STATISTICS(II) - REGRESSION ANALYSIS

1 Linear Regression

1.1 Simple Linear Regression

Example:

Student	X(cm)	Y(kg)
1	170	60
.....	.....	.....
10	181	69.2
11	180	?

If we only have data of Y, the best estimator will be $\widehat{Y} = \overline{Y} = \mathbb{E}Y$.

But if we have data of X, then $\mathbb{E}Y$ can be changed into $\mathbb{E}(Y|X)$.

Definition of Simple Linear Regression: (simple means X is 1-dimension)

In a Standard Model $Y = \alpha + \beta X + \epsilon$, Y represents dependent (response) variable, X represents independent (predictor) variable, and ϵ represents random error, such that $\mathbb{E}\epsilon = 0, \mathbb{V}\mathrm{ar}\epsilon = \sigma^2$.

Then $\mathbb{E}(Y|X) = \alpha + \beta X$, which is linear in X, and we defined it as $r(X)$, which we call regression function.

Data setting: (X_i, Y_i) i.i.d., $i = 1, 2, \cdots, n$, that is $Y_i = \alpha + \beta X_i + \epsilon_i$, where $\{\epsilon_i\}$ uncorrelated.

Hence, we have $\mathbb{E}\epsilon_i = 0, \mathbb{V}\mathrm{ar}\epsilon_i = \sigma^2$, and $\mathbb{E}Y_i = \alpha + \beta X_i, \mathbb{V}\mathrm{ar}Y_i = \sigma^2$.

So here comes the question:

Why do we use linear model to depict the relationship between X and Y?

For example, if X and Y follow a joint normal distribution, that is

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right),$$

where $\rho = \frac{\mathbb{C}\mathrm{ov}(X,Y)}{\sqrt{\mathbb{V}\mathrm{ar}X\mathbb{V}\mathrm{ar}Y}}$. Then we do a normalization, that is

$$\begin{cases} \dfrac{X - \mu_X}{\sigma_X} = Z_1 \\ \dfrac{Y - \mu_Y}{\sigma_Y} = \rho Z_1 + \sqrt{1 - \rho^2}\, Z_2 \end{cases},$$

where $Z_1 \perp Z_2 \sim N(0, 1)$

$$\begin{aligned}\mathbb{E}(Y|X) &= \mathbb{E}(\rho Z_1\sigma_Y + \sqrt{1-\rho^2}Z_2\sigma_Y + \mu_Y|X) \\ &= \mathbb{E}(\rho Z_1\sigma_Y + \sqrt{1-\rho^2}Z_2\sigma_Y + \mu_Y|Z_1) \\ &= \rho Z_1\sigma_Y + \mu_Y \\ &= \rho\sigma_Y\frac{X-\mu_X}{\sigma_X} + \mu_Y \\ &= aX + b\end{aligned}$$

Therefore, $\mathbb{E}(Y|X)$ is linear in X, and the question above gets an answer.

SSE (sum of squared errors / residual sum of squares)

$$SSE(\widehat{\alpha},\widehat{\beta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}X_i)^2$$

1.1.1 LSE (Least Square Estimator)

$$(\widehat{\alpha}_{LSE}, \widehat{\beta}_{LSE}) = argmin_{(\widehat{\alpha},\widehat{\beta})} SSE(\widehat{\alpha},\widehat{\beta})$$

We will use two methods to calculate the above equation.

Method 1

Take the partial derivatives of SSE on both $\widehat{\alpha}$ and $\widehat{\beta}$,

$$\begin{cases} \dfrac{\partial SSE}{\partial \widehat{\alpha}} = \displaystyle\sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}X_i) = 0 \\ \dfrac{\partial SSE}{\partial \widehat{\beta}} = \displaystyle\sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}X_i)X_i = 0 \end{cases}$$

With the first equation, we can get $\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{X}$, and put into the second equation,

$$\sum_{i=1}^{n}\left[(Y_i - \overline{Y}) - \widehat{\beta}(X_i - \overline{X})\right]X_i = 0$$

$$\therefore \widehat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})X_i}{\sum_{i=1}^{n}(X_i - \overline{X})X_i} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})} = \frac{S_{XY}}{S_{XX}}$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{X} = \overline{Y} - \frac{S_{XY}}{S_{XX}}\overline{X}$$

where $S_{XX} = \sum_{i=1}^{n}(X_i - \overline{X})^2$ and $S_{XY} = \sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})$.

Method 2

$$SSE = \sum_{i=1}^{n} e_i^2 = \|e\|^2, \text{ where } e = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} Y_1 - \widehat{Y}_1 \\ \vdots \\ Y_n - \widehat{Y}_n \end{pmatrix}, \text{ where } \widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}X_i$$

Since $\begin{pmatrix} \widehat{Y}_1 \\ \vdots \\ \widehat{Y}_n \end{pmatrix} \in \text{span}\left\{ \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \right\}$, then $\begin{pmatrix} Y_1 - \widehat{Y}_1 \\ \vdots \\ Y_n - \widehat{Y}_n \end{pmatrix}$ is not less than the distance from $\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ to $\text{span}\left\{ \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \right\}$. In order to get the minimum,

$\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \perp \text{span}\left\{ \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \right\}$.

$$\therefore \begin{cases} \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \perp \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \Rightarrow \sum_{i=1}^{n} (Y_i - \widehat{\alpha} - \widehat{\beta} X_i) = 0 \\ \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \perp \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \Rightarrow \sum_{i=1}^{n} (Y_i - \widehat{\alpha} - \widehat{\beta} X_i) X_i = 0 \end{cases},$$

which is the same as the result of taking the derivatives, so the answer is also

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{X} = \overline{Y} - \frac{S_{XY}}{S_{XX}}\overline{X}$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})X_i}{\sum_{i=1}^{n}(X_i - \overline{X})X_i} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})} = \frac{S_{XY}}{S_{XX}}$$

Properties of $\widehat{\alpha}$ and $\widehat{\beta}$

Since $\widehat{\beta} = \frac{S_{XY}}{S_{XX}} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{S_{XX}} Y_i - \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{S_{XX}} \overline{Y} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{S_{XX}} Y_i$,

then $\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{S_{XX}} Y_i \overline{X} = \sum_{i=1}^{n} (\frac{1}{n} - \frac{(X_i - \overline{X})}{S_{XX}} \overline{X}) Y_i$.

Therefore, $(\widehat{\alpha}, \widehat{\beta})$ is linear in Y_i.

And since $Y_i = \alpha + \beta X_i + \epsilon_i$, we have

$$\widehat{\beta} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{S_{XX}} (\beta X_i + \epsilon_i) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})X_i}{S_{XX}} \beta + \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{S_{XX}} \epsilon_i = \beta + \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{S_{XX}} \epsilon_i$$

$$\therefore \mathbb{E}\widehat{\beta} = \beta$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{X} = \alpha + \beta\overline{X} + \overline{\epsilon} - \left(\beta + \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{S_{XX}} \epsilon_i \right) \overline{X} = \alpha + \left(\frac{1}{n} - \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{S_{XX}} \overline{X} \right) \epsilon_i$$

$$\therefore \mathbb{E}\widehat{\alpha} = \alpha$$

Therefore, $(\widehat{\alpha}, \widehat{\beta})$ is unbiased.

Hence, $(\widehat{\alpha}, \widehat{\beta})$ is a linear unbiased estimator, and then we will introduce another estimator, BLUE.

1.1.2 BLUE (Best Linear Unbiased Estimator)

Definition: 'Best' stands for Variance Min, in other words, BLUE has the smallest variance among all linear unbiased estimators. Note that since it relies on the variance and biasedness of the parameters, BLUE has some requirements on the model, that is, the first and second moment of the parameters exist.

Then we are going to find out the BLUE of β. For a linear estimator $\widehat{\beta} = \sum_{i=1}^{n} d_i Y_i$, where $\{d_i\}$ are fixed numbers, our job is to find out what $\{d_i\}$ are. And since $\widehat{\beta}$ is unbiased, $\mathbb{E}\widehat{\beta} = \beta$, therefore

$$\sum_{i=1}^{n} d_i \mathbb{E}Y_i = \sum_{i=1}^{n} d_i(\alpha + \beta X_i) = \beta \quad \therefore \begin{cases} \sum_{i=1}^{n} d_i = 0 \\ \sum_{i=1}^{n} d_i X_i = 1 \end{cases}$$

Since $\mathbb{V}\text{ar}\widehat{\beta} = \sum_{i=1}^{n} d_i^2 \mathbb{V}\text{ar}Y_i = \sigma^2 \sum_{i=1}^{n} d_i^2$, then our job is to find out what $\{d_i\}$ are in the formula below.

$$\begin{aligned} & min \sum_{i=1}^{n} d_i^2 \\ s.t. & \sum_{i=1}^{n} d_i = 0 \\ & \sum_{i=1}^{n} d_i X_i = 1 \end{aligned}$$

Using Lagrange multiplier, we have
$L(d_i, a, b) = d_1^2 + \cdots + d_n^2 + a(d_1 + \cdots + d_n) + bd_1 X_1 + \cdots + d_n X_n$.

$$\therefore \begin{cases} 2d_1 + a + bX_1 = 0 \\ \cdots \\ 2d_n + a + bX_n = 0 \\ d_1 + \cdots + d_n = 0 \\ d_1 X_1 + \cdots + d_n X_n = 1 \end{cases}$$

With the first $n+1$ equations, we can get $a + b\overline{X} = 0$, that is, $a = -b\overline{X}$, and we multiply the first n equations with $X_1, \cdots, X_n$, respectively, then we can get

$$\begin{cases} 2d_1 X_1 + a + bX_1^2 = 0 \\ \cdots \\ 2d_n X_n + a + bX_n^2 = 0 \\ d_1 X_1 + \cdots + d_n X_n = 1 \end{cases} \Rightarrow 2 + an\overline{X} + b\sum_{i=1}^{n} X_i^2 = 0$$

$$\therefore 2 - b\overline{X}^2 n + b\sum_{i=1}^{n} X_i^2 = 0$$

$$\therefore b = -\frac{2}{-\overline{X}^2 n + \sum_{i=1}^{n} X_i^2} = -\frac{2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = -\frac{2}{S_{XX}}$$

Therefore, $a = \frac{2}{S_{XX}}\overline{X}$, and $d_i = \frac{X_i - \overline{X}}{S_{XX}}$, and consequently,
$\widehat{\beta} = \sum_{i=1}^{n} d_i Y_i = \sum_{i=1}^{n} \frac{X_i - \overline{X}}{S_{XX}} Y_i = \widehat{\beta}_{LSE}$. As we can see, the BLUE of simple linear regression is just the same as the LSE of that.

1.1.3 MLE (Maximum Likelihood Estimator)

Suppose that $\{\epsilon_i\} \sim N(0, \sigma^2)$, thus $\{\epsilon_i\}$ are independent, thus $\{Y_i\}$ are independent. Therefore,

$$Y_i|_{X_i} \sim N(\alpha + \beta X_i, \sigma^2)$$

$$\therefore pdf(Y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(Y_i - \alpha - \beta X_i)^2}{2\sigma^2}\right)$$

And because $\{Y_i\}$ is independent, therefore,

$$joint\ pdf = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}}exp\left(-\frac{(Y_i-\alpha-\beta X_i)^2}{2\sigma^2}\right)\right]$$

$$\therefore loglikelihood\ L(\sigma^2,\alpha,\beta) = -\sum_{i=1}^{n}\frac{(Y_i-\alpha-\beta X_i)^2}{2\sigma^2} - \frac{n}{2}log(\pi\sigma^2)$$

$$\therefore (\widehat{\sigma}^2,\widehat{\alpha},\widehat{\beta})_{MLE} = argmax_{(\widehat{\sigma}^2,\widehat{\alpha},\widehat{\beta})}L(\widehat{\sigma}^2,\widehat{\alpha},\widehat{\beta})$$

$$\therefore (\widehat{\alpha},\widehat{\beta})_{MLE} = argmax_{(\widehat{\alpha},\widehat{\beta})}\sum_{i=1}^{n}(Y_i-\widehat{\alpha}-\widehat{\beta}X_i)^2 = argmax_{(\widehat{\alpha},\widehat{\beta})}SSE(\widehat{\alpha},\widehat{\beta}) = (\widehat{\alpha},\widehat{\beta})_{LSE}$$

By way of conclusion, $(\widehat{\alpha},\widehat{\beta})_{LSE} = (\widehat{\alpha},\widehat{\beta})_{BLUE} = (\widehat{\alpha},\widehat{\beta})_{MLE}(\text{with } Y_i \sim Gaussian)$

Then we consider the distribution of $\begin{pmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{pmatrix}$.

1.1.3.1 Distribution of $\begin{pmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{pmatrix}$

Since $\begin{pmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{pmatrix}$ is linear in $\{Y_i\}$ and $Y_i|_{X_i} \sim N(\alpha+\beta X_i,\sigma^2)$ follow a normal distribution, $\begin{pmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{pmatrix}$ follow a normal distribution. And because $\begin{pmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{pmatrix}$ is unbiased, we can get

$$\begin{pmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{pmatrix} \sim N\left(\begin{pmatrix}\alpha\\ \beta\end{pmatrix}, \begin{pmatrix}\mathbb{V}\text{ar}\widehat{\alpha} & \mathbb{C}\text{ov}(\widehat{\alpha},\widehat{\beta})\\ \mathbb{C}\text{ov}(\widehat{\alpha},\widehat{\beta}) & \mathbb{V}\text{ar}\widehat{\beta}\end{pmatrix}\right)$$

Since $\widehat{\alpha} = \sum_{i=1}^{n}(\frac{1}{n} - \frac{(X_i-\overline{X})}{S_{XX}}\overline{X})Y_i$, then

$$\begin{aligned}\mathbb{V}\text{ar}\widehat{\alpha} &= \mathbb{V}\text{ar}\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{(X_i-\overline{X})}{S_{XX}}\overline{X}\right)Y_i = \sum_{i=1}^{n}\left(\frac{1}{n}-\frac{(X_i-\overline{X})}{S_{XX}}\overline{X}\right)^2\mathbb{V}\text{ar}Y_i\\ &= \sigma^2\sum_{i=1}^{n}\left(\frac{1}{n^2}+\frac{(X_i-\overline{X})^2\overline{X}^2}{S_{XX}^2}-\frac{2}{n}\frac{(X_i-\overline{X})}{S_{XX}}\overline{X}\right)\\ &= \sigma^2\left(\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}+0\right) = \sigma^2\left(\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}\right)\end{aligned}$$

Since $\widehat{\beta} = \sum_{i=1}^{n}\frac{(X_i-\overline{X})}{S_{XX}}Y_i$, then

$$\mathbb{V}\text{ar}\widehat{\beta} = \mathbb{V}\text{ar}\sum_{i=1}^{n}\frac{(X_i-\overline{X})}{S_{XX}}Y_i = \sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{S_{XX}^2}\mathbb{V}\text{ar}Y_i = \frac{\sigma^2}{S_{XX}}$$

$$\begin{aligned}\mathbb{C}\text{ov}(\widehat{\alpha},\widehat{\beta}) &= \mathbb{C}\text{ov}\left(\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{(X_i-\overline{X})}{S_{XX}}\overline{X}\right)Y_i, \sum_{j=1}^{n}\frac{(X_j-\overline{X})}{S_{XX}}Y_j\right)\\ &= \sum_{i,\,j}\left(\frac{1}{n}-\frac{(X_i-\overline{X})}{S_{XX}}\overline{X}\right)\frac{(X_j-\overline{X})}{S_{XX}}\mathbb{C}\text{ov}(Y_i,Y_j)\\ &= \sum_{i=1}^{n}\left(\frac{1}{n}-\frac{(X_i-\overline{X})}{S_{XX}}\overline{X}\right)\frac{(X_i-\overline{X})}{S_{XX}}\sigma^2\\ &= -\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{S_{XX}}\frac{\overline{X}}{S_{XX}}\sigma^2 = -\frac{\overline{X}}{S_{XX}}\sigma^2\end{aligned}$$

Therefore, $\begin{pmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{pmatrix} \sim N\left(\begin{pmatrix}\alpha\\ \beta\end{pmatrix}, \begin{pmatrix}\sigma^2(\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}) & -\frac{\overline{X}}{S_{XX}}\sigma^2\\ -\frac{\overline{X}}{S_{XX}}\sigma^2 & \frac{\sigma^2}{S_{XX}}\end{pmatrix}\right)$.

We want to consider a hypothesis testing: $H_0 : \beta = 0 \leftrightarrow H_1 : \beta \neq 0$. We now know that $\widehat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{XX}})$, that is, $\frac{\widehat{\beta}-\beta}{\sigma^2/S_{XX}} \sim N(0,1)$. So in order to conduct a hypothesis testing, we should use $\widehat{\sigma}^2$ to substitute σ^2.

1.1.3.2 Estimator of σ^2

Similarly, using MLE, we have

$$\begin{aligned}
\widehat{\sigma}^2_{MLE} &= argmax_{\widehat{\sigma}^2} L(\widehat{\sigma}^2, \widehat{\alpha}_{MLE}, \widehat{\beta}_{MLE}) \\
&= argmax_{\widehat{\sigma}^2} \left(-\sum_{i=1}^{n} \frac{(Y_i - \widehat{\alpha}_{MLE} - \widehat{\beta}_{MLE} X_i)^2}{2\widehat{\sigma}^2} - \frac{n}{2} log(\pi \widehat{\sigma}^2) \right) \\
\therefore \frac{\partial L}{\partial \widehat{\sigma}^2} &= \sum_{i=1}^{n} \frac{(Y_i - \widehat{\alpha}_{MLE} - \widehat{\beta}_{MLE} X_i)^2}{2\widehat{\sigma}^4} - \frac{n}{2}\frac{1}{\widehat{\sigma}^2} = 0 \\
\therefore \widehat{\sigma}^2_{MLE} &= \frac{1}{n}\sum_{i=1}^{n} (Y_i - \widehat{\alpha}_{MLE} - \widehat{\beta}_{MLE} X_i)^2
\end{aligned}$$

Then, we are going to check the biasedness of $\widehat{\sigma}^2_{MLE}$.

$$\begin{aligned}
\widehat{\sigma}^2_{MLE} &= \frac{1}{n}\sum_{i=1}^{n} (Y_i - \widehat{\alpha}_{MLE} - \widehat{\beta}_{MLE} X_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} (Y_i - \widehat{\alpha}_{LSE} - \widehat{\beta}_{LSE} X_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \frac{1}{n}\left\| \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} - \begin{pmatrix} \widehat{Y}_1 \\ \vdots \\ \widehat{Y}_n \end{pmatrix} \right\|^2 = \frac{1}{n}\|e\|^2
\end{aligned}$$

$$\because e \perp \begin{pmatrix} \widehat{Y}_1 \\ \vdots \\ \widehat{Y}_n \end{pmatrix}, \quad \therefore \widehat{\sigma}^2_{MLE} = \frac{1}{n}\|e\|^2 = \frac{1}{n}\left[\left\| \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} \widehat{Y}_1 \\ \vdots \\ \widehat{Y}_n \end{pmatrix} \right\|^2 \right] = \frac{1}{n}\sum_{i=1}^{n} (Y_i^2 - \widehat{Y}_i^2)$$

$$\therefore \mathbb{E}\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (\mathbb{E}Y_i^2 - \mathbb{E}\widehat{Y}_i^2)$$

$$\text{and } \because \mathbb{E}\widehat{Y}_i = \mathbb{E}(\widehat{\alpha} + \widehat{\beta} X_i) = \alpha + \beta X_i = \mathbb{E}(\alpha + \beta X_i + \epsilon_i) = \mathbb{E}Y_i$$

$$\begin{aligned}
\therefore \mathbb{E}\widehat{\sigma}^2_{MLE} &= \frac{1}{n}\sum_{i=1}^{n} (\mathbb{E}Y_i^2 - \mathbb{E}\widehat{Y}_i^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left[\left(\mathbb{E}Y_i^2 - (\mathbb{E}Y_i)^2\right) - \left(\mathbb{E}\widehat{Y}_i^2 - (\mathbb{E}\widehat{Y}_i)^2\right) \right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \left[\mathbb{E}(Y_i - \mathbb{E}Y_i)^2 - \mathbb{E}\left(\widehat{Y}_i - \mathbb{E}\widehat{Y}_i\right)^2 \right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{V}\text{ar}Y_i - \mathbb{V}\text{ar}\widehat{Y}_i \right)
\end{aligned}$$

$$\begin{aligned}
\because \mathbb{V}\text{ar}\widehat{Y}_i &= \mathbb{V}\text{ar}\widehat{\alpha} + \mathbb{V}\text{ar}\widehat{\beta} X_i^2 + 2\mathbb{C}\text{ov}(\widehat{\alpha}, \widehat{\beta}) X_i \\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right) + \frac{\sigma^2}{S_{XX}} X_i^2 - 2\frac{\overline{X}}{S_{XX}}\sigma^2 X_i \\
&= \frac{1}{n}\sigma^2 + \frac{\sigma^2}{S_{XX}}(X_i - \overline{X})^2
\end{aligned}$$

$$\therefore \mathbb{E}\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2 - \frac{1}{n}\sigma^2 - \frac{\sigma^2}{S_{XX}}(X_i - \overline{X})^2\right)$$
$$= \sigma^2 - \frac{1}{n}\sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-2}{n}\sigma^2$$

Therefore, $\widehat{\sigma}^2_{MLE}$ is biased, and we define a unbiased estimator of σ^2:

$$s^2 = \frac{n}{n-2}\widehat{\sigma}^2_{MLE} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}X_i)^2$$

Why is there an $n-2$ on the denominator, while it is an $n-1$ on that of one-dimensional MLE, that is,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

When $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, we have $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$, so if we define $P_i = X_i - \overline{X}$, then $\{P_i\}$ is not independent, its degree of freedom is n-1.

When $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}X_i)^2$, we have $\begin{cases} \sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}X_i) = 0 \\ \sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}X_i)X_i = 0 \end{cases}$, so there are two limits, and thus its degree of freedom is n-2. Next, we will discuss more about degree of freedom, we are going to look at the distribution of s^2.

1.1.3.3 Distribution of s^2

Theorem 1.1 $\{X_i\}_{i=1}^n \overset{iid}{\sim} N(\mu, \sigma^2)$, where μ, σ^2 are unknown. Previously, we got $\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$ and $\widehat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. There are three properties of $\overline{X}$ and s^2:

(1). $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

(2). $\frac{s^2(n-1)}{\sigma^2} \sim \chi^2(n-1)$

(3). $\overline{X} \perp s^2$

pf: The first property is very basic, so the proof is omitted.

For the second property, define $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, \sigma^2 I_n\right)$. We want to find an orthogonal matrix A, so that if we define an orthogonal transform $Y = AX$, then $\mathbb{V}\mathrm{ar}Y = \sigma^2 I_n$. Therefore, we construct

$$A = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2\cdot 1}} & \frac{-1}{\sqrt{2\cdot 1}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{3\cdot 2}} & \frac{1}{\sqrt{3\cdot 2}} & \frac{-1}{\sqrt{3\cdot 2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{pmatrix}$$

$$\therefore Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sqrt{n}\overline{X} \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

$$\text{and } Y = AX \sim N(A\mathbb{E}X, A\mathbb{C}\text{ov}(X)A^*) = N(A\mathbb{E}X, A\sigma^2 I_n A^*) = N(A\mathbb{E}X, \sigma^2 I_n)$$

$$\begin{aligned} \therefore Y_2^2 + \cdots + Y_n^2 &= Y_1^2 + \cdots + Y_n^2 - Y_1^2 = Y^*Y - n\overline{X}^2 = X^*A^*AX - n\overline{X}^2 \\ &= X^*X - n\overline{X}^2 = \sum_{i=1}^n (X_i^2 - \overline{X}^2) = \sum_{i=1}^n (X_i - \overline{X})^2 \end{aligned}$$

$$\text{and } \because Y_2, \cdots, Y_n \overset{iid}{\sim} N(0, \sigma^2)$$

$$\therefore \frac{s^2(n-1)}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} = \frac{Y_2^2 + \cdots + Y_n^2}{\sigma^2} \sim \chi^2(n-1)$$

Similarly, with the linear model, we have the following theorem.

Theorem 1.2

(1). $\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \sigma^2(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}) & -\frac{\overline{X}}{S_{XX}}\sigma^2 \\ -\frac{\overline{X}}{S_{XX}}\sigma^2 & \frac{\sigma^2}{S_{XX}} \end{pmatrix} \right)$

(2). $s^2 \sim \frac{1}{n-2}\sigma^2\chi^2(n-2)$

(3). $s^2 \perp \widehat{\alpha}, \widehat{\beta}$

pf: The first property has already been proved.

For the second property, $s^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \widehat{\alpha} - \widehat{\beta}X_i)^2$, define $e_i = Y_i - \widehat{\alpha} - \widehat{\beta}X_i$, then

$$\begin{aligned} e_i &= \alpha + \beta X_i + \epsilon_i - \left[\alpha + \sum_{j=1}^n \left(\frac{1}{n} - \frac{X_j - \overline{X}}{S_{XX}}\overline{X}\right)\epsilon_j\right] - \left[\beta + \sum_{j=1}^n \frac{X_j - \overline{X}}{S_{XX}}\epsilon_j\right] X_i \\ &= \sum_{j=1}^n \left(\delta_{ij} - \frac{1}{n} - \frac{(X_j - \overline{X})(X_i - \overline{X})}{S_{XX}}\right)\epsilon_j \\ &= \begin{pmatrix} \delta_{i1} - \frac{1}{n} - \frac{(X_1 - \overline{X})(X_i - \overline{X})}{S_{XX}} & \cdots & \delta_{in} - \frac{1}{n} - \frac{(X_n - \overline{X})(X_i - \overline{X})}{S_{XX}} \end{pmatrix} \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \end{aligned}$$

where $\delta_{ij} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$. Therefore,

$$\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = A \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} = \begin{pmatrix} \delta_{11} - \frac{1}{n} - \frac{(X_1 - \overline{X})(X_1 - \overline{X})}{S_{XX}} & \cdots & \delta_{1n} - \frac{1}{n} - \frac{(X_n - \overline{X})(X_1 - \overline{X})}{S_{XX}} \\ & \vdots & \\ \delta_{n1} - \frac{1}{n} - \frac{(X_1 - \overline{X})(X_n - \overline{X})}{S_{XX}} & \cdots & \delta_{nn} - \frac{1}{n} - \frac{(X_n - \overline{X})(X_n - \overline{X})}{S_{XX}} \end{pmatrix} \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

where $A_{ij} = \delta_{ij} - \frac{1}{n} - \frac{(X_j - \overline{X})(X_i - \overline{X})}{S_{XX}}$, so that $A_{ij} = A_{ji}$. Hence, A is symmetric.

Define $A = I_n - B$, so that $B_{ij} = \frac{1}{n} + \frac{(X_j - \overline{X})(X_i - \overline{X})}{S_{XX}}$, then

$$
\begin{aligned}
B_{ij}^2 &= \sum_{k=1}^{n} B_{ik}B_{kj} \\
&= \sum_{k=1}^{n} \left[\frac{1}{n} + \frac{(X_k - \overline{X})(X_i - \overline{X})}{S_{XX}}\right]\left[\frac{1}{n} + \frac{(X_j - \overline{X})(X_k - \overline{X})}{S_{XX}}\right] \\
&= \frac{1}{n} + 0 + \frac{(X_j - \overline{X})(X_i - \overline{X})}{S_{XX}^2}\sum_{k=1}^{n}(X_k - \overline{X})^2 \\
&= \frac{1}{n} + \frac{(X_j - \overline{X})(X_i - \overline{X})}{S_{XX}} = B_{ij}
\end{aligned}
$$

Therefore, B is idempotent, that is, $B = B^2$. Therefore,

$$
A^2 = (I - B)^2 = I + B^2 - 2B = I + B - 2B = I - B = A
$$

Therefore, A is idempotent, and thus, is always diagonalizable and its eigenvalues are either 0 or 1. Hence, we conduct an eigenvalue decomposition to A, that is $A = UDU^*$, where U is an orthogonal matrix and D is a diagonal matrix with either 0 or 1 on the diagonal. And since

$$
tr(A) = tr(I_n - B) = n - tr(B) = n - \sum_{i=1}^{n} B_{ii} = n - \sum_{i=1}^{n}\left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{S_{xx}}\right] = n - (1 + 1) = n - 2
$$

Therefore, $D = \text{diag}(1, \cdots, 1, 0, 0)$, so

$$
e^*e = \epsilon^* A^* A\epsilon = \epsilon^* A\epsilon = \epsilon^* UDU^*\epsilon = \xi^* D\xi = \sum_{i=1}^{n-2}\xi_i^2 \sim \sigma^2\chi^2(n-2)
$$

where $\xi = U^*\epsilon \sim N(0, \sigma^2 I_n)$. Therefore, $s^2 = \frac{1}{n-2}e^*e \sim \frac{1}{n-2}\sigma^2\chi^2(n-2)$.

For the third property,

$$
\begin{aligned}
\mathbb{C}\text{ov}(e_i, \widehat{\beta}) &= \mathbb{C}\text{ov}\left(\sum_{j=1}^{n}\left(\delta_{ij} - \frac{1}{n} - \frac{(X_j - \overline{X})(X_i - \overline{X})}{S_{XX}}\right)\epsilon_j, \beta + \sum_{j=1}^{n}\frac{X_j - \overline{X}}{S_{XX}}\epsilon_j\right) \\
&= \mathbb{C}\text{ov}\left(\sum_{j=1}^{n}\left(\delta_{ij} - \frac{1}{n} - \frac{(X_j - \overline{X})(X_i - \overline{X})}{S_{XX}}\right)\epsilon_j, \sum_{j=1}^{n}\frac{X_j - \overline{X}}{S_{XX}}\epsilon_j\right) \\
&= \sum_{j=1}^{n}\left(\delta_{ij} - \frac{1}{n} - \frac{(X_j - \overline{X})(X_i - \overline{X})}{S_{XX}}\right)\frac{X_j - \overline{X}}{S_{XX}}\sigma^2 \\
&= \left(\frac{X_i - \overline{X}}{S_{XX}} - 0 - \frac{X_i - \overline{X}}{S_{XX}}\right)\cdot 0 = 0
\end{aligned}
$$

And since $e_i = \sum_{j=1}^{n}\left(\delta_{ij} - \frac{1}{n} - \frac{(X_j - \overline{X})(X_i - \overline{X})}{S_{XX}}\right)\epsilon_j$ is Gaussian, and $\widehat{\beta}$ is also Gaussian, then $e_i \perp \widehat{\beta}$. Similarly, we will have $e_i \perp \widehat{\alpha}$.

1.1.4 Inference

1.1.4.1 Sum of Squares (Partitioning Variability)

SST (Total sum of squares)

$$
SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2
$$

It depends on the sample and is independent to the model.

SSR (Regression sum of squares)

$$SSR = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$$

It depends on the model. Its variation is due to the regression line, and can be explained by the model. There are two useful properties, the first one is

$$\overline{\widehat{Y}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_i = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\alpha} + \widehat{\beta}X_i) = \frac{1}{n}\sum_{i=1}^{n}(\overline{Y} - \widehat{\beta}\overline{X} + \widehat{\beta}X_i) = \overline{Y}$$

And the second one is

$$\begin{aligned} SSR &= \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{\alpha} + \widehat{\beta}X_i - \overline{Y})^2 = \sum_{i=1}^{n}(\overline{Y} - \widehat{\beta}\overline{X} + \widehat{\beta}X_i - \overline{Y})^2 \\ &= \sum_{i=1}^{n}\left(\widehat{\beta}(X_i - \overline{X})\right)^2 = \widehat{\beta}^2 S_{XX} = \frac{S_{XY}^2}{S_{XX}} \end{aligned}$$

SSE (Sum of squared errors / Residual sum of squares)

$$SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

There is a relationship between these sums of squares, that is, $SST = SSR + SSE$. Below is the proof.

$$\begin{aligned} SST &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) \\ &= SSE + SSR + 2\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)(\widehat{\alpha} + \widehat{\beta}X_i) - 2\sum_{i=1}^{n}\overline{Y}(Y_i - \widehat{Y}_i) \\ &= SSE + SSR + 2\sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}X_i)\widehat{\beta}X_i + 0 - 2\overline{Y}\,n(\overline{Y} - \overline{\widehat{Y}}) \\ &= SSE + SSR \end{aligned}$$

Degree of Freedom

Since, $SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = (n-2)s^2 \sim \sigma^2\chi^2(n-2)$, that is, the degree of freedom of SSE is $n-2$. And SST is linear with the sample variance of Y_i, so the degree of freedom of SST is $n-1$. Thus, the degree of freedom of SSR is 1.

Mean Square

Mean Square = Sum of Squares / Degree of Freedom, thus,

$$MSR = \frac{SSR}{df_R} = SSR, \;\; MSE = \frac{SSE}{df_E} = \frac{SSE}{n-2} = s^2,$$

that is, MSE is equal to s^2, which is the unbiased estimator of σ^2.

Quality if Fitted Model

Define $r^2 = \frac{SSR}{SST} \in [0,1]$ the Coefficient of Determination.

When $r^2 = 0$, $\frac{SSE}{SST} = \frac{SST - SSR}{SST} = 1 - \frac{SSR}{SST} = 1 - r^2 = 1$, that is, $SSE = SST$ reaches the maximum, which indicates that the model is bad fitted.

When $r^2 = 1$, $\frac{SSE}{SST} = \frac{SST-SSR}{SST} = 1 - \frac{SSR}{SST} = 1 - r^2 = 0$, that is, $SSE = 0$. Since $SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$, that is $\forall i,\ Y_i = \widehat{Y}_i$, which indicates that the model is well fitted.

1.1.4.2 Hypothesis Testing

Define $r(x) = \mathbb{E}(Y|X) = \alpha + \beta X$. If $\beta = 0$, then $r(X) \nsim X$, that is, X has no explanatory effect on Y. Hence, we construct a hypothesis testing that $H_0 :\ \beta = 0 \leftrightarrow H_1 :\ \beta \neq 0$.

Method 1

Use sum of squares. Since $SSR = \widehat{\beta}^2 S_{XX} \to \beta^2 S_{XX} (= 0, \text{under } H_0)$, therefore, if $SSR > c^*$, we reject H_0. Then we are going to use the test size α, say 0.05, and the distribution of SSR to determine the value of c^*.

Theorem 1.3

For any simple linear regression model, we will have:

(1). $SSE \sim \chi^2(n-2)\sigma^2$

(2). $SSR \sim \chi^2(1)\sigma^2(\text{under } H_0)$

(3). $SSE \perp SSR$

(4). $SST \sim \chi^2(n-1)\sigma^2(\text{under } H_0)$

pf: For (1), $SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = (n-2)s^2 \sim \sigma^2\chi^2(n-2)$.

For (2), $SSR = \frac{S_{XY}^2}{S_{XX}} = \left[\sum_{i=1}^{n}\frac{X_i - \overline{X}}{\sqrt{S_{XX}}}(Y_i - \overline{Y})\right]^2 = \left(\sum_{i=1}^{n}\frac{X_i - \overline{X}}{\sqrt{S_{XX}}}Y_i\right)^2$. Under $H_0 : \beta = 0,$ we have $Y_i = \alpha + \beta X_i + \epsilon_i = \alpha + \epsilon_i$. Therefore,

$$SSR = \left(\sum_{i=1}^{n}\frac{X_i - \overline{X}}{\sqrt{S_{XX}}}Y_i\right)^2 = \left(\sum_{i=1}^{n}\frac{X_i - \overline{X}}{\sqrt{S_{XX}}}(\alpha + \epsilon_i)\right)^2 = \left(\sum_{i=1}^{n}\frac{X_i - \overline{X}}{\sqrt{S_{XX}}}\epsilon_i\right)^2$$

Therefore, $\sqrt{SSR}$ linear in $\{\epsilon_i\}$, that is $\sqrt{SSR} \sim N(0, \mathbb{V}\text{ar}\sqrt{SSR})$. Since

$$\mathbb{V}\text{ar}\sqrt{SSR} = \mathbb{V}\text{ar}\left(\sum_{i=1}^{n}\frac{X_i - \overline{X}}{\sqrt{S_{XX}}}\epsilon_i\right) = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sqrt{S_{XX}}}\right)^2 \mathbb{V}\text{ar}\epsilon_i = \sigma^2\sum_{i=1}^{n}\frac{(X_i - \overline{X})^2}{S_{XX}} = \sigma^2$$

Therefore, $\sqrt{SSR} \sim N(0, \sigma^2)$, and thus, $SSR \sim \chi^2(1)\sigma^2(\text{under } H_0)$.

For (3), since $SSR = \widehat{\beta}^2 S_{XX},\ SSE = s^2$, and $s^2 \perp \widehat{\beta}$, then $SSE \perp SSR$.

With (1), (2), and (3), then (4) is obvious.

Now that we know $\frac{SSR}{\sigma^2} \sim \chi^2(1)(\text{under } H_0)$, however since σ^2 is unknown, we have to use s^2 to replace it. And because $\frac{s^2}{\sigma^2} = \frac{MSE}{\sigma^2} = \frac{SSE}{(n-2)\sigma^2} \sim \frac{\chi^2(n-2)\sigma^2}{(n-2)\sigma^2} = \frac{\chi^2(n-2)}{(n-2)}$, we define the F-statistic

$$F \triangleq \frac{MSR}{MSE} = \frac{SSR}{s^2} = \frac{\frac{SSR}{\sigma^2}}{\frac{s^2}{\sigma^2}} \sim \frac{\chi^2(1)}{\frac{\chi^2(n-2)}{(n-2)}} = F_{1,n-2}(\text{under } H_0)$$

Therefore, the reject region is $F > c^*$, where $\alpha = \mathbb{P}(\text{reject } H_0|H_0) = \mathbb{P}(F > c^*|H_0)$, that is $c^* = F_{1,n-2}(\alpha)$.

Method 2

If we can find the distribution of a function $f(\widehat{\beta}, \beta)$ which includes $\widehat{\beta} - \beta$, then we can design a hypothesis testing. And since $\widehat{\beta} \sim (\beta, \frac{\sigma^2}{S_{XX}})$, we have $\frac{\widehat{\beta}-\beta}{\sigma/\sqrt{S_{XX}}} \sim N(0,1)$. However, since σ is unknown, we have to use s to replace it. And because $\frac{s^2}{\sigma^2} \sim \frac{\chi^2(n-2)}{(n-2)}$, we define the t-statistic

$$t \triangleq \frac{\widehat{\beta} - \beta}{s/\sqrt{S_{XX}}} = \frac{\frac{\widehat{\beta}-\beta}{\sigma/\sqrt{S_{XX}}}}{s/\sigma} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-2)}{n-2}}} = t_{n-2}$$

Under H_0, $t = \frac{\widehat{\beta}-\beta}{s/\sqrt{S_{XX}}} = \frac{\widehat{\beta}}{s/\sqrt{S_{XX}}} \sim t_{n-2}$. Therefore, the reject region is $|t| > c^*$, where $\alpha = \mathbb{P}(\text{reject } H_0 | H_0) = \mathbb{P}(|t| > c^* | H_0)$, that is if $|t| > c^* = t_{n-2}(\alpha/2)$, we reject H_0.

the Equivalence of these Two Methods

For simple linear regression, under H_0, we will have $t^2 = \frac{\widehat{\beta}^2}{s^2/S_{XX}} = \frac{S_{XY}^2}{S_{XX}^2} \cdot \frac{S_{XX}}{s^2} = \frac{S_{XY}^2}{S_{XX}} \cdot \frac{1}{MSE} = \frac{MSR}{MSE} = F$. And since the reject region of these two methods are $|t| > t_{n-2}(\alpha/2)$ and $F > F_{1,n-2}(\alpha)$, we need to prove that $t_{n-2}^2(\alpha/2) = F_{1,n-2}(\alpha)$, so as to prove the equivalence of these two methods.

We define a random variable $X \sim t_{n-2}$, then $X^2 \sim F_{1,n-2}$. Hence,

$$\begin{aligned}
&\mathbb{P}(X^2 > F_{1,n-2}(\alpha)) = \alpha \\
&\mathbb{P}(X > t_{n-2}(\alpha/2)) = \alpha/2 \\
&\mathbb{P}(X^2 > t_{n-2}^2(\alpha/2)) = \mathbb{P}(X > t_{n-2}(\alpha/2)) + \mathbb{P}(X < -t_{n-2}(\alpha/2)) = \alpha \\
\therefore\ &\mathbb{P}(X^2 > F_{1,n-2}(\alpha)) = \alpha = \mathbb{P}(X^2 > t_{n-2}^2(\alpha/2))
\end{aligned}$$

Therefore, $t_{n-2}^2(\alpha/2) = F_{1,n-2}(\alpha)$. The equivalence of these two methods is proved.

1.1.4.3 Confidence Interval

Let's review how we calculate a confidence interval, say for x. First, we need an unbiased estimator of x, say $\widehat{x}$. Second, we need to derive the distribution of $\widehat{x}$, say $\widehat{x} \sim N(x, \mathbb{V}\text{ar}\widehat{x})$, which means $\frac{\widehat{x}-x}{\mathbb{V}\text{ar}\widehat{x}} \sim N(0,1)$. Third, since there is always σ in the expression of $\mathbb{V}\text{ar}\widehat{x}$, we need to use s to replace it. By replacing, we always divide $\frac{\widehat{x}-x}{\mathbb{V}\text{ar}\widehat{x}}$ by $\frac{s}{\sigma}$, which may turn a normal distribution to another distribution, say t-distribution. Finally, we have a function $f(\widehat{x} - x)$ following a certain distribution, and we define the confidence level $1-\alpha$, and solve the inequality $f(\widehat{x} - x)$ between the upper and lower $\alpha/2$ percentile of that distribution.

Mean Response

Suppose the first n observations are $(X_1, Y_1), \cdots, (X_n, Y_n)$, and define $\mu = \mathbb{E}(Y|X_{n+1})$. Then μ is called mean response, and we want to find the confidence interval of μ.

Since $\mu = \mathbb{E}(Y|X_{n+1}) = \alpha + \beta X_{n+1}$ is fixed but unknown, we define $\widehat{\mu} = \widehat{\alpha} + \widehat{\beta} X_{n+1}$, and we are going to find a function $f(\widehat{\mu} - \mu)$ following a certain distribution. Since

$$\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \sigma^2(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}) & -\frac{\overline{X}}{S_{XX}}\sigma^2 \\ -\frac{\overline{X}}{S_{XX}}\sigma^2 & \frac{\sigma^2}{S_{XX}} \end{pmatrix} \right)$$

$$
\begin{aligned}
\therefore \widehat{\mu} = \widehat{\alpha} + \widehat{\beta} X_{n+1} &= (1, X_{n+1}) \begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} \sim Gaussian \\
\mathbb{E}\widehat{\mu} &= \alpha + \beta X_{n+1} = \mu \\
\mathbb{V}\text{ar}\widehat{\mu} &= \mathbb{V}\text{ar}\widehat{\alpha} + \mathbb{V}\text{ar}(\widehat{\beta} X_{n+1}) + 2\mathbb{C}\text{ov}(\widehat{\alpha}, \widehat{\beta} X_{n+1})) \\
&= \sigma^2 (\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}) + X_{n+1}^2 \frac{\sigma^2}{S_{XX}} - 2X_{n+1} \frac{\overline{X}}{S_{XX}} \sigma^2 \\
&= \sigma^2 \left(\frac{1}{n} + \frac{1}{S_{XX}} (X_{n+1} - \overline{X})^2 \right) \\
\therefore \widehat{\mu} &\sim N \left(\mu, \sigma^2 \left(\frac{1}{n} + \frac{1}{S_{XX}} (X_{n+1} - \overline{X})^2 \right) \right) \\
\therefore \frac{\widehat{\mu} - \mu}{\sigma \sqrt{\frac{1}{n} + \frac{1}{S_{XX}} (X_{n+1} - \overline{X})^2}} &\sim N(0, 1)
\end{aligned}
$$

And since σ is unknown, we have to use s to replace it. Because $\frac{s^2}{\sigma^2} \sim \frac{\chi^2(n-2)}{(n-2)}$ and $\widehat{\mu} = \widehat{\alpha} + \widehat{\beta} X_{n+1} \perp s$, we will change the normal distribution to a t-distribution, that is

$$
\frac{\widehat{\mu} - \mu}{s \sqrt{\frac{1}{n} + \frac{1}{S_{XX}} (X_{n+1} - \overline{X})^2}} \sim t_{n-2}
$$

Then we define the confidence level $1 - \alpha$, so $\mathbb{P}(f(\widehat{\mu} - \mu) \in [L^*, U^*]) = 1 - \alpha$. With the t-distribution, then

$$
L^* = -t_{n-2}(\alpha/2),\ U^* = t_{n-2}(\alpha/2)
$$

Therefore, the confidence interval of μ with confidence level $1 - \alpha$ is

$$
\left[\widehat{\mu} - s \sqrt{\frac{1}{n} + \frac{1}{S_{XX}} (X_{n+1} - \overline{X})^2} \cdot t_{n-2}(\alpha/2),\ \widehat{\mu} + s \sqrt{\frac{1}{n} + \frac{1}{S_{XX}} (X_{n+1} - \overline{X})^2} \cdot t_{n-2}(\alpha/2) \right]
$$

Individual Response

Suppose the first n observations are $(X_1, Y_1), \cdots, (X_n, Y_n)$, and $Y_{n+1} = \alpha + \beta X_{n+1} + \epsilon_{n+1}$. Then Y_{n+1} is called individual response, and we want to find the confidence interval of Y_{n+1}.

Define $\widehat{Y}_{n+1} = \widehat{\alpha} + \widehat{\beta} X_{n+1}$, then $Y_{n+1} \perp \{X_1, \cdots, X_n, Y_1, \cdots, Y_n\}, \widehat{\alpha}, \widehat{\beta}, s^2$. Therefore,

$$
\widehat{Y}_{n+1} - Y_{n+1} = \widehat{\alpha} + \widehat{\beta} X_{n+1} - \alpha - \beta X_{n+1} - \epsilon_{n+1},
$$

where $\alpha + \beta X_{n+1}$ is fixed and $\widehat{\alpha} + \widehat{\beta} X_{n+1} \sim Gaussian,\ \epsilon_{n+1} \sim Gaussian$, and $\alpha + \widehat{\beta} X_{n+1} \perp \epsilon_{n+1}$. Therefore, $\widehat{Y}_{n+1} - Y_{n+1} \sim Gaussian$. And since $\mathbb{E}(\widehat{Y}_{n+1} - Y_{n+1}) = 0$, and

$$
\mathbb{V}\text{ar}(\widehat{Y}_{n+1} - Y_{n+1}) = \mathbb{V}\text{ar}(\widehat{\alpha} + \widehat{\beta} X_{n+1}) + \mathbb{V}\text{ar}(\epsilon_{n+1}) = \sigma^2 \left(\frac{1}{n} + \frac{1}{S_{XX}} (X_{n+1} - \overline{X})^2 \right) + \sigma^2,
$$

in which we use the result in the previous mean response that

$$
\mathbb{V}\text{ar}(\widehat{\alpha} + \widehat{\beta} X_{n+1}) = \mathbb{V}\text{ar}\widehat{\mu} = \sigma^2 \left(\frac{1}{n} + \frac{1}{S_{XX}} (X_{n+1} - \overline{X})^2 \right).
$$

Therefore,

$$
\frac{\widehat{Y}_{n+1} - Y_{n+1}}{\sigma \sqrt{\frac{1}{n} + \frac{1}{S_{XX}} (X_{n+1} - \overline{X})^2 + 1}} \sim N(0, 1).
$$

And since σ is unknown, we have to use s to replace it. Because $\frac{s^2}{\sigma^2} \sim \frac{\chi^2(n-2)}{(n-2)}$ and $\widehat{Y}_{n+1} - Y_{n+1} = f(\widehat{\alpha}, \widehat{\beta}, \epsilon_{n+1}) \perp s$, we will change the normal distribution to a t-distribution, that is

$$\frac{\widehat{Y}_{n+1} - Y_{n+1}}{s\sqrt{\frac{1}{n} + \frac{1}{S_{XX}}(X_{n+1} - \overline{X})^2 + 1}} \sim t_{n-2}.$$

Similarly, the confidence interval of Y_{n+1} with confidence level $1 - \alpha$ is

$$\left[\widehat{Y}_{n+1} - t_{n-2}(\alpha/2)s\sqrt{\frac{1}{n} + \frac{1}{S_{XX}}(X_{n+1} - \overline{X})^2 + 1},\ \widehat{Y}_{n+1} + t_{n-2}(\alpha/2)s\sqrt{\frac{1}{n} + \frac{1}{S_{XX}}(X_{n+1} - \overline{X})^2 + 1}\right]$$

As we can see, the difference is that there is an extra "$+1$" in the square in the confidence interval of Y_{n+1}, which suggests that the confidence interval of Y_{n+1} is a little longer than that of μ.

1.2 Multiple Linear Regression

Now that we have k independent random variables, that is, in the population level, a standard model will be $y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. Just like in Simple Linear Regression, y is also called dependent (response) variable in Multiple Linear Regression. And for each X_i, it is a 1×1 dimensional number. Therefore, we can write the standard model into matrix form, that is,

$$y = (1, X_1, \cdots, X_k) \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \epsilon = (1, X_1, \cdots, X_k)\,\beta + \epsilon,$$

where $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}_{(k+1)\times 1}$.

And for each observed data $y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i$, where $i = 1, \cdots, n$, and X_{ij} represents the i-th observation of the j-th explanatory variable. Thus, if we write all n observations together in a matrix form, that will be

$$Y_{n\times 1} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} = X_{n\times(k+1)}\beta_{(k+1)\times 1} + \epsilon_{n\times 1},$$

where $\epsilon \sim N(0, \sigma^2 I_n)$, and X is called design matrix or data matrix and $X[i, j]$ represents the i-th observation of the $(j-1)$-th explanatory variable.

Since $\epsilon \sim N(0, \sigma^2 I_n)$, $Y|_X \sim N(X\beta, \sigma^2 I_n)$, and therefore, define the regression plane as $r(x) = \mathbb{E}(Y|X) = X\beta$, and define the fitted value as $\widehat{Y} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix}$ and fitted regression plane as $X\widehat{\beta}$, and thus, the residual as $e = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} y_1 - \hat{y}_1 \\ \vdots \\ y_n - \hat{y}_n \end{pmatrix} = Y - \widehat{Y}$.

1.2.1 LSE in Multiple Linear Regression

We are going to find the LSE of β, which we defined as b. Suppose $\widehat{\beta}$ is an estimator of β, then

$$b = argmin_{\widehat{\beta}} SSE = argmin_{\widehat{\beta}} \|e\|^2 = argmin_{\widehat{\beta}} \|Y - \widehat{Y}\|^2 = argmin_{\widehat{\beta}} \|Y - X\widehat{\beta}\|^2$$
$$= argmin_{\widehat{\beta}} (Y - X\widehat{\beta})^T (Y - X\widehat{\beta}) = argmin_{\widehat{\beta}} Y^T Y - Y^T X\widehat{\beta} - \widehat{\beta}^T X^T Y + \beta^T X^T X\widehat{\beta}$$

$$\therefore \frac{\partial SSE}{\partial \widehat{\beta}} = 0 \Leftrightarrow X^T X\widehat{\beta} = X^T Y \Leftrightarrow \widehat{\beta} = (X^T X)^{-1}(X^T Y)$$

Similar to Simple Linear Regression, we have a linear space explanation for LSE. Since $SSE = \|Y - \widehat{Y}\|^2$, where $\widehat{Y} = X\widehat{\beta} = \sum_{i=0}^{k} \widehat{\beta}_i X_i \in span\{1, X_1, \cdots, X_k\}$, in which $X_0 = 1$. Therefore, $\widehat{Y}$ should be the projection from Y to $span\{1, X_1, \cdots, X_k\}$, which we defined as $\widehat{Y} = HY$, where H is the projection matrix, and $H = X(X^T X)^{-1} X^T$. H has the property that

$$\forall \text{ vector } Y, \quad \widehat{Y} = HY = X(X^T X)^{-1} X^T Y = X\widehat{\beta} \in span\{1, X_1, \cdots, X_k\}.$$

1.2.2 BLUE in Multiple Linear Regression

For any linear estimator of β, we can write in the form that $\widehat{\beta} = [(X^T X)^{-1} X^T + \Delta]Y$. If $\widehat{\beta}$ is unbiased, $\mathbb{E}\widehat{\beta} = [(X^T X)^{-1} X^T + \Delta]\mathbb{E}Y = [(X^T X)^{-1} X^T + \Delta]X\beta = \beta + \Delta X\beta = \beta$, that is $\Delta X = 0$. Then, we are going to calculate the covariance of $\widehat{\beta}$.

$$\begin{aligned}
\mathbb{C}\text{ov}(\widehat{\beta}) &= [(X^T X)^{-1} X^T + \Delta]\sigma^2 I[(X^T X)^{-1} X^T + \Delta]^T \\
&= \sigma^2[(X^T X)^{-1} X^T + \Delta][X(X^T X)^{-1} + \Delta^T] \\
&= \sigma^2[(X^T X)^{-1} X^T X(X^T X)^{-1} + \Delta X(X^T X)^{-1} + (X^T X)^{-1} X^T \Delta^T + \Delta\Delta^T] \\
(\because \Delta X = 0) &= \sigma^2[(X^T X)^{-1} + \Delta\Delta^T] \\
(\because \Delta\Delta^T \succeq 0) &\succeq \sigma^2 (X^T X)^{-1}
\end{aligned}$$

Therefore, $\mathbb{V}\text{ar}(\widehat{\beta}_i) \geq \sigma^2 (X^T X)^{-1}_{(i,i)}$, "=" $\Leftrightarrow \forall i,\ \Delta\Delta^T_{(i,i)} = \sum_j (\Delta_{(i,j)})^2 = 0 \Leftrightarrow \Delta = 0$, that is, $\widehat{\beta}_{BLUE} = [(X^T X)^{-1} X^T]Y = \widehat{\beta}_{LSE}$.

1.2.3 MLE in Multiple Linear Regression

As we suppose $\epsilon \sim N(0, \sigma^2 I_n)$, $Y|_X \sim N(X\beta, \sigma^2 I_n)$, we can derive the pdf of Y, that is,

$$l(\beta) = \frac{exp\left\{-\frac{1}{2}(Y - X\beta)^T (\sigma^2 I_n)^{-1}(Y - X\beta)\right\}}{\sqrt{(2\pi)^n |\sigma^2 I_n|}}$$

where $|\sigma^2 I_n|$ means the determinant of $\sigma^2 I_n$. Therefore,

$$\widehat{\beta}_{MLE} = argmax_{\widehat{\beta}}((Y - X\beta)^T (Y - X\beta)) = argmax_{\widehat{\beta}} SSE = \widehat{\beta}_{LSE} = \widehat{\beta}_{BLUE} = b$$

1.2.3.1 Estimator of σ^2

Similar to Simple Linear Regression, we want to estimate σ^2 with MLE, which is

$$\widehat{\sigma}^2_{MLE} = argmax_{\widehat{\sigma}^2} l(b, \widehat{\sigma}^2) = argmin_{\widehat{\sigma}^2} \frac{1}{2\widehat{\sigma}^2}(Y - Xb)^T (Y - Xb) + \frac{n}{2}\log\widehat{\sigma}^2$$
$$\frac{\partial}{\partial \widehat{\sigma}^2} = 0 \Leftrightarrow -\frac{1}{2\widehat{\sigma}^4}(Y - Xb)^T (Y - Xb) + \frac{n}{2\widehat{\sigma}^2} \Leftrightarrow \widehat{\sigma}^2 = \frac{1}{n}(Y - Xb)^T (Y - Xb)$$

With the projection matrix we defined previously that $H = X(X^T X)^{-1} X^T$, we have

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}(Y - Xb)^T (Y - Xb) = \frac{1}{n}(Y - HY)^T (Y - HY) = \frac{1}{n} Y^T (I - H)^T (I - H)Y$$

Before we derive the biasedness of $\widehat{\sigma}^2_{MLE}$, we first discuss some properties of the projection matrix H.

1. $H^T = H$, that is, H is symmetric, and $I - H$ is also symmetric;

2. $H^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$, that is, H is idempotent;
3. $(I-H)^2 = I - 2H + H^2 = I - 2H + H = I - H$, that is, $I - H$ is also idempotent;
4. $trH = trX(X^TX)^{-1}X^T = tr(X^TX)^{-1}X^TX = trI_{k+1} = k+1$;
5. $tr(I-H) = trI_n - trH = n - k - 1$;
6. $\forall x \in span\{1, X_1, \cdots, X_k\},\ Hx = x,\ (I-H)x = 0.$

With the above properties, we can now easily derive the expectation of $\widehat{\sigma}^2_{MLE}$,

$$\begin{aligned}
\mathbb{E}\widehat{\sigma}^2_{MLE} &= \mathbb{E}\frac{1}{n}Y^T(I-H)^T(I-H)Y \quad (\because I-H \text{ is symmetric and idempotent}) \\
&= \mathbb{E}\frac{1}{n}Y^T(I-H)Y \\
&= \mathbb{E}\frac{1}{n}(X\beta+\epsilon)^T(I-H)(X\beta+\epsilon) \quad (\because (I-H)X = 0) \\
&= \mathbb{E}\frac{1}{n}\epsilon^T(I-H)\epsilon \\
&= \mathbb{E}tr\frac{1}{n}(I-H)\epsilon\epsilon^T \\
&= \frac{1}{n}tr(I-H)\mathbb{E}\epsilon\epsilon^T \quad (\because \mathbb{E}\epsilon\epsilon^T = \mathbb{C}\text{ov}\epsilon + (\mathbb{E}\epsilon)^2 = \sigma^2 I_n + 0 = \sigma^2 I_n) \\
&= \frac{\sigma^2 tr(I-H)}{n} \quad (\because tr(I-H) = n-k-1) \\
&= \frac{n-k-1}{n}\sigma^2
\end{aligned}$$

Therefore, $\widehat{\sigma}^2_{MLE}$ is biased, but we can define an unbiased estimator,

$$s^2 = \frac{n}{n-k-1}\widehat{\sigma}^2_{MLE} = \frac{1}{n-k-1}(Y-Xb)^T(Y-Xb)$$

1.2.3.2 Distribution of b and s^2

Then, we would like to find out the distribution of b and s^2.

First, since $b = (X^TX)^{-1}X^TY$, and $Y \sim N(X\beta, \sigma^2 I_n)$, therefore,

$$\begin{aligned}
\mathbb{E}b &= (X^TX)^{-1}X^TX\beta = \beta \\
\mathbb{C}\text{ov}b &= (X^TX)^{-1}X^T\mathbb{C}\text{ov}YX(X^TX)^{-1} = \sigma^2(X^TX)^{-1} \\
\therefore b &\sim N(\beta, \sigma^2(X^TX)^{-1})
\end{aligned}$$

Second, $s^2 = \frac{1}{n-k-1}(Y-Xb)^T(Y-Xb) = \frac{1}{n-k-1}\epsilon^T(I-H)\epsilon$, and with eigenvalue decomposition, we have $I-H = UDU^T$, where $D = \text{diag}\{1, \cdots, 1, 0, \cdots, 0\}$, in which there are $(n-k-1)$ of 1's. And we define $\eta = U^T\epsilon \sim N(0, \sigma^2 I_n)$. Therefore,

$$\begin{aligned}
s^2 &= \frac{1}{n-k-1}\epsilon^T(I-H)\epsilon \\
&= \frac{1}{n-k-1}\epsilon^TUDU^T\epsilon \\
&= \frac{1}{n-k-1}\eta^TD\eta \\
&= \frac{1}{n-k-1}\sum_{i=1}^{n-k-1}\eta_i^2 \sim \frac{\sigma^2\chi^2(n-k-1)}{n-k-1}
\end{aligned}$$

Third, we are going to prove that $b \perp s^2$.

Since $b = (X^TX)^{-1}X^TY$, and $s^2 = \frac{1}{n-k-1}\|(I-H)Y\|_2^2 \triangleq f((I-H)Y)$, we are going to prove that $(X^TX)^{-1}X^TY \perp (I-H)Y$. And since

$$\begin{pmatrix} (X^TX)^{-1}X^TY \\ (I-H)Y \end{pmatrix} = \begin{pmatrix} (X^TX)^{-1}X^T \\ I-H \end{pmatrix} Y \sim Gaussian,$$

all we need to prove is that $\mathbb{C}\text{ov}\left((X^TX)^{-1}X^TY, (I-H)Y\right) = 0$. Since

$$\begin{aligned} \mathbb{C}\text{ov}\left((X^TX)^{-1}X^TY, (I-H)Y\right) &= (X^TX)^{-1}X^T\mathbb{C}\text{ov}Y(I-H)^T \\ &= \sigma^2(X^TX)^{-1}X^T(I-H)^T \\ &= \sigma^2(X^TX)^{-1}((I-H)X)^T \quad (\because (I-H)X = 0) \\ &= 0 \end{aligned}$$

Therefore, we proved that $b \perp s^2$.

1.2.4 Confidence Interval

Confidence Interval of b_i

First, we are going to look at the confidence interval of a single b_i. Since $b \sim N(\beta, \sigma^2(X^TX)^{-1})$, then $b_i \sim N(\beta_i, \sigma^2(X^TX)^{-1}_{[i+1,i+1]})$. Therefore, $\frac{b_i-\beta_i}{\sigma\sqrt{(X^TX)^{-1}_{[i+1,i+1]}}} \sim N(0,1)$. And since σ is unknown, we have to use s to replace it. Because $\frac{s^2}{\sigma^2} \sim \frac{\chi^2(n-k-1)}{n-k-1}$ and $b \perp s^2$, we will change the normal distribution to a t-distribution, that is

$$\frac{b_i - \beta_i}{s\sqrt{(X^TX)^{-1}_{[i+1,i+1]}}} \sim t_{n-k-1}$$

Therefore, the confidence interval of b_i is $\left[b_i \pm t_{n-k-1}(\alpha/2)s\sqrt{(X^TX)^{-1}_{[i+1,i+1]}}\right]$.

Mean Response

Second, just like in Simple Linear Regression, we can also derive the mean response if we have the first n observations $(X_1, Y_1), \cdots, (X_n, Y_n)$, where

$$X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

If we observe that $X_{n+1} = (1, X_{n+1,1}, \cdots, X_{n+1,k})$, as we previously addressed, we may use $\widehat{Y}_{n+1} = X_{n+1}\widehat{\beta}$ to estimate $\mu = \mathbb{E}Y_{n+1} = X_{n+1}\beta$. Since

$$\widehat{Y}_{n+1} = X_{n+1}\widehat{\beta} \sim N(X_{n+1}\beta, X_{n+1}\mathbb{C}\text{ov}\widehat{\beta}X_{n+1}^T) = N(X_{n+1}\beta, \sigma^2X_{n+1}(X^TX)^{-1}X_{n+1}^T)$$

$$\because \widehat{Y}_{n+1} \perp s, \quad \therefore \frac{\frac{\widehat{Y}_{n+1}-\mu}{\sigma\sqrt{X_{n+1}(X^TX)^{-1}X_{n+1}^T}}}{s/\sigma} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-k-1)}{n-k-1}}}$$

Therefore, the confidence interval of μ is $\left[\widehat{Y}_{n+1} \pm t_{n-k-1}(\alpha/2)s\sqrt{X_{n+1}(X^TX)^{-1}X_{n+1}^T}\right]$.

1.2.5 Inference

Now that we have the first n observations $(X_1, Y_1), \cdots, (X_n, Y_n)$, we can build a regression plane $X\widehat{\beta}$, and we want to know if it is well fitted.

1.2.5.1 Sum of Squares

Just like in Simple Linear Regression, we can also derive three sum of squares, SST, SSR, and SSE.

SST

$$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \begin{pmatrix} Y_1 - \overline{Y}, \cdots, Y_n - \overline{Y} \end{pmatrix} \begin{pmatrix} Y_1 - \overline{Y} \\ \vdots \\ Y_n - \overline{Y} \end{pmatrix} = (Y - 1_n\overline{Y})^T(Y - 1_n\overline{Y})$$

$$\because \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \begin{pmatrix} \frac{1}{n}, \cdots, \frac{1}{n} \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \frac{1}{n}1_n^T Y$$

$$\therefore SST = (Y - \frac{1}{n}1_n 1_n^T Y)^T(Y - \frac{1}{n}1_n 1_n^T Y) = Y^T(I - \frac{1}{n}1_n 1_n^T)^T(I - \frac{1}{n}1_n 1_n^T)Y$$

Since, $(I - 1_n 1_n^T)$ is symmetric and idempotent, that is,

$$(I - \frac{1}{n}1_n 1_n^T)^T(I - \frac{1}{n}1_n 1_n^T) = (I - \frac{1}{n}1_n 1_n^T)^2 = I - \frac{2}{n}1_n 1_n^T + \frac{1}{n^2}1_n 1_n^T 1_n 1_n^T = I - \frac{1}{n}1_n 1_n^T$$

$$\therefore SST = Y^T(I - \frac{1}{n}1_n 1_n^T)Y$$

SSR

$$SSR = \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{\widehat{Y}}\right)^2 = \left(\widehat{Y} - 1_n\overline{\widehat{Y}}\right)^T\left(\widehat{Y} - 1_n\overline{\widehat{Y}}\right)$$

$$\because \overline{\widehat{Y}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_i = \frac{1}{n}1_n^T\widehat{Y} = \frac{1}{n}1_n^T HY = \frac{1}{n}(H1_n)^T Y \overset{H \text{ is proj.}}{=} \frac{1}{n}1_n^T Y$$

$$\begin{aligned}
\therefore SSR &= \left(HY - \frac{1}{n}1_n 1_n^T Y\right)^T\left(HY - \frac{1}{n}1_n 1_n^T Y\right) \\
&= Y^T\left(H - \frac{1}{n}1_n 1_n^T\right)^T\left(H - \frac{1}{n}1_n 1_n^T\right)Y \\
&= Y^T\left(H - \frac{1}{n}1_n 1_n^T\right)\left(H - \frac{1}{n}1_n 1_n^T\right)Y \\
&= Y^T\left(H^2 - \frac{1}{n}1_n 1_n^T H - \frac{1}{n}H1_n 1_n^T + \frac{1}{n^2}1_n 1_n^T 1_n 1_n^T\right)Y \\
&= Y^T\left(H - \frac{1}{n}1_n 1_n^T - \frac{1}{n}1_n 1_n^T + \frac{1}{n^2}n1_n 1_n^T\right)Y \\
&= Y^T\left(H - \frac{1}{n}1_n 1_n^T\right)Y
\end{aligned}$$

SSE

$$SSE = (Y - \widehat{Y})^T(Y - \widehat{Y}) = Y^T(I - H)^T(I - H)Y = Y^T(I - H)Y$$

Therefore, to sum up,

$$\begin{cases}
SST = Y^T(I - \frac{1}{n}1_n 1_n^T)Y = \|(I - \frac{1}{n}1_n 1_n^T)Y\|_2^2, & df_T = n - 1 \\
SSR = Y^T(H - \frac{1}{n}1_n 1_n^T)Y = \|(H - \frac{1}{n}1_n 1_n^T)Y\|_2^2, & df_R = k \\
SSE = Y^T(I - H)Y = \|(I - H)Y\|_2^2, & df_E = n - k - 1
\end{cases}$$

And we can also define that $MSR = SSR/k,\ MSE = SSE/n - k - 1$. And then, we are going to prove that $SSR \perp SSE$. That is to prove $(H - \frac{1}{n}1_n 1_n^T)Y$ and $(I - H)Y$ are independent. Since

$$\begin{pmatrix} (H - \frac{1}{n}1_n 1_n^T)Y \\ (I - H)Y \end{pmatrix} = \begin{pmatrix} H - \frac{1}{n}1_n 1_n^T \\ I - H \end{pmatrix} Y \sim Gaussian,$$

all we need to prove is that $\mathbb{C}\text{ov}\left((H - \frac{1}{n}1_n 1_n^T)Y, (I - H)Y\right) = 0$. Since

$$\begin{aligned}\mathbb{C}\text{ov}\left((H-\frac{1}{n}1_n1_n^T)Y,(I-H)Y\right)&=(H-\frac{1}{n}1_n1_n^T)\mathbb{C}\text{ov}Y(I-H)^T\\&=\sigma^2(H-\frac{1}{n}1_n1_n^T)(I-H)\\&=\sigma^2(H-H^2-\frac{1}{n}1_n((I-H)1_n)^T)\\(\because(I-H)1_n=0,\ H=H^2)&=0\end{aligned}$$

Hence, $SSR\perp SSE$.

R-squared

Similarly, we can define the R-squared,

$$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}\in[0,1]$$

And $\sqrt{R^2}=\mathbb{C}\text{orr}(Y,\widehat{Y})$ is also called the multiple correlation coefficient. Here's the proof.

Since

$$\sum_{i=1}^n\widehat{Y}_i=1_n^T\widehat{Y}=1_n^THY=(H1_n)^TY=1_n^TY=\sum_{i=1}^nY_i,$$

and

$$\sum_{i=1}^n\widehat{Y}_i^2=\widehat{Y}^T\widehat{Y}=Y^TH^THY=Y^THY=Y^T\widehat{Y}=\sum_{i=1}^n\widehat{Y}_iY_i,$$

we have

$$\sum_{i=1}^n(\widehat{Y}_i-\overline{Y})^2=\sum_{i=1}^n(\widehat{Y}_i-\overline{Y})(Y_i-\overline{Y})$$

Therefore,

$$\begin{aligned}R^2&=\frac{\sum_{i=1}^n(\widehat{Y}_i-\overline{Y})^2}{\sum_{i=1}^n(Y_i-\overline{Y})^2}=\frac{\sum_{i=1}^n(\widehat{Y}_i-\overline{Y})^2}{\sum_{i=1}^n(Y_i-\overline{Y})^2}\cdot\frac{\sum_{i=1}^n(\widehat{Y}_i-\overline{Y})^2}{\sum_{i=1}^n(\widehat{Y}_i-\overline{Y})^2}\\&=\frac{\sum_{i=1}^n(\widehat{Y}_i-\overline{Y})(Y_i-\overline{Y})}{\sum_{i=1}^n(Y_i-\overline{Y})^2}\cdot\frac{\sum_{i=1}^n(\widehat{Y}_i-\overline{Y})(Y_i-\overline{Y})}{\sum_{i=1}^n(\widehat{Y}_i-\overline{Y})^2}=\rho^2(Y,\widehat{Y})\end{aligned}$$

Next, we wonder if R-squared is the bigger the better? Think about if we add a X_{k+1} which has no explanatory effect on Y. Then, how does R-squared change?

Since $span\{1,X_1,\cdots,X_k\}\leq span\{1,X_1,\cdots,X_{k+1}\}$, the SSE corresponding to the projection on $span\{1,X_1,\cdots,X_{k+1}\}$ will be smaller than that on $span\{1,X_1,\cdots,X_k\}$. So, R-squared will definitely increase if we add a new variable. Hence, we introduce a new metric, Adjusted R-squared,

$$R_a^2=1-\frac{MSE}{MST}=1-\frac{SSE/n-k-1}{SST/n-1}=1-\frac{MSE}{MST}\frac{n-1}{n-k-1}$$

1.2.5.2 Hypothesis Testing

We can do two kinds of hypothesis testing on a multiple linear regression model.

First, we can test on the whole regression model, that is we want to know if $(X_1,\cdots,X_k)$ as a whole has explanatory effect on Y. Hence, we set the null hypothesis as $H_0:\ \beta_0=\cdots=\beta_k=0$, and we reject H_0 if there exists $\beta_i\neq0$. Here we use F-test, that is,

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/n-k-1} \overset{SSR \perp SSE}{\sim} \frac{\chi^2(k)/k}{\chi^2(n-k-1)/n-k-1} = F_{k,n-k-1}$$

Therefore, if $F > F_{k,n-k-1}(\alpha)$, then we reject H_0.

Second, we can test on a single X_i. Since we've rejected the previous null hypothesis, we want to know which X_i has the greatest explanatory effect on Y. Hence, we set the null hypothesis as $H_0 : \beta_i = 0$. Since $b \sim N(\beta, \sigma^2(X^TX)^{-1})$, $b_i \sim N(\beta_i, \sigma^2(X^TX)^{-1}_{[i+1,i+1]})$. Here we use t-test, that is,

$$t = \frac{b_i}{s\sqrt{(X^TX)^{-1}_{[i+1,i+1]}}} \sim t_{n-k-1}$$

Therefore, if $|t| > t_{n-k-1}(\alpha/2)$, then we reject H_0.

1.2.5.3 Extra sum of squares

Considering a linear model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, with two explanatory variable X_1, X_2, if we have a priori knowledge that the explanatory effect of $X_1 - X_2$ is greater, we may change the model into $Y = \alpha_0 + \alpha_1(X_1 - X_2) + \epsilon$. So, we wonder when we can do this changing. The answer is that when we accept the null hypothesis, $H_0 : \beta_1 + \beta_2 = 0$. In this way, we can extend it to any multiple linear regression, $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$, which we defined as the full model. Now, if we set a null hypothesis, $H_0 : C_{m\times(k+1)}\beta_{(k+1)\times 1} = 0$, where $m \leq k+1$, then under H_0, there are $(k+1-m)$ free variables, including 1_n, that is, without loss of generality, we can suppose $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{k-m} \\ * \end{pmatrix}$, and denote $\alpha = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{k-m} \end{pmatrix}$ as the coefficient of $(k+1-m)$ free variables. Then, we will derive the reduced model,

$$Y = X\beta + \epsilon \overset{H_0}{=} Z\alpha + \epsilon = \beta_0 + \beta_1 X_1^* + \cdots + \beta_{k-m} X_{k-m}^* + \epsilon,$$

where $Z = (1_n, X_1^*, \cdots, X_{k-m}^*)$. For both the full model and the reduced model, we can get an SSR and an SSE, that is,

$$\begin{aligned}
&SSR_F = Y^T(H_F - \frac{1}{n}1_n 1_n^T)Y, \;\; H_F = X(X^TX)^{-1}X^T, \;\; X = (1_n, X_1, \cdots, X_k) \\
&SSE_F = Y^T(I - H_F)Y \\
&SSE_R = Y^T(H_R - \frac{1}{n}1_n 1_n^T)Y, \;\; H_R = X^*(X^{*T}X^*)^{-1}X^{*T}, \;\; X^* = (1_n, X_1^*, \cdots, X_{k-m}^*) \\
&SSE_R = Y^T(I - H_R)Y
\end{aligned}$$

Since variables in the full model are more than that in the reduced model, $SSR_F \geq SSR_R$. Hence, we define a statistic $SSR_F - SSR_R$, which is called **extra sum of squares**. Now we are going to simplify the form of the problem, we gather all the first k_1 free variables as X_1, and the rest k_2 non-free variables as X_2. Then, the full model will be $Y = 1_n\beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$, where $\beta = \begin{pmatrix} {\beta_0}^{1\times 1} \\ {\beta_1}^{k_1\times 1} \\ {\beta_2}^{k_2\times 1} \end{pmatrix}$. Then the SSR and SSE for the full model is

$$\begin{aligned}
SSR(X_1, X_2) &= Y^T(H_F - \frac{1}{n}1_n 1_n^T)Y \\
SSE(X_1, X_2) &= Y^T(I - H_F)Y
\end{aligned}$$

where $H_F = X(X^TX)^{-1}X^T$, in which $X = (1_n, X_1, X_2)$. And the SSR and SSE for the reduced model is

$$SSR(X_1) = Y^T(H_R - \frac{1}{n}1_n 1_n^T)Y$$
$$SSE(X_1) = Y^T(I - H_R)Y$$

where $H_R = X^*(X^{*T}X^*)^{-1}X^{*T}$, in which $X^* = (1_n, X_1)$.

Therefore, we can derive the expression of extra sum of squares,

$$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1) = SSR_F - SSR_R \geq 0$$

Meanwhile, $SST = SSR + SSE$, and SST does not change since it does not rely on the model. Therefore,

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2) = SSE_R - SSE_F \geq 0$$

Then, let's look at the distribution of $SSR(X_2|X_1)$. In the full model, $SSE_F = \|Y - Xb\|^2$, where $b = (X^TX)^{-1}X^TY$. When it comes to the reduced model, we also need to derive the least square estimator of β, that is,

$$\widehat{\beta}_{LSE} = argmin_{\widehat{\beta}}\|Y - X\widehat{\beta}\|^2 \quad s.t. C\widehat{\beta} = 0$$

Using Lagrange Multiplier Method, we can introduce a $m \times 1$ dimensional vector λ, that is,

$$f(\widehat{\beta}, \lambda) = \|Y - X\widehat{\beta}\|^2 + \lambda^T C\widehat{\beta} = Y^TY - 2Y^TX\widehat{\beta} + \widehat{\beta}^TX^TX\widehat{\beta} + \lambda^TC\widehat{\beta}$$
$$\begin{cases} \frac{\partial f}{\partial \widehat{\beta}} = -2X^TY + 2X^TX\widehat{\beta} + C^T\lambda = 0 & (1) \\ \frac{\partial f}{\partial \lambda} = C\widehat{\beta} = 0 & (2) \end{cases}$$
$$(1) \times C(X^TX)^{-1} \Rightarrow \lambda = [C(X^TX)^{-1}C^T]^{-1}2C(X^TX)^{-1}X^TY$$
$$\therefore \widehat{\beta} = (X^TX)^{-1}X^TY - (X^TX)^{-1}C^T[C(X^TX)^{-1}C^T]^{-1}C(X^TX)^{-1}X^TY = b - Ab$$

where $A = (X^TX)^{-1}C^T[C(X^TX)^{-1}C^T]^{-1}C$. Therefore, in the reduced model,

$$\begin{aligned} SSE_R &= \|Y - X\widehat{\beta}\|^2 = \|Y - Xb + XAb\|^2 \\ &= \|Y - Xb\|^2 + \|XAb\|^2 + 2(Y - Xb)^TXAb \\ &= SSE_F + b^TA^TX^TXAb + 2Y^T(I - H)^TXAb \\ &= SSE_F + b^TA^TX^TXAb + 2Y^T(I - H)XAb \quad (\because (I - H)X = 0) \\ &= SSE_F + b^TA^TX^TXAb \end{aligned}$$

$$\begin{aligned} &\therefore\ SSE_R - SSE_F \\ &=\ b^TA^TX^TXAb \\ &=\ b^TC^T[C(X^TX)^{-1}C^T]^{-1}C(X^TX)^{-1}X^TX(X^TX)^{-1}C^T[C(X^TX)^{-1}C^T]^{-1}Cb \\ &=\ b^TC^T[C(X^TX)^{-1}C^T]^{-1}C(X^TX)^{-1}C^T[C(X^TX)^{-1}C^T]^{-1}Cb \\ &=\ b^TC^T[C(X^TX)^{-1}C^T]^{-1}Cb \\ &=\ \|\Sigma^{-1/2}Cb\|^2 \end{aligned}$$

where $\Sigma = C(X^TX)^{-1}C^T$. Since $Cb \sim N(C\beta, C\sigma^2(X^TX)^{-1}C^T) \overset{H_0}{=} N(0, \sigma^2\Sigma)$, then

$$\Sigma^{-1/2}Cb \sim N(0, \Sigma^{-1/2}\sigma^2\Sigma\Sigma^{-1/2}) = N(0, \sigma^2I_m)$$

Hence, $SSE_R - SSE_F = \|\Sigma^{-1/2}Cb\|^2$ can be seen as the sum of m independent $N^2(0,1)\sigma^2$. That is to say,

$$b^TC^T[C(X^TX)^{-1}C^T]^{-1}Cb = \|\Sigma^{-1/2}Cb\|^2 \sim \sigma^2\chi^2(m)$$

And because $MSE_F \sim \sigma^2\chi^2(n-k-1)$, $F = \frac{SSE_R - SSE_F/m}{SSE_F/n-k-1} \sim F_{m,n-k-1}$, if the numerator and the denominator are independent. Since $SSE_R - SSE_F = \|\Sigma^{-1/2}Cb\|^2 = \|DY\|^2$, where $D = \Sigma^{-1/2}C(X^TX)^{-1}X^T$, and $SSE = \|(I-H)Y\|^2$, then $\begin{pmatrix} DY \\ (I-H)Y \end{pmatrix} \sim Gaussian$. Thus, we just need to prove $\mathbb{C}\text{ov}(DY, (I-H)Y) = 0$. Because, $(I-H)X = 0$,

$$\mathbb{C}\text{ov}(DY, (I-H)Y) = D\sigma^2 I_n (I-H)^T = \sigma^2\Sigma^{-1/2}C(X^TX)^{-1}X^T(I-H)^T = 0$$

Therefore, to sum up, the task is that we want to test a null hypothesis $H_0: C\beta = 0$, where C is a $m \times (k+1)$ dimensional matrix, and the solution is that we establish a statistic

$$F = \frac{SSE_R - SSE_F/m}{SSE_F/n-k-1} = \frac{b^TC^T[C(X^TX)^{-1}C^T]^{-1}Cb/m}{Y^T(I-H)Y/n-k-1} \sim F_{m,n-k-1}$$

and if $F > F_{m,n-k-1}(\alpha)$, we reject H_0. In fact, under certain condition, this task will become some specific hypothesis testing.

Eg. 1: Suppose $C = (0, \cdots, 0, 1, 0, \cdots, 0)$, where $C_{i+1} = 1$, then $H_0: C\beta = \beta_i = 0$, and

$$F = \frac{SSE_R - SSE_F/1}{SSE_F/n-k-1} = \frac{b^TC^T[C(X^TX)^{-1}C^T]^{-1}Cb}{MSE_F} = \frac{b_i^2}{s^2(X^TX)^{-1}_{[i+1,i+1]}} = t^2$$

where t is exactly the statistic when we previously tested on a single X_i.

Eg. 2: Suppose $C_{2\times(k+1)} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \end{pmatrix}$, then $H_0: \beta_1 = \beta_2 = 0$.

1.2.6 Model Selection

Now that we have a model $Y \sim (X_1, \cdots, X_k)$, we don't want the k being too big, that is we want to use less amount of X_i that can still explain Y well.

1.2.6.1 All Possible Regression Selection

Define $X = \{X_1, \cdots, X_k\}$ as all possible independent variables and $D \subset X$. We want to find a D with few variables but high fitting performance. There are several criteria that can assess the fitting performance, namely R^2, R_a^2, C_p, AIC.

(1). $R^2 = \frac{SSR}{SST}$. However, since the more variables the higher R^2 will be, we usually first select one X_i from $\{X_1, \cdots, X_k\}$ that has the highest R^2. For example, we select X_1 and then, we select one of $\{(X_1, X_2), \cdots, (X_1, X_k)\}$ that has the highest R^2, and vaguely observe if there is a large increase in R^2. As we can see, it is not a strict method.

(2). $R_a^2 = 1 - \frac{MSE}{MST}$. This time, we can simply find a $D \subset \{X_1, \cdots, X_k\}$ that maximize $R_a^2(D)$.

(3). $C_p = \frac{SSE(p)}{s^2} - [n - 2(p+1)]$, where $p = |D|,\ D \subset X$, $SSE(p)$ is the SSE when D is the set of independent variables, and s^2 is the MSE of the full model. Without loss of generality, suppose $D = \{X_1, \cdots, X_p\}$, and the reduced model will be $Y = \beta_0 + \beta_1X_1 + \cdots + \beta_pX_p + \epsilon$. Next, we are going to prove that C_p can be used to estimate $\frac{MSE(\widehat{Y})}{\sigma^2}$. Note that, here MSE is not SSE/df_E, but the mean square error, $MSE(\hat{\theta}) = \mathbb{E}\|\hat{\theta} - \theta\|_2^2$. Suppose $\widehat{Y} = \begin{pmatrix} \widehat{Y}_1 \\ \vdots \\ \widehat{Y}_n \end{pmatrix}$ and $\mu = \mathbb{E}(Y|X)$, then,

$$MSE(\widehat{Y}) = \mathbb{E}\|\widehat{Y} - \mu\|_2^2 = \mathbb{E}(\widehat{Y} - \mu)^T(\widehat{Y} - \mu) = \mathbb{E}(\widehat{Y}^T\widehat{Y} - 2\widehat{Y}^T\mu + \mu^T\mu)$$

Since $\widehat{Y} = H_p Y$, where $H_p = X(X^TX)^{-1}X^T$, $X = (1_n, X_1, \cdots, X_p)$, then,

$$\begin{aligned} MSE(\widehat{Y}) &= \mathbb{E}(Y^T H_p Y) - 2\mathbb{E}(Y^T H_p^T \mu) + \mu^T\mu \\ &= \mathbb{E}(trH_pYY^T) - 2\mu^T H_p\mu + \mu^T\mu \\ &= trH_p\mathbb{E}(YY^T) - 2\mu^T H_p\mu + \mu^T\mu \end{aligned}$$

Since

$$\begin{aligned} \mathbb{C}\text{ov}(Y) &= \mathbb{E}(Y - \mathbb{E}Y)(Y - \mathbb{E}Y)^T \\ &= \mathbb{E}YY^T - E[Y(\mathbb{E}Y)^T] - E[(\mathbb{E}Y)Y^T] + (\mathbb{E}Y)(\mathbb{E}Y)^T \\ &= \mathbb{E}YY^T - (\mathbb{E}Y)(\mathbb{E}Y)^T \end{aligned}$$

Therefore, $\mathbb{E}(YY^T) = \mathbb{C}\text{ov}(Y) + (\mathbb{E}Y)(\mathbb{E}Y)^T = \sigma^2 I_n + \mu\mu^T$, then,

$$\begin{aligned} MSE(\widehat{Y}) &= trH_p(\sigma^2 I_n + \mu\mu^T) - 2\mu^T H_p\mu + \mu^T\mu \\ &= \sigma^2(p+1) + \mu^T H_p\mu - 2\mu^T H_p\mu + \mu^T\mu \\ &= \sigma^2(p+1) - \mu^T H_p\mu + \mu^T\mu \end{aligned}$$

Since $SSE(p) = Y^T(I - H_p)Y$, then

$$\begin{aligned} &\mathbb{E}SSE(p) = tr(I - H_p)(\sigma^2 I_n + \mu\mu^T) = \sigma^2(n - p - 1) + \mu^T(I - H_p)\mu \\ \therefore\ &MSE(\widehat{Y}) = \sigma^2(p+1) + \mathbb{E}SSE(p) - \sigma^2(n - p - 1) = \mathbb{E}SSE(p) - \sigma^2[n - 2(p+1)] \\ \therefore\ &\frac{MSE(\widehat{Y})}{\sigma^2} = \frac{\mathbb{E}SSE(p)}{\sigma^2} - [n - 2(p+1)] \leftarrow \frac{SSE(p)}{s^2} - [n - 2(p+1)] = C_p \end{aligned}$$

Therefore, C_p can be used to estimate $\frac{MSE(\widehat{Y})}{\sigma^2}$, and we want C_p as small as possible. And since

$$\begin{aligned} MSE(\widehat{Y}) &= \mathbb{E}\|\widehat{Y} - \mu\|_2^2 = \mathbb{V}\text{ar}(\widehat{Y}) + \text{bias}^2(\widehat{Y}) \geq \mathbb{V}\text{ar}(\widehat{Y}) = \mathbb{V}\text{ar}(HY) \\ &= tr(\mathbb{C}\text{ov}(HY)) = tr(H\sigma^2 I_n H^T) = \sigma^2 tr(HH^T) = \sigma^2 tr(H) \\ &= \sigma^2(p+1) \\ \therefore\ &\frac{MSE(\widehat{Y})}{\sigma^2} \geq p + 1 \end{aligned}$$

That is, we want C_p as close to $p + 1$ as possible.

(4). $AIC = ln_p - p$, where ln_p is the maximum of log-likelihood in the model $Y \sim \{X_1, \cdots, X_p\}$, and we are going to select a model that maximize the AIC Statistic. Suppose $Y = X_p\beta_p + \epsilon$,

where $X_p = \{1_n, X_1, \cdots, X_p\}$, $\beta_p = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$. The joint pdf is

$$pdf = \frac{exp\left\{-\frac{1}{2}(Y - X_p\beta_p)^T(\sigma^2 I_n)^{-1}(Y - X_p\beta_p)\right\}}{\sqrt{(2\pi)^n|\sigma^2 I_n|}}$$

$$\therefore loglikelihood = \frac{-\frac{1}{2}(Y - X_p\beta_p)^T(Y - X_p\beta_p)}{\sigma^2} - \frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2)$$

$$\therefore \widehat{\beta}_{p\,MLE} = (X_p^TX_p)^{-1}X_p^TY, \quad \widehat{\sigma}_{p\,MLE}^2 = \frac{1}{n}\|Y - X_p\widehat{\beta}_{p\,MLE}\|^2$$

Therefore, the maximum of log-likelihood is

$$\begin{aligned} ln_p &= \frac{-\frac{1}{2}(Y - X_p\widehat{\beta}_{p\,MLE})^T(Y - X_p\widehat{\beta}_{p\,MLE})}{\frac{1}{n}\|Y - X_p\widehat{\beta}_{p\,MLE}\|^2} - \frac{n}{2}log(2\pi) - \frac{n}{2}log(\frac{1}{n}\|Y - X_p\widehat{\beta}_{p\,MLE}\|^2) \\ &= -\frac{n}{2}logSSE(p) + f(n) \end{aligned}$$

Hence, we can write $AIC = -\frac{n}{2}logSSE(p) - p$, and the problem becomes to find the D that maximizes $AIC(D)$.

Note that there is a different AIC expression widely found on the Internet, that is $AIC = 2p - 2ln(L)$, and the task is to minimize AIC, but since there are just differences between the sign, the total coefficient and the constant number, basically the two expressions work the same.

1.2.6.2 Sequential Selection Method

Since $X = \{X_1, \cdots, X_k\}$ has 2^k different subsets, when k goes very large, it would be very bothering to go over all the subsets. So, we are going to address this problem through sequential selection method.

Backward Selection

Step 1: Build the full model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$. Then, for $i = 1, \cdots, k$, we set a null hypothesis $H_0 : \beta_i = 0$, that is, we build a reduced model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{i-1} X_{i-1} + \beta_{i+1} X_{i+1} + \cdots + \beta_k X_k + \epsilon$, and we can derive a statistic, $F = \frac{SSE_R - SSE_F/1}{SSE_F/n-k-1} \sim F_{1,n-k-1}$. Thus, we will get k statistics, $F_1^{(1)}, \cdots, F_k^{(1)}$. Without loss of generality, we assume that $F_k^{(1)}$ is the smallest. If $F_k^{(1)} > F_{1,n-k-1}(\alpha)$, we reject H_0 since X_k has certain explanatory effect. Else, we drop X_k from the model.

Step 2: Build the full model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + \epsilon$. Then, for $i = 1, \cdots, k-1$, we set a null hypothesis $H_0 : \beta_i = 0$, that is, we build a reduced model, and we can derive a statistic, $F = \frac{SSE_R - SSE_F/1}{SSE_F/n-k} \sim F_{1,n-k}$. Thus, we will get $k-1$ statistics, $F_1^{(2)}, \cdots, F_{k-1}^{(2)}$. Without loss of generality, we assume that $F_{k-1}^{(2)}$ is the smallest. If $F_{k-1}^{(2)} > F_{1,n-k}(\alpha)$, we reject H_0 since X_{k-1} has certain explanatory effect. Else, we drop X_{k-1} from the model.

Similarly, we step until if $F_{k-i}^{(i+1)} > F_{1,n-k-1+i}$, then we choose $X_1, \cdots, X_{k-i}$ as the final choice.

Forward Selection

Step 1: Build the reduced model $Y = \beta_0 + \epsilon$. Then, for $i = 1, \cdots, k$, we set a null hypothesis $H_0 : \beta_i = 0$, that is, we build a full model $Y = \beta_0 + \beta_i X_i + \epsilon$, and we can derive a statistic, $F = \frac{SSE_R - SSE_F/1}{SSE_F/n-2} \sim F_{1,n-2}$. Thus, we will get k statistics, $F_1^{(1)}, \cdots, F_k^{(1)}$. Without loss of generality, we assume that $F_1^{(1)}$ is the largest. If $F_1^{(1)} > F_{1,n-2}(\alpha)$, we reject H_0 and add X_1 into the model. Else, we stop adding variables.

Step 2: Build the reduced model $Y = \beta_0 + \beta_1 X_1 + \epsilon$. Then, for $i = 2, \cdots, k$, we set a null hypothesis $H_0 : \beta_i = 0$, that is, we build a full model $Y = \beta_0 + \beta_1 X_1 + \beta_i X_i + \epsilon$, and we can derive a statistic, $F = \frac{SSE_R - SSE_F/1}{SSE_F/n-3} \sim F_{1,n-3}$. Thus, we will get $k-1$ statistics, $F_1^{(2)}, \cdots, F_{k-1}^{(2)}$. Without loss of generality, we assume that $F_1^{(1)}$ is the largest. If $F_1^{(2)} > F_{1,n-3}(\alpha)$, we reject H_0 and add X_2 into the model. Else, we stop adding variables.

Similarly, we step until if $F_1^{(i)} < F_{1,n-i-1}$, then we choose $X_1, \cdots, X_{i-1}$ as the final choice.

Stepwize Selection

Since there are problems about both backward selection and forward selection, for example, in forward selection, once a variable is added into the model, it will always be in the model. Thus, we can use stepwize selection to solve this problem.

1.2.7 Model Diagnostic Checking

Previously, we made 2 assumption on X and ϵ in the model $Y = X\beta + \epsilon$.

1. $\mathbb{V}\text{ar}\epsilon_i = \sigma^2$, which is homogeneous and uncorrelated, that is, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.
2. The columns of $X_{n\times(k+1)}$, $\{1_n, X_1, \cdots, X_k\}$, are linearly independent.

1.2.7.1 assumption of ϵ

For the assumption of ϵ, if it is not correct, we can use $e = Y - \widehat{Y}$ to estimate ϵ, and e has the following properties.

1. $e = (I-H)Y \sim N(0, (I-H)\sigma^2 I_n (I-H)) = N(0, \sigma^2(I-H))$
2. $e \perp \widehat{Y}$ $(\because \mathbb{C}\text{ov}(e, \widehat{Y}) = (I-H)\sigma^2 I_n H^T = \sigma^2(H - H^2) = 0)$
3. $1_n^T e = 1_n^T (I-H)Y = 0$ $(\because (I-H)1_n = 0)$

Case 1: if $\{\epsilon_i\}$ are not homogeneous

In one case, $Y_i = X_i^T\beta + \epsilon_i$, where $\mathbb{V}\text{ar}\epsilon_i = \sigma_i^2$, consider the easy condition of this case, that is, $\{\sigma_i\}$ are known. Define $Y_i^* = Y_i/\sigma_i$, $X_i^* = X_i/\sigma_i$, $\epsilon_i^* = \epsilon_i/\sigma_i$, then $Y_i^* = {X_i^*}^T\beta + \epsilon_i^*$, where $\mathbb{V}\text{ar}\epsilon_i^* = 1$. Consider the new model, we can get

$$
\begin{aligned}
\widehat{\beta}_{LSE} &= argmin_{\widehat{\beta}} \sum_{i=1}^{n} \left(Y_i^* - \widehat{Y}_i^*\right)^2 = argmin_{\widehat{\beta}} \sum_{i=1}^{n} \left(\frac{Y_i}{\sigma_i} - \frac{X_i^T}{\sigma_i}\widehat{\beta}\right)^2 \\
&= argmin_{\widehat{\beta}} \begin{pmatrix} \frac{Y_1}{\sigma_1} - \frac{X_1^T}{\sigma_1}\widehat{\beta} \\ \vdots \\ \frac{Y_n}{\sigma_n} - \frac{X_n^T}{\sigma_n}\widehat{\beta} \end{pmatrix}^T \begin{pmatrix} \frac{Y_1}{\sigma_1} - \frac{X_1^T}{\sigma_1}\widehat{\beta} \\ \vdots \\ \frac{Y_n}{\sigma_n} - \frac{X_n^T}{\sigma_n}\widehat{\beta} \end{pmatrix} \\
&= argmin_{\widehat{\beta}} \begin{pmatrix} Y_1 - X_1\widehat{\beta} \\ \vdots \\ Y_n - X_n\widehat{\beta} \end{pmatrix}^T \begin{pmatrix} \frac{1}{\sigma_1} & & \\ & \ddots & \\ & & \frac{1}{\sigma_n} \end{pmatrix}^T \begin{pmatrix} \frac{1}{\sigma_1} & & \\ & \ddots & \\ & & \frac{1}{\sigma_n} \end{pmatrix} \begin{pmatrix} Y_1 - X_1\widehat{\beta} \\ \vdots \\ Y_n - X_n\widehat{\beta} \end{pmatrix}
\end{aligned}
$$

Define $W = \begin{pmatrix} \frac{1}{\sigma_1} & & \\ & \ddots & \\ & & \frac{1}{\sigma_n} \end{pmatrix}$, and $A = W^TW = \begin{pmatrix} \frac{1}{\sigma_1^2} & & \\ & \ddots & \\ & & \frac{1}{\sigma_n^2} \end{pmatrix}$, then

$$
\begin{aligned}
\widehat{\beta}_{LSE} &= argmin_{\widehat{\beta}} (Y - X\widehat{\beta})^T W^T W (Y - X\widehat{\beta}) \\
&= argmin_{\widehat{\beta}} (Y - X\widehat{\beta})^T A (Y - X\widehat{\beta}) \\
&= argmin_{\widehat{\beta}} (Y^TA - \widehat{\beta}^T X^T A)(Y - X\widehat{\beta}) \\
&= argmin_{\widehat{\beta}} -2\widehat{\beta}^T X^T AY + \widehat{\beta}^T X^T AX\widehat{\beta}
\end{aligned}
$$

Find the derivative, that $2X^TAY + 2X^TAX\widehat{\beta} = 0$. Therefore, the weighted least square estimator of β is

$$\widehat{\beta}_{LSE} = (X^TAX)^{-1}X^TAY$$

Similarly, the weighted BLUE of β the also the same as th weighted LSE of β. As we define $\widehat{\beta}^* = [(X^TAX)^{-1}X^TA + \Delta]Y$, then since unbiasedness, we have $\Delta X = 0$, and since the minimum variance, we can derive that $\Delta = 0$, so that $\widehat{\beta}^* = (X^TAX)^{-1}X^TAY$.

In another case, we consider a kind of condition, say, $\mathbb{V}\text{ar}\epsilon_i = cx_{i1}^2$. Similarly, we define $Y_i^* = Y_i/x_{i1}$, $X_i^* = X_i/x_{i1}$, $\epsilon_i^* = \epsilon_i/x_{i1}$, then $Y_i^* = {X_i^*}^T\beta + \epsilon_i^*$, where $\mathbb{V}\text{ar}\epsilon_i^* = c$. Now we consider the general case, that is we want to find a mapping, $h(Y_i)$, such that $\mathbb{V}\text{ar}(h(Y_i)) = c$. Hence, with Taylor Formula, the model changes from $Y_i = \mathbb{E}Y_i + \epsilon_i$ to

$$h(Y_i) = h(\mathbb{E}Y_i) + h'(\mathbb{E}Y_i)\epsilon_i + o(\mathbb{E}Y_i)$$

Therefore, $\mathbb{V}\text{ar}(h(Y_i)) = [h'(\mathbb{E}Y_i)]^2\mathbb{V}\text{ar}\epsilon_i = c$, so $h'(\mathbb{E}Y_i) \propto \frac{1}{\sqrt{\mathbb{V}\text{ar}\epsilon_i}}$. That is, we can choose the mapping h according to $\mathbb{V}\text{ar}\epsilon_i$. But what if we don't know the value of $\mathbb{V}\text{ar}\epsilon_i$? This time, we simply use $\widehat{\beta} = (X^TX)^{-1}X^TY$ and see its properties.

$$\mathbb{E}\widehat{\beta} = (X^TX)^{-1}X^TX\beta = \beta$$
$$\mathbb{V}\text{ar}\widehat{\beta} = (X^TX)^{-1}X^T\begin{pmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2\end{pmatrix}X(X^TX)^{-1}$$

When n is large enough,

$$\begin{aligned}\mathbb{V}\text{ar}\widehat{\beta}_i &= \left[(X^TX)^{-1}X^T\begin{pmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2\end{pmatrix}X(X^TX)^{-1}\right]_{(i,i)} \\ &= \sum_{j=1}^{n}\left[(X^TX)^{-1}X^T\right]_{(i,j)}\sigma_j^2\left[X(X^TX)^{-1}\right]_{(j,i)} \\ &\le k\left[(X^TX)^{-1}X^TX(X^TX)^{-1}\right]_{(i,i)} \\ &= k\left[(X^TX)^{-1}\right]_{(i,i)}\end{aligned}$$

where $k = max_{j=1}^{n}\{\sigma_j^2\}$. Suppose $(X^TX)^{-1}$ is diagonal, though it usually isn't, then

$$\mathbb{V}\text{ar}\widehat{\beta}_i \le k\left[(X^TX)^{-1}\right]_{(i,i)} = \frac{k}{[X^TX]_{(i,i)}} = \frac{k}{\sum_{j=1}^{n}X_{ij}^2} \to 0$$

this is because when n goes large enough, $\sum_{j=1}^{n}X_{ij}^2$ will also goes large enough. Therefore, we can simply use $\widehat{\beta} = (X^TX)^{-1}X^TY$.

Case 2: if $\{\epsilon_i\}$ are not uncorrelated

First, how to define whether $\{\epsilon_i\}$ are uncorrelated or not?

For two random variables $\xi, \eta \in \mathbb{R}$, define $\rho = \frac{\mathbb{C}\text{ov}(\xi,\eta)}{\sqrt{\mathbb{V}\text{ar}(\xi)\mathbb{V}\text{ar}(\eta)}} \in [-1, 1]$.

1. If $\rho = 0$, then ξ, η are uncorrelated;
2. If $\rho \in (0, 1]$, then ξ, η are positive correlated;
3. If $\rho \in [-1, 0)$, then ξ, η are negative correlated;

Suppose $\xi_1, \cdots, \xi_n \overset{iid}{\sim} N(0, \sigma_\xi^2),\ \eta_1, \cdots, \eta_n \overset{iid}{\sim} N(0, \sigma_\eta^2)$, then

$$\mathbb{C}\text{ov}(\xi, \eta) = \mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta) \leftarrow \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi})(\eta_i - \bar{\eta}) = \frac{1}{n}\sum_{i=1}^{n}\xi_i\eta_i$$
$$\mathbb{V}\text{ar}(\xi) = \mathbb{E}(\xi - \mathbb{E}\xi)^2 \leftarrow \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2 = \frac{1}{n}\sum_{i=1}^{n}\xi_i^2$$
$$\mathbb{V}\text{ar}(\eta) = \mathbb{E}(\eta - \mathbb{E}\eta)^2 \leftarrow \frac{1}{n}\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2 = \frac{1}{n}\sum_{i=1}^{n}\eta_i^2$$

Therefore, we introduce the first-order coefficient of autocorrelation,

$$\rho = \frac{\mathbb{C}\text{ov}(\epsilon_i, \epsilon_{i-1})}{\sqrt{\mathbb{V}\text{ar}(\epsilon_i)\mathbb{V}\text{ar}(\epsilon_{i-1})}}$$
$$\therefore \rho \leftarrow \frac{\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\epsilon_{i-1}}{\frac{1}{n}\sum_{i=2}^{n}\epsilon_i^2} \leftarrow \frac{\frac{1}{n}\sum_{i=2}^{n}e_ie_{i-1}}{\frac{1}{n}\sum_{i=2}^{n}e_i^2} \triangleq (*)$$

and we define the DW-test statistic,

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=2}^{n} e_t^2} = \frac{2\sum_{t=2}^{n} e_t^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2} = 2 - 2(*)$$

Therefore,

1. If $\{\epsilon_i\}$ are uncorrelated, then $\rho = 0$, $DW = 2$;
2. If $\{\epsilon_i\}$ are positive correlated, then $\rho \in (0, 1]$, $DW \in [0, 2)$;
3. If $\{\epsilon_i\}$ are negative correlated, then $\rho \in [-1, 0)$, $DW =\in (2, 4]$;

AR(1)

Consider the model $\begin{cases} \epsilon_t = \rho\epsilon_{t-1} + v_t \\ |\rho| < 1 \\ Y = X\beta + \epsilon \end{cases}$, where $v_t \sim N(0, \sigma_v^2)$ is the white noise.

If ρ is known, then $y_i = x_i^T\beta + \epsilon_i$, we want to transform ϵ_i to v_i. Since $\rho y_{i-1} = \rho x_{i-1}^T\beta + \rho\epsilon_{i-1}$,

$$\tilde{y} \triangleq y_i - \rho y_{i-1} = (x_i - \rho x_{i-1})^T\beta + v_i \triangleq \tilde{x}_i^T\beta + v_i$$
$$\therefore \widehat{\beta} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\widetilde{Y}$$

If ρ is unknown, there are basically two methods, iteration and differencing.

The first method is to find an estimator of ρ. First, we derive an initial estimator, $\widehat{\beta}_0 = (X^TX)^{-1}X^TY$, it doesn't have to be very accurate. Then, we derive $\widehat{Y} = X\widehat{\beta}_0$ and use $e = Y - \widehat{Y} \to \epsilon$, thus, we can regard $\epsilon_t = \rho\epsilon_{t-1} + v_t$ as a single linear regression with $\alpha = 0$. In this way, we can derive ρ_0, an estimator of ρ, from the model $\begin{pmatrix} e_2 \\ \vdots \\ e_n \end{pmatrix} \sim \begin{pmatrix} e_1 \\ \vdots \\ e_{n-1} \end{pmatrix}$. Then, we can put ρ_0 into the condition above when ρ is known and derive $\widehat{\beta}_1$, and then ρ_1, and so on.

The second method is that we can use differencing when ρ is very big such that $\rho \approx 1$. In this way, $\epsilon_t \approx \epsilon_{t-1} + v_t$, thus $\tilde{y} \triangleq y_t - y_{t-1} = (x_t - x_{t-1})^T\beta + v_t \triangleq \tilde{x}_t^T\beta + v_t$, and $\widehat{\beta} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\widetilde{Y}$.

Also, we can leave the unknown ρ and define $y_t^* = y_t - \rho y_{t-1}$, $x_{ti}^* = x_{ti} - \rho x_{t-1i}$, then

$$\widehat{\beta} = argmin_{\widehat{\beta}} \sum_{t=1}^{n} \left[y_t^* - \beta_0(1-\rho) - \beta_1 x_{t1}^* - \cdots - \beta_k x_{tk}^*\right]^2$$

Let $\rho = \rho_i = \frac{i}{10}$, where $i = 0, 1, \cdots, 10$, and find the minimum of $SSE(\rho_i)$, then choose that ρ_i as the fix value of ρ.

1.2.7.2 assumption of X

For the assumption of X, if it is not correct, that is if the columns of $X = (X_0, X_1, \cdots, X_k)$, where $X_0 = 1_n$, are not linearly independent, there exist $\{c_i\}_{i=0}^k$ such that $\sum_{i=0}^k c_i X_i = 0$, or $\sum_{i=0}^k c_i X_i \approx 0$, which means approximately linearly dependent. Normalize $\{c_i\}_{i=0}^k$ to a unit vector $\{u_i\}_{i=0}^k$, and then consider the eigenvalue decomposition of X^TX, there exist

$$\lambda_i = u_i^T\lambda_i u_i = u_i^T X^T X u_i \approx 0$$

That is, $\lambda_{min}(X^TX) \approx 0$, then $\lambda_{max}[(X^TX)^{-1}] \to \infty$, which will result in huge variance in $\widehat{\beta}$. That's because $\mathbb{V}\text{ar}\widehat{\beta} = \sigma^2 tr(X^TX)^{-1} = \sigma^2\sum_i \lambda_i[(X^TX)^{-1}] \to \infty$. Therefore, we introduce Ridge Regression to solve this problem.

1.2.7.3 Ridge Regression

The idea of Ridge Regression is to add something to each eigenvalue of X^TX, that is,

$$\widehat{\beta}(\lambda) = (X^TX + \lambda I)^{-1}X^TY\ ,\ \ \lambda \geq 0$$

There are two explanations of the Ridge Regression.

On the one hand, $\widehat{\beta}(\lambda)$ is a shrinkage version of b, and we can look at some special conditions.

1. If $X^TX = I$, then $\widehat{\beta}(\lambda) = \frac{1}{\lambda+1}X^TY, b = X^TY$. Hence, $\widehat{\beta}(\lambda)/b < 1$.
2. If $\lambda = 0$, $\widehat{\beta}(\lambda) = b$.
3. If $\lambda \to \infty$, $\widehat{\beta}(\lambda) \to 0$.

As we can see, $\widehat{\beta}(\lambda)$ is a shrinkage version of b, and λ determine the extent of the shrinkage.

On the other hand, $\widehat{\beta}(\lambda)$ can be regarded as adding a L2-norm penalty, that is,

$$\begin{aligned}\widehat{\beta}(\lambda) &= argmin_{\widehat{\beta}}\left\{SSE(\widehat{\beta}) + \lambda\|\widehat{\beta}\|_2^2\right\} \\ &= argmin_{\widehat{\beta}}\left\{(Y - X\widehat{\beta})^T(Y - X\widehat{\beta}) + \lambda\widehat{\beta}^T\widehat{\beta}\right\}\end{aligned}$$

$$\frac{\partial}{\partial\widehat{\beta}} = 0 \Rightarrow -2X^TY + 2X^TX + 2\lambda I\widehat{\beta} = 0 \Rightarrow \widehat{\beta}(\lambda) = (X^TX + \lambda I)^{-1}X^TY$$

Note that if we put a L2-norm penalty, then it is called Ridge Regression. If we put a L1-norm penalty, then it is called Lasso Regression.

Next, we are going to find out the bias and the variance of $\widehat{\beta}(\lambda)$. If $\lambda \neq 0$,

$$\mathbb{E}\widehat{\beta}(\lambda) = (X^TX + \lambda I)^{-1}X^TX\beta \neq \beta$$

Therefore, $\widehat{\beta}(\lambda)$ is biased, and the bias is

$$\mathbb{E}\widehat{\beta}(\lambda) - \beta = (X^TX + \lambda I)^{-1}X^TX\beta - (X^TX + \lambda I)^{-1}(X^TX + \lambda I)\beta = -\lambda(X^TX + \lambda I)^{-1}\beta$$

Similar to the variance of a random variable, which is defined as $\mathbb{V}\text{ar}\xi = \mathbb{E}(\xi - \mathbb{E}\xi)^2$, we have

$$\begin{aligned}\mathbb{V}\text{ar}\widehat{\beta}(\lambda) &= \mathbb{E}\left\|\widehat{\beta}(\lambda) - \mathbb{E}\widehat{\beta}(\lambda)\right\|^2 \\ &= \mathbb{E}\left(\widehat{\beta}(\lambda) - \mathbb{E}\widehat{\beta}(\lambda)\right)^T\left(\widehat{\beta}(\lambda) - \mathbb{E}\widehat{\beta}(\lambda)\right) \\ &= \mathbb{E}tr\left(\widehat{\beta}(\lambda) - \mathbb{E}\widehat{\beta}(\lambda)\right)\left(\widehat{\beta}(\lambda) - \mathbb{E}\widehat{\beta}(\lambda)\right)^T \\ &= tr\mathbb{E}\left(\widehat{\beta}(\lambda) - \mathbb{E}\widehat{\beta}(\lambda)\right)\left(\widehat{\beta}(\lambda) - \mathbb{E}\widehat{\beta}(\lambda)\right)^T \\ &= tr\mathbb{C}\text{ov}\left(\widehat{\beta}(\lambda)\right) \\ &= tr\sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1} \\ &= \sigma^2 tr(X^TX + \lambda I)^{-2}X^TX\end{aligned}$$

Therefore, we can derive the following table.

	$\widehat{\beta}(\lambda)$	b
bias	$-\lambda(X^TX + \lambda I)^{-1}\beta$	0
Cov(·)	$\sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}$	$\sigma^2(X^TX)^{-1}$
Var(·)	$\sigma^2 tr(X^TX + \lambda I)^{-2}X^TX$	$\sigma^2 tr(X^TX)^{-1}$

Since $\lambda \geq 0$, $\mathbb{V}\text{ar}\widehat{\beta}(\lambda) = \sigma^2 tr(X^TX + \lambda I)^{-2}X^TX \leq \sigma^2 tr(X^TX)^{-1} = \mathbb{V}\text{ar}b$. This is because

$$
\begin{aligned}
\mathbb{V}\text{ar}\widehat{\beta}(\lambda) &= \sigma^2 tr(X^TX+\lambda I)^{-2}X^TX \\
&= \sigma^2 tr[U(D+\lambda I)U^T]^{-2}UDU^T \\
&= \sigma^2 trU(D+\lambda I)^{-2}U^TUDU^T \\
&= \sigma^2 trU(D+\lambda I)^{-2}DU^T \\
&= \sigma^2 tr(D+\lambda I)^{-2}D \\
&\leq \sigma^2 trD^{-2}D \\
&= \sigma^2 trUD^{-1}U^T \\
&= \sigma^2 tr(X^TX)^{-1} = \mathbb{V}\text{ar}b
\end{aligned}
$$

As we can see, $\widehat{\beta}(\lambda)$ is biased but it has a smaller variance. And if we look at the MSE of $\widehat{\beta}(\lambda)$,

$$
\begin{aligned}
MSE(\widehat{\beta}) &= \mathbb{E}\|\widehat{\beta}-\beta\|_2^2 = \mathbb{E}(\widehat{\beta}-\beta)^T(\widehat{\beta}-\beta) \\
&= \mathbb{E}(\widehat{\beta}-\mathbb{E}\widehat{\beta})^T(\widehat{\beta}-\mathbb{E}\widehat{\beta}) + (\mathbb{E}\widehat{\beta}-\beta)^T(\mathbb{E}\widehat{\beta}-\beta) \\
&= \mathbb{V}\text{ar}\widehat{\beta}(\lambda) + \|\text{bias}\widehat{\beta}(\lambda)\|_2^2 \\
&= \sigma^2 tr(X^TX+\lambda I)^{-2}X^TX + \lambda^2\beta^T(X^TX+\lambda I)^{-2}\beta \\
&= \sigma^2 tr(X^TX+\lambda I)^{-2}X^TX + \lambda^2 tr(X^TX+\lambda I)^{-2}\beta\beta^T \\
&= tr(X^TX+\lambda I)^{-2}\left[\sigma^2X^TX+\lambda^2\beta\beta^T\right]
\end{aligned}
$$

We want to know if there is a λ such that $MSE\left(\widehat{\beta}(\lambda)\right)$ is smaller than $MSE\,(b)$. Since

$$
\begin{aligned}
\frac{\partial}{\partial\lambda}MSE\left(\widehat{\beta}(\lambda)\right) &= -2tr(X^TX+\lambda I)^{-3}\left[\sigma^2X^TX+\lambda^2\beta\beta^T\right] + tr(X^TX+\lambda I)^{-2}\left[2\lambda\beta\beta^T\right] \\
\therefore \frac{\partial}{\partial\lambda}MSE\left(\widehat{\beta}(\lambda)\right)\Big|_{\lambda=0} &= -2tr(X^TX)^{-3}\sigma^2X^TX = -2\sigma^2tr(X^TX)^{-2} < 0
\end{aligned}
$$

Hence, there exists a λ such that $MSE\left(\widehat{\beta}(\lambda)\right)$ is smaller than $MSE\,(b)$.

And we can extend $\widehat{\beta}(\lambda) = (X^TX+\lambda I)^{-1}X^TY$ to other forms. Since the model now is

$$
X\widehat{\beta}(\lambda) = X(X^TX+\lambda I)^{-1}X^TY
$$

and since

$$
\begin{aligned}
(X^TX+\lambda I)X^T &= X^TXX^T+\lambda X^T = X^T(XX^T+\lambda I) \\
\therefore X^T(XX^T+\lambda I)^{-1} &= (X^TX+\lambda I)^{-1}(X^TX+\lambda I)X^T(XX^T+\lambda I)^{-1} \\
&= (X^TX+\lambda I)^{-1}X^T(XX^T+\lambda I)(XX^T+\lambda I)^{-1} = (X^TX+\lambda I)^{-1}X^T
\end{aligned}
$$

Hence, the model will become

$$
X\widehat{\beta}(\lambda) = XX^T(XX^T+\lambda I)^{-1}Y
$$

which only depends on XX^T. And we can regard XX^T as a kind of distance, that is,

$$
XX^T_{(i,j)} = x_i^Tx_j = \langle x_i, x_j\rangle \triangleq k(x_i,x_j)
$$

Define $X^TX = K$, which is called kernel. If we change to another kind of distance, then it is called Kernel Ridge Regression and the model becomes

$$
X\widehat{\beta}(\lambda) = K(K+\lambda I)^{-1}Y
$$

Usually, there are two kinds of kernels, one is Gaussian, that is,

$$
k(x_i,x_j) = exp\{-\alpha\|x_i-x_j\|^2\}
$$

And the other is Polynomial, that is,

$$k(x_i, x_j) = \langle x_i, x_j \rangle^d$$

1.2.7.4 Cross Validation

Now that we know how to calculate $\widehat{\beta}(\lambda)$ if λ is given, but how do we choose a good λ? Suppose there are n observations, the so-called training set, $(x_1^T, y_1), \cdots, (x_n^T, y_n) \sim (P_x, P_y)$, with which we derive the model $X\widehat{\beta}(\lambda)$ to fit another $m - n$ observations, the so-called testing set, $(x_{n+1}^T, y_{n+1}), \cdots, (x_m^T, y_m) \sim (P_x, P_y)$. We have two methods to carry out cross validation.

leave one out

Suppose there are n observations, $(x_1^T, y_1), \cdots, (x_n^T, y_n)$. Each time we use one observation as the testing set, that is, for $i = 1, \cdots, n$, set the training set as $(x_j^T, y_j)_{j \neq i}$ and the testing set as (x_i^T, y_i). For a fixed λ, $\widehat{\beta}_{(i)}(\lambda) = (X^T X + \lambda I)^{-1} X^T Y$, where $X = X^{(i)} = \begin{pmatrix} x_1^T \\ \vdots \\ x_{i-1}^T \\ x_{i+1}^T \\ \vdots \\ x_n^T \end{pmatrix}$, and $Y = Y^{(i)} = \begin{pmatrix} y_1 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_n \end{pmatrix}$. Then, the task becomes that for $\lambda \in [\lambda_0, \cdots, \lambda_m]$, we are going to find the minimum of

$$CV^{(1)}(\lambda) \triangleq \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \widehat{\beta}_{(i)}(\lambda))^2$$

K-fold CV

Suppose there are n observations, $(x_1^T, y_1), \cdots, (x_n^T, y_n)$. We separate them to k groups, that is,

$$\{1, \cdots, n\} = \cup_{i=1}^{k} c_i, \ \ |c_i| = \frac{n}{k}$$

Therefore, for $i = 1, \cdots, k$, set the training set as $(x_t^T, y_t)_{t \notin c_i}$ and the testing set as $(x_s^T, y_s)_{s \in c_i}$. For a fixed λ, $\widehat{\beta}_{(i)}(\lambda) = (X^T X + \lambda I)^{-1} X^T Y$, where $X = X^{(i)} = (x_t^T)_{t \notin c_i}$, and $Y = Y^{(i)} = (y_t)_{t \notin c_i}$. Then, the task becomes that for $\lambda \in [\lambda_0, \cdots, \lambda_m]$, we are going to find the minimum of

$$CV^{(k)}(\lambda) \triangleq \frac{1}{k} \sum_{i=1}^{k} \frac{\sum_{s \in c_i} (y_s - x_s^T \widehat{\beta}_{(i)}(\lambda))^2}{|c_i|}$$

1.2.8 Box-Cox Transformation

If we have already known that the relationship between X and Y is non-linear, we are going to use Box-Cox transformation to deal with this condition. The basic form of Box-Cox transformation is

$$Y(\lambda) = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ lnY & \lambda = 0 \end{cases},$$

such that $Y(\lambda) = X\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2 I_n)$. The reason of the form under $\lambda \neq 0$ is to make $Y(\lambda)$ continuous at $\lambda = 0$, that is,

$$\lim_{\lambda \to 0} Y(\lambda) = \lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} \overset{(1)}{=} \lim_{\lambda \to 0} \frac{y^\lambda lny}{1} = lny,$$

in which the equal sign at (1) is according to L'Hospital's Rule. Now that $Y_i(\lambda) \sim N(X_i^T \beta, \sigma^2)$, define

$$f(Y_i) = Y_i(\lambda) = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ lny_i & \lambda = 0 \end{cases}$$

Then, we can derive the pdf of Y_i as

$$\begin{aligned} \frac{\partial}{\partial y_i} P(Y_i \leq y_i) &= \frac{\partial}{\partial y_i} P(f(Y_i) \leq f(y_i)) = \frac{\partial}{\partial y_i} P(Y_i(\lambda)) \leq f(y_i)) \\ &= \frac{\partial}{\partial y_i} \int_{-\infty}^{f(y_i)} \text{pdf of } N(X_i^T \beta, \sigma^2)\, dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{ -\frac{(f(y_i) - X_i^T \beta)^2}{2\sigma^2} \right\} \frac{\partial f(y_i)}{\partial y_i} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{ -\frac{(Y_i(\lambda) - X_i^T \beta)^2}{2\sigma^2} \right\} y_i^{\lambda - 1} \end{aligned}$$

Thus, the joint pdf of $(Y_1, \cdots, Y_n)$ will be

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n exp\left\{ -\sum_{i=1}^n \frac{(Y_i(\lambda) - X_i^T \beta)^2}{2\sigma^2} \right\} \prod_{i=1}^n y_i^{\lambda - 1}$$

Thus, the log-likelihood function will be

$$\begin{gathered} L(\beta, \sigma^2, \lambda) = -\sum_{i=1}^n \frac{(Y_i(\lambda) - X_i^T \beta)^2}{2\sigma^2} - \frac{n}{2} log 2\pi\sigma^2 + log \prod_{i=1}^n y_i^{\lambda - 1} \\ \therefore (\widehat{\beta}, \widehat{\sigma}^2, \widehat{\lambda})_{MLE} = argmax_{(\widehat{\beta}, \widehat{\sigma}^2, \widehat{\lambda})} L(\widehat{\beta}, \widehat{\sigma}^2, \widehat{\lambda}) \end{gathered}$$

Suppose that we fix $\widehat{\lambda}$ first, then the form of $(\widehat{\beta}, \widehat{\sigma}^2)_{MLE}$ is the same as that in linear regression, except that Y is replaced by $Y(\lambda)$. Therefore,

$$\begin{aligned} &\widehat{\beta}_{MLE} = (X^T X)^{-1} X^T Y(\lambda) \\ &\widehat{\sigma}^2_{MLE} = \frac{1}{n} Y(\lambda)^T (I - H) Y(\lambda), \quad H = X(X^T X)^{-1} X^T \end{aligned}$$

Then we use the expression of $\widehat{\beta}_{MLE}$ and $\widehat{\sigma}^2_{MLE}$ to find $\widehat{\lambda}_{MLE}$, that is,

$$
\begin{aligned}
\widehat{\lambda}_{MLE} &= argmax_\lambda L(\widehat{\beta}_{MLE}, \widehat{\sigma}^2_{MLE}, \lambda) \\
&= argmax_\lambda -\frac{\|Y(\lambda) - X\widehat{\beta}_{MLE}\|^2}{2\widehat{\sigma}^2_{MLE}} - \frac{n}{2} log 2\pi\widehat{\sigma}^2_{MLE} + log\prod_{i=1}^{n} y_i^{\lambda-1} \\
&= argmax_\lambda -\frac{Y(\lambda)^T(I-H)Y(\lambda)}{\frac{2}{n}Y(\lambda)^T(I-H)Y(\lambda)} - \frac{n}{2} log\widehat{\sigma}^2_{MLE} + log\prod_{i=1}^{n} y_i^{\lambda-1} \\
&= argmax_\lambda -\frac{n}{2} log\widehat{\sigma}^2_{MLE} + log\prod_{i=1}^{n} y_i^{\lambda-1} \\
&= argmax_\lambda -\frac{n}{2} log\frac{1}{n}Y(\lambda)^T(I-H)Y(\lambda) + \frac{n}{2} log\prod_{i=1}^{n}(y_i^{\lambda-1})^{\frac{2}{n}} \\
&= argmax_\lambda -\frac{n}{2} log\frac{1}{n}\frac{Y(\lambda)^T(I-H)Y(\lambda)}{\prod_{i=1}^{n} y_i^{2\frac{\lambda-1}{n}}}
\end{aligned}
$$

Define $Z(\lambda) = \frac{Y(\lambda)}{\prod_{i=1}^{n} y_i^{\frac{\lambda-1}{n}}}$, then $Z_i(\lambda) = \frac{Y_i(\lambda)}{\prod_{j=1}^{n} y_j^{\frac{\lambda-1}{n}}}$. Hence,

$$
\begin{aligned}
\widehat{\lambda}_{MLE} &= argmax_\lambda -\frac{n}{2} log\frac{1}{n}Z(\lambda)^T(I-H)Z(\lambda) \\
&= argmax_\lambda -\frac{n}{2} log\frac{1}{n}SSE(Z(\lambda)) \\
&= argmin_\lambda SSE(Z(\lambda))
\end{aligned}
$$

To sum up, the whole process of Box-Cox Transformation is that for a series of different λ, we are going to calculate $Z(\lambda)$ and regard $Z(\lambda)$ as response variable, and then find the minimum of $SSE(Z(\lambda))$ and the corresponding $\widehat{\lambda}$, $\widehat{\beta}$ and $\widehat{\sigma}^2$.

2 Classification

In this part, we are going to focus on classification problems, that is, the response Y is no longer a continuous variable, instead, it becomes a discreet variable. And we are going to introduce two models to deal with this kind of problems, Logistic Regression and Possion Regression.

2.1 Logistic Regression

For the i-th observation of the response Y, $y_i = \begin{cases} 1 & p_i \\ 0 & 1-p_i \end{cases}$, where p_i is related to the i-th observation of k explanatory variables $x_{i1}, \cdots, x_{ik}$. In Linear Regression, the expectation of Y_i is

$$\mathbb{E}(Y_i|X_i) = X_i^T\beta \in (-\infty, +\infty)$$

However, in Logistic Regression, the expectation of Y_i is

$$\mathbb{E}(Y_i|X_i) = P(Y_i = 1) = p_i \in (0, 1)$$

Therefore, we need a function that can map $(0, 1)$ to $(-\infty, +\infty)$, and thus we introduce logit function and its inverse function, logistic function or sigmoid function,

$$
\begin{aligned}
logit(x) = \log\frac{x}{1-x} \in (-\infty, +\infty), \quad x \in (0, 1) \\
\sigma(x) = \frac{e^x}{1+e^x} \in (0, 1), \quad x \in (-\infty, +\infty) \\
\therefore p_i \underset{\sigma}{\overset{logit}{\leftrightarrows}} \log\frac{p_i}{1-p_i} = X_i^T\beta
\end{aligned}
$$

We first introduce two properties of the logistic function $\sigma(x)$.

$$(1).\ \sigma(-x)=\frac{e^{-x}}{1+e^{-x}}=\frac{1}{e^{x}+1}=1-\sigma(x)$$
$$(2).\ \sigma'(x)=\frac{e^{x}(1+e^{x})-e^{x}e^{x}}{(1+e^{x})^{2}}=\frac{e^{x}}{1+e^{x}}\frac{1}{1+e^{x}}=\sigma(x)\sigma(-x)=\sigma(x)(1-\sigma(x))$$

Then, we construct the model that $\{y_i\}_{i=1}^n$ are independent, and $y_i=\begin{cases}1 & p_i\\ 0 & 1-p_i\end{cases}$, where $p_i=\sigma(X_i^T\beta)$, in which $X_i=\begin{pmatrix}X_{i1}\\ \vdots\\ X_{ik}\end{pmatrix}$, $\beta=\begin{pmatrix}\beta_1\\ \vdots\\ \beta_k\end{pmatrix}$. Then, the pmf of y_i is $p_i^{y_i}(1-p_i)^{1-y_i}$, and the joint pdf of Y is $\prod_{i=1}^n p_i^{y_i}(1-p_i)^{1-y_i}$. Therefore, the log-likelihood function will be

$$\begin{aligned}l(\beta)&=\sum_{i=1}^{n}y_i\log p_i+(1-y_i)\log(1-p_i)\\&=\sum_{i=1}^{n}y_i\log\frac{p_i}{1-p_i}+\log(1-p_i)\\&=\sum_{i=1}^{n}y_iX_i^T\beta+\log(1-\sigma(X_i^T\beta))\end{aligned}$$

Consider $-l(\widehat{\beta})=\sum_{i=1}^n -y_i\log p_i-(1-y_i)\log(1-p_i)\overset{\Delta}{=}\sum_{i=1}^n d(y_i,\hat{y}_i)$, where $d(y_i,\hat{y}_i)$ is regarded as a kind of residual.

When $y_i=1$, $d(y_i,\hat{y}_i)=-\log p_i$.

1. If $p_i=1$, that is, $\hat{y}_i=1$, so the residual $d(y_i,\hat{y}_i)=0$, the estimation is good.
2. If $p_i=0$, that is, $\hat{y}_i=0$, so the residual $d(y_i,\hat{y}_i)=\infty$, the estimation is bad.

When $y_i=0$, $d(y_i,\hat{y}_i)=-\log(1-p_i)$.

1. If $1-p_i=1$, that is, $p_i=\hat{y}_i=0$, so the residual $d(y_i,\hat{y}_i)=0$, the estimation is good.
2. If $1-p_i=0$, that is, $p_i=\hat{y}_i=1$, so the residual $d(y_i,\hat{y}_i)=\infty$, the estimation is bad.

To sum up, the MLE of β, that is,

$$\widehat{\beta}_{MLE}=argmax_{\widehat{\beta}}\,l(\widehat{\beta})=argmin_{\widehat{\beta}}\,-l(\widehat{\beta})$$

can also be regarded as finding when the sum of residual reaches the minimum. Similarly, we want to use the derivatives to find $\widehat{\beta}_{MLE}$, that is,

$$\begin{aligned}l'(\widehat{\beta})&=\sum_{i=1}^{n}y_iX_i+\frac{-\sigma'(X_i^T\widehat{\beta})}{1-\sigma(X_i^T\widehat{\beta})}X_i\\&=\sum_{i=1}^{n}\left[y_i-\frac{\sigma(X_i^T\widehat{\beta})(1-\sigma(X_i^T\widehat{\beta}))}{1-\sigma(X_i^T\widehat{\beta})}\right]X_i\\&=\sum_{i=1}^{n}\left[y_i-\sigma(X_i^T\widehat{\beta})\right]X_i=0\end{aligned}$$

However, it is not easy to solve the above equation, so we change to another method. With Taylor Formula, we have

$$l(\beta) = l(\beta_0) + l'(\beta_0)^T(\beta - \beta_0) + \frac{1}{2}(\beta - \beta_0)^T l''(\beta_0)(\beta - \beta_0)$$

$$l'(\beta_0)^T = \sum_{i=1}^{n} \left[y_i - \sigma(X_i^T\beta_0)\right] X_i^T = (Y^T - P^T)X$$

$$l''(\beta_0) = \sum_{i=1}^{n} -\sigma'(X_i^T\beta_0)X_iX_i^T = -\sum_{i=1}^{n} \sigma(X_i^T\beta_0)(1 - \sigma(X_i^T\beta_0))X_iX_i^T$$
$$= -\sum_{i=1}^{n} P_i(1 - P_i)X_iX_i^T = -X^TDX$$

where

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, P = \begin{pmatrix} \sigma(X_1^T\beta_0) \\ \vdots \\ \sigma(X_n^T\beta_0) \end{pmatrix} = \begin{pmatrix} P_1 \\ \vdots \\ P_n \end{pmatrix}, D = \begin{pmatrix} P_1(1-P_1) & & \\ & \ddots & \\ & & P_n(1-P_n) \end{pmatrix}, X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix}$$

Therefore,

$$l(\beta) = l(\beta_0) + (Y^T - P^T)X(\beta - \beta_0) - \frac{1}{2}(\beta - \beta_0)^T X^T DX(\beta - \beta_0)$$

Define $h = \beta - \beta_0$, then $\widehat{h} = argmax_h(Y^T - P^T)Xh - \frac{1}{2}h^TX^TDXh = (X^TDX)^{-1}X^T(Y - P)$, thus,

$$\widehat{\beta}_{t+1} = \widehat{\beta}_t + h(\widehat{\beta}_t) = \widehat{\beta}_t + (X^TD(\widehat{\beta}_t)X)^{-1}X^T(Y - P(\widehat{\beta}_t))$$

2.2 Poisson Regression

For $\xi \sim poisson(\lambda)$, the pmf of ξ is $P(\xi = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, $k = 0, 1, 2, \cdots$, and $\mathbb{E}\xi = \lambda, \mathbb{V}\text{ar}\xi = k$.

For the i-th observation, $y_i \sim poisson(\lambda_i)$, with explanatory variable $X_i^T = (X_{i1}, \cdots, X_{ik})$. Since $\lambda_i = \mathbb{E}y_i$, we want to use a function of $X_i^T\beta$ to estimate λ_i. Here we use exponential function, that is,

$$\lambda_i = e^{X_i^T\beta} = exp(X_{i1}\beta_1 + \cdots + X_{ik}\beta_k)$$

Thus, $\lambda_i/\lambda_j = exp((X_i - X_j)^T\beta)$, then, if $\beta_i > 0$, y is positively correlated with X_i, and vice versa.

Next we are going to find $\widehat{\beta}$, an estimator of β, since we can predict Y_{n+1} using $poisson(e^{X_{n+1}^T\widehat{\beta}})$. Here we use MLE to estimate β, that is,

$$\widehat{\beta}_{MLE} = argmax_{\widehat{\beta}} \prod_{i=1}^{n} \frac{\lambda_i^k e^{-\lambda_i}}{Y_i!}$$

3 Analysis of Variance

3.1 Single-Factor Analysis of Variance

Factor means an independent random variable to be studied, and factor level means a particular form of the random variable.

Eg1. The influence of pricing on sales amount.

Eg2. The effect of three different drugs on one particular disease.

In eg1, the factor level A_i is quantitative, and in eg2, the factor level A_i is qualitative. Here we define the factor as A and its factor levels as $A_1, A_2, \cdots, A_r$. And for each factor level A_i, we have n_i observations $y_{i1}, \cdots, y_{in_i}$. We make a few settings:

1. All observations under the same factor level A_i follow the same distribution $N(\mu_i, \sigma^2)$.
2. All observations under all factor levels have the same variance σ^2.
3. Observations under different factor levels A_i, A_j follow different distributions, that is $\mu_i \neq \mu_j$.
4. $n_1, n_2, \cdots, n_r$ can be different, and $n_1 + n_2 + \cdots + n_r = n$.

Thus, for y_{ij}, which means the j-th observation under factor level A_i, $y_{ij} = \mu_i + \epsilon_{ij}$, where for all i, j, $\epsilon_{ij} \sim N(0, \sigma^2)$ independently and identically. Then, we can combine all y_{ij} together as follow,

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{r1} \\ \vdots \\ y_{rn_r} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_2 \\ \vdots \\ \mu_r \\ \vdots \\ \mu_r \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{r1} \\ \vdots \\ \epsilon_{rn_r} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_r \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{r1} \\ \vdots \\ \epsilon_{rn_r} \end{pmatrix}$$

That is, $Y = X_{n \times r}\beta + \epsilon$, where X is the design matrix, of which the i-th column indicates the i-th factor level. Then we are going to give the definition of a few notations that we will use later.

1. $y_{i\cdot}$
2. $\overline{y}_{i\cdot}$
3. $y_{\cdot\cdot}$
4. $\overline{y}_{\cdot\cdot}$

3.1.1 Estimators of the Means and the Variance

First, the LSE of μ_i. According to the definition, $\widehat{\mu}_i = argmin_{\theta_i} Residual^2$, and since we use $\widehat{\mu}_i$ to fit Y_{ij}, thus,

$$\widehat{\mu}_{i\,LSE} = argmin_{\widehat{\mu}_i} \sum_{j=1}^{n_i} (Y_{ij} - \widehat{\mu}_i)^2$$

Find the derivative, we have

$$2\sum_{j=1}^{n_i}(Y_{ij} - \widehat{\mu}_i) = 0 \Rightarrow \sum_{j=1}^{n_i} Y_{ij} = n_i\widehat{\mu}_i \Rightarrow \widehat{\mu}_{i\,LSE} = \overline{Y}_{i\cdot}$$

Second, the MLE of μ_i. The pdf of y_{ij} is $\frac{1}{\sqrt{2\pi\sigma^2}} exp[-\frac{(y_{ij}-\mu_i)^2}{2\sigma^2}]$, thus the joint pdf will be

$$\prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y_{ij}-\mu_i)^2}{2\sigma^2}\right] = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n_i} exp\left[-\sum_{j=1}^{n_i}\frac{(y_{ij}-\mu_i)^2}{2\sigma^2}\right]$$

Therefore, the log-likelihood function will be

$$l(\mu_i) = -\sum_{j=1}^{n_i}\frac{(y_{ij}-\mu_i)^2}{2\sigma^2} - \frac{n_i}{2}\log 2\pi\sigma^2$$

Hence, the MLE of μ_i will be

$$\widehat{\mu}_{i\,MLE} = argmax_{\widehat{\mu}_i} l(\widehat{\mu}_i) = argmin_{\widehat{\mu}_i} \sum_{j=1}^{n_i} (y_{ij} - \widehat{\mu}_i)^2 = \widehat{\mu}_{i\,LSE} = \overline{y}_{i\cdot}$$

Since $\widehat{\mu}_i = \overline{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \sim N(\mu_i, \frac{\sigma^2}{n_i})$, thus $\widehat{\mu}_i$ is unbiased. Also we can prove that $\widehat{\mu}_i$ has the minimum variance, thus $\widehat{\mu}_i$ is the BLUE of μ_i.

Third, the MLE of σ^2. The log-likelihood function of σ^2 is

$$l(\sigma^2) = -\sum_{i=1}^{r}\sum_{j=1}^{n_i} \frac{(y_{ij}-\mu_i)^2}{2\sigma^2} - \frac{n}{2}\log 2\pi\sigma^2$$

We will have

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\cdot})^2$$

According to Theorem 1.1 in simple linear regression, we have
$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \sigma^2\chi^2(n-1)$. So,

$$\mathbb{E}\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{r}(n_i-1)\mathbb{E}\left\{\frac{1}{n_i-1}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\cdot})^2\right\} = \frac{1}{n}\sum_{i=1}^{r}(n_i-1)\sigma^2 = \frac{n-r}{n}\sigma^2$$

Therefore, $\widehat{\sigma}^2$ is biased, and we introduce an unbiased estimator of σ^2,

$$s^2 = \frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\cdot})^2$$

3.1.2 Hypothesis Testing

Next we are going to do a hypothesis testing that $H_0 : \mu_1 = \cdots = \mu_r$. We first consider the sum of squares of $y_{ij} - \overline{y}_{\cdot\cdot}$, and we can divide it as $y_{ij} - \overline{y}_{i\cdot} + \overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}$. For the first half, $y_{ij} - \overline{y}_{i\cdot}$, we call it within group. For the second half, $\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}$, we call it between group. Then, we define

$$SST = \sum_{i,j}(y_{ij} - \overline{y}_{\cdot\cdot})^2 = \sum_{i,j}(y_{ij}-\overline{y}_{i\cdot})^2 + \sum_{i,j}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2 \triangleq S_e + S_A$$

Also, we can build an ANOVA Table as follow.

	Sum of Squares	**df**	**Mean Square**
within group	S_e	n-r	$S_e/n-r$
between group	S_A	r-1	$S_A/r-1$
total	SST	n-1	$SST/n-1$

Then, how to do the hypothesis testing? Since $S_A = \sum_{i,j}(a_i + \overline{\epsilon}_{i\cdot} - \overline{\epsilon}_{\cdot\cdot})^2$, where $a_i = \mu_i - \frac{1}{n}\sum_{i=1}^{r} n_i\mu_i$. Under H_0, all the μ_i are equal, so $a_i = 0$. If $\overline{y}_{1\cdot} = \cdots = \overline{y}_{r\cdot}$, $S_A = 0$. So we reject H_0 if S_A is large enough. Therefore, we are going to find the distribution of S_A under H_0. Define $\xi_i = \frac{\sqrt{n_i}\overline{\epsilon}_{i\cdot}}{\sigma} \sim N(0,1)$, then we can write $\overline{\epsilon}_{i\cdot}$ and $\overline{\epsilon}_{\cdot\cdot}$ as $\overline{\epsilon}_{i\cdot} = \xi_i\sigma/\sqrt{n_i}$ and $\overline{\epsilon}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{r} n_i\overline{\epsilon}_{i\cdot} = \frac{1}{n}\sum_{i=1}^{r}\sqrt{n_i}\xi_i\sigma$. Therefore,

$$
\begin{aligned}
S_A &= \sum_{i=1}^{r} n_i(\overline{\epsilon}_{i\cdot} - \overline{\epsilon}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{r} (\sqrt{n_i}\overline{\epsilon}_{i\cdot} - \sqrt{n_i}\overline{\epsilon}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{r} (\sigma\xi_{i\cdot} - \frac{1}{n}\sum_{j=1}^{r}\sqrt{n_i n_j}\xi_i\sigma)^2 \\
&= \cdots \\
&= \sigma^2(\xi_1, \cdots, \xi_r)(I-A)^T(I-A)\begin{pmatrix}\xi_1 \\ \vdots \\ \xi_r\end{pmatrix}
\end{aligned}
$$

where $A_{ij} = \frac{1}{n}\sqrt{n_i n_j}$, we can prove that A is symmetric and idempotent. And

$$
tr(I-A) = r - \sum_{i=1}^{r}\frac{1}{n}\sqrt{n_i n_i} = r-1
$$

Therefore, $S_A \sim \chi^2(r-1)\sigma^2$, that is $MS_A \sim \frac{\sigma^2\chi^2(r-1)}{r-1}$.
$MS_e = \frac{1}{n-r}\sum_{i,j}(y_{ij} - \overline{y}_{i\cdot})^2 = s^2 \sim \frac{\sigma^2\chi^2(n-r)}{n-r}$ is the unbiased estimator of σ^2. Because MS_A can be written as a function of sample means under each factor level and MS_e can be written as a function of sample variances under each factor level, we can get $MS_A \perp MS_e$. Therefore, we introduce the test statistic

$$
F = \frac{MS_A}{MS_e} \sim F_{r-1,n-r}
$$

Thus, we reject H_0 if $F = \frac{MS_A}{MS_e} > F_{r-1,n-r}(\alpha)$.

Similar to linear regression, we can use Extra Sum of Squares Principle to do the same thing. Define the full model as

$$
\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \vdots \\ y_{r1} \\ \vdots \\ y_{rn_r} \end{pmatrix} = \begin{pmatrix} 1_{n_1} & 0 & \cdots & 0 \\ 0 & 1_{n_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1_{n_r} \end{pmatrix}\begin{pmatrix}\mu_1 \\ \vdots \\ \mu_r\end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \vdots \\ \epsilon_{r1} \\ \vdots \\ \epsilon_{rn_r} \end{pmatrix}
$$

and the reduced model as

$$
\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \vdots \\ y_{r1} \\ \vdots \\ y_{rn_r} \end{pmatrix} = 1_n\mu + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \vdots \\ \epsilon_{r1} \\ \vdots \\ \epsilon_{rn_r} \end{pmatrix}
$$

and we can write the null hypothesis $\mu_1 = \cdots = \mu_r = \mu$ as

$$C\begin{pmatrix}\mu_1\\ \vdots\\ \mu_r\end{pmatrix}=\begin{pmatrix}1 & -1 & 0 & \cdots & 0 & 0\\ 0 & 1 & -1 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 1 & -1\end{pmatrix}\begin{pmatrix}\mu_1\\ \vdots\\ \mu_r\end{pmatrix}=0$$

Then, we can prove that $SSE_R=\sum_{i,j}(y_{ij}-\overline{y}_{..})^2=SST$, $SSE_F=\sum_{i,j}(y_{ij}-\overline{y}_{i.})^2=S_e$. Hence

$$\frac{SSE_R-SSE_F/m}{SSE_F/n-k-1}=\frac{SST-S_e/r-1}{S_e/n-r}=\frac{S_A/r-1}{S_e/n-r}\sim F_{r-1,n-r}$$

3.1.3 Confidence Interval

First, we are going to see the confidence interval of a single μ_i.

Since $\widehat{\mu}_i=\overline{y}_{i\cdot}\sim N(\mu_i,\frac{\sigma^2}{n_i})$, $\frac{\overline{y}_{i\cdot}-\mu_i}{\sigma/\sqrt{n_i}}\sim N(0,1)$, $\sqrt{MS_e/\sigma^2}\sim\sqrt{\frac{\chi^2(n-r)}{n-r}}$, and $\overline{y}_{i\cdot}\perp MS_e$, we have

$$\frac{\overline{y}_{i\cdot}-\mu_i}{\sqrt{MS_e/n_i}}=\frac{\frac{\overline{y}_{i\cdot}-\mu_i}{\sigma/\sqrt{n_i}}}{\sqrt{MS_e/\sigma^2}}\sim t_{n-r}$$

Therefore, the $1-\alpha$ confidence interval of μ_i is

$$\left[\overline{y}_{i\cdot}\pm t_{n-r}(\alpha/2)\sqrt{MS_e/n_i}\right]$$

Second, the confidence interval of the difference between μ_i and μ_j, defined as $D_{ij}=\mu_i-\mu_j$.

Since $\widehat{D}_{ij}=\overline{y}_{i\cdot}-\overline{y}_{j\cdot}\sim N(\mu_i-\mu_j,\frac{\sigma^2}{n_i}+\frac{\sigma^2}{n_j})$, similarly, the $1-\alpha$ confidence interval of D_{ij} is

$$\left[\overline{y}_{i\cdot}-\overline{y}_{j\cdot}\pm t_{n-r}(\alpha/2)\sqrt{MS_e\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}\right]$$

Third, any linear combination of $\{\mu_i\}$, defined as $A=\sum_{i=1}^r\mu_ic_i$.

Since $\widehat{A}=\sum_{i=1}^r\overline{y}_{i\cdot}c_i\sim N(A,\sum_{i=1}^r c_i^2\frac{\sigma^2}{n_i})$, similarly, the $1-\alpha$ confidence interval of A is

$$\left[\sum_{i=1}^r\overline{y}_{i\cdot}c_i\pm t_{n-r}(\alpha/2)\sqrt{MS_e\sum_{i=1}^r\frac{1}{n_i}c_i^2}\right]$$

3.2 Two-Factor Analysis of Variance

Here we have 2 factors A and B, and factor levels $\{A_i\}_{i=1}^a$ and $\{B_j\}_{j=1}^b$, then we have total ab different combinations of A_i,B_j called treatment.

$A\setminus B$	B_1	B_j	B_b	**Mean**
A_1	μ_{11}		μ_{1b}	$\overline{\mu}_{1\cdot}$
A_i		μ_{ij}		$\overline{\mu}_{i\cdot}$
A_a	μ_{a1}		μ_{ab}	$\overline{\mu}_{a\cdot}$
Mean	$\overline{\mu}_{\cdot 1}$	$\overline{\mu}_{\cdot j}$	$\overline{\mu}_{\cdot b}$	$\overline{\mu}_{\cdot\cdot}$

Define a_i as the main effect of A_i, that is, $a_i = \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot\cdot}$, b_j as the main effect of B_j, that is, $b_j = \overline{\mu}_{\cdot j} - \overline{\mu}_{\cdot\cdot}$, and treatment main effect as $\mu_{ij} - \overline{\mu}_{\cdot\cdot}$. We have $\sum_i a_i = \sum_j b_j = 0$. We can classify the model into two categories, according to whether the treatment main effect can be divided into the sum of factor levels main effect, namely without intersection and with intersection. On the one hand, without intersection, that is $\mu_{ij} - \overline{\mu}_{\cdot\cdot} = a_i + b_j$. On the other hand, with intersection, that is, $\mu_{ij} - \overline{\mu}_{\cdot\cdot} \neq a_i + b_j$, define

$$r_{ij} = \mu_{ij} - \overline{\mu}_{\cdot\cdot} - a_i - b_j = \mu_{ij} - \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot j} + \overline{\mu}_{\cdot\cdot}$$

then $\sum_i r_{ij} = \sum_j r_{ij} = 0$.

For the observations, we define y_{ijk} as the k-th observation under the treatment of i, j. We have the following notations, $y_{ij\cdot}, \overline{y}_{ij\cdot}, y_{i\cdot\cdot}, \overline{y}_{i\cdot\cdot}, y_{\cdot j\cdot}, \overline{y}_{\cdot j\cdot}, y_{\cdots}, \overline{y}_{\cdots}$. Similarly, we can use $\overline{y}_{ij\cdot}$ to estimate μ_{ij} , and so on. And then $\widehat{a}_i = \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots}$, $\hat{b}_j = \overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots}$, $\hat{r}_{ij} = \overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdots}$, and

$$\sum_i \widehat{a}_i = \sum_j \hat{b}_j = \sum_i \hat{r}_{ij} = \sum_i \hat{r}_{ij} = 0$$

3.2.1 Hypothesis Testing

Consider the sum of squares.

$$SST = \sum_{i,j,k}(y_{ijk} - \overline{y}_{\cdots})^2 = \sum_{i,j,k}(y_{ijk} - \overline{y}_{ij\cdot})^2 + \sum_{i,j,k}(\overline{y}_{ij\cdot} - \overline{y}_{\cdots})^2 = S_e + S_{AB}$$

$$S_e = \sum_{i,j}\sum_{k}(\epsilon_{ijk} - \overline{\epsilon}_{ij\cdot})^2 \sim \sum_{i,j}\sigma^2\chi^2(n-1) \sim \sigma^2\chi^2(ab(n-1))$$

$$S_{AB} = \sum_{i,j,k}(\overline{y}_{ij\cdot} - \overline{y}_{\cdots})^2 = \sum_{i,j,k}(\widehat{a}_i + \hat{b}_j + \hat{r}_{ij})^2 = \sum_{i,j,k}\widehat{a}_i^2 + \sum_{i,j,k}\hat{b}_j^2 + \sum_{i,j,k}\hat{r}_{ij}^2 = SS_A + SS_B + SS_{AB}$$

Next, we are going to do the first hypothesis testing $H_0 : r_{ij} = 0, \forall i = 1, \cdots, a, j = 1, \cdots, b$. Since

$$\begin{aligned}
SS_{AB} &= \sum_{i,j,k}\hat{r}_{ij}^2 = \sum_{i,j,k}(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdots})^2 \\
&= \sum_{i,j,k}(\mu_{ij} + \overline{\epsilon}_{ij\cdot} - \mu_{i\cdot} - \overline{\epsilon}_{i\cdot\cdot} - \mu_{\cdot j} - \overline{\epsilon}_{\cdot j\cdot} + \mu_{\cdot\cdot} + \overline{\epsilon}_{\cdots})^2 \\
&= \sum_{i,j,k}(r_{ij} + \overline{\epsilon}_{ij\cdot} - \overline{\epsilon}_{i\cdot\cdot} - \overline{\epsilon}_{\cdot j\cdot} + \overline{\epsilon}_{\cdots})^2 \\
&\overset{H_0}{=} \sum_{i,j,k}(\overline{\epsilon}_{ij\cdot} - \overline{\epsilon}_{i\cdot\cdot} - \overline{\epsilon}_{\cdot j\cdot} + \overline{\epsilon}_{\cdots})^2 \\
&= \sum_{i,j} n(\overline{\epsilon}_{ij\cdot} - \overline{\epsilon}_{i\cdot\cdot} - \overline{\epsilon}_{\cdot j\cdot} + \overline{\epsilon}_{\cdots})^2 \\
&= n\sum_{i,j} A_{ij}A_{ji}^T = ntrAA^T
\end{aligned}$$

where $A = P_a\epsilon_{a\times b}P_b = (I_a - \frac{1}{a}1_a1_a^T)\epsilon_{a\times b}(I_b - \frac{1}{b}1_b1_b^T)$, and P_a, P_b are symmetric and idempotent. Thus,

$$SS_{AB} = ntrP_a\epsilon P_bP_b^T\epsilon^T P_a^T = ntr\epsilon P_bP_b^T\epsilon^T P_a^T P_a = tr\sqrt{n}\epsilon P_b\sqrt{n}\epsilon^T P_a = tr\tilde{\epsilon}P_b\tilde{\epsilon}^T P_a$$

Define $P_a = Udiag_a\{1, \cdots, 1, 0\}U^T$, $P_b = Vdiag_b\{1, \cdots, 1, 0\}V^T$, then

$$\begin{aligned} SS_{AB} &= tr\tilde{\epsilon}Vdiag_b\{1,\cdots,1,0\}V^T\tilde{\epsilon}^TUdiag_a\{1,\cdots,1,0\}U^T \\ &= trU^T\tilde{\epsilon}Vdiag_b\{1,\cdots,1,0\}V^T\tilde{\epsilon}^TUdiag_a\{1,\cdots,1,0\} \\ &= tr\xi diag_b\{1,\cdots,1,0\}\xi^T diag_a\{1,\cdots,1,0\} \\ &= \sum_{i=1}^{a-1}\sum_{j=1}^{b-1}\xi_{ij}^2\cdot 1\cdot 1 \sim \sigma^2\chi^2((a-1)(b-1)) \end{aligned}$$

Therefore, we introduce the F-statistic

$$F = \frac{MS_{AB}}{MS_e} \sim \frac{\sigma^2\chi^2((a-1)(b-1))}{\sigma^2\chi^2((n-1)ab)} \sim F_{(a-1)(b-1),(n-1)ab}$$

And we reject H_0 if $F > F_{(a-1)(b-1),(n-1)ab}(\alpha)$.

If we accept the previous hypothesis, which means there is no intersection, then we are going to do the second hypothesis testing $H_0 : a_1 = \cdots = a_a = 0 \Leftrightarrow \overline{\mu}_{1\cdot} = \cdots = \overline{\mu}_{a\cdot} = \overline{\mu}_{\cdot\cdot}$.

Consider the full model

$$SST_F = S_{eF} + SS_{AF} + SS_{BF} + SS_{ABF}$$

and the reduced model, since $r_{ij} = 0$, we have $SS_{ABR} = 0$, thus,

$$SST_R = S_{eR} + SS_{AR} + SS_{BR}$$

And since $\overline{y}_{ijk} = \overline{\mu}_{\cdot\cdot} + a_i + b_j + r_{ij} + \epsilon_{ijk}$, then,

$$\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots} = \overline{\mu}_{\cdot\cdot} + a_i + \overline{b}_{\cdot} + \overline{r}_{i\cdot} + \overline{\epsilon}_{i\cdot\cdot} - \overline{\mu}_{\cdot\cdot} - \overline{a}_{\cdot} - \overline{b}_{\cdot} - \overline{r}_{\cdot\cdot} - \overline{\epsilon}_{\cdots} = a_i + \overline{\epsilon}_{i\cdot\cdot} - \overline{a}_{\cdot} - \overline{\epsilon}_{\cdots}$$

Therefore, $SS_{AR} = \sum_{i,j,k}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots})^2$ is independent with r_{ij}, that is, $SS_{AR} = SS_{AF}$, similarly, $SS_{BR} = SS_{BF}$. And since $SST_R = SST_F$, we can get $S_{eR} = S_{eF} + SS_{ABF}$. Next we will focus on the reduced model, so we omit the subscript R. We first consider S_e,

$$\because SST = \sum_{i,j,k}(y_{ijk} - \overline{y}_{\cdots})^2,\ SS_A = \sum_{i,j,k}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots})^2,\ SS_B = \sum_{i,j,k}(\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots})^2$$

$$\therefore S_e = \sum_{i,j,k}(y_{ijk} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdots})^2 = \sum_{i,j,k}(\epsilon_{ijk} - \overline{\epsilon}_{i\cdot\cdot} - \overline{\epsilon}_{\cdot j\cdot} + \overline{\epsilon}_{\cdots})^2$$

Define three projection matrix P_1, P_2, P_3, which are easy to be proved as symmetric and idempotent,

$$\epsilon = \begin{pmatrix} \epsilon_{111} \\ \vdots \\ \epsilon_{11n} \\ \epsilon_{121} \\ \vdots \\ \epsilon_{12n} \\ \vdots \\ \epsilon_{ab1} \\ \vdots \\ \epsilon_{abn} \end{pmatrix},\ P_1\epsilon = \begin{pmatrix} \overline{\epsilon}_{1\cdot\cdot} \\ \vdots \\ \overline{\epsilon}_{1\cdot\cdot} \\ \vdots \\ \overline{\epsilon}_{a\cdot\cdot} \\ \vdots \\ \overline{\epsilon}_{a\cdot\cdot} \end{pmatrix},\ P_2\epsilon = \begin{pmatrix} \overline{\epsilon}_{\cdot 1\cdot} \\ \vdots \\ \overline{\epsilon}_{\cdot 1\cdot} \\ \vdots \\ \overline{\epsilon}_{\cdot b\cdot} \\ \vdots \\ \overline{\epsilon}_{\cdot b\cdot} \end{pmatrix},\ P_3\epsilon = \begin{pmatrix} \overline{\epsilon}_{\cdots} \\ \vdots \\ \overline{\epsilon}_{\cdots} \end{pmatrix},$$

then, since

$$rank(I) = abn,\ rank(P_1) = rank(P_1\epsilon) = a,\ rank(P_2) = rank(P_2\epsilon) = b,\ rank(P_3) = rank(P_3\epsilon) = 1$$
$$\therefore rank(I - P_1 - P_2 + P_3) = abn - a - b + 1$$
$$\therefore S_e = \|\epsilon - P_1\epsilon - P_2\epsilon + P_3\epsilon\|^2 = \epsilon^T(I - P_1 - P_2 + P_3)\epsilon \sim \sigma^2\chi^2(abn - a - b + 1)$$

That is to say,

$$\mathbb{E}\frac{S_e}{abn - a - b + 1} = \sigma^2$$

Then, we are going to find the distribution of SS_A,

$$SS_A = \sum_{i,j,k}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot\cdot\cdot})^2 \overset{H_0}{=} \sum_{i,j,k}(\overline{\epsilon}_{i\cdot\cdot} - \overline{\epsilon}_{\cdot\cdot\cdot})^2 = \|P_1\epsilon - P_3\epsilon\|^2 = \epsilon^T(P_1 - P_3)\epsilon \sim \sigma^2\chi^2(a - 1)$$

Since $\begin{pmatrix} SS_A \\ S_e \end{pmatrix} = \begin{pmatrix} I - P_1 - P_2 + P_3 \\ P_1 - P_3 \end{pmatrix}\epsilon \sim Gaussian$, thus,

$$\mathbb{C}\text{ov}(SS_A, S_e) = \sigma^2(I - P_1 - P_2 + P_3)(P_1 - P_3) = 0$$

Therefore, $SS_A \perp S_e$, so we can introduce the F-statistic,

$$F = \frac{MS_A}{MS_e} \sim \frac{\sigma^2\chi^2(a - 1)}{\sigma^2\chi^2(abn - a - b + 1)} \sim F_{a-1,abn-a-b+1}$$

Similarly, we can also do hypothesis testing on $H_0 : b_1 = \cdots = b_b = 0 \Leftrightarrow \overline{\mu}_{\cdot 1} = \cdots = \overline{\mu}_{\cdot b} = \overline{\mu}_{\cdot\cdot}$, with the F-statistic,

$$F = \frac{MS_B}{MS_e} \sim \frac{\sigma^2\chi^2(b - 1)}{\sigma^2\chi^2(abn - a - b + 1)} \sim F_{b-1,abn-a-b+1}$$

3.2.2 Confidence Interval

Since the MS_e in the full model is different from that in the reduced model, so when we are required to give the confidence interval, we have to check whether the model is with or without intersection.

If the model has intersection, $MS_e = \sum_{i,j,k}(y_{ijk} - \overline{y}_{ij\cdot})^2/ab(n - 1)$, so the confidence interval of μ_{ij} is

$$\left[\overline{y}_{ij\cdot} \pm t_{ab(n-1)}(\alpha/2)\sqrt{MS_e/n}\right]$$

the confidence interval of $\mu_{i\cdot}$ is

$$\left[\overline{y}_{i\cdot\cdot} \pm t_{ab(n-1)}(\alpha/2)\sqrt{MS_e/bn}\right]$$

If the model does not have intersection,
$MS_e = \sum_{i,j,k}(y_{ijk} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdot\cdot\cdot})^2/(abn - a - b + 1)$, so the confidence interval of μ_{ij} is

$$\left[\overline{y}_{ij\cdot} \pm t_{abn-a-b+1}(\alpha/2)\sqrt{MS_e/n}\right]$$

the confidence interval of $\mu_{i\cdot}$ is

$$\left[\overline{y}_{i\cdot\cdot} \pm t_{abn-a-b+1}(\alpha/2)\sqrt{MS_e/bn}\right]$$