A linear regression model $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$ has a matrix-vector representation $\mathbf{y} = \mathbf{X}^T\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is a $p \times N$ matrix whose *i-th* column is $\mathbf{x}_i$. This representation will be used frequently.

In the least square regression, we estimate $\boldsymbol{\beta}$ as the minimizer of $\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$ and in matrix notation, the objective function becomes $(\mathbf{y} - \mathbf{X}^T\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}^T\boldsymbol{\beta})$ (they are the same as $\mathbf{a}^T\mathbf{a} = \sum_i a_i^2$).

Expanding this expression we have, $(\mathbf{y} - \mathbf{X}^T\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}^T\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}\mathbf{X}^T\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}\mathbf{y}$. Now to minimize, we take the derivative of this objective function with respect to $\boldsymbol{\beta}$ which gives $2(\mathbf{X}\mathbf{X}^T\boldsymbol{\beta} - \mathbf{X}\mathbf{y})$ and equate it to zero.

Thus, we need to solve $\mathbf{X}\mathbf{X}^T\boldsymbol{\beta} = \mathbf{X}\mathbf{y}$ which gets us $\hat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y}$. From matrix algebra, a useful result is that $\mathbf{A}\mathbf{B}^T = \sum_{i=1}^{N}\mathbf{a}_i\mathbf{b}_i^T$, where $\mathbf{A}$ is a matrix with N columns with $\mathbf{a}_i$ is the entry on the *i-th* columns and $\mathbf{B}$ is a matrix with N rows whose *i-th* row is $\mathbf{b}_i$. Thus, $\mathbf{X}\mathbf{X}^T = \sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T, \mathbf{X}\mathbf{y} = \sum_{i=1}^{N} y_i\mathbf{x}_i$.

We know that $\mathrm{Var}(\mathbf{z}) = \boldsymbol{\Sigma}$, then $\mathrm{Var}(\mathbf{A}\mathbf{z}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$.

Then, $\mathrm{Var}(\hat{\boldsymbol{\beta}})= (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathrm{Var}(\mathbf{y})\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}$. We have $\mathrm{Var}(\mathbf{y}) = \sigma^2\mathbf{I}_N$, in case of **homoscedastic errors**.

Thus, $\mathrm{Var}(\hat{\boldsymbol{\beta}}) =$
$(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathrm{Var}(\mathbf{y})\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\sigma^2\mathbf{I}_N\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} = \sigma^2(\mathbf{X}\mathbf{X}^T)^{-1}(\mathbf{X}\mathbf{X}^T)(\mathbf{X}\mathbf{X}^T)^{-1} = \sigma^2(\mathbf{X}\mathbf{X}^T)^{-1}$.

If $\mathbf{z} \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have $\mathbf{A}\mathbf{z} \sim \mathrm{MVN}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$.

Thus, under normality assumption, $\mathbf{y} \sim \mathrm{MVN}(\mathbf{X}^T\boldsymbol{\beta}_0, \sigma^2\mathbf{I})$ which leads to $\hat{\boldsymbol{\beta}} \sim \mathrm{MVN}(\boldsymbol{\beta}_0, \sigma^2(\mathbf{X}\mathbf{X}^T)^{-1})$.

Expression for estimation error: $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}^T\hat{\boldsymbol{\beta}} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y} = \{\mathbf{I}_N - \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\}\mathbf{y}$.

Let $\mathbf{H}_\mathbf{X} = \mathbf{I} - \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}$. There are many nice properties of $\mathbf{H}_\mathbf{X}$.

- $\mathbf{H}_\mathbf{X}^2 = \mathbf{H}_\mathbf{X}\mathbf{H}_\mathbf{X} = \mathbf{H}_\mathbf{X}$ (Verify!)
- $\mathbf{H}_\mathbf{X}\mathbf{X}^T = \mathbf{0}$.

The above two properties make $\mathbf{H}_\mathbf{X}$ the **Orthogonal projection matrix** with respect to $\mathbf{X}^T$.

Additional references They all convey the same message, but good to read as many references on this topic as possible. It is a super important concept and useful to solve various problems.

- `https://online.stat.psu.edu/stat462/node/132/`

- https://bookdown.org/josiesmith/qrmbook/introduction-to-multiple-regression.html
- https://www.stat.purdue.edu/~lingsong/teaching/2018spring/topic3.pdf
- In the current note, page 115 has a detailed matrix-vector representation for linear regression.