STAT/BIOST 571: Homework 3

Philip Pham

February 9, 2019

Problem 1: Quasilikelihood and semiparametric methods for the general linear model (14 points)

This question examines the effect of different correlation structures, designs, and sample sizes in fitting a general linear model in a quasi-likelihood and semiparametric framework. It is also an exercise in writing code systematically; please take care to break the required programming into small tasks, and write individual functions to do each of these tasks. Please write all code "by hand", using matrix algebra and simple moment-based estimators. You may find the `mvtnorm` *package helpful.*

For the marginal model

$$E(Y_{ij}|x_{ij}) = \beta_0 + \beta_1 x_{ij},$$

consider estimation by weighted least squares, where the cluster weights are the inverse of the estimated cluster covariance matrix. Calculate quasi-likelihood standard errors as if your assumed form of the covariance matrix is known to be correct (even if, in actuality, you have assumed the wrong form of the covariance) and semi-parametric standard errors using the sandwich estimator that accounts for clustering. All of the notation follows the lecture notes.

Throughout, the following are true in the data-generating mechanism

- $\beta_0 = 0$, $\beta_1 = 0.5$
- $\boldsymbol{Y}_i|\boldsymbol{X}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{R}_i)$ *with* $\sigma^2 = 1$.

The factors that will vary are

- *The number of clusters is 15, 30, or 60*
- *The design:*
  - *Design I has* $m_i = 4$*, for all clusters. In each cluster, we see* $\{x_{i1}, x_{i2}, x_{i3}, x_{i4}\} = \{7, 10, 13, 16\}$
  - *Design II has* $m_i = 3$ *for all clusters. We see equal numbers of clusters with* $\{x_{i1}, x_{i2}, x_{i3}\} = \{7, 10, 13\}, \{7, 10, 16\}, \{7, 13, 16\},$ *or* $\{10, 13, 16\}$
- *The true covariance and the assumed covariance matrices are of the form* $\sigma^2\boldsymbol{R}_i$*:*

- *For the true covariance, consider exchangeable and exponential correlation structures, with $\rho = 0.5$ or $\rho = 0.9$ (distances between observations in the exponential model based on x_{ij}).*
- *For the assumed covariance, consider these and additionally the uncorrelated homoscedastic covariance. Any covariance parameters should be estimated using moment-based methods.*

Solution: See the results in Table 1.

Let X_i be the covariates for each cluster with a column of 1s prepended. Let y_i be the cluster responses.

We first estimate β with the ordinary least squares (OLS) estimator, $\hat{\beta}_{\text{OLS}} = (X^\intercal X)^{-1} X^\intercal y$, where X is the concatenation of the cluster covariates and y is the concatenation of the cluster responses.

We can get the covariances for $\hat{\beta}$ with

$$\operatorname{cov}\left(\hat{\beta}\right) = \left(\sum_{i=1}^{n} X_i^\intercal W X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i^\intercal W \hat{\Sigma} W X_i\right) \left(\sum_{i=1}^{n} X_i^\intercal W X_i\right)^{-1}.$$

For the different estimation methods and assumed covariances, we vary the form of $\hat{\Sigma}$ and W, both of which are assumed to be the same for all clusters.

To obtain W, we first estimate σ^2 with

$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^{n} m_i - 2} \sum_{i=1}^{n} \left(y_i - X_i \hat{\beta}_{\text{OLS}}\right)^\intercal \left(y_i - X_i \hat{\beta}_{\text{OLS}}\right). \tag{1}$$

Let the standardized OLS residuals for each cluster be $\tilde{\epsilon}_i = \frac{y_i - X_i \hat{\beta}_{\text{OLS}}}{\hat{\sigma}}$.

If we assume that the correlation structure is exchangeable, we estimate

$$\hat{\rho}_{\text{exch}} = \frac{1}{\sum_{i=1}^{n} \sum_{j=1}^{m_i - 1} j} \sum_{i=1}^{n} \sum_{j=1}^{m_i - 1} \sum_{k=j+1}^{m_i} \left(\tilde{\epsilon}_i \tilde{\epsilon}_i^\intercal\right)_{jk}, \tag{2}$$

so $W_{ii}^{-1} = 1$ for all i and $W_{ij}^{-1} = \hat{\rho}_{\text{exch}}$ for all $i \neq j$.

If we assume that the correlation structure is exponential, we estimate

$$\hat{\rho}_{\text{expon}} = \frac{1}{\sum_{i=1}^{n} \sum_{j=1}^{m_i - 1} j} \sum_{i=1}^{n} \sum_{j=1}^{m_i - 1} \left(\tilde{\epsilon}_i \tilde{\epsilon}_i^\intercal\right)_{j,j+1}, \tag{3}$$

so $W_{ij}^{-1} = \hat{\rho}_{\text{expon}}^{|j-i|}$ for any i and j.

If we don't assume any correlation structure, $W = I$.

Now that we know W, we can estimate β with

$$\hat{\beta} = \frac{1}{\sum_{i=1}^{n} m_i} \sum_{i=1}^{n} m_i \left(X_i^\intercal W X_i\right)^{-1} X_i^\intercal W y_i. \tag{4}$$

			$\mathbf{SD}(\hat{\beta}_\mathbf{1})$			$\mathbf{E}(\widehat{\mathbf{SE}}_{\mathbf{1,QL}})$			$\mathbf{E}(\widehat{\mathbf{SE}}_{\mathbf{1,sand}})$		
			Assumed Corr			Assumed Corr			Assumed Corr		
n	Design	True Corr	Uncor	Exch	Expon	Uncor	Exch	Expon	Uncor	Exch	Expon
15	I	Exchangeable $\rho = 0.5$	0.03849	0.02825	0.03773	0.03807	0.02764	0.03718	0.02596	0.02596	0.02655
15	I	Exchangeable $\rho = 0.9$	0.03849	0.01468	0.02467	0.03733	0.01393	0.0235	0.01164	0.01164	0.01223
15	I	Exponential $\rho = 0.5$	0.03849	0.03153	0.03793	0.03811	0.03108	0.03749	0.0363	0.0363	0.03601
15	I	Exponential $\rho = 0.9$	0.03849	0.01736	0.02445	0.03731	0.01642	0.0232	0.02047	0.02047	0.02027
15	II	Exchangeable $\rho = 0.5$	0.04471	0.03459	0.04045	0.04393	0.03363	0.03952	0.03453	0.03104	0.03152
15	II	Exchangeable $\rho = 0.9$	0.04465	0.01899	0.025	0.04305	0.01797	0.02369	0.02588	0.0143	0.01475
15	II	Exponential $\rho = 0.5$	0.04464	0.03646	0.04027	0.04409	0.0358	0.0396	0.03974	0.03797	0.0379
15	II	Exponential $\rho = 0.9$	0.0447	0.0204	0.02468	0.04402	0.0197	0.02385	0.02892	0.01942	0.01922
30	I	Exchangeable $\rho = 0.5$	0.02722	0.01944	0.02664	0.02719	0.01932	0.02656	0.01864	0.01864	0.019
30	I	Exchangeable $\rho = 0.9$	0.02722	0.00947	0.0161	0.02677	0.0092	0.01567	0.00839	0.00839	0.0088
30	I	Exponential $\rho = 0.5$	0.02722	0.02213	0.02684	0.02705	0.02193	0.02664	0.02634	0.02634	0.02606
30	I	Exponential $\rho = 0.9$	0.02722	0.01133	0.01593	0.02709	0.01115	0.01569	0.01457	0.01457	0.01442
30	II	Exchangeable $\rho = 0.5$	0.0315	0.02398	0.0283	0.0311	0.02357	0.02786	0.02513	0.02277	0.02304
30	II	Exchangeable $\rho = 0.9$	0.03149	0.01208	0.016	0.03091	0.01171	0.01554	0.01901	0.01029	0.01062
30	II	Exponential $\rho = 0.5$	0.03147	0.02547	0.0282	0.03126	0.02521	0.02794	0.029	0.02775	0.02766
30	II	Exponential $\rho = 0.9$	0.03158	0.01322	0.01588	0.03143	0.01302	0.01564	0.02119	0.01408	0.0139
60	I	Exchangeable $\rho = 0.5$	0.01925	0.0138	0.01895	0.01916	0.0137	0.01885	0.01339	0.01339	0.01367
60	I	Exchangeable $\rho = 0.9$	0.01925	0.00642	0.01106	0.01898	0.0063	0.01085	0.006	0.006	0.00632
60	I	Exponential $\rho = 0.5$	0.01925	0.01556	0.01894	0.01921	0.01551	0.0189	0.01891	0.01891	0.01869
60	I	Exponential $\rho = 0.9$	0.01925	0.00793	0.01099	0.01909	0.00782	0.01084	0.01057	0.01057	0.01045
60	II	Exchangeable $\rho = 0.5$	0.02226	0.01664	0.01984	0.02223	0.01657	0.01978	0.01806	0.01633	0.01653
60	II	Exchangeable $\rho = 0.9$	0.02224	0.008	0.01078	0.02201	0.00786	0.0106	0.01373	0.00734	0.00757
60	II	Exponential $\rho = 0.5$	0.02224	0.01776	0.01979	0.02212	0.01763	0.01966	0.02032	0.01946	0.01938
60	II	Exponential $\rho = 0.9$	0.02223	0.0089	0.01066	0.02208	0.00879	0.01053	0.01505	0.00997	0.00987

Table 1: Table of standard deviations of the estimated slopes and average of model-based and sandwich-based standard error estimates

By default, for the general linear model, we would have that $\hat{\Sigma} = W^{-1}$, in which case, we have that

$$\operatorname{cov}\left(\hat{\beta}\right) = \left(\sum_{i=1}^{n} X_i^\intercal W X_i\right)^{-1}. \tag{5}$$

In the quasi-likelihood model, we have an additional dispersion factor α, so $\hat{\Sigma} = \hat{\alpha} W^{-1}$, where

$$\hat{\alpha} = \frac{1}{\sum_{i=1}^{n} m_i - 2} \left(y - X\hat{\beta}\right)^\intercal \left(y - X\hat{\beta}\right), \tag{6}$$

which results in the covariance

$$\operatorname{cov}\left(\hat{\beta}\right) = \hat{\alpha}\left(\sum_{i=1}^{n} X_i^\intercal W X_i\right)^{-1}. \tag{7}$$

For the sandwich estimator, we use the empirical estimate

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - X_i\hat{\beta}\right)^\intercal \left(y_i - X_i\hat{\beta}\right). \tag{8}$$

Problem 2: Efficiency of OLS for linear models with correlated data (6 points)

Review the example on slides 2.34 – 2.35, which can also be found on pages 60 – 62 of the Diggle et al. textbook. We will generalize this example by considering the mean model

$$E(Y_{ij}) = \beta_0 + \beta_1 x_j$$

for arbitrary $\boldsymbol{x} = (x_1, x_2, \ldots, x_5)$ *that is the same for all subjects, but which may or may not be equal to* $\boldsymbol{t} = (-2, -1, 0, 1, 2)$ *(as is the case in the original version of the example). Note that the correlation structure is still determined based on* t*, as in the original example, but now the mean model contains* x *rather than* t*.*

(a) Derive a general expression for the relative efficiency of OLS compared to the optimal GLS in estimating β_0 *and* β_1 *in this problem. Your formula should be valid for a homoscedastic exponential covariance matrix with arbitrary* ρ *and for arbitrary* $\boldsymbol{x}$*. That is, derive a general version of the expressions on the bottom of 2.44 that is valid for any choice of covariates. Note that it is acceptable for your solution to be written using matrix notation and matrix algebra.*

Solution: Let there be n subjects. Each subject i has m_i observations $x_i = \begin{pmatrix} x_{i1} & x_{i2} & \ldots & x_{im_i} \end{pmatrix}^\intercal$. Let X_i be the $m_i \times 2$ covariate matrix, where $X_{ij1} = 1$ and $X_{ij2} = x_{ij}$. Let X be the $\sum_{i=1}^n m_i \times 2$ matrix obtained by stacking $X_1, X_2, \ldots, X_n$.

The true model is $Y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$. Let the obvervations between subjects be independent with each other. Let Σ_i denote the covariance structure for within-subject observations, that is $\Sigma_{ijk} = \operatorname{cov}(\epsilon_{ij}, \epsilon_{ik})$. Let Σ be the $\sum_{i=1}^n m_i \times \sum_{i=1}^n m_i$ block diagonal matrix

$$\Sigma = \begin{pmatrix} \Sigma_1 & & & \\ & \Sigma_2 & & \\ & & \ddots & \\ & & & \Sigma_n \end{pmatrix}. \tag{9}$$

Let Y_i be the vector of observations for subject i. Let Y be obtained by concatenating the $Y_1, Y_2, \ldots, Y_n$. If $\beta = \begin{pmatrix} \beta_0 & \beta_1 \end{pmatrix}^\intercal$, we can write $Y = X\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma)$.

The OLS estimate for is

$$\hat{\beta}_{\text{OLS}} = (X^\intercal X)^{-1} X^\intercal Y = \beta + (X^\intercal X)^{-1} X^\intercal \epsilon, \tag{10}$$

which has covariance

$$\begin{aligned} \operatorname{cov}\left(\hat{\beta}_{\text{OLS}}\right) &= (X^\intercal X)^{-1} X^\intercal \operatorname{cov}(\epsilon) X (X^\intercal X)^{-1} \\ &= (X^\intercal X)^{-1} X^\intercal \Sigma X (X^\intercal X)^{-1}. \end{aligned} \tag{11}$$

The optimal GLS estimate, which can be derived by maximizing likelihood is

$$\hat{\beta}_{\text{GLS}} = \left(X^\intercal \Sigma^{-1} X\right)^{-1} X^\intercal \Sigma^{-1} Y = \beta + \left(X^\intercal \Sigma^{-1} X\right)^{-1} X^\intercal \Sigma^{-1} \epsilon, \tag{12}$$

which has covariance

$$\begin{aligned}\operatorname{cov}\left(\hat{\beta}_{\mathrm{GLS}}\right) &= \left(X^{\intercal}\Sigma^{-1}X\right)^{-1} X^{\intercal}\Sigma^{-1} \operatorname{cov}(\epsilon)\, \Sigma^{-1}X \left(X^{\intercal}\Sigma^{-1}X\right)^{-1} \\ &= \left(X^{\intercal}\Sigma^{-1}X\right)^{-1}. \end{aligned} \tag{13}$$

Much of these matrix multiplications can be written as summations:

$$\begin{aligned} X^{\intercal}X &= \sum_{i=1}^{n} X_i^{\intercal}X_i \\ X^{\intercal}\Sigma X &= \sum_{i=1}^{n} X_i^{\intercal}\Sigma_i X_i \\ X^{\intercal}\Sigma^{-1}X &= \sum_{i=1}^{n} X_i^{\intercal}\Sigma_i^{-1}X_i., \end{aligned}$$

which is probably computationally faster if n is very large.
In any case, using Equations 11 and 13, we can calculate relative efficiency

$$\begin{aligned} e\left(\hat{\beta}_0\right) &= \left[\left(X^{\intercal}\Sigma^{-1}X\right)^{-1}\right]_{11} \Big/ \left[(X^{\intercal}X)^{-1} X^{\intercal}\Sigma X \,(X^{\intercal}X)^{-1}\right]_{11} \\ e\left(\hat{\beta}_1\right) &= \left[\left(X^{\intercal}\Sigma^{-1}X\right)^{-1}\right]_{22} \Big/ \left[(X^{\intercal}X)^{-1} X^{\intercal}\Sigma X \,(X^{\intercal}X)^{-1}\right]_{22} \end{aligned}$$

by taking the corresponding entries of the covariance matrix.

(b) Reproduce the lines in the table on 2.35 that pertain to β_1 for the following choices of covariate vectors

$$\begin{aligned} \boldsymbol{x} &= (-2,-1,0,1,2) \\ \boldsymbol{x} &= (-1,-2,0,2,1) \\ \boldsymbol{x} &= (0,-1,1,3,2) \\ \boldsymbol{x} &= (0,-1,1,5,2) \end{aligned}$$

(c) Explain the key differences between the relative efficiencies you just calculated. Phrase your answers in a manner that will be understandable by a quantitavely sophisticated non-statistician (e.g., an epidemiologist collaborator).