Assignment 1

Problem 1: (10 points)
The two variable regression model $y = \alpha + \beta x + \varepsilon$.

1. Show that the least squares normal equations imply $\sum_i e_i = 0$ and $\sum_i x_i e_i = 0$.
2. Show that the solution for the constant term is $a = \bar{y} - b\bar{x}$.
3. Show that the solution for b is $b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Solution:

$$\begin{aligned}
\min \sum_i e_i^2 &= \min \sum_i (y_i - a - bx_i) \\
\frac{\partial \sum_i e_i^2}{\partial a} &= -2\sum_i (y_i - a - bx_i) = -2\sum_i e_i = 0 \Rightarrow \sum_i e_i = 0 \\
\frac{\partial \sum_i e_i^2}{\partial b} &= -2\sum_i (y_i - a - bx_i)x_i = -2\sum_i e_i x_i = 0 \Rightarrow \sum_i e_i x_i = 0
\end{aligned}$$

$$\begin{aligned}
\sum_i e_i &= 0 \\
\sum_i (y_i - a - bx_i) &= 0 \\
\sum_i y_i &= \sum_i a + \sum_i bx_i \\
\sum_i y_i &= na + \sum_i bx_i \\
\frac{1}{n}\sum_i y_i &= a + b\frac{1}{n}\sum_i x_i \\
\bar{y} &= a + b\bar{x}
\end{aligned}$$

$$\sum_i e_i x_i - \bar{x}\sum_i e_i = 0$$

$$\begin{aligned}
\sum_i (x_i - \bar{x}) e_i &= 0 \\
\sum_i (x_i - \bar{x})(y_i - a - bx_i) &= 0 \\
\sum_i (x_i - \bar{x})(y_i - \bar{y} + b\bar{x} - bx_i) &= 0 \\
\sum_i (x_i - \bar{x})(y_i - \bar{y} - b(x_i - \bar{x})) &= 0 \\
\sum_i (x_i - \bar{x})(y_i - \bar{y}) &= b \sum_i (x_i - \bar{x})(x_i - \bar{x}) \\
b &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}
\end{aligned}$$

Problem 3: (10 points)
A common strategy for handling a case in which an observation is missing data for one or more variables is to fill those missing variables with 0s and add a variable to the model that takes the value 1 for that one observation and 0 for all other observations. Show that this strategy is equivalent to discarding the observation as regards the computation of **b** but is does have an effect on R^2. Consider the special case in which X contains only a constant and one variable.

Solution:
The data matrix has the following design:

$$\begin{aligned}
X &= \begin{pmatrix} 1 & x & 0 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} X_1 & \begin{matrix} 0 \\ 1 \end{matrix} \end{pmatrix} = (X_1, X_2) \\
Y &= \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}
\end{aligned}$$

Now using Frisch-Waugh-Lovell Theorem:

$$\begin{aligned}
b_1 &= (X_1' M_2 X_1)^{-1} (X_1' M_2 Y) \\
M_2 &= I - X_2 (X_2' X_2)^{-1} X_2'
\end{aligned}$$

$$
\begin{aligned}
M_2 &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \left[\begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right]^{-1} \begin{pmatrix} 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}
\end{aligned}
$$

This matrix drops the last observation. Consequently b_1 is calculated without the last observation.

$$
R^2 = \frac{\left(\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum_i (y_i - \bar{y})\right)^2 \left(\sum_i (\hat{y}_i - \bar{\hat{y}})\right)^2}
$$

So R^2 is a function of $\bar{y}$. If we add an observation the mean will change (in general) and thereby changes the value of R^2.