Advanced Regression Methods for Independent Data

STAT/BIOST 570, 2020

Regression Models with Weaker Assumptions

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY *of* WASHINGTON

Regression Models with Weaker Assumptions

We now study certain types of inference under weaker sets of assumptions:

- Least squares without assuming a fully parametric model for $Y \mid \boldsymbol{x}$
- Assuming only a regression and variance function for $Y \mid \boldsymbol{x}$
- Assuming only a regression function for $Y \mid \boldsymbol{x}$
- What happens with fully parametric models when they are wrong?

We first study some key results from *estimating equations*[1]

[1] See Chapter 7 of Boos and Stefanski (2013) *Essential Statistical Inference*, Springer.

Estimating Equations

Estimating equations provide a very general framework for deriving estimators with desirable properties

- We assume that Y_i, $i = 1, \ldots, n$, are i.i.d. from distribution F
- We seek to estimate

$$\theta_0 : \quad \mathrm{E}_F[\boldsymbol{G}(\theta_0, Y)] = 0,$$

  where

  - $Y \sim F$: generic random variable (vector)
  - $\boldsymbol{\theta}$: vector of p parameters
  - $\boldsymbol{G}$: p-dimensional, continuously differentiable function of $\boldsymbol{\theta}$ and Y

Estimating Equations

Estimating equations provide a very general framework for deriving estimators with desirable properties

- We assume that Y_i, $i = 1, \ldots, n$, are i.i.d. from distribution F
- We seek to estimate

$$\boldsymbol{\theta}_0 : \quad \mathrm{E}_F[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)] = 0,$$

  where

  - $Y \sim F$: generic random variable (vector)
  - $\boldsymbol{\theta}$: vector of p parameters
  - $\boldsymbol{G}$: p-dimensional, continuously differentiable function of $\boldsymbol{\theta}$ and Y

Estimating Equations

$$\boldsymbol{G}(\boldsymbol{\theta}, Y) = \begin{pmatrix} G_1(\boldsymbol{\theta}, Y) \\ \vdots \\ G_p(\boldsymbol{\theta}, Y) \end{pmatrix} : \ \Theta \times \mathcal{Y} \to \mathbb{R}^p$$

Estimating Equations

- Given the data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, an *estimating function* is defined as
$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i)$$

- If we have that
$$\mathrm{E}[\boldsymbol{G}_n(\boldsymbol{\theta}_0)] = \mathbf{0},$$
we refer to $\boldsymbol{G}_n(\boldsymbol{\theta})$ as an *unbiased* estimating function, as it's unbiased for estimating $\mathbf{0}$

Estimating Equations

- Given the data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, an *estimating function* is defined as
$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i)$$

- If we have that
$$\mathrm{E}[\boldsymbol{G}_n(\boldsymbol{\theta}_0)] = \mathbf{0},$$
we refer to $\boldsymbol{G}_n(\boldsymbol{\theta})$ as an *unbiased* estimating function, as it's unbiased for estimating $\mathbf{0}$

Estimating Equations

- The corresponding *estimating equation* is given by

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \mathbf{0}$$

- The solution to this system of equations

$$\hat{\boldsymbol{\theta}}_n \quad : \quad \boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\hat{\boldsymbol{\theta}}_n, Y_i) = \mathbf{0}$$

is often referred to as *M-estimator* or *Z-estimator*

Estimating Equations

- The corresponding *estimating equation* is given by

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \mathbf{0}$$

- The solution to this system of equations

$$\hat{\boldsymbol{\theta}}_n \quad : \quad \boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\hat{\boldsymbol{\theta}}_n, Y_i) = \mathbf{0}$$

is often referred to as *M-estimator* or *Z-estimator*

M-Estimators

M-estimator is the term used when the above formulation is derived from a **M**aximization problem

- For known function h, say we seek to maximize

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{\theta}, Y_i)$$

- Often we find the maximizer by taking derivatives and setting them equal to zero

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(\boldsymbol{\theta}, Y_i)}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

  in which case $\boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \partial h(\boldsymbol{\theta}, Y_i)/\partial \boldsymbol{\theta}$

- Examples: maximum likelihood estimation: $h(\boldsymbol{\theta}, Y_i) = \log p(Y_i \mid \boldsymbol{\theta})$

M-Estimators

M-estimator is the term used when the above formulation is derived from a **M**aximization problem

- For known function h, say we seek to maximize

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{\theta}, Y_i)$$

- Often we find the maximizer by taking derivatives and setting them equal to zero

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(\boldsymbol{\theta}, Y_i)}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

  in which case $\boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \partial h(\boldsymbol{\theta}, Y_i)/\partial \boldsymbol{\theta}$

- Examples: maximum likelihood estimation: $h(\boldsymbol{\theta}, Y_i) = \log p(Y_i \mid \boldsymbol{\theta})$

M-Estimators

M-estimator is the term used when the above formulation is derived from a **M**aximization problem

- For known function h, say we seek to maximize

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{\theta}, Y_i)$$

- Often we find the maximizer by taking derivatives and setting them equal to zero

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(\boldsymbol{\theta}, Y_i)}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

  in which case $\boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \partial h(\boldsymbol{\theta}, Y_i)/\partial \boldsymbol{\theta}$

- Examples: maximum likelihood estimation: $h(\boldsymbol{\theta}, Y_i) = \log p(Y_i \mid \boldsymbol{\theta})$

Z-Estimators

- However, the theory of estimating equations does not care about where the estimating equation comes from
- *Z-estimator* is the term used more generally, as the estimator is derived as a **Z**ero or solution to an equation
- Example: method of moments:

$$\boldsymbol{G}(\theta, Y_i) = \begin{pmatrix} Y_i - \mathrm{E}[Y_i \mid \theta] \\ Y_i^2 - \mathrm{E}[Y_i^2 \mid \theta] \\ \vdots \\ Y_i^p - \mathrm{E}[Y_i^p \mid \theta] \end{pmatrix}$$

when θ is p-dimensional

Z-Estimators

- However, the theory of estimating equations does not care about where the estimating equation comes from
- *Z-estimator* is the term used more generally, as the estimator is derived as a **Z**ero or solution to an equation
- Example: method of moments:

$$\boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \begin{pmatrix} Y_i - \mathrm{E}[Y_i \mid \boldsymbol{\theta}] \\ Y_i^2 - \mathrm{E}[Y_i^2 \mid \boldsymbol{\theta}] \\ \vdots \\ Y_i^p - \mathrm{E}[Y_i^p \mid \boldsymbol{\theta}] \end{pmatrix}$$

when $\boldsymbol{\theta}$ is p-dimensional

Z-Estimators

- However, the theory of estimating equations does not care about where the estimating equation comes from
- *Z-estimator* is the term used more generally, as the estimator is derived as a **Z**ero or solution to an equation
- Example: method of moments:

$$\boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \begin{pmatrix} Y_i - \mathrm{E}[Y_i \mid \boldsymbol{\theta}] \\ Y_i^2 - \mathrm{E}[Y_i^2 \mid \boldsymbol{\theta}] \\ \vdots \\ Y_i^p - \mathrm{E}[Y_i^p \mid \boldsymbol{\theta}] \end{pmatrix}$$

when $\boldsymbol{\theta}$ is p-dimensional

Comments on Parameters

- A parameter θ is a characteristic of a probability distribution F (a super-population)
- If I give you F, then you would know how to compute θ (in principle)
- A parameter θ can be seen as a functional that maps distributions to, say, the real line, and we often write $\theta := \theta(F)$
  - The median is defined as (similarly for other quantiles)

$$m : \; F\{(-\infty, m]\} = 1/2$$

  - If F has a pdf $f(\cdot)$, then

$$\int h(y) f(y) \, dy$$

    is a parameter for any integrable function h (in particular, moments)

Comments on Parameters

- A parameter θ is a characteristic of a probability distribution F (a super-population)

- If I give you F, then you would know how to compute θ (in principle)

- A parameter θ can be seen as a functional that maps distributions to, say, the real line, and we often write $\theta := \theta(F)$

  - The median is defined as (similarly for other quantiles)

$$m : \;\; F\{(-\infty, m]\} = 1/2$$

  - If F has a pdf $f(\cdot)$, then

$$\int h(y) f(y) \; dy$$

    is a parameter for any integrable function h (in particular, moments)

Comments on Parameters

- A parameter θ is a characteristic of a probability distribution F (a super-population)
- If I give you F, then you would know how to compute θ (in principle)
- A parameter θ can be seen as a functional that maps distributions to, say, the real line, and we often write $\theta := \theta(F)$
  - The median is defined as (similarly for other quantiles)
    $$m : \ F\{(-\infty, m]\} = 1/2$$
  - If F has a pdf $f(\cdot)$, then
    $$\int h(y) f(y) \, dy$$
    is a parameter for any integrable function h (in particular, moments)

Comments on Parameters

- A parameter θ is a characteristic of a probability distribution F (a super-population)
- If I give you F, then you would know how to compute θ (in principle)
- A parameter θ can be seen as a functional that maps distributions to, say, the real line, and we often write $\theta := \theta(F)$
  - The median is defined as (similarly for other quantiles)

$$m : \ F\{(-\infty, m]\} = 1/2$$

  - If F has a pdf $f(\cdot)$, then

$$\int h(y) f(y) \, dy$$

is a parameter for any integrable function h (in particular, moments)

Comments on Parameters

- A parameter θ is a characteristic of a probability distribution F (a super-population)
- If I give you F, then you would know how to compute θ (in principle)
- A parameter θ can be seen as a functional that maps distributions to, say, the real line, and we often write $\theta := \theta(F)$
  - The median is defined as (similarly for other quantiles)

    $$m : \; F\{(-\infty, m]\} = 1/2$$

  - If F has a pdf $f(\cdot)$, then

    $$\int h(y) f(y) \; dy$$

    is a parameter for any integrable function h (in particular, moments)

Comments on Parameters

- In parametric models, a parameter θ appears explicitly in the functional form of the pdf $f(\cdot)$ of F

- With estimating equations, however, we can define parameters based on

$$\boldsymbol{\theta}_0 : \quad \mathrm{E}_F[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)] = 0,$$

  where E_F means that the expected value is taken w.r.t. Y, where $Y \sim F$

- Ultimately, we can think that $\boldsymbol{\theta}_0$ is a function of F, i.e., $\boldsymbol{\theta}_0 := \boldsymbol{\theta}_0(F)$

- *The power of estimating functions*: we can estimate $\boldsymbol{\theta}_0$ without having to fully specify a model for F

Comments on Parameters

- In parametric models, a parameter θ appears explicitly in the functional form of the pdf $f(\cdot)$ of F
- With estimating equations, however, we can define parameters based on
$$\boldsymbol{\theta}_0 : \quad \mathrm{E}_F[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)] = 0,$$
where E_F means that the expected value is taken w.r.t. Y, where $Y \sim F$
- Ultimately, we can think that $\boldsymbol{\theta}_0$ is a function of F, i.e., $\boldsymbol{\theta}_0 := \boldsymbol{\theta}_0(F)$
- *The power of estimating functions*: we can estimate θ_0 without having to fully specify a model for F

Comments on Parameters

- In parametric models, a parameter θ appears explicitly in the functional form of the pdf $f(\cdot)$ of F
- With estimating equations, however, we can define parameters based on
$$\boldsymbol{\theta}_0 : \quad \mathrm{E}_F[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)] = 0,$$
where E_F means that the expected value is taken w.r.t. Y, where $Y \sim F$
- Ultimately, we can think that $\boldsymbol{\theta}_0$ is a function of F, i.e., $\boldsymbol{\theta}_0 := \boldsymbol{\theta}_0(F)$
- *The power of estimating functions*: we can estimate $\boldsymbol{\theta}_0$ without having to fully specify a model for F

Consistency of Z-estimators

Result (Consistency of Z-estimators): Suppose that $\hat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation $\boldsymbol{G}_n(\boldsymbol{\theta}) = \mathbf{0}$, i.e. $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Then $\hat{\boldsymbol{\theta}}_n \rightarrow_p \boldsymbol{\theta}_0$.

See Theorem 7.1 of Boos and Stefanski (2013), similar to the proof of consistency of MLEs.

Consistency of Z-estimators

Result (Consistency of Z-estimators): Suppose that $\hat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation $\boldsymbol{G}_n(\boldsymbol{\theta}) = \mathbf{0}$, i.e. $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Then $\hat{\boldsymbol{\theta}}_n \to_p \boldsymbol{\theta}_0$.

See Theorem 7.1 of Boos and Stefanski (2013), similar to the proof of consistency of MLEs.

Asymptotic Normality of Z-Estimators

Result (Asymptotic Normality of Z-estimators): Suppose that $\hat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation $\boldsymbol{G}_n(\boldsymbol{\theta}) = \mathbf{0}$, i.e. $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Then

$$\sqrt{n}\ (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \quad \to_d \quad \mathrm{N}_p[\ \mathbf{0},\ \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^\top\ ]$$

where $\boldsymbol{A} = \boldsymbol{A}(\theta_0)$ with

$$\boldsymbol{A}(\theta) = \mathrm{E}\left[\boldsymbol{G}'(\theta, Y)\right] = \mathrm{E}\left\{\left[\frac{\partial}{\partial\theta_1}\boldsymbol{G}(\theta, Y), \ldots, \frac{\partial}{\partial\theta_p}\boldsymbol{G}(\theta, Y)\right]_{p\times p}\right\}$$

and

$$\boldsymbol{B} \quad = \quad \boldsymbol{B}(\theta_0) = \mathrm{E}[\boldsymbol{G}(\theta_0, Y)\boldsymbol{G}(\theta_0, Y)^\top] = \mathrm{cov}\{\boldsymbol{G}(\theta_0, Y)\}.$$

Proof outline: derive asymptotic properties of the estimating function, and then transfer these to the estimator.

Remark: Boos and Stefanski (2013) define $\boldsymbol{A}(\theta) = \mathrm{E}\left[-\boldsymbol{G}'(\theta, Y)\right]$, but the minus cancels in $\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^\top$

Asymptotic Normality of Z-Estimators

Result (Asymptotic Normality of Z-estimators): Suppose that $\hat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation $\boldsymbol{G}_n(\boldsymbol{\theta}) = \mathbf{0}$, i.e. $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Then

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \quad \to_d \quad \mathrm{N}_p[\,\mathbf{0},\ \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^\top\,]$$

where $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{\theta}_0)$ with

$$\boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[\boldsymbol{G}'(\boldsymbol{\theta}, Y)\right] = \mathrm{E}\left\{\left[\frac{\partial}{\partial\theta_1}\boldsymbol{G}(\boldsymbol{\theta}, Y), \ldots, \frac{\partial}{\partial\theta_p}\boldsymbol{G}(\boldsymbol{\theta}, Y)\right]_{p\times p}\right\}$$

and

$$\boldsymbol{B} \quad = \quad \boldsymbol{B}(\boldsymbol{\theta}_0) = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\boldsymbol{G}(\boldsymbol{\theta}_0, Y)^\top] = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\}.$$

Proof outline: derive asymptotic properties of the estimating function, and then transfer these to the estimator.

Remark: Boos and Stefanski (2013) define $\boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[-\boldsymbol{G}'(\boldsymbol{\theta}, Y)\right]$, but the minus cancels in $\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^\top$

Asymptotic Normality of Z-Estimators

Result (Asymptotic Normality of Z-estimators): Suppose that $\hat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation $\boldsymbol{G}_n(\boldsymbol{\theta}) = \mathbf{0}$, i.e. $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Then

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \quad \to_d \quad \mathrm{N}_p[\,\mathbf{0},\ \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^{\mathsf{T}}\,]$$

where $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{\theta}_0)$ with

$$\boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[\boldsymbol{G}'(\boldsymbol{\theta}, Y)\right] = \mathrm{E}\left\{\left[\frac{\partial}{\partial\theta_1}\boldsymbol{G}(\boldsymbol{\theta}, Y), \ldots, \frac{\partial}{\partial\theta_p}\boldsymbol{G}(\boldsymbol{\theta}, Y)\right]_{p\times p}\right\}$$

and

$$\boldsymbol{B} \quad = \quad \boldsymbol{B}(\boldsymbol{\theta}_0) = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\boldsymbol{G}(\boldsymbol{\theta}_0, Y)^{\mathsf{T}}] = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\}.$$

Proof outline: derive asymptotic properties of the estimating function, and then transfer these to the estimator.

Remark: Boos and Stefanski (2013) define $\boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[-\boldsymbol{G}'(\boldsymbol{\theta}, Y)\right]$, but the minus cancels in $\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^{\mathsf{T}}$

Asymptotic Normality of Z-Estimators

Result (Asymptotic Normality of Z-estimators): Suppose that $\hat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation $\boldsymbol{G}_n(\boldsymbol{\theta}) = \mathbf{0}$, i.e. $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Then

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \quad \to_d \quad \mathrm{N}_p[\, \mathbf{0},\; \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^{\intercal}\,]$$

where $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{\theta}_0)$ with

$$\boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[\boldsymbol{G}'(\boldsymbol{\theta}, Y)\right] = \mathrm{E}\left\{ \left[\frac{\partial}{\partial\theta_1}\boldsymbol{G}(\boldsymbol{\theta}, Y), \ldots, \frac{\partial}{\partial\theta_p}\boldsymbol{G}(\boldsymbol{\theta}, Y) \right]_{p\times p} \right\}$$

and

$$\boldsymbol{B} \;=\; \boldsymbol{B}(\boldsymbol{\theta}_0) = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\boldsymbol{G}(\boldsymbol{\theta}_0, Y)^{\intercal}] = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\}.$$

Proof outline: derive asymptotic properties of the estimating function, and then transfer these to the estimator.

Remark: Boos and Stefanski (2013) define $\boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[-\boldsymbol{G}'(\boldsymbol{\theta}, Y)\right]$, but the minus cancels in $\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^{\intercal}$

Asymptotic Normality of Z-Estimators

Proof outline: From a Taylor series expansion around $\boldsymbol{\theta}_0$ we get:

$$\mathbf{0} = \boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{G}_n(\boldsymbol{\theta}_0) + \boldsymbol{G}'_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \boldsymbol{R}_n$$

which implies

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \left[-\boldsymbol{G}'_n(\boldsymbol{\theta}_0)\right]^{-1}\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) + \sqrt{n}\boldsymbol{R}^*_n$$

As $n \to \infty$, by the WLLN,

$$\boldsymbol{G}'_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}'(\boldsymbol{\theta}_0, Y_i) \ \xrightarrow{p}\ \mathrm{E}\left[\boldsymbol{G}'(\boldsymbol{\theta}_0, Y)\right] = \boldsymbol{A}(\boldsymbol{\theta}_0)$$

and by the CLT,

$$\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}_p[\ \mathbf{0},\ \ \boldsymbol{B}(\boldsymbol{\theta}_0)\ ]$$

where $\boldsymbol{B}(\boldsymbol{\theta}_0) = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\} = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\boldsymbol{G}(\boldsymbol{\theta}_0, Y)^{\mathsf{T}}]$.

Showing conditions under which $\sqrt{n}\boldsymbol{R}^*_n = o_p(1)$ is 580's material.

Paste these results with the help of Slutsky and you get the result.

Asymptotic Normality of Z-Estimators

Proof outline: From a Taylor series expansion around $\boldsymbol{\theta}_0$ we get:

$$\mathbf{0} = \boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{G}_n(\boldsymbol{\theta}_0) + \boldsymbol{G}_n'(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \boldsymbol{R}_n$$

which implies

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = [-\boldsymbol{G}_n'(\boldsymbol{\theta}_0)]^{-1}\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) + \sqrt{n}\boldsymbol{R}_n^*$$

As $n \to \infty$, by the WLLN,

$$\boldsymbol{G}_n'(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}'(\boldsymbol{\theta}_0, Y_i) \xrightarrow{p} \mathrm{E}\left[\boldsymbol{G}'(\boldsymbol{\theta}_0, Y)\right] = \boldsymbol{A}(\boldsymbol{\theta}_0)$$

and by the CLT,

$$\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}_p[\ \mathbf{0},\ \boldsymbol{B}(\boldsymbol{\theta}_0)\ ]$$

where $\boldsymbol{B}(\boldsymbol{\theta}_0) = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\} = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\boldsymbol{G}(\boldsymbol{\theta}_0, Y)^{\mathsf{T}}]$.

Showing conditions under which $\sqrt{n}\boldsymbol{R}_n^* = o_p(1)$ is 580's material.

Paste these results with the help of Slutsky and you get the result.

Asymptotic Normality of Z-Estimators

Proof outline: From a Taylor series expansion around $\boldsymbol{\theta}_0$ we get:

$$\mathbf{0} = \boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{G}_n(\boldsymbol{\theta}_0) + \boldsymbol{G}_n'(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \boldsymbol{R}_n$$

which implies

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = [-\boldsymbol{G}_n'(\boldsymbol{\theta}_0)]^{-1}\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) + \sqrt{n}\boldsymbol{R}_n^*$$

As $n \to \infty$, by the WLLN,

$$\boldsymbol{G}_n'(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}'(\boldsymbol{\theta}_0, Y_i) \ \xrightarrow{p}\ \mathrm{E}\left[\boldsymbol{G}'(\boldsymbol{\theta}_0, Y)\right] = \boldsymbol{A}(\boldsymbol{\theta}_0)$$

and by the CLT,

$$\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}_p[\ \mathbf{0},\ \boldsymbol{B}(\boldsymbol{\theta}_0)\ ]$$

where $\boldsymbol{B}(\boldsymbol{\theta}_0) = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\} = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\boldsymbol{G}(\boldsymbol{\theta}_0, Y)^{\mathsf{T}}]$.

Showing conditions under which $\sqrt{n}\boldsymbol{R}_n^* = o_p(1)$ is 580's material.

Paste these results with the help of Slutsky and you get the result.

Asymptotic Normality of Z-Estimators

Proof outline: From a Taylor series expansion around $\boldsymbol{\theta}_0$ we get:

$$\mathbf{0} = \boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{G}_n(\boldsymbol{\theta}_0) + \boldsymbol{G}'_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \boldsymbol{R}_n$$

which implies

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = [-\boldsymbol{G}'_n(\boldsymbol{\theta}_0)]^{-1}\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) + \sqrt{n}\boldsymbol{R}^*_n$$

As $n \to \infty$, by the WLLN,

$$\boldsymbol{G}'_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}'(\boldsymbol{\theta}_0, Y_i) \ \xrightarrow{p} \ \mathrm{E}\left[\boldsymbol{G}'(\boldsymbol{\theta}_0, Y)\right] = \boldsymbol{A}(\boldsymbol{\theta}_0)$$

and by the CLT,

$$\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}_p[\ \mathbf{0},\ \ \boldsymbol{B}(\boldsymbol{\theta}_0)\ ]$$

where $\boldsymbol{B}(\boldsymbol{\theta}_0) = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\} = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\boldsymbol{G}(\boldsymbol{\theta}_0, Y)^{\mathsf{T}}]$.

Showing conditions under which $\sqrt{n}\boldsymbol{R}^*_n = o_p(1)$ is 580's material.

Paste these results with the help of Slutsky and you get the result.

Asymptotic Normality of Z-Estimators

Proof outline: From a Taylor series expansion around $\boldsymbol{\theta}_0$ we get:

$$\mathbf{0} = \boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{G}_n(\boldsymbol{\theta}_0) + \boldsymbol{G}'_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \boldsymbol{R}_n$$

which implies

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = [-\boldsymbol{G}'_n(\boldsymbol{\theta}_0)]^{-1}\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) + \sqrt{n}\boldsymbol{R}^*_n$$

As $n \to \infty$, by the WLLN,

$$\boldsymbol{G}'_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}'(\boldsymbol{\theta}_0, Y_i) \stackrel{p}{\to} \mathrm{E}\left[\boldsymbol{G}'(\boldsymbol{\theta}_0, Y)\right] = \boldsymbol{A}(\boldsymbol{\theta}_0)$$

and by the CLT,

$$\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) \stackrel{d}{\to} \mathrm{N}_p[\ \mathbf{0},\ \boldsymbol{B}(\boldsymbol{\theta}_0)\ ]$$

where $\boldsymbol{B}(\boldsymbol{\theta}_0) = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\} = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\boldsymbol{G}(\boldsymbol{\theta}_0, Y)^{\mathsf{T}}]$.

Showing conditions under which $\sqrt{n}\boldsymbol{R}^*_n = o_p(1)$ is 580's material.

Paste these results with the help of Slutsky and you get the result.

Asymptotic Normality of Z-Estimators

Proof outline: From a Taylor series expansion around $\boldsymbol{\theta}_0$ we get:

$$\mathbf{0} = \boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{G}_n(\boldsymbol{\theta}_0) + \boldsymbol{G}_n'(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \boldsymbol{R}_n$$

which implies

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = [-\boldsymbol{G}_n'(\boldsymbol{\theta}_0)]^{-1}\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) + \sqrt{n}\boldsymbol{R}_n^*$$

As $n \to \infty$, by the WLLN,

$$\boldsymbol{G}_n'(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}'(\boldsymbol{\theta}_0, Y_i) \ \xrightarrow{p}\ \mathrm{E}\left[\boldsymbol{G}'(\boldsymbol{\theta}_0, Y)\right] = \boldsymbol{A}(\boldsymbol{\theta}_0)$$

and by the CLT,

$$\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}_p[\ \mathbf{0},\ \ \boldsymbol{B}(\boldsymbol{\theta}_0)\ ]$$

where $\boldsymbol{B}(\boldsymbol{\theta}_0) = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\} = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}_0, Y)\boldsymbol{G}(\boldsymbol{\theta}_0, Y)^\mathsf{T}]$.

Showing conditions under which $\sqrt{n}\boldsymbol{R}_n^* = o_p(1)$ is 580's material.

Paste these results with the help of Slutsky and you get the result.

Comments on Estimating Functions

This result is *very* important!

- Using only the definition of $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n$, it tells us that as $n \to \infty$, $\hat{\boldsymbol{\theta}}_n$ is approximately normally distributed, around the 'right' mean $\boldsymbol{\theta}_0$
- The covariance $\mathrm{cov}(\hat{\boldsymbol{\theta}}_n) \approx \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^{\mathsf{T}}/n$ is known as the *sandwich formula*, and it goes to $\mathbf{0}_{p \times p}$ as $n \to \infty$

Comments on Estimating Functions

This result is *very* important!

- Using only the definition of $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n$, it tells us that as $n \to \infty$, $\hat{\boldsymbol{\theta}}_n$ is approximately normally distributed, around the 'right' mean $\boldsymbol{\theta}_0$

- The covariance $\text{cov}(\hat{\boldsymbol{\theta}}_n) \approx \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^{\mathsf{T}}/n$ is known as the *sandwich formula*, and it goes to $\mathbf{0}_{p\times p}$ as $n \to \infty$

Comments on Estimating Functions

This result is *very* important!

- It holds under very mild conditions; in particular, there is no assumed parametric model, and therefore no likelihood
- Whether asymptotic normality is accurate depends on the accuracy of the Central Limit Theorem: in many cases OK for n in the hundreds
- The $\boldsymbol{A}(\boldsymbol{\theta})$ and $\boldsymbol{B}(\boldsymbol{\theta})$ matrices are expectations over unknown distribution F. Different ways of evaluating these expectations lead to different estimating approaches!
- Note that $\boldsymbol{\theta}$ is unknown but we know that $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}_0$ (more on this later)

Comments on Estimating Functions

This result is *very* important!

- It holds under very mild conditions; in particular, there is no assumed parametric model, and therefore no likelihood
- Whether asymptotic normality is accurate depends on the accuracy of the Central Limit Theorem: in many cases OK for n in the hundreds
- The $\boldsymbol{A}(\boldsymbol{\theta})$ and $\boldsymbol{B}(\boldsymbol{\theta})$ matrices are expectations over unknown distribution F. Different ways of evaluating these expectations lead to different estimating approaches!
- Note that $\boldsymbol{\theta}$ is unknown but we know that $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}_0$ (more on this later)

Comments on Estimating Functions

This result is *very* important!

- It holds under very mild conditions; in particular, there is no assumed parametric model, and therefore no likelihood
- Whether asymptotic normality is accurate depends on the accuracy of the Central Limit Theorem: in many cases OK for n in the hundreds
- The $\boldsymbol{A}(\boldsymbol{\theta})$ and $\boldsymbol{B}(\boldsymbol{\theta})$ matrices are expectations over unknown distribution F. Different ways of evaluating these expectations lead to different estimating approaches!
- Note that $\boldsymbol{\theta}$ is unknown but we know that $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}_0$ (more on this later)

Comments on Estimating Functions

This result is *very* important!

- It holds under very mild conditions; in particular, there is no assumed parametric model, and therefore no likelihood
- Whether asymptotic normality is accurate depends on the accuracy of the Central Limit Theorem: in many cases OK for n in the hundreds
- The $\boldsymbol{A}(\boldsymbol{\theta})$ and $\boldsymbol{B}(\boldsymbol{\theta})$ matrices are expectations over unknown distribution F. Different ways of evaluating these expectations lead to different estimating approaches!
- Note that $\boldsymbol{\theta}$ is unknown but we know that $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}_0$ (more on this later)

Example: Mean

- Taking $\boldsymbol{G}(\theta, Y) = Y - \theta$, we obtain

$$\mathrm{E}_F[\boldsymbol{G}(\theta_0, Y)] = 0 \implies \mathrm{E}_F(Y) = \theta_0$$

  so, for any distribution F, θ_0 will represent the mean

- The Z-estimator based on $\{Y_i\}_{i=1}^n \overset{i.i.d.}{\sim} F$ is obtained from

$$\boldsymbol{G}_n(\theta, Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta) = 0 \implies \hat{\theta} = \bar{Y}$$

- On the other hand,

$$\boldsymbol{A}(\theta) = \mathrm{E}_F\left[\boldsymbol{G}'(\theta, Y)\right] = -1$$

  and

$$\boldsymbol{B}(\theta) = \mathrm{E}_F[\boldsymbol{G}(\theta, Y)\boldsymbol{G}(\theta, Y)^{\mathsf{T}}] = \mathrm{var}_F(Y)$$

- We finally have that

$$\bar{Y} \approx N[\,\mathrm{E}_F(Y),\ \mathrm{var}_F(Y)/n\,]$$

Example: Mean

- Taking $\boldsymbol{G}(\theta, Y) = Y - \theta$, we obtain

$$\mathrm{E}_F[\boldsymbol{G}(\theta_0, Y)] = 0 \implies \mathrm{E}_F(Y) = \theta_0$$

so, for any distribution F, θ_0 will represent the mean

- The Z-estimator based on $\{Y_i\}_{i=1}^n \overset{i.i.d.}{\sim} F$ is obtained from

$$\boldsymbol{G}_n(\theta, Y) = \frac{1}{n}\sum_{i=1}^n (Y_i - \theta) = 0 \implies \hat{\theta} = \bar{Y}$$

- On the other hand,

$$\boldsymbol{A}(\theta) = \mathrm{E}_F\left[\boldsymbol{G}'(\theta, Y)\right] = -1$$

and

$$\boldsymbol{B}(\theta) = \mathrm{E}_F[\boldsymbol{G}(\theta, Y)\boldsymbol{G}(\theta, Y)^{\mathsf{T}}] = \mathrm{var}_F(Y)$$

- We finally have that

$$\bar{Y} \approx N[\,\mathrm{E}_F(Y),\ \mathrm{var}_F(Y)/n\,]$$

Example: Mean

- Taking $\boldsymbol{G}(\theta, Y) = Y - \theta$, we obtain

$$\mathrm{E}_F[\boldsymbol{G}(\theta_0, Y)] = 0 \implies \mathrm{E}_F(Y) = \theta_0$$

so, for any distribution F, θ_0 will represent the mean

- The Z-estimator based on $\{Y_i\}_{i=1}^n \overset{i.i.d.}{\sim} F$ is obtained from

$$\boldsymbol{G}_n(\theta, Y) = \frac{1}{n}\sum_{i=1}^n (Y_i - \theta) = 0 \implies \hat{\theta} = \bar{Y}$$

- On the other hand,

$$\boldsymbol{A}(\theta) = \mathrm{E}_F\left[\boldsymbol{G}'(\theta, Y)\right] = -1$$

and

$$\boldsymbol{B}(\theta) = \mathrm{E}_F[\boldsymbol{G}(\theta, Y)\boldsymbol{G}(\theta, Y)^{\mathsf{T}}] = \mathrm{var}_F(Y)$$

- We finally have that

$$\bar{Y} \approx N[\ \mathrm{E}_F(Y),\ \mathrm{var}_F(Y)/n\ ]$$

Example: Mean and Variance

- Taking

$$\mathrm{E}_F[\boldsymbol{G}(\boldsymbol{\theta}, Y)] = \mathrm{E}_F \begin{pmatrix} Y - \theta_1 \\ (Y - \theta_1)^2 - \theta_2 \end{pmatrix} = \mathbf{0} \implies \begin{array}{l} \theta_1 = \mathrm{E}_F(Y) \\ \theta_2 = \mathrm{var}_F(Y) \end{array}$$

- Based on this definition of $\boldsymbol{G}(\boldsymbol{\theta}, Y)$ we obtain

$$\boldsymbol{A} = \mathrm{E}_F[\boldsymbol{G}'(\boldsymbol{\theta}, Y)] = \mathrm{E}_F \begin{pmatrix} -1 & 0 \\ -2(Y - \theta_1) & -1 \end{pmatrix} = -\boldsymbol{I}_2$$

and

$$\boldsymbol{B} = \mathrm{E}_F[\boldsymbol{G}(\boldsymbol{\theta}, Y)\boldsymbol{G}(\boldsymbol{\theta}, Y)^\top] = \begin{pmatrix} \theta_2 & \mathrm{E}_F[(Y - \theta_1)^3] \\ \mathrm{E}_F[(Y - \theta_1)^3] & \mathrm{E}_F[(Y - \theta_1)^4] - \theta_2^2 \end{pmatrix}$$

- The Z-estimators of (θ_1, θ_2) are the sample mean and variance $(\bar{Y}, s_n^2)$, and

$$\begin{pmatrix} \bar{Y} \\ s_n^2 \end{pmatrix} \approx N\left(\begin{pmatrix} \mathrm{E}_F(Y) \\ \mathrm{var}_F(Y) \end{pmatrix}, \boldsymbol{B}/n \right)$$

- See Sec 7.2.2 of Boos and Stefanski (2013) for more details

Example: Mean and Variance

- Taking

$$\mathrm{E}_F[\boldsymbol{G}(\boldsymbol{\theta}, Y)] = \mathrm{E}_F\begin{pmatrix} Y-\theta_1 \\ (Y-\theta_1)^2-\theta_2 \end{pmatrix} = \mathbf{0} \quad \Longrightarrow \quad \begin{array}{l} \theta_1 = \mathrm{E}_F(Y) \\ \theta_2 = \mathrm{var}_F(Y) \end{array}$$

- Based on this definition of $\boldsymbol{G}(\boldsymbol{\theta}, Y)$ we obtain

$$\boldsymbol{A} = \mathrm{E}_F[\boldsymbol{G}'(\boldsymbol{\theta}, Y)] = \mathrm{E}_F\begin{pmatrix} -1 & 0 \\ -2(Y-\theta_1) & -1 \end{pmatrix} = -\boldsymbol{I}_2$$

and

$$\boldsymbol{B} = \mathrm{E}_F[\boldsymbol{G}(\boldsymbol{\theta}, Y)\boldsymbol{G}(\boldsymbol{\theta}, Y)^\top] = \begin{pmatrix} \theta_2 & \mathrm{E}_F[(Y-\theta_1)^3] \\ \mathrm{E}_F[(Y-\theta_1)^3] & \mathrm{E}_F[(Y-\theta_1)^4]-\theta_2^2 \end{pmatrix}$$

- The Z-estimators of (θ_1, θ_2) are the sample mean and variance $(\bar{Y}, s_n^2)$, and

$$\begin{pmatrix} \bar{Y} \\ s_n^2 \end{pmatrix} \approx N\left(\begin{pmatrix} \mathrm{E}_F(Y) \\ \mathrm{var}_F(Y) \end{pmatrix}, \boldsymbol{B}/n\right)$$

- See Sec 7.2.2 of Boos and Stefanski (2013) for more details

Example: Mean and Variance

- Taking

$$\mathrm{E}_F[\boldsymbol{G}(\boldsymbol{\theta}, Y)] = \mathrm{E}_F \begin{pmatrix} Y - \theta_1 \\ (Y - \theta_1)^2 - \theta_2 \end{pmatrix} = \mathbf{0} \quad \Longrightarrow \quad \begin{aligned} \theta_1 &= \mathrm{E}_F(Y) \\ \theta_2 &= \mathrm{var}_F(Y) \end{aligned}$$

- Based on this definition of $\boldsymbol{G}(\boldsymbol{\theta}, Y)$ we obtain

$$\boldsymbol{A} = \mathrm{E}_F[\boldsymbol{G}'(\boldsymbol{\theta}, Y)] = \mathrm{E}_F \begin{pmatrix} -1 & 0 \\ -2(Y - \theta_1) & -1 \end{pmatrix} = -\boldsymbol{I}_2$$

and

$$\boldsymbol{B} = \mathrm{E}_F[\boldsymbol{G}(\boldsymbol{\theta}, Y)\boldsymbol{G}(\boldsymbol{\theta}, Y)^{\mathsf{T}}] = \begin{pmatrix} \theta_2 & \mathrm{E}_F[(Y - \theta_1)^3] \\ \mathrm{E}_F[(Y - \theta_1)^3] & \mathrm{E}_F[(Y - \theta_1)^4] - \theta_2^2 \end{pmatrix}$$

- The Z-estimators of (θ_1, θ_2) are the sample mean and variance $(\bar{Y}, s_n^2)$, and

$$\begin{pmatrix} \bar{Y} \\ s_n^2 \end{pmatrix} \approx N\left(\begin{pmatrix} \mathrm{E}_F(Y) \\ \mathrm{var}_F(Y) \end{pmatrix}, \boldsymbol{B}/n \right)$$

- See Sec 7.2.2 of Boos and Stefanski (2013) for more details

Example: Mean and Variance

- Taking

$$\mathrm{E}_F[\boldsymbol{G}(\boldsymbol{\theta}, Y)] = \mathrm{E}_F \begin{pmatrix} Y - \theta_1 \\ (Y - \theta_1)^2 - \theta_2 \end{pmatrix} = \mathbf{0} \quad \Longrightarrow \quad \begin{array}{l} \theta_1 = \mathrm{E}_F(Y) \\ \theta_2 = \mathrm{var}_F(Y) \end{array}$$

- Based on this definition of $\boldsymbol{G}(\boldsymbol{\theta}, Y)$ we obtain

$$\boldsymbol{A} = \mathrm{E}_F[\boldsymbol{G}'(\boldsymbol{\theta}, Y)] = \mathrm{E}_F \begin{pmatrix} -1 & 0 \\ -2(Y - \theta_1) & -1 \end{pmatrix} = -\boldsymbol{I}_2$$

  and

$$\boldsymbol{B} = \mathrm{E}_F[\boldsymbol{G}(\boldsymbol{\theta}, Y)\boldsymbol{G}(\boldsymbol{\theta}, Y)^\intercal] = \begin{pmatrix} \theta_2 & \mathrm{E}_F[(Y - \theta_1)^3] \\ \mathrm{E}_F[(Y - \theta_1)^3] & \mathrm{E}_F[(Y - \theta_1)^4] - \theta_2^2 \end{pmatrix}$$

- The Z-estimators of (θ_1, θ_2) are the sample mean and variance $(\bar{Y}, s_n^2)$, and

$$\begin{pmatrix} \bar{Y} \\ s_n^2 \end{pmatrix} \approx N \left(\begin{pmatrix} \mathrm{E}_F(Y) \\ \mathrm{var}_F(Y) \end{pmatrix}, \boldsymbol{B}/n \right)$$

- See Sec 7.2.2 of Boos and Stefanski (2013) for more details

Estimating Equations and Likelihood-Based Inference

The asymptotic distribution of MLEs can be derived as a particular case of the asymptotic distribution of Z-estimators

- Take the score as the basis for the estimating function

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \frac{1}{n}\boldsymbol{S}(\boldsymbol{\theta}),$$

with $\boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \partial \log p(Y_i \mid \boldsymbol{\theta})/\partial\boldsymbol{\theta}$

- We know that, *under the model that leads to the likelihood*,

$$\mathrm{E}[\boldsymbol{G}_n(\boldsymbol{\theta})] = \frac{1}{n}\mathrm{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \boldsymbol{0}$$

- And the MLE satisfies $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{0}$

Estimating Equations and Likelihood-Based Inference

The asymptotic distribution of MLEs can be derived as a particular case of the asymptotic distribution of Z-estimators

- Take the score as the basis for the estimating function

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \frac{1}{n}\boldsymbol{S}(\boldsymbol{\theta}),$$

with $\boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \partial \log p(Y_i \mid \boldsymbol{\theta})/\partial\boldsymbol{\theta}$

- We know that, *under the model that leads to the likelihood*,

$$\mathrm{E}[\boldsymbol{G}_n(\boldsymbol{\theta})] = \frac{1}{n}\mathrm{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \boldsymbol{0}$$

- And the MLE satisfies $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{0}$

Estimating Equations and Likelihood-Based Inference

The asymptotic distribution of MLEs can be derived as a particular case of the asymptotic distribution of Z-estimators

- Take the score as the basis for the estimating function

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \frac{1}{n}\boldsymbol{S}(\boldsymbol{\theta}),$$

  with $\boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \partial \log p(Y_i \mid \boldsymbol{\theta})/\partial\boldsymbol{\theta}$

- We know that, *under the model that leads to the likelihood*,

$$\mathrm{E}[\boldsymbol{G}_n(\boldsymbol{\theta})] = \frac{1}{n}\mathrm{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \boldsymbol{0}$$

- And the MLE satisfies $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{0}$

Estimating Equations and Likelihood-Based Inference

Reminder. The proof of $\mathsf{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathbf{0}$:

$$\begin{aligned}
\mathsf{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathsf{E}\left[\frac{\partial l}{\partial \boldsymbol{\theta}}\right] &= n\mathsf{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right] \\
&= n \int \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(y \mid \boldsymbol{\theta})\right] p(y \mid \boldsymbol{\theta}) dy \quad (\text{if } Y \sim p(\cdot \mid \boldsymbol{\theta})) \\
&= n \int \left[\frac{\partial}{\partial \boldsymbol{\theta}} p(y \mid \boldsymbol{\theta})\right] \frac{p(y \mid \boldsymbol{\theta})}{p(y \mid \boldsymbol{\theta})} dy \\
&= n \frac{\partial}{\partial \boldsymbol{\theta}} \int p(y \mid \boldsymbol{\theta}) dy \quad (\text{if OK to exchange } \int \text{ and } \partial/\partial \boldsymbol{\theta}) \\
&= \mathbf{0}.
\end{aligned}$$

So this result, in general, relies on your model being correctly specified, i.e. $Y \sim p(\cdot \mid \theta)$

Estimating Equations and Likelihood-Based Inference

Reminder. The proof of $\mathsf{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathbf{0}$:

$$
\begin{aligned}
\mathsf{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathsf{E}\left[\frac{\partial l}{\partial \boldsymbol{\theta}}\right] &= n\mathsf{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right] \\
&= n \int \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(y \mid \boldsymbol{\theta})\right] p(y \mid \boldsymbol{\theta}) dy \quad (\text{if } Y \sim p(\cdot \mid \boldsymbol{\theta})) \\
&= n \int \left[\frac{\partial}{\partial \boldsymbol{\theta}} p(y \mid \boldsymbol{\theta})\right] \frac{p(y \mid \boldsymbol{\theta})}{p(y \mid \boldsymbol{\theta})} dy \\
&= n \frac{\partial}{\partial \boldsymbol{\theta}} \int p(y \mid \boldsymbol{\theta}) dy \quad \left(\text{if OK to exchange } \int \text{ and } \partial/\partial \boldsymbol{\theta}\right) \\
&= \mathbf{0}.
\end{aligned}
$$

So this result, in general, relies on your model being correctly specified, i.e. $Y \sim p(\cdot \mid \boldsymbol{\theta})$

Estimating Equations and Likelihood-Based Inference

Reminder. The proof of $\mathrm{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathbf{0}$:

$$
\begin{aligned}
\mathrm{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathrm{E}\left[\frac{\partial l}{\partial \boldsymbol{\theta}}\right] &= n\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right] \\
&= n \int \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(y \mid \boldsymbol{\theta})\right] p(y \mid \boldsymbol{\theta}) dy \quad (\text{if } Y \sim p(\cdot \mid \boldsymbol{\theta})) \\
&= n \int \left[\frac{\partial}{\partial \boldsymbol{\theta}} p(y \mid \boldsymbol{\theta})\right] \frac{p(y \mid \boldsymbol{\theta})}{p(y \mid \boldsymbol{\theta})} dy \\
&= n \frac{\partial}{\partial \boldsymbol{\theta}} \int p(y \mid \boldsymbol{\theta}) dy \quad (\text{if OK to exchange } \int \text{ and } \partial/\partial \boldsymbol{\theta}) \\
&= \mathbf{0}.
\end{aligned}
$$

So this result, in general, relies on your model being correctly specified, i.e. $Y \sim p(\cdot \mid \theta)$

Estimating Equations and Likelihood-Based Inference

Reminder. The proof of $\mathrm{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathbf{0}$:

$$
\begin{aligned}
\mathrm{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathrm{E}\left[\frac{\partial l}{\partial \boldsymbol{\theta}}\right] &= n\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right] \\
&= n \int \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(y \mid \boldsymbol{\theta})\right] p(y \mid \boldsymbol{\theta}) dy \quad \text{(if } Y \sim p(\cdot \mid \boldsymbol{\theta})\text{)} \\
&= n \int \left[\frac{\partial}{\partial \boldsymbol{\theta}} p(y \mid \boldsymbol{\theta})\right] \frac{p(y \mid \boldsymbol{\theta})}{p(y \mid \boldsymbol{\theta})} dy \\
&= n \frac{\partial}{\partial \boldsymbol{\theta}} \int p(y \mid \boldsymbol{\theta}) dy \quad \text{(if OK to exchange } \textstyle\int \text{ and } \partial/\partial\boldsymbol{\theta}\text{)} \\
&= \mathbf{0}.
\end{aligned}
$$

So this result, in general, relies on your model being correctly specified, i.e. $Y \sim p(\cdot \mid \theta)$

Estimating Equations and Likelihood-Based Inference

Reminder. The proof of $\mathsf{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathbf{0}$:

$$
\begin{aligned}
\mathsf{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathsf{E}\left[\frac{\partial l}{\partial \boldsymbol{\theta}}\right] &= n\mathsf{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right] \\
&= n \int \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(y \mid \boldsymbol{\theta})\right] p(y \mid \boldsymbol{\theta}) dy \quad (\text{if } Y \sim p(\cdot \mid \boldsymbol{\theta})) \\
&= n \int \left[\frac{\partial}{\partial \boldsymbol{\theta}} p(y \mid \boldsymbol{\theta})\right] \frac{p(y \mid \boldsymbol{\theta})}{p(y \mid \boldsymbol{\theta})} dy \\
&= n \frac{\partial}{\partial \boldsymbol{\theta}} \int p(y \mid \boldsymbol{\theta}) dy \quad \left(\text{if OK to exchange } \int \text{ and } \partial/\partial \boldsymbol{\theta}\right) \\
&= \mathbf{0}.
\end{aligned}
$$

So this result, in general, relies on your model being correctly specified, i.e. $Y \sim p(\cdot \mid \boldsymbol{\theta})$

Estimating Equations and Likelihood-Based Inference

Taking $\boldsymbol{G}_n(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{\theta})/n$, the $\boldsymbol{A}$ and $\boldsymbol{B}$ in the asymptotic covariance of Z-estimators take the following form:

$$\boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[\boldsymbol{G}'(\theta, Y)\right] = \mathrm{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^\intercal \partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right],$$

and

$$\begin{aligned} \boldsymbol{B}(\theta) &= \mathrm{E}\left[\boldsymbol{G}(\theta, Y)\boldsymbol{G}(\theta, Y)^\intercal\right] \\ &= \mathrm{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right)^\intercal\right], \end{aligned}$$

- From the definition of Fisher information $\mathcal{I}_1(\boldsymbol{\theta}) = \boldsymbol{B}(\boldsymbol{\theta})$
- *Under the model that leads to the likelihood* we obtain $\mathcal{I}_1(\boldsymbol{\theta}) = -\boldsymbol{A}(\boldsymbol{\theta})$

Estimating Equations and Likelihood-Based Inference

Taking $\boldsymbol{G}_n(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{\theta})/n$, the $\boldsymbol{A}$ and $\boldsymbol{B}$ in the asymptotic covariance of Z-estimators take the following form:

$$\boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[\boldsymbol{G}'(\theta, Y)\right] = \mathrm{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^\mathsf{T} \partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right],$$

and

$$\begin{aligned}
\boldsymbol{B}(\boldsymbol{\theta}) &= \mathrm{E}\left[\boldsymbol{G}(\theta, Y)\boldsymbol{G}(\theta, Y)^\mathsf{T}\right] \\
&= \mathrm{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right)^\mathsf{T}\right],
\end{aligned}$$

- From the definition of Fisher information $\mathcal{I}_1(\boldsymbol{\theta}) = \boldsymbol{B}(\boldsymbol{\theta})$
- *Under the model that leads to the likelihood* we obtain $\mathcal{I}_1(\boldsymbol{\theta}) = -\boldsymbol{A}(\boldsymbol{\theta})$

Estimating Equations and Likelihood-Based Inference

Taking $\boldsymbol{G}_n(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{\theta})/n$, the $\boldsymbol{A}$ and $\boldsymbol{B}$ in the asymptotic covariance of Z-estimators take the following form:

$$\boldsymbol{A}(\boldsymbol{\theta}) = \mathsf{E}\left[\boldsymbol{G}'(\theta, Y)\right] = \mathsf{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^{\mathsf{T}} \partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right],$$

and

$$\begin{aligned}
\boldsymbol{B}(\boldsymbol{\theta}) &= \mathsf{E}\left[\boldsymbol{G}(\theta, Y)\boldsymbol{G}(\theta, Y)^{\mathsf{T}}\right] \\
&= \mathsf{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right)^{\mathsf{T}}\right],
\end{aligned}$$

- From the definition of Fisher information $\mathcal{I}_1(\boldsymbol{\theta}) = \boldsymbol{B}(\boldsymbol{\theta})$
- *Under the model that leads to the likelihood* we obtain $\mathcal{I}_1(\boldsymbol{\theta}) = -\boldsymbol{A}(\boldsymbol{\theta})$

Estimating Equations and Likelihood-Based Inference

- Therefore, *if the model is correctly specified*, the asymptotic variance of the MLE obtained as a Z-estimator is

$$\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^{\top}/n = \mathcal{I}_1(\boldsymbol{\theta})^{-1}/n$$

- Hence, *under the model*

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \;\rightarrow_d\; \mathrm{N}_p[\,\mathbf{0},\, \mathcal{I}_1(\boldsymbol{\theta})^{-1}\,]$$

- This result relies on $\mathcal{I}_1(\boldsymbol{\theta}) = \boldsymbol{B}(\boldsymbol{\theta}) = -\boldsymbol{A}(\boldsymbol{\theta})$, which *holds under the parametric model leading to the likelihood*

Estimating Equations and Likelihood-Based Inference

- Therefore, *if the model is correctly specified*, the asymptotic variance of the MLE obtained as a Z-estimator is

$$\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^{\top}/n = \mathcal{I}_1(\boldsymbol{\theta})^{-1}/n$$

- Hence, *under the model*

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \;\to_d\; \mathrm{N}_p[\,\mathbf{0},\ \mathcal{I}_1(\boldsymbol{\theta})^{-1}\,]$$

- This result relies on $\mathcal{I}_1(\boldsymbol{\theta}) = \boldsymbol{B}(\boldsymbol{\theta}) = -\boldsymbol{A}(\boldsymbol{\theta})$, which *holds under the parametric model leading to the likelihood*

Estimating Equations and Likelihood-Based Inference

Reminder. The proof[2] that

$$-\mathrm{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}^{\mathsf{T}}\partial\boldsymbol{\theta}}\log p(Y\mid\boldsymbol{\theta})\right]=\mathrm{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(Y\mid\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(Y\mid\boldsymbol{\theta})\right)^{\mathsf{T}}\right],$$

relies on being able to write down

$$\mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log p(y\mid\boldsymbol{\theta})\right]=\int\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log p(y\mid\boldsymbol{\theta})\right]p(y\mid\boldsymbol{\theta})dy,$$

which is valid if $Y\sim p(\cdot\mid\boldsymbol{\theta})$

- So this result, in general, relies on your model being correctly specified, i.e. $Y\sim p(\cdot\mid\boldsymbol{\theta})$
- Nevertheless, the theory of estimating equations will allow us to study the asymptotic behavior of MLEs under model misspecification, among many other things!

[2] See, e.g., Theorems 6.6 and 6.7 of Boos and Stefanski (2013), or p. 38 of Wakefield(2013)

Looking Forward

The theory of estimating equations will allow us to study:

- What happens with fully parametric models when they are wrong?
- Assuming only a regression and variance function for $Y \mid \boldsymbol{x}$
- Least squares without assuming a fully parametric model for $Y \mid \boldsymbol{x}$
- Assuming only a regression function for $Y \mid \boldsymbol{x}$

Behavior of MLEs Under Misspecification

In practice we can *compute* MLEs *regardless* of whether our model is correct – what are we estimating if the model is wrong?

- Let F be the true distribution, with density $f(\cdot)$
- Let $p(y \mid \boldsymbol{\theta})$ denote the density function of the *assumed* model
- In practice, we wouldn't know that the data come from F, so we still maximize the log-likelihood (dividing by n doesn't change the maximizer)

$$\frac{1}{n}\sum_{i=1}^{n} \log p(Y_i \mid \theta)$$

but now $\{Y_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} F$

Behavior of MLEs Under Misspecification

In practice we can *compute* MLEs *regardless* of whether our model is correct – what are we estimating if the model is wrong?

- Let F be the true distribution, with density $f(\cdot)$
- Let $p(y \mid \boldsymbol{\theta})$ denote the density function of the *assumed* model
- In practice, we wouldn't know that the data come from F, so we still maximize the log-likelihood (dividing by n doesn't change the maximizer)

$$\frac{1}{n}\sum_{i=1}^{n} \log p(Y_i \mid \boldsymbol{\theta})$$

but now $\{Y_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} F$

Behavior of MLEs Under Misspecification

- In the limit we have that

$$\frac{1}{n}\sum_{i=1}^{n}\log p(Y_i \mid \boldsymbol{\theta}) \rightarrow_{a.s.} \mathrm{E}_F[\log p(Y \mid \boldsymbol{\theta})] = \int \log p(y \mid \boldsymbol{\theta}) f(y) dy,$$

- We also obtain that $\hat{\boldsymbol{\theta}}_n \rightarrow_p \boldsymbol{\theta}_T$, where

$$\boldsymbol{\theta}_T = \arg\max_{\boldsymbol{\theta}} \mathrm{E}_F \left[\log p(Y \mid \boldsymbol{\theta})\right]$$

- We shall refer to $\boldsymbol{\theta}_T$ as the *pseudo-true* parameter, as this is the parameter that we will recover asymptotically

Behavior of MLEs Under Misspecification

- In the limit we have that

$$\frac{1}{n}\sum_{i=1}^{n}\log p(Y_i \mid \boldsymbol{\theta}) \rightarrow_{a.s.} \mathsf{E}_F[\log p(Y \mid \boldsymbol{\theta})] = \int \log p(y \mid \boldsymbol{\theta}) f(y)dy,$$

- We also obtain that $\hat{\boldsymbol{\theta}}_n \rightarrow_p \boldsymbol{\theta}_T$, where

$$\boldsymbol{\theta}_T = \arg\max_{\boldsymbol{\theta}} \mathsf{E}_F \left[\log p(Y \mid \boldsymbol{\theta})\right]$$

- We shall refer to $\boldsymbol{\theta}_T$ as the *pseudo-true* parameter, as this is the parameter that we will recover asymptotically

Behavior of MLEs Under Misspecification

- Now, note that we can write

$$
\begin{aligned}
\mathrm{E}_F[\log p(Y \mid \boldsymbol{\theta})] &= \mathrm{E}_F\left[\log f(Y) - \log f(Y) + \log p(Y \mid \boldsymbol{\theta})\right] \\
&= \mathrm{E}_F[\log f(Y)] - \mathrm{KL}[f(\cdot), p(\cdot \mid \boldsymbol{\theta})]
\end{aligned}
$$

where KL is the *Kullback-Leibler divergence*

$$
\mathrm{KL}[f(\cdot), p(\cdot \mid \boldsymbol{\theta})] = \int f(y) \log \frac{f(y)}{p(y \mid \boldsymbol{\theta})} \, \mathrm{d}y
$$

- Therefore,

$$
\boldsymbol{\theta}_T = \arg\min_{\boldsymbol{\theta}} \mathrm{KL}[f(\cdot), p(\cdot \mid \boldsymbol{\theta})],
$$

that is, the MLE asymptotically minimizes KL as a function of $\boldsymbol{\theta}$

- The MLE is that value of $\boldsymbol{\theta}$ that makes the *assumed model* diverge the least from the true distribution F

Behavior of MLEs Under Misspecification

- Now, note that we can write

$$\begin{aligned} \mathsf{E}_F[\log p(Y \mid \boldsymbol{\theta})] &= \mathsf{E}_F\left[\log f(Y) - \log f(Y) + \log p(Y \mid \boldsymbol{\theta})\right] \\ &= \mathsf{E}_F[\log f(Y)] - \mathrm{KL}[f(\cdot), p(\cdot \mid \boldsymbol{\theta})] \end{aligned}$$

  where KL is the *Kullback-Leibler divergence*

$$\mathrm{KL}[f(\cdot), p(\cdot \mid \boldsymbol{\theta})] = \int f(y) \log \frac{f(y)}{p(y \mid \boldsymbol{\theta})} \, \mathrm{d}y$$

- Therefore,

$$\boldsymbol{\theta}_T = \arg\min_{\boldsymbol{\theta}} \mathrm{KL}[f(\cdot), p(\cdot \mid \boldsymbol{\theta})],$$

  that is, the MLE asymptotically minimizes KL as a function of $\boldsymbol{\theta}$

- The MLE is that value of $\boldsymbol{\theta}$ that makes the *assumed model* diverge the least from the true distribution F

Behavior of MLEs Under Misspecification

- Now, note that we can write

$$\begin{aligned} \mathrm{E}_F[\log p(Y \mid \boldsymbol{\theta})] &= \mathrm{E}_F\left[\log f(Y) - \log f(Y) + \log p(Y \mid \boldsymbol{\theta})\right] \\ &= \mathrm{E}_F[\log f(Y)] - \mathrm{KL}[f(\cdot), p(\cdot \mid \boldsymbol{\theta})] \end{aligned}$$

where KL is the *Kullback-Leibler divergence*

$$\mathrm{KL}[f(\cdot), p(\cdot \mid \boldsymbol{\theta})] = \int f(y) \log \frac{f(y)}{p(y \mid \boldsymbol{\theta})} \, \mathrm{d}y$$

- Therefore,

$$\boldsymbol{\theta}_T = \arg\min_{\boldsymbol{\theta}} \mathrm{KL}[f(\cdot), p(\cdot \mid \boldsymbol{\theta})],$$

that is, the MLE asymptotically minimizes KL as a function of $\boldsymbol{\theta}$

- The MLE is that value of $\boldsymbol{\theta}$ that makes the *assumed model* diverge the least from the true distribution F

Behavior of MLEs Under Misspecification

- Taking derivatives of $\mathsf{E}_F\left[\log p(Y \mid \boldsymbol{\theta})\right]$ and interchanging with the integral, we find that $\boldsymbol{\theta}_T$ solves the system of equations at the population level:

$$\boldsymbol{\theta}_T : \quad \mathsf{E}_F \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right]\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_T} = \mathbf{0}$$

- We can use this as the basis for treating $\hat{\boldsymbol{\theta}}_n$ as a Z-estimator for $\boldsymbol{\theta}_T$

Behavior of MLEs Under Misspecification

Result: Suppose that $\hat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log p(Y_i \mid \boldsymbol{\theta}) = \mathbf{0}.$$

Then again

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_T) \rightarrow_d \mathrm{N}_p[\,\mathbf{0},\ \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{\top -1}\,]$$

where

$$\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{\theta}_T) = \mathrm{E}_T\left[\frac{\partial^2}{\partial\boldsymbol{\theta}^\top\partial\boldsymbol{\theta}}\log p(Y \mid \boldsymbol{\theta})\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_T}$$

and

$$\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{\theta}_T) = \mathrm{E}_T\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(Y \mid \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(Y \mid \boldsymbol{\theta})\right)^\top\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_T}$$

Behavior of MLEs Under Misspecification

- The above result characterizes the behavior of MLEs under model misspecification
- It can be used for inferences on the pseudo-true parameters $\boldsymbol{\theta}_T$
- However, $\boldsymbol{\theta}_T$ is *in general* difficult to interpret
  - Huber (1967) presented these results from a purely mathematical point of view: what is the MLE's asymptotic distribution if the model is misspecified?
  - David A. Freedman (2006)[3] criticizes using these results in practice: if your model is wrong, how do you know you care about $\boldsymbol{\theta}_T$?

[3] On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors" *The American Statistician*

Behavior of MLEs Under Misspecification

- With parametric models, we assume that the true distribution F has a density that can be written as $p(y \mid \boldsymbol{\theta})$ for some $\boldsymbol{\theta} \in \Theta$
- Ideally, we choose this model based on knowledge of the data, and therefore the interpretation of $\boldsymbol{\theta}$ is given when you specify the model
- On the other hand, reality is complicated, and so F may not have a density that can be written as $p(y \mid \boldsymbol{\theta})$ for some $\boldsymbol{\theta} \in \Theta$
- We saw that the MLE estimates the value of $\boldsymbol{\theta}_T$ that makes the assumed model "least divergent" from the true distribution
- The interpretation of $\boldsymbol{\theta}_T$ can also be obtained from reorganizing:

$$\boldsymbol{\theta}_T : \quad \mathrm{E}_F \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta}) \right] \Bigg|_{\boldsymbol{\theta} = \boldsymbol{\theta}_T} = \mathbf{0}$$

- This may or may not lead to an interpretation of $\boldsymbol{\theta}$ that matches what you intended when you specified your model

Behavior of MLEs Under Misspecification

- With parametric models, we assume that the true distribution F has a density that can be written as $p(y \mid \boldsymbol{\theta})$ for some $\boldsymbol{\theta} \in \Theta$
- Ideally, we choose this model based on knowledge of the data, and therefore the interpretation of $\boldsymbol{\theta}$ is given when you specify the model
- On the other hand, reality is complicated, and so F may not have a density that can be written as $p(y \mid \boldsymbol{\theta})$ for some $\boldsymbol{\theta} \in \Theta$
- We saw that the MLE estimates the value of $\boldsymbol{\theta}_T$ that makes the assumed model "least divergent" from the true distribution
- The interpretation of $\boldsymbol{\theta}_T$ can also be obtained from reorganizing:

$$\boldsymbol{\theta}_T : \quad \mathrm{E}_F\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta})\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_T} = \mathbf{0}$$

- This may or may not lead to an interpretation of $\boldsymbol{\theta}$ that matches what you intended when you specified your model

Behavior of MLEs Under Misspecification

- With parametric models, we assume that the true distribution F has a density that can be written as $p(y \mid \boldsymbol{\theta})$ for some $\boldsymbol{\theta} \in \Theta$
- Ideally, we choose this model based on knowledge of the data, and therefore the interpretation of $\boldsymbol{\theta}$ is given when you specify the model
- On the other hand, reality is complicated, and so F may not have a density that can be written as $p(y \mid \boldsymbol{\theta})$ for some $\boldsymbol{\theta} \in \Theta$
- We saw that the MLE estimates the value of $\boldsymbol{\theta}_T$ that makes the assumed model "least divergent" from the true distribution
- The interpretation of $\boldsymbol{\theta}_T$ can also be obtained from reorganizing:

$$\boldsymbol{\theta}_T : \quad \mathsf{E}_F \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta}) \right] \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_T} = \mathbf{0}$$

- This may or may not lead to an interpretation of $\boldsymbol{\theta}$ that matches what you intended when you specified your model

Example: Misspecified Exponential Dispersion Family

- The generic term of the log-likelihood *obtained* from the exponential dispersion family is given by

$$\log p(Y \mid \theta, \alpha) = \frac{Y\theta - b(\theta)}{\alpha} + c(Y, \alpha)$$

- From this,

$$\frac{\partial}{\partial \theta} \log p(Y \mid \theta, \alpha) = \frac{Y - b'(\theta)}{\alpha}$$

- Using this in an estimating equation, with $Y \sim F$, leads to

$$\mathrm{E}_F \left[\frac{\partial}{\partial \theta} \log p(Y \mid \theta, \alpha) \right] \Bigg|_{\boldsymbol{\theta} = \boldsymbol{\theta}_T} = 0 \quad \Longrightarrow \quad \mathrm{E}_F(Y) = b'(\theta_T)$$

- Therefore, using the exponential dispersion family to form the likelihood based on data $\{Y_i\}_{i=1}^n \stackrel{i.i.d.}{\sim} F$ leads to $b'(\hat{\theta})$ as a consistent estimator for $\mathrm{E}_F(Y)$, regardless of whether the model is misspecified!

Example: Misspecified Exponential Dispersion Family

- The generic term of the log-likelihood *obtained* from the exponential dispersion family is given by

$$\log p(Y \mid \theta, \alpha) = \frac{Y\theta - b(\theta)}{\alpha} + c(Y, \alpha)$$

- From this,

$$\frac{\partial}{\partial \theta} \log p(Y \mid \theta, \alpha) = \frac{Y - b'(\theta)}{\alpha}$$

- Using this in an estimating equation, with $Y \sim F$, leads to

$$\mathrm{E}_F \left[\frac{\partial}{\partial \theta} \log p(Y \mid \theta, \alpha) \right] \Bigg|_{\boldsymbol{\theta} = \boldsymbol{\theta}_T} = 0 \quad \Longrightarrow \quad \mathrm{E}_F(Y) = b'(\theta_T)$$

- Therefore, using the exponential dispersion family to form the likelihood based on data $\{Y_i\}_{i=1}^n \overset{i.i.d.}{\sim} F$ leads to $b'(\hat{\theta})$ as a consistent estimator for $\mathrm{E}_F(Y)$, regardless of whether the model is misspecified!

Example: Misspecified Exponential Dispersion Family

- We can also find that

$$\frac{\partial}{\partial \alpha} \log p(Y \mid \theta, \alpha) = -\frac{Y\theta - b(\theta)}{\alpha^2} + \frac{\partial}{\partial \alpha} c(Y, \alpha)$$

- Adding this to the previous estimating equation, with $Y \sim F$, leads to α_T being the value that solves

$$\theta_T \mathrm{E}_F(Y) - b(\theta_T) = \alpha^2 \mathrm{E}_F \left[\frac{\partial}{\partial \alpha} c(Y, \alpha) \right]$$

- Therefore, in general it is not clear whether α_T would be of interest when the model is misspecified

Example: Misspecified Weibull Model

- Jon A. Wellner in his Breslow Lecture (November 12, 2020)[4] presented the example of a misspecified Weibull(α, β) model, and he showed that the MLE of α can be interpretable even if the model is misspecified, whereas the MLE of β is more difficult to interpret.
- *Problem*: Explain where the MLE of α and β converge to if the Weibull model is misspecified.

[4] https://sites.stat.washington.edu/jaw/RESEARCH/TALKS/Breslow-Lecture.pdf

Sandwich Estimation

- The Z-estimator $\hat{\boldsymbol{\theta}}_n$ that solves

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \mathbf{0},$$

  based on independent and identically distributed observations, has asymptotic variance $\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^{\intercal}/n$

- We use this to define the *sandwich estimator* of $\mathrm{var}(\hat{\theta}_n)$ as

$$\widehat{\mathrm{var}}(\hat{\theta}_n) = \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{B}}(\hat{\boldsymbol{A}}^{-1})^{\intercal}/n,$$

  where

$$\hat{\boldsymbol{A}} := \hat{\boldsymbol{A}}(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}'(\hat{\theta}_n, Y_i),$$

  and

$$\hat{\boldsymbol{B}} := \hat{\boldsymbol{B}}(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\hat{\theta}_n, Y_i)\boldsymbol{G}(\hat{\theta}_n, Y_i)^{\intercal}.$$

Sandwich Estimation

- The Z-estimator $\hat{\boldsymbol{\theta}}_n$ that solves

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \mathbf{0},$$

  based on independent and identically distributed observations, has asymptotic variance $\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^\intercal/n$

- We use this to define the *sandwich estimator* of $\text{var}(\hat{\boldsymbol{\theta}}_n)$ as

$$\widehat{\text{var}}(\hat{\boldsymbol{\theta}}_n) = \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{B}}(\hat{\boldsymbol{A}}^{-1})^\intercal/n,$$

  where

$$\hat{\boldsymbol{A}} := \hat{\boldsymbol{A}}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}'(\hat{\boldsymbol{\theta}}_n, Y_i),$$

  and

$$\hat{\boldsymbol{B}} := \hat{\boldsymbol{B}}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\hat{\boldsymbol{\theta}}_n, Y_i)\boldsymbol{G}(\hat{\boldsymbol{\theta}}_n, Y_i)^\intercal.$$

Sandwich Estimation

- Note that by the weak law of large numbers $\hat{\boldsymbol{A}}(\boldsymbol{\theta}) \rightarrow_p \boldsymbol{A}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{B}}(\boldsymbol{\theta}) \rightarrow_p \boldsymbol{B}(\boldsymbol{\theta})$, for any fixed $\boldsymbol{\theta}$

- However, our estimators are $\hat{\boldsymbol{A}} := \hat{\boldsymbol{A}}(\hat{\theta}_n)$ and $\hat{\boldsymbol{B}} := \hat{\boldsymbol{B}}(\hat{\theta}_n)$, which depend on the Z- estimator $\hat{\theta}_n$

- Nevertheless, since $\hat{\theta}_n \rightarrow_p \theta_0$, Boos and Stefanski (2013) present Theorems 7.3 and 7.4, which guarantee

$$\hat{\boldsymbol{A}}(\hat{\theta}_n) \rightarrow_p \boldsymbol{A}(\theta_0) := \boldsymbol{A}, \quad \hat{\boldsymbol{B}}(\hat{\theta}_n) \rightarrow_p \boldsymbol{B}(\theta_0) := \boldsymbol{B},$$

and therefore

$$\widehat{\boldsymbol{A}}^{-1} \widehat{\boldsymbol{B}} (\widehat{\boldsymbol{A}}^{\top})^{-1} \rightarrow_p \boldsymbol{A}^{-1} \boldsymbol{B} (\boldsymbol{A}^{\top})^{-1}$$

Sandwich Estimation

- Note that by the weak law of large numbers $\hat{\boldsymbol{A}}(\boldsymbol{\theta}) \to_p \boldsymbol{A}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{B}}(\boldsymbol{\theta}) \to_p \boldsymbol{B}(\boldsymbol{\theta})$, for any fixed $\boldsymbol{\theta}$
- However, our estimators are $\hat{\boldsymbol{A}} := \hat{\boldsymbol{A}}(\hat{\boldsymbol{\theta}}_n)$ and $\hat{\boldsymbol{B}} := \hat{\boldsymbol{B}}(\hat{\boldsymbol{\theta}}_n)$, which depend on the Z- estimator $\hat{\boldsymbol{\theta}}_n$
- Nevertheless, since $\hat{\theta}_n \to_p \theta_0$, Boos and Stefanski (2013) present Theorems 7.3 and 7.4, which guarantee

$$\hat{A}(\hat{\theta}_n) \to_p \boldsymbol{A}(\theta_0) := \boldsymbol{A}, \quad \hat{B}(\hat{\theta}_n) \to_p \boldsymbol{B}(\theta_0) := \boldsymbol{B},$$

and therefore

$$\widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}(\widehat{\boldsymbol{A}}^{\top})^{-1} \to_p \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{\top})^{-1}$$

Sandwich Estimation

- Note that by the weak law of large numbers $\hat{\boldsymbol{A}}(\boldsymbol{\theta}) \to_p \boldsymbol{A}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{B}}(\boldsymbol{\theta}) \to_p \boldsymbol{B}(\boldsymbol{\theta})$, for any fixed $\boldsymbol{\theta}$

- However, our estimators are $\hat{\boldsymbol{A}} := \hat{\boldsymbol{A}}(\hat{\boldsymbol{\theta}}_n)$ and $\hat{\boldsymbol{B}} := \hat{\boldsymbol{B}}(\hat{\boldsymbol{\theta}}_n)$, which depend on the Z- estimator $\hat{\boldsymbol{\theta}}_n$

- Nevertheless, since $\hat{\boldsymbol{\theta}}_n \to_p \boldsymbol{\theta}_0$, Boos and Stefanski (2013) present Theorems 7.3 and 7.4, which guarantee

$$\hat{\boldsymbol{A}}(\hat{\boldsymbol{\theta}}_n) \to_p \boldsymbol{A}(\boldsymbol{\theta}_0) := \boldsymbol{A}, \quad \hat{\boldsymbol{B}}(\hat{\boldsymbol{\theta}}_n) \to_p \boldsymbol{B}(\boldsymbol{\theta}_0) := \boldsymbol{B},$$

  and therefore

$$\widehat{\boldsymbol{A}}^{-1} \widehat{\boldsymbol{B}} (\widehat{\boldsymbol{A}}^{\top})^{-1} \to_p \boldsymbol{A}^{-1} \boldsymbol{B} (\boldsymbol{A}^{\top})^{-1}$$

Sandwich Estimation and Robust Standard Errors

- The sandwich estimator provides a consistent estimator of the variance-covariance of Z-estimators in very broad situations

- For small sample sizes, the sandwich estimator may be unreliable, as it builds on an asymptotic argument: model-based estimators may be preferable for small to medium sized n, but you need to trust your model!

- The sandwich estimator is often called *robust* (people talk about *robust standard errors*), meaning *robustness to model departures*

Sandwich Estimation

How about when $\hat{\boldsymbol{\theta}}_n$ is the MLE of a misspecified model?

- The estimating function arises from the score from a likelihood function:

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}} l_i(\boldsymbol{\theta}),$$

  with $l_i(\boldsymbol{\theta}) = \log p(Y_i \mid \boldsymbol{\theta})$

- Plugging into the formulae above, the sandwich estimator is based on

$$\hat{\boldsymbol{A}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}^{\mathsf{T}}\partial\boldsymbol{\theta}} l_i(\boldsymbol{\theta})\bigg|_{\hat{\boldsymbol{\theta}}_n}$$

  and

$$\hat{\boldsymbol{B}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\theta}} l_i(\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}} l_i(\boldsymbol{\theta})\right)^{\mathsf{T}}\bigg|_{\hat{\boldsymbol{\theta}}_n}$$

- Inferences on the pseudo-true parameter $\boldsymbol{\theta}_T$ can be based on the sandwich variance estimator $\widehat{\text{var}}(\hat{\boldsymbol{\theta}}_n) = \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{B}}(\hat{\boldsymbol{A}}^{-1})^{\mathsf{T}}/n$

Sandwich Estimation

How about when $\hat{\boldsymbol{\theta}}_n$ is the MLE of a misspecified model?

- The estimating function arises from the score from a likelihood function:

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} l_i(\boldsymbol{\theta}),$$

with $l_i(\boldsymbol{\theta}) = \log p(Y_i \mid \boldsymbol{\theta})$

- Plugging into the formulae above, the sandwich estimator is based on

$$\hat{\boldsymbol{A}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta}^\mathsf{T} \partial \boldsymbol{\theta}} l_i(\boldsymbol{\theta}) \bigg|_{\hat{\boldsymbol{\theta}}_n}$$

and

$$\hat{\boldsymbol{B}} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\theta}} l_i(\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} l_i(\boldsymbol{\theta})\right)^\mathsf{T} \bigg|_{\hat{\boldsymbol{\theta}}_n}$$

- Inferences on the pseudo-true parameter $\boldsymbol{\theta}_T$ can be based on the sandwich variance estimator $\widehat{\text{var}}(\hat{\boldsymbol{\theta}}_n) = \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{B}}(\hat{\boldsymbol{A}}^{-1})^\mathsf{T}/n$

Regression and Estimating Equations

We now get back to the topic of this class: regression

- For Z-estimation and estimating equations, so far we relied on i.i.d. data
- In a regression context we focus on modeling $\mathrm{E}(Y \mid \boldsymbol{x})$, and typically focus on the randomness of Y conditional on covariates $\boldsymbol{x}$
- Nevertheless, if we can assume that our data are, say,

$$(Y_1, \boldsymbol{x}_1), \ldots, (Y_n, \boldsymbol{x}_n) \overset{i.i.d.}{\sim} F,$$

  then all the results on estimating equations apply to regression estimators
- If we want to treat the data as independent pairs $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^n$ where the covariates $\boldsymbol{x}_i$ are fixed, then more careful treatment is required

Ordinary Least Squares, Best Fitting Plane

Let's start with the simplest regression approach

- Consider the least squares estimator

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i \boldsymbol{\beta})^2 = \underset{\boldsymbol{\beta}}{\text{argmin}} \ (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

- Taking derivatives with respect to β and setting them to zero, leads to

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

which we had also obtained as the MLE in the *normal linear model*

Ordinary Least Squares, Best Fitting Plane

Let's start with the simplest regression approach

- Consider the least squares estimator

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i \boldsymbol{\beta})^2 = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

  where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

- Taking derivatives with respect to $\boldsymbol{\beta}$ and setting them to zero, leads to

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

  which we had also obtained as the MLE in the *normal linear model*

Ordinary Least Squares, Best Fitting Plane

- The system of equations that we solve to find the OLS estimator is

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\boldsymbol{\beta})\boldsymbol{x}_i^\top = \mathbf{0}$$

  which fits into the framework of estimating equations!

- We can take

$$\boldsymbol{G}(\boldsymbol{\beta}, Y, \boldsymbol{x}) = (Y - \boldsymbol{x}\boldsymbol{\beta})\boldsymbol{x}^\top$$

  as the basis for a Z-estimator for $\boldsymbol{\beta}$ based on $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^{n} \stackrel{i.i.d.}{\sim} F$

- If we do not impose any assumptions on the distribution F, this leads to the target of inference being

$$\boldsymbol{\beta}_0 : \quad \mathrm{E}_F[(Y - \boldsymbol{x}\boldsymbol{\beta}_0)\boldsymbol{x}^\top] = \mathbf{0},$$

  from which we find

$$\boldsymbol{\beta}_0 = \mathrm{E}_F(\boldsymbol{x}^\top\boldsymbol{x})^{-1}\mathrm{E}_F(Y\boldsymbol{x}^\top),$$

  with $(Y, \boldsymbol{x}) \sim F$

Ordinary Least Squares, Best Fitting Plane

- The system of equations that we solve to find the OLS estimator is

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\boldsymbol{\beta})\boldsymbol{x}_i^{\top} = \mathbf{0}$$

  which fits into the framework of estimating equations!

- We can take

$$\boldsymbol{G}(\boldsymbol{\beta}, Y, \boldsymbol{x}) = (Y - \boldsymbol{x}\boldsymbol{\beta})\boldsymbol{x}^{\top}$$

  as the basis for a Z-estimator for $\boldsymbol{\beta}$ based on $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} F$

- If we do not impose any assumptions on the distribution F, this leads to the target of inference being

$$\boldsymbol{\beta}_0 : \quad \mathrm{E}_F[(Y - \boldsymbol{x}\boldsymbol{\beta}_0)\boldsymbol{x}^{\top}] = \mathbf{0},$$

  from which we find

$$\boldsymbol{\beta}_0 = \mathrm{E}_F(\boldsymbol{x}^{\top}\boldsymbol{x})^{-1}\mathrm{E}_F(Y\boldsymbol{x}^{\top}),$$

  with $(Y, \boldsymbol{x}) \sim F$

Ordinary Least Squares, Best Fitting Plane

- The system of equations that we solve to find the OLS estimator is

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\boldsymbol{\beta})\boldsymbol{x}_i^\intercal = \mathbf{0}$$

  which fits into the framework of estimating equations!

- We can take

$$\boldsymbol{G}(\boldsymbol{\beta}, Y, \boldsymbol{x}) = (Y - \boldsymbol{x}\boldsymbol{\beta})\boldsymbol{x}^\intercal$$

  as the basis for a Z-estimator for $\boldsymbol{\beta}$ based on $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} F$

- If we do not impose any assumptions on the distribution F, this leads to the target of inference being

$$\boldsymbol{\beta}_0 : \quad \mathsf{E}_F[(Y - \boldsymbol{x}\boldsymbol{\beta}_0)\boldsymbol{x}^\intercal] = \mathbf{0},$$

  from which we find

$$\boldsymbol{\beta}_0 = \mathsf{E}_F(\boldsymbol{x}^\intercal\boldsymbol{x})^{-1}\mathsf{E}_F(Y\boldsymbol{x}^\intercal),$$

  with $(Y, \boldsymbol{x}) \sim F$

Ordinary Least Squares, Best Fitting Plane

- It can be seen that $\boldsymbol{\beta}_0$ is the solution to the population-level least squares problem

$$\boldsymbol{\beta}_0 = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \;\; \mathrm{E}_F[(Y - \boldsymbol{x}\boldsymbol{\beta}_0)^2]$$

- *Important*: $\boldsymbol{\beta}_0$ characterizes the best fitting plane, in a least squares sense
  - This is well defined across distributions for which $\mathrm{E}_F(\boldsymbol{x}^\top \boldsymbol{x})^{-1}\mathrm{E}_F(Y\boldsymbol{x}^\top)$ is finite
  - It *does not* rely on the assumption $\mathrm{E}(Y \mid \boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}$
  - This estimation approach is *nonparametric*, in the sense that it does not impose stringent assumptions on the distribution F

Ordinary Least Squares, Best Fitting Plane

- In this case, we obtain the $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices as

$$\boldsymbol{A} = \mathrm{E}[\boldsymbol{G}'(\boldsymbol{\beta}_0, Y, \boldsymbol{x})] = -\mathrm{E}[\boldsymbol{x}^\top \boldsymbol{x}]$$

$$\boldsymbol{B} = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\beta}_0, Y, \boldsymbol{x})\boldsymbol{G}(\boldsymbol{\beta}_0, Y, \boldsymbol{x})^\top] = \mathrm{E}[(Y - \boldsymbol{x}\boldsymbol{\beta}_0)^2 \boldsymbol{x}^\top \boldsymbol{x}]$$

- To compute the sandwich estimator we obtain

$$\hat{\boldsymbol{A}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{x}_i = \boldsymbol{X}^\top \boldsymbol{X}/n, \qquad \hat{\boldsymbol{B}} = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}})^2 \boldsymbol{x}_i^\top \boldsymbol{x}_i$$

- The sandwich variance estimator is

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{B}}(\hat{\boldsymbol{A}}^{-1})^\top/n,$$

  which can be used for inference on the parameters $\boldsymbol{\beta}_0$ of the best fitting plane

- However, entries of $\boldsymbol{\beta}_0$ are hard to interpret individually

Ordinary Least Squares, Best Fitting Line

Example: consider least squares with a single covariate: $\boldsymbol{x\beta} = \beta_0 + \beta_1 x$

Problem: Show that the OLS leads to

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \sum_{i=1}^n \sum_{j=1}^n \left(\frac{(x_i - x_j)^2}{\sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^2} \right) \left(\frac{y_i - y_j}{x_i - x_j} \right) \\
&= \sum_{i=1}^n \sum_{j=1}^n w_{ij} \mathrm{slope}_{ij}
\end{aligned}$$

- This expression shows that $\hat{\beta}_1$ is a weighted sum of the slopes of lines connecting all pairs of points
- Weights are proportional to squared x-distances between pairs of points
- What does $\hat{\beta}_1$ estimate?

Ordinary Least Squares, Best Fitting Line

- We can think of the *estimand* $\beta_{1,T}$ as a weighted sum in the (super)population:

$$\beta_{1,T} = \iint \frac{(x-x')^2}{\iint (x-x')^2 dF(x,y)dF(x',y')} \cdot \frac{(y-y')}{(x-x')} dF(x,y)dF(x',y'),$$

where you can think of

  - For X and Y continuous: $\int h(x,y)dF(x,y) = \iint h(x,y)f(x,y)dxdy$
  - For X and Y discrete: $\int h(x,y)dF(x,y) = \sum_{x,y} h(x,y)f(x,y)$

Ordinary Least Squares, Best Fitting Line

- $\beta_{1,T}$ could be a reasonable target of inference
  - Exact linearity of relationship between $\boldsymbol{x}$ and $\mathrm{E}(Y \mid \boldsymbol{x})$ seldom holds
  - The slope of the best fitting line is also interpretable as a weighted average slope, where points farther apart receive more weight
  - Could provide a reasonable answer to questions about average trend in the $\boldsymbol{x} - Y$ relationship
- Again, this is a *nonparametric* approach:
  - No distribution-family assumption
  - No mean-model assumption
  - Estimand does not depend on models
- Note, however, that the term *nonparametric regression* more commonly refers to flexible ways of modeling the regression function $\mathrm{E}(Y \mid x)$

Ordinary Least Squares, Best Fitting Line

- $\beta_{1,T}$ could be a reasonable target of inference
  - Exact linearity of relationship between $\boldsymbol{x}$ and $\mathrm{E}(Y \mid \boldsymbol{x})$ seldom holds
  - The slope of the best fitting line is also interpretable as a weighted average slope, where points farther apart receive more weight
  - Could provide a reasonable answer to questions about average trend in the $\boldsymbol{x} - Y$ relationship
- Again, this is a *nonparametric* approach:
  - No distribution-family assumption
  - No mean-model assumption
  - Estimand does not depend on models
- Note, however, that the term *nonparametric regression* more commonly refers to flexible ways of modeling the regression function $\mathrm{E}(Y \mid x)$

Ordinary Least Squares, Best Fitting Line

- $\beta_{1,T}$ could be a reasonable target of inference
  - Exact linearity of relationship between $\boldsymbol{x}$ and $\mathrm{E}(Y \mid \boldsymbol{x})$ seldom holds
  - The slope of the best fitting line is also interpretable as a weighted average slope, where points farther apart receive more weight
  - Could provide a reasonable answer to questions about average trend in the $\boldsymbol{x} - Y$ relationship
- Again, this is a *nonparametric* approach:
  - No distribution-family assumption
  - No mean-model assumption
  - Estimand does not depend on models
- Note, however, that the term *nonparametric regression* more commonly refers to flexible ways of modeling the regression function $\mathrm{E}(Y \mid \boldsymbol{x})$

OLS for the Linear Model

- In *linear regression* we work under the assumption
$$\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta},$$
or equivalently, $Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \mathrm{E}(\epsilon_i \mid \boldsymbol{x}_i) = 0$

- If, in addition, we assume homoscedasticity, that is,
$$\mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \mathrm{var}(\epsilon_i \mid \boldsymbol{x}_i) = \sigma^2$$
then we obtain a simplification of the previous result:
$$\boldsymbol{B} = \mathrm{E}[(Y - \boldsymbol{x}\boldsymbol{\beta}_0)^2\boldsymbol{x}^\top\boldsymbol{x}] = \mathrm{E}[\epsilon^2\boldsymbol{x}^\top\boldsymbol{x}] = \mathrm{E}[\mathrm{E}(\epsilon^2 \mid \boldsymbol{x})\boldsymbol{x}^\top\boldsymbol{x}] = \sigma^2\mathrm{E}(\boldsymbol{x}^\top\boldsymbol{x})$$

- This allows us to simplify
$$\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^\top/n = \sigma^2\mathrm{E}(\boldsymbol{x}^\top\boldsymbol{x})^{-1}/n,$$
and so we obtain the estimator
$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$$
where the estimator $\hat{\sigma}^2$ can be taken as
$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}})^2$$

OLS for the Linear Model

- In *linear regression* we work under the assumption
$$\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta},$$
or equivalently, $Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \mathrm{E}(\epsilon_i \mid \boldsymbol{x}_i) = 0$

- If, in addition, we assume homoscedasticity, that is,
$$\mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \mathrm{var}(\epsilon_i \mid \boldsymbol{x}_i) = \sigma^2$$
then we obtain a simplification of the previous result:
$$\boldsymbol{B} = \mathrm{E}[(Y - \boldsymbol{x}\boldsymbol{\beta}_0)^2\boldsymbol{x}^\top\boldsymbol{x}] = \mathrm{E}[\epsilon^2\boldsymbol{x}^\top\boldsymbol{x}] = \mathrm{E}[\mathrm{E}(\epsilon^2 \mid \boldsymbol{x})\boldsymbol{x}^\top\boldsymbol{x}] = \sigma^2\mathrm{E}(\boldsymbol{x}^\top\boldsymbol{x})$$

- This allows us to simplify
$$\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^\top/n = \sigma^2\mathrm{E}(\boldsymbol{x}^\top\boldsymbol{x})^{-1}/n,$$
and so we obtain the estimator
$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$$
where the estimator $\hat{\sigma}^2$ can be taken as
$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}})^2$$

OLS for the Linear Model

- Note that we had obtained the OLS and the variance estimator:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \qquad \widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

  under the *normal linear model*, where we in addition assumed normality of the response (of the errors)

- These estimators now appear using an *asymptotic* justification under a *semiparametric regression model*, where we assume
  - $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^n \overset{i.i.d.}{\sim} F$
  - $\text{E}(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{\beta}$
  - $\text{var}(Y_i \mid \boldsymbol{x}_i) = \sigma^2$

- However, these estimators do not depend on the assumption of i.i.d. data (not even needed to think of the covariates as random), and actually they enjoy certain optimality properties in finite samples

OLS for the Linear Model

- Note that we had obtained the OLS and the variance estimator:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \qquad \widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

  under the *normal linear model*, where we in addition assumed normality of the response (of the errors)

- These estimators now appear using an *asymptotic* justification under a *semiparametric regression model*, where we assume
  - $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^n \overset{i.i.d.}{\sim} F$
  - $\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{\beta}$
  - $\mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \sigma^2$

- However, these estimators do not depend on the assumption of i.i.d. data (not even needed to think of the covariates as random), and actually they enjoy certain optimality properties in finite samples

OLS for the Linear Model

Unbiasedness:

- If $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$, then

$$
\begin{aligned}
\mathrm{E}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) &= \mathrm{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \mid \boldsymbol{X}] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}
\end{aligned}
$$

- Notice that this only relies on correct specification of the mean model as $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$

OLS for the Linear Model

Variance:

- If $\text{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$, then

$$\begin{aligned} \text{var}(\hat{\beta} \mid \boldsymbol{X}) &= \text{var}\{(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} \mid \boldsymbol{X}\} \\ &= (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \text{var}(\boldsymbol{Y} \mid \boldsymbol{X}) \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \\ &= \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} \end{aligned}$$

- Notice that this only relies on the homoskedasticity assumption $\text{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$

OLS for the Linear Model: Gauss-Markov Theorem

Theorem (Gauss-Markov): Among all *linear, unbiased* estimators of $\boldsymbol{\beta}$ in the model that assumes

$$\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}, \qquad \mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n, \quad \sigma^2 < \infty,$$

the OLS estimator $\hat{\boldsymbol{\beta}}$ has minimum variance and is unique. Likewise, the estimator $\boldsymbol{z}^T\hat{\boldsymbol{\beta}}$ of $\boldsymbol{z}^T\boldsymbol{\beta}$ has minimum variance and is unique.

Note: no normality assumed, no i.i.d. data assumed, the result holds for finite sample sizes n, and the covariates $\boldsymbol{X}$ are seen as fixed

OLS for the Linear Model: Gauss-Markov Theorem

- Note that to talk about "minimization" of $\text{var}(\hat{\boldsymbol{\beta}})$, for $\dim(\boldsymbol{\beta}) > 1$, we need a way of ordering positive semi-definite matrices

- We use the so-called *Loewner partial order*, which says that $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite

- For covariance of estimators, $\text{var}(\tilde{\boldsymbol{\beta}}) \succeq \text{var}(\hat{\boldsymbol{\beta}})$ is equivalent to $\boldsymbol{z}^T\hat{\boldsymbol{\beta}}$ being optimal for estimating univariate contrasts $\boldsymbol{z}^T\boldsymbol{\beta}$:

$$\begin{aligned} \text{var}(\boldsymbol{z}^T\tilde{\boldsymbol{\beta}}) &= \boldsymbol{z}^T\text{var}(\tilde{\boldsymbol{\beta}})\boldsymbol{z} \\ \text{var}(\boldsymbol{z}^T\hat{\boldsymbol{\beta}}) &= \boldsymbol{z}^T\text{var}(\hat{\boldsymbol{\beta}})\boldsymbol{z}, \end{aligned}$$

  and so by definition of positive semi-definite matrices:

$$\text{var}(\tilde{\boldsymbol{\beta}}) \succeq \text{var}(\hat{\boldsymbol{\beta}})$$

  if and only if, for all vectors $\boldsymbol{z}$,

$$\text{var}(\boldsymbol{z}^T\hat{\boldsymbol{\beta}}) - \text{var}(\boldsymbol{z}^T\tilde{\boldsymbol{\beta}}) = \boldsymbol{z}^T[\text{var}(\tilde{\boldsymbol{\beta}}) - \text{var}(\hat{\boldsymbol{\beta}})]\boldsymbol{z} \geq 0$$

OLS for the Linear Model: Gauss-Markov Theorem

- Note that to talk about "minimization" of $\text{var}(\hat{\boldsymbol{\beta}})$, for $\dim(\boldsymbol{\beta}) > 1$, we need a way of ordering positive semi-definite matrices

- We use the so-called *Loewner partial order*, which says that $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite

- For covariance of estimators, $\text{var}(\tilde{\boldsymbol{\beta}}) \succeq \text{var}(\hat{\boldsymbol{\beta}})$ is equivalent to $\boldsymbol{z}^T\hat{\boldsymbol{\beta}}$ being optimal for estimating univariate contrasts $\boldsymbol{z}^T\boldsymbol{\beta}$:

$$\begin{aligned}\text{var}(\boldsymbol{z}^T\tilde{\boldsymbol{\beta}}) &= \boldsymbol{z}^T\text{var}(\tilde{\boldsymbol{\beta}})\boldsymbol{z} \\ \text{var}(\boldsymbol{z}^T\hat{\boldsymbol{\beta}}) &= \boldsymbol{z}^T\text{var}(\hat{\boldsymbol{\beta}})\boldsymbol{z},\end{aligned}$$

  and so by definition of positive semi-definite matrices:

$$\text{var}(\tilde{\boldsymbol{\beta}}) \succeq \text{var}(\hat{\boldsymbol{\beta}})$$

  if and only if, for all vectors $\boldsymbol{z}$,

$$\text{var}(\boldsymbol{z}^T\hat{\boldsymbol{\beta}}) - \text{var}(\boldsymbol{z}^T\tilde{\boldsymbol{\beta}}) = \boldsymbol{z}^T[\text{var}(\tilde{\boldsymbol{\beta}}) - \text{var}(\hat{\boldsymbol{\beta}})]\boldsymbol{z} \geq 0$$

OLS for the Linear Model: Gauss-Markov Theorem

Proof of Gauss-Markov: Linearity and unbiasedness of an estimator $\tilde{\boldsymbol{\beta}}$ means:

- $\tilde{\boldsymbol{\beta}} = \boldsymbol{C}\boldsymbol{Y}$, where $\boldsymbol{C}$ can depend on $\boldsymbol{X}$ but not on $\boldsymbol{Y}$
- $\mathrm{E}(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{X}) = \boldsymbol{\beta}$

Let $\Delta = \boldsymbol{C} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Then

$$\begin{aligned}
\mathrm{E}(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{X}) &= \mathrm{E}(\boldsymbol{C}\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} = ((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top + \Delta)\boldsymbol{X}\boldsymbol{\beta} \\
&= (\boldsymbol{I}_{k+1} + \Delta\boldsymbol{X})\boldsymbol{\beta}
\end{aligned}$$

$\Rightarrow \Delta\boldsymbol{X} = \mathbf{0}$, since $\mathrm{E}(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{X}) = \boldsymbol{\beta}$.

OLS for the Linear Model: Gauss-Markov Theorem

Proof of Gauss-Markov, cont'd:

$$\begin{aligned}
\mathrm{var}(\tilde{\beta} \mid \boldsymbol{X}) &= \mathrm{var}(\boldsymbol{C}\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{C}\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X})\boldsymbol{C}^\top = \boldsymbol{C}\boldsymbol{I}_n\boldsymbol{C}^\top\sigma^2 = \boldsymbol{C}\boldsymbol{C}^\top\sigma^2 \\
&= ((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^T + \Delta)((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^T + \Delta)^T\sigma^2 \\
&= ((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\Delta^T \\
&\qquad + \Delta\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \Delta\Delta^T)\sigma^2 \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\sigma^2 + \Delta\Delta^T\sigma^2 \\
&= \mathrm{var}(\hat{\beta} \mid \boldsymbol{X}) + \Delta\Delta^T\sigma^2
\end{aligned}$$

Since $\Delta\Delta^T$ is positive semidefinite, this is minimized when $\Delta = \mathbf{0}$ because

$$\mathrm{var}(\boldsymbol{z}^T\tilde{\beta}) = \boldsymbol{z}^T[((\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \Delta\Delta^T)\sigma^2]\boldsymbol{z}.$$

is minimized for $\Delta = \mathbf{0}$.

This also shows that if $\mathrm{var}(\tilde{\beta} \mid \boldsymbol{X}) = \mathrm{var}(\hat{\beta} \mid \boldsymbol{X})$ then
$\Delta = \boldsymbol{C} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top = \mathbf{0} \implies \tilde{\beta} = \boldsymbol{C}\boldsymbol{Y} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} = \hat{\beta}$.

OLS for the Linear Model: Gauss-Markov Theorem

Proof of Gauss-Markov, cont'd:

$$\begin{aligned}
\mathrm{var}(\tilde{\beta} \mid \boldsymbol{X}) &= \mathrm{var}(\boldsymbol{C}\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{C}\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X})\boldsymbol{C}^{\top} = \boldsymbol{C}\boldsymbol{I}_n\boldsymbol{C}^{\top}\sigma^2 = \boldsymbol{C}\boldsymbol{C}^{\top}\sigma^2 \\
&= ((\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^T + \Delta)((\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^T + \Delta)^T\sigma^2 \\
&= ((\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1} + (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\Delta^T \\
&\qquad + \Delta\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1} + \Delta\Delta^T)\sigma^2 \\
&= (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\sigma^2 + \Delta\Delta^T\sigma^2 \\
&= \mathrm{var}(\hat{\beta} \mid \boldsymbol{X}) + \Delta\Delta^T\sigma^2
\end{aligned}$$

Since $\Delta\Delta^T$ is positive semidefinite, this is minimized when $\Delta = \mathbf{0}$ because

$$\mathrm{var}(\boldsymbol{z}^T\tilde{\beta}) = \boldsymbol{z}^T[((\boldsymbol{X}^{\top}\boldsymbol{X})^{-1} + \Delta\Delta^T)\sigma^2]\boldsymbol{z}.$$

is minimized for $\Delta = \mathbf{0}$.

This also shows that if $\mathrm{var}(\tilde{\beta} \mid \boldsymbol{X}) = \mathrm{var}(\hat{\beta} \mid \boldsymbol{X})$ then

$\Delta = \boldsymbol{C} - (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top} = \mathbf{0} \implies \tilde{\beta} = \boldsymbol{C}\boldsymbol{Y} = (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y} = \hat{\beta}.$

OLS for the Linear Model: Gauss-Markov Theorem

Proof of Gauss-Markov, cont'd:

$$\begin{aligned}
\text{var}(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{X}) &= \text{var}(\boldsymbol{C}\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{C}\text{var}(\boldsymbol{Y} \mid \boldsymbol{X})\boldsymbol{C}^\intercal = \boldsymbol{C}\boldsymbol{I}_n\boldsymbol{C}^\intercal\sigma^2 = \boldsymbol{C}\boldsymbol{C}^\intercal\sigma^2 \\
&= ((\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^T + \Delta)((\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^T + \Delta)^T\sigma^2 \\
&= ((\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1} + (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\Delta^T \\
&\qquad + \Delta\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1} + \Delta\Delta^T)\sigma^2 \\
&= (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\sigma^2 + \Delta\Delta^T\sigma^2 \\
&= \text{var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) + \Delta\Delta^T\sigma^2
\end{aligned}$$

Since $\Delta\Delta^T$ is positive semidefinite, this is minimized when $\Delta = \mathbf{0}$ because

$$\text{var}(\boldsymbol{z}^T\tilde{\boldsymbol{\beta}}) = \boldsymbol{z}^T[((\boldsymbol{X}^\intercal\boldsymbol{X})^{-1} + \Delta\Delta^T)\sigma^2]\boldsymbol{z}.$$

is minimized for $\Delta = \mathbf{0}$.
This also shows that if $\text{var}(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{X}) = \text{var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X})$ then
$\Delta = \boldsymbol{C} - (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal = \mathbf{0} \implies \tilde{\boldsymbol{\beta}} = \boldsymbol{C}\boldsymbol{Y} = (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{Y} = \hat{\boldsymbol{\beta}}$.

OLS for the Linear Model: Gauss-Markov Theorem

- Thus, the OLS estimator $\hat{\beta}$ is the *B*est (i.e. lowest variance) *L*inear *U*nbiased *E*stimator (BLUE) of the coefficients in the linear regression model

- Normality of Y (or of errors) not required

- Unbiasedness is important: there exist biased estimators with lower MSE $=$ bias^2 $+$ variance
  - ridge regression
  - lasso
  - Bayesian estimators, etc.

OLS for the Linear Model: Gauss-Markov Theorem

The Gauss-Markov theorem can be interpreted as demonstrating the properties of OLS estimators in a *semiparametric* model:

- Minimal distributional assumptions:
  - $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) < \infty$;
  - $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$, where $\sigma^2 < \infty$
- Mean model: $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$
- $\hat{\boldsymbol{\beta}}$ estimates $\boldsymbol{\beta}$ in this mean model
- The Gauss-Markov Theorem says that among unbiased estimators of $\boldsymbol{\beta}$ in this semiparametric model, the OLS estimator has minimum variance

OLS for the Linear Model: Asymptotic Distribution

What is the asymptotic distribution of the OLS estimator when we see the covariates as fixed or as random but not i.i.d.?

We now consider using OLS for inference in a model with the following characteristics:

- Covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ fixed or not i.i.d.
- Mean model: $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{X}\beta$
- Possibly heteroskedastic, but uncorrelated responses (errors): $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \mathrm{diag}\{\sigma_i^2\}$

Equivalently:

$$Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \mathrm{E}(\epsilon_i \mid \boldsymbol{x}_i) = 0, \quad \mathrm{var}(\epsilon_i \mid \boldsymbol{x}_i) = \sigma_i^2, \quad \mathrm{cov}(\epsilon_i, \epsilon_j) = 0$$

OLS for the Linear Model: Asymptotic Distribution

Under this model, different authors provide conditions that guarantee asymptotic normality of the OLS $\hat{\boldsymbol{\beta}}_n$:

$$\sqrt{n}\boldsymbol{B}_n^{-1/2}\boldsymbol{A}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \quad \xrightarrow{d} \quad \mathrm{N}_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1})$$

where

$$\boldsymbol{A}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\boldsymbol{x}_i^\top \boldsymbol{x}_i), \qquad \boldsymbol{B}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[\sigma_i^2 \boldsymbol{x}_i^\top \boldsymbol{x}_i]$$

where $\sigma_i^2 = \mathrm{E}(\epsilon_i^2 \mid \boldsymbol{x}_i) = \mathrm{E}[(Y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 \mid \boldsymbol{x}_i]$

If the covariates are taken as fixed, then

$$\boldsymbol{A}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{x}_i = \boldsymbol{X}^\top \boldsymbol{X}/n, \qquad \boldsymbol{B}_n = \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 \boldsymbol{x}_i^\top \boldsymbol{x}_i = \boldsymbol{X}^\top \mathrm{diag}\{\sigma_i^2\}\boldsymbol{X}/n$$

OLS for the Linear Model: Asymptotic Distribution

Under this model, different authors provide conditions that guarantee asymptotic normality of the OLS $\hat{\boldsymbol{\beta}}_n$:

$$\sqrt{n}\boldsymbol{B}_n^{-1/2}\boldsymbol{A}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \quad \xrightarrow{d} \quad \mathrm{N}_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1})$$

where

$$\boldsymbol{A}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\boldsymbol{x}_i^\top \boldsymbol{x}_i), \qquad \boldsymbol{B}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[\sigma_i^2 \boldsymbol{x}_i^\top \boldsymbol{x}_i]$$

where $\sigma_i^2 = \mathrm{E}(\epsilon_i^2 \mid \boldsymbol{x}_i) = \mathrm{E}[(Y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 \mid \boldsymbol{x}_i]$

If the covariates are taken as fixed, then

$$\boldsymbol{A}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{x}_i = \boldsymbol{X}^\top \boldsymbol{X}/n, \qquad \boldsymbol{B}_n = \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 \boldsymbol{x}_i^\top \boldsymbol{x}_i = \boldsymbol{X}^\top \mathrm{diag}\{\sigma_i^2\}\boldsymbol{X}/n$$

OLS for the Linear Model: Asymptotic Distribution

This is used as the basis for estimating $\mathrm{var}(\hat{\boldsymbol{\beta}}_n)$ as:

$$\begin{aligned}\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_n) &= \hat{\boldsymbol{A}}_n^{-1}\hat{\boldsymbol{B}}_n(\hat{\boldsymbol{A}}_n^{-1})^\top/n \\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}[\boldsymbol{X}^\top\mathrm{diag}\{(Y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_n)^2\}\boldsymbol{X}](\boldsymbol{X}^\top\boldsymbol{X})^{-1},\end{aligned}$$

which might be found under the names of *sandwich*, *robust*, *heteroscedasticity-consistent*, *Huber-Eicker-White*, or *HC0* estimator

- This estimator was developed separately by three authors: Friedhelm Eicker, Peter J. Huber, and Halbert White
- This sandwich estimator is consistent for $\mathrm{var}(\hat{\boldsymbol{\beta}})$ under the linear model with heteroscedastic errors
- Notice that this estimator has exactly the same form as the sandwich estimator under the "best fitting plane" approach: the two approaches are numerically the same!
    - *Question*: If both approaches are numerically the same, what do we gain by assuming $\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$?

OLS for the Linear Model: Asymptotic Distribution

This is used as the basis for estimating $\mathrm{var}(\hat{\beta}_n)$ as:

$$\begin{aligned}\widehat{\mathrm{var}}(\hat{\beta}_n) &= \hat{\boldsymbol{A}}_n^{-1}\hat{\boldsymbol{B}}_n(\hat{\boldsymbol{A}}_n^{-1})^\intercal/n \\ &= (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}[\boldsymbol{X}^\intercal\mathrm{diag}\{(Y_i - \boldsymbol{x}_i\hat{\beta}_n)^2\}\boldsymbol{X}](\boldsymbol{X}^\intercal\boldsymbol{X})^{-1},\end{aligned}$$

which might be found under the names of *sandwich*, *robust*, *heteroscedasticity-consistent*, *Huber-Eicker-White*, or *HC0* estimator

- This estimator was developed separately by three authors: Friedhelm Eicker, Peter J. Huber, and Halbert White
- This sandwich estimator is consistent for $\mathrm{var}(\hat{\beta})$ under the linear model with heteroscedastic errors
- Notice that this estimator has exactly the same form as the sandwich estimator under the "best fitting plane" approach: the two approaches are numerically the same!
  - *Question*: If both approaches are numerically the same, what do we gain by assuming $\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i\beta$?

OLS for the Linear Model: Asymptotic Distribution

This is used as the basis for estimating var($\hat{\boldsymbol{\beta}}_n$) as:

$$\begin{aligned}\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_n) &= \hat{\boldsymbol{A}}_n^{-1}\hat{\boldsymbol{B}}_n(\hat{\boldsymbol{A}}_n^{-1})^\intercal/n \\ &= (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}[\boldsymbol{X}^\intercal\text{diag}\{(Y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_n)^2\}\boldsymbol{X}](\boldsymbol{X}^\intercal\boldsymbol{X})^{-1},\end{aligned}$$

which might be found under the names of *sandwich*, *robust*, *heteroscedasticity-consistent*, *Huber-Eicker-White*, or *HC0* estimator

- This estimator was developed separately by three authors: Friedhelm Eicker, Peter J. Huber, and Halbert White
- This sandwich estimator is consistent for var($\hat{\boldsymbol{\beta}}$) under the linear model with heteroscedastic errors
- Notice that this estimator has exactly the same form as the sandwich estimator under the "best fitting plane" approach: the two approaches are numerically the same!
    - *Question*: If both approaches are numerically the same, what do we gain by assuming $\text{E}(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$?

OLS for the Linear Model: Asymptotic Distribution

- The sandwich estimator $\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_n)$ is more variable than the estimator under homoscedasticity, $\hat{\sigma}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, in small samples

- Some small-sample fixes have been proposed:

  - HC1: replace $e_i^2 := (Y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}})^2$ with $\frac{n}{n-k-1}e_i^2$

  - HC2: replace e_i^2 with $\frac{e_i^2}{1-h_i}$, where h_i is the ith diagonal element of the "hat" matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$

  - More options in the `sandwich` `R` package

Each leads to consistent estimators of $\text{var}(\hat{\boldsymbol{\beta}})$, see MacKinnon and White (1985).

OLS for the Linear Model: Asymptotic Distribution

- Note that
$$\boldsymbol{B}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\sigma_i^2 \boldsymbol{x}_i^\top \boldsymbol{x}_i)$$
depends on n different variances $\sigma_i^2 = \mathrm{E}(\epsilon_i^2 \mid \boldsymbol{x}_i) = \mathrm{E}[(Y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 \mid \boldsymbol{x}_i]$

- Estimating $\hat{\boldsymbol{B}}_n = \boldsymbol{X}^\top \mathrm{diag}\{(Y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}})^2\}\boldsymbol{X}/n$ is *as if* we took $\hat{\sigma}_i^2 = (Y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}})^2$

- However, estimating n different variances σ_i^2 with n data points is hopeless

- How do theoreticians handled this situation?

- *Idea*: focus on estimating $\boldsymbol{B}_n$, not each individual σ_i^2, and set yourself up for success, i.e., put enough conditions under which you can do a good job at estimating $\boldsymbol{B}_n$ (and $\boldsymbol{A}_n$)

OLS for the Linear Model: Asymptotic Distribution

Some conditions for success: control $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ as $n \to \infty$

- Boos and Stefanski (2013, p. 316) for non-random covariates, assume these limits exist:

$$\lim_{n\to\infty} \boldsymbol{B}_n = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{x}_i^\top \boldsymbol{x}_i = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[(Y_i - \boldsymbol{x}_i \boldsymbol{\beta})^2] \boldsymbol{x}_i^\top \boldsymbol{x}_i$$

$$\lim_{n\to\infty} \boldsymbol{A}_n = \lim_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{x}_i$$

- White (1980) imposes conditions to control the behavior of

$$\boldsymbol{A}_n = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(\boldsymbol{x}_i^\top \boldsymbol{x}_i), \quad \boldsymbol{B}_n = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(\sigma_i^2 \boldsymbol{x}_i^\top \boldsymbol{x}_i) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(\epsilon_i^2 \boldsymbol{x}_i^\top \boldsymbol{x}_i)$$

(interestingly, White doesn't require that these converge to a limit)

OLS for the Linear Model: Asymptotic Distribution

Some conditions for success: control $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ as $n \to \infty$

- Boos and Stefanski (2013, p. 316) for non-random covariates, assume these limits exist:

$$\lim_{n\to\infty} \boldsymbol{B}_n = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 \boldsymbol{x}_i^\top \boldsymbol{x}_i = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[(Y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2]\boldsymbol{x}_i^\top \boldsymbol{x}_i$$

$$\lim_{n\to\infty} \boldsymbol{A}_n = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{x}_i$$

- White (1980) imposes conditions to control the behavior of

$$\boldsymbol{A}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\boldsymbol{x}_i^\top \boldsymbol{x}_i), \quad \boldsymbol{B}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\sigma_i^2 \boldsymbol{x}_i^\top \boldsymbol{x}_i) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\epsilon_i^2 \boldsymbol{x}_i^\top \boldsymbol{x}_i)$$

(interestingly, White doesn't require that these converge to a limit)

OLS for the Linear Model: Asymptotic Distribution

White (1980) imposed the following assumptions:

- $\mathrm{E}(\boldsymbol{x}_i \epsilon_i) = \mathbf{0}$ (particular cases: $\mathrm{E}(\epsilon_i | \boldsymbol{x}_i) = 0$ and $\epsilon_i \perp\!\!\!\perp \boldsymbol{x}_i$)
- There exist constants $\delta_1, \delta_2 > 0$ such that for all i, j, k, $\mathrm{E}(|\epsilon_i^2|^{1+\delta_1}) < \delta_2$, $\mathrm{E}(|X_{ij} X_{ik}|^{1+\delta_1}) < \delta_2$ ($\approx$ uniformly bounded error variances and covariance of covariates, stronger)
- $\boldsymbol{A}_n$ is non-singular for all $n > n_0$, such that $\det \boldsymbol{A}_n > \delta_1 > 0$ (eventually the average covariance matrix of covariates is non-singular, and elements of its inverse are uniformly bounded)
- Similar conditions for $\mathrm{E}(|\epsilon_i^2 X_{ij} X_{ik}|^{1+\delta_1}) < \delta_2$ and for $\boldsymbol{B}_n$ (elements of $\boldsymbol{B}_n$ and of its inverse are uniformly bounded)
- There exist $\delta_1, \delta_2 > 0$ such that for all i, j, k, l, $\mathrm{E}(|X_{ij}^2 X_{ik} X_{il}|^{1+\delta_1}) < \delta_2$

OLS for the Linear Model: Asymptotic Distribution

White (1980) imposed the following assumptions:

- $\mathrm{E}(\boldsymbol{x}_i \epsilon_i) = \mathbf{0}$ (particular cases: $\mathrm{E}(\epsilon_i | \boldsymbol{x}_i) = 0$ and $\epsilon_i \perp\!\!\!\perp \boldsymbol{x}_i$)
- There exist constants $\delta_1, \delta_2 > 0$ such that for all i, j, k, $\mathrm{E}(|\epsilon_i^2|^{1+\delta_1}) < \delta_2$, $\mathrm{E}(|X_{ij} X_{ik}|^{1+\delta_1}) < \delta_2$ ($\approx$ uniformly bounded error variances and covariance of covariates, stronger)
- $\boldsymbol{A}_n$ is non-singular for all $n > n_0$, such that $\det \boldsymbol{A}_n > \delta_1 > 0$ (eventually the average covariance matrix of covariates is non-singular, and elements of its inverse are uniformly bounded)
- Similar conditions for $\mathrm{E}(|\epsilon_i^2 X_{ij} X_{ik}|^{1+\delta_1}) < \delta_2$ and for $\boldsymbol{B}_n$ (elements of $\boldsymbol{B}_n$ and of its inverse are uniformly bounded)
- There exist $\delta_1, \delta_2 > 0$ such that for all i, j, k, l, $\mathrm{E}(|X_{ij}^2 X_{ik} X_{il}|^{1+\delta_1}) < \delta_2$

OLS for the Linear Model: Asymptotic Distribution

- *Lemma 1 of White (1980, Econometrica Vol. 48 No. 4):*

$$\sqrt{n}\boldsymbol{B}_n^{-1/2}\boldsymbol{A}_n(\hat{\beta}-\beta) \xrightarrow{d} \mathrm{N}_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1})$$

- *Theorem 1 of White (1980, Econometrica Vol. 48 No. 4):*
  Let $\hat{\boldsymbol{B}}_n = \boldsymbol{X}^\top \mathrm{diag}\{(Y_i - \boldsymbol{x}_i\hat{\beta})^2\}\boldsymbol{X}/n$. Under the assumptions above

  i) $|\hat{\boldsymbol{B}}_n - \boldsymbol{B}_n| \xrightarrow{a.s.} 0$

  ii) $|(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\hat{\boldsymbol{B}}_n(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1} - \boldsymbol{A}_n^{-1}\boldsymbol{B}_n\boldsymbol{A}_n^{-1}| \xrightarrow{a.s.} 0$

  iii) Under $H_0: \ \boldsymbol{C}\beta = \boldsymbol{b}$, for $\boldsymbol{C}$ of rank q

$$n(\boldsymbol{C}\hat{\beta}-\boldsymbol{b})^T[\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\hat{\boldsymbol{B}}_n(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\hat{\beta}-\boldsymbol{b}) \xrightarrow{d} \chi_q^2$$

OLS for the Linear Model: Asymptotic Distribution

- *Lemma 1 of White (1980, Econometrica Vol. 48 No. 4):*

$$\sqrt{n}\boldsymbol{B}_n^{-1/2}\boldsymbol{A}_n(\hat{\beta}-\beta) \xrightarrow{d} \mathrm{N}_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1})$$

- *Theorem 1 of White (1980, Econometrica Vol. 48 No. 4):*
  Let $\hat{\boldsymbol{B}}_n = \boldsymbol{X}^\top \mathrm{diag}\{(Y_i - \boldsymbol{x}_i\hat{\beta})^2\}\boldsymbol{X}/n$. Under the assumptions above

  i) $|\hat{\boldsymbol{B}}_n - \boldsymbol{B}_n| \xrightarrow{a.s.} 0$

  ii) $|(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\hat{\boldsymbol{B}}_n(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1} - \boldsymbol{A}_n^{-1}\boldsymbol{B}_n\boldsymbol{A}_n^{-1}| \xrightarrow{a.s.} 0$

  iii) Under $H_0: \ \boldsymbol{C}\beta = \boldsymbol{b}$, for $\boldsymbol{C}$ of rank q

$$n(\boldsymbol{C}\hat{\beta}-\boldsymbol{b})^T[\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\hat{\boldsymbol{B}}_n(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\hat{\beta}-\boldsymbol{b}) \xrightarrow{d} \chi^2_q$$

OLS for the Linear Model: Asymptotic Distribution

- *Lemma 1 of White (1980, Econometrica Vol. 48 No. 4):*

$$\sqrt{n}\boldsymbol{B}_n^{-1/2}\boldsymbol{A}_n(\hat{\beta}-\beta) \xrightarrow{d} \mathrm{N}_{k+1}(\mathbf{0}, \boldsymbol{I}_{k+1})$$

- *Theorem 1 of White (1980, Econometrica Vol. 48 No. 4):*
Let $\hat{\boldsymbol{B}}_n = \boldsymbol{X}^\top \mathrm{diag}\{(Y_i - \boldsymbol{x}_i\hat{\beta})^2\}\boldsymbol{X}/n$. Under the assumptions above

   i) $|\hat{\boldsymbol{B}}_n - \boldsymbol{B}_n| \xrightarrow{a.s.} 0$

   ii) $|(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\hat{\boldsymbol{B}}_n(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1} - \boldsymbol{A}_n^{-1}\boldsymbol{B}_n\boldsymbol{A}_n^{-1}| \xrightarrow{a.s.} 0$

   iii) Under $H_0: \ \boldsymbol{C}\beta = \boldsymbol{b}$, for $\boldsymbol{C}$ of rank q

$$n(\boldsymbol{C}\hat{\beta}-\boldsymbol{b})^T[\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\hat{\boldsymbol{B}}_n(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\hat{\beta}-\boldsymbol{b}) \xrightarrow{d} \chi^2_q$$

OLS for the Linear Model: Asymptotic Distribution

- *Lemma 1 of White (1980, Econometrica Vol. 48 No. 4):*

$$\sqrt{n}\boldsymbol{B}_n^{-1/2}\boldsymbol{A}_n(\hat{\beta}-\beta) \xrightarrow{d} \mathrm{N}_{k+1}(\mathbf{0}, \boldsymbol{I}_{k+1})$$

- *Theorem 1 of White (1980, Econometrica Vol. 48 No. 4):*
  Let $\hat{\boldsymbol{B}}_n = \boldsymbol{X}^\top \mathrm{diag}\{(Y_i - \boldsymbol{x}_i\hat{\beta})^2\}\boldsymbol{X}/n$. Under the assumptions above

  i) $|\hat{\boldsymbol{B}}_n - \boldsymbol{B}_n| \xrightarrow{a.s.} 0$

  ii) $|(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\hat{\boldsymbol{B}}_n(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1} - \boldsymbol{A}_n^{-1}\boldsymbol{B}_n\boldsymbol{A}_n^{-1}| \xrightarrow{a.s.} 0$

  iii) Under H_0 : $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{b}$, for $\boldsymbol{C}$ of rank q

$$n(\boldsymbol{C}\hat{\beta}-\boldsymbol{b})^T[\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\hat{\boldsymbol{B}}_n(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\hat{\beta}-\boldsymbol{b}) \xrightarrow{d} \chi^2_q$$

OLS for the Linear Model: Asymptotic Distribution

Key points in the proof of part **i)** $|\hat{\boldsymbol{B}}_n - \boldsymbol{B}_n| \xrightarrow{a.s.} 0$

- Lemma 2.3 of White (1980, Econometrica Vol. 48 No. 3): Let $Z_1, \ldots, Z_n$ be independent random variables. Under some conditions

$$\sup_{\theta} \left| \frac{1}{n} \sum_{i=1}^{n} q_i(Z_i, \theta) - \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[q_i(Z_i, \theta)] \right| \xrightarrow{a.s.} \mathbf{0}$$

- Lemma 2.6 of the same paper: If in addition $\hat{\theta} \xrightarrow{a.s.} \theta$ then

$$\left| \frac{1}{n} \sum_{i=1}^{n} q_i(Z_i, \hat{\theta}) - \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[q_i(Z_i, \theta)] \right| \xrightarrow{a.s.} \mathbf{0}$$

- Since $\hat{\boldsymbol{\beta}} \xrightarrow{a.s.} \boldsymbol{\beta}$, these are combined to show

$$\left| \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}})^2 \boldsymbol{x}_i^T \boldsymbol{x}_i - \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[\epsilon_i^2 \boldsymbol{x}_i^T \boldsymbol{x}_i] \right| \xrightarrow{a.s.} \mathbf{0}$$

where $\epsilon_i = Y_i - \boldsymbol{x}_i \boldsymbol{\beta}$. That is, $|\hat{\boldsymbol{B}}_n - \boldsymbol{B}_n| \xrightarrow{a.s.} 0$.

OLS for the Linear Model: Asymptotic Distribution

Connection of part **iii)** of White's theorem with other tests:

- *Under errors with equal variance* σ^2
  - $(\boldsymbol{X}^\top \boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma \xrightarrow{d} N_{k+1}(\mathbf{0}, \boldsymbol{I}_{k+1})$
  - Then under $H_0: \ \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{b}$

    $(RSS_{H_0} - RSS)/\sigma^2 = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{b})^T[\boldsymbol{C}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{b})/\sigma^2 \xrightarrow{d} \chi^2_q$
  - Also, $\hat{\sigma}^2 = RSS/(n-k-1) \xrightarrow{a.s.} \sigma^2$
  - Then under H_0 we have $(RSS_{H_0} - RSS)/\hat{\sigma}^2 \xrightarrow{d} \chi^2_q$
- *Under i.i.d. normal errors* we had found $(RSS_{H_0} - RSS)/q\hat{\sigma}^2 \sim F_{q,n-k-1}$, so what's the connection?
  - If $Z \sim F(m,n)$ then as $n \to \infty$, $mZ \xrightarrow{d} \chi^2_m$
  - Under H_0 we have $(RSS_{H_0} - RSS)/\hat{\sigma}^2 \sim qF_{q,n-k-1} \xrightarrow{d} \chi^2_q$

OLS for the Linear Model: Asymptotic Distribution

Connection of part **iii)** of White's theorem with other tests:

- *Under errors with equal variance* σ^2
  - $(\boldsymbol{X}^\top \boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma \xrightarrow{d} N_{k+1}(\mathbf{0}, \boldsymbol{I}_{k+1})$
  - Then under $H_0: \ \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{b}$

    $(RSS_{H_0} - RSS)/\sigma^2 = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{b})^T [\boldsymbol{C}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{b})/\sigma^2 \xrightarrow{d} \chi^2_q$
  - Also, $\hat{\sigma}^2 = RSS/(n-k-1) \xrightarrow{a.s.} \sigma^2$
  - Then under H_0 we have $(RSS_{H_0} - RSS)/\hat{\sigma}^2 \xrightarrow{d} \chi^2_q$
- *Under i.i.d. normal errors* we had found $(RSS_{H_0} - RSS)/q\hat{\sigma}^2 \sim F_{q,n-k-1}$, so what's the connection?
  - If $Z \sim F(m,n)$ then as $n \to \infty$, $mZ \xrightarrow{d} \chi^2_m$
  - Under H_0 we have $(RSS_{H_0} - RSS)/\hat{\sigma}^2 \sim qF_{q,n-k-1} \xrightarrow{d} \chi^2_q$

OLS for the Linear Model: Asymptotic Distribution

Key important pieces of White's results for inference:

- Under $H_0: \ \boldsymbol{C}\beta = \boldsymbol{b}$, for $\boldsymbol{C}$ of rank q

$$(\boldsymbol{C}\hat{\beta} - \boldsymbol{b})^T[\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\hat{\boldsymbol{B}}_n(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\hat{\beta} - \boldsymbol{b})/n \xrightarrow{d} \chi^2_q$$

- Confidence regions for $\boldsymbol{C}\beta$

$$\{\boldsymbol{C}\beta : \ \boldsymbol{C}^T(\hat{\beta}-\beta)^T[\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\hat{\boldsymbol{B}}_n(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{C}^T]^{-1}\boldsymbol{C}(\hat{\beta}-\beta)/n \leq \chi^2_q(1-\alpha)\}$$

OLS for the Linear Model: Asymptotic Distribution

Key important pieces of White's results for inference:

- Under $H_0: \ \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{b}$, for $\boldsymbol{C}$ of rank q

$$(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{b})^T[\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\hat{\boldsymbol{B}}_n(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{b})/n \xrightarrow{d} \chi^2_q$$

- Confidence regions for $\boldsymbol{C}\boldsymbol{\beta}$

$$\{\boldsymbol{C}\boldsymbol{\beta}: \ \boldsymbol{C}^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T[\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\hat{\boldsymbol{B}}_n(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{C}^T]^{-1}\boldsymbol{C}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/n \le \chi^2_q(1-\alpha)\}$$

OLS for the Linear Model: Asymptotic Distribution

Particular cases:

- Confidence interval for a single β_j, equivalent to

$$(\hat{\beta}_j - \mathrm{se}(\hat{\beta}_j) z_{1-\alpha/2}, \hat{\beta}_j + \mathrm{se}(\hat{\beta}_j) z_{1-\alpha/2})$$

  with $\mathrm{se}(\hat{\beta}_j)^2$ the jth element of the diagonal of $n(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \hat{\boldsymbol{B}}_n (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$

- Confidence region for $\boldsymbol{\beta}_S = (\beta_j : j \in S, S \subseteq 0 : k)$ can be formed by

$$\{\boldsymbol{\beta}_S : (\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S)^T \widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_S)^{-1} (\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S) \le \chi^2_q(1-\alpha)\}$$

  where

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_S) = n\boldsymbol{C}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \hat{\boldsymbol{B}}_n (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{C}^T$$

  with $\boldsymbol{C}$ a matrix that indicates how to extract the entries S of a vector of length $k+1$

OLS for the Linear Model: Asymptotic Distribution

Particular cases:

- Confidence interval for a single β_j, equivalent to

$$(\hat{\beta}_j - \mathrm{se}(\hat{\beta}_j) z_{1-\alpha/2}, \hat{\beta}_j + \mathrm{se}(\hat{\beta}_j) z_{1-\alpha/2})$$

  with $\mathrm{se}(\hat{\beta}_j)^2$ the jth element of the diagonal of $n(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \hat{\boldsymbol{B}}_n (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$

- Confidence region for $\boldsymbol{\beta}_S = (\beta_j : j \in S, S \subseteq 0 : k)$ can be formed by

$$\{\boldsymbol{\beta}_S : (\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S)^T \widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_S)^{-1} (\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S) \leq \chi^2_q(1-\alpha)\}$$

  where

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_S) = n\boldsymbol{C}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \hat{\boldsymbol{B}}_n (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{C}^T$$

  with $\boldsymbol{C}$ a matrix that indicates how to extract the entries S of a vector of length $k+1$

Summary: Properties of OLS Under Different Assumptions

We covered properties of OLS for different purposes:

OLS for the best fitting plane as a summary of the relationship of response Y_i and covariates $\boldsymbol{x}_i$

- Estimand $\boldsymbol{\beta}$ characterizes best fitting plane, but individual elements in $\boldsymbol{\beta}$ hard to interpret

- OLS $\hat{\boldsymbol{\beta}}_n$ satisfies

$$\sqrt{n}\boldsymbol{B}_n^{-1/2}\boldsymbol{A}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \quad \xrightarrow{d} \quad \mathrm{N}_{k+1}(\mathbf{0}, \boldsymbol{I}_{k+1})$$

- Variance of $\hat{\boldsymbol{\beta}}_n$ can be estimated as

$$\begin{aligned}\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_n) &= \hat{\boldsymbol{A}}_n^{-1}\hat{\boldsymbol{B}}_n(\hat{\boldsymbol{A}}_n^{-1})^\top/n \\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}[\boldsymbol{X}^\top\text{diag}\{(Y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_n)^2\}\boldsymbol{X}](\boldsymbol{X}^\top\boldsymbol{X})^{-1}\end{aligned}$$

Summary: Properties of OLS Under Different Assumptions

OLS for the parameters in the linear mean model (regression function), assumed to be $\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$, allowing for heteroskedastic errors:

$$Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \mathrm{E}(\epsilon_i \mid \boldsymbol{x}_i) = 0, \quad \mathrm{var}(\epsilon_i \mid \boldsymbol{x}_i) = \sigma_i^2, \quad \mathrm{cov}(\epsilon_i, \epsilon_j) = 0$$

- Estimand $\boldsymbol{\beta}$ characterizes the regression function, and individual elements in $\boldsymbol{\beta}$ can be interpretable as usual in linear regression
- OLS $\hat{\boldsymbol{\beta}}_n$ satisfies

$$\sqrt{n}\boldsymbol{B}_n^{-1/2}\boldsymbol{A}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \quad \xrightarrow{d} \quad \mathrm{N}_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1})$$

- Variance of $\hat{\boldsymbol{\beta}}_n$ can be estimated as

$$\begin{aligned}\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_n) &= \hat{\boldsymbol{A}}_n^{-1}\hat{\boldsymbol{B}}_n(\hat{\boldsymbol{A}}_n^{-1})^\intercal/n \\ &= (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}[\boldsymbol{X}^\intercal\mathrm{diag}\{(Y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_n)^2\}\boldsymbol{X}](\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\end{aligned}$$

Summary: Properties of OLS Under Different Assumptions

OLS for the parameters in the linear mean model (regression function), assumed to be $\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$, with homoskedastic errors:

$$Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \mathrm{E}(\epsilon_i \mid \boldsymbol{x}_i) = 0, \quad \mathrm{var}(\epsilon_i \mid \boldsymbol{x}_i) = \sigma^2, \quad \mathrm{cov}(\epsilon_i, \epsilon_j) = 0$$

- Individual elements in $\boldsymbol{\beta}$ can be interpretable as usual in linear regression since we assume $\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$
- OLS $\hat{\boldsymbol{\beta}}_n$ satisfies $(\boldsymbol{X}^\top\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})/\sigma \xrightarrow{d} \mathrm{N}_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1})$
- Variance of $\hat{\boldsymbol{\beta}}_n$ can be estimated as $\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ where the estimator $\hat{\sigma}^2$ can be taken as

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}})^2$$

- OLS $\hat{\boldsymbol{\beta}}_n$ is BLUE (Gauss-Markov)

Summary: Properties of OLS Under Different Assumptions

Back to the first part of the course, OLS for the parameters in the normal linear model:

$$Y_i = \boldsymbol{x}_i \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

- Individual elements in $\boldsymbol{\beta}$ can be interpretable as usual in linear regression since we assume $\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{\beta}$
- OLS $\hat{\boldsymbol{\beta}}_n$ derived as MLE
- OLS $\hat{\boldsymbol{\beta}}_n$ satisfies $(\boldsymbol{X}^\top \boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})/\sigma \sim \mathrm{N}_{k+1}(\mathbf{0}, \boldsymbol{I}_{k+1})$ exactly, in finite samples
- Variance of $\hat{\boldsymbol{\beta}}_n$ can be estimated as $\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ where the estimator $\hat{\sigma}^2$ can be taken as
$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}})^2$$
- Many other exact, finite sample distributional results useful for tests and confidence intervals