Unit 2: Simple Linear Regression Problems by Prof. Pei Liu

Vaibhav Singh

16 September 2021

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.

   Solution (a):

   Lets define the successful student as someone who is able to complete his graduation in time. So we can declare our target as a binary variable having either 1 (Corresponding to when student completes his degree on time) or 0 (when the student is not able to complete his degree on time). This makes our problem statement as a binary classification.

   (b) Is the target variable continuous or discrete-valued?

   Solution (b):

   Discrete since the target variable can be either 1 or 0.

   (c) State at least one possible variable that can act as the predictor for the target variable you chose in part (a).

   Solution (c):

   School Marks, or GPA score can be used as a predictor for target variable.

   (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 (d):

   From a general assumption that students scoring high marks in school tend to perform better in colleges, we expect to have high positive correlation between predictor variable and target variable, so we can say that a linear model is reasonable. The slope should be positive.

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 $\bar{y}$?

Solution (a):

- Mean $= (1/n)\sum_{i=1}^{n}(x_i)$, where n = 5
  $\bar{x}$=2
- Mean $= (1/n)\sum_{i=1}^{n}(y_i)$, where n = 5
  $\bar{y}$=6

(b) What are the sample variances and co-variances s_x^2, s_y^2 and s_{xy}?

Solution (b):

- Variance(s_x^2) $= (1/n)\sum_{i=1}^{n}(x_i - \bar{x})^2$, where n = 5
  s_x^2= 2
- Variance(s_y^2) $= (1/n)\sum_{i=1}^{n}(y_i - \bar{y})^2$, where n = 5
  s_y^2= 37.2
- Variance(s_{xy}) $= (1/n)\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, where n = 5
  s_{xy}= 8

(c) What are the least squares parameters for the regression line

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

Solution (c):

- $\beta_1 = s_{xy}/s_x^2$. $(8/2 = 4)$
- $\beta_0 = \bar{y} - \beta_1\bar{x}$ $(6 - 4 * 2 = -2)$

(These equations come from minimising the RSS as taught in class and slide(30/37)

(d) Using the linear model, what is the predicted value at $x = 2.5$?

Solution (d):

- $-2 + 2.5 * 4 = 8$

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.

Solution (a):

$$log(z(t)) = log(z_0 e^{-\alpha t}) \tag{2}$$

$$z' = log(z(t)) \tag{3}$$

$$z' = log(z_0) - \alpha t \tag{4}$$

Now $\beta_0 = log(z_0)$ and $\beta_1 = -\alpha$
Finally we have the linear regression equation in parameters $\beta_0,\ \beta_1$

$$z' = \beta_0 + \beta_1 t$$

(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.

Solution (b):

$$\beta_1 = s_{tz'}/s_{t^2}$$
$$\alpha = -s_{tz'}/s_{t^2}$$

$$\beta_0 = \bar{z'} - \beta_1 \bar{t}$$
$$z_0 = e^{(\bar{z'} - (\frac{s_{tz'}}{s_{t^2}})\bar{t})}$$

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

```
import numpy as np

def fit_model(t, z):
    t_mean = np.mean(t)
    z_mean = np.mean(z)

    t_variance = np.mean((t - t_mean)**2)
    zt_covariance = np.mean((t-t_mean)*(z-z_mean))

    alpha = -(zt_covariance/t_variance)

    z0 = np.exp(z_mean + alpha*t_mean)

    return z0, alpha

z_0 , alpha = fit_model(t,z)
```

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.

Solution (a):

$Cost = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
$Cost = \sum_{i=1}^{n}(y_i - \beta x_i)^2$

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

Solution (b):

$$\frac{\partial Cost}{\partial \beta} = \frac{\partial(\sum_{i=1}^{n}(y_i - \beta x)^2)}{\partial \beta} \tag{5}$$

Putting $\frac{\partial Cost}{\partial \beta} = 0$

We get, $-2\sum_{i=1}^{n}(y_i - \beta x_i)x_i = 0$
Solving further

$\sum_{i=1}^{n}(y_i x_i) = \beta \sum_{i=1}^{n}(x_i^2)$

$$\beta = \frac{\sum_{i=1}^{n}(y_i x_i)}{\sum_{i=1}^{n}(x_i^2)} \tag{6}$$

Comments/Remarks: