Authors: Wojciech Ptaś (7042843), Mhd Jawad Al Rahwanji (7038980)

Problem 5 - Ex Pluribus Unum

1. We have:

$$\bar{y} = \frac{K}{n} \sum_{i=\frac{n(k-1)}{K}+1}^{\frac{nk}{K}} y_i, \quad \bar{x} = \frac{K}{n} \sum_{i=\frac{n(k-1)}{K}+1}^{\frac{nk}{K}} x_i$$

   Our parameters would take the forms:

$$\hat{\beta}_0^{(k)} = \bar{y} - \hat{\beta}_1^{(k)} \bar{x}$$

$$\hat{\beta}_1^{(k)} = \frac{\sum_{i=\frac{n(k-1)}{K}+1}^{\frac{nk}{K}} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=\frac{n(k-1)}{K}+1}^{\frac{nk}{K}} (x_i - \bar{x})^2}$$

2. Our model prediction takes the form:

$$\hat{y}_i = \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_0^{(k)} + \hat{\beta}_1^{(k)} x_i$$

3. Gauss-Markov theorem states that an OLS model provides the best reducible error when compared to any other unbiased model. Since we are dividing our data to subsets and training an OLS on each we introduce variance because we aggregate the irreducible error multiple times which amplifies it.
4. Splitting the data with K = 3 would leave us with 4 datasets:

$$S_1 = [-1, -0.\overline{3}[$$

$$S_2 = [-0.\overline{3}, 0.\overline{3}[$$

$$S_3 = [0.\overline{3}, 1[$$

$$S_4 = [1, \infty[$$

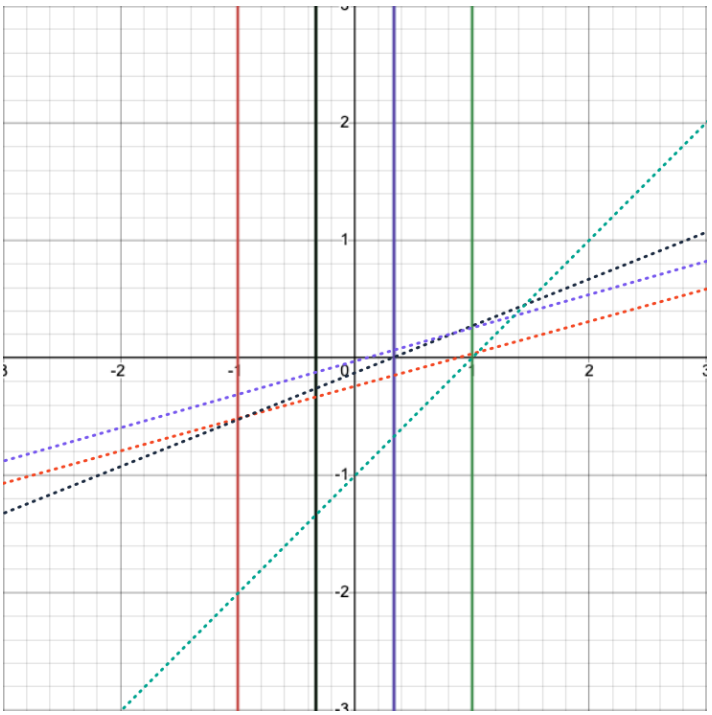

Fig. 1 — Illustration of each OLS per S_k

5. We have a different model for each dataset, thus, different coefficients per model. Our model would have the following form:

$$\hat{y}_0 = \hat{\beta}_0^{(k)} + \hat{\beta}_1^{(k)} x_0 : x_0 \in S_k$$

6. The problem here would be a potential lack of support for some $OLS^{(k)}$, if we're unlucky we may end up with a very low confidence model in some split. To fix this, we may use quantiles instead, going for equal height over equal width.
7. Technically yes, but it's not a good idea. If we recursively split each dimension into zones we'll very quickly run out of examples to cover each combination of zones. The higher dimensionality the more examples we require but in this case we would be doing the opposite and harming the model further by reducing support per model (despite using quantiles).