DSC 40A - Discussion 03

January 28, 2020

Recall that the least squares solutions to the problem of fitting a straight line, $h(x) = w_1 x + w_0$, to the data (x_i, y_i) are:

$$w_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$w_0 = \bar{y} - w_1 \bar{x}$$

Where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Problem 1.

Consider the problem of fitting a function of the form $h(x) = b_1 \sin x + b_0$ to the data $(x_1, y_1), \ldots, (x_n, y_n)$. What are the least squares solutions for b_1 and b_0?

Solution: Since h is linear in terms of b_1 and b_0, we can substitute $sin(x_i)$ for x_i.

$$b_1 = \frac{\sum_{i=1}^{n}(sin(x_i) - (1/n)\sum_{i=1}^{n} sin(x_i))(y_i - \bar{y})}{\sum_{i=1}^{n}(sin(x_i) - (1/n)\sum_{i=1}^{n} sin(x_i))^2}$$

$$b_0 = \bar{y} - b_1 \frac{1}{n}\sum_{i=1}^{n} sin(x_i)$$

Problem 2.

Suppose that the age and net worth of each representative in the US House of Representatives is collected, and least squares is used to fit a linear prediction function $H(x) = w_1 x + w_0$ for predicting a representative's worth from their age.

Now suppose that the world's richest person, Jeff Bezos (net worth \$110 billion), has decided to run for the US House of Representatives in a special election! Assume that Bezos replaces a congressperson who is the same age as him: 56 years old. The net worth of the replaced representative was \$1 million. Assume, too, that that the *variance* in the age of the House of Representatives is 10 years.[1] If a new linear predictor $H'(x) = w_1' x + w_0'$ is fit using the new data with Bezos included, what is the difference between the new slope w_1' and the old? You may assume that Bezos is the only new member of the House of Representatives.

For this problem, you'll need to use the fact that the US House of Representatives has 435 members and that their average age is 58 years.

[1] recall that the variance is $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Solution: First, let's simplify the numerator of the expression for the least squares solution for w_1.

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}(x_i-\bar{x})y_i-(x_i-\bar{x})\bar{y}\\ &= \sum_{i=1}^{n}(x_i-\bar{x})y_i-\bar{y}\sum_{i=1}^{n}(x_i-\bar{x})\\ &= \sum_{i=1}^{n}(x_i-\bar{x})y_i-0\\ &= \sum_{i=1}^{n}(x_i-\bar{x})y_i\end{aligned}$$

Also, notice that the denominator is $n*var(x)$.

Let (x_b, y_b) represent the age and net worth of Jeff Bezos and (x_r, y_r) represent the age and net worth of the person he replaced. Assume without loss of generality that Bezos replaced the jth congress member.

We can now write w_1'. Remember, Jeff Bezos replaced the jth congress member in the dataset used to estimate w_1'.

$$\begin{aligned}w_1' &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{n*var(x)} && \text{(least squares solution for linear model)}\\ &= \frac{(x_b-\bar{x})y_b+\sum_{i\neq j}(x_i-\bar{x})y_i}{n*var(x)} && \text{(separate Bezos)}\end{aligned}$$

Similarly, we can write w_1. Remember, the dataset used to estimate w_1 does not have Bezos but the person he replaced as the jth congress member.

$$\begin{aligned}w_1 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{n*var(x)} && \text{(least squares solution for linear model)}\\ &= \frac{(x_r-\bar{x})y_r+\sum_{i\neq j}(x_i-\bar{x})y_i}{n*var(x)} && \text{(separate the person Bezos replaces)}\end{aligned}$$

We need to find the difference between the two slopes.

$$\begin{aligned}w_1'-w_1 &= \frac{(x_b-\bar{x})y_b-(x_r-\bar{x})y_r+\sum_{i\neq j}(x_i-\bar{x})y_i-\sum_{i\neq j}(x_i-\bar{x})y_i}{n*var(x)}\\ &= \frac{(x_b-\bar{x})y_b-(x_r-\bar{x})y_r}{n*var(x)}\\ &= \frac{(56-58)11*10^{10}-(56-58)10^6}{435*10} && \text{(Substitute in } x_b, x_r, \bar{x}, y_b, y_r)\\ &= \frac{-2(11*10^{10}-10^6)}{4350}\\ &= \frac{-2*(109,999,000,000)}{4350}\\ w_1'-w_1 &= -50,574,252.9\end{aligned}$$

Problem 3.

A *Boolean feature* is one that is either true or false. For example, when predicting the price of a car, a useful feature might be whether or not the car has an automatic transmission. We can perform least squares

regression with Boolean features by "encoding" true and false as numbers: a common choice is to encode true as 1 and false as 0.

In this problem, suppose we have a data set $(x_1, y_1), \ldots, (x_n, y_n)$ of n cars, where the feature x_i is either 1 or 0 (has automatic transmission, or does not) and where y_i is the price of the car. Furthermore, suppose that n_1 of the cars have automatic transmissions, while n_0 do not. Assume for simplicity that the data are sorted so that the first n_1 cars do not have automatic transmissions while the rest do, so that $x_1, \ldots, x_{n_0} = 0$ and $x_{n_0+1}, \ldots, x_n = 1$.

a) Show that $\bar{x} = \frac{n_1}{n}$.

Solution:

$$
\begin{aligned}
\bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\
&= \frac{1}{n} \left(\sum_{i=1}^{n_0} x_i + \sum_{i=n_0+1}^{n} x_i \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^{n_0} 0 + \sum_{i=n_0+1}^{n} 1 \right) \\
&= \frac{1}{n} (n - n_0 - 1 + 1) \\
&= \frac{n - n_0}{n} \\
\bar{x} &= \frac{n_1}{n}
\end{aligned}
$$