The CrudeOil Example

Consider the following Crude Oil dataset. The independent variables are x_1: pressure (pounds per square inch) of injected steam and x_2: angle of steam injection (degrees). The dependent variable is y: percentage of oil recovery.

Crude Oil Dataset

x_1	x_2	y
1000	0	60.58
1000	15	72.72
1000	30	79.99
2000	0	66.83
2000	15	80.78
2000	30	89.78
3000	0	69.68
3000	15	80.31
3000	30	91.99

Notice that the percentage of oil recovery increases with the injected steam pressure and also with the steam injection angle.

Formulate a multiple regression model with dependent variable y and predictors x_1 and x_2. Then express this model as a matrix regression model and solve for the estimated regression parameters with

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Solution: Use this linear regression model with two predictors

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \tag{1}$$

then substitute the dataset values into (1) to obtain

$$\begin{aligned}
60.58 &= \beta_0 + \beta_1 1000 + \beta_2 0 + \epsilon_1 \\
72.72 &= \beta_0 + \beta_1 1000 + \beta_2 15 + \epsilon_2 \\
79.99 &= \beta_0 + \beta_1 1000 + \beta_2 30 + \epsilon_3 \\
66.83 &= \beta_0 + \beta_1 2000 + \beta_2 0 + \epsilon_4 \\
80.78 &= \beta_0 + \beta_1 2000 + \beta_2 15 + \epsilon_5 \\
89.78 &= \beta_0 + \beta_1 2000 + \beta_2 30 + \epsilon_6 \\
69.68 &= \beta_0 + \beta_1 3000 + \beta_2 0 + \epsilon_7 \\
80.31 &= \beta_0 + \beta_1 3000 + \beta_2 15 + \epsilon_8 \\
91.99 &= \beta_0 + \beta_1 3000 + \beta_2 30 + \epsilon_9
\end{aligned} \tag{2}$$

Now formulate (2) as the matrix regression model $Y = X\beta + \epsilon$:

$$\begin{pmatrix} 60.58 \\ 72.72 \\ 79.99 \\ 66.83 \\ 80.78 \\ 89.78 \\ 69.68 \\ 80.31 \\ 91.99 \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} 1000 \\ 1000 \\ 1000 \\ 2000 \\ 2000 \\ 2000 \\ 3000 \\ 3000 \\ 3000 \end{pmatrix} + \beta_2 \begin{pmatrix} 0 \\ 15 \\ 30 \\ 0 \\ 15 \\ 30 \\ 0 \\ 15 \\ 30 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \end{pmatrix}$$

$$\begin{pmatrix} 60.58 \\ 72.72 \\ 79.99 \\ 66.83 \\ 80.78 \\ 89.78 \\ 69.68 \\ 80.31 \\ 91.99 \end{pmatrix} = \begin{pmatrix} 1 & 1000 & 0 \\ 1 & 1000 & 15 \\ 1 & 1000 & 30 \\ 1 & 2000 & 0 \\ 1 & 2000 & 15 \\ 1 & 2000 & 30 \\ 1 & 3000 & 0 \\ 1 & 3000 & 15 \\ 1 & 3000 & 30 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \end{pmatrix},$$

$$\text{where } Y = \begin{pmatrix} 60.58 \\ 72.72 \\ 79.99 \\ 66.83 \\ 80.78 \\ 89.78 \\ 69.68 \\ 80.31 \\ 91.99 \end{pmatrix}, \; X = \begin{pmatrix} 1 & 1000 & 0 \\ 1 & 1000 & 15 \\ 1 & 1000 & 30 \\ 1 & 2000 & 0 \\ 1 & 2000 & 15 \\ 1 & 2000 & 30 \\ 1 & 3000 & 0 \\ 1 & 3000 & 15 \\ 1 & 3000 & 30 \end{pmatrix}, \text{ and } \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Finally find the estimated regression parameters β_0, β_1, and β_2. (These are the least squared estimators.)

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= \left[\begin{pmatrix} 1 & 1000 & 0 \\ 1 & 1000 & 15 \\ 1 & 1000 & 30 \\ 1 & 2000 & 0 \\ 1 & 2000 & 15 \\ 1 & 2000 & 30 \\ 1 & 3000 & 0 \\ 1 & 3000 & 15 \\ 1 & 3000 & 30 \end{pmatrix} \begin{pmatrix} 1 & 1000 & 0 \\ 1 & 1000 & 15 \\ 1 & 1000 & 30 \\ 1 & 2000 & 0 \\ 1 & 2000 & 15 \\ 1 & 2000 & 30 \\ 1 & 3000 & 0 \\ 1 & 3000 & 15 \\ 1 & 3000 & 30 \end{pmatrix}^T \right]^{-1} \begin{pmatrix} 1 & 1000 & 0 \\ 1 & 1000 & 15 \\ 1 & 1000 & 30 \\ 1 & 2000 & 0 \\ 1 & 2000 & 15 \\ 1 & 2000 & 30 \\ 1 & 3000 & 0 \\ 1 & 3000 & 15 \\ 1 & 3000 & 30 \end{pmatrix}^T \begin{pmatrix} 60.58 \\ 72.72 \\ 79.99 \\ 66.83 \\ 80.78 \\ 89.78 \\ 69.68 \\ 80.31 \\ 91.99 \end{pmatrix} \\ &= \begin{pmatrix} 9 & 18,000 & 135 \\ 18,000 & 42,000,000 & 270,000 \\ 135 & 270,000 & 3,375 \end{pmatrix}^{-1} \begin{pmatrix} 692.66 \\ 1,414,010.00 \\ 11,359.95 \end{pmatrix} = \begin{pmatrix} 56.620556 \\ 0.004782 \\ 0.718556 \end{pmatrix}, \end{aligned}$$

This means that, $\hat{\beta}_0 = 56.620556$, $\hat{\beta}_1 = 0.004782$, and $\hat{\beta}_2 = 0.718556$.

Verify these estimates with SAS and R. See the CrudeOil Example for details.