1. Consider the following definitions:
   - $S_x = \Sigma\ x_i$
   - $S_y = \Sigma\ y_i$
   - $S_{xx} = \Sigma\ x_i^2$
   - $S_{xy} = \Sigma\ x_i\ y_i$

   For univariate linear regression, the best fit line is given $y = \alpha + \beta x$ where the optimal values of α and β are:

   - $\beta_{opt} = (n\ S_{xy} - S_x\ S_y\ ) / (n\ S_{xx} - S_x^2)$; and
   - $\alpha_{opt} = (S_y / n) - \beta_{opt}\ (S_x / n)$.

   (Read the wiki page on "simple linear regression".)

   For the training data below (given in the numerical example of the wiki page):

x_i	1.52	1.55	1.57	1.60	1.63	1.65	1.68	1.70	1.73	1.75	1.78	1.80	1.83	**Height (m)**
y_i	54.48	55.84	57.20	58.57	59.93	61.29	63.11	64.47	66.28	68.10	69.92	72.19	74.46	**Mass (kg)**

   a) Find the optimal line.
   b) Find the correlation coefficient, *r*, between height and mass.
   c) Identify the relationship between the slope of the best fit line to *r*.

   Answer:
   a) Plug the formulas in. $\beta_{opt} = 61.27$ and $\alpha_{opt} = -39.06$. (See wiki page for more details.)
   b) Correlation coefficient *r* is given by $\text{covariance}_{xy} / (\sigma_x\ \sigma_y)$, where $\text{covariance}_{xy} = (S_{xy} / n) - (S_x / n)\ (S_y / n)$, and σ_x , σ_y are the standard deviation of x and y. Plug in the formula to get r.
   c) A key purpose of this question is to identify that the slope β_{opt} is proportional to the correlation coefficient *r,* as in: $\beta_{opt} = r * \sigma_y / \sigma_x$.

2. In the previous question, we relate linear regression to correlation. Suppose that A, B, C and D are features in a regression problem with the outcome variable being L.
   a) Is it possible that A, as a single feature, is correlated to L but A is not a significant variable in the multiple regression model for L?
   b) If A is a significant variable in the multiple regression model for L, is it possible that A, as a single feature, is correlated to L?
   c) If A is a significant variable in the multiple regression model for L, is it possible that A, as a single feature, is *not* correlated to L?

Answer:

a) It is possible. A may be correlated to L. But the correlation may be through another attribute B. Thus, in the multiple regression model, it is sufficient to have B significant, and A is not needed.
b) It is possible. B is such an example in the scenario shown in a).
c) It is possible. C on its own may not be correlated to L. But once B is significant in the multiple regression model, the model may be improved by including C to capture the interaction between B and C.

   For example, height (B) may be a predictor of mass (L) for people. Gender (C) is not correlated to mass. But as a predictor for mass, height and gender combined may give a better model than height alone.