Introduction to Machine Learning
Problem Solutions Unit 2: Simple Linear Regression

Prof. Sundeep Rangan

1. (a) There are many possible target variables: GPA, time to graduate, ...

   (b) Both of the above examples are continuous.

   (c) Some choices: SAT score, high-school GPA, high-school class rank. Note that others, like extra curricular activities, are non-numeric and harder to represent as a numeric feature vector.

   (d) For university GPA vs. high-school GPA, a linear model would be a good place to start and would probably have a positive correlation.

2. (a) The sample means are:

$$\bar{x} = \frac{1}{N}\sum_i x_i = 2, \quad \bar{y} = \frac{1}{N}\sum_i y_i = 6,$$

   where $N = 5$ are the number of samples.

   (b) The (biased) sample variances and co-variances are

$$s_x^2 = \frac{1}{N}\sum_i (x_i - \bar{x})^2 = 2, \quad s_y^2 = \frac{1}{N}\sum_i (y_i - \bar{y})^2 = 37.2$$
$$s_{xy} = \frac{1}{N}\sum_i (y_i - \bar{y})(x - \bar{x}) = 8$$

   (c) The LS parameters are

$$\beta_1 = \frac{s_{xy}}{s_x^2} = 4, \quad \beta_0 = \bar{y} - \beta_1\bar{x} = -2.$$

   (d) The predicted value at $x = 2.5$ is

$$\hat{y} = -2 + 4(2.5) = 8.$$

3. (a) Let $y_i = \ln z(t_i)$ and $x_i = t_i$, then

$$y_i = \ln z(t_i) = \ln\left[z_0 e^{-\alpha t_i}\right] = \ln z_0 - \alpha t_i,$$

   where we have used the properties that $\ln(ab) = \ln a + \ln b$ and $\ln(e^x) = x$. Thus, if we define $\beta_0 = \ln z_0$ and $\beta_1 = -\alpha$ we get that

$$y_i = \beta_0 + \beta_1 x_i,$$

   which is a linear model.

(b) We first make the transformations, then perform the LS solution:

$$\begin{aligned}
&y_i = \ln z(t_i), \quad x_i = t_i, \\
&\bar{x} = \frac{1}{N}\sum_i x_i, \quad \bar{y} = \frac{1}{N}\sum_i y_i, \\
&s_x^2 = \frac{1}{N}\sum_i (x_i - \bar{x})^2, \quad s_y^2 = \frac{1}{N}\sum_i (y_i - \bar{y})^2, \quad s_{xy} = \frac{1}{N}\sum_i (y_i - \bar{y})(x - \bar{x}), \\
&\beta_1 = \frac{s_{xy}}{s_x^2}, \quad \beta_0 = \bar{y} - \beta_1 \bar{x}.
\end{aligned}$$

Then, we invert the equations $\beta_0 = \ln z_0$ and $\beta_1 = -\alpha$ to get the parameters in the original model,

$$\alpha = -\beta_1, \quad z_0 = e^{\beta_0}.$$

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

```
# Transform the variables
x = t
z = np.log(z)

# Compute the sample means and the difference from the sample means
xm = np.mean(x)
ym = np.mean(y)
x1 = x − xm
y1 = y − ym

# Compute the variances and covariances
sxx = np.mean(x1**2)
sxy = np.mean(x1*y1)

# Compute the LS coefficients
b1 = sxy/sxx
b0 = ym−b1*xm

# Get back the coefficients in the original model
alpha = −b1
z0 = exp(b0)
```

4. (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. The RSS is

$$\text{RSS}(\beta) := \sum_{i=1}^{N} (y_i - \beta x_i)^2.$$

(b) Taking the derivative with respect to β we get

$$\frac{\partial \mathrm{RSS}(\beta)}{\partial \beta} = \sum_{i=1}^{N} 2(y_i - \beta x_i)(-x_i) = 0$$
$$\Rightarrow \beta \sum_i x_i^2 = \sum_i x_i y_i \Rightarrow \beta = \frac{\sum_i x_i y_i}{\sum_i x_i^2}.$$