Introduction to Machine Learning
Homework 1: Simple Linear Regression*

Prof. Linda Sellie

1. Question 2 on page 52 from "An Introduction to Statistical Learning."
2. A university admissions office wants to predict the success of students based on their application material. They have access to past student records as training data.
   (a) To formulate this as a supervised learning problem, identify a possible target variable. This should be some variable that measures success in a meaningful way and can be easily collected (in an automated manner) by the university. There is no one correct answer to this problem.
   (b) Is the target variable continuous or discrete-valued?
   (c) State at least one possible variable that can act as the predictor for the target variable you chose in part (a).
   (d) Before looking at the data, would a linear model for the data be reasonable? If so, what sign do you expect the slope to be?
3. Suppose that we are given data samples (x_i, y_i):

x_i	0	1	2	3	4
y_i	0	2	3	8	17

   (a) What are the sample means, $\bar{x}$ and $\bar{y}$?
   (b) What are the sample variances and co-variances s_x^2, s_y^2 and s_{xy}?
   (c) What are the least squares parameters for the regression line

$$y = \beta_0 + \beta_1 x + \epsilon.$$

   (d) Using the linear model, what is the predicted value at $x = 2.5$?
   (e) Compute the MSE.
   (f) Calculate the R^2 (*confidence of determination*) and discuss the meaning of the number calculated.
   (g) If one of the examples was changed to:

x_i	0	1	2	3	4
y_i	0	2	3	8	15

*Most of these questions came from (or were inspired by) Prof. Sundeep Rangan.

how does the value of the parameters change?

If the training data was changed further to:

x_i	0	1	2	3	4
y_i	0	2	3	8	9

how much would the parameters change (be qualitative(e.g. not very much increase, drastically decrease, etc...)?

4. A medical researcher wants to model, $z(t)$, the concentration of some chemical in the blood over time. She believes the concentration should decay exponentially in that

$$z(t) \approx z_0 e^{-\alpha t}, \tag{1}$$

for some parameters z_0 and α. To confirm this model, and to estimate the parameters z_0, α, she collects a large number of time-stamped samples $(t_i, z(t_i))$, $i = 1, \ldots, N$. Unfortunately, the model (**??**) is non linear, so she can't directly apply the linear regression formula.

   (a) Taking logarithms, show that we can rewrite the model in a form where the parameters z_0 and α appear linearly.

   (b) Using the transform in part (a), write the least-squares solution for the best estimates of the parameters z_0 and α from the data.

   (c) Write a few lines of python code that you would compute these estimates from vectors of samples `t` and `z`.

5. Consider a linear model of the form,

$$y \approx \beta x,$$

which is a linear model, but with the intercept forced to zero. This occurs in applications where we want to force the predicted value $\hat{y} = 0$ when $x = 0$. For example, if we are modeling $y =$ output power of a motor vs. $x =$ the input power, we would expect $x = 0 \Rightarrow y = 0$.

   (a) Given data (x_i, y_i), write a cost function representing the residual sum of squares (RSS) between y_i and the predicted value $\hat{y}_i$ as a function of β.

   (b) Taking the derivative with respect to β, find the β that minimizes the RSS.