Exercise: Partial derivative of the RSS

Define

$$RSS(\mathbf{w}) = ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 \tag{1}$$

1. Show that

$$\begin{aligned} \frac{\partial}{\partial w_k} RSS(\mathbf{w}) &= a_k w_k - c_k & (2) \\ a_k &= 2\sum_{i=1}^{n} x_{ik}^2 = 2||\mathbf{x}_{:,k}||^2 & (3) \\ c_k &= 2\sum_{i=1}^{n} x_{ik}(y_i - \mathbf{w}_{-k}^T \mathbf{x}_{i,-k}) = 2\mathbf{x}_{:,k}^T \mathbf{r}_k & (4) \end{aligned}$$

   where $\mathbf{w}_{-k} = \mathbf{w}$ without component k, $\mathbf{x}_{i,-k}$ is $\mathbf{x}_i$ without component k, and $\mathbf{r}_k = \mathbf{y} - \mathbf{w}_{-k}^T \mathbf{x}_{:,-k}$ is the residual due to using all the features except feature k. Hint: Partition the weights into those involving k and those not involving k.

2. Show that if $\frac{\partial}{\partial w_k} RSS(\mathbf{w}) = 0$, then

$$\hat{w}_k = \frac{\mathbf{x}_{:,k}^T \mathbf{r}_k}{||\mathbf{x}_{:,k}||^2} \tag{5}$$

   Hence when we sequentially add features, the optimal weight for feature k is computed by computing orthogonally projecting $\mathbf{x}_{:,k}$ onto the current residual.