

# Report

## SPS Coursework: An Unknown Signal

Emil Centiu, z118810

March 24, 2020

### 1 Least Squares Regression

Because we can't assume anything about the random error term (i.e. follows a normal distribution like in MLE), Least squares was the go-to method to find the best fitting line.

The goal is to deterministically minimise the sum of squared differences between the observed value of the dependent variable ( $y_i$ ) and the predicted value of the dependent variable ( $\hat{y}$ ), that is provided by the regression function. In other words, we need to find  $a, b$  such that the residual error  $R(a, b)$  is minimised, where

$$R(a, b) = \sum_{i=1}^N (\hat{y}_i - y_i)^2 = \sum_{i=1}^N ((a + bx_i) - y_i)^2$$

We can observe that  $R(a, b)$  plots as an elliptic paraboloid, so there is only one critical point, and according to Fermat's Theorem, it is a local extreme, hence the minimum value of the function. To find the pair  $(a, b)$ , we can just calculate the critical point from the partial derivatives with respect to each variable:

$$\frac{\delta R}{\delta a} = \frac{\delta}{\delta a} \sum_{i=1}^N (y_i^2 + a^2 + 2abx_i + b^2x_i^2 - 2y_ia - 2y_ibx_i) = \sum_{i=1}^N (2a + 2bx_i - 2y_i) = -2 \sum_{i=1}^N (y_i - (a + bx_i)) = 0$$

$$\frac{\delta R}{\delta b} = \frac{\delta}{\delta b} \sum_{i=1}^N (y_i^2 + a^2 + 2abx_i + b^2x_i^2 - 2y_ia - 2y_ibx_i) = \sum_{i=1}^N (2ax_i + 2bx_i^2 - 2y_ix_i) = -2 \sum_{i=1}^N x_i(y_i - (a + bx_i)) = 0$$

From the first equation, we can get  $a$ :

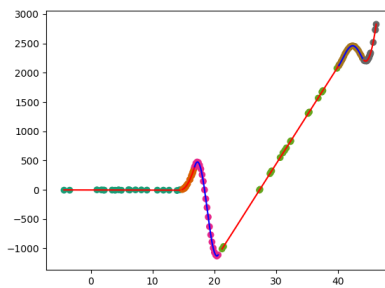
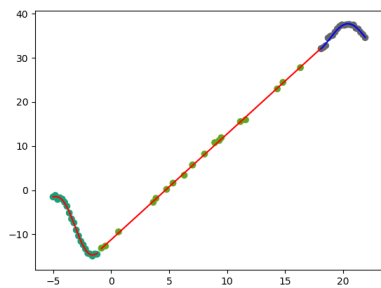
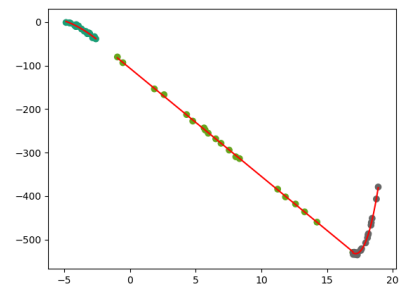
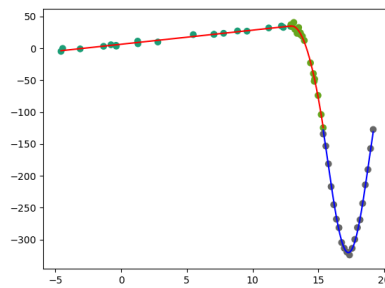
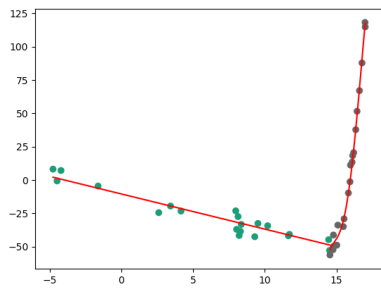
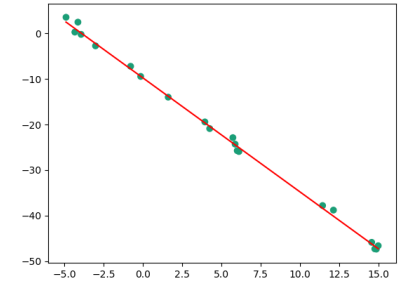
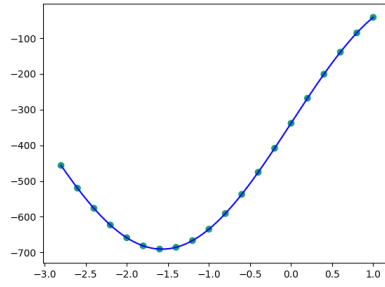
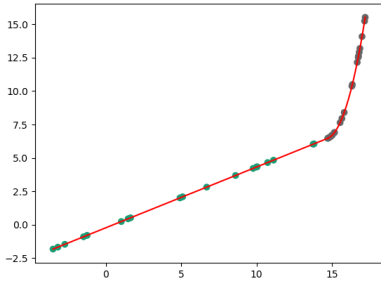
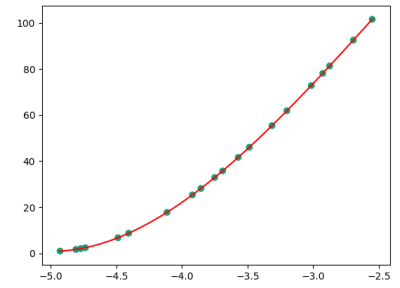
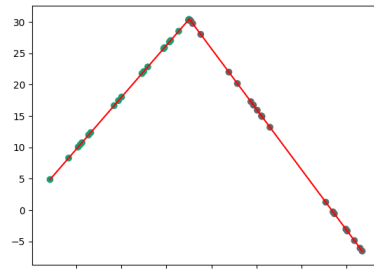
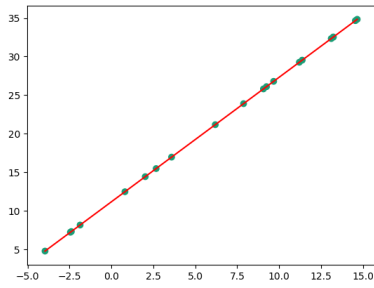
$$\begin{aligned} -2 \sum_{i=1}^N (y_i - (a + bx_i)) = 0 &\Leftrightarrow \sum_{i=1}^N (y_i - a - bx_i) = 0 \Leftrightarrow \sum_{i=1}^N y_i - Na + b \sum_{i=1}^N b_i = 0 \Leftrightarrow \\ a &= \frac{\sum_{i=1}^N y_i - b \sum_{i=1}^N x_i}{N} \Leftrightarrow a = \bar{y} - b\bar{x} \end{aligned}$$

where  $\bar{x}$  is the mean value of the  $x$  coordinates, and  $\bar{y}$ , the mean value of the  $y$  coordinates. An interesting observation is that the regression function always goes through the centroid (the point of coordinates  $(\bar{x}, \bar{y})$ ).

We can get  $b$  from the second equation, by substituting the value of  $a$ :

$$\begin{aligned} -2 \sum_{i=1}^N x_i(y_i - (a + bx_i)) = 0 &\Leftrightarrow \sum_{i=1}^N x_i(y_i - a - bx_i) = 0 \Leftrightarrow \sum_{i=1}^N x_iy_i - \sum_{i=1}^N x_ia - \sum_{i=1}^N x_i^2b = 0 \\ &\Leftrightarrow \sum_{i=1}^N x_iy_i - \sum_{i=1}^N x_i(\bar{y} - b\bar{x}) - \sum_{i=1}^N x_i^2b = 0 \Leftrightarrow \sum_{i=1}^N x_iy_i - \bar{y} \sum_{i=1}^N x_i + b\bar{x} \sum_{i=1}^N x_i - b \sum_{i=1}^N x_i^2 = 0 \\ &\Leftrightarrow \sum_{i=1}^N x_iy_i - \bar{y}N\bar{x} + N\bar{x}^2 - b \sum_{i=1}^N x_i^2 = 0 \Leftrightarrow b = \frac{\sum_{i=1}^N x_iy_i - N\bar{x}\bar{y}}{\sum_{i=1}^N x_i^2 - N\bar{x}^2} \end{aligned}$$

## 2 figures/plots



## 3 overfitting and new data

## 4 implementation