Answers to the questions

Yijin Zeng

1 Question A

1.1 Question a

For this question, let us assume ${x_1}^T x_1 > 0$, and ${x_2}^T x_2 > 0$, otherwise β_1 and β_2 are not well defined.

According to the definition of OLS regression, we have

$$\beta_1 = \underset{\beta}{\operatorname{argmax}}\ (y_1 - \beta x_1)^T (y_1 - \beta x_1) \tag{1}$$

Denote

$$L(\beta) = (y_1 - \beta x_1)^T (y_1 - \beta x_1),$$
$$\Rightarrow \frac{dL(\beta)}{d\beta} = -{x_1}^T (y_1 - \beta x_1).$$

Set $\frac{dL(\beta)}{d\beta} = 0$, we have

$$\beta = \frac{{x_1}^T y_1}{{x_1}^T x_1},$$

Notice that

$$\frac{d^2 L(\beta)}{d\beta^2} = {x_1}^T x_1 > 0.$$

Hence, according to equation (1), we have

$$\beta_1 = \frac{{x_1}^T y_1}{{x_1}^T x_1}.$$

Similarly, we can show that

$$\beta_2 = \frac{{x_2}^T y_2}{{x_2}^T x_2}.$$

Let us define $x = ({x_1}^T\ {x_2}^T)^T$ and $y = ({y_1}^T\ {y_2}^T)^T$. Then, following the similar strategy as deriving β_1, we obtain that

$$\beta = \frac{x^T y}{x^T x} = \frac{{x_1}^T y_1 + {x_2}^T y_2}{{x_1}^T x_1 + {x_2}^T x_2}.$$

Notice that

$$\beta_1 {x_1}^T x_1 = {x_1}^T y_1, \text{ and } \beta_2 {x_2}^T x_2 = {x_2}^T y_2.$$

Hence,

$$\begin{aligned}
\beta &= \frac{\beta_1 {x_1}^T x_1 + \beta_2 {x_2}^T x_2}{{x_1}^T x_1 + {x_2}^T x_2} \\
&= \beta_1 \frac{{x_1}^T x_1}{{x_1}^T x_1 + {x_2}^T x_2} + \beta_2 \frac{{x_2}^T x_2}{{x_1}^T x_1 + {x_2}^T x_2} \\
&= \beta_1 \frac{{x_1}^T x_1}{{x_1}^T x_1 + {x_2}^T x_2} + \beta_2 \left(1 - \frac{{x_1}^T x_1}{{x_1}^T x_1 + {x_2}^T x_2}\right) \qquad (2) \\
&= (\beta_1 - \beta_2) \frac{{x_1}^T x_1}{{x_1}^T x_1 + {x_2}^T x_2} + \beta_2
\end{aligned}$$

Therefore, we can see that for any given β_1 and β_2, β is a function of $\frac{{x_1}^T x_1}{{x_1}^T x_1 + {x_2}^T x_2} \in (0, 1)$. Hence, we have

$$\min \beta = \begin{cases} \beta_2 & \text{if } \beta_1 \geq \beta_2, \\ \beta_1 & \text{otherwise,} \end{cases}, \quad \max \beta = \begin{cases} \beta_1 & \text{if } \beta_1 \geq \beta_2, \\ \beta_2 & \text{otherwise.} \end{cases}$$

Thus, we conclude that

$$\min\{\beta_1, \beta_2\} \leq \beta \leq \max\{\beta_1, \beta_2\}.$$

1.2 Question b

To start, let us denote

$$x_1 = \begin{pmatrix} x_{1,1} \\ \vdots \\ x_{1,m_1} \end{pmatrix}, \ y_1 = \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,m_1} \end{pmatrix}, \ x_2 = \begin{pmatrix} x_{2,1} \\ \vdots \\ x_{2,n_1} \end{pmatrix}, \ y_2 = \begin{pmatrix} y_{2,1} \\ \vdots \\ y_{2,n_1} \end{pmatrix},$$

Then, we define

$$\bar{x}_1 = \frac{1}{m_1} \sum_{i=1}^{m_1} x_{1,i}, \ \bar{x}_2 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{2,i}, \ \bar{y}_1 = \frac{1}{m_1} \sum_{i=1}^{m_1} y_{1,i}, \ \bar{y}_2 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{2,i},$$

and

$$\bar{x} = \frac{1}{m_1 + n_1} \left(\sum_{i=1}^{m_1} x_{1,i} + \sum_{i=1}^{n_1} x_{2,i} \right), \ \bar{y} = \frac{1}{m_1 + n_1} \left(\sum_{i=1}^{m_1} y_{1,i} + \sum_{i=1}^{n_1} y_{2,i} \right).$$

If the interception terms are added for all three models, then using standard results of least square linear regression, we have

$$\beta_1 = \frac{\sum_{i=1}^{m_1} (x_{1,i} - \bar{x}_1)(y_{1,i} - \bar{y}_1)}{\sum_{i=1}^{m_1} (x_{1,i} - \bar{x}_1)^2}, \ \beta_2 = \frac{\sum_{i=1}^{n_1} (x_{2,i} - \bar{x}_2)(y_{2,i} - \bar{y}_2)}{\sum_{i=1}^{n_1} (x_{2,i} - \bar{x}_2)^2},$$

and

$$\beta = \frac{\sum_{i=1}^{m_1}(x_{1,i} - \bar{x})(y_{1,i} - \bar{y}) + \sum_{i=1}^{n_1}(x_{2,i} - \bar{x})(y_{2,i} - \bar{y})}{\sum_{i=1}^{m_1}(x_{1,i} - \bar{x})^2 + \sum_{i=1}^{n_1}(x_{2,i} - \bar{x})^2}$$

Now, in order to consider how, when β_1 and β_2 are fixed, the parameter β can change, let us consider a new data set $y_3 \in \mathbb{R}^{n_1}$ of the same size as y_2, such that for any $i \in \{1, \ldots, n_1\}$,

$$y_{3,i} = y_{2,i} + s_y,$$

i.e. y_3 is a shift of y_2. Since the new data set is only a shift of y_2, we have

$$\bar{y}_3 = \bar{y}_2 + s_y.$$

Importantly, the linear regression of y_3 on x_2 will have the same slope of the regression of y_2 on x_2, i.e.

$$\begin{aligned}
\beta_3 &= \frac{\sum_{i=1}^{n_1}(x_{2,i} - \bar{x}_2)(y_{3,i} - \bar{y}_3)}{\sum_{i=1}^{n_1}(x_{2,i} - \bar{x}_2)^2} \\
&= \frac{\sum_{i=1}^{n_1}(x_{2,i} - \bar{x}_2)(y_{2,i} + s_y - \bar{y}_2 - s_y)}{\sum_{i=1}^{n_1}(x_{2,i} - \bar{x}_2)^2} \\
&= \frac{\sum_{i=1}^{n_1}(x_{2,i} - \bar{x}_2)(y_{2,i} - \bar{y}_2)}{\sum_{i=1}^{n_1}(x_{2,i} - \bar{x}_2)^2} \\
&= \beta_2.
\end{aligned}$$

Denote

$$\bar{y}' = \frac{1}{m_1 + n_1}\left(\sum_{i=1}^{m_1} y_{1,i} + \sum_{i=1}^{n_1} y_{3,i}\right).$$

We have

$$\bar{y}' = \bar{y} + \frac{n_1}{n_1 + m_1} s_y.$$

For notation ease, let us denote

$$S_y = \frac{n_1}{n_1 + m_1} s_y.$$

Subsequently, if we fit a linear regression of y' onto x, where y' is a concatenation of y_1 and y_3, we have the parameter associated with x' equals to

$$\begin{aligned}
\beta' &= \frac{\sum_{i=1}^{m_1}(x_{1,i} - \bar{x})(y_{1,i} - \bar{y}') + \sum_{i=1}^{n_1}(x_{2,i} - \bar{x})(y_{3,i} - \bar{y}')}{\sum_{i=1}^{m_1}(x_{1,i} - \bar{x})^2 + \sum_{i=1}^{n_1}(x_{2,i} - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{m_1}(x_{1,i} - \bar{x})(y_{1,i} - \bar{y} - S_y) + \sum_{i=1}^{n_1}(x_{2,i} - \bar{x})(y_{3,i} - \bar{y} - S_y)}{\sum_{i=1}^{m_1}(x_{1,i} - \bar{x})^2 + \sum_{i=1}^{n_1}(x_{2,i} - \bar{x})^2}
\end{aligned}$$

Notice that the numerator of β' is such that

$$
\begin{aligned}
&\sum_{i=1}^{m_1}(x_{1,i}-\bar{x})(y_{1,i}-\bar{y}-S_y)+\sum_{i=1}^{n_1}(x_{2,i}-\bar{x})(y_{3,i}-\bar{y}-S_y)\\
&=\sum_{i=1}^{m_1}(x_{1,i}-\bar{x})(y_{1,i}-\bar{y})-S_y\sum_{i=1}^{m_1}(x_{1,i}-\bar{x})+\sum_{i=1}^{n_1}(x_{2,i}-\bar{x})(y_{3,i}-\bar{y})-S_y\sum_{i=1}^{n_1}(x_{2,i}-\bar{x})\\
&=\sum_{i=1}^{m_1}(x_{1,i}-\bar{x})(y_{1,i}-\bar{y})+\sum_{i=1}^{n_1}(x_{2,i}-\bar{x})(y_{3,i}-\bar{y})\\
&=\sum_{i=1}^{m_1}(x_{1,i}-\bar{x})(y_{1,i}-\bar{y})+\sum_{i=1}^{n_1}(x_{2,i}-\bar{x})(y_{2,i}+s_y-\bar{y})\\
&=\sum_{i=1}^{m_1}(x_{1,i}-\bar{x})(y_{1,i}-\bar{y})+\sum_{i=1}^{n_1}(x_{2,i}-\bar{x})(y_{2,i}-\bar{y})+s_y\sum_{i=1}^{n_1}(x_{2,i}-\bar{x}).
\end{aligned}
$$

Hence

$$\beta'=\beta+\frac{s_y\sum_{i=1}^{n_1}(x_{2,i}-\bar{x})}{\sum_{i=1}^{m_1}(x_{1,i}-\bar{x})^2+\sum_{i=1}^{n_1}(x_{2,i}-\bar{x})^2}.$$

Since s_y can be any real number, when $\sum_{i=1}^{n_1}(x_{2,i}-\bar{x})\neq 0$, we have

$$\beta'\in(-\infty,\infty).$$

If, however, $\sum_{i=1}^{n_1}(x_{2,i}-\bar{x})=0$, then shifting y_2 alone will not change the value of coefficient associated with x_2. In this case, one could still show that the coefficient of β' is unbounded by consider shifting both x_2 and y_2, while keeping the coefficient of shifted x_2 and y_2 unchanged.

Overall, since x_2, y_2 can be any data set such that the coefficient associated with x_2 equals to β_2, we have shown that if the interception terms are added for all three models, then β is unbounded, i.e. $\beta\in(-\infty,\infty)$.

1.3 Question c

In Question a, equation (2) shows that

$$\beta=\beta_1\frac{{x_1}^T x_1}{{x_1}^T x_1+{x_2}^T x_2}+\beta_2\left(1-\frac{{x_1}^T x_1}{{x_1}^T x_1+{x_2}^T x_2}\right),$$

i.e. β is a weighted average of β_1 and β_2, where the weights are determined by the variance of x_1 and x_2 given they are both drawn from i.i.d zero-mean normal distribution.

We are given that all pairs are drawn i.i.d. from a zero-mean 2D multivariate Gaussian distribution. Without loss of generality, let us denote that

$$\begin{bmatrix}X\\Y\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\sigma_x^2&\sigma_{xy}\\\sigma_{xy}&\sigma_y^2\end{bmatrix}\right).$$

Then one reasonable assumption is that:

$$x_1{}^T x_1 \approx m_1 \sigma_x^2,$$
$$x_2{}^T x_2 \approx n_1 \sigma_x^2.$$

This is reasonable because $x_1{}^T x_1$ and $x_2{}^T x_2$ are the maximum likelihood estimators of σ_x^2 given x_1 and x_2, separately. The estimators are unbiased and consistent, meaning they are more accurate, in probability, given larger sample sizes n_1 and m_1.

Thus, our guess for β is

$$\beta \approx \beta_1 \frac{m_1}{n_1 + m_1} + \beta_2 \frac{n_1}{n_1 + m_1}.$$

2 Question B

2.1 Question a

To start, let us consider the linear regression model:

$$Y = X_1\beta + \epsilon.$$

According to standard results of linear regression, we have

$$\hat{Y} = X_1(X_1^T X_1)^{-1} X_1^T Y.$$

Let us denote $P = X_1(X_1^T X_1)^{-1} X_1^T$ as the projection matrix of X_1. Then using the results of QR decomposition, we have

$$\begin{aligned} P &= QR(R^T Q^T QR)^{-1} R^T Q^T \\ &= QR(R^T R)^{-1} R^T Q^T \\ &= QQ^T. \end{aligned}$$

Hence, we have

$$\hat{Y} = PY = QQ^T Y.$$

The sum of squared residuals of the model equals to

$$\begin{aligned} e^2 &= (Y - \hat{Y})^T (Y - \hat{Y}) \\ &= ((I - P)Y)^T ((I - P)Y) \\ &= Y^T (I - P)^T (I - P) Y \\ &= Y^T (I - P) Y \\ &= Y^T (I - QQ^T) Y \\ &= Y^T Y - Y^T QQ^T Y. \end{aligned} \tag{3}$$

After obtaining the above results, let us now consider a new model

$$Y = X_1\beta_{\text{new}} + \gamma x_{2,j} + \epsilon. \tag{4}$$

In other words, the new model is the regression of Y on $(X_1\ x_{2,j})$. Let us assume $x_{2,j}$ is not in the column space of X_1, otherwise the model does not have a unique solution.

Let us denote

$$X_2 = (X_1\ x_{2,j}).$$

Then, following the Gram–Schmidt process for QR decomposition, we have the QR decomposition of X_2 is:

$$X_2 = Q_2 R_2 = (Q\ q_j) \begin{pmatrix} R & r_{1,j} \\ 0 & r_{2,j} \end{pmatrix}. \tag{5}$$

To see this, we can perform the Gram–Schmidt process for X_2, where each column of Q is generated by the column of X_2 minus its projection on the column space generated by the previous column of X_2, and R is an upper matrix where each column is the coefficients to represent the same column of X_2 using the all columns so far of R as basis.

Subsequently, using the result of equation (3), we have the sum of squared residuals of the new model equals to

$$\begin{aligned} e^2_{\text{new}} &= Y^T Y - Y^T Q_2 Q_2^T Y \\ &= Y^T Y - Y^T Q Q^T Y - Y^T q_j q_j^T Y \\ &= e^2 - Y^T q_j q_j^T Y. \end{aligned}$$

Hence, the reduce error by introducing $x_{2,j}$ equals to

$$e^2 - e^2_{\text{new}} = Y^T q_j q_j^T Y.$$

Therefore, to choose $x_{2,j}$ reduces the most residual sum of squares, we want it to maximize the absolute value of its inner product between Y and q_j.

For each j, q_j can be found by performing the Gram–Schmidt process again,

$$u_j = x_{2,j} - QQ^T x_{2,j}, \tag{6}$$
$$q_j = u_j / ||u_j||. \tag{7}$$

Therefore, one efficient strategy to obtain the additional variable to minimize the residual sum of squares is for any $j \in \{1, \ldots, f-k\}$, to consider the orthogonal component of $x_{2,j}$ to X_1 following equation (6), and standardize this component using equation (7), and choose the j which maximize the absolute value of the inner product with Y. More specifically, we should choose x_{2,j^*} such that

$$j^* = \underset{j \in \{1,\ldots,f-k\}}{\text{argmax}} \frac{|(x_{2,j} - QQ^T x_{2,j})^T Y|}{||x_{2,j} - QQ^T x_{2,j}||}. \tag{8}$$

Let us analyze the computational complexity of this strategy by using equation (8). Since we know the QR composition of X_1, we assume Q is known. As Q is of shape $n \times k$, $Q^T x_{2,j}$ requires computational complexity $O(nk)$, and then $QQ^T x_{2,j}$ also requires computational complexity $O(nk)$. After obtaining $QQ^T x_{2,j}$, computing $|(x_{2,j} - QQ^T x_{2,j})^T Y|$ and $||x_{2,j} - QQ^T x_{2,j}||$ both require computational complexity $O(n)$. Hence, evaluating equation (8) for each single $j \in \{1, \ldots, f-k\}$ requires $O(nk)$. Notice that there are $f-k$ number of j's to evaluate in total. This strategy requires computational time of $O((f-k)(nk))$.

If, however, we naively re-fitting a new model when $x_{2,j}$ is added for each $j \in \{1, \ldots, f-k\}$, we will have overall computational complexity of $O((f-k)(k+1)^2 n)$ since fitting one linear regression typically requires $O((k+1)^2 n)$. Therefore, our strategy using equation (8) is more efficient than naively refitting a new model every time, and it will be particularly useful when k is large.

2.2 Question b

Suppose we have chosen a variable $x_{2,j}$ according to (8), and we are interested in finding the coefficients in (4), i.e. β_{new}, and γ below:

$$Y = X_1\beta_{\text{new}} + \gamma x_{2,j} + \epsilon. \tag{9}$$

Denote $X_2 = (X_1\ x_{2,j})$, then according to standard results of linear regression, we know that

$$\begin{aligned} \begin{pmatrix} \beta_{\text{new}} \\ \gamma \end{pmatrix} &= (X_2^T X_2)^{-1} X_2^T Y \\ &= (R_2^T Q_2^T Q_2 R_2)^{-1} R_2^T Q_2^T Y \\ &= R_2^{-1} Q_2^T Y. \end{aligned}$$

Hence, we have

$$R_2 \begin{pmatrix} \beta_{\text{new}} \\ \gamma \end{pmatrix} = Q_2^T Y.$$

Using equation (5), we further obtain that:

$$\begin{aligned} &\begin{pmatrix} R & r_{1,j} \\ 0 & r_{2,j} \end{pmatrix} \begin{pmatrix} \beta_{\text{new}} \\ \gamma \end{pmatrix} = \begin{pmatrix} Q^T \\ q_{2,j}^T \end{pmatrix} Y \\ \Rightarrow &\gamma = (q_{2,j}^T Y)/(r_{2,j}),\ \beta_{\text{new}} = R^{-1}Q^T Y - R^{-1} r_{1,j}\gamma. \end{aligned} \tag{10}$$

In order to solve (10), we need to know the values of $r_{2,j}$ and $r_{1,j}$. Notice that from equation (5), we have

$$\begin{aligned} &x_{2,j} = Q r_{1,j} + q_j r_{2,j} \\ \Rightarrow &Q^T x_{2,j} = Q^T Q r_{1,j} + Q^T q_j r_{2,j}. \end{aligned}$$

Since Q is an orthonormal matrix and q_j is orthogonal to Q, we have

$$r_{1,j} = Q^T x_{2,j}, \tag{11}$$

which is already computed in Question (a).

Similarly, to compute $r_{2,j}$, we start from equation (5),

$$\begin{aligned} &x_{2,j} = Q r_{1,j} + q_j r_{2,j}, \\ \Rightarrow &q_j^T x_{2,j} = q_j^T Q r_{1,j} + q_j^T q_j r_{2,j}, \\ \Rightarrow &q_j^T x_{2,j} = r_{2,j}. \end{aligned} \tag{12}$$

Put equation (12) back to equation (10), we have γ equals to:

$$\gamma = (q_{2,j}^T Y)/(q_j^T x_{2,j}). \tag{13}$$

To solve β_{new}, we notice that

$$R^{-1}Q^TY = \beta,$$

and $R^{-1}r_{1,j}$ can be computed by solving the v below:

$$Rv = r_{1,j}, \tag{14}$$

which requires $O(k^2)$ since R is an upper triangle matrix. Hence we have

$$\beta_{\text{new}} = \beta - v\gamma, \tag{15}$$

where γ v are defined in (13), and (14), respectively.

Overall, we can compute the coefficients $(\beta_{\text{new}}, \gamma)$ of the new model defined in (9) using equations (13) and (15). The computational complexity of solving the two equations is $O(n + k^2)$ provided $Q^Tx_{2,j}$ has already been computed in Question (a).