General Linear Model

Biostatistics 653

Applied Statistics III: Longitudinal Data Analysis

Estimation Without Distributional Assumptions

- Now we take a step back and assume that we only know

$$E(Y_i) = X_i\beta, V(Y_i) = \Sigma$$

for i = 1,…,N. Then the weighted least squares estimator (a.k.a. generalized least squares estimator), $\hat{\beta}$, minimizes the function

$$Q_W(\beta) = \sum_{i=1}^{N}(Y_i - X_i\beta)^T W_i(Y_i - X_i\beta)$$

where $W_i = W(X_i, \theta)$ are positive definite and symmetric matrices of weights

Estimation Without Distributional Assumptions

- If the minimum exits, then it solves

$$\frac{\partial Q_W(\beta)}{\partial \beta} = -2\sum_{i=1}^{N} X_i^T W_i (Y_i - X_i \beta) = 0$$

so that

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1} \sum_{i=1}^{N} X_i^T W_i Y_i$$

Estimation Without Distributional Assumptions

- Note that the contribution of each data vector to $\hat{\beta}$ is being weighted. A popular choice for the weights is $W_i = \Sigma_i^{-1}$ (typically, one would use an estimate of Σ). In that case, data vectors with "more variation" would be weighted less.

- We might use other choices of the weights (for example, if the outcomes are blood pressure measurements, and investigators report a mean of three values at each time for each subject, we might want W_i to include information on the variability of those means at each time).

Estimation With Known Weights

- Now, assuming W_i is known (i.e. not estimated from data, so it is not a function of Y)

$$E(\hat{\beta}) = \left(\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1} \sum_{i=1}^{N} X_i^T W_i E(Y_i)$$
$$= \left(\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1} \sum_{i=1}^{N} X_i^T W_i X_i \beta = \beta$$

so that $\hat{\beta}$ is unbiased for any choice of weight functions $W_1, \cdots, W_N$. In addition, $\hat{\beta}$ is asymptotically normal.

Estimation With Known Weights

- The variance $V(\hat{\beta})$ is

$$= \left(\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1} \sum_{i=1}^{N} X_i^T W_i V(Y_i) W_i X_i \left(\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1}$$

$$= \left(\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1} \sum_{i=1}^{N} X_i^T W_i \Sigma W_i X_i \left(\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1}$$

Estimation Without Distributional Assumptions

- By the Gauss-Markov theorem, the WLS estimator that uses $W_i = \Sigma^{-1}$ has the smallest variance among all WLS estimators. This variance is given by

$$V\left(\hat{\beta}_{\Sigma^{-1}}\right) = \left(\sum_{i=1}^{N} X_i^T \Sigma^{-1} X_i\right)^{-1}$$

- Thus the optimal estimator depends on knowing Σ. In practice, we will typically have to estimate Σ using $\hat{\Sigma}$.

Estimation Without Distributional Assumptions

- If we choose $W_i = I$, we have

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i^T X_i\right)^{-1} \sum_{i=1}^{N} X_i^T Y_i$$

$$E\left(\hat{\beta}\right) = \beta$$

- This implies that ordinary least squares regression based on identity weights leads to unbiased estimation of β. However, calculating its variance assuming independence of all observations leads to incorrect inferences.

Estimation Without Distributional Assumptions

- Instead, when choosing $W_i = I$, we use the variance estimator

$$\mathrm{V}(\hat{\beta}_I) = \left(\sum_{i=1}^{N} X_i^T X_i\right)^{-1} \sum_{i=1}^{N} X_i^T \Sigma X_i \left(\sum_{i=1}^{N} X_i^T X_i\right)^{-1}$$

- In this case, $\hat{\beta}_I$ is known as the generalized estimating equations estimator of based on the working independence assumption.

Data-dependent Weight functions

- Most likely, we will not know W_i, and we need to estimate the weights. We could allow these weights to depend on covariates, obtaining

$$W_i = W(\theta; X_i)$$

- Because we will estimate θ, we will investigate the properties of using $\widehat{W}_i = W(\hat{\theta}; X_i)$ to estimate β.

- If $V(Y_i) = \Sigma$ for all i, we may wish to take $\widehat{W}_i = \Sigma^{-1}$, in which case W_i does not depend on X_i, and we take θ to include the parameters of Σ.

Data-dependent Weight functions

- Using $\widehat{W}_i = W(\hat{\theta}; X_i)$, we obtain

$$\hat{\beta}_{\widehat{W}} = \left(\sum_{i=1}^{N} X_i^T \widehat{W}_i X_i\right)^{-1} \sum_{i=1}^{N} X_i^T \widehat{W}_i Y_i$$

- $\hat{\beta}_{\widehat{W}}$ is not necessarily unbiased. In particular

$$E(\hat{\beta}_{\widehat{W}}) = E[\left(\sum_{i=1}^{N} X_i^T \widehat{W}_i X_i\right)^{-1} \sum_{i=1}^{N} X_i^T \widehat{W}_i Y_i]$$

- and the estimated weights $\widehat{W}_i$ may depend on the data

Data-dependent Weight functions

- We have asymptotic normality

$$\sqrt{N}\left(\hat{\beta}_{\widehat{W}} - \beta\right) \rightarrow^{d} N(0, C_w)$$

Where $C_w = \Gamma_W^{-1}\Omega_W\Gamma_W^{-1}, \Gamma_W = E\left(X_i^T W(\theta^*; X_i)X_i\right), \Omega_W = E\left(X_i^T W(\theta^*; X_i)\Sigma \mathrm{W}(\theta^*; X_i)\mathrm{X_i}\right)$, and $\theta^* = \lim_{N\rightarrow\infty} \widehat{\theta}$

- When the sample size N is finite, this result is just approximate. If N is moderately large, this is a good approximation. However, it is tough to determine how “large” is “large” enough.

Data-dependent Weight functions

- If we knew θ^* and used $W_i = W(\theta^*; X_i)$ to estimate β by $\hat{\beta}_W = \left(\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1} \sum_{i=1}^{N} X_i^T W_i Y_i$. Then, we also have

$$\sqrt{N}\left(\hat{\beta}_W - \beta\right) \to N(0, C_w)$$

- So that we get the same asymptotic distribution whether we estimate the weights or not.
- However, in finite samples, equality of the variances does not hold, and $\widehat{W}_i$ may be quite poorly estimated in small samples.

Estimation of Optimal Weights

- When $Var(Y_i) = \Sigma$ for all i, we can easily obtain a consistent estimator of Σ as follows.
- Let $\hat{\beta}_W$ be a weighted least squares estimator of β for fixed and known (but arbitrary) W. Then, under mild regularity conditions,

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - X_i\hat{\beta}_W\right)\left(Y_i - X_i\hat{\beta}_W\right)^T$$

is a consistent estimator of Σ.

- Then, we could use:

$$\widehat{W}_i = \hat{\Sigma}^{-1}$$

Estimation of Optimal Weights

- In this case, $\hat{\beta}_{\hat{\Sigma}^{-1}}$ is a two-step estimator: (1) Choose W and estimate $\hat{\beta}_W$ and $\hat{\Sigma}$; (2) Obtain $\hat{\beta}_{\hat{\Sigma}^{-1}}$ and $\hat{C}_{\hat{\Sigma}^{-1}}$.

- You could iterate to get the MLE's of β and Σ under normality in the balanced, complete data case.

- This estimator is called a generalized least squares estimator of β because we do not know Σ but must estimate its parameters.

Estimation of Optimal Weights

- Substituting $W_i = \Sigma^{-1}$, we still have asymptotic normality

$$\sqrt{N}\left(\hat{\beta}_{\hat{\Sigma}^{-1}} - \beta\right) \to N(0, C_{\Sigma^{-1}})$$

where $C_{\Sigma^{-1}} = \left[E\left(X_i^T \Sigma^{-1} X_i\right)\right]^{-1}$.

- Thus an estimate of the asymptotic variance of $\hat{\beta}_{\hat{\Sigma}^{-1}}$ is given by a consistent estimator of $C_{\Sigma^{-1}}$, so that

$$\widehat{Var}\left(\sqrt{N}\left(\hat{\beta}_{\hat{\Sigma}^{-1}} - \beta\right)\right) = C_{\hat{\Sigma}^{-1}} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i^T \hat{\Sigma}^{-1} X_i\right)^{-1}$$

$$\widehat{Var}\left(\hat{\beta}_{\hat{\Sigma}^{-1}}\right) = \left(\sum_{i=1}^{N} X_i^T \hat{\Sigma}^{-1} X_i\right)^{-1}$$

Estimation of Optimal Weights

- This is called the model-based variance because it assumes our models for the mean and covariance are correct (e.g., $\lim_{N \to \infty} \hat{\Sigma} = Var(Y_i)$).

- Note that there are n(n+1)/2 free parameters in Σ. For large n, we would need a very large N for the asymptotic distribution of $\hat{\beta}_{\hat{\Sigma}^{-1}}$ to be a good approximation of its finite sample distribution.

- When we can posit a more parsimonious model for $\Sigma = Var(Y_i)$, we can obtain asymptotically efficient WLS estimators of β with better small-sample properties.

Example: Compound Symmetry

- Suppose $\Sigma = \sigma_E^2 I + \sigma_B^2 J J^T$, where J is a column of 1s.

- In this case, we have only two parameters in Σ to estimate, which we can do consistently as follows.

- Choose some W and estimate $\hat{\beta}_W$ and $\hat{\Sigma}$ (unstructured). Then, a consistent estimator of σ_B^2 is obtained by averaging the off-diagonal elements of $\hat{\Sigma}$, and a consistent estimator of $\sigma_E^2 + \sigma_B^2$ is obtained by averaging the diagonal elements of $\hat{\Sigma}$.

- Thus $(\sigma_E^2 + \sigma_B^2)$ is estimated by $\frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n}\left(Y_{ij} - X_{ij}^T\hat{\beta}_W\right)^2$, and σ_B^2 is estimated by $\frac{2}{Nn(n-1)}\sum_{i=1}^{N}\sum_{1\le j<l\le n}(Y_{ij} - X_{ij}^T\hat{\beta}_W)(Y_{il} - X_{il}^T\hat{\beta}_W)$

Example: Compound Symmetry

- After estimating σ_E^2, σ_B^2, we can use $\widehat{W}_i = \left[\hat{\Sigma}_{(\hat{\sigma}_E^2, \hat{\sigma}_B^2)}\right]^{-1}$ to calculate $\hat{\beta}_{\widehat{W}}$ and $\hat{C}_{\widehat{W}}$.

- So $\hat{\beta}_{\widehat{W}}$ is a three-step estimator: (1) Fix W (possibly I) and estimate $\hat{\beta}_W$ and $\hat{\Sigma}$ (unstructured); (2) Estimate θ from $\hat{\Sigma}$, where θ contains the parameters in the model for $\Sigma = Var(Y)$, e.g. $\theta = (\sigma_E^2, \sigma_B^2)$ for compound symmetry; (3) Use $\widehat{\mathrm{W}}_i = \hat{\Sigma}_{\hat{\theta}}^{-1}$ and WLS to estimate $\hat{\beta}_{\widehat{W}}$ and calculate $\hat{C}_{\widehat{W}}$. We can iterate, though estimates generally will not converge to ML.

Example: Compound Symmetry

- The estimator $\hat{\beta}_{\widehat{W}}$ is consistent and asymptotically normally distributed. We call it a locally optimal weighted least squares (LOWLS) estimator because it has asymptotic variance equal to the lower bound of the variances of all WLS estimators when the model for $W^{-1} = Var(Y_i)$ is correctly specified. Even when the covariance model is not correctly specified, it remains consistent and asymptotically normally distributed, though the appropriate variance depends on whether the covariance model is correctly specified.
- These estimates are also known as generalized estimating equations (GEE) estimators.

Variance Estimation

- We know that

$$\sqrt{N}\left(\hat{\beta}_{\hat{\Sigma}^{-1}} - \beta\right) \rightarrow N(0, C_{\Sigma^{-1}})$$

- When $\hat{\Sigma}$ is a consistent estimator of Σ (that is, when the model used to estimate $\Sigma = Var(Y_i)$ is correctly specified, and our estimate of b is consistent). When this is true, a consistent estimator of $C_{\Sigma^{-1}}$ is given by the model-based variance:

$$C_{\hat{\Sigma}^{-1}} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i^T \hat{\Sigma}^{-1} X_i\right)^{-1}$$

- What happens if our model for Σ is wrong?

Variance Estimation

- For example, we could assume compound symmetry with constant variance over time, but in fact variances could increase over time, and correlations might decrease over time. If the model for $\Sigma \neq Var(Y_i)$, then $\sqrt{N}\left(\hat{\beta}_{\hat{\Sigma}^{-1}} - \beta\right)$ has the same asymptotic variance as $\sqrt{N}\left(\hat{\beta}_{\Sigma^{*-1}} - \beta\right)$, where $\Sigma^* = \lim_{N \to \infty} \hat{\Sigma}$

- This asymptotic variance equals

$$C_{\Sigma^{*-1}} = \Gamma_{\Sigma^{*-1}}^{-1} \Omega_{\Sigma^{*-1}} \Gamma_{\Sigma^{*-1}}^{-1}$$

Variance Estimation

- Consistent estimators of $\Gamma_{\Sigma^{*-1}}, \Omega_{\Sigma^{*-1}}$ are given by

$$\hat{\Gamma}_{\Sigma^{*-1}} = \frac{1}{N}\sum_{i=1}^{N} X_i^T \hat{\Sigma}^{-1} X_i$$

$$\hat{\Omega}_{\Sigma^{*-1}} = \frac{1}{N}\sum_{i=1}^{N} X_i^T \hat{\Sigma}^{-1}\left(Y_i - X_i\hat{\beta}\right)\left(Y_i - X_i\hat{\beta}\right)^T \hat{\Sigma}^{-1} X_i$$

- So a consistent estimator of $C_{\Sigma^{*-1}}$ is given by

$$\hat{C}_{\Sigma^{*-1}} = \hat{\Gamma}_{\Sigma^{*-1}}^{-1}\hat{\Omega}_{\Sigma^{*-1}}\hat{\Gamma}_{\Sigma^{*-1}}^{-1}$$

- This estimator is due to Huber (1967) and White (1980) and was also used by Liang and Zeger (1986) in developing GEE.

Variance Estimation

- The variance estimator $\hat{C}_{\Sigma^{-1}}$ is called the model-based variance estimator, as its consistency relies on the correct specification of Var(Yi). Sometimes it is called the naive variance estimate, though this isn't a great term.

- The variance estimator $\hat{C}_{\Sigma^{*-1}}$ is called the robust, empirical, or sandwich variance estimator and is consistent even when the model for Var(Yi) is mis-specified. This estimator is valid under more general conditions, though the model-based estimator has less variance (and no asymptotic bias) when the variance model is correctly specified. It has also been shown that the empirical variance estimator can be highly biased when the number of clusters (here, N) is small.

WLS Summary

- For the linear model for correlated data, we can use WLS to get consistent and asymptotically normally distributed estimates of $\boldsymbol{\beta}$, regardless of the choice of weights, as long as the model for the mean $E(\boldsymbol{Y_i}) = \boldsymbol{X_i\beta}$ is specified correctly.
- If $\mathrm{V}(\boldsymbol{Y_i}) = \boldsymbol{\Sigma}$ is known, and we use the optimal weights $\boldsymbol{W} = \boldsymbol{\Sigma^{-1}}$, then we get $\widehat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ that has the smallest variance among all weighted least squares estimators.

WLS Summary

- If our variance model is correct but we need to estimate the optimal weights, then $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{W}}}$ is consistent and asymptotically normally distributed, and we can use the *model-based estimator* of the variance, given by

$$\hat{V}\left(\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{W}}}\right) = \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{T}}\;\widehat{\boldsymbol{\Sigma}}^{\boldsymbol{-1}}\;\boldsymbol{X}_{\boldsymbol{i}}\right)^{-1}$$

WLS Summary

- If our model for $\boldsymbol{\Sigma}$ is wrong, but we still estimate the weights $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{\Sigma}}^{-\mathbf{1}}$, then $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{W}}}$ is still consistent and asymptotically normal, but we now should use the *robust estimator* of variance, given by

$$\hat{V}\left(\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{W}}}\right) = \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{T}}\, \widehat{\boldsymbol{W}}\, \boldsymbol{X}_{\boldsymbol{i}}\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{T}} \widehat{\boldsymbol{W}} \left(\boldsymbol{Y}_{\boldsymbol{i}} - \boldsymbol{X}_{\boldsymbol{i}} \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{W}}}\right)\left(\boldsymbol{Y}_{\boldsymbol{i}} - \boldsymbol{X}_{\boldsymbol{i}} \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{W}}}\right)^{\boldsymbol{T}} \widehat{\boldsymbol{W}} \boldsymbol{X}_{i}\right) \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{T}} \widehat{\boldsymbol{W}}\, \boldsymbol{X}_{\boldsymbol{i}}\right)^{-1}$$

- This estimator of variance is also referred to as empirical or sandwich estimator of variance.

WLS Summary

- The model-based variance estimator has less variance and no asymptotic bias when the variance model is correctly specified.
- The robust variance estimator is valid under more general conditions, but is less efficient and can be highly biased when N is small.
- In GEE literature, the model for $\boldsymbol{\Sigma}$ is often called the working covariance model.
- Most GEE software uses iterations, rather than the three-step or two-step estimators described here. The desire is to get closer to the optimal weights by iterating.

Hypothesis Tests

- We can use our variance estimators to construct Wald tests of $H_0\colon \boldsymbol{L}_{r\times p}\boldsymbol{\beta}_{p\times 1} = \boldsymbol{0}_{r\times 1}$ or to construct confidence intervals.
- Under H_0

$$\sqrt{N}\boldsymbol{L}\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}}^{-1}} \rightarrow MVN_r(0, \boldsymbol{L}\boldsymbol{C}_{\boldsymbol{\Sigma}^{*-1}}\boldsymbol{L}^T)$$

so that

$$N\widehat{\boldsymbol{\beta}}^T_{\widehat{\boldsymbol{\Sigma}}^{-1}}\boldsymbol{L}^T\left(\boldsymbol{L}\widehat{\boldsymbol{C}}_{\boldsymbol{\Sigma}^{*-1}}\boldsymbol{L}^T\right)^{-1}\boldsymbol{L}\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}}^{-1}} \sim \chi^2_r$$

and we reject H_0 when this test statistic is large.

Hypothesis Tests

- We might also be interested in likelihood ratio tests (LRT), which are also based on large sample theory. When N is not that large, the LRT tends to be more reliable than the Wald test.
- Likelihood ratio tests are valid tests when comparing nested models. For example, when we test whether interactions are significant, we can think of such a test as comparing two models: a bigger model containing the interaction terms, and a smaller one without them. When hypotheses are nested this way, we can compare the likelihoods of the two models directly.
- Note: In order for the test to be valid, we should use the same covariance structure in each model, and we should also should avoid using this test of mean parameters with estimates obtained with $\widehat{\boldsymbol{\Sigma}}_{REML}$.

Hypothesis Tests

- In particular, recall that

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto \prod_{i=1}^{N} f(\boldsymbol{Y}_i | \boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})$$

- We calculate the estimated value of this likelihood under the larger ("full") model and under the smaller ("reduced") model. We can call these values

$$\hat{L}_{full} = L_{full}(\widehat{\boldsymbol{\beta}}_{full}, \widehat{\boldsymbol{\Sigma}}_{full})$$
$$\hat{L}_{reduced} = L_{reduced}(\widehat{\boldsymbol{\beta}}_{reduced}, \widehat{\boldsymbol{\Sigma}}_{reduced})$$

- As long as we use the same form for , we can conduct a likelihood ratio test of the mean parameterization of the models by comparing the likelihood ratio statistic

$$T_{LRT} = -2(log\hat{L}_{reduced} - log\hat{L}_{full})$$

to a χ^2 distribution with degrees of freedom equal to the difference in number of parameters in the two models.