Chapter 7: Sufficiency

Tonglin Zhang

7.1 Measures of Quality of Estimators

Definition(*UMVUE or MVUE*) An estimator T is called a (uniform) minimum variance unbiased estimator (UMVUE or MVUE) of θ if T is unbiased and $V(T)$ is less than or equal to that of any other unbiased estimator of θ.

Two requirement of MVUE of θ is that:

(i) $\mathrm{E}(T) = \theta$;

(ii) $\mathrm{V}(T) \leq \mathrm{V}(T')$

if T' is an unbiased estimator of θ.

Definition. (Complete Sufficient Statistics). A statistic T is complete sufficient statistic (CSS) if it satisfies that $\mathrm{E}[g(T)] = 0$ iff $g(T) = 0$.

Theorem. Let $X_1, \cdots, X_n$ be random samples. Suppose $\theta \in \Theta$ is unknown parameters. Assume $\Theta \in R^k$ contains an open set (i.e., the dimension of Θ is k). Let T be sufficient statistic. If the dimension of T equals k, then T is CSS.

Theorem. (CSS and UMVUE). Let T be CSS. If $\mathrm{E}[g(T)] = k(\theta)$, then $g(T)$ is UMVUE of $k(\theta)$.

Definition. Let T be estimator of θ. The MSE of T is defined as

$$\mathrm{MSE}_\theta(T) = \mathrm{E}_\theta[(T-\theta)^2].$$

For two estimators T and T' of θ, we say T is better than T' if

$$\mathrm{MSE}_\theta(T) \leq \mathrm{MSE}_\theta(T').$$

We say T is the best estimator of θ if

$$\mathrm{MSE}_\theta(T) \leq \mathrm{MSE}_\theta(T')$$

for any estimator T' of θ.
Note: The best estimator may not exist. Therefore, one often studies the risk function, defined under the Bayesian perspective.

Bayesian Risk: Let $X = (X_1, \cdots, X_n)$ be the observations of random samples. Suppose $\theta \in \Theta$ is an unknown parameter. Let T be the statistic and we use $\delta = \delta(T)$ to estimate θ (called decision in Bayesian theory). Assume we choose a loss function $L(\theta, \delta)$. Let $f_\theta(t)$ be the pdf (or PMF) of T. Then, the risk function at θ is defined by

$$\begin{aligned} R(\theta, \delta) =& \mathrm{E}[L(\theta, \delta(X))] \\ =& \begin{cases} \int_{-\infty}^{\infty} L(\theta, \delta(t)) f_\theta(t) dt & \text{continuous case} \\ \sum_t L(\theta, \delta(t)) f_\theta(t) & \text{discrete case} \end{cases} \end{aligned}$$

Usually, we use the squared error loss as

$$L(\theta, \delta(t)) = (\delta(t) - \theta)^2.$$

We have the following criteria:

- Minimax principle: The maximum of the risk function is minimized, i.e., we need to find $\delta_0(t)$ so that

$$\max_\theta R(\theta, \delta_0) \leq \max_\theta R(\theta, \delta)$$

  for all δ. This is called the minimax estimator.

- Bayes principle: Given a prior density $\pi(\theta)$ for θ, the average risk is minimized, i.e., we need to find $\delta_0(t)$ so that

$$\int_\Theta R(\theta, \delta_0)\pi(\theta)d\theta \leq \int_\Theta R(\theta, \delta)\pi(\theta)d\theta.$$

  This is called the Bayes estimator.

Proposition. Under the squared error loss, the Bayes estimator is the posterior mean.

Example: Let $X_1, \cdots, X_n$ be iid *Poisson*(λ). Let $\pi(\lambda)$ is the PDF of $\Gamma(\alpha, \beta)$. Let $L(\delta, \lambda) = (\delta - \lambda)^2$. Find the MLE and the Bayesian estimator.

Solution: The MLE is $\hat{\lambda} = \bar{X}$. Note that $T = \sum_{i=1}^{n} X_i \sim$ *Poisson*$(n\lambda)$. The PMF of T is

$$f_n(t) = \frac{(n\lambda)^t}{t!} e^{-n\lambda}.$$

The joint PMF-PDF of T and λ is

$$\begin{aligned} f_n(t)\pi(\lambda) &= \frac{(n\lambda)^t}{t!} e^{-n\lambda} \frac{\beta^\alpha \lambda^{\alpha-1}}{\Gamma(\beta)} e^{-\beta\lambda} \\ &= \frac{\beta^\alpha n^t}{t!\Gamma(\beta)} \lambda^{t+\alpha-1} e^{-(n+\beta)\lambda}. \end{aligned}$$

The marginal PMF of T is

$$\bar{f}_n(t) = \int_0^\infty f_n(t)\pi(\lambda)d\lambda = \frac{\Gamma(t+\alpha)}{t!\Gamma(\beta)}\frac{\beta^\alpha n^t}{(n+\beta)^t+\alpha}.$$

The posterior PDF of λ is

$$q(\lambda|data) = \frac{f_n(t)\pi(\lambda)}{\bar{f}_n(t)} = \frac{(n+\beta)^{t+\alpha}}{\Gamma(t+\alpha)}\lambda^{t+\alpha-1}e^{-(n+\beta)\lambda},$$

where is $\Gamma(t+\alpha, n+\beta)$ distribution. Thus, the Bayesian estimator is the expected value of $\Gamma(t+\alpha, n+\beta)$ distribution, which is

$$\hat{\lambda}_{bayes} = \frac{T+\alpha}{n+\beta}.$$

Comparing it with $\hat{\lambda} = T/n$, they should be close if n is large.

Example: 7.1.2. Let $X_1, \cdots, X_{25}$ are iid from $N(\theta, 1)$. Consider an estimator of θ. We use $\bar{X}$, where $\bar{X} \sim N(\theta, 1/25)$.

(a) $\bar{X}$ is the UMVUE.

(b) Suppose the prior for θ is $N(0, 1)$. Then, the posterior density of θ is

$$N(\frac{t}{26}, 1/26).$$

Thus, the Bayesian decision under the squared error loss is

$$\delta(t) = \frac{t}{26}.$$

7.2 A Sufficient Statistics for a Parameter

Definition 7.2.1. Let $X_1, \cdots, X_n$ be a random sample from a PDF $f_\theta(x)$. Let Y be a statistic (or vector of statistic). Let $f_\theta(x_1, \cdots, x_n)$ be the joint PDF or PMF of $X_1, \cdots, X_n$, and $f_{Y,\theta}(y)$ be the PDF or PMF of Y. If

$$\frac{f_\theta(x_1, \cdots, x_n)}{f_{Y,\theta}(y)} = H(x_1, \cdots, x_n)$$

where H dose not depend on θ, then we call Y is sufficient statistic for θ.

Theorem 7.2.1: (Neyman Factorization Theorem). To check whether Y is sufficient statistic for θ, we just need to check the following formula:

$$f_\theta(x_1, \cdots, x_n) = k_1(y, \theta) k_2(x_1, \cdots, x_n)$$

where h_2 does not depend on θ. The above equation is "if and only if".

- Many books use the formula given by Neyman Factorization Theorem as the definition.
- You do not need to pay attention to the second function ($k_2(\cdot)$). It may be complicated.
- The first function $k_1(y, \theta)$ is important. We need to simply it.
- The size of y is usually much lower than the size of the data.
- Since our interest is θ, it is enough to look at $k_1(y, \theta)$ only.
- Note that the loglikelihood function is

$$\ell(\theta) = \log k_1(y, \theta) + \log k_2(x_1, \cdots, x_n).$$

The MLE does not depend on $k_2(\cdot)$.

Theorem. Let $X_1, \cdots, X_n$ be random samples. If the MLE of θ exists, then it is a function of SS only.

Example. Let $X_1, \cdots, X_n$ be iid $N(\mu, \sigma^2)$ and denote $\boldsymbol{\theta} = (\mu, \sigma^2)$. The likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}$$

$$= (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}[\sum_{i=1}^{n}(X_i - \bar{X})^2 - n(\bar{X}-\mu)^2]\right\}\left(\frac{1}{\sqrt{2\pi}}\right)^n.$$

Thus

$$\{\bar{X}, \sum_{i=1}^{n}(X_i - \bar{X})^2\}$$

is CSS. Because its size is identical to the size of $\boldsymbol{\theta}$, we conclude it is also CSS. Since $\bar{X}$ is an unbiased estimator of μ, it is also the UMVUE of μ. Because $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ is an unbiased estimator of σ^2, it is also the UMVUE of σ^2.

Example. Let $X_1, \cdots, X_n$ be iid $Poisson(\lambda)$. The likelihood function is

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \left(\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}\right) \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right).$$

Thus, $\sum_{i=1}^{n} X_i$ (or $\bar{X}$, equivalently) is SS. Since its size is identical the number of parameter (both are 1), we conclude it is also CSS. Because $\mathrm{E}(\bar{X}) = \lambda$, $\bar{X}$ is the UMVUE of θ.

Example. Let $X_1, \cdots, X_n$ be iid $Bernouli(p)$. The likelihood function is

$$L(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}.$$

Thus, $\bar{X}$ (or $\sum_{i=1}^n X_i$) is SS and also CSS. Since $\mathrm{E}(\bar{X}) = p$, $\bar{X}$ is also the UMVUE of p.

Example. Let $X_1, \cdots, X_n$ be iid $Uniform[0, \theta]$. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I(0 \le X_i \le \theta) = \left(\frac{1}{\theta^n} I(0 \le X_{(n)} \le \theta \right) \left(I(0 < X_{(1)}) \right).$$

Thus, $X_{(n)}$ is CSS. To find CSS, we need find the function $g(X_{(n)})$ such that $\mathrm{E}[g(X_{(n)})] = \theta$. By PDF of $X_{(n)}$

$$f_n(x) = \frac{n x_{n-1}}{\theta^n}.$$

We can choose $g(X_{(n)}) = cX_{(n)}$ for some c. Then,

$$\mathrm{E}(cX_{(n)}) = c\int_0^\theta x f_n(x)dx = c\int_0^\theta \frac{nx^n}{\theta^n}dx = c\frac{n}{n+1}\theta.$$

Let $c = (n+1)/n$. We obtain the UMVUE $[(n+1)/n]X_{(n)}$.

Example. Let $X_1, \cdots, X_n$ be iid $Uniform[\theta_1, \theta_2]$. The likelihood function of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1} \frac{1}{\theta_2 - \theta_1} I(\theta_1 < X_i < \theta_2) \\ &= (\frac{1}{(\theta_2 - \theta_1)^n}) I(X_{(1)} > \theta_1) I(X_{(n)} < \theta_2). \end{aligned}$$

Thus, $\{X_{(1)}, X_{(n)}\}$ is the set of SS, and also CSS. The derivation of UMVUE is omitted here because it is complicated.

Example. Example 7.2.5. Let $X_1, \cdots, X_n$ be iid with PDF $f_\theta(x) = \theta x^{\theta-1}$ for $0 < x < 1$ and $\theta > 0$. The likelihood function is

$$L(\theta) = \theta^n e^{(\theta-1)\sum_{i=1}^n \log X_i}.$$

Thus, $\{\sum_{i=1}^n \log X_i\}$ is SS and also CSS. For the UMVUE, we need

$$\begin{aligned}
\mathrm{E}(\log X_i) &= \int_0^1 (\log x)\theta x^{\theta-1} dx \\
&= \int_0^1 (\log x) dx^\theta \\
&= -\int_0^1 x^{\theta-1} dx \\
&= -\frac{1}{\theta}.
\end{aligned}$$

Thus, $-\sum_{i=1}^n \log X_i/n$ is the UMVUE of $1/\theta$. However, $-(\sum_{i=1}^n \log X_i/n)^{-1}$ is not the UMVUE of θ.

Definition. (Minimum Sufficient Statistics). The minimum sufficient statistic is the sufficient statistic with minimum dimensions. It is different from the complete sufficient statistic. If complete sufficient statistic exists, then it is a minimum sufficient statistic.

Note: MSS has computational advantage. We only need MSS in the computation. In the case when the size of data is large, we can focus on the derivation of MSS.

7.3 Properties of Sufficient Statistic and 7.4 Completeness and Uniqueness

Two important formulae: For any two random variables X_1 and X_2, we have

- $$\mathrm{E}(X_2) = \mathrm{E}[\mathrm{E}(X_2|X_1)].$$
- $$\mathrm{V}(X_2) = \mathrm{E}[\mathrm{V}(X_2|X_1)] + \mathrm{V}[\mathrm{E}(X_2|X_1)].$$

Theorem 7.3.1 (Rao-Blackwell). Let $X_1, \cdots, X_n$ be random samples. Suppose T_1 is a sufficient statistic and T_2 is a statistic which can represent as a function of T_1 and other random variables (i.e. $T_2 = g(T_1, \cdot)$). Assume $\varphi(T_2)$ is an unbiased estimator of θ. Then, we can find another estimator depending on T_1 only which has a less variance than $\varphi(T_2)$. In fact, this estimator is

$$\mathrm{E}[\varphi(T_2)|T_1].$$

Proof. First,

$$\mathrm{E}[\varphi(T_2)] = \theta \Rightarrow \mathrm{E}\{\mathrm{E}[\varphi(T_2)|T_1]\} = \theta.$$

Thus, $\mathrm{E}[\varphi(T_2)|T_1]$ is unbiased. Second,

$$\mathrm{V}[\varphi(T_2)] = \mathrm{E}\{\mathrm{V}[\varphi(T_2)|T_1]\} + \mathrm{V}\{\mathrm{E}[\varphi(T_2)|T_1]\} \geq \mathrm{V}\{\mathrm{E}[\varphi(T_2)|T_1]\}.$$

Thus, we have the conclusion. □

Theorem. (Uniqueness of UMVUE). Let T be CSS. Let U and V be unbiased estimator of θ. Then, $\mathrm{E}(U|T) = E(V|T)$ are UMVUE of θ.

Proof. This can be shown by the definition of CSS.

Example. Let $X_1, \cdots, X_n$ be iid $N(\mu, \sigma^2)$. Then X_1 is an unbiased estimator of μ. Based on this estimator, find a better estimator based on CSS $(\bar{X}, S^2)$. Note that $\bar{X}$ is the UMVUE of μ and S^2 is the UMVUE of S^2. The UMVUE of $\mu^2 + \sigma^2$ is $\sum_{i=1}^n X_i^2/n$. We have

$$\mathrm{E}(X_1|\bar{X}, S^2) = \bar{X}$$

and

$$\mathrm{E}(X_1^2|\bar{X}, S^2) = \frac{1}{n}\sum_{i=1}^n X_i^2.$$

Example. Let $X_1, \cdots, X_n$ be iid *Poisson*(λ). Then X_1 is an unbiased of λ. Then, $\mathrm{E}(X_1|\bar{X}) = \bar{X}$.

Example. Let $X_1, \cdots, X_n$ be iid *Uniform*$[0, \theta]$. Then, $2\bar{X}$ is an unbiased estimator of θ. Since

$$\mathrm{E}(2\bar{X}) = \theta,$$

we have

$$\mathrm{E}(2\bar{X}|X_{(n)}) = \frac{(n+1)X_{(n)}}{n}.$$

Linear Regression

Suppose we observe data y_i and $\mathbf{x}_i = (1, x_{i1}, \cdots, x_{i(p-1)})^\top$ for $i = 1, \cdots, n$, where $p-1$ is the number of explanaotry variables. Then, y_i is the ith observed value of the response and $\mathbf{x}_i$ is the ith observed vector of explanatory variables. Consider the linear regression model

$$\begin{aligned} y_i =& \beta_0 + x_{i1}\beta_1 + \cdots x_{i(p-1)}\beta_{p-1} + \epsilon_i \\ =& \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2), \end{aligned}$$

for $i = 1, \cdots, n$.

Let

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n(p-1)} \end{pmatrix}$$

and

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}.$$

Then, the model can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \ \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

Neither $\mathbf{X}$ nor $\boldsymbol{\beta}$ are random. Only ϵ is random. Thus, we have

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}).$$

Let $\theta = (\boldsymbol{\beta}, \sigma^2) = (\boldsymbol{\beta}, \phi)$. The loglikelihood function of $\boldsymbol{\theta}$ is

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\phi - \frac{1}{2\phi}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\phi \\ &\quad -\frac{1}{2\phi}\left[\sum_{i=1}^{n} y_i^2 - 2(\sum_{i=1}^{n} y_i\mathbf{x}_i)^\top\boldsymbol{\beta} + \boldsymbol{\beta}^\top\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)\boldsymbol{\beta}\right].\end{aligned}$$

Thus,

$$SS = \{\sum_{i=1}^{n} y_i^2, \sum_{i=1}^{n} y_i\mathbf{x}_i, \sum_{i=1}^{n} x_i\mathbf{x}_i^\top\}.$$

The size of SS is $(p+1)^2$, which is smaller than the size of the data.

By matrix, we have

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\phi - \frac{1}{2\phi}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

The equation function of $\boldsymbol{\beta}$ is

$$\dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = -\frac{1}{2\phi}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

The SSE of the model is

$$\begin{aligned} SSE =& \|\hat{\boldsymbol{\epsilon}}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \\ =& \|(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y}\|^2 \\ =& \mathbf{y}^\top(\mathbf{I} - \mathbf{P})\mathbf{y}, \end{aligned}$$

where $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{P})\mathbf{y}$ and $\mathbf{P} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

Definition. (Orthogonal projection matrix.) For any $n \times n$ matrix $\mathbf{M}$, we say it is an orthogonal projection matrix if $\mathbf{M}^\top \mathbf{M} = \mathbf{M}$.

Theorem. If $\mathbf{M}$ is an $n \times n$ orthogonal project matrix, then there exists an orthogonal matrix $\mathbf{Q}$ such that it can be expressed as

$$\mathbf{M} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top,$$

where $r = \mathrm{tr}(\mathbf{M})$, $\mathbf{D} = \mathrm{diag}(\mathbf{1}_r, \mathbf{0}_{n-r})$, $\mathbf{1}_r$ is the r-dimensional vector composed of 1 and $\mathbf{0}_{n-r}$ is the $(n-r)$-dimensional vector composed of 0.

Proof. Let λ be an eigenvalue of $\mathbf{M}$. Then, we can find $\mathbf{u}$ such that $\mathbf{Mu} = \lambda\mathbf{u}$. By $\mathbf{M} = \mathbf{M}^\top = \mathbf{M}^\top\mathbf{M} = \mathbf{M}^2$. We have

$$\lambda\mathbf{u} = \mathbf{Mu} = \mathbf{M}^2\mathbf{u} = \lambda\mathbf{Mu} = \lambda^2\mathbf{u}.$$

Thus, λ can only be 0 or 1. By eigen-decomposition of symmetric matrix, we can find orthogonal matrix $\mathbf{Q}$ (i.e., it is the matrix of eigen vectors), such that

$$\mathbf{M} = \mathbf{QDQ}^\top,$$

where $\mathbf{D}$ is the diagonal matrix of eigen values. By

$$\mathrm{tr}(\mathbf{M}) = \mathrm{tr}(\mathbf{QDQ}^\top) = \mathrm{tr}(\mathbf{DQ}^\top\mathbf{Q}) = \mathrm{tr}(\mathbf{D}),$$

we can permute the eigenvalues and eigenvectors to draw the conclusion. □

Theorem. If $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$, then for any orthogonal projection matrix $\mathbf{M}$ there is

$$\|\mathbf{Mz}\|^2 \sim \chi^2_r$$

with

$$r = \mathrm{tr}(\mathbf{M}).$$

Proof. Suppose that

$$\mathbf{M} = \mathbf{QDQ}^\top,$$

where $\mathbf{Q}$ and $\mathbf{D}$ are defined previously. We obtain

$$\begin{aligned}\|\mathbf{Mz}\|^2 =& \mathbf{z}^\top \mathbf{M}^\top \mathbf{Mz} = \mathbf{z}^\top \mathbf{Mz} \\ =& \mathbf{z}^\top \mathbf{QDQ}^\top \mathbf{z} = (\mathbf{Q}^\top \mathbf{z})^\top \mathbf{D}(\mathbf{Q}^\top \mathbf{z}),\end{aligned}$$

implying that we can define

$$\mathbf{y} = \mathbf{Q}^\top \mathbf{z}.$$

We obtain

$$\|\mathbf{Mz}\|^2 = \mathbf{z}^\top \mathbf{Mz} = \mathbf{y}^\top \mathbf{Dy}.$$

Note that $\mathrm{E}(\mathbf{y}) = \mathbf{0}$ and

$$\mathrm{V}(\mathbf{y}) = \mathbf{Q}^\top \mathrm{V}(\mathbf{z})\mathbf{Q} = \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}.$$

We have $\mathbf{y} \sim N(\mathbf{0}, \mathbf{I})$. By $\|\mathbf{Mz}\|^2 = \sum_{i=1}^{r} y_i^2$, we draw the conclusion. □

Corollary. If $\mathbf{z} \sim N(0, \mathbf{M})$ for an orthogonal projection matrix M, then

$$\|\mathbf{z}\|^2 \sim \chi^2_r,$$

where

$$r = \mathrm{tr}(\mathbf{M}).$$

Proof. Please show it by yourself. □

Theorem In the regression model, we have

(a) $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\epsilon}}$ are independent,

(b) $\|\hat{\boldsymbol{\epsilon}}\|^2 \sim \sigma^2 \chi^2_{n-p}$,

(c) $\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\|^2 \sim \sigma^2 \chi^2_p$.

Proof. By

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

and

$$\hat{\boldsymbol{\epsilon}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top)\mathbf{y},$$

we have

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\epsilon}}) =& (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathrm{Cov}(\mathbf{y}, \mathbf{y})(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \\
=& \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \\
=& \mathbf{0}_{p \times n}.
\end{aligned}$$

Thus, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\epsilon}}$ are independent. Then, we conclude (a).

Then, we have

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\epsilon}}) &= \mathrm{E}[(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y}] \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathrm{E}(\mathbf{y}) \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}.\end{aligned}$$

We also have

$$\begin{aligned}\mathrm{V}(\hat{\boldsymbol{\epsilon}}) =& \mathrm{V}[(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y}] \\ =& (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathrm{V}(\mathbf{y})(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) \\ =& \sigma^2(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top).\end{aligned}$$

Since $\mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is an orthogonal projection matrix and

$$\begin{aligned}\mathrm{tr}(\mathbf{I} - \mathbf{P}) =& \mathrm{tr}(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\\ =& \mathrm{tr}(\mathbf{I}) - \mathrm{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\\ =& \mathrm{tr}(\mathbf{I}_n) - \mathrm{tr}(\mathbf{I}_p)\\ =& n - p,\end{aligned}$$

we draw (b).

For (c), we have

$$\begin{aligned}\mathrm{E}(\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) =& \mathrm{E}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}]-\mathbf{X}\boldsymbol{\beta}\\ =& \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{E}(\mathbf{y})-\mathbf{X}\boldsymbol{\beta}\\ =& \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}-\mathbf{X}\boldsymbol{\beta}\\ =& \mathbf{0}\end{aligned}$$

and

$$\begin{aligned}\mathrm{V}(\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) = \mathrm{V}(\mathbf{X}\hat{\boldsymbol{\beta}}) =& \mathbf{V}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}]\\ =& \sigma^2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \sigma^2\mathbf{P}.\end{aligned}$$

Since $\mathbf{P}$ is an orthogonal projection matrix and $\mathrm{tr}(\mathbf{P}) = p$, we draw the conclusion.