear and the variability of the residuals is constant. There is no evidence of outliers in either scatterplot.