Foundations/Stats

STATS 743B

Cameron Roopnarine* Narayanaswamy Balakrishnan†

1st March 2023

LECTURE 1
11th January

Order Statistics

Let $X_1, X_2, \dots, X_n$ be a random sample of size n from a population with CDF $F(x)$ and PDF $f(x)$. Let $X_{1:n} \leq \cdots \leq X_{n:n}$ denote the corresponding order statistics obtained by arranging the X_i's in increasing (non-decreasing) order of magnitude. Then, their distributions, dependence properties, moments, characteristics, etc. can be made use of effectively to develop inferential methods.

Binomial Derivation

The CDF of $X_{r:n}$, for $r = 1, 2, \dots, n$, is

$$\begin{aligned}
F_{r:n}(x) &= \mathbb{P}(X_{r:n} \leq x) \\
&= \mathbb{P}(\text{at least } r \text{ of the } X_i\text{'s are} \leq x) \\
&= \sum_{i=r}^{n} \mathbb{P}(\text{exactly } i \text{ of } x_i\text{'s are} \leq x_i) && \text{because they are mutually exclusive} \\
&= \sum_{i=r}^{n} \binom{n}{i} \big(F(x)\big)^i \big(1 - F(x)\big)^{n-i} \\
&= I_{F(x)}(r, n-r+1),
\end{aligned}$$

where

$$I_p(a,b) = \frac{1}{B(a,b)} \int_0^p t^{a-1}(1-t)^{b-1} \, \mathrm{d}t, \text{ for } 0 < p < 1,$$

is the Incomplete Beta Ratio function.

Pearson's Identity

For $0 < p < 1$,

$$I_p(r, n-r+1) = \sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i}, \text{ for } r = 1, 2, \dots, n.$$

It connects the survival function of binomial distribution with the cumulative distribution function of beta distribution. (Proof by integration of parts)

*LaTeXer

†Instructor

REMARK 1

The expression of the CDF of $X_{r:n}$ in (1) holds whether the distribution $F(x)$ is continuous or discrete.

If the distribution is continuous, then the PDF of $X_{r:n}$, for $r = 1, \dots, n$, can be obtained from (1) as

$$\begin{aligned}
f_{r:n}(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F_{r:n}(x) \\
&= \frac{\mathrm{d}}{\mathrm{d}x} I_{F(x)}(r, n-r+1) \\
&= \frac{\mathrm{d}}{\mathrm{d}x} \frac{1}{B(r, n-r+1)} \int_0^{F(x)} t^{r-1}(1-t)^{n-r}\,\mathrm{d}t \\
&= \frac{1}{B(r, n-r+1)} \big(F(x)\big)^{r-1} \big(1-F(x)\big)^{n-r} f(x) \\
&= \frac{n!}{(r-1)!(n-r)!} \big(F(x)\big)^{r-1} \big(1-F(x)\big)^{n-r} f(x).
\end{aligned}$$

If the population distribution $F(x)$ is discrete, however, the above method can not be used. But, we can find the PDF of $X_{r:n}$ (for $r = 1, \dots, n$) as

$$\begin{aligned}
f_{r:n}(x) &= \mathbb{P}(X_{r:n} = x) \\
&= F_{r:n}(x) - F_{r:n}(x-) \\
&= \frac{1}{B(r, n-r+1)} \int_{F(x-)}^{F(x)} t^{r-1}(1-t)^{n-r}\,\mathrm{d}t.
\end{aligned}$$

Derivation from Jacobian

Now, let us focus on the case when the population distribution is continuous. In this case, with $f(x)$ as the PDF, due to independence of X_i's, their joint density is

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i), \quad x_i \in S.$$

Now, let us introduce the transformation

$$X_{1:n} = \min\{X_1, \dots, X_n\}, \dots, X_{n:n} = \max\{X_1, \dots, X_n\}.$$

Then, evidently, it is an $n!$-to-1 transformation, and so the joint density function of $(X_{1:n}, \dots, X_{n:n})$ is

$$f_{1,\dots,n:n}(x_1, \dots, x_n) = n! \prod_{i=1}^{n} f(x_i), \; x_1 < x_2 < \cdots < x_n.$$

From (4), we can obtain, by integrating out $(x_{r+1}, \dots, x_n)$ and $(x_1, \dots, x_{r-1})$, the PDF of $X_{r:n}$ $(r = 1, \dots, n)$ as follows:

$$\underset{x_r < x_{r+1} < \cdots < x_n}{\int\int \cdots \int} f(x_{r+1}) \cdots f(x_n)\,\mathrm{d}x_n\,\mathrm{d}x_{n-1} \cdots \mathrm{d}x_{r+1} = \frac{[1-F(x_r)]^{n-r}}{(n-r)!}$$

and

$$\underset{x_1 < x_2 < \cdots < x_r}{\int\int \cdots \int} f(x_1) \cdots f(x_{r-1})\,\mathrm{d}x_1\,\mathrm{d}x_2 \cdots \mathrm{d}x_{r-1} = \frac{[F(x_r)]^{r-1}}{(r-1)!}$$

so that we obtain the PDF of $X_{r:n}$ as

$$f_{r:n}(x_r) = \frac{n!}{(r-1)!(n-r)!}\big(F(x_r)\big)^{r-1}\big(1-F(x_r)\big)^{n-r} f(x_r).$$

Similarly, from (4), we can obtain, by integrating out $(x_{s+1}, \ldots, x_n)$, $(x_{r+1}, \ldots, x_{s-1})$ and $(x_1, \ldots, x_{r-1})$, the joint PDF of $(X_{r:n}, X_{s:n})$, for $1 \le r < s \le n$, as follows:

$$\underset{x_s < x_{s+1} < \cdots < x_n}{\int\int \cdots \int} f(x_{s+1}) \cdots f(x_n)\,\mathrm{d}x_n\,\mathrm{d}x_{n-1}\,\mathrm{d}x_{s+1} = \frac{[1-F(x_s)]^{n-s}}{(n-s)!},$$

$$\underset{x_r < x_{r+1} < \cdots < x_{s-1} < x_s}{\int\int \cdots \int} f(x_{r+1}) \cdots f(x_{s-1})\,\mathrm{d}x_{r+1}\,\mathrm{d}x_{r+2}\,\mathrm{d}x_{s-1} = \frac{[F(x_s)-F(x_r)]^{s-r-1}}{(s-r-1)!},$$

and

$$\underset{x_1 < x_2 < \cdots < x_{r-1} < x_r}{\int\int \cdots \int} f(x_1) \cdots f(x_{r-1})\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_{r-1} = \frac{[F(x_r)]^{r-1}}{(r-1)!},$$

so that we can obtain the joint PDF of $(X_{r:n}, X_{s:n})$, for $1 \le r < s \le n$ as

$$\begin{aligned} f_{r,s:n}(x_r, x_s) = {} & \frac{n!}{(r-1)!(s-r-1)!(n-s)!}\big(F(x_r)\big)^{r-1}\big(F(x_s)-F(x_r)\big)^{s-r-1}\big(1-F(x_s)\big)^{n-s} f(x_r) f(x_s), \\ & \text{for } x_r < x_s. \end{aligned}$$

Multinomial Derivation

For directly deriving the PDF of $X_{r:n}$, let us consider

$$\mathbb{P}(x \le X_{r:n} \le x+\Delta x) = \frac{n!}{(r-1)!1!(n-r)!}\big(F(x)\big)^{r-1}\big(F(x+\Delta x)-F(x)\big)^{1}\big(1-F(x+\Delta x)\big)^{n-r} + \mathcal{O}((\Delta x)^2),$$

where $\mathcal{O}((\Delta x)^2)$ corresponds to more than one x_i in the interval $(x, x+\Delta x)$. Then, we obtain the density of $X_{r:n}$ as follows:

$$\begin{aligned} f_{r:n}(x) &= \lim_{\Delta x \to 0} \frac{\mathbb{P}(x \le X_{r:n} \le x+\Delta x)}{\Delta x} \\ &= \frac{n!}{(r-1)!1!(n-r)!}\big(F(x)\big)^{r-1} \lim_{\Delta x \to 0}\Big[\big(F(x+\Delta x)-F(x)\big)^{1} \\ &\qquad \big(1-F(x+\Delta x)\big)^{n-r}\Big] + \mathcal{O}((\Delta x)^2) \\ &= \frac{n!}{(r-1)!(n-r)!}\big(F(x)\big)^{r-1} f(x)\big(1-F(x)\big)^{n-r}, \end{aligned}$$

exactly as before.

Similarly, for deriving the joint PDF of $(X_{r:n}, X_{s:n})$, for $1 \le r < s \le n$, let us consider the multinomial probability

$$\begin{aligned} & \mathbb{P}(x < X_{r:n} \le x+\Delta x, y < X_{s:n} \le y+\Delta y) \\ &= \frac{n!}{(r-1)!1!(s-r-1)!1!(n-s)!}\big(F(x)\big)^{r-1}\big(F(x+\Delta x)-F(x)\big)^{1} \\ &\quad \times \big(F(y)-F(x+\Delta x)\big)^{s-r-1}\big(F(y+\Delta y)-F(y)\big)^{1}\big(1-F(y+\Delta y)\big)^{n-s} \\ &\quad + \mathcal{O}((\Delta x)^2 \Delta y) \to \text{corresponds to more than one } X_i \text{ in } (x, x+\Delta x) \\ &\quad + \mathcal{O}(\Delta x (\Delta y)^2) \to \text{corresponds to more than one } X_i \text{ in } (y, y+\Delta y) \end{aligned}$$

Then, we obtain the joint density of $(X_{r:n}, X_{s:n})$ as follows:

$$\begin{aligned}
f_{r,s:n}(x,y) &= \lim_{\Delta x \to 0, \Delta y \to 0} \frac{\mathbb{P}(x < X_{r:n} \le x + \Delta x, y < X_{s:n} \le y + \Delta y)}{\Delta x \Delta y} \\
&= \frac{n!}{(r-1)!1!(s-r-1)!1!(n-s)!} \big(F(x)\big)^{r-1} \\
&\times \lim_{\Delta x \to 0} \left[\frac{\big(F(x+\Delta x) - F(x)\big)^1}{\Delta x} \big(F(y) - F(x+\Delta x)\big)^{s-r-1} \right] \\
&\times \lim_{\Delta y \to 0} \left[\frac{\big(F(y+\Delta y) - F(y)\big)^1}{\Delta y} \big(1 - F(y+\Delta y)\big)^{n-s} \right] + \mathcal{O}((\Delta x)^2 \Delta y) + \mathcal{O}(\Delta x (\Delta y)^2) \\
&= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \big(F(x)\big)^{r-1} f(x) \big(F(y) - F(x)\big)^{s-r-1} f(y) \big(1 - F(y)\big)^{n-s}, \ x < y,
\end{aligned}$$

exactly as before.

EXAMPLE 1

Let us consider Uniform$(0,1)$ distribution with

$$f(x) = 1 \text{ for } 0 < x < 1, \qquad F(x) = x \text{ for } 0 < x < 1.$$

Then, from (2), we have the PDF of $X_{r:n}$ (for $1 \le r \le n$) to be

$$f_{r:n}(x) = \frac{1}{B(r, n-r+1)} x^{r-1} (1-x)^{n-r} \text{ for } 0 < x < 1;$$

that is,

$$X_{r:n} \overset{\mathrm{d}}{=} \text{Beta}(r, n-r+1).$$

So, we readily have

$$\mathbb{E}[X_{r:n}] = \frac{r}{n+1} = \pi_r, \qquad \text{Var}(X_{r:n}) = \frac{r(n-r+1)}{(n+1)^2(n+2)} = \frac{\pi_r(1-\pi_r)}{n+2}.$$

Similarly, from (6), we have the joint PDF of $(X_{r:n}, X_{s:n})$ for $1 \le r < s \le n$, to be

$$f_{r,s:n}(x,y) = \frac{1}{B(r, s-r, n-s+1)} x^{r-1} (y-s)^{s-r-1} (1-y)^{n-s} \text{ for } 0 < x < y < 1,$$

which implies

$$(X_{r:n}, X_{s:n}) \overset{\mathrm{d}}{=} \text{BivBeta}(r, s-r, n-s+1).$$

From this, we can readily find, for $1 \le r < s \le n$,

$$\begin{aligned}
\text{Cov}(X_{r:n}, X_{s:n}) &= \mathbb{E}[X_{r:n} X_{s:n}] - \mathbb{E}[X_{r:n}]\, \mathbb{E}[X_{s:n}] \\
&= \frac{\pi(n-s+1)}{(n+1)^2(n+2)} \\
&= \frac{\pi_r(1-\pi_s)}{n+2};
\end{aligned}$$

observe that they are always positively correlated. Moreover,

$$\rho_{X_{r:n}, X_{s:n}} = \frac{\text{Cov}(X_{r:n}, X_{s:n})}{\sqrt{\text{Var}(X_{r:n})\, \text{Var}(X_{s:n})}} = \sqrt{\frac{\pi_r}{1-\pi_r} \times \frac{1-\pi_s}{\pi_s}},$$

free of n (just a function of proportions $\frac{r}{n+1}$ and $\frac{s}{n+1}$).

Probability Integral Transform

Suppose X is a continuous random variable with CDF $F(x)$ and PDF $f(x)$. Then, the transformed variable $U = F(x)$ is uniformly distributed over $(0, 1)$.

Proof: For $u \in (0, 1)$, consider

$$\begin{aligned}\mathbb{P}(U \leq u) &= \mathbb{P}(F(x) \leq u) \\ &= \mathbb{P}(X \leq Q(u)) \\ &= F(Q(u)) \\ &= u,\end{aligned}$$

where $Q(u)$ is the quantile function (i.e., it is F^{-1} in the case of absolute continuous function), which means $U = F(x)$ is Uniform$(0, 1)$.

REMARK 2

The above result will hold even if the population distribution is not absolutely continuous, but has discontinuities. All we have to do is take Q as the generalized quantile function with right inverse.
Since $F(x)$ is a non-decreasing function, if we have $(X_{1:n}, X_{2:n}, \ldots, X_{n:n})$ as order statistics from a continuous distribution with CDF $F(x)$, then the transformed variables $\big(F(X_{1:n}), F(X_{2:n}), \ldots, F(X_{n:n})\big)$ will be distributed as Uniform$(0, 1)$ order statistics, $(U_{1:n}, U_{2:n}, \ldots, U_{n:n})$ no matter what the distribution of $F(\,\cdot\,)$ is!

Probability-Probability Plot

One important application of the previous result is in model validation methods. Because

$$\big(F(X_{1:n}, \ldots, F(X_{n:n}))\big) \text{ and } (U_{1:n}, \ldots, U_{n:n})$$

have identical distributions no matter what the population distribution $F(\,\cdot\,)$ is, we can use it to examine whether an assumed $F(\,\cdot\,)$ is reasonable for the data at hand. This is done by a P-P plot as follows:

- Step 1: From the given data $(x_{1:n}, \ldots, x_{n:n})$, estimate the parameters of the assumed model $F(\,\cdot\,)$;
- Step 2: With the estimated parameter values, say $\hat{\theta}$, find the values of

  $$\big(F(x_{1:n}; \hat{\theta}), F(x_{2:n}; \hat{\theta}), \ldots, F(x_{n:n}; \hat{\theta})\big).$$

  These are the "empirical" (or observed) probabilities;
- Step 3: Plot these against the "theoretical" probabilities

  $$\left(\mathbb{E}[U_{1:n}] = \frac{1}{n+1}, \mathbb{E}[U_{2:n}] = \frac{2}{n+1}, \ldots, \mathbb{E}[U_{n:n}] = \frac{n}{n+1}\right).$$

  A near straight line fit would provide support for the assumed model $F(\,\cdot\,)$.

REMARK 3

One can also indicate variability at each point by estimating $\mathrm{Var}(F(x_{i:n}))$ (using delta method).

Quantile-Quantile Plot

Another important and related application is the Q-Q plot. In it, we invert the distributional relationship to use

$$(X_{1:n}, X_{2:n}, \dots, X_{n:n}) \text{ and } (F^{-1}(U_{1:n}), F^{-1}(U_{2:n}), \dots, F^{-1}(U_{n:n}))$$

have identical distribution, where $Q \equiv F^{-1}(\,\cdot\,)$ is the quantile function of the assumed model. Then, the $Q - Q$ plot proceeds as follows:

- Step 1: Determine the order statistics from the given data, $x_{1:n}, x_{2:n}, \dots, x_{n:n}$, which are the empirical quantiles;
- Plot them against the theoretical quantiles

  $$\left(F^{-1}\left(\frac{1}{n+1}\right), F^{-1}\left(\frac{2}{n+1}\right), \dots, F^{-1}\left(\frac{n}{n+1}\right)\right).$$

  Once again, a near straight line fit would provide support for the assumed model $F(\,\cdot\,)$.

REMARK 4

Once again, we can indicate the variability at each point by estimating $\mathrm{Var}(X_{i:n})$.

REMARK 5

In both cases, rather than making a qualitative assessment on "near straight line fit", can can make it quantitatively by using the correlation coefficient or any other "measure of fit" (correlation-type goodness-of-fit test).

REMARK 6

Note that estimation of the model parameters is avoided in Q-Q plot, but is necessary in a P-P plot!

Pivot 1

A pivot is a random variable, which is a function of the data and the parameter of the model, whose distribution is free of the parameter(s).

Pivots are essential quantities for developing inferential methods such as interval estimation, hypothesis tests, etc.

EXAMPLE 2

Let $X_1, \dots, X_n$ be a random sample from EXP(θ) distribution with PDF

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta},\ x > 0,\ \theta > 0.$$

Then, it is well-known that

$$\sum_{i=1}^{n} X_i \sim \text{GAM}(n, \theta),$$

where n is the shape parameter and θ is the scale parameter of the Gamma distribution. Hence,

$$Y = \frac{\sum_{i=1}^{n} X_i}{\theta}$$

is a pivot (for θ) since its distribution is GAM$(n, 1)$ which is free of the parameter θ.

EXAMPLE 3

Suppose $X_1, \dots, X_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$ distribution. Let $\bar{X}$ and S^2 denote the sample mean and sample variance, respectively. Then, it is known that

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ and } \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

So, if we consider

$$\bar{X} - \mu \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right),$$

it will not be a pivot if σ^2 is unknown, and it will be a pivot for μ only if σ^2 is known. However, if we consider

$$\frac{\bar{X} - \mu}{S/\sqrt{n}},$$

it will be a pivot (for μ) since its distribution will be Student's t distribution with $n-1$ degrees of freedom, as it is free of both μ and σ^2. Similarly,

$$\frac{(n-1)S^2}{\sigma^2}$$

will be a pivot (for σ^2) as its distribution is central χ^2 distribution with $n-1$ degrees of freedom, and is free of μ and σ^2.

EXAMPLE 4

Suppose $X_1, X_2, \dots, X_n$ is a random sample from U$(0, \theta)$ distribution. Let $X_{1:n}, \dots, X_{n:n}$ be the corresponding order statistics. Then,

$$T = \frac{X_{n:n}}{\theta}$$

is a pivot (for θ) since its density function is

$$f_T(t) = nt^{n-1}, \; 0 < t < 1,$$

which is free of θ.

REMARK 7

All the pivotal quantities discussed above are all "parametric pivots" as they are pivots for parameters of a specific parametric model assumed.

EXAMPLE 5

Let $X_1, \dots, X_n$ be a random sample from BERN(π) distribution, where $\pi \in (0,1)$ is the probability of success. Then, it is well-known that

$$Y = \sum_{i=1}^{n} X_i \sim \text{BIN}(n, \pi).$$

As the distribution of Y depends on the parameter π, it is not a pivot. In fact, no exact pivot exists in this case.

However, by the use of Central Limit Theorem, it is known that

$$Z = \frac{Y - n\pi}{\sqrt{n\pi(1-\pi)}} \overset{\text{asymp.}}{\sim} \mathcal{N}(0,1).$$

As the distribution of Z is, asymptotically, standard normal, which is free of π, it can serve as a pivot. But, note that it is only an approximate pivot for π.

Pivot for Population Quantile

Suppose we have a random sample from $\mathcal{N}(\mu, \sigma^2)$ distribution, and that we are interested in inferring about p-th quantile $\xi_p = \mu + \sigma z_p$, where z_p is the p-th quantile of the standard normal distribution. Then, it is evident that, with $\bar{X}$ and S as estimates of μ and σ respectively, then an estimate of ξ_p is

$$\hat{\xi}_p = \hat{\mu} + \hat{\sigma} z_p = \bar{X} + z_p S.$$

Then, the variable

$$\begin{aligned} T &= \frac{\hat{\xi}_p - \xi_p}{S} \\ &= \frac{(\bar{X} + z_p S) - (\mu + z_p \sigma)}{S} \\ &= \frac{(\bar{X} - \mu) + z_p (S - \sigma)}{S} \\ &= \frac{\bar{X} - \mu}{S} + z_p \left(1 - \frac{1}{S/\sigma}\right) \end{aligned}$$

is a pivot as its distribution is free of parameters μ and σ. Hence, this pivot could be used for developing inference for the p-th quantile ξ_p of the normal distribution.

EXERCISE 1

Can you think of a way to find its percentage points?

REMARK 8

Though the above derivation is shown for normal distribution, it can be done similarly for any member of location-scale family of distributions like Logistic, Laplace, Gumbel distributions.

Non-parametric Confidence Interval for Quantile

Now, let us assume we have a random sample $X_1, X_2, \dots, X_n$ from a distribution function $F(x)$ that is continuous. Let ξ_p be the p-th quantile of F. Then, $F(\xi_p) = \mathbb{P}(X \leq \xi_p) = p$. We are interested in a confidence interval for ξ_p, but without assuming a specific form of the distribution F, like normal!

Let $X_{1:n}, X_{2:n}, \dots, X_{n:n}$ denote the order statistics obtained from the sample. Let $X_{r:n}$ and $X_{s:n}$ be two selected order statistics, for $1 \leq r < s \leq n$. Then, we have:

$$\begin{aligned} \mathbb{P}(X_{r:n} \leq \xi_p \leq X_{s:n}) &= \mathbb{P}(F(X_{r:n}) \leq F(\xi_p) \leq F(X_{s:n})) \\ &= \mathbb{P}(U_{r:n} \leq p \leq U_{s:n}) \\ &= \mathbb{P}(p \leq U_{s:n}) - \mathbb{P}(p < U_{r:n}) \\ &= 1 - \mathbb{P}(U_{s:n} < p) - \big(1 - \mathbb{P}(U_{r:n} < p)\big) \\ &= \mathbb{P}(U_{r:n} < p) - \mathbb{P}(U_{s:n} < p) \\ &= \sum_{i=r}^{p} \binom{n}{i} p^i (1-p)^{n-i} - \sum_{i=s}^{n} \binom{n}{i} p^i (1-p)^{n-i} \\ &= \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}, \end{aligned}$$

where $U_{r:n}$ and $U_{s:n}$ are order statistics from Uniform$(0, 1)$ distribution. Thus, $(X_{r:n}, X_{s:n})$ is a non-parametric confidence interval for the p-th population quantile ξ_p, with its coverage probability not depending on F, but only on p and n.

So, for a given confidence level $1-\alpha$, all we need to do is, for a given sample size n and the quantile level p, we need to determine integers r and s such that

$$1 \leq r < s \leq n$$

and

$$\sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \approx 1-\alpha.$$

Note that $1-\alpha$ may not be achievable exactly as the binomial distribution is discrete and so has jumps.

REMARK 9

The choice of r and s may not be unique. So, if there is more than one choice of (r, s) satisfying the above conditions, then it would be meaningful to choose that pair (r, s) for which

$$s - r \text{is the smallest}$$

among all these choices satisfying the conditions. This would then correspond to the "narrowest non-parametric confidence interval for population quantile."

LECTURE 7

1st March

EXAMPLE 6: Uniform Distribution

Let $X_1, \ldots, X_n$ be a random sample from Uniform$(0, \theta)$ distribution. Then, we have seen before that $X_{n:n}$ is a sufficient statistic for θ.
Further, the PDF of $T = X_{n:n}$ is

$$f_T(t \mid \theta) = n\big(F(t)\big)^{n-1} f(t) = \frac{nt^{n-1}}{\theta^n}, \ 0 < t < \theta.$$

Now, let $g(\,\cdot\,)$ be a measurable function of t such that $\mathbb{E}_\theta(g(T)) = 0$ for all $\theta > 0$. Then,

$$\mathbb{E}_\theta(g(T)) = \frac{n}{\theta^n} G(\theta), \text{ where } G(\theta) = \int_0^\theta g(t) t^{n-1} \,\mathrm{d}t.$$

Note $G(\theta)$ is differentiable almost everywhere, and further

$$G'(\theta) = g(\theta)\theta^{n-1}.$$

If $G(\theta) = 0$ for all $\theta > 0$, then $G'(\theta) = 0$ for all $\theta > 0$ and so $g(\theta) = 0$ for all $\theta > 0$.
This shows that $T = X_{n:n}$ is a complete sufficient statistic for θ. Since $\mathbb{E}[T] = \mathbb{E}[X_{n:n}] = \frac{n}{n+1}\theta$, we find readily $\frac{n+1}{n}T = \frac{n+1}{n}X_{n:n}$ is an unbiased estimator of θ. Hence, it is an Uniformly Minimum Variance Unbiased Estimator of θ.

Optimal Linear Estimation

The celebrated Gauss-Markov theorem states that the OLS (Ordinary Least Squares) estimator possesses the smallest sampling variance, within the class of all linear unbiased estimators, if the error variables in the linear regression model are uncorrelated and have zero means and equal variances.

Note: The errors need not be normally distributed, nor be independent and identically distributed. They only need to be uncorrelated and with zero mean and homoscedastic (all equal) finite variance.

Note: The result, under the assumptions of normality and independence, was first established by Gauss. Much later, the result was proved by Markov under the weaker conditions of uncorrelated errors (and also without the assumption of normality).

Gauss-Markov Theorem. Consider the linear model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

with $\boldsymbol{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n\times k}$, $\boldsymbol{\beta} \in \mathbb{R}^k$, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$, or equivalently

$$y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i, \text{ for } i = 1, 2, \dots, n,$$

where β_{ij} are non-random parameters that are unobservable. X_{ij} are non-random explanatory variables that are observable, the errors ε_i are random, and consequently the dependent variables y_i $(i = 1, 2, \dots, n)$ are random. ε_i's are commonly referred to as "noise" or "disturbance," but most commonly as "error."

Note: To include a constant in the above model, one can simply take $X_{i0} = 1$ for all $i = 1, \dots, n$, in order to introduce β_0 as the unobservable intercept term. Then, the vector $\boldsymbol{\beta}$ will be $(\beta_0, \beta_1, \dots, \beta_k)^\top_{(k+1)\times 1}$, and $\mathbf{X} \in \mathbb{R}^{n\times(k+1)}$.

Assumptions:

(1) Errors are "centred," i.e., $\mathbb{E}[\varepsilon_i] = 0$ for $i = 1, \dots, n$;

(2) Errors are "homoscedastic," i.e., $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$ for $i = 1, \dots, n$;

(3) Errors are "uncorrelated," i.e., $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. In other words, the variance-covariance matrix of $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^\top$ as $\sigma^2\mathbf{I}$, where $\mathbf{I}$ is an identity matrix of order n.

Then, the OLS (Ordinary Least Squares) estimator of $\boldsymbol{\beta}$ is given by

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

provided $\mathbf{X}^\top\mathbf{X}$ is of full rank, and that this estimator is BLUE (Best Linear Unbiased Estimator).

Preamble: A linear estimator of the parameter β_j is of the form

$$\hat{\beta}_j = c_{1j}y_1 + \cdots + c_{nj}y_n$$

in which the coefficients $c_{1j}, \dots, c_{nj}$ are not allowed to depend on the coefficients β_j as they are unobservable, but are allowed to depend on X_{ij} as they are observable.

The dependence of the coefficients on X_{ij} would typically be non-linear; but the estimator is linear in each y_i and hence in each random error ε_i. This is why we call it "linear regression."

The estimator $\hat{\beta}_j$ will be unbiased if and only if

$$\mathbb{E}[\hat{\beta}_j] = \beta_j$$

regardless of X_{ij}. Now, let $L = \sum_{j=0}^{k} \ell_j\beta_j$ be some linear combination of the coefficients $\beta_0, \dots, \beta_k$. Then, the MSE (mean squared error) of the corresponding estimator $\hat{L}$ is

$$\mathbb{E}\left[\left(\sum_{j=0}^{k}\ell_j\hat{\beta}_j - \sum_{j=0}^{k}\ell_j\beta_j\right)^2\right] = \mathbb{E}\left[\left(\sum_{j=0}^{k}\ell_j(\hat{\beta}_j - \beta_j)\right)^2\right].$$

It is the expectation of the square of the weighted sum (across the parameters) of the differences between the estimators and the corresponding parameters to be estimated.

Observe that as we are considering the case in which all the parameter estimates are unbiased, the MSE is the same as the variance of the estimator of the linear combination L.

The BLUE of $\boldsymbol{\beta} = (\beta_0, \dots, \beta_k)^\top$ is the one in which the smallest MSE for every vector L of the linear combination of parameters. This is equivalent to saying

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}) - \mathrm{Var}(\hat{\boldsymbol{\beta}})$$

is a positive semi-definite matrix for any other linear unbiased estimator $\tilde{\boldsymbol{\beta}}$.

Derivation of OLS estimator: The MSE function that we need to minimize is

$$\begin{aligned} Q(\beta_0, \beta_1, \dots, \beta_k) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_k x_{ik})^2 \\ &= (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})_{1\times n}^{\top} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})_{n\times 1}. \end{aligned}$$

The first derivative is

$$\begin{aligned} \frac{\mathrm{d}Q}{\mathrm{d}\boldsymbol{\beta}} &= -2\mathbf{X}^{\top}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -2 \begin{bmatrix} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}) \\ \sum_{i=1}^{n} x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}) \\ \vdots \\ \sum_{i=1}^{n} x_{ik}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}) \end{bmatrix} \\ &= \mathbf{0}_{(k+1)\times 1}, \end{aligned}$$

where $\mathbf{X}$ is the design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}_{n\times(k+1)} \qquad \in \mathbb{R}^{n\times(k+1)}$$

and $n \geq k+1$ (so that $\mathbf{X}^{\top}\mathbf{X}$ of order $(k+1)\times(k+1)$ can be of full rank).

The solution from the equation

$$-2\mathbf{X}^{\top}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \iff (\mathbf{X}^{\top}\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^{\top}\boldsymbol{y}$$

is clearly

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

where $\mathbf{X}$ is the design matrix of as presented above.

Checking for minimum: The Hessian matrix of second derivatives is readily obtained to be

$$\begin{aligned} \mathcal{H} = \frac{\mathrm{d}^2 Q}{\mathrm{d}\boldsymbol{\beta}^2} &= \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{i1}x_{ik} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{pmatrix} \\ &= 2\mathbf{X}^{\top}\mathbf{X}. \end{aligned}$$

Assume that the columns of the design matrix $\mathbf{X}$ are linearly independent, so that $\mathbf{X}^{\top}\mathbf{X}$ is invertible. Let

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{c}_0 & \boldsymbol{c}_1 & \cdots & \boldsymbol{c}_k \end{bmatrix}.$$

Then,

$$\lambda_0\boldsymbol{c}_0 + \lambda_1\boldsymbol{c}_1 + \cdots + \lambda_k\boldsymbol{c}_k = 0 \iff \lambda_0 = \lambda_1 = \cdots = \lambda_k = 0.$$

Now, let $(\lambda_0, \lambda_1, \dots, \lambda_k) \in \mathbb{R}^{(k+1)\times 1}$ be an eigenvector of the Hessian matrix $\mathcal{H}$. Then,

$$\boldsymbol{\lambda} \neq \mathbf{0} \implies (\lambda_0\boldsymbol{c}_0 + \lambda_1\boldsymbol{c}_1 + \cdots + \lambda_k\boldsymbol{c}_k)^2 > 0.$$

So, we find

$$\begin{aligned}
\begin{bmatrix}\lambda_0 & \lambda_1 & \cdots & \lambda_k\end{bmatrix}\begin{bmatrix}\boldsymbol{c}_0\\ \boldsymbol{c}_1\\ \vdots\\ \boldsymbol{c}_k\end{bmatrix}\begin{bmatrix}\boldsymbol{c}_0 & \boldsymbol{c}_1 & \cdots & \boldsymbol{c}_k\end{bmatrix}\begin{bmatrix}\lambda_0\\ \lambda_1\\ \vdots\\ \lambda_k\end{bmatrix} &= \boldsymbol{\lambda}^\top(\tfrac{1}{2}\mathcal{H})\boldsymbol{\lambda} \qquad \text{since } \mathcal{H} = 2\mathbf{X}^\top\mathbf{X}\\
&= \tfrac{1}{2}\boldsymbol{\lambda}^\top\mathcal{H}\boldsymbol{\lambda}\\
&= \tfrac{1}{2}(\boldsymbol{\lambda}^\top\boldsymbol{\lambda})a \qquad \text{since } \boldsymbol{\lambda} \text{ is a vector of eigenvalues}\\
&> 0,
\end{aligned}$$

where a is the eigenvalue corresponding to the eigenvector $\boldsymbol{\lambda}$. Further,

$$\boldsymbol{\lambda}^\top\boldsymbol{\lambda} = \sum_{i=1}^{k}\lambda_i^2 > 0 \implies a > 0.$$

Finally, as eigenvalue $\boldsymbol{\lambda}$ is arbitrary, all eigenvalues of $\mathcal{H}$ are positive, and so $\mathcal{H}$ is positive definite. Hence, the OLS estimator

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

does indeed correspond to a global minimum.

Uniqueness of the OLS estimator: Assume

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}\boldsymbol{y}$$

be another linear estimator of $\boldsymbol{\beta}$ with

$$\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B},$$

where $\mathbf{B}$ is a $(k+1)\times n$ non-zero matrix. Then, we find

$$\begin{aligned}
\mathbb{E}[\tilde{\boldsymbol{\beta}}] &= \mathbb{E}[\mathbf{A}\boldsymbol{y}]\\
&= \mathbb{E}\big[\big\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B}\big\}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\big]\\
&= \big\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B}\big\}\mathbf{X}\boldsymbol{\beta} + \big\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B}\big\}\,\mathbb{E}[\boldsymbol{\varepsilon}]\\
&= \big\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B}\big\}\mathbf{X}\boldsymbol{\beta} \qquad \text{since } \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}\\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \mathbf{B}\mathbf{X}\boldsymbol{\beta}\\
&= (\mathbf{I}_{(k+1)\times(k+1)} + \mathbf{B}\mathbf{X})\boldsymbol{\beta},
\end{aligned}$$

where $\mathbf{I}_{(k+1)\times(k+1)}$ is an identity matrix of order $(k+1)\times(k+1)$. Thus, $\tilde{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ if and only if

$$\mathbf{B}\mathbf{X} = \mathbf{O}.$$

Now, let us consider

$$\begin{aligned}
\mathrm{Var}(\tilde{\boldsymbol{\beta}}) &= \mathrm{Var}(\mathbf{A}\boldsymbol{y})\\
&= \mathbf{A}\,\mathrm{Var}(\boldsymbol{y})\mathbf{A}^\top\\
&= \mathbf{A}(\sigma^2\mathbf{I})\mathbf{A}^\top\\
&= \sigma^2\mathbf{A}\mathbf{A}^\top\\
&= \sigma^2\big\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B}\big\}\big\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B}\big\}^\top\\
&= \sigma^2\big\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{B}\big\}\big\{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{B}^\top\big\}\\
&= \sigma^2\big\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{B}^\top + \mathbf{B}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{B}\mathbf{B}^\top\big\}\\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} + \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{B}\mathbf{X})^\top + \sigma^2(\mathbf{B}\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1} + \sigma^2\mathbf{B}\mathbf{B}^\top\\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} + \sigma^2\mathbf{B}\mathbf{B}^\top,\ \text{ due to unbiasedness condition}\\
&= \mathrm{Var}(\hat{\boldsymbol{\beta}}) + \sigma^2\mathbf{B}\mathbf{B}^\top,\ \text{ since } \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.
\end{aligned}$$

As $\mathbf{B}\mathbf{B}^\top$ is a positive semidefinite matrix, $\mathrm{Var}(\tilde{\boldsymbol{\beta}})$ exceeds $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ by a positive semidefinite matrix.

Another interpretation for this property: Let $\boldsymbol{\ell}^\top\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\ell}^\top\tilde{\boldsymbol{\beta}}$ be both linear unbiased estimators of $\boldsymbol{\ell}^\top\boldsymbol{\beta}$. Then,

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\ell}^\top\tilde{\boldsymbol{\beta}}) &= \boldsymbol{\ell}^\top\,\mathrm{Var}(\tilde{\boldsymbol{\beta}})\boldsymbol{\ell} \\
&= \sigma^2\boldsymbol{\ell}^\top\big\{(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{B}\mathbf{B}^\top\big\}\boldsymbol{\ell} \\
&= \sigma^2\boldsymbol{\ell}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{\ell} + \sigma^2\boldsymbol{\ell}^\top\mathbf{B}\mathbf{B}^\top\boldsymbol{\ell} \\
&= \mathrm{Var}(\boldsymbol{\ell}^\top\hat{\boldsymbol{\beta}}) + \sigma^2(\mathbf{B}^\top\boldsymbol{\ell})^\top(\mathbf{B}^\top\boldsymbol{\ell}) \\
&= \mathrm{Var}(\boldsymbol{\ell}^\top\hat{\boldsymbol{\beta}}) + \sigma^2\|\mathbf{B}^\top\boldsymbol{\ell}\| \\
&\geq \mathrm{Var}(\boldsymbol{\ell}^\top\hat{\boldsymbol{\beta}}).
\end{aligned}$$

Furthermore, equality holds if and only if $\mathbf{B}^\top\boldsymbol{\ell} = \mathbf{0}$. Then, in this case, we readily find

$$\begin{aligned}
\boldsymbol{\ell}^\top\tilde{\beta} &= \boldsymbol{\ell}^\top\big\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X} + \mathbf{B}\big\}\boldsymbol{y} \\
&= \boldsymbol{\ell}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y} + \boldsymbol{\ell}^\top\mathbf{B}\boldsymbol{y} \\
&= \boldsymbol{\ell}^\top\hat{\boldsymbol{\beta}} + (\mathbf{B}^\top\boldsymbol{\ell})^\top\boldsymbol{y},\ \text{since } \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y} \\
&= \boldsymbol{\ell}^\top\hat{\boldsymbol{\beta}},\ \text{since } \mathbf{B}^\top\boldsymbol{\ell} = \mathbf{0}.
\end{aligned}$$

This establishes that the equality holds if and only if $\boldsymbol{\ell}^\top\tilde{\boldsymbol{\beta}} = \boldsymbol{\ell}^\top\hat{\boldsymbol{\beta}}$, which shows the uniqueness of the OLS estimator as a Best Linear Unbiased Estimator.

Warning: The condition that the estimator be unbiased cannot be dropped, since biased estimators may exist with lower variance (and even with lower MSE).

To see this, consider the following example: Let $X_1, X_2, \ldots, X_n$ be a random sample from EXP(θ) distribution. Then, $\bar{X}$ is an unbiased estimator of θ with

$$\mathbb{E}[\bar{X}] = \theta \quad \text{and} \quad \mathrm{Var}(\bar{X}) = \frac{\theta^2}{n}.$$

Now, consider another linear estimator of θ to be

$$\tilde{\theta} = \sum_{i=1}^{n} X_i a_i.$$

Then, evidently, we have:

$$\begin{aligned}
\mathbb{E}[\tilde{\theta}] &= \sum_{i=1}^{n} a_i\,\mathbb{E}[X_i] = \theta\sum_{i=1}^{n} a_i, \\
\text{Bias}(\tilde{\theta}) &= \mathbb{E}[\tilde{\theta}] - \theta = \theta\left\{\sum_{i=1}^{n} a_i - 1\right\}, \\
\mathrm{Var}(\tilde{\theta}) &= \sum_{i=1}^{n} a_i^2\,\mathrm{Var}(X_i) = \theta^2\sum_{i=1}^{n} a_i^2, \text{MSE}(\tilde{\theta}) &= \mathrm{Var}(\tilde{\theta}) + \big(\text{Bias}(\tilde{\theta})\big)^2 \\
&= \theta^2\left\{\sum_{i=1}^{n} a_i^2 + \left(\sum_{i=1}^{n} a_i - 1\right)^2\right\}.
\end{aligned}$$

To minimize MSE($\tilde{\theta}$), we find the partial derivatives to be

$$\begin{aligned}
\frac{\partial \text{MSE}(\tilde{\theta})}{\partial a_i} &= \theta^2\left\{2a_i + 2\left(\sum_{j=1}^{n} a_j - 1\right)\right\} = 0 \\
\implies a_i &= -\left(\sum_{j=1}^{n} a_j - 1\right),\ i = 1, \ldots, n.
\end{aligned}$$

Adding these n equations,

$$\sum_{j=1}^{n} a_j = -n\Big(\sum_{j=1}^{n} a_j - 1\Big) = -n\sum_{j=1}^{n} a_j + n$$

and so

$$\sum_{j=1}^{n} a_j = \frac{n}{n+1}.$$

This immediately yields

$$a_i = -\Big(\sum_{j=1}^{n} a_j - 1\Big) = -\Big(\frac{n}{n+1} - 1\Big) = \frac{1}{n+1}, \text{ for } i = 1, \ldots, n.$$

Hence, the linear estimator $\tilde{\theta}$ becomes

$$\tilde{\theta} = \sum_{i=1}^{n} a_i X_i = \frac{1}{n+1}\sum_{i=1}^{n} X_i = \frac{n}{n+1}\bar{X}.$$

Observe that

$$\begin{aligned}
\text{Bias}(\tilde{\theta}) &= \frac{n}{n+1}\theta - \theta = -\frac{1}{n+1}\theta, \\
\text{Var}(\tilde{\theta}) &= \frac{n^2}{(n+1)^2}\,\text{Var}(\bar{X}) = \frac{n^2}{(n+1)^2}\frac{\theta^2}{n} = \frac{n}{(n+1)^2}\theta^2 \\
&< \frac{\theta^2}{n},
\end{aligned}$$

which is the variance of $\bar{X}$, the OLS estimator of θ. Furthermore,

$$\begin{aligned}
\text{MSE}(\tilde{\theta}) &= \text{Var}(\tilde{\theta}) + \big(\text{Bias}(\tilde{\theta})\big)^2 \\
&= \frac{n}{(n+1)^2}\theta^2 + \frac{\theta^2}{(n+1)^2} = \frac{\theta^2}{n+1} \\
&< \frac{\theta^2}{n},
\end{aligned}$$

the variance of $\bar{X}$ (the OLS estimator). This is an example to demonstrate why unbiasedness is needed in the Gauss-Markov theorem.

Things to keep in mind:

(1) In most applications of OLS, the regressors (parameters of interest) in the design matrix $\mathbf{X}$ are assumed to be fixed in repeated samples. This assumption may not be reasonable in some cases (like Econometrics). Instead, there are assumptions of Gauss-Markov theorem are stated conditional on $\mathbf{X}$;

(2) $\boldsymbol{y}$ is assumed to be a linear function of the variables specified in the model. But, the specification must be linear in its parameters; it does not mean that there must be a linear relationship between $\boldsymbol{y}$ and the independent variables $\mathbf{X}$. The independent variables can take on non-linear forms as long as the parameters are linear;

(3) Data transformations can be made use to convert an equation into a linear form. For example, "the power law model" of the form

$$y = \alpha X^{\beta} e^{\varepsilon}$$

can be transformed, by natural logarithms, to

$$\ln(y) = \ln(\alpha) + \beta\ln(X) + \varepsilon.$$

Similarly, the "Cobb-Douglas model" of the form

$$y = AL^{\alpha}K^{\alpha}$$

can be transformed, by natural logarithms, to

$$\ln(y) = \ln(A) + \alpha\ln(L) + (1-\alpha)\ln(K) + \varepsilon;$$

remember the parameters that minimize the residuals of the transformed model would not minimize the residuals of the original model;

(4) Recall we assumed that $\mathbf{X}$ must have full column rank, i.e.,

$$\operatorname{rank}(\mathbf{X}) = k+1;$$

otherwise, $\mathbf{X}^{\top}\mathbf{X}$ is not invertible and consequently the OLS estimator cannot be computed.

This gets violated in case of "perfect multicollinearity," a case when some explanatory variables are linearly dependent. This can happen, for example, in a "dummy variable trap," when a base dummy variable is not omitted resulting in perfect correlation between the dummy variable and the constant term.

Interestingly, "near multicollinearity" would result in unbiased estimation, even though with much lesser precision. The estimators would be less efficient and very sensitive to particular sets of data.

Multicollinearity can be detected from "variance inflation factor" (which is the variance of estimating a parameter in a model that includes multiple other parameters divided by the variance of a model using only that variable), or "conditional number" (which measures how much the output value can change for a small change in the input argument);

(5) "Spherical errors model" means that the outer product of the error vector must be spherical, i.e.,

$$\mathbb{E}[\boldsymbol{\varepsilon}^{\top}\boldsymbol{\varepsilon} \mid \mathbf{X}] = \operatorname{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix}_{n\times n} = \sigma^2\mathbf{I}, \text{ with } \sigma^2 > 0.$$

This means that the error term has uniform variance (i.e., homoscedasticity) and no serial dependence. Recall that in the case of multivariate normal distribution for $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, the equation $f(\boldsymbol{\varepsilon}) = c$ will correspond to a ball centred at $\mathbf{0}$ with radius σ in $\mathbb{R}^n$.

Homoscedasticity will get violated when there is "autocorrelation." In this case, the OLS estimator is still unbiased, but is inefficient.

When there is "spherical errors," the OLS estimator remains BLUE;

(6) Finally, it may be observed that the expectation of errors, conditioned on the regressors, is

$$\mathbb{E}[\varepsilon_i \mid \boldsymbol{X}] = \mathbb{E}[\varepsilon_i \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] = \mathbf{0},$$

where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})^{\top}$ is the data vector of regressors from the i^{th} observation, and then

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^{\top} & \boldsymbol{x}_2^{\top} & \cdots & \boldsymbol{x}_n^{\top} \end{pmatrix}^{\top}$$

is the design matrix (on the data matrix). So, the above statement means that $\boldsymbol{x}_i$ and ε_i are orthogonal to each other, i.e., their inner product

$$\mathbb{E}[\boldsymbol{x}_j \cdot \varepsilon_i] = \begin{pmatrix} \mathbb{E}[1 \cdot \varepsilon_i] \\ \mathbb{E}[x_{j1} \cdot \varepsilon_i] \\ \vdots \\ \mathbb{E}[x_{jk} \cdot \varepsilon_i] \end{pmatrix} = \mathbf{0}, \text{ for all } i, j.$$

This will get violated if the explanatory variables are stochastic, for example, when they are measured with error, or when they are "endogenous" (i.e., when an explanatory variable is correlated with error variable).

Instrument variable methods become useful in this case!