Modern Statistics **Lecture 12 - 3/31/2021**

Parametric Inference-II

Lecturer: Xiangyu Chang *Scribe: Yongqi Chen, Jilong Chen*

Edited by: Zhihong Liu

1 Recall

Multiple Regression

General format:

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\beta + \epsilon.$$

$$n > p, \text{Rank}(\mathbf{X}) = p.$$

The least squares estimate is:

$$\min \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta\|^2.$$

From the least squares estimate, we can get the estimator of β:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

We can know $\hat{\beta}$'s distribution:

$$\mathbb{E}[\epsilon|\mathbf{X}] = 0, \mathbb{V}[\epsilon|\mathbf{X}] = \sigma^2\mathbf{I} \Rightarrow \mathbb{E}[\hat{\beta}] = \beta, \mathbb{V}[\hat{\beta}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Example 1 *There exist p kinds of parameters: $\mathbf{X} = (X_1, \ldots, X_p)$, we suppose that X_1 = Income and X_2 = Cost, according to general logic, there may be $X_2 = \alpha X_1$, then we can use X_1 and X_2 to build a linear regression model (3 or more parameters are similar situations).*

2 Model Selection

If we have p parameters, we can build $2^p - 1 = \binom{p}{1} + \binom{p}{2} + \cdots + \binom{p}{p}$ linear regression models. Therefore, we need "Model Selection".

3 Unbiased estimate of σ^2

$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, therefore $\hat{\mathbf{Y}}$ is in $\mathbf{X}$'s column space. Suppose that $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}, \langle \mathbf{e}, \hat{\mathbf{Y}} \rangle = 0$, this has been proved in the last class. If $\mathbf{P}$ is a Projection Matrix, then $\mathbf{P}^T = \mathbf{P}, \mathbf{P}^2 = \mathbf{P} \Rightarrow \lambda = 1$ or 0. Property: $(\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$.

- $\mathbf{P}^2 = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{P}$.
- $\mathbf{P}(\mathbf{X}\mathbf{a}) = \mathbf{P}\mathbf{X}\mathbf{a} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{X}\mathbf{a}$. ($\mathbf{X}\mathbf{a}$ must be in $\mathbf{X}$'s column space)
- $\mathbf{P} = \mathbf{D}^T\mathbf{D}\mathbf{D}$. (Eigen Value Decomposition), $\mathbf{D} = \begin{pmatrix} 1 & & & & & \\ & \dots & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & \dots & \\ & & & & & 0 \end{pmatrix}$. (There are p "1"s)
- $\text{Rank}(\mathbf{P}) = \text{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$, $(\text{tr}(AB) = \text{tr}(BA))$. Namely: $\text{tr}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}) = \text{tr}(\mathbf{I}_p) = p$.
- $(\mathbf{I}-\mathbf{P})^2 = \mathbf{I}+\mathbf{P}^2-2\mathbf{P}\mathbf{I} = \mathbf{I}+\mathbf{P}-2\mathbf{P} = \mathbf{I}-\mathbf{P}$. ($\mathbf{I}-\mathbf{P}$ is a projection matrix too.) $\text{Rank}(\mathbf{I}-\mathbf{P}) = n-p$.

  $\mathbf{I}-\mathbf{P} = \mathbf{V}^T\mathbf{D}\mathbf{V}, \mathbf{D} = \begin{pmatrix} 1 & & & & & \\ & \dots & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & \dots & \\ & & & & & 0 \end{pmatrix}$. (There are $(n-p)$ "1"s)

Theorem 1 *Having the above properties, we can prove the following equation:*

$$\mathbb{E}[\mathbf{e}^T\mathbf{e}] = (n-p)\sigma^2.$$

Proof 1 *Since* $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I}-\mathbf{P})\mathbf{Y}$ *and* $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, *we have*

$$\mathbf{e}^T\mathbf{e} = \mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y} = (\mathbf{X}\beta+\epsilon)^T(\mathbf{I}-\mathbf{P})(\mathbf{X}\beta+\epsilon) = \epsilon^T(\mathbf{I}-\mathbf{P})\epsilon.$$

Denote $\mathbf{Z} = \mathbf{V}\epsilon$. *Since* $\mathbf{I}-\mathbf{P} = \mathbf{V}^T\mathbf{D}\mathbf{V}$, *then* $\mathbf{e}^T\mathbf{e} = (\mathbf{V}\epsilon)^T\mathbf{D}(\mathbf{V}\epsilon)$.

- $\mathbb{E}[\mathbf{Z}|\mathbf{X}] = 0$.
- $\mathbb{V}(\mathbf{Z}|\mathbf{X}) = \mathbf{V}^T\mathbb{V}(\epsilon)\mathbf{V} = \sigma^2\mathbf{V}^T\mathbf{V} = \sigma^2$.

Therefore, $\mathbf{Z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \in \mathbb{R}^n$, $\mathbb{E}[z_i] = 0, \mathbb{V}(z_i) = \sigma^2$. $\mathbb{E}[\mathbf{e}^T\mathbf{e}] = \mathbf{Z}^T\mathbf{D}\mathbf{Z} = \sum_{i=1}^{n-p}\mathbb{E}(z_i^2) = (n-p)\sigma^2$.

4 Inference

We suppose that:

$$\epsilon \mid \mathbf{X} \sim N(0, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \mathbf{Y} \mid \mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}).$$

Therefore:

$$\hat{\beta} \mid \mathbf{X} \sim N\left(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right).$$

This implies:

$$\frac{(\mathbf{X}^T\mathbf{X})^{1/2}(\beta - \hat{\beta})}{\sigma} \sim N(0, \mathbf{I}).$$

Let $\mathbf{Z} = \mathbf{V}\epsilon$. Then:

$$\mathbf{Z} \sim N(0, \sigma^2\mathbf{I}).$$

We also have:

$$\frac{\mathbf{e}^T\mathbf{e}}{n-p} = \sum_{i=1}^{n-p} z_i^2 \sim \sigma^2\chi^2(n-p).$$

This leads to:

$$\frac{\mathbf{e}^T\mathbf{e}}{(n-p)\sigma^2} \sim \chi^2(n-p).$$

Finally, we can derive the following distribution:

$$\frac{\sqrt{n-p}(\mathbf{X}^T\mathbf{X})^{1/2}(\beta - \hat{\beta})}{\|\mathbf{e}\|} \sim t(n-p).$$

5 Logistic Regression

5.1 Bernoulli

There exists a data set: $\{(x_i, y_i)\}_{i=1}^n, y_i \in \{0, 1\}$. We can believe that y has the distribution:

$$y_i|x_i \sim \text{Ber}(p_i).$$

$\exists f(x_i)$, such that $0 \le p_i = f(x_i) \le 1$. For example, Sigmoid Function: $f(x_i) = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$ (Link Function - link β and p_i). Logistic Regression has no RSS, so we can only use MLE to find the parameter p_i. MLE Function is:

$$\prod_{i=1}^n \left\{(p_i)^{y_i}(1-p_i)^{1-y_i}\right\}.$$

This becomes the optimization problem:

$$\max_\beta \sum_{i=1}^n \{y_i \log p_i + (1-y_i)\log(1-p_i)\} \Longleftrightarrow \max_\beta \sum_{i=1}^n \left\{y_i \log\frac{p_i}{1-p_i} + \log(1-p_i)\right\}.$$

Due to:

$$\log(1-p_i) = -\log(1+e^{x_i^T\beta}) \quad \text{and} \quad \log\frac{p_i}{1-p_i} = x_i^T\beta.$$

The above optimization problem is:

$$\max_\beta \left\{y_i x_i^T\beta - \log(1+e^{x_i^T\beta})\right\}.$$

(Which is called General Linear Model - GLM) Notice: This problem usually has no analytic solution, so we need to use Gradient Descent (GD) or Newton's method, etc., to solve it.

5.2 Poisson

When the data set is: $\{(x_i, y_i)\}_{i=1}^n, y_i \in \{0, 1, \ldots, \infty\}$. We can believe that y has the distribution:

$$y_i|x_i \sim \text{Poisson}(\lambda_i).$$

It also has a Link Function: $\lambda_i = x_i^T\beta$. After that, the process is similar, using MLE to solve for β.