The goal of this problem is to show that there are two equivalent expressions for the residual sum of squares in linear regression. Then we will compute the analytical solution of the linear regression problem and prove that the solution is corresponds to a global minimum.

Let our training data be $\{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_N}, y_N)\}$, where the vector $\mathbf{x_n} \in \mathbb{R}^D$ includes the D features and $y_n \in \mathbb{R}$ is the label of sample n.

The training data can be also written in a matrix/vector notation as:

$$\mathbf{X} = \begin{bmatrix} 1 & \mathbf{x_1}^T \\ & \vdots \\ 1 & \mathbf{x_N}^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1D} \\ & & \vdots & & \\ 1 & x_{N1} & x_{N2} & \ldots & x_{ND} \end{bmatrix} \in \mathbb{R}^{N\times(D+1)} \text{ and } \mathbf{y} = [y_1, \ldots, y_N]^T \in \mathbb{R}^N$$

where x_{nd} is the d^{th} feature of sample n.

Let's also assume that the weight of the linear regression model is written as $\mathbf{w} = [w_0,\ w_1,\ \ldots,\ w_D]$, where w_0 is the bias.

We will show that the following expressions of the RSS error are equivalent.

$$RSS(\mathbf{w}) = \sum_{n=1}^{N} \left[y_n - \left(w_0 + \sum_{d=1}^{D} w_d x_{nd} \right) \right]^2$$

$$RSS(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$$

$$\begin{aligned}
\mathbf{y} - \mathbf{X}\mathbf{w} &= \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\in\mathbb{R}^{N\times 1}} - \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1D} \\ 1 & x_{21} & x_{22} & \ldots & x_{2D} \\ & & \vdots & & \\ 1 & x_{N1} & x_{N2} & \ldots & x_{ND} \end{bmatrix}}_{\in\mathbb{R}^{N\times(D+1)}} \cdot \underbrace{\begin{bmatrix} w0 \\ w1 \\ w2 \\ \vdots \\ w_D \end{bmatrix}}_{\in\mathbb{R}^{(D+1)\times 1}} \\
&= \begin{bmatrix} y_1 - (w_0 + w_1 x_{11} + w_2 x_{12} + \ldots + w_D x_{1D}) \\ y_2 - (w_0 + w_1 x_{21} + w_2 x_{22} + \ldots + w_D x_{2D}) \\ \vdots \\ y_N - (w_0 + w_1 x_{N1} + w_2 x_{N2} + \ldots + w_D x_{ND}) \end{bmatrix} \\
&= \begin{bmatrix} y_1 - (w_0 + \sum_{d=1}^{D} w_d x_{1d}) \\ y_2 - (w_0 + \sum_{d=1}^{D} w_d x_{2d}) \\ \vdots \\ y_N - (w_0 + \sum_{d=1}^{D} w_d x_{Nd}) \end{bmatrix} \in \mathbb{R}^N
\end{aligned}$$

For any vector $\mathbf{x} = [x_1, \ldots, x_N]^T \in \mathbb{R}^N$ (i.e., column vector), $\mathbf{x}^T = [x_1, \ldots, x_N] \in \mathbb{R}^{1\times N}$ (i.e., row vector), and we have that $\mathbf{x}^T\mathbf{x} = x_1^2 + \ldots + x_N^2 \in \mathbb{R}$, therefore:

$$
\begin{aligned}
RSS(\mathbf{w}) &= (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) \\
&= \left(y_1 - \left(w_0 + \sum_{d=1}^{D} w_d x_{1d}\right)\right)^2 + \ldots + \left(y_N - \left(w_0 + \sum_{d=1}^{D} w_d x_{Nd}\right)\right)^2 \\
&= \sum_{n=1}^{N} \left(y_n - \left(w_0 + \sum_{d=1}^{D} w_d x_{nd}\right)\right)^2
\end{aligned}
$$

We will minimize the RSS error by setting its first derivative $\frac{\vartheta RSS(\mathbf{w})}{\vartheta \mathbf{w}}$ to 0.

We will first expand the vector/matrix expression of $RSS(\mathbf{w})$ and use the transpose of matrix product $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.

$$
\begin{aligned}
RSS(\mathbf{w}) &= (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T(\mathbf{X}\mathbf{w}) - (\mathbf{X}\mathbf{w})^T\mathbf{y} + (\mathbf{X}\mathbf{w})^T(\mathbf{X}\mathbf{w}) \\
&= \mathbf{y}^T\mathbf{y} - 2(\mathbf{X}\mathbf{w})^T\mathbf{y} + (\mathbf{X}\mathbf{w})^T(\mathbf{X}\mathbf{w}) \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T(\mathbf{X}^T\mathbf{y}) + \mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w}
\end{aligned}
$$

We then compute the first-order derivative $\frac{\vartheta RSS(\mathbf{w})}{\vartheta \mathbf{w}}$ of $RSS(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$.
We will take into account that $\frac{\theta(\boldsymbol{\alpha}^T\mathbf{x})}{\theta\mathbf{x}} = \frac{\theta(\mathbf{x}^T\boldsymbol{\alpha})}{\theta\mathbf{x}} = \boldsymbol{\alpha}$ and $\frac{\theta(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\theta\mathbf{x}} = 2\mathbf{A}\mathbf{x}$, therefore we have:

$$
\frac{\vartheta RSS(\mathbf{w})}{\vartheta \mathbf{w}} = -2(\mathbf{X}^T\mathbf{y}) + 2(\mathbf{X}^T\mathbf{X})\mathbf{w}
$$

We finally find the minimum of $RSS(\mathbf{w})$ by solving the equation:

$$
\begin{aligned}
&\frac{\vartheta RSS(\mathbf{w})}{\vartheta \mathbf{w}} = 0 \\
\Rightarrow & -2(\mathbf{X}^T\mathbf{y}) + 2(\mathbf{X}^T\mathbf{X})\mathbf{w} = 0 \\
\Rightarrow & (\mathbf{X}^T\mathbf{X})\mathbf{w} = (\mathbf{X}^T\mathbf{y}) \\
\Rightarrow & \mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}
$$

The analytic solution of the linear regression problem is $\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \in \mathbb{R}^D$.

We will finally use the second derivative test to prove that the solution $\mathbf{w}^*$ is unique.

We will show that $RSS(\mathbf{w})$ is a convex function by proving the the Hessian matrix of $RSS(\mathbf{w})$ is positive semi-definite.
The Hessian matrix of $RSS(\mathbf{w})$ is defined as:

$$
\begin{aligned}
\mathbf{H}_{RSS(\mathbf{w})} &= \frac{\theta^2 RSS(\mathbf{w})}{\theta\mathbf{w}^2} = \\
&= \frac{\theta}{\theta\mathbf{w}}\left(\frac{\theta RSS(\mathbf{w})}{\theta\mathbf{w}}\right) \\
&= \frac{\theta}{\theta\mathbf{w}}\left(-2(\mathbf{X}^T\mathbf{y}) + 2(\mathbf{X}^T\mathbf{X})\mathbf{w}\right) \quad = 2(\mathbf{X}^T\mathbf{X})
\end{aligned}
$$

For every $\mathbf{u} \in \mathbb{R}^D$ we have (by applying the transpose product rule and the definition of $l2$-norm):

$$
\mathbf{u}^T\mathbf{H}_{RSS(\mathbf{w})}\mathbf{u} = 2\mathbf{u}^T(\mathbf{X}^T\mathbf{X})\mathbf{u} = 2\mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u} = 2(\mathbf{X}\mathbf{u})^T\mathbf{X}\mathbf{u} = 2\|\mathbf{X}\mathbf{u}\|_2^2 \geq 0
$$

Therefore the Hessian $\mathbf{H}_{RSS(\mathbf{w})}$ of the RSS error is positive semi-definite, thus $RSS(\mathbf{w})$ is convex and any local optima is a global minimum. Therefore the solution $\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \in \mathbb{R}^D$ is a global minimum of the RSS error of the linear regression problem.