EECS 127/227AT Optimization Models in Engineering
Spring 2020 Discussion 3

1. Linear regression versus orthogonal distance regression

In this exercise, we explore two regression techniques:

- linear regression (LR), generally solved via least-squares, and
- orthogonal distance regression (ODR), solved using PCA.

We will examine these regression methods in turn and compare their possible use cases.

In general, regression is used to model the relationship between observed input data and corresponding output data. In this problem, we consider n input data points $\vec{a}_i \in \mathbb{R}^d$ and n corresponding output data points $b_i \in \mathbb{R}$. Note that the input comprises d real-valued features and the output is a real-valued scalar.

In the case of *linear regression* (LR), each output b_i is assumed to be a linear combination of the features of the input $\vec{a}_i$, i.e., $b_i \sim \vec{a}_i^\top \vec{x}$, where $\vec{x} \in \mathbb{R}^d$ is a d–dimensional vector of weights used in the linear combination. We define the LR computation as finding the $\vec{x}$ that minimizes the sum of the squared errors between the outputs b_i and the predicted outputs $\vec{a}_i^\top \vec{x}$ (least-squares), i.e., computing

$$\vec{x}^*_{\text{LR}} = \arg\min_{\vec{x}} \sum_i (\vec{a}_i^\top \vec{x} - b_i)^2.$$

Assume for the entirety of this problem that the data are centered, i.e., $\forall j = 1, \ldots, d$, $\sum_{i=1}^n a_{ij} = 0$ and $\sum_{i=1}^n b_i = 0$. This means means that all trendlines we compute (during LR, and later, ODR) will pass through the origin.

(a) Show that the LR computation can be formulated as a least squares problem of the form

$$\vec{x}^*_{\text{LR}} = \arg\min_{\vec{x}} \|A\vec{x} - \vec{b}\|_2^2.$$

State A and $\vec{b}$.

Solution: Data matrix is $A = \begin{bmatrix} \leftarrow \vec{a}_1^\top \rightarrow \\ \vdots \\ \leftarrow \vec{a}_i^\top \rightarrow \\ \vdots \\ \leftarrow \vec{a}_n^\top \rightarrow \end{bmatrix}$. Output vector is $\vec{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_i \\ \vdots \\ b_n \end{bmatrix}$

(b) We now consider an example LR computation in which $d = 1$ and $n = 3$. Let $a_1 = -1$, $a_2 = 0$, $a_3 = 1$ and $b_1 = -1$, $b_2 = -1$, $b_3 = 2$. Compute the best fit LR regression line — in this case, a 1–dimensional scalar — x^*_{LR}.

Solution:

$$x^*_{\text{LR}} = \arg\min_x (-x+1)^2 + (1)^2 + (x-2)^2 = \arg\min_x x^2 - 2x + 1 + 1 + x^2 - 4x + 4$$
$$= 2(x - \frac{3}{2})^2 + 6 - \frac{9}{2} = 2(x - \frac{3}{2})^2 + \frac{3}{2}$$
$$= \frac{3}{2}$$

This can also be solved using the standard LS formula $x^*_{\text{LR}} = (A^\top A)^\dagger A^\top b$

Teaching Notes: I recommend solving this using the standard LS formula.

(c) Now let $a_1 = -1$, $a_2 = -1$, $a_3 = 2$ and $b_1 = -1$, $b_2 = 0$, $b_3 = 1$. Compute the best fit LR regression value x^*_{LR}.

Solution:

$$x^*_{\text{LR}} = \operatorname{argmin}_x (1-x)^2 + (-1x)^2 + (-1+2x)^2 = \operatorname{argmin}_x x^2 - 2x + 1 + x^2 + 4x^2 - 4x + 1$$
$$= 6(x - \frac{1}{2})^2 + 2 - \frac{6}{4} = 6(x - \frac{1}{2})^2 + \frac{1}{2}$$
$$= \frac{1}{2}$$

(d) Note that the computations in (b) and (c) above are performed on the same values; inputs and outputs are simply switched. Plot these data points on the a–b plane and plot the trendlines corresponding to each x^*_{LR} from (b) and (c) above. Are these trendlines the same? Reason geometrically about why this is the case.

Solution: See figure 1: in the case where $d = 1$, the LR (b) finds the line that goes through the origin and minimizes the sum of the squares of the vertical distance (y-distance) from the points to the line. The LR (c) finds the line that goes through the origin and minimizes the sum of the squares of the horizontal distance (x-distance).

In some cases, we may not have a preference for which values are designated inputs or outputs and want to avoid differences in regression result based on our choice. In this case, we can use *orthogonal distance regression* (ODR) to compute the regression line that minimizes the sum of the squares of the orthogonal distances of each data point to the line.

To perform the ODR computation, we define vectors $\vec{z}_i$, each of which concatenates all inputs and outputs at observation point i, i.e.,

$$\vec{z}_i = \begin{bmatrix} \uparrow \\ \vec{a}_i \\ \downarrow \\ b_i \end{bmatrix}.$$

The ODR regression line is then the direction $\vec{x} \in \mathbb{R}^{d+1}$ such that the sum of squares of the (orthogonal) distances between the points $\vec{z}_i$ and their projections on the line passing through the origin along direction $\vec{x}$ are minimized.

$$\vec{x}^*_{\text{ODR}} = \arg\min_{\vec{x}:\|\vec{x}\|_2=1} \sum_i \|\vec{z}_i - \text{proj}_{\vec{x}}(\vec{z}_i)\|_2^2 \tag{1}$$
$$= \operatorname{argmin}_{\vec{x}:\|\vec{x}\|_2=1} \sum_i \min_{v_i} \|\vec{z}_i - v_i\vec{x}\|_2^2. \tag{2}$$

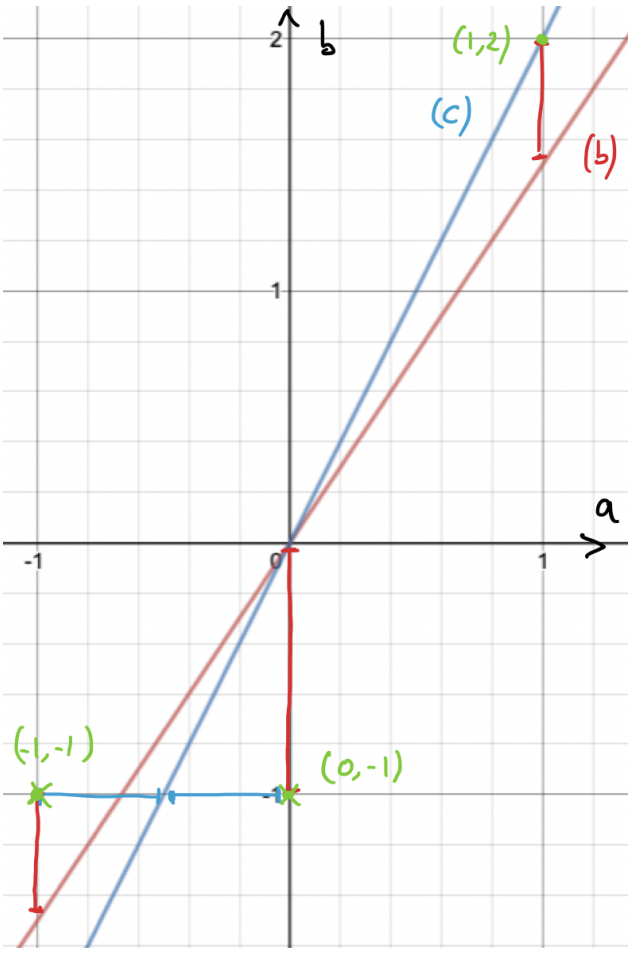

Figure 1: Illustration of solution of (b) and (c) on the $a - b$ plane

(e) Show that the ODR computation can be formulated as the problem of finding the eigenvector corresponding to the largest eigenvalue of $H = \sum_{i=1}^{n} \vec{z}_i \vec{z}_i^\top$, i.e.,

$$\vec{x}^*_{\text{ODR}} = \arg \max_{\vec{x}:\|\vec{x}\|_2=1} \vec{x}^\top H \vec{x}.$$

Solve ODR using the singular value decomposition (SVD) of the "augmented" data matrix:

$$Z = \begin{bmatrix} \leftarrow \vec{z}_1^\top \rightarrow \\ \vdots \\ \leftarrow \vec{z}_i^\top \rightarrow \\ \vdots \\ \leftarrow \vec{z}_n^\top \rightarrow \end{bmatrix}.$$

Solution: The inner minimization problem can be solved using projections to obtain,

$$\text{proj}_{\vec{x}}(\vec{z}_i) = \underset{v_i}{\text{argmin}} \|\vec{z}_i - v_i \vec{x}\|_2^2 = \vec{z}_i^\top \vec{x}$$

Substituting this into expression from original definition of ODR we have,

$$\begin{aligned}
\sum_i \|\vec{z}_i - (\vec{z}_i^\top \vec{x})\vec{x}\|_2^2 &= \sum_i \vec{z}_i^\top \vec{z}_i - 2\vec{x}^\top \vec{z}_i \vec{z}_i^\top \vec{x} + \vec{x}^\top \vec{z}_i \vec{z}_i^\top \vec{x} \vec{x}^\top \vec{x} \\
&= \left(\sum_i \vec{z}_i^\top \vec{z}_i\right) - \left(\sum_i \vec{x}^\top \vec{z}_i \vec{z}_i^\top \vec{x}\right) \\
&= \left(\sum_i \vec{z}_i^\top \vec{z}_i\right) - \vec{x}^\top \left(\sum_i \vec{z}_i \vec{z}_i^\top\right) \vec{x} \\
\underset{\vec{x},\|\vec{x}\|_2=1}{\text{argmin}} \sum_i \min_{v_i} \|\vec{z}_i - v_i \vec{x}\|_2^2 &= \underset{\vec{x},\|\vec{x}\|_2=1}{\text{argmin}} \left(\sum_i \vec{z}_i^\top \vec{z}_i\right) - \vec{x}^\top \left(\sum_i \vec{z}_i \vec{z}_i^\top\right) \vec{x} \\
&= \underset{\vec{x},\|\vec{x}\|_2=1}{\text{argmax}} \vec{x}^\top \left(\sum_i \vec{z}_i \vec{z}_i^\top\right) \vec{x} \\
&= \underset{\vec{x},\|\vec{x}\|_2=1}{\text{argmax}} \vec{x}^\top H \vec{x}
\end{aligned}$$

The solution is the eigenvector corresponding to the largest eigenvalue of H. Note that the second line follows because we enforce in our optimization that $\vec{x}$ is unit norm, and thus $\vec{x}^\top \vec{x} = 1$.

If $Z = \sum_{i=i}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$, $\vec{x}^*_{\text{ODR}} = \vec{u}_1$, and $H = ZZ^\top$.

(f) Considering $d = 1$ and $n = 3$, numerically find the results of the ODR when $a_1 = -1$, $a_2 = 0$, $a_3 = -1$ and $b_1 = -1$, $b_2 = -1$, $b_3 = 2$ using iPython. Are the results similar if a and b values are switched?

Teaching Notes: You have to do the other way around as well. It doesn't have to be done in discussion, but you can ask them to do it.

Solution:

$$\vec{x}^*_{\text{ODR}} = \begin{bmatrix} 1 \\ 1.87 \end{bmatrix}$$

```
X = np.array([-1, 0, 1])
Y = np.array([-1,-1,2])
Z = np.concatenate((X.reshape((1,3)),Y.reshape((1,3))), axis=0)
U,sigma,V = np.linalg.svd(Z)
print(U)
print(U[0][1]/U[0][0])
```

```
[[-0.47185793 -0.8816746 ]
 [-0.8816746   0.47185793]]
1.8685170918213294
```

Figure 2: iPython code to find ODR for question (f)

(g) Compare the two techniques. If your goal is to understand a invertible symmetric relationship between the "Parent height" and the "Child height", which method would you prefer? See figure 3.

Solution: ODR has the nice property that it treats the "measurement matrix" and the "measurements" symmetrically — so unlike Lr, it does not give you two different answers when you switch the roles of $\vec{a}$ and $\vec{b}$, in a simple scalar regression case. So if you really trust your "experiment" i.e. the values in the matrix A then you can use LR, but if you want to just understand the relationship between the (a_i, b_i) pairs in a symmetric way, then ODR will help with that. In the case of figure 3, since you want to be able to just understand the relationships between the two heights, ODR might be a better choice.

(h) For practice: Use the SVD of A to write an expression for the least squares solution of $A\vec{x} = \vec{b}$.

Solution: If $A = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^{\top} = U\Sigma V^{\top}$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$ (different σ_i, $\vec{u}_i$ and $\vec{v}_i$ than the previous question), we have $LR = \sum_{i=i}^{r} \frac{1}{\sigma_i} \vec{v}_i \vec{u}_i^{\top} \vec{b}$.

Teaching Notes: (h) is optional.

References

[Gal86] Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

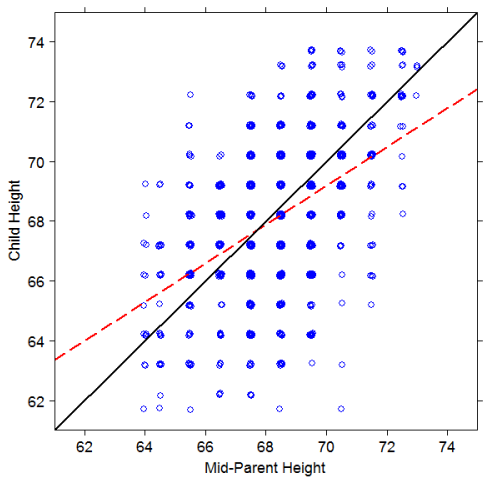

Figure 3: Illustration of the correlation between the heights of adults and their parents. In red is the result of linear regression. In black is the result of the orthogonal direction regression. This work has first been done by Galton in 1886 ([Gal86]). Using linear regression, Galton remarks that: "It appeared from these experiments that the offspring did not tend to resemble their parents in size, but always to be more mediocre than they – to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were small." Figure comes from `https://select-statistics.co.uk/blog/regression-to-the-mean-as-relevant-today-as-it-was-in-the-1900s/` The original text can be found at `https://www.jstor.org/stable/2841583`.