Classwork 03: Simple Linear Regression
CPE 490/590 ST

Instructor: Rahul Bhadani

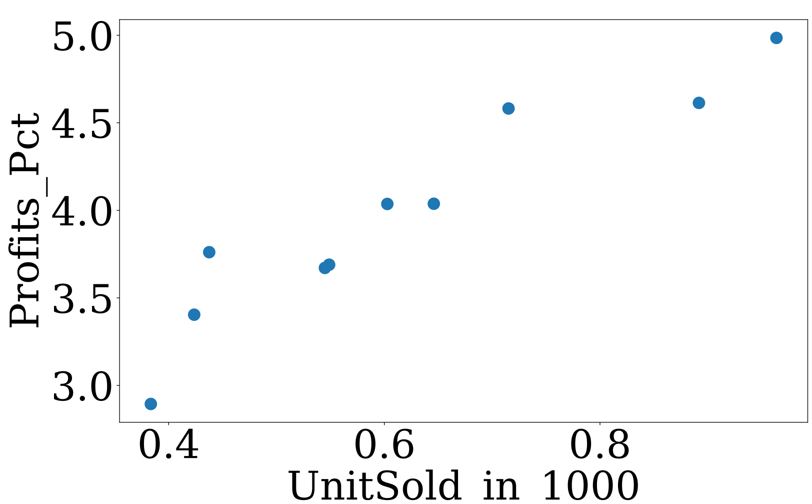
50 points

1 Simple Linear Regression

Consider the dataset that exhibits linear relationship between Units Sold in 1000 (predictor variable x), and profit in percentage (y):

UnitSold_in_1000 x_i	**Profits_Pct** y_i
0.54	3.69
0.72	4.58
0.6	4.03
0.54	3.67
0.42	3.40
0.64	4.03
0.43	3.76
0.89	4.61
0.96	4.98
0.38	2.89

$n = 10$.



Calculate the following quantity ($\sum_{i=1}^{10}$ is shorthanded as $\sum$) (10 points):

1. $\sum x_i = \boxed{}$

2. $\sum y_i = \boxed{}$

3. $\overline{x} = \boxed{}$

4. $\overline{y} = \boxed{}$

5. $\sum(x_i - \overline{x}) = \boxed{}$

6. $\sum(y_i - \overline{y}) = \boxed{}$

7. $\sum(x_i - \overline{x})^2 = \boxed{}$

8. $\sum(x_i y_i) = \boxed{}$

9. $\sum x_i^2 = \boxed{}$

10. $\sum y_i^2 = \boxed{}$

In an ideal scenario, the data was obtained from a linear model $Y = Xw_1 + w_0$. The given data comes follows the $y_i = x_i w_1 + w_0 + \epsilon_i$ where ϵ_i is the error introduced by gather data such that $\epsilon_i \sim \mathcal{N}(\mu = 0, \sigma = 0.3)$.

We will fit our data into the linear model $y_i = x_i w_1 + w_0$ with the hope of minimizing the cost function (or error):

$$Q = \frac{1}{n}\sum_i (y_i - \hat{y})^2 \tag{1}$$

where $\hat{y} = x_i \hat{w}_1 + \hat{w}_0$.

The least-square solution gives the normal equation by solving:

$$\begin{aligned} \sum y_i &= n\hat{w}_0 + \hat{w}_1 \sum x_i \\ \sum x_i y_i &= \hat{w}_0 \sum x_i + \hat{w}_1 \sum x_i^2 \end{aligned} \tag{2}$$

which gives

$$\hat{w}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \underline{\hspace{8cm}} = \boxed{\phantom{\rule{2cm}{2cm}}} \tag{3}$$

(4 Points)

and

$$\hat{w}_0 = \overline{y} - \hat{w}_1 \overline{x} = \boxed{\phantom{\rule{2cm}{2cm}}} \tag{4}$$

(2 Points)

Alternatively, we can take the derivative of the cost function which can be re-written as

$$Q = \frac{1}{n} \sum (y_i - (w_1 x_i + w_0))^2 \tag{5}$$

Taking the partial derivative with respect to w_1 and w_0:

$$\frac{\partial Q}{\partial w_1} = \tag{6}$$

(5 Points)

$$\frac{\partial Q}{\partial w_0} = \tag{7}$$

(5 Points)

Setting above to 0 gives the following set of equations:

$$
\begin{aligned}
w_1 &= \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} \\
w_0 &= \frac{1}{n}(\sum_{i=1}^{n} y_i) - w_1 \frac{1}{n}(\sum_{i=1}^{n} x_i)
\end{aligned} \tag{8}
$$

Putting the value in the above equations should give an estimated $\hat{w}_0$ and $\hat{w}_1$. Write down what you got in that Equation.

$$\hat{w}_1 = \tag{9}$$

$$\hat{w}_0 = \tag{10}$$

(5 Points)

Using $\hat{w}_0$ and $\hat{w}_1$, we can obtain $\hat{y}_i$, the estimated response. Using estimated response and actual response, we can calculate Sum of Square Error (SSE) and Total Square Sum (SSTO) as:

$$
\begin{aligned}
\text{SSE} &= \sum (y_i - \hat{y}_i)^2 = \\
\text{MSE} &= \frac{\text{SSE}}{n-2} = \\
\text{SSTO} &= \sum (y_i - \overline{y})^2 =
\end{aligned} \tag{11}
$$

(2 Points)

Calculate the goodness of fit using the coefficient of determination R^2:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SSTO}} = \tag{12}$$

(2 Points)

Based on the value of R^2 how good fit is the linear model on the given dataset?

(2 Points)

Next, we calculate the standard deviation on the estimated coefficients or weight ($\hat{w}_1$, and $\hat{w}_0$).

$$s^2[\hat{w}_1] = \frac{\text{MSE}}{\sum(x_i - \overline{x})^2} = \tag{13}$$

(2 Points)

$$s^2[\hat{w}_0] = \mathrm{MSE}\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}\right) = \tag{14}$$

(2 Points)

Now, we need to calculate the 95% confidence interval on the estimated $\hat{w}_1$ and $\hat{w}_0$ for which $\alpha = 0.05$ (use $t(0.975, 8) = 2.306$):

$$\hat{w}_1 \pm t(1 - 0.05/2, 10 - 2)s^2[\hat{w}_1] \tag{15}$$

(2 Points)

$$\hat{w}_0 \pm t(1 - 0.05/2, 10 - 2)s^2[\hat{w}_0] \tag{16}$$

(2 Points)

Now, consider a new data point $x_{new} = 0.47$ Units Sold (in 1000). Based on the linear regression model that we developed, what is the estimated/predicted profit percent for the given quantity **(5 Points)**?