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

Prof. Sundeep Rangan

1. A university admissions office wants to predict the success of students based on their application material. They have access to past student records to learn a good algorithm.

   (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?

   Solution:

   (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. 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 $x = 2.5$?

Solution:

(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 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 (1) 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`.

Solution:

(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. 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.

Solution:

(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

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

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

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

$$\Rightarrow \beta \sum_i x_i^2 = \sum_i x_i y_i \Rightarrow \beta = \frac{\sum_i x_i y_i}{\sum_i x_i^2}.$$