Advanced Regression Methods for Independent Data

STAT/BIOST 570, 2020

Regression Models with Weaker Assumptions II

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY of WASHINGTON

Nonlinear Models with Weaker Assumptions

We now consider semiparametric inference for *nonlinear models*:

- We assume
$$Y_i = \mu(\boldsymbol{x}_i, \boldsymbol{\beta}) + \epsilon_i,$$
where:
  - $\mu(\boldsymbol{x}_i, \boldsymbol{\beta}) = \mathsf{E}(Y_i \mid \boldsymbol{x}_i)$: the regression function is given by the functional form $\mu(\boldsymbol{x}_i, \boldsymbol{\beta})$, which is nonlinear in $\boldsymbol{\beta}$
  - $\mathsf{E}(\epsilon_i \mid \boldsymbol{x}_i) = 0$
- This formulation is equivalent to
$$Y_i \mid \boldsymbol{x}_i \sim H_i[\ \mu(\boldsymbol{x}_i, \boldsymbol{\beta})\ ],$$
for some unspecified distribution H_i with mean $\mu(\boldsymbol{x}_i, \boldsymbol{\beta})$

Nonlinear Models with Weaker Assumptions

Some notation:

- For simplicity, we shall write

$$\mu_i := \mu_i(\boldsymbol{\beta}) := \mu(\boldsymbol{x}_i, \boldsymbol{\beta})$$

- We use the vector notation

$$\boldsymbol{\mu}(\boldsymbol{\beta}) := \begin{pmatrix} \mu_1(\boldsymbol{\beta}) \\ \vdots \\ \mu_n(\boldsymbol{\beta}) \end{pmatrix}$$

Estimating Equations for Nonlinear Models

- Inferences on $\boldsymbol{\beta}$ will be obtained from estimating equations with the form:

$$\boldsymbol{G}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i^T(Y_i - \mu_i(\boldsymbol{\beta})) = \frac{1}{n}\boldsymbol{Z}^T(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{0},$$

  - $\boldsymbol{z}_i$: $k+1$ row vector (same length as $\boldsymbol{\beta}$) that might depend on $\boldsymbol{\beta}$ but not on Y_i
  - $\boldsymbol{Z}$: matrix containing the $\boldsymbol{z}_i$'s as its rows

- Note that *under the assumption* $\mu(\boldsymbol{x}_i, \boldsymbol{\beta}) = \mathrm{E}(Y_i \mid \boldsymbol{x}_i)$ we obtain

$$\mathrm{E}[\boldsymbol{G}(\boldsymbol{\beta}, Y_i, \boldsymbol{x}_i)] = \mathrm{E}[\boldsymbol{z}_i^T(Y_i - \mu_i(\boldsymbol{\beta}))] = \mathbf{0}$$

  which we use to simplify some of the sandwich formulae

- These are called *linear unbiased estimating equations*

Least Squares for Nonlinear Models

- Consider the least-squares objective

$$SS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \mu_i(\boldsymbol{\beta}))^2,$$

again with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)^T$

- The minimizer of $SS(\boldsymbol{\beta})$ solves

$$\boldsymbol{G}_n(\boldsymbol{\beta}) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} (Y_i - \mu_i(\boldsymbol{\beta})) = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta_0} (Y_i - \mu_i(\boldsymbol{\beta})) \\ \frac{\partial \mu_i}{\partial \beta_1} (Y_i - \mu_i(\boldsymbol{\beta})) \\ \vdots \\ \frac{\partial \mu_i}{\partial \beta_k} (Y_i - \mu_i(\boldsymbol{\beta})) \end{pmatrix} = \mathbf{0}.$$

Least Squares for Nonlinear Models

- Noting that $\boldsymbol{\mu} : \mathbb{R}^{k+1} \to \mathbb{R}^n$, define

$$\boldsymbol{D} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \boldsymbol{\mu}}{\partial \beta_0}, \quad \cdots, \quad \frac{\partial \boldsymbol{\mu}}{\partial \beta_k} \right) = \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_0} & \cdots & \frac{\partial \mu_1}{\partial \beta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mu_n}{\partial \beta_0} & \cdots & \frac{\partial \mu_n}{\partial \beta_k} \end{pmatrix}$$

  and note that

$$\mu_i'(\boldsymbol{\beta}) := \frac{\partial}{\partial \boldsymbol{\beta}} \mu_i(\boldsymbol{\beta}) = \boldsymbol{D}_{[i,]}^{\top} \quad (\text{with } \boldsymbol{D}_{[i,]} \text{ the } i\text{th row of } \boldsymbol{D}),$$

- We can then express the least squares estimating equation for a nonlinear model in matrix format as

$$\boldsymbol{G}_n(\boldsymbol{\beta}) = \frac{1}{n} \boldsymbol{D}^{\top} (\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{0}$$

- The linear model is a special case where $\boldsymbol{\mu}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{D} = \boldsymbol{X}$, so the estimating equation reduces to the known form

$$\boldsymbol{X}^{\top} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}$$

Least Squares for Nonlinear Models

Different authors provide conditions that guarantee asymptotic normality of the least-squares estimator $\hat{\boldsymbol{\beta}}_n$ in the nonlinear model:

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

where $\boldsymbol{B}_n$ and $\boldsymbol{A}_n$ are defined analogously as before, taking

$$\boldsymbol{G}(\boldsymbol{\beta}, Y_i, \boldsymbol{x}_i) = \mu_i'(\boldsymbol{\beta})(Y_i - \mu_i(\boldsymbol{\beta})),$$

see, e.g. Boos and Stefanski (2013, sec 7.5.3)

Least Squares for Nonlinear Models

- We obtain

$$\boldsymbol{A}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[\boldsymbol{G}'(\beta, Y_i, \boldsymbol{x}_i)]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{\mathrm{E}[Y_i - \mu_i(\boldsymbol{\beta})]\mu_i''(\boldsymbol{\beta}) - \mu_i'(\boldsymbol{\beta})\mu_i'(\boldsymbol{\beta})^T\}$$

where

$$\mu_i''(\boldsymbol{\beta}) = \frac{\partial^2}{\partial\boldsymbol{\beta}^{\top}\partial\boldsymbol{\beta}}\mu_i(\boldsymbol{\beta})$$

- Under the model assumption $\mu(\boldsymbol{x}_i, \boldsymbol{\beta}) = \mathrm{E}(Y_i \mid \boldsymbol{x}_i)$

$$\boldsymbol{A}_n = -\frac{1}{n}\sum_{i=1}^{n}\mu_i'(\boldsymbol{\beta})\mu_i'(\boldsymbol{\beta})^{\top} = -\boldsymbol{D}^{\top}\boldsymbol{D}/n$$

Least Squares for Nonlinear Models

- Similarly,

$$\begin{aligned}\boldsymbol{B}_n &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[\boldsymbol{G}(\boldsymbol{\beta}, Y_i, \boldsymbol{x}_i)\boldsymbol{G}(\boldsymbol{\beta}, Y_i, \boldsymbol{x}_i)^\mathsf{T}] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[(Y_i - \mu_i(\boldsymbol{\beta}))^2]\mu_i'(\boldsymbol{\beta})\mu_i'(\boldsymbol{\beta})^\mathsf{T} \\ &= \boldsymbol{D}^\mathsf{T}\mathrm{diag}\{\mathrm{E}[(Y_i - \mu_i(\boldsymbol{\beta}))^2]\}\boldsymbol{D}/n\end{aligned}$$

- If we assume homoskedasticity, that is, $Y_i = \mu(\boldsymbol{x}_i, \boldsymbol{\beta}) + \epsilon_i$ with $\mathrm{var}(\epsilon_i \mid \boldsymbol{x}_i) = \sigma^2$ then

$$\boldsymbol{B}_n = \sigma^2 \frac{1}{n}\sum_{i=1}^{n} \mu_i'(\boldsymbol{\beta})\mu_i'(\boldsymbol{\beta})^\mathsf{T} = \sigma^2 \boldsymbol{D}^\mathsf{T}\boldsymbol{D}/n$$

Least Squares for Nonlinear Models

The sandwich estimator of the variance of $\hat{\boldsymbol{\beta}}$ is then

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

where for $\hat{\boldsymbol{A}}_n$ we have

- If we allow $\mu(\boldsymbol{x}_i, \boldsymbol{\beta}) \neq \mathrm{E}(Y_i \mid \boldsymbol{x}_i)$, then

$$\hat{\boldsymbol{A}}_n = \frac{1}{n} \sum_{i=1}^{n} \{[Y_i - \mu_i(\hat{\boldsymbol{\beta}})]\mu_i''(\hat{\boldsymbol{\beta}}) - \mu_i'(\hat{\boldsymbol{\beta}})\mu_i'(\hat{\boldsymbol{\beta}})^T\},$$

  for example, if we want to estimate $\boldsymbol{\beta}_0 = \underset{\boldsymbol{\beta}}{\text{argmin}} \; \mathrm{E}_F\{[Y - \mu(\boldsymbol{\beta}, \boldsymbol{x})]^2\}$

- Under the model assumption $\mu(\boldsymbol{x}_i, \boldsymbol{\beta}) = \mathrm{E}(Y_i \mid \boldsymbol{x}_i)$

$$\hat{\boldsymbol{A}}_n = -\frac{1}{n} \sum_{i=1}^{n} \mu_i'(\hat{\boldsymbol{\beta}})\mu_i'(\hat{\boldsymbol{\beta}})^T = -\hat{\boldsymbol{D}}^{\mathsf{T}} \hat{\boldsymbol{D}} / n$$

Least Squares for Nonlinear Models

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

where for $\hat{\boldsymbol{B}}_n$ we have

- In general

$$\begin{aligned} \hat{\boldsymbol{B}}_n &= \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu_i(\hat{\boldsymbol{\beta}}))^2 \mu_i'(\hat{\boldsymbol{\beta}}) \mu_i'(\hat{\boldsymbol{\beta}})^{\mathsf{T}} \\ &= \hat{\boldsymbol{D}}^{\mathsf{T}} \mathrm{diag}\{(Y_i - \mu_i(\hat{\boldsymbol{\beta}}))^2\} \hat{\boldsymbol{D}} / n \end{aligned}$$

- If we assume homoskedasticity, that is, $Y_i = \mu(\boldsymbol{x}_i, \boldsymbol{\beta}) + \epsilon_i$ with $\mathrm{var}(\epsilon_i \mid \boldsymbol{x}_i) = \sigma^2$ then

$$\hat{\boldsymbol{B}}_n = \hat{\sigma}^2 \frac{1}{n} \sum_{i=1}^{n} \mu_i'(\hat{\boldsymbol{\beta}}) \mu_i'(\hat{\boldsymbol{\beta}})^{\mathsf{T}} = \hat{\sigma}^2 \hat{\boldsymbol{D}}^{\mathsf{T}} \hat{\boldsymbol{D}} / n$$

  with

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

Least Squares for Nonlinear Models

- Under the assumption $\mu(\boldsymbol{x}_i, \boldsymbol{\beta}) = \mathrm{E}(Y_i \mid \boldsymbol{x}_i)$, we finally have

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

- Under homoskedasticity

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\hat{\boldsymbol{D}}^\intercal \hat{\boldsymbol{D}})^{-1}$$

with

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

Estimating Equations for Nonlinear Models

Least squares is *not* the only option for obtaining estimating equations for nonlinear models

- Consider, for example,

$$\begin{aligned}
\boldsymbol{G}_n(\boldsymbol{\beta}) &= \frac{1}{n}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \\
&= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^\top (Y_i - \mu_i(\boldsymbol{\beta})) \\
&= \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} (Y_i - \mu_i(\boldsymbol{\beta})) \\ x_{1i}(Y_i - \mu_i(\boldsymbol{\beta})) \\ \vdots \\ x_{ki}(Y_i - \mu_i(\boldsymbol{\beta})) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}
\end{aligned}$$

  where $\mu_i(\boldsymbol{\beta})$ is a general nonlinear function in $\boldsymbol{\beta}$

- Note that these expressions correspond to the score equations of GLMs with canonical links and equal dispersion parameters $\alpha_i = \alpha$ (see slides4.pdf, p. 39), and also to the OLS equations with $\mu_i(\boldsymbol{\beta}) = \boldsymbol{x}_i\boldsymbol{\beta}$

Estimating Equations for Nonlinear Models

- Another estimating equation with this generic form we previously saw was

$$\boldsymbol{S}_n(\boldsymbol{\beta}) = \boldsymbol{D}^\top \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \boldsymbol{0},$$

  which we obtained from GLMs with general link functions, with specification of the mean and variance as

$$\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \mu_i(\boldsymbol{\beta}), \qquad \mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \alpha_i V(\mu_i),$$

  where $\alpha_i = \alpha/\phi_i$ with ϕ_i known, or in matrix form

$$\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{\mu}(\boldsymbol{\beta}), \qquad \mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \alpha \boldsymbol{V},$$

  with $\boldsymbol{V} = \mathrm{diag}\{V(\mu_i)/\phi_i\}$ (see slides4.pdf, p. 25)

Any of these approaches can be used to obtain a Z-estimator and its sandwich covariance matrix, analogously as before!

Estimating Equations for Nonlinear Models

- All the estimating equations seen above have the form:

$$\boldsymbol{G}_n(\boldsymbol{\beta}) = \frac{1}{n}\boldsymbol{Z}^T(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i^T(Y_i - \mu_i(\boldsymbol{\beta})) = \mathbf{0},$$

  where $\boldsymbol{z}_i$ is a row vector that might depend on $\boldsymbol{\beta}$, and $\boldsymbol{Z}$ is a matrix containing the $\boldsymbol{z}_i$'s as its rows

- How to choose $\boldsymbol{Z}$?
  - We might hope to choose $\boldsymbol{Z}$ in a way that assures good properties of our solutions/estimator $\hat{\boldsymbol{\beta}}$
  - One criterion is to select $\boldsymbol{Z}$ to minimize the large sample approximate var$(\hat{\boldsymbol{\beta}})$

Optimal Estimating Equations

- *Question:* If we know both the form of $\boldsymbol{\mu}(\boldsymbol{\beta})$ and also $\text{var}(\boldsymbol{Y} \mid \boldsymbol{X})$, can we derive estimating equations that produce consistent estimates of $\boldsymbol{\beta}$ with smaller variance?

- *Important Result:* Let $\text{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \alpha \boldsymbol{V}$. Among estimating functions of the form $\boldsymbol{G}_n(\boldsymbol{\beta}) = \boldsymbol{Z}^T(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))/n$, setting $\boldsymbol{Z}^T = \boldsymbol{D}^\top \boldsymbol{V}^{-1}$ yields an estimator $\hat{\boldsymbol{\beta}}$ with the smallest asymptotic variance.

- Therefore, if we knew $\text{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \alpha \boldsymbol{V}$, to get more precise estimates of $\boldsymbol{\beta}$, find $\hat{\boldsymbol{\beta}}$ that solves

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \frac{1}{n} \boldsymbol{D}^\top(\boldsymbol{\beta}) \boldsymbol{V}^{-1}(\boldsymbol{\beta})(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \boldsymbol{0}$$

Optimal Estimating Equations

- This result is due to Godambe and Heyde (1987, *International Statistical Review*) who "*state the Gauss-Markov theorem in the framework of estimating function theory*."
- See also Heyde (1997), *Quasi-Likelihood and its Application*, Springer.
- Additional assumptions by Godambe and Heyde (1987):
  - $\mathrm{E}(Y \mid \boldsymbol{x}) < \infty, \quad \mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{X}), \quad \mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) \propto \boldsymbol{V}.$
  - $\mathrm{E}[\frac{\partial \boldsymbol{G}_n}{\partial \boldsymbol{\beta}}]$ is of full rank and $\mathrm{E}[\boldsymbol{G}_n(\boldsymbol{\beta})\boldsymbol{G}_n(\boldsymbol{\beta})^T]$ and $\mathrm{E}[\boldsymbol{U}_n(\boldsymbol{\beta})\boldsymbol{U}_n(\boldsymbol{\beta})^T]$ are positive definite.

Optimal Estimating Equations

Proof outline of the result of Godambe and Heyde:

- Remember, the asymptotic variance of $\hat{\boldsymbol{\beta}}$, the solution of

$$\boldsymbol{G}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\beta}, Y_i, \boldsymbol{x}_i) = \boldsymbol{0},$$

is given by $\boldsymbol{A}_n^{-1}\boldsymbol{B}_n(\boldsymbol{A}_n^{-1})^T/n$, where $\boldsymbol{A}_n$ can be written as

$$\boldsymbol{A}_n := \boldsymbol{A}_G \quad = \quad \mathrm{E}\left[\frac{\partial \boldsymbol{G}_n}{\partial \boldsymbol{\beta}}\right] = \left(\mathrm{E}\left[\frac{\partial \boldsymbol{G}_n}{\partial \beta_0}\right], \ldots, \mathrm{E}\left[\frac{\partial \boldsymbol{G}_n}{\partial \beta_k}\right]\right),$$

with $\mathrm{E}\left[\frac{\partial \boldsymbol{G}_n}{\partial \beta_j}\right]$ a column of length $k+1$, and $\boldsymbol{B}_n$ can be written as

$$\begin{aligned}\boldsymbol{B}_n := \boldsymbol{B}_G \quad &= \quad \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\boldsymbol{G}(\boldsymbol{\beta}, Y_i, \boldsymbol{x}_i)\boldsymbol{G}(\boldsymbol{\beta}, Y_i, \boldsymbol{x}_i)^T\right] \\ &= \quad n\mathrm{E}\left[\boldsymbol{G}_n(\boldsymbol{\beta})\boldsymbol{G}_n(\boldsymbol{\beta})^T\right],\end{aligned}$$

since $\mathrm{E}[\boldsymbol{G}(\boldsymbol{\beta}, Y_i, \boldsymbol{x}_i)] = 0$ by assumption, and

$$\mathrm{E}[\boldsymbol{G}(\boldsymbol{\beta}, Y_i, \boldsymbol{x}_i)\boldsymbol{G}(\boldsymbol{\beta}, Y_j, \boldsymbol{x}_j)^T] = 0,$$

for all $i \neq j$ since the observations are independent.

Optimal Estimating Equations

Proof outline of the result of Godambe and Heyde, cont'd:

- The general linear unbiased estimating equation

  $$\boldsymbol{G}_n(\boldsymbol{\beta}) = \boldsymbol{Z}^T(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))/n = \boldsymbol{0},$$

  leads to

  $$\boldsymbol{A}_G = \boldsymbol{Z}^T\boldsymbol{D}/n,$$
  $$\boldsymbol{B}_G = \alpha\boldsymbol{Z}^T\boldsymbol{V}\boldsymbol{Z}/n,$$

  and asymptotic variance of the solution of $\boldsymbol{G}_n(\boldsymbol{\beta}) = \boldsymbol{0}$, $\hat{\boldsymbol{\beta}}_G$, as

  $$\text{var}(\hat{\boldsymbol{\beta}}_G) = \alpha(\boldsymbol{Z}^T\boldsymbol{D})^{-1}(\boldsymbol{Z}^T\boldsymbol{V}\boldsymbol{Z})(\boldsymbol{Z}^T\boldsymbol{D})^{-1\ T}$$

Optimal Estimating Equations

Proof outline of the result of Godambe and Heyde, cont'd:

- The optimal linear unbiased EE is actually defined up to an invertible constant matrix $\boldsymbol{C}$

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \boldsymbol{C}\boldsymbol{D}^\top \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))/n = \mathbf{0},$$

which leads to

$$\boldsymbol{A}_U = \boldsymbol{C}\boldsymbol{D}^\top \boldsymbol{V}^{-1}\boldsymbol{D}/n,$$

$$\boldsymbol{B}_U = \alpha \boldsymbol{C}\boldsymbol{D}^\top \boldsymbol{V}^{-1}\boldsymbol{D}\boldsymbol{C}^T/n,$$

and asymptotic variance of the solution of $\boldsymbol{U}_n(\boldsymbol{\beta}) = \mathbf{0}$, $\hat{\boldsymbol{\beta}}_U$, as

$$\text{var}(\hat{\boldsymbol{\beta}}_U) = \alpha(\boldsymbol{D}^\top \boldsymbol{V}^{-1}\boldsymbol{D})^{-1} = \alpha^2 \boldsymbol{C}^T \boldsymbol{B}_U^{-1}\boldsymbol{C}/n$$

Optimal Estimating Equations

Proof outline of the result of Godambe and Heyde, cont'd:

- Also, we obtain

$$\mathrm{E}\left[\boldsymbol{U}_n(\boldsymbol{\beta})\boldsymbol{G}_n(\boldsymbol{\beta})^T\right] = \alpha \boldsymbol{C}\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{Z}/n^2 = \alpha \boldsymbol{C}\boldsymbol{D}^T\boldsymbol{Z}/n^2 = \alpha \boldsymbol{C}\boldsymbol{A}_G^T/n$$

- Let

$$\boldsymbol{M} = \begin{pmatrix} \mathrm{E}[\boldsymbol{U}_n(\boldsymbol{\beta})\boldsymbol{U}_n(\boldsymbol{\beta})^T] & \mathrm{E}[\boldsymbol{U}_n(\boldsymbol{\beta})\boldsymbol{G}_n(\boldsymbol{\beta})^T] \\ \mathrm{E}[\boldsymbol{U}_n(\boldsymbol{\beta})\boldsymbol{G}_n(\boldsymbol{\beta})^T]^T & \mathrm{E}[\boldsymbol{G}_n(\boldsymbol{\beta})\boldsymbol{G}_n(\boldsymbol{\beta})^T] \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}_U/n & \alpha\boldsymbol{C}\boldsymbol{A}_G^T/n \\ \alpha\boldsymbol{A}_G\boldsymbol{C}^T/n & \boldsymbol{B}_G/n \end{pmatrix}$$

  be the variance-covariance matrix of $(\boldsymbol{U}_n(\boldsymbol{\beta}), \boldsymbol{G}_n(\boldsymbol{\beta}))^T$, since $\mathrm{E}[\boldsymbol{U}_n(\boldsymbol{\beta})] = \boldsymbol{0}, \mathrm{E}[\boldsymbol{G}_n(\boldsymbol{\beta})] = \boldsymbol{0}$.

- The Schur complement of $\boldsymbol{B}_U/n$ in $\boldsymbol{M}$ is

$$(\boldsymbol{B}_G - \alpha^2\boldsymbol{A}_G\boldsymbol{C}^T\boldsymbol{B}_U^{-1}\boldsymbol{C}\boldsymbol{A}_G^T)/n,$$

  which is non negative definite since $\boldsymbol{M}$ is non negative definite and $\boldsymbol{B}_U$ is assumed to be positive definite (property of the Schur complement)

Optimal Estimating Equations

Proof outline of the result of Godambe and Heyde, cont'd:

- Note that if

$$\boldsymbol{a}^T[\boldsymbol{B}_G - \alpha^2 \boldsymbol{A}_G \boldsymbol{C}^T \boldsymbol{B}_U^{-1} \boldsymbol{C} \boldsymbol{A}_G^T]\boldsymbol{a} \;\geq\; 0$$

  for all $\boldsymbol{a}$, then

$$(\boldsymbol{A}_G^T \boldsymbol{a})^T[(\boldsymbol{A}_G)^{-1}\boldsymbol{B}_G(\boldsymbol{A}_G^T)^{-1} - \alpha^2 \boldsymbol{C}^T \boldsymbol{B}_U^{-1} \boldsymbol{C}](\boldsymbol{A}_G^T \boldsymbol{a}) \;\geq\; 0,$$

  for all $\boldsymbol{a}' = \boldsymbol{A}_G^T \boldsymbol{a}$. Since $\boldsymbol{A}_G^T$ is invertible, $\boldsymbol{a}'$ can be any vector in $\mathbb{R}^{k+1}$.

- Therefore,

$$(\boldsymbol{A}_G)^{-1}\boldsymbol{B}_G(\boldsymbol{A}_G^T)^{-1}/n - \alpha^2 \boldsymbol{C}^T \boldsymbol{B}_U^{-1} \boldsymbol{C}/n = \mathrm{var}(\hat{\boldsymbol{\beta}}_G) - \mathrm{var}(\hat{\boldsymbol{\beta}}_U)$$

  is non negative definite.

This completes the proof of Godambe and Heyde's result.

Optimal Estimating Equations

These results lead to an interesting connection with GLMs:

- Regardless of where the Y_i's live, we could think of a generative model

  $$Y_i = \mu_i(\boldsymbol{\beta}) + \epsilon_i,$$

  which certainly holds for $\epsilon_i = Y_i - \mu_i(\boldsymbol{\beta})$, with $Y_i \mid \boldsymbol{x}_i \sim H_i[\mu_i(\boldsymbol{\beta})]$

- If the Y_i's are independent given covariates, so are the ϵ_i's

- $\mathrm{var}(\epsilon_i \mid \boldsymbol{x}_i) = \mathrm{var}(Y_i \mid \boldsymbol{x}_i)$ and say we assume $\mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \alpha V(\mu_i)$, where $\alpha V(\mu_i)$ coincides with one of the variances in a GLM

- So a GLM can be seen as a nonlinear model $Y_i = \mu_i(\boldsymbol{\beta}) + \epsilon_i$ with a specific mean and variance structure

- The optimal estimating equations of Godambe and Heyde (1987)

  $$\boldsymbol{U}_n(\boldsymbol{\beta}) = \boldsymbol{D}^\top \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})/n = \boldsymbol{0}$$

  correspond to the score equations under maximum likelihood estimation

Reminder: Key Parts of Likelihood Inference for GLMs

When it comes down to obtaining an MLE and its asymptotic distribution in GLMs, the following are the key parts:

- The MLE $\hat{\boldsymbol{\beta}}_n$ is obtained as the solution to

$$\boldsymbol{S}(\boldsymbol{\beta}) \quad = \quad \boldsymbol{D}^T \boldsymbol{V}^{-1} \left[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}) \right] / \alpha = \mathbf{0},$$

  which is derived from the log-likelihood, where

  - $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{\mu}(\boldsymbol{\beta})$ (assumption)
  - $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \alpha \boldsymbol{V}$ with $\boldsymbol{V} = \mathrm{diag}\{V(\mu_i)/\phi_i\}$ (assumption)
  - $\boldsymbol{D}$ is $n \times (k+1)$ with (i, j)th entry $\partial \mu_i / \partial \beta_j$, $i = 1, ..., n$, $j = 0, ..., k$

- The proof that the MLE $\hat{\boldsymbol{\beta}}_n$ has asymptotic distribution

$$\mathcal{I}_n(\boldsymbol{\beta})^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \rightarrow_d \ \mathrm{N}_{k+1}(\mathbf{0}, \boldsymbol{I}_{k+1})$$

  relies on $\mathrm{E}[\boldsymbol{S}(\boldsymbol{\beta})] = \mathbf{0}$ and $-\mathrm{E}[\boldsymbol{S}'(\boldsymbol{\beta})] = \mathrm{E}[\boldsymbol{S}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta})^{\top}] = \mathcal{I}_n(\boldsymbol{\beta})$, see, e.g., Theorem 6.6 in Boos and Stefanski (2013)

Reminder: Key Parts of Likelihood Inference for GLMs

In the context of GLMs:

- To obtain

$$\mathsf{E}[\boldsymbol{S}(\boldsymbol{\beta})] \;\;=\;\; \mathsf{E}\{\boldsymbol{D}^T\boldsymbol{V}^{-1}\left[\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{\beta})\right]/\alpha\} = \mathbf{0},$$

  we need $\boldsymbol{\mu}(\boldsymbol{\beta}) = \mathsf{E}(\boldsymbol{Y} \mid \boldsymbol{X})$

- To obtain $-\mathsf{E}[\boldsymbol{S}'(\boldsymbol{\beta})] = \mathsf{E}[\boldsymbol{S}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta})^\intercal]$, note

$$\begin{aligned}
\mathsf{E}[\boldsymbol{S}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta})^\intercal] &= \boldsymbol{D}^T\boldsymbol{V}^{-1}\mathsf{E}[(\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{\beta}))(\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{\beta}))^T]\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha^2 \\
&= \alpha\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha^2 \\
&= \boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha
\end{aligned}$$

  which holds if $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \alpha\boldsymbol{V}$

Reminder: Key Parts of Likelihood Inference for GLMs

- Similarly,

$$
\begin{aligned}
-\mathsf{E}[\boldsymbol{S}'(\boldsymbol{\beta})] &= -\mathsf{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}\{\boldsymbol{D}^{T}\boldsymbol{V}^{-1}\left[\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{\beta})\right]/\alpha\}\right] \\
&= -\boldsymbol{D}^{T}\boldsymbol{V}^{-1}\mathsf{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}\left[\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{\beta})\right]\right]/\alpha \\
&\quad - \frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}(\boldsymbol{D}^{T}\boldsymbol{V}^{-1})\mathsf{E}\left[\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{\beta})\right]/\alpha \\
&= \boldsymbol{D}^{T}\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha
\end{aligned}
$$

which depends on $\boldsymbol{\mu}(\boldsymbol{\beta}) = \mathsf{E}(\boldsymbol{Y} \mid \boldsymbol{X})$

Reminder: Key Parts of Likelihood Inference for GLMs

Key observations:

- The theory that supports

$$(\boldsymbol{D}^T \boldsymbol{V}^{-1} \boldsymbol{D}/\alpha)^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \;\to_d\; \mathrm{N}_{k+1}(\mathbf{0}, \boldsymbol{I}_{k+1})$$

  for $\hat{\boldsymbol{\beta}}_n$, the solution to

$$\boldsymbol{S}(\boldsymbol{\beta}) \;=\; \boldsymbol{D}^T \boldsymbol{V}^{-1} [\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})] /\alpha = \mathbf{0},$$

  relies mainly on the assumptions that $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{\mu}(\boldsymbol{\beta})$ and $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \alpha \boldsymbol{V}$

- We don't really rely on a specific form for $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{\mu}(\boldsymbol{\beta})$ and $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \alpha \boldsymbol{V}$

- When we use the asymptotic distribution of the MLEs in GLMs, *we don't really rely on the full model specification*: we only use the implied mean and variance functions!

Quasi-Likelihood

Given the previous point of view, there is no reason to restrict yourself to assuming $\text{var}(Y_i \mid \boldsymbol{x}_i) = \alpha V(\mu_i)$ with $\alpha V(\mu_i)$ coming from one of the variances in a GLM: *quasi-likelihood* builds on this idea!

- Proposed by R. W. M. Wedderburn (1974, *Biometrika*)
- An alternative to MLE, when we do not wish to commit to specifying the full distribution of the data

Quasi-Likelihood

- Let Y_i, $i = 1, \dots, n$, have expected values μ_i and variances $\alpha V(\mu_i)$, where $V(\cdot)$ is a known function and $\alpha > 0$ is a scalar

- We assume only a structure for the mean and the variance:

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) &= \boldsymbol{\mu}(\boldsymbol{\beta}) \\ \mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) &= \alpha \boldsymbol{V}[\boldsymbol{\mu}(\boldsymbol{\beta})] \end{aligned}$$

  where

  - $\boldsymbol{\mu}(\boldsymbol{\beta}) = [\mu_1(\boldsymbol{\beta}), \dots, \mu_n(\boldsymbol{\beta})]^\top$, with $\mu_i(\boldsymbol{\beta}) = \mu(\boldsymbol{x}_i, \boldsymbol{\beta})$ representing the regression function

  - $\boldsymbol{V} := \boldsymbol{V}[\boldsymbol{\mu}(\boldsymbol{\beta})] = \mathrm{diag}\{V[\mu_i(\boldsymbol{\beta})]\}$ so that the observations are uncorrelated, and
$$\mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \alpha V[\mu_i(\boldsymbol{\beta})]$$

Quasi-Likelihood

- We use the estimating function or *quasi-score*:

$$\tilde{\boldsymbol{S}}(\boldsymbol{\beta}) = \boldsymbol{D}^\top \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})/\alpha$$

- This quasi-score is such that:
  - $\mathrm{E}[\tilde{\boldsymbol{S}}(\boldsymbol{\beta})] = \mathbf{0}$
  - $\mathrm{var}[\tilde{\boldsymbol{S}}(\boldsymbol{\beta})] = \boldsymbol{D}^\top \boldsymbol{V}^{-1} \boldsymbol{D}/\alpha$
  - $-\mathrm{E}\left[\frac{\partial \tilde{\boldsymbol{S}}}{\partial \boldsymbol{\beta}}\right] = \mathrm{var}[\tilde{\boldsymbol{S}}(\boldsymbol{\beta})] = \boldsymbol{D}^\top \boldsymbol{V}^{-1} \boldsymbol{D}/\alpha$

  under the quasi-likelihood mean and variance assumptions

- These properties are analogous to those obtained under maximum likelihood, thereby the name *quasi*-likelihood.

Quasi-Likelihood

- The word “quasi” also refers to the fact that the score may or may not correspond to a likelihood derived from a probability distribution
- For example, the variance function $V(\mu) = \mu^2(1-\mu)^2$ does not correspond to a probability distribution (see, Wakefield (2013, p. 52))
- Model misspecification can also happen with quasi-likelihood!

Quasi-Likelihood

- The quasi-likelihood estimator satisfies

$$(\boldsymbol{D}^\top \boldsymbol{V}^{-1} \boldsymbol{D}/\alpha)^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \to_d \ \mathrm{N}_{k+1}(\boldsymbol{0}, \boldsymbol{I}_{k+1}),$$

  which is identical to what we obtain for GLMs under maximum likelihood

- The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_n$ is

$$\alpha(\boldsymbol{D}^\top \boldsymbol{V}^{-1} \boldsymbol{D})^{-1},$$

  which we use to obtain the estimator

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_n) = \hat{\alpha}(\hat{\boldsymbol{D}}^\top \hat{\boldsymbol{V}}^{-1} \hat{\boldsymbol{D}})^{-1},$$

  where $\hat{\boldsymbol{D}} = \boldsymbol{D}(\hat{\boldsymbol{\beta}}_n)$, $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\mu}})$, $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_n)$, where $\hat{\mu}_i = \hat{\mu}_i(\hat{\boldsymbol{\beta}}_n)$, and $\hat{\alpha}$ is an estimator of α

Quasi-Likelihood

- Since[1]
$$E[(\boldsymbol{Y}-\boldsymbol{\mu})^\top \boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})] = n\alpha,$$
an unbiased estimator of α would be (with $\boldsymbol{\mu}$ known)
$$\hat{\alpha} = (\boldsymbol{Y}-\boldsymbol{\mu})^\top \boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})/n$$
- For diagonal $\boldsymbol{V}$ and replacing $\boldsymbol{\mu}$ by $\hat{\boldsymbol{\mu}}$:
$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$
- As usual, it is common to use a degrees-of-freedom-corrected (but not, in general, unbiased) estimator
$$\hat{\alpha} = \frac{1}{n-k-1}\sum_{i=1}^{n}\frac{(Y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

[1] *Reminder*: suppose $\boldsymbol{Z}$ is an $n \times 1$ random variable with $E[\boldsymbol{Z}] = \boldsymbol{\mu}$, $\mathrm{var}(\boldsymbol{Z}) = \Sigma$ and $\boldsymbol{A}$ is a symmetric $n \times n$ matrix. Then
$$E[\boldsymbol{Z}^\top \boldsymbol{A}\boldsymbol{Z}] = \mathrm{tr}(\boldsymbol{A}\Sigma) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$$

Quasi-Likelihood and Overdispersion

- Quasi-likelihood is often introduced in the context of *overdispersion*: greater variability than expected under a given statistical model

- GLMs with $\alpha = 1$ in $\mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \alpha V(\mu_i)$ are particularly vulnerable to mean-variance model misspecification, since they do not have a way of absorbing extra variability of the response

- For instance, it is common to find overdispersion with respect to the Poisson ($\mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \mu_i$) and binomial ($\mathrm{var}(Y_i \mid \boldsymbol{x}_i) = n_i\mu_i(1 - \mu_i)$) GLMs, since their variances are entirely determined by their means

- In such cases, a simple implementation of quasi-likelihood corresponds to taking
  - $\mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \alpha\mu_i$ for overdispersed Poisson data
  - $\mathrm{var}(Y_i \mid \boldsymbol{x}_i) = \alpha n_i\mu_i(1 - \mu_i)$ for overdispersed binomial data

Sandwich Estimation from Misspecified GLMs

- The asymptotic variance-covariance matrices for $\hat{\boldsymbol{\beta}}_n$ under likelihood and quasi-likelihood are appropriate only if the first two moments are correctly specified

- Let us take the score or quasi-score to be

$$\tilde{\boldsymbol{S}}(\boldsymbol{\beta}) = \boldsymbol{D}^\top \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \quad \propto \quad \boldsymbol{D}^\top \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})/\alpha$$

  where we ignore α as it cancels in $\tilde{\boldsymbol{S}}(\boldsymbol{\beta}) = \mathbf{0}$

- In general, the asymptotic variance-covariance matrix for $\hat{\boldsymbol{\beta}}_n$ is

$$\boldsymbol{A}_n^{-1} \boldsymbol{B}_n (\boldsymbol{A}_n^\top)^{-1}/n$$

Sandwich Estimation from Misspecified GLMs

- In general,

$$
\begin{aligned}
n\boldsymbol{A}_n = \mathrm{E}[\tilde{\boldsymbol{S}}'(\boldsymbol{\beta})] &= \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}\{\boldsymbol{D}^T\boldsymbol{V}^{-1}[\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{\beta})]\}\right] \\
&= \boldsymbol{D}^T\boldsymbol{V}^{-1}\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}[\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{\beta})]\right] \\
&\qquad + \frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}(\boldsymbol{D}^T\boldsymbol{V}^{-1})\mathrm{E}[\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{\beta})] \\
&= -\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D} + \frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}(\boldsymbol{D}^T\boldsymbol{V}^{-1})\mathrm{E}[\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{\beta})]
\end{aligned}
$$

- If we assume $\boldsymbol{\mu}(\boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X})$, we can take $\boldsymbol{A}_n$ as

$$
\boldsymbol{A}_n = \boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}/n
$$

since the -1 cancels in the sandwich formula

Sandwich Estimation from Misspecified GLMs

Note that in:

$$nA_n = \mathrm{E}[\tilde{S}'(\boldsymbol{\beta})] = \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}\{\boldsymbol{D}^T\boldsymbol{V}^{-1}[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})]\}\right] = -\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D} + \frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}(\boldsymbol{D}^T\boldsymbol{V}^{-1})\mathrm{E}\left[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right],$$

we have that

$$\frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}(\boldsymbol{D}^T\boldsymbol{V}^{-1}) := \boldsymbol{Q}$$

is a $(k+1) \times (k+1) \times n$ tensor, so that $\boldsymbol{Q}[, , i]$ is a $(k+1) \times (k+1)$ matrix

$$\boldsymbol{Q}[, , i] = \frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}\left(\frac{1}{V[\mu_i(\boldsymbol{\beta})]}\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) = \left[\frac{\partial}{\partial \beta_0}\left(\frac{1}{V[\mu_i(\boldsymbol{\beta})]}\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right), \ldots, \frac{\partial}{\partial \beta_k}\left(\frac{1}{V[\mu_i(\boldsymbol{\beta})]}\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\right].$$

To see this, we can write down

$$\tilde{S}(\boldsymbol{\beta}) = \boldsymbol{D}^T\boldsymbol{V}^{-1}[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})] = \sum_{i=1}^{n}\frac{[Y_i - \mu_i(\boldsymbol{\beta})]}{V[\mu_i(\boldsymbol{\beta})]}\left(\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)$$

so that

$$\begin{aligned}
\mathrm{E}[\tilde{S}'(\boldsymbol{\beta})] &= -\sum_{i=1}^{n}\frac{1}{V[\mu_i(\boldsymbol{\beta})]}\left(\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{\mathsf{T}} + \sum_{i=1}^{n}\mathrm{E}\left[Y_i - \mu_i(\boldsymbol{\beta})\right]\frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}\left(\frac{1}{V[\mu_i(\boldsymbol{\beta})]}\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) \\
&= -\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D} + \frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}(\boldsymbol{D}^T\boldsymbol{V}^{-1})\mathrm{E}\left[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right].
\end{aligned}$$

Sandwich Estimation from Misspecified GLMs

- We can write down

$$\tilde{\boldsymbol{S}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{[Y_i - \mu_i(\boldsymbol{\beta})]}{V[\mu_i(\boldsymbol{\beta})]} \left(\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) := \sum_{i=1}^{n} \tilde{S}_i(\boldsymbol{\beta})$$

- In general,

$$\boldsymbol{B}_n = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[\tilde{S}_i(\boldsymbol{\beta}) \tilde{S}_i(\boldsymbol{\beta})^{\mathsf{T}}]$$

- If we assume $\boldsymbol{\mu}(\boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X})$, since we have uncorrelated data, we can write

$$\boldsymbol{B}_n = \frac{1}{n} \mathrm{E}[\tilde{\boldsymbol{S}}(\boldsymbol{\beta}) \tilde{\boldsymbol{S}}(\boldsymbol{\beta})^{\mathsf{T}}] = \frac{1}{n} \boldsymbol{D}^T \boldsymbol{V}^{-1} \mathrm{E}[(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))^T] \boldsymbol{V}^{-1} \boldsymbol{D}$$

  and

$$\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \mathrm{E}[(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))^T]$$

- If we assume $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \alpha \boldsymbol{V}$ then

$$\boldsymbol{B}_n = \alpha \boldsymbol{D}^T \boldsymbol{V}^{-1} \boldsymbol{V} \boldsymbol{V}^{-1} \boldsymbol{D}/n = \alpha \boldsymbol{D}^T \boldsymbol{V}^{-1} \boldsymbol{D}/n$$

Sandwich Estimation from Misspecified GLMs

- Defining $\hat{\boldsymbol{A}}_n$ and $\hat{\boldsymbol{B}}_n$ by plugging $\hat{\boldsymbol{\beta}}_n$ into the expressions for $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ leads to the sandwich estimator of the variance-covariance matrix for $\hat{\boldsymbol{\beta}}_n$ as

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_n) = \hat{\boldsymbol{A}}_n^{-1} \hat{\boldsymbol{B}}_n (\hat{\boldsymbol{A}}_n^{\mathsf{T}})^{-1}/n$$

- This sandwich estimator $\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_n)$ provides a consistent estimator of the variance $\hat{\boldsymbol{\beta}}_n$ and therefore asymptotically correct confidence interval coverage (as long as independence of responses holds)

- For more details see Kauermann and Carroll (2001, *JASA*)

Sandwich Estimation from Misspecified GLMs

- For instance, the most general version of $\hat{\boldsymbol{B}}_n$ for uncorrelated data is

$$
\begin{aligned}
\hat{\boldsymbol{B}}_n &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)\Bigg|_{\hat{\boldsymbol{\beta}}_n} \frac{[Y_i - \mu_i(\hat{\boldsymbol{\beta}}_n)]^2}{(V[\mu_i(\hat{\boldsymbol{\beta}}_n)])^2} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)\Bigg|_{\hat{\boldsymbol{\beta}}_n}^{\mathsf{T}} \\
&= \frac{1}{n}\hat{\boldsymbol{D}}^T\hat{\boldsymbol{V}}^{-1}\mathrm{diag}\{[Y_i - \mu_i(\hat{\boldsymbol{\beta}}_n)]^2\}\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{D}}
\end{aligned}
$$

- And the most general version of $\hat{\boldsymbol{A}}_n$ is

$$
\hat{\boldsymbol{A}}_n = -\frac{1}{n}\hat{\boldsymbol{D}}^T\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{D}} + \frac{1}{n}\left[\frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}(\boldsymbol{D}^T\boldsymbol{V}^{-1})\right]\Bigg|_{\hat{\boldsymbol{\beta}}_n}\left[\boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_n)\right]
$$

  which simplifies to $\hat{\boldsymbol{A}}_n = -\hat{\boldsymbol{D}}^T\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{D}}/n$ if the mean model is correctly specified, which we often assume for interpretability (again, we could ignore the -1 in $\hat{\boldsymbol{A}}_n$ since it cancels in the sandwich formula)

Final Comments

These results are very powerful!

- Say you want to fit a regression model where you assume $\mu(\boldsymbol{x}_i, \boldsymbol{\beta}) = \mathrm{E}(Y_i \mid \boldsymbol{x}_i)$

- The optimality theory of Godambe and Heyde (1987) tells you that, for the sake of efficiency, you should use as your estimating function the quasi-score:

$$\tilde{\boldsymbol{S}}(\boldsymbol{\beta}) \quad = \quad \boldsymbol{D}^T \boldsymbol{V}^{-1} [\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})] = \boldsymbol{0},$$

  where $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{\mu}(\boldsymbol{\beta})$ and $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X}) = \alpha \boldsymbol{V}$

- However, the theory of estimating equations has your back in case you misspecify $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{X})$!

- So, you can choose a *working variance* $\alpha \boldsymbol{V}$ to specify your estimating function, but use sandwich estimators, just in case you are wrong!

- A more general version of this idea is heavily used in *generalized estimating equations* (STAT/BIOST 571)