## 1 The Generalized Form of Least Squares Method

The purpose of this section is to show how we solve the OLS problem in vectors and matrices form. Firstly, we define

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{1}$$

Here, we use bold-face notation for representing vectors (lower case) and matrices (upper case), and non-bold-face notation for representing scalars.

Now, consider the goal of the least squares method, that is

$$\underset{\boldsymbol{w}}{\operatorname{argmin}} \boldsymbol{\epsilon}^{\boldsymbol{T}} \boldsymbol{\epsilon} \tag{2}$$

We know that $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}$, and we know that matrices multiplication has the property that $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$. Thus,

$$\begin{aligned} \boldsymbol{\epsilon}^{\boldsymbol{T}}\boldsymbol{\epsilon} &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T\boldsymbol{y} - (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T\boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{y}^T\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + (\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} \end{aligned} \tag{3}$$

Both the results of $\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^T\boldsymbol{y}$ and $\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{w}$ are $1 \times 1$ scalars, and we know that for a scalar $a$, its transpose is equal to itself ($a^T = a$). Based on these facts, we have $\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} = (\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y})^T = \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w}$. Thus,

$$\begin{aligned} \boldsymbol{\epsilon}^{\boldsymbol{T}}\boldsymbol{\epsilon} &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} \end{aligned} \tag{4}$$

To find the minimum value of $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$, let $\nabla\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = 0$. Therefore,

$$\begin{aligned} \nabla\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} &= \nabla(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}) \\ &= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{X}^T\boldsymbol{y} = 0 \end{aligned} \tag{5}$$

Then we have

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{6}$$

Notice that the $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exists only when $\boldsymbol{X}^T\boldsymbol{X}$ is a full rank matrix, that is the column vectors (or row vector because it is a square matrix here thus it does not matter) inside the

matrix are linear independent [1].

## 2 Hypothesis Testings on Coefficients

In this section, we are focusing on answering why the coefficients follow a t-distribution. From the previous section, we know that, if the following assumptions hold,

1. The linear relationship $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$ exists;
2. $\boldsymbol{X}$ is known and it is an invertible matrix with full rank, that is, there is no multi-collinearity among the features.
3. The noise (residual) is homoscedastic and uncorrelated, that is, $\text{Var}(\epsilon_1) = \text{Var}(\epsilon_2) = \cdots = \text{Var}(\epsilon_n) = \sigma^2$ and $\forall i \neq j, \text{Cov}(\epsilon_i, \epsilon_j) = 0$
4. The noise $\boldsymbol{\epsilon}$ follows gaussian distribution: $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\, \sigma^2 \boldsymbol{I})$

then we can fit a linear model $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{w}}$. $\hat{\boldsymbol{w}}$ is estimated by least square method, and the estimation is

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{w} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} \tag{7}$$

Now, the first thing to do is to figure out the distribution of $\hat{\boldsymbol{w}}$ (reflecting how it varies under repeated sampling). Since $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$. Also the linear combination of a gaussian distribution is also a gaussian distribution. Combining these two facts with conclusion (7), we know that $\hat{\boldsymbol{w}}$ also follows gaussian distribution. Furthermore, we have

$$\text{E}(\hat{\boldsymbol{w}}) = \text{E}(\boldsymbol{w}) + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{E}(\boldsymbol{\epsilon}) = \boldsymbol{w} \ ^{[2]} \tag{8}$$

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{w}}) &= \text{Var}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}) \\
&= \text{Var}(\boldsymbol{X}^{-1}(\boldsymbol{X}^T)^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}) \\
&= \text{Var}(\boldsymbol{X}^{-1}(\boldsymbol{X}^{-1})^T\boldsymbol{X}^T\boldsymbol{\epsilon}) \\
&= \text{Var}(\boldsymbol{X}^{-1}(\boldsymbol{X}\boldsymbol{X}^{-1})^T\boldsymbol{\epsilon}) \\
&= \text{Var}(\boldsymbol{X}^{-1}\boldsymbol{I}^T\boldsymbol{\epsilon}) \\
&= \text{Var}(\boldsymbol{X}^{-1}\boldsymbol{\epsilon}) \\
&= \boldsymbol{X}^{-1}\text{Var}(\boldsymbol{\epsilon})(\boldsymbol{X}^{-1})^T \\
&= \boldsymbol{X}^{-1}\sigma^2\boldsymbol{I}(\boldsymbol{X}^{-1})^T \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \quad ^{[3]}
\end{aligned} \tag{9}$$

[1] The is one of the reasons that why we need the features to meet the linear-independent condition.

[2] $\boldsymbol{w}$ is a constant vector, thus $\text{E}(\boldsymbol{w}) = \boldsymbol{w}$. $\text{E}(\boldsymbol{\epsilon}) = 0$ base on the fourth assumption.

[3] The derivative process is based on following facts: $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$; $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$; $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$; $\boldsymbol{AI} = \boldsymbol{A}$ where $\boldsymbol{I}$ is the identity matrix; $\text{Var}(aX + b) = a^2\text{Var}(X)$; $\text{Var}(\boldsymbol{Ax}) = \boldsymbol{A}\text{Var}(\boldsymbol{x})\boldsymbol{A}^T$.

Thus,

$$\hat{\boldsymbol{w}} \sim \mathcal{N}\Big(\boldsymbol{w},\, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\Big) \tag{10}$$

For the $j^{th}$ item $\hat{w}_j$ in $\hat{\boldsymbol{w}}$, its mean is $w_j$, and its variance is the $j^{th}$ item on the diagonal of matrix $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Thus,

$$\hat{w}_j \sim \mathcal{N}\Big(w_j,\, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}\Big) \tag{11}$$

Then, we have

$$\frac{\hat{w}_j - w_j}{\sqrt{\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}} \sim \mathcal{N}(0,\, 1) \tag{12}$$

**Lemma 2.1.** *if* $X \sim \mathcal{N}(0,\, 1)$*, and* $Y \sim \chi^2(n)$*, then* $t = \dfrac{X}{\sqrt{Y/n}} \sim t(n)$*, which is a t-distribution with* $n$ *degree of freedom.*

**Lemma 2.2.** *if* $Z_j \sim \mathcal{N}(0,\, 1)$*, then* $\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi_n^2$

Now, we are focusing on getting the unbiased estimation $\hat{\sigma}^2$ of $\sigma^2$ , and we are using the sample variance for estimating.

$$\begin{aligned} \hat{\sigma}^2 = S_\epsilon^2 &= \frac{1}{n-p-1}\sum_{j=1}^{n}(\epsilon_j - \bar{\epsilon})^2 \\ &= \frac{\boldsymbol{\epsilon^T\epsilon}}{n-p-1} \quad {}^{4} \end{aligned} \tag{13}$$

where $p$ is the number of predictors in the regression model. Thus,

$$\boldsymbol{\epsilon^T\epsilon} = (n-p-1)\hat{\sigma}^2 \tag{14}$$

We also know that $\epsilon_j \sim \mathcal{N}(0,\, \sigma^2)$. Therefore, $\dfrac{\epsilon_j}{\sigma} \sim \mathcal{N}(0,\, 1)$. Based on lemma 2.2,

$$\sum_{j=1}^{n}\left(\frac{\epsilon_j}{\sigma}\right)^2 = \frac{1}{\sigma^2}\boldsymbol{\epsilon^T\epsilon} \sim \chi^2_{n-p-1} \tag{15}$$

Then we know that

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1} \tag{16}$$

[4] Based on $\epsilon_j = y_j - \hat{y}_j$ and $E(\epsilon) = 0$.

Finally, based on conclusion (11), (12), conclusion (16), and lemma 2.1, we can prove that

$$\frac{\frac{\hat{w}_j - w_j}{\sqrt{\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}}}{\sqrt{\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2}\Big/(n-p-1)}} = \frac{\hat{w}_j - w_j}{\hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}} = \frac{\hat{w}_j - w_j}{\sqrt{\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}} \sim t_{n-p-1} \tag{17}$$

Since $\sqrt{\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}$ in (17) is the unbiased estimation of $\sqrt{\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}$, which is the standard error of $\hat{w_j}$. Notice that $\mathrm{SE}(\hat{w_j})$ is also the unbiased estimation of the standard error of $\hat{w_j}$. Therefore,

$$\frac{\hat{w}_j - w_j}{\mathrm{SE}(\hat{w_j})} \sim t_{n-p-1} \tag{18}$$

## 3 Multilinearity

In the previous section, we have already discussed why multilinearity will break our assumptions during solving OLS. In this section, we will use another perspective, model interpretation, for exploring why we need to eliminate multilinearity.

**Example 3.1.** Assume

$$\begin{aligned} X_1 &: \mathrm{TV} + 2 \times \mathrm{Newspaper} + \mathrm{Radio} \\ X_2 &: \mathrm{TV} \\ X_3 &: \mathrm{Radio} \\ X_4 &: \mathrm{Newspaper} \\ Y &: \mathrm{Sales} \end{aligned}$$

and we are building the regression model $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$. We will show that due to the existence of multilinearity, the advertisement effect from newspaper can have a positive influence on sales even if $\hat{\beta}_4 < 0$.

$$\begin{aligned} \hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 \\ &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + X_2 - X_2 + 2X_4 - 2X_4 + X_3 - X_3 \\ &= \hat{\beta}_0 + (\hat{\beta}_1 - 1)X_1 + (\hat{\beta}_2 - 1)X_2 + (\hat{\beta}_3 - 1)X_3 + (\hat{\beta}_4 - 2)X_4 \end{aligned} \tag{19}$$

Let $\hat{\beta}0' = \hat{\beta}0$, $\hat{\beta}1' = \hat{\beta}1 - 1$, $\hat{\beta}2' = \hat{\beta}2 - 1$, $\hat{\beta}3' = \hat{\beta}3 - 1$, $\hat{\beta}4' = \hat{\beta}4 - 2$. Thus, if the final model is

$$\hat{Y} = \hat{\beta}_0{}' + \hat{\beta}_1{}' X_1 + \hat{\beta}_2{}' X_2 + \hat{\beta}_3{}' X_3 + \hat{\beta}_4{}' X_4 \tag{20}$$

and $\hat{\beta_4}' < 0$, we cannot say the level of the advertisement effect from newspaper increases will make sales decrease based on $\hat{\beta_4}' < 0$ since $\hat{\beta_4}$ still have a chance to be a value that is greater than zero (e.g. $\hat{\beta_4} = 1$).

## 4 References

[1] J.W, Li, Regression Based on Least Square: The Matrices Prospective, May. 13, 2018. Accessed on: https://zhuanlan.zhihu.com/p/33899560
[2] J.W, Li, The Hypothesis Testing for Linear Regression: The T test and the F test, May. 14, 2018. Accessed on: https://zhuanlan.zhihu.com/p/36782834